IBM SPSS Modeler 18.0 소스, 프로세스 및 출력 노드

IBM

참고

이 정보와 이 정보가 지원하는 제품을 사용하기 전에, 399 페이지의 『주의사항』의 정보를 읽으십시오.

제품 정보

이 개정판은 새 개정판에 별도로 명시하지 않는 한, IBM SPSS Modeler의 버전 18, 릴리스 0, 수정 0 및 모든 후속 릴리스와 수정에 적용됩니다.

목차

서론

IBM® SPSS® Modeler는 IBM Corp. 엔터프라이즈 중심의 데이터 마이닝 워크벤치입니다. SPSS Modeler는 상세한 데이터 이해를 통해 조직이 고객과 시민과의 관계를 향상시킬 수 있도록 도움을 줍니다. 조직은 SPSS Modeler에서 확보한 통찰력을 통해 수익 창출이 가능한 고객을 보유하고, 교차 판매 기회를 식별하고, 새 고객을 모으고, 사기 행위를 적발하고, 위험을 줄이고, 정부 서비스 지원을 향상시킬 수 있습니다.

SPSS Modeler의 시각적 인터페이스를 통해 사용자는 보다 쉽게 비즈니스에 특정한 전문 지식을 적용할 수 있으므로, 더 강력한 예측 모델을 생성하고 솔루션 출시 시점을 단축할 수 있습니다. SPSS Modeler에서는 예측, 분류, 세분화, 연관 발견 알고리즘과 같은 많은 모델링 기법을 제공합니다. 모델을 작성하면 IBM SPSS Modeler Solution Publisher에서 의사결정자 또는 데이터베이스까지 엔터프라이즈 범위로 모델을 전달할 수 있습니다.

IBM Business Analytics 소개

IBM Business Analytics 소프트웨어는 의사 결정자가 비즈니스 성능을 개선하기 위해 신뢰하는 완전하고 일관되며 정확한 정보를 제공합니다. 비즈니스 지능, 예측 분석, 금융 성과와 전략 관리 및 분석 응용 프로그램의 종합 포트폴리오는 현재 성과와 앞으로의 결과를 예측하는 능력에 분명하고 즉각적이면서 실행 가능한 통찰력을 제공합니다. 풍부한 업계 솔루션, 입증된 사례 및 전문 서비스가 결합되어 어떠한 크기의 조직이라도 생산성을 극대화하고 자신있는 자동 결정을 내릴 수 있으며 더 나은 결과를 가져올 수 있습니다.

이 포트폴리오의 일부인 IBM SPSS Predictive Analytics 소프트웨어를 통해 조직은 미래의 사건을 예측하고 더 나은 비즈니스 결과를 얻기 위한 통찰력에 대해 적극적인 조치를 할 수 있습니다. 전 세계의 기업, 정부 및 학계 고객들은 고객을 매료시키고 유지하며 성장하게 만드는 동시에 불공정 행위를 줄이고 위험을 낮추는 IBM SPSS 기술의 경쟁 이점을 활용합니다. 일상 업무에서 IBM SPSS 소프트웨어를 활용한다면 예측형 기업으로 거듭날 수 있습니다. 즉 비즈니스 목표 달성을 위해 의사 결정의 방향을 정하고 이를 자동화하며 측정 가능한 경쟁 우위를 달성할 수 있습니다. 자세한 내용을 보거나 담당자에게 문의하려면 http://www.ibm.com/spss 사이트를 방문하십시오.

기술 지원

기술 지원은 유지 관리 고객에게 제공됩니다. IBM Corp. 제품 사용 및 지원된 하드웨어 환경 중 하나에 대해 설치하는 데 도움이 필요한 경우 기술 지원부로 문의하십시오. 기술 지원에 문의하려면 IBM Corp. 웹 사이트(http://www.ibm.com/support)를 참조하십시오. 지원을 요청하려면 본인의 신상과 소속 조직(회사) 및 지원 동의서를 제시해야 합니다.

제 1 장 IBM SPSS Modeler 정보

IBM SPSS Modeler는 비즈니스 전문 지식을 사용하여 예측 모형을 신속하게 개발하고 이를 비즈니스 운영에 배포하여 의사결정의 정확성을 향상시켜주는 데이터 마이닝 도구 세트입니다. 산업 표준 CRISP-DM 모델을 중심으로 디자인된 IBM SPSS Modeler는 데이터에서 보다 나아진 비즈니스 결과에 이르는 전체 데이터 마이닝 프로세스를 지원합니다.

IBM SPSS Modeler는 기계 학습, 인공지능 및 통계로부터 취한 다양한 모델링 방법을 제공합니다. 모델링 팔레트에서 사용할 수 있는 이러한 방법을 통해 데이터로부터 새로운 정보를 얻어서 예측 모형을 개발할 수 있습니다. 각각의 방법은 그것만의 장점이 있으며 특정한 문제점 유형에 가장 적합합니다.

SPSS Modeler는 독립형 제품으로 구매하거나 SPSS Modeler Server와 통합하여 클라이언트로 사용할 수 있습니다. 다음 절에 요약된 바와 같이 여러가지 추가 옵션도 사용할 수 있습니다. 자세한 정보는 http://www.ibm.com/software/analytics/spss/products/modeler/의 내용을 참조하십시오.

IBM SPSS Modeler 제품

IBM SPSS Modeler 제품군 및 연관 소프트웨어는 다음으로 구성됩니다.

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Collaboration and Deployment Services용 IBM SPSS Modeler Server 어댑터

IBM SPSS Modeler

SPSS Modeler는 개인용 컴퓨터에 설치하여 실행되는 기능적으로 완전한 버전의 제품입니다. 로컬 모드에서 독립형 제품으로 SPSS Modeler를 실행하거나 대형 데이터 세트에 대한 성능 향상을 위해 분산 모드에서 IBM SPSS Modeler Server와 함께 사용할 수 있습니다.

SPSS Modeler를 사용하여 프로그래밍하지 않고 신속하게 직관적으로 정확한 예측 모델을 작성할 수 있습니다. 고유한 시각적 인터페이스를 사용하면 데이터 마이닝 프로세스를 쉽게 시각화할 수 있습니다. 제품에 포함된 고급 분석 지원을 통해 데이터에서 이전에 숨겨진 패턴과 추세를 발견할 수 있습니다. 결과를 모델링하고 결과에 영향을 주는 요인을 이해하여 비즈니스 기회를 활용하고 위험을 줄일 수 있습니다.

SPSS Modeler는 두 개의 에디션(SPSS Modeler Professional과 SPSS Modeler Premium)으로 사용할 수 있습니다. 자세한 정보는 3 페이지의 『IBM SPSS Modeler 에디션』의 내용을 참조하십시오.

IBM SPSS Modeler Server

SPSS Modeler는 클라이언트/서버 설계를 사용하여 자원 집약적 작업에 대한 요청을 강력한 서버 소프트웨어로 분배하여 대형 데이터 세트에 대한 성능을 향상시킵니다.

SPSS Modeler Server는 하나 이상의 IBM SPSS Modeler 설치와 함께 서버 호스트의 분산 분석 모드에서 계속해서 실행되는 별도로 라이센스가 부여된 제품입니다. 이런 방법으로 클라이언트 컴퓨터로 데이터를 다운로드하지 않고 서버에서 메모리 집약적 작업을 수행할 수 있기 때문에 SPSS Modeler Server는 대형 데이터 세트에 대한 우수한 성능을 제공합니다. 또한 IBM SPSS Modeler Server는 SQL 최적화 및 In-Database 모델링 기능에 대한 지원을 제공하여 성능 및 자동화의 이점도 추가로 제공합니다.

IBM SPSS Modeler Administration Console

Modeler Administration Console은 옵션 파일을 통해서도 구성 가능한 다수의 SPSS Modeler Server 구성 옵션을 관리하기 위한 그래픽 애플리케이션입니다. 이 애플리케이션은 SPSS Modeler Server 설치를 모니터하고 구성하기 위한 콘솔 사용자 인터페이스를 제공하며 현재 SPSS Modeler Server 고객에게 무료로 제공됩니다. 이 애플리케이션은 Windows 컴퓨터에만 설치할 수 있지만 지원되는 플랫폼에 설치된 서버를 관리할 수 있습니다.

IBM SPSS Modeler Batch

데이터 마이닝은 일반적으로 대화식 처리인 반면, 그래픽 사용자 인터페이스가 없어도 명령행에서 SPSS Modeler를 실행할 수 있습니다. 예를 들어, 사용자 개입 없이 수행할 장기 실행 또는 반복 작업이 있습니다. SPSS Modeler Batch는 정규 사용자 인터페이스에 대한 액세스 없이 SPSS Modeler의 전체 분석 기능에 대한 지원을 제공하는 특수 버전의 제품입니다. SPSS Modeler Batch를 사용하려면 SPSS Modeler Server가 필요합니다.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher는 외부 런타임 엔진을 통해 실행하거나 외부 애플리케이션에 포함될 수 있는 SPSS Modeler 스트림의 패키지 버전을 작성할 수 있게 하는 도구입니다. 이런 방법으로 SPSS Modeler가 설치되지 않는 환경에 사용할 수 있도록 전체 SPSS Modeler 스트림을 출판하고 배포할 수 있습니다. SPSS Modeler Solution Publisher는 별도의 라이센스가 필요한 IBM SPSS Collaboration and Deployment Services - Scoring 서비스의 일부로 분배됩니다. 이 라이센스가 있으면 출판된 스트림을 실행할 수 있게 하는 SPSS Modeler Solution Publisher Runtime을 수신합니다.

SPSS Modeler Solution Publisher에 대한 자세한 정보는 IBM SPSS Collaboration and Deployment Services 문서를 참조하십시오. IBM SPSS Collaboration and Deployment Services Knowledge Center에는 "IBM SPSS Modeler Solution Publisher" 및 "IBM SPSS Analytics Toolkit" 섹션이 포함되어 있습니다.

IBM SPSS Collaboration and Deployment Services용 IBM SPSS Modeler Server 어댑터

SPSS Modeler와 SPSS Modeler Server가 IBM SPSS Collaboration and Deployment Services 리포지토리와 상호작용할 수 있게 하는 IBM SPSS Collaboration and Deployment Services용 어댑터를 상당수 사

용할 수 있습니다. 이런 방법으로 리포지토리에 배포된 SPSS Modeler 스트림을 여러 사용자가 공유하거나 씬 클라이언트 애플리케이션 IBM SPSS Modeler Advantage에서 액세스할 수 있습니다. 리포지토리를 호스팅하는 시스템에 어댑터를 설치하십시오.

IBM SPSS Modeler 에디션

SPSS Modeler는 다음 에디션으로 사용할 수 있습니다.

SPSS Modeler Professional

SPSS Modeler Professional은 CRM 시스템, 인구 통계, 구매 동작, 판매 데이터에서 추적된 동작 및 상호작용과 같은 대부분의 구조화된 데이터 유형에 대한 작업을 하는 데 필요한 모든 도구를 제공합니다.

SPSS Modeler Premium

SPSS Modeler Premium은 특수 데이터(예: 엔티티 분석 또는 소셜 네트워킹에 사용된 데이터) 및 비구조적 텍스트 데이터에 대한 작업을 하도록 SPSS Modeler Professional을 확장하는 별도로 라이센스가 부여된 제품입니다. SPSS Modeler Premium은 다음 구성요소로 구성됩니다.

IBM SPSS Modeler Entity Analytics는 IBM SPSS Modeler 예측 분석에 추가로 차원을 제공합니다. 예측 분석은 과거 데이터로부터 향후의 활동을 예측하는 것을 시도하는 반면, 엔티티 분석은 레코드 자체 내에서 ID 충돌을 해결함으로써 현재 데이터의 일관성 향상에 중점을 둡니다. ID는 모호성이 있을 수 있는 개별 조직, 개체 또는 다른 엔티티의 ID입니다. ID 확인은 고객 관계 관리, 사기 발견, 자금 세탁 그리고 자국 및 국제 보안을 비롯해 필드의 수에서 중요할 수 있습니다.

IBM SPSS Modeler Social Network Analysis는 관계에 대한 정보를 개인 및 그룹의 사회 행동을 특징화하는 필드로 변환합니다. IBM SPSS Modeler Social Network Analysis는 소셜 네트워크에 깔린 관계를 설명하는 데이터를 사용하여 네트워크에서 다른 사람의 행동에 영향을 미치는 사회 리더를 식별합니다. 또한 어떤 사람이 다른 네트워크 참가자에 의한 영향을 가장 많이 받는지 파악할 수 있습니다. 이러한 결과를 다른 측정과 결합함으로써 예측 모형의 토대인 개인에 대한 복합적인 프로파일을 만들 수 있습니다. 이 사회 정보가 포함된 모델은 그렇지 않은 모델보다 성능이 우수합니다.

IBM SPSS Modeler Text Analytics는 고급 언어 기술 및 자연어 처리(NLP)를 사용하여 다양한 비정형 텍스트 데이터를 빠르게 처리하고, 주요 개념을 추출 및 구성하고, 이러한 개념을 범주로 분류합니다. 추출된 개념과 범주는 인구 통계와 같은 기존의 구조화된 데이터와 결합할 수 있고 보다 나은 집중적인 의사결정을 내리기 위해 전체 IBM SPSS Modeler 데이터 마이닝 세트를 사용하여 모델링에 적용할 수 있습니다.

IBM SPSS Modeler 문서

SPSS Modeler의 도움말 메뉴에서 온라인 도움말 형식의 문서를 사용할 수 있습니다. 여기에는 SPSS Modeler, SPSS Modeler Server에 대한 문서는 물론 애플리케이션 안내서(자습서라고도 함) 및 기타 지원 자료도 포함됩니다.

설치 지시사항을 포함하여 각 제품에 대한 전체 문서는 제품 다운로드의 일부로 별도의 압축 폴더에 PDF 형식으로 제공됩니다. PDF 문서는 http://www.ibm.com/support/docview.wss?uid=swg27046871 웹 페이지에서 다운로드할 수도 있습니다.

두 형식의 문서는 SPSS Modeler Knowledge Center(http://www-01.ibm.com/support/knowledgecenter/SS3RA7_18.1.0)에서도 사용할 수 있습니다.

SPSS Modeler Professional 문서

SPSS Modeler Professional 문서 스위트(설치 지시사항은 제외)는 다음과 같습니다.

- **IBM SPSS Modeler 사용자 안내서.** 데이터 스트림 작성, 결측값 처리, CLEM 표현식 작성, 프로젝트 및 보고서에 대한 작업, IBM SPSS Collaboration and Deployment Services 또는 IBM SPSS Modeler Advantage에 배포하기 위한 스트림 패키지 방법을 포함하여 SPSS Modeler 사용에 대한 일반 소개입니다.
- **IBM SPSS Modeler 소스, 프로세스 및 출력 노드.** 여러 형식의 데이터를 읽고 처리하며, 출력하는 데 사용하는 모든 노드에 대한 설명입니다. 실질적으로 이는 모델링 노드 이외의 모든 노드를 의미합니다.
- **IBM SPSS Modeler 모델링 노드.** 데이터 마이닝 모델을 작성하는 데 사용하는 모든 노드에 대한 설명입니다. IBM SPSS Modeler는 기계 학습, 인공지능 및 통계로부터 취한 다양한 모델링 방법을 제공합니다.
- **IBM SPSS Modeler 알고리즘 안내서.** IBM SPSS Modeler에서 사용하는 모델링 방법의 수학적 토대에 대한 설명입니다. 이 안내서는 PDF 형식으로만 사용할 수 있습니다.
- **IBM SPSS Modeler 애플리케이션 안내서.** 이 안내서의 예제는 특정 모델링 방법과 기법을 중점적으로 간략히 소개합니다. 이 안내서의 온라인 버전을 도움말 메뉴에서도 사용할 수 있습니다. 자세한 정보는 5 페이지의 『애플리케이션 예제』 주제를 참조하십시오.
- **IBM SPSS Modeler Python 스크립팅 및 자동화.** 노드와 스트림을 조작하는 데 사용할 수 있는 특성을 포함하여 Python 스크립팅을 통한 시스템 자동화에 대한 정보입니다.
- **IBM SPSS Modeler 배포 안내서.** IBM SPSS Collaboration and Deployment Services Deployment Manager에서 작업 처리 단계로 IBM SPSS Modeler 스트림 실행에 대한 정보입니다.
- **IBM SPSS Modeler CLEF 개발자 안내서.** CLEF는 데이터 처리 루틴 또는 모델링 알고리즘과 같은 써드파티 프로그램을 IBM SPSS Modeler의 노드로 통합하는 기능을 제공합니다.
- **IBM SPSS Modeler In-Database 마이닝 안내서.** 데이터베이스의 능력을 사용하여 성능을 향상시키고 써드파티 알고리즘을 통해 분석 기능 범위를 확장하는 방법에 대한 정보입니다.
- **IBM SPSS Modeler Server 관리 및 성능 안내서.** IBM SPSS Modeler Server 구성 및 관리 방법에 대한 정보입니다.
- **IBM SPSS Modeler 관리 콘솔 사용자 안내서.** IBM SPSS Modeler Server 모니터링 및 구성을 위한 콘솔 사용자 인터페이스 설치 및 사용에 대한 정보입니다. 콘솔은 Deployment Manager 애플리케이션에 플러그인으로 구현됩니다.
- **IBM SPSS Modeler CRISP-DM 안내서.** SPSS Modeler에서 데이터 마이닝에 CRISP-DM 방법론을 사용하기 위한 단계별 안내서입니다.

- **IBM SPSS Modeler Batch 사용자 안내서.** 일괄처리 모드 실행 및 명령행 인수 세부사항을 포함하여 일괄처리 모드에서 IBM SPSS Modeler 사용을 위한 전체 안내서입니다. 이 안내서는 PDF 형식으로만 사용할 수 있습니다.

SPSS Modeler Premium 문서

SPSS Modeler Premium 문서 스위트(설치 지시사항은 제외)는 다음과 같습니다.

- **IBM SPSS Modeler Entity Analytics 사용자 안내서.** SPSS Modeler에서 엔티티 분석 사용에 대한 정보로, 리포지토리 설치 및 구성, 엔티티 분석 노드, 관리 작업에 대해 설명합니다.
- **IBM SPSS Modeler Social Network Analysis 사용자 안내서.** 그룹 분석 및 확산 분석을 포함하여 SPSS Modeler를 사용하여 소셜 네트워크 분석을 수행하기 위한 안내서입니다.
- **SPSS Modeler Text Analytics 사용자 안내서.** SPSS Modeler에서 텍스트 분석 사용에 대한 정보로, 텍스트 마이닝 노드, 대화식 워크벤치, 템플리트 및 기타 자원에 대해 설명합니다.

애플리케이션 예제

SPSS Modeler의 데이터 마이닝 도구가 광범위한 비즈니스 및 조직의 문제점을 해결하는 데 도움을 주는 가운데, 애플리케이션 예제는 특정 모델링 방법 및 기술에 대해 대상화된 간략한 소개를 제공합니다. 여기서 사용된 데이터 세트는 일부 데이터 마이너에서 관리하는 거대한 데이터 스토어보다 훨씬 작지만, 관련된 개념과 방법은 실제 애플리케이션으로 확장 가능합니다.

예제에 액세스하려면 SPSS Modeler의 도움말 메뉴에서 **애플리케이션 예제**를 클릭하십시오.

데이터 파일 및 샘플 스트림은 제품 설치 디렉토리 아래에 있는 Demos 폴더에 설치됩니다. 자세한 정보는 『Demos 폴더』의 내용을 참조하십시오.

데이터베이스 모델링 예제. *IBM SPSS Modeler In-Database 마이닝 안내서*의 예제를 참조하십시오.

스크립팅 예제. *IBM SPSS Modeler 스크립팅 및 자동화 안내서*의 예제를 참조하십시오.

Demos 폴더

애플리케이션 예에서 사용하는 데이터 파일 및 샘플 스트림은 제품 설치 디렉토리 아래의 Demos 폴더에 설치됩니다(예: C:\Program Files\IBM\SPSS\Modeler\<버전>\Demos). Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서, 또는 **파일 > 스트림 열기** 대화 상자의 최근 디렉토리 목록에서 Demos를 클릭해서도 이 폴더에 액세스할 수 있습니다.

라이센스 추적

SPSS Modeler를 사용할 때 라이센스 사용이 정기적으로 추적되고 로그됩니다. 로그되는 라이센스 메트릭은 *AUTHORIZED_USER* 및 *CONCURRENT_USER*이며 로그되는 메트릭의 유형은 SPSS Modeler에 대해 가진 라이센스의 유형에 의해 결정됩니다.

생성되는 로그 파일은 사용자가 라이센스 사용 보고서를 생성할 수 있는 IBM 라이센스 메트릭 도구에 의해 처리될 수 있습니다.

라이센스 로그 파일은 SPSS Modeler 클라이언트 로그 파일이 기록되는 디렉토리와 동일한 디렉토리(기본적으로 %ALLUSERSPROFILE%/IBM/SPSS/Modeler/<버전>/log)에 작성됩니다.

제 2 장 소스 노드

개요

소스 노드를 사용하면 플랫 파일, IBM SPSS Statistics(*.sav*), SAS, Microsoft Excel 및 ODBC 준수 관계형 데이터베이스 등을 포함하여 여러 가지 형식으로 저장된 데이터를 가져올 수 있습니다. 또한 사용자 입력 노드를 사용하여 합성 데이터를 생성할 수 있습니다.

소스 팔레트에는 다음 노드가 포함됩니다.

Analytic Server 소스를 사용하면 HDFS(Hadoop Distributed File System)에서 스트림을 실행할 수 있습니다. Analytic Server 데이터 소스의 정보는 텍스트 파일, 데이터베이스 등의 다양한 위치에서 제공될 수 있습니다. 자세한 정보는 13 페이지의 『Analytic Server 소스 노드』의 내용을 참조하십시오.

데이터베이스 노드를 사용하면 ODBC(Open Database Connectivity)를 사용하여 Microsoft SQL Server, DB2, Oracle 및 기타를 포함한 다양한 다른 패키지로부터 데이터를 가져올 수 있습니다. 자세한 정보는 18 페이지의 『데이터베이스 소스 노드』의 내용을 참조하십시오.

가변파일 노드는 자유 필드 텍스트 파일, 즉 레코드가 일정한 수의 필드를 포함하지만 변하는 문자를 포함하는 파일로부터 데이터를 읽습니다. 이 노드는 또한 고정 길이 헤더 텍스트와 특정 유형의 주석(Annotation)을 갖는 파일에도 유용합니다. 자세한 정보는 28 페이지의 『가변파일 노드』의 내용을 참조하십시오.

고정 파일 노드는 고정 필드 텍스트 파일, 즉 그의 필드가 구분되지 않고 동일한 위치에서 시작하며 고정된 길이의 파일로부터 데이터를 가져옵니다. 머신 생성 또는 레거시 데이터가 자주 고정 필드 형식으로 저장됩니다. 자세한 정보는 32 페이지의 『고정 파일 노드』의 내용을 참조하십시오.

통계량 파일 노드는 IBM SPSS Statistics에서 사용하는 *.sav* 또는 *.zsav* 파일 형식뿐 아니라 동일한 형식을 사용하는 IBM SPSS Modeler에 저장된 캐시 파일로부터 데이터를 읽습니다.

Data Collection 노드는 Data Collection 데이터 모델을 준수하는 시장 조사 소프트웨어에서 사용하는 다양한 형식에서 설문조사 데이터를 가져옵니다. 이 노드를 사용하려면 Data Collection Developer Library가 설치되어 있어야 합니다. 자세한 정보는 34 페이지의 『Data Collection 노드』의 내용을 참조하십시오.

IBM Cognos BI 소스 노드는 Cognos BI 데이터베이스에서 데이터를 가져옵니다.

IBM Cognos TM1 소스 노드는 Cognos TM1 데이터베이스에서 데이터를 가져옵니다.

SAS 파일 노드는 SAS 데이터를 IBM SPSS Modeler로 가져옵니다. 자세한 정보는 45 페이지의 『SAS 소스 노드』의 내용을 참조하십시오.

Excel 노드는 Microsoft Excel로부터 .xlsx 파일 형식으로 데이터를 가져옵니다. ODBC 데이터 소스는 필요하지 않습니다. 자세한 정보는 46 페이지의 『Excel 소스 노드』 주제를 참조하십시오.

XML 소스 노드는 XML 형식의 데이터를 스트림으로 가져옵니다. 단일 파일 또는 디렉토리의 모든 파일을 가져올 수 있습니다. 선택적으로 XML 구조를 읽을 스키마 파일을 지정할 수 있습니다.

사용자 입력 노드는 스크래치로부터 또는 기존 데이터를 변경하여 합성 데이터를 작성하는 쉬운 방법을 제공합니다. 이것은 예를 들어 모델링을 위한 검정 데이터 세트를 작성할 때 유용합니다. 자세한 정보는 49 페이지의 『사용자 입력 노드』의 내용을 참조하십시오.

시뮬레이션 생성 노드는 사용자가 지정한 통계 분포를 사용하는 스크래치로부터 또는 기존 히스토리 데이터에 대해 시뮬레이션 적합 노드를 실행하여 얻은 분포를 자동으로 사용하여 시뮬레이션된 데이터를 생성하는 쉬운 방법을 제공합니다. 이것은 모델 입력에 불확실성이 존재하는 상황에서 예측 모델의 결과를 평가하기 원할 때 유용합니다.

데이터 보기 노드는 IBM SPSS Collaboration and Deployment Services 분석 데이터 보기에서 정의된 데이터 소스에 액세스하는 데 사용할 수 있습니다. 분석 데이터 보기는 데이터에 액세스하는 데 필요한 표준 인터페이스를 정의하고 여러 물리적 데이터 소스를 해당 인터페이스와 연관시킵니다. 자세한 정보는 67 페이지의 『데이터 보기 노드』 주제를 참조하십시오.

맵 또는 공간 데이터를 데이터 마이닝 세션으로 가져오려면 지리 공간적 소스 노드를 사용하십시오. 자세한 정보는 69 페이지의 『지리 공간적 소스 노드』의 내용을 참조하십시오.

스트림을 시작하려면 스트림 캔버스에 소스 노드를 추가하십시오. 그런 다음, 노드를 두 번 클릭하여 대화 상자를 여십시오. 대화 상자의 다양한 탭을 사용하여 데이터를 읽고 필드 및 값을 보고 다양한 옵션(필터, 데이터 유형, 필드 역할 및 결측값 확인 등)을 설정할 수 있습니다.

필드 저장 공간 및 형식화 설정

고정 파일, 가변파일, XML 소스 및 사용자 입력 노드의 데이터 탭에 있는 옵션을 사용하여 IBM SPSS Modeler에서 필드를 가져오거나 작성할 때 필드의 저장 유형을 지정할 수 있습니다. 고정 파일, 가변파일 및 사용자 입력 노드의 경우 필드 형식화 및 다른 메타데이터도 지정할 수 있습니다.

다른 소스에서 읽히는 데이터의 경우, 저장 공간은 자동으로 결정되지만 채움 노드 또는 파생 노드에서 변환 함수(예: `to_integer`)를 사용하여 변경할 수 있습니다.

필드 현재 데이터 세트의 필드를 보고 선택하려면 **필드** 열을 사용하십시오.

대체 저장 공간 및 **입력 형식** 열의 옵션을 활성화하려면 **대체** 열의 선택란을 선택하십시오.

데이터 저장 공간

저장 공간은 데이터를 필드에 저장하는 방식을 설명합니다. 예를 들어, 값이 1 및 0인 필드는 정수 데이터를 저장합니다. 이는 데이터 사용에 대해 설명하고 저장 공간에 영향을 미치지 않는 측정 수준과 구별됩니다. 예를 들어, 값이 1 및 0인 정수 필드에 대한 측정 수준을 *플래그*로 설정할 수 있습니다. 이는 일반적으로 1 = *True* 및 0 = *False*을 표시합니다. 저장 공간은 소스에서 결정해야 하지만 측정 수준은 스트림의 어느 지점에서나 유형 노드를 사용하여 변경할 수 있습니다. 자세한 정보는 140 페이지의 『측정 수준』의 내용을 참조하십시오.

사용 가능한 저장 유형은 다음과 같습니다.

- **문자열** 영숫자 데이터라고도 하는 숫자가 아닌 데이터가 포함된 필드에 사용됩니다. 문자열은 *fred*, *Class 2* 또는 *1234* 등의 문자열 시퀀스를 포함할 수 있습니다. 문자열의 숫자는 계산에서 사용할 수 없습니다.
- **정수** 값이 정수인 필드입니다.
- **실수** 값은 10진수를 포함할 수 있는 숫자입니다(정수로 제한되지 않음). 표시 형식은 스트림 특성 대화 상자에서 지정되며 유형 노드(형식 탭)에서 개별 필드에 대해 대체될 수 있습니다.
- **날짜** 연도, 월 및 일 등의 표준 형식으로 지정된 날짜 값입니다(예: `2007-09-26`). 구체적인 형식은 스트림 특성 대화 상자에서 지정됩니다.
- **시간** 기간으로 측정된 시간입니다. 예를 들어, 1시간 26분 38초 동안 지속되는 서비스 호출은 스트림 특성 대화 상자에서 지정된 대로 현재 시간 형식에 따라 `01:26:38`로 표시될 수 있습니다.
- **시간소인** 스트림 특성 대화 상자의 현재 날짜 및 시간 형식에 따라 날짜와 시간 구성요소를 모두 포함하는 값입니다(예: `2007-09-26 09:04:00`). 시간소인 값을 별도의 날짜 및 시간 값 대신 단일 값으로 해석하기 위해 시간소인 값을 큰따옴표로 묶어야 할 수 있습니다. (이는 예를 들어, 사용자 입력 노드에서 값을 입력하는 경우에 적용됩니다.)
- **목록** 지리 공간 및 콜렉션이라는 새로운 측정 수준과 함께 SPSS Modeler 버전 17에서 도입된 목록 저장 공간 필드에는 단일 레코드에 대한 다중 값이 포함되어 있습니다. 모든 기타 저장 유형의 목록 버전이 있습니다.

표 1. 목록 저장 유형 아이콘

아이콘	저장 유형
[A]	문자열 목록
[ ]	정수 목록
[ ]	실수 목록
[ ]	시간 목록
[ ]	날짜 목록
[ ]	시간소인 목록
[[]]	0(영)보다 큰 깊이를 가진 목록

또한 콜렉션 측정 수준과 함께 사용하기 위해 다음과 같은 측정 수준의 목록 버전이 있습니다.

표 2. 목록 측정 수준 아이콘

아이콘	측정 수준
[ ]	연속형 목록
[ ]	범주형 목록
[ ]	플래그 목록
[ ]	명목 목록
[ ]	순서 목록

목록은 세 가지 소스 노드(Analytic Server, 지리 공간 또는 가변파일) 중 하나에서 SPSS Modeler로 가져오거나 파생 또는 채움 필드 작업 노드를 사용하여 스트림 내에서 작성할 수 있습니다.

목록 및 해당 콜렉션 및 지리 공간 측정 수준과의 상호작용에 대한 자세한 정보는 11 페이지의 『목록 저장 공간 및 연관된 측정 수준』의 내용을 참조하십시오.

저장 공간 변환. 채움 노드에서 `to_string` 및 `to_integer` 등의 다양한 변환 함수를 사용하여 필드에 대한 저장 공간을 변환할 수 있습니다. 자세한 정보는 165 페이지의 『채움 노드를 사용한 저장 공간 변환』의 내용을 참조하십시오. 변환 함수(및 날짜 또는 시간 값 등의 특정 입력 유형이 필요한 기타 함수)는 스트림 특성 대화 상자에서 지정된 현재 형식에 따라 다릅니다. 예를 들어, 값이 *Jan 2003*, *Feb 2003* 등인 문자열 필드를 날짜 저장 공간으로 변환하려면 **MON YYYY**를 스트림의 기본 날짜 형식으로 선택하십시오. 파생 계산 중

임시 변환의 경우 파생 노드에서도 변환 함수를 사용할 수 있습니다. 파생 노드를 사용하여 범주형 값을 가진 문자열 필드 코딩 변경 등의 기타 조작을 수행할 수도 있습니다. 자세한 정보는 164 페이지의 『파생 노드를 사용하여 값 코딩변경』의 내용을 참조하십시오.

혼합 데이터 읽기. 숫자 저장 공간(정수, 실수, 시간, 시간소인 또는 날짜)을 가진 필드에서 읽을 때 숫자가 아닌 값은 널값 또는 시스템 결측값으로 설정됩니다. 이는 일부 애플리케이션과는 달리 IBM SPSS Modeler가 필드 내에서 혼합 저장 유형을 허용하지 않기 때문입니다. 이를 방지하기 위해 혼합 데이터가 포함된 필드는 필요에 따라 소스 노드 또는 외부 애플리케이션에서 저장 유형을 변경하여 문자열로 읽어야 합니다.

필드 입력 형식(고정 파일, 가변파일 및 사용자 입력 노드에만 해당)

문자열 및 정수를 제외한 모든 저장 유형의 경우, 드롭 다운 목록을 사용하여 선택된 필드에 대해 형식화 옵션을 지정할 수 있습니다. 예를 들어, 다양한 로케일의 데이터를 병합할 때, 하나의 필드에 대해 소수점 구분 문자로 마침표(.)를 지정해야 하지만 다른 필드는 쉼표 구분 문자를 필요로 할 수 있습니다.

소스 노드에 지정된 입력 옵션은 스트림 특성 대화 상자에 지정된 형식화 옵션을 대체합니다(단, 나중에 스트림에서 지속되지 않음). 이러한 옵션은 데이터에 대한 지식을 기반으로 입력을 올바르게 구문 분석하는 데 사용됩니다. 지정된 형식은 IBM SPSS Modeler로 데이터를 읽은 후 데이터를 형식화하는 방법을 결정하는 데 사용되지 않고 IBM SPSS Modeler로 데이터를 읽을 때 데이터를 구문 분석하기 위한 지침으로 사용됩니다. 스트림의 다른 위치에서 필드별로 형식화를 지정하려면 유형 노드의 형식 탭을 사용하십시오. 자세한 정보는 152 페이지의 『필드 형식 설정 탭』의 내용을 참조하십시오.

옵션은 저장 유형에 따라 다릅니다. 예를 들어, 실수 저장 유형의 경우, 소수점 구분 문자로 **마침표(.)** 또는 **쉼표(,)**를 선택할 수 있습니다. 시간소인 필드의 경우, 드롭 다운 목록에서 **지정**을 선택하면 별도의 대화 상자가 열립니다. 자세한 정보는 153 페이지의 『필드 형식 옵션 설정』의 내용을 참조하십시오.

모든 저장 유형의 경우, **스트림 기본값**을 선택하여 가져오기에 스트림 기본값 설정을 사용할 수도 있습니다. 스트림 설정은 스트림 특성 대화 상자에 지정됩니다.

