

IBM SPSS Modeler 14.2 Source, Process, and Output Nodes



Note: Before using this information and the product it supports, read the general information under Notices on p. 432.

This edition applies to IBM SPSS Modeler 14 and to all subsequent releases and modifications until otherwise indicated in new editions.

Adobe product screenshot(s) reprinted with permission from Adobe Systems Incorporated.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.

Licensed Materials - Property of IBM

© Copyright IBM Corporation 1994, 2011.

U.S. Government Users Restricted Rights - Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Preface

IBM® SPSS® Modeler is the IBM Corp. enterprise-strength data mining workbench. SPSS Modeler helps organizations to improve customer and citizen relationships through an in-depth understanding of data. Organizations use the insight gained from SPSS Modeler to retain profitable customers, identify cross-selling opportunities, attract new customers, detect fraud, reduce risk, and improve government service delivery.

SPSS Modeler's visual interface invites users to apply their specific business expertise, which leads to more powerful predictive models and shortens time-to-solution. SPSS Modeler offers many modeling techniques, such as prediction, classification, segmentation, and association detection algorithms. Once models are created, IBM® SPSS® Modeler Solution Publisher enables their delivery enterprise-wide to decision makers or to a database.

About IBM Business Analytics

IBM Business Analytics software delivers complete, consistent and accurate information that decision-makers trust to improve business performance. A comprehensive portfolio of [business intelligence](#), [predictive analytics](#), [financial performance and strategy management](#), and [analytic applications](#) provides clear, immediate and actionable insights into current performance and the ability to predict future outcomes. Combined with rich industry solutions, proven practices and professional services, organizations of every size can drive the highest productivity, confidently automate decisions and deliver better results.

As part of this portfolio, IBM SPSS Predictive Analytics software helps organizations predict future events and proactively act upon that insight to drive better business outcomes. Commercial, government and academic customers worldwide rely on IBM SPSS technology as a competitive advantage in attracting, retaining and growing customers, while reducing fraud and mitigating risk. By incorporating IBM SPSS software into their daily operations, organizations become predictive enterprises – able to direct and automate decisions to meet business goals and achieve measurable competitive advantage. For further information or to reach a representative visit <http://www.ibm.com/spss>.

Technical support

Technical support is available to maintenance customers. Customers may contact Technical Support for assistance in using IBM Corp. products or for installation help for one of the supported hardware environments. To reach Technical Support, see the IBM Corp. web site at <http://www.ibm.com/support>. Be prepared to identify yourself, your organization, and your support agreement when requesting assistance.

Contents

1	About IBM SPSS Modeler	1
	IBM SPSS Modeler Server	1
	IBM SPSS Modeler Options	1
	IBM SPSS Text Analytics	2
	IBM SPSS Modeler Documentation	2
	Application Examples	3
	Demos Folder	4
2	Source Nodes	5
	Overview	5
	Enterprise View Node	6
	Setting Options for the Enterprise View Node	7
	Enterprise View Connections	9
	Choosing the DPD	10
	Choosing the Table	11
	Database Source Node	11
	Setting Database Node Options	13
	Adding a Database Connection	14
	Specifying preset values for a database connection	15
	Selecting a Database Table	16
	Querying the Database	17
	Variable File Node	18
	Setting Options for the Variable File Node	19
	Fixed File Node	21
	Setting Options for the Fixed File Node	21
	Setting Field Storage and Formatting	23
	Data Collection Node	26
	Data Collection Import File Options	27
	IBM SPSS Data Collection Import Metadata Properties	30
	Database Connection String	31
	Advanced Properties	32
	Importing Multiple Response Sets	32
	IBM SPSS Data Collection Column Import Notes	33
	IBM Cognos BI Source Node	34
	Cognos connections	36
	Cognos package selection	36
	SAS Source Node	37
	Setting Options for the SAS Source Node	38

Excel Source Node	39
XML Source Node	40
Selecting from Multiple Root Elements	42
Removing Unwanted Spaces from XML Source Data	43
User Input Node	44
Setting Options for the User Input Node	45
Common Source Node Tabs	50
Setting Measurement Levels in the Source Node	50
Filtering Fields from the Source Node	52

3 Record Operations Nodes 53

Overview of Record Operations	53
Select Node	54
Sample Node	55
Sample Node Options	56
Cluster and Stratify Settings	59
Sample Sizes for Strata	61
Balance Node	63
Setting Options for the Balance Node	64
Aggregate Node	65
Setting Options for the Aggregate Node	66
RFM Aggregate Node	67
Setting Options for the RFM Aggregate Node	68
Sort Node	69
Sort Optimization Settings	70
Merge Node	71
Types of Joins	72
Specifying a Merge Method and Keys	74
Selecting Data for Partial Joins	75
Filtering Fields from the Merge Node	76
Setting Input Order and Tagging	77
Merge Optimization Settings	78
Append Node	80
Setting Append Options	81
Distinct Node	81
Distinct Optimization Settings	83

4 *Field Operations Nodes*

85

Field Operations Overview	85
Automated Data Preparation	87
Fields Tab	90
Settings Tab	90
Field Settings.	91
Prepare Dates & Times	92
Exclude Fields	93
Preparing Inputs and Targets.	94
Construction and Feature Selection	96
Field Names.	98
Analysis Tab	99
Field Processing Summary	101
Fields	102
Action Summary	104
Predictive Power	105
Fields Table	106
Field Details	107
Action Details	109
Generating a Derive Node	112
Type Node	113
Measurement Levels	115
Converting Continuous Data.	117
What Is Instantiation?	118
Data Values.	119
Defining Missing Values	124
Checking Type Values	124
Setting the Field Role.	126
Copying Type Attributes.	127
Field Format Settings Tab.	128
Filtering or Renaming Fields	130
Setting Filtering Options.	131
Ensemble Node	136
Ensemble Node Settings	137
Derive Node	139
Setting Basic Options for the Derive Node	141
Deriving Multiple Fields	142
Setting Derive Formula Options	144
Setting Derive Flag Options	144
Setting Derive Set Options	146
Setting Derive State Options	147

Setting Derive Count Options	148
Setting Derive Conditional Options	149
Recoding Values with the Derive Node	150
Filler Node	151
Storage Conversion Using the Filler Node	153
Anonymize Node	154
Setting Options for the Anonymize Node	155
Anonymizing Field Values.	157
Reclassify Node	158
Setting Options for the Reclassify Node	159
Reclassifying Multiple Fields	161
Storage and Measurement Level for Reclassified Fields.	162
Binning Node	162
Setting Options for the Binning Node	163
Fixed-Width Bins	164
Tiles (Equal Count or Sum)	165
Rank Cases	168
Mean/Standard Deviation	169
Optimal Binning	170
Previewing the Generated Bins	171
RFM Analysis Node	173
RFM Analysis Node Settings	174
RFM Analysis Node Binning.	175
Partition Node	176
Partition Node Options.	177
Set to Flag Node	178
Setting Options for the Set to Flag Node.	179
Restructure Node	180
Setting Options for the Restructure Node.	181
Transpose Node	182
Setting Options for the Transpose Node	182
Time Intervals Node	186
Specifying Time Intervals.	187
Time Interval Build Options	188
Estimation Period.	190
Forecasts	191
Supported Intervals	194
History Node	205
Setting Options for the History Node	205
Field Reorder Node	206
Setting Field Reorder Options.	206

5 *Graph Nodes*

209

Common Graph Nodes Features	209
Aesthetics, Overlays, Panels, and Animation	210
Using the Output Tab	215
Using the Annotations Tab	215
3-D Graphs	215
Graphboard Node	216
Graphboard Basic Tab	218
Graphboard Detailed Tab	223
Graphboard Appearance Tab	224
Setting the Location of Templates and Stylesheets	225
Managing Templates and Stylesheets	227
Plot Node	229
Plot Node Tab	231
Plot Options Tab	234
Plot Appearance Tab	235
Using a Plot Graph	236
Distribution Node	237
Distribution Plot Tab	238
Distribution Appearance Tab	238
Using a Distribution Node	239
Histogram Node	242
Histogram Plot Tab	242
Histogram Options Tab	243
Histogram Appearance Tab	243
Using Histograms	244
Collection Node	245
Collection Plot Tab	246
Collection Options Tab	247
Collection Appearance Tab	248
Using a Collection Graph	249
Multiplot Node	250
Multiplot Plot Tab	251
Multiplot Appearance Tab	253
Using a Multiplot Graph	254
Web Node	254
Web Plot Tab	256
Web Options Tab	257
Web Appearance Tab	260
Using a Web Graph	261

Time Plot Node	266
Time Plot Tab	267
Time Plot Appearance Tab	268
Using a Time Plot Graph	269
Evaluation Node	269
Evaluation Plot Tab	274
Evaluation Options Tab	275
Evaluation Appearance Tab	277
Reading the Results of a Model Evaluation	279
Using an Evaluation Chart	280
Exploring Graphs	281
Using Bands	282
Using Regions	285
Using Marked Elements	287
Generating Nodes from Graphs	288
Editing Visualizations	291
General Rules for Editing Visualizations	292
Editing and Formatting Text	293
Changing Colors, Patterns, Dashings, and Transparency	294
Rotating and Changing the Shape and Aspect Ratio of Point Elements	295
Changing the Size of Graphic Elements	295
Specifying Margins and Padding	296
Formatting Numbers	297
Changing the Axis and Scale Settings	298
Editing Categories	299
Changing the Orientation Panels	301
Transforming the Coordinate System	302
Changing Statistics and Graphic Elements	303
Changing the Position of the Legend	306
Copying a Visualization and Visualization Data	306
Keyboard Shortcuts	307
Adding Titles and Footnotes	307
Using Graph Stylesheets	309
Applying Stylesheets	310
Printing, Saving, Copying, and Exporting Graphs	311

6 Output Nodes 314

Overview of Output Nodes	314
Managing Output	315

Viewing Output	316
Publish to Web	316
Viewing Output in an HTML Browser	319
Exporting Output	319
Selecting Cells and Columns	320
Table Node	321
Table Node Settings Tab	321
Table Node Format Tab	322
Output Node Output Tab.	322
Table Browser	324
Matrix Node	325
Matrix Node Settings Tab	326
Matrix Node Appearance Tab	327
Matrix Node Output Browser.	328
Analysis Node	330
Analysis Node Analysis Tab	330
Analysis Output Browser	332
Data Audit Node	335
Data Audit Node Settings Tab	336
Data Audit Quality Tab	338
Data Audit Output Browser	340
Transform Node	349
Transform Node Options Tab	350
Transform Node Output Tab	351
Transform Node Output Viewer	351
Statistics Node	355
Statistics Node Settings Tab	355
Statistics Output Browser	356
Means Node.	359
Comparing Means for Independent Groups	359
Comparing Means Between Paired Fields	360
Means Node Options.	361
Means Node Output Browser	362
Report Node	364
Report Node Template Tab.	365
Report Node Output Browser.	367
Set Globals Node	368
Set Globals Node Settings Tab.	368
IBM SPSS Statistics Helper Applications	369

7 Export Nodes

371

Overview of Export Nodes	371
Database Export Node	372
Database Node Export Tab.	372
Database Export Merge Options	373
Database Export Schema Options	375
Database Export Index Options	378
Database Export Advanced Options.	380
Flat File Export Node.	382
Flat File Export Tab.	382
IBM SPSS Data Collection Export Node	383
IBM Cognos BI Export Node	385
Cognos connection	385
ODBC connection	387
SAS Export Node	388
SAS Export Node Export Tab	389
Excel Export Node	389
Excel Node Export Tab.	390
XML Export Node	391
Writing XML Data	392
XML Mapping Records Options	393
XML Mapping Fields Options	393
XML Mapping Preview	394

8 IBM SPSS Statistics Nodes

395

IBM SPSS Statistics Nodes - Overview.	395
Statistics File Node.	396
Statistics Transform Node.	397
Statistics Transform Node - Syntax Tab	398
Allowable Syntax.	399
Statistics Model Node	401
Statistics Model Node - Model Tab	402
Statistics Model Node - Model Nugget Summary.	403
Statistics Output Node	405
Statistics Output Node - Syntax Tab	406
Statistics Output Node - Output Tab	408

Statistics Export Node	409
Statistics Export Node - Export Tab	409
Renaming or Filtering Fields for IBM SPSS Statistics	410

9 SuperNodes 412

Overview of SuperNodes	412
Types of SuperNodes	412
Source SuperNodes	412
Process SuperNodes	413
Terminal SuperNodes	414
Creating SuperNodes	414
Nesting SuperNodes	416
Examples of Valid SuperNodes	417
Examples of Invalid SuperNodes	418
Locking SuperNodes	419
Locking and Unlocking a SuperNode	420
Editing a Locked SuperNode	421
Editing SuperNodes	422
Modifying SuperNode Types	422
Annotating and Renaming SuperNodes	423
SuperNode Parameters	424
SuperNodes and Caching	428
SuperNodes and Scripting	429
Saving and Loading SuperNodes	430

Appendix

A Notices 432

Index 435

About IBM SPSS Modeler

IBM® SPSS® Modeler is a set of data mining tools that enable you to quickly develop predictive models using business expertise and deploy them into business operations to improve decision making. Designed around the industry-standard CRISP-DM model, SPSS Modeler supports the entire data mining process, from data to better business results.

SPSS Modeler offers a variety of modeling methods taken from machine learning, artificial intelligence, and statistics. The methods available on the Modeling palette allow you to derive new information from your data and to develop predictive models. Each method has certain strengths and is best suited for particular types of problems.

SPSS Modeler can be purchased as a standalone product, or used in combination with SPSS Modeler Server. A number of additional options are also available, as summarized in the following sections. For more information, see <http://www.ibm.com/software/analytics/spss/products/modeler/>.

IBM SPSS Modeler Server

SPSS Modeler uses a client/server architecture to distribute requests for resource-intensive operations to powerful server software, resulting in faster performance on larger data sets. Additional products or updates beyond those listed here may also be available. For more information, see <http://www.ibm.com/software/analytics/spss/products/modeler/>.

SPSS Modeler. SPSS Modeler is a functionally complete version of the product that is installed and run on the user's desktop computer. It can be run in local mode as a standalone product or in distributed mode along with IBM® SPSS® Modeler Server for improved performance on large data sets.

SPSS Modeler Server. SPSS Modeler Server runs continually in distributed analysis mode together with one or more IBM® SPSS® Modeler installations, providing superior performance on large data sets because memory-intensive operations can be done on the server without downloading data to the client computer. SPSS Modeler Server also provides support for SQL optimization and in-database modeling capabilities, delivering further benefits in performance and automation. At least one SPSS Modeler installation must be present to run an analysis.

IBM SPSS Modeler Options

The following components and features can be separately purchased and licensed for use with SPSS Modeler. Note that additional products or updates may also become available. For more information, see <http://www.ibm.com/software/analytics/spss/products/modeler/>.

- SPSS Modeler Server access, providing improved scalability and performance on large data sets, as well as support for SQL optimization and in-database modeling capabilities.

- SPSS Modeler Solution Publisher, for real-time or automated scoring outside the SPSS Modeler environment.
- Adapters to enable deployment to IBM SPSS Collaboration and Deployment Services or the thin-client application IBM SPSS Modeler Advantage.

IBM SPSS Text Analytics

IBM® SPSS® Text Analytics is a fully integrated add-on for SPSS Modeler that uses advanced linguistic technologies and Natural Language Processing (NLP) to rapidly process a large variety of unstructured text data, extract and organize the key concepts, and group these concepts into categories. Extracted concepts and categories can be combined with existing structured data, such as demographics, and applied to modeling using the full suite of IBM® SPSS® Modeler data mining tools to yield better and more focused decisions.

- The Text Mining node offers concept and category modeling, as well as an interactive workbench where you can perform advanced exploration of text links and clusters, create your own categories, and refine the linguistic resource templates.
- A number of import formats are supported, including blogs and other web-based sources.
- Custom templates, libraries, and dictionaries for specific domains, such as CRM and genomics, are also included.

Note: A separate license is required to access this component. For more information, see <http://www.ibm.com/software/analytics/spss/products/modeler/>.

IBM SPSS Modeler Documentation

Complete documentation in online help format is available from the Help menu of SPSS Modeler. This includes documentation for SPSS Modeler, SPSS Modeler Server, and SPSS Modeler Solution Publisher, as well as the Applications Guide and other supporting materials.

Complete documentation for each product in PDF format is available under the *\Documentation* folder on each product DVD.

- **IBM SPSS Modeler User's Guide.** General introduction to using SPSS Modeler, including how to build data streams, handle missing values, build CLEM expressions, work with projects and reports, and package streams for deployment to IBM SPSS Collaboration and Deployment Services, Predictive Applications, or IBM SPSS Modeler Advantage.
- **IBM SPSS Modeler Source, Process, and Output Nodes.** Descriptions of all the nodes used to read, process, and output data in different formats. Effectively this means all nodes other than modeling nodes.
- **IBM SPSS Modeler Modeling Nodes.** Descriptions of all the nodes used to create data mining models. IBM® SPSS® Modeler offers a variety of modeling methods taken from machine learning, artificial intelligence, and statistics.
- **IBM SPSS Modeler Algorithms Guide.** Descriptions of the mathematical foundations of the modeling methods used in SPSS Modeler.

- **IBM SPSS Modeler Applications Guide.** The examples in this guide provide brief, targeted introductions to specific modeling methods and techniques. An online version of this guide is also available from the Help menu. For more information, see the topic [Application Examples](#) on p. 3.
- **IBM SPSS Modeler Scripting and Automation.** Information on automating the system through scripting, including the properties that can be used to manipulate nodes and streams.
- **IBM SPSS Modeler Deployment Guide.** Information on running SPSS Modeler streams and scenarios as steps in processing jobs under IBM® SPSS® Collaboration and Deployment Services Deployment Manager.
- **IBM SPSS Modeler CLEF Developer's Guide.** CLEF provides the ability to integrate third-party programs such as data processing routines or modeling algorithms as nodes in SPSS Modeler.
- **IBM SPSS Modeler In-Database Mining Guide.** Information on how to use the power of your database to improve performance and extend the range of analytical capabilities through third-party algorithms.
- **IBM SPSS Modeler Server and Performance Guide.** Information on how to configure and administer IBM® SPSS® Modeler Server.
- **IBM SPSS Modeler Administration Console User Guide.** Information on installing and using the console user interface for monitoring and configuring SPSS Modeler Server. The console is implemented as a plug-in to the Deployment Manager application.
- **IBM SPSS Modeler Solution Publisher Guide.** SPSS Modeler Solution Publisher is an add-on component that enables organizations to publish streams for use outside of the standard SPSS Modeler environment.
- **IBM SPSS Modeler CRISP-DM Guide.** Step-by-step guide to using the CRISP-DM methodology for data mining with SPSS Modeler.

Application Examples

While the data mining tools in SPSS Modeler can help solve a wide variety of business and organizational problems, the application examples provide brief, targeted introductions to specific modeling methods and techniques. The data sets used here are much smaller than the enormous data stores managed by some data miners, but the concepts and methods involved should be scalable to real-world applications.

You can access the examples by clicking [Application Examples](#) on the Help menu in SPSS Modeler. The data files and sample streams are installed in the *Demos* folder under the product installation directory. For more information, see the topic [Demos Folder](#) on p. 4.

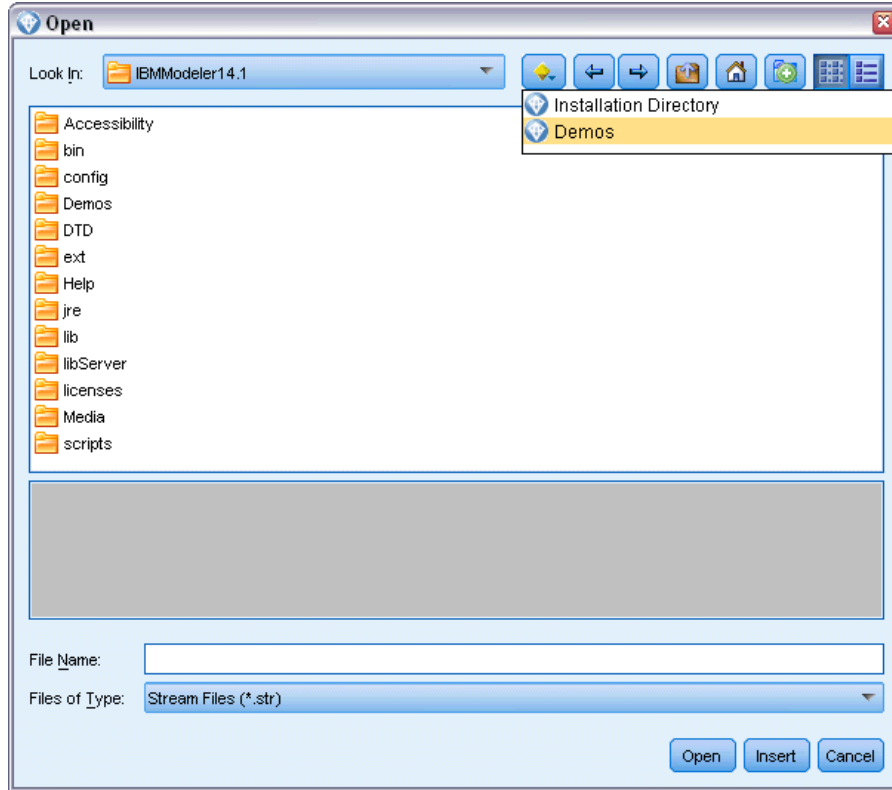
Database modeling examples. See the examples in the *IBM SPSS Modeler In-Database Mining Guide*.

Scripting examples. See the examples in the *IBM SPSS Modeler Scripting and Automation Guide*.

Demos Folder

The data files and sample streams used with the application examples are installed in the *Demos* folder under the product installation directory. This folder can also be accessed from the IBM SPSS Modeler 14.2 program group on the Windows Start menu, or by clicking *Demos* on the list of recent directories in the File Open dialog box.

Figure 1-1
Selecting the Demos folder from the list of recently-used directories



Source Nodes

Overview

Source nodes enable you to import data stored in a number of formats, including flat files, IBM® SPSS® Statistics (.sav), SAS, Microsoft Excel, and ODBC-compliant relational databases. You can also generate synthetic data using the User Input node.

The Sources palette contains the following nodes:



The Enterprise View node creates a connection to an IBM® SPSS® Collaboration and Deployment Services Repository, enabling you to read Enterprise View data into a stream and to package a model in a scenario that can be accessed from the repository by other users. For more information, see the topic [Enterprise View Node](#) on p. 6.



The Database node can be used to import data from a variety of other packages using ODBC (Open Database Connectivity), including Microsoft SQL Server, DB2, Oracle, and others. For more information, see the topic [Database Source Node](#) on p. 11.



The Variable File node reads data from free-field text files—that is, files whose records contain a constant number of fields but a varied number of characters. This node is also useful for files with fixed-length header text and certain types of annotations. For more information, see the topic [Variable File Node](#) on p. 18.



The Fixed File node imports data from fixed-field text files—that is, files whose fields are not delimited but start at the same position and are of a fixed length. Machine-generated or legacy data are frequently stored in fixed-field format. For more information, see the topic [Fixed File Node](#) on p. 21.



The Statistics File node reads data from the .sav file format used by SPSS Statistics, as well as cache files saved in IBM® SPSS® Modeler, which also use the same format. For more information, see the topic [Statistics File Node](#) in Chapter 8 on p. 396.



The IBM® SPSS® Data Collection node imports survey data from various formats used by market research software conforming to the Data Collection Data Model. The Data Collection Developer Library must be installed to use this node. For more information, see the topic [Data Collection Node](#) on p. 26.



The IBM Cognos BI source node imports data from Cognos BI databases. For more information, see the topic [IBM Cognos BI Source Node](#) on p. 34.



The SAS File node imports SAS data into SPSS Modeler. For more information, see the topic [SAS Source Node](#) on p. 37.



The Excel node imports data from any version of Microsoft Excel. An ODBC data source is not required. For more information, see the topic [Excel Source Node](#) on p. 39.



The XML source node imports data in XML format into the stream. You can import a single file, or all files in a directory. You can optionally specify a schema file from which to read the XML structure. For more information, see the topic [XML Source Node](#) on p. 40.



The User Input node provides an easy way to create synthetic data—either from scratch or by altering existing data. This is useful, for example, when you want to create a test dataset for modeling. For more information, see the topic [User Input Node](#) on p. 44.

To begin a stream, add a source node to the stream canvas. Next, double-click the node to open its dialog box. The various tabs in the dialog box allow you to read in data; view the fields and values; and set a variety of options, including filters, data types, field role, and missing-value checking.

Enterprise View Node

The Enterprise View node enables you to create and maintain a connection between an IBM® SPSS® Modeler session and an Enterprise View in a shared IBM® SPSS® Collaboration and Deployment Services Repository. Doing so enables you to read data from an Enterprise View into an SPSS Modeler stream and to package an SPSS Modeler model in a scenario that can be accessed by other users of the shared repository.

A **scenario** is a file containing an SPSS Modeler stream with specific nodes, models, and additional properties that enable it to be deployed to an IBM SPSS Collaboration and Deployment Services Repository for the purposes of scoring or automatic model refresh. The use of Enterprise View nodes with scenarios ensures that, in a multi-user situation, all users are working from the same data. A **connection** is a link from an SPSS Modeler session to an Enterprise View in the IBM SPSS Collaboration and Deployment Services Repository.

The **Enterprise View** is the complete set of the data belonging to an organization, irrespective of where the data is physically located. Each connection consists of a specific selection of a single **Application View** (subset of the Enterprise View tailored for a particular application), a **Data Provider Definition** (DPD—links the logical Application View tables and columns to a physical data source), and an **environment** (identifies which particular columns should be associated with defined business segments). The Enterprise View, Application Views, and DPD definitions are stored and versioned in the repository, although the actual data resides in one or more databases or other external sources.

Once a connection has been established, you specify an Application View **table** to work with in SPSS Modeler. In an Application View, a table is a logical view consisting of some or all columns from one or more physical tables in one or more physical databases. Thus the Enterprise View node enables records from multiple database tables to be seen as a single table in SPSS Modeler.

Requirements

- To use the Enterprise View node, an IBM SPSS Collaboration and Deployment Services Repository must first be installed and configured at your site, with an Enterprise View, Application Views, and DPDs already defined.

Note: A separate license is required to access an IBM® SPSS® Collaboration and Deployment Services repository. For more information, see <http://www.ibm.com/software/analytics/spss/products/deployment/cds/>

- In addition, the IBM® SPSS® Collaboration and Deployment Services Enterprise View Driver must be installed on each computer used to modify or run the stream. For Windows, simply install the driver on the computer where IBM® SPSS® Modeler or IBM® SPSS® Modeler Server is installed, and no further configuration of the driver is needed. On UNIX, a reference to the *pev.sh* script must be added to the startup script. Contact your local administrator for details on installing the IBM SPSS Collaboration and Deployment Services Enterprise View Driver.
- A DPD is defined against a particular ODBC datasource. To use a DPD from SPSS Modeler, you must have an ODBC datasource defined on the SPSS Modeler server host which has the same name, and which connects to the same data store, as the one referenced in the DPD.

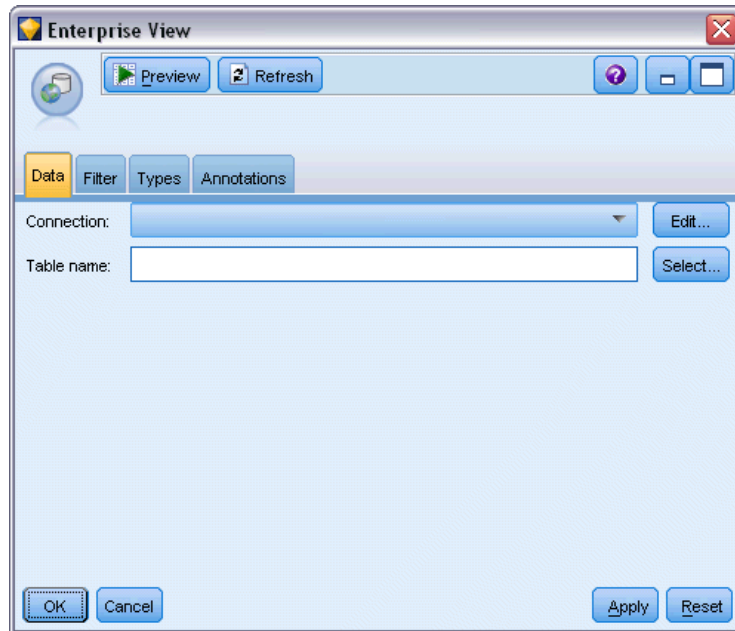
Setting Options for the Enterprise View Node

You can use the options on the Data tab of the Enterprise View dialog box to:

- Select an existing repository connection
- Edit an existing repository connection
- Create a new repository connection
- Select an Application View table

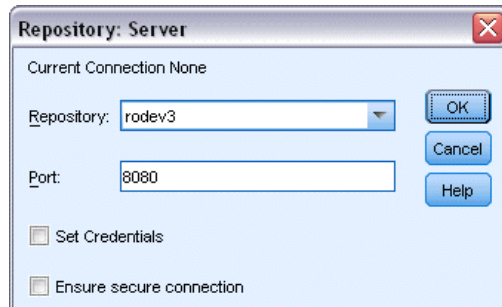
Refer to the *IBM® SPSS® Collaboration and Deployment Services Administrator's Guide* for details on working with repositories.

Figure 2-1
Adding a connection to an IBM SPSS Collaboration and Deployment Services Repository



Connection. The drop-down list provides options for selecting an existing repository connection, editing an existing connection, or adding a connection. If you are already logged in to a repository through IBM® SPSS® Modeler, choosing the Add/Edit a connection option displays the Enterprise View Connections dialog box, from where you can define or edit the required details for the current connection. If you are not logged in, this option displays the repository Login dialog box.

Figure 2-2
Logging in to a repository



For information on logging in to the repository, see the *SPSS Modeler User's Guide*.

Once a connection to a repository has been established, that connection remains in place until you exit from SPSS Modeler. A connection can be shared by other nodes within the same stream, but you must create a new connection for each new stream.

A successful login displays the Enterprise View Connections dialog box.

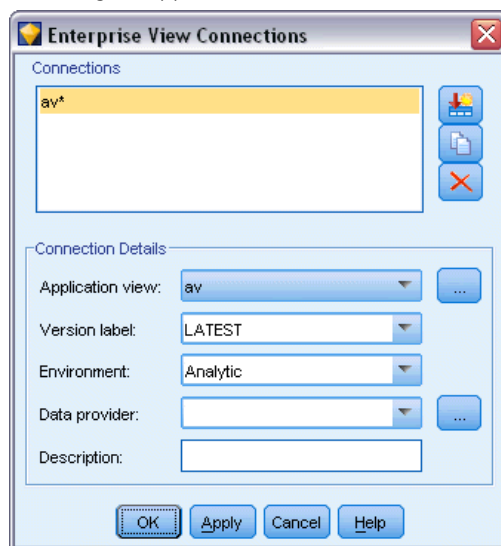
Table name. This field is initially empty and cannot be populated until you create a connection. If you know the name of the Application View table you would like to access, enter it in the Table Name field. Otherwise, click the Select button to open a dialog box listing the available Application View tables.

Enterprise View Connections

This dialog box enables you to define or edit the required details for the repository connection. You can specify the:

- Application View and version
- Environment
- Data Provider Definition (DPD)
- Connection description

Figure 2-3
Choosing an application view



Connections. Lists existing repository connections.

- **Add a new connection.** Displays the Retrieve Object dialog box, from where you can search for and select an Application View from the repository.
- **Copy the selected connection.** Makes a copy of a selected connection, saving you from having to browse again for the same Application View.
- **Delete the selected connection.** Deletes the selected connection from the list.

Connection Details. For the connection currently selected in the Connections pane, displays the Application View, version label, environment, DPD, and descriptive text.

- **Application view.** The drop-down list displays the selected application view, if any. If connections have been made to other Application Views in the current session, these also appear on the drop-down list. Click the adjacent Browse button to search for other Application Views in the repository.
- **Version label.** The drop-down field lists all defined version labels for the specified Application View. Version labels help identify specific repository object versions. For example, there may be two versions for a particular Application View. By using labels, you could specify the label TEST for the version used in the development environment and the label PRODUCTION for the version used in the production environment. Select an appropriate label.

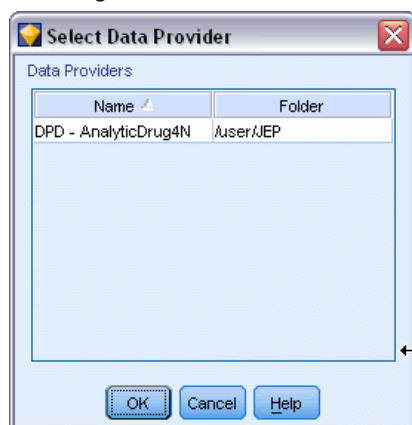
Note: Labels should not include the “[” character, otherwise the table name will not be displayed on the Data tab of the Enterprise View dialog box.

- **Environment.** The drop-down field lists all valid environments. The environment setting determines which DPDs are available, thus specifying which particular columns should be associated with defined business segments. For example, when Analytic is selected, only those Application View columns defined as Analytic are returned. The default environment is Analytic; you can also choose Operational.
- **Data provider.** The drop-down list displays the names of up to ten Data Provider Definitions for the selected Application View. Only DPDs that reference the selected Application View are shown. Click the adjacent Browse button to view the name and path of all DPDs related to the current Application View.
- **Description.** Descriptive text about the repository connection. This text will be used for the connection name—clicking OK causes the text to appear on the Connection drop-down list and title bar of the Enterprise View dialog box, and as the label of the Enterprise View node on the canvas.

Choosing the DPD

The Select Data Provider dialog box shows the name and path of all the DPDs that reference the current Application View.

Figure 2-4
Choosing a DPD



Application Views can have multiple DPDs in order to support the different stages of a project. For example, the historic data used to build a model may come from one database, while operational data comes from another.

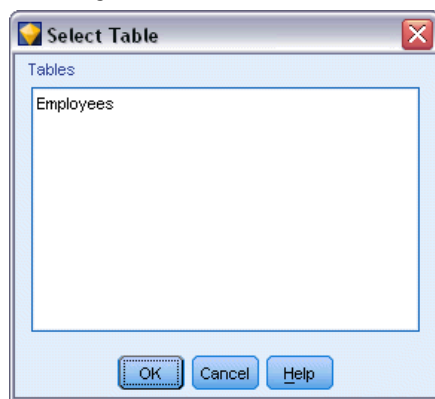
A DPD is defined against a particular ODBC datasource. To use a DPD from IBM® SPSS® Modeler, you must have an ODBC datasource defined on the SPSS Modeler server host which has the same name, and which connects to the same data store, as the one referenced in the DPD.

- ▶ To choose a DPD to work with, select its name on the list and click OK.

Choosing the Table

The Select Table dialog box lists all the tables that are referenced in the current Application View. The dialog box is empty if no connection has been made to an IBM SPSS Collaboration and Deployment Services Repository.

Figure 2-5
Choosing a table



- ▶ To choose a table to work with, select its name on the list and click OK.

Database Source Node

The Database source node can be used to import data from a variety of other packages using ODBC (Open Database Connectivity), including Microsoft SQL Server, DB2, Oracle, and others.

To read or write to a database, you must have an ODBC data source installed and configured for the relevant database, with read or write permissions as needed. The IBM® SPSS® Data Access Pack includes a set of ODBC drivers that can be used for this purpose, and these drivers are available on the IBM SPSS Data Access Pack installation disk shipped with this release. If you have questions about creating or setting permissions for ODBC data sources, contact your database administrator.

Database support in IBM® SPSS® Modeler is classified into three tiers, each representing a different level of support for SQL pushback and optimization, depending on the database vendor. The different levels of support are implemented by means of a number of system settings, which can be customized as part of a Services engagement.

The three tiers of database support are:

Table 2-1
Database support tiers

Support tier	Description
Tier 1	All possible SQL pushback is available, with database-specific SQL optimization.
Tier 2	Most SQL pushback is available, with non-database-specific SQL optimization.
Tier 3	No SQL pushback or optimization—only reading data from, and writing data to, the database.

Supported ODBC Drivers

For the latest information on which databases and ODBC drivers are supported and tested for use with SPSS Modeler 14.2, see the product compatibility matrices on the corporate Support site (<http://www.ibm.com/support>).

Where to Install Drivers

Note that ODBC drivers must be installed and configured on each computer where processing may occur.

- If you are running IBM® SPSS® Modeler in local (standalone) mode, the drivers must be installed on the local computer.
- If you are running SPSS Modeler in distributed mode against a remote IBM® SPSS® Modeler Server, the ODBC drivers need to be installed on the computer where SPSS Modeler Server is installed.
- If you need to access the same data sources from both SPSS Modeler and SPSS Modeler Server, the ODBC drivers must be installed on both computers.
- If you are running SPSS Modeler over Terminal Services, the ODBC drivers need to be installed on the Terminal Services server on which you have SPSS Modeler installed.
- If you are using the IBM® SPSS® Modeler Solution Publisher Runtime to run published streams on a separate computer, you also need to install and configure ODBC drivers on that computer.

Note: If you are using SPSS Modeler Server on UNIX to access a Teradata database you must use the ODBC Driver Manager that is installed with the Teradata ODBC driver. In order to make this change to SPSS Modeler Server please specify a value for ODBC_DRIVER_MANAGER_PATH near the top of the modelersrv.sh script where indicated by the comments. This environment variable needs to be set to the location of the ODBC Driver Manager that is shipped with the Teradata ODBC driver (/usr/odbc/lib in a Teradata ODBC driver default installation). You must restart SPSS Modeler Server for the change to take effect. For details of the SPSS Modeler Server platforms that offer support for Teradata access, and the Teradata ODBC driver version that is supported, see the corporate Support site at <http://www.ibm.com/support>.

Use the following general steps to access data from a database:

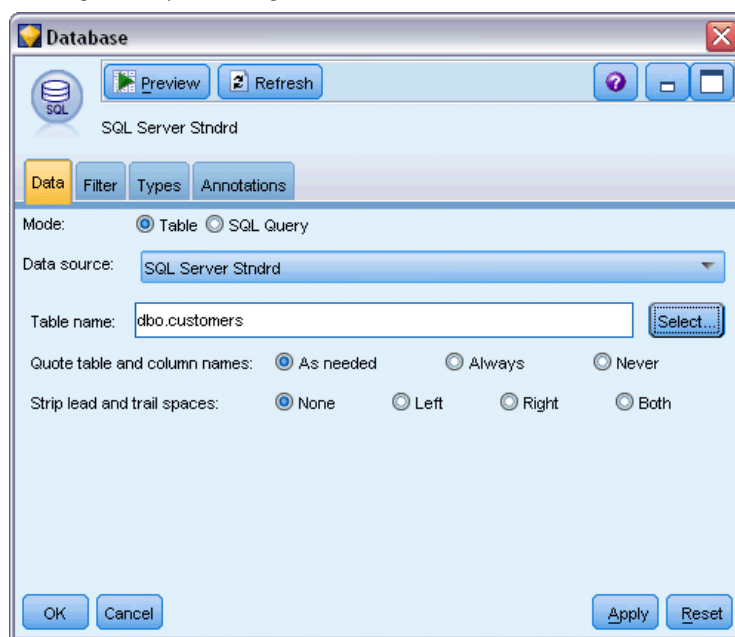
- ▶ Install an ODBC driver and configure a data source to the database you want to use.
- ▶ In the Database node dialog box, connect to a database using Table mode or SQL Query mode.

- ▶ Select a table from the database.
- ▶ Using the tabs in the Database node dialog box, you can alter usage types and filter data fields. These steps are described in more detail in the next several topics.

Setting Database Node Options

You can use the options on the Data tab of the Database source node dialog box to gain access to a database and read data from the selected table.

Figure 2-6
Loading data by selecting a table



Mode. Select Table to connect to a table using the dialog box controls. Select SQL Query to query the database selected below using SQL.

Data source. For both Table and SQL Query modes, you can enter a name in the Data Source field or select Add new database connection from the drop-down list.

The following options are used to connect to a database and select a table using the dialog box:

Table name. If you know the name of the table you would like to access, enter it in the Table Name field. Otherwise, click the Select button to open a dialog box listing the available tables.

Quote table and column names. Specify whether you want table and column names to be enclosed in quotation marks when queries are sent to the database (if, for example, they contain spaces or punctuation).

- The As needed option will quote table and field names *only* if they include nonstandard characters. Nonstandard characters include non-ASCII characters, space characters, and any non-alphanumeric character other than a full stop (.

- Select Never if you *never* want table and field names quoted.
- Select Always if you want *all* table and field names quoted.

Strip lead and trail spaces. Select options for discarding leading and trailing spaces in strings.

Note. Comparisons between strings that do and do not use SQL pushback may generate different results where trailing spaces exist.

Reading empty strings from Oracle. When reading from or writing to an Oracle database, be aware that, unlike IBM® SPSS® Modeler and unlike most other databases, Oracle treats and stores empty string values as equivalent to null values. This means that the same data extracted from an Oracle database may behave differently than when extracted from a file or another database, and the data may return different results.

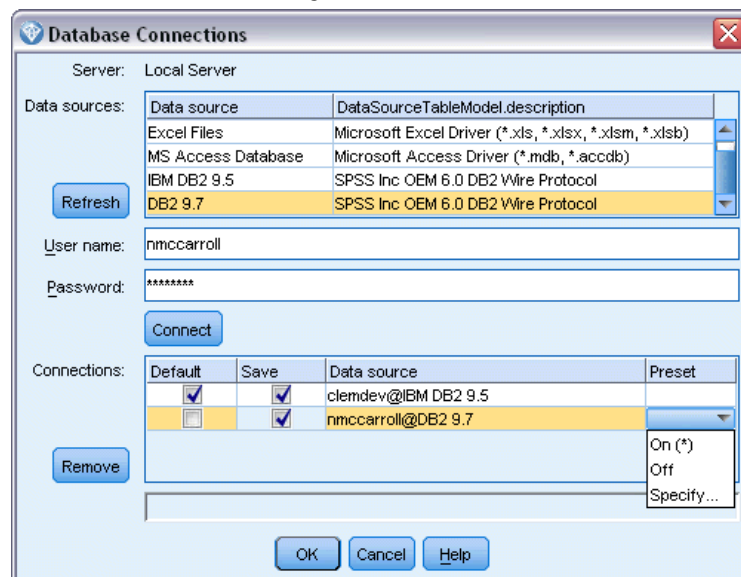
Adding a Database Connection

In order to open a database, you first have to select the data source to which you want to connect. On the Data tab, select Add new database connection from the Data Source drop-down list.

This opens the Database Connections dialog box.

Note: For an alternative way to open this dialog, from the main menu, choose: Tools > Databases...

Figure 2-7
Database Connections dialog box



Data sources. Lists the available data sources. Scroll down if you do not see the desired database. Once you have selected a data source and entered any passwords, click Connect. Click Refresh to update the list.

User name. If the data source is password protected, enter your user name.

Password. If the data source is password protected, enter your password.

Connections. Shows currently connected databases.

- **Default.** You can optionally choose one connection as the default. Doing so causes Database source or export nodes to have this connection predefined as their data source, though this can be edited if desired.
- **Save.** Optionally select one or more connections that you want to redisplay in subsequent sessions.
- **Data source.** The connection strings for the currently connected databases.
- **Preset.** Indicates (with a * character) whether preset values have been specified for the database connection. To specify preset values, click this column in the row corresponding to the database connection, and choose Specify from the list. For more information, see the topic [Specifying preset values for a database connection](#) on p. 15.

To remove connections, select one from the list and click Remove.

Once you have completed your selections, click OK.

Specifying preset values for a database connection

You can specify a number of default settings for a database connection.

IBM InfoSphere Warehouse settings

The InfoSphere Warehouse preset options all apply to database export.

Figure 2-8

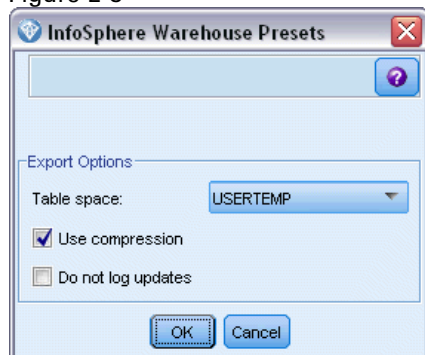


Table space. The tablespace to be used for export. Database administrators can create or configure tablespaces as partitioned. We recommend selecting one of these tablespaces (rather than the default one) to use for database export.

Use compression. If selected, creates tables for export with compression (for example, the equivalent of `CREATE TABLE MYTABLE(...) COMPRESS YES` in SQL).

Do not log updates. If selected, avoids logging when creating tables and inserting data (the equivalent of `CREATE TABLE MYTABLE(...) NOT LOGGED INITIALLY` in SQL).

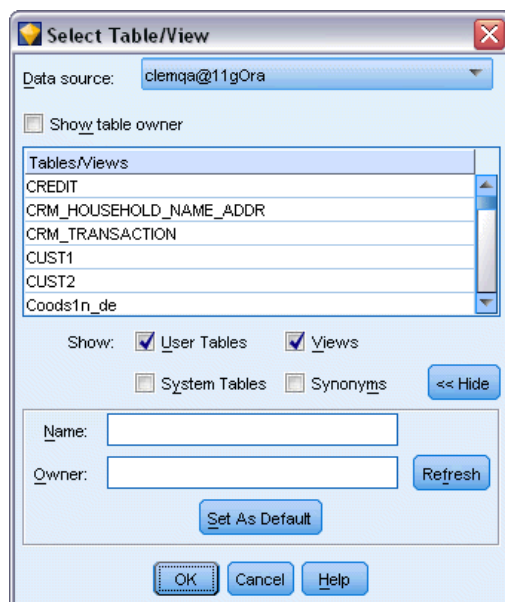
Settings for other databases

No presets are currently possible for other databases.

Selecting a Database Table

After you have connected to a data source, you can choose to import fields from a specific table or view. From the Data tab of the Database dialog box, you can either enter the name of a table in the Table Name field or click Select to open a dialog box listing the available tables and views.

Figure 2-9
Selecting a table from the currently connected database



Show table owner. Select if a data source requires that the owner of a table must be specified before you can access the table. Deselect this option for data sources that do not have this requirement.

Note: SAS and Oracle databases usually require you to show the table owner.

Tables/Views. Select the table or view to import.

Show. Lists the columns in the data source to which you are currently connected. Click one of the following options to customize your view of the available tables:

- Click **User Tables** to view ordinary database tables created by database users.
- Click **System Tables** to view database tables owned by the system (for example, tables that provide information about the database, such as details of indexes). This option can be used to view the tabs used in Excel databases. (Note that a separate Excel source node is also available. For more information, see the topic [Excel Source Node](#) on p. 39.)
- Click **Views** to view virtual tables based on a query involving one or more ordinary tables.
- Click **Synonyms** to view synonyms created in the database for any existing tables.

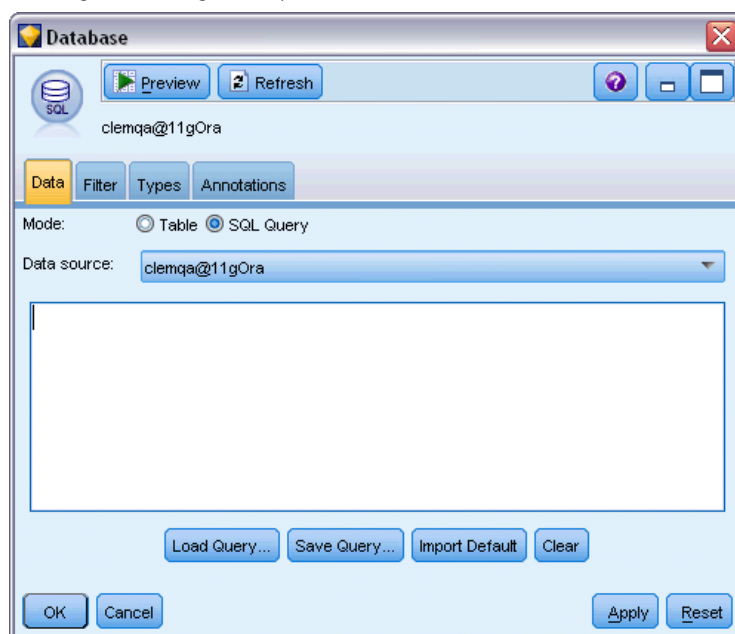
Name/Owner filters. These fields allow you to filter the list of displayed tables by name or owner. For example, type SYS to list only tables with that owner. For wildcard searches, an underscore () can be used to represent any single character and a percent sign (%) can represent any sequence of zero or more characters.

Set As Default. Saves the current settings as the default for the current user. These settings will be restored in the future when a user opens a new table selector dialog box *for the same data source name and user login only*.

Querying the Database

Once you have connected to a data source, you can choose to import fields using an SQL query. From the main dialog box, select SQL Query as the connection mode. This adds a query editor window in the dialog box. Using the query editor, you can create or load an SQL query whose result set will be read into the data stream. To cancel and close the query editor window, select Table as the connection mode.

Figure 2-10
Loading data using SQL queries



Load Query. Click to open the file browser, which you can use to load a previously saved query.

Save Query. Click to open the Save Query dialog box, which you can use to save the current query.

Import Default. Click to import an example SQL SELECT statement constructed automatically using the table and columns selected in the dialog box.

Clear. Clear the contents of the work area. Use this option when you want to start over.

Variable File Node

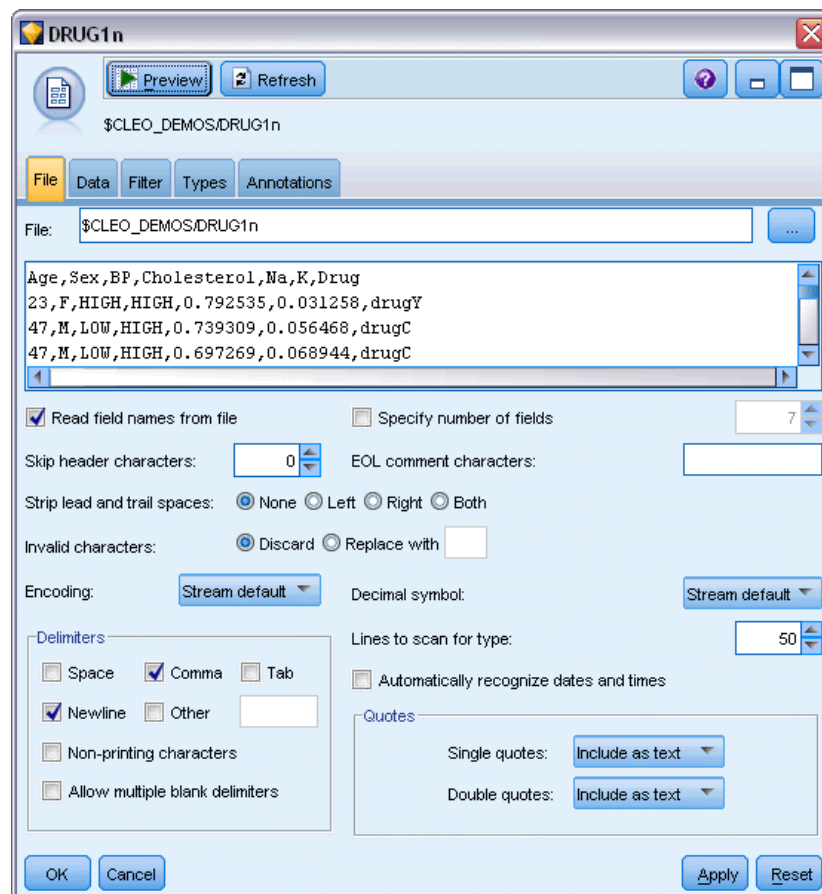
You can use Variable File nodes to read data from free-field text files (files whose records contain a constant number of fields but a varied number of characters), also known as delimited text files. This type of node is also useful for files with fixed-length header text and certain types of annotations. Records are read one at a time and passed through the stream until the entire file is read.

Notes for Reading in Delimited Text Data

- Records must be delimited by a newline character at the end of each line. The newline character must not be used for any other purpose (such as within any field name or value). Leading and trailing spaces should ideally be stripped off to save space, although this is not critical. Optionally these can be stripped out by the node.
- Fields must be delimited by a comma or other character that ideally is used only as a delimiter, meaning it does not appear within field names or values. If this is not possible, then all text fields can be wrapped in double quotes, provided that none of the field names or text values contains a double quote. If field names or values do contain double quotes, then text fields can be wrapped in single quotes as an alternative, again provided that single quotes are not used elsewhere within values. If neither single or double quotes can be used, then text values will need to be amended to remove or replace either the delimiter character, or single or double quotes.
- Each row, including the header row, should contain the same number of fields.
- The first line should contain the field names. If this is not the case, deselect Read field names from file to give each field a generic name such as *Field1*, *Field2*, and so on.
- The second line must contain the first record of data. There must no blank lines or comments.
- Numeric values must not include the thousands separator or grouping symbol—without the comma in 3,000.00, for example. The decimal indicator (period or full-stop in the U.S. or the UK) must only be used where appropriate.
- Date and time values should be in one of the formats recognized in the Stream Options dialog box, such as DD/MM/YYYY or HH:MM:SS. All dates and time fields in the file should ideally follow the same format, and any field containing a date must use the same format for all values within that field.

Setting Options for the Variable File Node

Figure 2-11
Variable File node dialog box



File. Specify the name of the file. You can enter a filename or click the ellipsis button (...) to select a file. The file path is shown once you have selected a file, and its contents are displayed with delimiters in the panel below it.

The sample text displayed from your data source can be copied and pasted into the following controls: EOL comment characters and user-specified delimiters. Use Ctrl-C and Ctrl-V to copy and paste.

Read field names from file. Selected by default, this option treats the first row in the data file as labels for the column. If your first row is not a header, deselect to automatically give each field a generic name, such as *Field1*, *Field2*, for the number of fields in the dataset.

Specify number of fields. Specify the number of fields in each record. The number of fields can be detected automatically as long as the records are new-line terminated. You can also set a number manually.

Skip header characters. Specify how many characters you want to ignore at the beginning of the first record.

EOL comment characters. Specify characters, such as # or !, to indicate annotations in the data. Wherever one of these characters appears in the data file, everything up to but not including the next new-line character will be ignored.

Strip lead and trail spaces. Select options for discarding leading and trailing spaces in strings on import.

Note. Comparisons between strings that do and do not use SQL pushback may generate different results where trailing spaces exist.

Invalid characters. Select Discard to remove invalid characters from the data source. Select Replace with to replace invalid characters with the specified symbol (one character only). Invalid characters are null characters or any character that does not exist in the encoding method specified.

Encoding. Specifies the text-encoding method used. You can choose between the system default, stream default, or UTF-8.

- The system default is specified in the Windows Control Panel or, if running in distributed mode, on the server computer.
- The stream default is specified in the Stream Properties dialog box.

Decimal symbol. Select the type of decimal separator used in your data source. The Stream default is the character selected from the Options tab of the stream properties dialog box. Otherwise, select either Period (.) or Comma (,) to read all data in this dialog box using the chosen character as the decimal separator.

Delimiters. Using the check boxes listed for this control, you can specify which characters, such as the comma (,), define field boundaries in the file. You can also specify more than one delimiter, such as “,|” for records that use multiple delimiters. The default delimiter is the comma.

Note: If the comma is also defined as the decimal separator, the default settings here will not work. In cases where the comma is both the field delimiter and the decimal separator, select Other in the Delimiters list. Then manually specify a comma in the entry field.

Select Allow multiple blank delimiters to treat multiple adjacent blank delimiter characters as a single delimiter. For example, if one data value is followed by four spaces and then another data value, this group would be treated as two fields rather than five.

Lines to scan for type. Specify how many lines to scan for specified data types.

Automatically recognize dates and times. To enable IBM® SPSS® Modeler to automatically attempt to recognize data entries as dates or times, select this check box. For example, this means that an entry such as 07-11-1965 will be identified as a date and 02:35:58 will be identified as a time; however, ambiguous entries such as 07111965 or 023558 will show up as integers since there are no delimiters between the numbers.

Note: To avoid potential data problems when using data files from previous versions of SPSS Modeler, this box is turned off by default for information saved in versions prior to 13.

Quotes. Using the drop-down lists, you can specify how single and double quotation marks are treated on import. You can choose to Discard all quotation marks, Include as text by including them in the field value, or Pair and discard to match pairs of quotation marks and remove them. If a quotation mark is unmatched, you will receive an error message. Both Discard and Pair and discard store the field value (without quotation marks) as a string.

At any point while you are working in this dialog box, click Refresh to reload fields from the data source. This is useful when you are altering data connections to the source node or when you are working between tabs in the dialog box.

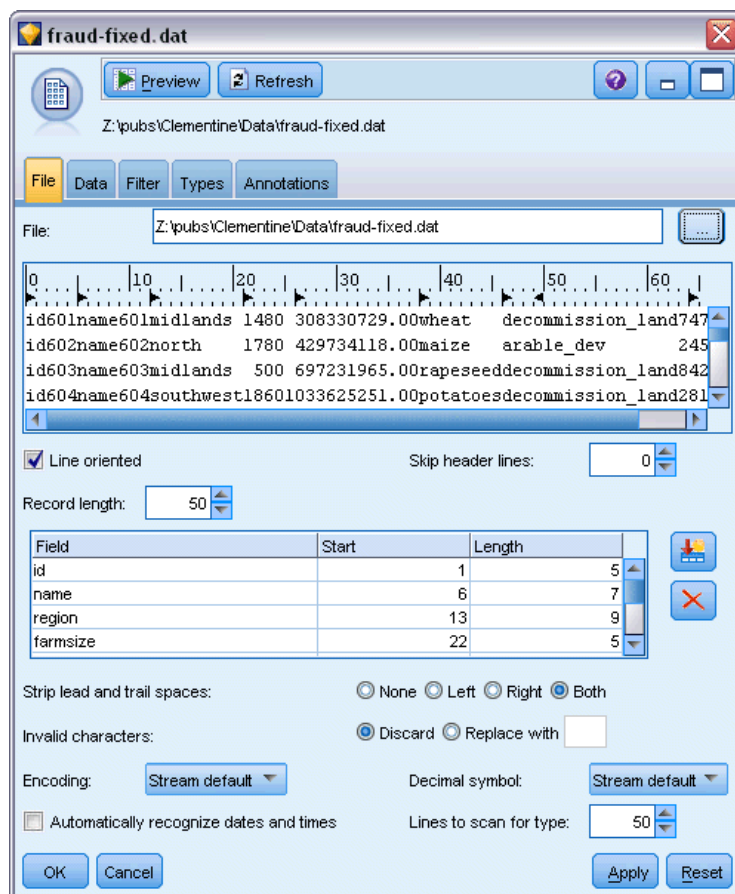
Fixed File Node

You can use Fixed File nodes to import data from fixed-field text files (files whose fields are not delimited but start at the same position and are of a fixed length). Machine-generated or legacy data are frequently stored in fixed-field format. Using the File tab of the Fixed File node, you can easily specify the position and length of columns in your data.

Setting Options for the Fixed File Node

The File tab of the Fixed File node enables you to bring data into IBM® SPSS® Modeler and to specify the position of columns and length of records. Using the data preview pane in the center of the dialog box, you can click to add arrows specifying the break points between fields.

Figure 2-12
Specifying columns in fixed-field data



File. Specify the name of the file. You can enter a filename or click the ellipsis button (...) to select a file. Once you have selected a file, the file path is shown and its contents are displayed with delimiters in the panel below.

The data preview pane can be used to specify column position and length. The ruler at the top of the preview window helps to measure the length of variables and to specify the break point between them. You can specify break point lines by clicking in the ruler area above the fields. Break points can be moved by dragging and can be discarded by dragging them outside of the data preview region.

- Each break-point line automatically adds a new field to the fields table below.
- Start positions indicated by the arrows are automatically added to the Start column in the table below.

Line oriented. Select if you want to skip the new-line character at the end of each record.

Skip header lines. Specify how many lines you want to ignore at the beginning of the first record. This is useful for ignoring column headers.

Record length. Specify the number of characters in each record.

Field. All fields that you have defined for this data file are listed here. There are two ways to define fields:

- Specify fields interactively using the data preview pane above.
- Specify fields manually by adding empty field rows to the table below. Click the button to the right of the fields pane to add new fields. Then, in the empty field, enter a field name, a start position, and a length. These options will automatically add arrows to the data preview pane, which can be easily adjusted.

To remove a previously defined field, select the field in the list and click the red delete button.

Start. Specify the position of the first character in the field. For example, if the second field of a record begins on the sixteenth character, you would enter 16 as the starting point.

Length. Specify how many characters are in the longest value for each field. This determines the cutoff point for the next field.

Strip lead and trail spaces. Select to discard leading and trailing spaces in strings on import.

Note. Comparisons between strings that do and do not use SQL pushback may generate different results where trailing spaces exist.

Invalid characters. Select Discard to remove invalid characters from the data input. Select Replace with to replace invalid characters with the specified symbol (one character only). Invalid characters are null (0) characters or any character that does not exist in the current encoding.

Encoding. Specifies the text-encoding method used. You can choose between the system default, stream default, or UTF-8.

- The system default is specified in the Windows Control Panel or, if running in distributed mode, on the server computer.
- The stream default is specified in the Stream Properties dialog box.

Decimal symbol. Select the type of decimal separator used in your data source. Stream default is the character selected from the Options tab of the stream properties dialog box. Otherwise, select either Period (.) or Comma (,) to read all data in this dialog box using the chosen character as the decimal separator.

Automatically recognize dates and times. To enable SPSS Modeler to automatically attempt to recognize data entries as dates or times, select this check box. For example, this means that an entry such as 07-11-1965 will be identified as a date and 02:35:58 will be identified as a time; however, ambiguous entries such as 07111965 or 023558 will show up as integers since there are no delimiters between the numbers.

Note: To avoid potential data problems when using data files from previous versions of SPSS Modeler, this box is turned off by default for information saved in versions prior to 13.

Lines to scan for type. Specify how many lines to scan for specified data types.

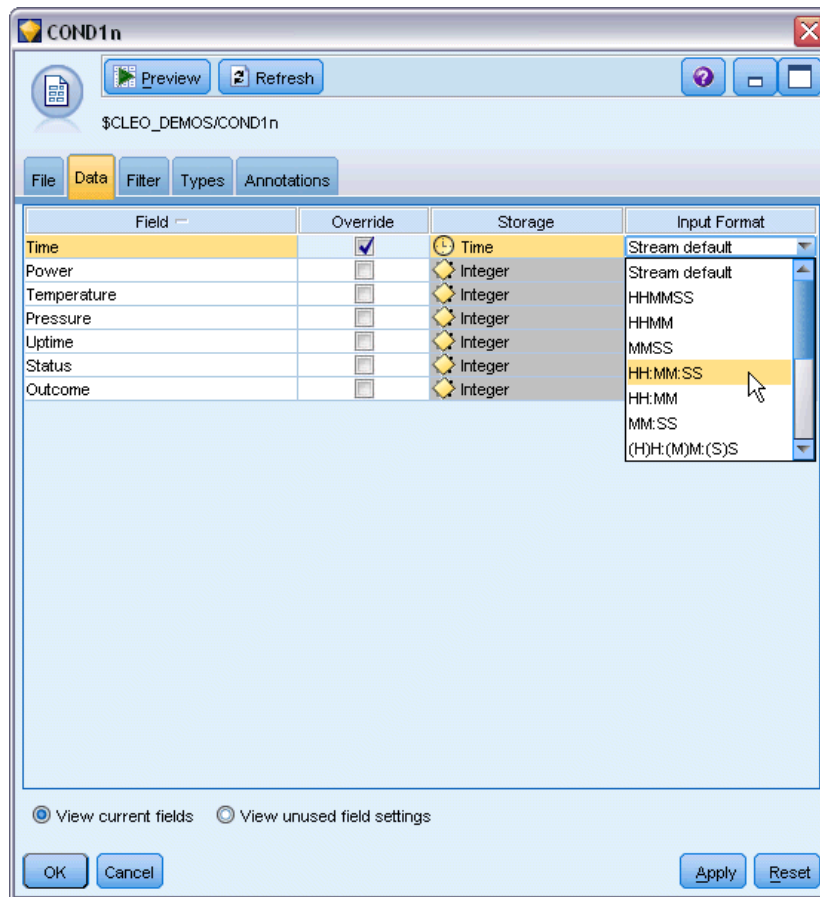
At any point while working in this dialog box, click **Refresh** to reload fields from the data source. This is useful when altering data connections to the source node or when working between tabs on the dialog box.

Setting Field Storage and Formatting

Options on the Data tab for Fixed File, Variable File, XML Source and User Input nodes allow you to specify the storage type for fields as they are imported or created in IBM® SPSS® Modeler. For Fixed File, Variable File and User Input nodes you can also specify the field formatting, and other metadata.

For data read from other sources, storage is determined automatically but can be changed using a conversion function, such as `to_integer`, in a Filler node or Derive node.

Figure 2-13
Overriding storage type and field formatting upon import



Field. Use the *Field* column to view and select fields in the current dataset.

Override. Select the check box in the *Override* column to activate options in the *Storage* and *Input Format* columns.

Data Storage

Storage describes the way data are stored in a field. For example, a field with values of 1 and 0 stores integer data. This is distinct from the measurement level, which describes the usage of the data, and does not affect storage. For example, you may want to set the measurement level for an integer field with values of 1 and 0 to *Flag*. This usually indicates that 1 = *True* and 0 = *False*. While storage must be determined at the source, measurement level can be changed using a Type node at any point in the stream. For more information, see the topic [Measurement Levels](#) in Chapter 4 on p. 115.

Available storage types are:

- **String.** Used for fields that contain non-numeric data, also called alphanumeric data. A string can include any sequence of characters, such as *fred*, *Class 2*, or *1234*. Note that numbers in strings cannot be used in calculations.

- **Integer.** A field whose values are integers.
- **Real.** Values are numbers that may include decimals (not limited to integers). The display format is specified in the Stream Properties dialog box and can be overridden for individual fields in a Type node (Format tab).
- **Date.** Date values specified in a standard format such as year, month, and day (for example, 2007-09-26). The specific format is specified in the Stream Properties dialog box.
- **Time.** Time measured as a duration. For example, a service call lasting 1 hour, 26 minutes, and 38 seconds might be represented as 01:26:38, depending on the current time format as specified in the Stream Properties dialog box.
- **Timestamp.** Values that include both a date and time component, for example 2007-09-26 09:04:00, again depending on the current date and time formats in the Stream Properties dialog box. Note that timestamp values may need to be wrapped in double-quotes to ensure they are interpreted as a single value rather than separate date and time values. (This applies for example when entering values in a User Input node.)

Storage conversions. You can convert storage for a field using a variety of conversion functions, such as `to_string` and `to_integer`, in a Filler node. For more information, see the topic [Storage Conversion Using the Filler Node](#) in Chapter 4 on p. 153. Note that conversion functions (and any other functions that require a specific type of input such as a date or time value) depend on the current formats specified in the Stream Properties dialog box. For example, if you want to convert a string field with values *Jan 2003*, *Feb 2003*, (and so forth) to date storage, select `MON YYYY` as the default date format for the stream. Conversion functions are also available from the Derive node, for temporary conversion during a derive calculation. You can also use the Derive node to perform other manipulations, such as recoding string fields with categorical values. For more information, see the topic [Recoding Values with the Derive Node](#) in Chapter 4 on p. 150.

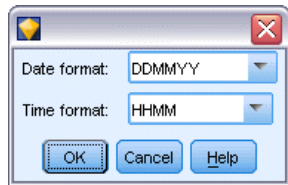
Reading in mixed data. Note that when reading in fields with numeric storage (either integer, real, time, timestamp, or date), any non-numeric values are set to null or system missing. This is because unlike some applications, SPSS Modeler does not allow mixed storage types within a field. To avoid this, any fields with mixed data should be read in as strings, either by changing the storage type in the source node or in the external application as necessary.

Field Input Format (Fixed File, Variable File and User Input nodes only)

For all storage types except String and Integer, you can specify formatting options for the selected field using the drop-down list. For example, when merging data from various locales, you may need to specify a period (.) as the decimal separator for one field, while another will require a comma separator.

Input options specified in the source node override the formatting options specified in the stream properties dialog box; however, they do not persist later in the stream. They are intended to parse input correctly based on your knowledge of the data. The specified formats are used as a guide for parsing the data as they are read into SPSS Modeler, not to determine how they should be formatted after being read into SPSS Modeler. To specify formatting on a per-field basis elsewhere in the stream, use the Format tab of a Type node. For more information, see the topic [Field Format Settings Tab](#) in Chapter 4 on p. 128.

Figure 2-14
Specifying date and time formats for timestamp fields



Options vary depending on the storage type. For example, for the Real storage type, you can select Period (.) or Comma (,) as the decimal separator. For timestamp fields, a separate dialog box opens when you select Specify from the drop-down list. For more information, see the topic [Setting Field Format Options](#) in Chapter 4 on p. 129.

For all storage types, you can also select Stream default to use the stream default settings for import. Stream settings are specified in the stream properties dialog box.

Additional Options

Several other options can be specified using the Data tab:

- To view storage settings for data that are no longer connected through the current node (train data, for example), select View unused field settings. You can clear the legacy fields by clicking Clear.
- At any point while working in this dialog box, click Refresh to reload fields from the data source. This is useful when you are altering data connections to the source node or when you are working between tabs on the dialog box.

Data Collection Node

Data Collection source nodes import survey data based on the IBM® SPSS® Data Collection Survey Reporter Developer Kit used by market research software from IBM Corp. This format distinguishes **case data**—the actual responses to questions gathered during a survey—from the **metadata** that describes how the case data is collected and organized. Metadata consists of information such as question texts, variable names and descriptions, multiple response variable definitions, translations of the text strings, and the definition of the structure of the case data.

Note: This node requires Data Collection Survey Reporter Developer Kit, which is distributed along with Data Collection software products from IBM Corp. For more information, see the Data Collection Web page at <http://www.ibm.com/software/analytics/spss/products/data-collection/survey-reporter-dev-kit/>. Aside from installing the Developer Kit, no additional configuration is required.

Comments

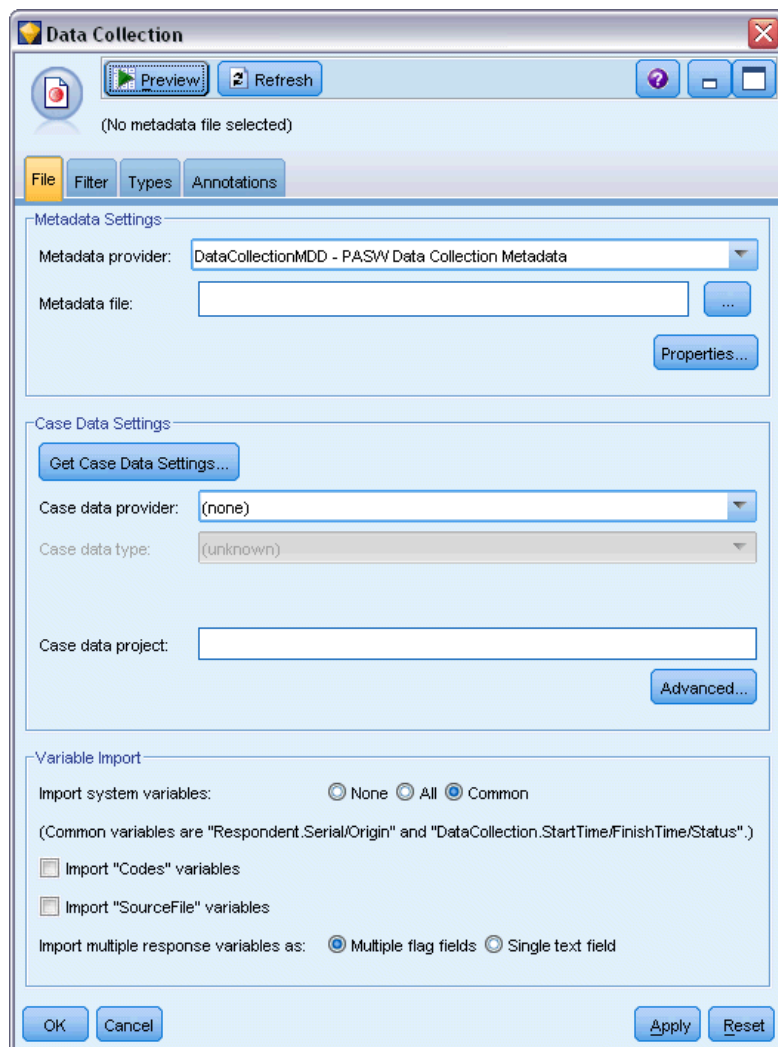
- Survey data is read from the flat, tabular VDATA format, or from sources in the hierarchical HDATA format if they include a metadata source (requires Data Collection 4.5 or higher).

- Types are instantiated automatically by using information from the metadata.
- When survey data is imported into IBM® SPSS® Modeler, questions are rendered as fields, with a record for each respondent.

Data Collection Import File Options

The File tab in the Data Collection node enables you to specify options for the metadata and case data you want to import.

Figure 2-15
Data Collection source node file options



Metadata Settings

Note: To see the full list of available provider file types, you need to install the IBM® SPSS® Data Collection Survey Reporter Developer Kit, available with the Data Collection software. For more information, see the Data Collection Web page at <http://www.ibm.com/software/analytics/spss/products/data-collection/survey-reporter-dev-kit/>

Metadata Provider. Survey data can be imported from a number of formats as supported by Data Collection Survey Reporter Developer Kit software. Available provider types include the following:

- **DataCollectionMDD.** Reads metadata from a questionnaire definition file (*.mdd*). This is the standard Data Collection Data Model format.
- **ADO Database.** Reads case data and metadata from ADO files. Specify the name and location of the *.adoinfo* file that contains the metadata. The internal name of this DSC is *mrADODsc*.
- **In2data Database.** Reads In2data case data and metadata. The internal name of this DSC is *mrI2dDsc*.
- **Data Collection Log File.** Reads metadata from a standard Data Collection log file. Typically, log files have a *.tmp* filename extension. However, some log files may have another filename extension. If necessary, you can rename the file so that it has a *.tmp* filename extension. The internal name of this DSC is *mrLogDsc*.
- **Quancept Definitions File.** Converts metadata to Quancept script. Specify the name of the Quancept *.qdi* file. The internal name of this DSC is *mrQdiDrsDsc*.
- **Quanvert Database.** Reads Quanvert case data and metadata. Specify the name and location of the *.qvinfo* or *.pkd* file. The internal name of this DSC is *mrQvDsc*.
- **Data Collection Participation Database.** Reads a project's Sample and History Table tables and creates derived categorical variables corresponding to the columns in those tables. The internal name of this DSC is *mrSampleReportingMDSC*.
- **Statistics File.** Reads case data and metadata from an IBM® SPSS® Statistics *.sav* file. Writes case data to an SPSS Statistics *.sav* file for analysis in SPSS Statistics. Writes metadata from an SPSS Statistics *.sav* file to an *.mdd* file. The internal name of this DSC is *mrSavDsc*.
- **Surveycraft File.** Reads SurveyCraft case data and metadata. Specify the name of the SurveyCraft *.vq* file. The internal name of this DSC is *mrSCDsc*.
- **Data Collection Scripting File.** Reads from metadata in an *mrScriptMetadata* file. Typically, these files have an *.mdd* or *.dms* filename extension. The internal name of this DSC is *mrScriptMDSC*.
- **Triple-S XML File.** Reads metadata from a Triple-S file in XML format. The internal name of this DSC is *mrTripleSDsc*.

Metadata properties. Optionally, select Properties to specify the survey version to import as well as the language, context, and label type to use. For more information, see the topic [IBM SPSS Data Collection Import Metadata Properties](#) on p. 30.

Case Data Settings

Note: To see the full list of available provider file types, you need to install the Data Collection Survey Reporter Developer Kit, available with the Data Collection software. For more information, see the Data Collection Web page at <http://www.ibm.com/software/analytics/spss/products/data-collection/survey-reporter-dev-kit/>

Get Case Data Settings. When reading metadata from *.mdd* files only, click Get Case Data Settings to determine what case data sources are associated with the selected metadata, along with the specific settings needed to access a given source. This option is available only for *.mdd* files.

Case Data Provider. The following provider types are supported:

- ADO Database. Reads case data using the Microsoft ADO interface. Select OLE-DB UDL for the case data type, and specify a connection string in the Case Data UDL field. For more information, see the topic [Database Connection String](#) on p. 31. The internal name of this component is *mrADODsc*.
- Delimited Text File (Excel). Reads case data from a comma-delimited (.CSV) file, such as can be output by Excel. The internal name is *mrCsvDsc*.
- Data Collection Data File. Reads case data from a native Data Collection Data Format file (Data Collection 4.5 onwards). The internal name is *mrDataFileDsc*.
- In2data Database. Reads case data and metadata from an In2data database (*.i2d*) file. The internal name is *mrI2dDsc*.
- Data Collection Log File. Reads case data from a standard Data Collection log file. Typically, log files have a *.tmp* filename extension. However, some log files may have another filename extension. If necessary, you can rename the file so that it has a *.tmp* filename extension. The internal name is *mrLogDsc*.
- Quantum Data File. Reads case data from any Quantum-format ASCII file (*.dat*). The internal name is *mrPunchDsc*.
- Quancept Data File. Reads case data from a Quancept *.drs*, *.drz*, or *.dru* file. The internal name is *mrQdiDrsDsc*.
- Quanvert Database. Reads case data from a Quanvert *qvinfo* or *.pkd* file. The internal name is *mrQvDsc*.
- Data Collection Database (MS SQL Server). Reads case data to a relational Microsoft SQL Server database. For more information, see the topic [Database Connection String](#) on p. 31. The internal name is *mrRdbDsc2*.
- Statistics File. Reads case data from an SPSS Statistics *.sav* file. The internal name is *mrSavDsc*.
- Surveycraft File. Reads case data from a SurveyCraft *.qdt* file. Both the *.vq* and *.qdt* files must be in the same directory, with read and write access for both files. This is not how they are created by default when using SurveyCraft, so one of the files needs to be moved to import SurveyCraft data. The internal name is *mrScDsc*.
- Triple-S Data File. Reads case data from a Triple-S data file, in either fixed-length or comma-delimited format. The internal name is *mrTripleDsc*.
- Data Collection XML. Reads case data from a Data Collection XML data file. Typically, this format may be used to transfer case data from one location to another. The internal name is *mrXmlDsc*.

Case Data Type. Specifies whether case data is read from a file, folder, OLE-DB UDL, or ODBC DSN, and updates the dialog box options accordingly. Valid options depend on the type of provider. For database providers, you can specify options for the OLE-DB or ODBC connection. For more information, see the topic [Database Connection String](#) on p. 31.

Case Data Project. When reading case data from a Data Collection database, you can enter the name of the project. For all other case data types, this setting should be left blank.

Variable Import

Import System Variables. Specifies whether system variables are imported, including variables that indicate interview status (in progress, completed, finish date, and so on). You can choose None, All, or Common.

Import “Codes” Variables. Controls import of variables that represent codes used for open-ended “Other” responses for categorical variables.

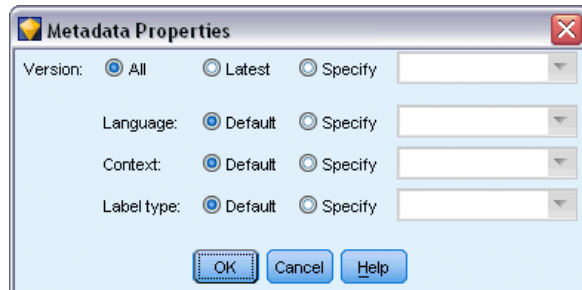
Import “SourceFile” Variables. Controls import of variables that contain filenames of images of scanned responses.

Import multi-response variables as. Multiple response variables can be imported as multiple flag fields (a multiple dichotomy set), which is the default method for new streams. Streams created in releases of IBM® SPSS® Modeler prior to 12.0 imported multiple responses into a single field, with values separate by commas. The older method is still supported to allow existing streams to run as they did previously, but updating older streams to use the new method is recommended. For more information, see the topic [Importing Multiple Response Sets](#) on p. 32.

IBM SPSS Data Collection Import Metadata Properties

When importing IBM® SPSS® Data Collection survey data, you can specify the survey version to import as well as the language, context, and label type to use. Note that only one language, context, and label type can be imported at a time.

Figure 2-16
IBM SPSS Data Collection Import Metadata Properties



Version. Each survey version can be regarded as a snapshot of the metadata used to collect a particular set of case data. As a questionnaire undergoes changes, multiple versions may be created. You can import the latest version, all versions, or a specific version.

- **All versions.** Select this option if you want to use a combination (superset) of all of the available versions. (This is sometimes called a superversion). When there is a conflict between the versions, the most recent versions generally take precedence over the older versions. For example, if a category label differs in any of the versions, the text in the latest version will be used.
- **Latest version.** Select this option if you want to use the most recent version.
- **Specify version.** Select this option if you want to use a particular survey version.

Choosing all versions is useful when, for example, you want to export case data for more than one version and there have been changes to the variable and category definitions that mean that case data collected with one version is not valid in another version. Selecting all of the versions for which you want to export the case data means that generally you can export the case data collected with the different versions at the same time without encountering validity errors due to the differences between the versions. However, depending on the version changes, some validity errors may still be encountered.

Language. Questions and associated text can be stored in multiple languages in the metadata. You can use the default language for the survey or specify a particular language. If an item is unavailable in the specified language, the default is used.

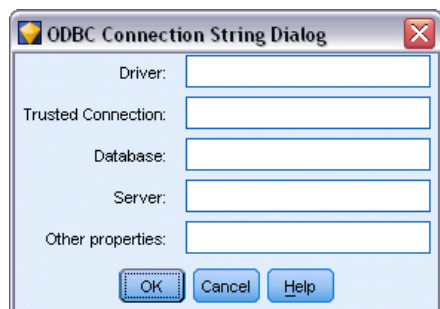
Context. Select the user context you want to use. The user context controls which texts are displayed. For example, select Question to display question texts or Analysis to display shorter texts suitable for displaying when analyzing the data.

Label type. Lists the types of labels that have been defined. The default is label, which is used for question texts in the Question user context and variable descriptions in the Analysis user context. Other label types can be defined for instructions, descriptions, and so forth.

Database Connection String

When using the IBM® SPSS® Data Collection node to import case data from a database via an OLE-DB or ODBC, select Edit from the File tab to access the Connection String dialog box, which enables you to customize the connection string passed to the provider in order to fine-tune the connection.

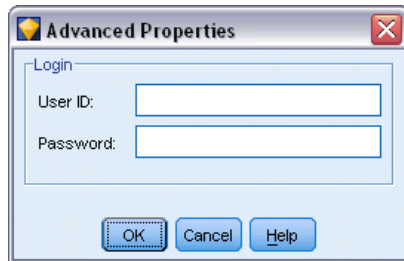
Figure 2-17
IBM SPSS Data Collection Import Connection String



Advanced Properties

When using the IBM® SPSS® Data Collection node to import case data from a database that requires an explicit login, select Advanced to provide a user ID and password to access the data source.

Figure 2-18
IBM SPSS Data Collection Import Advanced Properties



Importing Multiple Response Sets

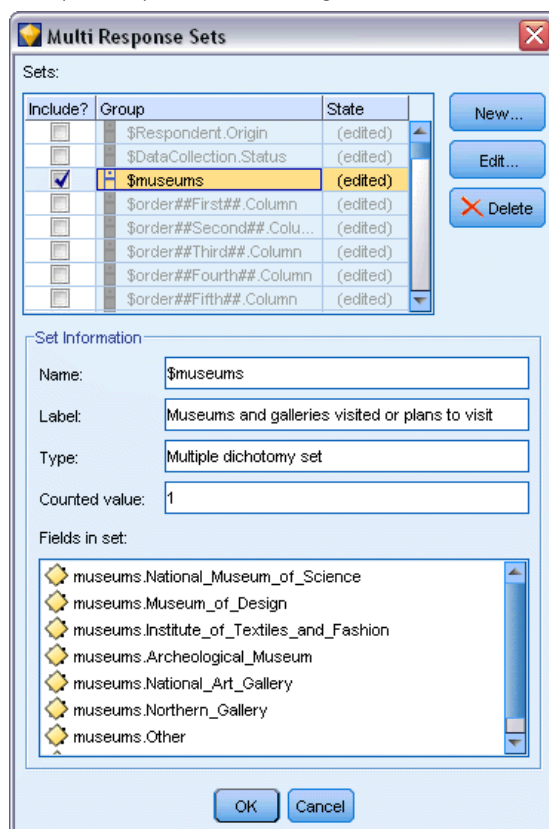
Multiple response variables can be imported from IBM® SPSS® Data Collection as multiple dichotomy sets, with a separate flag field for each possible value of the variable. For example, if respondents are asked to select which museums they have visited from a list, the set would include a separate flag field for each museum listed.

Figure 2-19
Multiple response question

Q14	Which museums or art galleries have you visited or do you intend to visit?
	SELECT ALL ANSWERS THAT APPLY.
National Museum of Science	<input type="checkbox"/>
Museum of Design	<input type="checkbox"/>
Institute of Textiles and Fashion	<input type="checkbox"/>
Archeological Museum	<input type="checkbox"/>
National Art Gallery	<input type="checkbox"/>
Northern Gallery	<input type="checkbox"/>
Other (Please write in)	<input type="checkbox"/>
Not answered	<input type="checkbox"/>

After importing the data, you can add or edit multiple response sets from any node that includes a Filter tab. For more information, see the topic [Editing Multiple Response Sets](#) in Chapter 4 on p. 134.

Figure 2-20
Multiple Response Sets dialog box



Importing Multiple Responses into a Single Field (for Streams Created in Previous Releases)

In older releases of IBM® SPSS® Modeler, rather than import multiple responses as described above, they were imported into a single field, with values separate by commas. This method is still supported in order to support for existing streams, but it is recommended that any such streams be updated to use the new method.

IBM SPSS Data Collection Column Import Notes

Columns from the IBM® SPSS® Data Collection data are read into IBM® SPSS® Modeler as summarized in the following table.

Data Collection Column Type	SPSS Modeler Storage	Measurement Level
Boolean flag (yes/no)	String	Flag (values 0 and 1)
Categorical	String	Nominal
Date or time stamp	Timestamp	Continuous
Double (floating point value within a specified range)	Real	Continuous
Long (integer value within a specified range)	Integer	Continuous

Data Collection Column Type	SPSS Modeler Storage	Measurement Level
Text (free text description)	String	Typeless
Level (indicates grids or loops within a question)	Doesn't occur in VDATA and is not imported into SPSS Modeler	
Object (binary data such as a facsimile showing scribbled text or a voice recording)	Not imported into SPSS Modeler	
None (unknown type)	Not imported into SPSS Modeler	
Respondent.Serial column (associates a unique ID with each respondent)	Integer	Typeless

To avoid possible inconsistencies between value labels read from metadata and actual values, all metadata values are converted to lower case. For example, the value label *E1720_years* would be converted to *e1720_years*.

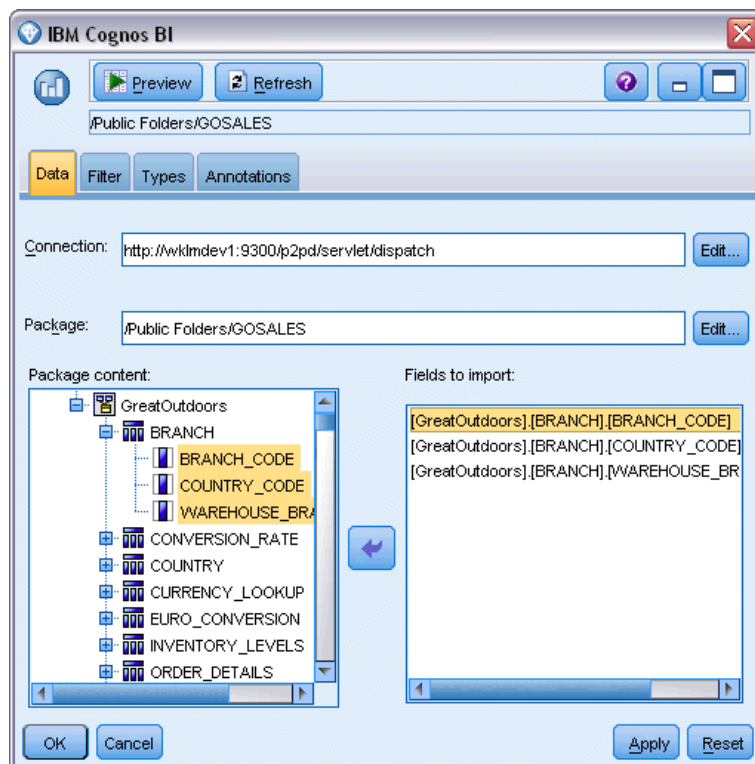
IBM Cognos BI Source Node

The IBM Cognos BI source node enables you to bring Cognos BI database data into your data mining session. In this way, you can combine the business intelligence features of Cognos with the predictive analytics capabilities of IBM® SPSS® Modeler. You can import relational, dimensionally-modeled relational (DMR) and OLAP data.

From a Cognos server connection, you first select a package from which to import data. A package contains a Cognos model and all of the folders, queries, reports, views, shortcuts, URLs, and job definitions associated with that model. A Cognos model defines business rules, data descriptions, data relationships, business dimensions and hierarchies, and other administrative tasks.

You then select the query subjects (which represent database tables) or individual query items (which represent table columns) that you want to import from the selected package. *Note:* The data to be imported must be in UTF-8 format.

Figure 2-21
Importing Cognos data



Connection. Click the Edit button to display a dialog box where you can define the details of a new Cognos connection from which to import data. If you are already logged in to a Cognos server through SPSS Modeler, you can also edit the details of the current connection. For more information, see the topic [Cognos connections](#) on p. 36.

Package. When you have established the Cognos server connection, click the Edit button next to this field to display a list of available packages from which to import content. For more information, see the topic [Cognos package selection](#) on p. 36.

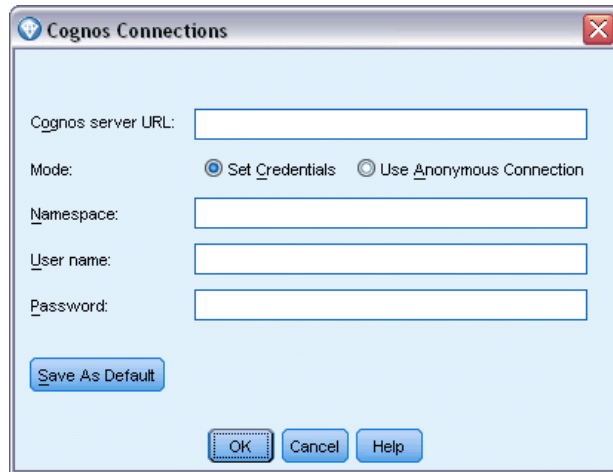
Package content. Displays the name of the selected package, together with the namespaces associated with the package. Double-click a namespace to display the objects that you can import. Select the objects you want to import and click the right arrow to move them into the Fields to import pane. Selecting a query subject imports all of its query items. Double-clicking a query subject expands it so that you can choose one or more of its individual query items. You can perform multiple selections with Ctrl-click (select individual items), Shift-click (select a block of items) and Ctrl-A (select all items).

Fields to import. Lists the database objects that will be imported into SPSS Modeler for processing. If you no longer require a particular object, select it and click the left arrow to return it to the Package content pane. You can perform multiple selections in the same way as for Package content.

Cognos connections

The Cognos Connections dialog box enables you to select the Cognos server from which to import or export data.

Figure 2-22
Cognos server selection



Cognos server URL. Type the URL of the Cognos server from which to import or export data. Contact your Cognos system administrator if you are not sure which URL to use.

Mode. Select Set Credentials if you want to log in with a specific Cognos namespace, username and password (for example, as an administrator). Select Use Anonymous connection to log in with no user credentials, in which case you do not fill in the other fields.

Namespace. Specify the Cognos security authentication provider used to log on to the server. The authentication provider is used to define and maintain users, groups, and roles, and to control the authentication process.

User name. Enter the Cognos user name with which to log on to the server.

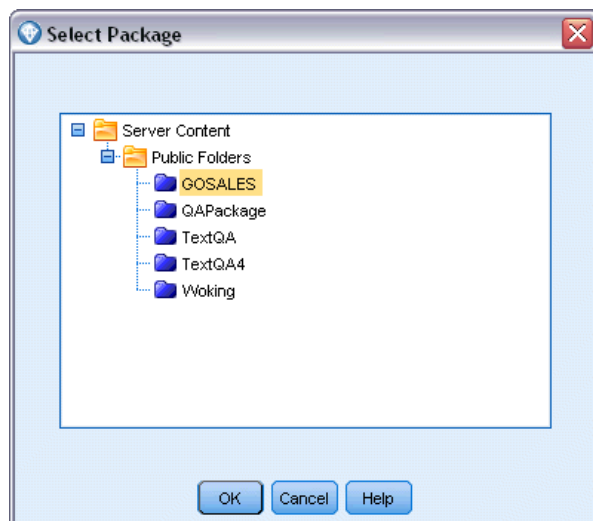
Password. Enter the password associated with the specified user name.

Save as Default. Click this button to store these settings as your default, to avoid having to re-enter them every time you open the node.

Cognos package selection

The Select Package dialog box enables you to select a Cognos package from which to import data.

Figure 2-23
Cognos package selection



Cognos objects. A list of the packages available from the chosen server. Select the package you want to use and click OK. You can select only one package per Cognos BI source node.

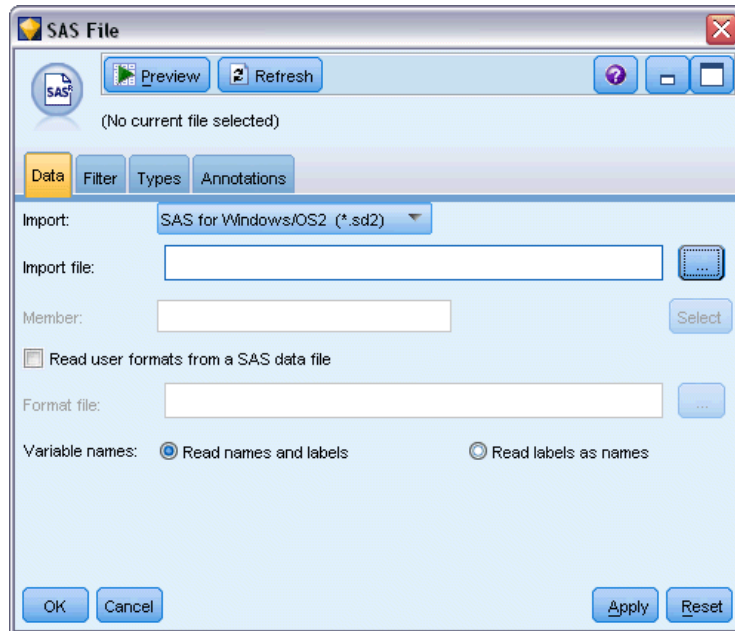
SAS Source Node

The SAS source node enables you to bring SAS data into your data mining session. You can import four types of files:

- SAS for Windows/OS2 (.sd2)
- SAS for UNIX (.ssd)
- SAS Transport File (.tpt)
- SAS version 7/8/9 (.sas7bdat)

When the data are imported, all variables are kept and no variable types are changed. All cases are selected.

Figure 2-24
Importing a SAS file



Setting Options for the SAS Source Node

Import. Select which type of SAS file to transport. You can choose SAS for Windows/OS2 (.sd2), SAS for UNIX (.SSD), SAS Transport File (.tpt), or SAS Version 7/8/9 (.sas7bdat).

Import file. Specify the name of the file. You can enter a filename or click the ellipsis button (...) to browse to the file's location.

Member. Select a member to import from the SAS transport file selected above. You can enter a member name or click Select to browse through all members in the file.

Read user formats from a SAS data file. Select to read user formats. SAS files store data and data formats (such as variable labels) in different files. Most often, you will want to import the formats as well. If you have a large dataset; however, you may want to deselect this option to save memory.

Format file. If a format file is required, this text box is activated. You can enter a filename or click the ellipsis button (...) to browse to the file's location.

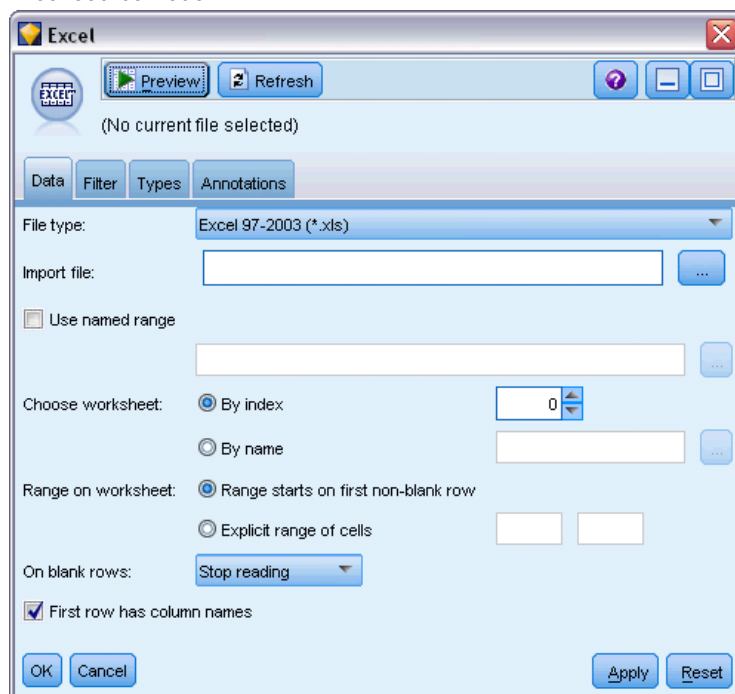
Variable names. Select a method of handling variable names and labels upon import from a SAS file. Metadata that you choose to include here persists throughout your work in IBM® SPSS® Modeler and may be exported again for use in SAS.

- **Read names and labels.** Select to read both variable names and labels into SPSS Modeler. By default, this option is selected and variable names are displayed in the Type node. Labels may be displayed in the Expression Builder, charts, model browsers, and other types of output, depending on the options specified in the stream properties dialog box.
- **Read labels as names.** Select to read the descriptive variable labels from the SAS file rather than the short field names and use these labels as variable names in SPSS Modeler.

Excel Source Node

The Excel source node enables you to import data from any version of Microsoft Excel.

Figure 2-25
Excel source node



File type. Select the Excel file type that you are importing.

Import file. Specifies the name and location of the spreadsheet file to import.

Use Named Range. Enables you to specify a named range of cells as defined in the Excel worksheet. Click the ellipses button (...) to choose from the list of available ranges. If a named range is used, other worksheet and data range settings are no longer applicable and are disabled as a result.

Choose worksheet. Specifies the worksheet to import, either by index or by name.

- **By index.** Specify the index value for the worksheet you want to import, beginning with 0 for the first worksheet, 1 for the second worksheet, and so on.
- **By name.** Specify the name of the worksheet you want to import. Click the ellipses button (...) to choose from the list of available worksheets.

Range on worksheet. You can import data beginning with the first non-blank row or with an explicit range of cells.

- **Range starts on first non-blank row.** Locates the first non-blank cell and uses this as the upper left corner of the data range.
- **Explicit range of cells.** Enables you to specify an explicit range by row and column. For example, to specify the Excel range A1:D5, you can enter A1 in the first field and D5 in the second (or alternatively, R1C1 and R5C4). All rows in the specified range are returned, including blank rows.

On blank rows. If more than one blank row is encountered, you can choose whether to Stop reading, or choose Return blank rows to continue reading all data to the end of the worksheet, including blank rows.

First row has column names. Indicates that the first row in the specified range should be used as field (column) names. If not selected, field names are generated automatically.

Field Storage and Measurement Level

When reading values from Excel, fields with numeric storage are read in with a measurement level of *Continuous* by default, and string fields are read in as *Nominal*. You can manually change the measurement level (continuous versus nominal) on the Type tab, but the storage is determined automatically (although it can be changed using a conversion function, such as `to_integer`, in a Filler node or Derive node if necessary). For more information, see the topic [Setting Field Storage and Formatting](#) on p. 23.

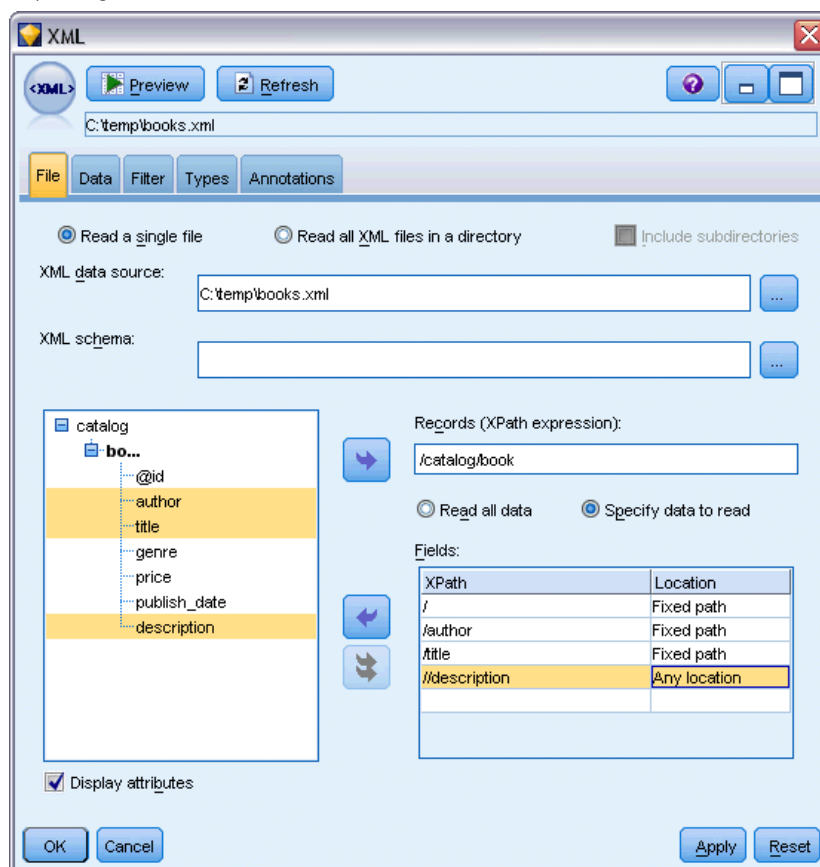
By default, fields with a mix of numeric and string values read in as numbers, which means that any string values will be set to null (system missing) values in IBM® SPSS® Modeler. This happens because—unlike Excel—SPSS Modeler does not allow mixed storage types within a field. To avoid this, you can manually set the cell format to Text in the Excel spreadsheet, which causes all values (including numbers) to read in as strings.

XML Source Node

The XML source node enables you to import data from a file in XML format into an IBM® SPSS® Modeler stream. XML is a standard language for data exchange, and for many organizations it is the format of choice for this purpose. For example, a government tax agency might want to analyze data from tax returns that have been submitted online and which have their data in XML format.

Importing XML data into an SPSS Modeler stream enables you to perform a wide range of predictive analytics functions on the source. The XML data is parsed into a tabular format in which the columns correspond to the different levels of nesting of the XML elements and attributes. The XML items are displayed in XPath format (see <http://www.w3.org/TR/xpath20/>).

Figure 2-26
Importing XML data



Read a single file. By default, SPSS Modeler reads a single file, which you specify in the XML data source field.

Read all XML files in a directory. Choose this option if you want to read all the XML files in a particular directory. Specify the location in the Directory field that appears. Select the Include subdirectories check box to additionally read XML files from all the subdirectories of the specified directory.

XML data source. Type the full path and file name of the XML source file you want to import, or use the Browse button to find the file.

XML schema. (Optional) Specify the full path and file name of an XSD or DTD file from which to read the XML structure, or use the Browse button to find this file. If you leave this field blank, the structure is read from the XML source file. An XSD or DTD file can have more than one root element. In this case, when you change the focus to a different field, a dialog is displayed where you choose the root element you want to use. For more information, see the topic [Selecting from Multiple Root Elements](#) on p. 42.

XML structure. A hierarchical tree showing the structure of the XML source file (or the schema, if you specified one in the XML schema field). To define a record boundary, select an element and click the right-arrow button to copy the item to the Records field.

Display attributes. Displays or hides the attributes of the XML elements in the XML structure field.

Records (XPath expression). Shows the XPath syntax for an element copied from the XML structure field. This element is then highlighted in the XML structure, and defines the record boundary. Each time this element is encountered in the source file, a new record is created. If this field is empty, the first child element under the root is used as the record boundary.

Read all data. By default, all data in the source file is read into the stream.

Specify data to read. Choose this option if you want to import individual elements, attributes or both. Choosing this option enables the Fields table where you can specify the data you want to import.

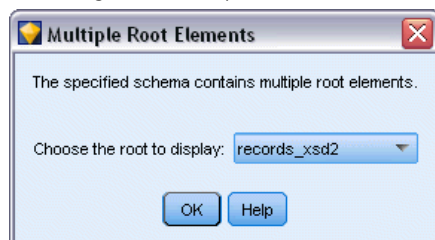
Fields. This table lists the elements and attributes selected for import, if you have selected the Specify data to read option. You can either type the XPath syntax of an element or attribute directly into the XPath column, or select an element or attribute in the XML structure and click the right-arrow button to copy the item into the table. To copy all the child elements and attributes of an element, select the element in the XML structure and click the double-arrow button.

- **XPath.** The XPath syntax of the items to be imported.
- **Location.** The location in the XML structure of the items to be imported. Fixed path shows the path of the item relative to the element highlighted in the XML structure (or the first child element under the root, if no element is highlighted). Any location denotes an item of the given name at any location in the XML structure. Custom is displayed if you type a location directly into the XPath column.

Selecting from Multiple Root Elements

While a properly formed XML file can only have a single root element, an XSD or DTD file can contain multiple roots. If one of the roots matches that in the XML source file, that root element is used, otherwise you need to select one to use.

Figure 2-27
Selecting from multiple root elements



Choose the root to display. Select the root element you want to use. The default is the first root element in the XSD or DTD structure.

Removing Unwanted Spaces from XML Source Data

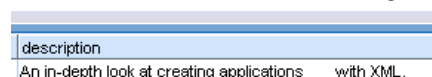
Line breaks in the XML source data may be implemented by a [CR][LF] character combination. In some cases these line breaks can occur in the middle of a text string, for example:

```
<description>An in-depth look at creating applications[CR][LF]
with XML.</description>
```

These line breaks may not be visible when the file is opened in some applications, for example a Web browser. However, when the data are read into the stream through the XML source node, the line breaks are converted to a series of space characters, for example:

Figure 2-28

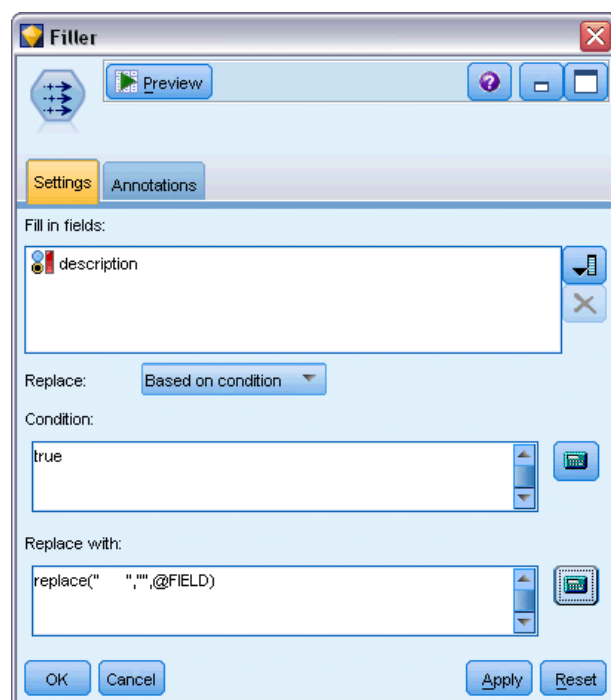
XML record with line break showing as spaces



You can correct this by using a Filler node to remove these unwanted spaces:

Figure 2-29

Filler node with settings to remove spaces



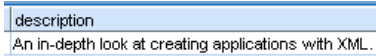
Here is an example of how you can achieve this:

- ▶ Attach a Filler node to the XML source node.
- ▶ Open the Filler node and use the field chooser to select the field with the unwanted spaces.
- ▶ Set Replace to Based on condition and set Condition to true.

- ▶ In the Replace with field, enter `replace(" ", "", @FIELD)` and click OK.
- ▶ Attach a Table node to the Filler node and run the stream.

In the Table node output, the text now appear as follows:

Figure 2-30
XML record with unwanted spaces removed



The screenshot shows a table with one row. The first column is labeled 'description' and contains the text 'An in-depth look at creating applications with XML.' The text is displayed without any leading or trailing spaces.

User Input Node

The User Input node provides an easy way for you to create synthetic data—either from scratch or by altering existing data. This is useful, for example, when you want to create a test dataset for modeling.

Creating Data from Scratch

The User Input node is available from the Sources palette and can be added directly to the stream canvas.

- ▶ Click the Sources tab of the nodes palette.
- ▶ Drag and drop or double-click to add the User Input node to the stream canvas.
- ▶ Double-click to open its dialog box and specify fields and values.

Note: User Input nodes that are selected from the Sources palette will be completely blank, with no fields and no data information. This enables you to create synthetic data entirely from scratch.

Generating Data from an Existing Data Source

Figure 2-31
User Input node generated from a stream node



You can also generate a User Input node from any nonterminal node in the stream:

- ▶ Decide at which point in the stream you want to replace a node.
- ▶ Right-click on the node that will feed its data into the User Input node and choose Generate User Input Node from the menu.
- ▶ The User Input node appears with all downstream processes attached to it, replacing the existing node at that point in your data stream. When generated, the node inherits all of the data structure and field type information (if available) from the metadata.

Note: If data have not been run through all nodes in the stream, then the nodes are not fully instantiated, meaning that storage and data values may not be available when replacing with a User Input node.

Setting Options for the User Input Node

The dialog box for a User Input node contains several tools you can use to enter values and define the data structure for synthetic data. For a generated node, the table on the Data tab contains field names from the original data source. For a node added from the Sources palette, the table is blank. Using the table options, you can perform the following tasks:

- Add new fields using the Add a New Field button at the right in the table.
- Rename existing fields.
- Specify data storage for each field.

- Specify values.
- Change the order of fields on the display.

Entering Data

For each field, you can specify values or insert values from the original dataset using the value picker button to the right of the table. See the rules described below for more information on specifying values. You can also choose to leave the field blank—fields left blank are filled with the system null (\$null\$).

Figure 2-32
Specifying storage type for fields in a generated User Input node



To specify string values, simply type them in the values column, separated by spaces:

Fred Ethel Martin

Strings that include spaces can be wrapped in double-quotes:

"Bill Smith" "Fred Martin" "Jack Jones"

For numeric fields, you can either enter multiple values in the same manner (listed with spaces between):

10 12 14 16 18 20

Or you can specify the same series of numbers by setting its limits (10, 20) and the steps in between (2). Using this method, you would type:

10,20,2

These two methods can be combined by embedding one within the other, such as:

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

This entry will produce the following values:

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

Date and time values can be entered using the current default format selected in the Stream Properties dialog box, for example:

```
11:04:00 11:05:00 11:06:00
```

```
2007-03-14 2007-03-15 2007-03-16
```

For timestamp values, which have both a date and time component, double-quotes must be used:

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

For additional details see comments on data storage below.

Generate data. Enables you to specify how the records are generated when you run the stream.

- **All combinations.** Generates records containing every possible combination of the field values, so each field value will appear in several records. This can sometimes generate more data than is wanted, so often you might follow this node with a sample node.
- **In order.** Generates records in the order in which the data field values are specified. Each field value only appears in one record. The total number of records is equal to the largest number of values for a single field. Where fields have fewer than the largest number, undefined (\$null\$) values are inserted.

For example, the following entries will generate the records listed in the tables below.

- **Age.** 30,60,10
- **BP.** LOW
- **Cholesterol.** NORMAL HIGH
- **Drug.** (left blank)

Generate data set to All combinations:

Age	BP	Cholesterol	Drug
30	LOW	NORMAL	\$null\$
30	LOW	HIGH	\$null\$
40	LOW	NORMAL	\$null\$
40	LOW	HIGH	\$null\$
50	LOW	NORMAL	\$null\$
50	LOW	HIGH	\$null\$
60	LOW	NORMAL	\$null\$
60	LOW	HIGH	\$null\$

Generate data set to In order:

Age	BP	Cholesterol	Drug
30	LOW	NORMAL	\$null\$
40	\$null\$	HIGH	\$null\$
50	\$null\$	\$null\$	\$null\$
60	\$null\$	\$null\$	\$null\$

Data Storage

Storage describes the way data are stored in a field. For example, a field with values of 1 and 0 stores integer data. This is distinct from the measurement level, which describes the usage of the data, and does not affect storage. For example, you may want to set the measurement level for an integer field with values of 1 and 0 to *Flag*. This usually indicates that 1 = *True* and 0 = *False*. While storage must be determined at the source, measurement level can be changed using a Type node at any point in the stream. For more information, see the topic [Measurement Levels](#) in Chapter 4 on p. 115.

Available storage types are:

- **String.** Used for fields that contain non-numeric data, also called alphanumeric data. A string can include any sequence of characters, such as *fred*, *Class 2*, or *1234*. Note that numbers in strings cannot be used in calculations.
- **Integer.** A field whose values are integers.
- **Real.** Values are numbers that may include decimals (not limited to integers). The display format is specified in the Stream Properties dialog box and can be overridden for individual fields in a Type node (Format tab).
- **Date.** Date values specified in a standard format such as year, month, and day (for example, 2007-09-26). The specific format is specified in the Stream Properties dialog box.
- **Time.** Time measured as a duration. For example, a service call lasting 1 hour, 26 minutes, and 38 seconds might be represented as 01:26:38, depending on the current time format as specified in the Stream Properties dialog box.
- **Timestamp.** Values that include both a date and time component, for example 2007-09-26 09:04:00, again depending on the current date and time formats in the Stream Properties dialog box. Note that timestamp values may need to be wrapped in double-quotes to ensure they are interpreted as a single value rather than separate date and time values. (This applies for example when entering values in a User Input node.)

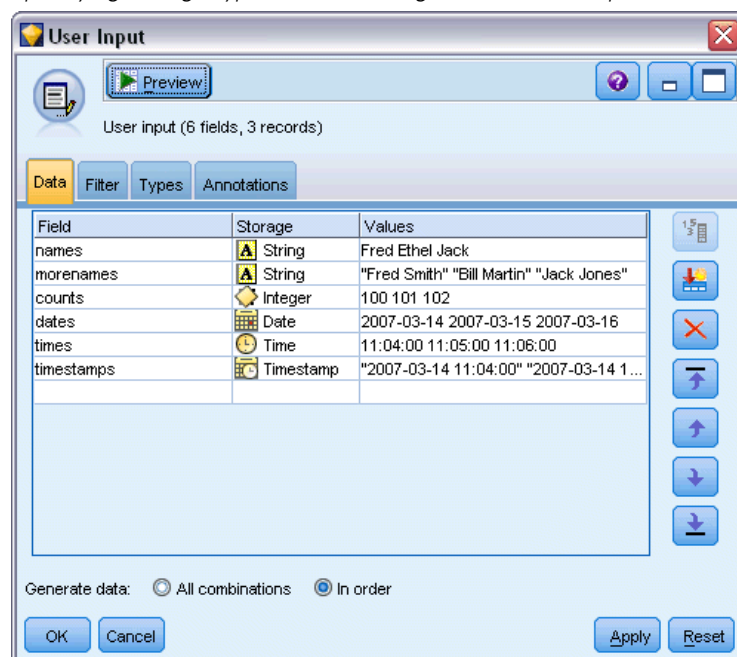
Storage conversions. You can convert storage for a field using a variety of conversion functions, such as `to_string` and `to_integer`, in a Filler node. For more information, see the topic [Storage Conversion Using the Filler Node](#) in Chapter 4 on p. 153. Note that conversion functions (and any other functions that require a specific type of input such as a date or time value) depend on the current formats specified in the Stream Properties dialog box. For example, if you want to convert a string field with values *Jan 2003*, *Feb 2003*, (and so forth) to date storage, select MON YYYY as the default date format for the stream. Conversion functions are also available from the Derive node, for temporary conversion during a derive calculation. You can also use the Derive node

to perform other manipulations, such as recoding string fields with categorical values. For more information, see the topic [Recoding Values with the Derive Node](#) in Chapter 4 on p. 150.

Reading in mixed data. Note that when reading in fields with numeric storage (either integer, real, time, timestamp, or date), any non-numeric values are set to null or system missing. This is because unlike some applications, IBM® SPSS® Modeler does not allow mixed storage types within a field. To avoid this, any fields with mixed data should be read in as strings, either by changing the storage type in the source node or in the external application as necessary.

Note: Generated User Input nodes may already contain storage information garnered from the source node if instantiated. An uninstantiated node does not contain storage or usage type information.

Figure 2-33
Specifying storage type for fields in a generated User Input node



Rules for Specifying Values

For symbolic fields, you should leave spaces between multiple values, such as:

HIGH MEDIUM LOW

For numeric fields, you can either enter multiple values in the same manner (listed with spaces between):

10 12 14 16 18 20

Or you can specify the same series of numbers by setting its limits (10, 20) and the steps in between (2). Using this method, you would type:

10,20,2

These two methods can be combined by embedding one within the other, such as:

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

This entry will produce the following values:

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

Common Source Node Tabs

The following options can be specified for all source nodes by clicking the corresponding tab:

- **Data tab.** Used to change the default storage type.
- **Filter tab.** Used to eliminate or rename data fields. This tab offers the same functionality as the Filter node. For more information, see the topic [Setting Filtering Options](#) in Chapter 4 on p. 131.
- **Types tab.** Used to set measurement levels. This tab offers the same functionality as the Type node.
- **Annotations tab.** Used for all nodes, this tab offers options to rename nodes, supply a custom ToolTip, and store a lengthy annotation.

Setting Measurement Levels in the Source Node

Field properties can be specified in a source node or in a separate Type node. The functionality is similar in both nodes. The following properties are available:

- **Field.** Double-click any field name to specify value and field labels for data in IBM® SPSS® Modeler. For example, field metadata imported from IBM® SPSS® Statistics can be viewed or modified here. Similarly, you can create new labels for fields and their values. The labels that you specify here are displayed throughout SPSS Modeler depending on the selections you make in the stream properties dialog box.
- **Measurement.** This is the measurement level, used to describe characteristics of the data in a given field. If all of the details of a field are known, it is called **fully instantiated**. For more information, see the topic [Measurement Levels](#) in Chapter 4 on p. 115.

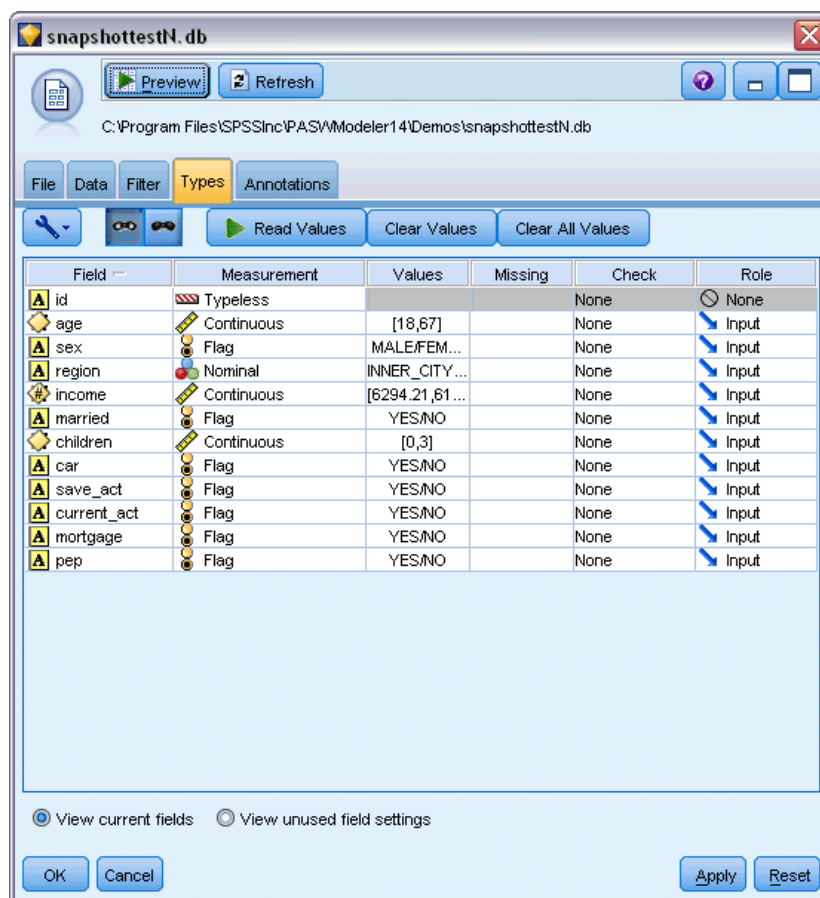
Note: The measurement level of a field is different from its storage type, which indicates whether the data are stored as strings, integers, real numbers, dates, times, or timestamps.

- **Values.** This column enables you to specify options for reading data values from the dataset, or use the Specify option to specify measurement levels and values in a separate dialog box. You can also choose to pass fields without reading their values. For more information, see the topic [Data Values](#) in Chapter 4 on p. 119.
- **Missing.** Used to specify how missing values for the field will be handled. For more information, see the topic [Defining Missing Values](#) in Chapter 4 on p. 124.

- **Check.** In this column, you can set options to ensure that field values conform to the specified values or ranges. For more information, see the topic [Checking Type Values](#) in Chapter 4 on p. 124.
- **Role.** Used to tell modeling nodes whether fields will be Input (predictor fields) or Target (predicted fields) for a machine-learning process. Both and None are also available roles, along with Partition, which indicates a field used to partition records into separate samples for training, testing, and validation. The value Split specifies that separate models will be built for each possible value of the field. For more information, see the topic [Setting the Field Role](#) in Chapter 4 on p. 126.

For more information, see the topic [Type Node](#) in Chapter 4 on p. 113.

