IBM SPSS Modeler 16 Quellen-, Prozess- und Ausgabeknoten

Hinweis

Vor Verwendung dieser Informationen und des darin beschriebenen Produkts sollten die Informationen unter „Bemerkungen" auf Seite 371 gelesen werden.

Produktinformation

Diese Ausgabe bezieht sich auf Version 16, Release 0, Modifikation 0 von IBM(r) SPSS(r) Modeler und alle nachfolgenden Releases und Modifikationen, bis dieser Hinweis in einer Neuausgabe geändert wird.

Diese Veröffentlichung ist eine Übersetzung des Handbuchs
IBM SPSS Modeler 16, Source, Process, and Output Nodes,
herausgegeben von International Business Machines Corporation, USA

Herausgegeben von:
TSC Germany
Kst. 2877
Oktober 2013

Inhaltsverzeichnis

Vorwort

IBM® SPSS Modeler ist die auf Unternehmensebene einsetzbare Data-Mining-Workbench von IBM. Mit SPSS Modeler können Unternehmen und Organisationen die Beziehungen zu ihren Kunden bzw. zu den Bürgern durch ein tief greifendes Verständnis der Daten verbessern. Organisationen verwenden die mithilfe von SPSS Modeler gewonnenen Erkenntnisse zur Bindung profitabler Kunden, zur Ermittlung von Cross-Selling-Möglichkeiten, zur Gewinnung neuer Kunden, zur Ermittlung von Betrugsfällen, zur Reduzierung von Risiken und zur Verbesserung der Verfügbarkeit öffentlicher Dienstleistungen.

Die visuelle Benutzerschnittstelle von SPSS Modeler erleichtert die Anwendung des spezifischen Fachwissens der Benutzer, was zu leistungsstärkeren Vorhersagemodellen führt und die Zeit bis zur Lösungserstellung verkürzt. SPSS Modeler bietet zahlreiche Modellierungsverfahren, beispielsweise Algorithmen für Vorhersage, Klassifizierung, Segmentierung und Assoziationserkennung. Nach der Modellerstellung ermöglicht IBM SPSS Modeler Solution Publisher die unternehmensweite Bereitstellung des Modells für Entscheidungsträger oder in einer Datenbank.

Informationen zu IBM Business Analytics

Die Software IBM Business Analytics liefert umfassende, einheitliche und korrekte Informationen, mit denen Entscheidungsträger die Unternehmensleistung verbessern können. Ein umfassendes Portfolio aus Anwendungen für Business Intelligence, Vorhersageanalyse, Finanz- und Strategiemanagement sowie Analysen bietet Ihnen sofort klare und umsetzbare Einblicke in die aktuelle Leistung und gibt Ihnen die Möglichkeit, zukünftige Ergebnisse vorherzusagen. Durch umfassende Branchenlösungen, bewährte Vorgehensweisen und professionellen Service können Unternehmen jeder Größe die Produktivität maximieren, Entscheidungen automatisieren und bessere Ergebnisse erzielen.

Als Teil dieses Portfolios unterstützt IBM SPSS Predictive Analytics-Software Unternehmen dabei, zukünftige Ereignisse vorherzusagen und proaktiv Maßnahmen zu ergreifen, um bessere Geschäftsergebnisse zu erzielen. Kunden aus Wirtschaft, öffentlichem Dienst und staatlichen Lehr- und Forschungseinrichtungen weltweit nutzen IBM SPSS-Technologie als Wettbewerbsvorteil für die Kundengewinnung, Kundenbindung und Erhöhung der Kundenumsätze bei gleichzeitiger Eindämmung der Betrugsmöglichkeiten und Minderung von Risiken. Durch die Einbindung von IBM SPSS-Software in ihre täglichen Operationen werden Organisationen zu "Predictive Enterprises", die Entscheidungen auf Geschäftsziele ausrichten und automatisieren und einen messbaren Wettbewerbsvorteil erzielen können. Wenn Sie weitere Informationen wünschen oder Kontakt zu einem Mitarbeiter aufnehmen möchten, besuchen Sie die Seite http://www.ibm.com/spss.

Technical Support

Kunden mit Wartungsvertrag können den Technical Support in Anspruch nehmen. Kunden können sich an den Technical Support wenden, wenn sie Hilfe bei der Arbeit mit IBM Produkten oder bei der Installation in einer der unterstützten Hardwareumgebungen benötigen. Zur Kontaktaufnahme mit dem Technical Support besuchen Sie die IBM Website unter http://www.ibm.com/support. Sie müssen bei der Kontaktaufnahme Ihren Namen, Ihre Organisation und Ihre Supportvereinbarung angeben.

Kapitel 1. Informationen zu IBM SPSS Modeler

IBM SPSS Modeler ist ein Set von Data-Mining-Tools, mit dem Sie auf der Grundlage Ihres Fachwissens schnell und einfach Vorhersagemodelle erstellen und zur Erleichterung der Entscheidungsfindung in die Betriebsabläufe einbinden können. Das Produkt IBM SPSS Modeler, das auf der Grundlage des den Industrienormen entsprechenden Modells CRISP-DM entwickelt wurde, unterstützt den gesamten Data Mining-Prozess, von den Daten bis hin zu besseren Geschäftsergebnissen.

IBM SPSS Modeler bietet eine Vielzahl von Modellbildungsmethoden, die aus dem maschinellen Lernen, der künstlichen Intelligenz und der Statistik stammen. Mit den in der Modellierungspalette verfügbaren Methoden können Sie aus Ihren Daten neue Informationen ableiten und Vorhersagemodelle erstellen. Jede Methode hat ihre speziellen Stärken und eignet sich besonders für bestimmte Problemtypen.

SPSS Modeler kann als Standalone-Produkt oder als Client in Verbindung mit SPSS Modeler Server erworben werden. Außerdem ist eine Reihe von Zusatzoptionen verfügbar, die in den folgenden Abschnitten kurz zusammengefasst werden. Weitere Informationen finden Sie unter http://www.ibm.com/software/analytics/spss/products/modeler/.

IBM SPSS Modeler-Produkte

Zur IBM SPSS Modeler-Produktfamilie und der zugehörigen Software gehören folgende Elemente.

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Modeler Server-Adapter für IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler ist eine funktionell in sich abgeschlossene Produktversion, die Sie auf Ihrem PC installieren und ausführen können. Sie können SPSS Modeler im lokalen Modus als Standalone-Produkt oder im verteilten Modus zusammen mit IBM SPSS Modeler Server verwenden, um bei Datasets die Leistung zu verbessern.

Mit SPSS Modeler können Sie schnell und intuitiv genaue Vorhersagemodelle erstellen, und das ohne Programmierung. Mithilfe der speziellen visuellen Benutzerschnittstelle können Sie den Data Mining-Prozess auf einfache Weise visualisieren. Mit der Unterstützung der in das Produkt eingebetteten erweiterten Analyseprozesse können Sie zuvor verborgene Muster und Trends in Ihren Daten aufdecken. Sie können Ergebnisse modellieren und Einblick in die Faktoren gewinnen, die Einfluss auf diese Ergebnisse haben, wodurch Sie in die Lage versetzt werden, Geschäftschancen zu nutzen und Risiken zu mindern.

SPSS Modeler ist in zwei Editionen erhältlich: SPSS Modeler Professional und SPSS Modeler Premium. Weitere Informationen finden Sie unter im Thema „IBM SPSS Modeler-Editionen" auf Seite 2.

IBM SPSS Modeler Server

SPSS Modeler verwendet eine Client/Server-Architektur zur Verteilung von Anforderungen für ressourcenintensive Vorgänge an leistungsstarke Serversoftware, wodurch bei größeren Datasets eine höhere Leistung erzielt werden kann.

SPSS Modeler Server ist ein separat lizenziertes Produkt, das durchgehend im verteilten Analysemodus auf einem Server-Host in Verbindung mit einer oder mehreren IBM SPSS Modeler-Installationen ausgeführt wird. Auf diese Weise bietet SPSS Modeler Server eine herausragende Leistung bei großen Datasets, da speicherintensive Vorgänge auf dem Server ausgeführt werden können, ohne Daten auf den Client-Computer herunterladen zu müssen. IBM SPSS Modeler Server bietet außerdem Unterstützung für SQL-Optimierung sowie Möglichkeiten zur Modellierung innerhalb der Datenbank, was weitere Vorteile hinsichtlich Leistung und Automatisierung mit sich bringt.

IBM SPSS Modeler Administration Console

Modeler Administration Console ist eine grafische Anwendung zur Verwaltung einer Vielzahl der SPSS Modeler Server-Konfigurationsoptionen, die auch mithilfe einer Optionsdatei konfiguriert werden können. Die Anwendung bietet eine Konsolenbenutzerschnittstelle zur Überwachung und Konfiguration der SPSS Modeler Server-Installationen und steht aktuellen SPSS Modeler Server-Kunden kostenlos zur Verfügung. Die Anwendung kann nur unter Windows installiert werden. Der von ihr verwaltete Server kann jedoch auf einer beliebigen unterstützten Plattform installiert sein.

IBM SPSS Modeler Batch

Das Data-Mining ist zwar in der Regel ein interaktiver Vorgang, es ist jedoch auch möglich, SPSS Modeler über eine Befehlszeile auszuführen, ohne dass die grafische Benutzerschnittstelle verwendet werden muss. Beispielsweise kann es sinnvoll sein, langwierige oder sich wiederholende Aufgaben ohne Eingreifen des Benutzers durchzuführen. SPSS Modeler Batch ist eine spezielle Version des Produkts, die die vollständigen Analysefunktionen von SPSS Modeler ohne Zugriff auf die reguläre Benutzerschnittstelle bietet. Zur Verwendung von SPSS Modeler Batch ist eine SPSS Modeler Server-Lizenz erforderlich.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher ist ein Tool, mit dem Sie eine gepackte Version eines SPSS Modeler-Streams erstellen können, der durch eine externe Runtime-Engine ausgeführt oder in eine externe Anwendung eingebettet werden kann. Auf diese Weise können Sie vollständige SPSS Modeler-Streams für die Verwendung in Umgebungen veröffentlichen und bereitstellen, in denen SPSS Modeler nicht installiert ist. SPSS Modeler Solution Publisher wird als Teil des Diensts für IBM SPSS Collaboration and Deployment Services - Scoring verteilt, für den eine separate Lizenz erforderlich ist. Mit dieser Lizenz erhalten Sie SPSS Modeler Solution Publisher Runtime, womit Sie die veröffentlichten Streams ausführen können.

IBM SPSS Modeler Server-Adapter für IBM SPSS Collaboration and Deployment Services

Für IBM SPSS Collaboration and Deployment Services ist eine Reihe von Adaptern verfügbar, mit denen SPSS Modeler und SPSS Modeler Server mit einem IBM SPSS Collaboration and Deployment Services-Repository interagieren können. Auf diese Weise kann ein im Repository bereitgestellter SPSS Modeler-Stream von mehreren Benutzern gemeinsam verwendet werden. Auch der Zugriff über die Thin-Client-Anwendung IBM SPSS Modeler Advantage ist möglich. Sie installieren den Adapter auf dem System, das als Host für das Repository fungiert.

IBM SPSS Modeler-Editionen

SPSS Modeler ist in den folgenden Editionen erhältlich.

SPSS Modeler Professional

SPSS Modeler Professional bietet sämtliche Tools, die Sie für die Arbeit mit den meisten Typen von strukturierten Daten benötigen, beispielsweise in CRM-Systemen erfasste Verhaltensweisen und Interaktionen, demografische Daten, Kaufverhalten und Umsatzdaten.

SPSS Modeler Premium

SPSS Modeler Premium ist ein separat lizenziertes Produkt, das SPSS Modeler Professional für die Arbeit mit spezialisierten Daten, wie beispielsweise Daten, die für Entitätsanalysen oder soziale Netze verwendet werden, sowie für die Arbeit mit unstrukturierten Textdaten erweitert. SPSS Modeler Premium umfasst die folgenden Komponenten.

IBM SPSS Modeler Entity Analytics adds an extra dimension to IBM SPSS Modeler predictive analytics. Whereas predictive analytics attempts to predict future behavior from past data, entity analytics focuses on improving the coherence and consistency of current data by resolving identity conflicts within the records themselves. An identity can be that of an individual, an organization, an object, or any other entity for which ambiguity might exist. Identity resolution can be vital in a number of fields, including customer relationship management, fraud detection, anti-money laundering, and national and international security.

IBM SPSS Modeler Social Network Analysis transformiert Informationen zu Beziehungen in Felder, die das Sozialverhalten von Einzelpersonen und Gruppen charakterisieren. Durch die Verwendung von Daten, die die Beziehungen beschreiben, die sozialen Netzen zugrunde liegen, ermittelt IBM SPSS Modeler Social Network Analysis Führungskräfte in sozialen Netzen, die das Verhalten anderer Personen im Netz beeinflussen. Außerdem können Sie feststellen, welche Personen am meisten durch andere Teilnehmer im Netz beeinflusst werden. Durch die Kombination dieser Ergebnisse mit anderen Maßen können Sie aussagekräftige Profile für Einzelpersonen erstellen, die Sie als Grundlage für Ihre Vorhersagemodelle verwenden können. Modelle, die diese sozialen Informationen berücksichtigen, sind leistungsstärker als Modelle, die dies nicht tun.

IBM SPSS Modeler Text Analytics verwendet hoch entwickelte linguistische Technologien und die Verarbeitung natürlicher Sprache (Natural Language Processing, NLP), um eine schnelle Verarbeitung einer großen Vielfalt an unstrukturierten Textdaten zu ermöglichen, um die Schlüsselkonzepte zu extrahieren und zu ordnen und um diese Konzepte in Kategorien zusammenzufassen. Extrahierte Konzepte und Kategorien können mit bestehenden strukturierten Daten, beispielsweise demografischen Informationen, kombiniert und mithilfe der vollständigen Suite der Data-Mining-Tools von IBM SPSS Modeler auf die Modellierung angewendet werden, um bessere und fokussiertere Entscheidungen zu ermöglichen.

IBM SPSS Modeler-Dokumentation

Eine Dokumentation im Onlinehilfeformat finden Sie im Hilfemenü von SPSS Modeler. Diese umfasst die Dokumentation für SPSS Modeler, SPSS Modeler Server und SPSS Modeler Solution Publisher sowie das Anwendungshandbuch und weiteres Material zur Unterstützung.

Die vollständige Dokumentation für die einzelnen Produkte (einschließlich Installationsanweisungen) steht im PDF-Format im Ordner *Documentation* auf der jeweiligen Produkt-DVD zur Verfügung. Installationsdokumente können auch aus dem Internet unter http://www-01.ibm.com/support/docview.wss?uid=swg27038316 heruntergeladen werden.

Dokumentation in beiden Formaten steht auch im SPSS Modeler Information Center unter http://publib.boulder.ibm.com/infocenter/spssmodl/v16r0m0/ zur Verfügung.

SPSS Modeler Professional-Dokumentation

Die SPSS Modeler Professional-Dokumentationssuite (ohne Installationsanweisungen) umfasst folgende Dokumente:

- **IBM SPSS Modeler Benutzerhandbuch.** Allgemeine Einführung in die Verwendung von SPSS Modeler, in der u. a. die Erstellung von Datenstreams, der Umgang mit fehlenden Werten, die Erstellung von CLEM-Ausdrücken, die Arbeit mit Projekten und Berichten sowie das Packen von Streams für die Bereitstellung in IBM SPSS Collaboration and Deployment Services, Predictive Applications (Vorhersageanwendungen) oder IBM SPSS Modeler Advantage beschrieben werden.

- **IBM SPSS Modeler Quellen-, Prozess- und Ausgabeknoten.** Beschreibung aller Knoten, die zum Lesen, zum Verarbeiten und zur Ausgabe von Daten in verschiedenen Formaten verwendet werden. Im Grunde sind sie alle Knoten, mit Ausnahme der Modellierungsknoten.
- **IBM SPSS Modeler Modellierungsknoten.** Beschreibungen sämtlicher für die Erstellung von Data-Mining-Modellen verwendeter Knoten. IBM SPSS Modeler bietet eine Vielzahl von Modellbildungsmethoden, die aus dem maschinellen Lernen, der künstlichen Intelligenz und der Statistik stammen.
- **IBM SPSS Modeler Algorithms Guide.** Beschreibung der mathematischen Grundlagen der in IBM SPSS Modeler verwendeten Modellierungsmethoden. Dieses Handbuch steht nur im PDF-Format zur Verfügung.
- **IBM SPSS Modeler Anwendungshandbuch.** Die Beispiele in diesem Handbuch bieten eine kurze, gezielte Einführung in bestimmte Modellierungsmethoden und -verfahren. Eine Online-Version dieses Handbuchs kann auch über das Hilfemenü aufgerufen werden. Weitere Informationen finden Sie unter im Thema „Anwendungsbeispiele" auf Seite 5.
- **IBM SPSS Modeler Handbuch für Scripterstellung und Automatisierung.** Informationen zur Automatisierung des Systems über Scripterstellung, einschließlich der Eigenschaften, die zur Bearbeitung von Knoten und Streams verwendet werden können.
- IBM SPSS Modeler Bereitstellungshandbuch. Informationen zum Ausführen von IBM SPSS Modeler-Streams und -Szenarios als Schritte bei der Verarbeitung von Jobs im IBM SPSS Collaboration and Deployment Services Deployment Manager.
- **IBM SPSS Modeler CLEF-Entwicklerhandbuch.** CLEF bietet die Möglichkeit, Drittanbieterprogramme, wie Datenverarbeitungsroutinen oder Modellierungsalgorithmen, als Knoten in IBM SPSS Modeler zu integrieren.
- **IBM SPSS Modeler Datenbankinternes Mining.** Informationen darüber, wie Sie Ihre Datenbank dazu einsetzen, die Leistung zu verbessern, und wie Sie die Palette der Analysefunktionen über Drittanbieteralgorithmen erweitern.
- **IBM SPSS Modeler Server Verwaltungs- und Leistungshandbuch.** Informationen zur Konfiguration und Verwaltung von IBM SPSS Modeler Server.
- **IBM SPSS Modeler Administration Console Benutzerhandbuch.** Informationen zur Installation und Nutzung der Konsolenbenutzerschnittstelle zur Überwachung und Konfiguration von IBM SPSS Modeler Server. Die Konsole ist als Plug-in für die Deployment Manager-Anwendung implementiert.
- IBM SPSS Modeler CRISP-DM Handbuch. Schritt-für-Schritt-Anleitung für das Data Mining mit SPSS Modeler unter Verwendung der CRISP-DM-Methode.
- IBM SPSS Modeler Batch Benutzerhandbuch. Vollständiges Handbuch für die Verwendung von IBM SPSS Modeler im Stapelmodus, einschließlich Details zur Ausführung des Stapelmodus und zu Befehlszeilenargumenten. Dieses Handbuch steht nur im PDF-Format zur Verfügung.

SPSS Modeler Premium-Dokumentation

Die SPSS Modeler Premium-Dokumentationssuite (ohne Installationsanweisungen) umfasst folgende Dokumente:

- IBM SPSS Modeler Entity Analytics Benutzerhandbuch. Informationen zur Verwendung von Entitätsanalysen mit SPSS Modeler, unter Behandlung der Repository-Installation und -Konfiguration, Entity Analytics-Knoten und Verwaltungsaufgaben.
- **IBM SPSS Modeler Social Network Analysis User Guide.** Ein Handbuch zur Durchführung einer sozialen Netzanalyse mit SPSS Modeler, einschließlich einer Gruppenanalyse und Diffusionsanalyse.
- SPSS Modeler Text Analytics Benutzerhandbuch. Informationen zur Verwendung von Textanalysen mit SPSS Modeler, unter Behandlung der Text Mining-Knoten, der interaktiven Workbench sowie von Vorlagen und anderen Ressourcen.
- IBM SPSS Modeler Text Analytics Administration Console Benutzerhandbuch. Informationen zur Installation und Nutzung der Konsolenbenutzerschnittstelle zur Überwachung und Konfiguration von IBM SPSS Modeler Server für die Verwendung mit SPSS Modeler Text Analytics . Die Konsole ist als Plug-in für die Deployment Manager-Anwendung implementiert.

Anwendungsbeispiele

Mit den Data-Mining-Tools in SPSS Modeler kann eine große Bandbreite an geschäfts- und unternehmensbezogenen Problemen gelöst werden; die Anwendungsbeispiele dagegen bieten jeweils eine kurze, gezielte Einführung in spezielle Modellierungsmethoden und -verfahren. Die hier verwendeten Datasets sind viel kleiner als die großen Datenbestände, die von einigen Data-Mining-Experten verwaltet werden müssen, die zugrunde liegenden Konzepte und Methoden sollten sich jedoch auch auf reale Anwendungen übertragen lassen.

Sie können auf die Beispiele zugreifen, indem Sie im Menü "Hilfe" in SPSS Modeler auf die Option **Anwendungsbeispiele** klicken. Die Datendateien und Beispielstreams wurden im Ordner *Demos*, einem Unterordner des Produktinstallationsverzeichnisses, installiert. Weitere Informationen finden Sie unter im Thema „Ordner "Demos"".

Beispiele für die Datenbankmodellierung. Die Beispiele finden Sie im IBM SPSS Modeler-Handbuch zum datenbankinternen Mining.

Scriptbeispiele. Die Beispiele finden Sie im *IBM SPSS Modeler Handbuch für Scripterstellung und Automatisierung*.

Ordner "Demos"

Die in den Anwendungsbeispielen verwendeten Datendateien und Beispielstreams wurden im Ordner *Demos*, einem Unterordner des Produktinstallationsverzeichnisses, installiert. Auf diesen Ordner können Sie auch über die Programmgruppe IBM SPSS Modeler im Windows-Startmenü oder durch Klicken auf *Demos* in der Liste der zuletzt angezeigten Verzeichnisse im Dialogfeld "Datei öffnen" zugreifen.

Kapitel 2. Quellenknoten

Übersicht

Mithilfe von Quellenknoten können Sie Daten importieren, die in einer Reihe von Formaten gespeichert sind, darunter Flatfiles, IBM SPSS Statistics (*.sav*), SAS, Microsoft Excel und ODBC-kompatible relationale Datenbanken. Mit dem Benutzereingabeknoten können Sie außerdem künstliche Daten generieren.

Die Palette "Datenquellen" enthält folgende Knoten:

Der Enterprise-Ansichtsknoten erstellt eine Verbindung mit einem IBM SPSS Collaboration and Deployment Services Repository, was es Ihnen ermöglicht, Enterprise-Ansichtsdaten in einen Stream einzulesen und ein Modell in ein Szenario zu packen, auf das andere Benutzer über das Repository zugreifen können. Weitere Informationen finden Sie im Thema „Enterprise-Ansichtsknoten" auf Seite 8.

Mit dem Datenansichtsknoten können Sie auf Datenquellen zugreifen, die in Analysedatenansichten von IBM SPSS Collaboration and Deployment Services definiert sind. Eine Analysedatenansicht definiert eine Standardschnittstelle für den Datenzugriff und ordnet dieser Schnittstelle mehrere physische Datenquellen zu. Weitere Informationen finden Sie im Thema „Datenansichtsknoten" auf Seite 11.

Mit dem Datenbankknoten lassen sich Daten aus einer Reihe von anderen Paketen importieren, die ODBC (Open Database Connectivity) verwenden, darunter u. a. Microsoft SQL Server, DB2 und Oracle. Weitere Informationen finden Sie im Thema „Datenbankquellenknoten" auf Seite 13.

Der Variablendateiknoten liest Daten aus Textdateien mit freien Feldern, also aus Dateien, deren Datensätze eine konstante Anzahl von Feldern, aber eine variable Anzahl von Zeichen enthalten. Dieser Knoten ist außerdem nützlich für Dateien mit fester Länge, Überschriftentext und bestimmten Anmerkungen. Weitere Informationen finden Sie im Thema „Knoten "Variable Datei"" auf Seite 20.

Der Knoten des Typs "Datei (fest)" importiert Daten aus Textdateien mit festen Feldern, also aus Dateien, deren Felder nicht begrenzt sind, sondern an derselben Position beginnen und eine feste Länge haben. Maschinell erzeugte Daten oder vorhandene Daten werden häufig im Format mit festen Feldern gespeichert. Weitere Informationen finden Sie im Thema „Knoten "Datei (fest)"" auf Seite 23.

Der Statistikdateiknoten liest Daten aus dem Dateiformat *.sav* oder *.zsav* ein, das von IBM SPSS Statistics verwendet wird, sowie in IBM SPSS Modeler gespeicherte Cachedateien, die ebenfalls dasselbe Format verwenden.

Der IBM SPSS Data Collection-Knoten importiert Daten aus zahlreichen in der Marktforschungssoftware verwendeten Formaten und passt sie dem IBM SPSS Data Collection-Datenmodell an. Um diesen Knoten verwenden zu können, muss die IBM SPSS Data Collection Developer Library installiert sein. Weitere Informationen finden Sie im Thema „Data Collection-Knoten" auf Seite 26.

Der IBM Cognos BI-Quellenknoten importiert Daten aus Cognos BI-Datenbanken.

Der IBM Cognos TM1-Quellenknoten importiert Daten aus Cognos TM1-Datenbanken.

Der SAS-Dateiknoten importiert SAS-Daten in IBM SPSS Modeler. Weitere Informationen finden Sie im Thema „SAS-Quellenknoten” auf Seite 41.

Der Excel-Knoten importiert Daten aus einer beliebigen Version von Microsoft Excel. Es ist keine ODBC-Datenquelle erforderlich. Weitere Informationen finden Sie im Thema „Excel-Quellenknoten” auf Seite 42.

Der XML-Quellenknoten importiert Daten im XML-Format in den Stream. Sie können eine einzelne Datei oder alle Dateien in einem Verzeichnis importieren. Optional können Sie eine Schemadatei angeben, aus der die XML-Struktur gelesen werden soll.

Der Simulationsgenerierungsknoten bietet eine einfache Möglichkeit, simulierte Daten entweder völlig neu anhand von durch den Benutzer angegebenen statistischen Verteilungen oder automatisch anhand der Verteilungen aus der Ausführung eines Simulationsanpassungsknotens für vorhandene historische Daten zu generieren. Dies ist hilfreich, wenn Sie das Ergebnis eines Vorhersagemodells bei einer Unsicherheit in den Modelleingaben auswerten möchten.

Der Benutzereingabeknoten bietet eine einfache Möglichkeit, künstliche Daten zu erstellen. Dazu können entweder neue Daten ohne Vorlage erstellt oder vorhandene Daten geändert werden. Diese Funktion ist nützlich, wenn Sie z. B. ein Testdataset für die Modellierung erstellen möchten. Weitere Informationen finden Sie im Thema „Benutzereingabeknoten” auf Seite 58.

Um mit dem Erstellen eines Streams zu beginnen, fügen Sie einen Quellenknoten zum Streamerstellungsbereich hinzu. Doppelklicken Sie dann auf den Knoten, um das zugehörige Dialogfeld zu öffnen. Auf den einzelnen Registerkarten im Dialogfeld können Sie Daten einlesen, Felder und Werte anzeigen und eine Vielzahl von Optionen festlegen, wie Filter, Datentypen, Feldrolle und die Überprüfung fehlender Werte.

Enterprise-Ansichtsknoten

Mit dem Enterprise-Ansichtsknoten können Sie eine Verbindung zwischen einer IBM SPSS Modeler-Sitzung und einer Enterprise-Ansicht in einem freigegebenen IBM SPSS Collaboration and Deployment Services Repository herstellen und aufrechterhalten. Dadurch können Sie Daten aus einer Enterprise-Ansicht in einen IBM SPSS Modeler-Stream einlesen und ein IBM SPSS Modeler-Modell in ein Szenario packen, auf das andere Benutzer des gemeinsam genutzten Repository Zugriff haben.

Ein **Szenario** ist eine Datei, die einen IBM SPSS Modeler-Stream mit bestimmten Knoten, Modellen und weiteren Eigenschaften enthält, die es möglich machen, eine Bereitstellung des Streams in einem IBM SPSS Collaboration and Deployment Services Repository vorzunehmen, um ein Scoring oder eine automatisierte Modellaktualisierung durchzuführen. Die Verwendung von Enterprise-Ansichtsknoten mit Szenarios gewährleistet, dass bei einer Konstellation mit mehreren Benutzern alle Benutzer auf der Grundlage derselben Daten arbeiten. Eine **Verbindung** ist eine Verknüpfung von einer IBM SPSS Modeler-Sitzung zu einer Enterprise-Ansicht im IBM SPSS Collaboration and Deployment Services Repository.

Die **Enterprise-Ansicht** ist die Gesamtmenge der Daten, die zu einer Organisation gehören, unabhängig davon, wo sich diese Daten physisch befinden. Jede Verbindung besteht aus einer Auswahl einer einzelnen **Anwendungsansicht** (Subset der Enterprise-Ansicht, die auf eine bestimmte Anwendung zugeschnitten ist), einer **Datenproviderdefinition** (DPD - stellt eine Verknüpfung zwischen den logischen Tabellen und Spalten der Anwendungsansicht und einer physischen Datenquelle her) und einer **Umgebung** (gibt an, welche Spalten jeweils den definierten Geschäftssegmenten zugeordnet werden sollen). Enterprise-Ansicht, Anwendungsansicht und PDP-Definitionen sind im Repository gespeichert (dort findet auch die Versionsverwaltung statt). Die tatsächlichen Daten befinden sich jedoch in einer oder mehreren Datenbanken oder in anderen externen Quellen.

Sobald eine Verbindung hergestellt wurde, geben Sie eine **Anwendungsansichtstabelle** für die Arbeit in IBM SPSS Modeler an. In einer Anwendungsansicht ist eine Tabelle eine logische Ansicht, die aus einigen oder allen Spalten aus einer oder mehreren physischen Tabellen in einer oder mehreren physischen Datenbanken besteht. So ermöglicht der Enterprise-Ansichtsknoten, dass Datensätze aus mehreren Datenbanktabellen in IBM SPSS Modeler als eine einzige Tabelle angezeigt werden.

Voraussetzungen

- Um den Enterprise-Ansichtsknoten verwenden zu können, muss zuvor IBM SPSS Collaboration and Deployment Services Repository an Ihrem Standort installiert und konfiguriert sein. Dabei müssen bereits eine Enterprise-Ansicht, Anwendungsansichten und DPDs definiert sein.

  Hinweis: Für den Zugriff auf ein IBM SPSS Collaboration and Deployment Services-Repository ist eine separate Lizenz erforderlich. Weitere Informationen finden Sie im Dokument http://www.ibm.com/software/analytics/spss/products/deployment/cds/
- Außerdem muss auf jedem Computer, der zur Bearbeitung oder Ausführung des Streams verwendet wird, der IBM SPSS Collaboration and Deployment Services Enterprise View Driver installiert sein. Unter Windows installieren Sie den Treiber einfach auf dem Computer, auf dem IBM SPSS Modeler bzw. IBM SPSS Modeler Server installiert ist. Es ist keine weitere Konfiguration des Treibers erforderlich. Unter UNIX muss ein Verweis auf das Script *pev.sh* zum Startscript hinzugefügt werden. Details zur Installation von IBM SPSS Collaboration and Deployment Services Enterprise View Driver erhalten Sie von Ihrem lokalen Administrator.
- Eine DPD wird anhand einer bestimmten ODBC-Datenquelle definiert. Um eine DPD aus IBM SPSS Modeler zu verwenden, muss eine ODBC-Datenquelle auf dem IBM SPSS Modeler Server-Host definiert sein, der denselben Namen trägt und der eine Verbindung zu demselben Datenspeicher herstellt wie die in der DPD referenzierte Datenquelle.

Festlegen der Optionen für den Enterprise-Ansichtsknoten

Mit den Optionen auf der Registerkarte "Daten" des Dialogfelds "Enterprise-Ansicht" haben Sie folgende Möglichkeiten:

- Auswahl einer bestehenden Repository-Verbindung
- Bearbeiten einer bestehenden Repository-Verbindung
- Erstellen einer neuen Repository-Verbindung
- Auswahl einer Anwendungsansichtstabelle

Einzelheiten zur Arbeit mit Repositorys finden Sie im *IBM SPSS Collaboration and Deployment Services-Administratorhandbuch*.

Verbindung. Die Dropdown-Liste bietet Optionen zur Auswahl einer bestehenden Repository-Verbindung, zum Bearbeiten einer bestehenden Repository-Verbindung bzw. zum Hinzufügen einer Verbindung. Wenn Sie bereits über IBM SPSS Modeler an einem Repository angemeldet sind, wird durch Auswahl der Option **Verbindung hinzufügen/bearbeiten** das Dialogfeld "Enterprise-Ansichtsverbindungen" angezeigt. In diesem Dialogfeld können Sie die erforderlichen Details für die aktuelle Verbindung definieren bzw. bearbeiten. Wenn Sie nicht angemeldet sind, zeigt diese Option das Anmeldedialogfeld für das Repository an.

Informationen zur Anmeldung im Repository finden Sie im *IBM SPSS Modeler-Benutzerhandbuch*.

Sobald eine Verbindung zu einem Repository hergestellt wurde, bleibt diese Verbindung erhalten, bis Sie IBM SPSS Modeler beenden. Eine Verbindung kann für andere Knoten innerhalb desselben Streams freigegeben werden; Sie müssen jedoch für jeden neuen Stream eine neue Verbindung erstellen.

Bei erfolgreicher Anmeldung wird das Dialogfeld "Enterprise-Ansichtsverbindungen" angezeigt.

Tabellenname. Dieses Feld ist ursprünglich leer und kann erst nach der Herstellung einer Verbindung mit Daten versehen werden. Wenn Ihnen der Name der Anwendungsansichtstabelle, auf die Sie zugreifen möchten, bekannt ist, geben Sie ihn in das Feld "Tabellenname" ein. Klicken Sie andernfalls auf die Schaltfläche **Auswählen**, um ein Dialogfeld mit einer Liste der verfügbaren Anwendungsansichtstabellen zu öffnen.

Enterprise-Ansichtsverbindungen

In diesem Dialogfeld können Sie die erforderlichen Details für die Repository-Verbindung definieren bzw. bearbeiten. Sie können folgende Elemente angeben:

- Anwendungsansicht und Version
- Umgebung
- Datenproviderdefinition (DPD)
- Verbindungsbeschreibung

Verbindungen. Listet bestehende Repository-Verbindungen auf.

- **Neue Verbindung hinzufügen.** Zeigt das Dialogfeld "Objekt abrufen" an, in dem Sie nach einer Anwendungsansicht aus dem Repository suchen und diese auswählen können.
- **Ausgewählte Verbindung kopieren.** Erstellt eine Kopie einer ausgewählten Verbindung, sodass Sie nicht erneut zu derselben Anwendungsansicht blättern müssen.
- **Ausgewählte Verbindung löschen.** Löscht die ausgewählte Verbindung aus der Liste.

Verbindungsdetails. Zeigt für die aktuell im Fenster "Verbindungen" ausgewählte Verbindung die Anwendungsansicht, die Versionsbeschriftung, die Umgebung, DPD sowie einen beschreibenden Text an.

- **Anwendungsansicht.** In der Dropdown-Liste wird gegebenenfalls die ausgewählte Anwendungsansicht angezeigt. Wenn in der aktuellen Sitzung Verbindungen zu anderen Anwendungsansichten hergestellt wurden, werden diese ebenfalls in der Dropdown-Liste angezeigt. Klicken Sie auf die angrenzende Schaltfläche "Durchsuchen", um nach anderen Anwendungsansichten im Repository zu suchen.
- **Versionsbeschriftung.** Im Dropdown-Feld werden alle definierten Versionsbeschriftungen für die angegebene Anwendungsansicht aufgeführt. Die Versionsbeschriftungen erleichtern die Kennzeichnung bestimmter Versionen von Repository-Objekten. Beispielsweise kann es zwei Versionen einer bestimmten Anwendungsansicht geben. Bei Verwendung von Beschriftungen können Sie beispielsweise die Beschriftung **TEST** für die Version angeben, die in der Entwicklungsumgebung verwendet wird, und die Beschriftung **PRODUKTION** für die in der Produktionsumgebung verwendete Version. Wählen Sie eine geeignete Beschriftung aus.

Hinweis: Beschriftungen sollten das Zeichen "[" nicht enthalten, da sonst der Tabellenname nicht auf der Registerkarte "Daten" im Dialogfeld "Enterprise-Ansicht" angezeigt wird.

- **Umgebung.** Im Dropdown-Feld werden alle gültigen Umgebungen aufgelistet. Die Umgebungseinstellung bestimmt, welche DPDs verfügbar sind, gibt also an, welche Spalten definierten Geschäftssegmenten zugeordnet werden sollen. Bei Auswahl von **Analytisch** beispielsweise, werden nur die Spalten der Anwendungsansicht zurückgegeben, die als **Analytisch** definiert sind. Die Standardumgebung lautet **Analytisch**; Sie können jedoch auch **Betrieb** auswählen.
- **Datenprovider.** In der Dropdown-Liste werden die Namen von bis zu zehn Datenproviderdefinitionen für die ausgewählte Anwendungsansicht aufgeführt. Nur DPDs, die auf die ausgewählte Anwendungsansicht verweisen, werden angezeigt. Klicken Sie auf die angrenzende Schaltfläche "Durchsuchen", um Namen und Pfad aller DPDs anzuzeigen, die sich auf die aktuelle Anwendungsansicht beziehen.
- **Beschreibung.** Beschreibender Text zur Repository-Verbindung. Dieser Text wird als Verbindungsname verwendet. Beim Klicken auf **OK** wird der Text in der Dropdown-Liste "Verbindung" und in der Titelleiste des Dialogfelds "Enterprise-Ansicht" sowie als Beschriftung des Enterprise-Ansichtsknotens im Erstellungsbereich angezeigt.

Auswählen der DPD

Im Dialogfeld "Datenprovider auswählen" werden Name und Pfad aller DPDs angezeigt, die auf die aktuelle Anwendungsansicht verweisen.

Anwendungsansichten können mehrere DPDs aufweisen, um die verschiedenen Phasen eines Projekts zu unterstützen. So können beispielsweise die zur Modellerstellung verwendeten historischen Daten aus einer bestimmten Datenbank stammen, die operativen Daten jedoch aus einer anderen.

Eine DPD wird anhand einer bestimmten ODBC-Datenquelle definiert. Um eine DPD aus IBM SPSS Modeler zu verwenden, muss eine ODBC-Datenquelle auf dem IBM SPSS Modeler Server-Host definiert sein, der denselben Namen trägt und der eine Verbindung zu demselben Datenspeicher herstellt wie die in der DPD referenzierte Datenquelle.

Um die zu verwendende DPD auszuwählen, markieren Sie ihren Namen auf der Liste und klicken Sie auf **OK**.

Auswählen der Tabelle

Im Dialogfeld "Tabelle auswählen" werden alle Tabellen aufgelistet, die in der aktuellen Anwendungsansicht referenziert werden. Das Dialogfeld ist leer, wenn keine Verbindung zu einem IBM SPSS Collaboration and Deployment Services Repository hergestellt wurde.

Um die zu verwendende Tabelle auszuwählen, markieren Sie ihren Namen auf der Liste und klicken Sie auf **OK**.

Datenansichtsknoten

Mit dem Datenansichtsknoten können Sie Daten, die in einer Analysedatenansicht von IBM SPSS Collaboration and Deployment Services definiert sind, in Ihren Stream einfügen. Eine Analysedatenansicht definiert eine Struktur zum Zugreifen auf Daten, die die Entitäten beschreiben, die in Vorhersagemodellen und Geschäftsregeln verwendet werden. Die Ansicht ordnet die Datenstruktur physischen Datenquellen für die Analyse zu.

Die Vorhersageanalyse benötigt Daten, die in Tabellen zusammengefasst sind, bei denen jede Zeile einer Entität entspricht, für die Vorhersagen gemacht werden. Jede Spalte in einer Tabelle stellt ein messbares Attribut der Entität dar. Einige Attribute können durch Aggregieren über die Werte für ein anderes Attribut abgeleitet werden. Die Zeilen einer Tabelle könnten z. B. Kunden darstellen und die Spalten könnten dem Kundennamen, dem Geschlecht, der Postleitzahl und der Anzahl der Einkäufe über 500 Dollar entsprechen, die der Kunde im letzten Jahr getätigt hat. Die letzte Spalte wird aus der Statistik der Kundenaufträge abgeleitet, die normalerweise in einer oder in mehreren zugehörigen Tabellen gespeichert ist.

Der Analyseprozess zur Vorhersage beinhaltet die Verwendung verschiedener Sets von Daten im gesamten Lebenszyklus eines Modells. Während der ursprünglichen Entwicklung eines Vorhersagemodells verwenden Sie historische Daten, die oft bekannte Ergebnisse für das vorhergesagte Ereignis aufweisen. Um die Modelleffektivität und -genauigkeit auszuwerten, validieren Sie ein in Frage kommendes Modell anhand anderer Daten. Nach der Validierung des Modells stellen Sie es für den Produktionseinsatz bereit, um für mehrere Entitäten in einem Stapelprozess oder für einzelne Entitäten in einem Echtzeitprozess Scores zu generieren. Wenn Sie das Modell mit Geschäftsregeln in einem Entscheidungsmanagementprozess kombinieren, verwenden Sie simulierte Daten, um die Ergebnisse der Kombination zu validieren. Auch wenn sich die verwendeten Daten in den verschiedenen Phasen des Modellentwicklungsprozesses unterscheiden, muss jedes Dataset trotzdem das gleiche Set von Attributen für das Modell bereitstellen. Das Attributset bleibt konstant. Die Datensätze, die analysiert werden, ändern sich.

Eine Analysedatenansicht besteht aus den folgenden Komponenten, die den spezialisierten Anforderungen der Vorhersageanalyse ansprechen:

- Ein Datenansichtsschema oder Datenmodell, das eine logische Schnittstelle für den Datenzugriff als Set von Attributen definiert, die in zugehörigen Tabellen zusammengefasst sind. Attribute im Modell können aus anderen Attributen abgeleitet werden.
- Mindestens ein Datenzugriffsplan, der die Datenmodellattribute mit physischen Werten bereitstellt. Sie steuern die Daten, die dem Datenmodell zur Verfügung stehen, indem Sie angeben, welcher Datenzugriffsplan für eine bestimmte Anwendung aktiv ist.

Wichtig: Um den Datenansichtsknoten verwenden zu können, muss zuerst ein IBM SPSS Collaboration and Deployment Services Repository an Ihrem Standort installiert sein. Die Analysedatenansicht, die vom Knoten referenziert wird, wird normalerweise mithilfe von IBM SPSS Collaboration and Deployment Services Deployment Manager erstellt und im Repository gespeichert.

Festlegen der Optionen für den Datenansichtsknoten

Mit den Optionen auf der Registerkarte **Daten** des Knotendialogfelds **Datenansicht** können Sie die Dateneinstellungen für eine Analysedatenansicht angeben, die im IBM SPSS Collaboration and Deployment Services Repository ausgewählt wurde.

Analysedatenansicht. Klicken Sie auf die Schaltfläche mit Auslassungspunkten (**...**), um eine Analysedatenansicht auszuwählen. Wenn Sie derzeit nicht mit einem Repository-Server verbunden sind, geben Sie die URL für den Server im Dialogfeld **Repository: Server** an, klicken Sie auf **OK** und geben Sie die Berechtigungsnachweise für Ihre Verbindung im Dialogfeld **Repository: Berechtigungsnachweise** an. Weitere Informationen zur Anmeldung beim Repository und zum Abrufen von Objekten finden Sie im IBM SPSS Modeler-Benutzerhandbuch.

Tabellenname. Wählen Sie eine Tabelle aus dem Datenmodell in der Analysedatenansicht aus. Jede Tabelle im Datenmodell stellt ein Konzept oder eine Entität dar, das bzw. die am Analyseprozess zur Vorhersage beteiligt ist. Felder für die Tabellen entsprechen Attributen der von den Tabellen dargestellten Entitäten. Wenn Sie z. B. Kundenbestellungen analysieren, könnte Ihr Datenmodell eine Tabelle für Kunden und eine Tabelle für Bestellungen enthalten. Die Kundentabelle könnte Attribute für die Kunden-ID, das Alter, das Geschlecht, den Familienstand und den Wohnsitz enthalten. Die Bestellungstabelle könnte Attribute für die Bestellungs-ID, die Anzahl der Artikel in der Bestellung, die Gesamtkosten und die ID für den Kunden enthalten, der die Bestellung abgegeben hat. Mithilfe des Attributs für die Kunden-ID könnten die Kunden in der Kundentabelle ihren Bestellungen in der Bestellungstabelle zugeordnet werden.

Datenzugriffsplan. Wählen Sie einen Datenzugriffsplan aus der Analysedatenansicht aus. Ein Datenzugriffsplan ordnet die Datenmodelltabellen in einer Analysedatenansicht physischen Datenquellen zu. Eine Analysedatenansicht enthält normalerweise mehrere Datenzugriffspläne. Wenn Sie den verwendeten Datenzugriffsplan ändern, ändern Sie die Daten, die von Ihrem Stream verwendet werden. Wenn die Analysedatenansicht beispielsweise einen Datenzugriffsplan zum Trainieren eines Modells und einen Datenzugriffsplan zum Testen eines Modells enthält, können Sie von Trainingsdaten zu Testdaten wechseln, indem Sie den verwendeten Datenzugriffsplan ändern.

Optionale Attribute. Wenn ein bestimmtes Attribut von der Anwendung, die die Analysedatenansicht verwendet, nicht benötigt wird, können Sie das Attribut als optional markieren. Anders als erforderliche Attribute können optionale Attribute Nullwerte enthalten. Sie müssen Ihre Anwendung möglicherweise anpassen, um die Handhabung von Nullwerten für optionale Attribute einzubeziehen. Beim Aufrufen einer Geschäftsregel, die in IBM Operational Decision Manager erstellt wurde, fragt IBM Analytical Decision Management beispielsweise den Regelservice ab, um zu bestimmen, welche Eingaben erforderlich sind. Wenn der Datensatz, der gescort werden soll, für mindestens eines der erforderlichen Felder des Regelservice einen Nullwert enthält, wird die Regel nicht aufgerufen und die Ausgabefelder der Regel werden mit Standardwerten gefüllt. Wenn ein optionales Feld einen Nullwert enthält, wird die Regel aufgerufen. Die Regel kann auf Nullwerte prüfen, um die Verarbeitung zu steuern.

Um Attribute als optional anzugeben, klicken Sie auf **Optionale Attribute** und wählen Sie die Attribute aus, die optional sind.

XML-Daten in Feld einschließen. Wählen Sie diese Option aus, um ein Feld zu erstellen, das die XML-Daten des ausführbaren Objektmodells für jede Datenzeile enthält. Diese Informationen sind erforderlich, wenn die Daten zusammen mit IBM Operational Decision Manager verwendet werden. Geben Sie den Namen für dieses neue Feld an.

Datenbankquellenknoten

Mit dem Datenbankquellenknoten lassen sich Daten aus einer Reihe von anderen Paketen importieren, die ODBC (Open Database Connectivity) verwenden, darunter u. a. Microsoft SQL Server, DB2 und Oracle.

Um in einer Datenbank zu lesen oder in ihr zu schreiben, muss eine ODBC-Datenquelle für die entsprechende Datenbank mit den erforderlichen Lese- und Schreibberechtigungen installiert und konfiguriert sein. Das IBM SPSS Data Access Pack umfasst eine Reihe von ODBC-Treibern, die zu diesem Zweck verwendet werden können. Diese Treiber stehen auf der IBM SPSS Data Access Pack-DVD oder auf der Download-Site zur Verfügung. Wenn Sie Fragen zur Erstellung oder Einstellung von Berechtigungen für ODBC-Datenquellen haben, wenden Sie sich an Ihren Datenbankadministrator.

Unterstützte ODBC-Treiber

Neueste Informationen zu Datenbanken und ODBC-Treibern, die für die Verwendung mit IBM SPSS Modeler 16 getestet wurden und unterstützt werden, finden Sie in den Produktkompatibilitätsdiagrammen auf der unternehmensweiten Support-Site unter http://www.ibm.com/support.

Installationsort der Treiber

Beachten Sie, dass die ODBC-Treiber auf jedem Computer installiert und konfiguriert werden müssen, auf dem eine Verarbeitung erfolgt.

- Wenn Sie IBM SPSS Modeler im lokalen Modus (Standalone-Modus) ausführen, müssen die Treiber auf dem lokalen Computer installiert sein.
- Wenn Sie IBM SPSS Modeler im verteilten Modus mit einer fernen IBM SPSS Modeler Server-Instanz ausführen, müssen die ODBC-Treiber auf dem Computer installiert sein, auf dem IBM SPSS Modeler Server installiert ist. Beachten Sie bei IBM SPSS Modeler Server auf UNIX-Systemen auch "Konfiguration von ODBC-Treibern auf UNIX-Systemen" weiter hinten in diesem Abschnitt.
- Wenn Sie von IBM SPSS Modeler und IBM SPSS Modeler Server auf dieselben Datenquellen zugreifen müssen, müssen die ODBC-Treiber auf beiden Computern installiert sein.
- Wenn Sie IBM SPSS Modeler über Terminaldienste ausführen, müssen die ODBC-Treiber auf dem Terminaldiensteserver installiert sein, auf dem Sie IBM SPSS Modeler installiert haben.

Wichtig: Wenn Sie IBM SPSS Modeler Server unter UNIX zum Zugriff auf eine Teradata-Datenbank verwenden und den von Teradata bereitgestellten ODBC-Treiber verwenden, müssen Sie den ODBC-Treiber-

manager verwenden, der mit diesem Treiber installiert wurde. (**Hinweis**: Sie müssen diese Änderungen nicht vornehmen, wenn Sie den SDAP-Teradata-Treiber verwenden.) Um diese Änderung an IBM SPSS Modeler Server vorzunehmen, geben Sie für ODBC_DRIVER_MANAGER_PATH einen Wert in der Nähe des oberen Bereichs des Scripts modelersrv.sh ein, wo dies durch die Kommentare angegeben wurde. Diese Umgebungsvariable muss auf den Speicherort des ODBC-Treibermanagers eingestellt werden, der mit dem Teradata ODBC-Treiber ausgeliefert wird (/usr/odbc/lib in einer Standardinstallation eines Teradata-ODBC-Treibers). Sie müssen IBM SPSS Modeler Server neu starten, damit die Änderung wirksam wird. Weitere Informationen zu den IBM SPSS Modeler Server-Plattformen, die Teradata-Zugriff unterstützen, sowie über die unterstützte Teradata ODBC-Treiberversion finden Sie auf der unternehmensweiten Support-Site unter http://www.ibm.com/support.

Konfiguration von ODBC-Treibern auf UNIX-Systemen

Standardmäßig ist der DataDirect-Treibermanager nicht für IBM SPSS Modeler Server auf UNIX-Systemen konfiguriert. Geben Sie folgende Befehle ein, um UNIX so zu konfigurieren, dass der DataDirect-Treibermanager geladen wird:

```
cd Modeler Server-Installationsverzeichnis/bin
rm -f libspssodbc.so
ln -s libspssodbc_datadirect.so libspssodbc.so
```

Dadurch wird die Standardverknüpfung entfernt und eine Verknüpfung zum DataDirect-Treibermanager erstellt.

Führen Sie die folgenden allgemeinen Schritte aus, um auf Daten einer Datenbank zuzugreifen:

1. Installieren Sie einen ODBC-Treiber und konfigurieren Sie eine Datenquelle für die zu verwendende Datenbank.
2. Stellen Sie im Dialogfeld des Datenbankknotens im Modus "Tabelle" oder "SQL-Abfrage" eine Verbindung zu einer Datenbank her.
3. Wählen Sie eine Tabelle aus der Datenbank.
4. Anhand der Registerkarten des Dialogfelds des Datenbankknotens können Sie Verwendungstypen ändern und Datenfelder filtern.

Diese Schritte werden in den nächsten Themenabschnitten ausführlicher beschrieben.

Festlegen von Optionen für Datenbankknoten

Mit den Optionen auf der Registerkarte "Daten" des Dialogfelds des Datenbankquellenknotens erhalten Sie Zugriff auf eine Datenbank und können Daten aus der ausgewählten Tabelle lesen.

Modalwert. Wählen Sie **Tabelle** aus, um mit den Steuerelementen des Dialogfelds eine Verbindung zu einer Tabelle herzustellen.

Wählen Sie **SQL-Abfrage**, um die unten ausgewählte Datenbank unter Verwendung von SQL abzufragen. Weitere Informationen finden Sie im Thema „Abfragen der Datenbank” auf Seite 19.

Datenquelle. Sowohl im Modus "Tabelle" als auch im Modus "SQL-Abfrage" können Sie einen Namen in das Feld "Datenquelle" eingeben oder die Option **Neue Datenbankverbindung hinzufügen** in der Dropdown-Liste auswählen.

Die folgenden Optionen dienen zur Verbindung mit einer Datenbank und zur Auswahl einer Tabelle anhand des Dialogfelds:

Tabellenname. Wenn Ihnen der Name der Tabelle, auf die Sie zugreifen möchten, bekannt ist, geben Sie ihn in das Feld "Tabellenname" ein. Klicken Sie andernfalls auf die Schaltfläche **Auswählen**, um ein Dialogfeld mit einer Liste der verfügbaren Tabellen zu öffnen.

Tabellen- und Spaltennamen in Anführungszeichen. Legen Sie fest, ob die Tabellen- und Spaltennamen in Anführungszeichen eingeschlossen werden sollen, wenn Abfragen an die Datenbank gesendet werden (wenn sie z. B. Leerzeichen oder Satzzeichen enthalten).

- Bei Auswahl der Option **Nach Bedarf** werden Tabellen- und Feldnamen *nur* in Anführungszeichen gesetzt, wenn sie Nichtstandardzeichen enthalten. Nichtstandardzeichen sind Nicht-ASCII-Zeichen, Leerzeichen und alle nicht alphanumerischen Zeichen außer einem Punkt (.).
- Wählen Sie **Nie**, wenn Tabellen- und Feldnamen *nie* in Anführungszeichen gesetzt werden sollen.
- Wählen Sie **Immer**, wenn *alle* Tabellen- und Feldnamen in Anführungszeichen gesetzt werden sollen.

Führende und abschließende Leerzeichen löschen. Wählen Sie die Optionen zum Verwerfen von führenden und nachfolgenden Leerzeichen in Zeichenfolgen aus.

Anmerkung. Vergleiche zwischen Zeichenfolgen, die SQL-Pushback verwenden oder nicht, können unterschiedliche Ergebnisse generieren, wenn nachfolgende Leerzeichen vorhanden sind.

Leere Zeichenfolgen aus Oracle lesen. Beim Lesen aus oder Schreiben in Oracle-Datenbanken sollten Sie darauf achten, dass Oracle im Gegensatz zu IBM SPSS Modeler und den meisten anderen Datenbanken leere Zeichenfolgewerte wie Nullwerte behandelt und speichert. Dies bedeutet, dass dieselben Daten sich unterschiedlich verhalten können und unterschiedliche Ergebnisse ausgeben können, je nachdem ob sie aus einer Oracle-Datenbank oder aus einer anderen Datenbank bzw. einer Datei extrahiert wurden.

Hinzufügen einer Datenbankverbindung

Um eine Datenbank zu öffnen, müssen Sie zunächst die Datenquelle auswählen, mit der Sie sich verbinden möchten. Wählen Sie auf der Registerkarte "Daten" in der Dropdown-Liste "Datenquelle" die Option **Neue Datenbankverbindung hinzufügen**.

Das Dialogfeld "Datenbankverbindungen" wird geöffnet.

Hinweis: Alternativ können Sie dieses Dialogfeld über das Hauptmenü öffnen, indem Sie die folgenden Befehle auswählen:

Tools > **Datenbanken...**

Datenquellen. Listet die verfügbaren Datenquellen auf. Führen Sie einen Bildlauf nach unten durch, wenn die gewünschte Datenbank nicht angezeigt wird. Nachdem Sie eine Datenquelle ausgewählt und gegebenenfalls Kennwörter eingegeben haben, klicken Sie auf **Verbinden**. Klicken Sie zum Aktualisieren der Liste auf **Aktualisieren**.

Benutzername. Wenn die Datenquelle kennwortgeschützt ist, geben Sie Ihren Benutzernamen ein.

Kennwort. Wenn die Datenquelle kennwortgeschützt ist, geben Sie Ihr Kennwort ein.

Verbindungen. Zeigt die momentan verbundenen Datenbanken an.

- **Standard.** Sie können eine Verbindung optional als Standard auswählen. Diese Verbindung ist daraufhin für Datenbankquellen- und Exportknoten als Datenquelle vordefiniert. Sie können diese Einstellung jederzeit ändern.
- **Speichern.** Wählen Sie optional eine oder mehr Verbindungen aus, die in nachfolgenden Sitzungen wieder eingeblendet werden soll.
- **Datenquelle.** Die Verbindungszeichenfolgen für die aktuell verbundenen Datenbanken.
- **Voreinstellung.** Zeigt an (mit dem Zeichen *), ob voreingestellte Werte für die Datenbankverbindung angegeben wurden. Um voreingestellte Werte anzugeben, klicken Sie in dieser Spalte auf die der Datenbankverbindung entsprechenden Reihe und wählen Sie aus der Liste "Angeben" aus. Weitere Informationen finden Sie im Thema „Angeben von voreingestellten Werten für eine Datenbankverbindung" auf Seite 16.

Zum Entfernen von Verbindungen wählen Sie eine Verbindung in der Liste aus und klicken Sie auf **Entfernen**.

Wenn Sie Ihre Auswahl getroffen haben, klicken Sie auf **OK**.

- Wenn Sie IBM SPSS Modeler im lokalen Modus (Standalone-Modus) ausführen, müssen die Treiber auf dem lokalen Computer installiert sein.
- Wenn Sie IBM SPSS Modeler im verteilten Modus mit einer fernen IBM SPSS Modeler Server-Instanz ausführen, müssen die ODBC-Treiber auf dem Computer installiert sein, auf dem IBM SPSS Modeler Server installiert ist. Beachten Sie bei IBM SPSS Modeler Server auf UNIX-Systemen auch "Konfiguration von ODBC-Treibern auf UNIX-Systemen" weiter hinten in diesem Abschnitt.
- Wenn Sie von IBM SPSS Modeler und IBM SPSS Modeler Server auf dieselben Datenquellen zugreifen müssen, müssen die ODBC-Treiber auf beiden Computern installiert sein.
- Wenn Sie IBM SPSS Modeler über Terminaldienste ausführen, müssen die ODBC-Treiber auf dem Terminaldiensteserver installiert sein, auf dem Sie IBM SPSS Modeler installiert haben.

Angeben von voreingestellten Werten für eine Datenbankverbindung

Bei einigen Datenbanken können Sie verschiedene Standardeinstellungen für die Datenbankverbindung angeben. Diese Einstellungen gelten alle für den Datenbankexport.

Diese Funktion wird von folgenden Datenbanktypen unterstützt:

- IBM InfoSphere Warehouse unter DB2 9.1 oder höher. Weitere Informationen finden Sie im Thema „Einstellungen für IBM DB2 InfoSphere Warehouse".
- SQL Server 2008 oder höher, Enterprise und Developer Edition. Weitere Informationen finden Sie im Thema „Einstellungen für SQL Server".
- Oracle 10g und 11gR1 oder höher, Enterprise oder Personal Edition. Weitere Informationen finden Sie im Thema „Einstellungen für Oracle" auf Seite 17.
- IBM Netezza, IBM DB2 unter z/OS und Teradata stellen eine Verbindung zu einer Datenbank oder zu einem Schema auf ähnliche Weise her. Weitere Informationen finden Sie im Thema „Einstellungen für IBM Netezza, IBM DB2 for z/OS, IBM DB2 LUW und Teradata" auf Seite 18.

Wenn Sie mit einer Datenbank oder einem Schema verbunden sind, die bzw. das diese Funktion nicht unterstützt, wird die Nachricht **Für diese Datenbankverbindung können keine Voreinstellungen konfiguriert werden** angezeigt.

Einstellungen für IBM DB2 InfoSphere Warehouse

Diese Einstellungen werden für IBM InfoSphere Warehouse unter DB2 9.1 oder höher angezeigt.

Tabellenbereich. Der Tabellenbereich, der für den Export verwendet wird. Datenbankadministratoren können Tabellenbereiche partioniert erstellen oder konfigurieren. Wir empfehlen, einen dieser Tabellenbereiche (anstelle des standardmäßig eingestellten) für den Datenbankexport zu verwenden.

Komprimierung verwenden. Bei Auswahl dieser Option werden Tabellen für den komprimierten Export erstellt (entspricht z. B. `CREATE TABLE MYTABLE(...) COMPRESS YES;` in SQL).

Aktualisierungen nicht protokollieren. Bei Auswahl dieser Option werden das Erstellen von Tabellen und Einfügen von Daten nicht protokolliert (entspricht `CREATE TABLE MYTABLE(...) NOT LOGGED INITIALLY;` in SQL).

Einstellungen für SQL Server

Diese Einstellungen werden für SQL Server 2008 oder höher, Enterprise und Developer Edition, angezeigt.

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt.

Komprimierung für. Wählen Sie die Komprimierungsstufe aus.

- **Zeile.** Aktiviert Komprimierung auf der Zeilenebene (z. B. die Entsprechung von `CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = ROW);` in SQL).
- **Seite.** Aktiviert Komprimierung auf der Seitenebene (z. B. `CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = PAGE);` in SQL).

Einstellungen für Oracle

Oracle 10g-Einstellungen

Diese Einstellungen werden für Oracle 10g, Enterprise oder Personal Edition, angezeigt.

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt. Für diese Version der Datenbank steht nur einfache Komprimierung zur Verfügung (beispielsweise `CREATE TABLE MYTABLE(...) COMPRESS;` in SQL).

Oracle 11gR1-Einstellungen

Diese Einstellungen werden für Oracle 11g, Enterprise oder Personal Edition, angezeigt.

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt.

Komprimierung für. Wählen Sie die Komprimierungsstufe aus.

- **Standard.** Aktiviert Standardkomprimierung (z. B. `CREATE TABLE MYTABLE(...) COMPRESS;` in SQL). In diesem Fall hat sie dieselbe Wirkung wie die Option **Direkte Ladevorgänge**.
- **Direkte Ladevorgänge.** Aktiviert Komprimierung ausschließlich für Masseneinfügevorgänge (direkter Pfad) (z. B. `CREATE TABLE MYTABLE(...)COMPRESS FOR DIRECT_LOAD OPERATIONS;` in SQL).
- **Alle Vorgänge.** Aktiviert Komprimierung für alle Vorgänge (z. B. `CREATE TABLE MYTABLE(...)COMPRESS FOR ALL OPERATIONS;` in SQL).

Oracle 11gR2-Einstellungen - Option "Basic" (Einfach)

Diese Einstellungen werden für Oracle 11g R2, Enterprise oder Personal Edition, bei Verwendung der Option "Basic" (Einfach) angezeigt.

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt.

Komprimierung für. Wählen Sie die Komprimierungsstufe aus.

- **Standard.** Aktiviert Standardkomprimierung (z. B. `CREATE TABLE MYTABLE(...) COMPRESS;` in SQL). In diesem Fall hat sie dieselbe Wirkung wie die Option **Einfach**.
- **Einfach.** Aktiviert einfache Komprimierung (z. B. `CREATE TABLE MYTABLE(...) COMPRESS BASIC;` in SQL).

Oracle 11gR2-Einstellungen - Option "Advanced" (Erweitert)

Diese Einstellungen werden für Oracle 11g R2, Enterprise oder Personal Edition, bei Verwendung der Option "Advanced" (Erweitert) angezeigt.

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt.

Komprimierung für. Wählen Sie die Komprimierungsstufe aus.

- **Standard.** Aktiviert Standardkomprimierung (z. B. `CREATE TABLE MYTABLE(...) COMPRESS;` in SQL). In diesem Fall hat sie dieselbe Wirkung wie die Option **Einfach**.
- **Einfach.** Aktiviert einfache Komprimierung (z. B. `CREATE TABLE MYTABLE(...) COMPRESS BASIC;` in SQL).
- **OLTP.** Aktiviert OLTP-Komprimierung (z. B. `CREATE TABLE MYTABLE(...)COMPRESS FOR OLTP;` in SQL).
- **Abfrage niedrig/hoch.** (Nur Exadata-Server) Aktiviert Hybrid Columnar Compression für Abfrage (z. B. `CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY LOW;` oder `CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY HIGH;` in SQL). Komprimierung für Abfragen ist in Data Warehousing-Umgebungen sinnvoll; `HIGH` (Hoch) bietet ein höheres Komprimierungsverhältnis als `LOW` (Niedrig).
- **Archiv niedrig/hoch.** (Nur Exadata-Server) Aktiviert Hybrid Columnar Compression für Archiv (z. B. `CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE LOW;` oder `CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE HIGH;` in SQL). Komprimierung für Archive ist sinnvoll zur Komprimierung von Daten, die lange Zeit gespeichert werden sollen; `HIGH` (Hoch) bietet ein höheres Komprimierungsverhältnis als `LOW` (Niedrig).

Einstellungen für IBM Netezza, IBM DB2 for z/OS, IBM DB2 LUW und Teradata

Wenn Sie Voreinstellungen für IBM Netezza, IBM DB2 for z/OS, IBM DB2 LUW oder Teradata festlegen, werden Sie aufgefordert, Folgendes auszuwählen:

Datenbank/Schema mit Server-Scoring-Adapter verwenden. Wenn ausgewählt, wird die Option **Datenbank/Schema mit Server-Scoring-Adapter** aktiviert.

Datenbank/Schema mit Server-Scoring-Adapter. Wählen Sie die erforderliche Verbindung aus der Dropdown-Liste aus.

Auswählen einer Datenbanktabelle

Nachdem Sie eine Verbindung zu einer Datenquelle hergestellt haben, können Sie wahlweise Felder aus einer bestimmten Tabelle oder Ansicht importieren. Auf der Registerkarte "Daten" des Dialogfelds "Datenbank" können Sie entweder den Namen einer Tabelle in das Feld "Tabellenname" eingeben oder auf **Auswählen** klicken, um das Dialogfeld "Tabelle/Ansicht auswählen" zu öffnen, in dem eine Liste der verfügbaren Tabellen und Ansichten angezeigt wird.

Tabelleneigner anzeigen. Wählen Sie diese Option, wenn für eine Datenquelle die Angabe des Tabellenbesitzers erforderlich ist, damit Sie auf die Tabelle zugreifen können. Inaktivieren Sie diese Option für Datenquellen, die die Angabe des Tabellenbesitzers nicht erfordern.

Hinweis: Für SAS- und Oracle-Datenbanken ist es in der Regel erforderlich, den Tabellenbesitzer anzuzeigen.

Tabellen/Ansichten. Wählen Sie die Tabelle oder Ansicht aus, die Sie importieren möchten.

Anzeigen. Listet die Spalten der Datenquelle auf, mit der Sie verbunden sind. Klicken Sie auf eine der folgenden Optionen, um Ihre Ansicht der verfügbaren Tabellen anzupassen:

- Klicken Sie auf **Benutzertabellen**, um gewöhnliche, von Datenbankbenutzern erstellte Datenbanktabellen anzuzeigen.
- Klicken Sie auf **Systemtabellen**, um systemeigene Datenbanktabellen anzuzeigen (dies sind z. B. Tabellen, die Informationen über die Datenbank wie Indexdetails enthalten). Mit dieser Option können Sie die in Excel-Datenbanken verwendeten Register anzeigen. (Beachten Sie, dass auch ein eigener Excel-Quellenknoten verfügbar ist. Weitere Informationen finden Sie im Thema „Excel-Quellenknoten” auf Seite 42.)
- Klicken Sie auf **Ansichten**, um virtuelle Tabellen basierend auf einer Abfrage, die eine oder mehrere gewöhnliche Tabellen betrifft, anzuzeigen.
- Klicken Sie auf **Synonyme**, um Synonyme anzuzeigen, die in der Datenbank für bereits vorhandene Tabellen erstellt wurden.

Namens-/Eignerfilter. Mit diesen Feldern können Sie die Liste der angezeigten Tabellen nach Name oder Besitzer filtern. Geben Sie z. B. `SYS` ein, um nur Tabellen mit diesem Besitzer aufzulisten. Bei Suchvorgänge mit Platzhalterzeichen steht ein Unterstrich (_) für ein einzelnes Zeichen und ein Prozentzeichen (%) für eine Folge von null oder mehr Zeichen.

Als Standard festlegen. Diese Option speichert die aktuellen Einstellungen als Standardwerte für den aktuellen Benutzer. Diese Einstellungen werden zukünftig wiederhergestellt, wenn ein Benutzer ein neues Dialogfeld zur Auswahl einer Tabelle öffnet. Dies gilt jedoch *nur für denselben Datenquellennamen und dieselbe Benutzeranmeldung*.

Abfragen der Datenbank

Sobald Sie eine Verbindung zu einer Datenquelle hergestellt haben, können Sie Felder anhand von SQL-Abfragen importieren. Wählen Sie im Hauptdialogfeld **SQL-Abfrage** als Verbindungsmodus. Dem Dialogfeld wird ein Fenster für den Abfrageeditor hinzugefügt. Mit dem Abfrageeditor können Sie eine oder mehrere SQL-Abfragen erstellen oder laden, deren Ergebnisset in den Datenstream eingelesen wird.

Wenn Sie mehrere SQL-Abfragen angeben, trennen Sie sie durch Semikolons (;) und achten Sie darauf, dass es nicht mehrere SELECT-Anweisungen gibt.

Um das Fenster des Abfrageeditors abzubrechen und zu schließen, wählen Sie **Tabelle** als Verbindungsmodus aus.

Sie können SPSS Modeler-Streamparameter (eine Art benutzerdefinierte Variable) in die SQL-Abfrage mit aufnehmen. Weitere Informationen finden Sie im Thema „Verwenden von Streamparametern in einer SQL-Abfrage".

Abfrage laden. Klicken Sie auf diese Option, um den Dateibrowser zu öffnen, mit dem Sie eine bereits gespeicherte Abfrage laden können.

Abfrage speichern. Klicken Sie auf diese Option, um das Dialogfeld "Abfrage speichern" zu öffnen. In diesem Dialogfeld können Sie die aktuelle Abfrage speichern.

Import Standardabfrage. Klicken Sie auf diese Option, um eine `SQL SELECT`-Beispielanweisung zu importieren, die automatisch anhand der im Dialogfeld ausgewählten Tabelle und Spalten erstellt wird.

Löschen. Löscht den Inhalt des Arbeitsbereichs. Verwenden Sie diese Option, wenn Sie neu beginnen möchten.

Text teilen. Die Standardoption **Nie** bedeutet, dass die Abfrage als Ganzes an die Datenbank gesendet wird. Sie können auch **Nach Bedarf** auswählen, was bedeutet, dass SPSS Modeler versucht, die Abfrage zu analysieren und zu ermitteln, ob SQL-Anweisungen vorhanden sind, die nacheinander an die Datenbank gesendet werden sollten.

Verwenden von Streamparametern in einer SQL-Abfrage

Beim Schreiben einer SQL-Abfrage für den Feldimport können Sie zuvor definierte SPSS Modeler-Streamparameter mit einschließen. Es werden sämtliche Arten von Streamparametern unterstützt.

In der folgenden Tabelle wird angezeigt, wie einige Beispiele für Streamparameter in der SQL-Abfrage interpretiert werden.

Tabelle 1. Beispiele für Streamparameter.

Name des Streamparameters (Beispiel)	Speicher	Wert des Streamparameters	Interpretiert als
PString	Zeichenfolge	ss	'ss'

Tabelle 1. Beispiele für Streamparameter (Forts.).

Name des Streamparameters (Beispiel)	Speicher	Wert des Streamparameters	Interpretiert als
PInt	Ganzzahl	5	5
PReal	Reelle Zahl	5.5	5.5
PTime	Zeit	23:05:01	t{'23:05:01'}
PDate	Datum	2011-03-02	d{'2011-03-02'}
PTimeStamp	Zeitmarke	2011-03-02 23:05:01	ts{'2011-03-02 23:05:01'}
PColumn	Unbekannt	IntValue	IntValue

In der SQL-Abfrage geben Sie einen Streamparameter auf dieselbe Weise an wie in einem CLEM-Ausdruck, nämlich durch `'$P-<Parametername>'`, wobei `<Parametername>` der für den Streamparameter definierte Name ist.

Beim Verweisen auf ein Feld muss der Speichertyp als "Unbekannt" definiert sein und der Parameterwert muss in Anführungszeichen eingeschlossen sein, falls er benötigt wird. Wenn Sie also unter Verwendung der in der Tabelle angezeigten Beispiele folgende SQL-Abfrage eingeben:

```
select "IntValue" from Table1 where "IntValue" < '$P-PInt';
```

würde sie ausgewertet als:

```
select "IntValue" from Table1 where "IntValue" < 5;
```

Wenn Sie auf das Feld `IntValue` mit dem Parameter `PColumn` verweisen, müssen Sie die Abfrage wie folgt angeben, um dasselbe Ergebnis zu erhalten:

```
select "IntValue" from Table1 where "'$P-PColumn'" < '$P-PInt';
```

Knoten "Variable Datei"

Mit Knoten des Typs "Datei (var.)" können Sie Daten aus Textdateien mit freien Feldern lesen (dies sind Dateien, deren Datensätze eine konstante Anzahl von Feldern und eine variable Anzahl von Zeichen enthalten). Diese Dateien sind auch als Textdateien mit Trennzeichen bekannt. Dieser Knotentyp ist außerdem nützlich für Dateien mit fester Länge, Überschriftentext und bestimmten Anmerkungen. Datensätze werden einzeln nacheinander eingelesen und durch den Stream geleitet, bis die gesamte Datei eingelesen ist.

Hinweise zum Einlesen von Textdaten mit Trennzeichen

- Datensätze müssen durch einen Zeilenumbruch am Ende jeder Zeile getrennt sein. Das Zeilenumbruchzeichen darf für keinen anderen Zweck verwendet werden (beispielsweise innerhalb von Feldnamen oder Werten). Führende und nachfolgende Leerzeichen sollten idealerweise entfernt werden, um Platz zu sparen. Dies ist jedoch nicht unbedingt erforderlich. Optional können sie auch durch den Knoten entfernt werden.
- Felder müssen durch ein Komma oder ein anderes Zeichen getrennt werden, das idealerweise ausschließlich als Trennzeichen verwendet wird, also nicht in Feldnamen oder Werten vorkommt. Wenn dies nicht möglich ist, können alle Textfelder in doppelte Anführungszeichen gesetzt werden, vorausgesetzt dass keiner der Feldnamen oder Textwerte ein doppeltes Anführungszeichen enthält. Wenn Feldnamen oder Werte doppelte Anführungszeichen enthalten, können die Textfelder alternativ in einfache Anführungszeichen gesetzt werden. Auch hier gilt natürlich wieder die Bedingung, dass einzelne Anführungszeichen nicht bereits an anderen Stellen in Werten verwendet werden. Wenn weder einfache

noch doppelte Anführungszeichen verwendet werden können, müssen die Textwerte geändert werden, um entweder das Trennzeichen oder die einfachen bzw. doppelten Anführungszeichen zu entfernen bzw. zu ersetzen.

- Alle Zeilen, einschließlich der Zeile für die Überschrift, sollten die gleiche Anzahl von Feldern enthalten.
- Die erste Zeile sollte die Feldnamen enthalten. Wenn dies nicht der Fall ist, müssen Sie die Auswahl von **Feldnamen aus Datei lesen** aufheben, um jedem Feld einen allgemeinen Namen zu geben, wie *Feld1, Feld2* usw.
- Die zweite Zeile muss den ersten Datensatz enthalten. Leerzeilen und Kommentare sind nicht zulässig.
- Numerische Werte dürfen kein Tausendertrennzeichen oder Gruppierungssymbol enthalten, daher muss 3.000,00 beispielsweise ohne Punkt geschrieben werden. Das Dezimaltrennzeichen (Komma in Deutschland) darf nur an den entsprechenden Stellen verwendet werden.
- Datums- und Zeitangaben sollten in einem der Format vorliegen, die vom Dialogfeld für die Streamoptionen erkannt werden, beispielsweise `TT/MM/JJJJ` oder `HH:MM:SS`. Alle Datums- und Zeitfelder in der Datei sollten idealerweise dasselbe Format verwenden und alle Felder, die ein Datum enthalten, müssen für alle Werte in diesem Feld dasselbe Format verwenden.

Festlegen der Optionen für Knoten "Variable Datei"

Die Optionen können Sie über die Registerkarte "Datei" des Dialogfelds für den Knoten "Variable Datei" festlegen.

Datei. Geben Sie den Namen der Datei an. Zur Auswahl einer Datei können Sie einen Dateinamen eingeben oder auf die Schaltfläche mit den Auslassungspunkten (**...**) klicken. Der Dateipfad wird angezeigt, sobald Sie eine Datei ausgewählt haben, und der entsprechende Inhalt wird mit Trennzeichen im Fenster darunter angezeigt.

Der von Ihrer Datenquelle angezeigte Beispieltext kann kopiert und in folgende Steuerelemente eingefügt werden: EOL-Kommentarzeichen und benutzerdefinierte Trennzeichen. Verwenden Sie zum Kopieren und Einfügen Strg-C und Strg-V.

Feldnamen aus Datei lesen. Diese standardmäßig ausgewählte Option behandelt die erste Zeile der Datendatei als Beschriftungen für die Spalte. Handelt es sich bei der ersten Zeile nicht um eine Überschrift, inaktivieren Sie die Option, damit jedes Feld im Dataset automatisch einen generischen Namen, wie *Feld1, Feld2*, erhält.

Anzahl der Felder eingeben. Geben Sie die Anzahl der Felder in jedem Datensatz an. Die Anzahl der Felder kann automatisch ermittelt werden, sofern sich am Ende der Datensätze ein Zeilenumbruch befindet. Sie können auch manuell eine Zahl angeben.

Führende Zeichen überspringen. Legen Sie fest, wie viele Zeichen am Anfang des ersten Datensatzes ignoriert werden sollen.

EOL-Kommentarzeichen. Geben Sie Zeichen wie # oder ! ein, um auf Anmerkungen in den Daten hinzuweisen. Wenn ein solches Zeichen in der Datendatei angezeigt wird, werden alle Daten bis zu diesem Zeichen, jedoch nicht einschließlich des nächsten Zeilenumbruchs, ignoriert.

Führende und abschließende Leerzeichen löschen. Wählen Sie die Optionen zum Verwerfen von führenden und nachfolgenden Leerzeichen in Zeichenfolgen beim Importieren aus.

Anmerkung. Vergleiche zwischen Zeichenfolgen, die SQL-Pushback verwenden oder nicht, können unterschiedliche Ergebnisse generieren, wenn nachfolgende Leerzeichen vorhanden sind.

Ungültige Zeichen. Wählen Sie **Verwerfen**, um ungültige Zeichen aus der Datenquelle zu entfernen. Wählen Sie **Ersetzen durch**, um ungültige Zeichen durch das angegebene Symbol (nur ein Zeichen) zu ersetzen. Ungültige Zeichen sind Nullzeichen bzw. alle Zeichen, die nicht in der angegebenen Codierungsmethode vorhanden sind.

Codierung. Gibt die verwendete Textcodierungsmethode an. Sie haben die Wahl zwischen der Systemstandardeinstellung, der Streamstandardeinstellung und UTF-8.

- Die Systemstandardeinstellung wird in der Windows-Systemsteuerung bzw. bei Ausführung im verteilten Modus auf dem Server-Computer angegeben.
- Der Streamstandard wird im Dialogfeld "Streameigenschaften" festgelegt.

Dezimaltrennzeichen. Wählen Sie das in Ihrer Datenquelle verwendete Dezimaltrennzeichen aus. Die **Streamstandardeinstellung** entspricht dem auf der Registerkarte "Optionen" des Dialogfelds "Streameigenschaften" ausgewählten Zeichen. Wählen Sie andernfalls entweder **Punkt (.)** oder **Komma (,)**, um alle Daten dieses Dialogfelds mit dem ausgewählten Zeichen als Dezimaltrennzeichen zu lesen.

Zeilentrennzeichen ist Zeichen für Zeilenvorschub. Wählen Sie diese Option aus, um das Zeichen für den Zeilenvorschub als Zeilentrennzeichen anstatt als Feldtrennzeichen zu verwenden. Dies kann beispielsweise dann nützlich sein, wenn es in einer Zeile eine ungerade Anzahl von Trennzeichen gibt, die einen Zeilenumbruch bewirken. Beachten Sie: Wenn Sie diese Option auswählen, können Sie in der Liste der Trennzeichen nicht die Option **Neue Zeile** auswählen.

Beachten Sie, dass bei Auswahl dieser Option alle Leerwerte am Ende von Datenzeilen entfernt werden.

Trennzeichen. Mit den für dieses Steuerelement aufgelisteten Kontrollkästchen können Sie angeben, welche Zeichen, z. B. das Komma (,), die Feldbegrenzungen in der Datei definieren. Außerdem können Sie mehr als ein Trennzeichen angeben, z. B. ", |" für Datensätze mit mehreren Trennzeichen. Das Standardtrennzeichen ist das Komma.

Hinweis: Wenn das Komma auch als Dezimaltrennzeichen definiert wurde, funktionieren die Standardeinstellungen nicht. Wenn das Komma sowohl als Feldtrennzeichen als auch als Dezimaltrennzeichen festgelegt ist, wählen Sie in der Liste "Trennzeichen" **Andere** aus. Geben Sie dann manuell ein Komma in das Eingabefeld ein.

Wählen Sie **Mehrere leere Trennzeichen zulassen** aus, um mehrere nebeneinander liegende leere Trennzeichen als ein einzelnes Trennzeichen zu behandeln. Beispiel: Wenn auf ein Datenwert vier Leerzeichen und ein weiterer Datenwert folgen, wird diese Gruppe als zwei statt fünf Felder betrachtet.

Nach Typ zu durchsuchende Zeilen und Spalten. Legen Sie fest, wie viele Zeilen und Spalten nach angegebenen Datentypen durchsucht werden sollen.

Datum und Uhrzeit automatisch erkennen. Aktivieren Sie dieses Kontrollkästchen, damit IBM SPSS Modeler automatisch versucht, Dateneinträge als Datum oder Uhrzeit zu erkennen. Das bedeutet beispielsweise, dass ein Eintrag wie 07-11-1965 als Datum erkannt wird und 02:35:58 als Uhrzeit. Zweideutige Einträge wie 07111965 oder 023558 werden jedoch als Ganzzahlen angezeigt, da die Zahlen nicht durch Trennzeichen getrennt sind.

Hinweis: Um mögliche Datenprobleme bei der Verwendung von Datendateien älterer Versionen von IBM SPSS Modeler zu vermeiden, ist dieses Kontrollkästchen standardmäßig für Informationen inaktiviert, die in älteren Versionen als 13 gespeichert wurden.

Anführungszeichen. Mit den Dropdown-Listen können Sie angeben, wie einfache und doppelte Anführungszeichen beim Importieren zu behandeln sind. Sie können alle Anführungszeichen **verwerfen**, **als Text einschließen**, d. h. in den Feldwert einschließen, oder **Paare bilden und verwerfen**, um Anführungszeichenpaare zu finden und zu löschen. Kann einem Anführungszeichen kein zweites Anführungs-

zeichen zugeordnet werden, wird eine Fehlernachricht ausgegeben. Sowohl die Option **Verwerfen** als auch **Paaren und verwerfen** speichert den Feldwert (ohne Anführungszeichen) als Zeichenfolge.

Klicken Sie bei der Bearbeitung dieses Dialogfelds zu einem beliebigen Zeitpunkt auf **Aktualisieren**, um Daten aus der Datenquelle neu zu laden. Diese Funktion ist nützlich, wenn Sie Datenverbindungen zum Quellenknoten ändern oder wenn Sie die verschiedenen Registerkarten des Dialogfelds bearbeiten.

Knoten "Datei (fest)"

Mit Knoten des Typs "Datei (fest)" können Sie Daten aus Textdateien mit festen Feldern importieren (dies sind Dateien, deren Felder nicht begrenzt sind, sondern an derselben Position beginnen und eine feste Länge haben). Maschinell erzeugte Daten oder Legacydaten werden häufig im Format mit festen Feldern gespeichert. Anhand der Registerkarte "Datei" des Knotens "Datei (fest)" können Sie problemlos die Position und Länge der Spalten Ihrer Daten angeben.

Festlegen der Optionen für den Knoten "Datei (fest)"

Auf der Registerkarte "Datei" des Knotens "Datei (fest)" können Sie Daten in IBM SPSS Modeler importieren und die Spaltenposition und Datensatzlänge angeben. Klicken Sie im Datenvorschaufenster in der Mitte des Dialogfelds, um Pfeile hinzuzufügen, mit denen die Haltepunkte zwischen den Feldern angegeben werden.

Datei. Geben Sie den Namen der Datei an. Zur Auswahl einer Datei können Sie einen Dateinamen eingeben oder auf die Schaltfläche mit den Auslassungspunkten (**...**) klicken. Sobald Sie eine Datei ausgewählt haben, wird der Dateipfad angezeigt und der entsprechende Inhalt wird mit Trennzeichen im Fenster unten angezeigt.

Im Datenvorschaufenster können Sie die Spaltenposition und Länge festlegen. Das Lineal am oberen Rand des Vorschaufensters unterstützt Sie beim Messen der Länge der Variablen und beim Festlegen des Haltepunkts zwischen den Variablen. Sie können Haltepunktlinien festlegen, indem Sie in den Linealbereich oberhalb der Felder klicken. Haltepunkte können durch Ziehen verschoben werden. Um sie zu verwerfen, ziehen Sie sie aus dem Datenvorschaubereich.

- Jede Haltepunktlinie fügt automatisch ein neues Feld zur Feldtabelle hinzu.
- Durch die Pfeile markierte Startpositionen werden automatisch zur Startspalte in der Tabelle unten hinzugefügt.

Zeilenorientiert. Wählen Sie diese Option aus, wenn Sie das Zeilenwechselzeichen am Ende jedes Datensatzes überspringen möchten.

Kopfzeilen überspringen. Legen Sie fest, wie viele Zeilen am Anfang des ersten Datensatzes ignoriert werden sollen. Diese Funktion ist nützlich, um Spaltenkopfzeilen zu ignorieren.

Datensatzlänge. Geben Sie die Zahl der Zeichen in jedem Datensatz an.

Feld. Alle Felder, die Sie für diese Datendatei definiert haben, werden hier aufgelistet. Es gibt zwei Methoden für das Definieren von Feldern:

- Felder interaktiv anhand des Datenvorschaufensters festlegen.
- Felder manuell durch Hinzufügen leerer Feldzeilen zur Tabelle unten festlegen. Klicken Sie auf die Schaltfläche rechts neben dem Feldfenster, um neue Felder hinzuzufügen. Geben Sie anschließend einen Feldnamen, eine Start-Position und eine Länge in das leere Feld ein. Mit diesen Optionen werden automatisch Pfeile zum Datenvorschaufenster hinzugefügt, die problemlos angepasst werden können.

Um ein bereits definiertes Feld zu löschen, wählen Sie das Feld in der Liste aus und klicken Sie auf die rote Löschschaltfläche.

Start. Legen Sie die Position des ersten Zeichens im Feld fest. Beispiel: Wenn das zweite Feld eines Datensatzes beim sechzehnten Zeichen beginnt, geben Sie 16 als Startwert ein.

Länge. Legen Sie fest, wie viele Zeichen sich im längsten Wert für jedes Feld befinden. Dadurch wird der Abbruchpunkt für das nächste Feld bestimmt.

Führende und abschließende Leerzeichen löschen. Wählen Sie diese Option aus, um führende und nachfolgende Leerzeichen in Zeichenfolgen beim Importieren zu verwerfen.

Anmerkung. Vergleiche zwischen Zeichenfolgen, die SQL-Pushback verwenden oder nicht, können unterschiedliche Ergebnisse generieren, wenn nachfolgende Leerzeichen vorhanden sind.

Ungültige Zeichen. Wählen Sie **Verwerfen** aus, um ungültige Zeichen aus der Dateneingabe zu entfernen. Wählen Sie **Ersetzen durch**, um ungültige Zeichen durch das angegebene Symbol (nur ein Zeichen) zu ersetzen. Ungültige Zeichen sind Nullzeichen (0) bzw. alle Zeichen, die nicht in der aktuellen Codierung vorhanden sind.

Codierung. Gibt die verwendete Textcodierungsmethode an. Sie haben die Wahl zwischen der Systemstandardeinstellung, der Streamstandardeinstellung und UTF-8.

- Die Systemstandardeinstellung wird in der Windows-Systemsteuerung bzw. bei Ausführung im verteilten Modus auf dem Server-Computer angegeben.
- Der Streamstandard wird im Dialogfeld "Streameigenschaften" festgelegt.

Dezimaltrennzeichen. Wählen Sie das in Ihrer Datenquelle verwendete Dezimaltrennzeichen aus. **Streamstandardeinstellung** entspricht dem auf der Registerkarte "Optionen" des Dialogfelds "Streameigenschaften" ausgewählten Zeichen. Wählen Sie andernfalls entweder **Punkt (.)** oder **Komma (,)**, um alle Daten dieses Dialogfelds mit dem ausgewählten Zeichen als Dezimaltrennzeichen zu lesen.

Datum und Uhrzeit automatisch erkennen. Aktivieren Sie dieses Kontrollkästchen, damit IBM SPSS Modeler automatisch versucht, Dateneinträge als Datum oder Uhrzeit zu erkennen. Das bedeutet beispielsweise, dass ein Eintrag wie 07-11-1965 als Datum erkannt wird und 02:35:58 als Uhrzeit. Zweideutige Einträge wie 07111965 oder 023558 werden jedoch als Ganzzahlen angezeigt, da die Zahlen nicht durch Trennzeichen getrennt sind.

Hinweis: Um mögliche Datenprobleme bei der Verwendung von Datendateien älterer Versionen von IBM SPSS Modeler zu vermeiden, ist dieses Kontrollkästchen standardmäßig für Informationen inaktiviert, die in älteren Versionen als 13 gespeichert wurden.

Nach Typ zu durchsuchende Zeilen. Legen Sie fest, wie viele Zeilen nach angegebenen Datentypen durchsucht werden sollen.

Klicken Sie bei der Bearbeitung dieses Dialogfelds zu einem beliebigen Zeitpunkt auf **Aktualisieren**, um Daten aus der Datenquelle neu zu laden. Diese Funktion ist nützlich, wenn Sie Datenverbindungen zum Quellenknoten ändern oder wenn Sie die verschiedenen Registerkarten des Dialogfelds bearbeiten.

Festlegen von Feldspeicher und Formatierung

Mit den Optionen auf der Registerkarte "Daten" für die Knoten "Datei (fest)" und "Datei (var.)", "XML-Quelle" und "Eingabe" können Sie den Speichertyp für Felder festlegen, die in IBM SPSS Modeler importiert oder erstellt werden. Für die Knoten "Datei (fest)", "Datei (var.)" und "Eingabe" können Sie außerdem die Feldformatierung und andere Metadaten festlegen.

Bei aus anderen Quellen eingelesenen Daten wird der Speichertyp automatisch ermittelt, kann jedoch mithilfe einer Konvertierungsfunktion, wie beispielsweise `to_integer`, in einem Füller- oder Ableitungsknoten geändert werden.

Feld. Mit der Spalte *Feld* zeigen Sie Felder im aktuellen Dataset an und wählen sie aus.

Überschreiben. Aktivieren Sie das Kontrollkästchen in der Spalte *Überschreiben*, um die Optionen in den Spalten *Speichertyp* und *Eingabeformat* zu aktivieren.

Datenspeichertyp

Der Speichertyp beschreibt die Art und Weise, wie Daten in einem Feld gespeichert werden. Beispiel: Ein Feld mit den Werten 1 und 0 speichert ganzzahlige Daten. Dies ist vom Messniveau zu unterscheiden, das die Verwendung der Daten beschreibt und sich nicht auf den Speichertyp auswirkt. Beispiel: Sie möchten das Messniveau für ein Feld ganzer Zahlen mit den Werten 1 und 0 auf *Flag* setzen. Das bedeutet normalerweise, dass 1=*True* und 0=*False* ist. Während der Speichertyp stets an der Quelle festgelegt werden muss, kann das Messniveau mithilfe eines Typknotens an jeder beliebigen Stelle im Stream geändert werden. Weitere Informationen finden Sie im Thema „Messniveaus" auf Seite 119.

Folgende Speichertypen sind verfügbar:

- **Zeichenfolge.** Wird für Felder verwendet, die nicht numerische Daten enthalten (auch als alphanumerische Daten bezeichnet). Eine Zeichenfolge kann jede beliebige Abfolge von Zeichen enthalten, beispielsweise *fred*, *Klasse 2* oder *1234*. Beachten Sie, dass die Zahlen in Zeichenfolgen nicht für Berechnungen verwendet werden können.
- **Ganze Zahl.** Ein Feld, bei dessen Werten es sich um ganze Zahlen handelt.
- **Reelle Zahl.** Bei den Werten handelt es sich um Zahlen, die Dezimalstellen enthalten können (nicht auf ganze Zahlen beschränkt). Das Anzeigeformat wird im Dialogfeld für die Streameigenschaften angegeben und kann für einzelne Felder in einem Typknoten überschrieben werden (Registerkarte "Format").
- **Datum.** In einem Standardformat, wie Jahr, Monat und Tag (z. B. `2007-09-26`), angegebene Datumswerte. Das jeweilige Format wird im Dialogfeld für die Streameigenschaften angegeben.
- **Uhrzeit.** Als Dauer gemessene Zeit. Beispielsweise kann ein Service-Call, der 1 Stunde, 26 Minuten und 38 Sekunden dauerte, als `01:26:38` angegeben werden, je nachdem, welches Zeitformat aktuell im Dialogfeld für die Streameigenschaften angegeben ist.
- **Zeitmarke.** Werte, die sowohl eine Datums- als auch eine Zeitkomponente enthalten, wie beispielsweise `2007-09-26 09:04:00`; auch hier wieder abhängig von den aktuellen Formaten für Datum und Zeit im Dialogfeld "Streameigenschaften". Beachten Sie, dass Zeitmarkenwerte gegebenenfalls in Anführungszeichen gesetzt werden müssen, um sicherzustellen, dass sie als Einzelwert interpretiert werden und nicht als gesonderte Datums- und Zeitwerte. (Dies gilt beispielsweise bei der Eingabe von Werten in einem Benutzereingabeknoten.)

Speichertypkonvertierung. Der Speichertyp für ein Feld kann mit verschiedenen Konvertierungsfunktionen, z. B. `to_string` und `to_integer`, in einem Füllerknoten geändert werden. Weitere Informationen finden Sie im Thema „Speichertypkonvertierung mithilfe des Füllerknotens" auf Seite 140. Beachten Sie, dass die Konvertierungsfunktionen (und alle anderen Funktionen, für die ein spezieller Eingabetyp, wie beispielsweise ein Wert für Datum oder Uhrzeit, erforderlich ist) von den aktuell im Dialogfeld "Streameigenschaften" angegebenen Formaten abhängen. Wenn Sie beispielsweise ein Zeichenfolgenfeld mit den Werten *Jan 2003*, *Feb 2003* (usw.) in einen Datumsspeicher konvertieren müssen, wählen Sie **MON JJJJ** als Standarddatumsformat für den Stream aus. Konvertierungsfunktionen sind auch im Ableitungsknoten zur temporären Konvertierung während einer Ableitungsberechnung verfügbar. Mit dem Ableitungsknoten können Sie auch andere Bearbeitungen vornehmen wie beispielsweise die Umcodierung von Zeichenfolgenfeldern mit kategorialen Werten. Weitere Informationen finden Sie im Thema „Umcodieren von Werten mit dem Ableitungsknoten" auf Seite 139.

Einlesen gemischter Daten. Beachten Sie, dass beim Einlesen von Feldern mit numerischem Speichertyp (ganze Zahl, reelle Zahl, Zeit, Zeitmarke oder Datum) alle nicht numerischen Werte auf null oder auf systemdefiniert fehlend gesetzt werden. Dies liegt daran, dass IBM SPSS Modeler im Gegensatz zu einigen anderen Anwendungen keine gemischten Speichertypen innerhalb eines Felds zulässt. Um dies zu ver-

meiden, sollten alle Felder mit gemischten Daten als Zeichenfolgen eingelesen werden, indem der Speichertyp im Quellenknoten oder in der externen Anwendung nach Bedarf geändert wird.

Feldeingabeformat (nur für die Knoten "Datei (fest)", "Datei (var.)" und "Eingabe")

Sie können für alle Speichertypen außer "Zeichenfolge" und "Ganze Zahl" anhand der Dropdown-Liste Formatierungsoptionen für das ausgewählte Feld festlegen. Beispiel: Beim Verbinden von Daten verschiedener Ländereinstellungen müssen Sie einen Punkt (.) als Dezimaltrennzeichen für ein Feld festlegen, während ein anderes Feld ein Komma als Trennzeichen erfordert.

Im Quellenknoten festgelegte Eingabeoptionen überschreiben die Formatierungsoptionen, die im Dialogfeld "Streameigenschaften" definiert sind. Sie sind jedoch später im Stream nicht persistent. Ihr Zweck besteht darin, Eingaben basierend auf Ihrem Wissen über die Daten korrekt zu analysieren. Die festgelegten Formate dienen als Richtlinie für die Analyse der Daten beim Einlesen in IBM SPSS Modeler und bestimmen nicht das Format nach dem Einlesen in IBM SPSS Modeler. Um die Formatierung für die einzelnen Felder an anderer Stelle im Stream festzulegen, verwenden Sie die Registerkarte "Format" eines Typknotens. Weitere Informationen finden Sie im Thema „Feldformat - Registerkarte "Einstellungen"" auf Seite 127.

Die Optionen sind je nach Speichertyp verschieden. Für den Speichertyp "Reelle Zahl" können Sie z. B. **Punkt (.)** oder **Komma (,)** als Dezimaltrennzeichen auswählen. Für Zeitmarkenfelder wird ein separates Dialogfeld geöffnet, wenn Sie in der Dropdown-Liste **Angeben** wählen. Weitere Informationen finden Sie im Thema „Festlegen der Feldformatierungsoptionen" auf Seite 128.

Für alle Speichertypen können Sie auch **Streamstandardeinstellung** wählen, um die Streamstandardeinstellungen für den Import zu verwenden. Streameinstellungen werden im Dialogfeld "Streameigenschaften" festgelegt.

Weitere Optionen

Auf der Registerkarte "Daten" können einige andere Optionen festgelegt werden:

- Zum Anzeigen von Speichertypeinstellungen für Daten, die nicht mehr über den aktuellen Knoten verbunden sind (z. B. Trainingsdaten), wählen Sie **Nicht verwendete Feldeinstellungen anzeigen**. Sie können die Legacyfelder löschen, indem Sie auf **Löschen** klicken.
- Klicken Sie bei der Bearbeitung dieses Dialogfelds zu einem beliebigen Zeitpunkt auf **Aktualisieren**, um Daten aus der Datenquelle neu zu laden. Diese Funktion ist nützlich, wenn Sie Datenverbindungen zum Quellenknoten ändern oder wenn Sie die verschiedenen Registerkarten des Dialogfelds bearbeiten.

Data Collection-Knoten

Data Collection-Quellenknoten importieren Umfragedaten auf der Basis von IBM SPSS Data Collection Survey Reporter Developer Kit, das von der Marktforschungssoftware von IBM verwendet wird. Bei diesem Format wird zwischen **Falldaten**, also den tatsächlichen Antworten auf Fragen, die während einer Umfrage gesammelt wurden, und den **Metadaten** unterschieden, die beschreiben, wie Falldaten gesammelt und organisiert werden. Metadaten bestehen aus Informationen wie Fragetexten, Variablennamen und -beschreibungen, Variablendefinitionen für Mehrfachantworten, Übersetzungen der verschiedenen Textzeichenfolgen und der Definition der Struktur der Falldaten.

Hinweis: Für diesen Knoten ist IBM SPSS Data Collection Survey Reporter Developer Kit erforderlich, das zusammen mit IBM SPSS Data Collection-Softwareprodukten von IBM verteilt wird. weitere Informationen finden Sie auf der Webseite von IBM SPSS Data Collection unter . Abgesehen von der Installation des Developer Kit ist keine weitere Konfiguration erforderlich.

Kommentare

- Umfragedaten werden aus dem einfachen VDATA-Tabellenformat eingelesen oder aus Datenquellen im hierarchischen HDATA-Format, sofern diese eine Metadatenquelle beinhalten (erfordert IBM SPSS Data Collection 4.5 oder höher).
- Die Typen werden automatisch mithilfe von Informationen aus den Metadaten instanziiert.
- Wenn Umfragedaten in IBM SPSS Modeler importiert werden, werden Fragen als Felder wiedergegeben, wobei für jeden Befragten ein Datensatz verwendet wird.

Dateioptionen für den Data Collection-Import

Auf der Registerkarte "Datei" im Data Collection-Knoten können Sie Optionen für die zu importierenden Metadaten und Falldaten angeben.

Metadateneinstellungen

Hinweis: Um die vollständige Liste der verfügbaren Providerdateitypen zu sehen, müssen Sie das IBM SPSS Data Collection Survey Reporter Developer Kit installieren, das mit der IBM SPSS Data Collection-Software zur Verfügung steht. Weitere Informationen finden Sie auf der IBM SPSS Data Collection-Webseite unter http://www.ibm.com/software/analytics/spss/products/data-collection/survey-reporter-devkit/.

Metadatenprovider. Die Umfragedaten können aus einer Reihe von Formaten importiert werden, die von der IBM SPSS Data Collection Survey Reporter Developer Kit-Software unterstützt werden. Folgende Provider-Typen werden unterstützt:

- **DataCollectionMDD.** Liest Metadaten aus einer Fragebogendefinitionsdatei (*.mdd*) ein. Dies ist das standardmäßige IBM SPSS Data Collection Data Model-Format.
- **ADO-Datenbank.** Liest Falldaten und Metadaten aus ADO-Dateien ein. Geben Sie Namen und Speicherort der *.adoinfo*-Datei an, die die Metadaten enthält. Der interne Name dieser DSC lautet *mrADODsc*.
- **In2data-Datenbank.** Liest In2data-Falldaten und Metadaten. Der interne Name dieser DSC lautet *mrI2dDsc*.
- **Datensammlungsprotokolldatei.** Liest Metadaten aus einer IBM SPSS Data Collection-Standardprotokolldatei. In der Regel haben Protokolldateien die Dateinamenerweiterung *.tmp*. Einige Protokolldateien können jedoch eine andere Dateinamenerweiterung aufweisen. Falls erforderlich, können Sie die Datei umbenennen, sodass sie die Dateinamenerweiterung *.tmp* erhält. Der interne Name dieser DSC lautet *mrLogDsc*.
- **Quancept-Definitionsdatei.** Konvertiert Metadaten in ein Quancept-Script. Geben Sie den Namen der Quancept-Datei (*.qdi*) an. Der interne Name dieser DSC lautet *mrQdiDrsDsc*.
- **Quanvert-Datenbank.** Liest Quanvert-Falldaten und Metadaten. Geben Sie Namen und Speicherort der Datei *.qvinfo* bzw. *.pkd* an. Der interne Name dieser DSC lautet *mrQvDsc*.
- **Datensammlungsteilnahmedatenbank.** Liest die Stichproben- und Verlaufstabellen eines Projekts ein und erstellt abgeleitete kategoriale Variablen, die den Spalten in diesen Tabellen entsprechen. Der interne Name dieser DSC lautet *mrSampleReportingMDSC*.
- **Statistikdatei.** Liest Falldaten und Metadaten aus einer IBM SPSS Statistics-Datei (*.sav*). Schreibt Falldaten zur Analyse in IBM SPSS Statistics in eine IBM SPSS Statistics-Datei (*.sav*). Schreibt Metadaten aus einer IBM SPSS Statistics-Datei (*.sav*) in eine *.mdd*-Datei. Der interne Name dieser DSC lautet *mrSavDsc*.
- **Surveycraft-Datei.** Liest SurveyCraft-Falldaten und Metadaten. Geben Sie den Namen der SurveyCraft-Datei (*.vq*) an. Der interne Name dieser DSC lautet *mrSCDsc*.
- **Datensammlungsscriptdatei.** Liest aus Metadaten in einer mrScriptMetadata-Datei. Typischerweise tragen diese Dateien die Dateinamenerweiterung *.mdd* bzw. *.dms*. Der interne Name dieser DSC lautet *mrScriptMDSC*.
- **Triple-S-XML-Datei.** Liest Metadaten aus einer Triple-S-Datei im XML-Format. Der interne Name dieser DSC lautet *mrTripleSDsc*.

Metadateneigenschaften. Wählen Sie optional **Eigenschaften** aus, um die zu importierende Umfrageversion sowie die Sprache, den Kontext und den Beschriftungstyp anzugeben, die verwendet werden sollen. Weitere Informationen finden Sie im Thema „IBM SPSS Data Collection-Import - Metadateneigenschaften” auf Seite 29.

Falldateneinstellungen

Hinweis: Um die vollständige Liste der verfügbaren Providerdateitypen zu sehen, müssen Sie das IBM SPSS Data Collection Survey Reporter Developer Kit installieren, das mit der IBM SPSS Data Collection-Software zur Verfügung steht. Weitere Informationen finden Sie auf der IBM SPSS Data Collection-Webseite unter http://www.ibm.com/software/analytics/spss/products/data-collection/survey-reporter-devkit/.

Falldateneinstellungen abrufen. Wenn Sie Metadaten ausschließlich aus *.mdd*-Dateien einlesen, klicken Sie auf **Falldateneinstellungen abrufen**, um zu bestimmen, welche Falldatenquellen den ausgewählten Metadaten zugeordnet sind, sowie die konkreten Einstellungen, die für den Zugriff auf eine bestimmte Quelle erforderlich sind. Diese Option ist nur für *.mdd*-Dateien verfügbar.

Falldatenprovider. Folgende Provider-Typen werden unterstützt:

- **ADO-Datenbank.** Liest Falldaten mithilfe der Microsoft ADO-Schnittstelle. Wählen Sie `OLE-DB UDL` für den Falldatentyp aus und geben Sie eine Verbindungszeichenfolge im Feld "Falldaten-UDL" an. Weitere Informationen finden Sie im Thema „Datenbankverbindungszeichenfolge” auf Seite 30. Der interne Name dieser component lautet *mrADODsc*.
- **Textdatei mit Trennzeichen (Excel).** Liest Falldaten aus einer kommagetrennten Datei (.CSV), wie sie von Excel ausgegeben werden kann. Der interne Name lautet *mrCsvDsc*.
- **Datensammlungsdatendatei.** Liest Falldaten aus einer Datei im systemeigenen IBM SPSS Data Collection-Datenformat (ab IBM SPSS Data Collection 4.5). Der interne Name lautet *mrDataFileDsc*.
- **In2data-Datenbank.** Liest Falldaten und Metadaten aus einer In2data-Datenbank (*.i2d*) ein. Der interne Name lautet *mrI2dDsc*.
- **Datensammlungsprotokolldatei.** Liest Falldaten aus einer Standard-IBM SPSS Data Collection-Protokolldatei. Typischerweise tragen Protokolldatei die Dateinamenerweiterung *.tmp*. Einige Protokolldateien können jedoch eine andere Dateinamenerweiterung aufweisen. Falls erforderlich, können Sie die Datei umbenennen, sodass sie die Dateinamenerweiterung *.tmp* erhält. Der interne Name lautet *mrLogDsc*.
- **Quantum-Datendatei.** Liest Daten aus einer ASCII-Datei im Quantum-Format (*.dat*). Der interne Name lautet *mrPunchDsc*.
- **Quancept-Datendatei.** Liest Daten aus einer Quancept-Datei (*.drs*, *.drz* bzw..*dru*). Der interne Name lautet *mrQdiDrsDsc*.
- **Quanvert-Datenbank.** Liest Falldaten aus einer Quanvert-Datei (*qvinfo* bzw. *.pkd*). Der interne Name lautet *mrQvDsc*.
- **Datensammlungsdatenbank (MS SQL Server).** Liest Falldaten in eine relationale Microsoft SQL Server-Datenbank ein. Weitere Informationen finden Sie im Thema „Datenbankverbindungszeichenfolge” auf Seite 30. Der interne Name lautet *mrRdbDsc2*.
- **Statistikdatei.** Liest Falldaten aus einer IBM SPSS Statistics-Datei (*.sav*). Der interne Name lautet *mrSavDsc*.
- **Surveycraft-Datei.** Liest Falldaten aus einer SurveyCraft-Datei (*.qdt*). Die *.vq-* und *.qdt*-Dateien müssen sich i demselben Verzeichnis befinden und es muss für beide Dateien Lese- und Schreibzugriff bestehen. Dies ist nicht die Standardvorgehensweise bei der Erstellung mit SurveyCraft. Daher muss eine der Dateien verschoben werden, um SurveyCraft-Daten importieren zu können. Der interne Name lautet *mrScDsc*.
- **Triple-S-Datendatei.** Liest Daten aus einer Triple-S-Datendatei, entweder im Format mit fester Länge oder im kommagetrennten Format. Der interne Name lautet *mr TripleDsc*.

- **Datensammlungs-XML.** Liest Falldaten aus einer IBM SPSS Data Collection XML-Datendatei. Typischerweise kann dieses Format zur Übertragung von Falldaten von einem Speicherort an einen anderen verwendet werden. Der interne Name lautet *mrXmlDsc*.

Falldatentyp. Gibt an, ob Falldaten aus einer Datei, einem Ordner, aus OLE-DB UDL oder ODBC DSN gelesen werden sollen, und aktualisiert die Dialogfeldoptionen entsprechend. Welche Optionen gültig sind, hängt vom Provider-Typ ab. Bei Datenbankprovidern können Sie Optionen für die OLE-DB- bzw. ODBC-Verbindung angeben. Weitere Informationen finden Sie im Thema „Datenbankverbindungszeichenfolge" auf Seite 30.

Falldatenprojekt. Beim Lesen von Falldaten aus einer IBM SPSS Data Collection-Datenbank können Sie den Namen des Projekts eingeben. Bei allen anderen Falldatentypen sollte diese Einstellung leer bleiben.

Variablenimport

Systemvariablen importieren. Gibt an, ob Systemvariablen importiert werden sollen, einschließlich Variablen, die den Befragungsstatus angeben (läuft, abgeschlossen, Fertigstellungsdatum usw.). Sie haben die Auswahl zwischen **Keine**, **Alle** und **Benutzerdefiniert**.

"Codes"-Variablen importieren. Steuert den Import von Variablen, die Codes darstellen, die für offene Antworten vom Typ "Andere" bei kategorialenVariablen verwendet werden.

"SourceFile"-Variablen importieren. Steuert den Import von Variablen, die Dateinamen oder Bilder von gescannten Antworten enthalten.

Mehrfachantwortvariablen importieren als. Mehrfachantwortvariablen können als mehrere Flagfelder (Set aus dichotomen Variablen) importiert werden (dies ist die Standardmethode für neue Streams). In Versionen von IBM SPSS Modeler vor 12.0 erstellte Streams importierten Mehrfachantworten in ein einzelnes Feld. Die Werte wurden dabei durch Kommas getrennt. Die ältere Methode wird weiterhin unterstützt, damit bestehende Streams weiterhin wie gehabt ausgeführt werden können, es wird jedoch empfohlen, ältere Streams für die Verwendung der neuen Methode zu aktualisieren. Weitere Informationen finden Sie im Thema „Importieren von Mehrfachantwortsets" auf Seite 30.

IBM SPSS Data Collection-Import - Metadateneigenschaften

Beim Import der IBM SPSS Data Collection-Umfragedaten können Sie im Dialogfeld "Metadateneigenschaften" die zu importierende Umfrageversion sowie die Sprache, den Kontext und den Beschriftungstyp angeben, die verwendet werden sollen. Beachten Sie, dass jeweils nur eine Sprache, ein Kontext und ein Beschriftungstyp importiert werden kann.

Version. Jede Umfrageversion lässt sich als Momentaufnahme der für die Sammlung eines bestimmten Falldatasets verwendeten Metadaten betrachten. Wenn sich ein Fragebogen ändert, können mehrere Versionen erstellt werden. Sie können die aktuellste Version, alle Versionen oder eine bestimmte Version erstellen.

- **Alle Versionen.** Wählen Sie diese Option, wenn eine Kombination (Obermenge) aller verfügbaren Versionen verwendet werden soll. (Dies wird manchmal als "Superversion" bezeichnet.) Bei einem Konflikt zwischen den Versionen haben die aktuelleren Versionen normalerweise Vorrang gegenüber den älteren Versionen. Wenn sich beispielsweise eine Kategoriebeschriftung in einer Version abweicht, wird der Text in der aktuellsten Version verwendet.
- **Neueste Version.** Wählen Sie diese Option, wenn Sie die aktuellste Version verwenden möchten.
- **Version angeben.** Wählen Sie diese Option, wenn Sie eine bestimmte Umfrageversion verwenden möchten.

Die Auswahl aller Versionen ist sinnvoll, wenn Sie beispielsweise Falldaten für mehrere Versionen exportieren möchten und Änderungen an den Variablen- und Kategoriedefinitionen durchgeführt wurden, die dazu führen, dass Falldaten, die mit einer bestimmten Version gesammelt wurden, in einer anderen Versi-

on nicht gültig sind. Die Auswahl aller Versionen, für die die Falldaten exportiert werden sollen, bedeutet, dass Sie im Allgemeinen die mit den verschiedenen Versionen gesammelten Falldaten gleichzeitig exportieren können, ohne dass Gültigkeitsfehler aufgrund der Unterschiede zwischen den Versionen gemeldet werden. Dennoch können, je nach den Versionsänderungen, dennoch einige Gültigkeitsfehler auftreten.

Sprache. Die Fragen und der zugehörige Text können in den Metadaten in mehreren Sprachen gespeichert werden. Sie können die Standardsprache für die Umfrage verwenden oder eine bestimmte Sprache angeben. Wenn ein Element in der angegebenen Sprache nicht verwendet wird, wird der Standard verwendet.

Kontext. Wählen Sie den zu verwendenden Benutzerkontext aus. Der Benutzerkontext regelt, welche Texte angezeigt werden. Wählen Sie beispielsweise **Frage**, um Fragetexte anzuzeigen, oder **Analyse**, um kürzere Texte anzuzeigen, die bei der Analyse der Daten für die Anzeige geeignet sind.

Beschriftungstyp. Listet die definierten Beschriftungstypen auf. Der Standard ist **Beschriftung**. Er wird für Fragetexte im Benutzerkontext "Frage" und für Variablenbeschreibungen im Benutzerkontext "Analyse" verwendet. Für Anweisungen, Beschreibungen usw. können weitere Beschriftungstypen definiert werden.

Datenbankverbindungszeichenfolge

Bei Verwendung des IBM SPSS Data Collection-Knotens zum Import von Falldaten aus einer Datenbank über OLE-DB oder ODBC wählen Sie **Bearbeiten** auf der Registerkarte "Datei" aus, um auf das Dialogfeld "Verbindungszeichenfolge" zuzugreifen, in dem Sie die Verbindungszeichenfolge anpassen können, die zur Feinabstimmung der Verbindung an den Provider übergeben wird.

Erweiterte Eigenschaften

Bei Verwendung des IBM SPSS Data Collection-Knotens zum Importieren von Falldaten aus einer Datenbank, für die eine explizite Anmeldung erforderlich ist, wählen Sie **Erweitert** aus, um eine Benutzer-ID und ein Kennwort für den Zugriff auf die Datenquelle anzugeben.

Importieren von Mehrfachantwortsets

Mehrfachantwortvariablen können aus IBM SPSS Data Collection als Sets aus dichotomen Variablen mit einem gesonderten Flagfeld für jeden möglichen Wert der Variablen importiert werden. Wenn die Befragten beispielsweise in einer Liste auswählen sollen, welche Museen sie besucht haben, enthält das Set ein gesondertes Flagfeld für jedes aufgeführte Museum.

Nach dem Import der Daten können Sie Mehrfachantwortsets über jeden Knoten, der die Registerkarte "Filter" enthält, hinzufügen und bearbeiten. Weitere Informationen finden Sie im Thema „Bearbeiten von Mehrfachantwortsets” auf Seite 131.

Importieren von Mehrfachantworten in ein einzelnes Feld (für in früheren Versionen erstellte Streams)

In älteren Versionen von IBM SPSS Modeler wurden Mehrfachantworten nicht wie oben beschrieben importiert, sondern in ein einzelnes Feld. Die Werte wurden dabei durch Kommas getrennt. Diese Methode wird weiterhin unterstützt, um bestehende Streams zu unterstützen, es wird jedoch empfohlen, alle derartigen Streams für die Verwendung der neuen Methode zu aktualisieren.

Anmerkungen zum Import von IBM SPSS Data Collection-Spalten

Spalten aus den IBM SPSS Data Collection-Daten werden wie in der folgenden Tabelle zusammengefasst in IBM SPSS Modeler eingelesen.

Tabelle 2. IBM SPSS Data Collection-Spaltenimport - Zusammenfassung

IBM SPSS Data Collection Spaltentyp	**IBM SPSS Modeler-Speicher**	**Messniveau**
Boolesches Flag (ja/nein)	Zeichenfolge	Flag (Werte 0 und 1)
Kategorial	Zeichenfolge	Nominal
Datum oder Zeitmarke	Zeitmarke	Stetig
Doppelt (Gleitkommawert innerhalb eines angegebenen Bereichs)	Reelle Zahl	Stetig
Lang (ganzzahliger Wert innerhalb eines angegebenen Bereichs)	Ganzzahl	Stetig
Text (Freitextbeschreibung)	Zeichenfolge	Ohne Typ
Ebene (gibt Raster oder Schleifen innerhalb einer Frage an)	Kommt in VDATA nicht vor und wird nicht in IBM SPSS Modeler importiert	
Objekt (Binärdaten wie beispielsweise ein Fax mit handgeschriebenem Text oder eine Tonaufnahme)	Wird nicht in IBM SPSS Modeler importiert	
Keine (unbekannter Typ)	Wird nicht in IBM SPSS Modeler importiert	
Respondent.Serial-Spalte (weist jedem Befragten eine eindeutige ID zu)	Ganzzahl	Ohne Typ

Um mögliche Inkonsistenzen zwischen Wertbeschriftungen aus Metadaten und tatsächlichen Werten zu vermeiden, werden alle Metadatenwerte in Kleinbuchstaben umgewandelt. So wird beispielsweise der Wert der Beschriftung *E1720_Jahre* in *e1720_jahre* konvertiert.

Analytic Server-Quelle

Mithilfe der Analytic Server-Quelle können Sie einen Stream in HDFS (Hadoop Distributed File System) ausführen. Die Informationen in einer Analytic Server-Datenquelle können aus verschiedenen Quellen stammen. Dazu gehören:

- Textdateien in HDFS
- Datenbanken
- HCatalog

Normalerweise wird ein Stream mit einer Analytic Server-Quelle in HDFS ausgeführt. Wenn ein Stream jedoch einen Knoten enthält, der nicht für die Ausführung in HDFS unterstützt wird, wird ein möglichst großer Teil des Streams mit einer Pushback-Operation zurück zu Analytic Server übertragen und SPSS Modeler Server versucht anschließend, den restlichen Stream zu verarbeiten. Sie müssen für sehr große Datasets eine Teilstichprobe durchführen, indem Sie z. B. einen Stichprobenknoten im Stream platzieren.

Datenquelle. Wenn Ihr SPSS Modeler Server-Administrator eine Verbindung hergestellt hat, wählen Sie eine Datenquelle aus, die die Daten enthält, die Sie verwenden wollen. Eine Datenquelle enthält die zu dieser Quelle zugehörigen Dateien und Metadaten. Klicken Sie auf **Auswählen**, um eine Liste der verfügbaren Datenquellen anzuzeigen. Weitere Informationen finden Sie im Thema „Analytic Server - Datenquelle auswählen" auf Seite 32.

Wenn Sie eine neue Datenquelle erstellen müssen oder eine vorhandene Datenquelle bearbeiten müssen, klicken Sie auf **Datenquelleneditor starten...**.

Analytic Server - Datenquelle auswählen

In der Tabelle "Datenquellen" wird eine Liste der verfügbaren Datenquellen angezeigt. Wählen Sie die Quelle aus, die Sie verwenden möchten, und klicken Sie auf **OK**.

Klicken Sie auf **Eigner anzeigen**, um den Datenquelleneigner anzuzeigen.

Mit **Filtern nach** können Sie die Datenquellenliste nach Schlüsselwort filtern, wodurch die Filterkriterien mit dem Datenquellennamen und der Datenquellenbeschreibung oder dem Eigner verglichen werden. Sie können eine Kombination aus Zeichenfolgewert und numerischem Wert oder das Platzhalterzeichen (%) als Filterkriterien eingeben. Bei dem Suchbegriff muss die Groß-/Kleinschreibung beachtet werden. Klicken Sie auf **Aktualisieren**, um die Tabelle "Quellentabellen" zu aktualisieren.

Analytic Server-Berechtigungsnachweise

Wenn sich Ihre Berechtigungsnachweise für den Zugriff auf Analytic Server von den Berechtigungsnachweisen für den Zugriff auf SPSS Modeler Server unterscheiden, müssen Sie die Analytic Server-Berechtigungsnachweise bei der Ausführung eines Streams auf Analytic Server eingeben. Wenn Sie Ihre Berechtigungsnachweise nicht kennen, wenden Sie sich an Ihren Serveradministrator.

Unterstützte Knoten

Viele SPSS Modeler-Knoten werden für die Ausführung in HDFS unterstützt, bei der Ausführung bestimmter Knoten gibt es jedoch möglicherweise einige Unterschiede und einige Knoten werden zurzeit nicht unterstützt. In diesem Thema wird die aktuelle Unterstützungsstufe detailliert beschrieben.

Allgemein

- Einige Zeichen, die normalerweise in einem Modeler-Feldnamen in Anführungszeichen zulässig sind, werden von Analytic Server nicht akzeptiert.
- Damit ein Modeler-Stream in Analytic Server ausgeführt werden kann, muss er mit mindestens einem Analytic Server-Quellenknoten beginnen und mit einem einzelnen Modellierungsknoten oder Analytic Server-Exportknoten enden. Zusammenführungen sind zulässig, eine Verzweigung jedoch nicht.
- Es wird empfohlen, den Speicher von stetigen Zielen als Speicher für reelle Zahlen und nicht als Speicher für ganze Zahlen festzulegen. Scoring-Modelle schreiben immer reelle Werte in die Ausgabedatendateien für stetige Ziele, während das Ausgabedatenmodell für die Scores dem Speicher des Ziels folgt. Wenn ein stetiges Ziel über einen Speicher für ganze Zahlen verfügt, gibt es daher eine Diskrepanz zwischen den geschriebenen Werten und dem Datenmodell für die Scores und diese Diskrepanz führt zu Fehlern, wenn Sie versuchen, die gescorten Daten zu lesen.

Quelle

- Ein Stream, der mit etwas anderem als einem Analytic Server-Quellenknoten beginnt, wird lokal ausgeführt.

Datensatzoperationen

Es werden alle Datensatzoperationen unterstützt. Weitere Hinweise zur Funktion dieser Knoten folgen.

Auswählen

- Unterstützt dieselbe Funktionsgruppe wie der Ableitungsknoten.

Stichprobe

- Stichprobenziehung auf Blockebene wird nicht unterstützt.
- Komplexe Methoden der Stichprobenziehung werden nicht unterstützt.

Aggregieren

- Zusammenhängende Schlüssel werden nicht unterstützt.

- Reihenfolgestatistiken (Median, 1. Quartil, 3. Quartil) werden nicht unterstützt.

Sortieren

- Die Registerkarte "Optimierung" wird nicht unterstützt.

In einer verteilten Umgebung gibt es eine begrenzte Anzahl von Operationen, bei dem die vom Sortierknoten eingerichtete Datensatzreihenfolge beibehalten wird.

- Ein Sortierknoten, auf den ein Exportknoten folgt, erstellt eine sortierte Datenquelle.
- Ein Sortierknoten, auf den ein Stichprobenknoten mit der **ersten** Datensatzstichprobenziehung folgt, gibt die ersten *N* Datensätze zurück.
- Ein Sortierknoten, auf den ein Modellierungsknoten mit dem Ziel **Für sehr große Datasets optimieren** (Neuronales Netz, Linear, C&R-Baum, Quest, CHAID) folgt, ist ein hilfreiches Muster für das Anzeigen von Datensätzen in zufälliger Reihenfolge durch das Sortieren nach einem abgeleiteten Zufallszahlschlüssel, um eine Verzerrung zu vermeiden, die im Modellerstellungsalgorithmus auftreten kann, wenn die ursprünglichen Datensätze geordnet werden.

Im Allgemeinen sollten Sie einen Sortierknoten so nah wie möglich bei den Operationen platzieren, die die sortierten Datensätze benötigen.

Zusammenführen

- Das Zusammenführen nach Reihenfolge wird nicht unterstützt.
- Das Zusammenführen nach Bedingung wird nicht unterstützt.
- Die Registerkarte "Optimierung" wird nicht unterstützt.
- Das Platzieren eines Stichprobenknotens oder eines Modellnuggets zwischen einem Analytic Server-Quellenknoten und einem Zusammenführungsknoten wird zurzeit nicht unterstützt. Normalerweise ist es möglich, einen Auswahlknoten anzugeben, um die Funktion des Stichprobenknotens zu ersetzen.
- Analytic Server führt bei Schlüsseln für leere Zeichenfolgen keinen Join durch. Wenn also einer der Schlüssel, mit dem Sie die Zusammenführung durchführen, leere Zeichenfolgen enthält, werden alle Datensätze, die die leere Zeichenfolge enthalten, aus der zusammengeführten Ausgabe gelöscht.
- Zusammenführungsoperationen sind relativ langsam. Wenn in HDFS Speicherplatz verfügbar ist, ist es unter Umständen weniger zeitintensiv, wenn Sie Ihre Datenquellen einmal zusammenzuführen und die zusammengeführte Quelle in den folgenden Streams verwenden, anstatt die Datenquellen in jedem Stream zusammenzuführen.

Feldoperationen

Die Knoten "Autom. Datenvorbereitung", "Typ", "Filter", "Ableiten", "Ensemble", "Füller", "Umcodieren", "Klassierung", "RFM-Analyse, "Partition", "Dichotom", "Umstrukturieren" und "Felder ordnen" werden unterstützt. Weitere Hinweise zur Funktion dieser Knoten folgen.

Autom. Datenvorbereitung

- Das Trainieren des Knotens wird nicht unterstützt. Die Anwendung der Transformationen in einem trainierten Knoten des Typs "Autom. Datenvorbereitung" auf neue Daten wird unterstützt.

Typ

- Die Spalte "Überprüfen" wird nicht unterstützt.
- Die Registerkarte "Format" wird nicht unterstützt.

Ableiten

- Alle Ableitungsfunktionen werden unterstützt, mit Ausnahme von Sequenzfunktionen.
- Aufteilungsfelder können nicht in demselben Stream abgeleitet werden, der sie als Aufteilungen verwendet. Sie müssen zwei Streams erstellen: einen, der das Aufteilungsfeld ableitet, und einen, der das Feld als Aufteilungen verwendet.

- Ein Flagfeld kann nicht allein in einem Vergleich verwendet werden. Das heißt, dass `if (flagField) then ... endif` einen Fehler verursacht. Als Fehlerumgehung kann `if (flagField=trueValue) then ... endif` verwendet werden.
- Wenn der Operator ** verwendet wird, wird empfohlen, den Exponenten als reelle Zahl anzugeben, z. B. x**2,0 anstelle von x**2, damit die Ergebnisse mit den Ergebnissen in Modeler übereinstimmen.

Füller

- Unterstützt dieselbe Funktionsgruppe wie der Ableitungsknoten.

Klassierung

Die folgende Funktion wird nicht unterstützt.

- Optimales Klassieren
- Ränge
- N-Perzentile -> Perzentilmethode: Summe der Werte
- N-Perzentile -> Bindungen: "In aktuellem beibehalten" und "Zufällig zuweisen"
- N-Perzentile ->Benutzerdef. N: Werte über 100 und jeder N-Wert, bei dem 100 % N ungleich null ist.

RFM-Analyse

- Die Option "In aktuellem beibehalten" für die Handhabung von Bindungen wird nicht unterstützt. RFM-Aktualitäts-, Häufigkeits- und Geldwertscores stimmen nicht immer mit denen überein, die von Modeler aus denselben Daten berechnet werden. Die Scorebereiche sind identisch, Scorezuweisungen (Klassennummern) können sich jedoch um 1 unterscheiden.

Grafiken

Alle Diagrammknoten werden unterstützt.

Modellierung

Es wird eine begrenzte Anzahl an Modellierungsknoten unterstützt. Beispiele für unterstützte Knoten: Linear, Neuronales Netz, C&RT, CHAID, Quest. Weitere Hinweise zur Funktion dieser Knoten folgen.

Linear Beim Erstellen von Modellen für große Daten ändern Sie das Ziel in der Regel in "Sehr große Datasets" oder geben Aufteilungen an.

- Fortlaufendes Training vorhandener PSM-Modelle wird nicht unterstützt.
- Das Modellerstellungsziel "Standard" wird nur empfohlen, wenn Aufteilungsfelder so definiert sind, dass die Anzahl an Datensätzen in den einzelnen Aufteilungen nicht "zu groß" ist, wobei die Definition von "zu groß" von der Leistungsstärke einzelner Knoten in Ihrem Hadoop-Cluster abhängt. Im Gegensatz dazu müssen Sie auch darauf bedacht sein, sicherzustellen, dass Aufteilungen nicht so fein definiert sind, dass zu wenige Datensätze für die Erstellung eines Modells vorhanden sind.
- Das Ziel "Boosting" wird nicht unterstützt.
- Das Ziel "Bagging" wird nicht unterstützt.
- Das Ziel "Sehr große Datasets" wird nicht empfohlen, wenn es wenige Datensätze gibt. Häufig wird dann entweder kein Modell oder ein vermindertes Modell erstellt. Sie haben möglicherweise auch Probleme, wenn die Eingabedatensätze nach einer Systematik geordnet werden, die gegen die Zufälligkeitsvoraussetzungen hinter den verwendeten Erstellungsalgorithmen für Ensemblemodelle verstößt.
- Die automatische Datenaufbereitung wird nicht unterstützt. Dies kann Probleme verursachen, wenn versucht wird, anhand von Daten mit vielen fehlenden Werten ein Modell zu erstellen. Normalerweise würden diese als Teil der automatischen Datenaufbe-

reitung imputiert. Als Problemumgehung kann ein Baummodell oder ein neuronales Netz mit der Einstellung "Erweitert" verwendet werden, um fehlende ausgewählte Werte zu imputieren.

- Die Genauigkeitsstatistik wird für aufgeteilte Modelle nicht berechnet.

Neuronales Netz

Beim Erstellen von Modellen für große Daten ändern Sie das Ziel in der Regel in "Sehr große Datasets" oder geben Aufteilungen an.

- Fortlaufendes Training vorhandener Standard- oder PSM-Modelle wird nicht unterstützt.
- Das Modellerstellungsziel "Standard" wird nur empfohlen, wenn Aufteilungsfelder so definiert sind, dass die Anzahl an Datensätzen in den einzelnen Aufteilungen nicht "zu groß" ist, wobei die Definition von "zu groß" von der Leistungsstärke einzelner Knoten in Ihrem Hadoop-Cluster abhängt. Im Gegensatz dazu müssen Sie auch darauf bedacht sein, sicherzustellen, dass Aufteilungen nicht so fein definiert sind, dass zu wenige Datensätze für die Erstellung eines Modells vorhanden sind.
- Das Ziel "Boosting" wird nicht unterstützt.
- Das Ziel "Bagging" wird nicht unterstützt.
- Das Ziel "Sehr große Datasets" wird nicht empfohlen, wenn es wenige Datensätze gibt. Häufig wird dann entweder kein Modell oder ein vermindertes Modell erstellt. Sie haben möglicherweise auch Probleme, wenn die Eingabedatensätze nach einer Systematik geordnet werden, die gegen die Zufälligkeitsvoraussetzungen hinter den verwendeten Erstellungsalgorithmen für Ensemblemodelle verstößt.
- Wenn in den Daten viele Werte fehlen, verwenden Sie die Einstellung "Erweitert", um fehlende Werte zu imputieren.
- Die Genauigkeitsstatistik wird für aufgeteilte Modelle nicht berechnet.

C&R-Baum, CHAID, Quest

Beim Erstellen von Modellen für große Daten ändern Sie das Ziel in der Regel in "Sehr große Datasets" oder geben Aufteilungen an.

- Fortlaufendes Training vorhandener PSM-Modelle wird nicht unterstützt.
- Das Modellerstellungsziel "Standard" wird nur empfohlen, wenn Aufteilungsfelder so definiert sind, dass die Anzahl an Datensätzen in den einzelnen Aufteilungen nicht "zu groß" ist, wobei die Definition von "zu groß" von der Leistungsstärke einzelner Knoten in Ihrem Hadoop-Cluster abhängt. Im Gegensatz dazu müssen Sie auch darauf bedacht sein, sicherzustellen, dass Aufteilungen nicht so fein definiert sind, dass zu wenige Datensätze für die Erstellung eines Modells vorhanden sind.
- Das Ziel "Boosting" wird nicht unterstützt.
- Das Ziel "Bagging" wird nicht unterstützt.
- Das Ziel "Sehr große Datasets" wird nicht empfohlen, wenn es wenige Datensätze gibt. Häufig wird dann entweder kein Modell oder ein vermindertes Modell erstellt. Sie haben möglicherweise auch Probleme, wenn die Eingabedatensätze nach einer Systematik geordnet werden, die gegen die Zufälligkeitsvoraussetzungen hinter den verwendeten Erstellungsalgorithmen für Ensemblemodelle verstößt.
- Interaktive Sitzungen werden nicht unterstützt.
- Die Genauigkeitsstatistik wird für aufgeteilte Modelle nicht berechnet.

Modellscoring

Die folgenden Modellnuggets werden für das Scoring unterstützt: C&RT, Quest, CHAID, Linear, Regression, Neuronales Netz, C5.0, Logistisch, Genlin, GLMM, Cox, SVM, Bayes-Netz, TwoStep, KNN, Entscheidungsliste, Diskriminanzanalyse, Selbstlernfunktion, Anomalieerkennung, Apriori, Carma, K-Means, Kohonen, R, Textmining.

- Raw Propensity und Adjusted Propensity werden nicht gescort. Als Problemumgehung können Sie denselben Effekt erzielen, indem Sie die Raw Propensity mithilfe eines Ableitungsknotens mit dem folgenden Ausdruck berechnen: if 'predicted-value' == 'value-of-interest' then 'prob-of-that-value' else 1-'prob-of-that-value' endif
- Beim Scoren eines Modells überprüft Analytic Server nicht, ob alle im Modell verwendeten Felder im Dataset vorhanden sind. Stellen Sie daher vor der Ausführung in Analytic Server sicher, dass dies der Fall ist.

R Die R-Syntax im Nugget sollte aus Operationen bestehen, die jeweils nur für einen Datensatz ausgeführt werden.

Ausgabe
Die Knoten "Matrix", "Analyse", "Data Audit", "Transformieren", "Statistik" und "Mittelwert" werden unterstützt.

Export Ein Stream kann mit einem Analytic Server-Quellenknoten beginnen und mit einem anderen Exportknoten als dem Analytic Server-Exportknoten enden, die Daten werden jedoch von HDFS in SPSS Modeler Server und schließlich an die Exportposition verschoben.

IBM Cognos BI-Quellenknoten

Mit dem IBM Cognos IB-Quellenknoten können Sie Cognos BI-Datenbankdaten oder einzelne Listenberichte in Ihre Data-Mining-Sitzung importieren. Auf diese Weise können Sie die Business Intelligence-Funktionen von Cognos mit den Vorhersageanalysefunktionen von IBM SPSS Modeler kombinieren. Sie können relationale, dimensional modellierte relationale (DMR) und OLAP-Daten importieren.

Wählen Sie über eine Cognos-Serververbindung zunächst einen Speicherort aus, aus dem Daten bzw. Berichte importiert werden sollen. Ein Speicherort enthält ein Cognos-Modell und alle Ordner, Abfragen, Ansichten, Verknüpfungen, URLs und Aufgabendefinitionen, die diesem Modell zuordnet sind. Ein Cognos-Modell definiert Unternehmensregeln, Datenbeschreibungen, Datenbeziehungen, Geschäftsdimensionen und -Hierarchien sowie andere administrative Aufgaben.

Wenn Sie Daten importieren, wählen Sie die zu importierenden Objekte aus dem ausgewählten Paket aus. Zu den importierbaren Objekten gehören Abfragesubjekte (die für Datenbanktabellen stehen) oder einzelne Abfrageelemente (die für Datenbankspalten stehen). Weitere Informationen finden Sie im Thema „Cognos-Objektsymbole”.

Wenn für das Paket Filter definiert wurden, können Sie einen oder mehrere davon importieren. Wenn ein von Ihnen importierter Filter importierten Daten zugeordnet ist, wird der betreffende Filter angewendet, bevor die Daten importiert werden. *Hinweis*: Die zu importierenden Daten müssen im UTF-8-Format vorliegen.

Wenn Sie einen Bericht importieren, wählen Sie ein Paket bzw. einen Ordner in einem Paket mit einem oder mehreren Berichten aus. Anschließend wählen Sie den Bericht aus, der importiert werden soll. *Hinweis*: Es können nur einzelne Listenberichte importiert werden; mehrere Listen werden nicht unterstützt.

Wenn Parameter definiert wurden, entweder für ein Datenobjekt oder für einen Bericht, können Sie Werte für diese Parameter angeben, bevor Sie das Objekt bzw. den Bericht importieren.

Cognos-Objektsymbole

Die verschiedenen Objekttypen, die Sie aus einer Cognos BI-Datenbank importieren können, werden durch unterschiedliche Symbole dargestellt, wie in der folgenden Tabelle zu sehen.

Tabelle 3. Cognos-Objektsymbole.

Symbol	Objekt
	Paket
	Namespace
	Abfragesubjekt
	Abfrageelement
	Maßdimension
	Maß
	Dimension
	Ebenenhierarchie
	Ebene
	Filter
	Bericht
	Eigenständige Berechnung

Importieren von Cognos-Daten

Beim Importieren von Daten aus einer IBM Cognos BI-Datenbank müssen Sie sicherstellen, dass **Modus** auf der Registerkarte "Daten" des IBM Cognos BI-Dialogfensters auf **Daten** gesetzt ist.

Verbindung. Klicken Sie auf die Schaltfläche **Bearbeiten**, um ein Dialogfeld anzuzeigen, in dem Sie die Details einer neuen Cognos-Verbindung definieren können, über die Daten bzw. Berichte importiert werden sollen. Wenn Sie bereits bei einem Cognos-Server über IBM SPSS Modeler angemeldet sind, können Sie auch die Dateils der aktuellen Verbindung bearbeiten. Weitere Informationen finden Sie im Thema „Cognos-Verbindungen” auf Seite 39.

Speicherort. Wenn Sie die Cognos-Serververbindung eingerichtet haben, klicken Sie auf die Schaltfläche **Bearbeiten** neben diesem Feld, um eine Liste der verfügbaren Pakete anzuzeigen, aus denen Sie Inhalte importieren können. Weitere Informationen finden Sie im Thema „Auswählen des Cognos-Standorts” auf Seite 39.

Inhalt. Zeigt den Namen des ausgewählten Pakets und die dem Paket zugewiesenen Namespaces an. Doppelklicken Sie auf einen Namespace, um die Objekte anzuzeigen, die Sie importieren können. Die verschiedenen Objekttypen sind durch unterschiedliche Symbole gekennzeichnet. Weitere Informationen finden Sie im Thema „Cognos-Objektsymbole” auf Seite 36.

Um ein Objekt für den Import auszuwählen, markieren Sie das Objekt und klicken Sie auf den oberen der beiden Rechtspfeile, um das Objekt in den Bereich **Zu importierende Felder** zu verschieben. Wenn Sie auf ein Abfragesubjekt klicken, werden alle entsprechenden Abfrageelemente importiert. Wenn Sie auf ein Abfragesubjekt doppelklicken, wird es erweitert, sodass Sie ein oder mehrere seiner individuellen Ab-

frageelemente auswählen können. Mit Strg-Klicken (individuelle Elemente auswählen), Umschalt-Klicken (mehrere Elemente auswählen) und Strg-A (alle Elemente auswählen) können Sie eine Mehrfachauswahl vornehmen.

Um einen anzuwendenden Filter auszuwählen (sofern für das Paket Filter definiert sind), navigieren Sie im Bereich "Inhalt" zu dem Filter, markieren Sie ihn und klicken Sie auf den unteren der beiden Rechtspfeile, um den Filter in den Bereich **Anzuwendende Filter** zu verschieben. Mit Strg-Klicken (einzelne Filter auswählen) und Umschalt-Klicken (zusammenhängenden Block von Filtern auswählen) können Sie eine Mehrfachauswahl vornehmen.

Zu importierende Felder. Listet die Datenbankobjekte auf, die laut Ihrer Auswahl zur Verarbeitung in IBM SPSS Modeler importiert werden. Wenn Sie ein bestimmtes Objekt nicht mehr benötigen, wählen Sie es aus und klicken Sie auf den Linkspfeil, um es wieder in den Bereich **Inhalt** zu verschieben. Sie können Mehrfachauswahlen auf dieselbe Weise wie für **Inhalt** vornehmen.

Anzuwendende Filter. Listet die Filter auf, die laut Ihrer Auswahl vor dem Import auf die Daten angewendet werden sollen. Wenn Sie einen bestimmten Filter nicht mehr benötigen, wählen Sie ihn aus und klicken Sie auf den Linkspfeil, um ihn wieder in den Bereich **Inhalt** zu verschieben. Sie können Mehrfachauswahlen auf dieselbe Weise wie für **Inhalt** vornehmen.

Parameter. Wenn diese Schaltfläche aktiviert ist, sind für das ausgewählte Objekt Parameter definiert. Mit Parametern können Sie vor dem Import der Daten Anpassungen vornehmen (beispielsweise eine parametrisierte Berechnung durchführen). Wenn Parameter definiert sind, jedoch keine Standardwerte angegeben wurden, wird auf der Schaltfläche ein Warndreieck angezeigt. Klicken Sie auf die Schaltfläche, um die Parameter anzuzeigen und gegebenenfalls zu bearbeiten. Wenn die Schaltfläche inaktiviert ist, sind für den Bericht keine Parameter definiert.

Daten vor dem Importieren aggregieren. Aktivieren Sie dieses Kontrollkästchen, wenn Sie statt Rohdaten aggregierte Daten importieren möchten.

Importieren von Cognos-Berichten

Beim Importieren von vordefinierten Berichten aus einer IBM Cognos BI-Datenbank müssen Sie sicherstellen, dass **Modus** auf der Registerkarte "Daten" des IBM Cognos BI-Dialogfensters auf **Bericht** gesetzt ist. *Hinweis*: Es können nur einzelne Listenberichte importiert werden; mehrere Listen werden nicht unterstützt.

Verbindung. Klicken Sie auf die Schaltfläche **Bearbeiten**, um ein Dialogfeld anzuzeigen, in dem Sie die Details einer neuen Cognos-Verbindung definieren können, über die Daten bzw. Berichte importiert werden sollen. Wenn Sie bereits bei einem Cognos-Server über IBM SPSS Modeler angemeldet sind, können Sie auch die Dateils der aktuellen Verbindung bearbeiten. Weitere Informationen finden Sie im Thema „Cognos-Verbindungen” auf Seite 39.

Speicherort. Wenn Sie die Cognos-Serververbindung eingerichtet haben, klicken Sie auf die Schaltfläche **Bearbeiten** neben diesem Feld, um eine Liste der verfügbaren Pakete anzuzeigen, aus denen Sie Inhalte importieren können. Weitere Informationen finden Sie im Thema „Auswählen des Cognos-Standorts” auf Seite 39.

Inhalt. Zeigt den Namen des ausgewählten Pakets bzw. des ausgewählten Ordners an, der Berichte enthält. Navigieren Sie zu einem Bericht, wählen Sie ihn aus und klicken Sie auf den Rechtspfeil, um den Bericht in das Feld **Zu importierender Bericht** zu verschieben.

Zu importierender Bericht. Gibt den Bericht an, der laut Ihrer Auswahl in IBM SPSS Modeler importiert wird. Wenn Sie den Bericht nicht mehr benötigen, wählen Sie ihn aus und klicken Sie auf den Linkspfeil, um ihn wieder in den Bereich **Inhalt** zu verschieben, oder verschieben Sie einen anderen Bericht in dieses Feld.

Parameter. Wenn diese Schaltfläche aktiviert ist, sind für den ausgewählten Bericht Parameter definiert. Sie können Parameter verwenden, um vor dem Import des Berichts Anpassungen vorzunehmen (z. B. Angabe eines Start- und Enddatums für Berichtdaten). Wenn Parameter definiert sind, jedoch keine Standardwerte angegeben wurden, wird auf der Schaltfläche ein Warndreieck angezeigt. Klicken Sie auf die Schaltfläche, um die Parameter anzuzeigen und gegebenenfalls zu bearbeiten. Wenn die Schaltfläche inaktiviert ist, sind für den Bericht keine Parameter definiert.

Cognos-Verbindungen

Im Dialogfeld "Cognos-Verbindungen" können Sie den Cognos BI-Server auswählen, von dem Sie Datenbankobjekte importieren bzw. an den Sie Datenbankobjekte exportieren möchten.

Cognos-Server-URL. Geben Sie die URL des Cognos BI-Servers ein, den Sie für die Import- bzw. Exportvorgänge verwenden möchten. Dies ist der Wert der Umgebungseigenschaft "External dispatcher URI" (Externe Dispatcher-URI) der IBM Cognos-Konfiguration auf dem Cognos BI-Server. Wenden Sie sich an Ihren Cognos-Systemadministrator, wenn Sie sich nicht sicher sind, welche URL Sie verwenden müssen.

Modalwert. Wählen Sie **Berechtigungsnachweise festlegen**, wenn Sie sich mit einem spezifischen Cognos-Namespace, Benutzernamen und Kennwort anmelden möchten (z. B. als Administrator). Wählen Sie **Anonyme Verbindung verwenden**, um sich ohne Benutzerberechtigungsnachweise anzumelden. Sie füllen in diesem Fall keine weiteren Felder aus.

Namespace. Geben Sie den Sicherheitsanbieter für die Authentifizierung bei Cognos an, mit dem Sie sich beim Server anmelden möchten. Der Authentifizierungsanbieter dient dazu, Benutzer, Gruppen und Rollen zu definieren und zu verwalten und den Authentifizierungsprozess zu steuern.

Benutzername. Geben Sie den Cognos-Benutzernamen ein, mit dem die Anmeldung beim Server erfolgen soll.

Kennwort. Geben Sie das Kennwort ein, das zum angegebenen Benutzernamen gehört.

Als Standard speichern. Klicken Sie auf diese Schaltfläche, um diese Einstellung als Standardeinstellungen wiederherzustellen, damit Sie sie nicht jedesmal, wenn Sie den Knoten öffnen, neu eingeben müssen.

Auswählen des Cognos-Standorts

Im Dialogfeld "Speicherort angeben" können Sie ein Cognos-Paket angeben, aus dem Daten importiert werden sollen, bzw. ein Paket bzw. einen Ordner, aus dem Berichte importiert werden sollen.

Öffentliche Ordner. Wenn Sie Daten importieren, werden hier die Pakete und Ordner aufgelistet, die auf dem ausgewählten Server zur Verfügung stehen. Wählen Sie das Paket aus, das Sie verwenden möchten, und klicken Sie auf **OK**. Sie können nur ein Paket pro Cognos BI-Quellenknoten auswählen.

Wenn Sie Berichte importieren, werden hier die Ordner und Pakete mit Berichten aufgelistet, auf denen ausgewählte Server zur Verfügung stehen. Wählen Sie ein Paket oder einen Berichtordner aus und klicken Sie auf **OK**. Sie können nur ein einziges Paket bzw. nur einen einzigen Berichtordner pro Cognos BI-Quellenknoten auswählen, die Berichtordner können jedoch andere Berichtordner sowie einzelne Berichte enthalten.

Angeben von Parametern für Daten bzw. Berichte

Wenn Parameter in Cognos BI definiert wurden, entweder für ein Datenobjekt oder für einen Bericht, können Sie Werte für diese Parameter angeben, bevor Sie das Objekt bzw. den Bericht importieren. Ein Beispiel für Parameter für einen Bericht wären die Anfangs- und Enddaten für die Berichtsinhalte.

Name. Der Name des Parameters laut Angabe in der Cognos BI-Datenbank.

Typ. Eine Beschreibung des Parameters.

Wert. Der dem Parameter zuzuweisende Wert. Doppelklicken Sie zur Eingabe bzw. Bearbeitung eines Werts auf die entsprechende Zelle in der Tabelle. Hier werden keine Werte validiert, etwaige ungültige Werte werden somit zur Laufzeit entdeckt.

Ungültige Parameter automatisch aus Tabelle entfernen. Diese Option ist standardmäßig ausgewählt und entfernt alle ungültigen Parameter, die im Datenobjekt bzw. Bericht gefunden werden.

IBM Cognos TM1-Quellenknoten

Mit dem IBM Cognos TM1-Quellenknoten können Sie Cognos TM1-Daten in Ihre Data-Mining-Sitzung importieren. Auf diese Weise können Sie die Unternehmensplanungsfunktionen von Cognos mit den Vorhersageanalysefunktionen von IBM SPSS Modeler kombinieren. Sie können eine abgeflachte Version der mehrdimensionalen OLAP-Cubedaten importieren.

Sie müssen die Daten in TM1 ändern, bevor die Daten importiert werden. *Hinweis*: Die zu importierenden Daten müssen im UTF-8-Format vorliegen.

Wählen Sie über eine Cognos-PM-Hubverbindung zunächst einen TM1-Server aus, aus dem die Daten importiert werden sollen. Ein Server enthält mindestens einen TM1-Cube. Wählen Sie anschließend den erforderlichen Cube und im Cube die Spalten und Zeilen aus, die Sie importieren möchten.

Anmerkung: Bevor Sie die TM1-Quellen- oder -Exportknoten in SPSS Modeler verwenden können, müssen Sie die folgenden drei Prozesse aus Modeler auf den TM1-Server kopieren : *ExportToSPSS.pro*, *ImportFromSPSS.pro* und *SPSSCreateNewMeasures.pro*. Um dem TM1-Server diese Prozesse hinzuzufügen, müssen Sie die Dateien in das Datenverzeichnis des TM1-Servers kopieren und den TM1-Server erneut starten. Diese Dateien sind im folgenden Verzeichnis verfügbar: <*Modeler-Installationsverzeichnis*>/ext/bin/pasw.tm1/scripts.

Importieren von Cognos TM1-Daten

Wählen Sie auf der Registerkarte "Daten" des Dialogfelds "IBM Cognos TM1" den relevanten TM1-PM-Hub und den zugehörigen Server, den Cube sowie Datendetails aus, um Daten aus einer IBM Cognos TM1-Datenbank zu importieren.

Anmerkung: Bevor Sie Daten importieren, müssen Sie in TM1 eine Vorverarbeitung durchführen, um sicherzustellen, dass die Daten in einem Format vorliegen, das von IBM SPSS Modeler erkannt werden kann. Dazu gehört das Filtern Ihrer Daten mithilfe des Subseteditors, um die Ansicht für den Import in die richtige Größe und Form zu bringen.

PM-System. Geben Sie die URL des Hubs ein, der den TM1-Server enthält, zu dem Sie eine Verbindung herstellen wollen.

TM1-Server. Wenn Sie die Cognos-Hubverbindung eingerichtet haben, wählen Sie den Server aus, der die zu importierenden Daten enthält, und klicken Sie auf **Anmeldung**. Wenn Sie bisher noch keine Verbindung zu diesem Server hergestellt haben, werden Sie zur Eingabe des Benutzernamens und des Kennworts aufgefordert. Alternativ können Sie einen anderen Server auswählen.

Zu importierende TM1-Cube-Ansicht auswählen. Zeigt den Namen der Cubes im TM1-Server an, aus denen Sie Daten importieren können. Doppelklicken Sie auf einen Cube, um die Ansichtsdaten anzuzeigen, die Sie importieren können.

Wählen Sie zur Auswahl der zu importierenden Daten die Ansicht aus und klicken Sie auf den Rechtspfeil, um die Ansicht in den Bereich **Zu importierende Ansicht** zu verschieben. Wenn die von Ihnen benötigte Ansicht nicht sichtbar ist, doppelklicken Sie auf einen Cube, um die entsprechende Ansichtsliste einzublenden.

Spaltendimension(en). Listet den Namen der Spaltendimension in den Daten auf, die Sie für den Import ausgewählt haben. Blättern Sie in der Liste der Ebenen und wählen Sie die erforderliche Liste aus.

Zeilendimension(en). Listet den Namen der Zeilendimension in den Daten auf, die Sie für den Import ausgewählt haben. Blättern Sie in der Liste der Ebenen und wählen Sie die erforderliche Liste aus.

Kontextdimension(en). Nur Anzeige. Zeigt die Kontextdimensionen für die ausgewählten Spalten und Zeilen an.

SAS-Quellenknoten

Hinweis: Diese Funktion ist in SPSS Modeler Professional und SPSS Modeler Premium verfügbar.

Mit dem SAS-Quellenknoten können Sie SAS-Daten in Ihre Data-Mining-Sitzung importieren. Sie können vier Dateitypen importieren:

- SAS für Windows/OS2 (*.sd2*)
- SAS für UNIX (*.ssd*)
- SAS-Transportdatei (*.tpt*)
- SAS Version 7/8/9 (*.sas7bdat*)

Beim Importieren der Daten werden alle Variablen beibehalten und kein Variablentyp wird geändert. Alle Fälle werden ausgewählt.

Festlegen von Optionen für den SAS-Quellenknoten

Importieren. Wählen Sie den zu transportierenden SAS-Dateityp aus. Zur Auswahl stehen die Optionen **SAS für Windows/OS2 (.sd2)**, **SAS für UNIX (.SSD)**, **SAS-Transportdatei (.tpt)** oder **SAS Version 7/8/9 (.sas7bdat)**.

Datei importieren. Geben Sie den Namen der Datei an. Sie können einen Dateinamen eingeben oder auf die Schaltfläche mit den Auslassungspunkten (**...**) klicken, um zum Speicherort der Datei zu navigieren.

Element. Wählen Sie ein Element für den Import aus der oben ausgewählten SAS-Transportdatei aus. Sie können einen Elementnamen eingeben oder auf **Auswählen** klicken, um durch alle Elemente in der Datei zu blättern.

Benutzerformate aus SAS-Datendatei lesen. Wählen Sie diese Option aus, um Benutzerformate zu lesen. SAS-Dateien speichern Daten und Datenformate (wie Variablenbeschriftungen) in verschiedenen Dateien. In den meisten Fällen sollen die Formate ebenfalls importiert werden. Bei einem großen Dataset ist es jedoch empfehlenswert, diese Option zu inaktivieren, um Speicher zu sparen.

Formatdatei. Wenn eine Formatdatei erforderlich ist, ist dieses Textfeld aktiviert. Sie können einen Dateinamen eingeben oder auf die Schaltfläche mit den Auslassungspunkten (**...**) klicken, um zum Speicherort der Datei zu navigieren.

Variablennamen. Wählen Sie eine Methode zur Behandlung von Variablennamen und -beschriftungen beim Importieren aus einer SAS-Datei aus. Metadaten, die Sie hier einschließen, bleiben während Ihrer Arbeit in IBM SPSS Modeler erhalten und können zur Verwendung in SAS wieder exportiert werden.

- **Namen und Beschriftungen lesen.** Wählen Sie diese Option aus, wenn sowohl Variablennamen als auch -beschriftungen in IBM SPSS Modeler eingelesen werden sollen. Standardmäßig ist diese Option

ausgewählt und Variablennamen werden im Typknoten angezeigt. Beschriftungen können je nach den im Dialogfeld "Streameigenschaften" angegebenen Optionen in Expression Builder, Diagrammen, Modellbrowsern und anderen Ausgabearten angezeigt werden.

- **Beschriftungen als Namen lesen.** Wählen Sie diese Option aus, um statt der kurzen Feldnamen die beschreibenden Variablenbeschriftungen aus der SAS-Datei zu lesen und diese Beschriftungen als Variablennamen in IBM SPSS Modeler zu verwenden.

Excel-Quellenknoten

Mit dem Excel-Quellenknoten können Sie Daten aus einer beliebigen Version von Microsoft Excel importieren.

Dateityp. Wählen Sie den Excel-Dateityp, den Sie importieren möchten.

Datei importieren. Gibt Namen und Speicherort der zu importierenden Tabellenkalkulationsdatei an.

Benannten Bereich verwenden. Ermöglicht die Angabe eines benannten Zellenbereichs, wie im Excel-Arbeitsblatt definiert. Klicken Sie auf die Schaltfläche mit den Auslassungspunkten (**...**), um eine Auswahl aus der Liste der verfügbaren Bereiche zu treffen. Wenn ein benannter Bereich verwendet wird, sind andere Einstellungen für Arbeitsblatt und Datenbereich nicht mehr anwendbar und werden daher inaktiviert.

Arbeitsblatt auswählen. Gibt das zu importierende Arbeitsblatt an, entweder nach Index oder nach Namen.

- **Nach Index.** Geben Sie den Indexwert für das zu importierende Arbeitsblatt an. Beginnen Sie mit 0 für das erste Arbeitsblatt, 1 für das zweite Arbeitsblatt usw.
- **Nach Name.** Geben Sie den Namen des importierenden Arbeitsblattes an. Klicken Sie auf die Schaltfläche mit den Auslassungspunkten (**...**), um eine Auswahl aus der Liste der verfügbaren Arbeitsblätter zu treffen.

Bereich auf Arbeitsblatt. Sie können Daten beginnend mit der ersten nicht leeren Zeile oder mit einem expliziten Zellenbereich importieren.

- **Bereich beginnt in erster nicht leere Zeile.** Sucht die erste nicht leere Zelle und verwendet diese als linke obere Ecke des Datenbereichs.
- **Eindeutiger Zellenbereich.** Ermöglicht die Angabe eines expliziten Bereichs nach Zeile und Spalte. Beispielsweise können Sie für den Excel-Bereich `A1:D5` in das erste Feld `A1` und in das zweite Feld `D5` eingeben (oder alternativ `R1C1` und `R5C4`). Alle Zeilen im angegebenen Bereich werden ausgegeben, einschließlich der Leerzeilen.

In leeren Zeilen. Wenn mehrere leere Zeilen gefunden werden, können Sie mit **Lesen stoppen** angeben, dass der Lesevorgang angehalten werden soll, oder mit **Leere Zeilen zurückgeben** festlegen, dass alle Daten bis zum Ende des Arbeitsblatts gelesen werden sollen, einschließlich Leerzeilen.

Erste Zeile enthält Spaltennamen. Gibt an, dass die erste Zeile im angegebenen Bereich als Feldnamen (Spaltennamen) verwendet werden soll. Wenn diese Option nicht ausgewählt ist, werden Feldnamen automatisch generiert.

Feld-Speichertyp und -Messniveau

Beim Lesen von Werten aus Excel werden Felder mit numerischem Speicher standardmäßig mit einem Messniveau von *Stetig* und Zeichenfolgenfelder als *Nominal* eingelesen. Sie können auf der Registerkarte "Typ" manuell das Messniveau ("Stetig" bzw. "Nominal") ändern, der Speichertyp wird jedoch automatisch bestimmt (allerdings kann er, falls erforderlich, mithilfe einer Konvertierungsfunktion, wie beispielsweise `to_integer`, in einem Füller- oder Ableitungsknoten geändert werden. Weitere Informationen finden Sie im Thema „Festlegen von Feldspeicher und Formatierung” auf Seite 24.

Standardmäßig werden Felder mit einer Mischung aus numerischen Werten und Zeichenfolgewerten als Zahlen eingelesen. Alle Zeichenfolgewerte werden also in IBM SPSS Modeler auf null (systemdefiniert fehlend) gesetzt. Dies liegt daran, dass IBM SPSS Modeler, anders als Excel, keine gemischten Speichertypen innerhalb eines Felds zulässt. Um dies zu vermeiden, können Sie das Zellenformat in der Excel-Tabelle manuell auf **Text** setzen. Dadurch werden alle Werte (einschließlich Zahlen) als Zeichenfolgen eingelesen.

XML-Quellenknoten

Hinweis: Diese Funktion ist in SPSS Modeler Professional und SPSS Modeler Premium verfügbar.

Mit dem XML-Quellenknoten können Sie die Daten aus einer Datei im XML-Format in einen IBM SPSS Modeler-Stream importieren. XML ist eine Standardsprache für den Datenaustausch und gilt für viele Unternehmen als das bevorzugte Format für diesen Zweck. So möchte beispielsweise eine Steuerbehörde Daten aus Steuererklärungen analysieren, die online und im XML-Format übermittelt wurden (siehe http://www.w3.org/standards/xml/).

Durch Importieren von XML-Daten in einen IBM SPSS Modeler-Stream können Sie zahlreiche Vorhersageanalysefunktionen an der Quelle ausführen. Die XML-Daten werden in ein Tabellenformat gegliedert, bei dem die Spalten den verschiedenen Verschachtelungsniveaus der XML-Elemente und Attribute entsprechen. Die XML-Objekte werden im XPath-Format angezeigt (siehe http://www.w3.org/TR/xpath20/).

Einzelne Datei lesen. Standardmäßig liest IBM SPSS Modeler eine einzelne Datei, die Sie im Feld **XML-Datenquelle** angeben.

Alle XML-Dateien in einem Verzeichnis lesen. Wenn Sie diese Option wählen, werden alle XML-Dateien in einem bestimmten Verzeichnis gelesen. Geben Sie die Position in dem Feld **Verzeichnis** an, das angezeigt wird. Aktivieren Sie das Kontrollkästchen **Unterverzeichnisse einschließen**, um zusätzlich XML-Dateien aus allen Unterverzeichnissen des angegebenen Verzeichnisses zu lesen.

XML-Datenquelle. Geben Sie den vollständigen Pfad und Dateinamen der XML-Quellendatei an, die Sie importieren möchten, oder nutzen Sie die Schaltfläche "Durchsuchen", um die Datei zu finden.

XML-Schema. (Optional) Geben Sie den vollständigen Pfad und Dateinamen einer XSD- oder DTD-Datei an, aus der die XML-Struktur gelesen werden soll, oder verwenden Sie die Schaltfläche "Durchsuchen", um diese Datei zu finden. Wenn Sie dieses Feld frei lassen, wird die Struktur aus der XML-Quellendatei gelesen. Eine XSD- oder DTD-Datei kann mehr als ein Stammelement besitzen. In diesem Fall wird ein Dialogfeld angezeigt, in dem Sie das gewünschte Stammelement auswählen, wenn Sie den Fokus auf ein anderes Feld wechseln. Weitere Informationen finden Sie im Thema „Auswählen aus mehreren Stammelementen” auf Seite 44.

XML-Struktur. Ein hierarchischer Baum, der die Struktur der XML-Quellendatei anzeigt (oder das Schema, sofern Sie eines im Feld **XML-Schema** angegeben haben). Zum Definieren einer Datensatzgrenze wählen Sie ein Element aus und klicken auf die Schaltfläche mit dem Rechtspfeil, um das Objekt in das Feld **Datensätze** zu kopieren.

Attribute anzeigen. Zeigt die Attribute der XML-Elemente in dem Feld **XML-Struktur** an oder blendet sie aus.

Datensätze (XPath-Ausdruck). Zeigt die XPath-Syntax für ein Element, das aus dem Feld "XML-Struktur" kopiert wurde. Dieses Element wird dann in der XML-Struktur hervorgehoben und definiert die Datensatzgrenze. Jedes Mal, wenn dieses Element in der Quellendatei gefunden wird, wird ein neuer Datensatz erstellt. Wenn das Feld leer ist, wird das erste untergeordnete Element unter dem Stamm als Datensatzgrenze verwendet.

Alle Daten lesen. Standardmäßig werden alle Daten in der Quellendatei in den Stream eingelesen.

Zu lesende Daten angeben. Wählen Sie diese Option, wenn Sie einzelne Elemente, Attribute oder beides importieren möchten. Durch Auswählen dieser Option wird die Feldertabelle aktiviert, in der Sie die zu importierenden Daten angeben können.

Felder. In dieser Tabelle werden die für den Import ausgewählten Elemente und Attribute angezeigt, wenn Sie die Option **Zu lesende Daten angeben** ausgewählt haben. Sie können die XPath-Syntax eines Elements oder Attributs entweder direkt in die Spalte "XPath" eingeben oder ein Element oder Attribut in der XML-Struktur auswählen und auf die Schaltfläche mit dem Rechtspfeil klicken, um das Objekt in die Tabelle zu kopieren. Zum Kopieren aller untergeordneten Elemente und Attribute eines Elements wählen Sie das Element in der XML-Struktur aus und klicken Sie auf die Schaltfläche mit dem Doppelpfeil.

- **XPath.** Die XPath-Syntax der zu importierenden Objekte.
- **Lage.** Die Position in der XML-Struktur der zu importierenden Objekte. **Fester Weg** zeigt den Weg des Objekts im Verhältnis zu dem in der XML-Struktur hervorgehobenen Element (oder dem ersten untergeordneten Element unter dem Stamm, wenn kein Element hervorgehoben ist). **Beliebiger Ort** kennzeichnet ein Objekt mit dem angegebenen Namen an einem beliebigen Ort in der XML-Struktur. **Benutzerdefiniert** wird angezeigt, wenn Sie den Ort direkt in die XPath-Spalte eingeben.

Auswählen aus mehreren Stammelementen

Während eine ordnungsgemäß erstellte XML-Datei nur ein einzelnes Stammelement besitzen kann, kann eine XSD- oder DTD-Datei mehrere Stammelemente enthalten. Wenn eine der Wurzeln mit der Wurzel in der XML-Quellendatei übereinstimmt, wird dieses Stammelement verwendet. Andernfalls müssen Sie das zu verwendende Stammelement auswählen.

Wählen Sie den anzuzeigenden Stamm. Wählen Sie das zu verwendende Stammelement aus. Standardmäßig wird das erste Stammelement in der XSD- oder DTD-Struktur verwendet.

Entfernen unerwünschter Leerzeichen aus XML-Quellendaten

Zeilenumbrüche können in XML-Quellendaten mit der Zeichenkombination `[CR][LF]` erzeugt werden. In manchen Fällen können diese Zeilenumbrüche mitten im Text auftreten, zum Beispiel:

```
<description>Ein tiefer Einblick in das Erstellen von Anwendungen[CR][LF]
mit XML.</description>
```

Diese Zeilenumbrüche sind unter Umständen nicht sichtbar, wenn die Datei in bestimmten Anwendungen, etwa einem Web-Browser, geöffnet wird. Wenn die Daten jedoch durch den XML-Quellenknoten in den Stream eingelesen werden, werden die Zeilenumbrüche in eine Reihe von Leerzeichen umgewandelt.

Sie können diese unerwünschten Leerzeichen mithilfe eines Füllerknotens beseitigen:

Wie Sie dazu vorgehen, erfahren Sie an einem Beispiel:

1. Hängen Sie einen Füllerknoten an den XML-Quellenknoten an.
2. Öffnen Sie den Füllerknoten und wählen Sie mithilfe der Feldauswahl das Feld mit den unerwünschten Leerzeichen aus.
3. Setzen Sie die Option **Ersetzen** auf **Anhand der Bedingung** und die Option **Bedingung** auf **wahr**.
4. Geben Sie `replace(" ","",@FIELD)` in das Feld **Ersetzen durch** ein und klicken Sie auf "OK".
5. Hängen Sie einen Tabellenknoten an den Füllerknoten an und führen Sie den Stream aus.

In der Ausgabe des Tabellenknotens wird der Text nun ohne die zusätzlichen Leerzeichen angezeigt.

Simulationsgenerierungsknoten

Der Simulationsgenerierungsknoten bietet eine einfache Möglichkeit, simulierte Daten entweder ohne historische Daten anhand von durch den Benutzer angegebenen statistischen Verteilungen oder automatisch anhand der Verteilungen aus der Ausführung eines Simulationsanpassungsknotens für vorhandene historische Daten zu generieren. Die Generierung simulierter Daten ist hilfreich, wenn Sie das Ergebnis eines Vorhersagemodells bei einer Unsicherheit in den Modelleingaben auswerten möchten.

Erstellen von Daten ohne historische Daten

Der Simulationsgenerierungsknoten ist von der Palette "Quellen" aus verfügbar und kann dem Streamerstellungsbereich direkt hinzugefügt werden.

1. Klicken Sie auf die Registerkarte **Datenquellen** der Knotenpalette.
2. Fügen Sie den Simulationsgenerierungsknoten durch Ziehen und Ablegen oder durch Doppelklicken dem Streamerstellungsbereich hinzu.
3. Öffnen Sie das entsprechende Dialogfeld durch Doppelklicken und geben Sie Felder, Speichertypen, statistische Verteilungen und Verteilungsparameter an.

Hinweis: In der Palette "Quellen" ausgewählte Simulationsgenerierungsknoten sind völlig leer und enthalten weder Felder noch Verteilungsinformationen. Somit können Sie simulierte Daten ohne historische Daten völlig neu erstellen.

Generieren simulierter Daten anhand vorhandener historischer Daten

Ein Simulationsgenerierungsknoten kann auch durch die Ausführung eines Endknotens zur Simulationsanpassung erstellt werden:

1. Klicken Sie mit der rechten Maustaste auf den Simulationsanpassungsknoten und wählen Sie im Menü **Ausführen** aus.
2. Der Simulationsgenerierungsknoten wird im Streamerstellungsbereich mit einem Aktualisierungslink zum Simulationsanpassungsknoten angezeigt.
3. Der Simulationsgenerierungsknoten übernimmt bei seiner Generierung alle Felder, Speichertypen und statistischen Verteilungsinformationen vom Simulationsanpassungsknoten.

Definieren eines Aktualisierungslinks zu einem Simulationsanpassungsknoten

Sie können einen Link zwischen einem Simulationsgenerierungsknoten und einem Simulationsanpassungsknoten erstellen. Dies ist hilfreich, wenn Sie mindestens ein Feld mit den Informationen der am besten angepassten Verteilung aktualisieren möchten, die bei der Anpassung an historische Daten bestimmt wird.

1. Klicken Sie mit der rechten Maustaste auf den Simulationsgenerierungsknoten.
2. Wählen Sie im Menü **Aktualisierungslink definieren** aus. Der Cursor nimmt die Form eines Verknüpfungscursors an.
3. Klicken Sie auf einen anderen Knoten. Wenn dieser Knoten ein Simulationsanpassungsknoten ist, wird ein Link erstellt. Wenn dieser Knoten kein Simulationsanpassungsknoten ist, wird kein Link erstellt und der Cursor nimmt wieder die Form eines normalen Cursors an.

Wenn sich die Felder im Simulationsanpassungsknoten von denen im Simulationsgenerierungsknoten unterscheiden, wird eine Nachricht angezeigt, die Sie darüber informiert, dass ein Unterschied vorhanden ist.

Wenn der verlinkte Simulationsgenerierungsknoten mithilfe des Simulationsanpassungsknotens aktualisiert wird, hängt das Ergebnis davon ab, ob in beiden Knoten dieselben Felder vorhanden sind und ob die Felder im Simulationsgenerierungsknoten entsperrt sind. In der folgenden Tabelle werden die Ergebnisse der Aktualisierung eines Simulationsanpassungsknotens angezeigt.

Tabelle 4. Ergebnisse der Aktualisierung eines Simulationsanpassungsknotens

	Feld in Simulation Anpassungsknoten	
Feld in Simulationsgenerierungsknoten	**Vorhanden**	**Fehlend**
Vorhanden und entsperrt.	Feld wird überschrieben.	Feld wird gelöscht.
Fehlend.	Feld wird hinzugefügt.	Keine Änderung.
Vorhanden und gesperrt.	Die Verteilung des Felds wird nicht überschrieben. Die Informationen im Dialogfeld **Details zur Anpassungsgüte** und die Korrelationen werden aktualisiert.	Das Feld wird nicht überschrieben. Die Korrelationen werden auf null gesetzt.
Das Kontrollkästchen **Bei erneuter Anpassung Min und Max nicht löschen** ist ausgewählt.	Das Feld wird überschrieben, mit Ausnahme der	Werte in der Spalte "Min,Max".
Das Kontrollkästchen **Korrelationen bei erneuter Anpassung nicht neu berechnen** ist ausgewählt.	Wenn das Feld entsperrt wird, wird es überschrieben.	Die Korrelationen werden nicht überschrieben.

Entfernen eines Aktualisierungslinks zu einem Simulationsanpassungsknoten

Sie können einen Link zwischen einem Simulationsgenerierungsknoten und einem Simulationsanpassungsknoten entfernen, indem Sie die folgenden Schritte ausführen:

1. Klicken Sie mit der rechten Maustaste auf den Simulationsgenerierungsknoten.
2. Wählen Sie im Menü **Aktualisierungslink entfernen** aus. Der Link wird entfernt.

Festlegen der Optionen für den Simulationsgenerierungsknoten

Mit den Optionen auf der Registerkarte "Daten" des Dialogfelds des Simulationsgenerierungsknotens können Sie folgende Aktionen ausführen:

- Die statistischen Verteilungsinformationen für die Felder anzeigen, angeben und bearbeiten
- Die Korrelationen zwischen den Feldern anzeigen, angeben und bearbeiten
- Die Anzahl der Iterationen und der zu simulierenden Fälle angeben

Element auswählen. Ermöglicht Ihnen, zwischen den drei Ansichten des Simulationsgenerierungsknotens zu wechseln: "Simulierte Felder", "Korrelationen" und "Erweiterte Optionen".

Ansicht "Simulierte Felder"

Wenn der Simulationsgenerierungsknoten über einen Simulationsanpassungsknoten anhand historischer Daten generiert oder aktualisiert wurde, können Sie in der Ansicht "Simulierte Felder" die statistischen Verteilungsinformationen für die einzelnen Felder anzeigen und bearbeiten. Die folgenden Informationen zu jedem Feld werden vom Simulationsanpassungsknoten in die Registerkarte **Typen** des Simulationsgenerierungsknotens kopiert:

- Messniveau
- Werte
- Fehlend
- Überprüfen
- Rolle

Wenn Sie nicht über historische Daten verfügen, können Sie Felder definieren und ihre Verteilungen angeben, indem Sie einen Speichertyp und einen Verteilungstyp auswählen und die erforderlichen Parame-

ter eingeben. Wenn Daten auf diese Weise generiert werden, bedeutet dies, dass Informationen zum Messniveau der einzelnen Felder erst verfügbar sind, wenn die Daten instanziiert werden, z. B. auf der Registerkarte **Typen** oder in einem Typknoten.