추가 옵션

데이터 탭을 사용하여 일부 다른 옵션을 지정할 수 있습니다.

- 현재 노드를 통해 더 이상 연결되지 않는 데이터에 대한 저장 공간 설정을 보려면(예를 들어, 훈련 데이터) **사용하지 않는 필드 설정 보기**를 선택하십시오. **지우기**를 클릭하여 레거시 필드를 지울 수 있습니다.
- 이 대화 상자에서 작업하는 중 언제든 **새로 고치기**를 클릭하여 데이터 소스에서 필드를 다시 로드할 수 있습니다. 이는 소스 노드와의 데이터 연결을 변경하거나 대화 상자의 탭들 사이에서 작업 중일 때 유용합니다.

목록 저장 공간 및 연관된 측정 수준

새 측정 수준인 지리 공간적 및 콜렉션에 대해 작업하기 위해 SPSS Modeler 버전 17에 도입된 목록 저장 공간 필드에는 단일 레코드에 대한 다중 값이 포함되어 있습니다. 목록은 꺽쇠 대괄호([])로 싸여 있습니다. 목록의 예로는 `[1,2,4,16]`과 `["abc", "def"]`가 있습니다.

목록은 세 가지 소스 노드(Analytic Server, 지리 공간 또는 가변파일) 중 하나에서 SPSS Modeler로 가져오거나 파생 또는 채움 필드 작업 노드를 사용하여 스트림 내에서 작성되거나 순위화된 조건 병합 방법 사용 시 병합 노드에 의해 생성될 수 있습니다.

목록은 깊이를 가지는 것으로 간주됩니다. 예를 들어, [1,3] 형식의 단일 대괄호로 묶인 항목을 가진 단순 목록은 깊이 0(영)을 사용하여 IBM SPSS Modeler에서 기록됩니다. 깊이 0(영)을 가진 단순 목록 외에도 목록 내 각각의 값이 목록 자체인 중첩된 목록을 사용할 수 있습니다.

중첩된 목록의 깊이는 연관된 측정 수준에 따라 다릅니다. 유형 없음의 경우에는 설정된 깊이 제한이 없고 콜렉션의 경우 깊이는 0(영)이며 지리 공간적 측정 수준의 경우 깊이는 중첩된 항목의 수에 따라 0(영)과 2 사이(경계값 포함)여야 합니다.

깊이가 0(영)인 목록의 경우 측정 수준을 지리 공간적 또는 콜렉션으로 설정할 수 있습니다. 이 수준은 모두 상위 측정 수준이므로 사용자는 값 대화 상자에서 측정 하위 수준 정보를 설정합니다. 콜렉션의 측정 하위 수준은 해당 목록에 있는 요소의 측정 수준을 결정합니다. 모든 측정 수준(유형 없음 및 지리 공간적 측정 수준 제외)을 콜렉션의 하위 수준으로 사용할 수 있습니다. 지리 공간적 측정 수준은 점, 선 스트링, 다각형, 다중 점, 다중 선 스트림 및 다중 다각형이라는 6개의 하위 수준을 가지고 있습니다. 자세한 정보는 142 페이지의 『지리 공간적 측정 수준』의 내용을 참조하십시오.

참고: 콜렉션 측정 수준은 깊이가 0(영)인 목록에만 사용할 수 있고 지리 공간적 측정 수준은 최대 깊이가 2인 목록에만 사용할 수 있으며 유형 없음 측정 수준은 모든 목록 깊이와 함께 사용할 수 있습니다.

다음 예제에서는 지리 공간적 측정 하위 수준 점 및 선 스트링의 구조를 사용하여 깊이가 0(영)인 목록과 중첩된 목록 사이의 차이를 보여줍니다.

- 지리 공간적 측정 하위 수준 점의 필드 깊이는 0(영)입니다.

  [1,3] 2개의 좌표

  [1,3,-1] 3개의 좌표
- 지리 공간적 측정 하위 수준 선 스트림의 필드 깊이는 1입니다.

  [[1,3], [5,0]] 2개의 좌표

  [[1,3,-1], [5,0,8]] 3개의 좌표

깊이가 0(영)인 점 필드는 각각의 값이 2개 또는 3개의 좌표로 구성되는 일반적인 목록입니다. 깊이가 1인 선 스트림 필드는 각각의 점이 추가적인 일련의 목록 값으로 구성되는 점의 목록입니다.

목록 작성에 대한 자세한 정보는 161 페이지의 『목록 또는 지리 공간적 필드 파생』의 내용을 참조하십시오.

지원되지 않는 제어 문자

SPSS Modeler의 일부 프로세스에서는 다양한 제어 문자가 포함된 데이터를 처리할 수 없습니다. 데이터가 이 문자를 사용하는 경우에는 다음 예제와 같은 오류 메시지가 표시될 수 있습니다.

```
Unsupported control characters found in values of field {0}
```

지원되지 않는 문자는 0x0부터 0x3F까지(0x3F 포함) 및 0x7F이지만 탭(0x9(\t)), 줄 바꾸기(0xA(\n)) 및 캐리지 리턴(0xD(\r)) 문자는 문제를 유발하지 않습니다.

지원되지 않는 문자와 관련된 오류 메시지가 표시되는 경우에는 스트림의 소스 노드 뒤에서 채움 노드 및 CLEM 표현식 **stripctrlchars**를 사용하여 해당 문자를 바꾸십시오.

Analytic Server 소스 노드

Analytic Server 소스를 사용하면 HDFS(Hadoop Distributed File System)에서 스트림을 실행할 수 있습니다. Analytic Server 데이터 소스의 정보는 다음과 같은 다양한 위치에서 제공될 수 있습니다.

- HDFS의 텍스트 파일
- 데이터베이스
- HCatalog

일반적으로 Analytic Server 소스를 가진 스트림은 HDFS에서 실행됩니다. 하지만 스트림에 HDFS에서의 실행에 대해 지원되지 않는 노드가 포함되어 있으면 가능한 많은 스트림이 Analytic Server에 푸시백된 후 SPSS Modeler Server가 나머지 스트림을 처리합니다. 예를 들어, 스트림 내에 표본 노드를 배치하여 매우 큰 데이터 세트의 부표본을 추출해야 합니다.

데이터 소스. SPSS Modeler Server 관리자가 연결을 설정했다고 가정하고 사용할 데이터가 포함된 데이터 소스를 선택합니다. 데이터 소스에는 해당 소스와 연관된 파일 및 메타데이터가 포함되어 있습니다. **선택**을 클릭하여 사용 가능한 데이터 소스의 목록을 표시하십시오. 자세한 정보는 『데이터 소스 선택』의 내용을 참조하십시오.

새 데이터 소스를 작성하거나 기존 데이터 소스를 편집해야 하는 경우에는 **데이터 소스 편집기 시작...**을 클릭하십시오.

데이터 소스 선택

데이터 소스 테이블에는 사용 가능한 데이터 소스의 목록이 표시됩니다. 사용할 소스를 선택한 후 **확인**을 클릭하십시오.

소유자 표시를 클릭하여 데이터 소스 소유자를 표시하십시오.

필터 기준을 사용하면 **키워드**에서 데이터 소스 목록을 필터링하여 **소유자** 또는 데이터 소스 이름 및 데이터 소스 설명에 대해 필터 기준을 확인할 수 있습니다. 아래에 설명된 문자열, 숫자 또는 와일드카드의 조합을 필터 기준으로 입력할 수 있습니다. 검색 문자열은 대소문자를 구분합니다. **새로 고치기**를 클릭하여 데이터 소스 테이블을 업데이트하십시오.

_ 밑줄을 사용하여 검색 문자열에서 단일 문자를 나타낼 수 있습니다.

% 퍼센트 부호를 사용하여 검색 문자열에서 0개 이상 문자의 시퀀스를 나타낼 수 있습니다.

신임 정보 수정

Analytic Server에 액세스하는 데 필요한 신임 정보가 SPSS Modeler Server에 액세스하는 데 필요한 신임 정보와 다른 경우에는 Analytic Server에서 스트림을 실행할 때 Analytic Server 신임 정보를 입력해야 합니다. 신임 정보를 모르는 경우에는 서버 관리자에게 문의하십시오.

지원되는 노드

여러 SPSS Modeler 노드는 HDFS에서 실행하도록 지원되지만 특정 노드 실행에는 일부 차이가 있을 수 있으며 일부는 현재 지원되지 않습니다. 이 주제에서는 현재 지원 수준에 대해 설명합니다.

일반

- 인용된 Modeler 필드 이름에서 일반적으로 허용 가능한 일부 문자를 Analytic Server에서 허용하지 않습니다.
- Modeler 스트림을 Analytic Server에서 실행하려면 하나 이상의 Analytic Server 소스 노드로 시작하고 단일 모델링 노드 또는 Analytic Server 내보내기 노드에서 종료되어야 합니다.
- 연속형 대상의 저장 공간을 정수가 아닌 실수로 설정하도록 권장됩니다. 스코어의 출력 데이터 모델은 대상의 저장 공간을 따르지만 스코어링 모델은 항상 실수 값을 연속형 대상의 출력 데이터 파일에 기록합니다. 따라서 연속형 대상에 정수 저장 공간이 있는 경우 기록된 값과 스코어의 데이터 모델이 일치하지 않으며 이러한 불일치로 인해 스코어링된 데이터를 읽으려고 시도할 때 오류가 발생합니다.

소스

- Analytic Server 소스 노드가 아닌 다른 노드로 시작하는 스트림이 로컬로 실행됩니다.

레코드 작업

스트리밍 TS 및 Space-Time-Box 노드를 제외한 모든 레코드 작업이 지원됩니다. 지원되는 노드 기능에 대한 추가 참고사항은 다음과 같습니다.

선택

- 파생 노드에서 지원하는 동일한 함수 세트를 지원합니다.
- 선택 노드를 삭제 옵션으로 사용하는 경우 규칙 세트에서 널 값이 있는 필드가 삭제됩니다. 예를 들어, 조건이 OCCUPATION = "Retired"인 행을 버리는 것인 경우, OCCUPATION = "Retired" 및 OCCUPATION = null인 모든 행이 버려집니다. "not(field = undef)"을 추가하도록 선택기준을 수정해야 합니다. 예를 들어, 선택 기준을 ((OCCUPATION =

"Retired) 및 not(OCCUPATION = undef))로 업데이트하십시오. 결과 세트는 OCCUPATION 필드가 널인 행을 포함합니다.

샘플

- 블록 수준 샘플링은 지원되지 않습니다.
- 복합 샘플링 방법은 지원되지 않습니다.

통합

- 인접 키는 지원되지 않습니다. 데이터를 정렬한 후 통합 노드에서 이 설정을 사용하도록 설정된 기존 스트림을 재사용하는 경우 정렬 노드를 제거하여 스트림을 변경하십시오.
- 주문 통계(중앙값, 첫 번째 사분위수, 세 번째 사분위수)가 대략적으로 계산되며 최적화 탭을 통해 지원됩니다.

정렬

- 최적화 탭은 지원되지 않습니다.

분배 환경에는 정렬 노드에서 설정한 레코드 순서를 따르는 제한된 수의 작업이 있습니다.

- 정렬 이후에 내보내기 노드가 있으면 정렬된 데이터 소스가 생성됩니다.
- 정렬 이후에 표본 노드가 있으며 **첫 번째** 레코드 샘플링이 있는 경우 처음 *N*개의 레코드가 리턴됩니다.

일반적으로 정렬된 레코드가 필요한 작업에 가능하면 가까이 정렬 노드를 배치해야 합니다.

병합

- 순서별 병합은 지원되지 않습니다.
- 최적화 탭은 지원되지 않습니다.
- Analytic Server는 비어 있는 문자열 키에서 결합되지 않습니다. 즉, 병합할 때 사용하는 하나의 키에 비어 있는 문자열이 있는 경우 비어 있는 문자열이 포함된 레코드가 병합된 출력에서 삭제됩니다.
- 병합 작업은 상대적으로 느립니다. HDFS에 사용 가능한 공간이 있는 경우 데이터 소스를 한 번 병합하고 다음 스트림에서 병합된 소스를 사용하는 것이 각 스트림에서 데이터 소스를 병합하는 것보다 훨씬 빠를 수 있습니다.

R 변환

노드의 R 구문은 record-at-a-time 작업으로 구성되어 있어야 합니다.

필드 작업

전치, 시간 간격, 히스토리 노드를 제외한 모든 필드 작업이 지원됩니다. 지원되는 노드 기능에 대한 추가 참고사항은 다음과 같습니다.

자동 데이터 준비

- 노드 학습은 지원되지 않습니다. 학습된 자동 데이터 준비 노드의 변환을 새 데이터에 적용하는 것이 지원됩니다.

유형

- 검사 열은 지원되지 않습니다.
- 형식화 탭은 지원되지 않습니다.

파생

- 시퀀스 함수를 제외한 모든 파생 함수가 지원됩니다.
- 분할 필드는 필드를 분할로 사용하는 동일한 스트림에서 파생될 수 없습니다. 두 개의 스트림을 작성해야 합니다. 분할 필드를 파생하는 스트림과 필드를 분할로 사용하는 스트림입니다.
- 플래그 필드는 비교에서 단독으로 사용할 수 없습니다. 즉, if (flagField) then ... endif는 오류를 발생시킵니다. 임시 해결책으로 if (flagField=trueValue) then ... endif를 사용할 수 있습니다.
- Modeler의 결과와 일치하도록 ** 연산자를 사용하여 지수를 실수로(예를 들어, x**2가 아닌 x**2.0으로) 지정하도록 권장됩니다.

채움

- 파생 노드에서 지원하는 동일한 함수 세트를 지원합니다.

구간화 다음 기능은 지원되지 않습니다.

- 최적 구간화
- 순위
- 분위수 -> 분위수 지정: 값의 합계
- 분위수 -> 경계값: 현재에서 유지 및 무작위 지정
- 분위수 -> 사용자 정의 N: 100 이상의 값 및 100 % N이 0이 아닌 모든 N 값

RFM 분석

- 경계값 처리를 위한 현재에서 유지 옵션은 지원되지 않습니다. RFM 최근성, 빈도, 통화 스코어는 항상 동일한 데이터에서 Modeler가 계산한 것과 일치하지 않습니다. 스코어 범위는 동일하지만 스코어 지정(구간 수)이 1씩 다를 수 있습니다.

그래프 모든 그래프 노드가 지원됩니다.

모델링 TCM, 트리-AS, C&R 트리, Quest, CHAID, 선형, 선형-AS, 신경망, GLE, LSVM, TwoStep-AS, 랜덤 트리, STP, 연관 규칙 모델링 노드가 지원됩니다. 이러한 노드의 기능에 대한 추가 참고사항은 다음과 같습니다.

선형 큰 데이터에서 모델을 작성할 때 일반적으로 매우 큰 데이터 세트로 목표를 변경하거나 분할을 지정하게 됩니다.

- 기존 PSM 모델의 연속 학습은 지원되지 않습니다.
- 표준 모델 작성 목표는 각 분할의 레코드 수가 너무 크지 않도록 분할 필드가 정의된 경우에만 권장됩니다. 여기서 "너무 큰"의 정의는 Hadoop 군집에서 개별 노드의 거듭제곱에 따라 다릅니다. 반대로 모델을 작성하는 데 레코드 수가 너무 적을 정도로 분할이 세밀하게 정의되지 않도록 주의해야 합니다.

- 부스팅 목표는 지원되지 않습니다.
- 배깅 목표는 지원되지 않습니다.
- 매우 큰 데이터 세트 목표는 레코드가 거의 없는 경우에 권장되지 않습니다. 모델을 작성하지 않거나 저하된 모델을 작성합니다.
- 자동 데이터 준비는 지원되지 않습니다. 결측값이 많은 데이터에서 모델을 작성하려고 할 때 이로 인해 문제가 발생할 수 있습니다. 일반적으로 자동 데이터 준비의 일부분으로 대치됩니다. 임시 해결책은 트리 모델 또는 신경망을 고급 설정으로 사용하여 선택된 결측값을 대치하는 것입니다.
- 분할 모델의 경우 정확도 통계가 계산되지 않습니다.

신경망 큰 데이터에서 모델을 작성할 때 일반적으로 매우 큰 데이터 세트로 목표를 변경하거나 분할을 지정하게 됩니다.

- 기존 표준 또는 PSM 모델의 연속 학습은 지원되지 않습니다.
- 표준 모델 작성 목표는 각 분할의 레코드 수가 너무 크지 않도록 분할 필드가 정의된 경우에만 권장됩니다. 여기서 "너무 큰"의 정의는 Hadoop 군집에서 개별 노드의 거듭제곱에 따라 다릅니다. 반대로 모델을 작성하는 데 레코드 수가 너무 적을 정도로 분할이 세밀하게 정의되지 않도록 주의해야 합니다.
- 부스팅 목표는 지원되지 않습니다.
- 배깅 목표는 지원되지 않습니다.
- 매우 큰 데이터 세트 목표는 레코드가 거의 없는 경우에 권장되지 않습니다. 모델을 작성하지 않거나 저하된 모델을 작성합니다.
- 데이터에 결측값이 많은 경우 고급 설정을 사용하여 결측값을 대치하십시오.
- 분할 모델의 경우 정확도 통계가 계산되지 않습니다.

C&R 트리, CHAID, Quest

큰 데이터에서 모델을 작성할 때 일반적으로 매우 큰 데이터 세트로 목표를 변경하거나 분할을 지정하게 됩니다.

- 기존 PSM 모델의 연속 학습은 지원되지 않습니다.
- 표준 모델 작성 목표는 각 분할의 레코드 수가 너무 크지 않도록 분할 필드가 정의된 경우에만 권장됩니다. 여기서 "너무 큰"의 정의는 Hadoop 군집에서 개별 노드의 거듭제곱에 따라 다릅니다. 반대로 모델을 작성하는 데 레코드 수가 너무 적을 정도로 분할이 세밀하게 정의되지 않도록 주의해야 합니다.
- 부스팅 목표는 지원되지 않습니다.
- 배깅 목표는 지원되지 않습니다.
- 매우 큰 데이터 세트 목표는 레코드가 거의 없는 경우에 권장되지 않습니다. 모델을 작성하지 않거나 저하된 모델을 작성합니다.
- 대화식 세션은 지원되지 않습니다.
- 분할 모델의 경우 정확도 통계가 계산되지 않습니다.

- 분할 필드가 존재하는 경우, Modeler에서 로컬로 빌드된 트리 모델은 Analytic Server에서 빌드한 트리 모델과 조금 다르며, 그로 인해 다른 점수를 냅니다. 두 경우의 알고리즘이 유효합니다. Analytic Server에 의해 사용되는 알고리즘이 더 최신입니다. 트리 알고리즘이 다수의 휴리스틱 규칙을 가지려는 경향이 있다는 사실을 고려하면 두 구성요소 사이의 차이는 정상입니다.

모델 스코어링

모델링에 지원되는 모든 모델은 스코어링에도 지원됩니다. 또한 다음 노드에 대해 로컬로 작성된 모델 너깃은 스코어링에 지원됩니다. C&RT, Quest, CHAID, 선형, 신경망(모델이 표준, 부스팅된 배깅이거나 매우 큰 데이터 세트에 해당하는지 여부에 관계없음), 회귀분석, C5.0, 로지스틱, Genlin, GLMM, Cox, SVM, Bayes Net, TwoStep, KNN, 의사결정 목록, 차별, 자체 학습, 이상 항목 발견, Apriori, Carma, K-평균, Kohonen, R, 텍스트 마이닝.

- 원시 또는 조정된 성향은 스코어링되지 않습니다. 임시 해결책으로 if 'predicted-value' == 'value-of-interest' then 'prob-of-that-value' else 1-'prob-of-that-value' endif 표현식에서 파생 노드를 사용하여 원시 성향을 수동으로 계산하면 동일한 결과를 얻을 수 있습니다.
- 모델을 스코어링할 때 Analytic Server는 모델에서 사용되는 모든 필드가 데이터 세트에 있는지 확인하지 않으므로 Analytic Server에서 실행하기 전에 이 사항이 올바른지 확인해야 합니다.

R 너깃의 R 구문은 record-at-a-time 작업으로 구성되어 있어야 합니다.

출력 교차표, 분석, 데이터 검토, 변환, 통계량, 평균, 테이블 노드가 지원됩니다. 지원되는 노드 기능에 대한 추가 참고사항은 다음과 같습니다.

데이터 검토

데이터 검토 노드는 연속형 필드의 모드를 생성합니다.

평균 평균 노드는 표준 오류 또는 95% 신뢰구간을 생성할 수 없습니다.

테이블 업스트림 작업의 결과가 들어 있는 임시 Analytic Server 데이터 소스를 기록하여 테이블 노드가 지원됩니다. 그런 다음, 테이블 노드는 해당 데이터 소스의 컨텐츠를 통해 페이징합니다.

내보내기

스트림은 Analytic Server 소스 노드로 시작하고 Analytic Server 내보내기 노드가 아닌 다른 내보내기 노드로 종료할 수 있지만 데이터는 HDFS에서 SPSS Modeler Server로 이동되고 마지막으로 내보내기 위치로 이동됩니다.

데이터베이스 소스 노드

데이터베이스 소스 노드를 사용하면 ODBC(Open Database Connectivity)를 사용하여 Microsoft SQL Server, DB2, Oracle 및 기타를 포함한 다양한 다른 패키지로부터 데이터를 가져올 수 있습니다.

데이터베이스를 읽거나 데이터베이스에 쓰려면 필요에 따라 읽기 또는 쓰기 권한을 가지고 관련 데이터베이스에 대해 ODBC 데이터 소스가 설치 및 구성되어 있어야 합니다. IBM SPSS Data Access Pack에는 이 용

도로 사용할 수 있는 ODBC 드라이버 세트가 포함되어 있으며 이 드라이버는 다운로드 사이트로부터 얻을 수 있습니다. ODBC 데이터 소스에 대한 작성 및 설정에 관한 문의사항이 있으면 데이터베이스 관리자에게 문의하십시오.

지원되는 ODBC 드라이버

IBM SPSS Modeler 18과 함께 사용하기 위해 지원되고 테스트되는 데이터베이스 및 ODBC 드라이버에 대한 최신 정보는 회사 지원 사이트(http://www.ibm.com/support)에서 제품 호환성 교차표를 참조하십시오.

드라이버 설치 위치

참고: 처리가 발생할 수 있는 각각의 컴퓨터에 ODBC 드라이버가 설치 및 구성되어 있어야 합니다.

- 로컬(독립형) 모드에서 IBM SPSS Modeler를 실행 중인 경우에는 로컬 컴퓨터에 드라이버가 설치되어야 합니다.
- 원격 IBM SPSS Modeler Server에 대해 분산 모드에서 IBM SPSS Modeler를 실행 중인 경우에는 IBM SPSS Modeler Server가 설치되는 컴퓨터에 ODBC 드라이버가 설치되어야 합니다. UNIX 시스템의 IBM SPSS Modeler Server에 대해서는 이 절의 뒷부분에 있는 "UNIX 시스템에서 ODBC 드라이버 구성"도 참조하십시오.
- IBM SPSS Modeler와 IBM SPSS Modeler Server 모두에서 동일한 소스에 액세스해야 하는 경우에는 두 컴퓨터 모두에 ODBC 드라이버가 설치되어야 합니다.
- 터미널 서비스를 통해 IBM SPSS Modeler를 실행 중인 경우에는 IBM SPSS Modeler를 설치한 터미널 서비스 서버에 ODBC 드라이버가 설치되어야 합니다.

데이터베이스의 데이터에 액세스

데이터베이스의 데이터에 액세스하려면 다음의 단계를 완료하십시오.

- ODBC 드라이버를 설치하고 사용할 데이터베이스에 대한 데이터 소스를 구성하십시오.
- 데이터베이스 노드 대화 상자에서 테이블 모드 또는 SQL 쿼리 모드를 사용하여 데이터베이스에 연결하십시오.
- 데이터베이스에서 테이블을 선택하십시오.
- 데이터베이스 노드 대화 상자의 탭을 사용하여 사용 유형을 변경하고 데이터 필드를 필터링할 수 있습니다.

선행 단계에 대한 자세한 내용은 관련 문서 주제에서 제공됩니다.

참고: SPSS Modeler에서 데이터베이스 스토어드 프로시저(SP)를 호출하는 경우에는 SP의 예상 출력 대신 `RowsAffected`라는 단일 출력 필드가 리턴되는 것을 볼 수 있습니다. 이는 ODBC가 SP의 출력 데이터 모델을 판별하기에 충분한 정보를 리턴하지 않는 경우 발생합니다. SPSS Modeler는 출력을 리턴하는 SP에 대해 제한된 지원만 제공하므로 SP를 사용하는 대신 SP에서 SELECT를 추출한 후 다음 조치 중 하나를 사용하는 것이 좋습니다.

- SELECT를 기반으로 보기를 작성하고 데이터베이스 소스 노드에서 보기 선택
- 데이터베이스 소스 노드에서 직접 SELECT 사용

데이터베이스 노드 옵션 설정

데이터베이스 소스 노드 대화 상자의 데이터 탭에 있는 옵션을 사용하여 데이터베이스에 대한 액세스를 얻고 선택된 테이블에서 데이터를 읽어올 수 있습니다.

모드. 대화 상자 제어를 사용하여 테이블에 연결하려면 **테이블**을 선택하십시오.

SQL을 사용하여 아래에서 선택된 데이터베이스를 쿼리하려면 **SQL 쿼리**를 선택하십시오. 자세한 정보는 27 페이지의 『데이터베이스 쿼리』의 내용을 참조하십시오.

데이터 소스. 테이블 모드와 SQL 쿼리 모드 모두에 대해 데이터 소스 필드에서 이름을 입력하거나 드롭 다운 목록에서 **새 데이터베이스 연결 추가**를 선택할 수 있습니다.

데이터베이스에 연결하고 대화 상자를 사용하여 테이블을 선택하는 데 사용되는 옵션은 다음과 같습니다.

테이블 이름. 액세스할 테이블의 이름을 알고 있는 경우에는 테이블 이름 필드에 해당 이름을 입력하십시오. 그렇지 않으면 **선택** 단추를 클릭하여 사용 가능한 테이블이 나열되는 대화 상자를 여십시오.

테이블 및 열 이름 따옴표로 묶기. 쿼리가 데이터베이스에 전송될 때 예를 들어, 쿼리에 공백 또는 구두점이 포함된 경우 테이블 및 열 이름을 따옴표로 묶을지 여부를 지정하십시오.

- **필요 시** 옵션은 테이블 및 필드 이름이 비표준 문자를 포함하는 경우에만 테이블 및 필드 이름을 따옴표로 묶습니다. 비표준 문자에는 비ASCII 문자, 공백 문자, 마침표(.) 이외의 영숫자가 아닌 문자가 포함됩니다.
- 모든 테이블 및 필드 이름을 따옴표로 묶으려면 **항상**을 선택하십시오.
- 테이블 및 필드 이름을 따옴표로 묶지 *않으려면* **사용 안 함**을 선택하십시오.

선행 및 후미 공백 제거문자열에서 선행 및 후미 공백을 삭제하는 옵션을 선택하십시오.

참고. SQL 푸시백을 사용하는 문자열과 사용하지 않는 문자열 간의 비교는 후미 공백이 존재하는 서로 다른 결과를 생성할 수 있습니다.

Oracle에서 빈 문자열 읽기. Oracle 데이터베이스에서 읽거나 Oracle 데이터베이스에 쓸 때 IBM SPSS Modeler 및 대부분의 다른 데이터베이스와 달리 Oracle은 빈 문자열 값을 널값과 같다고 간주하고 저장합니다. 이는 Oracle 데이터베이스에서 추출된 동일한 데이터가 파일 또는 다른 데이터베이스에서 추출된 경우와 다르게 동작하고 데이터가 다른 결과를 리턴할 수 있음을 의미합니다.

데이터베이스 연결 추가

데이터베이스를 열려면 먼저 연결할 데이터 소스를 선택하십시오. 데이터 탭의 데이터 소스 드롭 다운 목록에서 **새 데이터베이스 연결 추가**를 선택하십시오.

데이터베이스 연결 대화 상자가 열립니다.

참고: 이 대화 상자를 여는 다른 방법은 기본 메뉴에서 **도구 > 데이터베이스...**를 선택하는 것입니다.

데이터 소스 사용 가능한 데이터 소스를 나열합니다. 원하는 데이터베이스가 표시되지 않으면 아래로 스크롤하십시오. 데이터 소스를 선택하고 비밀번호를 입력한 후 **연결**을 클릭하십시오. **새로 고치기**를 클릭하여 목록을 업데이트하십시오.

사용자 이름 및 비밀번호 데이터 소스가 비밀번호로 보호되어 있으면 사용자 이름 및 연관된 비밀번호를 입력하십시오.

신임 정보 IBM SPSS Collaboration and Deployment Services에서 신임 정보가 구성되어 있는 경우 이 옵션을 선택하여 리포지토리에서 해당 신임 정보를 찾아볼 수 있습니다. 신임 정보의 사용자 이름 및 비밀번호는 데이터베이스에 액세스하기 위해 필요한 사용자 이름 및 비밀번호와 일치해야 합니다.

연결 현재 연결된 데이터베이스를 표시합니다.

- **기본값** 선택적으로 하나의 연결을 기본값으로 선택할 수 있습니다. 그러면 데이터베이스 소스 또는 내보내기 노드에 이 연결이 데이터 소스로 사전 정의되지만 이는 원하는 경우 편집할 수 있습니다.
- **저장** 선택적으로 후속 세션에서 다시 표시할 하나 이상의 연결을 선택하십시오.
- **데이터 소스** 현재 연결된 데이터베이스에 대한 연결 문자열입니다.
- **사전 설정** 사전 설정 값이 데이터베이스 연결에 대해 지정되었는지를 표시합니다(* 문자 사용). 사전 설정 값을 지정하려면 데이터베이스 연결에 해당하는 행에서 이 열을 클릭한 후 목록에서 지정을 선택하십시오. 자세한 정보는 23 페이지의 『데이터베이스 연결에 대한 사전 설정된 값 지정』의 내용을 참조하십시오.

연결을 제거하려면 목록에서 하나의 연결을 선택한 후 **제거**를 클릭하십시오.

선택을 완료한 후 **확인**을 클릭하십시오.

데이터베이스를 읽거나 데이터베이스에 쓰려면 필요에 따라 읽기 또는 쓰기 권한을 가지고 관련 데이터베이스에 대해 ODBC 데이터 소스가 설치 및 구성되어 있어야 합니다. IBM SPSS Data Access Pack에는 이 용도로 사용할 수 있는 ODBC 드라이버 세트가 포함되어 있으며 이 드라이버는 다운로드 사이트로부터 얻을 수 있습니다. ODBC 데이터 소스에 대한 작성 및 설정에 관한 문의사항이 있으면 데이터베이스 관리자에게 문의하십시오.

지원되는 ODBC 드라이버

IBM SPSS Modeler 18과 함께 사용하기 위해 지원되고 테스트되는 데이터베이스 및 ODBC 드라이버에 대한 최신 정보는 회사 지원 사이트(http://www.ibm.com/support)에서 제품 호환성 교차표를 참조하십시오.

드라이버 설치 위치

참고: 처리가 발생할 수 있는 각각의 컴퓨터에 ODBC 드라이버가 설치 및 구성되어 있어야 합니다.

- 로컬(독립형) 모드에서 IBM SPSS Modeler를 실행 중인 경우에는 로컬 컴퓨터에 드라이버가 설치되어야 합니다.

- 원격 IBM SPSS Modeler Server에 대해 분산 모드에서 IBM SPSS Modeler를 실행 중인 경우에는 IBM SPSS Modeler Server가 설치되는 컴퓨터에 ODBC 드라이버가 설치되어야 합니다. UNIX 시스템의 IBM SPSS Modeler Server에 대해서는 이 절의 뒷부분에 있는 "UNIX 시스템에서 ODBC 드라이버 구성"도 참조하십시오.
- IBM SPSS Modeler와 IBM SPSS Modeler Server 모두에서 동일한 소스에 액세스해야 하는 경우에는 두 컴퓨터 모두에 ODBC 드라이버가 설치되어야 합니다.
- 터미널 서비스를 통해 IBM SPSS Modeler를 실행 중인 경우에는 IBM SPSS Modeler를 설치한 터미널 서비스 서버에 ODBC 드라이버가 설치되어야 합니다.

UNIX 시스템에서 ODBC 드라이버 구성

기본적으로 DataDirect 드라이버 관리자는 UNIX 시스템의 IBM SPSS Modeler Server에 대해 구성되어 있지 않습니다. DataDirect 드라이버 관리자를 로드하도록 UNIX를 구성하려면 다음의 명령을 입력하십시오.

```
cd <modeler_server_install_directory>/bin
rm -f libspssodbc.so
ln -s libspssodbc_datadirect.so libspssodbc.so
```

그러면 기본 링크가 제거되고 DataDirect 드라이버 관리자에 대한 링크가 작성됩니다.

참고: 일부 데이터베이스의 경우 SAP HANA 또는 IBM DB2 CLI 드라이버를 사용하려면 UTF16 드라이버 랩퍼가 필요합니다. DashDB에는 IBM DB2 CLI 드라이버가 필요합니다. UTF16 드라이버 랩퍼에 대한 링크를 작성하려면 다음의 명령을 대신 입력하십시오.

```
rm -f libspssodbc.so
ln -s libspssodbc_datadirect_utf16.so libspssodbc.so
```

SPSS Modeler Server를 구성하려면 다음을 수행하십시오.

1. `modelersrv.sh`에 다음 행을 추가하여 IBM SPSS Data Access Pack `odbc.sh` 환경 파일을 제공하도록 SPSS Modeler Server 시작 스크립트 `modelersrv.sh`를 구성하십시오.

   ```
   . /<pathtoSDAPinstall>/odbc.sh
   ```

   여기서 `<pathtoSDAPinstall>`은 IBM SPSS Data Access Pack 설치의 전체 경로입니다.
2. SPSS Modeler Server를 다시 시작하십시오.

또한 `odbc.ini` 파일의 DSN에 다음 매개변수 정의를 추가하여 연결 중 버퍼 오버플로우를 방지하십시오(SAP HANA 및 IBM DB2의 경우에만).

```
DriverUnicodeType=1
```

참고: `libspssodbc_datadirect_utf16.so` 랩퍼는 다른 SPSS Modeler Server 지원 ODBC 드라이버와도 호환 가능합니다.

데이터베이스 연결에 대한 사전 설정된 값 지정

일부 데이터베이스의 경우 데이터베이스 연결에 대해 다수의 기본 설정을 지정할 수 있습니다. 이 설정은 모두 데이터베이스 내보내기에 적용됩니다.

이 기능을 지원하는 데이터베이스 유형은 다음과 같습니다.