Figure 2-34
Types tab options



When to Instantiate at the Source Node

There are two ways you can learn about the data storage and values of your fields. This **instantiation** can occur at either the source node, when you first bring data into IBM® SPSS® Modeler, or by inserting a Type node into the data stream.

Instantiating at the source node is useful when:

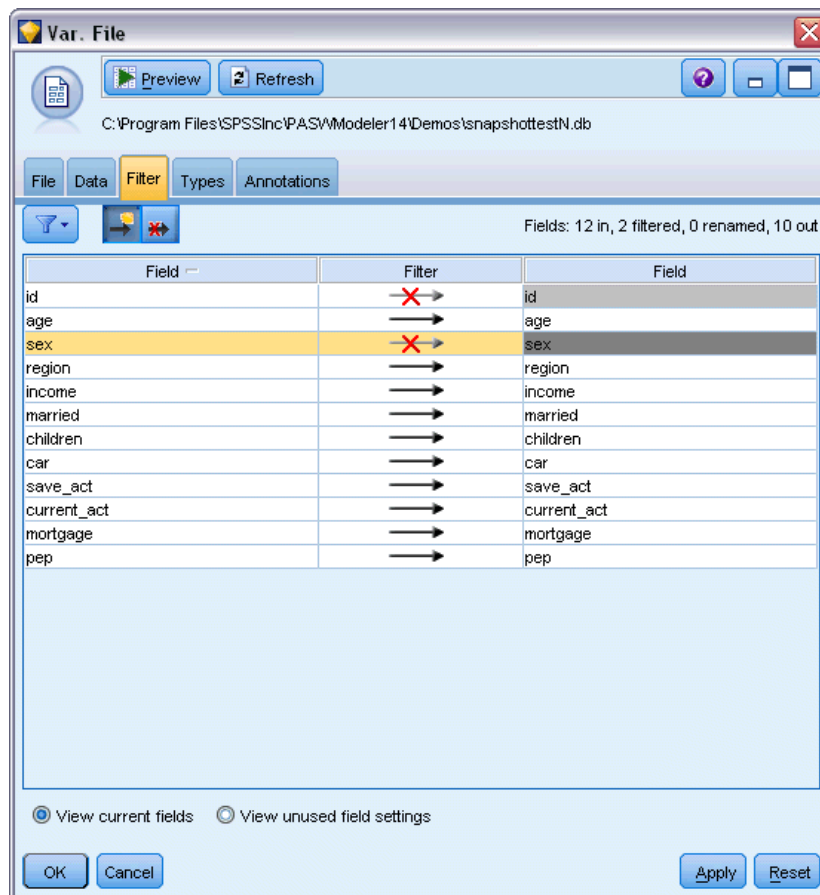
- The dataset is small.
- You plan to derive new fields using the Expression Builder (instantiating makes field values available from the Expression Builder).

Generally, if your dataset is not very large and you do not plan to add fields later in the stream, instantiating at the source node is the most convenient method.

Filtering Fields from the Source Node

The Filter tab on a source node dialog box enables you to exclude fields from downstream operations based on your initial examination of the data. This is useful, for example, if there are duplicate fields in the data or if you are already familiar enough with the data to exclude irrelevant fields. Alternatively, you can add a separate Filter node later in the stream. The functionality is similar in both cases. For more information, see the topic [Setting Filtering Options](#) in Chapter 4 on p. 131.

Figure 2-35
Filtering fields from the source node



Record Operations Nodes

Overview of Record Operations

Record operations nodes are used to make changes to data at the record level. These operations are important during the **Data Understanding** and **Data Preparation** phases of data mining because they allow you to tailor the data to your particular business need.

For example, based on the results of the data audit conducted using the Data Audit node (Output palette), you might decide that you would like customer purchase records for the past three months to be merged. Using a Merge node, you can merge records based on the values of a key field, such as *Customer ID*. Or you might discover that a database containing information about Web site hits is unmanageable with over one million records. Using a Sample node, you can select a subset of data for use in modeling.

The Record Operations palette contains the following nodes:



The Select node selects or discards a subset of records from the data stream based on a specific condition. For example, you might select the records that pertain to a particular sales region. For more information, see the topic [Select Node](#) on p. 54.



The Sample node selects a subset of records. A variety of sample types are supported, including stratified, clustered, and nonrandom (structured) samples. Sampling can be useful to improve performance, and to select groups of related records or transactions for analysis. For more information, see the topic [Sample Node](#) on p. 55.



The Balance node corrects imbalances in a dataset, so it conforms to a specified condition. The balancing directive adjusts the proportion of records where a condition is true by the factor specified. For more information, see the topic [Balance Node](#) on p. 63.



The Aggregate node replaces a sequence of input records with summarized, aggregated output records. For more information, see the topic [Aggregate Node](#) on p. 65.



The Recency, Frequency, Monetary (RFM) Aggregate node enables you to take customers' historical transactional data, strip away any unused data, and combine all of their remaining transaction data into a single row that lists when they last dealt with you, how many transactions they have made, and the total monetary value of those transactions. For more information, see the topic [RFM Aggregate Node](#) on p. 67.



The Sort node sorts records into ascending or descending order based on the values of one or more fields. For more information, see the topic [Sort Node](#) on p. 69.



The Merge node takes multiple input records and creates a single output record containing some or all of the input fields. It is useful for merging data from different sources, such as internal customer data and purchased demographic data. For more information, see the topic [Merge Node](#) on p. 71.



The Append node concatenates sets of records. It is useful for combining datasets with similar structures but different data. For more information, see the topic [Append Node](#) on p. 80.



The Distinct node removes duplicate records, either by passing the first distinct record to the data stream or by discarding the first record and passing any duplicates to the data stream instead. For more information, see the topic [Distinct Node](#) on p. 81.

Many of the nodes in the Record Operations palette require you to use a CLEM expression. If you are familiar with CLEM, you can type an expression in the field. However, all expression fields provide a button that opens the CLEM Expression Builder, which helps you create such expressions automatically.

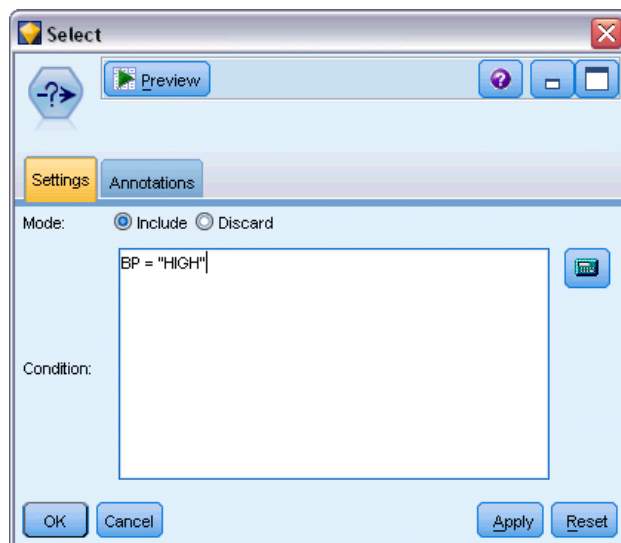
Figure 3-1
Expression Builder button



Select Node

You can use Select nodes to select or discard a subset of records from the data stream based on a specific condition, such as BP (blood pressure) = "HIGH".

Figure 3-2
Select node dialog box



Mode. Specifies whether records that meet the condition will be included or excluded from the data stream.

- **Include.** Select to include records that meet the selection condition.
- **Discard.** Select to exclude records that meet the selection condition.

Condition. Displays the selection condition that will be used to test each record, which you specify using a CLEM expression. Either enter an expression in the window or use the Expression Builder by clicking the calculator (Expression Builder) button to the right of the window.

If you choose to discard records based on a condition, such as the following:

```
(var1='value1' and var2='value2')
```

the Select node by default also discards records having null values for all selection fields. To avoid this, append the following condition to the original one:

```
and not(@NULL(var1) and @NULL(var2))
```

Select nodes are also used to choose a proportion of records. Typically, you would use a different node, the Sample node, for this operation. However, if the condition you want to specify is more complex than the parameters provided, you can create your own condition using the Select node. For example, you can create a condition such as:

```
BP = "HIGH" and random(10) <= 4
```

This will select approximately 40% of the records showing high blood pressure and pass those records downstream for further analysis.

Sample Node

You can use Sample nodes to select a subset of records for analysis, or to specify a proportion of records to discard. A variety of sample types are supported, including stratified, clustered, and nonrandom (structured) samples. Sampling can be used for several reasons:

- To improve performance by estimating models on a subset of the data. Models estimated from a sample are often as accurate as those derived from the full dataset, and may be more so if the improved performance allows you to experiment with different methods you might not otherwise have attempted.
- To select groups of related records or transactions for analysis, such as selecting all the items in an online shopping cart (or market basket), or all the properties in a specific neighborhood.
- To identify units or cases for random inspection in the interest of quality assurance, fraud prevention, or security.

Note: If you simply want to partition your data into training and test samples for purposes of validation, a Partition node can be used instead. For more information, see the topic [Partition Node](#) in Chapter 4 on p. 176.

Types of Samples

Clustered samples. Sample groups or clusters rather than individual units. For example, suppose you have a data file with one record per student. If you cluster by school and the sample size is 50%, then 50% of schools will be chosen and all students from each selected school will be picked. Students in unselected schools will be rejected. On average, you would expect about 50% of students to be picked, but because schools vary in size, the percentage may not be exact. Similarly, you could cluster shopping cart items by transaction ID to make sure that all items from selected transactions are maintained. For an example that clusters properties by town, see the *complexsample_property.str* sample stream.

Stratified samples. Select samples independently within non-overlapping subgroups of the population, or strata. For example, you can ensure that men and women are sampled in equal proportions, or that every region or socioeconomic group within an urban population is represented. You can also specify a different sample size for each stratum (for example, if you think that one group has been under-represented in the original data). For an example that stratifies properties by county, see the *complexsample_property.str* sample stream.

Systematic or 1-in-n sampling. When selection at random is difficult to obtain, units can be sampled systematically (at a fixed interval) or sequentially.

Sampling weights. Sampling weights are automatically computed while drawing a complex sample and roughly correspond to the “frequency” that each sampled unit represents in the original data. Therefore, the sum of the weights over the sample should estimate the size of the original data.

Sampling Frame

A sampling frame defines the potential source of cases to be included in a sample or study. In some cases, it may be feasible to identify every single member of a population and include any one of them in a sample—for example, when sampling items that come off a production line. More often, you will not be able to access every possible case. For example, you cannot be sure who will vote in an election until after the election has happened. In this case, you might use the electoral register as your sampling frame, even though some registered people won’t vote, and some people may vote despite not having been listed at the time you checked the register. Anybody not in the sampling frame has no prospect of being sampled. Whether your sampling frame is close enough in nature to the population you are trying to evaluate is a question that must be addressed for each real-life case.

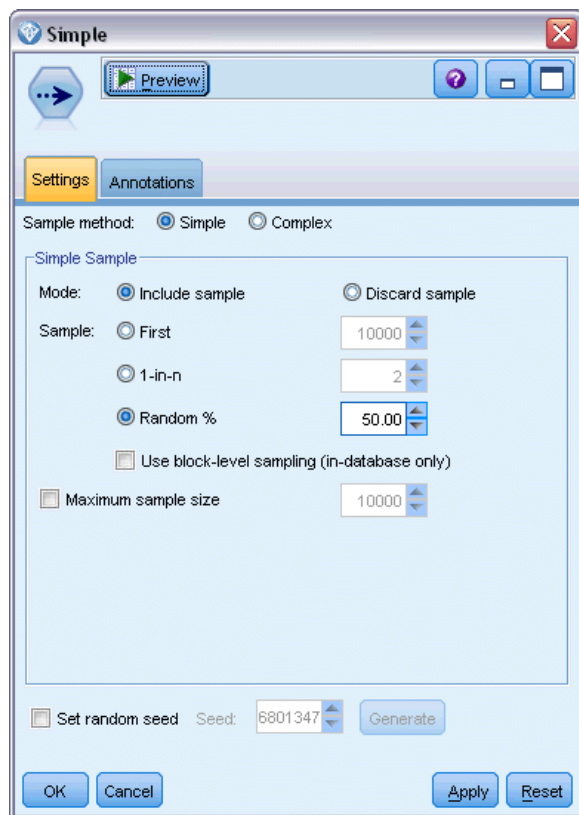
Sample Node Options

You can choose the Simple or Complex method as appropriate for your requirements.

Simple Sampling Options

The Simple method allows you to select a random percentage of records, select contiguous records, or select every *n*th record.

Figure 3-3
Simple sampling options



Mode. Select whether to pass (include) or discard (exclude) records for the following modes:

- **Include sample.** Includes selected records in the data stream and discards all others. For example, if you set the mode to Include sample and set the 1-in-n option to 5, then every fifth record will be included, yielding a dataset that is roughly one-fifth the original size. This is the default mode when sampling data, and the only mode when using the complex method.
- **Discard sample.** Excludes selected records and includes all others. For example, if you set the mode to Discard sample and set the 1-in-n option to 5, then every fifth record will be discarded. This mode is only available with the simple method.

Sample. Select the method of sampling from the following options:

- **First.** Select to use contiguous data sampling. For example, if the maximum sample size is set to 10000, then the first 10,000 records will be selected.
- **1-in-n.** Select to sample data by passing or discarding every n th record. For example, if n is set to 5, then every fifth record will be selected.
- **Random %.** Select to sample a random percentage of the data. For example, if you set the percentage to 20, then 20% of the data will either be passed to the data stream or discarded, depending on the mode selected. Use the field to specify a sampling percentage. You can also specify a seed value using the Set random seed control.

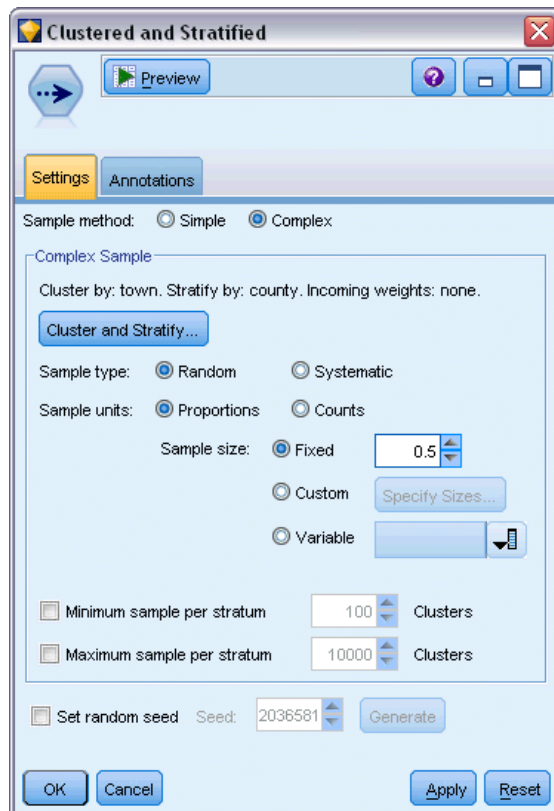
Use block level sampling (in-database only). This option is enabled only if you choose random percentage sampling when performing in-database mining on an Oracle or IBM DB2 database. In these circumstances, block-level sampling can be more efficient.

Maximum sample size. Specifies the maximum number of records to include in the sample. This option is redundant and therefore disabled when First and Include are selected. Also note that when used in combination with the Random % option, this setting may prevent certain records from being selected. For example, if you have 10 million records in your dataset, and you select 50% of records with a maximum sample size of three million records, then 50% of the first six million records will be selected, and the remaining four million records have no chance of being selected. To avoid this limitation, select the Complex sampling method, and request a random sample of three million records without specifying a cluster or stratify variable.

Complex Sampling Options

Complex sample options allow for finer control of the sample, including clustered, stratified, and weighted samples along with other options.

Figure 3-4
Complex sampling options



Cluster and stratify. Allows you to specify cluster, stratify, and input weight fields if needed. For more information, see the topic [Cluster and Stratify Settings](#) on p. 59.

Sample type.

- **Random.** Selects clusters or records randomly within each strata.
- **Systematic.** Selects records at a fixed interval. This option works like the *1 in n* method, except the position of the first record changes depending on a random seed. The value of *n* is determined automatically based on the sample size or proportion.

Sample units. You can select proportions or counts as the basic sample units.

Sample size. You can specify the sample size in several ways:

- **Fixed.** Allows you to specify the overall size of the sample as a count or proportion.
- **Custom.** Allows you to specify the sample size for each subgroup or strata. This option is only available if a stratification field has been specified in the Cluster and Stratify sub dialog box.
- **Variable.** Allows the user to pick a field that defines the sample size for each subgroup or strata. This field should have the same value for each record within a particular stratum; for example, if the sample is stratified by county, then all records with *county = Surrey* must have the same value. The field must be numeric and its values must match the selected sample units. For proportions, values should be greater than 0 and less than 1; for counts, the minimum value is 1.

Minimum sample per stratum. Specifies a minimum number of records (or minimum number of clusters if a cluster field is specified).

Maximum sample per stratum. Specifies a maximum number of records or clusters. If you select this option without specifying a cluster or stratify field, a random or systematic sample of the specified size will be selected.

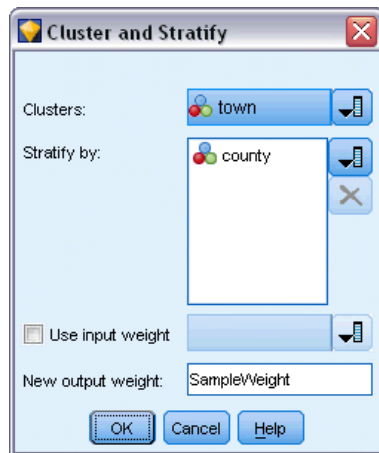
Set random seed. When sampling or partitioning records based on a random percentage, this option allows you to duplicate the same results in another session. By specifying the starting value used by the random number generator, you can ensure the same records are assigned each time the node is executed. Enter the desired seed value, or click the Generate button to automatically generate a random value. If this option is not selected, a different sample will be generated each time the node is executed.

Note: When using the Set random seed option with records read from a database, a Sort node may be required prior to sampling in order to ensure the same result each time the node is executed. This is because the random seed depends on the order of records, which is not guaranteed to stay the same in a relational database. For more information, see the topic [Sort Node](#) on p. 69.

Cluster and Stratify Settings

The Cluster and Stratify dialog box allows you to select cluster, stratification, and weight fields when drawing a complex sample.

Figure 3-5
Cluster and stratification field settings



Clusters. Specifies a categorical field used to cluster records. Records are sampled based on cluster membership, with some clusters included and others not. But if any record from a given cluster is included, all are included. For example, when analyzing product associations in shopping carts, you could cluster items by transaction ID to make sure that all items from selected transactions are maintained. Instead of sampling records—which would destroy information about what items are sold together—you can sample transactions to make sure that all records for selected transactions are preserved.

Stratify by. Specifies a categorical field used to stratify records so that samples are selected independently within non-overlapping subgroups of the population, or strata. If you select a 50% sample stratified by gender, for example, then two 50% samples will be taken, one for the men and one for the women. For example, strata may be socioeconomic groups, job categories, age groups, or ethnic groups, allowing you to ensure adequate sample sizes for subgroups of interest. If there are three times more women than men in the original dataset, this ratio will be preserved by sampling separately from each group. Multiple stratification fields can also be specified (for example, sampling product lines within regions or vice-versa).

Note: If you stratify by a field that has missing values (null or system missing values, empty strings, white space, and blank or user-defined missing values), then you cannot specify custom sample sizes for strata. If you want to use custom sample sizes when stratifying by a field with missing or blank values, then you need to fill them upstream.

Use input weight. Specifies a field used to weight records prior to sampling. For example, if the weight field has values ranging from 1 to 5, records weighted 5 are five times as likely to be selected. The values of this field will be overwritten by the final output weights generated by the node (see following paragraph).

New output weight. Specifies the name of the field where final weights are written if no input weight field is specified. (If an input weight field is specified, its values are replaced by the final weights as noted above, and no separate output weight field is created.) The output weight values indicate the number of records represented by each sampled record in the original data. The sum of the weight values gives an estimate of the sample size. For example, if a random 10% sample is taken, the output weight will be 10 for all records, indicating that each sampled record represents

roughly ten records in the original data. In a stratified or weighted sample, the output weight values may vary based on the sample proportion for each stratum.

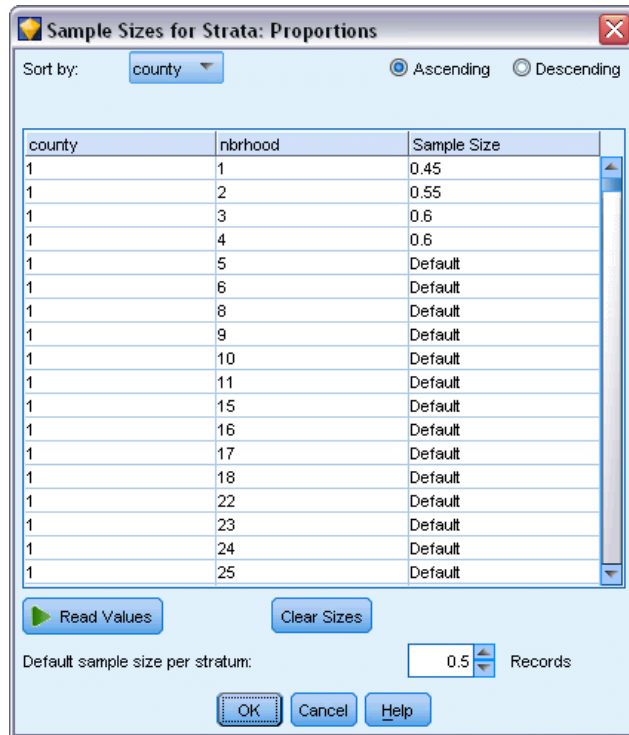
Comments

- Clustered sampling is useful if you cannot get a complete list of the population you want to sample, but can get complete lists for certain groups or clusters. It is also used when a random sample would produce a list of test subjects that it would be impractical to contact. For example, it would be easier to visit all farmers in one county than a selection of farmers scattered across every county in the nation.
- You can specify both cluster and stratify fields in order to sample clusters independently within each strata. For example, you could sample property values stratified by county, and cluster by town within each county. This will ensure that an independent sample of towns is drawn from within each county. Some towns will be included and others will not, but for each town that is included, all properties within the town are included.
- To select a random sample of units from within each cluster, you can string two Sample nodes together. For example, you could first sample townships stratified by county as described above. Then attach a second Sample node and select *town* as a stratify field, allowing you to sample a proportion of records from within each township.
- In cases where a combination of fields is required to uniquely identify clusters, a new field can be generated using a Derive node. For example, if multiple shops use the same numbering system for transactions, you could derive a new field that concatenates the shop and transaction IDs.

Sample Sizes for Strata

When drawing a stratified sample, the default option is to sample the same proportion of records or clusters from each stratum. If one group outnumbered another by a factor of 3, for example, you typically want to preserve the same ratio in the sample. If this is not the case, however, you can specify the sample size separately for each stratum.

Figure 3-6
Specifying sample sizes for strata



The Sample Sizes for Strata dialog box lists each value of the stratification field, allowing you to override the default for that stratum. If multiple stratification fields are selected, every possible combination of values is listed, allowing you to specify the size for each ethnic group within each city, for example, or each town within each county. Sizes are specified as proportions or counts, as determined by the current setting in the Sample node.

To Specify Sample Sizes for Strata

- ▶ In the Sample node, select Complex, and select one or more stratification fields. For more information, see the topic [Cluster and Stratify Settings](#) on p. 59.
- ▶ Select Custom, and select Specify Sizes.
- ▶ In the Sample Sizes for Strata dialog box, click the Read Values button at lower left to populate the display. If necessary, you may need to instantiate values in an upstream source or Type node. For more information, see the topic [What Is Instantiation?](#) in Chapter 4 on p. 118.
- ▶ Click in any row to override the default size for that stratum.

Notes on Sample Size

Custom sample sizes may be useful if different strata have different variances, for example, in order to make sample sizes proportional to the standard deviation. (If the cases within the stratum are more varied, you need to sample more of them to get a representative sample.) Or if

a stratum is small, you may wish to use a higher sample proportion to ensure that a minimum number of observations is included.

Note: If you stratify by a field that has missing values (null or system missing values, empty strings, white space, and blank or user-defined missing values), then you cannot specify custom sample sizes for strata. If you want to use custom sample sizes when stratifying by a field with missing or blank values, then you need to fill them upstream.

Balance Node

You can use Balance nodes to correct imbalances in datasets so they conform to specified test criteria. For example, suppose that a dataset has only two values—*low* or *high*—and that 90% of the cases are *low* while only 10% of the cases are *high*. Many modeling techniques have trouble with such biased data because they will tend to learn only the *low* outcome and ignore the *high* one, since it is more rare. If the data are well balanced with approximately equal numbers of *low* and *high* outcomes, models will have a better chance of finding patterns that distinguish the two groups. In this case, a Balance node is useful for creating a balancing directive that reduces cases with a *low* outcome.

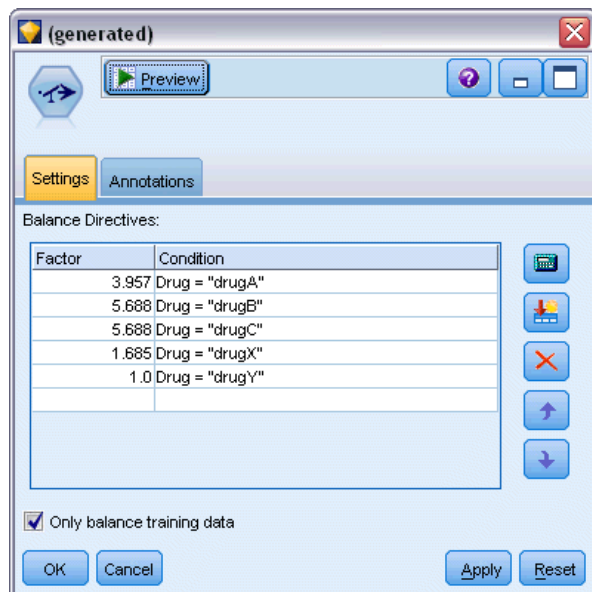
Balancing is carried out by duplicating and then discarding records based on the conditions you specify. Records for which no condition holds are always passed through. Because this process works by duplicating and/or discarding records, the original sequence of your data is lost in downstream operations. Be sure to derive any sequence-related values before adding a Balance node to the data stream.

Note: Balance nodes can be generated automatically from distribution charts and histograms. For example, you can balance your data to show equal proportions across all categories of a categorical field, as shown in a distribution plot.

Example. When building an RFM stream to identify recent customers who have positively responded to previous marketing campaigns, the marketing department of a sales company uses a Balance node to balance the differences between true and false responses in the data.

Setting Options for the Balance Node

Figure 3-7
Balance node settings



Record balancing directives. Lists the current balancing directives. Each directive includes both a factor and a condition that tells the software to “increase the proportion of records by a factor specified where the condition is true.” A factor lower than 1.0 means that the proportion of indicated records will be decreased. For example, if you want to decrease the number of records where drug Y is the treatment drug, you might create a balancing directive with a factor of 0.7 and a condition Drug = "drugY". This directive means that the number of records where drug Y is the treatment drug will be reduced to 70% for all downstream operations.

Note: Balance factors for reduction may be specified to four decimal places. Factors set below 0.0001 will result in an error, since the results do not compute correctly.

- **Create conditions** by clicking the button to the right of the text field. This inserts an empty row for entering new conditions. To create a CLEM expression for the condition, click the Expression Builder button.
- **Delete directives** using the red delete button.
- **Sort directives** using the up and down arrow buttons.

Only balance training data. If a partition field is present in the stream, this option balances data in the training partition only. In particular, this may be useful if generating adjusted propensity scores, which require an unbalanced testing or validation partition. If no partition field is present in the stream (or if multiple partition fields are specified), then this option is ignored and all data are balanced.

Aggregate Node

Aggregation is a data preparation task frequently used to reduce the size of a dataset. Before proceeding with aggregation, you should take time to clean the data, concentrating especially on missing values. Once you have aggregated, potentially useful information regarding missing values may be lost.

You can use an Aggregate node to replace a sequence of input records with summary, aggregated output records. For example, you might have a set of input records such as:

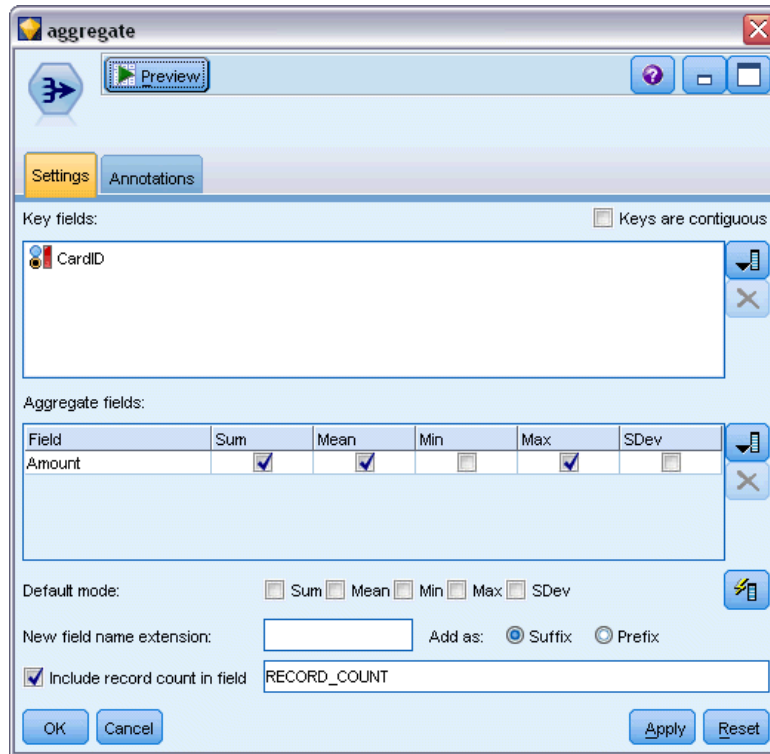
Age	Sex	Region	Branch	Sales
23	M	S	8	4
45	M	S	16	4
37	M	S	8	5
30	M	S	5	7
44	M	N	4	9
25	M	N	2	11
29	F	S	16	6
41	F	N	4	8
23	F	N	6	2
45	F	N	4	5
33	F	N	6	10

You can aggregate these records with *Sex* and *Region* as key fields. Then choose to aggregate *Age* with the mode Mean and *Sales* with the mode Sum. Select Include record count in field in the Aggregate node dialog box and your aggregated output would be:

Age	Sex	Region	Sales	RECORD_COUNT
35.5	F	N	25	4
34.5	M	N	20	2
29	F	S	6	1
33.75	M	S	20	4

Note: Fields such as *Branch* are automatically discarded when no aggregate mode is specified.

Figure 3-8
Aggregate node dialog box



Setting Options for the Aggregate Node

Key fields. Lists fields that can be used as keys for aggregation. Both numeric and symbolic fields can be used as keys. If you choose more than one key field, the values will be combined to produce a key value for aggregating records. One aggregated record will be generated for each unique key field. For example, if *Sex* and *Region* are your key fields, each unique combination of *M* and *F* with regions *N* and *S* (four unique combinations) will have an aggregated record. To add a key field, use the Field Chooser button to the right of the window.

Keys are contiguous. Select this option if you know that all records with the same key values are grouped together in the input (for example, if the input is sorted on the key fields). Doing so can improve performance.

Aggregate fields. Lists the numeric fields whose values will be aggregated as well as the selected modes of aggregation. To add fields to this list, use the Field Chooser button on the right.

Default mode. Specify the default aggregation mode to be used for newly added fields. If you frequently use the same aggregation, select one or more modes here and use the Apply to All button on the right to apply the selected modes to all fields listed above. The following aggregation modes are available:

- **Sum.** Select to return summed values for each key field combination.
- **Mean.** Select to return the mean values for each key field combination.

- **Min.** Select to return minimum values for each key field combination.
- **Max.** Select to return maximum values for each key field combination.
- **SDev.** Select to return the standard deviation for each key field combination.

New field name extension. Select to add a suffix or prefix, such as “1” or “new,” to duplicate aggregated fields. For example, the result of a minimum values aggregation on the field *Age* will produce a field name called *Age_Min_1* if you have selected the suffix option and specified “1” as the extension. *Note:* Aggregation extensions such as *_Min* or *_Max* are automatically added to the new field, indicating the type of aggregation performed. Select *Suffix* or *Prefix* to indicate your preferred extension style.

Include record count in field. Select to include an extra field in each output record called *Record_Count*, by default. This field indicates how many input records were aggregated to form each aggregate record. Create a custom name for this field by typing in the edit field.

Note: System null values are excluded when aggregates are computed, but they are included in the record count. Blank values, on the other hand, are included in both aggregation and record count. To exclude blank values, you can use a *Filler* node to replace blanks with null values. You can also remove blanks using a *Select* node.

Performance

Aggregations operations may benefit from enabling parallel processing.

RFM Aggregate Node

The Recency, Frequency, Monetary (RFM) Aggregate node enables you to take customers’ historical transactional data, strip away any unused data, and combine all of their remaining transaction data into a single row, using their unique customer ID as a key, that lists when they last dealt with you (recency), how many transactions they have made (frequency), and the total value of those transactions (monetary).

Before proceeding with any aggregation, you should take time to clean the data, concentrating especially on any missing values.

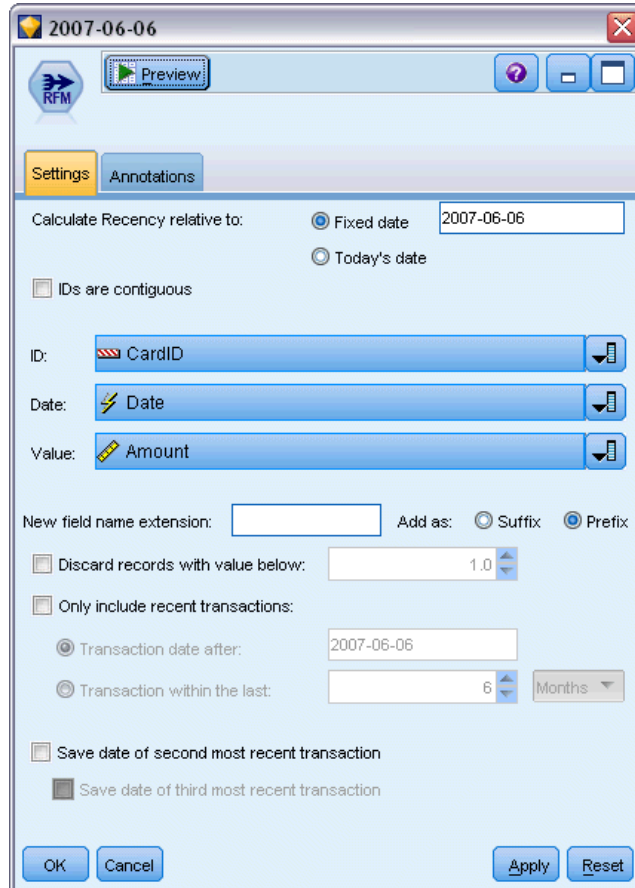
Once you have identified and transformed the data using the RFM Aggregate node, you may use an RFM Analysis node to carry out further analysis. For more information, see the topic [RFM Analysis Node](#) in Chapter 4 on p. 173.

Note that once the data file has been run through the RFM Aggregate node, it will not have any target values; therefore, before being able to use it as inputs for further predictive analysis with any modeling nodes such as C5.0 or CHAID, you will need to merge it with other customer data (for example, by matching the customer IDs). For more information, see the topic [Merge Node](#) on p. 71.

The RFM Aggregate and RFM Analysis nodes in IBM® SPSS® Modeler are set up to use independent binning; that is, they rank and bin data on each measure of recency, frequency, and monetary value, without regard to their values or the other two measures.

Setting Options for the RFM Aggregate Node

Figure 3-9
RFM aggregate settings



Calculate recency relative to. Specify the date from which the recency of transactions will be calculated. This may be either a Fixed date that you enter, or Today's date, as set by your system. Today's date is entered by default and is automatically updated when the node is executed.

IDs are contiguous. If your data are presorted so that all records with the same ID appear together in the data stream, select this option to speed up processing. If your data are not presorted (or you are not sure), leave this option unselected, and the node will sort the data automatically.

ID. Select the field to be used to identify the customer and their transactions. To display the fields from which you can select, use the Field Chooser button on the right.

Date. Select the date field to be used to calculate recency against. To display the fields from which you can select, use the Field Chooser button on the right.

Note that this requires a field with a storage of date, or timestamp, in the appropriate format to use as input. For example, if you have a string field with values like *Jan 2007*, *Feb 2007*, and so on, you can convert this to a date field using a Filler node and the `to_date()` function. For more information, see the topic [Storage Conversion Using the Filler Node](#) in Chapter 4 on p. 153.

Value. Select the field to be used to calculate the total monetary value of the customer's transactions. To display the fields from which you can select, use the Field Chooser button on the right. *Note:* This must be a numeric value.

New field name extension. Select to append either a suffix or prefix, such as "12_month", to the newly generated recency, frequency, and monetary fields. Select Suffix or Prefix to indicate your preferred extension style. For example, this may be useful when examining several time periods.

Discard records with value below. If required, you can specify a minimum value below which any transaction details are not used when calculating the RFM totals. The units of value relate to the Value field selected.

Include only recent transactions. If you are analyzing a large database, you can specify that only the latest records are used. You can choose to use the data recorded either after a certain date or within a recent period:

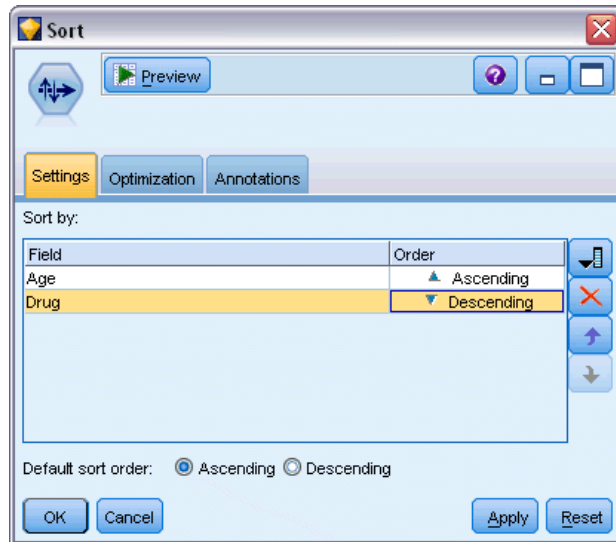
- **Transaction date after.** Specify the transaction date after which records will be included in your analysis.
- **Transaction within the last.** Specify the number and type of periods (days, weeks, months, or years) back from the Calculate recency relative to date after which records will be included in your analysis.

Save date of second most recent transaction. If you want to know the date of the second most recent transaction for each customer, select this box. In addition, you can then select the Save date of third most recent transaction box as well. For example, this can help you identify customers who may have carried out many transactions some considerable time ago, but only one recent transaction.

Sort Node

You can use Sort nodes to sort records into ascending or descending order based on the values of one or more fields. For example, Sort nodes are frequently used to view and select records with the most common data values. Typically, you would first aggregate the data using the Aggregate node and then use the Sort node to sort the aggregated data into descending order of record counts. Displaying these results in a table will allow you to explore the data and to make decisions, such as selecting the records of the top 10 best customers.

Figure 3-10
Sort node dialog box



Sort by. All fields selected to use as sort keys are displayed in a table. A key field works best for sorting when it is numeric.

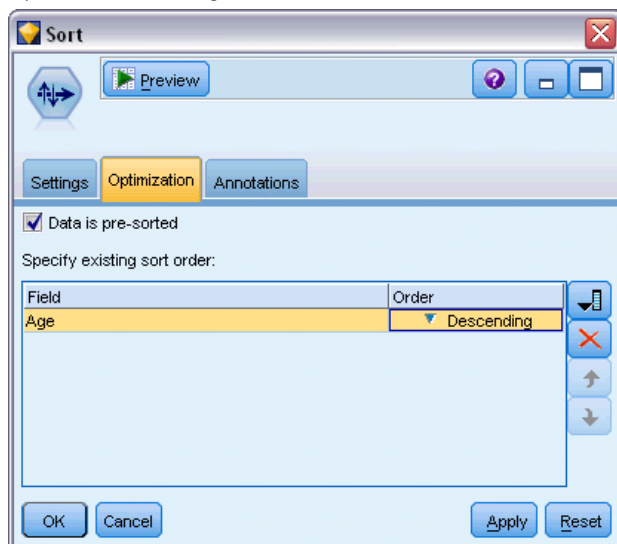
- **Add fields** to this list using the Field Chooser button on the right.
- **Select an order** by clicking the Ascending or Descending arrow in the table's *Order* column.
- **Delete fields** using the red delete button.
- **Sort directives** using the up and down arrow buttons.

Default sort order. Select either Ascending or Descending to use as the default sort order when new fields are added above.

Sort Optimization Settings

If you are working with data you know are already sorted by some key fields, you can specify which fields are already sorted, allowing the system to sort the rest of the data more efficiently. For example, you want to sort by *Age* (descending) and *Drug* (ascending) but know your data are already sorted by *Age* (descending).

Figure 3-11
Optimization settings



Data is presorted. Specifies whether the data are already sorted by one or more fields.

Specify existing sort order. Specify the fields that are already sorted. Using the Select Fields dialog box, add fields to the list. In the *Order* column, specify whether each field is sorted in ascending or descending order. If you are specifying multiple fields, make sure that you list them in the correct sorting order. Use the arrows to the right of the list to arrange the fields in the correct order. If you make a mistake in specifying the correct existing sort order, an error will appear when you run the stream, displaying the record number where the sorting is inconsistent with what you specified.

Note: Sorting speed may benefit from enabling parallel processing.

Merge Node

The function of a Merge node is to take multiple input records and create a single output record containing all or some of the input fields. This is a useful operation when you want to merge data from different sources, such as internal customer data and purchased demographic data. There are two ways to merge data:

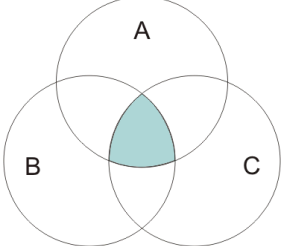
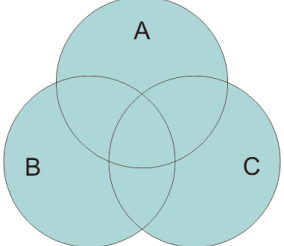
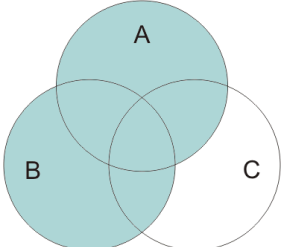
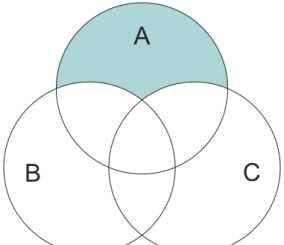
- **Merge by order** concatenates corresponding records from all sources in the order of input until the smallest data source is exhausted. It is important if using this option that you have sorted your data using a Sort node.
- **Merge using a key field**, such as *Customer ID*, to specify how to match records from one data source with records from the other(s). Several types of joins are possible, including inner join, full outer join, partial outer join, and anti-join. For more information, see the topic [Types of Joins](#) on p. 72.

Types of Joins

When using a key field for data merging, it is useful to spend some time thinking about which records will be excluded and which will be included. There are a variety of joins, which are discussed in detail below.

The two basic types of joins are referred to as inner and outer joins. These methods are frequently used to merge tables from related datasets based on common values of a key field, such as *Customer ID*. Inner joins allow for clean merging and an output dataset that includes only complete records. Outer joins also include complete records from the merged data, but they also allow you to include unique data from one or more input tables.

The types of joins allowed are described in greater detail below.

	<p>An inner join includes only records in which a value for the key field is common to all input tables. That is, unmatched records will not be included in the output dataset.</p>
	<p>A full outer join includes all records, both matching and nonmatching, from the input tables. Left and right outer joins are referred to as partial outer joins and are described below.</p>
	<p>A partial outer join includes all records matched using the key field as well as unmatched records from specified tables. (Or, to put it another way, all records from some tables and only matching records from others.) Tables (such as A and B shown here) can be selected for inclusion in the outer join using the Select button on the Merge tab. Partial joins are also called left or right outer joins when only two tables are being merged. Since IBM® SPSS® Modeler allows the merging of more than two tables, we refer to this as a partial outer join.</p>
	<p>An anti-join includes only unmatched records for the first input table (Table A shown here). This type of join is the opposite of an inner join and does not include complete records in the output dataset.</p>

For example, if you have information about farms in one dataset and farm-related insurance claims in another, you can match the records from the first source to the second source using the Merge options.

To determine if a customer in your farm sample has filed an insurance claim, use the inner join option to return a list showing where all IDs match from the two samples.

Figure 3-12

Sample output for an inner join merge

	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...

Using the full outer join option returns both matching and nonmatching records from the input tables. The system-missing value (\$null\$) will be used for any incomplete values.

Figure 3-13

Sample output for a full outer join merge

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimtype	claimvalue
1	id601	\$null\$	\$null\$	\$null\$	\$null\$	\$null\$	\$null\$	\$null\$	decomm...	74703.1C
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

A partial outer join includes all records matched using the key field as well as unmatched records from specified tables. The table displays all of the records matched from the ID field as well as the records matched from the first dataset.

Figure 3-14

Sample output for a partial outer join merge

	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

If you are using the anti-join option, the table returns only unmatched records for the first input table.

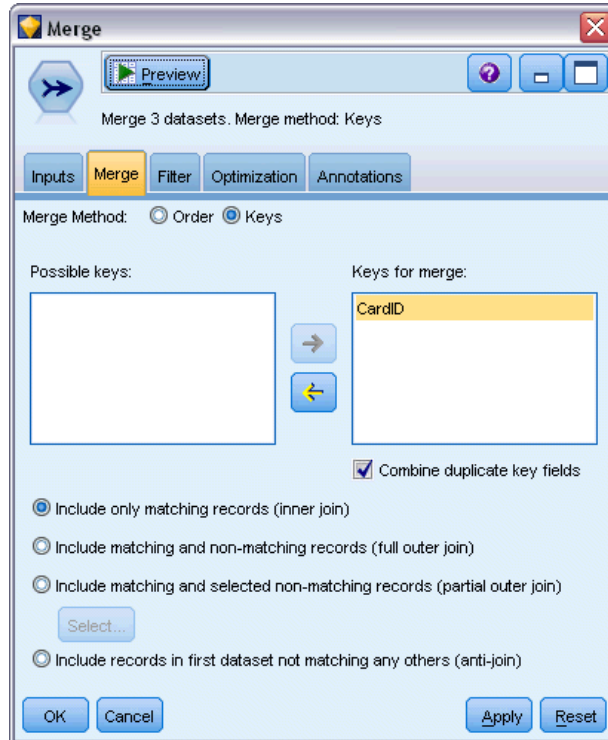
Figure 3-15

Sample output for an anti-join merge

	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

Specifying a Merge Method and Keys

Figure 3-16
Using the Merge tab to set merge method options



Merge Method. Select either Order or Keys to specify the method of merging records. Selecting Keys activates the bottom half of the dialog box.

- **Order.** Merges records by order such that the n th record from each input is merged to produce the n th output record. When any record runs out of a matching input record, no more output records are produced. This means that the number of records created is the number of records in the smallest dataset.
- **Keys.** Uses a key field, such as *Transaction ID*, to merge records with the same value in the key field. This is equivalent to a database “equi-join.” If a key value occurs more than once, all possible combinations are returned. For example, if records with the same key field value A contain differing values B , C , and D in other fields, the merged fields will produce a separate record for each combination of A with value B , A with value C , and A with value D .

Note: Null values are not considered identical in the merge-by-key method and will not join.

Possible keys. Lists only those fields with exactly matching field names in all input data sources. Select a field from this list and use the arrow button to add it as a key field used for merging records. More than one key field may be used. You can rename non-matching input fields by means of a Filter node, or the Filter tab of a source node.

Keys for merge. Lists all fields used to merge records from all input data sources based on values of the key fields. To remove a key from the list, select one and use the arrow button to return it to the Possible Keys list. When more than one key field is selected, the option below is enabled.

Combine duplicate key fields. When more than one key field is selected above, this option ensures that there is only one output field of that name. This option is enabled by default except in the case when streams have been imported from earlier versions of IBM® SPSS® Modeler. When this option is disabled, duplicate key fields must be renamed or excluded using the Filter tab in the Merge node dialog box.

Include only matching records (inner join). Select to merge only complete records.

Include matching and non-matching records (full outer join). Select to perform a “full outer join.” This means that if values for the key field are not present in all input tables, the incomplete records are still retained. The undefined value (\$null\$) is added to the key field and included in the output record.

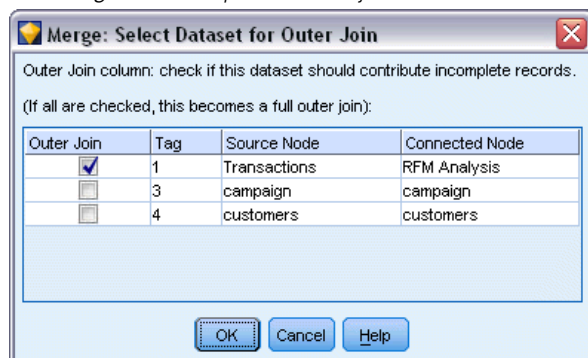
Include matching and selected non-matching records (partial outer join). Select to perform a “partial outer join” of the tables you select in a subdialog box. Click Select to specify tables for which incomplete records will be retained in the merge.

Include records in the first dataset not matching any others (anti-join). Select to perform a type of “anti-join,” where only nonmatching records from the first dataset are passed downstream. You can specify the order of input datasets using arrows on the Inputs tab. This type of join does not include complete records in the output dataset. For more information, see the topic [Types of Joins](#) on p. 72.

Selecting Data for Partial Joins

For a partial outer join, you must select the table(s) for which incomplete records will be retained. For example, you may want to retain all records from a Customer table while retaining only matched records from the Mortgage Loan table.

Figure 3-17
Selecting data for a partial outer join

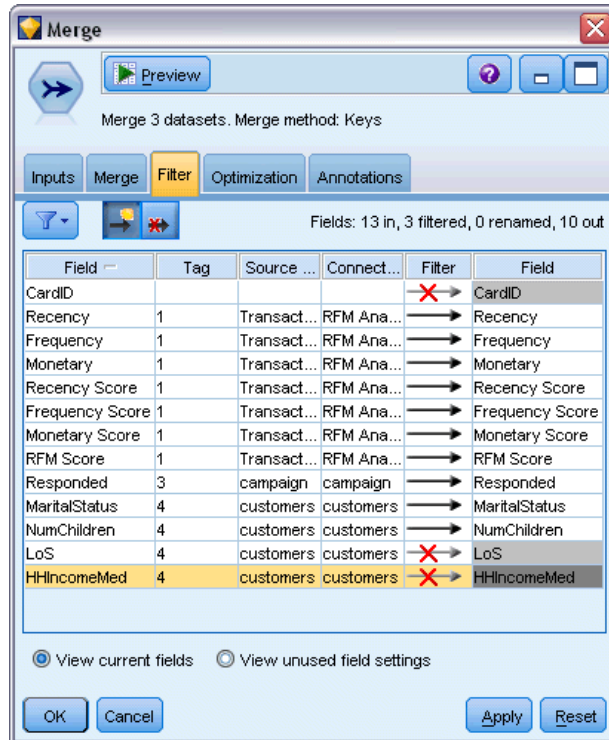


Outer Join column. In the *Outer Join* column, select datasets to include in their entirety. For a partial join, overlapping records will be retained as well as incomplete records for datasets selected here. For more information, see the topic [Types of Joins](#) on p. 72.

Filtering Fields from the Merge Node

Merge nodes include a convenient way of filtering or renaming duplicate fields as a result of merging multiple data sources. Click the Filter tab in the dialog box to select filtering options.

Figure 3-18
Filtering from the Merge node



The options presented here are nearly identical to those for the Filter node. There are, however, additional options not discussed here that are available on the Filter menu. For more information, see the topic [Filtering or Renaming Fields](#) in Chapter 4 on p. 130.

Field. Displays the input fields from currently connected data sources.

Tag. Lists the tag name (or number) associated with the data source link. Click the Inputs tab to alter active links to this Merge node.

Source node. Displays the source node whose data is being merged.

Connected node. Displays the node name for the node that is connected to the Merge node. Frequently, complex data mining requires several merge or append operations that may include the same source node. The connected node name provides a way of differentiating these.

Filter. Displays the current connections between input and output field. Active connections show an unbroken arrow. Connections with a red X indicate filtered fields.

Field. Lists the output fields after merging or appending. Duplicate fields are displayed in red. Click in the Filter field above to disable duplicate fields.

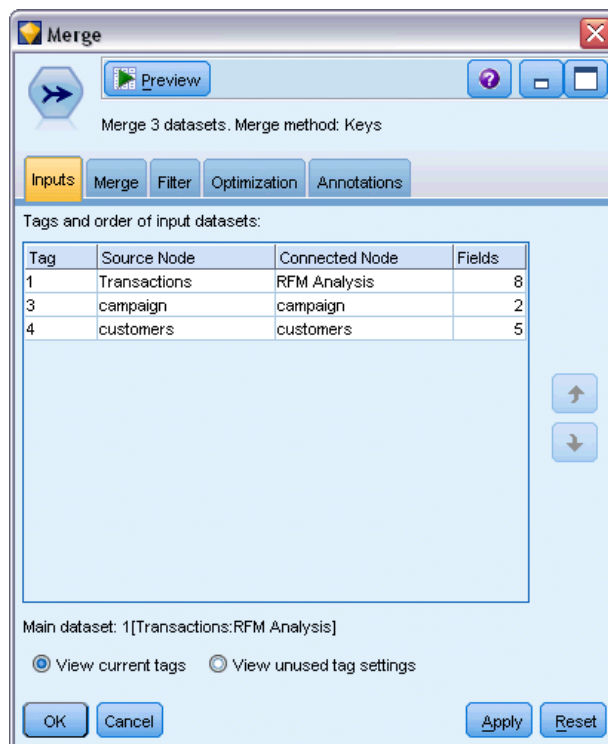
View current fields. Select to view information on fields selected to be used as key fields.

View unused field settings. Select to view information on fields that are not currently in use.

Setting Input Order and Tagging

Using the Inputs tab in the Merge and Append node dialog boxes, you can specify the order of input data sources and make any changes to the tag name for each source.

Figure 3-19
Using the Inputs tab to specify tags and input order



Tags and order of input datasets. Select to merge or append only complete records.

- Tag.** Lists current tag names for each input data source. Tag names, or **tags**, are a way of uniquely identifying the data links for the merge or append operation. For example, imagine water from various pipes that is combined at one point and flows through a single pipe. Data in IBM® SPSS® Modeler flows similarly, and the merging point is often a complex interaction between the various data sources. Tags provide a way of managing the inputs (“pipes”) to a Merge or Append node so that if the node is saved or disconnected, the links remain and are easily identifiable.

When you connect additional data sources to a Merge or Append node, default tags are automatically created using numbers to represent the order in which you connected the nodes. This order is unrelated to the order of fields in the input or output datasets. You can change the default tag by entering a new name in the *Tag* column.

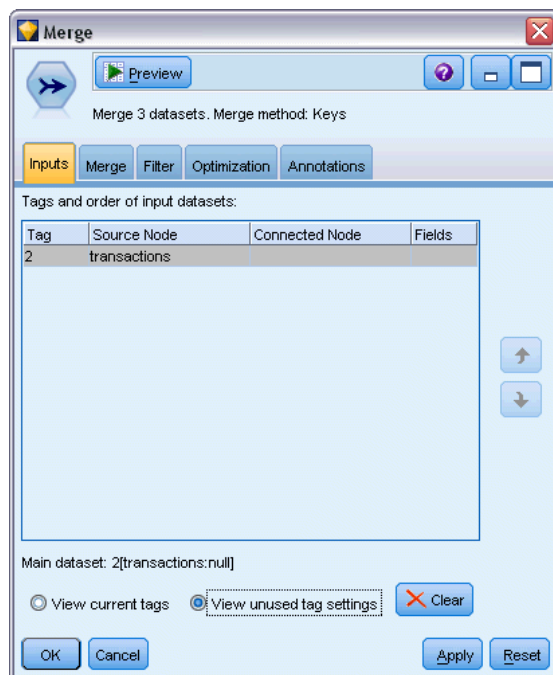
- Source Node.** Displays the source node whose data is being combined.

- **Connected Node.** Displays the node name for the node that is connected to the Merge or Append node. Frequently, complex data mining requires several merge operations that may include the same source node. The connected node name provides a way of differentiating these.
- **Fields.** Lists the number of fields in each data source.

View current tags. Select to view tags that are actively being used by the Merge or Append node. In other words, current tags identify links to the node that have data flowing through. Using the pipe metaphor, current tags are analogous to pipes with existing water flow.

View unused tag settings. Select to view tags, or links, that were previously used to connect to the Merge or Append node but are not currently connected with a data source. This is analogous to empty pipes still intact within a plumbing system. You can choose to connect these “pipes” to a new source or remove them. To remove unused tags from the node, click Clear. This clears all unused tags at once.

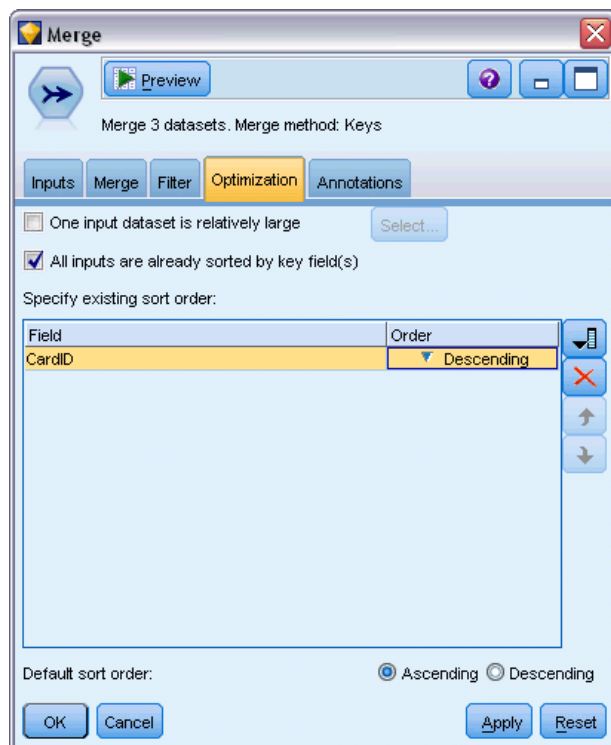
Figure 3-20
Removing unused tags from the Merge node



Merge Optimization Settings

The system provides two options that can help you merge data more efficiently in certain situations. These options allow you to optimize merging when one input dataset is significantly larger than the other datasets or when your data are already sorted by all or some of the key fields that you are using for the merge.

Figure 3-21
Optimization settings



One input dataset is relatively large. Select to indicate that one of the input datasets is much larger than the others. The system will cache the smaller datasets in memory and then perform the merge by processing the large dataset without caching or sorting it. You will commonly use this type of join with data designed using a star-schema or similar design, where there is a large central table of shared data (for example, in transactional data). If you select this option, click *Select* to specify the large dataset. Note that you can select only *one* large dataset. The following table summarizes which joins can be optimized using this method.

Type of Join	Can be optimized for a large input dataset?
Inner	Yes
Partial	Yes, if there are no incomplete records in the large dataset.
Full	No
Anti-join	Yes, if the large dataset is the first input.

All inputs are already sorted by key field(s). Select to indicate that the input data are already sorted by one or more of the key fields that you are using for the merge. Make sure *all* your input datasets are sorted.

Specify existing sort order. Specify the fields that are already sorted. Using the *Select Fields* dialog box, add fields to the list. You can select from only the key fields that are being used for the merge (specified in the *Merge* tab). In the *Order* column, specify whether each field is sorted in ascending or descending order. If you are specifying multiple fields, make sure that you list them in the correct sorting order. Use the arrows to the right of the list to arrange the fields in the

correct order. If you make a mistake in specifying the correct existing sort order, an error will appear when you run the stream, displaying the record number where the sorting is inconsistent with what you specified.

Note: Merging speed may benefit from enabling parallel processing.

Append Node

You can use Append nodes to concatenate sets of records. Unlike Merge nodes, which join records from different sources together, Append nodes read and pass downstream all of the records from one source until there are no more. Then the records from the next source are read using the same data structure (number of records, number of fields, and so on) as the first, or primary, input. When the primary source has more fields than another input source, the system null string (\$null\$) will be used for any incomplete values.

Append nodes are useful for combining datasets with similar structures but different data. For example, you might have transaction data stored in different files for different time periods, such as a sales data file for March and a separate one for April. Assuming that they have the same structure (the same fields in the same order), the Append node will join them together into one large file, which you can then analyze.

Note: In order to append files, the field measurement levels must be similar. For example, a *Nominal* field cannot be appended with a field whose measurement level is *Continuous*.

Figure 3-22
Append node dialog box showing field matching by name



Setting Append Options

Match fields by. Select a method to use when matching fields to append.

- **Position.** Select to append datasets based on the position of fields in the main data source. When using this method, your data should be sorted to ensure proper appending.
- **Name.** Select to append datasets based on the name of fields in the input datasets. Also select Match case to enable case sensitivity when matching field names.

Output Field. Lists the source nodes that are connected to the Append node. The first node on the list is the primary input source. You can sort the fields in the display by clicking on the column heading. This sorting does not actually reorder the fields in the dataset.

Include fields from. Select Main dataset only to produce output fields based on the fields in the main dataset. The main dataset is the first input, specified on the Inputs tab. Select All datasets to produce output fields for all fields in all datasets regardless of whether there is a matching field across all input datasets.

Tag records by including source dataset in field. Select to add an additional field to the output file whose values indicate the source dataset for each record. Specify a name in the text field. The default field name is *Input*.

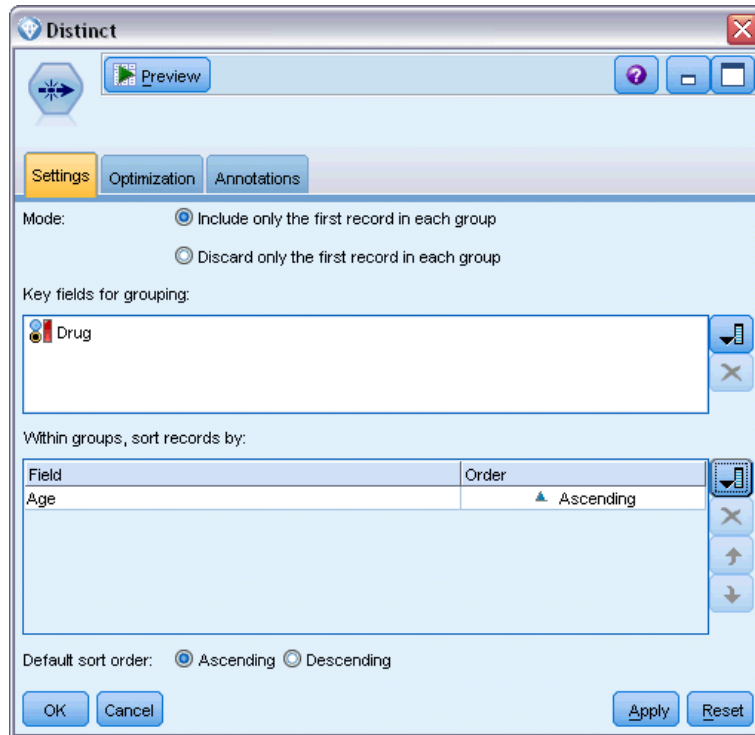
Distinct Node

Duplicate records in a data set must be removed before data mining can begin. For example, in a marketing database, individuals may appear multiple times with different address or company information. You can use the Distinct node to find, or remove, duplicate records in your data set.

Using the Distinct node, you can either remove duplicate records by passing the first distinct record to the data stream, or you can find duplicate records by discarding the first record and passing any duplicates to the data stream instead.

In addition, you can define a sort order within each distinct key value for the returned results. If you want a specific row returned for each distinct key, you must sort the records within the Distinct node rather than using an upstream Sort node (see “Sorting Records Within the Distinct Node” below).

Figure 3-23
Distinct node dialog box



Mode. Specify whether to include or exclude (discard) the first record.

- **Include only the first record in each group.** Includes the first distinct record in the data stream and removes any duplicates.
- **Discard only the first record in each group.** Discards the first distinct record found and passes any duplicate records to the data stream instead. This option is useful for *finding* duplicates in your data so that you can examine them later in the stream.

Key fields for grouping. Lists the field or fields used to determine whether records are identical. You can:

- Add fields to this list using the field picker button on the right.
- Delete fields from the list by using the red X (remove) button.

Within groups, sort records by. Lists the fields used to determine how records are sorted within each distinct key value, and whether they are sorted in ascending or descending order. You can:

- Add fields to this list using the field picker button on the right.
- Delete fields from the list by using the red X (remove) button.
- Move fields using the up or down buttons, if you are sorting by more than one field.

Default sort order. Specify whether, by default, records are sorted in Ascending or Descending order.

Sorting Records Within the Distinct Node

Using the Within groups, sort records by option within the Distinct node you can return a specific row for each distinct key, so there is no need to use a preceding Sort node. For example, assume we have the following data about the ages of prescription drug users:

Age	Drug
50	Drug A
71	Drug B
44	Drug A
65	Drug X
39	Drug A
75	Drug C
72	Drug Y
57	Drug X
79	Drug Y
69	Drug C
74	Drug B
85	Drug Y
69	Drug X

To find the oldest user of each drug we would set the mode to include only the first record in each group, use Drug as the key field, and use Age as the sort field, set in Descending order. The input order does not affect the result because the sort selection specifies which of the many rows for a given Drug is to be returned, and the final data output would be:

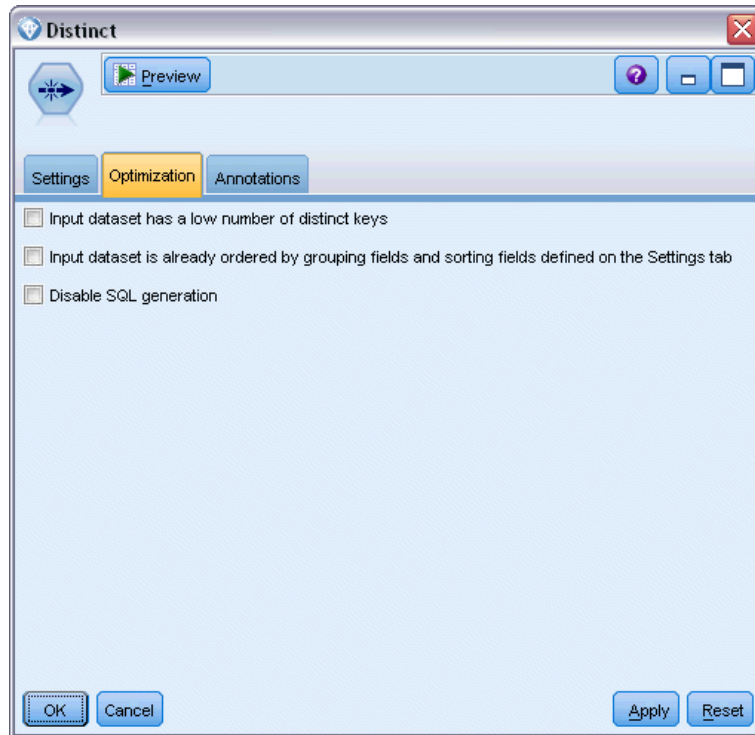
Age	Drug
50	Drug A
74	Drug B
75	Drug C
69	Drug X
85	Drug Y

Distinct Optimization Settings

If the data on which you are working has only a small number of records, or has already been sorted, you can optimize the way in which it is handled to enable IBM® SPSS® Modeler to process the data more efficiently.

Note: If you either select Input dataset has a low number of distinct keys, or use SQL generation for the node, any row within the distinct key value can be returned; to control which row is returned within a distinct key you need to specify the sort order by using the Within groups, sort records by fields on the Settings tab. The optimization options do not affect the results output by the Distinct node as long as you have specified a sort order on the Settings tab.

Figure 3-24
Optimization settings



Input dataset has a low number of distinct keys. Select this option if you have a small number of records, or a small number of unique values of the key field(s), or both. Doing so can improve performance.

Input dataset is already ordered by grouping fields and sorting fields on the Settings tab. Only select this option if your data is already sorted by all of the fields listed under Within groups, sort records by on the Settings tab, and if the ascending or descending sort order of the data is the same. Doing so can improve performance.

Disable SQL generation. Select this option to disable SQL generation for the node.

Field Operations Nodes

Field Operations Overview

After an initial data exploration, you will probably have to select, clean, or construct data in preparation for analysis. The Field Operations palette contains many nodes useful for this transformation and preparation.

For example, using a Derive node, you might create an attribute that is not currently represented in the data. Or you might use a Binning node to recode field values automatically for targeted analysis. You will probably find yourself using a Type node frequently—it allows you to assign a measurement level, values, and a modeling role for each field in the dataset. Its operations are useful for handling missing values and downstream modeling.

The Field Operations palette contains the following nodes:



The Automated Data Preparation (ADP) node can analyze your data and identify fixes, screen out fields that are problematic or not likely to be useful, derive new attributes when appropriate, and improve performance through intelligent screening and sampling techniques. You can use the node in fully automated fashion, allowing the node to choose and apply fixes, or you can preview the changes before they are made and accept, reject, or amend them as desired. For more information, see the topic [Automated Data Preparation](#) on p. 87.



The Type node specifies field metadata and properties. For example, you can specify a measurement level (continuous, nominal, ordinal, or flag) for each field, set options for handling missing values and system nulls, set the role of a field for modeling purposes, specify field and value labels, and specify values for a field. For more information, see the topic [Type Node](#) on p. 113.



The Filter node filters (discards) fields, renames fields, and maps fields from one source node to another. For more information, see the topic [Filtering or Renaming Fields](#) on p. 130.



The Derive node modifies data values or creates new fields from one or more existing fields. It creates fields of type formula, flag, nominal, state, count, and conditional. For more information, see the topic [Derive Node](#) on p. 139.



The Ensemble node combines two or more model nuggets to obtain more accurate predictions than can be gained from any one model. For more information, see the topic [Ensemble Node](#) on p. 136.



The Filler node replaces field values and changes storage. You can choose to replace values based on a CLEM condition, such as `@BLANK(@FIELD)`. Alternatively, you can choose to replace all blanks or null values with a specific value. A Filler node is often used together with a Type node to replace missing values. For more information, see the topic [Filler Node](#) on p. 151.



The Anonymize node transforms the way field names and values are represented downstream, thus disguising the original data. This can be useful if you want to allow other users to build models using sensitive data, such as customer names or other details. For more information, see the topic [Anonymize Node](#) on p. 154.



The Reclassify node transforms one set of categorical values to another. Reclassification is useful for collapsing categories or regrouping data for analysis. For more information, see the topic [Reclassify Node](#) on p. 158.



The Binning node automatically creates new nominal (set) fields based on the values of one or more existing continuous (numeric range) fields. For example, you can transform a continuous income field into a new categorical field containing groups of income as deviations from the mean. Once you have created bins for the new field, you can generate a Derive node based on the cut points. For more information, see the topic [Binning Node](#) on p. 162.



The Recency, Frequency, Monetary (RFM) Analysis node enables you to determine quantitatively which customers are likely to be the best ones by examining how recently they last purchased from you (recency), how often they purchased (frequency), and how much they spent over all transactions (monetary). For more information, see the topic [RFM Analysis Node](#) on p. 173.



The Partition node generates a partition field, which splits the data into separate subsets for the training, testing, and validation stages of model building. For more information, see the topic [Partition Node](#) on p. 176.



The Set to Flag node derives multiple flag fields based on the categorical values defined for one or more nominal fields. For more information, see the topic [Set to Flag Node](#) on p. 178.



The Restructure node converts a nominal or flag field into a group of fields that can be populated with the values of yet another field. For example, given a field named *payment type*, with values of *credit*, *cash*, and *debit*, three new fields would be created (*credit*, *cash*, *debit*), each of which might contain the value of the actual payment made. For more information, see the topic [Restructure Node](#) on p. 180.



The Transpose node swaps the data in rows and columns so that records become fields and fields become records. For more information, see the topic [Transpose Node](#) on p. 182.



The Time Intervals node specifies intervals and creates labels (if needed) for modeling time series data. If values are not evenly spaced, the node can pad or aggregate values as needed to generate a uniform interval between records. For more information, see the topic [Time Intervals Node](#) on p. 186.



The History node creates new fields containing data from fields in previous records. History nodes are most often used for sequential data, such as time series data. Before using a History node, you may want to sort the data using a Sort node. For more information, see the topic [History Node](#) on p. 205.



The Field Reorder node defines the natural order used to display fields downstream. This order affects the display of fields in a variety of places, such as tables, lists, and the Field Chooser. This operation is useful when working with wide datasets to make fields of interest more visible. For more information, see the topic [Field Reorder Node](#) on p. 206.

Several of these nodes can be generated directly from the audit report created by a Data Audit node. For more information, see the topic [Generating Other Nodes for Data Preparation](#) in Chapter 6 on p. 349.

Automated Data Preparation

Preparing data for analysis is one of the most important steps in any project—and traditionally, one of the most time consuming. Automated Data Preparation (ADP) handles the task for you, analyzing your data and identifying fixes, screening out fields that are problematic or not likely to be useful, deriving new attributes when appropriate, and improving performance through intelligent screening techniques. You can use the algorithm in fully **automatic** fashion, allowing it to choose and apply fixes, or you can use it in **interactive** fashion, previewing the changes before they are made and accept or reject them as desired.

Using ADP enables you to make your data ready for model building quickly and easily, without needing prior knowledge of the statistical concepts involved. Models will tend to build and score more quickly; in addition, using ADP improves the robustness of automated modeling processes, such as model refresh and champion / challenger.

Note: when ADP prepares a field for analysis, it creates a new field containing the adjustments or transformations, rather than replacing the existing values and properties of the old field. The old field is not used in further analysis; its role is set to None.

Example. An insurance company with limited resources to investigate homeowner’s insurance claims wants to build a model for flagging suspicious, potentially fraudulent claims. Before building the model, they will ready the data for modeling using automated data preparation. Since they want to be able to review the proposed transformations before the transformations are applied, they will use automated data preparation in interactive mode.

An automotive industry group keeps track of the sales for a variety of personal motor vehicles. In an effort to be able to identify over- and underperforming models, they want to establish a relationship between vehicle sales and vehicle characteristics. They will use automated data preparation to prepare the data for analysis, and build models using the data “before” and “after” preparation to see how the results differ.

Figure 4-1
Automated Data Preparation Objective tab

Recommends data preparation steps that will speed up model building and improve predictive power. This can include transforming, constructing and selecting features. The target can also be transformed.

What is your objective?

Each objective corresponds to a distinct default configuration on the Settings tab that you can further customize, if desired.

Balance speed & accuracy
 Optimize for speed
 Optimize for accuracy
 Customize analysis

Description

Balanced speed and accuracy adjusts the default setting to transform the data with an emphasis on building models with a balance of speed and accuracy.

What is your objective? Automated data preparation recommends data preparation steps that will affect the speed with which other algorithms can build models and improve the predictive power of those models. This can include transforming, constructing and selecting features. The target can also be transformed. You can specify the model-building priorities that the data preparation process should concentrate on.

- **Balance speed and accuracy.** This option prepares the data to give equal priority to both the speed with which data are processed by model-building algorithms and the accuracy of the predictions.
- **Optimize for speed.** This option prepares the data to give priority to the speed with which data are processed by model-building algorithms. When you are working with very large datasets, or are looking for a quick answer, select this option.
- **Optimize for accuracy.** This option prepares the data to give priority to the accuracy of predictions produced by model-building algorithms.
- **Custom analysis.** When you want to manually change the algorithm on the Settings tab, select this option. Note that this setting is automatically selected if you subsequently make changes to options on the Settings tab that are incompatible with one of the other objectives.

Training The Node

The ADP node is implemented as a process node and works in a similar way to the Type node; **training** the ADP node corresponds to instantiating the Type node. Once analysis has been performed, the specified transformations are applied to the data without further analysis as long as

the upstream data model does not change. Like the Type and Filter nodes, if the ADP node is disconnected it remembers the data model and transformations so that if it is reconnected it does not need to be retrained; this enables you to train it on a subset of typical data and then copy or deploy it for use it on live data as often as required.

Using the Toolbar

The toolbar enables you to run and update the display of the data analysis, and generate nodes that you can use in conjunction with the original data.

Figure 4-2
Automated Data Preparation - Toolbar



- **Generate** From this menu you can generate either a Filter or Derive node. Note that this menu is only available when there is an analysis shown on the Analysis tab.

The Filter node removes transformed input fields. If you configure the ADP node to leave the original input fields in the dataset, this will restore the original set of inputs allowing you to interpret the score field in terms of the inputs. For example, this may be useful if you want to produce a graph of the score field against various inputs.

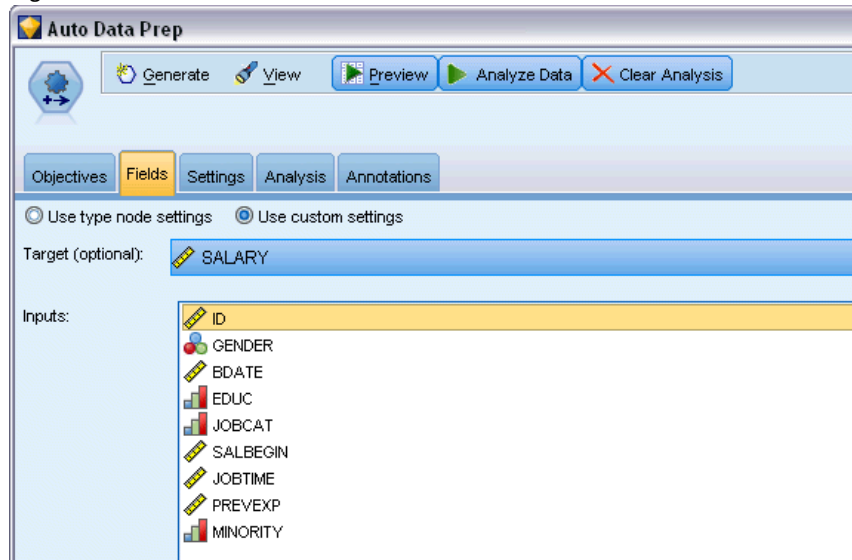
The Derive node can restore the original dataset and target units. You can only generate a Derive node when the ADP node contains an analysis which rescales a range target (that is, Box-Cox rescaling is selected on the Prepare Inputs & Target panel). You cannot generate a Derive node if the target is not a range, or if the Box-Cox rescaling is not selected. For more information, see the topic [Generating a Derive Node](#) on p. 112.
- **View** Contains options that control what is shown on the Analysis tab. This includes the graph editing controls and the display selections for both the main panel and linked views.
- **Preview** Displays a sample of the transformations that will be applied to the input data.
- **Analyze Data** Initiates an analysis using the current settings and displays the results on the Analysis tab.
- **Clear Analysis** Deletes the existing analysis (only available when a current analysis exists).

Node Status

The status of the ADP node on the IBM® SPSS® Modeler canvas is indicated by either an arrow or tick on the icon that shows whether or not analysis has taken place.

Fields Tab

Figure 4-3



Before you can build a model, you need to specify which fields you want to use as targets and as inputs. With a few exceptions, all modeling nodes will use field information from an upstream Type node. If you are using a Type node to select input and target fields, you don't need to change anything on this tab.

Use type node settings. This option tells the node to use field information from an upstream Type node. This is the default.

Use custom settings. This option tells the node to use field information specified here instead of that given in any upstream Type node(s). After selecting this option, specify the fields below as required.

Target. For models that require one or more target fields, select the target field or fields. This is similar to setting the field role to *Target* in a Type node.

Inputs. Select the input field or fields. This is similar to setting the field role to *Input* in a Type node.

Settings Tab

The Settings tab comprises several different groups of settings that you can modify to fine-tune how the algorithm processes your data. If you make any changes to the default settings that are incompatible with the other objectives, the Objective tab is automatically updated to select the Customize analysis option.

Field Settings

Figure 4-4
Automated Data Preparation - Field Settings

Field settings are not affected if you change your objective.

Use frequency field

Use weight field

How to handle fields that are excluded from modeling:

Filter out unused fields

Set the direction of unused fields to "None"

If the incoming fields do not match the existing analysis:

Stop execution and keep the existing analysis

Clear the existing analysis and analyze the new data

Use frequency field. This option enables you to select a field as a frequency weight. Use this if the records in your training data represent more than one unit each; for example, if you are using aggregated data. The field values should be the number of units represented by each record.

Use weight field. This option enables you to select a field as a case weight. Case weights are used to account for differences in variance across levels of the output field.

How to handle fields that are excluded from modeling. Specify what happens to excluded fields; you can choose either to filter them out of the data or simply set their *Role* to None.

If the incoming fields do not match the existing analysis. Specify what happens if one or more required input fields are missing from the incoming dataset when you execute a trained ADP node.

- **Stop execution and keep the existing analysis.** This stops the execution process, preserves the current analysis information, and displays an error.
- **Clear the existing analysis and analyze the new data.** This clears the existing analysis, analyzes the incoming data, and applies the recommended transformations to that data.

Prepare Dates & Times

Figure 4-5
Automated Data Preparation Prepare Dates & Times Settings

Prepare dates and times for modeling

Compute Duration

Compute elapsed time until reference date

Reference Date

Today's date

Fixed date

Date: 2009-04-21

Units for Date Duration

Automatic

Fixed units

Unit: Years

Compute elapsed time until reference time

Reference Time

Current time

Fixed time

Time: 10:36:38

Units for Time Duration

Automatic

Fixed units

Unit: Hours

Extract Cyclical Time Elements

Extract from dates:

Year Month Day

Extract from times:

Hour Minute Second

Many modeling algorithms are unable to directly handle date and time details; these settings enable you to derive new duration data that can be used as model inputs from dates and times in your existing data. The fields containing dates and times must be predefined with date or time storage types. The original date and time fields will not be recommended as model inputs following automated data preparation.

Prepare dates and times for modeling. Deselecting this option disables all other Prepare Dates & Times controls while maintaining the selections.

Compute elapsed time until reference date. This produces the number of years/months/days since a reference date for each variable containing dates.

- **Reference Date.** Specify the date from which the duration will be calculated with regard to the date information in the input data. Selecting Today's date means that the current system date is always used when ADP is executed. To use a specific date, select Fixed date and enter the required date. The current date is automatically entered in the Fixed date field when the node is first created.
- **Units for Date Duration.** Specify whether ADP should automatically decide on the date duration unit, or select from Fixed units of Years, Months, or Days.

Compute elapsed time until reference time. This produces the number of hours/minutes/seconds since a reference time for each variable containing times.

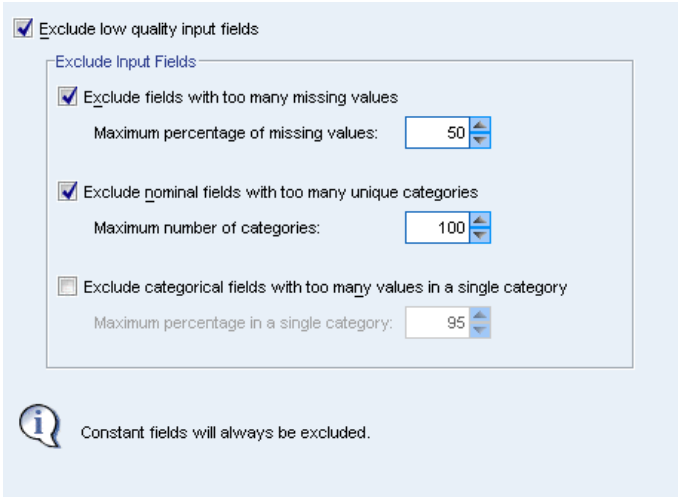
- **Reference Time.** Specify the time from which the duration will be calculated with regard to the time information in the input data. Selecting Current time means that the current system time is always used when ADP is executed. To use a specific time, select Fixed time and enter the required details. The current time is automatically entered in the Fixed time field when the node is first created.
- **Units for Time Duration.** Specify whether ADP should automatically decide on the time duration unit, or select from Fixed units of Hours, Minutes, or Seconds.

Extract Cyclical Time Elements. Use these settings to split a single date or time field into one or more fields. For example if you select all three date checkboxes, the input date field “1954-05-23” is split into three fields: 1954, 5, and 23, each using the suffix defined on the Field Names panel, and the original date field is ignored.

- **Extract from dates.** For any date inputs, specify if you want to extract years, months, days, or any combination.
- **Extract from times.** For any time inputs, specify if you want to extract hours, minutes, seconds, or any combination.

Exclude Fields

Figure 4-6
Automated Data Preparation Exclude Fields Settings




Exclude low quality input fields

Exclude Input Fields

Exclude fields with too many missing values
Maximum percentage of missing values: 50

Exclude nominal fields with too many unique categories
Maximum number of categories: 100

Exclude categorical fields with too many values in a single category
Maximum percentage in a single category: 95

 Constant fields will always be excluded.

Poor quality data can affect the accuracy of your predictions; therefore, you can specify the acceptable quality level for input features. All fields that are constant or have 100% missing values are automatically excluded.

Exclude low quality input fields. Deselecting this option disables all other Exclude Fields controls while maintaining the selections.

Exclude fields with too many missing values. Fields with more than the specified percentage of missing values are removed from further analysis. Specify a value greater than or equal to 0, which is equivalent to deselecting this option, and less than or equal to 100, though fields with all missing values are automatically excluded. The default is 50.

Exclude nominal fields with too many unique categories. Nominal fields with more than the specified number of categories are removed from further analysis. Specify a positive integer. The default is 100. This is useful for automatically removing fields containing record-unique information from modeling, like ID, address, or name.

Exclude categorical fields with too many values in a single category. Ordinal and nominal fields with a category that contains more than the specified percentage of the records are removed from further analysis. Specify a value greater than or equal to 0, equivalent to deselecting this option, and less than or equal to 100, though constant fields are automatically excluded. The default is 95.

Preparing Inputs and Targets

Because no data is ever in a perfect state for processing, you may want to adjust some of the settings before running an analysis. For example, this might include the removal of outliers, specifying how to handle missing values, or adjusting the type.

Note: If you change the values on this panel, the Objectives tab is automatically updated to select the Custom analysis option.

Figure 4-7
Automated Data Preparation - Input and Target Settings

Prepare the input and target fields for modeling

Adjust Type and Improve Data Quality

Inputs	Target	
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Adjust the type of numeric fields (ordinal and continuous)
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Reorder nominal fields to have smallest category first, largest last
<input type="checkbox"/>	<input type="checkbox"/>	Replace outlier values in continuous fields (recommended for input fields if they will be put on a common scale)
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Continuous fields: replace missing values with mean
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Nominal fields: replace missing values with mode
<input type="checkbox"/>	<input type="checkbox"/>	Ordinal fields: replace missing values with median

Maximum number of values for ordinal fields:

Minimum number of values for continuous fields:

Outlier cutoff value: (standard deviations)

Method for replacing outliers: Replace with cutoff value Delete value

Transform Continuous Field

Put all continuous input fields on a common scale (highly recommended if feature construction will be performed)

Rescaling method: Final mean: Final standard deviation:

Rescale a continuous target with a Box-Cox transformation to reduce skew

Final mean: Final standard deviation:

Prepare the input and target fields for modeling. Toggles all fields on the panel either on or off.

Adjust Type and Improve Data Quality. For inputs and the target you can specify several data transformations separately; this is because you may not want to change the values of the target. For example, a prediction of income in dollars is more meaningful than a prediction measured in log(dollars). In addition, if the target has missing values there is no predictive gain to filling

missing values, whereas filling missing values in inputs may enable some algorithms to process information that would otherwise be lost.

Additional settings for these transformations, such as the outlier cutoff value, are common to both the target and inputs.

You can select the following settings for either, or both, inputs and target:

- **Adjust the type of numeric fields.** Select this to determine if numeric fields with a measurement level of *Ordinal* can be converted to *Continuous*, or vice versa. You can specify the minimum and maximum threshold values to control the conversion.
- **Reorder nominal fields.** Select this to sort nominal (set) fields into order, from smallest to largest category.
- **Replace outlier values in continuous fields.** Specify whether to replace outliers; use this in conjunction with the Method for replacing outliers options below.
- **Continuous fields: replace missing values with mean.** Select this to replace missing values of continuous (range) features.
- **Nominal fields: replace missing values with mode.** Select this to replace missing values of nominal (set) features.
- **Ordinal fields: replace missing values with median.** Select this to replace missing values of ordinal (ordered set) features.

Maximum number of values for ordinal fields. Specify the threshold for redefining ordinal (ordered set) fields as continuous (range). The default is 10; therefore, if an ordinal field has more than 10 categories it is redefined as continuous (range).

Minimum number of values for continuous fields. Specify the threshold for redefining scale or continuous (range) fields as ordinal (ordered set). The default is 5; therefore, if a continuous field has fewer than 5 values it is redefined as ordinal (ordered set).

Outlier cutoff value. Specify the outlier cutoff criterion, measured in standard deviations; the default is 3.

Method for replacing outliers. Select whether outliers are to be replaced by either trimming (coerce) with the cutoff value, or to delete them and set them as missing values. Any outliers set to missing values follow the missing value handling settings selected above.

Put all continuous input fields on a common scale. To normalize continuous input fields, select this check box and choose the normalization method. The default is z-score transformation, where you can specify the Final mean, which has a default of 0, and the Final standard deviation, which has a default of 1. Alternatively, you can choose to use Min/max transformation and specify the minimum and maximum values, which have default values of 0 and 100 respectively.

This field is especially useful when you select Perform feature construction on the Construct & Select Features panel.

Rescale a continuous target with a Box-Cox transformation. To normalize a continuous (scale or range) target field, select this check box. The Box-Cox transformation has default values of 0 for the Final mean and 1 for the Final standard deviation.

Note: If you choose to normalize the target, the dimension of the target will be transformed. In this case you may need to generate a Derive node to apply an inverse transformation in order to turn the transformed units back into a recognizable format for further processing. For more information, see the topic [Generating a Derive Node](#) on p. 112.

Construction and Feature Selection

To improve the predictive power of your data, you can transform the input fields, or construct new ones based on the existing fields.

Note: If you change the values on this panel, the Objectives tab is automatically updated to select the Custom analysis option.

Figure 4-8

Automated Data Preparation - Transform, Construction, and Selection Settings

Transform, construct and select input fields to improve predictive power

Categorical Input Fields

Merge sparse categories to maximize association with target p-value: 0.05

Input fields that have only one category after supervised merging will be excluded.

When there is no target, merge sparse categories based on counts

Ordinal features Nominal features Minimum % of cases in any category: 10

Continuous Input Fields

Bin continuous fields while preserving predictive power (available only with a categorical target)

p-value: 0.05

Input fields that have only one category after binning will be excluded.

Feature Selection and Construction

Perform feature selection

p-value: 0.05

Feature selection applies to continuous input fields when the target is continuous, and to categorical inputs.

Perform feature construction

Feature construction applies to continuous input fields when the target is continuous or there is no target.

Transform, construct and select input fields to improve predictive power. Toggles all fields on the panel either on or off.

Merge sparse categories to maximize association with target. Select this to make a more parsimonious model by reducing the number of variables to be processed in association with the target. If required, change the probability value from the default of 0.05.

Note that if all categories are merged into one, the original and derived versions of the field are excluded because they have no value as a predictor.

When there is no target, merge sparse categories based on counts. If you are dealing with data that has no target, you can choose to merge sparse categories of either, or both, ordinal (ordered set) and nominal (set) features. Specify the minimum percentage of cases, or records, in the data that identifies the categories to be merged; the default is 10.

Categories are merged using the following rules:

- Merging is not performed on binary fields.
- If there are only two categories during merging, merging stops.
- If there is no original category, nor any category created during merging, with fewer than the specified minimum percent of cases, merging stops.

Bin continuous fields while preserving predictive power. Where you have data that includes a categorical target, you can bin continuous inputs with strong associations to improve processing performance. If required, change the probability value for the homogenous subsets from the default of 0.05.

If the binning operation results in a single bin for a particular field, the original and binned versions of the field are excluded because they have no value as a predictor.

Note: Binning in ADP differs from optimal binning used in other parts of IBM® SPSS® Modeler. Optimal binning uses entropy information to convert a continuous variable to a categorical variable; this needs to sort data and store it all in memory. ADP uses homogenous subsets to bin a continuous variable, this means that ADP binning does not need to sort data and does not store all data in memory. The use of the homogenous subset method to bin a continuous variable means that the number of categories after binning is always less than or equal to the number of categories of target.

Perform feature selection. Select this option to remove features with a low correlation coefficient. If required, change the probability value from the default of 0.05.

This option only applies to continuous input features where the target is continuous, and to categorical input features.

Perform feature construction. Select this option to derive new features from a combination of several existing features (which are then discarded from modeling).

This option only applies to continuous input features where the target is continuous, or where there is no target.

Field Names

Figure 4-9
Automated Data Preparation Name Fields Settings

Field Names are not affected if you change your objective.

Specify how to construct the names of transformed and constructed fields.

Transformed and Constructed Fields

Name extension for transformed target field:

Name extension for transformed input fields:

Root name for constructed features:

Durations Computed from Dates and Times

Name extensions for durations computed from dates

Years: Months: Days:

Name extensions for durations computed from times

Hours: Minutes: Seconds:

Cyclical Elements Extracted From Dates and Times

Name extensions for cyclical elements extracted from dates

Year: Month: Day:

Name extensions for cyclical elements extracted from times

Hour: Minute: Second:

To easily identify new and transformed features, ADP creates and applies basic new names, prefixes, or suffixes. You can amend these names to be more relevant to your own needs and data. If you wish to specify other labels, you will need to do this in a downstream Type node.

Transformed and Constructed Fields. Specify the name extensions to be applied to transformed target and input fields.

Note that in the ADP node, setting string fields to contain nothing may cause an error depending on how you chose to handle unused fields. If How to handle fields that are excluded from modeling is set to Filter out unused fields on the Field Settings panel of the Settings tab, the name extensions for inputs and the target can be set to nothing. The original fields are filtered out and the transformed fields saved over them; in this case the new transformed fields will have the same name as your original.

However if you chose Set the direction of unused fields to 'None', then empty, or null, name extensions for the target and inputs will cause an error because you will be trying to create duplicate field names.

In addition, specify the prefix name to be applied to any features that are constructed via the Select and Construct settings. The new name is created by attaching a numeric suffix to this prefix root name. The format of the number depends on how many new features are derived, for example:

- 1-9 constructed features will be named: feature1 to feature9.

- 10-99 constructed features will be named: feature01 to feature99.
- 100-999 constructed features will be named: feature001 to feature999, and so on.

This ensures that the constructed features will sort in a sensible order no matter how many there are.

Durations Computed from Dates and Times. Specify the name extensions to be applied to durations computed from both dates and times.

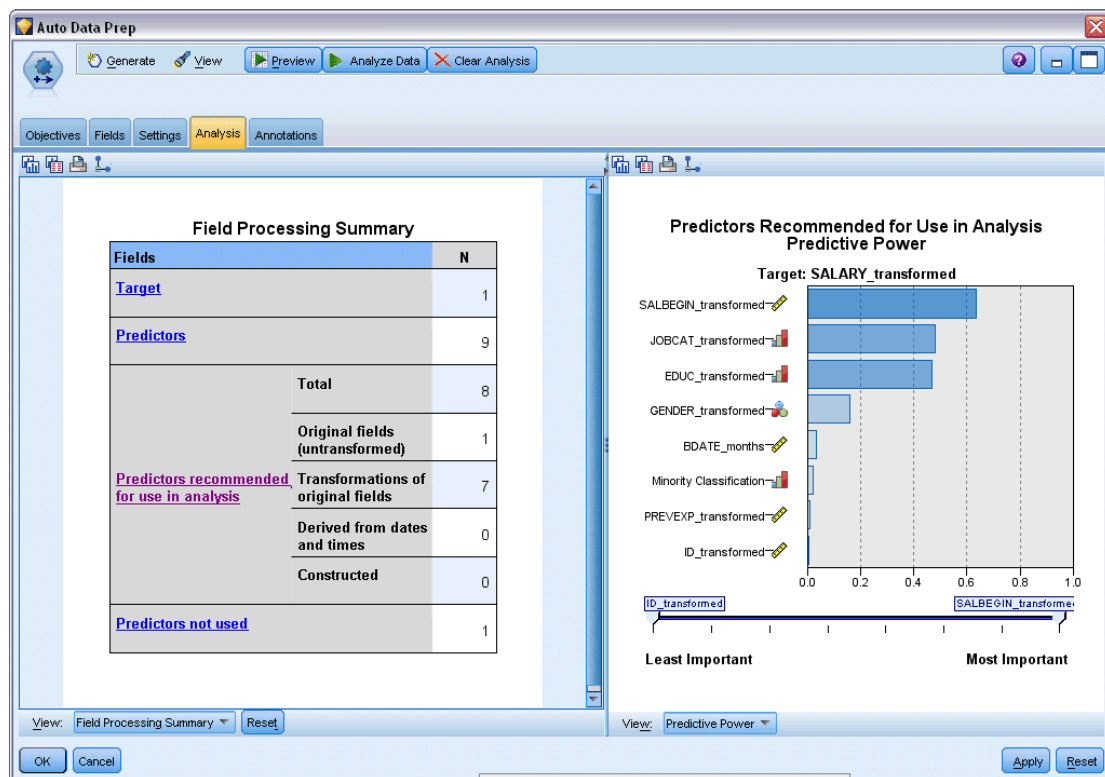
Cyclical Elements Extracted from Dates and Times. Specify the name extensions to be applied to cyclical elements extracted from both dates and times.

Analysis Tab

- ▶ When you are satisfied with the ADP settings, including any changes made on the Objective, Fields, and Settings tabs, click Analyze Data; the algorithm applies the settings to the data inputs and displays the results on the Analysis tab.

The Analysis tab contains both tabular and graphical output that summarizes the processing of your data and displays recommendations as to how the data may be modified or improved for scoring. You can then review and either accept or reject those recommendations.

Figure 4-10
Automated Data Preparation Analysis Tab



The Analysis tab is made up of two panels, the main view on the left and the linked, or auxiliary, view on the right. There are three main views:

- Field Processing Summary (the default). For more information, see the topic [Field Processing Summary](#) on p. 101.
- Fields. For more information, see the topic [Fields](#) on p. 102.
- Action Summary. For more information, see the topic [Action Summary](#) on p. 104.

There are four linked/auxiliary views:

- Predictive Power (the default). For more information, see the topic [Predictive Power](#) on p. 105.
- Fields Table. For more information, see the topic [Fields Table](#) on p. 106.
- Field Details. For more information, see the topic [Field Details](#) on p. 107.
- Action Details. For more information, see the topic [Action Details](#) on p. 109.

Links between views

Within the main view, underlined text in the tables controls the display in the linked view. Clicking on the text allows you to get details on a particular field, set of fields, or processing step. The link that you last selected is shown in a darker color; this helps you identify the connection between the contents of the two view panels.

Resetting the views

To redisplay the original Analysis recommendations and abandon any changes you have made to the Analysis views, click Reset at the bottom of the main view panel.

Field Processing Summary

Figure 4-11
Field Processing Summary

Fields	N
Target	1
Predictors	9
	Total 8
	Original fields (untransformed) 1
Predictors recommended for use in analysis	Transformations of original fields 7
	Derived from dates and times 0
	Constructed 0
Predictors not used	1

The Field Processing Summary table gives a snapshot of the projected overall impact of processing, including changes to the state of the features and the number of features constructed.

Note that no model is actually built, so there isn't a measure or graph of the change in overall predictive power before and after data preparation; instead, you can display graphs of the predictive power of individual recommended predictors.

The table displays the following information:


- The number of target fields.
- The number of original (input) predictors.
- The predictors recommended for use in analysis and modeling. This includes the total number of fields recommended; the number of original, untransformed, fields recommended; the number of transformed fields recommended (excluding intermediate versions of any field, fields derived from date/time predictors, and constructed predictors); the number of fields recommended that are derived from date/time fields; and the number of constructed predictors recommended.
- The number of input predictors not recommended for use in any form, whether in their original form, as a derived field, or as input to a constructed predictor.

Where any of the Fields information is underlined, click to display more details in a linked view. Details of the Target, Input features, and Input features not used are shown in the Fields Table linked view. For more information, see the topic [Fields Table](#) on p. 106. Features recommended for use in analysis are displayed in the Predictive Power linked view. For more information, see the topic [Predictive Power](#) on p. 105.








Fields

Figure 4-12
Fields

Fields

Target	
Name	Measurement Level
SALARY	

Predictors Include nonrecommended fields in table

Version to Use	Name	Measurement Level	Predictive Power
Transformed	SALBEGIN		0.64
Transformed	JOB CAT		0.48
Transformed	EDUC		0.47
Transformed	GENDER		0.16
Transformed	BDATE_Duration Months		0.03
Original	MINORITY		0.02
Transformed	PREVEXP		0.01

The Fields main view displays the processed fields and whether ADP recommends using them in downstream models. You can override the recommendation for any field; for example, to exclude constructed features or include features that ADP recommends excluding. If a field has been transformed, you can decide whether to accept the suggested transformation or use the original version.

The Fields view consists of two tables, one for the target and one for predictors that were either processed or created.

Target table

The Target table is only shown if a target is defined in the data.

The table contains two columns:

- **Name.** This is the name or label of the target field; the original name is always used, even if the field has been transformed.
- **Measurement Level.** This displays the icon representing the measurement level; hover the mouse over the icon to display a label (continuous, ordinal, nominal, and so on) that describes the data.

If the target has been transformed the Measurement Level column reflects the final transformed version. *Note:* you cannot turn off transformations for the target.

Predictors table

The Predictors table is always shown. Each row of the table represents a field. By default the rows are sorted in descending order of predictive power.

For ordinary features, the original name is always used as the row name. Both original and derived versions of date/time fields appear in the table (in separate rows); the table also includes constructed predictors.

Note that transformed versions of fields shown in the table always represent the final versions.

By default only recommended fields are shown in the Predictors table. To display the remaining fields, select the Include nonrecommended fields in table box above the table; these fields are then displayed at the bottom of the table.

The table contains the following columns:

- **Version to Use.** This displays a drop-down list that controls whether a field will be used downstream and whether to use the suggested transformations. By default, the drop-down list reflects the recommendations.

For ordinary predictors that have been transformed the drop-down list has three choices: Transformed, Original, and Do not use.

For untransformed ordinary predictors the choices are: Original and Do not use.

For derived date/time fields and constructed predictors the choices are: Transformed and Do not use.

For original date fields the drop-down list is disabled and set to Do not use.

Note: For predictors with both original and transformed versions, changing between Original and Transformed versions automatically updates the Measurement Level and Predictive Power settings for those features.

- **Name.** Each field's name is a link. Click on a name to display more information about the field in the linked view. For more information, see the topic [Field Details](#) on p. 107.
- **Measurement Level.** This displays the icon representing the data type; hover the mouse over the icon to display a label (continuous, ordinal, nominal, and so on) that describes the data.
- **Predictive Power.** Predictive power is displayed only for fields that ADP recommends. This column is not displayed if there is no target defined. Predictive power ranges from 0 to 1, with larger values indicating “better” predictors. In general, predictive power is useful for comparing predictors within an ADP analysis, but predictive power values should not be compared across analyses.

Action Summary

Figure 4-13
Action Summary

Action
Text Fields
Date and Time Predictors
Predictor Screening
Check Measurement Level
Outliers
Missing Values
Target
Categorical Predictors
Continuous Predictors

For each action taken by automated data preparation, input predictors are transformed and/or filtered out; fields that survive one action are used in the next. The fields that survive through to the last step are then recommended for use in modeling, whilst inputs to transformed and constructed predictors are filtered out.










The Action Summary is a simple table that lists the processing actions taken by ADP. Where any Action is underlined, click to display more details in a linked view about the actions taken. For more information, see the topic [Action Details](#) on p. 109.

Note: Only the original and final transformed versions of each field are shown, not any intermediate versions that were used during analysis.

Fields Table

Figure 4-15
Fields Table

Predictors

Name	Measurement Level
ID	 Continuous
GENDER	 Nominal
BDATE	 Continuous
EDUC	 Ordinal
JOB CAT	 Ordinal
SALBEGIN	 Continuous
JOB TIME	 Continuous
PREV EXP	 Continuous
MINORITY	 Ordinal

Displayed when you click Target, Predictors, or Predictors not used in the Field Processing Summary main view, the Fields Table view displays a simple table listing the relevant features.

The table contains two columns:

- **Name.** The predictor name.

For targets, the original name or label of the field is used, even if the target has been transformed.

For transformed versions of ordinary predictors, the name reflects your choice of suffix in the Field Names panel of the Settings tab; for example: *_transformed*.

For fields derived from dates and times, the name of the final transformed version is used; for example: *bdate_years*.

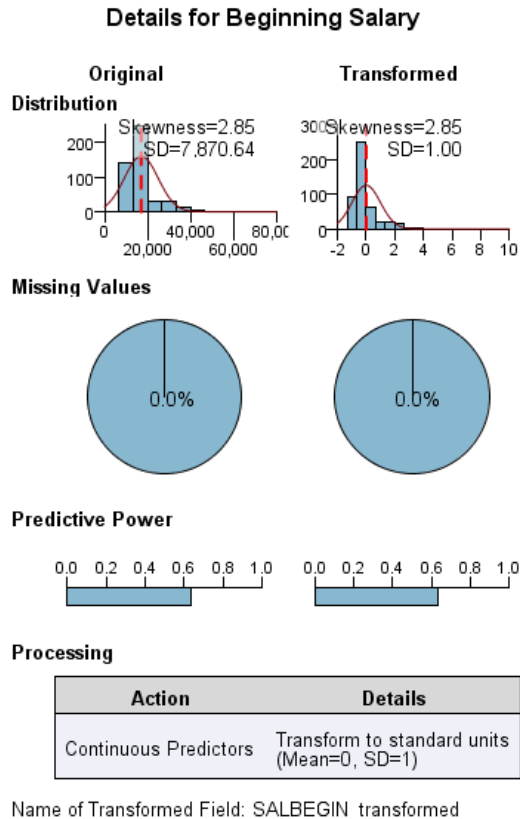
For constructed predictors, the name of the constructed predictor is used; for example: *Predictor1*.

- **Measurement Level.** This displays the icon representing the data type.

For the Target, the Measurement Level always reflects the transformed version (if the target has been transformed); for example, changed from ordinal (ordered set) to continuous (range, scale), or vice versa.

Field Details

Figure 4-16
Field Details



Displayed when you click any Name in the Fields main view, the Field Details view contains distribution, missing values, and predictive power charts (if applicable) for the selected field. In addition, the processing history for the field and the name of the transformed field are also shown (if applicable)

For each chart set, two versions are shown side by side to compare the field with and without transformations applied; if a transformed version of the field does not exist, a chart is shown for the original version only. For derived date or time fields and constructed predictors, the charts are only shown for the new predictor.

Note: If a field is excluded due to having too many categories only the processing history is shown.

Distribution Chart

Continuous field distribution is shown as a histogram, with a normal curve overlaid, and a vertical reference line for the mean value; categorical fields are displayed as a bar chart.

Histograms are labeled to show standard deviation and skewness; however, skewness is not displayed if the number of values is 2 or fewer or the variance of the original field is less than 10-20.

Hover the mouse over the chart to display either the mean for histograms, or the count and percentage of the total number of records for categories in bar charts.

Missing Value Chart

Pie charts compare the percentage of missing values with and without transformations applied; the chart labels show the percentage.

If ADP carried out missing value handling, the post-transformation pie chart also includes the replacement value as a label — that is, the value used in place of missing values.

Hover the mouse over the chart to display the missing value count and percentage of the total number of records.

Predictive Power Chart

For recommended fields, bar charts display the predictive power before and after transformation. If the target has been transformed, the calculated predictive power is in respect to the transformed target.

Note: Predictive power charts are not shown if no target is defined, or if the target is clicked in the main view panel.

Hover the mouse over the chart to display the predictive power value.

Processing History Table

The table shows how the transformed version of a field was derived. Actions taken by ADP are listed in the order in which they were carried out; however, for certain steps multiple actions may have been carried out for a particular field.

Note: This table is not shown for fields that have not been transformed.

The information in the table is broken down into two or three columns:

- **Action.** The name of the action. For example, Continuous Predictors. For more information, see the topic [Action Details](#) on p. 109.
- **Details.** The list of processing carried out. For example, Transform to standard units.
- **Function.** Only shown only for constructed predictors, this displays the linear combination of input fields, for example, $.06 * \text{age} + 1.21 * \text{height}$.

Action Details

Figure 4-17
ADP Analysis - Action Details

Step 9: Continuous Predictors

Transformation	Number of Predictors	Criteria	
		Mean	SD
Transform to standard units	5	0	1

Predictor Space Construction	N
Predictors constructed	0
Predictors excluded due to low association with target	1
Predictors excluded because they were constant after binning	0

Displayed when you select any underlined Action in the Action Summary main view, the Action Details linked view displays both action-specific and common information for each processing step that was carried out; the action-specific details are displayed first.

For each action, the description is used as the title at the top of the linked view. The action-specific details are displayed below the title, and may include details of the number of derived predictors, fields recast, target transformations, categories merged or reordered, and predictors constructed or excluded.

As each action is processed, the number of predictors used in the processing may change, for example as predictors are excluded or merged.

Note: If an action was turned off, or no target was specified, an error message is displayed in place of the action details when the action is clicked in the Action Summary main view.

There are nine possible actions; however, not all are necessarily active for every analysis.

Text Fields Table

The table displays the number of:

- Trailing blank values trimmed.
- Predictors excluded from analysis.

Date and Time Predictors Table

The table displays the number of:

- Durations derived from date and time predictors.
- Date and time elements.
- Derived date and time predictors, in total.

The reference date or time is displayed as a footnote if any date durations were calculated.

Predictor Screening Table

The table displays the number of the following predictors excluded from processing:

- Constants.
- Predictors with too many missing values.
- Predictors with too many cases in a single category.
- Nominal fields (sets) with too many categories.
- Predictors screened out, in total.

Check Measurement Level Table

The table displays the numbers of fields recast, broken down into the following:

- Ordinal fields (ordered sets) recast as continuous fields.
- Continuous fields recast as ordinal fields.
- Total number recast.

If no input fields (target or predictors) were continuous or ordinal, this is shown as a footnote.

Outliers Table

The table displays counts of how any outliers were handled.

- Either the number of continuous fields for which outliers were found and trimmed, or the number of continuous fields for which outliers were found and set to missing, depending on your settings in the Prepare Inputs & Target panel on the Settings tab.
- The number of continuous fields excluded because they were constant, after outlier handling.

One footnote shows the outlier cutoff value; while another footnote is shown if no input fields (target or predictors) were continuous.

Missing Values Table

The table displays the numbers of fields that had missing values replaced, broken down into:

- Target. This row is not shown if no target is specified.

- Predictors. This is further broken down into the number of nominal (set), ordinal (ordered set), and continuous.
- The total number of missing values replaced.

Target Table

The table displays whether the target was transformed, shown as:

- Box-Cox transformation to normality. This is further broken down into columns that show the specified criteria (mean and standard deviation) and Lambda.
- Target categories reordered to improve stability.

Categorical Predictors Table

The table displays the number of categorical predictors:

- Whose categories were reordered from lowest to highest to improve stability.
- Whose categories were merged to maximize association with the target.
- Whose categories were merged to handle sparse categories.
- Excluded due to low association with the target.
- Excluded because they were constant after merging.

A footnote is shown if there were no categorical predictors.

Continuous Predictors Table

There are two tables. The first displays one of the following number of transformations:

- Predictor values transformed to standard units. In addition, this shows the number of predictors transformed, the specified mean, and the standard deviation.
- Predictor values mapped to a common range. In addition, this shows the number of predictors transformed using a min-max transformation, as well as the specified minimum and maximum values.
- Predictor values binned and the number of predictors binned.

The second table displays the predictor space construction details, shown as the number of predictors:

- Constructed.
- Excluded due to a low association with the target.
- Excluded because they were constant after binning.
- Excluded because they were constant after construction.

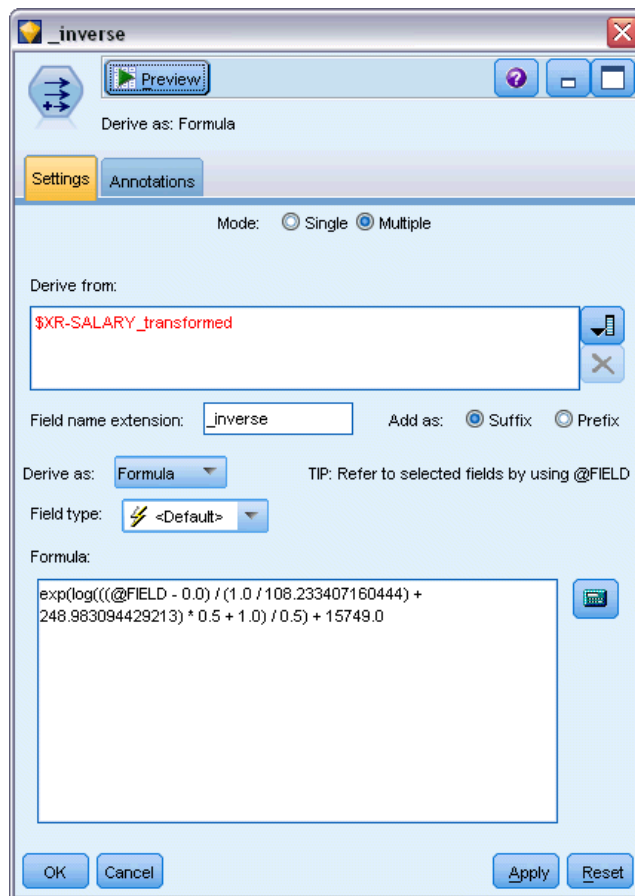
A footnote is shown if no continuous predictors were input.

Generating a Derive Node

When you generate a Derive node, it applies the inverse target transformation to the score field. By default the node enters the name of the score field that would be produced by an automodeler (such as Auto Classifier or Auto Numeric) or the Ensemble node. If a scale (range) target has been transformed the score field is shown in transformed units; for example, log(\$) instead of \$. In order to interpret and use the results, you must convert the predicted value back to the original scale.

Note: You can only generate a Derive node when the ADP node contains an analysis which rescales a range target (that is, Box-Cox rescaling is selected on the Prepare Inputs & Target panel). You cannot generate a Derive node if the target is not a range, or if the Box-Cox rescaling is not selected.

Figure 4-18
Derive node generated from the Automated Data Preparation node



The Derive node is created in Multiple mode and uses @FIELD in the expression so you can add the transformed target if required. For example using the following details:

- Target field name: response
- Transformed target field name: response_transformed
- Score field name: \$XR-response_transformed

The Derive node would create a new field: \$XR-response_transformed_inverse.

Note: If you are not using an automodeler or Ensemble node, you will need to edit the Derive node to transform the correct score field for your model.

Normalized continuous targets

By default, if you select the Rescale a continuous target with a Box-Cox transformation check box on the Prepare Inputs & Target panel this transforms the target and you create a new field that will be the target for your model building. For example if your original target was *response*, the new target will be *response_transformed*; models downstream of the ADP node will pick this new target up automatically.

However, this may cause issues, depending on the original target. For example, if the target was *Age*, the values of the new target will not be *Years*, but a transformed version of *Years*. This means you cannot look at the scores and interpret them since they aren't in recognizable units. In this case you can apply an inverse transformation that will turn your transformed units back into whatever they were meant to be. To do this:

- ▶ After clicking Analyze Data to run the ADP analysis, select *Derive Node* from the *Generate* menu.
- ▶ Place the Derive node after your nugget on the model canvas.

The Derive node will restore the score field to the original dimensions so that the prediction will be in the original *Years* values.

By default the Derive node transforms the score field generated by an auto-modeler or an ensemble model. If you are building an individual model, you need to edit the Derive node to derive from your actual score field. If you want to evaluate your model, you should add the transformed target to the Derive From field in the Derive node. This applies the same inverse transformation to the target and any downstream Evaluation or Analysis node will use the transformed data correctly as long as you switch those nodes to use field names instead of metadata.

If you also want to restore the original name, you can use a Filter node to remove the original target field if it's still there, and rename the target and score fields.

Type Node

Field properties can be specified in a source node or in a separate Type node. The functionality is similar in both nodes. The following properties are available:

- **Field.** Double-click any field name to specify value and field labels for data in IBM® SPSS® Modeler. For example, field metadata imported from IBM® SPSS® Statistics can be viewed or modified here. Similarly, you can create new labels for fields and their values. The labels

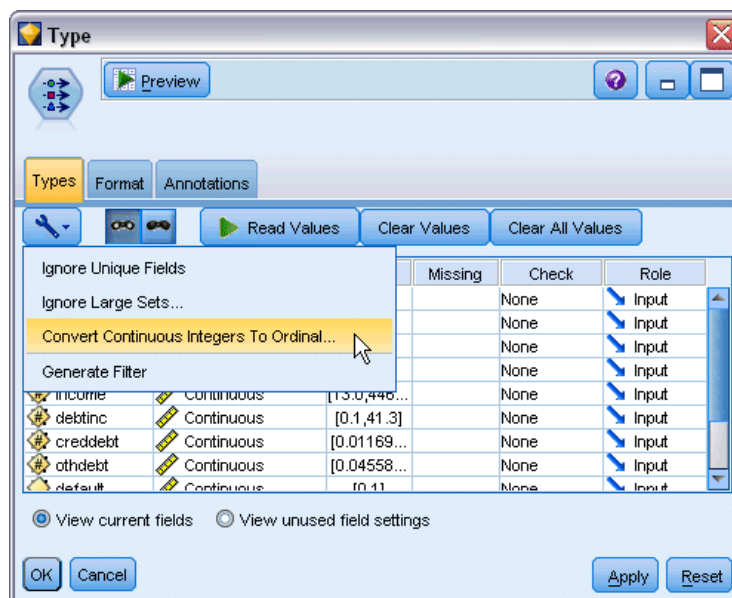
that you specify here are displayed throughout SPSS Modeler depending on the selections you make in the stream properties dialog box.

- **Measurement.** This is the measurement level, used to describe characteristics of the data in a given field. If all of the details of a field are known, it is called **fully instantiated**. For more information, see the topic [Measurement Levels](#) on p. 115.

Note: The measurement level of a field is different from its storage type, which indicates whether the data are stored as strings, integers, real numbers, dates, times, or timestamps.

- **Values.** This column enables you to specify options for reading data values from the dataset, or use the Specify option to specify measurement levels and values in a separate dialog box. You can also choose to pass fields without reading their values. For more information, see the topic [Data Values](#) on p. 119.
- **Missing.** Used to specify how missing values for the field will be handled. For more information, see the topic [Defining Missing Values](#) on p. 124.
- **Check.** In this column, you can set options to ensure that field values conform to the specified values or ranges. For more information, see the topic [Checking Type Values](#) on p. 124.
- **Role.** Used to tell modeling nodes whether fields will be Input (predictor fields) or Target (predicted fields) for a machine-learning process. Both and None are also available roles, along with Partition, which indicates a field used to partition records into separate samples for training, testing, and validation. The value Split specifies that separate models will be built for each possible value of the field. For more information, see the topic [Setting the Field Role](#) on p. 126.

Figure 4-19
Type node options



Several other options can be specified using the Type node window:

- Using the tools menu button, you can choose to Ignore Unique Fields once a Type node has been instantiated (either through your specifications, reading values, or running the stream). Ignoring unique fields will automatically ignore fields with only one value.

- Using the tools menu button, you can choose to Ignore Large Sets once a Type node has been instantiated. Ignoring large sets will automatically ignore sets with a large number of members.
- Using the tools menu button, you can choose to Convert Continuous Integers To Ordinal once a Type node has been instantiated. For more information, see the topic [Converting Continuous Data](#) on p. 117.
- Using the tools menu button, you can generate a Filter node to discard selected fields.
- Using the sunglasses toggle buttons, you can set the default for all fields to Read or Pass. The Types tab in the source node passes fields by default, while the Type node itself reads values by default.
- Using the Clear Values button, you can clear changes to field values made in this node (non-inherited values) and reread values from upstream operations. This option is useful for resetting changes that you may have made for specific fields upstream.
- Using the Clear All Values button, you can reset values for **all** fields read into the node. This option effectively sets the *Values* column to **Read** for all fields. This option is useful to reset values for all fields and reread values and types from upstream operations.
- Using the context menu, you can choose to Copy attributes from one field to another. For more information, see the topic [Copying Type Attributes](#) on p. 127.
- Using the View unused field settings option, you can view type settings for fields that are no longer present in the data or were once connected to this Type node. This is useful when reusing a Type node for datasets that have changed.



Measurement Levels






Measurement level (formerly known as “data type” or “usage type”) describes the usage of the data fields in IBM® SPSS® Modeler. The measurement level can be specified on the Types tab of a source or Type node. For example, you may want to set the measurement level for an integer field with values of 1 and 0 to *Flag*. This usually indicates that 1 = *True* and 0 = *False*.

Storage versus measurement. Note that the measurement level of a field is different from its storage type, which indicates whether data are stored as a string, integer, real number, date, time, or timestamp. While data types can be modified at any point in a stream using a Type node, storage must be determined at the source when reading data into SPSS Modeler (although it can subsequently be changed using a conversion function). For more information, see the topic [Setting Field Storage and Formatting](#) in Chapter 2 on p. 23.

Some modeling nodes indicate the permitted measurement level types for their input and target fields by means of icons on their Fields tab.