Die Ansicht "Simulierte Felder" enthält einige Tools, mit denen Sie die folgenden Tasks ausführen können:

- Felder hinzufügen und entfernen
- Reihenfolge der Felder in der Anzeige ändern
- Einen Speichertyp für jedes Feld angeben
- Eine statistische Verteilung für jedes Feld angeben
- Parameterwerte für die statistische Verteilung der einzelnen Felder angeben

Simulierte Felder. Diese Tabelle enthält eine leere Zeile, wenn der Simulationsgenerierungsknoten über die Palette "Quellen" dem Streamerstellungsbereich hinzugefügt wurde. Wenn diese Zeile bearbeitet wird, wird unten in der Tabelle eine neue leere Zeile hinzugefügt. Wenn der Simulationsgenerierungsknoten über einen Simulationsanpassungsknoten erstellt wurde, enthält diese Tabelle eine Zeile für jedes Feld der historischen Daten. Durch Klicken auf das Symbol **Neues Feld hinzufügen** können der Tabelle zusätzliche Zeilen hinzugefügt werden.

Die Tabelle "Simulierte Felder" besteht aus den folgenden Spalten:

- **Feld.** Enthält die Namen der Felder. Die Feldnamen können durch eine Eingabe in die Zellen bearbeitet werden.
- **Speichertyp.** Die Zellen in dieser Spalte enthalten eine Dropdown-Liste der Speichertypen. Verfügbare Speichertypen sind **Zeichenfolge**, **Ganzzahl**, **Reelle Zahl**, **Zeit**, **Datum** und **Zeitmarke**. Durch die Auswahl des Speichertyps wird festgelegt, welche Verteilungen in der Spalte "Verteilung" verfügbar sind. Wenn der Simulationsgenerierungsknoten über einen Simulationsanpassungsknoten erstellt wurde, wird der Speichertyp aus dem Simulationsanpassungsknoten kopiert.

  Hinweis: Für Felder mit Datum/Uhrzeitspeichertypen müssen Sie die Verteilungsparameter als ganze Zahlen angeben. Wenn Sie beispielsweise den 1. Januar 1970 als mittleres Datum angeben wollen, verwenden Sie die ganze Zahl 0. Die ganze Zahl mit Vorzeichen steht für die Anzahl der Sekunden nach (oder vor) Mitternacht am 1. Januar 1970.
- **Status.** Symbole in der Spalte "Status" geben den Anpassungsstatus für jedes Feld an.

Für das Feld wurde keine Verteilung angegeben oder mindestens ein Parameter fehlt. Um die Simulation auszuführen, müssen Sie eine Verteilung für dieses Feld angeben und gültige Werte für die Parameter eingeben.

Das Feld wird auf die Verteilung mit der besten Anpassung gesetzt.
Anmerkung: Dieses Symbol kann nur angezeigt werden, wenn der Simulationsgenerierungsknoten über einen Simulationsanpassungsknoten erstellt wird.

Die Verteilung mit der besten Anpassung wurde über das Unterdialogfeld **Details zur Anpassungsgüte** durch eine alternative Verteilung ersetzt. Weitere Informationen finden Sie im Thema „Details zur Anpassungsgüte” auf Seite 52.

Die Verteilung wurde manuell angegeben oder bearbeitet und kann einen Parameter enthalten, der auf mehreren Niveaus angegeben wurde.

- **Gesperrt.** Wenn ein simuliertes Feld durch Auswählen des Kontrollkästchens in der Spalte mit dem Sperrsymbol gesperrt wird, wird das Feld von der automatischen Aktualisierung durch einen verlinkten Simulationsanpassungsknoten ausgeschlossen. Dies ist besonders hilfreich, wenn Sie eine Verteilung manuell angeben und sicherstellen möchten, dass das Feld nicht von der automatischen Verteilungsanpassung betroffen ist, wenn ein verlinkter Simulationsanpassungsknoten ausgeführt wird.
- **Verteilung.** Die Zellen in dieser Spalte enthalten eine Dropdown-Liste der statistischen Verteilungen. Durch die Auswahl des Speichertyps wird festgelegt, welche Verteilungen in dieser Spalte für ein bestimmtes Feld verfügbar sind. Weitere Informationen finden Sie im Thema „Verteilungen” auf Seite 55.

  Anmerkung: Sie können die feste Verteilung nicht für jedes Feld angeben. Wenn jedes Feld in Ihren generierten Daten als fest definiert sein soll, können Sie den Benutzereingabeknoten gefolgt von einem Balancierungsknoten verwenden.
- **Parameter.** In dieser Spalte werden die Verteilungsparameter angezeigt, die den einzelnen angepassten Verteilungen zugeordnet sind. Mehrere Werte für einen Parameter werden durch Kommas getrennt. Wenn mehrere Werte für einen Parameter angegeben werden, werden mehrere Iterationen für die Simulation generiert. Weitere Informationen finden Sie im Thema „Iterationen” auf Seite 55. Wenn Parameter fehlen, spiegelt sich dies in dem Symbol wider, das in der Spalte "Status" angezeigt wird. Klicken Sie zur Angabe von Werten für die Parameter in der Zeile, die dem relevanten Feld entspricht, auf diese Spalte, und wählen Sie **Angeben** aus der Liste aus. Dadurch wird das Unterdialogfeld **Parameter angeben** geöffnet. Weitere Informationen finden Sie im Thema „Angeben von Parametern” auf Seite 53. Diese Spalte ist inaktiviert, wenn in der Spalte "Verteilung" "Empirisch" ausgewählt wird.
- **Min,Max.** In dieser Spalte können Sie für einige Verteilungen einen Minimalwert und/oder einen Maximalwert für die simulierten Daten angeben. Simulierte Daten, die kleiner als der Minimalwert und größer als der Maximalwert sind, werden zurückgewiesen, auch wenn sie für die angegebene Verteilung gültig wären. Klicken Sie zur Angabe von Minimal- und Maximalwerten in der Zeile, die dem relevanten Feld entspricht, auf diese Spalte, und wählen Sie **Angeben** aus der Liste aus. Dadurch wird das Unterdialogfeld **Parameter angeben** geöffnet. Weitere Informationen finden Sie im Thema „Angeben von Parametern” auf Seite 53. Diese Spalte ist inaktiviert, wenn in der Spalte "Verteilung" "Empirisch" ausgewählt wird.

Beste Anpassung verwenden. Ist nur aktiviert, wenn der Simulationsgenerierungsknoten automatisch über einen Simulationsanpassungsknoten anhand historischer Daten erstellt wurde und eine einzige Zeile in der Tabelle "Simulierte Felder" ausgewählt ist. Ersetzt die Informationen für das Feld in der ausgewählten Zeile durch die Informationen der Verteilung mit der besten Anpassung für das Feld. Wenn die Informationen in der ausgewählten Zeile bearbeitet wurden, werden die Informationen durch Klicken auf diese Schaltfläche auf die Verteilung mit der besten Anpassung zurückgesetzt, die über den Simulationsanpassungsknoten bestimmt wurde.

Details zur Anpassungsgüte. Ist nur aktiviert, wenn der Simulationsgenerierungsknoten automatisch über einen Simulationsanpassungsknoten erstellt wurde. Öffnet das Unterdialogfeld **Details zur Anpassungsgüte**. Weitere Informationen finden Sie im Thema „Details zur Anpassungsgüte” auf Seite 52.

Einige hilfreiche Aufgaben können mithilfe der Symbole auf der rechten Seite der Ansicht "Simulierte Felder" ausgeführt werden. Diese Symbole werden in der folgenden Tabelle beschrieben.

Tabelle 5. Symbole in der Ansicht "Simulierte Felder"

Symbol	QuickInfo	Beschreibung
	**Verteilungsparameter bearbeiten**	Nur aktiviert, wenn eine einzelne Zeile in der Tabelle "Simulierte Felder" ausgewählt ist. Öffnet das Unterdialogfeld **Parameter angeben** für die ausgewählte Zeile. Weitere Informationen finden Sie im Thema „Angeben von Parametern” auf Seite 53.

Tabelle 5. Symbole in der Ansicht "Simulierte Felder" (Forts.)

Symbol	QuickInfo	Beschreibung
	**Neues Feld hinzufügen**	Nur aktiviert, wenn eine einzelne Zeile in der Tabelle "Simulierte Felder" ausgewählt ist. Fügt unten in der Tabelle "Simulierte Felder" eine neue leere Zeile hinzu.
	**Mehrere Kopien erstellen**	Nur aktiviert, wenn eine einzelne Zeile in der Tabelle "Simulierte Felder" ausgewählt ist. Öffnet das Unterdialogfeld **Feld klonen**. Weitere Informationen finden Sie im Thema „Klonen eines Felds" auf Seite 51.
	**Ausgewähltes Feld löschen**	Löscht die ausgewählte Zeile aus der Tabelle "Simulierte Felder".
	**Ganz nach oben verschieben**	Ist nur aktiviert, wenn die ausgewählte Zeile nicht bereits die oberste Zeile der Tabelle "Simulierte Felder" ist. Verschiebt die ausgewählte Zeile in der Tabelle "Simulierte Felder" ganz nach oben. Diese Aktion wirkt sich auf die Reihenfolge der Felder in den simulierten Daten aus.
	**Nach oben**	Ist nur aktiviert, wenn die ausgewählte Zeile nicht die oberste Zeile der Tabelle "Simulierte Felder" ist. Verschiebt die ausgewählte Zeile in der Tabelle "Simulierte Felder" eine Position nach oben. Diese Aktion wirkt sich auf die Reihenfolge der Felder in den simulierten Daten aus.
	**Nach unten**	Ist nur aktiviert, wenn die ausgewählte Zeile nicht die unterste Zeile der Tabelle "Simulierte Felder" ist. Verschiebt die ausgewählte Zeile in der Tabelle "Simulierte Felder" eine Position nach unten. Diese Aktion wirkt sich auf die Reihenfolge der Felder in den simulierten Daten aus.
	**Ganz nach unten verschieben**	Ist nur aktiviert, wenn die ausgewählte Zeile nicht bereits die unterste Zeile der Tabelle "Simulierte Felder" ist. Verschiebt die ausgewählte Zeile in der Tabelle "Simulierte Felder" ganz nach unten. Diese Aktion wirkt sich auf die Reihenfolge der Felder in den simulierten Daten aus.

Bei erneuter Anpassung Min und Max nicht löschen. Ist diese Option ausgewählt, werden die Minimal- und Maximalwerte nicht überschrieben, wenn die Verteilungen durch die Ausführung eines verbundenen Simulationsanpassungsknotens aktualisiert werden.

Ansicht "Korrelationen"

Zwischen Eingabefeldern für Vorhersagemodelle liegen bekanntlich häufig Korrelationen vor, beispielsweise zwischen Größe und Gewicht. Korrelationen zwischen zu simulierenden Feldern müssen berücksichtigt werden, um sicherzustellen, dass diese Korrelationen in den simulierten Werten beibehalten werden.

Wenn der Simulationsgenerierungsknoten über einen Simulationsanpassungsknoten anhand historischer Daten generiert oder aktualisiert wurde, können Sie in der Ansicht "Korrelationen" die berechneten Korrelationen zwischen Feldpaaren anzeigen und bearbeiten. Wenn Sie nicht über historische Daten verfügen, können Sie die Korrelationen manuell anhand Ihres Wissens darüber, wie die Felder korreliert sind, angeben.

Anmerkung: Bevor Daten generiert werden, wird die Korrelationsmatrix automatisch überprüft, um festzustellen, ob sie positiv semidefinit ist und daher invertiert werden kann. Eine Matrix kann invertiert werden, wenn ihre Spalten linear unabhängig sind. Wenn die Korrelationsmatrix nicht invertiert werden kann, wird sie automatisch angepasst, um sie invertierbar zu machen.

Sie können die Korrelationen im Matrix- oder im Listenformat anzeigen.

Korrelationsmatrix. Zeigt die Korrelationen zwischen Feldpaaren in einer Matrix an. Die Feldnamen sind in alphabetischer Reihenfolge auf der linken Seite der Matrix und am oberen Rand der Matrix aufgelistet. Nur die Zellen unter der Diagonalen können bearbeitet werden. Es muss ein Wert zwischen -1,000 und 1,000 (inklusive) eingegeben werden. Die Zelle über der Diagonalen wird aktualisiert, wenn der Fokus von der gespiegelten Zelle unter der Diagonalen verschoben wird. Beide Zellen zeigen dann denselben Wert an. Die Zellen in der Diagonalen sind immer inaktiviert und verfügen immer über die Korrelation 1,000. Der Standardwert für alle anderen Zellen ist 0,000. Der Wert 0,000 gibt an, dass keine Korrelation zwischen dem zugeordneten Feldpaar vorhanden ist. In der Matrix sind nur stetige und ordinale Felder enthalten. Nominale und kategoriale Felder sowie Flagfelder und Felder, die der festen Verteilung zugewiesen sind, werden in der Tabelle nicht angezeigt.

Korrelationsliste. Zeigt die Korrelationen zwischen Feldpaaren in einer Tabelle an. Jede Zeile der Tabelle zeigt die Korrelation zwischen einem Feldpaar an. Es können keine Zeilen hinzugefügt oder gelöscht werden. Die Spalten mit den Überschriften "Feld 1" und "Feld 2" enthalten die Feldnamen, die nicht bearbeitet werden können. Die Spalte "Korrelation" enthält die Korrelationen, die bearbeitet werden können. Es muss ein Wert zwischen -1,000 und 1,000 (inklusive) eingegeben werden. Der Standardwert für alle Zellen ist 0,000. In der Liste sind nur stetige und ordinale Felder enthalten. Nominale und kategoriale Felder sowie Flagfelder und Felder, die der festen Verteilung zugewiesen sind, werden in der Liste nicht angezeigt.

Korrelationen zurücksetzen. Öffnet das Dialogfeld **Korrelationen zurücksetzen**. Wenn historische Daten verfügbar sind, können Sie eine der drei folgenden Optionen auswählen:

- **Angepasst.** Ersetzt die aktuellen Korrelationen durch die Korrelationen, die anhand der historischen Daten berechnet wurden.
- **Nullen.** Ersetzt die aktuellen Korrelationen durch Nullen.
- **Abbrechen.** Schließt das Dialogfeld. Die Korrelationen sind unverändert.

Wenn keine historischen Daten verfügbar sind, Sie jedoch Änderungen an den Korrelationen vorgenommen haben, können Sie auswählen, ob Sie die aktuellen Korrelationen durch Nullen ersetzen oder den Vorgang abbrechen möchten.

Anzeigen als. Wählen Sie **Tabelle** aus, um die Korrelationen als Matrix anzuzeigen. Wählen Sie **Liste** aus, um die Korrelationen als Liste anzuzeigen.

Korrelationen bei erneuter Anpassung nicht neu berechnen. Wählen Sie diese Option aus, wenn Sie Korrelationen manuell angeben und verhindern möchten, dass sie überschrieben werden, wenn Verteilungen mithilfe eines Simulationsanpassungsknotens und historischer Daten automatisch angepasst werden.

Angepasste Mehrwege-Kontingenztabelle für Eingaben mit einer kategorialen Verteilung verwenden. Standardmäßig sind alle Felder mit einer kategorialen Verteilung in einer Kontingenztabelle (oder einer Mehrwege-Kontingenztabelle, je nach der Anzahl der Felder mit einer kategorialen Verteilung) enthalten. Die Kontingenztabelle wird, ebenso wie die Korrelationen, bei der Ausführung eines Simulationsanpassungsknotens erstellt. Die Kontingenztabelle kannn nicht angezeigt werden. Wenn diese Option ausgewählt ist, werden Felder mit einer kategorialen Verteilung anhand der tatsächlichen Prozentsätze aus der Kontingenztabelle simuliert. Das heißt, alle Zuordnungen zwischen nominalen Feldern werden in den neuen, simulierten Daten erneut erstellt. Wenn diese Option inaktiviert ist, werden Felder mit kategorialen Verteilungen anhand der erwarteten Prozentsätze aus der Kontingenztabelle simuliert. Wenn Sie ein Feld ändern, wird das Feld aus der Kontingenztabelle entfernt.

Ansicht "Erweiterte Optionen"

Anzahl der zu simulierenden Fälle. Zeigt die Optionen für die Angabe der Anzahl der zu simulierenden Fälle und der Benennung von Iterationen an.

- **Maximale Anzahl an Fällen.** Gibt die maximale Anzahl an Fällen simulierter Daten und zugehöriger Zielwerte an, die generiert werden sollen.
- **Iterationen.** Diese Zahl wird automatisch berechnet und kann nicht bearbeitet werden. Eine Iteration wird automatisch jedes Mal erstellt, wenn für einen Verteilungsparameter mehrere Werte angegeben sind.
- **Zeilen insgesamt.** Ist nur aktiviert, wenn die Anzahl der Iterationen größer als 1 ist. Die Zahl wird automatisch anhand der angezeigten Gleichung berechnet und kann nicht bearbeitet werden.
- **Iterationsfeld erstellen.** Ist nur aktiviert, wenn die Anzahl der Iterationen größer als 1 ist. Wenn diese Option ausgewählt ist, ist das Feld **Name** aktiviert. Weitere Informationen finden Sie im Thema „Iterationen" auf Seite 55.
- **Name.** Ist nur aktiviert, wenn das Kontrollkästchen **Iterationsfeld erstellen** ausgewählt ist und die Anzahl der Iterationen größer als 1 ist. Bearbeiten Sie den Namen des Iterationsfelds durch eine Eingabe in dieses Textfeld. Weitere Informationen finden Sie im Thema „Iterationen" auf Seite 55.

Startwert für Zufallszahlen. Durch die Festlegung eines Startwerts für Zufallszahlen kann Ihre Simulation repliziert werden.

- **Ergebnisse replizieren.** Wenn diese Option ausgewählt ist, sind die Schaltfläche **Generieren** und das Feld **Startwert für Zufallszahlen** aktiviert.
- **Startwert für Zufallszahlen.** Nur aktiviert, wenn das Kontrollkästchen **Ergebnisse replizieren** ausgewählt ist. In diesem Feld können Sie eine ganze Zahl angeben, die als Startwert für Zufallszahlen verwendet werden soll. Der Standardwert ist 629111597.
- **Generieren.** Nur aktiviert, wenn das Kontrollkästchen **Ergebnisse replizieren** ausgewählt ist. Erstellt eine ganze Pseudozufallszahl zwischen 1 und 999999999 (inklusive) im Feld **Startwert für Zufallszahlen**.

Klonen eines Felds

Im Dialogfeld **Feld klonen** können Sie angeben, wie viele Kopien des ausgewählten Felds erstellt und wie die einzelnen Kopien benannt werden sollen. Es ist hilfreich, über mehrere Kopien von Feldern zu verfügen, wenn zusammengesetzte Effekte untersucht werden, z. B. Zins- oder Wachstumsraten über eine Reihe von aufeinanderfolgenden Zeiträumen.

Die Titelleiste des Dialogfelds enthält den Namen des ausgewählten Felds.

Anzahl der anzufertigenden Kopien. Enthält die Anzahl der Kopien des Felds, die erstellt werden sollen. Klicken Sie auf die Pfeile, um die Anzahl der Kopien anzugeben, die erstellt werden sollen. Die minimale Anzahl an Kopien ist 1 und das Maximum ist 512. Die Anzahl der Kopien ist anfänglich auf 10 gesetzt.

Kopiesuffixzeichen. Enthält die Zeichen, die am Ende des Feldnamens für jede Kopie hinzugefügt werden. Diese Zeichen trennen den Feldnamen von der Kopienummer. Die Suffixzeichen können durch eine Eingabe in diesem Feld bearbeitet werden. Dieses Feld kann leer gelassen werden. In diesem Fall stehen keine Zeichen zwischen dem Feldnamen und der Kopienummer. Das Standardzeichen ist ein Unterstrich.

Nummer der ersten Kopie. Enthält die Suffixnummer für die erste Kopie. Klicken Sie auf die Pfeile, um die Nummer der ersten Kopie auszuwählen. Das Minimum für die Nummer der ersten Kopie ist 1 und das Maximum ist 1000. Der Standardwert für die Nummer der ersten Kopie ist 1.

Schrittgröße der Kopienummern. Enthält das Inkrement für die Suffixnummern. Klicken Sie auf die Pfeile, um das Inkrement auszuwählen. Das minimale Inkrement ist 1 und das Maximum ist 255. Das Inkrement ist anfänglich auf 1 gesetzt.

Felder. Enthält eine Vorschau der Feldnamen für die Kopien. Sie wird aktualisiert, wenn eines der Felder des Dialogfelds **Feld klonen** bearbeitet wird. Dieser Text wird automatisch generiert und kann nicht bearbeitet werden.

OK. Generiert alle Kopien wie im Dialogfeld angegeben. Die Kopien werden der Tabelle "Simulierte Felder" im Dialogfeld des Simulationsgenerierungsknotens direkt unter der Zeile hinzugefügt, die das kopierte Feld enthält.

Abbrechen. Schließt das Dialogfeld. Alle vorgenommenen Änderungen werden verworfen.

Details zur Anpassungsgüte

Das Dialogfeld **Details zur Anpassungsgüte ist nur verfügbar, wenn der Simulationsgenerierungsknoten durch die Ausführung eines Simulationsanpassungsknotens erstellt oder aktualisiert wurde. Es zeigt die Ergebnisse der automatischem Verteilungsanpassung für das ausgewählte Feld an. Verteilungen werden nach Anpassungsgüte geordnet, wobei die am besten angepasste Verteilung zuerst aufgelistet wird. In diesem Dialogfeld können Sie die folgenden Tasks ausführen:**

- Die Verteilungen überprüfen, die an die historischen Daten angepasst sind.
- Eine der angepassten Verteilungen auswählen.

Feld. Enthält den Namen des ausgewählten Felds. Dieser Text kann nicht bearbeitet werden.

Behandeln als (Maß). Zeigt den Messtyp des ausgewählten Felds an. Dieser wird der Tabelle "Simulierte Felder" im Dialogfeld des Simulationsgenerierungsknotens entnommen. Der Messtyp kann durch Klicken auf den Pfeil und Auswählen eines Messtyps aus der Dropdown-Liste geändert werden. Es gibt drei Optionen: **Stetig**, **Nominal** und **Ordinal**.

Verteilungen. In der Tabelle "Verteilungen" werden alle Verteilungen angezeigt, die für den Messtyp geeignet sind. Die Verteilungen, die an die historischen Daten angepasst wurden, werden nach Anpassungsgüte geordnet, angefangen bei der am besten angepassten Verteilung bis zu der am schlechtesten angepassten Verteilung. Die Anpassungsgüte wird durch die Statistik zur Anpassungsgüte bestimmt, die im Simulationsanpassungsknoten ausgewählt wurde. Die Verteilungen, die nicht an die historischen Daten angepasst wurden, sind in der Tabelle unter den Verteilungen, die angepasst wurden, in alphabetische Reihenfolge aufgelistet.

Die Verteilungstabelle enthält die folgenden Spalten:

- **Verwenden.** Das ausgewählte Optionsfeld gibt an, welche Verteilung derzeit für das Feld ausgewählt ist. Sie können die am besten angepasste Verteilung außer Kraft setzen, indem Sie das Optionsfeld für die gewünschte Verteilung in der Spalte "Verwenden" auswählen. Wenn ein Optionsfeld in der Spalte

"Verwenden" ausgewählt wird, wird auch ein Diagramm der Verteilung angezeigt, das ein Histogramm (oder Balkendiagramm) der historischen Daten für das ausgewählte Feld überlagert. Es kann jeweils immer nur eine Verteilung ausgewählt werden.

- **Verteilung.** Enthält den Namen der Verteilung. Diese Spalte kann nicht bearbeitet werden.
- **Statistiken zur Anpassungsgüte.** Enthält die berechneten Statistiken zur Anpassungsgüte für die Verteilung. Diese Spalte kann nicht bearbeitet werden. Der Inhalt der Zelle hängt vom Messtyp des Felds ab:
  - **Stetig.** Enthält die Ergebnisse der Anderson-Darling- und Kolmogorov-Smirnoff-Tests. Die den Tests zugeordneten p-Werte werden ebenfalls angezeigt. Die Statistik zur Anpassungsgüte, die im Simulationsanpassungskonten als Kriterium für die Anpassungsgüte ausgewählt wurde, wird zuerst angezeigt und wird zum Ordnen der Verteilungen verwendet. Die Anderson-Darling-Statistiken werden als A=*a-Wert* P=*p-Wert* angezeigt. Die Kolmogorov-Smirnoff-Statistiken werden als K=*k-Wert* P=*p-Wert* angezeigt. Wenn eine Statistik nicht berechnet werden kann, wird anstatt einer Zahl ein Punkt angezeigt.
  - **Nominal/Ordinal.** Enthält die Ergebnisse des Chi-Quadrat-Tests. Der dem Test zugeordnete p-Wert wird ebenfalls angezeigt. Die Statistiken werden als Chi-Sq=*Wert* P=*p-Wert* angezeigt. Wenn die Verteilung nicht angepasst wurde, wird `Nicht angepasst` angezeigt. Wenn die Verteilung nicht mathematisch angepasst werden kann, wird `Kann nicht angepasst werden` angezeigt.

  Hinweis: Die Zelle ist für die empirische Verteilung immer leer.
- **Parameter.** Enthält die Verteilungsparameter, die jeder angepassten Verteilung zugeordnet sind. Die Parameter werden als *Parametername* = *Parameterwert* angezeigt, wobei die Parameter durch ein einzelnes Leerzeichen getrennt sind. Bei der kategorialen Verteilung sind die Parameternamen die Kategorien und die Parameterwerte sind die zugehörigen Wahrscheinlichkeiten. Wenn die Verteilung nicht an die historischen Daten angepasst wurde, ist die Zelle leer. Diese Spalte kann nicht bearbeitet werden.

Histogrammpiktogramm. Zeigt ein Diagramm der ausgewählten Verteilung an, das ein Histogramm der historischen Daten des ausgewählten Felds überlagert.

Verteilungspiktogramm. Zeigt eine Erläuterung und eine Abbildung der ausgewählten Verteilung an.

OK. Schließt das Dialogfeld und aktualisiert die Werte der Spalten "Messung", "Verteilung", "Parameter" und "Min,Max" der Tabelle "Simulierte Felder" für das ausgewählte Feld mit den Informationen aus der ausgewählten Verteilung. Das Symbol in der Spalte "Status" wird ebenfalls aktualisiert, damit widergespiegelt wird, ob die ausgewählte Verteilung die Verteilung mit der besten Anpassung an die Daten ist.

Abbrechen. Schließt das Dialogfeld. Alle vorgenommenen Änderungen werden verworfen.

Angeben von Parametern

Im Dialogfeld **Parameter angeben** können Sie die Parameterwerte für die Verteilung des ausgewählten Felds manuell angeben. Sie können auch eine andere Verteilung für das ausgewählte Feld auswählen.

Das Dialogfeld "Parameter angeben" kann auf drei Arten geöffnet werden:

- Doppelklicken Sie auf einen Feldnamen in der Tabelle "Simulierte Felder" im Dialogfeld des Simulationsgenerierungsknotens .
- Klicken Sie auf die Spalte "Parameter" oder "Min,Max" der Tabelle "Simulierte Felder" und wählen Sie **Angeben** aus der Liste aus.
- Wählen Sie in der Tabelle "Simulierte Felder" eine Zeile aus und klicken Sie anschließend auf das Symbol **Verteilungsparameter bearbeiten**.

Feld. Enthält den Namen des ausgewählten Felds. Dieser Text kann nicht bearbeitet werden.

Verteilung. Enthält die Verteilung des ausgewählten Felds. Diese wird der Tabelle "Simulierte Felder" entnommen. Die Verteilung kann durch Klicken auf den Pfeil und Auswählen einer Verteilung aus der Dropdown-Liste geändert werden. Die verfügbaren Verteilungen hängen vom Speichertyp des ausgewählten Felds ab.

Seiten. Diese Option ist nur verfügbar, wenn im Feld **Verteilung** die Dice-Verteilung ausgewählt ist. Klicken Sie auf die Pfeile, um die Anzahl der Seiten oder Kategorien anzugeben, in die das Feld aufgeteilt werden soll. Die minimale Anzahl an Standardabweichungen ist 2 und das Maximum ist 20. Die Anzahl der Seiten ist anfänglich auf 6 gesetzt.

Verteilungsparameter. Die Tabelle "Verteilungsparameter" enthält für jeden Parameter der ausgewählten Verteilung eine Zeile. Die Tabelle enthält zwei Spalten:

- **Parameter.** Enthält die Namen der Parameter. Diese Spalte kann nicht bearbeitet werden.
- **Wert(e).** Enthält die Werte der Parameter. Wenn der Simulationsgenerierungsknoten über einen Simulationsanpassungsknoten erstellt oder aktualisiert wurde, enthalten die Zellen in dieser Spalte die Parameterwerte, die durch die Anpassung der Verteilung an die historischen Daten bestimmt wurden. Wenn der Simulationsgenerierungsknoten über die Palette "Quellenknoten" dem Streamerstellungsbereich hinzugefügt wurde, sind die Zellen in dieser Spalte leer. Die Werte können durch eine Eingabe in die Zellen bearbeitet werden. Weitere Informationen zu den Parametern, die für die einzelnen Verteilungen erforderlich sind, und zu zulässigen Parameterwerten finden Sie im Thema „Verteilungen” auf Seite 55.

  Mehrere Werte für einen Parameter müssen durch Kommas getrennt werden. Wenn mehrere Werte für einen Parameter angegeben werden, werden mehrere Iterationen der Simulation definiert. Sie können nur für einen Parameter mehrere Werte angeben.

  Hinweis: Für Felder mit Datum/Uhrzeitspeichertypen müssen Sie die Verteilungsparameter als ganze Zahlen angeben. Wenn Sie beispielsweise den 1. Januar 1970 als mittleres Datum angeben wollen, verwenden Sie die ganze Zahl 0.

Hinweis: Wenn die Dice-Verteilung ausgewählt ist, sieht die Tabelle "Verteilungsparameter" etwas anders aus. Die Tabelle enthält für jede Seite (oder Kategorie) eine Zeile. Die Tabelle enthält die Spalte "Wert" und die Spalte "Wahrscheinlichkeit". Die Spalte "Wert" enthält für jede Kategorie eine Beschriftung. Die Standardwerte für die Beschriftungen sind die ganzen Zahlen 1 bis n, wobei n für die Anzahl der Seiten steht. Die Beschriftungen können durch eine Eingabe in die Zellen bearbeitet werden. In die Zellen kann ein beliebiger Wert eingegeben werden. Wenn Sie einen Wert verwenden möchten, der keine Zahl ist, muss der Speichertyp des Datenfelds in "Zeichenfolge" geändert werden, wenn der Speichertyp nicht bereits auf "Zeichenfolge" gesetzt ist. Die Spalte "Wahrscheinlichkeit" enthält die Wahrscheinlichkeit für die einzelnen Kategorien. Die Wahrscheinlichkeiten können nicht bearbeitet werden und werden als 1/n berechnet.

Vorschau. Zeigt ein Beispieldiagramm der Verteilung an, das auf den angegebenen Parametern basiert. Wenn für einen Parameter mindestens zwei Werte angegeben wurden, werden für jeden Wert des Parameters Beispieldiagramme angezeigt. Wenn für das ausgewählte Feld historische Daten verfügbar sind, überlagert das Diagramm der Verteilung ein Histogramm der historischen Daten.

Optionale Einstellungen. Verwenden Sie diese Optionen, um einen Minimalwert und/oder einen Maximalwert für die simulierten Daten anzugeben. Simulierte Daten, die kleiner als der Minimalwert und größer als der Maximalwert sind, werden zurückgewiesen, auch wenn sie für die angegebene Verteilung gültig wären.

- **Minimum angeben.** Wählen Sie diese Option aus, um das Feld **Zurückweisen von Werten unter** zu aktivieren. Das Kontrollkästchen wird inaktiviert, wenn die empirische Verteilung ausgewählt ist.
- **Zurückweisen von Werten unter.** Ist nur aktiviert, wenn **Minimum angeben** ausgewählt ist. Geben Sie einen Minimalwert für die simulierten Daten an. Alle simulierten Werte, die kleiner als dieser Wert sind, werden zurückgewiesen.

- **Maximum angeben.** Wählen Sie diese Option aus, um das Feld **Zurückweisen von Werten über** zu aktivieren. Das Kontrollkästchen wird inaktiviert, wenn die empirische Verteilung ausgewählt ist.
- **Zurückweisen von Werten über.** Ist nur aktiviert, wenn **Maximum angeben** ausgewählt ist. Geben Sie einen Maximalwert für die simulierten Daten an. Alle simulierten Werte, die größer als dieser Wert sind, werden zurückgewiesen.

OK. Schließt das Dialogfeld und aktualisiert die Werte der Spalten "Verteilung", "Parameter" und "Min-,Max" der Tabelle "Simulierte Felder" für das ausgewählte Feld. Das Symbol in der Spalte "Status" wird ebenfalls aktualisiert, damit die ausgewählte Verteilung widergespiegelt wird.

Abbrechen. Schließt das Dialogfeld. Alle vorgenommenen Änderungen werden verworfen.

Iterationen

Wenn Sie mehrere Werte für ein festes Feld oder einen Verteilungsparameter angegeben haben, wird für jeden angegebenen Wert ein unabhängiges Set simulierter Fälle - und somit eine separate Simulation - generiert. Damit können Sie die Auswirkung der Variation des Felds oder Parameters überprüfen. Jedes Set simulierter Fälle wird als *Iteration* bezeichnet. In den simulierten Daten werden die Iterationen gestapelt.

Wenn das Kontrollkästchen **Iterationsfeld erstellen** in der Ansicht "Erweiterte Optionen" des Dialogfelds des Simulationsgenerierungsknotens ausgewählt ist, wird den simulierten Daten ein Iterationsfeld als nominales Feld mit numerischem Speichertyp hinzugefügt. Der Name dieses Felds kann durch eine Eingabe in das Feld **Name** in der Ansicht "Erweiterte Optionen" bearbeitet werden. Dieses Feld enthält eine Beschriftung, die angibt, zu welcher Iteration die einzelnen simulierten Fälle gehören. Die Form der Beschriftung hängt vom Iterationstyp ab:

- **Iterieren eines festen Felds.** Die Beschriftung ist der Name des Felds, auf den ein Gleichheitszeichen folgt, auf das wiederum der Wert des Felds für diese Iteration folgt, d. h.

  Feldname = Feldwert
- **Iterieren eines Verteilungsparameters.** Die Beschriftung besteht aus dem Namen des Felds, gefolgt von einem Doppelpunkt, dem Namen des iterierten Parameters, einem Gleichheitszeichen und dem Wert des Parameters für diese Iteration, d. h.

  Feldname:*Parametername = Parameterwert*
- **Iterieren eines Verteilungsparameters für eine kategoriale Verteilung oder Bereichsverteilung.** Die Beschriftung besteht aus dem Namen des Felds, gefolgt von einem Doppelpunkt, der Angabe "Iteration" und der Iterationsnummer, d. h.

  Feldname: Iteration *Iterationsnummer*

Verteilungen

Sie können die Wahrscheinlichkeitsverteilung für ein beliebiges Feld manuell angeben, indem Sie das Dialogfeld **Parameter angeben** für dieses Feld öffnen, die gewünschte Verteilung aus der Liste **Verteilung** auswählen und die Verteilungsparameter in die Tabelle **Verteilungsparameter** eingeben. Im Folgenden finden Sie einige Hinweise zu bestimmten Verteilungen:

- **Kategorial.** Die kategoriale Verteilung beschreibt ein Eingabefeld mit einer festen Anzahl numerischer Werte, die als Kategorien bezeichnet werden. Jeder Kategorie ist eine bestimmte Wahrscheinlichkeit zugeordnet, sodass die Summe der Wahrscheinlichkeiten für alle Kategorien gleich 1 ist.

  Anmerkung: Wenn Sie Wahrscheinlichkeiten für die Kategorien angeben, deren Summe nicht 1 ist, erhalten Sie eine Warnung.
- **Negativ binomial - Fehler.** Beschreibt die Verteilung der Anzahl der Fehler in einer Folge von Versuchen, bevor eine angegebene Anzahl von Erfolgen beobachtet wird. Der Parameter *Schwellenwert* legt die angegebene Anzahl Erfolge fest und der Parameter *Wahrscheinlichkeit* gibt die Wahrscheinlichkeit des Erfolgs eines bestimmten Versuchs an.

- **Negativ binomial - Versuche.** Beschreibt die Verteilung der Anzahl der erforderlichen Versuche, bevor eine angegebene Anzahl von Erfolgen beobachtet wird. Der Parameter *Schwellenwert* legt die angegebene Anzahl Erfolge fest und der Parameter *Wahrscheinlichkeit* gibt die Wahrscheinlichkeit des Erfolgs eines bestimmten Versuchs an.
- **Bereich.** Diese Verteilung besteht aus einem Set von Intervallen, wobei den einzelnen Intervallen eine bestimmte Wahrscheinlichkeit zugewiesen wird, sodass die Summe der Wahrscheinlichkeiten für alle Intervalle gleich 1 ist. Werte innerhalb eines bestimmten Intervalls werden aus einer Gleichverteilung gezogen, die für dieses Intervall definiert ist. Intervalle werden durch die Eingabe eines Minimalwerts, eines Maximalwerts und einer zugeordneten Wahrscheinlichkeit angegeben.

  Beispiel: Sie vermuten, dass die Kosten für einen Rohstoff mit einer 40-prozentigen Wahrscheinlichkeit in den Bereich von 10 bis 15 Dollar pro Einheit fallen und mit einer 60-prozentigen Wahrscheinlichkeit in den Bereich von 15 bis 20 Dollar pro Einheit fallen. Sie würden die Kosten mit einer Bereichsverteilung modellieren, die aus den zwei Intervallen [10 - 15] und [15 - 20] besteht, wobei Sie die dem ersten Intervall zugeordnete Wahrscheinlichkeit auf 0.4 und die dem zweiten Intervall zugeordnete Wahrscheinlichkeit auf 0,6 setzen würden. Die Intervalle müssen nicht aneinander angrenzen und können sich sogar überschneiden. Sie könnten beispielsweise die Intervalle 10 bis 15 Dollar und 20 bis 25 Dollar oder 10 bis 15 Dollar und 13 bis 16 Dollar angegeben haben.
- **Weibull.** Der Parameter *Speicherort* ist ein optionaler Positionsparameter, der angibt, wo sich der Ursprung der Verteilung befindet.

Die folgende Tabelle zeigt die Verteilungen, die für die benutzerdefinierte Verteilungsanpassung verfügbar sind, und die zulässigen Werte für die Parameter. Einige dieser Verteilungen sind für die benutzerdefinierte Anpassung an bestimmte Speichertypen verfügbar, auch wenn sie nicht automatisch durch den Simulationsanpassungsknoten an diese Speichertypen angepasst werden.

Tabelle 6. Für benutzerdefinierte Anpassung unterstützter Verteilungen

Verteilung	Für benutzerdefinierte Anpassung unterstützter Speichertyp	Parameter	Parametergrenzwerte	Hinweise
Bernoulli	Ganze Zahl, reelle Zahl, Datum/Uhrzeit	Wahrscheinlichkeit	0 ≤ *Wahrscheinlichkeit* ≤ 1	
Beta	Ganze Zahl, reelle Zahl, Datum/Uhrzeit	Form 1 Form 2 Minimum Maximum	≥ 0 ≥ 0 < *Maximum* > *Minimum*	Minimum und Maximum sind optional.
Binomial	Ganze Zahl, reelle Zahl, Datum/Uhrzeit	Anzahl der Versuche (n) Wahrscheinlichkeit Minimum Maximum	> 0, ganze Zahl 0 ≤ *Wahrscheinlichkeit* ≤ 1 < *Maximum* > *Minimum*	Die Anzahl der Versuche muss eine ganze Zahl sein. Minimum und Maximum sind optional.
Kategorial	Ganze Zahl, reelle Zahl, Datum/Uhrzeit, Zeichenfolge	Kategoriename (oder Kategoriebeschriftung)	0 ≤ *Wert* ≤ 1	Der Wert ist die Wahrscheinlichkeit der Kategorie. Die Werte müssen die Summe 1 ergeben, andernfalls wird eine Warnung generiert.

Tabelle 6. Für benutzerdefinierte Anpassung unterstützter Verteilungen (Forts.)

Verteilung	Für benutzerdefinierte Anpassung unterstützter Speichertyp	Parameter	Parametergrenzwerte	Hinweise
Dice	Ganze Zahl, Zeichenfolge	Seiten	2 ≤ *Seiten* ≤ 20	Die Wahrscheinlichkeit jeder Kategorie (Seite) wird als 1/n berechnet, wobei n für die Anzahl der Seiten steht. Die Wahrscheinlichkeiten können nicht bearbeitet werden.
Empirisch	Ganze Zahl, reelle Zahl, Datum/Uhrzeit			Sie können die empirische Verteilung nicht bearbeiten oder als Typ auswählen. Die empirische Verteilung ist nur verfügbar, wenn historische Daten vorhanden sind.
Exponentiell	Ganze Zahl, reelle Zahl, Datum/Uhrzeit	Skala Minimum Maximum	> 0 < *Maximum* > *Minimum*	Minimum und Maximum sind optional.
Fest	Ganze Zahl, reelle Zahl, Datum/Uhrzeit, Zeichenfolge	Wert		Sie können die feste Verteilung nicht für jedes Feld angeben. Wenn jedes Feld in Ihren generierten Daten als fest definiert sein soll, können Sie den Benutzereingabeknoten gefolgt von einem Balancierungsknoten verwenden.
Gamma	Ganze Zahl, reelle Zahl, Datum/Uhrzeit	Form Skala Minimum Maximum	≥ 0 ≥ 0 < *Maximum* > *Minimum*	Minimum und Maximum sind optional.
Lognormal	Ganze Zahl, reelle Zahl, Datum/Uhrzeit	Form 1 Form 2 Minimum Maximum	≥ 0 ≥ 0 < *Maximum* > *Minimum*	Minimum und Maximum sind optional.
Negativ binomial - Fehler	Ganze Zahl, reelle Zahl, Datum/Uhrzeit	Schwellenwert Wahrscheinlichkeit Minimum Maximum	≥ 0 0 ≤ *Wahrscheinlichkeit* ≤ 1 < *Maximum* > *Minimum*	Minimum und Maximum sind optional.
Negativ binomial - Versuche	Ganze Zahl, reelle Zahl, Datum/Uhrzeit	Schwellenwert Wahrscheinlichkeit Minimum Maximum	≥ 0 0 ≤ *Wahrscheinlichkeit* ≤ 1 < *Maximum* > *Minimum*	Minimum und Maximum sind optional.

Tabelle 6. Für benutzerdefinierte Anpassung unterstützter Verteilungen (Forts.)

Verteilung	Für benutzerdefinierte Anpassung unterstützter Speichertyp	Parameter	Parametergrenzwerte	Hinweise
Normalverteilung	Ganze Zahl, reelle Zahl, Datum/Uhrzeit	Mittelwert Standardabweichung Minimum Maximum	≥ 0 > 0 < *Maximum* > *Minimum*	Minimum und Maximum sind optional.
Poisson	Ganze Zahl, reelle Zahl, Datum/Uhrzeit	Mittelwert Minimum Maximum	≥ 0 < *Maximum* > *Minimum*	Minimum und Maximum sind optional.
Bereich	Ganze Zahl, reelle Zahl, Datum/Uhrzeit	Anfang(X) Ende(X) Wahrscheinlichkeit(X)	0 ≤ *Wert* ≤ 1	X ist der Index der einzelnen Klassen. Die Summe der Wahrscheinlichkeitswerte muss 1 ergeben.
Triangular	Ganze Zahl, reelle Zahl, Datum/Uhrzeit	Modus Minimum Maximum	*Minimum* ≤ *Wert* ≤ *Maximum* < *Maximum* > *Minimum*	
Gleichverteilung	Ganze Zahl, reelle Zahl, Datum/Uhrzeit	Minimum Maximum	< *Maximum* > *Minimum*	
Weibull	Ganze Zahl, reelle Zahl, Datum/Uhrzeit	Rate Skala Speicherort Minimum Maximum	> 0 > 0 ≥ 0 < *Maximum* > *Minimum*	Speicherort, Minimum und Maximum sind optional.

Benutzereingabeknoten

Der Benutzereingabeknoten bietet eine einfache Möglichkeit, künstliche Daten zu erstellen. Dazu können entweder neue Daten ohne Vorlage erstellt oder vorhandene Daten geändert werden. Diese Funktion ist nützlich, wenn Sie z. B. ein Testdataset für die Modellierung erstellen möchten.

Erstellen von Daten ohne Vorlage

Der Benutzereingabeknoten ist von der Palette der Datenquellen aus verfügbar und kann direkt zum Streamerstellungsbereich hinzugefügt werden.

1. Klicken Sie auf die Registerkarte **Datenquellen** der Knotenpalette.
2. Fügen Sie den Benutzereingabeknoten durch Ziehen und Ablegen oder durch Doppelklicken zum Streamerstellungsbereich hinzu.
3. Öffnen Sie das Dialogfeld durch Doppelklicken und geben Sie Felder und Werte an.

Hinweis: In der Palette der Datenquellen ausgewählte Benutzereingabeknoten sind komplett leer und enthalten keine Felder oder Dateninformationen. So können Sie künstliche Daten vollkommen neu ohne Vorlage erstellen.

Erzeugen von Daten von einer vorhandenen Datenquelle aus

Einen Benutzereingabeknoten können Sie auch von jedem Nichtendknoten im Stream aus erzeugen:

1. Überlegen Sie sich, an welchem Punkt des Streams Sie einen Knoten ersetzen möchten.

2. Klicken Sie mit der rechten Maustaste auf den Knoten, der seine Daten in den Benutzereingabeknoten speist, und wählen Sie im Menü die Option **Benutzereingabeknoten generieren**.
3. Der Benutzereingabeknoten wird mit allen Prozessen weiter unten im Stream angezeigt und ersetzt den vorhandenen Knoten an dem ausgewählten Punkt Ihres Datenstreams. Nachdem der Knoten erzeugt wurde, übernimmt er die gesamte Datenstruktur und Feldtypinformationen (sofern verfügbar) von den Metadaten.

Hinweis: Wenn Daten nicht alle Knoten im Stream durchlaufen haben, sind die Knoten nicht vollständig instanziiert. Dies bedeutet, dass der Speichertyp und Datenwerte unter Umständen nicht verfügbar sind, wenn der Knoten durch einen Benutzereingabeknoten ersetzt wird.

Festlegen von Optionen für den Benutzereingabeknoten

Das Dialogfeld für einen Benutzereingabeknoten enthält mehrere Tools, mit denen Sie Werte eingeben und die Datenstruktur für künstliche Daten definieren können. Bei einem generierten Knoten enthält die Tabelle auf der Registerkarte "Daten" Feldnamen aus der ursprünglichen Datenquelle. Bei einem Knoten, der von der Palette der Datenquellen hinzugefügt wurde, ist die Tabelle leer. Anhand der Tabellenoptionen können Sie folgende Aufgaben durchführen:

- Neue Felder mit der Schaltfläche "Neues Feld hinzufügen" rechts neben der Tabelle hinzufügen
- Vorhandene Felder umbenennen
- Den Datenspeichertyp für jedes Feld festlegen
- Werte angeben
- Reihenfolge der Felder in der Anzeige ändern

Eingeben von Daten

Für jedes Feld können Sie Werte festlegen oder vom ursprünglichen Dataset aus über die Schaltfläche zur Wertauswahl rechts neben der Tabelle einfügen. Weitere Informationen zur Angabe von Werten finden Sie in den unten beschriebenen Regeln. Sie können das Feld auch leer lassen. Leere Felder werden mit dem systemdefinierten Nullwert (`$null$`) aufgefüllt.

Zeichenfolgewerte können Sie einfach durch Leerzeichen getrennt in die Wertspalte eingeben:

```
Fritz Tanja Martin
```

Zeichenfolgen, die Leerzeichen enthalten, können in doppelte Leerzeichen gesetzt werden:

```
"Willi Schmidt" "Fritz Martin" "Jochen Berger"
```

Bei numerischen Feldern können Sie mehrere Werte auf gleiche Weise eingeben (d. h. mit Leerzeichen):

```
10 12 14 16 18 20
```

Sie können jedoch diese Reihe von Werten auch angeben, indem Sie die Grenzwerte (10, 20) festlegen und die Schritte dazwischen (2). Bei dieser Methode lautet die Eingabe wie folgt:

```
10,20,2
```

Beide Methoden können auch miteinander kombiniert werden, indem die eine in die andere eingebettet wird. Beispiel:

```
1 5 7 10,20,2 21 23
```

Das Ergebnis dieser Eingabe sind folgende Werte:

```
1 5 7 10 12 14 16 18 20 21 23
```

Datums- und Zeitwerte können unter Verwendung des aktuellen Standardformats eingegeben werden, das im Dialogfeld "Streameigenschaften" ausgewählt wird. Beispiele:

```
11:04:00 11:05:00 11:06:00
2007-03-14 2007-03-15 2007-03-16
```

Bei Zeitmarkenwerten, die aus einer Datums- und einer Zeitkomponente bestehen, müssen doppelte Anführungszeichen verwendet werden:

```
"2007-03-14 11:04:00" "2007-03-14 11:05:00" "2007-03-14 11:06:00"
```

Weitere Details finden Sie weiter unten in den Kommentaren zum Datenspeichertyp.

Daten generieren. Ermöglicht die Angabe, wie die Datensätze generiert werden sollen, wenn Sie den Stream ausführen.

- **Alle Kombinationen**. Generiert Datensätze, die jede mögliche Kombination der Feldwerte enthalten, sodass jeder Feldwert in mehreren Datensätzen enthalten ist. Dadurch können zuweilen mehr Daten generiert werden als gewünscht. Daher wird nach diesem Knoten häufig ein Stichprobenknoten eingefügt.
- **In Reihenfolge**. Generiert Datensätze in der Reihenfolge, in der die Datenfeldwerte angegeben werden. Jeder Feldwert kommt in einem Datensatz jeweils nur einmal vor. Die Gesamtzahl der Datensätze entspricht der höchsten Anzahl von Werten für ein einzelnes Feld. Wenn die Felder weniger Werte enthalten als die größte Anzahl, werden nicht definierte Werte ($null$) eingefügt.

Beispiel anzeigen

Durch die folgenden Einträge werden beispielsweise die in den folgenden beiden Beispielen aufgeführten Datensätze generiert.

- **Alter.** `30,60,10`
- **BD.** `NIEDRIG`
- **Cholesterin.** `NORMAL HOCH`
- **Medikament.** (leer)

Tabelle 7. Feld "Daten generieren" auf "Alle Kombinationen" gesetzt.

Alter	BP	Cholesterol	Medikament
30	NIEDRIG	NORMAL	$null$
30	NIEDRIG	HOCH	$null$
40	NIEDRIG	NORMAL	$null$
40	NIEDRIG	HOCH	$null$
50	NIEDRIG	NORMAL	$null$
50	NIEDRIG	HOCH	$null$
60	NIEDRIG	NORMAL	$null$
60	NIEDRIG	HOCH	$null$

Tabelle 8. Feld "Daten generieren" auf "In Reihenfolge" gesetzt.

Alter	BP	Cholesterol	Medikament
30	NIEDRIG	NORMAL	$null$
40	$null$	HOCH	$null$
50	$null$	$null$	$null$
60	$null$	$null$	$null$

Datenspeichertyp

Der Speichertyp beschreibt die Art und Weise, wie Daten in einem Feld gespeichert werden. Beispiel: Ein Feld mit den Werten 1 und 0 speichert ganzzahlige Daten. Dies ist vom Messniveau zu unterscheiden, das die Verwendung der Daten beschreibt und sich nicht auf den Speichertyp auswirkt. Beispiel: Sie möchten das Messniveau für ein Feld ganzer Zahlen mit den Werten 1 und 0 auf *Flag* setzen. Das bedeutet normalerweise, dass 1=*True* und 0=*False* ist. Während der Speichertyp stets an der Quelle festgelegt werden muss, kann das Messniveau mithilfe eines Typknotens an jeder beliebigen Stelle im Stream geändert werden. Weitere Informationen finden Sie im Thema „Messniveaus" auf Seite 119.

Folgende Speichertypen sind verfügbar:

- **Zeichenfolge.** Wird für Felder verwendet, die nicht numerische Daten enthalten (auch als alphanumerische Daten bezeichnet). Eine Zeichenfolge kann jede beliebige Abfolge von Zeichen enthalten, beispielsweise *fred*, *Klasse 2* oder *1234*. Beachten Sie, dass die Zahlen in Zeichenfolgen nicht für Berechnungen verwendet werden können.
- **Ganze Zahl.** Ein Feld, bei dessen Werten es sich um ganze Zahlen handelt.
- **Reelle Zahl.** Bei den Werten handelt es sich um Zahlen, die Dezimalstellen enthalten können (nicht auf ganze Zahlen beschränkt). Das Anzeigeformat wird im Dialogfeld für die Streameigenschaften angegeben und kann für einzelne Felder in einem Typknoten überschrieben werden (Registerkarte "Format").
- **Datum.** In einem Standardformat, wie Jahr, Monat und Tag (z. B. `2007-09-26`), angegebene Datumswerte. Das jeweilige Format wird im Dialogfeld für die Streameigenschaften angegeben.
- **Uhrzeit.** Als Dauer gemessene Zeit. Beispielsweise kann ein Service-Call, der 1 Stunde, 26 Minuten und 38 Sekunden dauerte, als `01:26:38` angegeben werden, je nachdem, welches Zeitformat aktuell im Dialogfeld für die Streameigenschaften angegeben ist.
- **Zeitmarke.** Werte, die sowohl eine Datums- als auch eine Zeitkomponente enthalten, wie beispielsweise `2007-09-26 09:04:00`; auch hier wieder abhängig von den aktuellen Formaten für Datum und Zeit im Dialogfeld "Streameigenschaften". Beachten Sie, dass Zeitmarkenwerte gegebenenfalls in Anführungszeichen gesetzt werden müssen, um sicherzustellen, dass sie als Einzelwert interpretiert werden und nicht als gesonderte Datums- und Zeitwerte. (Dies gilt beispielsweise bei der Eingabe von Werten in einem Benutzereingabeknoten.)

Speichertypkonvertierung. Der Speichertyp für ein Feld kann mit verschiedenen Konvertierungsfunktionen, z. B. `to_string` und `to_integer`, in einem Füllerknoten geändert werden. Weitere Informationen finden Sie im Thema „Speichertypkonvertierung mithilfe des Füllerknotens" auf Seite 140. Beachten Sie, dass die Konvertierungsfunktionen (und alle anderen Funktionen, für die ein spezieller Eingabetyp, wie beispielsweise ein Wert für Datum oder Uhrzeit, erforderlich ist) von den aktuell im Dialogfeld "Streameigenschaften" angegebenen Formaten abhängen. Wenn Sie beispielsweise ein Zeichenfolgenfeld mit den Werten *Jan 2003*, *Feb 2003* (usw.) in einen Datumsspeicher konvertieren müssen, wählen Sie **MON JJJJ** als Standarddatumsformat für den Stream aus. Konvertierungsfunktionen sind auch im Ableitungsknoten zur temporären Konvertierung während einer Ableitungsberechnung verfügbar. Mit dem Ableitungsknoten können Sie auch andere Bearbeitungen vornehmen wie beispielsweise die Umcodierung von Zeichenfolgenfeldern mit kategorialen Werten. Weitere Informationen finden Sie im Thema „Umcodieren von Werten mit dem Ableitungsknoten" auf Seite 139.

Einlesen gemischter Daten. Beachten Sie, dass beim Einlesen von Feldern mit numerischem Speichertyp (ganze Zahl, reelle Zahl, Zeit, Zeitmarke oder Datum) alle nicht numerischen Werte auf null oder auf systemdefiniert fehlend gesetzt werden. Dies liegt daran, dass IBM SPSS Modeler im Gegensatz zu einigen anderen Anwendungen keine gemischten Speichertypen innerhalb eines Felds zulässt. Um dies zu vermeiden, sollten alle Felder mit gemischten Daten als Zeichenfolgen eingelesen werden, indem der Speichertyp im Quellenknoten oder in der externen Anwendung nach Bedarf geändert wird.

Hinweis: Generierte Benutzereingabeknoten enthalten unter Umständen bereits Speicherinformationen, die aus dem Quellenknoten gesammelt wurden, sofern der Knoten instanziiert wurde. Ein nicht instanziierter Knoten enthält keine Speichertyp- oder Verwendungstypinformationen.

Regeln für das Festlegen von Werten

Bei symbolischen Feldern sollten zwischen den Werten Leerzeichen stehen. Beispiel:

```
HOCH MITTEL NIEDRIG
```

Bei numerischen Feldern können Sie mehrere Werte auf gleiche Weise eingeben (d. h. mit Leerzeichen):

```
10 12 14 16 18 20
```

Sie können jedoch diese Reihe von Werten auch angeben, indem Sie die Grenzwerte (10, 20) festlegen und die Schritte dazwischen (2). Bei dieser Methode lautet die Eingabe wie folgt:

```
10,20,2
```

Beide Methoden können auch miteinander kombiniert werden, indem die eine in die andere eingebettet wird. Beispiel:

```
1 5 7 10,20,2 21 23
```

Das Ergebnis dieser Eingabe sind folgende Werte:

```
1 5 7 10 12 14 16 18 20 21 23
```

Allgemeine Registerkarten für Quellenknoten

Folgende Optionen können für alle Quellenknoten festgelegt werden, indem Sie auf die entsprechende Registerkarte klicken:

- **Registerkarte "Daten".** Dient zum Ändern des Standardspeichertyps.
- **Registerkarte "Filter".** Dient zum Entfernen oder Umbenennen von Datenfeldern. Die Registerkarte bietet dieselben Funktionen wie der Filterknoten. Weitere Informationen finden Sie im Thema „Festlegen der Filteroptionen" auf Seite 130.
- **Registerkarte "Typen".** Wird verwendet, um Messniveaus festzulegen. Die Registerkarte bietet dieselben Funktionen wie der Typknoten.
- **Registerkarte "Anmerkungen".** Wird für alle Knoten verwendet. Die Registerkarte bietet Optionen zum Umbenennen von Knoten, zum Anzeigen einer benutzerdefinierten QuickInfo und zum Speichern einer längeren Anmerkung.

Festlegen von Messniveaus im Quellenknoten

Die Feldeigenschaften können in einem Quellenknoten oder in einem separaten Typknoten angegeben werden. Die Funktionsweise ist bei beiden Knoten ähnlich. Folgende Eigenschaften stehen zur Verfügung:

- **Feld.** Doppelklicken Sie auf einen beliebigen Feldnamen, um Wert- und Feldbeschriftungen für Daten in IBM SPSS Modeler anzugeben. So können aus IBM SPSS Statistics beispielsweise importierte Feldmetadaten hier angezeigt oder geändert werden. Auf ähnliche Weise können Sie auch neue Beschriftungen für Felder und ihre Werte erstellen. Die Beschriftungen, die Sie hier angeben, werden überall in IBM SPSS Modeler angezeigt, je nach der von Ihnen im Dialogfeld "Streameigenschaften" getroffenen Auswahl.
- **Messung.** Dies ist das Messniveau, das zur Beschreibung der Eigenschaften von Daten in einem bestimmten Feld verwendet wird. Wenn alle Details eines Felds bekannt sind, wird es als **vollständig instanziiert** bezeichnet. Weitere Informationen finden Sie im Thema „Messniveaus" auf Seite 119.

  Hinweis: Das Messniveau eines Felds ist etwas anderes als sein Speichertyp, der angibt, ob die Daten als Zeichenfolge, ganze Zahl, reelle Zahl, Datum, Zeit oder Zeitmarke gespeichert werden sollen.

- **Werte.** In dieser Spalte können Sie Optionen zum Lesen von Datenwerten aus dem Dataset auswählen oder die Option **Angeben** verwenden, um Messniveaus und Werte in einem separaten Dialogfeld anzugeben. Sie können auch Felder übergeben, ohne ihre Werte zu lesen. Weitere Informationen finden Sie im Thema „Datenwerte" auf Seite 122.
- **Fehlend.** Wird verwendet, um anzugeben, wie fehlende Werte für das Feld behandelt werden. Weitere Informationen finden Sie im Thema „Definieren fehlender Werte" auf Seite 125.
- **Überprüfen.** In dieser Spalte können Sie Optionen festlegen, um sicherzustellen, dass die Feldwerte den angegebenen Werten oder Bereichen entsprechen. Weitere Informationen finden Sie im Thema „Überprüfen von Typenwerten" auf Seite 125.
- **Rolle.** Wird verwendet, um Modellierungsknoten mitzuteilen, ob es sich bei Feldern um **Eingabefelder** (Prädiktorfelder) oder **Zielfelder** (vorhergesagte Felder) für einen Maschinenlernprozess handelt. **Beides** und **Keine** sind auch verfügbare Rollen, zusammen mit **Partition**, das ein Feld bezeichnet, das für die Aufteilung von Datensätzen in separate Stichproben zu Training-, Test- und Validierungszwecken verwendet wird. Der Wert **Aufteilung** gibt an, dass für jeden möglichen Wert des Felds separate Modelle erstellt werden. Weitere Informationen finden Sie im Thema „Festlegen der Feldrolle" auf Seite 126.

Weitere Informationen finden Sie im Thema „Typknoten" auf Seite 118.

Zeitpunkt der Instanziierung am Quellenknoten

Es gibt zwei Möglichkeiten, Informationen über den Datenspeichertyp und die Werte Ihrer Felder abzurufen. Diese **Instanziierung** kann entweder am Quellenknoten erfolgen, wenn Sie Daten erstmals in IBM SPSS Modeler importieren, oder durch Einfügen eines Typknotens in den Datenstream.

Die Instanziierung am Quellenknoten ist in folgenden Fällen nützlich:

- Wenn das Dataset recht klein ist.
- Wenn Sie beabsichtigen, mit Expression Builder neue Felder abzuleiten (durch Instanziierung werden die Feldwerte von Expression Builder verfügbar gemacht).

Im Allgemeinen ist bei nicht allzu großen Datasets und wenn keine Felder später im Stream hinzugefügt werden sollen, eine Instanziierung am Quellenknoten die praktischste Methode.

Filtern von Feldern am Quellenknoten

Mit der Registerkarte "Filter" des Dialogfelds eines Quellenknotens können Sie Felder basierend auf Ihrer anfänglichen Untersuchung der Daten aus Vorgängen weiter unten im Stream ausschließen Diese Funktion ist nützlich, wenn z. B. doppelte Felder in den Daten vorhanden sind oder wenn Sie bereits ausreichend mit den Daten vertraut sind und irrelevante Felder ausschließen können. Alternativ können Sie weiter unten im Stream einen gesonderten Filterknoten einfügen. Die Funktionsweise ist in beiden Fällen ähnlich. Weitere Informationen finden Sie im Thema „Festlegen der Filteroptionen" auf Seite 130.

Kapitel 3. Datensatzoperationsknoten

Überblick über die Datensatzoperationen

Mit Datensatzoperationsknoten werden Änderungen an Daten auf der Datensatzebene vorgenommen. Diese Operationen sind wichtig während der **Datenverständnis**- und **Datenvorbereitungs**-Phase des Data Mining, da Sie damit die Daten für Ihre jeweiligen geschäftlichen Anforderungen zuschneiden können.

Sie könnten beispielsweise auf der Grundlage der Ergebnisse des Data Audit, das mit dem Data Audit-Knoten (Ausgabepalette) durchgeführt wurde, zu dem Schluss kommen, dass die Datensätze über die Einkäufe der Kunden für die letzten drei Monate zusammengeführt werden sollten. Mit einem Zusammenführungsknoten ("Mergen") können Sie Datensätze auf der Grundlage der Werte eines Schlüsselfelds, beispielsweise *Kunden-ID*, zusammenführen. Oder Sie könnten feststellen, dass eine Datenbank mit Informationen über Website-Aufrufe unüberschaubar ist, da sie mehr als eine Million Datensätze enthält. Mithilfe von Beispielknoten können Sie ein Subset von Daten für die Modellierung auswählen.

Die Palette "Datensatzoperationen" enthält folgende Knoten:

Der Auswahlknoten wählt auf der Grundlage einer bestimmten Bedingung ein Subset von Datensätzen aus einem Datenstream aus oder verwirft sie. Sie können beispielsweise die Datensätze auswählen, die zu einer bestimmten Verkaufsregion gehören.

Der Stichprobenknoten wählte ein Subset der Datensätze aus. Es wird eine Vielzahl von Stichprobentypen unterstützt, darunter geschichtete, gruppierte (Clusterstichproben) und nichtzufällige (strukturierte) Stichproben. Eine Stichprobenziehung kann nützlich zur Verbesserung der Leistungsfähigkeit und zur Auswahl von verwandten Datensätzen bzw. Transaktionen für die Analyse sein.

Der Balancierungsknoten korrigiert Unausgewogenheiten in einem Dataset, sodass dieses eine bestimmte Bedingung erfüllt. Die Balancierungsanweisung passt den Anteil der Datensätze, bei denen eine Bedingung wahr ist, um den angegebenen Faktor an.

Der Aggregatknoten ersetzt eine Sequenz von Eingabedatensätzen durch zusammengefasste, aggregierte Ausgabedatensätze.

Mit dem RFM-Aggregatknoten (Recency-, Frequency-, Monetary-Aggregat) können Sie Daten über die früheren Transaktionen von Kunden verwenden, alle nicht benötigten Daten entfernen und alle verbliebenen Transaktionsdaten zu einer einzigen Zeile zusammenfassen, die angibt, wann der betreffende Kunde zuletzt mit Ihnen in Geschäftskontakt stand, wie viele Transaktionen er vorgenommen hat und wie hoch der Gesamtwert dieser Transaktionen ist.

Der Sortierknoten sortiert Datensätze anhand der Werte eines oder mehrerer Felder in aufsteigender oder absteigender Reihenfolge.

Der Zusammenführungsknoten erstellt aus mehreren Eingabedatensätzen einen einzelnen Ausgabedatensatz mit einigen oder allen der Eingabefelder. Er wird zum Zusammenführen von Daten aus verschiedenen Quellen verwendet, beispielsweise Daten über Auslandskunden und erworbene demografische Daten.

Der Anhangknoten verkettet Gruppen von Datensätzen miteinander. Er ist insbesondere nützlich für die Kombination von Datasets mit ähnlicher Struktur, aber unterschiedlichen Daten.

Der Duplikatknoten entfernt doppelte Datensätze, entweder indem jeweils der erste Datensatz an den Datenstream übergeben wird oder aber indem der erste Datensatz verworfen wird und stattdessen etwaige Duplikate an den Stream übergeben werden.

Der Streaming-ZR-Knoten erstellt und scort Zeitreihenmodelle in einem Schritt, ohne dass ein Zeitintervallknoten benötigt wird.

Für viele der Knoten in der Palette "Datensatzoperationen" ist die Verwendung eines CLEM-Ausdrucks erforderlich. Wenn Sie mit CLEM vertraut sind, können Sie einen Ausdruck in das Feld eingeben. Alle Ausdrucksfelder enthalten jedoch eine Schaltfläche zum Öffnen des CLEM Expression Builder, mit dem solche Ausdrücke automatisch erstellt werden.

Abbildung 1. Schaltfläche für Expression Builder

Auswahlknoten

Mit Auswahlknoten können Sie ein Subset von Datensätzen aus dem Stream auswählen bzw. verwerfen. Dafür werden spezielle Bedingungen verwendet, beispielsweise `BD (Blutdruck) = "HOCH"`.

Modalwert. Gibt an, ob Datensätze, die die Bedingung erfüllen, in den Datenstream eingeschlossen oder daraus ausgeschlossen werden.

- **Einschließen.** Wählen Sie diese Option aus, um Datensätze einzuschließen, die die Auswahlbedingung erfüllen.
- **Verwerfen.** Wählen Sie diese Option aus, um Datensätze auszuschließen, die die Auswahlbedingung erfüllen.

Bedingung. Zeigt die Auswahlbedingung an, die zum Testen der einzelnen Datensätze verwendet wird, die Sie mithilfe eines CLEM-Ausdrucks angeben. Geben Sie entweder einen Ausdruck in das Fenster ein oder verwenden Sie den Expression Builder, den Sie mit der Taschenrechnerschaltfläche rechts neben dem Fenster aufrufen können.

Sie können Datensätze auf der Basis einer Bedingung wie der folgenden verwerfen:
`(Variable1='Wert1' and Variable2='Wert2')`

Der Auswahlknoten verwirft in diesem Fall standardmäßig auch Datensätze, die Nullwerte für alle Auswahlfelder enthalten. Um dies zu vermeiden, hängen Sie die folgende Bedingung an die Originalbedingung an:

```
and not(@NULL(Variable1) and @NULL(Variable2))
```

Auswahlknoten werden auch zur Auswahl eines Anteils der Datensätze verwendet. Normalerweise wird für diesen Vorgang ein anderer Knoten, der Stichprobenknoten, verwendet. Wenn die Bedingung, die Sie angeben möchten, jedoch komplexer ist als die zur Verfügung stehenden Parameter, können Sie mithilfe des Auswahlknotens Ihre eigene Bedingung erstellen. Sie können beispielsweise Bedingungen der folgenden Art erstellen:

```
BD = "HOCH" and random(10) <= 4
```

Dadurch werden ungefähr 40 % der Datensätze mit hohem Blutdruck ausgewählt und zur weiteren Analyse im Stream weitergegeben.

Stichprobenknoten

Mithilfe von Stichprobenknoten können Sie ein Subset der Datensätze für die Analyse auswählen oder einen Anteil von Datensätzen auswählen, der verworfen werden soll. Es wird eine Vielzahl von Stichprobentypen unterstützt, darunter geschichtete, gruppierte (Clusterstichproben) und nichtzufällige (strukturierte) Stichproben. Stichprobenziehungen können aus verschiedenen Gründen durchgeführt werden:

- Zur Verbesserung der Leistung durch Schätzung von Modellen anhand eines Subsets der Daten. Modelle, die aus einer Stichprobe geschätzt wurden, sind häufig ebenso genau wie Modelle, die aus dem vollständigen Datensatz abgeleitet werden. Das gilt insbesondere, wenn sie durch die verbesserte Leistungsfähigkeit in der Lage sind, mit unterschiedlichen Methoden zu experimentieren, die Sie andernfalls nicht ausprobiert hätten.
- Zur Auswahl von Gruppen verwandter Datensätze oder Transaktionen für die Analyse, beispielsweise alle Artikel in einem Online-Warenkorb oder alle Eigenschaften in einem bestimmten Umfeld.
- Zur Ermittlung von Einheiten oder Fällen zur zufälligen Untersuchung im Rahmen von Qualitätssicherung, Betrugsprävention oder Sicherheitsmaßnahmen.

Hinweis: Wenn Sie die Daten einfach nur zum Zwecke der Validierung in eine Trainings- und eine Teststichprobe unterteilen möchten, kann stattdessen ein Partitionsknoten verwendet werden. Weitere Informationen finden Sie im Thema „Partitionsknoten” auf Seite 154.

Typen von Stichproben

Clusterstichproben. Hierbei werden Gruppen bzw. Cluster als Stichprobe gezogen, nicht einzelne Einheiten. Nehmen Sie beispielsweise an, Sie haben eine Datendatei mit einem Datensatz pro Schüler. Wenn Sie nach Schule gruppieren und der Stichprobenumfang 50 % beträgt, werden 50 % der Schulen ausgewählt und aus jeder ausgewählten Schule werden alle Schüler ausgewählt. Die Schüler in den nicht ausgewählten Schulen werden verworfen. Durchschnittlich wäre zu erwarten, dass ungefähr 50 % der Schüler ausgewählt werden, da jedoch die Schulen unterschiedlich groß sind, wird dieser Prozentsatz vermutlich nicht genau erreicht. Auf ähnliche Weise können Sie Artikel in einem Warenkorb nach Transaktions-ID zu Clustern zusammenfassen, um sicherzustellen, dass alle Artikel aus ausgewählten Transaktionen verwendet werden Ein Beispiel, in dem Immobilien nach Gemeinde zu Clustern gruppiert werden, finden Sie im Beispielstream *complexsample_property.str*.

Geschichtete Stichproben. Hierbei werden die Stichproben unabhängig innerhalb von sich nicht überschneidenden Untergruppen der Grundgesamtheit, den sogenannten Schichten, ausgewählt. So können Sie beispielsweise sicherstellen, dass Männer und Frauen zu gleichen Anteilen ausgewählt werden oder dass jede Region oder sozioökonomische Gruppe innerhalb der Einwohner einer Stadt repräsentiert wird. Außerdem können Sie für jede Schicht einen anderen Stichprobenumfang angeben (z. B. wenn Sie annehmen, dass eine Gruppe in den ursprünglichen Daten unterrepräsentiert ist). Ein Beispiel, in dem Immobilien nach Bezirk geschichtet werden, finden Sie im Beispielstream *complexsample_property.str*.

Systematische Stichprobenziehung (Stichprobenziehung vom Typ "1 in n"). Wenn eine zufällige Auswahl schwer zu erzielen ist, können die Stichprobeneinheiten systematisch (in festgelegten Intervallen) oder sequenziell gezogen werden.

Stichprobengewichtungen. Stichprobengewichtungen werden beim Ziehen einer komplexen Stichprobe automatisch berechnet und entsprechen ungefähr der "Häufigkeit" der einzelnen gezogenen Einheiten in den ursprünglichen Daten. Daher sollte die Summe der Gewichtungen in der gesamten Stichprobe eine Schätzung des Umfangs der ursprünglichen Daten darstellen.

Stichprobenrahmen

Ein Stichprobenrahmen definiert die potenzielle Quelle der in eine Stichprobe oder Studie aufzunehmenden Fälle. In einigen Fällen kann es möglich sein, jedes einzelne Mitglied einer Grundgesamtheit zu ermitteln und jedes beliebige davon in eine Stichprobe aufzunehmen. Dies ist beispielsweise bei der Stichprobenziehung aus Artikeln von einem Fließband der Fall. In den meisten Fällen besteht jedoch nicht auf jeden möglichen Fall Zugriff. So können Sie beispielsweise nicht sicher sein, welche Personen bei einer Wahl abstimmen wird, bis die Wahl stattgefunden hat. In diesem Fall können Sie in den USA beispielsweise das Wählerregister als Stichprobenrahmen verwenden, auch wenn einige registrierte Personen nicht abstimmen werden und wenn einige Personen möglicherweise abstimmen, obwohl sie zu dem Zeitpunkt, als Sie Einsicht in das Register nahmen, noch nicht aufgeführt waren. Personen, die sich nicht im Stichprobenrahmen befinden, können auch nicht in die Stichprobe aufgenommen werden. Ob Ihr Stichprobenrahmen hinsichtlich seiner Natur hinreichend große Ähnlichkeit mit der Grundgesamtheit aufweist, die Sie evaluieren möchten, ist eine Frage, die für jeden realen Fall gesondert zu untersuchen ist.

Optionen für Stichprobenknoten

Sie können, je nach Anforderung, die Methode **Einfach** oder **Komplex** auswählen.

Einfache Stichproben - Optionen

Mit der Methode "Einfach" können Sie einen Zufallsprozentsatz von Datensätzen, zusammenhängende Datensätze oder einfach jeden *n-ten* Datensatz auswählen.

Modalwert. Wählen Sie aus, ob Datensätze für die folgenden Modi übergeben (eingeschlossen) oder verworfen (ausgeschlossen) werden sollen:

- **Stichprobe einschließen.** Nimmt die ausgewählten Datensätze in den Datenstream auf und verwirft alle anderen. Beispiel: Wenn Sie den Modus auf **Stichprobe einschließen** und die Option **1 in n** auf "5" setzen, wird jeder 5. Datensatz in den Datenstream aufgenommen und es ergibt sich ein Dataset mit ungefähr einem Fünftel der ursprünglichen Größe. Dies ist der Standardmodus bei der Stichprobenziehung von Daten und der einzige Modus, der bei der Methode "Komplex" zur Verfügung steht.
- **Stichprobe verwerfen.** Schließt die ausgewählten Datensätze aus und nimmt alle anderen auf. Beispiel: Wenn Sie den Modus auf **Stichprobe verwerfen** und die Option **1 in n** auf "5" setzen, wird jeder 5. Datensatz verworfen. Dieser Modus ist nur bei der Methode "Einfach" verfügbar.

Beispiel. Wählen Sie die Methode der Stichprobenziehung aus den folgenden Optionen aus:

- **Erste.** Verwenden Sie diese Option, um eine Stichprobenziehung mit zusammenhängenden Daten durchzuführen. Beispiel: Wenn die maximale Stichprobengröße auf "1000" gesetzt ist, werden die ersten 10.000 Datensätze ausgewählt.
- **1 in n.** Wählen Sie diese Option aus, um Stichproben zu ziehen, indem jeder *n*-te Datensatz übergeben bzw. verworfen wird. Wenn z. B. "*n*" auf "5" gesetzt ist, wird jeder 5. Datensatz ausgewählt.
- **Zufällig %.** Wählen Sie diese Option aus, um per Zufallsgenerator einen festgelegten Prozentsatz der Daten als Stichprobe zu ziehen. Beispiel: Wenn der Prozentsatz auf "20" gesetzt wird, werden 20 % der Daten entweder an den Datenstream übergeben oder verworfen, je nach dem ausgewählten Modus. Geben Sie mithilfe des Felds einen Prozentsatz für die Stichprobenziehung an. Mit dem Steuerelement **Startwert für Zufallsgenerator festlegen** können Sie außerdem einen Startwert bestimmen.

Sampling auf Blockebene verwenden (nur datenbankintern). Diese Option ist nur aktiviert, wenn Sie beim Durchführen von datenbankinternem Mining in einer Oracle- oder IBM DB2-Datenbank einen Zufallsprozentsatz für die Stichprobenbeziehung auswählen. In diesem Fall kann es effizienter sein, Sampling auf Blockebene zu verwenden.

Maximale Stichprobengröße. Gibt an, wie viele Datensätze maximal in die Stichprobe aufgenommen werden sollen. Diese Option ist redundant und wird daher inaktiviert, wenn **Erste** und **Einschließen** ausgewählt wurden. Beachten Sie auch, dass diese Einstellung bei Verwendung in Kombination mit der Option **Zufällig %** dazu führen kann, dass bestimmte Datensätze nicht ausgewählt werden. Wenn Ihr Dataset beispielsweise 10 Millionen Datensätze enthält und Sie bei einem maximalen Stichprobenumfang von 3 Millionen 50 % der Datensätze auswählen, führt das dazu, dass nur die ersten 6 Millionen Datensätze jeweils mit 50%iger Wahrscheinlichkeit ausgewählt werden und die restlichen vier Millionen Datensätze keine Chance haben, in die Stichprobe aufgenommen zu werden. Um diese Einschränkung zu vermeiden, müssen Sie die Methode **Komplex** für die Stichprobenziehung auswählen und eine Standardstichprobe von 3 Millionen Datensätzen anfordern, ohne eine Cluster- oder Schichtungsvariable anzugeben.

Komplexe Stichproben - Optionen

Die Optionen für komplexe Stichproben gestatten eine feinere Steuerung der Stichprobe. So können beispielsweise neben anderen Optionen gruppierte (Clusterstichproben), geschichtete und gewichtete Stichproben festgelegt werden.

Cluster und Schichtung. Mit dieser Option können Sie bei Bedarf Felder für Cluster, Schichtung und Eingabegewichtung angeben. Weitere Informationen finden Sie im Thema „Einstellungen unter "Cluster und Schichtung"" auf Seite 70.

Stichprobentyp.

- **Zufällig.** Wählt Cluster oder Datensätze innerhalb der einzelnen Schichten nach dem Zufallsprinzip aus.
- **Systematisch.** Wählt Datensätze in festen Intervallen aus. Diese Option hat dieselbe Wirkung wie die Methode *1 in n*, mit der Ausnahme, dass sich die Position des ersten Datensatzes in Abhängigkeit von einem Zufallsstartwert ändert. Der Wert von *n* wird automatisch auf der Grundlage der Stichprobengröße bzw. des Anteils ermittelt.

Stichprobeneinheiten. Sie können Anteile oder Anzahl (beobachtete Werte) als Grundeinheiten für die Stichprobe auswählen.

Stichprobenumfang. Es gibt mehrere Möglichkeiten zur Festlegung des Stichprobenumfangs:

- **Fest.** Hiermit können Sie den Gesamtumfang der Stichprobe als Anzahl (beobachtete Werte) oder Anteil angeben.
- **Benutzerdefiniert.** Ermöglicht die Angabe des Stichprobenumfangs für die einzelnen Untergruppen oder Schichten. Diese Option ist nur verfügbar, wenn im Unterdialogfeld "Cluster und Schichtung" ein Schichtungsfeld angegeben wurde.
- **Variable.** Ermöglicht dem Benutzer die Auswahl eines Felds, mit dem der Stichprobenumfang für die einzelnen Untergruppen oder Schichten definiert werden kann. Dieses Feld sollte für jeden Datensatz innerhalb einer bestimmten Schicht denselben Wert aufweisen; wenn die Stichprobe beispielsweise nach Bezirk geschichtet wird, müssen alle Datensätze mit *county = Surrey* denselben Wert aufweisen. Das Feld muss numerisch sein und seine Werte müssen mit den ausgewählten Stichprobeneinheiten übereinstimmen. Bei Anteilen sollten die Werte größer als 0 und kleiner als 1 sein; bei den beobachteten Werten ist der Mindestwert 1.

Minimale Stichprobe pro Schicht. Gibt die Mindestanzahl an Datensätzen (bzw. die Mindestanzahl an Clustern, wenn ein Clusterfeld angegeben wurde) an.

Maximale Stichprobe pro Schicht. Legt die maximale Anzahl an Datensätzen oder Clustern fest. Wenn Sie diese Option auswählen, ohne einen Cluster oder ein Schichtungsfeld anzugeben, wird eine Zufallsstichprobe oder systematische Stichprobe mit der angegebenen Größe ausgewählt.

Startwert für Zufallsgenerator festlegen. Bei der Stichprobenziehung oder Partitionierung von Datensätzen auf der Grundlage eines Zufallsprozentsatzes können Sie mit dieser Option dieselben Ergebnisse in einer anderen Sitzung replizieren. Wenn Sie den vom Zufallszahlengenerator verwendeten Startwert angeben, stellen Sie sicher, dass bei jeder Ausführung des Knotens dieselben Datensätze zugewiesen werden. Geben Sie den gewünschten Startwert ein oder klicken Sie auf die Schaltfläche **Generieren**, um automatisch einen Startwert zu generieren. Wenn diese Option nicht ausgewählt ist, wird bei jeder Ausführung des Knotens eine andere Stichprobe generiert.

Hinweis: Bei Verwendung der Option **Startwert für Zufallsgenerator festlegen** mit Datensätzen, die aus einer Datenbank eingelesen wurden, ist möglicherweise vor der Stichprobenziehung ein Sortierknoten erforderlich, um zu gewährleisten, dass bei jeder Ausführung des Knotens dasselbe Ergebnis erzielt wird. Dies liegt daran, dass der Startwert für den Zufallsgenerator von der Reihenfolge der Datensätze abhängt, die in relationalen Datenbanken nicht unbedingt gleich bleibt. Weitere Informationen finden Sie im Thema „Sortierknoten” auf Seite 78.

Einstellungen unter "Cluster und Schichtung"

Im Dialogfeld "Cluster und Schichtung" können Sie beim Ziehen einer komplexen Stichprobe Felder für Cluster, Schichtung und Gewichtung auswählen.

Cluster. Gibt ein kategoriales Feld an, das für die Gruppierung von Datensätzen verwendet wird. Datensätze werden anhand Ihrer Zugehörigkeit zu bestimmten Clustern bei der Stichprobenziehung berücksichtigt. Dabei werden bestimmte Cluster aufgenommen und andere nicht. Wenn jedoch ein Datensatz aus einem bestimmten Cluster aufgenommen wird, werden auch alle anderen aufgenommen. Bei der Analyse von Verbindungen zwischen Produkten in Warenkörben könnten Sie beispielsweise die Artikel nach Transaktions-ID gruppieren, um sicherzustellen, dass alle Artikel aus den ausgewählten Transaktionen verwendet werden. Anstatt bei der Stichprobenziehung einzelne Datensätze auszuwählen, wodurch Informationen darüber, welche Artikel gemeinsam verkauft wurden, verloren gehen würden, können Sie bei der Stichprobenziehung Transaktionen auswählen, um sicherzugehen, dass alle Datensätze der ausgewählten Transaktionen erhalten bleiben.

Schichten nach. Dient zur Angabe eines kategorialen Felds, mit dem Datensätze geschichtet werden können, sodass die Stichproben unabhängig innerhalb von sich nicht überschneidenden Untergruppen der Grundgesamtheit, den sogenannten Schichten, ausgewählt werden. Wenn Sie beispielsweise eine Stichprobe mit dem Umfang 50 % auswählen, die nach Geschlecht geschichtet ist, werden zwei 50-Prozent-Stichproben gezogen, eine für die Männer und eine für die Frauen. Weitere Beispiele für Schichten sind sozioökonomische Gruppen, Berufskategorien, Altersgruppen oder ethnische Gruppen. Mithilfe von Schichten können Sie angemessene Stichprobengrößen für relevante Untergruppen gewährleisten. Wenn im ursprünglichen Dataset dreimal mehr Frauen als Männer enthalten sind, wird dieses Verhältnis durch die separate Stichprobenziehung aus jeder der Gruppen beibehalten. Es können auch mehrere Schichtungsfelder angegeben werden (beispielsweise die Stichprobenziehung von Produktlinien innerhalb von Regionen oder umgekehrt).

Hinweis: Wenn Sie die Schichtung anhand eines Felds mit fehlenden Werten (null oder systemdefiniert fehlende Werte, leere Zeichenfolgen, leere Bereiche und Leerstellen oder benutzerdefiniert fehlende Werte) vornehmen, können Sie keine benutzerdefinierten Stichprobengrößen für die Schichten angeben. Wenn Sie bei der Schichtung nach einem Feld mit fehlenden Werten oder Leerwerten benutzerdefinierte Stichprobengrößen verwenden möchten, müssen Sie die fehlenden Werte weiter oben im Stream ergänzen.

Eingabegewichtung verwenden. Dient zur Angabe eines Felds, das zur Gewichtung von Datensätzen vor der Stichprobenziehung verwendet werden soll. Wenn beispielsweise das Gewichtungsfeld Werte im Bereich von 1 bis 5 aufweist, werden die Datensätze mit der Gewichtung 5 mit der 5-fachen Wahrscheinlich-

keit ausgewählt. Die Werte in diesem Feld werden mit den endgültigen Ausgabegewichtungen überschrieben, die vom Knoten generiert wurden (siehe folgender Absatz)

Neue Ausgabegewichtung. Gibt den Namen des Felds an, in das die endgültigen Gewichtungen geschrieben werden, wenn kein Feld für die Eingabegewichtung angegeben wurde. (Wenn ein Feld für die Eingabegewichtung angegeben wurde, werden seine Werte, wie oben angegeben, durch die endgültigen Gewichtungen ersetzt und es wird kein separates Feld für die Ausgabegewichtungen erstellt.) Die Werte für die Ausgabegewichtungen geben die Anzahl der Datensätze an, die durch die einzelnen Stichprobendatensätze in den ursprünglichen Daten repräsentiert werden. Die Summe der Gewichtungswerte ergibt eine Schätzung des Stichprobenumfangs. Wenn beispielsweise eine 10 % umfassende Zufallsstichprobe gezogen wurde, ist die Ausgabegewichtung für alle Datensätze 10, was anzeigt, dass jeder Datensatz in der Stichprobe ungefähr 10 Datensätze in den ursprünglichen Daten repräsentiert. Bei einer geschichteten oder gewichteten Stichprobe können die Werte der Ausgabegewichtungen je nach dem Stichprobenanteil der einzelnen Schichten variieren.

Kommentare

- Die Ziehung von Clusterstichproben ist sinnvoll, wenn Sie keine vollständige Auflistung der Grundgesamtheit, aus der die Stichprobe gezogen werden soll, beschaffen können, aber vollständige Listen für bestimmte Gruppen bzw. Cluster zugänglich sind. Außerdem wird sie verwendet, wenn eine Zufallsstichprobe zu einer Liste mit Testsubjekten führen würde, mit denen eine Kontaktaufnahme nicht praktikabel wäre. Es wäre beispielsweise einfacher, alle Bauern in einem bestimmten Landkreis bzw. Bezirk zu besuchen, als eine Auswahl von Bauern, die über alle Landkreise oder Bezirkes des Staates verstreut sind.
- Sie können auch sowohl ein Cluster- als auch ein Schichtungsfeld angeben, um Cluster innerhalb der einzelnen Schichten unabhängig voneinander zu ziehen. Sie können beispielsweise eine Stichprobe der Eigentumswerte ziehen, die nach Bezirk geschichtet ist, und dann innerhalb der einzelnen Bezirke Cluster bilden, die auf den Gemeinden beruhen. Dadurch wird sichergestellt, dass aus jedem Bezirk eine unabhängige Stichprobe der Gemeinden gezogen wird. Einige Gemeinden werden aufgenommen, andere dagegen nicht, aber bei jeder aufgenommenen Gemeinde werden alle Eigenschaften innerhalb der Gemeinde aufgenommen.
- Um eine Zufallsstichprobe der Einheiten aus den einzelnen Clustern zu ziehen, können Sie zwei Stichprobenknoten miteinander verknüpfen. So können Sie beispielsweise zuerst eine nach Bezirk geschichtete Stichprobe der Gemeinden ziehen. Anschließend können Sie einen zweiten Stichprobenknoten anfügen und *town* (Gemeinde) als Schichtungsfeld auswählen. Dadurch können Sie aus jeder Gemeinde einen Anteil an Datensätzen als Stichprobe ziehen.
- In Fällen, in denen für eine eindeutige Identifizierung der Cluster eine Kombination von Feldern erforderlich ist, können Sie mithilfe eines Ableitungsknotens ein neues Feld generieren. Beispiel: Wenn mehrere Läden dasselbe Nummerierungssystem für Transaktionen verwenden, könnten Sie ein neues Feld ableiten, das die Geschäfts- und die Transaktions-ID miteinander verknüpft.

Stichprobengrößen für Schichten

Beim Ziehen einer geschichteten Stichprobe besteht die Standardoption darin, aus jeder Schicht denselben Anteil an Datensätzen oder Clustern zu ziehen. Wenn eine Gruppe einer anderen beispielsweise zahlenmäßig um den Faktor 3 überlegen ist, soll dieses Verhältnis normalerweise in der Stichprobe erhalten bleiben. Andernfalls können Sie die Stichprobengröße für jede Schicht separat angeben.

Im Dialogfeld "Stichprobengrößen für Schichten" werden die einzelnen Werte des Schichtungsfelds aufgeführt, sodass Sie den Standardwert für die betreffende Schicht außer Kraft setzen können. Wenn mehrere Schichtungsfelder ausgewählt sind, wird jede mögliche Wertekombination aufgelistet, sodass Sie beispielsweise die Größe für die verschiedenen ethnischen Gruppen in den verschiedenen Städten oder die Größe der verschiedenen Gemeinden innerhalb der einzelnen Bezirke angeben können. Die Größen werden als Anteile oder als Anzahl der beobachteten Werte angegeben, je nachdem was in der aktuellen Einstellung im Stichprobenknoten festgelegt ist.

So geben Sie Stichprobengrößen für Schichten an:

1. Wählen Sie im Stichprobenknoten die Option **Komplex** und wählen Sie mindestens ein Schichtungsfeld aus. Weitere Informationen finden Sie im Thema „Einstellungen unter "Cluster und Schichtung"" auf Seite 70.
2. Wählen Sie die Option **Angepasst** und dann **Größen angeben**.
3. Klicken Sie im Dialogfeld "Stichprobengrößen für Schichten" auf die Schaltfläche **Werte lesen** links unten, um die Anzeige auszufüllen. Ggf. müssen Sie Werte in einem weiter oben liegenden Quellen- oder Typknoten instanziieren. Weitere Informationen finden Sie im Thema „Was ist Instanziierung?" auf Seite 121.
4. Klicken Sie in eine Zeile, um die Standardgröße für die betreffende Schicht zu überschreiben.

Hinweise zur Stichprobengröße

Benutzerdefinierte Stichprobengrößen können nützlich sein, wenn verschiedene Schichten eine unterschiedliche Varianz aufweisen, beispielsweise, um Stichprobengrößen proportional zur Standardabweichung zu machen. (Wenn die Fälle innerhalb der Schicht eine größere Variation aufweisen, müssen Sie mehr davon ziehen, um eine repräsentative Stichprobe zu erhalten.) Bei kleinen Schichten kann ein höherer Stichprobenanteil sinnvoll sein, um sicherzustellen, dass eine Mindestanzahl von Beobachtungen aufgenommen wird.

Hinweis: Wenn Sie die Schichtung anhand eines Felds mit fehlenden Werten (null oder systemdefiniert fehlende Werte, leere Zeichenfolgen, leere Bereiche und Leerstellen oder benutzerdefiniert fehlende Werte) vornehmen, können Sie keine benutzerdefinierten Stichprobengrößen für die Schichten angeben. Wenn Sie bei der Schichtung nach einem Feld mit fehlenden Werten oder Leerwerten benutzerdefinierte Stichprobengrößen verwenden möchten, müssen Sie die fehlenden Werte weiter oben im Stream ergänzen.

Balancierungsknoten

Mithilfe von Balancierungsknoten können Sie Unausgewogenheiten in den Datasets korrigieren, sodass Sie den angegebenen Testkriterien entsprechen. Beispiel: Angenommen, ein Dataset weist nur zwei Werte auf - *niedrig* und *hoch* - und 90 % der Fälle sind *niedrig* und nur 10 % der Fälle *hoch*. Bei vielen Modellierungsverfahren gibt es Schwierigkeiten mit solchen verzerrten Daten, weil sie in der Regel nur die *niedrigen* Ergebnisse berücksichtigen und die *hohen* ignorieren, da diese seltener sind. Bei ausgewogenen Daten mit ungefähr gleich vielen Ergebnissen vom Typ *niedrig* und *hoch* haben die Modelle eine bessere Chance, Muster zu finden, die zur Unterscheidung zwischen den beiden Gruppen dienen können. In diesem Fall kann mit einem Balancierungsknoten eine Balancierungsanweisung erstellt werden, die die Anzahl der Fälle mit dem Ergebnis vom Typ *niedrig* reduziert.

Die Balancierung erfolgt durch das Duplizieren und anschließende Verwerfen von Datensätzen auf der Grundlage der von Ihnen angegebenen Bedingungen. Datensätze, für die keine Bedingung gilt, werden immer übergeben. Da dieser Vorgang auf der Duplizierung und/oder dem Verwerfen von Datensätzen beruht, kann die ursprüngliche Sequenz Ihrer Daten in den nachgeordneten Operationen nicht erhalten bleiben. Daher müssen Sie alle sequenzbezogenen Werte ableiten, bevor Sie einen Balancierungsknoten zum Datenstream hinzufügen.

Hinweis: Balancierungsknoten können automatisch aus Verteilungsdiagrammen und Histogrammen generiert werden. Beispielsweise können Sie die Daten balancieren, sodass sie in allen Kategorien eines kategorialen Felds gleiche Anteile anzeigen, wie in einem Verteilungsdiagramm gezeigt.

Beispiel. Bei der Erstellung eines RFM-Streams zur Ermittlung aktueller Kunden, die positiv auf frühere Marketingkampagnen reagiert haben, verwendet die Marketingabteilung einer Vertriebsgesellschaft einen Balancierungsknoten, um die Unterschiede zwischen den Wahr- und den Falsch-Antworten in den Daten auszugleichen.

Festlegen der Optionen für den Balancierungsknoten

Anweisungen für Datensatzgewichtung. Listet die aktuellen Gewichtungsanweisungen auf. Zu jeder Anweisung gehören ein Faktor und eine Bedingung, die die Software anweist, den Anteil der Datensätze um einen angegebenen Faktor zu erhöhen, wenn die Bedingung wahr ist. Bei einem Faktor von unter 1,0 wird der Anteil der angegebenen Datensätze verringert. Beispiel: Angenommen, Sie möchten die Anzahl der Datensätze, bei denen die Behandlung mit Medikament Y erfolgt, verringern. Dann können Sie beispielsweise eine Gewichtungsanweisung mit dem Faktor 0,7 und der Bedingung `Medikament = "MedikamentY"` erstellen. Diese Anweisung führt dazu, dass die Anzahl der Datensätze, bei denen die Behandlung mit Medikament Y erfolgt, für alle nachgeordneten Operationen auf 70 % reduziert wird.

Hinweis: Balancierungsfaktoren für die Reduzierung können bis auf vier Dezimalstellen angegeben werden. Faktoren, die unter 0,0001 festgelegt werden, führen zu einem Fehler, da die Ergebnisse nicht ordnungsgemäß berechnet werden.

- **Zum Erstellen von Bedingungen** klicken Sie auf die Schaltfläche rechts neben dem Textfeld. Dadurch wird eine leere Zeile zur Eingabe neuer Bedingungen eingefügt. Um einen CLEM-Ausdruck für die Bedingung zu erstellen, klicken Sie auf die Schaltfläche "Expression Builder".
- **Zum Löschen von Anweisungen** verwenden Sie die rote Löschschaltfläche.
- **Zum Sortieren von Anweisungen** verwenden Sie die Schaltflächen mit den Aufwärts- und Abwärtspfeilen.

Balancierung nur für Trainingsdaten durchführen. Wenn im Stream ein Partitionsfeld vorhanden ist, wird bei Auswahl dieser Option die Balancierung ausschließlich in der Trainingspartition durchgeführt. Dies kann insbesondere bei der Generierung von Adjusted-Propensity-Scores nützlich sein, da diese eine unausgewogene Test- bzw. Validierungspartition erfordern. Wenn im Stream kein Partitionsfeld vorhanden ist (oder wenn mehrere Partitionsfelder angegeben sind), wird diese Option ignoriert und alle Daten werden balanciert.

Aggregatknoten

Aggregation ist eine Vorbereitungsaufgabe, die häufig zur Reduzierung der Größe eines Datasets verwendet wird. Bevor Sie mit der Aggregation fortfahren, sollten Sie sich die Zeit nehmen, die Daten zu bereinigen. Achten Sie dabei insbesondere auf fehlende Daten. Bei der Aggregierung können potenziell nützliche Informationen zu fehlenden Werten verloren gehen.

Mit einem Aggregatknoten können Sie eine Sequenz von Eingabedatensätzen mit aggregierten Übersichts-Ausgabedatensätzen ersetzen. Beispielsweise könnten Sie ein Set von Eingabeverkaufsdatensätzen haben, ähnlich den in der folgenden Tabelle gezeigten.

Tabelle 9. Beispiel für Eingabeverkaufsdatensätze

Alter	Geschlecht	Region	Zweigstelle	Verkäufe
23	M	S	8	4
45	M	S	16	4
37	M	S	8	5
30	M	S	5	7
44	M	N	4	9
25	M	N	Z	11
29	W	S	16	6
41	W	N	4	8
23	W	N	6	Z
45	W	N	4	5
33	W	N	6	10

Sie können diese Datensätze mit *Geschlecht* und *Region* als Schlüsselfelder aggregieren. Legen Sie anschließend fest, dass *Alter* mit dem Modus **Mittelwert** und *Umsatz* mit dem Modus **Summe** aggregiert werden soll. Wenn Sie anschließend im Dialogfeld des Aggregatknotens die Option **Datensatzanzahl einschließen in Feld** auswählen, lautet die aggregierte Ausgabe wie in der folgenden Tabelle gezeigt:

Tabelle 10. Beispiel für aggregierten Datensatz

Alter (Mittelwert)	Geschlecht	Region	Verkäufe (Summe)	Datensatzanzahl
35,5	W	N	25	4
29	W	S	6	E
34,5	M	N	20	Z
33,75	M	S	20	4

Daraus können Sie beispielsweise entnehmen, dass das Durchschnittsalter der vier weiblichen Angehörigen des Vertriebspersonals in der Region "Nord" 35,5 Jahre beträgt und dass sie insgesamt 25 Einheiten verkauft haben.

Hinweis: Felder wie *Zweigstelle* werden automatisch verworfen, wenn kein Aggregatmodus angegeben wurde.

Festlegen der Optionen für den Aggregatknoten

Im Aggregatknoten geben Sie Folgendes an:

- Ein oder mehrere Schlüsselfelder zur Verwendung als Kategorien für die Aggregation
- Ein oder mehrere Aggregatfelder, für die die Aggregatwerte berechnet werden sollen
- Ein oder mehrere Aggregatmodi (Aggregattypen), die für die einzelnen Aggregatfelder ausgegeben werden sollen

Sie können auch die Standardaggregationsmodi angeben, die für neu hinzugefügte Felder verwendet werden sollen, und Ausdrücke (ähnlich wie Formeln) verwenden, um die Aggregation zu kategorisieren.

Beachten Sie, dass Aggregationsvorgänge gegebenenfalls durch Parallelverarbeitung beschleunigt werden können, um eine bessere Leistung zu erzielen.

Schlüsselfelder. Listet Felder auf, die als Kategorien für die Aggregation verwendet werden können. Sowohl stetige (numerische) als auch kategoriale Felder können als Schlüssel verwendet werden. Bei Auswahl von mehreren Schlüsselfeldern werden die Werte kombiniert und ergeben einen Schlüsselwert für die Aggregation von Datensätzen. Für jedes eindeutige Schlüsselfeld wird jeweils ein (1) aggregierter Datensatz generiert. Bei den Schlüsselfeldern *Geschlecht* und *Region* beispielsweise erhält jede eindeutige Kombination von *M* und *W* mit den Regionen *N* und *S* (vier eindeutige Kombinationen) einen aggregierten Datensatz. Verwenden Sie zum Hinzufügen eines Schlüsselfelds die Feldauswahlschaltfläche auf der rechten Seite des Fensters.

Schlüssel sind zusammenhängend. Wählen Sie diese Option aus, wenn Sie wissen, dass alle Datensätze mit denselben Schlüsselwerten in der Eingabe als zusammenhängende Gruppe vorliegen (z. B. wenn die Eingabe nach Schlüsselfeldern sortiert ist). Dadurch lässt sich eventuell die Leistungsfähigkeit verbessern.

Das restliche Dialogfeld ist in zwei Hauptbereiche aufgeteilt - "Basisaggregate" und "Aggregatausdrücke".

Basisaggregate

Aggregatfelder. Listet die Felder auf, für die Werte aggregiert werden, sowie die ausgewählten Aggregationsmodi. Mit der Feldauswahlschaltfläche auf der rechten Seite können Sie Felder zu dieser Liste hinzufügen. Die folgenden Aggregationsmodi stehen zur Verfügung:

Hinweis: Einige Modi gelten nicht für nicht numerische Felder (z. B. **Summe** für Datums-/Zeitfelder). Modi, die bei einem ausgewählten Aggregatfeld nicht verwendet werden können, sind inaktiviert.

- **Summe.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination summierte Werte auszugeben. Die Summe der Werte über alle Fälle mit nicht fehlenden Werten.
- **Mittelwert.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination die Mittelwerte auszugeben. Der Mittelwert ist ein Lagemaß, bei dem es sich um den arithmetischen Durchschnitt (die Summe dividiert durch die Anzahl der Fälle) handelt.
- **Min.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination Mindestwerte auszugeben.
- **Max.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination Höchstwerte auszugeben.
- **Std.Abw.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination die Standardabweichung auszugeben. Die Standardabweichung ist ein Maß für die Streuung um den Mittelwert, definiert als positive Wurzel der Varianzmessung.
- **Median.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination die Medianwerte auszugeben. Der Median ist ein Lagemaß, das gegenüber Ausreißern unempfindlich ist (im Gegensatz zum Mittelwert, der durch wenige extrem niedrige oder hohe Werte beeinflusst werden kann). Auch als 50. Perzentil bzw. 2. Quartil bezeichnet.
- **Häufigkeiten.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination die Anzahl der Werte auszugeben, bei denen es sich nicht um Nullwerte handelt.
- **Varianz.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination die Varianzwerte auszugeben. Die Varianz ist ein Maß der Streuung um den Mittelwert. Sie ist gleich dem Quotienten aus der Summe der quadrierten Abweichung vom Mittelwert und der um 1 verringerten Fallanzahl.
- **1. Quartil.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination die Werte für das 1. Quartil (25. Perzentil) auszugeben.
- **3. Quartil.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination die Werte für das 3. Quartil (75. Perzentil) auszugeben.

Hinweis: Bei Ausführung eines Streams, der einen Aggregatknoten enthält, können sich bei einem SQL-Pushback an eine Oracle-Datenbank die Werte für das 1. und das 3. Quartil von den im nativen Modus zurückgegebenen Werten unterscheiden.

Standardmodus. Geben Sie den Standardaggregationsmodus an, der für neu hinzugefügte Felder verwendet werden soll. Wenn Sie häufig dieselbe Aggregation verwenden, wählen Sie hier einen oder mehrere Knoten aus und verwenden Sie die Schaltfläche "Auf alle anwenden" auf der rechten Seiten, um die ausgewählten Modi auf alle oben aufgeführten Felder zu übernehmen.

Neue Feldnamenerweiterung. Wählen Sie diese Option aus, um ein Suffix oder ein Präfix, beispielsweise "1" oder "neu" zu den duplizierten aggregierten Feldern hinzuzufügen. So führt eine Aggregation mit Mindestwerten beim Feld *Alter* zu einem Feld mit der Bezeichnung *Alter_Min_1*, wenn Sie Suffixoption ausgewählt und "1" als Erweiterung angegeben haben. *Hinweis:* Aggregationserweiterungen wie *_Min* oder *Max_* werden automatisch zum neuen Feld hinzugefügt und geben den Typ der durchgeführten Aggregation an. Wählen Sie **Suffix** bzw. **Präfix** aus, um die bevorzugte Erweiterungsart anzugeben.

Datensatzanzahl einschließen in Feld. Wählen Sie diese Option aus, um standardmäßig ein zusätzliches Feld mit der Bezeichnung *Datensatzanzahl* in jeden Ausgabedatensatz einzufügen. Dieses Feld gibt an, wie viele Eingabedatensätze aus den einzelnen Aggregatdatensätzen aggregiert wurden. Im Bearbeitungsfeld können Sie einen benutzerdefinierten Namen für dieses Feld angeben.

Hinweis: Systemdefinierte Nullwerte werden bei der Berechnung von Aggregaten ausgeschlossen, in der Datensatzanzahl sind sie jedoch enthalten. Leere Werte dagegen sind sowohl in der Aggregation als auch

in der Datensatzanzahl enthalten. Um leere Werte auszuschließen, ersetzen Sie mithilfe eines Füllerknotens Leerstellen durch Nullwerte. Außerdem können Sie Leerstellen mithilfe eines Auswahlknotens entfernen.

Aggregatausdrücke

Ausdrücke ähneln Formeln, die aus Werten, Feldnamen, Operatoren und Funktionen erstellt werden. Im Gegensatz zu Funktionen, die jeweils nur mit einem Datensatz arbeiten, arbeiten Aggregatausdrücke mit einer Gruppe, einem Set oder einer Datensatzsammlung.

Anmerkung: Sie können nur Aggregatausdrücke erstellen, wenn der Stream eine Datenbankverbindung enthält (mittels eines Datenbankquellenknotens).

Neue Ausdrücke werden als abgeleitete Felder erstellt. Für die Erstellung eines Ausdrucks verwenden Sie die Funktionen unter *Datenbankaggregate*, die in Expression Builder verfügbar sind.

Beachten Sie, dass eine Verbindung zwischen den **Schlüsselfeldern** und allen von Ihnen erstellten Aggregatausdrücken besteht, da die Aggregatausdrücke anhand des Schlüsselfelds gruppiert werden.

Gültige Aggregatausdrücke ergeben zusammengefasste Ergebnisse. Im Folgenden finden Sie einige Beispiele für gültige Aggregatausdrücke und die Regeln, die für sie gelten:

- Mithilfe von Skalarfunktionen können Sie mehrere Aggregationsfunktionen miteinander kombinieren, um ein einziges Aggregationsergebnis zu erhalten. Beispiel:
  ```
  max(C01) - min(C01)
  ```
- Eine Aggregationsfunktion kann mit dem Ergebnis mehrerer Skalarfunktionen arbeiten. Beispiel:
  ```
  sum (C01*C01)
  ```

RFM-Aggregatknoten

Mit dem Knoten "RFM-Aggregat" (Recency-, Frequency-, Monetary-Aggregat) können Sie Daten über die früheren Transaktionen von Kunden verwenden, alle nicht benötigten Daten entfernen und alle verbliebenen Transaktionsdaten zu einer einzigen Zeile zusammenfassen (mit der eindeutigen Kunden-ID als Schlüssel), die angibt, wann der betreffende Kunde zuletzt mit Ihnen in Geschäftskontakt stand (Recency - Aktualität), wie viele Transaktionen er vorgenommen hat (Frequency - Häufigkeit) und wie hoch der Gesamtwert dieser Transaktionen ist (Monetary - Geldwert).

Bevor Sie mit der Aggregation fortfahren, sollten Sie sich die Zeit nehmen, die Daten zu bereinigen. Achten Sie dabei insbesondere auf etwaige fehlende Daten.

Sobald Sie die Daten mithilfe des RFM-Aggregatknotens identifiziert und transformiert haben, können Sie mithilfe eines RFM-Analyseknotens weitere Analysen durchführen. Weitere Informationen finden Sie im Thema „Knoten "RFM-Analyse"" auf Seite 152.

Beachten Sie: Nach dem Durchlaufen des RFM-Aggregatknotens enthält die Datendatei keinerlei Zielwerte; daher können Sie sie erst dann als Eingabe für weitere Vorhersageanalysen mit anderen Modellierungsknoten, wie beispielsweise C5.0 oder CHAID verwenden, nachdem Sie sie mit anderen Kundendaten zusammengeführt haben (beispielsweise durch Abgleich der Kunden-IDs). Weitere Informationen finden Sie im Thema „Zusammenführungsknoten ("Mergen")" auf Seite 78.

Die Knoten "RFM-Aggregat" und "RFM-Analyse" in IBM SPSS Modeler sind für die Verwendung einer unabhängigen Klassierung eingerichtet. Damit werden also Daten für jedes der Maße Aktualität, Häufigkeit und Geldwert in Ränge eingeteilt und klassiert, ohne Berücksichtigung ihrer Werte oder der beiden anderen Maße.

Festlegen der Optionen für den RFM-Aggregatknoten

Die Registerkarte "Einstellungen" des Knoten "RFM-Aggregat" enthält die folgenden Felder.

Aktualität (Recency) berechnen relativ zu. Dient zur Angabe des Datums, ausgehend von dem die Aktualität der Transaktionen berechnet werden soll. Hierfür können Sie entweder unter **Festes Datum** ein Datum eingeben oder **Heutiges Datum** auswählen, wobei das aktuelle Datum laut Systemeinstellungen verwendet wird. Der Wert **Heutiges Datum** wird standardmäßig eingegeben und automatisch aktualisiert, wenn der Knoten ausgeführt wird.

IDs sind zusammenhängend. Wenn Ihre Daten vorsortiert sind, sodass alle Datensätze mit derselben ID zusammen im Datenstream erscheinen, wählen Sie diese Option aus, um die Verarbeitung zu beschleunigen. Wenn Ihre Daten nicht vorsortiert sind (oder Sie nicht sicher sind), lassen Sie diese Option inaktiviert. Die Daten werden dann vom Knoten automatisch sortiert.

ID. Dient zur Auswahl des für die Identifizierung des Kunden und seiner Transaktionen zu verwendenden Felds. Die zur Auswahl zur Verfügung stehenden Felder können Sie mit der Feldauswahlschaltfläche auf der rechten Seite anzeigen.

Datum. Dient zur Auswahl des Datumsfelds, das für die Berechnung der Aktualität (Recency) verwendet werden soll. Die zur Auswahl zur Verfügung stehenden Felder können Sie mit der Feldauswahlschaltfläche auf der rechten Seite anzeigen.

Beachten Sie, dass hierfür ein Feld mit dem Speichertyp "Datum" oder "Zeitmarke" im entsprechenden Format zur Verwendung als Eingabe erforderlich ist. Wenn Ihnen beispielsweise ein Zeichenfolgenfeld mit Werten wie *Jan 2000*, *Feb 2000* usw. vorliegt, können Sie dieses mithilfe eines Füllerknotens und der Funktion `to_date()` in ein Datumsfeld konvertieren. Weitere Informationen finden Sie im Thema „Speichertypkonvertierung mithilfe des Füllerknotens” auf Seite 140.

Wert. Dient zur Auswahl des Felds, das für die Berechnung des Gesamtwerts der Transaktionen des Kunden verwendet werden soll. Die zur Auswahl zur Verfügung stehenden Felder können Sie mit der Feldauswahlschaltfläche auf der rechten Seite anzeigen. *Hinweis:* Dies muss ein numerischer Wert sein.

Neue Feldnamenerweiterung. Wählen Sie diese Option aus, wenn Sie die neu erstellten Felder für Aktualität, Häufigkeit und Geldwert durch ein Suffix oder Präfix, wie beispielsweise "12_Monate", ergänzen möchten. Wählen Sie **Suffix** bzw. **Präfix** aus, um die bevorzugte Erweiterungsart anzugeben. Dies kann beispielsweise bei der Untersuchung mehrerer Zeitperioden nützlich sein.

Datensätze verwerfen mit Werten unter. Falls erforderlich, können Sie hier einen Mindestwert für die bei der Berechnung der RFM-Gesamtwerte verwendeten Transaktionsdetails angeben. Die für den Wert geltenden Einheiten beziehen sich auf das ausgewählte Feld **Wert**.

Nur aktuelle Transaktionen einschließen. Bei der Analyse großer Datenbanken können Sie angeben, dass nur die aktuellsten Datensätze verwendet werden sollen. Sie können auswählen, ob die nach einem bestimmten Datum oder innerhalb eines bestimmten Zeitraums protokollierten Daten verwendet werden sollen:

- **Transaktionsdatum nach.** Dient zur Angabe des Transaktionsdatums, nach dem die Datensätze in die Analyse aufgenommen werden sollen.
- **Transaktion innerhalb der letzten.** Hier können Sie anhand von Anzahl und Typ der Zeiträume (Tage, Wochen, Monate oder Jahre) angeben, wie weit ausgehend von **Aktualität (Recency) berechnen relativ zu** die in die Analyse aufzunehmenden Datensätze zurückliegen dürfen.

Datum der zweitaktuellsten Transaktion speichern. Aktivieren Sie dieses Kontrollkästchen, wenn Sie das Datum der zweitaktuellsten Transaktion für die einzelnen Kunden ermitteln möchten. Zusätzlich können Sie auch das Kontrollkästchen **Datum der drittaktuellsten Transaktion speichern** aktivieren. Dies kann

Ihnen beispielsweise dabei helfen, Kunden zu identifizieren, die möglicherweise vor längerer Zeit zahlreiche Transaktionen getätigt haben, aber nur eine aktuelle Transaktion.

Sortierknoten

Mit Sortierknoten können Sie Datensätze anhand der Werte eines oder mehrerer Felder in aufsteigender oder absteigender Reihenfolge kopieren. Sortierknoten werden beispielsweise häufig verwendet, um Datensätze mit den häufigsten Datenwerten anzuzeigen und auszuwählen. Üblicherweise werden die Daten zuerst mit dem Aggregatknoten aggregiert und die aggregierten Daten anschließend mit dem Sortierknoten in absteigender Reihenfolge nach Datensatzanzahl sortiert. Durch die Anzeige dieser Ergebnisse in einer Tabelle können Sie die Daten untersuchen und Entscheidungen treffen. Beispielsweise könnten Sie die Datensätze der 10 besten Kunden auswählen.

Die Registerkarte "Einstellungen" des Sortierknotens enthält die folgenden Felder.

Sortieren nach. Alle Felder, die zur Verwendung als Sortierschlüssel ausgewählt wurden, werden in einer Tabelle angezeigt. Schlüsselfelder sind am besten für die Sortierung geeignet, wenn sie numerisch sind.

- **Zum Hinzufügen von Feldern** zu dieser Liste verwenden Sie die Feldauswahlschaltfläche auf der rechten Seite.
- **Zur Auswahl einer Reihenfolge** klicken Sie auf den Pfeil **Aufsteigend** oder **Absteigend** in der Spalte *Reihenfolge* der Tabelle.
- **Zum Löschen von Feldern** verwenden Sie die rote Löschschaltfläche.
- **Zum Sortieren von Anweisungen** verwenden Sie die Schaltflächen mit den Aufwärts- und Abwärtspfeilen.

Standardsortierreihenfolge. Als Standardsortierreihenfolge, die für das Hinzufügen neuer Felder verwendet wird, können Sie entweder **Aufsteigend** oder **Absteigend** auswählen.

Optimierungseinstellungen für das Sortieren

Wenn Sie mit Daten arbeiten, die bereits nach bestimmten Schlüsselfeldern sortiert sind, können Sie diese Sortierungsfelder angeben, sodass die restlichen Daten effizienter sortiert werden können. Beispiel: Die Daten sollen nach *Alter* (absteigend) und *Medikament* (aufsteigend) sortiert werden; Sie wissen jedoch, dass die Daten bereits nach *Alter* (absteigend) sortiert sind.

Daten sind vorsortiert. Gibt an, ob die Daten bereits nach einem oder mehreren Feldern sortiert sind.

Bestehende Sortierreihenfolge angeben. Gibt die Felder an, die bereits sortiert sind. Über das Dialogfeld "Felder auswählen" nehmen Sie Felder in die Liste auf. Geben Sie in der Spalte *Reihenfolge* jeweils an, ob die Felder in aufsteigender oder absteigender Reihenfolge sortiert sind. Wenn Sie mehrere Felder angeben, achten Sie darauf, die Felder in der richtigen Sortierreihenfolge aufzuführen. Mit den Pfeilen rechts neben der Liste ordnen Sie die Felder in der richtigen Reihenfolge an. Wenn Sie die vorhandene Sortierreihenfolge nicht richtig angeben, tritt beim Ausführen des Streams ein Fehler auf. In der Fehlernachricht wird dabei die Nummer des Datensatzes angezeigt, bei dem die Sortierung nicht mit Ihren Angaben übereinstimmt.

Hinweis: Die Sortierung kann gegebenenfalls durch Parallelverarbeitung beschleunigt werden.

Zusammenführungsknoten ("Mergen")

Die Funktion von Zusammenführungsknoten ("Mergen") besteht darin, aus mehreren Eingabedatensätzen einen einzelnen Ausgabedatensatz mit allen oder einigen der Eingabefelder zu erstellen. Dies ist ein nützlicher Vorgang, wenn Daten aus verschiedenen Quellen, wie beispielsweise interne Kundendaten und käuflich erworbene demografische Daten, zusammengeführt werden sollen. Sie können Daten auf folgende Weisen zusammenführen:

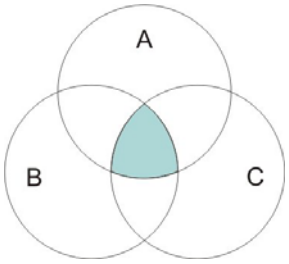
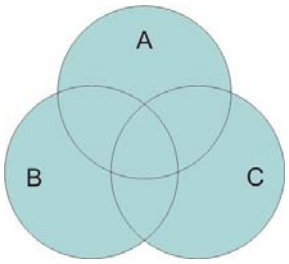
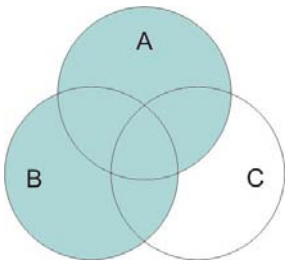
- **Beim Zusammenführen (Mergen) nach Reihenfolge** werden die entsprechenden Datensätze aus allen Quellen in der Reihenfolge der Eingabe miteinander verkettet, bis die kleinste Datenquelle erschöpft ist. Vor der Verwendung dieser Option müssen die Daten unbedingt mit einem Sortierknoten sortiert worden sein.
- **Das Zusammenführen (Mergen) mit einem Schlüsselfeld**, wie beispielsweise *Kunden-ID*, dient zur Angabe, wie die Datensätze aus einer Datenquelle mit Datensätzen aus den anderen Quellen abgeglichen werden können. Mehrere Arten von Joins sind möglich, beispielsweise Inner Join, Full Outer Join, Partieller Outer Join und Anti-Join. Weitere Informationen finden Sie im Thema „Jointypen".
- **Beim Zusammenführen (Mergen) nach Bedingung** können Sie eine Bedingung angegeben, die erfüllt sein muss, damit das Zusammenführen stattfindet. Sie können die Bedingung direkt im Knoten angeben, oder die Bedingung mithilfe des Expression Builder erstellen.

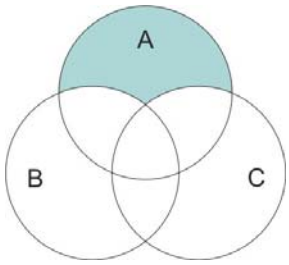
Jointypen

Bei der Verwendung eines Schlüsselfelds zum Zusammenführen von Daten sollten Sie vorher überlegen, welche Datensätze ausgeschlossen und welche eingeschlossen werden sollen. Es gibt eine Vielzahl von Joins, die unten im Detail erörtert werden.

Die beiden Join-Grundtypen heißen "Inner Join" und "Outer Join". Diese Methoden werden häufig zur Zusammenführung von Tabellen aus verwandten Datasets auf der Grundlage gemeinsamer Werte eines Schlüsselfelds, beispielsweise *Kunden-ID*, verwendet. Inner Joins ergeben eine saubere Zusammenführung und ein Ausgabedataset, das nur vollständige Datensätze enthält. Outer Joins beinhalten ebenfalls vollständige Datensätze aus den zusammengeführten Daten, doch sie ermöglichen auch die Aufnahme eindeutiger Daten aus einer oder mehreren Eingabetabellen.

Die zulässigen Jointypen sind weiter unten detaillierter beschrieben.

	Ein **Inner Join** enthält nur Datensätze, bei denen ein Wert für das Schlüsselfeld bei allen Eingabetabellen gleich ist. Nicht übereinstimmende Datensätze werden nicht in das Ausgabedataset aufgenommen.
	Bei einem **Full Outer Join** werden alle Datensätze (übereinstimmend und nicht übereinstimmend) aus den Eingabetabellen eingeschlossen. Linke und rechte Outer Joins werden als partielle Outer Joins bezeichnet und werden im Folgenden beschrieben.
	Ein **partieller Outer Join** enthält alle Datensätze, deren Übereinstimmung anhand des Schlüsselfelds abgeglichen wurde, sowie nicht übereinstimmende Datensätze aus den angegebenen Tabellen. (Oder anders gesagt: Alle Datensätze aus bestimmten Tabellen und nur passende Datensätze aus anderen Tabellen.) Tabellen (wie beispielsweise A und B in der Abbildung) können mithilfe der Schaltfläche "Auswahl" auf der Registerkarte "Verbinden" ausgewählt werden. Partielle Joins werden auch linke bzw. rechte Outer Joins genannt, wenn nur zwei Tabellen zusammengeführt werden. Da IBM SPSS Modeler die Zusammenführung von mehr als zwei Tabellen erlaubt, wird dieser Vorgang hier als partieller Outer Join bezeichnet.

	Bei **Anti-Join** werden nur nicht übereinstimmende Datensätze für die erste Eingabetabelle (Tabelle A in der Abbildung) aufgenommen. Bei diesem Jointyp handelt es sich um das Gegenteil eines Inner Join. Es werden keine vollständigen Datensätze in das Ausgabedataset aufgenommen.

Wenn beispielsweise Informationen über Bauernhöfe in einem Dataset vorliegen und Versicherungsansprüche zu Bauernhöfen in einem zweiten Dataset, dann können Sie die Datensätze aus der ersten Quelle mithilfe der Zusammenführungsoptionen mit den Datensätzen aus der zweiten Quelle abgleichen.

Um festzustellen, ob ein Kunde in diesem Bauernhof-Beispiel einen Versicherungsanspruch angemeldet hat, rufen Sie mit der Option "Inner Join" eine Liste mit allen IDs ab, die in beiden Datasets vorkommen.

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimtype	claimvalue
1	id604	name604	southwest	1860.000	103.0...	3.000	625251.000	potatoes	decomm...	281082.0...
2	id605	name605	north	1700.000	46.000	8.000	621148.000	wheat	decomm...	122006.0...
3	id620	name620	north	880.000	74.000	6.000	426988.000	rapeseed	arable_de	118885.0...

Abbildung 2. Beispielausgabe für eine Zusammenführung mit Inner Join

Bei einem Full Outer Join werden alle Datensätze (übereinstimmend und nicht übereinstimmend) aus den Eingabetabellen eingeschlossen. Bei unvollständigen Werten wird der systemdefiniert fehlende Wert (`$null$`) verwendet.

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimtype	claimvalu
1	id601	$null$	$null$	$null$	$null$	$null$	$null$	$null$	decomm...	74703.1(
2	id602	name602	north	1780.000	42.000	9.000	734118.000	maize	$null$	$nul
3	id604	name604	southwest	1860.000	103.0...	3.000	625251.000	potatoes	decomm...	281082.0
4	id605	name605	north	1700.000	46.000	8.000	621148.000	wheat	decomm...	122006.0
5	id606	$null$	$null$	$null$	$null$	$null$	$null$	$null$	arable_de	122135.0

Abbildung 3. Beispielausgabe für eine Zusammenführung mit Full Outer Join

Ein partieller Outer Join enthält alle Datensätze, deren Übereinstimmung anhand des Schlüsselfelds abgeglichen wurde, sowie nicht übereinstimmende Datensätze aus den angegebenen Tabellen. Die Tabelle zeigt alle Datensätze, die mit dem ID-Feld übereinstimmen, sowie alle Datensätze, die mit dem ersten Dataset übereinstimmen.

	id	claimtype	claimvalue	name	region	farmsize	rainfall	landquality	farmincome	maincrop
1	id602	$null$	$null$	name602	north	1780.000	42.000	9.000	734118.000	maize
2	id604	decomm...	281082.0...	name604	southwest	1860.000	103.0...	3.000	625251.000	potatoes
3	id605	decomm...	122006.0...	name605	north	1700.000	46.000	8.000	621148.000	wheat
4	id607	$null$	$null$	name607	southeast	1820.000	29.000	6.000	211605.000	maize
5	id608	$null$	$null$	name608	southeast	1640.000	108.0...	7.000	1167040.0...	maize
6	id609	$null$	$null$	name609	southwest	1600.000	101.0...	5.000	756755.000	wheat
7	id615	$null$	$null$	name615	midlands	920.000	86.000	6.000	442554.000	potatoes
8	id618	$null$	$null$	name618	southeast	1180.000	98.000	3.000	368646.000	maize

Abbildung 4. Beispielausgabe für eine Zusammenführung mit partiellem Outer Join

Bei Anti-Join gibt die Tabelle nur nicht übereinstimmende Datensätze für die erste Eingabetabelle aus.

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop
1	id602	name602	north	1780.000	42.000	9.000	734118.000	maize
2	id607	name607	southeast	1820.000	29.000	6.000	211605.000	maize
3	id608	name608	southeast	1640.000	108.0...	7.000	1167040.0...	maize
4	id609	name609	southwest	1600.000	101.0...	5.000	756755.000	wheat
5	id615	name615	midlands	920.000	86.000	6.000	442554.000	potatoes
6	id618	name618	southeast	1180.000	98.000	3.000	368646.000	maize
7	id619	name619	north	840.000	64.000	8.000	457552.000	potatoes

Abbildung 5. Beispielausgabe für eine Zusammenführung mit Anti-Join

Angeben eines Zusammenführungsverfahrens und von Schlüsseln

Die Registerkarte "Verbinden" des Zusammenführungsknotens enthält die folgenden Felder.

Zusammenführungsmethode. Wählen Sie entweder **Reihenfolge** oder **Schlüssel** aus, um die Methode für die Zusammenführung von Datensätzen anzugeben. Durch die Auswahl von **Schlüssel** wird der untere Teil des Dialogfelds aktiviert.

- **Reihenfolge.** Führt Datensätze nach der Reihenfolge zusammen. Beispielsweise wird der *n*-te Datensatz aus jeder Eingabe zusammengeführt, um den *n*-ten Ausgabedatensatz zu erstellen. Wenn für einen Datensatz kein übereinstimmender Eingabedatensatz mehr vorhanden ist, werden keine weiteren Ausgabedatensätze erstellt. Die Anzahl der erstellten Datensätze ist also gleich der Anzahl der Datensätze im kleinsten Dataset.
- **Schlüssel.** Verwendet ein Schlüsselfeld wie *Transaktions-ID*, um Datensätze mit demselben Wert im Schlüsselfeld zusammenzuführen. Dies entspricht einem Datenbank-"Equi-Join". Wenn ein Schlüsselwert mehrmals vorkommt, werden alle möglichen Kombinationen ausgegeben. Beispiel: Wenn Datensätze mit demselben Schlüsselfeldwert *A* verschiedene Werte für *B*, *C* und *D* in anderen Feldern enthalten, erstellen die zusammengeführten Felder einen separaten Datensatz für die einzelnen Kombinationen von *A* mit Wert *B*, *A* mit Wert *C* und *A* mit Wert *D*.

  Hinweis: Nullwerte werden bei der Zusammenführung nach Schlüssel nicht als identisch betrachtet und werden nicht zusammengeführt.
- **Bedingung.** Verwenden Sie diese Option, um eine Bedingung für die Zusammenführung anzugeben. Weitere Informationen finden Sie im Thema „Angeben von Bedingungen für das Zusammenführen” auf Seite 82.

Mögliche Felder. Listet nur die Felder mit identischen Namen in allen Eingabedatenquellen auf. Wählen Sie ein Feld aus dieser Liste aus und fügen Sie es mithilfe der Pfeilschaltflächen als Schlüsselfeld für die Zusammenführung von Datensätzen hinzu. Es können mehrere Schlüsselfelder verwendet werden. Sie können nicht übereinstimmende Eingabefelder mittels eines Filterknotens oder über die Registerkarte "Filter" eines Quellenknotens umbenennen.

Verwendete Schlüsselfelder. Listet alle Felder auf, die für die Zusammenführung der Datensätze aus allen Eingabedatenquellen auf der Grundlage der Schlüsselfeldwerte verwendet werden. Um einen Schlüssel aus der Liste zu entfernen, wählen Sie ihn aus und verschieben Sie ihn mithilfe der Pfeilschaltfläche zurück in die Liste "Mögliche Schlüsselfelder". Bei Auswahl mehrerer Schlüssel wird die unten stehende Option aktiviert.

Doppelte Schlüsselfelder kombinieren. Wenn oben mehrere Felder ausgewählt wurden, gewährleistet diese Option, dass nur ein einziges Ausgabefeld dieses Namens vorhanden ist. Die aktivierte Option ist standardmäßig aktiviert, außer wenn Streams aus früheren Versionen von IBM SPSS Modeler importiert wurden. Wenn diese Option inaktiviert ist, müssen doppelte Schlüsselfelder mithilfe der Registerkarte "Filter" im Dialogfeld des Zusammenführungsknotens ("Mergen") umbenannt oder ausgeschlossen werden.

Nur übereinstimmende Datensätze einschließen (Inner Join). Wählen Sie diese Option aus, um nur vollständige Datensätze zusammenzuführen.

Übereinstimmende und nicht übereinstimmende Datensätze einschließen (Full Outer Join). Wählen Sie diese Option aus, um einen Full Outer Join durchzuführen. Wenn also Werte für das Schlüsselfeld nicht in allen Eingabetabellen vorhanden sind, werden die unvollständigen Datensätze dennoch beibehalten. Der nicht definierte Wert (`$null$`) wird zum Schlüsselfeld hinzugefügt und in den Ausgabedatensatz aufgenommen.

Partieller Outer Join (Übereinstimmende und teilweise nicht übereinstimmende Datensätze einschließen). Mit dieser Option können Sie einen partiellen Outer Join der Tabellen durchführen, die Sie in einem Unterdialogfeld ausgewählt haben. Klicken Sie auf **Auswählen**, um die Tabellen anzugeben, für die die unvollständigen Datensätze in der Zusammenführung beibehalten werden.

Anti-Join (Datensätze in erstes Dataset aufnehmen, die nicht mit anderen übereinstimmen). Wählen Sie diese Option aus, um eine Art "Anti-Join" durchzuführen, bei dem nur nicht übereinstimmende Datensätze aus dem ersten Dataset an die nachgeordneten Knoten übergeben werden. Mithilfe der Pfeile auf der Registerkarte "Eingaben" können Sie die Reihenfolge der Eingabedatasets angeben. Bei diesem Jointyp werden keine vollständigen Datensätze in das Ausgabedataset eingeschlossen. Weitere Informationen finden Sie im Thema „Jointypen” auf Seite 79.

Auswählen von Daten für partielle Joins

Für einen partiellen Outer Join müssen Sie die Tabelle(n) auswählen, für die unvollständige Datensätze beibehalten werden sollen. Beispielsweise könnten Sie alle Datensätze aus einer Kundentabelle beibehalten, jedoch nur übereinstimmende Datensätze aus der Tabelle "Hypothekendarlehen".

Spalte "Outer Join". Wählen Sie in der Spalte *Outer Join* die Datasets aus, die vollständig aufgenommen werden sollen. Bei einem partiellen Join werden überlappende Datensätze sowie unvollständige Datensätze für die hier ausgewählten Datasets beibehalten. Weitere Informationen finden Sie im Thema „Jointypen” auf Seite 79.

Angeben von Bedingungen für das Zusammenführen

Wenn Sie das Zusammenführungsverfahren auf **Bedingung** setzen, können Sie eine oder mehrere Bedingungen angegeben, die erfüllt sein müssen, damit das Zusammenführen stattfindet.

Sie können die Bedingungen entweder direkt in das Feld "Bedingung" eingeben oder sie mithilfe des Expression Builder erstellen, indem Sie auf das Rechnersymbol rechts neben dem Feld klicken.

Filtern von Feldern aus dem Zusammenführungsknoten ("Mergen")

Zusammenführungsknoten bieten eine bequeme Möglichkeit zum Filtern oder Umbenennen doppelter Felder, die beim Zusammenführen mehrerer Datenquellen entstehen. Klicken Sie auf die Registerkarte **Filter** im Dialogfeld, um Filteroptionen auszuwählen.

Die hier verfügbaren Optionen sind annähernd mit denen für den Filterknoten identisch. Im Filtermenü stehen jedoch zusätzliche Optionen zur Verfügung, die hier nicht erörtert werden. Weitere Informationen finden Sie im Thema „Filtern oder Umbenennen von Feldern” auf Seite 129.

Feld. Zeigt die Eingabefelder aus den aktuell verbundenen Datenquellen an.

Tag. Listet den Tagnamen (bzw. die Nummer) auf, der der Datenquellenverknüpfung zugeordnet ist. Klicken Sie auf die Registerkarte **Eingaben**, um aktive Verknüpfungen mit diesem Zusammenführungsknoten ("Mergen") zu ändern.

Quellenknoten. Zeigt den Quellenknoten an, dessen Daten zusammengeführt werden.

Verbundener Knoten. Zeigt den Knotennamen für den Knoten an, der mit dem Zusammenführungsknoten ("Mergen") verbunden ist. Häufig sind beim komplexen Data Mining mehrere Zusammenführungs- bzw. Anhangsvorgänge erforderlich, die denselben Quellenknoten beinhalten können. Der Name des verbundenen Knotens bietet eine Möglichkeit zur Unterscheidung

Filter. Zeigt die aktiven Verbindungen zwischen Eingabe- und Ausgabefeld an. Aktive Verbindungen weisen einen nicht unterbrochenen Pfeil auf. Verbindungen mit einem roten X weisen auf gefilterte Felder hin.

Feld. Führt die Ausgabefelder nach dem Zusammenführen oder Anhängen auf. Doppelte Felder werden in roter Farbe angezeigt. Klicken Sie auf das Filterfeld oben, um doppelte Felder zu inaktivieren.

Aktuelle Felder anzeigen. Verwenden Sie diese Option, um Informationen zu den Feldern auszuwählen, die als Schlüsselfelder verwendet werden sollen.

Nicht verwendete Feldeinstellungen anzeigen. Wählen Sie diese Option aus, um Informationen zu Feldern auszuwählen, die derzeit nicht verwendet werden.

Festlegen der Eingabereihenfolge und Tagkennzeichnung

Auf der Registerkarte "Eingaben" der Dialogfelder der Zusammenführungs- und Anhangknoten können Sie die Reihenfolge der Eingabedatenquellen angeben und etwaige Änderungen an den Tagnamen für die einzelnen Quellen vornehmen.

Tags und Reihenfolge der Eingabedatasets. Wählen Sie diese Option aus, um nur vollständige Datensätze anzuhängen.

- **Tag.** Listet die aktuellen Tag-Namen für die einzelnen Eingabedatenquellen auf. Tagnamen bzw. **Tags** dienen zur eindeutigen Kennzeichnung der Datenverknüpfungen für das Zusammenführen oder Anhängen. Stellen Sie sich Wasser aus verschiedenen Leitungen vor, das an einem bestimmten Punkt zusammengeführt wird und ab da durch eine einzige Leitung fließt. Die Daten in IBM SPSS Modeler fließen in ähnlicher Weise und der Zusammenführungspunkt ist häufig eine komplexe Interaktion zwischen den verschiedenen Datenquellen. Tags bieten eine Möglichkeit zur Verwaltung der Eingaben ("Leitungen") für einen Zusammenführungs- oder Anhangknoten, sodass beim Speichern oder Trennen des Knotens die Links erhalten bleiben und leicht zu identifizieren sind.

Wenn Sie weitere Datenquellen mit einem Zusammenführungs- oder Anhangknoten verbinden, werden automatisch Standardtags erstellt, bei denen die Reihenfolge, in der die Knoten verbunden wurden, durch Zahlen gekennzeichnet wird. Diese Reihenfolge steht in keinem Zusammenhang mit den Feldern im Ein- oder Ausgabedataset. Sie können das Standardtag ändern, indem Sie in der Spalte *Tag* einen neuen Namen eingeben.

- **Quellenknoten.** Zeigt den Quellenknoten an, dessen Daten zusammengefasst werden.
- **Verbundener Knoten.** Zeigt den Knotennamen für den Knoten an, der mit dem Zusammenführungs- oder Anhangknoten verbunden ist. Häufig sind beim komplexen Data Mining mehrere Zusammenführungsvorgänge erforderlich, die denselben Quellenknoten beinhalten können. Der Name des verbundenen Knotens bietet eine Möglichkeit zur Unterscheidung
- **Felder.** Führt die Anzahl der Felder in den einzelnen Datenquellen auf.

Aktuelle Tags anzeigen. Wählen Sie diese Option aus, um Tags anzuzeigen, die aktiv vom Zusammenführungs- oder Anhangknoten verwendet werden. Anders ausgedrückt: Die aktuellen Tags kennzeichnen Links zu dem Knoten, bei denen ein Datenfluss vorliegt. In der Leitungsmetapher entsprechen die aktuellen Tags den Leitungen, in denen Wasser fließt.

Nicht verwendete Tageinstellungen anzeigen. Wählen Sie diese Option aus, um Tags bzw. Links anzuzeigen, die zuvor für Verbindungen mit dem Zusammenführungs- bzw. Anhangknoten verwendet wurden, die derzeit jedoch nicht mit einer Datenquelle verbunden sind. Dies entspricht leeren Leitungen,

die in einem Leitungssystem noch intakt sind. Sie können diese "Leitungen" mit einer neuen Quelle verbinden oder sie entfernen. Um nicht verwendete Tags aus dem Knoten zu entfernen, klicken Sie auf **Löschen**. Dadurch werden alle nicht verwendeten Tags sofort gelöscht.

Optimierungseinstellungen für das Zusammenführen

Es stehen zwei Optionen zur Auswahl, mit denen Sie die Daten in bestimmten Situationen effektiver zusammenführen. Diese Optionen optimieren die Zusammenführung, wenn ein Eingabedataset deutlich größer ist als die anderen Datasets oder wenn Ihre Daten bereits nach einigen oder allen Schlüsselfeldern sortiert sind, die beim Zusammenführen herangezogen werden.

Ein Eingabedataset ist relativ groß. Mit dieser Option geben Sie an, dass eines der Eingabedatasets deutlich größer ist als die anderen Datasets. Die kleineren Datasets werden zwischengespeichert. Bei der anschließenden Zusammenführung wird das große Dataset ohne Zwischenspeichern und ohne Sortieren verarbeitet. Dieser Jointyp bietet sich in der Regel für Daten an, die ein Sternen-Schema oder einen ähnlichen Aufbau besitzen, bei dem eine große, zentrale Tabelle mit gemeinsam genutzten Daten vorliegt (z. B. bei Transaktionsdaten). Wenn Sie diese Option aktivieren, klicken Sie auf **Auswählen** und geben Sie das große Dataset an. Hierbei ist zu beachten, dass Sie nur *ein* großes Dataset festlegen können. Die nachstehende Tabelle bietet einen Überblick über die Jointypen, die mit dieser Methode optimiert werden können.

Tabelle 11. Übersicht über Optimierungsmöglichkeiten für Joins

Jointyp	Optimierung für ein großes Eingabedataset möglich?
Inner	Ja
Partiell...	Ja, wenn das große Dataset keine unvollständigen Datensätze enthält.
Full	Nein
Anti-Join	Ja, wenn das große Dataset die erste Eingabe bildet.

Alle Datensätze sind bereits durch Schlüsselfelder sortiert. Mit dieser Option geben Sie an, dass die Eingabedaten bereits nach mindestens einem der Schlüsselfelder sortiert sind, die beim Zusammenführen herangezogen werden sollen. Stellen Sie sicher, dass *alle* Eingabedatasets sortiert sind.

Bestehende Sortierreihenfolge angeben. Gibt die Felder an, die bereits sortiert sind. Über das Dialogfeld "Felder auswählen" nehmen Sie Felder in die Liste auf. Sie können nur unter den Schlüsselfeldern wählen, die für die Zusammenführung verwendet werden (auf der Registerkarte "Verbinden" angegeben). Geben Sie in der Spalte *Reihenfolge* jeweils an, ob die Felder in aufsteigender oder absteigender Reihenfolge sortiert sind. Wenn Sie mehrere Felder angeben, achten Sie darauf, die Felder in der richtigen Sortierreihenfolge aufzuführen. Mit den Pfeilen rechts neben der Liste ordnen Sie die Felder in der richtigen Reihenfolge an. Wenn Sie die vorhandene Sortierreihenfolge nicht richtig angeben, tritt beim Ausführen des Streams ein Fehler auf. In der Fehlernachricht wird dabei die Nummer des Datensatzes angezeigt, bei dem die Sortierung nicht mit Ihren Angaben übereinstimmt.

Abhängig davon, ob bei der von der Datenbank verwendeten Kollationsmethode die Groß- und Kleinschreibung berücksichtigt wird, funktioniert die Optimierung möglicherweise nicht ordnungsgemäß, wenn eine oder mehrere Eingaben von der Datenbank sortiert werden. Wenn Sie beispielsweise zwei Eingaben verwenden, wobei bei der einen zwischen Groß- und Kleinschreibung unterschieden wird und bei der anderen nicht, könnten die Ergebnisse der Sortierung voneinander abweichen. Die Zusammenführungsoptimierung führt dazu, dass die Datensätze gemäß ihrer sortierten Reihenfolge verarbeitet werden. Wenn die Eingaben mittels verschiedener Kollationsmethoden sortiert wurden, meldet der Zusammenführungsknoten daher einen Fehler und zeigt die Nummer des Datensatzes an, in dem die Sortierung inkonsistent ist. Wenn alle Eingaben aus derselben Quelle stammen oder mithilfe von sich gegenseitig einschließenden Kollationen sortiert wurden, können die Datensätze erfolgreich zusammengeführt werden.

Hinweis: Die Zusammenführung kann gegebenenfalls durch Parallelverarbeitung beschleunigt werden.

Anhangknoten

Mit Anhangknoten können Sie Sets von Datensätzen miteinander verketten. Anders als bei Zusammenführungsknoten ("Mergen"), in denen Datensätze aus verschiedenen Quellen miteinander verbunden werden, lesen Anhangknoten alle Datensätze aus einer Quelle und geben Sie nach unten im Stream weiter, bis keine mehr vorhanden sind. Anschließend werden die Datensätze aus der nächsten Quelle unter Verwendung derselben Datenstruktur (Anzahl der Datensätze, Anzahl der Felder usw.) wie bei der ersten Eingabe (Primäreingabe) gelesen. Wenn die Primärquelle mehr Felder aufweist als eine andere Eingabequelle, wird die systemdefinierte Nullzeichenfolge (`$null$`) für alle unvollständigen Werte verwendet.

Anhangknoten sind sinnvoll für die Kombination von Datasets mit ähnlicher Struktur, aber unterschiedlichen Daten. Sie könnten beispielsweise Transaktionsdaten für verschiedene Zeiträume in verschiedenen Dateien gespeichert haben (z. B. zwei Absatzdatendateien für März und April). Angenommen, diese Dateien weisen dieselbe Struktur (dieselben Felder in derselben Reihenfolge) auf, werden sie mit dem Anhangknoten in einer großen Datei zusammengefasst, die anschließend analysiert werden kann.

Hinweis: Zum Anhängen von Dateien sind ähnliche Feldmessniveaus erforderlich. Beispiel: Einem Feld des Typs *Nominal* kann kein Feld angehängt werden, dessen Messniveau *Stetig* ist.

Festlegen der Anhangoptionen

Feldübereinstimmung ermitteln nach. Dient zur Auswahl einer Methode für die Abgleichung der anzuhängenden Felder.

- **Position.** Wählen Sie diese Option aus, um Datasets auf der Grundlage der Position der Felder in der Hauptdatenquelle anzuhängen. Bei Verwendung dieser Methode sollten Ihre Daten sortiert sein, um einen ordnungsgemäßen Anhang zu gewährleisten.
- **Name.** Wählen Sie diese Option aus, um Datasets auf der Grundlage des Namens der Felder in den Eingabedatasets anzuhängen. Wählen Sie außerdem **Groß-/Kleinschreibung beachten**, wenn beim Abgleichen der Feldnamen die Groß- und Kleinschreibung berücksichtigt werden soll.

Ausgabefeld. Listet die Quellenknoten auf, die mit dem Anhangknoten verbunden sind. Der erste Knoten in der Liste ist die primäre Eingabequelle. Die Felder in der Anzeige können durch Klicken auf den Spaltentitel sortiert werden. Bei dieser Sortierung wird keine tatsächliche Umordnung der Felder im Dataset durchgeführt.

Felder einschließen aus. Wählen Sie die Option **Nur Hauptdatenquelle** aus, um Ausgabefelder auf der Grundlage der Felder in der Hauptdatenquelle zu erstellen. Die Hauptdatenquelle ist die erste Eingabe, die auf der Registerkarte "Eingaben" angegeben wurde. Wählen Sie **Alle Datasets** aus, um Ausgabefelder für alle Felder in allen Datasets zu erstellen, unabhängig davon, ob ein Feld vorhanden ist, das in allen Eingabedatasets übereinstimmt.

Datensätze durch Einschließen des Quellendatasets im Feld markieren. Wählen Sie diese Option aus, um ein zusätzliches Feld zur Ausgabedatei hinzuzufügen, dessen Werte das Quellendataset für die einzelnen Datensätze anzeigen. Geben Sie im Textfeld einen Namen an. Der Standardname des Felds lautet *Eingabe*.

Duplikatknoten

Doppelte Datensätze in einem Dataset müssen entfernt werden, bevor mit dem Data Mining begonnen werden kann. In einer Marketingdatenbank beispielsweise werden einzelne Personen möglicherweise mehrfach mit unterschiedlichen Adress- oder Firmendaten aufgeführt. Mit dem Duplikatknoten können Sie nach doppelten Datensätzen in Ihrem Dataset suchen und diese entfernen oder Sie können aus verschiedenen Datensätzen einen einzigen zusammengesetzten Datensatz erstellen.

Mit dem Duplikatknoten können Sie doppelte Datensätze entweder entfernen, indem jeweils der erste Datensatz an den Datenstream übergeben wird, oder aber nach doppelten Datensätzen suchen, indem der erste Datensatz verworfen wird und stattdessen etwaige Duplikate an den Stream übergeben werden.

Zusätzlich können Sie für jeden eindeutigen Schlüsselwert eine Sortierreihenfolge für die Ergebnisse festlegen. Wenn für jeden eindeutigen Schlüssel eine bestimmte Zeile angezeigt werden sollen, müssen Sie die Datensätze innerhalb des Duplikatknotens sortieren, anstatt einen weiter oben liegenden Sortierknoten zu verwenden (siehe "Sortieren von Datensätzen innerhalb des Duplikatknotens" weiter unten).

Modalwert. Geben Sie an, ob ein zusammengesetzter Datensatz erstellt werden soll oder ob der erste Datensatz aufgenommen oder ausgeschlossen (verworfen) werden soll.

- **Zusammengesetzten Datensatz für jede Gruppe erstellen.** Bietet Ihnen die Möglichkeit, nicht numerische Felder zu aggregieren. Wenn diese Option ausgewählt wird, steht die Registerkarte "Kombiniert" zur Verfügung, auf der Sie angeben können, wie die zusammengesetzten Datensätze erstellt werden sollen. Weitere Informationen finden Sie in „Einstellungen für unterschiedliche Zusammensetzung” auf Seite 88.
- **Nur jeweils den ersten Datensatz in jeder Gruppe aufnehmen.** Nimmt den jeweils ersten Datensatz in den Datenstream auf und entfernt alle Duplikate.
- **Nur jeweils den ersten Datensatz in jeder Gruppe verwerfen.** Verwirft jeweils den ersten gefundenen Datensatz und übergibt stattdessen etwaige doppelte Datensätze an den Datenstream. Mit dieser Option können Duplikate in den Daten *gefunden* werden, um sie später im Stream zu untersuchen.

Schlüsselfelder zur Gruppierung. Listet die Felder auf, die verwendet werden, um zu bestimmen, ob die Datensätze identisch sind. Sie verfügen über folgende Möglichkeiten:

- Zum Hinzufügen von Feldern zu dieser Liste verwenden Sie die Feldauswahlschaltfläche auf der rechten Seite.
- Zum Löschen von Feldern aus der Liste verwenden Sie die Schaltfläche mit dem roten X (Löschschaltfläche).

Datensätze innerhalb von Gruppen sortieren nach. Listet die Felder auf, die verwendet werden, um zu bestimmen, wie Datensätze innerhalb jedes eindeutigen Schlüsselwerts sortiert werden und ob sie in auf- oder absteigender Reihenfolge sortiert werden. Sie verfügen über folgende Möglichkeiten:

- Zum Hinzufügen von Feldern zu dieser Liste verwenden Sie die Feldauswahlschaltfläche auf der rechten Seite.
- Zum Löschen von Feldern aus der Liste verwenden Sie die Schaltfläche mit dem roten X (Löschschaltfläche).
- Verschieben Sie Felder mit den Schaltflächen "Nach oben" oder "Nach unten", wenn Sie nach mehr als einem Feld sortieren.

Standardsortierreihenfolge. Legen Sie fest, ob Datensätze standardmäßig **Aufsteigend** oder **Absteigend** sortiert werden sollen.

Sortieren von Datensätzen innerhalb des Duplikatknotens

Mithilfe der Option **Datensätze innerhalb von Gruppen sortieren nach** innerhalb des Duplikatknotens können Sie für jeden eindeutigen Schlüssel eine bestimmte Zeile anzeigen. Die Verwendung eines vorangehenden Sortierknotens ist nicht erforderlich. Nehmen wir zum Beispiel an, wir besitzen folgende Daten über das Alter von Menschen, die verschreibungspflichtige Medikamente einnehmen:

Tabelle 12. Daten der Anwender verschreibungspflichtiger Medikamente.

Alter	Medikament
50	Medikament A
71	Medikament B

Tabelle 12. Daten der Anwender verschreibungspflichtiger Medikamente (Forts.).

Alter	Medikament
44	Medikament A
65	Medikament X
39	Medikament A
75	Medikament C
72	Medikament Y
57	Medikament X
79	Medikament Y
69	Medikament C
74	Medikament B
85	Medikament Y
69	Medikament X

Um den ältesten Anwender eines jeden Medikaments zu ermitteln, würden wir den Modalwert auf "Nur jeweils den ersten Datensatz in jeder Gruppe aufnehmen" setzen, "Medikament" als Schlüsselfeld und "Alter" als Sortierfeld verwenden und eine absteigende Sortierreihenfolge auswählen. Die Reihenfolge der Eingaben hat keine Auswirkungen auf das Ergebnis, denn die Sortiereinstellungen legen fest, welche der Zeilen für ein bestimmtes Medikament angezeigt wird. Die endgültige Datenausgabe sähe also wie folgt aus:

Tabelle 13. Sortierte Daten der Benutzer verschreibungspflichtiger Medikamente.

Alter	Medikament
50	Medikament A
74	Medikament B
75	Medikament C
69	Medikament X
85	Medikament Y

Eindeutige Optimierungseinstellungen

Wenn die Daten, an denen Sie arbeiten, nur eine kleine Anzahl an Datensätzen umfassen oder bereits sortiert wurden, können Sie ihre Behandlung so optimieren, dass IBM SPSS Modeler die Daten effizienter verarbeitet.

Hinweis: Wenn Sie entweder **Der Eingabedatensatz weist eine geringe Anzahl unterschiedlicher Schlüssel auf** auswählen oder die SQL-Generierung für den Knoten verwenden, kann jede Zeile innerhalb des eindeutigen Schlüsselwerts angezeigt werden; um zu kontrollieren, welche Zeile innerhalb eines eindeutigen Schlüssels angezeigt wird, müssen Sie die Sortierreihenfolge mithilfe des Felds **Datensätze innerhalb von Gruppen sortieren nach** auf der Registerkarte "Einstellungen" festlegen. Die Optimierungsoptionen haben keine Auswirkungen auf die Ergebnisausgabe des Duplikatknotens, solange Sie auf der Registerkarte "Einstellungen" eine Sortierreihenfolge festgelegt haben.

Der Eingabedatensatz weist eine geringe Anzahl unterschiedlicher Schlüssel auf. Wählen Sie diese Option, wenn Sie über eine kleine Anzahl an Datensätzen oder eine kleine Anzahl an eindeutigen Werten der Schlüsselfelder oder beides verfügen. Dadurch lässt sich eventuell die Leistungsfähigkeit verbessern.

Eingabedataset ist bereits nach Gruppier- und Sortierfeldern auf der Registerkarte "Einstellungen" sortiert. Wählen Sie diese Option nur aus, wenn Ihre Daten bereits nach allen Feldern sortiert sind, die auf

der Registerkarte "Einstellungen" unter **Datensätze innerhalb von Gruppen sortieren nach** aufgelistet sind, und wenn die auf- und absteigende Sortierreihenfolge der Daten identisch ist. Dadurch lässt sich eventuell die Leistungsfähigkeit verbessern.

SQL-Erzeugung inaktivieren. Wählen Sie diese Option aus, um die SQL-Erzeugung für den Knoten zu inaktivieren.

Einstellungen für unterschiedliche Zusammensetzung

Wenn die Daten, an denen Sie arbeiten, aus mehreren Datensätzen bestehen, z. B. für dieselbe Person, können Sie die Art und Weise, wie die Daten verarbeitet werden, optimieren, indem Sie einen einzelnen zusammengesetzten Datensatz, oder Aggregatdatensatz, zur Verarbeitung erstellen. Wenn Sie IBM SPSS Modeler Entity Analytics installiert haben, können Sie damit auch doppelte Datensätze, die über SPSS Entity Analytics ausgegeben werden, kombinieren oder glätten.

Anmerkung: Diese Registerkarte ist nur verfügbar, wenn Sie auf der Registerkarte "Einstellungen" **Zusammengesetzten Datensatz für jede Gruppe erstellen** auswählen.

Nehmen wir beispielsweise an, dass SPSS Entity Analytics drei Datensätze als identische Entität markiert, wie in der folgenden Tabelle dargestellt.

Tabelle 14. Beispiel für mehrere Datensätze für dieselbe Entität.

$EA-ID	Name	Alter	Bank	Höchste Ausbildung	Gesamtschulden
0003	Jan Jäger	27	K	Abitur	27000
0003	Johannes Jäger	35	N	Diplom	42000
0003	Jannis Jäger	27	D	Dr. phil.	7000

Wir möchten diese drei Datensätze zu einem einzigen Datensatz aggregieren, den wir dann nachfolgend verwenden. Wir könnten den Aggregatknoten verwenden, um die Gesamtschulden zu addieren und ein Durchschnittsalter zu berechnen. Wir können jedoch nicht den Durchschnitt von Details wie Namen, Banken usw. ermitteln. Wenn wir angeben, welche Details verwendet werden sollen, um einen zusammengesetzten Datensatz zu erstellen, können wir einen einzelnen Datensatz ableiten.

Aus unserer Tabelle könnten wir einen zusammengesetzten Datensatz erstellen, indem wir die folgenden Details auswählen.

- Bei **Name** verwenden wir den ersten Datensatz.
- Bei **Alter** nehmen wir das höchste Alter.
- Bei **Bank** verketten wir alle Werte ohne Trennzeichen.
- Bei **Höchste Ausbildung** nehmen wir die erste Ausbildung, die wir in der Liste finden (Dr. phil., Diplom, Abitur)
- Bei **Schulden** nehmen wir die Gesamtsumme.

Durch Kombinieren (bzw. Aggregieren) dieser Details erhalten wir einen einzigen zusammengesetzten Datensatz, der die folgenden Details enthält.

- Name: Jan Jäger
- Alter: 35
- Bank: KND
- Höchste Ausbildung: Dr. phil.
- Schulden: 76000

Dadurch erhalten wir das beste Bild von Jan Jäger, einer gebildeten Person mit dem Titel Dr. phil., die mindestens 35 Jahre alt ist und mindestens drei bekannte Bankkonten sowie hohe Gesamtschulden hat.

Einstellungsoptionen für die Registerkarte "Kombiniert"

Feld. In dieser Spalte werden aller Felder, außer den Schlüsselfeldern im Datenmodell, in ihrer natürlichen Sortierreihenfolge angezeigt. Wenn keine Verbindung zum Knoten hergestellt ist, werden keine Felder angezeigt. Um die Zeilen alphabetisch nach Feldnamen zu sortieren, klicken Sie auf die Spaltenüberschrift. Sie können mehrere Zeilen auswählen, indem Sie bei gedrückter Umschalttaste oder bei gedrückter Steuertaste darauf klicken. Wenn Sie mit der rechten Maustaste auf ein Feld klicken, wird außerdem ein Menü mit folgenden Auswahlmöglichkeiten angezeigt: Alle Zeilen auswählen, Zeile nach aufsteigendem oder absteigendem Feldnamen oder Wert sortieren, Felder nach Maß oder Speichertyp auswählen oder einen Wert auswählen, um denselben Eintrag für **Füllen mit Werten auf der Basis von** automatisch jeder ausgewählten Zeile hinzuzufügen.

Füllen mit Werten auf der Basis von. Wählen Sie den Werttyp aus, der für den zusammengesetzten Datensatz für das **Feld** verwendet werden soll. Die verfügbaren Optionen hängen vom Feldtyp ab.

- Für numerische Bereichsfelder können Sie aus den folgenden Optionen auswählen:
  - Erster Datensatz in Gruppe
  - Letzter Datensatz in Gruppe
  - Gesamt
  - Mittelwert
  - Minimum
  - Maximum
  - Benutzerdefiniert
- Für Zeit- oder Datumsfelder können Sie aus den folgenden Optionen auswählen:
  - Erster Datensatz in Gruppe
  - Letzter Datensatz in Gruppe
  - Frühester
  - Letzter
  - Benutzerdefiniert
- Für Zeichenfolgefelder oder Felder ohne Typ können Sie aus den folgenden Optionen auswählen:
  - Erster Datensatz in Gruppe
  - Letzter Datensatz in Gruppe
  - Erster alphanumerisch
  - Letzter alphanumerisch
  - Benutzerdefiniert

In jedem Fall können Sie die Option **Benutzerdefiniert** verwenden, um eine bessere Kontrolle darüber auszuüben, welcher Wert zum Füllen des zusammengesetzten Datensatzes verwendet wird. Weitere Informationen finden Sie in „Unterschiedliche Zusammensetzung - Registerkarte "Benutzerdefiniert"".

Datensatzanzahl einschließen in Feld. Wählen Sie diese Option aus, um standardmäßig ein zusätzliches Feld mit der Bezeichnung "Datensatzanzahl" in jeden Ausgabedatensatz einzufügen. Dieses Feld gibt an, wie viele Eingabedatensätze aus den einzelnen Aggregatdatensätzen aggregiert wurden. Um einen benutzerdefinierten Namen für diesen Feldtyp zu erstellen, geben Sie Ihren Eintrag in das Bearbeitungsfeld ein.

Unterschiedliche Zusammensetzung - Registerkarte "Benutzerdefiniert"

Das Dialogfeld "Benutzerdefinierte Füllung" bietet Ihnen eine bessere Kontrolle darüber, welcher Wert zum Vervollständigen des neuen zusammengesetzten Datensatzes verwendet wird. Wenn Sie nur eine einzelne Feldzeile auf der Registerkarte "Kombiniert" anpassen, beachten Sie, dass Sie vor der Verwendung dieser Option Ihre Daten zuerst instanziieren müssen.

Anmerkung: Dieses Dialogfeld ist nur verfügbar, wenn Sie den Wert "Benutzerdefiniert" in der Spalte **Füllen mit Werten auf der Basis von** auf der Registerkarte "Kombiniert" auswählen.

Je nach Feldtyp können Sie eine der folgenden Optionen auswählen.

- **Nach Häufigkeit auswählen.** Wählen Sie einen Wert auf der Basis der Häufigkeit aus, mit der er im Datensatz vorkommt.

  Anmerkung: Nicht verfügbar für Felder mit dem Typ "Stetig", "Ohne Typ" oder "Datum/Uhrzeit".
  - **Verwenden.** Wählen Sie entweder "Am häufigsten" oder "Am seltensten" aus.
  - **Bindungen.** Wenn es mindestens zwei Datensätze gibt, die mit der gleichen Häufigkeit vorkommen, geben Sie an, wie der erforderliche Datensatz ausgewählt werden soll. Sie können eine der vier folgenden Optionen auswählen: "Erste verwenden", "Letzte verwenden", "Niedrigste verwenden" oder "Höchste verwenden".
- **Schließt Wert ein (T/F).** Wählen Sie diese Option aus, um ein Feld in ein Flag zu konvertieren, das angibt, ob mindestens einer der Datensätze in einer Gruppe einen angegebenen Wert aufweist. Sie können dann den **Wert** aus der Liste der Werte für das ausgewählte Feld auswählen.

  Anmerkung: Nicht verfügbar, wenn auf der Registerkarte "Kombiniert" mehr als eine Zeile "Feld" ausgewählt wird.
- **Erste Übereinstimmung in Liste.** Wählen Sie diese Option aus, um anzugeben, welcher Wert für den zusammengesetzten Datensatz Priorität hat. Sie können dann eines der **Elemente** aus der Liste der Elemente für das ausgewählte Feld auswählen.

  Anmerkung: Nicht verfügbar, wenn auf der Registerkarte "Kombiniert" mehr als eine Zeile "Feld" ausgewählt wird.
- **>Werte verketten.** Wählen Sie diese Option aus, um alle Werte in einer Gruppe beizubehalten, indem Sie zu einer Zeichenfolge verkettet werden. Sie müssen ein Trennzeichen angeben, das zwischen den einzelnen Werten verwendet werden soll.

  Anmerkung: Dies ist die einzige verfügbare Option, wenn Sie mindestens eine Feldzeile mit dem Typ "Stetig", "Ohne Typ" oder "Datum/Uhrzeit" auswählen.
- **Trennzeichen verwenden.** Sie können auswählen, ob Sie ein **Leerzeichen** oder ein **Komma** als Trennzeichenwert in der verketteten Zeichenfolge verwenden möchten. Sie können auch im Feld **Andere** selbst ein Zeichen eingeben, das als Trennzeichenwert verwendet werden soll.

  Anmerkung: Nur verfügbar, wenn Sie die Option **Werte verketten** auswählen.

Streaming-ZR-Knoten

Mit dem Streaming-ZR-Knoten (ZR - Zeitreihe) können Zeitreihenmodelle in einem Schritt erstellt und gescort werden und dieser Knoten kann mit oder ohne vorgeordneten Zeitintervallknoten verwendet werden. Für jedes Zielfeld wird ein separates Zeitreihenmodell erstellt, der generierten Modellpalette werden jedoch keine Modellnuggets hinzugefügt und die Modellinformationen können nicht durchsucht werden.

Es ist nicht immer erforderlich, einen Zeitintervallknoten mit dem Streaming-ZR-Knoten zu verwenden, wenn die folgenden Bedingungen erfüllt sind:

- Die Daten weisen gleichmäßige Abstände auf.
- Die Daten weisen Zeitreihenbeschriftungen auf.
- Es müssen keine zusätzlichen Datensätze hinzugefügt werden.
- Die Daten weisen eine *Standardperiodizität* auf, z. B. fünf Tage pro Woche oder acht Stunden pro Tag.

Bei den Methoden zur Modellierung von Zeitreihendaten ist ein einheitliches Intervall zwischen den Messungen erforderlich; fehlende Werte werden durch leere Zeilen dargestellt. Falls Ihre Daten diese Anforderung nicht bereits erfüllen, können Sie Werte mithilfe eines vorgeordneten Zeitintervallknotens entsprechend transformieren. Weitere Informationen finden Sie im Thema „Zeitintervallknoten" auf Seite 159.

Außerdem ist bei Zeitreihendaten zu beachten:

- Die Felder müssen numerisch sein.
- Datumsfelder können nicht als Eingaben verwendet werden.
- Partitionen werden ignoriert.

Der Streaming-ZR-Knoten berechnet Schätzungen für exponentielle Glättung, univariate ARIMA-Modelle (Autoregressive Integrated Moving Average - autoregressiver integrierter gleitender Durchschnitt) und multivariate ARIMA-Modelle (Transferfunktionsmodelle) für Zeitreihendaten und erstellt Vorhersagen auf der Grundlage der Zeitreihendaten. Außerdem ist ein Expert Modeler verfügbar, der automatisch das am besten angepasste ARIMA-Modell bzw. das am besten angepasste Modell mit exponentiellem Glätten für mindestens ein Zielfeld ermittelt. Weitere Informationen zu Zeitreihenmodellen im Allgemeinen finden Sie im Thema *Zeitreihenmodelle* im Dokument *IBM SPSS Modeler 16 Modellierungsknoten*.

Der Streaming-ZR-Knoten wird für die Verwendung in einer Streaming-Bereitstellungsumgebung über IBM SPSS Modeler Solution Publisher mit dem Scoring-Service für IBM SPSS Collaboration and Deployment Services oder IBM InfoSphere Warehouse unterstützt.

Streaming-ZR-Knoten - Feldoptionen

Auf der Registerkarte "Felder" geben Sie an, welche Felder bei der Erstellung des Modells verwendet werden sollen. Bevor Sie ein Modell erstellen können, müssen Sie festlegen, welche Felder als Ziele und als Eingaben verwendet werden sollen. Normalerweise verwendet der Streaming-ZR-Knoten Feldinformationen aus einem vorgeordneten Typknoten. Wenn Sie einen Typknoten verwenden, um Eingabe- und Zielfelder auszuwählen, brauchen Sie auf dieser Registerkarte keine Änderungen vorzunehmen.

Typknoteneinstellungen verwenden. Diese Option weist den Knoten an, die Feldinformationen von einem weiter oben liegenden Typknoten zu verwenden. Dies ist die Standardeinstellung.

Benutzerdefinierte Einstellungen verwenden. Diese Option weist den Knoten an, die hier angegebenen Feldinformationen anstelle der in einem weiter oben liegenden Typknoten angegebenen zu verwenden. Geben Sie nach Auswahl dieser Option die unten stehenden Felder an. Beachten Sie, dass als Datumswerte gespeicherte Felder nicht als Ziel- oder Eingabefelder zulässig sind.

- **Ziele.** Wählen Sie ein oder mehrere Zielfelder aus. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Ziel* festlegen. Zielfelder für Zeitreihenmodelle müssen ein Messniveau des Typs *Stetig* aufweisen. Für jedes Zielfeld wird ein separates Modell erstellt. Für Zielfelder kommen alle angegebenen *Eingabe*-Felder mit Ausnahme des jeweiligen Zielfelds selbst als mögliche Eingaben in Betracht. Daher kann dasselbe Feld in beiden Listen vorkommen; ein solches Feld wird als mögliche Eingabe für alle Modelle verwendet, außer für das Modell, bei dem es ein Zielfeld ist.
- **Eingaben.** Wählen Sie die Eingabefelder aus. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Eingabe* festlegen. Eingabefelder für Zeitreihenmodelle müssen numerisch sein.

Streaming-ZR-Knoten - Modelloptionen

Methode. Sie haben die Wahl zwischen Expert Modeler, exponentiellem Glätten und ARIMA. Wählen Sie **Kriterien** aus, um Optionen für die ausgewählte Methode anzugeben.

- **Expert Modeler.** Wählen Sie diese Option aus, um den Expert Modeler zu verwenden, der automatisch das jeweils am besten angepasste Modell für die einzelnen Zeitreihen ermittelt.
- **Exponentielles Glätten.** Mit dieser Option können Sie ein benutzerdefiniertes Modell mit exponentiellem Glätten angeben.

- **ARIMA.** Mit dieser Option können Sie ein ARIMA-Modell angeben.

Felder für obere und untere Konfidenzgrenzen ableiten. Wählen Sie diese Option aus, um Felder für Konfidenzintervalle für die Modellvorhersagen und Residuenautokorrelationen zu generieren. Wenn diese Option ausgewählt ist, ist das Steuerelement **Breite der Konfidenzgrenze (%)** aktiviert.

Breite der Konfidenzgrenze (%). Geben Sie die Konfidenzintervalle an, die berechnet werden. Es kann ein beliebiger positiver Wert unter 100 angegeben werden. In der Standardeinstellung wird ein Konfidenzintervall von 95 % verwendet.

Zeitintervalloptionen

In diesem Abschnitt des Dialogfelds können Sie auswählen, ob Sie Spezifikationen für Schätzungen und Vorhersagen aus einem vorgeordneten Zeitintervallknoten verwenden oder diese Einstellungen für den Streaming-ZR-Knoten angeben möchten.

Einstellungen aus Zeitintervallknoten verwenden. Wählen Sie diese Option aus, um Informationen zu Spezifikationen für Schätzungen und Vorhersagen zu verwenden, die an einem vorgeordneten Zeitintervallknoten vorgenommen werden.

Die erste Zeile der Informationen gibt an, ob Datensätze aus dem Modell ausgeschlossen oder als Holdouts verwendet werden.

Die zweite Zeile bietet Informationen zu den im Zeitintervallknoten angegebenen Vorhersageperioden.

Wenn in der ersten Zeile **Kein Zeitintervall definiert** steht, bedeutet dies, dass kein Zeitintervallknoten eingebunden ist. Dies führt zu einem Fehler beim Versuch, den Stream auszuführen. Wenn Sie **Einstellungen aus Zeitintervallknoten verwenden** auswählen, müssen Sie einen Zeitintervallknoten vor dem Streaming-ZR-Knoten einfügen.

Einstellungen angeben. Wählen Sie diese Option aus, um die Einstellungen für Schätzungen und Vorhersagen anzugeben, wenn vor dem Streaming-ZR-Knoten kein Zeitintervallknoten vorhanden ist.

Schätzen bis. Wählen Sie **Letzte gültige Beobachtung** aus, um das Modell anhand aller Datensätze bis zur letzten gültigen Beobachtung zu schätzen. Die Datensätze, die für die Vorhersage verwendet werden, beginnen bei der ersten Beobachtung, die nicht gültig ist. *Gültige* Beobachtungen sind die Beobachtungen, bei denen alle Zielfelder Werte ungleich null aufweisen. Wählen Sie **Vom letzten Datensatz rückwärts zählen** aus, um mehrere der aktuellsten Datensätze für die Vorhersage zu reservieren. Die aktuellsten Datensätze weisen möglicherweise Werte für die Zielfelder auf, dies ist jedoch nicht erforderlich. Wenn **Vom letzten Datensatz rückwärts zählen** ausgewählt ist, ist das Steuerelement **Offset** aktiviert, bei dem Sie die Anzahl der Datensätze angeben können, die nur für die Vorhersage verwendet werden sollen. Der Standardwert für die Anzahl der Datensätze beträgt 4 und das Maximum beträgt 99.

Anmerkung: Die Datensätze, die für die Vorhersage verwendet werden sollen, müssen Werte für alle Eingabefelder aufweisen.

Zeitintervall. Geben Sie das Zeitintervall an, das für die Schätzung und die Vorhersage verwendet wird. Die folgenden Optionen sind verfügbar: "Jahre", "Quartale", "Monate", "Wochen", "Tage", "Stunden", "Minuten" und "Sekunden".

Zeitreihen - Expert Modeler-Kriterien

Modelltyp. Die folgenden Optionen sind verfügbar:

- **Alle Modelle.** Der Expert Modeler berücksichtigt sowohl ARIMA-Modelle als auch Modelle mit exponentiellem Glätten.

- **Nur Modelle mit exponentiellem Glätten.** Der Expert Modeler berücksichtigt nur Modelle mit exponentiellem Glätten.
- **Nur ARIMA-Modelle.** Der Expert Modeler berücksichtigt nur ARIMA-Modelle.

Expert Modeler berücksichtigt saisonale Modelle. Diese Option ist nur verfügbar, wenn für die Arbeitsdatei eine Periodizität definiert wurde. Wenn diese Option aktiviert ist, berücksichtigt der Expert Modeler sowohl saisonale als auch nicht saisonale Modelle. Wenn diese Option inaktiviert ist, berücksichtigt der Expert Modeler nur nicht saisonale Modelle.

Ereignisse und Interventionen. Mit dieser Option können Sie bestimmte Eingabefelder als Ereignis- bzw. Interventionsfelder kennzeichnen. Dadurch wird angegeben, dass das betreffende Feld Zeitreihendaten enthält, die von Ereignissen (vorhersagbare wiederkehrende Situationen, z. B. Werbeaktionen) oder Interventionen (einmalige Vorfälle, z. B. Stromausfall, Streik) betroffen sind. Expert Modeler berücksichtigt keine frei wählbaren Transferfunktionen für Eingaben, die als Ereignis- bzw. Interventionsfelder gekennzeichnet sind.

Eingabefelder müssen das Messniveau *Flag*, *Nominal* oder *Ordinal* aufweisen und müssen numerisch sein (z. B. "1"/"0" und nicht "Wahr"/"Falsch" für ein Flagfeld), um in dieser Liste angezeigt zu werden.

Ausreißer

Ausreißer automatisch erkennen. In der Standardeinstellung wird keine automatische Erkennung von Ausreißern durchgeführt. Wählen Sie diese Option aus, um die automatische Erkennung von Ausreißern durchzuführen, und wählen Sie anschließend die gewünschten Ausreißertypen aus. Weitere Informationen finden Sie im Thema „Umgang mit Ausreißern” auf Seite 96.

Zeitreihen - Kriterien für exponentielles Glätten

Modelltyp. Modelle für das exponentielle Glätten werden als saisonal oder nicht saisonal klassifiziert[1]. Saisonale Modelle sind nur verfügbar, wenn die Periodizität saisonal ist. Es gibt folgende saisonale Periodizitäten: zyklische Perioden, Jahre, Quartale, Monate, Tage pro Woche, Stunden pro Tag, Minuten pro Tag und Sekunden pro Tag.