- IBM InfoSphere Warehouse. 자세한 정보는 『IBM DB2 InfoSphere Warehouse에 대한 설정』의 내용을 참조하십시오.
- SQL Server Enterprise 및 Developer Edition. 자세한 정보는 『SQL Server에 대한 설정』의 내용을 참조하십시오.
- Oracle Enterprise 또는 Personal Edition. 자세한 정보는 24 페이지의 『Oracle에 대한 설정』의 내용을 참조하십시오.
- IBM Netezza, IBM DB2 for z/OS 및 Teradata는 모두 비슷한 방식으로 데이터베이스 또는 스키마에 연결됩니다. 자세한 정보는 24 페이지의 『IBM Netezza, IBM DB2 for z/OS, IBM DB2 LUW 및 Teradata에 대한 설정』의 내용을 참조하십시오.

이 기능을 지원하지 않는 데이터베이스 또는 스키마에 연결되는 경우에는 **이 데이터베이스 연결에는 사전 설정을 구성할 수 없음** 메시지가 표시됩니다.

IBM DB2 InfoSphere Warehouse에 대한 설정

이 설정은 IBM InfoSphere Warehouse에 대해 표시됩니다.

테이블스페이스 내보내기에 사용할 테이블스페이스입니다. 데이터베이스 관리자는 테이블스페이스를 파티션되도록 작성하거나 구성할 수 있습니다. 기본 테이블스페이스 대신 내보내기에 사용할 이 테이블스페이스 중 하나를 선택하는 것이 좋습니다.

압축 사용. 선택된 경우 압축을 사용하여 내보낼 테이블을 작성합니다(예: SQL의 `CREATE TABLE MYTABLE(...) COMPRESS YES;`와 동등).

업데이트를 로그하지 않음 선택된 경우에는 테이블 작성 및 데이터 삽입 시 로깅을 방지합니다(SQL의 `CREATE TABLE MYTABLE(...) NOT LOGGED INITIALLY;`와 동등).

SQL Server에 대한 설정

이 설정은 SQL Server Enterprise 및 Developer Edition에 대해 표시됩니다.

압축 사용. 선택된 경우 압축을 사용하여 내보낼 테이블을 작성합니다.

압축. 압축의 수준을 선택하십시오.

- **행.** 행 수준 압축을 사용으로 설정합니다(예: SQL의 `CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = ROW);`와 동등).
- **페이지.** 페이지 수준 압축을 사용으로 설정합니다(예: SQL의 `CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = PAGE);`).

Oracle에 대한 설정

Oracle 설정 - 기본 옵션

이 설정은 기본 옵션을 사용하는 Oracle Enterprise 또는 Personal Edition에 대해 표시됩니다.

압축 사용. 선택된 경우 압축을 사용하여 내보낼 테이블을 작성합니다.

압축. 압축의 수준을 선택하십시오.

- **기본값.** 기본 압축을 사용으로 설정합니다(예: SQL의 `CREATE TABLE MYTABLE(...) COMPRESS;`). 이 케이스에서 이는 **기본** 옵션과 동일한 효과를 가집니다.
- **기본.** 기본 압축을 사용으로 설정합니다(예: SQL의 `CREATE TABLE MYTABLE(...) COMPRESS BASIC;`).

Oracle 설정 - 고급 옵션

이 설정은 고급 옵션을 사용하는 Oracle Enterprise 또는 Personal Edition에 대해 표시됩니다.

압축 사용. 선택된 경우 압축을 사용하여 내보낼 테이블을 작성합니다.

압축. 압축의 수준을 선택하십시오.

- **기본값.** 기본 압축을 사용으로 설정합니다(예: SQL의 `CREATE TABLE MYTABLE(...) COMPRESS;`). 이 케이스에서 이는 **기본** 옵션과 동일한 효과를 가집니다.
- **기본.** 기본 압축을 사용으로 설정합니다(예: SQL의 `CREATE TABLE MYTABLE(...) COMPRESS BASIC;`).
- **OLTP.** OLTP 압축을 사용으로 설정합니다(예: SQL의 `CREATE TABLE MYTABLE(...)COMPRESS FOR OLTP;`).
- **쿼리 낮음/높음.** (Exadata 서버 전용) 쿼리에 대해 HCC(Hybrid Columnar Compression)를 사용으로 설정합니다(예: SQL의 `CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY LOW;` 또는 `CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY HIGH;`). 쿼리에 대한 압축은 데이터 웨어하우징 환경에서 유용합니다. `HIGH`는 `LOW`보다 높은 압축 비율을 제공합니다.
- **아카이브 낮음/높음.** (Exadata 서버 전용) 아카이브에 대해 HCC(Hybrid Columnar Compression)를 사용으로 설정합니다(예: SQL의 `CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE LOW;` 또는 `CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE HIGH;`). 아카이브에 대한 압축은 장기간 저장될 데이터를 압축하는 경우에 유용합니다. `HIGH`는 `LOW`보다 높은 압축 비율을 제공합니다.

IBM Netezza, IBM DB2 for z/OS, IBM DB2 LUW 및 Teradata에 대한 설정

IBM Netezza, IBM DB2 for z/OS, IBM DB2 LUW 또는 Teradata에 대한 사전 설정을 지정하면 다음 항목 중에서 선택하라는 프롬프트가 표시됩니다.

서버 스코어링 어댑터 데이터베이스 사용 또는 **서버 스코어링 어댑터 스키마 사용**. 선택된 경우 **서버 스코어링 어댑터 데이터베이스** 또는 **서버 스코어링 어댑터 스키마** 옵션을 사용할 수 있게 합니다.

서버 스코어링 어댑터 데이터베이스 또는 **서버 스코어링 어댑터 스키마** 드롭 다운 목록에서 필요한 연결을 선택하십시오.

또한 Teradata의 경우 쿼리 밴딩 세부사항을 설정하여 워크로드 관리, 쿼리 조합, 식별 및 분석, 데이터베이스 사용 추적 등의 항목을 지원하기 위한 추가적인 메타데이터를 제공할 수도 있습니다.

쿼리 밴딩 스펠링. Teradata 데이터베이스 연결에 대해 작업하는 전체 시간에 대해 한 번 쿼리 밴딩을 설정할 경우(**세션의 경우**) 또는 스트림을 실행할 때마다 쿼리 밴딩을 설정할 경우(**트랜잭션의 경우**) 선택하십시오.

참고: 스트림에서 쿼리 밴딩을 설정하는 경우에는 해당 스트림을 다른 시스템으로 복사하면 밴딩이 유실됩니다. 이를 방지하기 위해 스크립팅을 사용하여 스트림을 실행하고 스크립트에 키워드 *querybanding*을 사용하여 필요한 설정을 적용할 수 있습니다.

필요한 데이터베이스 권한

SPSS Modeler 데이터베이스 기능이 올바르게 작동하게 하려면 다음과 같은 항목에 대한 액세스를 사용된 모든 사용자 ID에 부여하십시오.

DB2 LUW

SYSIBM.SYSDUMMY1

SYSIBM.SYSFOREIGNKEYS

SYSIBM.SYSINDEXES

SYSIBM.SYSKEYCOLUSE

SYSIBM.SYSKEYS

SYSIBM.SYSPARMS

SYSIBM.SYSRELS

SYSIBM.SYSROUTINES

SYSIBM.SYSROUTINES_SRC

SYSIBM.SYSSYNONYMS

SYSIBM.SYSTABCONST

SYSIBM.SYSTABCONSTPKC

SYSIBM.SYSTABLES

SYSIBM.SYSTRIGGERS

SYSIBM.SYSVIEWDEP

SYSIBM.SYSVIEWS

SYSCAT.TABLESPACES

SYSCAT.SCHEMATA

DB2/z SYSIBM.SYSDUMMY1

SYSIBM.SYSFOREIGNKEYS

SYSIBM.SYSINDEXES

SYSIBM.SYSKEYCOLUSE

SYSIBM.SYSKEYS

SYSIBM.SYSPARMS

SYSIBM.SYSRELS

SYSIBM.SYSROUTINES

SYSIBM.SYSROUTINES_SRC

SYSIBM.SYSSYNONYMS

SYSIBM.SYSTABCONST

SYSIBM.SYSTABCONSTPKC

SYSIBM.SYSTABLES

SYSIBM.SYSTRIGGERS

SYSIBM.SYSVIEWDEP

SYSIBM.SYSVIEWS

SYSIBM.SYSDUMMYU

SYSIBM.SYSPACKSTMT

Netezza

_V_FUNCTION

_V_DATABASE

Teradata

DBC.Functions

DBC.USERS

데이터베이스 테이블 선택

데이터 소스에 연결한 후 특정 테이블 또는 보기에서 필드를 가져오도록 선택할 수 있습니다. 데이터베이스 대화 상자의 데이터 탭에서 테이블 이름 필드에 테이블의 이름을 입력하거나 **선택**을 클릭하여 사용 가능한 테이블 및 보기를 나열하는 테이블/보기 선택 대화 상자를 열 수 있습니다.

테이블 소유자 표시. 사용자가 테이블에 액세스하려면 먼저 테이블의 소유자를 지정하도록 데이터 소스에서 요구하는 경우 선택하십시오. 이 요구사항을 가지고 있지 않은 데이터 소스의 경우에는 이 옵션을 선택 취소하십시오.

참고: SAS 및 Oracle 데이터베이스는 일반적으로 테이블 소유자를 표시하도록 요구합니다.

테이블/보기. 가져올 테이블 또는 보기를 선택하십시오.

표시. 현재 연결된 데이터 소스의 열을 나열합니다. 다음 옵션 중 하나를 클릭하여 사용 가능한 테이블의 보기를 사용자 정의하십시오.

- **사용자 테이블**을 클릭하여 데이터베이스 사용자가 작성한 일반적인 데이터베이스 테이블을 보십시오.
- **시스템 테이블**을 클릭하여 시스템이 소유한 데이터베이스 테이블을 보십시오(예: 인덱스 세부사항 등의 데이터베이스에 대한 정보를 제공하는 테이블). 이 옵션은 Excel 데이터베이스에서 사용되는 탭을 보는 데 사용할 수 있습니다. (별도의 Excel 소스 노드도 사용할 수 있습니다. 자세한 정보는 46 페이지의 『Excel 소스 노드』의 내용을 참조하십시오.)
- **보기**를 클릭하여 하나 이상의 일반적인 테이블과 관련된 쿼리를 기반으로 가상 테이블을 보십시오.
- **동의어**를 클릭하여 기존 테이블에 대해 데이터베이스에서 작성된 동의어를 보십시오.

이름/소유자 필터. 이 필드를 사용하면 이름 또는 소유자별로 표시된 테이블의 목록을 필터링할 수 있습니다. 예를 들어, SYS를 입력하여 해당 소유자를 가진 테이블만 나열하십시오. 와일드카드 검색의 경우 밑줄(_)은 단일 문자를 나타내는 데 사용할 수 있고 퍼센트 부호(%)는 0개 이상 문자의 시퀀스를 나타낼 수 있습니다.

기본값으로 설정. 현재 설정을 현재 사용자의 기본값으로 저장합니다. 이 설정은 사용자가 나중에 *동일한 데이터 소스 이름 및 사용자 로그인 전용* 새 테이블 선택기 대화 상자를 열 때 복원됩니다.

데이터베이스 쿼리

일단 데이터 소스에 연결되면 SQL 쿼리를 사용하여 필드를 가져올 수 있습니다. 주 대화 상자에서 연결 모드로 **SQL 쿼리**를 선택하십시오. 그러면 쿼리 편집기 창이 대화 상자에 추가됩니다. 쿼리 편집기를 사용하면 결과 세트를 데이터 스트림으로 읽어 올 하나 이상의 SQL 쿼리를 작성하거나 로드할 수 있습니다.

다중 SQL 쿼리를 지정하는 경우, 세미콜론(;)으로 구분하고 다중 SELECT문이 없는지 확인하십시오.

쿼리 편집기 창을 취소하거나 닫으려면 연결 모드로 **테이블**을 선택하십시오.

SQL 쿼리에 SPSS Modeler 스트림 매개변수(사용자 정의 변수의 유형)를 포함할 수 있습니다. 자세한 정보는 28 페이지의 『SQL 쿼리에서 스트림 매개변수 사용』의 내용을 참조하십시오.

쿼리 로드. 이전에 저장한 쿼리를 로드하는 데 사용할 수 있는 파일 브라우저를 열려면 클릭하십시오.

쿼리 저장. 현재 쿼리를 저장하는 데 사용할 수 있는 쿼리 저장 대화 상자를 열려면 클릭하십시오.

기본값 가져오기. 대화 상자에서 선택된 테이블 및 열을 사용하여 자동으로 구성된 SQL SELECT 명령문 예를 가져오려면 클릭하십시오.

지우기. 작업 영역의 내용을 지웁니다. 다시 시작하려면 이 옵션을 사용하십시오.

텍스트 분할. 기본 옵션인 **사용 안 함**은 쿼리가 데이터베이스에 하나로 전송됨을 의미합니다. 또는 SPSS Modeler에서 쿼리를 구문 분석하려고 시도하고 데이터베이스에 하나씩 차례로 전송되어야 하는 SQL문이 있는지 식별하는 **필요시 사용**을 선택할 수 있습니다.

SQL 쿼리에서 스트림 매개변수 사용

필드를 가져오기 위해 SQL 쿼리를 작성하는 경우 이전에 정의된 SPSS Modeler 스트림 매개변수를 포함할 수 있습니다. 모든 유형의 스트림 매개변수가 지원됩니다.

다음 표에는 SQL 쿼리에서 스트림 매개변수의 일부 예제가 해석되는 방식을 보여줍니다.

표 3. 스트림 매개변수의 예.

스트림 매개변수 이름(예)	저장 공간	스트림 매개변수 값	해석
PString	문자열	ss	'ss'
PInt	정수	5	5
PReal	실수	5.5	5.5
PTime	시간	23:05:01	t{'23:05:01'}
PDate	날짜	2011-03-02	d{'2011-03-02'}
PTimeStamp	시간소인	2011-03-02 23:05:01	ts{'2011-03-02 23:05:01'}
PColumn	알 수 없음	IntValue	IntValue

SQL 쿼리에서는 '$P-<parameter_name>'이라는 CLEM 표현식과 동일한 방식으로 스트림 매개변수를 지정합니다. 여기서 <parameter_name>은 스트림 매개변수에 대해 정의된 이름입니다.

필드를 참조하는 경우에는 저장 유형이 알 수 없음으로 정의되어야 하며 필요한 경우 매개변수 값을 따옴표로 묶어야 합니다. 따라서 표에 있는 예제를 사용하여 다음의 SQL 쿼리를 입력한 경우:

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

다음과 같이 평가됩니다.

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

PColumn 매개변수를 사용하여 IntValue 필드를 참조해야 하는 경우에는 동일한 결과를 얻기 위해 다음과 같이 쿼리를 지정해야 합니다.

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

가변파일 노드

가변파일 노드를 사용하여 자유 필드 텍스트 파일(해당 레코드에 일정한 수의 숫자와 일정하지 않은 수의 문자가 포함되는 파일이며 분리 텍스트 파일이라고도 함)에서 데이터를 읽을 수 있습니다. 이 유형의 노드는 또한 고정 길이 헤더 텍스트와 특정 유형의 주석(Annotation)이 있는 파일에도 유용합니다. 레코드는 한 번에 하나씩 읽히며 전체 파일이 읽혀질 때까지 스트림을 통해 전달됩니다.

지리 공간적 데이터 읽기 관련 주의사항

노드에 지리 공간적 데이터가 포함되고 노드가 플랫 파일에서 내보내기로 작성된 경우, 일부 추가 단계를 수행하여 지리 공간적 메타데이터를 설정해야 합니다. 자세한 정보는 31 페이지의 『가변파일 노드에 지리 공간적 데이터 가져오기』의 내용을 참조하십시오.

분리 텍스트 데이터 읽기 주의사항

- 레코드는 각 행의 끝에서 줄 바꾸기 문자를 사용하여 구분해야 합니다. 다른 목적으로(예를 들어, 필드 이름 또는 값 내에서) 줄 바꾸기 문자를 사용해서는 안 됩니다. 꼭 필요한 것은 아니지만 공간을 절약하기 위해 선행 및 후미 공백을 제거하는 것이 좋습니다. 선택적으로 노드에서 이러한 공백을 제거할 수 있습니다.
- 필드는 쉼표 또는 다른 문자(필드 이름 또는 값에서 사용되지 않고 구분자로만 사용되는 문자가 좋음)로 구분해야 합니다. 이것이 불가능한 경우, 큰따옴표가 포함된 필드 이름 또는 텍스트 값이 없는 한, 모든 텍스트 필드를 큰따옴표로 묶을 수 있습니다. 필드 이름 또는 값에 큰따옴표가 있으면, 값에서 작은따옴표가 사용되지 않는 한, 대안으로서 텍스트 필드를 작은따옴표로 묶을 수 있습니다. 큰따옴표와 작은따옴표 둘 다 사용할 수 없는 경우에는 텍스트 값을 수정하여 구분 문자나 작은따옴표 또는 큰따옴표를 제거 또는 대체해야 합니다.
- 헤더 행을 포함하여 각 행은 같은 수의 필드를 포함해야 합니다.
- 첫 번째 행은 필드 이름을 포함해야 합니다. 그렇지 않은 경우, **파일에서 필드 이름 읽기**를 선택 취소하여 각 필드에 Field1, Field2 등의 일반 이름을 부여하십시오.
- 두 번째 행은 데이터의 첫 번째 레코드를 포함해야 합니다. 공백 행 또는 주석이 없어야 합니다.
- 숫자 값에는 천단위 구분 문자 또는 그룹화 기호가 없어야 합니다(예를 들어, 3,000.00에서 쉼표가 없어야 함). 소수점 표시기(US 또는 UK에서는 마침표)는 적합한 곳에서만 사용해야 합니다.
- 날짜 및 시간 값은 스트림 옵션 대화 상자에서 인식되는 형식 중 하나여야 합니다(예: `DD/MM/YYYY` 또는 `HH:MM:SS`). 파일의 모든 날짜 및 시간 필드는 동일한 형식을 따르는 것이 좋으며, 날짜를 포함하는 필드는 해당 필드 내의 모든 값에 동일한 형식을 사용해야 합니다.

가변파일 노드의 옵션 설정

가변파일 대화 상자의 파일 탭에 있는 옵션을 설정합니다.

파일 파일의 이름을 지정하십시오. 파일 이름을 입력하거나 생략 기호 단추(**...**)를 클릭하여 파일을 선택할 수 있습니다. 파일을 선택하면 파일 경로가 표시되고 아래의 분할창에 해당 컨텐츠가 구분자와 함께 표시됩니다.

데이터 소스에서 표시되는 표본 텍스트는 복사하여 EOL 주석 문자 및 사용자 지정 구분자에 붙여넣을 수 있습니다. Ctrl-C 및 Ctrl-V를 사용하여 복사하고 붙여넣으십시오.

파일에서 필드 이름 읽기 기본적으로 선택되는 이 옵션은 데이터 파일의 첫 번째 행을 열의 레이블로 처리합니다. 첫 번째 행이 헤더가 아닌 경우에는 이 옵션을 선택 취소하십시오. 그러면 데이터 세트의 필드 수만큼 각 필드에 일반 이름(예: *Field1, Field2*)이 자동으로 부여됩니다.

필드 수 지정. 각 레코드의 필드 수를 지정하십시오. 레코드가 줄 바꾸기로 종료되는 경우 필드 수를 자동으로 발견할 수 있습니다. 수동으로 수를 설정할 수도 있습니다.

헤더 문자 건너뛰기. 첫 번째 레코드의 시작 부분에서 무시할 문자 수를 지정하십시오.

EOL 주석 문자. 문자(예: # 또는 !)를 지정하여 데이터에서 주석(Annotation)을 표시하십시오. 데이터 파일에서 이 문자 중 하나가 표시되는 경우마다 다음 줄 바꾸기 문자를 제외하고 이 문자까지의 모든 문자가 무시됩니다.

선행 및 후미 공백 제거. 가져올 때 문자열의 선행 및 후미 공백을 삭제하는 옵션을 선택하십시오.

참고. SQL 푸시백을 사용하는 문자열과 사용하지 않는 문자열 간의 비교는 후미 공백이 존재하는 서로 다른 결과를 생성할 수 있습니다.

유효하지 않은 문자. 데이터 소스에서 유효하지 않은 문자를 제거하려면 **삭제**를 선택하십시오. 유효하지 않은 문자를 지정된 기호(한 문자만)로 바꾸려면 **바꿀 문자열**을 선택하십시오. 유효하지 않은 문자는 널 문자이거나 지정된 인코딩 방법에 존재하지 않는 문자입니다.

인코딩. 사용되는 텍스트 인코딩 방법을 지정합니다. 시스템 기본값, 스트림 기본값 또는 UTF-8 중에서 선택할 수 있습니다.

- 시스템 기본값은 Windows 제어판에 지정되어 있거나 분산 모드에서 실행 중인 경우 서버 컴퓨터에 지정되어 있습니다.
- 스트림 기본값은 스트림 특성 대화 상자에서 지정됩니다.

소수점 기호 데이터 소스에서 사용되는 소수점 구분 문자의 유형을 선택하십시오. **스트림 기본값**은 스트림 특성 대화 상자의 옵션 탭에서 선택된 문자입니다. 또는 **마침표(.)** 또는 **쉼표(,)**를 선택하고 선택된 문자를 소수점 구분 문자로 사용하여 이 대화 상자의 모든 데이터를 읽을 수 있습니다.

행 구분자가 줄 바꾸기 문자임 필드 구분자 대신, 행 구분자로 줄 바꾸기 문자를 사용하려면 이 옵션을 선택하십시오. 이 옵션은 하나의 행에 줄 바꿈을 초래하는 홀수 개의 구분자가 있는 경우에 유용할 수 있습니다. 이 옵션을 선택하는 경우 구분자 목록에서 **줄 바꾸기**를 선택할 수 없습니다.

참고: 이 옵션을 선택하면 데이터 행 끝에 있는 공백 값이 제거됩니다.

구분자. 이 제어에 대해 나열된 선택란을 사용하면 파일에서 필드 경계를 정의하는 문자(예: 쉼표(,))를 지정할 수 있습니다. 여러 구분자를 사용하는 레코드의 경우 둘 이상의 구분자(예: ", |")를 지정할 수도 있습니다. 기본 구분자는 쉼표입니다.

참고: 쉼표가 소수점 구분 문자로도 정의되어 있는 경우에는 여기서의 기본 설정이 작동하지 않습니다. 쉼표가 필드 구분자인 동시에 소수점 구분 문자인 경우에는 구분자 목록에서 **기타**를 선택하십시오. 그런 다음 입력 필드에서 수동으로 쉼표를 지정하십시오.

인접한 여러 공백 구분 문자를 하나의 구문자로 처리하려면 **여러 공백 구분자 허용**을 선택하십시오. 예를 들어, 한 데이터 값 뒤에 4개의 공백이 있고 그 뒤에 다른 데이터 값이 있는 경우 이 그룹은 5개가 아니라 2개의 필드로 처리됩니다.

열 및 유형에 대해 스캔할 행 수 지정된 데이터 유형을 찾기 위해 스캔할 행 및 열 수를 지정하십시오.

날짜 및 시간 자동 인식 IBM SPSS Modeler에서 데이터 항목을 날짜 또는 시간으로 자동으로 인식할 수 있게 하려면 이 선택란을 선택하십시오. 예를 들어, 이는 07-11-1965와 같은 항목이 날짜로 식별되고 02:35:58은 시간으로 식별됨을 의미합니다. 그러나 07111965 또는 023558과 같은 모호한 항목은 숫자 사이에 구분자가 없으므로 정수로 표시됩니다.

참고: 이전 버전의 IBM SPSS Modeler에 있는 데이터 파일을 사용할 때 발생할 수 있는 잠재적 데이터 문제점을 피하기 위해, 13 이전의 버전에 저장된 정보에 대해서는 기본적으로 이 선택란이 꺼져 있습니다.

대괄호를 목록으로 처리 이 선택란을 선택하면, 여는 대괄호와 닫는 대괄호 사이에 있는 데이터는 해당 컨텐츠에 쉼표 및 큰따옴표와 같은 구분 문자가 포함된 경우에도 단일 값으로 처리됩니다. 예를 들어, 2차원 또는 3차원 지리 공간적 데이터가 포함될 수 있으며, 이 경우 대괄호 안에 있는 좌표는 단일 목록 항목으로 처리됩니다. 자세한 정보는 『가변파일 노드에 지리 공간적 데이터 가져오기』의 내용을 참조하십시오.

따옴표. 드롭 다운 목록을 사용하면 가져오기를 수행할 때 작은따옴표와 큰따옴표가 처리되는 방식을 지정할 수 있습니다. 모든 따옴표를 **삭제**하거나 필드 값에 포함시켜 **텍스트로 포함**하거나 **대응 및 삭제**하여 따옴표 쌍을 일치시킨 후 제거하도록 선택할 수 있습니다. 따옴표가 일치되지 않는 경우 오류 메시지가 수신됩니다. **삭제**와 **대응 및 삭제**는 모두 필드 값(따옴표 제외)을 문자열로 저장합니다.

참고: 대응 및 삭제를 사용하는 경우에는 공백이 유지됩니다. **삭제**를 사용하는 경우에는 따옴표 내부 및 외부의 후미 공백이 제거됩니다(예: **' " ab c" , "d ef " , " gh i " '**는 **'ab c, d ef, gh i'**라는 결과를 생성함). **텍스트로 포함**을 사용하는 경우에는 따옴표가 정상적인 문자로 처리되므로 선행 및 후미 공백은 자연스럽게 제거됩니다.

이 대화 상자에서 작업하는 중 언제든 **새로 고치기**를 클릭하여 데이터 소스에서 필드를 다시 로드할 수 있습니다. 이는 소스 노드와의 데이터 연결을 변경하거나 대화 상자의 탭들 사이에서 작업 중일 때 유용합니다.

가변파일 노드에 지리 공간적 데이터 가져오기

노드가 지리 공간적 데이터를 포함하고 플랫 파일에서 내보내기로 작성되었으며 작성된 스트림에서 사용되는 경우, 노드는 지리 공간적 메타데이터를 유지하며 추가 구성 단계가 필요하지 않습니다.

그러나, 노드를 내보내어 다른 스트림에서 사용하는 경우 지리 공간적 목록 데이터는 문자열 형식으로 자동으로 변환됩니다. 일부 추가 단계를 수행하여 목록 저장 유형 및 연관된 지리 공간적 메타데이터를 복원해야 합니다.

목록에 대한 자세한 정보는 11 페이지의 『목록 저장 공간 및 연관된 측정 수준』의 내용을 참조하십시오.

지리 공간적 메타데이터로 설정할 수 있는 세부사항에 대한 자세한 정보는 142 페이지의 『지리 공간적 측정 수준』의 내용을 참조하십시오.

지리 공간적 메타데이터를 설정하려면 다음 단계를 사용하십시오.

1. 가변파일 노드의 파일 탭에서 **대괄호를 목록으로 처리** 선택란을 선택하십시오. 이 선택란을 선택하면, 여는 대괄호와 닫는 대괄호 사이에 있는 데이터는 해당 컨텐츠에 쉼표 및 큰따옴표와 같은 구분 문자가 포함된 경우에도 단일 값으로 처리됩니다. 이 선택란을 선택하지 않는 경우, 데이터는 문자열 저장 유형으로 읽혀지고 필드의 모든 쉼표는 분리자로 처리되어 데이터 구조가 올바르지 않게 해석됩니다.
2. 데이터에 작은따옴표 또는 큰따옴표가 포함된 경우, 필요에 따라 **작은따옴표** 및 **큰따옴표** 필드에서 **짝 짓기 및 삭제** 옵션을 선택하십시오.

3. 가변파일 노드의 데이터 탭에서, 지리 공간적 데이터 필드에 대해, **대체** 선택란을 선택하고 **저장** 유형을 문자열에서 목록으로 변경하십시오.
4. 기본적으로, 목록 **저장** 유형은 *실수* 목록으로서 설정되고 목록 필드의 기본 값 저장 유형은 *실수*로 설정됩니다. 기본 값 저장 유형 또는 깊이를 변경하려면 **지정...**을 클릭하여 저장 공간 부속 대화 상자를 표시하십시오.
5. 저장 공간 부속 대화 상자에서 다음 설정을 수정할 수 있습니다.
   - **저장 공간** 데이터 필드의 전체 저장 유형을 지정하십시오. 기본적으로 저장 유형이 목록으로 설정됩니다. 그러나 드롭 다운 목록에 기타 모든 저장 유형(문자열, 정수, 실수, 날짜, 시간 및 시간소인)이 포함됩니다. 목록 이외의 저장 유형을 선택하는 경우에는 **값 저장 공간** 및 **깊이** 옵션을 사용할 수 없습니다.
   - **값 저장 공간** 필드 전체가 아니라 목록에 있는 요소의 저장 유형을 지정하십시오. 지리 공간적 필드를 가져올 때 관련된 유일한 저장 유형은 실수 및 정수입니다. 기본 설정은 실수입니다.
   - **깊이** 목록 필드의 깊이를 지정하십시오. 필요한 깊이는 지리 공간적 필드의 유형에 따라 다르며 다음 기준을 따릅니다.
     - 점 – 0
     - LineString – 1
     - 다각형 – 1
     - 다중 점 – 1
     - 다중 LineString – 2
     - 다중 다각형 – 2

     참고: 다시 목록으로 변환할 지리 공간적 필드의 유형과 그러한 종류의 필드에 필요한 깊이를 알아야 합니다. 이 정보가 올바르게 설정되어 있지 않으면 해당 필드를 사용할 수 없습니다.
6. 가변파일 노드의 유형 탭에서 지리 공간적 데이터 필드에 대한 **측정** 셀에 올바른 측정 수준이 포함되는지 확인하십시오. 측정 수준을 변경하려면 **측정** 셀에서 **지정...**을 클릭하여 값 대화 상자를 표시하십시오.
7. 값 대화 상자에서 목록에 대한 **측정**, **저장 공간** 및 **깊이**가 표시됩니다. **값 및 레이블 지정** 옵션을 선택하고 **유형** 드롭 다운 목록에서 **측정**에 올바른 유형을 선택하시시오. **유형**에 따라, 추가 세부사항(예: 데이터가 2차원을 나타내는지 또는 3차원을 나타내는지 여부와 사용되는 좌표계)을 요구하는 프롬프트가 표시될 수 있습니다.

고정 파일 노드

고정 파일 노드를 사용하여 고정 필드 텍스트 파일(필드가 구분되지 않지만 동일한 위치에서 시작하며 길이가 고정된 파일)에서 데이터를 가져올 수 있습니다. 머신 생성 또는 레거시 데이터가 자주 고정 필드 형식으로 저장됩니다. 고정 파일 노드의 파일 탭을 사용하면 데이터에서 열의 위치 및 길이를 쉽게 지정할 수 있습니다.

고정 파일 노드에 대한 옵션 설정

고정 파일 노드의 파일 탭을 사용하면 데이터를 IBM SPSS Modeler로 가져오고 레코드의 길이 및 열의 위치를 지정할 수 있습니다. 대화 상자의 가운데에 있는 데이터 미리보기 분할창을 사용하면 클릭하여 필드 간 중단점을 지정하는 화살표를 추가할 수 있습니다.

파일. 파일의 이름을 지정하십시오. 파일 이름을 입력하거나 생략 기호 단추(**...**)를 클릭하여 파일을 선택할 수 있습니다. 파일을 선택하고 나면 파일 경로가 표시되고 해당 컨텐츠가 구분자와 함께 아래의 패널에 표시됩니다.

데이터 미리보기 분할창은 열 위치 및 길이를 지정하는 데 사용할 수 있습니다. 미리보기 창의 맨 위에 있는 눈금자를 사용하면 변수의 길이를 측정하고 변수 간 중단점을 지정할 수 있습니다. 필드 위의 눈금자 영역을 클릭하여 중단점 행을 지정할 수 있습니다. 중단점은 끌어서 이동할 수 있으며 데이터 미리보기 영역 밖으로 끌어서 삭제할 수 있습니다.

- 각각의 중단점 행은 아래의 필드 테이블에 새 필드를 자동으로 추가합니다.
- 화살표에 의해 표시된 시작 위치는 아래 테이블의 시작 열에 자동으로 추가됩니다.

행 지향. 각 레코드의 끝에서 줄 바꾸기 문자를 건너뛰려면 선택하십시오.

헤더 행 건너뛰기. 첫 번째 레코드의 시작 부분에서 무시할 행 수를 지정하십시오. 이는 열 헤더를 무시하는 경우에 유용합니다.

레코드 길이. 각 레코드의 문자 수를 지정하십시오.

필드. 이 데이터 파일에 대해 정의한 모든 필드가 여기에 나열됩니다. 두 가지 방법으로 필드를 정의합니다.

- 위의 데이터 미리보기 분할창을 사용하여 대화식으로 필드를 지정하십시오.
- 아래의 테이블에 비어 있는 필드 행을 추가하여 수동으로 필드를 지정하십시오. 필드 분할창 오른쪽의 단추를 클릭하여 새 필드를 추가하십시오. 그런 다음 비어 있는 필드에서 필드 이름, 시작 위치 및 길이를 입력하십시오. 이 옵션은 데이터 미리보기 분할창에 화살표를 자동으로 추가하며 이는 쉽게 조정될 수 있습니다.

이전에 정의된 필드를 제거하려면 목록에서 해당 필드를 선택한 후 빨간색 삭제 단추를 클릭하십시오.

시작. 필드에서 첫 번째 문자의 위치를 지정하십시오. 예를 들어, 레코드의 두 번째 필드가 16번째 문자에서 시작하면 16을 시작점으로 입력합니다.

길이. 각 필드에 대해 가장 긴 값에 있는 문자의 수를 지정하십시오. 이는 다음 필드의 절사 지점을 결정합니다.

선행 및 후미 공백 제거. 가져올 때 문자열에서 선행 및 후미 공백을 삭제하려면 선택하십시오.

참고. SQL 푸시백을 사용하는 문자열과 사용하지 않는 문자열 간의 비교는 후미 공백이 존재하는 서로 다른 결과를 생성할 수 있습니다.

유효하지 않은 문자. 데이터 입력에서 유효하지 않은 문자를 제거하려면 **삭제**를 선택하십시오. 유효하지 않은 문자를 지정된 기호(한 문자만)로 바꾸려면 **바꿀 문자열**을 선택하십시오. 유효하지 않은 문자는 널(0) 문자 또는 현재 인코딩에 존재하지 않는 모든 문자입니다.

인코딩. 사용되는 텍스트 인코딩 방법을 지정합니다. 시스템 기본값, 스트림 기본값 또는 UTF-8 중에서 선택할 수 있습니다.

- 시스템 기본값은 Windows 제어판에 지정되어 있거나 분산 모드에서 실행 중인 경우 서버 컴퓨터에 지정되어 있습니다.
- 스트림 기본값은 스트림 특성 대화 상자에서 지정됩니다.