Measurement level icons

Icon	Measurement level
	Default
	Continuous

Icon	Measurement level
	Categorical
	Flag
	Nominal
	Ordinal
	Typeless

The following measurement levels are available:

- **Default.** Data whose storage type and values are unknown (for example, because they have not yet been read) are displayed as <Default>.
- **Continuous.** Used to describe numeric values, such as a range of 0–100 or 0.75–1.25. A continuous value can be an integer, real number, or date/time.
- **Categorical.** Used for string values when an exact number of distinct values is unknown. This is an **uninstantiated** data type, meaning that all possible information about the storage and usage of the data is not yet known. Once data have been read, the measurement level will be *Flag*, *Nominal*, or *Typeless*, depending on the maximum number of members for nominal fields specified in the Stream Properties dialog box.
- **Flag.** Used for data with two distinct values that indicate the presence or absence of a trait, such as true and false, Yes and No or 0 and 1. The values used may vary, but one must always be designated as the “true” value, and the other as the “false” value. Data may be represented as text, integer, real number, date, time, or timestamp.
- **Nominal.** Used to describe data with multiple distinct values, each treated as a member of a set, such as small/medium/large. Nominal data can have any storage—numeric, string, or date/time. Note that setting the measurement level to *Nominal* does not automatically change the values to string storage.
- **Ordinal.** Used to describe data with multiple distinct values that have an inherent order. For example, salary categories or satisfaction rankings can be typed as ordinal data. The order is defined by the natural sort order of the data elements. For example, 1, 3, 5 is the default sort order for a set of integers, while HIGH, LOW, NORMAL (ascending alphabetically) is the order for a set of strings. The ordinal measurement level enables you to define a set of categorical data as ordinal data for the purposes of visualization, model building, and export to other applications (such as IBM® SPSS® Statistics) that recognize ordinal data as a distinct type. You can use an ordinal field anywhere that a nominal field can be used. Additionally, fields of any storage type (real, integer, string, date, time, and so on) can be defined as ordinal.
- **Typeless.** Used for data that does not conform to any of the above types, for fields with a single value, or for nominal data where the set has more members than the defined maximum. It is also useful for cases in which the measurement level would otherwise be a set with many members (such as an account number). When you select Typeless for a field, the role is automatically set to None, with Record ID as the only alternative. The default maximum size for sets is 250 unique values. This number can be adjusted or disabled on the Options tab of the Stream Properties dialog box, which can be accessed from the Tools menu.

You can manually specify measurement levels, or you can allow the software to read the data and determine the measurement level based on the values that it reads.

Alternatively, where you have several continuous data fields that should be treated as categorical data, you can choose an option to convert them. For more information, see the topic [Converting Continuous Data](#) on p. 117.

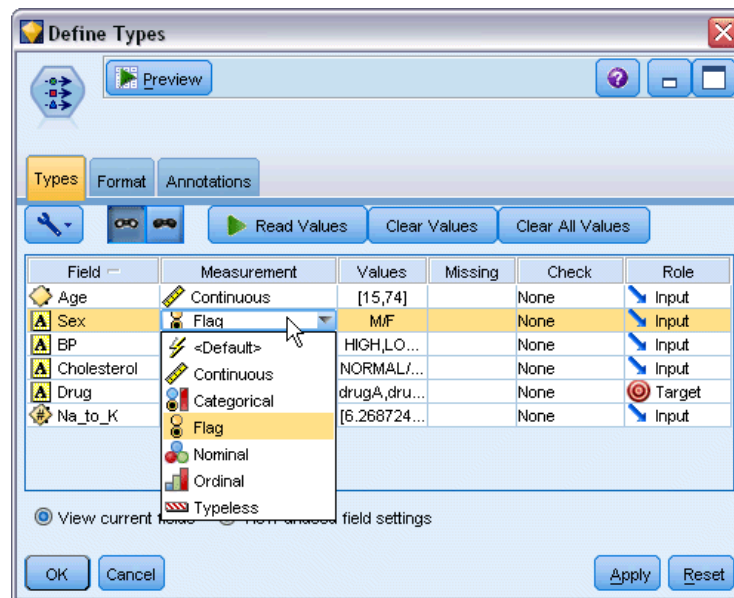
To Use Auto-Typing

- ▶ In either a Type node or the Types tab of a source node, set the *Values* column to <Read> for the desired fields. This will make metadata available to all nodes downstream. You can quickly set all fields to <Read> or <Pass> using the sunglasses buttons on the dialog box.
- ▶ Click Read Values to read values from the data source immediately.

To Manually Set the Measurement Level for a Field

- ▶ Select a field in the table.
- ▶ From the drop-down list in the *Measurement* column, select a measurement level for the field.
- ▶ Alternatively, you can use Ctrl-A or Ctrl-click to select multiple fields before using the drop-down list to select a measurement level .

Figure 4-20
Manually setting measurement levels

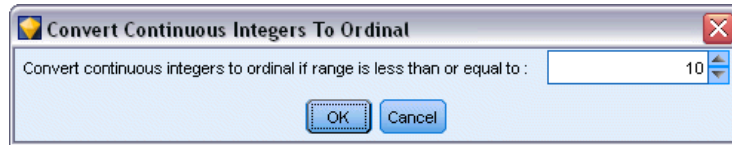


Converting Continuous Data

Treating categorical data as continuous can have a serious effect on the quality of a model, especially if it's the target field; for example, producing a regression model rather than a binary model. To prevent this you can convert integer ranges to categorical types such as *Ordinal* or *Flag*.

- ▶ From the Operations and Generate menu button (with the tool symbol), select Convert Continuous Integers To Ordinal. The conversion values dialog is displayed.

Figure 4-21
Conversion values dialog



- ▶ Specify the size of range that will be automatically converted; this applies to any range up to and including the size you enter.
- ▶ Click OK. The affected ranges are converted to either *Flag* or *Ordinal* and displayed on the Types tab of the Type node .

Results of the Conversion

- Where a *Continuous* field with integer storage is changed to *Ordinal*, the lower and upper values are expanded to include all of the integer values from the lower value to the upper value. For example, if the range is 1, 5, the set of values is 1, 2, 3, 4, 5.
- If the *Continuous* field changes to *Flag*, the lower and upper values become the false and true values of the flag field.

What Is Instantiation?

Instantiation is the process of reading or specifying information, such as storage type and values for a data field. In order to optimize system resources, instantiating is a user-directed process—you tell the software to read values by specifying options on the Types tab in a source node or by running data through a Type node.

- Data with unknown types are also referred to as **uninstantiated**. Data whose storage type and values are unknown are displayed in the *Measurement* column of the Types tab as <Default>.
- When you have some information about a field's storage, such as string or numeric, the data are called **partially instantiated**. Categorical or Continuous are partially instantiated measurement levels. For example, Categorical specifies that the field is symbolic, but you don't know whether it is nominal, ordinal, or flag.
- When all of the details about a type are known, including the values, a **fully instantiated** measurement level—nominal, ordinal, flag, or continuous—is displayed in this column. *Note:* The *continuous* type is used for both partially instantiated and fully instantiated data fields. Continuous data can be either integers or real numbers.

During the execution of a data stream with a Type node, uninstantiated types immediately become partially instantiated, based on the initial data values. Once all of the data have passed through the node, all data become fully instantiated unless values were set to <Pass>. If execution is interrupted, the data will remain partially instantiated. Once the Types tab has been instantiated, the values of a field are static at that point in the stream. This means that any upstream changes will not affect the values of a particular field, even if you rerun the stream. To change or update

the values based on new data or added manipulations, you need to edit them in the Types tab itself or set the value for a field to <Read> or <Read +>.

When to Instantiate

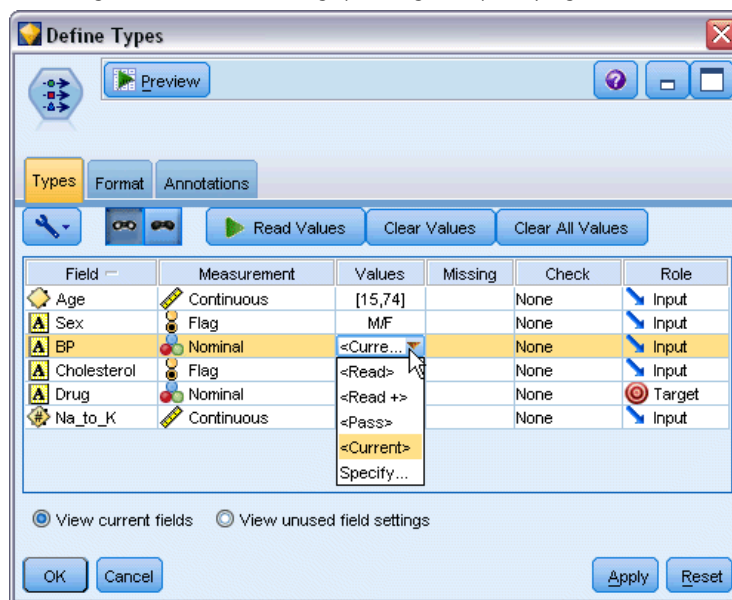
Generally, if your dataset is not very large and you do not plan to add fields later in the stream, instantiating at the source node is the most convenient method. However, instantiating in a separate Type node is useful when:

- The dataset is large, and the stream filters a subset prior to the Type node.
- Data have been filtered in the stream.
- Data have been merged or appended in the stream.
- New data fields are derived during processing.

Data Values

Using the *Values* column of the Types tab, you can read values automatically from the data, or you can specify measurement levels and values in a separate dialog box.

Figure 4-22
Selecting methods for reading, passing, or specifying data values



The options available from this drop-down list provide the following instructions for auto-typing:

Option	Function
<Read>	Data will be read when the node is executed.
<Read+>	Data will be read and appended to the current data (if any exist).
<Pass>	No data are read.

Option	Function
<Current>	Keep current data values.
Specify...	A separate dialog box is launched for you to specify values and measurement level options.

Executing a Type node or clicking Read Values will auto-type and read values from your data source based on your selection. These values can also be specified manually using the Specify option or by double-clicking a cell in the *Field* column.

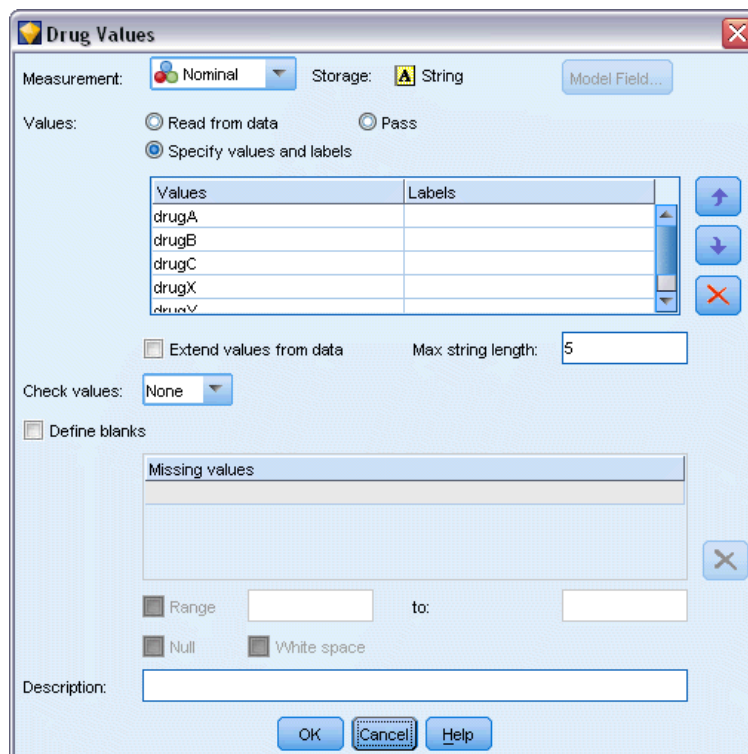
Once you have made changes for fields in the Type node, you can reset value information using the following buttons on the dialog box toolbar:

- Using the Clear Values button, you can clear changes to field values made in this node (non-inherited values) and reread values from upstream operations. This option is useful for resetting changes that you may have made for specific fields upstream.
- Using the Clear All Values button, you can reset values for **all** fields read into the node. This option effectively sets the *Values* column to **Read** for all fields. This option is useful to reset values for all fields and reread values and measurement levels from upstream operations.

Using the Values Dialog Box

Clicking the *Values* or *Missing* column of the Types tab displays a drop-down list of predefined values. Choosing the *Specify* option on this list opens a separate dialog box where you can set options for reading, specifying, labeling, and handling values for the selected field.

Figure 4-23
Setting options for data values



Many of the controls are common to all types of data. These common controls are discussed here.

Measurement. Displays the currently selected measurement level. You can change the setting to reflect the way that you intend to use data. For instance, if a field called *day_of_week* contains numbers representing individual days, you may want to change this to nominal data in order to create a distribution node that examines each category individually.

Storage. Displays the storage type if known. Storage types are unaffected by the measurement level that you choose. To alter the storage type, you can use the Data tab in Fixed File and Variable File source nodes, or a conversion function in a Filler node.

Model Field. For fields generated as a result of scoring a model nugget, model field details can also be viewed. These include the name of the target field as well as the role of the field in modeling (whether a predicted value, probability, propensity, etc.).

Values. Select a method to determine values for the selected field. Selections that you make here override any selections that you made earlier from the *Values* column of the Type node dialog box. Choices for reading values include:

- **Read from data.** Select to read values when the node is executed. This option is the same as <Read>.
- **Pass.** Select not to read data for the current field. This option is the same as <Pass>.
- **Specify values and labels.** Options here are used to specify values and labels for the selected field. Used in conjunction with value checking, this option allows you to specify values based on your knowledge of the current field. This option activates unique controls for each type of field. Options for values and labels are covered individually in subsequent topics. *Note:* You cannot specify values or labels for a field whose measurement level is *Typeless* or <Default>.
- **Extend values from data.** Select to append the current data with the values that you enter here. For example, if *field_1* has a range from (0,10), and you enter a range of values from (8,16), the range is extended by adding the 16, without removing the original minimum. The new range would be (0,16). Choosing this option automatically sets the auto-typing option to <Read+>.

Check values. Select a method of coercing values to conform to the specified continuous, flag, or nominal values. This option corresponds to the *Check* column in the Type node dialog box, and settings made here override those in the dialog box. Used in conjunction with the Specify Values option, value checking allows you to conform values in the data with expected values. For example, if you specify values as 1, 0 and then use the Discard option, you can discard all records with values other than 1 or 0.

Define blanks. Select to activate the controls below that enable you to declare missing values or blanks in your data.

- **Missing values table.** Allows you to define specific values (such as 99 or 0) as blanks. The value should be appropriate for the storage type of the field.
- **Range.** Used to specify a range of missing values, for example, ages 1–17 or greater than 65. If a bound value is left blank then the range will be unbounded; for example, if a lower bound of 100 is specified with no upper bound, then all values greater than or equal to 100 will be defined as missing. The bound values are inclusive; for example, a range with a lower bound of 5 and an upper bound of 10 will include 5 and 10 in the range definition. A missing value

range can be defined for any storage type, including date/time and string (in which case the alphabetic sort order will be used to determine whether a value is within the range).

- **Null/White space.** You can also specify system **nulls** (displayed in the data as \$null\$) and **white space** (string values with no visible characters) as blanks. Note that the Type node also treats empty strings as white space for purposes of analysis, although they are stored differently internally and may be handled differently in certain cases.

Note: To code blanks as undefined or \$null\$, you should use the Filler node.

Description. Use this text box to specify a field label. These labels appear in a variety of locations, such as in graphs, tables, output, and model browsers, depending on selections you make in the stream properties dialog box.

Specifying Values and Labels for Continuous Data

The *Continuous* measurement level is used for numeric fields. There are three storage types for continuous data:

- Real
- Integer
- Date/Time

The same dialog box is used to edit all continuous fields; the storage type is displayed for reference only.

Figure 4-24
Options for specifying continuous values and their labels

Measurement: Storage:

Values: Read from data Pass
 Specify values and labels

Lower:

Upper:

Specifying Values

The following controls are unique to continuous fields and are used to specify a range of values:

Lower. Specify a lower limit for the value range.

Upper. Specify an upper limit for the value range.

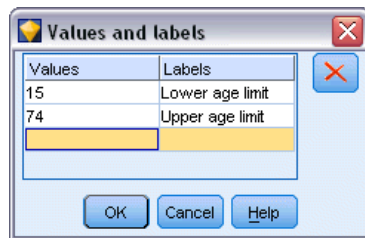
Specifying Labels

You can specify labels for any value of a range field. Click the Labels button to open a separate dialog box for specifying value labels.

Values and Labels Subdialog Box

Clicking Labels in the Values dialog box for a range field opens a new dialog box in which you can specify labels for any value in the range.

Figure 4-25
Providing labels (optional) for range values



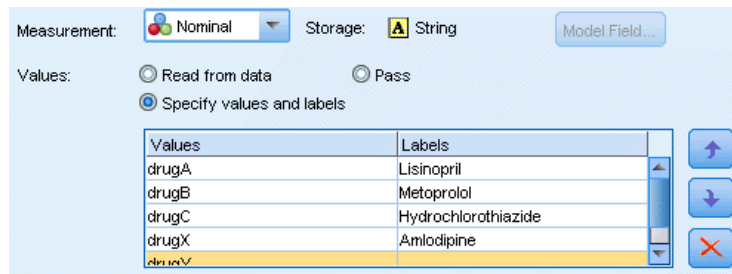
You can use the *Values* and *Labels* columns in this table to define value and label pairs. Currently defined pairs are shown here. You can add new label pairs by clicking in an empty cell and entering a value and its corresponding label. *Note:* Adding value/value-label pairs to this table will not cause any new values to be added to the field. Instead, it simply creates metadata for the field value.

The labels that you specify in the Type node are displayed in many places (as ToolTips, output labels, and so on), depending on selections that you make in the stream properties dialog box.

Specifying Values and Labels for Nominal and Ordinal Data

Nominal (set) and ordinal (ordered set) measurement levels indicate that the data values are used discretely as a member of the set. The storage types for a set can be string, integer, real number, or date/time.

Figure 4-26
Options for specifying nominal values and labels



The following controls are unique to nominal and ordinal fields and are used to specify values and labels:

Values. The *Values* column in the table allows you to specify values based on your knowledge of the current field. Using this table, you can enter expected values for the field and check the dataset's conformity to these values using the Check Values drop-down list. Using the arrow and delete buttons, you can modify existing values as well as reorder or delete values.

Labels. The *Labels* column enables you to specify labels for each value in the set. These labels appear in a variety of locations, such as graphs, tables, output, and model browsers, depending on selections that you make in the stream properties dialog box.

Specifying Values for a Flag

Flag fields are used to display data that have two distinct values. The storage types for flags can be string, integer, real number, or date/time.

Figure 4-27
Options for specifying flag field values

The screenshot shows a dialog box for configuring a flag field. At the top, 'Measurement' is set to 'Flag' and 'Storage' is 'String'. Below this, there are two radio buttons: 'Read from data' (unselected) and 'Specify values and labels' (selected). Under 'Specify values and labels', there are two rows of input fields. The first row is for the 'True' state, with a value of 'M' and a label of 'Male'. The second row is for the 'False' state, with a value of 'F' and a label of 'Female'. A 'Model Field...' button is visible in the top right corner.

True. Specify a flag value for the field when the condition is met.

False. Specify a flag value for the field when the condition is not met.

Labels. Specify labels for each value in the flag field. These labels appear in a variety of locations, such as graphs, tables, output, and model browsers, depending on selections that you make in the stream properties dialog box.

Defining Missing Values

The Missing column of the Types tab indicates whether missing value handling has been defined for a field. The possible settings are:

On (*). Indicates that missing values handling is defined for this field. This could be done by means of a downstream Filler node, or through an explicit specification using the Specify option (see below).

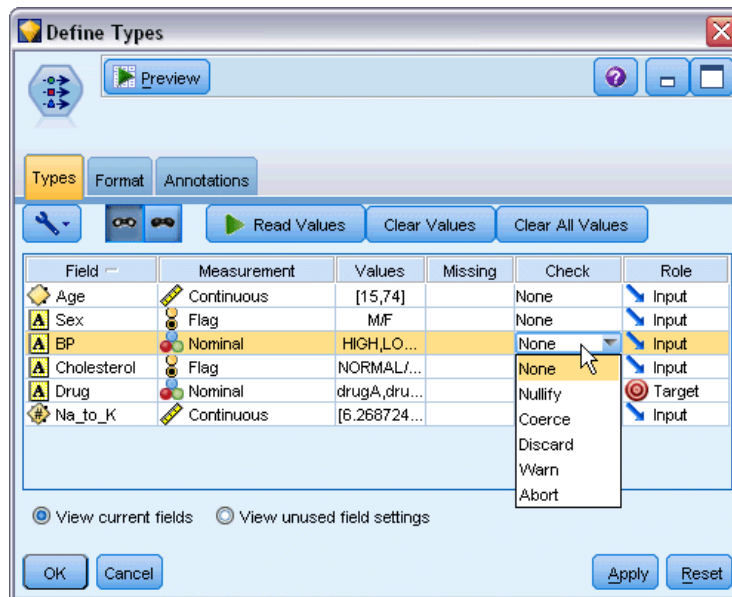
Off. The field has no missing value handling defined.

Specify. Choose this option to display a dialog where you can declare explicit values to be considered as missing values for this field.

Checking Type Values

Turning on the Check option for each field examines all values in that field to determine whether they comply with the current type settings or the values that you have specified in the Specify Values dialog box. This is useful for cleaning up datasets and reducing the size of a dataset within a single operation.

Figure 4-28
Selecting Check options for the selected field



The setting of the *Check* column in the Type node dialog box determines what happens when a value outside of the type limits is discovered. To change the Check settings for a field, use the drop-down list for that field in the *Check* column. To set the Check settings for all fields, click in the *Field* column and press Ctrl-A. Then use the drop-down list for any field in the *Check* column.

The following Check settings are available:

None. Values will be passed through without checking. This is the default setting.

Nullify. Change values outside of the limits to the system null (\$null\$).

Coerce. Fields whose measurement levels are fully instantiated will be checked for values that fall outside the specified ranges. Unspecified values will be converted to a legal value for that measurement level using the following rules:

- For flags, any value other than the true and false value is converted to the false value.
- For sets (nominal or ordinal), any unknown value is converted to the first member of the set's values.
- Numbers greater than the upper limit of a range are replaced by the upper limit.
- Numbers less than the lower limit of a range are replaced by the lower limit.
- Null values in a range are given the midpoint value for that range.

Discard. When illegal values are found, the entire record is discarded.

Warn. The number of illegal items is counted and reported in the stream properties dialog box when all of the data have been read.

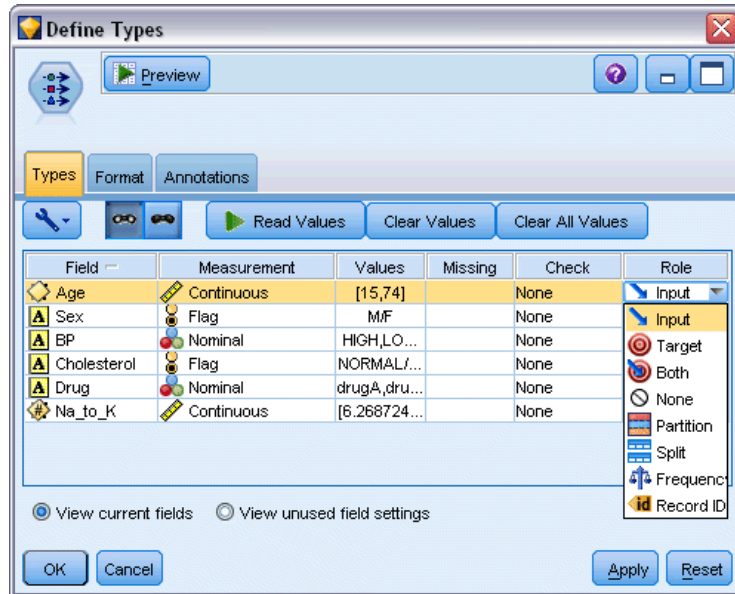
Abort. The first illegal value encountered terminates the running of the stream. The error is reported in the stream properties dialog box.

Setting the Field Role

The role of a field specifies how it is used in model building—for example, whether a field is an input or target (the thing being predicted).

Note: The Partition, Frequency and Record ID roles can each be applied to a single field only.

Figure 4-29
Setting field role options for the Type node



The following roles are available:

Input. The field will be used as an input to machine learning (a predictor field).

Target. The field will be used as an output or target for machine learning (one of the fields that the model will try to predict).

Both. The field will be used as both an input and an output by the Apriori node. All other modeling nodes will ignore the field.

None. The field will be ignored by machine learning. Fields whose measurement level has been set to Typeless are automatically set to None in the *Role* column.

Partition. Indicates a field used to partition the data into separate samples for training, testing, and (optional) validation purposes. The field must be an instantiated set type with two or three possible values (as defined in the Field Values dialog box). The first value represents the training sample, the second represents the testing sample, and the third (if present) represents the validation sample. Any additional values are ignored, and flag fields cannot be used. Note that to use the partition in an analysis, partitioning must be enabled on the Model Options tab in the appropriate model-building or analysis node. Records with null values for the partition field are excluded from the analysis when partitioning is enabled. If multiple partition fields have been defined in the stream, a single partition field must be specified on the Fields tab in each applicable modeling

node. If a suitable field doesn't already exist in your data, you can create one using a Partition node or Derive node. For more information, see the topic [Partition Node](#) on p. 176.

Split. (Nominal, ordinal and flag fields only) Specifies that a model is to be built for each possible value of the field.

Frequency. (Numeric fields only) Setting this role enables the field value to be used as a frequency weighting factor for the record. This feature is supported by C&R Tree, CHAID, QUEST and Linear models only; all other nodes ignore this role. Frequency weighting is enabled by means of the Use frequency weight option on the Fields tab of those modeling nodes that support the feature.

Record ID. The field will be used as the unique record identifier. This feature is ignored by most nodes; however it is supported by Linear models, and is required for the IBM Netezza in-database mining nodes.

Copying Type Attributes

You can easily copy the attributes of a type, such as values, checking options, and missing values from one field to another:

- ▶ Right-click on the field whose attributes you want to copy.
- ▶ From the context menu, choose Copy.
- ▶ Right-click on the field(s) whose attributes you want to change.
- ▶ From the context menu, choose Paste Special. *Note:* You can select multiple fields using the Ctrl-click method or by using the Select Fields option from the context menu.

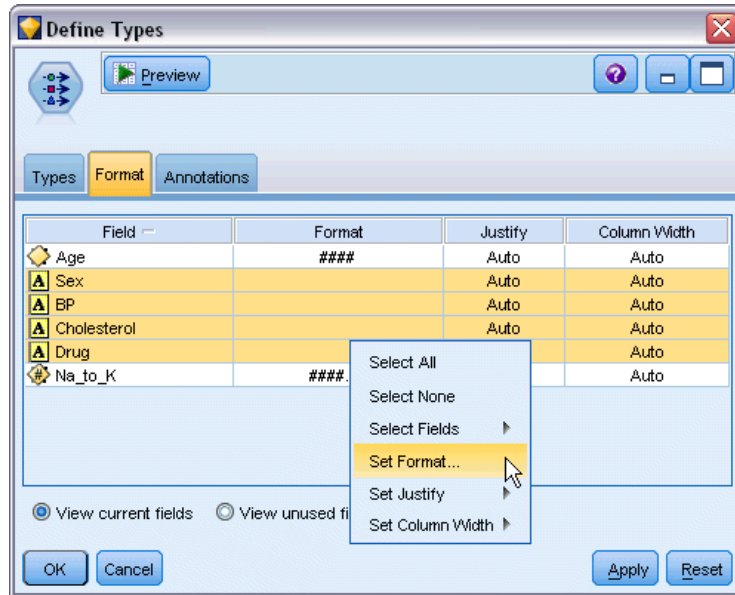
A new dialog box opens, from which you can select the specific attributes that you want to paste. If you are pasting into multiple fields, the options that you select here will apply to all target fields.

Paste the following attributes. Select from the list below to paste attributes from one field to another.

- **Type.** Select to paste the measurement level.
- **Values.** Select to paste the field values.
- **Missing.** Select to paste missing value settings.
- **Check.** Select to paste value checking options.
- **Role.** Select to paste the role of a field.

Field Format Settings Tab

Figure 4-30
Type node, Format tab



The Format tab on the Table and Type nodes shows a list of current or unused fields and formatting options for each field. Following is a description of each column in the field formatting table:

Field. This shows the name of the selected field.

Format. By double-clicking a cell in this column, you can specify formatting for fields on an individual basis using the dialog box that opens. For more information, see the topic [Setting Field Format Options](#) on p. 129. Formatting specified here overrides formatting specified in the overall stream properties.

Note: The Statistics Export and Statistics Output nodes export *.sav* files that include per-field formatting in their metadata. If a per-field format is specified that is not supported by the IBM® SPSS® Statistics *.sav* file format, then the node will use the SPSS Statistics default format.

Justify. Use this column to specify how the values should be justified within the table column. The default setting is Auto, which left-justifies symbolic values and right-justifies numeric values. You can override the default by selecting Left, Right, or Center.

Column Width. By default, column widths are automatically calculated based on the values of the field. To override the automatic width calculation, click a table cell and use the drop-down list to select a new width. To enter a custom width not listed here, open the Field Formats subdialog box by double-clicking a table cell in the *Field* or *Format* column. Alternatively, you can right-click on a cell and choose Set Format.

View current fields. By default, the dialog box shows the list of currently active fields. To view the list of unused fields, select View unused fields settings.

Context menu. The context menu for this tab provides various selection and setting update options.

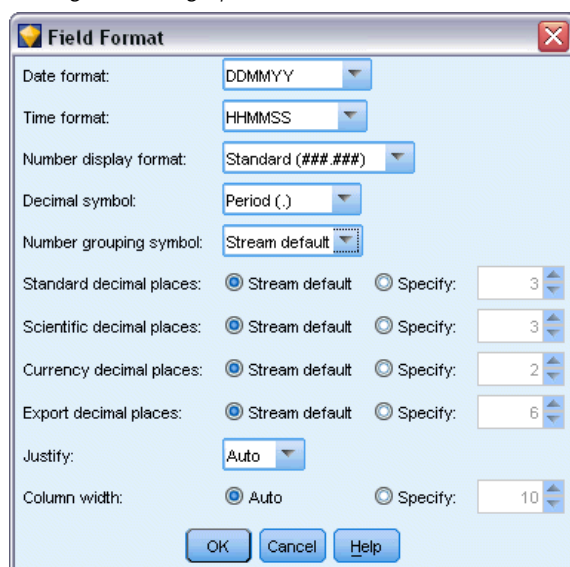
- **Select All.** Selects all fields.

- **Select None.** Clears the selection.
- **Select Fields.** Selects fields based on type or storage characteristics. Options are Select Categorical, Select Continuous (numeric), Select Typeless, Select Strings, Select Numbers, or Select Date/Time. For more information, see the topic [Measurement Levels](#) on p. 115.
- **Set Format.** Opens a subdialog box for specifying date, time, and decimal options on a per-field basis.
- **Set Justify.** Sets the justification for the selected field(s). Options are Auto, Center, Left, or Right.
- **Set Column Width.** Sets the field width for selected fields. Specify Auto to read width from the data. Or you can set field width to 5, 10, 20, 30, 50, 100, or 200.

Setting Field Format Options

Field formatting is specified on a subdialog box available from the Format tab on the Type and Table nodes. If you have selected more than one field before opening this dialog box, then settings from the first field in the selection are used for all. Clicking OK after making specifications here will apply these settings to all fields selected on the Format tab.

Figure 4-31
Setting formatting options for one or more fields



The following options are available on a per-field basis. Many of these settings can also be specified in the stream properties dialog box. Any settings made at the field level override the default specified for the stream.

Date format. Select a date format to be used for date storage fields or when strings are interpreted as dates by CLEM date functions.

Time format. Select a time format to be used for time storage fields or when strings are interpreted as times by CLEM time functions.

Number display format. You can choose from standard (#####.###), scientific (#.###E+##), or currency display formats (\$###.###).

Decimal symbol. Select either a comma (,) or period (.) as a decimal separator.

Grouping symbol. For number display formats, select the symbol used to group values (For example, the comma in 3,000.00). Options include none, period, comma, space, and locale-defined (in which case the default for the current locale is used).

Decimal places (standard, scientific, currency, export). For number display formats, specifies the number of decimal places to be used when displaying, printing, or exporting real numbers. This option is specified separately for each display format. The export format applies only to fields with real storage.

Justify. Specifies how the values should be justified within the column. The default setting is Auto, which left-justifies symbolic values and right-justifies numeric values. You can override the default by selecting left, right, or center.

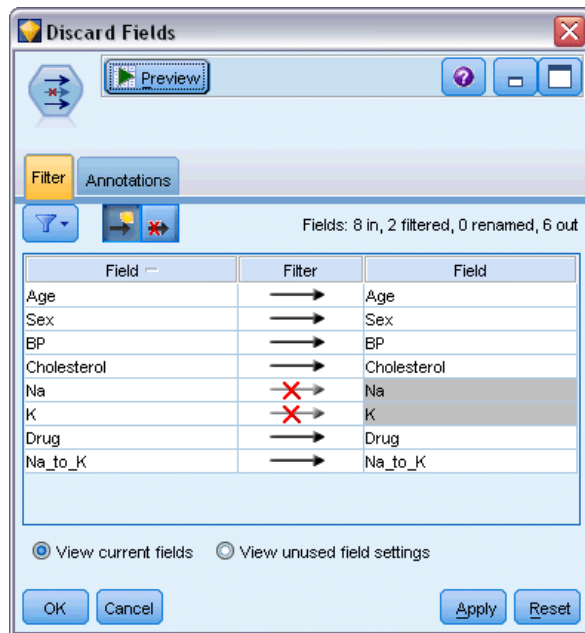
Column width. By default, column widths are automatically calculated based on the values of the field. You can specify a custom width in intervals of five using the arrows to the right of the list box.

Filtering or Renaming Fields

You can rename or exclude fields at any point in a stream. For example, as a medical researcher, you may not be concerned about the potassium level (field-level data) of patients (record-level data); therefore, you can filter out the *K* (potassium) field. This can be done using a separate Filter node or using the Filter tab on a source or output node. The functionality is the same regardless of which node it is accessed from.

- From source nodes, such as Variable File, Fixed File, Statistics File and XML, you can rename or filter fields as the data are read into IBM® SPSS® Modeler.
- Using a Filter node, you can rename or filter fields at any point in the stream.
- From Statistics Export, Statistics Transform, Statistics Model, and Statistics Output nodes, you can filter or rename fields to conform to IBM® SPSS® Statistics naming standards. For more information, see the topic [Renaming or Filtering Fields for IBM SPSS Statistics](#) in Chapter 8 on p. 410.
- You can use the Filter tab in any of the above nodes to define or edit multiple response sets. For more information, see the topic [Editing Multiple Response Sets](#) on p. 134.
- Finally, you can use a Filter node to map fields from one source node to another.

Figure 4-32
Setting Filter node options



Setting Filtering Options

The table used on the Filter tab shows the name of each field as it comes into the node as well as the name of each field as it leaves. You can use the options in this table to rename or filter out fields that are duplicates or are unnecessary for downstream operations.

- **Field.** Displays the input fields from currently connected data sources.
- **Filter.** Displays the filter status of all input fields. Filtered fields include a red X in this column, indicating that this field will not be passed downstream. Click in the *Filter* column for a selected field to turn filtering on and off. You can also select options for multiple fields simultaneously using the Shift-click method of selection.
- **Field.** Displays the fields as they leave the Filter node. Duplicate names are displayed in red. You can edit field names by clicking in this column and entering a new name. Or, remove fields by clicking in the *Filter* column to disable duplicate fields.

All columns in the table can be sorted by clicking on the column header.

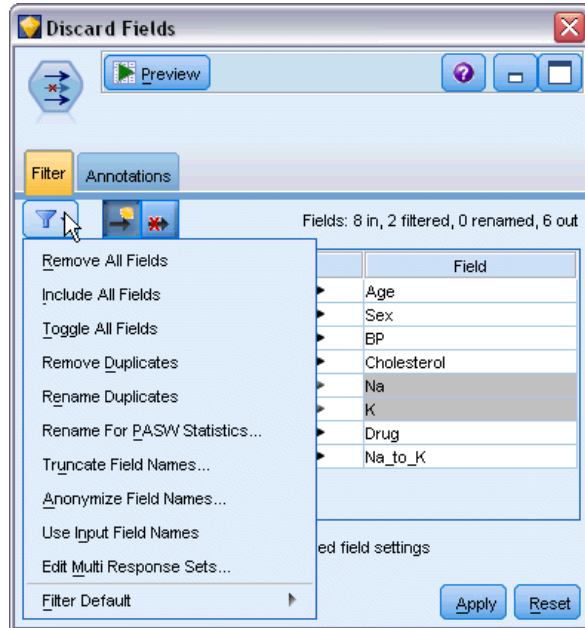
View current fields. Select to view fields for datasets actively connected to the Filter node. This option is selected by default and is the most common method of using Filter nodes.

View unused field settings. Select to view fields for datasets that were once but are no longer connected to the Filter node. This option is useful when copying Filter nodes from one stream to another or when saving and reloading Filter nodes.

Filter Button Menu

Click the Filter button in the upper left corner of the dialog box to access a menu that provides a number of shortcuts and other options.

Figure 4-33
Filter menu options



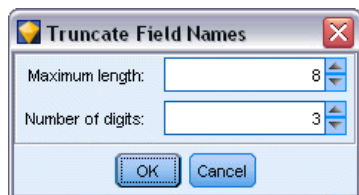
You can choose to:

- Remove all fields.
- Include all fields.
- Toggle all fields.
- Remove duplicates. *Note:* Selecting this option removes all occurrences of the duplicate name, including the first one.
- Rename fields and multiple response sets to conform with other applications. For more information, see the topic [Renaming or Filtering Fields for IBM SPSS Statistics](#) in Chapter 8 on p. 410.
- Truncate field names.
- Anonymize field and multiple response set names.
- Use input field names.
- Edit Multiple Response Sets. For more information, see the topic [Editing Multiple Response Sets](#) on p. 134.
- Set the default filter state.

You can also use the arrow toggle buttons at the top of the dialog box to specify whether you want to include or discard fields by default. This is useful for large datasets where only a few fields are to be included downstream. For example, you can select only the fields you want to keep and specify that all others should be discarded (rather than individually selecting all of the fields to discard).

Truncating Field Names

Figure 4-34
Truncate Field Names dialog box



From the Filter button menu (upper left corner of the Filter tab), you can choose to truncate field names.

Maximum length. Specify a number of characters to limit the length of field names.

Number of digits. If field names, when truncated, are no longer unique, they will be further truncated and differentiated by adding digits to the name. You can specify the number of digits used. Use the arrow buttons to adjust the number.

For example, the table below illustrates how field names in a medical dataset are truncated using the default settings (maximum length=8 and number of digits=2).

Field Names	Truncated Field Names
Patient Input 1	Patien01
Patient Input 2	Patien02
Heart Rate	HeartRat
BP	BP

Anonymizing Field Names

Figure 4-35
Anonymize Field Names dialog box



You can anonymize field names from any node that includes a Filter tab by clicking the Filter button menu in the upper left corner and choosing Anonymize Field Names. Anonymized field names consist of a string prefix plus a unique numeric-based value.

Anonymize names of. Choose Selected fields only to anonymize only the names of fields already selected on the Filter tab. The default is All fields, which anonymizes all field names.

Field names prefix. The default prefix for anonymized field names is anon_; choose Custom and type your own prefix if you want a different one.

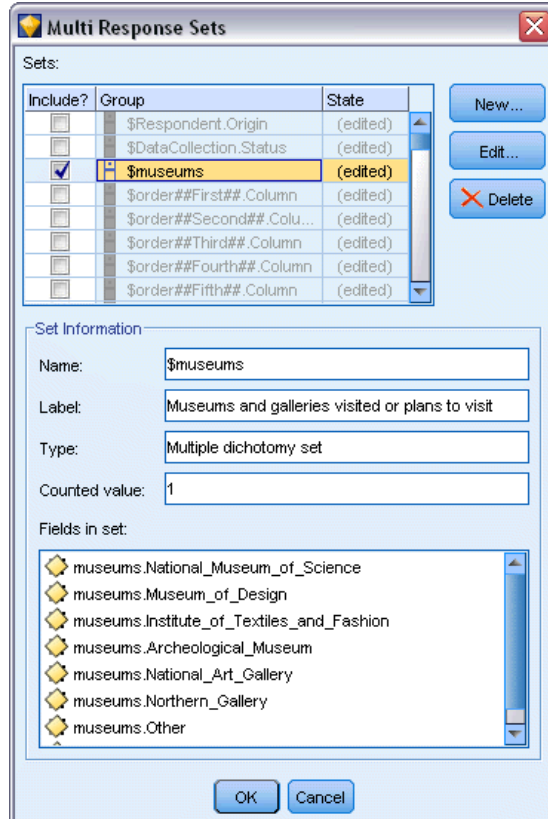
Anonymize multiple response sets. Anonymizes the names of multiple response sets in the same manner as fields. For more information, see the topic [Editing Multiple Response Sets](#) on p. 134.

To restore the original field names, choose Use Input Field Names from the filter button menu.

Editing Multiple Response Sets

You can add or edit multiple response sets from any node that includes a Filter tab by clicking the Filter button menu in the upper left corner and choosing Edit Multiple Response Sets.

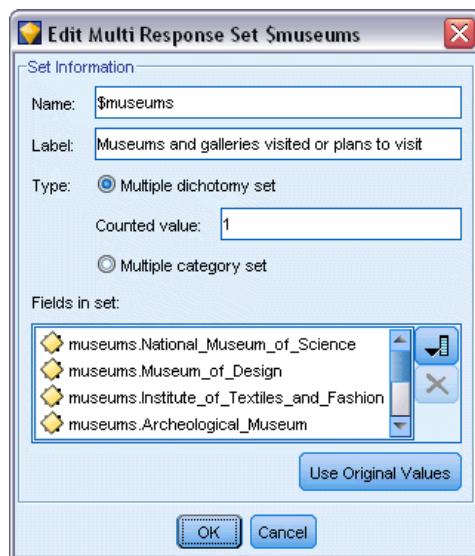
Figure 4-36
Multiple Response Sets dialog box



Multiple response sets are used to record data that can have more than one value for each case—for example, when asking survey respondents which museums they have visited or which magazines they read. Multiple response sets can be imported into IBM® SPSS® Modeler using a Data Collection source node or a Statistics File source node and can be defined in SPSS Modeler using a Filter node.

- Click **New** to create a new multiple response set, or click **Edit** to modify an existing set.

Figure 4-37
Editing a multiple response set



Name and label. Specifies the name and description for the set.

Type. Multiple response questions can be handled in one of two ways:

- **Multiple dichotomy set.** A separate flag field is created for each possible response; thus, if there are 10 magazines, there are 10 flag fields, each of which can have values, such as 0 or 1 for *true* or *false*. The counted value allows you to specify which value is counted as true. This method is useful when you want to allow respondents to choose all options that apply.
- **Multiple category set.** A nominal field is created for each response up to the maximum number of answers from a given respondent. Each nominal field has values representing the possible answers, such as 1 for *Time*, 2 for *Newsweek*, and 3 for *PC Week*. This method is most useful when you want to limit the number of answers—for example, when asking respondents to choose the three magazines they read most frequently.

Fields in set. Use the icons on the right to add or remove fields.

Figure 4-38
Multiple response question

Q14 Which museums or art galleries have you visited or do you intend to visit?
SELECT ALL ANSWERS THAT APPLY.

National Museum of Science

Museum of Design

Institute of Textiles and Fashion

Archeological Museum

National Art Gallery

Northern Gallery

Other (Please write in) _____

Not answered

Comments

- All fields included in a multiple response set must have the same storage.
- Sets are distinct from the fields they include. For example, deleting a set will not cause the fields it includes to be deleted—merely the links between those fields. The set is still visible upstream from the point of deletion but is not visible downstream.
- If fields are renamed using a Filter node (directly on the tab or by choosing the Rename for IBM® SPSS® Statistics, Truncate, or Anonymize options on the Filter menu), any references to these fields used in multiple response sets will also be updated. However, any fields in a multiple response set that are dropped by the Filter node will not be removed from the multiple response set. Such fields, although no longer visible in the stream, are still referenced by the multiple response set; this could be a consideration when exporting, for example.

Ensemble Node

The Ensemble node combines two or more model nuggets to obtain more accurate predictions than can be gained from any of the individual models. By combining predictions from multiple models, limitations in individual models may be avoided, resulting in a higher overall accuracy. Models combined in this manner typically perform at least as well as the best of the individual models and often better.

This combining of nodes happens automatically in the Auto Classifier, Auto Numeric and Auto Cluster automated modeling nodes.

After using an Ensemble node, you can use an Analysis node or Evaluation node to compare the accuracy of the combined results with each of the input models. To do this, make sure the Filter out fields generated by ensembled models option is not selected on the Settings tab in the Ensemble node.

Output Fields

Each Ensemble node generates a field containing the combined scores. The name is based on the specified target field and prefixed with $\$XF_$, $\$XS_$, or $\$XR_$, depending on the field measurement level—flag, nominal (set), or continuous (range), respectively. For example, if the target is a flag field named *response*, the output field would be $\$XF_response$.

Confidence or propensity fields. For flag and nominal fields, additional confidence or propensity fields are created based on the ensemble method, as detailed in the following table:

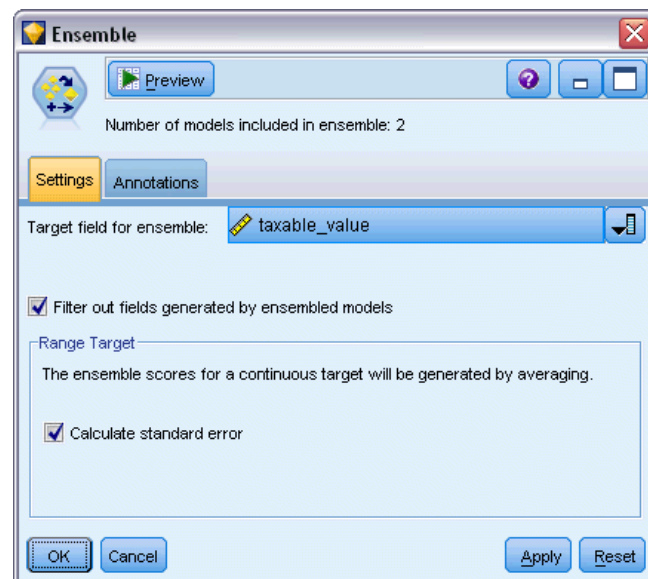
Ensemble method	Field name
Voting Confidence-weighted voting Raw-propensity-weighted voting Raw-propensity-weighted voting Highest confidence wins	$\$XFC_<field>$
Average raw propensity	$\$XFRP_<field>$
Average adjusted raw propensity	$\$XFAP_<field>$

Ensemble Node Settings

Target field for ensemble. Select a single field that is used as the target by two or more upstream models. The upstream models can use flag, nominal, or continuous targets, but at least two of the models must share the same target in order to combine scores.

Filter out fields generated by ensembled models. Removes from the output all of the additional fields generated by the individual models that feed into the Ensemble node. Select this check box if you are interested only in the combined score from all of the input models. Ensure that this option is deselected if, for example, you want to use an Analysis node or Evaluation node to compare the accuracy of the combined score with that of each of the individual input models.

Figure 4-39
Ensemble node with a continuous field selected as the target



Available settings depend on the measurement level of the field selected as the target.

Continuous Targets

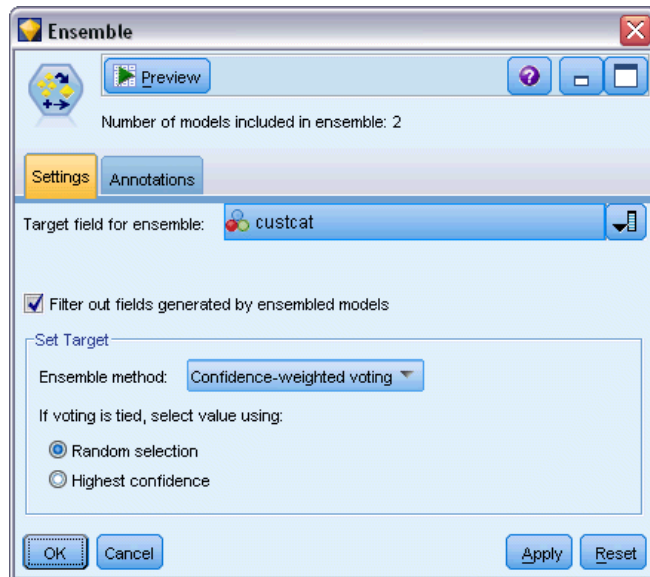
For a continuous target, scores will be averaged. This is the only available method for combining scores.

When averaging scores or estimates, the Ensemble node uses a standard error calculation to work out the difference between the measured or estimated values and the true values, and to show how close those estimates matched. Standard error calculations are generated by default for new models; however, you can deselect the check box for existing models, for example, if they are to be regenerated.

Categorical Targets

For categorical targets, a number of methods are supported, including **voting**, which works by tallying the number of times each possible predicted value is chosen and selecting the value with the highest total. For example, if three out of five models predict *yes* and the other two predict *no*, then *yes* wins by a vote of 3 to 2. Alternatively, votes can be **weighted** based on the confidence or propensity value for each prediction. The weights are then summed, and the value with the highest total is again selected. The confidence for the final prediction is the sum of the weights for the winning value divided by the number of models included in the ensemble.

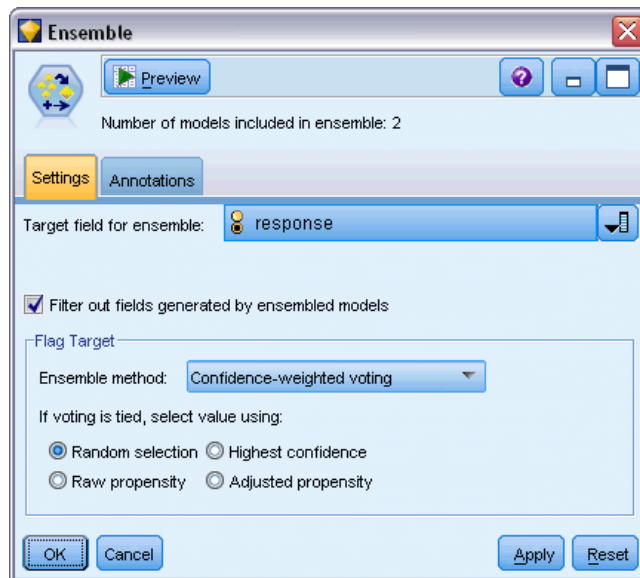
Figure 4-40
Ensemble node with a nominal field selected as the target



All categorical fields. For both flag and nominal fields, the following methods are supported:

- Voting
- Confidence-weighted voting
- Highest confidence wins

Figure 4-41
Ensemble node with a flag field selected as the target



Flag fields only. For flag fields only, a number of methods based on propensity are also available:

- Raw propensity-weighted voting
- Adjusted propensity-weighted voting
- Average raw propensity
- Average adjusted propensity

Voting ties. For voting methods, you can specify how ties are resolved.

- **Random selection.** One of the tied values is chosen at random.
- **Highest confidence.** The tied value that was predicted with the highest confidence wins. Note that this is not necessarily the same as the highest confidence of all predicted values.
- **Raw or adjusted propensity (flag fields only).** The tied value that was predicted with the largest absolute propensity, where the absolute propensity is calculated as:

$$\frac{\text{abs}(0.5 - \text{propensity})}{2}$$

Or, in the case of adjusted propensity:

$$\text{abs}(0.5 - \text{adjusted propensity}) * 2$$

Derive Node

One of the most powerful features in IBM® SPSS® Modeler is the ability to modify data values and derive new fields from existing data. During lengthy data mining projects, it is common to perform several derivations, such as extracting a customer ID from a string of Web log data

or creating a customer lifetime value based on transaction and demographic data. All of these transformations can be performed, using a variety of field operations nodes.

Several nodes provide the ability to derive new fields:



The Derive node modifies data values or creates new fields from one or more existing fields. It creates fields of type formula, flag, nominal, state, count, and conditional. For more information, see the topic [Derive Node](#) on p. 139.



The Reclassify node transforms one set of categorical values to another. Reclassification is useful for collapsing categories or regrouping data for analysis. For more information, see the topic [Reclassify Node](#) on p. 158.



The Binning node automatically creates new nominal (set) fields based on the values of one or more existing continuous (numeric range) fields. For example, you can transform a continuous income field into a new categorical field containing groups of income as deviations from the mean. Once you have created bins for the new field, you can generate a Derive node based on the cut points. For more information, see the topic [Binning Node](#) on p. 162.



The Set to Flag node derives multiple flag fields based on the categorical values defined for one or more nominal fields. For more information, see the topic [Set to Flag Node](#) on p. 178.



The Restructure node converts a nominal or flag field into a group of fields that can be populated with the values of yet another field. For example, given a field named *payment type*, with values of *credit*, *cash*, and *debit*, three new fields would be created (*credit*, *cash*, *debit*), each of which might contain the value of the actual payment made. For more information, see the topic [Restructure Node](#) on p. 180.



The History node creates new fields containing data from fields in previous records. History nodes are most often used for sequential data, such as time series data. Before using a History node, you may want to sort the data using a Sort node. For more information, see the topic [History Node](#) on p. 205.

Using the Derive Node

Using the Derive node, you can create six types of new fields from one or more existing fields:

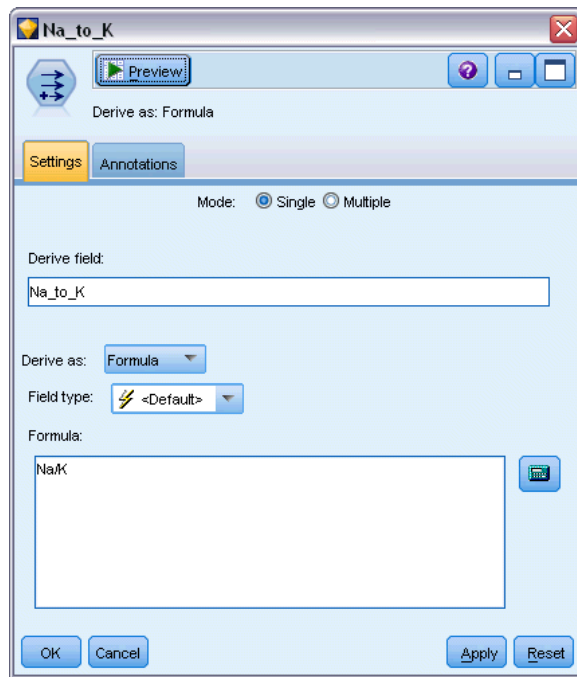
- **Formula.** The new field is the result of an arbitrary CLEM expression.
- **Flag.** The new field is a flag, representing a specified condition.
- **Nominal.** The new field is nominal, meaning that its members are a group of specified values.
- **State.** The new field is one of two states. Switching between these states is triggered by a specified condition.
- **Count.** The new field is based on the number of times that a condition has been true.
- **Conditional.** The new field is the value of one of two expressions, depending on the value of a condition.

Each of these nodes contains a set of special options in the Derive node dialog box. These options are discussed in subsequent topics.

Setting Basic Options for the Derive Node

At the top of the dialog box for Derive nodes are a number of options for selecting the type of Derive node that you need.

Figure 4-42
Derive node dialog box



Mode. Select Single or Multiple, depending on whether you want to derive multiple fields. When Multiple is selected, the dialog box changes to include options for multiple Derive fields.

Derive field. For simple Derive nodes, specify the name of the field that you want to derive and add to each record. The default name is $DeriveN$, where N is the number of Derive nodes that you have created thus far during the current session.

Derive as. Select a type of Derive node, such as Formula or Nominal, from the drop-down list. For each type, a new field is created based on the conditions that you specify in the type-specific dialog box.

Selecting an option from the drop-down list will add a new set of controls to the main dialog box according to the properties of each Derive node type.

Field type. Select a measurement level, such as continuous, categorical, or flag, for the newly derived node. This option is common to all forms of Derive nodes.

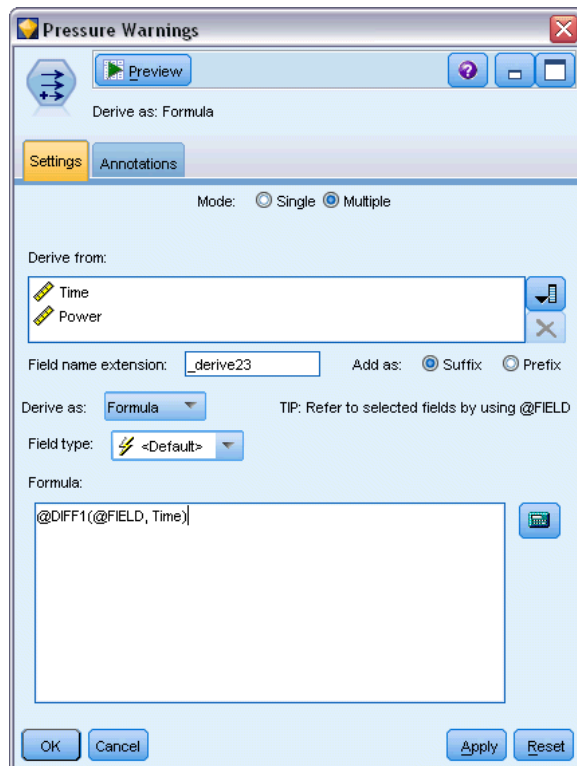
Note: Deriving new fields often requires the use of special functions or mathematical expressions. To help you create these expressions, an Expression Builder is available from the dialog box for all types of Derive nodes and provides rule checking as well as a complete list of CLEM expressions.

Deriving Multiple Fields

Setting the mode to Multiple within a Derive node gives you the capability to derive multiple fields based on the same condition within the same node. This feature saves time when you want to make identical transformations on several fields in your dataset. For example, if you want to build a regression model predicting current salary based on beginning salary and previous experience, it might be beneficial to apply a log transformation to all three skewed variables. Rather than add a new Derive node for each transformation, you can apply the same function to all fields at once. Simply select all fields from which to derive a new field and then type the derive expression using the @FIELD function within the field parentheses.

Note: The @FIELD function is an important tool for deriving multiple fields at the same time. It allows you to refer to the contents of the current field or fields without specifying the exact field name. For instance, a CLEM expression used to apply a log transformation to multiple fields is $\log(@FIELD)$.

Figure 4-43
Deriving multiple fields



The following options are added to the dialog box when you select Multiple mode:

Derive from. Use the Field Chooser to select fields from which to derive new fields. One output field will be generated for each selected field. *Note:* Selected fields do not need to be the same storage type; however, the Derive operation will fail if the condition is not valid for *all* fields.

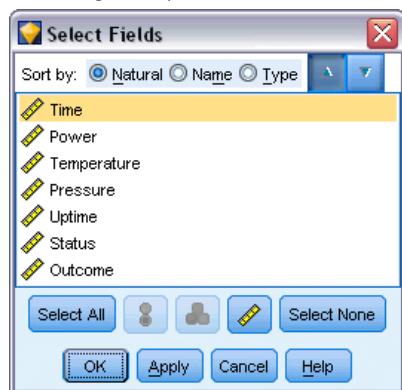
Field name extension. Type the extension that you would like added to the new field name(s). For example, for a new field containing the log of *Current Salary*, you could add the extension *log_* to the field name, producing *log_Current Salary*. Use the radio buttons to choose whether to add the extension as a prefix (at the beginning) or as a suffix (at the end) of the field name. The default name is *DeriveN*, where *N* is the number of Derive nodes that you have created thus far during the current session.

As in the single-mode Derive node, you now need to create an expression to use for deriving a new field. Depending on the type of Derive operation selected, there are a number of options to create a condition. These options are discussed in subsequent topics. To create an expression, you can simply type in the formula field(s) or use the Expression Builder by clicking the calculator button. Remember to use the @FIELD function when referring to manipulations on multiple fields.

Selecting Multiple Fields

For all nodes that perform operations on multiple input fields, such as Derive (multiple mode), Aggregate, Sort, Multiplot, and Time Plot, you can easily select multiple fields using the Select Fields dialog box.

Figure 4-44
Selecting multiple fields



Sort by. You can sort available fields for viewing by selecting one of the following options:

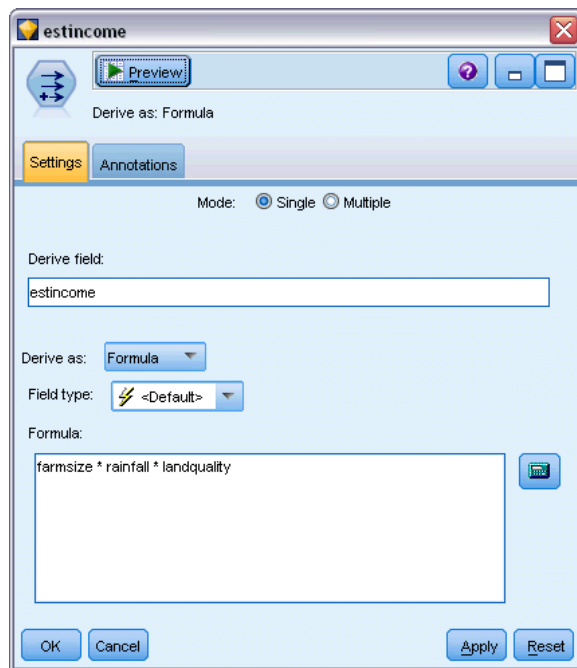
- **Natural.** View the order of fields as they have been passed down the data stream into the current node.
- **Name.** Use alphabetical order to sort fields for viewing.
- **Type.** View fields sorted by their measurement level. This option is useful when selecting fields with a particular measurement level.

Select fields from the list one at a time or use the Shift-click and Ctrl-click methods to select multiple fields. You can also use the buttons below the list to select groups of fields based on their measurement level, or to select or deselect all fields in the table.

Setting Derive Formula Options

Derive Formula nodes create a new field for each record in a dataset based on the results of a CLEM expression. Note that this expression cannot be conditional. To derive values based on a conditional expression, use the flag or conditional type of Derive node.

Figure 4-45
Setting options for a Derive Formula node

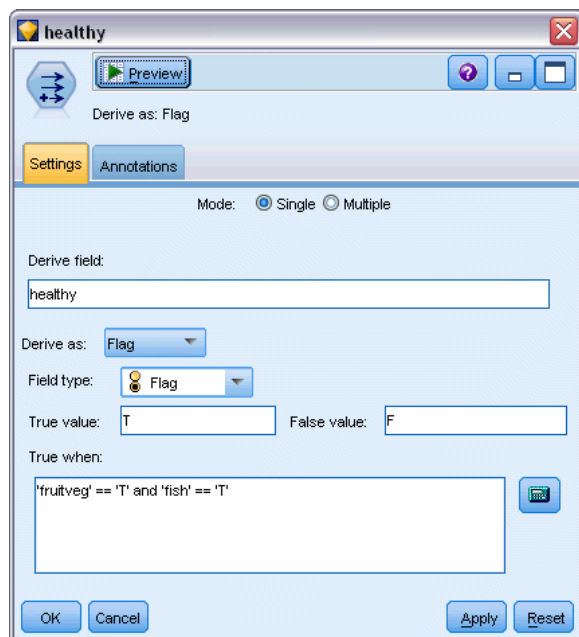


Formula. Specify a formula using the CLEM language to derive a value for the new field.

Setting Derive Flag Options

Derive Flag nodes are used to indicate a specific condition, such as high blood pressure or customer account inactivity. A flag field is created for each record, and when the true condition is met, the flag value for true is added in the field.

Figure 4-46
Deriving a flag field



True value. Specify a value to include in the flag field for records that match the condition specified below. The default is T.

False value. Specify a value to include in the flag field for records that do *not* match the condition specified below. The default is F.

True when. Specify a CLEM condition to evaluate certain values of each record and give the record a true value or a false value (defined above). Note that the true value will be given to records in the case of non-false numeric values.

Note: To return an empty string, you should type opening and closing quotes with nothing between them, such as "". Empty strings are often used, for example, as the false value in order to enable true values to stand out more clearly in a table. Similarly, quotes should be used if you want a string value that would otherwise be treated as a number

Example

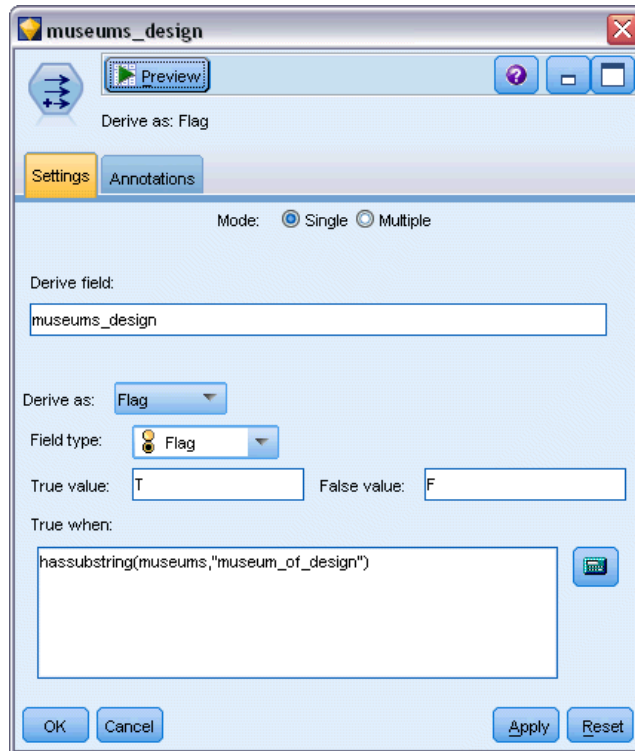
In releases of IBM® SPSS® Modeler prior to 12.0, multiple responses were imported into a single field, with values separate by commas.

museums
museum_of_design,institute_of_textiles_and_fashion
museum_of_design
archeological_museum
\$null\$
national_art_gallery,national_museum_of_science,other

To prepare this data for analysis, you could use the `hassubstring` function to generate a separate flag field for each response with an expression such as:

```
hassubstring(museums,"museum_of_design")
```

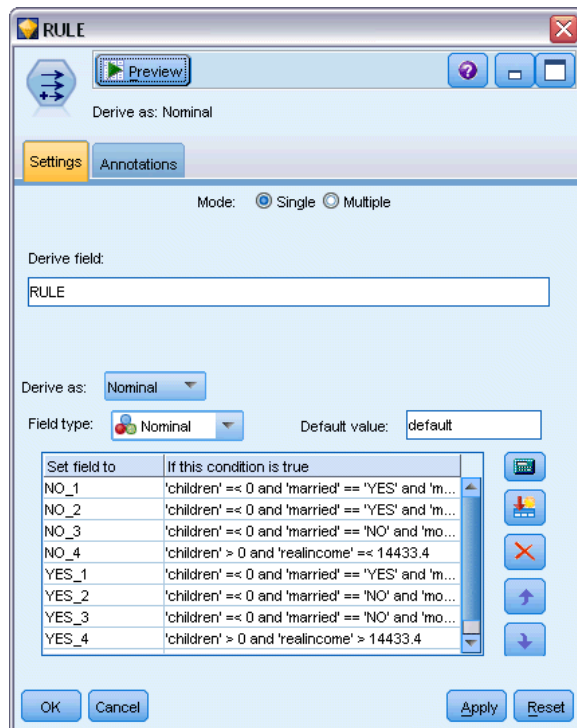
Figure 4-47
Deriving a flag field using the `hassubstring` function



Setting Derive Set Options

Derive Set nodes are used to execute a set of CLEM conditions in order to determine which condition each record satisfies. As a condition is met for each record, a value (indicating which set of conditions was met) will be added to the new, derived field.

Figure 4-48
Using a Derive Set node



Default value. Specify a value to be used in the new field if none of the conditions are met.

Set field to. Specify a value to enter in the new field when a particular condition is met. Each value in the list has an associated condition that you specify in the adjacent column.

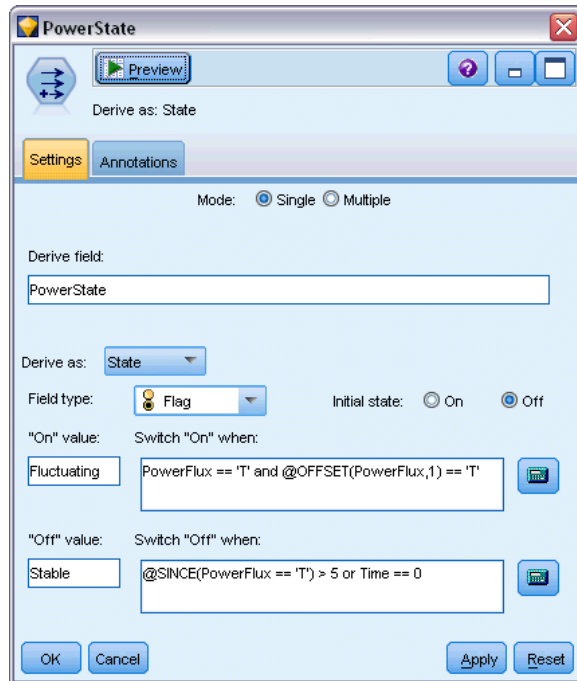
If this condition is true. Specify a condition for each member in the set field to list. Use the Expression Builder to select from available functions and fields. You can use the arrow and delete buttons to reorder or remove conditions.

A condition works by testing the values of a particular field in the dataset. As each condition is tested, the values specified above will be assigned to the new field to indicate which, if any, condition was met. If none of the conditions are met, the default value is used.

Setting Derive State Options

Derive State nodes are somewhat similar to Derive Flag nodes. A Flag node sets values depending on the fulfillment of a *single* condition for the current record, but a Derive State node can change the values of a field depending on how it fulfills *two independent* conditions. This means that the value will change (turn on or off) as each condition is met.

Figure 4-49
Using a Derive State node



Initial state. Select whether to give each record of the new field the On or Off value initially. Note that this value can change as each condition is met.

On value. Specify the value for the new field when the On condition is met.

Switch On when. Specify a CLEM condition that will change the state to On when the condition is true. Click the calculator button to open the Expression Builder.

Off value. Specify the value for the new field when the Off condition is met.

Switch Off when. Specify a CLEM condition that will change the state to Off when the condition is false. Click the calculator button to open the Expression Builder.

Note: To specify an empty string, you should type opening and closing quotes with nothing between them, such as "". Similarly, quotes should be used if you want a string value that would otherwise be treated as a number.

Setting Derive Count Options

A Derive Count node is used to apply a series of conditions to the values of a numeric field in the dataset. As each condition is met, the value of the derived count field is increased by a set increment. This type of Derive node is useful for time series data.

Figure 4-50
Count options in the Derive node dialog box



Initial value. Sets a value used on execution for the new field. The initial value must be a numeric constant. Use the arrow buttons to increase or decrease the value.

Increment when. Specify the CLEM condition that, when met, will change the derived value based on the number specified in Increment by. Click the calculator button to open the Expression Builder.

Increment by. Set the value used to increment the count. You can use either a numeric constant or the result of a CLEM expression.

Reset when. Specify a condition that, when met, will reset the derived value to the initial value. Click the calculator button to open the Expression Builder.

Setting Derive Conditional Options

Derive Conditional nodes use a series of If-Then-Else statements to derive the value of the new field.

Figure 4-51
Using a conditional Derive node

The screenshot shows a dialog box titled "PowerChange". At the top, there is a "Preview" button and a "Derive as: Conditional" label. Below this, there are "Settings" and "Annotations" tabs. The "Mode" is set to "Single". The "Derive field" is "PowerChange". The "Derive as" dropdown is set to "Conditional", and the "Field type" dropdown is set to "<Default>". The "If:" field contains the expression "Time <= 5". The "Then:" field contains "0.0". The "Else:" field contains "@AVE(PowerInc, 5)". At the bottom, there are buttons for "OK", "Cancel", "Apply", and "Reset".

If. Specify a CLEM condition that will be evaluated for each record upon execution. If the condition is true (or non-false, in the case of numbers), the new field is given the value specified below by the Then expression. Click the calculator button to open the Expression Builder.

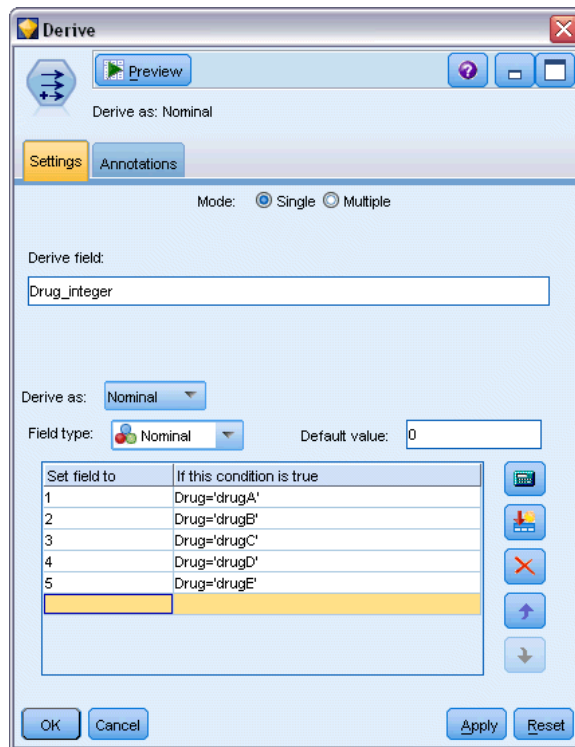
Then. Specify a value or CLEM expression for the new field when the If statement above is true (or non-false). Click the calculator button to open the Expression Builder.

Else. Specify a value or CLEM expression for the new field when the If statement above is false. Click the calculator button to open the Expression Builder.

Recoding Values with the Derive Node

Derive nodes can also be used to recode values, for example by converting a string field with categorical values to a numeric nominal (set) field.

Figure 4-52
Recoding string values

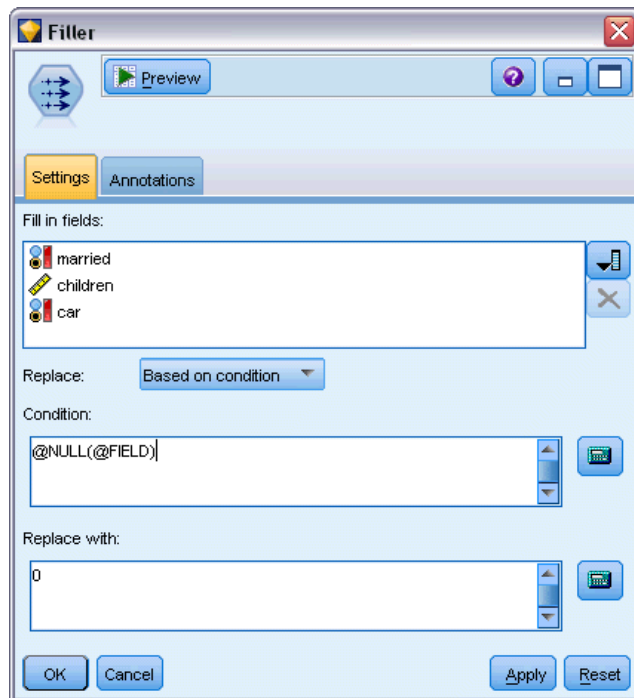


- ▶ For Derive As, select the type of field (Nominal, Flag, etc.) as appropriate.
- ▶ Specify the conditions for recoding values. For example, you could set the value to 1 if Drug='drugA', 2 if Drug='drugB', and so on.

Filler Node

Filler nodes are used to replace field values and change storage. You can choose to replace values based on a specified CLEM condition, such as @BLANK(FIELD). Alternatively, you can choose to replace all blanks or null values with a specific value. Filler nodes are often used in conjunction with the Type node to replace missing values. For example, you can fill blanks with the mean value of a field by specifying an expression such as @GLOBAL_MEAN. This expression will fill all blanks with the mean value as calculated by a Set Globals node.

Figure 4-53
Filler node dialog box



Fill in fields. Using the Field Chooser (button to the right of the text field), select fields from the dataset whose values will be examined and replaced. The default behavior is to replace values depending on the Condition and Replace with expressions specified below. You can also select an alternative method of replacement using the Replace options below.

Note: When selecting multiple fields to replace with a user-defined value, it is important that the field types are similar (all numeric or all symbolic).

Replace. Select to replace the values of the selected field(s) using one of the following methods:

- **Based on condition.** This option activates the Condition field and Expression Builder for you to create an expression used as a condition for replacement with the value specified.
- **Always.** Replaces all values of the selected field. For example, you could use this option to convert the storage of income to a string using the following CLEM expression: (to_string(income)).
- **Blank values.** Replaces all user-specified blank values in the selected field. The standard condition @BLANK(@FIELD) is used to select blanks. *Note:* You can define blanks using the Types tab of the source node or with a Type node.
- **Null values.** Replaces all system null values in the selected field. The standard condition @NULL(@FIELD) is used to select nulls.
- **Blank and null values.** Replaces both blank values and system nulls in the selected field. This option is useful when you are unsure whether or not nulls have been defined as missing values.

Condition. This option is available when you have selected the Based on condition option. Use this text box to specify a CLEM expression for evaluating the selected fields. Click the calculator button to open the Expression Builder.

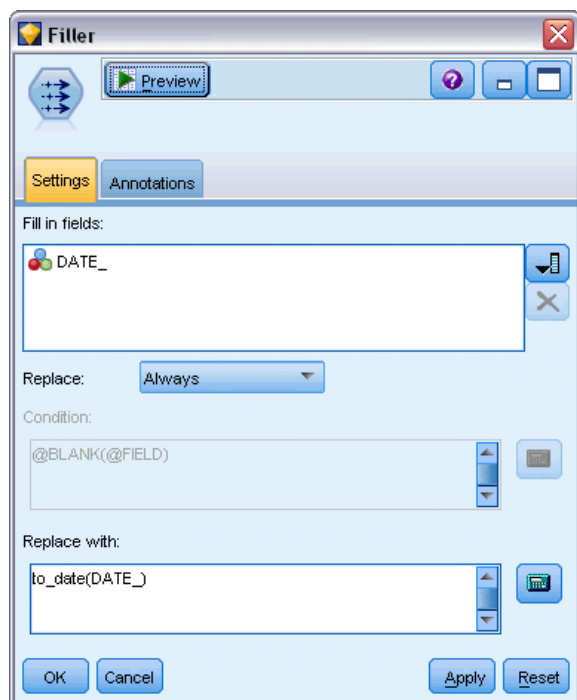
Replace with. Specify a CLEM expression to give a new value to the selected fields. You can also replace the value with a null value by typing undef in the text box. Click the calculator button to open the Expression Builder.

Note: When the field(s) selected are string, you should replace them with a string value. Using the default 0 or another numeric value as the replacement value for string fields will result in an error.

Storage Conversion Using the Filler Node

Using the Replace condition of a Filler node, you can easily convert the field storage for single or multiple fields. For example, using the conversion function `to_integer`, you could convert `income` from a string to an integer using the following CLEM expression: `to_integer(income)`.

Figure 4-54
Using a Filler node to convert field storage



You can view available conversion functions and automatically create a CLEM expression using the Expression Builder. From the Functions drop-down list, select Conversion to view a list of storage conversion functions. The following conversion functions are available:

- `to_integer(ITEM)`
- `to_real(ITEM)`
- `to_number(ITEM)`
- `to_string(ITEM)`

- `to_time(ITEM)`
- `to_timestamp(ITEM)`
- `to_date(ITEM)`
- `to_datetime(ITEM)`

Converting date and time values. Note that conversion functions (and any other functions that require a specific type of input such as a date or time value) depend on the current formats specified in the stream options dialog box. For example if you want to convert a string field with values *Jan 2003*, *Feb 2003*, etc. to date storage, select `MON YYYY` as the default date format for the stream.

Conversion functions are also available from the Derive node, for temporary conversion during a derive calculation. You can also use the Derive node to perform other manipulations such as recoding string fields with categorical values. For more information, see the topic [Recoding Values with the Derive Node](#) on p. 150.

Anonymize Node

The Anonymize node enables you to disguise field names, field values, or both when working with data that are to be included in a model downstream of the node. In this way, the generated model can be freely distributed (for example, to Technical Support) with no danger that unauthorized users will be able to view confidential data, such as employee records or patients' medical records.

Depending on where you place the Anonymize node in the stream, you may need to make changes to other nodes. For example, if you insert an Anonymize node upstream from a Select node, the selection criteria in the Select node will need to be changed if they are acting on values that have now become anonymized.

The method to be used for anonymizing depends on various factors. For field names and all field values except Continuous measurement levels, the data are replaced by a string of the form:

prefix_Sn

where *prefix_* is either a user-specified string or the default string `anon_`, and *n* is an integer value that starts at 0 and is incremented for each unique value (for example, `anon_S0`, `anon_S1`, etc.).

Field values of type Continuous must be transformed because numeric ranges deal with integer or real values rather than strings. As such, they can be anonymized only by transforming the range into a different range, thus disguising the original data. Transformation of a value *x* in the range is performed in the following way:

$$A*(x + B)$$

where:

A is a scale factor, which must be greater than 0.

B is a translation offset to be added to the values.

Example

In the case of a field *AGE* where the scale factor *A* is set to 7 and the translation offset *B* is set to 3, the values for *AGE* are transformed into:

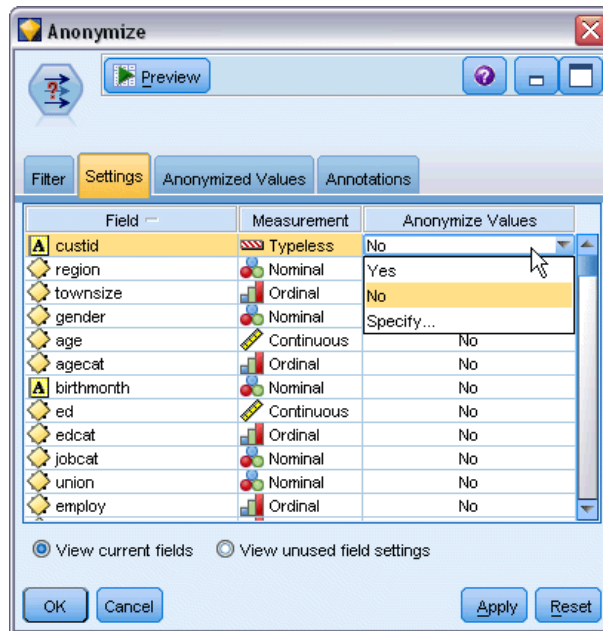
$$7*(AGE + 3)$$

Setting Options for the Anonymize Node

Here you can choose which fields are to have their values disguised further downstream.

Note that the data fields must be instantiated upstream from the Anonymize node before anonymize operations can be performed. You can instantiate the data by clicking the Read Values button on a Type node or on the Types tab of a source node.

Figure 4-55
Setting anonymize options



Field. Lists the fields in the current dataset. If any field names have already been anonymized, the anonymized names are shown here.

Measurement. The measurement level of the field.

Anonymize Values. Select one or more fields, click this column, and choose Yes to anonymize the field value using the default prefix anon_; choose Specify to display a dialog box in which you can either enter your own prefix or, in the case of field values of type *Continuous*, specify whether the transformation of field values is to use random or user-specified values. Note that *Continuous* and non-*Continuous* field types cannot be specified in the same operation; you must do this separately for each type of field.

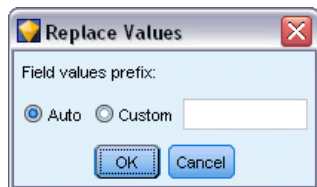
View current fields. Select to view fields for datasets actively connected to the Anonymize node. This option is selected by default.

View unused field settings. Select to view fields for datasets that were once but are no longer connected to the node. This option is useful when copying nodes from one stream to another or when saving and reloading nodes.

Specifying How Field Values Will Be Anonymized

The Replace Values dialog box lets you choose whether to use the default prefix for anonymized field values or to use a custom prefix. Clicking OK in this dialog box changes the setting of Anonymize Values on the Settings tab to Yes for the selected field or fields.

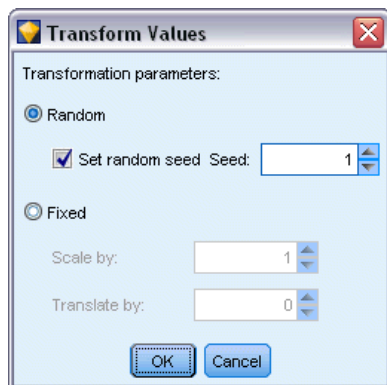
Figure 4-56
Replace Values dialog box



Field values prefix. The default prefix for anonymized field values is anon_; choose Custom and enter your own prefix if you want a different one.

The Transform Values dialog box is displayed only for fields of type Continuous and allows you to specify whether the transformation of field values is to use random or user-specified values.

Figure 4-57
Transform Values dialog box



Random. Choose this option to use random values for the transformation. Set random seed is selected by default; specify a value in the Seed field, or use the default value.

Fixed. Choose this option to specify your own values for the transformation.

- **Scale by.** The number by which field values will be multiplied in the transformation. Minimum value is 1; maximum is normally 10, but this may be lowered to avoid overflow.
- **Translate by.** The number that will be added to field values in the transformation. Minimum value is 0; maximum is normally 1000, but this may be lowered to avoid overflow.

Anonymizing Field Values

Fields that have been selected for anonymization on the Settings tab have their values anonymized:

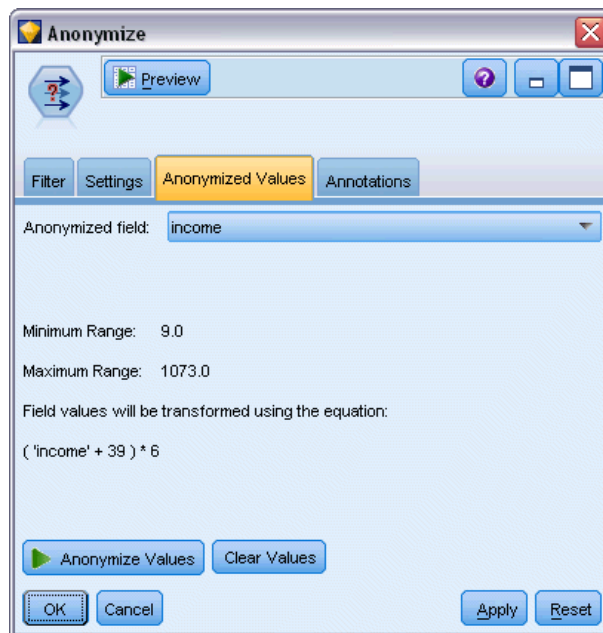
- When you run the stream containing the Anonymize node
- When you preview the values

To preview the values, click the Anonymize Values button on the Anonymized Values tab. Next, select a field name from the drop-down list.

If the measurement level is Continuous, the display shows the:

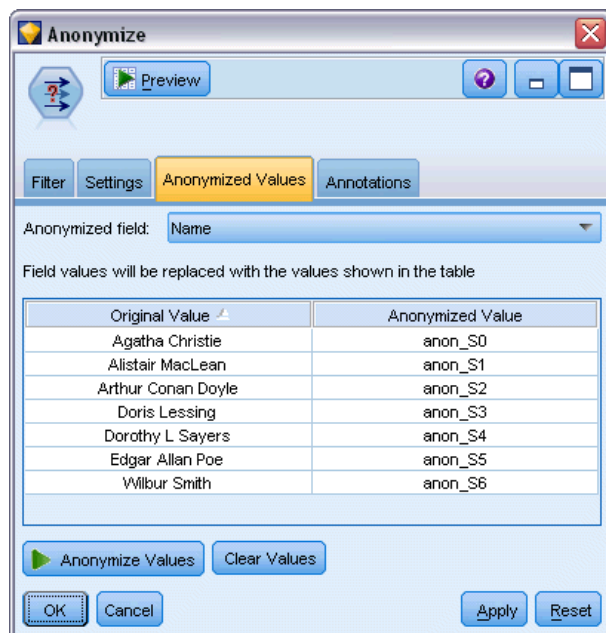
- Minimum and maximum values of the original range
- Equation used to transform the values

Figure 4-58
Anonymizing field values



If the measurement level is anything other than Continuous, the screen displays the original and anonymized values for that field.

Figure 4-59
Anonymizing field values



If the display appears with a yellow background, this indicates that either the setting for the selected field has changed since the last time the values were anonymized or that changes have been made to the data upstream of the Anonymize node such that the anonymized values may no longer be correct. The current set of values is displayed; click the Anonymize Values button again to generate a new set of values according to the current setting.

Anonymize Values. Creates anonymized values for the selected field and displays them in the table. If you are using random seeding for a field of type Continuous, clicking this button repeatedly creates a different set of values each time.

Clear Values. Clears the original and anonymized values from the table.

Reclassify Node

The Reclassify node enables the transformation from one set of categorical values to another. Reclassification is useful for collapsing categories or regrouping data for analysis. For example, you could reclassify the values for *Product* into three groups, such as *Kitchenware*, *Bath and Linens*, and *Appliances*. Often, this operation is performed directly from a Distribution node by grouping values and generating a Reclassify node. For more information, see the topic [Using a Distribution Node](#) in Chapter 5 on p. 239.

Reclassification can be performed for one or more symbolic fields. You can also choose to substitute the new values for the existing field or generate a new field.

Before using a Reclassify node, consider whether another Field Operations node is more appropriate for the task at hand:

- To transform numeric ranges into sets using an automatic method, such as ranks or percentiles, you should use a Binning node. For more information, see the topic [Binning Node](#) on p. 162.

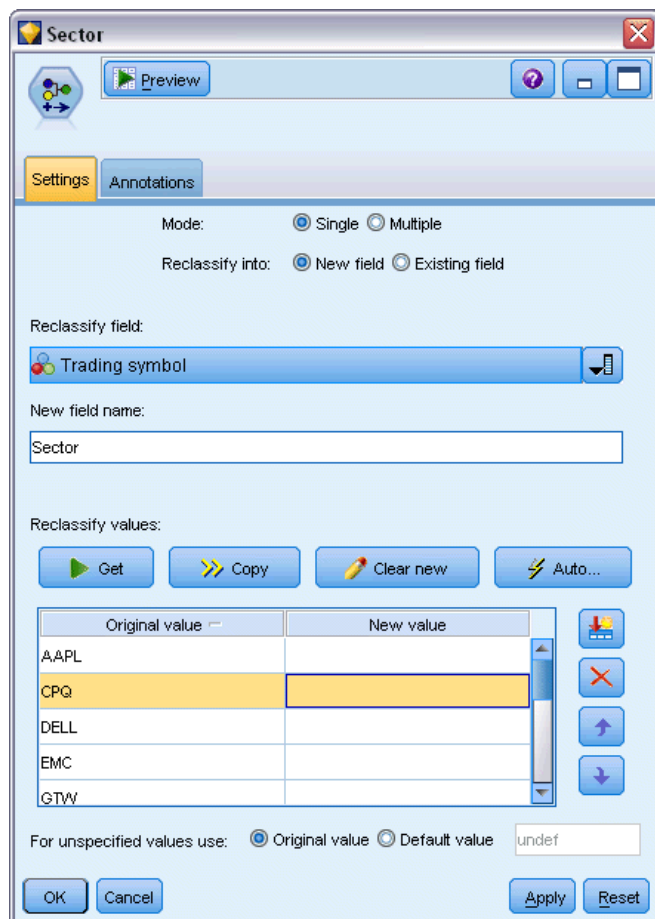
- To classify numeric ranges into sets manually, you should use a Derive node. For example, if you want to collapse salary values into specific salary range categories, you should use a Derive node to define each category manually.
- To create one or more flag fields based on the values of a categorical field, such as *Mortgage_type*, you should use a Set to Flag node.
- To convert a categorical field to numeric storage, you can use a Derive node. For example, you could convert *No* and *Yes* values to 0 and 1, respectively. For more information, see the topic [Recoding Values with the Derive Node](#) on p. 150.

Setting Options for the Reclassify Node

There are three steps to using the Reclassify node:

- ▶ First, select whether you want to reclassify multiple fields or a single field.
- ▶ Next, choose whether to recode into the existing field or create a new field.
- ▶ Then, use the dynamic options in the Reclassify node dialog box to map sets as desired.

Figure 4-60
Reclassify node dialog box



Mode. Select **Single** to reclassify the categories for one field. Select **Multiple** to activate options enabling the transformation of more than one field at a time.

Reclassify into. Select **New field** to keep the original nominal field and derive an additional field containing the reclassified values. Select **Existing field** to overwrite the values in the original field with the new classifications. This is essentially a “fill” operation.

Once you have specified mode and replacement options, you must select the transformation field and specify the new classification values using the dynamic options on the bottom half of the dialog box. These options vary depending on the mode you have selected above.

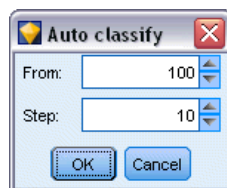
Reclassify field(s). Use the **Field Chooser** button on the right to select one (**Single mode**) or more (**Multiple mode**) categorical fields.

New field name. Specify a name for the new nominal field containing recoded values. This option is available only in **Single mode** when **New field** is selected above. When **Existing field** is selected, the original field name is retained. When working in **Multiple mode**, this option is replaced with controls for specifying an extension added to each new field. For more information, see the topic [Reclassifying Multiple Fields](#) on p. 161.

Reclassify values. This table enables a clear mapping from old set values to those you specify here.

- **Original value.** This column lists existing values for the select field(s).
 - **New value.** Use this column to type new category values or select one from the drop-down list. When you automatically generate a **Reclassify** node using values from a **Distribution chart**, these values are included in the drop-down list. This allows you to quickly map existing values to a known set of values. For example, healthcare organizations sometimes group diagnoses differently based upon network or locale. After a merger or acquisition, all parties will be required to reclassify new or even existing data in a consistent fashion. Rather than manually typing each target value from a lengthy list, you can read the master list of values in to IBM® SPSS® Modeler, run a **Distribution chart** for the *Diagnosis* field, and generate a **Reclassify (values)** node for this field directly from the chart. This process will make all of the target **Diagnosis** values available from the **New Values** drop-down list.
- ▶ Click **Get** to read original values for one or more fields selected above.
 - ▶ Click **Copy** to paste original values over to the *New value* column for fields that have not been mapped yet. The unmapped original values are added to the drop-down list.
 - ▶ Click **Clear new** to erase all specifications in the *New value* column. *Note:* This option does not erase the values from the drop-down list.
 - ▶ Click **Auto** to automatically generate consecutive integers for each of the original values. Only integer values (no real values, such as 1.5, 2.5, and so on) can be generated.

Figure 4-61
Auto-classification dialog box



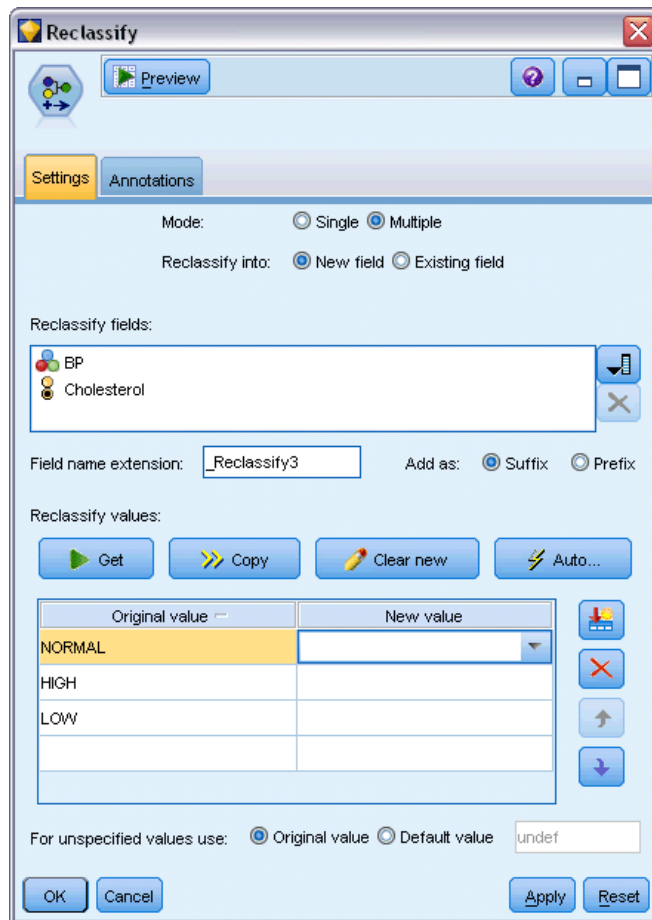
For example, you can automatically generate consecutive product ID numbers for product names or course numbers for university class offerings. This functionality corresponds to the Automatic Recode transformation for sets in IBM® SPSS® Statistics.

For unspecified values use. This option is used for filling unspecified values in the new field. You can either choose to keep the original value by selecting Original value or specify a default value.

Reclassifying Multiple Fields

To map category values for more than one field at a time, set the mode to Multiple. This enables new settings in the Reclassify dialog box, which are described below.

Figure 4-62
Dynamic dialog box options for reclassifying multiple fields



Reclassify fields. Use the Field Chooser button on the right to select the fields that you want to transform. Using the Field Chooser, you can select all fields at once or fields of a similar type, such as nominal or flag.

Field name extension. When recoding multiple fields simultaneously, it is more efficient to specify a common extension added to all new fields rather than individual field names. Specify an extension such as `_recode` and select whether to append or prepend this extension to the original field names.

Storage and Measurement Level for Reclassified Fields

The Reclassify node always creates a nominal field from the recode operation. In some cases, this may change the measurement level of the field when using the Existing field reclassification mode.

The new field's storage (how data are *stored* rather than how they are *used*) is calculated based on the following Settings tab options:

- If unspecified values are set to use a default value, the storage type is determined by examining both the new values as well as the default value and determining the appropriate storage. For example, if all values can be parsed as integers, the field will have the integer storage type.
- If unspecified values are set to use the original values, the storage type is based on the storage of the original field. If all of the values can be parsed as the storage of the original field, then that storage is preserved; otherwise, the storage is determined by finding the most appropriate storage type encompassing both old and new values. For example, reclassifying an integer set { 1, 2, 3, 4, 5 } with the reclassification 4 => 0, 5 => 0 generates a new integer set { 1, 2, 3, 0 }, whereas with the reclassification 4 => "Over 3", 5 => "Over 3" will generate the string set { "1", "2", "3", "Over 3" }.

Note: If the original type was uninstantiated, the new type will be also be uninstantiated.

Binning Node

The Binning node enables you to automatically create new nominal fields based on the values of one or more existing continuous (numeric range) fields. For example, you can transform a continuous income field into a new categorical field containing income groups of equal width, or as deviations from the mean. Alternatively, you can select a categorical "supervisor" field in order to preserve the strength of the original association between the two fields.

Binning can be useful for a number of reasons, including:

- **Algorithm requirements.** Certain algorithms, such as Naive Bayes, Logistic Regression, require categorical inputs.
- **Performance.** Algorithms such as multinomial logistic may perform better if the number of distinct values of input fields is reduced. For example, use the median or mean value for each bin rather than using the original values.
- **Data Privacy.** Sensitive personal information, such as salaries, may be reported in ranges rather than actual salary figures in order to protect privacy.

A number of binning methods are available, Once you have created bins for the new field, you can generate a Derive node based on the cut points.

Before using a Binning node, consider whether another technique is more appropriate for the task at hand:

- To manually specify cut points for categories, such as specific predefined salary ranges, use a Derive node. For more information, see the topic [Derive Node](#) on p. 139.
- To create new categories for existing sets, use a Reclassify node. For more information, see the topic [Reclassify Node](#) on p. 158.

Missing Value Handling

The Binning node handles missing values in the following ways:

- **User-specified blanks.** Missing values specified as blanks are included during the transformation. For example, if you designated –99 to indicate a blank value using the Type node, this value will be included in the binning process. To ignore blanks during binning, you should use a Filler node to replace the blank values with the system null value.
- **System-missing values (\$null\$).** Null values are ignored during the binning transformation and remain nulls after the transformation.

The Settings tab provides options for available techniques. The View tab displays cut points established for data previously run through the node.

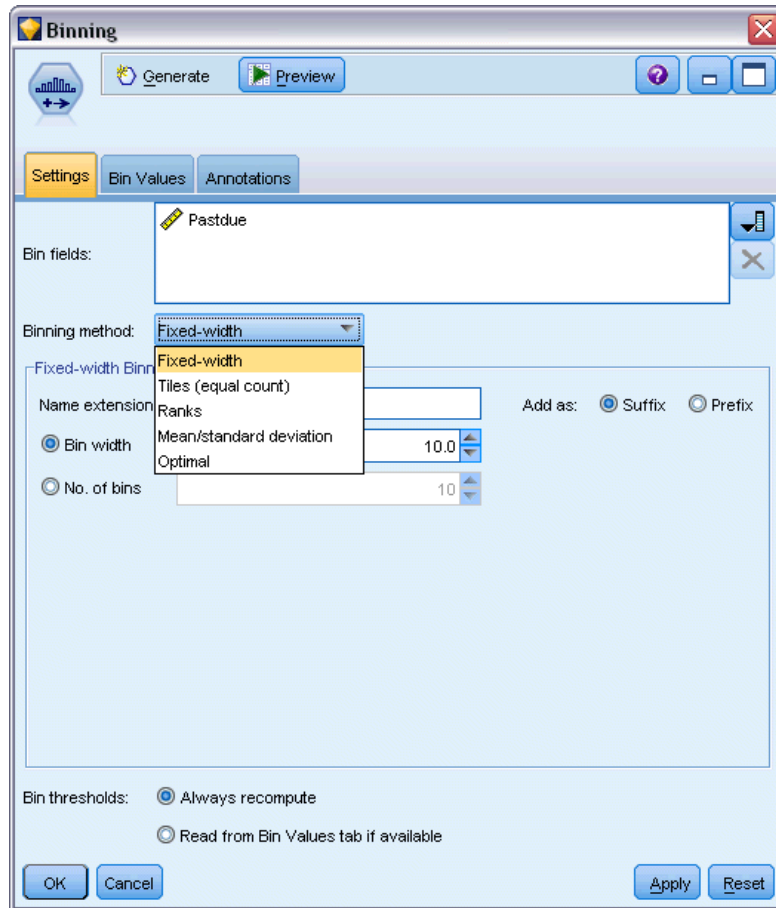
Setting Options for the Binning Node

Using the Binning node, you can automatically generate bins (categories) using the following techniques:

- Fixed-width binning
- Tiles (equal count or sum)
- Mean and standard deviation
- Ranks
- Optimized relative to a categorical “supervisor” field

The bottom half of the dialog box changes dynamically depending on the binning method you select.

Figure 4-63
Binning node dialog box, Settings tab



Bin fields. Continuous (numeric range) fields pending transformation are displayed here. The Binning node enables you to bin multiple fields simultaneously. Add or remove fields using the buttons on the right.

Binning method. Select the method used to determine cut points for new field bins (categories).

Bin thresholds. Specify whether cut points and bin allocations are always recomputed when the node is executed, or that they are only computed as necessary (for example, when new data has been added). If you select Read from Bin Values tab if available, you can edit the upper and lower cut points for the different bins on the Bin Values tab.

The following topics discuss options for the available methods of binning.

Fixed-Width Bins

When you choose Fixed-width as the binning method, a new set of options is displayed in the dialog box.

Figure 4-64
Binning node dialog box (Settings tab) with options for fixed-width bins

Fixed-width Binning

Name extension: Add as: Suffix Prefix

Bin width

No. of bins

Name extension. Specify an extension to use for the generated field(s). *_BIN* is the default extension. You may also specify whether the extension is added to the start (Prefix) or end (Suffix) of the field name. For example, you could generate a new field called *income_BIN*.

Bin width. Specify a value (integer or real) used to calculate the “width” of the bin. For example, you can use the default value, 10, to bin the field *Age*. Since *Age* has a range from 18–65, the generated bins would be the following:

Table 4-1
Bins for *Age* with range 18–65

Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6
>=13 to <23	>=23 to <33	>=33 to <43	>=43 to <53	>=53 to <63	>=63 to <73

The start of bin intervals is calculated using the lowest scanned value minus half of the bin width (as specified). For example, in the bins shown above, 13 is used to start the intervals according to the following calculation: $18 [lowest\ data\ value] - 5 [0.5 \times (Bin\ width\ of\ 10)] = 13$.

No. of bins. Use this option to specify an integer used to determine the number of fixed-width bins (categories) for the new field(s).

Once you have executed the Binning node in a stream, you can view the bin thresholds generated by clicking the Preview tab in the Binning node dialog box. For more information, see the topic [Previewing the Generated Bins](#) on p. 171.

Tiles (Equal Count or Sum)

The tile binning method creates nominal fields that can be used to split scanned records into percentile groups (or quartiles, deciles, and so on) so that each group contains the same number of records, or the sum of the values in each group is equal. Records are ranked in ascending order based on the value of the specified bin field, so that records with the lowest values for the selected bin variable are assigned a rank of 1, the next set of records are ranked 2, and so on. The threshold values for each bin are generated automatically based on the data and tiling method used.

Figure 4-65
Binning node dialog box (Settings tab) with options for equal count bins

Tile name extension. Specify an extension used for field(s) generated using standard p-tiles. The default extension is `_TILE` plus N , where N is the tile number. You may also specify whether the extension is added to the start (Prefix) or end (Suffix) of the field name. For example, you could generate a new field called `income_BIN4`.

Custom tile extension. Specify an extension used for a custom tile range. The default is `_TILEN`. Note that N in this case will *not* be replaced by the custom number.

Available p-tiles are:

- **Quartile.** Generate 4 bins, each containing 25% of the cases.
- **Quintile.** Generate 5 bins, each containing 20% of the cases.
- **Decile.** Generate 10 bins, each containing 10% of the cases.
- **Vingtile.** Generate 20 bins, each containing 5% of the cases.
- **Percentile.** Generate 100 bins, each containing 1% of the cases.
- **Custom N.** Select to specify the number of bins. For example, a value of 3 would produce 3 banded categories (2 cut points), each containing 33.3% of the cases.

Note that if there are fewer discrete values in the data than the number of tiles specified, all tiles will not be used. In such cases, the new distribution is likely to reflect the original distribution of your data.

Tiling method. Specifies the method used to assign records to bins.

- **Record count.** Seeks to assign an equal number of records to each bin.
- **Sum of values.** Seeks to assign records to bins such that the sum of the values in each bin is equal. When targeting sales efforts, for example, this method can be used to assign prospects to decile groups based on value per record, with the highest value prospects in the top bin. For example, a pharmaceutical company might rank physicians into decile groups based on the number of prescriptions they write. While each decile would contain approximately the same number of scripts, the number of individuals contributing those scripts would not be the same, with the individuals who write the most scripts concentrated in decile 10. Note

that this approach assumes that all values are greater than zero, and may yield unexpected results if this is not the case.

Ties. A tie condition results when values on either side of a cut point are identical. For example, if you are assigning deciles and more than 10% of records have the same value for the bin field, then all of them cannot fit into the same bin without forcing the threshold one way or another. Ties can be moved up to the next bin or kept in the current one but must be resolved so that all records with identical values fall into the same bin, even if this causes some bins to have more records than expected. The thresholds of subsequent bins may also be adjusted as a result, causing values to be assigned differently for the same set of numbers based on the method used to resolve ties.

- **Add to next.** Select to move the tie values up to the next bin.
- **Keep in current.** Keeps tie values in the current (lower) bin. This method may result in fewer total bins being created.
- **Assign randomly.** Select to allocate the tie values randomly to a bin. This attempts to keep the number of records in each bin at an equal amount.

Example: Tiling by Record Count

The table below illustrates how simplified field values are ranked as quartiles when tiling by record count. Note the results vary depending on the selected ties option.

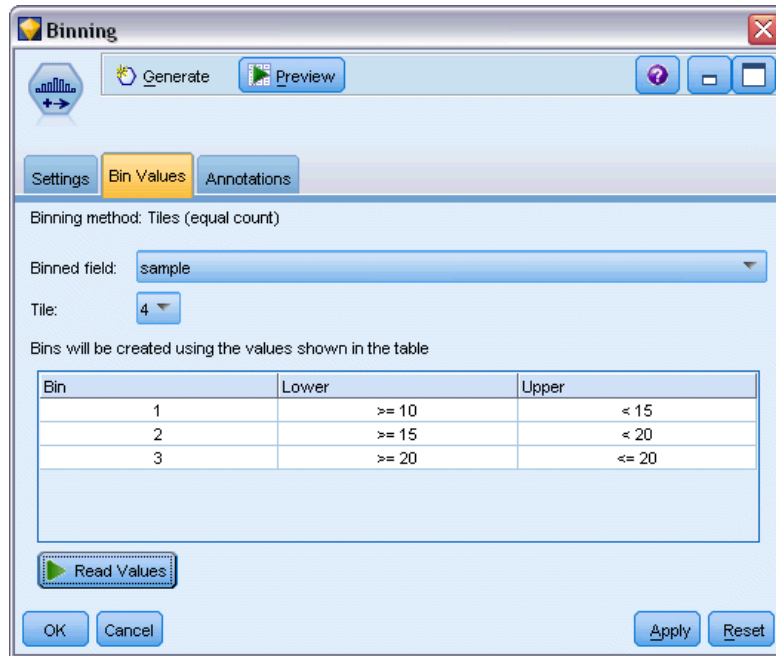
Values	Add to Next	Keep in Current
10	1	1
13	2	1
15	3	2
15	3	2
20	4	3

The number of items per bin is calculated as:

total number of value / number of tiles

In the simplified example above, the desired number of items per bin is 1.25 (5 values / 4 quartiles). The value 13 (being value number 2) straddles the 1.25 desired count threshold and is therefore treated differently depending on the selected ties option. In Add to Next mode, it is added into bin 2. In Keep in Current mode, it is left in bin 1, pushing the range of values for bin 4 outside that of existing data values. As a result, only three bins are created, and the thresholds for each bin are adjusted accordingly.

Figure 4-66
Thresholds for generated bins

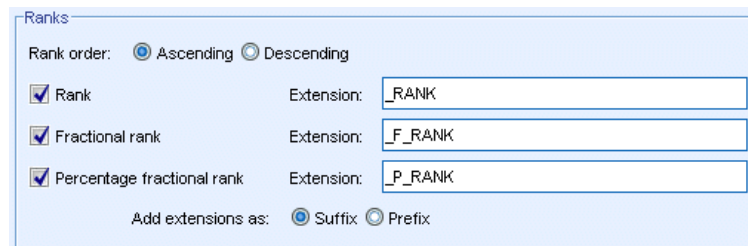


Note: The speed of binning by tiles may benefit from enabling parallel processing.

Rank Cases

When you choose Ranks as the binning method, a new set of options is displayed in the dialog box.

Figure 4-67
Binning node dialog box (Settings tab) with options for ranks



Ranking creates new fields containing ranks, fractional ranks, and percentile values for numeric fields depending on the options specified below.

Rank order. Select Ascending (lowest value is marked 1) or Descending (highest value is marked 1).

Rank. Select to rank cases in ascending or descending order as specified above. The range of values in the new field will be 1– N , where N is the number of discrete values in the original field. Tied values are given the average of their rank.

Fractional rank. Select to rank cases where the value of the new field equals rank divided by the sum of the weights of the nonmissing cases. Fractional ranks fall in the range of 0–1.

Percentage fractional rank. Each rank is divided by the number of records with valid values and multiplied by 100. Percentage fractional ranks fall in the range of 1–100.

Extension. For all rank options, you can create custom extensions and specify whether the extension is added to the start (Prefix) or end (Suffix) of the field name. For example, you could generate a new field called *income_P_RANK*.

Mean/Standard Deviation

When you choose Mean/standard deviation as the binning method, a new set of options is displayed in the dialog box.

Figure 4-68

Binning node dialog box (Settings tab) with options for mean/standard deviation

This method generates one or more new fields with banded categories based on the values of the mean and standard deviation of the distribution of the specified field(s). Select the number of deviations to use below.

Name extension. Specify an extension to use for the generated field(s). *_SDBIN* is the default extension. You may also specify whether the extension is added to the start (Prefix) or end (Suffix) of the field name. For example, you could generate a new field called *income_SDBIN*.

- **+/- 1 standard deviation.** Select to generate three bins.
- **+/- 2 standard deviations.** Select to generate five bins.
- **+/- 3 standard deviations.** Select to generate seven bins.

For example, selecting +/-1 standard deviation results in the three bins as calculated below:

Bin 1	Bin 2	Bin 3
$x < (\text{Mean} - \text{Std. Dev})$	$(\text{Mean} - \text{Std. Dev}) \leq x \leq (\text{Mean} + \text{Std. Dev})$	$x > (\text{Mean} + \text{Std. Dev})$

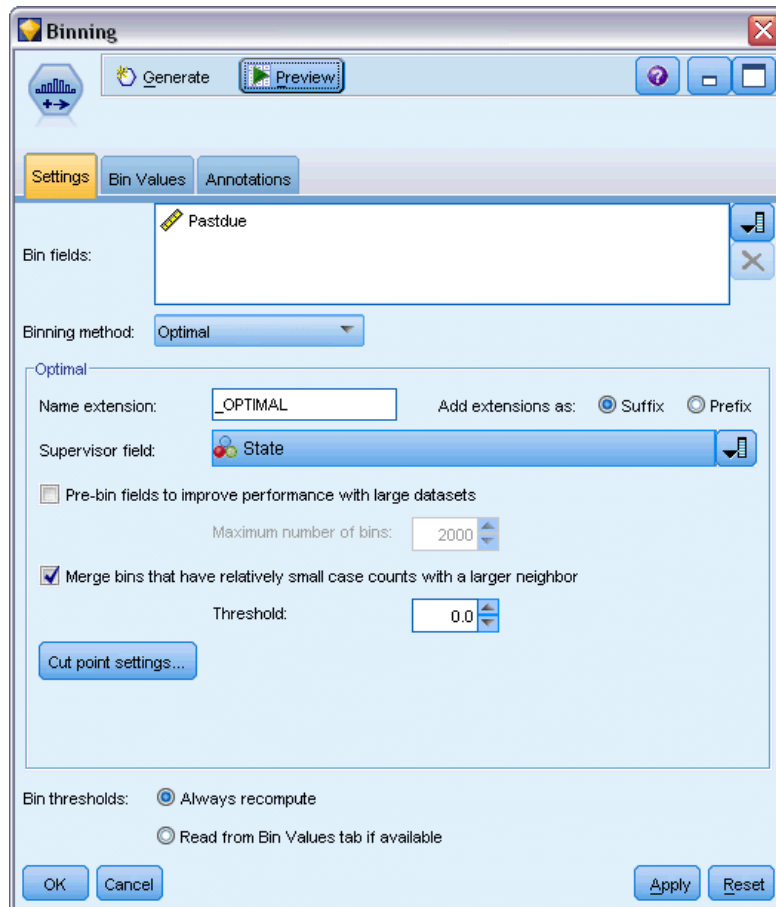
In a normal distribution, 68% of the cases fall within one standard deviation of the mean, 95% within two standard deviations, and 99% within three standard deviations. Note, however, that creating banded categories based on standard deviations may result in some bins being defined outside the actual data range and even outside the range of possible data values (for example, a negative salary range).

Optimal Binning

If the field you want to bin is strongly associated with another categorical field, you can select the categorical field as a “supervisor” field in order to create the bins in such a way as to preserve the strength of the original association between the two fields.

For example, suppose you have used cluster analysis to group states based on delinquency rates for home loans, with the highest rates in the first cluster. In this case, you might choose *Percent past due* and *Percent of foreclosures* as the Bin fields and the cluster membership field generated by the model as the supervisor field.

Figure 4-69
Options for optimal or supervised binning



Name extension. Specify an extension to use for the generated field(s) and whether to add it at the start (Prefix) or end (Suffix) of the field name. For example, you could generate a new field called *pastdue_OPTIMAL* and another called *inforeclosure_OPTIMAL*.

Supervisor field. A categorical field used to construct the bins.

Pre-bin fields to improve performance with large datasets. Indicates if preprocessing should be used to streamline optimal binning. This groups scale values into a large number of bins using a simple unsupervised binning method, represents values within each bin by the mean, and adjusts the case weight accordingly before proceeding with supervised binning. In practical terms, this

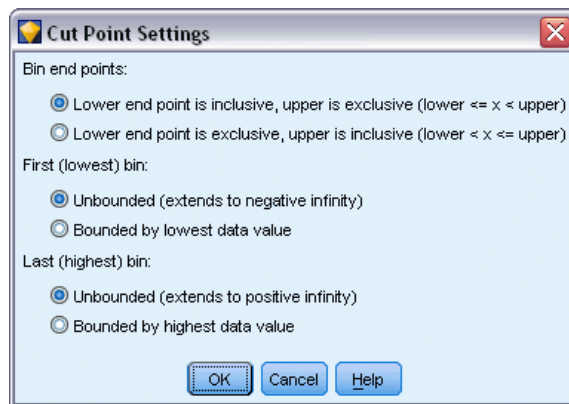
method trades a degree of precision for speed and is recommended for large datasets. You can also specify the maximum number of bins that any variable should end up with after preprocessing when this option is used.

Merge bins that have relatively small case counts with a larger neighbor. If enabled, indicates that a bin is merged if the ratio of its size (number of cases) to that of a neighboring bin is smaller than the specified threshold; note that larger thresholds may result in more merging.

Cut Point Settings

The Cut Point Settings dialog box enables you to specify advanced options for the optimal binning algorithm. These options tell the algorithm how to calculate the bins using the target field.

Figure 4-70
Cut point settings for optimal binning



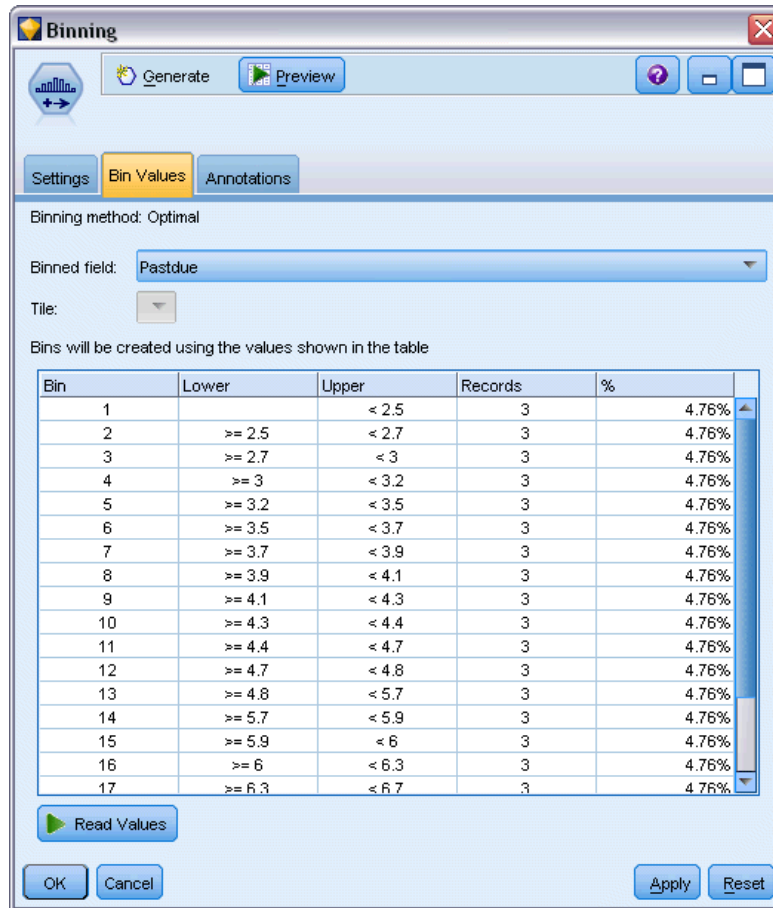
Bin end points. You can specify whether the lower or upper end points should be inclusive ($\text{lower} \leq x$) or exclusive ($\text{lower} < x$).

First and last bins. For both the first and last bin, you can specify whether the bins should be unbounded (extending toward positive or negative infinity) or bounded by the lowest or highest data points.

Previewing the Generated Bins

The Bin Values tab in the Binning node allows you to view the thresholds for generated bins. Using the Generate menu, you can also generate a Derive node that can be used to apply these thresholds from one dataset to another.

Figure 4-71
Binning node dialog box, Bin Values tab



Binned field. Use the drop-down list to select a field for viewing. Field names shown use the original field name for clarity.

Tile. Use the drop-down list to select a tile, such as 10 or 100, for viewing. This option is available only when bins have been generated using the tile method (equal count or sum).

Bin thresholds. Threshold values are shown here for each generated bin, along with the number of records that fall into each bin. For the optimal binning method only, the number of records in each bin is shown as a percentage of the whole. Note that thresholds are not applicable when the rank binning method is used.

Read Values. Reads binned values from the dataset. Note that thresholds will also be overwritten when new data are run through the stream.

Generating a Derive Node

You can use the Generate menu to create a Derive node based on the current thresholds. This is useful for applying established bin thresholds from one set of data to another. Furthermore, once these split points are known, a Derive operation is more efficient (meaning faster) than a Binning operation when working with large datasets.

RFM Analysis Node

The Recency, Frequency, Monetary (RFM) Analysis node enables you to determine quantitatively which customers are likely to be the best ones by examining how recently they last purchased from you (recency), how often they purchased (frequency), and how much they spent over all transactions (monetary).

The reasoning behind RFM analysis is that customers who purchase a product or service once are more likely to purchase again. The categorized customer data is separated into a number of bins, with the binning criteria adjusted as you require. In each of the bins, customers are assigned a score; these scores are then combined to provide an overall RFM score. This score is a representation of the customer's membership in the bins created for each of the RFM parameters. This binned data may be sufficient for your needs, for example, by identifying the most frequent, high-value customers; alternatively, it can be passed on in a stream for further modeling and analysis.

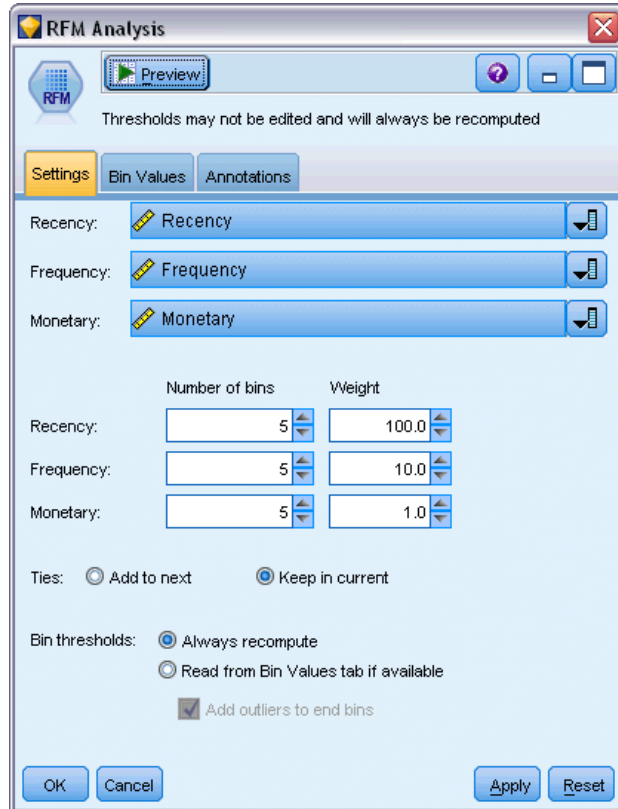
Note, however, that although the ability to analyze and rank RFM scores is a useful tool, you must be aware of certain factors when using it. There may be a temptation to target customers with the highest rankings; however, over-solicitation of these customers could lead to resentment and an actual fall in repeat business. It is also worth remembering that customers with low scores should not be neglected but instead may be cultivated to become better customers. Conversely, high scores alone do not necessarily reflect a good sales prospect, depending on the market. For example, a customer in bin 5 for recency, meaning that they have purchased very recently, may not actually be the best target customer for someone selling expensive, longer-life products such as cars or televisions.