- **Einfach.** Dieses Modell eignet sich für Zeitreihen ohne Trend oder Saisonalität. Der einzige relevante Glättungsparameter für dieses Modell ist das Niveau. Einfaches exponentielles Glätten weist die größte Ähnlichkeit mit einem ARIMA-Modell mit Autoregression der Ordnung null, Differenzenbildung der Ordnung eins, gleitendem Durchschnitt der Ordnung eins und ohne Konstante auf.
- **Linearer Trend nach Holt.** Dieses Modell eignet sich für Zeitreihen mit linearem Trend und ohne Saisonalität. Die dafür relevanten Glättungsparameter sind Niveau und Trend, die bei diesem Modell nicht durch die Werte des jeweils anderen Parameters eingeschränkt sind. Das Holt-Modell ist allgemeiner als das Brown-Modell, die Berechnung von Schätzungen für große Zeitreihen kann allerdings mehr Zeit in Anspruch nehmen. Das exponentielle Glätten nach Holt weist die größte Ähnlichkeit mit einem ARIMA-Modell mit Autoregression der Ordnung null, Differenzenbildung der Ordnung zwei und gleitendem Durchschnitt der Ordnung zwei auf.
- **Linearer Trend nach Brown.** Dieses Modell eignet sich für Zeitreihen mit linearem Trend und ohne Saisonalität. Die dafür relevanten Glättungsparameter sind Niveau und Trend. Bei diesem Modell wird jedoch davon ausgegangen, dass diese gleich sind. Das Brown-Modell ist daher ein Spezialfall des Holt-Modells. Das exponentielle Glätten nach Brown weist die größte Ähnlichkeit mit einem ARIMA-Modell mit Autoregression der Ordnung null, Differenzenbildung der Ordnung zwei und gleitendem Durchschnitt der Ordnung zwei auf, wobei der Koeffizient der zweiten Ordnung des gleitenden Durchschnitts die Hälfte des quadrierten Koeffizienten für die erste Ordnung beträgt.
- **Gedämpfter Trend.** Dieses Modell eignet sich für Zeitreihen mit auslaufendem linearem Trend und ohne Saisonalität. Die dafür relevanten Glättungsparameter sind Niveau, Trend und gedämpfter Trend.

1. Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting*, 4, 1-28.

Gedämpftes exponentielles Glätten weist die größte Ähnlichkeit mit einem ARIMA-Modell mit Autoregression der Ordnung eins, Differenzenbildung der Ordnung eins und gleitendem Durchschnitt der Ordnung zwei auf.

- **Einfach saisonal.** Dieses Modell eignet sich für Zeitreihen ohne Trend und mit einem saisonalen Effekt, der im Zeitverlauf konstant bleibt. Die dafür relevanten Glättungsparameter sind Niveau und Saison. Saisonales exponentielles Glätten weist die größte Ähnlichkeit mit einem ARIMA-Modell mit Autoregression der Ordnung null, Differenzenbildung der Ordnung eins, saisonaler Differenzenbildung der Ordnung eins und den Ordnungen 1, p und p+1 für den gleitenden Durchschnitt auf, wobei p die Anzahl der Perioden in einem saisonalen Intervall ist. Für Monatsdaten gilt: $p = 12$.
- **Additives Winters-Modell.** Dieses Modell eignet sich für Zeitreihen mit linearem Trend und mit einem saisonalen Effekt, der im Zeitverlauf konstant bleibt. Die dafür relevanten Glättungsparameter sind Niveau, Trend und Saison. Das exponentielle Glätten nach dem additiven Winters-Modell weist die größte Ähnlichkeit mit einem ARIMA-Modell mit Autoregression der Ordnung null, Differenzenbildung der Ordnung eins, saisonaler Differenzenbildung der Ordnung eins und p+1 Ordnungen für den gleitenden Durchschnitt auf, wobei p die Anzahl der Perioden in einem saisonalen Intervall ist. Für Monatsdaten gilt: $p = 12$.
- **Multiplikatives Winters-Modell.** Dieses Modell eignet sich für Zeitreihen mit linearem Trend und mit einem saisonalen Effekt, der sich mit der Größenordnung der Zeitreihe ändert. Die dafür relevanten Glättungsparameter sind Niveau, Trend und Saison. Exponentielles Glätten mit dem multiplikativen Winters-Modell weist keine Ähnlichkeit zu irgendeinem ARIMA-Modell auf.

Zieltransformation. Sie können für jede abhängige Variable eine Transformation angeben, die vor deren Modellierung durchgeführt werden soll. Weitere Informationen finden Sie im Thema *Reihentransformationen* des Dokuments *IBM SPSS Modeler 16 Modellierungsknoten*.

- **Keine.** Es wird keine Transformation durchgeführt.
- **Quadratwurzel.** Quadratwurzeltransformation wird ausgeführt.
- **Natürlicher Logarithmus.** Transformation mit natürlichem Logarithmus wird ausgeführt.

Zeitreihen - ARIMA-Kriterien

Mit dem Zeitreihenknoten können Sie benutzerdefinierte nicht saisonale oder saisonale ARIMA-Modelle - auch als Box-Jenkins-Modelle bekannt - mit oder ohne festes Set von Eingabevariablen (Prädiktorvariablen) erstellen[2]. Sie können Transferfunktionen für bestimmte oder alle Eingabevariablen definieren und die automatische Erkennung von Ausreißern oder einer bestimmten Gruppe von Ausreißern festlegen.

Alle angegebenen Eingabevariablen werden explizit in das Modell aufgenommen. Im Gegensatz dazu werden beim Expert Modeler Eingabevariablen nur aufgenommen, wenn sie eine statistisch signifikante Beziehung zu der Zielvariablen aufweisen.

Modell

Über die Registerkarte "Modelle" können Sie die Struktur eines benutzerdefinierten ARIMA-Modells festlegen.

ARIMA-Ordnungen. Geben Sie Werte für die verschiedenen ARIMA-Komponenten des Modells in die entsprechenden Zellen des Strukturrasters ein. Alle Werte müssen nicht negative ganze Zahlen sein. Bei autoregressiven Komponenten und Komponenten des gleitenden Durchschnitts stellt der Wert die höchste Ordnung dar. Alle positiven niedrigeren Ordnungen werden in das Modell eingeschlossen. Wenn Sie beispielsweise 2 angeben, enthält das Modell die Ordnungen 2 und 1. Die Zellen in der Spalte "Saisonal" sind nur verfügbar, wenn für die Arbeitsdatei eine Periodizität definiert wurde.

2. Box, G. E. P., G. M. Jenkins, and G. C. Reinsel. 1994. *Time series analysis: Forecasting and control*, 3rd ed. Englewood Cliffs, N.J.: Prentice Hall.

- **Autoregressiv (p).** Die Anzahl autoregressiver Ordnungen im Modell. Autoregressive Ordnungen geben die zurückliegenden Werte der Zeitreihe an, die für die Vorhersage der aktuellen Werte verwendet werden. Eine autoregressive Ordnung von 2 gibt beispielsweise an, dass die Werte der Zeitreihe, die zwei Zeitperioden zurückliegt, für die Vorhersage der aktuellen Werte verwendet wird.
- **Differenz (d).** Gibt die Ordnung der Differenzierung an, die vor dem Schätzen der Modelle auf die Zeitreihe angewendet wurde. Differenzierung ist erforderlich, wenn Trends vorhanden sind. (Zeitreihen mit Trends sind normalerweise nicht stationär, und bei der ARIMA-Modellierung wird Stationarität angenommen.) Mithilfe der Differenzierung werden die Effekte der Trends entfernt. Die Ordnung der Differenzierung entspricht dem Grad des Trends der Zeitreihe: Differenzierung erster Ordnung erklärt lineare Trends, Differenzierung zweiter Ordnung erklärt quadratische Trends usw.
- **Gleitender Durchschnitt (q).** Die Anzahl von Ordnungen des gleitenden Durchschnitts im Modell. Ordnungen des gleitenden Durchschnitts geben an, wie Abweichungen vom Mittelwert der Zeitreihe für zurückliegende Werte zum Vorhersagen der aktuellen Werte verwendet werden. Ordnungen des gleitenden Durchschnitts von 1 und 2 geben beispielsweise an, dass beim Vorhersagen der aktuellen Werte der Zeitreihe Abweichungen vom Mittelwert der Zeitreihe von den beiden letzten Zeitperioden berücksichtigt werden sollen.

Saisonale Ordnungen. Saisonale autoregressive Komponenten, Komponenten des gleitenden Durchschnitts und Differenzierungskomponenten entsprechen im Prinzip ihren nicht saisonalen Gegenstücken. Bei saisonalen Ordnungen werden die Werte der aktuellen Zeitreihe jedoch von Werten zurückliegender Zeitreihen beeinflusst, die um eine oder mehrere saisonalen Perioden getrennt sind. Bei monatlichen Daten (saisonale Periode von 12) beispielsweise bedeutet eine saisonale Ordnung von 1, dass der Wert der aktuellen Zeitreihe durch den Zeitreihenwert beeinflusst wird, der 12 Perioden vor dem aktuellen liegt. Eine saisonale Ordnung von 1 entspricht bei monatlichen Daten einer nicht saisonalen Ordnung von 12.

Zieltransformation. Sie können für jede Zielvariable eine Transformation angeben, die vor deren Modellierung durchgeführt werden soll. Weitere Informationen finden Sie im Thema *Reihentransformationen* des Dokuments *IBM SPSS Modeler 16 Modellierungsknoten*.

- **Keine.** Es wird keine Transformation durchgeführt.
- **Quadratwurzel.** Quadratwurzeltransformation wird ausgeführt.
- **Natürlicher Logarithmus.** Transformation mit natürlichem Logarithmus wird ausgeführt.

Konstante in Modell einschließen. Der Einschluss einer Konstanten ist das Standardverfahren, sofern Sie nicht sicher wissen, dass der Gesamtmittelwert der Zeitreihe 0 ist. Bei der Anwendung von Differenzierung empfiehlt es sich, die Konstante auszuschließen.

Transferfunktionen

Auf der Registerkarte "Transferfunktionen" können Sie Transferfunktionen für einige oder alle Eingabefelder definieren. Mithilfe von Transferfunktionen können Sie angeben, auf welche Weise frühere Werte der betreffenden Felder für die Vorhersage zukünftiger Werte der Ziel-Zeitreihe verwendet werden sollen.

Die Registerkarte wird nur angezeigt, wenn Eingabefelder (bei denen die Rolle auf *Eingabe* gesetzt ist) entweder im Typknoten oder auf der Registerkarte "Felder" des Streaming-ZR-Knotens angegeben sind (wählen Sie **Benutzerdefinierte Einstellungen verwenden - Eingaben** aus).

In der Liste oben werden alle Eingabefelder angezeigt. Die übrigen Informationen in diesem Dialogfeld hängen davon ab, welches Eingabefeld in der Liste ausgewählt wurde.

Transferfunktionsordnungen. Geben Sie Werte für die verschiedenen Komponenten der Transferfunktion in die entsprechenden Zellen des Strukturrasters ein. Alle Werte müssen nicht negative Ganzzahlen sein. Bei Zähler- und Nennerkomponenten stellt der Wert die höchste Ordnung dar. Alle positiven niedrigeren Ordnungen werden in das Modell eingeschlossen. Darüber hinaus wird die Ordnung 0 bei Zählerkomponenten immer eingeschlossen. Wenn Sie beispielsweise 2 für den Zähler angeben, enthält das Modell die

Ordnungen 2, 1 und 0. Wenn Sie 3 für den Nenner angeben, enthält das Modell die Ordnungen 3, 2 und 1. Die Zellen in der Spalte "Saisonal" sind nur verfügbar, wenn für die Arbeitsdatei eine Periodizität definiert wurde.

Zähler. Die Zählerordnung der Transferfunktion gibt an, welche zurückliegenden Werte aus der ausgewählten unabhängigen Zeitreihe (Prädiktor-Zeitreihe) zum Vorhersagen der aktuellen Werte der abhängigen Zeitreihe verwendet werden. Ein Zähler-Term von 1 gibt beispielsweise an, dass der Wert einer unabhängigen Zeitreihe, die eine Periode zurückliegt, und der aktuelle Wert der unabhängigen Zeitreihe zum Vorhersagen des aktuellen Werts der einzelnen abhängigen Zeitreihen verwendet werden.

Nenner. Die Nennerordnung der Transferfunktion gibt an, wie Abweichungen vom Mittelwert der Zeitreihe für zurückliegende Werte der ausgewählten unabhängigen Zeitreihe (Prädiktor-Zeitreihe) zum Vorhersagen der aktuellen Werte der abhängigen Zeitreihe verwendet werden. Ein Nenner-Term von 1 gibt beispielsweise an, dass beim Vorhersagen der aktuellen Werte für die einzelnen abhängigen Zeitreihen Abweichungen vom Mittelwert einer unabhängigen Zeitreihe berücksichtigt werden sollen, die eine Zeitperiode zurückliegt.

Differenz. Gibt die Ordnung der Differenzierung an, die vor dem Schätzen der Modelle auf die ausgewählte unabhängige Zeitreihe (Prädiktoren) angewendet wurde. Wenn Trends vorhanden sind, ist die Differenzierung erforderlich, um die Effekte der Trends zu entfernen.

Saisonale Ordnungen. Saisonale Zähler-, Nenner- und Differenzierungskomponenten entsprechen im Prinzip ihren nicht saisonalen Gegenstücken. Bei saisonalen Ordnungen werden die Werte der aktuellen Zeitreihe jedoch von Werten zurückliegender Zeitreihen beeinflusst, die um eine oder mehrere saisonalen Perioden getrennt sind. Bei monatlichen Daten (saisonale Periode von 12) beispielsweise bedeutet eine saisonale Ordnung von 1, dass der Wert der aktuellen Zeitreihe durch den Zeitreihenwert beeinflusst wird, der 12 Perioden vor dem aktuellen liegt. Eine saisonale Ordnung von 1 entspricht bei monatlichen Daten einer nicht saisonalen Ordnung von 12.

Verzögerung Wenn eine Verzögerung festgelegt wird, verzögert sich der Einfluss des Eingabefelds um die Anzahl der angegebenen Intervalle. Bei einer Verzögerung mit dem Wert 5 beeinflusst der Wert des Eingabefelds zum Zeitpunkt t die Vorhersagen erst nach dem Ablauf von fünf Perioden ($t + 5$).

Transformation. Die Angabe einer Transferfunktion für ein Set von unabhängigen Variablen enthält auch eine optionale Transformation, die für diese Variablen ausgeführt werden soll.

- **Keine.** Es wird keine Transformation durchgeführt.
- **Quadratwurzel.** Quadratwurzeltransformation wird ausgeführt.
- **Natürlicher Logarithmus.** Transformation mit natürlichem Logarithmus wird ausgeführt.

Umgang mit Ausreißern

Auf der Registerkarte "Ausreißer" ist eine Reihe von Möglichkeiten für die Behandlung von Ausreißern in den Daten verfügbar[3].

Ausreißer nicht erkennen oder modellieren. In der Standardeinstellung werden Ausreißer weder erkannt noch modelliert. Wählen Sie diese Option aus, um die Erkennung und Modellierung von Ausreißern zu inaktivieren.

Ausreißer automatisch erkennen. Wählen Sie diese Option aus, um eine automatische Erkennung von Ausreißern durchzuführen, und wählen Sie mindestens einen der gezeigten Ausreißertypen aus.

Typ der zu ermittelnden Ausreißer. Wählen Sie die Ausreißertypen aus, die erkannt werden sollen. Folgende Typen werden unterstützt:

3. Pena, D., G. C. Tiao, and R. S. Tsay, eds. 2001. *A course in time series analysis*. New York: John Wiley and Sons.

- Additiv (Standard)
- Verschiebung im Niveau (Standard)
- Innovativ
- Transient
- Saisonal additiv
- Lokaler Trend
- Additiver Bereich

Weitere Informationen finden Sie im Thema *Ausreißer* des Dokuments *IBM SPSS Modeler 16 Modellierungsknoten*.

Streaming-ZR-Knoten - Bereitstellungsoptionen

Wenn Sie den Streaming-ZR-Knoten in einer Streaming-Bereitstellungsumgebung über IBM SPSS Modeler Solution Publisher verwenden, wird das Modell bei jedem Scoren von Datensätzen mithilfe eines Zeitreihenmodells standardmäßig neu erstellt und es werden neue Vorhersagen gemacht. Sie können dieses Standardverhalten mithilfe der Optionen auf der Registerkarte "Bereitstellung" ändern, um anzugeben, wann die Zeitreihenmodelle neu erstellt werden.

Anmerkung: Die Einstellungen auf der Registerkarte "Bereitstellung" werden bei der normalen Verwendung des Streaming-ZR-Knotens in IBM SPSS Modeler ignoriert.

Vollständige Neuerstellung bei Bedarf erzwingen (ändert möglicherweise den Modelltyp. Wählen Sie diese Option aus, wenn die Zeitreihenmodelle nur neu erstellt werden sollen, wenn eine angegebene Anzahl neuer Datensätze vorhanden ist. Wenn das Modell für jeden dritten oder vierten Datensatz und nicht für jeden neuen Datensatz neu erstellt wird, kann dadurch die Leistung verbessert werden. Bei der Auswahl dieser Option werden die Steuerelemente unter **Minimale Anzahl neuer Datensätze erforderlich für Neuerstellung** aktiviert.

>Minimale Anzahl neuer Datensätze erforderlich für Neuerstellung. Sie können auswählen, wann die Zeitreihenmodelle neu erstellt werden (auf der Basis der Anzahl neuer Datensätze). Wählen Sie eine der folgenden Optionen aus:

- Wenn **Prozentsatz der Datensätze verwenden (%)** ausgewählt ist, werden die Zeitreihenmodelle neu erstellt, wenn ein angegebener Prozentsatz der Gesamtzahl der Datensätze neue Datensätze sind. Der Standardprozentsatz beträgt 10 %.
- Wenn **Absolute Anzahl der Datensätze verwenden** ausgewählt ist, wird das Zeitreihenmodell neu erstellt, wenn eine angegebene Anzahl von Datensätzen neue Datensätze sind. Die minimale Anzahl neuer Datensätze, die für das Modell angegeben werden kann, ist 1 und das Maximum ist 100. Der Standardwert für die Anzahl der Datensätze beträgt 4.

Für Angabe neuer Datensätze zu verwendendes Feld. Wählen Sie das Feld aus, das beobachtet wird, um zu erkennen, wann neue Datensätze vorhanden sind. Das Feld, das ausgewählt wird, muss das Feld sein, das die Datensätze in der Zeitreihe ordnet. Es können nur stetige Felder ausgewählt werden.

STB-Knoten

Space-Time-Boxes (STB) sind eine Erweiterung der räumlichen Positionen in einer Geohashtabelle. Genauer ist eine STB eine alphanumerische Zeichenfolge, die einen regelmäßig geformten Bereich von Raum und Zeit darstellt.

Die STB dr5ru7|2013-01-01 00:00:00|2013-01-01 00:15:00 besteht beispielsweise aus den folgenden drei Teilen:

- Geohash - dr5ru7
- Startzeitmarke - 2013-01-01 00:00:00

- Endzeitmarke - 2013-01-01 00:15:00

Sie könnten zum Beispiel Informationen zu Raum und Zeit verwenden, um die Konfidenz zu verbessern, dass zwei Entitäten identisch sind, da sie sich praktisch zur selben Zeit am selben Ort befinden. Sie könnten auch die Genauigkeit der Beziehungsidentifikation verbessern, indem Sie zeigen würden, dass zwei Entitäten aufgrund ihrer räumlichen und zeitlichen Nähe zusammengehören.

Ihren Anforderungen entsprechend können Sie den Modus **Einzelne Datensätze** oder **Aufenthalte** auswählen. Beide Modi erfordern dieselben grundlegenden Details, die im Folgenden aufgeführt sind:

Breitengradfeld. Wählen Sie das Feld aus, das den Breitengrad angibt.

Längengradfeld. Wählen Sie das Feld aus, das den Längengrad angibt.

Zeitmarkenfeld. Wählen Sie das Feld aus, das die Uhrzeit oder das Datum angibt.

Optionen für einzelne Datensätze

Mit diesem Modus können Sie einem Datensatz ein zusätzliches Feld hinzufügen, um seine Position zu einer bestimmten Zeit anzugeben.

Ableiten. Wählen Sie mindestens eine Dichte von Raum und Zeit aus, aus der das neue Feld abgeleitet werden soll. Weitere Informationen finden Sie in „Definieren der STB-Dichte" auf Seite 99.

Feldnamenerweiterung. Geben Sie die Erweiterung ein, die zu den neuen Feldnamen hinzugefügt werden soll. Sie können auswählen, ob Sie diese Erweiterung als **Suffix** oder als **Präfix** hinzufügen möchten.

Aufenthaltsoptionen

Einen Aufenthalt kann man sich als Ort und/oder Zeit vorstellen, an dem bzw. zu der eine Entität dauernd oder wiederholt gefunden wird. Dies könnte beispielsweise verwendet werden, um ein Fahrzeug zu identifizieren, das regelmäßige Transporte durchführt, und um Abweichungen von der Norm zu ermitteln.

STB-Dichte. Wählen Sie die Dichte von Raum und Zeit aus, aus der das neue Feld abgeleitet werden soll. Weitere Informationen finden Sie in „Definieren der STB-Dichte" auf Seite 99.

Feld "Entitäts-ID". Wählen Sie die Entität aus, die als Aufenthalts-ID verwendet werden soll.

Minimale Anzahl von Ereignissen. Ein Ereignis ist eine Zeile in den Daten. Wählen Sie die minimale Anzahl an Vorkommen eines Ereignisses für die Entität aus, damit die Entität den Status "Aufenthalt" aufweist.

Aufenthaltszeit beträgt mindestens. Geben Sie die minimale Dauer an, während derer die Entität sich an demselben Ort aufhalten muss. Damit kann beispielsweise ausgeschlossen werden, dass ein Auto, das an einer Ampel wartet, den Status "Aufenthalt" aufweist.

Überschreiten von STB-Grenzen durch Aufenthalte zulassen. Wenn diese Option ausgewählt ist, ist die Definition eines Aufenthalts weniger streng und sie könnte beispielsweise eine Entität einschließen, die sich in mehreren Space-Time-Boxes aufhält. Wenn Ihre STBs beispielsweise als ganze Stunden definiert sind, würde bei der Auswahl dieser Option eine Entität, die sich eine Stunde lang aufhält, als gültig erkannt, auch wenn die Stunde aus 30 Minuten vor Mitternacht und 30 Minuten nach Mitternacht bestehen würde. Wenn diese Option nicht ausgewählt ist, müssen sich 100 % der Aufenthaltszeit in einer einzigen Space-Time-Box befinden.

Mindestanteil an Ereignissen in qualifizierender Time-Box (%). Ist nur verfügbar, wenn **Überschreiten von STB-Grenzen durch Aufenthalte zulassen** ausgewählt ist. Steuern Sie mit dieser Option den Grad, bis zu dem sich ein in einer STB gemeldeter Aufenthalt tatsächlich mit einer anderen STB überschneiden kann. Wählen Sie den Mindestanteil der Ereignisse aus, die zur Identifikation eines Aufenthalts in einer einzigen STB auftreten müssen. Wenn dieser Wert auf 25 % gesetzt ist und der Anteil der Ereignisse 26 % ist, ist dies als Aufenthalt qualifiziert.

Definieren der STB-Dichte

Wählen Sie die Größe (Dichte) Ihrer Space-Time-Boxes (STB) aus, indem Sie die physische Fläche und die abgelaufene Zeit angeben, die in jede STB eingeschlossen werden sollen.

Geo-Dichte. Wählen Sie die Größe der Fläche aus, die in jede STB eingeschlossen werden soll.

Zeitintervall. Wählen Sie die Anzahl der Stunden aus, die in jede STB eingeschlossen werden sollen.

Feldname. Dieser Name mit dem Präfix "STB" wird auf der Basis Ihrer Auswahl in den vorhergehenden zwei Feldern vervollständigt.

Kapitel 4. Feldoperationsknoten

Überblick über Feldoperationen

Nach der ersten Datenexploration steht in der Regel die Auswahl, die Bereinigung oder die Konstruktion von Daten als Vorbereitung für die Analyse an. Die Palette "Feldfunktionen" enthält viele Knoten, die für diese Transformation und Vorbereitung nützlich sind.

So können Sie mit einem Ableitungsknoten ein Attribut erstellen, das noch nicht in den Daten repräsentiert wird. Oder Sie können mit einem Klassierknoten automatisch Feldwerte für eine gezielte Analyse neu codieren. Typknoten werden häufig verwendet, da sie die Möglichkeit bieten, für jedes Feld im Dataset ein Messniveau, Werte und eine Modellierungsrolle zuzuweisen. Diese Operationen sind nützlich für den Umgang mit fehlenden Werten und für die Downstream-Modellierung.

Die Palette "Feldfunktionen" enthält folgende Knoten:

Der ADP-Knoten (Automated Data Preparation - Automatisierte Datenaufbereitung) kann Ihre Daten analysieren und Korrekturen identifizieren, problematische oder vermutlich überflüssige Felder ausschließen, wie erforderlich neue Attribute ableiten und die Leistung durch intelligente Prüf- und Stichprobenverfahren verbessern. Sie können den Knoten vollständig automatisiert nutzen, damit er Korrekturen wählen und anwenden kann. Sie können die Änderungen aber auch prüfen, bevor sie durchgeführt werden, und wie gewünscht akzeptieren, ablehnen oder ändern.

Der Typknoten gibt Feldmetadaten und Eigenschaften an. Sie können beispielsweise ein Messniveau (stetig, nominal, ordinal oder Flag) für die einzelnen Felder angeben, Optionen für den Umgang mit fehlenden Werten und systemdefinierten Nullwerten festlegen, die Rolle eines Felds zu Modellierungszwecken festlegen, Feld- und Wertbeschriftungen angeben oder die Werte für ein Feld angeben.

Der Filterknoten filtert (verwirft) Felder, benennt Felder um und ordnet Felder von einem Quellenknoten einem anderen zu.

Der Ableitungsknoten ändert Datenwerte oder erstellt neue Felder aus einem oder mehreren bestehenden Feldern. Er erstellt Felder vom Typ "Formel", "Flag", "Nominal", "Status", "Anzahl" und "Bedingt".

Der Ensemble-Knoten kombiniert zwei oder mehr Modellnuggets, um genauere Vorhersagen zu erzielen, als aus einem dieser Modelle allein gewonnen werden können.

Der Füllerknoten ersetzt Feldwerte und ändert den Speichertyp. Sie können auswählen, dass die Werte auf der Grundlage einer CLEM-Bedingung wie beispielsweise `@BLANK(@FIELD)` ersetzt werden sollen. Alternativ können Sie auswählen, dass alle Leerstellen oder Nullwerte mit einem bestimmten Wert ersetzt werden sollen. Füllerknoten werden häufig zusammen mit einem Typknoten verwendet, um fehlende Werte zu ersetzen.

Der Anonymisierungsknoten ändert die Art und Weise, wie Feldnamen und -werte weiter unten im Stream dargestellt werden, und verschleiert damit die ursprünglichen Daten. Dies kann sinnvoll sein, wenn andere Benutzer in die Lage versetzt werden sollen, Modelle unter Verwendung vertraulicher Daten wie beispielsweise Kundennamen zu erstellen.

Der Umcodierungsknoten transformiert ein Set kategorialer Werte in ein anderes. Die Umcodierung dient zur Reduzierung von Kategorien bzw. Neugruppierung von Daten für die Analyse.

Der Klassierknoten erstellt automatisch neue nominale Felder (Setfelder) auf der Grundlage der Werte eines oder mehrerer bestehender stetiger Felder (numerischer Bereich). Sie können beispielsweise ein stetiges Einkommensfeld in ein neues kategoriales Feld transformieren, das Einkommensgruppen als Abweichungen vom Mittelwert enthält. Nach der Erstellung von Klassen für das neue Feld können Sie einen Ableitungsknoten anhand der Trennwerte generieren.

Mit dem RFM-Analyseknoten (Recency-, Frequency-, Monetary-Analyse) können Sie quantitativ ermitteln, welche Kunden wahrscheinlich die besten sind, indem Sie untersuchen, wann sie zuletzt etwas von Ihnen erworben haben (Recency - Aktualität), wie häufig sie eingekauft haben (Frequency - Häufigkeit) und wie viel sie für alle Transaktionen zusammengenommen ausgegeben haben (Monetary - Geldwert).

Der Partitionsknoten erstellt ein Partitionsfeld, das Daten in getrennte Subsets für die Trainings-, Test- und Validierungsphase der Modellerstellung aufteilt.

Der Dichotomknoten leitet mehrere Flagfelder auf der Grundlage der kategorialen Werte ab, die für ein oder mehrere nominale Felder definiert sind.

Der Knoten "Umstrukturieren" wandelt ein nominales Feld oder ein Flagfeld in eine Gruppe von Feldern um, die mit den Werten aus einem weiteren Feld ausgefüllt werden können. Beispiel: Aus einem Feld mit dem Namen *Zahlungsart* und den Werten *Kreditkarte*, *Bar* und *EC-Karte* werden drei neue Felder erstellt (*Kreditkarte*, *Bar*, *EC-Karte*), die jeweils den Wert der tatsächlichen Zahlung enthalten.

Der Transponierknoten vertauscht die Daten in Zeilen und Spalten, sodass aus Datensätzen Felder und aus Feldern Datensätze werden.

Der Zeitintervallknoten gibt Intervalle an und erstellt (bei Bedarf) Beschriftungen für die Modellierung von Zeitreihendaten. Wenn die Werte nicht gleichmäßig verteilt sind, kann der Knoten nach Bedarf Werte auffüllen oder aggregieren, um ein gleichmäßiges Intervall zwischen den Datensätzen zu erzeugen.

Der Verlaufsknoten erstellt neue Felder mit Daten aus Feldern in vorangegangenen Datensätzen. Verlaufsknoten werden am häufigsten für sequenzielle Daten, beispielsweise Zeitreihendaten, verwendet. Vor der Verwendung eines Verlaufsknotens sollten die Daten mithilfe eines Sortierknotens sortiert werden.

Der Knoten "Felder ordnen" definiert die natürliche Reihenfolge, die bei der Anzeige der nachgeordneten Felder verwendet wird. Diese Reihenfolge betrifft die Anzeige von Feldern an unterschiedlichen Stellen, beispielsweise in Tabellen, Listen und in der Feldauswahl. Dieser Vorgang dient beispielsweise dazu, um bei der Arbeit mit umfangreichen Datasets die relevanten Felder deutlicher hervorzuheben.

Mehrere dieser Knoten können direkt über den von einem Data Audit-Knoten erstellten Audit-Bericht generiert werden. Weitere Informationen finden Sie im Thema „Erzeugen von anderen Knoten zur Datenvorbereitung” auf Seite 302.

Automatisierte Datenaufbereitung

Die Aufbereitung von Daten zur Analyse ist einer der wichtigsten Schritte in jedem Projekt und gewöhnlich auch einer der zeitaufwendigsten. Die automatisierte Datenaufbereitung (ADP) übernimmt diese Aufgabe für Sie. Sie analysiert Ihre Daten und identifiziert Problemlösungen, findet problematische oder wahrscheinlich nicht nützliche Felder, leitet zum passenden Zeitpunkt neue Attribute ab und verbessert die Leistungsfähigkeit durch intelligente Screening-Methoden. Sie können den Algorithmus **vollautomatisch** verwenden und so Problemlösungen auswählen und anwenden oder Sie können ihn **interaktiv** verwenden und so die Änderungen in einer Vorschau betrachten, bevor sie vorgenommen werden, und sie nach Bedarf akzeptieren oder ablehnen.

Mit ADP können Sie Ihre Daten schnell und einfach für die Modellerstellung aufbereiten, ohne über Vorkenntnisse der dazugehörigen statistischen Konzepte verfügen zu müssen. Modelle lassen sich damit schneller erstellen und scoren; zudem verbessert sich mit ADP die Robustheit automatisierter Modellierungsprozesse wie der Modellaktualisierung und von Champion/Challenger.

Hinweis: Wenn die ADP ein Feld für die Analyse vorbereitet, erstellt sie ein neues Feld, das die Anpassungen oder Transformationen enthält, anstatt die bestehenden Werte und Eigenschaften des alten Felds zu ersetzen. Das alte Feld wird bei der weiteren Analyse nicht verwendet; seine Rolle wird auf "Keine" gesetzt.

Beispiel. Eine Versicherungsgesellschaft mit beschränkten Ressourcen für die Untersuchung der Versicherungsansprüche von Hauseigentümern möchte ein Modell zur Kennzeichnung verdächtiger, potenziell betrügerischer Ansprüche erstellen. Vor Erstellung des Modells bereiten sie die Daten für die Modellierung mithilfe der automatisierten Datenaufbereitung vor. Da sie die vorgeschlagenen Transformationen zunächst überprüfen möchten, bevor die Transformationen angewendet werden, nutzen sie die automatisierte Datenaufbereitung im interaktiven Modus.

Eine Gruppe in der Kraftfahrzeugindustrie erfasst die Verkaufszahlen verschiedener Personenkraftwagen. Um starke und schwache Modelle identifizieren zu können, soll eine Beziehung zwischen den Fahrzeugverkaufszahlen und den Fahrzeugeigenschaften hergestellt werden. Zur Vorbereitung der Daten für die Analyse wird die automatisierte Datenaufbereitung verwendet. Es werden Modelle mit Daten "vor" und "nach" der Aufbereitung erstellt, um zu sehen, wie sich die Ergebnisse unterscheiden.

Was ist Ihr Ziel? Die automatisierte Datenaufbereitung empfiehlt Schritte zur Datenaufbereitung, die sich auf die Geschwindigkeit auswirken, mit der andere Algorithmen Modelle erstellen können und die Vorhersagekraft dieser Modelle verbessern. Diese können die Transformation, Erstellung und Auswahl von Funktionen beinhalten. Das Ziel kann ebenfalls transformiert werden. Sie können die Prioritäten der Modellerstellung festlegen, auf die sich die Datenaufbereitung konzentrieren sollte.

- **Geschwindigkeit und Genauigkeit ausbalancieren.** Diese Option bereitet die Daten auf und sorgt dabei für eine ausgeglichene Priorität zwischen der Geschwindigkeit, mit der Daten durch die Modellerstellung verarbeitet werden, und der Genauigkeit der Vorhersagen.

- **Für Geschwindigkeit optimieren.** Diese Option bereitet die Daten auf und gibt dabei der Geschwindigkeit Vorrang, mit der Daten durch Modellerstellungsalgorithmen verarbeitet werden. Wählen Sie diese Option aus, wenn Sie mit sehr großen Datasets arbeiten oder nach einer schnellen Antwort suchen.
- **Für Genauigkeit optimieren.** Diese Option bereitet die Daten auf und gibt dabei der Genauigkeit der durch Modellerstellungsalgorithmen erzeugten Vorhersagen Vorrang.
- **Angepasste Analyse.** Wählen Sie diese Option, wenn Sie den Algorithmus auf der Registerkarte "Einstellungen" manuell ändern wollen. Beachten Sie, dass diese Einstellung automatisch ausgewählt wird, wenn Sie anschließend Änderungen auf der Registerkarte "Einstellungen" vornehmen, die mit einem der anderen Ziele nicht kompatibel sind.

Knoten-Training

Der ADP-Knoten wurde als Prozessknoten implementiert und arbeitet ähnlich wie der Typknoten; **Training** des ADP-Knotens entspricht der Instantiierung des Typknotens. Sobald die Analyse durchgeführt wurde, werden die angegebenen Transformationen ohne weitere Analyse auf die Daten angewendet, solange sich das vorgelagerte Datenmodell nicht ändert. Wenn die Verbindung zum ADP-Knoten getrennt wird, speichert dieser wie die Typ- und Filterknoten das Datenmodell und die Transformationen und muss so nicht erneut trainiert werden, wenn die Verbindung wiederhergestellt wird; dadurch können Sie ihn auf ein Subset typischer Daten trainieren und anschließend kopieren oder so oft wie nötig auf Live-Daten bereitstellen.

Verwendung der Symbolleiste

Mit der Symbolleiste können Sie die Anzeige der Datenanalyse ausführen und aktualisieren sowie Knoten generieren, die Sie zusammen mit den Originaldaten verwenden können.

- **Erzeugen** In diesem Menü können Sie entweder einen Filter- oder einen Ableitungsknoten erzeugen. Beachten Sie, dass dieses Menü nur verfügbar ist, wenn auf der Registerkarte "Analyse" eine Analyse angezeigt wird.

  Der Filterknoten entfernt transformierte Eingabefelder. Wenn Sie den ADP-Knoten so konfigurieren, dass die Originaleingabefelder im Dataset beibehalten werden, wird dadurch das Originaleingabeset wiederhergestellt und Sie können das Wertfeld bezüglich der Eingaben interpretieren. Dies ist beispielsweise dann nützlich, wenn Sie eine Grafik des Wertfelds anhand mehrerer Eingaben erzeugen möchten.

  Der Ableitungsknoten kann das Originaldataset und die Originalzieleinheiten wiederherstellen. Sie können einen Ableitungsknoten nur dann erzeugen, wenn der ADP-Knoten eine Analyse enthält, die ein Bereichsziel neu skaliert (d. h. die Box-Cox-Neuskalierung ist im Feld "Eingaben und Ziel vorbereiten" ausgewählt). Sie können keinen Ableitungsknoten erzeugen, wenn das Ziel kein Bereich ist oder wenn die Box-Cox-Neuskalierung nicht ausgewählt ist. Weitere Informationen finden Sie im Thema „Erzeugen eines Ableitungsknotens" auf Seite 117.
- **Ansicht** Enthält Optionen, die steuern, was auf der Registerkarte "Analyse" angezeigt wird. Dazu zählen die Steuerungen zur Bearbeitung von Grafiken sowie die Anzeigen für das Hauptfenster und verknüpfte Ansichten.
- **Vorschau** Zeigt ein Muster der Transformationen an, die auf die Eingabedaten angewendet werden.
- **Daten analysieren** Startet eine Analyse mit den aktuellen Einstellungen und zeigt die Ergebnisse auf der Registerkarte "Analyse" an.
- **Analyse löschen** Löscht die bestehende Analyse (nur verfügbar, wenn eine aktuelle Analyse vorhanden ist).

Knotenstatus

Der Status des ADP-Knotens im Erstellungsbereich von IBM SPSS Modeler wird entweder durch einen Pfeil oder ein Häkchen auf dem Symbol verdeutlicht, das anzeigt, ob eine Analyse durchgeführt wurde oder nicht.

Registerkarte "Felder"

Bevor Sie ein Modell erstellen können, müssen Sie festlegen, welche Felder als Ziele und als Eingaben verwendet werden sollen. Von wenigen Ausnahmen abgesehen, verwenden alle Modellierungsknoten die Feldinformationen des oberhalb liegenden Typknotens. Wenn Sie einen Typknoten verwenden, um Eingabe- und Zielfelder auszuwählen, brauchen Sie auf dieser Registerkarte keine Änderungen vorzunehmen.

Typknoteneinstellungen verwenden. Diese Option weist den Knoten an, die Feldinformationen von einem weiter oben liegenden Typknoten zu verwenden. Dies ist die Standardeinstellung.

Benutzerdefinierte Einstellungen verwenden. Diese Option weist den Knoten an, die hier angegebenen Feldinformationen anstelle der in einem weiter oben liegenden Typknoten angegebenen zu verwenden. Geben Sie nach Auswahl dieser Option wie erforderlich die unten stehenden Felder an.

Ziel. Wählen Sie die Zielfelder für Modelle aus, die eines oder mehrere Zielfelder benötigen. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Ziel* festlegen.

Eingaben. Wählen Sie das/die Eingabefeld(er) aus. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Eingabe* festlegen.

Registerkarte "Einstellungen"

Die Registerkarte "Einstellungen" enthält mehrere unterschiedliche Gruppen von Einstellungen, die Sie ändern können, um genau festzulegen, wie der Algorithmus Ihre Daten verarbeiten soll. Wenn Sie an den Standardeinstellungen Änderungen vornehmen, die mit den anderen Zielen nicht kompatibel sind, wird auf der Registerkarte "Ziel" automatisch die Option **Analyse anpassen** ausgewählt.

Feldeinstellungen

Häufigkeitsfeld verwenden. Mit dieser Option könen Sie ein Feld als Häufigkeitsgewichtung auswählen. Wenden Sie diese Option an, wenn die Datensätze in Ihren Trainingsdaten jeweils mehr als eine Einheit darstellen; dies ist zum Beispiel bei aggregierten Daten der Fall. Die Feldwerte sollten die Anzahl der Einheiten sein, die von jedem Datensatz repräsentiert werden.

Gewichtungsfeld verwenden. Mit dieser Option könen Sie ein Feld als Fallgewichtung auswählen. Fallgewichtungen werden verwendet, um Differenzen in der Varianz zwischen den Ebenen des Ausgabefelds zu berücksichtigen.

Handhabung von Feldern, die von der Modellierung ausgeschlossen sind. Geben Sie an, was mit ausgeschlossenen Feldern geschehen soll. Sie können wählen, ob sie aus den Daten herausgefiltert werden sollen oder einfach ihre *Rolle* auf **Keine** gesetzt werden soll.

Hinweis: Dieser Vorgang wird auch auf ein transformiertes Ziel angewendet. Wenn beispielsweise die neu abgeleitete Version des Ziels als das Feld **Ziel** verwendet wird, wird das Originalziel entweder gefiltert oder auf **Keine** gesetzt.

Wenn die eingehenden Felder nicht mit der vorhandenen Analyse übereinstimmen. Geben Sie an, was geschehen soll, falls eines oder mehrere erforderliche Eingabefelder im eingehenden Datensatz fehlen, wenn Sie einen trainierten ADP-Knoten ausführen.

- **Ausführung anhalten und vorhandene Analyse beibehalten.** Dieser Befehl stoppt den Ausführungsvorgang, speichert die gegenwärtigen Analysedaten und zeigt eine Fehlernachricht an.
- **Vorhandene Analyse löschen und neue Daten analysieren.** Dadurch wird die vorhandene Analyse gelöscht, die eingehenden Daten werden analysiert und die empfohlenen Transformationen werden auf diese Daten angewendet.

Aufbereiten von Datum und Uhrzeit

Viele Modellierungsalgorithmen sind nicht in der Lage, Datums- und Zeitangaben direkt zu behandeln; mit diesen Einstellungen können Sie neue Laufzeitdaten ableiten, die Sie in Ihren bestehenden Daten als Modelleingaben aus Datums- und Zeitangaben verwenden können. Die Felder mit Datums- und Zeitangaben müssen mit Datums- oder Zeitspeichertypen vordefiniert sein. Die ursprünglichen Datums- und Zeitfelder werden nicht als Modelleingaben nach der automatisierten Datenaufbereitung empfohlen.

Datums- und Zeitangaben für Modellierung aufbereiten. Durch Inaktivieren dieser Option werden alle anderen Datums- und Zeiteingaben inaktiviert und die Auswahl beibehalten.

Verstrichene Zeit bis zum Referenzdatum berechnen. Errechnet die Anzahl der Jahre/Monate/Tage seit einem Referenzdatum für jede Variable, die Datumsangaben enthält.

- **Referenzdatum.** Geben Sie das Datum an, ab dem die Dauer bezüglich der Datumsinformationen in den Eingabedaten berechnet wird. Durch die Auswahl von **Heutiges Datum** wird das aktuelle Systemdatum stets verwendet, wenn ADP ausgeführt wird. Um ein bestimmtes Datum zu verwenden, wählen Sie **Festes Datum** und geben Sie das erforderliche Datum ein. Das aktuelle Datum wird automatisch im Feld **Festes Datum** eingegeben, wenn der Knoten zum ersten Mal erstellt wird.
- **Einheiten für Datumsdauer.** Legen Sie fest, ob ADP die Einheit der Datumsdauer automatisch bestimmen soll, oder wählen Sie **Feste Einheiten** für Jahre, Monate oder Tage.

Verstrichene Zeit bis zur Referenzzeit berechnen. Errechnet die Anzahl der Stunden/Minuten/Sekunden seit einer Referenzzeit für jede Variable, die Uhrzeiten enthält.

- **Referenzzeit.** Geben Sie die Zeit an, ab der die Dauer bezüglich der Zeitinformationen in den Eingabedaten berechnet wird. Durch die Auswahl von **Aktuelle Uhrzeit** wird die aktuelle Systemzeit stets verwendet, wenn ADP ausgeführt wird. Um eine bestimmte Uhrzeit zu verwenden, wählen Sie **Feste Uhrzeit** und geben Sie die erforderlichen Daten ein. Die aktuelle Uhrzeit wird automatisch im Feld **Feste Uhrzeit** eingegeben, wenn der Knoten zum ersten Mal erstellt wird.
- **Einheiten für Zeitdauer.** Legen Sie fest, ob ADP die Einheit der Zeitdauer automatisch bestimmen soll, oder wählen Sie **Feste Einheiten** für Stunden, Minuten oder Sekunden.

Zyklische Zeitelemente extrahieren. Verwenden Sie diese Einstellungen, um ein einzelnes Datums- oder Zeitfeld in ein oder mehrere Felder aufzuteilen. Wenn Sie zum Beispiel alle drei Datumskontrollkästchen auswählen, wird das Eingabedatumsfeld "1954-05-23" in drei Felder aufgeteilt: 1954, 5 und 23, wobei jedes Feld das unter **Feldnamen** definierte Suffix verwendet und das ursprüngliche Datumsfeld ignoriert wird.

- **Aus Datumsangaben extrahieren.** Legen Sie für eine beliebige Datumseingabe fest, ob Sie Jahre, Monate, Tage oder eine Kombination daraus extrahieren möchten.
- **Aus Zeitangaben extrahieren.** Legen Sie für eine beliebige Zeiteingabe fest, ob Sie Stunden, Minuten, Sekunden oder eine Kombination daraus extrahieren möchten.

Felder ausschließen

Schlechte Datenqualität kann sich negativ auf die Genauigkeit Ihrer Vorhersagen auswirken; Sie können daher die akzeptable Qualitätsstufe für Eingabemerkmale festlegen. Alle konstanten oder 100 % an fehlenden Werten aufweisenden Felder werden automatisch ausgeschlossen.

Eingabefelder mit niedriger Qualität ausschließen. Durch Inaktivieren dieser Option werden alle anderen Befehle "Felder ausschließen" inaktiviert und die Auswahl beibehalten.

Felder mit zu vielen fehlenden Werten ausschließen. Felder mit mehr als dem angegebenen Prozentsatz an fehlenden Werten werden aus der weiteren Analyse ausgeschlossen. Geben Sie einen Wert größer oder gleich 0 ein, was dem Inaktivieren dieser Option entspricht, und einen Wert kleiner oder gleich 100, sodass die Felder mit allen fehlenden Werten automatisch ausgeschlossen werden. Der Standardwert ist 50.

Nominale Felder mit zu vielen eindeutigen Kategorien ausschließen. Nominale Felder mit mehr als der angegebenen Anzahl an Kategorien werden aus der weiteren Analyse ausgeschlossen. Geben Sie eine positive Ganzzahl ein. Der Standardwert ist 100. Dies ist nützlich für das automatische Entfernen von Feldern aus der Modellierung, die eine datensatzeindeutige Information enthalten, wie zum Beispiel eine ID, eine Adresse oder einen Namen.

Kategoriale Felder mit zu vielen Werten in einer einzelnen Kategorie ausschließen. Ordinale und nominale Felder mit einer Kategorie, die mehr als die angegebene Prozentzahl an Datensätzen enthält, werden aus der weiteren Analyse ausgeschlossen. Geben Sie einen Wert größer oder gleich 0 ein, was dem Inaktivieren dieser Option entspricht, und einen Wert kleiner oder gleich 100, sodass konstante Felder automatisch ausgeschlossen werden. Der Standardwert ist 95.

Vorbereiten von Eingaben und Zielen

Da sich Daten nie in einem perfekten Zustand für die Verarbeitung befinden, kann es hilfreich sein, vor dem Ausführen einer Analyse einige Einstellungen anzupassen. Dazu können zum Beispiel das Ein- oder Ausschließen von Ausreißern gehören, Angaben über den Umgang mit fehlenden Werten oder das Anpassen des Typs.

Hinweis: Wenn Sie die Werte in diesem Feld ändern, wird die Registerkarte **Ziele** automatisch auf die Auswahl der Option **Benutzerdefinierte Analyse** aktualisiert.

Eingabe- und Zielfelder für Modellierung vorbereiten. Schaltet alle Felder in dem Eingabefeld entweder an oder aus.

Typ anpassen und Datenqualität verbessern. Für die Eingaben und das Ziel können mehrere Datentransformationen separat angegeben werden, da es wünschenswert sein kann, die Zielwerte nicht zu ändern. So kann zum Beispiel eine Vorhersage über das Einkommen in Dollar aussagekräftiger sein als eine Vorhersage, die als Log(Dollar) angegeben wird. Wenn das Ziel außerdem fehlende Werte aufweist, ergibt sich für die Vorhersage kein Nutzen daraus, die fehlenden Werte zu ergänzen, wogegen das Ergänzen fehlender Werte bei den Eingaben durchaus dazu führen kann, dass einige Algorithmen Informationen aufbereiten können, die andernfalls verloren gehen würden.

Weitere Einstellungen für diese Transformationen, wie zum Beispiel der Ausreißertrennwert, sind sowohl für das Ziel als auch die Eingaben üblich.

Sie können die folgenden Einstellungen für entweder die Eingaben oder das Ziel oder für beides vornehmen:

- **Typ numerischer Felder anpassen.** Damit können Sie bestimmen, ob die numerischen Felder mit einem Messniveau *Ordinal* auf *Stetig* konvertiert werden können oder umgekehrt. Sie können die minimalen und maximalen Schwellenwerte für die Konversion angeben.
- **Nominale Felder neu sortieren.** Mit dieser Option können Sie nominale Felder (Setfelder) der Reihe nach sortieren, von der kleinsten zur größten Kategorie.
- **Ausreißerwerte in stetigen Feldern ersetzen.** Geben Sie an, ob Ausreißer ersetzt werden sollen. Nutzen Sie diese Option in den Verbindung mit dem **Verfahren zum Ersetzen von Ausreißern** unten.
- **Stetige Felder: Fehlende Werte durch Mittelwert ersetzen.** Mit dieser Option können Sie fehlende Werte stetiger Merkmale (Bereich) ersetzen.
- **Nominale Felder: Fehlende Werte durch Modalwert ersetzen.** Mit dieser Option können Sie fehlende Werte nominaler Merkmale (Set) ersetzen.
- **Ordinale Felder: Fehlende Werte durch Median ersetzen.** Mit dieser Option können Sie fehlende Werte ordinaler Merkmale (sortiertes Set) ersetzen.

Maximale Anzahl an Werten für ordinale Felder. Geben Sie den Schwellenwert an, bei dem ordinale Felder (sortierte Setfelder) in stetige Felder (Setfelder) umdefiniert werden sollen. Der Standardwert ist 10. Wenn also ein ordinales Feld mehr als 10 Kategorien aufweist, wird es als stetig (Bereich) umdefiniert.

Minimale Anzahl an Werten für stetige Felder. Geben Sie den Schwellenwert an, bei dem Skalenfelder oder stetige (Bereich) Felder in ordinale Felder (sortierte Setfelder) umdefiniert werden sollen. Der Standardwert ist 5. Wenn also ein stetiges Feld weniger als 5 Kategorien aufweist, wird es als ordinal (sortiertes Set) umdefiniert.

Ausreißertrennwert. Geben Sie das in Standardabweichungen gemessene Ausreißertrennwertkriterium an. Der Standardwert ist 3.

Methode für Ersatz von Ausreißern. Wählen Sie aus, ob Ausreißer durch Trimmen (Setzen) auf den Trennwert ersetzt werden sollen oder ob sie gelöscht und als fehlende Werte angegeben werden sollen. Jeder als fehlender Wert eingestufte Ausreißer unterliegt den oben ausgewählten Einstellungen für die Behandlung fehlender Werte.

Alle stetigen Eingabefelder auf gemeinsame Skala setzen. Um stetige Eingabefelder zu normalisieren, kreuzen Sie dieses Kontrollkästchen an und wählen das Normalisierungsverfahren aus. Standardmäßig ist die **Z-Wert-Transformation** eingestellt, bei der Sie den **Endgültigen Mittelwert** angeben können, der den Standardwert 0 hat, und die **Endgültige Standardabweichung**, die den Standardwert 1 hat. Alternativ können Sie die **Min/Max-Transformation** auswählen und die minimalen und maximalen Werte angeben, die standardmäßig auf 0 beziehungsweise 100 eingestellt sind.

Dieses Feld ist besonders nützlich, wenn Sie **Merkmalerstellung durchführen** im Merkmalbereich "Erstellen& auswählen" angeben.

Stetiges Ziel mit Box-Cox-Transformation neu skalieren. Um ein stetiges Zielfeld (Skala oder Bereich) zu normalisieren, wählen Sie dieses Kontrollkästchen aus. Bei der Box-Cox-Transformation sind standardmäßig die Werte 0 für den **Endgültigen Mittelwert** und 1 für die **Endgültige Standardabweichung** eingestellt.

Hinweis: Bei einer Normalisierung des Ziels wird die Dimension des Ziels transformiert. In diesem Fall müssen Sie u. U. einen Ableitungsknoten für die Anwendung einer inversen Transformation erstellen, um die transformierten Einheiten wieder in ein zur weiteren Verarbeitung erkennbares Format zu bringen. Weitere Informationen finden Sie im Thema „Erzeugen eines Ableitungsknotens” auf Seite 117.

Auswahl von Erstellung und Funktion

Um die Vorhersagekraft Ihrer Daten zu verbessern, können Sie die Eingabefelder transformieren oder basierend auf den bestehenden Feldern neue erstellen.

Hinweis: Wenn Sie die Werte in diesem Feld ändern, wird die Registerkarte **Ziele** automatisch auf die Auswahl der Option **Benutzerdefinierte Analyse** aktualisiert.

Eingabefelder transformieren, erstellen und auswählen, um Vorhersagekraft zu verbessern. Schaltet alle Felder in dem Eingabefeld entweder an oder aus.

Dünn besetzte Kategorien für maximale Zuordnung zusammenführen. Mit dieser Option erstellen Sie ein sparsameres Modell, indem die Anzahl der zu verarbeitenden Variablen in Zusammenhang mit dem Ziel reduziert wird. Ändern Sie bei Bedarf den Wahrscheinlichkeitswert von der Standardeinstellung 0,05.

Hinweis: Wenn alle Kategorien zu einer verschmolzen werden, werden die ursprünglichen und abgeleiteten Versionen des Felds ausgeschlossen, da sie als Einflussgrößen keinen Wert haben.

Wenn es kein Ziel gibt, dünn besetzte Kategorien basierend auf Häufigkeiten zusammenführen. Wenn Sie Daten verarbeiten, die kein Ziel aufweisen, können Sie auswählen, dünn besetzte Kategorien von ordinalen (sortiertes Set) oder nominalen (Set) Merkmalen oder beiden zusammenzuführen. Geben Sie den minimalen Prozentsatz an Fällen oder Datensätzen in den Daten an, der die zusammenzuführenden Kategorien identifiziert. Der Standardwert ist 10.

Kategorien werden mithilfe der folgenden Regeln zusammengeführt:

- Das Zusammenführen erfolgt nicht bei binären Feldern.
- Wenn es bei der Zusammenführung nur zwei Kategorien gibt, stoppt die Zusammenführung.
- Wenn es keine originale Kategorie oder eine während des Zusammenführens erzeugte Kategorie gibt, die weniger als den angegebenen minimalen Prozentsatz an Fällen aufweist, stoppt die Zusammenführung.

Stetige Felder klassieren, Vorhersagekraft bewahren. Wenn die Daten ein kategoriales Ziel enthalten, können Sie stetige Eingaben mit starkem Zusammenhang einteilen, um die Verarbeitungsleistung zu verbessern. Ändern Sie bei Bedarf den Wahrscheinlichkeitswert für die homogenen Subsets von der Standardeinstellung 0,05.

Wenn in dem Klassierungsvorgang eine einzelne Klassierung für ein bestimmtes Feld durchgeführt wird, werden die Original- und eingeteilten Versionen des Felds ausgeschlossen, da sie keinen Wert als Prädiktor aufweisen.

Hinweis: Die Klassierung in ADP unterscheidet sich von der optimalen Klassierung, die in anderen Teilen von IBM SPSS Modeler verwendet wird. Bei der optimalen Klassierung werden Entropieinformationen verwendet, um eine stetige Variable in eine kategoriale Variable umzuwandeln; dazu müssen Daten sortiert und im Arbeitsspeicher abgelegt werden. ADP verwendet homogene Subsets zum Klassieren einer stetigen Variable, das bedeutet, dass die ADP-Klassierung keine Daten sortieren und im Arbeitsspeicher ablegen muss. Der Einsatz von homogenen Subsets zum Klassieren einer stetigen Variablen bedeutet, dass die Anzahl der Kategorien nach der Klassierung immer kleiner oder gleich der Anzahl der Kategorien des Ziels ist.

Merkmalauswahl durchführen. Wählen Sie diese Option aus, um Merkmale mit einem niedrigen Korrelationskoeffizienten zu entfernen. Ändern Sie bei Bedarf den Wahrscheinlichkeitswert von der Standardeinstellung 0,05.

Diese Option gilt nur für stetige Eingabemerkmale mit stetigem Ziel und kategoriale Eingabemerkmale.

Merkmalerstellung durchführen. Wählen Sie diese Option aus, um neue Merkmale von einer Kombination aus mehreren bestehenden Merkmalen abzuleiten (die in der Modellierung nicht weiter beachtet werden).

Diese Option gilt nur für stetige Eingabemerkmalen mit stetigem Ziel oder Eingabemerkmalen, in denen kein Ziel vorhanden ist.

Feldnamen

Zur einfachen Identifikation neuer und transformierter Funktionen erstellt ADP allgemeine neue Namen, Präfixe oder Suffixe und wendet diese an. Sie können diese Namen ändern und ihnen mehr Aussagekraft für Ihre eigenen Anforderungen und Daten geben. Andere Beschriftungen müssen in einem nachgelagerten Typknoten angegeben werden.

Transformierte und erstellte Felder. Geben Sie die Namenserweiterungen an, die auf transformierte Ziel- und Eingabefelder angewendet werden sollen.

Beachten Sie, dass in einem Zeichenfolgeknoten die Einstellung leerer Zeichenfolgefelder je nach gewählter Behandlung nicht verwendeter Felder einen Fehler verursachen kann. Wenn **Behandlung von aus der Modellierung ausgeschlossenen Feldern** unter "Feldeinstellungen" auf der Registerkarte "Einstellungen" auf **Nicht verwendete Felder filtern** gesetzt ist, können die Namenserweiterungen für Eingaben und das Ziel auf "Nichts" gesetzt werden. Die Originalfelder werden durch Filterung ausgeschlossen und die transformierten Felder darüber gespeichert; in diesem Fall haben die transformierten Felder den gleichen Namen wie Ihr Original.

Wenn Sie jedoch **Richtung nicht verwendeter Felder auf "Keine" setzen** auswählen, werden leere Namenserweiterungen (Nullnamenserweiterungen) für das Ziel und die Eingaben einen Fehler verursachen, weil Sie versuchen, doppelt vorhandene Feldnamen zu erstellen.

Geben Sie außerdem über die Einstellungen "Auswählen und erstellen" den Präfixnamen an, der auf erstellte Funktionen angewendet werden soll. Der neue Name wird erstellt, indem ein numerisches Suffix an diesen Präfixstammnamen angehängt wird. Das Zahlenformat hängt davon ab, wie viele neue Merkmale abgeleitet werden, zum Beispiel:

- 1-9 erstellte Merkmale werden wie folgt benannt: Merkmal1 bis Merkmal9.
- 10-99 erstellte Merkmale werden wie folgt benannt: Merkmal01 bis Merkmal99.
- 100-999 erstellte Merkmale werden wie folgt benannt: Merkmal001 bis Merkmal999.

So wird gewährleistet, dass die erstellten Merkmale ungeachtet ihrer Anzahl in einer vernünftigen Reihenfolge sortiert werden.

Dauer aus Daten und Zeiten berechnet. Geben Sie die Namenserweiterungen an, die auf die aus Datums- und Zeitangaben berechnete Dauer angewendet werden sollen.

Aus Daten und Zeiten extrahierte zyklische Elemente. Geben Sie die Namenserweiterungen an, die auf die aus Datums- und Zeitangaben extrahierten zyklischen Elemente angewendet werden sollen.

Registerkarte "Analyse"

1. Wenn Sie mit den ADP-Einstellungen einschließlich aller in den Registerkarten "Ziel", "Felder" und "Einstellungen" vorgenommenen Änderungen zufrieden sind, klicken Sie auf **Daten analysieren**. Der Algorithmus wendet die Eingabedaten an und zeigt die Ergebnisse auf der Registerkarte "Analyse" an.

Die Registerkarte "Analyse" enthält Ausgaben in Grafik- und Tabellenform, die die Verarbeitung Ihrer Daten zusammenfassen, und zeigt Empfehlungen an, wie die Daten möglicherweise bearbeitet oder zum Scoring verbessert werden können. Anschließend können Sie diese Empfehlungen überprüfen und entweder akzeptieren oder ablehnen.

Die Registerkarte "Analyse" besteht aus zwei Bereichen, der Hauptansicht im linken Bereich und der verknüpften oder Hilfsansicht im rechten Bereich. Es gibt drei Hauptansichten:

- Feldverarbeitungsübersicht (Standard). Weitere Informationen finden Sie im Thema „Feldverarbeitungsübersicht” auf Seite 111.
- Felder. Weitere Informationen finden Sie im Thema „Felder” auf Seite 111.
- Aktionsübersicht. Weitere Informationen finden Sie im Thema „Aktionsübersicht” auf Seite 112.

Es gibt vier verknüpfte/Hilfsansichten:

- Vorhersagekraft (Standard). Weitere Informationen finden Sie im Thema „Vorhersagekraft” auf Seite 113.
- Feldertabelle. Weitere Informationen finden Sie im Thema „Feldertabelle” auf Seite 113.
- Felddetails. Weitere Informationen finden Sie im Thema „Felddetails” auf Seite 113.
- Aktionsdetails. Weitere Informationen finden Sie im Thema „Aktionsdetails” auf Seite 115.

Verknüpfungen zwischen Ansichten

In der Hauptansicht steuert unterstrichener Text in den Tabellen die Anzeige in der verknüpften Ansicht. Wenn Sie auf den Text klicken, erhalten Sie Informationen über ein bestimmtes Feld, ein Set von Feldern oder einen Verarbeitungsschritt. Der zuletzt von Ihnen ausgewählte Link wird in einer dunkleren Farbe angezeigt; dies hilft Ihnen dabei, die Verbindung zwischen den Inhalten der beiden Ansichtsbereiche zu identifizieren.

Zurücksetzen der Ansichten

Klicken Sie auf **Zurücksetzen** im unteren Bereich der Hauptansicht, um die ursprünglichen Empfehlungen der Analyse erneut anzuzeigen und alle in den Analyseansichten vorgenommenen Änderungen rückgängig zu machen.

Feldverarbeitungsübersicht

Die Tabelle "Feldverarbeitungsübersicht" gibt Ihnen eine Momentaufnahme des projizierten Gesamteinflusses der Verarbeitung, einschließlich Änderungen des Status der Merkmale und der Anzahl der erstellten Merkmale.

Beachten Sie, dass dabei kein Modell erstellt wird und somit kein Maß oder keine Grafik der Veränderung der Gesamtvorhersagekraft vor und nach der Datenaufbereitung vorhanden ist; Sie können stattdessen Grafiken der Vorhersagekraft einzelner empfohlener Prädiktoren anzeigen.

Die Tabelle zeigt folgende Informationen an:

- Die Anzahl der Zielfelder.
- Die Anzahl der ursprünglichen Prädiktoren (Eingabeprädiktoren).
- Die zur Verwendung in der Analyse und Modellierung empfohlenen Prädiktoren. Dazu gehören die Gesamtzahl empfohlener Felder, die Anzahl der ursprünglichen, nicht transformierten empfohlenen Felder, die Anzahl transformierter empfohlener Felder (außer Zwischenversionen, von Datums- und Zeitprädiktoren abgeleitete Felder und erstellte Prädiktoren), die Anzahl der von Datums- und Zeitfelder abgeleiteten empfohlenen Felder sowie die Anzahl erstellter empfohlener Prädiktoren.
- Die Anzahl der Eingabeprädiktoren, die in keiner Form empfohlen werden, sei es in ihrer ursprünglichen Form, als abgeleitetes Feld oder als Eingabe in einem erstellten Prädiktor.

Klicken Sie auf die unterstrichenen Informationen unter **Felder**, um weitere Informationen in einer verknüpften Ansicht anzuzeigen. In der verknüpften Ansicht "Feldertabelle" erhalten Sie Informationen über **Ziel**, **Eingabemerkmale** und **Nicht verwendte Eingabemerkmale**. Weitere Informationen finden Sie im Thema „Feldertabelle” auf Seite 113. **Empfohlene Merkmale für den Einsatz in Analysen** werden in der verknüpften Ansicht "Vorhersagekraft" angezeigt. Weitere Informationen finden Sie im Thema „Vorhersagekraft” auf Seite 113.

Felder

In der Hauptansicht "Felder" werden die verarbeiteten Felder angezeigt sowie, ob ADP diese zur Verwendung in nachgelagerten Modellen empfiehlt. Sie können die Empfehlung für jedes Feld überschreiben, zum Beispiel, um erstellte Funktionen auszuschließen oder Funktionen einzuschließen, von denen ADP empfiehlt, sie auszuschließen. Wenn ein Feld transformiert wurde, können Sie entscheiden, ob Sie die vorgeschlagene Transformation akzeptieren oder die Originalversion verwenden möchten.

Die Felderansicht besteht aus zwei Tabellen, eine für das Ziel und eine für Prädiktoren, die entweder verarbeitet oder erstellt wurden.

Tabelle "Ziel"

Die Tabelle **Ziel** wird nur angezeigt, wenn in den Daten ein Ziel definiert wurde.

Die Tabelle enthält zwei Spalten:

- **Name.** Dies ist der Name oder die Beschriftung des Zielfelds. Der Originalname wird immer verwendet, auch wenn das Feld transformiert wurde.
- **Messniveau.** Hier erscheint das Symbol für das entsprechende Messniveau; fahren Sie mit der Maus über das Symbol, um eine Beschriftung (stetig, ordinal, nominal usw.) anzuzeigen, die die Daten beschreibt.

Wenn das Ziel transformiert wurde, gibt die Spalte **Messniveau** die endgültige transformierte Version an. *Hinweis:* Transformationen für das Ziel können nicht abgeschaltet werden.

Tabelle "Prädiktoren"

Die Tabelle **Prädiktoren** wird immer angezeigt. Jede Zeile der Tabelle repräsentiert ein Feld. Standardmäßig sind die Zeilen nach absteigender Vorhersagekraft sortiert.

Bei gewöhnlichen Merkmalen wird der Originalname immer als Zeilenname verwendet. Sowohl Original- als auch abgeleitete Versionen von Datums-/Zeitfeldern werden in der Tabelle (in getrennten Zeilen) angezeigt; die Tabelle enthält auch erstellte Prädiktoren.

Beachten Sie, dass transformierte Versionen von in der Tabelle angezeigten Feldern immer die Endversionen darstellen.

Standardmäßig werden in der Tabelle "Prädiktoren" nur empfohlene Felder angezeigt. Um die restlichen Felder anzuzeigen, wählen Sie das Feld **Nicht empfohlene Felder in Tabelle einschließen** über der Tabelle aus; diese Felder werden dann am Ende der Tabelle angezeigt.

Die Tabelle enthält folgende Spalten:

- **Zu verwendende Version.** Hier wird eine Dropdown-Liste angezeigt, die festlegt, ob ein Feld nachgelagert verwendet wird oder ob die vorgeschlagenen Transformationen verwendet werden sollen. Standardmäßig werden in der Dropdown-Liste die Empfehlungen wiedergegeben.

  Für gewöhnliche Prädiktoren, die transformiert wurden, stehen in der Dropdown-Liste drei Optionen zur Auswahl: **Transformiert**, **Original** und **Nicht verwenden**.

  Für nicht transformierte gewöhnliche Prädiktoren sind folgende Auswahlmöglichkeiten verfügbar: **Original** und **Nicht verwenden**.

  Für abgeleitete Datums-/Zeitfelder und erstellte Prädiktoren sind folgende Auswahlmöglichkeiten verfügbar: **Transformiert** und **Nicht verwenden**.

  Für Originaldatumsfelder ist die Dropdown-Liste inaktiviert und auf **Nicht verwenden** gesetzt.

  Hinweis: Für Prädiktoren mit Originalversionen und transformierten Versionen werden bei einem Wechsel zwischen den Versionen **Original** und **Transformiert** automatisch die Einstellungen **Messniveau** und **Vorhersagekraft** für diese Funktionen aktualisiert.
- **Name.** Jeder Feldname ist ein Link. Klicken Sie auf den Namen, um in der verknüpften Ansicht weitere Informationen über das Feld anzuzeigen. Weitere Informationen finden Sie im Thema „Felddetails" auf Seite 113.
- **Messniveau.** Hier erscheint das Symbol für den entsprechenden Datentyp; fahren Sie mit der Maus über das Symbol, um eine Beschriftung (stetig, ordinal, nominal usw.) anzuzeigen, die die Daten beschreibt.
- **Vorhersagekraft.** Die Vorhersagekraft wird nur für Felder angezeigt, die von ADP empfohlen werden. Diese Spalte wird nicht angezeigt, wenn kein Ziel definiert wurde. Die Vorhersagekraft reicht von 0 bis 1, wobei größere Werte "bessere" Einflussgrößen andeuten. Im Allgemeinen ist die Vorhersagekraft für den Vergleich von Einflussgrößen in einer ADP-Analyse nützlich, doch sollten Vorhersagekraft-Werte nicht in Analysen verglichen werden.

Aktionsübersicht

Bei jeder von der automatisierten Datenaufbereitung vorgenommenen Aktion werden Eingabeprädiktoren transformiert und/oder herausgefiltert. Felder, die in einer Aktion erhalten bleiben, werden in der nächsten verwendet. Die Felder, die bis zum letzten Schritt erhalten bleiben, werden dann für die Modellierung empfohlen, während Eingaben zu transformierten und erstellten Prädiktoren durch Filterung ausgeschlossen werden.

Die Aktionsübersicht ist eine einfache Tabelle, in der die von der ADP vorgenommenen Verarbeitungsaktionen aufgelistet sind. Klicken Sie auf den unterstrichenen Link **Aktion**, um in einer verknüpften Ansicht weitere Informationen über die durchgeführten Schritte anzuzeigen. Weitere Informationen finden Sie im Thema „Aktionsdetails” auf Seite 115.

Hinweis: Es werden nur die Original- und endgültigen transformierten Versionen jedes Felds angezeigt, jedoch keine während der Analyse verwendeten Zwischenversionen.

Vorhersagekraft

Wird standardmäßig bei der ersten Ausführung der Analyse angezeigt. Wenn Sie dagegen **Empfohlene Prädiktoren für den Einsatz in Analysen** in der Hauptansicht "Feldverarbeitungsübersicht" auswählen, zeigt das Diagramm die Vorhersagekraft der empfohlenen Prädiktoren an. Felder werden nach Vorhersagekraft sortiert, wobei das Feld mit dem höchsten Wert zuerst erscheint.

Bei transformierten Versionen gewöhnlicher Prädiktoren gibt der Feldname Ihre Suffixauswahl im Bereich "Feldnamen" auf der Registerkarte "Einstellungen" an, z. B.: *_transformiert*.

Die Symbole für das Messniveau werden nach den einzelnen Feldnamen angezeigt.

Die Vorhersagekraft jedes empfohlenen Prädiktors wird entweder aus einer linearen Regression oder einem Naive Bayes-Modell berechnet, abhängig davon, ob das Ziel stetig oder kategorial ist.

Feldertabelle

Die Feldertabelle wird angezeigt, wenn Sie in der Hauptansicht "Feldverarbeitungsübersicht" auf **Ziel**, **Prädiktoren** oder **Nicht verwendete Prädiktoren** klicken, und enthält eine einfache Tabelle, die die wichtigsten Funktionen auflistet.

Die Tabelle enthält zwei Spalten:

- **Name.** Der Name des Prädiktors.

  Für Ziele wird der Originalname oder die Originalbeschriftung des Felds verwendet, selbst wenn das Ziel transformiert wurde.

  Bei transformierten Versionen gewöhnlicher Prädiktoren gibt der Name Ihre Suffixauswahl im Bereich "Feldnamen" auf der Registerkarte "Einstellungen" an, z. B.: *_transformiert*.

  Bei aus Datums- und Zeitangaben abgeleiteten Feldern wird der Name der endgültigen transformierten Version verwendet, z. B.: *gebdat_jahre*.

  Bei erstellten Prädiktoren wird der Name des erstellten Prädiktors verwendet, z. B.: *Prädiktor1*.
- **Messniveau.** Hier erscheint das Symbol für den entsprechenden Datentyp.

  Für das Ziel gibt das **Messniveau** stets die transformierte Version wieder (wenn das Ziel transformiert wurde), z. B. bei einem Wechsel von ordinal (sortiertes Set) zu stetig (Bereich, Skala) oder umgekehrt.

Felddetails

Die Ansicht "Felddetails" wird angezeigt, wenn Sie auf **Name** in der Hauptansicht "Felder" klicken, und enthält Informationen über Verteilung, fehlende Werte und (falls zutreffend) Vorhersagekraft-Diagramme für das ausgewählte Feld. Außerdem wird der Verarbeitungsverlauf für das Feld und der Name des transformierten Felds angezeigt (falls zutreffend).

Für jedes Diagrammset werden nebeneinander zwei Versionen angezeigt, um das Feld mit und ohne angewendete Transformationen zu vergleichen. Wenn keine transformierte Version des Felds vorhanden ist, wird nur ein Diagramm für die Originalversion angezeigt. Für abgeleitete Datums- und Zeitfelder sowie erstellte Prädiktoren werden die Diagramme nur für den neuen Prädiktor angezeigt.

Hinweis: Wenn ein Feld wegen zu vieler Kategorien ausgeschlossen wurde, wird nur der Verarbeitungsverlauf angezeigt.

Verteilungsdiagramm

Die Verteilung stetiger Felder wird als Histogramm angezeigt, mit einer überlagerten Normalverteilungskurve und einer vertikalen Referenzlinie für den Mittelwert; kategoriale Felder werden als Balkendiagramm angezeigt.

Die Histogramme werden nach Standardabweichung und Schiefe beschriftet, allerdings wird Letztere nicht angezeigt, wenn die Anzahl der Werte kleiner gleich 2 oder die Varianz des originalen Felds kleiner als 10-20 ist.

Fahren Sie mit der Maus über das Diagramm, um entweder den Mittelwert für Histogramme oder die Zählung und den Prozentsatz der Gesamtzahl der Datensätze für Kategorien in Balkendiagrammen anzuzeigen.

Diagramm fehlender Werte

Kreisdiagramme vergleichen den Prozentsatz fehlender Werte mit und ohne angewendete Transformationen; die Diagrammbeschriftungen zeigen den Prozentsatz an.

Wenn ADP die Behandlung fehlender Werte durchgeführt hat, enthält das Kreisdiagramm nach der Transformation auch den Ersatzwert als Beschriftung, d. h. den anstelle von fehlenden Werten verwendeten Wert.

Fahren Sie mit der Maus über das Diagramm, um die Zählung der fehlenden Werte und den Prozentsatz der Gesamtzahl an Datensätzen anzuzeigen.

Vorhersagekraft-Diagramme

Für empfohlene Felder zeigen Balkendiagramme die Vorhersagekraft vor und nach der Transformation an. Wenn das Ziel transformiert wurde, steht die berechnete Vorhersagekraft in Beziehung zum transformierten Ziel.

Hinweis: Die Vorhersagekraftdiagramme werden nicht angezeigt, wenn kein Ziel definiert wurde oder wenn Sie in der Hauptansicht auf das Ziel klicken.

Fahren Sie mit der Maus über das Diagramm, um den Wert der Vorhersagekraft anzuzeigen.

Tabelle "Verarbeitungsverlauf"

Die Tabelle zeigt, wie die transformierte Version eines Felds abgeleitet wurde. Von ADP durchgeführte Aktionen werden in der Reihenfolge ihrer Ausführung aufgelistet. Bei bestimmten Schritten wurden jedoch u. U. mehrere Aktionen für ein spezielles Feld durchgeführt.

Hinweis: Die Tabelle wird nur für transformierte Felder angezeigt.

Die Informationen in der Tabelle erscheinen in zwei oder drei Spalten:

- **Aktion.** Der Name der Aktion. Zum Beispiel "Stetige Prädiktoren". Weitere Informationen finden Sie im Thema „Aktionsdetails” auf Seite 115.
- **Details.** Die Liste der durchgeführten Verarbeitung. Zum Beispiel "Zu Standardeinheiten transformieren".
- **Funktion.** Diese Spalte erscheint nur bei erstellten Prädiktoren und zeigt die lineare Kombination von Eingabefeldern an, z. B. 0,06*Alter + 1,21*Größe.

Aktionsdetails

Die verknüpfte Ansicht "Aktionsdetails" wird angezeigt, wenn Sie in der Hauptansicht "Aktionsübersicht" auf den unterstrichenen Link **Aktion** klicken, und enthält sowohl aktionsspezifische als auch allgemeine Informationen über jeden durchgeführten Verarbeitungsschritt. Die aktionsspezifischen Informationen erscheinen stets zuerst.

Für jede Aktion wird die Beschreibung als Titel im oberen Bereich der verknüpften Ansicht verwendet. Die aktionsspezifischen Informationen erscheinen unter dem Titel und enthalten u. U. Details zur Anzahl abgeleiteter Prädiktoren, zu umgewandelten Feldern, zu Zieltransformationen, zu zusammengeführten oder neu sortierten Kategorien und zu erstellten oder ausgeschlossenen Prädiktoren.

Bei der Verarbeitung jeder Aktion kann sich die für die Verarbeitung verwendete Anzahl an Prädiktoren ändern, wenn beispielsweise Prädiktoren ausgeschlossen oder zusammengeführt werden.

Hinweis: Wenn eine Aktion inaktiviert oder kein Ziel angegeben wurde, wird anstelle der Aktionsdetails eine Fehlernachricht angezeigt, wenn Sie in der Hauptansicht "Aktionsübersicht" auf die Aktion klicken.

Es gibt neun mögliche Aktionen, davon sind allerdings nicht alle notwendigerweise für jede Analyse aktiv.

Tabelle "Textfelder"

Die Tabelle zeigt folgende Anzahl:
- Entfernte leere nachstehende Werte.
- Von der Analyse ausgeschlossene Prädiktoren.

Tabelle "Datums- und Uhrzeitprädiktoren"

Die Tabelle zeigt folgende Anzahl:
- Aus Datums- und Uhrzeitprädiktoren abgeleitete Dauer.
- Datums- und Uhrzeitelemente.
- Insgesamt abgeleitete Datums- und Uhrzeitprädiktoren.

Das Referenzdatum oder die -uhrzeit wird als Fußnote angezeigt, wenn eine Datumsdauer berechnet wurde.

Tabelle "Screening von Prädiktoren"

Die Tabelle zeigt die Anzahl folgender von der Verarbeitung ausgeschlossener Prädiktoren:
- Konstanten.
- Prädiktoren mit zu vielen fehlenden Werten.
- Prädiktoren mit zu vielen Fällen in einer einzelnen Kategorie.
- Nominale Felder (Sets) mit zu vielen Kategorien.
- Insgesamt ausgeschlossene Prädiktoren.

Tabelle "Messniveau überprüfen"

Die Tabelle zeigt die Anzahl umgewandelter Felder und teilt sich wie folgt auf:
- In stetige Feldern umgewandelte ordinale Felder (sortierte Sets).
- In ordinale Felder umgewandelte stetige Felder.
- Anzahl an Umwandlungen insgesamt.

Wenn keine Eingabefelder (Ziel oder Prädiktoren) stetig oder ordinal waren, wird dies in einer Fußnote angezeigt.

Tabelle "Ausreißer"

Die Tabelle zeigt, ob und wie Ausreißer behandelt wurden.

- Entweder die Anzahl stetiger Felder, für die Ausreißer gefunden und entfernt wurden, oder die Anzahl stetiger Felder, für die Ausreißer gefunden und als fehlend eingestuft wurden, je nach Ihren Einstellungen im Feld "Eingaben & Ziel vorbereiten" auf der Registerkarte "Einstellungen".
- Die Anzahl stetiger Felder, die ausgeschlossen wurden, weil sie nach der Ausreißer-Behandlung konstant waren.

Der Ausreißertrennwert wird in einer Fußnote vermerkt. Eine weitere Fußnote wird angezeigt, wenn keine Eingabefelder (Ziel oder Prädiktoren) stetig waren.

Tabelle "Fehlende Werte"

Die Tabelle zeigt die Anzahl an Feldern, in denen fehlende Werte ersetzt wurden, und teilt sich wie folgt auf:

- Ziel. Diese Zeile wird nicht angezeigt, wenn kein Ziel angegeben wurde.
- Prädiktoren. Dies teilt sich weiter auf in Anzahl an "nominal (Set)", "ordinal (sortiertes Set)" und "stetig".
- Die gesamte Anzahl ersetzter fehlender Werte.

Tabelle "Ziel"

Die Tabelle zeigt wie folgt, ob das Ziel tranformiert wurde:

- Box-Cox-Transformation in Normalverteilung. Dies teilt sich weiter in Spalten auf, die die angegebenen Kriterien (Mittelwert und Standardabweichung) und Lambda zeigen.
- Zielkategorien zur Verbesserung der Stabilität neu sortiert.

Tabelle "Kategoriale Prädiktoren"

Die Tabelle zeigt folgende Anzahl kategorialer Prädiktoren:

- Wessen Kategorien wurden zur Verbesserung der Stabilität in aufsteigender Reihenfolge neu sortiert.
- Wessen Kategorien wurden zur Maximierung des Zielzusammenhangs zusammengeführt.
- Wessen Kategorien wurden zur Behandlung dünn besetzter Kategorien zusammengeführt.
- Wegen niedrigem Zielzusammenhang ausgeschlossen.
- Ausgeschlossen, weil nach der Zusammenführung konstant.

Wenn es keine kategorialen Prädiktoren gab, wird dies durch eine Fußnote vermerkt.

Tabelle "Stetige Prädiktoren"

Es gibt zwei Tabellen. Die erste zeigt eine der folgenden Transformationen:

- Zu Standardeinheiten transformierte Prädiktorwerte. Zusätzlich werden hier die Anzahl transformierter Prädiktoren, der angegebene Mittelwert und die Standardabweichung angezeigt.
- Einem gemeinsamen Bereich zugeordnete Prädiktorwerte. Zusätzlich werden hier die Anzahl der mithilfe der min./max. Transformation transformierten Prädiktoren sowie die angegebenen Mindest- und Höchstwerte angezeigt.
- Klassierte Prädiktorwerte und die Anzahl klassierter Prädiktoren.

Die zweite Tabelle enthält Informationen über die Erstellung von Prädiktorbereichen, die als Anzahl folgender Prädiktoren angezeigt werden:

- Erstellt.
- Wegen niedrigem Zielzusammenhang ausgeschlossen.
- Ausgeschlossen, weil nach der Klassierung konstant.
- Ausgeschlossen, weil nach der Erstellung konstant.

Wenn keine stetigen Prädiktoren eingegeben wurden, wird dies durch eine Fußnote vermerkt.

Erzeugen eines Ableitungsknotens

Wenn Sie einen Ableitungsknoten erstellen, wendet dieser die inverse Zieltransformation auf das Wertfeld an. Standardmäßig gibt der Knoten den Namen des Wertfelds ein, das mithilfe eines Automodeler-Knotens (zum Beispiel Auto Classifier oder Auto Numeric) oder dem Ensemble-Knoten erstellt werden würde. Wenn ein metrisches (Bereichs-)Ziel transformiert wurde, wird das Wertfeld in transformierten Einheiten angezeigt, zum Beispiel log($) anstelle von $. Um die Ergebnisse interpretieren und verwenden zu können, müssen Sie den vorhergesagten Wert wieder in das ursprüngliche metrische Maß zurückkonvertieren.

Hinweis: Sie können einen Ableitungsknoten nur dann generierern, wenn der ADP-Knoten eine Analyse enthält, die ein Bereichsziel neu skaliert (d. h., die Box-Cox-Neuskalierung ist im Feld "Eingaben und Ziel vorbereiten" ausgewählt). Sie können keinen Ableitungsknoten erzeugen, wenn das Ziel kein Bereich ist oder wenn die Box-Cox-Neuskalierung nicht ausgewählt ist.

Der Ableitungsknoten wird im Mehrfachmodus erstellt und verwendet @FIELD im Ausdruck, damit Sie das transformierte Ziel gegebenenfalls hinzufügen können. Es können zum Beispiel folgende Informationen verwendet werden:

- Zielfeldname: response
- Feldname des transformierten Ziels: response_transformed
- Wertfeldname: $XR-response_transformed

Der Ableitungsknoten würde folgendes neues Feld erstellen: $XR-response_transformed_inverse.

Hinweis: Wenn Sie keinen Automodeler- oder Ensemble-Knoten verwenden, müssen Sie den Ableitungsknoten so bearbeiten, dass dieser das korrekte Wertfeld für Ihr Modell transformiert.

Normalisierte stetige Ziele

Wenn Sie das Kontrollkästchen **Stetiges Ziel mit einer Box-Cox-Transformation neu skalieren** im Feld "Eingaben & Ziel vorbereiten" auswählen, wird dadurch standardmäßig das Ziel transformiert und Sie können ein neues Feld erstellen, das für Ihre Modellerstellung als Ziel fungiert. Wenn zum Beispiel Ihr ursprüngliches Ziel *Antwort* war, heißt das neue Ziel *Antwort_transformiert*; dem ADP-Knoten nachgelagerte Modelle nehmen dieses Ziel automatisch auf.

Dies kann jedoch je nach ursprünglichem Ziel Probleme verursachen. Wenn das Ziel zum Beispiel *Alter* war, werden die Werte des neuen Ziels nicht *Jahre*, sondern eine transformierte Version *Jahre* sein. Somit können Sie die Werte nicht betrachten und interpretieren, da sie nicht in erkennbaren Einheiten vorliegen. In diesem Fall können Sie eine inverse Transformation anwenden, durch die Ihre transformierten Einheiten wieder in ihr ursprünglich gewünschtes Format zurückkonvertiert werden. Gehen Sie dazu wie folgt vor:

1. Klicken Sie zunächst auf **Daten analysieren**, um die ADP-Analyse auszuführen, und wählen Sie dann *Ableitungsknoten* im Menü *Erzeugen* aus.
2. Setzen Sie den Ableitungsknoten nach Ihrem Nugget im Modellerstellungsbereich.

Der Ableitungsknoten stellt die ursprünglichen Dimensionen des Wertfelds wieder her, sodass die Vorhersage in den ursprünglichen Werten *Jahre* erscheint.

Der Ableitungsknoten transformiert standardmäßig das durch ein Automodeler- oder Ensemblemodell erzeugte Wertfeld. Wenn Sie ein einzelnes Modell erstellen, müssen Sie den Ableitungsknoten so bearbeiten, dass dieser aus Ihrem tatsächlichen Wertfeld ableitet. Wenn Sie Ihr Modell evaluieren möchten, sollten Sie das transformierte Ziel dem Feld **Ableiten aus** im Ableitungsknoten hinzufügen. Dadurch wird die gleiche inverse Transformation auf das Ziel angewendet und jeder nachgelagerte Evaluierungs- oder Analyseknoten wird die transformierten Daten korrekt verwenden, vorausgesetzt, Sie stellen diese Knoten so ein, dass sie Feldnamen anstelle von Metadaten verwenden.

Wenn Sie zudem den Originalnamen wiederherstellen möchten, können Sie einen Filterknoten verwenden, um das eventuell noch vorhandene ursprüngliche Zielfeld zu entfernen, und die Ziel- und Wertfelder umbenennen.

Typknoten

Die Feldeigenschaften können in einem Quellenknoten oder in einem separaten Typknoten angegeben werden. Die Funktionsweise ist bei beiden Knoten ähnlich. Folgende Eigenschaften stehen zur Verfügung:

- **Feld.** Doppelklicken Sie auf einen beliebigen Feldnamen, um Wert- und Feldbeschriftungen für Daten in IBM SPSS Modeler anzugeben. So können aus IBM SPSS Statistics beispielsweise importierte Feldmetadaten hier angezeigt oder geändert werden. Auf ähnliche Weise können Sie auch neue Beschriftungen für Felder und ihre Werte erstellen. Die Beschriftungen, die Sie hier angeben, werden überall in IBM SPSS Modeler angezeigt, je nach der von Ihnen im Dialogfeld "Streameigenschaften" getroffenen Auswahl.
- **Messung.** Dies ist das Messniveau, das zur Beschreibung der Eigenschaften von Daten in einem bestimmten Feld verwendet wird. Wenn alle Details eines Felds bekannt sind, wird es als **vollständig instanziiert** bezeichnet. Weitere Informationen finden Sie im Thema „Messniveaus” auf Seite 119.

  Hinweis: Das Messniveau eines Felds ist etwas anderes als sein Speichertyp, der angibt, ob die Daten als Zeichenfolge, ganze Zahl, reelle Zahl, Datum, Zeit oder Zeitmarke gespeichert werden sollen.
- **Werte.** In dieser Spalte können Sie Optionen zum Lesen von Datenwerten aus dem Dataset auswählen oder die Option **Angeben** verwenden, um Messniveaus und Werte in einem separaten Dialogfeld anzugeben. Sie können auch Felder übergeben, ohne ihre Werte zu lesen. Weitere Informationen finden Sie im Thema „Datenwerte” auf Seite 122.
- **Fehlend.** Wird verwendet, um anzugeben, wie fehlende Werte für das Feld behandelt werden. Weitere Informationen finden Sie im Thema „Definieren fehlender Werte” auf Seite 125.
- **Überprüfen.** In dieser Spalte können Sie Optionen festlegen, um sicherzustellen, dass die Feldwerte den angegebenen Werten oder Bereichen entsprechen. Weitere Informationen finden Sie im Thema „Überprüfen von Typenwerten” auf Seite 125.
- **Rolle.** Wird verwendet, um Modellierungsknoten mitzuteilen, ob es sich bei Feldern um **Eingabefelder** (Prädiktorfelder) oder **Zielfelder** (vorhergesagte Felder) für einen Maschinenlernprozess handelt. **Beides** und **Keine** sind auch verfügbare Rollen, zusammen mit **Partition**, das ein Feld bezeichnet, das für die Aufteilung von Datensätzen in separate Stichproben zu Training-, Test- und Validierungszwecken verwendet wird. Der Wert **Aufteilung** gibt an, dass für jeden möglichen Wert des Felds separate Modelle erstellt werden. Weitere Informationen finden Sie im Thema „Festlegen der Feldrolle” auf Seite 126.

Mehrere andere Optionen können im Fenster "Typknoten" angegeben werden:

- Mithilfe der Schaltfläche **Felder mit nur 1 Wert ignorieren** im Menü "Extras" können Sie festlegen, dass Felder mit nur einem Wert ignoriert werden sollen, sobald ein Typknoten als Instanz erstellt wurde (entweder über Ihre Spezifikationen, aus eingelesenen Werten oder durch die Ausführung des Streams). Bei Auswahl von "Felder mit nur 1 Wert ignorieren" werden Felder mit nur einem Wert automatisch ignoriert.

- Mithilfe der Schaltfläche **Große Sets ignorieren** im Menü "Extras" knnen Sie festlegen, dass große Sets ignoriert werden sollen, sobald ein Typknoten als Instanz erstellt wurde. Bei Auswahl von "Große Sets ignorieren" werden automatisch Sets mit sehr vielen Mitgliedern ignoriert.
- Mithilfe der Schaltfläche **Stetige Ganzzahlen in Ordinalzahlen umwandeln** im Menü "Extras" können Sie Ganzzahlenbereiche in Sets umwandeln, sobald ein Typknoten als Instanz erstellt wurde. Weitere Informationen finden Sie im Thema „Stetige Daten umwandeln” auf Seite 121.
- Mithilfe der entsprechenden Schaltfläche im Menü "Extras" können Sie einen Filterknoten zum Verwerfen der ausgewählten Felder generieren.
- Mithilfe der Umschalttaste mit der Sonnenbrille können Sie den Standard für alle Felder auf "Lesen" oder "Übergeben" setzen. Die Registerkarte "Typen" im Quellenknoten übergibt Felder standardmäßig, während der Typknoten standardmäßig Werte liest.
- Mit der Schaltfläche **Werte löschen** können Sie die in diesem Knoten vorgenommenen Änderungen an den Feldwerten löschen (nicht übernommene Werte) und die Werte aus den vorgeordneten Operationen erneut lesen. Diese Operation dient zum Zurücksetzen von Änderungen, die Sie für bestimmte, aufwärts liegende Felder vorgenommen haben.
- Mit der Schaltfläche **Alle Werte löschen** können Sie die Werte für **alle** in den Knoten eingelesenen Felder zurücksetzen. Diese Option setzt die Spalte *Werte* für alle Felder effektiv auf **Lesen**. Mit dieser Option können Sie die Werte für alle Felder zurücksetzen und die Werte und Typen aus den vorgeordneten Operationen erneut lesen.
- Über das Kontextmenü (**Kopieren**) können Sie Attribute aus einem Feld in ein anderes kopieren. Weitere Informationen finden Sie im Thema „Kopieren von Typattributen” auf Seite 127.
- Mithilfe der Option **Nicht verwendete Feldeinstellungen anzeigen** können Sie Typeinstellungen für Felder anzeigen, die nicht mehr in den Daten vorliegen oder die zuvor mit diesem Typknoten verbunden waren. Dies ist sinnvoll, wenn Sie einen Typknoten für Datasets, die sich geändert haben, erneut verwenden möchten.

Messniveaus

Das Messniveau (früher "Datentyp" oder "Verwendung" genannt) beschreibt die Nutzung der Datenfelder in IBM SPSS Modeler. Das Messniveau kann auf der Registerkarte "Typen" eines Quellenknotens oder Typknotens festgelegt werden. Beispiel: Sie möchten das Messniveau für ein Feld ganzer Zahlen mit den Werten 1 und 0 auf *Flag* setzen. Das bedeutet normalerweise, dass 1=*True* und 0=*False* ist.

Speicherung versus Messung. Das Messniveau eines Felds unterscheidet sich dessen Speichertyp, der angibt, ob die Daten als Zeichenfolge, ganze Zahl, reelle Zahl, Datum, Zeit oder Zeitmarke gespeichert werden sollen. Während die Datentypen mithilfe eines Typknotens an jeder beliebigen Stelle im Stream geändert werden können, muss der Speichertyp beim Lesen der Daten in IBM SPSS Modeler stets an der Quelle festgelegt werden (kann jedoch später mithilfe einer Konvertierungsfunktion geändert werden). Weitere Informationen finden Sie im Thema „Festlegen von Feldspeicher und Formatierung” auf Seite 24.

Bei einigen Modellierungsknoten werden die zulässigen Messniveautypen für die Eingabe- und Zielfelder durch Symbole auf der Registerkarte "Felder" angegeben.

Messniveausymbole

Tabelle 15. Messniveausymbole

Symbol	Messniveau
	Default
	Stetig
	Kategorial

Tabelle 15. Messniveausymbole (Forts.)

Symbol	Messniveau
	Flag
	Nominal
	Ordinal
	Ohne Typ

Die folgenden Messniveaus stehen zur Verfügung:

- **Standard.** Daten, deren Speichertyp und Werte unbekannt sind (z. B. weil sie noch nicht gelesen wurden), werden als **<Standard>** angezeigt.
- **Stetig.** Wird zur Beschreibung numerischer Werte verwendet, beispielsweise des Bereichs 0-100 oder 0,75-1,25. Ein stetiger Wert kann eine ganze Zahl, eine reelle Zahl oder ein Datum/eine Uhrzeit sein.
- **Kategorial.** Wird für Zeichenfolgewerte verwendet, wenn eine exakte Anzahl unterschiedlicher Werte nicht bekannt ist. Dies ist ein Datentyp **ohne Instanz**, was bedeutet, dass nicht alle möglichen Informationen über Speicherung und Verwendung der Daten bereits bekannt sind. Nach dem Lesen der Daten ist das Messniveau *Flag*, *Nominal* oder *Ohne Typ*(abhängig von der maximalen Anzahl an Mitgliedern für nominale Felder, die im Dialogfeld "Streameigenschaften" angegeben wurde).
- **Flag.** Wird für Daten mit zwei unterschiedlichen Werten verwendet, die auf das Vorhandensein bzw. Nichtvorhandensein eines Merkmals hinweisen, wie etwa `true` (wahr) und `false` (falsch), `Yes` (Ja) und `No` (Nein) oder `0` und `1`. Die verwendeten Werte können abweichen, aber ein Wert muss immer als "wahr" und der andere als "falsch" festgelegt sein. Die Daten können als Text, ganze Zahl, reelle Zahl, Datum/Uhrzeit oder Zeitmarke dargestellt sein.
- **Nominal.** Wird verwendet, um Daten mit mehreren unterschiedlichen Werten zu beschreiben, von denen jeder als Mitglied eines Sets behandelt wird, beispielsweise `klein/mittel/groß`. Nominale Daten können jeden beliebigen Speichertyp aufweisen: "Numerisch", "Zeichenfolge" oder "Datum/Uhrzeit". Hinweis: Durch das Setzen des Messniveaus auf *Nominal* werden nicht automatisch die Werte auf Zeichenfolgenspeicherung geändert.
- **Ordinal.** Wird zur Beschreibung von Daten mit mehreren unterschiedlichen Werten verwendet, die eine natürliche Reihenfolge aufweisen. Gehaltskategorien oder Zufriedenheitsbewertungen beispielsweise können als ordinale Daten klassifiziert werden. Die Reihenfolge wird durch die natürliche Sortierfolge der Datenelemente definiert. So ist `1, 3, 5` die Standardsortierreihenfolge für eine Menge von ganzen Zahlen, während `HOCH, NIEDRIG, NORMAL` (aufsteigende alphabetische Reihenfolge) die Reihenfolge für eine Menge von Zeichenfolgen ist. Mit dem ordinalen Messniveau können Sie eine Menge kategorialer Daten als ordinale Daten festlegen - zum Zwecke der Visualisierung, Modellerstellung und zum Export in andere Anwendungen, beispielsweise IBM SPSS Statistics, die ordinale Daten als gesonderten Typ erkennen. Sie können ein ordinales Feld überall dort verwenden, wo sich ein nominales Feld verwenden lässt. Außerdem können Felder jedes beliebigen Speichertyps ("Reelle Zahl", "Ganze Zahl", "Zeichenfolge", "Datum", "Zeit" usw.) als "ordinal" definiert werden.
- **Ohne Typ.** Wird für Daten verwendet, die keinem der oben angegebenen Typen entsprechen, für Felder mit einem einzelnen Wert bzw . für nominale Daten, in denen das Set mehr Mitglieder als das definierte Maximum enthält. Dieser Typ ist auch sinnvoll in Fällen, in denen das Messniveau ansonsten ein Set mit zu vielen Mitgliedern wäre (beispielsweise eine Kontonummer). Bei Auswahl von **Ohne Typ** für ein Feld wird für die Rolle automatisch **Keine** festgelegt, wobei **Datensatz-ID** die einzige Alternative ist. Die Standardmaximalgröße für Sets liegt bei 250 eindeutigen Werten. Diese Zahl kann auf der Registerkarte "Optionen" des Dialogfelds "Streameigenschaften" (Zugriff über das Menü "Tools") angepasst oder inaktiviert werden.

Sie können die Messniveaus wahlweise manuell festlegen oder auch die Daten durch die Software einlesen und dann das Messniveau auf der Grundlage der eingelesenen Werte automatisch bestimmen lassen.

Alternativ können Sie bei mehreren stetigen Datenfeldern, die als kategoriale Daten behandelt werden sollen, eine Option auswählen, um sie umzuwandeln. Weitere Informationen finden Sie im Thema „Stetige Daten umwandeln".

So verwenden Sie die automatische Typfestlegung:

1. Setzen Sie in einem Typknoten oder auf der Registerkarte "Typen" eines Quellenknotens die Spalte *Werte* für die gewünschten Felder auf **<Lesen>**. Dadurch werden die Metadaten für alle abwärts liegenden Knoten verfügbar. Mit den Sonnenbrillenschaltflächen im Dialogfeld können Sie schnell und einfach alle Felder auf **<Lesen>** oder **<Übergeben>** setzen.
2. Klicken Sie auf **Werte lesen**, um sofort die Werte aus der Datenquelle zu lesen.

So legen Sie das Messniveau für ein Feld manuell fest:

1. Wählen Sie ein Feld in der Tabelle aus.
2. Wählen Sie in der Dropdown-Liste in der Spalte *Messung* ein Messniveau für das Feld aus.
3. Alternativ können Sie mit Strg-A oder Strg-Klicken mehrere Felder auswählen, bevor Sie in der Dropdown-Liste ein Messniveau auswählen.

Stetige Daten umwandeln

Die Behandlung von kategorialen Daten als stetige Daten kann schwerwiegende Auswirkungen auf die Qualität eines Modells haben, besonders wenn es sich dabei um das Zielfeld handelt. So könnte beispielsweise ein Regressionsmodell anstelle eines binären Modells erzeugt werden. Um dies zu vermeiden, können Sie Ganzzahlenbereiche in kategoriale Typen wie *Ordinal* oder *Flag* umwandeln.

1. Wählen Sie aus dem Menüfeld "Operationen und Generieren" (mit dem Toolsymbol) die Option **Stetige Ganzzahlen in Ordinalzahlen umwandeln**. Das Dialogfeld "Umwandlungswerte" wird angezeigt.
2. Geben Sie die Größe des Bereichs an, der automatisch umgewandelt wird. Dies gilt für jeden Bereich bis zu der von Ihnen angegebenen Größe.
3. Klicken Sie auf **OK**. Die betroffenen Bereiche werden in *Flag* oder *Ordinal* umgewandelt und auf der Registerkarte "Typen" des Typknotens angezeigt.

Ergebnisse der Umwandlung

- Wenn ein *stetiges* Feld mit Speichertyp "Ganze Zahl" in *Ordinal* umgewandelt wird, werden die unteren und oberen Werte erweitert, damit alle ganzzahligen Werte mit einbezogen werden. Wenn der Bereich die Werte 1 und 5 umfasst, besteht das Werteset aus 1, 2, 3, 4 und 5.
- Wenn das *stetige* Feld in *Flag* umgewandelt wird, werden der obere und untere Wert zum Wahr- und Falsch-Wert des Flagfeldes.

Was ist Instanziierung?

Instanziierung ist der Prozess des Lesens oder Angebens von Informationen, beispielsweise des Speichertyps und der Werte für ein Datenfeld. Zur Optimierung der Systemressourcen handelt es sich bei der Instanziierung um einen benutzergesteuerten Prozess - Sie weisen die Software an, Datenwerte zu lesen, indem Sie die entsprechenden Optionen auf der Registerkarte "Typen" in einem Quellenknoten angeben bzw. indem Sie Daten durch einen Typknoten laufen lassen.

- Daten mit unbekannten Typen werden auch als **ohne Instanz** bezeichnet. Daten mit unbekanntem Speichertyp und unbekannten Werten werden in der Spalte *Messung* der Registerkarte "Typen" als **<Standard>** angezeigt.
- Wenn ein Teil der Informationen über den Speichertyp eines Felds vorliegt, beispielsweise "Zeichenfolge" oder "Numerisch", werden die betreffenden Daten als **teilweise instanziiert** bezeichnet. **Kategorial** oder **Stetig** sind teilweise instanziierte Messniveaus. **Kategorial** beispielsweise gibt an, dass das Feld symbolisch ist, jedoch nicht bekannt ist, ob es nominal, ordinal oder Flag ist.

- Wenn alle Details über einen Typ bekannt sind, einschließlich der Werte, wird ein **vollständig instanziiertes** Messniveau - nominal, ordinal, Flag oder stetig - in dieser Spalte angezeigt. *Hinweis:* Der Typ *stetig* wird sowohl für teilweise als auch für vollständig instanziierte Datenfelder verwendet. Bei stetigen Daten kann es sich entweder um ganze Zahlen oder um reelle Zahlen handeln.

Während der Ausführung eines Datenstreams mit einem Typknoten werden Typen ohne Instanz sofort auf der Grundlage der ursprünglichen Datenwerte teilweise instanziiert. Sobald alle Daten den Knoten durchlaufen haben, werden alle Daten vollständig instanziiert, es sei denn, einige Werte wurden auf **<Übergeben>** gesetzt. Wenn die Ausführung unterbrochen wird, bleiben die Daten teilweise instanziiert. Sobald die Registerkarte "Typen" als Instanz erstellt wurde, sind die Werte eines Felds an dieser Stelle im Stream statisch. Das bedeutet, dass Änderungen weiter oben im Stream sich nicht auf die Werte eines bestimmten Felds auswirken, selbst wenn der Stream erneut ausgeführt wird. Um die Werte auf der Grundlage neuer Daten oder weiterer Bearbeitungen zu ändern bzw. zu aktualisieren, müssen Sie sie auf der Registerkarte "Typen" bearbeiten oder den Wert für ein Feld auf**<Lesen>** oder **<Lesen +>** setzen.

Zeitpunkt der Instanziierung

Im Allgemeinen ist bei nicht allzu großen Datasets und wenn keine Felder später im Stream hinzugefügt werden sollen, eine Instanziierung am Quellenknoten die praktischste Methode. Die Instanziierung in einem separaten Typknoten ist jedoch in folgenden Fällen ratsam:

- Das Dataset ist groß und der Stream filtert ein Subset vor dem Typknoten.
- Im Stream wurden Daten gefiltert.
- Im Stream wurden Daten zusammengeführt oder angehängt.
- Während der Verarbeitung werden neue Datenfelder abgeleitet.

Datenwerte

Mithilfe der Spalte *Werte* der Registerkarte "Typen" können Sie die Werte automatisch aus den Daten einlesen oder Sie können Messniveaus und Werte in einem separaten Dialogfeld angeben.

Die in der Dropdown-Liste "Werte" verfügbaren Optionen stellen Anweisungen für automatische Typfestlegung bereit, wie in der folgenden Tabelle gezeigt.

Tabelle 16. Anweisungen für automatische Typfestlegung

Option	Funktion
<Lesen>	Die Daten werden gelesen, wenn der Knoten ausgeführt wird.
<Lesen +>	Die Daten werden gelesen und an die aktuellen Daten angehängt (sofern vorhanden).
<Übergeben>	Es werden keine Daten gelesen.
<Aktuell>	Aktuelle Werte werden beibehalten.
Angeben...	Ein separates Dialogfeld wird gestartet, in dem Sie Werte und Optionen für Messniveaus angeben können.

Durch Ausführen eines Typknotens oder Klicken auf **Werte lesen** werden Werte aus Ihrer Datenquelle auf der Grundlage ihrer Auswahl automatisch einem Typ zugewiesen und gelesen. Diese Werte können auch mithilfe der Option "Angeben" oder durch Doppelklicken in eine Zelle in der Spalte *Feld* manuell angegeben werden.

Nachdem Sie die Änderungen für die Felder im Typknoten vorgenommen haben, können Sie die Werteinformationen mithilfe der folgenden Schaltflächen in der Symbolleiste des Dialogfelds zurücksetzen:

- Mit der Schaltfläche **Werte löschen** können Sie die in diesem Knoten vorgenommenen Änderungen an den Feldwerten löschen (nicht übernommene Werte) und die Werte aus den vorgeordneten Operationen erneut lesen. Diese Operation dient zum Zurücksetzen von Änderungen, die Sie für bestimmte, aufwärts liegende Felder vorgenommen haben.

- Mit der Schaltfläche **Alle Werte löschen** können Sie die Werte für **alle** in den Knoten eingelesenen Felder zurücksetzen. Diese Option setzt die Spalte *Werte* für alle Felder effektiv auf **Lesen**. Mit dieser Option können Sie die Werte für alle Felder zurücksetzen und die Werte und Messniveaus aus den vorgeordneten Operationen erneut lesen.

Verwenden des Dialogfelds "Werte"

Durch Klicken auf die Spalte *Werte* oder *Fehlend* der Registerkarte "Typen" wird eine Dropdown-Liste mit vordefinierten Werten angezeigt. Durch Auswahl der Option *Angeben* wird ein gesondertes Dialogfeld geöffnet, in dem Sie die Optionen zum Lesen, Angeben, Beschriften und Behandeln der Werte für das ausgewählte Feld festlegen können.

Viele der Steuerelemente sind für alle Datentypen gleich. Diese gemeinsamen Steuerelemente werden hier erörtert.

Messung. Zeigt das aktuell ausgewählte Messniveau an. Sie können die Einstellung so ändern, wie die Daten verwendet werden sollen. Beispiel: Wenn ein Feld mit dem Titel *Tag_der_Woche* Zahlen enthält, die für die einzelnen Tage stehen, können Sie dieses in nominale Daten ändern, um einen Verteilungsknoten zu erstellen, der jede Kategorie einzeln untersucht.

Speichertyp. Zeigt den Speichertyp an, sofern dieser bekannt ist. Speichertypen werden vom gewählten Messniveau nicht beeinflusst. Zum Ändern des Speichertyps können Sie die Registerkarte "Daten" in den Quellenknoten "Datei (fest)" und "Datei (var.)" oder eine Konvertierungsfunktion in einem Füllerknoten verwenden.

Modellfeld. Bei Feldern, die beim Scoren eines Modellnuggets generiert wurden, können auch Details zum Modell-Feld angezeigt werden. Dazu gehören der Name des Zielfelds sowie die Rolle des Felds bei der Modellierung (ob es sich um einen vorhergesagten Wert, eine Wahrscheinlichkeit, Neigung usw. handelt).

Werte. Dient zur Auswahl einer Methode zur Bestimmung der Werte für das ausgewählte Feld. Die hier vorgenommene Auswahl setzt alle Optionen außer Kraft, die Sie zuvor in der Spalte *Werte* des Dialogfelds "Typknoten" vorgenommen haben. Zu den Auswahlmöglichkeiten zum Lesen von Werten gehören:

- **Aus Daten lesen.** Wählen Sie diese Option aus, um Daten lesen zu lassen, wenn der Knoten ausgeführt wird. Diese Option entspricht **<Lesen>**.
- **Übergeben.** Wählen Sie diese Option aus, wenn keine Daten für das aktuelle Feld gelesen werden sollen. Diese Option entspricht **<Übergeben>**.
- **Werte und Beschriftungen angeben.** Die Optionen hier werden zur Angabe von Werten und Beschriftungen für das ausgewählte Feld verwendet. In Verbindung mit der Werteprüfung ermöglicht diese Option die Angabe von Werten auf der Grundlage Ihrer Kenntnisse über das aktuelle Feld. Diese Option aktiviert eindeutige Steuerelemente für jeden Feldtyp. Die Optionen für Werte und Beschriftungen werden einzeln in den nachfolgenden Themenabschnitten behandelt. *Hinweis:* Werte und Beschriftungen können nicht für ein Feld angegeben werden, dessen Messniveau *Ohne Typ* oder <*Standard*> lautet.
- **Werte aus Daten erweitern.** Wählen Sie diese Option aus, um die aktuellen Daten mit den hier eingegebenen Werten zu ergänzen. Wenn beispielsweise *Feld_1* den Bereich (0,10) aufweist und Sie den Wertebereich (8,16) eingeben, wird der Bereich durch Hinzufügen von 16 erweitert, wobei der ursprüngliche Mindestwert nicht verändert wird. Der neue Bereich ist also (0,16). Durch Auswahl dieser Option wird die Option für die automatische Typfestlegung auf **<Lesen+>** gesetzt.

Werte prüfen. Dient zur Auswahl einer Methode, mit der erzwungen wird, dass die Werte den angegebenen stetigen Werten, Flagwerten oder nominalen Werten entsprechen. Diese Option entspricht der Spalte *Überprüfen* im Dialogfeld "Typknoten" und die hier vorgenommenen Einstellungen setzen die Einstellungen im Dialogfeld außer Kraft. In Verbindung mit der Option "Werte angeben" ermöglicht die Wertprüfung den Abgleich der Werte in den Daten mit den erwarteten Werten. Beispiel: Wenn Sie die Werte als "1, 0" angeben und anschließend die Option **Verwerfen** verwenden, können Sie alle Datensätze verwerfen, die andere Werte aufweisen als 1 oder 0.

Fehlende Werte definieren. Wählen Sie diese Option aus, um die unten angegebenen Steuerelemente zu aktivieren, mit denen Sie fehlende Werte oder Leerzeichen in Ihren Daten angeben können.

- **Tabelle fehlender Werte.** Ermöglicht die Definition bestimmter Werte (z. B. 99 oder 0) als Leerstellen. Der Wert sollte für den Speichertyp des Felds geeignet sein.
- **Bereich.** Wird verwendet, um den Bereich fehlender Werte anzugeben, beispielsweise der Altersbereich 1-17 oder älter als 65. Wenn ein Begrenzungswert leer gelassen wird, bleibt der Bereich ohne Begrenzung. Beispiel: Wenn als Untergrenze 100 angegeben wird, jedoch keine Obergrenze, werden alle Werte größer-gleich 100 als fehlend definiert. Die Begrenzungswerte werden mit eingeschlossen, d. h., bei einer Untergrenze von 5 und einer Obergrenze von 10 sind 5 und 10 in der Bereichsdefinition enthalten. Ein Bereich fehlender Werte kann für jeden Speichertyp definiert werden, einschließlich "Datum/Uhrzeit" und "Zeichenfolge" (in diesem Fall wird die alphabetische Sortierung verwendet, um zu bestimmen, ob ein Wert im Bereich liegt).
- **Null/Leerer Bereich.** Außerdem können Sie systemdefinierte **Nullen** (in den Daten als `$null$` angezeigt) und **leere Bereiche** (Zeichenfolgewerte ohne sichtbare Zeichen) als Leerstellen angeben. Beachten Sie, dass leere Zeichenfolgen zum Zweck der Analyse im Typknoten als leere Bereiche behandelt werden, auch wenn sie intern anders gespeichert und in bestimmten Fällen anders behandelt werden.

Hinweis: Um Leerstellen als nicht definiert oder `$null$` zu codieren, sollten Sie den Füllerknoten verwenden.

Beschreibung. Verwenden Sie dieses Textfeld zur Angabe einer Feldbeschriftung. Diese Beschriftungen werden an verschiedenen Stellen angezeigt, beispielsweise in Diagrammen, Tabellen, Ausgaben und Modellbrowsern, je nach der im Dialogfeld "Streameigenschaften" getroffenen Auswahl.

Angabe von Werten und Beschriftungen für stetige Daten

Das *stetige* Messniveau wird für numerische Felder verwendet. Es gibt drei Speichertypen für stetige Daten:

- Reelle Zahl
- Ganzzahl
- Datum/Uhrzeit

Dasselbe Dialogfeld wird zur Bearbeitung aller stetigen Felder verwendet; der Speichertyp wird nur als Referenz angezeigt.

Angabe von Werten

Folgende Steuerelemente stehen nur bei stetigen Feldern zur Verfügung und werden zur Angabe von Wertebereichen verwendet:

Minimum. Geben Sie eine Untergrenze für den Wertebereich ein.

Maximum. Geben Sie eine Obergrenze für den Wertebereich ein.

Angabe von Beschriftungen

Sie können Beschriftungen für jeden Wert eines Bereichsfelds angeben. Klicken Sie auf die Schaltfläche **Beschriftungen**, um ein gesondertes Dialogfeld zur Angabe der Wertbeschriftungen zu öffnen.

Unterdialogfeld "Werte und Beschriftungen"

Durch Klicken auf **Beschriftungen** im Dialogfeld "Werte" für ein Bereichsfeld wird ein neues Dialogfeld geöffnet, in dem Sie Beschriftungen für jeden Wert im Bereich angeben können.

Mit den Spalten *Werte* und *Beschriftungen* in dieser Tabelle können Sie Wert-/Beschriftungspaare definieren. Die derzeit definierten Paare werden hier angezeigt. Sie können neue Beschriftungspaare hinzufügen, indem Sie in eine leere Zelle klicken und einen Wert und seine zugehörige Beschriftung eingeben. *Hin-*

weis: Durch das Hinzufügen von Wert/Wert-Beschriftungspaaren in dieser Tabelle werden dem Feld keine neuen Werte hinzugefügt. Stattdessen werden einfach Metadaten für den Feldwert erstellt.

Die im Typknoten angegebenen Beschriftungen werden an verschiedenen Stellen angezeigt (als QuickInfo, Ausgabebeschriftungen usw.), je nach der im Dialogfeld "Streameigenschaften" getroffenen Auswahl.

Angabe von Werten und Beschriftungen für nominale und ordinale Daten

Nominale (Set) und ordinale Messniveaus (sortiertes Set) zeigen an, dass die Datenwerte diskret als Setmitglieder verwendet werden. Als Speichertypen für Sets sind "Zeichenfolge", "Ganze Zahl", "Reelle Zahl" und "Datum/Uhrzeit" möglich.

Folgende Steuerelemente stehen nur bei nominalen und ordinalen Feldern zur Verfügung und werden zur Angabe von Werten und Beschriftungen verwendet:

Werte. Mit der Spalte *Werte* in der Tabelle können Sie Werte auf der Grundlage Ihrer Kenntnisse über das aktuelle Feld angeben. Mithilfe dieser Tabelle können Sie erwartete Werte für das Feld eingeben und dann mit der Dropdown-Liste "Werte überprüfen" testen, ob das Dataset diesen Werten entspricht. Mit den Pfeilschaltflächen und der Löschschaltfläche können Sie bestehende Werte bearbeiten sowie Werte neu sortieren und löschen.

Beschriftungen. In der Spalte *Beschriftungen* können Sie Beschriftungen für jeden Wert im Set angeben. Diese Beschriftungen werden an verschiedenen Stellen angezeigt, beispielsweise in Diagrammen, Tabellen, Ausgaben und Modellbrowsern, je nach der im Dialogfeld "Streameigenschaften" getroffenen Auswahl.

Angabe von Werten für ein Flag

Flagfelder werden zur Anzeige von Daten verwendet, die zwei unterschiedliche Werte aufweisen. Als Speichertypen für Flags sind "Zeichenfolge", "Ganze Zahl", "Reelle Zahl" und "Datum/Uhrzeit" möglich.

Wahr. Dient zur Angabe eines Flagwerts für das Feld, in dem die Bedingung erfüllt ist.

Falsch. Dient zur Angabe eines Flagwerts für das Feld, in dem die Bedingung nicht erfüllt ist.

Beschriftungen. Dient zur Angabe von Beschriftungen für die einzelnen Werte im Flagfeld. Diese Beschriftungen werden an verschiedenen Stellen angezeigt, beispielsweise in Diagrammen, Tabellen, Ausgaben und Modellbrowsern, je nach der im Dialogfeld "Streameigenschaften" getroffenen Auswahl.

Definieren fehlender Werte

Die Spalte **Fehlend** der Registerkarte "Typen" zeigt an, ob der Umgang mit fehlenden Werten für ein Feld definiert wurde. Folgende möglichen Optionen stehen zur Auswahl:

Ein (*). Zeigt an, dass der Umgang mit fehlenden Werten für dieses Feld definiert ist. Dies kann durch einen abwärts gelegenen Füllerknoten oder durch eine ausdrückliche Angabe mithilfe der Option "Angaben" (siehe unten) erfolgen.

Aus. Für das Feld ist kein Umgang mit fehlenden Werten definiert.

Angeben. Wählen Sie diese Option aus, um ein Dialogfeld anzuzeigen, in dem Sie explizite Werte festlegen können, die als fehlende Werte für dieses Feld betrachtet werden.

Überprüfen von Typenwerten

Durch Aktivierung der Option "Überprüfen" für die einzelnen Felder werden alle Werte im betreffenden Feld untersucht, um zu ermitteln, ob sie den aktuellen Typeneinstellungen bzw. den im Dialogfeld "Werte angeben" angegebenen Werten entsprechen. Dies ist sinnvoll bei der Bereinigung von Datasets und zur Verringerung der Größe eines Datasets innerhalb einer einzelnen Operation.

Durch die Einstellung der Spalte *Überprüfung* im Dialogfeld "Typknoten" wird bestimmt, was geschieht, wenn ein Wert außerhalb der Typengrenzen gefunden wird. Die Überprüfungseinstellungen für ein Feld können Sie mithilfe der Dropdown-Liste für das betreffende Feld in der Spalte *Überprüfen* ändern. Um die Überprüfungseinstellungen für alle Felder festzulegen, klicken Sie in die Spalte *Feld* und drücken Sie Strg-A. Verwenden Sie anschließend die Dropdown-Liste für alle Felder in der Spalte *Überprüfen*.

Folgende Überprüfungseinstellungen stehen zur Verfügung:

Keine. Die Werte werden ohne Überprüfung übergeben. Dies ist die Standardeinstellung.

Auf Nullwert setzen. Ändert Werte außerhalb der Grenzen auf den systemdefinierten Nullwert (`$null$`).

Erzwingen. Felder mit vollständig instanziierten Messniveaus werden auf Werte überprüft, die außerhalb der angegebenen Bereiche liegen. Nicht spezifizierte Werte werden mithilfe der folgenden Regeln in einen zulässigen Wert für das betreffende Messniveau konvertiert:

- Bei Flags werden alle Werte, die nicht "wahr" oder "falsch" sind, in den Wert "falsch" konvertiert.
- Bei Sets (nominal oder ordinal) werden alle unbekannten Werte in das erste Mitglied der Werte des Sets konvertiert.
- Zahlen, deren Wert die Obergrenze eines Bereichs überschreitet, werden durch den Wert der Obergrenze ersetzt.
- Zahlen, deren Wert die Untergrenze eines Bereichs unterschreitet, werden durch den Wert der Untergrenze ersetzt.
- Nullwerten in einem Bereich wird der Wert des Mittelpunkts für den betreffenden Bereich zugewiesen.

Verwerfen. Wenn unzulässige Werte gefunden werden, wird der gesamte Datensatz verworfen.

Warnen. Die Anzahl der unzulässigen Elemente wird gezählt und im Dialogfeld "Streameigenschaften" gemeldet, nachdem alle Daten gelesen wurden.

Abbrechen. Beim ersten unzulässigen Wert, der gefunden wird, wird die Ausführung des Streams abgebrochen. Der Fehler wird im Dialogfeld "Streameigenschaften" gemeldet.

Festlegen der Feldrolle

Die Rolle eines Felds gibt an, wie es bei der Modellerstellung verwendet werden soll, beispielsweise ob es sich bei einem Feld um eine Eingabe oder um ein Ziel (das vorhergesagte Element) handelt.

Hinweis: Die Rollen "Partition", "Häufigkeit" und "Datensatz-ID" können jeweils nur auf ein einziges Feld angewendet werden.

Folgende Rollen stehen zur Verfügung:

Eingabe. Das Feld wird als Eingabe für das Maschinenlernen verwendet (Prädiktorfeld).

Ziel. Das Feld wird als Ausgabe bzw. Ziel für das Maschinenlernen verwendet (eines der Felder, die das Modell vorherzusagen versucht).

Beides. Das Feld wird vom Apriori-Knoten sowohl als Prädiktor als auch als Ziel verwendet. Alle anderen Modellierungsknoten ignorieren das Feld.

Keine. Dieses Feld wird vom Maschinenlernen ignoriert. Felder, deren Messniveau auf **Ohne Typ** gesetzt wurde, werden in der Spalte *Rolle* automatisch auf **Keine** gesetzt.

Partitionieren. Gibt ein Feld an, das zur Partitionierung der Daten in getrennte Stichproben für Trainings, Test- und (optional) Validierungszwecken verwendet wird. Dieses Feld muss ein instanziierter Set-Typ

mit zwei oder drei möglichen Werten sein (wie im Dialogfeld "Feldwerte" definiert). Der erste Wert steht für die Trainingsstichprobe, der zweite für die Teststichprobe und der dritte (sofern vorhanden) für die Validierungsstichprobe. Alle weiteren Werte werden ignoriert und Flagfelder können nicht verwendet werden. Um die Partition in einer Analyse zu verwenden, muss auf der Registerkarte "Modelloptionen" des entsprechenden Modellerstellungs- oder Analyseknotens die Partitionierung aktiviert sein. Datensätze mit Nullwerten für das Partitionsfeld werden aus der Analyse ausgeschlossen, wenn die Partitionierung aktiviert ist. Wenn mehrere Partitionsfelder im Stream definiert wurden, muss in jedem Modellierungsknoten, der die Partitionierung verwendet, auf der Registerkarte "Felder" ein einzelnes Partitionsfeld ausgewählt werden. Wenn in Ihren Daten noch kein geeignetes Feld vorhanden ist, können Sie mithilfe eines Partitionierungs- oder Ableitungsknotens eines erstellen. Weitere Informationen finden Sie im Thema „Partitionsknoten" auf Seite 154.

Aufteilen. (Nur nominale, ordinale und Flagfelder.) Legt fest, dass ein Modell für jeden möglichen Wert des Felds erstellt werden soll.

Häufigkeit. (Nur numerische Felder.) Durch Festlegen dieser Rolle kann der Feldwert als Häufigkeitsgewichtungsfaktor für den Datensatz verwendet werden. Dieses Merkmal wird nur von CRT-, CHAID-, QUEST- und linearen Modellen unterstützt; alle anderen Knoten ignorieren diese Rolle. Die Häufigkeitsgewichtung wird mithilfe der Option **Häufigkeitsgewichtung anwenden** auf der Registerkarte "Felder" der Modellierungsknoten aktiviert, die dieses Merkmal unterstützen.

Datensatz-ID. Das Feld wird als eindeutige ID für einen Datensatz verwendet. Diese Funktion wird von den meisten Knoten ignoriert. Sie wird jedoch von linearen Modellen unterstützt und ist für die IBM Netezza-Knoten zum datenbankinternen Mining erforderlich.

Kopieren von Typattributen

Die Attribute eines Typs, wie beispielsweise Werte, Überprüfungsoptionen und fehlende Werte, können problemlos zwischen Feldern kopiert werden:

1. Klicken Sie mit der rechten Maustaste auf das Feld, dessen Attribute kopiert werden sollen.
2. Wählen Sie im Kontextmenü die Option **Kopieren** aus.
3. Klicken Sie mit der rechten Maustaste auf die Felder, deren Attribute geändert werden sollen.
4. Wählen Sie im Kontextmenü die Option **Inhalte einfügen** aus. *Hinweis:* Sie können durch Klicken bei gedrückter Steuertaste oder über die Option **Felder auswählen** mehrere Felder aus dem Kontextmenü auswählen.

Es wird ein neues Dialogfeld geöffnet, in dem Sie die speziellen Attribute auswählen können, die eingefügt werden sollen. Beim Einfügen in mehrere Felder gelten die hier ausgewählten Optionen für alle Zielfelder.

Folgende Attribute einfügen. Wählen Sie das entsprechende Element aus der unten stehenden Liste aus, um Attribute aus einem Feld in ein anderes einzufügen.

- **Typ.** Wählen Sie diese Option aus, um das Messniveau einzufügen.
- **Werte.** Wählen Sie diese Option aus, um die Feldwerte einzufügen.
- **Fehlende Werte.** Wählen Sie diese Option aus, um die Einstellungen für fehlende Werte einzufügen.
- **Überprüfen.** Wählen Sie diese Option aus, um die Überprüfungsoptionen einzufügen.
- **Rolle.** Wählen Sie diese Option aus, um die Rolle eines Felds einzufügen.

Feldformat - Registerkarte "Einstellungen"

Die Registerkarte "Format" an den Tabellen- und Typknoten zeigt eine Liste der aktuellen oder nicht verwendeten Felder sowie Formatierungsoptionen für die einzelnen Felder. Im Folgenden finden Sie eine Beschreibung der einzelnen Spalten in der Feldformatierungstabelle:

Feld. Hier wird der Name des ausgewählten Felds angezeigt.

Format. Durch Doppelklicken auf eine Zelle in dieser Spalte können Sie die Formatierung für die einzelnen Felder anhand des eingeblendeten Dialogfelds angeben. Weitere Informationen finden Sie im Thema „Festlegen der Feldformatierungsoptionen". Die hier angegebene Formatierung setzt die in den allgemeinen Streameigenschaften angegebene Formatierung außer Kraft.

Hinweis: Die Knoten "Statistics-Export" und "Statistics-Ausgabe" exportieren *SAV*-Dateien, die Formatierungen für die einzelnen Felder in ihren Metadaten enthalten. Wenn das Format für die einzelnen Felder nicht vom Format der *SAV*-Dateien von IBM SPSS Statistics unterstützt wird, verwendet der Knoten das IBM SPSS Statistics-Standardformat.

Ausrichten: In dieser Spalte können Sie angeben, wie die Werte innerhalb der Tabellenspalte ausgerichtet werden sollen. Die Standardeinstellung ist **Automatisch**, was bedeutet, dass Symbolwerte links und numerische Werte rechts ausgerichtet werden. Sie können die Standardeinstellung durch Auswahl von **Links**, **Rechts** oder **Mitte** außer Kraft setzen.

Spaltenbreite: Standardmäßig wird die Spaltenbreite anhand der Werte des Felds automatisch berechnet. Um die automatische Berechnung der Spaltenbreite außer Kraft zu setzen, klicken Sie auf eine Tabellenzelle und wählen Sie mithilfe der Dropdown-Liste eine neue Breite aus. Zur Eingabe von benutzerdefinierten Breiten, die hier nicht aufgeführt sind, öffnen Sie das Unterdialogfeld "Feldformat", indem Sie auf eine Tabellenzelle in der Spalte Feld oder Format doppelklicken. Alternativ können Sie mit der rechten Maustaste auf eine Zelle klicken und **Format festlegen** auswählen.

Aktuelle Felder anzeigen. Standardmäßig wird im Dialogfeld die Liste der derzeit aktiven Felder angezeigt. Um die Liste der nicht verwendeten Felder anzuzeigen, wählen Sie **Nicht verwendete Feldeinstellungen anzeigen**.

Kontextmenü. Das Kontextmenü für diese Registerkarte bietet verschiedene Optionen für Auswahl und Einstellungsaktualisierung. Klicken Sie mit der rechten Maustaste in eine Spalte, um dieses Menü anzuzeigen.

- **Alles auswählen.** Wählt alle Felder aus.
- **Nichts auswählen.** Hebt die Auswahl auf.
- **Felder auswählen.** Wählt Felder anhand von Typ- oder Speichertypeigenschaften aus. Zur Auswahl stehen: **Kategorialen Wert auswählen**, **Stetigen Wert auswählen** (numerisch), **Element ohne Typ auswählen**, **Zeichenfolgen auswählen**, **Zahlen auswählen** sowie **Datum/Uhrzeit auswählen**. Weitere Informationen finden Sie im Thema „Messniveaus" auf Seite 119.
- **Format festlegen.** Öffnet ein Unterdialogfeld, in dem für die einzelnen Felder Optionen für Datum, Uhrzeit und Dezimaltrennzeichen angegeben werden können.
- **Ausrichtung festlegen.** Legt die Ausrichtung für die ausgewählten Felder fest. Zur Auswahl stehen: **Automatisch**, **Mitte**, **Links** und **Rechts**.
- **Spaltenbreite festlegen.** Legt die Feldbreite für ausgewählte Felder fest. Geben Sie **Automatisch** an, um die Breite aus den Daten einzulesen. Alternativ können Sie die Feldbreite auf folgende Werte festlegen: 5, 10, 20, 30, 50, 100 oder 200.

Festlegen der Feldformatierungsoptionen

Die Feldformatierung wird in einem Unterdialogfeld angegeben, das über die Registerkarte "Format" in den Typ- und Tabellenknoten aufgerufen werden kann. Wenn Sie vor dem Öffnen dieses Dialogfelds mehrere Felder ausgewählt haben, werden die Einstellungen aus dem ersten Feld in der Auswahl für alle Felder verwendet. Wenn Sie auf **OK** klicken, nachdem Sie hier Angaben gemacht haben, werden diese Einstellungen für alle Felder übernommen, die auf der Registerkarte "Format" ausgewählt wurden.

Die folgenden Optionen stehen für die einzelnen Felder zur Verfügung. Viele dieser Einstellungen können auch im Dialogfeld "Streameigenschaften" angegeben werden. Alle auf der Feldebene vorgenommenen Einstellungen haben Vorrang gegenüber der für den Stream angegebenen Standardeinstellung.

Datumsformat. Wählen Sie ein Datumsformat aus, das für die Datumsspeicherfelder verwendet werden soll oder wenn Zeichenfolgen von den CLEM-Datumsfunktionen als Datumsangaben interpretiert werden.

Zeitformat. Wählen Sie ein Zeitformat aus, das für die Zeitspeicherfelder verwendet werden soll oder wenn Zeichenfolgen von den CLEM-Zeitfunktionen als Zeitangaben interpretiert werden.

Zahlenanzeigeformat. Sie können aus den Anzeigeformaten Standard (**####.###**), Wissenschaftlich (**#.###E+##**) und Währung (**$###.##**) wählen.

Dezimalzeichen. Wählen Sie als Dezimaltrennzeichen entweder Komma (,) oder Punkt (.) aus.

Symbol für Zifferngruppierung. Wählen Sie bei Zahlenanzeigeformaten aus, welches Symbol zur Gruppierung der Werte verwendet werden soll (z. B. der Punkt in *3.000,00*). Folgende Optionen stehen zur Auswahl: "Keine", "Punkt", "Komma", "Leerzeichen" und "Durch Ländereinstellung definiert" (in diesem Fall wird der Standardwert für die aktuelle Ländereinstellung verwendet).

Dezimalstellen (Standard, wissenschaftlich, Währung, Export). Gibt bei Zahlenanzeigeformaten an, wie viele Dezimalstellen bei der Anzeige reeller Zahlen verwendet werden sollen. Diese Option wird getrennt für jedes Anzeigeformat angegeben.

Ausrichten: Gibt an, wie die Werte innerhalb der Spalte ausgerichtet werden sollen. Die Standardeinstellung ist **Automatisch**, was bedeutet, dass Symbolwerte links und numerische Werte rechts ausgerichtet werden. Sie können die Standardeinstellung durch Auswahl von "Links", "Rechts" oder "Mitte" außer Kraft setzen.

Spaltenbreite. Standardmäßig wird die Spaltenbreite anhand der Werte des Felds automatisch berechnet. Mit den Pfeilen rechts neben dem Listenfeld können Sie eine benutzerdefinierte Breite in Fünferschritten angeben.

Filtern oder Umbenennen von Feldern

Das Umbenennen und Ausschließen von Feldern ist an jedem beliebigen Punkt in einem Stream möglich. Beispiel: Bei einer medizinischen Studie ist möglicherweise der Kaliumspiegel (Daten der Feldebene) der Patienten (Daten der Datensatzebene) nicht relevant. Daher können Sie das Feld *K* (Kalium) herausfiltern. Dies ist mithilfe eines gesonderten Filterknotens oder mithilfe der Registerkarte "Filter" in einem Quellen- oder Ausgabeknoten möglich. Die Funktionen sind immer dieselben, unabhängig davon, von welchem Knoten aus der Zugriff erfolgt.

- An Quellenknoten wie beispielsweise "Datei (var.)", "Datei (fest)", "Statistics-Datei" und "XML-Datei" können Sie Felder beim Einlesen der Daten in IBM SPSS Modeler umbenennen oder filtern.
- Mit einem Filterknoten können Sie Felder an jeder Stelle des Streams umbenennen oder filtern.
- Über die Knoten "Statistics-Export", "Statistics-Transformation", "Statistics-Modell" und "Statistics-Ausgabe" können Sie Felder filtern oder umbenennen, die den IBM SPSS Statistics-Benennungsstandards entsprechen. Weitere Informationen finden Sie im Thema „Umbenennen oder Filtern von Feldern für IBM SPSS Statistics” auf Seite 362.
- Mit der Registerkarte "Filter" in einem der oben angegebenen Knoten können Sie Mehrfachantwortsets definieren bzw. bearbeiten. Weitere Informationen finden Sie im Thema „Bearbeiten von Mehrfachantwortsets” auf Seite 131.
- Schließlich können Sie mit einem Filterknoten Felder aus einem Quellenknoten einem anderen Quellenknoten zuweisen.

Festlegen der Filteroptionen

Die auf der Registerkarte "Filter" verwendete Tabelle zeigt den Namen der einzelnen Felder, die in den Knoten eintreten, sowie den Namen jedes Felds, das den Knoten verlässt. Mit den Optionen in diese Tabelle können Sie Felder umbenennen oder herausfiltern, die doppelt vorhanden oder für die Operationen weiter hinten im Stream nicht erforderlich sind.

- **Feld.** Zeigt die Eingabefelder aus den aktuell verbundenen Datenquellen an.
- **Filter.** Zeigt den Filterstatus aller Eingabefelder an. Gefilterte Felder weisen in dieser Spalte ein rotes "X" auf, das darauf hinweist, dass das Feld nicht an die späteren Operationen im Stream übergeben wird. Klicken Sie in die Spalte *Filter* für ein bestimmtes Feld, um die Filterfunktion zu aktivieren bzw. zu inaktivieren. Außerdem können Sie mit der Auswahl durch Umschalt-Klicken Optionen für mehrere Felder gleichzeitig auswählen.
- **Feld.** Zeigt die Felder an, die den Filterknoten verlassen. Doppelte Namen werden in roter Farbe angezeigt. Durch Klicken auf diese Spalte und Eingabe eines neuen Namens können Sie Feldnamen bearbeiten. Alternativ können Sie Felder entfernen, indem Sie doppelte Felder durch Klicken in die Spalte *Filter* inaktivieren.

Alle Spalten in der Tabelle können durch Klicken auf den Spaltentitel sortiert werden.

Aktuelle Felder anzeigen. Wählen Sie diese Option aus, um die Felder für Datasets anzuzeigen, die aktiv mit dem Filterknoten verbunden sind. Diese Option wird standardmäßig ausgewählt und ist die häufigste Methode der Verwendung von Filterknoten.

Nicht verwendete Feldeinstellungen anzeigen. Wählen Sie diese Option aus, um die Felder für Datasets anzuzeigen, die zu einem früheren Zeitpunkt (jetzt jedoch nicht mehr) mit dem Filterknoten verbunden waren. Diese Option wird vor allem beim Kopieren von Filterknoten aus einem Stream in einen anderen oder beim Speichern und erneuten Laden von Filterknoten eingesetzt.

Schaltfläche "Filter" - Menü

Klicken Sie auf die Schaltfläche "Filter" links oben im Dialogfeld, um auf ein Menü zuzugreifen, das eine Reihe von Verknüpfungen und anderen Optionen enthält.

Sie haben folgende Möglichkeiten:

- Alle Felder entfernen.
- Alle Felder einschließen.
- Alle Felder umschalten.
- Duplikate entfernen. *Hinweis:* Bei Auswahl dieser Option werden alle Vorkommen des mehrfach vorhandenen Namens entfernt, einschließlich des ersten.
- Feldnamen und Mehrfachantwortsets umbenennen, sodass sie anderen Anwendungen entsprechen. Weitere Informationen finden Sie im Thema „Umbenennen oder Filtern von Feldern für IBM SPSS Statistics” auf Seite 362.
- Feldnamen verkürzen.
- Namen von Feldern und Mehrfachantwortsets anonymisieren.
- Eingabefeldnamen verwenden.
- Mehrfachantwortsets bearbeiten. Weitere Informationen finden Sie im Thema „Bearbeiten von Mehrfachantwortsets” auf Seite 131.
- Standardfilterstatus festlegen.

Außerdem können Sie mit den Pfeilschaltflächen oben im Dialogfeld festlegen, ob Felder standardmäßig eingeschlossen oder verworfen werden sollen. Dies ist sinnvoll für große Datasets, bei denen nur einige

nachgeordnete Felder verwendet werden sollen. Wählen Sie beispielsweise nur die beizubehaltenden Felder aus und legen Sie fest, dass alle anderen Felder verworfen werden sollen, anstatt die zu verwerfenden Felder einzeln auszuwählen.

Verkürzen von Feldnamen

Im Menü der Schaltfläche "Filter" (links oben auf der Registerkarte "Filter") können Sie auswählen, dass Feldnamen abgeschnitten werden sollen.

Maximale Länge. Dient zur Angabe einer Anzahl von Zeichen zur Begrenzung der Länge von Feldnamen.

Anzahl der Stellen. Wenn Feldnamen nach dem Verkürzen nicht mehr eindeutig sind, werden sie noch weiter verkürzt und durch Hinzufügen von Ziffern zum Namen unterschieden. Sie können angeben, wie viele Ziffern verwendet werden sollen. Mithilfe der Tabellenpfeile können Sie die Zahl einstellen.

Beispiel: Die folgende Tabelle illustriert das Verkürzen von Feldnamen in einem medizinischen Dataset mithilfe der Standardeinstellungen ("Maximale Länge" = 8 und "Anzahl der Ziffern" = 2).

Tabelle 17. Feldnamenverkürzung

Feldnamen	Verkürzte Feldnamen
Patienteneingabe 1	Patien01
Patienteneingabe 2	Patien02
Herzfrequenz	Herzfreq
BD	BD

Anonymisieren von Feldnamen

Feldnamen können aus jedem Knoten anonymisiert werden, der die Registerkarte "Filter" enthält. Klicken Sie dazu links oben auf das Menü der Schaltfläche "Filter" und wählen Sie die Option **Feldnamen anonymisieren** aus. Anonymisierte Feldnamen bestehen aus einem Zeichenfolgenpräfix sowie einem eindeutigen Wert auf numerischer Basis.

Namen anonymisieren. Wählen Sie **Nur ausgewählte Felder**, um nur die Namen der Felder zu anonymisieren, die bereits auf der Registerkarte "Felder" ausgewählt wurden. Die Standardvorgabe lautet **Alle Felder**; dabei werden alle Feldnamen anonymisiert.

Feldnamenpräfix. Das Standardpräfix für anonymisierte Feldnamen lautet **anon_**. Falls Sie ein anderes Präfix verwenden möchten, wählen Sie die Option **Benutzerdefiniert** und geben Sie das gewünschte Präfix ein.

Mehrfachantwortsets anonymisieren. Anonymisiert die Namen von Mehrfachantwortsets auf dieselbe Weise wie Felder. Weitere Informationen finden Sie im Thema „Bearbeiten von Mehrfachantwortsets".

Um die ursprünglichen Feldnamen wiederherzustellen, wählen Sie im Schatlflächenmenü "Filter" die Option **Eingabefeldnamen verwenden**.

Bearbeiten von Mehrfachantwortsets

Mehrfachantwortsets können aus jedem Knoten hinzugefügt bzw. bearbeitet werden, der die Registerkarte "Filter" enthält. Klicken Sie dazu links oben auf das Menü der Schaltfläche "Filter" und wählen Sie die Option **Mehrfachantwortsets bearbeiten** aus.

Mehrfachantwortsets dienen zur Aufzeichnung von Daten, die für jeden Fall mehrere Werte aufweisen können. Dies ist beispielsweise der Fall, wenn die Teilnehmer an einer Umfrage gefragt werden, welche Museen sie besucht haben oder welche Zeitschriften sie lesen. Mehrfachantwortsets können mithilfe eines

Data Collection-Quellenknotens oder eines Statistikdatei-Quellenknotens in IBM SPSS Modeler importiert werden und in IBM SPSS Modeler mithilfe eines Filterknotens definiert werden.

Klicken Sie auf **Neu**, um ein neues Mehrfachantwortset zu erstellen, oder klicken Sie auf **Bearbeiten**, um ein bestehendes Set zu bearbeiten.

Name und Beschriftung. Gibt den Namen und die Beschreibung für das Set an.

Typ. Fragen mit Mehrfachantworten können auf zwei verschiedene Weisen verarbeitet werden:

- **Set von dichotomen Variablen.** Für jede mögliche Antwort wird ein separates Flagfeld erstellt. Bei 10 Zeitschriften werden also 10 Flagfelder erstellt, die jeweils Werte wie 0 und 1 für *wahr* bzw. *falsch* aufweisen können. Unter "Gezählter Wert" können Sie angeben, welcher Wert als "wahr" gezählt werden soll. Diese Methode ist sinnvoll, wenn die Befragten die Möglichkeit haben sollen, alle zutreffenden Optionen auszuwählen.
- **Set von kategorialen Variablen.** Für jede Antwort wird ein nominales Feld mit der maximalen Anzahl an Antwortmöglichkeiten für den Befragten erstellt. Jedes nominale Feld weist Werte für die möglichen Antworten auf, wie beispielsweise *1* für *Spiegel*, *2* für *Focus* und *3* für *Bunte*. Diese Methode ist dann am sinnvollsten, wenn Sie die Anzahl der Antwortmöglichkeiten einschränken möchten, beispielsweise, wenn die Befragten die drei Zeitschriften angeben sollen, die sie am häufigsten lesen.

Felder im Set. Mithilfe der Symbole auf der rechten Seite können Sie Felder hinzufügen bzw. entfernen.

Kommentare

- Alle Felder in einem Mehrfachantwortset müssen denselben Speichertyp aufweisen.
- Es muss zwischen den Sets und den darin enthaltenen Feldern unterschieden werden. So werden beispielsweise durch das Löschen eines Sets nicht die darin enthaltenen Felder gelöscht, sondern lediglich die Verknüpfungen zwischen diesen Feldern. Das Set ist oberhalb vom Löschpunkt weiterhin sichtbar, nicht jedoch weiter unten im Stream.
- Wenn Felder mithilfe eines Filterknotens (unmittelbar auf der Registerkarte oder durch Auswahl der Optionen "Umbenennen für IBM SPSS Statistics", "**Verkürzen**", oder "**Anonymisieren**" im Filtermenü), werden alle Verweise auf diese Felder in Mehrfachantwortsets ebenfalls aktualisiert. Felder in einem Mehrfachantwortset, die vom Filterknoten verworfen werden, werden jedoch nicht aus dem Mehrfachantwortset entfernt. Derartige Felder sind zwar nicht mehr im Stream sichtbar, werden jedoch weiterhin vom Mehrfachantwortset referenziert. Dies ist beispielsweise beim Export zu berücksichtigen.

Ensemble-Knoten

Der Ensemble-Knoten kombiniert zwei oder mehr Modellnuggets, um genauere Vorhersagen zu erzielen, als aus einem dieser Modelle allein gewonnen werden können. Durch die Kombination der Vorhersagen aus mehreren Modellen können Beschränkungen in einzelnen Modellen vermieden werden, was zu einer höheren Gesamtgenauigkeit führt. Auf diese Weise kombinierte Modelle bringen normalerweise eine mindestens ebenso gute Leistung wie die besten Einzelmodelle und sind häufig sogar noch besser.

Diese Kombination von Knoten geschieht automatisch in den automatisierten Modellierungsknoten "Automatisches Klassifikationsmerkmal", "Auto-Numerisch" und "Autom. Cluster".

Nach der Verwendung eines Ensemble-Knotens können Sie mithilfe eines Analyse- oder Evaluierungsknotens die Genauigkeit der kombinierten Ergebnisse mit den Ergebnissen der einzelnen als Eingabe verwendeten Modelle vergleichen. Hierbei darf die Option **Von Ensemblemodellen generierte Felder herausfiltern** auf der Registerkarte "Einstellungen" des Ensemble-Knotens nicht ausgewählt sein.

Ausgabefelder (OutputFields)

Jeder Ensemble-Knoten generiert ein Feld mit den kombinierten Scores. Der Name beruht auf dem angegebenen Zielfeld und trägt das Präfix *$XF_*, *$XS_* oder *$XR_*, je nach Messniveau des Felds (Flag, nominal (Set) bzw. stetig (Bereich)). Beispiel: Wenn das Ziel ein Flagfeld mit dem Namen *Antwort* ist, erhält das Ausgabefeld den Namen *$XF_Antwort*.

Konfidenz- bzw. Neigungsfelder. Bei Flagfeldern und nominalen Feldern werden zusätzliche Konfidenz- bzw. Neigungsfelder, die auf der EnsemblemMethode beruhen, wie in der folgenden Tabelle beschrieben.

Tabelle 18. Erstellung von Feldern für Ensemble-Methode.

Ensemble-Methode	Feldname
Voting Nach Konfidenz gewichtetes Voting Nach Raw Propensity gewichtetes Voting Nach Adjusted Propensity gewichtetes Voting Höchste Konfidenz hat Vorrang	*$XFC_<Feld>*
Durchschnittliche Raw Propensity	*$XFRP_<Feld>*
Durchschnittliche Adjusted Propensity	*$XFAP_<Feld>*

Ensemble-Knoten - Einstellungen

Zielfeld für Ensemble. Dient zur Auswahl eines einzelnen Felds, das von zwei oder mehr weiter oben im Stream gelegenen Modellen als Ziel verwendet wird. Die weiter oben im Stream gelegenen Modelle können Ziele vom Typ "Flag", "nominal" oder "stetig" verwenden, aber mindestens zwei der Modelle müssen dasselbe Ziel verwenden, damit eine Kombination der Scores möglich ist.

Von Ensemblemodellen generierte Felder herausfiltern. Entfernt alle zusätzlichen Felder aus der Ausgabe, die von den Einzelmodellen generiert wurden, die in den Ensemble-Knoten eingespeist werden. Aktivieren Sie dieses Kontrollkästchen, wenn Sie ausschließlich am kombinierten Score aus allen Eingabemodellen interessiert sind. Diese Option muss inaktiviert sein, wenn Sie beispielsweise einen Analyseknoten oder einen Evaluierungsknoten verwenden möchten, um die Genauigkeit des kombinierten Score mit der Genauigkeit bei den einzelnen Eingabemodellen zu vergleichen.

Die verfügbaren Einstellungen hängen vom Messniveau des Felds ab, das als Ziel ausgewählt ist.

Stetige Ziele

Bei stetigen Zielen wird der Durchschnitt aus den Scores gebildet. Dies ist die einzige verfügbare Methode für die Kombination von Scores.

Bei der Generierung von Durchschnittsscores oder Schätzungen verwendet der Ensemble-Knoten eine Standardfehlerberechnung, um den Unterschied zwischen den gemessenen oder geschätzten Werten und den wahren Werten zu berechnen und um anzuzeigen, wie hoch die Übereinstimmung dieser Schätzungen war. Standardfehlerberechnungen werden für neue Modelle standardmäßig generiert; Sie können das Kontrollkästchen jedoch für existierende Modelle inaktivieren, wenn sie beispielsweise neu generiert werden sollen.

Kategoriale Ziele

Bei kategorialen Zielen werden mehrere Methoden unterstützt, darunter **Voting** (Abstimmung). Dabei wird zusammengerechnet, wie häufig jeder mögliche vorhergesagte Wert ausgewählt wurde; der Wert mit der höchsten Gesamtsumme wird dann verwendet. Beispiel: Wenn drei von fünf Modellen *Ja* vorhersagen und die anderen beiden *Nein*, dann gewinnt *Ja* mit 3 zu 2 "Stimmen". Alternativ können die Stimmen beim Voting auf der Grundlage des Konfidenz- oder Neigungswerts der einzelnen Vorhersagen **gewichtet** werden. Die Gewichtungen werden dann summiert und es wird wiederum der Wert mit dem höchsten

Gesamtergebnis ausgewählt. Die Konfidenz für die endgültige Vorhersage ist die Summe der Gewichtungen für den Siegerwert dividiert durch die Anzahl der im Ensemble enthaltenen Modelle.

Alle kategorialen Felder. Für Flagfelder und nominale Felder werden folgende Methoden unterstützt:

- Voting
- Nach Konfidenz gewichtetes Voting
- Höchste Konfidenz hat Vorrang

Nur Flagfelder. Wenn ausschließlich Flagfelder vorliegen, steht außerdem eine Reihe von Methoden zur Verfügung, die auf Neigung beruhen:

- Nach Raw Propensity gewichtetes Voting
- Nach Adjusted Propensity gewichtetes Voting
- Durchschnittliche Raw Propensity
- Durchschnittliche Adjusted Propensity

Gleichstand beim Voting. Bei Voting-Methoden können Sie auswählen, wie Gleichstände aufgelöst werden sollen.

- **Zufallsauswahl.** Einer der gebundenen Werte (Werte mit Gleichstand) wird nach dem Zufallsprinzip ausgewählt.
- **Höchste Konfidenz.** Der gebundene Wert, der mit der höchsten Konfidenz vorhergesagt wurde, gewinnt. Beachten Sie, dass es sich hierbei nicht unbedingt um die höchste Konfidenz aller vorhergesagten Werte handelt.
- **Raw Propensity oder Adjusted Propensity (nur bei Flagfeldern).** Der gebundene Wert, der mit der höchsten absoluten Neigung vorhergesagt wurde. Dabei berechnet sich die absolute Neigung wie folgt:

```
abs(0,5 - Propensity) *
Z
```

Oder bei Adjusted Propensity:

```
abs(0,5 - Adjusted Propensity) * 2
```

Ableitungsknoten

Eine der leistungsstärksten Funktionen in IBM SPSS Modeler ist die Möglichkeit, Datenwerte zu ändern und neue Felder aus bestehenden Daten abzuleiten. Bei längeren Data-Mining-Projekten werden zumeist mehrere Ableitungen durchgeführt, beispielsweise die Extraktion einer Kunden-ID aus einer Zeichenfolge mit Webprotokolldaten oder das Erstellen eines Kundenkapitalwerts auf der Basis von Transaktionsdaten und demografischen Daten. Alle diese Transformationen können mit einer Reihe von Feldoperationsknoten durchgeführt werden.

Mehrere Knoten bieten die Möglichkeit zur Ableitung neuer Felder:

Der Ableitungsknoten ändert Datenwerte oder erstellt neue Felder aus einem oder mehreren bestehenden Feldern. Er erstellt Felder vom Typ "Formel", "Flag", "Nominal", "Status", "Anzahl" und "Bedingt".

Der Umcodierungsknoten transformiert ein Set kategorialer Werte in ein anderes. Die Umcodierung dient zur Reduzierung von Kategorien bzw. Neugruppierung von Daten für die Analyse.

Der Klassierknoten erstellt automatisch neue nominale Felder (Setfelder) auf der Grundlage der Werte eines oder mehrerer bestehender stetiger Felder (numerischer Bereich). Sie können beispielsweise ein stetiges Einkommensfeld in ein neues kategoriales Feld transformieren, das Einkommensgruppen als Abweichungen vom Mittelwert enthält. Nach der Erstellung von Klassen für das neue Feld können Sie einen Ableitungsknoten anhand der Trennwerte generieren.

Der Dichotomknoten leitet mehrere Flagfelder auf der Grundlage der kategorialen Werte ab, die für ein oder mehrere nominale Felder definiert sind.

Der Knoten "Umstrukturieren" wandelt ein nominales Feld oder ein Flagfeld in eine Gruppe von Feldern um, die mit den Werten aus einem weiteren Feld ausgefüllt werden können. Beispiel: Aus einem Feld mit dem Namen *Zahlungsart* und den Werten *Kreditkarte*, *Bar* und *EC-Karte* werden drei neue Felder erstellt (*Kreditkarte*, *Bar*, *EC-Karte*), die jeweils den Wert der tatsächlichen Zahlung enthalten.

Der Verlaufsknoten erstellt neue Felder mit Daten aus Feldern in vorangegangenen Datensätzen. Verlaufsknoten werden am häufigsten für sequenzielle Daten, beispielsweise Zeitreihendaten, verwendet. Vor der Verwendung eines Verlaufsknotens sollten die Daten mithilfe eines Sortierknotens sortiert werden.

Verwenden des Ableitungsknotens

Mithilfe des Ableitungsknotens können Sie sechs Typen neuer Felder aus einem oder mehreren Feldern erstellen:

- **Formel.** Das neue Feld ist das Ergebnis eines beliebigen CLEM-Ausdrucks.
- **Flag.** Bei dem neuen Feld handelt es sich um ein Flag, das für eine angegebene Bedingung steht.
- **Nominal.** Bei dem neuen Feld handelt es sich um ein nominales Feld, was bedeutet, dass es eine Gruppe angegebener Werte als Mitglieder besitzt.
- **Status.** Das neue Feld weist einen von zwei Statuswerten auf. Der Wechsel zwischen diesen Statuswerten wird durch eine angegebene Bedingung ausgelöst.
- **Häufigkeiten.** Dieses neue Feld gibt an, wie oft eine Bedingung wahr war.
- **Bedingt.** Das neue Feld gibt den Wert eines von zwei Ausdrücken an, je nach dem Wert einer Bedingung.

Jeder dieser Knoten enthält ein Reihe von speziellen Optionen im Dialogfeld "Ableitungsknoten". Diese Optionen werden in den nachfolgenden Themenabschnitten erörtert.

Festlegen der Grundoptionen für den Ableitungsknoten

Oben im Dialogfeld für Ableitungsknoten steht eine Reihe von Optionen zur Verfügung, mit denen Sie den Typ des von Ihnen benötigten Ableitungsknotens auswählen können.

Modalwert. Wählen Sie **Einfach** oder **Mehrere**, je nachdem, ob Sie ein Feld oder mehrere Felder ableiten möchten. Bei Auswahl von **Mehrere** ändert sich das Dialogfeld. Es enthält nun Optionen für mehrere Ableitungsfelder.

Ableitungsfeld. Geben Sie bei einfachen Ableitungsknoten den Namen des Felds an, das Sie ableiten und zu den einzelnen Datensätzen hinzufügen möchten. Der Standardname lautet "Ableiten*N*". Dabei steht *N* für die Anzahl der Ableitungsknoten, die Sie bisher während der aktuellen Sitzung erstellt haben.

Ableitungstyp. Wählen Sie in der Dropdown-Liste einen Typ für den Ableitungsknoten aus, beispielsweise "Formel" oder "Nominal". Für jeden Typ wird auf der Grundlage der im typenspezifischen Dialogfeld angegebenen Bedingungen ein neues Feld erstellt.

Bei Auswahl einer Option in der Dropdown-Liste wird eine neue Gruppe von Steuerelementen zum Hauptdialogfeld hinzugefügt, die von den Eigenschaften jedes Ableitungsknotentyps abhängen.

Feldtyp. Wählen Sie ein Messniveau für den neu abgeleiteten Knoten aus, beispielsweise "Stetig", "Kategorial" oder "Flag". Diese Option haben alle Arten von Ableitungsknoten gemeinsam.

Hinweis: Für die Ableitung neuer Felder müssen häufig besondere Funktionen oder mathematische Ausdrücke verwendet werden. Um Ihnen die Erstellung dieser Ausdrücke zu erleichtern, steht im Dialogfeld für alle Typen von Ableitungsknoten ein Expression Builder zur Verfügung, mit dem Sie die Regeln überprüfen können und der außerdem eine vollständige Liste der CLEM-Ausdrücke bietet.

Ableiten mehrerer Felder

Wenn Sie den Modus innerhalb eines Ableitungsknotens auf **Mehrere** setzen, können Sie anhand derselben Bedingung innerhalb desselben Knotens mehrere Felder ableiten. Diese Funktion spart Zeit, wenn identische Transformationen für mehrere Felder im Dataset durchgeführt werden sollen. Beispiel: Wenn Sie ein Regressionsmodell erstellen möchten, das auf der Grundlage des Anfangsgehalts und der bisherigen Berufserfahrung das aktuelle Gehalt vorhersagt, kann es sinnvoll sein, für alle drei schiefen Variablen eine Log-Transformation durchzuführen. Anstatt für jede Transformation einen neuen Ableitungsknoten hinzuzufügen, können Sie dieselbe Funktion gleichzeitig zu allen Feldern hinzufügen. Wählen Sie einfach alle Felder aus, aus denen ein neues Feld abgeleitet werden soll, und geben Sie dann den Ableitungsausdruck mithilfe der Funktion `@FIELD` innerhalb der Feldklammern ein.

Hinweis: Die Funktion `@FIELD` ist ein wichtiges Tool zur gleichzeitigen Ableitung mehrerer Felder. Damit können Sie auf den Inhalt des aktuellen Felds bzw. der aktuellen Felder Bezug nehmen, ohne den genauen Feldnamen angeben zu müssen. Beispiel: Ein CLEM-Ausdruck, der zur Anwendung einer Log-Transformation auf mehrere Felder verwendet wird, ist `log(@FIELD)`.

Folgende Optionen werden zum Dialogfeld hinzugefügt, wenn Sie den Modus **Mehrere** auswählen:

Ableiten aus. Verwenden Sie die Feldauswahlschaltfläche zur Auswahl von Feldern, aus denen neue Felder abgeleitet werden sollen. Für jedes ausgewählte Feld wird ein Ausgabefeld generiert. *Hinweis:* Die ausgewählten Felder müssen nicht denselben Speichertyp aufweisen. Allerdings schlägt der Ableitungsvorgang fehl, wenn die Bedingung nicht für *alle* Felder gültig ist.

Feldnamenerweiterung. Geben Sie die Erweiterung ein, die zu den neuen Feldnamen hinzugefügt werden soll. Beispiel: Bei einem neuen Feld, das den Logarithmus von *Aktuelles Gehalt* enthält, könnten Sie den Feldnamen mit *log_* erweitern, wodurch sich *log_Aktuelles Gehalt* ergibt. Mit den Optionsfeldern können Sie auswählen, ob die Erweiterung als Präfix (am Anfang) oder als Suffix (am Ende) des Feldnamens eingefügt werden soll. Der Standardname lautet "Ableiten*N*". Dabei steht *N* für die Anzahl der Ableitungsknoten, die Sie bisher während der aktuellen Sitzung erstellt haben.

Wie beim Ableitungsknoten im Einzelmodus müssen Sie jetzt einen Ausdruck erstellen, der zur Ableitung eines neuen Felds verwendet wird. Je nach dem Typ der ausgewählten Ableitungsoperation steht eine Reihe von Optionen zum Erstellen einer Bedingung zur Verfügung. Diese Optionen werden in den nachfolgenden Themenabschnitten erörtert. Um einen Ausdruck zu erstellen, können Sie einfach Eingaben in den Formularfeldern vornehmen oder durch Klicken auf die Schaltfläche für den Taschenrechner den Expression Builder verwenden. Denken Sie daran, die Funktion `@FIELD` zu verwenden, wenn es um Bearbeitungen in mehreren Feldern geht.

Auswählen mehrerer Felder

Bei allen Knoten, die Operationen in mehreren Eingabefeldern durchführen, wie "Ableiten" (Mehrfachmodus), "Aggregieren", "Sortieren", "Multiplot" und "Zeitdiagramm", können Sie mithilfe des Dialogfelds "Felder auswählen" schnell und einfach mehrere Felder auswählen.

Sortieren nach. Sie können die verfügbaren Felder für die Anzeige sortieren. Dazu stehen folgende Optionen zur Verfügung:

- **Natürlich.** Zeigt die Felder in der Reihenfolge an, in der Sie über den Datenstream an den aktuellen Knoten übergeben wurden.
- **Name.** Die Felder werden für die Anzeige alphabetisch sortiert.
- **Typ.** Die Felder werden in der Anzeige nach Messniveau sortiert. Diese Option ist bei der Auswahl von Feldern mit einem bestimmten Messniveau nützlich.

Sie können die Felder in der Liste einzeln auswählen oder mithilfe von Umschalt-Klicken bzw. Strg-Klicken mehrere Felder gleichzeitig auswählen. Außerdem können Sie mit den Schaltflächen unter der Liste Gruppen von Feldern anhand ihres Messniveaus oder alle Felder in der Tabelle auswählen bzw. die Auswahl aller Felder aufheben.

Festlegen der Formelableitungsoptionen

Formelableitungsknoten erstellen ein neues Feld für jeden Datensatz im Dataset auf der Grundlage der Ergebnisse eines CLEM-Ausdrucks. Beachten Sie, dass dieser Ausdruck nicht bedingt sein kann. Um Werte auf der Grundlage eines bedingten Ausdrucks abzuleiten, verwenden Sie den Typ "Flag" oder "Bedingt" des Ableitungsknotens.

Formel. Geben Sie mithilfe der CLEM-Sprache eine Formel an, um einen Wert für das neue Feld abzuleiten.

Festlegen der Flagableitungsoptionen

Flagableitungsknoten werden verwendet, um eine bestimmte Bedingung anzugeben, beispielsweise hohen Blutdruck oder Inaktivität auf dem Kundenkonto. Ein Flagfeld wird für jeden Datensatz erstellt und wenn die Bedingung "Wahr" erfüllt ist, wird der Flag-Wert für "Wahr" in das Feld eingetragen.

Wahr-Wert. Dient zur Angabe einen Werts, der für Datensätze, die die unten angegebene Bedingung erfüllen, in das Flagfeld aufgenommen werden soll. Der Standardwert lautet "T".

Falsch-Wert. Dient zur Angabe eines Werts, der für Datensätze, die die unten angegebene Bedingung *nicht* erfüllen, in das Flagfeld aufgenommen werden soll. Der Standardwert lautet "F".

Wahr, wenn. Dient zur Angabe einer CLEM-Bedingung zur Evaluierung bestimmter Werte jedes Datensatzes und zur Zuweisung eines Wahr- oder Falsch-Werts (oben definiert) für den Datensatz. Beachten Sie: Datensätzen wird bei nicht falschen numerischen Werten der Wahr-Wert zugewiesen.

Hinweis: Wenn eine leere Zeichenfolge ausgegeben werden soll, müssen Sie öffnende und schließende Anführungszeichen ohne etwas dazwischen eingeben, d. h. `""`. Leere Zeichenfolgen werden beispielsweise häufig als Falsch-Wert verwendet, damit die Wahr-Werte deutlicher in der Tabelle sichtbar sind. Anführungszeichen sollten außerdem verwendet werden, wenn ein Zeichenfolgewert gewünscht wird, der ansonsten als Zahl behandelt werden würde.

Beispiel

In Versionen von IBM SPSS Modeler vor 12.0 wurden Mehrfachantworten in ein einzelnes Feld importiert. Die Werte wurden dabei durch Kommas getrennt. Beispiel:

```
museum_of_design,institute_of_textiles_and_fashion
museum_of_design
archeological_museum
$null$
national_art_gallery,national_museum_of_science,other
```

Um diese Daten für die Analyse vorzubreiten, können Sie mithilfe der Funktion `hassubstring` ein gesondertes Flagfeld für jede Antwort mit einem Ausdruck der folgenden Art erstellen:

```
hassubstring(museums,"museum_of_design")
```

Festlegen der Nominalableitungsoptionen

Nominalableitungsknoten dienen zur Ausführung eines Sets von CLEM-Bedingungen, um zu ermitteln, welche Bedingung die einzelnen Datensätze erfüllen. Wenn eine Bedingung für jeden Datensatz erfüllt ist, wird ein Wert (der angibt, welches Bedingungsset erfüllt war) in das neue, abgeleitete Feld eingetragen.

Standardwert. Geben Sie einen Wert an, der im neuen Feld verwendet werden soll, wenn keine der Bedingungen erfüllt ist.

Feld setzen auf. Dient zur Angabe eines Werts, der in das neue Feld eingetragen werden soll, wenn eine bestimmte Bedingung erfüllt ist. Jedem Wert in der Liste ist eine Bedingung zugeordnet, die in der benachbarten Spalte anzugeben ist.

Wenn diese Bedingung wahr ist. Dient zur Angabe einer Bedingung für jedes Mitglied im Setfeld. Mit dem Expression Builder können Sie eine Auswahl aus den verfügbaren Funktionen und Feldern treffen. Mit den Pfeilschaltflächen und der Löschschaltfläche können Sie Bedingungen neu ordnen bzw. entfernen.

Bei Bedingungen werden die Werte eines bestimmten Felds im Dataset gestestet. Beim Testen der einzelnen Bedingungen werden die oben angegebenen Werte dem neuen Feld zugewiesen, um anzuzeigen, welche Bedingung erfüllt wurde. Wenn keine der Bedingungen erfüllt wurde, wird der Standardwert verwendet.

Festlegen der Statusableitungsoptionen

Statusableitungsknoten weisen eine gewisse Ähnlichkeit mit Flagableitungsknoten auf. Flagknoten setzen Werte abhängig von der Erfüllung einer *einzelnen* Bedingung für den aktuellen Datensatz fest, Statusableitungsknoten dagegen können die Werte eines Felds abhängig davon ändern, wie es *zwei unabhängige* Bedingungen erfüllt. Das bedeutet, dass sich der Wert ändert (Schalten auf "Ein" bzw. "Aus"), je nachdem, ob die Bedingung erfüllt ist.

Anfänglicher Status. Dient zur Auswahl, ob jedem Datensatz des neuen Felds ursprünglich der Wert **Ein** oder **Aus** zugewiesen werden soll. Beachten Sie, dass dieser Wert sich ändern kann, wenn die einzelnen Bedingungen erfüllt werden.

"Ein"-Wert. Dient zur Angabe des Werts für das neue Feld, wenn die Bedingung für "Ein" erfüllt ist.

Auf "Ein" schalten, wenn. Dient zur Angabe einer CLEM-Bedingung, die den Wert in "Ein" ändert, wenn die Bedingung wahr ist. Klicken Sie auf die Schaltfläche für den Taschenrechner, um Expression Builder zu öffnen.

"Aus"-Wert. Dient zur Angabe des Werts für das neue Feld, wenn die Bedingung für "Aus" erfüllt ist.

Auf "Aus" schalten, wenn. Dient zur Angabe einer CLEM-Bedingung, die den Wert in "Aus" ändert, wenn die Bedingung falsch ist. Klicken Sie auf die Schaltfläche für den Taschenrechner, um Expression Builder zu öffnen.

Hinweis: Um eine leere Zeichenfolge anzugeben, müssen Sie öffnende und schließende Anführungszeichen ohne etwas dazwischen eingeben, d. h. `""`. Anführungszeichen sollten außerdem verwendet werden, wenn ein Zeichenfolgewert gewünscht wird, der ansonsten als Zahl behandelt werden würde.

Festlegen der Anzahlableitungsoptionen

Anzahlableitungsknoten werden verwendet, um eine Reihe von Bedingungen auf die Werte eines numerischen Felds im Dataset anzuwenden. Wenn die einzelnen Bedingungen erfüllt sind, erhöht sich der Wert des abgeleiteten Anzahlfelds um ein festgelegtes Inkrement. Diese Art von Ableitungsknoten ist sinnvoll für Zeitreihendaten.

Anfangswert. Legt einen Wert fest, der bei der Ausführung für das neue Feld verwendet wird. Beim anfänglichen Wert muss es sich um eine numerische Konstante handeln. Mithilfe der Pfeilschaltflächen können Sie den Wert erhöhen oder verringern.

Erhöhen, wenn. Dient zur Angabe der CLEM-Bedingung, bei deren Erfüllung der abgeleitete Wert anhand der in "Erhöhen um" angegebenen Zahl geändert wird. Klicken Sie auf die Schaltfläche für den Taschenrechner, um Expression Builder zu öffnen.

Erhöhen um. Dient zur Festlegung des zur Erhöhung der Anzahl verwendeten Werts. Sie können entweder eine numerische Konstante oder das Ergebnis eines CLEM-Ausdrucks verwenden.

Zurücksetzen bei Dient zur Angabe einer Bedingung, bei deren Erfüllung der abgeleitete Wert auf den anfänglichen Wert zurückgesetzt wird. Klicken Sie auf die Schaltfläche für den Taschenrechner, um Expression Builder zu öffnen.

Festlegen der Optionen für bedingte Ableitung

Knoten für die bedingte Ableitung verwenden eine Reihe von Wenn-Dann-Sonst-Anweisungen zur Ableitung des Werts des neuen Felds.

Wenn. Dient zur Angabe einer CLEM-Bedingung, die bei Ausführung für jeden Datensatz evaluiert wird. Wenn die Bedingung wahr (bzw. bei Zahlen: nicht falsch) ist, wird dem neuen Feld der unten durch den Dann-Ausdruck angegebene Wert zugewiesen. Klicken Sie auf die Schaltfläche für den Taschenrechner, um Expression Builder zu öffnen.

Dann. Dient zur Angabe eines Werts bzw. CLEM-Ausdrucks, der für das neue Feld gilt, wenn die oben stehende Wenn-Anweisung wahr (bzw. nicht falsch) ist. Klicken Sie auf die Schaltfläche für den Taschenrechner, um Expression Builder zu öffnen.

Sonst. Dient zur Angabe eines Werts bzw. CLEM-Ausdrucks, der für das neue Feld gilt, wenn die oben stehende Wenn-Anweisung falsch ist. Klicken Sie auf die Schaltfläche für den Taschenrechner, um Expression Builder zu öffnen.

Umcodieren von Werten mit dem Ableitungsknoten

Mit Ableitungsknoten können auch Werte umcodiert werden, beispielsweise durch Konvertieren eines Zeichenfolgenfelds mit kategorialen Werten in ein numerisches nominales Feld (Setfeld).

1. Wählen Sie unter "Ableitungstyp" den entsprechenden Feldtyp aus (z. B. "Nominal" oder "Flag").
2. Legen Sie die Bedingungen für die Umcodierung der Werte fest. Geben Sie beispielsweise an, dass der Wert 1 zugewiesen werden soll, wenn gilt: `Drug='drugA'`, der Wert 2 für `Drug='drugB'` usw.

Füllerknoten

Füllerknoten werden verwendet, um Feldwerte zu ersetzen und den Speichertyp zu ändern. Sie können auswählen, dass die Werte auf der Grundlage einer angegebenen CLEM-Bedingung ersetzt werden sollen, beispielsweise `@BLANK(FIELD)`. Alternativ können Sie auswählen, dass alle Leerstellen oder Nullwerte mit einem bestimmten Wert ersetzt werden sollen. Füllerknoten werden zum Ersetzen fehlender Werte häufig in Verbindung mit dem Typknoten verwendet. Beispielsweise können Sie Leerstellen mit dem Mittelwert eines Felds ausfüllen, indem Sie einen Ausdruck wie `@GLOBAL_MEAN` angeben. Dieser Ausdruck füllt alle Leerzeichen mit dem durch einen Globalwerteknoten berechneten Mittelwert.

Felder ausfüllen. Mit der Feldauswahlschaltfläche rechts neben dem Textfeld können Sie Felder aus den Datasets auswählen, deren Werte untersucht und ersetzt werden. Standardmäßig werden die Werte in Abhängigkeit von den unten angegebenen Ausdrücken "Bedingung" und "Ersetzen durch" ersetzt. Sie können jedoch auch eine Alternative Ersetzungsmethode auswählen. Verwenden Sie dazu die unten stehenden Ersetzungsoptionen.

Hinweis: Bei Auswahl mehrerer Felder für die Ersetzung mit einem benutzerdefinierten Wert müssen alle Feldtypen ähnliche sein (alle numerisch oder alle symbolisch).