소수점 기호. 데이터 소스에서 사용되는 소수점 구분 문자의 유형을 선택하십시오. **스트림 기본값**은 스트림 특성 대화 상자의 옵션 탭에서 선택된 문자입니다. 그렇지 않으면 **마침표(.)** 또는 **쉼표(,)**를 선택하여 선택된 문자를 소수점 구분 문자로 사용하여 이 대화 상자의 모든 데이터를 읽으십시오.

날짜 및 시간 자동 인식.IBM SPSS Modeler가 자동으로 데이터 항목을 날짜 또는 시간으로 인식할 수 있게 하려면 이 선택란을 선택하십시오. 예를 들어, 이는 07-11-1965와 같은 항목은 날짜로 식별되고 02:35:58은 시간으로 식별됨을 의미합니다. 하지만 07111965 또는 023558과 같이 모호한 항목은 숫자 사이에 구분자가 없기 때문에 정수로 표시됩니다.

참고: 이전 IBM SPSS Modeler 버전의 데이터 파일을 사용할 때 잠재적인 데이터 문제점을 방지하기 위해 13 이전의 버전에서 저장된 정보에 대해서는 이 선택란이 기본적으로 비활성화되어 있습니다.

유형에 대해 스캔할 행. 지정된 데이터 유형에 대해 스캔할 행 수를 지정하십시오.

이 대화 상자에서 작업하는 동안 어느 지점에서나 **새로 고치기**를 클릭하여 데이터 소스의 필드를 재로드하십시오. 이는 소스 노드에 대한 데이터 연결을 변경하거나 대화 상자의 탭 사이에서 작업할 때 유용합니다.

Data Collection 노드

Data Collection 소스 노드는 Data Collection 제품과 함께 제공되는 Survey Reporter Developer Kit을 기반으로 설문조사 데이터를 가져옵니다. 이 형식은 *케이스 데이터*(설문조사 중에 수집된 질문에 대한 실제 응답)를 케이스 데이터의 수집 및 구성 방식에 대해 설명하는 *메타데이터*와 구별합니다. 메타데이터는 케이스 데이터의 구조 정의, 질문 텍스트, 변수 이름 및 설명, 다중 응답 변수 정의, 텍스트 문자열의 변환 등의 정보로 구성됩니다.

참고: 이 노드를 사용하려면 Data Collection 제품과 함께 배포되는 Survey Reporter Developer Kit이 필요합니다. 이 Developer Kit을 설치하는 것 외에는 추가적인 구성이 필요하지 않습니다.

설명

- 설문조사 데이터는 표 형식 플랫 VDATA 형식에서 읽어오며 메타데이터 소스를 포함하는 경우에는 계층적 HDATA 형식의 소스에서 읽어옵니다.
- 유형은 메타데이터의 정보를 사용하여 자동으로 인스턴스화됩니다.

- 설문조사 데이터를 SPSS Modeler로 가져오면 질문이 각 응답자에 대한 레코드가 포함된 필드로 렌더링됩니다.

Data Collection 파일 가져오기 옵션

Data Collection 노드의 파일 탭에서는 가져올 메타데이터 및 케이스 데이터에 대한 옵션을 지정할 수 있습니다.

메타데이터 설정

참고: 사용 가능한 제공자 파일 유형의 전체 목록을 보려면 Data Collection 소프트웨어와 함께 사용 가능한 Survey Reporter Developer Kit을 설치해야 합니다.

메타데이터 제공자. Data Collection Survey Reporter Developer Kit에서 지원하는 대로 다수의 형식에서 설문조사 데이터를 가져올 수 있습니다. 사용 가능한 제공자 유형은 다음과 같습니다.

- **DataCollectionMDD.** 질문지 정의 파일(*.mdd*)에서 메타데이터를 읽어옵니다. 표준 Data Collection 데이터 모델 형식입니다.
- **ADO 데이터베이스.** ADO 파일에서 케이스 데이터 및 메타데이터를 읽어옵니다. 메타데이터가 포함된 *.adoinfo* 파일의 이름 및 위치를 지정하십시오. 이 DSC의 내부 이름은 *mrADODsc*입니다.
- **In2data 데이터베이스.** In2data 케이스 데이터 및 메타데이터를 읽습니다. 이 DSC의 내부 이름은 *mrI2dDsc*입니다.
- **Data Collection 로그 파일.** 표준 Data Collection 로그 파일에서 메타데이터를 읽어옵니다. 일반적으로 로그 파일의 파일 이름 확장자는 *.tmp*입니다. 하지만 일부 로그 파일의 파일 이름 확장자는 다를 수 있습니다. 필요할 경우 *.tmp* 파일 이름 확장자를 가지도록 파일의 이름을 바꿀 수 있습니다. 이 DSC의 내부 이름은 *mrLogDsc*입니다.
- **Quancept 정의 파일.** 메타데이터를 Quancept 스크립트로 변환합니다. Quancept *.qdi* 파일의 이름을 지정하십시오. 이 DSC의 내부 이름은 *mrQdiDrsDsc*입니다.
- **Quanvert 데이터베이스.** Quanvert 케이스 데이터 및 메타데이터를 읽습니다. *.qvinfo* 또는 *.pkd* 파일의 이름 및 위치를 지정하십시오. 이 DSC의 내부 이름은 *mrQvDsc*입니다.
- **Data Collection 참여 데이터베이스.** 프로젝트의 표본 및 히스토리 테이블 테이블을 읽고 해당 테이블의 열에 해당하는 파생된 범주형 변수를 작성합니다. 이 DSC의 내부 이름은 *mrSampleReportingMDSC*입니다.
- **Statistics 파일.** IBM SPSS Statistics *.sav* 파일에서 케이스 데이터 및 메타데이터를 읽어옵니다. IBM SPSS Statistics에서 분석을 위해 케이스 데이터를 IBM SPSS Statistics *.sav* 파일에 씁니다. IBM SPSS Statistics *.sav* 파일의 메타데이터를 *.mdd* 파일에 씁니다. 이 DSC의 내부 이름은 *mrSavDsc*입니다.
- **Surveycraft 파일.** SurveyCraft 케이스 데이터 및 메타데이터를 읽습니다. SurveyCraft *.vq* 파일의 이름을 지정하십시오. 이 DSC의 내부 이름은 *mrSCDsc*입니다.
- **Data Collection 스크립팅 파일.** mrScriptMetadata 파일의 메타데이터에서 읽어옵니다. 일반적으로 이 파일의 파일 이름 확장자는 *.mdd* 또는 *.dms*입니다. 이 DSC의 내부 이름은 *mrScriptMDSC*입니다.

- **Triple-S XML 파일.** XML 형식의 Triple-S 파일에서 메타데이터를 읽어옵니다. 이 DSC의 내부 이름은 *mrTripleSDsc*입니다.

메타데이터 특성. 선택적으로 **특성**을 선택하여 가져올 설문조사 버전과 사용할 언어, 컨텍스트 및 레이블 유형을 지정하십시오. 자세한 정보는 37 페이지의 『Data Collection 가져오기 메타데이터 특성』의 내용을 참조하십시오.

케이스 데이터 설정

참고: 사용 가능한 제공자 파일 유형의 전체 목록을 보려면 Data Collection 소프트웨어와 함께 사용 가능한 Survey Reporter Developer Kit을 설치해야 합니다.

케이스 데이터 설정 가져오기. *.mdd* 파일에서만 메타데이터를 읽어오는 경우에는 **케이스 데이터 설정 가져오기**를 클릭하여 지정된 소스에 액세스하기 위해 필요한 특정 설정과 함께 선택된 메타데이터와 연관되는 케이스 데이터 소스를 판별하십시오. 이 옵션은 *.mdd* 파일에만 사용할 수 있습니다.

케이스 데이터 제공자. 다음과 같은 제공자 유형이 지원됩니다.

- **ADO 데이터베이스.** Microsoft ADO 인터페이스를 사용하여 케이스 데이터를 읽습니다. 케이스 데이터 유형에 대해 OLE-DB UDL을 선택하고 케이스 데이터 UDL 필드에서 연결 문자열을 지정하십시오. 자세한 정보는 38 페이지의 『데이터베이스 연결 문자열』의 내용을 참조하십시오. 이 구성요소의 내부 이름은 *mrADODsc*입니다.
- **구분된 텍스트 파일(Excel).** 쉼표로 구분된(.CSV) 파일에서 케이스 데이터를 읽어와서 Excel로 출력할 수 있습니다. 내부 이름은 *mrCsvDsc*입니다.
- **Data Collection 데이터 파일.** 원시 Data Collection 데이터 형식 파일에서 케이스 데이터를 읽어옵니다. 내부 이름은 *mrDataFileDsc*입니다.
- **In2data 데이터베이스.** In2data 데이터베이스(*.i2d*) 파일에서 케이스 데이터 및 메타데이터를 읽어옵니다. 내부 이름은 *mrI2dDsc*입니다.
- **Data Collection 로그 파일.** 표준 Data Collection 로그 파일에서 케이스 데이터를 읽어옵니다. 일반적으로 로그 파일의 파일 이름 확장자는 *.tmp*입니다. 하지만 일부 로그 파일의 파일 이름 확장자는 다를 수 있습니다. 필요할 경우 *.tmp* 파일 이름 확장자를 가지도록 파일의 이름을 바꿀 수 있습니다. 내부 이름은 *mrLogDsc*입니다.
- **Quantum 데이터 파일.** Quantum 형식 ASCII 파일(*.dat*)에서 케이스 데이터를 읽어옵니다. 내부 이름은 *mrPunchDsc*입니다.
- **Quancept 데이터 파일.** Quancept *.drs*, *.drz* 또는 *.dru* 파일에서 케이스 데이터를 읽어옵니다. 내부 이름은 *mrQdiDrsDsc*입니다.
- **Quanvert 데이터베이스.**Quanvert *qvinfo* 또는 *.pkd* 파일에서 케이스 데이터를 읽어옵니다. 내부 이름은 *mrQvDsc*입니다.
- **Data Collection 데이터베이스(MS SQL Server).** 케이스 데이터를 관계형 Microsoft SQL Server 데이터베이스로 읽어옵니다. 자세한 정보는 38 페이지의 『데이터베이스 연결 문자열』의 내용을 참조하십시오. 내부 이름은 *mrRdbDsc2*입니다.

- **Statistics 파일.** IBM SPSS Statistics *.sav* 파일에서 케이스 데이터를 읽어옵니다. 내부 이름은 *mrSavDsc* 입니다.
- **Surveycraft 파일.** SurveyCraft *.qdt* 파일에서 케이스 데이터를 읽어옵니다. *.vq* 파일과 *.qdt* 파일은 모두 두 파일 모두에 대한 읽기 및 쓰기 액세스를 가진 동일한 디렉토리에 있어야 합니다. SurveyCraft를 사용할 때 기본적으로 이 방식으로 작성되지 않으므로 SurveyCraft 데이터를 가져오기 위해 파일 중 하나를 이동해야 합니다. 내부 이름은 *mrScDsc*입니다.
- **Triple-S 데이터 파일.** 고정 길이 또는 쉼표로 구분된 형식으로 Triple-S 데이터 파일에서 케이스 데이터를 읽어옵니다. 내부 이름은 *mr TripleDsc*입니다.
- **Data Collection XML.** Data Collection XML 데이터 파일에서 케이스 데이터를 읽어옵니다. 일반적으로 이 형식은 한 위치에서 다른 위치로 케이스 데이터를 전송하는 데 사용할 수 있습니다. 내부 이름은 *mrXmlDsc*입니다.

케이스 데이터 유형. 케이스 데이터를 파일, 폴더, OLE-DB UDL, ODBC DSN 중 어디에서 읽어오는지를 지정하고 대화 상자 옵션을 적절하게 업데이트합니다. 유효한 옵션은 제공자의 유형에 따라 다릅니다. 데이터베이스 제공자의 경우 OLE-DB 또는 ODBC 연결에 대한 옵션을 지정할 수 있습니다. 자세한 정보는 38 페이지의 『데이터베이스 연결 문자열』의 내용을 참조하십시오.

케이스 데이터 프로젝트. Data Collection 데이터베이스에서 케이스 데이터를 읽어오는 경우 프로젝트의 이름을 입력할 수 있습니다. 다른 모든 케이스 데이터 유형의 경우 이 설정은 공백이어야 합니다.

변수 가져오기

시스템 변수 가져오기. 인터뷰 상태(진행 중, 완료됨, 완료 날짜 등)를 표시하는 변수를 포함한 시스템 변수를 가져오는지 여부를 지정합니다. **없음**, **모두** 또는 **공통**을 선택할 수 있습니다.

"Codes" 변수 가져오기. 범주형 변수에 대한 오픈 엔드 "기타" 반응에 사용되는 코드를 나타내는 변수의 가져오기를 제어합니다.

"SourceFile" 변수 가져오기. 스캔된 반응의 이미지 파일 이름이 포함된 변수의 가져오기를 제어합니다.

다중 응답 변수 가져오기. 다중 응답 변수를 새 스트림에 대한 기본 방법인 다중 플래그 필드(다중 이분형 세트)로 가져올 수 있습니다. 12.0 이전의 IBM SPSS Modeler 릴리스에서 작성된 스트림은 값을 쉼표로 구분하여 다중 응답을 단일 필드로 가져왔습니다. 기존 스트림이 이전과 마찬가지로 실행될 수 있게 이전의 방법이 계속 지원되지만 새로운 방법을 사용하도록 이전 스트림을 업데이트하는 것이 좋습니다. 자세한 정보는 38 페이지의 『다중 응답 세트 가져오기』의 내용을 참조하십시오.

Data Collection 가져오기 메타데이터 특성

Data Collection 설문조사 데이터를 가져오는 경우 메타데이터 특성 대화 상자에서 가져올 설문조사 버전과 사용할 언어, 컨텍스트 및 레이블 유형을 지정할 수 있습니다. 한 번에 하나의 언어, 컨텍스트 및 레이블 유형만 가져올 수 있습니다.

버전. 각각의 설문조사 버전을 특정 케이스 데이터 세트를 수집하는 데 사용되는 메타데이터의 스냅샷으로 간주할 수 있습니다. 질문지가 변경됨에 따라 여러 버전이 작성될 수 있습니다. 최신 버전, 모든 버전 또는 특정 버전을 가져올 수 있습니다.

- **모든 버전.** 사용 가능한 모든 버전의 조합(수퍼 세트)을 사용하려면 이 옵션을 선택하십시오. (이를 수퍼 버전이라고도 함). 버전 간 충돌이 있는 경우에는 일반적으로 최신 버전이 이전 버전보다 우선합니다. 예를 들어, 버전에서 범주 레이블이 다른 경우에는 최신 버전의 텍스트가 사용됩니다.
- **최신 버전.** 최신 버전을 사용하려면 이 옵션을 사용하십시오.
- **버전 지정.** 특정 설문조사 버전을 사용하려면 이 옵션을 선택하십시오.

예를 들어, 둘 이상의 버전에 대한 케이스 데이터를 내보내려고 하는데 한 버전으로 수집된 케이스 데이터가 다른 버전에서 유효하지 않음을 의미하는 변수 및 범주 정의에 대한 변경사항이 작성된 경우에는 모든 버전을 선택하는 것이 유용합니다. 케이스 데이터를 내보낼 모든 버전을 선택하는 것은 일반적으로 버전 간 차이로 인한 유효성 오류 없이 동시에 다양한 버전으로 수집된 케이스 데이터를 내보낼 수 있음을 의미합니다. 하지만 버전 변경사항에 따라 일부 유효성 오류가 여전히 발생할 수 있습니다.

언어. 질문 및 연관된 텍스트는 여러 언어로 메타데이터에 저장할 수 있습니다. 설문조사의 기본 언어를 사용하거나 특정 언어를 지정할 수 있습니다. 지정된 언어로 항목을 사용할 수 없는 경우에는 기본값이 사용됩니다.

컨텍스트. 사용할 사용자 컨텍스트를 선택하십시오. 사용자 컨텍스트는 표시되는 텍스트를 제어합니다. 예를 들어, 질문 텍스트를 표시하려면 **질문**을 선택하고 데이터 분석 시 표시하기에 적합한 더 짧은 텍스트를 표시하려면 **분석**을 선택하십시오.

레이블 유형. 정의된 레이블의 유형을 나열합니다. 기본값은 **레이블**이며 이는 분석 사용자 컨텍스트의 변수 설명과 질문 사용자 컨텍스트의 질문 텍스트에 사용됩니다. 지시사항, 설명 등에 대해 다른 레이블 유형을 정의할 수 있습니다.

데이터베이스 연결 문자열

Data Collection 노드를 사용하여 OLE-DB 또는 ODBC를 통해 데이터베이스에서 케이스 데이터를 가져오는 경우 파일 탭에서 **편집**을 선택하여 연결 문자열 대화 상자에 액세스하십시오. 이 대화 상자에서는 연결을 미세 조정하기 위해 제공자에게 전달된 연결 문자열을 사용자 정의할 수 있습니다.

고급 특성

Data Collection 노드를 사용하여 명시적 로그인이 필요한 데이터베이스로부터 케이스 데이터를 가져오는 경우 **고급**을 선택하여 데이터 소스에 액세스하는 데 필요한 사용자 ID 및 비밀번호를 제공하십시오.

다중 응답 세트 가져오기

변수의 가능한 각각의 값에 대해 별도의 플래그 필드를 가진 다중 이분형 세트로 Data Collection에서 다중 응답 변수를 가져올 수 있습니다. 예를 들어, 응답자가 목록에서 방문한 박물관을 선택하도록 요청되는 경우 세트에는 나열된 각 박물관에 대한 별도의 플래그 필드가 포함됩니다.

데이터를 가져온 후 필터 탭을 포함하는 노드에서 다중 응답 세트를 추가하거나 편집할 수 있습니다. 자세한 정보는 156 페이지의 『다중 응답 세트 편집』의 내용을 참조하십시오.

단일 필드로 다중 응답 가져오기(이전 릴리스에서 작성된 스트림을 위해)

SPSS Modeler의 이전 릴리스에서는 위에 설명된 대로 다중 응답을 가져오는 대신 쉼표로 구분된 값을 사용하여 단일 필드로 다중 응답을 가져왔습니다. 이 방법은 기존 스트림을 지원하기 위해 여전히 지원되지만 새로운 방법을 사용하도록 해당 스트림을 업데이트하는 것이 좋습니다.

Data Collection 열 가져오기 참고

Data Collection 데이터의 열은 다음 표에 요약된 대로 SPSS Modeler로 읽어옵니다.

표 4. Data Collection 열 가져오기 요약

Data Collection 열 유형	**SPSS ModelerStorage**	측정 수준
부울 플래그(예/아니오)	문자열	플래그(값 0 및 1)
범주형	문자열	명목
날짜 또는 시간소인	시간소인	연속형
Double(지정된 범위 내의 부동 소수점 값)	실수	연속형
Long(지정된 범위 내의 정수 값)	정수	연속형
텍스트(자유 텍스트 설명)	문자열	유형 없음
수준(질문 내 눈금 또는 루프를 표시함)	VDATA에서는 발생하지 않으며 SPSS Modeler로 가져오지 않음	
오브젝트(낙서 텍스트 또는 음성 녹음을 표시하는 팩스 등의 이분형 데이터)	SPSS Modeler로 가져오지 않음	
없음(알 수 없는 유형)	SPSS Modeler로 가져오지 않음	
Respondent.Serial 열(고유 ID를 각 반응자와 연관시킴)	정수	유형 없음

실제 값과 메타데이터에서 읽어온 값 레이블 사이의 가능한 불일치를 방지하기 위해 모든 메타데이터 값이 소문자로 변환됩니다. 예를 들어, 값 레이블 *E1720_years*는 *e1720_years*로 변환됩니다.

IBM Cognos BI 소스 노드

IBM Cognos BI 소스 노드를 사용하면 Cognos BI 데이터베이스 데이터 또는 단일 목록 보고서를 데이터 마이닝 세션으로 가져올 수 있습니다. 이 방식으로 Cognos의 비즈니스 인텔리전스 기능을 IBM SPSS Modeler의 예측 분석 기능과 결합할 수 있습니다. 관계형, DMR(Dimensionally-Modeled Relational) 및 OLAP 데이터를 가져올 수 있습니다.

Cognos 서버 연결에서 먼저 데이터 또는 보고서를 가져올 위치를 선택하십시오. 한 위치에는 하나의 Cognos 모델과 해당 모델과 연관된 모든 폴더, 쿼리, 보고서, 보기, 단축키, URL 및 작업 정의가 포함되어 있습니다. Cognos 모델은 비즈니스 규칙, 데이터 설명, 데이터 관계, 비즈니스 차원 및 계층, 기타 관리 작업을 정의합니다.

데이터를 가져오는 경우에는 선택된 패키지에서 가져올 오브젝트를 선택하십시오. 가져올 수 있는 오브젝트에는 쿼리 제목(데이터베이스 테이블을 나타냄) 또는 개별 쿼리 항목(테이블 열을 나타냄)이 포함됩니다. 자세한 정보는 『Cognos 오브젝트 아이콘』의 내용을 참조하십시오.

패키지에 정의된 필터가 있으면 이 필터 중 하나 이상을 가져올 수 있습니다. 가져오는 필터가 가져온 데이터와 연관되는 경우에는 데이터를 가져오기 전에 필터가 적용됩니다. *참고*: 가져올 데이터는 UTF-8 형식이어야 합니다.

보고서를 가져오는 경우에는 하나 이상의 보고서가 포함된 패키지(또는 패키지 내의 폴더)를 선택하십시오. 그런 다음 가져올 개별 보고서를 선택하십시오. *참고*: 단일 목록 보고서만 가져올 수 있으며 다중 목록은 지원되지 않습니다.

데이터 오브젝트 또는 보고서에 대해 매개변수가 정의된 경우에는 오브젝트 또는 보고서를 가져오기 전에 이 매개변수에 대한 값을 지정할 수 있습니다.

Cognos 오브젝트 아이콘

Cognos BI 데이터베이스에서 가져올 수 있는 다양한 유형의 오브젝트는 다음 표에서와 같이 다양한 아이콘으로 표시됩니다.

표 5. Cognos 오브젝트 아이콘.

아이콘	오브젝트
	패키지
	네임스페이스
	쿼리 제목
	쿼리 항목
	측도 차원
	측도
	차원
	수준 계층
	수준
	필터
	보고서
	독립형 계산

Cognos 데이터 가져오기

IBM Cognos BI 데이터베이스에서 데이터를 가져오려면 IBM Cognos BI 대화 상자의 데이터 탭에서 **모드**가 **데이터**로 설정되어 있는지 확인하십시오.

연결. 편집 단추를 클릭하여 데이터 또는 보고서를 가져올 새 Cognos 연결의 세부사항을 정의할 수 있는 대화 상자를 표시하십시오. IBM SPSS Modeler를 통해 이미 Cognos 서버에 로그인한 경우에는 현재 연결의 세부사항도 편집할 수 있습니다. 자세한 정보는 42 페이지의 『Cognos 연결』의 내용을 참조하십시오.

위치. Cognos 서버 연결을 설정한 경우 이 필드 옆의 **편집** 단추를 클릭하여 컨텐츠를 가져올 사용 가능한 패키지의 목록을 표시하십시오. 자세한 정보는 43 페이지의 『Cognos 위치 선택』의 내용을 참조하십시오.

컨텐츠. 선택된 패키지와 연관된 네임스페이스와 함께 선택된 패키지의 이름을 표시합니다. 네임스페이스를 두 번 클릭하여 가져올 수 있는 오브젝트를 표시하십시오. 다양한 오브젝트 유형이 다양한 아이콘으로 표시됩니다. 자세한 정보는 40 페이지의 『Cognos 오브젝트 아이콘』의 내용을 참조하십시오.

가져올 오브젝트를 선택하려면 오브젝트를 선택한 후 두 개의 오른쪽 화살표의 상단을 클릭하여 오브젝트를 **가져올 필드** 분할창으로 이동하십시오. 쿼리 제목을 선택하면 해당 쿼리 항목을 모두 가져옵니다. 쿼리 제목을 두 번 클릭하면 해당 개별 쿼리 항목을 하나 이상 선택할 수 있도록 쿼리 제목이 펼쳐집니다. Ctrl+클릭(개별 항목 선택), Shift+클릭(항목 블록 선택) 및 Ctrl+A(모든 항목 선택)를 사용하여 다중 선택을 수행할 수 있습니다.

적용할 필터를 선택하려면(패키지에 필터가 정의된 경우) 컨텐츠 분할창의 필터로 이동하여 필터를 선택한 후 두 개의 오른쪽 화살표의 하단을 클릭하여 필터를 **적용할 필터** 분할창으로 이동하십시오. Ctrl+클릭(개별 필터 선택) 및 Shift+클릭(필터 블록 선택)을 사용하여 다중 선택을 수행할 수 있습니다.

가져올 필드. 처리를 위해 IBM SPSS Modeler로 가져오기 위해 선택한 데이터베이스 오브젝트를 나열합니다. 특정 오브젝트가 더 이상 필요하지 않은 경우에는 해당 오브젝트를 선택한 후 왼쪽 화살표를 클릭하여 해당 오브젝트를 **컨텐츠** 분할창으로 리턴하십시오. **컨텐츠**의 경우와 동일한 방식으로 다중 선택을 수행할 수 있습니다.

적용할 필터. 데이터를 가져오기 전에 데이터에 적용하기 위해 선택한 필터를 나열합니다. 특정 필터가 더 이상 필요하지 않은 경우에는 해당 필터를 선택한 후 왼쪽 화살표를 클릭하여 해당 필터를 **컨텐츠** 분할창으로 리턴하십시오. **컨텐츠**의 경우와 동일한 방식으로 다중 선택을 수행할 수 있습니다.

매개변수. 이 단추를 사용할 수 있으면 선택된 오브젝트에 매개변수가 정의되어 있습니다. 데이터를 가져오기 전에 매개변수를 사용하여 조정을 수행할 수 있습니다(예: 매개변수화된 계산 수행). 매개변수가 정의되어 있지만 기본값이 제공되지 않은 경우에는 이 단추에 경고 삼각형이 표시됩니다. 이 단추를 클릭하여 매개변수를 표시하고 선택적으로 편집하십시오. 이 단추가 사용 안함으로 설정되면 보고서에 매개변수가 정의되어 있지 않습니다.

가져오기 전에 데이터 통합. 원시 데이터 대신 통합 데이터를 가져오려면 이 선택란을 선택하십시오.

Cognos 보고서 가져오기

IBM Cognos BI 데이터베이스로부터 사전 정의된 보고서를 가져오려면 IBM Cognos BI 대화 상자의 데이터 탭에서 **모드**가 **보고서**로 설정되어 있는지 확인하십시오. 참고: 단일 목록 보고서만 가져올 수 있으며 다중 목록은 지원되지 않습니다.

연결. 편집 단추를 클릭하여 데이터 또는 보고서를 가져올 새 Cognos 연결의 세부사항을 정의할 수 있는 대화 상자를 표시하십시오. IBM SPSS Modeler를 통해 이미 Cognos 서버에 로그인한 경우에는 현재 연결의 세부사항도 편집할 수 있습니다. 자세한 정보는 『Cognos 연결』의 내용을 참조하십시오.

위치. Cognos 서버 연결을 설정한 경우 이 필드 옆의 **편집** 단추를 클릭하여 컨텐츠를 가져올 사용 가능한 패키지의 목록을 표시하십시오. 자세한 정보는 43 페이지의 『Cognos 위치 선택』의 내용을 참조하십시오.

내용. 보고서가 포함된 선택된 패키지 또는 폴더의 이름을 표시합니다. 특정 보고서로 이동하여 선택한 후 오른쪽 화살표를 클릭하여 보고서를 **가져올 보고서** 필드로 가져오십시오.

가져올 보고서. IBM SPSS Modeler로 가져오도록 선택한 보고서를 표시합니다. 보고서가 더 이상 필요하지 않은 경우에는 해당 보고서를 선택한 후 왼쪽 화살표를 클릭하여 **컨텐츠** 분할창에 리턴하거나 다른 보고서를 이 필드로 가져오십시오.

매개변수. 이 단추가 사용으로 설정되면 선택된 보고서에 매개변수가 정의되어 있습니다. 매개변수를 사용하여 보고서를 가져오기 전에 조정(예: 보고서 데이터의 시작 및 종료 날짜 지정)을 수행할 수 있습니다. 매개변수가 정의되어 있지만 기본값이 제공되지 않은 경우에는 이 단추에 경고 삼각형이 표시됩니다. 이 단추를 클릭하여 매개변수를 표시하고 선택적으로 편집하십시오. 이 단추가 사용 안함으로 설정되면 보고서에 매개변수가 정의되어 있지 않습니다.

Cognos 연결

Cognos Connections 대화 상자에서는 데이터베이스 오브젝트를 가져오거나 내보낼 Cognos BI 서버를 선택할 수 있습니다.

Cognos 서버 URL 가져오거나 내보낼 Cognos BI 서버의 URL을 입력하십시오. 이는 Cognos BI 서버에서 IBM Cognos Configuration의 "외부 디스패처 URI" 환경 특성의 값입니다. 사용할 URL이 확실하지 않으면 Cognos 시스템 관리자에게 문의하십시오.

모드 특정 Cognos 네임스페이스, 사용자 이름 및 비밀번호로 로그인하려는 경우(예: 관리자로서) **신임 정보 설정**을 선택하십시오. 사용자 신임 정보 없이 로그인하려면 **익명 연결 사용**을 선택하십시오. 이 경우에는 기타 필드를 채우지 않습니다.

또는 IBM SPSS Collaboration and Deployment Services 리포지토리에 IBM Cognos 신임 정보가 저장되어 있는 경우, 사용자 이름 및 비밀번호 정보를 입력하거나 익명 연결을 작성하는 대신 이 신임 정보를 사용할 수 있습니다. 기존 신임 정보를 사용하려면 **저장된 신임 정보**를 선택하고 **신임 정보 이름**을 입력하거나 찾아보십시오.

Cognos 네임스페이스는 IBM SPSS Collaboration and Deployment Services의 도메인을 통해 모델링됩니다.

네임스페이스 ID 서버에 로그인하는 데 사용되는 Cognos 보안 인증 제공자를 지정하십시오. 인증 제공자는 사용자, 그룹 및 역할을 정의 및 유지보수하고 인증 프로세스를 제어하는 데 사용됩니다. 이는 네임스페이스 이름이 아니라 네임스페이스 ID입니다(ID는 이름과 항상 동일하지는 않음).

사용자 이름 서버에 로그온하는 데 사용하는 Cognos 사용자 이름을 입력하십시오.

비밀번호 지정된 사용자 이름과 연관된 비밀번호를 입력하십시오.

기본값으로 저장 노드를 열 때마다 설정을 다시 입력하지 않아도 되도록 이 설정을 기본값으로 저장하려면 이 단추를 클릭하십시오.

Cognos 위치 선택

위치 지정 대화 상자에서는 데이터를 가져올 Cognos 패키지 또는 보고서를 가져올 패키지 또는 폴더를 선택할 수 있습니다.

공용 폴더. 데이터를 가져오는 경우 이는 선택된 서버에서 사용 가능한 패키지 및 폴더를 나열합니다. 사용할 패키지를 선택한 후 **확인**을 클릭하십시오. Cognos BI 소스 노드당 하나의 패키지만 선택할 수 있습니다.

보고서를 가져오는 경우 이는 선택된 서버에서 사용 가능한 보고서가 포함된 폴더 및 패키지를 나열합니다. 패키지 또는 보고서 폴더를 선택한 후 **확인**을 클릭하십시오. Cognos BI 소스 노드당 하나의 패키지 또는 보고서 폴더만 선택할 수 있지만 보고서 폴더는 다른 보고서 폴더 및 개별 보고서를 포함할 수 있습니다.

데이터 또는 보고서에 대한 매개변수 지정

Cognos BI에서 데이터 오브젝트 또는 보고서에 대한 매개변수가 정의된 경우에는 해당 오브젝트 또는 보고서를 가져오기 전에 이 매개변수에 대한 값을 지정할 수 있습니다. 보고서에 대한 매개변수의 예로는 보고서 컨텐츠의 시작 및 종료 날짜가 있습니다.

이름. Cognos BI 데이터베이스에서 지정되는 매개변수 이름입니다.

유형. 매개변수에 대한 설명입니다.

값. 매개변수에 할당할 값입니다. 값을 입력하거나 편집하려면 테이블에서 해당 셀을 두 번 클릭하십시오. 값은 여기서 검증되지 않으므로 유효하지 않은 값은 런타임 시 발견됩니다.

테이블에서 유효하지 않은 매개변수 자동으로 제거. 이 옵션은 기본적으로 선택되며 데이터 오브젝트 또는 보고서에서 발견된 유효하지 않은 매개변수를 모두 제거합니다.

IBM Cognos TM1 소스 노드

IBM Cognos TM1 소스 노드에서는 Cognos TM1 데이터를 데이터 마이닝 세션으로 가져올 수 있습니다. 이러한 방식으로 Cognos의 엔터프라이즈 계획 기능과 IBM SPSS Modeler의 예측 분석 기능을 결합할 수 있습니다. 다차원 OLAP 큐브 데이터의 플랫 버전을 가져올 수 있습니다.

참고: TM1 사용자에게는 큐브 쓰기 권한, 차원 읽기 권한 및 차원 요소 쓰기 권한이 필요합니다.

데이터를 가져오기 전에 TM1의 데이터를 수정해야 합니다.

참고: 가져올 데이터는 UTF-8 형식이어야 합니다.

IBM Cognos TM1 관리 호스트 연결에서 먼저 데이터를 가져올 TM1 서버를 선택합니다. 서버에는 하나 이상의 TM1 큐브가 있습니다. 그런 다음 필요한 큐브를 선택하고 해당 큐브 내에서 가져올 열 및 행을 선택합니다.

참고: SPSS Modeler에서 TM1 소스 또는 내보내기 노드를 사용하려면 먼저 `tm1s.cfg` 파일에서 일부 설정을 유효화해야 합니다. 이 파일은 TM1 서버의 루트 디렉토리에 있는 TM1 서버 구성 파일입니다.

- HTTPPortNumber - 유효한 포트 번호를 설정합니다. 일반적으로 1 - 65535입니다.
- UseSSL - 참으로 설정하면 HTTPS가 전송 프로토콜로 사용됩니다. 이 경우 TM1 인증을 SPSS Modeler Server JRE로 가져와야 합니다.

IBM Cognos TM1 데이터 가져오기

IBM Cognos TM1 데이터베이스에서 데이터를 가져오려면, IBM Cognos TM1 대화 상자의 데이터 탭에서 관련 TM1 관리 호스트, 연관된 서버, 큐브, 데이터 세부사항을 선택하십시오.

참고: 데이터를 가져오기 전, TM1 내에서 몇 가지 사전 처리를 수행하여 데이터가 IBM SPSS Modeler에서 인식 가능한 형식이 되도록 해야 합니다. 여기에는 보기가 가져오기에 올바른 *크기* 및 형태가 되도록 서브세트 편집기를 사용하여 데이터를 필터링하는 작업이 포함됩니다.