Note: Depending on how your data are stored, you may need to precede the RFM Analysis node with an RFM Aggregate node to transform the data into a usable format. For example, input data must be in customer format, with one row per customer; if the customers' data are in transactional form, use an RFM Aggregate node upstream to derive the recency, frequency, and monetary fields. For more information, see the topic [RFM Aggregate Node](#) in Chapter 3 on p. 67.

The RFM Aggregate and RFM Analysis nodes in IBM® SPSS® Modeler are set up to use independent binning; that is, they rank and bin data on each measure of recency, frequency, and monetary value, without regard to their values or the other two measures.

RFM Analysis Node Settings

Figure 4-72
Setting RFM Analysis options



Recency. Using the Field Chooser (button to the right of the text box), select the recency field. This may be a date, timestamp, or simple number. Note that when a date or timestamp represents the date of the most recent transaction, the highest value is considered the most recent; where a number is specified, it represents the time elapsed since the most recent transaction and the lowest value is considered as the most recent.

Note: If the RFM Analysis node is preceded in the stream by an RFM Aggregate node, the Recency, Frequency, and Monetary fields generated by the RFM Aggregate node should be selected as inputs in the RFM Analysis node.

Frequency. Using the Field Chooser, select the frequency field to be used.

Monetary. Using the Field Chooser, select the monetary field to be used.

Number of bins. For each of the three output types, select the number of bins to be created. The default is 5.

Note: The minimum number of bins is 2, and the maximum is 9.

Weight. By default, the highest importance when calculating scores is given to the recency data, followed by frequency, and then monetary. If required, you can amend the weighting affecting one or several of these to change which is given the highest importance.

The RFM score is calculated as follows: (Recency score x Recency weight) + (Frequency score x Frequency weight) + (Monetary score x Monetary weight).

Ties. Specify how identical (tied) scores are to be binned. The options are:

- **Add to next.** Select to move the tie values up to the next bin.
- **Keep in current.** Keeps tie values in the current (lower) bin. This method may result in fewer total bins being created. (This is the default value.)

Bin thresholds. Specify whether the RFM scores and bin allocations are always recomputed when the node is executed or that they are computed only as necessary (for example, when new data has been added). If you select Read from Bin Values tab if available you can edit the upper and lower cut points for the different bins on the Bin Values tab.

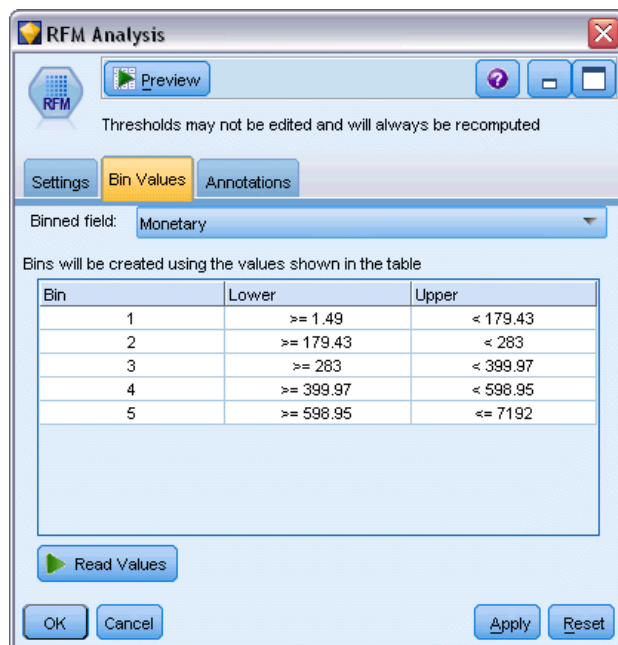
When executed, the RFM Analysis node bins the raw recency, frequency, and monetary fields and adds the following new fields to the dataset:

- Recency score. A rank (bin value) for Recency
- Frequency score. A rank (bin value) for Frequency
- Monetary score. A rank (bin value) for Monetary
- RFM score. The weighted sum of the recency, frequency, and monetary scores.

Add outliers to end bins. If you select this check box, records that lie below the lower bin are added to the lower bin, and records above the highest bin are added to the highest bin—otherwise, they are given a null value. This box is available only if you select Read from Bin Values tab if available.

RFM Analysis Node Binning

Figure 4-73
Setting RFM Analysis bin values



The Bin Values tab allows you to view, and in certain cases amend, the thresholds for generated bins.

Note: You can only amend values on this tab if you select Read from Bin Values tab if available on the Settings tab.

Binned field. Use the drop-down list to select a field for dividing into bins. The available values are those selected on the Settings tab.

Bin values table. Threshold values are shown here for each generated bin. If you select Read from Bin Values tab if available on the Settings tab, you can amend the upper and lower cut points for each bin by double-clicking on the relevant cell.

Read Values. Reads binned values from the dataset and populates the bin values table. Note that if you select Always recompute on the Settings tab, the bin thresholds will be overwritten when new data are run through the stream.

Partition Node

Partition nodes are used to generate a partition field that splits the data into separate subsets or samples for the training, testing, and validation stages of model building. By using one sample to generate the model and a separate sample to test it, you can get a good indication of how well the model will generalize to larger datasets that are similar to the current data.

The Partition node generates a nominal field with the role set to Partition. Alternatively, if an appropriate field already exists in your data, it can be designated as a partition using a Type node. In this case, no separate Partition node is required. Any instantiated nominal field with two or three values can be used as a partition, but flag fields cannot be used. For more information, see the topic [Setting the Field Role](#) on p. 126.

Multiple partition fields can be defined in a stream, but if so, a single partition field must be selected on the Fields tab in each modeling node that uses partitioning. (If only one partition is present, it is automatically used whenever partitioning is enabled.)

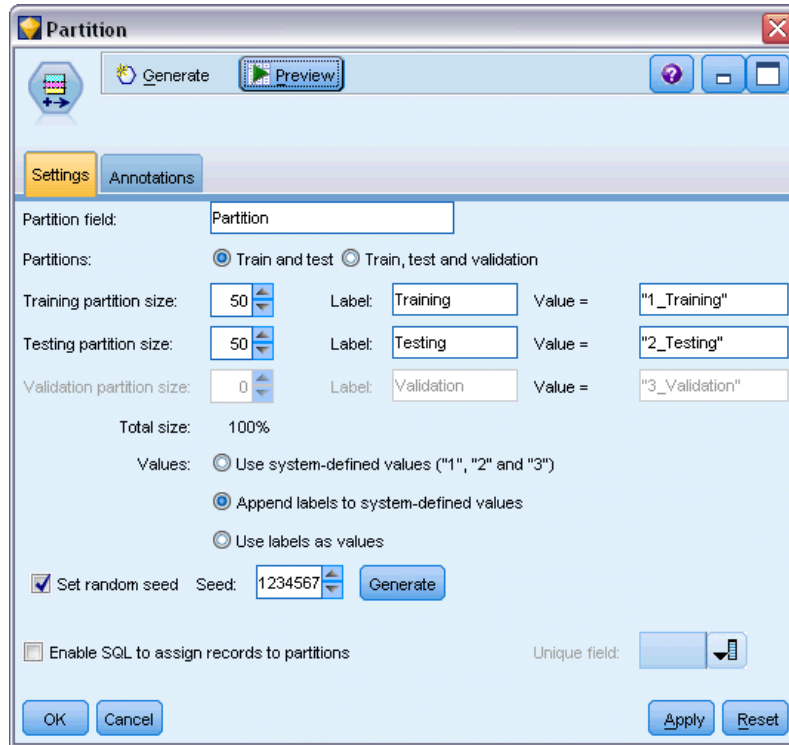
Enabling partitioning. To use the partition in an analysis, partitioning must be enabled on the Model Options tab in the appropriate model-building or analysis node. Deselecting this option makes it possible to disable partitioning without removing the field.

To create a partition field based on some other criterion such as a date range or location, you can also use a Derive node. For more information, see the topic [Derive Node](#) on p. 139.

Example. When building an RFM stream to identify recent customers who have positively responded to previous marketing campaigns, the marketing department of a sales company uses a Partition node to split the data into training and test partitions.

Partition Node Options

Figure 4-74
Partition node dialog box, Settings tab



Partition field. Specifies the name of the field created by the node.

Partitions. You can partition the data into two samples (train and test) or three (train, test, and validation).

- **Train and test.** Partitions the data into two samples, allowing you to train the model with one sample and test with another.
- **Train, test, and validation.** Partitions the data into three samples, allowing you to train the model with one sample, test and refine the model using a second sample, and validate your results with a third. This reduces the size of each partition accordingly, however, and may be most suitable when working with a very large dataset.

Partition size. Specifies the relative size of each partition. If the sum of the partition sizes is less than 100%, then the records not included in a partition will be discarded. For example, if a user has 10 million records and has specified partition sizes of 5% training and 10% testing, after running the node, there should be roughly 500,000 training and one million testing records, with the remainder having been discarded.

Values. Specifies the values used to represent each partition sample in the data.

- **Use system-defined values ("1," "2," and "3").** Uses an integer to represent each partition; for example, all records that fall into the training sample have a value of 1 for the partition field. This ensures the data will be portable between locales and that if the partition field is reinstated elsewhere (for example, reading the data back from a database), the sort order

is preserved (so that 1 will still represent the training partition). However, the values do require some interpretation.

- **Append labels to system-defined values.** Combines the integer with a label; for example, training partition records have a value of *I_Training*. This makes it possible for someone looking at the data to identify which value is which, and it preserves sort order. However, values are specific to a given locale.
- **Use labels as values.** Uses the label with no integer; for example, *Training*. This allows you to specify the values by editing the labels. However, it makes the data locale-specific, and instantiation of a partition column will put the values in their natural sort order, which may not correspond to their “semantic” order.

Set random seed. When sampling or partitioning records based on a random percentage, this option allows you to duplicate the same results in another session. By specifying the starting value used by the random number generator, you can ensure the same records are assigned each time the node is executed. Enter the desired seed value, or click the **Generate** button to automatically generate a random value. If this option is not selected, a different sample will be generated each time the node is executed.

Note: When using the **Set random seed** option with records read from a database, a **Sort** node may be required prior to sampling in order to ensure the same result each time the node is executed. This is because the random seed depends on the order of records, which is not guaranteed to stay the same in a relational database. For more information, see the topic [Sort Node](#) in Chapter 3 on p. 69.

Enable SQL to assign records to partitions. (For Tier 1 databases only) Check this box to use SQL pushback to assign records to partitions. From the **Unique field** drop-down, choose a field with unique values (such as an ID field) to ensure that records are assigned in a random but repeatable way.

Database tiers are explained in the description of the Database source node. For more information, see the topic [Database Source Node](#) in Chapter 2 on p. 11.

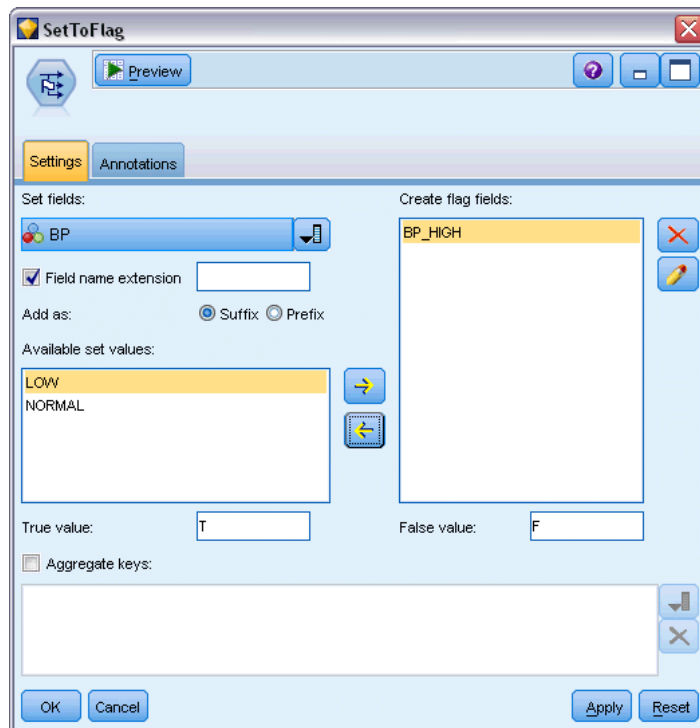
Generating Select Nodes

Using the **Generate** menu in the Partition node, you can automatically generate a Select node for each partition. For example, you could select all records in the training partition to obtain further evaluation or analyses using only this partition.

Set to Flag Node

The **Set to Flag** node is used to derive flag fields based on the categorical values defined for one or more nominal fields. For example, your dataset might contain a nominal field, *BP* (blood pressure), with the values *High*, *Normal*, and *Low*. For easier data manipulation, you might create a flag field for high blood pressure, which indicates whether or not the patient has high blood pressure.

Figure 4-75
Creating a flag field for high blood pressure



Setting Options for the Set to Flag Node

Set fields. Lists all data fields with a measurement level of *Nominal* (set). Select one from the list to display the values in the set. You can choose from these values to create a flag field. Note that data must be fully instantiated using an upstream source or Type node before you can see the available nominal fields (and their values). For more information, see the topic [Type Node](#) on p. 113.

Field name extension. Select to enable controls for specifying an extension that will be added as a suffix or prefix to the new flag field. By default, new field names are automatically created by combining the original field name with the field value into a label, such as *Fieldname_fieldvalue*.

Available set values. Values in the set selected above are displayed here. Select one or more values for which you want to generate flags. For example, if the values in a field called *blood_pressure* are *High*, *Medium*, and *Low*, you can select *High* and add it to the list on the right. This will create a field with a flag for records with a value indicating high blood pressure.

Create flag fields. The newly created flag fields are listed here. You can specify options for naming the new field using the field name extension controls.

True value. Specify the true value used by the node when setting a flag. By default, this value is T.

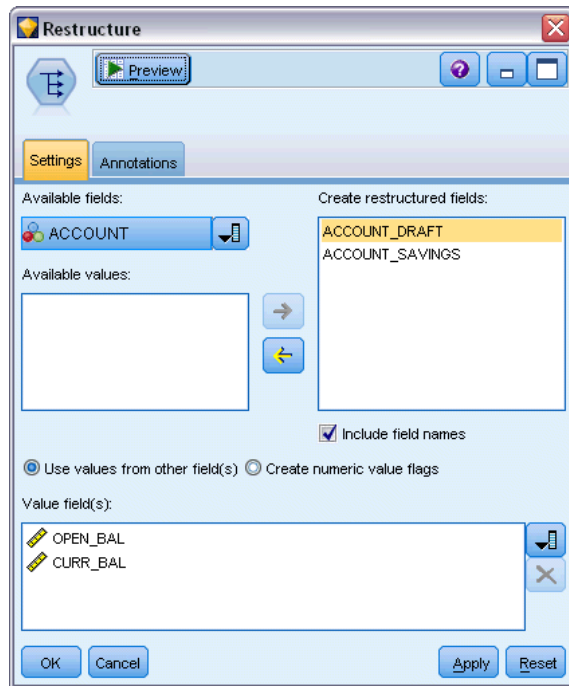
False value. Specify the false value used by the node when setting a flag. By default, this value is F.

Aggregate keys. Select to group records together based on key fields specified below. When Aggregate keys is selected, all flag fields in a group will be “turned on” if *any* record was set to true. Use the Field Chooser to specify which key fields will be used to aggregate records.

Restructure Node

The Restructure node can be used to generate multiple fields based on the values of a nominal or flag field. The newly generated fields can contain values from another field or numeric flags (0 and 1). The functionality of this node is similar to that of the Set to Flag node. However, it offers more flexibility. It allows you to create fields of any type (including numeric flags), using the values from another field. You can then perform aggregation or other manipulations with other nodes downstream. (The Set to Flag node lets you aggregate fields in one step, which may be convenient if you are creating flag fields.)

Figure 4-76
Generating restructured fields for Account



For example, the following dataset contains a nominal field, *Account*, with the values *Savings* and *Draft*. The opening balance and current balance are recorded for each account, and some customers have multiple accounts of each type. Let's say you want to know whether each customer has a particular account type, and if so, how much money is in each account type. You use the Restructure node to generate a field for each of the *Account* values, and you select *Current_Balance* as the value. Each new field is populated with the current balance for the given record.

Table 4-2
Sample data before restructuring

CustID	Account	Open_Bal	Current_Bal
12701	Draft	1000	1005.32
12702	Savings	100	144.51
12703	Savings	300	321.20
12703	Savings	150	204.51
12703	Draft	1200	586.32

Table 4-3
Sample data after restructuring

CustID	Account	Open_Bal	Current_Bal	Account_Draft_ Current_Bal	Account_Savings_ Current_Bal
12701	Draft	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	Draft	1200	586.32	586.32	\$null\$

Using the Restructure Node with the Aggregate Node

In many cases, you may want to pair the Restructure node with an Aggregate node. In the previous example, one customer (with the ID 12703) has three accounts. You can use an Aggregate node to calculate the total balance for each account type. The key field is *CustID*, and the aggregate fields are the new restructured fields, *Account_Draft_Current_Bal* and *Account_Savings_Current_Bal*. The following table shows the results.

Table 4-4
Sample data after restructuring and aggregation

CustID	Record_Count	Account_Draft_Current_ Bal_Sum	Account_Savings_Current_ Bal_Sum
12701	1	1005.32	\$null\$
12702	1	\$null\$	144.51
12703	3	586.32	525.71

Setting Options for the Restructure Node

Available fields. Lists all data fields with a measurement level of *Nominal* (set) or *Flag*. Select one from the list to display the values in the set or flag, then choose from these values to create the restructured fields. Note that data must be fully instantiated using an upstream source or Type node before you can see the available fields (and their values). For more information, see the topic [Type Node](#) on p. 113.

Available values. Values in the set selected above are displayed here. Select one or more values for which you want to generate restructured fields. For example, if the values in a field called *Blood Pressure* are *High*, *Medium*, and *Low*, you can select *High* and add it to the list on the right. This will create a field with a specified value (see below) for records with a value of *High*.

Create restructured fields. The newly created restructured fields are listed here. By default, new field names are automatically created by combining the original field name with the field value into a label, such as *Fieldname_fieldvalue*.

Include field name. Deselect to remove the original field name as a prefix from the new field names.

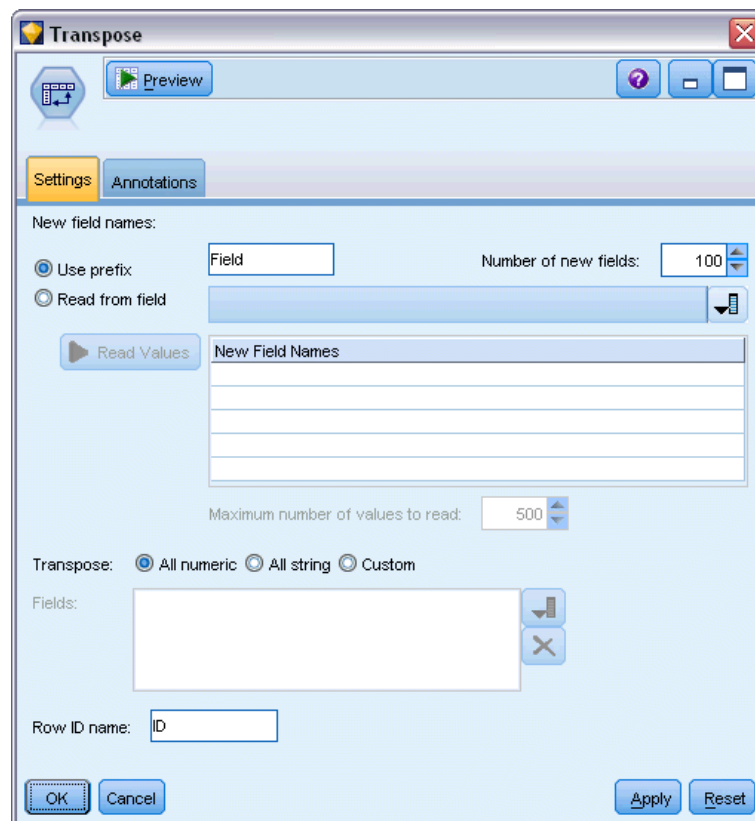
Use values from other fields. Specify one or more fields whose values will be used to populate the restructured fields. Use the Field Chooser to select one or more fields. For each field chosen, one new field is created. The value field name is appended to the restructured field name—for example, *BP_High_Age* or *BP_Low_Age*. Each new field inherits the type of the original value field.

Create numeric value flags. Select to populate the new fields with numeric value flags (0 for false and 1 for true), rather than using a value from another field.

Transpose Node

By default, columns are fields and rows are records or observations. If necessary, you can use a Transpose node to swap the data in rows and columns so that fields become records and records become fields. For example, if you have time series data where each series is a row rather than a column, you can transpose the data prior to analysis.

Figure 4-77
Transpose node, Settings tab



Setting Options for the Transpose Node

New Field Names

New field names can be generated automatically based on a specified prefix or read from an existing field in the data.

Use prefix. This option generates new field names automatically based on the specified prefix (*Field1*, *Field2*, and so on). You can customize the prefix as needed. With this option, you must specify the number of fields to be created, regardless of the number of rows in the original data.

For example, if Number of new fields is set to 100, all data beyond the first 100 rows will be discarded. If there are fewer than 100 rows in the original data, some fields will be null. (You can increase the number of fields as needed, but the purpose of this setting is to avoid transposing a million records into a million fields, which would produce an unmanageable result.)

For example, suppose you have data with series in rows and a separate field (column) for each month. You can transpose this so that each series is in a separate field, with a row for each month.

Figure 4-78

Original data with series in rows

	Jan	Feb	Mar	Apr
1	1	3	5	7
2	2	4	6	8

Figure 4-79

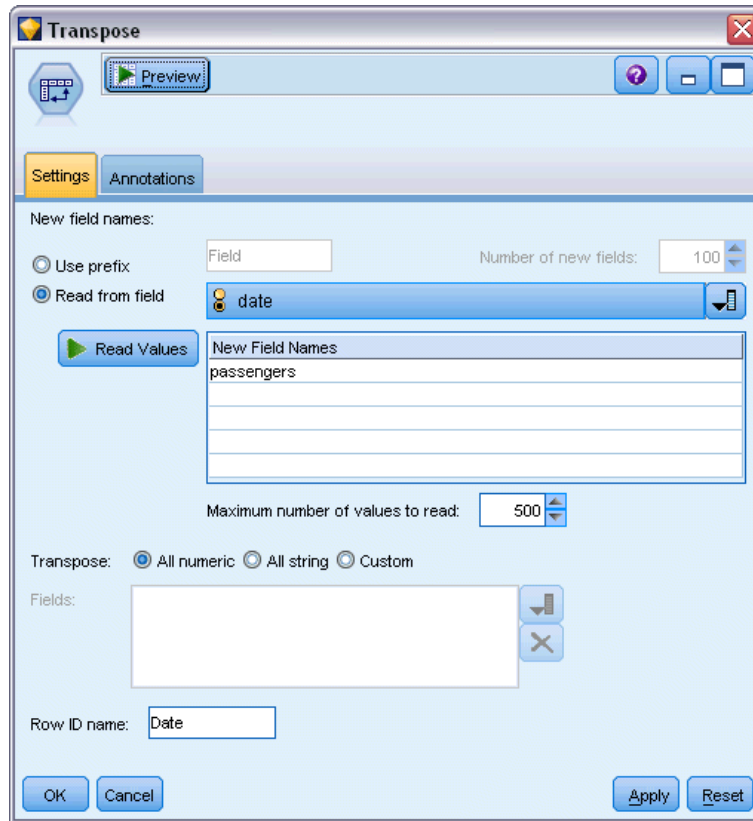
Transposed data with series in columns

	Month	Field1	Field2
1	Jan	1	2
2	Feb	3	4
3	Mar	5	6
4	Apr	7	8

Note: To produce the results shown, the Number of New Fields option was changed from 100 to 2, and the row ID name was changed from ID to Month (see below).

Read from field. Reads field names from an existing field. With this option, the number of new fields is determined by the data, up to the specified maximum. Each value of the selected field becomes a new field in the output data. The selected field can have any storage type (integer, string, date, and so on), but in order to avoid duplicate field names, each value of the selected field must be unique (in other words, the number of values should match the number of rows). If duplicate field names are encountered, a warning is displayed.

Figure 4-80
Reading field names from an existing field



- **Read Values.** If the selected field has not been instantiated, select this option to populate the list of new field names. If the field has already been instantiated, then this step is not necessary.
- **Maximum number of values to read.** When reading fields names from the data, an upper limit is specified in order to avoid creating an inordinately large number of fields. (As noted above, transposing one million records into one million fields would produce an unmanageable result.)

For example, if the first column in your data specifies the name for each series, you can use these values as fields names in the transposed data.

Figure 4-81
Original data with series in a single row

	date	1949-01-01	1949-02-01	1949-04-01	1949-05-01	1949-06-01	1949-07-01	1949-08-01
1	passengers	112.000	118.000	129.000	121.000	135.000	148.000	148.000

Figure 4-82
Transposed data with series in columns

	Date	passengers
1	1949-01-01	112.000
2	1949-02-01	118.000
3	1949-04-01	129.000
4	1949-05-01	121.000
5	1949-06-01	135.000
6	1949-07-01	148.000
7	1949-08-01	148.000
8	1949-09-01	136.000
9	1949-10-01	119.000
10	1949-11-01	104.000
11	1949-12-01	118.000
12	1950-01-01	115.000
13	1950-02-01	126.000
14	1950-03-01	141.000
15	1950-04-01	135.000
16	1950-05-01	125.000
17	1950-06-01	149.000
18	1950-07-01	170.000
19	1950-08-01	170.000
20	1950-09-01	158.000

Transpose. By default, only continuous (numeric range) fields are transposed (either integer or real storage). Optionally, you can choose a subset of numeric fields or transpose string fields instead. However, all transposed fields must be of the same storage type—either numeric or string but not both—since mixing the input fields would generate mixed values within each output column, which violates the rule that all values of a field must have the same storage. Other storage types (date, time, timestamp) cannot be transposed.

- All numeric.** Transposes all numeric fields (integer or real storage). The number of rows in the output matches the number of numeric fields in the original data.

- **All string.** Transposes all string fields.
- **Custom.** Allows you to select a subset of numeric fields. The number of rows in the output matches the number of fields selected. *Note:* This option is available only for numeric fields.

Row ID name. Specifies the name of the row ID field created by the node. The values of this field are determined by the names of the fields in the original data.

Tip: When transposing time series data from rows to columns, if your original data includes a row, such as date, month, or year, that labels the period for each measurement, be sure to read these labels into IBM® SPSS® Modeler as field names (as demonstrated in the above examples, which show the month or date as field names in the original data, respectively) rather than including the label in the first row of data. This will avoid mixing labels and values in each column (which would force numbers to be read as strings, since storage types cannot be mixed within a column).

Time Intervals Node

The Time Intervals node allows you to specify intervals and generate labels for time series data to be used in a Time Series modeling or a Time Plot node for estimating or forecasting. A full range of time intervals is supported, from seconds to years. For example, if you have a series with daily measurements beginning January 3, 2005, you can label records starting on that date, with the second row being January 4, and so on. You can also specify the periodicity—for example, five days per week or eight hours per day.

In addition, you can specify the range of records to be used for estimating. You can choose whether to exclude the earliest records in the series and whether to specify holdouts. Doing so enables you to test the model by holding out the most recent records in the time series data in order to compare their known values with the estimated values for those periods.

You can also specify how many time periods into the future you want to forecast, and you can specify future values for use in forecasting by downstream Time Series modeling nodes.

The Time Intervals node generates a *TimeLabel* field in a format appropriate to the specified interval and period along with a *TimeIndex* field that assigns a unique integer to each record. A number of additional fields may also be generated, depending on the selected interval or period (such as the minute or second within which a measurement falls).

You can pad or aggregate values as needed to ensure that measurements are equally spaced. Methods for modeling time series data require a uniform interval between each measurement, with any missing values indicated by empty rows. If your data do not already meet this requirement, the node can transform them to do so.

Comments

- Periodic intervals may not match real time. For example, a series based on a standard five-day work week would treat the gap between Friday and Monday as a single day.
- The Time Intervals node assumes that each series is in a field or column, with a row for each measurement. If necessary you can transpose your data to meet this requirement. For more information, see the topic [Transpose Node](#) on p. 182.
- For series that are not equally spaced, you can specify a field that identifies the date or time for each measurement. Note that this requires a date, time, or timestamp field in the appropriate format to use as input. If necessary, you can convert an existing field (such as a string

label field) to this format using a Filler node. For more information, see the topic [Storage Conversion Using the Filler Node](#) on p. 153.

- When viewing details for the generated label and index fields, turning on the display of value labels is often helpful. For example, when viewing a table with values generated for monthly data, you can click the value labels icon on the toolbar to see *January, February, March*, and so on, rather than *1, 2, 3*, and so on.

Figure 4-83
Value labels icon



Specifying Time Intervals

The Intervals tab allows you to specify the interval and periodicity for building or labeling the series. The specific settings depend on the selected interval. For example, if you choose Hours per day, you can specify the number of days per week, the day each week begins, the number of hours in each day, and the hour each day begins. For more information, see the topic [Supported Intervals](#) on p. 194.

Figure 4-84
Time-interval settings for an hourly series

Time Intervals

Preview

Periodicity: 24

Intervals | Build | Estimation | Forecast | Annotations

Time Interval: Hours per day

Number of days per week: 7 | Week begins on: Monday

Number of hours in a day: 24 | Day begins at: 00:00

Start labeling from first record Build from data

Year: 2000 | Month: January | Day: 1

Time: 00:00

New field name extension: \$TI_ | Add as: Prefix Suffix

Date format: YYYY-MM-DD | Time format: HH:MM:SS

OK | Cancel | Apply | Reset

Labeling or Building the Series

You can label records consecutively or build the series based on a specified date, timestamp, or time field.

- **Start labeling from the first record.** Specify the starting date and/or time to label consecutive records. If labeling hours per day, for example, you would specify the date and hour when the series begins, with a single record for each hour thereafter. Aside from adding labels, this method does not change the original data. Instead, it assumes that records are already equally spaced, with a uniform interval between each measurement. Any missing measurements must be indicated by empty rows in the data.
- **Build from data.** For series that are not equally spaced, you can specify a field that identifies the date or time for each measurement. Note that this requires a date, time, or timestamp field in the appropriate format to use as input. For example if you have a string field with values like *Jan 2000*, *Feb 2000*, etc., you can convert this to a date field using a Filler node. For more information, see the topic [Storage Conversion Using the Filler Node](#) on p. 153. The Build from data option also transforms the data to match the specified interval by padding or aggregating records as needed, for example, by “rolling up” weeks into months, or by replacing missing records with blanks or extrapolated values. You can specify the functions used to pad or aggregate records on the Build tab. For more information, see the topic [Time Interval Build Options](#) on p. 188.

New field name extension. Allows you to specify a prefix or suffix that is applied to all fields generated by the node. For example, using the default *\$TI_* prefix, the fields created by the node would be named *\$TI_TimeIndex*, *\$TI_TimeLabel*, and so on

Date format. Specifies the format for the *TimeLabel* field created by the node, as applicable to the current interval. Availability of this options depends on the current selection.

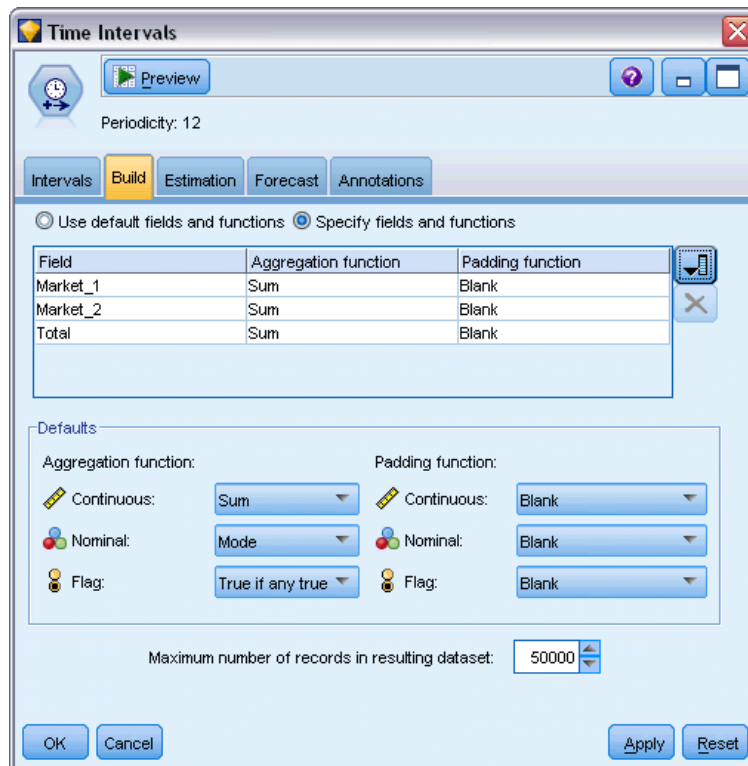
Time format. Specifies the format for the *TimeLabel* field created by the node, as applicable to the current interval. Availability of this options depends on the current selection.

Time Interval Build Options

The Build tab in the Time Intervals node allows you to specify options for aggregating or padding fields to match the specified interval. These settings apply only when the Build from data option is selected on the Intervals tab. For example, if you have a mix of weekly and monthly data, you could aggregate or “roll up” the weekly values to achieve a uniform monthly interval. Alternatively, you could set the interval to weekly and pad the series by inserting blank values for any weeks that are missing, or by extrapolating missing values using a specified padding function.

When you pad or aggregate data, any existing date or timestamp fields are effectively superseded by the generated *TimeLabel* and *TimeIndex* fields and are dropped from the output. Typeless fields are also dropped. Fields that measure time as a duration are preserved—such as a field that measures the length of a service call rather than the time the call started—as long as they are stored internally as time fields rather than timestamp. For more information, see the topic [Setting Field Storage and Formatting](#) in Chapter 2 on p. 23. Other fields are aggregated based on the options specified in the Build tab.

Figure 4-85
Time Intervals node, Build tab



- **Use default fields and functions.** Specifies that all fields should be aggregated or padded as needed, with the exception of date, timestamp, and typeless fields as noted above. The default function is applied based on the measurement level—for example, continuous fields are aggregated using the mean, while nominal fields use the mode. You can change the default for one or more measurement levels in the lower part of the dialog box.
- **Specify fields and functions.** Allows you to specify the fields to pad or aggregate, and the function used for each. Any fields not selected are dropped from the output. Use the icons on the right side to add or remove fields from the table, or click the cell in the appropriate column to change the aggregation or padding function used for that field to override the default. Typeless fields are excluded from the list and cannot be added to the table.

Defaults. Specifies the aggregation and padding functions used by default for different types of fields. These defaults are applied when Use defaults is selected and are also applied as the initial default for any new fields added to the table. (Changing the defaults does not change any of the existing settings in the table but does apply to any fields added subsequently.)

Aggregation function. The following aggregation functions are available:

- **Continuous.** Available functions for continuous fields include Mean, Sum, Mode, Min, and Max.
- **Nominal.** Options include Mode, First, and Last. First means the first non-null value (sorted by date) in the aggregation group; last means the last non-null value in the group.
- **Flag.** Options include True if any true, Mode, First, and Last.

Padding function. The following padding functions are available:

- **Continuous.** Options include Blank and Mean of most recent points, which means the mean of the three most recent non-null values prior to the time period that will be created. If there are not three values, the new value is blank. Recent values only include actual values; a previously created padded value is not considered in the search for a non-null value.
- **Nominal.** Blank and Most recent value. Most recent refers to the most recent non-null value prior to the time period that will be created. Again, only actual values are considered in the search for a recent value.
- **Flag.** Options include Blank, True, and False.

Maximum number of records in resulting dataset. Specifies an upper limit to the number of records created, which can otherwise become quite large, particularly when the time interval is set to seconds (whether deliberately or otherwise). For example, a series with only two values (Jan. 1, 2000 and Jan. 1, 2001) would generate 31,536,000 records if padded out to seconds (60 seconds x 60 minutes x 24 hours x 365 days). The system will stop processing and display a warning if the specified maximum is exceeded.

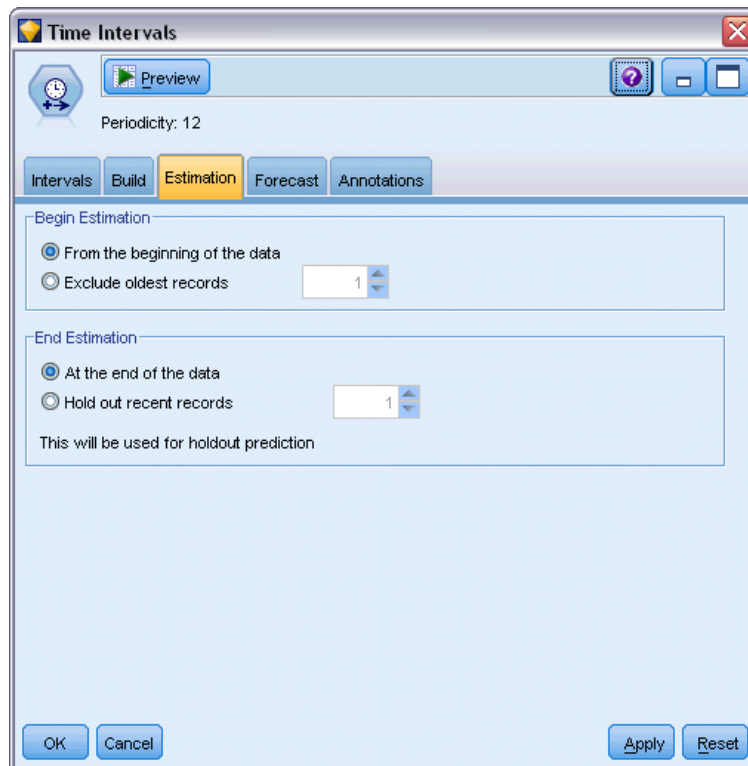
Count Field

When aggregating or padding values, a new *Count* field is created that indicates the number of records involved in determining the new record. If four weekly values were aggregated into a single month, for example, the count would be 4. For a padded record, the count is 0. The name of the field is *Count*, plus the prefix or suffix specified on the Interval tab.

Estimation Period

The Estimation tab of the Time Intervals node allows you to specify the range of records used in model estimation, as well as any holdouts. These settings may be overridden in downstream modeling nodes as needed, but specifying them here may be more convenient than specifying them for each node individually.

Figure 4-86
Time Intervals node, Estimation tab



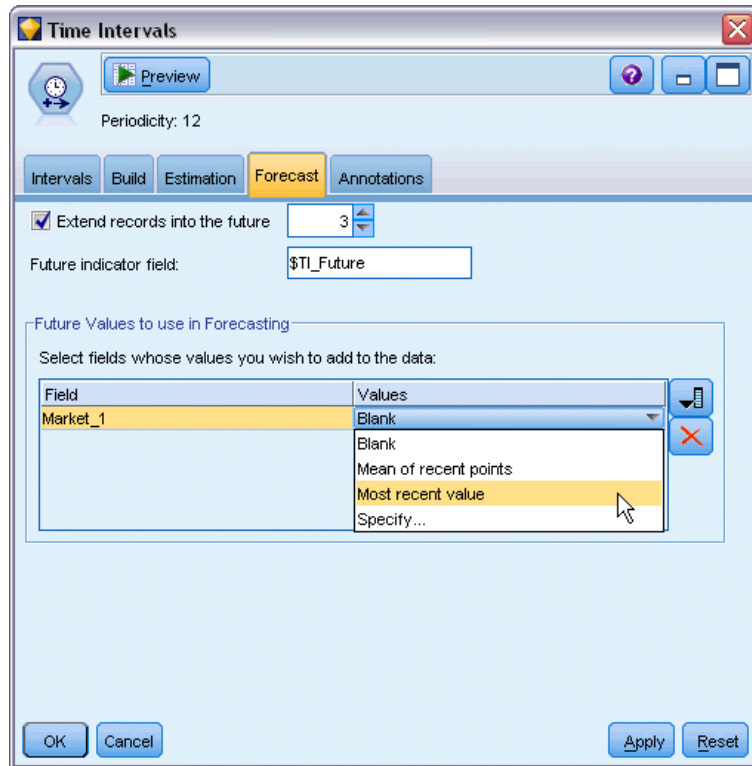
Begin Estimation. You can begin the estimation period at the beginning of the data or exclude older values that may be of limited use in forecasting. Depending on the data, you may find that shortening the estimation period can speed up performance (and reduce the amount of time spent on data preparation) with no significant loss in forecasting accuracy.

End Estimation. You can estimate the model using all records up to the end of the data, or “hold out” the most recent records in order to evaluate the model. In the latter case, you are effectively “forecasting” values that are already known, allowing you to compare observed and predictive values to gauge the model’s effectiveness.

Forecasts

The Forecast tab of the Time Intervals node allows you to specify the number of records you want to forecast and to specify future values for use in forecasting by downstream Time Series modeling nodes. These settings may be overridden in downstream modeling nodes as needed, but specifying them here may be more convenient than specifying them for each node individually.

Figure 4-87
Time Intervals node, Forecast tab



Extend records into the future. Specifies the number of records to forecast beyond the estimation period. Note that these records may or may not be “forecasts” depending on the number of holdouts that are specified on the Estimation tab.

Future indicator field. Label of the generated field that indicates whether a record contains forecast data. Default value for the label is *\$TI_Future*.

Future Values to Use in Forecasting. For each record that you want to forecast (excluding holdouts), if you are using predictor fields (with the role set to *Input*), you must specify estimated values for the forecast period for each predictor. You can either specify values manually, or choose from a list.

- **Field.** Click the field selector button and choose any fields that may be used as predictors. Note that fields selected here may or may not be used in modeling; to actually use a field as a predictor, it must be selected in a downstream modeling node. This dialog box simply gives you a convenient place to specify future values so they can be shared by multiple downstream modeling nodes without specifying them separately in each node. Also note that the list of available fields may be constrained by selections on the Build tab. For example, if Specify fields and functions is selected on the Build tab, any fields not aggregated or padded are dropped from the stream and cannot be used in modeling.

Note: If future values are specified for a field that is no longer available in the stream (because it has been dropped or because of updated selections made on the Build tab), the field is shown in red on the Forecast tab.

- Values.** For each field, you can choose from a list of functions, or click Specify to either enter values manually or choose from a list of predefined values. If the predictor fields relate to items that are under your control, or which are otherwise knowable in advance, you should enter values manually. For example, if you are forecasting next month's revenues for a hotel based on the number of room reservations, you could specify the number of reservations you actually have for that period. Conversely, if a predictor field relates to something outside your control, such as a stock price, you could use a function such as the most recent value or the mean of recent points.

The available functions depend on the measurement level of the field.

Measurement level	Functions
Continuous or Nominal field	Blank Mean of recent points Most recent value Specify
Flag field	Blank Most recent value True False Specify

Mean of recent points—Calculates the future value from the mean of the last three data points.

Most recent value—Sets the future value to that of the most recent data point.

True/False—Sets the future value of a flag field to True or False as specified.

Specify—Opens a dialog box for specifying future values manually, or choosing them from a predefined list.

Figure 4-88
Specifying future values for predictors



Future Values

Here you can specify future values for use in forecasting by downstream Time Series modeling nodes. These settings may be overridden in downstream modeling nodes as needed, but specifying them here may be more convenient than specifying them for each node individually.

You can enter values manually, or click the selector button on the right side of the dialog box to choose from a list of values defined for the current field.

The number of future values that you can specify corresponds to the number of records by which you are extending the time series into the future.

Supported Intervals

The Time Intervals node supports a full range of intervals from seconds to years, as well as cyclic (for example, seasonal) and non-cyclic periods. You specify the interval in the Time Interval field on the Intervals tab.

Periods

Select Periods to label an existing, non-cyclic series that doesn't match any of the other specified intervals. The series must already be in the correct order, with a uniform interval between each measurement. The Build from data option is not available when this interval is selected.

Figure 4-89
Time-interval settings for non-cyclic periods

Sample Output

Records are labeled incrementally based on the specified starting value (*Period 1, Period 2, and so on*). New fields are created as follows:

\$TI_TimeIndex (Integer)	\$TI_TimeLabel (String)	\$TI_Period (Integer)
1	Period 1	1
2	Period 2	2
3	Period 3	3
4	Period 4	4
5	Period 5	5

Cyclic Periods

Select Cyclic Periods to label an existing series with a repeating cycle that doesn't fit one of the standard intervals. For example, you could use this option if you have only 10 months in your fiscal year. The series must already be in the correct order, with a uniform interval between each measurement. (The Build from data option is not available when this interval is selected.)

Figure 4-90
Time-interval settings for cyclic periods

Time Interval: Cyclic Periods

Number of periods per cycle: 12

Start labeling from first record Build from data

Cycle: 1

Period: 1

New field name extension: \$TI_ Add as: Prefix Suffix

Sample Output

Records are labeled incrementally based on the specified starting cycle and period (*Cycle 1, Period 1, Cycle 1, Period 2*, and so on). For example, with the number of periods per cycle set to 3, new fields are created as follows:

\$TI_TimeIndex (Integer)	\$TI_TimeLabel (String)	\$TI_Cycle (Integer)	\$TI_Period (Integer)
1	Cycle 1, Period 1	1	1
2	Cycle 1, Period 2	1	2
3	Cycle 1, Period 3	1	3
4	Cycle 2, Period 1	2	1
5	Cycle 2, Period 2	2	2

Years

For years, you can specify the starting year to label consecutive records or select Build from data to specify a timestamp or date field that identifies the year for each record.

Figure 4-91
Time-interval settings for a yearly series

Time Interval: Years

Start labeling from first record Build from data

Year: 2000

New field name extension: \$TI_ Add as: Prefix Suffix

Sample Output

New fields are created as follows:

\$TI-TimeIndex (Integer)	\$TI-TimeLabel (String)	\$TI-Year (Integer)
1	2000	2000
2	2001	2001
3	2002	2002
4	2003	2003
5	2004	2004

Quarters

For a quarterly series, you can specify the month when the fiscal year begins. You can also specify the starting quarter and year (for example, Q1 2000) to label consecutive records or select Build from data to choose a timestamp or date field that identifies the quarter and year for each record.

Figure 4-92

Time-interval settings for quarterly series

Time Interval:

Fiscal year begins:

Start labeling from first record Build from data

Year: Quarter:

New field name extension: Add as: Prefix Suffix

Sample Output

For a fiscal year starting in January, new fields would be created and populated as follows:

\$TI-TimeIndex (Integer)	\$TI-TimeLabel (String)	\$TI-Year (Integer)	\$TI-Quarter (Integer with labels)
1	Q1 2000	2000	1 (Q1)
2	Q2 2000	2000	2 (Q2)
3	Q3 2000	2000	3 (Q3)
4	Q4 2000	2000	4 (Q4)
5	Q1 2001	2001	1 (Q1)

If the year starts in a month other than January, new fields are as below (assuming a fiscal year starting in July). To view the labels that identify the months for each quarter, turn on the display of value labels by clicking the toolbar icon.

Figure 4-93
Value labels icon



\$TI-TimeIndex (Integer)	\$TI-TimeLabel (String)	\$TI-Year (Integer)	\$TI-Quarter (Integer with labels)
1	Q1 2000/2001	1	1 (Q1 Jul-Sep)
2	Q2 2000/2001	1	2 (Q2 Oct-Dec)
3	Q3 2000/2001	1	3 (Q3 Jan-Mar)
4	Q4 2000/2001	1	4 (Q4 Apr-Jun)
5	Q1 2001/2002	2	1 (Q1 Jul-Sep)

Months

You can select the starting year and month to label consecutive records or select Build from data to choose a timestamp or date field that indicates the month for each record.

Figure 4-94
Time-interval settings for a monthly series

Time Interval:

Start labeling from first record Build from data

Year: Month:

New field name extension: Add as: Prefix Suffix

Sample Output

New fields are created as follows:

\$TI-TimeIndex (Integer)	\$TI-TimeLabel (Date)	\$TI-Year (Integer)	\$TI-Months (Integer with labels)
1	Jan 2000	2000	1 (January)
2	Feb 2000	2000	2 (February)
3	Mar 2000	2000	3 (March)
4	Apr 2000	2000	4 (April)
5	May 2000	2000	5 (May)

Weeks (Non-Periodic)

For a weekly series, you can select the day of the week on which the cycle begins.

Note that weeks can be only non-periodic because different months, quarters, and even years do not necessarily have the same number of weeks. However, time-stamped data can be easily aggregated or padded to a weekly level for non-periodic models.

Figure 4-95
Time-interval settings for a weekly series

Sample Output

New fields are created as follows:

\$TI-TimeIndex (Integer)	\$TI-TimeLabel (Date)	\$TI-Week (Integer)
1	1999-12-27	1
2	2000-01-03	2
3	2000-01-10	3
4	2000-01-17	4
5	2000-01-24	5

The *\$TI-TimeLabel* field for a week shows the first day of that week. In the preceding table, the user starts labeling from January 1, 2000. However, the week starts on Monday, and January 1, 2000, is a Saturday. Thus, the week that includes January 1 starts on December 27, 1999, and is the label of the first point.

The Date format determines the strings produced for the *\$TI-TimeLabel* field.

Days per Week

For daily measurements that fall into a weekly cycle, you can specify the number of days per week and the day each week begins. You can specify a starting date to label consecutive records or select Build from data to choose a timestamp or date field that indicates the date for each record.

Figure 4-96
Time-interval settings for a daily series

Time Interval: Days per week

Number of days per week: 7 Week begins on: Monday

Start labeling from first record Build from data

Year: 2000 Month: January Day: 1

New field name extension: \$TI_ Add as: Prefix Suffix

Date format: YYYY-MM-DD

Sample Output

New fields are created as follows:

\$TI-TimeIndex (Integer)	\$TI-TimeLabel (Date)	\$TI-Week (Integer)	\$TI-Day (Integer with labels)
1	Jan 5 2005	1	3 (Wednesday)
2	Jan 6 2005	1	4 (Thursday)
3	Jan 7 2005	1	5 (Friday)
4	Jan 10 2005	2	1 (Monday)
5	Jan 11 2005	2	2 (Tuesday)

Note: The week always starts at 1 for the first time period and does not cycle based on the calendar. Thus, week 52 is followed by week 53, 54, and so on. The week does not reflect the week of the year, just the number of weekly increments in the series.

Days (Non-Periodic)

Choose non-periodic days if you have daily measurements that do not fit into a regular weekly cycle. You can specify a starting date to label consecutive records or select Build from data to choose a timestamp or date field that indicates the date for each record.

Figure 4-97
Time-interval settings for a daily series (non-periodic)

Time Interval: Days (non-periodic)

Start labeling from first record Build from data

Year: 2000 Month: January Day: 1

New field name extension: \$TI_ Add as: Prefix Suffix

Date format: YYYY-MM-DD

Sample Output

New fields are created as follows:

\$TI-TimeIndex (Integer)	\$TI-TimeLabel (Date)
1	Jan 5 2005
2	Jan 6 2005
3	Jan 7 2005
4	Jan 8 2005
5	Jan 9 2005

Hours per Day

For hourly measurements that fit into a daily cycle, you can specify the number of days per week, the number of hours in the day (such as an eight-hour workday), the day when the week begins, and the hour when each day begins. Hours can be specified down to the minute based on a 24-hour clock (for example, 14:05 = 2:05 p.m.)

Figure 4-98
Time-interval settings for an hourly series

You can specify the starting date and time to label consecutive records or select Build from data to choose a timestamp field that identifies the date and time for each record.

Sample Output

New fields are created as follows:

\$TI-TimeIndex (Integer)	\$TI-TimeLabel (Timestamp)	\$TI-Day (Integer with labels)	\$TI-Hour (Integer with labels)
1	Jan 5 2005 8:00	3 (Wednesday)	8 (8:00)
2	Jan 5 2005 9:00	3 (Wednesday)	9 (9:00)
3	Jan 5 2005 10:00	3 (Wednesday)	10 (10:00)
4	Jan 5 2005 11:00	3 (Wednesday)	11 (11:00)
5	Jan 5 2005 12:00	3 (Wednesday)	12 (12:00)

Hours (Non-Periodic)

Choose this option if you have hourly measurements that do not fit into a regular daily cycle. You can specify the starting time to label consecutive records or select Build from data to choose a timestamp or time field that indicates the time for each record.

Figure 4-99
Time-interval settings for yearly data

Hours are based on a 24-hour clock (13:00 = 1:00 p.m.), and do not wrap (hour 25 follows hour 24).

Sample Output

New fields are created as follows:

\$TI-TimeIndex (Integer)	\$TI-TimeLabel (String)	\$TI-Hour (Integer with labels)
1	8:00	8 (8:00)
2	9:00	9 (9:00)
3	10:00	10 (10:00)
4	11:00	11 (11:00)
5	12:00	12 (12:00)

Minutes per Day

For measurements by the minute that fall into a daily cycle, you can specify the number of days per week, the day the week begins, the number of hours in the day, and the time the day begins. Hours are specified based on a 24-hour clock and can be specified down to the minute and second using colons (for example, 2:05:17 p.m. = 14:05:17). You can also specify the number of minutes to increment (every minute, every two minutes, and so on, where the increment must be a value that divides evenly into 60).

Figure 4-100
Time-interval settings for minutes per day

You can specify the starting date and time to label consecutive records or select Build from data to choose a timestamp field that identifies the date and time for each record.

Sample Output

New fields are created as follows:

\$TI-TimeIndex (Integer)	\$TI-TimeLabel (Timestamp)	\$TI-Minute
1	2005-01-05 08:00:00	0
2	2005-01-05 08:01:00	1
3	2005-01-05 08:02:00	2
4	2005-01-05 08:03:00	3
5	2005-01-05 08:04:00	4

Minutes (Non-Periodic)

Choose this option if you have measurements by the minute that do not fit into a regular daily cycle. You can specify the number of minutes to increment (every minute, every two minutes, and so on, where the specified value must be a number that divides evenly into 60).

Figure 4-101
Time-interval settings for minutes (non-periodic)

You can specify the starting time to label consecutive records or select Build from data to choose a timestamp or time field that identifies the time for each record.

Sample Output

New fields are created as follows:

\$TI-TimeIndex (Integer)	\$TI-TimeLabel (String)	\$TI-Minute
1	8:00	0
2	8:01	1
3	8:02	2
4	8:03	3
5	8:04	4

- The *TimeLabel* string is created by using a colon between the hour and minute. The hour does not wrap—hour 25 follows hour 24.
- Minutes increment by the value specified in the dialog box. For example, if the increment is 2, the *TimeLabel* will be 8:00, 8:02, and so on, and the minutes will be 0, 2, and so on.

Seconds per Day

For second intervals that fall into a daily cycle, you can specify the number of days per week, the day the week begins, the number of hours in the day, and the time the day begins. Hours are specified based on a 24-hour clock and can be specified down to the minute and second using colons (For example, 2:05:17 p.m. = 14:05:17). You can also specify the number of seconds to increment (every second, every two seconds, and so on, where the specified value must be a number that divides evenly into 60).

Figure 4-102
Time-interval settings for seconds per day

Time Interval: Seconds per day Increment by: 1

Number of days per week: 7 Week begins on: Monday

Number of hours in a day: 24 Day begins at: 00:00

Start labeling from first record Build from data

Year: 2000 Month: January Day: 1

Time: 00:00:00

New field name extension: \$TI_ Add as: Prefix Suffix

Date format: YYYY-MM-DD Time format: HH:MM:SS

You can specify the date and time to start labeling consecutive records or select Build from data to choose a timestamp field that specifies the date and time for each record.

Sample Output

New fields are created as follows:

\$TI-TimeIndex (Integer)	\$TI-TimeLabel (Timestamp)	\$TI-Minute	\$TI-Second
1	2005-01-05 08:00:00	0	0
2	2005-01-05 08:00:01	0	1
3	2005-01-05 08:00:02	0	2
4	2005-01-05 08:00:03	0	3
5	2005-01-05 08:00:04	0	4

Seconds (Non-Periodic)

Choose this option if you have measurements taken by the second that do not fit into a regular daily cycle. You can specify the number of seconds to increment (every second, every two seconds, and so on, where the specified value must be a number that divides evenly into 60).

Figure 4-103

Time-interval settings for seconds (non-periodic)

Specify the time to start labeling consecutive records or select *Build from data* to choose a timestamp or time field the identifies the time for each record.

Sample Output

New fields are created as follows:

\$TI-TimeIndex (Integer)	\$TI-TimeLabel (String)	\$TI-Minute	\$TI-Second
1	8:00:00	0	0
2	8:00:01	0	1
3	8:00:02	0	2
4	8:00:03	0	3
5	8:00:04	0	4

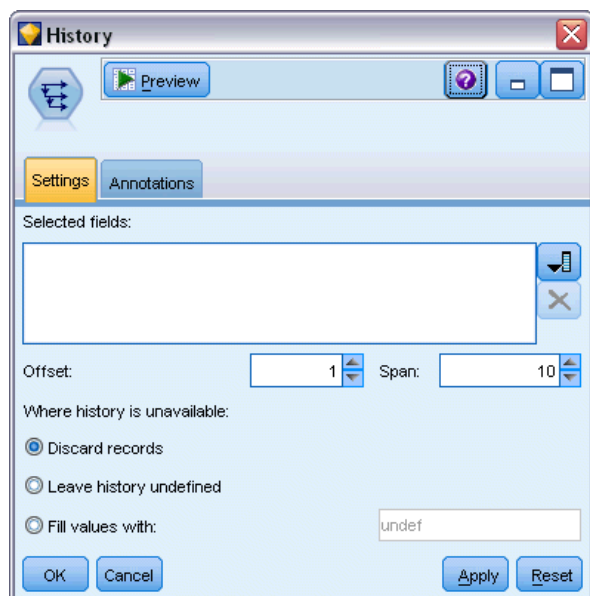
- The *TimeLabel* string is created by using a colon between the hour and minute and between minute and second. The hour does not wrap—after hour 24 will be hour 25.
- Seconds increment by whatever number is specified as the increment. If the increment is 2, the *TimeLabel* will be 8:00:00, 8:00:02, and so on, and the seconds will be 0, 2, and so on.

History Node

History nodes are most often used for sequential data, such as time series data. They are used to create new fields containing data from fields in previous records. When using a History node, you may want to have data that is presorted by a particular field. You can use a Sort node to do this.

Setting Options for the History Node

Figure 4-104
History node dialog box



Selected fields. Using the Field Chooser (button to the right of the text box), select the fields for which you want a history. Each selected field is used to create new fields for all records in the dataset.

Offset. Specify the latest record prior to the current record from which you want to extract historical field values. For example, if Offset is set to 3, as each record passes through this node, the field values for the third record previous will be included in the current record. Use the Span settings to specify how far back records will be extracted from. Use the arrows to adjust the offset value.

Span. Specify how many prior records from which you want to extract values. For example, if Offset is set to 3 and Span is set to 5, each record that passes through the node will have five fields added to it for each field specified in the Selected Fields list. This means that when the node is processing record 10, fields will be added from record 7 through record 3. Use the arrows to adjust the span value.

Where history is unavailable. Select one of the following options for handling records that have no history values. This usually refers to the first several records at the top of the dataset, for which there are no previous records to use as a history.

- **Discard records.** Select to discard records where no history value is available for the field selected.

- **Leave history undefined.** Select to keep records where no history value is available. The history field will be filled with an undefined value, displayed as `$null$`.
- **Fill values with.** Specify a value or string to be used for records where no history value is available. The default replacement value is `undef`, the system null. Null values are displayed using the string `$null$`.

When selecting a replacement value, keep in mind the following rules in order for proper execution to occur:

- Selected fields should be of the same storage type.
- If all of the selected fields have numeric storage, the replacement value must be parsed as an integer.
- If all of the selected fields have real storage, the replacement value must be parsed as a real.
- If all of the selected fields have symbolic storage, the replacement value must be parsed as a string.
- If all of the selected fields have date/time storage, the replacement value must be parsed as a date/time field.

If any of the above conditions are not met, you will receive an error when executing the History node.

Field Reorder Node

The Field Reorder node enables you to define the natural order used to display fields downstream. This order affects the display of fields in a variety of places, such as tables, lists, and the Field Chooser. This operation is useful, for example, when working with wide datasets to make fields of interest more visible.

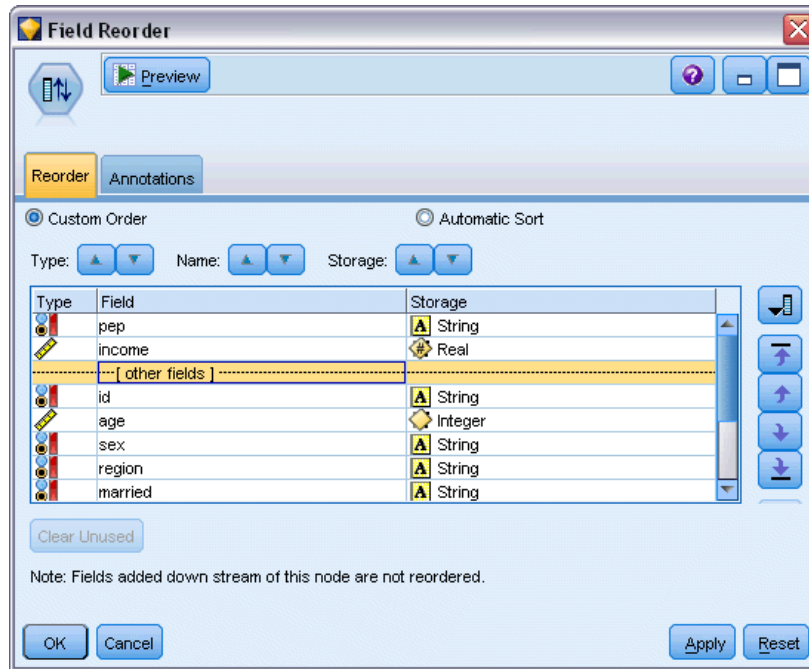
Setting Field Reorder Options

There are two ways to reorder fields: custom ordering and automatic sorting.

Custom Ordering

Select Custom Order to enable a table of field names and types where you can view all fields and use arrow buttons to create a custom order.

Figure 4-105
Reordering to display fields of interest first



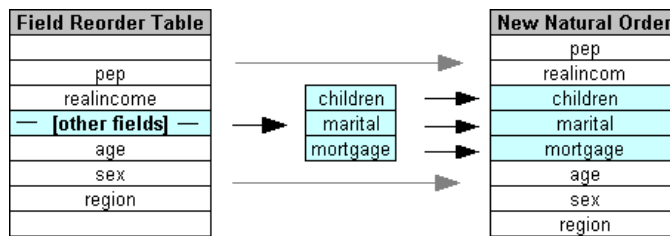
To reorder fields:

- ▶ Select a field in the table. Use the Ctrl-click method to select multiple fields.
- ▶ Use the simple arrow buttons to move the field(s) up or down one row.
- ▶ Use the line-arrow buttons to move the field(s) to the bottom or top of the list.
- ▶ Specify the order of fields not included here by moving up or down the divider row, indicated as [other fields].

Other fields. The purpose of the [other fields] divider row is to break the table into two halves.

- Fields appearing above the divider row will be ordered (as they appear in the table) at the top of all natural orders used to display the fields downstream of this node.
- Fields appearing below the divider row will be ordered (as they appear in the table) at the bottom of all natural orders used to display the fields downstream of this node.

Figure 4-106
Diagram illustrating how “other fields” are incorporated into the new field order



- All other fields not appearing in the field reorder table will appear between these “top” and “bottom” fields as indicated by the placement of the divider row.

Additional custom sorting options include:

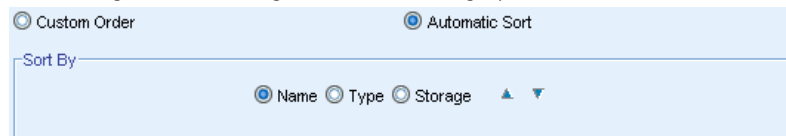
- Sort fields in ascending or descending order by clicking on the arrows above each column header (Type, Name, and Storage). When sorting by column, fields not specified here (indicated by the [other fields] row) are sorted last in their natural order.
- Click Clear Unused to delete all unused fields from the Field Reorder node. Unused fields are displayed in the table with a red font. This indicates that the field has been deleted in upstream operations.
- Specify ordering for any new fields (displayed with a lightning icon to indicate a new or unspecified field). When you click OK or Apply, the icon disappears.

Note: If fields are added upstream after a custom order has been applied, the new fields will be appended at the bottom of the custom list.

Automatic Sorting

Select Automatic Sort to specify a parameter for sorting. The dialog box options dynamically change to provide options for automatic sorting.

Figure 4-107
Reordering all fields using automatic sorting options



Sort By. Select one of three ways to sort fields read into the Reorder node. The arrow buttons indicate whether the order will be ascending or descending. Select one to make a change.

- Name
- Type
- Storage

Fields added upstream of the Field Reorder node after auto-sort has been applied will automatically be placed in their proper position based on the sort type selected.

Graph Nodes

Common Graph Nodes Features

Several phases of the data mining process use graphs and charts to explore data brought into IBM® SPSS® Modeler. For example, you can connect a Plot or Distribution node to a data source to gain insight into data types and distributions. You can then perform record and field manipulations to prepare the data for downstream modeling operations. Another common use of graphs is to check the distribution and relationships between newly derived fields.

Once you have configured the options for a graph node, you can execute it from within the dialog box or as part of a stream. In the generated graph window, you can generate Derive (Set and Flag) and Select nodes based on a selection or region of data, effectively “subsetting” the data. For example, you might use this powerful feature to identify and exclude outliers.

The Graphs palette contains the following nodes:



The Graphboard node offers many different types of graphs in one single node. Using this node, you can choose the data fields you want to explore and then select a graph from those available for the selected data. The node automatically filters out any graph types that would not work with the field choices. For more information, see the topic [Graphboard Node](#) on p. 216.



The Plot node shows the relationship between numeric fields. You can create a plot by using points (a scatterplot) or lines. For more information, see the topic [Plot Node](#) on p. 229.



The Distribution node shows the occurrence of symbolic (categorical) values, such as mortgage type or gender. Typically, you might use the Distribution node to show imbalances in the data, which you could then rectify using a Balance node before creating a model. For more information, see the topic [Distribution Node](#) on p. 237.



The Histogram node shows the occurrence of values for numeric fields. It is often used to explore the data before manipulations and model building. Similar to the Distribution node, the Histogram node frequently reveals imbalances in the data. For more information, see the topic [Histogram Plot Tab](#) on p. 242.



The Collection node shows the distribution of values for one numeric field relative to the values of another. (It creates graphs that are similar to histograms.) It is useful for illustrating a variable or field whose values change over time. Using 3-D graphing, you can also include a symbolic axis displaying distributions by category. For more information, see the topic [Collection Plot Tab](#) on p. 246.



The Multiplot node creates a plot that displays multiple *Y* fields over a single *X* field. The *Y* fields are plotted as colored lines; each is equivalent to a Plot node with Style set to Line and X Mode set to Sort. Multiplots are useful when you want to explore the fluctuation of several variables over time. For more information, see the topic [Multiplot Node](#) on p. 250.



The Web node illustrates the strength of the relationship between values of two or more symbolic (categorical) fields. The graph uses lines of various widths to indicate connection strength. You might use a Web node, for example, to explore the relationship between the purchase of a set of items at an e-commerce site. For more information, see the topic [Web Node](#) on p. 254.



The Time Plot node displays one or more sets of time series data. Typically, you would first use a Time Intervals node to create a *TimeLabel* field, which would be used to label the x axis. For more information, see the topic [Time Plot Node](#) on p. 266.



The Evaluation node helps to evaluate and compare predictive models. The evaluation chart shows how well models predict particular outcomes. It sorts records based on the predicted value and confidence of the prediction. It splits the records into groups of equal size (**quantiles**) and then plots the value of the business criterion for each quantile from highest to lowest. Multiple models are shown as separate lines in the plot. For more information, see the topic [Evaluation Node](#) on p. 269.

Once you have added a graph node to a stream, you can double-click it to open a dialog box for specifying options. Most graphs contain a number of unique options presented on one or more tabs. There are also several tab options common to all graphs. The following sections contain more information about these common options.

Once you have configured the options for a graph node, you can execute it from within the dialog box or as part of a stream. In the generated graph window, you can generate Derive (Set and Flag) and Select nodes based on a selection or region of data, effectively “subsetting” the data. For example, you might use this powerful feature to identify and exclude outliers.

Aesthetics, Overlays, Panels, and Animation

Overlays and Aesthetics

Aesthetics (and overlays) add dimensionality to a visualization. The effect of an aesthetic (grouping, clustering, or stacking) depends on the visualization type, the type of field/variable, and the graphic element type and statistic. For example, a categorical field/variable for color may be used to group points in a scatterplot or to create the stacks in a stacked bar chart. A continuous numeric range for color may also be used to indicate the range’s values for each point in a scatterplot.

You should experiment with the aesthetics and overlays to find one that fulfills your needs. The following descriptions may help you pick the right one.

Note: Not all aesthetics or overlays are available for all visualization types.

- **Color.** When color is defined by a categorical variable/field, it splits the visualization based on the individual categories, one color for each category. When color is a continuous numeric range, it varies the color based on the value of the range field/variable. If the graphic element (for example, a bar or box) represents more than one record/case and a range field/variable is used for color, the color varies based on the *mean* of the range field/variable.
- **Shape.** Shape is defined by a categorical variable/field that splits the visualization into elements of different shapes, one for each category.

- **Transparency.** When transparency is defined by a categorical variable/field, it splits the visualization based on the individual categories, one transparency level for each category. When transparency is a continuous numeric range, it varies the transparency based on the value of the range field/variable. If the graphic element (for example, a bar or box) represents more than one record/case and a range field/variable is used for transparency, the color varies based on the *mean* of the range field/variable. At the largest value, the graphic elements are fully transparent. At the smallest value, they are fully opaque.
- **Data Label.** Data labels are defined by any type of field/variable whose values are used to create labels that are attached to the graphic elements.
- **Size.** When size is defined by a categorical variable/field, it splits the visualization based on the individual categories, one size for each category. When size is a continuous numeric range, it varies the size based on the value of the range field/variable. If the graphic element (for example, a bar or box) represents more than one record/case and a range field/variable is used for size, the size varies based on the *mean* of the range field/variable.

Figure 5-1
Graph with a color overlay aesthetic

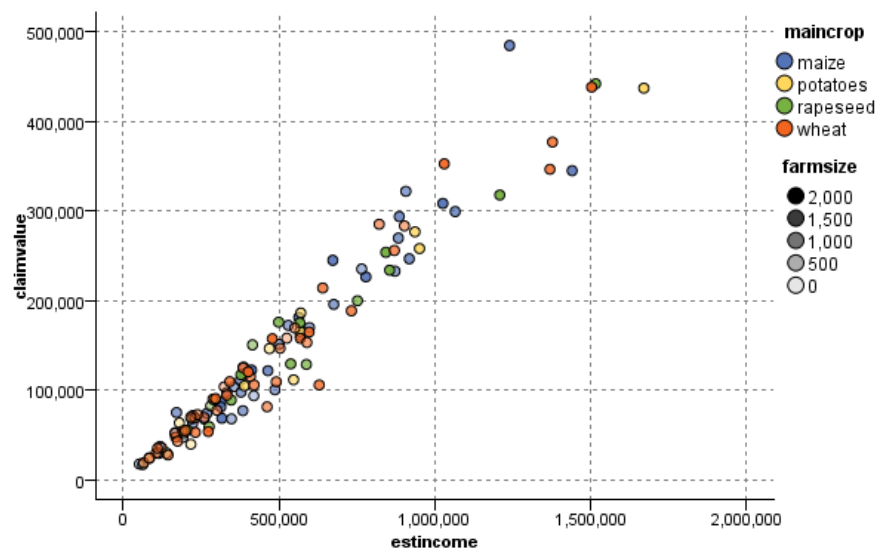
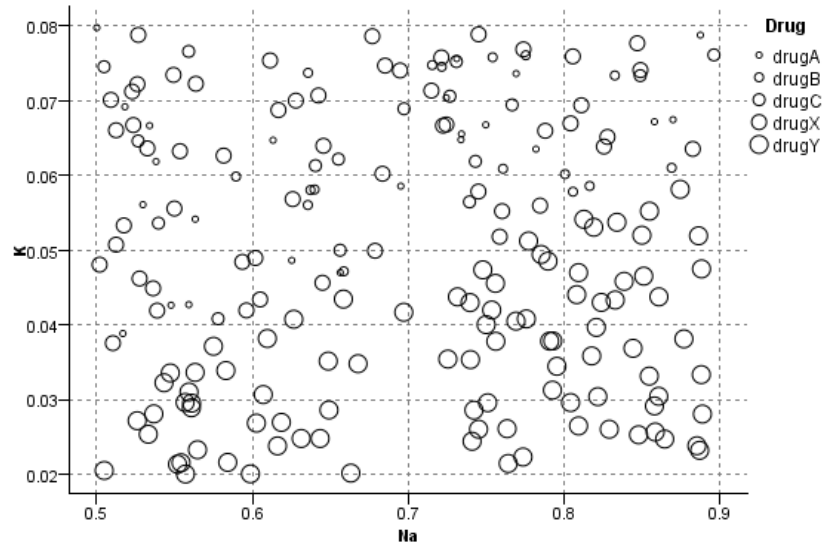


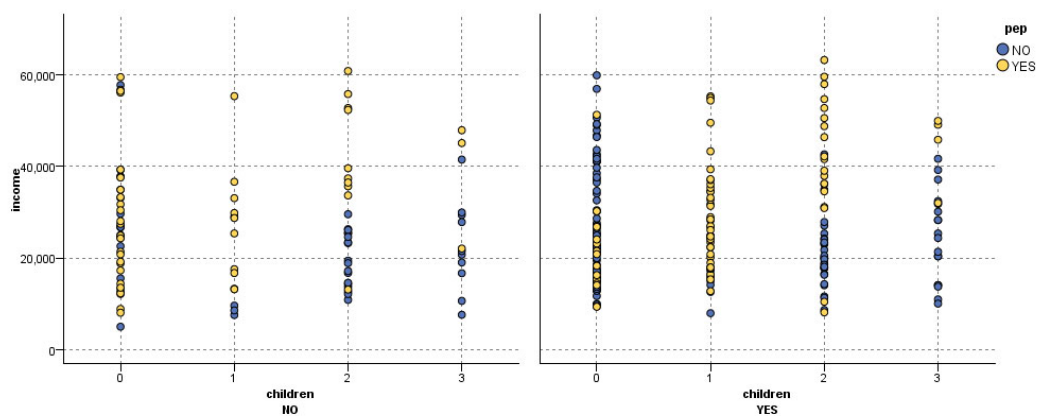
Figure 5-2
Graph with a size overlay aesthetic



Paneling and Animation

Paneling. Paneling, also known as faceting, creates a table of graphs. One graph is generated for each category in the paneling fields/variables, but all panels appear simultaneously. Paneling is useful for checking whether the visualization is subject to the conditions of the paneling fields/variables. For example, you may panel a histogram by gender to determine whether the frequency distributions are equal across males and females. That is, you can check whether salary is subject to gender differences. Select a categorical field/variable for paneling.

Figure 5-3
Graph with panels by married status (YES/NO)



Animation. Animation resembles paneling in that multiple graphs are created from the values of the animation field/variable, but these graphs are not shown together. Rather, you use the controls in Explore mode to animate the output and flip through a sequence of individual graphs. Furthermore, unlike paneling, animation does not require a categorical field/variable. You can specify a continuous field/variable whose values are split up into ranges automatically. You can vary the size of the range with the animation controls in explore mode. Not all visualizations offer animation.

Figure 5-4

Animated plot using a variable with three categories, slider at low blood pressure

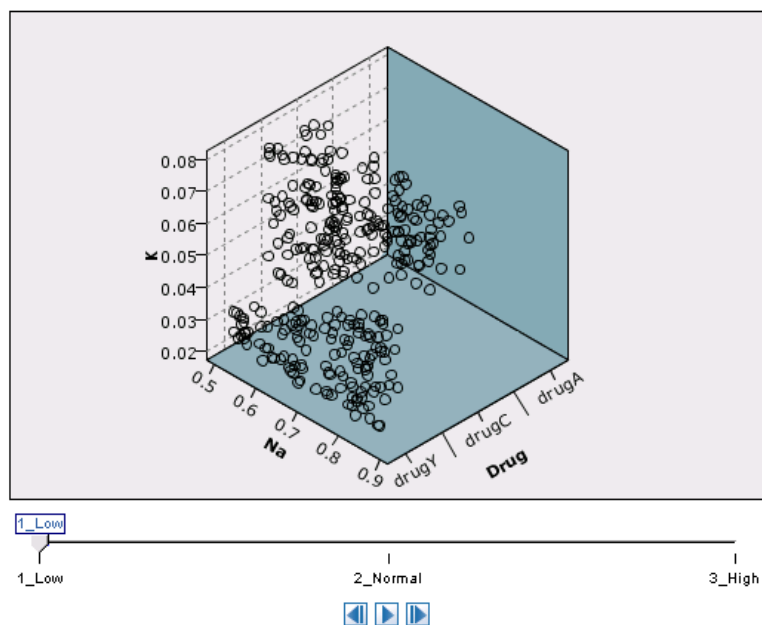


Figure 5-5
Animated plot using a variable with three categories, slider at normal blood pressure

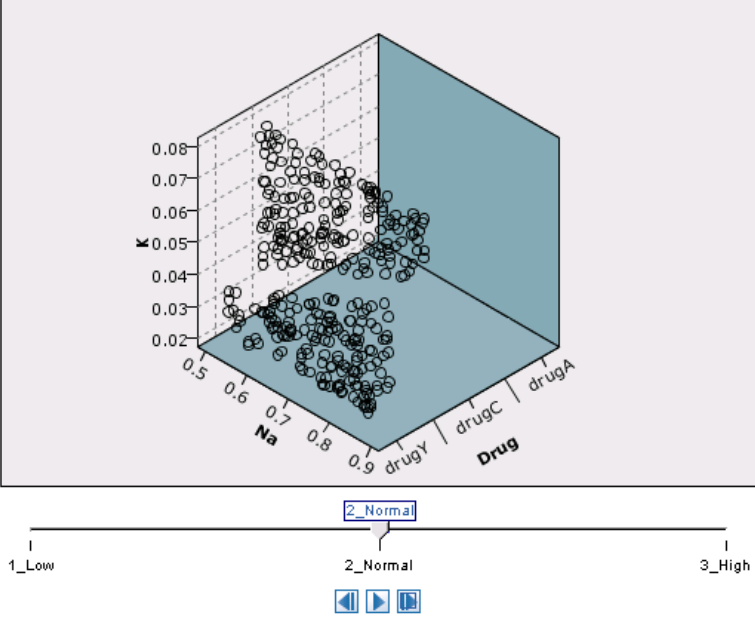
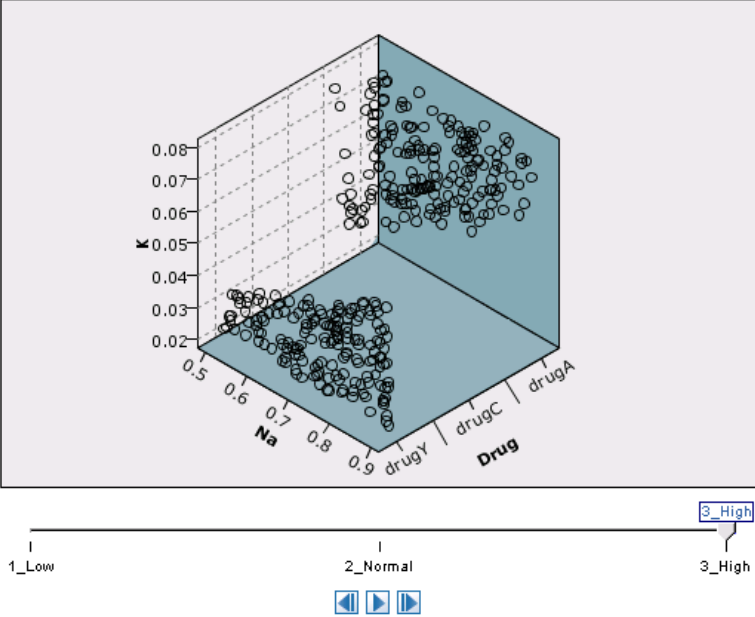


Figure 5-6
Animated plot using a variable with three categories, slider at high blood pressure



Using the Output Tab

For all graph types, you can specify the following options for the filename and display of generated graphs.

Note: Distribution node graphs have additional settings.

Output name. Specifies the name of the graph produced when the node is executed. Auto chooses a name based on the node that generates the output. Optionally, you can select Custom to specify a different name.

Output to screen. Select to generate and display the graph in a new window.

Output to file. Select to save the output as a file.

- **Output Graph.** Select to produce output in a graph format. Available only in Distribution nodes.
- **Output Table.** Select to produce output in a table format. Available only in Distribution nodes.
- **Filename.** Specify a filename used for the generated graph or table. Use the ellipsis button (...) to specify a specific file and location.
- **File type.** Specify the file type in the drop-down list. For all graph nodes, except the Distribution node with an Output Table option, the available graph file types are:

- Bitmap (.bmp)	- PNG (.png)	- Output object (.cou)
- JPEG (.jpg)	- HTML (.html)	- ViZml document (.xml) for use in other IBM® SPSS® Statistics applications.

For the Output Table option in the Distribution node, the available file types are:

- Tab delimited data (.tab)	- Comma delimited data (.csv)	- HTML (.html)	- Output object (.cou)
--------------------------------	----------------------------------	----------------	------------------------

Paginate output. When saving output as HTML, this option is enabled to allow you to control the size of each HTML page. (Applies only to the Distribution node.)

Lines per page. When Paginate output is selected, this option is enabled to allow you to determine the length of each HTML page. The default setting is 400 rows. (Applies only to the Distribution node.)

Using the Annotations Tab

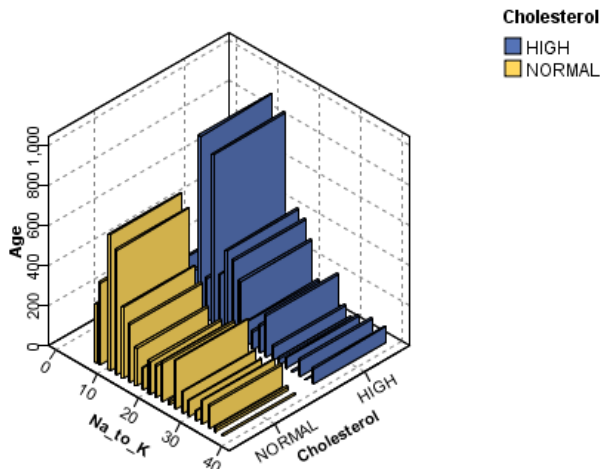
Used for all nodes, this tab offers options to rename nodes, supply a custom ToolTip, and store a lengthy annotation.

3-D Graphs

Plots and collection graphs in IBM® SPSS® Modeler have the ability to display information on a third axis. This provides you with additional flexibility in visualizing your data to select subsets or deriving new fields for modeling.

Once you have created a 3-D graph, you can click on it and drag your mouse to rotate it and view it from any angle.

Figure 5-7
Collection graph with x, y, and z axes



There are two ways of creating 3-D graphs in SPSS Modeler: plotting information on a third axis (true 3-D graphs) and displaying graphs with 3-D effects. Both methods are available for plots and collections.

To Plot Information on a Third Axis

- ▶ In the graph node dialog box, click the Plot tab.
- ▶ Click the 3-D button to enable options for the z axis.
- ▶ Use the Field Chooser button to select a field for the z axis. In some cases, only symbolic fields are allowed here. The Field Chooser will display the appropriate fields.

To Add 3-D Effects to a Graph

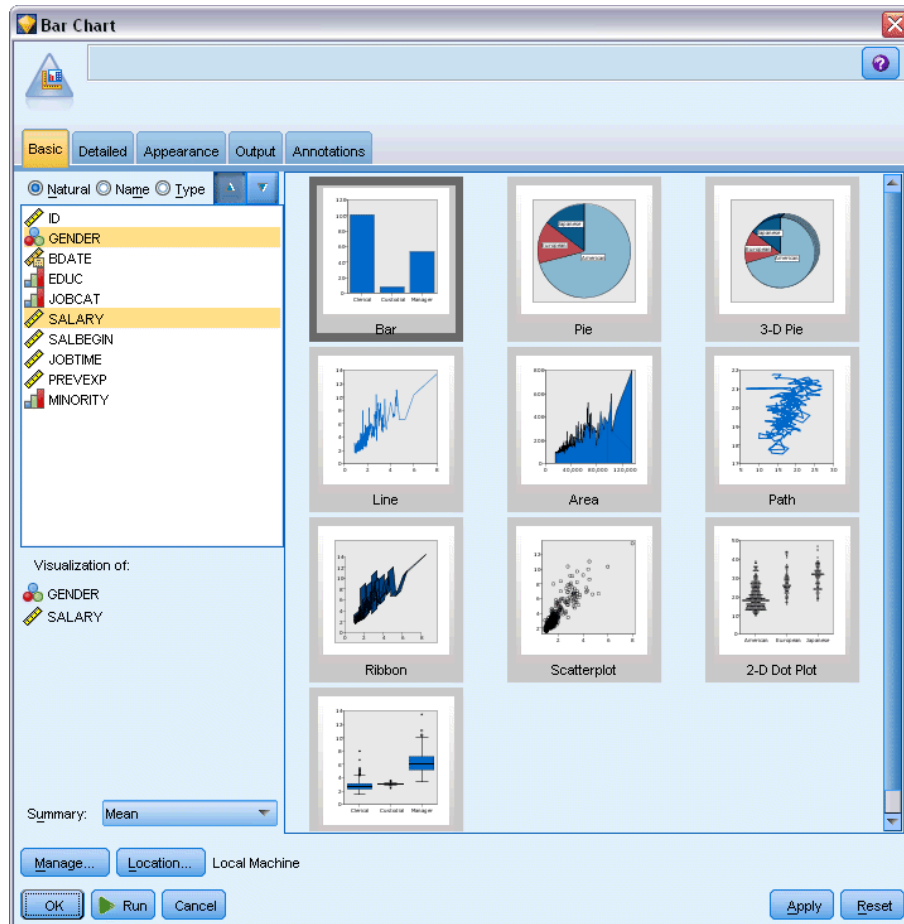
- ▶ Once you have created a graph, click the Graph tab in the output window.
- ▶ Click the 3-D button to switch the view to a three-dimensional graph.

Graphboard Node

The Graphboard node allows you to choose from many different graphs outputs (bar charts, pie charts, histograms, scatterplots, heatmaps, etc.) in one single node. You begin, in the first tab, by choosing the data fields you want to explore, and then the node presents you with a choice of graph types that work for your data. The node automatically filters out any graph types that would not work with the field choices. You can define detailed, or more advanced graph options in the Detailed tab.

Note: You must connect the Graphboard node to a stream with data in order to edit the node or select graph types.

Figure 5-8
Graphboard node dialog box, Basic tab



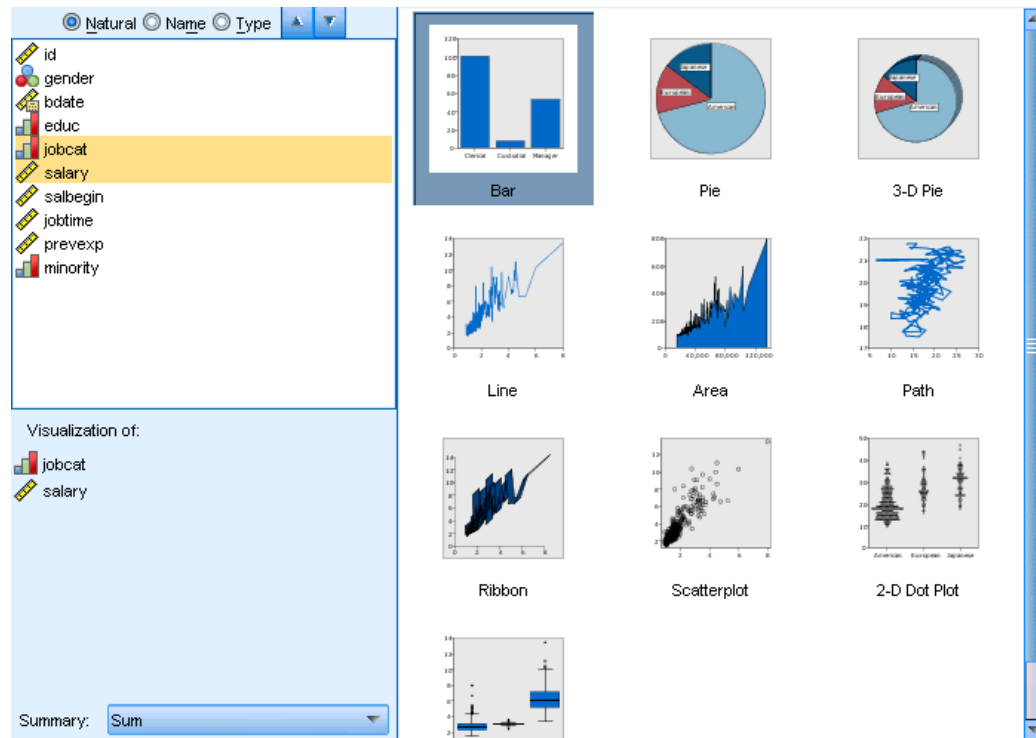
There are two buttons that allow you to control which visualization templates (and stylesheets) are available:

Manage. Manage visualization templates and stylesheets on your computer. You can import, export, rename, and delete visualization templates and stylesheets on your local machine. For more information, see the topic [Managing Templates and Stylesheets](#) on p. 227.

Location. Change the location in which visualization templates and stylesheets are stored. The current location is listed to the right of the button. For more information, see the topic [Setting the Location of Templates and Stylesheets](#) on p. 225.

Graphboard Basic Tab

Figure 5-9
Basic tab



If you aren't sure which visualization type would best represent your data, use the Basic tab. When you select your data, you are then presented with a subset of visualization types that are appropriate for the data.

- ▶ Select one or more fields/variables from the list. Use Ctrl+Click to select multiple fields/variables.

Note that the measurement level of the variable determines the type of visualizations that are available. You can change the measurement level by right-clicking the field/variable in the list and choosing an option. For more information about the available measurement level types, see [Variable Types](#) on p. 220.

- ▶ Select a visualization type.
- ▶ For certain visualizations, you can choose a summary statistic. Different subsets of statistics are available depending on whether the statistic is count-based or calculated from a continuous field/variable. The available statistics also depend on the template itself. A full list of statistics that may be available follows the next step.
- ▶ If you want to define more options, such as optional aesthetics and panel fields/variables, click Detailed. For more information, see the topic [Graphboard Detailed Tab](#) on p. 223.

Summary Statistics Calculated from a Continuous Field/Variable

- **Mean.** A measure of central tendency. The arithmetic average, the sum divided by the number of cases.
- **Median.** The value above and below which half of the cases fall, the 50th percentile. If there is an even number of cases, the median is the average of the two middle cases when they are sorted in ascending or descending order. The median is a measure of central tendency not sensitive to outlying values (unlike the mean, which can be affected by a few extremely high or low values).
- **Mode.** The most frequently occurring value. If several values share the greatest frequency of occurrence, each of them is a mode.
- **Minimum.** The smallest value of a numeric variable.
- **Maximum.** The largest value of a numeric variable.
- **Range.** The difference between the minimum and maximum values.
- **Mid Range.** The middle of the range, that is, the value whose difference from the minimum is equal to its difference from the maximum.
- **Sum.** The sum or total of the values, across all cases with nonmissing values.
- **Cumulative Sum.** The cumulative sum of the values. Each graphic element shows the sum for one subgroup plus the total sum of all previous groups.
- **Percent Sum.** The percentage within each subgroup based on a summed variable compared to the sum across all groups.
- **Cumulative Percent Sum.** The cumulative percentage within each subgroup based on a summed variable compared to the sum across all groups. Each graphic element shows the percentage for one subgroup plus the total percentage of all previous groups.
- **Variance.** A measure of dispersion around the mean, equal to the sum of squared deviations from the mean divided by one less than the number of cases. The variance is measured in units that are the square of those of the variable itself.
- **Standard Deviation.** A measure of dispersion around the mean. In a normal distribution, 68% of cases fall within one standard deviation of the mean and 95% of cases fall within two standard deviations. For example, if the mean age is 45, with a standard deviation of 10, 95% of the cases would be between 25 and 65 in a normal distribution.
- **Standard Error.** A measure of how much the value of a test statistic varies from sample to sample. It is the standard deviation of the sampling distribution for a statistic. For example, the standard error of the mean is the standard deviation of the sample means.
- **Kurtosis.** A measure of the extent to which observations cluster around a central point. For a normal distribution, the value of the kurtosis statistic is zero. Positive kurtosis indicates that, relative to a normal distribution, the observations are more clustered about the center of the distribution and have thinner tails until the extreme values of the distribution, at which point the tails of the leptokurtic distribution are thicker relative to a normal distribution. Negative kurtosis indicates that, relative to a normal distribution, the observations cluster less and

have thicker tails until the extreme values of the distribution, at which point the tails of the platykurtic distribution are thinner relative to a normal distribution.

- **Skewness.** A measure of the asymmetry of a distribution. The normal distribution is symmetric and has a skewness value of 0. A distribution with a significant positive skewness has a long right tail. A distribution with a significant negative skewness has a long left tail. As a guideline, a skewness value more than twice its standard error is taken to indicate a departure from symmetry.

The following region statistics may result in more than one graphic element per subgroup. When using the interval, area, or edge graphic elements, a region statistic results in one graphic element showing the range. All other graphic elements result in two separate elements, one showing the start of the range and one showing the end of the range.




- **Region: Range.** The range of values between the minimum and maximum values.
- **Region: 95% Confidence Interval of Mean.** A range of values that has a 95% chance of including the population mean.
- **Region: 95% Confidence Interval of Individual.** A range of values that has a 95% chance of including the predicted value given the individual case.
- **Region: 1 Standard Deviation above/below Mean.** A range of values between 1 **standard deviation** above and below the **mean**.
- **Region: 1 Standard Error above/below Mean.** A range of values between 1 **standard error** above and below the **mean**.











Count-Based Summary Statistics

- **Count.** The number of rows/cases.
- **Cumulative Count.** The cumulative number of rows/cases. Each graphic element shows the count for one subgroup plus the total count of all previous groups.
- **Percent of Count.** The percentage of rows/cases in each subgroup compared to the total number of rows/cases.
- **Cumulative Percent of Count.** The cumulative percentage of rows/cases in each subgroup compared to the total number of rows/cases. Each graphic element shows the percentage for one subgroup plus the total percentage of all previous groups.

Variable Types

Icons appear next to variables in variable lists to indicate the variable type and data type. Icons also identify multiple response sets.

Measurement Level	Data Type			
	Numeric	String	Date	Time
Continuous		n/a		

Ordered Set				
Set				
Multiple response set, multiple categories				
Multiple response set, multiple dichotomies				

Measurement Level

A variable's measurement level is important when you create a visualization. Following is a description of the measurement levels. You can temporarily change the measurement level by right-clicking a variable in a variable list and choosing an option. In most cases, you need to consider only the two broadest classifications of variables, categorical and continuous:

Categorical. Data with a limited number of distinct values or categories (for example, gender or religion). Categorical fields/variables can be string (alphanumeric) or numeric variables that use numeric codes to represent categories (for example, 0 = *male* and 1 = *female*). Also referred to as qualitative data. Sets, ordered sets, and flags are all categorical variables.

-
-
-

Continuous. Data measured on an interval or ratio scale, where the data values indicate both the order of values and the distance between values. For example, a salary of \$72,195 is higher than a salary of \$52,398, and the distance between the two values is \$19,797. Also referred to as quantitative, scale, or numeric range data.

Categorical variables define categories in the visualization, typically to draw separate graphic elements or to group graphic elements. Continuous variables are often summarized within categories of categorical variables. For example, a default visualization of income for gender categories would display the mean income for males and the mean income for females. The raw values for continuous variables can also be plotted, as in a scatterplot. For example, a scatterplot may show the current salary and beginning salary for each case. A categorical variable could be used to group the cases by gender.

Data Types

Measurement level isn't the only property of a variable that determines its type. A variable is also stored as a specific data type. Possible data types are strings (non-numeric data such as letters), numeric values (real numbers), and dates. Unlike the measurement level, a variable's data type cannot be changed temporarily. You must change the way the data are stored in the original data set.

Multiple Response Sets

Some data files support a special kind of “variable” called a **multiple response set**. Multiple response sets aren’t really “variables” in the normal sense. Multiple response sets use multiple variables to record responses to questions where the respondent can give more than one answer. Multiple response sets are treated like categorical variables, and most of the things you can do with categorical variables, you can also do with multiple response sets.

Multiple response sets can be multiple dichotomy sets or multiple category sets.

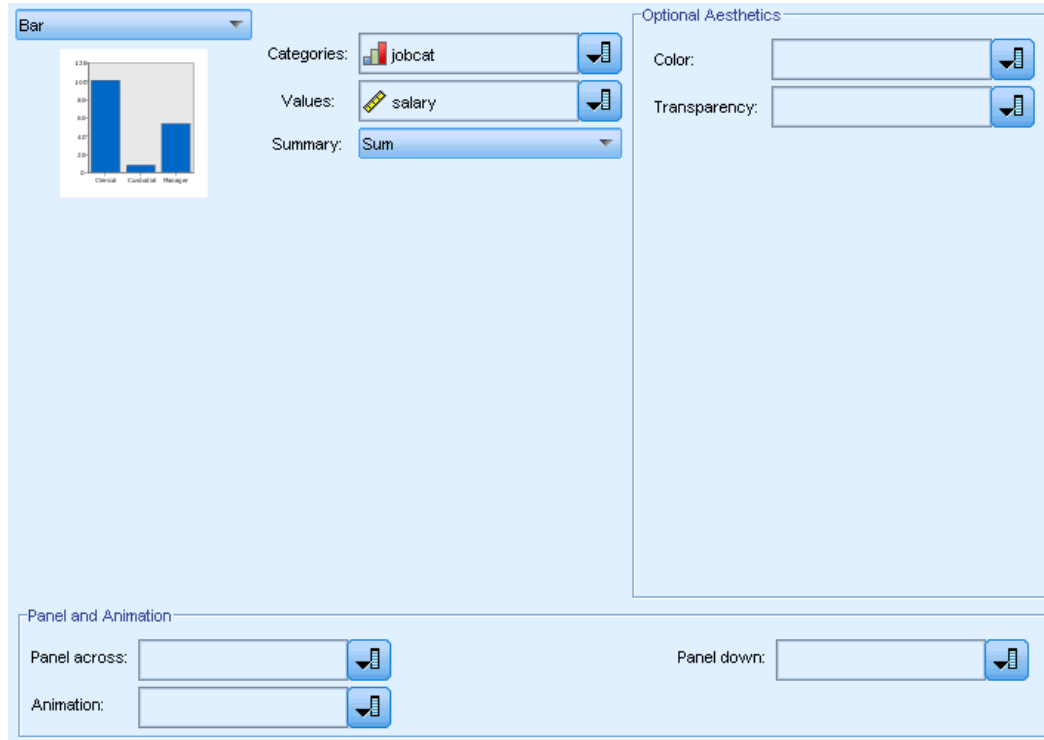
Multiple dichotomy sets. A multiple dichotomy set typically consists of multiple dichotomous variables: variables with only two possible values of a yes/no, present/absent, checked/not checked nature. Although the variables may not be strictly dichotomous, all of the variables in the set are coded the same way.

For example, a survey provides five possible responses to the question, “Which of the following sources do you rely on for news?” The respondent can indicate multiple choices by checking a box next to each choice. The five responses become five variables in the data file, coded 0 for *No* (not checked) and 1 for *Yes* (checked).

Multiple category sets. A multiple category set consists of multiple variables, all coded the same way, often with many possible response categories. For example, a survey item states, “Name up to three nationalities that best describe your ethnic heritage.” There may be hundreds of possible responses, but for coding purposes the list is limited to the 40 most common nationalities, with everything else relegated to an “other” category. In the data file, the three choices become three variables, each with 41 categories (40 coded nationalities and one “other” category).

Graphboard Detailed Tab

Figure 5-10
Detailed tab



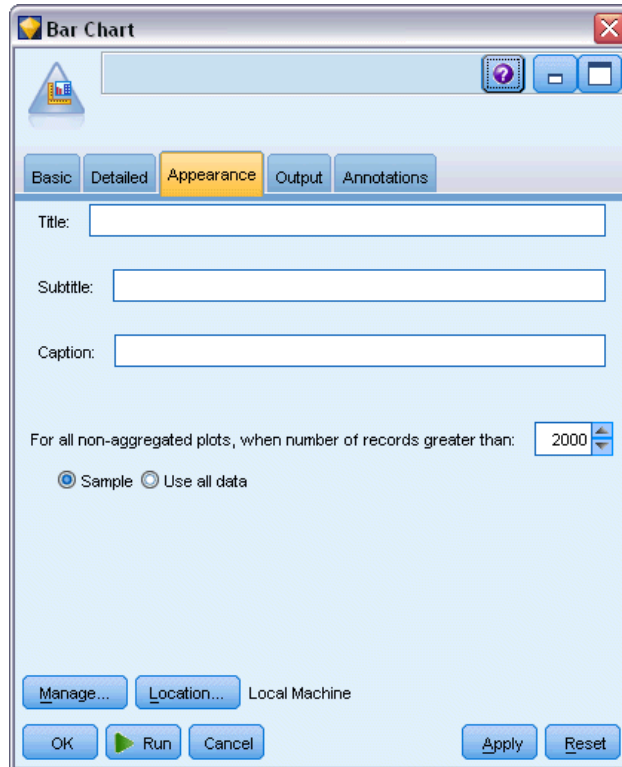
Use the Detailed tab when you know what type of visualization you want to create or when you want to add optional aesthetics, panels, and/or animation to a visualization.

- ▶ If you selected a visualization type on the Basic tab, it will be displayed. Otherwise, choose one from the drop-down list.
- ▶ To the immediate right of the visualization's thumbnail image are controls for specifying the variables/fields required for the visualization type. You must specify all of these fields/variables.
- ▶ For certain visualizations, you can select a summary statistic. In some cases (such as with bar charts), you can use one of these summary options for the transparency aesthetic. For descriptions of the summary statistics, see [Graphboard Basic Tab](#) on p. 218.
- ▶ You can select one or more of the optional aesthetics. These can add dimensionality by allowing you to include other fields/variables in the visualization. For example, you may use a field/variable to vary the size of points in a scatterplot. For more information about optional aesthetics, see [Aesthetics, Overlays, Panels, and Animation](#) on p. 210. Please note that the transparency aesthetic is not supported through scripting.
- ▶ You can select one or more of the paneling or animation options. For more information about paneling and animation options, see [Aesthetics, Overlays, Panels, and Animation](#) on p. 210.

Graphboard Appearance Tab

You can specify appearance options before graph creation.

Figure 5-12
Appearance tab settings for a Graphboard node



General Appearance Options

Title. Enter the text to be used for the graph's title.

Subtitle. Enter the text to be used for the graph's subtitle.

Caption. Enter the text to be used for the graph's caption.

Sampling. Specify a method for larger datasets. You can specify a maximum dataset size or use the default number of records. Performance is enhanced for large datasets when you select the Sample option. Alternatively, you can choose to plot all data points by selecting Use all data, but you should note that this may dramatically decrease the performance of the software.

Stylesheet Appearance Options

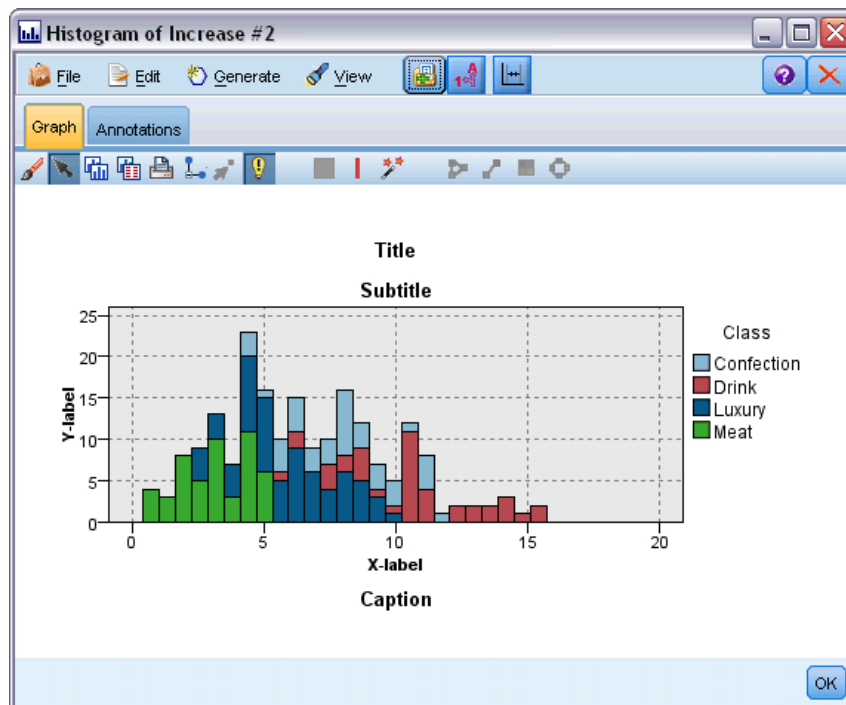
There are two buttons that allow you to control which visualization templates (and stylesheets) are available:

Manage. Manage visualization templates and stylesheets on your computer. You can import, export, rename, and delete visualization templates and stylesheets on your local machine. For more information, see the topic [Managing Templates and Stylesheets](#) on p. 227.

Location. Change the location in which visualization templates and stylesheets are stored. The current location is listed to the right of the button. For more information, see the topic [Setting the Location of Templates and Stylesheets](#) on p. 225.

The following example shows where appearance options are placed on a graph. (*Note:* Not all graphs use all these options.)

Figure 5-13
Position of various graph appearance options

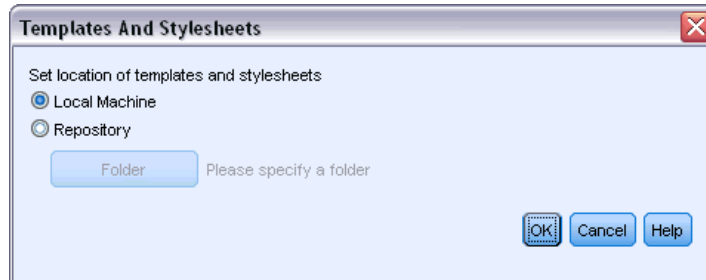


Setting the Location of Templates and Stylesheets

Visualization templates and stylesheets are stored in a specific local folder or in the IBM® SPSS® Collaboration and Deployment Services Repository. When selecting templates and stylesheets, only the built-in templates/stylesheets and templates/stylesheets in this location are displayed. By keeping all templates and stylesheets in one place, IBM Corp. applications can easily access them. For information about adding additional templates and stylesheets to this location, see [Managing Templates and Stylesheets](#) on p. 227.

How to Set the Location of Templates and Stylesheets

Figure 5-14
Templates and Stylesheets dialog box



- ▶ In a template or stylesheet dialog box, click Location... to display the Templates and Stylesheets dialog box.
- ▶ Select an option for the default location for templates and stylesheets:
 - Local Machine.** Templates and stylesheets are located in a specific folder on your local computer. On Windows XP, this folder is *C:\Documents and Settings\<user>\Application Data\SPSSInc\Graphboard*. The folder cannot be changed.
 - IBM® SPSS® Collaboration and Deployment Services Repository.** Templates and stylesheets are located in a user-specified folder in the IBM SPSS Collaboration and Deployment Services Repository. To identify the specific folder, click Folder. For more information, see [Using the IBM SPSS Collaboration and Deployment Services Repository as the Template and Stylesheet Location](#) on p. 226.
- ▶ Click OK.

Using the IBM SPSS Collaboration and Deployment Services Repository as the Template and Stylesheet Location

Visualization templates and stylesheets can be stored in the IBM® SPSS® Collaboration and Deployment Services Repository. This location is a specific folder in the IBM SPSS Collaboration and Deployment Services Repository. If this is set as the default location, any templates and stylesheets in this location are available for selection.

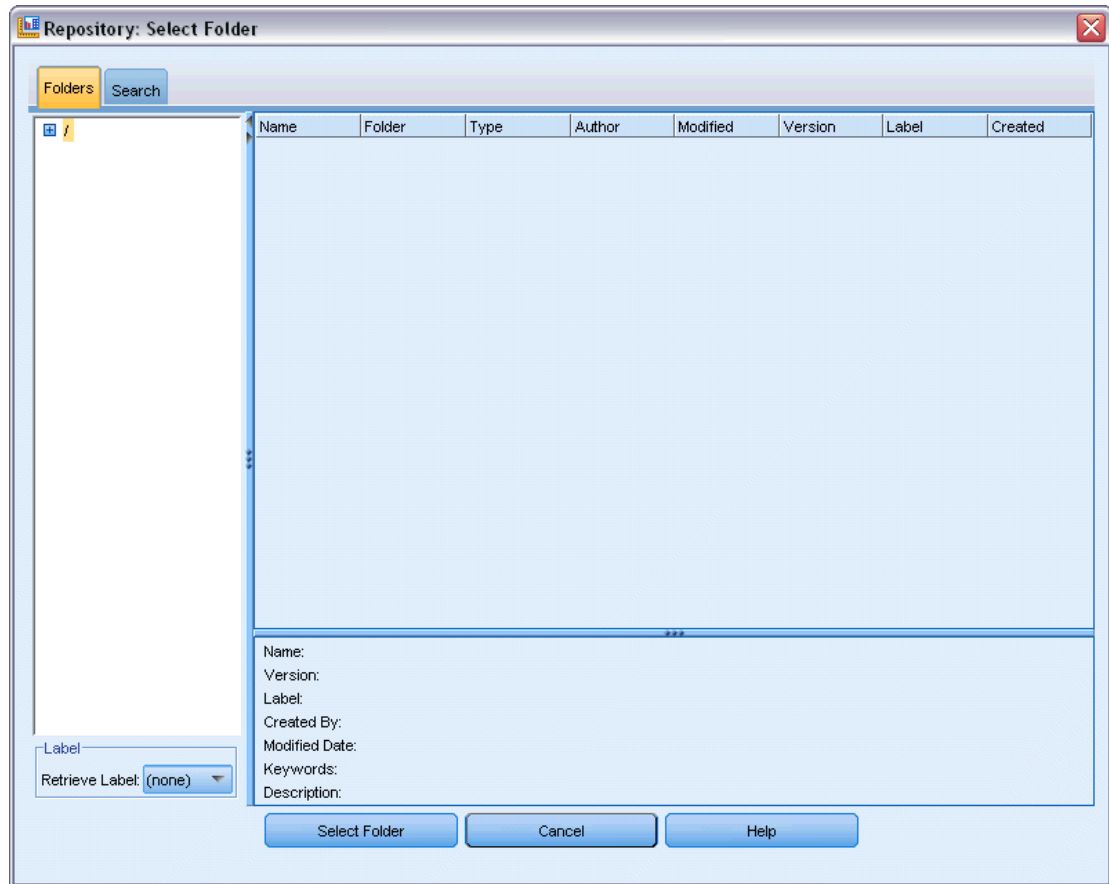
How to Set a Folder in IBM SPSS Collaboration and Deployment Services Repository as the Location for Templates and Stylesheets

- ▶ In a dialog box with a Location button, click Location....
- ▶ Select IBM® SPSS® Collaboration and Deployment Services Repository.
- ▶ Click Folder.

Note: If you are not already connected to the IBM SPSS Collaboration and Deployment Services Repository, you will be prompted for connection information.

- ▶ In the Select Folder dialog box, select the folder in which templates and stylesheets are stored.

Figure 5-15
Select Folder dialog box



- ▶ If desired, select a label from Retrieve Label. Only templates and stylesheets with that label will be displayed.
- ▶ If you are looking for a folder that contains a particular template or stylesheet, you may want to search for the template or stylesheet on the Search tab. The Select Folder dialog box automatically selects the folder in which the found template or stylesheet is located.
- ▶ Click Select Folder.

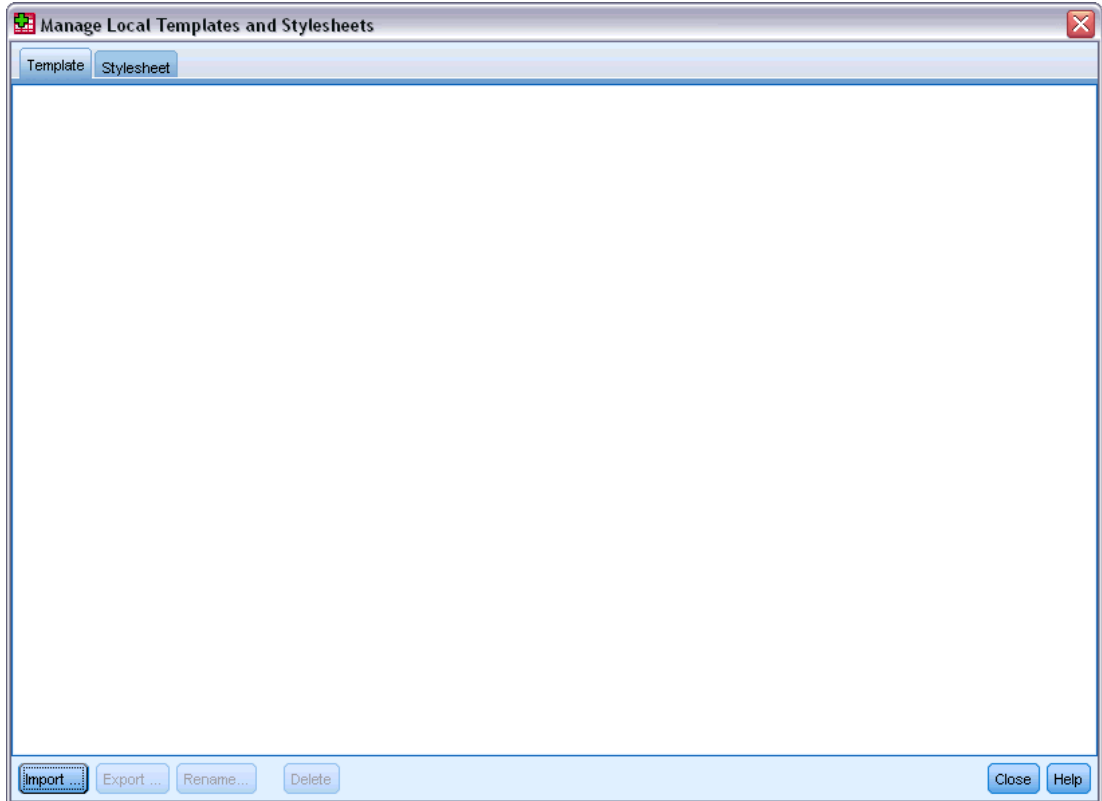
Managing Templates and Stylesheets

You can manage the templates and stylesheets in the local location on your computer by using the Manage Templates and Stylesheets dialog box. This dialog box allows you to import, export, rename, and delete visualization templates and stylesheets in the local location on your computer.

- ▶ Click Manage... in one of the dialog boxes where you select templates or stylesheets.

Manage Templates and Stylesheets Dialog Box

Figure 5-16
Manage Templates and Stylesheets dialog box



The Templates tab lists all the local templates. The Stylesheets tab lists all the local stylesheets, as well as displaying example visualizations with sample data. You can select one of the stylesheets to apply its styles to the example visualizations. For more information, see the topic [Applying Stylesheets](#) on p. 310.

The following buttons operate on whichever tab is currently activated.

Import. Import a visualization template or stylesheet from the file system. Importing a template or stylesheet makes it available to the IBM Corp. application. If another user sent you a template or stylesheet, you import the template or stylesheet before using it in your application.

Export. Export a visualization template or stylesheet to the file system. Export a template or stylesheet when you want to send it to another user.

Rename. Rename the selected visualization template or stylesheet. You cannot change a template name to one that's already used.

Delete. Delete the selected visualization template(s) or stylesheet(s). You can select multiple templates stylesheets with Ctrl-Click. There is no undo action for deleting so proceed with caution.

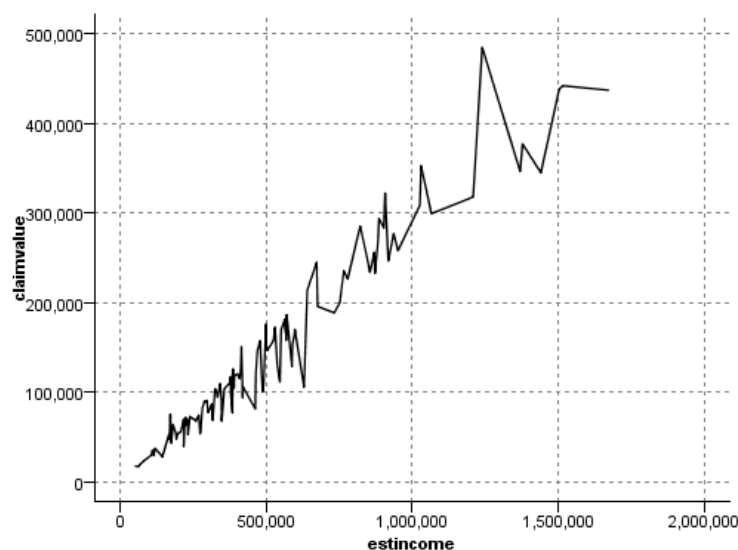
Plot Node

Plot nodes show the relationship between numeric fields. You can create a plot using points (also known as a scatterplot), or you can use lines. You can create three types of line plots by specifying an X Mode in the dialog box.

X Mode = Sort

Setting X Mode to Sort causes data to be sorted by values for the field plotted on the x axis. This produces a single line running from left to right on the graph. Using a nominal field as an overlay produces multiple lines of different hues running from left to right on the graph.

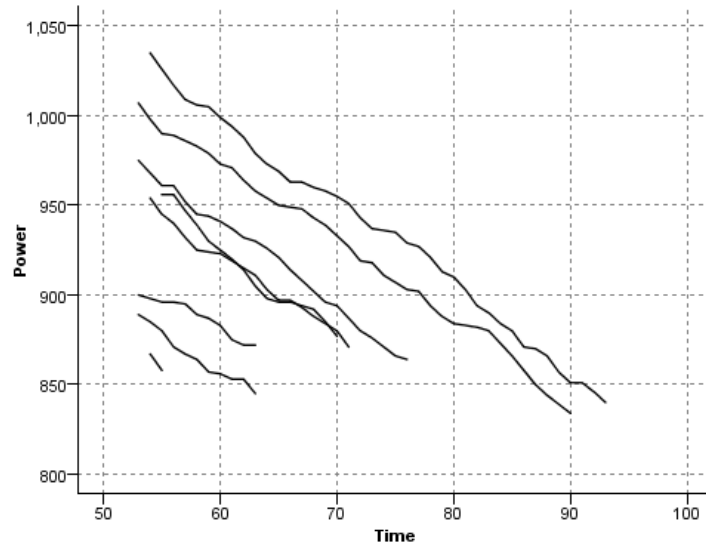
Figure 5-17
Line plot with X Mode set to Sort



X Mode = Overlay

Setting X Mode to Overlay creates multiple line plots on the same graph. Data are not sorted for an overlay plot; as long as the values on the x axis increase, data will be plotted on a single line. If the values decrease, a new line begins. For example, as x moves from 0 to 100, the y values will be plotted on a single line. When x falls below 100, a new line will be plotted in addition to the first one. The finished plot might have numerous plots useful for comparing several series of y values. This type of plot is useful for data with a periodic time component, such as electricity demand over successive 24-hour periods.

Figure 5-18
Line plot with X Mode set to Overlay

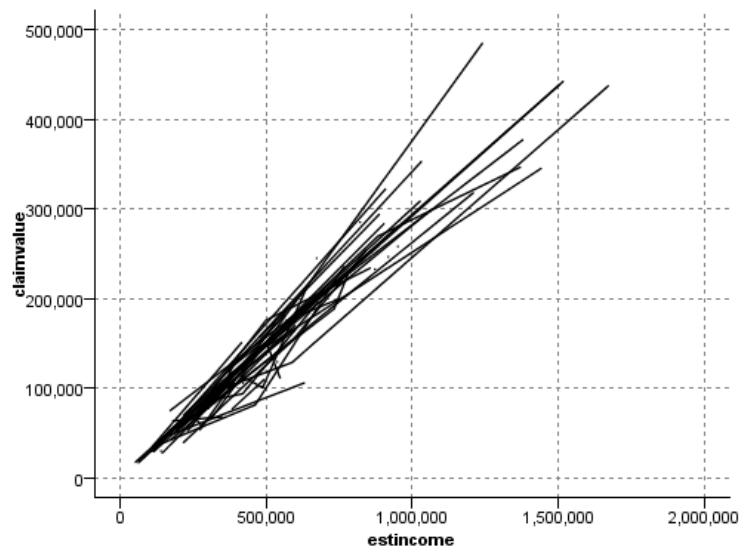


X Mode = As Read

Setting X Mode to As Read plots x and y values as they are read from the data source. This option is useful for data with a time series component where you are interested in trends or patterns that depend on the order of the data. You may need to sort the data before creating this type of plot. It may also be useful to compare two similar plots with X Mode set to Sort and As Read in order to determine how much of a pattern depends on the sorting.