Ersetzen. Hier können Sie auswählen, mit welcher der folgenden Methoden die Werte der ausgewählten Felder ersetzt werden sollen:

- **Anhand der Bedingung.** Diese Option aktiviert das Feld "Bedingung" und Expression Builder, damit Sie einen Ausdruck erstellen können, der als Bedingung für die Ersetzung mit dem angegebenen Wert verwendet werden kann.
- **Immer.** Ersetzt alle Werte für das ausgewählte Feld. Beispielsweise können Sie mit dieser Option den Speichertyp für "income" mit folgendem CLEM-Ausdruck in eine Zeichenfolge konvertieren: `(to_string(income))`.
- **Leere Werte.** Ersetzt alle benutzerdefinierten leeren Werte im ausgewählten Feld. Die Standardbedingung `@BLANK(@FIELD)` wird zur Auswahl von Leerstellen verwendet. *Hinweis:* Mit der Registerkarte "Typen" im Quellenknoten oder mit einem Typknoten können Sie Leerstellen definieren.
- **Nullwerte.** Ersetzt alle systemdefinierten Nullwerte im ausgewählten Feld. Die Standardbedingung `@NULL(@FIELD)` wird zur Auswahl von Nullwerten verwendet.
- **Leere Werte und Nullwerte.** Ersetzt sowohl leere Werte als auch systemdefinierte Nullen im ausgewählten Feld. Diese Option ist hilfreich, wenn Sie sich nicht sicher sind, ob Nullen als fehlende Werte definiert sind oder nicht.

Bedingungen. Diese Option ist verfügbar, wenn Sie die Option **Anhand der Bedingung** ausgewählt haben. In diesem Textfeld können Sie einen CLEM-Ausdruck zur Evaluierung der ausgewählten Felder angeben. Klicken Sie auf die Schaltfläche für den Taschenrechner, um Expression Builder zu öffnen.

Ersetzen durch. Dient zur Angabe eines CLEM-Ausdrucks, um den ausgewählten Feldern einen neuen Wert zuzuweisen. Außerdem können Sie den Wert durch einen Nullwert ersetzen, indem Sie `undef` in das Textfeld eingeben. Klicken Sie auf die Schaltfläche für den Taschenrechner, um Expression Builder zu öffnen.

Hinweis: Wenn die ausgewählten Felder den Typ "Zeichenfolge" aufweisen, sollten Sie sie mit einem Zeichenfolgewert ersetzen. Die Verwendung des Standardwerts 0 oder eines anderen numerischen Werts als Ersatzwert für Zeichenfolgenfelder führt zu einem Fehler.

Speichertypkonvertierung mithilfe des Füllerknotens

Mithilfe der Bedingung "Ersetzen" eines Füllerknotens können Sie problemlos den Feldspeicher für ein einzelnes Feld oder für mehrere Felder ändern. Beispiel: Mithilfe der Konvertierungsfunktion `to_integer` könnten Sie *income* von einer Zeichenfolge in eine ganze Zahl konvertieren. Dazu wird folgender CLEM-Ausdruck verwendet: `to_integer(income)`

Sie können die verfügbaren Funktionen anzeigen und mit Expression Builder automatisch einen CLEM-Ausdruck erstellen. Wählen Sie in der Dropdown-Liste "Funktionen" die Option **Konvertierung** aus, um eine Liste der Funktionen für die Konvertierung des Speichertyps anzuzeigen. Folgende Konvertierungsfunktionen stehen zur Verfügung:

- `to_integer(ELEMENT)`
- `to_real(ELEMENT)`
- `to_number(ELEMENT)`
- `to_string(ELEMENT)`
- `to_time(ELEMENT)`
- `to_timestamp(ELEMENT)`
- `to_date(ELEMENT)`
- `to_datetime(ELEMENT)`

Konvertieren von Datums- und Zeitwerten. Beachten Sie, dass die Konvertierungsfunktionen (und alle anderen Funktionen, für die ein spezieller Eingabetyp, wie beispielsweise ein Wert für Datum oder Uhrzeit, erforderlich ist) von den aktuell im Dialogfeld für die Streamoptionen angegebenen Formaten abhängen. Wenn Sie beispielsweise ein Zeichenfolgenfeld mit den Werten *Jan 2003*, *Feb 2003* usw. in einen Datumsspeicher konvertieren möchten, wählen Sie **MON JJJJ** als Standarddatumsformat für den Stream aus.

Konvertierungsfunktionen sind auch im Ableitungsknoten zur temporären Konvertierung während einer Ableitungsberechnung verfügbar. Mit dem Ableitungsknoten können Sie auch andere Bearbeitungen vornehmen, wie beispielsweise die Umcodierung von Zeichenfolgenfeldern mit kategorialen Werten. Weitere Informationen finden Sie im Thema „Umcodieren von Werten mit dem Ableitungsknoten” auf Seite 139.

Anonymisierungsknoten

Mit dem Anonymisierungsknoten können Sie Feldnamen und/oder Feldwerte verschleiern, wenn Sie mit Daten arbeiten, die in im Knoten nachgeordnetes Modell aufgenommen werden sollen. Auf diese Weise kann das generierte Modell frei verteilt werden (beispielsweise an den Technical Support), ohne dass die Gefahr besteht, dass unbefugte Benutzer vertrauliche Daten wie beispielsweise Personalakten oder Patientenakten anzeigen können.

Je nachdem, wo Sie den Anonymisierungsknoten im Stream platzieren, müssen Sie möglicherweise Änderungen an anderen Knoten vornehmen. Wenn Sie beispielsweise einen Anonymisierungsknoten oberhalb eines Auswahlknotens im Stream einfügen, müssen die Auswahlkriterien des Auswahlknotens geändert werden, wenn sie für Werte gelten sollen, die nun anonymisiert wurden.

Die für die Anonymisierung verwendete Methode beruht auf mehreren Faktoren. Bei Feldnamen und allen Feldwerten mit Ausnahme von stetigen Messniveaus werden die Daten durch eine Zeichenfolge der folgenden Form ersetzt:

*Präfix*_S*n*

Dabei ist *Präfix*_ entweder eine vom Benutzer angegebene Zeichenfolge oder die Standardzeichenfolge `anon_` und *n* ist ein ganzzahliger Wert, der bei 0 beginnt und für jeden eindeutigen Wert erhöht wird (z. B. `anon_S0`, `anon_S1` usw.).

Feldwerte mit dem Typ "Stetig" müssen transformiert werden, da sich numerische Bereiche mit ganzen oder reellen Zahlen befassen und nicht mit Zeichenfolgen. Daher können sie nur durch Transformation des Bereichs in einen anderen Bereich anonymisiert werden, wodurch die ursprünglichen Daten verschleiert werden. Die Transformation von Wert *x* im Bereich wird wie folgt durchgeführt:

`A*(`*x* `+ B)`

Dabei gilt:

A ist ein Skalierungsfaktor, der größer als 0 sein muss.

B ist ein Verschiebungsoffset, das zu den Werten addiert wird.

Beispiel

Bei einem Feld *ALTER*, bei dem der Skalierungsfaktor *A* auf 7 und das Verschiebungsoffset *B* auf 3 gesetzt ist, werden die Werte für *ALTER* wie folgt transformiert:

```
7*(ALTER + 3)
```

Festlegen der Optionen für den Anonymisierungsknoten

Hier können Sie auswählen, bei welchen Feldern die Werte weiter unten im Stream verschleiert werden sollen.

Beachten Sie, dass die Datenfelder oberhalb des Anonymisierungsknotens instanziiert werden müssen, damit Anonymisierungsoperationen durchgeführt werden können. Sie können die Daten durch Klicken auf die Schaltfläche **Werte lesen** in einem Typknoten bzw. auf der Registerkarte "Typen" eines Quellenknotens instanziieren.

Feld. Listet die Felder im aktuellen Dataset auf. Wenn bereits Feldnamen anonymisiert wurden, werden die anonymisierten Namen hier angezeigt.

Messung. Das Messniveau des Felds.

Werte anonymisieren. Wählen Sie ein oder mehrere Felder aus, klicken Sie auf diese Spalte und wählen Sie die Option **Ja**, um die Feldwerte mit dem Standardpräfix **anon_** zu anonymisieren; wählen Sie **Angeben**, um ein Dialogfeld anzuzeigen, in dem Sie entweder Ihr eigenes Präfix eingeben oder - bei Feldwerten des Typs *Stetig* - angeben können, ob bei der Transformation der Feldwerte Zufallswerte oder vom Benutzer angegebene Werte verwendet werden sollen. Beachten Sie, dass *stetige* und nicht-*stetige* Feldtypen nicht in derselben Operation zusammen mit Daten eines anderen Typs angegeben werden können; Sie müssen diesen Vorgang separat für die einzelnen Feldtypen durchführen.

Aktuelle Felder anzeigen. Wählen Sie diese Option aus, um die Felder für Datasets anzuzeigen, die aktiv mit dem Anonymisierungsknoten verbunden sind. Diese Option ist standardmäßig aktiviert.

Nicht verwendete Feldeinstellungen anzeigen. Wählen Sie diese Option aus, um die Felder für Datasets anzuzeigen, die zu einem früheren Zeitpunkt (jetzt jedoch nicht mehr) mit dem Knoten verbunden waren. Diese Option wird vor allem beim Kopieren von Knoten aus einem Stream in einen anderen oder beim Speichern und erneuten Laden von Knoten eingesetzt.

Angabe der Vorgehensweise bei der Anonymisierung von Feldwerten

Im Dialogfeld "Werte ersetzen" können Sie auswählen, ob das Standardpräfix für anonymisierte Feldwerte oder ein benutzerdefiniertes Präfix verwendet werden soll. Wenn Sie in diesem Dialogfeld auf **OK** klicken, ändert sich die Einstellung von "Werte anonymisieren" auf der Registerkarte "Einstellungen" für die ausgewählten Felder in **Ja**.

Feldwertepräfix. Das Standardpräfix für anonymisierte Feldwerte lautet **anon_**. Falls Sie ein anderes Präfix verwenden möchten, wählen Sie die Option **Benutzerdefiniert** und geben Sie das gewünschte Präfix ein.

Das Dialogfeld "Werte transformieren" wird nur für Felder des Typs "Stetig" angezeigt und ermöglicht Ihnen anzugeben, ob bei der Transformation der Feldwerte Zufallswerte oder vom Benutzer angegebene Werte verwendet werden sollen.

Zufällig. Wählen Sie diese Option aus, um Zufallswerte für die Transformation zu verwenden. **Startwert für Zufallsgenerator festlegen** ist standardmäßig ausgewählt; geben Sie einen Wert im Feld **Startwert** an oder verwenden Sie den Standardwert.

Fest. Wählen Sie diese Option aus, um Ihre eigenen Werte für die Transformation anzugeben.

- **Skalieren um.** Der Wert, mit dem die Feldwerte in der Transformation multipliziert werden. Der Mindestwert ist 1; der Höchstwert ist normalerweise 10; er kann jedoch gesenkt werden, um einen Überlauf zu vermeiden.
- **Verschieben um.** Der Wert, der in der Transformation zu den Feldwerten addiert wird. Der Mindestwert ist 0; der Höchstwert ist normalerweise 1000. Er kann jedoch gesenkt werden, um einen Überlauf zu vermeiden.

Anonymisieren von Feldwerten

Bei den auf der Registerkarte "Einstellungen" für die Anonymisierung ausgewählten Feldern werden die Werte in folgenden Fällen anonymisiert:

- Wenn Sie den Stream ausführen, der den Anonymisierungsknoten enthält
- Wenn Sie eine Vorschau der Werte anzeigen

Um eine Vorschau der Werte anzuzeigen, klicken Sie auf der Registerkarte "Anonymisierte Werte" auf die Schaltfläche **Werte anonymisieren**. Wählen Sie als Nächstes einen Feldnamen in der Dropdown-Liste aus.

Beim Messniveau "Stetig" werden folgende Elemente angezeigt:

- Mindest- und Höchstwert des ursprünglichen Bereichs
- Die zur Transformation der Werte verwendete Gleichung

Bei einem anderen Messniveau als "Stetig" werden der ursprüngliche und der anonymisierte Wert für das betreffende Feld angezeigt.

Wenn die Anzeige einen gelben Hintergrund aufweist, deutet dies darauf hin, dass sich entweder die Einstellung für das ausgewählte Feld seit der letzten Anonymisierung geändert hat oder dass Änderungen an den Daten oberhalb des Anonymisierungsknotens vorgenommen wurden, sodass die anonymisierten Werte möglicherweise nicht mehr korrekt sind. Das aktuelle Werteset wird angezeigt. Klicken Sie erneut auf die Schaltfläche **Werte anonymisieren**, um ein neues Werteset entsprechend der aktuellen Einstellung zu generieren.

Werte anonymisieren. Erstellt anonymisierte Werte für das ausgewählte Feld und zeigt diese in der Tabelle an. Bei Verwendung von Zufallsstartwerten für ein Feld vom Typ "Stetig" wird durch Klicken auf diese Schaltfläche jedes Mal ein anderes Werteset erstellt.

Werte löschen. Löscht die ursprünglichen und die anonymisierten Werte aus der Tabelle.

Umcodierungsknoten

Der Umcodierungsknoten ermöglicht die Transformation eines Sets kategorialer Werte in ein anderes. Die Umcodierung dient zur Reduzierung von Kategorien bzw. Neugruppierung von Daten für die Analyse. Beispielsweise können Sie die Werte für *Produkt* in drei Gruppen umcodieren, wie zum Beispiel *Küchenzubehör*, *Bad und Bettwäsche* sowie *Elektrogeräte*. Diese Operation wird häufig direkt aus einem Verteilungsknoten ausgeführt. Dazu werden die Werte gruppiert und ein Umcodierungsknoten wird erstellt. Weitere Informationen finden Sie im Thema „Verwendung von Verteilungsknoten" auf Seite 223.

Die Umcodierung kann für ein oder mehrere symbolische Felder durchgeführt werden. Außerdem können Sie festlegen, dass die neuen Werte für das bestehende Feld eingesetzt werden sollen, oder ein neues Feld generieren.

Einsatzmöglichkeiten für Umcodierungsknoten

Vor der Verwendung eines Umcodierungsknotens sollten Sie überlegen, ob ein anderer Feldoperationsknoten für die betreffende Aufgabe geeigneter ist:

- Um numerische Bereiche automatisch in Sets (z. B Ränge oder Prozentsätze) umzuwandeln, sollten Sie einen Klassierknoten verwenden. Weitere Informationen finden Sie im Thema „Klassierknoten" auf Seite 146.
- Wenn Sie numerische Bereiche manuell in Sets umwandeln möchten, sollten Sie einen Ableitungsknoten verwenden. Beispiel: Angenommen, Sie möchten Gehaltswerte in spezielle Gehaltsbereichskategorien zusammenfassen, dann sollten Sie jede Kategorie manuell mithilfe eines Ableitungsknotens definieren.
- Um eines oder mehrere Flagfelder auf der Grundlage der Werte eines kategorialen Felds, beispielsweise *Hypothektyp*, zu erstellen, sollten Sie einen Dichotomknoten verwenden.
- Soll ein kategoriales Feld in ein Feld mit numerischem Speichertyp konvertiert werden, verwenden Sie einen Ableitungsknoten. So können Sie beispielsweise *Nein* und *Ja* in die Werte 0 und 1 konvertieren. Weitere Informationen finden Sie im Thema „Umcodieren von Werten mit dem Ableitungsknoten" auf Seite 139.

Festlegen der Optionen für den Umcodierungsknoten

Die Verwendung des Umcodierungsknotens erfolgt in drei Schritten:

1. Wählen Sie zunächst aus, ob Sie mehrere Felder umcodieren möchten oder nur ein einziges Feld.
2. Wählen Sie als Nächstes aus, ob die Umcodierung in das bestehende Feld erfolgen oder ob ein neues Feld erstellt werden soll.
3. Verwenden Sie schließlich die dynamischen Optionen im Dialogfeld "Umcodierungsknoten", um die Sets wunschgemäß zuzuordnen.

   Modalwert. Wählen Sie **Einfach** aus, um die Kategorien für ein einzelnes Feld umzucodieren. Wählen Sie **Mehrere** aus, um Optionen zu aktivieren, die die Transformation von mehreren Feldern gleichzeitig erlauben.

   Umcodieren in. Wählen Sie **Neues Feld** aus, um das ursprüngliche nominale Feld beizubehalten und ein weiteres Feld abzuleiten, das die umcodierten Werte enthält. Wählen Sie die Option **Vorhandenes Feld**, um die Werte im ursprünglichen Feld mit den neuen Klassifikationen zu überschreiben. Dies ist im Grunde ein "Füll"-Vorgang.

   Nach der Angabe des Modus und der Ersetzungsoptionen müssen Sie das Transformationsfeld auswählen und mithilfe der dynamischen Optionen in der unteren Hälfte des Dialogfelds die neuen Klassifikationswerte angeben. Diese Optionen variieren in Abhängigkeit vom oben ausgewählten Modus.

   Umcodierungsfeld(er). Mit der Feldauswahlschaltfläche auf der rechten Seite können Sie eines (Modus "Einfach") oder mehrere (Modus "Mehrere") kategoriale Felder auswählen.

   Neuer Feldname. Dient zur Angabe eines Namens für das neue nominale Feld mit den umcodierten Werten. Diese Option ist nur im Modus "Einfach" verfügbar, wenn oben **Neues Feld** ausgewählt wurde. Bei Auswahl von **Vorhandenes Feld** wird der ursprüngliche Feldname beibehalten. Im Modus "Mehrere" wird diese Option mit Steuerelementen zur Angabe einer Erweiterung für jedes neue Feld ersetzt. Weitere Informationen finden Sie im Thema „Umcodieren mehrerer Felder" auf Seite 145.

   Werte umcodieren. Diese Tabelle ermöglicht eine klare Zuordnung von alten Set-Werten zu den hier angegebenen.

   - **Ursprünglicher Wert.** Diese Spalte listet bestehende Werte für die Auswahlfelder auf.
   - **Neuer Wert.** In dieser Spalte können Sie neue Kategoriewerte eingeben oder einen aus der Dropdown-Liste auswählen. Wenn Sie automatisch einen Umcodierungsknoten mit Werten aus einem Verteilungsdiagramm generieren, sind diese Werte in der Dropdown-Liste enthalten. Dadurch können Sie schnell und einfach bestehende Werte einem bekannten Set von Werten zuordnen. Beispiel: Gesundheitsorganisationen gruppieren Diagnosen manchmal unterschiedlich je nach Netz oder Ländereinstellung. Nach einer Fusion oder Übernahme müssen alle Beteiligten die neuen oder sogar die bereits vorhandenen Daten einheitlich klassifizieren. Anstatt jeden Zielwert aus einer langen Lis-

te einzeln manuell einzugeben, können Sie die Master-Liste der Werte in IBM SPSS Modeler einlesen, ein Verteilungsdiagramm für das Feld *Diagnose* ausführen und einen Umcodierungsknoten für dieses Feld direkt aus dem Diagramm erstellen. Dadurch werden alle Zielwerte für die Diagnose in der Dropdown-Liste "Neue Werte" verfügbar.

4. Klicken Sie auf **Ermitteln**, um die ursprünglichen Werte für ein oder mehrere oben ausgewählte Felder zu lesen.
5. Klicken Sie auf **Kopieren**, um für noch nicht zugeordnete Felder die ursprünglichen Werte in die Spalte *Neuer Wert* einzufügen. Die nicht zugeordneten ursprünglichen Werte werden in die Dropdown-Liste aufgenommen.
6. Klicken Sie auf **Neue löschen**, um alle Spezifikationen in der Spalte *Neuer Wert* zu löschen. *Hinweis:* Mit dieser Option werden die Werte nicht aus der Dropdown-Liste gelöscht.
7. Klicken Sie auf **Automatisch**, um automatisch aufeinander folgende ganze Zahlen für jeden der ursprünglichen Werte zu erstellen. Nur ganzzahlige Werte (keine reellen Werte wie 1,5; 2,5 usw.) können generiert werden.

Sie können beispielsweise automatisch fortlaufende Produkt-IDs für Produktnamen erstellen oder Kursnummern für das Lehrangebot einer Universität. Diese Funktion entspricht der Transformation "Automatisch umcodieren" für Sets in IBM SPSS Statistics.

Für nicht spezifizierte Werte verwenden. Diese Option wird verwendet, um nicht spezifizierte Werte in das neue Feld einzutragen. Sie können entweder auswählen, dass der ursprüngliche Wert beibehalten werden soll, indem Sie **Originalwert** auswählen, oder einen Standardwert angeben.

Umcodieren mehrerer Felder

Um die Kategoriewerte für mehrere Felder gleichzeitig zuzuordnen, legen Sie als Modus **Mehrere** fest. Dadurch werden neue Einstellungen im Dialogfeld "Umcodieren" aktiviert. Diese werden im Folgenden beschrieben.

Felder umcodieren. Mit der Feldauswahlschaltfläche auf der rechten Seite können Sie die Felder auswählen, die transformiert werden sollen. Mithilfe der Schaltfläche "Feldauswahl" können Sie alle Felder gleichzeitig auswählen oder Felder mit gleichem Typ wie "Nominal" oder "Flag".

Feldnamenerweiterung. Bei der gleichzeitigen Umcodierung mehrerer Felder ist es effizienter, anstatt einzelner Feldnamen eine gemeinsame Erweiterung anzugeben, um die alle neue Felder ergänzt werden. Geben Sie eine Erweiterung, wie beispielsweise `_umkod`, an und wählen Sie aus, ob diese Erweiterung an den Anfang oder an das Ende der ursprünglichen Dateinamen gestellt werden soll.

Speichertyp und Messniveau für umcodierte Felder

Der Umcodierungsknoten erstellt bei der Umcodierung immer ein nominales Feld. In einigen Fällen kann sich dadurch das Messniveau ändern, wenn der Umcodierungssmodus **Vorhandenes Feld** verwendet wird.

Der Speichertyp des neuen Felds (wie die Daten *gespeichert*, nicht wie sie *verwendet* werden) wird anhand der folgenden Optionen auf der Registerkarte "Einstellungen" berechnet:

- Wenn bei nicht spezifizierten Werten die Verwendung eines Standardwerts festgelegt ist, wird der Speichertyp durch Untersuchung der neuen Werte und des Standardwerts sowie durch die Bestimmung des geeigneten Speichers ermittelt. Beispiel: Wenn alle Werte als ganze Zahlen analysiert werden können, weist das Feld den Speichertyp "Ganze Zahl" auf.
- Wenn bei nicht spezifizierten Werten die Verwendung der ursprünglichen Werte festgelegt ist, beruht der Speichertyp auf dem Speichertyp des ursprünglichen Felds. Wenn alle Werte als Speichertyp des ursprünglichen Felds analysiert werden können, wird dieser Speichertyp beibehalten; andernfalls wird der Speichertyp ermittelt, indem der geeignetste Speichertyp gesucht wird, der sowohl die alten als auch die neuen Werte umfasst. Beispiel: Bei der Umcodierung eines Sets mit ganzen Zahlen { 1, 2, 3, 4,

5 } mit der Umcodierung 4 => 0, 5 => 0 wird ein neues Set mit ganzen Zahlen { 1, 2, 3, 0 } generiert, während die Umcodierung 4 => "Over 3", 5 => "Over 3" das Zeichenfolgeset { "1", "2", "3", "Over 3" } generiert.

Hinweis: Wenn der ursprüngliche Typ nicht instanziiert war, ist auch der neue Typ nicht instanziiert.

Klassierknoten

Mit dem Klassierknoten können Sie automatisch neue nominale Felder auf der Grundlage eines oder mehrerer bestehender stetiger Felder (numerischer Bereich) erstellen. Sie können beispielsweise ein stetiges Einkommensfeld in ein neues kategoriales Feld transformieren, das Einkommensgruppen gleicher Breite oder als Abweichungen vom Mittelwert enthält. Alternativ können Sie ein kategoriales "Supervisorfeld" auswählen, damit die Stärke der ursprünglichen Assoziation zwischen den beiden Feldern erhalten bleibt.

Die Durchführung der Klassierung kann aus einer Reihe von Gründen nützlich sein. Hier einige Beispiele:

- **Algorithmusanforderungen.** Für bestimmte Algorithmen, beispielsweise "Naive Bayes" und "Logistische Regression", sind kategoriale Eingaben erforderlich.
- **Leistung.** Die Leistung von Algorithmen wie "Multinomiale logistische Regression" kann eventuell gesteigert werden, wenn die Anzahl der unterschiedlichen Werte der Eingabefelder reduziert wird. Sie könnten beispielsweise statt der ursprünglichen Werte den Median oder den Mittelwert für jede Klasse verwenden.
- **Datenschutz.** Vertrauliche persönliche Daten, wie beispielsweise Gehälter, können anstatt als tatsächliche Werte in Bereichen angegeben werden, um dem Datenschutz gerecht zu werden.

Es stehen mehrere Klassierungsmethoden zur Verfügung. Nach der Erstellung von Klassen für das neue Feld können Sie einen Ableitungsknoten anhand der Trennwerte generieren.

Einsatzmöglichkeiten für Klassierknoten

Vor der Verwendung eines Klassierknotens sollten Sie überlegen, ob ein anderes Verfahren für die betreffende Aufgabe geeigneter ist:

- Zur manuellen Angabe von Trennwerten für Kategorien, beispielsweise vordefinierte Gehaltsbereiche, verwenden Sie einen Ableitungsknoten. Weitere Informationen finden Sie im Thema „Ableitungsknoten” auf Seite 134.
- Zur Erstellung neuer Kategorien für bestehende Sets verwenden Sie einen Umcodierungsknoten. Weitere Informationen finden Sie im Thema „Umcodierungsknoten” auf Seite 143.

Umgang mit fehlenden Werten

Der Klassierknoten behandelt fehlende Werte folgendermaßen:

- **Vom Benutzer angegebene Leerstellen.** Fehlende Werte, die als Leerstellen angegeben sind, werden während der Transformierung aufgenommen. Wenn Sie beispielsweise -99 mithilfe des Typknotens als Leerwert gekennzeichnet haben, dann wird dieser Wert in den Klassiervorgang aufgenommen. Um Leerstellen beim Klassieren zu ignorieren, sollten Sie mithilfe eines Füllerknotens die Leerwerte durch den systemdefinierten Nullwert ersetzen.
- **Systemdefiniert fehlende Werte ($null$).** Nullwerte werden während der Klassiertransformation ignoriert und bleiben nach der Transformation weiterhin Nullwerte.

Auf der Registerkarte "Einstellungen" finden Sie Optionen für verfügbare Verfahren. Auf der Registerkarte "Ansicht" werden die Trennwerte angezeigt, die für die Daten ermittelt wurden, die den Knoten zuvor durchlaufen haben.

Festlegen der Optionen für den Klassierknoten

Mit dem Klassierknoten können Sie mit folgenden Verfahren automatisch Klassen (Kategorien) generieren:

- Klassieren mit fester Breite
- N-Perzentile (gleiche Anzahl oder gleiche Summe)
- Mittelwert und Standardabweichung
- Ränge
- Optimiert in Bezug auf ein kategoriales "Supervisorfeld"

Der untere Teil des Dialogfelds ändert sich dynamisch in Abhängigkeit von der ausgewählten Klassiermethode.

Klassenfelder. Stetige Felder (numerischer Bereich) mit ausstehender Transformation werden hier angezeigt. Mit dem Klassierknoten können Sie mehrere Felder gleichzeitig klassieren. Zum Hinzufügen bzw. Entfernen von Feldern dienen die Schaltflächen auf der rechten Seite.

Klassierungsmethode. Dient zur Auswahl der Methode, die zur Ermittlung der Trennwerte für die neuen Feldklassen (Kategorien) verwendet werden. In den nachfolgenden Themenabschnitten werden die jeweils in den einzelnen Fällen verfügbaren Optionen behandelt.

Klassenschwellenwerte. Gibt an, wie die Klassenschwellenwerte berechnet werden.

- **Immer neu berechnen.** Trennwerte und Klassenzuordnungen werden jedes Mal neu berechnet, wenn der Knoten ausgeführt wird.
- **Von Registerkarte "Klassenwerte" lesen, sofern verfügbar.** Trennwerte und Klassenzuordnungen werden nur bei Bedarf berechnet (beispielsweise, wenn neue Daten hinzugefügt wurden).

In den folgenden Themenabschnitten werden Optionen für die verfügbaren Klassiermethoden erörtert.

Klassen mit fester Breite

Wenn Sie als Klassiermethode **Feste Breite** auswählen, wird im Dialogfeld ein neues Optionsset angezeigt.

Namenserweiterung. Dient zur Angabe einer Erweiterung für die generierten Felder. *_BIN* ist die Standarderweiterung. Außerdem können Sie angeben, ob die Erweiterung am Anfang (**Präfix**) oder am Ende (**Suffix**) des Feldnamens eingefügt werden soll. Sie könnten beispielsweise ein neues Feld namens *Einkommen_BIN* erstellen.

Klassenbreite. Geben Sie einen Wert an (ganzzahlig oder reell), der zur Berechnung der "Breite" der Klasse verwendet werden soll. Sie können beispielsweise den Standardwert, 10, verwenden, um das Feld *Alter* zu klassieren. Da *Alter* einen Bereich von 18-65 umfasst, würden folgende Klassen generiert:

Tabelle 19. Klassen für Alter im Bereich 18-65

Klasse 1	Klasse 2	Klasse 3	Klasse 4	Klasse 5	Klasse 6
>=13 bis <23	>=23 bis <33	>=33 bis <43	>=43 bis <53	>=53 bis <63	>=63 bis <73

Der Start der Klassenintervalle wird aus dem niedrigsten gescannten Wert minus der Hälfte der (angegebenen) Klassenbreite berechnet. Beispiel: In den oben angegebenen Klassen wird der Wert 13 verwendet, um die Intervalle gemäß folgender Berechnung zu starten: 18 *[niedrigster Datenwert]* - 5 *[0,5 × (Klassenbreite von 10)]* = 13.

Anzahl der Klassen. Mit dieser Option können Sie eine ganze Zahl angeben, die zur Bestimmung der Anzahl der Klassen (Kategorien) mit fester Breite für die neuen Felder verwendet wird.

Nach der Ausführung des Klassierknotens in einem Stream können Sie die Klassenschwellenwerte anzeigen, die durch Klicken auf die Registerkarte **Vorschau** im Dialogfeld "Klassierknoten" generiert wurden. Weitere Informationen finden Sie im Thema „Vorschau der generierten Klassen” auf Seite 151.

N-Perzentile (gleiche Anzahl oder gleiche Summe)

Mit der Klassiermethode für n-Perzentil erstellen Sie nominale Felder, mit denen die gescannten Datensätze so in Perzentilgruppen (oder Quartilgruppen, Dezilgruppen usw.) aufgeteilt werden können, dass jede Gruppe dieselbe Anzahl an Datensätzen aufweist oder dass die Summe der Werte in den einzelnen Gruppen gleich ist. Die Datensätze werden in aufsteigender Reihenfolge gemäß dem Wert des angegebenen Klassenfelds eingestuft. Datensätze mit dem niedrigsten Wert für die ausgewählte Klassenvariable erhalten somit den Rang 1, die nächste Gruppe von Datensätzen den Rang 2 usw. Die Schwellenwerte für die einzelnen Klassen werden automatisch auf der Grundlage der Daten und der verwendeten N-Perzentil-Methode erzeugt.

Namenserweiterung für N-Perzentile. Geben Sie eine Erweiterung an, die für die mithilfe von Standard-N-Perzentilen generierten Felder verwendet wird. Die Standarderweiterung ist *_TILE* plus *N*; dabei steht *N* für die Nummer des Perzentils. Außerdem können Sie angeben, ob die Erweiterung am Anfang (**Präfix**) oder am Ende (**Suffix**) des Feldnamens eingefügt werden soll. Sie könnten beispielsweise ein neues Feld namens *Einkommen_BIN4* erstellen.

Namenserweiterung für benutzerdef. N-Perzentil. Dient zur Angabe einer Erweiterung für einen benutzerdefinierten N-Perzentil-Bereich. Die Standarderweiterung lautet *_TILEN*. *N* wird in diesem Fall *nicht* durch die benutzerdefinierte Zahl ersetzt.

Folgende N-Perzentile stehen zur Verfügung:

- **Quartil.** Generiert vier Klassen, die jeweils 25 % der Fälle enthalten.
- **Quintil.** Generiert fünf Klassen, die jeweils 20 % der Fälle enthalten.
- **Dezil.** Generiert zehn Klassen, die jeweils 10 % der Fälle enthalten.
- **Vingtil.** Generiert 20 Klassen, die jeweils 5 % der Fälle enthalten.
- **Perzentil.** Generiert 100 Klassen, die jeweils 1% der Fälle enthalten.
- **Benutzerdef. N.** Wählen Sie diese Option aus, um die Anzahl der Klassen festzulegen. Der Wert 3 beispielsweise ergibt 3 in Bereiche eingeteilte Kategorien (2 Trennwerte), die jeweils 33,3 % der Fälle enthalten.

Falls weniger diskrete Werte in den Daten vorhanden sind als N-Perzentile angegeben wurden, werden nicht alle N-Perzentile verwendet. In diesen Fällen spiegelt die neue Verteilung vermutlich die ursprüngliche Verteilung der Daten wider.

Perzentilmethode. Legt fest, welche Methode für die Zuweisung der Datensätze zu den Klassen verwendet wird.

- **Datensatzanzahl.** Versucht, jeder Klasse eine gleich große Anzahl an Datensätzen zuzuweisen.
- **Summe.** Versucht die Datensätze so zu den Klassen zuzuweisen, dass die Summe der Werte in jeder Klasse gleich groß ist. Bei der Zielausrichtung von Absatzbemühungen sind Sie mit dieser Methode beispielsweise in der Lage, die Interessenten gemäß dem Wert je Datensatz zu Dezilgruppen zuzuweisen, wobei die Interessenten mit den höchsten Werten zur obersten Klasse gehören. Beispiel: Ein Pharmaunternehmen stuft die Ärzte gemäß der Anzahl ihrer Verschreibungen in Dezilgruppen ein. Jedes Dezil umfasst in etwa dieselbe Anzahl an Verschreibungen; die Anzahl der Personen, die diese Verschreibungen ausgestellt haben, ist jedoch nicht identisch. Die Personen mit den meisten Verschreibungen würden sich dabei in Dezil 10 wiederfinden. Hinweis: Bei dieser Vorgehensweise wird angenommen, dass alle Werte größer als null sind; ist dies nicht der Fall, können unerwartete Ergebnisse eintreten.

Bindungen. Eine Bindungsbedingung entsteht, wenn beide Seiten eines Trennwerts identisch sind. Wenn Sie beispielsweise Dezile zuweisen und mehr als 10 % der Datensätze denselben Wert im Klassenfeld aufweisen, können nicht alle Datensätze in derselben Klasse untergebracht werden, ohne den Schwellenwert entsprechend nach oben oder nach unten zu verschieben. Die Bindungen können wahlweise aufwärts in die nächste Klasse verschoben oder auch in der aktuellen Klasse beibehalten werden; die Bindungen müssen jedoch in jedem Fall aufgelöst werden, sodass alle Datensätze mit identischen Werten in dieselbe Klasse fallen, auch wenn dadurch einige Klassen mehr Datensätze erhalten als erwartet. Auch die Schwellenwerte der nachfolgenden Klassen müssen angepasst werden, sodass die Werte für dieselbe Zahlengruppe unterschiedlich zugewiesen werden, je nach der verwendeten Methode zum Auflösen der Bindungen.

- **Zu nächstem hinzu.** Wählen Sie diese Option aus, um die Bindungswerte nach oben zur nächsten Klasse zu verschieben.
- **In aktuellem beibehalten.** Hiermit werden die Bindungswerte in der aktuellen (niedrigeren) Klasse belassen. Bei dieser Methode werden insgesamt gegebenenfalls weniger Klassen erstellt.
- **Zufällig zuweisen.** Wählen Sie diese Option aus, um die Bindungswerte nach dem Zufallsprinzip einer Klasse zuzuordnen. Dadurch wird versucht, die Anzahl der Datensätze in jeder Klasse gleich zu halten.

Beispiel: N-Perzentil-Einteilung nach Anzahl der Datensätze

Die folgende Tabelle zeigt, wie vereinfachte Feldwerte bei der N-Perzentil-Einteilung nach Anzahl der Datensätze als Quartile eingestuft werden. Die Ergebnisse sind dabei abhängig von der ausgewählten Bindungsoption.

Tabelle 20. Beispiel für N-Perzentil-Einteilung nach Anzahl der Datensätze.

Werte	Zu nächstem hinzu	In aktuellem beibehalten
10	E	E
13	Z	E
15	3	Z
15	3	Z
20	4	3

Die Anzahl der Elemente pro Klasse wird folgendermaßen berechnet:

```
Gesamtzahl der Werte/Anzahl der N-Perzentile
```

In dem vereinfachten Beispiel oben ist die erwünschte Anzahl der Elemente pro Klasse 1,25 (5 Werte/ 4 Quartile). Der Wert 13 (Wert Nummer 2) überspannt den gewünschten Schwellenwert für die Anzahl (1,25) und wird daher, je nach der ausgewählten Bindungsoption, unterschiedlich behandelt. Im Modus **Zu nächstem hinzu** wird er in Klasse 2 aufgenommen. Im Modus **In aktuellem beibehalten** wird er in Klasse 1 belassen, wodurch der Wertebereich für Klasse 4 so weit verschoben wird, dass er außerhalb des Bereichs der vorhandenen Datenwerte liegt. Daher werden nur drei Klassen erstellt und die Schwellenwerte für jede Klasse werden entsprechend angepasst, wie in der folgenden Tabelle gezeigt.

Tabelle 21. Beispielergebnis der Klassierung.

Klasse	Unterer Bereich	Oberer Bereich
E	>=10	<15
Z	>=15	<20
3	>=20	<=20

Hinweis: Die Geschwindigkeit beim Klassieren nach N-Perzentilen kann gegebenenfalls durch Parallelverarbeitung gesteigert werden.

Bilden von Rangfolgen

Wenn Sie als Klassiermethode **Ränge** auswählen, wird im Dialogfeld ein neues Optionsset angezeigt.

Bei der Rangbildung werden neue Felder erstellt, die Ränge, Bruchzahlränge und Perzentilwerte für numerische Felder enthalten, je nach den unten angegebenen Optionen.

Rangordnung. Wählen Sie **Aufsteigend** (der niedrigste Wert wird mit "1" gekennzeichnet) oder **Absteigend** (der höchste Wert wird mit "1" gekennzeichnet).

Rang. Mit dieser Option weisen Sie den Fällen in aufsteigender bzw. absteigender Reihenfolge (oben angegeben) Ränge zu. Der Bereich der Werte im neuen Feld ist 1-*N*. Dabei ist *N* die Anzahl der diskreten Werte im ursprünglichen Feld. Gebundenen Werten wird der Durchschnitt ihres Ranges zugewiesen.

Relativer Rang. Mit dieser Option weisen Sie Fällen Ränge zu, wobei der Wert des neuen Felds gleich dem Rang dividiert durch die Summe der Gewichtungen der nicht fehlenden Fälle ist. Relative Ränge fallen in den Bereich 0-1.

Prozentsatz Bruchzahlrang. Die einzelnen Ränge werden durch die Anzahl der Datensätze mit gültigen Werten dividiert und mit 100 multipliziert. Als Prozentsatz angegebene Bruchzahlränge fallen in den Bereich 1-100.

Erweiterung. Bei allen Rangoptionen können Sie benutzerdefinierte Erweiterungen erstellen und angeben, ob die Erweiterung am Anfang (**Präfix**) oder am Ende (**Suffix**) des Feldnamens eingefügt werden soll. Sie könnten beispielsweise ein neues Feld namens *Einkommen_P_RANK* erstellen.

Mittelwert/Standardabweichung

Wenn Sie als Klassiermethode **Mittelwert/Standardabweichung** auswählen, wird im Dialogfeld ein neues Optionsset angezeigt.

Mit dieser Methode werden ein oder mehrere neue Felder mit in Bereiche eingeteilten Kategorien erstellt, die auf den Werten für Mittelwert und Standardabweichung der Verteilung für die angegebenen Felder beruhen. Wählen Sie die Anzahl der zu verwendenden Abweichungen aus.

Namenserweiterung. Dient zur Angabe einer Erweiterung für die generierten Felder. *_SDBIN* ist die Standarderweiterung. Außerdem können Sie angeben, ob die Erweiterung am Anfang (**Präfix**) oder am Ende (**Suffix**) des Feldnamens eingefügt werden soll. Sie könnten beispielsweise ein neues Feld namens *Einkommen_SDBIN* erstellen.

- **+/- 1 Standardabweichung.** Mit dieser Option werden drei Klassen generiert.
- **+/- 2 Standardabweichung.** Mit dieser Option werden fünf Klassen generiert.
- **+/- 3 Standardabweichung.** Mit dieser Option werden sieben Klassen generiert.

Wenn Sie beispielsweise "+/-1 Standardabweichung" auswählen, werden drei Klassen generiert, die wie folgt berechnet werden:

Tabelle 22. Beispiel für Standardabweichungsklassen.

Klasse 1	Klasse 2	Klasse 3
`x < (Mittelwert - Std.abw)`	`(Mittelwert - Std.Abw) <= x <= (Mittelwert + Std.Abw)`	`x > (Mittelwert + Std.Abw.)`

Bei einer Normalverteilung liegen 68 % der Fälle innerhalb einer Standardabweichung vom Mittelwert, 95 % innerhalb von zwei Standardabweichungen und 99 % innerhalb von drei Standardabweichungen. Das Erstellen von in Bereiche eingeteilten Kategorien auf der Grundlage der Standardabweichungen kann

jedoch zu definierten Bereichen außerhalb des tatsächlichen Datenbereichs und sogar außerhalb des Bereichs der möglichen Datenwerte führen (z. B. ein negativer Gehaltsbereich).

Optimales Klassieren

Wenn das zu klassierende Feld eine starke Assoziation mit einem anderen kategorialen Feld aufweist, können Sie das kategoriale Feld als "Supervisorfeld" auswählen, damit die Klassen so erstellt werden, dass die Stärke der ursprünglichen Assoziation zwischen den beiden Feldern beibehalten wird.

Beispiel: Angenommen, Sie haben mithilfe der Clusteranalyse Statuswerte auf der Grundlage der Säumnisquoten für Eigenheimkredite gruppiert, mit den höchsten Quoten im ersten Cluster. In diesem Fall können Sie *Prozent nach Fälligkeit* und *Prozent der Zwangsvollstreckungen* als Klassenfelder und das vom Modell generierte Feld für die Clusterzugehörigkeit als Supervisorfeld auswählen.

Namenserweiterung. Geben Sie eine Erweiterung für die generierten Felder an und legen Sie fest, ob die Erweiterung am Anfang (**Präfix**) oder am Ende (**Suffix**) des Feldnamens eingefügt werden soll. Sie könnten beispielsweise ein neues Feld namens *überfällig_OPTIMAL* und ein weiteres namens *Zwangsvollstreckung_OPTIMAL* generieren.

Supervisorfeld. Ein kategoriales Feld, das zur Erstellung der Klassen verwendet wird.

Vorklassierung von Feldern durchführen, um die Leistung bei großen Datasets zu verbessern. Gibt an, ob eine Vorverarbeitung durchgeführt werden soll, um die optimale Klassierung zu rationalisieren. Bei dieser Gruppe werden Skalenwerte mithilfe einer einfachen nicht überwachten Klassiermethode in eine große Anzahl von Klassen gruppiert, die Werte innerhalb der einzelnen Klassen werden durch den Mittelwert repräsentiert und die Fallgewichtung wird entsprechend angepasst, bevor mit dem überwachten Klassieren fortgefahren wird. In der Praxis bedeutet dies, dass bei diesem Verfahren zugunsten einer höheren Geschwindigkeit gewisse Einbußen bei der Präzision in Kauf genommen werden. Es empfiehlt sich für große Datasets. Außerdem können Sie angeben, wie viele Klassen eine Variable nach der Vorverarbeitung maximal aufweisen soll, wenn diese Option verwendet wird.

Klassen mit relativ kleinen Fallzahlen mit einem größeren Nachbarn zusammenführen. Wenn diese Option aktiviert ist, wird eine Klasse zusammengeführt, falls das Verhältnis zwischen ihrer Größe (Anzahl der Fälle) und der Größe einer benachbarten Klasse kleiner ist als der angegebene Schwellenwert; beachten Sie, dass größere Schwellenwerte eine stärkere Zusammenführung mit sich bringen können.

Trennwerteinstellungen

Im Dialogfeld "Trennwerteinstellungen" können Sie erweiterte Optionen für den Algorithmus "Optimales Klassieren" angeben. Diese Optionen legen fest, wie der Algorithmus die Klassen unter Verwendung des Zielfelds berechnen soll.

Klassengrenzen. Sie können angeben, ob der untere bzw. obere Endpunkt eingeschlossen (Minimum <= *x*) oder ausgeschlossen (Minimum < *x*) werden soll.

Erste und letzte Klasse. Für die erste und letzte Klasse können Sie angeben, ob die Klassen keine Begrenzung aufweisen (also gegen positiv bzw. negativ unendlich streben) sollen oder ob sie durch den niedrigsten bzw. höchsten Datenpunkt begrenzt werden sollen.

Vorschau der generierten Klassen

Auf der Registerkarte "Klassenwerte" im Klassierknoten können Sie die Schwellenwerte für die generierten Klassen anzeigen. Mithilfe des Menüs "Generieren" können Sie außerdem einen Ableitungsknoten generieren, mit dem Sie diese Schwellenwerte von einem Dataset auf ein anderes anwenden können.

Klassiertes Feld. Mithilfe der Dropdown-Liste können Sie ein Feld für die Anzeige auswählen. Für die angezeigten Feldnamen werden die ursprünglichen Feldnamen verwendet, um Verwirrung zu vermeiden.

N-Perzentil. Mithilfe der Dropdown-Liste können Sie ein N-Perzentil, beispielsweise 10 oder 100, für die Anzeige auswählen. Diese Option ist nur bei Klassen verfügbar, die mit der N-Perzentil-Methode (gleiche Anzahl oder gleiche Summe) generiert wurden.

Klassenschwellenwerte. Hier werden Schwellenwerte für die einzelnen generierten Klassen sowie die Anzahl der Datensätze angezeigt, die auf die einzelnen Klassen entfallen. Nur bei der Methode "Optimales Klassieren" wird die Anzahl der Datensätze in jeder Klasse als Prozentsatz der Gesamtzahl angezeigt. Beachten Sie, dass die Schwellenwerte beim Klassieren nach Rang nicht zum Einsatz kommen.

Werte lesen. Liest klassierte Werte aus dem Dataset. Beachten Sie, dass Schwellenwerte auch überschrieben werden, wenn neue Daten durch den Stream geleitet werden.

Erzeugen eines Ableitungsknotens

Im Menü "Generieren" können Sie einen Ableitungsknoten auf der Grundlage der aktuellen Schwellenwerte erstellen. Dies ist sinnvoll, um bewährte Klassenschwellenwerte aus einem Dataset auf ein anderes anzuwenden. Außerdem ist, sobald diese Aufteilungspunkte bekannt sind, eine Ableitung bei großen Datasets effizienter (d. h. schneller) als ein Klassiervorgang.

Knoten "RFM-Analyse"

Mit dem Knoten "RFM-Analyse" (Recency-, Frequency-, Monetary-Analyse) können Sie quantitativ ermitteln, welche Kunden wahrscheinlich die besten sind, indem Sie untersuchen, wann sie zuletzt etwas von Ihnen erworben haben (Recency - Aktualität), wie häufig sie eingekauft haben (Frequency - Häufigkeit) und wie viel sie für alle Transaktionen zusammengenommen ausgegeben haben (Monetary - Geldwert).

Der RDM-Analyse liegt zugrunde, dass Kunden, die einmal ein Produkt bzw. eine Dienstleistung erworben haben, dies mit größerer Wahrscheinlichkeit erneut tun. Die kategorisierten Kundendaten werden in eine Reihe von Klassen aufgeteilt, wobei die Klassierkriterien nach Bedarf angepasst werden können. In jeder Klasse wird den Kunden ein Score zugewiesen; diese Scores werden dann zu einem RFM-Gesamtscore kombiniert. Der Score stellt die Zugehörigkeit des Kunden zu den für die einzelnen RFM-Parameter erstellten Klassen dar. Die klassierten Daten reichen möglicherweise für Ihre Bedürfnisse aus, indem sie beispielsweise die häufigsten Kunden mit den höchsten Werten ermitteln. Alternativ können sie zur weiteren Modellierung und Analyse einem Stream übergeben werden.

Beachten Sie: So nützlich die Möglichkeit zur Analyse und Rangeinteilung von RFM-Scores auch ist, müssen Sie sich bei der Verwendung doch bestimmter Faktoren bewusst sein. Es besteht die Versuchung, verstärkt auf die Kunden mit den höchsten Rangwertungen zuzugehen. Eine übermäßige Umwerbung dieser Kunden kann jedoch auch zu Verstimmungen und einen Rückgang in den Wiederholungsgeschäften führen. Außerdem sollte nicht vergessen werden, dass Kunden mit niedrigen Scores nicht vernachlässigt sollten, sondern dass es sinnvoller sein kann, sie zu pflegen, damit sie bessere Kunden werden. Umgekehrt deuten hohe Scores alleine, je nach Markt, noch nicht unbedingt auf gute Absatzchancen hin. So ist ein Kunde in Klasse 5 für "Recency", der also vor sehr kurzer Zeit etwas erworben hat, nicht unbedingt der beste Zielkunde für Unternehmen, die teure, langlebige Produkte verkaufen, wie Autos oder Fernseher.

Hinweis: Je nachdem, wie Ihre Daten gespeichert sind, müssen Sie möglicherweise dem RFM-Analyseknoten einen RFM-Aggregatknoten vorschalten, um die Daten in ein brauchbares Format umzuwandeln. So müssen Eingabedaten beispielsweise im Kundenformat vorliegen, mit einer Zeile pro Kunden; wenn die Daten des Kunden in Transaktionsform vorliegen, können Sie durch Verwendung eines Knotens vom Typ "RFM-Aggregat" weiter oben im Stream die Felder für Aktualität, Häufigkeit und Geldwert ableiten. Weitere Informationen finden Sie im Thema „RFM-Aggregatknoten” auf Seite 76.

Die Knoten "RFM-Aggregat" und "RFM-Analyse" in IBM SPSS Modeler sind für die Verwendung einer unabhängigen Klassierung eingerichtet. Damit werden also Daten für jedes der Maße Aktualität, Häufigkeit und Geldwert in Ränge eingeteilt und klassiert, ohne Berücksichtigung ihrer Werte oder der beiden anderen Maße.

Knoten "RFM-Analyse" - Einstellungen

Aktualität. Mithilfe der Schaltfläche für die Feldauswahl (rechts neben dem Textfeld) können Sie das Aktualitätsfeld auswählen. Dabei kann es sich um ein Datum, eine Zeitmarke oder eine einfache Zahl handeln. Beachten Sie: Wenn ein Datum oder eine Zeitmarke das Datum der aktuellsten Transaktion angibt, wird der höchste Wert als der aktuellste betrachtet; wenn eine Nummer angegeben ist, steht sie für die Zeit, die seit der aktuellsten Transaktion verstrichen ist, und der niedrigste Wert wird als der aktuellste betrachtet.

Hinweis: Wenn dem Knoten "RFM-Analyse" (Recency, Frequency, Monetary) im Stream der Knoten "RFM-Aggregat" vorangeht, sollten die vom Knoten "RFM-Aggregat" generierten Felder "Aktualität", "Häufigkeit" und "Geldwert" im Knoten "RFM-Analyse" als Eingaben ausgewählt werden.

Häufigkeit. Wählen Sie mithilfe der Feldauswahl das zu verwendende Häufigkeitsfeld aus.

Geldwert. Wählen Sie mithilfe der Feldauswahl das zu verwendende Feld für den Geldwert aus.

Anzahl der Klassen. Wählen Sie für jeden der drei Ausgabetypen aus, wie viele Klassen erstellt werden sollen. Der Standardwert ist 5.

Hinweis: Die Mindestzahl beträgt 2, die Höchstzahl 9 Klassen.

Gewichtung. Standardmäßig erhalten bei der Berechnung der Scores die Aktualitätsdaten die größte Bedeutsamkeit; danach folgt die Häufigkeit und dann erst der Geldwert. Falls erforderlich können Sie die Gewichtung für eines oder mehrere dieser Elemente bearbeiten, um die Reihenfolge der Bedeutsamkeit zu ändern.

Der RFM-Score wird wie folgt berechnet: (Aktualitätsscore x Aktualitätsgewichtung) + (Häufigkeitsscore x Häufigkeitsgewicht) + (Geldwertscore x Geldwertgewichtung).

Bindungen. Gibt an, wie identische (gebundene) Scores klassiert werden sollen. Folgende Optionen stehen zur Auswahl:

- **Zu nächstem hinzu.** Wählen Sie diese Option aus, um die Bindungswerte nach oben zur nächsten Klasse zu verschieben.
- **In aktuellem beibehalten.** Hiermit werden die Bindungswerte in der aktuellen (niedrigeren) Klasse belassen. Bei dieser Methode werden insgesamt gegebenenfalls weniger Klassen erstellt. (Dies ist der Standardwert.)

Klassenschwellenwerte. Dient zur Angabe, ob RFM-Scores und Klassenzuordnungen bei jeder Ausführung des Knotens neu berechnet werden sollen oder nur nach Bedarf (z. B. wenn neue Daten hinzugefügt wurden). Bei Auswahl von **Von Registerkarte "Klassenwerte" lesen, sofern verfügbar** können Sie die oberen und unteren Trennwerte für die verschiedenen Klassen auf der Registerkarte "Klassenwerte" ändern.

Bei Ausführung klassiert der Knoten RFM-Analyse die Felder mit den Rohwerten für Aktualität, Häufigkeit und Geldwert und fügt folgende neue Felder zum Dataset hinzu:

- Aktualitätsscore. Ein Rang (Klassenwert) für "Aktualität"
- Häufigkeitsscore. Ein Rang (Klassenwert) für "Häufigkeit"
- Geldwertscore. Ein Rang (Klassenwert) für "Geldwert"
- RFM-Score. Die gewichtete Summe des Aktualitäts-, Häufigkeits- und Geldwertscores.

Ausreißer in Endklassen aufnehmen. Bei Auswahl dieses Kontrollkästchens werden Datensätze, die unterhalb der untersten Klasse liegen, zur untersten Klasse hinzugefügt und Datensätze oberhalb der höchs-

ten Klasse werden in die höchste Klasse aufgenommen; andernfalls erhalten sie einen Nullwert. Dieses Feld steht nur bei Auswahl von **Von Registerkarte "Klassenwerte" lesen, sofern verfügbar** zur Verfügung.

Knoten "RFM-Analyse" - Klassierung

Auf der Registerkarte "Klassenwerte" können Sie die Schwellenwerte für die generierten Klassen anzeigen und in bestimmten Fällen bearbeiten.

Hinweis: Die Werte auf dieser Registerkarte können nur bearbeitet werden, wenn auf der Registerkarte "Einstellungen" die Option **Von Registerkarte "Klassenwerte" lesen, sofern verfügbar** ausgewählt wurde.

Klassiertes Feld. Mithilfe der Dropdown-Liste können Sie ein Feld für die Aufteilung in Klassen auswählen. Verfügbar sind die auf der Registerkarte "Einstellungen" ausgewählten Werte.

Tabelle der Klassenwerte. Hier werden die Schwellenwerte für jeden generierten Bin angezeigt. Bei Auswahl von **Von Registerkarte "Klassenwerte" lesen, sofern verfügbar** auf der Registerkarte "Einstellungen" können Sie die oberen und unteren Trennwerte für die verschiedenen Klassen ändern, indem Sie auf die entsprechende Zelle doppelklicken.

Werte lesen. Liest klassierte Werte aus dem Dataset ein und füllt die Tabelle der Klassenwerte aus. Beachten Sie: Bei Auswahl von **Immer neu berechnen** auf der Registerkarte "Einstellungen" werden die Schwellenwerte der Klassen überschrieben, wenn neue Daten den Stream durchlaufen.

Partitionsknoten

Partitionsknoten werden zur Generierung eines Partitionsfelds verwendet, das Daten in getrennte Subsets bzw. Stichproben für die Trainings, Test- und Validierungsphase der Modellerstellung aufteilt. Indem Sie mit einer Stichprobe das Modell erstellen und es mit einer separaten Stichprobe testen, erhalten Sie einen guten Hinweis dafür, wie gut das Modell sich für größere Datasets verallgemeinern lässt, die den aktuellen Daten ähneln.

Der Partitionsknoten generiert ein nominales Feld, dessen Rolle auf **Partition** eingestellt ist. Wenn ein geeignetes Feld bereits in Ihren Daten vorhanden ist, kann dieses alternativ mithilfe eines Typknotens als Partition gekennzeichnet werden. In diesem Fall ist kein gesonderter Partitionsknoten erforderlich. Jedes instanziierte nominale Feld mit zwei oder drei Werten kann verwendet werden, nicht jedoch Flagfelder. Weitere Informationen finden Sie im Thema „Festlegen der Feldrolle” auf Seite 126.

In einem Stream können mehrere Partitionsfelder definiert werden. Wenn dies geschieht, muss allerdings bei jedem Modellierungsknoten, der Partitionierung verwendet, ein einzelnes Partitionsfeld ausgewählt werden. (Wenn nur eine einzige Partition vorhanden ist, wird diese immer automatisch verwendet, wenn die Partitionierung aktiviert ist.)

Aktivieren der Partitionierung. Um die Partition in einer Analyse zu verwenden, muss auf der Registerkarte "Modelloptionen" des entsprechenden Modellerstellungs- oder Analyseknotens die Partitionierung aktiviert sein. Wenn die Auswahl der Option aufgehoben ist, kann die Partitionierung inaktiviert werden, ohne das Feld zu entfernen.

Um ein Partitionsfeld auf der Grundlage eines anderen Kriteriums, wie beispielsweise Datumsbereich oder Standort, zu erstellen, können Sie auch einen Ableitungsknoten verwenden. Weitere Informationen finden Sie im Thema „Ableitungsknoten” auf Seite 134.

Beispiel. Bei der Erstellung eines RFM-Streams zur Ermittlung aktueller Kunden, die positiv auf frühere Marketingkampagnen reagiert haben, verwendet die Marketingabteilung einer Vertriebsgesellschaft einen Partitionsknoten, um die Daten in Trainings- und Test-Partitionen zu unterteilen.

Partitionsknotenoptionen

Partitionsfeld. Gibt den Namen des vom Knoten erstellten Felds an.

Partitionen. Sie können die Daten in zwei Stichproben (Trainieren und testen) oder drei Stichproben (Trainieren, testen und validieren) partitionieren.

- **Trainieren und testen.** Partitioniert die Daten in zwei Stichproben, sodass Sie das Modell mit einer Stichprobe trainieren und mit der zweiten Stichprobe testen können.
- **Trainieren, testen und validieren.** Partitioniert die Daten in drei Stichproben, sodass Sie das Modell mit einer Stichprobe trainieren und mit der zweiten Stichprobe testen und verfeinern können und schließlich die Ergebnisse mit der dritten Stichprobe validieren können. Dadurch wird allerdings die Größe der einzelnen Partitionen entsprechend verringert. Außerdem ist dieses Verfahren wohl für sehr große Datasets am besten geeignet.

Partitionsgröße. Gibt die relative Größe der einzelnen Partitionen an. Wenn die Summe der Partitionsgrößen weniger als 100 % beträgt, werden die Datensätze, die nicht in einer Partition enthalten sind, verworfen. Beispiel: Ein Benutzer hat 10 Millionen Datensätze und eine Partitionsgröße von 5 % für das Training und von 10 % für das Testen angegeben. Nach der Ausführung des Knotens sollten ca. 500.000 Trainings- und ca. 1 Million Testdatensätze vorhanden sein. Die restlichen Datensätze müssten verworfen worden sein.

Werte. Gibt die Werte an, die für die einzelnen Partitionsstichproben in den Daten verwendet werden.

- **Systemdefinierte Werte verwenden ("1", "2" und "3").** Verwendet eine ganze Zahl für jede Partition. Beispiel: Alle Datensätze, die in der Trainingsstichprobe enthalten sind, weisen den Wert 1 für das Partitionsfeld auf. Dadurch wird sichergestellt, dass die Daten zwischen verschiedenen Ländereinstellungen übertragbar sind und dass bei einer Reinstanziierung des Partitionsfelds an einer anderen Stelle (beispielsweise beim erneuten Einlesen der Daten aus einer Datenbank) die Sortierreihenfolge beibehalten wird (sodass 1 noch immer für die Trainingspartition steht). Die Werte bedürfen jedoch einiger Interpretation.
- **Beschriftungen an systemdefinierte Werte anhängen.** Kombiniert die ganze Zahl mit einer Beschriftung. Beispiel: Trainingspartitions-Datensätze, die den Wert *1_Training* aufweisen. Dadurch kann leicht erkannt werden, wozu die einzelnen Werte gehören, und gleichzeitig wird die Sortierreihenfolge beibehalten. Die Werte sind jedoch für eine bestimmte Ländereinstellung spezifisch.
- **Beschriftungen als Werte verwenden.** Verwendet die Beschriftung ohne ganze Zahl, beispielsweise *Training*. Dadurch können die Werte durch Bearbeitung der Beschriftungen angegeben werden. Dadurch werden die Daten jedoch von der Ländereinstellung abhängig und bei der erneuten Instanziierung einer Partitionsspalte werden die Werte in die natürliche Sortierreihenfolge gebracht, die nicht unbedingt mit ihrer "semantischen" Reihenfolge übereinstimmen muss.

Startwert für Zufallsgenerator festlegen. Bei der Stichprobenziehung oder Partitionierung von Datensätzen auf der Grundlage eines Zufallsprozentsatzes können Sie mit dieser Option dieselben Ergebnisse in einer anderen Sitzung replizieren. Wenn Sie den vom Zufallszahlengenerator verwendeten Startwert angeben, stellen Sie sicher, dass bei jeder Ausführung des Knotens dieselben Datensätze zugewiesen werden. Geben Sie den gewünschten Startwert ein oder klicken Sie auf die Schaltfläche **Generieren**, um automatisch einen Startwert zu generieren. Wenn diese Option nicht ausgewählt ist, wird bei jeder Ausführung des Knotens eine andere Stichprobe generiert.

Hinweis: Bei Verwendung der Option **Startwert für Zufallsgenerator festlegen** mit Datensätzen, die aus einer Datenbank eingelesen wurden, ist möglicherweise vor der Stichprobenziehung ein Sortierknoten erforderlich, um zu gewährleisten, dass bei jeder Ausführung des Knotens dasselbe Ergebnis erzielt wird. Dies liegt daran, dass der Startwert für den Zufallsgenerator von der Reihenfolge der Datensätze abhängt, die in relationalen Datenbanken nicht unbedingt gleich bleibt. Weitere Informationen finden Sie im Thema „Sortierknoten" auf Seite 78.

SQL aktivieren für die Zuweisung von Datensätzen zu Partitionen. (nur für Tier-1-Datenbanken) Aktivieren Sie dieses Kontrollkästchen, um SQL-Pushback für die Zuweisung von Datensätzen zu Partitionen zu verwenden. Wählen Sie in der Dropdown-Liste **Eindeutiges Feld** ein Feld mit eindeutigen Werten (z. B. ein ID-Feld), um sicherzustellen, dass Datensätze zufällig aber auf wiederholbare Weise zugewiesen werden.

Datenbankstufen werden in der Beschreibung des Quellenknotens "Datenbank" erläutert. Weitere Informationen finden Sie im Thema „Datenbankquellenknoten" auf Seite 13.

Generieren von Auswahlknoten

Mithilfe des Menüs "Generieren" im Partitionsknoten können Sie automatisch einen Auswahlknoten für jede Partition erstellen. Beispielsweise können Sie alle Datensätze in der Trainingspartition auswählen, um unter Verwendung nur dieser Partition eine weitere Evaluierung bzw. weitere Analysen zu erstellen.

Dichotomknoten

Der Dichotomknoten wird zur Ableitung von Flagfeldern auf der Grundlage der Kategoriewerte verwendet, die für ein oder mehrere nominale Felder definiert sind. Ihr Dataset könnte beispielsweise das nominale Feld *Blutdruck* mit den Werten *Hoch*, *Normal* und *Niedrig* enthalten. Zur Erleichterung der Datenbearbeitung können Sie ein Flagfeld für hohen Blutdruck erstellen, das angibt, ob der Patient unter hohem Blutdruck leidet oder nicht.

Festlegen der Optionen für den Dichotomknoten

Setfelder. Listet alle Datenfelder mit einem Messniveau von *Nominal* (Set) auf. Wählen Sie eines aus der Liste aus, um die Werte im Set anzuzeigen. Sie können eine Auswahl aus diesen Werten treffen, um ein Flagfeld zu erstellen. Dabei müssen die Daten mithilfe eines aufwärts liegenden Quellen- oder Typknotens vollständig instanziiert werden, bevor die verfügbaren nominalen Felder (und die zugehörigen Werte) sichtbar werden. Weitere Informationen finden Sie im Thema „Typknoten" auf Seite 118.

Feldnamenerweiterung. Bei Auswahl dieser Option werden Steuerelemente zur Angabe einer Erweiterung aktiviert, die dem neuen Flagfeld als Suffix oder Präfix hinzugefügt werden kann. Standardmäßig werden neue Feldnamen automatisch erstellt, indem der ursprüngliche Feldname mit dem Feldwert zu einer Beschriftung kombiniert wird, wie beispielsweise *Feldname_Feldwert*.

Verfügbare Setwerte. Die Werte im oben ausgewählten Set werden hier angezeigt. Wählen Sie einen oder mehrere Werte aus, für die Flags generiert werden sollen. Wenn die Werte in einem Feld namens *Blutdruck* beispielsweise *Hoch*, *Mittel* und *Niedrig* lauten, dann können Sie *Hoch* auswählen und zu der Liste auf der rechten Seite hinzufügen. Dadurch wird ein Feld mit einem Flag für Datensätze erstellt, die einen Wert enthalten, der auf einen hohen Blutdruck hinweist.

Flagfelder erstellen. Die neu erstellten Flagfelder werden hier aufgeführt. Mit den Steuerelementen für die Feldnamenerweiterung können Sie Optionen zur Benennung des neuen Felds angeben.

Wahr-Wert. Dient zur Angabe des Wahr-Werts, den der Knoten zum Festlegen eines Flags verwendet. Standardmäßig lautet dieser Wert **T**.

Falsch-Wert. Dient zur Angabe des Falsch-Werts, den der Knoten zum Festlegen eines Flags verwendet. Standardmäßig lautet dieser Wert **F**.

Aggregationsschlüssel. Mit dieser Option werden Gruppen anhand der unten angegebenen Schlüsselfelder zu Gruppen zusammengefasst. Bei Auswahl von **Aggregationsschlüssel** werden alle Flagfelder in einer Gruppe aktiviert, wenn *ein beliebiger* Datensatz auf "wahr" gesetzt wurde. Mit der Feldauswahlschaltfläche können Sie angeben, welche Schlüsselfelder zur Aggregierung von Datensätzen verwendet werden.

Umstrukturierungsknoten

Mit dem Umstrukturierungsknoten erzeugen Sie mehrere Felder auf der Grundlage der Werte eines nominalen oder Flagfelds. Diese neu erzeugten Felder können Werte aus einem anderen Feld enthalten oder auch ein numerisches Flag (0 oder 1). Dieser Knoten weist ähnliche Funktionen auf wie der Dichotomknoten. Er bietet jedoch größere Flexibilität. Hiermit können Sie Felder mit einem beliebigen Typ (auch numerische Flags) anhand der Werte aus einem anderen Feld anlegen. Anschließend können Sie die Aggregation oder andere Bearbeitungsschritte für andere abwärts liegende Knoten durchführen. (Mit dem Dichotomknoten können Sie Felder in einem einzigen Schritt aggregieren, was beim Erstellen von Flagfeldern nützlich ist.)

Das folgende Dataset enthält beispielsweise ein nominales Feld *Account* mit den Werten *Savings* und *Draft*. Für jedes Konto werden der Anfangssaldo und der aktuelle Saldo festgehalten; einige Kunden besitzen mehrere Konten von jedem Typ. Angenommen, Sie möchten erfahren, ob ein Kunde ein Konto mit einem bestimmten Typ besitzt und, wenn ja, wie hoch der Saldo in jedem Kontentyp ist. Mit dem Umstrukturierungsknoten erzeugen Sie je ein Feld für die Werte für *Account* und Sie wählen den Wert *Current_Balance* aus. In jedes neue Feld wird der aktuelle Saldo für den jeweiligen Datensatz eingetragen.

Tabelle 23. Beispieldaten vor der Umstrukturierung.

CustID	Account	Open_Bal	Current_Bal
12701	Text	1000	1005,32
12702	Savings	100	144,51
12703	Savings	300	321,20
12703	Savings	150	204,51
12703	Text	1200	586,32

Tabelle 24. Beispieldaten nach der Umstrukturierung.

CustID	Account	Open_Bal	Current_Bal	Account_Draft_ Current_Bal	Account_Savings _Current_Bal
12701	Text	1000	1005,32	1005,32	$null$
12702	Savings	100	144,51	$null$	144,51
12703	Savings	300	321,20	$null$	321,20
12703	Savings	150	204,51	$null$	204,51
12703	Text	1200	586,32	586,32	$null$

Verwenden des Umstrukturierungsknotens mit dem Aggregatknoten

In vielen Fällen soll der Umstrukturierungsknoten mit einem Aggregatknoten gekoppelt werden. Im obigen Beispiel besitzt ein Kunde (mit der ID 12703) drei Konten. Mit einem Aggregatknoten können Sie den Gesamtsaldo für jeden Kontentyp berechnen. Das Schlüsselfeld ist *CustID* und die Aggregatfelder sind die soeben umstrukturierten Felder *Account_Draft_Current_Bal* und *Account_Savings_Current_Bal*. Die nachstehende Tabelle zeigt die Ergebnisse.

Tabelle 25. Beispieldaten nach der Umstrukturierung und Aggregation.

CustID	Record_Count	Account_Draft_Current_ Bal_Sum	Account_Savings_Current_ Bal_Sum
12701	E	1005,32	$null$
12702	E	$null$	144,51
12703	3	586,32	525,71

Festlegen von Optionen für den Umstrukturierungsknoten

Verfügbare Felder. Listet alle Datenfelder mit einem Messniveau von *Nominal* (Set) oder *Flag* auf. Wählen Sie ein Feld in der Liste aus, sodass die Werte im Set oder im Flag angezeigt werden, und wählen Sie die gewünschten Werte für die Erstellung der umstrukturierten Felder aus. Dabei müssen die Daten mithilfe eines aufwärts liegenden Quellen- oder Typknotens vollständig instanziiert werden, bevor die verfügbaren Felder (und die zugehörigen Werte) sichtbar werden. Weitere Informationen finden Sie im Thema „Typknoten" auf Seite 118.

Verfügbare Werte. Die Werte im oben ausgewählten Set werden hier angezeigt. Wählen Sie einen oder mehrere Werte aus, für die umstrukturierte Felder generiert werden sollen. Wenn die Werte im Feld *Blutdruck* beispielsweise *Hoch*, *Mittel* und *Niedrig* lauten, dann können Sie *Hoch* auswählen und zu der Liste auf der rechten Seite hinzufügen. Dadurch wird ein Feld mit einem bestimmten Wert (siehe unten) für Datensätze erstellt, die den Wert *Hoch* enthalten.

Umstrukturierte Felder erstellen. Hier werden die soeben erstellten, umstrukturierten Felder aufgeführt. Standardmäßig werden neue Feldnamen automatisch erstellt, indem der ursprüngliche Feldname mit dem Feldwert zu einer Beschriftung kombiniert wird, wie beispielsweise *Feldname_Feldwert*.

Feldnamen einschließen. Inaktivieren Sie diese Option, wenn der ursprüngliche Feldname den neuen Feldnamen nicht als Präfix vorangestellt werden soll.

Werte aus anderen Feldern verwenden. Geben Sie mindestens ein Feld an, dessen Werte in die umstrukturierten Felder eingetragen werden sollen. Mit der Feldauswahl können Sie das oder die gewünschten Felder bestimmen. Für jedes angegebene Feld wird ein neues Feld erstellt. Der Name des Feldes, aus dem die Werte stammen, wird an den Namen des umstrukturierten Feldes angehängt, beispielsweise *Blutdruck_Hohes_Alter* oder *Blutdruck_Niedriges_Alter*. Jedes neue Feld übernimmt den Typ des Felds, aus dem der ursprüngliche Wert stammt.

Flags für numerische Werte erstellen. Wenn Sie diese Option auswählen, wird kein Wert aus einem anderen Feld übernommen, sondern die neuen Felder werden mit numerischen Wertflags (0 für Falsch, 1 für Wahr) gefüllt.

Transponierknoten

Standardmäßig bestehen die Spalten aus Feldern und die Zeilen aus Datensätzen oder Beobachtungen. Falls notwendig, können Sie mithilfe eines Transponierknotens die Daten in Zeilen und Spalten vertauschen, sodass aus Feldern Datensätze und aus Datensätzen Felder werden. Wenn Sie beispielsweise Zeitreihendaten verwenden, in der die Zeitreihen jeweils eine Zeile darstellen (also keine Spalte), können Sie die Daten vor der Analyse transponieren.

Festlegen von Optionen für Transponierknoten

Neue Feldnamen

Neue Feldnamen können automatisch auf der Grundlage eines angegebenen Präfixes generiert oder aus einem bestehenden Feld in den Daten eingelesen werden.

Präfix verwenden. Mit dieser Option werden die neuen Feldnamen automatisch auf der Grundlage des angegebenen Präfixes erstellt (*Field1*, *Field2* usw.). Sie können das Präfix ganz nach Bedarf anpassen. Bei dieser Option muss die Anzahl der zu erstellenden Felder angegeben werden, unabhängig von der Anzahl der Zeilen in den ursprünglichen Daten. Wenn Sie unter **Anzahl neuer Felder** beispielsweise den Wert 100 festlegen, werden alle Daten ab der 101. Zeile verworfen. Enthalten die Originaldaten weniger als 100 Zeilen, bleiben einige Felder leer. (Sie können die Anzahl der Felder ganz nach Bedarf anpassen. Mit dieser Einstellung soll vermieden werden, dass z. B. eine Million Datensätze in eine Million Felder transponiert werden, was zu unüberschaubaren Ergebnissen führen würde.)

Beispiel: Angenommen, Ihnen liegen Daten mit Zeitreihen in den Zeilen und einem separaten Feld (Spalte) für jeden Monat vor. Sie können diese Daten transponieren, sodass jede Zeitreihe in einem separaten Feld (mit einer Zeile für jeden Monat) vorliegt.

Lesen aus Feld. Liest die Feldnamen aus einem vorhandenen Feld. Bei dieser Option wird die Anzahl der neuen Felder durch die Daten bestimmt (bis zum angegebenen Höchstwert). Jeder Wert im ausgewählten Feld wird zu einem neuen Feld in den Ausgabedaten. Das ausgewählte Feld kann einen beliebigen Speichertyp besitzen (ganze Zahl, Zeichenfolge, Datum usw.); um doppelte Feldnamen zu vermeiden, müssen die Werte im ausgewählten Feld jedoch eindeutig sein. (Die Anzahl der Werte soll also mit der Anzahl der Zeilen übereinstimmen.) Falls doppelte Feldnamen auftreten, wird eine Warnnachricht angezeigt.

- **Werte lesen.** Falls das ausgewählte Feld nicht instanziiert wurde, wählen Sie diese Option, damit die Liste der neuen Feldnamen gefüllt wird. Wurde das Feld bereits instanziiert, ist dieser Schritt nicht notwendig.
- **Maximale Anzahl zu lesender Werte.** Beim Einlesen von Feldnamen aus den Daten wird eine Obergrenze angegeben, um das Erstellen einer übermäßig großen Anzahl von Feldern zu verhindern. (Wie bereits beschrieben, würde das Transponieren von einer Million Datensätzen in eine Million Felder zu unüberschaubaren Ergebnissen führen.)

Wenn beispielsweise in der ersten Spalte der Daten der Name für die einzelnen Zeitreihen angegeben wird, können diese Werte in den transponierten Daten als Feldnamen verwendet werden.

Transponieren. Standardmäßig werden nur stetige Felder (numerischer Bereich) transponiert (ganze Zahl oder reelle Zahl als Speichertyp). Optional können Sie stattdessen ein Subset numerischer Felder auswählen oder auch Zeichenfolgenfelder transponieren. Alle transponierten Felder müssen allerdings denselben Speichertyp aufweisen (entweder numerisch oder Zeichenfolge, nicht jedoch beide Typen), weil durch das Mischen der Eingabefelder gemischte Werte in den Ausgabespalten entstünden, was die Regel verletzt, dass alle Werte eines Feldes denselben Speichertyp besitzen müssen. Andere Speichertypen (Datum, Zeit, Zeitmarke) können nicht transponiert werden.

- **Nur numerisch.** Transponiert alle numerischen Felder (ganze Zahl oder reelle Zahl als Speichertyp). Die Anzahl der Zeilen in der Ausgabe entspricht der Anzahl der numerischen Felder in den ursprünglichen Daten.
- **Nur Zeichenfolgen.** Transponiert alle Zeichenfolgenfelder.
- **Benutzerdefiniert.** Ermöglicht die Auswahl eines Subsets von numerischen Feldern. Die Anzahl der Zeilen in der Ausgabe entspricht der Anzahl der ausgewählten Felder. *Hinweis:* Diese Option ist nur für numerische Felder verfügbar.

Zeilen-ID-Name. Gibt den Namen des vom Knoten erstellten Zeilen-ID-Felds an. Die Werte für dieses Feld ergeben sich aus den Namen der Felder in den ursprünglichen Daten.

Tipp: Wenn beim Transponieren von Zeitreihendaten von Zeilen in Spalten die ursprünglichen Daten eine Zeile (z. B. Datum, Monat, Jahr) enthalten, die die Zeitperiode für die einzelnen Messungen angibt, müssen Sie sicherstellen, dass diese Beschriftungen als Feldnamen in IBM SPSS Modeler eingelesen werden (wie in den oben stehenden Beispielen beschrieben, bei denen Monat bzw. Datum als Feldnamen in den ursprünglichen Daten angezeigt werden) und nicht etwa in die erste Datenzeile aufgenommen werden. Dadurch wird eine Vermischung von Beschriftungen und Werten in den einzelnen Spalten vermieden (die dazu führen würde, dass Zahlen als Zeichenfolgen gelesen würden, da innerhalb einer Spalte nicht verschiedene Speichertypen vorkommen dürfen).

Zeitintervallknoten

Mit dem Zeitintervallknoten können Sie Intervalle angeben und Beschriftungen für Zeitreihendaten generieren, die in einem Zeitreihenmodellierungsknoten, einem Streaming-ZR-Knoten oder einem Zeitdiagrammknoten zu Schätz- oder Vorhersagezwecken verwendet werden sollen. Die unterstützten Zeitintervalle reichen dabei von Sekunden bis hin zu Jahren. Wenn Sie beispielsweise eine Zeitreihe mit täglichen Messungen ausführen, die am 3. Januar 2005 begann, können Sie die Datensätze ab diesem Datum be-

schriften. Die zweite Zeile stünde dann für den 4. Januar usw. Darüber hinaus können Sie die Periodizität bestimmen, z. B. fünf Tage pro Woche oder acht Stunden pro Tag.

Des Weiteren können Sie den Bereich der Datensätze angeben, der für die Schätzung verwendet werden soll. Sie können auswählen, ob die frühesten Datensätze in der Zeitreihe ausgeschlossen werden sollen und ob Holdouts angegeben werden sollen. Dadurch können Sie das Modell testen, indem Sie die aktuellsten Datensätze in den Zeitreihendaten zurückhalten, um ihre bekannten Werte mit den geschätzten Werten für die betreffenden Zeitperioden zu vergleichen.

Außerdem können Sie angeben, für wie viele in die Zukunft reichende Zeiträume die Vorhersage erstellt werden soll, und Sie können zukünftige Werte angeben, die für die Vorhersage durch nachgeordnete Zeitreihenmodellierungsknoten oder Streaming-ZR-Knoten verwendet werden sollen.

Im Zeitintervallknoten wird ein *TimeLabel*-Feld in einem geeigneten Format für das angegebene Intervall und die Periode generiert, außerdem ein *TimeIndex*-Feld, mit dem jedem Datensatz eine eindeutige ganze Zahl zugewiesen wird. Auch einige zusätzliche Felder werden gegebenenfalls erzeugt, abhängig vom ausgewählten Intervall bzw. der Periodizität (z. B. die Minute oder Sekunde, in die eine Messung fällt).

Sie können die Werte auffüllen oder aggregieren, um so sicherzustellen, dass die Messungen in gleichmäßigen Zeitabständen erfolgen. Bei den Methoden zur Modellierung von Zeitreihendaten ist ein einheitliches Intervall zwischen den Messungen erforderlich; fehlende Werte werden durch leere Zeilen dargestellt. Falls Ihre Daten diese Bedingung nicht bereits erfüllen, können Sie sie mithilfe dieses Knotens entsprechend transponieren.

Kommentare

- Die periodischen Intervalle stimmen unter Umständen nicht mit der Echtzeit überein. Bei einer Reihe, die auf einer normalen Fünf-Tage-Arbeitswoche beruht, würde die Lücke zwischen Freitag und Montag als ein einziger Tag behandelt.
- Im Zeitintervallknoten wird vorausgesetzt, dass sich jede Zeitreihe in einem Feld oder einer Spalte befindet und dabei jeweils eine Zeile für jede Messung vorliegt. Falls notwendig, können Sie die Daten so transponieren, dass diese Anforderung erfüllt ist. Weitere Informationen finden Sie im Thema „Transponierknoten” auf Seite 158.
- Bei Zeitreihen mit ungleichmäßigen Abständen können Sie ein Feld definieren, aus dem das Datum oder die Uhrzeit für die jeweilige Messung hervorgeht. Beachten Sie, dass hierfür ein Datums-, Zeit- oder Zeitmarkenfeld im entsprechenden Format zur Verwendung als Eingabe erforderlich ist. Falls erforderlich, können Sie ein bestehendes Feld (beispielsweise ein Beschriftungsfeld vom Typ "Zeichenfolge") mithilfe eines Füllerknotens in dieses Format konvertieren. Weitere Informationen finden Sie im Thema „Speichertypkonvertierung mithilfe des Füllerknotens” auf Seite 140.
- Beim Betrachten der Details für die erzeugten Beschriftungs- und Indexfelder hilft es häufig, wenn Sie die Anzeige der Wertbeschriftungen aktivieren. Wenn Sie beispielsweise eine Tabelle mit Werten betrachten, die für monatliche Daten erzeugt wurde, können Sie mit dem Symbol für die Wertbeschriftungen in der Symbolleiste die Beschriftungen *Januar, Februar, März* usw. einblenden statt *1, 2, 3* usw.

Abbildung 6. Symbol für Wertbeschriftungen

Festlegen von Zeitintervallen

Auf der Registerkarte "Intervalle" bestimmen Sie das Intervall und die Periodizität zum Aufbauen oder Beschriften der Zeitreihe. Die jeweiligen Einstellungen sind abhängig vom ausgewählten Intervall. Bei der Option **Stunden pro Tag** können Sie beispielsweise die Anzahl der Tage in der Woche festlegen, außerdem die Anzahl der Stunden pro Tag sowie die Stunde, zu der der Tag beginnt. Weitere Informationen finden Sie im Thema „Unterstützte Intervalle” auf Seite 164.

Beschriften oder Aufbauen der Zeitreihe

Sie können die Datensätze nacheinander beschriften oder auch die Zeitreihe auf der Grundlage eines angegebenen Datums-, Zeitmarken- oder Zeitfelds erstellen.

- **Beschriftung bei erstem Datensatz beginnen.** Legen Sie das Anfangsdatum und/oder die Anfangsuhrzeit fest, mit der die aufeinander folgenden Datensätze beschriftet werden sollen. Bei einer Beschriftung pro Tag geben Sie beispielsweise das Datum und die Stunde an, zu der die Zeitreihe beginnt. Für jede nachfolgende Stunde wird dann ein neuer Datensatz angelegt. Bei dieser Methode werden lediglich die Beschriftungen hinzugefügt; ansonsten bleiben die ursprünglichen Daten unverändert. Es wird stattdessen angenommen, dass die Datensätze bereits gleiche Abstände zueinander aufweisen, dass also gleiche Intervalle zwischen den Messungen vorliegen. Fehlende Messwerte müssen durch leere Zeilen in den Daten ersetzt werden.
- **Aus Daten erstellen.** Bei Zeitreihen mit ungleichmäßigen Abständen können Sie ein Feld definieren, aus dem das Datum oder die Uhrzeit für die jeweilige Messung hervorgeht. Beachten Sie, dass hierfür ein Datums-, Zeit- oder Zeitmarkenfeld im entsprechenden Format zur Verwendung als Eingabe erforderlich ist. Wenn Ihnen beispielsweise ein Zeichenfolgenfeld mit Werten wie *Jan 2000*, *Feb 2000* usw. vorliegt, können Sie dieses mithilfe eines Füllerknotens in ein Datumsfeld konvertieren. Weitere Informationen finden Sie im Thema „Speichertypkonvertierung mithilfe des Füllerknotens" auf Seite 140. Mit der Option **Aus Daten erstellen** werden die Daten ebenfalls gemäß dem angegebenen Intervall transformiert, indem Datensätze je nach Bedarf aufgefüllt oder aggregiert werden, beispielsweise durch Zusammenfassen von Wochen zu Monaten oder durch Ersetzen fehlender Datensätze durch Leerwerte oder extrapolierte Werte. Auf der Registerkarte "Aufbauen" legen Sie die Funktionen fest, mit denen die Datensätze aufgefüllt oder aggregiert werden. Weitere Informationen finden Sie im Thema „Aufbauoptionen für Zeitintervalle".

Neue Feldnamenerweiterung. Hiermit legen Sie ein Präfix oder ein Suffix fest, das für alle durch den Knoten erzeugten Felder übernommen wird. Wenn Sie beispielsweise das Standardpräfix *$TI_* verwenden, erhalten die durch den Knoten erzeugten Felder die Bezeichnung *$TI_TimeIndex*, *$TI_TimeLabel* usw.

Datumsformat. Bestimmt das Format für das durch den Knoten erstellte *TimeLabel*-Feld gemäß dem aktuellen Intervall. Die Verfügbarkeit dieser Option ist abhängig von der aktuellen Auswahl.

Zeitformat. Bestimmt das Format für das durch den Knoten erstellte *TimeLabel*-Feld gemäß dem aktuellen Intervall. Die Verfügbarkeit dieser Option ist abhängig von der aktuellen Auswahl.

Aufbauoptionen für Zeitintervalle

Auf der Registerkarte "Aufbauen" im Zeitintervallknoten legen Sie Optionen fest, mit denen die Felder gemäß dem angegebenen Intervall aggregiert oder aufgefüllt werden. Diese Einstellungen gelten nur dann, wenn auf der Registerkarte "Intervalle" die Option **Aus Daten erstellen** aktiviert ist. Liegt beispielsweise eine Mischung aus Wochen- und Monatsdaten vor, können Sie die Wochenwerte so aggregieren (zusammenfassen), dass ein gleichmäßiges monatliches Intervall entsteht. Alternativ können Sie ein Wochenintervall festlegen und die Zeitreihe auffüllen, indem Sie Leerwerte für die fehlenden Wochen einfügen oder fehlende Werte mithilfe einer bestimmten Auffüllfunktion extrapolieren.

Beim Auffüllen und Aggregieren der Daten werden die vorhandenen Datums- oder Zeitmarkenfelder effektiv durch die erzeugten *TimeLabel*- und *TimeIndex*-Felder überschrieben und aus der Ausgabe entfernt. Auch Felder ohne Typ werden herausgenommen. Felder, mit denen ein Zeitraum gemessen wird (z. B. ein Feld, das nicht den Zeitpunkt festhält, zu dem ein Service-Call begann, sondern die Dauer dieses Gesprächs), werden beibehalten, sofern sie intern als Zeitfelder statt als Zeitmarkenfelder gespeichert sind. Weitere Informationen finden Sie im Thema „Festlegen von Feldspeicher und Formatierung" auf Seite 24. Die anderen Felder werden gemäß den Optionen auf der Registerkarte "Aufbauen" aggregiert.

- **Standardfelder und -funktionen verwenden.** Gibt an, dass alle Felder je nach Bedarf aggregiert oder aufgefüllt werden sollen, ausgenommen Datums- und Zeitmarkenfelder sowie Felder ohne Typ, wie oben beschrieben. Die Standardfunktion wird gemäß dem Messniveau angewendet. Stetige Felder wer-

den beispielsweise anhand des Mittelwerts aggregiert, nominale Felder dagegen anhand des Modus. Im unteren Teil des Dialogfelds können Sie die Standardeinstellungen für die Messniveaus ändern.

- **Felder und Funktionen angeben.** Mit dieser Option können Sie die aufzufüllenden oder zu aggregierenden Felder auswählen und auch die jeweilige Funktion festlegen. Alle nicht ausgewählten Felder werden aus der Ausgabe herausgenommen. Mit den Symbolen rechts können Sie Felder in die Tabelle aufnehmen oder daraus entfernen. Wenn Sie auf eine Zelle in einer bestimmten Spalte klicken, können Sie die Aggregations- oder Auffüllfunktion für dieses Feld ändern und somit die Standardeinstellung außer Kraft setzen. Felder ohne Typ werden aus der Liste ausgeschlossen und können nicht zur Tabelle hinzugefügt werden.

Standard. Bestimmt die Aggregations- und Auffüllfunktionen, die standardmäßig für die verschiedenen Feldtypen verwendet werden. Diese Standardeinstellungen werden angewendet, wenn Sie die Option **Standards verwenden** aktivieren, und gelten auch als anfängliche Standardeinstellungen für alle Felder, die neu in die Tabelle aufgenommen werden. (Wenn Sie die Standardeinstellungen ändern, wirkt sich dies nicht auf die Einstellungen in der Tabelle aus, sondern nur auf die nachfolgend hinzugefügten Felder.)

Aggregationsfunktion. Die folgenden Aggregationsfunktionen stehen zur Verfügung:

- **Stetig.** Verfügbare Funktionen für stetige Felder: **Mittelwert**, **Summe**, **Modus**, **Min** und **Max**.
- **Nominal.** Verfügbare Optionen: **Modus**, **Erste** und **Letzte**. "Erste" bezieht sich auf den ersten Wert ungleich null in der (nach Datum sortierten) Aggregationsgruppe, "Letzte" entsprechend auf den letzten Wert ungleich null in dieser Gruppe.
- **Flag.** Verfügbare Optionen: **Wahr, wenn beliebige wahr**, **Modus**, **Erste** und **Letzte**.

Auffüllfunktion. Die folgenden Auffüllfunktionen stehen zur Verfügung:

- **Stetig.** Verfügbare Optionen: **Leer** und **Mittelwert der zuletzt verwendeten Punkte**; hierbei wird der Mittelwert der drei jüngsten Werte ungleich null vor der zu erstellenden Zeitperiode gebildet. Falls weniger als drei Werte vorliegen, ist der neue Wert leer. Zu den letzten Werten zählen nur "echte" Werte; zuvor erstellte aufgefüllte Werte werden bei der Suche nach Werten ungleich null nicht berücksichtigt.
- **Nominal. Leer** und **Zuletzt verwendeter Wert**. "Zuletzt verwendet" bezieht sich auf den jüngsten Wert ungleich null vor der zu erstellenden Zeitperiode. Auch hier werden nur echte Werte berücksichtigt.
- **Flag.** Verfügbare Optionen: **Leer**, **Wahr** und **Falsch**.

Maximale Anzahl der Datensätze in resultierendem Dataset. Bestimmt die maximal zulässige Anzahl an erstellten Datensätzen, die ansonsten sehr groß würde, insbesondere dann, wenn Sekunden (absichtlich oder versehentlich) als Zeitintervall eingestellt wurden. Bei einer Zeitreihe mit nur zwei Werten (1. Jan. 2000 und 1. Jan. 2001) würden entsprechend 31.536.000 Datensätze erzeugt, wenn die Daten auf Sekunden aufgefüllt würden (60 Sekunden x 60 Minuten x 24 Stunden x 365 Tage). Sobald der angegebene Höchstwert erreicht ist, wird die Verarbeitung unterbrochen und eine Warnnachricht wird angezeigt.

Anzahlfeld

Beim Aggregieren oder Auffüllen von Werten wird ein neues *Anzahlfeld* erzeugt, aus dem die Anzahl der Datensätze hervorgeht, die beim Ermitteln des neuen Datensatzes berücksichtigt werden. Wenn Sie beispielsweise vier Wochenwerte zu einem einzigen Monat aggregieren, wäre die Anzahl gleich 4. Bei einem aufgefüllten Datensatz ist die Anzahl gleich 0. Der Name des Felds setzt sich aus der Bezeichnung *Anzahl* und dem auf der Registerkarte "Intervall" angegebenen Präfix oder Suffix zusammen.

Schätzperiode

Auf der Registerkarte "Schätzung" des Zeitintervallknotens können Sie den Bereich der in der Modellschätzung verwendeten Datensätze sowie etwaige Holdouts angeben. Diese Einstellungen können bei Bedarf in nachgeordneten Modellierungsknoten überschrieben werden, es ist jedoch zumeist praktischer, sie hier anzugeben, als für jeden Knoten einzeln.

Schätzung beginnen. Sie können die Schätzperiode am Anfang der Daten beginnen oder ältere Werte ausschließen, die bei der Vorhersage nur von begrenztem Nutzen sind. Je nach den jeweils vorliegenden Daten kann eine Verkürzung der Schätzperiode die Leistung beschleunigen (und den für die Datenvorbereitung erforderlichen Zeitaufwand verkürzen), ohne dass signifikante Einbußen bei der Vorhersagegenauigkeit auftreten.

Schätzung beenden. Sie können das Modell mit allen Datensätzen bis zum Ende der Daten schätzen oder die aktuellsten Datensätze als Holdout zurückhalten, um damit das Modell zu evaluieren. In letzterem Fall "sagen" Sie im Grunde Werte "vorher", die bereits bekannt sind. Auf diese Weise können Sie die beobachteten und die vorhergesagten Werte vergleichen, um die Effektivität des Modells abzuschätzen.

Vorhersagen

Auf der Registerkarte "Vorhersage" des Zeitintervallknotens können Sie die Anzahl der Datensätze angeben, die Sie vorhersagen möchten, sowie die zukünftigen Werte, die bei der Vorhersage durch die nachgeordneten Zeitreihenmodellierungsknoten oder Streaming-ZR-Knoten verwendet werden sollen. Diese Einstellungen können bei Bedarf in nachgeordneten Knoten überschrieben werden, es ist jedoch zumeist praktischer, sie hier anzugeben, als für jeden Knoten einzeln.

Datensätze auf die Zukunft ausdehnen. Gibt an, wie viele Datensätze über die Schätzperiode hinaus vorhergesagt werden sollen. Beachten Sie, dass es von der Anzahl der auf der Registerkarte "Schätzung" angegebenen Holdouts abhängt, ob diese Datensätze tatsächlich "vorhergesagt" werden.

Zukunftsindikatorfeld. Beschriftung des generierten Felds, das angibt, ob ein Datensatz Vorhersagedaten enthält. Der Standardwert für die Beschriftung lautet *$TI_Zukunft*.

Zukünftige Werte für Verwendung in Vorhersagen. Für jeden Datensatz, den Sie vorhersagen möchten (ausgenommen Holdouts) müssen Sie bei Verwendung von Prädiktorfeldern (mit der Rolle auf *Eingabe* eingestellt) für jeden Prädiktor geschätzte Werte für die Vorhersageperiode angeben. Sie können die Werte entweder manuell angeben oder aus einer Liste auswählen.

- **Feld.** Klicken Sie auf die Feldauswahlschaltfläche und wählen Sie alle Felder aus, die als Prädiktor verwendet werden können. Beachten Sie, dass die hier ausgewählten Felder nicht unbedingt bei der Modellierung verwendet werden. Damit ein Feld tatsächlich als Prädiktor verwendet wird, muss es in einem nachgeordneten Modellierungsknoten ausgewählt werden. Dieses Dialogfeld bietet einfach eine praktische Möglichkeit zur Angabe zukünftiger Werte, damit diese gemeinsam von mehreren nachgeordneten Modellierungsknoten verwendet werden können und nicht separat in jedem Knoten angegeben werden müssen. Beachten Sie, dass die Liste der verfügbaren Felder durch die auf der Registerkarte "Aufbauen" getroffene Auswahl eingeschränkt sein kann. Wenn beispielsweise auf der Registerkarte "Aufbauen" die Option **Felder und Funktionen angeben** ausgewählt wurde, werden alle Felder, die nicht aggregiert oder aufgefüllt wurden, aus dem Stream verworfen und können nicht bei der Modellierung verwendet werden.

  Hinweis: Wenn für ein Feld, das nicht mehr im Stream verfügbar ist (weil es verworfen wurde oder weil auf der Registerkarte "Aufbauen" eine neue Auswahl getroffen wurde), zukünftige Werte angegeben werden, wird das Feld auf der Registerkarte "Vorhersage" in roter Farbe angezeigt.
- **Werte.** Sie können bei jedem Feld aus einer Liste von Funktionen wählen oder auf **Angeben** klicken, um entweder Werte manuell einzugeben oder eine Auswahl aus einer Liste vordefinierter Werte zu treffen. Wenn die Prädiktorfelder sich auf Elemente beziehen, über die Sie die Kontrolle haben oder die anderweitig im Voraus bekannt sind, sollten Sie die Werte manuell eingeben. Wenn Sie beispielsweise die Einnahmen des nächsten Monats für ein Hotel auf der Grundlage der Anzahl der Zimmerreservierungen vorhersagen, können Sie die Anzahl der Reservierungen angeben, die Ihnen tatsächlich für diesen Zeitraum vorliegen. Wenn ein Prädiktorfeld sich dagegen auf Daten bezieht, über die Sie keine Kontrolle haben, wie beispielsweise der Preis einer Aktie, können Sie eine Funktion verwenden, wie beispielsweise den aktuellsten Wert oder den Mittelwert der aktuellsten Punkte.

  Die verfügbaren Funktionen hängen vom Messniveau des Felds ab.

Tabelle 26. Für Messniveaus verfügbare Funktionen.

Messniveau	Funktionen
Stetiges oder nominales Feld	leer Mittelwert der zuletzt verwendeten Punkte Zuletzt verwendeter Wert Angeben
Flagfelder	leer Zuletzt verwendeter Wert Wahr Falsch Angeben

Mittelwert der zuletzt verwendeten Punkte - Berechnet den zukünftigen Wert aus dem Mittelwert der letzten drei Datenpunkte.

Zuletzt verwendeter Wert - Legt den Wert des aktuellsten Datenpunkts als zukünftigen Wert fest.

Wahr/Falsch - Setzt den zukünftigen Wert eines Flagfelds je nach Angabe auf "Wahr" bzw. "Falsch".

Angeben - Öffnet ein Dialogfeld, in dem Sie zukünftige Werte manuell eingeben oder aus einer vordefinierten Liste auswählen können.

Zukünftige Werte

Hier können Sie zukünftige Werte angeben, die für die Vorhersage durch nachgeordnete Zeitreihenmodellierungsknoten oder Streaming-ZR-Knoten verwendet werden sollen. Diese Einstellungen können bei Bedarf in den nachgeordneten Knoten überschrieben werden, es ist jedoch zumeist praktischer, sie hier anzugeben, als für jeden Knoten einzeln.

Sie können die Werte manuell eingeben oder auf die Auswahlschaltfläche rechts neben dem Dialogfeld klicken, um eine Auswahl aus einer Liste von Werten zu treffen, die für das aktuelle Feld definiert wurden.

Die Anzahl der zukünftigen Werte, die Sie angeben können, entspricht der Anzahl der Datensätze, um die Sie die Zeitreihe auf die Zukunft ausdehnen.

Unterstützte Intervalle

Der Zeitintervallknoten unterstützt die vollständige Palette an Intervallen von Sekunden bis hin zu Jahren, sowie zyklische (z. B. saisonale) und nicht zyklische Perioden. Das Intervall wird auf der Registerkarte "Intervalle" im Feld "Zeitintervall" angegeben.

Perioden

Mit der Option **Perioden** beschriften Sie eine vorhandene, nicht zyklische Zeitreihe, die nicht unter die anderen angegebenen Intervalle fällt. Die Zeitreihe muss bereits die richtige Reihenfolge aufweisen und ein einheitliches Intervall zwischen den einzelnen Messungen besitzen. Bei Auswahl dieses Intervalls ist die Option **Aus Daten erstellen** nicht verfügbar.

Beispiel

Die Datensätze werden inkrementell gemäß dem angegebenen Anfangswert beschriftet (*Periode 1*, *Periode 2* usw.). Neue Felder werden wie in der folgenden Tabelle dargestellt erstellt.

Tabelle 27. Beispiele für das Erstellen neuer Felder

$TI_TimeIndex (ganze Zahl)	$TI_TimeLabel (Zeichenfolge)	$TI_Period (ganze Zahl)
E	Periode 1	E
Z	Periode 2	Z
3	Periode 3	3

Tabelle 27. Beispiele für das Erstellen neuer Felder (Forts.)

$TI_TimeIndex (ganze Zahl)	$TI_TimeLabel (Zeichenfolge)	$TI_Period (ganze Zahl)
4	Periode 4	4
5	Periode 5	5

Zyklische Perioden

Mit der Option **Zyklische Perioden** beschriften Sie eine vorhandene Zeitreihe mit einem wiederholenden Zyklus, der nicht unter die Standardintervalle fällt. Verwenden Sie diese Option beispielsweise, wenn Ihr Geschäftsjahr nur 10 Monate umfasst. Die Zeitreihe muss bereits die richtige Reihenfolge aufweisen und ein einheitliches Intervall zwischen den einzelnen Messungen besitzen. (Bei Auswahl dieses Intervalls ist die Option **Aus Daten erstellen** nicht verfügbar.)

Beispiel

Die Datensätze werden inkrementell gemäß dem angegebenen Anfangszyklus und der Periode beschriftet (*Zyklus 1*, *Periode 1*, *Zyklus 1*, *Periode 2* usw.). Wenn beispielsweise 3 Perioden pro Zyklus festgelegt sind, werden neue Felder erstellt, wie in der folgenden Tabelle gezeigt.

Tabelle 28. Beispiele für das Erstellen neuer Felder.

$TI_TimeIndex (ganze Zahl)	$TI_TimeLabel (Zeichenfolge)	$TI_Cycle (ganze Zahl)	$TI_Period (ganze Zahl)
E	Zyklus 1, Periode 1	E	E
Z	Zyklus 1, Periode 2	E	Z
3	Zyklus 1, Periode 3	E	3
4	Zyklus 2, Periode 1	Z	E
5	Zyklus 2, Periode 2	Z	Z

Jahre

Bei den Jahren können Sie das Anfangsjahr festlegen, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option **Aus Daten erstellen** ein Zeitmarken- oder Datumsfeld festlegen, aus dem das Jahr für die einzelnen Datensätze hervorgeht.

Beispiel

Neue Felder werden wie in der folgenden Tabelle dargestellt erstellt.

Tabelle 29. Beispiele für das Erstellen neuer Felder

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Zeichenfolge)	$TI-Year (ganze Zahl)
E	2000	2000
Z	2001	2001
3	2002	2002
4	2003	2003
5	2004	2004

Quartale

Bei einer Zeitreihe im Vierteljahresabstand können Sie den Monat festlegen, in dem das Geschäftsjahr beginnt. Darüber hinaus können Sie das Anfangsquartal und -jahr festlegen (z. B. Q1 2000), sodass die nach-

folgenden Datensätze entsprechend beschriftet werden, oder mit der Option **Aus Daten erstellen** ein Zeitmarken- oder Datumsfeld festlegen, aus dem das Quartal und das Jahr für die einzelnen Datensätze hervorgeht.

Beispiel

Bei einem Geschäftsjahr, das im Januar beginnt, werden neue Felder erstellt und gefüllt, wie in der folgenden Tabelle gezeigt.

Tabelle 30. Beispiele für das Erstellen neuer Felder

$TI-TimeIndex (ganze Zahl)	**$TI-TimeLabel (Zeichenfolge)**	**$TI-Year (ganze Zahl)**	**$TI-Quarter (ganze Zahl mit Beschriftungen)**
E	Q1 2000	2000	1 (Q1)
Z	Q2 2000	2000	2 (Q2)
3	Q3 2000	2000	3 (Q3)
4	Q4 2000	2000	4 (Q4)
5	Q1 2001	2001	1 (Q1)

Beginnt das Geschäftsjahr in einem anderen Monat (z. B. im Juli), werden neue Felder wie in der folgenden Tabelle beschrieben erstellt. Sollen die Beschriftungen für die Monate in den einzelnen Quartalen angezeigt werden, aktivieren Sie die Anzeige der Beschriftungen durch Klicken auf das entsprechende Symbol in der Symbolleiste.

Abbildung 7. Symbol für Wertbeschriftungen

Tabelle 31. Beispiele für das Erstellen neuer Felder

$TI-TimeIndex (ganze Zahl)	**$TI-TimeLabel (Zeichenfolge)**	**$TI-Year (ganze Zahl)**	**$TI-Quarter (ganze Zahl mit Beschriftungen)**
E	Q1 2000/2001	E	1 (Q1 Jul-Sep)
Z	Q2 2000/2001	E	2 (Q2 Okt-Dez)
3	Q3 2000/2001	E	3 (Q3 Jan-Mär)
4	Q4 2000/2001	E	4 (Q4 Apr-Jun)
5	Q1 2001/2002	Z	1 (Q1 Jul-Sep)

Monate

Sie können das Anfangsjahr und den Anfangsmonat festlegen, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option **Aus Daten erstellen** ein Zeitmarken- oder Datumsfeld festlegen, aus dem der Monat für die einzelnen Datensätze hervorgeht.

Beispiel

Neue Felder werden wie in der folgenden Tabelle dargestellt erstellt.

Tabelle 32. Beispiele für das Erstellen neuer Felder

$TI-TimeIndex (ganze Zahl)	**$TI-TimeLabel (Datum)**	**$TI-Year (ganze Zahl)**	**$TI-Months (ganze Zahl mit Beschriftungen)**
E	Jan 2000	2000	1 (Januar)

Tabelle 32. Beispiele für das Erstellen neuer Felder (Forts.)

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Datum)	$TI-Year (ganze Zahl)	$TI-Months (ganze Zahl mit Beschriftungen)
Z	Feb 2000	2000	2 (Februar)
3	Mär 2000	2000	3 (März)
4	Apr 2000	2000	4 (April)
5	Mai 2000	2000	5 (Mai)

Wochen (nicht periodisch)

Bei einer Zeitreihe im Wochenabstand können Sie den Tag der Woche angeben, an dem der Zyklus beginnt.

Beachten Sie, dass Wochen nur nicht periodisch sein können, da verschiedene Monate, Quartale und Jahre nicht unbedingt jeweils dieselbe Anzahl an Wochen aufweisen. Daten mit Zeitmarke können bei nicht periodischen Modellen jedoch leicht auf Wochenebene aggregiert oder aufgefüllt werden.

Beispiel

Neue Felder werden wie in der folgenden Tabelle dargestellt erstellt.

Tabelle 33. Beispiele für das Erstellen neuer Felder

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Datum)	$TI-Week (ganze Zahl)
E	1999-12-27	E
Z	2000-01-03	Z
3	2000-01-10	3
4	2000-01-17	4
5	2000-01-24	5

Im Feld *$TI-TimeLabel* für eine Woche wird der erste Tag der betreffenden Woche angezeigt. In der vorherigen Tabelle begann der Benutzer mit der Beschriftung beim 1. Januar 2000. Die Woche beginnt jedoch am Montag und der 1. Januar 2000 ist ein Samstag. Somit beginnt die Woche, in der der 1. Januar liegt, am 27. Dezember 1999 und wird als Beschriftung des ersten Punkts verwendet.

Das Datumsformat bestimmt, welche Zeichenfolgen für das Feld *$TI-TimeLabel* erstellt werden.

Tage pro Woche

Bei täglichen Messungen, die in einen Wochenzyklus fallen, können Sie die Anzahl der Tage pro Woche angeben und auch den Tag festlegen, an dem die Woche beginnt. Sie können das Anfangsdatum festlegen, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option **Aus Daten erstellen** ein Zeitmarken- oder Datumsfeld festlegen, aus dem das Datum für die einzelnen Datensätze hervorgeht.

Beispiel

Neue Felder werden wie in der folgenden Tabelle dargestellt erstellt.

Tabelle 34. Beispiele für das Erstellen neuer Felder

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Datum)	$TI-Week (ganze Zahl)	$TI-Day (ganze Zahl mit Beschriftungen)
E	Jan 5 2005	E	3 (Mittwoch)

Tabelle 34. Beispiele für das Erstellen neuer Felder (Forts.)

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Datum)	$TI-Week (ganze Zahl)	$TI-Day (ganze Zahl mit Beschriftungen)
Z	Jan 6 2005	E	4 (Donnerstag)
3	Jan 7 2005	E	5 (Freitag)
4	Jan 10 2005	Z	1 (Montag)
5	Jan 11 2005	Z	2 (Dienstag)

Hinweis: Die Woche beginnt stets mit Tag 1 in der ersten Zeitperiode und stimmt nicht mit der Kalenderzählung überein. Auf Woche 52 folgt daher Woche 53, dann Woche 54 usw. Die Woche entspricht nicht der Woche im Jahr, sondern bezeichnet lediglich die Anzahl der Wocheninkremente in der Zeitreihe.

Tage (nicht periodisch)

Wählen Sie die Option "Tage (nicht periodisch)" bei täglichen Messungen, die nicht in einen normalen Wochenzyklus fallen. Sie können das Anfangsdatum festlegen, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option **Aus Daten erstellen** ein Zeitmarken- oder Datumsfeld festlegen, aus dem das Datum für die einzelnen Datensätze hervorgeht.

Beispiel

Neue Felder werden wie in der folgenden Tabelle dargestellt erstellt.

Tabelle 35. Beispiele für das Erstellen neuer Felder

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Datum)
E	Jan 5 2005
Z	Jan 6 2005
3	Jan 7 2005
4	Jan 8 2005
5	Jan 9 2005

Stunden pro Tag

Bei stündlichen Messungen, die in einen normalen Tageszyklus fallen, können Sie die Anzahl der Tage pro Woche festlegen, außerdem die Anzahl der Stunden pro Tag (z. B. ein Acht-Stunden-Arbeitstag), den Tag, an dem die Woche beginnt, sowie die Stunde, an dem jeder Tag beginnt. Die Stunden können minutengenau im 24-Stunden-System angegeben werden (z. B. 14:05).

Sie können das Anfangsdatum und die Anfangszeit festlegen, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option **Aus Daten erstellen** ein Zeitmarkenfeld festlegen, aus dem das Datum und die Uhrzeit für die einzelnen Datensätze hervorgeht.

Beispiel

Neue Felder werden wie in der folgenden Tabelle dargestellt erstellt.

Tabelle 36. Beispiele für das Erstellen neuer Felder

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Zeitmarke)	$TI-Day (ganze Zahl mit Beschriftungen)	$TI-Hour (ganze Zahl mit Beschriftungen)
E	Jan 5 2005 8:00	3 (Mittwoch)	8 (8:00)
Z	Jan 5 2005 9:00	3 (Mittwoch)	9 (9:00)
3	Jan 5 2005 10:00	3 (Mittwoch)	10 (10:00)

Tabelle 36. Beispiele für das Erstellen neuer Felder (Forts.)

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Zeitmarke)	$TI-Day (ganze Zahl mit Beschriftungen)	$TI-Hour (ganze Zahl mit Beschriftungen)
4	Jan 5 2005 11:00	3 (Mittwoch)	11 (11:00)
5	Jan 5 2005 12:00	3 (Mittwoch)	12 (12:00)

Stunden (nicht periodisch)

Wählen Sie diese Option bei stündlichen Messungen, die nicht in einen normalen Tageszyklus fallen. Sie können die Anfangszeit festlegen, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option **Aus Daten erstellen** ein Zeitmarken- oder Zeitfeld festlegen, aus dem die Uhrzeit für die einzelnen Datensätze hervorgeht.

Die Stundenangaben beruhen auf dem 24-Stunden-System; auf die 24. Stunde folgt dabei die 25. Stunde, die Uhrzeit wird also nicht auf 1:00 zurückgesetzt.

Beispiel

Neue Felder werden wie in der folgenden Tabelle dargestellt erstellt.

Tabelle 37. Beispiele für das Erstellen neuer Felder

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Zeichenfolge)	$TI-Hour (ganze Zahl mit Beschriftungen)
E	8:00	8 (8:00)
Z	9:00	9 (9:00)
3	10:00	10 (10:00)
4	11:00	11 (11:00)
5	12:00	12 (12:00)

Minuten pro Tag

Bei minutengenauen Messungen, die in einen Tageszyklus fallen, können Sie beispielsweise die Anzahl der Tage in der Woche festlegen, außerdem die Anzahl der Stunden pro Tag sowie die Uhrzeit, zu der der Tag beginnt. Die Uhrzeiten können mithilfe von Doppelpunkten minuten- und sekundengenau im 24-Stunden-System angegeben werden (z. B. 14:05:17). Darüber hinaus können Sie den Zeitraum in Minuten pro Inkrement angeben (minütlich, alle zwei Minuten usw.); das Inkrement muss dabei ein Wert sein, durch den 60 ohne Rest dividiert werden kann.

Sie können das Anfangsdatum und die Anfangszeit festlegen, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option **Aus Daten erstellen** ein Zeitmarkenfeld festlegen, aus dem das Datum und die Uhrzeit für die einzelnen Datensätze hervorgeht.

Beispiel

Neue Felder werden wie in der folgenden Tabelle dargestellt erstellt.

Tabelle 38. Beispiele für das Erstellen neuer Felder

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Zeitmarke)	$TI-Minute
E	2005-01-05 08:00:00	0
Z	2005-01-05 08:01:00	E
3	2005-01-05 08:02:00	Z
4	2005-01-05 08:03:00	3

Tabelle 38. Beispiele für das Erstellen neuer Felder (Forts.)

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Zeitmarke)	$TI-Minute
5	2005-01-05 08:04:00	4

Minuten (nicht periodisch)

Wählen Sie diese Option bei minütlichen Messungen, die nicht in einen normalen Tageszyklus fallen. Sie können den Zeitraum in Minuten pro Inkrement angeben (minütlich, alle zwei Minuten usw.); das Inkrement muss dabei ein Wert sein, durch den 60 ohne Rest dividiert werden kann.

Sie können die Anfangszeit festlegen, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option **Aus Daten erstellen** ein Zeitmarken- oder Zeitfeld festlegen, aus dem die Uhrzeit für die einzelnen Datensätze hervorgeht.

Beispiel

Neue Felder werden wie in der folgenden Tabelle dargestellt erstellt.

Tabelle 39. Beispiele für das Erstellen neuer Felder

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Zeichenfolge)	$TI-Minute
E	8:00	0
Z	8:01	E
3	8:02	Z
4	8:03	3
5	8:04	4

- In der *TimeLabel*-Zeichenfolge sind die Stunden- und Minutenangaben durch einen Doppelpunkt getrennt. Auf die 24. Stunde folgt dabei die 25. Stunde; die Uhrzeit wird also nicht auf 1:00 zurückgesetzt.
- Die Minuten werden gemäß dem im Dialogfeld angegebenen Wert hochgezählt. Beim Inkrement 2 erhält das *TimeLabel*-Feld beispielsweise die Werte 8:00, 8:02 usw.; die Minutenangaben lauten entsprechend 0, 2 usw.

Sekunden pro Tag

Bei Sekundenintervallen, die in einen Tageszyklus fallen, können Sie beispielsweise die Anzahl der Tage in der Woche festlegen, außerdem die Anzahl der Stunden pro Tag sowie die Uhrzeit, zu der der Tag beginnt. Die Uhrzeiten können mithilfe von Doppelpunkten minuten- und sekundengenau im 24-Stunden-System angegeben werden (z. B. 14:05:17). Darüber hinaus können Sie den Zeitraum in Sekunden pro Inkrement angeben (sekündlich, alle zwei Sekunden usw.); das Inkrement muss dabei ein Wert sein, durch den 60 ohne Rest dividiert werden kann.

Sie können das Anfangsdatum und die Anfangszeit festlegen, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option **Aus Daten erstellen** ein Zeitmarkenfeld festlegen, aus dem das Datum und die Uhrzeit für die einzelnen Datensätze hervorgeht.

Beispiel

Neue Felder werden wie in der folgenden Tabelle dargestellt erstellt.

Tabelle 40. Beispiele für das Erstellen neuer Felder

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Zeitmarke)	$TI-Minute	$TI-Second
E	2005-01-05 08:00:00	0	0

Tabelle 40. Beispiele für das Erstellen neuer Felder (Forts.)

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Zeitmarke)	$TI-Minute	$TI-Second
Z	2005-01-05 08:00:01	0	E
3	2005-01-05 08:00:02	0	Z
4	2005-01-05 08:00:03	0	3
5	2005-01-05 08:00:04	0	4

Sekunden (nicht periodisch)

Wählen Sie diese Option bei sekündlichen Messungen, die nicht in einen normalen Tageszyklus fallen. Sie können den Zeitraum in Sekunden pro Inkrement angeben (sekündlich, alle zwei Sekunden usw.); das Inkrement muss dabei ein Wert sein, durch den 60 ohne Rest dividiert werden kann.

Legen Sie die Anfangszeit fest, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder wählen Sie mit der Option **Aus Daten erstellen** ein Zeitmarken- oder Zeitfeld, aus dem die Uhrzeit für die einzelnen Datensätze hervorgeht.

Beispiel

Neue Felder werden wie in der folgenden Tabelle dargestellt erstellt.

Tabelle 41. Beispiele für das Erstellen neuer Felder

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Zeichenfolge)	$TI-Minute	$TI-Second
E	8:00:00	0	0
Z	8:00:01	0	E
3	8:00:02	0	Z
4	8:00:03	0	3
5	8:00:04	0	4

- In der *TimeLabel*-Zeichenfolge sind die Stunden- und Minutenangaben sowie die Minuten- und Sekundenangaben jeweils durch einen Doppelpunkt getrennt. Auf die 24. Stunde folgt dabei die 25. Stunde; die Uhrzeit wird also nicht auf 1:00 zurückgesetzt.
- Die Sekunden werden gemäß dem Wert hochgezählt, der als Inkrement festgelegt ist. Beim Inkrement 2 erhält das *TimeLabel*-Feld beispielsweise die Werte 8:00:00, 8:00:02 usw.; die Sekundenangaben lauten entsprechend 0, 2 usw.

Verlaufsknoten

Verlaufsknoten werden am häufigsten für sequenzielle Daten, beispielsweise Zeitreihendaten, verwendet. Sie dienen zum Erstellen neuer Felder mit Daten aus Feldern in vorangegangenen Datensätzen. Bei Verwendung eines Verlaufsknotens sind meist Daten sinnvoll, die anhand eines bestimmten Felds vorsortiert sind. Dies lässt sich mit einem Sortierknoten erreichen.

Festlegen der Optionen für den Verlaufsknoten

Ausgewählte Felder. Mithilfe der Feldauswahlschaltfläche rechts neben dem Textfeld können Sie die Felder auswählen, für die der Verlauf erstellt werden soll. Alle ausgewählten Felder dienen zur Erstellung neuer Felder für alle Datensätze im Dataset.

Offset. Dient zur Angabe des jüngsten Datensatzes vor dem aktuellsten Datensatz, aus dem Verlaufsfeldwerte extrahiert werden sollen. Wenn für "Offset" beispielsweise der Wert "3" festgelegt ist, werden, wenn

die einzelnen Datensätze diesen Knoten durchlaufen, die Feldwerte für den dritten vorangegangenen Datensatz in den aktuellen Datensatz aufgenommen. Verwenden Sie die Einstellungen für die Spanne, um anzugeben, wie weit zurückliegende Datensätze für die Extraktion verwendet werden sollen. Mithilfe der Pfeile können Sie den Offset-Wert einstellen.

Spanne. Geben Sie an, aus wie vielen früheren Datensätzen Werte extrahiert werden sollen. Beispiel: Wenn für "Offset" der Wert "3" und für "Spanne" der Wert "5" angegeben wurde, werden jedem Datensatz, der den Knoten durchläuft, fünf Felder für jedes in der Liste "Ausgewählte Felder" angegebene Feld hinzugefügt. Wenn der Knoten also beispielsweise Datensatz 10 verarbeitet, werden für Datensatz 7 bis Datensatz 3 Felder hinzugefügt. Mithilfe der Pfeile können Sie den Wert für die Spanne einstellen.

Wenn Verlauf nicht verfügbar. Wählen Sie eine der folgenden Optionen für die Behandlung von Datensätzen aus, die keine Verlaufswerte aufweisen. Dies bezieht sich normalerweise auf die ersten Datensätze oben im Dataset, für die es keine vorangegangenen Datensätze gibt, die als Verlauf dienen könnten.

- **Datensätze verwerfen.** Wählen Sie diese Option aus, um Datensätze zu verwerfen, wenn für das ausgewählte Feld kein Verlaufswert verfügbar ist.
- **Verlauf undefiniert lassen.** Wählen Sie diese Option aus, um Datensätze beizubehalten, wenn für das ausgewählte Feld kein Verlaufswert verfügbar ist. Das Verlaufsfeld wird mit einem nicht definierten Wert ausgefüllt, der als `$null$` angezeigt wird.
- **Werte füllen mit.** Geben Sie einen Wert oder eine Zeichenfolge an, die für Datensätze verwendet werden soll, wenn kein Verlaufswert verfügbar ist. Der Standardersatzwert ist *undef*, die systemdefinierte Null. Nullwerte werden mit der Zeichenfolge `$null$` angezeigt.

Beachten Sie bei der Auswahl eines Ersatzwerts folgende Regeln, damit eine ordnungsgemäße Ausführung erfolgen kann:

- Die ausgewählten Felder sollten denselben Speichertyp aufweisen.
- Wenn alle ausgewählten Felder einen numerischen Speichertyp aufweisen, muss der Ersatzwert als ganze Zahl analysiert werden.
- Wenn alle ausgewählten Felder einen reellen Speichertyp aufweisen, muss der Ersatzwert als reelle Zahl analysiert werden.
- Wenn alle ausgewählten Felder einen symbolischen Speichertyp aufweisen, muss der Ersatzwert als Zeichenfolge analysiert werden.
- Wenn alle ausgewählten Felder den Speichertyp "Datum/Uhrzeit" aufweisen, muss der Ersatzwert als Datums-/Uhrzeit-Feld analysiert werden.

Wenn eine der oben angegebenen Bedingungen nicht erfüllt ist, wird bei der Ausführung des Verlaufsknotens eine Fehlernachricht ausgegeben.

Knoten "Felder ordnen"

Mit dem Knoten "Felder ordnen" können Sie die natürliche Reihenfolge definieren, die bei der Anzeige der nachgeordneten Felder verwendet wird. Diese Reihenfolge betrifft die Anzeige von Feldern an unterschiedlichen Stellen, beispielsweise in Tabellen, Listen und in der Feldauswahl. Dieser Vorgang dient beispielsweise dazu, um bei der Arbeit mit umfangreichen Datasets die relevanten Felder deutlicher hervorzuheben.

Festlegen der Optionen für "Felder ordnen"

Es gibt zwei Methoden für das Ordnen von Feldern: benutzerdefinierte Anordnung und automatische Sortierung.

Benutzerdefinierte Anordnung

Wählen Sie die Option **Benutzerdef. Anordnung**, um eine Tabelle mit Feldnamen und -typen zu definieren, in der Sie alle Felder anzeigen und mithilfe der Pfeilschaltflächen eine benutzerdefinierte Anordnung erstellen können.

So können Sie Felder neu anordnen:

1. Wählen Sie ein Feld in der Tabelle aus. Durch Klicken bei gedrückter Steuertaste können Sie mehrere Felder auswählen.
2. Mithilfe der einfachen Pfeilschaltflächen können Sie die Felder eine Zeile nach oben bzw. unten verschieben.
3. Mithilfe der Schaltflächen mit Pfeil und Balken können Sie die Felder ganz nach unten oder oben in der Liste verschieben.
4. Die Reihenfolge der hier nicht angegebenen Felder können Sie angeben, indem Sie die mit `[andere Felder]` bezeichnete Trennzeile nach oben bzw. unten verschieben.

Weitere Informationen zu [andere Felder]

Andere Felder. Die Trennzeile **[andere Felder]** dient dazu, die Tabelle in zwei Hälften aufzuteilen.

- Die oberhalb der Trennzeile angezeigten Felder werden (wie in der Tabelle zu sehen) oberhalb aller natürlichen Reihenfolgen angeordnet, die zur Anzeige der Felder unterhalb dieses Knotens verwendet werden.
- Die unterhalb der Trennzeile angezeigten Felder werden (wie in der Tabelle zu sehen) unterhalb aller natürlichen Reihenfolgen angeordnet, die zur Anzeige der Felder unterhalb dieses Knotens verwendet werden.

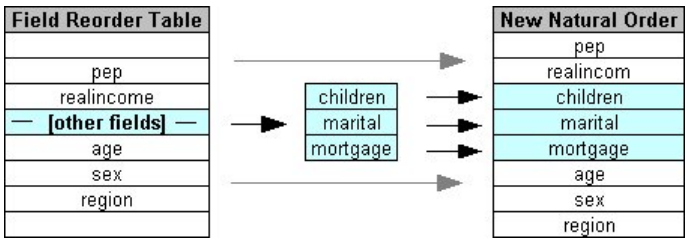


Abbildung 8. Diagramm zur Darstellung der Integration von "anderen Feldern" in die neue Feldanordnung

- Alle anderen Felder, die nicht in der Felderordnungstabelle angezeigt werden, werden zwischen den genannten "oberhalb" und "unterhalb" liegenden Feldern (durch die Lage der Trennzeile gekennzeichnet) angezeigt.

Weitere benutzerdefinierte Sortieroptionen:

- Sie können Felder in aufsteigender oder absteigender Reihenfolge sortieren, indem Sie auf die Pfeile oberhalb der einzelnen Spaltentitel klicken (**Typ**, **Name** und **Speichertyp**). Beim Sortieren nach Spalte werden die hier nicht angegebenen Felder (durch die Zeile `[andere Felder]` angezeigt) in ihrer natürlichen Reihenfolge sortiert.
- Klicken Sie auf **Nicht verwendete löschen**, um alle nicht verwendeten Felder aus dem Knoten "Felder ordnen" zu löschen. Nicht verwendete Felder werden in der Tabelle mit roter Schriftfarbe angezeigt. Diese zeigt an, dass das Feld in vorgeordneten Operationen gelöscht wurde.
- Sie können die Anordnung für alle neuen Felder angeben (angezeigt mit einem Blitzsymbol, das auf ein neues oder nicht spezifiziertes Feld hinweist). Wenn Sie auf **OK** oder **Anwenden** klicken, wird das Symbol entfernt.

Hinweis: Wenn weiter oben im Stream Felder hinzugefügt werden, nachdem eine benutzerdefinierte Anordnung durchgeführt wurde, werden die neuen Felder unten in der benutzerdefinierten Liste hinzugefügt.

Automatische Sortierung

Wählen Sie **Autom. Sort.** aus, um einen Parameter für die Sortierung anzugeben. Die Dialogfeldoptionen ändern sich dynamisch, um Optionen für die automatische Sortierung zu bieten.

Sortieren nach. Dient zur Auswahl einer der drei Methoden, die für die Sortierung der in den Knoten "Felder ordnen" eingelesenen Felder zur Verfügung stehen. Die Pfeilschaltflächen zeigen an, ob auf- oder absteigende Sortierreihenfolge verwendet wird. Wählen Sie eine Option aus, um eine Änderung vorzunehmen.

- Name
- Typ
- Speicher

Felder, die nach der Durchführung der automatischen Sortierung oberhalb des Knotens "Felder ordnen" hinzugefügt werden, werden automatisch in die richtige Position (je nach dem ausgewählten Sortiertyp) gebracht.

Kapitel 5. Diagrammknoten

Häufig verwendete Funktionen von Diagrammknoten

Die in IBM SPSS Modeler eingebrachten Daten werden in verschiedenen Phasen des Data-Mining-Prozesses mithilfe von Diagrammen und Grafiken untersucht. Verbinden Sie beispielsweise einen Plot- oder einen Verteilungsknoten mit einer Datenquelle und informieren Sie sich über die Datentypen und die Verteilungen. Anschließend können Sie Datensätze und Felder bearbeiten und die Daten so für Downstream-Modellierungsoperationen vorbereiten. Darüber hinaus werden Diagramme häufig zum Prüfen der Verteilung und der Beziehungen zwischen neu abgeleiteten Feldern eingesetzt.

Die Palette "Diagramme" enthält die folgenden Knoten:

Der Diagrammtafelknoten bietet viele verschiedene Diagrammtypen in einem einzigen Knoten. Bei Verwendung dieses Knotens können Sie die Datenfelder auswählen, die Sie untersuchen möchten, und anschließend eines der für die ausgewählten Daten verfügbaren Diagramme auswählen. Der Knoten filtert automatisch alle Diagrammtypen heraus, die nicht für die Feldauswahl geeignet sind.

Der Plotknoten zeigt die Beziehung zwischen numerischen Feldern an. Sie können einen Plot mithilfe von Punkten (Streudiagramm) oder mit Linien erstellen.

Der Verteilungsknoten zeigt das Auftreten symbolischer (kategorialer) Werte wie beispielsweise Hypothekenart oder Geschlecht. Verteilungsknoten eignen sich insbesondere zum Aufzeigen von Unausgewogenheiten in den Daten, die mithilfe eines Balancierungsknotens vor dem Erstellen eines Modells ausgeglichen werden können.

Der Histogrammknoten zeigt das Auftreten bestimmter Werte in numerischen Feldern. Damit werden häufig die Daten vor der weiteren Bearbeitung und der Modellerstellung untersucht. Ähnlich wie der Verteilungsknoten kann der Histogrammknoten oft Unausgewogenheiten in den Daten aufdecken.

Der Sammlungsknoten zeigt die Verteilung der Werte für ein numerisches Feld im Verhältnis zu den Werten eines anderen an. (Er erstellt histogrammähnliche Diagramme.) Er eignet sich besonders für die Darstellung einer Variablen oder eines Felds, dessen Werte sich mit der Zeit verändern. Mithilfe eines 3-D-Diagramms können Sie außerdem eine symbolische Achse anlegen, auf der die Verteilungen nach Kategorie aufgetragen sind.

Ein Multiplot erstellt ein Plot, bei dem mehrere *Y*-Felder über einem einzelnen *X*-Feld dargestellt werden. Die *Y*-Felder werden als farbige Linien geplottet, die jeweils einem Plotknoten mit dem Stil **Linie** und dem X-Modus **Sortieren** entsprechen. Multiplots sind hilfreich, wenn die Fluktuation mehrerer Variablen im Laufe der Zeit untersucht werden soll.

Der Netzdiagrammknoten zeigt die Stärke der Beziehung zwischen den Werten aus mindestens zwei symbolischen (kategorialen) Feldern. Im Diagramm wird die Verbindungsstärke durch unterschiedlich breite Linien angezeigt. Mit Netzdiagrammknoten können Sie beispielsweise die Beziehung zwischen dem Kauf einer Gruppe von Artikeln auf einer e-Commerce-Website untersuchen.

Der Zeitdiagrammknoten zeigt ein oder mehrere Sets mit Zeitreihendaten an. Normalerweise wird zuerst mithilfe eines Zeitintervallknotens ein *TimeLabel*-Feld erstellt, das dann zur Beschriftung der *X*-Achse verwendet wird.

Der Evaluierungsknoten erleichtert die Evaluierung und den Vergleich von Vorhersagemodellen. Das Evaluierungsdiagramm zeigt, wie gut Modelle bestimmte Ergebnisse vorhersagen. Die Datensätze werden auf der Grundlage des vorhergesagten Werts und des Konfidenzwerts für die Vorhersage sortiert. Die Datensätze werden in gleich große Gruppen (**Quantile**) aufgeteilt. Anschließend wird der Wert des Geschäftskriteriums für jedes Quantil geplottet, vom höchsten Wert bis zum niedrigsten Wert. Mehrere Modelle werden als separate Linien im Plot dargestellt.

Nachdem Sie einen Diagrammknoten zu einem Stream hinzugefügt haben, können Sie durch Doppelklicken auf den Knoten ein Dialogfeld zur Angabe von Optionen öffnen. Die meisten Diagramme enthalten eine Reihe spezieller Optionen, die auf einer oder mehreren Registerkarten gruppiert sind. Des Weiteren stehen einige Registerkartenoptionen zur Verfügung, die allen Diagrammen gemeinsam sind. In den nachstehenden Abschnitten finden Sie weitere Informationen zu diesen gemeinsamen Optionen.

Nachdem Sie die Optionen für einen Diagrammknoten konfiguriert haben, können Sie ihn im Dialogfeld oder als Teil eines Streams ausführen. Im generierten Diagrammfenster können Sie Ableitungsknoten (Set und Flag) und Auswahlknoten auf der Grundlage einer Datenauswahl bzw. eines Datenbereichs generieren, wodurch ein Subset der Daten erstellt wird. Diese leistungsstarke Funktion kann beispielsweise zur Ermittlung und zum Ausschluss von Ausreißern verwendet werden.

Formatierungen, Überlagerungen, Fenster und Animation

Überlagerungen und Formatierungen

Formatierung (und Überlagerungen) verleihen einer Visualisierung Dimensionalität. Die Wirkung einer Formatierung (Gruppierung, Clusterbildung oder Stapelung) hängt vom Visualisierungstyp, dem Feld-/Variablentyp sowie dem Grafikelementtyp und der Statistik ab. So kann beispielsweise ein kategoriales Feld für die Farbe verwendet werden, um Punkte in einem Streudiagramm zu gruppieren oder die Stapel in einem gestapelten Balkendiagramm zu erstellen. Ein fortlaufender Zahlenbereich für Farbe kann verwendet werden, um die Werte des Bereichs für jeden Punkt in einem Streudiagramm anzuzeigen.

Sie sollten die verschiedenen Formatierungsmöglichkeiten und Überlagerungen ausprobieren, um diejenige zu finden, die Ihren Bedürfnissen am besten entspricht. Die folgenden Beschreibungen helfen Ihnen bei der richtigen Auswahl.

Hinweis: Nicht alle Formatierungen und Überlagerungen stehen für alle Visualisierungstypen zur Verfügung.

- **Farbe.** Wenn die Farbe durch ein kategoriales Feld festgelegt wird, wird die Visualisierung für die einzelnen Kategorien aufgeteilt und Kategorie erhält eine andere Farbe. Wenn Farbe ein fortlaufender Zahlenbereich ist, variiert die Farbe basierend auf dem Wert des Bereichsfelds. Wenn das Grafikelement (z. B. ein Balken oder eine Box) für mehrere Datensätze/Fälle steht und ein Bereichsfeld für die Farbe verwendet wird, variiert die Farbe je nach dem *Mittelwert* des Bereichsfelds.
- **Form.** Form wird durch ein kategoriales Feld definiert, das die Visualisierung in Elemente mit unterschiedlichen Formen aufteilt, eine für jede Kategorie.
- **Transparenz.** Wenn Transparenz durch ein kategoriales Feld definiert ist, wird die Visualisierung für die einzelnen Kategorien aufgeteilt und jede Kategorie erhält eine Transparenzstufe. Wenn Transparenz ein fortlaufender Zahlenbereich ist, variiert die Transparenz basierend auf dem Wert des Bereichsfelds. Wenn das Grafikelement (zum Beispiel ein Balken oder ein Rechteck) mehr als einen Datensatz/Fall re-

präsentiert und ein Bereichsfeld für Transparenz verwendet wird, variiert die Farbe basierend auf dem *Mittelwert* des Bereichsfelds. Beim größten Wert sind die Grafikelemente völlig transparent. Beim kleinsten Wert sind sie völlig undurchsichtig.

- **Datenbeschriftung.** Datenbeschriftungen werden durch einen beliebigen Feldtyp definiert, dessen Werte zur Erstellung von Beschriftungen verwendet werden, die den Grafikelementen beigefügt werden.
- **Größe.** Wenn Größe durch ein kategoriales Feld definiert ist, teilt sie die Visualisierung basierend auf den individuellen Kategorien auf, eine Größe für jede Kategorie. Wenn Größe ein fortlaufender Zahlenbereich ist, variiert die Größe basierend auf dem Wert des Bereichsfelds. Wenn das Grafikelement (zum Beispiel ein Balken oder ein Rechteck) mehr als einen Datensatz/Fall repräsentiert und ein Bereichsfeld für Größe verwendet wird, variiert die Größe basierend auf dem *Mittelwert* des Bereichsfelds.

Einteilung in Felder und Animation

Einteilung in Felder. Durch die Einteilung in Felder wird eine Tabelle von Diagrammen erstellt. Für jede Kategorie in den Feldern der Einteilung in Felder wird ein Diagramm generiert, aber alle diese Felder werden gleichzeitig angezeigt. Die Einteilung in Felder eignet sich für eine Prüfung, ob die Visualisierung den Bedingungen der Felder entspricht. Sie können zum Beispiel ein Histogramm nach Geschlecht in Felder aufteilen, um zu ermitteln, ob die Häufigkeitsverteilungen bei Männern und Frauen gleich sind. So können Sie prüfen, ob das Gehalt Geschlechtsunterschieden unterworfen ist. Wählen Sie ein kategoriales Feld für die Einteilung in Felder aus.

Animation. Die Animation ähnelt der Einteilung in Felder dahin gehend, dass aus den Werten des Animationsfelds mehrere Diagramme erstellt werden. Diese Diagramme werden jedoch nicht gemeinsam angezeigt. Stattdessen können Sie mithilfe der Steuerelemente im Explorationsmodus die Ausgabe animieren und eine Folge einzelner Diagramme durchblättern. Außerdem ist für die Animation im Gegensatz zur Einteilung in Felder kein kategoriales Feld erforderlich. Sie können ein stetiges Feld angeben, dessen Werte automatisch in Bereiche aufgeteilt werden. Sie können die Größe des Bereichs mithilfe der Animationssteuerelemente im Explorationsmodus variieren. Nicht alle Visualisierungen bieten Animation.

Registerkarte "Ausgabe"

Bei allen Diagrammtypen können Sie die nachstehenden Optionen für den Dateinamen und die Anzeige der erzeugten Diagramme festlegen.

Hinweis: Für die Diagramme von Verteilungsknoten gelten zusätzliche Einstellungen.

Ausgabename. Bestimmt den Namen des Diagramms, das beim Ausführen des Knotens erstellt wird. Mit **Auto** wird ein Name auf der Grundlage des Knotens bestimmt, mit dem die Ausgabe erzeugt wird. Optional können Sie auch **Angepasst** auswählen und einen anderen Namen angeben.

Ausgabe auf Bildschirm. Hiermit lassen Sie das Diagramm in einem neuen Fenster erzeugen und anzeigen.

Ausgabe in Datei. Hiermit wird die Ausgabe als Datei gespeichert.

- **Ausgabediagramm.** Hiermit erstellen Sie Ausgaben in einem Diagrammformat. Nur bei Verteilungsknoten verfügbar.
- **Ausgabetabelle.** Hiermit erstellen Sie Ausgaben in einem Tabellenformat. Nur bei Verteilungsknoten verfügbar.
- **Dateiname.** Geben Sie einen Dateinamen für das erzeugte Diagramm bzw. die erzeugte Tabelle an. Mit der Auslassungsschaltfläche **(...)** legen Sie eine Datei und einen Pfad fest.
- **Dateityp.** Dient zur Auswahl des Dateityps in der Dropdown-Liste. Für alle Diagrammknoten mit Ausnahme des Verteilungsknotens mit der Option **Ausgabetabelle** stehen folgende Dateitypen für Diagramme zur Verfügung:

  - Bitmap (*.bmp*)

- PNG (*.png*)
- Ausgabeobjekt (*.cou*)
- JPEG (*.jpg*)
- HTML (*.html*)
- ViZml-Dokument (*.xml*) zur Verwendung in anderen IBM SPSS Statistics-Anwendungen.

Für die Option **Ausgabetabelle** im Verteilungsknoten stehen folgende Dateitypen zur Verfügung:

- Tabstoppgetrennte Daten (*.tab*)
- Kommagetrennte Daten (*.csv*)
- HTML (*.html*)
- Ausgabeobjekt (*.cou*)

Ausgabe paginieren. Beim Speichern der Ausgabe als HTML-Datei wird diese Option zur Verfügung gestellt, damit Sie die Größe der einzelnen HTML-Seiten festlegen können. (Gilt nur für den Verteilungsknoten.)

Zeilen pro Seite. Bei Auswahl von **Ausgabe paginieren** wird diese Option zur Verfügung gestellt, damit Sie die Länge der einzelnen HTML-Seiten festlegen können. Die Standardeinstellung sind 400 Zeilen. (Gilt nur für den Verteilungsknoten.)

Registerkarte "Anmerkungen"

Wird für alle Knoten verwendet. Die Registerkarte bietet Optionen zum Umbenennen von Knoten, zum Anzeigen einer benutzerdefinierten QuickInfo und zum Speichern einer längeren Anmerkung.

3-D-Diagramme

Bei Plots und Sammlungsdiagrammen in IBM SPSS Modeler können Daten auf einer dritten Achse dargestellt werden. Auf diese Weise erhalten Sie eine noch größere Flexibilität bei der Visualisierung der Daten zur Auswahl von Subsets oder zum Ableiten neuer Felder für die Modellierung.

Nachdem Sie ein 3-D-Diagramm erstellt haben, können Sie darauf klicken und mit der Maus ziehen, um es zu drehen und aus jedem beliebigen Winkel zu betrachten.

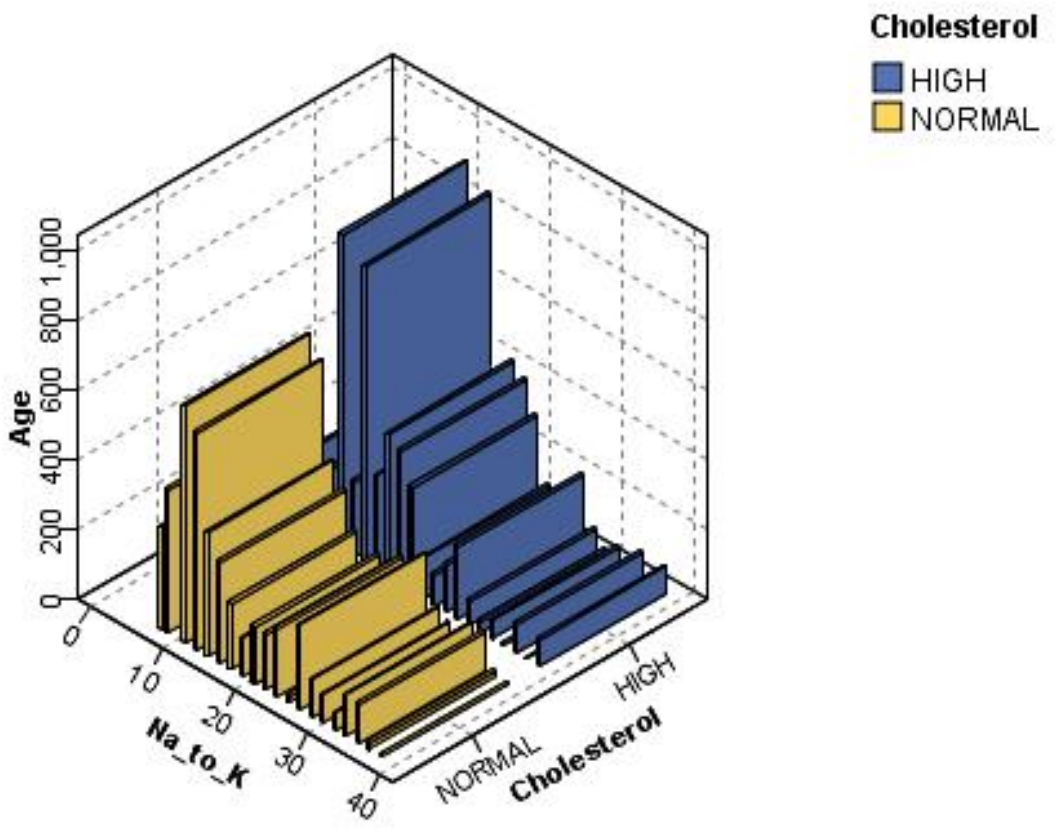


Abbildung 9. Sammlungsdiagramm mit X-, Y und Z-Achse

Für das Erstellen von 3-D-Diagrammen in IBM SPSS Modeler stehen zwei Verfahren zur Auswahl: Daten auf einer dritten Achse plotten (echte 3-D-Diagramme) oder Diagramme mit 3-D-Effekten anzeigen lassen. Beide Verfahren sind sowohl für Plots als auch für Sammlungen verfügbar.

So plotten Sie Daten auf einer dritten Achse:

1. Klicken Sie im Dialogfeld des Diagrammknotens auf die Registerkarte **Plot**.
2. Klicken Sie auf die Schaltfläche "3-D". Die Optionen für die *Z*-Achse werden aktiviert.
3. Wählen Sie mit der Feldauswahlschaltfläche ein Feld für die *Z*-Achse aus. In bestimmten Fällen sind hier nur symbolische Felder zulässig. In der Feldauswahl werden die entsprechenden Felder aufgeführt.

So statten Sie ein Diagramm mit 3-D-Effekten aus:

1. Erstellen Sie ein Diagramm und klicken Sie im Ausgabefenster auf die Registerkarte **Diagramm**.
2. Klicken Sie auf die Schaltfläche "3-D". Die Ansicht wechselt zu einem dreidimensionalen Diagramm.

Diagrammtafelknoten

Der Diagrammtafelknoten ermöglicht die Auswahl aus vielen verschiedenen Diagrammausgaben (Balkendiagramme, Kreisdiagramme, Histogramme, Streudiagramme, Heat-Maps usw.) in einem einzigen Knoten. Auf der ersten Registerkarte wählen Sie zunächst die zu untersuchenden Datenfelder aus. Anschließend stellt Ihnen der Knoten eine Reihe von Diagrammtypen zur Auswahl, die für Ihre Daten geeignet sind. Der Knoten filtert automatisch alle Diagrammtypen heraus, die nicht für die Feldauswahl geeignet sind. Detailliertere bzw. erweiterte Diagrammoptionen können Sie auf der Registerkarte "Detailliert" definieren.

Hinweis: Sie müssen den Diagrammtafelknoten mit einem Stream mit Daten verbinden, um den Knoten bearbeiten oder Diagrammtypen auswählen zu können.

Es gibt zwei Schaltflächen, über die Sie steuern können, welche Visualisierungsvorlagen (und Style-Sheets und Zuordnungen) verfügbar sind:

Verwalten. Verwalten Sie Visualisierungsvorlagen, Style-Sheets und Karten auf Ihrem Computer. Sie können Visualisierungsvorlagen, Style-Sheets und Karten auf Ihrem lokalen System importieren, exportieren, umbenennen und löschen. Weitere Informationen finden Sie im Thema „Verwalten von Vorlagen, Style-Sheets und Kartendateien" auf Seite 206.

Speicherort. Ändern Sie den Speicherort von Visualisierungsvorlagen und, Style-Sheets und Karten. Der aktuelle Speicherort wird rechts neben der Schaltfläche angezeigt. Weitere Informationen finden Sie im Thema „Festlegen des Speicherorts für Vorlagen, Style-Sheets und Karten" auf Seite 206.

Diagrammtafel - Registerkarte "Basis"

Wenn Sie sich nicht sicher sind, welcher Visualisierungstyp Ihre Daten am besten darstellt, sollten Sie die Registerkarte "Basis" verwenden. Wenn Sie Ihre Daten auswählen, wird ein Subset von Visualisierungstypen angezeigt, die für die Daten geeignet sind. Beispiele finden Sie unter „Diagrammtafel - Beispiele" auf Seite 195.

1. Wählen Sie mindestens ein Feld/eine Variable in der Liste aus. Wenn Sie bei gedrückter Steuertaste klicken, können Sie mehrere Felder auswählen.

   Beachten Sie, dass das Messniveau des Felds den Typ der verfügbaren Visualisierungen bestimmt. Sie können das Messniveau ändern, indem Sie in der Liste mit der rechten Maustaste auf das Feld klicken und eine Option auswählen. Weitere Informationen zu den verfügbaren Messniveautypen finden Sie unter „Feldtypen (Variablentypen)" auf Seite 182.
2. Wählen Sie einen Visualisierungstyp aus. Beschreibungen der verfügbaren Typen finden Sie unter „Verfügbare integrierte Visualisierungstypen für Diagrammtafeln" auf Seite 186.
3. Für bestimmte Visualisierungen können Sie eine Auswertungsstatistik auswählen. Je nachdem, ob die Statistik häufigkeitsbasiert oder aus einem stetigen Feld berechnet ist, stehen andere Statistiksubsets zur Verfügung. Welche Statistiken verfügbar sind, hängt auch von der Vorlage selbst ab. Eine vollständige Liste der Statistiken, die verfügbar sein könnten, folgt im nächsten Schritt.
4. Wenn Sie weitere Optionen wie optionale Formatierungen und Felder definieren möchten, klicken Sie auf **Detailliert**. Weitere Informationen finden Sie im Thema „Diagrammtafel - Registerkarte "Detailliert"" auf Seite 184.

Aus einem stetigen Feld berechnete Auswertungsstatistiken

- *Mittelwert*. Ein Lagemaß (zentrale Tendenz). Die Summe der Ränge, geteilt durch die Zahl der Fälle.
- *Median*. Wert, über und unter dem jeweils die Hälfte der Fälle liegt; 50. Perzentil. Bei einer geraden Anzahl von Fällen ist der Median der Mittelwert der beiden mittleren Fälle, wenn diese auf- oder absteigend sortiert sind. Der Median ist ein Lagemaß, das gegenüber Ausreißern unempfindlich ist (im Gegensatz zum Mittelwert, der durch wenige extrem niedrige oder hohe Werte beeinflusst werden kann).

- *Modalwert*. Der am häufigsten auftretende Wert. Wenn mehrere Werte gleichermaßen die größte Häufigkeit aufweisen, ist jeder von ihnen ein Modalwert.
- *Minimum*. Der kleinste Wert einer numerischen Variablen.
- *Maximum*. Der größte Wert einer numerischen Variablen.
- *Bereich*. Differenz zwischen Mindest- und Höchstwert.
- *Mittelbereich*. Der Mittelpunkt des Bereichs, also der Wert, dessen Differenz vom Mindestwert gleich seiner Differenz vom Höchstwert ist.
- *Summe*. Die Summe der Werte über alle Fälle mit nicht fehlenden Werten.
- *Kumulative Summe*. Die kumulative Summe der Werte. In jedem Grafikelement wird die Summe für eine Untergruppe plus der Gesamtsumme aller früheren Gruppen angezeigt.
- *Prozent Summe*. Der Prozentsatz innerhalb der einzelnen Untergruppen, beruhend auf einem summierten Feld im Vergleich zur Summe über alle Gruppen hinweg.
- *Kumulativer Prozentwert Summe*. Der kumulativer Prozentsatz innerhalb jeder Untergruppe basierend auf einem summierten Feld im Vergleich zur Summe über alle Gruppen hinweg. In jedem Grafikelement wird die der Prozentsatz für eine Untergruppe plus dem Gesamtprozentsatz aller früheren Gruppen angezeigt.
- *Varianz*. Ein Maß der Streuung um den Mittelwert, gleich der Summe der quadrierten Abweichungen vom Mittelwert geteilt durch eins weniger als die Anzahl der Fälle. Die Maßeinheit der Varianz ist das Quadrat der Maßeinheiten der Variablen.
- *Standardabweichung*. Ein Maß für die Streuung um den Mittelwert. In einer Normalverteilung liegen 68 % der Fälle innerhalb von einer Standardabweichung des Mittelwerts und 95 % der Fälle innerhalb von zwei Standardabweichungen. Wenn beispielsweise für das Alter der Mittelwert 45 und die Standardabweichung 10 beträgt, liegen bei einer Normalverteilung 95 % der Fälle im Bereich zwischen 25 und 65.
- *Standardfehler*. Ein Maß für die Abweichung des Werts einer Teststatistik zwischen Stichproben. Dies ist die Standardabweichung der Stichprobenverteilung einer Statistik. So ist z. B. der Standardfehler des Mittelwerts die Standardabweichung des Stichprobenmittelwerts.
- *Kurtosis*. Ein Maß dafür, wie sehr die Beobachtungen um einen zentralen Punkt gruppiert sind. Bei einer Normalverteilung ist der Wert der Kurtosis gleich 0. Bei positiver Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung enger um das Zentrum der Verteilung gruppiert und haben dünnere Flanken bis hin zu den Extremwerten der Verteilung. Ab dort sind die Flanken der leptokurtischen Verteilung im Vergleich zu einer Normalverteilung dicker. Bei negativer Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung weniger eng gruppiert und haben dickere Flanken bis hin zu den Extremwerten der Verteilung. Ab dort sind die Flanken der platykurtischen Verteilung im Vergleich zu einer Normalverteilung dünner.
- *Schiefe*. Ein Maß der Asymmetrie der Verteilung. Die Normalverteilung ist symmetrisch, ihre Schiefe hat den Wert 0. Eine Verteilung mit einer deutlichen positiven Schiefe läuft nach rechts lang aus (lange rechte Flanke). Eine Verteilung mit einer deutlichen negativen Schiefe läuft nach links lang aus (lange linke Flanke). Als Faustregel kann man verwenden, dass ein Schiefewert, der mehr als doppelt so groß ist wie sein Standardfehler, als Abweichung von der Symmetrie gilt.

Die folgenden Bereichsstatistiken können zu mehreren Grafikelementen pro Untergruppe führen. Bei Verwendung der Grafikelemente für Intervalle, Flächen oder Ränder führt eine Bereichsstatistik zu einem Grafikelement, das den Bereich zeigt. Alle anderen Grafikelemente führen zu zwei getrennten Elementen, einem, das den Beginn des Bereichs zeigt, und einem mit dem Endbereich.

- **Bereich: Bereich**. Der Bereich der Werte zwischen dem Mindest- und dem Höchstwert.
- **Bereich: 95-%-Konfidenzintervall für den Mittelwert**. Ein Wertebereich, der mit einer Wahrscheinlichkeit von 95 % den Mittelwert der Grundgesamtheit enthält.
- **Bereich: 95-%-Konfidenzintervall für einzelne Fälle**. Ein Wertebereich, der mit einer Wahrscheinlichkeit von 95 % den Wert vorhergesagten Wert für den Einzelfall enthält.
- **Bereich: 1 Standardabweichung über/unter dem Mittelwert**. Ein Wertebereich zwischen 1 *Standardabweichung* oberhalb und unterhalb des*Mittelwerts*.

- **Bereich: 1 Standardfehler über/unter dem Mittelwert**. Ein Wertebereich zwischen 1 *Standardfehler* oberhalb und unterhalb des*Mittelwerts*.

Anzahlbasierte Auswertungsstatistiken

- **Anzahl**. Die Anzahl der Zeilen/Fälle.
- **Kumulative Anzahl**. Die kumulative Anzahl der Zeilen/Fälle. In jedem Grafikelement wird die Anzahl für eine Untergruppe plus der Gesamtanzahl aller früheren Gruppen angezeigt.
- **Prozent Anzahl**. Der Prozentsatz der Zeilen/Fälle in jeder Untergruppe im Vergleich zur Gesamtzahl der Zeilen/Fälle.
- **Kumulativer Prozentwert Anzahl**. Der kumulative Prozentsatz der Zeilen/Fälle in jeder Untergruppe im Vergleich zur Gesamtzahl der Zeilen/Fälle. In jedem Grafikelement wird die der Prozentsatz für eine Untergruppe plus dem Gesamtprozentsatz aller früheren Gruppen angezeigt.

Feldtypen (Variablentypen)

Neben den Feldern in den Feldlisten werden Symbole angezeigt, die den Feld- und Datentyp angeben. Symbole kennzeichnen auch Mehrfachantwortsets.

Tabelle 42. Messniveausymbole.

Messniveau	Numerisch	Zeichenfolge	Datum	Zeit
Stetig		entfällt		
Sortiertes Set				
Set				

Tabelle 43. Symbole für Mehrfachantwortsets.

Typ des Mehrfachantwortsets	Symbol
Mehrfachantwortset, mehrere Kategorien	
Mehrfachantwortset, mehrere Dichotomien	

Messniveau

Für die Visualisierungserstellung ist das Messniveau eines Felds wichtig. Im Folgenden finden Sie eine Beschreibung der Messniveaus. Sie können das Messniveau vorübergehend ändern, indem Sie mit der rechten Maustaste auf ein Feld in der Feldliste klicken und eine Option auswählen. In den meisten Fällen müssen Sie nur die beiden allgemeinsten Klassifizierungen der Felder, kategorial und kontinuierlich (stetig), berücksichtigen:

Kategorial. Daten mit einer begrenzten Anzahl von eindeutigen Werten bzw. Kategorien (beispielsweise Geschlecht oder Religion). Bei kategorialen Feldern kann es sich um Zeichenfolgefelder (alphanumerisch) oder um numerische Felder handeln, bei denen numerische Codes zum Darstellen der Kategorien verwendet werden (beispielsweise 0 = *Männlich* und 1 = *Weiblich*). Auch als qualitative Daten bezeichnet. Sets, sortierte Sets und Flags sind kategoriale Felder.

- *Set*. Ein Feld/eine Variable, deren Werte Kategorien darstellen, die sich nicht in eine natürliche Reihenfolge bringen lassen, z. B. die Firmenabteilung, in der eine Person arbeitet. Beispiele für nominale Variablen sind Region, Postleitzahl oder Religionszugehörigkeit. Diese Art von Variable wird auch als nominale Variable bezeichnet.
- *Geordnetes Set*. Ein Feld/eine Variable, dessen bzw. deren Werte für Kategorien stehen, die eine natürliche Reihenfolge aufweisen (z. B. Grad der Zufriedenheit mit Kategorien von sehr unzufrieden bis sehr zufrieden). Geordnete Sets treten beispielsweise bei Einstellungsscores (Zufriedenheit oder Vertrauen) und bei Präferenzbeurteilungen auf. Diese Art von Variable wird auch als ordinale Variable bezeichnet.
- *Flag*. Ein Feld/eine Variable mit zwei unterschiedlichen Werten wie Ja und Nein oder 1 und 2. Auch als dichotome oder binäre Variable bezeichnet.

Stetig. Daten, die auf einer Intervall- oder Verhältnisskala gemessen werden und bei denen die Datenwerte sowohl die Reihenfolge der Werte als auch die Distanz zwischen den Werten festlegen. So ist beispielsweise ein Gehalt von $ 72.195 höher als ein Gehalt von $ 52.398 und die Distanz zwischen den Werten beträgt $ 19.797. Auch als quantitative Daten, Skalendaten oder Daten vom Typ numerischer Bereich bezeichnet.

Mithilfe von kategorialen Feldern werden Kategorien in der Visualisierung definiert. Normalerweise werden damit separate Grafikelemente gezeichnet oder Grafikelemente gruppiert. Stetige Felder werden häufig innerhalb der Kategorien von kategorialen Feldern zusammengefasst. So wird beispielsweise in einer Standardvisualisierung, in der das Einkommen nach Geschlecht kategorisiert ist, das durchschnittliche Einkommen von Männern und das durchschnittliche Einkommen von Frauen aufgeführt. Die Rohwerte für stetige Felder können auch in einem Streudiagramm dargestellt werden. So zeigt beispielsweise ein Streudiagramm das aktuelle Gehalt und das Anfangsgehalt für jeden Fall an. Ein kategoriales Feld kann dazu verwendet werden, um die Fälle nach Geschlecht zu gruppieren.

Datentypen

Das Messniveau ist nicht die einzige Eigenschaft eines Felds, die dessen Typ bestimmt. Ein Feld wird zudem als spezieller Datentyp gespeichert. Mögliche Datentypen sind Zeichenfolgen (nicht numerische Daten wie Buchstaben), numerische Werte (reelle Zahlen) und Datumsangaben. Im Gegensatz zum Messniveau kann der Datentyp eines Felds nicht vorübergehend geändert werden. Sie müssen die Art und Weise, wie die Daten im ursprünglichen Dataset gespeichert werden, ändern.

Mehrfachantwortsets

In bestimmten Datendateien kann eine besondere Art von "Feld" verwendet werden, die als **Mehrfachantwortset** bezeichnet wird. Bei Mehrfachantwortsets handelt es sich nicht um "Felder" im üblichen Sinn. Mehrfachantwortsets verwenden mehrere Felder, um Antworten auf Fragen aufzuzeichnen, auf welche der Befragte mehr als eine Antwort geben kann. Sie werden wie kategoriale Felder behandelt und bieten weitestgehend dieselben Möglichkeiten wie kategoriale Felder.

Mehrfachantwortsets können Sets aus dichotomen Variablen oder Sets aus kategorialen Variablen sein.

Set von dichotomen Variablen. Ein Set von dichotomen Variablen besteht in der Regel aus mehreren dichotomen Feldern. Dies sind Felder mit nur zwei möglichen Werten der Art Ja/Nein, Anwesend/Abwesend, Markiert/Nicht markiert. Auch wenn die Felder evtl. nicht streng dichotom sind, werden alle Felder im Set gleich codiert.

Beispielsweise gibt eine Umfrage auf die Frage "Welche der folgenden Quellen nutzen Sie für Nachrichten?" fünf mögliche Antworten vor. Der Befragte kann mehrere Antworten angeben, indem er das Kästchen neben jeder Auswahl markiert. Die fünf Antworten entsprechen fünf Feldern in der Datendatei, wobei 0 für *Nein* (nicht angekreuzt) und 1 für *Ja* (angekreuzt) steht.

Sets aus kategorialen Variablen. Ein Set von kategorialen Variablen besteht aus mehreren Feldern, die alle auf dieselbe Weise codiert wurden, häufig mit zahlreichen möglichen Antwortkategorien. Beispielsweise fordert eine Umfrage auf: "Nennen Sie drei Nationalitäten, die Ihre ethnische Herkunft am besten beschreiben." Hier sind hunderte von Antworten möglich, doch zu Codierungszwecken ist die Liste auf die 40 häufigsten Nationalitäten beschränkt und alle anderen werden der Kategorie "Andere" zugeordnet. In der Datendatei werden die Auswahlmöglichkeiten zu drei Feldern, wobei jedes über 41 Kategorien verfügt (40 codierte Nationalitäten und eine Kategorie "Andere").

Diagrammtafel - Registerkarte "Detailliert"

Verwenden Sie die Registerkarte "Detailliert", wenn Sie wissen, welche Art von Visualisierung Sie erstellen möchten, oder wenn Sie optionale Formatierungen, Fenster und/oder eine Animation zu einer Visualisierung hinzufügen möchten. Beispiele finden Sie unter „Diagrammtafel - Beispiele" auf Seite 195.

1. Wenn Sie einen Visualisierungstyp auf der Registerkarte "Einfach" ausgewählt haben, wird dieser angezeigt. Wählen Sie andernfalls einen Typ aus der Dropdown-Liste aus. Weitere Informationen zu Visualisierungstypen finden Sie unter „Verfügbare integrierte Visualisierungstypen für Diagrammtafeln" auf Seite 186.
2. Direkt rechts neben dem Piktogramm der Visualisierung befinden sich Steuerelemente für die Angabe der Felder (Variablen), die für den Visualisierungstyp erforderlich sind. Sie müssen alle diese Felder festlegen.
3. Bei bestimmten Visualisierungen können Sie eine Übersichtsstatistik auswählen. In einigen Fällen (beispielsweise bei Balkendiagrammen) können Sie eine dieser Auswertungsoptionen für die Transparenzformatierung verwenden. Beschreibungen der statistischen Funktionen finden Sie unter „Diagrammtafel - Registerkarte "Basis"" auf Seite 180.
4. Sie können eine oder mehrere der optionalen Formatierungen auswählen. Damit können Sie die Dimensionalität erhöhen, da Sie weitere Felder in die Visualisierung aufnehmen können. So können Sie beispielsweise ein Feld verwenden, um die Größe der Punkte in einem Streudiagramm zu verändern. Weitere Informationen zu optionalen Formatierungen finden Sie unter „Formatierungen, Überlagerungen, Fenster und Animation" auf Seite 176. Bitte beachten Sie, dass die Transparenzformatierung nicht von Scripts unterstützt wird.
5. Wenn Sie eine Kartenvisualisierung erstellen, zeigt die Gruppe **Kartendateien** die Kartendatei(en) an, die verwendet wird/werden. Wenn eine Standardkartendatei festgelegt ist, wird diese Datei angezeigt. Wenn Sie die Kartendatei ändern wollen, klicken Sie auf **Kartendatei auswählen**, um das Dialogfeld "Karten auswählen" anzuzeigen. In diesem Dialogfeld können Sie auch die Standardkartendatei angeben. Weitere Informationen finden Sie im Thema „Auswählen von Kartendateien für Kartenvisualisierungen".
6. Sie können eine oder mehrere der Optionen für die Einteilung in Felder oder die Animation auswählen. Weitere Informationen über die Optionen für die Einteilung in Felder oder die Animation finden Sie unter „Formatierungen, Überlagerungen, Fenster und Animation" auf Seite 176.

Auswählen von Kartendateien für Kartenvisualisierungen

Wenn Sie eine Vorlage für die Kartenvisualisierung auswählen, benötigen Sie eine Kartendatei, die die geografischen Informationen zum Zeichnen der Karte definiert. Wenn eine Standardkartendatei festgelegt ist, wird diese für die Kartenvisualisierung verwendet. Klicken Sie zur Auswahl einer anderen Kartendatei auf der Registerkarte "Detailliert" auf **Kartendatei auswählen**. Dadurch wird das Dialogfeld "Karten auswählen" angezeigt.

Im Dialogfeld "Karten auswählen" können Sie eine primäre Kartendatei und eine Referenzkartendatei auswählen. Die Kartendateien definieren die geografischen Informationen zum Zeichnen der Karte. Ihre Anwendung wird mit einem Set von Standardkartendateien installiert. Wenn Sie andere ESRI-Shapefiles besitzen, die Sie verwenden möchten, müssen Sie die Shapefiles zunächst in SMZ-Dateien konvertieren. Weitere Informationen finden Sie im Thema „Umwandeln und Verteilen von Kartenshapefiles" auf Seite 207. Klicken Sie nach der Konvertierung der Karte im Dialogfeld "Vorlagenauswahl" auf **Verwalten...** , um die Karte in das Verwaltungssystem zu importieren, damit sie im Dialogfeld "Karten auswählen" zur Verfügung steht.

Im Folgenden finden Sie einige Punkte, die bei der Angabe von Kartendateien berücksichtigt werden sollten:

- Für alle Kartenvorlagen ist mindestens eine Kartendatei erforderlich.
- Die Kartendatei verknüpft normalerweise ein Kartenschlüsselattribut mit dem Datenschlüssel.
- Wenn für die Vorlage kein Kartenschlüssel erforderlich ist, der mit einem Datenschlüssel verknüpft ist, so sind eine Referenzkartendatei und Felder erforderlich, die Koordinaten (z. B. Länge- und Breitengrad) zum Zeichnen von Elementen auf der Referenzkarte angeben.
- Überlagerungskartenvorlagen benötigen zwei Karten: eine primäre Kartendatei und eine Referenzkartendatei. Die Referenzkarte wird zuerst gezeichnet, liegt also hinter der primären Kartendatei.

Weitere Informationen zur Kartenterminologie, wie Attribute und Strukturen, finden Sie unter „Wichtige Konzepte im Zusammenhang mit Karten" auf Seite 208.

Kartendatei. Sie können jede beliebige Kartendatei auswählen, die sich im Verwaltungssystem befindet. Dazu gehören vorinstallierte Kartendateien sowie von Ihnen selbst importierte Kartendateien. Weitere Informationen zum Verwalten von Kartendateien finden Sie unter „Verwalten von Vorlagen, Style-Sheets und Kartendateien" auf Seite 206.

Kartenschlüssel. Geben Sie das Attribut an, das Sie als Schlüssel verwenden möchten, der die Kartendatei mit dem Datenschlüssel verknüpft.

Kartendatei und Einstellungen als Standardwert speichern. Aktivieren Sie dieses Kontrollkästchen, wenn Sie die ausgewählte Kartendatei standardmäßig verwenden möchten. Wenn Sie eine Standardkartendatei festgelegt haben, brauchen Sie nicht bei jeder Erstellung einer Kartenvisualisierung eine Kartendatei anzugeben.

Datenschlüssel. Mit diesem Steuerelement wird derselbe Wert aufgelistet, der auch auf der Registerkarte "Detailliert" der Vorlagenauswahl angezeigt wird. Er wird hier zur Arbeitserleichterung angegeben, für den Fall, dass Sie aufgrund der von Ihnen ausgewählten Kartendatei den Schlüssel ändern müssen.

Alle Kartenmerkmale in der Visualisierung auswählen. Wenn diese Option aktiviert ist, werden alle Strukturen (Merkmale) in der Karte in der Visualisierung gerendert, selbst wenn kein zugehöriger Datenschlüsselwert vorhanden ist. Wenn Sie nur die Merkmale anzeigen möchten, für die Daten vorhanden sind, inaktivieren Sie diese Option. Durch Kartenschlüssel aus der Liste **Nicht übereinstimmende Kartenschlüssel** angegebene Merkmale werden nicht in der Visualisierung gerendert.

Karten- und Datenwerte vergleichen Kartenschlüssel und Datenschlüssel werden jeweils miteinander verknüpft, um die Kartenvisualisierung zu erstellen. Der Kartenschlüssel und der Datenschlüssel sollten dieselbe Domäne verwenden (z. B. Länder und Regionen). Klicken Sie auf **Vergleichen**, um zu testen, ob die Werte von Datenschlüssel und Kartenschlüssel übereinstimmen. Das angezeigte Symbol informiert Sie über den Status des Vergleichs. Diese Symbole werden nachfolgend beschrieben. Wenn es nach dem Vergleich Datenschlüsselwerte ohne entsprechende Kartenschlüsselwerte gibt, werden die Datenschlüsselwerte in der Liste **Nicht übereinstimmende Datenschlüssel** angezeigt. In der Liste **Nicht übereinstimmende Kartenschlüssel** können Sie außerdem sehen, zu welchen Kartenschlüsselwerten es keine übereinstimmenden Datenschlüsselwerte gibt. Wenn **Alle Kartenmerkmale in der Visualisierung anzeigen** nicht aktiviert ist, werden durch diese Kartenschlüsselwerte angegebene Merkmale nicht wiedergegeben.

Tabelle 44. Vergleichssymbole.

Symbol	Beschreibung
	Es wurde kein Vergleich durchgeführt. Dies ist der Standardzustand vor dem Klicken auf **Vergleichen**. Sie sollten mit Bedacht vorgehen, da Sie nicht wissen, ob die Werte von Datenschlüssel und Kartenschlüssel übereinstimmen.

Tabelle 44. Vergleichssymbole (Forts.).

Symbol	Beschreibung
	Es wurde ein Vergleich durchgeführt und die Datenschlüssel- und Kartenschlüsselwerte stimmen vollständig überein. Für jeden Wert des Datenschlüssels wird vom Kartenschlüssel eine zugehörige Struktur angegeben.
	Es wurde ein Vergleich durchgeführt und einige Datenschlüssel- und Kartenschlüsselwerte stimmen nicht überein. Für einige Datenschlüsselwerte wird vom Kartenschlüssel keine zugehörige Struktur angegeben. Sie sollten mit Bedacht vorgehen. Wenn Sie fortfahren, sind in der Kartenvisualisierung nicht alle Datenwerte enthalten.
	Es wurde ein Vergleich durchgeführt und Datenschlüssel- und Kartenschlüsselwerte stimmen nicht überein. Sie sollten einen anderen Datenschlüssel oder einen anderen Kartenschlüssel auswählen, da keine Karte gerendert wird, wenn Sie fortfahren.

Verfügbare integrierte Visualisierungstypen für Diagrammtafeln

Sie können mehrere unterschiedliche Visualisierungstypen erstellen. Alle folgenden integrierten Typen sind auf den Registerkarten "Basis" und "Detailliert" verfügbar. Einige der Beschreibungen für die Vorlagen (insbesondere die Kartenvorlagen) geben die Felder (Variablen) an, die mithilfe von **Sondertext** auf der Registerkarte "Detailliert" festgelegt wurden.

Tabelle 45. Verfügbare Grafiktypen.

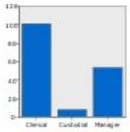
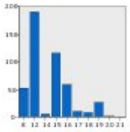
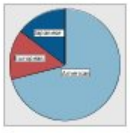
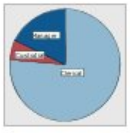
Diagrammsymbol	Beschreibung	Diagrammsymbol	Beschreibung
	**Balken** Berechnet eine Auswertungsstatistik für ein kontinuierliches numerisches Feld und zeigt die Ergebnisse für jede Kategorie eines kategorialen Felds in Form von Balken an. *Erfordert*: Ein kategoriales Feld und ein kontinuierliches Feld.		**Balken der Häufigkeiten** Zeigt den Anteil der Zeilen/Fälle in jeder Kategorie eines kategorialen Felds als Balken an. Sie können dieses Diagramm auch über den Diagrammknoten "Verteilung" erstellen. Dieser Knoten bietet einige zusätzliche Optionen. Weitere Informationen finden Sie im Thema „Verteilungsknoten” auf Seite 222. *Erfordert*: Ein einzelnes kategoriales Feld.
	**Kreis** Berechnet die Summe eines kontinuierlichen numerischen Felds und zeigt den Anteil dieser Summe in jeder Kategorie eines kategorialen Felds in Form eines Kreisausschnitts an. *Erfordert*: Ein kategoriales Feld und ein kontinuierliches Feld.		**Kreisdiagramm der Häufigkeiten** Zeigt den Anteil der Zeilen/Fälle in jeder Kategorie eines kategorialen Felds als Kreisausschnitte an. *Erfordert*: Ein einzelnes kategoriales Feld.

Tabelle 45. Verfügbare Grafiktypen (Forts.).

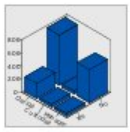
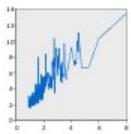
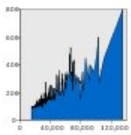
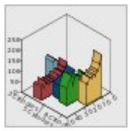
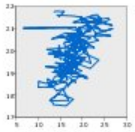
Diagrammsymbol	Beschreibung	Diagrammsymbol	Beschreibung
	**3-D-Balken** Berechnet eine Auswertungsstatistik für ein kontinuierliches numerisches Feld und zeigt die Ergebnisse für den Schnittpunkt von Kategorien zwischen zwei kategorialen Feldern an. *Erfordert*: Zwei kategoriale Felder und ein kontinuierliches Feld.		**3-D-Kreisdiagramm** Dasselbe wie ein Kreisdiagramm, allerdings mit zusätzlichem 3-D-Effekt. *Erfordert*: Ein kategoriales Feld und ein kontinuierliches Feld.
	**Linie** Berechnet eine Auswertungsstatistik für ein Feld für jeden Wert eines anderen Felds und verbindet die Werte durch eine Linie. Sie können dieses Diagramm auch über den Diagrammknoten "Diagramme" erstellen. Dieser Knoten bietet einige zusätzliche Optionen. Weitere Informationen finden Sie im Thema „Plotknoten” auf Seite 215. *Erfordert*: Zwei Felder beliebigen Typs.		**Fläche** Berechnet eine Auswertungsstatistik für ein Feld für jeden Wert eines anderen Felds und verbindet die Werte durch eine Fläche. Der Unterschied zwischen einem Linien- und einem Flächendiagramm ist minimal, da die Fläche durch eine Linie dargestellt wird, unter der der Bereich farbig markiert ist. Wenn Sie jedoch eine Farbformatierung verwenden, wird die Linie einfach getrennt und der wird Bereich gestapelt. *Erfordert*: Zwei Felder beliebigen Typs.
	**3-D-Fläche** Zeigt die Werte eines Felds im Verhältnis zu den Werten eines anderen Felds an, indem die Werte durch ein kategoriales Feld getrennt werden. Für jede Kategorie wird ein Flächenelement erstellt. *Erfordert*: Ein kategoriales Feld und zwei Felder beliebigen Typs.		**Pfad** Zeigt die Werte eines Felds im Verhältnis zu den Werten eines anderen Felds an, indem die Werte in der Reihenfolge, in der sie im ursprünglichen Dataset auftreten, mit einer Linie verbunden werden. Die Einhaltung der Reihenfolge ist der wesentliche Unterschied zwischen einem Pfad- und einem Liniendiagramm. *Erfordert*: Zwei Felder beliebigen Typs.