TM1에서 가져오는 영(0) 값은 "널" 값으로 취급됩니다(TM1은 공백과 0 값을 구별하지 않음). 또한 *정규 차원*의 비수치 데이터(또는 메타데이터)를 IBM SPSS Modeler로 가져올 수 있습니다. 그러나 비수치 측도 가져오기는 현재 지원되지 않습니다.

관리 호스트 연결할 TM1 서버가 설치된 관리 호스트의 URL을 입력하십시오. 관리 호스트는 모든 TM1 서버에 대한 단일 URL로 정의됩니다. 이 URL에서, 사용하는 환경에 설치되어 실행 중인 모든 IBM Cognos TM1 서버를 검색하고 액세스할 수 있습니다.

TM1 서버 관리 호스트에 연결한 경우, 가져올 데이터가 있는 서버를 선택하고 **로그인**을 클릭하십시오. 이전에 이 서버에 연결한 적이 없는 경우, **사용자 이름** 및 **비밀번호** 입력을 요구하는 프롬프트가 표시됩니다. 또는 이전에 입력하여 **저장된 신임 정보**로 저장한 로그인 세부사항을 검색할 수 있습니다.

가져올 TM1 큐브 보기 선택 데이터를 가져올 수 있는 TM1 서버 내의 큐브 이름을 표시합니다. 큐브를 두 번 클릭하면 가져올 수 있는 보기 데이터가 표시됩니다.

참고: 차원이 있는 큐브만 IBM SPSS Modeler로 가져올 수 있습니다.

가져올 데이터를 선택하려면, 보기를 선택하고 오른쪽 화살표를 클릭하여 보기를 **가져올 보기** 분할창으로 이동시키십시오. 필요한 보기가 표시되지 않으면 큐브를 두 번 클릭하여 보기 목록을 펼치십시오.

행 차원. 가져오기로 선택한 데이터의 행 차원 이름을 나열합니다. 수준 목록에서 화면 이동하여 필요한 수준을 선택하십시오.

열 차원 가져오기로 선택한 데이터의 열 차원 이름을 나열합니다. 수준 목록에서 화면 이동하여 필요한 수준을 선택하십시오.

컨텍스트 차원. 선택된 열 및 행과 관련된 컨텍스트 차원을 표시합니다.

SAS 소스 노드

참고: 이 기능은 SPSS Modeler Professional 및 SPSS Modeler Premium에서 사용 가능합니다.

SAS 소스 노드를 사용하면 SAS 데이터를 데이터 마이닝 세션으로 가져올 수 있습니다. 다음과 같은 네 가지 유형의 파일을 가져올 수 있습니다.

- SAS for Windows/OS2(*.sd2*)
- SAS for UNIX(*.ssd*)
- SAS 전송 파일(*.tpt*)
- SAS 버전 7/8/9(*.sas7bdat*)

데이터를 가져올 때 모든 변수가 유지되고 변수 유형이 변경되지 않습니다. 모든 케이스가 선택됩니다.

SAS 소스 노드에 대한 옵션 설정

가져오기. 전송할 SAS 파일의 유형을 선택합니다. **SAS for Windows/OS2(.sd2)**, **SAS for UNIX(.SSD)**, **SAS Transport File(.tpt)** 또는 **SAS 버전 7/8/9(.sas7bdat)**를 선택할 수 있습니다.

파일 가져오기. 파일의 이름을 지정합니다. 파일 이름을 입력하거나 생략 기호 단추(**...**)를 클릭하여 파일 위치를 찾아보십시오.

멤버. 위에서 선택한 SAS 전송 파일에서 가져올 멤버를 선택합니다. 멤버 이름을 입력하거나 **선택**을 클릭하여 파일 내의 모든 멤버를 찾아볼 수 있습니다.

SAS 데이터 파일에서 사용자 형식 읽기. 사용자 형식을 읽으려면 선택하십시오. SAS 파일은 데이터 및 데이터 형식(변수 레이블 등)을 다른 파일에 저장합니다. 대부분의 경우 형식도 함께 가져오려고 할 것입니다. 그러나 큰 데이터 세트가 있는 경우에는 이 옵션을 선택 취소하여 메모리를 절약할 수 있습니다.

형식 파일. 형식 파일이 필요하면 이 텍스트 상자가 활성화됩니다. 파일 이름을 입력하거나 생략 기호 단추(**...**)를 클릭하여 파일 위치를 찾아보십시오.

변수 이름. SAS 파일에서 가져올 때 변수 이름 및 레이블을 처리하는 방법을 선택합니다. 여기에 포함시키려고 선택하는 메타데이터는 IBM SPSS Modeler 내의 작업 전체에서 지속되며 SAS에서 사용하도록 다시 내보낼 수 있습니다.

- **이름 및 레이블 읽기.** 변수 이름 및 레이블을 둘 다 IBM SPSS Modeler에 읽어들일 수 있습니다. 기본적으로 이 옵션이 선택되고 변수 이름이 유형 노드에 표시됩니다. 레이블은 스트림 특성 대화 상자에서 지정된 옵션에 따라 표현식 작성기, 도표, 모델 브라우저 및 기타 유형의 출력에 표시될 수 있습니다.
- **이름으로 레이블 읽기.** SAS 파일에서 짧은 필드 이름이 아니라 설명 변수 레이블을 읽고 해당 레이블을 IBM SPSS Modeler에서 변수 이름으로 사용하려면 선택하십시오.

Excel 소스 노드

Excel 소스 노드를 사용하면 .xlsx 파일 형식으로 Microsoft Excel에서 데이터를 가져올 수 있습니다.

파일 유형. 가져오는 Excel 파일 유형을 선택하십시오.

파일 가져오기. 가져올 스프레드시트 파일의 이름 및 위치를 지정합니다.

이름 지정된 범위 사용. Excel 워크시트에서 정의된 대로 셀의 이름 지정된 범위를 지정할 수 있게 합니다. 생략 기호 단추(**...**)를 클릭하여 사용 가능한 범위 목록에서 선택하십시오. 이름 지정된 범위가 사용되는 경우에는 기타 워크시트 및 데이터 범위 설정은 더 이상 적용할 수 없으며 결과적으로 사용 안함으로 설정됩니다.

워크시트 선택. 가져올 워크시트를 지정합니다(인덱스별 또는 이름별).

- **인덱스별.** 가져올 워크시트에 대한 인덱스 값을 지정하십시오(첫 번째 워크시트에 대해 0으로 시작하여 두 번째 워크시트에 대해 1을 지정하는 등의 방식임).
- **이름별.** 가져올 워크시트의 이름을 지정하십시오. 생략 기호 단추(**...**)를 클릭하여 사용 가능한 워크시트 목록에서 선택하십시오.

워크시트의 범위. 첫 번째 비공백 행 또는 명시적 셀 범위로 시작하여 데이터를 가져올 수 있습니다.

- **첫 번째 비공백 행에서 범위 시작.** 첫 번째 비공백 셀을 찾아서 데이터 범위의 왼쪽 상단으로 사용합니다.
- **셀의 명시적 범위.** 행 및 열을 기준으로 명시적 범위를 지정할 수 있게 합니다. 예를 들어, Excel 범위 `A1:D5`를 지정하려면 첫 번째 필드에 `A1`을 입력하고 두 번째 필드에 `D5`를 입력하십시오(또는 `R1C1`과 `R5C4`). 공백 행을 포함하여 지정된 범위의 모든 행이 리턴됩니다.

공백 행에서. 둘 이상의 공백 행이 발생하는 경우에는 **읽기를 중지**할지 아니면 **공백행을 리턴**하여 공백 행을 포함한 모든 데이터를 워크시트의 끝까지 계속 읽을지 선택할 수 있습니다.

첫 번째 행에 열 이름이 있음. 지정된 범위의 첫 번째 행을 필드(열) 이름으로 사용해야 함을 나타냅니다. 선택되지 않은 경우에는 필드 이름이 자동으로 생성됩니다.

필드 저장 공간 및 측정 수준

Excel에서 값을 읽어오는 경우에는 기본적으로 *연속형* 측정 수준으로 숫자 저장 공간을 가진 필드를 읽어오며 문자열 필드는 *명목*으로 읽어옵니다. 유형 탭에서 수동으로 측정 수준(연속형 대 명목)을 변경할 수 있지만 저

장 공간은 자동으로 결정됩니다(필요한 경우에는 채움 노드 또는 파생 노드에서 to_integer 등의 변환 함수를 사용하여 변경할 수 있음). 자세한 정보는 9 페이지의 『필드 저장 공간 및 형식화 설정』의 내용을 참조하십시오.

기본적으로 숫자 값과 문자열 값의 혼합을 가진 필드는 숫자로 읽어오며 이는 모든 문자열 값이 IBM SPSS Modeler에서 널(시스템 결측) 값으로 설정됨을 의미합니다. 이는 Excel과 달리 IBM SPSS Modeler는 필드 내에서 혼합 저장 유형을 허용하지 않기 때문에 발생합니다. 이를 방지하려면 Excel 스프레드시트에서 셀 형식을 텍스트로 수동으로 설정하십시오. 그러면 모든 값(숫자 포함)을 문자열로 읽어옵니다.

XML 소스 노드

참고: 이 기능은 SPSS Modeler Professional 및 SPSS Modeler Premium에서 사용 가능합니다.

XML 소스 노드를 사용하여 XML 형식 파일의 데이터를 IBM SPSS Modeler 스트림으로 가져오십시오. XML은 데이터 교환을 위한 표준 언어이며 이러한 목적을 위해 다수의 조직에서 선택하는 형식입니다. 예를 들어, 국세청에서 온라인으로 제출된 XML 형식의 세금 환급 데이터를 분석하려 할 수 있습니다(http://www.w3.org/standards/xml/ 참조).

XML 데이터를 IBM SPSS Modeler 스트림으로 가져오면 소스에 대해 광범위한 예측 분석 함수를 수행할 수 있습니다. XML 데이터는 테이블 형식으로 구문 분석됩니다. 테이블 형식에서 열은 XML 요소 및 속성의 다양한 중첩 수준에 해당합니다. XML 항목은 XPath 형식으로 표시됩니다(http://www.w3.org/TR/xpath20/ 참조).

단일 파일 읽기 기본적으로, SPSS Modeler는 **XML 데이터 소스** 필드에 지정하는 하나의 파일을 읽습니다.

디렉토리의 모든 XML 파일 읽기 특정 디렉토리에 있는 모든 XML 파일을 읽으려면 이 옵션을 선택하십시오. 표시되는 **디렉토리** 필드에 위치를 지정하십시오. 지정된 디렉토리의 모든 서브디렉토리에 있는 파일을 추가로 읽으려면 **서브디렉토리 포함** 선택란을 선택하십시오.

XML 데이터 소스 가져올 XML 소스 파일의 전체 경로 및 파일 이름을 입력하거나 찾아보기 단추를 사용하여 파일을 찾으십시오.

XML 스키마(선택사항) XML 구조를 읽을 XSD 또는 DTD 파일의 전체 경로 및 파일 이름을 지정하거나 찾아보기 단추를 사용하여 이 파일을 찾으십시오. 이 필드를 비워 두면 XML 소스 파일에서 구조를 읽습니다. XSD 또는 DTD 파일은 둘 이상의 루트 요소를 포함할 수 있습니다. 이 경우, 초점을 다른 필드로 바꾸면 사용할 루트 요소를 선택하는 대화 상자가 표시됩니다. 자세한 정보는 48 페이지의 『여러 루트 요소에서 선택』의 내용을 참조하십시오.

참고: SPSS Modeler는 XSD 표시기를 무시합니다.

XML 구조 XML 소스 파일의 구조를 표시하는 계층 구조 트리(또는 **XML 스키마** 필드에 스키마를 지정한 경우 스키마)입니다. 레코드 경계를 정의하려면 요소를 선택하고 오른쪽 화살표 단추를 클릭하여 해당 항목을 **레코드** 필드로 복사하십시오.

속성 표시 XML 구조 필드에 있는 XML 요소의 속성을 표시하거나 숨깁니다.

레코드(XPath 표현식) XML 구조 필드에서 복사된 요소의 XPath 구문을 표시합니다. 그러면 이 요소가 XML 구조에서 강조표시되며 레코드 경계를 정의합니다. 소스 파일에서 이 요소를 만날 때마다 새 레코드가 작성됩니다. 이 필드가 비어 있으면 루트 아래의 첫 번째 하위 요소가 레코드 경계로 사용됩니다.

모든 데이터 읽기 기본적으로 소스 파일의 모든 데이터를 스트림으로 읽습니다.

읽을 데이터 지정 개별 요소나 속성 또는 둘 다를 가져오려면 이 옵션을 선택하십시오. 이 옵션을 선택하면 가져올 데이터를 지정할 수 있는 필드 테이블이 사용 가능하게 됩니다.

필드 이 테이블은 **읽을 데이터 지정** 옵션을 선택한 경우 가져오기를 위해 선택한 요소 및 속성을 나열합니다. 요소 또는 속성의 XPath 구문을 XPath 열에 직접 입력하거나 XML 구조에서 요소 또는 속성을 선택하고 오른쪽 화살표 단추를 클릭하여 해당 항목을 테이블로 복사할 수 있습니다. 요소의 모든 하위 요소 및 속성을 복사하려면 XML 구조에서 해당 요소를 선택하고 이중 화살표 단추를 클릭하십시오.

- **XPath** 가져올 항목의 XPath 구문입니다.
- **위치** 가져올 항목의 XML 구조에서의 위치입니다. **고정 경로**는 XML 구조에서 강조표시된 요소(또는 강조표시된 요소가 없는 경우 루트 아래의 첫 번째 하위 요소)에 상대적인 항목의 경로를 표시합니다. **임의 위치**는 XML 구조의 임의 위치에 지정된 이름의 항목이 있음을 표시합니다. **사용자 정의**는 XPath 열에 직접 위치를 입력하는 경우에 표시됩니다.

여러 루트 요소에서 선택

적절한 양식의 XML 파일은 하나의 루트 요소만 포함할 수 있는 반면, XSD 또는 DTD 파일은 여러 루트를 포함할 수 있습니다. 루트 중 하나가 XML 소스 파일의 루트와 일치하는 경우 해당 루트 요소가 사용되고, 그렇지 않은 경우에는 사용할 루트 요소를 선택해야 합니다.

표시할 루트 선택. 사용할 루트 요소를 선택하십시오. 기본값은 XSD 또는 DTD 구조의 첫 번째 루트 요소입니다.

XML 소스 데이터에서 원하지 않는 공백 제거

XML 소스 데이터의 행 바꾸기는 [CR][LF] 문자 조합으로 구현될 수 있습니다. 일부 경우에 행 바꾸기는 텍스트 문자열 가운데 올 수 있습니다.

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

파일이 일부 애플리케이션(예: 웹 브라우저)에서 열릴 때는 이러한 행 바꾸기가 표시되지 않을 수 있습니다. 그러나 XML 소스 노드를 통해 스트림으로 데이터를 읽으면 행 바꾸기가 일련의 공백 문자로 변환됩니다.

채움 노드를 사용하여 원하지 않는 공백을 제거함으로써 이러한 문제를 정정할 수 있습니다.

다음은 이러한 목적을 달성하기 위한 방법의 예입니다.

1. XML 소스 노드에 채움 노드를 연결하십시오.

2. 채움 노드를 열고 필드 선택기를 사용하여 원하지 않는 공백이 있는 필드를 선택하십시오.
3. **대체**를 **조건 기반**으로 설정하고 **조건**을 **true**로 설정하십시오.
4. **대체 대상** 필드에 `replace(" ","",@FIELD)`를 입력하고 확인을 클릭하십시오.
5. 테이블 노드를 채움 노드에 연결하고 스트림을 실행하십시오.

이제 테이블 노드 출력에서 텍스트가 추가 공백 없이 표시됩니다.

사용자 입력 노드

사용자 입력 노드는 처음부터 또는 기존 데이터를 변경하여 합성 데이터를 작성하는 쉬운 방법을 제공합니다. 이것은 예를 들어 모델링을 위한 검정 데이터 세트를 작성할 때 유용합니다.

처음부터 데이터 작성

사용자 입력 노드는 소스 팔레트에서 사용할 수 있으며 스트림 캔버스에 바로 추가할 수 있습니다.

1. 노드 팔레트의 **소스** 탭을 클릭하십시오.
2. 끌어서 놓기 또는 두 번 클릭으로 사용자 입력 노드를 스트림 캔버스에 추가하십시오.
3. 두 번 클릭하여 해당 대화 상자를 열고 필드 및 값을 지정하십시오.

참고: 소스 팔레트에서 선택한 사용자 입력 노드는 완전히 비어 있고 필드 및 데이터 정보가 없습니다. 이로 인해 완전히 처음부터 합성 데이터를 작성할 수 있습니다.

기존 데이터 소스에서 데이터 생성

스트림의 비터미널 노드에서 사용자 입력 노드를 생성할 수도 있습니다.

1. 스트림에서 노드를 대체할 위치를 결정하십시오.
2. 해당 데이터를 사용자 입력 노드에 공급할 노드를 마우스 오른쪽 단추로 클릭하고 메뉴에서 **사용자 입력 노드 생성**을 선택하십시오.
3. 모든 다운스트림 프로세스가 연결된 사용자 입력 노드가 표시되어 데이터 스트림의 해당 위치에 있는 기존 노드를 대체합니다. 노드는 생성될 때 메타데이터에서 모든 데이터 구조 및 필드 유형 정보(사용 가능한 경우)를 상속합니다.

참고: 데이터가 스트림의 모든 노드를 거치지 않았다면 해당 노드들은 완전히 인스턴스화되지 않았으며, 이는 사용자 입력 노드로 대체할 때 저장 공간 및 데이터 값이 사용 불가능할 수도 있음을 의미합니다.

사용자 입력 노드의 옵션 설정

사용자 입력 노드에 대한 대화 상자에는 값을 입력하고 합성 데이터의 데이터 구조를 정의하는 데 사용할 수 있는 여러 도구가 있습니다. 생성된 노드의 경우, 데이터 탭의 테이블에 원래 데이터 소스의 필드 이름이 있습니다. 소스 팔레트에서 추가된 노드의 경우에는 테이블이 비어 있습니다. 테이블 옵션을 사용하여 다음 태스크를 수행할 수 있습니다.

- 테이블의 오른쪽에 있는 새 필드 추가 단추를 사용하여 새 필드를 추가합니다.

- 기존 필드의 이름을 바꿉니다.
- 각 필드의 데이터 저장 공간을 지정합니다.
- 값을 지정합니다.
- 표시 화면에서 필드 순서를 변경합니다.

데이터 입력

필드마다 테이블 오른쪽에 있는 값 선택도구 단추를 사용하여 원래 데이터 세트의 값을 삽입하거나 값을 지정할 수 있습니다. 값 지정에 대한 자세한 정보는 아래에 설명된 규칙을 참조하십시오. 필드를 비워 둘 수도 있습니다. 빈 필드는 시스템 널($null$)로 채워집니다.

문자열 값을 지정하려면 값 열에 문자열 값을 공백으로 분리하여 입력하면 됩니다.

```
Fred Ethel Martin
```

공백을 포함하는 문자열은 큰따옴표로 묶을 수 있습니다.

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

숫자 필드의 경우, 여러 값을 동일한 방식으로(사이에 공백을 사용하여 나열) 입력할 수 있습니다.

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

또는 동일한 숫자 열을 해당 한계(10, 20)와 숫자 사이의 단계(2)를 설정하여 지정할 수 있습니다. 이 방법을 사용하는 경우 다음과 같이 입력합니다.

```
10,20,2
```

이 두 방법은 다음과 같이 하나를 다른 하나에 임베드하여 결합할 수 있습니다.

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

이와 같이 입력하는 경우 다음 값이 생성됩니다.

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

스트림 특성 대화 상자에서 선택된 현재 기본 형식을 사용하여 날짜 및 시간 값을 입력할 수 있습니다. 예를 들어, 다음과 같습니다.

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

날짜 및 시간 구성요소를 모두 포함하는 시간소인 값의 경우 큰따옴표를 사용해야 합니다.

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

추가 세부사항은 아래의 데이터 저장 공간에 대한 주석을 참조하십시오.

데이터 생성. 스트림을 실행할 때 레코드가 생성되는 방법을 지정할 수 있습니다.

- **모든 조합**. 가능한 모든 필드 값 조합을 포함하는 레코드를 생성하므로 각 필드 값이 여러 레코드에 표시됩니다. 가끔 필요한 양보다 많은 데이터를 생성할 수 있으므로 종종 이 노드 뒤에 표본 노드를 사용할 수도 있습니다.
- **순서대로**. 데이터 필드 값이 지정된 순서대로 레코드를 생성합니다. 각 필드 값이 하나의 레코드에만 표시됩니다. 총 레코드 수는 단일 필드의 값 최대 수와 같습니다. 필드에 값 최대 수보다 적은 수의 값이 있는 경우 정의되지 않은($null$) 값이 삽입됩니다.

예 표시

예를 들어, 다음과 같이 입력하면 다음 두 개의 표 예에 나열된 레코드를 생성합니다.

- **연령.** 30,60,10
- **BP.** 낮음
- **콜레스테롤.** 정상 높음
- **약제.** (비워 둠)

표 6. 데이터 생성 필드를 모든 조합으로 설정.

연령	BP	콜레스테롤	약제
30	낮음	정상	$null$
30	낮음	높음	$null$
40	낮음	정상	$null$
40	낮음	높음	$null$
50	낮음	정상	$null$
50	낮음	높음	$null$
60	낮음	정상	$null$
60	낮음	높음	$null$

표 7. 데이터 생성 필드를 순서대로로 설정.

연령	BP	콜레스테롤	약제
30	낮음	정상	$null$
40	$null$	높음	$null$
50	$null$	$null$	$null$
60	$null$	$null$	$null$

데이터 저장 공간

저장 공간은 데이터를 필드에 저장하는 방식을 설명합니다. 예를 들어, 값이 1 및 0인 필드는 정수 데이터를 저장합니다. 이는 데이터 사용에 대해 설명하고 저장 공간에 영향을 미치지 않는 측정 수준과 구별됩니다. 예를 들어, 값이 1 및 0인 정수 필드에 대한 측정 수준을 *플래그*로 설정할 수 있습니다. 이는 일반적으로 1 = *True* 및 0 = *False*을 표시합니다. 저장 공간은 소스에서 결정해야 하지만 측정 수준은 스트림의 어느 지점에서나 유형 노드를 사용하여 변경할 수 있습니다. 자세한 정보는 140 페이지의 『측정 수준』의 내용을 참조하십시오.

사용 가능한 저장 유형은 다음과 같습니다.

- **문자열** 영숫자 데이터라고도 하는 숫자가 아닌 데이터가 포함된 필드에 사용됩니다. 문자열은 *fred*, *Class 2* 또는 *1234* 등의 문자열 시퀀스를 포함할 수 있습니다. 문자열의 숫자는 계산에서 사용할 수 없습니다.
- **정수** 값이 정수인 필드입니다.
- **실수** 값은 10진수를 포함할 수 있는 숫자입니다(정수로 제한되지 않음). 표시 형식은 스트림 특성 대화 상자에서 지정되며 유형 노드(형식 탭)에서 개별 필드에 대해 대체될 수 있습니다.
- **날짜** 연도, 월 및 일 등의 표준 형식으로 지정된 날짜 값입니다(예: 2007-09-26). 구체적인 형식은 스트림 특성 대화 상자에서 지정됩니다.
- **시간** 기간으로 측정된 시간입니다. 예를 들어, 1시간 26분 38초 동안 지속되는 서비스 호출은 스트림 특성 대화 상자에서 지정된 대로 현재 시간 형식에 따라 01:26:38로 표시될 수 있습니다.
- **시간소인** 스트림 특성 대화 상자의 현재 날짜 및 시간 형식에 따라 날짜와 시간 구성요소를 모두 포함하는 값입니다(예: 2007-09-26 09:04:00). 시간소인 값을 별도의 날짜 및 시간 값 대신 단일 값으로 해석하기 위해 시간소인 값을 큰따옴표로 묶어야 할 수 있습니다. (이는 예를 들어, 사용자 입력 노드에서 값을 입력하는 경우에 적용됩니다.)
- **목록** 지리 공간 및 콜렉션이라는 새로운 측정 수준과 함께 SPSS Modeler 버전 17에서 도입된 목록 저장 공간 필드에는 단일 레코드에 대한 다중 값이 포함되어 있습니다. 모든 기타 저장 유형의 목록 버전이 있습니다.

표 8. 목록 저장 유형 아이콘

아이콘	저장 유형
[A]	문자열 목록
[◇]	정수 목록
[◈]	실수 목록
[◷]	시간 목록
[▦]	날짜 목록
[▣]	시간소인 목록
[[]]	0(영)보다 큰 깊이를 가진 목록

또한 콜렉션 측정 수준과 함께 사용하기 위해 다음과 같은 측정 수준의 목록 버전이 있습니다.

표 9. 목록 측정 수준 아이콘

아이콘	측정 수준
[✎]	연속형 목록

표 9. 목록 측정 수준 아이콘 (계속)

아이콘	측정 수준
[ ]	범주형 목록
[ ]	플래그 목록
[ ]	명목 목록
[ ]	순서 목록

목록은 세 가지 소스 노드(Analytic Server, 지리 공간 또는 가변파일) 중 하나에서 SPSS Modeler로 가져오거나 파생 또는 채움 필드 작업 노드를 사용하여 스트림 내에서 작성할 수 있습니다.

목록 및 해당 콜렉션 및 지리 공간 측정 수준과의 상호작용에 대한 자세한 정보는 11 페이지의 『목록 저장 공간 및 연관된 측정 수준』의 내용을 참조하십시오.

저장 공간 변환. 채움 노드에서 `to_string` 및 `to_integer` 등의 다양한 변환 함수를 사용하여 필드에 대한 저장 공간을 변환할 수 있습니다. 자세한 정보는 165 페이지의 『채움 노드를 사용한 저장 공간 변환』의 내용을 참조하십시오. 변환 함수(및 날짜 또는 시간 값 등의 특정 입력 유형이 필요한 기타 함수)는 스트림 특성 대화 상자에서 지정된 현재 형식에 따라 다릅니다. 예를 들어, 값이 *Jan 2003*, *Feb 2003* 등인 문자열 필드를 날짜 저장 공간으로 변환하려면 **MON YYYY**를 스트림의 기본 날짜 형식으로 선택하십시오. 파생 계산 중 임시 변환의 경우 파생 노드에서도 변환 함수를 사용할 수 있습니다. 파생 노드를 사용하여 범주형 값을 가진 문자열 필드 코딩 변경 등의 기타 조작을 수행할 수도 있습니다. 자세한 정보는 164 페이지의 『파생 노드를 사용하여 값 코딩변경』의 내용을 참조하십시오.

혼합 데이터 읽기. 숫자 저장 공간(정수, 실수, 시간, 시간소인 또는 날짜)을 가진 필드에서 읽을 때 숫자가 아닌 값은 널값 또는 시스템 결측값으로 설정됩니다. 이는 일부 애플리케이션과는 달리 IBM SPSS Modeler가 필드 내에서 혼합 저장 유형을 허용하지 않기 때문입니다. 이를 방지하기 위해 혼합 데이터가 포함된 필드는 필요에 따라 소스 노드 또는 외부 애플리케이션에서 저장 유형을 변경하여 문자열로 읽어야 합니다.

참고: 생성된 사용자 입력 노드는 인스턴스화된 경우 소스 노드에서 얻은 저장 공간 정보를 이미 포함할 수도 있습니다. 인스턴스화되지 않은 노드에는 저장 공간 또는 사용 유형 정보가 포함되지 않습니다.

값 지정 규칙

기호 필드의 경우 다음과 같이 값과 값 사이에 공백을 두어야 합니다.

```
HIGH MEDIUM LOW
```

숫자 필드의 경우, 여러 값을 동일한 방식으로(사이에 공백을 사용하여 나열) 입력할 수 있습니다.

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

또는 동일한 숫자 열을 해당 한계(10, 20)와 숫자 사이의 단계(2)를 설정하여 지정할 수 있습니다. 이 방법을 사용하는 경우 다음과 같이 입력합니다.

```
10,20,2
```

이 두 방법은 다음과 같이 하나를 다른 하나에 임베드하여 결합할 수 있습니다.

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

이와 같이 입력하는 경우 다음 값이 생성됩니다.

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

시뮬레이션 생성 노드

시뮬레이션 생성 노드는 히스토리 데이터 없이 사용자가 지정한 통계 분포를 사용하거나 기존 히스토리 데이터에 대해 시뮬레이션 적합 노드를 실행하여 얻은 분포를 자동으로 사용하여 시뮬레이션된 데이터를 생성하는 쉬운 방법을 제공합니다. 시뮬레이션된 데이터를 생성하는 것은 모델 입력에 불확실성이 존재하는 상황에서 예측 모델의 결과를 평가하기 원할 때 유용합니다.

히스토리 데이터 없이 데이터 작성

소스 팔레트에서 시뮬레이션 생성 노드를 사용할 수 있으며 스트림 캔버스에 직접 추가할 수 있습니다.

1. 노드 팔레트의 **소스** 탭을 클릭하십시오.
2. 시뮬레이션 생성 노드를 스트림 캔버스에 추가하려면 끌어서 놓거나 두 번 클릭하십시오.
3. 두 번 클릭하여 대화 상자를 열고 필드, 저장 유형, 통계 분포 및 분포모수를 지정하십시오.

참고: 소스 팔레트에서 선택된 시뮬레이션 생성 노드는 필드 및 분포 정보 없이 완전히 비어 있습니다. 따라서 사용자가 히스토리 데이터 없이 시뮬레이션 데이터를 작성할 수 있습니다.

기존 히스토리 데이터를 사용하여 시뮬레이션된 데이터 생성

시뮬레이션 생성 노드는 시뮬레이션 적합 터미널 노드를 실행하는 방법으로도 작성할 수 있습니다.

1. 시뮬레이션 적합 노드를 마우스 오른쪽 단추로 클릭하고 메뉴에서 **실행**을 선택하십시오.
2. 시뮬레이션 생성 노드는 시뮬레이션 적합 노드에 대한 업데이트 링크가 있는 스트림 캔버스에 표시됩니다.
3. 시뮬레이션 생성 노드는 생성될 때 모든 필드, 저장 유형 및 통계 분포 정보를 시뮬레이션 적합 노드에서 상속합니다.

시뮬레이션 적합 노드에 대한 업데이트 링크 정의

시뮬레이션 생성 노드 및 시뮬레이션 적합 노드 사이에 링크를 작성할 수 있습니다. 링크 작성은 히스토리 데이터에 대한 맞춤에 의해 판별된 가장 적합한 분포 정보가 있는 하나 이상의 필드를 업데이트할 때 유용합니다.

1. 시뮬레이션 생성 노드를 마우스 오른쪽 단추로 클릭하십시오.

2. 메뉴에서 **업데이트 링크 정의**를 선택하십시오. 커서가 링크 커서로 변경됩니다.
3. 또 다른 노드를 클릭하십시오. 이 노드가 시뮬레이션 적합 노드이면 링크가 작성된 것입니다. 이 노드가 시뮬레이션 적합 노드가 아니면 링크가 작성되지 않은 것이며 커서가 다시 일반 커서로 변경됩니다.

시뮬레이션 적합 노드의 필드가 시뮬레이션 생성 노드의 필드와 다른 경우, 차이가 있음을 알리는 메시지가 표시됩니다.

링크된 시뮬레이션 생성 노드를 업데이트하기 위해 시뮬레이션 적합 노드가 사용된 경우, 결과는 동일한 필드가 두 노드에 모두 존재하는지, 필드가 시뮬레이션 생성 노드에서 잠겨있지 않은지 여부에 따라 다릅니다. 시뮬레이션 적합 노드의 업데이트 결과는 다음 표에 표시됩니다.

표 10. 시뮬레이션 적합 노드의 업데이트 결과

	시뮬레이션의 필드 적합 노드	
시뮬레이션 생성 노드의 필드	**존재**	**결측값**
존재하며 잠겨있지 않음	필드를 덮어씁니다.	필드가 삭제됩니다.
결측	필드가 추가됩니다.	변경되지 않습니다.
존재하며 잠겨있음	필드의 분포를 덮어쓰지 않습니다. 세부사항 적합 대화 상자의 정보 및 상관관계가 업데이트됩니다.	필드를 덮어쓰지 않습니다. 상관계수는 0으로 설정됩니다.
재적합 시 최소/최대 지우지 않음 확인 상자가 선택됩니다.	최소, 최대 열의 값과 별개로 필드를 덮어씁니다.	
재적합 시 상관계수를 재계산하지 않음 확인 상자가 선택됩니다.	필드가 잠겨있지 않으면 덮어씁니다. 상관계수는 덮어쓰지 않습니다.	

시뮬레이션 적합 노드에 대한 업데이트 링크 제거

다음 단계를 수행하여 시뮬레이션 생성 노드 및 시뮬레이션 적합 노드 사이에 링크를 제거할 수 있습니다.

1. 시뮬레이션 생성 노드를 마우스 오른쪽 단추로 클릭하십시오.
2. 메뉴에서 **업데이트 링크 제거**를 선택하십시오. 링크가 제거됩니다.

시뮬레이션 생성 노드에 대한 옵션 설정

시뮬레이션 생성 노드 대화 상자의 데이터 탭에 대한 옵션을 사용하여 다음을 수행할 수 있습니다.

- 필드에 대한 통계 분포 정보를 보고 지정하고 편집할 수 있습니다.
- 필드 사이의 상관관계를 보고 지정하고 편집할 수 있습니다.
- 시뮬레이션할 반복계산 및 케이스 수를 지정합니다.

항목 선택. 시뮬레이션 생성 노드의 세 가지 뷰(시뮬레이션한 필드, 상관관계 및 고급 옵션) 사이에서 전환할 수 있습니다.

시뮬레이션한 필드 보기

시뮬레이션 생성 노드가 히스토리 데이터가 있는 시뮬레이션 적합 노드에서 생성되거나 업데이트된 경우, 시뮬레이션한 필드 보기에서 각 필드에 대한 통계 분포 정보를 보고 편집할 수 있습니다. 각 필드에 대한 다음 정보가 시뮬레이션 적합 노드로부터 시뮬레이션 생성 노드의 **유형** 탭에 복사됩니다.

- 측정 수준
- 값
- 결측값
- 검사
- 역할

히스토리 데이터가 없으면 저장 유형 및 분포 유형을 선택하고 필수 모수를 입력하여 필드를 정의하고 분포를 지정할 수 있습니다. 이런 방법으로 데이터를 생성하면 **유형** 탭 또는 유형 노드 등에서 데이터가 인스턴스화될 때까지 각 필드의 측정 수준에 대한 정보를 사용할 수 없습니다.