Figure 5-19

Line plot shown earlier as Sort, executed again with X Mode set to As Read

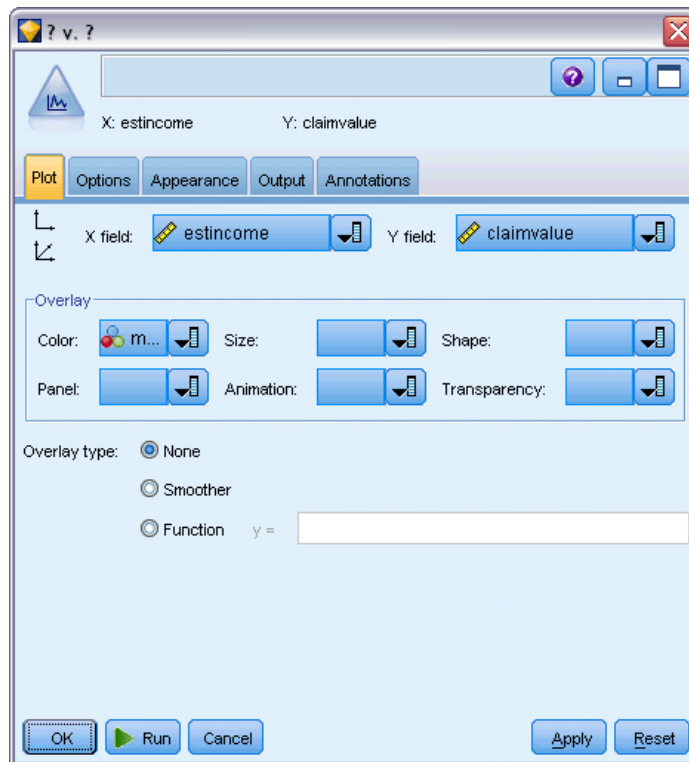


You can also use the Graphboard node to produce scatterplots and line plots. However, you have more options to choose from in this node.

Plot Node Tab

Plots show values of a *Y* field against values of an *X* field. Often, these fields correspond to a dependent variable and an independent variable, respectively.

Figure 5-20
Plot tab settings for a Plot node



X field. From the list, select the field to display on the horizontal x axis.

Y field. From the list, select the field to display on the vertical y axis.

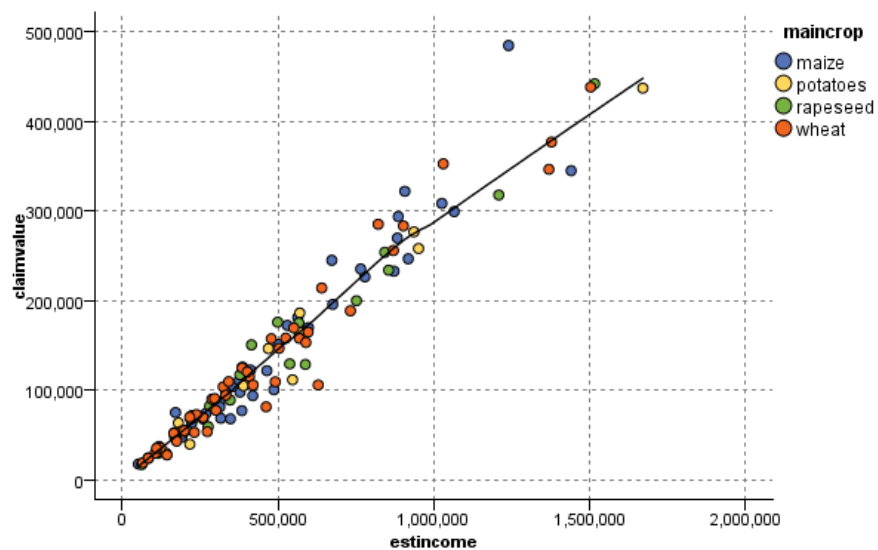
Z field. When you click the 3-D chart button, you can then choose a field from the list to display on the z axis.

Overlay. There are several ways to illustrate categories for data values. For example, you can use *maincrop* as a color overlay to indicate the *estincome* and *claimvalue* values for the main crop grown by claim applicants. For more information, see the topic [Aesthetics, Overlays, Panels, and Animation](#) on p. 210.

Overlay type. Specifies whether an overlay function or smoother is displayed. The smoother and overlay functions are always calculated as a function of y .

- **None.** No overlay is displayed.
- **Smoother.** Displays a smoothed fit line computed using locally weighted iterative robust least squares regression (LOESS). This method effectively computes a series of regressions, each focused on a small area within the plot. This produces a series of “local” regression lines that are then joined to create a smooth curve.

Figure 5-21
Plot with a LOESS smoother overlay



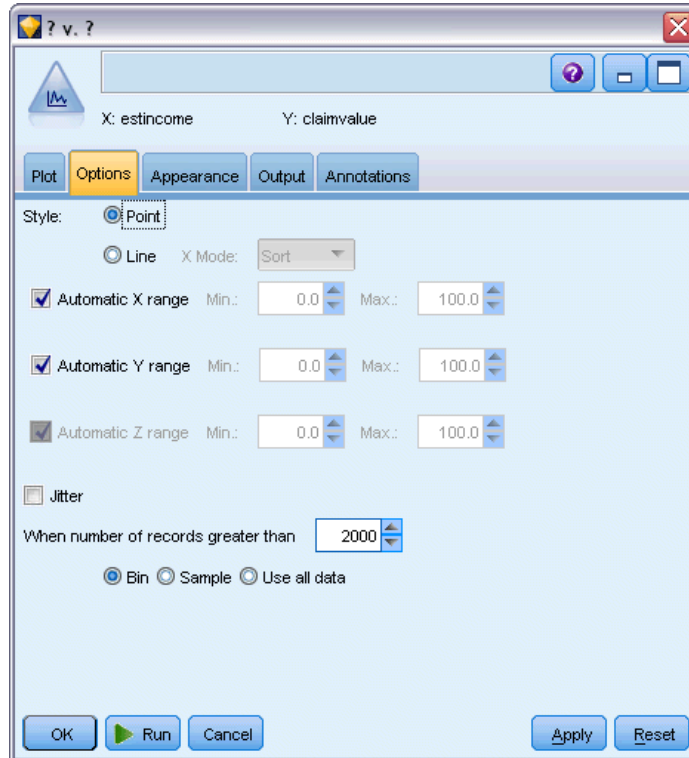
- **Function.** Select to specify a known function to compare to actual values. For example, to compare actual versus predicted values, you can plot the function $y = x$ as an overlay. Specify a function for $y =$ in the text box. The default function is $y = x$, but you can specify any sort of function, such as a quadratic function or an arbitrary expression, in terms of x .

Note: Overlay functions are not available for a panel or animation graph.

Once you have set options for a plot, you can run the plot directly from the dialog box by clicking Run. You may, however, want to use the Options tab for additional specifications, such as binning, X Mode, and style.

Plot Options Tab

Figure 5-22
Options tab settings for a Plot node



Style. Select either Point or Line for the plot style. Selecting Line activates the X Mode control. Selecting Point will use a plus symbol (+) as the default point shape. Once the graph is created, you can change the point shape and alter its size.

X Mode. For line plots, you must choose an X Mode to define the style of the line plot. Select Sort, Overlay, or As Read. For Overlay or As Read, you should specify a maximum dataset size used to sample the first n records. Otherwise, the default 2,000 records will be used.

Automatic X range. Select to use the entire range of values in the data along this axis. Deselect to use an explicit subset of values based on your specified Min and Max values. Either enter values or use the arrows. Automatic ranges are selected by default to enable rapid graph building.

Automatic Y range. Select to use the entire range of values in the data along this axis. Deselect to use an explicit subset of values based on your specified Min and Max values. Either enter values or use the arrows. Automatic ranges are selected by default to enable rapid graph building.

Automatic Z range. Only when a 3-D graph is specified on the Plot tab. Select to use the entire range of values in the data along this axis. Deselect to use an explicit subset of values based on your specified Min and Max values. Either enter values or use the arrows. Automatic ranges are selected by default to enable rapid graph building.

Jitter. Also known as **agitation**, jitter is useful for point plots of a dataset in which many values are repeated. In order to see a clearer distribution of values, you can use jitter to distribute the points randomly around the actual value.

Note to users of earlier versions of SPSS Modeler: The jitter value used in a plot uses a different metric in this release of IBM® SPSS® Modeler. In earlier versions, the value was an actual number, but it is now a proportion of the frame size. This means that agitation values in old streams are likely to be too large. For this release, any nonzero agitation values will be converted to the value 0.2.

Maximum number of records to plot. Specify a method for plotting large datasets. You can specify a maximum dataset size or use the default 2,000 records. Performance is enhanced for large datasets when you select the Bin or Sample options. Alternatively, you can choose to plot all data points by selecting Use all data, but you should note that this may dramatically decrease the performance of the software.

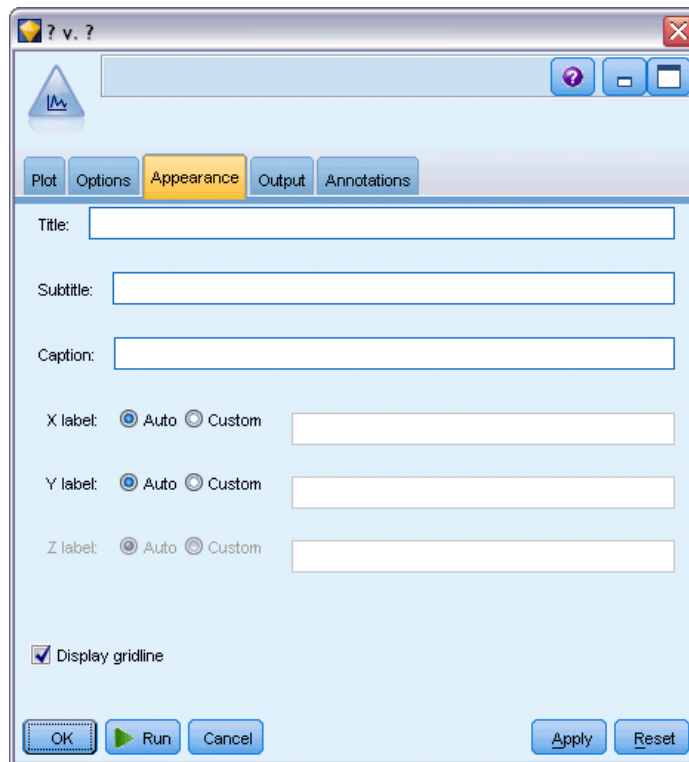
Note: When X Mode is set to Overlay or As Read, these options are disabled and only the first n records are used.

- **Bin.** Select to enable binning when the dataset contains more than the specified number of records. Binning divides the graph into fine grids before actually plotting and counts the number of points that would appear in each of the grid cells. In the final graph, one point is plotted per cell at the bin centroid (average of all point locations in the bin). The size of the plotted symbols indicates the number of points in that region (unless you have used size as an overlay). Using the centroid and size to represent the number of points makes the binned plot a superior way to represent large datasets, because it prevents overplotting in dense regions (undifferentiated masses of color) and reduces symbol artifacts (artificial patterns of density). Symbol artifacts occur when certain symbols (particularly the plus symbol [+]) collide in a way that produces dense areas not present in the raw data.
- **Sample.** Select to randomly sample the data to the number of records entered in the text field. The default is 2,000.

Plot Appearance Tab

You can specify appearance options before graph creation.

Figure 5-23
Appearance tab settings for a Plot node



Title. Enter the text to be used for the graph's title.

Subtitle. Enter the text to be used for the graph's subtitle.

Caption. Enter the text to be used for the graph's caption.

X label. Either accept the automatically generated x -axis (horizontal) label or select Custom to specify a label.

Y label. Either accept the automatically generated y -axis (vertical) label or select Custom to specify a label.

Z label. Available only for 3-D graphs, either accept the automatically generated z -axis label or select Custom to specify a custom label.

Display gridline. Selected by default, this option displays a gridline behind the plot or graph that enables you to more easily determine region and band cutoff points. Gridlines are always displayed in white unless the graph background is white; in this case, they are displayed in gray.

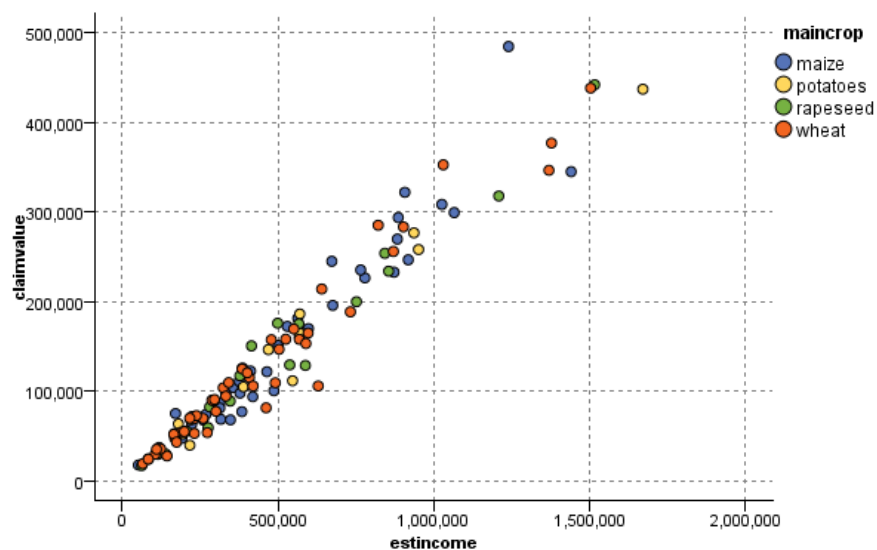
Using a Plot Graph

Plots and multiplots are essentially plots of X against Y . For example, if you are exploring potential fraud in agricultural grant applications (as illustrated in *fraud.str* in the *Demos* folder of your IBM® SPSS® Modeler installation), you might want to plot the income claimed on the

application versus the income estimated by a neural net. Using an overlay, such as crop type, will illustrate whether there is a relationship between claims (value or number) and type of crop.

Figure 5-24

Plot of the relationship between estimated income and claim value with main crop type as an overlay



Since plots, multiplots, and evaluation charts are two-dimensional displays of Y against X , it is easy to interact with them by defining regions, marking elements, or even drawing bands. You can also generate nodes for the data represented by those regions, bands, or elements. For more information, see the topic [Exploring Graphs](#) on p. 281.

Distribution Node

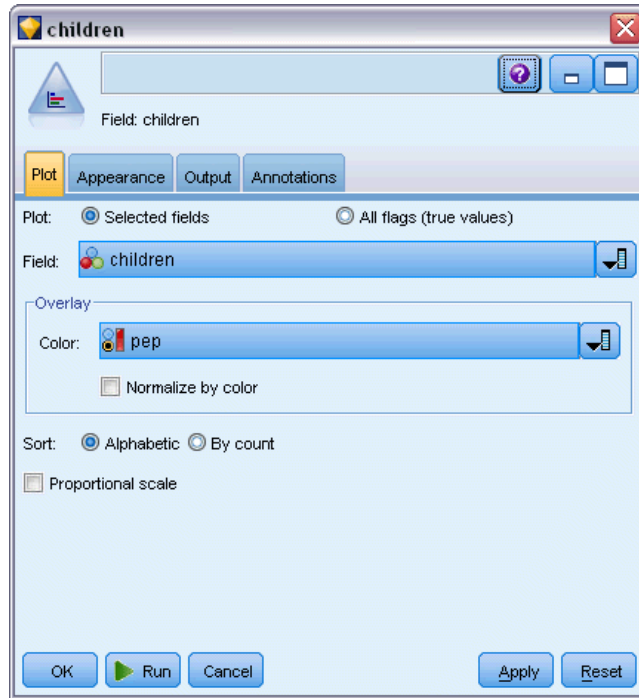
A distribution graph or table shows the occurrence of symbolic (non-numeric) values, such as mortgage type or gender, in a dataset. A typical use of the Distribution node is to show imbalances in the data that can be rectified by using a Balance node before creating a model. You can automatically generate a Balance node using the Generate menu in the distribution graph or table window.

You can also use the Graphboard node to produce bar of counts graphs. However, you have more options to choose from in this node.

Note: To show the occurrence of numeric values, you should use a Histogram node.

Distribution Plot Tab

Figure 5-25
Plot tab settings for a Distribution node



Plot. Select the type of distribution. Select *Selected fields* to show the distribution of the selected field. Select *All flags (true values)* to show the distribution of true values for flag fields in the dataset.

Field. Select a nominal or flag field for which to show the distribution of values. Only fields that have not been explicitly set as numeric appear on the list.

Overlay. Select a nominal or flag field to use as a color overlay, illustrating the distribution of its values within each value of the specified field. For example, you can use marketing campaign response (*pep*) as an overlay for number of children (*children*) to illustrate responsiveness by family size. For more information, see the topic [Aesthetics, Overlays, Panels, and Animation](#) on p. 210.

Normalize by color. Select to scale bars so that all bars take up the full width of the graph. The overlay values equal a proportion of each bar, making comparisons across categories easier.

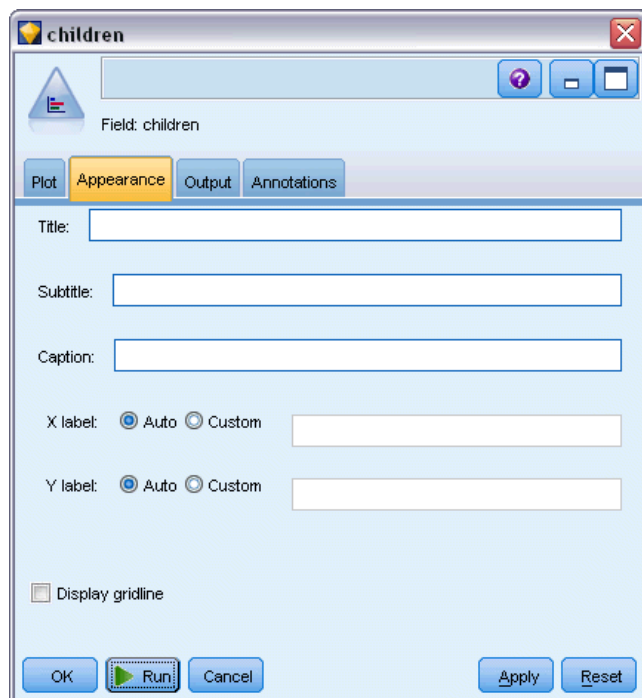
Sort. Select the method used to display values in the distribution graph. Select *Alphabetic* to use alphabetical order or *By count* to list values in decreasing order of occurrence.

Proportional scale. Select to scale the distribution of values so that the value with the largest count fills the full width of the plot. All other bars are scaled against this value. Deselecting this option scales bars according to the total counts of each value.

Distribution Appearance Tab

You can specify appearance options before graph creation.

Figure 5-26
Appearance tab settings



Title. Enter the text to be used for the graph's title.

Subtitle. Enter the text to be used for the graph's subtitle.

Caption. Enter the text to be used for the graph's caption.

X label. Either accept the automatically generated x -axis (horizontal) label or select Custom to specify a label.

Y label. Either accept the automatically generated y -axis (vertical) label or select Custom to specify a label.

Display gridline. Selected by default, this option displays a gridline behind the plot or graph that enables you to more easily determine region and band cutoff points. Gridlines are always displayed in white unless the graph background is white; in this case, they are displayed in gray.

Using a Distribution Node

Distribution nodes are used to show the distribution of symbolic values in a dataset. They are frequently used before manipulation nodes to explore the data and correct any imbalances. For example, if instances of respondents without children occur much more frequently than other types of respondents, you might want to reduce these instances so that a more useful rule can be generated in later data mining operations. A Distribution node will help you to examine and make decisions about such imbalances.

The Distribution node is unusual in that it produces both a graph and a table to analyze your data.

Figure 5-27

Distribution graph showing the number of people with or without children who responded to a marketing campaign

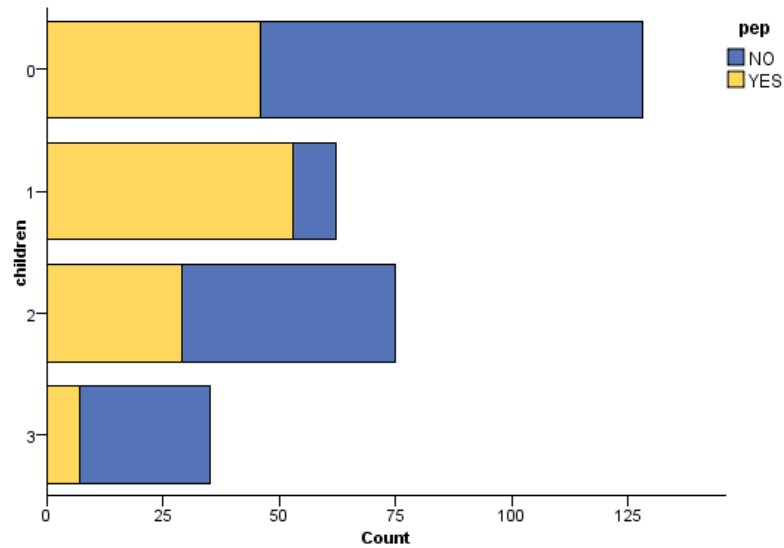
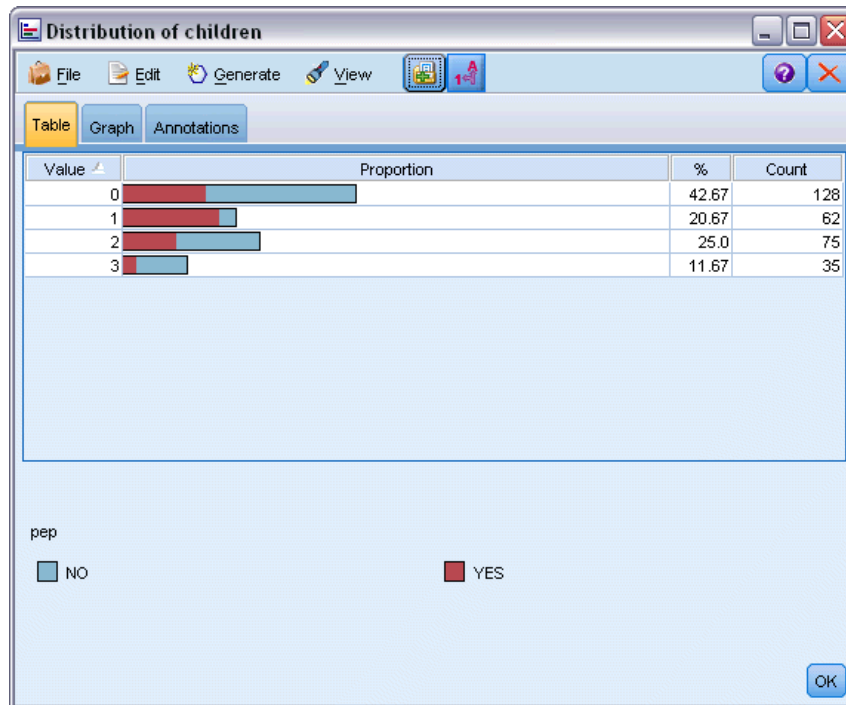


Figure 5-28

Distribution table showing the proportion of people with or without children who responded to a marketing campaign



Once you have created a distribution table and graph and examined the results, you can use options from the menus to group values, copy values, and generate a number of nodes for data preparation. In addition, you can copy or export the graph and table information for use in other applications, such as MS Word or MS PowerPoint. For more information, see the topic [Printing, Saving, Copying, and Exporting Graphs](#) on p. 311.

To Select and Copy Values from a Distribution Table

- ▶ Click and hold the mouse button while dragging it over the rows to select a set of values. You can also use the Edit menu to Select All values.
- ▶ From the Edit menu, choose Copy Table or Copy Table (inc. field names).
- ▶ Paste to the clipboard or into the desired application.

Note: The bars do not get copied directly. Instead, the table values are copied. This means that overlaid values will not be displayed in the copied table.

To Group Values from a Distribution Table

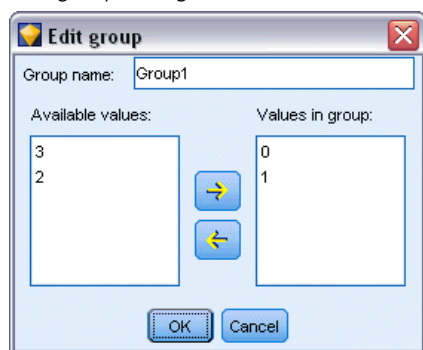
- ▶ Select values for grouping using the Ctrl+click method.
- ▶ From the Edit menu, choose Group.

Note: When you group and ungroup values, the graph on the Graph tab is automatically redrawn to show the changes.

You can also:

- Ungroup values by selecting the group name in the distribution list and choosing Ungroup from the Edit menu.
- Edit groups by selecting the group name in the distribution list and choosing Edit group from the Edit menu. This opens a dialog box where values can be shifted to and from the group.

Figure 5-29
Edit group dialog box



Generate Menu Options

You can use options on the Generate menu to select a subset of data, derive a flag field, regroup values, reclassify values, or balance the data from either a graph or table. These operations generate a data preparation node and place it on the stream canvas. To use the generated node,

connect it to an existing stream. For more information, see the topic [Generating Nodes from Graphs](#) on p. 288.

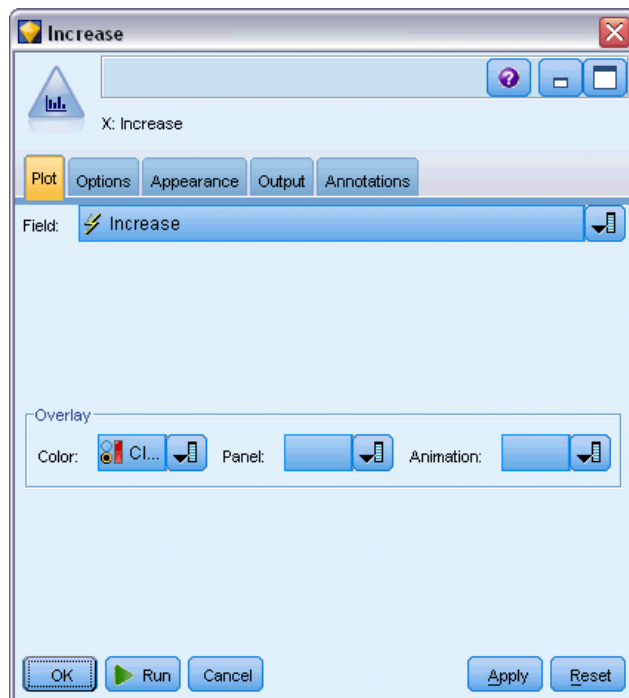
Histogram Node

Histogram nodes show the occurrence of values for numeric fields. They are often used to explore the data before manipulations and model building. Similar to the Distribution node, Histogram nodes are frequently used to reveal imbalances in the data. While you can also use the Graphboard node to produce a histogram, you have more options to choose from in this node.

Note: To show the occurrence of values for symbolic fields, you should use a Distribution node.

Histogram Plot Tab

Figure 5-30
Plot tab settings for a Histogram node

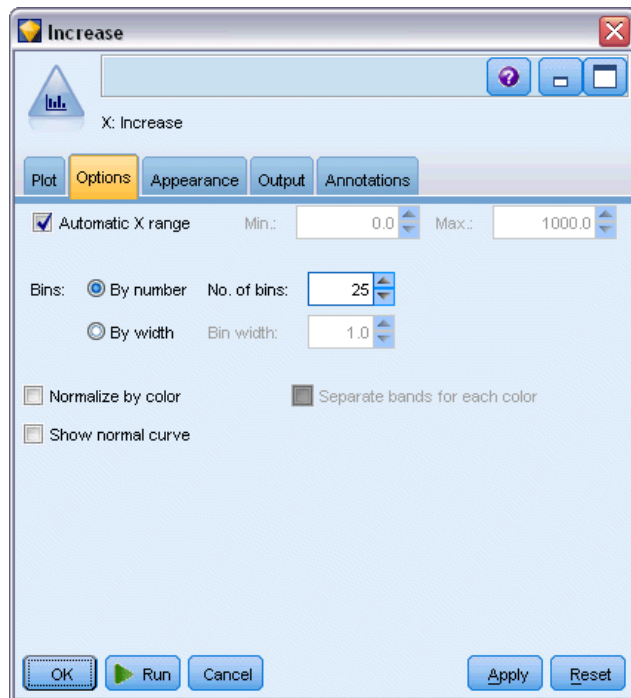


Field. Select a numeric field for which to show the distribution of values. Only fields that have not been explicitly defined as symbolic (categorical) will be listed.

Overlay. Select a symbolic field to show categories of values for the specified field. Selecting an overlay field converts the histogram to a stacked chart with colors used to represent different categories of the overlay field. Using the Histogram node, there are three types of overlays: color, panel, and animation. For more information, see the topic [Aesthetics, Overlays, Panels, and Animation](#) on p. 210.

Histogram Options Tab

Figure 5-31
Options tab settings for a Histogram node



Automatic X range. Select to use the entire range of values in the data along this axis. Deselect to use an explicit subset of values based on your specified Min and Max values. Either enter values or use the arrows. Automatic ranges are selected by default to enable rapid graph building.

Bins. Select either By number or By width.

- Select By number to display a fixed number of bars whose width depends on the range and the number of bins specified. Indicate the number of bins to be used in the graph in the No. of bins option. Use the arrows to adjust the number.
- Select By width to create a graph with bars of a fixed width. The number of bins depends on the specified width and the range of values. Indicate the width of the bars in the Bin width option.

Normalize by color. Select to adjust all bars to the same height, displaying overlaid values as a percentage of the total cases in each bar.

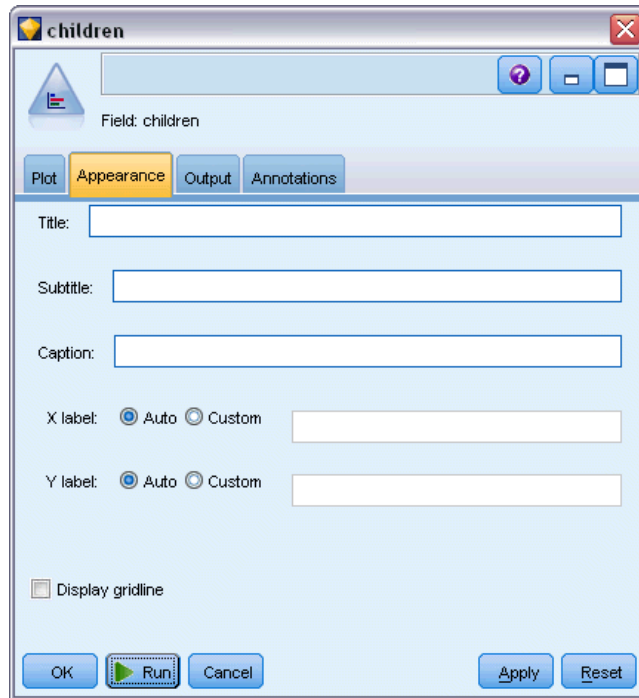
Show normal curve. Select to add a normal curve to the graph showing the mean and variance of the data.

Separate bands for each color. Select to display each overlaid value as a separate band on the graph.

Histogram Appearance Tab

You can specify appearance options before graph creation.

Figure 5-32
Appearance tab settings for most graph nodes



Title. Enter the text to be used for the graph's title.

Subtitle. Enter the text to be used for the graph's subtitle.

Caption. Enter the text to be used for the graph's caption.

X label. Either accept the automatically generated x -axis (horizontal) label or select Custom to specify a label.

Y label. Either accept the automatically generated y -axis (vertical) label or select Custom to specify a label.

Display gridline. Selected by default, this option displays a gridline behind the plot or graph that enables you to more easily determine region and band cutoff points. Gridlines are always displayed in white unless the graph background is white; in this case, they are displayed in gray.

Using Histograms

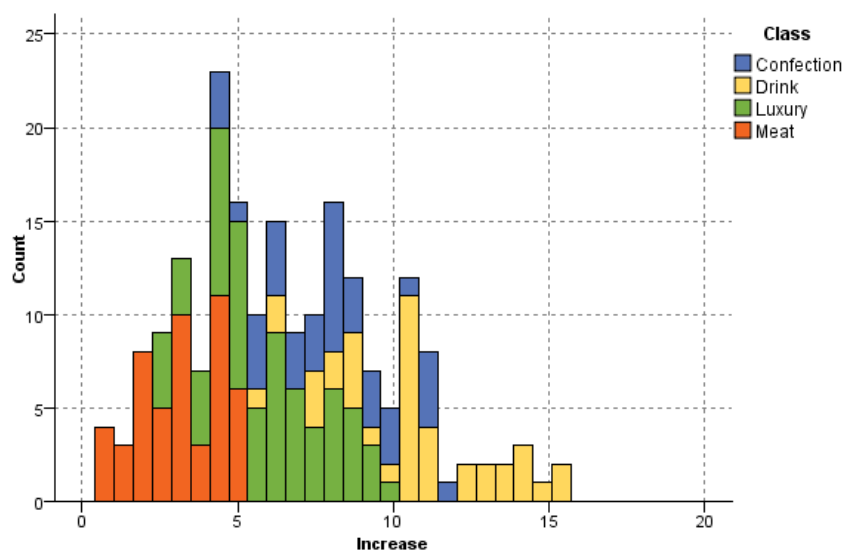
Histograms show the distribution of values in a numeric field whose values range along the x axis. Histograms operate similarly to collections graphs. Collections show the distribution of values for one numeric field *relative to the values of another*, rather than the occurrence of values for a single field.

Once you have created a graph, you can examine the results and define bands to split values along the x axis or define regions. You can also mark elements within the graph. For more information, see the topic [Exploring Graphs](#) on p. 281.

You can use options on the Generate menu to create Balance, Select, or Derive nodes using the data in the graph or more specifically within bands, regions, or marked elements. This type of graph is frequently used before manipulation nodes to explore the data and correct any imbalances by generating a Balance node from the graph to use in the stream. You can also generate a Derive Flag node to add a field showing which band each record falls into or a Select node to select all records within a particular set or range of values. Such operations help you to focus on a particular subset of data for further exploration. For more information, see the topic [Generating Nodes from Graphs](#) on p. 288.

Figure 5-33

Histogram showing the distribution of increased purchases by category due to promotion

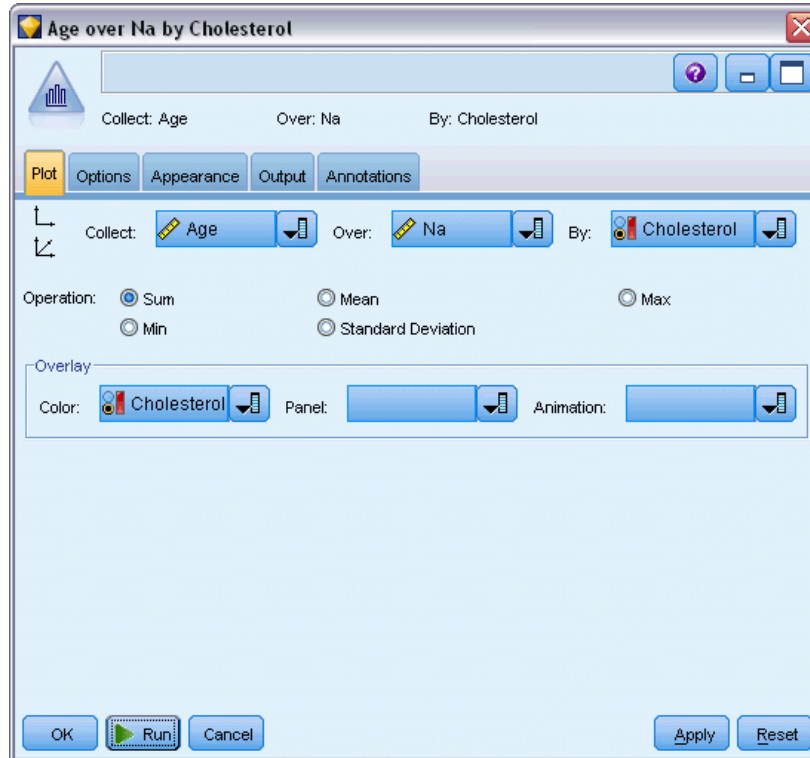


Collection Node

Collections are similar to histograms except that collections show the distribution of values for one numeric field relative to the values of another, rather than the occurrence of values for a single field. A collection is useful for illustrating a variable or field whose values change over time. Using 3-D graphing, you can also include a symbolic axis displaying distributions by category. Two dimensional Collections are shown as stacked bar charts, with overlays where used. For more information, see the topic [Aesthetics, Overlays, Panels, and Animation](#) on p. 210.

Collection Plot Tab

Figure 5-34
Plot tab settings for a Collection node



Collect. Select a field whose values will be collected and displayed over the range of values for the field specified in Over. Only fields that have not been defined as symbolic are listed.

Over. Select a field whose values will be used to display the field specified in Collect.

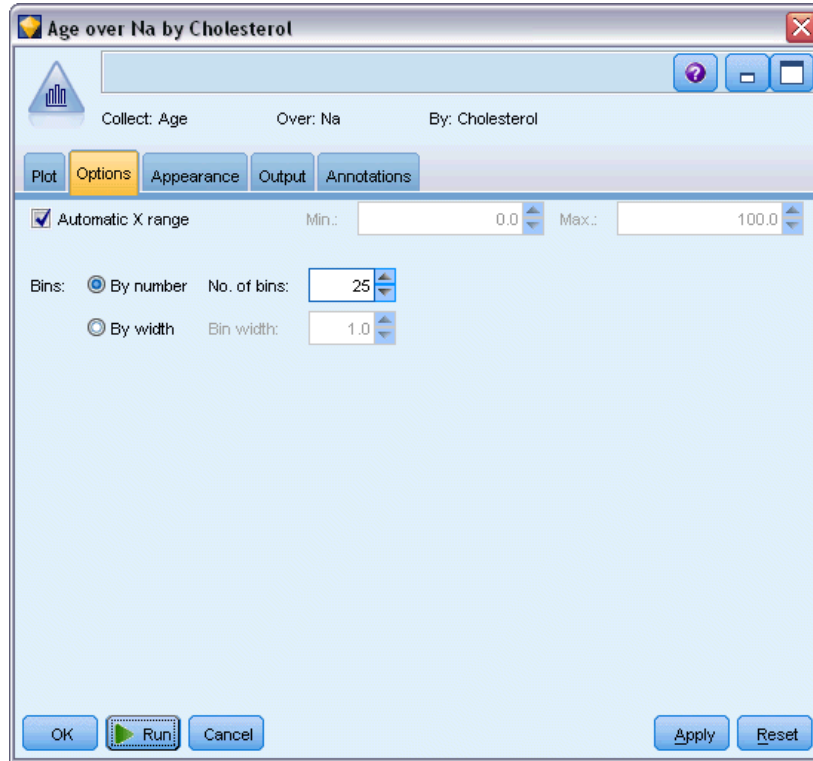
By. Enabled when creating a 3-D graph, this option allows you to select a nominal or flag field used to display the collection field by categories.

Operation. Select what each bar in the collection graph represents. Options include Sum, Mean, Max, Min, and Standard Deviation.

Overlay. Select a symbolic field to show categories of values for the selected field. Selecting an overlay field converts the collection and creates multiple bars of varying colors for each category. This node has three types of overlays: color, panel, and animation. For more information, see the topic [Aesthetics, Overlays, Panels, and Animation](#) on p. 210.

Collection Options Tab

Figure 5-35
Options tab settings for a Collection node



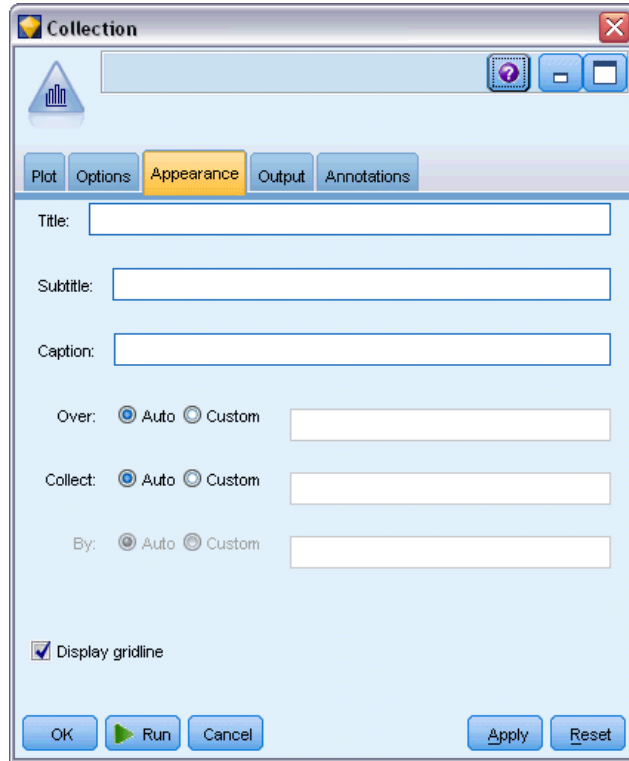
Automatic X range. Select to use the entire range of values in the data along this axis. Deselect to use an explicit subset of values based on your specified Min and Max values. Either enter values or use the arrows. Automatic ranges are selected by default to enable rapid graph building.

Bins. Select either By number or By width.

- Select By number to display a fixed number of bars whose width depends on the range and the number of bins specified. Indicate the number of bins to be used in the graph in the No. of bins option. Use the arrows to adjust the number.
- Select By width to create a graph with bars of a fixed width. The number of bins depends on the specified width and the range of values. Indicate the width of the bars in the Bin width option.

Collection Appearance Tab

Figure 5-36
Appearance tab settings for a Collection node



You can specify appearance options before graph creation.

Title. Enter the text to be used for the graph's title.

Subtitle. Enter the text to be used for the graph's subtitle.

Caption. Enter the text to be used for the graph's caption.

Over label. Either accept the automatically generated label, or select Custom to specify a label.

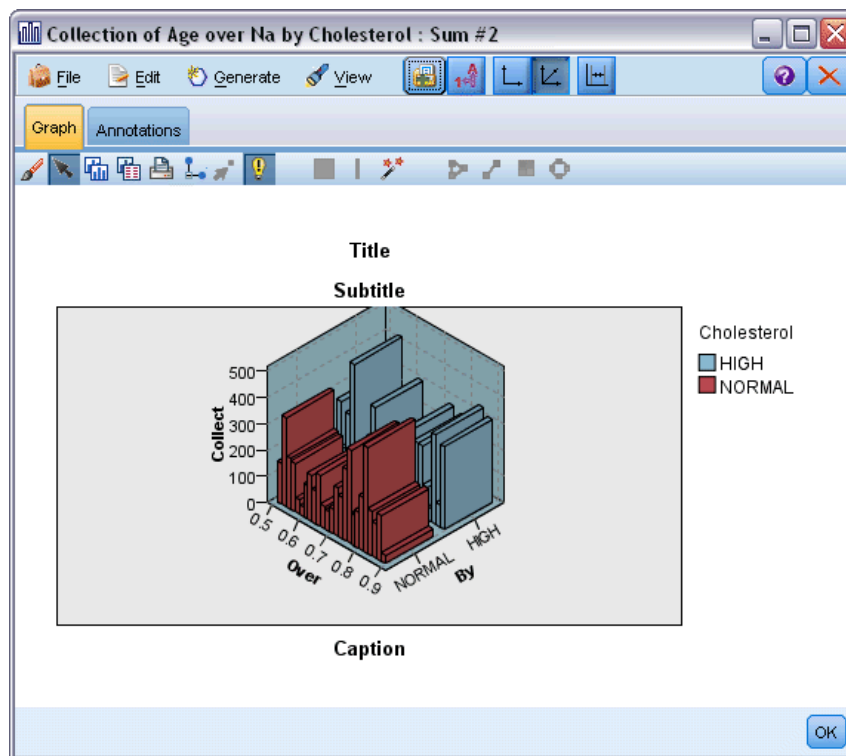
Collect label. Either accept the automatically generated label, or select Custom to specify a label.

By label. Either accept the automatically generated label, or select Custom to specify a label.

Display gridline. Selected by default, this option displays a gridline behind the plot or graph that enables you to more easily determine region and band cutoff points. Gridlines are always displayed in white unless the graph background is white; in this case, they are displayed in gray.

The following example shows where the appearance options are placed on a 3-D version of the graph.

Figure 5-37
Position of graph appearance options on 3-D Collection graph



Using a Collection Graph

Collections show the distribution of values for one numeric field *relative to the values of another*, rather than the occurrence of values for a single field. Histograms operate similarly to collections graphs. Histograms show the distribution of values in a numeric field whose values range along the x axis.

Once you have created a graph, you can examine the results and define bands to split values along the x axis or define regions. You can also mark elements within the graph. For more information, see the topic [Exploring Graphs](#) on p. 281.

You can use options on the Generate menu to create Balance, Select, or Derive nodes using the data in the graph or more specifically within bands, regions, or marked elements. This type of graph is frequently used before manipulation nodes to explore the data and correct any imbalances by generating a Balance node from the graph to use in the stream. You can also generate a Derive Flag node to add a field showing which band each record falls into or a Select node to select all records within a particular set or range of values. Such operations help you to focus on a particular subset of data for further exploration. For more information, see the topic [Generating Nodes from Graphs](#) on p. 288.

Figure 5-38
3-D collection graph showing sum of Na_to_K over Age for both high and normal cholesterol levels

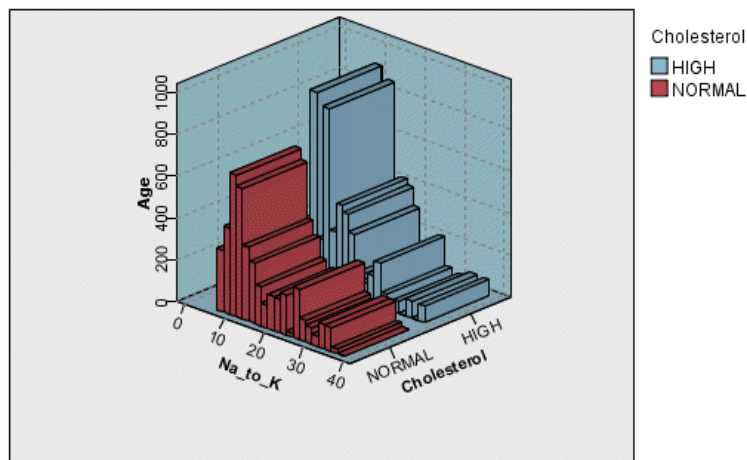
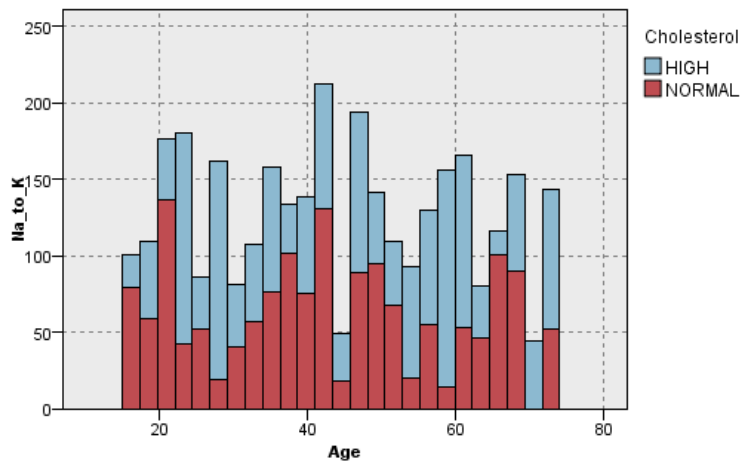


Figure 5-39
Collection graph without z axis displayed but with Cholesterol as color overlay

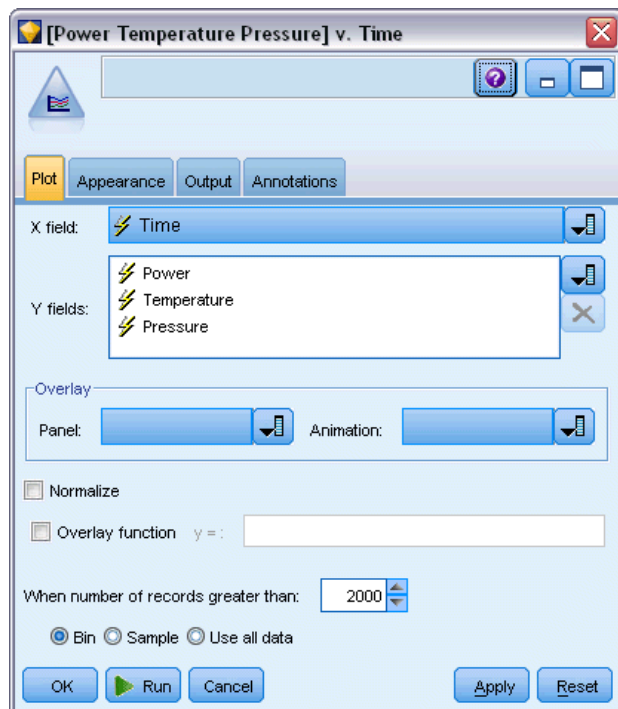


Multiplot Node

A multiplot is a special type of plot that displays multiple Y fields over a single X field. The Y fields are plotted as colored lines and each is equivalent to a Plot node with Style set to Line and X Mode set to Sort. Multiplots are useful when you have time sequence data and want to explore the fluctuation of several variables over time.

Multiplot Plot Tab

Figure 5-40
Plot tab settings for a Multiplot node



X field. From the list, select the field to display on the horizontal x axis.

Y fields. Select one or more fields from the list to display over the range of X field values. Use the Field Chooser button to select multiple fields. Click the delete button to remove fields from the list.

Overlay. There are several ways to illustrate categories for data values. For example, you might use an animation overlay to display multiple plots for each value in the data. This is useful for sets containing upwards of 10 categories. When used for sets with more than 15 categories, you may notice a decrease in performance. For more information, see the topic [Aesthetics, Overlays, Panels, and Animation](#) on p. 210.

Normalize. Select to scale all Y values to the range 0–1 for display on the graph. Normalizing helps you explore the relationship between lines that might otherwise be obscured due to differences in the range of values for each series and is recommended when plotting multiple lines on the same graph, or when comparing plots in side-by-side panels. (Normalizing is not necessary when all data values fall within a similar range.)

Figure 5-41

Standard multiplot showing power-plant fluctuation over time (note that without normalizing, the plot for Pressure is impossible to see)

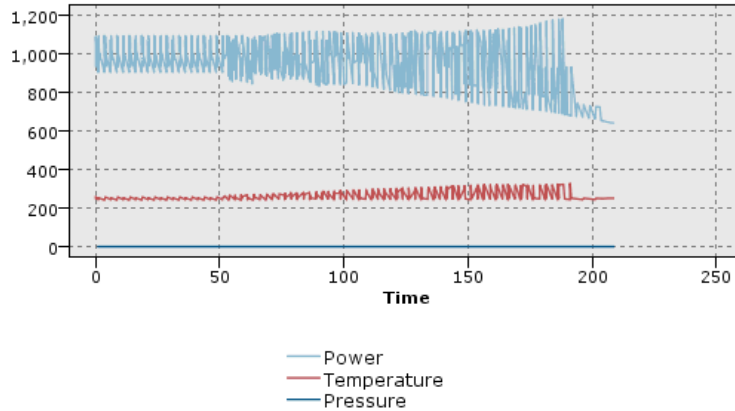
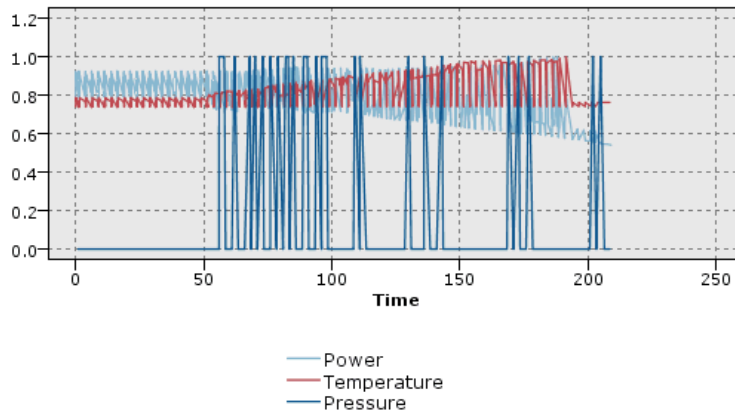


Figure 5-42

Normalized multiplot showing a plot for Pressure



Overlay function. Select to specify a known function to compare to actual values. For example, to compare actual versus predicted values, you can plot the function $y = x$ as an overlay. Specify a function for $y =$ in the text box. The default function is $y = x$, but you can specify any sort of function, such as a quadratic function or an arbitrary expression, in terms of x .

Note: Overlay functions are not available for a panel or animation graph.

When number of records greater than. Specify a method for plotting large datasets. You can specify a maximum dataset size or use the default 2,000 points. Performance is enhanced for large datasets when you select the Bin or Sample options. Alternatively, you can choose to plot all data points by selecting Use all data, but you should note that this may dramatically decrease the performance of the software.

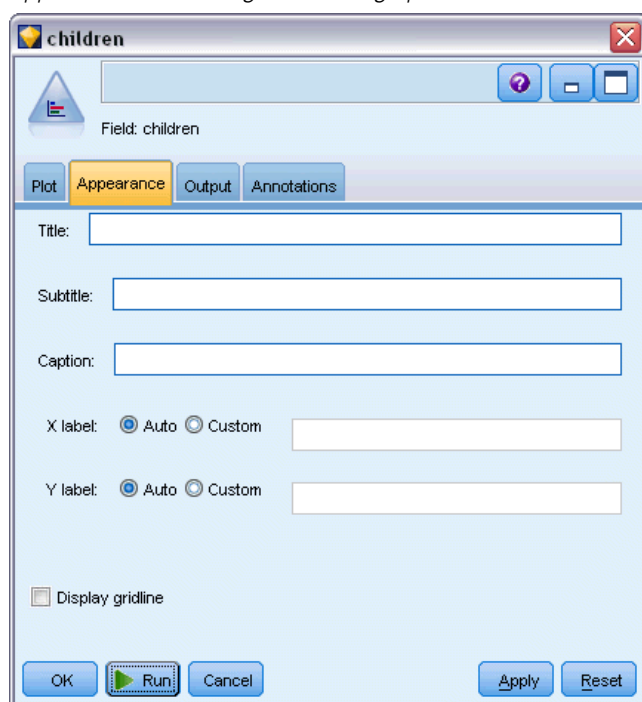
Note: When X Mode is set to Overlay or As Read, these options are disabled and only the first n records are used.

- **Bin.** Select to enable binning when the dataset contains more than the specified number of records. Binning divides the graph into fine grids before actually plotting and counts the number of connections that would appear in each of the grid cells. In the final graph, one connection is used per cell at the bin centroid (average of all connection points in the bin).
- **Sample.** Select to randomly sample the data to the specified number of records.

Multiplot Appearance Tab

You can specify appearance options before graph creation.

Figure 5-43
Appearance tab settings for most graph nodes



Title. Enter the text to be used for the graph's title.

Subtitle. Enter the text to be used for the graph's subtitle.

Caption. Enter the text to be used for the graph's caption.

X label. Either accept the automatically generated *x*-axis (horizontal) label or select Custom to specify a label.

Y label. Either accept the automatically generated *y*-axis (vertical) label or select Custom to specify a label.

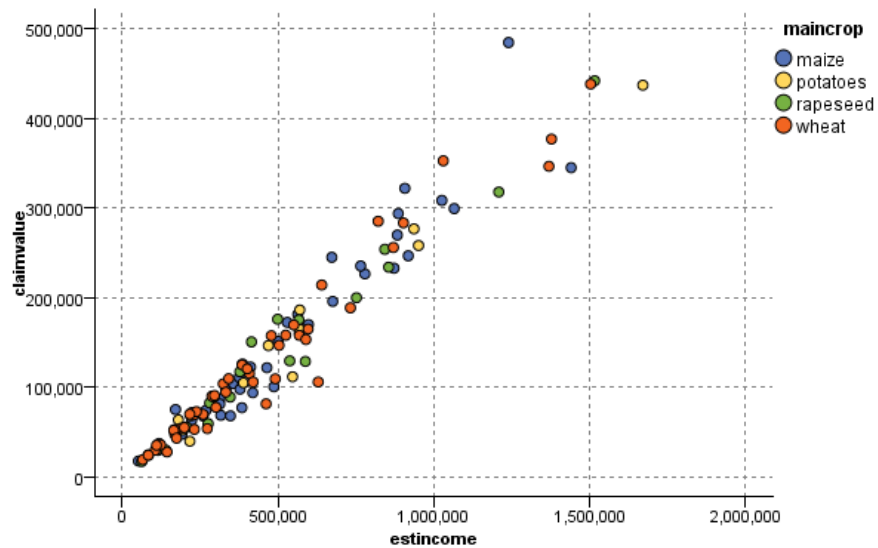
Display gridline. Selected by default, this option displays a gridline behind the plot or graph that enables you to more easily determine region and band cutoff points. Gridlines are always displayed in white unless the graph background is white; in this case, they are displayed in gray.

Using a Multiplot Graph

Plots and multiplots are essentially plots of X against Y . For example, if you are exploring potential fraud in agricultural grant applications (as illustrated in *fraud.str* in the *Demos* folder of your IBM® SPSS® Modeler installation), you might want to plot the income claimed on the application versus the income estimated by a neural net. Using an overlay, such as crop type, will illustrate whether there is a relationship between claims (value or number) and type of crop.

Figure 5-44

Plot of the relationship between estimated income and claim value with main crop type as an overlay

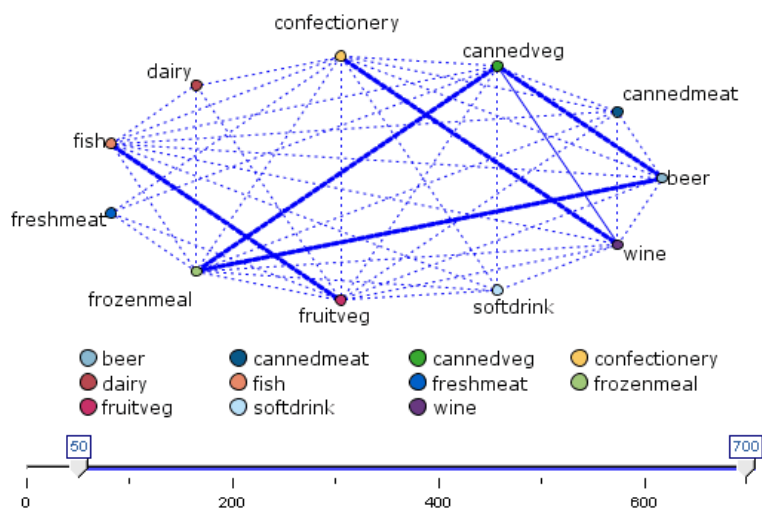


Since plots, multiplots, and evaluation charts are two-dimensional displays of Y against X , it is easy to interact with them by defining regions, marking elements, or even drawing bands. You can also generate nodes for the data represented by those regions, bands, or elements. For more information, see the topic [Exploring Graphs](#) on p. 281.

Web Node

Web nodes show the strength of relationships between values of two or more symbolic fields. The graph displays connections using varying types of lines to indicate connection strength. You can use a Web node, for example, to explore the relationship between the purchase of various items at an e-commerce site or a traditional retail outlet.

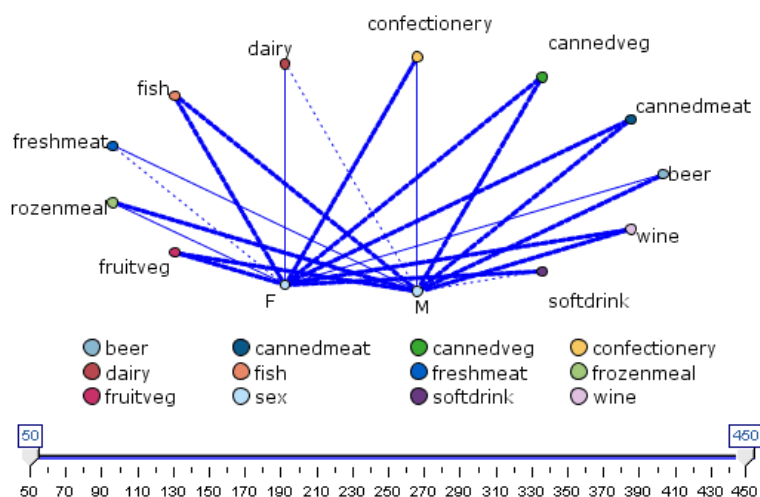
Figure 5-45
Web graph showing relationships between the purchase of grocery items



Directed Webs

Directed Web nodes are similar to Web nodes in that they show the strength of relationships between symbolic fields. However, directed web graphs show connections only from one or more From fields to a single To field. The connections are unidirectional in the sense that they are one-way connections.

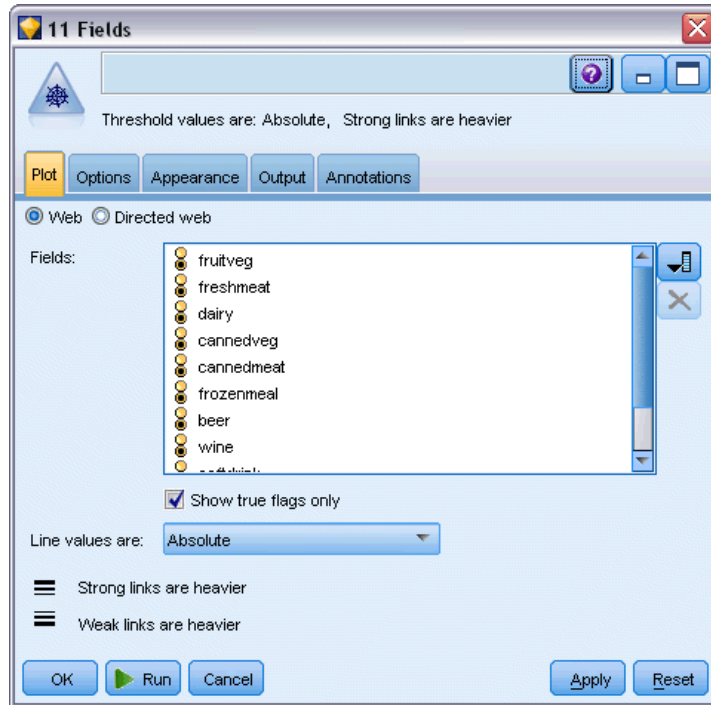
Figure 5-46
Directed web graph showing the relationship between the purchase of grocery items and gender



Like Web nodes, the graph displays connections using varying types of lines to indicate connection strength. You can use a Directed Web node, for example, to explore the relationship between gender and a proclivity for certain purchase items.

Web Plot Tab

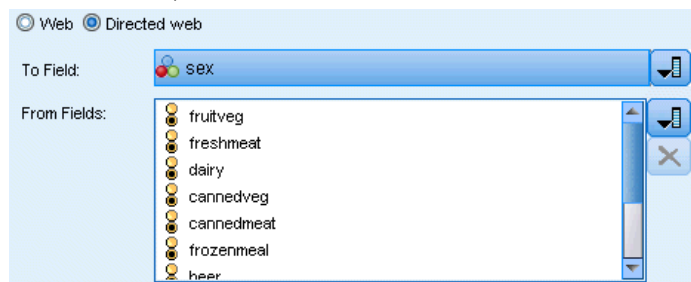
Figure 5-47
Plot tab settings for a Web node



Web. Select to create a web graph illustrating the strength of relationships between all specified fields.

Directed web. Select to create a directional web graph illustrating the strength of relationships between multiple fields and the values of one field, such as gender or religion. When this option is selected, a To Field is activated and the Fields control below is renamed From Fields for additional clarity.

Figure 5-48
Directed web options



To Field (directed web only). Select a flag or nominal field used for a directed web. Only fields that have not been explicitly set as numeric are listed.

Fields/From Fields. Select fields to create a web graph. Only fields that have not been explicitly set as numeric are listed. Use the Field Chooser button to select multiple fields or select fields by type.

Note: For a directed web, this control is used to select From fields.

Show true flags only. Select to display only true flags for a flag field. This option simplifies the web display and is often used for data where the occurrence of positive values is of special importance.

Line values are. Select a threshold type from the drop-down list.

- Absolute sets thresholds based on the number of records having each pair of values.
- Overall percentages shows the absolute number of cases represented by the link as a proportion of all of the occurrences of each pair of values represented in the web graph.
- Percentages of smaller field/value and Percentages of larger field/value indicate which field/value to use for evaluating percentages. For example, suppose 100 records have the value *drugY* for the field *Drug* and only 10 have the value *LOW* for the field *BP*. If seven records have both values *drugY* and *LOW*, this percentage is either 70% or 7%, depending on which field you are referencing, smaller (*BP*) or larger (*Drug*).

Note: For directed web graphs, the third and fourth options above are not available. Instead, you can select Percentage of “To” field/value and Percentage of “From” field/value.

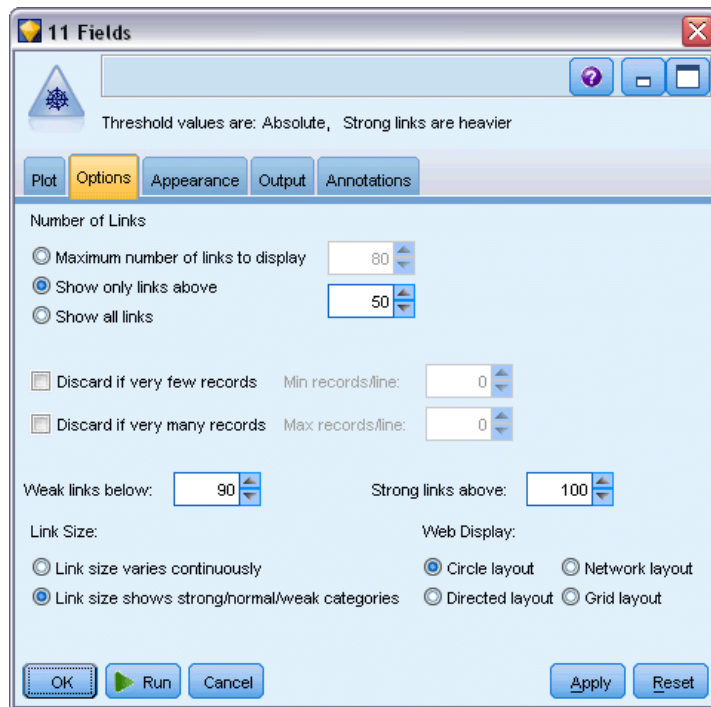
Strong links are heavier. Selected by default, this is the standard way of viewing links between fields.

Weak links are heavier. Select to reverse the meaning of links displayed in bold lines. This option is frequently used for fraud detection or examination of outliers.

Web Options Tab

The Options tab for Web nodes contains a number of additional options to customize the output graph.

Figure 5-49
Options tab settings for a Web node



Number of Links. The following options are used to control the number of links displayed in the output graph. Some of these options, such as *Weak links above* and *Strong links above*, are also available in the output graph window. You can also use a slider control in the final graph to adjust the number of links displayed.

- **Maximum number of links to display.** Specify a number indicating the maximum number of links to show on the output graph. Use the arrows to adjust the value.
- **Show only links above.** Specify a number indicating the minimum value for which to show a connection in the web. Use the arrows to adjust the value.
- **Show all links.** Specify to display all links regardless of minimum or maximum values. Selecting this option may increase processing time if there are a large number of fields.

Discard if very few records. Select to ignore connections that are supported by too few records. Set the threshold for this option by entering a number in *Min. records/line*.

Discard if very many records. Select to ignore strongly supported connections. Enter a number in *Max. records/line*.

Weak links below. Specify a number indicating the threshold for weak connections (dotted lines) and regular connections (normal lines). All connections below this value are considered weak.

Strong links above. Specify a threshold for strong connections (heavy lines) and regular connections (normal lines). All connections above this value are considered strong.

Link Size. Specify options for controlling the size of links:

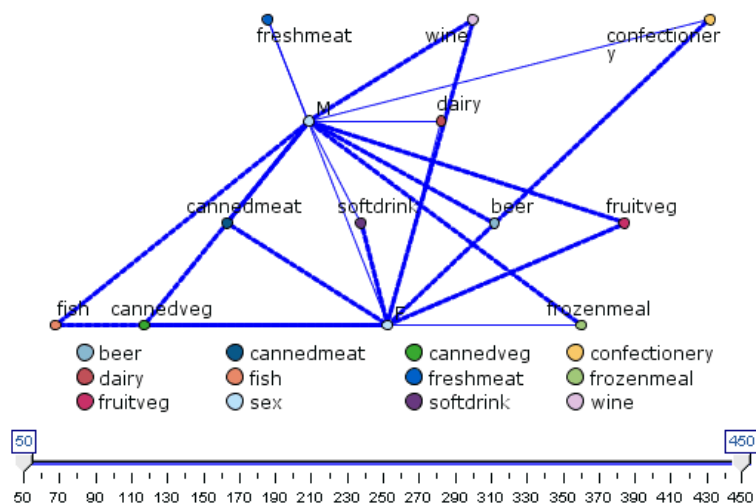
- **Link size varies continuously.** Select to display a range of link sizes reflecting the variation in connection strengths based on actual data values.
- **Link size shows strong/normal/weak categories.** Select to display three strengths of connections—strong, normal, and weak. The cutoff points for these categories can be specified above as well as in the final graph.

Web Display. Select a type of web display:

- **Circle layout.** Select to use the standard web display.
- **Network layout.** Select to use an algorithm to group together the strongest links. This is intended to highlight strong links using spatial differentiation as well as weighted lines.
- **Directed Layout.** Select to create a directed web display that uses the To Field selection from the Plot tab as the focus for the direction.
- **Grid Layout.** Select to create a web display that is laid out in a regularly spaced grid pattern.

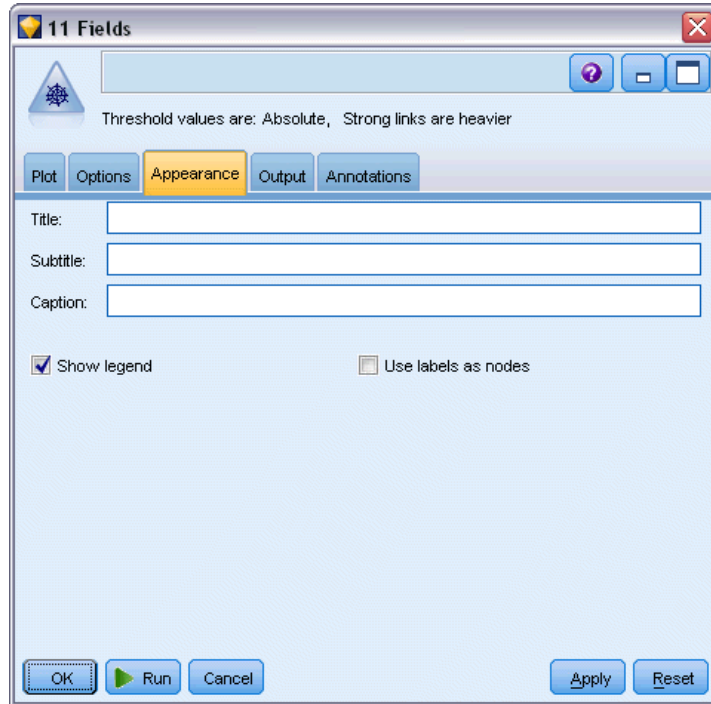
Figure 5-50

Web graph showing strong connections from frozenmeal and cannedveg to other grocery items



Web Appearance Tab

Figure 5-51
Appearance tab settings for a Web node



You can specify appearance options before graph creation.

Title. Enter the text to be used for the graph's title.

Subtitle. Enter the text to be used for the graph's subtitle.

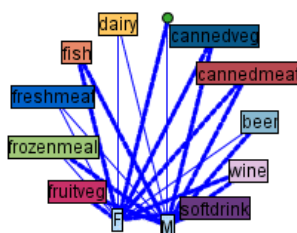
Caption. Enter the text to be used for the graph's caption.

Show legend. You can specify whether the legend is displayed. For plots with a large number of fields, hiding the legend may improve the appearance of the plot.

Use labels as nodes. You can include the label text within each node rather than displaying adjacent labels. For plots with a small number of fields, this may result in a more readable chart.

Figure 5-52
Web graph showing labels as nodes

Relationship between gender and grocery purchases



Using a Web Graph

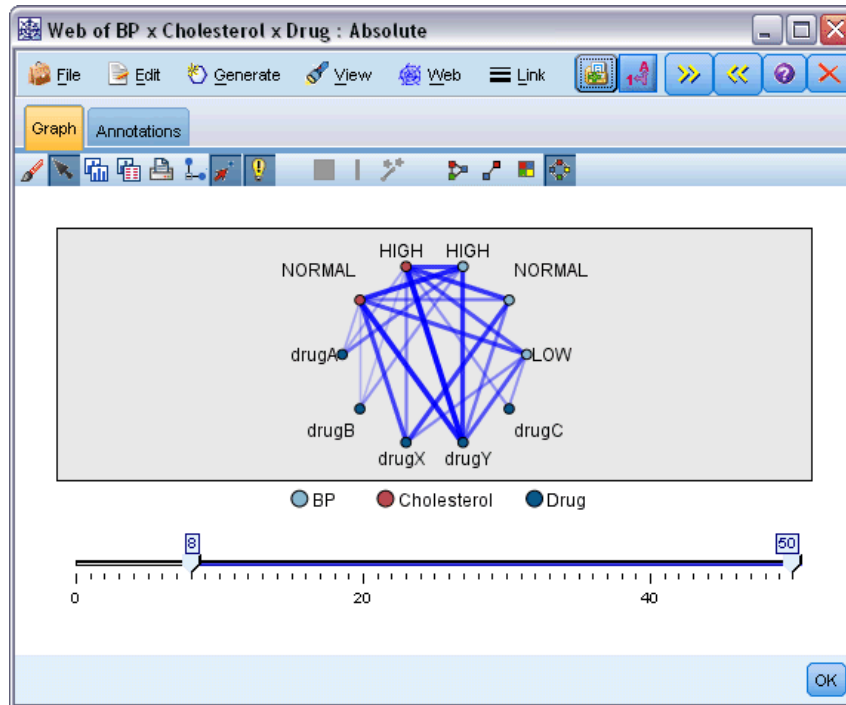
Web nodes are used to show the strength of relationships between values of two or more symbolic fields. Connections are displayed in a graph with varying types of lines to indicate connections of increasing strength. You can use a Web node, for example, to explore the relationship between cholesterol levels, blood pressure, and the drug that was effective in treating the patient's illness.

- Strong connections are shown with a heavy line. This indicates that the two values are strongly related and should be further explored.
- Medium connections are shown with a line of normal weight.
- Weak connections are shown with a dotted line.
- If no line is shown between two values, this means either that the two values never occur in the same record or that this combination occurs in a number of records below the threshold specified in the Web node dialog box.

Once you have created a Web node, there are several options for adjusting the graph display and generating nodes for further analysis.

Figure 5-53

Web graph indicating a number of strong relationships, such as normal blood pressure with DrugX and high cholesterol with DrugY



For both Web nodes and Directed Web nodes, you can:

- Change the layout of the web display.
- Hide points to simplify the display.
- Change the thresholds controlling line styles.
- Highlight lines between values to indicate a “selected” relationship.
- Generate a Select node for one or more “selected” records or a Derive Flag node associated with one or more relationships in the web.

To Adjust Points

- **Move** points by clicking the mouse on a point and dragging it to the new location. The web will be redrawn to reflect the new location.
- **Hide** points by right-clicking on a point in the web and choosing Hide or Hide and Replan from the context menu. Hide simply hides the selected point and any lines associated with it. Hide and Replan redraws the web, adjusting for any changes you have made. Any manual moves are undone.
- **Show** all hidden points by choosing Reveal All or Reveal All and Replan from the Web menu in the graph window. Selecting Reveal All and Replan redraws the web, adjusting to include all previously hidden points and their connections.

To Select, or "Highlight," Lines

Selected lines are highlighted in red.

- ▶ To select a single line, left-click the line.
- ▶ To select multiple lines, do one of the following:
 - Using the cursor, draw a circle around the points whose lines you want to select.
 - Hold down the Ctrl key and left-click the individual lines you want to select.

You can deselect all selected lines by clicking the graph background, or by choosing Clear Selection from the Web menu in the graph window.

To View the Web Using a Different Layout

- ▶ From the Web menu, choose Circle Layout, Network Layout, Directed Layout, or Grid Layout to change the layout of the graph.

To Turn the Links Slider on or off

- ▶ From the View menu, choose Links Slider.

To Select or Flag Records for a Single Relationship

- ▶ Right-click on the line representing the relationship of interest.
- ▶ From the context menu, choose Generate Select Node For Link or Generate Derive Node For Link.

A Select node or Derive node is automatically added to the stream canvas with the appropriate options and conditions specified:

- The Select node selects all records in the given relationship.
- The Derive node generates a flag indicating whether the selected relationship holds true for records in the entire dataset. The flag field is named by joining the two values in the relationship with an underscore, such as *LOW_drugC* or *drugC_LOW*.

To Select or Flag Records for a Group of Relationships

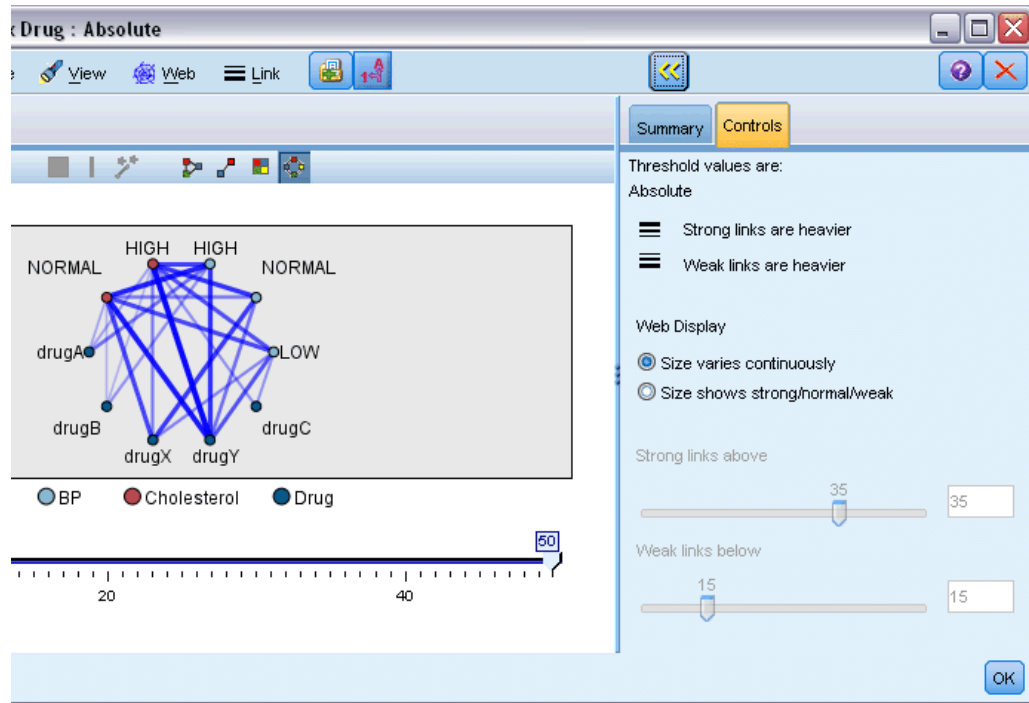
- ▶ Select the line(s) in the web display representing relationships of interest.
- ▶ From the Generate menu in the graph window, choose Select Node ("And"), Select Node ("Or"), Derive Node ("And"), or Derive Node ("Or").
 - The "Or" nodes give the disjunction of conditions. This means that the node will apply to records for which any of the selected relationships hold.
 - The "And" nodes give the conjunction of conditions. This means that the node will apply only to records for which all selected relationships hold. An error occurs if any of the selected relationships are mutually exclusive.

After you have completed your selection, a Select node or Derive node is automatically added to the stream canvas with the appropriate options and conditions specified.

Adjusting Web Thresholds

After you have created a web graph, you can adjust the thresholds controlling line styles using the toolbar slider to change the minimum visible line. You can also view additional threshold options by clicking the yellow double-arrow button on the toolbar to expand the web graph window. Then click the Controls tab to view additional options.

Figure 5-54
Expanded window featuring display and threshold options



Threshold values are. Shows the type of threshold selected during creation in the Web node dialog box.

Strong links are heavier. Selected by default, this is the standard way of viewing links between fields.

Weak links are heavier. Select to reverse the meaning of links displayed in bold lines. This option is frequently used for fraud detection or examination of outliers.

Web Display. Specify options for controlling the size of links in the output graph:

- **Size varies continuously.** Select to display a range of link sizes reflecting the variation in connection strengths based on actual data values.
- **Size shows strong/normal/weak.** Select to display three strengths of connections—strong, normal, and weak. The cutoff points for these categories can be specified above as well as in the final graph.

Strong links above. Specify a threshold for strong connections (heavy lines) and regular connections (normal lines). All connections above this value are considered strong. Use the slider to adjust the value or enter a number in the field.

Weak links below. Specify a number indicating the threshold for weak connections (dotted lines) and regular connections (normal lines). All connections below this value are considered weak. Use the slider to adjust the value or enter a number in the field.

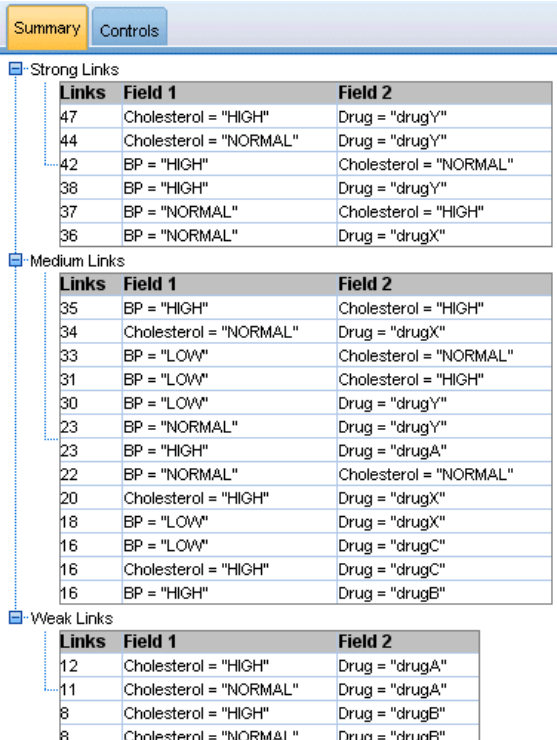
After you have adjusted the thresholds for a web, you can replan, or redraw, the web display with the new threshold values through the web menu located on the web graph toolbar. Once you have found settings that reveal the most meaningful patterns, you can update the original settings in the Web node (also called the Parent Web node) by choosing Update Parent Node from the Web menu in the graph window.

Creating a Web Summary

You can create a web summary document that lists strong, medium, and weak links by clicking the yellow double-arrow button on the toolbar to expand the web graph window. Then click the Summary tab to view tables for each type of link. Tables can be expanded and collapsed using the toggle buttons for each.

Figure 5-55

Web summary listing connections between blood pressure, cholesterol, and drug type



Summary		
Controls		
Strong Links		
Links	Field 1	Field 2
47	Cholesterol = "HIGH"	Drug = "drugY"
44	Cholesterol = "NORMAL"	Drug = "drugY"
42	BP = "HIGH"	Cholesterol = "NORMAL"
38	BP = "HIGH"	Drug = "drugY"
37	BP = "NORMAL"	Cholesterol = "HIGH"
36	BP = "NORMAL"	Drug = "drugX"
Medium Links		
Links	Field 1	Field 2
35	BP = "HIGH"	Cholesterol = "HIGH"
34	Cholesterol = "NORMAL"	Drug = "drugX"
33	BP = "LOW"	Cholesterol = "NORMAL"
31	BP = "LOW"	Cholesterol = "HIGH"
30	BP = "LOW"	Drug = "drugY"
23	BP = "NORMAL"	Drug = "drugY"
23	BP = "HIGH"	Drug = "drugA"
22	BP = "NORMAL"	Cholesterol = "NORMAL"
20	Cholesterol = "HIGH"	Drug = "drugX"
18	BP = "LOW"	Drug = "drugX"
16	BP = "LOW"	Drug = "drugC"
16	Cholesterol = "HIGH"	Drug = "drugC"
16	BP = "HIGH"	Drug = "drugB"
Weak Links		
Links	Field 1	Field 2
12	Cholesterol = "HIGH"	Drug = "drugA"
11	Cholesterol = "NORMAL"	Drug = "drugA"
8	Cholesterol = "HIGH"	Drug = "drugB"
8	Cholesterol = "NORMAL"	Drug = "drugB"

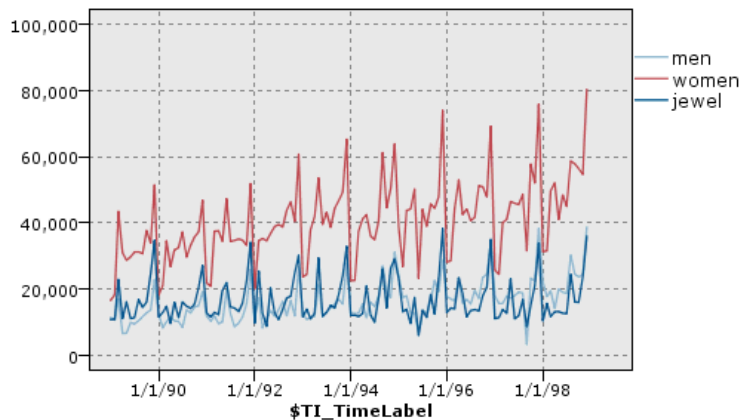
To print the summary, choose the following from the menu in the web graph window:

File > Print Summary

Time Plot Node

Time Plot nodes allow you to view one or more time series plotted over time. The series you plot must contain numeric values and are assumed to occur over a range of time in which the periods are uniform. You usually use a Time Intervals node before a Time Plot node to create a *TimeLabel* field, which is used by default to label the x axis in the graphs. For more information, see the topic [Time Intervals Node](#) in Chapter 4 on p. 186.

Figure 5-56
Plotting sales of men's and women's clothing and jewelry over time

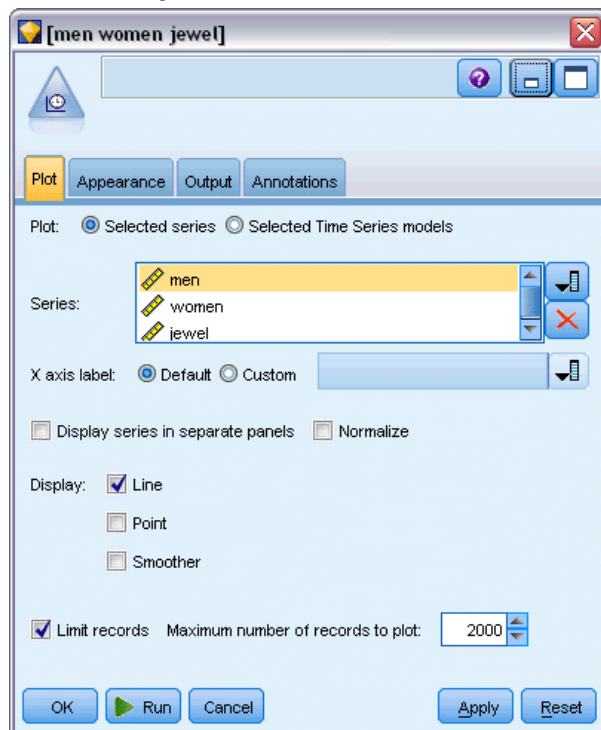


Creating Interventions and Events

You can create Event and Intervention fields from the time plot by generating a derive (flag or nominal) node from the context menus. For example, you could create an event field in the case of a rail strike, where the drive state is True if the event happened and False otherwise. For an Intervention field, for a price rise for example, you could use a derive count to identify the date of the rise, with 0 for the old price and 1 for the new price. For more information, see the topic [Derive Node](#) in Chapter 4 on p. 139.

Time Plot Tab

Figure 5-57
Plot tab settings for a Time Plot node



Plot. Provides a choice of how to plot time series data.

- **Selected series.** Plots values for selected time series. If you select this option when plotting confidence intervals, deselect the Normalize check box.
- **Selected Time Series models.** Used in conjunction with a Time Series model, this option plots all the related fields (actual and predicted values, as well as confidence intervals) for one or more selected time series. This option disables some other options in the dialog box. This is the preferred option if plotting confidence intervals.

Series. Select one or more fields with time series data you want to plot. The data must be numeric.

X axis label. Choose either the default label or a single field to use as the label for the x axis in plots. If you choose Default, the system uses the TimeLabel field created from a Time Intervals node upstream or sequential integers if there is no Time Intervals node. For more information, see the topic [Time Intervals Node](#) in Chapter 4 on p. 186.

Display series in separate panels. Specifies whether each series is displayed in a separate panel. Alternatively, if you do not choose to panel, all time series are plotted on the same graph, and smoothers will not be available. When plotting all time series on the same graph, each series will be represented by a different color.

Normalize. Select to scale all Y values to the range 0–1 for display on the graph. Normalizing helps you explore the relationship between lines that might otherwise be obscured due to differences in the range of values for each series and is recommended when plotting multiple lines on the

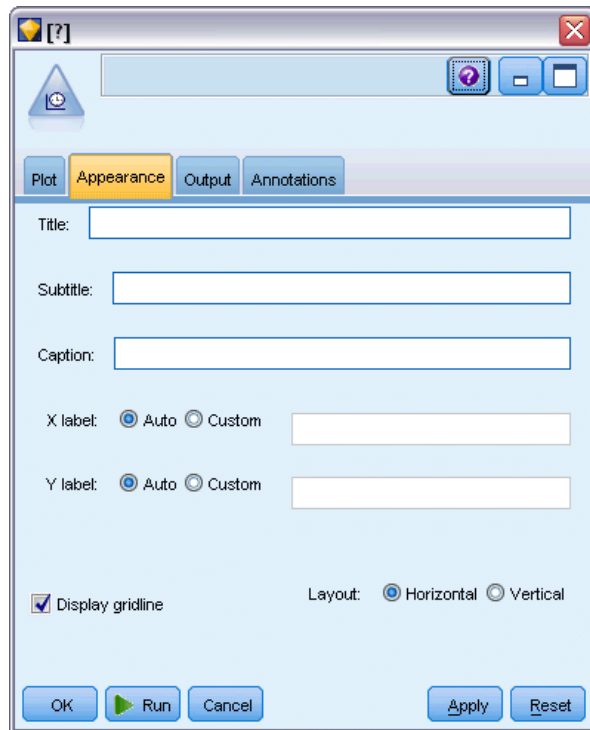
same graph, or when comparing plots in side-by-side panels. (Normalizing is not necessary when all data values fall within a similar range.)

Display. Select one or more elements to display in your plot. You can choose from lines, points, and (LOESS) smoothers. Smoothers are available only if you display the series in separate panels. By default, the line element is selected. Make sure you select at least one plot element before you execute the graph node; otherwise, the system will return an error stating that you have selected nothing to plot.

Limit records. Select this option if you want to limit the number of records plotted. Specify the number of records, read from the beginning of your data file, that will be plotted in the Maximum number of records to plot option. By default this number is set to 2,000. If you want to plot the last n records in your data file, you can use a Sort node prior to this node to arrange the records in descending order by time.

Time Plot Appearance Tab

Figure 5-58
Appearance tab settings for a Time Plot node



You can specify appearance options before graph creation.

Title. Enter the text to be used for the graph's title.

Subtitle. Enter the text to be used for the graph's subtitle.

Caption. Enter the text to be used for the graph's caption.

X label. Either accept the automatically generated x-axis (horizontal) label or select Custom to specify a label.

Y label. Either accept the automatically generated y-axis (vertical) label or select Custom to specify a label.

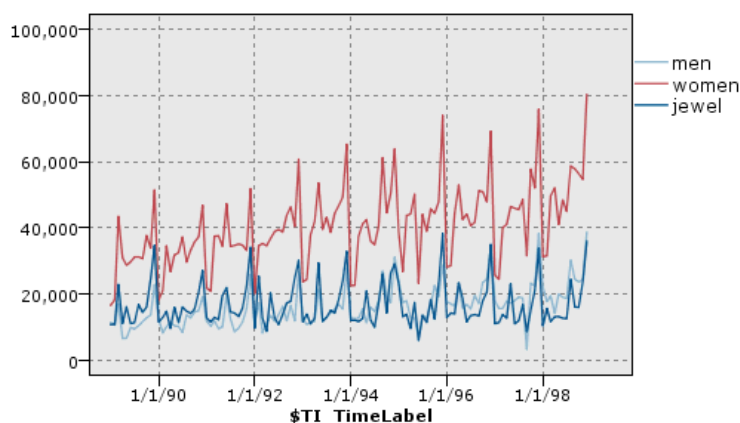
Display gridline. Selected by default, this option displays a gridline behind the plot or graph that enables you to more easily determine region and band cutoff points. Gridlines are always displayed in white unless the graph background is white; in this case, they are displayed in gray.

Layout. For time plots only, you can specify whether time values are plotted along a horizontal or vertical axis.

Using a Time Plot Graph

Once you have created a time plot graph, there are several options for adjusting the graph display and generating nodes for further analysis. For more information, see the topic [Exploring Graphs](#) on p. 281.

Figure 5-59
Plotting sales of men's and women's clothing and jewelry over time



Once you have created a time plot, defined bands, and examined the results, you can use options on the Generate menu and the context menu to create Select or Derive nodes. For more information, see the topic [Generating Nodes from Graphs](#) on p. 288.

Evaluation Node

The Evaluation node offers an easy way to evaluate and compare predictive models to choose the best model for your application. Evaluation charts show how models perform in predicting particular outcomes. They work by sorting records based on the predicted value and confidence of the prediction, splitting the records into groups of equal size (**quantiles**), and then plotting the value of the business criterion for each quantile, from highest to lowest. Multiple models are shown as separate lines in the plot.

Outcomes are handled by defining a specific value or range of values as a **hit**. Hits usually indicate success of some sort (such as a sale to a customer) or an event of interest (such as a specific medical diagnosis). You can define hit criteria on the Options tab of the dialog box, or you can use the default hit criteria as follows:

- **Flag** output fields are straightforward; hits correspond to *true* values.
- For **Nominal** output fields, the first value in the set defines a hit.
- For **Continuous** output fields, hits equal values greater than the midpoint of the field's range.

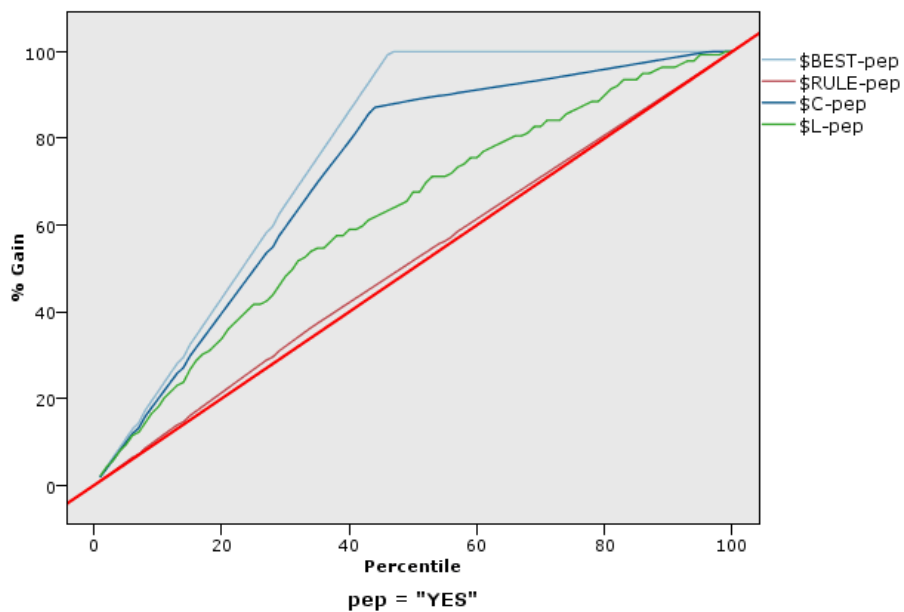
There are five types of evaluation charts, each of which emphasizes a different evaluation criterion.

Gains Charts

Gains are defined as the proportion of total hits that occurs in each quantile. Gains are computed as $(\text{number of hits in quantile} / \text{total number of hits}) \times 100\%$.

Figure 5-60

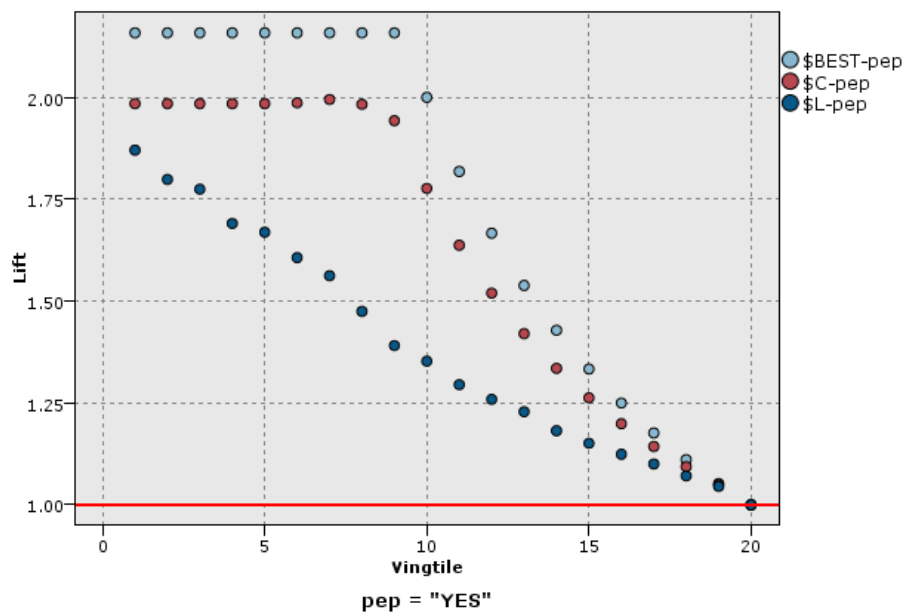
Gains chart (cumulative) with baseline, best line, and business rule displayed



Lift Charts

Lift compares the percentage of records in each quantile that are hits with the overall percentage of hits in the training data. It is computed as $(\text{hits in quantile} / \text{records in quantile}) / (\text{total hits} / \text{total records})$.

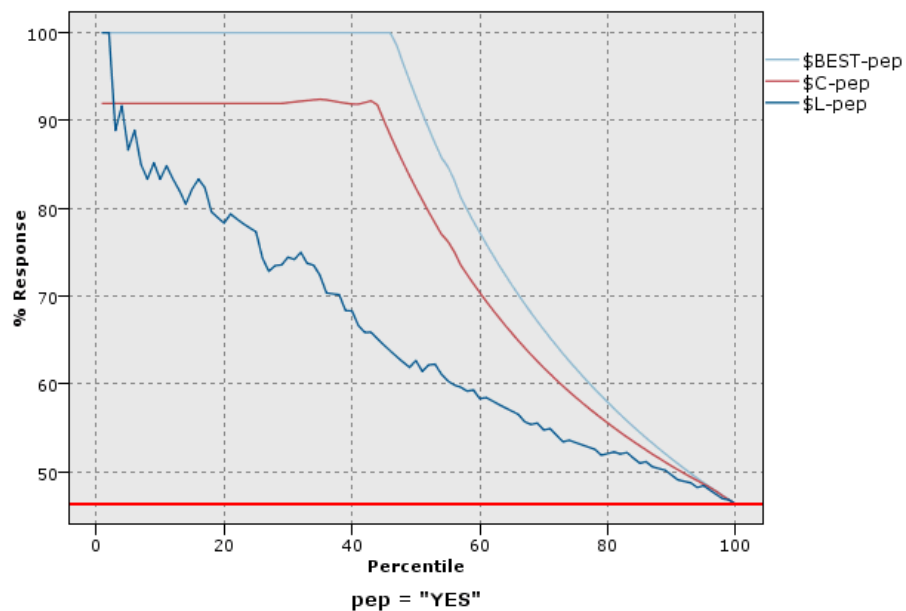
Figure 5-61
Lift chart (cumulative) using points and best line



Response Charts

Response is simply the percentage of records in the quantile that are hits. Response is computed as $(\text{hits in quantile} / \text{records in quantile}) \times 100\%$.

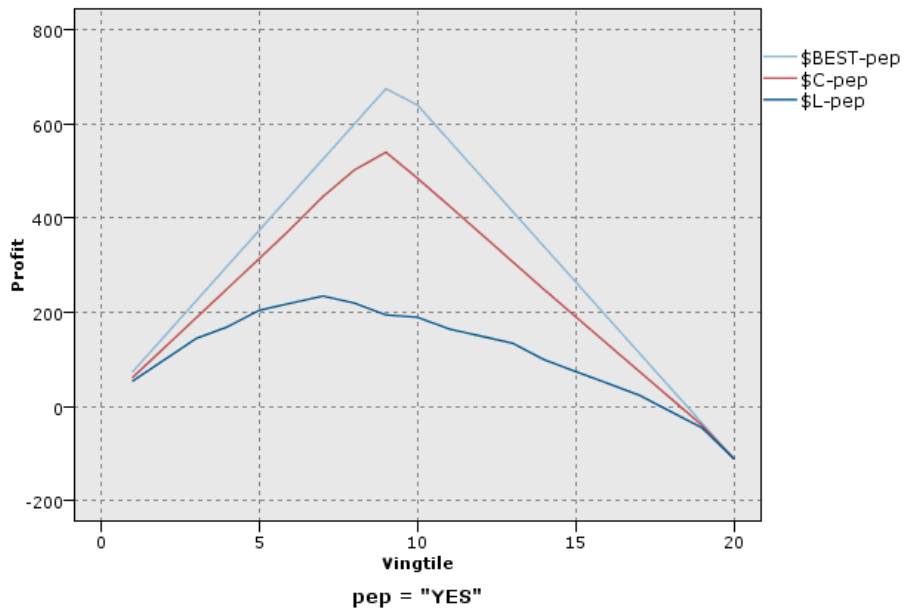
Figure 5-62
Response chart (cumulative) with best line



Profit Charts

Profit equals the **revenue** for each record minus the **cost** for the record. Profits for a quantile are simply the sum of profits for all records in the quantile. Revenues are assumed to apply only to hits, but costs apply to all records. Profits and costs can be fixed or can be defined by fields in the data. Profits are computed as (sum of revenue for records in quantile – sum of costs for records in quantile).

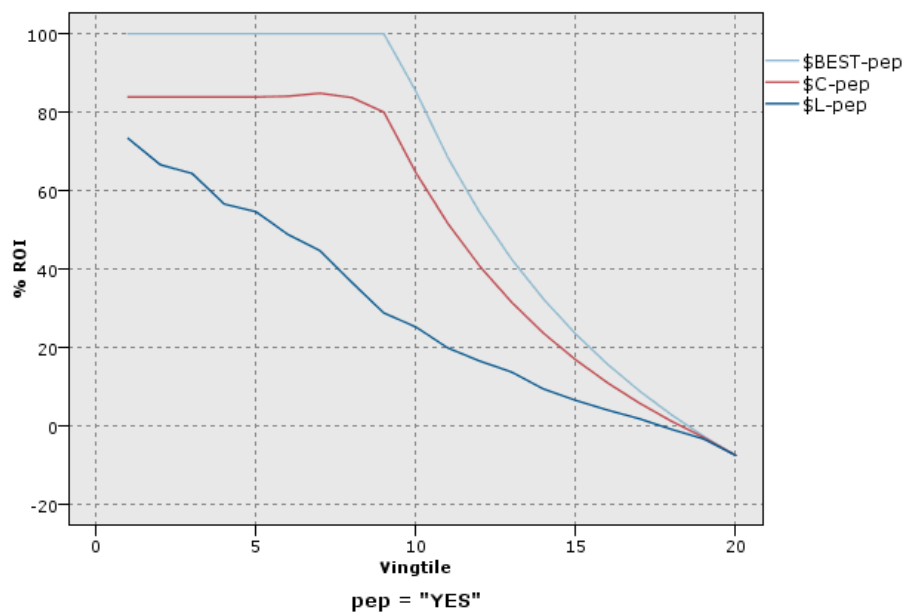
Figure 5-63
Profit chart (cumulative) with best line



ROI Charts

ROI (return on investment) is similar to profit in that it involves defining revenues and costs. ROI compares profits to costs for the quantile. ROI is computed as (profits for quantile / costs for quantile) \times 100%.

Figure 5-64
ROI chart (cumulative) with best line



Evaluation charts can also be cumulative, so that each point equals the value for the corresponding quantile plus all higher quantiles. Cumulative charts usually convey the overall performance of models better, whereas noncumulative charts often excel at indicating particular problem areas for models.

Evaluation Plot Tab

Figure 5-65
Plot tab settings for an Evaluation node

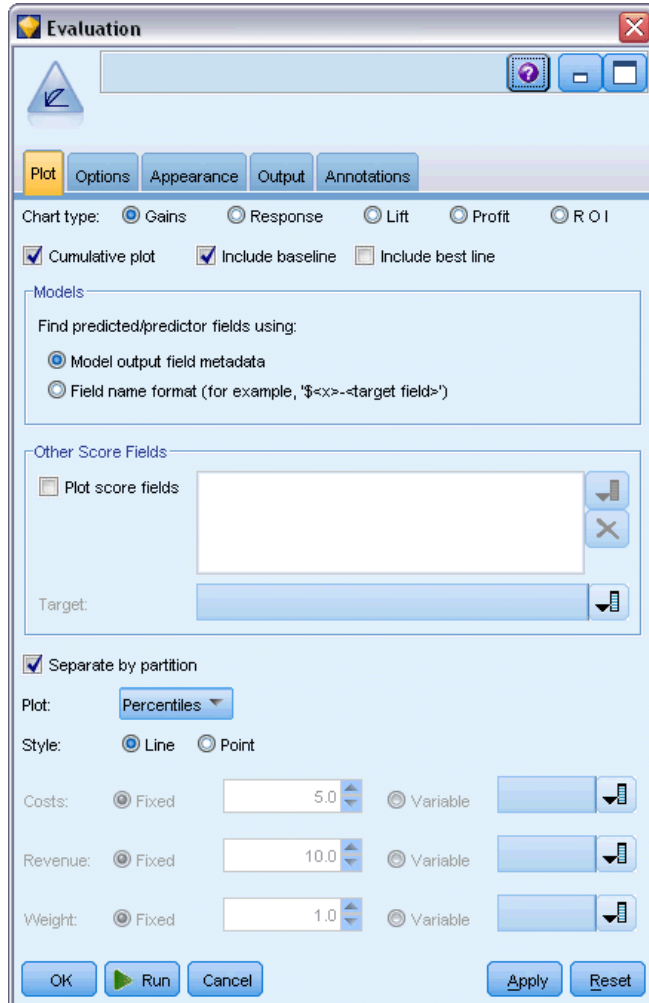


Chart type. Select one of the following types: Gains, Response, Lift, Profit, or ROI (return on investment).

Cumulative plot. Select to create a cumulative chart. Values in cumulative charts are plotted for each quantile plus all higher quantiles.

Include baseline. Select to include a baseline in the plot, indicating a perfectly random distribution of hits where confidence becomes irrelevant. (Include baseline is not available for Profit and ROI charts.)

Include best line. Select to include a best line in the plot, indicating perfect confidence (where hits = 100% of cases).

Find predicted/predictor fields using. Select either Model output field metadata to search for the predicted fields in the graph using their metadata, or select Field name format to search for them by name.

Plot score fields. Select this check box to enable the score fields chooser. Then select one or more range, or continuous, score fields; that is, fields which are not strictly predictive models but which might be useful to rank records in terms of propensity to be a hit. The Evaluation node can compare any combination of one or more score fields with one or more predictive models. A typical example might be to compare several RFM fields with your best predictive model.

Target. Select the target field using the field chooser. Choose any instantiated flag or nominal field with two or more values.

Note: This target field is only applicable to score fields (predictive models define their own targets), and is ignored if a custom hit criterion is set on the Options tab.

Split by partition. If a partition field is used to split records into training, test, and validation samples, select this option to display a separate evaluation chart for each partition. For more information, see the topic [Partition Node](#) in Chapter 4 on p. 176.

Note: When splitting by partition, records with null values in the partition field are excluded from the evaluation. This will never be an issue if a Partition node is used, since Partition nodes do not generate null values.

Plot. Select the size of quantiles to plot in the chart from the drop-down list. Options include Quartiles, Quintiles, Deciles, Vingtiles, Percentiles, and 1000-tiles.

Style. Select Line or Point.

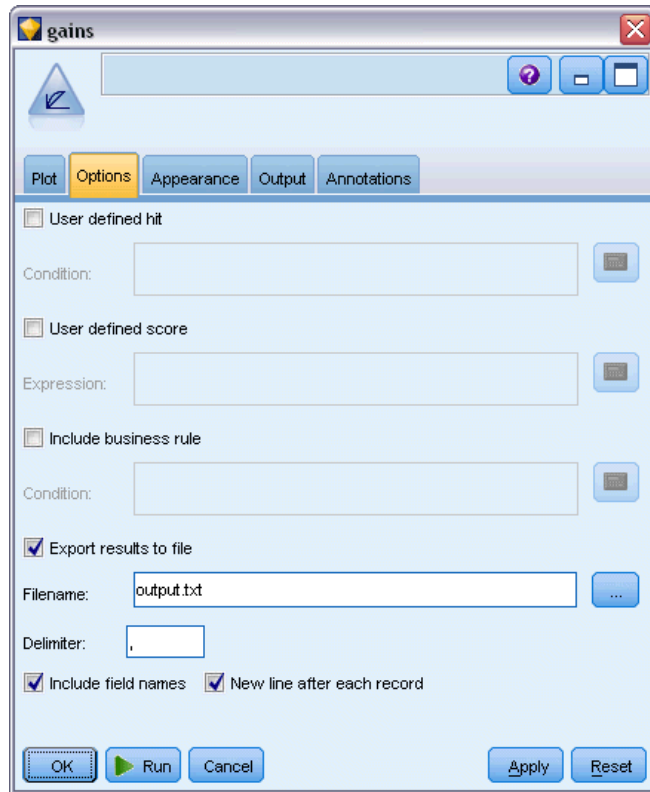
Profit and ROI charts. For Profit and ROI charts, additional controls allow you to specify costs, revenue, and weights.

- **Costs.** Specify the cost associated with each record. You can select Fixed or Variable costs. For fixed costs, specify the cost value. For variable costs, click the Field Chooser button to select a field as the cost field.
- **Revenue.** Specify the revenue associated with each record that represents a hit. You can select Fixed or Variable costs. For fixed revenue, specify the revenue value. For variable revenue, click the Field Chooser button to select a field as the revenue field.
- **Weight.** If the records in your data represent more than one unit, you can use frequency weights to adjust the results. Specify the weight associated with each record, using Fixed or Variable weights. For fixed weights, specify the weight value (the number of units per record). For variable weights, click the Field Chooser button to select a field as the weight field.

Evaluation Options Tab

The Options tab for evaluation charts provides flexibility in defining hits, scoring criteria, and business rules displayed in the chart. You can also set options for exporting the results of the model evaluation.

Figure 5-66
Options tab settings for an Evaluation node



User defined hit. Select to specify a custom condition used to indicate a hit. This option is useful for defining the outcome of interest rather than deducing it from the type of target field and the order of values.

- **Condition.** When User defined hit is selected above, you must specify a CLEM expression for a hit condition. For example, `@TARGET = "YES"` is a valid condition indicating that a value of *Yes* for the target field will be counted as a hit in the evaluation. The specified condition will be used for all target fields. To create a condition, type in the field or use the Expression Builder to generate a condition expression. If the data are instantiated, you can insert values directly from the Expression Builder.

User defined score. Select to specify a condition used for scoring cases before assigning them to quantiles. The default score is calculated from the predicted value and the confidence. Use the Expression field to create a custom scoring expression.

- **Expression.** Specify a CLEM expression used for scoring. For example, if a numeric output in the range 0–1 is ordered so that lower values are better than higher, you might define a hit as `@TARGET < 0.5` and the associated score as `1 - @PREDICTED`. The score expression must result in a numeric value. To create a condition, type in the field or use the Expression Builder to generate a condition expression.

Include business rule. Select to specify a rule condition reflecting criteria of interest. For example, you may want to display a rule for all cases where `mortgage = "Y"` and `income >= 33000`. Business rules are drawn on the chart and labeled in the key as *Rule*.

- **Condition.** Specify a CLEM expression used to define a business rule in the output chart. Simply type in the field or use the Expression Builder to generate a condition expression. If the data are instantiated, you can insert values directly from the Expression Builder.

Export results to file. Select to export the results of the model evaluation to a delimited text file. You can read this file to perform specialized analyses on the calculated values. Set the following options for export:

- **Filename.** Enter the filename for the output file. Use the ellipsis button (...) to browse to the desired folder.
- **Delimiter.** Enter a character, such as a comma or space, to use as the field delimiter.

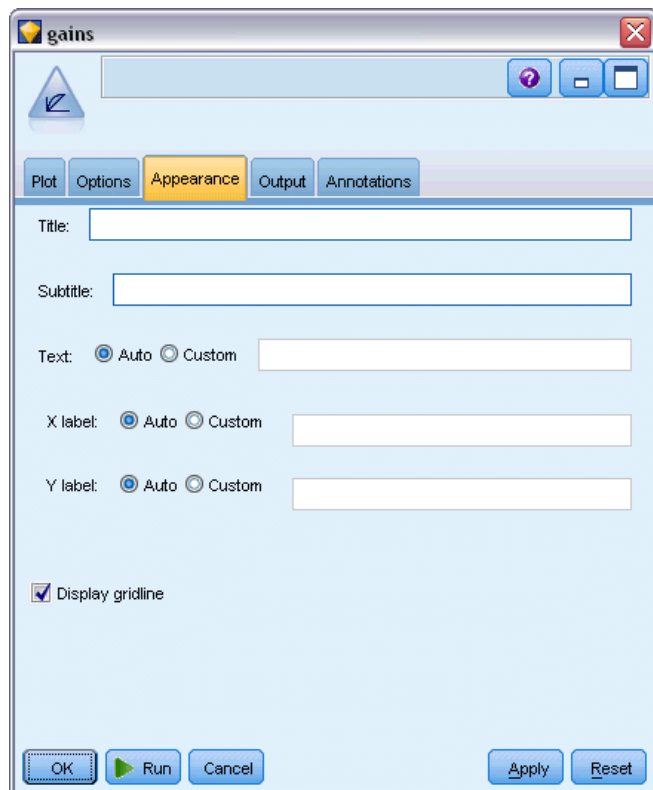
Include field names. Select this option to include field names as the first line of the output file.

New line after each record. Select this option to begin each record on a new line.

Evaluation Appearance Tab

You can specify appearance options before graph creation.

Figure 5-67
Appearance tab settings for an Evaluation node



Title. Enter the text to be used for the graph's title.

Subtitle. Enter the text to be used for the graph's subtitle.

Text. Either accept the automatically generated text label, or select Custom to specify a label.

X label. Either accept the automatically generated x-axis (horizontal) label or select Custom to specify a label.

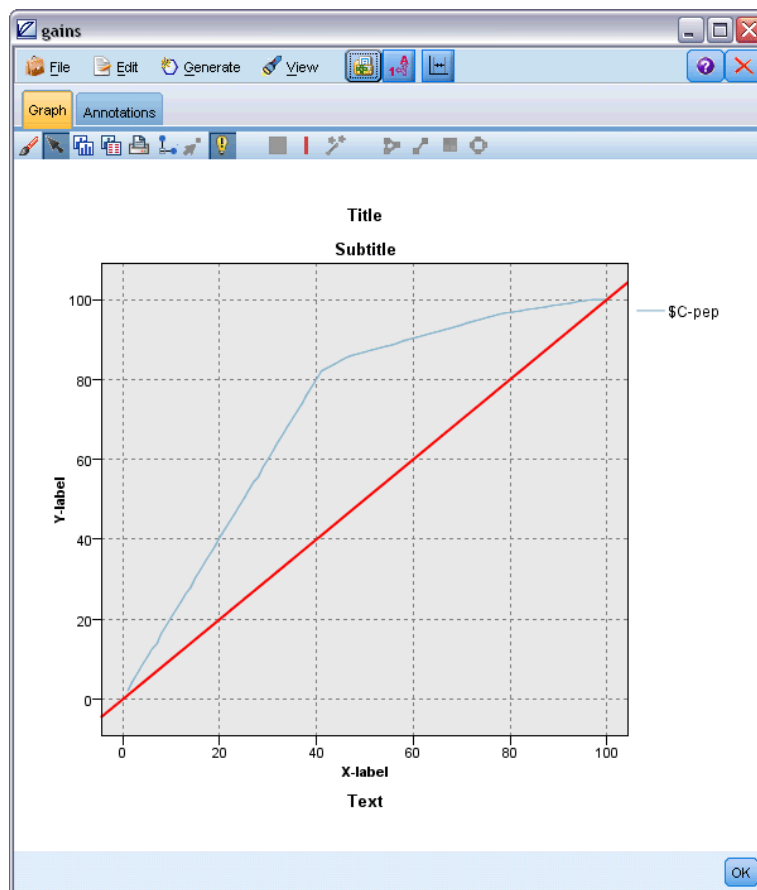
Y label. Either accept the automatically generated y-axis (vertical) label or select Custom to specify a label.

Display gridline. Selected by default, this option displays a gridline behind the plot or graph that enables you to more easily determine region and band cutoff points. Gridlines are always displayed in white unless the graph background is white; in this case, they are displayed in gray.

The following example shows where the appearance options are placed on the graph.

Figure 5-68

Position of graph appearance options on Evaluation graph



Reading the Results of a Model Evaluation

The interpretation of an evaluation chart depends to a certain extent on the type of chart, but there are some characteristics common to all evaluation charts. For cumulative charts, higher lines indicate better models, especially on the left side of the chart. In many cases, when comparing multiple models the lines will cross, so that one model will be higher in one part of the chart and another will be higher in a different part of the chart. In this case, you need to consider what portion of the sample you want (which defines a point on the x axis) when deciding which model to choose.

Most of the noncumulative charts will be very similar. For good models, noncumulative charts should be high toward the left side of the chart and low toward the right side of the chart. (If a noncumulative chart shows a sawtooth pattern, you can smooth it out by reducing the number of quantiles to plot and re-executing the graph.) Dips on the left side of the chart or spikes on the right side can indicate areas where the model is predicting poorly. A flat line across the whole graph indicates a model that essentially provides no information.

Gains charts. Cumulative gains charts always start at 0% and end at 100% as you go from left to right. For a good model, the gains chart will rise steeply toward 100% and then level off. A model that provides no information will follow the diagonal from lower left to upper right (shown in the chart if Include baseline is selected).

Lift charts. Cumulative lift charts tend to start above 1.0 and gradually descend until they reach 1.0 as you go from left to right. The right edge of the chart represents the entire dataset, so the ratio of hits in cumulative quantiles to hits in data is 1.0. For a good model, lift should start well above 1.0 on the left, remain on a high plateau as you move to the right, and then trail off sharply toward 1.0 on the right side of the chart. For a model that provides no information, the line will hover around 1.0 for the entire graph. (If Include baseline is selected, a horizontal line at 1.0 is shown in the chart for reference.)

Response charts. Cumulative response charts tend to be very similar to lift charts except for the scaling. Response charts usually start near 100% and gradually descend until they reach the overall response rate (total hits / total records) on the right edge of the chart. For a good model, the line will start near or at 100% on the left, remain on a high plateau as you move to the right, and then trail off sharply toward the overall response rate on the right side of the chart. For a model that provides no information, the line will hover around the overall response rate for the entire graph. (If Include baseline is selected, a horizontal line at the overall response rate is shown in the chart for reference.)

Profit charts. Cumulative profit charts show the sum of profits as you increase the size of the selected sample, moving from left to right. Profit charts usually start near 0, increase steadily as you move to the right until they reach a peak or plateau in the middle, and then decrease toward the right edge of the chart. For a good model, profits will show a well-defined peak somewhere in the middle of the chart. For a model that provides no information, the line will be relatively straight and may be increasing, decreasing, or level depending on the cost/revenue structure that applies.

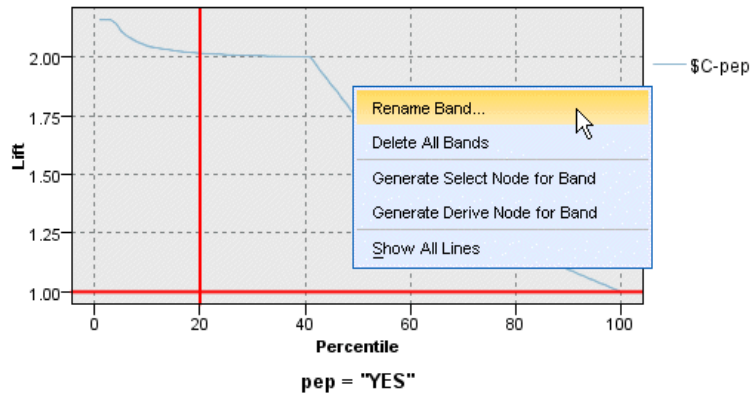
ROI charts. Cumulative ROI (return on investment) charts tend to be similar to response charts and lift charts except for the scaling. ROI charts usually start above 0% and gradually descend until they reach the overall ROI for the entire dataset (which can be negative). For a good model, the line should start well above 0%, remain on a high plateau as you move to the right, and then trail

off rather sharply toward the overall ROI on the right side of the chart. For a model that provides no information, the line should hover around the overall ROI value.

Using an Evaluation Chart

Using the mouse to explore an evaluation chart is similar to using a histogram or collection graph. The x axis represents model scores across the specified quantiles, such as vingtiles or deciles.