Tabelle 45. Verfügbare Grafiktypen (Forts.).

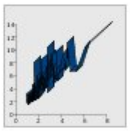
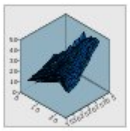
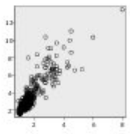
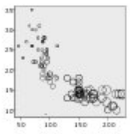
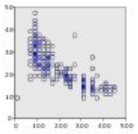
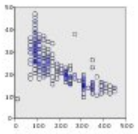
Diagrammsymbol	Beschreibung	Diagrammsymbol	Beschreibung
	**Band** Berechnet eine Auswertungsstatistik für ein Feld für jeden Wert eines anderen Felds und verbindet die Werte durch ein Band. Ein Band ist im Wesentlichen eine Linie mit 3-D-Effekten. Es handelt sich dabei nicht um ein echtes 3-D-Diagramm. *Erfordert*: Zwei Felder beliebigen Typs.		**Oberfläche** Zeigt die Werte von drei Feldern im Verhältnis zueinander an, indem die Werte mit einer Oberfläche verbunden werden. *Erfordert*: Drei Felder beliebigen Typs.
	**Streudiagramm** Zeigt die Werte eines Felds im Verhältnis zu den Werten eines anderen Felds an. Dieses Diagramm kann den Zusammenhang zwischen den Feldern (falls vorhanden) verdeutlichen. Sie können Streudiagramme auch über den Diagrammknoten "Diagramme" herstellen. Dieser Knoten bietet einige zusätzliche Optionen. Weitere Informationen finden Sie im Thema „Plotknoten” auf Seite 215. *Erfordert*: Zwei Felder beliebigen Typs.		**Blasendiagramm** Wie das allgemeine Streudiagramm zeigt das Blasendiagramm die Werte eines Felds im Verhältnis zu den Werten eines anderen Felds an. Der Unterschied liegt darin, dass die Werte eines dritten Felds verwendet werden, um die Größe der einzelnen Diagramme zu variieren. *Erfordert*: Drei Felder beliebigen Typs.
	**Klassiertes Streudiagramm** Wie das allgemeine Streudiagramm zeigt das Blasendiagramm die Werte eines Felds im Verhältnis zu den Werten eines anderen Felds an. Der Unterschied besteht darin, dass ähnliche Werte zu Gruppen zusammengefasst werden und dass die Farb- oder Größenformatierung verwendet wird, um die Anzahl der Fälle in den einzelnen Klassen anzugeben. *Erfordert*: Zwei kontinuierliche Felder.		**In Hexadezimalklassen unterteiltes Streudiagramm** Siehe die Beschreibung für das klassierte Streudiagramm. Der Unterschied liegt in der Form der zugrunde liegenden Klassen, die die Form von Sechsecken und nicht von Kreisen aufweisen. Das resultierende, in Hexadezimalklassen unterteilte Streudiagramm ist dem klassierten Streudiagramm ähnlich. Die Anzahl der Werte in jeder Klasse unterscheidet sich jedoch in den einzelnen Diagrammen aufgrund der Form der zugrunde liegenden Klassen. *Erfordert*: Zwei kontinuierliche Felder.

Tabelle 45. Verfügbare Grafiktypen (Forts.).

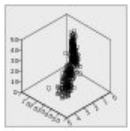
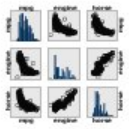
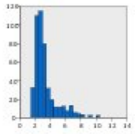
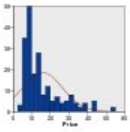
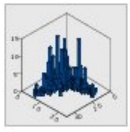
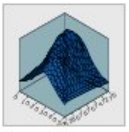
Diagrammsymbol	Beschreibung	Diagrammsymbol	Beschreibung
	**3-D-Streudiagramm** Zeigt die Werte von drei Feldern im Verhältnis zueinander an. Dieses Diagramm kann den Zusammenhang zwischen den Feldern (falls vorhanden) verdeutlichen. Sie können 3-D-Streudiagramme auch über den Diagrammknoten "Diagramme" erstellen. Dieser Knoten bietet einige zusätzliche Optionen. Weitere Informationen finden Sie im Thema „Plotknoten" auf Seite 215. *Erfordert*: Drei Felder beliebigen Typs.		**Streudiagrammmatrix (SPLOM)** Zeigt für jedes Feld die Werte eines Felds im Verhältnis zu den Werten eines anderen Felds an. Eine SPLOM entspricht einer Tabelle von Streudiagrammen. Die SPLOM enthält zudem ein Histogramm für jedes Feld. *Erfordert*: Mindestens zwei kontinuierliche Felder.
	**Histogramm** Zeigt die Häufigkeitsverteilung eines Felds an. Mit einem Histogramm können Sie den Verteilungstyp bestimmen und feststellen, ob die Verteilung Datenabweichungen enthält. Sie können dieses Diagramm auch über den Diagrammknoten "Histogramm" erstellen. Dieser Knoten bietet einige zusätzliche Optionen. Weitere Informationen finden Sie im Thema „Histogramm - Registerkarte "Plot"" auf Seite 226. *Erfordert*: Ein einzelnes Feld eines beliebigen Typs.		**Histogramm mit Normalverteilung** Zeigt die Häufigkeitsverteilung eines kontinuierlichen Felds mit einer überlagerten Normalverteilungskurve an. *Erfordert*: Ein einzelnes kontinuierliches Feld.
	**3-D-Histogramm** Zeigt die Häufigkeitsverteilung von zwei kontinuierlichen Feldern an. *Erfordert*: Zwei kontinuierliche Felder.		**3-D-Dichte** Zeigt die Häufigkeitsverteilung von zwei kontinuierlichen Feldern an. Dieses Diagramm ist dem 3-D-Histogramm ähnlich. Der einzige Unterschied besteht darin, dass anstatt von Balken die Oberfläche für die Anzeige der Verteilung verwendet wird. *Erfordert*: Zwei kontinuierliche Felder.

Tabelle 45. Verfügbare Grafiktypen (Forts.).

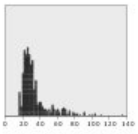
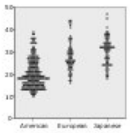
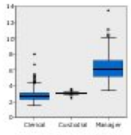
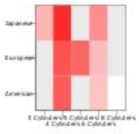
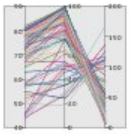
Diagrammsymbol	Beschreibung	Diagrammsymbol	Beschreibung
	**Punktdiagramm** Zeigt die einzelnen Fälle/Zeilen an und stapelt sie am jeweils richtigen Datenpunkt auf der *X*-Achse. Dieses Diagramm zeigt wie das Histogramm die Verteilung der Daten an. Der Unterschied besteht darin, dass jeder Fall bzw. jede Zeile und nicht die aggregierten Häufigkeiten für eine bestimmte Klasse (einen Wertebereich) angezeigt werden. *Erfordert*: Ein einzelnes Feld eines beliebigen Typs.		**2-D-Punktdiagramm** Zeigt die einzelnen Fälle/Zeilen an und stapelt sie für jede Kategorie eines kategorialen Felds am richtigen Datenpunkt auf der *Y*-Achse. *Erfordert*: Ein kategoriales Feld und ein kontinuierliches Feld.
	**Boxplot** Berechnet die fünf Statistiken (Minimum, erstes Quartil, Median, drittes Quartil und Maximum) für ein kontinuierliches Feld für jede Kategorie eines kategorialen Felds. Die Ergebnisse werden als Boxplot-/Schemaelemente dargestellt. Mit den Boxplots können Sie feststellen, wie die Verteilung kontinuierlicher Daten innerhalb der Kategorien variiert. *Erfordert*: Ein kategoriales Feld und ein kontinuierliches Feld.		**Verteilung** Berechnet den Mittelwert für ein kontinuierliches Feld für den Schnittpunkt von Kategorien zwischen zwei kategorialen Feldern. *Erfordert*: Zwei kategoriale Felder und ein kontinuierliches Feld.
	**Parallel** Erstellt parallele Achsen für jedes Feld und zieht für jede Zeile bzw. jeden Fall in den Daten eine Linie durch den Feldwert. *Erfordert*: Mindestens zwei kontinuierliche Felder.		**Choroplethenkarte der Häufigkeiten** Berechnet die Anzahl für jede Kategorie eines kategorialen Felds (**Datenschlüssel**) und zeichnet eine Karte, bei der Farbsättigung zur Darstellung der Häufigkeiten in den Kartenstrukturen verwendet wird, die den Kategorien entsprechen. *Erfordert*: Ein kategoriales Feld. Eine Karte, deren Schlüssel mit den **Datenschlüsselkategorien** übereinstimmt.

Tabelle 45. Verfügbare Grafiktypen (Forts.).

Diagrammsymbol	Beschreibung	Diagrammsymbol	Beschreibung
	**Choroplethenkarte der Mittelwerte/Mediane/Summen** Berechnet Mittelwert, Median oder Summe eines kategorialen Felds (**Farbe**) für jede Kategorie eines kategorialen Felds (**Datenschlüssel**) und zeichnet eine Karte, bei der Farbsättigung zur Darstellung der berechneten Statistiken in den Kartenstrukturen verwendet wird, die den Kategorien entsprechen. *Erfordert*: Ein kategoriales Feld und ein kontinuierliches Feld. Eine Karte, deren Schlüssel mit den **Datenschlüsselkategorien** übereinstimmt.		**Choroplethenkarte der Werte** Zeichnet eine Karte, bei der Farbe zur Darstellung der Werte eines kategorialen Felds (**Farbe**) für die Kartenstrukturen verwendet wird, die Werten entsprechen, die durch ein anderes kategoriales Feld (**Datenschlüssel**) definiert sind. Wenn mehrere kategoriale Werte des Felds "Farbe" für die einzelnen Strukturen vorliegen, wird der Modalwert verwendet. *Erfordert*: Zwei kategoriale Felder. Eine Karte, deren Schlüssel mit den **Datenschlüsselkategorien** übereinstimmt.
	**Koordinaten auf einer Choroplethenkarte der Häufigkeiten** Ähnlich wie "Choroplethenkarte der Häufigkeiten", mit dem Unterschied, dass zwei weitere kontinuierliche Felder (**Längengrad** und **Breitengrad**) vorhanden sind, die Koordinaten zum Zeichnen von Punkten auf der Choroplethenkarte angeben. *Erfordert*: Ein kategoriales Feld und zwei stetige Felder. Eine Karte, deren Schlüssel mit den **Datenschlüsselkategorien** übereinstimmt.		**Koordinaten auf einer Choroplethenkarte der Mittelwerte/Mediane/Summen** Ähnlich wie "Choroplethenkarte der Mittelwerte/Mediane/Summen", mit dem Unterschied, dass zwei weitere kontinuierliche Felder (**Längengrad** und **Breitengrad**) vorhanden sind, die Koordinaten zum Zeichnen von Punkten auf der Choroplethenkarte angeben. *Erfordert*: Ein kategorisches Feld und drei stetige Felder. Eine Karte, deren Schlüssel mit den **Datenschlüsselkategorien** übereinstimmt.

Tabelle 45. Verfügbare Grafiktypen (Forts.).

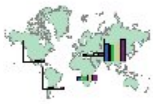
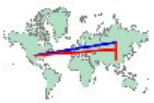
Diagrammsymbol	Beschreibung	Diagrammsymbol	Beschreibung
	**Koordinaten auf einer Choroplethenkarte der Werte** Ähnlich wie "Chloropethenkarte der Werte", mit dem Unterschied, dass zwei weitere kontinuierliche Felder (**Längengrad** und **Breitengrad**) vorhanden sind, die Koordinaten zum Zeichnen von Punkten auf der Choroplethenkarte angeben. *Erfordert*: Zwei kategoriale Feld und zwei stetige Felder. Eine Karte, deren Schlüssel mit den **Datenschlüsselkategorien** übereinstimmt.		**Balken mit Zählerwerten auf einer Karte** Berechnet den Anteil an Zeilen/Fällen in den einzelnen Kategorien eines kategorialen Felds (**Kategorien**) für jede Kartenstruktur (**Datenschlüssel**) und zeichnet eine Karte und die Balkendiagramme in der Mitte der einzelnen Kartenstrukturen. *Erfordert*: Zwei kategoriale Felder. Eine Karte, deren Schlüssel mit den **Datenschlüsselkategorien** übereinstimmt.
	**Balken auf einer Karte** Berechnet eine Auswertungsstatistik für ein kontinuierliches Feld (**Werte**) und zeigt die Ergebnisse für jede Kategorie eines kategorialen Felds (**Kategorien**) für jede Kartenstruktur (**Datenschlüssel**) als Balkendiagramme in der Mitte der einzelnen Kartenstrukturen an. *Erfordert*: Zwei kategoriale Felder und ein kontinuierliches Feld. Eine Karte, deren Schlüssel mit den **Datenschlüsselkategorien** übereinstimmt.		**Kreisdiagramm mit Zählerwerten auf einer Karte** Zeigt den Anteil an Zeilen/Fällen in den einzelnen Kategorien eines kategorialen Felds (**Kategorien**) für jede Kartenstruktur (**Datenschlüssel**) an und zeichnet eine Karte sowie die Anteile als Ausschnitte eines Kreisdiagramms in der Mitte der einzelnen Kartenstrukturen. *Erfordert*: Zwei kategoriale Felder. Eine Karte, deren Schlüssel mit den **Datenschlüsselkategorien** übereinstimmt.
	**Kreisdiagramm auf einer Karte** Berechnet die Summe eines kontinuierlichen Felds (**Werte**) in den einzelnen Kategorien eines kategorialen Felds (**Kategorien**) für jede Kartenstruktur (**Datenschlüssel**) und zeichnet eine Karte sowie die Summen als Ausschnitte eines Kreisdiagramms in der Mitte der einzelnen Kartenstrukturen. *Erfordert*: Zwei kategoriale Felder und ein kontinuierliches Feld. Eine Karte, deren Schlüssel mit den **Datenschlüsselkategorien** übereinstimmt.		**Liniendiagramm auf einer Karte** Berechnet eine Auswertungsstatistik für ein kontinuierliches Feld (**Werte**) für jeden Wert eines anderen Felds (**X**) für jede Kartenstruktur (**Datenschlüssel**) und zeichnet eine Karte sowie die Liniendiagramme, die die Werte verbinden, in der Mitte der einzelnen Kartenstrukturen. *Erfordert*: Ein kategoriales Feld und zwei Felder beliebigen Typs. Eine Karte, deren Schlüssel mit den **Datenschlüsselkategorien** übereinstimmt.

Tabelle 45. Verfügbare Grafiktypen (Forts.).

Diagrammsymbol	Beschreibung	Diagrammsymbol	Beschreibung
	**Koordinaten auf einer Bezugskarte** Zeichnet eine Karte und Punkte mithilfe kontinuierlicher Felder (**Längengrad** und **Breitengrad**), die Koordinaten für die Punkte angeben. *Erfordert*: Zwei Bereichsfelder. Eine Kartendatei.		**Pfeile auf einer Bezugskarte** Zeichnet eine Karte und Pfeile mithilfe kontinuierlicher Felder, die die Startpunkte (**Start Längengrad** und **Start Breitengrad**) und Endpunkte (**Ende Längengrad** und **Ende Breitengrad**) für die einzelnen Pfeile angeben. Jeder Datensatz/Fall in den Daten führt zu einem Pfeil auf der Karte. *Erfordern*: Vier stetige Felder. Eine Kartendatei.
	**Punktüberlagerungskarte** Zeichnet eine Referenzkarte und überlagert diese mit einer weiteren Punktkarte, wobei die Farbe der Punktstrukturen durch ein kategoriales Feld (**Farbe**) festgelegt ist. *Erfordert*: Zwei kategoriale Felder. Eine Punktkarte, deren Schlüssel mit den **Datenschlüsselkategorien** übereinstimmt. Eine Referenzkartendatei.		**Polygonüberlagerungskarte** Zeichnet eine Referenzkarte und überlagert diese mit einer weiteren Polygonkarte, wobei die Farbe der Polygonstrukturen durch ein kategoriales Feld (**Farbe**) festgelegt ist. *Erfordert*: Zwei kategoriale Felder. Eine Polygonkarte, deren Schlüssel mit den **Datenschlüsselkategorien** übereinstimmt. Eine Referenzkartendatei.
	**Linienüberlagerungskarte** Zeichnet eine Referenzkarte und überlagert diese mit einer weiteren Linienkarte, wobei die Farbe der Linienstrukturen durch ein kategoriales Feld (**Farbe**) festgelegt ist. *Erfordert*: Zwei kategoriale Felder. Eine Linienkarte, deren Schlüssel mit den **Datenschlüsselkategorien** übereinstimmt. Eine Referenzkartendatei.		

Erstellen von Kartenvisualisierungen

Für viele Visualisierungen müssen Sie nur zwei Dinge auswählen: die relevanten Felder (Variablen) und eine Vorlage zur Visualisierung dieser Felder. Es sind keine weiteren Entscheidungen oder Aktionen erforderlich. Bei Kartenvisualisierungen ist mindesten ein weiterer Schritt nötig, nämlich die Auswahl einer Kartendatei, die die geografischen Informationen für die Kartenvisualisierung definiert.

Die Grundschritte zur Erstellung einer einfachen Karte lauten wie folgt:

1. Wählen Sie die relevanten Felder auf der Registerkarte "Einfach" aus. Informationen dazu, welcher Typ und welche Anzahl von Feldern für verschiedene Kartenvisualisierungen erforderlich sind, finden Sie unter „Verfügbare integrierte Visualisierungstypen für Diagrammtafeln” auf Seite 186.
2. Wählen Sie eine Kartenvorlage aus.
3. Klicken Sie auf die Registerkarte "Detailliert".
4. Vergewissern Sie sich, dass **Datenschlüssel** und die anderen erforderlichen Dropdown-Listen auf die richtigen Felder gesetzt sind.
5. Klicken Sie in der Gruppe "Kartendateien" auf **Kartendatei auswählen**.
6. Wählen Sie im Dialogfeld "Karten auswählen" die Kartendatei und den Kartenschlüssel aus. Die Werte des Kartenschlüssels müssen mit den unter **Datenschlüssel** für das Feld angegebenen Werten übereinstimmen. Mit der Schaltfläche **Vergleichen** können diese Werte verglichen werden. Wenn Sie eine Überlagerungskartenvorlage auswählen, müssen Sie auch eine Referenzkarte auswählen. Die Referenzkarte ist nicht mit den Daten verknüpft. Sie wird als Hintergrund für die Hauptkarte verwendet. Weitere Informationen zum Dialogfeld "Karten auswählen" finden Sie unter „Auswählen von Kartendateien für Kartenvisualisierungen” auf Seite 184.
7. Klicken Sie auf **OK**, um das Dialogfeld "Karten auswählen" zu schließen.
8. Klicken Sie in der Auswahlfunktion für Diagrammtafelvorlagen auf **Ausführen**, um die Kartenvisualisierung zu erstellen.

Diagrammtafel - Beispiele

Dieser Abschnitt enthält einige unterschiedliche Beispiele zur Veranschaulichung der verfügbaren Optionen. Die Beispiele liefern zudem Informationen für die Interpretation der resultierenden Visualisierungen.

In diesen Beispielen wird der Stream *graphboard.str* verwendet, der auf die Datendateien *employee_data.sav*, *customer_subset.sav* und *worldsales.sav* verweist. Diese Dateien finden Sie im Ordner *Demos* jeder IBM SPSS Modeler Client-Installation. Sie können über die Programmgruppe "IBM SPSS Modeler" im Windows-Startmenü darauf zugreifen. Die Datei *graphboard.str* befindet sich im Ordner *streams*.

Es wird empfohlen, die Beispiele in der vorgegebenen Reihenfolge zu lesen. Die nachfolgenden Beispiele bauen auf den vorherigen Beispielen auf.

Beispiel: Balkendiagramm mit Auswertungsstatistik

Wir erstellen ein Balkendiagramm, das ein kontinuierliches numerisches Feld bzw. eine kontinuierliche numerische Variable für jede Kategorie eines Sets bzw. einer kategorialen Variable zusammenfasst. Insbesondere erstellen wir ein Balkendiagramm, das das mittlere Gehalt für Männer und Frauen darstellt.

In diesem und einigen folgenden Beispielen wird die Datei *Employee data* verwendet, bei der es sich um ein hypothetisches Dataset mit Informationen über die Mitarbeiter eines Unternehmens handelt.

1. Fügen Sie einen Statistics-Quellenknoten hinzu, der auf *employee_data.sav* verweist.
2. Fügen Sie einen Diagrammtafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.
3. Wählen Sie auf der Registerkarte "Basis" die Optionen *Gender* (Geschlecht) und *Current Salary* (Aktuelles Gehalt) aus. (Wenn Sie bei gedrückter Steuertaste klicken, können Sie mehrere Felder bzw. Variablen auswählen.)
4. Wählen Sie **Balken** aus.
5. Wählen Sie aus der Dropdown-Liste "Auswertung" den Eintrag **Mittelwert** aus.
6. Klicken Sie auf **Ausführen**.
7. Klicken Sie in der eingeblendeten Anzeige auf die Symbolleistenschaltfläche "Feld- und Wertbeschriftungen anzeigen" (die zweite in der Zweiergruppe in der Mitte der Symbolleiste).

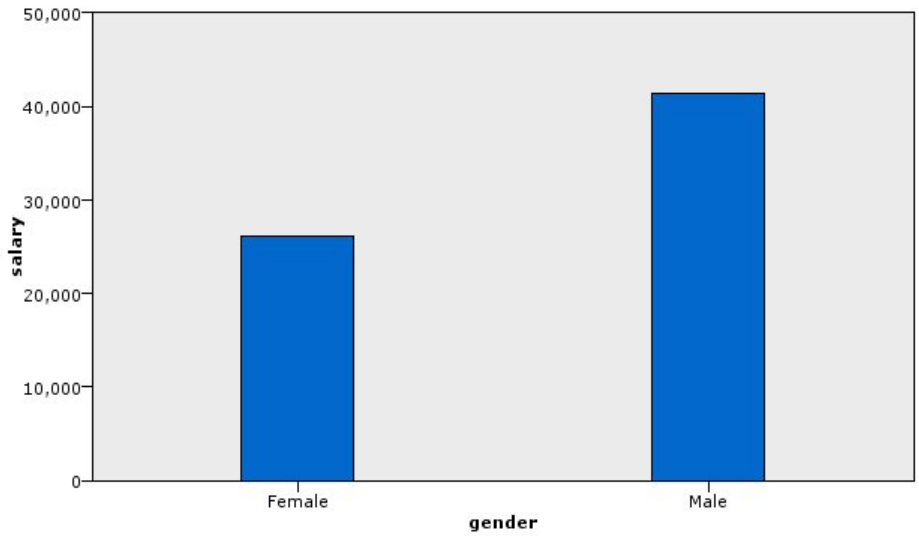


Abbildung 10. Balkendiagramm mit Auswertungsstatistik

Wir stellen Folgendes fest:

- Basierend auf der Höhe der Balken ist klar, dass das mittlere Gehalt von Männern über dem von Frauen liegt.

Beispiel: Gruppiertes Balkendiagramm mit Auswertungsstatistik

Wir erstellen nun ein gruppiertes Balkendiagramm, um festzustellen, ob der Unterschied im mittleren Gehalt zwischen Männern und Frauen von der Art der Tätigkeit abhängig ist. Vielleicht arbeiten Frauen im Durchschnitt mehr als Männer in bestimmten Tätigkeitsarten.

Hinweis: In diesem Beispiel wird die Datendatei *Employee data* verwendet.

1. Fügen Sie einen Diagrammtafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.
2. Wählen Sie auf der Registerkarte "Basis" die Optionen *Employment Category* (Art der Tätigkeit) und *Current Salary* (Aktuelles Gehalt) aus. (Wenn Sie bei gedrückter Steuertaste klicken, können Sie mehrere Felder bzw. Variablen auswählen.)
3. Wählen Sie **Balken** aus.
4. Wählen Sie aus der Liste "Auswertung" den Eintrag **Mittelwert** aus.
5. Klicken Sie auf die Registerkarte "Detailliert". Beachten Sie, dass Ihre Auswahl auf der vorherigen Registerkarte hier berücksichtigt wird.
6. Wählen Sie in der Gruppe "Optionale Formatierungen" die Option *Geschlecht* aus der Dropdown-Liste "Farbe" aus.
7. Klicken Sie auf **Ausführen**.

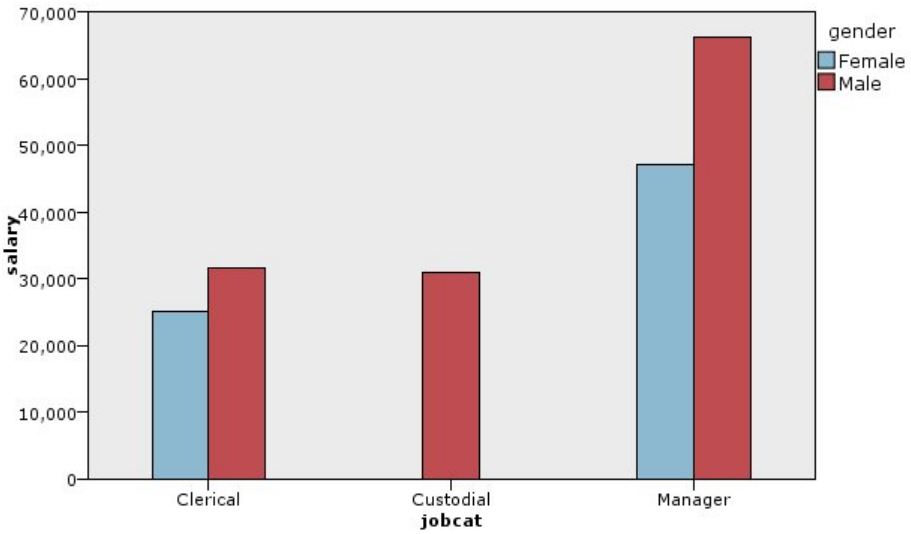


Abbildung 11. Gruppiertes Balkendiagramm

Wir stellen Folgendes fest:

- Der Unterschied im mittleren Gehalt für die einzelnen Arten der Tätigkeiten scheint nicht genauso groß zu sein wie im Balkendiagramm, in dem das mittlere Gehalt für alle Männer und Frauen verglichen wurde. Vielleicht gibt es eine unterschiedliche Anzahl an Männern und Frauen in den einzelnen Gruppen. Das könnten wir überprüfen, indem wir ein Balkendiagramm der Häufigkeiten erstellen.
- Unabhängig von der Art der Tätigkeit ist das mittlere Gehalt für Männer immer größer als das der Frauen.

Beispiel: Unterteiltes Histogramm

Wir erstellen ein Histogramm, das nach Geschlechtern unterteilt ist, um die Häufigkeitsverteilung der Gehälter von Männern und Frauen vergleichen zu können. Die Häufigkeitsverteilung zeigt, wie viele Fälle

bzw. Zeilen innerhalb eines bestimmten Gehaltsbereichs liegen. Mit dem unterteilten Histogramm können wir den Unterschied bei den Gehältern von Männern und Frauen genauer analysieren.

Hinweis: In diesem Beispiel wird die Datendatei *Employee data* verwendet.

1. Fügen Sie einen Diagrammtafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.
2. Wählen Sie auf der Registerkarte "Basis" die Option *Current Salary* (Aktuelles Gehalt) aus.
3. Wählen Sie **Histogramm** aus.
4. Klicken Sie auf die Registerkarte "Detailliert".
5. Wählen Sie in der Gruppe "Aufteilungen und Animation" die Option *gender* (Geschlecht) aus der Dropdown-Liste "Aufteilen nach" aus.
6. Klicken Sie auf **Ausführen**.

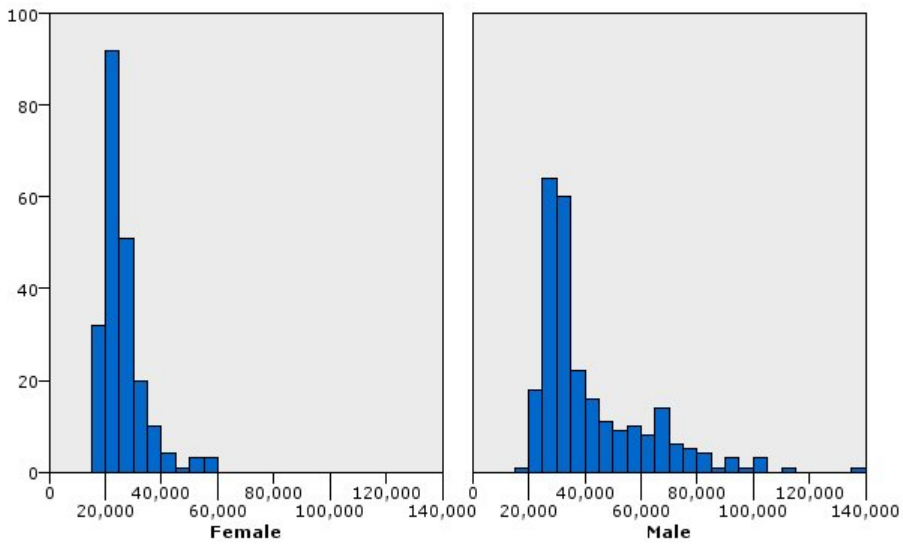


Abbildung 12. Unterteiltes Histogramm

Wir stellen Folgendes fest:

- Keine der beiden Häufigkeitsverteilungen ist eine Normalverteilung. Das heißt, die Histogramme stellen keine Glockenkurve dar, wie es bei einer Normalverteilung der Fall wäre.
- Die höheren Balken befinden sich auf der linken Seite des Diagramms. Das bedeutet, dass sowohl mehr Männer als auch mehr Frauen geringere als höhere Gehälter haben.
- Die Häufigkeitsverteilungen der Gehälter der Männer und der Frauen sind nicht identisch. Beachten Sie die Form der Histogramme. Es gibt mehr Männer, die höhere Gehälter erhalten, als Frauen, die höhere Gehälter erhalten.

Beispiel: Unterteiltes Punktdiagramm

Wie ein Histogramm zeigt auch ein Punktdiagramm die Verteilung eines kontinuierlichen numerischen Bereichs an. Im Gegensatz zu Histogrammen, die die Häufigkeiten für klassierte Datenbereiche darstellen, zeigen Punktdiagramme jede Zeile bzw. jeden Fall in den Daten an. Daher bietet ein Punktdiagramm im Vergleich zu Histogrammen eine höhere Granularität. Für die Analyse von Häufigkeitsverteilungen könnte ein Punktdiagramm sogar der geeignetere Ausgangspunkt sein.

Hinweis: In diesem Beispiel wird die Datendatei *Employee data* verwendet.

1. Fügen Sie einen Diagrammtafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.
2. Wählen Sie auf der Registerkarte "Basis" die Option *Current Salary* (Aktuelles Gehalt) aus.

3. Wählen Sie die Option **Punktdiagramm** aus.
4. Klicken Sie auf die Registerkarte "Detailliert".
5. Wählen Sie in der Gruppe "Aufteilungen und Animation" die Option *gender* (Geschlecht) aus der Dropdown-Liste "Aufteilen nach" aus.
6. Klicken Sie auf **Ausführen**.
7. Maximieren Sie das angezeigte Ausgabefenster, um das Diagramm besser sehen zu können.

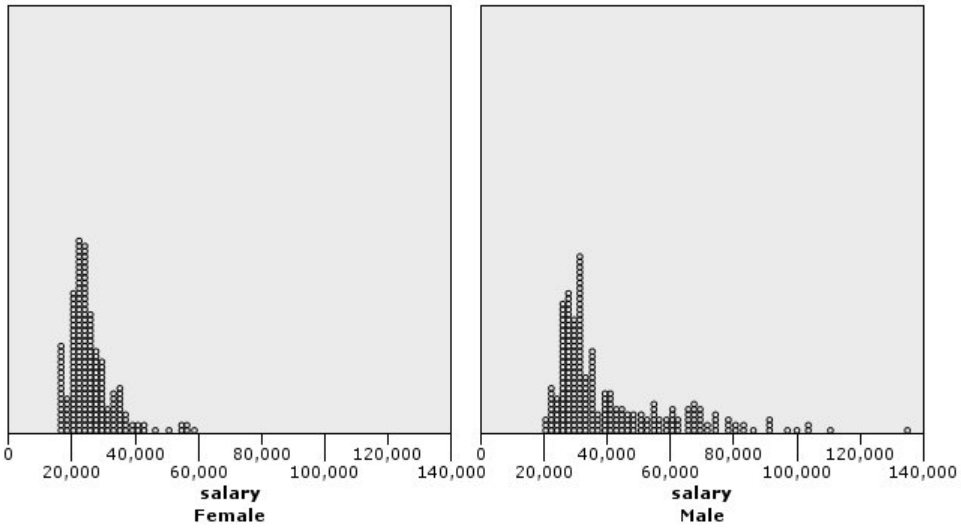


Abbildung 13. Unterteiltes Punktdiagramm

Im Vergleich zum Histogramm (siehe „Beispiel: Unterteiltes Histogramm" auf Seite 196) stellen wir Folgendes fest:

- Die Spitze bei 20.000, die sich im Histogramm für Frauen ergab, ist im Punktdiagramm weniger deutlich ausgeprägt. Es sind viele Fälle und Zeilen um diesen Wert konzentriert, die meisten von ihnen liegen jedoch näher an 25.000. Dieses Granularitätsniveau ist im Histogramm nicht ersichtlich.
- Obwohl das Histogramm für Männer darauf hindeutet, dass das mittlere Gehalt für Männer nach 40.000 gleichmäßig abnimmt, zeigt das Punktdiagramm, dass die Verteilung nach diesem Wert bis 80.000 relativ einheitlich ist. Bei jedem Gehaltswert in diesem Bereich gibt es drei oder mehr Männer, die dieses Gehalt beziehen.

Beispiel: Boxplot

Ein Boxplot ist eine weitere sinnvolle Visualisierung, um darzustellen, wie die Daten verteilt sind. Ein Boxplot enthält mehrere statistische Messgrößen, die wir nach der Erstellung der Visualisierung kennenlernen werden.

Hinweis: In diesem Beispiel wird die Datendatei *Employee data* verwendet.

1. Fügen Sie einen Diagrammtafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.
2. Wählen Sie auf der Registerkarte "Basis" die Optionen *Gender* (Geschlecht) und *Current Salary* (Aktuelles Gehalt) aus. (Wenn Sie bei gedrückter Steuertaste klicken, können Sie mehrere Felder bzw. Variablen auswählen.)
3. Wählen Sie **Boxplot** aus.
4. Klicken Sie auf **Ausführen**.

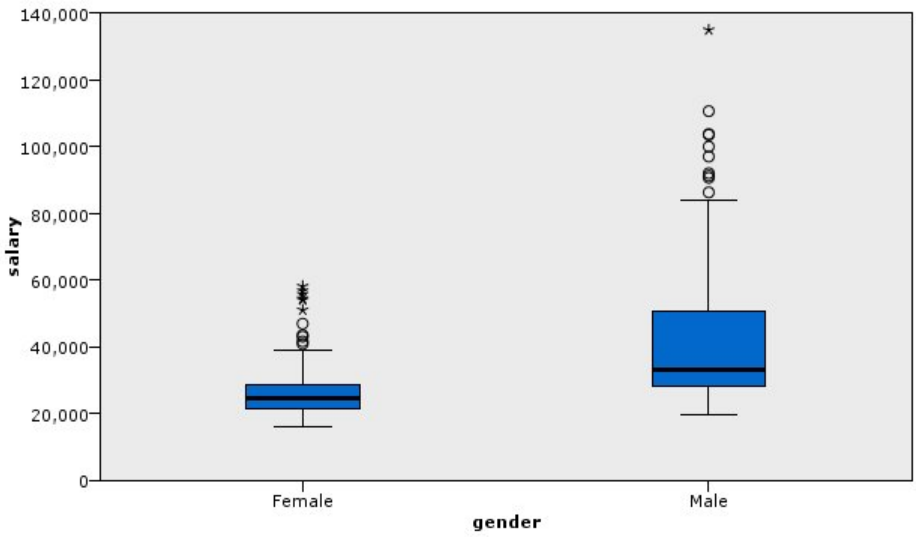


Abbildung 14. Boxplot

Machen wir uns zunächst mit den einzelnen Bereichen des Boxplots vertraut:

- Die dunkle Linie in der Mitte der Boxen ist der Median des Gehalts (*salary*), Die Hälfte der Fälle bzw. Zeilen besitzt einen höheren Wert als der Median und die andere Hälfte einen geringeren Wert. Wie der Mittelwert ist der Median eine Messgröße für Lagemaße. Im Gegensatz zum Mittelwert haben Fälle bzw. Zeilen mit Extremwerten weniger Einfluss auf den Median. In diesem Beispiel ist der Median kleiner als der Mittelwert (siehe „Beispiel: Balkendiagramm mit Auswertungsstatistik" auf Seite 195). Der Unterschied zwischen dem Mittelwert und dem Median deutet an, dass einige Fälle bzw. Zeilen mit Extremwerten den Mittelwert anheben. Das heißt, es gibt ein paar Angestellte, die große Gehälter beziehen.
- Im unteren Bereich der Box wird das 25. Perzentil dargestellt. 25 Prozent der Fälle/Zeilen haben Werte unter dem 25. Perzentil. Im oberen Bereich der Box wird das 75. Perzentil dargestellt. 25 Prozent der Fälle/Zeilen haben Werte über dem 75. Perzentil. Das bedeutet, dass 50 % der Fälle/Zeilen innerhalb der Box liegen. Die Box ist für Frauen wesentlich kürzer als für Männer. Das deutet darauf hin, dass das Gehalt (*salary*) bei Frauen weniger variiert als bei Männern. Der obere und untere Bereich der Box werden häufig als **Hinges** bezeichnet.
- Die T-Balken, die von den Boxen ausgehen, werden als **Fühler** oder **Whisker** bezeichnet. Die Länge beträgt das 1,5-Fache der Höhe der Box oder falls keine Fälle bzw. Zeilen mit Werten in diesem Bereich vorhanden sind, wird die Länge durch den maximalen bzw. minimalen Wert festgelegt. Bei einer Normalverteilung der Daten wird erwartet, dass circa 95 % der Daten innerhalb der Fühler liegen. In diesem Beispiel sind die Fühler bei Frauen kürzer als bei den Männern. Auch das deutet darauf hin, dass das Gehalt (*salary*) bei Frauen weniger variiert als bei Männern.
- Die Punkte sind **Ausreißer**. Ausreißer sind Werte, die nicht innerhalb der Fühler liegen. Ausreißer sind Extremwerte. Die Sterne sind **extreme Ausreißer**. Das sind all jene Fälle/Zeilen, deren Werte mehr als dreimal so groß sind wie die Höhe der Boxen. Es sind mehrere Ausreißer bei Frauen und Männern vorhanden. Berücksichtigen Sie, dass der Mittelwert größer als der Median ist. Der höhere Mittelwert wird von diesen Ausreißern verursacht.

Beispiel: Kreisdiagramm

Wir verwenden nun ein anderes Dataset, um andere Visualisierungstypen kennenzulernen. Das Dataset *customer_subset* ist eine hypothetische Datendatei mit Informationen über Kunden.

Zunächst erstellen wir ein Kreisdiagramm, um zu ermitteln, welche Anteile der Kunden in verschiedenen geografischen Regionen zu finden sind.

1. Fügen Sie einen Statistics-Quellenknoten hinzu, der auf *customer_subset.sav* verweist.

2. Fügen Sie einen Diagrammtafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.
3. Wählen Sie auf der Registerkarte "Basis" die Option *Geographic indicator* (Geografischer Indikator) aus.
4. Wählen Sie **Kreisdiagramm der Häufigkeiten** aus.
5. Klicken Sie auf **Ausführen**.

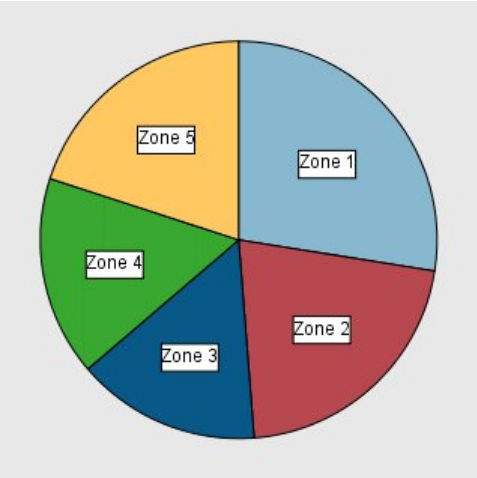


Abbildung 15. Kreisdiagramm

Wir stellen Folgendes fest:

- In Zone 1 leben mehr Kunden als in allen anderen Zonen.
- Die Kunden sind gleichmäßig auf die anderen Zonen verteilt.

Beispiel: Heat-Map

Wir erstellen nun eine kategoriale Heat-Map, um das mittlere Einkommen für Kunden in unterschiedlichen geografischen Regionen und Altersgruppen zu überprüfen.

Hinweis: Für dieses Beispiel wird die Datei *customer_subset* verwendet.

1. Fügen Sie einen Diagrammtafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.
2. Wählen Sie auf der Registerkarte "Basis" die Optionen *Geographic indicator* (Geografischer Indikator), *Age category* (Alterskategorie) and *Household income in thousands* (Haushaltseinkommen in Tausend) in der genannten Reihenfolge aus. (Wenn Sie bei gedrückter Steuertaste klicken, können Sie mehrere Felder bzw. Variablen auswählen.)
3. Wählen Sie **Heat-Map** aus.
4. Klicken Sie auf **Ausführen**.
5. Klicken Sie im angezeigten Ausgabefenster auf die Symbolleistenschaltfläche "Feld- und Wertbeschriftungen anzeigen" (die rechte der beiden Schaltflächen in der Mitte der Symbolleiste).

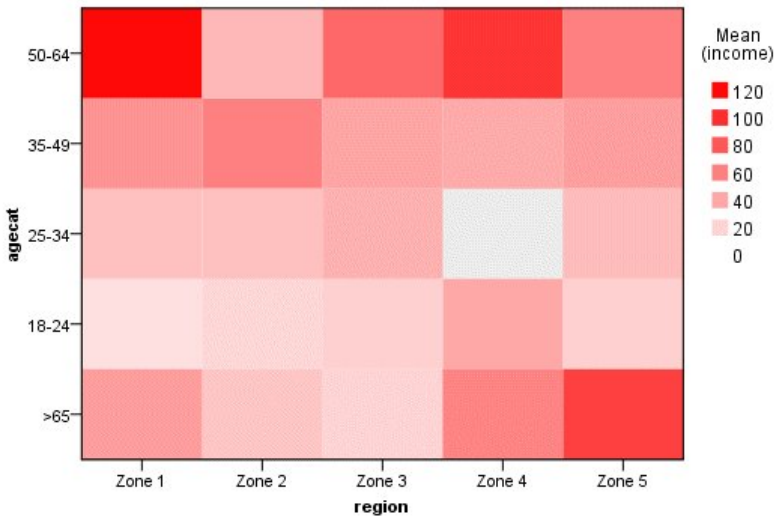


Abbildung 16. Kategoriale Heat-Map

Wir stellen Folgendes fest:

- Eine Heat-Map verhält sich wie eine Tabelle, in der anstelle von Zahlen Farben verwendet werden, um die Werte der Zellen darzustellen. Ein kräftiges Rot steht für den höchsten Wert, während Grau den niedrigsten Wert darstellt. Der Wert der einzelnen Zellen ist der Mittelwert des stetigen Felds bzw. der stetigen Variablen für jedes Kategoriepaar.
- Mit Ausnahme der Zonen 2 und 5 hat die Kundengruppe, deren Alter zwischen 50 und 64 liegt, ein höheres mittleres Haushaltseinkommen als die anderen Gruppen.
- In Zone 4 gibt es keine Kunden im Alter von 25 bis 34.

Beispiel: Streudiagrammmatrix (SPLOM)

Wir erstellen eine Streudiagrammmatrix aus mehreren unterschiedlichen Variablen, um feststellen zu können, ob Zusammenhänge zwischen den Variablen im Dataset bestehen.

Hinweis: Für dieses Beispiel wird die Datei *customer_subset* verwendet.

1. Fügen Sie einen Diagrammtafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.
2. Wählen Sie auf der Registerkarte "Basis" die Optionen *Age in years* (Alter in Jahren), *Household income in thousands* (Haushaltseinkommen in Tausend) und *Credit card debt in thousands* (Schulden auf Kreditkarte in Tausend) aus. (Wenn Sie bei gedrückter Steuertaste klicken, können Sie mehrere Felder bzw. Variablen auswählen.)
3. Wählen Sie **SPLOM** aus.
4. Klicken Sie auf **Ausführen**.
5. Maximieren Sie das Ausgabefenster, um die Matrix besser sehen zu können.

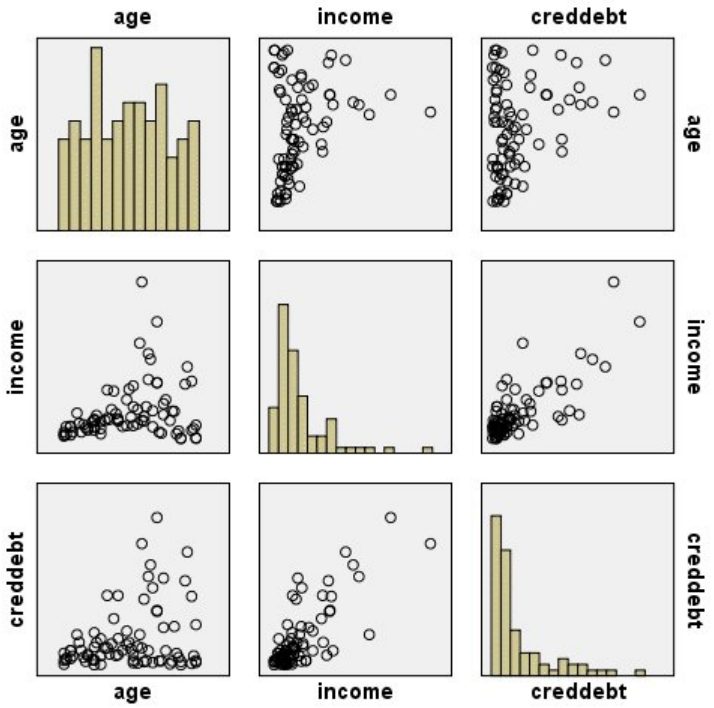


Abbildung 17. Streudiagrammmatrix (SPLOM)

Wir stellen Folgendes fest:

- Die auf der Diagonale angezeigten Histogramme stellen die Verteilung jeder Variablen in der SPLOM dar. Das Histogramm für *age* (Alter) wird in der oberen linken Zelle, das für *income* (Einkommen) in der mittleren Zelle und das für *creddebt* (Kreditkartenschulden) in der Zelle unten rechts dargestellt. Keine der Variablen weist eine Normalverteilung auf. Das heißt, keines der Histogramme ähnelt einer Glockenkurve. Beachten Sie auch, dass die Histogramme für *income* (Einkommen) und *creddebt* (Kreditkartenschulden) positiv schief sind.
- Es scheint keine Beziehung zwischen *age* (Alter) und den anderen Variablen zu geben.
- Zwischen *income* (Einkommen) und *creddebt* (Kreditkartenschulden) besteht ein lineares Verhältnis. Das heißt, die Kreditkartenschulden (*creddebt*) steigen, wenn das Einkommen (*income*) zunimmt. Gegebenenfalls können eigene Streudiagramme für diese Variablen und andere zugehörige Variablen erstellt werden, um die Beziehungen genauer zu untersuchen.

Beispiel: Choroplethenkarten (Farbkarten) von Summen

Nun erstellen wir eine Kartenvisualisierung. Im anschließenden Beispiel erstellen wir dann eine Variation dieser Visualisierung. Beim Dataset handelt es sich um *worldsales*. Dieses ist eine hypothetische Datendatei, die Verkaufserlöse nach Kontinent und Produkt enthält.

1. Fügen Sie einen Diagrammtafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.
2. Wählen Sie auf der Registerkarte "Basis" die Optionen *Continent* (Kontinent) und *Revenue* (Ertrag) aus. (Wenn Sie bei gedrückter Steuertaste klicken, können Sie mehrere Felder bzw. Variablen auswählen.)
3. Wählen Sie **Choroplethenkarte von Summen** aus.
4. Klicken Sie auf die Registerkarte "Detailliert".
5. Wählen Sie in der Gruppe "Optionale Formatierungen" die Option *Continent* (Kontinent) aus der Dropdown-Liste der Datenbeschriftungen aus.
6. Klicken Sie in der Gruppe "Kartendateien" auf **Kartendatei auswählen**.
7. Stellen Sie sicher, dass **Karte** im Dialogfeld "Karten auswählen" mit *Continents* (Kontinente) und **Kartenschlüssel** mit *CONTINENT* (KONTINENT) festgelegt ist.

8. Klicken Sie in den Gruppen "Karten- und Datenwerte vergleichen" auf **Vergleichen**, um sicherzustellen, dass die Kartenschlüssel mit den Datenschlüsseln übereinstimmen. In diesem Beispiel weisen alle Datenschlüsselwerte entsprechende Kartenschlüssel und Funktionen auf. Außerdem wird angezeigt, dass für Ozeanien keine Daten vorliegen.
9. Klicken Sie im Dialogfeld "Karten auswählen" auf **OK**.
10. Klicken Sie auf **Ausführen**.

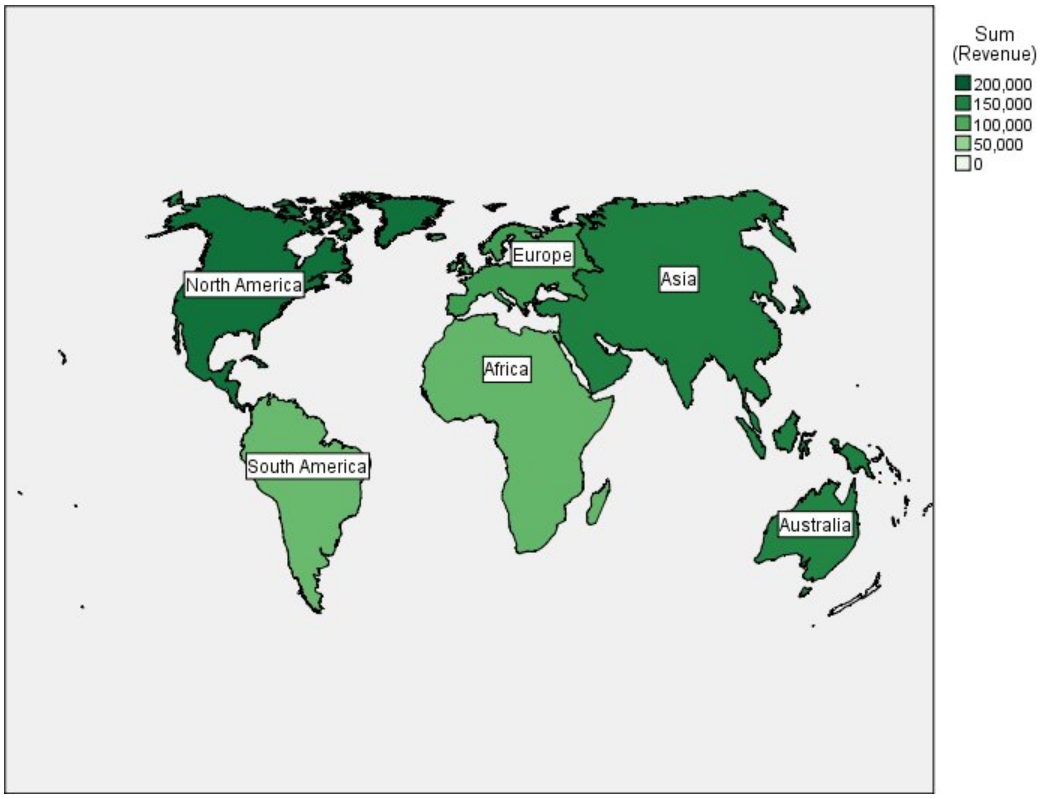


Abbildung 18. Choroplethenkarte von Summen

Mit dieser Kartenvisualisierung können wir mühelos erkennen, dass der Ertrag in Nordamerika am höchsten und in Südamerika sowie in Afrika am niedrigsten ist. Jeder Kontinent ist beschriftet, da wir für die Datenbeschriftungsformatierung die Option *Continent* (Kontinent) ausgewählt haben.

Beispiel: Balkendiagramme auf einer Karte

In diesem Beispiel wird verdeutlicht, wie sich der Ertrag auf jedem Kontinent nach Produkt aufteilen lässt.

Hinweis: In diesem Beispiel wird die Datendatei *worldsales* verwendet.

1. Fügen Sie einen Diagrammtafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.
2. Wählen Sie auf der Registerkarte "Basis" die Optionen *Continent* (Kontinent), *Product* (Produkt) und *Revenue* (Ertrag) aus. (Wenn Sie bei gedrückter Steuertaste klicken, können Sie mehrere Felder bzw. Variablen auswählen.)
3. Wählen Sie **Balken auf einer Karte** aus.
4. Klicken Sie auf die Registerkarte "Detailliert".

   Beim Verwenden mehrerer Felder eines bestimmten Typs ist es wichtig zu prüfen, dass jedes Feld dem richtigen Abschnitt zugewiesen ist.
5. Wählen Sie in der Dropdown-Liste "Kategorien" den Eintrag *Product* (Produkt) aus.
6. Wählen Sie in der Dropdown-Liste "Werte" den Eintrag *Revenue* (Ertrag) aus.

7. Wählen Sie in der Dropdown-Liste "Datenschlüssel" den Eintrag *Continent* (Kontinent) aus.
8. Wählen Sie in der Dropdown-Liste "Zusammenfassung" den Eintrag *Sum* (Summe) aus.
9. Klicken Sie in der Gruppe "Kartendateien" auf **Kartendatei auswählen**.
10. Stellen Sie sicher, dass **Karte** im Dialogfeld "Karten auswählen" mit *Continents* (Kontinente) und **Kartenschlüssel** mit *CONTINENT* (KONTINENT) festgelegt ist.
11. Klicken Sie in den Gruppen "Karten- und Datenwerte vergleichen" auf **Vergleichen**, um sicherzustellen, dass die Kartenschlüssel mit den Datenschlüsseln übereinstimmen. In diesem Beispiel weisen alle Datenschlüsselwerte entsprechende Kartenschlüssel und Funktionen auf. Außerdem wird angezeigt, dass für Ozeanien keine Daten vorliegen.
12. Klicken Sie im Dialogfeld "Karten auswählen" auf **OK**.
13. Klicken Sie auf **Ausführen**.
14. Maximieren Sie das angezeigte Ausgabefenster, um die Anzeige besser sehen zu können.

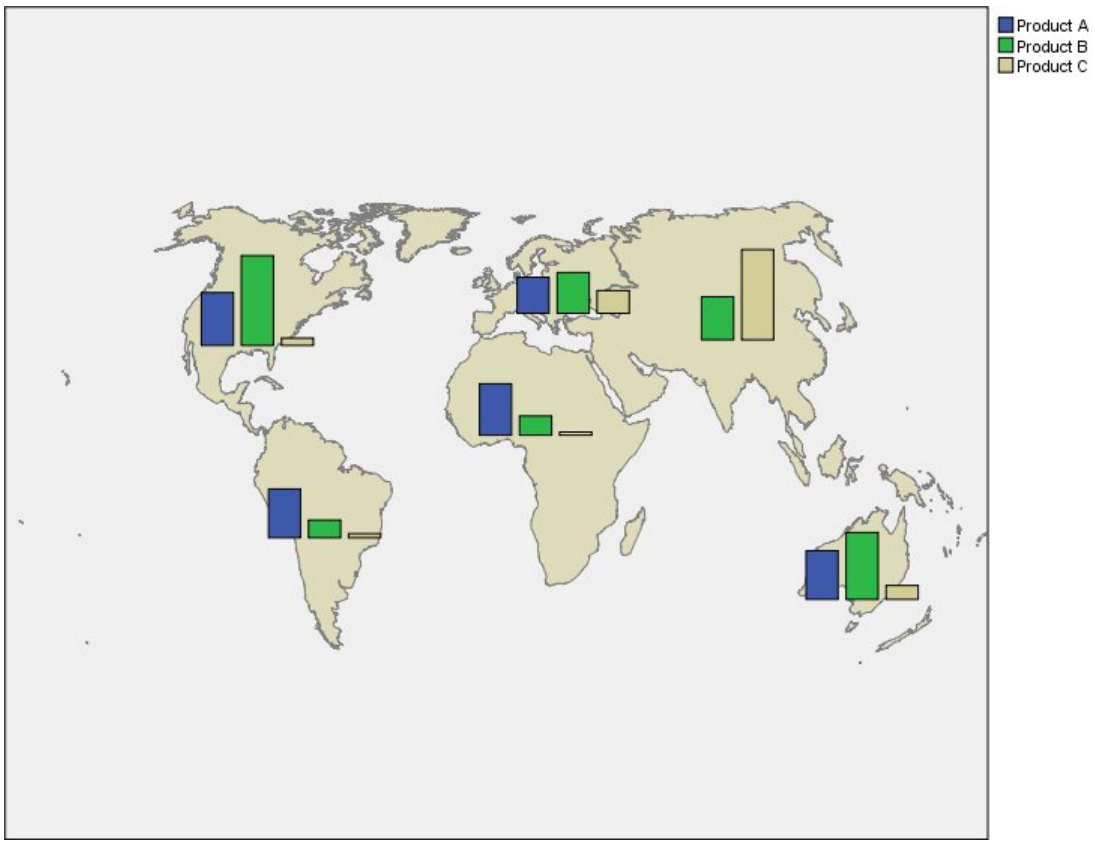


Abbildung 19. Balkendiagramme auf einer Karte

Wir stellen Folgendes fest:

- Die Verteilung des Gesamtertrags über Produkte hinweg ist in Südamerika und Afrika sehr ähnlich.
- *Product C* (Produkt C) erzeugt mit Ausnahme von Asien am wenigsten Ertrag.
- *Product A* (Produkt A) erbringt in Asien keinen bzw. nur minimalen Ertrag.

Diagrammtafel - Registerkarte "Darstellung"

Vor der Diagrammerstellung können Sie Darstellungsoptionen angeben.

Allgemeine Darstellungsoptionen

Titel. Dient zur Eingabe des Texts, der als Titel des Diagramms verwendet werden soll.

Untertitel. Dient zur Eingabe des Texts, der als Untertitel des Diagramms verwendet werden soll.

Titelzeile. Dient zur Eingabe des Texts, der als Titelzeile des Diagramms verwendet werden soll.

Stichprobenziehung Geben Sie eine Methode für umfangreichere Datasets an. Sie können eine maximal zulässige Größe für das Dataset angeben oder den Standardwert für die Anzahl an Datensätzen verwenden. Bei umfangreichen Datasets steigt die Leistung, wenn Sie die Option **Stichprobe** aktivieren. Alternativ können Sie mit **Alle Daten verwenden** alle Datenpunkte gleichzeitig plotten lassen; dies kann sich jedoch beträchtlich auf die Leistung der Software auswirken.

Optionen für die Style-Sheet-Darstellung

Es gibt zwei Schaltflächen, über die Sie steuern können, welche Visualisierungsvorlagen (und Style-Sheets und Zuordnungen) verfügbar sind:

Verwalten. Verwalten Sie Visualisierungsvorlagen, Style-Sheets und Karten auf Ihrem Computer. Sie können Visualisierungsvorlagen, Style-Sheets und Karten auf Ihrem lokalen System importieren, exportieren, umbenennen und löschen. Weitere Informationen finden Sie im Thema „Verwalten von Vorlagen, Style-Sheets und Kartendateien" auf Seite 206.

Speicherort. Ändern Sie den Speicherort von Visualisierungsvorlagen und, Style-Sheets und Karten. Der aktuelle Speicherort wird rechts neben der Schaltfläche angezeigt. Weitere Informationen finden Sie im Thema „Festlegen des Speicherorts für Vorlagen, Style-Sheets und Karten" auf Seite 206.

Das folgende Beispiel zeigt, wo sich die Darstellungsoptionen in einem Diagramm befinden. (*Hinweis:* Nicht in jedem Diagramm steht jede Option zur Verfügung.)

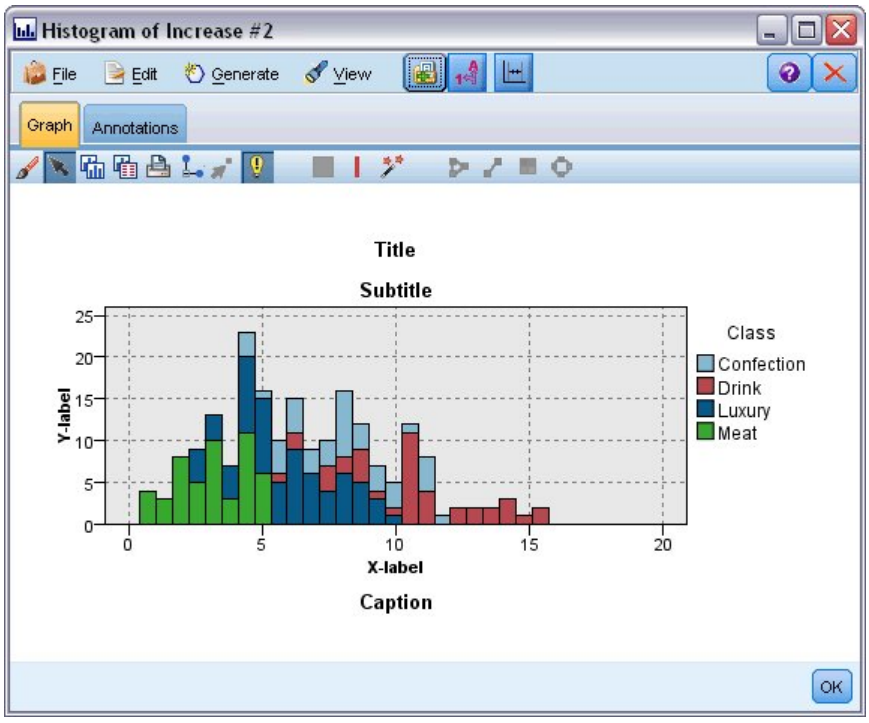


Abbildung 20. Position verschiedener Diagrammdarstellungsoptionen

Festlegen des Speicherorts für Vorlagen, Style-Sheets und Karten

Visualisierungsvorlagen, Visualisierungs-Style-Sheets und Kartendateien werden in einem speziellen lokalen Ordner oder im IBM SPSS Collaboration and Deployment Services Repository gespeichert. Bei der Auswahl von Vorlagen, Style-Sheets und Karten werden nur die in diesem Speicherort integrierten Elemente angezeigt. Wenn Sie alle Vorlagen, Style-Sheets und Karten an einer Position speichern, können die IBM SPSS-Anwendungen problemlos darauf zugreifen. Informationen zum Hinzufügen weiterer Vorlagen, Style-Sheets und Karten zu diesem Speicherort finden Sie unter „Verwalten von Vorlagen, Style-Sheets und Kartendateien".

So legen Sie den Speicherort von Vorlagen, Style-Sheets und Karten fest

1. Klicken Sie im Dialogfeld für eine Vorlage bzw. ein Style-Sheet auf die Schaltfläche **Kategorie...**, um das Dialogfeld "Vorlagen, Style-Sheets und Karten" anzuzeigen.
2. Wählen Sie eine Option für den Standardspeicherort für Vorlagen, Style-Sheets und Kartendateien aus:

   Lokales System. Vorlagen, Style-Sheets und Kartendateien werden in einem speziellen Ordner auf Ihrem lokalen Computer gespeichert. Unter Windows XP befindet sich dieser Ordner unter folgendem Pfad: *C:\Dokumente und Einstellungen\<Benutzername>\Anwendungsdaten\SPSSInc\Graphboard*. Der Ordner kann nicht geändert werden.

   IBM SPSS Collaboration and Deployment Services Repository. Vorlagen, Style-Sheets und Kartendateien werden in einem vom Benutzer angegebenen Ordner im IBM SPSS Collaboration and Deployment Services Repository gespeichert. Um den Ordner anzugeben, klicken Sie auf **Ordner**. Weitere Informationen finden Sie unter „IBM SPSS Collaboration and Deployment Services Repository als Speicherort für Vorlagen, Style-Sheets und Kartendateien".
3. Klicken Sie auf **OK**.

IBM SPSS Collaboration and Deployment Services Repository als Speicherort für Vorlagen, Style-Sheets und Kartendateien

Visualisierungsvorlagen und Style-Sheets können im IBM SPSS Collaboration and Deployment Services Repository gespeichert werden. Der Speicherort ist ein bestimmter Ordner im IBM SPSS Collaboration and Deployment Services Repository. Wird er als Standardposition festgelegt, stehen alle an dieser Position gespeicherten Vorlagen, Style-Sheets und Kartendateien zur Auswahl zur Verfügung.

So legen Sie einen Ordner in IBM SPSS Collaboration and Deployment Services Repository als Speicherort für Vorlagen, Style-Sheets und Kartendateien fest

1. Klicken Sie in einem Dialogfeld mit der Schaltfläche "Kategorie..." auf **Kategorie...**.
2. Wählen Sie IBM SPSS Collaboration and Deployment Services Repository.
3. Klicken Sie auf **Ordner**.

   Hinweis: Wenn Sie nicht bereits mit einem IBM SPSS Collaboration and Deployment Services Repository verbunden sind, werden Sie zur Eingabe von Verbindungsdaten aufgefordert.
4. Wählen Sie im Dialogfeld "Ordner wählen" den Ordner aus, in dem Vorlagen, Style-Sheets und Kartendateien gespeichert werden.
5. Optional können Sie über **Beschriftung abrufen** eine Beschriftung auswählen. Es werden nur Vorlagen, Style-Sheets und Kartendateien mit dieser Beschriftung angezeigt.
6. Wenn Sie einen Ordner mit einer bestimmten Vorlage oder Kartendatei oder einem bestimmten Style-Sheet suchen, können Sie die Vorlage, das Style-Sheet oder die Kartendatei mithilfe der Registerkarte "Suchen" ausfindig machen. Im Dialogfeld "Ordner auswählen" wird automatisch der Ordner ausgewählt, in dem sich die gefundene Vorlage oder Kartendatei oder das Style-Sheet befindet.
7. Klicken Sie auf **Ordner auswählen**.

Verwalten von Vorlagen, Style-Sheets und Kartendateien

Sie können die Vorlagen, Style-Sheets und Kartendateien am lokalen Speicherort auf Ihrem Computer über das Dialogfeld "Vorlagen, Style-Sheets und Karten verwalten" verwalten. Über dieses Dialogfeld

können Sie Visualisierungsvorlagen, Style-Sheets und Kartendateien am lokalen Speicherort auf Ihrem Computer importieren, exportieren, umbenennen und löschen.

Klicken Sie auf **Verwalten...** in einem der Dialogfelder, in dem Sie Vorlagen, Style-Sheets oder Karten auswählen.

Dialogfeld "Vorlagen, Style-Sheets und Karten verwalten"

Auf der Registerkarte "Vorlage" werden alle lokalen Vorlagen aufgelistet. Auf der Registerkarte "Style-Sheet" werden alle lokalen Style-Sheets aufgelistet und Beispielvisualisierungen mit Beispieldaten angezeigt. Sie können ein Style-Sheet auswählen, um dessen Stile auf die Beispielvisualisierungen anzuwenden. Weitere Informationen finden Sie im Thema „Zuweisen von Style-Sheets" auf Seite 278. Auf der Registerkarte "Karte" werden alle lokalen Kartendateien aufgelistet. Auf dieser Registerkarte wird außerdem Folgendes angezeigt: die Kartenschlüssel einschließlich Beispielwerten, ein Kommentar, sofern einer beim Erstellen der Karte angegeben wurde, sowie eine Vorschau der Karte.

Die folgenden Schaltflächen gelten für die jeweils aktive Registerkarte.

Importieren. Dient zum Importieren einer Visualisierungsvorlage, eines Style-Sheets oder einer Kartendatei aus dem Dateisystem. Der Import einer Vorlage, eines Style-Sheets oder eine Kartendatei macht das betreffende Element in der IBM SPSS-Anwendung verfügbar. Wenn Ihnen ein anderer Benutzer eine Vorlage, ein Style-Sheet oder eine Kartendatei sendet, müssen Sie die Datei importieren, bevor Sie sie in Ihrer Anwendung verwenden.

Exportieren. Dient zum Exportieren einer Visualisierungsvorlage, eines Style-Sheets oder einer Kartendatei aus dem Dateisystem. Exportieren Sie eine Visualisierungsvorlage, ein Style-Sheet bzw. eine Kartendatei, wenn Sie sie einem anderen Benutzer senden möchten.

Umbenennen. Dient zum Umbenennen der ausgewählten Visualisierungsvorlage, des Style-Sheets oder der Kartendatei. Sie können einen Namen nicht in einen Namen ändern, der bereits verwendet wird.

Kartenschlüssel exportieren. Dient zum Exportieren der Kartenschlüssel als Datei mit kommagetrennten Werten (CSV-Datei). Diese Schaltfläche ist nur auf der Registerkarte "Karte" aktiviert.

Löschen. Dient zum Löschen der ausgewählten Visualisierungsvorlagen, Style-Sheets oder Kartendateien. Sie können mehrere Vorlagen, Style-Sheets oder Kartendateien durch Klicken bei gedrückter Steuertaste auswählen. Das Löschen kann nicht rückgängig gemacht werden. Handeln Sie daher mit Bedacht.

Umwandeln und Verteilen von Kartenshapefiles

Mit der Auswahlfunktion für Diagrammtafelvorlagen können Sie Kartenvisualisierungen aus der Kombination aus einer Visualisierungsvorlage und einer SMZ-Datei erstellen. SMZ-Dateien ähneln ESRI Shapefiles (Dateiformat SHP) dahingehend, dass sie die geografischen Informationen zum Zeichnen einer Karte (z. B. Ländergrenzen) enthalten. Sie sind jedoch für Kartenvisualisierungen optimiert. Die Auswahlfunktion für Diagrammtafelvorlagen ist mit einer ausgewählten Anzahl an SMZ-Dateien vorinstalliert. Wenn Sie bereits eine ESRI-Shapefile besitzen, die Sie für Kartenvisualisierungen verwenden möchten, müssen Sie zunächst die Shapefile mithilfe des Dienstprogramms zur Konvertierung von Karten in eine SMZ-Datei konvertieren. Das Dienstprogramm zur Konvertierung von Karten unterstützt ESRI-Shapefiles mit Punkten, Mehrfachlinien oder Polygonen (Formtypen 1, 3 und 5), die eine einzige Schicht enthalten.

Zusätzlich zur Konvertierung von ESRI-Shapefiles können Sie mit dem Dienstprogramm zur Konvertierung von Karten auch den Detaillierungsgrad der Karte ändern, Strukturbeschriftungen ändern, Strukturen zusammenführen und Strukturen verschieben sowie zahlreiche weitere optionale Änderungen vornehmen. Außerdem können Sie mit dem Dienstprogramm zur Konvertierung von Karten auch bestehende SMZ-Dateien (einschließlich der vorinstallierten) ändern.

Bearbeiten vorinstallierter SMZ-Dateien

1. Exportieren Sie die SMZ-Datei aus dem Verwaltungssystem. Weitere Informationen finden Sie im Thema „Verwalten von Vorlagen, Style-Sheets und Kartendateien" auf Seite 206.
2. Verwenden Sie das Dienstprogramm zur Konvertierung von Karten zum Öffnen und Bearbeiten der exportierten SMZ-Datei. Es wird empfohlen, die Datei unter einem anderen Namen zu speichern. Weitere Informationen finden Sie im Thema „Verwenden des Dienstprogramms zur Konvertierung von Karten" auf Seite 209.
3. Importieren Sie die geänderte SMZ-Datei in das Verwaltungssystem. Weitere Informationen finden Sie im Thema „Verwalten von Vorlagen, Style-Sheets und Kartendateien" auf Seite 206.

Zusätzliche Ressourcen für Kartendateien

Geodaten im Dateiformat SHP, die für Ihre kartenbezogenen Anforderungen verwendet werden können, werden von vielen privaten und öffentlichen Quellen angeboten. Informieren Sie sich auf den Websites Ihrer jeweiligen Regierung, wenn Sie nach kostenlosen Daten suchen. Viele der Vorlagen dieses Produkts basieren auf öffentlich verfügbaren Daten, die von GeoCommons ( ) und der Volkszählungsbehörde U.S. Census Bureau (http://www.census.gov) bezogen wurden.

WICHTIGER HINWEIS: Alle Informationen zu Produkten anderer Anbieter stammen von den Anbietern der aufgeführten Produkte, deren veröffentlichten Ankündigungen oder anderen allgemein verfügbaren Quellen. IBM hat diese Produkte nicht getestet und kann daher keine Aussagen zu Leistung, Kompatibilität oder anderen Merkmalen machen. Fragen zu den Leistungsmerkmalen von Produkten anderer Anbieter sind an den
jeweiligen Anbieter zu richten. Verweise in diesen Informationen auf Websites anderer Anbieter werden lediglich als Service für den Kunden bereitgestellt und stellen keinerlei Billigung des Inhalts dieser Websites dar. Das Material auf diesen Websites ist kein Bestandteil des Materials dieses IBM Programms, sofern dies nicht in einer Datei mit Hinweisen zu diesem IBM Programm anders vermerkt ist. Die Verwendung des Materials dieser Websites erfolgt auf eigene Gefahr.

Wichtige Konzepte im Zusammenhang mit Karten

Sie sollten sich mit einigen wichtigen Konzepten im Zusammenhang mit Shapefiles vertraut machen, um das Dienstprogramm zur Konvertierung von Karten effektiver nutzen zu können.

Eine **Shapefile** stellt die geografischen Informationen bereit, die zum Zeichnen einer Karte erforderlich sind. Es gibt drei Typen von Shapefiles, die vom Dienstprogramm zur Konvertierung von Karten unterstützt werden:

- **Punkt.** Die Shapefile gibt die Position von Punkten (z. B. Städte) an.
- **Mehrfachlinie.** Die Shapefile gibt Wegverläufe und deren Position (z. B. Flüsse) an.
- **Polygon.** Die Shapefile gibt begrenzte Regionen und ihre Position (z. B. Länder) an.

Am häufigsten werden Polygonshapefiles verwendet. Choroplethenkarten werden aus Polygonshapefiles erstellt. Bei Choroplethenkarten werden Farben zur Darstellung von Werten innerhalb einzelner Polygone (Regionen) verwendet. Punkt- und Mehrfachlinienshapefiles werden üblicherweise auf eine Polygonshapefile gelegt (Überlagerung). Ein Beispiel hierfür ist eine Punktshapefile mit deutschen Städten, das eine Shapefile der deutschen Bundesländer überlagert.

Shapefiles bestehen aus **Strukturen**. Strukturen sind die einzelnen geografischen Elemente. Beispiele für Strukturen sind Länder, Bundesländer, Städte usw. Die Shapefile enthält auch Daten zu den Strukturen. Diese Daten werden in **Attributen** gespeichert. Attribute ähneln Feldern oder Variablen in einer Datendatei. Es gibt mindestens ein Attribut, das der **Kartenschlüssel** für die Struktur ist. Beim Kartenschlüssel kann es sich um eine Beschriftung, wie beispielsweise den Namen eines Landes oder Bundeslandes handeln. Der Kartenschlüssel ist das Element, das Sie mit einer Variablen bzw. einem Feld in einer Datendatei verknüpfen, um eine Kartenvisualisierung zu erstellen.

Beachten Sie, dass in der SMZ-Datei nur das Schlüsselattribut bzw. die Schlüsselattribute beibehalten werden können. Das Dienstprogramm zur Konvertierung von Karten unterstützt nicht die Speicherung zusätzlicher Attribute. Daher müssen Sie mehrere SMZ-Dateien erstellen, wenn Sie eine Zusammenfassung in verschiedenen Schichten durchführen möchten. Wenn Sie beispielsweise Bundesländer und Landkreise zusammenfassen möchten, benötigen Sie getrennte SMZ-Dateien: eine mit einem Schlüssel für die Bundesländer und eine mit einem Schlüssel für die Landkreise.

Verwenden des Dienstprogramms zur Konvertierung von Karten

So starten Sie das Dienstprogramm zur Konvertierung von Karten:

Wählen Sie die folgenden Befehle aus den Menüs aus:

Tools > **Dienstprogramm zur Konvertierung von Karten**

Das Dienstprogramm zur Konvertierung von Karten enthält vier Hauptbildschirme (Schritte). Einer der Schritte beinhaltet auch Unterschritte für detailliertere Festlegungen zur Bearbeitung der Kartendatei.

Schritt 1 - Ziel und Quellendatei auswählen

Zunächst müssen Sie eine Kartendatei als Quelle und ein Ziel für die konvertierte Kartendatei auswählen. Sie benötigen sowohl die *.shp* als auch die *.dbf*-Datei für die Shapefile.

Wählen Sie die zu konvertierende SHP-Datei (ESRI) oder SMZ-Datei aus. Navigieren Sie zu einer bestehenden Kartendatei auf Ihrem Computer. Dies ist die Datei, in die die Konvertierung erfolgt und die als SMZ-Datei gespeichert wird. Die *.dbf*-Datei für die Shapefile *muss* in demselben Verzeichnis gespeichert werden und einen Basisdateinamen tragen, der mit dem Namen der *.shp*-Datei übereinstimmt. Die *.dbf*-Datei ist erforderlich, da sie Attributinformationen für die *.shp*-Datei enthält.

Geben Sie Ziel und Dateinamen für die konvertierte Kartendatei an. Geben Sie einen Pfad und einen Dateinamen für die SMZ-Datei ein, die aus der ursprünglichen Kartenquellendatei erstellt wird.

- **In die Vorlagenauswahl importieren.** Zusätzlich zum Speichern einer Datei im Dateisystem können Sie die Karte optional zur Liste "Verwalten" der Vorlagenauswahl hinzufügen. Wenn Sie diese Option auswählen, steht die Karte in der Vorlagenauswahl automatisch für auf Ihrem Computer installierte IBM SPSS-Produkte zur Verfügung. Wenn Sie die Karte zu diesem Zeitpunkt nicht in die Vorlagenauswahl importieren, müssen Sie den Import zu einem späteren Zeitpunkt manuell nachholen. Weitere Informationen zum Importieren von Karten in das Verwaltungssystem der Vorlagenauswahl finden Sie unter „Verwalten von Vorlagen, Style-Sheets und Kartendateien” auf Seite 206.

Schritt 2 - Kartenschlüssel auswählen

Nun wählen Sie aus, welche Kartenschlüssel in die SMZ-Datei aufgenommen werden sollen. Sie können anschließend einige Optionen ändern, die beeinflussen, wie die Karte gerendert wird. Die anschließenden Schritte im Dienstprogramm zur Konvertierung von Karten beinhalten eine Vorschau der Karte. Die ausgewählten Rendering-Optionen werden zur Erzeugung der Kartenvorschau verwendet.

Primären Kartenschlüssel auswählen. Wählen Sie das Attribut aus, das als primärer Schlüssel zur Angabe und Beschriftung von Strukturen in der Karte dient. Der primäre Schlüssel einer Weltkarte könnte beispielsweise das Attribut für die Ländernamen sein. Der primäre Schlüssel verknüpft außerdem die Daten mit den Kartenstrukturen. Achten Sie also darauf, dass die Werte (Beschriftungen) des ausgewählten Attributs mit den Werten in Ihren Daten übereinstimmen. Bei der Auswahl eines Attributs werden Beispielbeschriftungen angezeigt. Wenn Sie diese Beschriftungen ändern möchten, können Sie dies in einem späteren Schritt tun.

Einzuschließende Alternativschlüssel auswählen. Markieren Sie neben dem primären Kartenschlüssel auch alle anderen Schlüsselattribute, die in die generierte SMZ-Datei aufgenommen werden sollen. So können beispielsweise manche Attribute übersetzte Beschriftungen enthalten. Wenn Sie Daten erwarten, die in anderen Sprachen codiert sind, ist es sinnvoll, diese Attribute beizubehalten. Beachten Sie, dass Sie

nur diejenigen zusätzlichen Schlüssel auswählen können, die für dieselben Strukturen stehen wie der primäre Schlüssel. Wenn es sich beim primären Schlüssel beispielsweise um die vollständigen Namen der US-Bundesstaaten handelt, können Sie nur diejenigen Alternativschlüssel auswählen, die ebenfalls für US-Bundesstaaten stehen, also beispielsweise die Kürzel der Bundesstaaten.

Karte automatisch glätten. Shapefiles mit Polygonen enthalten üblicherweise zu viele Datenpunkte und zu viele Details für statistische Kartenvisualisierungen. Die überschüssigen Details können ablenkend wirken und die Leistung beeinträchtigen. Durch Glätten können Sie den Detaillierungsgrad verringern und die Karte verallgemeinern. Die Karte sieht dadurch prägnanter aus lässt sich schneller rendern. Wenn die Karte automatisch geglättet wird, beträgt der maximale Winkel 15 Grad und der beizubehaltende Bereich beträgt 99. Informationen zu diesen Einstellungen finden Sie unter „Glätten der Karte". Beachten Sie, dass Sie die Gelegenheit haben, später in einem anderen Schritt eine weitere Glättung durchzuführen.

Grenzen zwischen sich berührenden Polygonen der gleichen Struktur entfernen. Einige Strukturen können Unterstrukturen enthalten, deren Grenzen innerhalb der relevanten Hauptstrukturen liegen. Beispielsweise kann eine Weltkarte der Kontinente interne Grenzen für die Länder auf den einzelnen Kontinenten aufweisen. Wenn Sie diese Option auswählen, werden die internen Grenzen nicht in der Karte angezeigt. Beim Beispiel mit der Weltkarte der Kontinente werden durch die Auswahl dieser Option die Ländergrenzen entfernt, während die Grenzen der Kontinente beibehalten werden.

Schritt 3 - Karte bearbeiten

Nachdem Sie die grundlegenden Optionen für die Karte angegeben haben, können Sie konkretere Optionen bearbeiten. Diese Änderungen sind optional. Dieser Schritt des Dienstprogramms zur Konvertierung von Karten führt Sie durch die zugehörigen Aufgaben und zeigt eine Vorschau der Karte an, mit der Sie die Änderungen überprüfen können. Je nach Shapefiletyp (Punkt, Mehrfachlinie oder Polygon) und Koordinatensystem stehen einige Aufgaben möglicherweise nicht zur Verfügung.

Für alle Aufgaben werden auf der linken Seite des Dienstprogramms zur Konvertierung von Karten die folgenden allgemeinen Steuerelemente angezeigt:

Beschriftungen auf Karte anzeigen. Standardmäßig werden in der Vorschau keine Strukturbeschriftungen angezeigt. Sie können auswählen, dass die Beschriftungen angezeigt werden sollen. Die Beschriftungen erleichtern zwar möglicherweise die Identifizierung der Strukturen, sie können jedoch die Direktauswahl auf der Vorschaukarte behindern. Aktivieren Sie diese Option, wenn Sie sie benötigen, beispielsweise für die Bearbeitung der Strukturbeschriftungen.

Farben für Vorschaukarte auswählen. Standardmäßig werden auf der Vorschaukarte Bereiche mit einer Volltonfarbe angezeigt. Alle Strukturen haben dieselbe Farbe. Sie können festlegen, dass den einzelnen Kartenstrukturen eine Reihe verschiedener Farben zugewiesen wird. Diese Option kann die Unterscheidung verschiedener Strukturen auf der Karte erleichtern. Sie ist besonders hilfreich, wenn Sie Strukturen zusammenführen und sehen möchten, wie die neuen Strukturen in der Vorschau dargestellt werden.

Für alle Aufgaben wird außerdem auf der rechten Seite des Dienstprogramms zur Konvertierung von Karten das folgende allgemeine Steuerelement angezeigt:

Rückgängig. Klicken Sie auf **Rückgängig**, um den vorherigen Zustand wiederherzustellen. Es können maximal 100 Änderungen rückgängig gemacht werden.

Glätten der Karte: Shapefiles mit Polygonen enthalten üblicherweise zu viele Datenpunkte und zu viele Details für statistische Kartenvisualisierungen. Die überschüssigen Details können ablenkend wirken und die Leistung beeinträchtigen. Durch Glätten können Sie den Detaillierungsgrad verringern und die Karte verallgemeinern. Die Karte sieht dadurch prägnanter aus lässt sich schneller rendern. Diese Option steht nicht für Punkt- und Mehrfachlinienkarten zur Verfügung.

Max. Winkel. Der maximale Winkel, dessen Wert zwischen 1 und 20 liegen muss, gibt die Toleranz für die Glättung von Punktesets an, die nahezu linear sind. Ein größerer Wert bietet größere Toleranz für die

lineare Glättung. Dabei wird anschließend eine größere Anzahl an Punkten verworfen, was zu einer verallgemeinerteren Karte führt. Zur Anwendung der linearen Glättung überprüft das Dienstprogramm zur Konvertierung von Karten den Innenwinkel, der jeweils durch drei Punkte auf der Karte gebildet wird. Wenn 180 minus dem Winkel kleiner ist als der angegebene Wert, verwirft das Dienstprogramm zur Konvertierung von Karten den mittleren Punkt. Anders ausgedrückt: das Dienstprogramm zur Konvertierung von Karten überprüft, ob die von den drei Punkten gebildete Linie annähernd gerade ist. Wenn dies der Fall ist, behandelt das Dienstprogramm zur Konvertierung von Karten die Linie als Gerade zwischen den Endpunkten und verwirft den mittleren Punkt.

Beizubehaltender Prozentsatz. Der beizubehaltende Prozentsatz, bei dem es sich um einen Wert zwischen 90 und 100 handeln muss, legt fest, welche Menge an Landbereich beim Glätten der Karte beibehalten werden soll. Diese Option betrifft nur diejenigen Strukturen, die mehrere Polygone aufweisen, wie es der Fall ist, wenn eine Struktur Inseln beinhaltet. Wenn der Gesamtbereich einer Struktur minus eines Polygons größer ist als der angegebene Prozentsatz des ursprünglichen Bereichs, verwirft das Dienstprogramm zur Konvertierung von Karten das Polygon aus der Karte. Das Dienstprogramm zur Konvertierung von Karten entfernt niemals sämtliche Polygone für die Struktur. Es bleibt also stets mindestens ein Polygon für die Struktur erhalten, unabhängig vom angewendeten Glättungsumfang.

Klicken Sie nach der Auswahl des maximalen Winkels und des beizubehaltenden Prozentsatzes auf **Zuweisen**. Die Vorschau wird mit den Glättungsänderungen aktualisiert. Wenn die Karte weiterer Glättung bedarf, wiederholen Sie den Vorgang, bis der gewünschte Glättungsgrad erreicht ist. Beachten Sie, dass die mögliche Glättung begrenzt ist. Beim wiederholten Glätten wird irgendwann ein Punkt erreicht, an dem keine weitere Glättung auf die Karte angewendet werden kann.

Strukturbeschriftungen bearbeiten: Sie können die Strukturbeschriftungen nach Bedarf bearbeiten (beispielsweise so, dass sie den erwarteten Daten entsprechen) und auch die Position der Beschriftungen auf der Karte ändern. Auch wenn Sie nicht glauben, dass sie die Beschriftungen ändern müssen, sollten Sie sie überprüfen, bevor Sie Visualisierungen aus der Karte erstellen. Da in der Vorschau Beschriftungen nicht standardmäßig angezeigt werden, kann es auch sinnvoll sein, die Option **Beschriftungen auf Karte anzeigen** auszuwählen, um sie sichtbar zu machen.

Schlüssel. Wählen Sie den Schlüssel aus, der die zu überprüfenden/bearbeitenden Strukturbeschriftungen enthält.

Strukturen. In dieser Liste werden die im ausgewählten Schlüssel enthaltenen Strukturbeschriftungen angezeigt. Zur Bearbeitung der Beschriftung doppelklicken Sie auf die Liste. Wenn Beschriftungen in der Karte angezeigt werden, können Sie auch direkt in der Kartenvorschau auf die Strukturbeschriftungen doppelklicken. Wenn Sie die Beschriftungen mit einer Datendatei vergleichen möchten, klicken Sie auf **Vergleichen**.

X/Y. Diese Textfelder geben den aktuellen Mittelpunkt der Beschriftung für die ausgewählte Struktur in der Karte an. Die Einheiten werden in den Koordinaten der Karte angezeigt. Dabei kann es sich um lokale kartesische Koordinaten handeln (z. B. das State Plane Coordinate System für die USA) oder um geografische Koordinaten (wobei **X** den Längengrad und **Y** den Breitengrad angibt). Geben Sie Koordinaten für die neue Position der Beschriftung an. Wenn Beschriftungen angezeigt werden, können Sie die Beschriftungen auch durch Klicken und Ziehen auf der Karte verschieben. Die Textfelder werden entsprechend der neuen Position aktualisiert.

Vergleichen. Wenn Sie über eine Datendatei mit Datenwerten verfügen, die den Strukturbeschriftungen für einen bestimmten Schlüssel entsprechen sollen, klicken Sie auf **Vergleichen**, um das Dialogfeld "Mit externer Datenquelle vergleichen" anzuzeigen. In diesem Dialogfeld können Sie die Datendatei öffnen und ihre Werte direkt mit den Strukturbeschriftungen des Kartenschlüssels vergleichen.

Dialogfeld "Mit externer Datenquelle vergleichen": Im Dialogfeld "Mit externer Datenquelle vergleichen" können Sie eine Datei mit tabstoppgetrennten Werten (Erweiterung *.txt*), eine Datei mit kommagetrennten Werten (Erweiterung *.csv*) oder eine für IBM SPSS Statistics formatierte Datendatei (Erweiterung *.sav*) öff-

nen. Wenn die Datei geöffnet ist, können Sie in der Datendatei ein Feld auswählen, das mit den Strukturbeschriftungen in einem bestimmten Kartenschlüssel verglichen werden soll. Anschließend können Sie etwaige Diskrepanzen in der Kartendatei korrigieren.

Felder in der Datendatei. Wählen Sie das Feld aus, dessen Werte mit den Strukturbeschriftungen verglichen werden sollen. Wenn die erste Zeile der *.txt*- bzw. *.csv*-Datei deskriptive Beschriftungen für die einzelnen Felder enthält, klicken Sie auf **Erste Zeile als Spaltenbeschriftungen verwenden**. Andernfalls werden die einzelnen Felder durch ihre Position in der Datendatei angegeben (z. B. "Spalte 1", "Spalte 2" usw.).

Schlüssel für Vergleich. Wählen Sie den Kartenschlüssel aus, dessen Strukturbeschriftungen mit den Feldwerten der Datendatei verglichen werden sollen.

Vergleichen. Klicken Sie, wenn Sie mit dem Vergleich der Werte beginnen möchten.

Vergleichsergebnisse. Standardmäßig werden in der Tabelle "Vergleichsergebnisse" nur die nicht zugeordneten Feldwerte in der Datendatei aufgelistet. Die Anwendung versucht, eine zugehörige Strukturbeschriftung zu finden, normalerweise durch Prüfung auf eingefügte oder fehlende Leerzeichen. Klicken Sie auf die Dropdown-Liste in der Spalte *Kartenbeschriftung*, um die Strukturbeschriftung in der Kartendatei mit dem angezeigten Feldwert abzugleichen. Wenn Ihre Kartendatei keine entsprechende Strukturbeschriftung enthält, wählen Sie die Option *Ohne Zuordnung belassen* aus. Wenn Sie alle Feldwerte anzeigen möchten, auch diejenigen, die bereits mit einer Strukturbeschriftung übereinstimmen, wählen Sie die Option **Nur nicht zugeordnete Fälle anzeigen** ab. Dies kann sinnvoll sein, um bestimmte Zuordnungen außer Kraft zu setzen.

Sie können jede Struktur nur einmal für die Zuordnung zu einem Feldwert verwenden. Wenn Sie mehrere Strukturen einem einzelnen Feldwert zuordnen möchten, können Sie die Strukturen zusammenführen und anschließend die neue, zusammengeführte Struktur dem Feldwert zuordnen. Weitere Informationen zum Zusammenführen von Strukturen finden Sie unter „Strukturen zusammenführen”.

Strukturen zusammenführen: Das Zusammenführen von Strukturen ist nützlich, um größere Regionen in einer Karte zu erstellen. Wenn Sie beispielsweise eine Karte der deutschen Bundesländer konvertieren, können Sie die Bundesländer (die Strukturen in diesem Beispiel) zu größeren Nord-, Süd-, Ost- und Westregionen zusammenführen.

Schlüssel. Wählen Sie den Kartenschlüssel aus, der die Strukturbeschriftungen enthält, mit denen Sie die zusammenzuführenden Strukturen identifizieren können.

Strukturen. Klicken Sie auf die erste zusammenzuführende Struktur. Klicken Sie bei gedrückter Steuertaste auf die anderen Strukturen, die zusammengeführt werden sollen. Beachten Sie, dass die Strukturen auch in der Kartenvorschau ausgewählt werden. Neben der Auswahl aus der Liste haben Sie auch die Möglichkeit, direkt in der Kartenvorschau auf die Strukturen zu klicken und dann beim Klicken die Steuertaste gedrückt zu halten.

Klicken Sie nach der Auswahl der zusammenzuführenden Strukturen auf **Zusammenführen**, um das Dialogfeld "Zusammengeführte Struktur benennen" anzuzeigen, in dem Sie eine Beschriftung auf die neue Struktur anwenden können. Sie können nach dem Zusammenführen der Strukturen auch die Option **Farben für Vorschaukarte auswählen** aktivieren, um sich zu vergewissern, dass die Ergebnisse Ihren Erwartungen entsprechen.

Nach dem Zusammenführen der Strukturen können Sie auch die Beschriftung für die neue Struktur verschieben. Dies ist in der Aufgabe *Strukturbeschriftungen bearbeiten* möglich. Weitere Informationen finden Sie im Thema „Strukturbeschriftungen bearbeiten” auf Seite 211.

Dialogfeld "Zusammengeführte Struktur benennen": Im Dialogfeld "Zusammengeführte Struktur benennen" können Sie der neu zusammengeführten Struktur Beschriftungen zuweisen.

In der Tabelle "Beschriftungen" werden Informationen für die einzelnen Schlüssel in der Kartendatei angezeigt. Außerdem können Sie dort den Schlüsseln jeweils eine Beschriftung zuweisen.

Neue Beschriftung. Geben Sie eine neue Beschriftung für die zusammengeführte Struktur ein, die dem betreffenden Kartenschlüssel zugewiesen werden soll.

Schlüssel. Der Kartenschlüssel, dem Sie die neue Beschriftung zuweisen.

Alte Beschriftungen. Die Beschriftungen für die Strukturen, die zu der neuen Struktur zusammengeführt werden.

Grenzen zwischen sich berührenden Polygonen entfernen. Aktivieren Sie diese Option, um die Grenzen aus den zusammengeführten Strukturen zu entfernen. Wenn Sie beispielsweise Bundesländer zu geografischen Regionen zusammengeführt haben, können Sie mit dieser Option die Grenzen zwischen den einzelnen Bundesländern entfernen.

Strukturen verschieben: Sie können Strukturen in der Karte verschieben. Dies kann nützlich sein, wenn Sie Strukturen zusammenbringen möchten, beispielsweise das Festlandsterritorium eines Landes und die zugehörigen abgelegenen Inseln.

Schlüssel. Wählen Sie den Kartenschlüssel aus, der die Strukturbeschriftungen enthält, mit denen Sie die zu verschiebenden Strukturen identifizieren können.

Strukturen. Klicken Sie auf die zu verschiebende Struktur. Beachten Sie, dass die Struktur in der Kartenvorschau ausgewählt wird. Sie können auch direkt in der Kartenvorschau auf die Struktur klicken.

X/Y. Diese Textfelder geben den aktuellen Mittelpunkt der Struktur in der Karte an. Die Einheiten werden in den Koordinaten der Karte angezeigt. Dabei kann es sich um lokale kartesische Koordinaten handeln (z. B. das State Plane Coordinate System für die USA) oder um geografische Koordinaten (wobei **X** den Längengrad und **Y** den Breitengrad angibt). Geben Sie die Koordinaten für die neue Position der Struktur an. Sie können Strukturen auch durch Klicken und Ziehen auf der Karte verschieben. Die Textfelder werden entsprechend der neuen Position aktualisiert.

Strukturen löschen: Sie können unerwünschte Strukturen aus der Karte löschen. Dies kann nützlich sein, wenn Sie die Karte überschaubarer gestalten möchten, indem Sie Strukturen löschen, die bei der Kartenvisualisierung nicht von Belang sind.

Schlüssel. Wählen Sie den Kartenschlüssel aus, der die Strukturbeschriftungen enthält, mit denen Sie die zu löschenden Strukturen identifizieren können.

Strukturen. Klicken Sie auf die zu löschende Struktur. Wenn Sie mehrere Strukturen gleichzeitig löschen möchten, klicken Sie bei gedrückter Steuertaste auf die weiteren Strukturen. Beachten Sie, dass die Strukturen auch in der Kartenvorschau ausgewählt werden. Neben der Auswahl aus der Liste haben Sie auch die Möglichkeit, direkt in der Kartenvorschau auf die Strukturen zu klicken und dann beim Klicken die Steuertaste gedrückt zu halten.

Einzelne Elemente löschen: Neben dem Löschen ganzer Strukturen können Sie einige der einzelnen Elemente löschen, aus denen sich die Strukturen zusammensetzen, beispielsweise Seen und kleine Inseln. Diese Option steht nicht für Punktkarten zur Verfügung.

Elemente. Klicken Sie auf die zu löschenden Elemente. Wenn Sie mehrere Elemente gleichzeitig löschen möchten, klicken Sie bei gedrückter Steuertaste auf die weiteren Elemente. Beachten Sie, dass die Elemente auch in der Kartenvorschau ausgewählt werden. Neben der Auswahl aus der Liste haben Sie auch die Möglichkeit, direkt in der Kartenvorschau auf die Elemente zu klicken und dann beim Klicken die Steuertaste gedrückt zu halten. Da die Liste der Elementnamen nicht selbsterklärend ist (den einzelnen Ele-

menten wird jeweils eine Nummer innerhalb der Struktur zugewiesen), sollten Sie die Auswahl in der Kartenvorschau überprüfen, um sich zu vergewissern, dass Sie die gewünschten Elemente ausgewählt haben.

Projektion festlegen:

Die Kartenprojektion gibt an, wie die dreidimensionale Erde in zwei Dimensionen dargestellt wird. Projektionen verursachen stets Verzerrungen. Allerdings sind, je nachdem, ob eine Weltkarte betrachtet wird oder eine regional begrenzte Karte, einige Projektionen besser geeignet als andere. Außerdem wird bei einigen Projektionen die Form der ursprünglichen Strukturen beibehalten. Projektionen, bei denen die Form beibehalten wird, sind konforme Projektionen. Diese Option steht nur für Karten mit geografischen Koordinaten (Längen- und Breitengrade) zur Verfügung.

Im Gegensatz zu anderen Optionen im Dienstprogramm zur Konvertierung von Karten kann die Projektion auch nach der Erstellung einer Kartenvisualisierung geändert werden.

Projektion. Wählen Sie eine Kartenprojektion aus. Bei der Erstellung einer Weltkarte oder einer Karte für eine Erdhalbkugel sollten Sie die Projektionstypen *Lokal*, *Mercator* oder *Winkel-Tripel* verwenden. Für kleinere Gebiete sollten Sie die Projektionstypen *Lokal*, *Lambert, konisch, konform* oder *Mercator, diagonal* verwenden. Bei allen Projektionen wird der WGS83-Ellipsoid als Bezugshöhe verwendet.

- Die Projektion vom Typ **Lokal** wird immer dann verwendet, wenn die Karte mit einem lokalen Koordinatensystem erstellt wurde, beispielsweise dem State Plane Coordinate System für die USA. Diese Koordinatensysteme werden statt durch geografische Koordinaten (Längen- und Breitengrade) durch kartesische Koordinaten definiert. Beim Projektionstyp "Lokal" befinden sich die horizontalen und vertikalen Linien in gleichmäßigen Abständen in einem kartesischen Koordinatensystem. Projektionen vom Typ "Lokal" sind nicht konform.
- Die Projektion vom Typ **Mercator** ist eine konforme Projektion für Weltkarten. Die horizontalen und vertikalen Linien sind gerade und stehen immer im rechten Winkel zueinander. Beachten Sie, dass sich die Mercator-Projektion ins Unendliche ausdehnt, wenn sie sich dem Nord- bzw. Südpol nähert. Daher kann sie nicht verwendet werden, wenn auf der Karte der Nord- bzw. Südpol enthalten ist. Die Verzerrung wird umso größer, je mehr sich die Karte diesen Grenzen nähert.
- Die Projektion vom Typ **Winkel-Tripel** ist eine nicht konforme Projektion für Weltkarten. Sie ist zwar nicht konform, bietet jedoch einen guten Ausgleich zwischen Form und Größe. Abgesehen von Äquator und Nullmeridian sind alle Linien gekrümmt. Wenn auf Ihrer Weltkarte der Nord- bzw. Südpol enthalten ist, ist dies eine gute Wahl für die Projektion.
- Wie der Name andeutet, handelt es sich beim Projektionstyp **Lambert, konisch, konform** (winkeltreue Kegelprojektion nach Lambert) um eine konforme Projektion. Diese wird für Karten von Kontinenten oder kleineren Landmassen verwendet, deren Ost-West-Ausdehnung größer ist als die Nord-Süd-Ausdehnung.
- Der Projektionstyp **Mercator, diagonal** ist eine weitere konforme Projektion für Karten von Kontinenten oder kleineren Landmassen. Diese Projektion eignet sich besonders für Landmassen, bei denen die Nord-Süd-Ausdehnung größer ist als die Ost-West-Ausdehnung.

Schritt 4 - Fertigstellen

Nun können Sie einen Kommentar hinzufügen, der die Kartendatei beschreibt, und eine Beispieldatendatei aus den Kartenschlüsseln erstellen.

Kartenschlüssel. Wenn die Kartendatei mehrere Schlüssel enthält, wählen Sie den Kartenschlüssel aus, dessen Strukturbeschriftungen in der Vorschau angezeigt werden soll. Wenn Sie eine Datendatei aus der Karte erstellen, werden diese Beschriftungen als Datenwerte verwendet.

Kommentar. Geben Sie einen Kommentar ein, der die Karte beschreibt oder zusätzliche Informationen angibt, die für Ihre Benutzer relevant sein könnten, beispielsweise die Quellen für die ursprünglichen Shapefiles. Der Kommentar wird im Verwaltungssystem der Auswahlfunktion für Diagrammtafelvorlagen angezeigt.

Aus Strukturbeschriftungen Dataset erstellen. Aktivieren Sie diese Option, wenn Sie eine Datei aus den angezeigten Strukturbeschriftungen erstellen möchten. Durch Klicken auf **Durchsuchen...** können Sie einen Speicherort und einen Dateinamen angeben. Wenn Sie die Erweiterung *.txt* hinzufügen, wird die Datei als Datei mit tabstoppgetrennten Werten gespeichert. Wenn Sie die Erweiterung *.csv* hinzufügen, wird die Datei als Datei mit kommagetrennten Werten gespeichert. Wenn Sie die Erweiterung *.sav* hinzufügen, wird die Datei im IBM SPSS Statistics-Format gespeichert. Wenn Sie keine Angabe machen, wird standardmäßig die Erweiterung SAV verwendet.

Verteilen der Kartendateien

Im ersten Schritt des Dienstprogramms zur Konvertierung von Karten haben Sie ein Verzeichnis zur Speicherung der konvertierten SMZ-Datei ausgewählt. Außerdem haben Sie möglicherweise ausgewählt, dass die Karte der Auswahlfunktion für Diagrammtafelvorlagen hinzugefügt werden soll. Wenn Sie sich für die Speicherung im Verwaltungssystem entschieden haben, steht Ihnen die Karte in jedem IBM SPSS-Produkt zur Verfügung, das auf demselben Computer ausgeführt wird.

Um die Karte an andere Benutzer zu verteilen, müssen Sie diesen die SMZ-Datei zukommen lassen. Die Benutzer können dann mit dem Verwaltungssystem die Karte importieren. Sie können einfach die Datei senden, deren Speicherort Sie in Schritt 1 angegeben haben. Wenn Sie eine Datei senden möchten, die sich im Verwaltungssystem befindet, müssen Sie sie zuerst exportieren:

1. Klicken Sie in der Vorlagenauswahl auf **Verwalten...**
2. Klicken Sie auf die Registerkarte "Karte".
3. Wählen Sie die Karte aus, die Sie verteilen möchten.
4. Klicken Sie auf **Exportieren...** und wählen Sie ein Verzeichnis aus, in dem die Datei gespeichert werden soll.

Nun können Sie die Kartendatei an andere Benutzer senden. Die Benutzer müssen diesen Vorgang umkehren und die Karte in das Verwaltungssystem importieren.

Plotknoten

Plotknoten zeigen die Beziehung zwischen numerischen Feldern. Sie können einen Plot mithilfe von Punkten (auch als Streudiagramm bezeichnet) oder mit Linien erstellen. Mit einem X-Modus im Dialogfeld stehen drei Arten von Liniendiagrammen zur Verfügung.

X-Modus = Sortieren

Beim X-Modus **Sortieren** werden die Daten nach den Werten für das Feld sortiert, das auf der *X*-Achse geplottet wird. So entsteht eine einzelne Linie, die von links nach rechts im Diagramm verläuft. Wenn Sie ein nominales Feld als Überlagerung verwenden, entstehen mehrere Linien mit verschiedenen Farbtönen, die von links nach rechts im Diagramm verlaufen.

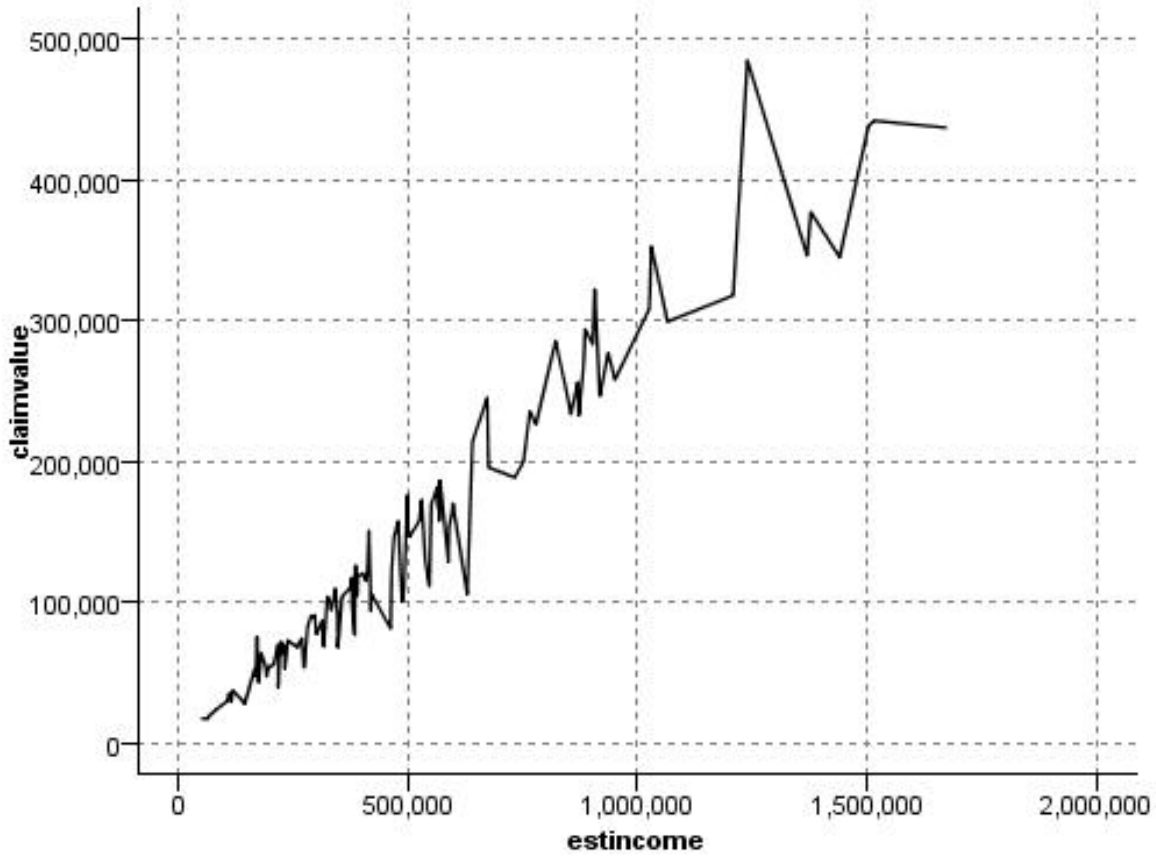


Abbildung 21. Liniendiagramm mit X-Modus "Sortieren"

X-Modus = Überlagern

Beim X-Modus **Überlagern** werden mehrere Liniendiagramme in einem einzigen Diagramm erstellt. Die Daten werden beim Überlagerungsplot nicht sortiert. Solange die Werte auf der *X*-Achse steigen, werden die Daten auf einer einzelnen Linie geplottet. Sobald die Werte fallen, beginnt eine neue Linie. Beispiel: Wenn *x* von 0 auf 100 steigt, werden die *y*-Werte auf einer einzelnen Linie aufgetragen. Sobald *x* unter 100 fällt, wird eine neue Linie zusätzlich zur ersten Linie geplottet. Der fertige Plot umfasst gegebenenfalls verschiedene Plots, mit denen mehrere Serien von *y*-Werten bequem miteinander verglichen werden können. Diese Art von Plot eignet sich für Daten mit einer Zeitraum-Komponente, z. B. für den Strombedarf über mehrere aufeinander folgende 24-Stunden-Zeiträume.

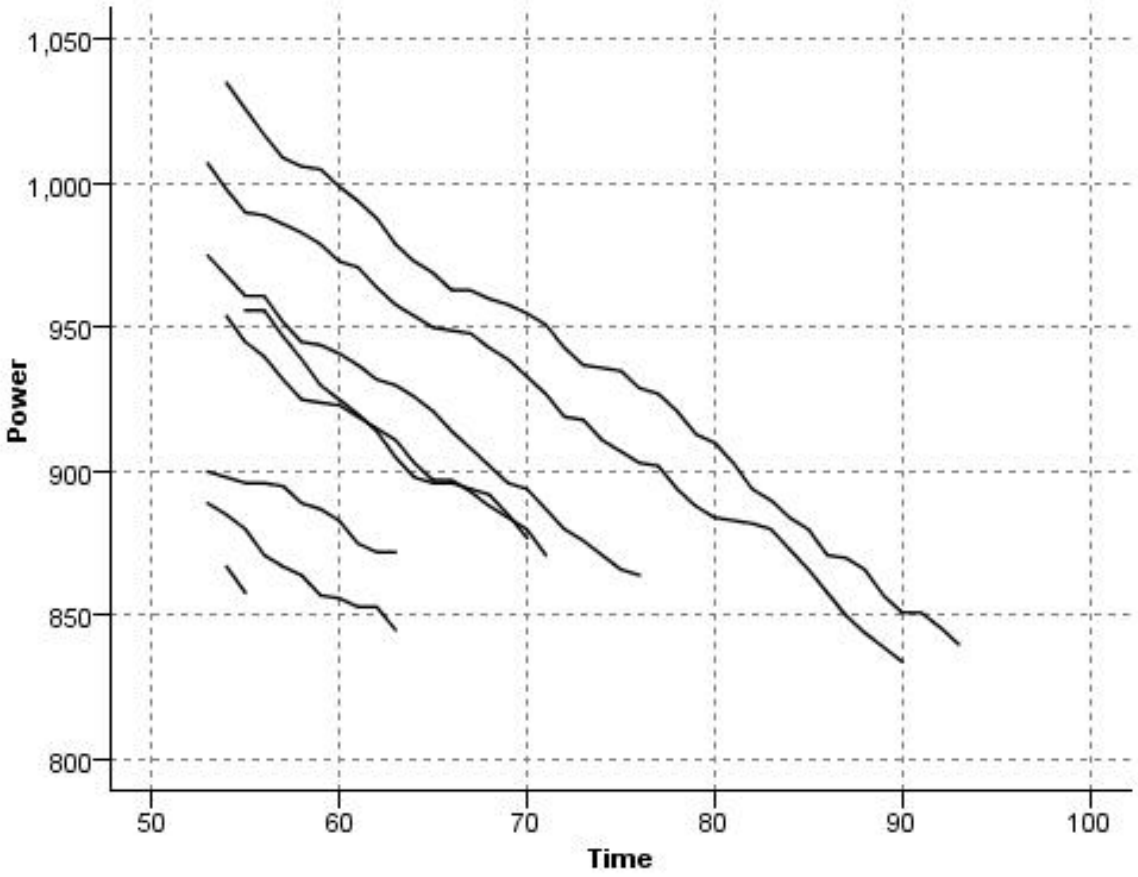


Abbildung 22. Liniendiagramm mit X-Modus "Überlagern"

X-Modus = Wie gelesen

Beim X-Modus **Wie gelesen** werden die *x*- und *y*-Werte so geplottet, wie sie aus der Datenquelle gelesen wurden. Diese Option eignet sich für Daten mit einer Zeitreihen-Komponente, bei der Sie sich für Trends oder Muster interessieren, die sich aus der Reihenfolge der Daten ergeben. Unter Umständen sollten Sie die Daten sortieren, bevor Sie diese Art von Plot erstellen. Außerdem ist es möglich, zwei ähnliche Plots mit dem X-Modus **Sortieren** und **Wie gelesen** zu vergleichen, um so zu ermitteln, inwieweit ein Muster von der Sortierung abhängig ist.

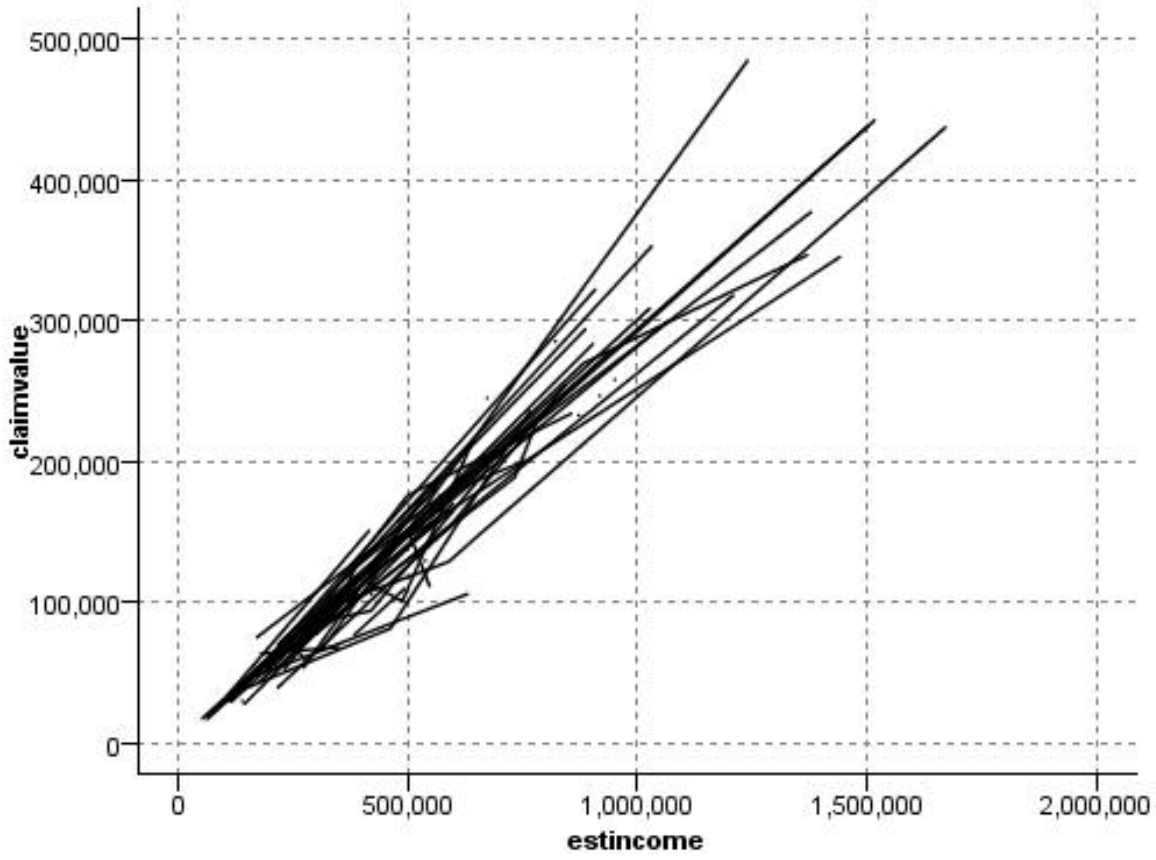


Abbildung 23. Liniendiagramm, oben mit X-Modus "Sortieren" dargestellt, nun erneut mit X-Modus "Wie gelesen" ausgeführt

Sie können auch den Diagrammtafelknoten zur Erstellung von Streudiagrammen und Liniendiagrammen verwenden. In diesem Knoten stehen jedoch mehr Optionen zur Auswahl. Weitere Informationen finden Sie im Thema „Verfügbare integrierte Visualisierungstypen für Diagrammtafeln" auf Seite 186.

Registerkarte des Plotknotens

Plots zeigen die Werte eines *Y*-Felds gegen die Werte eines *X*-Felds. Häufig entsprechen diese Felder einer abhängigen Variablen bzw. einer unabhängigen Variablen.

X-Feld. Wählen Sie ein Feld in der Liste aus, das auf der horizontalen *X*-Achse dargestellt werden soll.

Y-Feld. Wählen Sie ein Feld in der Liste aus, das auf der vertikalen *Y*-Achse dargestellt werden soll.

Z-Feld. Wenn Sie auf die Schaltfläche für das 3-D-Diagramm klicken, steht ein Feld in der Liste zur Auswahl, das auf der *Z*-Achse dargestellt werden kann.

Überlagerung. Die Kategorien für die Datenwerte können auf unterschiedliche Weise dargestellt werden. Verwenden Sie beispielsweise *maincrop* (Hauptfeldfrucht) als Farbüberlagerung, um so die Werte *estincome* (Geschätztes Einkommen) und *claimvalue* (Förderungswert) für die Hauptfeldfrucht darzustellen, die von den Anspruchstellern gezogen wird. Weitere Informationen finden Sie im Thema „Formatierungen, Überlagerungen, Fenster und Animation" auf Seite 176.

Überlagerungstyp. Gibt an, ob eine Überlagerungsfunktion oder ein Glättungselement angezeigt werden soll. Die Glättungs- und die Überlagerungsfunktion werden immer als Funktion von y berechnet.

- **Keine.** Es wird keine Überlagerung angezeigt.
- **Glättungselement.** Zeigt eine geglättete Anpassungslinie an, die mithilfe einer Regression mit lokal gewichteten iterativen robusten kleinsten Quadraten (LOESS) berechnet wurde. Bei dieser Methode wird im Grunde eine Reihe von Regressionen berechnet, wobei sich jede auf einen kleinen Bereich innerhalb des Plots konzentriert. Dies führt zu einer Reihe "lokaler" Regressionslinien, die anschließend zu einer glatten Kurve zusammengefügt werden.

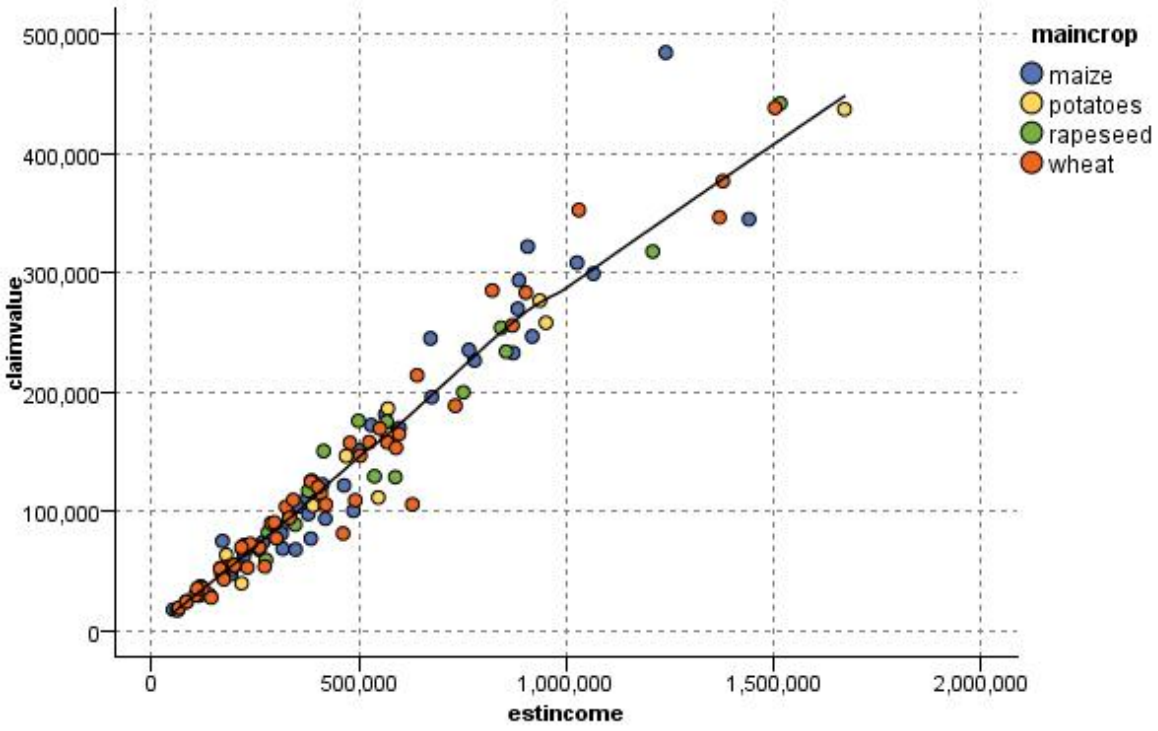


Abbildung 24. Plot mit LOESS-Smoother-Überlagerung

- **Funktion.** Hiermit geben Sie eine bekannte Funktion an, die mit tatsächlichen Werten verglichen werden soll. Um beispielsweise die Istwerte mit den vorhergesagten Werten zu vergleichen, plotten Sie die Funktion y = x als Überlagerung. Geben Sie eine Funktion für **y =** im Textfeld an. Die Standardfunktion lautet y = x; Sie können jedoch auch andere Funktionen festlegen, z. B. quadratische Funktionen oder beliebige Ausdrücke im Hinblick auf x.

  Hinweis: Überlagerungsfunktionen sind für Fenster- und Animationsdiagramme nicht verfügbar.

Sobald Sie die Optionen für einen Plot festgelegt haben, können Sie den Plot direkt aus dem Dialogfeld heraus starten. Klicken Sie hierzu auf **Ausführen**. Auf der Registerkarte "Optionen" können Sie jedoch zusätzliche Optionen angeben, z. B. Klassieren, X-Modus oder Stil.

Plot - Registerkarte "Optionen"

Stil. Wählen Sie **Punkt** oder **Linie** als Plotstil aus. Bei Auswahl von **Linie** wird das Steuerelement **X-Modus** aktiviert. Bei Auswahl von **Punkt** wird ein Pluszeichen (+) als Standardpunktform verwendet. Nach der Erstellung des Diagramms können Sie die Punktform und ihre Größe ändern.

X-Modus. Bei Liniendiagrammen definieren Sie den Stil mithilfe eines X-Modus. Wählen Sie **Sortieren**, **Überlagern** oder **Wie gelesen** aus. Bei **Überlagern** und **Wie gelesen** sollten Sie eine maximal zulässige Datasetgröße angeben, mit der die ersten *n* Datensätze in eine Stichprobe aufgenommen werden. Ansonsten werden die standardmäßig festgelegten 2.000 Datensätze verwendet.

Automatischer X-Bereich. Hiermit geben Sie an, dass der gesamte Wertebereich in den Daten entlang dieser Achse verwendet werden soll. Um nur ein explizites Subset von Werten auf der Grundlage der angegebenen Werte für **Min** und **Max** zu verwenden, inaktivieren Sie diese Option. Geben Sie die gewünschten Werte ein oder stellen Sie sie mit den Pfeilen ein. Automatische Bereiche sind standardmäßig aktiviert, um so den raschen Aufbau der Diagramme zu gewährleisten.

Automatischer Y-Bereich. Hiermit geben Sie an, dass der gesamte Wertebereich in den Daten entlang dieser Achse verwendet werden soll. Um nur ein explizites Subset von Werten auf der Grundlage der angegebenen Werte für **Min** und **Max** zu verwenden, inaktivieren Sie diese Option. Geben Sie die gewünschten Werte ein oder stellen Sie sie mit den Pfeilen ein. Automatische Bereiche sind standardmäßig aktiviert, um so den raschen Aufbau der Diagramme zu gewährleisten.

Automatischer Z-Bereich. Nur bei Angabe eines 3-D-Diagramms auf der Registerkarte "Plot". Hiermit geben Sie an, dass der gesamte Wertebereich in den Daten entlang dieser Achse verwendet werden soll. Um nur ein explizites Subset von Werten auf der Grundlage der angegebenen Werte für **Min** und **Max** zu verwenden, inaktivieren Sie diese Option. Geben Sie die gewünschten Werte ein oder stellen Sie sie mit den Pfeilen ein. Automatische Bereiche sind standardmäßig aktiviert, um so den raschen Aufbau der Diagramme zu gewährleisten.

Fluktuation. Auch als **Bewegung** bezeichnet. Die Fluktuation eignet sich für Punktplots von Datasets, in denen sich zahlreiche Werte wiederholen. Um eine deutlichere Verteilung der Werte zu erzielen, können Sie mit der Fluktuation die Punkte zufällig um den tatsächlichen Wert herum verteilen.

Hinweis an Benutzer früherer Versionen von IBM SPSS Modeler: In diesem Release von IBM SPSS Modeler wird für den in einem Plot verwendeten Fluktuationswert eine andere Metrik verwendet. In früheren Versionen bestand der Wert aus einer tatsächlichen Zahl; nun wird ein Teil der Rahmengröße herangezogen. Dies bedeutet, dass die Bewegungswerte aus alten Streams wahrscheinlich zu groß sind. Bei dieser Version werden alle Bewegungswerte ungleich null durch den Wert 0,2 ersetzt.

Maximale Anzahl der Datensätze für Plot. Geben Sie eine Methode für das Plotten umfangreicher Datasets an. Sie können wahlweise eine maximal zulässige Größe für das Dataset angeben oder den Standardwert von 2.000 Datensätzen verwenden. Bei umfangreichen Datasets steigt die Leistung, wenn Sie die Option **Klasse** oder **Stichprobe** aktivieren. Alternativ können Sie mit **Alle Daten verwenden** alle Datenpunkte gleichzeitig plotten lassen; dies kann sich jedoch beträchtlich auf die Leistung der Software auswirken.

Hinweis: Beim X-Modus **Überlagern** oder **Wie gelesen** sind diese Optionen inaktiviert und es werden nur die ersten *n* Datensätze verwendet.

- **Klasse.** Hiermit aktivieren Sie die Klassierung, wenn das Dataset mehr Datensätze enthält als die angegebene Anzahl. Bei der Klassierung wird das Diagramm vor dem eigentlichen Plotten in feinmaschige Raster aufgeteilt und es wird die Anzahl der Punkte gezählt, die in die einzelnen Rasterzellen fallen würden. Im endgültigen Diagramm wird je ein Punkt pro Zelle im Klassierschwerpunkt (Durchschnitt aller Punktpositionen in der Klasse) geplottet. Die Größe der geplotteten Symbole weist auf die Anzahl der Punkte im betreffenden Bereich hin (sofern Sie die Größe nicht als Überlagerung verwenden). Durch die Methode, den Schwerpunkt und die Größe zur Darstellung der Anzahl an Punkten heranzuziehen, eignet sich der klassierte Plot besonders gut für die Darstellung umfangreicher Datasets, weil ein übermäßiges Plotten (unidentifizierbare Farbansammlungen) in dicht besetzten Bereichen vermieden wird und auch Symbolartefakte (künstliche Dichtemuster) verringert werden. Symbolartefakte treten auf, wenn bestimmte Symbole (insbesondere das Pluszeichen +) auf eine Weise kollidieren, bei der dichte Bereiche entstehen, die in den eigentlichen Rohdaten nicht vorhanden sind.

- **Beispiel.** Aus den Daten wird eine zufällige Stichprobe mit der im Textfeld eingegebenen Anzahl an Datensätzen zusammengestellt. Der Standardwert ist 2.000.

Plot - Registerkarte "Darstellung"

Vor der Diagrammerstellung können Sie Darstellungsoptionen angeben.

Titel. Dient zur Eingabe des Texts, der als Titel des Diagramms verwendet werden soll.

Untertitel. Dient zur Eingabe des Texts, der als Untertitel des Diagramms verwendet werden soll.

Titelzeile. Dient zur Eingabe des Texts, der als Titelzeile des Diagramms verwendet werden soll.

X-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *X*-Achsenbeschriftung (horizontal) oder wählen Sie **Angepasst** aus, um eine Beschriftung anzugeben.

Y-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *Y*-Achsenbeschriftung (vertikal) oder wählen Sie **Angepasst** aus, um eine Beschriftung anzugeben.

Z-Beschriftung. Nur bei 3-D-Diagrammen: Akzeptieren Sie entweder die automatisch generierte *Z*-Achsenbeschriftung oder wählen Sie **Angepasst** aus, um eine benutzerdefinierte Beschriftung anzugeben.

Rasterlinie anzeigen. Diese Option ist standardmäßig aktiviert. Hiermit lassen Sie Rasterlinien hinter dem Plot oder dem Diagramm einblenden, was die Bestimmung der Bereichs- und Bandabschnittpunkte erleichtert. Rasterlinien werden stets in weißer Farbe angezeigt; bei einem weißen Diagrammhintergrund erfolgt die Anzeige in Grau.

Verwenden eines Plotdiagramms

Plots und Multiplots sind im Grunde genommen Plots von *X* in Abhängigkeit von *Y*. Wenn Sie beispielsweise potenzielle Betrugsfälle in Bewerbungen um landwirtschaftliche Subventionen untersuchen (wie in *fraud.str* im Ordner *Demos* der IBM SPSS Modeler-Installation dargestellt), soll beispielsweise das in der Bewerbung angegebene Einkommen in Abhängigkeit von dem Einkommen geplottet werden, das mithilfe eines neuronalen Netzes geschätzt wurde. Aus einer Überlagerung, z. B. dem Feldfruchttyp, geht hervor, ob eine Beziehung zwischen den Forderungen (Wert oder Anzahl) und der Art der Feldfrucht besteht.

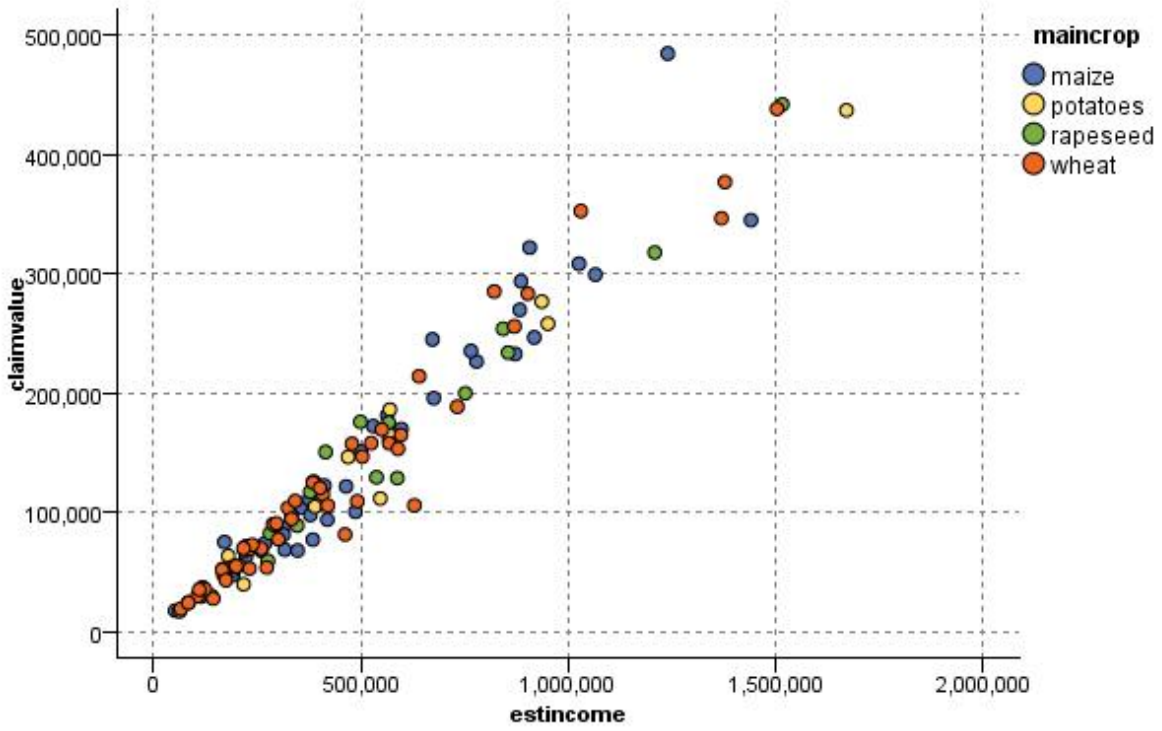


Abbildung 25. Plot der Beziehung zwischen geschätztem Einkommen und Forderungswert mit Hauptfeldfruchttyp als Überlagerung

Plots, Multiplots und Evaluierungsdiagramme sind zweidimensionale Darstellungen von *Y* gegen *X*. Die Arbeit mit diesen Diagrammen ist daher denkbar unkompliziert: Sie können ganz einfach Bereiche definieren, Elemente markieren und sogar Abschnitte einzeichnen. Außerdem können Sie Knoten für die durch diese Bereiche, Abschnitte bzw. Elemente dargestellten Daten generieren. Weitere Informationen finden Sie im Thema „Exploration von Diagrammen” auf Seite 254.

Verteilungsknoten

Verteilungsdiagramme bzw. -tabellen zeigen das Auftreten symbolischer (nicht numerischer) Werte, z. B. Hypothekenart oder Geschlecht, in einem Dataset. Verteilungsknoten eignen sich beispielsweise zum Aufzeigen von Unausgewogenheiten in den Daten, die mithilfe eines Balancierungsknotens vor dem Erstellen eines Modells ausgeglichen werden können. Über das Menü "Generieren" im Fenster eines Verteilungsdiagramms bzw. einer Verteilungstabelle können Sie automatisch einen Balancierungsknoten generieren lassen.

Sie können auch den Diagrammtafelknoten zur Erstellung von Diagrammen vom Typ "Balken für Häufigkeiten" verwenden. In diesem Knoten stehen jedoch mehr Optionen zur Auswahl. Weitere Informationen finden Sie im Thema „Verfügbare integrierte Visualisierungstypen für Diagrammtafeln” auf Seite 186.

Hinweis: Sie sollten einen Histogrammknoten verwenden, um das Vorkommen von numerischen Werten anzuzeigen.

Verteilung - Registerkarte "Plot"

Diagramm. Wählen Sie den Typ der Verteilung aus. Mit **Ausgewählte Felder** lassen Sie die Verteilung für das ausgewählte Feld anzeigen. Mit **Alle Flags (wahre Werte)** rufen Sie die Verteilung der Wahr-Werte für die Flagfelder im Dataset ab.

Feld. Wählen Sie ein nominales Feld oder ein Flagfeld aus, für das die Verteilung der Werte dargestellt werden soll. Die Liste enthält nur Felder, die nicht explizit als numerisch definiert wurden.

Überlagerung. Wählen Sie ein nominales Feld oder ein Flagfeld aus, das als Farbüberlagerung verwendet werden soll, um so die Verteilung der zugehörigen Werte innerhalb der einzelnen Werte für das angegebene Feld darzustellen. Mithilfe der Reaktionen auf eine Marketingkampagne (*pep*) als Überlagerung für die Anzahl der Kinder (*children*) können Sie beispielsweise die Ansprechbarkeit nach Familiengröße darstellen lassen. Weitere Informationen finden Sie im Thema „Formatierungen, Überlagerungen, Fenster und Animation" auf Seite 176.

Nach Farbe normalisieren. Die Balken werden so skaliert, dass alle Balken die volle Breite des Diagramms einnehmen. Die Überlagerungswerte entsprechen einem Anteil jedes Balkens, sodass Vergleiche zwischen den Kategorien erleichtert werden.

Sortieren. Wählen Sie die Methode aus, mit der die Werte im Verteilungsdiagramm dargestellt werden sollen. Mit der Option **Alphabetisch** werden die Werte in alphabetischer Reihenfolge angezeigt, mit **Nach Anzahl** dagegen absteigend nach der Anzahl der Vorkommen.

Anteilsskala. Die Verteilung der Werte wird so skaliert, dass der Wert mit der größten Anzahl die volle Breite des Plots einnimmt. Alle anderen Balken werden gemäß diesem Wert skaliert. Wenn Sie diese Option inaktivieren, werden die Balken gemäß der Gesamtanzahl der einzelnen Werte skaliert.

Verteilung - Registerkarte "Darstellung"

Vor der Diagrammerstellung können Sie Darstellungsoptionen angeben.

Titel. Dient zur Eingabe des Texts, der als Titel des Diagramms verwendet werden soll.

Untertitel. Dient zur Eingabe des Texts, der als Untertitel des Diagramms verwendet werden soll.

Titelzeile. Dient zur Eingabe des Texts, der als Titelzeile des Diagramms verwendet werden soll.

X-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *X*-Achsenbeschriftung (horizontal) oder wählen Sie **Angepasst** aus, um eine Beschriftung anzugeben.

Y-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *Y*-Achsenbeschriftung (vertikal) oder wählen Sie **Angepasst** aus, um eine Beschriftung anzugeben.

Rasterlinie anzeigen. Diese Option ist standardmäßig aktiviert. Hiermit lassen Sie Rasterlinien hinter dem Plot oder dem Diagramm einblenden, was die Bestimmung der Bereichs- und Bandabschnittpunkte erleichtert. Rasterlinien werden stets in weißer Farbe angezeigt; bei einem weißen Diagrammhintergrund erfolgt die Anzeige in Grau.

Verwendung von Verteilungsknoten

Verteilungsknoten zeigen die Verteilung symbolischer Werte in einem Dataset. Diese Knoten werden häufig als Vorstufe für Bearbeitungsknoten eingesetzt, um die Daten zu untersuchen und eventuelle Unausgewogenheiten zu bereinigen. Wenn beispielsweise Instanzen mit Antwortenden ohne Kinder viel häufiger auftreten als andere Typen von Teilnehmern, können Sie diese Instanzen verringern, sodass in

späteren Data-Mining-Operationen eine nützlichere Regel aufgestellt werden kann. Mit einem Verteilungsknoten können Sie diese Unausgewogenheiten untersuchen und über die weitere Vorgehensweise entscheiden.

Der Verteilungsknoten ist dahingehend ungewöhnlich, dass er sowohl ein Diagramm als auch eine Tabelle zur Datenanalyse erstellt.

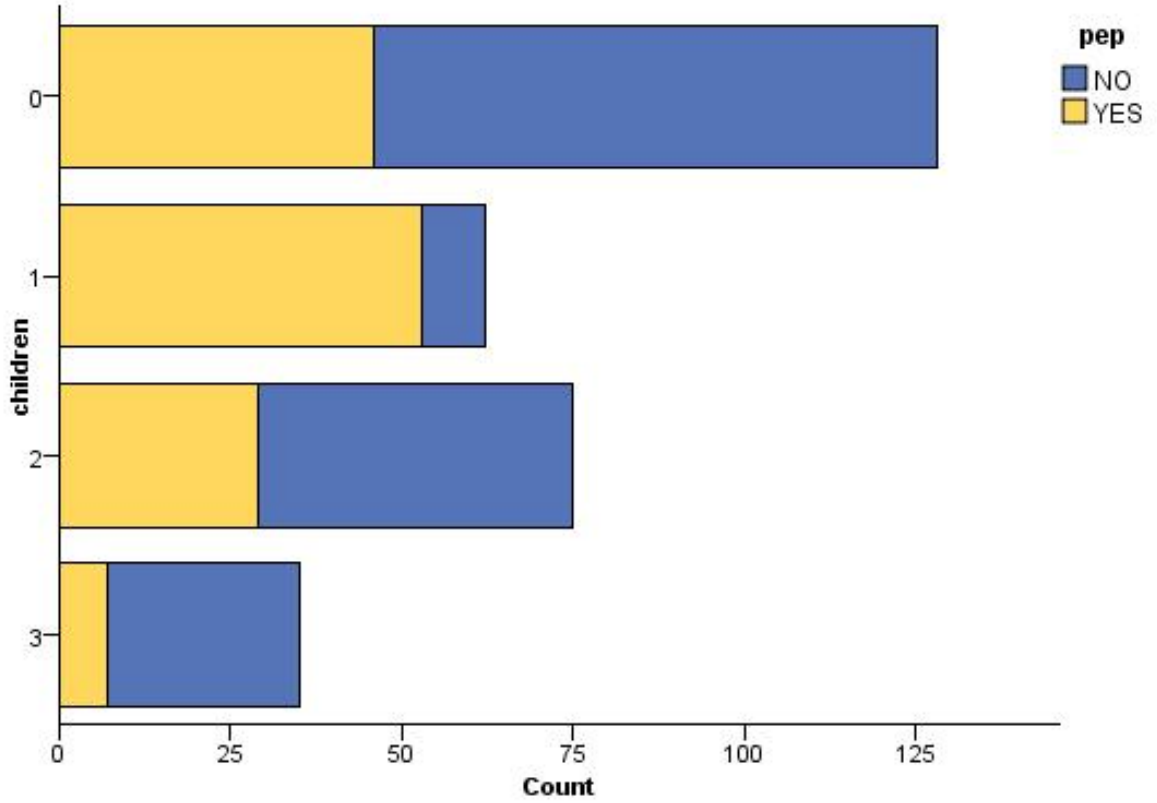


Abbildung 26. Verteilungsdiagramm zur Anzahl der Teilnehmer mit Kindern bzw. ohne Kinder, die auf eine Marketing-Kampagne reagiert haben

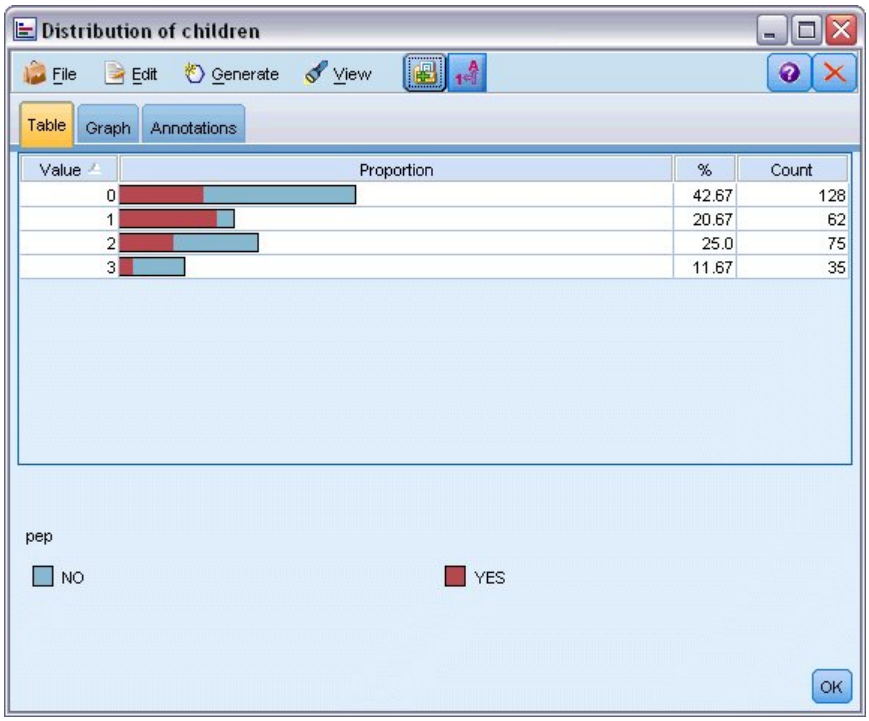


Abbildung 27. Verteilungstabelle zum Anteil der Teilnehmer mit Kindern bzw. ohne Kinder, die auf eine Marketing-Kampagne reagiert haben

Sobald Sie eine Verteilungstabelle und ein Verteilungsdiagramm erstellt und die Ergebnisse untersucht haben, können Sie mit den Optionen in den Menüs die Werte gruppieren, bestimmte Werte kopieren und eine Reihe von Knoten zur Datenvorbereitung erzeugen. Außerdem können Sie die Diagramm- und Tabelleninformationen zur Verwendung in anderen Anwendungen, wie beispielsweise MS Word oder MS PowerPoint, kopieren bzw. exportieren. Weitere Informationen finden Sie im Thema „Drucken, Speichern, Kopieren und Exportieren von Diagrammen” auf Seite 279.

So können Sie Werte in einer Verteilungstabelle auswählen und kopieren:

1. Klicken Sie mit der Maus, halten Sie die Maustaste gedrückt und wählen Sie eine Reihe von Werten durch Ziehen aus. Sie können auch mit dem Befehl **Alles auswählen** im Menü "Bearbeiten" alle Werte gleichzeitig auswählen.
2. Wählen Sie im Menü "Bearbeiten" die Option **Tabelle kopieren** oder **Tabelle kopieren (einschl. Feldnamen)**.
3. Übernehmen Sie die Daten in die Zwischenablage oder fügen Sie sie in die gewünschte Anwendung ein.

   Hinweis: Die Balken werden nicht direkt kopiert. Stattdessen werden die Tabellenwerte kopiert. In der kopierten Tabelle werden also keine überlagerten Werte dargestellt.

So gruppieren Sie Werte aus einer Verteilungstabelle:

1. Wählen Sie mehrere Werte durch Klicken bei gedrückter Steuertaste zur Gruppierung aus.
2. Wählen Sie im Menü "Bearbeiten" die Option **Gruppieren**.

Hinweis: Beim Gruppieren von Werten bzw. beim Aufheben der Gruppierung wird das Diagramm auf der Registerkarte "Diagramm" automatisch unter Berücksichtigung der Änderungen neu erstellt.

Weitere Möglichkeiten:

- Heben Sie die Gruppierung der Werte auf. Wählen Sie hierzu den Namen der Gruppe in der Verteilungsliste aus und wählen Sie im Menü "Bearbeiten" den Befehl **Gruppierung aufheben**.

- Bearbeiten Sie die Gruppen. Wählen Sie hierzu den Namen der Gruppe in der Verteilungsliste aus und wählen Sie im Menü "Bearbeiten" den Befehl **Gruppe bearbeiten**. Ein Dialogfeld wird geöffnet, in dem Sie die Werte in die Gruppe hinein und aus dieser hinaus verschieben können.

Optionen im Menü "Generieren"

Mit den Optionen im Menü "Generieren" können Sie ein Subset mit Daten auswählen, ein Flagfeld ableiten oder die Daten aus einem Diagramm bzw. einer Tabelle balancieren. Bei diesen Funktionen wird ein Datenvorbereitungsknoten erzeugt und in den Streamerstellungsbereich platziert. Um den erzeugten Knoten nutzen zu können, verbinden Sie ihn mit einem vorhandenen Stream. Weitere Informationen finden Sie im Thema „Generieren von Knoten aus Diagrammen” auf Seite 261.

Histogrammknoten

Histogrammknoten zeigen das Auftreten bestimmter Werte in numerischen Feldern. Hiermit werden häufig die Daten vor der weiteren Bearbeitung und der Modellerstellung untersucht. Ähnlich wie Verteilungsknoten werden Histogrammknoten oft dazu herangezogen, Unausgewogenheiten in den Daten zu erkennen. Sie können Histogramme zwar auch mit dem Diagrammtafelknoten erstellen, in diesem Knoten stehen Ihnen jedoch mehr Optionen zur Auswahl. Weitere Informationen finden Sie im Thema „Verfügbare integrierte Visualisierungstypen für Diagrammtafeln” auf Seite 186.

Hinweis: Soll das Auftreten von Werten für symbolische Felder aufgezeigt werden, verwenden Sie einen Verteilungsknoten.

Histogramm - Registerkarte "Plot"

Feld. Wählen Sie ein numerisches Feld aus, für das die Verteilung der Werte dargestellt werden soll. Die Liste enthält nur solche Felder, die nicht explizit als symbolisch (kategorial) definiert wurden.

Überlagerung. Wählen Sie ein symbolisches Feld aus, mit dem die Kategorien der Werte für das angegebene Feld dargestellt werden sollen. Bei einem Überlagerungsfeld wird das Histogramm in ein Stapeldiagramm umgewandelt, bei dem die verschiedenen Kategorien des Überlagerungsfelds mithilfe von Farben gekennzeichnet sind. Bei Verwendung des Histogrammknotens gibt es drei Überlagerungstypen: Farbe, Fenster und Animation. Weitere Informationen finden Sie im Thema „Formatierungen, Überlagerungen, Fenster und Animation” auf Seite 176.

Histogramm - Registerkarte "Optionen"

Automatischer X-Bereich. Hiermit geben Sie an, dass der gesamte Wertebereich in den Daten entlang dieser Achse verwendet werden soll. Um nur ein explizites Subset von Werten auf der Grundlage der angegebenen Werte für **Min** und **Max** zu verwenden, inaktivieren Sie diese Option. Geben Sie die gewünschten Werte ein oder stellen Sie sie mit den Pfeilen ein. Automatische Bereiche sind standardmäßig aktiviert, um so den raschen Aufbau der Diagramme zu gewährleisten.

Klassen. Wählen Sie entweder **Nach Anzahl** oder **Nach Breite**.

- Mit der Option **Nach Anzahl** lassen Sie eine feste Anzahl von Balken anzeigen, deren Breite vom angegebenen Bereich und der angegebenen Anzahl an Buckets abhängig ist. Geben Sie in der Option **Anzahl der Klassen** an, wie viele Klassen im Diagramm verwendet werden sollen. Mithilfe der Pfeile können Sie die Anzahl einstellen.
- Mit **Nach Breite** erstellen Sie ein Diagramm, dessen Balken eine feste Breite besitzen. Die Anzahl der Klassen ergibt sich aus der festgelegten Breite und dem Wertebereich. Geben Sie in der Option **Klassenbreite** die Breite der Balken an.

Nach Farbe normalisieren. Alle Balken werden auf dieselbe Höhe gebracht. Überlagerte Werte werden dabei als Prozentsatz der Gesamtanzahl an Fällen in jedem Balken dargestellt.

Normalverteilungskurve anzeigen. Wählen Sie diese Option aus, um eine Normalverteilungskurve in das Diagramm aufzunehmen, die Mittelwert und Varianz der Daten anzeigt.

Getrennte Abschnitte für jede Farbe. Jeder überlagerte Wert wird als getrennter Abschnitt im Diagramm dargestellt.

Histogramm - Registerkarte "Darstellung"

Vor der Diagrammerstellung können Sie Darstellungsoptionen angeben.

Titel. Dient zur Eingabe des Texts, der als Titel des Diagramms verwendet werden soll.

Untertitel. Dient zur Eingabe des Texts, der als Untertitel des Diagramms verwendet werden soll.

Titelzeile. Dient zur Eingabe des Texts, der als Titelzeile des Diagramms verwendet werden soll.

X-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *X*-Achsenbeschriftung (horizontal) oder wählen Sie **Angepasst** aus, um eine Beschriftung anzugeben.

Y-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *Y*-Achsenbeschriftung (vertikal) oder wählen Sie **Angepasst** aus, um eine Beschriftung anzugeben.

Rasterlinie anzeigen. Diese Option ist standardmäßig aktiviert. Hiermit lassen Sie Rasterlinien hinter dem Plot oder dem Diagramm einblenden, was die Bestimmung der Bereichs- und Bandabschnittpunkte erleichtert. Rasterlinien werden stets in weißer Farbe angezeigt; bei einem weißen Diagrammhintergrund erfolgt die Anzeige in Grau.

Histogramme

Histogramme zeigen die Verteilung der Werte in einem numerischen Feld, dessen Werte an der *X*-Achse dargestellt werden. Histogramme funktionieren ähnlich wie Sammlungsdiagramme. Bei Sammlungen wird die Verteilung der Werte für ein numerisches Feld *relativ zu den Werten eines anderen Felds* dargestellt, also nicht das Auftreten von Werten für ein einziges Feld.

Sobald Sie ein Diagramm erstellt haben, können Sie die Ergebnisse untersuchen und Abschnitte festlegen, um die Werte entlang der *X*-Achse aufzuspalten bzw. Regionen zu definieren. Außerdem können Sie Elemente innerhalb des Diagramms markieren. Weitere Informationen finden Sie im Thema „Exploration von Diagrammen” auf Seite 254.

Mit Optionen im Menü "Generieren" können Sie Balancierungs- Auswahl- und Ableitungsknoten erstellen. Hierfür werden die Daten im Diagramm bzw. genauer die Daten innerhalb bestimmter Abschnitte, Bereiche oder markierter Elemente verwendet. Dieser Diagrammtyp wird häufig als Vorbereitung auf Bearbeitungsknoten eingesetzt, um die Daten zu untersuchen und etwaige Unausgewogenheiten mithilfe eines Balancierungsknotens, der aus dem Diagrammknoten heraus erzeugt wird, auszugleichen. Darüber hinaus können Sie einen Flagableitungsknoten erzeugen und so ein Feld hinzufügen, aus dem hervorgeht, in welchen Abschnitt die einzelnen Datensätze fallen, oder auch einen Auswahlknoten, mit dem Sie alle Datensätze in einem bestimmten Set oder Wertebereich auswählen. Diese Funktionen sorgen dafür, dass ein bestimmtes Subset an Daten zur näheren Exploration im Mittelpunkt verbleibt. Weitere Informationen finden Sie im Thema „Generieren von Knoten aus Diagrammen” auf Seite 261.

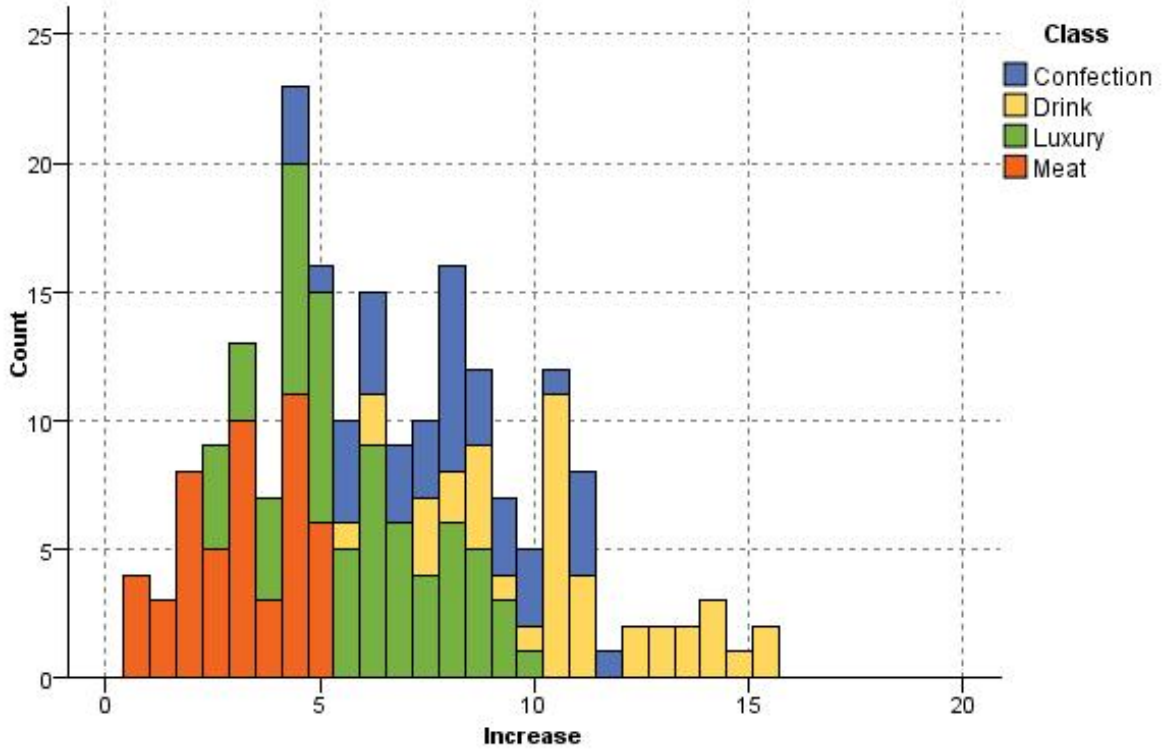


Abbildung 28. Histogramm mit der Verteilung gesteigerter Käufe nach Kategorie aufgrund einer Werbeaktion

Sammlungsknoten

Sammlungen sind nahezu mit Histogrammen identisch, mit dem Unterschied, dass bei Sammlungen die Verteilung der Werte für ein numerisches Feld relativ zu den Werten eines anderen Felds dargestellt wird, also nicht das Auftreten von Werten für ein einziges Feld. Eine Sammlung eignet sich besonders für die Darstellung einer Variablen oder eines Felds, dessen Werte sich mit der Zeit verändern. Mithilfe eines 3-D-Diagramms können Sie außerdem eine symbolische Achse anlegen, auf der die Verteilungen nach Kategorie aufgetragen sind. Zweidimensionale Sammlungen werden als gestapelte Balkendiagramme (gegebenenfalls mit Überlagerungen) angezeigt. Weitere Informationen finden Sie im Thema „Formatierungen, Überlagerungen, Fenster und Animation" auf Seite 176.

Sammlung - Registerkarte "Plot"

Sammeln. Wählen Sie ein Feld aus, dessen Werte gesammelt und über den Wertebereich für das unter **Über** angegebene Feld dargestellt werden sollen. Die Liste enthält nur solche Felder, die nicht als symbolisch definiert sind.

Über. Wählen Sie ein Feld aus, dessen Werte für die Darstellung des unter **Sammeln** angegebenen Felds herangezogen werden sollen.

Nach. Mit dieser Option (beim Erstellen eines 3-D-Diagramms aktiviert) können Sie ein nominales Feld oder ein Flagfeld auswählen, mit dem das Sammlungsfeld nach Kategorien dargestellt werden soll.

Funktion. Legen Sie fest, was die einzelnen Balken im Sammlungsdiagramm enthalten sollen. Die folgenden Optionen stehen zur Auswahl: **Summe**, **Mittelwert**, **Max**, **Min** und **Standardabweichung**.

Überlagerung. Wählen Sie ein symbolisches Feld aus, mit dem die Kategorien der Werte für das ausgewählte Feld dargestellt werden sollen. Wenn Sie ein Überlagerungsfeld auswählen, wird die Sammlung umgewandelt und es entstehen mehrere Balken in verschiedenen Farben für die einzelnen Kategorien. Für diesen Knoten gibt es drei Überlagerungstypen: Farbe, Fenster und Animation. Weitere Informationen finden Sie im Thema „Formatierungen, Überlagerungen, Fenster und Animation" auf Seite 176.

Sammlung - Registerkarte "Optionen"

Automatischer X-Bereich. Hiermit geben Sie an, dass der gesamte Wertebereich in den Daten entlang dieser Achse verwendet werden soll. Um nur ein explizites Subset von Werten auf der Grundlage der angegebenen Werte für **Min** und **Max** zu verwenden, inaktivieren Sie diese Option. Geben Sie die gewünschten Werte ein oder stellen Sie sie mit den Pfeilen ein. Automatische Bereiche sind standardmäßig aktiviert, um so den raschen Aufbau der Diagramme zu gewährleisten.

Klassen. Wählen Sie entweder **Nach Anzahl** oder **Nach Breite**.

- Mit der Option **Nach Anzahl** lassen Sie eine feste Anzahl von Balken anzeigen, deren Breite vom angegebenen Bereich und der angegebenen Anzahl an Buckets abhängig ist. Geben Sie in der Option **Anzahl der Klassen** an, wie viele Klassen im Diagramm verwendet werden sollen. Mithilfe der Pfeile können Sie die Anzahl einstellen.
- Mit **Nach Breite** erstellen Sie ein Diagramm, dessen Balken eine feste Breite besitzen. Die Anzahl der Klassen ergibt sich aus der festgelegten Breite und dem Wertebereich. Geben Sie in der Option **Klassenbreite** die Breite der Balken an.

Sammlung - Registerkarte "Darstellung"

Vor der Diagrammerstellung können Sie Darstellungsoptionen angeben.

Titel. Dient zur Eingabe des Texts, der als Titel des Diagramms verwendet werden soll.

Untertitel. Dient zur Eingabe des Texts, der als Untertitel des Diagramms verwendet werden soll.

Titelzeile. Dient zur Eingabe des Texts, der als Titelzeile des Diagramms verwendet werden soll.

Beschriftung "Über". Akzeptieren Sie entweder die automatisch generierte Beschriftung oder wählen Sie **Angepasst** aus, um eine benutzerdefinierte Beschriftung anzugeben.

Beschriftung "Sammeln". Akzeptieren Sie entweder die automatisch generierte Beschriftung oder wählen Sie **Angepasst** aus, um eine benutzerdefinierte Beschriftung anzugeben.

Beschriftung "Nach". Akzeptieren Sie entweder die automatisch generierte Beschriftung oder wählen Sie **Angepasst** aus, um eine benutzerdefinierte Beschriftung anzugeben.

Rasterlinie anzeigen. Diese Option ist standardmäßig aktiviert. Hiermit lassen Sie Rasterlinien hinter dem Plot oder dem Diagramm einblenden, was die Bestimmung der Bereichs- und Bandabschnittpunkte erleichtert. Rasterlinien werden stets in weißer Farbe angezeigt; bei einem weißen Diagrammhintergrund erfolgt die Anzeige in Grau.

Das folgende Beispiel zeigt, wo die Darstellungsoptionen bei einer 3-D-Version des Diagramms platziert sind.

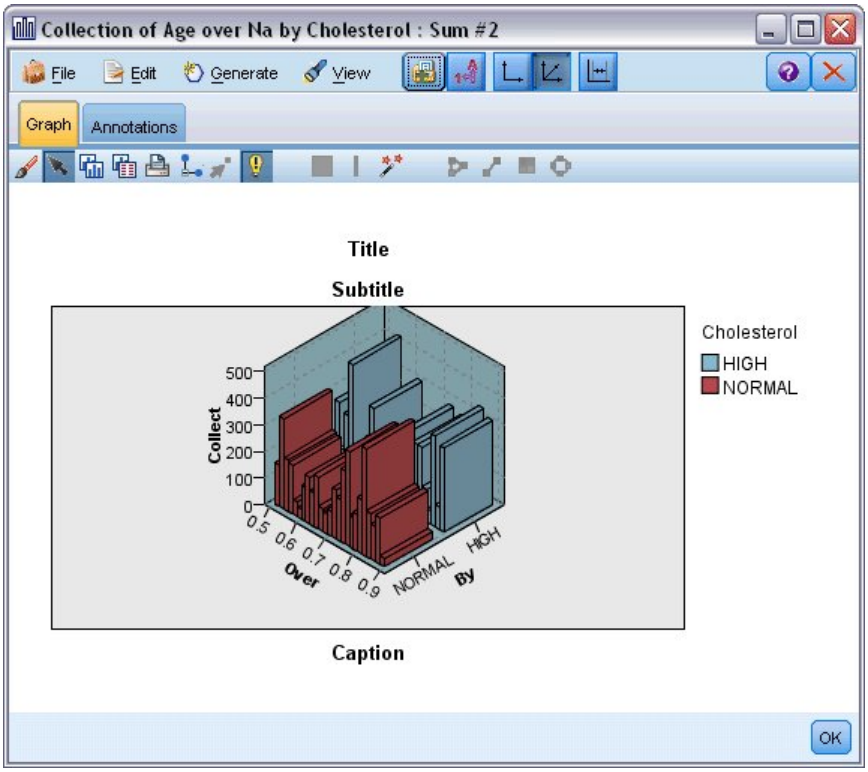


Abbildung 29. Position der Diagrammdarstellungsoptionen bei einem 3-D-Sammlungsdiagramm

Verwenden eines Sammlungsdiagramms

Bei Sammlungen wird die Verteilung der Werte für ein numerisches Feld *relativ zu den Werten eines anderen Felds* dargestellt, also nicht das Auftreten von Werten für ein einziges Feld. Histogramme funktionieren ähnlich wie Sammlungsdiagramme. Histogramme zeigen die Verteilung der Werte in einem numerischen Feld, dessen Werte an der *X*-Achse dargestellt werden.

Sobald Sie ein Diagramm erstellt haben, können Sie die Ergebnisse untersuchen und Abschnitte festlegen, um die Werte entlang der *X*-Achse aufzuspalten bzw. Regionen zu definieren. Außerdem können Sie Elemente innerhalb des Diagramms markieren. Weitere Informationen finden Sie im Thema „Exploration von Diagrammen” auf Seite 254.

Mit Optionen im Menü "Generieren" können Sie Balancierungs- Auswahl- und Ableitungsknoten erstellen. Hierfür werden die Daten im Diagramm bzw. genauer die Daten innerhalb bestimmter Abschnitte, Bereiche oder markierter Elemente verwendet. Dieser Diagrammtyp wird häufig als Vorbereitung auf Bearbeitungsknoten eingesetzt, um die Daten zu untersuchen und etwaige Unausgewogenheiten mithilfe eines Balancierungsknotens, der aus dem Diagrammknoten heraus erzeugt wird, auszugleichen. Darüber hinaus können Sie einen Flagableitungsknoten erzeugen und so ein Feld hinzufügen, aus dem hervorgeht, in welchen Abschnitt die einzelnen Datensätze fallen, oder auch einen Auswahlknoten, mit dem Sie alle Datensätze in einem bestimmten Set oder Wertebereich auswählen. Diese Funktionen sorgen dafür, dass ein bestimmtes Subset an Daten zur näheren Exploration im Mittelpunkt verbleibt. Weitere Informationen finden Sie im Thema „Generieren von Knoten aus Diagrammen” auf Seite 261.

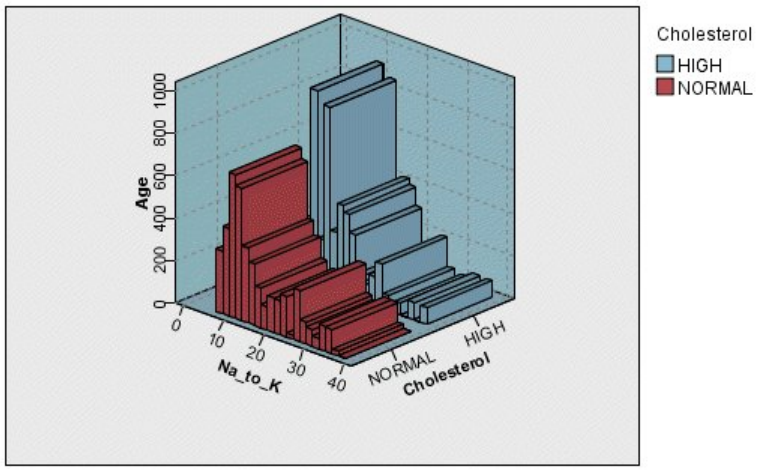


Abbildung 30. 3-D-Sammlungsdiagramm für die Summe von "Verh_Na/K" über "Alter" für hohe und normale Cholesterinspiegel.

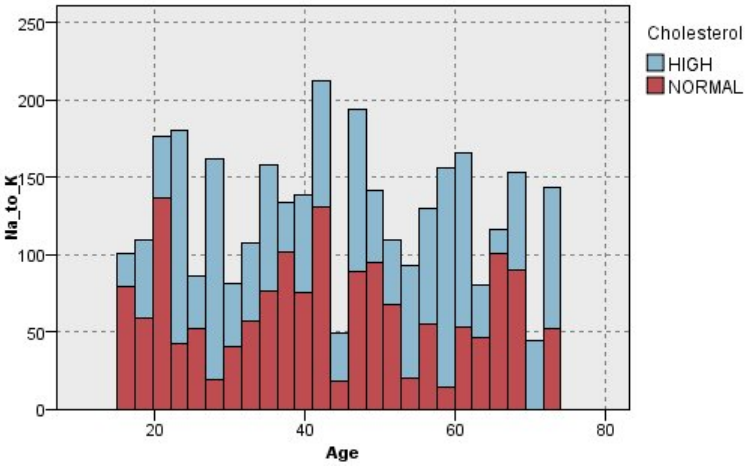


Abbildung 31. Sammlungsdiagramm ohne Anzeige der Z-Achse, jedoch mit "Cholesterin" als Farbüberlagerung

Multiplotknoten

Ein Multiplot ist eine besondere Art eines Plots, bei dem mehrere *Y*-Felder über einem einzelnen *X*-Feld dargestellt werden. Die *Y*-Felder werden als farbige Linien geplottet, die jeweils einem Plotknoten mit dem Stil **Linie** und dem X-Modus **Sortieren** entsprechen. Multiplots eignen sich für Zeitsequenzdaten, bei denen die Fluktuation mehrerer Variablen im Lauf der Zeit untersucht werden soll.

Multiplot - Registerkarte "Plot"

X-Feld. Wählen Sie ein Feld in der Liste aus, das auf der horizontalen *X*-Achse dargestellt werden soll.

Y-Feld. Wählen Sie mindestens ein Feld in der Liste aus, das über den Bereich der *X*-Feldwerte dargestellt werden soll. Mit der Feldauswahlschaltfläche können Sie mehrere Felder auswählen. Mit der Schaltfläche "Löschen" können Sie Felder wieder aus der Liste entfernen.

Überlagerung. Die Kategorien für die Datenwerte können auf unterschiedliche Weise dargestellt werden. Lassen Sie beispielsweise mehrere Plots für die einzelnen Werte in den Daten mithilfe einer Animationsüberlagerung darstellen. Dies ist nützlich für Sets mit mehr als 10 Kategorien. Bei Sets mit mehr als 15

Kategorien kann die Leistung beeinträchtigt werden. Weitere Informationen finden Sie im Thema „Formatierungen, Überlagerungen, Fenster und Animation" auf Seite 176.

Normalisieren. Wählen Sie diese Option aus, um alle *Y*-Werte zur Darstellung im Diagramm auf den Bereich 0-1 zu skalieren. Durch Normalisieren können Sie die Beziehung zwischen Linien untersuchen, die im Diagramm ansonsten aufgrund von Unterschieden im Wertebereich für die einzelnen Reihen verdeckt sind. Das Normalisieren wird bei der Darstellung mehrerer Linien in demselben Diagramm und für den Vergleich von Plots in nebeneinander angeordneten Teilfenstern empfohlen. (Eine Normalisierung ist nicht erforderlich, wenn alle Datenwerte in einen ähnlichen Bereich fallen.)

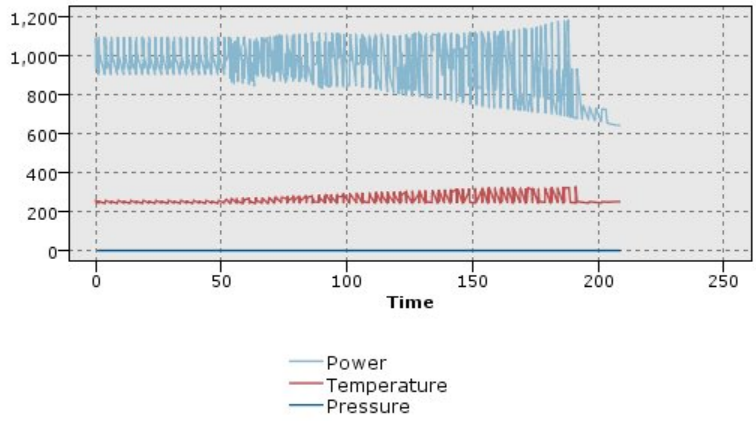


Abbildung 32. Standard-Multiplot mit der Kraftwerksfluktuation im Lauf der Zeit (Hinweis: Ohne Normalisierung ist der Plot für den Druck nicht sichtbar)

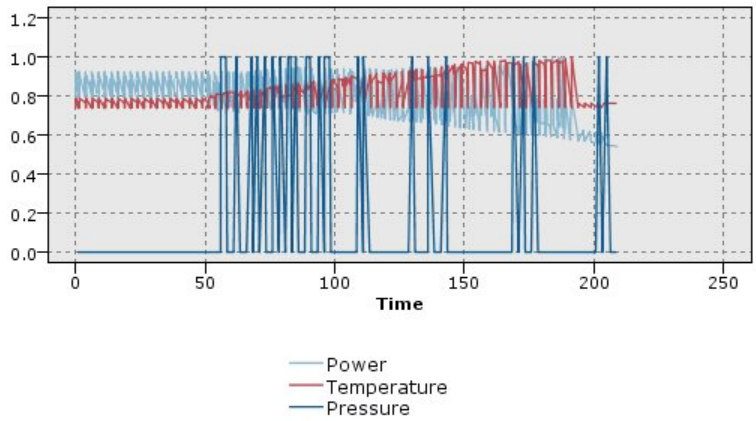


Abbildung 33. Normalisierter Multiplot mit einem Plot für den Druck

Überlagerungsfunkion. Hiermit geben Sie eine bekannte Funktion an, die mit tatsächlichen Werten verglichen werden soll. Um beispielsweise die Istwerte mit den vorhergesagten Werten zu vergleichen, plotten Sie die Funktion `y = x` als Überlagerung. Geben Sie eine Funktion für **y =** im Textfeld an. Die Standardfunktion lautet `y = x`; Sie können jedoch auch andere Funktionen festlegen, z. B. quadratische Funktionen oder beliebige Ausdrücke im Hinblick auf `x`.

Hinweis: Überlagerungsfunktionen sind für Fenster- und Animationsdiagramme nicht verfügbar.

Wenn Anzahl der Datensätze größer als. Geben Sie eine Methode für das Plotten umfangreicher Datasets an. Sie können wahlweise eine maximal zulässige Größe für das Dataset angeben oder den Standardwert von 2.000 Punkten verwenden. Bei umfangreichen Datasets steigt die Leistung, wenn Sie die Option **Klasse** oder **Stichprobe** aktivieren. Alternativ können Sie mit **Alle Daten verwenden** alle Datenpunkte gleichzeitig plotten lassen; dies kann sich jedoch beträchtlich auf die Leistung der Software auswirken.