시뮬레이션한 필드 보기는 다음 태스크를 수행하는 데 사용할 수 있는 여러 가지 도구를 포함합니다.

- 필드 추가 및 제거
- 표시에서 필드의 순서 변경
- 각 필드에 대한 저장 유형 지정
- 각 필드에 대한 통계 분포 지정
- 각 필드의 통계 분포에 대한 모수 값 지정

시뮬레이션한 필드. 소스 팔레트에서 스트림 캔버스로 시뮬레이션 생성 노드가 추가된 경우, 이 표에 한 개의 빈 행이 추가됩니다. 이 행을 편집하면 비어 있는 새 행이 표의 아래쪽에 추가됩니다. 시뮬레이션 생성 노드가 시뮬레이션 적합 노드로부터 작성된 경우, 이 표에 히스토리 데이터의 각 필드에 대해 한 행이 포함됩니다. **새 필드 추가** 아이콘을 클릭하여 표에 추가 행을 추가할 수 있습니다.

시뮬레이션한 필드 표는 다음 열로 구성됩니다.

- **필드.** 필드의 이름을 포함합니다. 필드 이름은 셀에 입력하여 편집할 수 있습니다.
- **저장 공간.** 이 열의 셀은 저장 유형의 드롭 다운 목록을 포함합니다. 사용가능한 저장 유형은 **문자열**, **정수**, **실수**, **시간**, **날짜** 및 **시간소인**입니다. 저장 유형을 선택하면 분포 열에서 사용 가능한 분포가 결정됩니다. 시뮬레이션 생성 노드가 시뮬레이션 적합 노드로부터 작성된 경우, 시뮬레이션 적합 노드로부터 저장 유형이 복사됩니다.

  참고: 날짜/시간 저장 유형이 있는 필드의 경우, 분포모수를 정수로 지정해야 합니다. 예를 들어, 1970년 1월 1일의 평균 날짜를 지정하려면 정수 0을 사용하십시오. 부호 있는 정수는 1970년 1월 1일 자정 이후 또는 이전의 초 수를 나타냅니다.
- **상태.** 상태 열의 아이콘은 각 필드에 대한 적합 상태를 표시합니다.

필드에 대해 분포가 지정되지 않았거나 하나 이상의 분포모수가 결측값입니다. 시뮬레이션을 실행하려면 이 필드에 대한 분포를 지정하고 모수에 대한 유효한 값을 입력해야 합니다.

필드가 가장 가까운 적합 분포로 설정됩니다.
참고: 이 아이콘은 시뮬레이션 생성 노드가 시뮬레이션 적합 노드로부터 생성된 경우에만 표시될 수 있습니다.

가장 가까운 적합 분포가 세부사항 적합 하위 대화 상자의 대체 분포로 대체되었습니다. 자세한 정보는 61 페이지의 『세부사항 적합』의 내용을 참조하십시오.

분포가 수동으로 지정되거나 편집되었으며 두 개 이상의 수준에서 지정된 모수를 포함할 수 있습니다.

- **잠겨있음.** 잠금 아이콘이 있는 열에서 선택란을 선택하여 시뮬레이션한 필드를 잠그면 링크된 시뮬레이션 적합 노드에 의한 자동 업데이트에서 필드가 제외됩니다. 이 방법은 수동으로 분포를 지정하고 링크된 시뮬레이션 적합 노드가 실행될 때 자동 분포 적합에 의해 영향을 받지 않도록 할 때 가장 유용합니다.
- **분포.** 이 열의 셀은 통계 분포의 드롭 다운 목록을 포함합니다. 저장 유형을 선택하면 지정된 필드에 대한 해당 열에서 사용 가능한 분포가 결정됩니다. 자세한 정보는 65 페이지의 『분포』의 내용을 참조하십시오.

  참고: 모든 필드에 대해 고정 분포를 지정할 수는 없습니다. 생성된 데이터의 모든 필드가 고정되도록 하려면 뒤에 균형 노드가 있는 사용자 입력 노드를 사용하십시오.
- **매개변수.** 맞춰진 각 분포와 연관된 분포모수가 이 열에 표시됩니다. 모수의 다중 값은 콤마로 구분됩니다. 모수에 대한 다중 값을 지정하면 시뮬레이션에 대한 다중 반복이 생성됩니다. 자세한 정보는 64 페이지의 『반복』의 내용을 참조하십시오. 모수가 결측값이면 상태 열에 표시되는 아이콘에서 반영됩니다. 모수에 대한 값을 지정하려면 관심 있는 필드에 대한 행에서 이 열을 클릭하고 목록에서 **지정**을 선택하십시오. 그러면 매개변수 지정 하위 대화 상자가 열립니다. 자세한 정보는 62 페이지의 『매개변수 지정』의 내용을 참조하십시오. 분포 열에서 경험적 분포가 선택되면 이 열을 사용하지 않습니다.
- **최소, 최대.** 일부 분포의 경우, 이 열에서 시뮬레이션한 데이터에 대한 최소값, 최대값 또는 둘 다를 지정할 수 있습니다. 최소값보다 작고 최대값보다 큰 시뮬레이션한 데이터는 지정된 분포에 유효한 경우라도 거부됩니다. 최소값 및 최대값을 지정하려면 관심 있는 필드에 해당되는 행에서 이 열을 클릭하고 목록에서 **지정**을 선택하십시오. 그러면 매개변수 지정 하위 대화 상자가 열립니다. 자세한 정보는 62 페이지의 『매개변수 지정』의 내용을 참조하십시오. 분포 열에서 경험적 분포가 선택되면 이 열을 사용하지 않습니다.

가장 가까운 적합 사용. 시뮬레이션 생성 노드가 히스토리 데이터가 있는 시뮬레이션 적합 노드에서 자동으로 생성되고 시뮬레이션한 필드 표에서 단일 행이 선택된 경우에만 사용 가능합니다. 선택된 행의 필드에 대한 정보를 필드에 대한 가장 가까운 적합 분포의 정보로 바꾸십시오. 선택된 행의 정보를 편집한 경우, 이 단추를 누르면 시뮬레이션 적합 노드에서 판별된 가장 가까운 적합 분포로 정보가 다시 설정됩니다.

세부사항 적합. 시뮬레이션 생성 노드가 시뮬레이션 적합 노드에서 자동으로 생성된 경우에만 사용 가능합니다. 그러면 세부사항 적합 하위 대화 상자가 열립니다. 자세한 정보는 61 페이지의 『세부사항 적합』의 내용을 참조하십시오.

시뮬레이션한 필드 보기의 오른쪽에 있는 아이콘을 사용하여 여러 가지 유용한 태스크를 수행할 수 있습니다. 이러한 아이콘에 대해서는 다음 표에서 설명합니다.

표 11. 시뮬레이션한 필드 보기의 아이콘.

아이콘	도구팁	설명
	분포 모수 편집	시뮬레이션한 필드 표에서 단일 행이 선택된 경우에만 사용 가능합니다. 선택된 행에 대한 모수 지정 하위 대화 상자가 열립니다. 자세한 정보는 62 페이지의 『매개변수 지정』의 내용을 참조하십시오.
	새 필드 추가	시뮬레이션한 필드 표에서 단일 행이 선택된 경우에만 사용 가능합니다. 시뮬레이션한 필드 표의 아래쪽까지 비어 있는 새 행을 추가합니다.
	다중 사본 생성	시뮬레이션한 필드 표에서 단일 행이 선택된 경우에만 사용 가능합니다. 그러면 필드 복제 하위 대화 상자가 열립니다. 자세한 정보는 61 페이지의 『복제 필드』의 내용을 참조하십시오.
	선택된 필드 삭제	시뮬레이션한 필드 표에서 선택된 행이 삭제됩니다.
	맨 위로 이동	선택된 행이 시뮬레이션한 필드 표의 맨 위 행이 아닌 위치에 있는 경우에만 사용 가능합니다. 선택된 행을 시뮬레이션한 필드 표의 맨 위쪽으로 이동합니다. 이 조치는 시뮬레이션한 데이터의 필드 순서에 영향을 미칩니다.
	위로 이동	선택된 행이 시뮬레이션한 필드 표의 맨 위 행이 아닌 위치에 있는 경우에만 사용 가능합니다. 선택된 행을 시뮬레이션한 필드 표의 한 위치 위쪽으로 이동합니다. 이 조치는 시뮬레이션한 데이터의 필드 순서에 영향을 미칩니다.
	아래로 이동	선택된 행이 시뮬레이션한 필드 표의 맨 아래 행이 아닌 위치에 있는 경우에만 사용 가능합니다. 선택된 행을 시뮬레이션한 필드 표의 한 위치 아래쪽으로 이동합니다. 이 조치는 시뮬레이션한 데이터의 필드 순서에 영향을 미칩니다.

표 11. 시뮬레이션한 필드 보기의 아이콘 (계속).

아이콘	도구팁	설명
	맨 아래로 이동	선택된 행이 시뮬레이션한 필드 표의 맨 아래 행이 아닌 위치에 있는 경우에만 사용 가능합니다. 선택된 행을 시뮬레이션한 필드 표의 맨 아래쪽으로 이동합니다. 이 조치는 시뮬레이션한 데이터의 필드 순서에 영향을 미칩니다.

재적합 시 최소 및 최대 지우지 않음. 이 옵션을 선택하면 분포가 연결된 시뮬레이션 적합 노드를 실행하여 업데이트될 때 최소값 및 최대값을 덮어쓰지 않습니다.

상관관계 보기

예측 모형에 대한 입력 필드는 종종 상관관계가 있는 것으로 알려져 있습니다. 예를 들어, 높이 및 가중치가 있습니다. 시뮬레이션한 값이 이러한 상관관계를 유지하기 위해서는 시뮬레이션되는 필드 사이의 상관관계를 고려해야 합니다.

시뮬레이션 생성 노드가 히스토리 데이터가 있는 시뮬레이션 적합 노드에서 생성되거나 업데이트된 경우, 상관관계 보기에서 필드 쌍 사이의 계산된 상관관계를 보고 편집할 수 있습니다. 히스토리 데이터가 없으면 필드가 상관된 방법에 대한 지식을 기반으로 하여 수동으로 상관관계를 지정할 수 있습니다.

참고: 데이터가 생성되기 전에 상관행렬이 자동으로 positive semi-definite인지 여부를 확인하므로 도치될 수 있습니다. 행렬이 선형 독립변수이면 도치될 수 있습니다. 상관행렬을 도치할 수 없으면 자동으로 도치될 수 있도록 조정됩니다.

행렬 또는 목록 형식으로 상관관계를 표시할 수 있습니다.

상관행렬. 행렬에서 필드 쌍 사이의 상관관계를 표시합니다. 필드 이름이 행렬의 맨 위에서 아래 왼쪽으로 문자순으로 나열됩니다. 대각선 아래의 셀만 편집할 수 있습니다. -1.000 이상이며 1.000 이하인 값을 입력해야 합니다. 초점이 대각선 아래의 미러링된 셀에서 변경될 때 대각선 위 셀이 업데이트되고 두 셀 모두 동일한 값을 표시합니다. 대각선 셀은 항상 사용하지 않으며 항상 상관관계가 1.000입니다. 모든 기타 셀에 대한 기본값은 0.000입니다. 0.000 값은 연관된 필드 쌍 사이에 상관관계가 없음을 표시합니다. 연속형 및 순서 필드만 행렬에 포함됩니다. 고정 분포에 지정된 명목형, 범주형 및 플래그 필드는 표에 표시되지 않습니다.

상관관계 목록. 표에서 필드 쌍 사이의 상관관계를 표시합니다. 표의 각 행은 필드 쌍 사이의 상관관계를 표시합니다. 행은 추가되거나 삭제될 수 없습니다. 머리말 필드 1 및 필드 2가 있는 열은 편집할 수 없는 필드 이름을 포함합니다. 상관관계 열은 편집할 수 있는 상관관계를 포함하며 -1.000 이상이며 1.000 이하인 값을 입력해야 합니다. 모든 셀에 대한 기본값은 0.000입니다. 연속형 및 순서 필드만 목록에 포함됩니다. 고정 분포에 지정된 명목형, 범주형 및 플래그 필드는 목록에 표시되지 않습니다.

상관관계 재설정. 상관관계 재설정 대화 상자가 열립니다. 히스토리 데이터를 사용할 수 있으면 다음 세 가지 옵션 중 하나를 선택할 수 있습니다.

- **맞춰짐.** 현재 상관관계를 히스토리 데이터를 사용하여 계산한 상관관계로 바꿉니다.

- **0값.** 현재 상관관계를 0으로 바꿉니다.
- **취소.** 대화 상자를 닫습니다. 상관관계가 변경되지 않습니다.

히스토리 데이터를 사용할 수 없으나 상관관계를 변경한 경우에는 현재 상관관계를 0값으로 대체하거나 취소할 수 있습니다.

표시 형식. 상관관계를 행렬로 표시하려면 **표**를 선택하십시오. 상관관계를 목록으로 표시하려면 **목록**을 선택하십시오.

재적합 시 상관관계를 재계산하지 않음. 수동으로 상관관계를 지정하고 시뮬레이션 적합 노드 및 히스토리 데이터를 사용하여 자동으로 분포를 맞출 때 덮어쓰지 않도록 하려면 이 옵션을 선택하십시오.

범주형 분포가 있는 입력에 대해 맞춰진 다방향 분할표 사용. 기본적으로 범주형 분포가 있는 모든 필드는 분할표(또는 범주형 분포가 있는 필드의 수에 따라 다방향 분할표)에 포함됩니다. 분할표는 상관관계와 같이 시뮬레이션 적합 노드가 실행될 때 구성됩니다. 분할표는 볼 수 없습니다. 이 옵션을 선택하면 범주형 분포가 있는 필드가 분할표의 실제 퍼센트를 사용하여 시뮬레이션됩니다. 즉, 명목 필드 사이의 모든 연관이 시뮬레이션한 새 데이터에서 다시 작성됩니다. 이 옵션을 선택 취소하면 범주형 분포가 있는 필드가 분할표의 예측 퍼센트를 사용하여 시뮬레이션됩니다. 필드를 수정하면 분할표에서 필드가 제거됩니다.

고급 옵션 보기

시뮬레이션할 케이스 수. 시뮬레이션할 케이스 수 및 반복의 이름을 설정하는 방법을 지정할 수 있는 옵션이 표시됩니다.

- **최대 케이스 수.** 생성할 시뮬레이션한 데이터 및 연관된 목표 값의 최대수를 지정합니다. 기본값은 100,00, 최소값은 1000, 최대값은 2,147,483,647입니다.
- **반복.** 이 수는 자동으로 계산되며 편집할 수 없습니다. 반복은 분포 모수에 다중 값이 지정될 때마다 자동으로 계산됩니다.
- **전체 행.** 반복계산 수가 1보다 큰 경우에만 사용 가능합니다. 수는 표시된 방정식을 사용하여 자동으로 계산되며 편집할 수 없습니다.
- **반복 필드 생성.** 반복계산 수가 1보다 큰 경우에만 사용 가능합니다. 선택하면 **이름** 필드를 사용할 수 있습니다. 자세한 정보는 64 페이지의 『반복』의 내용을 참조하십시오.
- **이름. 반복 필드 생성** 확인 상자가 선택되었으며 반복계산 수가 1보다 큰 경우에만 사용 가능합니다. 이 텍스트 필드를 입력하여 반복계산 수를 편집하십시오. 자세한 정보는 64 페이지의 『반복』의 내용을 참조하십시오.

난수 시드. 난수 시드를 설정하면 시뮬레이션을 복제할 수 있습니다.

- **결과 복제.** 선택하면 **생성** 단추 및 **난수 시드** 필드를 사용할 수 있습니다.
- **난수 시드. 복제 결과** 확인 상자를 선택한 경우에만 사용 가능합니다. 이 필드에서 난수 시드로 사용할 정수를 지정할 수 있습니다. 기본값은 629111597입니다.
- **생성. 복제 결과** 확인 상자를 선택한 경우에만 사용 가능합니다. **난수 시드** 필드에 1 이상 999999999 이하의 의사 난수 정수를 생성합니다.

복제 필드

복제 필드 대화 상자에서는 선택한 필드의 사본 작성 수 및 각 사본의 이름을 결정하는 방식을 지정할 수 있습니다. 복합 효과를 조사할 때 필드의 다중 사본이 있는 경우에 유용합니다. 예를 들어, 수많은 연속된 시간 주기에 걸친 이율 또는 성장 비율 등입니다.

대화 상자의 제목 표시줄에는 선택된 필드의 이름이 포함됩니다.

작성할 사본 수. 작성할 필드의 사본 수를 포함합니다. 작성할 사본 수를 선택하려면 화살표를 클릭하십시오. 최소 사본 수는 1이며 최대값은 512입니다. 사본 수는 처음에 10으로 설정됩니다.

사본 접미문자. 각 사본에 대한 필드 이름의 끝에 추가되는 문자를 포함합니다. 이러한 문자는 필드 이름과 사본 수를 구분합니다. 이 필드를 입력하여 접미문자를 편집할 수 있습니다. 이 필드는 비어 있는 상태로 둘 수 있습니다. 이 경우, 필드 이름과 사본 수 사이에 문자가 없습니다. 기본 문자는 밑줄입니다.

초기 사본 번호. 첫 번째 사본에 대한 접미문자를 포함합니다. 초기 사본 수를 선택하려면 화살표를 클릭하십시오. 최소 초기 사본 수는 1이며 최대값은 1000입니다. 기본 초기 사본 수는 1입니다.

사본 수 단계. 접미문자 수에 대한 증분을 포함합니다. 증분을 선택하려면 화살표를 클릭하십시오. 최소 증분은 1이며 최대값은 255입니다. 증분은 초기에 1로 설정됩니다.

필드. 사본에 대한 필드 이름 미리보기를 포함합니다. 이는 복제 필드 대화 상자의 필드 중 임의의 필드가 편집될 때 업데이트됩니다. 이 텍스트는 자동으로 생성되고 편집할 수 없습니다.

확인. 대화 상자에서 지정된 대로 모든 사본을 생성합니다. 시뮬레이션 생성 노드 대화 상자의 시뮬레이션한 필드 표에서 복사된 필드를 포함하는 행 바로 아래에 사본이 추가됩니다.

취소. 대화 상자를 닫습니다. 작성된 변경사항을 삭제합니다.

세부사항 적합

세부사항 적합 대화 상자는 시뮬레이션 적합 노드를 실행하여 시뮬레이션 생성 노드가 작성되었거나 업데이트된 경우에만 사용 가능합니다. 이 대화 상자는 선택된 필드에 대한 자동 분포 적합의 결과를 표시합니다. 분포는 적합도에 의해 순서가 지정되며 가장 가까운 적합 분포가 첫 번째로 나열됩니다. 이 대화 상자에서는 다음 태스크를 수행할 수 있습니다.

- 히스토리 데이터에 맞춰진 분포 검사
- 맞춰진 분포 중 한 개 선택

필드. 선택된 필드의 이름을 포함합니다. 이 텍스트는 편집할 수 없습니다.

다음으로 처리(측도). 선택된 필드의 측정 유형을 표시합니다. 시뮬레이션 생성 노드 대화 상자의 시뮬레이션한 필드 표에서 가져옵니다. 측정 유형은 화살표를 클릭하여 변경할 수 있으며 드롭 다운 목록에서 측정 유형을 선택할 수 있습니다. **연속형**, **명목형** 및 **순서**라는 세 가지 옵션이 있습니다.

분포. 분포 표는 측정 유형에 적합한 모든 분포를 표시합니다. 히스토리 데이터에 맞춰진 분포는 가장 적합한 것부터 가장 적합하지 않은 것까지 적합도에 의해 순서가 지정됩니다. 적합도는 시뮬레이션 적합 노드에서 선택된 적합 통계량에 의해 판별됩니다. 히스토리 데이터에 맞춰지지 않은 분포는 표에서 맞춰진 분포 아래에 문자순으로 나열됩니다.

분포 테이블에는 다음 열이 포함됩니다.

- **사용.** 선택된 단일 선택 단추는 현재 필드에 대해 선택된 분포를 표시합니다. 사용 열에서 원하는 분포에 대한 단일 선택 단추를 선택하여 가장 가까운 적합 분포를 대체할 수 있습니다. 또한 사용 열에서 단일 선택 단추를 선택하면 선택한 필드에 대한 히스토리 데이터의 히스토그램 또는 막대형 차트에 겹쳐진 분포의 도표가 표시됩니다. 한 번에 하나의 분포만 선택할 수 있습니다.
- **분포.** 분포의 이름을 포함합니다. 이 열은 편집할 수 없습니다.
- **통계량 적합.** 분포에 대해 계산되는 적합 통계량을 포함합니다. 이 열은 편집할 수 없습니다. 셀의 내용은 필드의 측정 유형에 따라 결정됩니다.
  - **연속형.** Anderson-Darling 및 Kolmogorov-Smirnoff 검정의 결과를 포함합니다. 검정과 연관된 p-값도 표시됩니다. 시뮬레이션 적합 노드에서 적합도 기준으로 선택된 적합 통계량이 첫 번째로 표시되고 분포 순서 지정에 사용됩니다. Anderson-Darling 통계량은 A=*aval* P=*pval*로 표시됩니다. Kolmogorov-Smirnoff 통계량은 K=*kval* P=*pval*로 표시됩니다. 통계를 계산할 수 없으면 숫자 대신 점이 표시됩니다.
  - **명목형 및 순서.** 카이제곱 검정의 결과를 포함합니다. 검정과 연관된 p-값도 표시됩니다. 통계량은 카이제곱=*val* P=*pval*로 표시됩니다. 분포가 맞춰지지 않으면 맞춰지지 않음이 표시됩니다. 분포가 수학적으로 맞춰지지 않으면 적합할 수 없음이 표시됩니다.

  참고: 경험적 분포에 대한 셀은 항상 비어 있습니다.
- **매개변수.** 맞춰진 각 분포와 연관된 분포모수를 포함합니다. 모수는 *parameter_name = parameter_value*로 표시되고 단일 공백으로 구분됩니다. 범주형 분포의 경우, 모수 이름은 범주형이며 모수값은 확률과 연관됩니다. 분포가 히스토리 데이터에 맞춰지지 않으면 셀이 비어 있습니다. 이 열은 편집할 수 없습니다.

히스토그램 썸네일. 선택한 필드에 대한 히스토리 데이터의 히스토그램에 겹쳐진 선택된 분포의 도표가 표시됩니다.

분포 썸네일. 선택된 분포에 대한 설명을 표시합니다.

확인. 대화 상자를 닫고 선택된 필드에 대한 시뮬레이션한 필드 표의 측정, 분포, 모수 및 최소, 최대 열의 값을 선택한 분포의 정보로 업데이트합니다. 상태 열의 아이콘도 선택된 분포가 데이터에 가장 가까운 적합이 있는 분포인지 여부를 반영하도록 업데이트됩니다.

취소. 대화 상자를 닫습니다. 작성된 변경사항을 삭제합니다.

매개변수 지정

매개변수 지정 대화 상자에서 선택한 필드의 분포에 대한 모수값을 수동으로 지정할 수 있습니다. 또한 선택한 필드에 대해 다른 분포를 선택할 수도 있습니다.

매개변수 지정 대화 상자는 세 가지 방법으로 열 수 있습니다.

- 시뮬레이션 생성 노드 대화 상자의 시뮬레이션한 필드 표에서 필드 이름을 두 번 클릭하십시오.
- 시뮬레이션한 필드 표의 모수 또는 최소, 최대 열을 클릭하고 목록에서 **지정**을 선택하십시오.
- 시뮬레이션한 필드 표에서 행을 선택한 다음 **분포모수 편집** 아이콘을 클릭하십시오.

필드. 선택된 필드의 이름을 포함합니다. 이 텍스트는 편집할 수 없습니다.

분포. 선택된 필드의 분포를 포함합니다. 이는 시뮬레이션한 필드 표에서 가져옵니다. 분포는 화살표를 클릭하여 변경할 수 있으며 드롭 다운 목록에서 분포를 선택할 수 있습니다. 사용가능한 분포는 선택된 필드의 저장 유형에 따라 다릅니다.

방향. 이 옵션은 **분포** 필드에서 Dice 분포를 선택한 경우에만 사용 가능합니다. 필드를 분할할 방향 또는 범주의 수를 선택하려면 화살표를 클릭하십시오. 최소 방향 수는 2이며 최대수는 20입니다. 방향 수는 처음에 6으로 설정됩니다.

분포모수. 분포모수 표에는 선택한 분포의 각 모수에 대해 한 행이 포함됩니다.

참고: 분포가 형상 모수 α = k이며 역 척도 모수 β = 1/θ인 비율 모수를 사용합니다.
이 테이블에는 두 개의 열이 있습니다.

- **모수.** 모수의 이름을 포함합니다. 이 열은 편집할 수 없습니다.
- **값.** 모수의 값을 포함합니다. 시뮬레이션 생성 노드가 시뮬레이션 적합 노드로부터 작성되거나 업데이트된 경우, 분포를 히스토리 데이터에 맞춰서 결정된 모수값이 이 열의 셀에 포함됩니다. 소스 노드 팔레트에서 스트림 캔버스로 시뮬레이션 생성 노드가 추가된 경우, 이 열의 셀이 비어 있습니다. 값은 셀에 입력하여 편집할 수 있습니다. 각 분포에 필요한 모수 및 허용 가능한 모수값에 대한 자세한 정보는 65 페이지의 『분포』의 내용을 참조하십시오.

  모수에 대한 다중 값은 콤마로 구분되어야 합니다. 모수에 대한 다중 값을 지정하면 시뮬레이션의 다중 반복이 정의됩니다. 한 모수에 대해서만 다중 값을 지정할 수 있습니다.

  참고: 날짜/시간 저장 유형이 있는 필드의 경우, 분포모수를 정수로 지정해야 합니다. 예를 들어, 1970년 1월 1일의 평균 날짜를 지정하려면 정수 0을 사용하십시오.

참고: Dice 분포를 선택하는 경우, 분포모수 표가 약간 다릅니다. 표에 각 방향 또는 범주에 대해 한 행이 포함됩니다. 표에 값 열 및 확률 열이 포함됩니다. 값 열은 각 범주에 대한 레이블을 포함합니다. 레이블의 기본값은 1-N 정수이며 여기서, N은 방향의 수입니다. 레이블은 셀에 입력하여 편집할 수 있습니다. 셀에 임의의 값을 입력할 수 있습니다. 숫자가 아닌 값을 사용하려는 경우, 저장 유형이 아직 문자열로 설정되지 않았으면 데이터 필드의 저장 유형을 문자열로 변경해야 합니다. 확률 열은 각 범주의 확률을 포함합니다. 확률은 편집할 수 없으며 1/N로 계산됩니다.

미리보기. 지정된 모수를 기준으로 하여 분포의 표본 도표를 표시합니다. 한 모수에 대해 두 개 이상의 값을 지정하면 모수의 각 값에 대한 표본 도표가 표시됩니다. 선택한 필드에 대해 히스토리 데이터를 사용할 수 있으면 분포의 도표가 히스토리 데이터의 히스토그램에 겹쳐집니다.

선택적 설정. 이 옵션을 사용하여 시뮬레이션한 데이터에 대한 최소값, 최대값 또는 둘 다를 지정할 수 있습니다. 최소값보다 작고 최대값보다 큰 시뮬레이션한 데이터는 지정된 분포에 유효한 경우라도 거부됩니다.

- **최소값 지정. 아래 값 거부** 필드를 사용하려면 선택하십시오. 경험적 분포를 선택하면 이 확인 상자를 사용할 수 없습니다.
- **아래 값 거부. 최소값 지정**이 선택된 경우에만 사용 가능합니다. 시뮬레이션한 데이터에 대한 최소값을 입력하십시오. 이 값 보다 작은 시뮬레이션한 모든 값이 거부됩니다.
- **최대값 지정. 위 값 거부** 필드를 사용하려면 선택하십시오. 경험적 분포를 선택하면 이 확인 상자를 사용할 수 없습니다.
- **위 값 거부. 최대값 지정**이 선택된 경우에만 사용 가능합니다. 시뮬레이션한 데이터에 대한 최대값을 입력하십시오. 이 값 보다 큰 시뮬레이션한 모든 값이 거부됩니다.

확인. 대화 상자를 닫고 선택된 필드에 대한 시뮬레이션한 필드 표의 분포, 모수 및 최소, 최대 열의 값을 업데이트합니다. 상태 열의 아이콘도 선택된 분포를 반영하도록 업데이트됩니다.

취소. 대화 상자를 닫습니다. 작성된 변경사항을 삭제합니다.

반복

고정 필드 또는 분포모수에 대해 둘 이상의 값을 지정한 경우, 지정된 각 값에 대해 별도의 시뮬레이션에 대해 효과적인 시뮬레이션된 케이스의 독립변수 집합이 생성됩니다. 그러면 필드 또는 모수 변형의 효과를 조사할 수 있습니다. 각 시뮬레이션된 케이스 집합을 *반복*이라고 합니다. 시뮬레이션된 데이터에서 반복이 수직누적됩니다.

시뮬레이션 생성 노드 대화 상자의 고급 옵션 보기에서 **반복 작성 필드** 확인 상자를 선택하면 반복 필드가 시뮬레이션된 데이터에 숫자 저장 공간이 있는 명목 필드로 추가됩니다. 이 필드의 이름은 고급 옵션 보기의 **이름** 필드에 입력하여 편집할 수 있습니다. 이 필드에는 시뮬레이션된 각 케이스가 속한 반복을 표시하는 레이블이 포함됩니다. 레이블의 양식은 반복의 유형에 따라 다릅니다.

- **고정 필드 반복.** 레이블이 필드의 이름이며 뒤에 등호 부호가 표시되고 그 뒤에 해당 반복에 대한 필드의 값이 표시됩니다. 다음과 같습니다.

  field_name = *field_value*

- **분포모수 반복.** 레이블이 필드의 이름이며 뒤에 콜론이 오고 그 뒤에 반복 모수의 이름이 오고 그 뒤에 등호 부호가 표시되고 그 뒤에 해당 반복에 대한 필드의 값이 표시됩니다. 다음과 같습니다.

  field_name:*parameter_name* = *parameter_value*

- **범주형 또는 범위 분포에 대한 분포모수 반복.** 레이블이 필드의 이름이며 뒤에 콜론이 표시되고 그 뒤에 "Iteration"이 표시되고 그 뒤에 반복 수가 표시됩니다. 다음과 같습니다.

  field_name: Iteration *iteration_number*

분포

해당 필드에 대한 매개변수 지정 대화 상자를 열고 **분포** 목록에서 원하는 분포를 선택하고 **분포모수** 표에 분포모수를 입력하여 임의의 필드에 대한 확률 분포를 지정할 수 있습니다. 다음은 특정 분포에 대한 몇 가지 참고사항입니다.

- **범주형.** 범주형 분포는 범주라고 하는 고정 숫자의 숫자 값이 있는 입력 필드를 설명합니다. 각 범주에는 연관된 확률이 있으며 모든 범주의 확률 합계가 1이 됩니다.

  참고: 합계가 1이 되지 않는, 범주에 대한 확률을 지정하면 경고가 수신될 수 있습니다.
- **음이항 - 실패.** 지정된 수의 성공이 관측되기 전에 시도 시퀀스의 실패 수의 분포를 설명합니다. *Threshold* 매개변수는 지정된 수의 성공이며 *Probability* 매개변수는 지정된 시도 내의 성공 확률입니다.
- **음이항 - 시행.** 지정된 수의 성공이 관측되기 전에 요구되는 시도 수의 분포를 설명합니다. *Threshold* 매개변수는 지정된 수의 성공이며 *Probability* 매개변수는 지정된 시도 내의 성공 확률입니다.
- **범위.** 이 분포는 각 구간에 확률이 지정된 구간 집합으로 구성되며 전체 구간에 걸친 확률의 값이 1이 됩니다. 지정된 구간 내의 값은 해당 구간에 대해 정의된 균일 분포에서 파생됩니다. 구간은 최소값, 최대값 및 연관된 확률을 입력하여 지정됩니다.

  예를 들어, 원자재의 비용이 단위당 $10 - $15 범위에 해당될 가능성이 40%이며 단위당 $15 - $20 범위에 해당될 가능성이 60%라고 가정해 봅시다. 그러면 첫 번째 구간과 연관된 확률을 0.4로 설정하고 두 번째 구간과 연관된 확률을 0.6으로 설정하여 두 개의 구간 [10 - 15] 및 [15 - 20]으로 구성되는 범위 분포로 비용을 모델링할 수 있습니다. 구간이 연속될 필요가 없으며 겹칠 수도 있습니다. 예를 들어, $10 - $15 및 $20 - $25 또는 $10 - $15 및 $13 - $16의 구간으로 지정할 수도 있습니다.
- **와이블.** *Location* 매개변수가 선택적 위치 매개변수이며 분포의 원점 위치를 지정합니다.

다음은 사용자 정의 분포 맞춤에 대해 사용할 수 있는 분포 및 매개변수에 허용 가능한 값을 표시하는 표입니다. 이러한 분포 중 일부는 시뮬레이션 적합 노드에 의해 저장 유형에 자동으로 맞춰지지 않는 경우에도 특정 저장 유형에 대해 사용자 정의 맞춤을 수행하는 데 사용 가능합니다.

표 12. 사용자 정의 맞춤에 대해 사용 가능한 분포

분포	사용자 정의 맞춤에 대해 지원되는 저장 유형	매개변수	매개변수 한계	참고
베르누이	정수, 실수, 날짜/시간	확률	0 ≤ *Probability* ≤ 1	
베타	정수, 실수, 날짜/시간	모양 1 모양 2 최소값최대	≥ 0 ≥ 0 < *Maximum* > *Minimum*	최소값 및 최대값은 선택적입니다.
이항	정수, 실수, 날짜/시간	시행 수(n) 확률 최소값최대	> 0, 정수 0 ≤ *Probability* ≤ 1 < *Maximum* > *Minimum*	시행 수는 정수여야 합니다. 최소값 및 최대값은 선택적입니다.
범주형	정수, 실수, 날짜/시간, 문자열	범주 이름 (또는 레이블)	0 ≤ *Value* ≤ 1	값은 범주의 확률입니다. 값의 합계가 1이 되어야 합니다. 그렇지 않으면 경고가 생성됩니다.