Figure 5-69
Working with an evaluation chart

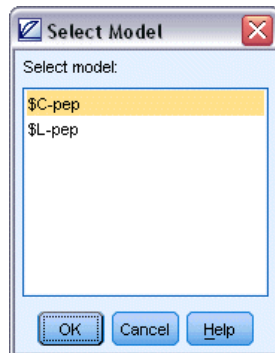


You can partition the x axis into bands just as you would for a histogram by using the splitter icon to display options for automatically splitting the axis into equal bands. For more information, see the topic [Exploring Graphs](#) on p. 281. You can manually edit the boundaries of bands by selecting Graph Bands from the Edit menu.

Once you have created an evaluation chart, defined bands, and examined the results, you can use options on the Generate menu and the context menu to automatically create nodes based upon selections in the graph. For more information, see the topic [Generating Nodes from Graphs](#) on p. 288.

When generating nodes from an evaluation chart, you will be prompted to select a single model from all available models in the chart.

Figure 5-70
Selecting a model for node generation



Select a model and click OK to generate the new node onto the stream canvas.

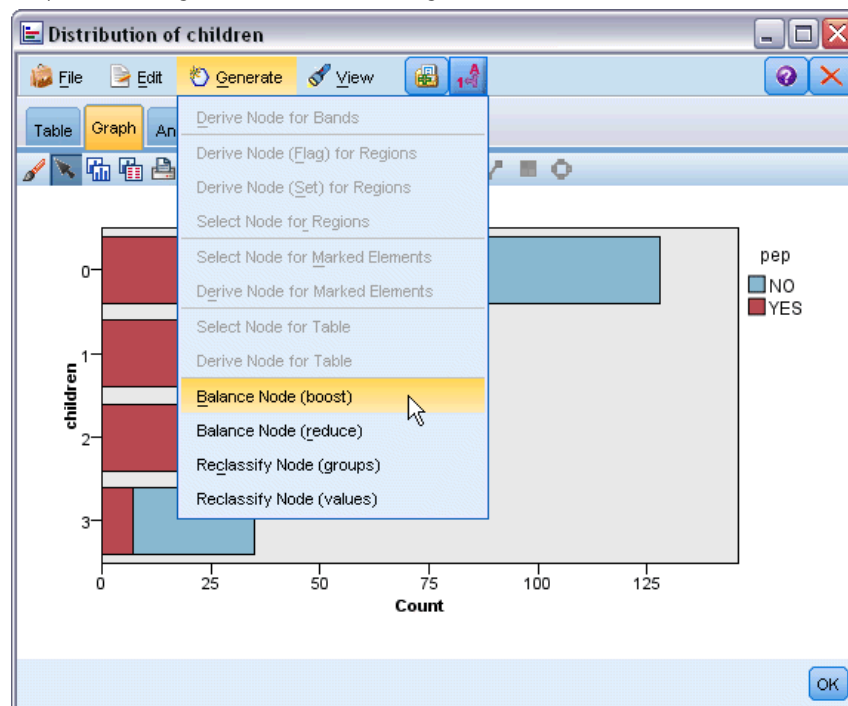
Exploring Graphs

While Edit mode allows you to edit the graph's layout and look, Explore mode allows you to analytically explore the data and values represented by the graph. The main goal of exploration is to analyze the data and then identify values using bands, regions, and marking to generate Select, Derive, or Balance nodes. To select this mode, choose View > Explore Mode from the menus (or click the toolbar icon).

While some graphs can use all of the exploration tools, others accept only one. Explore mode includes:

- Defining and editing bands, which are used to split the values along a scale x axis. For more information, see the topic [Using Bands](#) on p. 282.
- Defining and editing regions, which are used to identify a group of values within the rectangular area. For more information, see the topic [Using Regions](#) on p. 285.
- Marking and unmarking elements to hand select the values that could be used to generate a Select or Derive node. For more information, see the topic [Using Marked Elements](#) on p. 287.
- Generating nodes using the values identified by bands, regions, marked elements, and web links to use in your stream. For more information, see the topic [Generating Nodes from Graphs](#) on p. 288.

Figure 5-71
Graph with the generate menu showing



Using Bands

In any graph with a scale field on the x axis, you can draw vertical band lines to split the range of values on the x axis. If a graph has multiple panels, a band line drawn in one panel is also represented in the other panels as well.

Not all graphs accept bands. Some of those graphs which can have bands include: histograms, bar charts and distributions, plots (line, scatter, time, etc.), collections, and evaluation charts. In graphs with paneling, bands appears in all panels. And in some cases in a SPLOM, you will see a horizontal band line since the axis on which the field/variable band was drawn has been flipped.

Figure 5-72
Graph with three bands

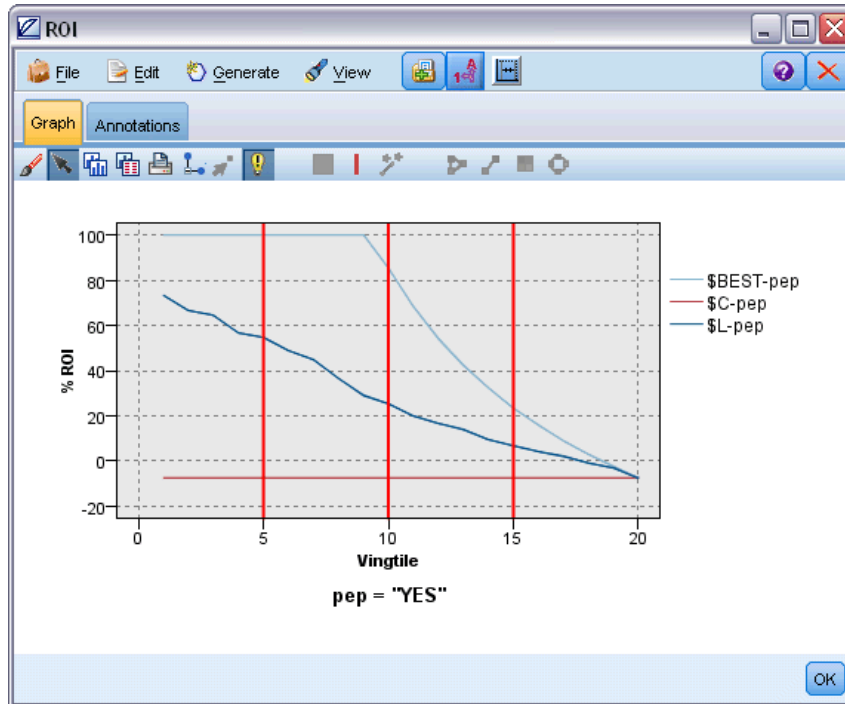
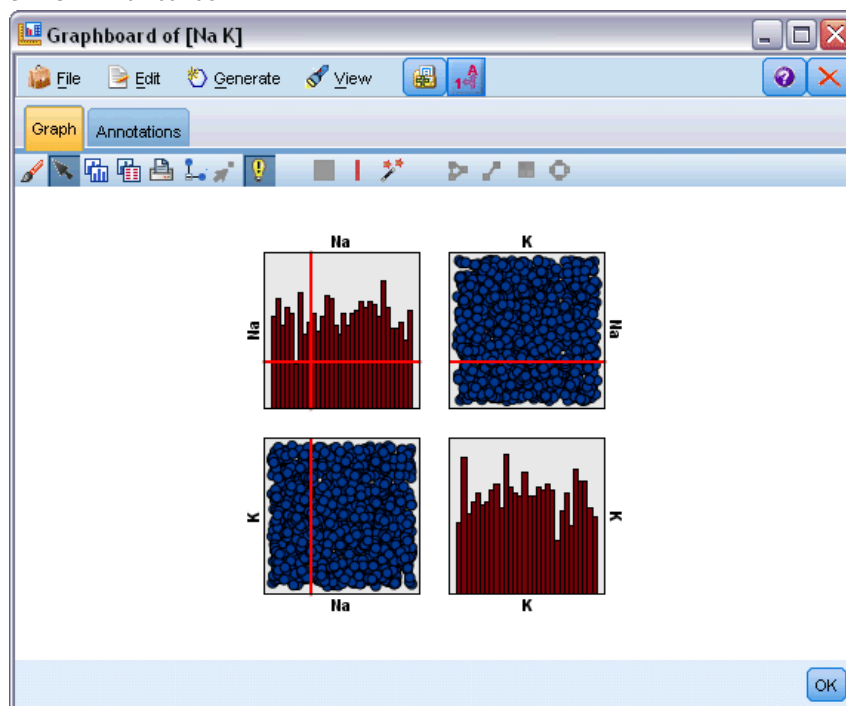


Figure 5-73
SPLOM with bands



Defining Bands

In a graph without bands, adding a band line splits the graph into two bands. The band line value represents the starting point, also referred to as the lower bound, of the second band when reading the graph from left to right. Likewise, in a graph with two bands, adding a band line splits one of those bands into two, which results in three bands. By default, bands are named *bandN*, where *N* equals the number of bands from left to right on the *x* axis.

Once you have defined a band, you can drag-and-drop the band to reposition it on the *x* axis. You can see more shortcuts by right-clicking within a band for tasks such as renaming, deleting, or generating nodes for that specific band.

To define bands:

- ▶ Verify that you are in Explore mode. From the menus, choose View > Explore Mode.
- ▶ In the Explore mode toolbar, click the Draw Band button.

Figure 5-74
Draw Bands toolbar button



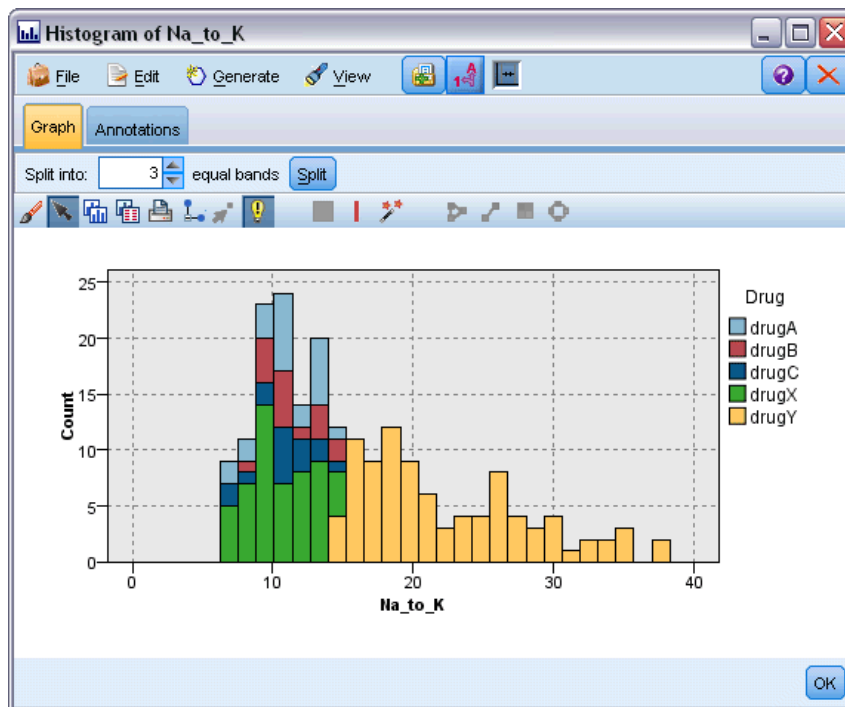
- ▶ In a graph that accepts bands, click the *x*-axis value point at which you want to define a band line.

Note: Alternatively, click the Split Graph into Bands toolbar icon and enter the number of equal bands you want and click Split.

Figure 5-75
 Splitter icon used to expand the toolbar with options for splitting into bands



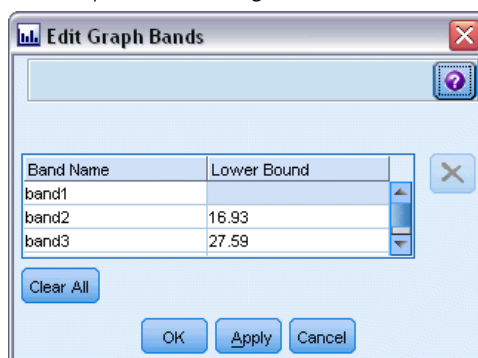
Figure 5-76
 Creating equal bands toolbar with bands enabled



Editing, Renaming, and Deleting Bands

You can edit the properties of existing bands in the Edit Graph Bands dialog box or through context menus in the graph itself.

Figure 5-77
 Edit Graph Bands dialog box



To edit bands:

- ▶ Verify that you are in Explore mode. From the menus, choose View > Explore Mode.
- ▶ In the Explore mode toolbar, click the Draw Band button.
- ▶ From the menus, choose Edit > Graph Bands. The Edit Graph Bands dialog box opens.
- ▶ If you have multiple fields in your graph (such as with SPLOM graphs), you can select the field you want in the drop-down list.
- ▶ Add a new band by typing a name and lower bound. Press the Enter key to begin a new row.
- ▶ Edit a band's boundary by adjusting the Lower Bound value.
- ▶ Rename a band by entering a new band name.
- ▶ Delete a band by selecting the line in the table and clicking the delete button.
- ▶ Click OK to apply your changes and close the dialog box.

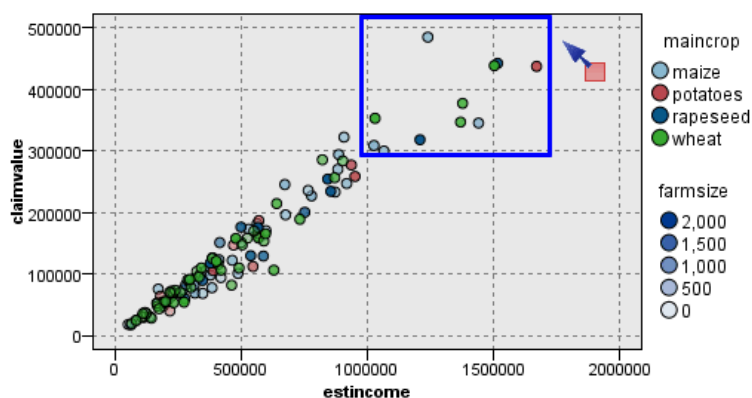
Note: Alternatively, you can delete and rename bands directly in the graph by right-clicking the band's line and choosing the option you want from the context menus.

Using Regions

In any graph with two scale (or range) axes, you can draw regions to group values within a rectangular area you draw, called a region. A **region** is an area of the graph described by its minimum and maximum *X* and *Y* values. If a graph has multiple panels, a region drawn in one panel is also represented in the other panels as well.

Not all graphs accept regions. Some of those graphs that accept regions include: plots (line, scatter, bubble, time, etc.), SPLOM, and collections. These regions are drawn in *X,Y* space and cannot, therefore, be defined in 1-D, 3-D, or animated plots. In graphs with paneling, regions appear in all panels. With a scatterplot matrix (SPLOM), a corresponding region will appear in the corresponding upper plots but not on the diagonal plots since they show only one scale field.

Figure 5-78
Defining a region of high claim values



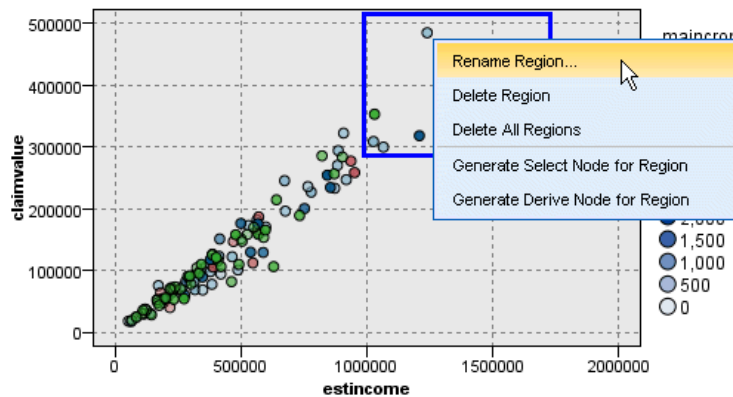
Defining Regions

Wherever you define a region, you are creating a grouping of values. By default, each new region is called *Region<N>*, where *N* corresponds to the number of regions already created.

Once you have defined a region, you can right-click the region line to get some basic shortcuts. However, you can see many other shortcuts by right-clicking inside the region (not on the line) for tasks such as renaming, deleting, or generating Select and Derive nodes for that specific region.

You can select subsets of records on the basis of their inclusion in a particular region or in one of several regions. You can also incorporate region information for a record by producing a Derive node to flag records based on their inclusion in a region. For more information, see the topic [Generating Nodes from Graphs](#) on p. 288.

Figure 5-79
Exploring the region of high claim values



To define regions:

- ▶ Verify that you are in Explore mode. From the menus, choose View > Explore Mode.
- ▶ In the Explore mode toolbar, click the Draw Region button.

Figure 5-80
Draw Region toolbar button

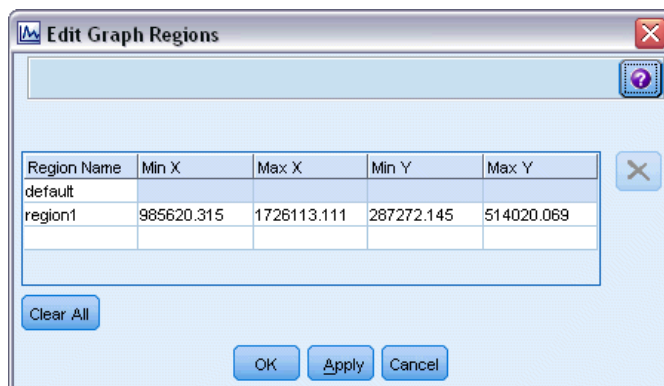


- ▶ In a graph that accepts regions, click and drag your mouse to draw the rectangular region.

Editing, Renaming and Deleting Regions

You can edit the properties of existing regions in the Edit Graph Regions dialog box or through context menus in the graph itself.

Figure 5-81
Specifying properties for the defined regions



To edit regions:

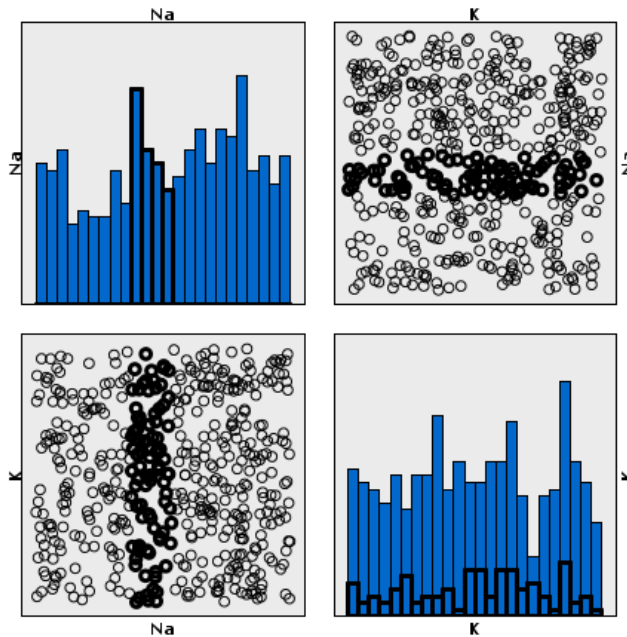
- ▶ Verify that you are in Explore mode. From the menus, choose View > Explore Mode.
- ▶ In the Explore mode toolbar, click the Draw Region button.
- ▶ From the menus, choose Edit > Graph Regions. The Edit Graph Regions dialog box opens.
- ▶ If you have multiple fields in your graph (for example, SPLOM graphs), you must define the field for the region in the *Field A* and *Field B* columns.
- ▶ Add a new region on a new line by typing a name, selecting field names (if applicable) and defining the maximum and minimum boundaries for each field. Press the Enter key to begin a new row.
- ▶ Edit existing region boundaries by adjusting the Min and Max values for *A* and *B*.
- ▶ Rename a region by changing the region's name in the table.
- ▶ Delete a region by selecting the line in the table and clicking the delete button.
- ▶ Click OK to apply your changes and close the dialog box.

Note: Alternatively, you can delete and rename regions directly in the graph by right-clicking the region's line and choosing the option you want from the context menus.

Using Marked Elements

You can mark elements, such as bars, slices, and points, in any graph. Lines, areas, and surfaces cannot be marked in graphs other than time plot, multiplot, and evaluation graphs since lines refers to fields in those cases. Whenever you mark an element, you are essentially highlighting all of the data represented by that element. In any graph where the same case is represented in more than one place (such as SPLOM), marking is synonymous with brushing. You can mark elements in graphs and even within bands and regions. Whenever you mark an element and then go back into Edit mode, the marking still remains visible.

Figure 5-82
Marking elements in a SPLOM



You can mark and unmark elements by clicking on elements in the graph. When you first click an element to mark it, the element appears with a thick border color to indicate that it has been marked. If you click the element again, the border disappears and the element is no longer marked. To mark multiple elements, you can either hold down the Ctrl key while clicking elements, or you can drag the mouse around each of the elements you want marked using the “magic wand”. Remember that if you click another area or element without holding down the Ctrl key, all previously marked elements are cleared.

You can generate Select and Derive nodes from the marked elements in your graph. For more information, see the topic [Generating Nodes from Graphs](#) on p. 288.

To mark elements:

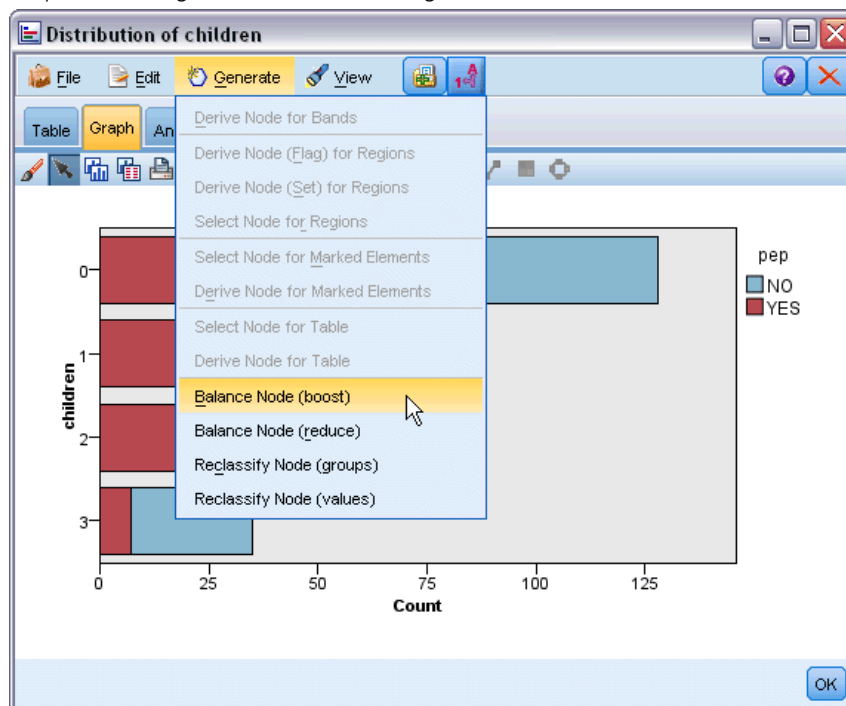
- ▶ Verify that you are in Explore mode. From the menus, choose View > Explore Mode.
- ▶ In the Explore mode toolbar, click the Mark Elements button.
- ▶ Either click on the element you require, or click and drag your mouse to draw a line around the region containing multiple elements.

Generating Nodes from Graphs

One of the most powerful features offered by IBM® SPSS® Modeler graphs is the ability to generate nodes from a graph or a selection within the graph. For example, from a time plot graph, you can generate Derive and Select nodes based on a selection or region of data, effectively “subsetting” the data. For example, you might use this powerful feature to identify and exclude outliers.

Whenever you can draw a band, you can also generate a Derive node. In graphs with two scale axes, you can generate Derive or Select nodes from the regions drawn in your graph. In graphs with marked elements, you can generate Derive nodes, Select nodes, and in some cases Filter nodes from these elements. Balance node generation is enabled for any graph showing a distribution of counts.

Figure 5-83
Graph with the generate menu showing



Whenever you generate a node, it is placed on the stream canvas directly so that you can connect it to an existing stream. The following nodes can be generated from graphs: Select, Derive, Balance, Filter, and Reclassify.

Select Nodes

Select nodes can be generated to test for inclusion of the records within a region and exclusion of all records falling outside the region or the reverse for downstream processing.

- **For bands.** You can generate a Select node that includes or excludes the records within that band. Select node for Bands only is only available through contextual menus since you need to select which band to use in the Select node.
- **For regions.** You can generate a Select node that includes or excludes the records within a region.
- **For marked elements.** You can generate Select nodes to capture the records corresponding to the marked elements or web graph links.

Derive Nodes

Derive nodes can be generated from regions, bands, and marked elements. All graphs can produce Derive nodes. In the case of evaluation charts, a dialog box for selecting the model appears. In the case of web graphs, Derive Node (“And”) and Derive Node (“Or”) are possible.

- **For bands.** You can generate a Derive node that produces a category for each interval marked on the axis, using the band names listed in the Edit Bands dialog box as category names.
- **For regions.** You can generate a Derive node (Derive as flag) that creates a flag field called *in_region* with the flags set to *T* for records inside any region and *F* for records outside all regions. You can also generate a Derive node (Derive as set) that produces a set with a value for each region with a new field called *region* for each record, which takes as its value the name of the region into which the records fall. Records falling outside all regions receive the name of the default region. Value names become the region names listed in the Edit regions dialog box.
- **For marked elements.** You can generate a Derive node that calculates a flag that is *True* for all marked elements and *False* for all other records.

Balance Nodes

Balance nodes can be generated to correct imbalances in the data, such as reducing the frequency of common values (use Balance Node (reduce) menu option) or boosting the occurrence of infrequent values (use Balance Node (boost) menu option). Balance node generation is enabled for any graph showing a distribution of counts, such as Histogram, Dot, Collection, Bar of Counts, Pie of Counts, and Multiplot.

Filter Nodes

Filter nodes can be generated to rename or filter fields based on the lines or nodes marked in the graph. In the case of evaluation charts, the best fit line does not generate a filter node.

Reclassify Nodes

Reclassify nodes can be generated to recode values. This option is used for distribution graphs. You can generate a Reclassify node for **groups** to recode specific values of a displayed field depending upon their inclusion in a group (select groups using Ctrl+click on the Tables tab). You can also generate a reclassify node for **values** to recode data into an existing set of numerous values, such as reclassifying data into a standard set of values in order to merge financial data from various companies for analysis.

Note: If the values are predefined, you can read them into SPSS Modeler as a flat file and use a distribution to display all values. Then generate a Reclassify (values) node for this field directly from the chart. Doing so will put all the target values in the Reclassify node’s *New values* column (drop-down list).

Generating Nodes from Graphs

You can use the Generate menu in the graph output window to generate nodes. The generated node will be placed on the stream canvas. To use the node, connect it to an existing stream.

To generate a node from a graph:

- ▶ Verify that you are in Explore mode. From the menus, choose View > Explore Mode.
- ▶ In the Explore mode toolbar, click the Region button.
- ▶ Define bands, regions, or any marked elements needed to generate your node.
- ▶ From the Generate menu, choose the kind of node you want to produce. Only those which are possible are enabled.

Note: Alternatively, you can also generate nodes directly from the graph by right-clicking and choosing the generate option you want from the context menus.

Editing Visualizations

While Explore mode allows you to analytically explore the data and values represented by the visualization, Edit mode allows you to change the visualization's layout and look. For example, you can change the fonts and colors to match your organization's style guide. To select this mode, choose View > Edit Mode from the menus (or click the toolbar icon).

In Edit mode, there are several toolbars that affect different aspects of the visualization's layout. If you find that there are any you don't use, you can hide them to increase the amount of space in the dialog box in which the graph is displayed. To select or deselect toolbars, click on the relevant toolbar name on the View menu.

Note: To add further detail to your visualizations, you can apply title, footnote, and axis labels. For more information, see the topic [Adding Titles and Footnotes](#) on p. 307.

You have several options for editing a visualization in **Edit mode**. You can:

- Edit text and format it.
- Change the fill color, transparency, and pattern of frames and graphic elements.
- Change the color and dashing of borders and lines.
- Rotate and change the shape and aspect ratio of point elements.
- Change the size of graphic elements (such as bars and points).
- Adjust the space around items by using margins and padding.
- Specify formatting for numbers.
- Change the axis and scale settings.
- Sort, exclude, and collapse categories on a categorical axis.
- Set the orientation of panels.
- Apply transformations to a coordinate system.
- Change statistics, graphic element types, and collision modifiers.
- Change the position of the legend.
- Apply visualization stylesheets.

The following topics describe how to perform these various tasks. It is also recommended that you read the general rules for editing graphs.

How to Switch to Edit Mode

- ▶ From the menus choose:
View > Edit Mode

General Rules for Editing Visualizations**Edit Mode**

All edits are done in Edit mode. To enable Edit mode, from the menus choose:
View > Edit Mode

Selection

The options available for editing depend on selection. Different toolbar and properties palette options are enabled depending on what is selected. Only the enabled items apply to the current selection. For example, if an axis is selected, the Scale, Major Ticks, and Minor Ticks tabs are available in the properties palette.

Here are some tips for selecting items in the visualization:

- Click an item to select it.
- Select a graphic element (such as points in a scatterplot or bars in a bar chart) with a single click. After initial selection, click again to narrow the selection to groups of graphic elements or a single graphic element.
- Press Esc to deselect everything.

Palettes

When an item is selected in the visualization, the various palettes are updated to reflect the selection. The palettes contain controls for making edits to the selection. Palettes may be toolbars or a panel with multiple controls and tabs. Palettes can be hidden, so ensure the necessary palette is displayed for making edits. Check the View menu for palettes that are currently displayed.

You can reposition the palettes by clicking and dragging the empty space in a toolbar palette or the left side of other palettes. Visual feedback lets you know where you can dock the palette. For non-toolbar palettes, you can also click the close button to hide the palette and the undock button to display the palette in a separate window. Click the help button to display help for the specific palette.

Automatic Settings

Some settings provide an -auto- option. This indicates that automatic values are applied. Which automatic settings are used depends on the specific visualization and data values. You can enter a value to override the automatic setting. If you want to restore the automatic setting, delete the current value and press Enter. The setting will display -auto- again.

Removing/Hiding Items

You can remove/hide various items in the visualization. For example, you can hide the legend or axis label. To delete an item, select it and press Delete. If the item does not allow deletion, nothing will happen. If you accidentally delete an item, press Ctrl+Z to undo the deletion.

State

Some toolbars reflect the state of the current selection, others don't. The properties palette always reflects state. If a toolbar does *not* reflect state, this is mentioned in the topic that describes the toolbar.

Editing and Formatting Text

You can edit text in place and change the formatting of an entire text block. Note that you can't edit text that is linked directly to data values. For example, you can't edit a tick label because the content of the label is derived from the underlying data. However, you can format any text in the visualization.

How to Edit Text in Place

- ▶ Double-click the text block. This action selects all the text. All toolbars are disabled at this time, because you cannot change any other part of the visualization while editing text.
- ▶ Type to replace the existing text. You can also click the text again to display a cursor. Position the cursor in the desired position and enter the additional text.

How to Format Text

- ▶ Select the frame containing the text. Do not double-click the text.
- ▶ Format text using the font toolbar. If the toolbar is not enabled, make sure only the *frame* containing the text is selected. If the text itself is selected, the toolbar will be disabled.

Figure 5-84
Font toolbar



You can change the font:

- Color
- Family (for example, Arial or Verdana)
- Size (the unit is pt unless you indicate a different unit, such as pc)
- Weight
- Alignment relative to the text frame

Formatting applies to all the text in a frame. You can't change the formatting of individual letters or words in any particular block of text.

Changing Colors, Patterns, Dashings, and Transparency

Many different items in a visualization have a fill and border. The most obvious example is a bar in a bar chart. The color of the bars is the fill color. They may also have a solid, black border around them.

There are other less obvious items in the visualization that have fill colors. If the fill color is transparent, you may not know there is a fill. For example, consider the text in an axis label. It appears as if this text is “floating” text, but it actually appears in a frame that has a transparent fill color. You can see the frame by selecting the axis label.

Any frame in the visualization can have a fill and border style, including the frame around the whole visualization. Also, any fill has an associated opacity/transparency level that can be adjusted.

How to Change the Colors, Patterns, Dashing, and Transparency

- ▶ Select the item you want to format. For example, select the bars in a bar chart or a frame containing text. If the visualization is split by a categorical variable or field, you can also select the group that corresponds to an individual category. This allows you to change the default aesthetic assigned to that group. For example, you can change the color of one of the stacking groups in a stacked bar chart.
- ▶ To change the fill color, the border color, or the fill pattern, use the color toolbar.

Figure 5-85
Color toolbar

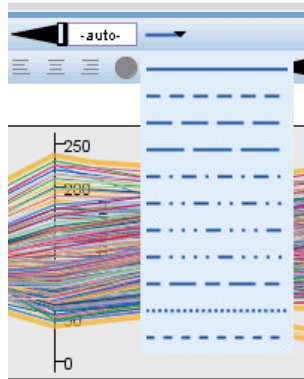


Note: This toolbar does not reflect the state of the current selection.

To change a color or fill, you can click the button to select the displayed option or click the drop-down arrow to choose another option. For colors, notice there is one color that looks like white with a red, diagonal line through it. This is the transparent color. You could use this, for example, to hide the borders on bars in a histogram.

- The first button controls the fill color.
 - The second button controls the border color.
 - The third button controls the fill pattern. The fill pattern uses the border color. Therefore, the fill pattern is visible only if there is a visible border color.
 - The fourth control is a slider and text box that control the opacity of the fill color and pattern. A lower percentage means less opacity and more transparency. 100% is fully opaque (no transparency).
- ▶ To change the dashing of a border or line, use the line toolbar.

Figure 5-86
Line toolbar



Note: This toolbar does not reflect the state of the current selection.

As with the other toolbar, you can click the button to select the displayed option or click the drop-down arrow to choose another option.

Rotating and Changing the Shape and Aspect Ratio of Point Elements

You can rotate point elements, assign a different predefined shape, or change the aspect ratio (the ratio of width to height).

How to Modify Point Elements

- ▶ Select the point elements. You cannot rotate or change the shape and aspect ratio of individual point elements.
- ▶ Use the symbol toolbar to modify the points.

Figure 5-87
Symbol toolbar



- The first button allows you to change the shape of the points. Click the drop-down arrow and select a predefined shape.
- The second button allows you to rotate the points to a specific compass position. Click the drop-down arrow and then drag the needle to the desired position.
- The third button allows you to change the aspect ratio. Click the drop-down arrow and then click and drag the rectangle that appears. The shape of the rectangle represents the aspect ratio.

Changing the Size of Graphic Elements

You can change the size of the graphic elements in the visualization. These include bars, lines, and points among others. If the graphic element is sized by a variable or field, the specified size is the *minimum* size.

How to Change the Size of the Graphic Elements

- ▶ Select the graphic elements you want to resize.
- ▶ Use the slider or enter a specific size for the option available on the symbol toolbar. The unit is pixels unless you indicate a different unit (see below for a full list of unit abbreviations). You can also specify a percentage (such as 30%), which means that a graphic element uses the specified percentage of the available space. The available space depends on the graphic element type and the specific visualization.

Table 5-1
Valid unit abbreviations

Abbreviation	Unit
cm	centimeter
in	inch
mm	millimeter
pc	pica
pt	point
px	pixel

Figure 5-88
Size control on symbol toolbar



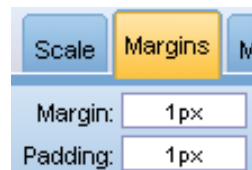
Specifying Margins and Padding

If there is too much or too little spacing around or inside a frame in the visualization, you can change its margin and padding settings. The **margin** is the amount of space between the frame and other items around it. The **padding** is the amount of space between the border of the frame and the *contents* of the frame.

How to Specify Margins and Padding

- ▶ Select the frame for which you want to specify margins and padding. This can be a text frame, the frame around the legend, or even the data frame displaying the graphic elements (such as bars and points).
- ▶ Use the Margins tab on the properties palette to specify the settings. All sizes are in pixels unless you indicate a different unit (such as cm or in).

Figure 5-89
Margins tab



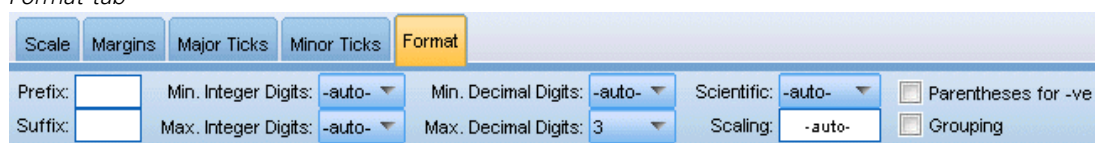
Formatting Numbers

You can specify the format for numbers in tick labels on a continuous axis or data value labels displaying a number. For example, you can specify that numbers displayed in the tick labels are shown in thousands.

How to Specify Number Formats

- ▶ Select the continuous axis tick labels or the data value labels if they contain numbers.
- ▶ Click the Format tab on the properties palette.

Figure 5-90
Format tab



- ▶ Select the desired number formatting options:

Prefix. A character to display at the beginning of the number. For example, enter a dollar sign (\$) if the numbers are salaries in U.S. dollars.

Suffix. A character to display at the end of the number. For example, enter a percentage sign (%) if the numbers are percentages.

Min. Integer Digits. Minimum number of digits to display in the integer part of a decimal representation. If the actual value does not contain the minimum number of digits, the integer part of the value will be padded with zeros.

Max. Integer Digits. Maximum number of digits to display in the integer part of a decimal representation. If the actual value exceeds the minimum number of digits, the integer part of the value will be replaced with asterisks.

Min. Decimal Digits. Minimum number of digits to display in the decimal part of a decimal or scientific representation. If the actual value does not contain the minimum number of digits, the decimal part of the value will be padded with zeros.

Max. Decimal Digits. Maximum number of digits to display in the decimal part of a decimal or scientific representation. If the actual value exceeds the minimum number of digits, the decimal is rounded to the appropriate number of digits.

Scientific. Whether to display numbers in scientific notation. Scientific notation is useful for very large or very small numbers. -auto- lets the application determine when scientific notation is appropriate.

Scaling. A scale factor, which is a number by which the original value is divided. Use a scale factor when the numbers are large, but you don't want the label to extend too much to accommodate the number. If you change the number format of the tick labels, be sure to edit the axis title to indicate how the number should be interpreted. For example, assume your scale axis displays salaries and the labels are 30,000, 50,000, and 70,000. You might enter a scale factor of 1000 to display 30, 50, and 70. You should then edit the scale axis title to include the text in thousands.

Parentheses for -ve. Whether parentheses should be displayed around negative values.

Grouping. Whether to display a character between groups of digits. Your computer’s current locale determines which character is used for digit grouping.

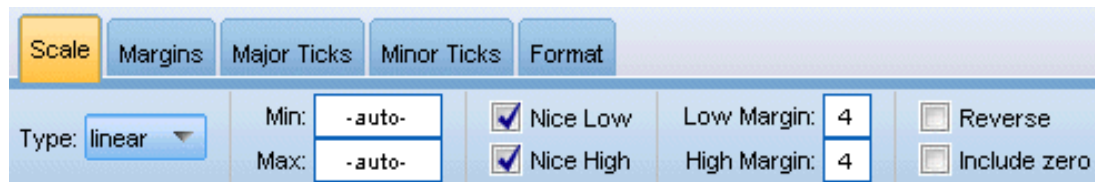
Changing the Axis and Scale Settings

There are several options for modifying axes and scales.

How to Change Axis and Scale Settings

- ▶ Select any part of the axis (for example, the axis label or tick labels).
- ▶ Use the Scale, Major Ticks, and Minor Ticks tabs on the properties palette to change the axis and scale settings.

Figure 5-91
Properties palette



Scale tab

Note: The Scale tab does not appear for graphs where the data is pre-aggregated (for example, histograms).

Type. Specifies whether the scale is linear or transformed. Scale transformations help you understand the data or make assumptions necessary for statistical inference. On scatterplots, you might use a transformed scale if the relationship between the independent and dependent variables or fields is nonlinear. Scale transformations can also be used to make a skewed histogram more symmetric so that it resembles a normal distribution. Note that you are transforming only the scale on which the data are displayed; you are not transforming the actual data.

- **linear.** Specifies a linear, untransformed scale.
- **log.** Specifies a base-10 log transformed scale. To accommodate zero and negative values, this transformation uses a modified version of the log function. This “safe log” function is defined as $\text{sign}(x) * \log(1 + \text{abs}(x))$. So $\text{safeLog}(-99)$ equals:

$$\text{sign}(-99) * \log(1 + \text{abs}(-99)) = -1 * \log(1 + 99) = -1 * 2 = -2$$
- **power.** Specifies a power transformed scale, using an exponent of 0.5. To accommodate negative values, this transformation uses a modified version of the power function. This “safe power” function is defined as $\text{sign}(x) * \text{pow}(\text{abs}(x), 0.5)$. So $\text{safePower}(-100)$ equals:

$$\text{sign}(-100) * \text{pow}(\text{abs}(-100), 0.5) = -1 * \text{pow}(100, 0.5) = -1 * 10 = -10$$

Min/Max/Nice Low/Nice High. Specifies the range for the scale. Selecting Nice Low and Nice High allows the application to select an appropriate scale based on the data. The minimum and maximum are “nice” because they are typically whole values greater or less than the maximum

and minimum data values. For example, if the data range from 4 to 92, a nice low and high for scale may be 0 and 100 rather than the actual data minimum and maximum. Be careful that you don't set a range that is too small and hides important items. Also note that you cannot set an explicit minimum and maximum if the Include zero option is selected.

Low Margin/High Margin. Create margins at the low and/or high end of the axis. The margin appears perpendicular to the selected axis. The unit is pixels unless you indicate a different unit (such as cm or in). For example, if you set the High Margin to 5 for the vertical axis, a horizontal margin of 5 px runs along the top of the data frame.

Reverse. Specifies whether the scale is reversed.

Include zero. Indicates that the scale should include 0. This option is commonly used for bar charts to ensure the bars begin at 0, rather than a value near the height of the smallest bar. If this option is selected, Min and Max are disabled because you cannot set a custom minimum and maximum for the scale range.

Major Ticks/Minor Ticks Tabs

Ticks or tick marks are the lines that appear on an axis. These indicate values at specific intervals or categories. **Major ticks** are the tick marks with labels. These are also longer than other tick marks. **Minor ticks** are tick marks that appear between the major tick marks. Some options are specific to the tick type, but most options are available for major and minor ticks.

Show ticks. Specifies whether major or minor ticks are displayed on a graph.

Show gridlines. Specifies whether gridlines are displayed at the major or minor ticks. **Gridlines** are lines that cross a whole graph from axis to axis.

Position. Specifies the position of the tick marks relative to the axis.

Length. Specifies the length of the tick marks. The unit is pixels unless you indicate a different unit (such as cm or in).

Base. *Applies only to major ticks.* Specifies the value at which the first major tick appears.

Delta. *Applies only to major ticks.* Specifies the difference between major ticks. That is, major ticks will appear at every n th value, where n is the delta value.

Divisions. *Applies only to minor ticks.* Specifies the number of minor tick divisions between major ticks. The number of minor ticks is one less than the number of divisions. For example, assume that there are major ticks at 0 and 100. If you enter 2 as the number of minor tick divisions, there will be *one* minor tick at 50, dividing the 0–100 range and creating *two* divisions.

Editing Categories

You can edit the categories on a categorical axis in several ways:

- Change the sort order for displaying the categories.
- Exclude specific categories.
- Add a category that does not appear in the data set.
- Collapse/combine small categories into one category.

How to Change the Sort Order of Categories

- ▶ Select a categorical axis. The Categories palette displays the categories on the axis.

Note: If the palette is not visible, make sure that you have it enabled. From the View menu in IBM® SPSS® Modeler, choose Categories.

- ▶ In the Categories palette, select a sorting option from the drop-down list:

Custom. Sort categories based on the order in which they appear in the palette. Use the arrow buttons to move categories to the top of the list, up, down, and to the bottom of the list.

Data. Sort categories based on the order in which they occur in the dataset.

Name. Sort categories alphabetically, using the names as displayed in the palette. This may be either the value or label, depending on whether the toolbar button to display values and labels is selected.

Value. Sort categories by the underlying data value, using the values displayed in parentheses in the palette. Only data sources with metadata (such as IBM® SPSS® Statistics data files) support this option.

Statistic. Sort categories based on the calculated statistic for each category. Examples of statistics include counts, percentages, and means. This option is available only if a statistic is used in the graph.

How to Add a Category

By default, only categories that appear in the data set are available. You can add a category to the visualization if needed.

- ▶ Select a categorical axis. The Categories palette displays the categories on the axis.

Note: If the palette is not visible, make sure that you have it enabled. From the View menu in SPSS Modeler, choose Categories.

- ▶ In the Categories palette, click the add category button:

Figure 5-92
Add category button



- ▶ In the Add a new category dialog box, enter a name for the category.
- ▶ Click OK.

How to Exclude Specific Categories

- ▶ Select a categorical axis. The Categories palette displays the categories on the axis.

Note: If the palette is not visible, make sure that you have it enabled. From the View menu in SPSS Modeler, choose Categories.

- ▶ In the Categories palette, select a category name in the Include list, and then click the X button. To move the category back, select its name in the Excluded list, and then click the arrow to the right of the list.

How to Collapse/Combine Small Categories

You can combine categories that are so small you don't need to display them separately. For example, if you have a pie chart with many categories, consider collapsing categories with a percentage less than 10. Collapsing is available only for statistics that are additive. For example, you can't add means together because means are not additive. Therefore, combining/collapsing categories using a mean is not available.

- ▶ Select a categorical axis. The Categories palette displays the categories on the axis.
Note: If the palette is not visible, make sure that you have it enabled. From the View menu in SPSS Modeler, choose Categories.
- ▶ In the Categories palette, select Collapse and specify a percentage. Any categories whose percentage of the total is less than the specified number are combined into one category. The percentage is based on the statistic shown in the chart. Collapsing is available only for count-based and summation (sum) statistics.

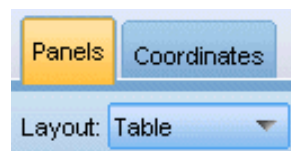
Changing the Orientation Panels

If you are using panels in your visualization, you can change their orientation.

How to Change the Orientation of the Panels

- ▶ Select any part of the visualization.
- ▶ Click the Panels tab on the properties palette.

Figure 5-93
Panels tab



- ▶ Select an option from Layout:
 - Table.** Lays out panels like a table, in that there is a row or column assigned to every individual value.
 - Transposed.** Lays out panels like a table, but also swaps the original rows and columns. This option is not the same as transposing the graph itself. Note that the *x* axis and the *y* axis are unchanged when you select this option.
 - List.** Lays out panels like a list, in that each cell represents a combination of values. Columns and rows are no long assigned to individual values. This option allows the panels to wrap if needed.

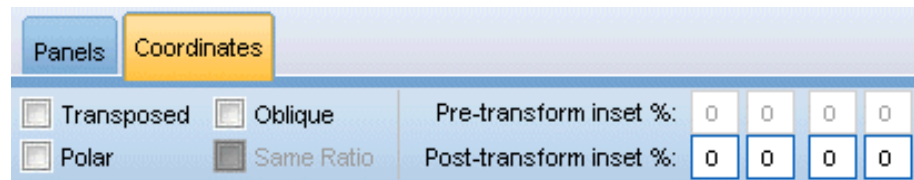
Transforming the Coordinate System

Many visualizations are displayed in a flat, rectangular coordinate system. You can transform the coordinate system as needed. For example, you can apply a polar transformation to the coordinate system, add oblique drop shadow effects, and transpose the axes. You can also undo any of these transformations if they are already applied to the current visualization. For example, a pie chart is drawn in a polar coordinate system. If desired, you can undo the polar transformation and display the pie chart as a single stacked bar in a rectangular coordinate system.

How to Transform the Coordinate System

- ▶ Select the coordinate system that you want to transform. You select the coordinate system by selecting the frame around the individual graph.
- ▶ Click the Coordinates tab on the properties palette.

Figure 5-94
Coordinates tab



- ▶ Select the transformations that you want to apply to the coordinate system. You can also deselect a transformation to undo it.

Transposed. Changing the orientation of the axes is called **transposing**. It is similar to swapping the vertical and horizontal axes in a 2-D visualization.

Polar. A polar transformation draws the graphic elements at a specific angle and distance from the center of the graph. A pie chart is a 1-D visualization with a polar transformation that draws the individual bars a specific angles. A radar chart is a 2-D visualization with a polar transformation that draws graphic elements a specific angles and distances from the center of the graph. A 3-D visualization would also include an additional depth dimension.

Oblique. An oblique transformation adds a 3-D effect to the graphic elements. This transformation adds depth to the graphic elements, but the depth is purely decorative. It is not influenced by particular data values.

Same Ratio. Applying the same ratio specifies that the same distance on each scale represents the same difference in data values. For example, 2cm on both scales represent a difference of 1000.

Pre-transform inset %. If axes are clipped after the transformation, you may want to add insets to the graph before applying the transformation. The insets shrink the dimensions by a certain percentage before any transformations are applied to the coordinate system. You have control over the lower *x*, upper *x*, lower *y*, and upper *y* dimensions, in that order.

Post-transform inset %. If you want to change the aspect ratio of the a graph, you can add insets to the graph after applying the transformation. The insets shrink the dimensions by a certain percentage after any transformations are applied to the coordinate system. These insets can also be

applied even if no transformation is applied to the graph. You have control over the lower x , upper x , lower y , and upper y dimensions, in that order.

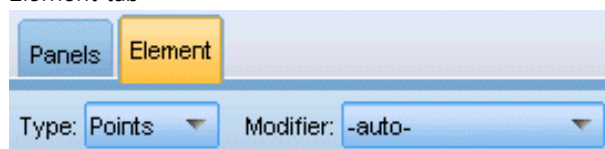
Changing Statistics and Graphic Elements

You can convert a to another type, change the statistic used to draw the graphic element, or specify the collision modifier that determines what happens when graphic elements overlap.

How to Convert a Graphic Element

- ▶ Select the graphic element that you want to convert.
- ▶ Click the Element tab on the properties palette.

Figure 5-95
Element tab



- ▶ Select a new graphic element type from the Type list.

Graphic Element Type	Description
Point	A marker identifying a specific data point. A point element is used in scatterplots and other related visualizations.
Interval	A rectangular shape drawn at a specific data value and filling the space between an origin and another data value. An interval element is used in bar charts and histograms.
Line	A line that connects data values.
Path	A line that connects data values in the order they appear in the dataset.
Area	A line that connects data elements with the area between the line and an origin filled in.
Polygon	A multi-sided shape enclosing a data region. A polygon element could be used in a binned scatterplot or a map.
Schema	An element consisting of a box with whiskers and markers indicating outliers. A schema element is used for boxplots.

How to Change the Statistic

- ▶ Select the graphic element whose statistic you want to change.
- ▶ Click the Element tab on the properties palette.
- ▶ From the Summary drop-down list, select a new statistic. Note that selecting a statistic aggregates the data. If instead you want the visualization to display unaggregated data, select (no statistic) from the Summary list.

Summary Statistics Calculated from a Continuous Field/Variable

- **Mean.** A measure of central tendency. The arithmetic average, the sum divided by the number of cases.
- **Median.** The value above and below which half of the cases fall, the 50th percentile. If there is an even number of cases, the median is the average of the two middle cases when they are sorted in ascending or descending order. The median is a measure of central tendency not sensitive to outlying values (unlike the mean, which can be affected by a few extremely high or low values).
- **Mode.** The most frequently occurring value. If several values share the greatest frequency of occurrence, each of them is a mode.
- **Minimum.** The smallest value of a numeric variable.
- **Maximum.** The largest value of a numeric variable.
- **Range.** The difference between the minimum and maximum values.
- **Mid Range.** The middle of the range, that is, the value whose difference from the minimum is equal to its difference from the maximum.
- **Sum.** The sum or total of the values, across all cases with nonmissing values.
- **Cumulative Sum.** The cumulative sum of the values. Each graphic element shows the sum for one subgroup plus the total sum of all previous groups.
- **Percent Sum.** The percentage within each subgroup based on a summed variable compared to the sum across all groups.
- **Cumulative Percent Sum.** The cumulative percentage within each subgroup based on a summed variable compared to the sum across all groups. Each graphic element shows the percentage for one subgroup plus the total percentage of all previous groups.
- **Variance.** A measure of dispersion around the mean, equal to the sum of squared deviations from the mean divided by one less than the number of cases. The variance is measured in units that are the square of those of the variable itself.
- **Standard Deviation.** A measure of dispersion around the mean. In a normal distribution, 68% of cases fall within one standard deviation of the mean and 95% of cases fall within two standard deviations. For example, if the mean age is 45, with a standard deviation of 10, 95% of the cases would be between 25 and 65 in a normal distribution.
- **Standard Error.** A measure of how much the value of a test statistic varies from sample to sample. It is the standard deviation of the sampling distribution for a statistic. For example, the standard error of the mean is the standard deviation of the sample means.
- **Kurtosis.** A measure of the extent to which observations cluster around a central point. For a normal distribution, the value of the kurtosis statistic is zero. Positive kurtosis indicates that, relative to a normal distribution, the observations are more clustered about the center of the distribution and have thinner tails until the extreme values of the distribution, at which point the tails of the leptokurtic distribution are thicker relative to a normal distribution. Negative kurtosis indicates that, relative to a normal distribution, the observations cluster less and

have thicker tails until the extreme values of the distribution, at which point the tails of the platykurtic distribution are thinner relative to a normal distribution.

- **Skewness.** A measure of the asymmetry of a distribution. The normal distribution is symmetric and has a skewness value of 0. A distribution with a significant positive skewness has a long right tail. A distribution with a significant negative skewness has a long left tail. As a guideline, a skewness value more than twice its standard error is taken to indicate a departure from symmetry.

The following region statistics may result in more than one graphic element per subgroup. When using the interval, area, or edge graphic elements, a region statistic results in one graphic element showing the range. All other graphic elements result in two separate elements, one showing the start of the range and one showing the end of the range.

- **Region: Range.** The range of values between the minimum and maximum values.
- **Region: 95% Confidence Interval of Mean.** A range of values that has a 95% chance of including the population mean.
- **Region: 95% Confidence Interval of Individual.** A range of values that has a 95% chance of including the predicted value given the individual case.
- **Region: 1 Standard Deviation above/below Mean.** A range of values between 1 **standard deviation** above and below the **mean**.
- **Region: 1 Standard Error above/below Mean.** A range of values between 1 **standard error** above and below the **mean**.

Count-Based Summary Statistics

- **Count.** The number of rows/cases.
- **Cumulative Count.** The cumulative number of rows/cases. Each graphic element shows the count for one subgroup plus the total count of all previous groups.
- **Percent of Count.** The percentage of rows/cases in each subgroup compared to the total number of rows/cases.
- **Cumulative Percent of Count.** The cumulative percentage of rows/cases in each subgroup compared to the total number of rows/cases. Each graphic element shows the percentage for one subgroup plus the total percentage of all previous groups.

How to Specify the Collision Modifier

The collision modifier determines what happens when graphic elements overlap.

- ▶ Select the graphic element for which you want to specify the collision modifier.
- ▶ Click the Element tab on the properties palette.
- ▶ From the Modifier drop-down list, select a collision modifier. -auto- lets the application determine which collision modifier is appropriate for the graphic element type and statistic.

Overlay. Draw graphic elements on top of each other when they have the same value.

Stack. Stack graphic elements that would normally be superimposed when they have the same data values.

Dodge. Move graphic elements next to other graphic elements that appear at the same value, rather than superimposing them. The graphic elements are arranged symmetrically. That is, the graphic elements are moved to opposite sides of a central position. Dodging is very similar to clustering.

Pile. Move graphic elements next to other graphic elements that appear at the same value, rather than superimposing them. The graphic elements are arranged asymmetrically. That is, the graphic elements are piled on top of one another, with the graphic element on the bottom positioned at a specific value on the scale.

Jitter (normal). Randomly reposition graphic elements at the same data value using a normal distribution.

Jitter (uniform). Randomly reposition graphic elements at the same data value using a uniform distribution.

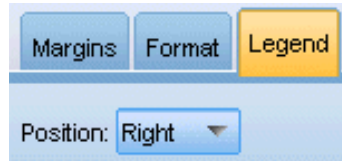
Changing the Position of the Legend

If a graph includes a legend, the legend is typically displayed to the right of a graph. You can change this position if needed.

How to Change the Legend Position

- ▶ Select the legend.
- ▶ Click the Legend tab on the properties palette.

Figure 5-96
Legend tab



- ▶ Select a position.

Copying a Visualization and Visualization Data

The General palette includes buttons for copying the visualization and its data.

Figure 5-98
Copy visualization button



Copying the visualization. This action copies the visualization to the clipboard as an image. Multiple image formats are available. When you paste the image into another application, you can choose a “paste special” option to select one of the available image formats for pasting.

Figure 5-99
Copy visualization data button



Copying the visualization data. This action copies the underlying data that is used to draw the visualization. The data is copied to the clipboard as plain text or HTML-formatted text. When you paste the data into another application, you can choose a “paste special” option to choose one of these formats for pasting.

Keyboard Shortcuts

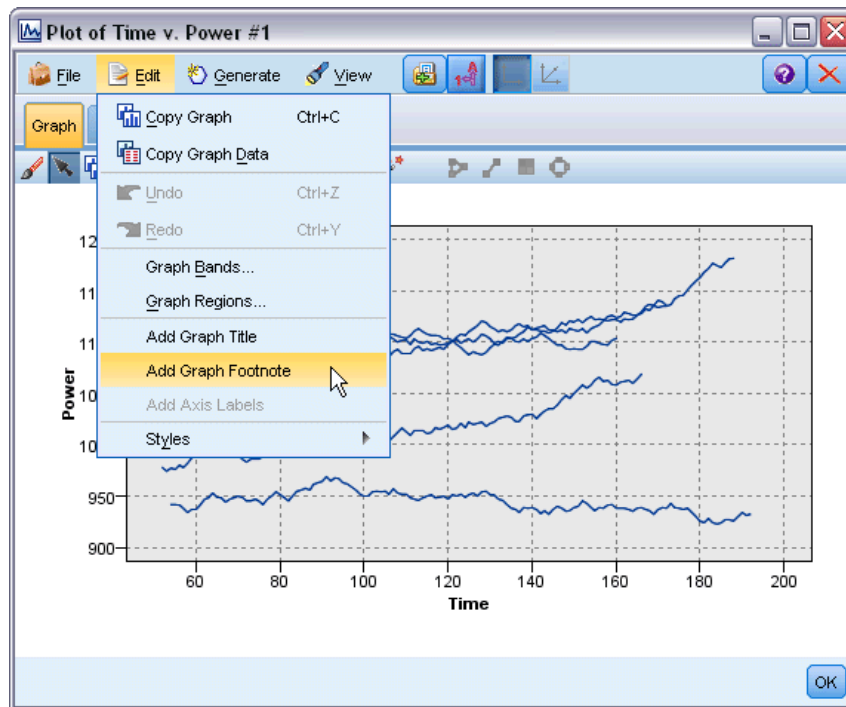
Table 5-2
Keyboard shortcuts

Shortcut Key	Function
Ctrl+Space	Toggle between Explore and Edit mode
Delete	Delete a visualization item
Ctrl+Z	Undo
Ctrl+Y	Redo
F2	Display outline for selecting items in the graph

Adding Titles and Footnotes

For all graph types you can add unique title, footnote, or axis labels to help identify what is shown in the graph.

Figure 5-100
Adding a graph footnote



Adding Titles to Graphs

- ▶ From the menus, choose Edit > Add Graph Title. A text box containing <TITLE> is displayed above the graph.
- ▶ Verify that you are in Edit mode. From the menus, choose View > Edit Mode.
- ▶ Double-click the <TITLE> text.
- ▶ Type the required title and press Return.

Adding Footnotes to Graphs

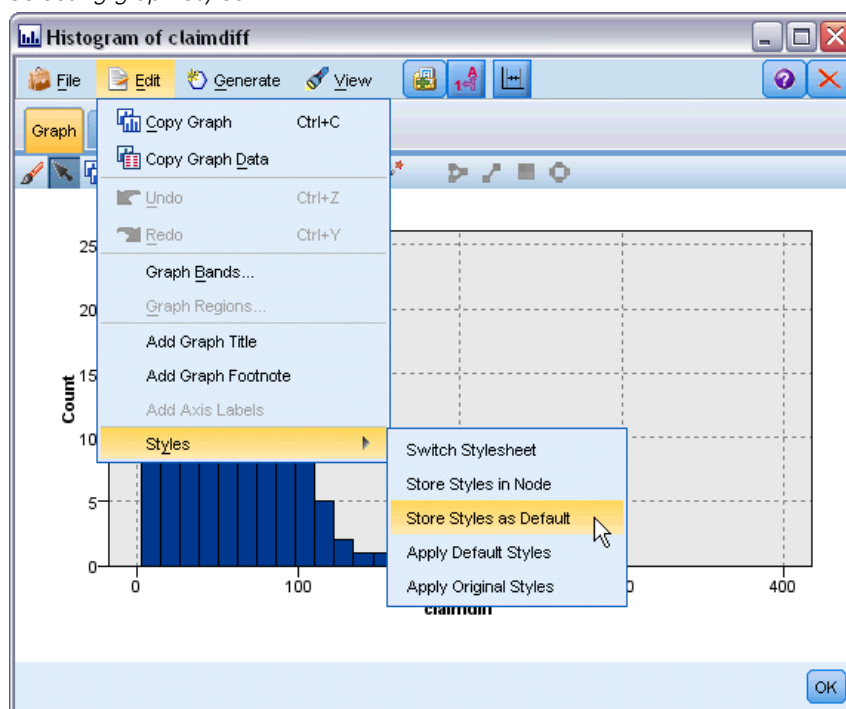
- ▶ From the menus, choose Edit > Add Graph Footnote. A text box containing <FOOTNOTE> is displayed below the graph.
- ▶ Verify that you are in Edit mode. From the menus, choose View > Edit Mode.
- ▶ Double-click the <FOOTNOTE> text.
- ▶ Type the required title and press Return.

Using Graph Stylesheets

Basic graph display information, such as colors, fonts, symbols, and line thickness, are controlled by a stylesheet. There is a default stylesheet supplied with IBM® SPSS® Modeler; however, you can make changes to it if you need. For example, you may have a corporate color scheme for presentations that you want used in your graphs. For more information, see the topic [Editing Visualizations](#) on p. 291.

In the graph nodes, you can use the Edit mode to make style changes to the look of a graph. You can then use the Edit > Styles menu to save the changes as a stylesheet to apply to all graphs that you subsequently generate from the current graph node or as a new default stylesheet for all graphs that you produce using SPSS Modeler.

Figure 5-101
Selecting graph styles



There are five stylesheet options available from the Styles option on the Edit menu:

- **Switch Stylesheet.** This displays a list of different, stored, stylesheets that you can select to change the look of your graphs. For more information, see the topic [Applying Stylesheets](#) on p. 310.
- **Store Styles in Node.** This stores modifications to the selected graph's styles so that they are applied to any future graphs created from the same graph node in the current stream.
- **Store Styles as Default.** This stores modifications to the selected graph's styles so that they are applied to all future graphs created from any graph node in any stream. After selecting this option, you can use Apply Default Styles to change any other existing graphs to use the same styles.

- **Apply Default Styles.** This changes the selected graph's styles to those that are currently saved as the default styles.
- **Apply Original Styles.** This changes a graph's styles back to the ones supplied as the original default.

Applying Stylesheets

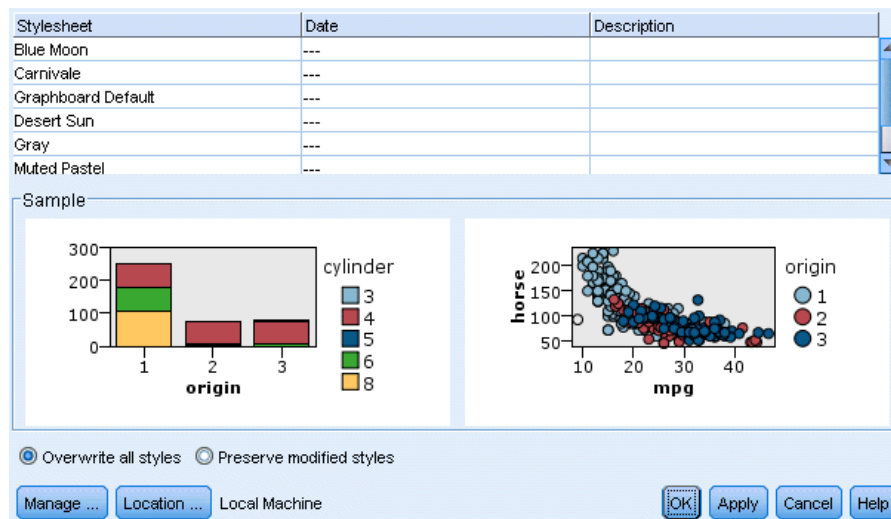
You can apply a visualization stylesheet that specifies stylistic properties of the visualization. For example, the stylesheet can define fonts, dashings, and colors, among other options. To a certain extent, stylesheets provide a shortcut for edits that you would have to perform manually. Note, however, that a stylesheet is limited to *style* changes. Other changes such as the position of the legend or the scale range are not stored in the stylesheet.

How to Apply a Stylesheet

- ▶ From the menus choose:
Edit > Styles > Switch Stylesheet
- ▶ Use the Switch Stylesheet dialog box to select a stylesheet.
- ▶ Click Apply to apply the stylesheet to the visualization without closing the dialog. Click OK to apply the stylesheet and close the dialog box.

Switch/Select Stylesheet Dialog Box

Figure 5-102
Switch Stylesheet dialog box



The table at the top of the dialog box lists all of the visualization stylesheets that are currently available. Some stylesheets are pre-installed, while others may have been created in the IBM® SPSS® Visualization Designer (a separate product).

The bottom of the dialog box shows example visualizations with sample data. Select one of the stylesheets to apply its styles to the example visualizations. These examples can help you determine how the stylesheet will affect your actual visualization.

The dialog box also offers the following options.

Existing styles. By default, a stylesheet can overwrite all the styles in the visualization. You can change this behavior.

- **Overwrite all styles.** When applying the stylesheet, overwrite all styles in the visualization, including those styles modified in the visualization during the current editing session.
- **Preserve modified styles.** When applying the stylesheet, overwrite only those styles that were *not* modified in the visualization during the current editing session. Styles that were modified during the current editing session are preserved.

Manage. Manage visualization templates and stylesheets on your computer. You can import, export, rename, and delete visualization templates and stylesheets on your local machine. For more information, see the topic [Managing Templates and Stylesheets](#) on p. 227.

Location. Change the location in which visualization templates and stylesheets are stored. The current location is listed to the right of the button. For more information, see the topic [Setting the Location of Templates and Stylesheets](#) on p. 225.

Printing, Saving, Copying, and Exporting Graphs

Each graph has a number of options that allow you to save or print the graph or export it to another format. Most of these options are available from the File menu. In addition, from the Edit menu, you can choose to copy the graph or the data within it for use in another application.

Figure 5-103
File menu and toolbar for graph windows



Printing

- ▶ To print the graph, use the Print menu item or button. Before you print, you can use Page Setup and Print Preview to set print options and preview the output.

Saving Graphs

- ▶ To save the graph to an IBM® SPSS® Modeler output file (*.cou), choose File > Save or File > Save As from the menus.

or

To save the graph in the repository, choose File > Store Output from the menus.

Copying Graphs

- ▶ To copy the graph for use in another application, such as MS Word or MS PowerPoint, choose Edit > Copy Graph from the menus.

Copying Data

- ▶ To copy the data for use in another application, such as MS Excel or MS Word, choose Edit > Copy Data from the menus. By default, the data will be formatted as HTML. Use Paste Special in the other application to view other formatting options when pasting.

Exporting Graphs

The Export Graph option enables you to export the graph in one of the following formats: Bitmap (*.bmp*), JPEG (*.jpg*), PNG (*.png*), HTML (*.html*), or ViZml document (*.xml*) for use in other IBM® SPSS® Statistics applications

- ▶ To export graphs, choose File > Export Graph from the menus and then choose the format.

Exporting Tables

The Export Table option enables you to export the table in one of the following formats: tab delimited (*.tab*), comma delimited (*.csv*), or HTML (*.html*)

- ▶ To export tables, choose File > Export Table from the menus and then choose the format.

Output Nodes

Overview of Output Nodes

Output nodes provide the means to obtain information about your data and models. They also provide a mechanism for exporting data in various formats to interface with your other software tools.

The following output nodes are available:



The Table node displays the data in table format, which can also be written to a file. This is useful anytime that you need to inspect your data values or export them in an easily readable form. For more information, see the topic [Table Node](#) on p. 321.



The Matrix node creates a table that shows relationships between fields. It is most commonly used to show the relationship between two symbolic fields, but it can also show relationships between flag fields or numeric fields. For more information, see the topic [Matrix Node](#) on p. 325.



The Analysis node evaluates predictive models' ability to generate accurate predictions. Analysis nodes perform various comparisons between predicted values and actual values for one or more model nuggets. They can also compare predictive models to each other. For more information, see the topic [Analysis Node](#) on p. 330.



The Data Audit node provides a comprehensive first look at the data, including summary statistics, histograms and distribution for each field, as well as information on outliers, missing values, and extremes. Results are displayed in an easy-to-read matrix that can be sorted and used to generate full-size graphs and data preparation nodes. For more information, see the topic [Data Audit Node](#) on p. 335.



The Transform node allows you to select and visually preview the results of transformations before applying them to selected fields. For more information, see the topic [Transform Node](#) on p. 349.



The Statistics node provides basic summary information about numeric fields. It calculates summary statistics for individual fields and correlations between fields. For more information, see the topic [Statistics Node](#) on p. 355.



The Means node compares the means between independent groups or between pairs of related fields to test whether a significant difference exists. For example, you could compare mean revenues before and after running a promotion or compare revenues from customers who did not receive the promotion with those who did. For more information, see the topic [Means Node](#) on p. 359.



The Report node creates formatted reports containing fixed text as well as data and other expressions derived from the data. You specify the format of the report using text templates to define the fixed text and data output constructions. You can provide custom text formatting by using HTML tags in the template and by setting options on the Output tab. You can include data values and other conditional output by using CLEM expressions in the template. For more information, see the topic [Report Node](#) on p. 364.



The Set Globals node scans the data and computes summary values that can be used in CLEM expressions. For example, you can use this node to compute statistics for a field called *age* and then use the overall mean of *age* in CLEM expressions by inserting the function `@GLOBAL_MEAN(age)`. For more information, see the topic [Set Globals Node](#) on p. 368.

Managing Output

The Output manager shows the charts, graphs, and tables generated during an IBM® SPSS® Modeler session. You can always reopen an output by double-clicking it in the manager—you do not have to rerun the corresponding stream or node.

To View the Output Manager

- ▶ Open the View menu and choose Managers. Click the Outputs tab.

Figure 6-1
Output manager



From the Output manager, you can:

- Display existing output objects, such as histograms, evaluation charts, and tables.
- Rename output objects.
- Save output objects to disk or to the IBM® SPSS® Collaboration and Deployment Services Repository (if available).
- Add output files to the current project.
- Delete unsaved output objects from the current session.
- Open saved output objects or retrieve them from the IBM SPSS Collaboration and Deployment Services Repository (if available).

To access these options, right-click anywhere on the Outputs tab.

Viewing Output

On-screen output is displayed in an output browser window. The output browser window has its own set of menus that allow you to print or save the output, or export it to another format. Note that specific options may vary depending on the type of output.

Printing, saving, and exporting data. More information is available as follows:

- To print the output, use the Print menu option or button. Before you print, you can use Page Setup and Print Preview to set print options and preview the output.
- To save the output to an IBM® SPSS® Modeler output file (.cou), choose Save or Save As from the File menu.
- To save the output in another format, such as text or HTML, choose Export from the File menu. For more information, see the topic [Exporting Output](#) on p. 319.
- To save the output in a shared repository so that other users can view it using the IBM® SPSS® Collaboration and Deployment Services Deployment Portal, choose Publish to Web from the File menu. Note that this option requires a separate license for IBM® SPSS® Collaboration and Deployment Services.

Selecting cells and columns. The Edit menu contains various options for selecting, deselecting, and copying cells and columns, as appropriate for the current output type. For more information, see the topic [Selecting Cells and Columns](#) on p. 320.

Generating new nodes. The Generate menu allows you to generate new nodes based on the contents of the output browser. The options vary depending on the type of output and the items in the output that are currently selected. For details about the node-generation options for a particular type of output, see the documentation for that output.

Publish to Web

The Publish to Web feature enables you to publish certain types of stream output to a central shared IBM® SPSS® Collaboration and Deployment Services Repository that forms the basis of IBM® SPSS® Collaboration and Deployment Services. If you use this option, other users who need to view this output can do so by using Internet access and an IBM SPSS Collaboration and Deployment Services account—they do not need to have IBM® SPSS® Modeler installed.

Note: A separate license is required to access an IBM SPSS Collaboration and Deployment Services repository. For more information, see <http://www.ibm.com/software/analytics/spss/products/deployment/cds/>

The following table lists the SPSS Modeler nodes that support the Publish to Web feature. Output from these nodes is stored in the IBM SPSS Collaboration and Deployment Services Repository in output object (.cou) format, and can be viewed directly in the IBM® SPSS® Collaboration and Deployment Services Deployment Portal.

Other types of output can be viewed only if the relevant application (e.g. SPSS Modeler, for stream objects) is installed on the user's machine.

Table 6-1
Nodes supporting Publish to Web

Node Type	Node
Graphs	all
Output	Table
	Matrix
	Data Audit
	Transform
	Means
	Analysis
	Statistics
	Report (HTML)
IBM® SPSS® Statistics	Statistics Output

Publishing Output to the Web

To publish output to the Web:

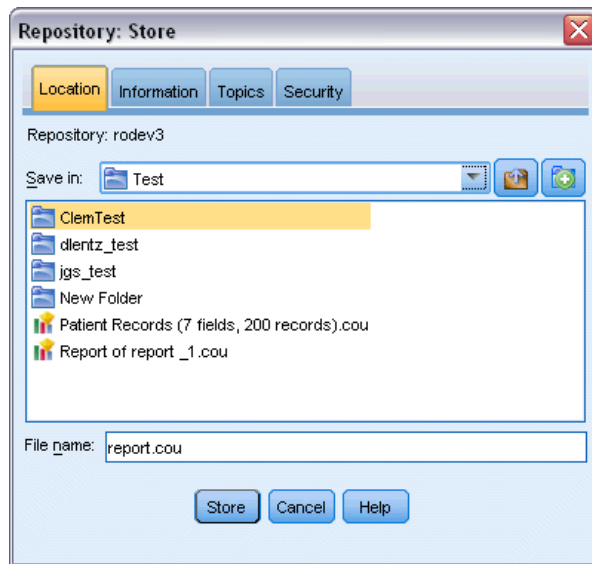
- ▶ In an IBM® SPSS® Modeler stream, execute one of the nodes listed in the table. Doing so creates an output object (for example, a table, matrix or report object) in a new window.
- ▶ From the output object window, choose:
File > Publish to Web

Note: If you just want to export simple HTML files for use with a standard Web browser, choose Export from the File menu and select HTML.

- ▶ Connect to the IBM® SPSS® Collaboration and Deployment Services Repository.

When you have connected successfully, the Repository: Store dialog is displayed, offering a number of storage options.

Figure 6-2
Repository:Store dialog



- ▶ When you have chosen the storage options you want, click Store.

Viewing Published Output Over the Web

You must have an IBM SPSS Collaboration and Deployment Services account set up in order to use this feature. If you have the relevant application installed for the object type you want to view (for example, IBM® SPSS® Modeler or IBM® SPSS® Statistics), the output is displayed in the application itself rather than in the browser.

Note: A separate license is required to access IBM® SPSS® Collaboration and Deployment Services. For more information, see <http://www.ibm.com/software/analytics/spss/products/deployment/cds/>.

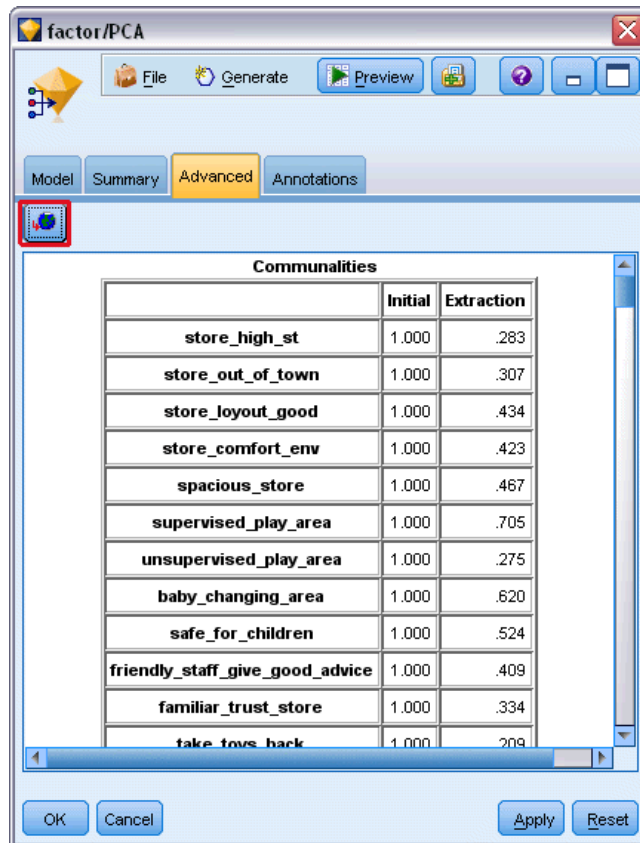
To view published output over the Web:

- ▶ Point your browser to `http://<repos_host>:<repos_port>/peb` where `repos_host` and `repos_port` are the hostname and port number for the IBM SPSS Collaboration and Deployment Services host.
- ▶ Enter the login details for your IBM SPSS Collaboration and Deployment Services account.
- ▶ Click on Content Repository.
- ▶ Navigate to, or search for, the object you want to view.
- ▶ Click on the object name. For some object types, such as graphs, there may be a delay while the object is rendered in the browser.

Viewing Output in an HTML Browser

From the Advanced tab on the Linear, Logistic, and PCA/Factor model nuggets, you can view the displayed information in a separate browser, such as Internet Explorer. The information is output as HTML, enabling you to save it and reuse it elsewhere, such as on a corporate intranet, or Internet site.

Figure 6-3
Launch button on Advanced tab of model nugget



To display the information in a browser, click the launch button, situated below the model icon in the top left of the Advanced tab of the model nugget.

Exporting Output

In the output browser window, you may choose to export the output to another format, such as text or HTML. The export formats vary depending on the type of output, but in general are similar to the file type options available if you select Save to file in the node used to generate the output.

To Export Output

- ▶ In the output browser, open the File menu and choose Export. Then choose the file type that you want to create:
 - **Tab Delimited (*.tab).** This option generates a formatted text file containing the data values. This style is often useful for generating a plain-text representation of the information that can be imported into other applications. This option is available for the Table, Matrix, and Means nodes.
 - **Comma Delimited (*.dat).** This option generates a comma-delimited text file containing the data values. This style is often useful as a quick way to generate a data file that can be imported into spreadsheets or other data analysis applications. This option is available for the Table, Matrix, and Means nodes.
 - **Transposed Tab Delimited (*.tab).** This option is identical to the Tab Delimited option, but the data is transposed so that rows represent fields and the columns represent records.
 - **Transposed Comma Delimited (*.dat).** This option is identical to the Comma Delimited option, but the data is transposed so that rows represent fields and the columns represent records.
 - **HTML (*.html).** This option writes HTML-formatted output to a file or files.

Selecting Cells and Columns

Figure 6-4
Table browser window

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimt
1	id602	name602	north	1780	42	9	734118.000	maize	arable
2	id606	name606	southeast	1580	42	7	445785.000	maize	arable
3	id607	name607	southeast	1820	29	6	211605.000	maize	arable
4	id608	name608	southeast	1640	108	7	1167040.0...	maize	arable
5	id610	name610	southeast	600	80	6	267928.000	wheat	arable
6	id611	name611	southeast	980	38	6	222703.000	maize	arable
7	id613	name613	southeast	440	86	3	115544.000	potatoes	arable
8	id614	name614	southeast	1260	90	8	900243.000	maize	arable
9	id616	name616	midlands	1660	36	9	490617.000	rapeseed	arable
10	id620	name620	north	880	74	6	426988.000	rapeseed	arable
11	id621	name621	southwest	1160	105	4	299274.000	maize	arable
12	id622	name622	southeast	1500	61	7	687736.000	wheat	arable
13	id623	name623	southeast	1260	17	8	170279.000	maize	arable
14	id626	name626	midlands	1580	109	8	1286430.0...	wheat	arable
15	id627	name627	southeast	500	93	3	102720.000	rapeseed	arable
16	id628	name628	southeast	880	15	5	70439.800	wheat	arable
17	id630	name630	midlands	680	81	4	221391.000	potatoes	arable
18	id636	name636	southeast	1160	21	8	185939.000	potatoes	arable
19	id637	name637	midlands	940	106	6	622450.000	maize	arable
20	id638	name638	midlands	1480	64	6	586185.000	wheat	arable

A number of nodes, including the Table node, Matrix node, and Means node, generate tabular output. These output tables can be viewed and manipulated in similar ways, including selecting cells, copying all or part of the table to the Clipboard, generating new nodes based on the current selection, and saving and printing the table.

Selecting cells. To select a cell, click it. To select a rectangular range of cells, click one corner of the desired range, drag the mouse to the other corner of the range, and release the mouse button. To select an entire column, click the column heading. To select multiple columns, use Shift-click or Ctrl-click on column headings.

When you make a new selection, the old selection is cleared. By holding down the Ctrl key while selecting, you can add the new selection to any existing selection instead of clearing the old selection. You can use this method to select multiple, noncontiguous regions of the table. The Edit menu also contains the Select All and Clear Selection options.

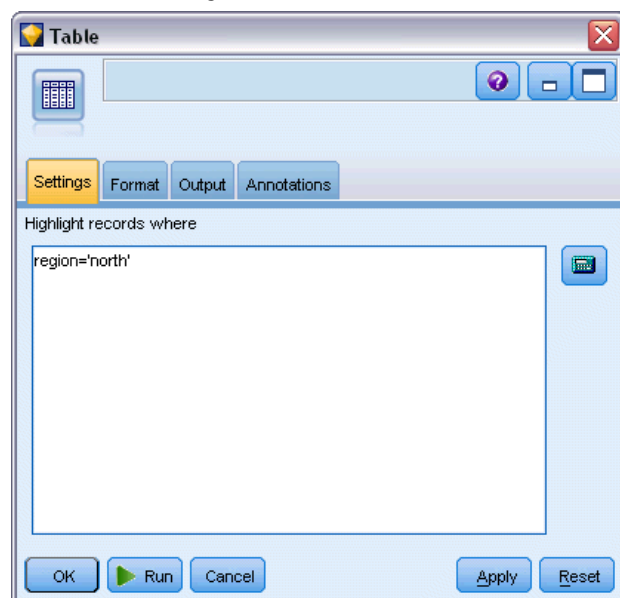
Reordering columns. The Table node and Means node output browsers allow you to move columns in the table by clicking a column heading and dragging it to the desired location. You can move only one column at a time.

Table Node

The Table node creates a table that lists the values in your data. All fields and all values in the stream are included, making this an easy way to inspect your data values or export them in an easily readable form. Optionally, you can highlight records that meet a certain condition.

Table Node Settings Tab

Figure 6-5
Table node: Settings tab



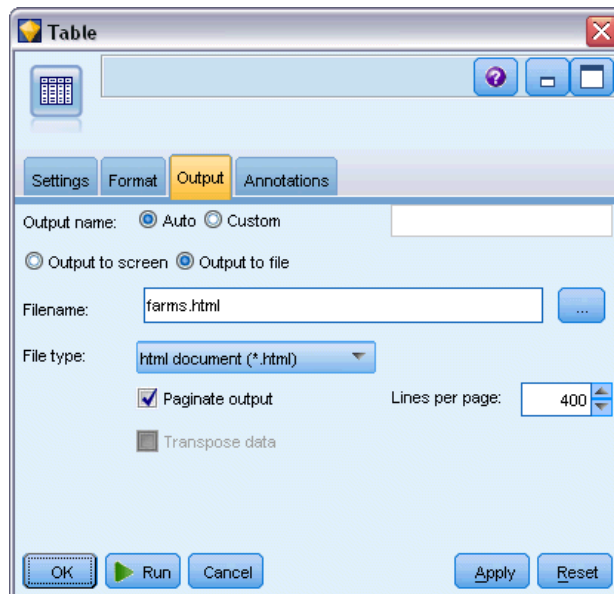
Highlight records where. You can highlight records in the table by entering a CLEM expression that is true for the records to be highlighted. This option is enabled only when Output to screen is selected.

Table Node Format Tab

The Format tab contains options used to specify formatting on a per-field basis. This tab is shared with the Type node. For more information, see the topic [Field Format Settings Tab](#) on p. 128.

Output Node Output Tab

Figure 6-6
Output node Output tab



For nodes that generate table-style output, the Output tab lets you specify the format and location of the results.

Output name. Specifies the name of the output produced when the node is executed. Auto chooses a name based on the node that generates the output. Optionally, you can select Custom to specify a different name.

Output to screen (the default). Creates an output object to view online. The output object will appear on the Outputs tab of the manager window when the output node is executed.

Output to file. Saves the output to a file when the node is executed. If you choose this option, enter a filename (or navigate to a directory and specify a filename using the File Chooser button) and select a file type. Note that some file types may be unavailable for certain types of output.

Data is output in the system default encoding format, which is specified in the Windows Control Panel or, if running in distributed mode, on the server computer.

- **Data (tab delimited) (*.tab).** This option generates a formatted text file containing the data values. This style is often useful for generating a plain-text representation of the information that can be imported into other applications. This option is available for the Table, Matrix, and Means nodes.
- **Data (comma delimited) (*.dat).** This option generates a comma-delimited text file containing the data values. This style is often useful as a quick way to generate a data file that can be imported into spreadsheets or other data analysis applications. This option is available for the Table, Matrix, and Means nodes.
- **HTML (*.html).** This option writes HTML-formatted output to a file or files. For tabular output (from the Table, Matrix, or Means nodes), a set of HTML files contains a contents panel listing field names and the data in an HTML table. The table may be split over multiple HTML files if the number of rows in the table exceeds the Lines per page specification. In this case, the contents panel contains links to all table pages and provides a means of navigating the table. For non-tabular output, a single HTML file is created containing the results of the node.

Note: If the HTML output contains only formatting for the first page, select Paginate output and adjust the Lines per page specification to include all output on a single page. Or if the output template for nodes such as the Report node contains custom HTML tags, be sure you have specified Custom as the format type.
- **Text File (*.txt).** This option generates a text file containing the output. This style is often useful for generating output that can be imported into other applications, such as word processors or presentation software. This option is not available for some nodes.
- **Output object (*.cou).** Output objects saved in this format can be opened and viewed in IBM® SPSS® Modeler, added to projects, and published and tracked using the IBM® SPSS® Collaboration and Deployment Services Repository.

Output view. For the Means node, you can specify whether simple or advanced output is displayed by default. Note you can also toggle between these views when browsing the generated output. For more information, see the topic [Means Node Output Browser](#) on p. 362.

Format. For the Report node, you can choose whether output is automatically formatted or formatted using HTML included in the template. Select Custom to allow HTML formatting in the template.

Title. For the Report node, you can specify optional title text that will appear at the top of the report output.

Highlight inserted text. For the Report node, select this option to highlight text generated by CLEM expressions in the Report template. For more information, see the topic [Report Node Template Tab](#) on p. 365. This option is not recommended when using Custom formatting.

Lines per page. For the Report node, specify a number of lines to include on each page during Auto formatting of the output report.

Transpose data. This option transposes the data before export, so that rows represent fields and the columns represent records.

Note: For large tables, the above options can be somewhat inefficient, especially when working with a remote server. In such cases, using a File output node provides much better performance. For more information, see the topic [Flat File Export Node](#) in Chapter 7 on p. 382.

Table Browser

Figure 6-7
Table browser window

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimt
1	id602	name602	north	1780	42	9	734118.000	maize	arable
2	id606	name606	southeast	1580	42	7	445785.000	maize	arable
3	id607	name607	southeast	1820	29	6	211605.000	maize	arable
4	id608	name608	southeast	1640	108	7	1167040.0...	maize	arable
5	id610	name610	southeast	600	80	6	267928.000	wheat	arable
6	id611	name611	southeast	980	38	6	222703.000	maize	arable
7	id613	name613	southeast	440	86	3	115544.000	potatoes	arable
8	id614	name614	southeast	1260	90	8	900243.000	maize	arable
9	id616	name616	midlands	1660	36	9	490617.000	rapeseed	arable
10	id620	name620	north	880	74	6	426988.000	rapeseed	arable
11	id621	name621	southwest	1160	105	4	299274.000	maize	arable
12	id622	name622	southeast	1500	61	7	687736.000	wheat	arable
13	id623	name623	southeast	1260	17	8	170279.000	maize	arable
14	id626	name626	midlands	1580	109	8	1286430.0...	wheat	arable
15	id627	name627	southeast	500	93	3	102720.000	rapeseed	arable
16	id628	name628	southeast	880	15	5	70439.800	wheat	arable
17	id630	name630	midlands	680	81	4	221391.000	potatoes	arable
18	id636	name636	southeast	1160	21	8	185939.000	potatoes	arable
19	id637	name637	midlands	940	106	6	622450.000	maize	arable
20	id638	name638	midlands	1480	64	6	586185.000	wheat	arable

The table browser displays tabular data and allows you to perform standard operations including selecting and copying cells, reordering columns, and saving and printing the table. For more information, see the topic [Selecting Cells and Columns](#) on p. 320. These are the same operations that you can carry out when previewing the data in a node.

Exporting table data. You can export data from the table browser by choosing:
File > Export

For more information, see the topic [Exporting Output](#) on p. 319.

Data is exported in the system default encoding format, which is specified in the Windows Control Panel or, if running in distributed mode, on the server computer.

Searching the table. The search button (with the binoculars icon) on the main toolbar activates the search toolbar, allowing you to search the table for specific values. You can search forward or backward in the table, you can specify a case-sensitive search (the Aa button), and you can interrupt a search-in-progress with the interrupt search button.

Figure 6-8
Table with search controls activated

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimt
29	id669	name669	southwest	1840	80	7	1072440.0...	wheat	arable
30	id671	name671	southeast	1020	51	5	245851.000	wheat	arable
31	id672	name672	southeast	1000	65	4	234890.000	maize	arable
32	id673	name673	midlands	900	66	6	380620.000	maize	arable
33	id675	name675	north	700	92	6	401818.000	maize	arable
34	id676	name676	southeast	740	46	7	248335.000	wheat	arable
35	id677	name677	midlands	1460	63	3	211222.000	rapeseed	arable
36	id679	name679	midlands	1380	21	8	170604.000	wheat	arable
37	id682	name682	midlands	1140	100	5	592811.000	potatoes	arable
38	id685	name685	southwest	600	48	4	108645.000	maize	arable
39	id688	name688	southwest	1480	75	3	335648.000	wheat	arable
40	id689	name689	southeast	1160	108	3	374262.000	maize	arable
41	id691	name691	southwest	920	109	9	925974.000	wheat	arable
42	id693	name693	southeast	500	76	5	181057.000	wheat	arable
43	id696	name696	southeast	1300	23	9	274389.000	maize	arable
44	id699	name699	southeast	1520	49	3	217542.000	maize	arable
45	id704	name704	southeast	1840	103	8	1588890.0...	rapeseed	arable
46	id705	name705	midlands	1800	38	7	472370.000	wheat	arable

Generating new nodes. The Generate menu contains node generation operations.

- **Select Node ("Records").** Generates a Select node that selects the records for which any cell in the table is selected.
- **Select ("And").** Generates a Select node that selects records containing *all* of the values selected in the table.
- **Select ("Or").** Generates a Select node that selects records containing *any* of the values selected in the table.
- **Derive ("Records").** Generates a Derive node to create a new flag field. The flag field contains *T* for records for which any cell in the table is selected and *F* for the remaining records.
- **Derive ("And").** Generates a Derive node to create a new flag field. The flag field contains *T* for records containing *all* of the values selected in the table and *F* for the remaining records.
- **Derive ("Or").** Generates a Derive node to create a new flag field. The flag field contains *T* for records containing *any* of the values selected in the table and *F* for the remaining records.

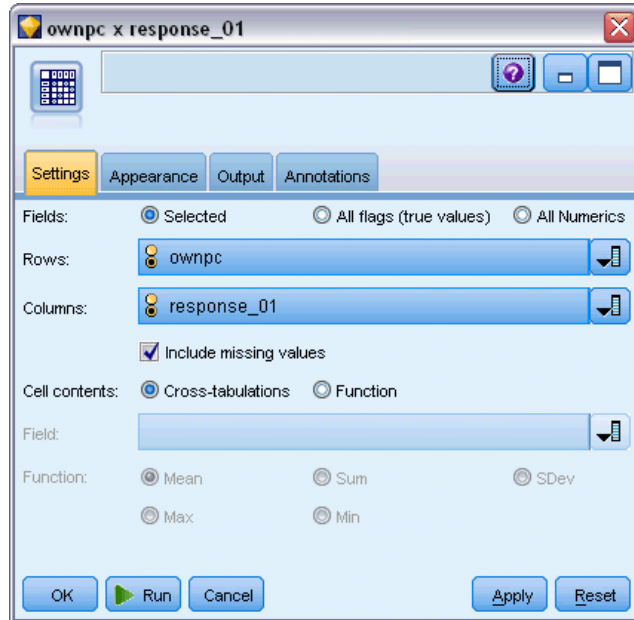
Matrix Node

The Matrix node allows you to create a table that shows relationships between fields. It is most commonly used to show the relationship between two categorical fields (flag, nominal, or ordinal), but it can also be used to show relationships between continuous (numeric range) fields.

Matrix Node Settings Tab

The Settings tab lets you specify options for the structure of the matrix.

Figure 6-9
Matrix node: Settings tab



Fields. Select a field selection type from the following options:

- **Selected.** This option allows you to select a categorical field for the rows and one for the columns of the matrix. The rows and columns of the matrix are defined by the list of values for the selected categorical field. The cells of the matrix contain the summary statistics selected below.
- **All flags (true values).** This option requests a matrix with one row and one column for each flag field in the data. The cells of the matrix contain the counts of double positives for each flag combination. In other words, for a row corresponding to *bought bread* and a column corresponding to *bought cheese*, the cell at the intersection of that row and column contains the number of records for which both *bought bread* and *bought cheese* are true.
- **All numerics.** This option requests a matrix with one row and one column for each numeric field. The cells of the matrix represent the sum of the cross-products for the corresponding pair of fields. In other words, for each cell in the matrix, the values for the row field and the column field are multiplied for each record and then summed across records.

Include missing values. Includes user-missing (blank) and system missing (\$null\$) values in the row and column output. For example, if the value *N/A* has been defined as user-missing for the selected column field, a separate column labeled *N/A* will be included in the table (assuming this value actually occurs in the data) just like any other category. If this option is deselected, the *N/A* column is excluded regardless of how often it occurs.

Note: The option to include missing values applies only when selected fields are cross-tabulated. Blank values are mapped to \$null\$ and are excluded from aggregation for the function field when the mode is Selected and the content is set to Function and for all numeric fields when the mode is set to All Numerics.

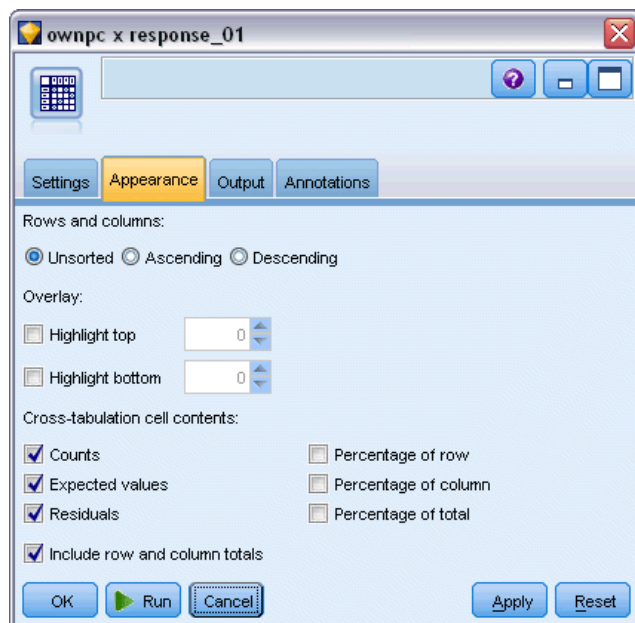
Cell contents. If you have chosen Selected fields above, you can specify the statistic to be used in the cells of the matrix. Select a count-based statistic, or select an overlay field to summarize values of a numeric field based on the values of the row and column fields.

- **Cross-tabulations.** Cell values are counts and/or percentages of how many records have the corresponding combination of values. You can specify which cross-tabulation summaries you want using the options on the Appearance tab. The global chi-square value is also displayed along with the significance. For more information, see the topic [Matrix Node Output Browser](#) on p. 328.
- **Function.** If you select a summary function, cell values are a function of the selected overlay field values for cases having the appropriate row and column values. For example, if the row field is *Region*, the column field is *Product*, and the overlay field is *Revenue*, then the cell in the *Northeast* row and the *Widget* column will contain the sum (or average, minimum, or maximum) of revenue for widgets sold in the northeast region. The default summary function is Mean. You can select another function for summarizing the function field. Options include Mean, Sum, SDev (standard deviation), Max (maximum), and Min (minimum).

Matrix Node Appearance Tab

The Appearance tab allows you to control sorting and highlighting options for the matrix, as well as statistics presented for cross-tabulation matrices.

Figure 6-10
Matrix node: Appearance tab



Rows and columns. Controls the sorting of row and column headings in the matrix. The default is Unsorted. Select Ascending or Descending to sort row and column headings in the specified direction.

Overlay. Allows you to highlight extreme values in the matrix. Values are highlighted based on cell counts (for cross-tabulation matrices) or calculated values (for function matrices).

- **Highlight top.** You can request the highest values in the matrix to be highlighted (in red). Specify the number of values to highlight.
- **Highlight bottom.** You can also request the lowest values in the matrix to be highlighted (in green). Specify the number of values to highlight.

Note: For the two highlighting options, ties can cause more values than requested to be highlighted. For example, if you have a matrix with six zeros among the cells and you request Highlight bottom 5, all six zeros will be highlighted.

Cross-tabulation cell contents. For cross-tabulations, you can specify the summary statistics contained in the matrix for cross-tabulation matrices. These options are not available when either the All Numerics or Function option is selected on the Settings tab.

- **Counts.** Cells include the number of records with the row value that have the corresponding column value. This is only default cell content.
- **Expected values.** The expected value for number of records in the cell, assuming that there is no relationship between the rows and columns. Expected values are based on the following formula:

$$p(\text{row value}) * p(\text{column value}) * \text{total number of records}$$

- **Residuals.** The difference between observed and expected values.
- **Percentage of row.** The percentage of all records with the row value that have the corresponding column value. Percentages sum to 100 within rows.
- **Percentage of column.** The percentage of all records with the column value that have the corresponding row value. Percentages sum to 100 within columns.
- **Percentage of total.** The percentage of all records having the combination of column value and row value. Percentages sum to 100 over the whole matrix.
- **Include row and column totals.** Adds a row and a column to the matrix for column and row totals.
- **Apply Settings.** (Output Browser only) Enables you to make changes to the appearance of the Matrix node output without having to close and reopen the Output Browser. Make the changes on this tab of the Output Browser, click this button and then select the Matrix tab to see the effect of the changes.

Matrix Node Output Browser

The matrix browser displays cross-tabulated data and allows you to perform operations on the matrix, including selecting cells, copying the matrix to the Clipboard in whole or in part, generating new nodes based on the matrix selection, and saving and printing the matrix. The matrix browser may also be used to display output from certain models, such as Naive Bayes models from Oracle.

Figure 6-11
Matrix browser

Matrix of ownpc by response_01

response_01

ownpc		0	1	Total
0	Count	1611	225	1836
	Expected	1682.510	153.490	1836
	Residual	-71.510	71.510	0
1	Count	2971	193	3164
	Expected	2899.490	264.510	3164
	Residual	71.510	-71.510	0
Total	Count	4582	418	5000
	Expected	4582	418	5000
	Residual	0	0	0

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 57.452, df = 1, probability = 0

The File and Edit menus provide the usual options for printing, saving, and exporting output, and for selecting and copying data. For more information, see the topic [Viewing Output](#) on p. 316.

Chi-square. For a cross-tabulation of two categorical fields, the global Pearson chi-square is also displayed below the table. This test indicates the probability that the two fields are unrelated, based on the difference between observed counts and the counts you would expect if no relationship exists. For example, if there were no relationship between customer satisfaction and store location, you would expect similar satisfaction rates across all stores. But if customers at certain stores consistently report higher rates than others, you might suspect it wasn't a coincidence. The greater the difference, the smaller the probability that it was the result of chance sampling error alone.

- The chi-square test indicates the probability that the two fields are unrelated, in which case any differences between observed and expected frequencies are the result of chance alone. If this probability is very small—typically less than 5%—then the relationship between the two fields is said to be significant.
- If there is only one column or one row (a one-way chi-square test), the degrees of freedom is the number of cells minus one. For a two-way chi-square, the degrees of freedom is the number of rows minus one times the number of columns minus one.
- Use caution when interpreting the chi-square statistic if any of the expected cell frequencies are less than five.
- The chi-square test is available only for a cross-tabulation of two fields. (When All flags or All numerics is selected on the Settings tab, this test is not displayed.)

Generate menu. The Generate menu contains node generation operations. These operations are available only for cross-tabulated matrices, and you must have at least one cell selected in the matrix.

- **Select Node.** Generates a Select node that selects the records that match any selected cell in the matrix.
- **Derive Node (Flag).** Generates a Derive node to create a new flag field. The flag field contains *T* for records that match any selected cell in the matrix and *F* for the remaining records.
- **Derive Node (Set).** Generates a Derive node to create a new nominal field. The nominal field contains one category for each contiguous set of selected cells in the matrix.

Analysis Node

The Analysis node allows you to evaluate the ability of a model to generate accurate predictions. Analysis nodes perform various comparisons between predicted values and actual values (your target field) for one or more model nuggets. Analysis nodes can also be used to compare predictive models to other predictive models.

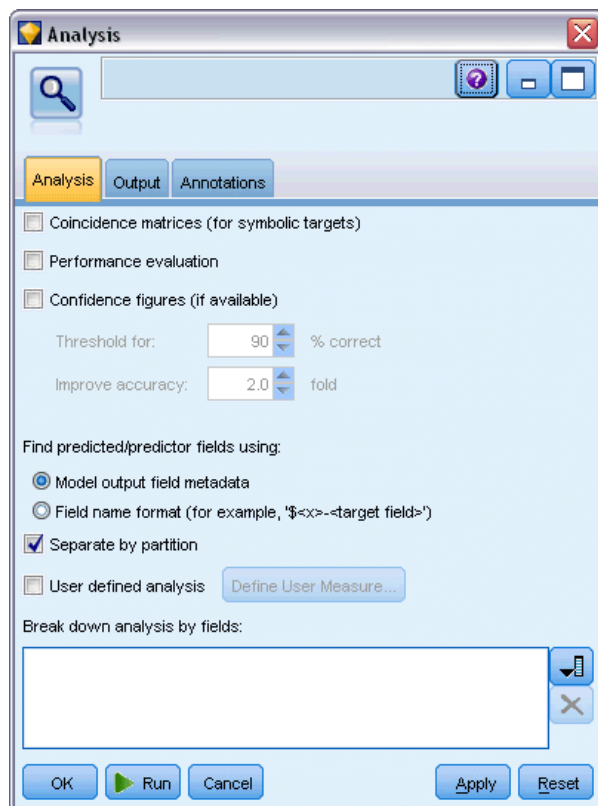
When you execute an Analysis node, a summary of the analysis results is automatically added to the Analysis section on the Summary tab for each model nugget in the executed stream. The detailed analysis results appear on the Outputs tab of the manager window or can be written directly to a file.

Note: Because Analysis nodes compare predicted values to actual values, they are only useful with supervised models (those that require a target field). For unsupervised models such as clustering algorithms, there are no actual results available to use as a basis for comparison.

Analysis Node Analysis Tab

The Analysis tab allows you to specify the details of the analysis.

Figure 6-12
Analysis node: Analysis tab



Coincidence matrices (for symbolic or categorical targets). Shows the pattern of matches between each generated (predicted) field and its target field for categorical targets (either flag, nominal, or ordinal). A table is displayed with rows defined by actual values and columns defined by predicted values, with the number of records having that pattern in each cell. This is useful for identifying systematic errors in prediction. If there is more than one generated field related to the same output field but produced by different models, the cases where these fields agree and disagree are counted and the totals are displayed. For the cases where they agree, another set of correct/wrong statistics is displayed.

Performance evaluation. Shows performance evaluation statistics for models with categorical outputs. This statistic, reported for each category of the output field(s), is a measure of the average information content (in bits) of the model for predicting records belonging to that category. It takes the difficulty of the classification problem into account, so accurate predictions for rare categories will earn a higher performance evaluation index than accurate predictions for common categories. If the model does no better than guessing for a category, the performance evaluation index for that category will be 0.

Confidence figures (if available). For models that generate a confidence field, this option reports statistics on the confidence values and their relationship to predictions. There are two settings for this option:

- **Threshold for.** Reports the confidence level above which the accuracy will be the specified percentage.
- **Improve accuracy.** Reports the confidence level above which the accuracy is improved by the specified factor. For example, if the overall accuracy is 90% and this option is set to 2.0, the reported value will be the confidence required for 95% accuracy.

Find predicted/predictor fields using. Determines how predicted fields are matched to the original target field.

- **Model output field metadata.** Matches predicted fields to the target based on model field information, allowing a match even if a predicted field has been renamed. Model field information can also be accessed for any predicted field from the Values dialog box using a Type node. For more information, see the topic [Using the Values Dialog Box](#) in Chapter 4 on p. 120.
- **Field name format.** Matches fields based on the naming convention. For example predicted values generated by a C5.0 model nugget for a target named *response* must be in a field named *\$C-response*.

Separate by partition. If a partition field is used to split records into training, test, and validation samples, select this option to display results separately for each partition. For more information, see the topic [Partition Node](#) in Chapter 4 on p. 176.

Note: When separating by partition, records with null values in the partition field are excluded from the analysis. This will never be an issue if a Partition node is used, since Partition nodes do not generate null values.

User defined analysis. You can specify your own analysis calculation to be used in evaluating your model(s). Use CLEM expressions to specify what should be computed for each record and how to combine the record-level scores into an overall score. Use the functions @TARGET and @PREDICTED to refer to the target (actual output) value and the predicted value, respectively.

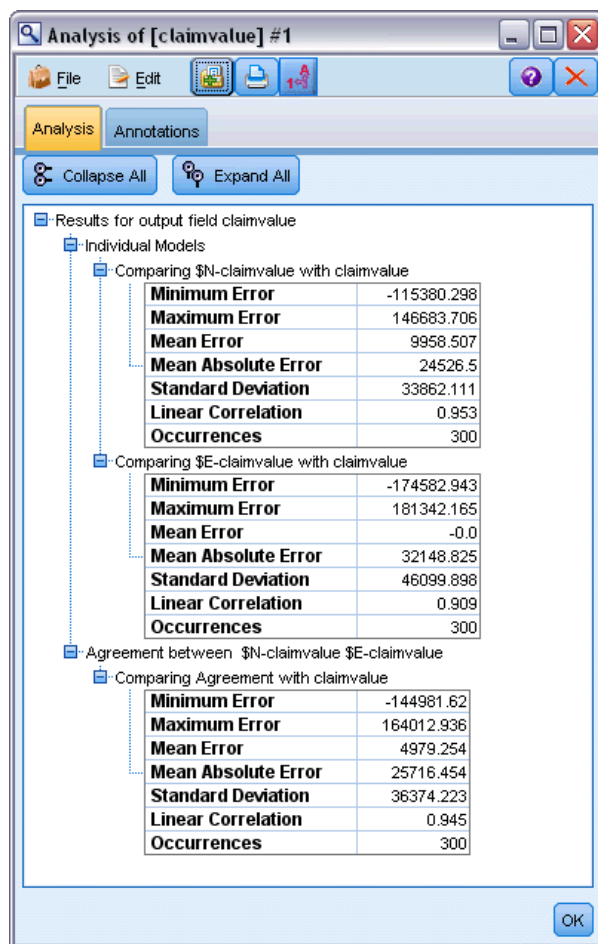
- **If.** Specify a conditional expression if you need to use different calculations depending on some condition.
- **Then.** Specify the calculation if the If condition is true.
- **Else.** Specify the calculation if the If condition is false.
- **Use.** Select a statistic to compute an overall score from the individual scores.

Break down analysis by fields. Shows the categorical fields available for breaking down the analysis. In addition to the overall analysis, a separate analysis will be reported for each category of each breakdown field.

Analysis Output Browser

The analysis output browser lets you see the results of executing the Analysis node. The usual saving, exporting, and printing options are available from the File menu. For more information, see the topic [Viewing Output](#) on p. 316.

Figure 6-13
Analysis output browser



When you first browse Analysis output, the results are expanded. To hide results after viewing them, use the expander control to the left of the item to collapse the specific results you want to hide or click the Collapse All button to collapse all results. To see results again after collapsing them, use the expander control to the left of the item to show the results or click the Expand All button to show all results.

Results for output field. The Analysis output contains a section for each output field for which there is a corresponding prediction field created by a generated model.

Comparing. Within the output field section is a subsection for each prediction field associated with that output field. For categorical output fields, the top level of this section contains a table showing the number and percentage of correct and incorrect predictions and the total number of records in the stream. For numeric output fields, this section shows the following information:

- **Minimum Error.** Shows the minimum error (difference between observed and predicted values).
- **Maximum Error.** Shows the maximum error.

- **Mean Error.** Shows the average (mean) of errors across all records. This indicates whether there is a systematic **bias** (a stronger tendency to overestimate than to underestimate, or vice versa) in the model.
- **Mean Absolute Error.** Shows the average of the absolute values of the errors across all records. Indicates the average magnitude of error, independent of the direction.
- **Standard Deviation.** Shows the standard deviation of the errors.
- **Linear Correlation.** Shows the linear correlation between the predicted and actual values. This statistic varies between -1.0 and 1.0 . Values close to $+1.0$ indicate a strong positive association, so that high predicted values are associated with high actual values and low predicted values are associated with low actual values. Values close to -1.0 indicate a strong negative association, so that high predicted values are associated with low actual values, and vice versa. Values close to 0.0 indicate a weak association, so that predicted values are more or less independent of actual values. *Note:* A blank entry here indicates that linear correlation cannot be computed in this case, because either the actual or predicted values are constant.
- **Occurrences.** Shows the number of records used in the analysis.

Coincidence Matrix. For categorical output fields, if you requested a coincidence matrix in the analysis options, a subsection appears here containing the matrix. The rows represent actual observed values, and the columns represent predicted values. The cell in the table indicates the number of records for each combination of predicted and actual values.

Performance Evaluation. For categorical output fields, if you requested performance evaluation statistics in the analysis options, the performance evaluation results appear here. Each output category is listed with its performance evaluation statistic.

Confidence Values Report. For categorical output fields, if you requested confidence values in the analysis options, the values appear here. The following statistics are reported for model confidence values:

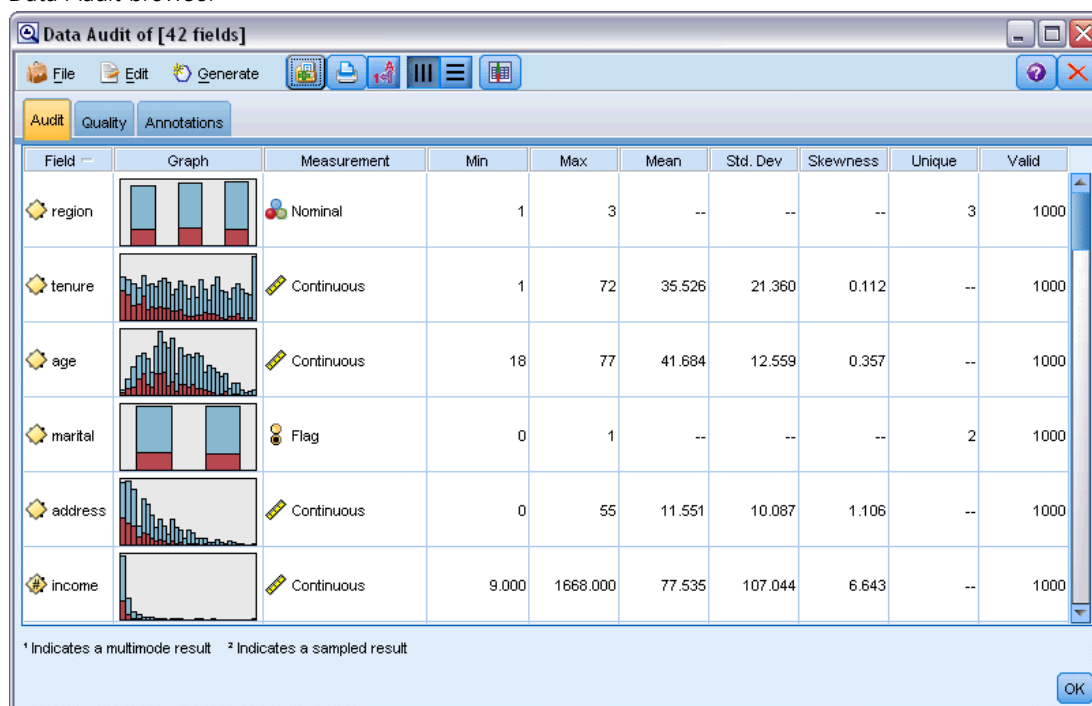
- **Range.** Shows the range (smallest and largest values) of confidence values for records in the stream data.
- **Mean Correct.** Shows the average confidence for records that are classified correctly.
- **Mean Incorrect.** Shows the average confidence for records that are classified incorrectly.
- **Always Correct Above.** Shows the confidence threshold above which predictions are always correct and shows the percentage of cases meeting this criterion.
- **Always Incorrect Below.** Shows the confidence threshold below which predictions are always incorrect and shows the percentage of cases meeting this criterion.
- **X% Accuracy Above.** Shows the confidence level at which accuracy is $X\%$. X is approximately the value specified for **Threshold for** in the Analysis options. For some models and datasets, it is not possible to choose a confidence value that gives the exact threshold specified in the options (usually due to clusters of similar cases with the same confidence value near the threshold). The threshold reported is the closest value to the specified accuracy criterion that can be obtained with a single confidence value threshold.
- **X Fold Correct Above.** Shows the confidence value at which accuracy is X times better than it is for the overall dataset. X is the value specified for **Improve accuracy** in the Analysis options.

Agreement between. If two or more generated models that predict the same output field are included in the stream, you will also see statistics on the **agreement** between predictions generated by the models. This includes the number and percentage of records for which the predictions agree (for categorical output fields) or error summary statistics (for continuous output fields). For categorical fields, it includes an analysis of predictions compared to actual values for the subset of records on which the models agree (generate the same predicted value).

Data Audit Node

The Data Audit node provides a comprehensive first look at the data you bring into IBM® SPSS® Modeler, presented in an easy-to-read matrix that can be sorted and used to generate full-size graphs and a variety of data preparation nodes.

Figure 6-14
Data Audit browser



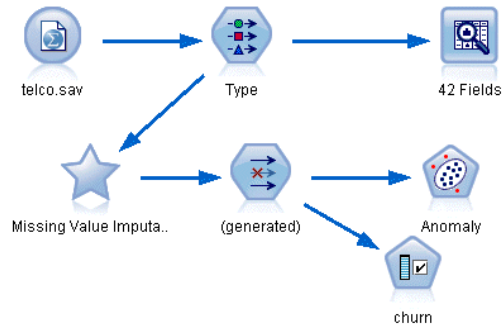
- The Audit tab displays a report that provides summary statistics, histograms, and distribution graphs that may be useful in gaining a preliminary understanding of the data. The report also displays the storage icon before the field name.
- The Quality tab in the audit report displays information about outliers, extremes, and missing values, and offers tools for handling these values.

Using the Data Audit Node

The Data Audit node can be attached directly to a source node or downstream from an instantiated Type node. You can also generate a number of data preparation nodes based on the results. For example, you can generate a Filter node that excludes fields with too many missing values to be

useful in modeling, and generate a SuperNode that imputes missing values for any or all of the fields that remain. This is where the real power of the audit comes in, allowing you not only to assess the current state of your data, but to take action based on the assessment.

Figure 6-15
Stream with Missing Values Supernode

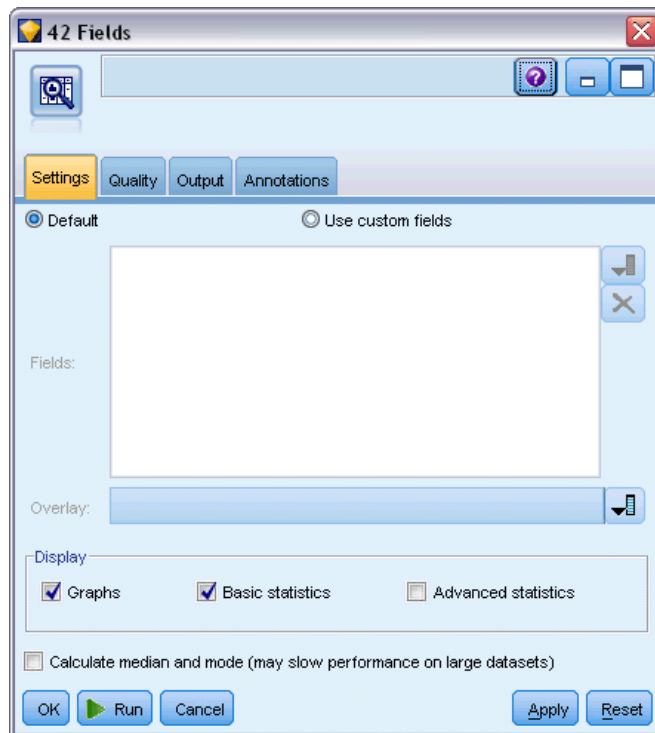


Screening or sampling the data. Because an initial audit is particularly effective when dealing with “big data,” a Sample node may be used to reduce processing time during the initial exploration by selecting only a subset of records. The Data Audit node can also be used in combination with nodes such as Feature Selection and Anomaly Detection in the exploratory stages of analysis.

Data Audit Node Settings Tab

The Settings tab allows you to specify basic parameters for the audit.

Figure 6-16
Data Audit node: Settings tab



Default. You can simply attach the node to your stream and click Run to generate an audit report for all fields based on default settings, as follows:

- If there are no Type node settings, all fields are included in the report.
- If there are Type settings (regardless of whether or not they are instantiated), all *Input*, *Target*, and *Both* fields are included in the display. If there is a single *Target* field, use it as the Overlay field. If there is more than one *Target* field specified, no default overlay is specified.

Use custom fields. Select this option to manually select fields. Use the field chooser button on the right to select fields individually or by type.

Overlay field. The overlay field is used in drawing the thumbnail graphs shown in the audit report. In the case of a continuous (numeric range) field, bivariate statistics (covariance and correlation) are also calculated. If a single *Target* field is present based on Type node settings, it is used as the default overlay field as described above. Alternatively, you can select Use custom fields in order to specify an overlay.

Display. Allows you to specify whether graphs are available in the output, and to choose the statistics displayed by default.

- **Graphs.** Displays a graph for each selected field; either a distribution (bar) graph, histogram, or scatterplot as appropriate for the data. Graphs are displayed as thumbnails in the initial report, but full-sized graphs and graph nodes can also be generated. For more information, see the topic [Data Audit Output Browser](#) on p. 340.
- **Basic/Advanced statistics.** Specifies the level of statistics displayed in the output by default. While this setting determines the initial display, all statistics are available in the output regardless of this setting. For more information, see the topic [Display Statistics](#) on p. 342.

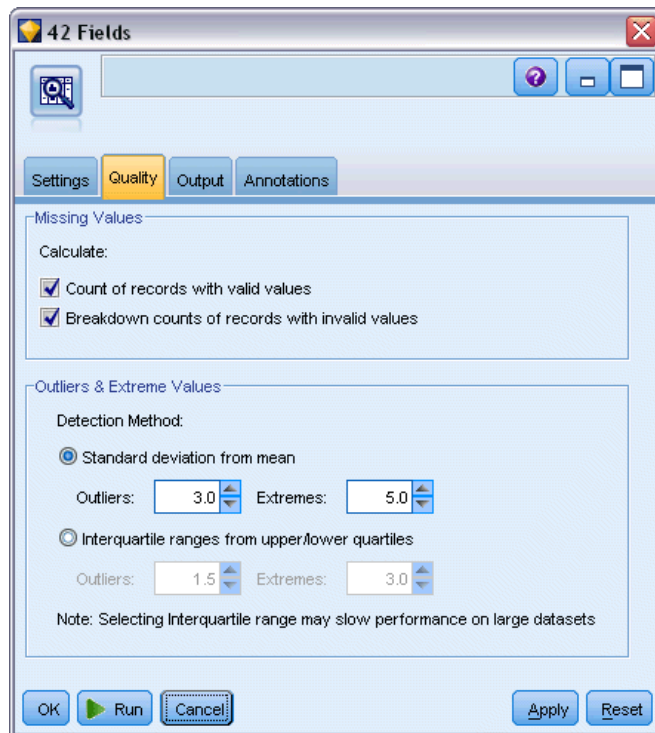
Median and mode. Calculates the median and mode for all fields in the report. Note that with large datasets, these statistics may increase processing time, since they take longer than others to compute. In the case of the median only, the reported value may be based on a sample of 2000 records (rather than the full dataset) in some cases. This sampling is done on a per-field basis in cases where memory limits would otherwise be exceeded. When sampling is in effect, the results will be labeled as such in the output (*Sample Median* rather than just *Median*). All statistics other than the median are always computed using the full dataset.

Empty or typeless fields. When used with instantiated data, typeless fields are not included in the audit report. To include typeless fields (including empty fields), select *Clear All Values* in any upstream *Type* nodes. This ensures that data are not instantiated, causing all fields to be included in the report. For example, this may be useful if you want to obtain a complete list of all fields or generate a *Filter* node that will exclude those that are empty. For more information, see the topic [Filtering Fields with Missing Data](#) on p. 347.