Hinweis: Beim X-Modus **Überlagern** oder **Wie gelesen** sind diese Optionen inaktiviert und es werden nur die ersten *n* Datensätze verwendet.

- **Klasse.** Hiermit aktivieren Sie die Klassierung, wenn das Dataset mehr Datensätze enthält als die angegebene Anzahl. Beim Klassieren wird das Diagramm vor dem eigentlichen Plotten in feinmaschige Raster aufgeteilt und es wird die Anzahl der Verbindungen gezählt, die in die einzelnen Rasterzellen fallen würden. Im endgültigen Diagramm wird je eine Verbindung pro Zelle im Klassierschwerpunkt (Durchschnitt aller Verbindungspunkte in der Klasse) verwendet.
- **Beispiel.** Es wird eine zufällige Stichprobe mit der angegebenen Anzahl von Datensätzen aus den Daten gebildet.

Multiplot - Registerkarte "Darstellung"

Vor der Diagrammerstellung können Sie Darstellungsoptionen angeben.

Titel. Dient zur Eingabe des Texts, der als Titel des Diagramms verwendet werden soll.

Untertitel. Dient zur Eingabe des Texts, der als Untertitel des Diagramms verwendet werden soll.

Titelzeile. Dient zur Eingabe des Texts, der als Titelzeile des Diagramms verwendet werden soll.

X-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *X*-Achsenbeschriftung (horizontal) oder wählen Sie **Angepasst** aus, um eine Beschriftung anzugeben.

Y-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *Y*-Achsenbeschriftung (vertikal) oder wählen Sie **Angepasst** aus, um eine Beschriftung anzugeben.

Rasterlinie anzeigen. Diese Option ist standardmäßig aktiviert. Hiermit lassen Sie Rasterlinien hinter dem Plot oder dem Diagramm einblenden, was die Bestimmung der Bereichs- und Bandabschnittpunkte erleichtert. Rasterlinien werden stets in weißer Farbe angezeigt; bei einem weißen Diagrammhintergrund erfolgt die Anzeige in Grau.

Verwenden eines Multiplots

Plots und Multiplots sind im Grunde genommen Plots von *X* in Abhängigkeit von *Y*. Wenn Sie beispielsweise potenzielle Betrugsfälle in Bewerbungen um landwirtschaftliche Subventionen untersuchen (wie in *fraud.str* im Ordner *Demos* der IBM SPSS Modeler-Installation dargestellt), soll beispielsweise das in der Bewerbung angegebene Einkommen in Abhängigkeit von dem Einkommen geplottet werden, das mithilfe eines neuronalen Netzes geschätzt wurde. Aus einer Überlagerung, z. B. dem Feldfruchttyp, geht hervor, ob eine Beziehung zwischen den Forderungen (Wert oder Anzahl) und der Art der Feldfrucht besteht.

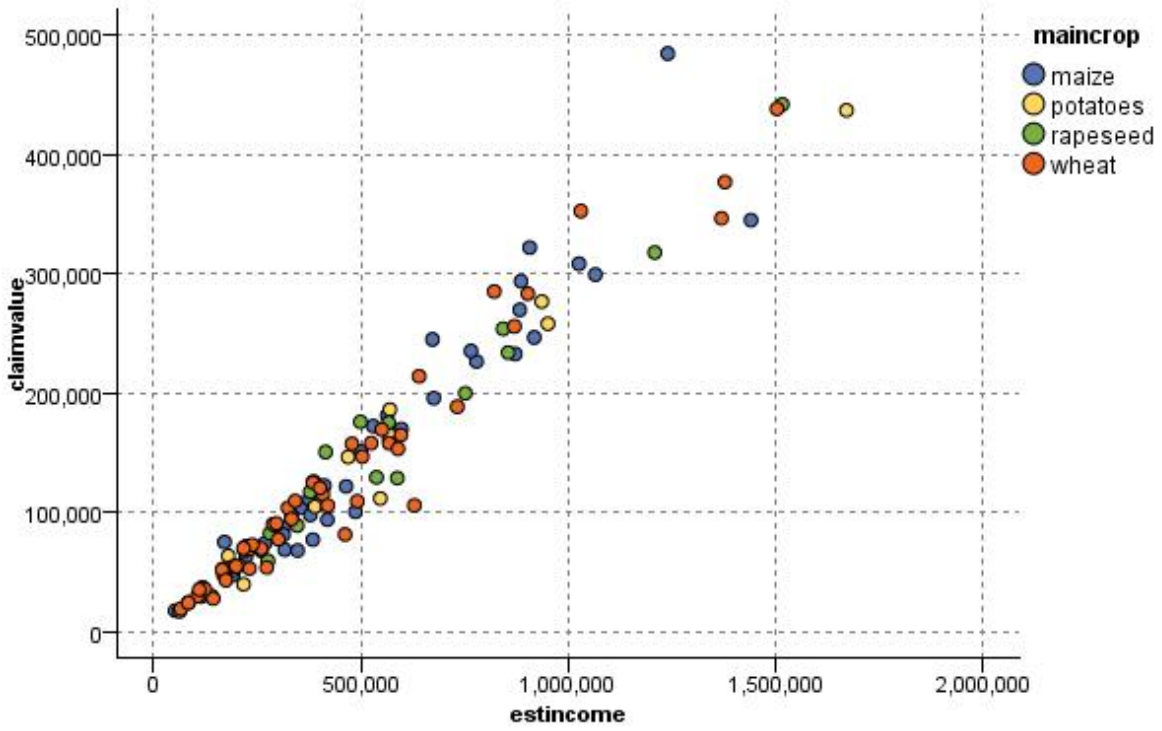


Abbildung 34. Plot der Beziehung zwischen geschätztem Einkommen und Forderungswert mit Hauptfeldfruchttyp als Überlagerung

Plots, Multiplots und Evaluierungsdiagramme sind zweidimensionale Darstellungen von *Y* gegen *X*. Die Arbeit mit diesen Diagrammen ist daher denkbar unkompliziert: Sie können ganz einfach Bereiche definieren, Elemente markieren und sogar Abschnitte einzeichnen. Außerdem können Sie Knoten für die durch diese Bereiche, Abschnitte bzw. Elemente dargestellten Daten generieren. Weitere Informationen finden Sie im Thema „Exploration von Diagrammen" auf Seite 254.

Netzdiagrammknoten

Netzdiagrammknoten zeigen die Stärke der Beziehung zwischen den Werten aus mindestens zwei symbolischen Feldern. Die Verbindungen werden mithilfe verschiedener Linientypen im Diagramm dargestellt, aus denen die Stärke der jeweiligen Verbindung hervorgeht. Mit Netzdiagrammknoten können Sie beispielsweise die Beziehung zwischen dem Kauf verschiedener Artikel auf einer e-Commerce-Website oder in einem traditionellen Einzelhandelsgeschäft untersuchen.

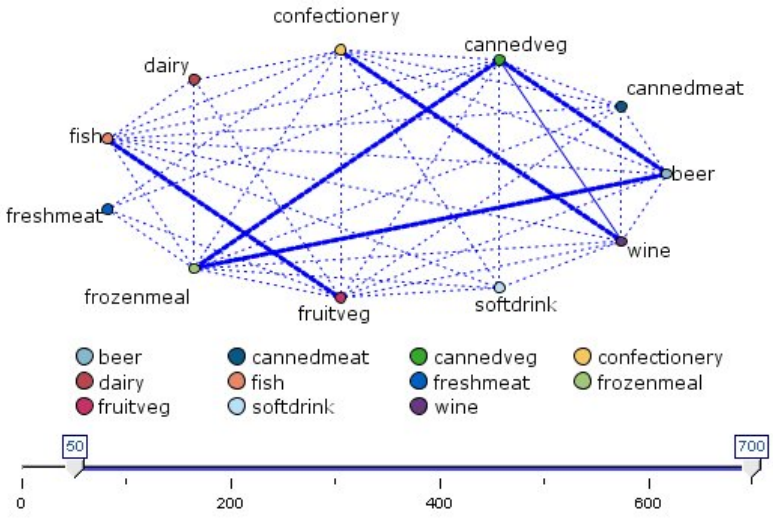


Abbildung 35. Netzdiagramm mit Beziehungen zwischen dem Kauf von Lebensmitteln

Gerichtete Netzdiagramme

Gerichtete Netzdiagrammknoten zeigen wie die Netzdiagrammknoten die Stärke der Beziehungen zwischen symbolischen Feldern. In gerichteten Netzdiagrammen sind jedoch nur die Verbindungen von mindestens einem Ausgangsfeld zu einem einzelnen Zielfeld ersichtlich. Die Verbindungen sind unidirektional, verlaufen also nur als "Einbahnstraßen".

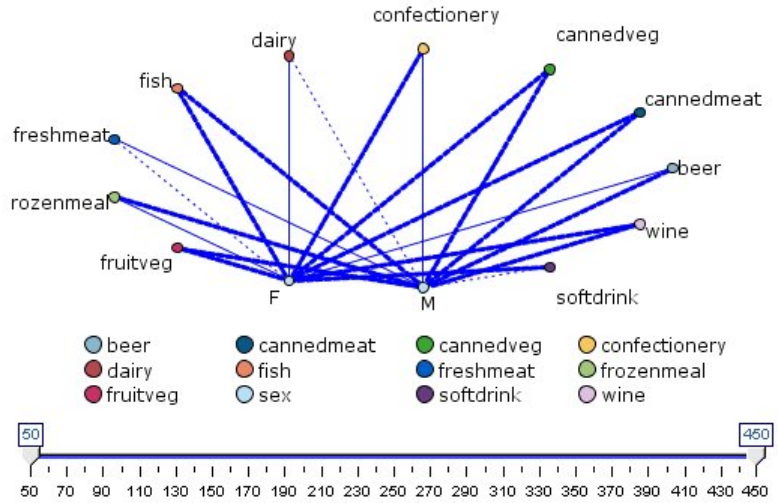


Abbildung 36. Gerichtetes Netzdiagramm mit der Beziehung zwischen dem Kauf von Lebensmitteln und dem Geschlecht

Wie bei Netzdiagrammknoten werden die Verbindungen mithilfe verschiedener Linientypen im Diagramm dargestellt, aus denen die Stärke der jeweiligen Verbindung hervorgeht. Mit einem gerichteten Netzdiagrammknoten können Sie beispielsweise die Beziehung zwischen dem Geschlecht des Käufers und der Neigung zum Kauf bestimmter Artikel untersuchen.

Netzdiagramm - Registerkarte "Plot"

Netzdiagramm. Hiermit erstellen Sie ein Netzdiagramm, das die Stärke der Beziehungen zwischen allen angegebenen Feldern verdeutlicht.

Gerichtetes Netzdiagramm. Ein gerichtetes Netzdiagramm wird erstellt, aus dem die Stärke der Beziehungen zwischen mehreren Feldern und den Werten eines einzigen Felds (z. B. Geschlecht oder Religion) hervorgeht. Wenn diese Option ausgewählt ist, wird ein "Zielfeld" aktiviert und das nachfolgende Steuerelement für die Felder wird zur näheren Verdeutlichung in Ausgangsfelder umbenannt.

Zielfeld (nur bei gerichteten Netzdiagrammen). Wählen Sie ein Flagfeld oder ein nominales Feld für ein zielgerichtetes Netzdiagramm aus. Die Liste enthält nur Felder, die nicht explizit als numerisch definiert sind.

Felder/Ausgangsfelder. Wählen Sie die gewünschten Felder für die Erstellung des gerichteten Netzdiagramms aus. Die Liste enthält nur solche Felder, die nicht explizit als numerisch definiert sind. Mit der Feldauswahlschaltfläche können Sie mehrere Felder auswählen. Alternativ können Sie die Felder nach Typ auswählen.

Hinweis: Bei gerichteten Netzdiagrammen dient dieses Steuerelement zur Auswahl der Ausgangsfelder.

Nur wahre Flags anzeigen. Es werden nur wahre Flags für ein Flagfeld angezeigt. Diese Option vereinfacht die Netzdiagrammanzeige und wird häufig bei Daten verwendet, bei denen das Auftreten positiver Werte von besonderer Bedeutung ist.

Zeilenwerte sind. Wählen Sie einen Schwellenwerttyp aus der Dropdown-Liste aus.

- Mit **Absolut** werden die Schwellenwerte auf der Grundlage der Anzahl an Datensätzen festgelegt, in denen die einzelnen Wertepaare vorkommen.
- Mit **Prozent insgesamt** rufen Sie die absolute Anzahl der Fälle ab, die im Zusammenhang als Anteil am Gesamtauftreten der einzelnen Wertepaare im Netzdiagramm dargestellt werden.
- Aus den Feldern **Prozentsätze vom kleineren Feld/Wert** und **Prozentsätze vom größeren Feld/Wert** geht hervor, welches Feld bzw. welcher Wert für die Evaluierung der Prozentsätze herangezogen werden soll. Beispiel: 100 Datensätze besitzen den Wert *MedY* für das Feld *Medikament*, nur 10 Datensätze dagegen den Wert *NIEDRIG* im Feld *BP*. Wenn sieben Datensätze sowohl den Wert *MedY* als auch *NIEDRIG* aufweisen, beträgt der Prozentsatz entsprechend 70 % oder 7 %, abhängig davon, welches Feld referenziert wird, also kleiner (*BP*) oder größer (*Medikament*).

Hinweis: Bei gerichteten Netzdiagrammen sind die dritte und vierte oben genannte Option nicht verfügbar. Stattdessen können Sie die Optionen **Prozentsätze vom Feld/Wert "Bis"** und **Prozentsätze vom Feld/Wert "Von"** auswählen.

Starke Zusammenhänge sind bedeutsamer. Diese Option ist standardmäßig aktiviert und ist die Standarddarstellung der Zusammenhänge zwischen Feldern.

Schwache Zusammenhänge sind bedeutsamer. Hiermit kehren Sie die Bedeutung der als fett gedruckte Linien dargestellten Zusammenhänge um. Diese Option wird häufig im Rahmen der Betrugserkennung oder bei der Untersuchung von Ausreißern herangezogen.

Netzdiagramm - Registerkarte "Optionen"

Bei Netzdiagrammknoten enthält die Registerkarte "Optionen" eine Reihe weiterer Optionen, mit denen Sie das Ausgabediagramm anpassen können.

Anzahl der Zusammenhänge. Mit den nachstehenden Optionen wird die Anzahl der im Ausgabediagramm dargestellten Zusammenhänge festgelegt. Ein Teil dieser Optionen, z. B. **Schwache Zusammenhänge unter** oder **Starke Zusammenhänge über**, stehen auch im Ausgabediagrammfenster zur Verfügung. Des Weiteren können Sie die Anzahl der angezeigten Zusammenhänge mit einem Schieberegler im fertigen Diagramm einstellen.

- **Maximale Anzahl der anzuzeigenden Zusammenhänge.** Geben Sie die maximale Anzahl der Zusammenhänge ein, die im Ausgabediagramm dargestellt werden sollen. Mithilfe der Pfeile können Sie den Wert einstellen.
- **Nur Zusammenhänge anzeigen über.** Geben Sie den Mindestwert an, ab dem eine Verbindung im Netzdiagramm dargestellt werden soll. Mithilfe der Pfeile können Sie den Wert einstellen.
- **Alle Zusammenhänge anzeigen.** Alle Zusammenhänge werden angezeigt, unabhängig von den Mindest- und Höchstwerten. Bei dieser Option steigt gegebenenfalls die Verarbeitungszeit an, wenn eine große Anzahl an Feldern vorliegt.

Bei sehr wenigen Datensätzen verwerfen. Zusammenhänge, die durch zu wenige Datensätze gestützt sind, werden ignoriert. Um den Schwellenwert für diese Option festzulegen, geben Sie den gewünschten Wert in das Feld **Minimale Anzahl Datensätze/Zeilen** ein.

Bei sehr vielen Datensätzen verwerfen. Stark gestützte Verbindungen werden ignoriert. Geben Sie den gewünschten Wert in das Feld **Max. Anzahl Datensätze/Zeilen** ein.

Schwache Zusammenhänge unter. Bestimmen Sie einen Schwellenwert für schwache Verbindungen (gepunktete Linien) und normale Verbindungen (normale Linien). Alle Verbindungen unterhalb dieses Werts gelten als schwach.

Starke Zusammenhänge über. Geben Sie einen Schwellenwert für starke Verbindungen (dicke Linien) und normale Verbindungen (normale Linien) an. Alle Verbindungen oberhalb dieses Werts gelten als stark.

Zusammenhangsstärke. Legen Sie Optionen zur Steuerung der Stärke von Zusammenhängen fest:

- **Zusammenhangsstärke schwankt fortlaufend.** Es wird ein Bereich von Zusammenhangsstärken dargestellt, der die Schwankungen bei der Verbindungsstärke auf der Grundlage der tatsächlichen Datenwerte wiedergibt.
- **Zusammenhangsstärke zeigt starke/normale/schwache Kategorien.** Es werden drei Verbindungsstärken dargestellt (stark, normal und schwach). Die Trennwerte für diese Kategorien können wahlweise oben bestimmt werden oder auch im fertigen Diagramm.

Netzdiagrammanzeige. Wählen Sie einen Typ für die Netzdiagrammanzeige aus:

- **Kreislayout.** Die standardmäßige Netzdiagrammanzeige wird verwendet.
- **Netzlayout.** Die stärksten Zusammenhänge werden mithilfe eines Algorithmus gruppiert. Auf diese Weise werden starke Zusammenhänge durch räumliche Differenzierung und durch gewichtete Linien hervorgehoben.
- **Gerichtetes Layout.** Wählen Sie diese Option aus, um eine gerichtete Webanzeige zu erstellen, die die Auswahl **Zielfeld** als Schwerpunkt für die Richtung verwendet.
- **Rasterlayout.** Wählen Sie diese Option aus, um eine Webanzeige zu erstellen, die in einem Rastermuster mit regelmäßigen Abständen ausgelegt ist.

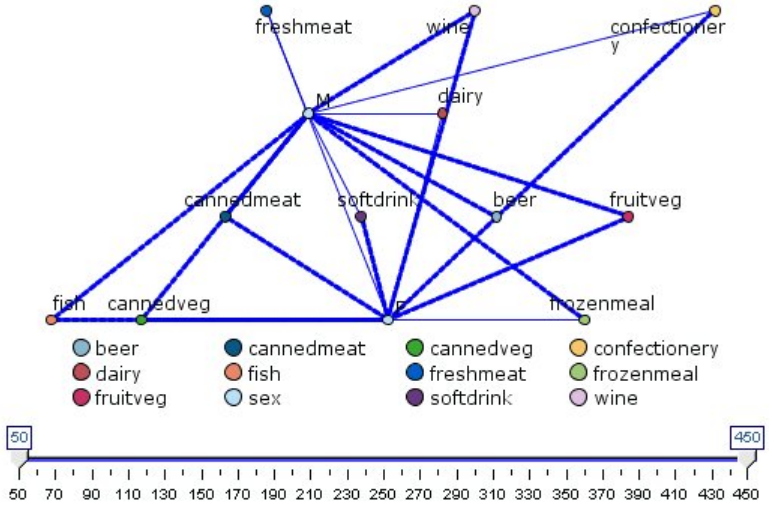


Abbildung 37. Netzdiagramm mit starken Verbindungen von "Tiefkühlware" und "Gemüse in Dosen" zu anderen Lebensmitteln

Netzdiagramm - Registerkarte "Darstellung"

Vor der Diagrammerstellung können Sie Darstellungsoptionen angeben.

Titel. Dient zur Eingabe des Texts, der als Titel des Diagramms verwendet werden soll.

Untertitel. Dient zur Eingabe des Texts, der als Untertitel des Diagramms verwendet werden soll.

Titelzeile. Dient zur Eingabe des Texts, der als Titelzeile des Diagramms verwendet werden soll.

Legende anzeigen. Dient zur Angabe, ob die Legende angezeigt werden soll oder nicht. Bei Plots mit zahlreichen Feldern wird die Darstellung gegebenenfalls verbessert, wenn Sie die Legende ausblenden.

Beschriftungen als Knoten anzeigen. Gibt an, dass die Beschriftungen nicht seitlich aufgeführt werden sollen, sondern direkt in jedem Knoten. Bei Plots mit einer geringen Anzahl an Feldern kann so die Übersichtlichkeit des Diagramms verbessert werden.

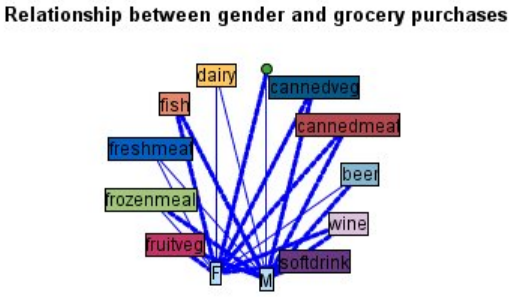


Abbildung 38. Netzdiagramm mit Beschriftungen als Knoten

Verwenden eines Netzdiagramms

Netzdiagrammknoten zeigen die Stärke der Beziehung zwischen den Werten aus mindestens zwei symbolischen Feldern. Die Verbindungen werden in Form von verschiedenen Linientypen, mit denen die größer werdende Stärke der Verbindungen verdeutlicht wird, in einem Diagramm dargestellt. Mit einem Netzdiagrammknoten können Sie beispielsweise die Beziehung zwischen Cholesterolspiegel, Blutdruck und dem eingenommenen Medikament bei der Behandlung des Patienten untersuchen.

- Starke Verbindungen sind mit einer dicken Linie gekennzeichnet. Dies bedeutet, dass die beiden Werte eng zusammenhängen und näher untersucht werden sollten.
- Mittelstarke Verbindungen sind als normal dicke Linien dargestellt.
- Schwache Verbindungen sind mit einer gepunkteten Linie gekennzeichnet.
- Befindet sich keine Linie zwischen zwei Werten, bedeutet dies entweder, dass die betreffenden Werte niemals gemeinsam in einem einzigen Datensatz auftreten oder dass diese Kombination in einer Anzahl an Datensätzen vorliegt, die unter dem im Dialogfeld für den Netzdiagrammknoten festgelegten Schwellenwert liegt.

Sobald Sie einen Netzdiagrammknoten erstellt haben, stehen verschiedene Optionen zur Auswahl, mit denen Sie die Darstellung des Diagramms anpassen und Knoten für die weitere Analyse erzeugen können.

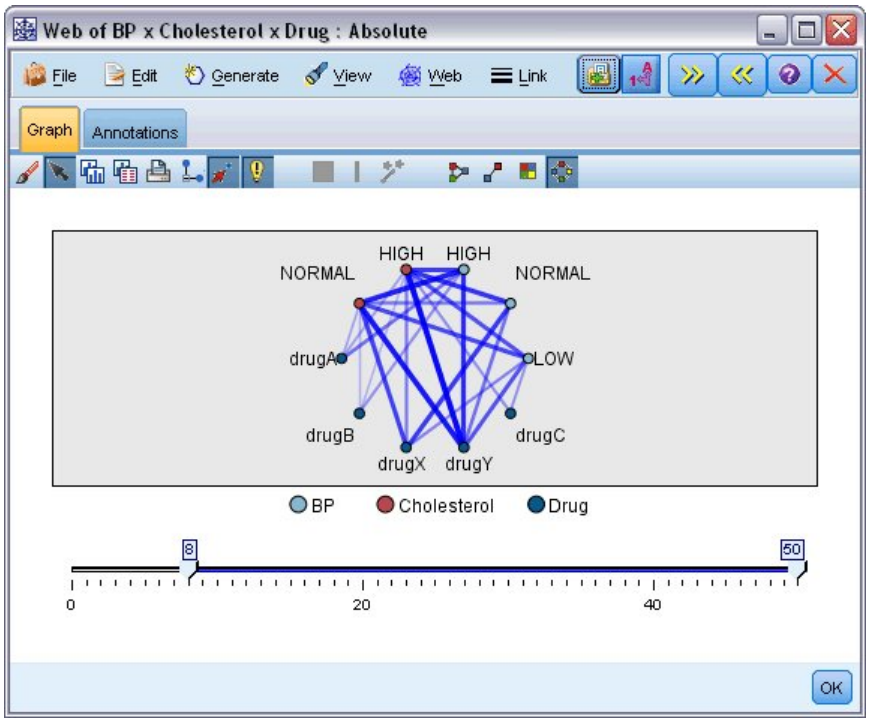


Abbildung 39. Netzdiagramm mit einer Reihe starker Verbindungen, z. B. normaler Blutdruck und MedX oder hoher Cholesterolspiegel und MedY

Bei Netzdiagrammknoten und gerichteten Netzdiagrammknoten stehen die folgenden Möglichkeiten zur Auswahl:

- Ändern Sie das Layout der Netzdiagrammanzeige.
- Blenden Sie verschiedene Punkte aus, um so die Darstellung zu vereinfachen.
- Ändern Sie die Schwellenwerte für die Linienstile.
- Heben Sie Linien zwischen Werten hervor und kennzeichnen Sie so eine "ausgewählte" Beziehung.
- Erzeugen Sie einen Auswahlknoten für einen oder mehrere "ausgewählte" Datensätze oder auch einen Flagableitungsknoten, der mit mindestens einer Beziehung im Netzdiagramm assoziiert ist.

So passen Sie die Punkte an:

- Punkte **verschieben**: Klicken Sie mit der Maus auf einen Punkt und ziehen Sie diesen an die gewünschte Position. Das Netzdiagramm wird neu gezeichnet, um so die neue Position wiederzugeben.
- Punkte **ausblenden**: Klicken Sie mit der rechten Maustaste auf einen Punkt im Netzdiagramm und wählen Sie im Kontextmenü die Option **Ausblenden** oder **Ausblenden und neu zeichnen**. Bei der Option **Ausblenden** lassen Sie lediglich den ausgewählten Punkt und die zugehörigen Linien ausblenden. Bei der Option **Ausblenden und neu zeichnen** wird das Netzdiagramm neu gezeichnet, sodass die vorgenommenen Änderungen ersichtlich werden. Alle manuell vorgenommenen Verschiebungen werden rückgängig gemacht.
- Alle ausgeblendeten Punkte **anzeigen**: Wählen Sie im Diagrammfenster im Menü "Netzdiagramm" den Befehl **Alle anzeigen** oder **Alle anzeigen und neu zeichnen**. Mit der Option **Alle anzeigen und neu zeichnen** lassen Sie das Netzdiagramm neu zeichnen, sodass auch alle bislang ausgeblendeten Punkte und deren Verbindungen wieder sichtbar werden.

So können Sie Linien auswählen oder "hervorheben":

Ausgewählte Linien werden in roter Farbe hervorgehoben.

1. Klicken Sie zum Auswählen einer einzelnen Linie bei gedrückter linker Maustaste auf die Linie.
2. Um mehrere Linien auszuwählen, führen Sie eine der folgenden Aktionen aus:

- Ziehen Sie mithilfe des Cursors einen Kreis um die Punkte auf, deren Linien Sie auswählen möchten.
- Halten Sie die Steuertaste gedrückt und klicken Sie mit der linken Maustaste auf die einzelnen Linien, die Sie auswählen möchten.

Sie können die Auswahl aller Linien aufheben, indem Sie in den Diagrammhintergrund klicken oder **Auswahl aufheben** aus dem Web-Menü im Diagrammfenster wählen.

So lassen Sie das Netzdiagramm mithilfe eines anderen Layouts anzeigen:

Wählen Sie im Menü "Web" **Kreislayout**, **Netzlayout**, **Gerichtetes Layout** oder **Rasterlayout** aus, um das Layout des Diagramms zu ändern.

So schalten Sie den Links-Schieberegler ein bzw. aus.

Wählen Sie im Menü "Ansicht" die Option **Links-Schieberegler**.

So können Sie Datensätze für eine einzelne Beziehung auswählen oder mit einem Flag versehen:

1. Klicken Sie mit der rechten Maustaste auf die Linie für die relevante Beziehung.
2. Wählen Sie im Kontextmenü die Option **Auswahlknoten für Zusammenhang generieren** oder **Ableitungsknoten für Zusammenhang generieren**.

In den Streamerstellungsbereich wird automatisch ein Auswahlknoten oder Ableitungsknoten mit den richtigen Optionen und Bedingungen aufgenommen.

- Mit dem Auswahlknoten werden alle Datensätze in der betreffenden Beziehung ausgewählt.
- Der Ableitungsknoten erzeugt ein Flag, aus dem hervorgeht, ob die ausgewählte Beziehung für Datensätze im gesamten Dataset gilt. Der Name des Flagfelds besteht aus den beiden Werten in der Beziehung, getrennt durch einen Unterstrich, z. B. *NIEDRIG_MedC* oder *MedC_NIEDRIG*.

So können Sie Datensätze für eine Gruppe von Beziehungen auswählen oder mit einem Flag versehen:

1. Wählen Sie die Linie(n) für die relevanten Beziehungen in der Netzdiagrammanzeige aus.
2. Wählen Sie im Diagrammfenster im Menü "Generieren" den Befehl **Auswahlknoten ("UND")**, **Auswahlknoten ("ODER")**, **Ableitungsknoten ("UND")** oder **Ableitungsknoten ("ODER")**.

- Bei den "ODER"-Knoten werden die Bedingungen getrennt voneinander betrachtet. Der Knoten gilt also für alle Datensätze, bei denen mindestens eine der ausgewählten Beziehungen vorliegt.
- Bei den "UND"-Knoten werden die Bedingungen gemeinsam betrachtet. Der Knoten gilt also für alle Datensätze, bei denen alle ausgewählten Beziehungen vorliegen. Falls ausgewählte Beziehungen sich gegenseitig ausschließen, tritt ein Fehler auf.

Sobald die Auswahl abgeschlossen ist, wird automatisch ein Auswahlknoten oder Ableitungsknoten mit den richtigen Optionen und Bedingungen in den Streamerstellungsbereich aufgenommen.

Anpassen der Netzdiagrammschwellenwerte

Sobald Sie ein Netzdiagramm erstellt haben, können Sie die Schwellenwerte für die Linienstile mit dem Schieberegler einstellen und so die minimale noch sichtbare Linie ändern. Des Weiteren können zusätzliche Optionen für die Schwellenwerte abgerufen werden. Klicken Sie hierzu auf den gelben Doppelpfeil in der Symbolleiste. Das Netzdiagrammfenster wird erweitert. Klicken Sie anschließend auf die Registerkarte **Steuerelemente** und wählen Sie die gewünschten zusätzlichen Optionen.

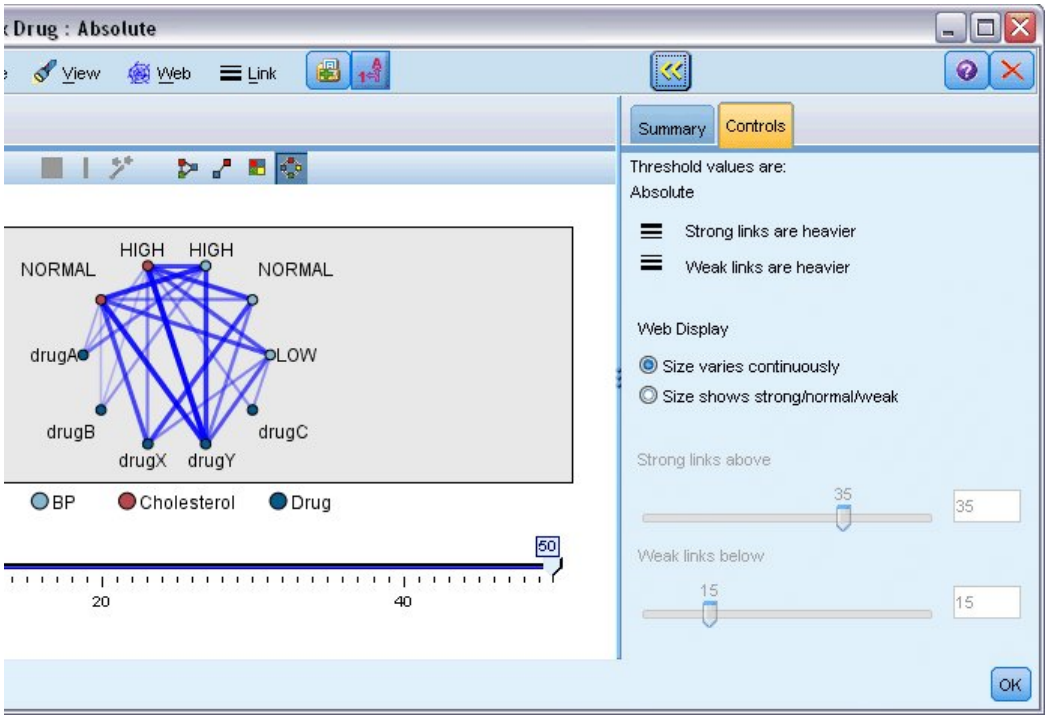


Abbildung 40. Erweitertes Fenster mit Darstellungs- und Schwellenwertoptionen

Einstellung für Schwellenwert. Dies ist der Typ des Schwellenwerts, den Sie beim Erstellen im Dialogfeld des Netzdiagrammknotens ausgewählt haben.

Starke Zusammenhänge sind bedeutsamer. Diese Option ist standardmäßig aktiviert und ist die Standarddarstellung der Zusammenhänge zwischen Feldern.

Schwache Zusammenhänge sind bedeutsamer. Hiermit kehren Sie die Bedeutung der als fett gedruckte Linien dargestellten Zusammenhänge um. Diese Option wird häufig im Rahmen der Betrugserkennung oder bei der Untersuchung von Ausreißern herangezogen.

Netzdiagrammanzeige. Legen Sie Optionen zur Steuerung der Stärke von Zusammenhängen im Ausgabediagramm fest:

- **Größe schwankt fortlaufend.** Es wird ein Bereich von Zusammenhangsstärken dargestellt, der die Schwankungen bei der Verbindungsstärke auf der Grundlage der tatsächlichen Datenwerte wiedergibt.

- **Größe zeigt stark/mittel/schwach.** Es werden drei Verbindungsstärken dargestellt (stark, normal und schwach). Die Trennwerte für diese Kategorien können wahlweise oben bestimmt werden oder auch im fertigen Diagramm.

Starke Zusammenhänge über. Geben Sie einen Schwellenwert für starke Verbindungen (dicke Linien) und normale Verbindungen (normale Linien) an. Alle Verbindungen oberhalb dieses Werts gelten als stark. Stellen Sie den Wert mit dem Schieberegler ein oder geben Sie einen Wert in das Feld ein.

Schwache Zusammenhänge unter. Bestimmen Sie einen Schwellenwert für schwache Verbindungen (gepunktete Linien) und normale Verbindungen (normale Linien). Alle Verbindungen unterhalb dieses Werts gelten als schwach. Stellen Sie den Wert mit dem Schieberegler ein oder geben Sie einen Wert in das Feld ein.

Wenn Sie die Schwellenwerte für ein Netzdiagramm angepasst haben, können Sie die Netzdiagrammanzeige mit den neuen Schwellenwerten aktualisieren (neu zeichnen). Verwenden Sie hierzu das Menü in der Symbolleiste des Netzdiagramms. Sobald Sie die richtigen Einstellungen gefunden haben, die zu den aussagekräftigsten Mustern führen, können Sie die ursprünglichen Einstellungen im Netzdiagrammknoten (auch als "übergeordneter Netzdiagrammknoten" bezeichnet) aktualisieren. Wählen Sie hierzu im Diagrammfenster im Menü "Netzdiagramm" den Befehl **Übergeordneten Knoten aktualisieren**.

Erstellen einer Netzdiagrammübersicht

Sie können eine Netzdiagrammübersicht anlegen, in der die starken, mittleren und schwachen Linien aufgeführt werden. Klicken Sie hierzu auf den gelben Doppelpfeil in der Symbolleiste. Das Netzdiagrammfenster wird erweitert. Klicken Sie anschließend auf die Registerkarte **Übersicht**. Hier werden Tabellen für die einzelnen Arten der Zusammenhänge aufgeführt. Mit den Umschalttasten können Sie die Tabellen erweitern und reduzieren.

Um die Übersicht auszudrucken, wählen Sie Folgendes aus dem Menü im Netzdiagrammfenster:

Datei > **Übersicht drucken**

Zeitdiagrammknoten

Mit Zeitdiagrammknoten können Sie die Darstellung einer oder mehrerer Zeitreihen über einen bestimmten Zeitraum anzeigen. Die geplotteten Reihen müssen numerische Werte enthalten. Außerdem wird vorausgesetzt, dass sie in einem Zeitbereich mit einheitlichen Abschnitten auftreten. Normalerweise wird vor einem Zeitdiagrammknoten ein Zeitintervallknoten verwendet, um ein *TimeLabel*-Feld zu erstellen, das standardmäßig zur Beschriftung der *X*-Achse im Diagramm verwendet wird. Weitere Informationen finden Sie im Thema „Zeitintervallknoten” auf Seite 159.

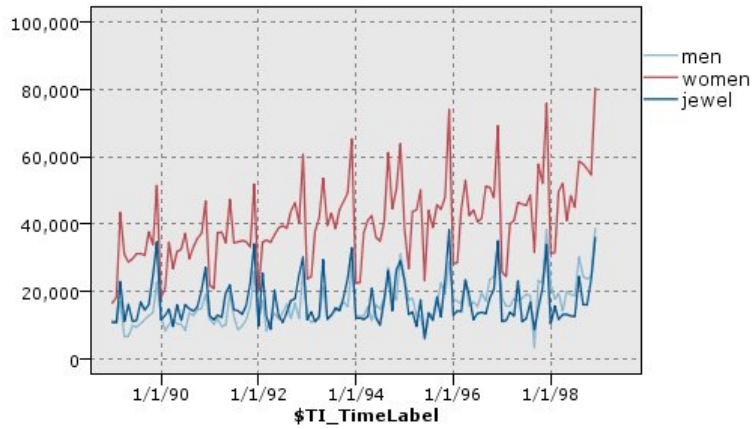


Abbildung 41. Darstellung der Verkäufe bei Herren- und Damenbekleidung und Schmuck im Laufe der Zeit

Erstellen von Interventionen und Ereignissen

Sie können Ereignis- und Interventionsfelder aus dem Zeitdiagramm erstellen, indem Sie einen Ableitungsknoten (Flag oder nominal) aus den Kontextmenüs generieren. Sie können beispielsweise ein Ereignisfeld für den Fall eines Bahnstreiks erstellen, der den Ableitungsstatus "Wahr" aufweist, wenn das Ereignis eingetreten ist, und ansonsten den Ableitungsstatus "Falsch". Bei einem Interventionsfeld, beispielsweise für eine Preiserhöhung, könnten Sie eine Ableitungsanzahl verwenden, um das Datum der Erhöhung anzugeben; dabei wird der Wert "0" für den alten und "1" für den neuen Preis verwendet. Weitere Informationen finden Sie im Thema „Ableitungsknoten” auf Seite 134.

Zeit - Registerkarte "Plot"

Diagramm. Bietet eine Auswahl für das Plotten der Zeitreihendaten.

- **Ausgewählte Zeitreihe.** Plottet Werte für ausgewählte Zeitreihen. Wenn Sie diese Option beim Plotten von Konfidenzintervallen auswählen, müssen Sie das Kontrollkästchen **Normalisieren** inaktivieren.
- **Ausgewählte Zeitreihenmodelle**. In Verbindung mit einem Zeitreihenmodell plottet diese Option alle verwandten Felder (tatsächliche und vorhergesagte Werte sowie die Konfidenzintervalle) für mindestens eine ausgewählte Zeitreihe. Mit dieser Option werden einige anderen Optionen im Dialogfeld inaktiviert. Diese Option eignet sich besonders für das Plotten von Konfidenzintervallen.

Zeitreihen. Wählen Sie mindestens ein Feld mit Zeitreihendaten für den Plot aus. Die Daten müssen numerisch sein.

X-Achsenbeschriftung. Wählen Sie entweder die Standardbeschriftung oder ein einzelnes Feld aus, das als Beschriftung für die *X*-Achse in Plots dienen soll. Bei Auswahl von "Standard" verwendet das System das aus einem aufwärts gelegenen Zeitintervallknoten erstellte TimeLabel-Feld bzw. aufeinanderfolgende ganze Zahlen, wenn kein Zeitintervallknoten vorhanden ist. Weitere Informationen finden Sie im Thema „Zeitintervallknoten” auf Seite 159.

Reihe in gesonderten Fenstern anzeigen. Gibt an, ob jede Reihe in einem gesonderten Fenster angezeigt werden soll. Wenn Sie nicht verschiedene Fenster verwenden möchten, werden alternativ alle Zeitreihen i demselben Diagramm dargestellt und es stehen keine Glättungselemente zur Verfügung. Wenn alle Zeitreihen in demselben Diagramm dargestellt werden, wird jede Reihe in einer anderen Farbe angezeigt.

Normalisieren. Wählen Sie diese Option aus, um alle *Y*-Werte zur Darstellung im Diagramm auf den Bereich 0-1 zu skalieren. Durch Normalisieren können Sie die Beziehung zwischen Linien untersuchen, die im Diagramm ansonsten aufgrund von Unterschieden im Wertebereich für die einzelnen Reihen verdeckt sind. Das Normalisieren wird bei der Darstellung mehrerer Linien in demselben Diagramm und für den

Vergleich von Plots in nebeneinander angeordneten Teilfenstern empfohlen. (Eine Normalisierung ist nicht erforderlich, wenn alle Datenwerte in einen ähnlichen Bereich fallen.)

Anzeigen. Wählen Sie mindestens ein Element aus, das in Ihrem Plot angezeigt werden soll. Sie können aus Linien, Punkten und Glättungselementen (LOESS-Smoother) wählen. Glättungselemente sind nur verfügbar, wenn die Reihen in gesonderten Fenstern angezeigt werden. Standardmäßig ist das Linienelement ausgewählt. Denken Sie daran, mindestens ein Plotelement auszuwählen, bevor Sie den Grafikknoten ausführen; andernfalls gibt das System eine Fehlernachricht aus, die angibt, dass Sie keine Elemente für den Plot ausgewählt haben.

Datensätze beschränken. Wählen Sie diese Option, wenn Sie die Anzahl der zu plottenden Datensätze begrenzen möchten. Geben Sie unter der Option **Maximale Anzahl der Datensätze für Plot** die Anzahl der zu plottenden Datensätze ein (Lesebeginn ist der Anfang der Datendatei). Dieser Wert ist standardmäßig auf 2.000 gesetzt. Wenn Sie die letzten *n* Datensätze in Ihrer Datendatei plotten möchten, können Sie vor diesen Knoten einen Sortierknoten schalten, um die Datensätze in zeitlich absteigender Reihenfolge zu ordnen.

Zeitdiagramm - Registerkarte "Darstellung"

Vor der Diagrammerstellung können Sie Darstellungsoptionen angeben.

Titel. Dient zur Eingabe des Texts, der als Titel des Diagramms verwendet werden soll.

Untertitel. Dient zur Eingabe des Texts, der als Untertitel des Diagramms verwendet werden soll.

Titelzeile. Dient zur Eingabe des Texts, der als Titelzeile des Diagramms verwendet werden soll.

X-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *X*-Achsenbeschriftung (horizontal) oder wählen Sie **Angepasst** aus, um eine Beschriftung anzugeben.

Y-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *Y*-Achsenbeschriftung (vertikal) oder wählen Sie **Angepasst** aus, um eine Beschriftung anzugeben.

Rasterlinie anzeigen. Diese Option ist standardmäßig aktiviert. Hiermit lassen Sie Rasterlinien hinter dem Plot oder dem Diagramm einblenden, was die Bestimmung der Bereichs- und Bandabschnittpunkte erleichtert. Rasterlinien werden stets in weißer Farbe angezeigt; bei einem weißen Diagrammhintergrund erfolgt die Anzeige in Grau.

Layout. Nur bei Zeitdiagrammen können Sie angeben, ob die Zeitwerte entlang einer horizontalen oder einer vertikalen Achse dargestellt werden sollen.

Verwenden eines Zeitdiagramms

Sobald Sie ein Zeitdiagramm erstellt haben, stehen verschiedene Optionen zur Auswahl, mit denen Sie die Darstellung des Diagramms anpassen und Knoten für die weitere Analyse erzeugen können. Weitere Informationen finden Sie im Thema „Exploration von Diagrammen” auf Seite 254.

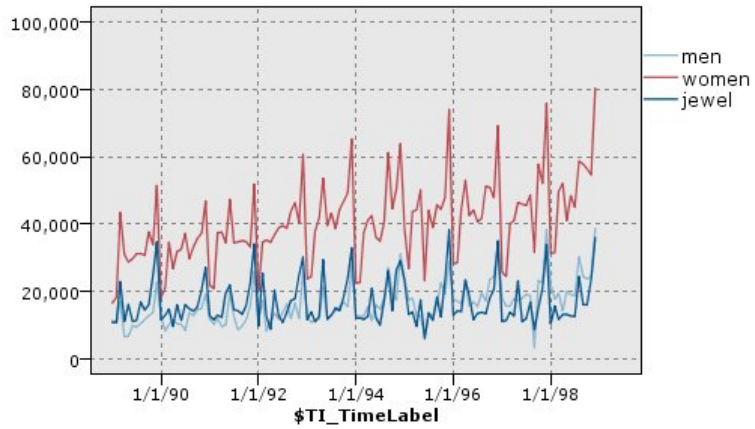


Abbildung 42. Darstellung der Verkäufe bei Herren- und Damenbekleidung und Schmuck im Laufe der Zeit

Sobald Sie ein Zeitdiagramm erstellt, Abschnitte definiert und die Ergebnisse untersucht haben, können Sie mit den Optionen im Menü "Generieren" und im Kontextmenü verschiedene Auswahl- und Ableitungsknoten erstellen. Weitere Informationen finden Sie im Thema „Generieren von Knoten aus Diagrammen” auf Seite 261.

Evaluierungsknoten

Der Evaluierungsknoten eröffnet eine unkomplizierte Möglichkeit, Vorhersagemodelle auszuwerten und miteinander zu vergleichen, um so das am besten geeignete Modell für die Anwendung zu ermitteln. Evaluierungsdiagramme zeigen die Leistung der Modelle beim Vorhersagen bestimmter Ergebnisse. Hierzu werden Datensätze auf der Grundlage des vorhergesagten Werts und der Konfidenz der Vorhersage sortiert. Die Datensätze werden dabei in gleich große Gruppen (**Quantile**) aufgeteilt; anschließend wird der Wert des Geschäftskriteriums für jedes Quantil dargestellt, vom höchsten Wert bis zum niedrigsten Wert. Mehrere Modelle werden als separate Linien im Diagramm dargestellt.

Zur Handhabung der Ergebnisse wird ein bestimmter Wert oder Wertebereich als **Treffer** definiert. Ein Treffer weist in der Regel auf einen gewissen Erfolg hin (z. B. auf einen Verkauf an einen Kunden) oder auf ein relevantes Ereignis (z. B. auf eine bestimmte medizinische Diagnose). Auf der Registerkarte "Optionen" des Dialogfelds können Sie Trefferkriterien definieren oder Sie können die standardmäßigen Trefferkriterien verwenden:

- **Flag**-Ausgabefelder sind unkompliziert; ein Treffer steht für *wahre* Werte.
- Bei **Nominal**-Ausgabefeldern definiert der erste Wert im Set einen Treffer.
- Bei **Stetig**-Ausgabefeldern entspricht ein Treffer einem Wert, der größer ist als der Mittelpunkt des Bereichs für das betreffende Feld.

Es stehen sechs Typen von Evaluierungsdiagrammen zur Auswahl, bei denen der Schwerpunkt jeweils auf einem anderen Auswertungskriterium liegt.

Gewinndiagramme

Gewinne sind definiert als der Anteil an allen Treffern, der in den einzelnen Quantilen vorliegt. Die Gewinne werden wie folgt berechnet: (Anzahl der Treffer im Quantil / Gesamtanzahl der Treffer) × 100 %.

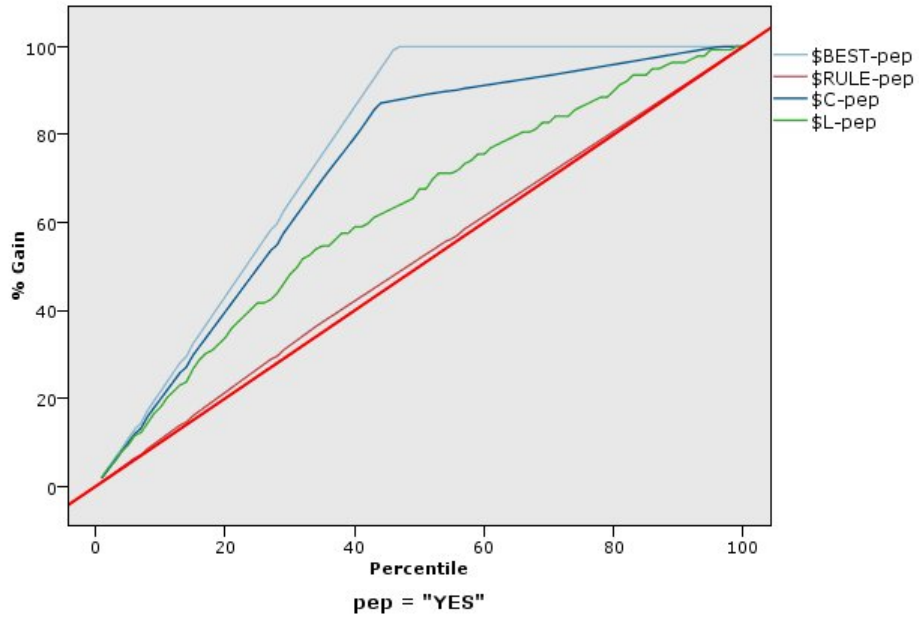


Abbildung 43. Gewinndiagramm (kumulativ) mit Basis, bester Linie und Geschäftsregel

Liftdiagramme

Beim Lift wird der Prozentsatz der Datensätze in jedem Quantil, die als Treffer gelten, mit dem Gesamtprozentsatz der Treffer in den Trainingsdaten verglichen. Die Berechnung läuft wie folgt ab: (Treffer im Quantil / Datensätze im Quantil) / (Gesamtanzahl der Treffer / Gesamtanzahl der Datensätze).

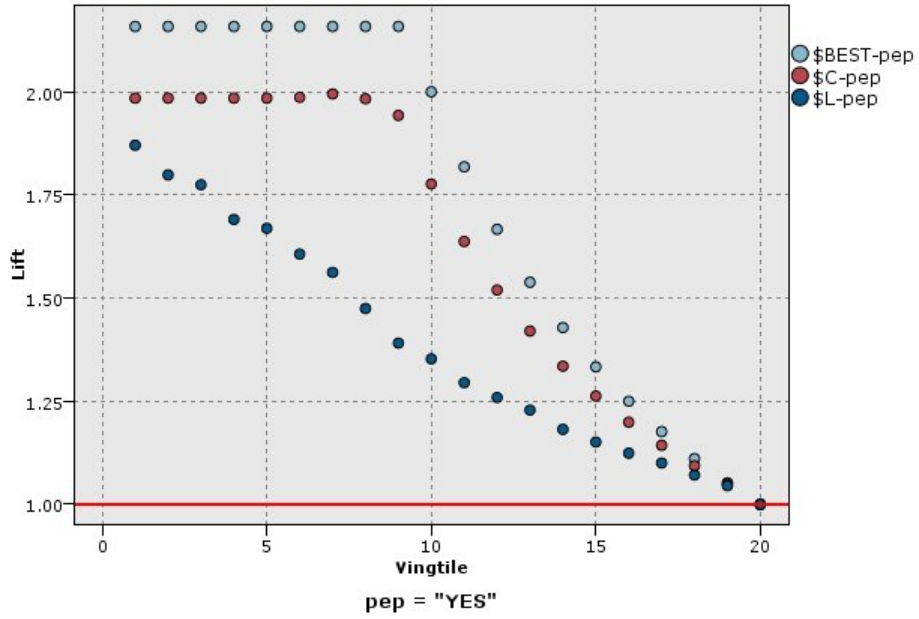


Abbildung 44. Liftdiagramm (kumulativ) mit Punkten und bester Linie

Trefferdiagramme

Treffer bezeichnen einfach den Prozentsatz der Datensätze im Quantil, die als Treffer gelten. Die Treffer werden wie folgt berechnet: (Treffer im Quantil/Datensätze im Quantil) × 100 %.

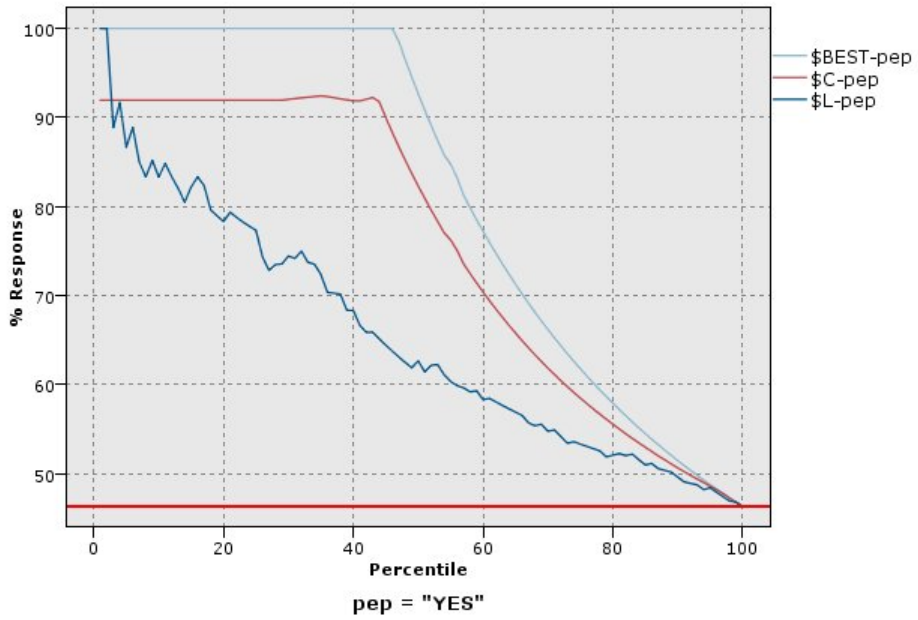


Abbildung 45. Trefferdiagramm (kumulativ) mit bester Linie

Profitdiagramme

Der Profit entspricht dem **Umsatz** für jeden Datensatz abzüglich der **Kosten** für den betreffenden Datensatz. Die Profite für ein Quantil entsprechen einfach der Summe der Profite für alle Datensätze im Quantil. Umsätze gelten definitionsgemäß nur für Treffer, Kosten dagegen für alle Datensätze. Die Profite und Kosten können fest sein oder auch durch Felder in den Daten definiert werden. Die Profite werden wie folgt berechnet: (Summe des Ertrags für die Datensätze im Quantil – Summe der Kosten für die Datensätze im Quantil).

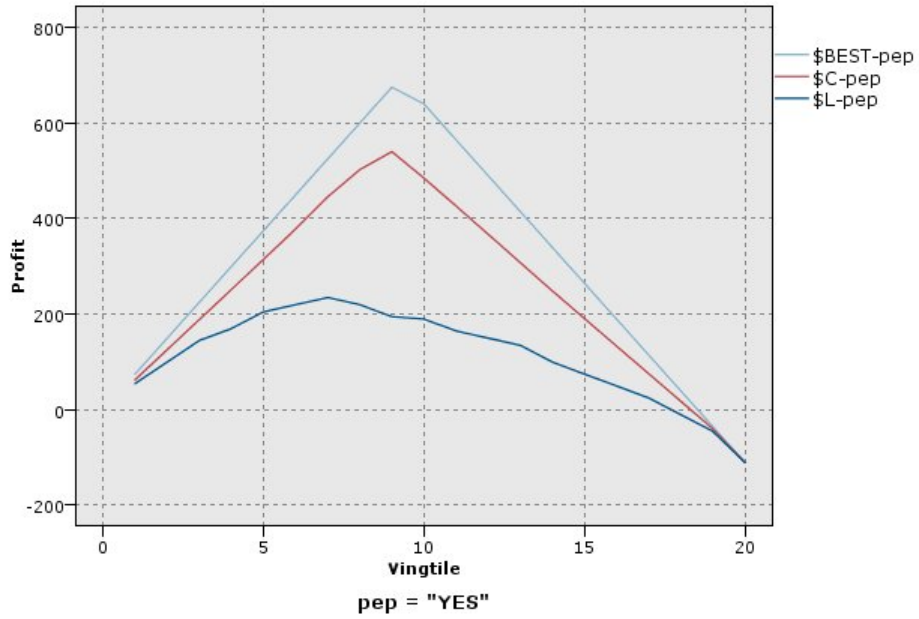


Abbildung 46. Profitdiagramm (kumulativ) mit bester Linie

ROI-Diagramme

Der ROI (Return-on-Investment) weist gewisse Ähnlichkeiten mit dem Profit auf; auch hier wird eine Definition für Umsätze und Kosten herangezogen. Beim ROI werden die Profite mit den Kosten für das Quantil verglichen. Der ROI wird wie folgt berechnet: (Profite für das Quantil/Kosten für das Quantil) × 100 %.

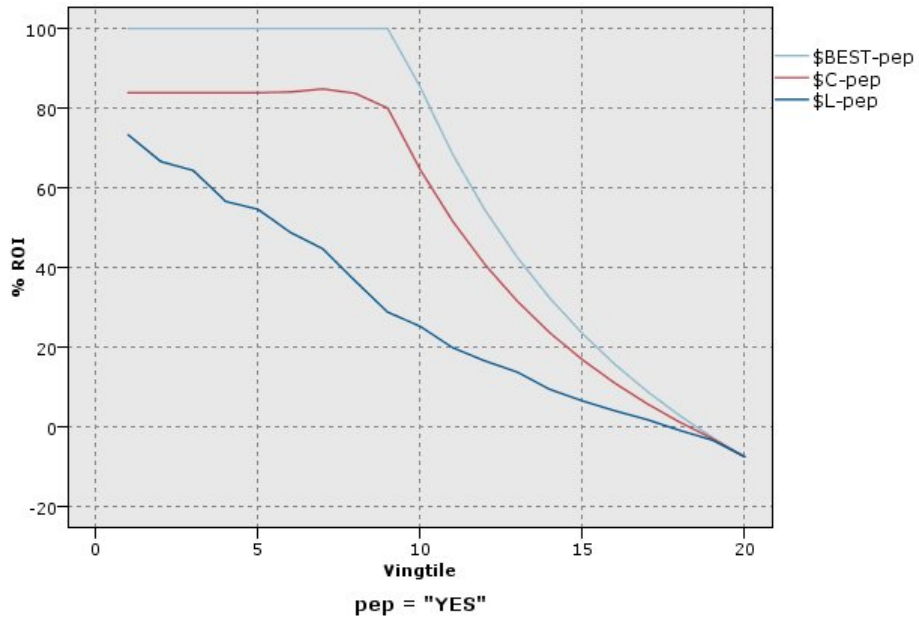


Abbildung 47. ROI-Diagramm (kumulativ) mit bester Linie

ROC-Diagramme

ROC (Receiver Operator Characteristic) kann nur mit binären Klassifikationsmerkmalen verwendet werden. Mithilfe von ROC können Klassifikationsmerkmale auf der Basis ihrer Leistung visualisiert, organisiert und ausgewählt werden. In einem ROC-Diagramm wird die "Wahr Positiv"-Rate (oder Sensitivität) gegenüber der "Falsch positiv"-Rate des Klassifikationsmerkmals dargestellt. In einem ROC-Diagramm werden die relativen Kompromisse zwischen Vorteilen (wahr positiv) und Kosten (falsch positiv) dargestellt. "Wahr positiv" ist eine Instanz, die einen Treffer darstellt und als Treffer klassifiziert wird. Daher wird die "Wahr positiv"-Rate als Anzahl der als "wahr positiv" erkannten Instanzen geteilt durch die Anzahl der Instanzen, die tatsächlich Treffer darstellen, berechnet. "Falsch positiv" ist eine Instanz, die ein Fehlschlag ist und als Treffer klassifiziert wird. Daher wird die "Falsch positiv"-Rate als Anzahl der als "falsch positiv" erkannten Instanzen geteilt durch die Anzahl der Instanzen, die tatsächlich Fehlschläge darstellen, berechnet.

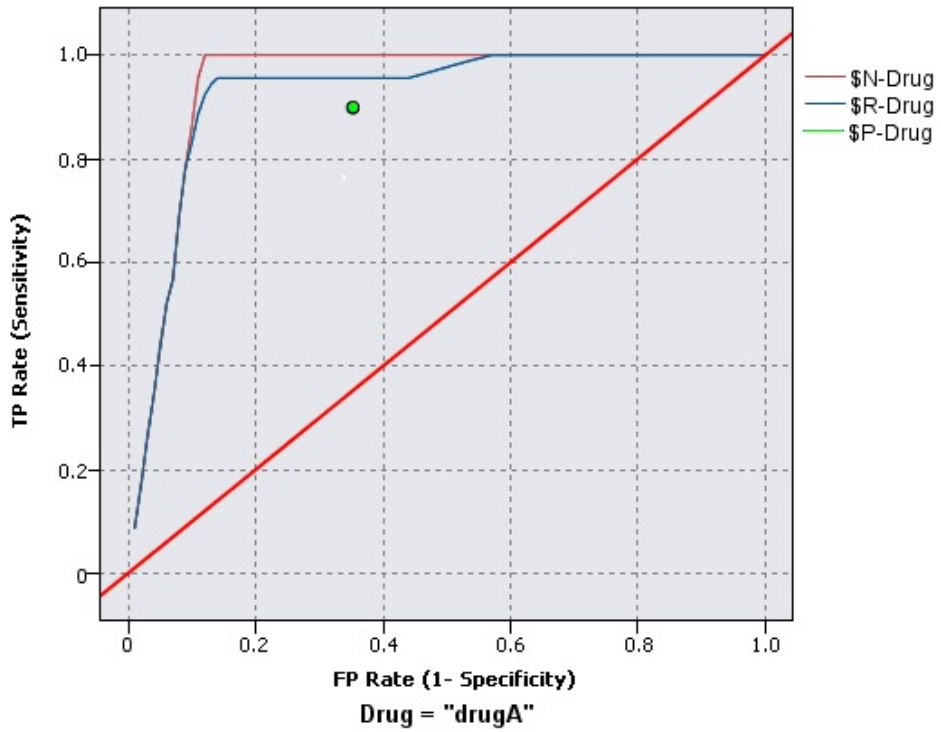


Abbildung 48. ROI-Diagramm mit bester Linie

Auch Evaluierungsdiagramme können kumulativ sein, sodass jeder Punkt dem Wert für das entsprechende Quantil zuzüglich aller höheren Quantile entspricht. Kumulative Diagramme geben die Gesamtleistung von Modellen in der Regel besser wieder; nicht kumulative Diagramme weisen dagegen häufig auf bestimmte Problembereiche in den Modellen hin.

Evaluierung - Registerkarte "Diagramm"

Diagrammtyp. Wählen Sie einen der folgenden Typen aus: **Gewinne**, **Antwort**, **Lift**, **Profit**, **ROI** (Return-on-Investment) oder **ROC** (Receiver Operator Characteristic).

Kumulatives Diagramm. Es wird ein kumulatives Diagramm erstellt. Die Werte in kumulativen Diagrammen werden für jedes Quantil zuzüglich aller höheren Quantile dargestellt. (**Kumulatives Diagramm** ist für ROC-Diagramme nicht verfügbar.)

Basis einschließen. In das Diagramm wird eine Basis aufgenommen, die auf eine völlig zufällige Verteilung der Treffer hinweist, sodass die Konfidenz irrelevant wird. (Bei Profit- und ROI-Diagrammen ist die Option **Basis einschließen** nicht verfügbar.)

Beste Linie einschließen. In das Diagramm wird eine beste Linie aufgenommen, die auf völlige Konfidenz hinweist (Treffer in 100 % aller Fälle). (**Beste Linie einschließen** ist für ROC-Diagramme nicht verfügbar.)

Für alle Diagrammtypen Profitkriterien verwenden. Wählen Sie diese Option aus, um bei der Berechnung der Evaluierungsmaße die Profitkriterien (Kosten, Ertrag und Gewichtung) anstatt der normalen Anzahl der Treffer zu verwenden. Für Modelle mit bestimmten numerischen Zielen, z. B. ein Modell, das den Ertrag vorhersagt, der als Reaktion auf ein Angebot durch einen Kunden erzielt wird, stellt der Wert des Zielfeldes ein besseres Maß der Leistung des Modells dar als die Anzahl der Treffer. Bei Auswahl dieser Option werden die Felder **Kosten**, **Ertrag** und **Gewichtung** für Gewinn-, Treffer- und Liftdiagramme aktiviert. Um die Profitkriterien für diese drei Diagrammtypen verwenden zu können, wird empfohlen, **Ertrag** als Zielfeld festzulegen, **Kosten** auf 0,0 zu setzen, sodass der Profit dem Ertrag entspricht, und für eine benutzerdefinierte Trefferbedingung "wahr" anzugeben, sodass alle Datensätze als Treffer gezählt werden. (**Für alle Diagrammtypen Profitkriterien verwenden** ist für ROC-Diagramme nicht verfügbar.)

Vorhergesagte Felder/Prädiktorfelder finden mithilfe von. Wählen Sie entweder **Metadaten des Modellausgabefelds**, um anhand der zugehörigen Metadaten nach den vorhergesagten Feldern im Diagramm zu suchen, oder wählen Sie **Format für Feldnamen** , um nach diesen Feldern anhand ihres Namens zu suchen.

Scorefelder grafisch darstellen. Aktivieren Sie dieses Kontrollkästchen, um das Auswahltool für Scorefelder zu aktivieren. Wählen Sie anschließend mindestens ein Bereichsscorefeld bzw. stetiges Scorefeld aus, also Felder, bei denen es sich nicht um Vorhersagemodelle im strengen Wortsinn handelt, die jedoch möglicherweise bei der Rangordnung der Datensätze hinsichtlich ihrer Trefferneigung nützlich sein könnten. Der Evaluierungsknoten kann alle Kombinationen von einem oder mehreren Scorefeldern mit einem oder mehreren Vorhersagemodellen vergleichen. Ein typisches Beispiel kann der Vergleich mehrerer RFM-Felder mit dem besten vorhandenen Vorhersagemodell sein.

Ziel. Wählen Sie mithilfe der Feldauswahl das Zielfeld aus. Sie können jedes beliebige instanziierte Flagfeld oder nominales Feld mit mindestens zwei Werten auswählen.

Hinweis: Dieses Zielfeld gilt nur für das Scoren von Feldern (Vorhersagemodelle definieren ihre eigenen Ziele) und wird ignoriert, wenn auf der Registerkarte "Optionen" ein Trefferkriterium festgelegt wurde.

Nach Partition aufteilen. Wenn Datensätze mithilfe eines Partitionsfelds in Trainings-, Test- und Validierungsstichproben aufgeteilt werden, lassen Sie mit dieser Option ein separates Evaluierungsdiagramm für die einzelnen Partitionen anzeigen. Weitere Informationen finden Sie im Thema „Partitionsknoten” auf Seite 154.

Hinweis: Wenn Sie eine Partition aufteilen, werden Datensätze mit Nullwerten im Partitionsfeld von der Auswertung ausgeschlossen. Dieses Problem tritt nicht auf, wenn Sie einen Partitionsknoten verwenden, weil diese Knoten keine Nullwerte erzeugen.

Diagramm. Wählen Sie die Größe der Quantile, die im Diagramm dargestellt werden sollen, in der Dropdown-Liste aus. Die folgenden Optionen stehen zur Auswahl: **Quartile**, **Quintile**, **Dezile**, **Vingtile**, **Perzentile** und **1000-tile**. (**Diagramm** ist für ROC-Diagramme nicht verfügbar.)

Stil. Wählen Sie die Option **Linie** oder **Punkt**.

Bei allen Diagrammtypen außer ROI-Diagrammen stehen weitere Steuerelemente zur Verfügung, mit denen Sie die Kosten, den Ertrag und die Gewichtungen festlegen können.

- **Kosten.** Geben Sie die Kosten für die einzelnen Datensätze an. Wählen Sie die Option **Fest** oder **Variabel** für den Ertrag. Bei festen Kosten geben Sie den Wert der Kosten ein. Bei variablen Kosten klicken Sie auf die Feldauswahlschaltfläche und bestimmen Sie ein Feld als Kostenfeld. (**Kosten** ist für ROC-Diagramme nicht verfügbar.)
- **Ertrag.** Geben Sie den Ertrag für die einzelnen Datensätze ein, die als Treffer gelten. Wählen Sie die Option **Fest** oder **Variabel** für den Ertrag. Bei einem festen Ertrag geben Sie den Wert des Ertrags ein. Bei einem variablen Ertrag klicken Sie auf die Feldauswahlschaltfläche und bestimmen Sie ein Feld als Ertragsfeld. (**Ertrag** ist für ROC-Diagramme nicht verfügbar.)
- **Gewichtung.** Wenn die Datensätze in den Daten für mehrere Einheiten stehen, können Sie die Ergebnisse mithilfe der Häufigkeitsgewichtungen anpassen. Geben Sie die Gewichtung für die einzelnen Datensätze im Feld **Fest** oder **Variabel** an. Bei einer festen Gewichtung geben Sie den Wert für die Gewichtung an (die Anzahl der Einheiten pro Datensatz). Bei variablen Gewichtungen klicken Sie auf die Schaltfläche für die Feldauswahl und bestimmen ein Feld als Gewichtungsfeld. (**Gewichtung** ist für ROC-Diagramme nicht verfügbar.)

Evaluierung - Registerkarte "Optionen"

Auf der Registerkarte "Optionen" für Evaluierungsdiagramme können Sie die Treffer, die Scoring-Kriterien und die Geschäftsregeln für das Diagramm auf flexible Weise festlegen. Des Weiteren stehen Optionen zur Verfügung, mit denen Sie die Ergebnisse der Modellauswertung exportieren.

Benutzerdefinierter Treffer. Geben Sie eine benutzerdefinierte Bedingung für einen Treffer an. Diese Option ist von Nutzen, wenn das relevante Ereignis definiert werden soll (also nicht aus dem Typ des Zielfelds und der Reihenfolge der Werte abgeleitet).

- **Bedingung.** Wenn Sie oben die Option **Benutzerdefinierter Treffer** wählen, muss ein CLEM-Ausdruck für eine Trefferbedingung festgelegt werden. Beispiel: `@TARGET = "JA"` ist eine gültige Bedingung, aus der hervorgeht, dass der Wert *Ja* im Zielfeld als Treffer bei der Auswertung gezählt wird. Die angegebene Bedingung wird für alle Zielfelder herangezogen. Geben Sie die gewünschte Bedingung in das Feld ein oder erzeugen Sie einen Bedingungsausdruck mit Expression Builder. Falls die Daten instanziiert wurden, können Sie die Werte direkt aus Expression Builder einfügen.

Benutzerdefinierter Score. Geben Sie eine Bedingung ein, mit der die Fälle gescort werden, bevor sie Quantilen zugeordnet werden. Der Standardscore wird aus dem vorhergesagten Wert und der Konfidenz berechnet. Im Feld "Ausdruck" können Sie einen benutzerdefinierten Scoring-Ausdruck erstellen.

- **Ausdruck.** Geben Sie einen CLEM-Ausdruck für das Scoring an. Wenn beispielsweise die numerische Ausgabe im Bereich 0-1 so geordnet wird, dass niedrige Werte besser eingestuft werden als hohe Werte, können Sie einen Treffer als `@TARGET < 0,5` definieren und den zugehörigen Score als `1 - @PREDICTED`. Der Score-Ausdruck muss in einem numerischen Wert resultieren. Geben Sie die gewünschte Bedingung in das Feld ein oder erzeugen Sie einen Bedingungsausdruck mit Expression Builder.

Geschäftsregel einschließen. Geben Sie eine Regelbedingung gemäß den relevanten Kriterien an. Beispiel: Es soll eine Regel für alle Fälle angezeigt werden, in denen gilt: `Kreditrate= "J" und Einkommen >= 33000`. Geschäftsregeln werden im Diagramm gezeichnet und im Schlüssel als *Regel* gekennzeichnet. (**Geschäftsregel einschließen** wird für ROC-Diagramme nicht unterstützt.)

- **Bedingung.** Geben Sie einen CLEM-Ausdruck zur Definition einer Geschäftsregel im Ausgabediagramm an. Geben Sie den gewünschten Bedingungsausdruck in das Feld ein oder erzeugen Sie einen Bedingungsausdruck mit Expression Builder. Falls die Daten instanziiert wurden, können Sie die Werte direkt aus Expression Builder einfügen.

Ergebnisse in Datei exportieren. Die Ergebnisse der Modellauswertung werden in eine Textdatei mit Trennzeichen exportiert. Sie können diese Datei einlesen und spezielle Analysen für die berechneten Werte vornehmen. Legen Sie die folgenden Optionen für den Export fest:

- **Dateiname.** Geben Sie den Dateinamen für die Ausgabedatei ein. Mit der Auslassungsschaltfläche **(...)** wechseln Sie zum gewünschten Ordner.

- **Trennzeichen.** Geben Sie das Zeichen ein (z. B. Komma oder Leerschritt), das als Feldtrennzeichen verwendet werden soll.

Feldnamen einschließen. Die Feldnamen werden in die erste Zeile der Ausgabedatei eingetragen.

Neue Zeile nach jedem Datensatz. Jeder Datensatz beginnt in einer neuen Zeile.

Evaluierung - Registerkarte "Darstellung"

Vor der Diagrammerstellung können Sie Darstellungsoptionen angeben.

Titel. Dient zur Eingabe des Texts, der als Titel des Diagramms verwendet werden soll.

Untertitel. Dient zur Eingabe des Texts, der als Untertitel des Diagramms verwendet werden soll.

Text. Akzeptieren Sie entweder die automatisch generierte Beschriftung oder wählen Sie **Angepasst**, um eine benutzerdefinierte Beschriftung anzugeben.

X-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *X*-Achsenbeschriftung (horizontal) oder wählen Sie **Angepasst** aus, um eine Beschriftung anzugeben.

Y-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *Y*-Achsenbeschriftung (vertikal) oder wählen Sie **Angepasst** aus, um eine Beschriftung anzugeben.

Rasterlinie anzeigen. Diese Option ist standardmäßig aktiviert. Hiermit lassen Sie Rasterlinien hinter dem Plot oder dem Diagramm einblenden, was die Bestimmung der Bereichs- und Bandabschnittpunkte erleichtert. Rasterlinien werden stets in weißer Farbe angezeigt; bei einem weißen Diagrammhintergrund erfolgt die Anzeige in Grau.

Lesen der Ergebnisse einer Modellauswertung

Die Interpretation eines Evaluierungsdiagramms ist zu einem gewissen Grad abhängig vom jeweiligen Diagrammtyp; einige Merkmale sind jedoch allen Evaluierungsdiagrammen gemeinsam. Bei kumulativen Diagrammen weisen höhere Linien auf bessere Modelle hin, insbesondere auf der linken Seite des Diagramms. Werden mehrere Modelle miteinander verglichen, schneiden sich die Linien häufig, sodass ein Modell in einem Teil des Diagramms höher ist und ein anderes Modell in einem anderen Diagrammteil. In diesem Fall sollten Sie den erforderlichen Teil der Stichprobe berücksichtigen (mit dem ein Punkt auf der *X*-Achse definiert wird), wenn Sie sich für ein bestimmtes Modell entscheiden.

Die meisten nicht kumulativen Diagramme sind einander sehr ähnlich. Bei guten Modellen sind nicht kumulative Diagramme auf der linken Seite des Diagramms hoch und auf der rechten Seite des Diagramms niedrig. (Zeigt ein nicht kumulatives Diagramm ein Sägezahnmuster, können Sie das Diagramm glätten, indem Sie die Anzahl der darzustellenden Quantile verringern und das Diagramm neu zeichnen lassen.) Ein Abfall auf der linken Seite des Diagramms oder eine Spitze auf der rechten Seite weist auf Bereiche hin, in denen das Modell nur wenig aussagekräftig ist. Eine gerade Linie über das ganze Diagramm entsteht, wenn ein Modell im Grunde genommen keinerlei Informationen liefert.

Gewinndiagramme. Kumulative Gewinndiagramme beginnen stets bei 0 %, verlaufen von links nach rechts und enden bei 100 %. Bei einem guten Modell steigt die Gewinntabelle steil in Richtung 100 % an und flacht dann ab. Bei einem Modell ohne Informationsgehalt verläuft eine diagonale Linie von links unten nach rechts oben. (Dies ist im Diagramm sichtbar, wenn Sie die Option **Basis einschließen** aktiviert haben.)

Liftdiagramme. Kumulative Liftdiagramme beginnen in der Regel bei einem Wert über 1,0 und fallen von links nach rechts allmählich ab. Die rechte Kante des Diagramms entspricht dem gesamten Dataset; das Verhältnis der Treffer in den kumulativen Quantilen zu den Treffern in den Daten beträgt 1,0. Bei einem guten Modell sollte der Lift auf der linken Seite deutlich über 1,0 beginnen, von links nach rechts auf ei-

nem hohen Niveau verbleiben und dann auf der rechten Seite des Diagramms abrupt auf 1,0 fallen. Bei einem Modell ohne Informationsgehalt liegt die Linie im gesamten Diagramm bei einem Wert um 1,0. (Falls die Option **Basis einschließen** aktiviert ist, wird im Diagramm eine horizontale Linie bei 1,0 als Referenz eingeblendet.)

Trefferdiagramme. Kumulative Trefferdiagramme besitzen große Ähnlichkeit mit Liftdiagrammen, mit Ausnahme der Skalierung. Trefferdiagramme beginnen in der Regel bei einem Wert nahe 100 % und fallen dann allmählich auf die Gesamttrefferrate (Gesamtanzahl der Treffer / Gesamtanzahl der Datensätze) auf der rechten Seite des Diagramms ab. Bei einem guten Modell beginnt die Linie auf der linken Seite genau oder nahe bei 100 %, von links nach rechts auf einem hohen Niveau verbleiben und dann auf der rechten Seite des Diagramms abrupt auf die Gesamttrefferrate fallen. Bei einem Modell ohne Informationsgehalt liegt die Linie im gesamten Diagramm bei einem Wert um die Gesamttrefferrate. (Falls die Option **Basis einschließen** aktiviert ist, wird im Diagramm eine horizontale Linie bei der Gesamttrefferrate als Referenz eingeblendet.)

Profitdiagramme. Kumulative Profitdiagramme zeigen die Summe der Profite, wenn Sie die Größe der ausgewählten Stichprobe (von links nach rechts) erhöhen. Profitdiagramme beginnen in der Regel in der Nähe von 0, steigen dann von links nach rechts stetig bis zu einer Spitze oder einem hohen Niveau in der Mitte an und fallen dann zur rechten Kante des Diagramms hin ab. Bei einem guten Modell zeigen die Profite eine klar ausgeprägte Spitze im Mittelteil des Diagramms. Bei einem Modell ohne Informationsgehalt verläuft die Linie relativ gerade; die Linie kann ansteigen, abfallen oder auf demselben Niveau verbleiben, abhängig von der vorliegenden Kosten-Umsatz-Struktur.

ROI-Diagramme. Kumulative ROI-Diagramme (Return-on-Investment) verlaufen in der Regel ähnlich wie Trefferdiagramme und Liftdiagramme, mit Ausnahme der Skalierung. ROI-Diagramme beginnen in der Regel bei einem Wert oberhalb von 0 % und fallen dann allmählich auf den Gesamt-ROI für das gesamte Dataset ab; dieser Wert kann durchaus auch negativ sein. Bei einem guten Modell sollte die Linie auf der linken Seite deutlich über 0 % beginnen, von links nach rechts auf einem hohen Niveau verbleiben und dann auf der rechten Seite des Diagramms relativ abrupt auf den Gesamt-ROI abfallen. Bei einem Modell ohne Informationsgehalt liegt die Linie im gesamten Diagramm beim Gesamt-ROI.

ROC-Diagramme. ROC-Kurven haben im Allgemeinen die Form eines kumulativen Gewinndiagramms. Von links nach rechts gesehen beginnt die Kurve bei der Koordinate (0,0) und endet bei der Koordinate (1,1). Ein Diagramm, das zur Koordinate (0,1) hin steil ansteigt und dann flach wird, gibt ein gutes Klassifikationsmerkmal an. Bei einem Modell, das Instanzen nach dem Zufallsprinzip als Treffer oder Fehlschläge klassifiziert, verläuft eine diagonale Linie von links unten nach rechts oben. (Dies ist im Diagramm sichtbar, wenn Sie die Option **Basis einschließen** aktiviert haben.) Wenn für ein Modell kein Konfidenzfeld angegeben ist, wird das Modell als einzelner Punkt dargestellt. Das Klassifikationsmerkmal mit dem optimalen Schwellenwert für die Klassifikation befindet sich am nächsten bei der Koordinate (0,1), bzw. in der linken oberen Ecke des Diagramms. Diese Position repräsentiert eine hohe Anzahl von Instanzen, die ordnungsgemäß als Treffer klassifiziert wurden, und eine niedrige Anzahl von Instanzen, die nicht ordnungsgemäß als Treffer klassifiziert wurden. Punkte über der diagonalen Linie stehen für gute Klassifikationsergebnisse. Punkte unter der diagonalen Linie stehen für schlechte Klassifikationsergebnisse, die schlechter sind, als wenn die Instanzen nach dem Zufallsprinzip klassifiziert würden.

Verwenden eines Evaluierungsdiagramms

Evaluierungsdiagramme können auf ähnliche Weise mithilfe der Maus untersucht werden wie Histogramme oder Sammlungsdiagramme. Die *X*-Achse bezeichnet die Modellscores in den angegebenen Quantilen (z. B. Vingtile oder Dezile).

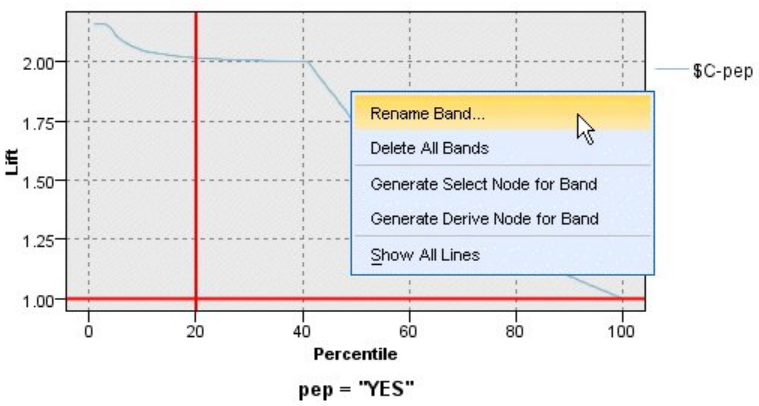


Abbildung 49. Arbeiten mit einem Evaluierungsdiagramm

Sie können die *X*-Achse wie bei Histogrammen in Abschnitte aufteilen. Greifen Sie über das Aufteilungssymbol auf die Optionen zu, mit denen die Achse automatisch in gleich große Abschnitte aufgeteilt wird. Weitere Informationen finden Sie im Thema „Exploration von Diagrammen". Sollen die Grenzen der Abschnitte manuell bearbeitet werden, wählen Sie im Menü "Bearbeiten" den Befehl **Diagrammabschnitte**.

Sobald Sie ein Evaluierungsdiagramm erstellt, Abschnitte definiert und die Ergebnisse untersucht haben, können Sie mit den Optionen im Menü "Generieren" und im Kontextmenü automatisch verschiedene Knoten auf der Grundlage der Auswahl im Diagramm erstellen. Weitere Informationen finden Sie im Thema „Generieren von Knoten aus Diagrammen" auf Seite 261.

Wenn Sie Knoten aus einem Evaluierungsdiagramm heraus erstellen, werden Sie aufgefordert, ein einzelnes Modell aus den verfügbaren Modellen im Diagramm auszuwählen.

Wählen Sie ein Modell aus und klicken Sie auf **OK**. Der neue Knoten wird im Streamerstellungsbereich erzeugt.

Exploration von Diagrammen

Während Sie im Bearbeitungsmodus Layout und Erscheinungsbild des Diagramms bearbeiten können, können Sie im Interaktionsmodus eine analytische Exploration der im Diagramm dargestellten Daten und Werte vornehmen. Das Hauptziel der Exploration besteht in der Analyse der Daten und der anschließenden Identifizierung von Werten mithilfe von Abschnitten, Bereichen und Markierungen zum Generieren von Auswahl-, Ableitungs- oder Balancierungsknoten. In diesen Modus wechseln Sie, indem Sie **Ansicht > Explorationsmodus** aus den Menüs auswählen (oder auf das entsprechende Symbol in der Symbolleiste klicken).

Bei einigen Diagrammen können alle Explorationstools verwendet werden, bei anderen dagegen ist nur ein einziges verfügbar. Der Interaktionsmodus umfasst folgende Aktionen:

- Definieren und Bearbeiten von Abschnitten, die zur Aufteilung der Werte entlang einer *x*-Skalenachse verwendet werden. Weitere Informationen finden Sie im Thema „Verwenden von Abschnitten" auf Seite 255.
- Definieren und Bearbeiten von Bereichen, die zur Identifizierung einer Gruppe von Werten innerhalb der rechteckigen Fläche verwendet werden. Weitere Informationen finden Sie im Thema „Verwenden von Bereichen" auf Seite 258.
- Markieren von Elementen (und Aufheben von Markierungen) zur manuellen Auswahl von Werten, die zum Generieren eines Auswahl- oder Ableitungsknotens verwendet werden könnten. Weitere Informationen finden Sie im Thema „Verwenden markierter Elemente" auf Seite 260.

- Generieren von Knoten mithilfe der durch Abschnitte, Bereiche, markierte Elemente und Netzlinks identifizierten Werte zur Verwendung im Stream. Weitere Informationen finden Sie im Thema „Generieren von Knoten aus Diagrammen” auf Seite 261.

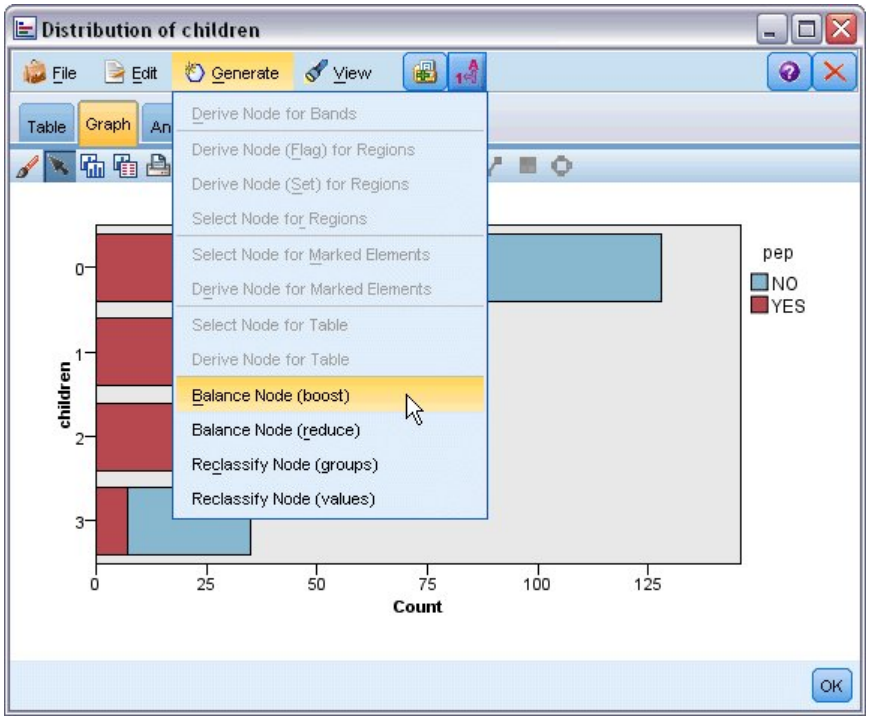


Abbildung 50. Diagramm mit Menü zum Generieren

Verwenden von Abschnitten

In jedem Diagramm mit einem metrischen Feld auf der *X*-Achse können Sie vertikale Abschnittslinien zeichnen, um den Wertebereich auf der *X*-Achse aufzuteilen. Bei aus mehreren Fenstern bestehenden Diagrammen wird eine Abschnittslinie, die in einem Fenster gezogen wird, auch in den anderen Fenstern angezeigt.

Nicht bei allen Diagrammen sind Abschnitte zulässig. Zu den Diagrammen, bei denen Abschnitte zulässig sind, gehören Histogramme, Balkendiagramme und Verteilungen, Plots (Linien-, Streu-, Zeitdiagramme usw.), Sammlungen und Evaluierungsdiagramme. Bei Diagrammen mit Einteilung in Felder werden die Abschnitte in allen Feldern angezeigt. In SPLOMs wird außerdem manchmal eine horizontale Abschnittslinie angezeigt, da die Achse, auf der der Feld-/Variablenabschnitt gezeichnet wurde, vertauscht wurde.

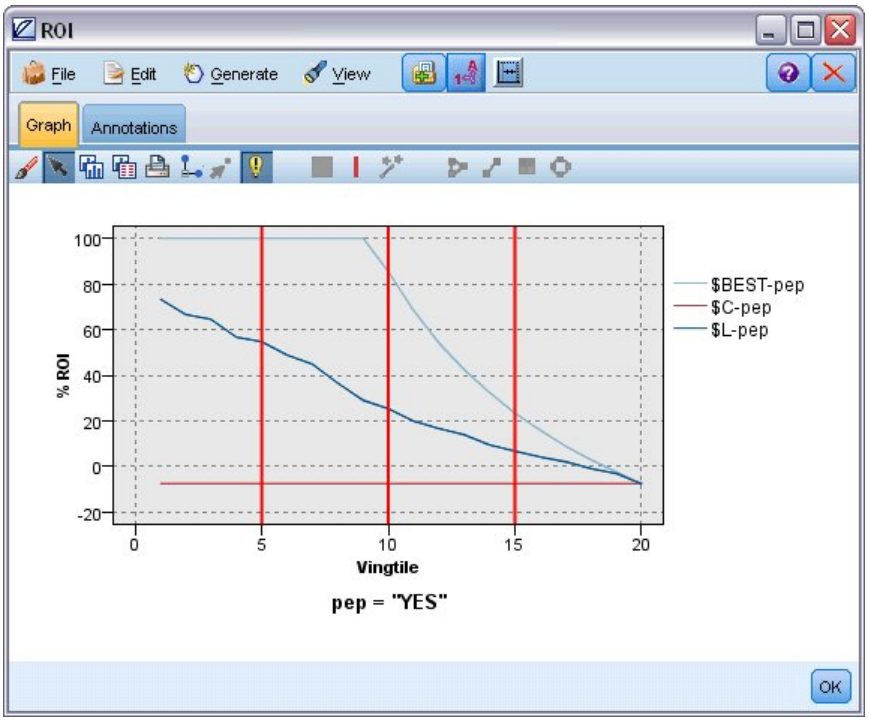


Abbildung 51. Diagramm mit drei Abschnitten

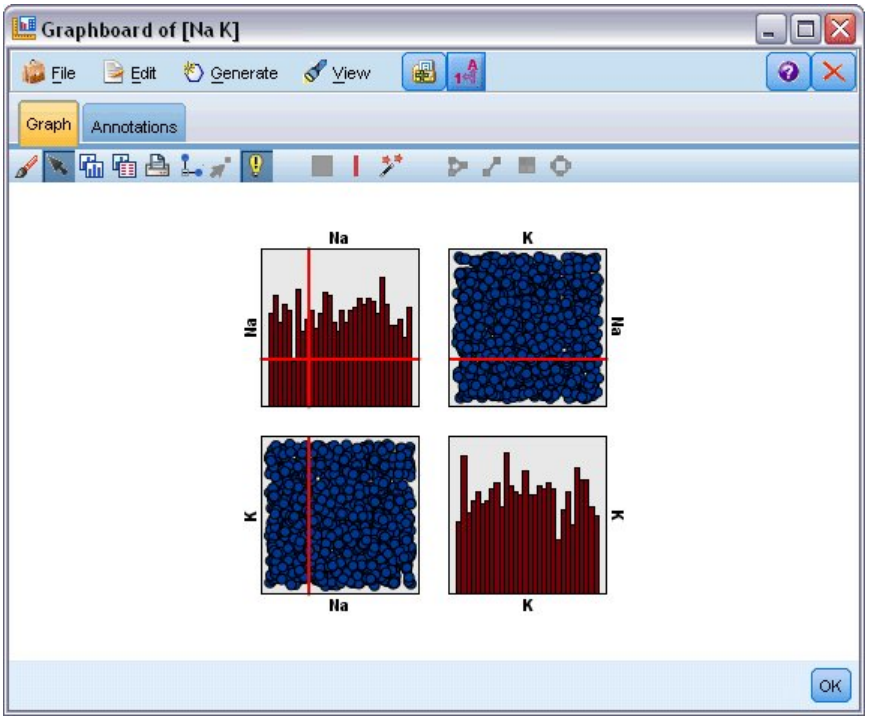


Abbildung 52. SPLOM mit Abschnitten

Definieren von Abschnitten

Diagramme ohne Abschnitte werden durch Einfügen einer Abschnittslinie in zwei Abschnitte aufgeteilt. Der Wert der Abschnittslinie stellt den Ausgangspunkt (auch als Untergrenze bezeichnet) des zweiten Ab-

schnitts dar, wenn das Diagramm von links nach rechts gelesen wird. Bei Diagrammen mit zwei Abschnitten wird durch Einfügen einer Abschnittslinie einer der Abschnitte geteilt, wodurch sich drei Abschnitte ergeben. Standardmäßig erhalten die Abschnitte die Bezeichnung *AbschnittN*, wobei *N* für die Anzahl der Abschnitte von links nach rechts auf der *X*-Achse steht.

Nach der Festlegung eines Abschnitts können Sie ihn durch Ziehen und Ablegen auf der *X*-Achse neu positionieren. Weitere Schnellverfahren können für Aufgaben wie Umbenennen, Löschen oder Generieren von Knoten für den betreffenden Abschnitt durch Rechtsklick innerhalb des Abschnitts angezeigt werden.

So definieren Sie Abschnitte:

1. Vergewissern Sie sich, dass Sie sich im Interaktionsmodus befinden. Wählen Sie in den Menüs die Optionsfolge **Ansicht > Interaktionsmodus**.
2. Klicken Sie in der Symbolleiste des Interaktionsmodus auf die Schaltfläche "Abschnitt zeichnen".

Abbildung 53. Symbolleistenschaltfläche "Abschnitt zeichnen".

3. Klicken Sie bei Diagrammen, bei denen Abschnitte zulässig sind, auf den Wertepunkt der *X*-Achse, an dem eine Abschnittslinie definiert werden soll.

Hinweis: Alternativ können Sie auf die Symbolleistenschaltfläche **Diagramm in Abschnitte teilen** klicken, die Anzahl der gewünschten gleich großen Abschnitte eingeben und auf **Aufteilen** klicken.

Abbildung 54. Aufteilungssymbol, mit dem die Symbolleiste um Optionen zum Aufteilen in Abschnitte erweitert wird

Abbildung 55. Symbolleiste zum Erstellen gleich großer Abschnitte mit aktivierten Abschnitten

Bearbeiten, Umbenennen und Löschen von Abschnitten

Die Eigenschaften bestehender Abschnitte können im Dialogfeld "Diagrammabschnitte bearbeiten" oder über die Kontextmenüs im Diagramm selbst bearbeitet werden.

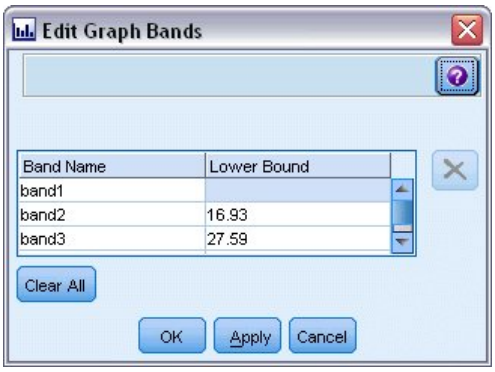


Abbildung 56. Dialogfeld "Diagrammabschnitte bearbeiten"

So bearbeiten Sie Abschnitte:

1. Vergewissern Sie sich, dass Sie sich im Interaktionsmodus befinden. Wählen Sie in den Menüs die Optionsfolge **Ansicht > Interaktionsmodus**.
2. Klicken Sie in der Symbolleiste des Interaktionsmodus auf die Schaltfläche "Abschnitt zeichnen".
3. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Diagrammabschnitte**. Das Dialogfeld "Diagrammabschnitte bearbeiten" wird geöffnet.
4. Wenn das Diagramm mehrere Felder enthält (beispielsweise bei SPLOM-Diagrammen), können Sie das gewünschte Feld in der Dropdown-Liste auswählen.
5. Sie können einen neuen Abschnitt hinzufügen, indem Sie einen Namen und eine Untergrenze eingeben. Drücken Sie die Eingabetaste, um eine neue Zeile zu beginnen.
6. Sie können die Grenze eines Abschnitts durch Anpassung des Werts für **Untergrenze** bearbeiten.
7. Abschnitte können durch Eingabe eines neuen Abschnittsnamens umbenannt werden.
8. Sie können Abschnitte löschen, indem Sie die Linie in der Tabelle auswählen und auf die Schaltfläche "Löschen" klicken.
9. Klicken Sie auf **OK**, um die Änderungen zu übernehmen und das Dialogfeld zu schließen.

Hinweis: Alternativ können Sie Abschnitte direkt im Diagramm löschen und umbenennen, indem Sie mit der rechten Maustaste auf die Linie des Abschnitts klicken und die gewünschte Option aus den Kontextmenüs auswählen.

Verwenden von Bereichen

In Diagrammen mit zwei Skalenachsen (oder Bereichsachsen) können Sie Bereiche zeichnen, um Werte innerhalb einer von Ihnen gezeichneten rechteckigen Fläche, dem sogenannten Bereich, zu gruppieren. Ein **Bereich** ist ein Teil des Diagramms, der durch einen bestimmten Mindest- und Höchstwert für *X* und *Y* beschrieben wird. Bei aus mehreren Fenstern bestehenden Diagrammen wird ein Bereich, der in einem Fenster gezeichnet wird, auch in den anderen Fenstern angezeigt.

Nicht bei allen Diagrammen sind Bereiche zulässig. Zu den Diagrammen, bei denen Bereiche zulässig sind, gehören: Plots (Linien-, Streu-, Blasen-, Zeitdiagramme usw.), SPLOM und Sammlungen. Diese Bereiche werden im X/Y-Raum gezeichnet und können daher nicht in 1-D-Plots, 3-D-Plots und animierten Plots definiert werden. Bei Diagrammen mit Einteilung in Felder werden die Bereiche in allen Feldern angezeigt. Bei einer Streudiagrammmatrix (SPLOM) wird ein Bereich in den zugehörigen oberen Plots angezeigt, nicht jedoch in den diagonalen Plots, da diese nur ein einziges metrisches Feld zeigen.

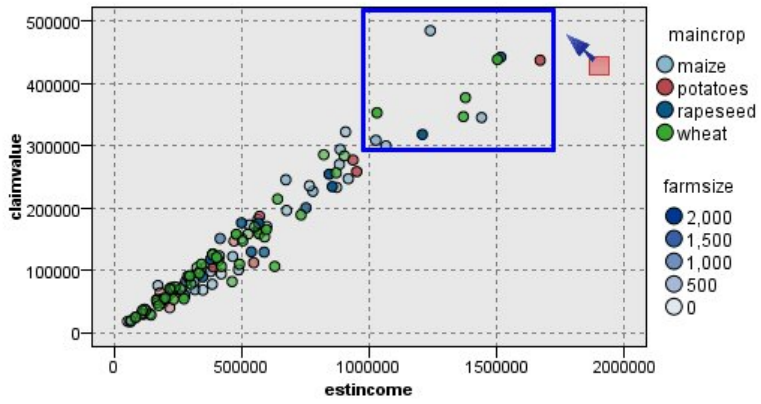


Abbildung 57. Definieren eines Bereichs mit hohen Forderungswerten

Definieren von Bereichen

Beim Definieren von Bereichen wird immer eine Gruppierung von Werten erstellt. Standardmäßig erhält jeder neue Bereich die Bezeichnung *Bereich<N>*, wobei *N* für die Anzahl der bereits erstellten Bereiche steht.

Nachdem Sie einen Bereich definiert haben, können Sie mit der rechten Maustaste auf die Bereichslinie klicken, um einige grundlegende Schnellverfahren anzuzeigen. Durch Rechtsklick innerhalb des Abschnitts (nicht auf die Linie) können zahlreiche weitere Schnellverfahren für Aufgaben wie Umbenennen, Löschen oder Generieren von Auswahl- und Ableitungsknoten für den betreffenden Bereich angezeigt werden.

Sie können Subsets von Datensätzen auf der Grundlage dessen auswählen, ob diese Datensätze in einem bestimmten Bereich oder in einem von mehreren Bereichen liegen. Des Weiteren können Sie Bereichsinformationen für einen Datensatz aufnehmen, indem Sie einen Ableitungsknoten erstellen, um Datensätze mit einem Flag zu versehen, basierend darauf, ob sie in einem bestimmten Bereich liegen. Weitere Informationen finden Sie im Thema „Generieren von Knoten aus Diagrammen" auf Seite 261.

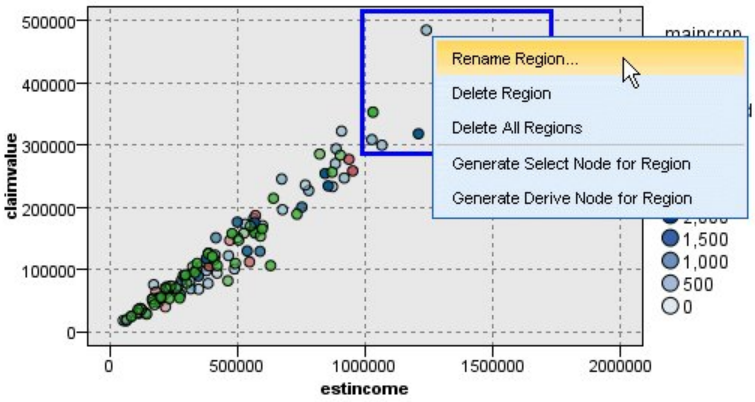


Abbildung 58. Untersuchen des Bereichs mit hohen Forderungswerten

So definieren Sie Bereiche:

1. Vergewissern Sie sich, dass Sie sich im Interaktionsmodus befinden. Wählen Sie in den Menüs die Optionsfolge **Ansicht > Interaktionsmodus**.
2. Klicken Sie in der Symbolleiste des Interaktionsmodus auf die Schaltfläche "Bereich zeichnen".

Abbildung 59. Symbolleistenschaltfläche "Bereich zeichnen".

3. Bei Diagrammen, bei denen Bereiche zulässig sind, können Sie den rechteckigen Bereich durch Klicken und Ziehen mit der Maus zeichnen.

Bearbeiten, Umbenennen und Löschen von Bereichen

Die Eigenschaften bestehender Bereiche können im Dialogfeld "Diagrammbereiche bearbeiten" oder über die Kontextmenüs im Diagramm selbst bearbeitet werden.

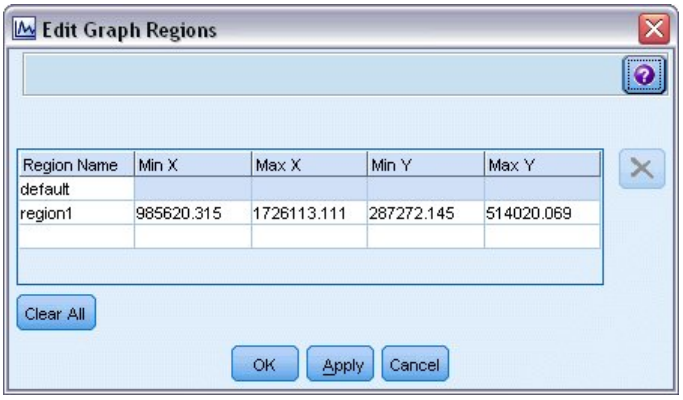


Abbildung 60. Festlegen von Eigenschaften für die definierten Bereiche

So bearbeiten Sie Bereiche:

1. Vergewissern Sie sich, dass Sie sich im Interaktionsmodus befinden. Wählen Sie in den Menüs die Optionsfolge **Ansicht > Interaktionsmodus**.
2. Klicken Sie in der Symbolleiste des Interaktionsmodus auf die Schaltfläche "Bereich zeichnen".
3. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Diagrammbereiche**. Das Dialogfeld "Diagrammbereiche bearbeiten" wird geöffnet.
4. Wenn das Diagramm mehrere Felder enthält (z. B. bei SPLOM-Diagrammen), müssen Sie das Feld für den Bereich in den Spalten *Feld A* und *Feld B* festlegen.
5. Ein neuer Bereich auf einer neuen Linie kann durch Eingabe eines Namens, (gegebenenfalls) Auswahl von Feldnamen und Festlegen der Ober- und Untergrenzen für jedes Feld hinzugefügt werden. Drücken Sie die Eingabetaste, um eine neue Zeile zu beginnen.
6. Mit den Werten **Min** und **Max** für *A* und *B* können Sie bestehende Bereichsgrenzen bearbeiten.
7. Zum Umbenennen von Bereichen wählen Sie den Namen des Bereichs in der Tabelle aus.
8. Sie können Bereiche löschen, indem Sie die Linie in der Tabelle auswählen und auf die Schaltfläche "Löschen" klicken.
9. Klicken Sie auf **OK**, um die Änderungen zu übernehmen und das Dialogfeld zu schließen.

Hinweis: Alternativ können Sie Bereiche direkt im Diagramm löschen und umbenennen, indem Sie mit der rechten Maustaste auf die Linie des Bereichs klicken und die gewünschte Option aus den Kontextmenüs auswählen.

Verwenden markierter Elemente

In jedem Diagramm können Elemente, wie Balken, Ausschnitte und Punkte, markiert werden. Linien, Flächen und Oberflächen können nur in Zeitdiagrammen, Multiplots und Evaluierungsdiagrammen markiert werden, da die Linien sich in diesen Fällen auf Felder beziehen. Bei der Markierung eines Elements he-

ben sie im Grunde alle Daten hervor, für die dieses Element steht. Bei Diagrammen, bei denen derselbe Fall an mehreren Stellen dargestellt wird (wie bei SPLOM) ist Markieren dasselbe wie Einfärben. Sie können Elemente in Diagrammen und sogar innerhalb von Abschnitten und Bereichen markieren. Wenn Sie ein Element markieren und anschließend wieder in den Bearbeitungsmodus wechseln, bleibt die Markierung weiterhin sichtbar.

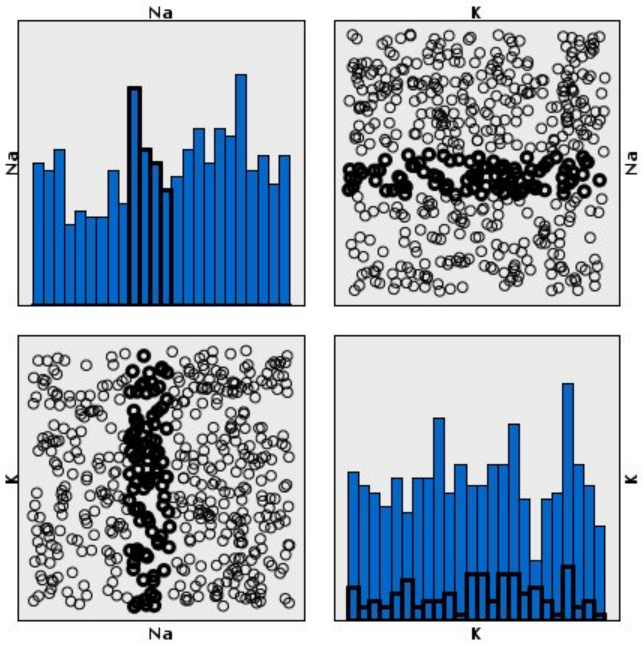


Abbildung 61. Markieren von Elementen in SPLOMs

Die Markierung von Elementen wird durch Klicken auf das jeweilige Element im Diagramm eingefügt und aufgehoben. Wenn Sie auf ein Element klicken, um es zu markieren, wird das Element mit einem dicken farbigen Rahmen angezeigt. Wenn Sie erneut auf das Element klicken, verschwindet der Rahmen und die Markierung des Elements ist aufgehoben. Um mehrere Elemente zu markieren, können Sie entweder beim Klicken auf die Elemente die Steuertaste gedrückt halten oder die Maus mit der Zauberstabfunktion über jedes der zu markierenden Elemente ziehen. Beachten Sie: Wenn Sie auf eine weitere Fläche oder ein weiteres Element klicken, ohne die Steuertaste gedrückt zu halten, wird die Markierung aller bisher ausgewählten Elemente aufgehoben.

Aus den markierten Elementen im Diagramm können Auswahl- und Ableitungsknoten generiert werden. Weitere Informationen finden Sie im Thema „Generieren von Knoten aus Diagrammen".

So markieren Sie Elemente:

1. Vergewissern Sie sich, dass Sie sich im Interaktionsmodus befinden. Wählen Sie in den Menüs die Optionsfolge **Ansicht > Interaktionsmodus**.
2. Klicken Sie in der Symbolleiste des Interaktionsmodus auf die Schaltfläche "Elemente markieren".
3. Klicken Sie auf das gewünschte Element oder klicken Sie und ziehen Sie mithilfe der Maus einen Rahmen um mehrere Elemente auf.

Generieren von Knoten aus Diagrammen

Eine der leistungsstärksten Funktionen von IBM SPSS Modeler-Diagrammen ist die Möglichkeit, Knoten aus einem Diagramm oder einer Auswahl innerhalb des Diagramms zu generieren. So können Sie beispielsweise aus einem Zeitdiagramm Ableitungs- und Auswahlknoten auf der Grundlage einer Datenaus-

wahl bzw. eines Datenbereichs generieren, wodurch ein Subset der Daten erstellt wird. Diese leistungsstarke Funktion kann beispielsweise zur Ermittlung und zum Ausschluss von Ausreißern verwendet werden.

Immer, wenn ein Abschnitt gezeichnet werden kann, kann auch ein Ableitungsknoten erstellt werden. Bei Diagrammen mit zwei Skalenachsen können Sie Ableitungs- bzw. Auswahlknoten aus den in Ihrem Diagramm gezeichneten Bereichen generieren. Bei Diagrammen mit markierten Elementen können Sie Ableitungsknoten, Auswahlknoten und in einigen Fällen Filterknoten aus diesen Elementen generieren. Die Generierung von Balancierungsknoten ist für alle Diagramme aktiviert, die eine Verteilung von Häufigkeiten (Anzahlwerten) anzeigen.

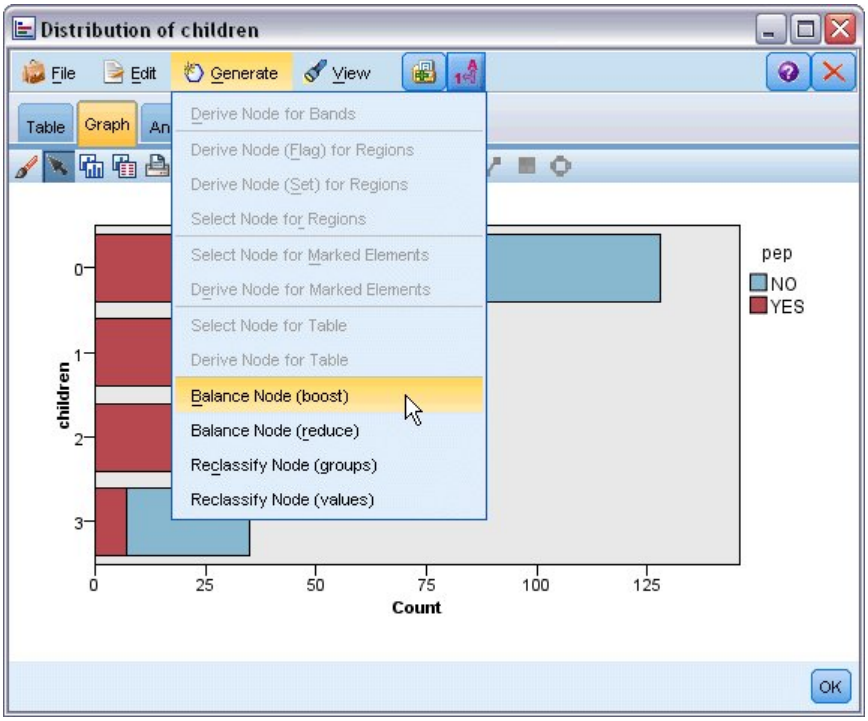


Abbildung 62. Diagramm mit Menü zum Generieren

Beim Generieren von Knoten wird der neue Knoten jeweils direkt im Streamerstellungsbereich platziert, sodass Sie ihn mit einem bestehenden Stream verbinden können. Aus Diagrammen können die folgenden Knotentypen generiert werden: Auswahl-, Ableitungs-, Balancierungs-, Filter- und Umcodierungsknoten.

Auswahlknoten

Auswahlknoten können generiert werden, um für die Verarbeitung im weiteren Streamverlauf einen Test auf Einschluss der Datensätze innerhalb eines Bereichs und Ausschluss aller Datensätze außerhalb des Bereichs (oder umgekehrt) durchzuführen.

- **Bei Abschnitten**. Sie können einen Auswahlknoten generieren, der die Datensätze innerhalb des betreffenden Abschnitts ein- bzw. ausschließt. **Auswahlknoten nur für Abschnitte** ist nur über Kontextmenüs verfügbar, da Sie auswählen müssen, welcher Abschnitt im Auswahlknoten verwendet werden soll.
- **Bei Bereichen**. Sie können einen Auswahlknoten generieren, der die Datensätze innerhalb des betreffenden Bereichs ein- bzw. ausschließt.
- **Bei markierten Elementen**. Sie können Auswahlknoten generieren, um die Datensätze zu erfassen, die den markierten Elementen bzw. Netzdiagrammzusammenhängen entsprechen.

Ableitungsknoten

Ableitungsknoten können aus Bereichen, Abschnitten und markierten Elementen generiert werden. Alle Diagramme können Ableitungsknoten erstellen. Bei Evaluierungsdiagrammen wird ein Dialogfeld zur Auswahl des Modells angezeigt. Bei Netzdiagrammen sind **Ableitungsknoten ("UND")** und **Ableitungsknoten ("ODER")** möglich.

- **Bei Abschnitten**. Sie können einen Ableitungsknoten generieren, der für jedes auf der Achse markierte Intervall eine Kategorie erstellt. Verwenden Sie hierzu die im Dialogfeld "Abschnitte bearbeiten" aufgeführten Abschnittsnamen als Kategorienamen.
- **Bei Bereichen**. Sie können einen Ableitungsknoten (**Ableitungstyp: Flag**) erstellen, der ein Flagfeld mit der Bezeichnung *Bereich (Flag)* erstellt, bei dem die Flags auf *T* (Datensätze in einem Bereich) bzw. *F* (Datensätze außerhalb aller Bereiche) gesetzt sind. Außerdem können Sie einen Ableitungsknoten (**Ableitungstyp: Set**) generieren, der ein Set mit einem Wert für jede Region mit einem neuen Feld *Bereich* für jeden Datensatz erstellt. Dabei wird der Name des Bereichs, in den die Datensätze fallen, als Wert verwendet. Datensätze, die außerhalb aller Bereiche liegen, erhalten den Namen des Standardbereichs. Wertenamen werden im Bearbeitungsdialogfeld für Bereiche als Bereichsnamen aufgelistet.
- **Bei markierten Elementen**. Sie können einen Ableitungsknoten generieren, der ein Flag berechnet, das für alle markierten Elemente *Wahr* und für alle anderen Datensätze *Falsch* ist.

Balancierungsknoten

Balancierungsknoten können generiert werden, um Unausgewogenheiten in den Daten zu korrigieren, beispielsweise durch Verringerung des Auftretens häufiger Werte (Menüoption **Balancierungsknoten (verringern)**) oder durch Erhöhen des Auftretens seltener Werte (Menüoption **Balancierungsknoten (erhöhen)**). Die Generierung von Balancierungsknoten wird für alle Diagramme aktiviert, die eine Häufigkeitsverteilung anzeigen, wie Histogramm, Punkt, Sammlung, Balken für Häufigkeiten, Kreis für Häufigkeiten und Multiplot.

Filterknoten

Filterknoten können generiert werden, um Felder anhand der im Diagramm markierten Linien bzw. Knoten umzubenennen oder zu filtern. Bei Evaluierungsdiagrammen generiert die Linie für die beste Anpassung keinen Filterknoten.

Umcodierungsknoten

Umcodierungsknoten können zum Umcodieren von Werten generiert werden. Diese Option wird für Verteilungsdiagramme verwendet. Sie können einen Umcodierungsknoten für **Gruppen** generieren, um bestimmte Werte eines angezeigten Felds in Abhängigkeit davon neu zu codieren, ob sie in einer Gruppe enthalten sind (mehrere Gruppen können durch Klicken bei gedrückter Steuertaste auf der Registerkarte "Tabellen" ausgewählt werden). Außerdem können Sie einen Umcodierungsknoten für **Werte** generieren, um Daten in ein bestehendes Set mit mehreren Werten neu zu codieren. Ein Beispiel hierfür ist die Umcodierung von Daten in ein Standardset von Werten, um Finanzdaten von mehreren Unternehmen zu Analysezwecken zusammenzuführen.

Hinweis: Sind die Werte bereits vordefiniert, können Sie sie als Flatfile in IBM SPSS Modeler einlesen und dann mithilfe einer Verteilung alle Werte anzeigen lassen. Anschließend können Sie einen Umcodierungsknoten (Werte) direkt aus dem Diagramm heraus erzeugen. Dadurch werden alle Zielwerte in die Spalte *Neue Werte* (Dropdown-Liste) im Umcodierungsknoten eingefügt.

Generieren von Knoten aus Diagrammen

Mithilfe des Menüs "Generieren" im Diagrammausgabefenster können Knoten generiert werden. Der generierte Knoten wird in den Streamerstellungsbereich platziert. Um den Knoten nutzen zu können, verbinden Sie ihn mit einem vorhandenen Stream.

So erzeugen Sie einen Knoten aus einem Diagramm:

1. Vergewissern Sie sich, dass Sie sich im Interaktionsmodus befinden. Wählen Sie in den Menüs die Optionsfolge **Ansicht > Interaktionsmodus**.
2. Klicken Sie in der Symbolleiste des Interaktionsmodus auf die Schaltfläche "Bereich".
3. Legen Sie die Abschnitte, Bereiche und markierten Elemente fest, die Sie zur Generierung des gewünschten Knotens benötigen.
4. Wählen Sie im Menü "Generieren" den gewünschten Knotentyp aus. Nur die zulässigen Knoten sind aktiviert.

Hinweis: Alternativ können Sie Knoten direkt im Diagramm generieren, indem Sie mit der rechten Maustaste klicken und die gewünschte Generierungsoption aus den Kontextmenüs auswählen.

Bearbeiten von Visualisierungen

Während Sie im Sondierungsmodus die durch die Visualisierung dargestellten Daten und Werte erforschen, ermöglicht Ihnen der Bearbeitungsmodus, das Layout und Aussehen der Visualisierung zu ändern. Sie können beispielsweise die Schriftarten und die Farben so ändern, dass sie den Stilvorgaben Ihres Unternehmens entsprechen. In diesen Modus wechseln Sie, indem Sie **Ansicht > Bearbeitungsmodus** aus den Menüs wählen (oder auf das entsprechende Symbol in der Symbolleiste klicken).

Im Bearbeitungsmodus stehen mehrere Symbolleisten zur Verfügung, mit denen sich die verschiedenen Aspekte des Visualisierungslayouts beeinflussen lassen. Wenn Sie einige der Symbolleisten nicht benötigen, können Sie sie ausblenden und so in dem Dialogfeld, in dem die Grafik angezeigt wird, mehr Platz schaffen. Um die Symbolleisten auszuwählen bzw. ihre Auswahl aufzuheben, klicken Sie im Menü "Ansicht" auf den Namen der entsprechenden Symbolleiste.

Hinweis: Um Ihre Visualisierungen mit weiteren Details zu versehen, können Sie Titel, Fußnoten und Achsenbeschriftungen zuweisen. Weitere Informationen finden Sie im Thema „Hinzufügen von Titeln und Fußnoten” auf Seite 277.

Zur Bearbeitung einer Visualisierung im **Bearbeitungsmodus** sind mehrere Optionen verfügbar. Sie verfügen über folgende Möglichkeiten:

- Bearbeiten und formatieren Sie den Text.
- Ändern Sie die Füllfarbe, die Transparenz und das Muster von Rahmen und Grafikelementen.
- Ändern Sie die Farbe und das Strichmuster von Rahmen und Linien.
- Drehen und ändern Sie die Form und das Verhältnis von Punktelementen.
- Ändern Sie die Größe von Grafikelementen (beispielsweise Balken und Punkte).
- Passen Sie den Platz um Elemente an, indem Sie Ränder und Abstände verwenden.
- Geben Sie Formate für Zahlen an.
- Ändern Sie die Einstellungen für Achsen und Skalierungen.
- Sortieren Sie Kategorien auf einer kategorialen Achse oder schließen Sie Kategorien aus oder blenden Sie sie aus.
- Legen Sie die Ausrichtung von Feldern fest.
- Weisen Sie einem Koordinatensystem Transformationen zu.
- Ändern Sie Statistiken, Grafikelementtypen und Kollisionsmodifikatoren.
- Ändern Sie die Position der Legende.
- Weisen Sie Visualisierungs-Style-Sheets hinzu.

In den folgenden Themen wird beschrieben, wie diese verschiedenen Aufgaben ausgeführt werden. Es wird auch empfohlen, dass Sie die allgemeinen Regeln zur Bearbeitung von Grafiken lesen.

So wechseln Sie in den Bearbeitungsmodus:

Wählen Sie die folgenden Befehle aus den Menüs aus:

Ansicht > **Bearbeitungsmodus**

Allgemeine Regeln zur Bearbeitung von Visualisierungen

Bearbeitungsmodus

Alle Bearbeitungen finden im Bearbeitungsmodus statt. Wählen Sie zum Aktivieren des Bearbeitungsmodus die folgenden Befehle aus den Menüs aus:

Ansicht > **Bearbeitungsmodus**

Auswahl

Die verfügbaren Optionen für die Bearbeitung hängen von der Auswahl ab. Abhängig von der Auswahl sind auch unterschiedliche Optionen der Symbolleiste und der Toolpalette aktiviert. Nur die aktivierten Elemente gelten für die aktuelle Auswahl. Wenn beispielsweise eine Achse ausgewählt ist, stehen im Fenster "Eigenschaften" die Registerkarten "Skala", "Hauptteilstriche" und "Hilfsteilstriche" zur Verfügung.

Hier finden Sie einige Tipps zur Auswahl von Elementen in der Visualisierung:

- Klicken Sie auf ein Element, um es auszuwählen.
- Wählen Sie ein Grafikelement (z. B. Punkte in einem Streudiagramm oder Balken in einem Balkendiagramm) mit einem einzigen Klick aus. Klicken Sie nach der anfänglichen Auswahl erneut, um die Auswahl auf Gruppen von Grafikelementen oder ein einziges Grafikelement einzuschränken.
- Drücken Sie Esc, um die Auswahl vollständig aufzuheben.

Paletten

Wenn ein Element in der Visualisierung ausgewählt ist, werden die verschiedenen Paletten passend zur Auswahl aktualisiert. Die Paletten enthalten Steuerungen für die Bearbeitung der Auswahl. Paletten können Symbolleisten oder ein Fenster mit mehreren Steuerungen und Registerkarten sein. Paletten können ausgeblendet sein, stellen Sie also sicher, dass die erforderliche Palette für die jeweilige Bearbeitung angezeigt ist. Prüfen Sie im Menü "Ansicht", welche Paletten derzeit angezeigt werden.

Sie können die Paletten umpositionieren, indem Sie in den leeren Bereich in der Symbolleistenpalette oder an der linken Seite anderer Paletten klicken und ihn an eine andere Stelle ziehen. Die Stellen, an denen Sie die Palette andocken können, werden visuell gekennzeichnet. Für Paletten, die keine Systemleisten sind, können Sie auch auf das Schließfeld klicken, um die Palette auszublenden, und auf die Schaltfläche zum Loslösen, um die Palette in einem separaten Fenster anzuzeigen. Klicken Sie auf die Schaltfläche "Hilfe", um Hilfe zur jeweiligen Palette zu erhalten.

Automatische Einstellungen

Für einige Einstellungen ist die Option **-auto-** verfügbar. Dies gibt an, dass automatisch Werte angewendet werden. Welche automatischen Einstellungen verwendet werden, hängt von der jeweiligen Visualisierung und den Datenwerten ab. Sie können einen Wert eingeben, um die automatische Einstellung zu überschreiben. Wenn Sie die automatische Einstellung wiederherstellen möchten, löschen Sie den aktuellen Wert und drücken Sie die Eingabetaste. Die Einstellung zeigt wieder **-auto-** an.

Ein- und Ausblenden von Elementen

Sie können zahlreiche Elemente in der Visualisierung ein- oder ausblenden. Sie können beispielsweise die Legende oder die Achsenbeschriftung ausblenden. Um ein Element zu löschen, wählen Sie es aus und drücken Sie die Löschtaste. Wenn das Element nicht gelöscht werden darf, geschieht nichts. Wenn Sie ein Element versehentlich löschen, drücken Sie Strg+Z, um den Löschvorgang zu widerrufen.

Status

Einige Symbolleisten zeigen den Status der aktuellen Auswahl an, andere nicht. Die Eigenschaftenpalette zeigt immer den Status an. Wenn eine Symbolleiste den Status *nicht* reflektiert, ist dies im Thema zur Beschreibung der Symbolleiste vermerkt.

Bearbeiten und Formatieren von Text

Sie können Text an seiner Position bearbeiten und die Formatierung eines ganzen Textblocks ändern. Beachten Sie, dass Sie keinen Text bearbeiten können, der direkt mit Datenwerten verknüpft ist. Sie können beispielsweise keine Teilstrichbeschriftungen bearbeiten, da der Inhalt einer Beschriftung aus den zugrunde liegenden Daten ermittelt wird. Sie können jedoch jeden Text in der Visualisierung formatieren.

So bearbeiten Sie Text direkt:

1. Doppelklicken Sie auf den Textblock. Durch diese Aktion wird der gesamte Text ausgewählt. Alle Symbolleisten werden inaktiviert, da Sie während der Bearbeitung von Text keinen anderen Bereich der Visualisierung ändern können.
2. Geben Sie den Text ein, der den vorhandenen ersetzen soll. Sie können auch erneut auf den Text klicken, damit ein Cursor angezeigt wird. Positionieren Sie den Cursor an der gewünschten Stelle und geben Sie weiteren Text ein.

So formatieren Sie Text:

1. Wählen Sie den Rahmen aus, der den Text enthält. Doppelklicken Sie nicht auf den Text.
2. Formatieren Sie den Text mithilfe der Symbolleiste für Schriftarten. Wenn die Symbolleiste nicht aktiviert ist, stellen Sie sicher, dass nur der *Rahmen*, der den Text enthält, ausgewählt ist. Wenn der Text ausgewählt ist, ist die Symbolleiste inaktiviert.

Folgende Schriftänderungen sind möglich:

- Farbe
- Schriftfamilie (z. B. Arial oder Verdana)
- Größe (Die Einheit ist Punkt, sofern Sie keine andere Einheit angeben.)
- Stärke
- Ausrichtung in Relation zum Textrahmen

Die Formatierung gilt für den gesamten Text im Rahmen. Die Formatierung einzelner Buchstaben oder Wörter in einem bestimmten Textblock lässt sich nicht ändern.

Ändern von Farben, Mustern, Strichmustern und Transparenz

Viele verschiedene Elemente in einer Visualisierung verfügen über eine Füllung und einen Rahmen. Ein deutliches Beispiel ist ein Balken in einem Balkendiagramm. Die Farbe der Balken ist die Füllfarbe. Sie können auch von einem durchgehenden schwarzen Rahmen eingefasst sein.

In der Visualisierung gibt es weniger deutliche Beispiele mit Füllfarben. Wenn die Füllfarbe transparent ist, wissen Sie womöglich gar nicht, dass eine Füllung vorhanden ist. Betrachten Sie beispielsweise den Text einer Achsenbeschriftung. Dieser Text sieht wie frei positionierter Fließtext aus, aber in Wirklichkeit befindet er sich in einem Rahmen mit transparenter Füllfarbe. Sie können den Rahmen sehen, wenn Sie die Achsenbeschriftung auswählen.

Jeder Rahmen in einer Visualisierung kann über eine Füllung und einen Umrandungsstil verfügen, auch der Rahmen um die gesamte Visualisierung. Zudem ist mit jedem Füllmuster ein Grad der Lichtdurchlässigkeit/Transparenz verbunden, der sich anpassen lässt.

So ändern Sie Farben, Muster, Strichmuster und Transparenz:

1. Wählen Sie das Element aus, das Sie formatieren möchten. Wählen Sie beispielsweise die Balken in einem Balkendiagramm oder einen Textrahmen aus. Wenn die Visualisierung durch eine kategoriale Variable oder ein Feld getrennt ist, können Sie auch die Gruppe auswählen, die einer individuellen Kategorie entspricht. Hiermit können Sie die Standardformatierung für diese Gruppe ändern. Sie können beispielsweise die Farbe von einer der Stapelgruppen in einem gestapelten Balkendiagramm ändern.
2. Verwenden Sie die Symbolleiste "Farbe", um die Füllfarbe, die Rahmenfarbe oder das Füllmuster zu ändern.

   Hinweis: Diese Symbolleiste reflektiert nicht den Status der aktuellen Auswahl.

   Sie können eine Farbe oder Füllung ändern, indem Sie auf die Schaltfläche klicken, um die angezeigte Option auszuwählen, oder auf den Dropdown-Pfeil, um eine andere Option zu wählen. Beachten Sie, dass eine der Farboptionen wie eine rote diagonale Linie auf weißem Grund aussieht. Dies ist die transparente Farbe. Sie können sie beispielsweise verwenden, um den Rahmen von Balken in einem Histogramm auszublenden.
   - Die erste Schaltfläche steuert die Füllfarbe. Wenn die Farbe einem stetigen oder ordinalen Feld zugeordnet ist, ändert diese Schaltfläche die Füllfarbe für die Farbe, die mit dem höchsten Wert in den Daten verknüpft ist. Sie können die Registerkarte "Farbe" im Fenster "Eigenschaften" verwenden, um die Farbe zu ändern, die mit dem niedrigsten Wert und mit fehlenden Daten verknüpft ist. Die Farbe der Elemente ändert sich mit zunehmenden Werten für die zugrunde liegenden Daten inkrementell von der Farbe für den untersten Wert bis zur Farbe für den obersten Wert.
   - Die zweite Schaltfläche steuert die Rahmenfarbe.
   - Die dritte Schaltfläche steuert das Füllmuster. Das Füllmuster verwendet die Rahmenfarbe. Daher ist das Füllmuster nur sichtbar, wenn es eine sichtbare Rahmenfarbe gibt.
   - Die vierte Steuerung besteht aus einem Schieberegler und einem Textfeld und dient zur Steuerung der Opazität der Füllfarbe und des Füllmusters. Ein geringerer Prozentsatz bedeutet weniger Opazität und mehr Transparenz. 100 % ist völlig undurchlässig (keine Transparenz).
3. Um das Strichmuster eines Rahmens oder einer Linie zu ändern, verwenden Sie die Liniensymbolleiste.

   Hinweis: Diese Symbolleiste reflektiert nicht den Status der aktuellen Auswahl.

   Wie bei der anderen Symbolleiste können Sie auf die Schaltfläche klicken, um die angezeigte Option auszuwählen, oder auf den Dropdown-Pfeil, um eine andere Option zu wählen.

Drehen und Ändern der Form und des Verhältnisses von Punktelementen

Sie können Punktelemente drehen, ihnen eine andere vordefinierte Form zuweisen oder das Seitenverhältnis (Breite zu Höhe) ändern.

So ändern Sie Punktelemente:

1. Wählen Sie die Punktelemente aus. Sie können die Form und das Seitenverhältnis einzelner Punktelemente nicht drehen oder ändern.
2. Verwenden Sie die Piktogramm-Symbolleiste, um die Punkte zu ändern.
   - Mit der ersten Schaltfläche können Sie die Form der Punkte ändern. Klicken Sie auf den Dropdown-Pfeil und wählen Sie eine vordefinierte Form aus.
   - Mit der zweiten Schaltfläche können Sie die Punkte an eine bestimmte Kompassposition drehen. Klicken Sie auf den Pfeil nach unten und ziehen Sie dann die Nadel an die gewünschte Position.
   - Mit der dritten Schaltfläche können Sie das Seitenverhältnis ändern. Klicken Sie auf den Dropdown-Pfeil und ziehen Sie dann das angezeigte Rechteck in die gewünschte Form. Die Form des Rechtecks repräsentiert das Seitenverhältnis.

Ändern der Größe von Grafikelementen

Sie können die Größe der Grafikelemente in der Visualisierung ändern. Hierzu gehören Balken, Punkte, Linien usw. Wenn die Größe des Grafikelements durch eine Variable oder ein Feld eingestellt wird, handelt es sich dabei um die *Mindestgröße*.

So ändern Sie die Größe von Grafikelementen:

1. Wählen Sie die Grafikelemente aus, deren Größe Sie ändern möchten.
2. Verwenden Sie den Schieberegler oder geben Sie eine bestimmte Größe für die Option in der Piktogrammsymbolleiste ein. Die Einheit ist Pixel, sofern Sie keine andere Einheit angeben. (Eine vollständige Liste der Einheitenabkürzungen siehe unten.) Sie können auch einen Prozentwert (z. B. 30 %) eingeben, um festzulegen, dass ein Grafikelement den angegebenen prozentualen Anteil des verfügbaren Platzes einnimmt. Der verfügbare Platz hängt vom Typ des Grafikelements und der spezifischen Visualisierung ab.

Tabelle 46. Gültige Abkürzungen für Einheiten

Abkürzung	**Einheit**
cm	Zentimeter
in	Zoll
mm	Millimeter
pc	Pica
pt	Punkt
px	Pixel

Festlegen von Rändern und Abständen

Wenn um oder im Rahmen in der Visualisierung zu viel oder zu wenig Platz ist, können Sie seine Rand- und Abstandeinstellungen ändern. Der **Rand** ist die Menge an Platz zwischen dem Rahmen und benachbarten anderen Elementen. Der **Abstand** ist die Menge an Platz zwischen der Begrenzungslinie des Rahmens und dem *Inhalt* des Rahmens.

So legen Sie Ränder und Abstände fest:

1. Wählen Sie den Rahmen aus, für den Sie Ränder und Abstände festlegen möchten. Dabei kann es sich um einen Textrahmen, den Rahmen um eine Legende oder sogar den Datenrahmen handeln, der die Grafikelemente (z. B. Balken und Punkte) anzeigt.
2. Verwenden Sie die Registerkarte "Ränder" der Palette "Eigenschaften", um die Einstellungen anzugeben. Die Größeneinheit ist Pixel, sofern Sie keine andere Einheit angeben (z. B. cm oder in).

Formatieren von Zahlen

Sie können das Format für Zahlen in Teilstrichbeschriftungen auf einer fortlaufenden Achse sowie bei Datenwertbeschriftungen mit Zahlen angeben. Beispielsweise können Sie festlegen, dass in den Teilstrichbeschriftungen angegebene Zahlen in Tausendern gezeigt werden.

So geben Sie Zahlenformate an:

1. Wählen Sie die Teilstrichbeschriftungen der fortlaufenden Achse oder die Datenwertbeschriftungen aus, wenn sie Zahlen enthalten.
2. Klicken Sie im Fenster "Eigenschaften" auf die Registerkarte **Format**.
3. Wählen Sie die gewünschten Zahlenformatoptionen aus:

   Präfix. Ein Zeichen, das vor der Zahl angezeigt werden soll. Geben Sie beispielsweise ein Dollarzeichen ($) ein, wenn es sich bei den Zahlen um Gehälter in US-Dollar handelt.

Suffix. Ein Zeichen, das nach der Zahl angezeigt werden soll. Geben Sie beispielsweise ein Prozentzeichen (%) ein, wenn es sich bei den Zahlen um Prozentsätze handelt.

Min. Ganzzahlstellen. Mindestanzahl an Stellen, die im ganzzahligen Teil der Dezimaldarstellung angezeigt werden sollen. Wenn der tatsächliche Wert nicht über die Mindestanzahl an Stellen verfügt, wird der ganzzahlige Teil des Wertes mit Nullen aufgefüllt.

Max. Ganzzahlstellen. Höchstanzahl an Stellen, die im ganzzahligen Teil der Dezimaldarstellung angezeigt werden sollen. Wenn der tatsächliche Wert die Höchstanzahl an Stellen überschreitet, wird der ganzzahlige Teil des Wertes durch Sterne ersetzt.

Min. Dezimalstellen. Mindestanzahl an Stellen, die im Dezimalteil der Dezimal- oder wissenschaftlichen Darstellung angezeigt werden sollen. Wenn der tatsächliche Wert nicht über die Mindestanzahl an Stellen verfügt, wird der Dezimalteil des Wertes mit Nullen aufgefüllt.

Max. Dezimalstellen. Höchstanzahl an Stellen, die im Dezimalteil der Dezimal- oder wissenschaftlichen Darstellung angezeigt werden sollen. Wenn der tatsächliche Wert die Höchstanzahl an Stellen überschreitet, wird der Dezimalwert auf die passende Anzahl an Stellen gerundet.

Wissenschaftlich. Ob Zahlen in wissenschaftlicher Notation angezeigt werden sollen. Die wissenschaftliche Notation ist für sehr große oder sehr kleine Zahlen sinnvoll. **-auto-** überlässt es der Anwendung zu entscheiden, ob wissenschaftliche Notation angemessen ist.

Skalierung. Ein Skalierungsfaktor, der eine Zahl ist, durch die der Originalwert dividiert wird. Verwenden Sie einen Skalierungsfaktor, wenn die Zahlen groß sind, Sie jedoch nicht möchten, dass die Teilstrichbeschriftung durch die Anzeige der großen Zahl übermäßig lang wird. Wenn Sie das Zahlenformat der Teilstrichbeschriftungen ändern, bearbeiten Sie auch den Achsentitel, um anzugeben, wie die Zahl zu interpretieren ist. Nehmen wir an, auf der Skalenachse werden Gehälter angezeigt und die Beschriftungen lauten 30.000, 50.000 und 70.000. Hier können Sie einen Skalierungsfaktor von `1.000` eingeben, um 30, 50 und 70 anzuzeigen. In diesem Fall wäre es sinnvoll, den Skalenachsentitel so zu bearbeiten, dass er Text "in Tausend" enthält.

Klammern für -ve. Ob negative Werte eingeklammert werden sollen.

Gruppierung. Ob ein bestimmtes Zeichen zwischen Zifferngruppen angezeigt werden soll. Die aktuelle Ländereinstellung Ihres Computers bestimmt, welches Zeichen zur Zifferngruppierung verwendet wird.

Ändern der Einstellungen für Achsen und Skalen

Für das Ändern von Achsen und Skalen stehen mehrere Optionen zur Verfügung.

So ändern Sie die Achsen- und Skaleneinstellungen

1. Wählen Sie einen beliebigen Teil der Achse aus (z. B. die Achsenbeschriftung oder Teilstrichbeschriftungen).
2. Verwenden Sie die Registerkarten "Skala", "Hauptteilstriche" und "Hilfsteilstriche" im Fenster "Eigenschaften", um die Einstellungen für Achsen und Skalen zu ändern.

Registerkarte "Skala"

Hinweis: Die Registerkarte "Skala" wird bei Diagrammen mit voraggregierten Daten (z. B. Histogrammen) nicht angezeigt.

Typ. Legt fest, ob die Skala linear oder transformiert ist. Skalentransformationen sind hilfreich beim Verständnis der Daten und bei den für Statistiken erforderlichen Annahmen. Bei Streudiagrammen können Sie eine transformierte Skala verwenden, wenn die Beziehung zwischen den unabhängigen und den abhängigen Variablen oder Feldern nicht linear ist. Skalentransformationen können außerdem verwendet werden, um ein schiefes Histogramm symmetrischer zu machen, sodass es eher wie eine Normalverteilung aussieht. Beachten Sie, dass Sie lediglich die Skala transformieren, auf der die Daten angezeigt werden, nicht jedoch die eigentlichen Daten.

- **linear.** Gibt eine lineare, nicht transformierte Skala an.

- **Log.** Gibt eine mit dem Logarithmus zur Basis 10 transformierte Skala an. Um Nullwerte und negative Werte zu berücksichtigen, verwendet diese Transformation eine modifizierte Version der Logarithmusfunktion. Diese sichere Logarithmusfunktion wird definiert als `sign(x) * log(1 + abs(x))`. Also ist `safeLog(-99)` gleich:

  ```
  sign(-99) * log(1 + abs(-99)) = -1 * log(1 + 99) = -1 * 2 = -2
  ```

- **Exponent.** Gibt eine exponententransformierte Skala mit einem Exponenten von 0,5 an. Um negative Werte zu berücksichtigen, verwendet diese Transformation eine modifizierte Version der Exponentialfunktion. Diese sichere Exponentialfunktion wird definiert als `sign(x) * pow(abs(x), 0,5)`. Also ist `safePower(-100)` gleich:

  ```
  sign(-100) * pow(abs(-100), 0,5) = -1* pow(100, 0,5) = -1 * 10 = -10
  ```

Min/Max/Guter unterer Wert/Guter oberer Wert. Gibt den Bereich für die Skala an. Wenn **Guter unterer Wert** und **Guter oberer Wert** ausgewählt sind, kann die Anwendung auf der Basis der Daten eine geeignete Skala auswählen. Es handelt sich um "gute" untere und obere Werte, da sie in der Regel um einige Werte größer oder kleiner als die Höchst- und Mindestdatenwerte sind. Wenn die Daten beispielsweise von 4 bis 92 reichen, kann ein guter unterer und oberer Wert für die Skala 0 bzw. 100 sein, anstatt die tatsächlichen Mindest- und Höchstwerte aus den Daten zu übernehmen. Achten Sie darauf, dass Sie keinen Bereich festlegen, der zu klein ist und wichtige Elemente verdeckt. Beachten Sie auch, dass Sie keinen expliziten Mindest- und Höchstwert festlegen können, wenn die Option **Null einschließen** aktiviert ist.

Unterer Rand/Oberer Rand. Erstellen Sie Ränder am unteren und/oder oberen Ende der Achse. Der Rand erscheint senkrecht zur ausgewählten Achse. Die Größeneinheit ist Pixel, sofern Sie keine andere Einheit angeben (z. B. cm oder in). Wenn Sie beispielsweise den **oberen Rand** für die vertikale Achse als 5 festlegen, verläuft ein horizontaler Rand in der Breite von 5 Pixel am oberen Rand des Datenrahmens.

Umkehren. Legt fest, ob die Skala umgekehrt ist.

Null einschließen. Gibt an, dass die Skala 0 einschließen soll. Diese Option wird in der Regel für Balkendiagramme verwendet, um sicherzustellen, dass die Balken bei 0 beginnen anstatt bei einem Wert im Bereich der Höhe des kleinsten Balkens. Wenn diese Option aktiviert ist, sind **Min** und **Max** inaktiviert, da Sie keinen benutzerdefinierten Mindest- oder Höchstwert für den Skalenbereich festlegen können.

Registerkarten "Hauptteilstriche" und "Hilfsteilstriche"

Teilstriche oder **Einteilungsstriche** sind die Striche auf einer Achse. Sie geben Werte in bestimmten Intervallen oder Kategorien an. **Hauptteilstriche** sind die Einteilungsstriche mit Beschriftung. Sie sind auch länger als andere Einteilungsstriche. **Hilfsteilstriche** sind Einteilungsstriche zwischen den Hauptteilstrichen. Einige Optionen hängen von der Art der Einteilungsstriche ab, aber die meisten Optionen sind für Hauptteilstriche und Hilfsteilstriche verfügbar.

Teilstriche anzeigen. Gibt an, ob in einem Diagramm Haupt- oder Hilfsteilstriche angezeigt werden sollen.

Rasterlinien einblenden. Gibt an, ob an den Haupt- oder Hilfsteilstrichen Rasterlinien angezeigt werden. **Rasterlinien** sind Linien, die auf einem ganzen Diagramm parallel zu den Achsen verlaufen.

Position. Legt die Position der Teilstriche relativ zur Achse fest.

Länge. Legt die Länge der Teilstriche fest. Die Größeneinheit ist Pixel, sofern Sie keine andere Einheit angeben (z. B. cm oder in).

Basis. *Gilt nur für Hauptteilstriche.* Gibt den Wert an, bei dem der erste Hauptteilstrich erscheint.

Delta *Gilt nur für Hauptteilstriche.* Gibt den Abstand zwischen Hauptteilstrichen an. Das bedeutet, dass ein Hauptteilstrich bei jedem *n*. Wert erscheint, wobei *n* den Deltawert angibt.

Unterteilungen. *Gilt nur für Hilfsteilstriche.* Gibt die Anzahl der Hilfsteilstriche zwischen Hauptteilstrichen an. Die Anzahl der Hilfsteilstriche ist um 1 weniger als die Anzahl der Unterteilungen. Angenommen, bei 0 und 100 befinden sich Hauptteilstriche. Wenn Sie für die Anzahl der Hilfsteilstrichunterteilungen 2 eingeben, wird *ein* Hilfsteilstrich bei *50* angezeigt, der den Bereich von 0 bis 100 in *zwei* Bereiche unterteilt.

Bearbeiten von Kategorien

Sie können die Kategorien auf einer kategorialen Achse auf mehrere Arten bearbeiten:

- Die Sortierfolge für die Anzeige der Kategorien kann geändert werden.
- Bestimmte Kategorien können ausgeschlossen werden.
- Sie können Kategorien hinzufügen, die nicht im Dataset angezeigt werden.
- Kleinere Kategorien lassen sich in einer Kategorie zusammenfassen oder kombinieren.

So ändern Sie die Sortierfolge von Kategorien:

1. Wählen Sie eine Kategorienachse aus. Die Kategorienpalette zeigt die Kategorien auf der Achse an.
   Hinweis: Wenn das Fenster nicht sichtbar ist, stellen Sie sicher, dass Sie es aktiviert haben. Wählen Sie in IBM SPSS Modeler im Menü "Ansicht" die Option **Kategorien** aus.
2. Wählen Sie in der Kategorienpalette eine Sortierreihenfolge aus der Dropdown-Liste aus.
   Benutzerdefiniert. Sortierung der Kategorien in der Reihenfolge, in der sie in der Palette angezeigt werden. Verschieben Sie die Kategorien mithilfe der Pfeilschaltflächen an den Beginn der Liste, nach oben, nach unten oder an das Ende der Liste.
   Daten. Sortierung der Kategorien in der Reihenfolge, in der sie im Dataset angezeigt werden.
   Name. Sortierung der Kategorien in alphabetischer Reihenfolge nach den Namen, die im Fenster angezeigt werden. Dabei kann es sich um den Wert oder die Beschriftung handeln, abhängig davon, ob in der Symbolleiste die Schaltfläche zur Anzeige von Werten oder Beschriftungen ausgewählt ist.
   Wert. Sortierung der Kategorien nach dem zugrunde liegenden Datenwert unter Verwendung der Werte, die im Fenster eingeklammert angezeigt werden. Nur Datenquellen mit Metadaten (z. B. IBM SPSS Statistics Datendateien) unterstützen diese Option.
   Statistik Sortierung von Kategorien auf der Basis der berechneten Statistik für jede Kategorie. Beispiele für Statistiken sind Anzahl, Prozentsatz und Mittelwert. Diese Option ist nur verfügbar, wenn im Diagramm eine Statistik verwendet wird.

So fügen Sie eine Kategorie hinzu:

Standardmäßig sind nur Kategorien verfügbar, die im Dataset angezeigt werden. Sie können der Visualisierung bei Bedarf eine Kategorie hinzufügen.

1. Wählen Sie eine Kategorienachse aus. Die Kategorienpalette zeigt die Kategorien auf der Achse an.
   Hinweis: Wenn das Fenster nicht sichtbar ist, stellen Sie sicher, dass Sie es aktiviert haben. Wählen Sie in IBM SPSS Modeler im Menü "Ansicht" die Option **Kategorien**.
2. Klicken Sie im Fenster "Kategorien" auf die Schaltfläche "Kategorie hinzufügen":

Abbildung 63. Schaltfläche "Kategorie hinzufügen"

3. Geben Sie im Dialogfeld "Neue Kategorie hinzufügen" einen Namen für die Kategorie ein.
4. Klicken Sie auf **OK**.

So schließen Sie bestimmte Kategorien aus:

1. Wählen Sie eine Kategorienachse aus. Die Kategorienpalette zeigt die Kategorien auf der Achse an.

Hinweis: Wenn das Fenster nicht sichtbar ist, stellen Sie sicher, dass Sie es aktiviert haben. Wählen Sie in IBM SPSS Modeler im Menü "Ansicht" die Option **Kategorien**.

2. Wählen Sie in der Kategorienpalette in der Liste "Einschließen" einen Kategorienamen aus und klicken Sie dann auf die Schaltfläche "X". Um die Kategorie wieder zurückzuverschieben, wählen Sie ihren Namen in der Liste "Ausgeschlossen" aus und klicken dann auf den Pfeil rechts neben der Liste.

So kombinieren Sie kleinere Kategorien bzw. fassen sie zusammen:

Sie können Kategorien zusammenfassen, die so klein sind, dass sie nicht gesondert angezeigt werden müssen. Bei einem Kreisdiagramm mit vielen Kategorien könnten beispielsweise alle Kategorien mit einem Prozentsatz von weniger als 10 Prozent zusammengefasst werden. Eine Zusammenfassung ist nur für additive Statistiken verfügbar. Sie können beispielsweise keine Mittelwerte addieren, da Mittelwerte nicht additiv sind. Daher ist das Kombinieren bzw. Zusammenfassen von Kategorien mithilfe eines Mittelwerts nicht möglich.

1. Wählen Sie eine Kategorienachse aus. Die Kategorienpalette zeigt die Kategorien auf der Achse an.
   Hinweis: Wenn das Fenster nicht sichtbar ist, stellen Sie sicher, dass Sie es aktiviert haben. Wählen Sie in IBM SPSS Modeler im Menü "Ansicht" die Option **Kategorien**.
2. Wählen Sie in der Kategorienpalette die Option **Reduzieren** aus und geben Sie einen Prozentsatz an. Alle Kategorien, deren prozentualer Anteil unter dem angegebenen Wert liegt, werden zu einer Kategorie zusammengefasst. Der Prozentsatz beruht auf der im Diagramm gezeigten Statistik. Die Zusammenfassung ist nur für häufigkeitsbasierte und Summenstatistiken verfügbar.

Ändern der Ausrichtung von Feldern

Wenn Sie in Ihrer Visualisierung Felder verwenden, können Sie deren Ausrichtung ändern.

So ändern Sie die Ausrichtung der Felder:

1. Klicken Sie auf einen beliebigen Bereich der Visualisierung.
2. Klicken Sie im Fenster "Eigenschaften" auf die Registerkarte **Felder**.
3. Wählen Sie eine Option aus **Layout**:

Tabelle. Ordnet Felder wie eine Tabelle an, indem jedem einzelnen Wert eine Zeile oder Spalte zugewiesen wird.

Transponiert. Ordnet Fenster wie eine Tabelle an und vertauscht außerdem die Daten der ursprünglichen Zeilen und Spalten. Diese Option ist nicht identisch mit dem Transponieren des Diagramms. Beachten Sie, dass die *X*-Achse und die *Y*-Achse unverändert bleiben, wenn Sie diese Option auswählen.

Liste. Ordnet Fenster wie eine Liste an, in der jede Zelle eine Kombination von Werten darstellt. Spalten und Zeilen sind keinen einzelnen Werten mehr zugewiesen. Diese Option ermöglicht, dass die Felder bei Bedarf umbrechen.

Transformieren des Koordinatensystems

Viele Visualisierungen werden in einem flachen, rechteckigen Koordinatensystem angezeigt. Sie können das Koordinatensystem wie erforderlich umformen. Sie können dem Koordinatensystem beispielsweise eine polare Transformation zuweisen, schräge Schatteneffekte hinzufügen und die Achsen transponieren. Sie können diese Transformationen auch rückgängig machen, wenn sie der aktuellen Visualisierung bereits zugewiesen sind. Beispiel: Ein Kreisdiagramm ist in einem Polarkoordinatensystem gezeichnet. Sie können die Polartransformation rückgängig machen und das Kreisdiagramm als einzelnes gestapeltes Balkendiagramm in einem rechteckigen Koordinatensystem anzeigen.

So transformieren Sie das Koordinatensystem:

1. Wählen Sie das zu transformierende Koordinatensystem aus. Sie wählen das Koordinatensystem aus, indem Sie den Rahmen um das individuelle Diagramm auswählen.

2. Klicken Sie im Fenster "Eigenschaften" auf die Registerkarte **Koordinaten**.
3. Wählen Sie die Transformationen aus, die Sie dem Koordinatensystem zuweisen möchten. Sie können die Auswahl einer Transformation auch aufheben, um sie rückgängig zu machen.

   Transponiert. Das Ändern der Ausrichtung der Achsen wird als **Transponieren** bezeichnet. Es gleicht dem Vertauschen der vertikalen und horizontalen Achsen in einer 2-D-Visualisierung.

   Polar. Eine Polartransformation zeichnet die Grafikelemente in einem bestimmten Winkel und Abstand vom Mittelpunkt des Diagramms. Ein Kreisdiagramm ist eine 1-D-Visualisierung mit einer Polartransformation, die einzelne Balken in bestimmten Winkeln zeichnet. Ein Radardiagramm ist eine 2-D-Visualisierung mit einer Polartransformation, die Grafikelemente in bestimmten Winkeln und Abständen von der Mitte des Diagramms zeichnet. Eine 3-D-Visualisierung würde eine zusätzliche Tiefendimension umfassen.

   Schräg. Eine schräge Transformation fügt den Grafikelementen einen 3-D-Effekt hinzu. Diese Transformation verleiht den Grafikelementen mehr Tiefe, die aber rein dekorativen Zwecken dient. Sie wird durch keine bestimmten Datenwerte beeinflusst.

   Gleiches Verhältnis. Wenn Sie dasselbe Verhältnis zuweisen, repräsentiert derselbe Abstand auf jeder Skale denselben Abstand in den Datenwerten. Beispielsweise repräsentieren 2 cm auf beiden Skalen einen Abstand von 1.000.

   % Einsatz vor der Transformation. Wenn Achsen nach der Transformation abgeschnitten sind, sollten Sie dem Diagramm Einsätze hinzufügen, bevor Sie die Transformation zuweisen. Die Einsätze schrumpfen die Abmessungen um einen bestimmten Prozentsatz, bevor dem Koordinatensystem etwaige Transformationen zugewiesen werden. Sie können die unteren *x*-, oberen *x*-, unteren *y*- und oberen *y*-Abmessungen in dieser Reihenfolge steuern.

   % Einsatz nach der Transformation. Wenn Sie das Seitenverhältnis eines Diagramms ändern möchten, können Sie dem Diagramm Einsätze hinzufügen, nachdem die Transformation angewendet wurde. Die Einsätze schrumpfen die Abmessungen um einen bestimmten Prozentsatz, nachdem dem Koordinatensystem etwaige Transformationen zugewiesen wurden. Diese Einsätze können auch zugewiesen werden, wenn dem Diagramm keine Transformation zugewiesen wird. Sie können die unteren *x*-, oberen *x*-, unteren *y*- und oberen *y*-Abmessungen in dieser Reihenfolge steuern.

Ändern von Statistiken und Grafikelementen

Sie können ein Grafikelement in einen anderen Typ umwandeln, die zum Zeichnen des Grafikelements verwendete Statistik ändern oder den Kollisionsmodifikator angeben, der bestimmt, was geschieht, wenn sich Grafikelemente überlagern.

So konvertieren Sie ein Grafikelement:

1. Wählen Sie das Grafikelement aus, das Sie konvertieren möchten.
2. Klicken Sie im Fenster "Eigenschaften" auf die Registerkarte **Element**.
3. Wählen Sie einen neuen Grafikelementtyp aus der Liste "Typ".

Tabelle 47. Grafikelementtypen

Grafikelementtyp	Beschreibung
Punkt	Eine Markierung, die den jeweiligen Datenpunkt identifiziert. Ein Punktelement wird in Streudiagrammen und anderen verwandten Visualisierungen verwendet.
Intervallskala	Eine rechteckige, an einem bestimmten Datenwert gezeichnete Form, die den Bereich zwischen einem Ursprung und einem anderen Datenwert ausfüllt. Ein Intervallelement wird in Balkendiagrammen und Histogrammen verwendet.
Linie	Eine Linie, die Datenwerte verbindet.
Pfad	Eine Linie, die Datenwerte in der Reihenfolge ihres Auftretens im Dataset verbindet.

Tabelle 47. Grafikelementtypen (Forts.)

Grafikelementtyp	Beschreibung
Fläche	Eine Linie, die Datenelemente mit der Fläche zwischen der Linie und einem Ursprung verbindet.
Polygon	Eine mehrseitige Form, die einen Datenbereich umschließt. Ein Polygonelement könnte in einem klassierten Streudiagramm oder in einer Karte verwendet werden.
Schema	Ein Element, bestehend aus einer Rechteck mit sogenannten Whiskers und Markierungen, die auf Ausreißer hinweisen. Ein Schemaelement wird für Boxplots verwendet.

So ändern Sie die Statistik:

1. Wählen Sie das Grafikelement aus, dessen Statistik Sie ändern möchten.
2. Klicken Sie im Fenster "Eigenschaften" auf die Registerkarte **Element**.
3. Wählen Sie aus der Dropdown-Liste "Auswertung" eine neue Statistik. Beachten Sie, dass bei der Auswahl einer Statistik die Daten zusammengefasst werden. Wenn die Visualisierung stattdessen nicht zusammengefasste Daten anzeigen soll, wählen Sie **(keine Statistik)** aus der Liste "Auswertung".

Aus einem stetigen Feld berechnete Auswertungsstatistiken

- *Mittelwert*. Ein Lagemaß (zentrale Tendenz). Die Summe der Ränge, geteilt durch die Zahl der Fälle.
- *Median*. Wert, über und unter dem jeweils die Hälfte der Fälle liegt; 50. Perzentil. Bei einer geraden Anzahl von Fällen ist der Median der Mittelwert der beiden mittleren Fälle, wenn diese auf- oder absteigend sortiert sind. Der Median ist ein Lagemaß, das gegenüber Ausreißern unempfindlich ist (im Gegensatz zum Mittelwert, der durch wenige extrem niedrige oder hohe Werte beeinflusst werden kann).
- *Modalwert*. Der am häufigsten auftretende Wert. Wenn mehrere Werte gleichermaßen die größte Häufigkeit aufweisen, ist jeder von ihnen ein Modalwert.
- *Minimum*. Der kleinste Wert einer numerischen Variablen.
- *Maximum*. Der größte Wert einer numerischen Variablen.
- *Bereich*. Differenz zwischen Mindest- und Höchstwert.
- *Mittelbereich*. Der Mittelpunkt des Bereichs, also der Wert, dessen Differenz vom Mindestwert gleich seiner Differenz vom Höchstwert ist.
- *Summe*. Die Summe der Werte über alle Fälle mit nicht fehlenden Werten.
- *Kumulative Summe*. Die kumulative Summe der Werte. In jedem Grafikelement wird die Summe für eine Untergruppe plus der Gesamtsumme aller früheren Gruppen angezeigt.
- *Prozent Summe*. Der Prozentsatz innerhalb der einzelnen Untergruppen, beruhend auf einem summierten Feld im Vergleich zur Summe über alle Gruppen hinweg.
- *Kumulativer Prozentwert Summe*. Der kumulativer Prozentsatz innerhalb jeder Untergruppe basierend auf einem summierten Feld im Vergleich zur Summe über alle Gruppen hinweg. In jedem Grafikelement wird die der Prozentsatz für eine Untergruppe plus dem Gesamtprozentsatz aller früheren Gruppen angezeigt.
- *Varianz*. Ein Maß der Streuung um den Mittelwert, gleich der Summe der quadrierten Abweichungen vom Mittelwert geteilt durch eins weniger als die Anzahl der Fälle. Die Maßeinheit der Varianz ist das Quadrat der Maßeinheiten der Variablen.
- *Standardabweichung*. Ein Maß für die Streuung um den Mittelwert. In einer Normalverteilung liegen 68 % der Fälle innerhalb von einer Standardabweichung des Mittelwerts und 95 % der Fälle innerhalb von zwei Standardabweichungen. Wenn beispielsweise für das Alter der Mittelwert 45 und die Standardabweichung 10 beträgt, liegen bei einer Normalverteilung 95 % der Fälle im Bereich zwischen 25 und 65.
- *Standardfehler*. Ein Maß für die Abweichung des Werts einer Teststatistik zwischen Stichproben. Dies ist die Standardabweichung der Stichprobenverteilung einer Statistik. So ist z. B. der Standardfehler des Mittelwerts die Standardabweichung des Stichprobenmittelwerts.

- *Kurtosis*. Ein Maß dafür, wie sehr die Beobachtungen um einen zentralen Punkt gruppiert sind. Bei einer Normalverteilung ist der Wert der Kurtosis gleich 0. Bei positiver Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung enger um das Zentrum der Verteilung gruppiert und haben dünnere Flanken bis hin zu den Extremwerten der Verteilung. Ab dort sind die Flanken der leptokurtischen Verteilung im Vergleich zu einer Normalverteilung dicker. Bei negativer Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung weniger eng gruppiert und haben dickere Flanken bis hin zu den Extremwerten der Verteilung. Ab dort sind die Flanken der platykurtischen Verteilung im Vergleich zu einer Normalverteilung dünner.
- *Schiefe*. Ein Maß der Asymmetrie der Verteilung. Die Normalverteilung ist symmetrisch, ihre Schiefe hat den Wert 0. Eine Verteilung mit einer deutlichen positiven Schiefe läuft nach rechts lang aus (lange rechte Flanke). Eine Verteilung mit einer deutlichen negativen Schiefe läuft nach links lang aus (lange linke Flanke). Als Faustregel kann man verwenden, dass ein Schiefewert, der mehr als doppelt so groß ist wie sein Standardfehler, als Abweichung von der Symmetrie gilt.

Die folgenden Bereichsstatistiken können zu mehreren Grafikelementen pro Untergruppe führen. Bei Verwendung der Grafikelemente für Intervalle, Flächen oder Ränder führt eine Bereichsstatistik zu einem Grafikelement, das den Bereich zeigt. Alle anderen Grafikelemente führen zu zwei getrennten Elementen, einem, das den Beginn des Bereichs zeigt, und einem mit dem Endbereich.

- **Bereich: Bereich**. Der Bereich der Werte zwischen dem Mindest- und dem Höchstwert.
- **Bereich: 95-%-Konfidenzintervall für den Mittelwert**. Ein Wertebereich, der mit einer Wahrscheinlichkeit von 95 % den Mittelwert der Grundgesamtheit enthält.
- **Bereich: 95-%-Konfidenzintervall für einzelne Fälle**. Ein Wertebereich, der mit einer Wahrscheinlichkeit von 95 % den Wert vorhergesagten Wert für den Einzelfall enthält.
- **Bereich: 1 Standardabweichung über/unter dem Mittelwert**. Ein Wertebereich zwischen 1 *Standardabweichung* oberhalb und unterhalb des*Mittelwerts*.
- **Bereich: 1 Standardfehler über/unter dem Mittelwert**. Ein Wertebereich zwischen 1 *Standardfehler* oberhalb und unterhalb des*Mittelwerts*.

Anzahlbasierte Auswertungsstatistiken

- **Anzahl**. Die Anzahl der Zeilen/Fälle.
- **Kumulative Anzahl**. Die kumulative Anzahl der Zeilen/Fälle. In jedem Grafikelement wird die Anzahl für eine Untergruppe plus der Gesamtanzahl aller früheren Gruppen angezeigt.
- **Prozent Anzahl**. Der Prozentsatz der Zeilen/Fälle in jeder Untergruppe im Vergleich zur Gesamtzahl der Zeilen/Fälle.
- **Kumulativer Prozentwert Anzahl**. Der kumulative Prozentsatz der Zeilen/Fälle in jeder Untergruppe im Vergleich zur Gesamtzahl der Zeilen/Fälle. In jedem Grafikelement wird die der Prozentsatz für eine Untergruppe plus dem Gesamtprozentsatz aller früheren Gruppen angezeigt.

So geben Sie den Kollisionsmodifikator an:

Der Kollisionsmodifikator bestimmt, was geschieht, wenn sich Grafikelemente überlagern.

1. Wählen Sie das Grafikelement aus, für das Sie den Kollisionsmodifikator angeben möchten.
2. Klicken Sie im Fenster "Eigenschaften" auf die Registerkarte **Element**.
3. Wählen Sie aus dem Dropdown-Listenfeld "Modifikator" einen Kollisionsmodifikator aus. **-auto-** überlässt es der Anwendung zu bestimmen, welcher Kollisionsmodifikator für den Grafikelementtyp und die Statistik geeignet ist.

   Überlagert. Zeichnet Grafikelemente mit demselben Wert übereinander.

   Stapeln. Stapelt Grafikelemente, die einander normalerweise überdecken würden, wenn sie dieselben Datumswerte besitzen.

Ausweichen. Verschiebt Grafikelemente neben andere Grafikelemente, die am gleichen Wert erscheinen, anstatt sie übereinanderzudecken. Die Grafikelemente werden symmetrisch angeordnet. D. h. die Grafikelemente werden an entgegengesetzten Seiten einer zentralen Position verschoben. Ausweichen ist dem Clustering sehr ähnlich.

Übereinander. Verschiebt Grafikelemente neben andere Grafikelemente, die am gleichen Wert erscheinen, anstatt sie übereinanderzudecken. Die Grafikelemente werden asymmetrisch angeordnet. D. h., die Grafikelemente werden schräg übereinander angeordnet, wobei das untere Grafikelement an einem bestimmten Wert auf der Skala positioniert ist.

Streuen (normal). Positioniert Grafikelemente an demselben Datenwert anhand einer Normalverteilung zufällig um.

Streuen (uniform). Positioniert Grafikelemente an demselben Datenwert anhand einer Gleichverteilung zufällig um.

Ändern der Position der Legende

Wenn ein Diagramm eine Legende umfasst, wird diese in der Regel rechts neben dem Diagramm angezeigt. Sie können diese Position bei Bedarf ändern.

So ändern Sie die Position der Legende:

1. Wählen Sie die Legende aus.
2. Klicken Sie im Fenster "Eigenschaften" auf die Registerkarte **Legende**.
3. Wählen Sie eine Position aus.

Kopieren von Visualisierungen und Visualisierungsdaten

Die Palette "Allgemein" enthält Schaltflächen zum Kopieren der Visualisierung und ihrer Daten.

Abbildung 64. Schaltfläche "Visualisierung kopieren"

Kopieren der Visualisierung. Diese Aktion kopiert die Visualisierung als Bild in die Zwischenablage. Mehrere Bildformate stehen zur Verfügung. Wenn Sie das Bild in eine andere Anwendung einfügen, können Sie eine Option "Inhalte einfügen" auswählen, um eines der verfügbaren Bildformate zum Einfügen festzulegen.

Abbildung 65. Schaltfläche "Visualisierungsdaten kopieren"

Kopieren der Visualisierungsdaten. Diese Aktion kopiert die zugrunde liegenden Daten, anhand denen die Visualisierung gezeichnet wird. Die Daten werden als einfacher Text oder HTML-Text in die Zwischenablage kopiert. Wenn Sie die Daten in eine andere Anwendung einfügen, können Sie eine Option "Inhalte einfügen" wählen, um eines dieser Formate zum Einfügen festzulegen.

Tastenkombinationen

Tabelle 48. Tastenkombinationen

Tastenkombination	Funktion
STRG + LEERTASTE	Umschalten zwischen Interaktions- und Bearbeitungsmodus

Tabelle 48. Tastenkombinationen (Forts.)

Tastenkombination	Funktion
Löschen	Löschen eines Visualisierungselements
Strg+Z	Rückgängig
Strg+Y	Wiederholen
F2	Anzeigen des Umrisses zur Auswahl von Elementen im Diagramm

Hinzufügen von Titeln und Fußnoten

Bei allen Diagrammtypen können Sie einen eindeutigen Titel, eine Fußnote oder Achsenbeschriftungen hinzufügen, um deutlicher zu machen, was im Diagramm angezeigt wird.

Hinzufügen von Titeln zu Diagrammen

1. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Diagrammtitel hinzufügen** aus. Ein Textfeld, das **<TITLE>** enthält, wird oberhalb des Diagramms angezeigt.
2. Vergewissern Sie sich, dass Sie sich im Bearbeitungsmodus befinden. Wählen Sie in den Menüs die Optionsfolge **Ansicht > Bearbeitungsmodus** aus.
3. Doppelklicken Sie auf den Text **<TITLE>**.
4. Geben Sie den gewünschten Titel ein und drücken Sie die Eingabetaste.

Hinzufügen von Fußnoten zu Diagrammen

1. Wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Diagrammfußnote hinzufügen** aus. Ein Textfeld, das **<FOOTNOTE>** enthält, wird unterhalb des Diagramms angezeigt.
2. Vergewissern Sie sich, dass Sie sich im Bearbeitungsmodus befinden. Wählen Sie in den Menüs die Optionsfolge **Ansicht > Bearbeitungsmodus** aus.
3. Doppelklicken Sie auf den Text **<FOOTNOTE>**.
4. Geben Sie den gewünschten Titel ein und drücken Sie die Eingabetaste.

Verwenden von Diagramm-Style-Sheets

Die Grundlegenden Informationen zur Anzeige von Diagrammen, wie Farben, Schriftarten, Symbole und Linienstärke werden über ein Style-Sheet festgelegt. Im Lieferumfang von IBM SPSS Modeler ist ein Standard-Style-Sheet enthalten; Sie können jedoch, falls erforderlich, Änderungen daran vornehmen. Beispielsweise gilt möglicherweise in Ihrem Unternehmen ein Farbschema für Präsentationen, das Sie auch in Ihren Diagrammen verwenden möchten. Weitere Informationen finden Sie im Thema „Bearbeiten von Visualisierungen” auf Seite 264.

In den Diagrammknoten können Sie mithilfe des Bearbeitungsmodus Stiländerungen am Erscheinungsbild eines Diagramms vornehmen. Anschließend können Sie im Menü **Bearbeiten > Stile** die Änderungen als Style-Sheet speichern, das auf alle Diagramme angewendet wird, die Sie danach aus dem Diagrammknoten generieren, oder als neues Standard-Style-Sheet, das für alle Diagramme gilt, die Sie mit IBM SPSS Modeler erstellen.

Im Menü "Bearbeiten" stehen über die Option **Stile** fünf Style-Sheet-Optionen zur Verfügung:

- **Style-Sheet wechseln.** Damit wird eine Liste von verschiedenen gespeicherten Style-Sheets angezeigt, die Sie auswählen können, um das Erscheinungsbild Ihrer Diagramme zu ändern. Weitere Informationen finden Sie im Thema „Zuweisen von Style-Sheets” auf Seite 278.
- **Stile im Knoten speichern.** Dadurch werden Änderungen an den Stilen des ausgewählten Diagramms gespeichert, um sie auf alle zukünftigen Diagramme anzuwenden, die über denselben Diagrammknoten im aktuellen Stream erstellt werden.
- **Stile als Standard speichern.** Dadurch werden Änderungen an den Stilen des ausgewählten Diagramms gespeichert, um sie auf alle zukünftigen Diagramme anzuwenden, die über beliebigen Dia-

grammknoten in einem beliebigen Stream erstellt werden. Nach Auswahl dieser Option können Sie mit **Standardstile übernehmen** den Stil auch auf alle bestehenden Diagramme anwenden.

- **Standardstile übernehmen.** Ändert die Stile des ausgewählten Diagramms auf die derzeit gespeicherten Standardstile.
- **Ursprüngliche Stile übernehmen.** Ändert die Stile eines Diagramms zurück auf die ursprünglichen, mitgelieferten Standardstile.

Zuweisen von Style-Sheets

Sie können ein Visualisierungs-Style-Sheet anwenden, das stilistische Eigenschaften der Visualisierung festlegt. Das Style-Sheet kann beispielsweise Schriftarten, Striche, Farben und weitere Optionen definieren. Bis zu einem gewissen Grad bieten Style-Sheets einen Direktzugriff zu Änderungen, die Sie sonst manuell vornehmen müssten. Beachten Sie jedoch, dass sich mit einem Style-Sheet nur Änderungen am *Stil* vornehmen lassen. Sonstige Änderungen wie die Position der Legende oder der Skalenbereich werden nicht in Style-Sheets gespeichert.

So weisen Sie ein Style-Sheet zu

1. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Bearbeiten > **Stile** > **Style-Sheet wechseln**
2. Wählen Sie ein Style-Sheet im Dialogfeld "Style-Sheet wechseln" aus.
3. Klicken Sie auf **Anwenden**, um das Style-Sheet auf die Visualisierung anzuwenden, ohne das Dialogdfeld zu schließen. Klicken Sie auf **OK**, um das Style-Sheet anzuwenden und das Dialogfeld zu schließen.

Dialogfeld "Style-Sheet wechseln/wählen"

Die Tabelle im oberen Bereich des Dialogfelds enthält alle derzeit verfügbaren Visualisierungs-Style-Sheets. Einige Style-Sheets sind vorinstalliert, andere hingegen wurden möglicherweise in IBM SPSS Visualization Designer (separates Produkt) erstellt.

Im unteren Bereich des Dialogfelds befinden sich Beispielvisualisierungen mit Beispieldaten. Wählen Sie ein Style-Sheet aus, um dessen Stile auf die Beispielvisualisierungen anzuwenden. Mithilfe dieser Beispiele können Sie feststellen, welche Auswirkungen das Style-Sheet auf die Visualisierung haben wird.

Das Dialogfeld bietet zudem folgende Optionen.

Vorhandene Stile. Standardmäßig kann ein Style-Sheet alle Stile der Visualisierung überschreiben. Sie können dieses Standardverhalten ändern.

- **Alle Stile überschreiben**. Wenn Sie das Style-Sheet zuweisen, werden alle Stile in der Visualisierung einschließlich der während der aktuellen Bearbeitung der Visualisierung vorgenommenen Änderungen überschrieben.
- **Geänderte Stile beibehalten**. Bei der Anwendung des Style-Sheets werden nur die Stile überschrieben, die *nicht* während des aktuellen Änderungsvorgangs in der Visualisierung bearbeitet wurden. Stile, die während der aktuellen Bearbeitung geändert wurden, werden beibehalten.

Verwalten. Verwalten Sie Visualisierungsvorlagen, Style-Sheets und Karten auf Ihrem Computer. Sie können Visualisierungsvorlagen, Style-Sheets und Karten auf Ihrem lokalen System importieren, exportieren, umbenennen und löschen. Weitere Informationen finden Sie im Thema „Verwalten von Vorlagen, Style-Sheets und Kartendateien” auf Seite 206.