표 12. 사용자 정의 맞춤에 대해 사용 가능한 분포 (계속)

분포	사용자 정의 맞춤에 대해 지원되는 저장 유형	매개변수	매개변수 한계	참고
Dice	정수, 문자열	방향	2 ≤ *Sides* ≤ 20	각 범주(방향)의 확률이 1/N로 계산되며 N은 방향의 수입니다. 확률을 편집할 수 없습니다.
경험적 분포	정수, 실수, 날짜/시간			경험적 분포를 편집하거나 유형으로 선택할 수 없습니다. 경험적 분포는 히스토리 데이터가 있는 경우에만 사용할 수 있습니다.
지수	정수, 실수, 날짜/시간	척도최소값최대값	> 0 < *Maximum* > *Minimum*	최소값 및 최대값은 선택적입니다.
고정	정수, 실수, 날짜/시간, 문자열	값		모든 필드에 대해 고정 분포를 지정할 수는 없습니다. 생성된 데이터의 모든 필드가 고정되도록 하려면 뒤에 균형 노드가 있는 사용자 입력 노드를 사용하십시오.
감마	정수, 실수, 날짜/시간	형태 척도최소값최대	≥ 0 ≥ 0 < *Maximum* > *Minimum*	최소값 및 최대값은 선택적입니다. 분포가 형상 모수 α = k이며 역 척도 모수 β = 1/θ인 비율 모수를 사용합니다.
로그정규	정수, 실수, 날짜/시간	모양 1 모양 2 최소값최대	≥ 0 ≥ 0 < *Maximum* > *Minimum*	최소값 및 최대값은 선택적입니다.
음이항 - 실패	정수, 실수, 날짜/시간	임계값 확률 최소값최대	≥ 0 0 ≤ *Probability* ≤ 1 < *Maximum* > *Minimum*	최소값 및 최대값은 선택적입니다.
음이항 - 시행	정수, 실수, 날짜/시간	임계값 확률 최소값최대	≥ 0 0 ≤ *Probability* ≤ 1 < *Maximum* > *Minimum*	최소값 및 최대값은 선택적입니다.
정규	정수, 실수, 날짜/시간	평균표준 편차 최소값최대	≥ 0 > 0 < *Maximum* > *Minimum*	최소값 및 최대값은 선택적입니다.
포아송	정수, 실수, 날짜/시간	평균최소값최대	≥ 0 < *Maximum* > *Minimum*	최소값 및 최대값은 선택적입니다.
범위	정수, 실수, 날짜/시간	시작(X) 끝(X) 확률(X)	0 ≤ *Value* ≤ 1	X는 각 구간의 지수입니다. 확률 값의 합계가 1이 되어야 합니다.
삼각	정수, 실수, 날짜/시간	모드 최소값최대	*Minimum* ≤ *Value* ≤ *Maximum* < *Maximum* >	*Minimum*

표 12. 사용자 정의 맞춤에 대해 사용 가능한 분포 (계속)

분포	사용자 정의 맞춤에 대해 지원되는 저장 유형	매개변수	매개변수 한계	참고
균일	정수, 실수, 날짜/시간	최소값최대	< *Maximum* > *Minimum*	
와이블	정수, 실수, 날짜/시간	비율 척도위치최소값최대	> 0 > 0 ≥ 0 < *Maximum* > *Minimum*	위치, 최대값 및 최소값은 선택적입니다.

데이터 보기 노드

데이터 보기 노드를 사용하여 스트림의 IBM SPSS Collaboration and Deployment Services 분석 데이터 보기에서 정의된 데이터를 포함하십시오. 분석 데이터 보기는 예측 모형 및 비즈니스 규칙에서 사용된 엔티티를 설명하는 데이터에 액세스하기 위한 구조를 정의합니다. 이 보기는 데이터 구조를, 분석을 위한 실제 데이터 소스와 연관시킵니다.

예측 분석에는 각 행이 예측이 작성되는 엔티티에 해당하는 테이블로 구성된 데이터가 필요합니다. 테이블의 각 열은 엔티티의 측정 가능한 속성을 표시합니다. 일부 속성은 다른 속성의 값을 통합하여 파생될 수 있습니다. 예를 들어, 테이블의 행은 고객 이름, 성별, 우편번호, 고객이 이전 연도에 $500 이상 구입한 횟수에 해당하는 열로 고객을 표시할 수 있습니다. 마지막 열은 일반적으로 하나 이상의 관련 테이블에 저장되는 고객 주문 히스토리에서 파생됩니다.

예측 분석 프로세스에서는 모델의 라이프사이클 전체에서 여러 가지 다른 데이터 세트를 사용합니다. 예측 모형의 초기 개발 중 예측 중인 이벤트의 알려진 결과가 있는 히스토리 데이터를 사용합니다. 모델 효과와 정확도를 평가하려면 다른 데이터와 비교하여 후보 모델의 유효성을 검증합니다. 모델의 유효성을 검증한 후 프로덕션 사용에 배치하여 일괄처리 프로세스에 있는 여러 엔티티 또는 실시간 프로세스에 있는 단일 엔티티의 스코어를 생성합니다. 모델을 의사결정 관리 프로세스의 비즈니스 규칙과 결합하는 경우 시뮬레이션된 데이터를 사용하여 조합 결과의 유효성을 검증합니다. 그러나 사용되는 데이터가 모델 개발 프로세스 단계에서 다른 경우에도 각 데이터 세트는 모델에 대해 동일한 속성 세트를 제공해야 합니다. 속성 세트는 계속 일정하지만 분석 중인 데이터 레코드는 변경됩니다.

분석 데이터 보기는 예측 분석의 특수 요구를 해결하는 다음과 같은 구성요소로 구성되어 있습니다.

- 관련 테이블로 구성된 속성 세트로 데이터에 액세스하기 위한 논리 인터페이스를 정의하는 데이터 보기 스키마 또는 데이터 모델. 모델의 속성은 다른 속성에서 파생될 수 있습니다.
- 데이터 모델 속성에 실제 값을 제공하는 하나 이상의 데이터 액세스 계획. 특정 애플리케이션에 대해 활성화된 데이터 액세스 계획을 지정하여 데이터 모델에 사용 가능한 데이터를 제어합니다.

중요사항: 데이터 보기 노드를 사용하려면 먼저 사이트에서 IBM SPSS Collaboration and Deployment Services Repository를 설치하고 구성해야 합니다. 노드에서 참조하는 분석 데이터 보기는 일반적으로 IBM SPSS Collaboration and Deployment Services Deployment Manager를 사용하여 리포지토리에서 작성되고 저장됩니다.

데이터 보기 노드에 대한 옵션 설정

데이터 보기 노드 대화 상자의 **데이터** 탭에 있는 옵션을 사용하여 IBM SPSS Collaboration and Deployment Services Repository에서 선택된 분석 데이터 보기에 대한 데이터 설정을 지정하십시오.

분석 데이터 보기. 생략 기호 단추(**...**)를 클릭하여 분석 데이터 보기를 선택하십시오. 리포지토리 서버에 현재 연결되어 있지 않은 경우에는 리포지토리: 서버 대화 상자에서 서버에 대한 URL을 지정하고 **확인**을 클릭한 후 리포지토리: 신임 정보 대화 상자에서 연결 신임 정보를 지정하십시오. 리포지토리에 로그인하고 오브젝트를 검색하는 것에 대한 자세한 정보는 IBM SPSS Modeler 사용자 안내서를 참조하십시오.

테이블 이름. 분석 데이터 보기의 데이터 모델에서 테이블을 선택하십시오. 데이터 모델의 각 테이블은 예측 분석 프로세스에 관련된 개념 또는 엔티티를 표시합니다. 테이블에 대한 필드는 테이블이 나타내는 엔티티의 속성에 해당합니다. 예를 들어, 고객 주문을 분석하는 경우 데이터 모델에는 고객에 대한 테이블과 주문에 대한 테이블이 포함될 수 있습니다. 고객 테이블에는 고객 ID, 연령, 성별, 혼인 여부, 거주 국가에 대한 속성이 있을 수 있습니다. 주문 테이블에는 주문 ID, 주문의 품목 수, 총 비용, 주문한 고객의 ID에 대한 속성이 있을 수 있습니다. 고객 ID 속성을 사용하여 고객 테이블에 있는 고객을 주문 테이블에 있는 해당 고객의 주문과 연관시킬 수 있습니다.

데이터 액세스 계획. 분석 데이터 보기에서 데이터 액세스 계획을 선택하십시오. 데이터 액세스 계획은 분석 데이터 보기의 데이터 모델 테이블을 실제 데이터 소스와 연관시킵니다. 분석 데이터 보기는 일반적으로 여러 데이터 액세스 계획을 포함하고 있습니다. 사용 중인 데이터 액세스 계획을 변경하는 경우 스트림에서 사용하는 데이터를 변경합니다. 예를 들어, 분석 데이터 보기에 모델 훈련을 위한 데이터 액세스 계획과 모델 검정을 위한 데이터 액세스 계획이 포함되어 있는 경우 사용 중인 데이터 액세스 계획을 변경하여 훈련 데이터에서 검정 데이터로 전환할 수 있습니다.

선택적 속성. 분석 데이터 보기를 사용하는 애플리케이션에 특정 속성이 필요하지 않은 경우 해당 속성을 선택사항으로 표시할 수 있습니다. 필수 속성과는 달리 선택적 속성은 널값을 포함할 수 있습니다. 선택적 속성에 대한 널값 처리를 포함하도록 애플리케이션을 조정해야 할 수 있습니다. 예를 들어, IBM Operational Decision Manager에서 작성된 비즈니스 규칙을 호출할 때 IBM Analytical Decision Management는 규칙 서비스를 쿼리하여 필수인 입력을 판별합니다. 스코어링할 레코드에 규칙 서비스의 필수 필드에 대한 널값이 포함되어 있는 경우 해당 규칙은 호출되지 않으며 규칙의 출력 필드는 기본값으로 채워집니다. 선택적 필드에 널값이 포함되어 있으면 해당 규칙이 호출됩니다. 이 규칙은 널값을 확인하여 처리를 제어합니다.

속성을 선택사항으로 지정하려면 **선택적 속성**을 클릭한 후 선택사항인 속성을 선택하십시오.

필드에 XML 데이터 포함. 데이터의 각 행에 대해 실행 가능 오브젝트 모델 XML 데이터가 포함된 필드를 작성하려면 이 옵션을 선택하십시오. 데이터가 IBM Operational Decision Manager와 함께 사용될 경우 이 정보는 필수입니다. 이 새 필드의 이름을 지정하십시오.

지리 공간적 소스 노드

지리 공간적 소스 노드를 사용하여 맵 또는 지리 공간적 데이터를 데이터 마이닝 세션으로 가져옵니다. 다음 두 방법 중 하나로 데이터를 가져올 수 있습니다.

- 모양 파일(.shp)로
- 맵 파일을 포함하는 계층 구조 파일 시스템이 포함된 ESRI 서버에 연결하여

  참고: 공용 맵 서비스에만 연결할 수 있습니다.

STP(Spatio-Temporal Prediction) 모델은 예측에 맵 또는 공간 요소를 포함할 수 있습니다. 이 모델에 대한 자세한 정보는 Modeler 모델링 노드 안내서(ModelerModelingNodes.pdf)의 시계열 모델 절에서 "STP(Spatio-Temporal Prediction) 모델링 노드" 주제를 참조하십시오.

지리 공간적 소스 노드에 대한 옵션 설정

데이터 소스 유형 **모양 파일**(.shp)에서 데이터를 가져오거나 **맵 서비스**에 연결할 수 있습니다.

모양 파일을 사용하는 경우에는 파일 이름 및 파일 경로를 입력하거나 파일을 찾아서 선택하십시오. 파일은 로컬 디렉토리에 있거나 맵핑된 드라이브에서 액세스해야 합니다. UNC(Uniform Naming Convention) 경로를 사용하여 파일에 액세스할 수는 없습니다.

참고: 모양 데이터에는 .shp 파일과 .dbf 파일이 모두 필요합니다. 두 파일은 동일한 이름을 가지고 동일한 폴더에 있어야 합니다. .shp 파일을 선택하면 .dbf 파일을 자동으로 가져옵니다. 또한 모양 데이터에 대한 좌표계를 지정하는 .prj 파일이 있을 수 있습니다.

맵 서비스를 사용하는 경우에는 서비스에 대한 URL을 입력한 후 **연결**을 클릭하십시오. 서비스에 연결하고 나면 해당 서비스 내 레이어가 **사용 가능한 맵** 분할창의 트리 구조에서 대화 상자의 맨 아래에 표시됩니다. 트리를 펼쳐서 필요한 레이어를 선택하십시오.

참고: 공용 맵 서비스에만 연결할 수 있습니다.

지리 공간적 데이터의 자동 정의

기본적으로 SPSS Modeler는 가능한 경우 올바른 메타데이터를 가진 소스 노드에서 지리 공간적 데이터 필드를 자동으로 정의합니다. 메타데이터는 점 또는 다각형 등의 지리 공간적 필드의 측정 수준과 필드에서 사용하는 좌표계(원점(예: 위도 0, 경도 0) 및 측정 단위 등의 세부사항 포함)를 포함할 수 있습니다. 측정 수준에 대한 자세한 정보는 142 페이지의 『지리 공간적 측정 수준』의 내용을 참조하십시오.

모양 파일을 구성하는 .shp 및 .dbf 파일에는 키로 사용되는 공통 식별자 필드가 포함되어 있습니다. 예를 들어, .shp 파일은 국가(식별자로 사용되는 국가명 필드 포함)를 포함하고 .dbf 파일은 식별자로도 사용되는 국가명이 포함된 해당 국가에 대한 정보를 포함할 수 있습니다.

참고: 좌표계가 기본 SPSS Modeler 좌표계와 동일하지 않은 경우에는 필요한 좌표계를 사용하도록 데이터를 재투영해야 할 수 있습니다. 자세한 정보는 190 페이지의 『재투영 노드』의 내용을 참조하십시오.

공통 소스 노드 탭

다음은 해당되는 탭을 클릭하여 모든 소스에 지정할 수 있는 옵션입니다.

- **데이터 탭.** 기본 저장 유형을 변경하는 데 사용됩니다.
- **필터 탭.** 데이터 필드를 제거하거나 이름을 변경하는 데 사용됩니다. 이 탭은 필터 노드와 같은 기능을 제공합니다.자세한 정보는 154 페이지의 『필터링 옵션 설정』의 내용을 참조하십시오.
- **유형 탭.** 측정 수준을 설정하는 데 사용됩니다. 이 탭은 유형 노드와 같은 기능을 제공합니다.
- **주석(Annotation) 탭.** 이 탭은 모든 노드에 사용되며 노드의 이름을 바꾸고 사용자 맞춤 도구팁을 제공하며 긴 주석(Annotation)을 저장하기 위한 옵션을 제공합니다.

소스 노드에서 측정 수준 설정

필드 특성은 소스 노드 또는 별도의 유형 노드에서 지정할 수 있습니다. 두 노드 모두에서 기능은 비슷합니다. 사용 가능한 특성은 다음과 같습니다.

- **필드** 필드 이름을 두 번 클릭하여 IBM SPSS Modeler에서 데이터에 대한 값 및 필드 레이블을 지정하십시오. 예를 들어, IBM SPSS Statistics에서 가져온 필드 메타데이터를 여기서 보거나 수정할 수 있습니다. 마찬가지로 필드 및 해당 값에 대한 새 레이블을 작성할 수 있습니다. 여기서 지정하는 레이블은 스트림 특성 대화 상자에서 작성하는 선택사항에 따라 IBM SPSS Modeler 전체에 표시됩니다.
- **측정** 지정된 필드의 데이터 특성을 설명하는 데 사용되는 측정 수준입니다. 필드의 모든 세부사항이 알려져 있는 경우 이를 **완전히 인스턴스화**되어 있다고 합니다. 자세한 정보는 140 페이지의 『측정 수준』의 내용을 참조하십시오.

  참고: 필드의 측정 수준은 데이터가 문자열, 정수, 실수, 날짜, 시간소인, 목록 중 어느 것으로 저장되는지를 표시하는 해당 저장 유형과 다릅니다.
- **값** 이 열에서는 데이터 세트에서 데이터 값을 읽어오기 위한 옵션을 지정하거나 **지정** 옵션을 사용하여 별도의 대화 상자에서 측정 수준 및 값을 지정할 수 있습니다. 해당 값을 읽지 않고 필드를 통과하도록 선택할 수도 있습니다. 자세한 정보는 145 페이지의 『데이터 값』의 내용을 참조하십시오.

  참고: 해당 **필드** 항목에 목록이 포함되어 있으면 이 열에서 셀을 수정할 수 없습니다.
- **결측** 필드의 결측값이 처리되는 방식을 지정하는 데 사용됩니다. 자세한 정보는 149 페이지의 『결측값 정의』의 내용을 참조하십시오.

  참고: 해당 **필드** 항목에 목록이 포함되어 있으면 이 열에서 셀을 수정할 수 없습니다.
- **확인** 이 열에서는 필드 값이 지정된 값 또는 범위를 준수하는지 확인하는 옵션을 설정할 수 있습니다. 자세한 정보는 150 페이지의 『유형 값 검사』의 내용을 참조하십시오.

  참고: 해당 **필드** 항목에 목록이 포함되어 있으면 이 열에서 셀을 수정할 수 없습니다.
- **역할** 필드가 시스템 학습 프로세스에 대해 **입력**(예측변수 필드)인지 아니면 **목표**(예측 필드)인지를 모델링 노드에 알리는 데 사용됩니다. 레코드를 훈련, 검정 및 검증을 위한 별도의 표본으로 파티셔닝하는 데 사용

되는 필드를 표시하는 **파티션**과 함께 **모두** 및 **없음** 역할도 사용할 수 있습니다. **분할** 값은 필드의 가능한 각각의 값에 대해 별도의 모델이 작성되도록 지정합니다. 자세한 정보는 150 페이지의 『필드 역할 설정』의 내용을 참조하십시오.

자세한 정보는 138 페이지의 『유형 노드』의 내용을 참조하십시오.

소스 노드에서 인스턴스화할 시기

두 가지 방법으로 필드의 값 및 데이터 저장 공간에 대해 학습할 수 있습니다. 이 **인스턴스화**는 처음으로 데이터를 IBM SPSS Modeler로 가져올 때 소스 노드에서 발생하거나 유형 노드를 데이터 스트림에 삽입하여 발생할 수 있습니다.

소스 노드에서 인스턴스화는 다음 경우에 유용합니다.

- 데이터 세트가 작은 경우
- 표현식 작성기를 사용하여 새 필드를 파생시키려고 계획하는 경우(인스턴스화를 수행하면 표현식 작성기에서 필드 값을 사용할 수 있게 됨)

일반적으로 데이터 세트가 그다지 크지 않고 나중에 스트림에서 필드를 추가하지 않을 계획인 경우에는 소스 노드에서 인스턴스화가 가장 편리한 방법입니다.

소스 노드에서 필드 필터링

소스 노드 대화 상자의 필터 탭을 사용하면 데이터의 초기 검사를 기준으로 하여 다운스트림 작업에서 필드를 제외할 수 있습니다. 예를 들어, 데이터에 중복 필드가 있거나 관련이 없는 필드를 제외할 수 있을 정도로 데이터에 익숙한 경우에 유용합니다. 또는 나중에 스트림에 별도의 필터 노드를 추가할 수도 있습니다. 기능은 두 경우 모두 유사합니다. 자세한 정보는 154 페이지의 『필터링 옵션 설정』의 내용을 참조하십시오.

제 3 장 레코드 작업 노드

레코드 작업 개요

레코드 작업 노드는 레코드 수준에서 데이터를 변경하는 데 사용됩니다. 이러한 작업은 데이터 마이닝의 **데이터 이해** 및 **데이터 준비** 단계에서 중요합니다. 사용자가 데이터를 사용자의 특정 비즈니스 요구에 맞출 수 있기 때문입니다.

예를 들어, 데이터 검토 노드(출력 팔레트)를 사용하여 수행된 데이터 검토의 결과를 기준으로 하여 지난 삼 개월 동안의 고객 구매 레코드를 합치도록 결정할 수 있습니다. 합치기 노드를 사용하여 *고객 ID* 등의 키 필드 값을 기준으로 하여 레코드를 합칠 수 있습니다. 또는 웹 사이트 방문에 대한 정보를 포함하는 데이터베이스의 레코드가 백만 개 이상이 되면 관리할 수 없다는 것을 발견하게 될 수도 있습니다. 표본 노드를 사용하여 모델링에 사용할 데이터 서브세트를 선택할 수 있습니다.

레코드 작업 팔레트에는 다음 노드가 포함됩니다.

선택 노드는 특정 조건을 기반으로 데이터 스트림에서 레코드의 서브세트를 선택 또는 삭제합니다. 예를 들어, 특정 영업 지역에 관련된 레코드를 선택할 수 있습니다.

표본 노드는 레코드의 서브세트를 선택합니다. 층화, 수평배열, 비임의(구조화) 표본을 포함하여 다양한 표본 유형이 지원됩니다. 표본추출은 성능을 개선하고 분석을 위해 관련 레코드나 트랜잭션 집단을 선택하는 데 유용할 수 있습니다.

균형 노드는 데이터 세트의 불균형을 정정하므로, 데이터 세트가 지정된 조건을 준수합니다. 균형 지시문이 조건이 지정된 요인만큼 참인 레코드 비율을 조정합니다.

통합 노드는 입력 레코드의 시퀀스를 요약되고 통합된 출력 레코드로 대체합니다.

RFM(Recency, Frequency, Monetary) 통합 노드를 사용하면 고객의 히스토리 트랜잭션 데이터를 취하고 모든 사용하지 않은 데이터를 제거하고 마지막으로 다룬 시기, 작성한 트랜잭션 수, 해당 트랜잭션의 구매총액을 나열하는 단일 행으로 모든 나머지 트랜잭션 데이터를 결합할 수 있습니다.

정렬 노드는 하나 이상의 필드의 값을 기반으로 레코드를 내림차순 또는 오름차순으로 정렬합니다.

병합 노드는 다중 입력 레코드를 취하고 입력 필드의 일부 또는 모두를 포함하는 단일 출력 레코드를 작성합니다. 내부 고객 데이터 및 구매한 인구통계학적 데이터 같은 상이한 소스의 데이터 병합에 유용합니다.

붙여쓰기 노드는 레코드 세트를 연결합니다. 비슷한 구조를 갖지만 상이한 데이터를 갖는 데이터 세트 결합에 유용합니다.

고유 노드는 첫 번째 고유 레코드를 데이터 스트림으로 전달하거나 첫 번째 레코드를 삭제하고 대신 모든 중복을 데이터 스트림으로 전달하여 중복 레코드를 제거합니다.

스트리밍 시계열 노드는 한 단계로 시계열 모델을 작성하고 스코어링합니다. 로컬 또는 분산 환경의 데이터와 함께 노드를 사용할 수 있으며 분산 환경에서는 IBM SPSS Analytic Server의 기능을 이용할 수 있습니다.

레코드 작업 팔레트 내의 많은 노드에 대해 사용자가 CLEM 표현식을 사용해야 합니다. CLEM에 익숙한 경우에는 필드에 표현식을 입력할 수 있습니다. 단, 모든 표현식 필드가 CLEM 표현식 작성기를 여는 단추를 제공하므로 이를 사용하여 해당 표현식을 자동으로 작성하는 것이 도움이 됩니다.

그림 1. 표현식 작성기 단추

선택 노드

선택 노드를 사용하여 BP(혈압) = "높음"과 같이 특정 조건을 기준으로 하여 데이터 스트림에서 레코드의 서브세트를 선택 또는 삭제할 수 있습니다.

모드. 조건을 충족하는 레코드를 데이터 스트림에 대해 포함할 것인지 제외할 것인지 지정합니다.

- **포함.** 선택 조건을 충족하는 레코드를 포함하려면 선택하십시오.
- **삭제.** 선택 조건을 충족하는 레코드를 제외하려면 선택하십시오.

조건. CLEM 표현식을 사용하여 지정하는 각 레코드를 검정하기 위해 사용할 선택 조건을 표시합니다. 창에 표현식을 입력하거나 창의 오른쪽에 있는 계산기(표현식 작성기) 단추를 클릭하여 표현식 작성기를 사용하십시오.

다음과 같이 조건을 기준으로 하여 레코드를 삭제하도록 선택합니다.

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

기본적으로 선택 노드는 모든 선택 필드에 대해 널값을 가진 레코드도 삭제합니다. 이런 현상을 피하려면 다음 조건을 원본에 추가하십시오.

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

레코드의 비율을 선택하는 데도 선택 노드가 사용됩니다. 일반적으로 이 작업에 대해서는 다른 노드인 표본 노드를 사용합니다. 그러나 지정할 조건이 제공된 모수보다 복잡한 경우, 선택 노드를 사용하여 직접 조건을 작성할 수 있습니다. 예를 들어 다음과 같은 조건을 작성할 수 있습니다.

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

그러면 높은 혈압을 나타내는 40% 정도의 레코드가 선택되고 해당 레코드가 추가 분석을 위해 다운스트림에 전달됩니다.

표본 노드

표본 노드를 사용하여 분석할 레코드 서브세트를 선택하거나 삭제할 레코드 비율을 지정할 수 있습니다. 층화, 수평배열, 비임의(구조화) 표본을 포함하여 다양한 표본 유형이 지원됩니다. 표본추출은 다음과 같은 여러 가지 이유로 사용될 수 있습니다.

- 데이터 서브세트에서 모형을 추정하여 성능을 개선하기 위해서 사용됩니다. 표본에서 추정된 모형은 종종 전체 데이터 세트에서 파생된 모형만큼 정확하며 이전에는 불가능했던 다양한 방법을 개선된 성능을 사용하여 시도함으로써 더 정확한 결과를 얻을 수도 있습니다.
- 분석할 관련 레코드 또는 트랜잭션 집단을 선택하기 위해서 사용됩니다. 예를 들어, 온라인 장바구니에서 모든 항목을 선택하거나 특정 인접 항목에서 모든 특성을 선택하는 것이 있습니다.
- 품질 보장, 사기 방지 또는 보안 등의 목적으로 임의 검사를 수행하기 위해 단위 또는 케이스를 식별하기 위해서 사용됩니다.

참고: 검증 목적으로 데이터를 학습 및 검정 표본으로 파티션만 하면 되는 경우에는 파티션 노드를 대신 사용하십시오. 자세한 정보는 181 페이지의 『파티션 노드』의 내용을 참조하십시오.

표본 유형

군집 표본. 개별 단위가 아니라 표본 집단 또는 군집입니다. 예를 들어, 한 학생당 하나의 레코드가 있는 데이터 파일이 있다고 가정합니다. 학교 기준으로 군집화하고 표본 *크기*가 50%인 경우, 50%의 학교가 선택되고 선택된 각 학교의 모든 학생이 선택됩니다. 선택되지 않은 학교의 학생은 유효하지 않습니다. 평균적으로 약 50%의 학생이 선택될 것으로 예상하나 학교 *크기*가 다양하므로 퍼센트가 정확하지 않을 수 있습니다. 이와 유사하게, 장바구니 항목을 트랜잭션 ID 기준으로 군집화하여 선택된 트랜잭션의 모든 항목이 유지되는지 확인할 수 있습니다. 마을 기준으로 군집화되는 예를 보려면 *complexsample_property.str* 샘플 스트림을 참조하십시오.

층화 표본. 모집단 밀도 또는 계층의 겹치지 않는 하위 그룹 내에서 표본을 독립적으로 선택합니다. 예를 들어, 여성 및 남성이 동일한 비율로 선택되도록 하거나 도시 인구의 모든 지역 또는 사회 경제적 집단이 표시되도록 할 수 있습니다. 또한 각 계층별로 다른 표본 *크기*를 지정할 수 있습니다. 예를 들어, 한 집단이 원 데이

터에서 적게 표시된다고 생각할 수 있습니다. 지역 기준으로 특성을 층화하는 예를 보려면 *complexsample_property.str* 샘플 스트림을 참조하십시오.

계통 또는 n중1 표본추출. 임의선택이 어려운 경우 계통적(고정된 간격) 또는 순차적으로 단위를 표본추출할 수 있습니다.

표본추출 가중치. 표본추출 가중치는 복합 표본을 그릴 때 자동으로 계산되며 각 표본 단위가 원 데이터에서 나타내는 "빈도"와 거의 일치합니다. 따라서 표본에 대한 가중치 합계가 원 데이터의 *크기*를 추정해야 합니다.

표본추출 프레임

표본추출 프레임은 표본 또는 연구에 포함되는 케이스의 잠재적 소스를 정의합니다. 어떤 경우에는 모집단의 각각의 모든 멤버를 식별하고 그 각각을 표본에 포함시킬 수 있습니다. 예를 들어, 생산 공정에서 만들어내는 항목을 표본추출하는 경우입니다. 모든 가능한 케이스에 액세스할 수 없는 경우가 더 많습니다. 예를 들어, 선거가 시행될 때까지는 누가 선거에서 투표할 것인지 확인할 수 없습니다. 이 경우, 등록한 사람 중 일부가 투표하지 않으며 사용자가 등록을 확인하는 시점에 나열되지 않았음에도 투표하는 사람이 있더라도 선거 등록을 표본추출 프레임으로 사용할 수 있습니다. 표본추출 프레임 내에 없는 사람은 표본추출에 포함될 가능성이 없습니다. 표본추출 프레임이 평가하려고 시도하는 모집단의 성격에 충분히 근접한지 여부는 실생활의 각 케이스에서 반드시 다루어야 하는 안건입니다.

표본 노드 옵션

요구 사항에 따라 **단순** 또는 **복합** 방법을 선택할 수 있습니다.

단순 표본추출 옵션

단순 방법을 사용하면 레코드를 임의의 비율로 선택하거나 연속된 레코드를 선택하거나 모든 *n번째* 레코드를 선택할 수 있습니다.

모드. 다음 방법에 대해 레코드를 전달(포함)하거나 삭제(제외)할 수 있습니다.

- **표본 포함.** 선택된 레코드를 데이터 스트림에 포함시키고 모든 기타를 삭제합니다. 예를 들어, 모드를 **표본 포함**으로 설정하고 **n중1** 옵션을 5로 설정하면 매 다섯 번째 레코드가 포함되고 대략 원래 *크기*의 1/5인 데이터 세트가 생성됩니다. 이 모드가 데이터 표본추출의 기본 모드이며 복합 방법을 사용할 때의 유일한 모드입니다.
- **표본 삭제.** 선택된 레코드를 제외하고 모든 기타를 포함합니다. 예를 들어, 모드를 **표본 삭제**로 설정하고 **n중1** 옵션을 5로 설정하면 매 다섯 번째 레코드가 삭제됩니다. 이 모드는 단순 방법에서만 사용할 수 있습니다.

표본. 다음 옵션 중에서 표본추출 방법을 선택합니다.

- **처음.** 연속된 데이터 표본추출을 사용하도록 선택합니다. 예를 들어, 최대 표본 *크기*가 10000으로 설정되면 처음 10,000개의 레코드가 선택됩니다.
- **n중1.** 매 *n*번째 레코드를 전달하거나 삭제하여 데이터를 표본추출하도록 선택합니다. 예를 들어, *n*이 5로 설정되면 매 다섯 번째 레코드가 선택됩니다.

- **임의 %.** 데이터의 임의 퍼센트를 표본 추출하도록 선택합니다. 예를 들어, 퍼센트를 20으로 설정하면 선택한 모드에 따라 데이터의 20%가 데이터 스트림에 전달되거나 삭제됩니다. 표본추출 퍼센트를 지정하려면 필드를 사용하십시오. 또한 **난수 시작값 설정** 제어를 사용하여 시드 값을 지정할 수 있습니다.

  블록 수준 표본추출(In-Database 전용)을 사용하십시오. 이 옵션은 사용자가 Oracle 또는 IBM DB2 데이터베이스에서 In-Database 마이닝을 수행할 때 임의 퍼센트 표본추출을 선택한 경우에만 사용 가능합니다. 이러한 상황에서는 블록 수준 표본추출이 더 효율적일 수 있습니다.

  참고: 동일한 임의 표본 설정을 실행할 때마다 정확히 동일한 수의 행이 리턴되지는 않습니다. 그 이유는 각 입력 레코드의 확률이 표본에 포함되는 N/100이며(여기서, N은 사용자가 노드에서 지정한 **임의 %**임) 확률은 독립변수입니다. 따라서 결과가 정확히 N%가 아닙니다.

최대 표본 크기. 표본에 포함될 최대 레코드 수를 지정합니다. 이 옵션은 중복이므로 **First** 및 **Include**가 선택될 때는 사용하지 않습니다. 또한 **임의 %** 옵션과 함께 사용되는 경우에는 이 설정이 특정 레코드가 선택되는 것을 방지할 수 있습니다. 예를 들어, 데이터 세트에 천만 개의 레코드가 있으며 최대 표본 크기가 삼백만 레코드인 50%의 레코드를 선택한 경우, 첫 번째 육백만 개의 레코드의 50%가 선택되며 나머지 사백만 개의 레코드는 선택될 기회가 없습니다. 이 제한을 방지하기 위해 **복합** 표본추출 방법을 선택하고 군집 또는 층화 변수를 지정하지 말고 삼백만 개의 레코드의 임의 표본을 요청하십시오.

복합 표본추출 옵션

복합 표본 옵션을 사용하면 기타 옵션과 함께 군집, 층화 및 가중치 표본 등을 사용하여 표본을 더 미세하게 제어할 수 있습니다.

군집 및 층화. 필요에 따라 군집, 층화 및 입력 가중 필드를 지정할 수 있습니다. 자세한 정보는 78 페이지의 『군집 및 층화 설정』의 내용을 참조하십시오.

표본 유형.

- **임의.** 각 계층 내에서 군집 또는 레코드를 임의로 선택합니다.
- **계통.** 고정된 간격으로 레코드를 선택합니다. 이 옵션은 *n중1* 방법과 동일하게 작동하나 난수 시드에 따라 첫 번째 레코드의 위치가 변경된다는 점만 다릅니다. *n*의 값은 표본 크기 또는 비율을 기준으로 하여 자동으로 결정됩니다.

표본 단위. 기본 표본 단위로 비율 또는 빈도를 선택할 수 있습니다.

표본 크기. 여러 가지 방법으로 표본 크기를 지정할 수 있습니다.

- **고정됨.** 개수 또는 비율로 표본의 전체 크기를 지정할 수 있습니다.
- **사용자 정의.** 각 하위 그룹 또는 계층의 표본 크기를 지정할 수 있습니다. 이 옵션은 군집 및 층화 하위 대화 상자에서 층화 필드가 지정된 경우에만 사용 가능합니다.
- **변수.** 각 하위 그룹 또는 계층의 표본 크기를 정의하는 필드를 사용자가 선택할 수 있도록 허용합니다. 이 필드는 특정 계층 내의 각 레코드에 대해 동일한 값을 가져야 합니다. 예를 들어, 표본이 지역별로 계층화

된 경우, *county = Surrey* 내의 모든 레코드가 동일한 값을 가져야 합니다. 필드는 숫자여야 하며 해당 값이 선택된 표본 단위와 일치해야 합니다. 비율의 경우, 값이 0보다 크고 1보다 작아야 하며 개수의 경우 최소값이 1입니다.

계층 당 최소 표본. 최소 레코드 수를 지정합니다. 군집 필드가 지정된 경우에는 최소 군집 수가 지정됩니다.

계층 당 최대 표본. 레코드 또는 군집의 최대 수를 지정합니다. 군집 또는 층화 필드를 지정하지 않고 이 옵션을 선택하면 지정된 *크기*의 임의 또는 계통 표본이 선택됩니다.