Data Audit Quality Tab

The *Quality* tab in the *Data Audit* node provides options for handling missing values, outliers, and extreme values.

Figure 6-17
Data Audit node Quality tab



Missing Values

- **Count of records with valid values.** Select this option to show the number of records with valid values for each evaluated field. Note that null (undefined) values, blank values, white spaces and empty strings are always treated as invalid values.
- **Breakdown counts of records with invalid values.** Select this option to show the number of records with each type of invalid value for each field.

Outliers and Extreme Values

Detection method for outliers and extreme values. Two methods are supported.:

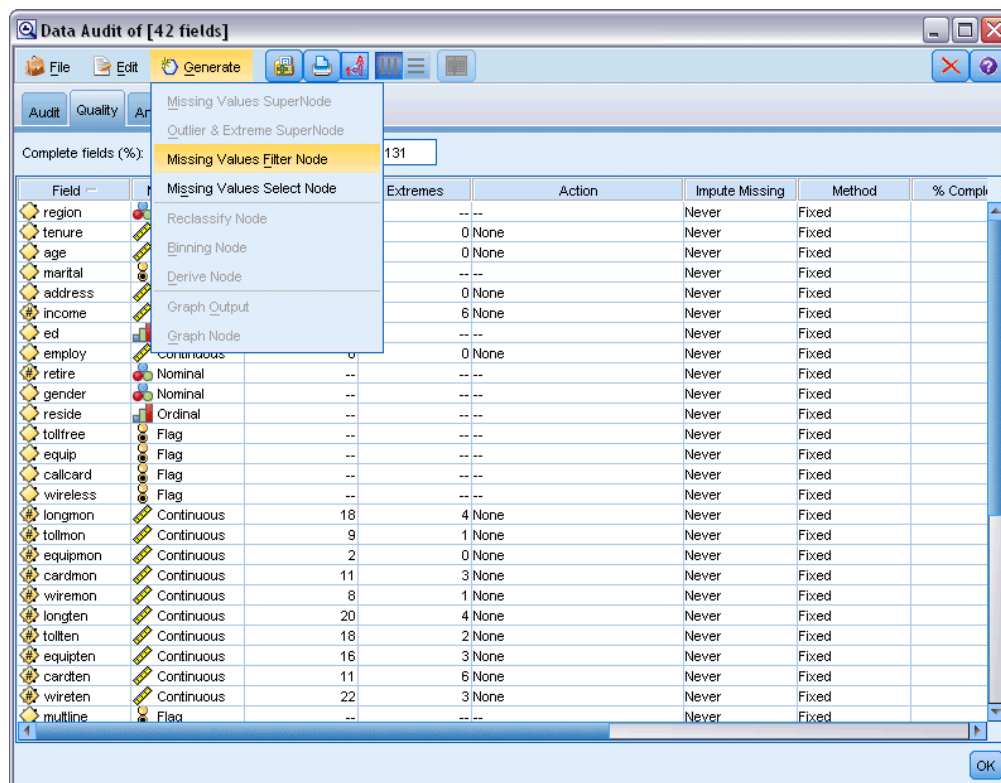
Standard deviation from the mean. Detects outliers and extremes based on the number of standard deviations from the mean. For example, if you have a field with a mean of 100 and a standard deviation of 10, you could specify 3.0 to indicate that any value below 70 or above 130 should be treated as an outlier.

Interquartile range. Detects outliers and extremes based on the interquartile range, which is the range within which the two central quartiles fall (between the 25th and 75th percentiles). For example, based on the default setting of 1.5, the lower threshold for outliers would be $Q1 - 1.5 * IQR$ and the upper threshold would be $Q3 + 1.5 * IQR$. Note that using this option may slow performance on large datasets.

Data Audit Output Browser

The Data Audit browser is a powerful tool for gaining overview of your data. The Audit tab displays thumbnail graphs, storage icons, and statistics for all fields, while the Quality tab displays information about outliers, extremes, and missing values. Based on the initial graphs and summary statistics, you might decide to recode a numeric field, derive a new field, or reclassify the values of a nominal field. Or you may want to explore further using more sophisticated visualization. You can do this right from the audit report browser using the Generate menu to create any number of nodes that can be used to transform or visualize the data.

Figure 6-18
Generating a Missing Values Filter node



- Sort columns by clicking on the column header, or reorder columns using drag and drop. Most standard output operations are also supported. For more information, see the topic [Viewing Output](#) on p. 316.
- View values and ranges for fields by double-clicking a field in the Measurement or Unique columns.
- Use the toolbar or Edit menu to show or hide value labels, or to choose the statistics you want to display. For more information, see the topic [Display Statistics](#) on p. 342.
- Verify the storage icons to the left of the field names. Storage describes the way data are stored in a field. For example, a field with values of 1 and 0 stores integer data. This is distinct from the measurement level, which describes the usage of the data, and does not affect storage. For more information, see the topic [Setting Field Storage and Formatting](#) in Chapter 2 on p. 23.

Viewing and Generating Graphs

If no overlay is selected, the Audit tab displays either bar charts (for nominal or flag fields) or histograms (continuous fields).

Figure 6-19
Excerpt of audit results without an overlay field

Field	Graph	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
region		Set	1	3	--	--	--	3	1000
tenure		Range	1	72	35.526	21.360	0.112	--	1000

For a nominal or flag field overlay, the graphs are colored by the values of the overlay.

Figure 6-20
Excerpt of audit results with a nominal field overlay

Field	Graph	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
region		Set	1	3	--	--	--	3	1000
tenure		Range	1	72	35.526	21.360	0.112	--	1000

For a continuous field overlay, two-dimensional scatterplots are generated rather than one-dimensional bars and histograms. In this case, the x axis maps to the overlay field, enabling you to see the same scale on all x axes as you read down the table.

Figure 6-21
Excerpt of audit results with a continuous field overlay

Field	Graph	Type	Min	Max	Mean	Correlation	Correlation T	Correlation T df.
region		Set	1	3	--	--	--	--
tenure		Range	1	72	35.526	0.490	17.768	998.000

- For Flag or Nominal fields, hold the mouse cursor over a bar to display the underlying value or label in a ToolTip.
- For Flag or Nominal fields, use the toolbar to toggle the orientation of thumbnail graphs from horizontal to vertical.
- To generate a full-sized graph from any thumbnail, double-click on the thumbnail, or select a thumbnail and choose Graph Output from the Generate menu. *Note:* If a thumbnail graph was based on sampled data, the generated graph will contain all cases if the original data stream is still open.

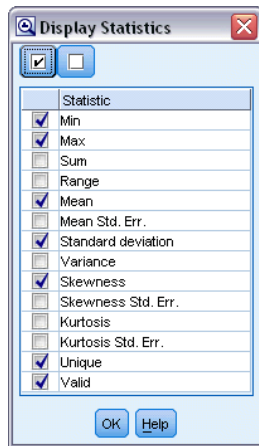
You can only generate a graph if the Data Audit node that created the output is connected to the stream.

- To generate a matching graph node, select one or more fields on the Audit tab and choose Graph Node from the Generate menu. The resulting node is added to the stream canvas and can be used to re-create the graph each time the stream is run.
- If an overlay set has more than 100 values, a warning is raised and the overlay is not included.

Display Statistics

The Display Statistics dialog box allows you to choose the statistics displayed on the Audit tab. The initial settings are specified in the Data Audit node. For more information, see the topic [Data Audit Node Settings Tab](#) on p. 336.

Figure 6-22
Display Statistics



Minimum. The smallest value of a numeric variable.

Maximum. The largest value of a numeric variable.

Sum. The sum or total of the values, across all cases with nonmissing values.

Range. The difference between the largest and smallest values of a numeric variable, the maximum minus the minimum.

Mean. A measure of central tendency. The arithmetic average, the sum divided by the number of cases.

Standard Error of Mean. A measure of how much the value of the mean may vary from sample to sample taken from the same distribution. It can be used to roughly compare the observed mean to a hypothesized value (that is, you can conclude the two values are different if the ratio of the difference to the standard error is less than -2 or greater than +2).

standard deviation. A measure of dispersion around the mean, equal to the square root of the variance. The standard deviation is measured in the same units as the original variable.

Variance. A measure of dispersion around the mean, equal to the sum of squared deviations from the mean divided by one less than the number of cases. The variance is measured in units that are the square of those of the variable itself.

Skewness. A measure of the asymmetry of a distribution. The normal distribution is symmetric and has a skewness value of 0. A distribution with a significant positive skewness has a long right tail. A distribution with a significant negative skewness has a long left tail. As a guideline, a skewness value more than twice its standard error is taken to indicate a departure from symmetry.

Standard Error of Skewness. The ratio of skewness to its standard error can be used as a test of normality (that is, you can reject normality if the ratio is less than -2 or greater than +2). A large positive value for skewness indicates a long right tail; an extreme negative value indicates a long left tail.

Kurtosis. A measure of the extent to which observations cluster around a central point. For a normal distribution, the value of the kurtosis statistic is zero. Positive kurtosis indicates that, relative to a normal distribution, the observations are more clustered about the center of the distribution and have thinner tails until the extreme values of the distribution, at which point the tails of the leptokurtic distribution are thicker relative to a normal distribution. Negative kurtosis indicates that, relative to a normal distribution, the observations cluster less and have thicker tails until the extreme values of the distribution, at which point the tails of the platykurtic distribution are thinner relative to a normal distribution.

Standard Error of Kurtosis. The ratio of kurtosis to its standard error can be used as a test of normality (that is, you can reject normality if the ratio is less than -2 or greater than +2). A large positive value for kurtosis indicates that the tails of the distribution are longer than those of a normal distribution; a negative value for kurtosis indicates shorter tails (becoming like those of a box-shaped uniform distribution).

Unique. Evaluates all effects simultaneously, adjusting each effect for all other effects of any type.

Valid. Valid cases having neither the system-missing value, nor a value defined as user-missing.

Median. The value above and below which half of the cases fall, the 50th percentile. If there is an even number of cases, the median is the average of the two middle cases when they are sorted in ascending or descending order. The median is a measure of central tendency not sensitive to outlying values (unlike the mean, which can be affected by a few extremely high or low values).

Mode. The most frequently occurring value. If several values share the greatest frequency of occurrence, each of them is a mode.

Note that median and mode are suppressed by default in order to improve performance but can be selected on the Settings tab in the Data Audit node. For more information, see the topic [Data Audit Node Settings Tab](#) on p. 336.

Statistics for Overlays

If a continuous (numeric range) overlay field is in use, the following statistics are also available:

Covariance. An unstandardized measure of association between two variables, equal to the cross-product deviation divided by $N-1$.

Data Audit Browser Quality Tab

Figure 6-23
Quality report in the Data Audit browser

The screenshot shows the 'Data Audit of [42 fields]' window with the 'Quality' tab selected. At the top, there are two input fields: 'Complete fields (%)' with the value '90.47619' and 'Complete records (%)' with the value '13.1'. Below these is a table with the following columns: Field, Measurement, Outliers, Extremes, Action, Impute Missing, and Method. The table lists 20 fields with their respective measurements, outlier and extreme counts, and the actions and imputation methods applied to them.

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method
region	Nominal	--	--		Never	Fixed
tenure	Continuous	0	0	None	Never	Fixed
age	Continuous	0	0	None	Never	Fixed
marital	Flag	--	--		Never	Fixed
address	Continuous	12	0	None	Never	Fixed
income	Continuous	9	6	None	Never	Fixed
ed	Ordinal	--	--		Never	Fixed
employ	Continuous	8	0	None	Never	Fixed
retire	Nominal	--	--		Never	Fixed
gender	Nominal	--	--		Never	Fixed
reside	Ordinal	--	--		Never	Fixed
tollfree	Flag	--	--		Never	Fixed
equip	Flag	--	--		Never	Fixed
callcard	Flag	--	--		Never	Fixed
wireless	Flag	--	--		Never	Fixed
longmon	Continuous	18	4	None	Never	Fixed
tollmon	Continuous	9	1	None	Never	Fixed
equipmon	Continuous	2	0	None	Never	Fixed
cardmon	Continuous	11	3	None	Never	Fixed

The Quality tab in the Data Audit browser displays the results of the data quality analysis and allows you to specify treatments for outliers, extremes, and missing values.

Imputing Missing Values

The audit report lists the percentage of complete records for each field, along with the number of valid, null, and blank values. You can choose to impute missing values for specific fields as appropriate, and then generate a SuperNode to apply these transformations.

- In the Impute Missing column, specify the type of values you want to impute, if any. You can choose to impute blanks, nulls, both, or specify a custom condition or expression that selects the values to impute.

There are several types of missing values recognized by IBM® SPSS® Modeler:

- **Null or system-missing values.** These are nonstring values that have been left blank in the database or source file and have not been specifically defined as “missing” in a source or Type node. System-missing values are displayed as \$null\$. Note that empty strings are not considered nulls in SPSS Modeler, although they may be treated as nulls by certain databases.

- **Empty strings and white space.** Empty string values and white space (strings with no visible characters) are treated as distinct from null values. Empty strings are treated as equivalent to white space for most purposes. For example, if you select the option to treat white space as blanks in a source or Type node, this setting applies to empty strings as well.
 - **Blank or user-defined missing values.** These are values such as unknown, 99, or -1 that are explicitly defined in a source node or Type node as missing. Optionally, you can also choose to treat nulls and white space as blanks, which allows them to be flagged for special treatment and to be excluded from most calculations. For example, you can use the @BLANK function to treat these values, along with other types of missing values, as blanks. For more information, see the topic [Using the Values Dialog Box](#) in Chapter 4 on p. 120.
- In the Method column, specify the method you want to use.

The following methods are available for imputing missing values:

Fixed. Substitutes a fixed value (either the field mean, midpoint of the range, or a constant that you specify).

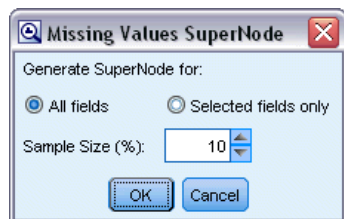
Random. Substitutes a random value based on a normal or uniform distribution.

Expression. Allows you to specify a custom expression. For example, you could replace values with a global variable created by the Set Globals node.

Algorithm. Substitutes a value predicted by a model based on the C&RT algorithm. For each field imputed using this method, there will be a separate C&RT model, along with a Filler node that replaces blanks and nulls with the value predicted by the model. A Filter node is then used to remove the prediction fields generated by the model.

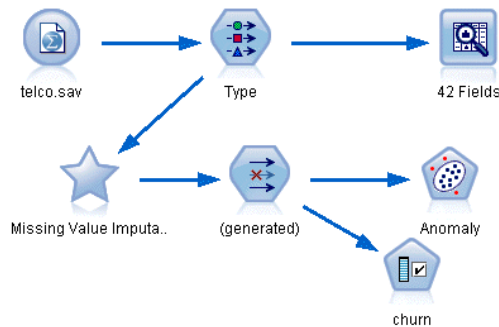
- To generate a Missing Values SuperNode, from the menus choose:
Generate > Missing Values SuperNode

Figure 6-24
Missing Values SuperNode dialog box



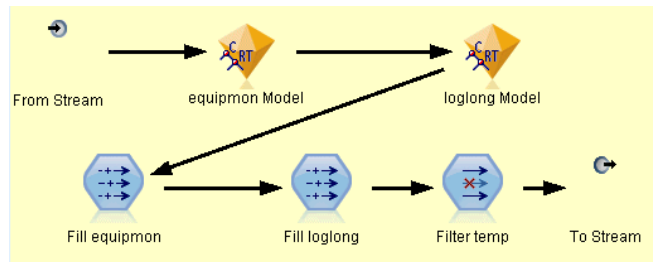
- Select All fields or Selected fields only, and specify a sample size if desired. (The specified sample is a percentage; by default, 10% of all records are sampled.)
- Click OK to add the generated SuperNode to the stream canvas.
- Attach the SuperNode to the stream to apply the transformations.

Figure 6-25
Adding the SuperNode to the stream



Within the SuperNode, a combination of model nugget, Filler, and Filter nodes is used as appropriate. To understand how it works, you can edit the SuperNode and click Zoom In, and you can add, edit, or remove specific nodes within the SuperNode to fine-tune the behavior.

Figure 6-26
Zooming in on the SuperNode



Handling Outliers and Extreme Values

The audit report lists number of outliers and extremes is listed for each field based on the detection options specified in the Data Audit node. For more information, see the topic [Data Audit Quality Tab](#) on p. 338. You can choose to coerce, discard, or nullify these values for specific fields as appropriate, and then generate a SuperNode to apply the transformations.

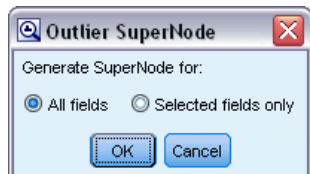
- ▶ In the Action column, specify handling for outliers and extremes for specific fields as desired.

The following actions are available for handling outliers and extremes:

- **Coerce.** Replaces outliers and extreme values with the nearest value that would not be considered extreme. For example if an outlier is defined to be anything above or below three standard deviations, then all outliers would be replaced with the highest or lowest value within this range.
- **Discard.** Discards records with outlying or extreme values for the specified field.
- **Nullify.** Replaces outliers and extremes with the null or system-missing value.
- **Coerce outliers / discard extremes.** Discards extreme values only.
- **Coerce outliers / nullify extremes.** Nullifies extreme values only.

- ▶ To generate the SuperNode, from the menus choose:
Generate > Outlier & Extreme SuperNode

Figure 6-27
Outlier SuperNode dialog box



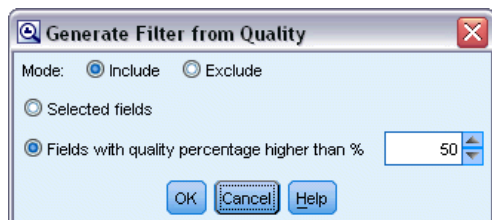
- ▶ Select All fields or Selected fields only, and then click OK to add the generated SuperNode to the stream canvas.
- ▶ Attach the SuperNode to the stream to apply the transformations.

Optionally, you can edit the SuperNode and zoom in to browse or make changes. Within the SuperNode, values are discarded, coerced, or nullified using a series of Select and/or Filler nodes as appropriate.

Filtering Fields with Missing Data

From the Data Audit browser, you can create a new Filter node based on the results of the Quality analysis.

Figure 6-28
Generate Filter from Quality dialog box



Mode. Select the desired operation for specified fields, either Include or Exclude.

- **Selected fields.** The Filter node will include/exclude the fields selected on the Quality tab. For example you could sort the table on the % Complete column, use Shift-click to select the least complete fields, and then generate a Filter node that excludes these fields.
- **Fields with quality percentage higher than.** The Filter node will include/exclude fields where the percentage of complete records is greater than the specified threshold. The default threshold is 50%.

Filtering Empty or Typeless Fields

Note that after data values have been instantiated, typeless or empty fields are excluded from the audit results and from most other output in IBM® SPSS® Modeler. These fields are ignored for purposes of modeling, but may bloat or clutter the data. If so, you can use the Data Audit browser to generate a Filter node from that removes these fields from the stream.

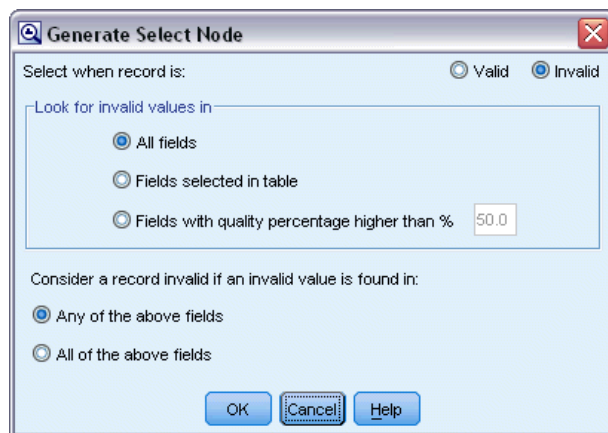
- ▶ To make sure that all fields are included in the audit, including empty or typeless fields, click **Clear All Values** in the upstream source or Type node, or set Values to `<Pass>` for all fields.
- ▶ In the Data Audit browser, sort on the % Complete column, select the fields that have zero valid values (or some other threshold) and use the Generate menu to produce a Filter node which can be added to the stream.

Selecting Records with Missing Data

From the Data Audit browser, you can create a new Select node based on the results of the quality analysis.

- ▶ In the Data Audit browser, choose the Quality tab.
- ▶ From the menu, choose:
Generate > Missing Values Select Node

Figure 6-29
Generate Select node dialog box



Select when record is. Specify whether records should be kept when they are Valid or Invalid.

Look for invalid values in. Specify where to check for invalid values.

- **All fields.** The Select node will check all fields for invalid values.
- **Fields selected in table.** The Select node will check only the fields currently selected in the Quality output table.
- **Fields with quality percentage higher than.** The Select node will check fields where the percentage of complete records is greater than the specified threshold. The default threshold is 50%.

Consider a record invalid if an invalid value is found in. Specify the condition for identifying a record as invalid.

- **Any of the above fields.** The Select node will consider a record invalid if *any* of the fields specified above contains an invalid value for that record.
- **All of the above fields.** The Select node will consider a record invalid only if *all* of the fields specified above contain invalid values for that record.

Generating Other Nodes for Data Preparation

A variety of nodes used in data preparation can be generated directly from the Data Audit browser, including Reclassify, Binning, and Derive nodes. For example:

- You can derive a new field based on the values of *claimvalue* and *farmincome* by selecting both in the audit report and choosing Derive from the Generate menu. The new node is added to the stream canvas.
- Similarly, you may determine, based on audit results, that recoding *farmincome* into percentile-based bins provides a more focused analysis. To generate a Binning node, select the field row in the display and choose Binning from the Generate menu.

Once a node is generated and added to the stream canvas, you must attach it to the stream and open the node to specify options for the selected field(s).

Transform Node

Normalizing input fields is an important step before using traditional scoring techniques, such as regression, logistic regression, and discriminant analysis. These techniques carry assumptions about normal distributions of data that may not be true for many raw data files. One approach to dealing with real-world data is to apply transformations that move a raw data element toward a more normal distribution. In addition, normalized fields can easily be compared with each other—for example, income and age are on totally different scales in a raw data file but when normalized, the relative impact of each can be easily interpreted.

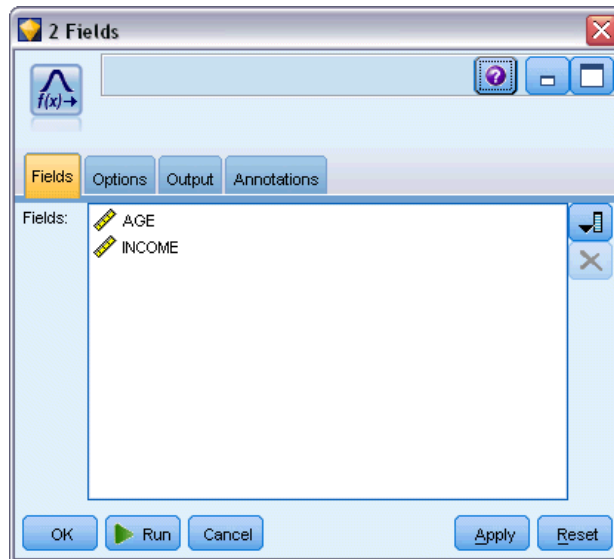
The Transform Node provides an output viewer that enables you to perform a rapid visual assessment of the best transformation to use. You can see at a glance whether variables are normally distributed and, if necessary, choose the transformation you want and apply it. You can pick multiple fields and perform one transformation per field.

After selecting the preferred transformations for the fields, you can generate Derive or Filler nodes that perform the transformations and attach these nodes to the stream. The Derive node creates new fields, while the Filler node transforms the existing ones. For more information, see the topic [Generating Graphs](#) on p. 353.

Transform Node Fields Tab

On the Fields tab, you can specify which fields of the data you want to use for viewing possible transformations and applying them. Only numeric fields can be transformed. Click the field selector button and select one or more numeric fields from the list displayed.

Figure 6-30
Transform node: Fields tab



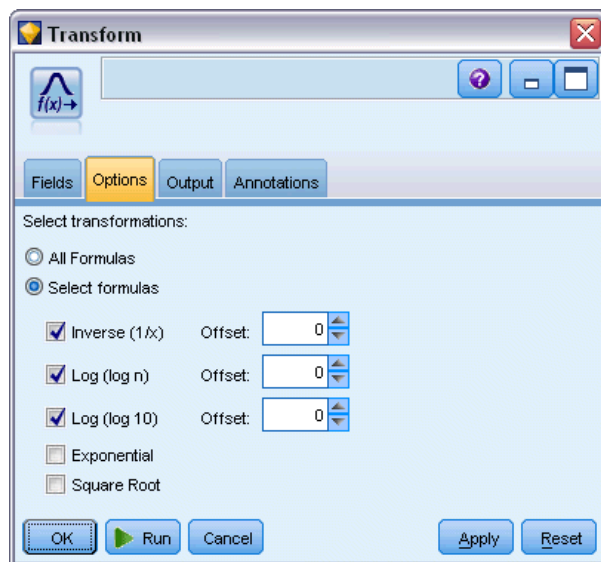
Transform Node Options Tab

The Options tab allows you to specify the type of transformations you want to include. You can choose to include all available transformations or select transformations individually.

In the latter case, you can also enter a number to offset the data for the inverse and log transformations. Doing so is useful in situations where a large proportion of zeros in the data would bias the mean and standard deviation results.

For example, assume that you have a field named *BALANCE* that has some zero values in it, and you want to use the inverse transformation on it. To avoid undesired bias, you would select Inverse ($1/x$) and enter 1 in the Use a data offset field. (Note that this offset is not related to that performed by the @OFFSET sequence function in IBM® SPSS® Modeler.)

Figure 6-31
Transform node: Options tab



All formulas. Indicates that all available transformations should be calculated and shown in the output.

Select formulas. Allows you to select the different transformations to be calculated and shown in the output.

- **Inverse (1/x).** Indicates that the inverse transformation should be displayed in the output.
- **Log (log n).** Indicates that the \log_n transformation should be displayed in the output.
- **Log (log 10).** Indicates that the \log_{10} transformation should be displayed in the output.
- **Exponential.** Indicates that the exponential transformation (e^x) should be displayed in the output.
- **Square Root.** Indicates that the square root transformation should be displayed in the output.

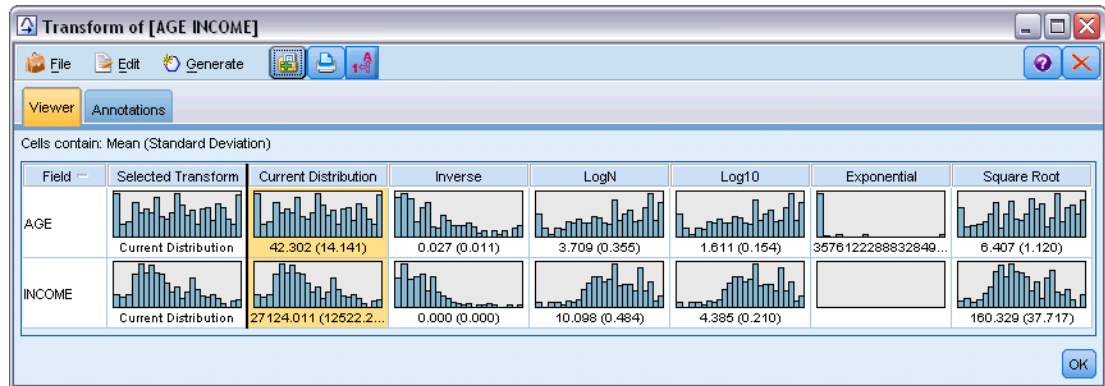
Transform Node Output Tab

The Output tab lets you specify the format and location of the output. You can choose to display the results on the screen, or send them to one of the standard file types. For more information, see the topic [Output Node Output Tab](#) on p. 322.

Transform Node Output Viewer

The output viewer enables you to see the results of executing the Transform Node. The viewer is a powerful tool that displays multiple transformations per field in thumbnail views of the transformation, enabling you to compare fields quickly. You can use options on its File menu to save, export, or print the output. For more information, see the topic [Viewing Output](#) on p. 316.

Figure 6-32
Viewing available transformations per field



For each transformation (other than Selected Transform), a legend is displayed underneath in the format:

Mean (Standard deviation)

Generating Nodes for the Transformations

The output viewer provides a useful starting point for your data preparation. For example, you might want to normalize the field *AGE* so that you can use a scoring technique (such as logistic regression or discriminant analysis) that assumes a normal distribution. Based upon the initial graphs and summary statistics, you might decide to transform the *AGE* field according to a particular distribution (for example, log). After selecting the preferred distribution, you can then generate a derive node with a standardized transformation to use for scoring.

You can generate the following field operations nodes from the output viewer:

- Derive
- Filler

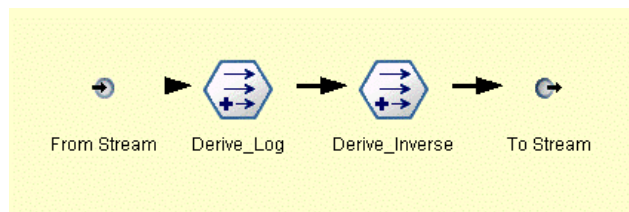
A Derive node creates new fields with the desired transformations, while the Filler node transforms existing fields. The nodes are placed on the canvas in the form of a SuperNode.

If you select the same transformation for different fields, a Derive or Filler node contains the formulas for that transformation type for all the fields to which that transformation applies. For example, assume that you have selected the following fields and transformations to generate a Derive node:

Field	Transformation
<i>AGE</i>	Current Distribution
<i>INCOME</i>	Log
<i>OPEN_BAL</i>	Inverse
<i>BALANCE</i>	Inverse

The following nodes are contained in the SuperNode:

Figure 6-33
SuperNode on canvas



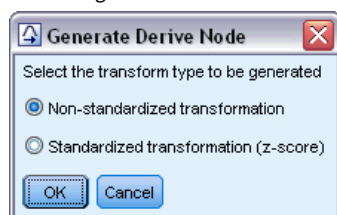
In this example, the Derive_Log node has the log formula for the *INCOME* field, and the Derive_Inverse node has the inverse formulas for the *OPEN_BAL* and *BALANCE* fields.

To Generate a Node

- ▶ For each field in the output viewer, select the desired transformation.
- ▶ From the Generate menu, choose Derive Node or Filler Node as desired.

Doing so displays the Generate Derive Node or Generate Filler Node dialog box, as appropriate.

Figure 6-34
Choosing standardized or non-standardized transformation



Choose Non-standardized transformation or Standardized transformation (z-score) as desired. The second option applies a *z* score to the transformation; *z* scores represent values as a function of distance from the mean of the variable in standard deviations. For example, if you apply the log transformation to the *AGE* field and choose a standardized transformation, the final equation for the generated node will be:

$$(\log(\text{AGE}) - \text{Mean}) / \text{SD}$$

Once a node is generated and appears on the stream canvas:

- ▶ Attach it to the stream.
- ▶ For a SuperNode, optionally double-click the node to view its contents.
- ▶ Optionally double-click a Derive or Filler node to modify options for the selected field(s).

Generating Graphs

You can generate full-size histogram output from a thumbnail histogram in the output viewer.

To Generate a Graph

- ▶ Double-click a thumbnail graph in the output viewer.

or

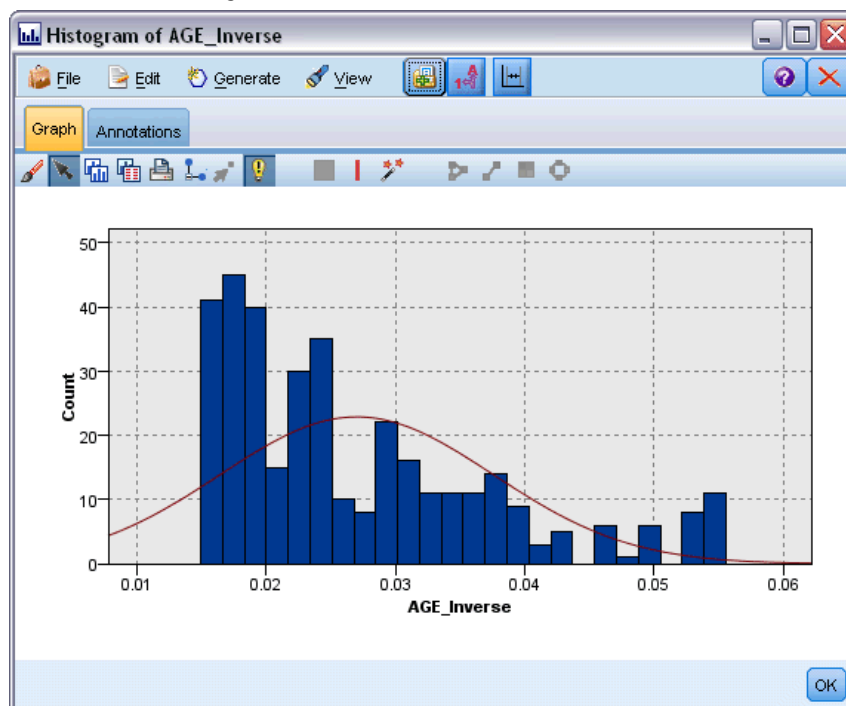
- ▶ Select a thumbnail graph in the output viewer.
- ▶ From the Generate menu, choose Graph output.

Doing so displays the histogram with a normal distribution curve overlaid. This enables you to compare how closely each available transformation matches a normal distribution.

Note: You can only generate a graph if the Transform node that created the output is connected to the stream.

Figure 6-35

Transformation histogram with normal distribution curve overlaid



Other Operations

From the output viewer, you can also:

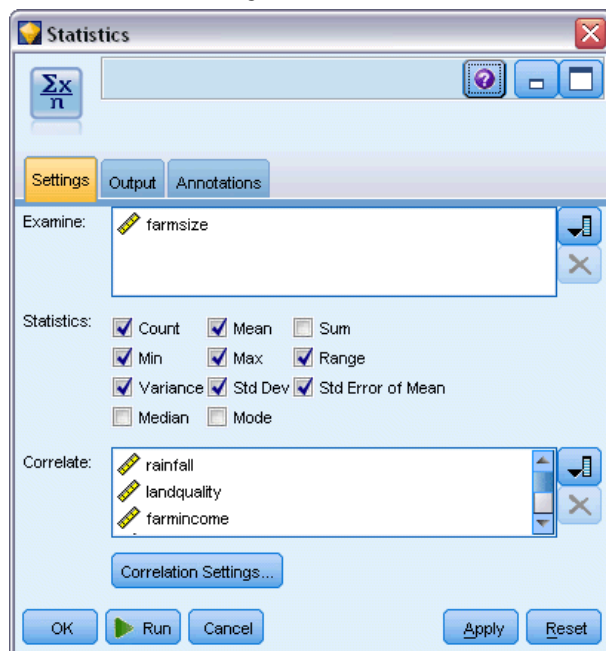
- Sort the output grid by the Field column.
- Export the output to an HTML file. For more information, see the topic [Exporting Output](#) on p. 319.

Statistics Node

The Statistics node gives you basic summary information about numeric fields. You can get summary statistics for individual fields and correlations between fields.

Statistics Node Settings Tab

Figure 6-36
Statistics node: Settings tab



Examine. Select the field or fields for which you want individual summary statistics. You can select multiple fields.

Statistics. Select the statistics to report. Available options include Count, Mean, Sum, Min, Max, Range, Variance, Std Dev, Std Error of Mean, Median, and Mode.

Correlate. Select the field or fields that you want to correlate. You can select multiple fields. When correlation fields are selected, the correlation between each Examine field and the correlation field(s) will be listed in the output.

Correlation Settings. You can specify options for displaying the strength of correlations in the output.

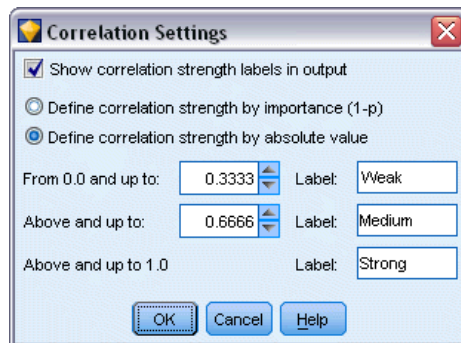
Correlation Settings

IBM® SPSS® Modeler can characterize correlations with descriptive labels to help highlight important relationships. The **correlation** measures the strength of relationship between two continuous (numeric range) fields. It takes values between -1.0 and 1.0 . Values close to $+1.0$ indicate a strong positive association so that high values on one field are associated with high values on the other and low values are associated with low values. Values close to -1.0 indicate a

strong negative association so that high values for one field are associated with low values for the other, and vice versa. Values close to 0.0 indicate a weak association, so that values for the two fields are more or less independent.

You can control display of correlation labels, change the thresholds that define the categories, and change the labels used for each range. Because the way you characterize correlation values depends greatly on the problem domain, you may want to customize the ranges and labels to fit your specific situation.

Figure 6-37
Correlation Settings dialog box



Show correlation strength labels in output. This option is selected by default. Deselect this option to omit the descriptive labels from the output.

Correlation Strength. There are two options for defining and labeling the strength of correlations:

- **Define correlation strength by importance (1-p).** Labels correlations based on importance, defined as 1 minus the significance, or 1 minus the probability that the difference in means could be explained by chance alone. The closer this value comes to 1, the greater the chance that the two fields are *not* independent—in other words, that some relationship exists between them. Labeling correlations based on importance is generally recommended over absolute value because it accounts for variability in the data—for example, a coefficient of 0.6 may be highly significant in one dataset and not significant at all in another. By default, importance values between 0.0 and 0.9 are labeled as *Weak*, those between 0.9 and 0.95 are labeled as *Medium*, and those between 0.95 and 1.0 are labeled as *Strong*.
- **Define correlation strength by absolute value.** Labels correlations based on the absolute value of the Pearson's correlation coefficient, which ranges between -1 and 1 , as described above. The closer the absolute value of this measure comes to 1, the stronger the correlation. By default, correlations between 0.0 and 0.3333 (in absolute value) are labeled as *Weak*, those between 0.3333 and 0.6666 are labeled as *Medium*, and those between 0.6666 and 1.0 are labeled as *Strong*. Note, however, that the significance of any given value is difficult to generalize from one dataset to another; for this reason, defining correlations based on probability rather than absolute value is recommended in most cases.

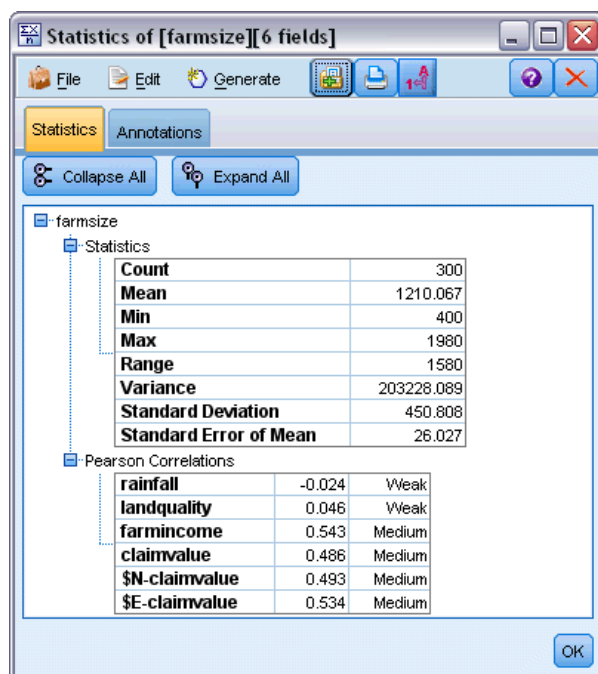
Statistics Output Browser

The Statistics node output browser displays the results of the statistical analysis and allows you to perform operations, including selecting fields, generating new nodes based on the selection, and saving and printing the results. The usual saving, exporting, and printing options are available

from the File menu, and the usual editing options are available from the Edit menu. For more information, see the topic [Viewing Output](#) on p. 316.

When you first browse Statistics output, the results are expanded. To hide results after viewing them, use the expander control to the left of the item to collapse the specific results you want to hide or click the Collapse All button to collapse all results. To see results again after collapsing them, use the expander control to the left of the item to show the results or click the Expand All button to show all results.

Figure 6-38
Statistics output browser



The output contains a section for each *Examine* field, containing a table of the requested statistics.

- **Count.** The number of records with valid values for the field.
- **Mean.** The average (mean) value for the field across all records.
- **Sum.** The sum of values for the field across all records.
- **Min.** The minimum value for the field.
- **Max.** The maximum value for the field.
- **Range.** The difference between the minimum and maximum values.
- **Variance.** A measure of the variability in the values of a field. It is calculated by taking the difference between each value and the overall mean, squaring it, summing across all of the values, and dividing by the number of records.
- **Standard Deviation.** Another measure of variability in the values of a field, calculated as the square root of the variance.
- **Standard Error of Mean.** A measure of the uncertainty in the estimate of the field's mean if the mean is assumed to apply to new data.

- **Median.** The “middle” value for the field; that is, the value that divides the upper half of the data from the lower half of the data (based on values of the field).
- **Mode.** The most common single value in the data.

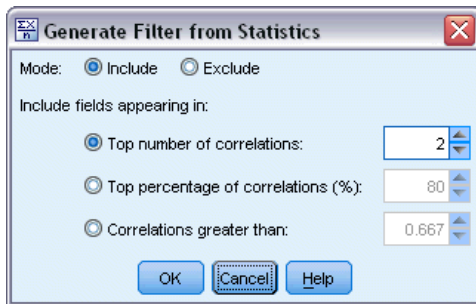
Correlations. If you specified correlate fields, the output also contains a section listing the Pearson correlation between the Examine field and each correlate field, and optional descriptive labels for the correlation values. For more information, see the topic [Correlation Settings](#) on p. 355.

Generate menu. The Generate menu contains node generation operations.

- **Filter.** Generates a Filter node to filter out fields that are uncorrelated or weakly correlated with other fields.

Generating a Filter Node from Statistics

Figure 6-39
Generate Filter from Statistics dialog box



The Filter node generated from a Statistics output browser will filter fields based on their correlations with other fields. It works by sorting the correlations in order of absolute value, taking the largest correlations (according to the criterion set in the dialog box), and creating a filter that passes all fields that appear in any of those large correlations.

Mode. Decide how to select correlations. Include causes fields appearing in the specified correlations to be retained. Exclude causes the fields to be filtered.

Include/Exclude fields appearing in. Define the criterion for selecting correlations.

- **Top number of correlations.** Selects the specified number of correlations and includes/excludes fields that appear in any of those correlations.
- **Top percentage of correlations (%).** Selects the specified percentage ($n\%$) of correlations and includes/excludes fields that appear in any of those correlations.
- **Correlations greater than.** Selects correlations greater in absolute value than the specified threshold.

Means Node

The Means node compares the means between independent groups or between pairs of related fields to test whether a significant difference exists. For example, you can compare mean revenues before and after running a promotion or compare revenues from customers who didn't receive the promotion with those who did.

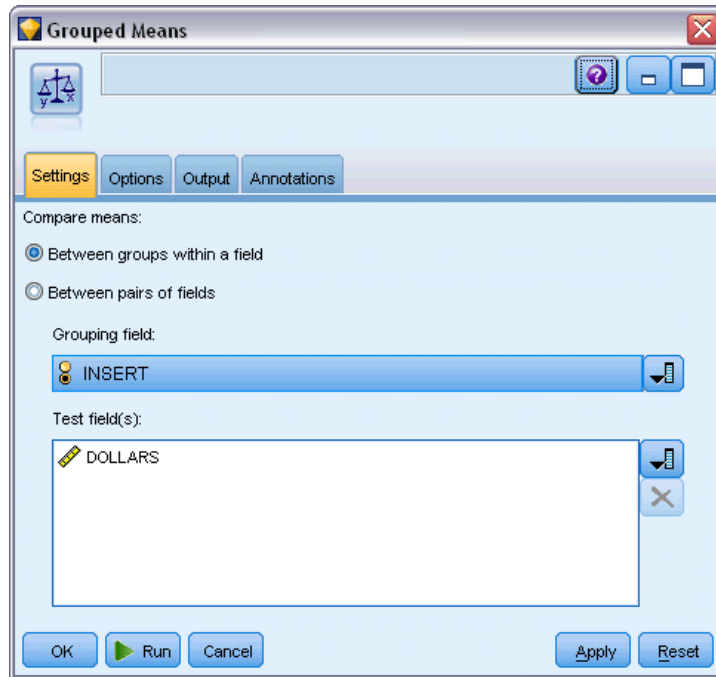
You can compare means in two different ways, depending on your data:

- **Between groups within a field.** To compare independent groups, select a test field and a grouping field. For example, you could exclude a sample of "holdout" customers when sending a promotion and compare mean revenues for the holdout group with all of the others. In this case, you would specify a single test field that indicates the revenue for each customer, with a flag or nominal field that indicates whether they received the offer. The samples are independent in the sense that each record is assigned to one group or another, and there is no way to link a specific member of one group to a specific member of another. You can also specify a nominal field with more than two values to compare the means for multiple groups. When executed, the node calculates a one-way ANOVA test on the selected fields. In cases where there are only two field groups, the one-way ANOVA results are essentially the same as an independent-samples t test. For more information, see the topic [Comparing Means for Independent Groups](#) on p. 359.
- **Between pairs of fields.** When comparing means for two related fields, the groups must be paired in some way for the results to be meaningful. For example, you could compare the mean revenues from the same group of customers before and after running a promotion or compare usage rates for a service between husband-wife pairs to see if they are different. Each record contains two separate but related measures that can be compared meaningfully. When executed, the node calculates a paired-samples t test on each field pair selected. For more information, see the topic [Comparing Means Between Paired Fields](#) on p. 360.

Comparing Means for Independent Groups

Select Between groups within a field in the Means node to compare the mean for two or more independent groups.

Figure 6-40
Comparing means between groups within one field



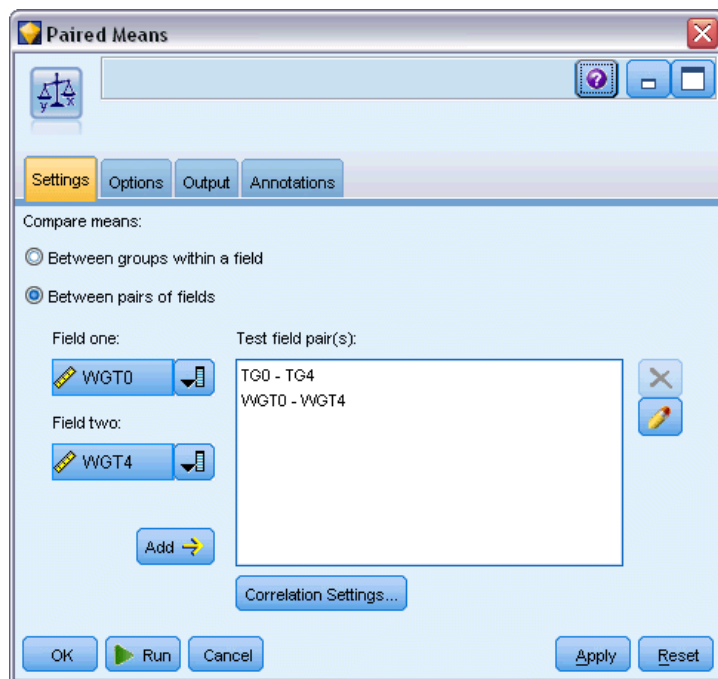
Grouping field. Select a numeric flag or nominal field with two or more distinct values that divides records into the groups you want to compare, such as those who received an offer versus those who did not. Regardless of the number of test fields, only one grouping field can be selected.

Test fields. Select one or more numeric fields that contain the measures you want to test. A separate test will be conducted for each field you select. For example, you could test the impact of a given promotion on usage, revenue, and churn.

Comparing Means Between Paired Fields

Select *Between pairs of fields* in the Means node to compare means between separate fields. The fields must be related in some way for the results to be meaningful, such as revenues before and after a promotion. Multiple field pairs can also be selected.

Figure 6-41
Comparing means between paired fields



Field one. Select a numeric field that contains the first of the measures you want to compare. In a before-and-after study, this would be the “before” field.

Field two. Select the second field you want to compare.

Add. Adds the selected pair to the Test field pair(s) list.

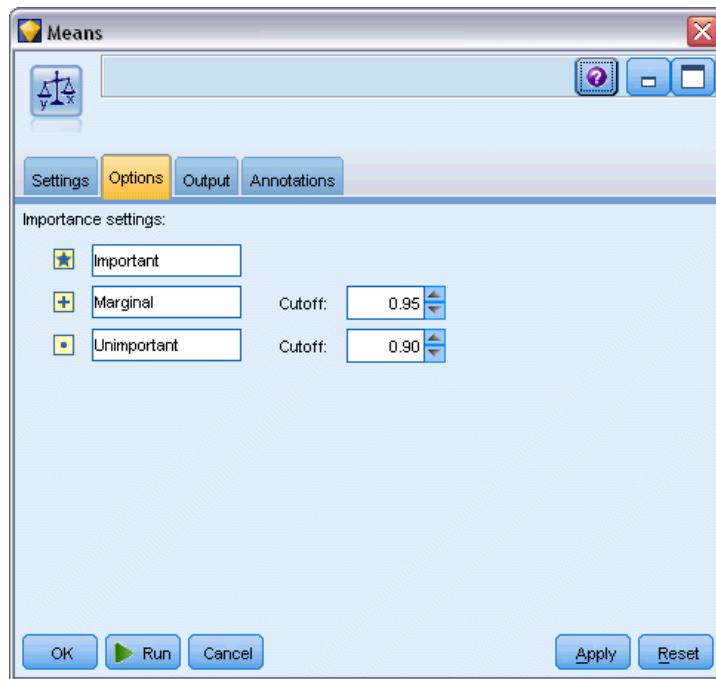
Repeat field selections as needed to add multiple pairs to the list.

Correlation settings. Allows you to specify options for labeling the strength of correlations. For more information, see the topic [Correlation Settings](#) on p. 355.

Means Node Options

The Options tab allows you to set the threshold p values used to label results as important, marginal, or unimportant. You can also edit the label for each ranking. Importance is measured on a percentage scale and can be broadly defined as 1 minus the probability of obtaining a result (such as the difference in means between two fields) as extreme as or more extreme than the observed result by chance alone. For example, a p value greater than 0.95 indicates less than a 5% chance that the result could be explained by chance alone.

Figure 6-42
Importance settings



Importance labels. You can edit the labels used to label each field pair or group in the output. The default labels are *important*, *marginal*, and *unimportant*.

Cutoff values. Specifies the threshold for each rank. Typically p values greater than 0.95 would rank as important, while those lower than 0.9 would be unimportant, but these thresholds can be adjusted as needed.

Note: Importance measures are available in a number of nodes. The specific computations depend on the node and on the type of target and input fields used, but the values can still be compared, since all are measured on a percentage scale.

Means Node Output Browser

The Means output browser displays cross-tabulated data and allows you to perform standard operations including selecting and copying the table one row at a time, sorting by any column, and saving and printing the table. For more information, see the topic [Viewing Output](#) on p. 316.

The specific information in the table depends on the type of comparison (groups within a field or separate fields).

Sort by. Allows you to sort the output by a specific column. Click the up or down arrow to change the direction of the sort. Alternatively, you can click on any column heading to sort by that column. (To change the direction of the sort within the column, click again.)

View. You can choose Simple or Advanced to control the level of detail in the display. The advanced view includes all of the information from the simple view but with additional details provided.

Means Output Comparing Groups within a Field

When comparing groups within a field, the name of the grouping field is displayed above the output table, and means and related statistics are reported separately for each group. The table includes a separate row for each test field.

Figure 6-43
Comparing groups within a field

Field	Standard*	New Promoti...	Importance
\$ spent during promotional period	1566.389	1637.500	0.976 ★ Important

The following columns are displayed:

- **Field.** Lists the names of the selected test fields.
- **Means by group.** Displays the mean for each category of the grouping field. For example, you might compare those who received an special offer (*New Promotion*) with those who didn't (*Standard*). In the advanced view, the standard deviation, standard error, and count are also displayed.
- **Importance.** Displays the importance value and label. For more information, see the topic [Means Node Options](#) on p. 361.

Advanced Output

In the advanced view, the following additional columns are displayed.

- **F-Test.** This test is based on the ratio of the variance between the groups and the variance within each group. If the means are the same for all groups, you would expect the F ratio to be close to 1 since both are estimates of the same population variance. The larger this ratio, the greater the variation between groups and the greater than chance that a significant difference exists.
- **df.** Displays the degrees of freedom.

Means Output Comparing Pairs of Fields

When comparing separate fields, the output table includes a row for each selected field pair.

Figure 6-44
Comparing pairs of fields

Field One	Field Two	Mean One*	Mean Two*	Correlation	Mean Differe...	Importance
Triglyceride	Final triglyceride	138.438	124.375	-0.286 Weak	14.062	0.751 <input type="checkbox"/> Unimportant
Weight	Final weight	198.375	190.312	0.996 Strong	8.062	1.000 <input checked="" type="checkbox"/> Important

- **Field One/Two.** Displays the name of the first and second field in each pair. In the advanced view, the standard deviation, standard error, and count are also displayed.
- **Mean One/Two.** Displays the mean for each field, respectively.
- **Correlation.** Measures the strength of relationship between two continuous (numeric range) fields. Values close to +1.0 indicate a strong positive association, and values close to -1.0 indicate a strong negative association. For more information, see the topic [Correlation Settings](#) on p. 355.
- **Mean Difference.** Displays the difference between the two field means.
- **Importance.** Displays the importance value and label. For more information, see the topic [Means Node Options](#) on p. 361.

Advanced Output

Advanced output adds the following columns:

95% Confidence Interval. Lower and upper boundaries of the range within which the true mean is likely to fall in 95% of all possible samples of this size from this population.

T-Test. The t statistic is obtained by dividing the mean difference by its standard error. The greater the absolute value of this statistic, the greater the probability that the means are not the same.

df. Displays the degrees of freedom for the statistic.

Report Node

The Report node allows you to create formatted reports containing fixed text, as well as data and other expressions derived from the data. You specify the format of the report by using text templates to define the fixed text and the data output constructions. You can provide custom text

formatting using HTML tags in the template and by setting options on the Output tab. Data values and other conditional output are included in the report using CLEM expressions in the template.

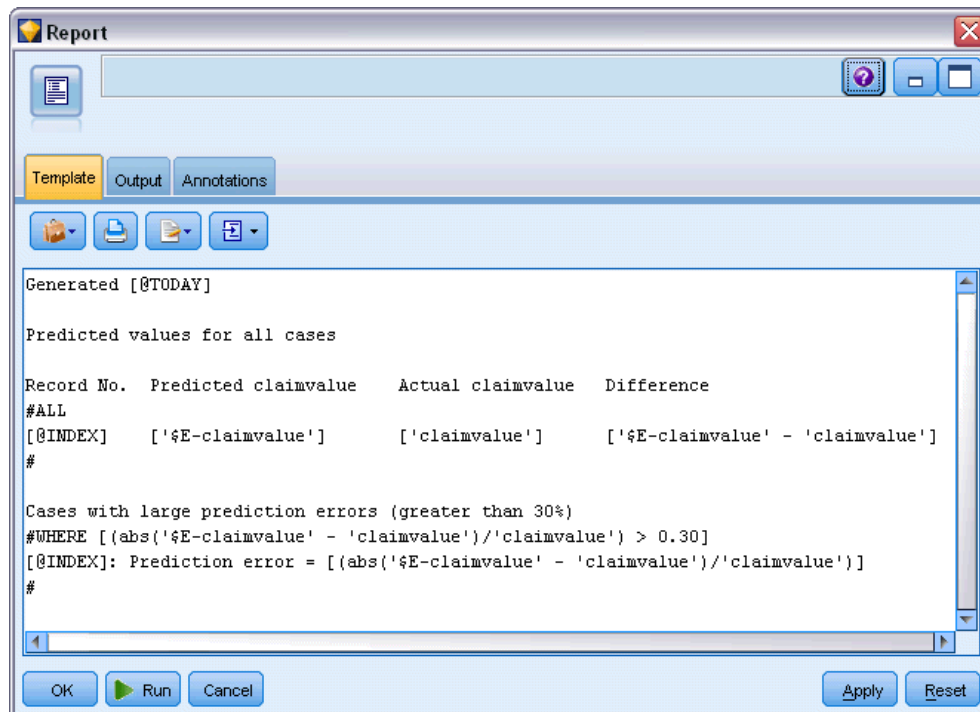
Alternatives to the Report Node

The Report node is most typically used to list records or cases output from a stream, such as all records meeting a certain condition. In this regard, it can be thought of as a less-structured alternative to the Table node.

- If you want a report that lists field information or anything else that is defined in the stream rather than the data itself (such as field definitions specified in a Type node), then a script can be used instead.
- To generate a report that includes multiple output objects (such as a collection of models, tables, and graphs generated by one or more streams) and that can be output in multiple formats (including text, HTML, and Microsoft Word/Office), an IBM® SPSS® Modeler project can be used.
- To produce a list of field names without using scripting, you can use a Table node preceded by a Sample node that discards all records. This produces a table with no rows, which can be transposed on export to produce a list of field names in a single column. (Select Transpose data on the Output tab in the Table node to do this.)

Report Node Template Tab

Figure 6-45
Report node: Template tab



Creating a template. To define the contents of the report, you create a template on the Report node Template tab. The template consists of lines of text, each of which specifies something about the contents of the report, and some special tag lines used to indicate the scope of the content lines. Within each content line, CLEM expressions enclosed in square brackets ([]) are evaluated before the line is sent to the report. There are three possible scopes for a line in the template:

Fixed. Lines that are not marked otherwise are considered fixed. Fixed lines are copied into the report only once, after any expressions that they contain are evaluated. For example, the line

```
This is my report, printed on [@TODAY]
```

would copy a single line to the report, containing the text and the current date.

Global (iterate ALL). Lines contained between the special tags #ALL and # are copied to the report once for each record of input data. CLEM expressions (enclosed in brackets) are evaluated based on the current record for each output line. For example, the lines

```
#ALL
For record [@INDEX], the value of AGE is [AGE]
#
```

would include one line for each record indicating the record number and age.

To generate a list of all records:

```
#ALL
[Age] [Sex] [Cholesterol] [BP]
#
```

Conditional (iterate WHERE). Lines contained between the special tags #WHERE *<condition>* and # are copied to the report once for each record where the specified condition is true. The condition is a CLEM expression. (In the WHERE condition, the brackets are optional.) For example, the lines

```
#WHERE [SEX = 'M']
Male at record no. [@INDEX] has age [AGE].
#
```

will write one line to the file for each record with a value of *M* for sex. The complete report will contain the fixed, global, and conditional lines defined by applying the template to the input data.

You can specify options for displaying or saving results using the Output tab, common to various types of output nodes. For more information, see the topic [Output Node Output Tab](#) on p. 322.

Outputting Data in HTML or XML Format

You can include HTML or XML tags directly in the template in order to write reports in either of these formats. For example, the following template produces an HTML table.

```
This report is written in HTML.
Only records where Age is above 60 are included.
```

```
<HTML>
<TABLE border="2">
```

```

<TR>
  <TD>Age</TD>
  <TD>BP</TD>
  <TD>Cholesterol</TD>
  <TD>Drug</TD>
</TR>

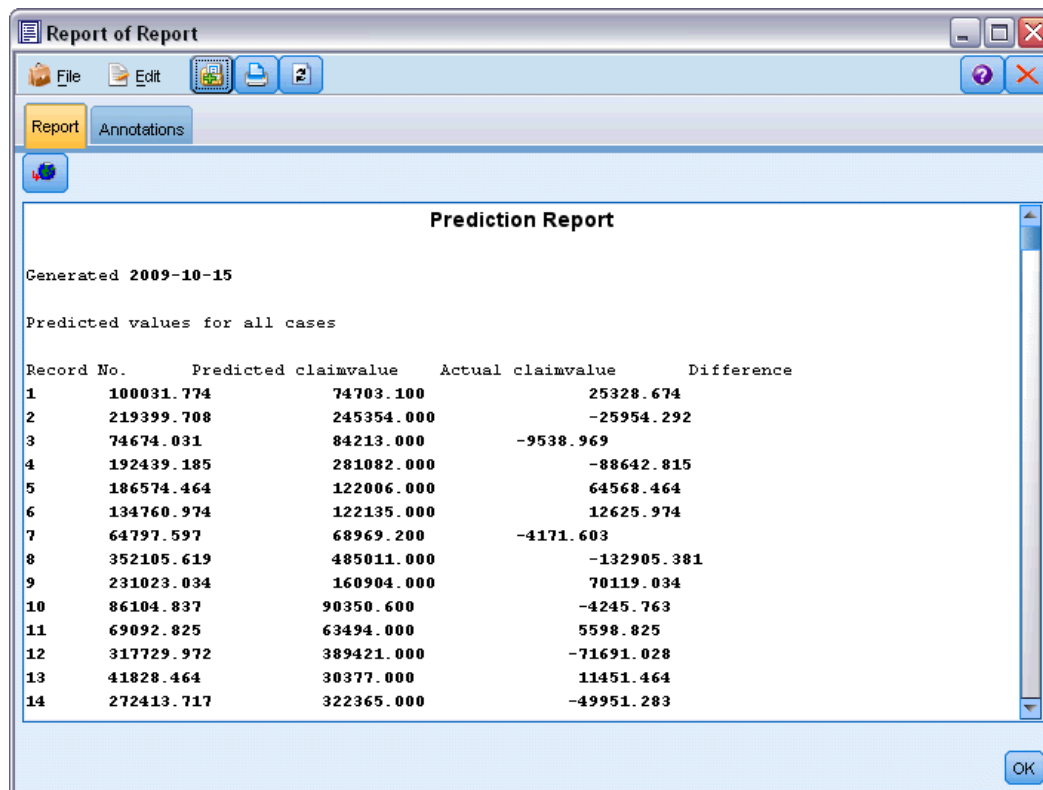
#WHERE Age > 60
<TR>
  <TD>[Age]</TD>
  <TD>[BP]</TD>
  <TD>[Cholesterol]</TD>
  <TD>[Drug]</TD>
</TR>
#
</TABLE>
</HTML>

```

Report Node Output Browser

The report browser shows you the contents of the generated report. The usual saving, exporting, and printing options are available from the File menu, and the usual editing options are available from the Edit menu. For more information, see the topic [Viewing Output](#) on p. 316.

Figure 6-46
Report browser

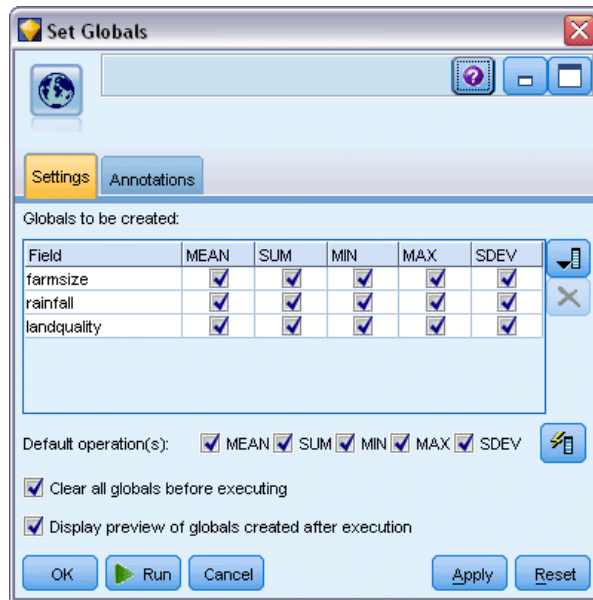


Set Globals Node

The Set Globals node scans the data and computes summary values that can be used in CLEM expressions. For example, you can use a Set Globals node to compute statistics for a field called *age* and then use the overall mean of *age* in CLEM expressions by inserting the function @GLOBAL_MEAN(*age*).

Set Globals Node Settings Tab

Figure 6-47
Set Globals node: Settings tab



Globals to be created. Select the field or fields for which you want globals to be available. You can select multiple fields. For each field, specify the statistics to compute by making sure that the statistics you want are selected in the columns next to the field name.

- **MEAN.** The average (mean) value for the field across all records.
- **SUM.** The sum of values for the field across all records.
- **MIN.** The minimum value for the field.
- **MAX.** The maximum value for the field.
- **SDEV.** The standard deviation, which is a measure of variability in the values of a field and is calculated as the square root of the variance.

Default operation(s). The options selected here will be used when new fields are added to the Globals list above. To change the default set of statistics, select or deselect statistics as appropriate. You can also use the Apply button to apply the default operations to all fields in the list.

Clear all globals before executing. Select this option to remove all global values before calculating new values. If this option is not selected, newly calculated values replace older values, but globals that are not recalculated remain available, as well.

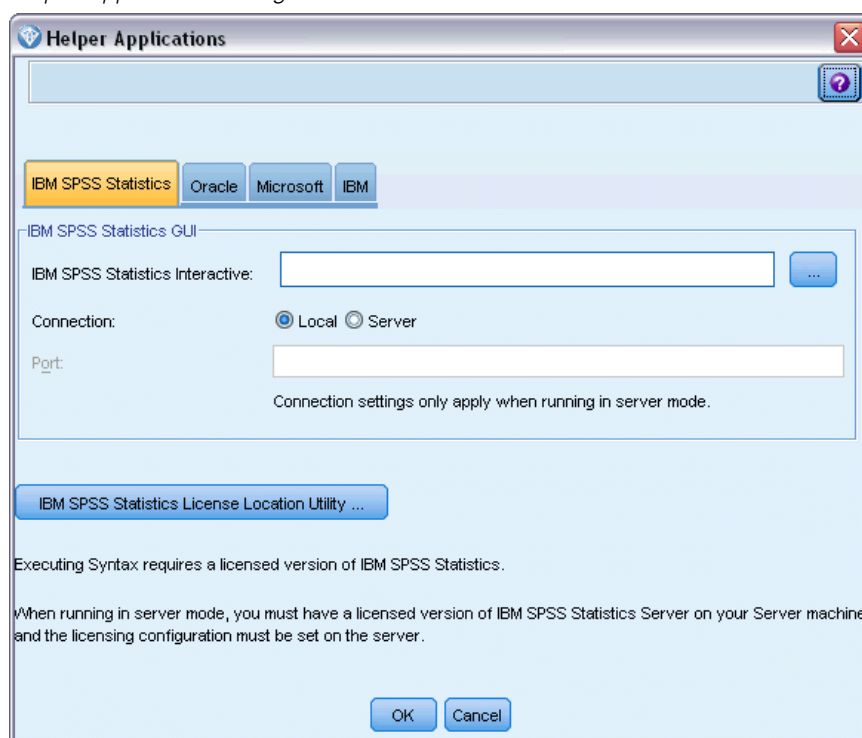
Display preview of globals created after execution. If you select this option, the Globals tab of the stream properties dialog box will appear after execution to display the calculated global values.

IBM SPSS Statistics Helper Applications

If a compatible version of IBM® SPSS® Statistics is installed and licensed on your computer, you can configure IBM® SPSS® Modeler to process data with SPSS Statistics functionality using the Statistics Transform, Statistics Model, Statistics Output, or Statistics Export nodes.

- To configure SPSS Modeler to work with SPSS Statistics and other applications, choose:
Tools > Options > Helper Applications

Figure 6-48
Helper Applications dialog box



IBM SPSS Statistics Interactive. Enter the full path and name of the command (for example, *C:\Program Files\IBM\SPSS\Statistics\<nn>\stats.exe*) to be used when launching SPSS Statistics directly on a data file produced by the Statistics Export node. For more information, see the topic [Statistics Export Node](#) in Chapter 8 on p. 409.

Connection. If SPSS Statistics Server is located on the same host as IBM® SPSS® Modeler Server, you can enable a connection between the two applications, which increases efficiency by leaving data on the server during analysis. Select Server to enable the Port option below. The default setting is Local.

Port. Specify the server port for SPSS Statistics Server.

IBM SPSS Statistics License Location Utility. To enable SPSS Modeler to use the Statistics Transform, Statistics Model, and Statistics Output nodes, you must have a copy of SPSS Statistics installed and licensed on the computer where the stream is run. Additionally, if you are running in distributed mode against a remote SPSS Modeler Server, you must also have a copy of SPSS Statistics client installed and licensed on your SPSS Modeler client computer.

- If running SPSS Modeler in local (standalone) mode, the licensed copy of SPSS Statistics must be on the local computer. Click this button to specify the location of the local SPSS Statistics installation you want to use for licensing.
- In addition, if running in distributed mode against a remote SPSS Modeler Server, you also need a licensed version of SPSS Statistics on the server computer, and the license configuration must also be done on the server. To do this, from the command prompt, change to the SPSS Modeler Server *bin* directory and, for Windows, run:

```
statisticsutility -location=<path to IBM SPSS Statistics Server license file>/bin
```

Alternatively, for UNIX, run:

```
./statisticsutility -location=<path to IBM SPSS Statistics Server license file>/bin
```

where *<path to SPSS Statistics Server license file>* is the installation directory of a licensed SPSS Statistics server.

If you do not have a licensed copy of SPSS Statistics on your local machine, you can still run the Statistics File node against a licensed SPSS Statistics server, but attempts to run other SPSS Statistics nodes will display an error message.

Comments

If you have trouble running the SPSS Statistics procedure nodes, consider the following tips:

- If field names used in SPSS Modeler are longer than eight characters (for versions prior to SPSS Statistics 12.0), longer than 64 characters (for SPSS Statistics 12.0 and subsequent versions), or contain invalid characters, it is necessary to rename or truncate them before reading them into SPSS Statistics. For more information, see the topic [Renaming or Filtering Fields for IBM SPSS Statistics](#) in Chapter 8 on p. 410.
- If SPSS Statistics was installed after SPSS Modeler, you may need to specify the SPSS Statistics license location, as explained above.

Export Nodes

Overview of Export Nodes

Export nodes provide a mechanism for exporting data in various formats to interface with your other software tools.

The following export nodes are available:



The Database export node writes data to an ODBC-compliant relational data source. In order to write to an ODBC data source, the data source must exist and you must have write permission for it. For more information, see the topic [Database Export Node](#) on p. 372.



The Flat File export node outputs data to a delimited text file. It is useful for exporting data that can be read by other analysis or spreadsheet software. For more information, see the topic [Flat File Export Node](#) on p. 382.



The Statistics Export node outputs data in IBM® SPSS® Statistics *.sav* format. The *.sav* files can be read by SPSS Statistics Base and other products. This is also the format used for cache files in IBM® SPSS® Modeler. For more information, see the topic [Statistics Export Node](#) in Chapter 8 on p. 409.



The IBM® SPSS® Data Collection export node outputs data in the format used by Data Collection market research software. The Data Collection Data Library must be installed to use this node. For more information, see the topic [IBM SPSS Data Collection Export Node](#) on p. 383.



The SAS export node outputs data in SAS format, to be read into SAS or a SAS-compatible software package. Three SAS file formats are available: SAS for Windows/OS2, SAS for UNIX, or SAS Version 7/8. For more information, see the topic [SAS Export Node](#) on p. 388.



The Excel export node outputs data in Microsoft Excel format (*.xls*). Optionally, you can choose to launch Excel automatically and open the exported file when the node is executed. For more information, see the topic [Excel Export Node](#) on p. 389.



The XML export node outputs data to a file in XML format. You can optionally create an XML source node to read the exported data back into the stream. For more information, see the topic [XML Export Node](#) on p. 391.

Database Export Node

You can use Database nodes to write data to ODBC-compliant relational data sources, which are explained in the description of the Database source node. For more information, see the topic [Database Source Node](#) in Chapter 2 on p. 11.

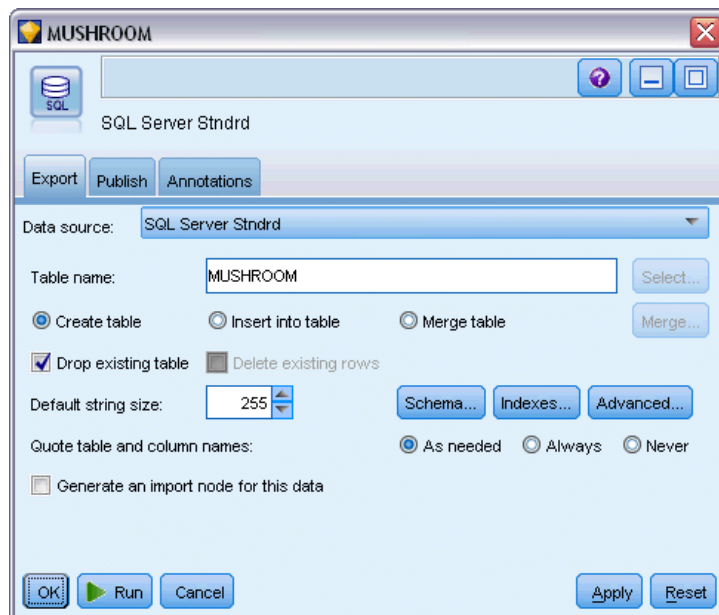
Use the following general steps to write data to a database:

- ▶ Install an ODBC driver and configure a data source to the database you want to use.
- ▶ On the Database node Export tab, specify the data source and table you want to write to. You can create a new table or insert data into an existing one.
- ▶ Specify additional options as needed.

These steps are described in more detail in the next several topics.

Database Node Export Tab

Figure 7-1
Database export node, Export tab



Data source. Shows the selected data source. Enter the name or select it from the drop-down list. If you don't see the desired database in the list, select [Add new database connection](#) and locate your database from the Database Connections dialog box. For more information, see the topic [Adding a Database Connection](#) in Chapter 2 on p. 14.

Table name. Enter the name of the table to which you want to send the data. If you select the Insert into table option, you can select an existing table in the database by clicking the Select button.

Create table. Select this option to create a new database table or to overwrite an existing database table.

Insert into table. Select this option to insert the data as new rows in an existing database table.

Merge table. (Where available) Select this option to update selected database columns with values from corresponding source data fields. Selecting this option enables the Merge button, which displays a dialog from where you can map source data fields to database columns.

Drop existing table. Select this option to delete any existing table with the same name when creating a new table.

Delete existing rows. Select this option to delete existing rows from the table before exporting, when inserting into a table.

Note: If either of the two options above are selected, you will receive an Overwrite warning message when you execute the node. To suppress the warnings, deselect Warn when a node overwrites a database table on the Notifications tab of the User Options dialog box.

Default string size. Fields you have marked as typeless in an upstream Type node are written to the database as string fields. Specify the size of strings to be used for typeless fields.

Click Schema to open a dialog box where you can set SQL data types for your fields, and specify the primary key for purposes of database indexing. For more information, see the topic [Database Export Schema Options](#) on p. 375.

Click Indexes to specify options for indexing the exported table in order to improve database performance. For more information, see the topic [Database Export Schema Options](#) on p. 375.

Click Advanced to specify bulk loading and database commit options. For more information, see the topic [Database Export Advanced Options](#) on p. 380.

Quote table and column names. Select options used when sending a CREATE TABLE statement to the database. Tables or columns with spaces or nonstandard characters must be quoted.

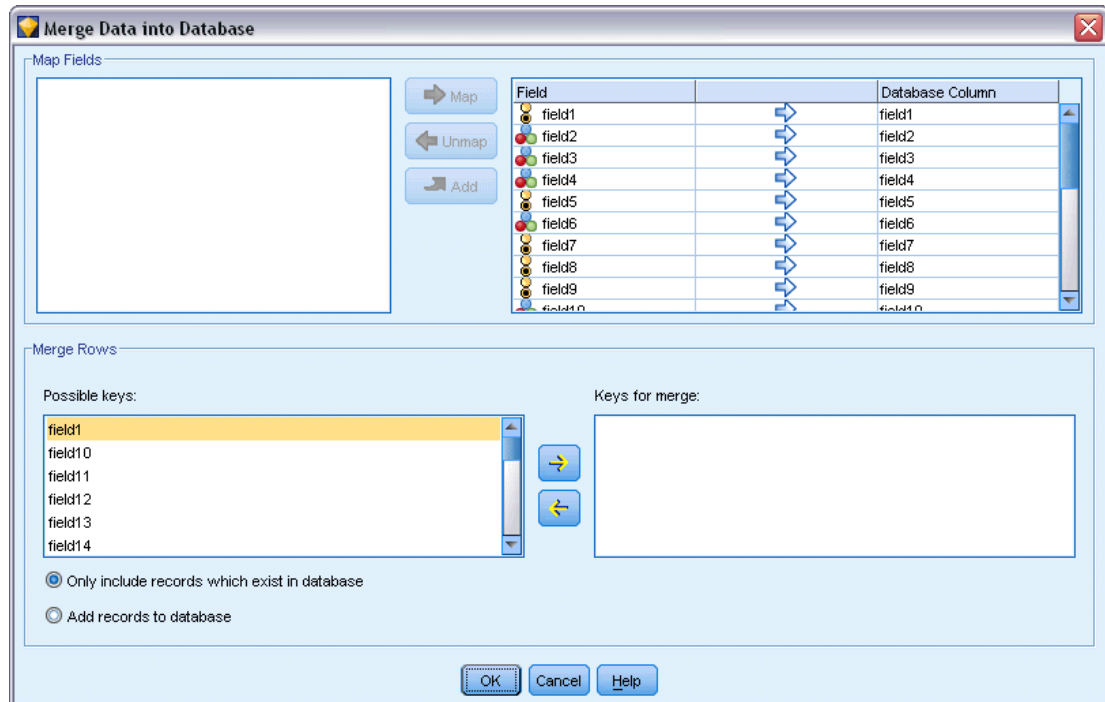
- **As needed.** Select to allow IBM® SPSS® Modeler to automatically determine when quoting is needed on an individual basis.
- **Always.** Select to always enclose table and column names in quotes.
- **Never.** Select to disable the use of quotes.

Generate an import node for this data. Select to generate a Database source node for the data as exported to the specified data source and table. Upon execution, this node is added to the stream canvas.

Database Export Merge Options

This dialog enables you to map fields from the source data onto columns in the target database table. Where a source data field is mapped to a database column, the column value is replaced with the source data value when the stream is run. Unmapped source fields are left unchanged in the database.

Figure 7-2
Mapping source data fields to database columns



Map Fields. This is where you specify the mapping between source data fields and database columns. Source data fields with the same name as columns in the database are mapped automatically.

- **Map.** Maps a source data field selected in the field list on the left of the button to a database column selected in the list on the right. You can map more than one field at a time, but the number of entries selected in both lists must be the same.
- **Unmap.** Removes the mapping for one or more selected database columns. This button is activated when you select a field or database column in the table on the right of the dialog.
- **Add.** Adds one or more source data fields selected in the field list on the left of the button to the list on the right ready for mapping. This button is activated when you select a field in the list on the left and no field with that name exists in the list on the right. Clicking this button maps the selected field to a new database column with the same name. The word <NEW> is displayed after the database column name to indicate that this is a new field.

Merge Rows. You use a key field, such as *Transaction ID*, to merge records with the same value in the key field. This is equivalent to a database “equi-join.” Key values must be those of primary keys; that is, they must be unique, and cannot contain null values.

- **Possible keys.** Lists all fields found in all input data sources. Select one or more fields from this list and use the arrow button to add them as key fields for merging records. Any map field with a corresponding mapped database column is available as a key, except that fields added as new database columns (shown with <NEW> after the name) are not available.

- **Keys for merge.** Lists all fields used to merge records from all input data sources based on values of the key fields. To remove a key from the list, select one and use the arrow button to return it to the Possible Keys list. When more than one key field is selected, the option below is enabled.
- **Only include records which exist in database.** Performs a partial join; if the record is in the database and the stream, the mapped fields will be updated.
- **Add records to database.** Performs an outer join; all records in the stream will be merged (if the same record exists in the database) or added (if the record does not yet exist in the database).

To map a source data field to a new database column

- ▶ Click the source field name in the list on the left, under Map Fields.
- ▶ Click the Add button to complete the mapping.

To map a source data field to an existing database column

- ▶ Click the source field name in the list on the left, under Map Fields.
- ▶ Click the column name under Database Column on the right.
- ▶ Click the Map button to complete the mapping.

To remove a mapping

- ▶ In the list on the right, under Field, click the name of the field for which you want to remove the mapping.
- ▶ Click the Unmap button.

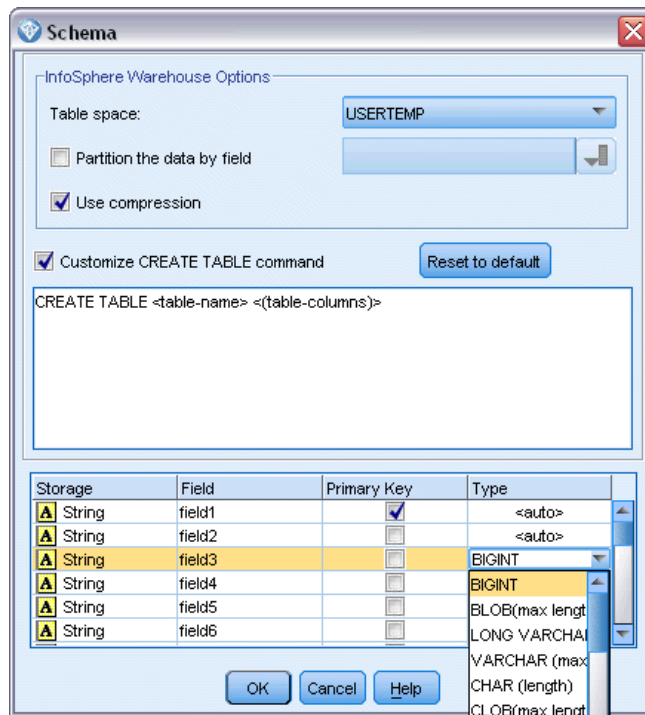
To deselect a field in any of the lists

- ▶ Hold down the CTRL key and click the field name.

Database Export Schema Options

The database export Schema dialog box allows you to set options for export to an InfoSphere Warehouse database, set SQL data types for your fields, specify which fields are primary keys, and customize the CREATE TABLE statement generated upon export.

Figure 7-3
Database export Schema dialog box



The dialog box has several parts:

- The section at the top contains options for export to an InfoSphere Warehouse database (IBM DB2 9.7 or later). This section is not displayed if you are not connected to such a database.
- The text field in the center displays the template used to generate the `CREATE TABLE` command, which by default follows the format:
`CREATE TABLE <table-name> <(table columns)>`
- The table in the lower portion allows you to specify the SQL data type for each field and to indicate which fields are primary keys as discussed below. The dialog box automatically generates the values of the `<table-name>` and `<(table columns)>` parameters based on the specifications in the table.

Setting InfoSphere Warehouse options

You can specify a number of settings for export to an InfoSphere Warehouse database.

Table space. The tablespace to be used for export. Database administrators can create or configure tablespaces as partitioned. We recommend selecting one of these tablespaces (rather than the default one) to use for database export.

Partition the data by field. Specifies the input field to be used for partitioning.

Use compression. If selected, creates tables for export with compression (for example, the equivalent of `CREATE TABLE MYTABLE(...) COMPRESS YES` in SQL).

Customizing CREATE TABLE statements

Using the top portion of this dialog box, you can add extra database-specific options to the CREATE TABLE statement.

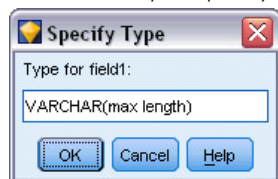
- ▶ Select the Customize CREATE TABLE command check box to activate the text window.
- ▶ Add any database-specific options to the statement. Be sure to retain the text <table-name> and (<table-columns>) parameters because these are substituted for the real table name and column definitions by IBM® SPSS® Modeler.

Setting SQL data types

By default, SPSS Modeler allows the database server to assign SQL data types automatically. To override the automatic type for a field, find the row corresponding to the field and select the desired type from the drop-down list in the *Type* column of the schema table. You can use Shift-click to select more than one row.

For types that take a length, precision, or scale argument (BINARY, VARBINARY, CHAR, VARCHAR, NUMERIC, and NUMBER), you should specify a length rather than allow the database server to assign an automatic length. For example, specifying a sensible value, such as VARCHAR(25), for length ensures that the storage type in SPSS Modeler will be overwritten if that is your intention. To override the automatic assignment, select Specify from the Type drop-down list and replace the type definition with the desired SQL type definition statement.

Figure 7-4
Database output Specify Type dialog box



The easiest way to do this is to first select the type that is closest to the desired type definition and then select Specify to edit that definition. For example, to set the SQL data type to VARCHAR(25), first set the type to VARCHAR(length) from the Type drop-down list, and then select Specify and replace the text length with the value 25.

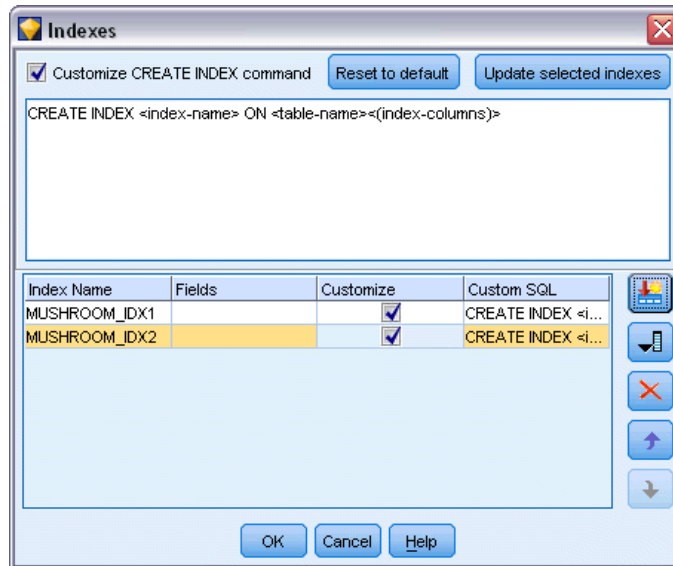
Primary keys

If one or more columns in the exported table must have a unique value or combination of values for every row, you can indicate this by selecting the Primary Key check box for each field that applies. Most databases will not allow the table to be modified in a manner that invalidates a primary key constraint and will automatically create an index over the primary key to help enforce this restriction. (Optionally, you can create indexes for other fields in the Indexes dialog box. For more information, see the topic [Database Export Index Options](#) on p. 378.)

Database Export Index Options

The Indexes dialog box allows you to create indexes on database tables exported from IBM® SPSS® Modeler. You can specify the field sets you want to include and customize the **CREATE INDEX** command, as needed.

Figure 7-5
Database output Indexes dialog box



The dialog box has two parts:

- The text field at the top displays a template that can be used to generate one or more **CREATE INDEX** commands, which by default follows the format:

```
CREATE INDEX <index-name> ON <table-name>
```

- The table in the lower portion of the dialog box allows you to add specifications for each index you want to create. For each index, specify the index name and the fields or columns to include. The dialog box automatically generates the values of the <index-name> and <table-name> parameters accordingly.

For example, the generated SQL for a single index on the fields *empid* and *deptid* might look like this:

```
CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID)
```

You can add multiple rows to create multiple indexes. A separate **CREATE INDEX** command is generated for each row.

Customizing the **CREATE INDEX** Command

Optionally, you can customize the **CREATE INDEX** command for all indexes or for a specific index only. This gives you the flexibility to accommodate specific database requirements or options and to apply customizations to all indexes or only specific ones, as needed.

- Select Customize CREATE INDEX command at the top of the dialog box to modify the template used for all indexes added subsequently. Note that changes will not automatically apply to indexes that have already been added to the table.
- Select one or more rows in the table and then click Update selected indexes at the top of the dialog box to apply the current customizations to all selected rows.
- Select the Customize check box in each row to modify the command template for that index only.

Note that the values of the <index-name> and <table-name> parameters are generated automatically by the dialog box based on the table specifications and cannot be edited directly.

BITMAP KEYWORD. If you are using an Oracle database, you can customize the template to create a bitmap index rather than a standard index, as follows:

```
CREATE BITMAP INDEX <index-name> ON <table-name>
```

Bitmap indexes may be useful for indexing columns with a small number of distinct values. The resulting SQL might look this:

```
CREATE BITMAP INDEX MYTABLE_IDX1 ON MYTABLE(COLOR)
```

UNIQUE keyword. Most databases support the UNIQUE keyword in the CREATE INDEX command. This enforces a uniqueness constraint similar to a primary key constraint on the underlying table.

```
CREATE UNIQUE INDEX <index-name> ON <table-name>
```

Note that for fields actually designated as primary keys, this specification is not necessary. Most databases will automatically create an index for any fields specified as primary key fields within the CREATE TABLE command, so explicitly creating indexes on these fields is not necessary. For more information, see the topic [Database Export Schema Options](#) on p. 375.

FILLFACTOR keyword. Some physical parameters for the index can be fine-tuned. For example, SQL Server allows the user to trade off the index size (after initial creation) against the costs of maintenance as future changes are made to the table.

```
CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID) WITH FILLFACTOR=20
```

Other Comments

- If an index already exists with the specified name, index creation will fail. Any failures will initially be treated as warnings, allowing subsequent indexes to be created and then re-reported as an error in the message log after all indexes have been attempted.
- For best performance, indexes should be created after data has been loaded into the table. Indexes must contain at least one column.
- Before executing the node, you can preview the generated SQL in the message log.
- For temporary tables written to the database (that is, when node caching is enabled) the options to specify primary keys and indexes are not available. However the system may create indexes on the temporary table as appropriate, depending on how the data is used in downstream nodes. For example, if cached data is subsequently joined by a *DEPT* column, it would make sense to index the cached tabled on this column.

Indexes and Query Optimization

In some database management systems, once a database table has been created, loaded, and indexed, a further step is required before the optimizer is able to utilize the indexes to speed up query execution on the new table. For example, in Oracle, the cost-based query optimizer requires that a table be analyzed before its indexes can be used in query optimization. The internal ODBC properties file for Oracle (not user-visible) contains an option to make this happen, as follows:

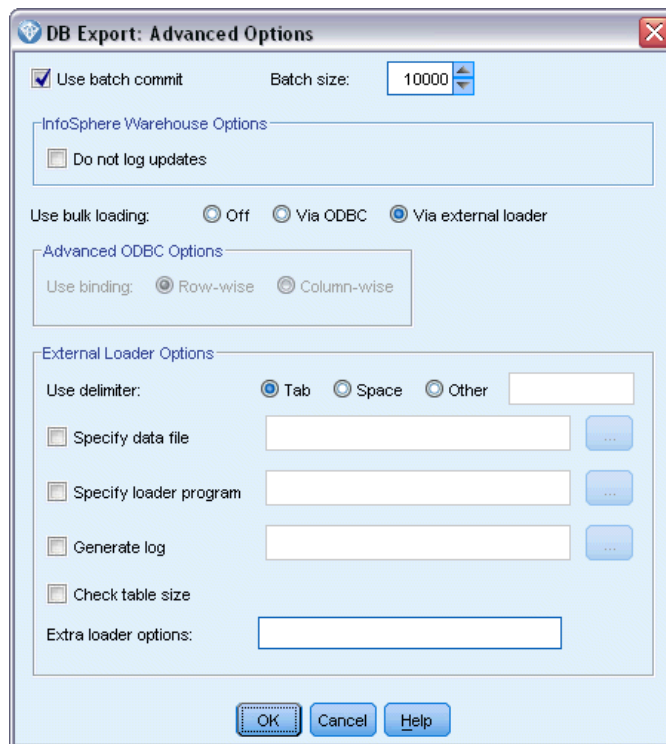
```
# Defines SQL to be executed after a table and any associated indexes
# have been created and populated
table_analysis_sql, 'ANALYZE TABLE <table-name> COMPUTE STATISTICS'
```

This step is executed whenever a table is created in Oracle (regardless of whether primary keys or indexes are defined). If necessary, the ODBC properties file for additional databases can be customized in a similar way - contact Support for assistance.

Database Export Advanced Options

When you click the Advanced button from the Database export node dialog box, a new dialog box opens in which you can specify technical details for exporting results to a database.

Figure 7-6
Specifying advanced options for database export



Use batch commit. Select to turn off row-by-row commits to the database.

Batch size. Specifies the number of records to send to the database before committing to memory. Lowering this number provides greater data integrity at the cost of slower transfer speeds. You may want to fine-tune this number for optimal performance with your database.

InfoSphere Warehouse options. Displayed only if you are connected to an InfoSphere Warehouse database (IBM DB2 9.7 or later). Do not log updates enables you to avoid logging events when creating tables and inserting data.

Use bulk loading. Specifies a method for bulk loading data to the database directly from IBM® SPSS® Modeler. Some experimentation may be required to select which bulk load options are appropriate for a particular scenario.

- **Via ODBC.** Select to use the ODBC API to execute multiple-row inserts with greater efficiency than normal export to the database. Choose from row-wise or column-wise binding in the options below.
- **Via external loader.** Select to use a custom bulk loader program specific to your database. Selecting this option activates a variety of options below.

Advanced ODBC Options. These options are available only when Via ODBC is selected. Note that this functionality may not be supported by all ODBC drivers.

- **Row-wise.** Select row-wise binding to use the `SQLBulkOperations` call for loading data into the database. Row-wise binding typically improves speed compared to the use of parameterized inserts that insert data on a record-by-record basis.
- **Column-wise.** Select to use column-wise binding for loading data into the database. Column-wise binding improves performance by binding each database column (in a parameterized `INSERT` statement) to an array of N values. Executing the `INSERT` statement once causes N rows to be inserted into the database. This method can dramatically increase performance.

External Loader Options. When Via external loader is specified, a variety of options are displayed for exporting the dataset to a file and specifying and executing a custom loader program to load the data from that file into the database. SPSS Modeler can interface with external loaders for many popular database systems. Several scripts have been included with the software and are available along with technical documentation under the *scripts* subdirectory. Note that in order to use this functionality, Python 2.7 must be installed on the same machine as SPSS Modeler or IBM® SPSS® Modeler Server, and the `python_exe_path` parameter must be set in the *options.cfg* file.

- **Use delimiter.** Specifies which delimiter character should be used in the exported file. Select Tab to delimit with tab and Space to delimit with spaces. Select Other to specify another character, such as a comma (,).
- **Specify data file.** Select to enter the path to use for the data file written during bulk loading. By default, a temporary file is created in the temp directory on the server.
- **Specify loader program.** Select to specify a bulk loading program. By default, the software searches the *scripts* subdirectory of the SPSS Modeler installation for a Python script to execute for a given database. Several scripts have been included with the software and are available along with technical documentation under the *scripts* subdirectory.
- **Generate log.** Select to generate a log file to the specified directory. The log file contains error information and is useful if the bulk load operation fails.

- **Check table size.** Select to perform table checking that ensures that the increase in table size corresponds to the number of rows exported from SPSS Modeler.
- **Extra loader options.** Specifies additional arguments to the loader program. Use double quotes for arguments containing spaces.

Double quotes are included in optional arguments by escaping with a backslash. For example, the option specified as `-comment "This is a \"comment\""` includes both the `-comment` flag and the comment itself rendered as `This is a "comment"`.

A single backslash can be included by escaping with another backslash. For example, the option specified as `-specialdir "C:\\Test Scripts\\"` includes the flag `-specialdir` and the directory rendered as `C:\Test Scripts\`.

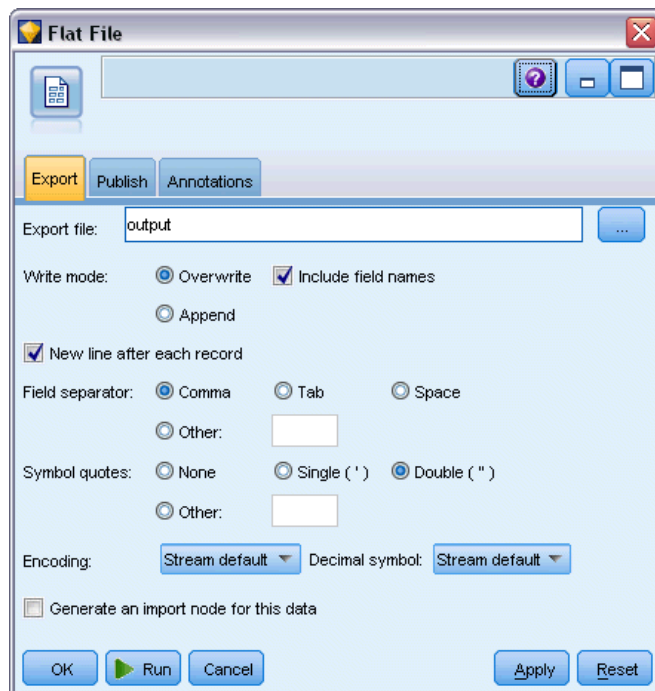
Flat File Export Node

The Flat File export node allows you to write data to a delimited text file. This is useful for exporting data that can be read by other analysis or spreadsheet software.

Note: You cannot write files in the old cache format, because IBM® SPSS® Modeler no longer uses that format for cache files. SPSS Modeler cache files are now saved in IBM® SPSS® Statistics.sav format, which you can write using a Statistics export node. For more information, see the topic [Statistics Export Node](#) in Chapter 8 on p. 409.

Flat File Export Tab

Figure 7-7
Flat File node, Export tab



Export file. Specifies the name of the file. Enter a filename or click the file chooser button to browse to the file's location.

Write mode. If Overwrite is selected, any existing data in the specified file will be overwritten. If Append is selected, output will be added to the end of the existing file, preserving any data it contains.

- **Include field names.** If this option is selected, field names will be written to the first line of the output file. This option is available only for the Overwrite write mode.

New line after each record. If this option is selected, each record will be written on a new line in the output file.

Field separator. Specifies the character to insert between field values in the generated text file. Options are Comma, Tab, Space, and Other. If you select Other, enter the desired delimiter character(s) in the text box.

Symbol quotes. Specifies the type of quoting to use for values of symbolic fields. Options are None (values are not quoted), Single ('), Double ("), and Other. If you select Other, enter the desired quoting character(s) in the text box.

Encoding. Specifies the text-encoding method used. You can choose between the system default, stream default, or UTF-8.

- The system default is specified in the Windows Control Panel or, if running in distributed mode, on the server computer.
- The stream default is specified in the Stream Properties dialog box.

Decimal symbol. Specifies how decimals should be represented in the data.

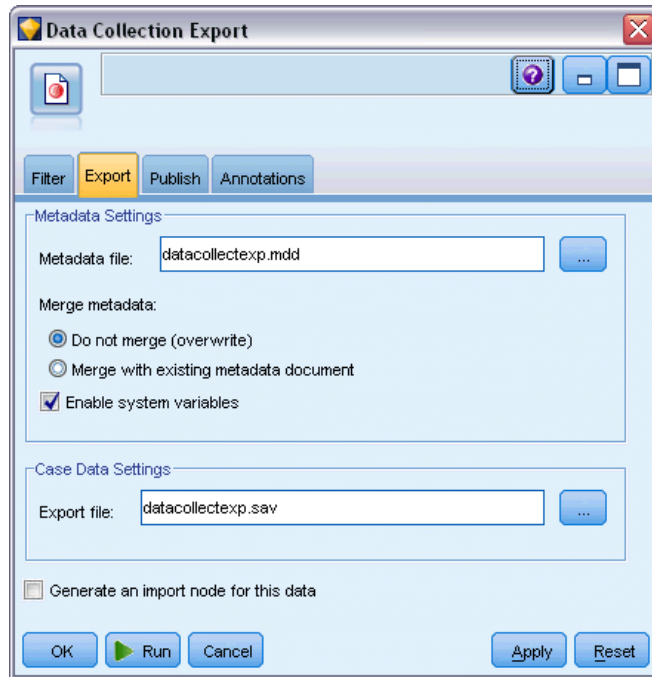
- **Stream default.** The decimal separator defined by the current stream's default setting will be used. This will normally be the decimal separator defined by the computer's locale settings.
- **Period (.).** The period character will be used as the decimal separator.
- **Comma (,).** The comma character will be used as the decimal separator.

Generate an import node for this data. Select this option to automatically generate a Variable File source node that will read the exported data file. For more information, see the topic [Variable File Node](#) in Chapter 2 on p. 18.

IBM SPSS Data Collection Export Node

The IBM® SPSS® Data Collection export node saves data in the format used by Data Collection market research software, based on the Data Collection Data Model. This format distinguishes case data—the actual responses to questions gathered during a survey—from the metadata that describes how the case data is collected and organized. Metadata consists of information such as question texts, variable names and descriptions, multiple-response sets, translations of the various texts, and definitions of the structure of the case data. For more information, see the topic [Data Collection Node](#) in Chapter 2 on p. 26.

Figure 7-8
IBM SPSS Data Collection export node, Export tab



Note: This node requires Data Collection Data Model version 4.0 or higher, which is distributed with Data Collection software. For more information, see the Data Collection Web page at <http://www.ibm.com/software/analytics/spss/products/data-collection/>. Aside from installing the Data Model, no additional configuration is required.

Metadata file. Specifies the name of the questionnaire definition file (*.mdd*) where the exported metadata will be saved. A default questionnaire is created based on field type information. For example, a nominal (set) field could be represented as a single question with the field description used as the question text and a separate check box for each defined value.

Merge metadata. Specifies whether the metadata will overwrite existing versions or be merged with existing metadata. If the merge option is selected, a new version is created each time the stream is run. This makes it possible to track versions of a questionnaire as it undergoes changes. Each version can be regarded as a snapshot of the metadata used to collect a particular set of case data.

Enable system variables. Specifies whether system variables are included in the exported *.mdd* file. These include variables such as *Respondent.Serial*, *Respondent.Origin*, and *DataCollection.StartTime*.

Case data settings. Specifies the IBM® SPSS® Statistics data (*.sav*) file where case data is exported. Note that all the restrictions on variable and value names apply here, so for example you may need to switch to the Filter tab and use the “Rename for SPSS Statistics” option on the Filter options menu to correct invalid characters in field names.

Generate an import node for this data. Select this option to automatically generate a Data Collection source node that will read the exported data file.

Multiple response sets. Any multiple response sets defined in the stream will automatically be preserved when the file is exported. You can view and edit multiple response sets from any node with a Filter tab. For more information, see the topic [Editing Multiple Response Sets](#) in Chapter 4 on p. 134.

IBM Cognos BI Export Node

The IBM Cognos BI Export node enables you to export data from an IBM® SPSS® Modeler stream to Cognos BI, in UTF-8 format. In this way, Cognos BI can make use of transformed or scored data from SPSS Modeler. For example, you could use Cognos BI Report Studio to create a report based on the exported data, including the predictions and confidence values. The report could then be saved on the Cognos BI server and distributed to Cognos BI users.

Note: You can export only relational data, not OLAP data.

To export data to Cognos BI, you need to specify the following:

- Cognos connection - the connection to the Cognos BI server
- ODBC connection - the connection to the Cognos data server that the Cognos BI server uses

The connections must point to the same database and must use the same username for the connection.

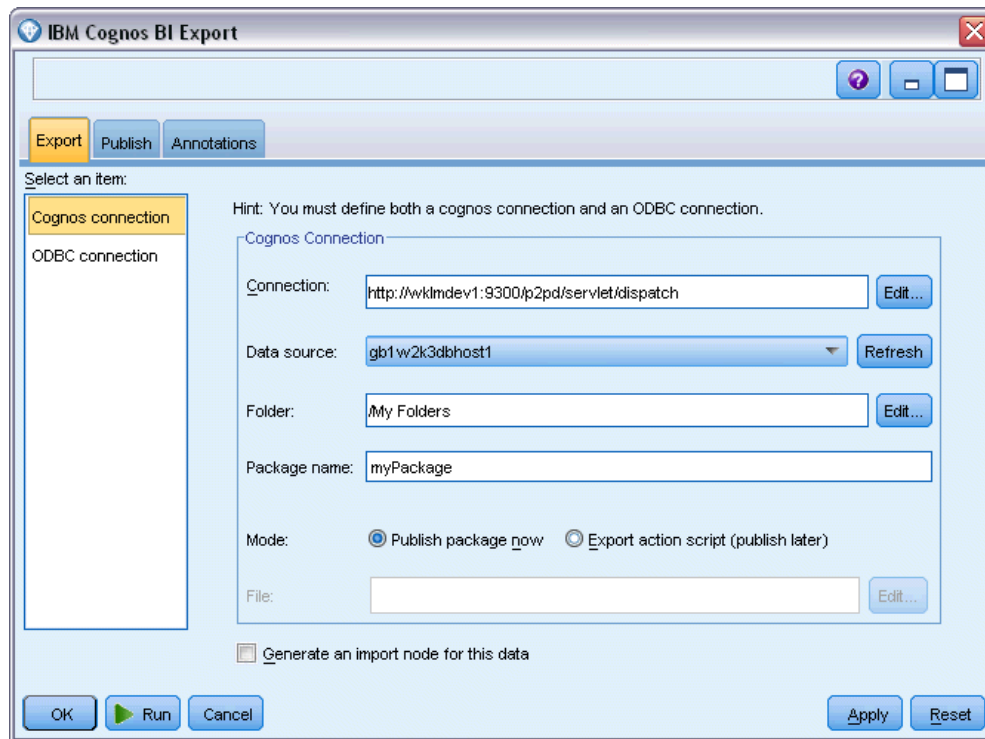
You export the actual stream data to the data server, and the package metadata to the Cognos BI server.

As with any other export node, you can also use the Publish tab of the node dialog box to publish the stream for deployment using IBM® SPSS® Modeler Solution Publisher.

Cognos connection

This is where you specify the connection to the Cognos BI server that you want to use for the export. The procedure involves exporting the metadata to a new package on the Cognos BI server, while the stream data is exported to the Cognos data server.

Figure 7-9
Exporting Cognos data



Connection. Click the Edit button to display a dialog box where you can define the URL and other details of the Cognos BI server to which you want to export the data. If you are already logged in to a Cognos BI server through IBM® SPSS® Modeler, you can also edit the details of the current connection. For more information, see the topic [Cognos connections](#) in Chapter 2 on p. 36.

Data source. The name of the Cognos data source (typically a database) to which you are exporting the data. The drop-down list shows all the Cognos data sources that you can access from the current connection. Click the Refresh button to update the list.

Folder. The path and name of the folder on the Cognos BI server where the export package is to be created.

Package name. The name of the package in the specified folder that is to contain the exported metadata. This must be a new package with a single query subject; you cannot export to an existing package.

Mode. Specifies how you want to perform the export:

- **Publish package now.** (default) Performs the export operation as soon as you click Run.
- **Export action script.** Creates an XML script that you can run later (for example, using Framework Manager) to perform the export. Type the path and file name for the script in the File field, or use the Edit button to specify the name and location of the script file.

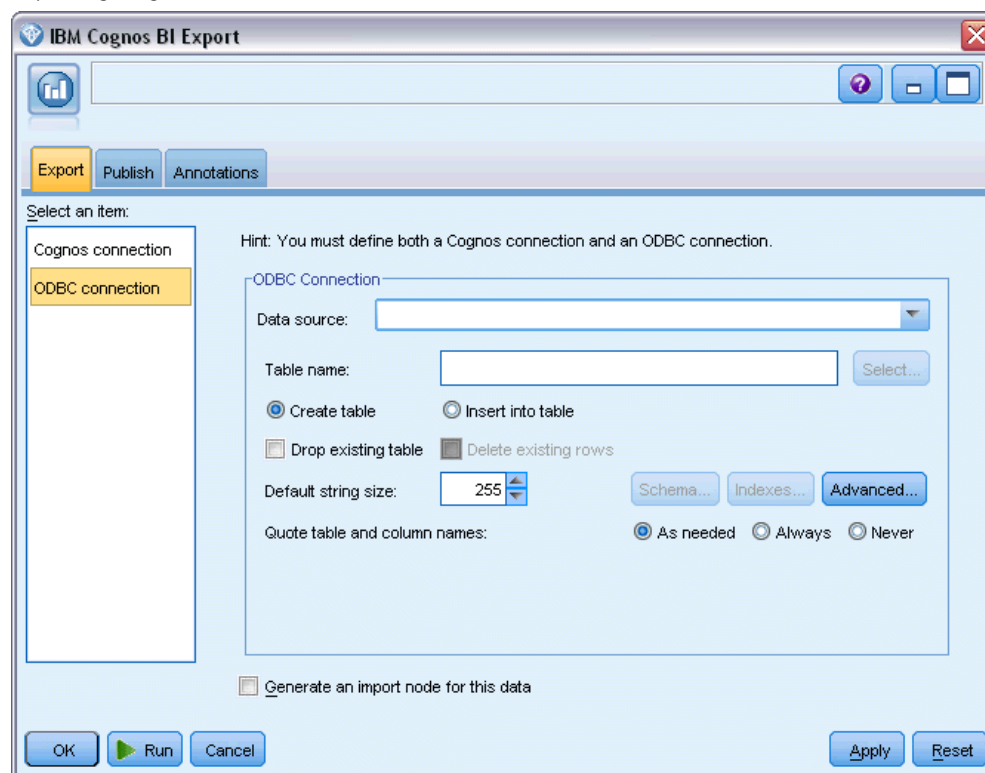
Generate an import node for this data. Select to generate a source node for the data as exported to the specified data source and table. When you click Run, this node is added to the stream canvas.

ODBC connection

Here you specify the connection to the Cognos data server (that is, the database) to which the stream data is to be exported.

Note: You must ensure that the data source you specify here points to the same one specified on the Cognos connections panel. You must also ensure that the same username is used for the Cognos and ODBC connections.

Figure 7-10
Exporting Cognos data



Data source. Shows the selected data source. Enter the name or select it from the drop-down list. If you don't see the desired database in the list, select [Add new database connection](#) and locate your database from the Database Connections dialog box. For more information, see the topic [Adding a Database Connection](#) in Chapter 2 on p. 14.

Table name. Enter the name of the table to which you want to send the data. If you select the Insert into table option, you can select an existing table in the database by clicking the Select button.

Create table. Select this option to create a new database table or to overwrite an existing database table.

Insert into table. Select this option to insert the data as new rows in an existing database table.

Merge table. (Where available) Select this option to update selected database columns with values from corresponding source data fields. Selecting this option enables the Merge button, which displays a dialog from where you can map source data fields to database columns.

Drop existing table. Select this option to delete any existing table with the same name when creating a new table.

Delete existing rows. Select this option to delete existing rows from the table before exporting, when inserting into a table.

Note: If either of the two options above are selected, you will receive an Overwrite warning message when you execute the node. To suppress the warnings, deselect Warn when a node overwrites a database table on the Notifications tab of the User Options dialog box.

Default string size. Fields you have marked as typeless in an upstream Type node are written to the database as string fields. Specify the size of strings to be used for typeless fields.

Click Schema to open a dialog box where you can set SQL data types for your fields, and specify the primary key for purposes of database indexing. For more information, see the topic [Database Export Schema Options](#) on p. 375.

Click Indexes to specify options for indexing the exported table in order to improve database performance. For more information, see the topic [Database Export Schema Options](#) on p. 375.

Click Advanced to specify bulk loading and database commit options. For more information, see the topic [Database Export Advanced Options](#) on p. 380.

Quote table and column names. Select options used when sending a CREATE TABLE statement to the database. Tables or columns with spaces or nonstandard characters must be quoted.

- **As needed.** Select to allow IBM® SPSS® Modeler to automatically determine when quoting is needed on an individual basis.
- **Always.** Select to always enclose table and column names in quotes.
- **Never.** Select to disable the use of quotes.

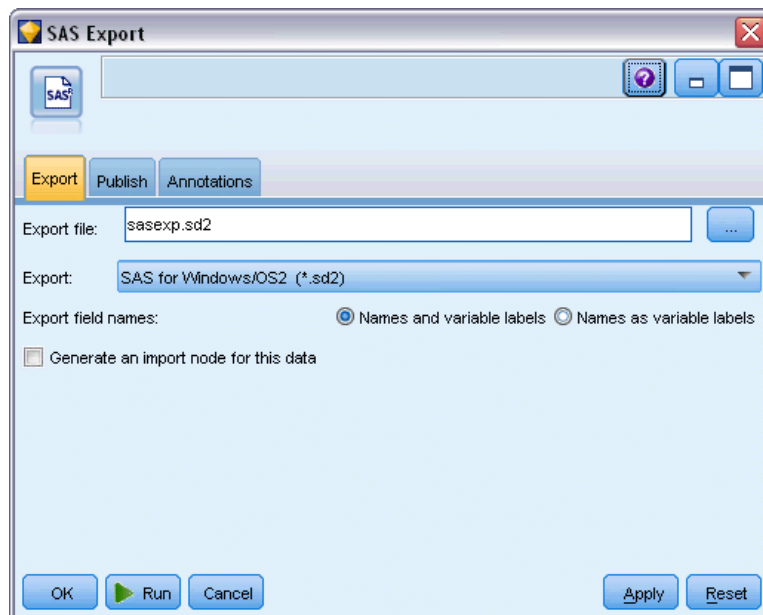
Generate an import node for this data. Select to generate a source node for the data as exported to the specified data source and table. When you click Run, this node is added to the stream canvas.

SAS Export Node

The SAS export node allows you to write data in SAS format to be read into SAS or a SAS-compatible software package. You can export in three SAS file formats: SAS for Windows/OS2, SAS for UNIX, or SAS Version 7/8.

SAS Export Node Export Tab

Figure 7-11
SAS export node, Export tab



Export file. Specify the name of the file. Enter a filename or click the file chooser button to browse to the file's location.

Export. Specify the export file format. Options are SAS for Windows/OS2, SAS for UNIX, or SAS Version 7/8.

Export field names. Select options for exporting field names and labels from IBM® SPSS® Modeler for use with SAS.

- **Names and variable labels.** Select to export both SPSS Modeler field names and field labels. Names are exported as SAS variable names, while labels are exported as SAS variable labels.
- **Names as variable labels.** Select to use the SPSS Modeler field names as variable labels in SAS. SPSS Modeler allows characters in field names that are invalid in SAS variable names. To prevent possibly creating invalid SAS names, select Names and variable labels instead.

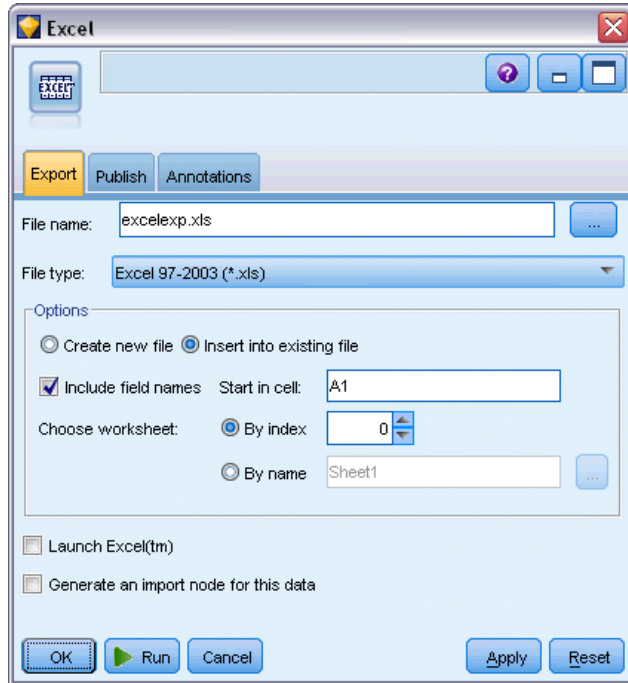
Generate an import node for this data. Select this option to automatically generate a SAS source node that will read the exported data file. For more information, see the topic [SAS Source Node](#) in Chapter 2 on p. 37.

Excel Export Node

The Excel export node outputs data in Microsoft Excel format (.xls). Optionally, you can choose to automatically launch Excel and open the exported file when the node is executed.

Excel Node Export Tab

Figure 7-12
Excel export node, Export tab



File name. Enter a filename or click the file chooser button to browse to the file's location. The default filename is *excelexp.xls*.

File type. Select the Excel file type that you are exporting.

Create new file. Creates a new Excel file.

Insert into existing file. Content is replaced beginning at the cell designated by the Start in cell field. Other cells in the spreadsheet are left with their original content.

Include field names. Specifies whether field names should be included in the first row of the worksheet.

Start in cell. The cell location used for the first export record (or first field name if Include field names is checked). Data are filled to the right and down from this initial cell.

Choose worksheet. Specifies the worksheet to which you want to export the data. You can identify the worksheet either by index or by name:

- **By index.** If you are creating a new file, specify a number from 0 to 9 to identify the worksheet to which you want to export, beginning with 0 for the first worksheet, 1 for the second worksheet, and so on. You can use values of 10 or higher only if a worksheet already exists at this position.
- **By name.** If you are creating a new file, specify the name used for the worksheet. If you are inserting into an existing file, the data are inserted into this worksheet if it exists, otherwise a new worksheet with this name is created.

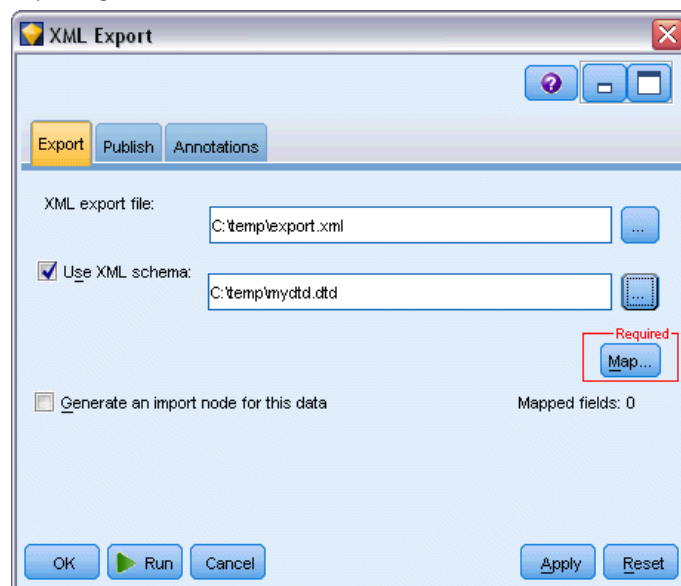
Launch Excel. Specifies whether Excel is automatically launched on the exported file when the node is executed. Note that when running in distributed mode against IBM® SPSS® Modeler Server, the output is saved to the server file system, and Excel is launched on the Client with a copy of the exported file.

Generate an import node for this data. Select this option to automatically generate an Excel source node that will read the exported data file. For more information, see the topic [Excel Source Node](#) in Chapter 2 on p. 39.

XML Export Node

The XML Export node enables you to output data in XML format, using UTF-8 encoding. You can optionally create an XML source node to read the exported data back into the stream.

Figure 7-13
Exporting XML data



XML export file. The full path and file name of the XML file to which you want to export the data.

Use XML schema. Select this check box if you want to use a schema or DTD to control the structure of the exported data. Doing so activates the Map button, described below.

If you do not use a schema or DTD, the following default structure is used for the exported data:

```
<records>
  <record>
    <fieldname1>value</fieldname1>
    <fieldname2>value</fieldname2>
    :
    <fieldnameN>value</fieldnameN>
  </record>
  <record>
  :
  :
  </record>
  :
  :
</records>
```

Spaces in a field name are replaced with underscores; for example, “My Field” becomes <My_Field>.

Map. If you have chosen to use an XML schema, this button opens a dialog where you can specify which part of the XML structure should be used to start each new record. For more information, see the topic [XML Mapping Records Options](#) on p. 393.

Mapped fields. Indicates the number of fields that have been mapped.

Generate an import node for this data. Select this option to automatically generate an XML source node that will read the exported data file back into the stream. For more information, see the topic [XML Source Node](#) in Chapter 2 on p. 40.

Writing XML Data

When an XML element is specified, the field value is placed inside the element tag:

```
<element>value</element>
```

When an attribute is mapped, the field value is placed as the value for the attribute:

```
<element attribute="value">
```

If a field is mapped to an element above the <records> element, the field is written only once, and will be a constant for all records. The value for this element will come from the first record.

If a null value is to be written, it is done by specifying empty content. For elements, this is:

```
<element></element>
```

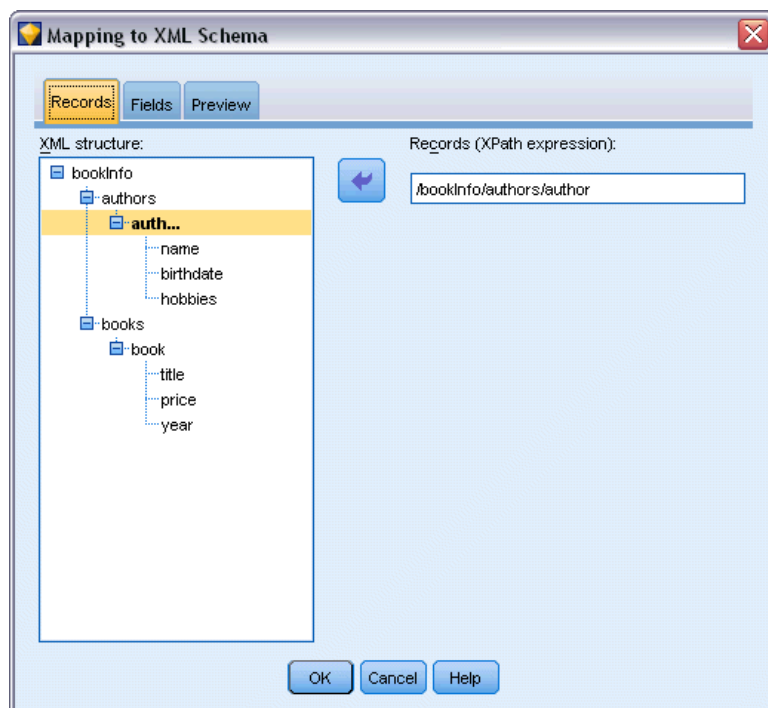
For attributes, it is:

```
<element attribute="">
```

XML Mapping Records Options

The Records tab enables you to specify which part of the XML structure to use to start each new record. In order to map correctly onto a schema, you need to specify the record delimiter.

Figure 7-14
XML mapping records



XML structure. A hierarchical tree showing the structure of the XML schema specified on the previous screen.

Records (XPath expression). To set the record delimiter, select an element in the XML structure and click the right-arrow button. Each time this element is encountered in the source data, a new record is created in the output file.