Speicherort. Ändern Sie den Speicherort von Visualisierungsvorlagen und, Style-Sheets und Karten. Der aktuelle Speicherort wird rechts neben der Schaltfläche angezeigt. Weitere Informationen finden Sie im Thema „Festlegen des Speicherorts für Vorlagen, Style-Sheets und Karten” auf Seite 206.

Drucken, Speichern, Kopieren und Exportieren von Diagrammen

Jedes Diagramm weist eine Reihe von Optionen auf, mit denen Sie das Diagramm speichern oder drucken oder in ein anderes Format exportieren können. Die meisten dieser Optionen stehen über das Menü "Datei" zur Verfügung. Außerdem können Sie im Menü "Bearbeiten" auswählen, dass das Diagramm oder die enthaltenen Daten zur Verwendung in einer anderen Anwendung kopiert werden soll.

Drucken

Zum Drucken des Diagramms können Sie das Menüelement bzw. die Schaltfläche **Drucken** verwenden. Vor dem Drucken können Sie mithilfe von **Seite einrichten** und **Druckvorschau** die Druckoptionen festlegen und eine Vorschau der Ausgabe anzeigen.

Speichern von Diagrammen

Um ein Diagramm in einer IBM SPSS Modeler-Ausgabedatei (**.cou*) zu speichern, müssen Sie in den Menüs die Optionsfolge **Datei > Speichern** bzw. **Datei > Speichern unter** auswählen.

ODER

Um das Diagramm im Repository zu speichern, wählen Sie in den Menüs die Optionsfolge **Datei > Ausgabe speichern**.

Kopieren von Diagrammen

Um das Diagramm zur Verwendung in einer anderen Anwendung, beispielsweise in MS Word oder MS PowerPoint, zu kopieren, wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Diagramm kopieren**.

Kopieren von Daten

Um die Daten zur Verwendung in einer anderen Anwendung, beispielsweise in MS Excel oder MS Word, zu kopieren, wählen Sie in den Menüs die Optionsfolge **Bearbeiten > Daten kopieren**. Standardmäßig werden die Daten als HTML formatiert. Verwenden Sie in der anderen Anwendung die Option **Inhalte einfügen**, damit beim Einfügen andere Formatierungsoptionen angezeigt werden.

Exportieren von Diagrammen

Mit der Option Diagramm exportieren können Sie das Diagramm in einem der folgenden Formate exportieren: Bitmap (*.bmp*), JPEG (*.jpg*), PNG (*.png*), HTML (*.html*) oder ViZml-Dokument (*.xml*) zur Verwendung in anderen IBM SPSS Statistics-Anwendungen.

Wählen Sie zum Exportieren von Diagrammen in den Menüs die Optionsfolge **Datei > Diagramm exportieren** aus und wählen Sie dann das Format aus.

Exportieren von Tabellen

Mit der Option **Tabelle exportieren** können Sie die Tabelle in einem der folgenden Formate exportieren: tabstoppgetrennt (*.tab*), kommagetrennt (*.csv*) oder HTML (*.html*).

Wählen Sie zum Exportieren von Tabellen in den Menüs die Optionsfolge **Datei > Tabelle exportieren** aus und wählen Sie dann das Format aus.

Kapitel 6. Ausgabeknoten

Überblick über Ausgabeknoten

Mit Ausgabeknoten erhalten Sie Informationen zu Ihren Daten und Modellen. Sie bieten außerdem einen Mechanismus zum Exportieren von Daten in verschiedenen Formaten, sodass Sie diese Daten auch mit anderen Software-Tools nutzen können.

Folgende Ausgabeknoten stehen zur Verfügung:

Der Tabellenknoten zeigt die Daten in Tabellenform an, die auch in eine Datei geschrieben werden kann. Diese Vorgehensweise empfiehlt sich immer dann, wenn die Datenwerte überprüft oder in leicht lesbarer Form exportiert werden sollen.

Der Matrixknoten erstellt eine Tabelle, die die Beziehungen zwischen den Feldern aufzeigt. Dieser Knoten dient am häufigsten zur Darstellung der Beziehung zwischen zwei symbolischen Feldern, kann jedoch auch zum Aufzeigen der Beziehungen zwischen Flagfeldern oder numerischen Feldern herangezogen werden.

Der Analyseknoten evaluiert die Fähigkeit von Vorhersagemodellen, genaue Vorhersagen zu generieren. Mit Analyseknoten werden verschiedene Vergleiche zwischen den vorhergesagten Werten und den tatsächlichen Werten für ein oder mehrere Modellnuggets angestellt. Sie können außerdem Vorhersagemodelle miteinander vergleichen.

Der Data Audit-Knoten bietet einen umfassenden ersten Einblick in die Daten mit statistischen Funktionen, Histogrammen und der Verteilung für die einzelnen Felder sowie Informationen zu Ausreißern, fehlenden Werten und Extremwerten. Die Ergebnisse werden in einer übersichtlichen Matrix dargestellt, die sortiert werden kann und als Grundlage für die Erzeugung normal großer Diagramme und Datenvorbereitungsknoten dient.

Mit dem Transformationsknoten können Sie die Ergebnisse von Transformationen auswählen und in einer Vorschau anzeigen, bevor Sie sie auf ausgewählte Felder anwenden.

Der Statistikknoten liefert grundlegende Übersichtsdaten zu numerischen Feldern. Er berechnet Übersichtsstatistiken für einzelne Felder und für die Korrelationen zwischen den Feldern.

Der Mittelwertknoten vergleicht die Mittelwerte zwischen unabhängigen Gruppen oder zwischen Paaren von in Bezug stehenden Feldern, um zu testen, ob ein signifikanter Unterschied vorliegt. So können Sie beispielsweise die Einnahmen vor und nach der Durchführung einer Werbeaktion vergleichen oder die Einnahmen, die von Kunden stammen, die keine Werbezettel erhielten, mit den Einnahmen von Kunden vergleichen, die von der Werbeaktion erreicht wurden.

Der Berichtknoten erstellt formatierte Berichte, die sowohl festen Text als auch Daten und andere aus den Daten abgeleitete Ausdrücke enthalten. Das Format des Berichts wird mithilfe von Textvorlagen festgelegt, mit denen der feste Text und die Datenausgabekonstruktionen definiert werden. Sie können eine benutzerdefinierte Textformatierung angeben; hierzu stehen HTML-Tags in der Vorlage sowie Optionen auf der Registerkarte "Ausgabe" zur Verfügung. Sie können Datenwerte und andere bedingte Ausgaben mithilfe von CLEM-Ausdrücken in der Vorlage aufnehmen.

Mit dem Globalwerteknoten werden die Daten gescannt und Übersichtswerte berechnet, die in CLEM-Ausdrücken herangezogen werden können. Mit diesem Knoten können Sie beispielsweise die Statistiken für das Feld *Alter* berechnen und dann den Gesamtmittelwert für *Alter* in CLEM-Ausdrücken verwenden. Fügen Sie hierzu die Funktion `@GLOBAL_MEAN(alter)` ein.

Der Simulationsanpassungsknoten prüft die statistische Verteilung der Daten in jedem Feld und generiert (oder aktualisiert) einen Simulationsgenerierungsknoten, wobei jedem Feld die am besten angepasste Verteilung zugeordnet wird. Mit dem Simulationsgenerierungsknoten können dann simulierte Daten generiert werden.

Der Simulationsevaluierungsknoten wertet ein angegebenes vorhergesagtes Zielfeld aus und stellt Verteilungs- und Korrelationsinformationen zum Zielfeld dar.

Verwalten der Ausgabe

Der Ausgabemanager zeigt die Diagramme, Grafiken und Tabellen an, die während einer IBM SPSS Modeler-Sitzung erstellt wurden. Sie können eine Ausgabe jederzeit erneut öffnen, indem Sie im Manager darauf doppelklicken. Der entsprechende Stream bzw. Knoten muss nicht erneut ausgeführt werden.

So zeigen Sie den Ausgabemanager an:

Öffnen Sie das Menü "Ansicht" und wählen Sie **Manager**. Klicken Sie auf die Registerkarte **Ausgaben**.

Im Ausgabemanager haben Sie folgende Möglichkeiten:

- Vorhandene Ausgabeobjekte anzeigen, beispielsweise Histogramme, Evaluierungsdiagramme oder Tabellen.
- Ausgabeobjekte umbenennen.
- Ausgabeobjekte auf Datenträger oder im IBM SPSS Collaboration and Deployment Services Repository (sofern verfügbar) speichern.
- Ausgabedateien zum aktuellen Projekt hinzufügen.
- Ungespeicherte Ausgabeobjekte aus der aktuellen Sitzung löschen.
- Gespeicherte Ausgabeobjekte öffnen oder aus dem IBM SPSS Collaboration and Deployment Services Repository (sofern verfügbar) abrufen.

Um auf diese Optionen zuzugreifen, klicken Sie mit der rechten Maustaste auf eine beliebige Stelle auf der Registerkarte "Ausgaben".

Anzeigen der Ausgabe

Die Bildschirmausgabe wird in einem Ausgabebrowserfenster angezeigt. Das Ausgabebrowserfenster weist einen eigenen Satz an Menüs auf, mit denen Sie die Ausgabe drucken, speichern oder in ein anderes Format exportieren können. Beachten Sie, dass die spezifischen Optionen je nach Ausgabetyp unterschiedlich sein können.

Drucken, Speichern und Exportieren von Daten. Folgende weitere Informationen sind verfügbar:

- Zum Drucken der Ausgabe können Sie die Menüoption bzw. Schaltfläche **Drucken** verwenden. Vor dem Drucken können Sie mithilfe von **Seite einrichten** und **Druckvorschau** die Druckoptionen festlegen und eine Vorschau der Ausgabe anzeigen.
- Zum Speichern der Ausgabe in einer IBM SPSS Modeler-Ausgabedatei (*.cou*) wählen Sie im Menü "Datei" die Option **Speichern** bzw. **Speichern unter**.
- Um die Ausgabe in einem anderen Format zu speichern, beispielsweise als Text oder HTML, wählen Sie im Menü "Datei" die Option **Exportieren**. Weitere Informationen finden Sie im Thema „Exportieren von Ausgaben” auf Seite 285.
- Um die Ausgabe in einem gemeinsam genutzten Repository zu speichern, damit andere Benutzer sie mit dem IBM SPSS Collaboration and Deployment Services Deployment Portal anzeigen können, wählen Sie **Im Web veröffentlichen** aus dem Menü "Datei". Beachten Sie, dass für IBM SPSS Collaboration and Deployment Services eine gesonderte Lizenz erforderlich ist.

Auswählen von Zellen und Spalten. Das Menü "Bearbeiten" enthält verschiedene Optionen, mit denen Zellen und Spalten, ausgewählt oder kopiert werden können bzw. ihre Auswahl aufgehoben werden, je nachdem was für den aktuellen Ausgabetyp geeignet ist. Weitere Informationen finden Sie im Thema „Auswählen von Zellen und Spalten” auf Seite 285.

Neue Knoten generieren. Im Menü "Generieren" können Sie neue Knoten auf der Grundlage des Inhalts des Ausgabebrowsers generieren. Die Optionen variieren je nach Ausgabetyp und aktuell in der Ausgabe ausgewählten Elementen. Einzelheiten zu den Knotengenerierungsoptionen für einen bestimmten Ausgabetyp finden Sie in der Dokumentation für die betreffende Ausgabe.

Veröffentlichen im Web

Mit der Funktion "Im Web veröffentlichen" können Sie bestimmte Typen von Streamausgaben in einem zentralen, gemeinsam genutzten IBM SPSS Collaboration and Deployment Services Repository veröffentlichen, das die Grundlage von IBM SPSS Collaboration and Deployment Services bildet. Wenn Sie diese Option verwenden, können andere Benutzer diese Ausgabe über einen Internet-Zugang und ein IBM SPSS Collaboration and Deployment Services-Konto ansehen; sie müssen nicht über eine Installation von IBM SPSS Modeler verfügen.

Hinweis: Für den Zugriff auf ein IBM SPSS Collaboration and Deployment Services-Repository ist eine separate Lizenz erforderlich. Weitere Informationen finden Sie im Dokument http://www.ibm.com/software/analytics/spss/products/deployment/cds/

Die folgende Tabelle listet die IBM SPSS Modeler-Knoten auf, die die Funktion "Im Web veröffentlichen" unterstützen. Ausgabe von diesen Knoten wird im IBM SPSS Collaboration and Deployment Services Repository in Ausgabeobjekt-Format (*.cou*) gespeichert und kann direkt in der IBM SPSS Collaboration and Deployment Services Deployment Portal angezeigt werden.

Andere Ausgabetypen können nur angezeigt werden, wenn die betreffende Anwendung (z. B. IBM SPSS Modeler für Streamobjekte) auf dem System des Benutzers installiert ist.

Tabelle 49. Knoten, die "Im Web veröffentlichen" unterstützen

Knotentyp	Knoten
Grafiken	all

Tabelle 49. Knoten, die "Im Web veröffentlichen" unterstützen (Forts.)

Knotentyp	Knoten
Ausgabe	Tabelle
	Matrix
	Data Audit
	Transformieren
	Mittelwerte
	Analyse
	Statistics
	Bericht (HTML)
IBM SPSS Statistics	Statistikausgabe

Veröffentlichen von Ausgabeobjekten im Web

So veröffentlichen Sie Ausgabeobjekte im Web:

1. Führen Sie in einem IBM SPSS Modeler-Stream einen der in der Tabelle aufgeführten Knoten aus. Damit wird ein Ausgabeobjekt (z. B. eine Tabelle, eine Matrix oder ein Berichtobjekt) in einem neuen Fenster erstellt.
2. Treffen Sie im Ausgabeobjektfenster folgende Auswahl:

   Datei > **Im Web veröffentlichen**

   Hinweis: Sollen einfache HTML-Dateien zur Verwendung mit einem Standard-Web-Browser exportiert werden, wählen Sie **Exportieren** aus dem Menü "Datei" und dann **HTML** aus.
3. Bauen Sie eine Verbindung zum IBM SPSS Collaboration and Deployment Services Repository auf.

   Wenn die Verbindung erfolgreich aufgebaut wurde, wird das Dialogfeld "Repository: Speichern" angezeigt.
4. Wenn Sie die gewünschten Speicheroptionen ausgewählten haben, klicken Sie auf **Speichern**.

Anzeigen von veröffentlichten Ausgabedaten im Web

Für die Verwendung dieser Funktion muss ein IBM SPSS Collaboration and Deployment Services-Konto eingerichtet sein. Wenn die entsprechende Anwendung für den anzuzeigenden Objekttyp installiert ist (z. B. IBM SPSS Modeler oder IBM SPSS Statistics), wird die Ausgabe in der Anwendung, nicht im Browser angezeigt.

Hinweis: Für den Zugriff auf IBM SPSS Collaboration and Deployment Services ist eine separate Lizenz erforderlich. Weitere Informationen finden Sie unter http://www.ibm.com/software/analytics/spss/products/deployment/cds/.

So zeigen Sie veröffentlicht Ausgabeobjekte im Web an:

1. Geben Sie in Ihrem Browser die Adresse *http://<repos_host>:<repos_port>/peb*

   ein.Dabei bezeichnen *repos_host* und *repos_port* den Hostnamen bzw. die Portnummer für den IBM SPSS Collaboration and Deployment Services-Host.
2. Geben Sie die Anmeldedaten für Ihr IBM SPSS Collaboration and Deployment Services-Konto ein.
3. Klicken Sie auf **Inhalt Repository**.
4. Navigieren Sie zu dem Objekt, das Sie anzeigen möchten, oder suchen Sie nach dem Objekt.
5. Klicken Sie auf den Objektnamen. Für einige Objekttypen wie z. B. Diagramme tritt eventuell eine Verzögerung ein, während das Objekt im Browser gerendert wird.

Anzeigen der Ausgabe in einem HTML-Browser

Auf der Registerkarte "Erweitert" der Modellnuggets "Lineare Regression", "Logistische Regression" und "Faktor" können Sie die angezeigten Informationen in einem separaten Browser öffnen, beispielsweise im Internet Explorer. Die Informationen werden als HTML ausgegeben, was es ermöglicht, sie zu speichern und an anderer Stelle wiederzuverwenden, beispielsweise in einem unternehmensweiten Intranet oder auf einer Internet-Site.

Um die Informationen in einem Browser anzuzeigen, klicken Sie auf die Startschaltfläche unterhalb des Modellsymbols links oben auf der Registerkarte "Erweitert" des Modellnuggets.

Exportieren von Ausgaben

Im Ausgabebrowserfenster können Sie auswählen, dass die Ausgabe in einem anderen Format, beispielsweise Text oder HTML, exportiert werden soll. Die Exportformate variieren je nach Ausgabetyp, im Allgemeinen sind ähneln sie jedoch den Optionen für den Dateityp, die verfügbar sind, wenn Sie in dem zur Generierung der Ausgabe verwendeten Knoten die Option **In Datei speichern** auswählen.

So exportieren Sie Ausgaben:

1. Öffnen Sie im Ausgabebrowser das Menü "Datei" und wählen Sie die Option **Exportieren**. Wählen Sie anschließend den zu erstellenden Dateityp:
   - **Tabstoppgetrennt (*.tab).** Mit dieser Option erzeugen Sie eine formatierte Textdatei mit den Datenwerten. Dieser Stil eignet sich häufig für das Erzeugen einer Textdarstellung der Daten, die dann in andere Anwendungen importiert werden kann. Diese Option ist für Tabellen-, Matrix- und Mittelwertknoten verfügbar.
   - **Kommagetrennt (*.dat).** Mit dieser Option erzeugen Sie eine Textdatei mit den Datenwerten; diese Werte sind durch Komma voneinander getrennt. Dieser Stil eignet sich häufig für die rasche Erzeugung einer Datendatei, die in Tabellenkalkulationen oder andere Anwendungen zur Datenanalyse importiert werden kann. Diese Option ist für Tabellen-, Matrix- und Mittelwertknoten verfügbar.
   - **Transponiert tabstoppgetrennt (*.tab).** Diese Option ist mit der Option "Tabstoppgetrennt" identisch, die Daten werden jedoch transponiert, sodass die Zeilen Felder und die Spalten Datensätze darstellen.
   - **Transponiert kommagetrennt (*.dat).** Diese Option ist mit der Option "Kommagetrennt" identisch, die Daten werden jedoch transponiert, sodass die Zeilen Felder und die Spalten Datensätze darstellen.
   - **HTML (*.html).** Mit dieser Option wird die Ausgabe im HTML-Format in eine oder mehrere Dateien geschrieben.

Auswählen von Zellen und Spalten

Eine Reihe von Knoten, darunter der Tabellenknoten, der Matrixknoten und der Mittelwertknoten generieren eine Tabellenausgabe. Diese Ausgabetabellen können auf ähnliche Weise angezeigt und bearbeitet werden. Zu den Bearbeitungsmöglichkeiten gehören die Aufnahme ausgewählter Zellen, das Kopieren der gesamten Tabelle oder von Teilen davon in die Zwischenablage, das Erstellen neuer Knoten auf der Grundlage der aktuellen Auswahl sowie das Speichern und Drucken der Tabelle.

Zellen auswählen. Um eine Zelle auszuwählen, klicken Sie darauf. Soll ein rechteckiger Zellbereich ausgewählt werden, klicken Sie auf eine Ecke des gewünschten Bereichs. Halten Sie die Maustaste gedrückt, ziehen Sie die Maus auf die diagonal gegenüberliegende Ecke des Bereichs und lösen Sie die Maustaste wieder. Um eine ganze Spalte auszuwählen, klicken Sie auf die Spaltenüberschrift. Mit Umschalt-Klicken bzw. Strg-Klicken auf Spaltenüberschriften können Sie mehrere Spalten gleichzeitig auswählen.

Sobald Sie eine neue Auswahl treffen, wird die bisherige Auswahl wieder aufgehoben. Wenn Sie die Steuertaste beim Auswählen gedrückt halten, wird jedoch nicht die bisherige Auswahl aufgehoben, sondern

die neue Auswahl wird zur vorhandenen Auswahl hinzugefügt. Auf diese Weise können Sie mehrere, nicht zusammenhängende Bereiche der Tabelle auswählen. Das Menü "Bearbeiten" enthält außerdem die Optionen **Alles auswählen** und **Auswahl aufheben**.

Spalten neu ordnen. Mit den Ausgabebrowsern des Tabellenknotens und des Mittelwertknotens können Sie Spalten in der Tabelle verschieben. Klicken Sie hierzu auf eine Spaltenüberschrift und ziehen Sie sie an die gewünschte Position. Es ist nicht möglich, mehrere Spalten gleichzeitig zu verschieben.

Tabellenknoten

Der Tabellenknoten erstellt eine Tabelle mit den in Ihren Daten enthaltenen Werten. Da alle Felder und alle Werte im Stream enthalten sind, können hiermit Datenwerte schnell und einfach überprüft oder in leicht lesbarer Form exportiert werden. Wahlweise können Sie auch Datensätze hervorheben, die eine bestimmte Bedingung erfüllen.

Registerkarte "Einstellungen" beim Tabellenknoten

Datensätze hervorheben, wenn. Sie können Datensätze in der Tabelle hervorheben, indem Sie einen CLEM-Ausdruck eingeben, der für die betreffenden Datensätze wahr ist. Diese Option ist nur dann aktiviert, wenn Sie die Option **Ausgabe auf Bildschirm** aktiviert haben.

Registerkarte "Format" beim Tabellenknoten

Die Registerkarte "Format" enthält Optionen, mit denen Sie die Formatierung für die einzelnen Felder festlegen. Diese Registerkarte wird auch beim Typknoten verwendet. Weitere Informationen finden Sie im Thema „Feldformat - Registerkarte "Einstellungen"" auf Seite 127.

Registerkarte "Ausgabe" beim Ausgabeknoten

Bei Knoten, die Ausgaben in Tabellenform generieren, können Sie mithilfe der Registerkarte "Ausgabe" Format und Standort der Ergebnisse angeben.

Ausgabename. Bestimmt den Namen der Ausgabe, die beim Ausführen des Knotens erstellt wird. Mit **Auto** wird ein Name auf der Grundlage des Knotens bestimmt, mit dem die Ausgabe generiert wird. Optional können Sie auch **Angepasst** auswählen und einen anderen Namen angeben.

Ausgabe auf Bildschirm (Standardeinstellung)**.** Erstellt ein Ausgabeobjekt für die Online-Anzeige. Das Ausgabeobjekt wird auf der Registerkarte "Ausgaben" im Manager-Fenster dargestellt, wenn der Ausgabeknoten ausgeführt wird.

Ausgabe in Datei. Speichert die Ausgabe in einer Datei, wenn der Knoten ausgeführt wird. Wenn Sie diese Option wählen, geben Sie einen Dateinamen an (oder wechseln Sie zu einem Verzeichnis und geben Sie einen Dateinamen mithilfe der Feldauswahlschaltfläche an) und wählen Sie einen Dateityp aus. Beachten Sie, dass einige Dateitypen möglicherweise nicht für bestimmte Ausgabetypen verfügbar sind.

Daten werden im Standardcodierungsformat des Systems ausgegeben, das in der Windows-Systemsteuerung bzw. bei Ausführung im verteilten Modus auf dem Server-Computer angegeben wird.

- **Daten (tabstoppgetrennt) (*.tab).** Mit dieser Option generieren Sie eine formatierte Textdatei mit den Datenwerten. Dieser Stil eignet sich häufig für das Generieren einer Textdarstellung der Daten, die dann in andere Anwendungen importiert werden kann. Diese Option ist für Tabellen-, Matrix- und Mittelwertknoten verfügbar.
- **Daten (kommagetrennt) (*.dat).** Mit dieser Option generieren Sie eine Textdatei mit den Datenwerten; diese Werte sind durch Komma voneinander getrennt. Dieser Stil eignet sich häufig für die rasche Generierung einer Datendatei, die in Tabellenkalkulationen oder andere Anwendungen zur Datenanalyse importiert werden kann. Diese Option ist für Tabellen-, Matrix- und Mittelwertknoten verfügbar.
- **HTML (*.html).** Mit dieser Option wird die Ausgabe im HTML-Format in eine oder mehrere Dateien geschrieben. Bei Tabellenausgaben (aus dem Tabellen-, Matrix- oder Mittelwertknoten) enthält eine Rei-

he von HTML-Dateien einen Inhaltsbereich, in dem die Feldnamen aufgeführt werden; die Daten befinden sich in einer HTML-Tabelle. Die Tabelle wird gegebenenfalls auf mehrere HTML-Dateien aufgeteilt, wenn die Anzahl der Zeilen in der Tabelle die Angaben unter **Zeilen pro Seite** überschreitet. In diesem Fall enthält der Inhaltsbereich Links zu allen Tabellenseiten und dient als Mittel zur Navigation in der Tabelle. Bei einer nicht tabellenförmigen Ausgabe wird eine einzige HTML-Datei mit den Ergebnissen des Knotens erstellt.

Hinweis: Falls die HTML-Ausgabe nur die Formatierung für die erste Seite enthält, wählen Sie die Option **Ausgabe paginieren** aus und passen Sie die Angaben unter **Zeilen pro Seite** an, sodass die gesamte Ausgabe auf einer einzigen Seite erfolgt. Falls die Ausgabevorlage für die Knoten (z. B. für den Berichtknoten) benutzerdefinierte HTML-Tags enthält, können Sie alternativ den Formattyp **Benutzerdefiniert** auswählen.

- **Textdatei (*.txt).** Mit dieser Option generieren Sie eine Textdatei mit der Ausgabe. Dieser Stil eignet sich häufig zum Generieren einer Ausgabe, die dann in andere Anwendungen importiert werden kann, z. B. in eine Textverarbeitung oder in Präsentations-Software. Diese Option ist für einige Knoten nicht verfügbar.
- **Ausgabeobjekt (*.cou).** Die in diesem Format gespeicherten Ausgabeobjekte können in IBM SPSS Modeler geöffnet und angezeigt, zu Projekten hinzugefügt sowie mit demIBM SPSS Collaboration and Deployment Services Repository veröffentlicht und verfolgt werden.

Ausgabeansicht. Für den Mittelwertknoten können Sie angeben, ob standardmäßig eine einfache oder eine erweiterte Ausgabe angezeigt werden soll. Beachten Sie, dass Sie auch zwischen diesen beiden Ansichten umschalten können, während Sie die generierte Ausgabe durchsuchen. Weitere Informationen finden Sie im Thema „Mittelwertknoten - Ausgabebrowser” auf Seite 309.

Format. Beim Berichtknoten können Sie auswählen, ob die Ausgabe automatisch formatiert oder mit HTML aus der Vorlage formatiert werden soll. Mit der Option **Angepasst** ermöglichen Sie die HTML-Formatierung in der Vorlage.

Titel. Beim Berichtknoten können Sie optional einen Titeltext angeben, der oben in der Berichtausgabe eingefügt werden soll.

Eingefügten Text hervorheben. Beim Berichtknoten lassen Sie mit dieser Option den Text hervorheben, der durch CLEM-Ausdrücke in der Berichtvorlage generiert wurde. Weitere Informationen finden Sie im Thema „Registerkarte "Vorlage" beim Berichtknoten” auf Seite 311. Diese Option wird nicht empfohlen, wenn Sie die Formatierung **Angepasst** verwenden.

Zeilen pro Seite. Geben Sie beim Berichtknoten die Anzahl der Zeilen an, die bei der Formatierung des Ausgabeberichts mit der Option **Auto** auf jeder Seite untergebracht werden sollen.

Daten transponieren. Mit dieser Option werden die Daten vor dem Export transponiert, sodass die Zeilen Felder und die Spalten Datensätze darstellen.

Hinweis: Bei umfangreichen Tabellen sind die obigen Optionen eher ineffizient; dies gilt insbesondere dann, wenn Sie mit einem fernen Server arbeiten. In solchen Fällen liefert ein Dateiausgabeknoten deutlich bessere Leistungen. Weitere Informationen finden Sie im Thema „Flatfile-Exportknoten” auf Seite 343.

Tabellenbrowser

Der Tabellenbrowser zeigt Daten in Tabellenform an und ermöglicht Ihnen die Durchführung von Standardoperationen, beispielsweise das Auswählen und Kopieren von Zellen, das Neuordnen von Spalten und das Speichern und Drucken der Tabelle. Weitere Informationen finden Sie im Thema „Auswählen von Zellen und Spalten” auf Seite 285. Dabei handelt es sich um die gleichen Operationen, die Sie bei der Vorschau der Daten in einem Knoten ausführen können.

Tabellendaten exportieren. Sie können Daten aus dem Tabellenbrowser über folgende Optionsfolge exportieren:

Datei > **Exportieren**

Weitere Informationen finden Sie im Thema „Exportieren von Ausgaben” auf Seite 285.

Daten werden im Standardcodierungsformat des Systems exportiert, das in der Windows-Systemsteuerung bzw. bei Ausführung im verteilten Modus auf dem Server-Computer angegeben wird.

Tabelle durchsuchen. Mit der Schaltfläche "Suche" (das Fernglassymbol) in der Hauptsymbolleiste aktivieren Sie die Symbolleiste für Suchvorgänge, um so bestimmte Werte in der Tabelle zu suchen. Sie können vorwärts oder rückwärts in der Tabelle suchen, die Suche unter Beachtung der Groß-/Kleinschreibung starten (Schaltfläche **Aa**) sowie einen laufenden Suchvorgang mit der Schaltfläche zum Unterbrechen der Suche anhalten.

Neue Knoten generieren. Das Menü "Generieren" enthält Funktionen zum Erzeugen von Knoten.

- **Auswahlknoten ("Datensätze").** Erzeugt einen Auswahlknoten, mit dem die Datensätze ausgewählt werden, für die eine beliebige Zelle in der Tabelle markiert ist.
- **Auswahlknoten ("UND").** Erzeugt einen Auswahlknoten, mit dem die Datensätze ausgewählt werden, die *alle* in der Tabelle markierten Werte enthält.
- **Auswahlknoten ("ODER").** Erzeugt einen Auswahlknoten, mit dem die Datensätze ausgewählt werden, die *einen* der in der Tabelle markierten Werte enthält.
- **Ableitungsknoten ("Datensätze").** Erzeugt einen Ableitungsknoten, mit dem ein neues Flagfeld erstellt wird. Das Flagfeld besitzt den Wert *T* für Datensätze, für die eine beliebige Zelle in der Tabelle ausgewählt ist, bzw. den Wert *F* für die verbleibenden Datensätze.
- **Ableitungsknoten ("UND").** Erzeugt einen Ableitungsknoten, mit dem ein neues Flagfeld erstellt wird. Das Flagfeld besitzt den Wert *T* für Datensätze, die *alle* in der Tabelle ausgewählten Werte enthalten, bzw. *F* für die verbleibenden Datensätze.
- **Ableitungsknoten ("ODER").** Erzeugt einen Ableitungsknoten, mit dem ein neues Flagfeld erstellt wird. Das Flagfeld besitzt den Wert *T* für Datensätze, die *einen* in der Tabelle ausgewählten Wert enthalten, bzw. *F* für die verbleibenden Datensätze.

Matrixknoten

Mit dem Matrixknoten können Sie eine Tabelle erstellen, die die Beziehungen zwischen den Feldern aufzeigt. Dieser Knoten dient am häufigsten zum Darstellen der Beziehung zwischen zwei kategorialen Feldern (Flag, nominal oder ordinal), kann jedoch auch zum Aufzeigen der Beziehungen zwischen stetigen Feldern (numerischer Bereich) herangezogen werden.

Registerkarte "Einstellungen" beim Matrixknoten

Auf der Registerkarte "Einstellungen" legen Sie Optionen für die Struktur der Matrix fest.

Felder. Wählen Sie einen Feldauswahltyp aus den folgenden Optionen aus:

- **Ausgewählt.** Bei dieser Option können Sie je ein kategoriales Feld für die Zeilen und Spalten in der Matrix auswählen. Die Zeilen und Spalten der Matrix werden durch die Liste der Werte für das ausgewählte kategoriale Feld bestimmt. Die Zellen der Matrix enthalten die unten ausgewählten Übersichtsstatistiken.
- **Alle Flags (wahre Werte).** Bei dieser Option wird eine Matrix mit je einer Zeile und einer Spalte für jedes Flagfeld in den Daten angefordert. Die Zellen der Matrix enthalten die Anzahl der doppeltpositiven Werte für jede Flagkombination. Beispiel: Bei einer Zeile für *Brot gekauft* und einer Spalte für *Käse gekauft* enthält die Zelle am Schnittpunkt dieser Zeile und Spalte die Anzahl der Datensätze, bei denen sowohl *Brot gekauft* als auch *Käse gekauft* wahr sind.
- **Nur numerisch.** Bei dieser Option wird eine Matrix mit je einer Zeile und einer Spalte für jedes numerische Feld angefordert. Die Zellen der Matrix stellen die Summe der Kreuzprodukte für das entspre-

chende Feldpaar dar. Für jede Zelle in der Matrix werden also die Werte aus dem Zeilenfeld und dem Spaltenfeld für jeden Datensatz multipliziert und dann über die Datensätze hinweg summiert.

Fehlende Werte einschließen. Schließt benutzerdefiniert fehlende Werte (leer) und systemdefiniert fehlende Werte ($null$) in die Zeilen- und Spaltenausgabe ein. Wenn beispielsweise der Wert *N/A* für das ausgewählte Spaltenfeld als benutzerdefiniert fehlend definiert wurde, wird eine gesonderte Spalte mit der Beschriftung *N/A* wie jede andere Kategorie in die Tabelle aufgenommen (vorausgesetzt, dieser Wert kommt tatsächlich in den Daten vor). Wenn die Auswahl dieser Option aufgehoben wird, wird die Spalte *k. A* ausgeschlossen, egal wie oft sie vorkommt.

Hinweis: Die Option zur Aufnahme fehlender Werte gilt nur, wenn die ausgewählten Felder als Kreuztabelle vorliegen. Leere Werte werden $null$ zugeordnet und aus der Aggregation für das Funktionsfeld ausgeschlossen, wenn der Modus **Ausgewählt** ist und der Inhalt auf **Funktion** gesetzt ist. Der Ausschluss für alle numerischen Felder erfolgt, wenn der Modus auf **Alle numerischen Werte** gesetzt ist.

Zelleninhalte. Wenn Sie oben die Option **Ausgewählte Felder** aktiviert haben, können Sie die Statistik angeben, die in den Zellen der Matrix verwendet werden soll. Wählen Sie eine anzahlbasierte Statistik aus oder wählen Sie ein Überlagerungsfeld aus, mit dem die Werte aus einem numerischen Feld auf der Grundlage der Werte der Zeilen- und Spaltenfelder zusammengefasst werden.

- **Kreuztabellen.** Die Zellenwerte bestehen aus der Anzahl und/oder dem Prozentsatz der Datensätze, die die entsprechende Wertekombination aufweisen. Mit den Optionen auf der Registerkarte "Darstellung" können Sie die gewünschten Kreuztabellenübersichten auswählen. Der globale Chi-Quadrat-Wert wird ebenso zusammen mit der Signifikanz angezeigt. Weitere Informationen finden Sie im Thema „Matrixknoten - Ausgabebrowser” auf Seite 290.
- **Funktion.** Wenn Sie eine Übersichtsfunktion auswählen, bilden die Zellenwerte eine Funktion der Werte aus dem ausgewählten Überlagerungsfeld für Fälle, die die entsprechenden Zeilen- und Spaltenwerte besitzen. Beispiel: Sie verwenden das Zeilenfeld *Region*, das Spaltenfeld *Produkt* und das Überlagerungsfeld *Einkommen*. In diesem Fall enthält die Zelle in der Zeile *Nordosten* und der Spalte *Dings* die Summe (bzw. den Durchschnitt, den Mindestwert oder den Höchstwert) für den Umsatz aus Geräten, die in der Region Nordost verkauft wurden. Die Standardeinstellung für die Übersichtsfunktion lautet **Mittelwert**. Sie können eine andere Funktion auswählen, mit der das Funktionsfeld zusammengefasst werden soll. Die folgenden Optionen stehen zur Auswahl: **Mittelwert**, **Summe**, **Std.abw.** (Standardabweichung), **Max** (Maximum) sowie **Min** (Minimum).

Registerkarte "Darstellung" beim Matrixknoten

Auf der Registerkarte "Darstellung" steuern Sie die Optionen zum Sortieren und Hervorheben für die Matrix sowie die Statistiken, die für Kreuztabellenmatrizen angezeigt werden.

Zeilen und Spalten. Steuert die Sortierung der Zeilen- und Spaltenüberschriften in der Matrix. Die Standardeinstellung lautet **Nicht sortiert**. Mit der Option **Aufsteigend** oder **Absteigend** lassen Sie die Zeilen- und Spaltenüberschrift in die angegebene Richtung sortieren.

Überlagerung. Ermöglicht das Hervorheben von Extremwerten in der Matrix. Die Werte werden auf der Grundlage der Zellenanzahl (bei Kreuztabellenmatrizen) oder der berechneten Werte (bei Funktionsmatrizen) hervorgehoben.

- **Oberen Rand hervorheben.** Sie können die höchsten Werte in der Matrix hervorheben lassen (in Rot). Geben Sie die Anzahl der hervorzuhebenden Werte an.
- **Unteren Rand hervorheben.** Sie können auch die niedrigsten Werte in der Matrix hervorheben lassen (in Grün). Geben Sie die Anzahl der hervorzuhebenden Werte an.

Hinweis: Bei den beiden Hervorhebungsoptionen können Bindungen dazu führen, dass mehr Werte als angefordert hervorgehoben werden. Wenn Sie beispielsweise eine Matrix mit sechs Nullwerten in den Zellen verwenden und die Option **Unteren Rand hervorheben** mit dem Wert 5 auswählen, werden alle sechs Nullwerte hervorgehoben.

Kreuztabellen-Zelleninhalte. Bei Kreuztabellen können Sie die Übersichtsstatistiken in der Matrix für Kreuztabellenmatrizen angeben. Diese Optionen sind nicht verfügbar, wenn die Option **Alle numerischen Werte** oder **Funktion** auf der Registerkarte "Einstellungen" ausgewählt wurde.

- **Häufigkeiten.** Die Zellen umfassen die Anzahl der Datensätze mit dem Zeilenwert, die auch den zugehörigen Spaltenwert aufweisen. Dies gilt nur für den Standardzelleninhalt.
- **Erwartete Werte.** Der erwartete Wert für die Anzahl der Datensätze in der Zelle, unter der Annahme, dass keine Beziehung zwischen Zeilen und Spalten besteht. Die erwarteten Werte beruhen auf der folgenden Formel:

  p(*Zeilenwert*) * p(*Spaltenwert*) * *Gesamtzahl der Datensätze*
- **Residuen.** Die Differenz zwischen beobachteten und erwarteten Werten.
- **Prozentsatz für Zeile.** Der Prozentsatz aller Datensätze mit dem Zeilenwert, die auch den zugehörigen Spaltenwert aufweisen. Die Prozentsätze für die Zeilen ergeben insgesamt den Wert 100.
- **Prozentsatz für Spalte.** Der Prozentsatz aller Datensätze mit dem Spaltenwert, die auch den zugehörigen Zeilenwert aufweisen. Die Prozentsätze für die Spalten ergeben insgesamt den Wert 100.
- **Prozentsatz für Gesamtsumme.** Der Prozentsatz aller Datensätze, die die angegebene Kombination aus Spaltenwert und Zeilenwert aufweisen. Die Prozentsätze in der gesamten Matrix ergeben insgesamt den Wert 100.
- **Zeilen- und Spaltengesamtsummen einschließen.** Fügt eine Reihe und eine Spalte zur Matrix hinzu, in denen die Gesamtsummen der Zeilen und Spalten eingetragen werden.
- **Einstellungen anwenden.** (nur Ausgabebrowser) Ermöglicht die Vornahme von Änderungen am Erscheinungsbild der Ausgabe des Matrixknotens, ohne dass der Ausgabebrowser geschlossen und erneut geöffnet werden muss. Nehmen Sie die Änderungen auf dieser Registerkarte des Ausgabebrowsers vor, klicken Sie auf diese Schaltfläche und wählen Sie dann die Registerkarte "Matrix", um die Auswirkungen der Änderungen anzuzeigen.

Matrixknoten - Ausgabebrowser

Im Matrixbrowser werden Kreuztabellendaten angezeigt. Hier können Sie verschiedene Vorgänge für die Matrix ausführen, z. B. Zellen auswählen, Matrix ganz oder teilweise in die Zwischenablage kopieren, neue Knoten auf der Grundlage der Auswahl in der Matrix generieren sowie die Matrix speichern und drucken. Mit dem Matrixbrowser können Sie außerdem die Ausgabe bestimmter Modelle anzeigen lassen, z. B. Naive Bayes-Modelle aus Oracle.

Die Menüs "Datei" und "Bearbeiten" bieten die üblichen Optionen zum Drucken, Speichern und Exportieren von Ausgaben sowie zum Auswählen und Kopieren von Daten. Weitere Informationen finden Sie im Thema „Anzeigen der Ausgabe” auf Seite 283.

Chi-Quadrat. Für eine Kreuztabelle zweier kategorialer Felder wird außerdem das globale Pearson-Chi-Quadrat unterhalb der Tabelle angezeigt. Dieser Test gibt die Wahrscheinlichkeit an, dass die beiden Felder unabhängig sind. Die Grundlage hierfür ist die Differenz zwischen der beobachteten Anzahl und den Anzahlwerten, die zu erwarten sind, wenn keine Beziehung vorhanden ist. Beispiel: Wenn es keinen Zusammenhang zwischen Kundenzufriedenheit und Geschäftsstandort gibt, sind ähnliche Zufriedenheitsquoten in allen Geschäften zu erwarten. Wenn jedoch die Kunden in bestimmten Geschäften durchgängig höhere Zufriedenheitsquoten aufweisen als andere, ist zu vermuten, dass es sich nicht um einen Zufall handelt. Je größer die Differenz, desto kleiner ist die Wahrscheinlichkeit, dass es sich lediglich um die Folge eines Fehlers bei der Zufallsstichprobennahme handelt.

- Der Chi-Quadrat-Test gibt die Wahrscheinlichkeit an, dass die beiden Felder unabhängig sind. In diesem Fall sind die Differenzen zwischen beobachteten und erwarteten Häufigkeiten ausschließlich auf den Zufall zurückzuführen. Wenn diese Wahrscheinlichkeit sehr gering ist - üblicherweise unter 5 % - wird die Beziehung zwischen den beiden Feldern als signifikant bezeichnet.
- Wenn es nur eine einzige Spalte oder Zeile gibt (einfacher Chi-Quadrat-Test) ist der Wert für die Freiheitsgrade die Anzahl der Zellen minus 1. Bei einem zweifachen Chi-Quadrat ist der Wert für die Freiheitsgrade gleich der Anzahl der Zeilen minus 1 mal der Anzahl der Spalten minus 1.

- Lassen Sie bei der Interpretation der Chi-Quadrat-Statistik Vorsicht walten, wenn eine der erwarteten Zellenhäufigkeiten unter 5 liegt.
- Der Chi-Quadrat-Test ist nur für eine Kreuztabelle aus zwei Feldern verfügbar. (Wenn **Alle Flags** oder **Alle numerischen Werte** auf der Registerkarte "Einstellungen" ausgewählt wurde, wird dieser Test nicht angezeigt.)

Menü "Generieren". Das Menü "Generieren" enthält Funktionen zum Erzeugen von Knoten. Diese Funktionen sind nur bei Kreuztabellenmatrizen verfügbar; in der Matrix muss dabei mindestens eine Zelle ausgewählt sein.

- **Auswahlknoten.** Erzeugt einen Auswahlknoten, mit dem die Datensätze ausgewählt werden, die mit einer beliebigen ausgewählten Zelle in der Matrix übereinstimmen.
- **Ableitungsknoten (Flag).** Erzeugt einen Ableitungsknoten, mit dem ein neues Flagfeld erstellt wird. Das Flagfeld besitzt den Wert *T* für Datensätze, die mit einer beliebigen Zelle in der Matrix übereinstimmen, bzw. den Wert *F* für die verbleibenden Datensätze.
- **Ableitungsknoten (Set).** Erzeugt einen Ableitungsknoten, mit dem ein neues nominales Feld erstellt wird. Das nominale Feld enthält je eine Kategorie für jedes zusammenhängende Set ausgewählter Zellen in der Matrix.

Analyseknoten

Mit dem Analyseknoten können Sie die Fähigkeit eines Modells zur Erzeugung genauer Vorhersagen evaluieren. Mit Analyseknoten werden verschiedene Vergleiche zwischen den vorhergesagten Werten und den tatsächlichen Werten (Ihr Zielfeld) für ein oder mehrere Modellnuggets angestellt. Analyseknoten können außerdem zum Vergleich von Vorhersagemodellen mit anderen Vorhersagemodellen dienen.

Wenn Sie einen Analyseknoten ausführen, wird auf der Registerkarte "Übersicht" unter "Analyse" automatisch eine Zusammenfassung der Analyseergebnisse für jedes Modellnugget im ausgeführten Stream eingetragen. Die ausführlichen Analyseergebnisse werden auf der Registerkarte "Ausgabe" des Manager-Fensters angezeigt oder können direkt in eine Datei geschrieben werden.

Hinweis: Da Analyseknoten vorhergesagte Felder mit tatsächlichen Werten vergleichen, sind sie nur in überwachten Modellen (Modelle, die ein Zielfeld erfordern) sinnvoll. Bei nicht überwachten Modellen wie beispielsweise Clusteralgorithmen, stehen keine tatsächlichen Ergebnisse als Grundlage für einen Vergleich zur Verfügung.

Registerkarte "Analyse" beim Analyseknoten

Auf der Registerkarte "Analyse" können Sie die Details für die Analyse angeben.

Fehlklassifizierungstabellen (für symbolische bzw. kategoriale Ziele). Zeigt für kategoriale Ziele (Flag, nominal oder ordinal) das Muster der Übereinstimmungen zwischen jedem erzeugten (vorhergesagten) Feld und dem zugehörigen Zielfeld an. Es wird eine Tabelle eingeblendet, bei der die Zeilen durch tatsächliche Werte definiert sind und die Spalten durch vorhergesagte Werte; hierbei wird die Anzahl der Datensätze ersichtlich, die dieses Muster in den einzelnen Zellen aufweisen. Mit dieser Option können systematische Fehler bei der Vorhersage erkannt werden. Falls mehrere erzeugte Felder zu einem bestimmten Ausgabefeld gehören, jedoch durch unterschiedliche Modelle erstellt wurden, werden die Fälle gezählt, in denen diese Felder übereinstimmen bzw. nicht übereinstimmen, und die Gesamtsummen werden angezeigt. In den Fällen mit Übereinstimmung wird eine weitere Gruppe mit Richtig/Falsch-Statistiken eingeblendet.

Leistungsauswertung. Zeigt die Leistungsauswertungsstatistik für Modelle mit kategorialen Ausgaben. Diese Statistik wird für jede Kategorie des oder der Ausgabefelder erstellt und ist ein Maß für den durchschnittlichen Informationsgehalt (in Bit) des Modells bei der Vorhersage für Datensätze, die zu der betreffenden Kategorie gehören. Hierbei wird die Schwierigkeit des Klassifizierungsproblems berücksichtigt; genaue Vorhersagen für seltene Kategorien erhalten somit einen höheren Leistungsauswertungsindex als

genaue Vorhersagen für häufig auftretende Kategorien. Liefert das Modell quasi nur "geratene" Werte für eine Kategorie, erhält diese Kategorie den Leistungsauswertungsindex 0.

Auswertungsmetrik (AUC & Gini, nur binäre Klassifikationsmerkmale). Diese Option dokumentiert für binäre Klassifikationsmerkmale die AUC-Auswertungsmetrik (AUC - Area Under Curve, Fläche unter der Kurve) und die Auswertungsmetrik für den Gini-Koeffizienten. Diese beiden Auswertungsmetriken werden für jedes binäre Modell zusammen berechnet. Die Werte der Metriken werden in einer Tabelle im Analyseausgabebrowser dokumentiert.

Die AUC-Auswertungsmetrik wird als Fläche unter einer ROC-Kurve (ROC - Receiver Operator Characteristic) berechnet und ist eine skalare Darstellung der erwarteten Leistung eines Klassifikationsmerkmals. Die Fläche unter der Kurve liegt immer zwischen 0 und 1, wobei eine höhere Zahl für ein besseres Klassifikationsmerkmal steht. Eine diagonale ROC-Kurve zwischen den Koordinaten (0,0) und (1,1) stellt ein zufälliges Klassifikationsmerkmal dar und die Fläche unter der Kurve beträgt 0,5. Daher hat ein realistisches Klassifikationsmerkmal keine Fläche unter der Kurve unter 0,5.

Die Auswertungsmetrik für den Gini-Koeffizienten wird manchmal als Alternative zur AUC-Auswertungsmetrik verwendet und die beiden Maße sind eng verwandt. Der Gini-Koeffizient wird als doppelte Fläche zwischen der ROC-Kurve und der Diagonalen oder als Gini = 2AUC - 1 berechnet. Der Gini-Koeffizient liegt immer zwischen 0 und 1, wobei eine höhere Zahl für ein besseres Klassifikationsmerkmal steht. Der Gini-Koeffizient ist in dem unwahrscheinlichen Fall, dass die ROC-Kurve unter der Diagonalen liegt, negativ.

Konfidenzzahlen (falls verfügbar). Bei Modellen, bei denen ein Konfidenzfeld erzeugt wird, lassen Sie mit dieser Option die Statistik zu den Konfidenzwerten und deren Beziehung zu den Vorhersagen zusammenstellen. Für diese Option stehen zwei Einstellungen zur Auswahl:

- **Schwellenwert für.** Meldet das Konfidenzniveau, ab dem die Genauigkeit den angegebenen Prozentsatz erreicht.
- **Genauigkeit verbessern.** Meldet das Konfidenzniveau, ab dem die Genauigkeit um den angegebenen Faktor verbessert wird. Beispiel: Die Gesamtgenauigkeit liegt bei 90 % und für diese Option wurde der Wert 2,0 angegeben. Der gemeldete Wert entspricht somit der Konfidenz für eine Genauigkeit von 95 %.

Vorhergesagte Felder/Prädiktorfelder finden mithilfe von. Bestimmt, wie vorhergesagte Felder dem ursprünglichen Zielfeld zugeordnet werden sollen.

- **Metadaten des Modellausgabefelds.** Ordnet vorhergesagte Felder auf der Grundlage der Informationen zum Modell-Feld dem Ziel zu und ergibt auch dann eine Übereinstimmung, wenn ein vorhergesagtes Ziel umbenannt wurde. Die Informationen zum Modell-Feld können für jedes vorhergesagte Feld mithilfe eines Typknotens über das Dialogfeld "Werte" aufgerufen werden. Weitere Informationen finden Sie im Thema „Verwenden des Dialogfelds "Werte"" auf Seite 123.
- **Format für Feldnamen.** Ordnet Felder anhand der Namensgebungskonventionen zu. So müssen sich beispielsweise vorhergesagte Werte, die von einem C5.0-Modellnugget für das Ziel *Antwort* erstellt wurden, in einem Feld mit der Bezeichnung *$C-Antwort* befinden.

Nach Partition trennen. Wenn Datensätze mithilfe eines Partitionsfelds in Trainings-, Test- und Validierungsstichproben aufgeteilt werden, lassen Sie mit dieser Option die Ergebnisse für die einzelnen Partitionen separat anzeigen. Weitere Informationen finden Sie im Thema „Partitionsknoten" auf Seite 154.

Hinweis: Beim Aufteilen nach Partition werden Datensätze mit Nullwerten im Partitionsfeld von der Analyse ausgeschlossen. Dieses Problem tritt nicht auf, wenn Sie einen Partitionsknoten verwenden, weil diese Knoten keine Nullwerte erzeugen.

Benutzerdefinierte Analyse. Sie können eine eigene Analyseberechnung angeben, mit der die Modelle ausgewertet werden sollen. Legen Sie die zu berechnenden Werte für die einzelnen Datensätze mithilfe von CLEM-Ausdrücken fest und geben Sie an, wie die Scores auf der Ebene der Datensätze zu einem Ge-

samtwert zusammengefasst werden sollen. Mit den Funktionen **@TARGET** und **@PREDICTED** verweisen Sie auf den Zielwert (tatsächliche Ausgabe) bzw. auf den vorhergesagten Wert.

- **Wenn.** Legen Sie einen bedingten Ausdruck fest, wenn die auszuführende Berechnung von einer bestimmten Bedingungen abhängt.
- **Dann.** Geben Sie die Berechnung an, die ausgeführt werden soll, wenn die Wenn-Bedingung wahr ist.
- **Sonst.** Geben Sie die Berechnung an, die ausgeführt werden soll, wenn die Wenn-Bedingung falsch ist.
- **Verwenden.** Wählen Sie eine Statistik aus, mit der ein Gesamtscore aus den Einzelscores berechnet werden soll.

Analyse nach Feldern aufschlüsseln. Zeigt die kategorialen Felder, die für die Aufschlüsselung der Analyse zur Verfügung stehen. Neben der Gesamtanalyse wird je eine separate Analyse für die einzelnen Kategorien in jedem Aufschlüsselungsfeld erstellt.

Analyseausgabebrowser

Im Analyseausgabebrowser werden die Ergebnisse aus der Ausführung des Analyseknotens angezeigt. Das Menü "Datei" enthält die üblichen Befehle zum Speichern, Exportieren und Drucken. Weitere Informationen finden Sie im Thema „Anzeigen der Ausgabe” auf Seite 283.

Beim Öffnen des Analyseausgabebrowsers werden die Ergebnisse erweitert. Um die Ergebnisse nach der Betrachtung wieder auszublenden, können Sie mit dem Erweiterungssteuerelement links neben dem gewünschten Element die Ergebnisse reduzieren. Alternativ können Sie mit der Schaltfläche **Alles ausblenden** alle Ergebnisse ausblenden. Um die für Sie relevanten Ergebnisse nach dem Reduzieren wieder anzeigen zu lassen, erweitern Sie die gewünschten Ergebnisse mithilfe des Erweiterungssteuerelements auf der linken Seite oder klicken Sie auf die Schaltfläche **Alles anzeigen**, um alle Ergebnisse anzuzeigen.

Ergebnisse für Zielfeld. Die Analyseausgabe enthält je einen Abschnitt für die einzelnen Ausgabefelder, für die ein zugehöriges Vorhersagefeld vorliegt, das durch ein erzeugtes Modell erstellt wurde.

Vergleich. Der Abschnitt mit den Ausgabefeldern enthält je einen Unterabschnitt für jedes Vorhersagefeld, das mit dem betreffenden Ausgabefeld verknüpft ist. Bei kategorialen Ausgabefeldern wird im oberen Bereich dieses Abschnitts eine Tabelle angezeigt, aus der die Anzahl und der Prozentsatz der richtigen und falschen Vorhersagen sowie die Gesamtanzahl der Datensätze im Stream hervorgeht. Bei numerischen Ausgabefeldern enthält dieser Abschnitt die folgenden Informationen:

- **Minimaler Fehler.** Zeigt den minimalen Fehler (Differenz zwischen beobachteten und vorhergesagten Werten).
- **Maximaler Fehler.** Zeigt den maximalen Fehler.
- **Mittlerer Fehler.** Zeigt den Durchschnitt (Mittelwert) der Fehler über alle Datensätze hinweg. Hieraus geht hervor, ob ein systematischer **Fehler** (eine stärkere Tendenz zu Überbewertungen als zu Unterbewertungen und umgekehrt) im Modell vorliegt.
- **Mittlerer absoluter Fehler.** Zeigt den Durchschnitt der absoluten Werte der Fehler über alle Datensätze hinweg. Weist auf die durchschnittliche Fehlergröße hin, unabhängig von der Richtung.
- **Standardabweichung.** Zeigt die Standardabweichung der Fehler.
- **Lineare Korrelation.** Zeigt die lineare Korrelation zwischen den vorhergesagten und den tatsächlichen Werten. Diese Statistik reicht von -1,0 bis 1,0. Werte nahe +1,0 weisen auf eine starke positive Assoziation hin; dies bedeutet, dass hohe vorhergesagte Werte mit hohen tatsächlichen Werten verknüpft sind und entsprechend niedrige vorhergesagte Werte mit niedrigen tatsächlichen Werten. Werte nahe -1,0 weisen auf eine starke negative Assoziation hin; dies bedeutet, dass hohe vorhergesagte Werte mit niedrigen tatsächlichen Werten verknüpft sind und umgekehrt. Werte nahe 0,0 weisen auf eine schwache Assoziation hin; dies bedeutet, dass die vorhergesagten Werte relativ unabhängig von den tatsächlichen Werten sind. *Hinweis:* Ein leerer Eintrag hier gibt an, dass eine lineare Korrelation in diesem Fall nicht berechnet werden kann, da entweder die tatsächlichen oder die vorhergesagten Werte Konstanten sind.
- **Vorkommen.** Zeigt die Anzahl der Datensätze, die für die Analyse herangezogen wurden.

Fehlklassifizierungstabelle. Bei kategorialen Ausgabefeldern wird hier ein Unterabschnitt mit der Matrix angezeigt, wenn Sie in den Analyseoptionen eine Fehlklassifizierungstabelle angefordert haben. Die Zeilen stehen für tatsächlich beobachtete Werte, die Spalten für vorhergesagte Werte. Die Zelle in der Tabelle gibt die Anzahl der Datensätze für jede Kombination aus vorhergesagten und tatsächlichen Werten an.

Leistungsauswertung. Bei kategorialen Ausgabefeldern werden hier die Ergebnisse der Leistungsauswertung angezeigt, wenn Sie die Leistungsauswertungsstatistik in den Analyseoptionen angefordert haben. Jede Ausgabekategorie wird gemeinsam mit der zugehörigen Leistungsauswertungsstatistik aufgeführt.

Übersicht Konfidenzwerte für. Bei kategorialen Ausgabefeldern werden hier die Konfidenzwerte angezeigt, wenn Sie diese Werte in den Analyseoptionen angefordert haben. Für Modellkonfidenzwerte werden die folgenden Statistiken zusammengestellt:

- **Bereich.** Zeigt den Bereich (kleinster und größter Wert) der Konfidenzwerte für die Datensätze in den Streamdaten.
- **Mittelwert korrekt.** Zeigt die durchschnittliche Konfidenz für Datensätze, die als korrekt klassifiziert wurden.
- **Mittelwert inkorrekt.** Zeigt die durchschnittliche Konfidenz für Datensätze, die als inkorrekt klassifiziert wurden.
- **Immer korrekt über.** Zeigt den Schwellenwert für die Konfidenz, ab dem die Vorhersagen immer korrekt sind, sowie den Prozentsatz der Fälle, die dieses Kriterium erfüllen.
- **Immer korrekt unter.** Zeigt den Schwellenwert für die Konfidenz, bis zu dem die Vorhersagen immer korrekt sind, sowie den Prozentsatz der Fälle, die dieses Kriterium erfüllen.
- **X % Genauigkeit über.** Zeigt das Konfidenzniveau, bei dem die Genauigkeit bei *x* % liegt. *x* ist hierbei etwa gleich dem Wert, den Sie in den Analyseoptionen unter **Schwellenwert für** angegeben haben. Bei bestimmten Modellen und Datasets ist es nicht möglich, einen Konfidenzwert auszuwählen, der genau gleich dem in den Optionen angegebenen Schwellenwert ist (in der Regel aufgrund von Clustern ähnlicher Fälle, die denselben Konfidenzwert nahe dem Schwellenwert aufweisen). Der angezeigte Schwellenwert ist der bestmögliche Näherungswert an das angegebene Kriterium für die Genauigkeit, der mit einem einzigen Schwellenwert für den Konfidenzwert erzielt werden kann.
- **X-fach korrekt über.** Zeigt den Konfidenzwert, bei dem die Genauigkeit *x*-mal besser ist als für das gesamte Dataset. *x* ist hierbei etwa gleich dem Wert, den Sie in den Analyseoptionen unter **Genauigkeit verbessern** angegeben haben.

Übereinstimmung zwischen Enthält der Stream zwei oder mehr erzeugte Modelle, die eine Vorhersage für dasselbe Ausgabefeld enthalten, wird auch eine Statistik über die **Übereinstimmung** zwischen den durch die Modelle abgegebenen Vorhersagen angezeigt. Hierzu gehören die Anzahl und der Prozentsatz der Datensätze, bei denen die Vorhersagen übereinstimmen (bei kategorialen Ausgabefeldern), bzw. die Fehlerübersichtsstatistik (für stetige Ausgabefelder). Bei kategorialen Feldern wird eine Analyse der Vorhersagen im Vergleich zu den tatsächlichen Werten für das Subset der Datensätze angezeigt, bei denen die Modelle miteinander übereinstimmen (also denselben vorhergesagten Wert erzeugen).

Auswertungsmetriken. Wenn Sie für binäre Klassifikationsmerkmale Auswertungsmetriken in den Analyseoptionen angefordert haben, werden die Werte der AUC-Auswertungsmetrik und der Auswertungsmetrik für den Gini-Koeffizienten in einer Tabelle in diesem Abschnitt angezeigt. Die Tabelle enthält eine Zeile für jedes Modell für binäre Klassifikationsmerkmale. Die Tabelle mit den Auswertungsmetriken wird für jedes Ausgabefeld und nicht für jedes Modell angezeigt.

Data Audit-Knoten

Mit dem Data Audit-Knoten können Sie einen umfassenden ersten Blick auf die Daten werfen, die Sie in IBM SPSS Modeler einbringen. Die Ausgabe erfolgt in einer leicht lesbaren Matrix, die sortiert und zur Erstellung von normal großen Diagrammen und einer Vielzahl von Datenvorbereitungsknoten verwendet werden kann.

- Die Registerkarte "Audit" enthält einen Bericht mit Übersichtsstatistiken, Histogrammen und Verteilungsdiagrammen, die dabei helfen können, einen ersten Einblick in die Daten zu gewinnen. Der Bericht zeigt außerdem vor dem Feldnamen ein Symbol für den Speichertyp an.
- Auf der Registerkarte "Qualität" das Auditberichts finden Sie Informationen zu Ausreißern, Extremwerten und fehlenden Werten sowie Tools für den Umgang mit diesen Werten.

Verwenden des Data Audit-Knotens

Der Data Audit-Knoten kann direkt an einen Quellenknoten angehängt oder unterhalb eines instanziierten Typknotens eingefügt werden. Außerdem können Sie auf der Grundlage der Ergebnisse eine Reihe von Datenvorbereitungsknoten generieren. Beispielsweise können Sie einen Filterknoten generieren, der Felder ausschließt, die so viele fehlende Werte aufweisen, dass sie bei der Modellierung nicht mehr sinnvoll einsetzbar sind, und einen Superknoten generieren, der fehlende Werte für bestimmte oder alle verbleibenden Felder imputiert. Hier zeigt sich die wahre Stärke des Audits: Sie können nicht nur den aktuellen Status Ihrer Daten einschätzen, sondern auch aufgrund dieser Einschätzung aktiv werden.

Screening oder Stichprobennahme der Daten. Das anfängliche Audit ist besonders wirkungsvoll, wenn große Datenmengen anstehen. Aus diesem Grund kann ein Stichprobenknoten verwendet werden, mit dem Sie die Verarbeitungszeit während dieser ersten Exploration verkürzen können, indem Sie nur ein Subset der Datensätze auswählen. Der Data Audit-Knoten kann auch in Verbindung mit Knoten wie "Merkmalauswahl" und "Anomalieerkennung" in den Erkundungsphasen der Analyse verwendet werden.

Registerkarte "Einstellungen" beim Data Audit-Knoten

Auf der Registerkarte "Einstellungen" können Sie grundlegende Parameter für das Audit angeben.

Standard. Sie können beispielsweise den Knoten einfach an den Stream anhängen und auf **Ausführen** klicken. Auf diese Weise wird wie folgt ein Audit-Bericht für alle Felder erzeugt, der auf Standardeinstellungen beruht.

- Enthält der Typknoten keine Einstellungen, werden alle Felder in den Bericht aufgenommen.
- Falls Typeinstellungen vorhanden sind (unabhängig davon, ob diese instanziiert wurden oder nicht), werden alle Felder vom Typ *Eingabe*, *Ziel* und *Beides* in der Anzeige dargestellt. Liegt ein einzelnes Feld *Ziel* vor, wird dieses als Überlagerungsfeld herangezogen. Bei mehreren Feldern vom Typ *Ziel* wird keine Standardüberlagerung festgelegt.

Angepasste Felder verwenden. Mit dieser Option können Sie manuell Felder auswählen. Wählen Sie die Felder mit der Feldauswahlschaltfläche rechts einzeln oder nach Typ aus.

Überlagerungsfeld. Das Überlagerungsfeld dient zum Zeichnen der Piktogramme, die im Audit-Bericht angezeigt werden. Bei einem stetigen Feld (numerischer Bereich) werden außerdem bivariate Statistiken (Kovarianz und Korrelation) berechnet. Wenn ein einzelnes Feld vom Typ *Ziel* vorliegt, das auf den Einstellungen für den Knotentyp beruht, wird dieses, wie oben beschrieben, als Standard-Überlagerungsfeld verwendet. Alternativ können Sie **Benutzerdef. Felder verwenden** auswählen, um eine Überlagerung anzugeben.

Anzeigen. Ermöglicht die Angabe, ob Diagramme in der Ausgabe verfügbar sein sollen, sowie die Auswahl der standardmäßig anzuzeigenden Statistiken.

- **Diagramme.** Zeigt ein Diagramm für jedes ausgewählte Feld an. Dabei kann es sich um ein Verteilungsdiagramm (Balkendiagramm), ein Histogramm oder ein Streudiagramm handeln, je nachdem, was für die Daten geeignet ist. Die Diagramme werden im ursprünglichen Bericht als Piktogramm angezeigt, es können jedoch auch Diagramme in normaler Größe sowie Diagrammknoten generiert werden. Weitere Informationen finden Sie im Thema „Data Audit-Ausgabebrowser” auf Seite 296.
- **Basisstatistiken/Erweiterte Statistiken.** Gibt an, wie detailliert die Statistiken standardmäßig in der Ausgabe angezeigt werden sollen. Diese Einstellung legt zwar die ursprüngliche Anzeige fest, es sind

jedoch unabhängig von dieser Einstellung alle Statistiken in der Ausgabe verfügbar. Weitere Informationen finden Sie im Thema „Statistik anzeigen” auf Seite 297.

Median und Modus. Berechnet Median und Modus für alle Felder im Bericht. Beachten Sie, dass diese Statistiken bei großen Datasets die Verarbeitungszeit erhöhen können, da ihre Berechnung länger dauert als die anderer Statistiken. Beim Median (und nur dort) kann der gemeldete Wert in einigen Fällen auf einer Stichprobe von 2.000 Datensätzen (anstelle des vollständigen Datasets) beruhen. Diese Stichprobennahme erfolgt in Fällen, bei denen ansonsten die Arbeitsspeichergrenzen überschritten würden, auf der Grundlage einzelner Felder. Wenn die Stichprobennahme aktiviert wird, werden die Ergebnisse auch so in der Ausgabe beschriftet (*Median Stichpr.* anstatt *Median*). Alle anderen Statistiken als der Median werden immer mit dem vollständigen Dataset berechnet.

Leere Felder bzw. Felder ohne Typ. Bei Verwendung mit instanziierten Daten werden Felder ohne Typ nicht in den Audit-Bericht aufgenommen. Um Felder ohne Typ (einschließlich leerer Felder) aufzunehmen, wählen Sie in allen weiter oben im Stream liegenden Typknoten die Option **Alle Werte löschen**. Dadurch wird sichergestellt, dass keine Daten instanziiert und somit alle Felder in den Bericht aufgenommen werden. Dies kann beispielsweise dann nützlich sein, wenn Sie eine vollständige Liste aller Felder benötigen oder einen Filterknoten generieren möchten, der alle leeren Felder ausschließt. Weitere Informationen finden Sie im Thema „Filtern von Feldern mit fehlenden Daten” auf Seite 301.

Data Audit - Registerkarte "Qualität"

Die Registerkarte "Qualität" im Data Audit-Knoten bietet Optionen für den Umgang mit fehlenden Werten, Ausreißern und Extremwerten.

Missing Values (Fehlende Werte)

- **Anzahl der Datensätze mit gültigen Werten.** Mit dieser Option rufen Sie die Anzahl der Datensätze ab, die gültige Werte für alle ausgewählten Felder enthalten. Beachten Sie, dass Nullwerte (nicht definierte Werte), Leerwerte, leere Bereiche und leere Zeichenfolgen stets als ungültige Werte behandelt werden.
- **Aufgeschlüsselte Anzahl der Datensätze mit ungültigen Werten.** Mit dieser Option ermitteln Sie die Anzahl der Datensätze sowie die verschiedenen Typen der ungültigen Werte für jedes Feld ab.

Ausreißer und Extremwerte

Erkennungsmethode für Ausreißer und Extremwerte. Es werden zwei Methoden unterstützt:

Standardabweichung vom Mittelwert. Erkennt Ausreißer und Extremwerte anhand der Anzahl an Standardabweichungen vom Mittelwert. Nehmen wir beispielsweise an, Sie haben ein Feld mit einem Mittelwert von 100 und einer Standardabweichung von 10. In diesem Fall könnten Sie 3,0 angeben, um festzulegen, dass jeder Wert unter 70 und über 130 als Ausreißer behandelt werden soll.

Interquartilbereich. Erkennt Ausreißer und Extremwerte anhand des Interquartilbereichs, also des Bereichs, in den die beiden mittleren Quartile fallen (zwischen dem 25. und dem 75. Perzentil). Auf der Grundlage der Standardeinstellung 1,5 beispielsweise wäre der untere Schwellenwert für Ausreißer Q1 - 1,5 * IQB und der obere Schwellenwert wäre Q3 + 1,5 * IQB. Beachten Sie, dass diese Option bei großen Datasets zu Geschwindigkeitseinbußen führen kann.

Data Audit-Ausgabebrowser

Der Data Audit-Browser ist ein leistungsstarkes Tool, mit dem Sie einen Überblick über Ihre Daten gewinnen können. Auf der Registerkarte "Audit" werden Piktogramme, Symbole für den Speichertyp und Statistiken für alle Felder angezeigt, auf der Registerkarte "Qualität" finden Sie Informationen zu Ausreißern, Extremwerten und fehlenden Werten. Auf der Grundlage der anfänglichen Diagramme und der Übersichtsstatistik können Sie beispielsweise ein numerisches Feld neu codieren, ein neues Feld ableiten oder auch die Werte eines nominalen Felds umcodieren. Des Weiteren können Sie die Exploration mithilfe ei-

ner ausgereiften Visualisierung fortsetzen. Dies können Sie über den Browser für Audit-Berichte erreichen, indem Sie mithilfe des Menüs "Generieren" eine Reihe von Knoten erzeugen, die zum Transformieren bzw. Visualisieren der Daten verwendet werden können.

- Sortieren Sie die Spalten durch Klicken auf die gewünschte Spaltenüberschrift oder ordnen Sie die Spalten durch Ziehen und Ablegen neu. Außerdem werden die meisten Standardausgabeoperationen unterstützt. Weitere Informationen finden Sie im Thema „Anzeigen der Ausgabe" auf Seite 283.
- Werte und Bereiche für Felder abrufen: Doppelklicken Sie auf ein Feld in der Spalte "Messung" oder "Eindeutig".
- Verwenden Sie die Symbolleiste oder das Menü "Bearbeiten", um Wertbeschriftungen ein- bzw. auszublenden bzw. um die anzuzeigenden Statistiken auszuwählen. Weitere Informationen finden Sie im Thema „Statistik anzeigen".
- Überprüfen Sie die Speichertypsymbole links neben den Feldnamen. Der Speichertyp beschreibt die Art und Weise, wie Daten in einem Feld gespeichert werden. Beispiel: Ein Feld mit den Werten 1 und 0 speichert ganzzahlige Daten. Dies ist vom Messniveau zu unterscheiden, das die Verwendung der Daten beschreibt und sich nicht auf den Speichertyp auswirkt. Weitere Informationen finden Sie im Thema „Festlegen von Feldspeicher und Formatierung" auf Seite 24.

Anzeigen und Generieren von Diagrammen

Ist keine Überlagerung ausgewählt, werden auf der Registerkarte "Audit" entweder Balkendiagramme (für nominale oder Flagfelder) oder Histogramme (stetige Felder) angezeigt.

Bei Überlagerung mit einem nominalen oder Flagfeld werden die Diagramme gemäß den Werten der Überlagerung farbig gekennzeichnet.

Bei Überlagerung mit einem stetigen Feld werden keine eindimensionalen Balkendiagramme und Histogramme generiert, sondern zweidimensionale Streudiagramme. In diesem Fall wird die *X*-Achse dem Überlagerungsfeld zugeordnet, sodass bei allen *X*-Achsen in der gesamten Tabelle jeweils dieselbe Skala verwendet wird.

- Halten Sie bei Flagfeldern oder nominalen Feldern den Mauscursor über einen Balken, um den zugrunde liegenden Wert bzw. die Beschriftung in einer QuickInfo anzuzeigen.
- Verwenden Sie bei Flagfeldern oder nominalen Feldern die Symbolleiste, um die Ausrichtung der Piktogramme von horizontal zu vertikal zu ändern.
- Um ein normal großes Diagramm aus einem Piktogramm zu generieren, doppelklicken Sie auf dem Piktogramm oder wählen Sie ein Piktogramm aus und wählen Sie im Menü "Generieren" die Option **Diagrammausgabe**. *Hinweis:* Beruht ein Piktogramm auf Stichprobendaten, werden alle Fälle in das generierte Diagramm aufgenommen, wenn der ursprüngliche Datenstream noch geöffnet ist.

  Sie können auch ein Diagramm generieren, wenn der Data-Audit-Knoten, der die Ausgabe erstellt hat, mit dem Stream verbunden ist.
- Um einen passenden Diagrammknoten zu generieren, wählen Sie auf der Registerkarte "Audit" mindestens ein Feld aus und wählen Sie im Menü "Generieren" die Option **Diagrammknoten**. Der so entstehende Knoten wird dem Streamerstellungsbereich hinzugefügt und kann verwendet werden, um das Diagramm bei jeder Ausführung des Streams neu zu erstellen.
- Enthält eine Überlagerung mehr als 100 Werte, wird eine Warnnachricht eingeblendet und die Überlagerung wird nicht berücksichtigt.

Statistik anzeigen

Im Dialogfeld" Statistik anzeigen" können Sie die auf der Registerkarte "Audit" anzuzeigenden Statistiken auswählen. Die ursprünglichen Einstellungen werden im Data Audit-Knoten angegeben. Weitere Informationen finden Sie im Thema „Registerkarte "Einstellungen" beim Data Audit-Knoten" auf Seite 295.

Minimum. Der kleinste Wert einer numerischen Variablen.

Maximum. Der größte Wert einer numerischen Variablen.

Summe. Die Summe der Werte über alle Fälle mit nicht fehlenden Werten.

Bereich. Die Differenz zwischen den größten und kleinsten Werten einer numerischen Variablen; Maximalwert minus Minimalwert.

Mittelwert. Ein Lagemaß (zentrale Tendenz). Die Summe der Ränge, geteilt durch die Zahl der Fälle.

Standardfehler des Mittelwerts. Ein Maß dafür, wie stark der Mittelwert von Stichprobe zu Stichprobe in derselben Verteilung variieren kann. Dieser Wert kann für einen ungefähren Vergleich des beobachteten Mittelwerts mit einem hypothetischen Wert verwendet werden. (Es kann geschlossen werden, dass die beiden Werte unterschiedlich sind, wenn das Verhältnis der Differenz zum Standardfehler kleiner als -2 oder größer als +2 ist.)

Standardabweichung. Ein Maß für die Streuung um den Mittelwert, definiert als Quadratwurzel aus der Varianz. Die Standardabweichung wird in denselben Einheiten gemessen wie die ursprüngliche Variable.

Varianz. Ein Maß der Streuung um den Mittelwert, gleich der Summe der quadrierten Abweichungen vom Mittelwert geteilt durch eins weniger als die Anzahl der Fälle. Die Maßeinheit der Varianz ist das Quadrat der Maßeinheiten der Variablen.

Schiefe. Ein Maß der Asymmetrie der Verteilung. Die Normalverteilung ist symmetrisch, ihre Schiefe hat den Wert 0. Eine Verteilung mit einer deutlichen positiven Schiefe läuft nach rechts lang aus (lange rechte Flanke). Eine Verteilung mit einer deutlichen negativen Schiefe läuft nach links lang aus (lange linke Flanke). Als Faustregel kann man verwenden, dass ein Schiefewert, der mehr als doppelt so groß ist wie sein Standardfehler, als Abweichung von der Symmetrie gilt.

Standardfehler der Schiefe. Der Quotient aus der Schiefe und deren Standardfehler kann als Test auf Normalverteilung verwendet werden. (Sie können die Normalverteilung ausschließen, wenn der Quotient unter -2 oder über +2 liegt.) Ein großer positiver Wert für die Schiefe bedeutet, dass die Verteilung eine lange rechte Flanke hat; ein extremer negativer Wert bedeutet, dass sie eine lange linke Flanke hat.

Kurtosis. Ein Maß dafür, wie sehr die Beobachtungen um einen zentralen Punkt gruppiert sind. Bei einer Normalverteilung ist der Wert der Kurtosis gleich 0. Bei positiver Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung enger um das Zentrum der Verteilung gruppiert und haben dünnere Flanken bis hin zu den Extremwerten der Verteilung. Ab dort sind die Flanken der leptokurtischen Verteilung im Vergleich zu einer Normalverteilung dicker. Bei negativer Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung weniger eng gruppiert und haben dickere Flanken bis hin zu den Extremwerten der Verteilung. Ab dort sind die Flanken der platykurtischen Verteilung im Vergleich zu einer Normalverteilung dünner.

Standardfehler der Kurtosis. Der Quotient aus der Kurtosis und deren Standardfehler kann als Test auf Normalverteilung verwendet werden. (Sie können die Normalverteilung ausschließen, wenn der Quotient unter -2 oder über +2 liegt.) Ein großer positiver Wert für die Kurtosis deutet darauf hin, dass die Flanken der Verteilung länger sind als bei einer Normalverteilung; ein negativer Wert bedeutet, dass sie kürzer sind (etwa wie bei einer kastenförmigen, gleichförmigen Verteilung).

Eindeutig. Bewertet alle Effekte gleichzeitig; damit werden alle Effekte an alle sonstigen Effekte jedweden Typs angepasst.

Gültig. Gültige Fälle, d. h. solche, die weder den systemdefiniert fehlenden Wert noch einen benutzerdefiniert fehlenden Wert aufweisen.

Median. Wert, über und unter dem jeweils die Hälfte der Fälle liegt; 50. Perzentil. Bei einer geraden Anzahl von Fällen ist der Median der Mittelwert der beiden mittleren Fälle, wenn diese auf- oder absteigend sortiert sind. Der Median ist ein Lagemaß, das gegenüber Ausreißern unempfindlich ist (im Gegensatz zum Mittelwert, der durch wenige extrem niedrige oder hohe Werte beeinflusst werden kann).

Modalwert. Der am häufigsten auftretende Wert. Wenn mehrere Werte gleichermaßen die größte Häufigkeit aufweisen, ist jeder von ihnen ein Modalwert.

Beachten Sie, dass Median und Modus standardmäßig unterdrückt sind, um die Leistungsfähigkeit zu erhöhen. Diese Statistiken können jedoch im Data Audit-Knoten auf der Registerkarte "Einstellungen" ausgewählt werden. Weitere Informationen finden Sie im Thema „Registerkarte "Einstellungen" beim Data Audit-Knoten" auf Seite 295.

Statistiken für Überlagerungen

Wenn ein stetiges Überlagerungsfeld (numerischer Bereich) verwendet wird, stehen außerdem folgende Statistiken zur Verfügung:

Kovarianz. Ein nicht standardisiertes Maß für den Zusammenhang zwischen zwei Variablen. Es ist gleich der Kreuzproduktabweichung geteilt durch N-1.

Registerkarte "Qualität" beim Data Audit-Browser

Auf der Registerkarte "Qualität" des Data Audit-Browsers werden die Ergebnisse der Datenqualitätsanalyse angezeigt. Außerdem können Sie hier angeben, wie Ausreißer, Extremwerte und fehlende Werte behandelt werden sollen.

Fehlende Werte imputieren: Der Audit-Bericht listet den Prozentsatz vollständiger Datensätze für die einzelnen Felder auf, dazu die Anzahl der gültigen Werte, der Nullwerte und der leeren Werte. Sie können je nach Bedarf fehlende Werte für bestimmte Felder imputieren und anschließend einen Superknoten generieren, um die Transformationen anzuwenden.

1. Geben Sie in der Spalte **Fehlende Werte imputieren** die zu imputierenden Wertetypen an, sofern vorhanden. Sie können festlegen, dass Leerstellen, Nullen oder beides imputiert werden sollen, oder eine benutzerdefinierte Bedingung bzw. einen benutzerdefinierten Ausdruck angeben, der die zu imputierenden Werte auswählt.

   In Clementine gibt es mehrere Arten von fehlenden Werten, die von IBM SPSS Modeler erkannt werden:

   - **Nullwerte oder systemdefiniert fehlende Werte.** Bei diesen Werten handelt es sich um Nicht-Zeichenfolgewerte, die in der Datenbank bzw. der Quellendatei leer gelassen und nicht speziell in einem Quellen- oder Typknoten als "fehlend" definiert wurden. Systemdefiniert fehlende Werte werden als **$null$** angezeigt. Beachten Sie, dass leere Zeichenfolgen in IBM SPSS Modeler nicht als Nullen betrachtet werden, auch wenn sie von bestimmten Datenbanken (siehe unten) als Nullen behandelt werden können.
   - **Leere Zeichenfolgen und leere Bereiche.** Leere Zeichenfolgewerte und leere Bereiche (Zeichenfolgen ohne sichtbare Zeichen) werden anders als Nullwerte behandelt. Leere Zeichenfolgen werden in den meisten Fällen als äquivalent mit leeren Bereichen (Leerzeichen) behandelt. Beispiel: Wenn Sie die Option auswählen, dass leere Bereiche in einem Quellen- oder Typknoten als Leerstellen behandelt werden sollen, gilt diese Einstellung auch für leere Zeichenfolgen.
   - **Leere oder benutzerdefiniert fehlende Werte.** Es handelt sich hierbei um Werte wie `unbekannt`, `99` oder `-1`, die in einem Quellen- oder Typknoten ausdrücklich als fehlend definiert sind. Optional können Sie auch auswählen dass Nullen und leere Bereiche als Leerzeichen behandelt werden sollen. Dadurch können sie mit Flags für eine spezielle Behandlung versehen und aus den meisten Berechnungen ausgeschlossen werden. Beispielsweise können Sie die Funktion `@BLANK` verwenden, um diese Werte gemeinsam mit anderen Arten von fehlenden Werten als Leerstellen zu behandeln.
2. Geben Sie in der Spalte **Methode** die zu verwendende Methode an.

   Folgende Methoden stehen zum Imputieren fehlender Werte zur Verfügung:

   Fest. Ersetzt einen festen Wert (Feldmittelwert, Mittelpunkt des Bereichs oder eine von Ihnen angegebene Konstante).

   Zufällig. Ersetzt einen Zufallswert auf der Grundlage einer Normal- oder Gleichverteilung.

Ausdruck. Ermöglicht die Angabe eines benutzerdefinierten Ausdrucks. Beispielsweise könnten Sie Werte durch eine globale Variable ersetzen, die vom Globalwerteknoten erstellt wurde.

Algorithmus. Ersetzt einen von einem Modell vorhergesagten Wert auf der Grundlage eines C&RT-Algorithmus. Für jedes Feld, das unter Verwendung dieser Methode imputiert wurde, gibt es ein separates C&RT-Modell sowie einen Füllerknoten, der Leerstellen und Nullen durch den vom Modell vorhergesagten Wert ersetzt. Anschließend werden die vom Modell generierten Vorhersagefelder mithilfe eines Filterknotens entfernt.

3. Um einen Superknoten für fehlende Werte zu generieren, wählen Sie folgende Optionsfolge aus den Menüs aus:

   Generieren > **Superknoten für fehlende Werte**

   Das Dialogfeld "Superknoten für fehlende Werte" wird angezeigt.
4. Wählen Sie **Alle Felder** oder **Nur ausgewählte Felder** aus und geben Sie bei Bedarf einen Stichprobenumfang an. (Die Stichprobe wird als Prozentsatz angegeben; standardmäßig werden 10 % aller Datensätze in die Stichprobe aufgenommen.)
5. Klicken Sie auf **OK**, um den generierten Superknoten zum Streamerstellungsbereich hinzuzufügen.
6. Fügen Sie den Superknoten zum Stream hinzu, um die Transformationen anzuwenden.

Innerhalb des Superknotens wird je nach Bedarf eine Kombination aus Modellnugget, Füllerknoten und Filterknoten verwendet. Um Einblicke in die Funktionsweise zu gewinnen, können Sie den Superknoten bearbeiten und auf **Vergrößern** klicken. Außerdem können Sie einzelne Knoten im Superknoten hinzufügen, bearbeiten bzw. entfernen, um eine Feineinstellung des Verhaltens vorzunehmen.

Umgang mit Ausreißern und Extremwerten: Im Audit-Bericht wird die Anzahl der Ausreißer und Extremwerte für die einzelnen Felder auf der Grundlage der im Data Audit-Knoten angegebenen Erkennungsoptionen aufgeführt. Weitere Informationen finden Sie im Thema „Data Audit - Registerkarte "Qualität"" auf Seite 296. Sie können je nach Bedarf diese Werte für bestimmte Felder erzwingen, verwerfen oder auf Nullwert setzen und anschließend einen Superknoten generieren, um die Transformationen anzuwenden.

1. Geben Sie in der Spalte **Aktion** den gewünschten Umgang mit Ausreißern und Extremwerten für bestimmte Felder an.

   Für den Umgang mit Ausreißern und Extremwerten stehen folgende Aktionen zur Auswahl:

   - **Erzwingen.** Ersetzt Ausreißer und Extremwerte durch den nächsten Wert, der nicht als Extremwert betrachtet würde. Wenn beispielsweise alle Werte als Ausreißer gelten, die den Bereich von drei Standardabweichungen über- bzw. unterschreiten, werden alle Ausreißer mit dem höchsten bzw. niedrigsten Wert innerhalb dieses Bereichs ersetzt.
   - **Verwerfen.** Verwirft Datensätze mit Ausreißern bzw. Extremwerten für das angegebene Feld.
   - **Auf Nullwert setzen.** Ersetzt Ausreißer und Extremwerte durch den Nullwert bzw. den systemdefiniert fehlenden Wert.
   - **Ausreißer erzwingen/Extremwerte verwerfen.** Verwirft nur Extremwerte.
   - **Ausreißer erzwingen/Extremwerte auf Nullwert setzen.** Setzt nur Extremwerte auf Nullwert.
2. Um den Superknoten zu generieren, wählen Sie folgende Optionsfolge aus den Menüs aus:

   Generieren > **Superknoten für Ausreißer & Extremwerte**

   Das Dialogfeld "Ausreißersuperknoten" wird angezeigt.
3. Wählen Sie **Alle Felder** oder **Nur ausgewählte Felder** aus und klicken Sie dann auf **OK**, um den generierten Superknoten zum Streamerstellungsbereich hinzuzufügen.
4. Fügen Sie den Superknoten zum Stream hinzu, um die Transformationen anzuwenden.

Optional können Sie den Superknoten bearbeiten und vergrößern, um ihn zu durchsuchen oder Änderungen vorzunehmen. Innerhalb des Superknotens werden Werte mithilfe einer Reihe von Auswahl- und/oder Füllerknoten verworfen, erzwungen oder auf null gesetzt.

Filtern von Feldern mit fehlenden Daten: Über den Data Audit-Browser können Sie mithilfe des Dialogfelds "Filter aus Qualität generieren" einen neuen Filterknoten auf der Grundlage der Ergebnisse aus der Qualitätsanalyse erstellen.

Modalwert. Wählen Sie die gewünschte Funktion für die ausgewählten Felder aus (**Einschließen** oder **Ausschließen**).

- **Ausgewählte Felder.** Mit dem Filterknoten werden die Felder eingeschlossen bzw. ausgeschlossen, die in der Qualitätstabelle ausgewählt wurden. Beispielsweise können Sie die Tabelle anhand der Spalte **% Vollständig** sortieren, durch Klicken bei gedrückter Umschalttaste die am wenigsten vollständigen Felder auswählen und anschließend einen Filterknoten generieren, der diese Felder ausschließt.
- **Felder mit einem Qualitätsprozentsatz größer als.** Mit dem Filterknoten werden die Felder eingeschlossen bzw. ausgeschlossen, bei dem der Prozentsatz der abgeschlossenen Datensätze höher ist als der angegebene Schwellenwert. Der Standardschwellenwert beträgt 50 %.

Filtern von leeren Feldern bzw. Feldern ohne Typ

Beachten Sie: Nach dem Instanziieren von Datenwerten werden Felder ohne Typ bzw. leere Felder aus den Audit-Ergebnissen und den meisten anderen Ausgaben in IBM SPSS Modeler ausgeschlossen. Diese Felder werden zum Zweck der Modellierung ignoriert, können die Daten jedoch aufblähen bzw. unübersichtlich werden lassen. In diesem Fall können Sie mit dem Data Audit-Browser einen Filterknoten generieren, der diese Felder aus dem Stream entfernt.

1. Um sicherzustellen, dass alle Felder in das Audit aufgenommen werden, einschließlich leerer Felder und Feldern ohne Typen, wählen Sie im weiter oben im Stream liegenden Quellen- oder Typknoten die Option **Alle Werte löschen** oder setzen Sie für alle Felder "Werte" auf <*Übergeben*>.
2. Sortieren Sie die Daten im Data Audit-Browser anhand der Spalte **% Vollständig**, wählen Sie die Felder mit 0 (oder anderer Schwellenwert) gültigen Werten aus und erstellen Sie über das Menü "Generieren" einen Filterknoten, der zum Stream hinzugefügt werden kann.

Auswählen von Datensätzen mit fehlenden Daten: Über den Data Audit-Browser können Sie einen neuen Auswahlknoten auf der Grundlage der Ergebnisse aus der Qualitätsanalyse erstellen.

1. Wählen Sie im Data Audit-Browser die Registerkarte "Qualität" aus.
2. Wählen Sie die folgenden Befehle aus dem Menü aus:

   Generieren > **Auswahlknoten für fehlende Werte**

   Das Dialogfeld "Auswahlknoten generieren" wird angezeigt.

Auswählen, wenn Datensatz. Geben Sie an, ob die Datensätze beibehalten werden sollen, wenn diese **Gültig** oder **Ungültig** sind.

Ungültige Werte suchen in. Geben Sie an, wo nach ungültigen Werten gesucht werden soll.

- **Alle Felder.** Mit dem Auswahlknoten werden alle Felder auf ungültige Werte geprüft.
- **Ausgewählte Felder in Tabelle.** Mit dem Auswahlknoten werden nur die Felder geprüft, die derzeit in der Qualitätsausgabetabelle ausgewählt sind.
- **Felder mit einem Qualitätsprozentsatz größer als.** Mit dem Auswahlknoten werden alle Felder geprüft, bei dem der Prozentsatz der abgeschlossenen Datensätze höher ist als der angegebene Schwellenwert. Der Standardschwellenwert beträgt 50 %.

Datensatz als ungültig betrachten, wenn ein ungültiger Wert vorliegt in. Legen Sie die Bedingung fest, unter der ein Datensatz als ungültig betrachtet wird.

- **Eines der aufgeführten Felder.** Mit dem Auswahlknoten wird ein Datensatz als ungültig betrachtet, wenn *eines* der oben angegebenen Felder einen ungültigen Wert für diesen Datensatz enthält.
- **Alle aufgeführten Felder.** Mit dem Auswahlknoten wird ein Datensatz als ungültig betrachtet, wenn *alle* oben angegebenen Felder einen ungültigen Wert für diesen Datensatz enthalten.

Erzeugen von anderen Knoten zur Datenvorbereitung

Zahlreiche Knoten für die Datenvorbereitung können direkt über den Data Audit-Browser erzeugt werden, beispielsweise Umcodierungs-, Klassier- und Ableitungsknoten. Beispiel:

- Soll ein neues Feld auf der Grundlage der Werte für *Schadensersatzforderung* und *Grundwert* erzeugt werden, wählen Sie beide Felder im Audit-Bericht aus und wählen Sie dann im Menü "Generieren" den Befehl **Ableiten**. Der neue Knoten wird in den Streamerstellungsbereich aufgenommen.
- Unter Umständen stellen Sie auf der Grundlage der Audit-Ergebnisse fest, dass eine Umcodierung von *Grundwert* in perzentilbasierte Klassen eine stärker zielgerichtete Analyse ergäbe. Um einen Klassierknoten zu erzeugen, wählen Sie die Feldzeile in der Anzeige aus und wählen Sie dann im Menü "Generieren" den Befehl **Klassieren**.

Sobald Sie einen Knoten erzeugt und zum Streamerstellungsbereich hinzugefügt haben, muss dieser Knoten an den Stream angehängt und geöffnet werden; legen Sie dann die Optionen für das oder die ausgewählten Felder fest.

Transformationsknoten

Die Normalisierung der Eingabefelder ist ein wichtiger Schritt vor der Anwendung herkömmlicher Scoring-Verfahren wie Regression, logistische Regression und Diskriminanzanalyse. Bei diesen Verfahren wird von Annahmen über die Normalverteilung von Daten ausgegangen, die für viele Rohdatendateien möglicherweise nicht gelten. Ein Ansatz für den Umgang mit realen Daten besteht in der Anwendung von Transformationen, die ein Rohdatenelement mehr in Richtung einer Normalverteilung verschieben. Außerdem können normalisierte Felder leicht miteinander verglichen werden. So befinden sich Einkommen und Alter in einer Rohdatendatei auf vollständig unterschiedlichen Skalen, nach einer Normalisierung jedoch lassen sich die relativen Auswirkungen der beiden Datenarten leicht interpretieren.

Der Transformationsknoten enthält einen Ausgabeviewer, mit dem Sie eine schnelle Sichtprüfung zur Ermittlung der besten Transformation durchführen können Sie sehen auf einen Blick, ob die Variablen normal verteilt sind und können, falls erforderlich, die gewünschte Transformation auswählen und anwenden. Sie können mehrere Felder auswählen und pro Feld jeweils eine Transformation durchführen.

Nach Auswahl der bevorzugten Transformationen für die Felder können Sie Ableitungs- oder Füllerknoten generieren, die die Transformationen durchführen, und diese Knoten zum Stream hinzufügen. Der Ableitungsknoten erstellt neue Felder, während der Füllerknoten die bestehenden transformiert. Weitere Informationen finden Sie im Thema „Generieren von Diagrammen” auf Seite 305.

Registerkarte "Felder" beim Transformationsknoten

Auf der Registerkarte "Felder" können Sie angeben, welche Datenfelder zur Anzeige und Anwendung möglicher Transformationen verwendet werden sollen. Eine Transformation ist nur bei numerischen Feldern möglich. Klicken Sie auf die Feldauswahlschaltfläche und wählen Sie mindestens ein numerisches Feld aus der angezeigten Liste aus.

Registerkarte "Optionen" beim Transformationsknoten

Auf der Registerkarte "Optionen" können Sie den Typ der einzuschließenden Transformationen angeben. Sie können auswählen, dass alle verfügbaren Transformationen eingeschlossen werden sollen, oder alle Transformationen einzeln auswählen.

Im letzteren Fall können Sie außerdem einen Wert für den Offset der Daten für Kehrwert- und Logarithmustransformationen eingeben. Dies ist nützlich in Situationen, bei denen ein großer Anteil von Nullen in den Daten die Ergebnisse für Mittelwert und Standardabweichung verzerren würde.

Beispiel: Angenommen, Sie haben ein Feld namens *BALANCE* mit einigen 0-Werten und möchten die inversen Transformation darauf anwenden. Um unerwünschte Verzerrungen zu vermeiden, wählen Sie

Kehrwert (1/x) aus und geben im Feld **Datenoffset verwenden** den Wert 1 ein. (Beachten Sie, dass dieses Offset nichts mit dem Offset zu tun hat, das durch die Sequenzfunktion `@OFFSET` in IBM SPSS Modeler ausgeführt wird.)

Alle Formeln. Gibt an, dass alle verfügbaren Transformationen berechnet und in der Ausgabe angezeigt werden sollen.

Formeln auswählen. Ermöglicht die Auswahl anderer Transformationen für Berechnung und Anzeige in der Ausgabe.

- **Invers (1/x).** Gibt an, dass die inversen Transformation berechnet und in der Ausgabe angezeigt werden soll.
- **Log (log n).** Gibt an, dass die Transformation $\log_n$ berechnet und in der Ausgabe angezeigt werden soll.
- **Log (log 10).** Gibt an, dass die Transformation $\log_{10}$ berechnet und in der Ausgabe angezeigt werden soll.
- **Exponentiell.** Gibt an, dass die exponentielle Transformation (e^x) berechnet und in der Ausgabe angezeigt werden soll.
- **Quadratwurzel.** Gibt an, dass die Quadratwurzeltransformation berechnet und in der Ausgabe angezeigt werden soll.

Registerkarte "Ausgabe" beim Transformationsknoten

Auf der Registerkarte "Ausgabe" legen Sie Format und Position der Ausgabe fest. Sie können auswählen, dass die Ergebnisse auf dem Bildschirm angezeigt werden sollen, oder sie an einen der Standarddateitypen senden. Weitere Informationen finden Sie im Thema „Registerkarte "Ausgabe" beim Ausgabeknoten” auf Seite 286.

Transformationsknoten - Ausgabeviewer

Im Ausgabeviewer werden die Ergebnisse aus der Ausführung des Transformationsknotens angezeigt. Der Viewer ist ein leistungsstarkes Tool, das mehrere Transformationen pro Feld in Piktogrammen der Transformation anzeigt, sodass ein schneller Vergleich der Felder möglich ist. Mit den Optionen im zugehörigen Menü "Datei" können Sie die Ausgaben speichern, exportieren bzw. drucken. Weitere Informationen finden Sie im Thema „Anzeigen der Ausgabe” auf Seite 283.

Für jede Transformation (mit Ausnahme von "Ausgewählte Transformation") wird unterhalb eine Legende mit folgendem Format angezeigt:

```
Mittelwert (Standardabweichung)
```

Generieren von Knoten für die Transformationen

Der Ausgabeviewer bildet einen soliden Ausgangspunkt für die Vorbereitung der Daten. Beispielsweise können Sie das Feld *ALTER* normalisieren, um die Möglichkeit zu erhalten, ein Scoring-Verfahren (z. B. logistische Regression oder Diskriminanzanalyse) anzuwenden, das eine Normalverteilung voraussetzt. Auf der Grundlage der ursprünglichen Diagramme und Übersichtsstatistiken könnten Sie sich entscheiden, das Feld *ALTER* gemäß einer bestimmten Verteilung zu transformieren (z. B. .log). Nach Auswahl der bevorzugten Verteilung können Sie anschließend einen Ableitungsknoten mit einer standardisierten Transformation für die Verwendung beim Scoring generieren.

Sie können folgende Feldoperationsknoten aus dem Ausgabeviewer generieren:

- Ableiten
- Füller

Ein Ableitungsknoten erstellt neue Felder mit den gewünschten Transformationen, während der Füllerknoten bestehende Felder transformiert. Die Knoten werden in Form eines Superknotens im Erstellungsbereich platziert.

Wenn Sie dieselbe Transformation für verschiedene Felder auswählen, enthält ein Ableitungs- bzw. Füllerknoten die Formeln für den betreffenden Transformationstyp für alle Felder, für die die Transformation gilt. Nehmen wir beispielsweise an, dass Sie die in der folgenden Tabelle aufgeführten Felder und Transformationen ausgewählt haben, um einen Ableitungsknoten zu generieren.

Tabelle 50. Beispiel für die Generierung eines Ableitungsknotens.

Feld	**Transformation**
AGE	Aktuelle Verteilung
INCOME	Log
OPEN_BAL	Invers
BALANCE	Invers

Folgende Knoten sind im Superknoten enthalten:

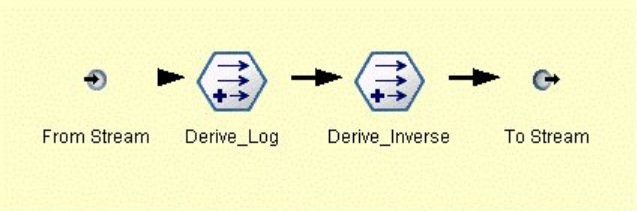


Abbildung 66. Superknoten im Erstellungsbereich

In diesem Beispiel weist der Knoten Derive_Log die log-Formel für das Feld *INCOME* und der Knoten Derive_Inverse weist die Kehrwertformeln für die Felder *OPEN_BAL* und *BALANCE* auf.

So generieren Sie einen Knoten:

1. Wählen Sie für jedes Feld im AusgabeViewer die gewünschte Transformation.
2. Wählen Sie im Menü "Generieren" die Option **Ableitungsknoten** bzw. **Füllerknoten**.

   Auf diese Weise wird das Dialogfeld "Ableitungsknoten generieren" bzw. "Füllerknoten generieren" angezeigt.

   Wählen Sie **Nicht standardisierte Transformation** bzw. **Standardisierte Transformation (Z-Score)**. Die zweite Option wendet einen *z*-Score auf die Transformation an; *z*-Scores stellen Werte als Funktion der Distanz vom Mittelwert der Variablen in Standardabweichungen dar. Wenn Sie beispielsweise die logarithmische Transformation auf das Feld *ALTER* anwenden und eine standardisierte Transformation auswählen, lautet die endgültige Gleichung für den generierten Knoten wie folgt:

   ```
   (log(AGE)-Mittwelwert)/SD
   ```

Gehen Sie wie folgt vor, sobald ein Knoten generiert wurde und im Streamerstellungsbereich angezeigt wird:

1. Fügen Sie ihn zum Stream hinzu.
2. Bei einem Superknoten können Sie optional auf den Knoten doppelklicken, um seinen Inhalt anzuzeigen.
3. Optional können Sie auf einen Ableitungs- oder Füllerknoten doppelklicken, um die Optionen für die ausgewählten Felder zu ändern.

Generieren von Diagrammen: Sie können aus einem Histogramm, das als Piktogramm erscheint, im Ausgabeviewer eine Histogrammausgabe in normaler Größe generieren.

So generieren Sie ein Diagramm:

1. Doppelklicken Sie auf ein Piktogramm im Ausgabeviewer.
   oder
   Wählen Sie ein Piktogramm im Ausgabeviewer aus.
2. Wählen Sie im Menü "Generieren" den Befehl **Diagrammausgabe** aus.

Dadurch wird das Histogramm mit einer überlagerten Normalverteilungskurve angezeigt. So können Sie vergleichen, wie eng die einzelnen Transformationen mit einer Normalverteilung übereinstimmen.

Hinweis: Sie können auch ein Diagramm generieren, wenn der Transformationsknoten, der die Ausgabe erstellt hat, mit dem Stream verbunden ist.

Andere Operationen: Im Ausgabeviewer haben Sie außerdem folgende Möglichkeiten:

- Sortieren des Ausgaberasters nach der Spalte "Feld".
- Exportieren der Ausgabe in eine HTML-Datei. Weitere Informationen finden Sie im Thema „Exportieren von Ausgaben" auf Seite 285.

Statistikknoten

Der Statistikknoten liefert grundlegende Übersichtsdaten zu numerischen Feldern. Die Übersichtsstatistiken können für einzelne Felder und für die Korrelationen zwischen den Feldern abgerufen werden.

Registerkarte "Einstellungen" beim Statistikknoten

Prüfen. Wählen Sie das oder die Felder aus, für die eine separate Übersichtsstatistik zusammengestellt werden soll. Sie können mehrere Felder auswählen.

Statistiken. Wählen Sie die zu bildenden Statistiken aus. Die folgenden Optionen stehen zur Auswahl: **Anzahl**, **Mittelwert**, **Summe**, **Min**, **Max**, **Bereich**, **Varianz**, **Std.abw.**, **Standardfehler für Mittelwert**, **Median** und **Modus**.

Korrelieren. Wählen Sie das oder die zu korrelierenden Felder aus. Sie können mehrere Felder auswählen. Wenn Sie Korrelationsfelder auswählen, wird die Korrelation zwischen jedem Untersuchungsfeld und den Korrelationsfeldern in der Ausgabe aufgeführt.

Korrelationseinstellungen. Sie können Optionen zur Anzeige der Korrelationsstärke in der Ausgabe angeben.

Korrelationseinstellungen

Bei IBM SPSS Modeler können Korrelationen mit deskriptiven Beschriftungen versehen werden, um so wichtige Beziehungen hervorzuheben. Die **Korrelation** bezeichnet die Stärke der Beziehung zwischen zwei stetigen Feldern (numerischer Bereich). Zulässige Werte liegen im Bereich zwischen -1,0 und 1,0. Werte nahe +1,0 weisen auf eine starke positive Assoziation hin; dies bedeutet, dass hohe Werte in einem Feld mit hohen Werten im zweiten Feld verknüpft sind und entsprechend niedrige Werte mit niedrigen Werten. Werte nahe -1.0 weisen auf eine starke negative Assoziation hin. Dies bedeutet, dass hohe Werte in einem Feld mit niedrigen Werten im zweiten Feld verknüpft sind und umgekehrt. Werte nahe 0,0 weisen auf eine schwache Assoziation hin. Dies bedeutet, dass die Werte der beiden Felder relativ unabhängig voneinander sind.

Über das Dialogfeld **Korrelationseinstellungen** können Sie die Anzeige der Korrelationsbeschriftungen festlegen, die Schwellenwerte ändern, die die Kategorien definieren, und die für die einzelnen Bereiche

verwendeten Beschriftungen ändern. Die Charakterisierung der Korrelationswerte ist stark abhängig von der Problemdomäne. Aus diesem Grund sollten Sie die Bereiche und Beschriftungen an die jeweils gegebene Situation anpassen.

Korrelationsstärkebeschriftungen in Ausgabe anzeigen. Diese Option ist standardmäßig aktiviert. Wenn die deskriptiven Beschriftungen nicht in die Ausgabe aufgenommen werden sollen, inaktivieren Sie diese Option.

Korrelationsstärke. Es gibt zwei Optionen zur Definition und Beschriftung der Korrelationsstärke:

- **Korrelationsstärke nach Wichtigkeit definieren (1-p).** Gibt die Korrelationen auf der Grundlage der Wichtigkeit an, definiert als 1 minus Signifikanz oder 1 minus die Wahrscheinlichkeit, dass sich die Differenz bei den Mittelwerten durch den Zufall allein erklären lässt. Je näher dieser Wert bei 1 liegt, desto größer ist die Wahrscheinlichkeit, dass die beiden Felder *nicht* unabhängig sind und dass sie in Beziehung zueinander stehen. Die Angabe der Korrelationen auf der Grundlage der Wichtigkeit wird normalerweise gegenüber dem absoluten Wert empfohlen, da hierbei die Variabilität in den Daten berücksichtigt wird. So kann es beispielsweise sein, das ein Koeffizient von 0,6 in einem Dataset extrem signifikant ist, in einem anderen jedoch ganz und gar nicht. Standardmäßig werden Wichtigkeitswerte zwischen 0,0 und 0,9 als *Schwach* beschriftet, Wichtigkeitswerte zwischen 0,9 und 0,95 als *Mittel* und Wichtigkeitswerte zwischen 0,95 und 1,0 entsprechend als *Stark*.
- **Korrelationsstärke nach absolutem Wert definieren.** Beschriftet Korrelationen auf der Grundlage des absoluten Werts des Korrelationskoeffizienten nach Pearson, der zwischen -1 und 1 liegt, wie oben beschrieben. Je näher der absolute Wert dieses Maßes bei 1 liegt, desto stärker ist die Korrelation. Standardmäßig werden Korrelationen zwischen 0,0 und 0,3333 (absoluter Wert) als *Schwach* beschriftet, Korrelationen zwischen 0,3333 und 0,6666 als *Mittel* und Korrelationen zwischen 0,6666 und 1,0 entsprechend als *Stark*. Beachten Sie jedoch, dass sich die Signifikanz eines Werts schlecht über mehrere Datasets hinweg verallgemeinern lässt; aus diesem Grund wird in den meisten Fällen empfohlen, Korrelationen auf der Grundlage der Wahrscheinlichkeit und nicht auf der Grundlage ihres absoluten Werts zu definieren.

Statistikausgabebrowser

Der Ausgabebrowser des Statistikknotens enthält die Ergebnisse der statistischen Analyse und ermöglicht die Ausführung verschiedener Funktionen, z. B. Felder auswählen, neue Knoten auf der Grundlage der Auswahl erzeugen sowie die Ergebnisse speichern und drucken. Das Menü "Datei" enthält die üblichen Befehle zum Speichern, Exportieren und Drucken, das Menü "Bearbeiten" die üblichen Bearbeitungsfunktionen. Weitere Informationen finden Sie im Thema „Anzeigen der Ausgabe” auf Seite 283.

Beim Öffnen des Statistikausgabebrowsers werden die Ergebnisse erweitert. Um die Ergebnisse nach der Betrachtung wieder auszublenden, können Sie mit dem Erweiterungssteuerelement links neben dem gewünschten Element die Ergebnisse reduzieren. Alternativ können Sie mit der Schaltfläche **Alles ausblenden** alle Ergebnisse ausblenden. Um die für Sie relevanten Ergebnisse nach dem Reduzieren wieder anzeigen zu lassen, erweitern Sie die gewünschten Ergebnisse mithilfe des Erweiterungssteuerelements auf der linken Seite oder klicken Sie auf die Schaltfläche **Alles anzeigen**, um alle Ergebnisse anzuzeigen.

Die Ausgabe umfasst je einen Abschnitt für die einzelnen Felder *Prüfen* mit einer Tabelle der angeforderten Statistiken.

- **Häufigkeiten.** Anzahl der Datensätze mit gültigen Werten für das Feld.
- **Mittelwert.** Durchschnittlicher Wert (Mittelwert) für das Feld über alle Datensätze hinweg.
- **Summe.** Summe der Werte für das Feld über alle Datensätze hinweg.
- **Min.** Mindestwert für das Feld.
- **Max.** Höchstwert für das Feld.
- **Bereich.** Differenz zwischen Mindest- und Höchstwert.

- **Varianz.** Maß für die Schwankungen der Werte eines Felds. Dieser Wert wird wie folgt berechnet: Zunächst wird die Differenz zwischen jedem Wert und dem Gesamtmittelwert ermittelt. Diese Differenz wird quadriert, über alle Werte summiert und dann durch die Anzahl der Datensätze dividiert.
- **Standardabweichung.** Ein weiteres Maß für die Schwankungen der Werte eines Felds, berechnet als die Quadratwurzel der Varianz.
- **Standardfehler des Mittelwerts.** Maß für die Unsicherheit bei der Schätzung des Mittelwerts für das Feld, falls der Mittelwert vermutungsgemäß für neue Daten gilt.
- **Median.** "Mittlerer" Wert für das Feld, also der Wert, der die obere Hälfte der Daten von der unteren Hälfte trennt (auf der Grundlage der Werte des Felds).
- **Modalwert.** Am häufigsten auftretender einzelner Wert in den Daten.

Korrelationen. Wenn Sie Korrelationsfelder angegeben haben, enthält die Ausgabe außerdem einen Abschnitt, in dem die Korrelation nach Pearson zwischen dem Prüffeld und jedem Korrelationsfeld aufgeführt wird und auch die optionalen deskriptiven Beschriftungen für die Korrelationswerte genannt werden. Weitere Informationen finden Sie im Thema „Korrelationseinstellungen” auf Seite 305.

Menü "Generieren". Das Menü "Generieren" enthält Funktionen zum Erzeugen von Knoten.

- **Filter.** Erzeugt einen Filterknoten, mit dem die Felder herausgefiltert werden, für die keine oder nur eine schwache Korrelation mit anderen Feldern besteht.

Erzeugen eines Filterknotens aus den Statistiken heraus

Mit dem aus einem Statistikausgabebrowser heraus erzeugten Filterknoten werden die Felder auf der Grundlage ihrer Korrelation mit anderen Feldern gefiltert. Hierzu werden die Korrelationen nach dem absoluten Wert sortiert. Anschließend werden die größten Korrelationen ermittelt (gemäß dem im Dialogfeld "Filter aus Statistiken generieren" festgelegten Kriterium); dann wird ein Filter erstellt, der alle Felder passieren lässt, die in einer dieser großen Korrelationen auftreten.

Modalwert. Legen Sie fest, auf welche Weise die Korrelationen ausgewählt werden sollen. Mit der Option **Einschließen** werden alle Felder beibehalten, die in den angegebenen Korrelationen auftreten. Mit **Ausschließen** werden die Felder gefiltert.

Felder ein-/ausschließen in. Definieren Sie das Kriterium zum Auswählen der Korrelationen.

- **Obere Anzahl von Korrelationen.** Die angegebene Anzahl an Korrelationen wird ausgewählt; Felder, die in einer dieser Korrelationen auftreten, werden dabei eingeschlossen bzw. ausgeschlossen.
- **Oberer Prozentsatz von Korrelationen (%).** Der angegebene Prozentsatz (*n* %) an Korrelationen wird ausgewählt; Felder, die in einer dieser Korrelationen auftreten, werden dabei eingeschlossen bzw. ausgeschlossen.
- **Korrelationen größer als.** Hiermit werden Korrelationen ausgewählt, deren absoluter Wert größer ist als der angegebene Schwellenwert.

Mittelwertknoten

Der Mittelwertknoten vergleicht die Mittelwerte zwischen unabhängigen Gruppen oder zwischen Paaren von in Bezug stehenden Feldern, um zu testen, ob ein signifikanter Unterschied vorliegt. So können Sie beispielsweise die Einnahmen vor und nach der Durchführung einer Werbeaktion vergleichen oder die Einnahmen, die von Kunden stammen, die keine Werbezettel erhielten, mit den Einnahmen von Kunden vergleichen, die von der Werbeaktion erreicht wurden.

Sie können Mittelwerte auf zwei verschiedene Weisen vergleichen, je nach den verwendeten Daten:

- **Zwischen Gruppen innerhalb eines Felds.** Um unabhängige Gruppen vergleichen zu können, wählen Sie ein Testfeld und ein Gruppierungsfeld aus. Sie können beispielsweise eine Stichprobe von "Holdout"-Kunden beim Versenden von Werbesendungen ausschließen und die durchschnittlichen Einnahmen durch die Standhalte-Gruppe mit den anderen Kunden vergleichen. In diesem Fall geben Sie ein einzelnes Testfeld an, das die Einnahmen für jeden Kunden anzeigt, sowie ein Flagfeld bzw. nomi-

nales Feld, das angibt, ob der jeweilige Kunde das Angebot erhalten hat. Die Stichproben sind unabhängig in dem Sinn, dass jeder Datensatz entweder der einen oder anderen Gruppe zugewiesen wird und es nicht möglich ist, ein bestimmtes Mitglied einer Gruppe einem bestimmten Mitglied einer anderen Gruppe zuzuweisen. Sie können auch ein nominales Feld mit mehr als zwei Werten angeben, um die Mittelwerte für mehrere Gruppen zu vergleichen. Bei der Ausführung berechnet der Knoten einen einfachen ANOVA-Test für die ausgewählten Felder. In Fällen, in denen es nur zwei Feldgruppen gibt, stimmen die Ergebnisse der einfachen ANOVA im Wesentlichen mit denen eines *t*-Tests mit unabhängigen Stichproben überein. Weitere Informationen finden Sie im Thema „Vergleich der Mittelwerte für unabhängige Gruppen".

- **Zwischen Feldpaaren.** Beim Vergleich der Mittelwerte für verwandte Felder müssen Gruppenpaare gebildet werden, um aussagekräftige Ergebnisse zu erhalten. Sie könnten beispielsweise den Mittelwert der Einnahmen aus derselben Gruppe von Kunden vor und nach der Durchführung einer Werbeaktion vergleichen oder die Nutzungsquoten für einen Service zwischen Paaren aus Ehemann und Ehefrau vergleichen, um zu sehen, ob Unterschiede vorliegen. Jeder Datensatz enthält zwei verschiedene, jedoch miteinander verwandte Maße, bei denen ein aussagekräftiger Vergleich möglich ist. Bei der Ausführung berechnet der Knoten einen *t*-Test bei Stichproben mit paarigen Werten für jedes ausgewählte Feldpaar. Weitere Informationen finden Sie im Thema „Vergleich der Mittelwerte zwischen paarigen Feldern".

Vergleich der Mittelwerte für unabhängige Gruppen

Wählen Sie **Zwischen Gruppen innerhalb eines Felds** im Mittelwertknoten, um die Mittelwerte für zwei oder mehrere unabhängige Gruppen miteinander zu vergleichen.

Gruppierungsfeld. Wählen Sie ein numerisches Flagfeld oder ein nominales Feld mit zwei oder mehr verschiedenen Werten aus, das Datensätze in die Gruppen einteilt, die verglichen werden sollen, beispielsweise in die Gruppe der Personen, die ein Angebot erhalten haben, und die Gruppe von Personen, bei denen dies nicht der Fall ist. Unabhängig von der Anzahl der Testfelder kann nur ein Gruppierungsfeld ausgewählt werden.

Testfelder. Wählen Sie mindestens ein numerisches Feld aus, das die zu testenden Maße enthält. Für jedes ausgewählte Feld wird ein separater Test ausgeführt. Sie können beispielsweise die Auswirkungen einer bestimmten Werbeaktion auf Nutzung, Ertrag und Abwanderung testen.

Vergleich der Mittelwerte zwischen paarigen Feldern

Wählen Sie im Mittelwertknoten die Option **Zwischen Feldpaaren**, um die Mittelwerte zwischen unterschiedlichen Feldern zu vergleichen. Die Felder müssen in einem bestimmten Bezug zueinander stehen, damit die Ergebnisse aussagekräftig sind, beispielsweise die Einnahmen vor und nach einer Werbeaktion. Es können auch mehrere Feldpaare ausgewählt werden.

Feld eins. Wählen Sie ein numerisches Feld aus, das das erste Maß enthält, das verglichen werden soll. In einer Vorher-Nachher-Studie wäre dies das Feld "Vorher".

Feld zwei. Wählen Sie das zweite Feld für den Vergleich aus.

Hinzufügen. Fügt das ausgewählte Paar zur Liste der Testfeldpaare hinzu.

Wiederholen Sie die Feldauswahl nach Bedarf, um mehrere Paare zur Liste hinzuzufügen.

Korrelationseinstellungen. Ermöglicht die Angabe von Optionen zur Beschriftung der Korrelationsstärke. Weitere Informationen finden Sie im Thema „Korrelationseinstellungen" auf Seite 305.

Mittelwertknoten - Optionen

Auf der Registerkarte "Optionen" können Sie Schwellenwerte für *p*-Werte festlegen, die verwendet werden, um Ergebnisse als "Bedeutsam", "Marginal" oder "Unbedeutend" zu beschriften. Außerdem können

Sie die Beschriftung für jede Einstufung bearbeiten. Die Wichtigkeit wird auf einer Prozentskala gemessen und lässt sich grob wie folgt definieren: 1 minus die Wahrscheinlichkeit, ein Ergebnis (beispielsweise die Differenz der Mittelwerte zwischen zwei Feldern) zu erhalten, das nur allein Zufall mindestens so extrem ist wie das beobachtete Ergebnis. Ein *p*-Wert größer als 0,95 beispielsweise zeigt an, dass eine Wahrscheinlichkeit von weniger als 5 % besteht, dass sich das Ergebnis allein durch Zufall erklären lässt.

Wichtigkeitsbeschriftungen. Sie können die Beschriftungen für die einzelnen Feldpaare bzw. -gruppen in der Ausgabe bearbeiten. Die Standardbeschriftungen lauten *Bedeutsam*, *Marginal* und *Unbedeutend*.

Trennwerte. Geben den Schwellenwert für jeden Rang an. Üblicherweise werden *p*-Werte über 0,95 als bedeutsam eingestuft, Werte unter 0,9 als unbedeutend. Diese Schwellenwerte lassen sich jedoch nach Bedarf anpassen.

Hinweis: Wichtigkeitsmaße sind in einer Reihe von Knoten verfügbar. Die speziellen Berechnungen hängen vom Knoten und vom verwendeten Ziel- und Eingabefeldtyp ab, die Werte können jedoch weiterhin verglichen werden, da sie auf einer Prozentskala gemessen werden.

Mittelwertknoten - Ausgabebrowser

Der Browser für die Mittelwertausgabe zeigt Daten als Kreuztabellen an und ermöglicht die Ausführung von Standardoperationen wie Auswählen und Kopieren der Tabelle Zeile für Zeile, Sortieren nach einer beliebigen Spalte sowie Speichern und Drucken der Tabelle. Weitere Informationen finden Sie im Thema „Anzeigen der Ausgabe" auf Seite 283.

Die spezifischen Informationen in der Tabelle hängen vom Vergleichstyp (Gruppen innerhalb eines Felds oder gesonderte Felder) ab.

Sortieren nach. Ermöglicht die Sortierung der Ausgabe nach einer bestimmten Spalte. Klicken Sie auf den nach oben bzw. nach unten weisenden Pfeil, um die Sortierrichtung zu ändern. Alternativ können Sie auf die Überschrift einer Spalte klicken, um eine Sortierung nach dieser Spalte vorzunehmen. (Wenn Sie die Sortierrichtung innerhalb der Spalte ändern möchten, klicken Sie noch einmal.)

Ansicht. Sie haben die Wahl zwischen **Einfach** und **Erweitert**, um den Detaillierungsgrad der Anzeige zu steuern. Die erweiterte Ansicht enthält alle Informationen der einfachen Ansicht sowie weitere Einzelheiten.

Mittelwertausgabe zum Vergleich von Gruppen innerhalb eines Felds

Beim Vergleich von Gruppen innerhalb eines Felds wird der Name des Gruppierungsfelds oberhalb der Ausgabetabelle angezeigt; außerdem werden Mittelwerte und verwandte Statistiken separat für die einzelnen Gruppen gemeldet. Die Tabelle enthält eine gesonderte Zeile für jedes Testfeld.

Folgende Spalten werden angezeigt:

- **Feld.** Hier werden die Namen der ausgewählten Testfelder angegeben.
- **Mittelwert nach Gruppe.** Zeigt den Mittelwert für die einzelnen Kategorien des Gruppierungsfelds an. Sie könnten beispielsweise die Personen, die ein Sonderangebot (*New Promotion*) erhalten haben, mit denen vergleichen, die keines erhalten haben (*Standard*). In der erweiterten Ansicht werden außerdem Standardabweichung, Standardfehler und Anzahl angezeigt.
- **Wichtigkeit.** Zeigt Wert und Beschriftung für die Wichtigkeit an. Weitere Informationen finden Sie im Thema „Mittelwertknoten - Optionen" auf Seite 308.

Erweiterte Ausgabe

In der erweiterten Ansicht werden folgende zusätzlichen Spalten angezeigt.

- **F-Test.** Dieser Test beruht auf dem Quotienten aus der Varianz zwischen den Gruppen und der Varianz innerhalb der einzelnen Gruppen. Wenn die Mittelwerte für alle Gruppen gleich sind, ist zu erwarten, dass das *F*-Verhältnis nahe bei 1 liegt, da beides Schätzungen derselben Populationsvarianz sind. Je

größer dieser Quotient, desto größer ist die Variation zwischen den Gruppen und desto größer ist die Wahrscheinlichkeit, dass eine signifikante Differenz vorliegt.

- **df.** Zeigt die Freiheitsgrade (Degrees of Freedom) an.

Mittelwertausgabe zum Vergleich von Feldpaaren

Beim Vergleich zwischen verschiedenen Feldern enthält die Ausgabetabelle eine Zeile für jedes ausgewählte Feldpaar.

- **Feld eins/zwei.** Zeigt den Namen des ersten und zweiten Felds in jedem Paar an. In der erweiterten Ansicht werden außerdem Standardabweichung, Standardfehler und Anzahl angezeigt.
- **Mittelwert eins/zwei.** Zeigt den Mittelwert für das jeweilige Feld an.
- **Korrelation.** Misst die Stärke der Beziehung zwischen zwei stetigen Feldern (numerischer Bereich). Werte in der Nähe von +1,0 deuten auf eine starke positive Assoziation hin und Werte in der Nähe von -1,0 deuten auf eine starke negative Assoziation hin. Weitere Informationen finden Sie im Thema „Korrelationseinstellungen" auf Seite 305.
- **Mittelwertdifferenz.** Zeigt die Differenz zwischen den beiden Feldmittelwerten an.
- **Wichtigkeit.** Zeigt Wert und Beschriftung für die Wichtigkeit an. Weitere Informationen finden Sie im Thema „Mittelwertknoten - Optionen" auf Seite 308.

Erweiterte Ausgabe

Bei der erweiterten Ausgabe sind folgende Spalten hinzugefügt:

95%-Konfidenzintervall. Unter- und Obergrenze des Bereichs, in dem der wahre Mittelwert statistisch gesehen in 95 % aller möglichen Stichproben dieser Größe aus dieser Grundgesamtheit fällt.

T-Test. Die *t*-Statistik wird berechnet, indem die mittlere Differenz durch ihren Standardfehler dividiert wird. Je größer der absolute Wert dieser Statistik, desto größer ist die Wahrscheinlichkeit, dass die Mittelwerte nicht identisch sind.

df. Zeigt die Freiheitsgrade (Degrees of Freedom) für die Statistik an.

Berichtknoten

Mit dem Berichtknoten erstellen Sie formatierte Berichte, die sowohl festen Text als auch Daten und andere aus den Daten abgeleitete Ausdrücke enthält. Das Format des Berichts wird mithilfe von Textvorlagen festgelegt, mit denen der feste Text und die Datenausgabekonstruktionen definiert werden. Sie können eine benutzerdefinierte Textformatierung angeben; hierzu stehen HTML-Tags in der Vorlage sowie Optionen auf der Registerkarte "Ausgabe" zur Verfügung. Datenwerte und andere bedingte Ausgaben werden mithilfe von CLEM-Ausdrücken in der Vorlage in den Bericht aufgenommen.

Alternativen zum Berichtknoten

Der Berichtknoten wird normalerweise verwendet, um ausgegebene Datensätze oder Fälle aus einem Stream aufzulisten, beispielsweise alle Datensätze, die eine bestimmte Bedingung erfüllen. In dieser Hinsicht kann er als weniger strukturierte Alternative zum Tabellenknoten betrachtet werden.

- Wenn Sie einen Bericht wünschen, der Feldinformationen oder andere Elemente auflistet, die im Stream definiert wurden, und nicht die Daten selbst (beispielsweise die in einem Typknoten angegebenen Felddefinitionen), kann stattdessen ein Script verwendet werden.
- Um einen Bericht zu generieren, der mehrere Ausgabeobjekte enthält (z. B. eine Sammlung von Modellen, Tabellen und Diagrammen, die von einem oder mehreren Streams generiert wurden) und in mehreren Formaten (z. B. Textformat, HTML und Microsoft Word/Office) ausgegeben werden kann, können Sie ein IBM SPSS Modeler-Projekt verwenden.
- Um eine Liste von Feldnamen ohne Verwendung von Scripts zu erstellen, können Sie einen Tabellenknoten verwenden, dem ein Stichprobenknoten vorangeht, der alle Datensätze verwirft. Auf diese Wei-

se wird eine Tabelle ohne Zeilen erstellt, die beim Export transponiert werden kann, um eine Liste von Feldnamen in einer einzelnen Spalte zu erzeugen. (Wählen Sie dazu im Tabellenknoten auf der Registerkarte "Ausgabe" die Option **Daten transponieren**.)

Registerkarte "Vorlage" beim Berichtknoten

Erstellen einer Vorlage. Um den Inhalt des Berichts zu definieren, erstellen Sie eine Vorlage auf der Registerkarte "Vorlage" im Berichtknoten. Die Vorlage besteht aus Textzeilen, die jeweils Angaben zum Inhalt des Berichts enthalten, sowie aus einigen Zeilen mit Sondertags, aus denen der Bereich der Inhaltszeilen hervorgeht. In jeder Inhaltszeile werden zunächst CLEM-Ausdrücke in eckigen Klammern ([]) ausgewertet, bevor die betreffende Zeile an den Bericht gesendet wird. Für die Zeilen in der Vorlage stehen jeweils drei Bereiche zur Auswahl:

Fest. Zeilen, die nicht anderweitig gekennzeichnet sind, werden als fest betrachtet. Feste Zeilen werden nur einmal in den Bericht kopiert, sobald alle in diesen Zeilen enthaltenen Ausdrücke ausgewertet wurden. Beispiel: Mit der Zeile

```
Dies ist mein Bericht, gedruckt am [@TODAY]
```

wird eine einzelne Zeile in den Bericht kopiert, die den angegebenen Text und das aktuelle Datum enthält.

Global (Alles iterieren). Die Zeilen zwischen den Sondertags `#ALL` und `#` werden je einmal für jeden Datensatz mit Eingabedaten in den Bericht kopiert. Die CLEM-Ausdrücke (in Klammern) werden auf der Grundlage des jeweils aktuellen Datensatzes für jede Ausgabezeile ausgewertet. Beispiel: Mit den Zeilen

```
#ALL
Für Datensatz [@INDEX] lautet der Wert von ALTER [ALTER]
#
```

wird je eine Zeile für jeden Datensatz eingefügt, aus der die Nummer des Datensatzes und das Alter hervorgeht.

So generieren Sie eine Liste aller Datensätze:

```
#ALL
[Alter] [Geschlecht] [Cholesterin]  [BD]
#
```

Bedingt (Iterieren, wenn). Die Zeilen zwischen den Sondertags `#WHERE` <*Bedingung*> und `#` werden je einmal für jeden Datensatz, bei dem die angegebene Bedingung wahr ist, in den Bericht kopiert. Die Bedingung besteht aus einem CLEM-Ausdruck. (Die eckigen Klammern bei der `WHERE`-Bedingung sind optional.) Beispiel:

```
#WHERE [SEX = 'M']
Der Mann in Datensatz Nr. [@INDEX] ist [AGE] Jahre alt.
#
```

Mit den obigen Zeilen wird je eine Zeile für jeden Datensatz in die Datei geschrieben, bei dem der Wert *M* für das Geschlecht vorliegt. Der vollständige Datensatz enthält die festen, globalen und bedingten Zeilen, die durch Anwendung der Vorlage auf die Eingabedaten definiert wurden.

Auf der Registerkarte "Ausgabe" können Sie Optionen für das Anzeigen und Speichern der Ergebnisse festlegen, die verschiedenen Arten von Ausgabeknoten gemeinsam sind. Weitere Informationen finden Sie im Thema „Registerkarte "Ausgabe" beim Ausgabeknoten” auf Seite 286.

Ausgabe von Daten im HTML- oder XML-Format

Sie können HTML- oder XML-Tags direkt in die Vorlage einfügen, um Berichte in einem dieser Formate zu schreiben. Die folgende Vorlage beispielsweise führt zu einer HTML-Tabelle.

```
Dieser Bericht wurde in HTML geschrieben.
Dabei werden nur Fälle eingeschlossen, bei denen die Variable "Age" (Alter) größer 60 ist.

<HTML>
  <TABLE border="2">
   <TR>
     <TD>Alter</TD>
     <TD>BD</TD>
     <TD>Cholesterin</TD>
     <TD>Medikament</TD>
   </TR>

  #WHERE Alter > 60
    <TR>
       <TD>[Alter]</TD>
       <TD>[BD]</TD>
       <TD>[Cholesterin]</TD>
       <TD>[Medikament]</TD>
     </TR>
  #
  </TABLE>
</HTML>
```

Browser für Berichtknotenausgabe

Der Berichtsbrowser zeigt den Inhalt des erzeugten Berichts. Das Menü "Datei" enthält die üblichen Befehle zum Speichern, Exportieren und Drucken, das Menü "Bearbeiten" die üblichen Bearbeitungsfunktionen. Weitere Informationen finden Sie im Thema „Anzeigen der Ausgabe” auf Seite 283.

Globalwerteknoten

Mit dem Globalwerteknoten werden die Daten gescannt und Übersichtswerte berechnet, die in CLEM-Ausdrücken herangezogen werden können. Mit einem Globalwerteknoten können Sie beispielsweise die Statistiken für das Feld *age* (Alter) berechnen und dann den Gesamtmittelwert für *age* (Alter) in CLEM-Ausdrücken verwenden. Fügen Sie hierzu die Funktion `@GLOBAL_MEAN(age)` ein.

Registerkarte "Einstellungen" beim Globalwerteknoten

Zu erstellende Globalwerte. Wählen Sie das oder die Felder aus, für die Globalwerte verfügbar sein sollen. Sie können mehrere Felder auswählen. Geben Sie die zu berechnenden Statistiken für jedes Feld an. Wählen Sie hierzu die gewünschten Statistiken in den Spalten neben dem Feldnamen aus.

- **Mittelwert.** Durchschnittlicher Wert (Mittelwert) für das Feld über alle Datensätze hinweg.
- **Summe.** Summe der Werte für das Feld über alle Datensätze hinweg.
- **Min.** Mindestwert für das Feld.
- **Max.** Höchstwert für das Feld.
- **Std.Abw.** Die Standardabweichung, ein Maß für die Variabilität der Werte eines Felds, berechnet als die Quadratwurzel der Varianz.

Standardoperation(en). Die hier ausgewählten Optionen werden verwendet, wenn Sie weitere Felder zur obigen Liste der Globalwerte hinzufügen. Um die Standardgruppe der Statistiken zu ändern, wählen Sie die gewünschten Statistiken aus oder heben Sie die Auswahl bestimmter Statistiken wieder auf. Mit der Schaltfläche **Anwenden** können Sie zudem die Standardoperationen auf alle Felder in der Liste gleichzeitig anwenden.

Alle Globalwerte vor Ausführung löschen. Vor der Berechnung neuer Globalwerte werden alle vorhandenen Globalwerte gelöscht. Ist diese Option nicht ausgewählt, ersetzen die neuen berechneten Werte zwar die bisherigen Werte, die nicht neu berechneten Globalwerte bleiben jedoch weiterhin verfügbar.

Vorschau der Globalwerte anzeigen, die nach der Ausführung erstellt wurden. Mit dieser Option wird nach der Ausführung das Dialogfeld "Streameigenschaften" mit der Registerkarte "Globalwerte" geöffnet; hier werden die berechneten Globalwerte angezeigt.

Simulationsanpassungsknoten

Der Simulationsanpassungsknoten passt ein Set möglicher statistischer Verteilungen an die einzelnen Felder in den Daten an. Die Anpassung der einzelnen Verteilungen an ein Feld wird mithilfe eines Kriteriums für die Anpassungsgüte bewertet. Wenn ein Simulationsanpassungsknoten ausgeführt wird, wird ein Simulationsgenerierungsknoten erstellt (oder ein vorhandener Knoten aktualisiert). Jedes Feld wird seiner am besten angepassten Verteilung zugewiesen. Mit dem Simulationsgenerierungsknoten können dann simulierte Daten für jedes Feld generiert werden.

Der Simulationsanpassungsknoten ist zwar ein Endknoten, jedoch fügt er weder der generierten Modellpalette ein Modell, noch der Registerkarte "Ausgaben" eine Ausgabe oder ein Diagramm hinzu und er exportiert auch keine Daten.

Anmerkung: Wenn die historischen Daten dünn besetzt sind (d. h., wenn viele Werte fehlen), kann es für die Anpassungskomponente schwierig sein, genügend gültige Werte für die Anpassung von Verteilungen an die Daten zu finden. Um die Verteilungen an die Daten anzupassen, benötigt die Anpassungskomponente 2000 gültige Werte. In Fällen, in denen die Daten dünn besetzt sind, sollten Sie vor der Anpassung entweder die dünn besetzten Felder entfernen, wenn sie nicht erforderlich sind, oder die fehlenden Werte imputieren. Mithilfe der Optionen auf der Registerkarte **Qualität** des Data Audit-Knotens können Sie die Anzahl vollständiger Datensätze anzeigen, ermitteln, welche Felder dünn besetzt sind und eine Imputationsmethode auswählen. Wenn nicht genügend Datensätze für die Verteilungsanpassung vorhanden sind, können Sie die Anzahl der Datensätze mithilfe eines Balancierungsknotens erhöhen.

Verwenden eines Simulationsanpassungsknotens für die automatische Erstellung eines Simulationsgenerierungsknotens

Wenn der Simulationsanpassungsknoten zum ersten Mal ausgeführt wird, wird ein Simulationsgenerierungsknoten mit einem Aktualisierungslink zum Simulationsanpassungsknoten erstellt. Wenn der Simulationsanpassungsknoten erneut ausgeführt wird, wird nur ein neuer Simulationsgenerierungsknoten erstellt, wenn der Aktualisierungslink entfernt wurde. Mit einem Simulationsanpassungsknoten kann auch ein verbundener Simulationsgenerierungsknoten aktualisiert werden. Das Ergebnis hängt davon ab, ob in beiden Knoten dieselben Felder vorhanden sind und ob die Felder im Simulationsgenerierungsknoten entsperrt sind. Weitere Informationen finden Sie im Thema „Simulationsgenerierungsknoten" auf Seite 45.

Ein Simulationsanpassungsknoten kann nur einen Aktualisierungslink zu einem Simulationsgenerierungsknoten aufweisen. Führen Sie die folgenden Schritte aus, um einen Aktualisierungslink zu einem Simulationsgenerierungsknoten zu definieren:

1. Klicken Sie mit der rechten Maustaste auf den Simulationsanpassungsknoten.
2. Wählen Sie im Menü **Aktualisierungslink definieren** aus.
3. Klicken Sie auf den Simulationsgenerierungsknoten, zu dem Sie einen Aktualisierungslink definieren möchten.

Um einen Aktualisierungslink zwischen einem Simulationsanpassungsknoten und einem Simulationsgenerierungsknoten zu entfernen, klicken Sie mit der rechten Maustaste auf den Aktualisierungslink und wählen Sie **Verknüpfung entfernen** aus.

Verteilungsanpassung

Eine statistische Verteilung ist die theoretische Häufigkeit für das Vorkommen von Werten, die eine Variable annehmen kann. Im Simulationsanpassungsknoten wird ein Set theoretischer statistischer Verteilungen mit jedem Datenfeld verglichen. Die Verteilungen, die für die Anpassung zur Verfügung stehen, werden im Thema „Verteilungen" auf Seite 55 beschrieben. Die Parameter der theoretischen Verteilung werden so

angepasst, dass sich entsprechend einer Messung der Anpassungsgüte (Anderson-Darling-Kriterium oder Kolmogorov-Smirnov-Kriterium) die beste Anpassung an die Daten ergibt. Die Ergebnisse der Verteilungsanpassung durch den Simulationsanpassungsknoten zeigen, welche Verteilungen angepasst wurden, die besten Schätzungen der Parameter für die einzelnen Verteilungen und wie gut jede Verteilung an die Daten angepasst ist. Während der Verteilungsanpassung werden auch Korrelationen zwischen Feldern mit numerischen Speichertypen und Kontingenzen zwischen Feldern mit einer kategorialen Verteilung berechnet. Die Ergebnisse der Verteilungsanpassung werden für die Erstellung eines Simulationsgenerierungsknotens verwendet.

Bevor Verteilungen an Ihre Daten angepasst werden, werden die ersten 1000 Datensätze auf fehlende Werte überprüft. Wenn zu viele Werte fehlen, ist keine Verteilungsanpassung möglich. Ist dies der Fall, müssen Sie entscheiden, ob eine der folgenden Optionen angemessen ist:

- Verwenden Sie einen vorgeordneten Knoten, um Datensätze mit fehlenden Werten zu entfernen.
- Verwenden Sie einen vorgeordneten Knoten, um Werte für fehlende Werte zu imputieren.

Die Verteilungsanpassung schließt benutzerdefiniert fehlende Werte nicht aus. Wenn Ihre Daten 'benutzerdefiniert fehlende' Werte aufweisen und diese Werte aus der Verteilungsanpassung ausgeschlossen werden sollen, sollten Sie diese Werte als 'systemdefiniert fehlend' festlegen.

Die Rolle eines Felds wird nicht berücksichtigt, wenn die Verteilungen angepasst werden. Felder mit der Rolle **Ziel** werden beispielsweise genauso wie Felder mit den Rollen **Eingabe**, **Keine**, **Beide**, **Partition**, **Aufteilen**, **Häufigkeit** und **ID** behandelt.

Felder werden während der Verteilungsanpassung entsprechend ihrem Speichertyp und Messniveau unterschiedlich behandelt. Die Behandlung von Feldern während der Verteilungsanpassung wird in der folgenden Tabelle beschrieben.

Tabelle 51. Verteilungsanpassung entsprechend Speichertyp und Messniveau von Feldern

<table>
<tr><th rowspan="2">Speichertyp</th><th colspan="6">Messniveau</th></tr>
<tr><th>Stetig</th><th>Kategorial</th><th>Flag</th><th>Nominal</th><th>Ordinal</th><th>Ohne Typ</th></tr>
<tr><td>Zeichenfolge</td><td>Unmöglich</td><td colspan="3">Kategoriale Verteilungen, Dice-Verteilungen und feste Verteilungen werden angepasst.</td><td rowspan="7">Feld wird ignoriert und nicht an den Simulationsgenerierungsknoten übergeben.</td></tr>
<tr><td>Ganzzahl</td><td rowspan="6">Alle Verteilungen werden angepasst. Korrelationen und Kontingenzen werden berechnet.</td><td colspan="3" rowspan="6">Die kategoriale Verteilung wird angepasst. Korrelationen werden nicht berechnet.</td><td rowspan="6">Binomial-, negative Binomial- und Poisson-Verteilungen werden angepasst und Korrelationen werden berechnet.</td></tr>
<tr><td>Reelle Zahl</td></tr>
<tr><td>Zeit</td></tr>
<tr><td>Datum</td></tr>
<tr><td>Zeitmarke</td></tr>
</table>

Tabelle 51. Verteilungsanpassung entsprechend Speichertyp und Messniveau von Feldern (Forts.)

Speichertyp	Messniveau	
Unbekannt	Entsprechender Speichertyp wird aus den Daten bestimmt.	

Felder mit dem Messniveau "ordinal" werden wie stetige Felder behandelt und werden in die Korrelationstabelle im Simulationsgenerierungsknoten eingefügt. Wenn eine andere Verteilung als eine Binomial-, negative Binomial- oder Poisson-Verteilung an ein ordinales Feld angepasst werden soll, müssen Sie das Messniveau des Felds in "stetig" ändern. Wenn Sie zuvor für jeden Wert eines ordinalen Felds eine Beschriftung definiert haben und dann das Messniveau in "stetig" ändern, gehen die Beschriftungen verloren.

Felder mit Einzelwerten werden während der Verteilungsanpassung nicht anders als Felder mit mehreren Werten behandelt. Felder mit dem Speichertyp "Zeit", "Datum" oder "Zeitmarke" werden als numerische Felder behandelt.

Anpassen von Verteilungen an Aufteilungsfelder

Wenn Ihre Daten ein Aufteilungsfeld enthalten und die Verteilungsanpassung für jede Aufteilung gesondert ausgeführt werden soll, müssen Sie die Daten mithilfe eines vorgeordneten Umstrukturierungsknotens transformieren. Generieren Sie mithilfe des Umstrukturierungsknotens für jeden Wert des Aufteilungsfelds ein neues Feld. Diese umstrukturierten Daten können dann für die Verteilungsanpassung im Simulationsanpassungsknoten verwendet werden.

Simulationsanpassungsknoten - Registerkarte "Einstellungen"

Quellenknotenname. Sie können den Namen des generierten (oder aktualisierten) Simulationsgenerierungsknotens automatisch generieren, indem Sie **Auto** auswählen. Der automatisch generierte Name ist der im Simulationsanpassungsknoten angegebene Name, wenn ein benutzerdefinierter Name angegeben wurde (oder "Simulationsgenerierung", wenn kein benutzerdefinierter Name im Simulationsanpassungsknoten angegeben wurde). Wählen Sie **Benutzerdefiniert** aus, um im Textfeld daneben einen benutzerdefinierten Namen anzugeben. Wenn das Textfeld nicht bearbeitet wird, ist der benutzerdefinierte Standardname "Simulationsgenerierung".

Anpassungsoptionen. Mit diesen Optionen können Sie angeben, wie die Verteilungen an die Felder angepasst werden und wie die Anpassung der Verteilungen bewertet wird.

- **Anzahl der Fälle für Stichproben.** Gibt die Anzahl der Fälle an, die verwendet werden sollen, wenn Verteilungen an die Felder im Dataset angepasst werden. Wählen Sie **Alle Fälle** aus, um Verteilungen an alle Datensätze in den Daten anzupassen. Wenn Ihr Dataset sehr groß ist, sollten Sie in Erwägung ziehen, die Anzahl der für die Verteilungsanpassung verwendeten Fälle zu begrenzen. Wählen Sie **Auf erste n Fälle begrenzen** aus, um nur die ersten n Fälle zu verwenden. Klicken Sie auf die Pfeile, um die Anzahl der Fälle anzugeben, die verwendet werden sollen. Sie können auch einen vorgeordneten Knoten verwenden, um eine Zufallsstichprobe von Datensätzen für die Verteilungsanpassung zu ziehen.
- **Kriterien für Anpassungsgüte (nur stetige Felder).** Wählen Sie für stetige Felder entweder den Anderson-Darling-Test oder den Kolmogorov-Smirnoff-Test für die Anpassungsgüte aus, um Verteilungen bei der Anpassung von Verteilungen an Felder einzustufen. Der Anderson-Darling-Test ist standardmäßig ausgewählt und wird besonders empfohlen, wenn Sie die bestmögliche Anpassung in den Flankenbereichen sicherstellen möchten. Beide Statistiken werden für jede mögliche Verteilung berechnet, es wird jedoch nur die ausgewählte Statistik verwendet, um die Verteilungen zu ordnen und die am besten angepasste Verteilung zu bestimmen.

- **Klassen (nur empirische Verteilung).** Für stetige Felder ist die empirische Verteilung die kumulative Verteilungsfunktion der historischen Daten. Dies ist die Wahrscheinlichkeit für jeden Wert oder Wertebereich und wird direkt aus den Daten abgeleitet. Sie können die Anzahl der Klassen, die für die Berechnung der empirischen Verteilung für stetige Felder verwendet werden, angeben, indem Sie auf die Pfeile klicken. Der Standardwert ist 100 und das Maximum ist 1000.
- **Gewichtungsfeld (optional).** Wenn Ihr Dataset ein Gewichtungsfeld enthält, klicken Sie auf das Symbol für die Feldauswahlfunktion und wählen Sie das Gewichtungsfeld aus der Liste aus. Das Gewichtungsfeld wird dann aus dem Verteilungsanpassungsprozess ausgeschlossen. In der Liste werden alle Felder im Dataset angezeigt, die das Messniveau "stetig" aufweisen. Sie können nur ein Gewichtungsfeld auswählen.

Simulationsevaluierungsknoten

Der Simulationsevaluierungsknoten ist ein Endknoten, der ein angegebenes Feld auswertet, eine Verteilung des Felds bereitstellt und Diagramme von Verteilungen und Korrelationen erstellt. Dieser Knoten wird hauptsächlich zur Auswertung stetiger Felder verwendet. Daher ergänzt er das Evaluierungsdiagramm, das durch einen Evaluierungsknoten generiert wird, und ist für die Auswertung diskreter Felder hilfreich. Ein weiterer Unterschied besteht darin, dass der Simulationsevaluierungsknoten eine einzige Vorhersage in mehreren Iterationen auswertet, während der Evaluierungsknoten mehrere Vorhersagen mit jeweils einer einzigen Iteration auswertet. Iterationen werden generiert, wenn für einen Verteilungsparameter im Simulationsgenerierungsknoten mehrere Werte angegeben sind. Weitere Informationen finden Sie im Thema „Iterationen” auf Seite 55.

Der Simulationsevaluierungsknoten ist für die Verwendung mit Daten bestimmt, die aus dem Simulationsanpassungsknoten und dem Simulationsgenerierungsknoten erhalten wurden. Der Knoten kann jedoch zusammen mit jedem anderen Knoten verwendet werden. Zwischen dem Simulationsgenerierungsknoten und dem Simulationsevaluierungsknoten kann eine beliebige Anzahl an Verarbeitungsschritten eingefügt werden.

Wichtig: Der Simulationsevaluierungsknoten benötigt mindestens 1000 Datensätze mit gültigen Werten für das Zielfeld.

Simulationsevaluierungsknoten - Registerkarte "Einstellungen"

Auf der Registerkarte "Einstellungen" des Simulationsevaluierungsknotens können Sie die Rolle jedes Felds in Ihrem Dataset angeben und die Ausgabe anpassen, die von der Simulation generiert wird.

Element auswählen. Ermöglicht Ihnen, zwischen den drei Ansichten des Simulationsevaluierungsknotens zu wechseln: "Felder", "Dichtefunktionen" und "Ausgaben".

Ansicht "Felder"

Zielfeld. Dies ist ein erforderliches Feld. Klicken Sie auf den Pfeil, um das Zielfeld Ihres Datasets aus der Dropdown-Liste auszuwählen. Das ausgewählte Feld kann entweder ein stetiges, ordinales oder nominales Messniveau aufweisen, es kann jedoch kein Datumsmessniveau oder nicht angegebenes Messniveau enthalten.

Iterationsfeld (optional). Wenn Ihre Daten ein Iterationsfeld aufweisen, das angibt, zu welcher Iteration die einzelnen Datensätze in Ihren Daten gehören, müssen Sie es hier auswählen. Dies bedeutet, dass jede Iteration gesondert ausgewertet wird. Es können nur Felder mit einem stetigen, ordinalen oder nominalen Messniveau ausgewählt werden.

Eingabedaten sind bereits nach Iteration sortiert. Nur aktiviert, wenn im Feld **Iterationsfeld (optional)** ein Iterationsfeld angegeben ist. Wählen Sie diese Option nur aus, wenn Sie sicher sind, dass Ihre Eingabedaten bereits nach dem in **Iterationsfeld (optional)** angegebenen Iterationsfeld sortiert wurden.

Maximale Anzahl der Iterationen für Diagramm. Nur aktiviert, wenn im Feld **Iterationsfeld (optional)** ein Iterationsfeld angegeben ist. Klicken Sie auf die Pfeile, um die Anzahl der Iterationen anzugeben, die dargestellt werden sollen. Durch die Angabe dieser Anzahl wird vermieden, dass versucht wird, zu viele Iterationen in einem einzigen Diagramm darzustellen, was die Interpretation des Diagramms erschweren würde. Die niedrigste Stufe, die für die maximale Anzahl der Iterationen festgelegt werden kann, ist 2, die höchste Stufe ist 50. Die maximale Anzahl der Iterationen für ein Diagramm ist anfänglich auf 10 gesetzt.

Eingabefelder für Korrelations-Tornado. Das Korrelations-Tornado-Diagramm ist ein Balkendiagramm, das die Korrelationskoeffizienten zwischen dem angegebenen Ziel und jeder der angegebenen Eingaben anzeigt. Klicken Sie auf das Symbol für die Feldauswahlfunktion, um die Eingabefelder, die in das Tornado-Diagramm eingeschlossen werden sollen, aus einer Liste der verfügbaren simulierten Eingaben auszuwählen. Es können nur Eingabefelder mit einem stetigen oder ordinalen Messniveau ausgewählt werden. Nominale Eingabefelder, Eingabefelder ohne Typ und Datumseingabefelder sind in der Liste nicht verfügbar und können nicht ausgewählt werden.

Ansicht "Dichtefunktionen"

Mit den Optionen in dieser Ansicht können Sie die Ausgabe für Wahrscheinlichkeitsdichtefunktionen und kumulative Verteilungsfunktionen für stetige Ziele sowie Balkendiagramme mit vorhergesagten Werten für kategoriale Ziele anpassen.

Dichtefunktionen. Dichtefunktionen sind das wichtigste Mittel zur Überprüfung des Ergebnissets aus Ihrer Simulation.

- **Wahrscheinlichkeitsdichtefunktion (PDF).** Wählen Sie diese Option aus, um eine Wahrscheinlichkeitsdichtefunktion für das Zielfeld zu erstellen. Die Wahrscheinlichkeitsdichtefunktion zeigt die Verteilung der Zielwerte an. Anhand der Wahrscheinlichkeitsdichtefunktion können Sie die Wahrscheinlichkeit bestimmen, mit der sich das Ziel in einem bestimmten Bereich befindet. Für kategoriale Ziele (Ziele mit einem nominalen oder ordinalen Messniveau) wird ein Balkendiagramm generiert, das den Prozentsatz der Fälle anzeigt, die in die einzelnen Kategorien des Ziels fallen.
- **Kumulative Verteilungsfunktion (CDF).** Wählen Sie diese Option aus, um eine kumulative Verteilungsfunktion für das Zielfeld zu erstellen. Die kumulative Verteilungsfunktion zeigt die Wahrscheinlichkeit an, mit der der Wert des Ziels kleiner-gleich einem angegebenen Wert ist. Sie ist nur für stetige Ziele verfügbar.

Bezugslinien (fortlaufend). Diese Optionen sind aktiviert, wenn **Wahrscheinlichkeitsdichtefunktion (PDF)** und/oder **Kumulative Verteilungsfunktion (CDF)** ausgewählt ist. Mit diesen Optionen können Sie verschiedene feste vertikale Bezugslinien den Wahrscheinlichkeitsdichtefunktionen und kumulativen Verteilungsfunktionen hinzufügen.

- **Mittelwert.** Wählen Sie diese Option aus, um eine Bezugslinie am Mittelwert des Zielfelds hinzuzufügen.
- **Median.** Wählen Sie diese Option aus, um eine Bezugslinie am Medianwert des Zielfelds hinzuzufügen.
- **Standardabweichungen.** Wählen Sie diese Option aus, um Bezugslinien bei plus und minus einer angegebenen Anzahl an Standardabweichungen vom Mittelwert des Zielfelds hinzuzufügen. Durch die Auswahl dieser Option wird das Feld **Anzahl** daneben aktiviert. Klicken Sie auf die Pfeile, um die Anzahl der Standardabweichungen anzugeben. Die minimale Anzahl an Standardabweichungen ist 1 und das Maximum ist 10. Die Anzahl der Standardabweichungen ist anfänglich auf 3 gesetzt.
- **Perzentile.** Wählen Sie diese Option aus, um Bezugslinien an zwei Perzentilwerten der Verteilung des Zielfelds hinzuzufügen. Durch die Auswahl dieser Option werden die Textfelder **Unten** und **Oben** daneben aktiviert. Wenn Sie beispielsweise in das Textfeld **Oben** den Wert 90 eingeben würden, würde am 90. Perzentil des Ziels eine Bezugslinie hinzugefügt. Dies ist der Wert, unter den 90 % der Beobachtungen fallen. Entsprechend stellt der Wert 10 im Textfeld **Unten** das 10. Perzentil des Ziels dar, also den Wert, unter den 10 % der Beobachtungen fallen.

- **Benutzerdefinierte Bezugslinien.** Wählen Sie diese Option aus, um Bezugslinien an bestimmten Werten auf der horizontalen Achse des Zielfelds hinzuzufügen. Durch die Auswahl dieser Option wird die Tabelle **Werte** daneben aktiviert. Jedes Mal, wenn Sie in die Tabelle **Werte** eine gültige Zahl eingeben, wird unten an die Tabelle eine neue leere Zeile angefügt. Eine *gültige* Zahl ist eine Zahl innerhalb des Wertebereichs des Zielfelds.

Anmerkung: Wenn mehrere Dichtefunktionen oder Verteilungsfunktionen (aus mehreren Iterationen) in einem einzigen Diagramm angezeigt werden, werden Bezugslinien (keine benutzerdefinierten Linien) gesondert auf jede Funktion angewendet.

Kategoriales Ziel (nur PDF). Diese Optionen sind nur aktiviert, wenn **Wahrscheinlichkeitsdichtefunktion (PDF)** ausgewählt ist.

- **Zu meldende Kategorienwerte.** Für Modelle mit kategorialen Zielfeldern ist das Ergebnis des Modells ein Set vorhergesagter Wahrscheinlichkeiten (eine für jede Kategorie), mit denen der Zielwert in die einzelnen Kategorien fällt. Die Kategorie mit der höchsten Wahrscheinlichkeit wird als vorhergesagte Kategorie übernommen und bei der Generierung des Balkendiagramms für die Wahrscheinlichkeitsdichtefunktion verwendet. Wählen Sie **Vorhergesagte Kategorie** aus, um das Balkendiagramm zu generieren. Wählen Sie **Vorhergesagte Wahrscheinlichkeiten** aus, um Histogramme der Verteilung vorhergesagter Wahrscheinlichkeiten für jede der Kategorien des Zielfelds zu generieren. Sie können auch **Beide** auswählen, um beide Diagrammtypen zu generieren.
- **Gruppierung für Sensitivitätsanalyse.** Simulationen, die Sensitivitätsanalyse-Iterationen enthalten, generieren für jede Iteration, die durch die Analyse definiert wird, ein unabhängiges Zielfeld (oder vorhergesagtes Zielfeld aus einem Modell). Für jeden Wert des Verteilungsparameters, der variiert wird, ist eine Iteration vorhanden. Wenn Iterationen vorhanden sind, wird das Balkendiagramm der vorhergesagten Kategorie für ein kategoriales Zielfeld als gruppiertes Balkendiagramm angezeigt, das die Ergebnisse für alle Iterationen enthält. Wählen Sie entweder **Kategorien zusammen gruppieren** oder **Iterationen zusammen gruppieren** aus.

Ansicht "Ausgaben"

Perzentilwerte der Zielverteilungen. Mit diesen Optionen können Sie eine Tabelle mit Perzentilwerten der Zielverteilungen erstellen und die Perzentile angeben, die angezeigt werden sollen.

Tabelle mit Perzentilwerten erstellen. Wählen Sie diese Option für stetige Zielfelder aus, um eine Tabelle mit angegebenen Perzentilen der Zielverteilungen zu erhalten. Wählen Sie eine der folgenden Optionen aus, um die Perzentile anzugeben:

- **Quartile.** Quartile sind das 25., 50. und 75. Perzentil der Zielfeldverteilung. Die Beobachtungen sind in vier gleich große Gruppen unterteilt.
- **Intervalle.** Wenn die Anzahl der gleich großen Gruppen nicht vier sein soll, wählen Sie **Intervalle** aus. Durch die Auswahl dieser Option wird das Feld **Anzahl** daneben aktiviert. Klicken Sie auf die Pfeile, um die Anzahl der Intervalle anzugeben. Die minimale Anzahl an Intervallen ist 2 und das Maximum ist 100. Die Anzahl der Intervalle ist anfänglich auf 10 gesetzt.
- **Benutzerdefinierte Perzentile.** Wählen Sie **Benutzerdefinierte Perzentile** aus, um einzelne Perzentile anzugeben, z. B. das 99. Perzentil. Durch die Auswahl dieser Option wird die Tabelle **Werte** daneben aktiviert. Jedes Mal, wenn Sie in die Tabelle **Werte** eine gültige Zahl zwischen 1 und 100 eingeben, wird unten an die Tabelle eine neue leere Zeile angefügt.

Simulationsevaluierungsknoten - Ausgabe

Wenn der Simulationsevaluierungsknoten ausgeführt wird, wird die Ausgabe dem Ausgabemanager hinzugefügt. Im Ausgabebrowser für die Simulationsevaluierung werden die Ergebnisse der Ausführung des Simulationsevaluierungsknotens angezeigt. Das Menü **Datei** enthält die üblichen Befehle zum Speichern, Exportieren und Drucken, das Menü **Bearbeiten** die üblichen Bearbeitungsfunktionen. Weitere Informationen finden Sie im Thema „Anzeigen der Ausgabe" auf Seite 283. Das Menü **Ansicht** ist nur aktiviert, wenn eines der Diagramme ausgewählt ist. Es ist nicht für die Verteilungstabelle oder für Informations-

ausgaben aktiviert. Im Menü **Ansicht** können Sie **Bearbeitungsmodus** auswählen, um das Layout und das Erscheinungsbild des Diagramms zu ändern, oder **Explorationsmodus**, um die Daten und Werte, die durch das Diagramm dargestellt werden, zu durchsuchen. Im statischen Modus werden die Bezugslinien (und Schieberegler) im Diagramm an ihren aktuellen Positionen fixiert, sodass sie nicht verschoben werden können. Der statische Modus ist der einzige Modus, in dem Sie das Diagramm mit seinen Bezugslinien kopieren, exportieren oder drucken können. Klicken Sie im Menü **Ansicht** auf **Statischer Modus**, um diesen Modus auszuwählen.

Das Ausgabebrowserfenster für die Simulationsevaluierung besteht aus zwei Anzeigen. Auf der linken Seite des Fensters befindet sich ein Navigationsfenster, in dem Piktogrammdarstellungen der Diagramme angezeigt werden, die bei der Ausführung des Simulationsevaluierungsknotens generiert wurden. Wenn ein Piktogramm ausgewählt wird, wird die Diagrammausgabe in der Anzeige auf der rechten Seite des Fensters angezeigt.

Navigationsfenster

Das Navigationsfenster des Ausgabebrowsers enthält Piktogramme der Diagramme, die aus einer Simulation generiert werden. Welche Piktogramme im Navigationsfenster angezeigt werden, hängt vom Messniveau des Zielfelds und von den Optionen ab, die im Knotendialogfeld "Simulationsevaluierung" ausgewählt werden. Beschreibungen der Piktogramme sind in der folgenden Tabelle angegeben.

Tabelle 52. Navigationsfensterpiktogramme

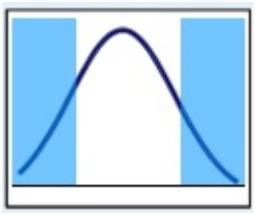
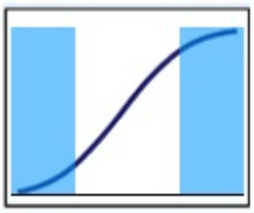
Piktogramm	Beschreibung	Kommentare
	Wahrscheinlichkeitsdichtefunktion	Dieses Piktogramm wird nur angezeigt, wenn das Messniveau des Zielfelds stetig ist und in der Ansicht "Dichtefunktionen" des Knotendialogfelds "Simulationsevaluierung" **Wahrscheinlichkeitsdichtefunktion (PDF)** ausgewählt ist. Wenn das Messniveau des Zielfelds "Kategorial" ist, wird dieses Piktogramm nicht angezeigt.
	Kumulative Verteilungsfunktion	Dieses Piktogramm wird nur angezeigt, wenn das Messniveau des Zielfelds stetig ist und in der Ansicht "Dichtefunktionen" des Knotendialogfelds "Simulationsevaluierung" **Kumulative Verteilungsfunktion (CDF)** ausgewählt ist. Wenn das Messniveau des Zielfelds "Kategorial" ist, wird dieses Piktogramm nicht angezeigt.

Tabelle 52. Navigationsfensterpiktogramme (Forts.)

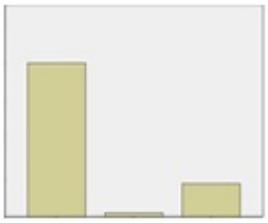
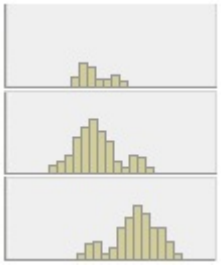
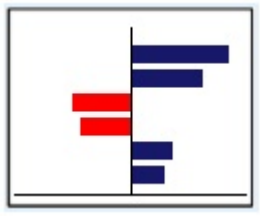
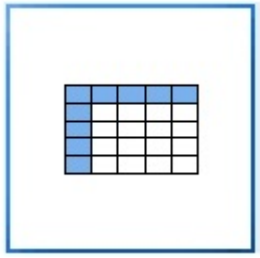
Piktogramm	Beschreibung	Kommentare
	Vorhergesagte Kategoriewerte	Dieses Piktogramm wird nur angezeigt, wenn das Messniveau des Zielfelds kategorial ist, in der Ansicht "Dichtefunktionen" des Knotendialogfelds "Simulationsevaluierung" **Wahrscheinlichkeitsdichtefunktion (PDF)** ausgewählt ist und im Bereich **Zu meldende Kategorienwerte** entweder **Vorhergesagte Kategorie** oder **Beide** ausgewählt ist. Wenn das Messniveau des Zielfelds "Stetig" ist, wird dieses Piktogramm nicht angezeigt.
	Vorhergesagte Kategoriewahrscheinlichkeiten	Dieses Piktogramm wird nur angezeigt, wenn das Messniveau des Zielfelds kategorial ist, in der Ansicht "Dichtefunktionen" des Knotendialogfelds "Simulationsevaluierung" **Wahrscheinlichkeitsdichtefunktion (PDF)** ausgewählt ist und im Bereich **Zu meldende Kategorienwerte** entweder **Vorhergesagte Wahrscheinlichkeiten** oder **Beide** ausgewählt ist. Wenn das Messniveau des Zielfelds "Stetig" ist, wird dieses Piktogramm nicht angezeigt.
	Tornado-Diagramme	Dieses Piktogramm wird nur angezeigt, wenn in der Ansicht "Felder" des Knotendialogfelds "Simulationsevaluierung" im Feld **Eingabefelder für Korrelations-Tornado** mindestens ein Eingabefeld ausgewählt ist.
	Verteilungstabelle	Dieses Piktogramm wird nur angezeigt, wenn das Messniveau des Zielfelds stetig ist und in der Ansicht "Ausgaben" des Knotendialogfelds "Simulationsevaluierung" **Tabelle mit Perzentilwerten erstellen** ausgewählt ist. Das Menü **Ansicht** ist für dieses Diagramm inaktiviert. Wenn das Messniveau des Zielfelds "Kategorial" ist, wird dieses Piktogramm nicht angezeigt.

Tabelle 52. Navigationsfensterpiktogramme (Forts.)

Piktogramm	Beschreibung	Kommentare
	Informationen	Dieses Piktogramm wird immer angezeigt. Das Menü **Ansicht** ist für diese Ausgabe inaktiviert.

Diagrammausgabe

Welche Typen von Ausgabediagrammen verfügbar sind, hängt vom Messniveau des Zielfelds, von den Optionen, die im Knotendialogfeld "Simulationsevaluierung" ausgewählt werden, sowie davon ab, ob ein Iterationsfeld verwendet wird. Einige der Diagramme, die aus einer Simulation generiert werden, verfügen über interaktive Funktionen, mit denen Sie die Anzeige anpassen können. Interaktive Funktionen sind durch Klicken auf **Diagrammoptionen** verfügbar. Bei allen Simulationsdiagrammen handelt es sich um Diagrammtafelvisualisierungen.

Diagramme der Wahrscheinlichkeitsdichtefunktionen für stetige Ziele. In diesem Diagramm werden sowohl die Wahrscheinlichkeit als auch die Häufigkeit angezeigt, wobei sich die Wahrscheinlichkeitsskala auf der linken vertikalen Achse und die Häufigkeitsskala auf der rechten vertikalen Achse befindet. Das Diagramm enthält zwei verschiebbare vertikale Bezugslinien, die das Diagramm in separate Bereiche aufteilen. In der Tabelle unter dem Diagramm wird der Prozentsatz der Verteilung in den einzelnen Bereichen angezeigt. Wenn mehrere Dichtefunktionen in demselben Diagramm angezeigt werden (aufgrund von Iterationen), enthält die Tabelle eine separate Zeile für die Wahrscheinlichkeiten, die jeder Dichtefunktion zugeordnet sind, und eine zusätzlichen Spalte, die den Iterationsnamen und eine Farbe enthält, die jeder Dichtefunktion zugeordnet ist. Die Iterationen sind in der Tabelle entsprechend der Iterationsbeschriftung in alphabetischer Reihenfolge aufgelistet. Wenn keine Iterationsbeschriftung verfügbar ist, wird statt dessen der Iterationswert verwendet. Die Tabelle kann nicht bearbeitet werden.

Jede der Bezugslinien verfügt über einen Schieberegler (umgekehrtes Dreieck), mit dem Sie die Linie einfach verschieben können. Jeder Schieberegler verfügt über eine Beschriftung, die seine aktuelle Position angibt. Standardmäßig sind die Schieberegler am 5. und am 95. Perzentil der Verteilung positioniert. Wenn mehrere Iterationen vorhanden sind, werden die Schieberegler am 5. und 95. Perzentil der ersten Iteration positioniert, die in der Tabelle aufgeführt ist. Sie können die Linien nicht so verschieben, dass sie sich kreuzen.

Durch Klicken auf **Diagrammoptionen** sind eine Reihe zusätzlicher Funktionen verfügbar. Insbesondere können Sie die Position der Schieberegler explizit festlegen, feste Bezugslinien hinzufügen und die Diagrammansicht von einer stetigen Kurve in ein Histogramm ändern. Weitere Informationen finden Sie im Thema „Diagrammoptionen” auf Seite 323. Klicken Sie mit der rechten Maustaste auf das Diagramm, um es zu kopieren oder zu exportieren.

Diagramme der kumulativen Verteilungsfunktionen für stetige Ziele. Dieses Diagramm enthält dieselben beiden verschiebbaren vertikalen Bezugslinien und dieselbe zugehörige Tabelle, die für das Diagramm der Wahrscheinlichkeitsdichtefunktionen beschrieben sind. Die Schiebeleistensteuerungen und die Tabelle verhalten sich so wie bei der Wahrscheinlichkeitsdichtefunktion, wenn mehrere Iterationen vorhanden sind. Die Farben, mit denen angegeben wird, welche Dichtefunktion zu den einzelnen Iterationen gehört, werden auch für die Verteilungsfunktionen verwendet.

Dieses Diagramm bietet ebenfalls Zugriff auf das Dialogfeld **Diagrammoptionen**, in dem Sie explizit die Position der Schieberegler festlegen, feste Bezugslinien hinzufügen und angeben können, ob die kumulative Verteilungsfunktion als steigende Funktion (Standardeinstellung) oder fallende Funktion dargestellt

werden soll. Weitere Informationen finden Sie im Thema „Diagrammoptionen” auf Seite 323. Klicken Sie mit der rechten Maustaste auf das Diagramm, um es zu kopieren, exportieren oder bearbeiten. Durch Auswahl von **Bearbeiten** wird das Diagramm in einem nicht verankerten Fenster des Diagrammtafel-Editors geöffnet.

Diagramm der vorhergesagten Kategoriewerte für kategoriale Ziele. In einem Balkendiagramm werden die vorhergesagten Werte für kategoriale Zielfelder angezeigt. Die vorhergesagten Werte werden als Prozentsatz des Zielfelds angezeigt, von dem vorhergesagt wird, dass es in die einzelnen Kategorien fällt. Für kategoriale Zielfelder mit Sensitivitätsanalyse-Iterationen werden Ergebnisse für die vorhergesagte Zielkategorie als gruppiertes Balkendiagramm angezeigt, das die Ergebnisse für alle Iterationen enthält. Das Diagramm wird nach Kategorie oder nach Iteration gruppiert, je nachdem, welche Option in der Ansicht "Dichtefunktionen" des Knotendialogfelds "Simulationsevaluierung" im Bereich **Gruppierung für Sensitivitätsanalyse** ausgewählt ist. Klicken Sie mit der rechten Maustaste auf das Diagramm, um es zu kopieren, exportieren oder bearbeiten. Durch Auswahl von **Bearbeiten** wird das Diagramm in einem nicht verankerten Fenster des Diagrammtafel-Editors geöffnet.

Diagramm der vorhergesagten Kategoriewahrscheinlichkeiten für kategoriale Ziele. Für kategoriale Zielfelder wird die Verteilung vorhergesagter Wahrscheinlichkeiten für jede der Kategorien des Ziels in einem Histogramm angezeigt. Für kategoriale Zielfelder mit Sensitivitätsanalyse-Iteration werden die Histogramme nach Kategorie oder nach Iteration angezeigt, je nachdem, welche Option in der Ansicht "Dichtefunktionen" des Knotendialogfelds "Simulationsevaluierung" im Bereich **Gruppierung für Sensitivitätsanalyse** ausgewählt ist. Wenn die Histogramme nach Kategorie gruppiert werden, können Sie in einer Dropdown-Liste, die die Iterationsbeschriftungen enthält, auswählen, welche Iteration angezeigt werden soll. Sie können die Iteration, die angezeigt werden soll, auch auswählen, indem Sie mit der rechten Maustaste auf das Diagramm klicken und die Iteration im Untermenü **Iteration** auswählen. Wenn die Histogramme nach Iteration gruppiert werden, können Sie in einer Dropdown-Liste, die die Kategorienamen enthält, auswählen, welche Kategorie angezeigt werden soll. Sie können auch auswählen, welche Kategorie angezeigt werden soll, indem Sie mit der rechten Maustaste auf das Diagramm klicken und die Kategorie im Untermenü **Kategorie** auswählen.

Dieses Diagramm ist nur für ein Subset von Modellen verfügbar und im Modellnugget muss die Option für die Generierung aller Gruppenwahrscheinlichkeiten ausgewählt sein. Im logistischen Modellnugget müssen Sie beispielsweise **Alle Wahrscheinlichkeiten ausgeben** auswählen. Die folgenden Modellnuggets unterstützen diese Option:

- Logistisch, SVM, Bayes, Neuronales Netz, KNN
- DB2-/ISW-Modelle für datenbankinternes Mining für logistische Regression, Entscheidungsbäume und Naive Bayes

Standardmäßig ist die Option zum Generieren aller Gruppenwahrscheinlichkeiten in diesen Modellnuggets nicht ausgewählt.

Tornado-Diagramme. Das Tornado-Diagramm ist ein Balkendiagramm, das die Sensitivität des Zielfelds für jede der angegebenen Eingaben anzeigt. Die Sensitivität wird durch die Korrelation des Ziels mit jeder Eingabe gemessen. Der Titel des Diagramms enthält den Namen des Zielfelds. Jeder Balken im Diagramm stellt die Korrelation zwischen dem Zielfeld und einem Eingabefeld dar. Die simulierten Eingaben, die im Diagramm enthalten sind, sind die Eingaben, die in der Ansicht "Felder" des Knotendialogfelds "Simulationsevaluierung" im Feld **Eingabefelder für Korrelations-Tornado** ausgewählt sind. Jeder Balken ist mit dem Korrelationswert beschriftet. Balken werden nach dem absoluten Wert der Korrelationen, vom größten Wert zum kleinsten, geordnet. Wenn Iterationen vorhanden sind, wird für jede Iteration ein separates Diagramm generiert. Jedes Diagramm hat einen Untertitel, der den Namen der Iteration enthält.

Verteilungstabelle. Diese Tabelle enthält den Wert des Zielfelds, unter den der angegebenen Prozentsatz der Beobachtungen fällt. Die Tabelle enthält eine Zeile für jeden Perzentilwert, der in der Ansicht "Ausgaben" des Knotendialogfelds "Simulationsevaluierung" angegeben ist. Die Perzentilwerte können Quartile,

eine andere Anzahl von Perzentilen in gleichmäßigen Abständen oder einzeln angegebene Perzentile sein. Die Verteilungstabelle enthält eine Spalte für jede Iteration.

Informationen. Dieser Abschnitt enthält eine Gesamtzusammenfassung der Felder und Datensätze, die in der Evaluierung verwendet werden. Er zeigt auch die Eingabefelder und die Datensatzanzahl für jede Iteration an.