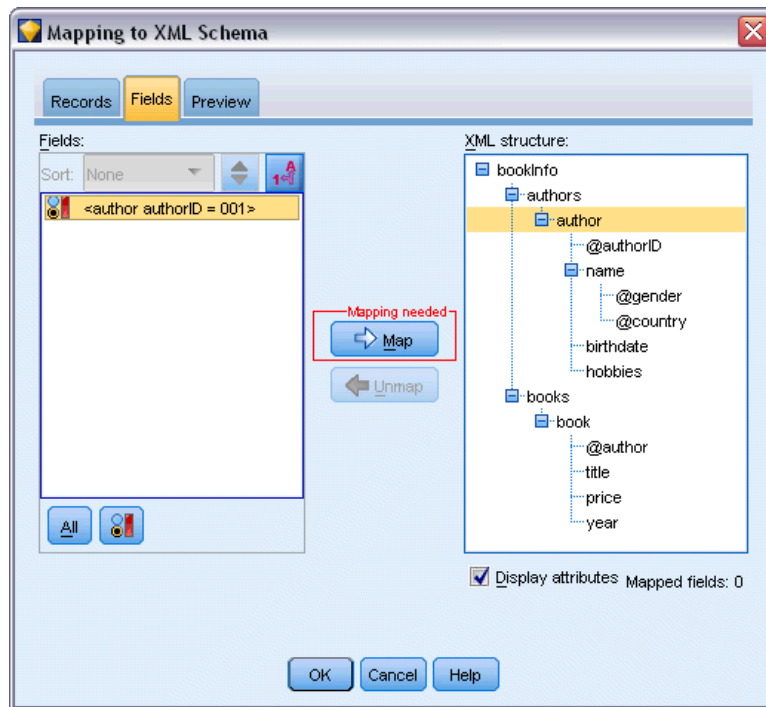
Note: If you select the root element in the XML structure, only a single record can be written, and all other records are skipped.

XML Mapping Fields Options

The Fields tab is used to map fields in the data set to elements or attributes in the XML structure when a schema file is used.

Field names that match an element or attribute name are automatically mapped as long as the element or attribute name is unique. Thus if there is both an element and an attribute with the name field1, there is no automatic mapping. If there is only one item in the structure named field1, a field with that name in the stream is automatically mapped.

Figure 7-15
XML mapping fields



Fields. The list of fields in the model. Select one or more fields as the source part of the mapping. You can use the buttons at the bottom of the list to select all fields, or all fields with a particular measurement level.

XML structure. Select an element in the XML structure as the map target. To create the mapping, click Map. The mapping is then displayed. The number of fields that have been mapped in this way is displayed below this list.

To remove a mapping, select the item in the XML structure list and click Unmap.

Display attributes. Displays or hides the attributes, if any, of the XML elements in the XML structure.

XML Mapping Preview

On the Preview tab, click Update to see a preview of the XML that will be written.

If the mapping is incorrect, return to the Records or Fields tab to correct the errors and click Update again to see the result.

IBM SPSS Statistics Nodes

IBM SPSS Statistics Nodes - Overview

To complement IBM® SPSS® Modeler and its data mining abilities, IBM® SPSS® Statistics provides you with the ability to carry out further statistical analysis and data management.

If you have a compatible, licensed copy of SPSS Statistics installed, you can connect to it from SPSS Modeler and carry out complex, multistep data manipulation and analysis not otherwise supported by SPSS Modeler. For the advanced user there is also the option to further modify the analyses by using command syntax. See the Release Notes for information on version compatibility.

If available, the SPSS Statistics nodes are shown on a dedicated part of the nodes palette.

Note: We recommend that you instantiate your data in a Type node before using the SPSS Statistics Transform, Model, or Output nodes. This is also a requirement when using the AUTORECODE syntax command.

The SPSS Statistics palette contains the following nodes:



The Statistics File node reads data from the *.sav* file format used by SPSS Statistics, as well as cache files saved in SPSS Modeler, which also use the same format. For more information, see the topic [Statistics File Node](#) on p. 396.



The Statistics Transform node runs a selection of SPSS Statistics syntax commands against data sources in SPSS Modeler. This node requires a licensed copy of SPSS Statistics. For more information, see the topic [Statistics Transform Node](#) on p. 397.



The Statistics Model node enables you to analyze and work with your data by running SPSS Statistics procedures that produce PMML. This node requires a licensed copy of SPSS Statistics. For more information, see the topic [Statistics Model Node](#) on p. 401.



The Statistics Output node allows you to call an SPSS Statistics procedure to analyze your SPSS Modeler data. A wide variety of SPSS Statistics analytical procedures is available. This node requires a licensed copy of SPSS Statistics. For more information, see the topic [Statistics Output Node](#) on p. 405.

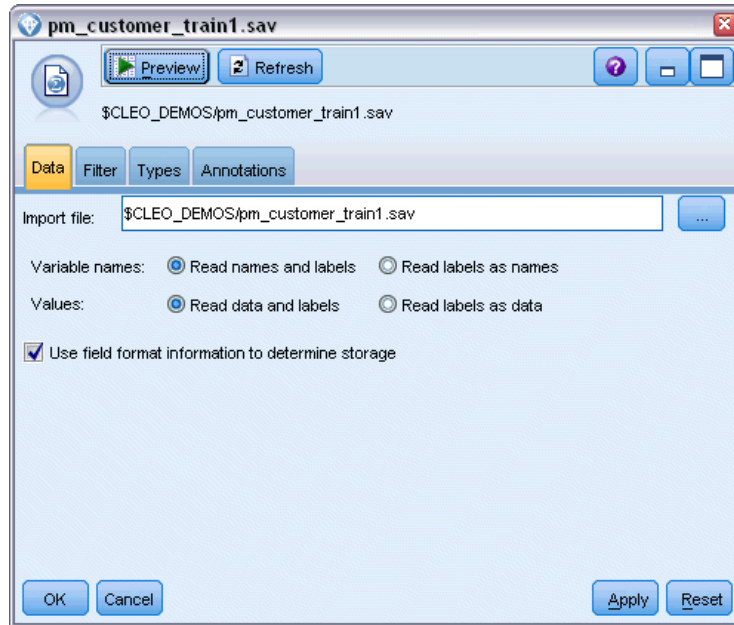


The Statistics Export node outputs data in SPSS Statistics *.sav* format. The *.sav* files can be read by SPSS Statistics Base and other products. This is also the format used for cache files in SPSS Modeler. For more information, see the topic [Statistics Export Node](#) on p. 409.

Statistics File Node

You can use the Statistics File node to read data directly from a saved IBM® SPSS® Statistics file (.sav). This format is now used to replace the cache file from earlier versions of IBM® SPSS® Modeler. If you would like to import a saved cache file, you should use the SPSS Statistics File node.

Figure 8-1
Importing a .sav file



Import file. Specify the name of the file. You can enter a filename or click the ellipsis button (...) to select a file. The file path is shown once you have selected a file.

Variable names. Select a method of handling variable names and labels upon import from an SPSS Statistics .sav file. Metadata that you choose to include here persists throughout your work in SPSS Modeler and may be exported again for use in SPSS Statistics.

- **Read names and labels.** Select to read both variable names and labels into SPSS Modeler. By default, this option is selected and variable names are displayed in the Type node. Labels may be displayed in charts, model browsers, and other types of output, depending on the options specified in the stream properties dialog box. By default, the display of labels in output is disabled.
- **Read labels as names.** Select to read the descriptive variable labels from the SPSS Statistics .sav file rather than the short field names, and use these labels as variable names in SPSS Modeler.

Values. Select a method of handling values and labels upon import from an SPSS Statistics .sav file. Metadata that you choose to include here persists throughout your work in SPSS Modeler and may be exported again for use in SPSS Statistics.

- **Read data and labels.** Select to read both actual values and value labels into SPSS Modeler. By default, this option is selected and values themselves are displayed in the Type node. Value labels may be displayed in the Expression Builder, charts, model browsers, and other types of output, depending on the options specified in the stream properties dialog box.
- **Read labels as data.** Select if you want to use the value labels from the *.sav* file rather than the numerical or symbolic codes used to represent the values. For example, selecting this option for data with a gender field whose values of 1 and 2 actually represent *male* and *female*, respectively, will convert the field to a string and import *male* and *female* as the actual values.

It is important to consider missing values in your SPSS Statistics data before selecting this option. For example, if a numeric field uses labels only for missing values (0 = *No Answer*, -99 = *Unknown*), then selecting the option above will import only the value labels *No Answer* and *Unknown* and will convert the field to a string. In such cases, you should import the values themselves and set missing values in a Type node.

Use field format information to determine storage. If this box is checked, field values that are formatted in the *.sav* file as integers (i.e., fields specified as *Fn.0* in the Variable View in SPSS Statistics) are imported using integer storage. All other field values except strings are imported as real numbers.

If this box is unchecked (default), all field values except strings are imported as real numbers, whether formatted in the *.sav* file as integers or not.

Multiple response sets. Any multiple response sets defined in the SPSS Statistics file will automatically be preserved when the file is imported. You can view and edit multiple response sets from any node with a Filter tab. For more information, see the topic [Editing Multiple Response Sets](#) in Chapter 4 on p. 134.

Statistics Transform Node

The Statistics Transform node allows you to complete data transformations using IBM® SPSS® Statistics command syntax. This makes it possible to complete a number of transformations not supported by IBM® SPSS® Modeler and allows automation of complex, multistep transformations, including the creation of a number of fields from a single node. It resembles the Statistics Output node, except that the data are returned to SPSS Modeler for further analysis, whereas, in the Output node the data are returned as the requested output objects, such as graphs or tables.

You must have a compatible version of SPSS Statistics installed and licensed on your computer to use this node. For more information, see the topic [IBM SPSS Statistics Helper Applications](#) in Chapter 6 on p. 369. See the Release Notes for compatibility information.

If necessary, you can use the Filter tab to filter or rename fields so they conform to SPSS Statistics naming standards. For more information, see the topic [Renaming or Filtering Fields for IBM SPSS Statistics](#) on p. 410.

Syntax Reference. For details on specific SPSS Statistics procedures, see the *SPSS Statistics Command Syntax Reference* guide, included with your copy of the SPSS Statistics software. To view the guide from the Syntax tab, choose the Syntax editor option and click the Launch SPSS Statistics Syntax Help button.

Note: Not all SPSS Statistics syntax is supported by this node. For more information, see the topic [Allowable Syntax](#) on p. 399.

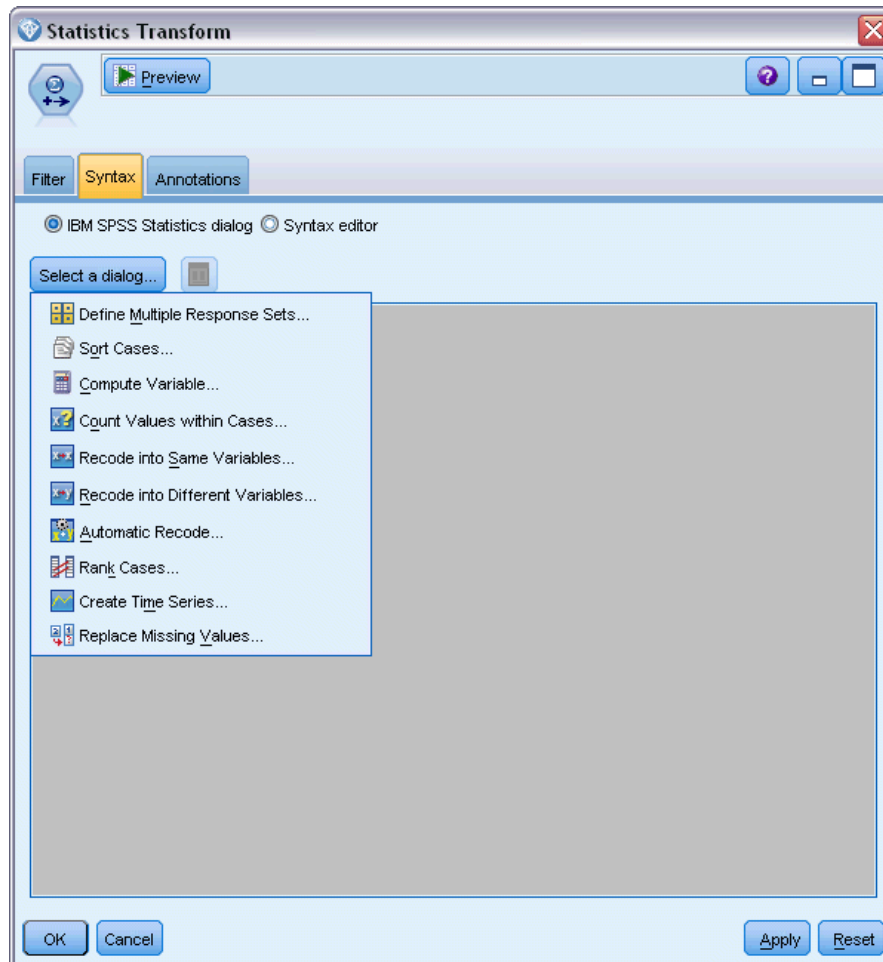
Statistics Transform Node - Syntax Tab

IBM SPSS Statistics dialog option

If you are unfamiliar with IBM® SPSS® Statistics syntax, the simplest way to create syntax in IBM® SPSS® Modeler is to first run the command in SPSS Statistics, copy the syntax into the Statistics Transform node in SPSS Modeler, and run the stream.

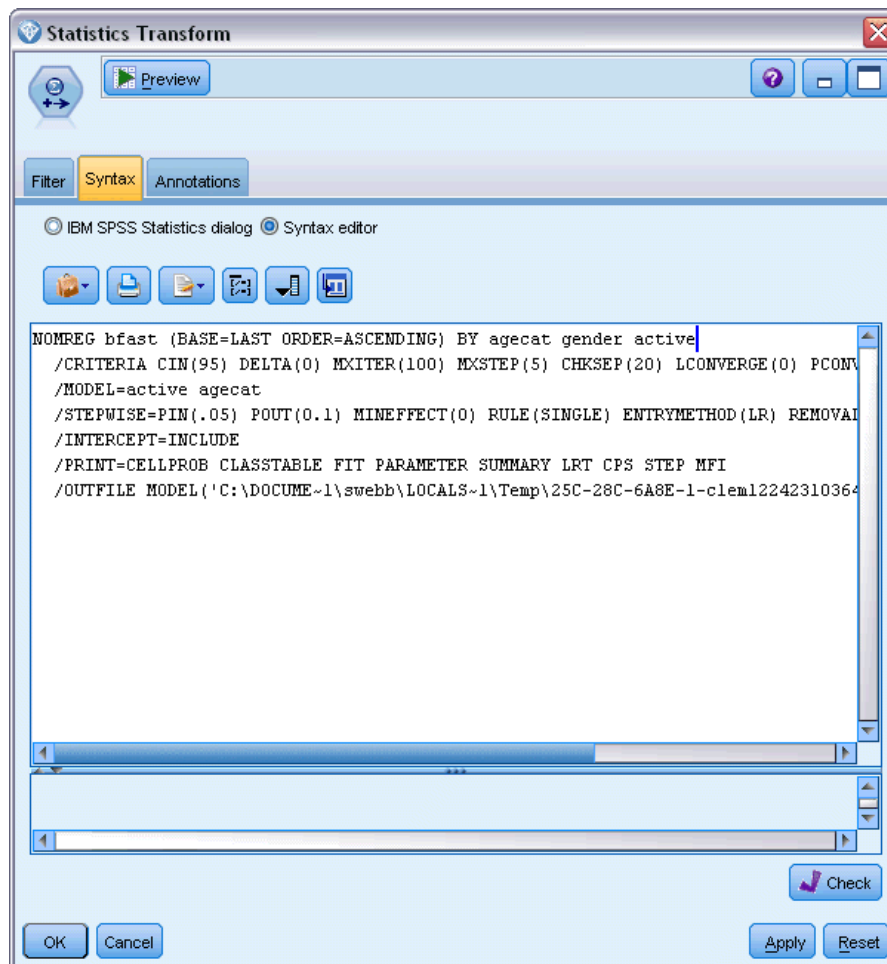
To do this, click the Select a dialog... button and choose the relevant SPSS Statistics procedure from the list. For more information, see the topic [Statistics Model Node - Model Tab](#) on p. 402.

Figure 8-2
Statistics Transform node, dialog selection



IBM SPSS Statistics Syntax editor option

Figure 8-3
Statistics Transform node, syntax editor



Check. After you have entered your syntax commands in the upper part of the dialog box, use this button to validate your entries. Any incorrect syntax is identified in the bottom part of the dialog box.

To ensure that the checking process does not take too long, when you validate the syntax, it is checked against a representative sample of your data to ensure that the entries are valid instead of checking against the entire dataset.

Allowable Syntax

If you have a lot of legacy syntax from IBM® SPSS® Statistics or are familiar with the data preparation features of SPSS Statistics, you can use the Statistics Transform node to run many of your existing transformations. As a guideline, the node enables you to transform data in predictable ways—for example, by running looped commands or by changing, adding, sorting, filtering, or selecting data.

Examples of the commands that can be carried out are:

- Compute random numbers according to a binomial distribution:

```
COMPUTE newvar = RV.BINOM(10000,0.1)
```

- Recode a variable into a new variable:

```
RECODE Age (Lowest thru 30=1) (30 thru 50=2) (50 thru Highest=3) INTO AgeRecoded
```

- Replace missing values:

```
RMV Age_1=SMEAN(Age)
```

The SPSS Statistics syntax that is supported by the Statistics Transform node is listed in the following table:

Command Name

```
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  
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
```

Command Name

MATRIX-END MATRIX
MISSING VALUES
N OF CASES
NUMERIC
PERMISSIONS
PRESERVE
RANK
RECODE
RENAME VARIABLES
RESTORE
RMV
SAMPLE
SELECT IF
SET
SORT CASES
SORT CASES
STRING
SUBTITLE
TEMPORARY
TITLE
UPDATE
V2C
VALIDATEDATA
VALUE LABELS
VARIABLE ATTRIBUTE
VARSTOCASES
VECTOR

Statistics Model Node

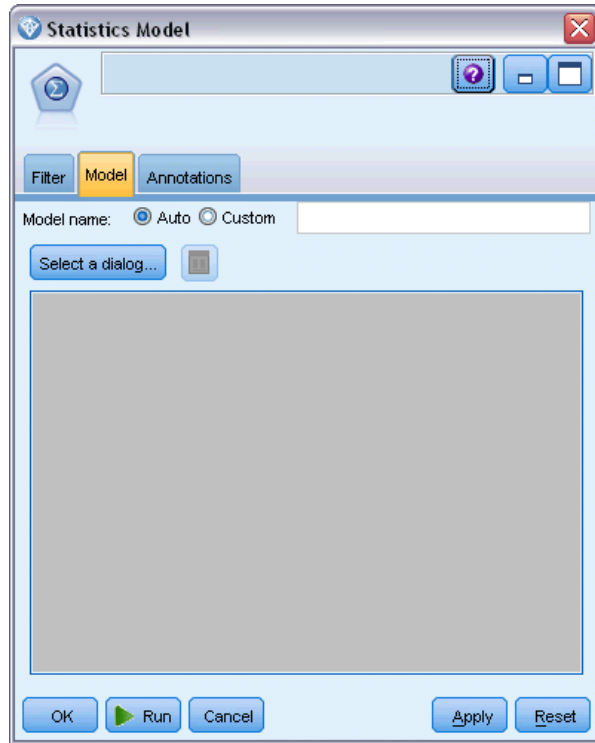
The Statistics Model node allows you to analyze and work with your data by running IBM® SPSS® Statistics procedures that produce PMML. The model nuggets you create can then be used in the usual way within IBM® SPSS® Modeler streams for scoring, and so on.

You must have a compatible version of SPSS Statistics installed and licensed on your computer to use this node. For more information, see the topic [IBM SPSS Statistics Helper Applications](#) in Chapter 6 on p. 369. See the Release Notes for compatibility information.

The SPSS Statistics analytical procedures that are available depend on the type of license you have.

Statistics Model Node - Model Tab

Figure 8-4
Statistics Model node, Model tab

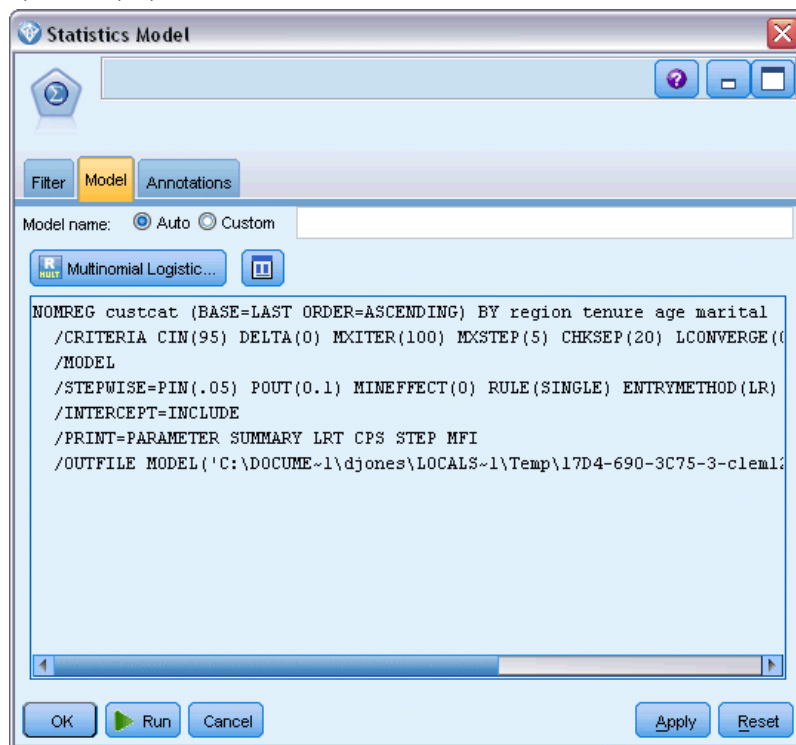


Model name. You can generate the model name automatically based on the target or ID field (or model type in cases where no such field is specified) or specify a custom name.

Select a dialog. Click to display a list of available IBM® SPSS® Statistics procedures that you can select and run. The list shows only those procedures that produce PMML and for which you are licensed, and does not include user-written procedures.

- ▶ Click on the required procedure; the relevant SPSS Statistics dialog is displayed.
- ▶ In the SPSS Statistics dialog, enter the details for the procedure.
- ▶ Click OK to return to the Statistics Model node; the SPSS Statistics syntax is displayed in the Model tab.

Figure 8-5
Syntax displayed on the Model tab

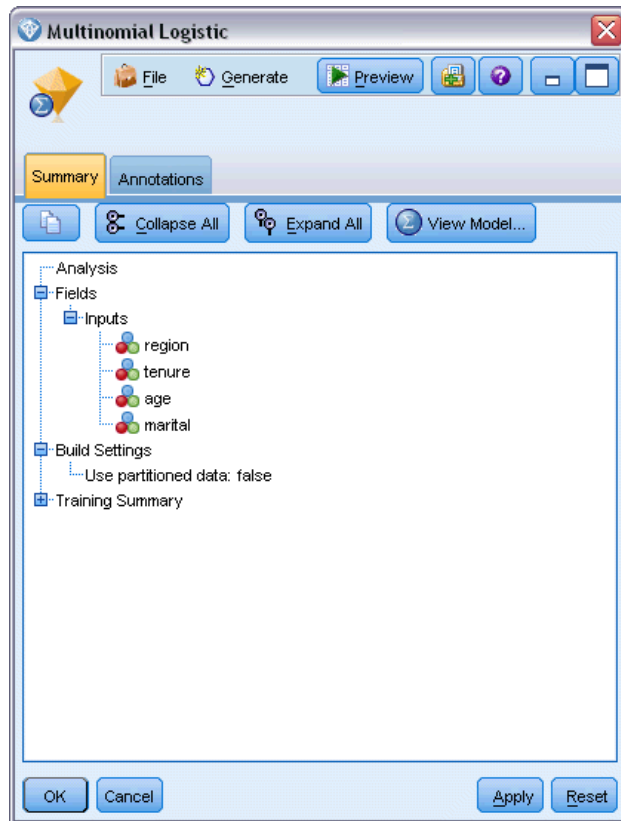


- To return to the SPSS Statistics dialog at any time, for example to modify your query, click the SPSS Statistics dialog display button to the right of the procedure selection button.

Statistics Model Node - Model Nugget Summary

When you run the Statistics Model node, it executes the associated IBM® SPSS® Statistics procedure and creates a model nugget that you can use in IBM® SPSS® Modeler streams for scoring.

Figure 8-6
Statistics Model nugget, Summary tab



The Summary tab of the model nugget displays information about the fields, build settings, and model estimation process. Results are presented in a tree view that can be expanded or collapsed by clicking specific items.

The View Model button displays the results in a modified form of the SPSS Statistics Output Viewer. For more information about this viewer, see the SPSS Statistics documentation.

The usual exporting and printing options are available from the File menu. For more information, see the topic [Viewing Output](#) in Chapter 6 on p. 316.

Figure 8-7
Statistics Model nugget, Advanced tab

The screenshot shows the PASW Statistics Output Viewer window for 'output1003.spv'. The left pane shows a tree view of the output, with 'Model Fitting Information' selected. The main pane displays the following data:

Model Fitting Information				
Model	Model Fitting Criteria	Likelihood Ratio Tests		
		-2 Log Likelihood	Chi-Square	df
Intercept Only	2.737E3			
Final	2.142E3	594.986	399	.000

Pseudo R-Square	
Cox and Snell	.448
Nagelkerke	.479
McFadden	.215

Likelihood Ratio Tests		
Effect	Model Fitting Criteria	Likelihood Ratio Tests

The status bar at the bottom right indicates 'H: 120, W: 300 pt.'

Statistics Output Node

The Statistics Output node allows you to call an IBM® SPSS® Statistics procedure to analyze your IBM® SPSS® Modeler data. You can view the results in a browser window or save results in the SPSS Statistics output file format. A wide variety of SPSS Statistics analytical procedures is accessible from SPSS Modeler.

You must have a compatible version of SPSS Statistics installed and licensed on your computer to use this node. For more information, see the topic [IBM SPSS Statistics Helper Applications](#) in Chapter 6 on p. 369. See the Release Notes for compatibility information.

If necessary, you can use the Filter tab to filter or rename fields so they conform to SPSS Statistics naming standards. For more information, see the topic [Renaming or Filtering Fields for IBM SPSS Statistics](#) on p. 410.

Syntax Reference. For details on specific SPSS Statistics procedures, see the *SPSS Statistics Command Syntax Reference* guide, included with your copy of the SPSS Statistics software. To view the guide from the Syntax tab, choose the Syntax editor option and click the Launch SPSS Statistics Syntax Help button.

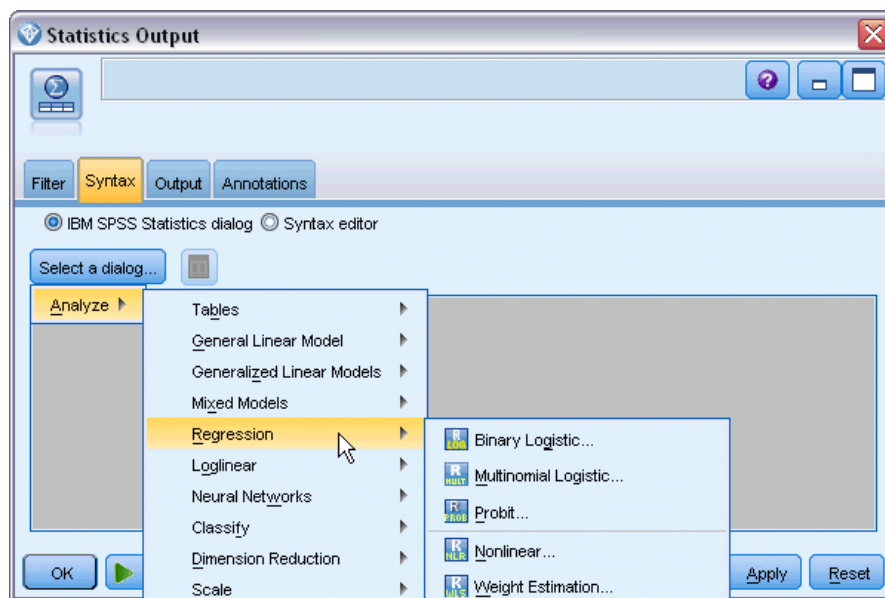
Statistics Output Node - Syntax Tab

Use this tab to create syntax for the IBM® SPSS® Statistics procedure you want to use to analyze your data. Syntax is composed of two parts: a **statement** and associated **options**. The statement specifies the analysis or operation to be performed and the fields to be used. The options specify everything else, including which statistics to display, derived fields to save, and so on.

IBM SPSS Statistics dialog option

If you are unfamiliar with SPSS Statistics syntax, the simplest way to create syntax in IBM® SPSS® Modeler is to first run the command in SPSS Statistics, copy the syntax into the SPSS Statistics Output node in SPSS Modeler, and run the stream.

Figure 8-8
Statistics Output node, dialog selection



To create the syntax:

- ▶ Click the Select a dialog button.
- ▶ Choose one of the options:
 - **Analyze.** Lists the contents of the SPSS Statistics Analyze menu; choose the procedure you want to use.
 - **Other.** If shown, lists dialogs created by the Custom Dialog Builder in SPSS Statistics, as well as any other SPSS Statistics dialogs that do not appear on the Analyze menu and for which you have a licence. If there are no applicable dialogs, this option is not shown.

Note: The Automatic Data Preparation dialogs are not shown.

If you have an SPSS Statistics custom dialog that creates new fields, these fields cannot be used in SPSS Modeler because the Statistics Output node is a terminal node.

- ▶ The rest of the process is similar to that used for the Statistics Model node. For more information, see the topic [Statistics Model Node - Model Tab](#) on p. 402.

Syntax editor option

To save syntax that has been created for a frequently used procedure:

- ▶ Click the File Options button (the first one on the toolbar).
- ▶ Choose Save or Save As from the menu.
- ▶ Save the file as an *.sps* file.

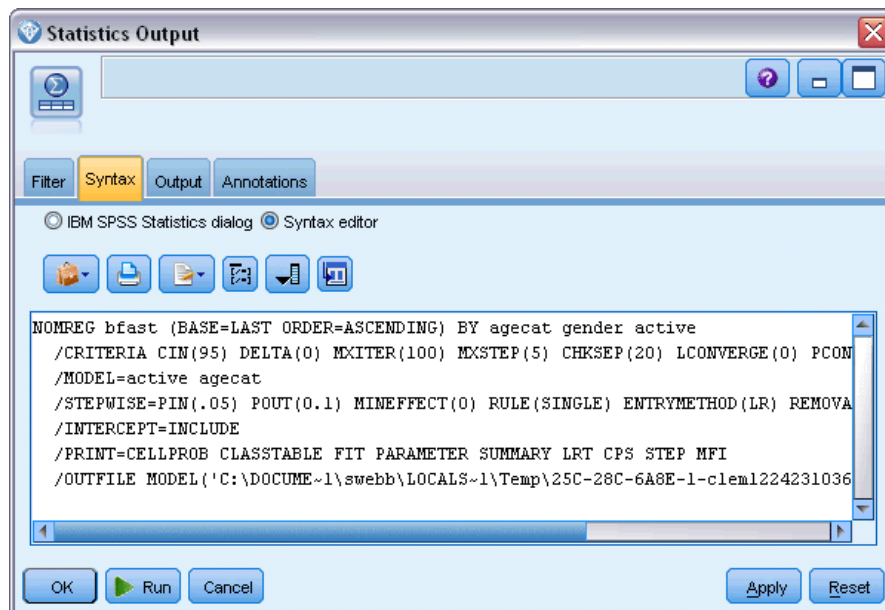
To use previously created syntax files, replacing the current contents, if any, of the Syntax editor:

- ▶ Click the File Options button (the first one on the toolbar).
- ▶ Choose Open from the menu.
- ▶ Select an *.sps* file to paste its contents into the Output node Syntax tab.

To insert previously saved syntax without replacing the current contents:

- ▶ Click the File Options button (the first one on the toolbar).
- ▶ Choose Insert from the menu
- ▶ Select an *.sps* file to paste its contents into the Output node at the point specified by the cursor.

Figure 8-9
Statistics Output node, Syntax editor

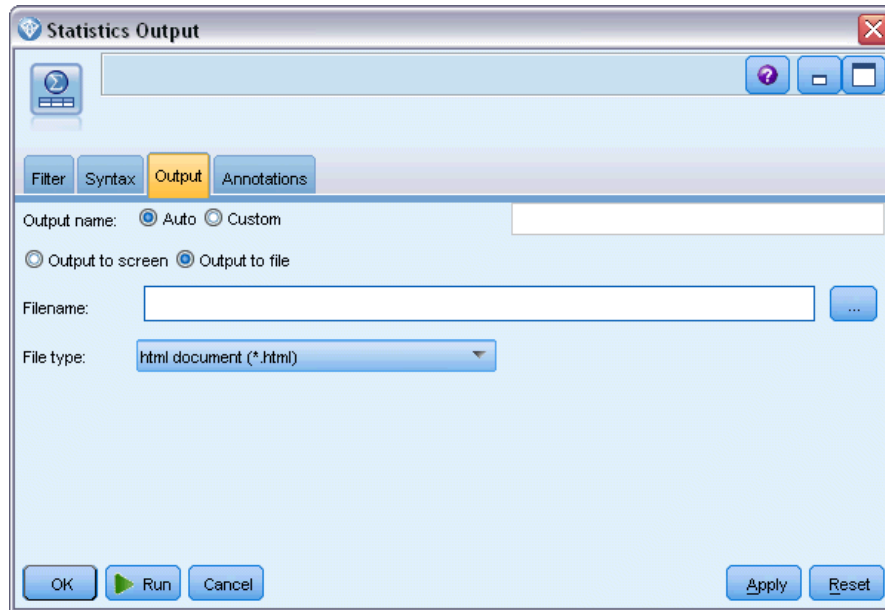


When you click Run, the results are shown in the SPSS Statistics Output Viewer. For more information about the viewer, please see the SPSS Statistics documentation.

Statistics Output Node - Output Tab

The Output tab lets you specify the format and location of the output. You can choose to display the results on the screen or send them to one of the available file types.

Figure 8-10
Statistics Output node, Output tab



Output name. Specifies the name of the output produced when the node is executed. Auto chooses a name based on the node that generates the output. Optionally, you can select Custom to specify a different name.

Output to screen (the default). Creates an output object to view online. The output object will appear on the Outputs tab of the manager window when the output node is executed.

Output to file. Saves the output to a file when you run the node. If you choose this option, enter a filename in the Filename field (or navigate to a directory and specify a filename using the File Chooser button) and select a file type.

File type. Choose the type of file to which you want to send the output.

- **HTML document (*.html).** Writes the output in HTML format.
- **SPSS Statistics Viewer File (*.spv).** Writes the output in a format that can be read by the IBM® SPSS® Statistics Output Viewer.
- **SPSS Statistics Web Reports File (*.spw).** Writes the output in SPSS Statistics Web Reports format, which can be published to an IBM SPSS Collaboration and Deployment Services repository and subsequently viewed in a Web browser. For more information, see the topic [Publish to Web](#) in Chapter 6 on p. 316.

Statistics Export Node

The Statistics Export node allows you to export data in IBM® SPSS® Statistics *.sav* format. SPSS Statistics *.sav* files can be read by SPSS Statistics Base and other modules. This is also the format used for IBM® SPSS® Modeler cache files.

Mapping SPSS Modeler field names to SPSS Statistics variable names can sometimes cause errors because SPSS Statistics variable names are limited to 64 characters and cannot include certain characters, such as spaces, dollar signs (\$), and dashes (-). There are two ways to adjust for these restrictions:

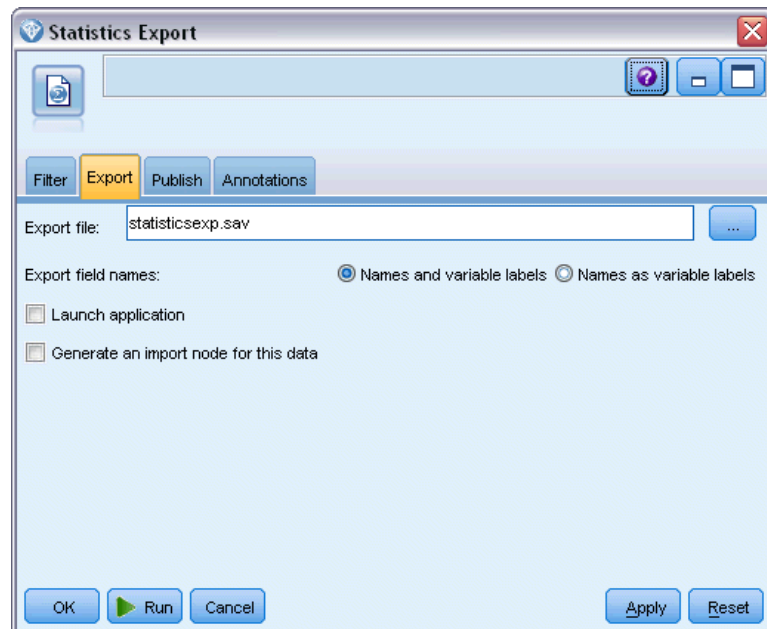
- You can rename fields conforming to SPSS Statistics variable name requirements by clicking the Filter tab. For more information, see the topic [Renaming or Filtering Fields for IBM SPSS Statistics](#) on p. 410.
- Choose to export both field names and labels from SPSS Modeler.

Note: SPSS Modeler writes *.sav* files in Unicode UTF-8 format. SPSS Statistics only supports files in Unicode UTF-8 format from release 16.0 onwards. To prevent the possibility of data corruption *.sav* files saved with Unicode encoding should not be used in releases of SPSS Statistics prior to 16.0. For more information please see the SPSS Statistics help.

Multiple response sets. Any multiple response sets defined in the stream will automatically be preserved when the file is exported. You can view and edit multiple response sets from any node with a Filter tab. For more information, see the topic [Editing Multiple Response Sets](#) in Chapter 4 on p. 134.

Statistics Export Node - Export Tab

Figure 8-11
Statistics Export node, Export tab



Export file. Specifies the name of the file. Enter a filename or click the file chooser button to browse to the file's location.

Export field names. Specifies a method of handling variable names and labels upon export from IBM® SPSS® Modeler to an IBM® SPSS® Statistics *.sav* file.

- **Names and variable labels.** Select to export both SPSS Modeler field names and field labels. Names are exported as SPSS Statistics variable names, while labels are exported as SPSS Statistics variable labels.
- **Names as variable labels.** Select to use the SPSS Modeler field names as variable labels in SPSS Statistics. SPSS Modeler allows characters in field names that are invalid in SPSS Statistics variable names. To prevent possibly creating invalid SPSS Statistics names, select Names as variable labels instead, or use the Filter tab to adjust field names.

Launch Application. If SPSS Statistics is installed on your computer, you can select this option to invoke the application directly on the saved data file. Options for launching the application must be specified in the Helper Applications dialog box. For more information, see the topic [IBM SPSS Statistics Helper Applications](#) in Chapter 6 on p. 369. To simply create an SPSS Statistics *.sav* file without opening an external program, deselect this option.

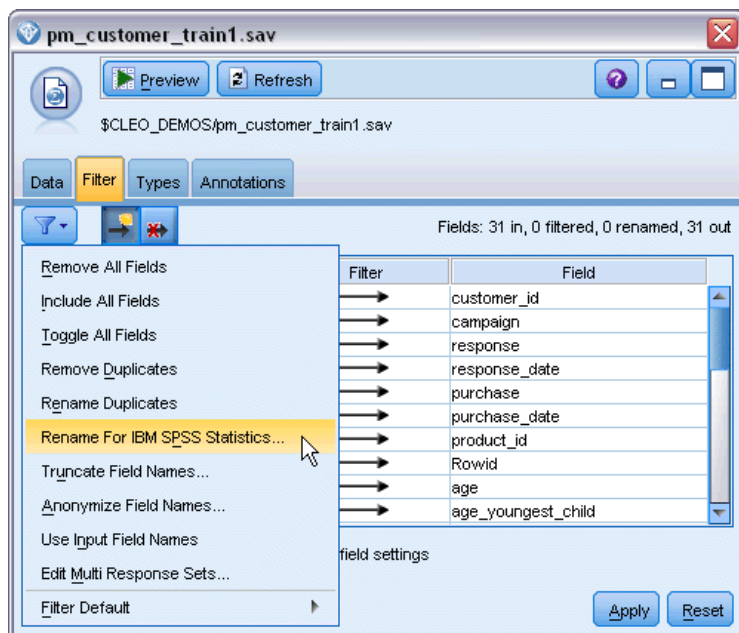
Generate an import node for this data. Select this option to automatically generate a Statistics File source node that will read the exported data file. For more information, see the topic [Statistics File Node](#) on p. 396.

Renaming or Filtering Fields for IBM SPSS Statistics

Before exporting or deploying data from IBM® SPSS® Modeler to external applications such as IBM® SPSS® Statistics, it may be necessary to rename or adjust field names. The Statistics Transform, Statistics Output, and Statistics Export dialog boxes contain a Filter tab to facilitate this process.

A basic description of Filter tab functionality is discussed elsewhere. For more information, see the topic [Setting Filtering Options](#) in Chapter 4 on p. 131. This topic provides tips for reading data into SPSS Statistics.

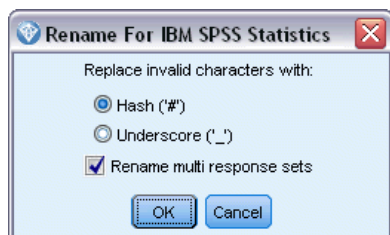
Figure 8-12
Renaming fields for IBM SPSS Statistics on the Filter tab of the Statistics File node



To adjust field names to conform to SPSS Statistics naming conventions:

- ▶ On the Filter tab, click the Filter Options Menu toolbar button (the first one on the toolbar).
- ▶ Select Rename For SPSS Statistics.

Figure 8-13
Renaming fields



- ▶ On the Rename For SPSS Statistics dialog, you can choose to replace invalid characters in filenames with either a Hash (#) character or an Underscore (_).

Rename multi response sets. Select this option if you want to adjust the names of multiple response sets, which can be imported into SPSS Modeler using a Statistics File source node. They are used to record data that can have more than one value for each case, such as in survey responses.

SuperNodes

Overview of SuperNodes

One of the reasons that IBM® SPSS® Modeler’s visual programming interface is so easy to learn is that each node has a clearly defined function. However, for complex processing, a long sequence of nodes may be necessary. Eventually, this may clutter the stream canvas and make it difficult to follow stream diagrams. There are two ways to avoid the clutter of a long and complex stream:

- You can split a processing sequence into several streams that feed one into the other. The first stream, for example, creates a data file that the second uses as input. The second creates a file that the third uses as input, and so on. You can manage these multiple streams by saving them in a **project**. A project provides organization for multiple streams and their output. However, a project file contains only a reference to the objects it contains, and you will still have multiple stream files to manage.
- You can create a **SuperNode** as a more streamlined alternative when working with complex stream processes.

SuperNodes group multiple nodes into a single node by encapsulating sections of a data stream. This provides numerous benefits to the data miner:

- Streams are neater and more manageable.
- Nodes can be combined into a business-specific SuperNode.
- SuperNodes can be exported to libraries for reuse in multiple data mining projects.

Types of SuperNodes

SuperNodes are represented in the data stream by a star icon. The icon is shaded to represent the type of SuperNode and the direction in which the stream must flow to or from it.

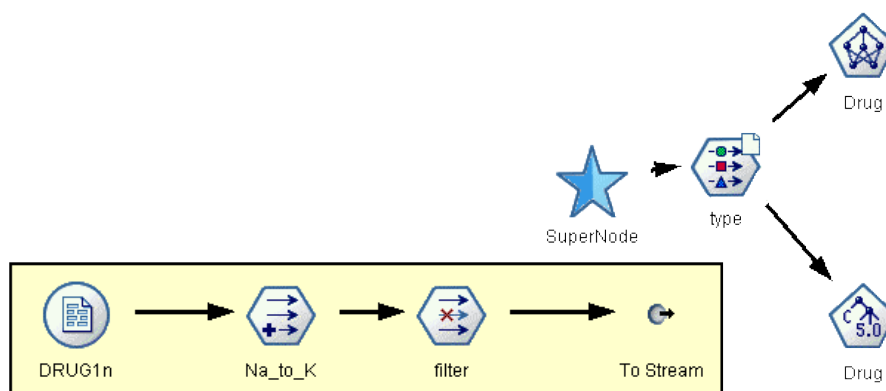
There are three types of SuperNodes:

- Source SuperNodes
- Process SuperNodes
- Terminal SuperNodes

Source SuperNodes

Source SuperNodes contain a data source just like a normal source node and can be used anywhere that a normal source node can be used. The left side of a source SuperNode is shaded to indicate that it is “closed” on the left and that data must flow downstream *from* a SuperNode.

Figure 9-1
Source SuperNode with zoomed-in version imposed over stream

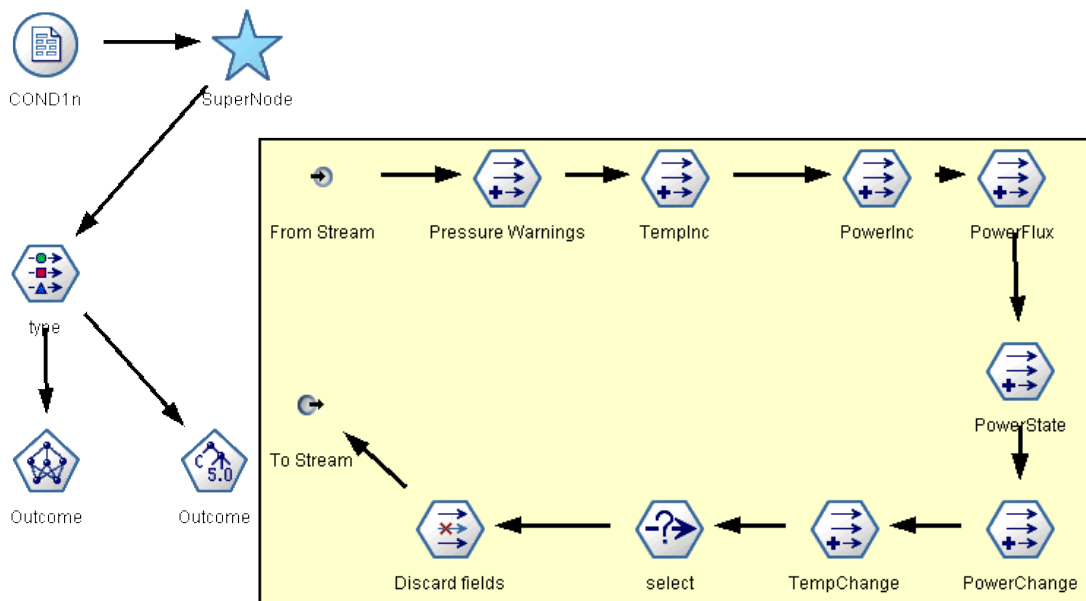


Source SuperNodes have only one connection point on the right, showing that data leaves the SuperNode and flows to the stream.

Process SuperNodes

Process SuperNodes contain only process nodes and are unshaded to show that data can flow both *in* and *out* of this type of SuperNode.

Figure 9-2
Process SuperNode with zoomed-in version imposed over stream



Process SuperNodes have connection points on both the left and right, showing that data enters the SuperNode and leaves to flow back to the stream. Although SuperNodes can contain additional stream fragments and even extra streams, both connection points must flow through a single path connecting the *From Stream* and *To Stream* points.

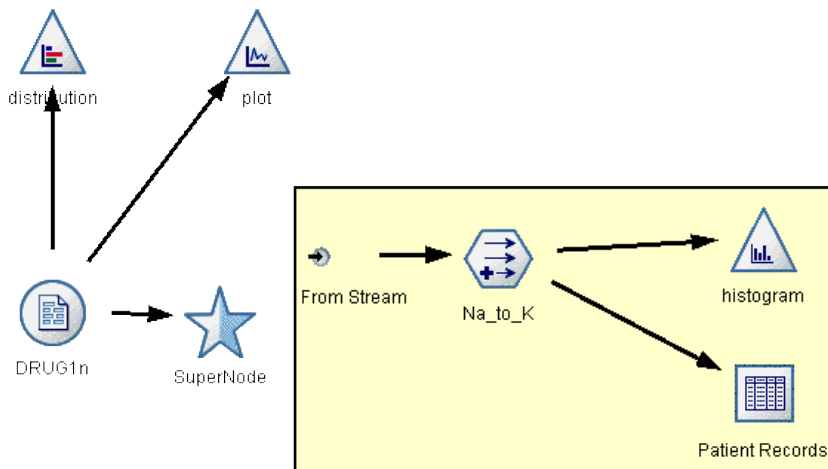
Note: Process SuperNodes are also sometimes referred to as *Manipulation SuperNodes*.

Terminal SuperNodes

Terminal SuperNodes contain one or more terminal nodes (plot, table, and so on) and can be used in the same manner as a terminal node. A terminal SuperNode is shaded on the right side to indicate that it is “closed” on the right and that data can flow only *into* a terminal SuperNode.

Figure 9-3

Terminal SuperNode with zoomed-in version imposed over stream



Terminal SuperNodes have only one connection point on the left, showing that data enters the SuperNode from the stream and terminates inside the SuperNode.

Terminal SuperNodes can also contain scripts that are used to specify the order of execution for all terminal nodes inside the SuperNode. For more information, see the topic [SuperNodes and Scripting](#) on p. 429.

Creating SuperNodes

Creating a SuperNode “shrinks” the data stream by encapsulating several nodes into one node. Once you have created or loaded a stream on the canvas, there are several ways to create a SuperNode.

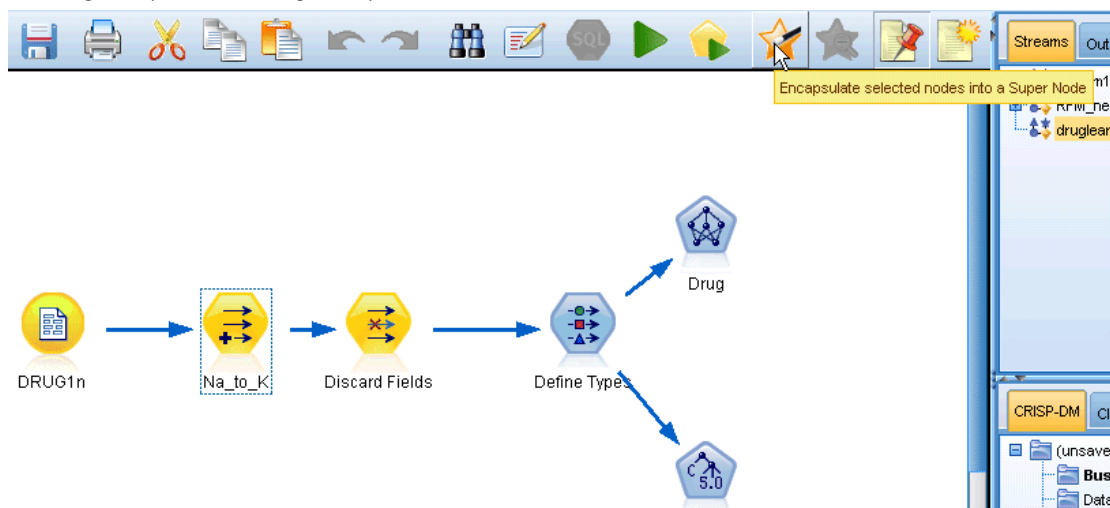
Multiple Selection

The simplest way to create a SuperNode is by selecting all of the nodes that you want to encapsulate:

- ▶ Use the mouse to select multiple nodes on the stream canvas. You can also use Shift-click to select a stream or section of a stream. *Note:* Nodes that you select must be from a continuous or forked stream. You cannot select nodes that are not adjacent or connected in some way.

- ▶ Then, using one of the following three methods, encapsulate the selected nodes:
 - Click the SuperNode icon (shaped like a star) on the toolbar.
 - Right-click the SuperNode, and from the context menu choose:
Create SuperNode > From Selection
 - From the SuperNode menu, choose:
Create SuperNode > From Selection

Figure 9-4
Creating a SuperNode using multiple selection



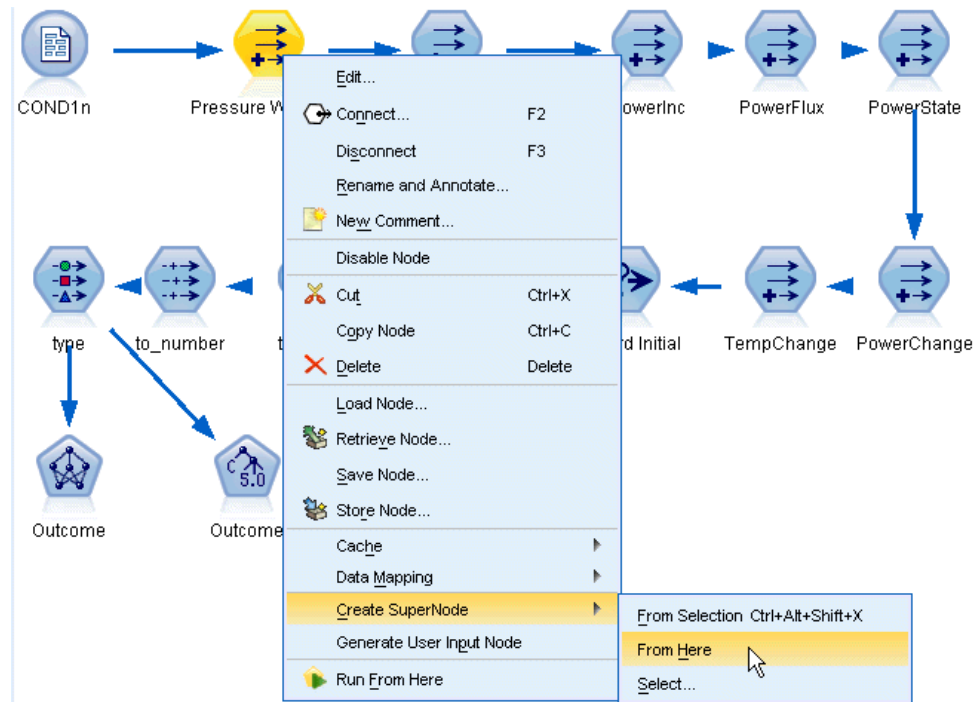
All three of these options encapsulate the nodes into a SuperNode shaded to reflect its type—source, process, or terminal—based on its contents.

Single Selection

You can also create a SuperNode by selecting a single node and using menu options to determine the start and end of the SuperNode or encapsulating everything downstream of the selected node.

- ▶ Click the node that determines the start of encapsulation.
- ▶ From the SuperNode menu, choose:
Create SuperNode > From Here

Figure 9-5
Creating a SuperNode using the context menu for single selection



SuperNodes can also be created more interactively by selecting the start and end of the stream section to encapsulate nodes:

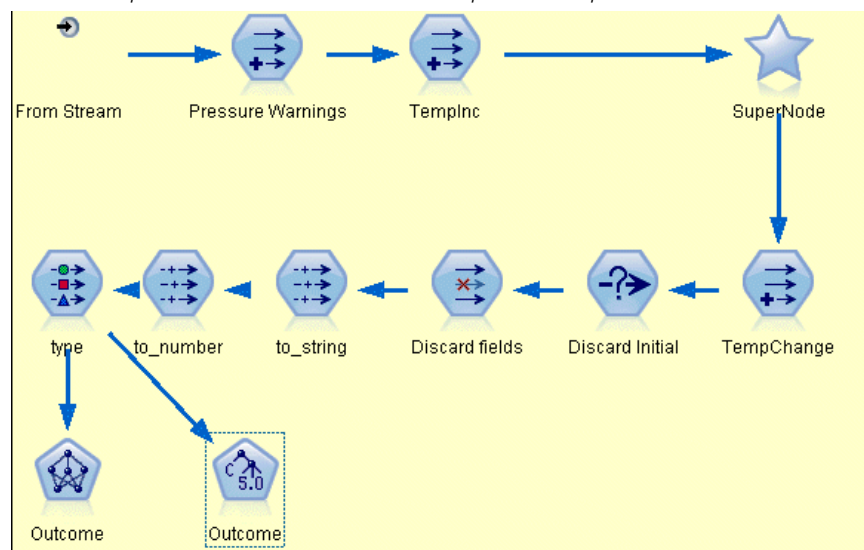
- ▶ Click on the first or last node that you want to include in the SuperNode.
- ▶ From the SuperNode menu, choose:
Create SuperNode > Select...
- ▶ Alternatively, you can use the context menu options by right-clicking the desired node.
- ▶ The cursor becomes a SuperNode icon, indicating that you must select another point in the stream. Move either upstream or downstream to the “other end” of the SuperNode fragment and click on a node. This action will replace all nodes in between with the SuperNode star icon.

Note: Nodes that you select must be from a continuous or forked stream. You cannot select nodes that are not adjacent or connected in some way.

Nesting SuperNodes

SuperNodes can be nested within other SuperNodes. The same rules for each type of SuperNode (source, process, and terminal) apply to nested SuperNodes. For example, a process SuperNode with nesting must have a continuous data flow through all nested SuperNodes in order for it to remain a process SuperNode. If one of the nested SuperNodes is terminal, then data would no longer flow through the hierarchy.

Figure 9-6
Process SuperNode nested within another process SuperNode



Terminal and source SuperNodes can contain other types of nested SuperNodes, but the same basic rules for creating SuperNodes apply.

Examples of Valid SuperNodes

Almost anything you create in IBM® SPSS® Modeler can be encapsulated in a SuperNode. Following are examples of valid SuperNodes:

Figure 9-7
Valid process SuperNode with two connections in a valid stream flow

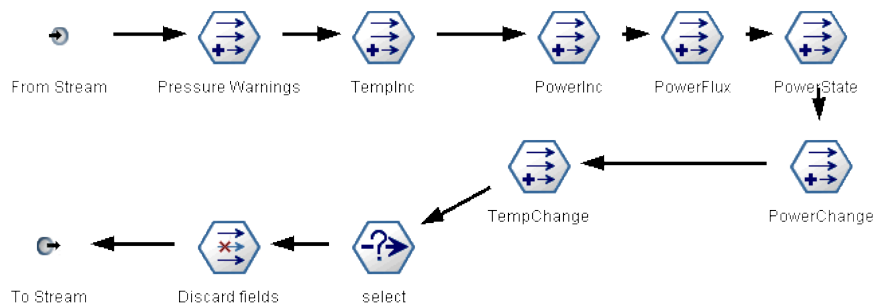


Figure 9-8
Valid terminal SuperNode including separate stream used to test generated models

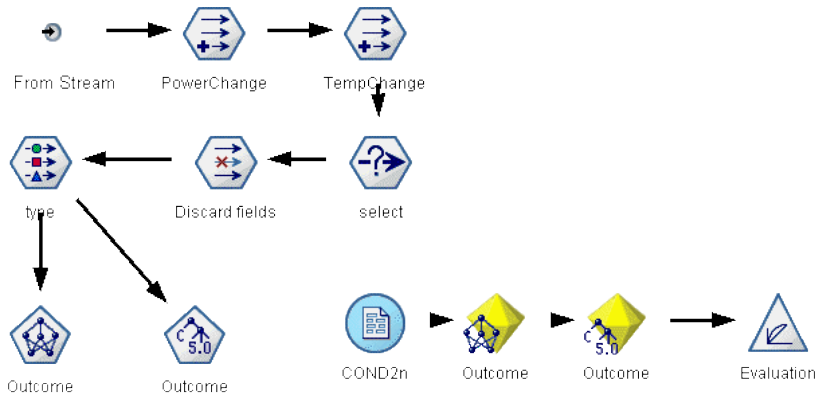
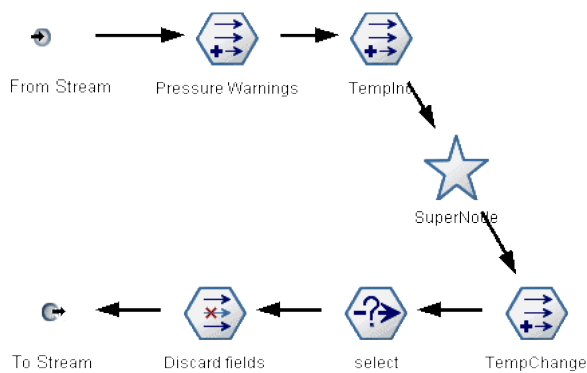


Figure 9-9
Valid process SuperNode containing a nested SuperNode



Examples of Invalid SuperNodes

The most important aspect of creating valid SuperNodes is to ensure that data flows linearly through the SuperNode connections. If there are two connections (a process SuperNode), then data must flow in a stream from the beginning connector to the ending connector. Similarly, a source SuperNode must allow data to flow from the source node to the single connector that brings data back to the zoomed-out stream.

Figure 9-10
Invalid source SuperNode: Source node not connected to the data flow path

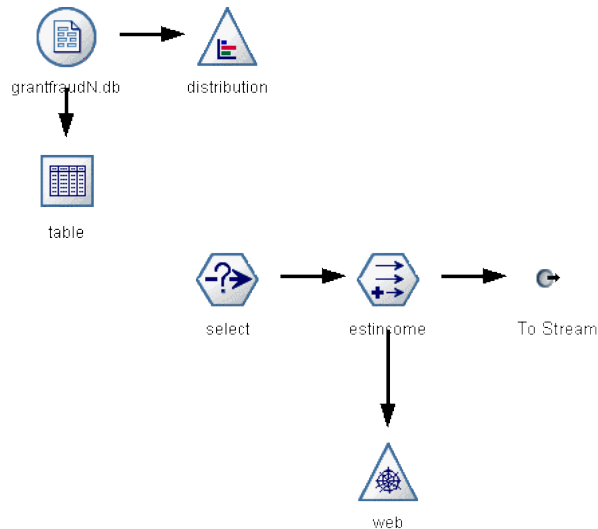
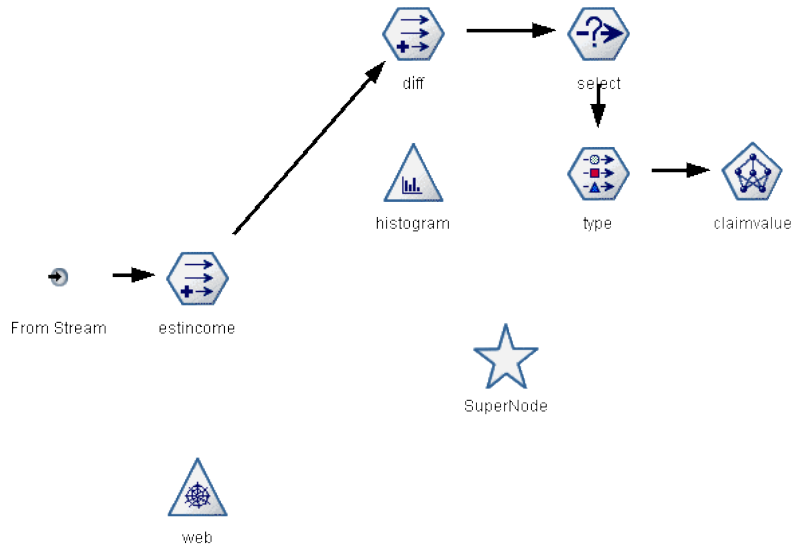


Figure 9-11
Invalid terminal SuperNode: Nested SuperNode not connected to the data flow path



Locking SuperNodes

Once you have created a SuperNode, you can lock it with a password to prevent it from being amended. For example, you might do this if you are creating streams, or parts of streams, as fixed-value templates for use by others in your organization who have less experience with setting up IBM® SPSS® Modeler enquiries.

When a SuperNode is locked users can still enter values on the Parameters tab for any parameters that have been defined, and a locked SuperNode can be executed without entering the password.

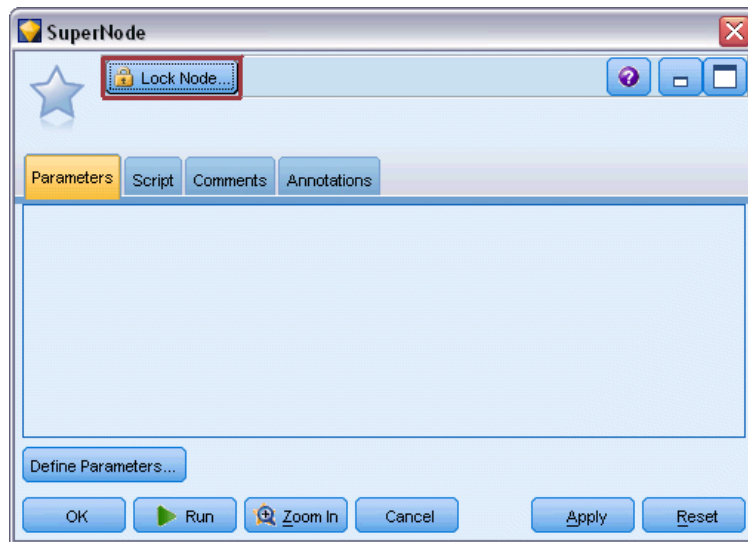
Note: Locking and unlocking cannot be performed using scripts.

Locking and Unlocking a SuperNode

Warning: Lost passwords cannot be recovered.

You can lock or unlock a SuperNode from any of the three tabs.

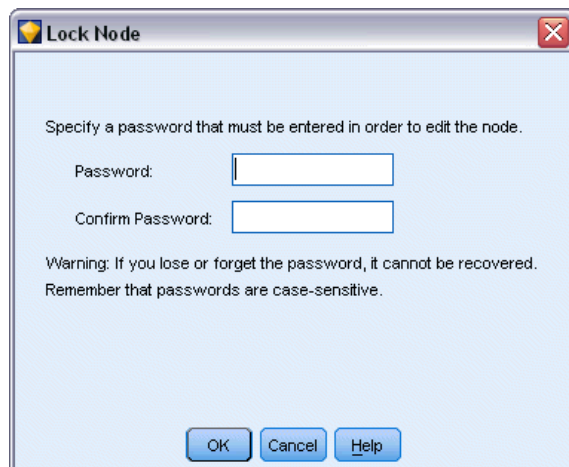
Figure 9-12
Locking a SuperNode



Click Lock Node.

Enter and confirm the password.

Figure 9-13
Enter and confirm SuperNode password



- ▶ Click OK.

A password protected SuperNode is identified on the stream canvas by a small padlock symbol to the top-left of the SuperNode icon.

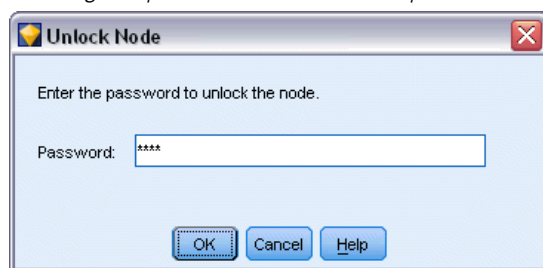
Figure 9-14
Locked source SuperNode as part of a stream



Unlock a SuperNode

- ▶ To permanently remove the password protection, click Unlock Node; you are prompted for the password.

Figure 9-15
Entering the password to unlock a SuperNode

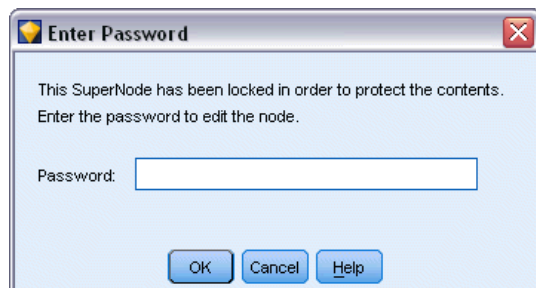


- ▶ Enter the password and click OK; the SuperNode is no longer password protected and the padlock symbol no longer shows next to the icon in the stream.

Editing a Locked SuperNode

If you attempt to either define parameters or zoom in to display a locked SuperNode you are prompted to enter the password.

Figure 9-16
Entering the password to zoom in to or edit a SuperNode



- ▶ Enter the password and click OK.

You are now able to edit the parameter definitions and zoom in and out as often as you require until you close the stream that the SuperNode is in.

Note that this does not remove the password protection, it only allows you access to work with the SuperNode. For more information, see the topic [Locking and Unlocking a SuperNode](#) on p. 420.

Editing SuperNodes

Once you have created a SuperNode, you can examine it more closely by zooming in to it; if the SuperNode is locked, you will be prompted to enter the password. For more information, see the topic [Editing a Locked SuperNode](#) on p. 421.

To view the contents of a SuperNode, you can use the zoom-in icon from the IBM® SPSS® Modeler toolbar, or the following method:

- ▶ Right-click a SuperNode.
- ▶ From the context menu, choose Zoom In.

The contents of the selected SuperNode will be displayed in a slightly different SPSS Modeler environment, with connectors showing the flow of data through the stream or stream fragment. At this level on the stream canvas, there are several tasks that you that can perform:

- Modify the SuperNode type—source, process, or terminal.
- Create parameters or edit the values of a parameter. Parameters are used in scripting and CLEM expressions.
- Specify caching options for the SuperNode and its subnodes.
- Create or modify a SuperNode script (terminal SuperNodes only).

Modifying SuperNode Types

In some circumstances, it is useful to alter the type of a SuperNode. This option is available only when you are zoomed in to a SuperNode, and it applies only to the SuperNode at that level.

The three types of SuperNodes are:

Source SuperNode	One connection going out
Process SuperNode	Two connections: one coming in and one going out
Terminal SuperNode	One connection coming in

To Change the Type of a SuperNode

- ▶ Be sure that you are zoomed in to the SuperNode.
- ▶ From the SuperNode menu, choose SuperNode Type, and then choose the type.

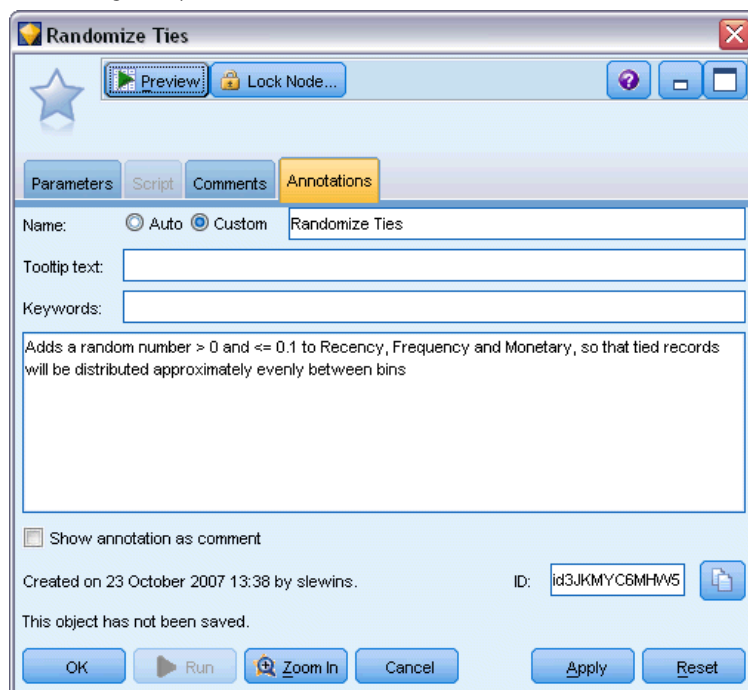
Annotating and Renaming SuperNodes

You can rename a SuperNode as it appears in the stream as well as write annotations used in a project or report. To access these properties:

- ▶ Right-click a SuperNode (zoomed out) and choose Rename and Annotate.
- ▶ Alternatively, from the SuperNode menu, choose Rename and Annotate. This option is available in both zoomed-in and zoomed-out modes.

In both cases, a dialog box opens with the Annotations tab selected. Use the options here to customize the name displayed on the stream canvas and provide documentation regarding SuperNode operations.

Figure 9-17
Annotating a SuperNode



Using Comments with SuperNodes

If you create a SuperNode from a commented node or nugget, you must include the comment in the selection to create the SuperNode if you want the comment to appear in the SuperNode. If you omit the comment from the selection, the comment will remain disconnected on the stream when the SuperNode is created.

When you expand a SuperNode that included comments, the comments are reinstated to where they were before the SuperNode was created.

When you expand a SuperNode that included commented objects, but the comments were not included in the SuperNode, the objects are reinstated to where they were, but the comments are not reattached.

SuperNode Parameters

In IBM® SPSS® Modeler, you have the ability to set user-defined variables, such as *Minvalue*, whose values can be specified when used in scripting or CLEM expressions. These variables are called **parameters**. You can set parameters for streams, sessions, and SuperNodes. Any parameters set for a SuperNode are available when building CLEM expressions in that SuperNode or any nested nodes. Parameters set for nested SuperNodes are not available to their parent SuperNode.

There are two steps to creating and setting parameters for SuperNodes:

- Define parameters for the SuperNode.
- Then, specify the value for each parameter of the SuperNode.

These parameters can then be used in CLEM expressions for any encapsulated nodes.

Defining SuperNode Parameters

Parameters for a SuperNode can be defined in both zoomed-out and zoomed-in modes. The parameters defined apply to all encapsulated nodes. To define the parameters of a SuperNode, you first need to access the Parameters tab of the SuperNode dialog box. Use one of the following methods to open the dialog box:

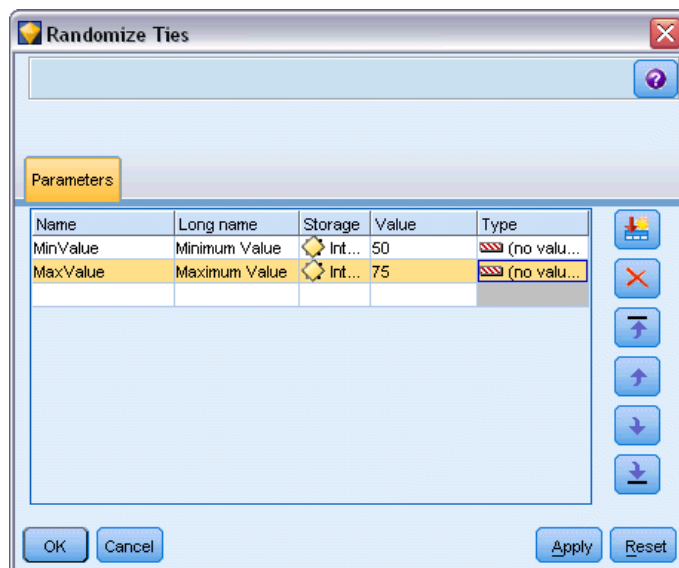
- Double-click a SuperNode in the stream.
- From the SuperNode menu, choose Set Parameters.
- Alternatively, when zoomed in to a SuperNode, choose Set Parameters from the context menu.

Once you have opened the dialog box, the Parameters tab is visible with any previously defined parameters.

To Define a New Parameter

- ▶ Click the Define Parameters button to open the dialog box.

Figure 9-18
Defining parameters for a SuperNode



Name. Parameter names are listed here. You can create a new parameter by entering a name in this field. For example, to create a parameter for the minimum temperature, you could type `minvalue`. Do not include the `$P-` prefix that denotes a parameter in CLEM expressions. This name is also used for display in the CLEM Expression Builder.

Long name. Lists the descriptive name for each parameter created.

Storage. Select a storage type from the list. Storage indicates how the data values are stored in the parameter. For example, when working with values containing leading zeros that you want to preserve (such as 008), you should select `String` as the storage type. Otherwise, the zeros will be stripped from the value. Available storage types are `string`, `integer`, `real`, `time`, `date`, and `timestamp`. For date parameters, note that values must be specified using ISO standard notation as shown in the next paragraph.

Value. Lists the current value for each parameter. Adjust the parameter as required. Note that for date parameters, values must be specified in ISO standard notation (that is, `YYYY-MM-DD`). Dates specified in other formats are not accepted.

Type (optional). If you plan to deploy the stream to an external application, select a measurement level from the list. Otherwise, it is advisable to leave the *Type* column as is. If you want to specify value constraints for the parameter, such as upper and lower bounds for a numeric range, select `Specify` from the list.

Note that long name, storage, and type options can be set for parameters through the user interface only. These options cannot be set using scripts.

Click the arrows at the right to move the selected parameter further up or down the list of available parameters. Use the delete button (marked with an *X*) to remove the selected parameter.

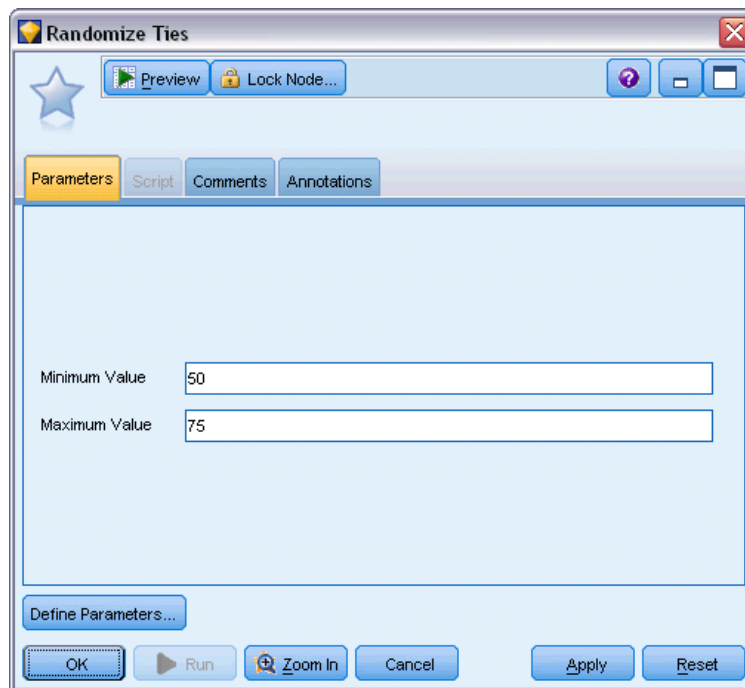
Setting Values for SuperNode Parameters

Once you have defined parameters for a SuperNode, you can specify values using the parameters in a CLEM expression or script.

To Specify the Parameters of a SuperNode

- ▶ Double-click the SuperNode icon to open the SuperNode dialog box.
- ▶ Alternatively, from the SuperNode menu, choose Set Parameters.
- ▶ Click the Parameters tab. *Note:* The fields in this dialog box are the fields defined by clicking the Define Parameters button on this tab.
- ▶ Enter a value in the text box for each parameter that you have created. For example, you can set the value *minvalue* to a particular threshold of interest. This parameter can then be used in numerous operations, such as selecting records above or below this threshold for further exploration.

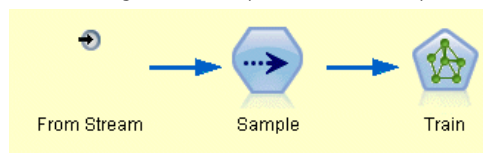
Figure 9-19
Specifying parameters for a SuperNode



Using SuperNode Parameters to Access Node Properties

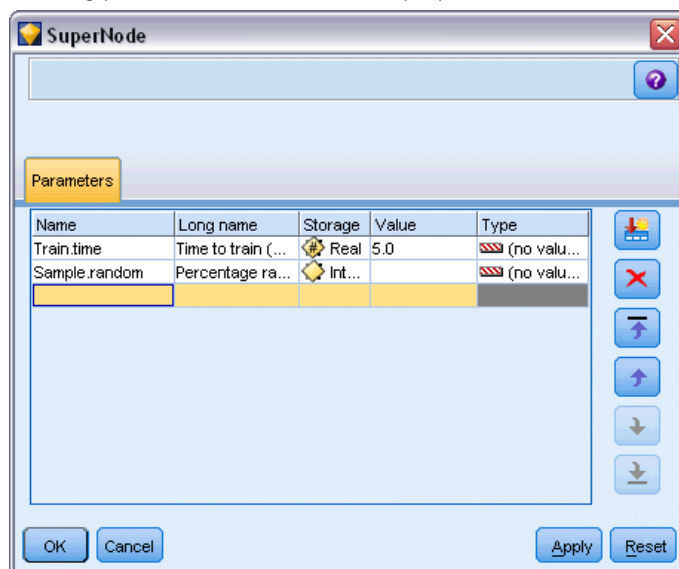
SuperNode parameters can also be used to define node properties (also known as **slot parameters**) for encapsulated nodes. For example, suppose you want to specify that a SuperNode train an encapsulated Neural Net node for a certain length of time using a random sample of the data available. Using parameters, you can specify values for the length of time and percentage sample.

Figure 9-20
Stream fragment encapsulated in a SuperNode



This example SuperNode contains a Sample node called *Sample* and a Neural Net node called *Train*. You can use the node dialog boxes to specify the Sample node's **Sample** setting as Random % and the Neural Net node's **Stop on** setting to Time. Once these options are specified, you can access the node properties with parameters and specify specific values for the SuperNode. In the SuperNode dialog box, click Define Parameters and create the following parameters:

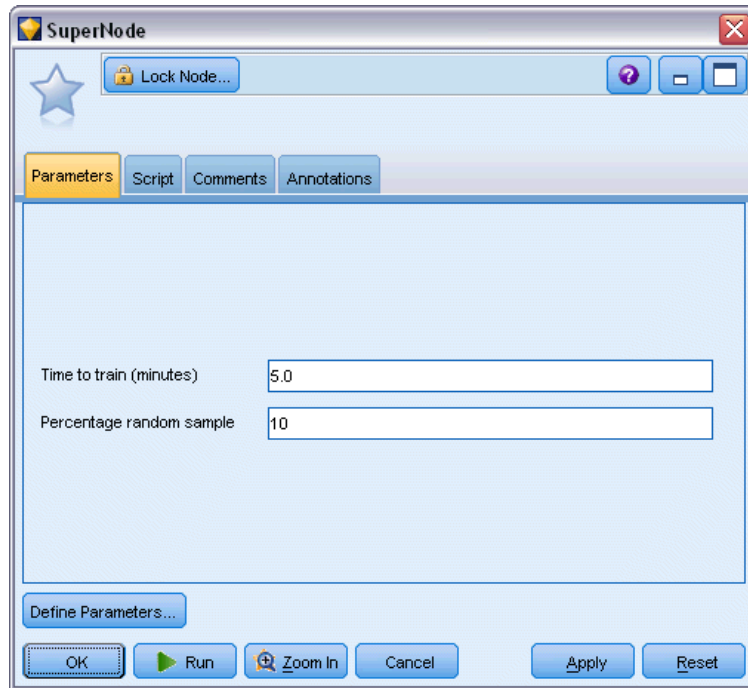
Figure 9-21
Defining parameters to access node properties



Note: The parameter names, such as *Sample.random*, use correct syntax for referring to node properties, where *Sample* represents the name of the node and *random* is a node property.

Once you have defined these parameters, you can easily modify values for the two Sample and Neural Net node properties without reopening each dialog box. Instead, simply select Set Parameters from the SuperNode menu to access the Parameters tab of the SuperNode dialog box, where you can specify new values for Random % and Time. This is particularly useful when exploring the data during numerous iterations of model building.

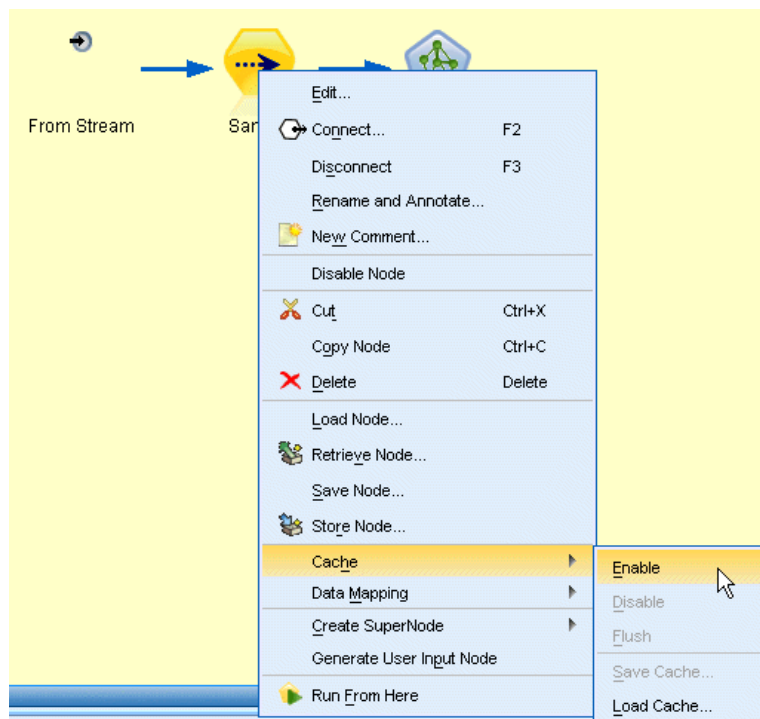
Figure 9-22
Specifying values for node properties on the Parameters tab in the SuperNode dialog box



SuperNodes and Caching

From within a SuperNode, all nodes except terminal nodes can be cached. Caching is controlled by right-clicking a node and choosing one of several options from the Cache context menu. This menu option is available both from outside a SuperNode and for the nodes encapsulated within a SuperNode.

Figure 9-23
Selecting caching options for a SuperNode



There are several guidelines for SuperNode caches:

- If any of the nodes encapsulated in a SuperNode have caching enabled, the SuperNode will also.
- Disabling the cache on a SuperNode disables the cache for *all* encapsulated nodes.
- Enabling caching on a SuperNode actually enables the cache on the last cacheable subnode. In other words, if the last subnode is a Select node, the cache will be enabled for that Select node. If the last subnode is a terminal node (which does not allow caching), the next node upstream that supports caching will be enabled.
- Once you have set caches for the subnodes of a SuperNode, any activities upstream from the cached node, such as adding or editing nodes, will flush the caches.

SuperNodes and Scripting

You can use the IBM® SPSS® Modeler scripting language to write simple programs that manipulate and execute the contents of a terminal SuperNode. For instance, you might want to specify the order of execution for a complex stream. As an example, if a SuperNode contains a Set Globals node that needs to be executed before a Plot node, you can create a script that executes the Set Globals node first. Values calculated by this node, such as the average or standard deviation, can then be used when the Plot node is executed.

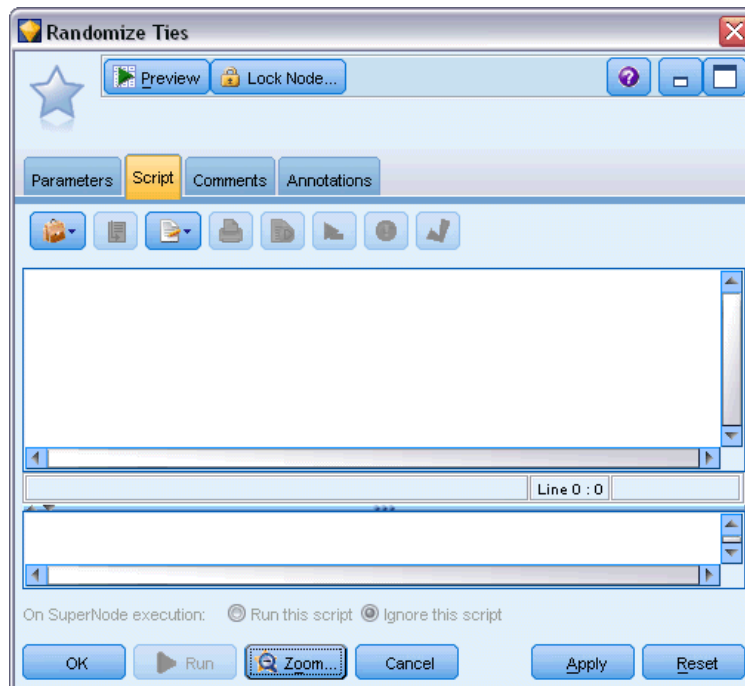
The Script tab of the SuperNode dialog box is available only for terminal SuperNodes.

To Open the Scripting Dialog Box for a Terminal SuperNode

- ▶ Right-click the SuperNode canvas and choose SuperNode Script.
- ▶ Alternatively, in both zoomed-in and zoomed-out modes, you can choose SuperNode Script from the SuperNode menu.

Note: SuperNode scripts are executed only with the stream and SuperNode when you have selected Run this script in the dialog box.

Figure 9-24
Creating a script for a SuperNode



Specific options for scripting and its use within SPSS Modeler are discussed in the *Scripting and Automation Guide*, which is available on the SPSS ModelerDVD.

Saving and Loading SuperNodes

One of the advantages of SuperNodes is that they can be saved and reused in other streams. When saving and loading SuperNodes, note that they use an *.slb* extension.

To Save a SuperNode

- ▶ Zoom in on the SuperNode.
- ▶ From the SuperNode menu, choose Save SuperNode.
- ▶ Specify a filename and directory in the dialog box.
- ▶ Select whether to add the saved SuperNode to the current project.

- ▶ Click Save.

To Load a SuperNode

- ▶ From the Insert menu in the IBM® SPSS® Modeler window, choose SuperNode.
- ▶ Select a SuperNode file (.slb) from the current directory or browse to a different one.
- ▶ Click Load.

Note: Imported SuperNodes have the default values for all of their parameters. To change the parameters, double-click on a SuperNode on the stream canvas.

Notices

This information was developed for products and services offered worldwide.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, U.S.A.

For license inquiries regarding double-byte character set (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japan.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Software Group, Attention: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Trademarks

IBM, the IBM logo, ibm.com, and SPSS are trademarks of IBM Corporation, registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at <http://www.ibm.com/legal/copytrade.shtml>.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

Other product and service names might be trademarks of IBM or other companies.

- 1-in-*n* sampling, 56
- 3-D graphs, 215

- adding
 - records, 65
- adjusted propensity scores
 - balancing data, 64
- ADO databases
 - importing, 28
- aesthetics
 - in visualizations, 210
- Aggregate node
 - overview, 65
 - parallel processing, 67
 - performance, 67
 - setting options, 66
- aggregating records, 179
- aggregating time series data, 188
- analysis browser
 - interpreting, 332
- Analysis node, 330
 - analysis tab, 330
 - output tab, 322
- animation
 - in visualizations, 213
- animation in graphs, 210, 212
- Anonymize node
 - creating anonymized values, 157
 - overview, 154
 - setting options, 155
- anonymizing field names, 133
- ANOVA
 - Means node, 359
- anti-join, 72
- Append node
 - field matching, 81
 - overview, 80
 - setting options, 81
 - tagging fields, 77
- application examples, 2
- ascending order, 69
- assigning data types, 50, 85, 113
- association plots, 254
- audit
 - Data Audit node, 335
 - initial data audit, 335
- Auto Data Prep node, 87
- auto- settings, 292
- auto-typing, 117, 119
- automated data preparation
 - action details, 109
 - action summary, 104
 - derive node generation, 112
 - exclude fields, 93
 - exclusion of unused fields, 91
 - field analysis, 102
 - field details, 107
 - field processing summary, 101
 - field settings, 91
 - fields, 90
 - fields table, 106
 - links between views, 100
 - model view, 99
 - name fields, 98
 - normalize continuous target, 112
 - objectives, 87
 - predictive power, 105
 - prepare dates and times, 92
 - reset views, 100
 - unused field exclusion, 91
- automatic data preparation
 - construction, 96
 - feature selection, 96
 - input preparation, 94
 - normalize continuous target, 95
 - preparing inputs, 94
 - preparing targets, 94
 - selection of features, 96
 - target preparation, 94
- automatic date recognition, 20, 23
- automatic recode, 158–159

- balance factors, 64
- Balance node
 - generating from graphs, 288
 - overview, 63
 - setting options, 64
- bands in graphs, 282
- baseline
 - evaluation chart options, 274
- best line
 - evaluation chart options, 274
- biased data, 63
- Binning node
 - equal counts, 165
 - equal sums, 165
 - fixed-width bins, 164
 - mean/standard deviation bins, 169
 - optimal, 170
 - overview, 162
 - previewing bins, 171
 - ranks, 168
 - setting options, 163
- BITMAP indexes
 - database tables, 379
- Blank function
 - padding time series, 190
- blank handling, 50, 113, 120
 - Binning node, 163
 - filling values, 151
- blank rows
 - Excel files, 39

- blank values
 - in Matrix tables, 326
- blanks, 344
 - in Matrix tables, 326
- bulk loading, 380
- business rule
 - evaluation chart options, 275

- cache
 - SuperNodes, 428
- cache file node, 396
- case data
 - Data Collection source node, 26–27
- categorical data, 116–117
- cell ranges
 - Excel files, 39
- charts
 - saving output, 322
- checking types, 124
- chi-square
 - Matrix node, 328
- clear values, 50
- CLEM expressions, 53
- cluster, 306
- clustered samples, 55–56, 59
- Codes variables
 - IBM SPSS Data Collection source node, 30
- coercing values, 124
- Cognos, *see* IBM Cognos BI, 36
- coincidence matrix
 - Analysis node, 330
- Collection node , 245
 - appearance tab, 248
 - options tab, 246–247
 - using the graph, 249
- collision modifiers, 303
- color
 - in visualizations, 210
- color graph overlay, 210
- column order
 - table browser, 320, 324
- column width
 - for fields, 128
- column-wise binding, 380
- combining data, 80
 - from multiple files, 71
- comma, 20, 128
- comma-delimited files
 - exporting, 319, 389
 - saving, 322
- comment characters
 - in variable files, 19
- comments
 - using with SuperNodes, 423
- commit size, 380
- compute durations
 - automated data preparation, 92
- concatenating records, 80
- conditions
 - specifying a series, 148
- confidence intervals
 - Means node, 363–364
- connections
 - to IBM SPSS Collaboration and Deployment Services Repository, 6
- contiguous data sampling, 56
- contiguous keys, 66
- continuous data, 116–117, 122
- convert measurement levels, 117
- converting sets to flags, 178, 180
- coordinate systems
 - transforming, 302
- copying type attributes, 127
- copying visualizations, 306
- correlations, 355
 - absolute value, 355
 - descriptive labels, 355
 - Means node, 364
 - probability, 355
 - significance, 355
 - statistics output, 356
- costs
 - evaluation charts, 275
- Count field
 - padding or aggregating time series, 190
 - Time Intervals node, 190
- counts
 - Binning node, 165
 - statistics output, 356
- CREATE INDEX command, 378
- creating
 - new fields, 139, 141
- CRISP-DM
 - data understanding, 5
- CRISP-DM process model
 - data preparation, 85
- cross-tabulation
 - Matrix node, 326–327
- CSV data
 - importing, 28
- currency display format, 129
- cut points
 - Binning node, 162
- cyclic periods
 - Time Intervals node, 194
- cyclical time elements
 - automated data preparation, 92

- daily measurements
 - Time Intervals node, 198–199
- DAT files
 - exporting, 319, 389
 - saving, 322

- data
 - aggregating, 65
 - anonymizing, 154
 - audit, 335
 - exploring, 335
 - preparation, 53
 - storage, 24, 48, 151, 153
 - storage type, 120
 - understanding, 53
- data audit browser
 - Edit menu, 340
 - File menu, 340
 - generating graphs, 349
 - generating nodes, 349
- Data Audit node, 335
 - output tab, 322
 - settings tab, 336
- Data Collection source node, 26–27
 - log files, 28
 - metadata files, 28
- Data Collection survey data
 - importing, 26–27
- data labels
 - in visualizations, 211
- Data Provider Definition, 6
- data quality
 - Data Audit browser, 344
- data sources
 - database connections, 14
- data types, 21, 50, 85, 113, 115
 - instantiation, 118
- database
 - bulk loading, 380
 - support tiers, 11
- database connections
 - defining, 14
 - preset values, 15
- Database export node, 372
 - data source, 372
 - export tab, 372
 - indexing tables, 378
 - mapping source data fields to database columns, 373
 - merge options, 373
 - schema, 375
 - table name, 372
- Database source node, 11
 - query editor, 17
 - selecting tables and views, 16
 - SQL queries, 13
- date recognition, 20, 23
- date storage format, 24, 48
- date/time, 115
- dates
 - setting formats, 128–129
- decile bins, 165
- decimal places
 - display formats, 129
- decimal symbol, 19–20, 128
 - Flat File export node, 382
 - number display formats, 130
- decreasing data, 54–55
- degrees of freedom
 - Matrix node, 328
 - Means node, 363–364
- deleting
 - output objects, 315
 - visualization stylesheets, 227
 - visualization templates, 227
- delimited text data, 18
- delimiters, 19–20, 380
- Derive node
 - conditional, 149
 - converting field storage, 150
 - count, 148
 - flag, 144
 - formula, 144
 - generating from a Binning node, 171
 - generating from automated data preparation, 112
 - generating from bins, 162
 - generating from graphs, 288
 - generating from Web graph links, 263
 - multiple derive, 142
 - overview, 139
 - recoding values, 150
 - set, 146
 - setting options, 141
 - state, 147
- descending order, 69
- directed layout for web graphs , 259
- direction of fields, 50, 113, 126
- discarding
 - fields, 130
- disguising data for use in a model, 154
- display formats
 - currency, 129
 - decimal places, 129
 - grouping symbol, 129
 - numbers, 129
 - scientific, 129
- Distinct node
 - optimization settings, 83
 - overview, 81
 - sorting records, 83
- distribution, 242
- Distribution node , 237
 - appearance tab, 238
 - plot tab, 238
 - using the graph, 239
 - using the table, 239
- documentation, 2
- dodge, 306
- DPD, 6
- dummy coding, 178

- duplicate
 - fields, 71, 131
 - records, 81
- duration computation
 - automated data preparation, 92
- editing graphs
 - size of graphic elements, 295
- editing visualizations, 291
 - adding 3-D effects, 302
 - automatic settings, 292
 - axes, 298
 - categories, 299
 - collapsing categories, 299
 - colors and patterns, 294
 - combining categories, 299
 - dashing, 294
 - excluding categories, 299
 - legend position, 306
 - margins, 296
 - number formats, 297
 - padding, 296
 - panels, 301
 - point aspect ratio, 295
 - point rotation, 295
 - point shape, 295
 - rules, 292
 - scales, 298
 - selection, 292
 - sorting categories, 299
 - text, 293
 - transforming coordinate systems, 302
 - transparency, 294
 - transpose, 301–302
- employee_data.sav data file, 397
- encapsulating nodes, 414
- encoding, 20, 22, 383
- Ensemble node
 - combining scores, 136
 - output fields, 136
- Enterprise View node, 6
- EOL characters, 19
- equal counts
 - Binning node, 165
- estimation period, 190
- evaluating models, 330
- Evaluation node, 269
 - appearance tab, 277
 - business rule, 275
 - hit condition, 275
 - options tab, 275
 - plot tab, 274
 - reading results, 279
 - score expression, 275
 - using the graph, 280
- events
 - creating, 266
- examples
 - Applications Guide, 2
 - overview, 3
- Excel
 - launching from IBM SPSS Modeler, 390
- Excel export node, 389–390
- Excel files
 - exporting, 389–390
- Excel Import node
 - generating from output, 390
- Excel source node, 39
- execution
 - specifying the order of, 429
- expected values
 - Matrix node, 327
- exploring data
 - Data Audit node, 335
- exploring graphs, 281
 - graph bands, 282
 - magic wand, 287
 - marking elements, 287
 - regions, 285
- export decimal places, 129
- export nodes, 371
- exporting
 - output, 319
 - SuperNodes, 430
 - visualization stylesheets, 227
 - visualization templates, 227
- exporting data
 - DAT files, 389
 - flat file format, 382
 - IBM Cognos BI Export node, 36, 385, 387
 - SAS format, 388
 - text, 389
 - to a database, 372
 - to Excel, 389–390
 - to IBM SPSS Statistics, 409
 - XML format, 391
- Expression Builder, 53
- extension
 - derived field, 142
- F* statistic
 - Means node, 363
- faceting
 - in visualizations, 212
- false values, 124
- field attributes, 127
- field derivation formula, 144
- field names, 133
 - anonymizing, 133
 - data export, 372, 382, 389, 409
- field operations nodes, 85
 - generating from a data audit, 349
- Field Reorder node, 206
 - automatic sorting, 208

- custom ordering, 206
 - setting options, 206
- field storage
 - converting, 150
- field types, 50, 113
- fields
 - anonymizing data, 154
 - delimiters, 20
 - deriving multiple fields, 142
 - field and value labels, 50, 113, 122
 - reordering, 206
 - selecting multiple, 143
 - transposing, 182
- Filler node
 - overview, 151
- FILLFACTOR keyword
 - indexing database tables, 379
- Filter node
 - multiple response sets, 134
 - overview, 130
 - setting options, 131
- filtering fields, 76, 130
 - for IBM SPSS Statistics, 410
- First function
 - time series aggregation, 189
- fiscal year
 - Time Intervals node, 196
- Fixed File node
 - automatic date recognition, 23
 - overview, 21
 - setting options, 21
- fixed-field text data, 21
- flag data, 116
- flag type, 115, 124
- Flat File export node, 382
 - export tab, 382
- flat files, 18
- format files, 38
- formats
 - data, 23, 128
- fractional ranks, 168
- free-field text data, 18
- frequencies
 - Binning node, 165
- gains charts, 269, 279
- generating flags, 179, 181
- generating nodes from graphs, 288
 - Balance nodes, 290
 - Derive nodes, 290
 - Filter nodes, 290
 - Reclassify nodes, 290
 - Select nodes, 289
- global values, 368
- graph nodes, 209
 - animation, 210, 212
 - Collection, 245
 - Distribution, 237
 - Evaluation, 269
 - Graphboard, 216
 - Histogram, 242
 - Multiplot, 250
 - overlays, 210
 - panels, 210, 212
 - Plot, 229
 - Time Plot, 266
 - Web, 254
- Graphboard node , 216
 - appearance tab, 224
- graphic elements
 - changing, 303
 - collision modifiers, 305
 - converting, 303
 - types, 303
- graphs
 - 3-D, 215
 - annotations tab, 215
 - axis labels, 307
 - bands, 282
 - collections, 245
 - copying, 311
 - default color scheme, 309
 - deleting regions, 286
 - distributions, 237
 - evaluation charts, 269
 - exploring, 281
 - exporting, 311
 - footnote, 307
 - from Graphboard, 216
 - generating from a data audit, 349
 - generating nodes, 288
 - histograms, 242
 - multiplot, 250
 - output tabs, 215
 - plots, 229
 - printing, 311
 - regions, 285
 - rotating a 3-D image, 215
 - saving, 311
 - saving edited layouts, 309
 - saving layout changes, 309
 - saving output, 322
 - size of graphic elements, 295
 - stylesheet, 309
 - time series, 266
 - title, 307
 - webs, 254
- grouping symbol
 - number display formats, 130
- grouping values, 239
- handling missing values, 85
- hassubstring function, 145

-
- HDATA format
 - Data Collection source node, 26
 - helper applications, 369
 - Histogram node, 242
 - appearance tab, 243
 - plot tab, 242–243
 - using the graph, 244
 - History node, 205
 - overview, 205
 - hits
 - evaluation chart options, 275
 - holdouts
 - time series modeling, 190
 - hourly measurements
 - Time Intervals node, 200–201
 - HTML
 - saving output, 323
 - HTML output
 - Report node, 366
 - view in browser, 319

 - IBM Cognos BI Export node, 36, 385, 387
 - IBM Cognos BI source node, 34, 36
 - IBM SPSS Collaboration and Deployment Services Repository
 - connecting to, 6
 - using as location for visualization templates and stylesheets, 226
 - IBM SPSS Data Collection export node, 383
 - IBM SPSS Data Collection source node, 33
 - database connection settings, 31–32
 - label types, 30
 - language, 30
 - multiple response sets, 32
 - IBM SPSS Modeler, 1
 - documentation, 2
 - IBM SPSS Statistics
 - launching from IBM SPSS Modeler, 369, 405, 409
 - license location, 369
 - valid field names, 410
 - IBM SPSS Statistics data files
 - importing survey data, 28
 - IBM SPSS Statistics models, 401
 - about, 401
 - advanced nugget details, 403
 - model nugget, 403
 - model options, 402
 - IBM SPSS Statistics nodes, 395
 - IBM SPSS Statistics Output node
 - Output tab, 408
 - IBM SPSS Text Analytics, 2
 - if-then-else statements, 149
 - imbalanced data, 63
 - importance
 - comparing means, 361
 - Means node, 363–364
 - importing
 - SuperNodes, 430
 - visualization stylesheets, 227
 - visualization templates, 227
 - In2data databases
 - importing, 28
 - incomplete records, 74
 - indexing database tables, 378
 - inner join, 72
 - instantiation, 50, 113, 115, 118–119
 - source node, 51
 - integer ranges, 122
 - integer storage format, 24, 48
 - intervals
 - time series data, 186
 - interventions
 - creating, 266

 - jitter, 306
 - jittering, 234
 - joining datasets, 80
 - joins, 71–72, 74
 - partial outer, 75
 - justification
 - for fields, 128

 - key fields, 66, 179
 - key method, 71
 - key value for aggregation, 66

 - label fields
 - labeling records in output, 126
 - label types
 - IBM SPSS Data Collection source node, 30
 - labels, 124
 - exporting, 389, 410
 - importing, 38, 396
 - in visualizations, 211
 - specifying, 50, 113, 120, 122–124
 - lagged data, 205
 - language
 - IBM SPSS Data Collection source node, 30
 - large databases, 53
 - performing a data audit, 335
 - Last function
 - time series aggregation, 189
 - legal notices, 432
 - legend
 - position, 306
 - lift charts, 269, 279
 - line plots, 229, 250
 - links
 - Web node, 257
 - locally weighted least squares regression
 - Plot node, 232

- locking SuperNodes, 419–420
- LOESS smoother
 - Plot node, 232
- lowess smoother *See* LOESS smoother
 - Plot node, 232
- magic wand in graphs, 287
- main dataset, 81
- managers
 - outputs tab, 315
- mapping fields, 373
- market research data
 - Data Collection source node, 26
 - IBM SPSS Data Collection source node, 32
 - importing, 27, 33
- marking elements, 285, 287
- matrix browser
 - Generate menu, 328
- Matrix node, 325
 - appearance tab, 327
 - column percentages, 327
 - cross-tabulation, 327
 - highlighting, 327
 - output browser, 328
 - output tab, 322
 - row percentages, 327
 - settings tab, 326
 - sorting rows and columns, 327
- matrix output
 - saving as text, 323
- Max function
 - time series aggregation, 189
- maximum
 - Set Globals node, 368
 - statistics output, 356
- maximum value for aggregation, 66
- MDD documents
 - importing, 28
- mean
 - Binning node, 169
 - Set Globals node, 368
 - statistics output, 356
- Mean function
 - time series aggregation, 189
- Mean of most recent function
 - padding time series, 190
- mean value for aggregation, 66
- mean value for records, 65
- mean/standard deviation
 - used to bin fields, 169
- means
 - comparing, 359–360, 362
- Means node, 359
 - importance, 361
 - independent groups, 359
 - output browser, 362–363
 - output tab, 322
- paired fields, 360
- measurement level, 50, 113
 - changing in visualizations, 218
 - defined, 115
 - in visualizations, 220
- median
 - statistics output, 356
- member (SAS import)
 - setting, 38
- Merge node, 72
 - filtering fields, 76
 - optimization settings, 78
 - overview, 71
 - setting options, 74
 - tagging fields, 77
- merge options, database export, 373
- metadata, 50, 113, 120
 - Data Collection source node, 26–27
- Microsoft Excel source node, 39
- Min function
 - time series aggregation, 189
- minimum
 - Set Globals node, 368
 - statistics output, 356
- minimum value for aggregation, 66
- minute increments
 - Time Intervals node, 201–202
- missing values, 85, 120, 124
 - filling, 344
 - handling, 344
 - in Aggregate nodes, 65
 - in Matrix tables, 326
- mode
 - statistics output, 356
- Mode function
 - time series aggregation, 189
- model evaluation, 269
- model options
 - Statistics Model node, 402
- model view
 - in automated data preparation, 99
- modeling roles
 - specifying for fields, 50, 113, 126
- models
 - anonymizing data for, 154
- modifying data values, 139
- monthly data
 - Time Intervals node, 197
- Most recent function
 - padding time series, 190
- multiple category sets, 134
- multiple derive, 142
- multiple dichotomy sets, 134
- multiple fields
 - selecting, 143
- multiple inputs, 71

- multiple response sets
 - Data Collection source node, 26–27
 - defining, 134
 - deleting, 134
 - IBM SPSS Data Collection source node, 32–33
 - IBM SPSS Statistics source node, 397
 - in visualizations, 220
 - multiple category sets, 134
 - multiple dichotomy sets, 134
- Multiplot node, 250
 - appearance tab, 253
 - plot tab, 251
 - using the graph, 254
- natural order
 - altering, 206
- network layout for web graphs, 259
- node properties, 426
- nominal data, 116, 123
- nonrandom samples, 55–56
- normalize continuous target, 95, 112
- normalize values
 - graph nodes, 251, 267
- null values
 - in Matrix tables, 326
 - mixed data, 25, 48
- nulls, 120, 344
 - in Matrix tables, 326
- number display formats, 129
- ODBC
 - bulk loading via, 380
 - connection for IBM Cognos BI Export node, 387
 - Database source node, 11
- ODBC export node. *See* Database export node, 372
- one-way ANOVA
 - Means node, 359
- opening
 - output objects, 315
- optimal binning, 170
- options
 - IBM SPSS Statistics, 369
- Oracle, 11
- order merging, 71
- order of execution
 - specifying, 429
- order of input data, 77
- ordering data, 69, 206
- ordinal data, 116, 123
- outer join, 72
- output
 - exporting, 319
 - generating new nodes from, 316
 - HTML, 319
 - printing, 316
 - saving, 316
- output files
 - saving, 322
- output formats, 322
- output manager, 315
- output nodes, 314, 321, 325, 330, 335, 355, 364, 368, 405
 - output tab, 322
 - publish to web, 316
- overlays for graphs, 210
- overwriting database tables, 372
- p* value
 - importance, 361
- padding time series data, 188
- palettes
 - displaying, 292
 - hiding, 292
 - moving, 292
- panel graph overlay, 210, 212
- paneling, 210, 212
 - in visualizations, 212
- parallel processing
 - Aggregate node, 67
 - merging, 78
 - sorting, 71
- parameters
 - node properties, 426
 - setting for SuperNodes, 424
 - SuperNodes, 424, 426
- partial joins, 72, 75
- partition fields, 50, 113, 126, 176–177
- Partition node, 176–177
- partitioning data, 176–177
 - Analysis node, 330
 - evaluation charts, 275
- Pearson chi-square
 - Matrix node, 328
- Pearson correlations
 - Means node, 364
 - statistics output, 356
- percentile bins, 165
- performance
 - Aggregate node, 67
 - Binning nodes, 171
 - Derive nodes, 171
 - merging, 78
 - sampling data, 55
 - sorting, 71
- performance evaluation statistic, 330
- period, 128
- periodicity
 - time series data, 186
- periods
 - Time Intervals node, 194
- Plot node, 229
 - appearance tab, 235
 - options tab, 234
 - plot tab, 231

- using the graph, 236
- plotting associations, 254
- point plots, 229, 250
- polar coordinates, 302
- preset values, database connection, 15
- primary key fields
 - Database export node, 377
- printing output, 316
- profit charts, 269, 279
- propensity scores
 - balancing data, 64
- properties
 - for fields, 128
 - node, 426
- publish to web, 316
- Python
 - bulk loading scripts, 380

- quality browser
 - generating filter nodes, 347
 - generating select nodes, 348
- quality report
 - Data Audit browser, 344
- Quancept data
 - importing, 28
- Quantum data
 - importing, 28
- Quanvert databases
 - importing, 28
- quarterly data
 - Time Intervals node, 196
- quartile bins, 165
- queries
 - Database source node, 11, 13
- query editor
 - Database source node, 17
- quintile bins, 165
- quotation marks
 - importing text files, 20
- quotes
 - for database export, 372

- random seed value
 - sampling records, 59, 178
- range
 - statistics output, 356
- ranges, 115
 - missing values, 120
- rank cases, 168
- real ranges, 122
- real storage format, 24, 48
- recency
 - setting relative date, 68
- Reclassify node, 159, 161
 - generating from a distribution, 239
 - overview, 158, 162
- recode, 158–159, 162

- record
 - counts, 66
 - labels, 126
 - length, 21
- record operations nodes, 53
 - Time Intervals node, 186
- records
 - merging, 71
 - transposing, 182
- regions in graphs, 285
- renaming
 - fields for export, 410
 - visualization stylesheets, 227
 - visualization templates, 227
- renaming output objects, 315
- replacing field values, 151
- report browser, 367
- Report node, 364
 - output tab, 322
 - template tab, 365
- reports
 - saving output, 322
- residuals
 - Matrix node, 327
- response charts, 269, 279
- Restructure node, 180–181
 - with Aggregate node, 181
- restructuring data, 180
- revenue
 - evaluation charts, 275
- RFM Aggregate node
 - independent binning, 67, 173
 - nested binning, 67, 173
 - overview, 67
 - setting options, 68
- RFM Analysis node
 - binning values, 175
 - independent binning, 67, 173
 - nested binning, 67, 173
 - overview, 173
 - settings, 174
- ROI
 - charts, 269, 279
- roles
 - specifying for fields, 50, 113, 126
- rolling up time series data, 188
- rotating 3-D graphs, 215
- row-wise binding, 380

- Sample node
 - clustered samples, 55–56, 59
 - nonrandom samples, 55–56
 - random samples, 55–56
 - Sample Sizes for Strata, 61
 - sampling frame, 55
 - stratified samples, 55–56, 59, 61
 - systematic samples, 55–56

- weighted samples, 59
- sampling data, 61
- sampling frame, 55
- SAS
 - setting import options, 38
- SAS export node, 388–389
- SAS source node
 - .sd2 (SAS) files, 37
 - .ssd (SAS) files, 37
 - .tpt (SAS) files, 37
 - transport files, 37
- .sav files, 396
- saving
 - output, 316
 - output objects, 315, 322
- scale factors, 64
- scatterplots, 229, 250
- scenario, 6
- schema
 - Database export node, 375
- scientific display format, 129
- scoring
 - evaluation chart options, 275
- scripting
 - SuperNodes, 429
- .sd2 (SAS) files, 37
- searching
 - table browser, 324
- second increments
 - Time Intervals node, 203–204
- seed value
 - sampling and records, 59, 178
- Select node
 - generating from graphs, 288
 - generating from Web graph links, 263
 - overview, 54
- selecting rows (cases), 54
- selecting values, 282, 285, 287
- Set Globals node, 368
 - settings tab, 368
- set random seed
 - sampling records, 59, 178
- Set to Flag node, 178–179
- set type, 115
- sets
 - converting to flags, 178, 180
 - transforming, 159, 161
- shape
 - in visualizations, 210
- shape graph overlay, 210
- significance
 - correlation strength, 355
- size
 - in visualizations, 211
- size graph overlay, 210
- .slb files, 430
- smoother
 - Plot node, 232
- Sort node
 - optimization settings, 70
 - overview, 69
- sorting
 - Distinct node, 83
 - fields, 206
 - presorted fields, 70, 83
 - records, 69
- source nodes
 - Database source node, 11
 - Enterprise View node, 6
 - Excel source node, 39
 - Fixed File node, 21
 - IBM Cognos BI source node, 34, 36
 - instantiating types, 51
 - overview, 5
 - SAS source node, 37
 - Statistics File node, 396
 - User Input node, 44–45
 - Variable File node, 18
 - XML source node, 40
- SourceFile variables
 - IBM SPSS Data Collection source node, 30
- SPSS Modeler Server, 1
- SQL queries
 - Database source node, 11, 13, 17
- .ssd (SAS) files, 37
- stack, 305
- standard deviation
 - Binning node, 169
 - Set Globals node, 368
 - statistics output, 356
- standard deviation for aggregation, 66
- standard error of mean
 - statistics output, 356
- statistics
 - Data Audit node, 335
 - descriptions, 219, 304
 - editing in visualizations, 303
 - Matrix node, 326
- statistics browser
 - Generate menu, 356
 - generating filter nodes, 358
 - interpreting, 356
- Statistics Export node, 409
 - Export tab, 409
- Statistics File node, 396
- Statistics node, 355
 - correlation labels, 355
 - correlations, 355
 - output tab, 322
 - settings tab, 355
 - statistics, 355
- Statistics Output node, 405
 - Syntax tab, 406

- Statistics Transform node, 397
 - allowable syntax, 399
 - setting options, 398
 - Syntax tab, 398
- storage, 120
 - converting, 150–151, 153
- storage formats, 23
- stratified samples, 55–56, 59, 61
- string storage format, 24, 48
- stylesheets
 - deleting, 227
 - exporting, 227
 - importing, 227
 - renaming, 227
- sum
 - Set Globals node, 368
 - statistics output, 356
- Sum function
 - time series aggregation, 189
- summary data, 65
- summary statistics
 - Data Audit node, 335
- summed values, 66
- SuperNode parameters, 424, 426
- SuperNodes, 412
 - creating, 414
 - creating caches for, 428
 - editing, 422
 - loading, 430
 - locking, 419–420
 - nesting, 416
 - password protection, 419–421
 - process SuperNodes, 413
 - saving, 430
 - scripting, 429
 - setting parameters, 424
 - source SuperNodes, 412
 - terminal SuperNodes, 414
 - types of, 412
 - unlocking, 420
 - using comments with, 423
 - zooming in, 422
- supervised binning, 170
- survey data
 - Data Collection source node, 26
 - importing, 27, 32–33
- Surveycraft data
 - importing, 28
- syntax tab
 - Statistics Output node, 406
- synthetic data
 - User Input node, 44
- system missing values
 - in Matrix tables, 326
- System variables
 - IBM SPSS Data Collection source node, 30
- system-missing values, 344
- systematic samples, 55–56
- t* test
 - independent samples, 359
 - Means node, 359–360, 364
 - paired samples, 360
- table browser
 - Generate menu, 324
 - reordering columns, 320, 324
 - searching, 324
 - selecting cells, 320, 324
- Table node, 321
 - column justification, 128
 - column width, 128
 - format tab, 128
 - output settings, 321
 - output tab, 322
 - settings tab, 321
- tables
 - joining, 72
 - saving as text, 323
 - saving output, 322
- tabular output
 - reordering columns, 320
 - selecting cells, 320
- tags, 71, 77
- templates
 - deleting, 227
 - exporting, 227
 - importing, 227
 - renaming, 227
 - Report node, 365
- test samples
 - partitioning data, 176–177
- text
 - data, 18, 21
 - delimited, 18
 - encoding, 20, 22, 383
- text files, 18
 - exporting, 389
- thresholds
 - viewing bin thresholds, 171
- tiers, database support, 11
- ties
 - Binning node, 165
- tiles
 - Binning node, 165
- time
 - setting formats, 128
- time formats, 129
- Time Intervals node, 187–188, 190
 - overview, 186
- Time Plot node, 266
 - appearance tab, 268
 - plot tab, 267
 - using the graph, 269
- time series, 205

- time series data
 - aggregating, 186, 188
 - building from data, 188
 - defining, 186–188, 190
 - estimation period, 190
 - holdouts, 190
 - intervals, 187
 - labeling, 186–188, 190
 - padding, 186, 188
- time storage format, 24, 48
- TimeIndex field
 - Time Intervals node, 188
- TimeLabel field
 - Time Intervals node, 188
- timestamp, 115
- timestamp storage format, 24, 48
- .*tpt* (SAS) files, 37
- trademarks, 433
- training samples
 - balancing, 64
 - partitioning data, 176–177
- Transform node, 349
- transformations
 - reclassify, 158, 162
 - recode, 158, 162
- transparency
 - in visualizations, 211
- transparency in graphs, 210
- transport files
 - SAS source node, 37
- Transpose node, 182
 - field names, 182
 - numeric fields, 182
 - string fields, 182
- transposing data, 182
- Triple-S data
 - importing, 28
- True if any true function
 - time series aggregation, 189
- true values, 124
- truncating field names, 131, 133
- type, 23
- type attributes, 127
- Type node
 - blank handling, 120
 - clearing values, 50
 - column justification, 128
 - column width, 128
 - continuous data, 122
 - copying types, 127
 - flag field type, 124
 - format tab, 128
 - nominal data, 123
 - ordinal data, 123
 - overview, 113
 - setting modeling role, 126
 - setting options, 115, 117
- typeless data, 116
- unbiased data, 63
- undefined values, 74
- UNIQUE keyword
 - indexing database tables, 379
- unique records, 81
- unlocking SuperNodes, 420
- unused field exclusion
 - automated data preparation, 91
- usage type, 23, 115
- User Input node
 - overview, 44
 - setting options, 45
- user-missing values, 344
 - in Matrix tables, 326
- UTF-8 encoding, 20, 22, 383
- validation samples
 - partitioning data, 176–177
- value labels
 - Statistics File node, 396
- values
 - field and value labels, 50, 113, 120
 - reading, 119
 - specifying, 120
- Variable File node, 18
 - automatic date recognition, 20
 - setting options, 19
- variable labels
 - Statistics Export node, 409
 - Statistics File node, 396
- variable names
 - data export, 372, 382, 389, 409
- variable types
 - in visualizations, 220
- variance
 - statistics output, 356
- VDATA format
 - Data Collection source node, 26
- viewing
 - HTML output in browser, 319
- vingtile bins, 165
- visualization
 - graphs and charts, 209
- visualization stylesheets
 - applying, 310
 - deleting, 227
 - exporting, 227
 - importing, 227
 - location, 225
 - renaming, 227
- visualization templates
 - deleting, 227
 - exporting, 227
 - importing, 227
 - location, 225

- renaming, 227
- visualizations
 - axes, 298
 - categories, 299
 - colors and patterns, 294
 - copying, 306
 - dashings, 294
 - edit mode, 291
 - editing, 291
 - legend position, 306
 - margins, 296
 - number formats, 297
 - padding, 296
 - panels, 299, 301
 - point aspect ratio, 295
 - point rotation, 295
 - point shape, 295
 - scales, 298
 - text, 293
 - transforming coordinate systems, 302
 - transparency, 294
 - transpose, 299, 301–302
- Web node , 254
 - adjusting points, 262
 - adjusting thresholds, 264
 - appearance tab, 260
 - change layout, 263
 - defining links, 257
 - links slider, 263
 - options tab, 257
 - plot tab, 256
 - slider, 263
 - using the graph, 261
 - web summary, 265
- weekly data
 - Time Intervals node, 197
- weighted samples, 59
- weights
 - evaluation charts, 275
- worksheets
 - importing from Excel, 39
- XLS files
 - exporting, 390
- XML export node, 391
- XML output
 - Report node, 366
- XML source node, 40
- XPath syntax, 40
- yearly data
 - Time Intervals node, 195
- zooming, 422