Diagrammoptionen

Im Dialogfeld "Diagrammoptionen" können Sie die Anzeige aktivierter Diagramme von Wahrscheinlichkeitsdichtefunktionen und kumulativen Verteilungsfunktionen anpassen, die aus einer Simulation generiert wurden.

Ansicht. Die Dropdown-Liste **Ansicht** gilt nur für das Diagramm der Wahrscheinlichkeitsdichtefunktion. Sie können damit die Diagrammansicht zwischen einer stetigen Kurve und einem Histogramm umschalten. Diese Funktion ist inaktiviert, wenn mehrere Dichtefunktionen (aus mehreren Iterationen) in demselben Diagramm angezeigt werden. Wenn mehrere Dichtefunktionen vorhanden sind, können die Dichtefunktionen nur als stetige Kurven angezeigt werden.

Reihenfolge. Die Dropdown-Liste **Reihenfolge** gilt nur für das Diagramm der kumulativen Verteilungsfunktion. Sie gibt an, ob die kumulative Verteilungsfunktion als steigende Funktion (Standardeinstellung) oder fallende Funktion angezeigt wird. Bei der Anzeige als fallende Funktion gibt der Wert der Funktion an einem bestimmten Punkt auf der horizontalen Achse die Wahrscheinlichkeit an, mit der das Zielfeld rechts von diesem Punkt liegt.

Schieberpositionen. Das Textfeld **Oberer** enthält die aktuelle Position der rechten verschiebbaren Bezugslinie. Das Textfeld **Unterer** enthält die aktuelle Position der linken verschiebbaren Bezugslinie. Sie können die Positionen der verschiebbaren Bezugslinien explizit festlegen, indem Sie Werte in die Textfelder **Oberer** und **Unterer** eingeben. Der Wert im Textfeld **Unterer** muss unbedingt kleiner als der Wert im Textfeld **Oberer** sein. Sie können die linke Bezugslinie entfernen, indem Sie **-Unendlichkeit** auswählen und somit die Position auf negativ unendlich setzen. Durch diese Aktion wird das Textfeld **Unterer** inaktiviert. Sie können die rechte Bezugslinie entfernen, indem Sie **Unendlichkeit** auswählen und somit die Position auf unendlich setzen. Durch diese Aktion wird das Textfeld **Oberer** inaktiviert. Sie können nicht beide Bezugslinie entfernen. Durch die Auswahl von **-Unendlichkeit** wird das Kontrollkästchen **Unendlichkeit** inaktiviert und umgekehrt.

Bezugslinien. Sie können verschiedene feste vertikale Bezugslinien den Wahrscheinlichkeitsdichtefunktionen und kumulativen Verteilungsfunktionen hinzufügen.

- **Mittelwert.** Sie können eine Bezugslinie am Mittelwert des Zielfelds hinzuzufügen.
- **Median.** Sie können eine Bezugslinie am Median des Zielfelds hinzuzufügen.
- **Standardabweichungen.** Sie können Bezugslinien bei plus und minus einer angegebenen Anzahl an Standardabweichungen vom Mittelwert des Zielfelds hinzufügen. Sie können die Anzahl der zu verwendenden Standardabweichungen in das Textfeld daneben eingeben. Die minimale Anzahl an Standardabweichungen ist 1 und das Maximum ist 10. Die Anzahl der Standardabweichungen ist anfänglich auf 3 gesetzt.
- **Perzentile.** Sie können Bezugslinien an einem oder zwei Perzentilwerten der Verteilung für das Zielfeld hinzufügen, indem Sie Werte in die Textfelder **Unten** und **Oben** eingeben. Der Wert 95 im Textfeld **Oben** stellt beispielsweise das 95. Perzentil dar, den Wert, unter den 95 % der Beobachtungen fallen. Ebenso stellt der Wert 5 im Textfeld **Unten** das 5. Perzentil dar, den Wert, unter den 5 % der Beobachtungen fallen. Für das Textfeld **Unten** ist der minimale Perzentilwert 0 und das Maximum ist 49. Für das Textfeld **Oben** ist der minimale Perzentilwert 50 und das Maximum ist 100.
- **Benutzerdefinierte Positionen.** Sie können Bezugslinien an bestimmten Werten auf der horizontalen Achse des Zielfelds hinzufügen. Sie können benutzerdefinierte Bezugslinien entfernen, indem Sie den Eintrag aus dem Raster entfernen.

Wenn Sie auf **OK** klicken, werden die Schieberegler, die Beschriftungen über den Schiebereglern, Bezugslinien und die Tabelle unter dem Diagramm so aktualisiert, dass die Optionen berücksichtigt werden, die im Dialogfeld "Diagrammoptionen" ausgewählt sind. Klicken Sie auf **Abbrechen**, um das Dialogfeld zu schließen, ohne Änderungen vorzunehmen. Bezugslinien können entfernt werden, indem die zugehörige Auswahloption im Dialogfeld "Diagrammoptionen" inaktiviert wird und auf **OK** geklickt wird.

Anmerkung: Wenn mehrere Dichtefunktionen oder Verteilungsfunktionen (aufgrund der Ergebnisse aus Sensitivitätsanalyse-Iterationen) in einem einzigen Diagramm angezeigt werden, werden Bezugslinien (keine benutzerdefinierten Linien) gesondert auf jede Funktion angewendet. Es werden nur die Bezugslinien für die erste Iteration angezeigt. Die Bezugslinienbeschriftungen enthalten die Iterationsbeschriftung. Die Iterationsbeschriftung wird aus einem vorgeordneten Knoten abgeleitet, normalerweise aus einem Simulationsgenerierungsknoten. Wenn keine Iterationsbeschriftung verfügbar ist, wird statt dessen der Iterationswert verwendet. Die Option **Mittelwert**, **Median**, **Standardabweichungen** und **Perzentile** sind für kumulative Verteilungsfunktionen mit mehreren Iterationen inaktiviert.

IBM SPSS Statistics-Hilfsanwendungen

Wenn eine kompatible Version von IBM SPSS Statistics auf dem Computer installiert und lizenziert ist, können Sie IBM SPSS Modeler so konfigurieren, dass Daten mit IBM SPSS Statistics-Funktionen über den Statistics-Transformations-, den Statistics-Modell-, den Statistics-Ausgabe- oder den Statistics-Exportknoten verarbeitet werden.

Informationen zur Produktkompatibilität mit der aktuellen Version von IBM SPSS Modeler finden Sie auf der unternehmensweiten Support-Site unter *http://www.ibm.com/support*.

Um IBM SPSS Modeler für die Zusammenarbeit mit IBM SPSS Statistics und anderen Anwendungen zu konfigurieren, wählen Sie:

Tools > **Optionen** > **Hilfsanwendungen**

IBM SPSS Statistics Interactive. Geben Sie den vollständigen Pfad und Namen des Befehls (z. B. *C:\Programme\IBM\SPSS\Statistics\<nn>\stats.exe*) ein, der verwendet werden soll, wenn IBM SPSS Statistics direkt für eine vom Statistikexportknoten erzeugte Datendatei gestartet wird. Weitere Informationen finden Sie im Thema „Statistikexportknoten” auf Seite 361.

Verbindung. Wenn sich IBM SPSS Statistics Server auf demselben Host befindet wie IBM SPSS Modeler Server, können Sie eine Verbindung zwischen diesen beiden Anwendungen aktivieren, mit der die Effizienz gesteigert wird, weil Daten während der Analyse auf dem Server belassen werden. Mit **Server** aktivieren Sie unten die Option **Port**. Die Standardeinstellung lautet **Lokal**.

Port. Bestimmen Sie den Server-Port für IBM SPSS Statistics Server.

IBM SPSS Statistics-Standortdienstprogramm. Damit IBM SPSS Modeler den Statistics-Transformations-, den Statistcs-Modell- und den Statistics-Ausgabeknoten verwenden kann, muss auf dem Computer, auf dem der Stream ausgeführt wird, eine Kopie von IBM SPSS Statistics installiert und lizenziert sein.

- Wenn Sie IBM SPSS Modeler im lokalen Modus (Standalone-Modus) ausführen, muss sich die lizenzierte Kopie von IBM SPSS Statistics auf dem lokalen Computer befinden. Klicken Sie auf diese Schaltfläche, um den Standort der lokalen IBM SPSS Statistics-Installation anzugeben, die für die Lizenzierung verwendet werden soll.
- Bei einer Ausführung im verteilten Modus unter einer fernenIBM SPSS Modeler Server-Instanz müssen Sie außerdem auf dem IBM SPSS Modeler Server-Host ein Dienstprogramm ausführen, um die Datei *statistics.ini* zu erstellen, die für IBM SPSS Statistics den Installationspfad für IBM SPSS Modeler Server angibt. Wechseln Sie hierzu an der Eingabeauforderung in das IBM SPSS Modeler Server-Verzeichnis *bin* und führen Sie für Windows den folgenden Befehl aus:

  ```
  statisticsutility -location=<IBM SPSS Statistics-Installationspfad>/
  ```

Alternativ führen Sie unter UNIX folgenden Befehl aus:

```
./statisticsutility -location=<IBM SPSS Statistics-Installationspfad>/bin
```

Wenn sich keine lizenzierte Kopie von IBM SPSS Statistics auf Ihrem lokalen System befindet, können Sie den Statistikdateiknoten trotzdem mithilfe eines lizenzierten IBM SPSS Statistics-Servers ausführen. Das Ausführen anderer IBM SPSS Statistics-Knoten wird jedoch zu Fehlernachrichten führen.

Kommentare

Wenn Schwierigkeiten beim Ausführen von IBM SPSS Statistics-Prozedurknoten auftreten, beachten Sie die folgenden Tipps:

- Falls die Feldnamen in IBM SPSS Modeler länger als acht (bei Versionen vor IBM SPSS Statistics) bzw. 64 Zeichen sind (bei IBM SPSS Statistics 12.0 und späteren Versionen) oder ungültige Zeichen enthalten, müssen diese Namen vor dem Einlesen in IBM SPSS Statistics geändert oder gekürzt werden. Weitere Informationen finden Sie im Thema „Umbenennen oder Filtern von Feldern für IBM SPSS Statistics” auf Seite 362.
- Wenn IBM SPSS Statistics nach IBM SPSS Modeler installiert wurde, müssen Sie möglicherweise die Position von IBM SPSS Statistics angeben, wie oben erläutert.

Kapitel 7. Exportknoten

Überblick über Exportknoten

Exportknoten bieten einen Mechanismus zum Exportieren von Daten in verschiedenen Formaten, sodass Sie diese Daten auch mit anderen Software-Tools nutzen können.

Folgende Exportknoten stehen zur Verfügung:

Der Datenbankexportknoten schreibt Daten in eine ODBC-kompatible relationale Datenquelle. Um Daten in eine ODBC-Datenquelle schreiben zu können, muss die betreffende Datenquelle bereits vorhanden sein und Sie benötigen Schreibzugriff dafür.

Der Flatfile-Exportknoten gibt Daten in einer Textdatei mit Trennzeichen aus. Diese Vorgehensweise eignet sich für das Exportieren von Daten, die von anderen Analyse- oder Tabellenkalkulationsprogrammen gelesen werden sollen.

Der Statistikexportknoten gibt Daten im IBM SPSS Statistics-Format *.sav* oder *.zsav* aus. Die *.sav*- oder *.zsav*-Dateien können von IBM SPSS Statistics Base und anderen Produkten gelesen werden. Dieses Format wird auch für Cache-Dateien in IBM SPSS Modeler verwendet.

Der IBM SPSS Data Collection-Exportknoten gibt Daten in dem von der Marktforschungssoftware IBM SPSS Data Collection verwendeten Format aus. Um diesen Knoten verwenden zu können, muss die IBM SPSS Data Collection Data Library installiert sein.

Der IBM Cognos BI-Exportknoten exportiert Daten in einem Format, das von Cognos BI-Datenbanken gelesen werden kann.

Der IBM Cognos TM1-Exportknoten exportiert Daten in einem Format, das von Cognos TM1-Datenbanken gelesen werden kann.

Mit dem SAS-Exportknoten werden Daten in das SAS-Format ausgegeben, die dann in SAS oder in SAS-kompatible Softwarepakete eingelesen werden können. Drei SAS-Dateiformate sind verfügbar: SAS für Windows/OS2, SAS für UNIX sowie SAS Version 7/8.

Der Excel-Exportknoten gibt Daten im Microsoft Excel-Format (*.xls*) aus. Optional können Sie auswählen, dass bei der Ausführung des Knotens Excel automatisch gestartet und die exportierte Datei geöffnet werden soll.

Der XML-Exportknoten gibt Daten an eine Datei im XML-Format aus. Optional können Sie einen XML-Quellenknoten erstellen, um die exportierten Daten wieder in der Stream einzulesen.

Datenbankexportknoten

Mithilfe von Datenbankknoten können Sie Daten in ODBC-konforme relationale Datenquellen schreiben, die in der Beschreibung des Quellenknotens "Datenbank" erläutert werden. Weitere Informationen finden Sie im Thema „Datenbankquellenknoten” auf Seite 13.

Führen Sie die folgenden allgemeinen Schritte aus, um Daten in eine Datenbank zu schreiben:

1. Installieren Sie einen ODBC-Treiber und konfigurieren Sie eine Datenquelle für die zu verwendende Datenbank.
2. Geben Sie auf der Registerkarte "Exportieren" des Datenbankknotens die Datenquelle und die Tabelle an, in die geschrieben werden soll. Sie können eine neue Tabelle erstellen oder Daten in eine bestehende Tabelle einfügen.
3. Geben Sie nach Bedarf weitere Optionen an.

Diese Schritte werden in den nächsten Themenabschnitten ausführlicher beschrieben.

Registerkarte "Exportieren" beim Datenbankknoten

Datenquelle. Zeigt die ausgewählte Datenquelle. Geben Sie den Namen ein oder wählen Sie einen Eintrag in der Dropdown-Liste aus. Wird die gewünschte Datenbank nicht in der Liste aufgeführt, wählen Sie **Neue Datenbankverbindung hinzufügen** und wechseln Sie im Dialogfeld "Datenbankverbindungen" zu dieser Datenbank. Weitere Informationen finden Sie im Thema „Hinzufügen einer Datenbankverbindung” auf Seite 15.

Tabellenname. Geben Sie den Namen der Tabelle ein, an die die Daten gesendet werden sollen. Bei der Option **In Tabelle einfügen** können Sie eine vorhandene Tabelle in der Datenbank auswählen, indem Sie auf die Schaltfläche **Auswählen** klicken.

Tabellen erstellen. Mit dieser Option können Sie eine neue Datenbanktabelle anlegen oder eine vorhandene Datenbanktabelle überschreiben.

In Tabelle einfügen. Mit dieser Option fügen Sie die Daten als neue Zeilen in eine vorhandene Datenbanktabelle ein.

Tabellen zusammenführen. (Wenn verfügbar) Aktivieren Sie diese Option, um ausgewählte Datenbankspalten mit Werten aus entsprechenden Quellendatenfeldern zu aktualisieren. Wenn Sie diese Option auswählen, wird die Schaltfläche **Zusammenführen** aktiviert, die ein Dialogfeld öffnet, in dem Sie Quellendatenfelder zu Datenbankspalten zuordnen können.

Vorhandene Tabelle löschen. Wenn Sie eine neue Tabelle erstellen, lassen Sie mit dieser Option alle vorhandenen Tabellen löschen, die denselben Namen besitzen wie die neu zu erstellende Tabelle.

Vorhandene Zeilen löschen. Wenn Sie Daten in eine Tabelle einfügen, lassen Sie mit dieser Option vorhandene Zeilen vor dem Exportieren aus der Tabelle löschen.

Hinweis: Wenn eine oben genannten Optionen ausgewählt ist, wird eine **Überschreibungswarnung** ausgegeben, sobald Sie den Knoten ausführen. Sollen diese Warnungen unterdrückt werden, inaktivieren Sie im Dialogfeld "Benutzeroptionen" auf der Registerkarte "Benachrichtigungen" die Option **Warnen, wenn eine Datenbanktabelle durch einen Knoten überschrieben wird**.

Standardzeichenfolgegröße. Felder, die Sie als "Ohne Typ" in einem aufwärts liegenden Typknoten gekennzeichnet haben, werden als Zeichenfolgenfelder in die Datenbank geschrieben. Geben Sie die Größe der Zeichenfolgen an, die für Felder ohne Typ verwendet werden sollen.

Klicken Sie auf **Schema**, um ein Dialogfeld zu öffnen, in dem Sie verschiedene Exportoptionen festlegen können (für Datenbanken, die diese Funktion unterstützen), und geben Sie den Primärschlüssel für die Datenbankindizierung an. Weitere Informationen finden Sie im Thema „Schemaoptionen für den Datenbankexport" auf Seite 330.

Klicken Sie auf **Indizes**, um Optionen für die Indizierung der exportierten Tabelle anzugeben und damit die Datenbankleistung zu verbessern. Weitere Informationen finden Sie im Thema „Indexoptionen für den Datenbankexport" auf Seite 333.

Klicken Sie auf **Erweitert**, um Optionen für das Massenladen und die Datenbankübertragung festzulegen. Weitere Informationen finden Sie im Thema „Erweiterte Optionen für den Datenbankexport" auf Seite 335.

Tabellen- und Spaltennamen in Anführungszeichen. Wählen Sie die Optionen aus, die beim Senden der Anweisung `CREATE TABLE` an die Datenbank verwendet werden sollen. Enthält der Name von Tabellen und Spalten ein Leerzeichen oder ein Sonderzeichen, muss der Name in Anführungszeichen gesetzt werden.

- **Nach Bedarf.** Hiermit lassen Sie automatisch von Fall zu Fall durch IBM SPSS Modeler feststellen, ob Anführungszeichen erforderlich sind oder nicht.
- **Immer.** Die Tabellen- und Spaltennamen werden immer in Anführungszeichen eingeschlossen.
- **Nie.** Es werden keine Anführungszeichen verwendet.

Importknoten für diese Daten generieren. Es wird ein Datenbankquellenknoten für die Daten erzeugt, die in die angegebene Datenquelle und Tabelle exportiert wurden. Beim Ausführen wird dieser Knoten in den Streamerstellungsbereich aufgenommen.

Zusammenführungsoptionen für den Datenbankexport

Dieses Dialogfeld ermöglicht Ihnen, Felder aus den Quellendaten zu Spalten in der Zieldatenbanktabelle zuzuordnen Beim Zuordnen eines Quellendatenfelds zu einer Datenbankspalte wird der Spaltenwert durch den Quellendatenwert ersetzt, wenn der Stream ausgeführt wird. Nicht zugeordnete Quellenfelder bleiben in der Datenbank unverändert.

Felder zuordnen. Hier geben Sie die Zuordnung zwischen Quellendatenfeldern und Datenbankspalten an. Quellendatenfelder mit demselben Namen wie Spalten in der Datenbank werden automatisch zugeordnet.

- **Zuordnen.** Ordnet ein Quellendatenfeld, das in der Feldliste links neben der Schaltfläche ausgewählt wurde, einer Datenbankspalte zu, die in der Liste rechts ausgewählt wurde. Sie können mehrere Felder gleichzeitig zuordnen, aber die Anzahl der ausgewählten Einträge muss in beiden Listen gleich sein.
- **Zuordnung aufheben.** Entfernt die Zuordnung für ein oder mehrere ausgewählte Datenbankspalten. Diese Schaltfläche wird aktiviert, wenn Sie ein Feld oder eine Datenbankspalte in der Tabelle im rechten Bereich des Dialogfelds auswählen.
- **Hinzufügen.** Fügt ein oder mehr Quellendatenfelder, die in der Feldliste links neben der Schaltfläche ausgewählt wurden, der Liste rechts zu, die bereit für die Zuordnung ist. Diese Schaltfläche wird aktiviert, wenn Sie ein Feld in der Liste im linken Bereich auswählen und in der Liste im rechten Bereich kein Feld mit diesem Namen vorhanden ist. Wenn Sie auf die Schaltfläche klicken, wird das ausgewählte Feld einer neuen Datenbankspalte mit dem gleichen Namen zugeordnet. Das Wort **<NEU>** wird hinter dem Namen der Datenbankspalte angezeigt, um anzuzeigen, dass es sich um ein neues Feld handelt.

Zeilen zusammenführen. Sie verwenden ein Schlüsselfeld wie *Transaktions-ID*, um Datensätze mit demselben Wert im Schlüsselfeld zusammenzuführen. Dies entspricht einem Datenbank-"Equi-Join". Schlüsselwerte müssen zu Primärschlüsseln gehören, das heißt, sie müssen eindeutig sein und dürfen keine Nullwerte enthalten.

- **Mögliche Felder.** Listet alle Felder auf, die in allen Eingabedatenquellen gefunden wurden. Wählen Sie ein oder mehrere Felder aus dieser Liste und fügen Sie sie mithilfe der Pfeilschaltfläche als Schlüsselfeld für die Zusammenführung von Datensätzen hinzu. Jedes Zuordnungsfeld mit einer entsprechenden zugeordneten Datenbankspalte ist als Schlüssel verfügbar, lediglich Felder, die als neue Datenbankspalten hinzugefügt wurden (durch **<NEU>** nach dem Namen gekennzeichnet) sind nicht verfügbar.
- **Verwendete Schlüsselfelder.** Listet alle Felder auf, die für die Zusammenführung der Datensätze aus allen Eingabedatenquellen auf der Grundlage der Schlüsselfeldwerte verwendet werden. Um einen Schlüssel aus der Liste zu entfernen, wählen Sie ihn aus und verschieben Sie ihn mithilfe der Pfeilschaltfläche zurück in die Liste "Mögliche Schlüsselfelder". Bei Auswahl mehrerer Schlüssel wird die unten stehende Option aktiviert.
- **Nur Datensätze einschließen, die in der Datenbank vorhanden sind.** Führt einen partiellen Join aus. Wenn sich der Datensatz in der Datenbank und im Stream befindet, werden die zugeordneten Felder aktualisiert.
- **Datensätze zur Datenbank hinzufügen.** Führt einen Outer Join aus. Alle Datensätze im Stream werden zusammengeführt (wenn derselbe Datensatz in der Datenbank vorhanden ist) oder hinzugefügt (wenn der Datensatz noch nicht in der Datenbank existiert).

So ordnen Sie ein Quellendatenfeld einer neuen Datenbankspalte zu:

1. Klicken Sie auf den Quellenfeldnamen in der Liste links unter **Felder zuordnen**.
2. Klicken Sie auf die Schaltfläche **Hinzufügen**, um die Zuordnung abzuschließen.

So ordnen Sie ein Quellendatenfeld einer vorhandenen Datenbankspalte zu:

1. Klicken Sie auf den Quellenfeldnamen in der Liste links unter **Felder zuordnen**.
2. Klicken Sie rechts unter **Datenbankspalte** auf den Spaltennamen.
3. Klicken Sie auf die Schaltfläche **Zuordnen**, um die Zuordnung abzuschließen.

So entfernen Sie eine Zuordnung:

1. Klicken Sie in der Liste rechts unter "Feld" auf den Namen des Felds, für das Sie die Zuordnung entfernen möchten.
2. Klicken Sie auf die Schaltfläche **Zuordnung aufheben**.

So wählen Sie ein Feld in einer beliebigen Liste aus

Klicken Sie bei gedrückter Steuertaste auf den Feldnamen.

Schemaoptionen für den Datenbankexport

Im Dialogfeld "Schema" für den Datenbankexport können Sie Optionen für den Datenbankexport festlegen (für Datenbanken, die diese Optionen unterstützen), die SQL-Datentypen für die Felder bestimmen, angeben, bei welchen Feldern es sich um Primärschlüssel handelt, sowie die beim Exportieren erstellte Anweisung `CREATE TABLE` anpassen.

Das Dialogfeld besteht aus verschiedenen Teilen:

- Der Abschnitt oben (sofern angezeigt) enthält Optionen für den Export in eine Datenbank, die diese Optionen unterstützt. Dieser Bereich wird nicht angezeigt, wenn Sie nicht mit einer solchen Datenbank verbunden sind.
- Das Textfeld in der Mitte zeigt die zur Generierung des Befehls `CREATE TABLE` Vorlage an, die standardmäßig folgendes Format aufweist:

```
CREATE TABLE <table-name> <(table columns)>
```

- Mit der Tabelle im unteren Bereich können Sie den SQL-Datentyp für die einzelnen Felder festlegen und angeben, bei welchen Feldern es sich um Primärschlüssel handelt (siehe unten). Das Dialogfeld generiert automatisch die Werte der Parameter `<table-name>` und `<(table columns)>` auf der Grundlage der Spezifikationen in der Tabelle.

Festlegen der Optionen für den Datenbankexport

Wenn dieser Abschnitt angezeigt wird, können Sie eine Reihe von Einstellungen für den Export in die Datenbank angeben. Diese Funktion wird von folgenden Datenbanktypen unterstützt:

- IBM InfoSphere Warehouse unter DB2 9.1 oder höher. Weitere Informationen finden Sie im Thema „Optionen für IBM DB2 InfoSphere Warehouse" auf Seite 332.
- SQL Server 2008 oder höher, Enterprise und Developer Edition. Weitere Informationen finden Sie im Thema „Optionen für SQL Server" auf Seite 332.
- Oracle 10g und 11gR1 oder höher, Enterprise oder Personal Edition. Weitere Informationen finden Sie im Thema „Optionen für Oracle" auf Seite 332.

Anpassen der Anweisung CREATE TABLE

Im Textfeldbereich dieses Dialogfelds können Sie zusätzliche datenbankspezifische Optionen in die Clementine-Anweisung `CREATE TABLE` aufnehmen.

1. Aktivieren Sie das Kontrollkästchen **Befehl CREATE TABLE anpassen**, damit das Textfenster zur Verfügung gestellt wird.
2. Ergänzen Sie die Anweisung mit den gewünschten datenbankspezifischen Optionen. Die Textparameter `<table-name>` und `(<table-columns>)` müssen beibehalten werden, weil diese Parameter in IBM SPSS Modeler durch den tatsächlichen Tabellennamen und die entsprechende Spaltendefinition ersetzt werden.

Festlegen von SQL-Datentypen

Standardmäßig ist es bei IBM SPSS Modeler möglich, dass die SQL-Datentypen automatisch durch den Datenbankserver zugewiesen werden. Soll der automatisch festgelegte Typ für ein Feld überschrieben werden, wechseln Sie zur zugehörigen Zeile für das Feld und wählen Sie den gewünschten Typ in der Schematabelle in der Dropdown-Liste in der Spalte *Type* aus. Durch Drücken der Umschalttaste und Klicken können Sie mehr als eine Reihe auswählen.

Bei Typen, für die ein Längen-, Genauigkeits- oder Skalenargument erforderlich ist (`BINARY`, `VARBINARY`, `CHAR`, `VARCHAR`, `NUMERIC` und `NUMBER`), sollten Sie selbst eine Länge zuweisen lassen, also nicht die automatische Längenzuweisung durch den Datenbankserver nutzen. Wenn Sie beispielsweise einen angemessenen Wert für die Länge festlegen, z. B. `VARCHAR(25)`, ist sichergestellt, dass der Speichertyp in IBM SPSS Modeler überschrieben wird (falls Sie dies wünschen). Soll die automatische Zuweisung überschrieben werden, wählen Sie in der Dropdown-Liste "Typ" den Eintrag **Angeben** und ersetzen Sie die Typdefinition durch die gewünschte Anweisung für die SQL-Typdefinition.

Die einfachste Methode besteht darin, zunächst den Typ auszuwählen, der der gewünschten Typdefinition am nächsten kommt, dann die Option **Angeben** zu wählen und schließlich die zugehörige Definition zu bearbeiten. Um den SQL-Datentyp beispielsweise auf `VARCHAR(25)` zu setzen, wählen Sie zunächst in der Dropdown-Liste "Typ" den Eintrag **VARCHAR(length)** aus. Wählen Sie dann **Angeben** und ersetzen Sie die Textlänge durch den Wert 25.

Primärschlüssel

Wenn eine oder mehrere Spalten in der exportierten Tabelle einem eindeutigen Wert bzw. eine eindeutige Wertekombination für jede Zeile aufweisen muss, können Sie dies angeben, indem Sie für jedes betroffene

Feld das Kontrollkästchen **Primärschlüssel** aktivieren. Bei den meisten Datenbanken ist es nicht zulässig, die Tabelle auf eine Weise zu ändern, die eine Primärschlüsselbeschränkung ungültig macht. Bei diesen Datenbanken wird zur Durchsetzung dieser Einschränkung automatisch ein Index über dem Primärschlüssel erstellt. (Optional können Sie im Dialogfeld "Indizes" Indizes für andere Felder erstellen). Weitere Informationen finden Sie im Thema „Indexoptionen für den Datenbankexport” auf Seite 333.)

Optionen für IBM DB2 InfoSphere Warehouse

Tabellenbereich. Der Tabellenbereich, der für den Export verwendet wird. Datenbankadministratoren können Tabellenbereiche partioniert erstellen oder konfigurieren. Wir empfehlen, einen dieser Tabellenbereiche (anstelle des standardmäßig eingestellten) für den Datenbankexport zu verwenden.

Daten nach Feld partitionieren. Legt das Eingabefeld für die Partionierung fest.

Komprimierung verwenden. Bei Auswahl dieser Option werden Tabellen für den komprimierten Export erstellt (entspricht z. B. `CREATE TABLE MYTABLE(...) COMPRESS YES;` in SQL).

Optionen für SQL Server

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt.

Komprimierung für. Wählen Sie die Komprimierungsstufe aus.

- **Zeile.** Aktiviert Komprimierung auf der Zeilenebene (z. B. die Entsprechung von `CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = ROW);` in SQL).
- **Seite.** Aktiviert Komprimierung auf der Seitenebene (z. B. `CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = PAGE);` in SQL).

Optionen für Oracle

Oracle 10g-Einstellungen

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt. Für diese Version der Datenbank steht nur einfache Komprimierung zur Verfügung (beispielsweise `CREATE TABLE MYTABLE(...) COMPRESS;` in SQL).

Oracle 11gR1-Einstellungen

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt.

Komprimierung für. Wählen Sie die Komprimierungsstufe aus.

- **Standard.** Aktiviert Standardkomprimierung (z. B. `CREATE TABLE MYTABLE(...) COMPRESS;` in SQL). In diesem Fall hat sie dieselbe Wirkung wie die Option **Direkte Ladevorgänge**.
- **Direkte Ladevorgänge.** Aktiviert Komprimierung ausschließlich für Masseneinfügevorgänge (direkter Pfad) (z. B. `CREATE TABLE MYTABLE(...)COMPRESS FOR DIRECT_LOAD OPERATIONS;` in SQL).
- **Alle Vorgänge.** Aktiviert Komprimierung für alle Vorgänge (z. B. `CREATE TABLE MYTABLE(...)COMPRESS FOR ALL OPERATIONS;` in SQL).

Oracle 11gR2-Einstellungen - Option "Basic" (Einfach)

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt.

Komprimierung für. Wählen Sie die Komprimierungsstufe aus.

- **Standard.** Aktiviert Standardkomprimierung (z. B. `CREATE TABLE MYTABLE(...) COMPRESS;` in SQL). In diesem Fall hat sie dieselbe Wirkung wie die Option **Einfach**.

- **Einfach.** Aktiviert einfache Komprimierung (z. B. `CREATE TABLE MYTABLE(...) COMPRESS BASIC;` in SQL).

Oracle 11gR2-Einstellungen - Option "Advanced" (Erweitert)

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt.

Komprimierung für. Wählen Sie die Komprimierungsstufe aus.

- **Standard.** Aktiviert Standardkomprimierung (z. B. `CREATE TABLE MYTABLE(...) COMPRESS;` in SQL). In diesem Fall hat sie dieselbe Wirkung wie die Option **Einfach**.
- **Einfach.** Aktiviert einfache Komprimierung (z. B. `CREATE TABLE MYTABLE(...) COMPRESS BASIC;` in SQL).
- **OLTP.** Aktiviert OLTP-Komprimierung (z. B. `CREATE TABLE MYTABLE(...)COMPRESS FOR OLTP;` in SQL).
- **Abfrage niedrig/hoch.** (Nur Exadata-Server) Aktiviert Hybrid Columnar Compression für Abfrage (z. B. `CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY LOW;` oder `CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY HIGH;` in SQL). Komprimierung für Abfragen ist in Data Warehousing-Umgebungen sinnvoll; `HIGH` (Hoch) bietet ein höheres Komprimierungsverhältnis als `LOW` (Niedrig).
- **Archiv niedrig/hoch.** (Nur Exadata-Server) Aktiviert Hybrid Columnar Compression für Archiv (z. B. `CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE LOW;` oder `CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE HIGH;` in SQL). Komprimierung für Archive ist sinnvoll zur Komprimierung von Daten, die lange Zeit gespeichert werden sollen; `HIGH` (Hoch) bietet ein höheres Komprimierungsverhältnis als `LOW` (Niedrig).

Indexoptionen für den Datenbankexport

Über das Dialogfeld "Indizes" können Sie Indizes für Datenbanktabellen erstellen, die aus IBM SPSS Modeler exportiert wurden. Sie können die einzuschließenden Feldsets angeben und den Befehl `CREATE INDEX` nach Bedarf anpassen.

Das Dialogfeld besteht aus zwei Teilen:

- Das Textfeld im oberen Teil zeigt eine Vorlage an, die zur Generierung eines oder mehrerer Befehle vom Typ `CREATE TABLE` verwendet werden kann. Diese Vorlage weist standardmäßig folgendes Format auf:

  ```
  CREATE INDEX <Indexname> ON <Tabellenname>
  ```

- Die Tabelle im unteren Bereich des Dialogfelds ermöglicht die Angabe von Spezifikationen für jeden Index, der erstellt werden soll. Geben Sie für jeden Index den Indexnamen und die einzuschließenden Felder bzw. Spalten an. Das Dialogfeld generiert automatisch die Werte der Parameter `<index-name>` und `<table-name>`.

  Beispielsweise kann die generierte SQL für einen einzelnen Index für die Felder *empid* und *deptid* wie folgt aussehen:

  ```
  CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID)
  ```

  Sie können mehrere Zeilen hinzufügen, um mehrere Indizes zu erstellen. Für jede Zeile wird ein gesonderter `CREATE INDEX`-Befehl generiert.

Anpassen des Befehls CREATE INDEX

Optional können Sie den Befehl `CREATE INDEX` für alle Indizes oder nur für einen bestimmten Index anpassen. Dadurch haben Sie die Flexibilität, spezielle Datenbankanforderungen oder -optionen zu berücksichtigen und Anpassungen nach Bedarf auf alle oder nur auf bestimmte Indizes anzuwenden.

- Wählen Sie **Befehl CREATE INDEX anpassen** oben im Dialogfeld, um die Vorlage, die für alle danach hinzugefügten Indizes verwendet wird, anzupassen. Beachten Sie, dass die Änderungen nicht automatisch auf Indizes angewendet werden, die bereits zur Tabelle hinzugefügt wurden.

- Wählen Sie mindestens eine Zeile in der Tabelle aus und klicken Sie oben im Dialogfeld auf **Ausgewählte Indizes aktualisieren**, um die aktuellen Anpassungen auf alle ausgewählten Zeilen anzuwenden.
- Aktivieren Sie das Kontrollkästchen **Anpassen** in den einzelnen Zeilen, um die Befehlsvorlage nur für den betreffenden Index zu ändern.

Beachten Sie, dass die Werte der Parameter `<Indexname>` und `<Tabellenname>` vom Dialogfeld automatisch auf der Grundlage der Tabellenangaben generiert werden und nicht direkt bearbeitet werden können.

BITMAP KEYWORD. Bei Verwendung einer Oracle-Datenbank können Sie die Vorlage so anpassen, dass statt eines Standardindex ein Bitmapindex erstellt wird. Dies geschieht wie folgt:

```
CREATE BITMAP INDEX <Indexname> ON <Tabellenname>
```

Bitmapindizes sind nützlich für die Indizierung von Spalten mit einer kleinen Anzahl unterschiedlicher Werte. Die entstehende SQL sieht etwa folgendermaßen aus:

```
CREATE BITMAP INDEX MYTABLE_IDX1 ON MYTABLE(COLOR)
```

Schlüsselwort UNIQUE. Die meisten Datenbanken unterstützen das Schlüsselwort `UNIQUE` im Befehl `CREATE INDEX`. Dadurch wird eine Eindeutigkeitsbeschränkung ähnlich einer Primärschlüsselbeschränkung in der zugrunde liegenden Tabelle erzwungen.

```
CREATE UNIQUE INDEX <Indexname> ON <Tabellenname>
```

Beachten Sie, dass diese Angabe für Felder, die tatsächlich als Primärschlüssel angegeben sind, nicht erforderlich ist. Die meisten Datenbanken erstellen automatisch einen Index für alle Felder, die im Befehl `CREATE TABLE` als Primärschlüsselfelder festgelegt wurden. Eine explizite Erstellung von Indizes für diese Felder ist also nicht erforderlich. Weitere Informationen finden Sie im Thema „Schemaoptionen für den Datenbankexport” auf Seite 330.

Schlüsselwort FILLFACTOR. Für einige physische Parameter des Index können Feineinstellungen vorgenommen werden. Beispielsweise ermöglicht SQL Server es dem Benutzer, die Indexgröße (nach der ursprünglichen Erstellung) gegen die Wartungskosten bei zukünftigen Änderungen an der Tabelle abzuwägen.

```
CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID) WITH FILLFACTOR=20
```

Weitere Kommentare

- Wenn bereits ein Index mit dem angegebenen Namen vorhanden ist, schlägt die Indexerstellung fehl. Alle Fehlschläge werden zunächst als Warnungen behandelt, sodass die nachfolgenden Indizes erstellt werden können. Nachdem die Erstellung aller Indizes versucht wurde, werden diese Fehlschläge dann im Nachrichtenprotokoll als Fehler gemeldet.
- Um eine bestmögliche Leistung zu erzielen, sollten die Indizes erstellt werden, nachdem Daten in die Tabelle geladen wurden. Indizes müssen mindestens eine Spalte enthalten.
- Vor der Ausführung des Knotens können Sie die generierte SQL im Nachrichtenprotokoll anzeigen.
- Für temporäre Tabellen, die in die Datenbank geschrieben wurden (d. h. wenn der Knoten-Cache aktiviert ist) sind die Optionen zur Angabe von Primärschlüsseln und Indizes nicht verfügbar. Das System kann jedoch nach Bedarf Indizes in der temporären Tabelle erstellen, je nachdem, wie die Daten in nachgeordneten Knoten verwendet werden sollen. Wenn die Daten im Cache beispielsweise anschließend durch eine *DEPT*-Spalte verbunden werden, ist es sinnvoll, die im Cache gespeicherte Tabelle auf dieser Spalte zu indizieren.

Indizes und Abfrageoptimierung

In einigen Datenbankverwaltungssystemen ist nach dem Erstellen, Laden und Indizieren einer Datenbanktabelle ein weiterer Schritt erforderlich, bevor das Optimierungsprogramm die Indizes zur Beschleunigung der Abfrageausführung in der neuen Tabelle nutzen kann. In Oracle beispielsweise erfordert das kostenbasierte Abfrageoptimierungsprogramm, dass eine Tabelle analysiert wird, bevor ihre Indizes für die Abfrageoptimierung verwendet werden können. Die interne ODBC-Eigenschaftendatei für Oracle (für den Benutzer nicht sichtbar) enthält hierfür eine Option:

```
# Defines SQL to be executed after a table and any associated indexes
# have been created and populated
table_analysis_sql, 'ANALYZE TABLE <table-name> COMPUTE STATISTICS'
```

Dieser Schritt wird bei jeder Erstellung einer Tabelle in Oracle ausgeführt (unabhängig davon, ob Primärschlüssel oder Indizes definiert sind). Falls erforderlich, kann die ODBC-Eigenschaftsdatei für zusätzliche Datenbanken auf ähnliche Weise angepasst werden. Falls Sie Unterstützung benötigen, wenden Sie sich an den Technical Support.

Erweiterte Optionen für den Datenbankexport

Wenn Sie im Dialogfeld für den Datenbankexportknoten auf die Schaltfläche "Erweitert" klicken, wird ein neues Dialogfeld angezeigt, in dem Sie die technischen Einzelheiten für das Exportieren der Ergebnisse in eine Datenbank festlegen können.

Stapelübertragung verwenden. Hiermit inaktivieren Sie die zeilenweise Übertragung an die Datenbank.

Stapelgröße. Hier können Sie die Anzahl der Datensätze angeben, die an die Datenbank gesendet werden sollen, bevor die Übertragung in den Speicher erfolgt. Wenn Sie hier einen niedrigeren Wert angeben, erzielen Sie eine größere Datenintegrität, jedoch auf Kosten der Übertragungsgeschwindigkeit. Nehmen Sie gegebenenfalls Feineinstellungen an diesem Wert vor, um so die optimale Leistung der Datenbank zu erreichen.

InfoSphere Warehouse-Optionen. Wird nur angezeigt, wenn Sie mit einer InfoSphere Warehouse-Datenbank verbunden sind (IBM DB2 9.7 oder höher). **Do not log updates** (Aktualisierungen nicht protokollieren) erlaubt Ihnen, das Protokollieren von Ereignissen zu inaktivieren, wenn Sie Tabellen erstellen und Daten einfügen.

Bulk Loading verwenden. Gibt eine Methode an, mit der die Daten per Massenladen direkt aus IBM SPSS Modeler in die Datenbank übernommen werden. Möglicherweise müssen Sie ein wenig herumexperimentieren, um herauszufinden, welche Massenladeoperationen für ein bestimmtes Szenario angemessen sind.

- **Über ODBC.** Mit dieser Option lassen Sie Einfügungen mehrerer Zeilen durch die ODBC-API vornehmen; dies ist effizienter als der normale Export in die Datenbank. Wählen Sie die zeilenweise oder spaltenweise Bindung in den Optionen im unteren Bereich.
- **Über externes Ladeprogramm.** Es wird ein benutzerdefiniertes Massenladeprogramm verwendet, das speziell auf die Datenbank abgestimmt ist. Wenn Sie diese Option aktivieren, wird im unteren Bereich eine Reihe von Optionen eingeblendet.

Erweiterte ODBC-Optionen. Diese Optionen stehen nur dann zur Verfügung, wenn Sie die Option **Über ODBC** aktiviert haben. Beachten Sie, dass diese Funktionalität möglicherweise nicht von allen ODBC-Treibern unterstützt wird.

- **Zeilenweise.** Bei der zeilenweisen Bindung werden die Daten über den Aufruf `SQLBulkOperations` in die Datenbank geladen. Mit der zeilenweisen Bindung verbessern Sie in der Regel die Geschwindigkeit im Vergleich zu parametrisierten Einfügungen, bei denen die Daten jeweils Datensatz für Datensatz eingefügt werden.
- **Spaltenweise.** Die Daten werden mithilfe der spaltenweisen Bindung in die Datenbank geladen. Bei der spaltenweisen Bindung wird die Leistung gesteigert, indem die einzelnen Datenbankspalten (in ei-

ner parametrisierten INSERT-Anweisung) zu einem Array mit *n* Werten verbunden werden. Wenn Sie die INSERT-Anweisung einmal ausführen, werden *N* Zeilen in die Datenbank eingefügt. Diese Methode kann die Leistung drastisch erhöhen.

Optionen für externes Ladeprogramm. Wenn Sie die Option **Über externes Ladeprogramm** aktiviert haben, wird eine Reihe von Optionen eingeblendet, mit denen Sie das Dataset in eine Datei exportieren und die Daten dann mithilfe eines benutzerdefinierten Ladeprogramms aus dieser Datei in die Datenbank laden können. IBM SPSS Modeler kann mit externen Ladeprogrammen für viele beliebte Datenbanksysteme zusammenarbeiten. In der Software sind mehrere Scripts enthalten. Diese sind außerdem zusammen mit technischer Dokumentation im Unterverzeichnis *scripts* enthalten. Beachten Sie: Um diese Funktionen verwenden zu können, muss Python 2.7 auf demselben Computer installiert sein wie IBM SPSS Modeler oder IBM SPSS Modeler Server und der Parameter python_exe_path muss in der Datei *options.cfg* festgelegt sein. Weitere Informationen finden Sie im Thema „Programmierung des Massenladeprogramms".

- **Trennzeichen verwenden.** Hier können Sie angeben, welches Trennzeichen in der exportierten Datei verwendet werden soll. Bei der Option **Tabulator** erfolgt die Trennung mit Tabulatoren, bei der Option **Leerzeichen** entsprechend mit Leerzeichen. Mit der Option **Andere** können Sie ein anderes Zeichen angeben, beispielsweise ein Komma (,).
- **Datendatei angeben.** Geben Sie den Pfad für die Datendatei ein, die beim Massenladen geschrieben wird. Standardmäßig wird eine temporäre Datei im Verzeichnis "temp" auf dem Server angelegt.
- **Ladeprogramm angeben.** Hiermit können Sie ein Programm für das Massenladen auswählen. Standardmäßig wird das Unterverzeichnis *scripts* der IBM SPSS Modeler-Installation nach einem Python-Script durchsucht, das für eine bestimmte Datenbank ausgeführt werden soll. In der Software sind mehrere Scripts enthalten. Diese sind außerdem zusammen mit technischer Dokumentation im Unterverzeichnis *scripts* enthalten.
- **Protokoll generieren.** Im angegebenen Verzeichnis wird eine Protokolldatei generiert. Die Protokolldatei enthält Fehlerinformationen und ist von Nutzen, falls das Massenladen fehlschlägt.
- **Tabellengröße prüfen.** Hiermit lassen Sie eine Tabellenprüfung vornehmen, mit der Sie sicherstellen, dass der Anstieg der Tabellengröße mit der Anzahl der Zeilen übereinstimmt, die aus IBM SPSS Modeler exportiert wurden.
- **Weitere Optionen für Ladeprogramm.** Hier können Sie weitere Argumente für das Ladeprogramm festlegen. Argumente, die ein Leerzeichen enthalten, müssen in doppelte Anführungszeichen eingeschlossen werden.

Sollen doppelte Anführungszeichen in optionale Argumente eingefügt werden, stellen Sie den Anführungszeichen jeweils einen umgekehrten Schrägstrich voran. Die Option `-comment "Dies ist ein \"Kommentar\""` umfasst sowohl das Flag `-comment` als auch den Kommentar selbst, dessen Ausgabe `Dies ist ein "Kommentar"` lautet.

Um einen umgekehrten Schrägstrich einzufügen, stellen Sie diesem einen weiteren umgekehrten Schrägstrich voran. Die Option `-specialdir "C:\\Testscripts\\"` beispielsweise enthält das Flag `-specialdir` und das als `C:\Testscripts\` wiedergegebene Verzeichnis.

Programmierung des Massenladeprogramms

Der Datenbankexportknoten beinhaltet Optionen für das Massenladen im Dialogfeld "Erweiterte Optionen". Mit Massenladeprogrammen können Daten aus einer Textdatei in eine Datenbank geladen werden.

Mit der Option **Bulk Loading verwenden - Über externes Ladeprogramm** wird IBM SPSS Modeler für drei Aktionen konfiguriert:

- Erstellen aller erforderlichen Datenbanktabellen.
- Exportieren der Daten in eine Textdatei.
- Aufrufen eines Massenladeprogramms, um die Daten aus dieser Datei in die Datenbanktabelle zu laden.

Normalerweise handelt es sich bei dem Massenladeprogramm nicht um das Datenbankladeprogramm selbst (beispielsweise das Dienstprogramm sqlldr von Oracle), sondern um ein kleines Script bzw. ein kleines Programm, das die richtigen Argumente bildet, alle erforderlichen datenbankspezifischen Hilfsdateien (beispielsweise eine Steuerdatei) erstellt und anschließend das Datenbank-Ladeprogramm aufruft. Anhand der Informationen in den folgenden Abschnitten können Sie ein bestehendes Massenladeprogramm bearbeiten.

Alternativ können Sie Ihr eigenes Programm für das Massenladen schreiben. Weitere Informationen finden Sie im Thema „Entwickeln von Massenladeprogrammen" auf Seite 341.

Scripts für das Massenladen

IBM SPSS Modeler wird mit einer Reihe von Massenladeprogrammen für verschiedene Datenbanken ausgeliefert, die mithilfe von Pythons-Scripts implementiert werden. Wenn Sie einen Stream, der einen Datenbankexportknoten enthält ausführen, während die Option **Über externes Ladeprogramm** ausgewählt ist, erstellt IBM SPSS Modeler die Datenbanktabelle (sofern erforderlich) über ODBC, exportiert die Daten in eine temporäre Datei auf dem Host, auf dem IBM SPSS Modeler Server ausgeführt wird, und ruft anschließend das Massenladescript auf. Dieses Script führt dann Dienstprogramme aus, die vom DBMS-Anbieter bereitgestellt wurden, um Daten aus den temporären Dateien in die Datenbank hochzuladen.

Hinweis: Die IBM SPSS Modeler-Installation enthält keinen Python-Laufzeitinterpreter, weshalb eine gesonderte Installation von Python erforderlich ist. Weitere Informationen finden Sie im Thema „Erweiterte Optionen für den Datenbankexport" auf Seite 335.

Für die in der folgenden Tabelle aufgeführten Datenbanken werden (im Ordner *\scripts* des IBM SPSS Modeler-Installationsverzeichnisses) Scripts bereitgestellt.

Tabelle 53. Bereitgestellte Massenladescripts.

Datenbank	Scriptname	Weitere Informationen
IBM DB2	*db2_loader.py*	Weitere Informationen finden Sie im Thema „Massenladen von Daten in IBM DB2-Datenbanken".
IBM Netezza	*netezza_loader.py*	Weitere Informationen finden Sie im Thema „Massenladen von Daten in IBM Netezza-Datenbanken" auf Seite 338.
Oracle	*oracle_loader.py*	Weitere Informationen finden Sie im Thema „Massenladen von Daten in Oracle-Datenbanken" auf Seite 339.
SQL Server	*mssql_loader.py*	Weitere Informationen finden Sie im Thema „Massenladen von Daten in SQL Server-Datenbanken" auf Seite 339.
Teradata	*teradata_loader.py*	Weitere Informationen finden Sie im Thema „Massenladen von Daten in Teradata-Datenbanken" auf Seite 340.

Massenladen von Daten in IBM DB2-Datenbanken

Die folgenden Punkte können Ihnen bei der Konfiguration für das Massenladen von IBM SPSS Modeler in eine IBM DB2-Datenbank mithilfe der Optionen für das externe Ladeprogramm im Dialogfeld "DB-Export: Erweiterte Optionen" behilflich sein.

Sicherstellen, dass das Dienstprogramm DB2 Command Line Processor (CLP)installiert ist

Das Script *db2_loader.py* ruft den DB2 LOAD-Befehl auf. Vergewissern Sie sich, dass der Befehlszeilenprozessor (*db2* unter UNIX, *db2cmd* unter Windows) auf dem Server installiert ist, auf dem *db2_loader.py* ausgeführt werden soll (üblicherweise der Host, auf dem IBM SPSS Modeler Server ausgeführt wird).

Überprüfen, ob der Aliasname der lokalen Datenbank mit dem tatsächlichen Datenbanknamen übereinstimmt

Der Aliasname der lokalen DB2-Datenbank ist der Name, der von der DB2-Client-Software verwendet wird, um auf eine Datenbank in einer lokalen oder entfernten DB2-Instanz zu verweisen. Wenn der Aliasname der lokalen Datenbank vom Namen der Remote-Datenbank abweicht, geben Sie zusätzlich folgende Option für das Ladeprogramm an:

```
-alias <Alias_der_lokalen_Datenbank>
```

Hier ein Beispiel: Die Remote-Datenbank trägt den Namen STARS und befindet sich auf dem Host GALAXY, der Alias der lokalen DB2-Datenbank auf dem Host, auf dem IBM SPSS Modeler Server ausgeführt wird, ist jedoch STARS_GALAXY. Verwenden Sie die zusätzliche Ladeprogrammoption

```
-alias STARS_GALAXY
```

Datencodierung mit Nicht-ASCII-Zeichen

Wenn Sie Massenladen von Daten durchführen, die nicht im ASCII-Format vorliegen, sollten Sie sicherstellen, dass die Codeseitenvariable im Konfigurationsabschnitt von *db2_loader.py* in Ihrem System richtig eingerichtet ist.

Leere Zeichenfolgen

Leere Zeichenfolgen werden als NULL-Werte in die Datenbank exportiert.

Massenladen von Daten in IBM Netezza-Datenbanken

Die folgenden Punkte können Ihnen bei der Konfiguration für das Massenladen von IBM SPSS Modeler in eine IBM Netezza-Datenbank mithilfe der Optionen für das externe Ladeprogramm im Dialogfeld "DB-Export: Erweiterte Optionen" behilflich sein.

Sicherstellen, dass das Netezza-Dienstprogramm "nzload" installiert ist

Das Script *netezza_loader.py* ruft das Netezza-Dienstprogramm *nzload* auf. Vergewissern Sie sich, dass *nzload* installiert und ordnungsgemäß auf dem Server konfiguriert ist, auf dem *netezza_loader.py* ausgeführt werden soll.

Exportieren von Nicht-ASCII-Daten

Wenn Ihr Bericht Daten enthält, die nicht im ASCII-Format vorliegen, müssen Sie möglicherweise dem Feld **Weitere Optionen für Ladeprogramm** im Dialogfeld "DB-Export: Erweiterte Optionen" die Angabe `-encoding UTF8` hinzufügen. Dadurch sollte sichergestellt werden, dass Nicht-ASCII-Daten ordnungsgemäß hochgeladen werden.

Daten in den Formaten "Datum", "Zeit" und "Zeitmarke"

Setzen Sie in den Streameigenschaften das Datumsformat auf **TT-MM-JJJJ** und das Zeitformat auf **HH:MM:SS**.

Leere Zeichenfolgen

Leere Zeichenfolgen werden als NULL-Werte in die Datenbank exportiert.

Andere Spaltenreichenfolge in Stream- und Zieltabelle beim Einfügen von Daten in eine bestehende Tabelle

Wenn die Spaltenreihenfolge im Stream von der in der Zieltabelle abweicht, werden Datenwerte in die falschen Spalten eingefügt. Verwenden Sie einen Knoten vom Typ "Felder ordnen", um sicherzustellen, dass die Reihenfolge der Spalten im Stream mit der Reihenfolge in der Zieltabelle übereinstimmt. Weitere Informationen finden Sie im Thema „Knoten "Felder ordnen"" auf Seite 172.

Verfolgen des nzload-Fortschritts

Fügen Sie bei Ausführung von IBM SPSS Modeler im lokalen Modus dem Feld **Weitere Optionen für Ladeprogramm** im Dialogfeld "DB-Export: Erweiterte Optionen" die Angabe `-sts` hinzu, um nach jeweils 10.000 Zeilen im vom Dienstprogramm *nzload* geöffneten Befehlsfenster Statusnachrichten anzuzeigen.

Massenladen von Daten in Oracle-Datenbanken

Die folgenden Punkte können Ihnen bei der Konfiguration für das Massenladen von IBM SPSS Modeler in eine Oracle-Datenbank mithilfe der Optionen für das externe Ladeprogramm im Dialogfeld "DB-Export: Erweiterte Optionen" behilflich sein.

Sicherstellen, dass das Oracle-Dienstprogramm "sqlldr" installiert ist

Das Script *oracle_loader.py* ruft das Oracle-Dienstprogramm *sqlldr* auf. Beachten Sie, dass *sqlldr* nicht automatisch in Oracle Client enthalten ist. Vergewissern Sie sich, dass *sqlldr* auf dem Server installiert ist, auf dem *oracle_loader.py* ausgeführt werden soll.

SID bzw. Service-Name der Datenbank angeben

Wenn Sie Daten an einen nicht lokalen Oracle-Server exportieren oder Ihr lokaler Oracle-Server mehrere Datenbanken enthält, müssen Sie im Feld **Weitere Optionen für Ladeprogramm** im Dialogfeld "DB-Export: Erweiterte Optionen" Folgendes angeben, um die SID bzw. den Service-Namen weiterzugeben:

```
-database <SID>
```

Bearbeiten des Konfigurationsabschnitts in oracle_loader.py

Bearbeiten Sie unter UNIX-Systemen (und optional: Windows-Systemen) den Konfigurationsabschnitt zu Beginn des Scripts *oracle_loader.py*. Hier können gegebenenfalls Werte für die Umgebungsvariablen ORACLE_SID, NLS_LANG, TNS_ADMIN und ORACLE_HOME angegeben werden, sowie der vollständige Pfad zum Dienstprogramm *sqlldr*.

Daten in den Formaten "Datum", "Zeit" und "Zeitmarke"

In den Streameigenschaften sollten Sie normalerweise das Datumsformat auf **JJJJ-MM-TT** und das Zeitformat auf **HH:MM:SS** setzen.

Wenn Sie ein von den oben genannten Werten abweichendes Datums- und Zeitformat verwenden müssen, lesen Sie in Ihrer Oracle-Dokumentation nach und bearbeiten Sie die Scriptdatei *oracle_loader.py*.

Datencodierung mit Nicht-ASCII-Zeichen

Wenn Sie Massenladen von Daten durchführen, die nicht im ASCII-Format vorliegen, sollten Sie sicherstellen, dass die Umgebungsvariable NLS_LANG in Ihrem System richtig eingerichtet ist. Dieses wird vom Oracle-Ladedienstprogramm *sqlldr* gelesen. Beispielsweise ist der richtige Wert für NLS_LANG für Shift-JIS unter Windows Japanese_Japan.JA16SJIS. Weitere Details zu NLS_LANG finden Sie in Ihrer Oracle-Dokumentation.

Leere Zeichenfolgen

Leere Zeichenfolgen werden als NULL-Werte in die Datenbank exportiert.

Massenladen von Daten in SQL Server-Datenbanken

Die folgenden Punkte können Ihnen bei der Konfiguration für das Massenladen von IBM SPSS Modeler in eine SQL Server-Datenbank mithilfe der Optionen für das externe Ladeprogramm im Dialogfeld "DB-Export: Erweiterte Optionen" behilflich sein.

Sicherstellen, dass das SQL Server-Dienstprogramm "bcp.exe" installiert ist

Das Script *mssql_loader.py* ruft das SQL Server-Dienstprogramm *bcp.exe* auf. Vergewissern Sie sich, dass *bcp.exe* auf dem Server installiert ist, auf dem *mssql_loader.py* ausgeführt werden soll.

Die Verwendung von Leerzeichen als Trennzeichen funktioniert nicht

Vermeiden Sie, im Dialogfeld "DB-Export: Erweiterte Optionen" Leerzeichen als Trennzeichen auszuwählen.

Option "Tabellengröße prüfen" empfohlen

Es wird empfohlen, die Option **Tabellengröße prüfen** im Dialogfeld "DB-Export: Erweiterte Optionen" zu aktivieren. Fehler beim Massenladevorgang werden nicht immer erkannt und durch die Aktivierung dieser Option wird eine zusätzliche Prüfung durchgeführt, um sicherzustellen, dass die richtige Anzahl an Zeilen geladen wurde.

Leere Zeichenfolgen

Leere Zeichenfolgen werden als NULL-Werte in die Datenbank exportiert.

Massenladen von Daten in Teradata-Datenbanken

Die folgenden Punkte können Ihnen bei der Konfiguration für das Massenladen von IBM SPSS Modeler in eine Teradata-Datenbank mithilfe der Optionen für das externe Ladeprogramm im Dialogfeld "DB-Export: Erweiterte Optionen" behilflich sein.

Sicherstellen, dass das Teradata-Dienstprogramm "fastload" installiert ist

Das Script *teradata_loader.py* ruft das Teradata-Dienstprogramm *fastload* auf. Vergewissern Sie sich, dass *fastload* installiert und ordnungsgemäß auf dem Server konfiguriert ist, auf dem *teradata_loader.py* ausgeführt werden soll.

Massenladen von Daten ist nur in leere Tabellen möglich

Als Ziele für das Massenladen sind nur leere Tabellen möglich. Wenn eine Zieltabelle bereits vor dem Massenladevorgang Daten enthält, kann der Vorgang nicht durchgeführt werden.

Daten in den Formaten "Datum", "Zeit" und "Zeitmarke"

Setzen Sie in den Streameigenschaften das Datumsformat auf **JJJJ-MM-TT** und das Zeitformat auf **HH:MM:SS**.

Leere Zeichenfolgen

Leere Zeichenfolgen werden als NULL-Werte in die Datenbank exportiert.

Teradata-Prozess-ID (tdpid)

Standardmäßig exportiert *fastload* Daten mit `tdpid=dbc` in das Teradata-System. Normalerweise gibt es einen Eintrag in der HOSTS-Datei, der `dbccop1` mit der IP-Adresse des Teradate-Servers verknüpft. Wenn Sie einen anderen Server verwenden möchten, geben Sie Folgendes im Feld **Weitere Optionen für Ladeprogramm** im Dialogfeld "DB-Export: Erweiterte Optionen" ein, um die `tdpid` dieses Servers zu übergeben:

```
-tdpid <id>
```

Leerzeichen in Tabellen- und Spaltennamen

Wenn ein Tabellen- oder Spaltenname Leerzeichen enthält schlägt der Massenladevorgang fehl. Benennen Sie nach Möglichkeit die Tabellen bzw. Spalten um, um die Leerzeichen zu entfernen.

Entwickeln von Massenladeprogrammen

In diesem Thema wird erläutert, wie Sie ein Massenladeprogramm entwickeln können, das über IBM SPSS Modeler ausgeführt werden kann, um Daten aus einer Textdatei in eine Datenbank zu laden.

Verwenden von Python zum Erstellen von Massenladeprogrammen

Standardmäßig sucht IBM SPSS Modeler anhand des Datenbanktyps nach einem Standardmassenladeprogramm. Siehe Tabelle 53 auf Seite 337.

Sie können das Script *test_loader.py* zur Unterstützung der Entwicklung von Massenladeprogrammen verwenden. Weitere Informationen finden Sie im Thema „Testen von Massenladeprogrammen" auf Seite 343.

An das Massenladeprogramm übergebene Objekte

IBM SPSS Modeler schreibt zwei Dateien, die an das Massenladeprogramm übergeben werden.

- **Datendatei.** Enthält die zu ladenden Daten im Textformat.
- **Schemadatei.** Dies ist eine XML-Datei, die die Namen und Typen der Spalten beschreibt und Informationen darüber bereitstellt, wie die Datendatei formatiert ist (beispielsweise, welches Zeichen als Trennzeichen zwischen Feldern verwendet wird).

Zusätzlich übergibt IBM SPSS Modeler weitere Informationen, wie Tabellenname, Benutzername und Kennwort, als Argumente, wenn das Massenladeprogramm aufgerufen wird.

Hinweis: Um IBM SPSS Modeler einen erfolgreichen Abschluss des Vorgangs zu signalisieren, sollte das Massenladeprogramm die Schemadatei löschen.

An das Massenladeprogramm übergebene Argumente

Die Parameter, die an das Programm übergeben werden, sind in der folgenden Tabelle aufgelistet.

Tabelle 54. An das Massenladeprogramm übergebene Argumente.

Argument	Beschreibung
`schemafile`	Pfad zur Schemadatei
`data file`	Pfad zur Datendatei
`servername`	Name des DBMS-Servers; kann leer sein
`databasename`	Name der Datenbank innerhalb des DBMS-Servers; kann leer sein
`username`	Benutzername für die Anmeldung bei der Datenbank
`password`	Kennwort für die Anmeldung bei der Datenbank
`tablename`	Name der zu ladenden Tabelle
`ownername`	Name des Tabellenbesitzers (auch als Schemaname bezeichnet)
`logfilename`	Name der Protokolldatei (wenn leer, wird keine Protokolldatei erstellt)
`rowcount`	Anzahl der Zeilen im Dataset

Alle im Feld **Weitere Optionen für Ladeprogramm** im Dialogfeld "DB-Export: Erweiterte Optionen" angegebenen Optionen werden nach diesen Standardargumenten an das Massenladeprogramm übergeben.

Format der Datendatei

Daten werden im Textformat in die Datendatei geschrieben. Dabei werden die einzelnen Felder durch ein Trennzeichen getrennt, das im Dialogfeld "DB-Export: Erweiterte Optionen" angegebenen wurde. Im Folgenden sehen Sie ein Beispiel für das Erscheinungsbild einer tabstoppgetrennten Datendatei.

```
48  F  HIGH    NORMAL  0.692623  0.055369  drugA
15  M  NORMAL  HIGH    0.678247  0.040851  drugY
37  M  HIGH    NORMAL  0.538192  0.069780  drugA
35  F  HIGH    HIGH    0.635680  0.068481  drugA
```

Die Datei wird in der von IBM SPSS Modeler Server verwendeten lokalen Codierung geschrieben (bzw. von IBM SPSS Modeler, wenn keine Verbindung zu IBM SPSS Modeler Server besteht. Ein Teil der Formatierung wird über die Streameinstellungen von IBM SPSS Modeler festgelegt.

Format der Schemadatei

Die Schemadatei ist eine XML-Datei, die zur Beschreibung der Datendatei dient. Im Folgenden sehen Sie ein Beispiel, das zur oben stehenden Datendatei gehört.

```
<?xml version="1.0" encoding="UTF-8"?>
<DBSCHEMA version="1.0">
 <table delimiter="\t" commit_every="10000" date_format="YYYY-MM-DD" time_format="HH:MM:SS"
append_existing="false" delete_datafile="false">
  <column name="Age" encoded_name="416765" type="integer"/>
  <column name="Sex" encoded_name="536578" type="char" size="1"/>
  <column name="BP" encoded_name="4250" type="char" size="6"/>
  <column name="Cholesterol" encoded_name="43686F6C65737465726F6C" type="char" size="6"/>
  <column name="Na" encoded_name="4E61" type="real"/>
  <column name="K" encoded_name="4B" type="real"/>
  <column name="Drug" encoded_name="44727567" type="char" size="5"/>
 </table>
</DBSCHEMA>
```

In den folgenden Tabellen werden die Attribute der Elemente `<table>` und `<column>` der Schemadatei aufgeführt.

Tabelle 55. Attribute des Elements <table>.

Attribut	Beschreibung
`delimiter`	Das Feldtrennzeichen (Tabulator wird als "\t" dargestellt).
`commit_every`	Das Intervall für die Stapelgröße (wie im Dialogfeld "DB-Export: Erweiterte Optionen").
`date_format`	Das zur Darstellung von Datumswerten verwendete Format.
`time_format`	Das zur Darstellung von Zeitwerten verwendete Format.
`append_existing`	`true`, wenn die zu ladende Tabelle bereits Daten enthält; ansonsten `false`.
`delete_datafile`	`true`, wenn das Massenladeprogramm die Datendatei nach Abschluss des Ladevorgangs löschen soll.

Tabelle 56. Attribute des Elements <column>.

Attribut	Beschreibung
`name`	Der Spaltenname.
`encoded_name`	Der Spaltenname, in dieselbe Codierung konvertiert wie die Datendatei und ausgegeben als Reihe von zweistelligen Hexadezimalzahlen.
`type`	Der Datentyp der Spalte: `integer`, `real`, `char`, `time`, `date` oder `datetime`.
`size`	Für den Datentyp `char` die maximale Breite der Spalte in Zeichen.

Testen von Massenladeprogrammen

Sie können die Massenladefunktion mithilfe des Testscripts *test_loader.py* testen, das im Ordner *\scripts* des Installationsverzeichnisses von IBM SPSS Modeler enthalten ist. Dies ist nützlich für Entwicklung, Debugging und Fehlerbehandlung von Massenladeprogrammen oder Scripts zur Verwendung mit IBM SPSS Modeler.

Gehen Sie zur Verwendung des Testscripts wie folgt vor.

1. Führen Sie das Script *test_loader.py* aus, um die Schemadatei und die Datendatei in die Dateien *schema.xml* und *data.txt* zu kopieren und eine Windows-Stapeldatei (*test.bat*) zu erstellen.
2. Bearbeiten Sie die Datei *test.bat*, um das zu testende Massenladeprogramm bzw. Script auszuwählen.
3. Führen Sie die Datei *test.bat* über eine Befehls-Shell aus, um das ausgewählte Massenladeprogramm bzw. Script zu testen.

Hinweis: Bei der Ausführung von *test.bat* werden nicht tatsächlich Daten in die Datenbank geladen.

Flatfile-Exportknoten

Mit dem Knoten "Flatfile" schreiben Sie Daten in eine Textdatei, die mit Trennzeichen getrennt ist. Diese Vorgehensweise eignet sich für das Exportieren von Daten, die durch andere Analyse- oder Tabellenkalkulationsprogramme gelesen werden sollen.

Hinweis: Es ist nicht möglich, Dateien im alten Cacheformat zu schreiben, weil IBM SPSS Modeler dieses Format nicht mehr für Cachedateien verwendet. IBM SPSS Modeler-Cachedateien werden nun im IBM SPSS Statistics-Format (*.sav*) gespeichert. Dieses Format kann mit einem Statistics-Exportknoten geschrieben werden. Weitere Informationen finden Sie im Thema „Statistikexportknoten" auf Seite 361.

Registerkarte "Exportieren" beim Flatfile-Knoten

Datei exportieren. Hier können Sie den Namen der Datei angeben. Geben Sie einen Dateinamen an oder klicken Sie auf die Feldauswahlschaltfläche und wechseln Sie zum Pfad der gewünschten Datei.

Schreibmodus. Wenn die Option **Überschreiben** aktiviert ist, werden alle vorhandenen Daten in der angegebenen Datei überschrieben. Ist die Option **Anhängen** aktiviert, wird die Ausgabe an die vorhandene Datei angehängt; die bereits vorhandenen Daten in dieser Datei werden also beibehalten.

- **Feldnamen einschließen.** Bei dieser Option werden die Dateinamen in die erste Zeile der Ausgabedatei geschrieben. Diese Option ist nur für den Schreibmodus **Überschreiben** verfügbar.

Neue Zeile nach jedem Datensatz. Bei dieser Option wird jeder Datensatz in eine eigene Zeile in der Ausgabedatei geschrieben.

Feldtrennzeichen. Dient zur Angabe des Zeichens, das als Trennzeichen zwischen den Feldwerten in der erzeugten Textdatei eingefügt werden soll. Die folgenden Optionen stehen zur Auswahl: **Komma**, **Tabulator**, **Leerzeichen** und **Andere**. Wenn Sie die Option **Andere** wählen, geben Sie das oder die gewünschten Trennzeichen in das Textfeld ein.

Symbolanführungszeichen. Hier können Sie die Art der Anführungszeichen angeben, die für Werte in symbolischen Feldern verwendet werden sollen. Die folgenden Optionen stehen zur Auswahl: **Keine** (die Werte werden nicht in Anführungszeichen eingeschlossen), **Einfach (')**, **Doppelt (")** und **Andere**. Wenn Sie die Option **Andere** wählen, geben Sie das oder die gewünschten Anführungszeichen in das Textfeld ein.

Codierung. Gibt die verwendete Textcodierungsmethode an. Sie haben die Wahl zwischen der Systemstandardeinstellung, der Streamstandardeinstellung und UTF-8.

- Die Systemstandardeinstellung wird in der Windows-Systemsteuerung bzw. bei Ausführung im verteilten Modus auf dem Server-Computer angegeben.

- Der Streamstandard wird im Dialogfeld "Streameigenschaften" festgelegt.

Dezimalzeichen. Hier können Sie das Trennzeichen für die Dezimalstellen in den Daten festlegen.

- **Streamstandardeinstellung.** Das Dezimaltrennzeichen, das durch die Standardeinstellung des aktuellen Streams definiert ist, wird verwendet. In der Regel ist dies das Dezimaltrennzeichen aus den Ländereinstellungen des Computers.
- **Punkt (.).** Als Dezimaltrennzeichen wird ein Punkt verwendet.
- **Komma (,).** Als Dezimaltrennzeichen wird ein Komma verwendet.

Importknoten für diese Daten generieren. Mit dieser Option lassen Sie automatisch einen Quellenknotens für variable Dateien erzeugen, mit dem die exportierte Datendatei eingelesen wird. Weitere Informationen finden Sie im Thema „Knoten "Variable Datei"" auf Seite 20.

IBM SPSS Data Collection-Exportknoten

Der IBM SPSS Data Collection-Exportknoten speichert Daten in dem von der Marktforschungssoftware IBM SPSS Data Collection (beruht auf IBM SPSS Data Collection Data Model) verwendeten Format. Bei diesem Format wird zwischen Falldaten, also den tatsächlichen Antworten auf Fragen, die während einer Umfrage gesammelt werden, und Metadaten unterschieden, die beschreiben, wie der Fall gesammelt und organisiert wird. Metadaten bestehen aus Informationen wie Fragetexten, Variablennamen und -beschreibungen, Mehrfachantwortsets, Übersetzungen der verschiedenen Texte und der Definition der Struktur der Falldaten. Weitere Informationen finden Sie im Thema „Data Collection-Knoten" auf Seite 26.

Hinweis: Für diesen Knoten ist IBM SPSS Data Collection Data Model Version 4.0 oder höher erforderlich, das mit Software von IBM SPSS Data Collection ausgeliefert wird. Weitere Informationen finden Sie auf der IBM SPSS Data Collection-Webseite unter http://www.ibm.com/software/analytics/spss/products/data-collection/. Abgesehen von der Installation von Data Model sind keine weiteren Konfigurationen erforderlich.

Metadatendatei. Gibt den Namen der Fragebogendefinitionsdatei (*.mdd*) an, in der die exportierten Metadaten gespeichert werden sollen. Auf der Grundlage der Informationen zum Feldtyp wird ein Standardfragebogen erstellt. So kann beispielsweise ein nominales Feld (Setfeld) als einzelne Frage dargestellt werden, wobei die Feldbeschreibung als Fragetext verwendet wird und für jeden definierten Wert ein gesondertes Kontrollkästchen vorhanden ist.

Metadaten zusammenführen. Hiermit können Sie angeben, ob die Metadaten die bestehenden Versionen überschreiben oder mit den bestehenden Metadaten zusammengeführt werden sollen. Wenn die Zusammenführungsoption ausgewählt wird, wird bei jeder Ausführung des Streams eine neue Version erstellt. Dadurch können die verschiedenen Versionen eines Fragebogens, der Änderungen unterzogen wird, dokumentiert werden. Jede Version lässt sich als Momentaufnahme der für die Sammlung eines bestimmten Falldatasets verwendeten Metadaten betrachten.

Systemvariablen aktivieren. Gibt an, ob Systemvariablen in die exportierte *.mdd*-Datei aufgenommen werden sollen. Dazu gehören Variablen wie *Respondent.Serial*, *Respondent.Origin* und *DataCollection.StartTime*.

Falldateneinstellungen. Gibt die IBM SPSS Statistics-Datendatei (*.sav*) an, in die die Falldaten exportiert werden sollen. Beachten Sie, dass hier alle Einschränkungen für Variablen- und Wertenamen gelten. So kann es beispielsweise erforderlich sein, auf die Registerkarte "Filter" zu wechseln und im Menü für Filteroptionen die Option "Umbenennen für IBM SPSS Statistics" zu verwenden, um ungültige Zeichen in Feldnamen zu korrigieren.

Importknoten für diese Daten generieren. Mit dieser Option lassen Sie automatisch einen IBM SPSS Data Collection-Quellenknoten generieren, mit dem die exportierte Datendatei eingelesen wird.

Mehrfachantwortsets. Etwaige im Stream definierte Mehrfachantwortsets bleiben beim Export der Datei automatisch erhalten. Mehrfachantwortsets können über jeden Knoten, der die Registerkarte "Filter" enthält, angezeigt und bearbeitet werden. Weitere Informationen finden Sie im Thema „Bearbeiten von Mehrfachantwortsets” auf Seite 131.

Analytic Server-Export

Mit dem Analytic Server-Export können Sie Daten aus Ihrer Analyse in eine vorhandene Analytic Server-Datenquelle schreiben. Dies können z. B. Textdateien in HDFS (Hadoop Distributed File System) oder eine Datenbank sein.

Normalerweise beginnt ein Stream mit einem Analytic Server-Exportknoten auch mit Analytic Server-Quellenknoten und er wird an Analytic Server übergeben und in HDFS ausgeführt. Alternativ kann ein Stream mit "lokalen" Datenquellen mit einem Analytic Server-Exportknoten enden, damit relativ kleine Datasets (nicht mehr als 100.000 Datensätze) für die Verwendung mit Analytic Server hochgeladen werden können.

Datenquelle. Wählen Sie eine Datenquelle aus, die die Daten enthält, die Sie verwenden wollen. Eine Datenquelle enthält die zu dieser Quelle zugehörigen Dateien und Metadaten. Klicken Sie auf **Auswählen**, um eine Liste der verfügbaren Datenquellen anzuzeigen. Weitere Informationen finden Sie im Thema „Analytic Server - Datenquelle auswählen” auf Seite 32.

Wenn Sie eine neue Datenquelle erstellen müssen oder eine vorhandene Datenquelle bearbeiten müssen, klicken Sie auf **Datenquelleneditor starten...**.

Modalwert. Wählen Sie **Anhängen** aus, um der vorhandenen Datenquelle Daten hinzuzufügen, oder **Überschreiben**, um den Inhalt der Datenquelle zu ersetzen.

Importknoten für diese Daten generieren. Es wird ein Quellenknoten für die Daten erzeugt, die in die angegebene Datenquelle exportiert wurden. Dieser Knoten wird dem Streamerstellungsbereich hinzugefügt.

IBM Cognos BI-Exportknoten

Mit dem IBM Cognos BI-Exportknoten können Sie Daten aus einem IBM SPSS Modeler-Stream im UTF-8 Format in Cognos BI exportieren. Auf diese Weise kann Cognos BI transformierte oder gescorte Daten aus IBM SPSS Modeler verwenden. Sie können mit Cognos BI Report Studio beispielsweise einen Bericht basierend auf den exportierten Daten erstellen, einschließlich Vorhersagen und Konfidenzwerten. Der Bericht könnte dann auf dem Cognos BI-Server gespeichert und an Cognos BI-Benutzer verteilt werden.

Hinweis: Sie können nur relationale Daten exportieren, keine OLAP-Daten.

Um Daten in Cognos BI zu exportieren, müssen Sie folgende Eingaben machen:

- Cognos-Verbindung - die Verbindung zum Cognos BI-Server
- ODBC-Verbindung - die Verbindung zum Cognos-Datenserver des Cognos BI-Servers

Innerhalb der Cognos-Verbindung geben Sie die zu verwendende Cognos-Datenquelle an. Diese Datenquelle muss dieselben Anmeldedaten verwenden wie die ODBC-Datenquelle.

Sie exportieren die tatsächlichen Streamdaten in den Daten-Server und die Paket-Metadaten in den Cognos BI-Server.

Wie bei allen Exportknoten können Sie auch die Registerkarte "Veröffentlichen" des Knoten-Dialogfelds verwenden, um den Stream mit IBM SPSS Modeler Solution Publisher zu veröffentlichen und bereitzustellen.

Cognos-Verbindung

Hier können Sie angeben, welche Verbindung zum Cognos BI-Server Sie für den Export verwenden möchten. Die Prozedur beinhaltet den Export der Metadaten in ein neues Paket auf dem Cognos BI-Server, während die Streamdaten in den Cognos-Daten-Server exportiert werden.

Verbindung. Klicken Sie auf die Schaltfläche **Bearbeiten**, um ein Dialogfeld anzuzeigen, in dem Sie die URL und andere Details für den Cognos BI-Server festlegen können, auf den die Daten exportiert werden sollen. Wenn Sie bereits bei einem Cognos BI-Server über IBM SPSS Modeler angemeldet sind, können Sie auch die Dateils der aktuellen Verbindung bearbeiten. Weitere Informationen finden Sie im Thema „Cognos-Verbindungen" auf Seite 39.

Datenquelle. Der Name der Cognos-Datenquelle (normalerweise eine Datenbank), in die Sie die Daten exportieren. Die Dropdown-Liste zeigt alle Cognos-Datenquellen an, auf die Sie über die aktuelle Verbindung zugreifen können. Klicken Sie zum Aktualisieren der Liste auf die Schaltfläche **Aktualisieren**.

Ordner. Der Pfad und Name des Ordners auf dem Cognos BI-Server, in dem das Export-Paket erstellt werden soll.

Paketname. Der Name des Pakets in einem angegebenen Ordner, der die exportierten Metadaten enthalten soll. Dabei muss es sich um neues Paket mit einem einzigen Abfragesubjekt handeln; Sie können nicht in vorhandene Pakete exportieren.

Modalwert. Legt fest, wie der Export durchgeführt werden soll:

- **Paket jetzt veröffenlichen.** (Standard) Führt den Exportvorgang aus, sobald Sie auf **Ausführen** klicken.
- **Aktionsscript exportieren.** Erstellt ein XML-Script, das Sie später ausführen können (z. B. mit Framework Manager), um den Export durchzuführen. Geben Sie den Pfad und Dateinamen für das Script in das Feld **Datei** ein, oder klicken Sie auf die Schaltfläche **Bearbeiten**, um Namen und Speicherort der Scriptdatei festzulegen.

Importknoten für diese Daten generieren. Es wird ein Quellenknoten für die Daten erzeugt, die in die angegebene Datenquelle und Tabelle exportiert wurden. Wenn Sie auf **Ausführen** klicken, wird dieser Knoten in den Streamerstellungsbereich aufgenommen.

ODBC-Verbindung

Hier geben Sie die Verbindung für den Cognos-Daten-Server (also die Datenbank) an, an den die Streamdaten exportiert werden sollen.

Hinweis: Stellen Sie sicher, dass die Datenquelle, die Sie hier angeben, auf dieselbe Datenquelle wie im Bereich **Cognos-Verbindungen** verweist. Außerdem müssen Sie sicherstellen, dass die Cognos-Datenquelle dieselben Anmeldedaten verwendet wie die ODBC-Datenquelle.

Datenquelle. Zeigt die ausgewählte Datenquelle. Geben Sie den Namen ein oder wählen Sie einen Eintrag in der Dropdown-Liste aus. Wird die gewünschte Datenbank nicht in der Liste aufgeführt, wählen Sie **Neue Datenbankverbindung hinzufügen** und wechseln Sie im Dialogfeld "Datenbankverbindungen" zu dieser Datenbank. Weitere Informationen finden Sie im Thema „Hinzufügen einer Datenbankverbindung" auf Seite 15.

Tabellenname. Geben Sie den Namen der Tabelle ein, an die die Daten gesendet werden sollen. Bei der Option **In Tabelle einfügen** können Sie eine vorhandene Tabelle in der Datenbank auswählen, indem Sie auf die Schaltfläche **Auswählen** klicken.

Tabellen erstellen. Mit dieser Option können Sie eine neue Datenbanktabelle anlegen oder eine vorhandene Datenbanktabelle überschreiben.

In Tabelle einfügen. Mit dieser Option fügen Sie die Daten als neue Zeilen in eine vorhandene Datenbanktabelle ein.

Tabellen zusammenführen. (Wenn verfügbar) Aktivieren Sie diese Option, um ausgewählte Datenbankspalten mit Werten aus entsprechenden Quellendatenfeldern zu aktualisieren. Wenn Sie diese Option auswählen, wird die Schaltfläche **Zusammenführen** aktiviert, die ein Dialogfeld öffnet, in dem Sie Quellendatenfelder zu Datenbankspalten zuordnen können.

Vorhandene Tabelle löschen. Wenn Sie eine neue Tabelle erstellen, lassen Sie mit dieser Option alle vorhandenen Tabellen löschen, die denselben Namen besitzen wie die neu zu erstellende Tabelle.

Vorhandene Zeilen löschen. Wenn Sie Daten in eine Tabelle einfügen, lassen Sie mit dieser Option vorhandene Zeilen vor dem Exportieren aus der Tabelle löschen.

Hinweis: Wenn eine oben genannten Optionen ausgewählt ist, wird eine **Überschreibungswarnung** ausgegeben, sobald Sie den Knoten ausführen. Sollen diese Warnungen unterdrückt werden, inaktivieren Sie im Dialogfeld "Benutzeroptionen" auf der Registerkarte "Benachrichtigungen" die Option **Warnen, wenn eine Datenbanktabelle durch einen Knoten überschrieben wird**.

Standardzeichenfolgegröße. Felder, die Sie als "Ohne Typ" in einem aufwärts liegenden Typknoten gekennzeichnet haben, werden als Zeichenfolgenfelder in die Datenbank geschrieben. Geben Sie die Größe der Zeichenfolgen an, die für Felder ohne Typ verwendet werden sollen.

Klicken Sie auf **Schema**, um ein Dialogfeld zu öffnen, in dem Sie verschiedene Exportoptionen festlegen können (für Datenbanken, die diese Funktion unterstützen), und geben Sie den Primärschlüssel für die Datenbankindizierung an. Weitere Informationen finden Sie im Thema „Schemaoptionen für den Datenbankexport" auf Seite 330.

Klicken Sie auf **Indizes**, um Optionen für die Indizierung der exportierten Tabelle anzugeben und damit die Datenbankleistung zu verbessern. Weitere Informationen finden Sie im Thema „Indexoptionen für den Datenbankexport" auf Seite 333.

Klicken Sie auf **Erweitert**, um Optionen für das Massenladen und die Datenbankübertragung festzulegen. Weitere Informationen finden Sie im Thema „Erweiterte Optionen für den Datenbankexport" auf Seite 335.

Tabellen- und Spaltennamen in Anführungszeichen. Wählen Sie die Optionen aus, die beim Senden der Anweisung `CREATE TABLE` an die Datenbank verwendet werden sollen. Enthält der Name von Tabellen und Spalten ein Leerzeichen oder ein Sonderzeichen, muss der Name in Anführungszeichen gesetzt werden.

- **Nach Bedarf.** Hiermit lassen Sie automatisch von Fall zu Fall durch IBM SPSS Modeler feststellen, ob Anführungszeichen erforderlich sind oder nicht.
- **Immer.** Die Tabellen- und Spaltennamen werden immer in Anführungszeichen eingeschlossen.
- **Nie.** Es werden keine Anführungszeichen verwendet.

Importknoten für diese Daten generieren. Es wird ein Quellenknoten für die Daten erzeugt, die in die angegebene Datenquelle und Tabelle exportiert wurden. Wenn Sie auf **Ausführen** klicken, wird dieser Knoten in den Streamerstellungsbereich aufgenommen.

IBM Cognos TM1-Exportknoten

Mit dem IBM Cognos TM1-Exportknoten können Sie Daten aus einem IBM SPSS Modeler-Stream in Cognos TM1 exportieren. Auf diese Weise kann Cognos BI transformierte oder gescorte Daten aus IBM SPSS Modeler verwenden.

Hinweis: Sie können nur Maße, keine Kontextdimensionsdaten exportieren. Alternativ können Sie dem Cube neue Elemente hinzufügen.

Um Daten in Cognos BI zu exportieren, sind folgende Angaben erforderlich:

- Verbindung zum Cognos TM1-Server
- Cube, in den die Daten exportiert werden
- Zuordnung der SPSS-Datennamen zu den entsprechenden TM1-Dimensionen und -Maßen

Wie bei allen Exportknoten können Sie auch die Registerkarte "Veröffentlichen" des Knoten-Dialogfelds verwenden, um den Stream mit IBM SPSS Modeler Solution Publisher zu veröffentlichen und bereitzustellen.

Verbinden mit einem Cognos TM1-Cube zum Exportieren von Daten

Der erste Schritt beim Exportieren von Daten in eine IBM Cognos TM1-Datenbank besteht darin, den relevanten TM1-PM-Hub sowie den zugehörigen Server und Cube auf der Registerkarte **Verbindung** des Dialogfelds "IBM Cognos TM1" auszuwählen.

Anmerkung: Sie müssen sicherstellen, dass die Datenstruktur der Ansicht, in die Sie Daten exportieren, der Datenstruktur der Ansicht entspricht, aus der die Daten importiert wurden. Die für den Export verwendete Ansicht kann sich auf einem anderen Cube als die für den Import verwendete Ansicht befinden, sofern beide Ansichten dieselbe Struktur haben.

PM-System. Geben Sie die URL des Hubs ein, der den TM1-Server enthält, zu dem Sie eine Verbindung herstellen wollen.

TM1-Server. Wenn Sie die Cognos-Hubverbindung eingerichtet haben, wählen Sie den Server aus, der die zu importierenden Daten enthält, und klicken Sie auf **Anmeldung**. Wenn Sie bisher noch keine Verbindung zu diesem Server hergestellt haben, werden Sie zur Eingabe des Benutzernamens und des Kennworts aufgefordert. Alternativ können Sie einen anderen Server auswählen.

Zu exportierenden TM1-Cube auswählen. Zeigt den Namen der Cubes im TM1-Server an, in die Sie Daten exportieren können. Doppelklicken Sie auf einen Cube, um die Ansichten in diesem Cube anzuzeigen.

Wählen Sie zur Auswahl der zu exportierenden Daten den Cube aus und klicken Sie auf den Rechtspfeil, um den Cube in das Feld **In Cube exportieren** zu verschieben. Wenn Sie den Cube ausgewählt haben, gleichen Sie die TM1-Felder mithilfe der Registerkarte **Zuordnung** mit den relevanten SPSS-Feldern ab.

Zuordnen von Cognos TM1-Daten für den Export

Nachdem Sie den TM1-PM-Hub sowie den zugehörigen Server und Cube auf der Registerkarte "Verbindung" des Dialogfelds "IBM Cognos TM1" ausgewählt haben, ermitteln Sie, welche SPSS-Felder den zugehörigen TM1-Feldern zugeordnet sind.

Wenn möglich werden die TM1-Felder automatisch den relevanten SPSS-Feldern zugeordnet.

Anmerkung: Sie müssen sicherstellen, dass die Datenstruktur der Ansicht, in die Sie Daten exportieren, der Datenstruktur der Ansicht entspricht, aus der die Daten importiert wurden. Die für den Export verwendete Ansicht kann sich auf einem anderen Cube als die für den Import verwendete Ansicht befinden, sofern beide Ansichten dieselbe Struktur haben.

Felder. Listet den in der SPSS-Datendatei angegebenen Datenfeldnamen für die Daten auf, die für den Export verfügbar sind. Verwenden Sie bei Bedarf die Schaltflächen am Ende der Liste, um die angezeigten Felder zu ändern. Sie können z. B. alle Felder, nur stetige Felder oder kategoriale Felder anzeigen.

TM1-Dimensionen. Zeigt den Cube, der auf der Registerkarte **Verbindung** ausgewählt ist, zusammen mit seinen regulären Dimensionen und Maßdimensionen an. Wenn diese Angabe nicht automatisch erfolgt, wählen Sie den Namen des TM1-Elements aus, das dem SPSS-Datenfeld zugeordnet werden soll.

SAS-Exportknoten

Hinweis: Diese Funktion ist in SPSS Modeler Professional und SPSS Modeler Premium verfügbar.

Mit dem SAS-Exportknoten können Sie Daten im SAS-Format schreiben, die dann in SAS oder in SAS-kompatible Programme eingelesen werden können. Sie können Daten in drei SAS-Dateiformaten exportieren: SAS für Windows/OS2, SAS für UNIX und SAS Version 7/8.

Registerkarte "Exportieren" beim SAS-Exportknoten

Datei exportieren. Geben Sie den Namen der Datei an. Geben Sie einen Dateinamen an oder klicken Sie auf die Feldauswahlschaltfläche und wechseln Sie zum Pfad der gewünschten Datei.

Exportieren. Legen Sie das Exportdateiformat fest. Die folgenden Optionen stehen zur Auswahl: **SAS für Windows/OS2**, **SAS für UNIX** sowie **SAS Version 7/8**.

Feldnamen exportieren. Wählen Sie die Optionen zum Exportieren der Feldnamen und Beschriftungen aus IBM SPSS Modeler, die in SAS genutzt werden sollen.

- **Namen und Variablenbeschriftungen.** Hiermit werden sowohl die Feldnamen als auch die Feldbeschriftungen aus IBM SPSS Modeler exportiert. Die Namen werden als SAS-Variablennamen exportiert, die Beschriftungen entsprechend als SAS-Variablenbeschriftungen.
- **Namen als Variablenbeschriftungen.** Mit dieser Einstellung werden die IBM SPSS Modeler-Feldnamen in SAS als Variablenbeschriftungen verwendet. Bei IBM SPSS Modeler können verschiedene Zeichen in den Feldnamen verwendet werden, die bei SAS-Variablennamen nicht gültig sind. Um die mögliche Bildung ungültiger SAS-Namen zu vermeiden, wählen Sie stattdessen die Option **Namen und Variablenbeschriftungen**.

Importknoten für diese Daten generieren. Mit dieser Option lassen Sie automatisch einen SAS-Quellenknoten erzeugen, mit dem die exportierte Datendatei eingelesen wird. Weitere Informationen finden Sie im Thema „SAS-Quellenknoten" auf Seite 41.

Excel-Exportknoten

Der Excel-Exportknoten gibt Daten im Microsoft Excel-Format (*.xls*) aus. Optional können Sie auswählen, dass bei der Ausführung des Knotens Excel automatisch gestartet und die exportierte Datei geöffnet werden soll.

Registerkarte "Exportieren" beim Excel-Knoten

Dateiname: Geben Sie einen Dateinamen an oder klicken Sie auf die Feldauswahlschaltfläche und wechseln Sie zum Pfad der gewünschten Datei. Der Standarddateiname lautet *excelexp.xls*.

Dateityp. Wählen Sie den Excel-Dateityp, den Sie exportieren möchten.

Neue Datei erstellen. Erstellt eine neue Excel-Datei.

In vorhandene Datei einfügen. Der Inhalt wird ersetzt, beginnend in der Zelle, die durch das Feld **Start in Zelle** angegeben ist. Andere Zellen im Arbeitsblatt behalten ihren ursprünglichen Inhalt.

Feldnamen einschließen. Gibt an, ob Feldnamen in die erste Zeile des Arbeitsblattes eingefügt werden sollen.

Beginn in Zelle. Die für den ersten Exportdatensatz (bzw. den ersten Feldnamen, falls **Feldnamen einschließen** aktiviert ist) verwendete Zellenposition. Daten werden ab dieser Anfangszelle nach rechts und unten gefüllt.

Arbeitsblatt auswählen. Legt das Arbeitsblatt fest, an das Sie die Daten exportieren möchten. Sie können das Arbeitsblatt nach Index oder nach Name identifizieren:

- **Nach Index.** Wenn Sie eine neue Datei anlegen, bestimmen Sie eine Zahl zwischen 0 und 9, um das Arbeitsblatt zu identifizieren, das Sie exportieren möchten, beginnend mit 0 für das erste Arbeitsblat, 1 für das zweite usw. Werte von 10 und darüber können Sie nur verwenden, wenn an dieser Position bereits ein Arbeitsblatt vorhanden ist.
- **Nach Name.** Wenn Sie eine neue Datei anlegen, bestimmen Sie den Namen, der für das Arbeitsblatt verwendet wird. Beim Einfügen in eine bestehende Datei werden die Daten in dieses Arbeitsblatt eingefügt, falls vorhanden, andernfalls wird ein neues Arbeitsblatt mit diesem Namen erstellt.

Excel starten. Gibt an, ob Excel bei der Ausführung des Knotens automatisch für die exportierte Datei gestartet werden soll. Beachten Sie, dass bei der Ausführung im verteilten Modus für IBM SPSS Modeler Server die Ausgabe im Dateisystem des Servers gespeichert und Excel auf dem Client mit einer Kopie der exportierten Datei gestartet wird.

Importknoten für diese Daten generieren. Mit dieser Option lassen Sie automatisch einen Excel-Quellenknoten erzeugen, mit dem die exportierte Datendatei eingelesen wird. Weitere Informationen finden Sie im Thema „Excel-Quellenknoten" auf Seite 42.

XML-Exportknoten

Mit dem XML-Exportknoten können Sie unter Verwendung der UTF-8-Codierung Daten im XML-Format ausgeben. Optional können Sie einen XML-Quellenknoten erstellen, um die exportierten Daten wieder in der Stream einzulesen.

XML-Exportdatei. Der vollständige Pfad und Dateiname der XML-Datei, an die Sie die Daten exportieren möchten.

XML-Schema verwenden. Aktivieren Sie dieses Kontrollkästchen, wenn Sie ein Schema oder DTD verwenden möchten, um die Struktur der exportierten Daten zu kontrollieren. Dadurch wird die unten beschriebene Schaltfläche **Zuordnen** aktiviert.

Wenn Sie kein Schema oder DTD verwenden, wird die folgende Standardstruktur für die exportierten Daten verwendet:

```
<records>
     <record>
          <Feldname1>Wert</Feldname1>
          <Feldname2>Wert</Feldname2>
      :
          <FeldnameN>Wert</FeldnameN>
     </record>
     <record>
      :
      :
     </record>
      :
      :
</records>
```

Leerfelder in einem Feldnamen werden durch Unterstriche ersetzt, so wird zum Beispiel "Mein Feld" zu `<Mein_Feld>`.

Zuordnen. Falls Sie ein XML-Schema verwenden, öffnet diese Schaltfläche ein Dialogfeld, in dem Sie angeben können, welcher Teil der XML-Struktur für den Beginn jedes neuen Datensatzes verwendet werden soll. Weitere Informationen finden Sie im Thema „XML-Zuordnungsdatensätze - Optionen".

Zugeordnete Felder. Zeigt die Anzahl der Felder an, die zugeordnet wurden

Importknoten für diese Daten generieren. Mit dieser Option lassen Sie automatisch einen XML-Quellenknoten erzeugen, mit dem die exportierte Datendatei wieder in den Stream eingelesen wird. Weitere Informationen finden Sie im Thema „XML-Quellenknoten" auf Seite 43.

Schreiben von XML-Daten

Wenn ein XML-Element angegeben wird, wird der Feldwert im Elementtag platziert:

```
<Element>Wert</Element>
```

Wenn ein Attribut zugeordnet wird, wird der Feldwert als Wert für das Attribut platziert:

```
<Element Attribut="Wert">
```

Wenn ein Feld einem Element oberhalb des Elements `<Datensätze>` zugeordnet wird, wird das Feld nur einmal beschrieben und fungiert als Konstante für alle Datensätze. Der Wert für dieses Element kommt aus dem ersten Datensatz.

Wenn ein Nullwert geschrieben werden muss, geschieht dies durch Angeben von leerem Inhalt. Bei Elementen ist das:

```
<Element></Element>
```

Bei Attributen ist es:

```
<Element Attribut="">
```

XML-Zuordnungsdatensätze - Optionen

Auf der Registerkarte "Datensätze" können Sie angeben, welcher Teil der XML-Struktur für den Beginn jedes neuen Datensatzes verwendet werden soll. Um korrekte Zuordnungen auf ein Schema durchführen zu können, müssen Sie das Datensatztrennzeichen festlegen.

XML-Struktur. Ein hierarchischer Baum zeigt die Sturktur des XML-Schemas, das auf dem vorangegangenen Bildschirm festgelegt wurde.

Datensätze (XPath-Ausdruck). Zum Festlegen des Datensatztrennzeichens wählen Sie ein Element in der XML-Struktur aus und klicken Sie auf die Schaltfläche mit dem Rechtspfeil. Jedes Mal, wenn dieses Element in den Quellendaten gefunden wird, wird in der Ausgabedatei ein neuer Datensatz erstellt.

Hinweis: Wenn Sie das Stammelement in der XML-Struktur auswählen, kann nur ein einziger Datensatz geschrieben werden, während alle anderen Datensätze übergangen werden.

XML-Zuordnungsfelder - Optionen

Die Registerkarte "Felder" dient zum Zuordnen von Feldern im Dataset zu Elementen oder Attributen in der XML-Struktur, wenn eine Schemadatei verwendet wird.

Feldnamen, die einem Element- oder Attributnamen entsprechen, werden automatisch zugeordnet, so lange der Element- oder Attributname eindeutig ist. Wenn es also sowohl ein Element als auch ein Attribut mit dem Namen `Feld1` gibt, kann keine automatische Zuordnung erfolgen. Wenn es nur ein Objekt in der Struktur mit der Bezeichnung `Feld1` gibt, wird ein Feld mit diesem Namen in dem Stream automatisch zugeordnet.

Felder. Die Liste der Felder im Modell. Wählen Sie eines oder mehrere Felder als Quellenteil der Zuordnung aus. Sie können die Schaltflächen am Ende der Liste verwenden, um alle Felder oder alle Felder mit einem bestimmten Messniveau auszuwählen.

XML-Struktur. Wählen Sie ein Element in der XML-Struktur als Zuordnungsziel aus. Klicken Sie auf "Zuordnen", um die Zuordnung zu erstellen. Anschließend wird die Zuordnung angezeigt. Die Anzahl der so zugeordneten Felder wird unterhalt dieser Liste angezeigt.

Um eine Zuordnung aufzuheben, wählen Sie das Objekt in der XML-Struktur aus und klicken Sie auf **Zuordnung aufheben**.

Attribute anzeigen. Zeigt die Attribute der XML-Elemente in der XML-Struktur an oder blendet diese aus, sofern vorhanden.

XML-Zuordnungsvorschau

Klicken Sie auf der Registerkarte "Vorschau" auf **Aktualisieren**, um eine Vorschau der XML-Datei zu sehen, die geschrieben wird.

Falls die Zuordnung nicht korrekt ist, kehren Sie zur Registerkarte "Datensätze" oder "Felder" zurück, um die Fehler zu korrigieren, und klicken erneut auf **Aktualisieren**, um sich das Ergebnis anzusehen.

Kapitel 8. IBM SPSS Statistics-Knoten

Überblick über IBM SPSS Statistics-Knoten

Zur Ergänzung von IBM SPSS Modeler und seinen Data-Mining-Funktionen bietet Ihnen IBM SPSS Statistics die Möglichkeit, weiterführende statistische Analysen durchzuführen und Daten zu verwalten.

Wenn Sie eine kompatible, lizenzierte Kopie von IBM SPSS Statistics installiert haben, können Sie von IBM SPSS Modeler eine Verbindung aufbauen und komplexe, aus mehreren Schritten bestehende Datenbearbeitungen und Analysen ausführen, die andernfalls von IBM SPSS Modeler nicht unterstützt werden. Für den erfahrenen Benutzer gibt es auch die Option, die Analysen mithilfe von Befehlssyntax weiter anzupassen. In den Versionshinweisen finden Sie Informationen zur Kompatibilität von Versionen.

Wenn verfügbar, werden die IBM SPSS Statistics-Knoten auf einem eigenen Teil der Knotenpalette angezeigt.

Hinweis: Es wird empfohlen, dass Sie Ihre Daten in einem Typenknoten instanziieren, bevor Sie die Transformations-, Modell- oder Ausgabeknoten von IBM SPSS Statistics verwenden. Dies ist auch eine Voraussetzung für die Verwendung des Syntaxbefehls AUTORECODE.

Die Palette "IBM SPSS Statistics" enthält die folgenden Knoten:

Der Statistikdateiknoten liest Daten aus dem Dateiformat *.sav* oder *.zsav* ein, das von IBM SPSS Statistics verwendet wird, sowie in IBM SPSS Modeler gespeicherte Cachedateien, die ebenfalls dasselbe Format verwenden.

Der Statistics-Transformationsknoten führt eine Auswahl von IBM SPSS Statistics-Syntaxbefehlen an Datenquellen in IBM SPSS Modeler aus. Für diesen Knoten ist eine lizenzierte Kopie von IBM SPSS Statistics erforderlich.

Mithilfe des Statistics-Modellknotens können Sie Ihre Daten analysieren und bearbeiten, indem Sie IBM SPSS Statistics-Prozeduren ausführen, die PMML erzeugen. Für diesen Knoten ist eine lizenzierte Kopie von IBM SPSS Statistics erforderlich.

Mit dem Statistics-Ausgabeknoten können Sie eineIBM SPSS Statistics-Prozedur aufrufen, um Ihre IBM SPSS Modeler-Daten zu analysieren. Es stehen zahlreiche IBM SPSS Statistics-Analyseprozeduren zur Verfügung. Für diesen Knoten ist eine lizenzierte Kopie von IBM SPSS Statistics erforderlich.

Der Statistikexportknoten gibt Daten im IBM SPSS Statistics-Format *.sav* oder *.zsav* aus. Die *.sav*- oder *.zsav*-Dateien können von IBM SPSS Statistics Base und anderen Produkten gelesen werden. Dieses Format wird auch für Cache-Dateien in IBM SPSS Modeler verwendet.

Hinweis: Wenn Ihre Kopie von SPSS Statistics nur für einen einzigen Benutzer lizenziert ist und Sie einen Stream mit mindestens zwei Verzweigungen ausführen, von denen jede einen SPSS Statistics-Knoten ent-

hält, erhalten Sie möglicherweise einen Lizenzierungsfehler. Dieser Fall tritt dann ein, wenn die SPSS Statistics-Sitzung für eine Verzweigung noch nicht beendet wurde, bevor die Sitzung für eine andere Verzweigung zu starten versucht. Überarbeiten Sie den Stream nach Möglichkeit so, dass nicht mehrere Verzweigungen mit SPSS Statistics-Knoten parallel ausgeführt werden.

Statistikdateiknoten

Mit dem Statistikdateiknoten können Sie Daten direkt aus einer gespeicherten IBM SPSS Statistics-Datei (*.sav* oder *.zsav*) lesen. Dieses Format ersetzt nun die Cachedatei aus früheren Versionen von IBM SPSS Modeler. Wenn Sie eine gespeicherte Cachedatei importieren möchten, verwenden Sie am besten den IBM SPSS Statistics-Dateiknoten.

Datei importieren. Geben Sie den Namen der Datei an. Zur Auswahl einer Datei können Sie einen Dateinamen eingeben oder auf die Schaltfläche mit den Auslassungspunkten (**...**) klicken. Der Dateipfad wird angezeigt, sobald Sie eine Datei ausgewählt haben.

Datei ist kennwortverschlüsselt. Wählen Sie dieses Feld aus, wenn Sie wissen, dass die Datei kennwortgeschützt ist. Sie werden aufgefordert, das **Kennwort** einzugeben. Wenn die Datei kennwortgeschützt ist und Sie das Kennwort nicht eingeben, wird bei dem Versuch, zu einer anderen Registerkarte zu wechseln, die Daten zu aktualisieren, eine Vorschau des Knoteninhalts anzuzeigen oder einen Stream auszuführen, der den Knoten enthält, ein Warnhinweis angezeigt.

Anmerkung: Kennwortgeschützte Dateien können nur von IBM SPSS Modeler Version 16 oder höher geöffnet werden.

Variablennamen. Wählen Sie eine Methode zur Behandlung von Variablennamen und Beschriftungen beim Importieren aus einer Datei im IBM SPSS Statistics-Format *.sav* oder *.zsav* aus. Metadaten, die Sie hier einschließen, bleiben während Ihrer Arbeit in IBM SPSS Modeler erhalten und können zur Verwendung in IBM SPSS Statistics wieder exportiert werden.

- **Namen und Beschriftungen lesen.** Wählen Sie diese Option aus, wenn sowohl Variablennamen als auch -beschriftungen in IBM SPSS Modeler eingelesen werden sollen. Standardmäßig ist diese Option ausgewählt und Variablennamen werden im Typknoten angezeigt. Beschriftungen können je nach den im Dialogfeld "Streameigenschaften" angegebenen Optionen in Diagrammen, Modellbrowsern und anderen Ausgabearten angezeigt werden. Standardmäßig ist die Anzeige von Beschriftungen in der Ausgabe inaktiviert.
- **Beschriftungen als Namen lesen.** Wählen Sie diese Option aus, um statt der kurzen Feldnamen die beschreibenden Variablenbeschriftungen aus der Datei im IBM SPSS Statistics-Format *.sav* oder *.zsav* zu lesen und diese Beschriftungen als Variablennamen in IBM SPSS Modeler zu verwenden.

Werte. Wählen Sie eine Methode zur Behandlung von Werten und Beschriftungen beim Importieren aus einer Datei im IBM SPSS Statistics-Format *.sav* oder *.zsav* aus. Metadaten, die Sie hier einschließen, bleiben während Ihrer Arbeit in IBM SPSS Modeler erhalten und können zur Verwendung in IBM SPSS Statistics wieder exportiert werden.

- **Daten und Beschriftungen lesen.** Wählen Sie diese Option aus, um sowohl die tatsächlichen Werte als auch die Wertbeschriftungen in IBM SPSS Modeler einzulesen. Standardmäßig ist diese Option ausgewählt und die Werte werden im Typknoten angezeigt. Wertbeschriftungen können je nach den im Dialogfeld "Streameigenschaften" angegebenen Optionen in Expression Builder, Diagrammen, Modellbrowsern und anderen Ausgabearten angezeigt werden.
- **Beschriftungen als Daten lesen.** Wählen Sie diese Option aus, wenn Sie statt der numerischen oder symbolischen Codes, mit denen die Werte dargestellt werden, die Wertbeschriftungen aus der Datei im Format *.sav* oder *.zsav* verwenden möchten. Bei Auswahl dieser Option z. B. für Daten mit dem Feld "Geschlecht", dessen Werte 1 und 2 für *männlich* und *weiblich* stehen, wird das Feld in eine Zeichenfolge konvertiert und *männlich* und *weiblich* werden als tatsächliche Werte importiert.

  Vor Auswahl dieser Option müssen Sie Ihre IBM SPSS Statistics-Daten auf fehlende Werte prüfen. Wenn ein numerisches Feld beispielsweise Beschriftungen nur für fehlende Werte verwendet (0 = *Keine Antwort*, 99 = *Unbekannt*), werden bei Auswahl der obigen Option nur die Wertbeschriftungen *Keine*

Antwort und *Unbekannt* importiert und das Feld wird in eine Zeichenfolge konvertiert. In diesem Fall sollten Sie die Werte selbst importieren und fehlende Werte in einem Typknoten festlegen.

Speichertyp anhand Feldformatinformationen bestimmen. Wenn dieses Kontrollkästchen abgewählt ist, werden Feldwerte, die in der *SAV*-Datei als ganze Zahlen formatiert sind (d. h. Felder, die in der Variablenansicht in IBM SPSS Statistics als F*n*.0 angegeben sind), mit dem Speichertyp "Ganze Zahl" importiert. Alle übrigen Feldwerte mit Ausnahme von Zeichenfolgen werden als reelle Zahlen importiert.

Wenn das Kontrollkästchen ausgewählt ist (Standard), werden alle Feldwerte außer Zeichenfolgen als reelle Zahlen importiert, unabhängig davon, ob sie in der *SAV*-Datei als Ganzzahlen formatiert sind.

Mehrfachantwortsets. Etwaige in der IBM SPSS Statistics-Datei definierte Mehrfachantwortsets bleiben beim Import der Datei automatisch erhalten. Mehrfachantwortsets können über jeden Knoten, der die Registerkarte "Filter" enthält, angezeigt und bearbeitet werden. Weitere Informationen finden Sie im Thema „Bearbeiten von Mehrfachantwortsets” auf Seite 131.

Statistics-Transformationsknoten

Mit dem Statistics-Transformationsknoten können Sie Datentransformationen mithilfe der IBM SPSS Statistics-Befehlssyntax durchführen. Dadurch kann eine Reihe von Transformationen durchgeführt werden, die von IBM SPSS Modeler nicht unterstützt werden, und die Automatisierung komplexer, aus mehreren Schritten bestehenden Transformationen ist möglich, einschließlich der Erstellung von Feldern aus einem einzelnen Knoten. Er ähnelt dem Statistics-Ausgabeknoten, mit der Ausnahme, dass die Daten zur weiteren Analyse an IBM SPSS Modeler ausgegeben werden, während die Daten beim Ausgabeknoten als angeforderte Ausgabeobjekte ausgegeben werden, beispielsweise als Diagramme oder Tabellen.

Um diesen Knoten verwenden zu können, muss eine kompatible Version von IBM SPSS Statistics auf Ihrem Computer installiert und lizenziert sein. Weitere Informationen finden Sie im Thema „IBM SPSS Statistics-Hilfsanwendungen” auf Seite 324. In den Releaseinformationen finden Sie Informationen zur Kompatibilität.

Falls erforderlich, können Sie mithilfe der Registerkarte "Filter" Felder filtern oder umbenennen, sodass sie den IBM SPSS Statistics-Benennungsstandards entsprechen. Weitere Informationen finden Sie im Thema „Umbenennen oder Filtern von Feldern für IBM SPSS Statistics” auf Seite 362.

Syntaxreferenz. Einzelheiten zu bestimmten IBM SPSS Statistics-Prozeduren finden Sie in der IBM SPSS Statistics-Befehlssyntaxreferenz, die in Ihrer Kopie der IBM SPSS Statistics-Software enthalten ist. Um das Handbuch anzuzeigen, wählen Sie auf der Registerkarte "Syntax" die Option **Syntaxeditor** aus und klicken Sie auf die Schaltfläche "IBM SPSS Statistics-Syntaxhilfe starten".

Hinweis: Dieser Knoten unterstützt nicht die gesamte IBM SPSS Statistics-Syntax. Weitere Informationen finden Sie im Thema „Zulässige Syntax” auf Seite 356.

Statistics-Transformationsknoten - Registerkarte "Syntax"

IBM SPSS Statistics-Dialogoption

Wenn Sie nicht mit der IBM SPSS Statistics-Syntax für eine Prozedur vertraut sind, ist dies die einfachste Methode zur Erstellung von Syntax in IBM SPSS Modeler: Wählen Sie die Option **IBM SPSS Statistics-Dialog**, wählen Sie das Dialogfeld für die Prozedur aus, füllen Sie das Dialogfeld aus und klicken Sie auf "OK". Dadurch wird die Syntax auf der Registerkarte "Syntax" des IBM SPSS Statistics-Knotens abgelegt, den Sie in IBM SPSS Modeler verwenden. Anschließend können Sie den Stream ausführen, um die Ausgabe aus der Prozedur zu erhalten.

IBM SPSS Statistics-Option "Syntaxeditor"

Überprüfen. Nachdem Sie Ihre Syntaxbefehle im oberen Bereich des Dialogfelds eingegeben haben, können Sie mit dieser Schaltfläche Ihre Einträge überprüfen. Etwaige falsche Syntax wird im unteren Teil des Dialogfelds angegeben.

Um sicherzustellen, dass die Überprüfung nicht zu lange dauert, wird bei der Syntaxvalidierung eine Überprüfung anhand einer repräsentativen Stichprobe Ihrer Daten durchgeführt, um sicherzustellen, dass die Einträge gültig sind. Auf eine Überprüfung anhand des gesamten Datasets wird verzichtet.

Zulässige Syntax

Wenn Sie viel alte Syntax aus IBM SPSS Statistics verwenden oder mit den Datenvorbereitungsfunktionen von IBM SPSS Statistics vertraut sind, können Sie viele Ihrer bestehenden Transformationen mithilfe des Statistics-Transformationsknotens ausführen. Grob gesagt, ermöglicht der Knoten die Transformation von Daten auf vorhersehbare Weise, beispielsweise durch die Ausführung von Schleifenbefehlen oder durch Ändern, Hinzufügen, Sortieren, Filtern oder Auswählen von Daten.

Hier einige Beispiele für Befehle, die ausgeführt werden können:

- Berechnung von Zufallszahlen gemäß einer Binomialverteilung:

  ```
  COMPUTE newvar = RV.BINOM(10000,0.1)
  ```

- Umcodieren einer Variablen in eine neue Variable:

  ```
  RECODE Age (Lowest thru 30=1)  (30 thru 50=2)  (50 thru Highest=3)  INTO  AgeRecoded
  ```

- Ersetzen fehlender Werte:

```
RMV Age_1=SMEAN(Age)
```

Die vom Statistics-Transformationsknoten unterstützte IBM SPSS Statistics-Syntax ist nachfolgend aufgelistet.

Befehlsname

```
ADD VALUE LABELS
APPLY DICTIONARY
AUTORECODE
BREAK
CD
CLEAR MODEL PROGRAMS
CLEAR TIME PROGRAM
CLEAR TRANSFORMATIONS
COMPUTE
COUNT
CREATE
DATE
DEFINE-!ENDDEFINE
DELETE VARIABLES
DO IF
DO REPEAT
ELSE
ELSE IF
END CASE
END FILE
END IF
```

Befehlsname

END INPUT PROGRAM

END LOOP

END REPEAT

EXECUTE

FILE HANDLE

FILE LABEL

FILE TYPE-END FILE TYPE

FILTER

FORMATS

IF

INCLUDE

INPUT PROGRAM-END INPUT PROGRAM

INSERT

LEAVE

LOOP-END LOOP

MATRIX-END MATRIX

MISSING VALUES

N OF CASES

NUMERIC

PERMISSIONS

PRESERVE

RANK

RECODE

RENAME VARIABLES

RESTORE

RMV

SAMPLE

SELECT IF

SET

SORT CASES

STRING

SUBTITLE

TEMPORARY

TITLE

UPDATE

V2C

VALIDATEDATA

VALUE LABELS

VARIABLE ATTRIBUTE

VARSTOCASES

VECTOR

Statistics-Modellknoten

Mithilfe des Statistics-Modellknotens können Sie Ihre Daten analysieren und bearbeiten, indem Sie IBM SPSS Statistics-Prozeduren ausführen, die PMML erzeugen. Die Modellnuggets, die Sie erzeugen, können dann wie üblich in IBM SPSS Modeler-Streams zum Scoring usw. verwendet werden.

Um diesen Knoten verwenden zu können, muss eine kompatible Version von IBM SPSS Statistics auf Ihrem Computer installiert und lizenziert sein. Weitere Informationen finden Sie im Thema „IBM SPSS Statistics-Hilfsanwendungen" auf Seite 324. In den Releaseinformationen finden Sie Informationen zur Kompatibilität.

Die verfügbaren IBM SPSS Statistics-Analyseprozeduren hängen von Ihrer Lizenz ab.

Statistics-Modellknoten - Registerkarte "Modell"

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Dialogfeld auswählen. Klicken Sie, um eine Liste verfügbarer IBM SPSS Statistics-Prozeduren anzuzeigen, die Sie auswählen und ausführen können. In der Liste werden nur die Prozeduren aufgeführt, die PMML produzieren und für die Sie eine Lizenz besitzen. Nicht enthalten sind benutzerdefinierte Prozeduren.

1. Klicken Sie auf die gewünschte Prozedur. Das entsprechende IBM SPSS Statistics-Dialogfeld wird geöffnet.
2. Geben Sie im IBM SPSS Statistics-Dialogfeld die Details für die Prozedur ein.
3. Klicken Sie auf **OK**, um in den Statistics-Modellknoten zurückzukehren; die IBM SPSS Statistics-Syntax wird auf der Registerkarte "Modell" angezeigt.
4. Um zu einem beliebigen Zeitpunkt in das IBM SPSS Statistics-Dialogfeld zurückzukehren, z. B. um Ihre Abfrage zu ändern, klicken Sie auf die Schaltfläche zum Anzeigen des IBM SPSS Statistics-Dialogfelds rechts neben der Schaltfläche zur Prozedurauswahl.

Statistics-Modellknoten - Modellnugget-Übersicht

Wenn Sie den Statistics-Modellknoten ausführen, führt dieser die zugehörige IBM SPSS Statistics-Prozedur aus und erstellt ein Modellnugget, das Sie zum Scoring in IBM SPSS Modeler-Streams verwenden können.

Auf der Registerkarte "Übersicht" für das Modellnugget werden Informationen über die Felder, die Aufbaueinstellungen und die Modellschätzung angezeigt. Die Ergebnisse werden in einer Baumansicht dargestellt, die durch Klicken auf bestimmte Elemente erweitert bzw. reduziert werden kann.

Die Schaltfläche **Modell anzeigen** zeigt die Ergebnisse in einer modifizierten Variante des IBM SPSS Statistics-Ausgabeviewer. Weitere Informationen zu diesem Viewer finden Sie in der IBM SPSS Statistics-Dokumentation.

Das Menü "Datei" enthält die üblichen Befehle zum Exportieren und Drucken. Weitere Informationen finden Sie im Thema „Anzeigen der Ausgabe" auf Seite 283.

Statistics-Ausgabeknoten

Mit dem Statistics-Ausgabeknoten können Sie eine IBM SPSS Statistics-Prozedur aufrufen, um Ihre IBM SPSS Modeler-Daten zu analysieren. Lassen Sie die Ergebnisse in einem Browserfenster anzeigen oder speichern Sie sie im IBM SPSS Statistics-Ausgabedateiformat. In IBM SPSS Statistics stehen zahlreiche IBM SPSS Modeler-Analyseprozeduren zur Verfügung.

Um diesen Knoten verwenden zu können, muss eine kompatible Version von IBM SPSS Statistics auf Ihrem Computer installiert und lizenziert sein. Weitere Informationen finden Sie im Thema „IBM SPSS Statistics-Hilfsanwendungen” auf Seite 324. In den Releaseinformationen finden Sie Informationen zur Kompatibilität.

Falls erforderlich, können Sie mithilfe der Registerkarte "Filter" Felder filtern oder umbenennen, sodass sie den IBM SPSS Statistics-Benennungsstandards entsprechen. Weitere Informationen finden Sie im Thema „Umbenennen oder Filtern von Feldern für IBM SPSS Statistics” auf Seite 362.

Syntaxreferenz. Einzelheiten zu bestimmten IBM SPSS Statistics-Prozeduren finden Sie in der IBM SPSS Statistics-Befehlssyntaxreferenz, die in Ihrer Kopie der IBM SPSS Statistics-Software enthalten ist. Um das Handbuch anzuzeigen, wählen Sie auf der Registerkarte "Syntax" die Option **Syntaxeditor** aus und klicken Sie auf die Schaltfläche "IBM SPSS Statistics-Syntaxhilfe starten".

Statistics-Ausgabeknoten - Registerkarte "Syntax"

Mit dieser Registerkarte können Sie die Syntax für die IBM SPSS Statistics-Prozedur erstellen, mit der Sie Ihre Daten analysieren möchten. Die Syntax besteht aus zwei Teilen: einer **Anweisung** und den zugehörigen **Optionen**. Die Anweisung bezeichnet die auszuführende Analyse oder Option sowie die zu verwendenden Felder. In den Optionen sind alle anderen Angaben festgelegt, z. B. die anzuzeigende Statistik oder die zu speichernden abgeleiteten Felder.

IBM SPSS Statistics-Dialogoption

Wenn Sie nicht mit der IBM SPSS Statistics-Syntax für eine Prozedur vertraut sind, ist dies die einfachste Methode zur Erstellung von Syntax in IBM SPSS Modeler: Wählen Sie die Option **IBM SPSS Statistics-Dialog**, wählen Sie das Dialogfeld für die Prozedur aus, füllen Sie das Dialogfeld aus und klicken Sie auf "OK". Dadurch wird die Syntax auf der Registerkarte "Syntax" des IBM SPSS Statistics-Knotens abgelegt, den Sie in IBM SPSS Modeler verwenden. Anschließend können Sie den Stream ausführen, um die Ausgabe aus der Prozedur zu erhalten.

Sie können optional einen Statistikdatei-Quellenknoten zum Importieren der resultierenden Daten generieren. Dies ist beispielsweise dann nützlich, wenn eine Prozedur zusätzlich zur Anzeige der Ausgabe Felder, beispielsweise für Scores, in das aktive Dataset schreibt.

So erstellen Sie die Syntax:

1. Klicken Sie auf die Schaltfläche **Dialogfeld auswählen**.
2. Wählen Sie eine der Optionen aus:
   - **Analysieren.** Listet den Inhalt des IBM SPSS Statistics-Analysemenüs auf; wählen Sie die Prozedur aus, die Sie verwenden möchten.
   - **Sonstige.** Wenn diese Option angezeigt wird, werden dort Dialogfelder aufgelistet, die Sie mit dem Custom Dialog Builder in IBM SPSS Statistics erstellt haben, sowie andere IBM SPSS Statistics-Dialogfelder, die nicht im Analysemenü erscheinen und für die Sie eine Lizenz besitzen. Wenn keine Dialogfelder zur Verfügung stehen, wird diese Option nicht angezeigt.

Hinweis: Die Dialogfelder zur automatischen Datenaufbereitung werden nicht angezeigt.

Bei einem benutzerdefinierten IBM SPSS Statistics-Dialogfeld, das neue Felder erstellt, können diese Felder nicht in IBM SPSS Modeler verwendet werden, da es sich bei dem Statistics-Ausgabeknoten um einen Endknoten handelt.

Optional können Sie das Kontrollkästchen **Importknoten für resultierende Daten generieren** aktivieren, um einen Statistikdatei-Quellenknoten zu erstellen, mit dem die resultierenden Daten in einen anderen Stream importiert werden können. Der Knoten wird auf dem Bildschirmerstellungsbereich abgelegt, wobei die in der *SAV*-Datei enthaltenen Daten im Feld **Datei** angegeben werden (Standardspeicherort ist das Installationsverzeichnis von IBM SPSS Modeler).

Option "Syntaxeditor"

Gehen Sie wie folgt vor, um Syntax zu speichern, die für eine häufig verwendete Prozedur erstellt wurde:

1. Klicken Sie auf die Schaltfläche "Dateioptionen" (die erste in der Symbolleiste).
2. Wählen Sie im Menü **Speichern** oder **Speichern unter**.
3. Speichern Sie die Datei im Format *sps*.

Gehen Sie wie folgt vor, um früher erstellte Syntaxdateien zu verwenden und den aktuellen Inhalt des Syntaxeditors zu ersetzen, falls vorhanden:

1. Klicken Sie auf die Schaltfläche "Dateioptionen" (die erste in der Symbolleiste).
2. Wählen Sie **Öffnen** im Menü aus.
3. Wählen Sie eine *.sps*-Datei aus, um den Inhalt dieser Datei in die Registerkarte "Syntax" für den Ausgabeknoten einzufügen.

Gehen Sie wie folgt vor, um früher gespeicherte Syntax ohne Ersetzen des aktuellen Inhalts einzufügen:

1. Klicken Sie auf die Schaltfläche "Dateioptionen" (die erste in der Symbolleiste).
2. Wählen Sie **Einfügen** im Menü aus.
3. Wählen Sie eine *.sps*-Datei aus, um den Inhalt dieser Datei für den Ausgabeknoten an der Cursorposition einzufügen.

Optional können Sie das Kontrollkästchen **Importknoten für resultierende Daten generieren** aktivieren, um einen Statistikdatei-Quellenknoten zu erstellen, mit dem die resultierenden Daten in einen anderen Stream importiert werden können. Der Knoten wird auf dem Bildschirmerstellungsbereich abgelegt, wobei die in der *SAV*-Datei enthaltenen Daten im Feld **Datei** angegeben werden (Standardspeicherort ist das Installationsverzeichnis von IBM SPSS Modeler).

Beim Klicken auf **Ausführen** werden die Ergebnisse im IBM SPSS Statistics-Ausgabeviewer angezeigt. Weitere Informationen zum Viewer finden Sie in der IBM SPSS Statistics-Dokumentation.

Statistics-Ausgabeknoten - Registerkarte "Ausgabe"

Auf der Registerkarte "Ausgabe" legen Sie Format und Position der Ausgabe fest. Sie können auswählen, dass die Ergebnisse auf dem Bildschirm angezeigt werden sollen, oder sie an einen der verfügbaren Dateitypen senden.

Ausgabename. Bestimmt den Namen der Ausgabe, die beim Ausführen des Knotens erstellt wird. Mit **Auto** wird ein Name auf der Grundlage des Knotens bestimmt, mit dem die Ausgabe erzeugt wird. Optional können Sie auch **Angepasst** auswählen und einen anderen Namen angeben.

Ausgabe auf Bildschirm (Standardeinstellung)**.** Erstellt ein Ausgabeobjekt für die Online-Anzeige. Das Ausgabeobjekt wird auf der Registerkarte "Ausgaben" im Manager-Fenster dargestellt, wenn der Ausgabeknoten ausgeführt wird.

Ausgabe in Datei. Speichert die Ausgabe in einer Datei, wenn der Knoten ausgeführt wird. Wenn Sie diese Option wählen, geben Sie einen Dateinamen im Feld **Dateiname** an (oder wechseln Sie zu einem Verzeichnis und geben Sie einen Dateinamen mithilfe der Feldauswahlschaltfläche an) und wählen Sie einen Dateityp aus.

Dateityp. Wählen Sie den Dateiyp aus, an den Sie die Ausgabe senden möchten.

- **HTML-Dokument (*.html).** Schreibt die Ausgabe im HTML-Format.
- **IBM SPSS Statistics Viewer-Datei (*.spv).** Schreibt die Ausgabe in einem Format, das vom IBM SPSS Statistics-Ausgabeviewer gelesen werden kann.
- **IBM SPSS Statistics Web Reports-Datei (*.spw).** Schreibt die Ausgabe in einem IBM SPSS Statistics Web Reports-Format, das in einem IBM SPSS Collaboration and Deployment Services-Repository veröffentlicht und anschließend in einem Web-Browser angezeigt werden kann. Weitere Informationen finden Sie im Thema „Veröffentlichen im Web” auf Seite 283.

Hinweis: Wenn Sie **Ausgabe auf Bildschirm** auswählen, hat die IBM SPSS Statistics-OMS-Anweisung `VIEWER=NO` keine Auswirkung. Außerdem sind die Scripterstellungs-APIs (Modul *Basic* und *Python SpssClient*) in IBM SPSS Modeler nicht verfügbar.

Statistikexportknoten

Mit dem Statistikexportknoten können Sie die Daten im IBM SPSS Statistics-Format *.sav* speichern. *SAV*-Dateien von IBM SPSS Statistics können von IBM SPSS Statistics Base und anderen Modulen gelesen werden. Dieses Format wird auch für die IBM SPSS Modeler-Cache-Dateien verwendet.

Beim Zuordnen von IBM SPSS Modeler-Feldnamen für IBM SPSS Statistics-Variablennamen entstehen hin und wieder Fehler, weil die IBM SPSS Statistics-Variablennamen maximal 64 Zeichen umfassen dürfen und bestimmte Zeichen nicht zulässig sind, beispielsweise Leerzeichen, Dollarzeichen ($) und Gedankenstriche (-). Diese Einschränkungen können auf zweierlei Weise umgangen werden:

- Benennen Sie die Felder so um, dass die Namen den Anforderungen für IBM SPSS Statistics-Variablennamen genügen. Klicken Sie hierzu auf die Registerkarte "Filter". Weitere Informationen finden Sie im Thema „Umbenennen oder Filtern von Feldern für IBM SPSS Statistics” auf Seite 362.
- Legen Sie fest, dass sowohl die Feldnamen als auch die Beschriftungen aus IBM SPSS Modeler exportiert werden sollen.

Hinweis: IBM SPSS Modeler schreibt *SAV*-Dateien im Unicode-Format UTF-8. IBM SPSS Statistics unterstützt nur Dateien im Unicode-Format UTF-8 aus Version 16.0 und höher. Um die Möglichkeit beschädigter Daten zu vermeiden, sollten *SAV*-Dateien nicht in IBM SPSS Statistics-Versionen vor 16.0 verwendet werden. Weitere Informationen finden Sie in der Hilfe zu IBM SPSS Statistics.

Mehrfachantwortsets. Etwaige im Stream definierte Mehrfachantwortsets bleiben beim Export der Datei automatisch erhalten. Mehrfachantwortsets können über jeden Knoten, der die Registerkarte "Filter" enthält, angezeigt und bearbeitet werden. Weitere Informationen finden Sie im Thema „Bearbeiten von Mehrfachantwortsets” auf Seite 131.

Statistikexportknoten - Registerkarte "Exportieren"

Datei exportieren. Hier können Sie den Namen der Datei angeben. Geben Sie einen Dateinamen an oder klicken Sie auf die Feldauswahlschaltfläche und wechseln Sie zum Pfad der gewünschten Datei.

Dateityp. Wählen Sie aus, ob die Datei im normalen Format *.sav* oder im komprimierten Format *.zsav* gespeichert werden soll.

Datei mit Kennwort verschlüsseln. Wählen Sie dieses Feld aus, um die Datei mit einem Kennwort zu verschlüsseln. Sie werden aufgefordert, das **Kennwort** in einem separaten Dialogfeld einzugeben und zu bestätigen.

Anmerkung: Kennwortgeschützte Dateien können nur von IBM SPSS Modeler Version 16 oder höher oder von IBM SPSS Statistics Version 21 oder höher geöffnet werden.

Feldnamen exportieren. Dient zur Angabe einer Methode zur Behandlung von Variablennamen und Beschriftungen beim Exportieren aus IBM SPSS Modeler in eine Datei im IBM SPSS Statistics-Format *.sav* oder *.zsav*.

- **Namen und Variablenbeschriftungen.** Hiermit werden sowohl die Feldnamen als auch die Feldbeschriftungen aus IBM SPSS Modeler exportiert. Die Namen werden als IBM SPSS Statistics-Variablennamen exportiert, die Beschriftungen entsprechend als IBM SPSS Statistics-Variablenbeschriftungen.
- **Namen als Variablenbeschriftungen.** Mit dieser Einstellung werden die IBM SPSS Modeler-Feldnamen in IBM SPSS Statistics als Variablenbeschriftungen verwendet. Bei IBM SPSS Modeler können verschiedene Zeichen in den Feldnamen verwendet werden, die bei IBM SPSS Statistics-Variablennamen nicht gültig sind. Um die mögliche Bildung ungültiger IBM SPSS Statistics-Namen zu vermeiden, wählen Sie stattdessen die Option **Beschriftungen** oder passen die Feldnamen auf der Registerkarte "Filter" an.

Anwendung starten. Wenn IBM SPSS Statistics oder AnswerTree auf dem Computer installiert ist, können Sie die Anwendung mit dieser Option direkt für die gespeicherte Datendatei aufrufen. Die Optionen zum Starten der Anwendung müssen im Dialogfeld "Hilfsanwendungen" angegeben werden. Weitere Informationen finden Sie im Thema „IBM SPSS Statistics-Hilfsanwendungen” auf Seite 324. Soll lediglich eine Datei im IBM SPSS Statistics-Format *.sav* oder *.zsav* erstellt werden, ohne ein externes Programm zu öffnen, inaktivieren Sie diese Option.

Importknoten für diese Daten generieren. Mit dieser Option lassen Sie automatisch einen Quellenknoten für eine Statistikdatei erzeugen, mit dem die exportierte Datendatei eingelesen wird. Weitere Informationen finden Sie im Thema „Statistikdateiknoten” auf Seite 354.

Umbenennen oder Filtern von Feldern für IBM SPSS Statistics

Vor dem Exportieren oder Bereitstellen von Daten aus IBM SPSS Modeler in externe Anwendungen wie IBM SPSS Statistics müssen die Feldnamen gegebenenfalls umbenannt oder angepasst werden. Die Dialogfelder "Statistiktransformation", "Statistikausgabe" und "Statistikexport" beinhalten jeweils die Registerkarte "Filter", mit der dieser Vorgang erleichtert wird.

Eine ausführliche Beschreibung der Funktionen auf der Registerkarte "Filter" finden Sie an anderer Stelle in diesem Handbuch. Weitere Informationen finden Sie im Thema „Festlegen der Filteroptionen” auf Seite 130. In diesem Thema finden Sie Tipps zum Einlesen von Daten in IBM SPSS Statistics.

Führen Sie folgende Schritte aus, um die Feldnamen an das IBM SPSS Statistics-Namensschema anzupassen:

1. Klicken Sie auf der Registerkarate "Filter" auf die Symbolleistenschaltfläche "Optionen im Filtermenü" (die erste in der Symbolleiste).
2. Wählen Sie "Umbenennen für IBM SPSS Statistics" aus.
3. Im Dialogfeld "Umbenennen für IBM SPSS Statistics" können Sie auswählen, ob ungültige Zeichen in Dateinamen durch ein Rautenzeichen (#) oder einen Unterstrich (_) ersetzt werden.

Umbenennen von Mehrfachantwortsets. Wählen Sie diese Option aus, wenn Sie die Namen von Mehrfachantwortsets bearbeiten wollen, die mithilfe eines Statistikdatei-Quellenknoten in IBM SPSS Modeler importiert werden können. Sie werden zum Aufzeichnen von Daten verwendet, die mehr als einen Wert für jeden Fall haben, wie beispielsweise bei Umfrageantworten.

Kapitel 9. Superknoten

Überblick über Superknoten

Einer der Gründe, warum der Umgang mit der visuellen Programmierschnittstelle von IBM SPSS Modeler so leicht zu erlernen ist, liegt darin, dass jeder Knoten eine klar definierte Funktion erfüllt. Für eine komplexe Verarbeitung ist jedoch eventuell eine lange Sequenz von Knoten erforderlich. Dadurch können die Elemente im Streamerstellungsbereich unübersichtlich werden und es kann schwierig werden, den Streamdiagrammen zu folgen. Es gibt zwei Methoden zur Vermeidung eines langen und komplexen Streams:

- Sie können eine Verarbeitungssequenz in mehrere Streams aufteilen, die einander als Datengrundlage dienen. So könnte der erste Stream beispielsweise eine Datendatei erstellen, die der zweite Stream als Eingabe verwendet. Der zweite erstellt eine Datei, die der dritte Stream als Eingabe verwendet, usw. Diese Streams können Sie verwalten, indem Sie sie in einem **Projekt** speichern. Ein Projekt kann mehrere Streams und deren Ausgaben organisieren. Projektdateien enthalten jedoch nur einen Verweis auf die Objekte, die sie enthalten, und Sie müssen noch immer mehrere Streamdateien verwalten.
- Sie können einen **Superknoten** als als effizientere Alternative bei der Arbeit mit komplexen Streamprozessen erstellen.

Superknoten fassen mehrere Knoten zu einem einzigen zusammen, indem sie Bereiche eines Datenstreams verkapseln. Dies bietet zahlreiche Vorteile für das Data Mining:

- Streams sind überschaubarer und können besser verwaltet werden.
- Knoten können zu einem geschäftsspezifischen Superknoten zusammengefasst werden.
- Superknoten können in Bibliotheken exportiert und in mehreren Data-Mining-Projekten wieder verwendet werden.

Typen von Superknoten

Superknoten werden im Datenstream durch ein sternförmiges Symbol angezeigt. Das Symbol ist schattiert, um den Superknotentyp und die Richtung anzugeben, in der der Stream zum Superknoten hin bzw. von ihm weg fließen muss.

Es gibt drei Typen von Superknoten:

- Quellensuperknoten
- Prozesssuperknoten
- Endsuperknoten

Quellensuperknoten

Quellensuperknoten enthalten eine Datenquelle, genau wie ein normaler Quellenknoten, und können an jeder Stelle verwendet werden, an der auch ein normaler Quellenknoten eingesetzt werden kann. Die linke Seite eines Quellensuperknotens ist schattiert, um anzuzeigen, dass er auf der linken Seite "geschlossen" ist und dass die Daten *vom* Superknoten nach unten im Stream fließen müssen.

Quellensuperknoten weisen nur einen einzigen Verbindungspunkt auf der rechten Seite auf, der anzeigt, dass die Daten den Superknoten verlassen und nach unten im Stream fließen.

Prozesssuperknoten

Prozesssuperknoten enthalten nur Prozessknoten und sind nicht schattiert, um anzuzeigen, dass die Daten bei diesem Superknotentyp sowohl in den Knoten *hinein*- als auch aus ihm *heraus*fließen können.

Prozesssuperknoten weisen sowohl links als auch rechts Verbindungspunkte auf, was anzeigt, dass die Daten in den Superknoten eintreten und ihn dann wieder verlassen und wieder in den Stream eintreten. Superknoten können zwar zusätzliche Streamfragmente und sogar zusätzliche Streams enthalten, beide Verbindungspunkte müssen aber durch einen einzelnen Pfad fließen, der die Punkte *Aus Stream* und *Bis Stream* verbindet.

Hinweis: Prozesssuperknoten werden manchmal als *Manipulationssuperknoten* bezeichnet.

Endsuperknoten

Endsuperknoten enthalten mindestens einen Endknoten (Diagramm, Tabelle usw.) und können auf dieselbe Weise verwendet werden wie Endknoten. Die rechte Seite eines Quellensuperknotens ist schattiert, um anzuzeigen, dass er auf der rechten Seite "geschlossen" ist und dass die Daten nur *in* den Endsuperknoten fließen können.

Quellensuperknoten weisen nur einen einzigen Verbindungspunkt auf der rechten Seite auf, der anzeigt, dass die Daten aus dem Stream in den Superknoten eintreten und dort enden.

Endsuperknoten können auch Scripts enthalten, die für alle Knoten innerhalb des Superknotens die Reihenfolge der Ausführung festlegen. Weitere Informationen finden Sie im Thema „Superknoten und Scripts” auf Seite 370.

Erstellen von Superknoten

Beim Erstellen von Superknoten wird der Datenstream reduziert, indem mehrere Knoten zu einem Knoten gekapselt werden. Nach dem Erstellen bzw. Laden eines Streams im Erstellungsbereich gibt es mehrere Möglichkeiten zum Erstellen eines Superknotens.

Mehrfachauswahl

Die einfachste Methode zum Erstellen eines Superknotens besteht in der Auswahl aller Knoten, die gekapselt werden sollen:

1. Mithilfe der Maus können Sie mehrere Knoten im Streamerstellungsbereich auswählen. Außerdem können Sie bei gedrückter Umschalttaste auf einen Stream oder einen Abschnitt eines Streams klicken, um ihn ausuwählen. *Hinweis*: Die ausgewählten Knoten müssen aus einem kontinuierlichen oder gegabelten Stream stammen. Knoten, die nicht aneinander angrenzen oder in irgendeiner Weise verbunden sind, können nicht ausgewählt werden.
2. Anschließend verkapseln Sie die ausgewählten Knoten unter Verwendung einer der folgenden drei Methoden:
   - Klicken Sie auf das Superknotensymbol (sternförmig) in der Symbolleiste.
   - Klicken Sie mit der rechten Maustaste auf den Superknoten und wählen Sie aus dem Kontextmenü folgende Optionen:

     Superknoten erstellen > **Aus Auswahl**
   - Wählen Sie im Superknotenmenü folgende Befehlsfolge aus:

     Superknoten erstellen > **Aus Auswahl**

Bei allen drei Optionen werden die Knoten in einem Superknoten gekapselt, dessen Typ (Quellen-, Prozess- oder Endsuperknoten) durch die Schattierung angezeigt wird. Die Grundlage dafür bildet der jeweilige Inhalt.

Einzelauswahl

Außerdem können Sie einen Superknoten erstellen, indem Sie einen einzelnen Knoten auswählen und mithilfe von Menüoptionen den Start und das Ende des Superknotens festlegen oder alle Elemente verkapseln, die im Stream hinter dem ausgewählten Knoten liegen.

1. Klicken Sie auf den Knoten, der den Start der Verkapselung bestimmt.
2. Wählen Sie im Superknotenmenü folgende Befehlsfolge aus:

   Superknoten erstellen > **Ab hier**

Superknoten können außerdem auf mehr interaktive Weise erstellt werden. Dazu wählen Sie den Start und das Ende des Streamabschnitts aus, um die Knoten zu verkapseln:

1. Klicken Sie auf den ersten oder letzten Knoten, der in den Superknoten aufgenommen werden soll.
2. Wählen Sie im Superknotenmenü folgende Befehlsfolge aus:

   Superknoten erstellen > **Auswählen...**
3. Alternativ können Sie die Optionen des Kontextmenüs verwenden. Klicken Sie dazu mit der rechten Maustaste auf den gewünschten Knoten.
4. Der Cursor wird zu einem Superknotensymbol, wodurch angezeigt wird, dass ein weiterer Punkt im Stream ausgewählt werden muss. Gehen Sie entweder nach unten oder nach oben im Stream zum "anderen Ende" des Superknotenfragments und klicken Sie auf einen Knoten. Dadurch werden alle dazwischenliegenden Knoten durch das Sternsymbol des Superknotens ersetzt.

Hinweis: Die ausgewählten Knoten müssen aus einem kontinuierlichen oder gegabelten Stream stammen. Knoten, die nicht aneinander angrenzen oder in irgendeiner Weise verbunden sind, können nicht ausgewählt werden.

Verschachteln von Superknoten

Superknoten können innerhalb von anderen Superknoten verschachtelt werden. Die Regeln für die einzelnen Superknotentypen (Quellen-, Prozess- und Endsuperknoten) gelten uneingeschränkt auch für verschachtelte Superknoten. Bei einem Prozesssuperknoten mit Verschachtelung muss ein kontinuierlicher Datenfluss durch alle verschachtelten Superknoten vorliegen, damit er ein Prozesssuperknoten bleiben kann. Wenn es sich bei einem der verschachtelten Superknoten um einen Endsuperknoten handelt, fließen die Daten nicht mehr durch die Hierarchie.

End- und Quellensuperknoten können andere Typen verschachtelter Superknoten enthalten, doch die grundlegenden Regeln für das Erstellen von Superknoten gelten weiterhin.

Sperren von Superknoten

Nach dem Erstellen eines Superknotens können Sie ihn mit einem Kennwort schützen, um eine Änderung zu verhindern. Dies ist z. B. möglich, wenn Sie Streams oder Teile von Streams als Vorlagen mit festem Wert für andere Benutzer in Ihrer Organisation erstellen, die weniger erfahren mit dem Einrichten von IBM SPSS Modeler-Abfragen sind.

Für einen gesperrten Superknoten können Benutzer auf der Registerkarte "Parameter" immer noch Werte für Parameter eingeben, die definiert wurden. Außerdem kann ein gesperrter Superknoten ohne Eingabe des Kennworts ausgeführt werden.

Hinweis: Das Sperren und Entsperren mithilfe von Scripts ist nicht möglich.

Sperren und Entsperren eines Superknotens

Warnung: Vergessene Kennwörter können nicht wiederhergestellt werden.

Sie können einen Superknoten über eine der drei Registerkarten sperren oder entsperren.

1. Klicken Sie auf **Knoten sperren**.
2. Geben Sie das Kennwort ein und bestätigen Sie es.
3. Klicken Sie auf **OK**.

Ein kennwortgeschützter Superknoten wird im Streamerstellungsbereich durch ein kleines Vorhängeschlosssymbol in der oberen linken Ecke des Superknotensymbols markiert.

Entsperren eines Superknotens

1. Sie entfernen den Kennwortschutz permanent, indem Sie auf **Knoten entsperren** klicken. Sie werden dann zur Eingabe des Kennworts aufgefordert.
2. Geben Sie das Kennwort ein und klicken Sie auf **OK**. Der Superknoten ist nun nicht mehr kennwortgeschützt und an seinem Symbol im Stream wird kein Vorhängeschloss-Symbol mehr angezeigt.

Bearbeiten eines gesperrten Superknotens

Wenn Sie versuchen, Parameter für einen gesperrten Superknoten zu definieren oder zu zoomen, um einen gesperrten Superknoten anzuzeigen, werden Sie aufgefordert, das Kennwort einzugeben.

Geben Sie das Kennwort ein und klicken Sie auf **OK**.

Sie können nun die Parameterdefinitionen bearbeiten und wie gewünscht zoomen, bis Sie den Stream schließen, in dem sich der Superknoten befindet.

Beachten Sie, dass damit nicht der Kennwortschutz entfernt wird, sondern Ihnen nur ermöglicht wird, mit dem Superknoten zu arbeiten. Weitere Informationen finden Sie im Thema „Sperren und Entsperren eines Superknotens" auf Seite 365.

Bearbeiten von Superknoten

Nach dem Erstellen eines Superknotens können Sie ihn genauer untersuchen, indem Sie ihn vergrößern. Wenn der Superknoten gesperrt ist, werden Sie zur Eingabe des Kennworts aufgefordert. Weitere Informationen finden Sie im Thema „Bearbeiten eines gesperrten Superknotens".

Wenn Sie den Inhalt eines Superknotens anzeigen möchten, können Sie dazu entweder das Vergrößerungssymbol der IBM SPSS Modeler-Symbolleiste oder die folgende Methode verwenden:

1. Klicken Sie mit der rechten Maustaste auf einen Superknoten.
2. Wählen Sie im Kontextmenü die Option **Vergrößern** aus.

Der Inhalt des ausgewählten Superknotens wird in einer leicht abweichenden IBM SPSS Modeler-Umgebung angezeigt, in der die Verbindungen den Fluss der Daten durch den Stream bzw. das Streamfragment anzeigen. Auf dieser Ebene im Streamerstellungsbereich können Sie mehrere Aufgaben durchführen:

- Ändern des Superknotentyps - Quellen-, Prozess- oder Endsuperknoten.
- Erstellen von Parametern bzw. Bearbeiten der Werte eines Parameters. Parameter werden zur Scripterstellung und für CLEM-Ausdrücke verwendet.
- Festlegen von Caching-Optionen für den Superknoten und seine Unterknoten.
- Erstellen oder Bearbeiten eines Superknotenscripts (nur bei Endsuperknoten).

Ändern der Superknotentypen

Unter gewissen Umständen kann es sinnvoll sein, den Typ eines Superknotens zu ändern. Diese Funktion ist nur verfügbar, wenn Sie die Ansicht des Superknotens vergrößert haben, und sie bezieht sich nur auf den Superknoten auf dieser Stufe. Die drei Superknotentypen werden in der folgenden Tabelle erläutert.

Tabelle 57. Superknotentypen.

Superknotentyp	Beschreibung
Quellensuperknoten	Eine ausgehende Verbindung
Prozesssuperknoten	Zwei Verbindungen: eine eingehende und eine ausgehende
Endsuperknoten	Eine eingehende Verbindung

So ändern Sie den Typ eines Superknotens:

1. Stellen Sie sicher, dass sie die Ansicht des Superknotens vergrößert haben.
2. Wählen Sie im Superknotenmenü die Option **Superknotentyp** und wählen Sie anschließend den Typ aus.

Anmerkungen für Superknoten und Umbenennen von Superknoten

Sie können einen Superknoten umbenennen, wenn er im Stream angezeigt wird, sowie Anmerkungen schreiben, die in einem Projekt oder Bericht verwendet werden. So können Sie auf diese Eigenschaften zugreifen:

- Klicken Sie mit der rechten Maustaste auf einen Superknoten (verkleinert) und wählen Sie die Option **Umbenennen und mit Anmerkung versehen**.
- Alternativ wählen Sie im Superknotenmenü die Option **Umbenennen und mit Anmerkung versehen**. Diese Option ist sowohl im vergrößerten als auch im verkleinerten Modus verfügbar.

In beiden Fällen wird ein Dialogfeld geöffnet, bei dem die Registerkarte "Anmerkungen" ausgewählt ist. Mit den hier verfügbaren Optionen können Sie den im Streamerstellungsbereich angezeigten Namen anpassen und eine Dokumentation zu den Superknotenoperationen bereitstellen.

Verwenden von Kommentaren mit Superknoten

Wenn Sie aus einem kommentierten Knoten oder Nugget einen Superknoten erstellen, müssen Sie den Kommentar in die Auswahl aufnehmen, falls dieser im Superknoten erscheinen soll. Wenn Sie den Kommentar aus der Auswahl weglassen, bleibt der Kommentar im Stream, nachdem der Superknoten erstellt wurde.

Wenn Sie einen Superknoten erweitern, der Kommentare enthielt, werden die Kommentare wieder an ihrer ursprünglichen Position (vor der Erstellung des Superknotens) eingesetzt.

Wenn Sie einen Superknoten erweitern, der kommentierte Objekte enthielt, aber die Kommentare nicht in den Superknoten aufgenommen wurden, werden die Objekte wieder an ihrer ursprünglichen Position eingesetzt, aber die Kommentare werden nicht erneut verknüpft.

Superknotenparameter

In IBM SPSS Modeler haben Sie die Möglichkeit, benutzerdefinierte Variablen festzulegen, beispielsweise *Minvalue*, deren Werte bei der Verwendung in der Scripterstellung oder in CLEM-Ausdrücken angegeben werden können. Diese Variablen heißen **Parameter**. Sie können Parameter für Streams, Sitzungen und Superknoten festlegen. Alle für einen Superknoten festgelegten Parameter sind bei der Erstellung von CLEM-Ausdrücken in diesem Superknoten oder etwaigen verschachtelten Knoten verfügbar. Die für verschachtelte Superknoten festgelegten Parameter stehen nicht für den übergeordneten Superknoten zur Verfügung.

Es gibt zwei Schritte zum Erstellen und Festlegen von Parametern für Superknoten:

1. Definieren Sie die Parameter für den Superknoten.
2. Geben Sie anschließend den Wert für die einzelnen Parameter des Superknotens an.

Diese Parameter können dann in CLEM-Ausdrücken für alle verschachtelten Knoten verwendet werden.

Festlegen von Superknotenparametern

Parameter für einen Superknoten können sowohl im vergrößerten als auch im verkleinerten Modus definiert werden. Die definierten Parameter gelten für alle gekapselten Knoten. Zur Definition der Parameter eines Superknotens müssen Sie zunächst im Dialogfeld des Superknotens die Registerkarte "Parameter" aufrufen. Das Dialogfeld lässt sich auf folgende Weisen öffnen:

- Doppelklicken Sie auf einen Superknoten im Stream.
- Wählen Sie im Superknotenmenü den Befehl **Parameter festlegen** aus.
- Alternativ wählen Sie bei vergrößerter Superknotenansicht die Option **Parameter festlegen** aus dem Kontextmenü.

Nach dem Öffnen des Dialogfelds wird die Registerkarte "Parameter" mit allen zuvor definierten Parametern angezeigt.

So definieren Sie einen neuen Parameter:

Klicken Sie auf die Schaltfläche **Parameter definieren**, um das Dialogfeld zu öffnen.

Name. Hier werden die Parameternamen aufgelistet. Sie können einen neuen Parameter erstellen, indem Sie in diesem Feld einen Namen eingeben. Um beispielsweise einen Parameter für die Mindesttemperatur zu erstellen, könnten Sie `minvalue` eingeben. Verwenden Sie nicht das Präfix `$P-`, das Parameter in CLEM-Ausdrücken kennzeichnet. Dieser Name wird auch zur Anzeige im CLEM Expression Builder verwendet.

Langer Name. Listet den beschreibenden Namen für die einzelnen erstellten Parameter auf.

Speichertyp. Wählen Sie einen Speichertyp aus der Liste aus. Der Speichertyp gibt an, wie die Datenwerte im Parameter gespeichert werden. Wenn Sie z. B. mit Werten arbeiten, die führende Nullen enthalten und die Sie beibehalten möchten (wie 008), sollten Sie **Zeichenfolge** als Speichertyp wählen. Andernfalls werden die Nullen vom Wert abgezogen. Verfügbare Speichertypen sind "Zeichenfolge", "Ganze Zahl", "Reelle Zahl", "Uhrzeit", "Datum" und "Zeitmarke". Beachten Sie, dass bei Datumsparametern die Werte gemäß der im nächsten Absatz erläuterten ISO-Standardnotierung eingegeben werden müssen.

Wert. Listet den aktuellen Wert für die einzelnen Parameter auf. Ändern Sie den Parameter wie gewünscht. Datumsparameter müssen in ISO-Standardnotierung angegeben werden (d. h. in der Form `JJJJ-MM-TT`). Datumsangaben in anderen Formaten sind nicht zulässig.

Typ (optional). Wenn Sie den Stream für eine externe Anwendung bereitstellen möchten, wählen Sie aus der Liste ein Messniveau aus. Andernfalls sollten Sie die Spalte *Typ* so belassen, wie sie ist. Wenn Sie Wertbeschränkungen für den Parameter festlegen möchten, z. B. die Ober- und Untergrenze für einen numerischen Bereich, wählen Sie **Angeben** aus der Liste aus.

Die Optionen "Langer Name", "Speichertyp" und "Typ" können für Parameter nur über die Benutzerschnittstelle festgelegt werden. Die Festlegung dieser Optionen mithilfe von Scripts ist nicht möglich.

Klicken Sie auf die Pfeile rechts, um den ausgewählten Parameter in der Liste verfügbarer Parameter weiter nach oben oder weiter nach unten zu verschieben. Verwenden Sie die Schaltfläche zum Löschen (mit einem *X* markiert), um den ausgewählten Parameter zu entfernen.

Festlegen von Werten für Superknotenparameter

Nach der Definition von Parametern für einen Superknoten können Sie mithilfe der Parameter in einem CLEM-Ausdruck bzw. einem -Script Werte angeben.

So geben Sie die Parameter eines Superknotens an:

1. Doppelklicken Sie auf das Superknotensymbol, um das Dialogfeld für den Superknoten zu öffnen.
2. Alternativ können Sie im Superknotenmenü den Befehl **Parameter festlegen** auswählen.
3. Klicken Sie auf die Registerkarte **Parameter**. *Hinweis:* Bei den Feldern in diesem Dialogfeld handelt es sich um die Felder, die durch Klicken auf die Schaltfläche **Parameter definieren** auf dieser Registerkarte definiert wurden.
4. Geben Sie für jeden erstellten Parameter einen Wert in das Textfeld ein. Beispielsweise können Sie den Wert *minvalue* auf einen bestimmten Schwellenwert festsetzen. Dieser Parameter kann anschließend in

verschiedenen Operationen verwendet werden, beispielsweise bei der Auswahl der Datensätze oberhalb oder unterhalb dieses Schwellenwerts zur weiteren Exploration.

Verwenden von Superknotenparametern zum Zugriff auf Knoteneigenschaften

Superknotenparameter können außerdem zur Definition von Knoteneigenschaften (auch als **Slotparameter** bezeichnet) für gekapselte Knoten verwendet werden. Beispiel: Angenommen, Sie möchten festlegen, dass ein Superknoten einen gekapselten Netzknoten eine bestimmte Zeit lang mithilfe einer Zufallsstichprobe der verfügbaren Daten trainiert. Mithilfe von Parametern können Sie Werte für die Zeitdauer und und den Prozentsatz der Stichprobe angeben.

Der Beispielsuperknoten enthält einen Stichprobenknoten mit der Bezeichnung *Sample* (Stichprobe) und einem Netzknoten mit der Bezeichnung *Train* (Trainieren). Mit den Knotendialogfeldern können Sie für die Einstellung **Stichprobe** des Stichprobenknotens **Zufällig %** und für die Einstellung **Stopp bei** des Netzknotens **Zeit** festlegen. Nachdem diese Optionen angegeben wurden, können Sie auf die Knoteneigenschaften mit Parametern zugreifen und bestimmte Werte für den Superknoten angeben. Klicken Sie im Dialogfeld "Superknoten" auf **Parameter definieren** und erstellen Sie die Parameter, die in der folgenden Tabelle aufgelistet sind.

Tabelle 58. Zu erstellende Parameter

Parameter	Wert	Langer Name
Train.time	5	Zeit zum Trainieren (Minuten)
Sample.random	10	Prozentsatz der Zufallsstichprobe

Hinweis: Bei den Parameternamen, beispielsweise *Sample.random*, wird korrekte Syntax für Referenzen auf Knoteneigenschaften verwendet. Dabei steht *Sample* für den Namen des Knotens und *random* ist eine Knoteneigenschaft.

Nach der Definition dieser Parameter können Sie problemlos diese Werte für die beiden Stichproben- und Netzknoten-Eigenschaften angeben, ohne dass die einzelnen Dialogfelder erneut geöffnet werden müssen. Wählen Sie stattdessen einfach im Superknotenmenü die Option **Parameter festlegen** aus, um über das Dialogfeld des Superknotens die Registerkarte "Parameter" aufzurufen, auf der Sie neue Werte für **Zufällig %** und **Zeit** angeben können. Dies ist besonders nützlich bei der Exploration der Daten während mehrerer Iterationen der Modellerstellung.

Superknoten und Caching

Aus einem Superknoten können alle Knoten außer Endknoten im Cache gespeichert werden. Das Caching wird durch Rechtsklicken auf einen Knoten und Auswahl einer von mehreren Optionen aus dem Cache-Kontextmenü gesteuert. Diese Menüoption ist sowohl von außerhalb eines Superknotens als auch für die in einem Superknoten gekapselten Knoten verfügbar.

Es gibt verschiedene Richtlinien für Superknotencaches:

- Wenn bei mindestens einem der in einem Superknoten gekapselten Knoten Caching aktiviert ist, ist es auch beim Superknoten aktiviert.
- Wenn der Cache für einen Superknoten inaktiviert wird, wird der Cache auch für *alle* gekapselten Knoten inaktiviert.
- Beim Aktivieren des Caching bei einem Superknoten wird der Cache tatsächlich für den letzten Caching-fähigen Superknoten aktiviert. Anders ausgedrückt: Wenn der letzte Superknoten ein Auswahlknoten ist, wird der Cache für diesen Auswahlknoten aktiviert. Wenn es sich beim letzten Unterknoten um einen Endknoten handelt (bei dem Caching nicht möglich ist), wird der nächste Knoten weiter oben im Stream, der Caching unterstützt, aktiviert.
- Nach der Festlegung von Caches für die Unterknoten eines Superknotens werden die Caches bei jeder Aktivität oberhalb des Knotens geleert, für den die Cachespeicherung erfolgte, wie beispielsweise Hinzufügen und Bearbeiten von Knoten.

Superknoten und Scripts

Sie können mithilfe der IBM SPSS Modeler-Scriptsprache einfache Programme schreiben, mit denen der Inhalt eines Endsuperknotens bearbeitet und ausgeführt werden kann. Beispielsweise können Sie für komplexe Streams die Reihenfolge der Ausführung festlegen. Wenn ein Superknoten beispielsweise einen Globalwerteknoten enthält, der vor einem Plotknoten ausgeführt werden muss, können Sie ein Script erstellen, mit dem zuerst der Globalwerteknoten ausgeführt wird. Die durch diesen Knoten berechneten Werte, wie Durchschnitt oder Standardabweichung, können anschließend bei der Ausführung des Plotknotens verwendet werden.

Die Registerkarte "Script" des Dialogfelds "Superknoten" ist nur für Endsuperknoten verfügbar.

So öffnen Sie das Scriptdialogfeld für einen Endsuperknoten:

- Klicken Sie mit der rechten Maustaste auf den Erstellungsbereich des Superknotens und wählen Sie die Option **Superknotenscript** aus:
- Alternativ können Sie sowohl im vergrößerten als auch im verkleinerten Modus im Superknotenmenü die Option **Superknotenscript** auswählen.

Hinweis: Superknotenscripts werden nur mit dem Stream und dem Superknoten ausgeführt, wenn im Dialogfeld **Dieses Script ausführen** ausgewählt wurde.

Spezifische Optionen für die Scripts und ihre Verwendung in IBM SPSS Modeler finden Sie im *Handbuch für Scripterstellung und Automatisierung* auf der IBM SPSS Modeler DVD.

Speichern und Laden von Superknoten

Einer der Vorteile von Superknoten besteht darin, dass sie gespeichert und in anderen Streams wieder verwendet werden können. Beim Speichern und Laden von Superknoten werden *.slb*-Erweiterungen verwendet.

So speichern Sie einen Superknoten:

1. Vergrößern Sie den Superknoten.
2. Wählen Sie im Superknotenmenü den Befehl **Superknoten speichern** aus.
3. Geben Sie im Dialogfeld einen Dateinamen und ein Verzeichnis an.
4. Wählen Sie aus, ob der gespeicherte Superknoten zum aktuellen Projekt hinzugefügt werden soll.
5. Klicken Sie auf **Speichern**.

So laden Sie einen Superknoten:

1. Wählen Sie im Menü "Einfügen" im IBM SPSS Modeler-Fenster die Option **Superknoten** aus.
2. Wählen Sie eine Superknotendatei (*.slb*) aus dem aktuellen Verzeichnis aus oder wechseln Sie zu einem anderen Verzeichnis.
3. Klicken Sie auf **Laden**.

Hinweis: Bei importierten Superknoten werden für alle Parameter Standardwerte verwendet. Zum Ändern der Parameter doppelklicken Sie auf einen Superknoten im Streamerstellungsbereich.

Bemerkungen

Diese Informationen wurden für weltweit angebotene Produkte und Dienstleistungen erarbeitet.

Möglicherweise bietet IBM die in dieser Dokumentation beschriebenen Produkte, Services oder Funktionen in anderen Ländern nicht an. Informationen über die gegenwärtig im jeweiligen Land verfügbaren Produkte und Services sind beim zuständigen IBM Ansprechpartner erhältlich. Hinweise auf IBM Lizenzprogramme oder andere IBM Produkte bedeuten nicht, dass nur Programme, Produkte oder Services von IBM verwendet werden können. Anstelle der IBM Produkte, Programme oder Services können auch andere, ihnen äquivalente Produkte, Programme oder Services verwendet werden, solange diese keine gewerblichen oder anderen Schutzrechte von IBM verletzen. Die Verantwortung für den Betrieb von Produkten, Programmen und Services anderer Anbieter liegt beim Kunden.

Für in diesem Handbuch beschriebene Erzeugnisse und Verfahren kann es IBM Patente oder Patentanmeldungen geben. Mit der Auslieferung dieses Handbuchs ist keine Lizenzierung dieser Patente verbunden. Lizenzanforderungen sind schriftlich an folgende Adresse zu richten (Anfragen an diese Adresse müssen auf Englisch formuliert werden):

IBM Director of Licensing
IBM Europe, Middle East & Africa
Tour Descartes
2, avenue Gambetta
92066 Paris La Defense
France

Trotz sorgfältiger Bearbeitung können technische Ungenauigkeiten oder Druckfehler in dieser Veröffentlichung nicht ausgeschlossen werden. Die hier enthaltenen Informationen werden in regelmäßigen Zeitabständen aktualisiert und als Neuausgabe veröffentlicht. IBM kann ohne weitere Mitteilung jederzeit Verbesserungen und/oder Änderungen an den in dieser Veröffentlichung beschriebenen Produkten und/oder Programmen vornehmen.

Verweise in diesen Informationen auf Websites anderer Anbieter werden lediglich als Service für den Kunden bereitgestellt und stellen keinerlei Billigung des Inhalts dieser Websites dar. Das über diese Websites verfügbare Material ist nicht Bestandteil des Materials für dieses IBM Produkt. Die Verwendung dieser Websites geschieht auf eigene Verantwortung.

Werden an IBM Informationen eingesandt, können diese beliebig verwendet werden, ohne dass eine Verpflichtung gegenüber dem Einsender entsteht.

Lizenznehmer des Programms, die Informationen zu diesem Produkt wünschen mit der Zielsetzung: (i) den Austausch von Informationen zwischen unabhängig voneinander erstellten Programmen und anderen Programmen (einschließlich des vorliegenden Programms) sowie (ii) die gemeinsame Nutzung der ausgetauschten Informationen zu ermöglichen, wenden sich an folgende Adresse:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
USA

Die Bereitstellung dieser Informationen kann unter Umständen von bestimmten Bedingungen - in einigen Fällen auch von der Zahlung einer Gebühr - abhängig sein.

Die Lieferung des in diesem Dokument beschriebenen Lizenzprogramms sowie des zugehörigen Lizenzmaterials erfolgt auf der Basis der IBM Rahmenvereinbarung bzw. der Allgemeinen Geschäftsbedingungen von IBM, der IBM Internationalen Nutzungsbedingungen für Programmpakete oder einer äquivalenten Vereinbarung.

Alle in diesem Dokument enthaltenen Leistungsdaten stammen aus einer kontrollierten Umgebung. Die Ergebnisse, die in anderen Betriebsumgebungen erzielt werden, können daher erheblich von den hier erzielten Ergebnissen abweichen. Einige Daten stammen möglicherweise von Systemen, deren Entwicklung noch nicht abgeschlossen ist. Eine Gewährleistung, dass diese Daten auch in allgemein verfügbaren Systemen erzielt werden, kann nicht gegeben werden. Darüber hinaus wurden einige Daten unter Umständen durch Extrapolation berechnet. Die tatsächlichen Ergebnisse können davon abweichen. Benutzer dieses Dokuments sollten die entsprechenden Daten in ihrer spezifischen Umgebung prüfen.

Alle Informationen zu Produkten anderer Anbieter stammen von den Anbietern der aufgeführten Produkte, deren veröffentlichten Ankündigungen oder anderen allgemein verfügbaren Quellen. IBM hat diese Produkte nicht getestet und kann daher keine Aussagen zu Leistung, Kompatibilität oder anderen Merkmalen machen. Fragen zu den Leistungsmerkmalen von Produkten anderer Anbieter sind an den jeweiligen Anbieter zu richten.

Aussagen über Pläne und Absichten von IBM unterliegen Änderungen oder können zurückgenommen werden und repräsentieren nur die Ziele von IBM.

Diese Veröffentlichung enthält Beispiele für Daten und Berichte des alltäglichen Geschäftsablaufs. Sie sollen nur die Funktionen des Lizenzprogramms illustrieren und können Namen von Personen, Firmen, Marken oder Produkten enthalten. Alle diese Namen sind frei erfunden; Ähnlichkeiten mit tatsächlichen Namen und Adressen sind rein zufällig.

Marken

IBM, das IBM Logo und ibm.com sind Marken oder eingetragene Marken der IBM Corp in den USA und/oder anderen Ländern. Weitere Produkt- und Servicenamen können Marken von IBM oder anderen Unternehmen sein. Eine aktuelle Liste der IBM Marken finden Sie auf der Webseite „Copyright and trademark information" unter www.ibm.com/legal/copytrade.shtml.

Intel, das Intel-Logo, Intel Inside, das Intel Inside-Logo, Intel Centrino, das Intel Centrino-Logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium und Pentium sind Markten oder eingetragene Marken der Intel Corporation oder der Tochtergesellschaften des Unternehmens in den USA und anderen Ländern.

Linux ist eine eingetragene Marke von Linus Torvalds in den USA, anderen Ländern oder beidem.

Microsoft, Windows, Windows NT und das Windows-Logo sind Marken der Microsoft Corporation in den USA und/oder anderen Ländern.

UNIX ist eine eingetragene Marke von The Open Group in den USA und anderen Ländern.

Java und alle auf Java basierenden Marken und Logos sind Marken oder eingetragene Marken der Oracle Corporation und/oder ihrer verbundenen Unternehmen.

Weitere Produkt- und Servicenamen können Marken von IBM oder anderen Unternehmen sein.

Glossar

B

Bereich. Die Differenz zwischen den größten und kleinsten Werten einer numerischen Variablen; Maximalwert minus Minimalwert.

E

Eindeutig. Bewertet alle Effekte gleichzeitig; damit werden alle Effekte an alle sonstigen Effekte jedweden Typs angepasst.

G

Gültig. Gültige Fälle, d. h. solche, die weder den systemdefiniert fehlenden Wert noch einen benutzerdefiniert fehlenden Wert aufweisen.

K

Kovarianz. Ein nicht standardisiertes Maß für den Zusammenhang zwischen zwei Variablen. Es ist gleich der Kreuzproduktabweichung geteilt durch N-1.

Kurtosis. Ein Maß dafür, wie sehr die Beobachtungen um einen zentralen Punkt gruppiert sind. Bei einer Normalverteilung ist der Wert der Kurtosis gleich 0. Bei positiver Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung enger um das Zentrum der Verteilung gruppiert und haben dünnere Flanken bis hin zu den Extremwerten der Verteilung. Ab dort sind die Flanken der leptokurtischen Verteilung im Vergleich zu einer Normalverteilung dicker. Bei negativer Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung weniger eng gruppiert und haben dickere Flanken bis hin zu den Extremwerten der Verteilung. Ab dort sind die Flanken der platykurtischen Verteilung im Vergleich zu einer Normalverteilung dünner.

M

Maximum. Der größte Wert einer numerischen Variablen.

Mittelwert. Ein Lagemaß (zentrale Tendenz). Die Summe der Ränge, geteilt durch die Zahl der Fälle.

Median. Wert, über und unter dem jeweils die Hälfte der Fälle liegt; 50. Perzentil. Bei einer geraden Anzahl von Fällen ist der Median der Mittelwert der beiden mittleren Fälle, wenn diese auf- oder absteigend sortiert sind. Der Median ist ein Lagemaß, das gegenüber Ausreißern unempfindlich ist (im Gegensatz zum Mittelwert, der durch wenige extrem niedrige oder hohe Werte beeinflusst werden kann).

Minimum. Der kleinste Wert einer numerischen Variablen.

Modalwert. Der am häufigsten auftretende Wert. Wenn mehrere Werte gleichermaßen die größte Häufigkeit aufweisen, ist jeder von ihnen ein Modalwert.

S

Schiefe. Ein Maß der Asymmetrie der Verteilung. Die Normalverteilung ist symmetrisch, ihre Schiefe hat den Wert 0. Eine Verteilung mit einer deutlichen positiven Schiefe läuft nach rechts lang aus (lange rechte Flanke). Eine Verteilung mit einer deutlichen negativen Schiefe läuft nach links lang aus (lange linke Flanke). Als Faustregel kann man verwenden, dass ein Schiefewert, der mehr als doppelt so groß ist wie sein Standardfehler, als Abweichung von der Symmetrie gilt.

Standardabweichung. Ein Maß für die Streuung um den Mittelwert, definiert als Quadratwurzel aus der Varianz. Die Standardabweichung wird in denselben Einheiten gemessen wie die ursprüngliche Variable.

Standardabweichung. Ein Maß für die Streuung um den Mittelwert. In einer Normalverteilung liegen 68 % der Fälle innerhalb von einer Standardabweichung des Mittelwerts und 95 % der Fälle innerhalb von zwei Standardabweichungen. Wenn beispielsweise für das Alter der Mittelwert 45 und die Standardabweichung 10 beträgt, liegen bei einer Normalverteilung 95 % der Fälle im Bereich zwischen 25 und 65.

Standardfehler. Ein Maß für die Abweichung des Werts einer Teststatistik zwischen Stichproben. Dies ist die Standardabweichung der Stichprobenverteilung einer Statistik. So ist z. B. der Standardfehler des Mittelwerts die Standardabweichung des Stichprobenmittelwerts.

Standardfehler der Kurtosis. Der Quotient aus der Kurtosis und deren Standardfehler kann als Test auf Normalverteilung verwendet werden. (Sie können die Normalverteilung ausschließen, wenn der Quotient unter -2 oder über +2 liegt.) Ein großer positiver Wert für die Kurtosis deutet darauf hin, dass die Flanken der Verteilung länger sind als bei einer Normalverteilung; ein negativer Wert bedeutet, dass sie kürzer sind (etwa wie bei einer kastenförmigen, gleichförmigen Verteilung).

Standardfehler des Mittelwerts. Ein Maß dafür, wie stark der Mittelwert von Stichprobe zu Stichprobe in derselben Verteilung variieren kann. Dieser Wert kann für einen ungefähren Vergleich des beobachteten Mittelwerts mit einem hypothetischen Wert verwendet werden. (Es kann geschlossen werden, dass die beiden Werte unterschiedlich sind, wenn das Verhältnis der Differenz zum Standardfehler kleiner als -2 oder größer als +2 ist.)

Standardfehler der Schiefe. Der Quotient aus der Schiefe und deren Standardfehler kann als Test auf Normalverteilung verwendet werden. (Sie können die Normalverteilung ausschließen, wenn der Quotient unter -2 oder über +2 liegt.) Ein großer positiver Wert für die Schiefe bedeutet, dass die Verteilung eine lange rechte Flanke hat; ein extremer negativer Wert bedeutet, dass sie eine lange linke Flanke hat.

Summe. Die Summe der Werte über alle Fälle mit nicht fehlenden Werten.

V

Varianz. Ein Maß der Streuung um den Mittelwert, gleich der Summe der quadrierten Abweichungen vom Mittelwert geteilt durch eins weniger als die Anzahl der Fälle. Die Maßeinheit der Varianz ist das Quadrat der Maßeinheiten der Variablen.

Index

Sonderzeichen

Numerische Stichwörter

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Z