난수 시드 설정. 난수 비율에 따라 레코드에 대해 표본 추출 및 파티셔닝을 수행하는 경우 이 옵션을 사용하면 다른 세션에서 동일한 결과를 복제할 수 있습니다. 난수 생성기에서 사용하는 시작값을 지정하면 노드를 실행할 때마다 동일한 레코드를 지정하도록 보장할 수 있습니다. 원하는 시드 값을 입력하거나 **생성** 단추를 클릭하여 자동으로 난수 값을 생성하십시오. 이 옵션을 선택하지 않으면 노드를 실행할 때마다 다른 표본이 생성됩니다.

참고: **난수 시드 설정** 옵션을 데이터베이스에서 읽은 레코드와 함께 사용할 경우에는 노드를 실행할 때마다 동일한 결과가 보장되도록 표본추출 이전에 정렬 노드가 필요할 수 있습니다. 난수 시드는 레코드 순서에 의존하여 관계형 데이터베이스에서는 동일하게 보장되지 않기 때문입니다. 자세한 정보는 86 페이지의 『정렬 노드』의 내용을 참조하십시오.

군집 및 층화 설정

군집 및 층화 대화 상자를 사용하면 복합 표본을 그릴 때 군집, 층화 및 가중 필드를 선택할 수 있습니다.

군집. 레코드를 군집화하는 데 사용할 범주형 필드를 지정합니다. 레코드는 일부 군집은 포함되고 다른 일부는 제외된 형태로 소속군집을 기준으로 하여 표본화됩니다. 단, 지정된 군집의 어느 한 레코드가 포함되면 모든 레코드가 포함됩니다. 예를 들어, 장바구니와 관련하여 제품을 분석할 때 항목을 트랜잭션 ID 기준으로 군집화하여 선택된 트랜잭션의 모든 항목이 유지되는지 확인할 수 있습니다. 함께 판매된 항목에 대한 정보를 영구 삭제하는 표본추출 레코드 대신 트랜잭션에 대한 표본추출을 수행하여 선택된 트랜잭션에 대한 모든 레코드를 유지할 수 있습니다.

계층화 기준. 모집단 밀도 또는 계층의 겹치지 않는 하위 그룹 내에서 표본이 독립적으로 선택되도록 레코드를 층화하는 데 사용되는 범주형 필드를 지정합니다. 성별 기준으로 층화된 50% 표본을 선택하면 남성에 대해 하나, 여성에 대해 하나의 두 개의 50% 표본이 선택됩니다. 예를 들어, 계층은 사회 경제적인 그룹, 작업 범주, 연령 그룹 또는 인종 그룹일 수 있으며 사용자가 관심 있는 부집단에 대한 적절한 표본 *크기*를 선택할 수 있도록 해줍니다. 원래 데이터 세트에 남성보다 여성이 세 배 많은 경우, 이 비율은 각 그룹에서 별도로 표본추출할 때 유지됩니다. 다중 층화 필드도 지정할 수 있습니다. 예를 들어, 지역 내의 제품군 표본추출 또는 그 반대의 경우도 가능합니다.

참고: 결측값(널 또는 시스템 결측값, 비어 있는 문자열, 공백 또는 사용자 정의 결측값)이 있는 필드를 기준으로 하여 계층화하는 경우, 계층에 대한 고객 표본 *크기*를 지정할 수 없습니다. 결측값 또는 공백값이 있는 필드를 기준으로 하여 계층화하는 경우에 사용자 정의 표본 *크기*를 사용하려면 해당 값을 채워야 합니다.

입력 가중치 사용. 표본추출 전에 레코드를 가중하는 데 사용되는 필드를 지정합니다. 예를 들어, 가중 필드에 1에서 5 범위의 값이 있는 경우, 5로 가중된 레코드는 5배 많이 선택되는 경향이 있습니다. 노드에 의해 생성된 최종 출력 가중값이 이 필드의 값을 덮어씁니다. 다음 단락을 참조하십시오.

새 출력 가중치. 입력 가중 필드가 지정되지 않은 경우에 최종 가중치가 작성되는 필드의 이름을 지정합니다. (입력 가중 필드가 지정되면 위에서 설명한 대로 그 값이 최종 가중값에 의해 교체되고 별도의 출력 가중 필드가 작성되지 않습니다.) 출력 가중값은 원 데이터 내의 표본추출된 각 레코드가 나타내는 레코드 수를 표시합니다. 가중값의 합계는 표본 *크기*에 대한 추정값을 제공합니다. 예를 들어, 10%의 임의 표본이 사용되면 출력 가중치가 모든 레코드에 대해 10이 됩니다. 즉, 표본추출된 각 레코드가 원 데이터에서의 약 10개의 레코드를 나타냅니다. 층화 또는 가중된 표본에서 출력 가중값은 각 계층에 대한 표본 비율에 따라 다릅니다.

설명

- 수평배열 표본추출은 표본추출할 모집단 분포의 전체 목록을 가져올 수 없으나 특정 그룹 또는 군집에 대한 전체 목록은 가져올 수 있는 경우에 유용합니다. 임의 표본이 접촉하기에 비실용적인 검정 개체를 생성하는 경우에도 유용합니다. 예를 들어, 국가 내의 모든 지역에 흩어져 있는 농부를 선택하기 보다 한 지역 내의 모든 농부를 방문하는 것이 더 쉬울 것입니다.
- 각 계층 내에서 독립적으로 군집을 표본추출하기 위해 군집 및 층화 필드를 둘 다 지정할 수 있습니다. 예를 들어, 지역별로 계층화된 특성 값을 표본추출하고 각 지역 내의 마을별로 군집화할 수 있습니다. 그러면 마을의 독립적 표본이 각 지역 내에 그려집니다. 일부 마을은 포함될 것이며 기타 마을은 포함되지 않을 것이나 포함되는 각 마을에 대해 마을 내의 모든 특성이 포함될 것입니다.
- 각 군집 내에서 임의 표본 단위를 선택하기 위해 두 표본 노드를 함께 연결할 수 있습니다. 예를 들어, 먼저 위에서 설명한 대로 지역 기준으로 계층화된 마을을 표본추출할 수 있습니다. 그런 다음 두 번째 표본 노드를 첨부하고 층화 필드로 *마을*을 선택하여 각 마을의 레코드의 비율을 표본추출할 수 있습니다.
- 군집을 고유하게 식별하기 위해 필드의 조합이 필요한 경우, 파생 노드를 사용하여 새 필드가 생성될 수 있습니다. 예를 들어, 여러 가게가 트랜잭션에 대해 동일한 번호 매기기 시스템을 사용하는 경우, 가게와 트랜잭션 ID를 연결하는 새 필드를 파생시킬 수 있습니다.

계층에 대한 표본 크기

층화 표본을 그릴 때 기본 옵션은 각 계층의 레코드 또는 군집에서 동일한 비율로 표본을 추출하는 것입니다. 예를 들어, 한 집단이 다른 집단보다 그 수가 세 배 많은 경우, 일반적으로 표본에서도 동일한 비율을 유지하려고 할 것입니다. 그러나 이런 케이스가 아니라면 각 계층에 대해 별도로 표본 *크기*를 지정할 수 있습니다.

계층에 대한 표본 *크기* 대화 상자에는 층화 필드의 각 값이 나열되어 있어서 사용자가 해당 계층의 기본값을 대체할 수 있습니다. 다중 층화 필드를 선택하면 가능한 모든 값 조합이 나열되어 사용자가 각 도시 내의 각 인종 집단에 대한 *크기* 또는 각 지역 내의 각 마을에 대한 *크기* 등을 지정할 수 있습니다. *크기*는 표본 노드의 현재 설정에 의해 결정되는 비율 또는 빈도로 지정됩니다.

계층에 대한 표본 *크기*를 지정하려면 다음을 수행하십시오.

1. 표본 노드에서 **복합**을 선택하고 하나 이상의 층화 필드를 선택하십시오. 자세한 정보는 78 페이지의 『군집 및 층화 설정』의 내용을 참조하십시오.

2. **사용자 정의**를 선택하고 *크기* **지정**을 선택하십시오.
3. 계층에 대한 표본 *크기* 대화 상자의 왼쪽 하단에 있는 **값 읽기** 단추를 클릭하여 표시를 채우십시오. 필요한 경우 업스트림 소스 또는 유형 노드에서 값을 인스턴스화할 수 있습니다. 자세한 정보는 144 페이지의 『인스턴스화 개념』의 내용을 참조하십시오.
4. 임의의 행을 클릭하여 해당 계층에 대한 기본 *크기*를 대체하십시오.

표본 *크기*에 대한 참고

다양한 계층에 다양한 분산이 있는 경우, 예를 들어, 표준 편차에 비례하는 표본 *크기*를 작성하는 경우에 사용자 정의 표본 *크기*가 유용합니다. (계층 내의 케이스가 더 다양한 경우에는 더 많은 표본을 추출하여 대표 표본을 얻어야 합니다.) 계층이 작은 경우에는 높은 표본 비율을 사용하여 관측값의 최소수가 포함되도록 할 수 있습니다.

참고: 결측값(널 또는 시스템 결측값, 비어 있는 문자열, 공백 또는 사용자 정의 결측값)이 있는 필드를 기준으로 하여 계층화하는 경우, 계층에 대한 고객 표본 *크기*를 지정할 수 없습니다. 결측값 또는 공백값이 있는 필드를 기준으로 하여 계층화하는 경우에 사용자 정의 표본 *크기*를 사용하려면 해당 값을 채워야 합니다.

균형 노드

균형 노드를 사용하면 지정된 테스트 기준을 준수하도록 데이터 세트에서 불균형을 정정할 수 있습니다. 예를 들어, 데이터 세트에 *low* 또는 *high*의 두 값만 있으며 케이스의 10%만 *high*일 때 케이스의 90%가 *low*라고 가정하십시오. 여러 모델링 기술은 거의 드물지만 *low* 결과만 학습하고 *high* 결과는 무시하는 경향이 있으므로 여러 모델링 기술을 사용할 때 이러한 편향 데이터에 문제가 있습니다. 데이터가 대략적으로 동일한 수의 *low* 및 *high* 결과와 올바른 균형을 유지하는 경우 모델은 두 그룹을 구별하는 패턴을 찾을 가능성이 높습니다. 이 경우 균형 노드는 *low* 결과의 케이스를 줄이는 균형 지시문을 작성하는 데 유용합니다.

사용자가 지정하는 조건에 기반하여 레코드를 복제한 후 버려서 균형이 수행됩니다. 어떤 조건도 없는 레코드는 항상 전달됩니다. 레코드를 복제하거나 버려서 이 프로세스가 작동되므로 원래 데이터 시퀀스는 다운스트림 작업에서 유실됩니다. 데이터 스트림에 균형 노드를 추가하기 전에 시퀀스 관련 값을 파생시켜야 합니다.

참고: 분포 차트 및 히스토그램에서 균형 노드가 자동으로 생성될 수 있습니다. 예를 들어, 분포 도표에 표시되어 있는 바와 같이 범주형 필드의 모든 범주에 동일한 비율을 표시하도록 데이터의 균형을 유지할 수 있습니다.

예. 이전 마케팅 캠페인에 긍정적으로 반응한 최근 고객을 식별하기 위해 RFM 스트림을 작성하는 경우 판매 회사의 마케팅 부서는 균형 노드를 사용하여 데이터에 있는 true 반응과 false 반응의 차이에 대한 균형을 유지합니다.

균형 노드의 옵션 설정

레코드 균형 지시문. 현재 균형 지시문을 나열합니다. 각 지시문에는 "조건이 true인 경우 계수를 지정하여 레코드 비율을 증가시키도록" 소프트웨어에 지시하는 조건과 계수가 포함됩니다. 1.0 미만의 계수는 표시된 레코드의 비율이 감소됨을 의미합니다. 예를 들어, Y 약물이 치료 약물인 레코드 수를 줄이려면 계수 0.7과 조건

Drug = "drugY"를 사용하여 균형 지시문을 작성할 수 있습니다. 이 지시문은 Y 약물이 치료 약물인 레코드 수가 모든 다운스트림 작업에 대해 70%로 감소됨을 의미합니다.

참고: 감소를 위한 균형 계수가 4자리의 소수 자리로 지정될 수 있습니다. 계수가 0.0001 미만으로 설정된 경우 결과가 올바르게 계산되지 않으므로 오류가 발생합니다.

- 텍스트 필드 오른쪽의 단추를 클릭하여 **조건 작성**을 수행하십시오. 이 경우 새 조건을 입력할 비어 있는 행이 삽입됩니다. 조건으로 CLEM 표현식을 작성하려면 표현식 작성기 단추를 클릭하십시오.
- 빨간색 삭제 단추를 사용하여 **지시문 삭제**를 수행하십시오.
- 위로 및 아래로 화살표 단추를 사용하여 **지시문 정렬**을 수행하십시오.

균형 학습 데이터만. 파티션 필드가 스트림에 있는 경우 이 옵션은 학습 파티션의 데이터 균형만 유지합니다. 특히 조정된 성향 스코어를 생성하는 경우에 유용할 수 있으며 불균형 테스트 또는 검증 파티션이 필요합니다. 파티션 필드가 스트림에 없는 경우(또는 여러 파티션 필드가 지정된 경우) 이 옵션은 무시되고 모든 데이터의 균형이 유지됩니다.

통합 노드

통합은 데이터 세트의 *크기*를 줄이기 위해 자주 사용되는 데이터 준비 작업입니다. 통합을 계속하기 전에 오래 걸려도 데이터를 정리해야 하며 특히 결측값에 집중해야 합니다. 통합했으면 결측값과 관련된 잠재적으로 유용한 정보가 유실될 수 있습니다.

통합 노드를 사용하여 입력 레코드 시퀀스를 요약, 통합 출력 레코드로 대체할 수 있습니다. 예를 들어, 다음 표와 같은 입력 판매 레코드 세트가 있을 수 있습니다.

표 13. 판매 레코드 입력 예

연령	성별	지역	지점	판매
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

성별 및 *지역*을 키 필드로 사용하여 이러한 레코드를 통합할 수 있습니다. 그런 다음 *연령*을 모드 **평균**과 통합하고 *판매*를 모드 **합계**와 통합하도록 선택하십시오. 통합 노드 대화 상자에서 **필드에 레코드 수 포함**을 선택하면 통합 출력은 다음 표와 같습니다.

표 14. 통합 레코드 예

연령(평균)	성별	지역	판매(합계)	레코드 수
35.5	F	N	25	4
29	F	S	6	1
34.5	M	N	20	2
33.75	M	S	20	4

이를 통해 예를 들어, 북부 지역에 있는 네 명의 여성 판매 담당자의 평균 연령이 35.5이며 총 판매 합계가 25 단위임을 알 수 있습니다.

참고: 통합 모드가 지정되지 않은 경우 *지점*과 같은 필드를 자동으로 버립니다.

통합 노드의 옵션 설정

통합 노드에서 다음을 지정하십시오.

- 통합의 범주로 사용할 하나 이상의 키 필드
- 통합 값을 계산할 하나 이상의 통합 필드
- 각 통합 필드에 대해 출력할 하나 이상의 통합 모드(통합 유형)

새로 추가된 필드에 사용할 기본 통합 모드를 지정하고 통합을 분류하기 위한 표현식(공식과 비슷함)을 사용할 수 있습니다.

성능 향상을 위해 병렬 처리를 사용하면 통합 작업에 유용할 수 있다는 점을 참고하십시오.

키 필드. 통합의 범주로 사용할 수 있는 필드를 나열합니다. 연속형(숫자) 및 범주형 필드를 모두 키로 사용할 수 있습니다. 두 개 이상의 키 필드를 선택하는 경우 값을 결합하여 레코드를 통합하기 위한 키 값을 생성합니다. 각 고유 키 필드마다 하나의 통합 레코드가 생성됩니다. 예를 들어 *Sex* 및 *Region*이 키 필드인 경우, *N*과 *S*를 갖는 *M*과 *F*의 각각의 고유한 조합(4가지 고유 조합)이 통합 레코드를 갖습니다. 키 필드를 추가하려면 창 오른쪽의 필드 선택기 단추를 사용하십시오.

대화 상자의 나머지는 두 개의 기본 영역인 **기본 통합** 및 **통합 표현식**으로 구분됩니다.

기본 통합

통합 필드. 값이 통합되는 필드와 선택된 통합 모드를 나열합니다. 이 목록에 필드를 추가하려면 오른쪽의 필드 선택기 단추를 사용하십시오. 다음 통합 모드를 사용할 수 있습니다.

참고: 일부 모드는 숫자가 아닌 필드에 적용할 수 없습니다(예: 날짜/시간 필드의 **합계**). 선택된 통합 필드와 함께 사용할 수 없는 모드가 사용 안함으로 설정되어 있습니다.

- **합계.** 각 키 필드 조합의 합계 값을 리턴하려면 이를 선택하십시오. 합계는 결측값이 있는 모든 케이스에서 값의 총계입니다.
- **평균.** 각 키 필드 조합의 평균 값을 리턴하려면 이를 선택하십시오. 평균은 중심 경향의 측도이며 산술 평균(합계를 케이스 수로 나눈 값)입니다.

- **최소값.** 각 키 필드 조합의 최소값을 리턴하려면 이를 선택하십시오.
- **최대값.** 각 키 필드 조합의 최대값을 리턴하려면 이를 선택하십시오.
- **SDev.** 각 키 필드 조합의 표준 편차를 리턴하려면 이를 선택하십시오. 표준 편차는 평균 주변의 산포도이며 분산 측정의 제곱근입니다.
- **중앙값.** 각 키 필드 조합의 중앙값을 리턴하려면 이를 선택하십시오. 중앙값은 벗어난 값의 영향을 받지 않는 중심 경향 측도이며 평균과 달리 상한 극단값 또는 하한 극단값의 영향을 받을 수 있습니다. 50번째 백분위수 또는 두 번째 사분위수라고도 합니다.
- **개수.** 각 키 필드 조합의 널이 아닌 값 수를 리턴하려면 이를 선택하십시오.
- **분산.** 각 키 필드 조합의 분산 값을 리턴하려면 이를 선택하십시오. 분산은 평균 주변의 산포도이며 평균을 케이스 수에서 1을 뺀 값으로 나눈 값의 제곱 편차 합계와 같습니다.
- **첫 번째 사분위수.** 각 키 필드 조합의 첫 번째 사분위수(25번째 백분위수) 값을 리턴하려면 이를 선택하십시오.
- **세 번째 사분위수.** 각 키 필드 조합의 세 번째 사분위수(75번째 백분위수) 값을 리턴하려면 이를 선택하십시오.

참고: 통합 노드가 있는 스트림을 실행하는 경우 SQL을 Oracle 데이터베이스에 푸시백할 때 첫 번째 및 세 번째 사분위수에 대해 리턴되는 값이 원시 모드에서 리턴되는 값과 다를 수 있습니다.

기본 모드. 새로 추가된 필드에 대해 사용할 기본 통합 모드를 지정하십시오. 동일한 통합을 자주 사용하는 경우 여기에서 하나 이상의 모드를 선택하고 오른쪽의 모두에 적용 단추를 사용하여 위에 나열된 모든 필드에 선택된 모드를 적용하십시오.

새 필드 이름 확장자. 통합 필드를 복제하기 위해 접미부 또는 접두부(예: "1" 또는 "new")를 추가하려면 이를 선택하십시오. 예를 들어, 필드 연령에서 최소값 통합의 결과는 접미부 옵션을 선택하고 확장자로 "1"을 지정한 경우 *Age_Min_1*이라는 필드 이름을 생성합니다. 참고: *_Min* 또는 *Max_*와 같은 통합 확장자가 새 필드에 자동으로 추가되어 수행된 통합의 유형을 표시합니다. 원하는 확장자 스타일을 표시하려면 **접미부** 또는 **접두부**를 선택하십시오.

필드에 레코드 수 포함. 기본적으로 *Record_Count*라는 각 출력 레코드에 추가 필드를 포함하려면 이를 선택하십시오. 이 필드는 각 통합 레코드를 구성하기 위해 통합된 입력 레코드 수를 표시합니다. 편집 필드에 입력하여 이 필드의 사용자 정의 이름을 작성하십시오.

참고: 통합이 계산될 때 시스템 널값이 제외되지만 레코드 수에는 포함됩니다. 반면, 공백 값은 통합 및 레코드 수에 모두 포함됩니다. 공백 값을 제외하려면 채움 노드를 사용하여 공백을 널값으로 대체할 수 있습니다. 선택 노드를 사용하여 공백을 제거할 수도 있습니다.

통합 표현식

표현식은 값, 필드 이름, 연산자, 함수에서 작성되는 공식과 비슷합니다. 한 번에 하나의 레코드에서 작동되는 함수와 달리 통합 표현식은 레코드의 그룹, 세트 또는 콜렉션에서 작동됩니다.

참고: 스트림에 데이터베이스 연결이 포함된 경우에만 통합 표현식을 작성할 수 있습니다(데이터베이스 소스 노드 사용).

새 표현식이 파생 필드로 작성됩니다. 표현식을 작성하려면 표현식 작성기에서 사용 가능한 *데이터베이스 통합* 함수를 사용하십시오.

표현식 작성기에 대한 자세한 정보는 IBM SPSS Modeler 사용자 안내서(ModelerUsersGuide.pdf)를 참조하십시오.

통합 표현식이 키 필드에 따라 그룹화되어 있으므로 **키 필드**와 통합 표현식 간 연결이 있다는 점을 참고하십시오.

유효한 통합 표현식은 결과를 통합하기 위해 평가되는 표현식입니다. 유효한 통합 표현식에 대한 두 개의 예와 이를 제어하는 규칙은 다음과 같습니다.

- 스칼라 함수를 사용하면 여러 통합 함수를 함께 결합하여 단일 통합 결과를 생성할 수 있습니다. 예:

  ```
  max(C01) - min(C01)
  ```

- 통합 함수는 여러 스칼라 함수의 결과에서 작동될 수 있습니다. 예:

  ```
  sum (C01*C01)
  ```

최적화 설정 통합

최적화 탭에서 다음을 지정하십시오.

키가 연속적입니다. 동일한 키 값을 갖는 모든 레코드가 입력에서 함께 그룹화됨(예를 들어, 입력이 키 필드에서 정렬됨)을 아는 경우 이 옵션을 선택하십시오. 그렇게 하면 성능이 개선될 수 있습니다.

중앙값 및 사분위수의 근사치 허용. Analytic Server에서 데이터를 처리할 때 주문 통계(중앙값, 첫 번째 사분위수, 세 번째 사분위수)는 현재 지원되지 않습니다. Analytic Server를 사용하는 경우 데이터를 구간화한 후 구간에 대한 분포에 기반하여 통계의 추정값을 계산하여 계산되는 통계 대신 이러한 통계에 대한 근사 값을 사용하려면 이 선택란을 선택하십시오. 기본적으로 이 옵션은 선택되어 있지 않습니다.

구간 수. 중앙값 및 사분위수의 근사치 허용 선택란을 선택하는 경우에만 사용할 수 있습니다. 통계를 추정할 때 사용할 구간 수를 선택하십시오. 구간 수는 **최대 오차 %**에 영향을 줍니다. 기본적으로 구간 수는 1000이며 범위의 0.1 퍼센트의 최대 오차에 해당합니다.

RFM 통합 노드

RFM(최근, 빈도, 구매총액) 통합 노드를 사용하면 고객의 히스토리 트랜잭션 데이터를 사용하고 모든 사용하지 않은 데이터를 제거하고 고유 고객 ID를 키로 사용하여 마지막으로 다룬 시기(최근), 작성한 트랜잭션 수(빈도), 해당 트랜잭션의 총 값(구매총액)을 나열하는 단일 행으로 모든 나머지 트랜잭션 데이터를 결합할 수 있습니다.

통합을 진행하기 전에 특히 결측값에 집중하여 데이터를 정리하는 시간이 필요합니다.

일단 RFM 통합 노드를 사용하여 데이터를 식별하고 변환한 후에는 RFM 분석 노드를 사용하여 추가 분석을 수행할 수 있습니다. 자세한 정보는 177 페이지의 『RFM 분석 노드』의 내용을 참조하십시오.

일단 RFM 통합 노드를 통해 데이터 파일이 실행된 후에는 어떠한 목표 값도 갖지 않습니다. 따라서 이를 C5.0 또는 CHAID 등의 모델링 노드를 사용하는 추가 예측 분석에 대한 입력으로 사용하려면 다른 고객 데이터와 합쳐야 합니다. 예를 들어, 고객 ID 일치 등의 방법이 있습니다. 자세한 정보는 87 페이지의 『합치기 노드』의 내용을 참조하십시오.

IBM SPSS Modeler의 RFM 통합 및 RFM 분석 노드는 독립적 구간화를 사용하기 위해 설정됩니다. 즉, 해당 값 또는 다른 두 측도에 관계없이 RFM(Recency, Frequency, Monetary) 값의 각 측도에 대한 데이터를 순위화하고 구간화합니다.

RFM 통합 노드에 대한 옵션 설정

RFM 통합 노드의 설정 탭은 다음 필드를 포함합니다.

상대적인 최근 계산. 트랜잭션의 최근성이 계산되는 날짜를 지정하십시오. 사용자가 입력하는 **고정 날짜** 또는 시스템에서 설정된 **오늘 날짜**일 수 있습니다. **오늘 날짜**는 기본적으로 입력되며 노드가 실행될 때 자동으로 업데이트됩니다.

연속적 ID. ID가 동일한 모든 레코드를 데이터 스트림에서 함께 표시하도록 데이터를 사전 정렬한 경우 이 옵션을 선택하여 처리 속도를 높입니다. 데이터가 사전 정렬되지 않았거나 정렬 여부가 확실하지 않은 경우 이 옵션을 선택하지 않은 상태로 두면 노드가 데이터를 자동으로 정렬합니다.

ID. 고객 및 고객의 트랜잭션을 식별하기 위해 사용할 필드를 선택하십시오. 선택할 수 있는 필드를 표시하려면 오른쪽의 필드 선택기 단추를 사용하십시오.

날짜. 최근성을 계산하기 위해 사용할 날짜 필드를 선택하십시오. 선택할 수 있는 필드를 표시하려면 오른쪽의 필드 선택기 단추를 사용하십시오.

입력으로 사용하려면 적절한 형식의 날짜 또는 시간소인 저장 공간이 있는 필드가 필요합니다. 예를 들어, 값이 *Jan 2007*, *Feb 2007* 등인 문자열 필드가 있는 경우 이를 채움 노드 및 `to_date()` 함수를 사용하여 날짜 필드로 변환할 수 있습니다. 자세한 정보는 165 페이지의 『채움 노드를 사용한 저장 공간 변환』의 내용을 참조하십시오.

값. 고객 트랜잭션의 구매총액 값을 계산하는 데 사용할 필드를 선택하십시오. 선택할 수 있는 필드를 표시하려면 오른쪽의 필드 선택기 단추를 사용하십시오. *참고*: 숫자 값이어야 합니다.

새 필드 이름 확장자. 접미문자 또는 접두문자를 첨부할 것인지 선택하십시오. 예를 들어, 새로 생성된 최근, 빈도 및 구매총액 필드에 "12_month"를 첨부할 수 있습니다. 선호하는 확장자 유형을 표시하려면 **접미문자** 또는 **접두문자**를 선택하십시오. 예를 들어, 여러 시간 주기를 검사할 때 유용합니다.

아래의 값을 가진 레코드 삭제. 필요한 경우, RFM 총계를 계산할 때 그 아래의 모든 트랜잭션 세부사항을 사용하지 않는 최소값을 지정할 수 있습니다. 값의 단위는 선택된 **Value** 필드와 관계가 있습니다.

최근 트랜잭션만 포함. 큰 데이터베이스를 분석하는 경우, 최근 레코드만 사용되도록 지정할 수 있습니다. 특정 날짜 이후 또는 최근 주기 내에 기록된 데이터만 사용하도록 선택할 수 있습니다.

- **다음 날짜 이후의 트랜잭션.** 그 이후의 레코드가 분석에 포함될 트랜잭션 날짜를 지정하십시오.
- **마지막 이내의 트랜잭션.** 그 이후의 레코드가 분석에 포함될 날짜에 **상대적인 최근 계산**으로부터의 기간(일, 주, 월 또는 년)의 유형과 수를 지정하십시오.

두 번째 가장 최근 트랜잭션 날짜 저장. 각 고객에 대한 두 번째 최근 트랜잭션의 날짜를 알려면 이 선택란을 선택하십시오. 또한 **세 번째 가장 최근 트랜잭션의 날짜** 선택란도 선택할 수 있습니다. 예를 들어, 상당한 정도의 기간 이전에는 많은 트랜잭션을 수행했으나 최근 트랜잭션은 한 번뿐인 고객을 식별하는 데 도움을 줍니다.

정렬 노드

정렬 노드를 사용하여 하나 이상의 필드의 값을 기반으로 레코드를 내림차순 또는 오름차순으로 정렬할 수 있습니다. 예를 들어, 가장 일반적인 데이터 값을 가진 레코드를 보고 선택하기 위해 정렬 노드를 사용하는 경우가 가장 빈번합니다. 일반적으로 먼저 통합 노드를 사용하여 데이터를 통합한 다음 정렬 노드를 사용하여 통합된 데이터를 레코드 개수의 내림차순으로 정렬합니다. 표에 이런 결과를 표시하면 데이터를 탐색하여 최고의 고객 열 명의 레코드를 선택하는 것 등을 포함하여 의사결정을 내릴 수 있습니다.

정렬 노드의 설정 탭은 다음 필드를 포함합니다.

정렬 기준. 정렬 키로 사용하기 위해 선택한 모든 필드가 표에 표시됩니다. 키 필드는 숫자 필드일 때 정렬이 가장 효과적입니다.

- 오른쪽의 필드 선택기 단추를 사용하여 이 목록에 **필드를 추가**하십시오.
- 표의 **정렬** 열에서 **오름차순** 또는 **내림차순** 화살표를 클릭하여 *순서를 선택*하십시오.
- 빨간색 삭제 단추를 사용하여 **필드를 삭제**하십시오.
- 위로 및 아래로 화살표 단추를 사용하여 **지시문을 정렬**하십시오.

기본 정렬 순서. 새 필드가 추가될 때 기본 정렬 순서로 사용할 **오름차순** 또는 **내림차순**을 선택하십시오.

참고: 모델 스트림 아래로 고유 노드가 있으면 정렬 노드가 적용되지 않습니다. 고유 노드에 대한 자세한 정보는 96 페이지의 『고유 노드』의 내용을 참조하십시오.

정렬 최적화 설정

일부 키 필드를 기준으로 하여 이미 정렬된 데이터를 사용하여 작업하는 경우, 이미 정렬된 필드를 지정하여 시스템이 데이터의 나머지를 더 효율적으로 정렬하도록 설정할 수 있습니다. 예를 들어, *연령*(내림차순) 및 *약물*(오름차순)로 정렬하려고 하나 데이터가 이미 *연령*(내림차순)으로 정렬되었음을 알고 있습니다.

데이터가 사전 정렬됨. 데이터가 이미 하나 이상의 필드에 의해 정렬되는지 여부를 지정합니다.

기존 정렬 순서 지정. 이미 정렬된 필드를 지정하십시오. 필드 선택 대화 상자를 사용하여 목록에 필드를 추가하십시오. *순서* 열에서 각 필드가 오름차순 또는 내림차순으로 정렬되는지 여부를 지정하십시오. 여러 필드를 지정하는 경우에는 올바른 정렬 순서로 필드를 나열하는지 확인하십시오. 목록 오른쪽의 화살표를 사용하여 올

바른 순서로 필드를 배열하십시오. 올바른 기존 정렬 순서를 지정하지 못하면 스트림을 실행할 때 오류가 표시되고 사용자가 지정한 정렬과 일치하지 않는 레코드 수가 표시됩니다.

참고: 병렬 처리를 사용할 경우 정렬 속도가 개선될 수 있습니다.

합치기 노드

병합 노드의 기능은 다중 입력 레코드를 가져와서 입력 필드 중 일부 또는 전부가 포함된 단일 출력 레코드를 작성하는 것입니다. 이는 내부 고객 데이터 및 구매한 인구 통계 데이터 등의 다양한 소스의 데이터를 병합하려고 할 때 유용한 조작입니다. 다음의 방법으로 데이터를 병합할 수 있습니다.

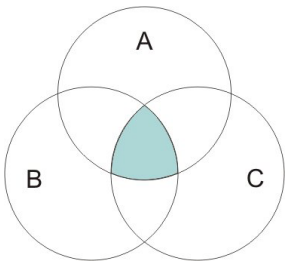
- **순서**별 병합은 가장 작은 데이터 소스가 고갈될 때까지 입력 순서로 모든 소스의 해당 레코드를 연결합니다. 이는 정렬 노드를 사용하여 데이터를 정렬한 이 옵션을 사용하는 경우 중요합니다.
- 한 데이터 소스의 레코드를 다른 데이터 소스의 레코드와 일치시키는 방법을 지정하기 위해 **키** 필드(예: *고객 ID*)를 사용하여 병합. 내부 결합, 전체 외부 결합, 부분 외부 결합, 안티 결합을 포함한 여러 유형의 결합을 사용할 수 있습니다. 자세한 정보는 『결합의 유형』의 내용을 참조하십시오.
- **조건**별 병합은 병합을 수행하기 위해 충족할 조건을 지정할 수 있음을 의미합니다. 노드에서 작업 조건을 지정하거나 표현식 작성기를 사용하여 조건을 작성할 수 있습니다.
- **순위가 매겨진 조건**별 병합은 병합을 수행하기 위해 충족할 조건과 오름차순으로 정렬하는 순위화 표현식을 지정하는 왼쪽 외부 결합입니다. 지리 공간적 데이터를 병합하는 데 가장 자주 사용되며 노드에서 직접 조건을 지정하거나 표현식 작성기를 사용하여 조건을 작성할 수 있습니다.

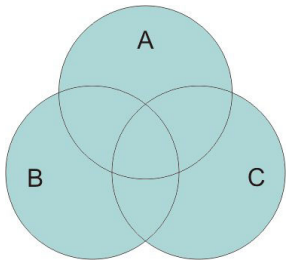
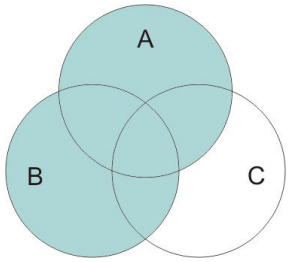
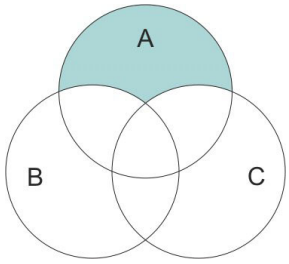
결합의 유형

데이터 병합을 위해 키 필드를 사용하는 경우 제외할 레코드와 포함할 레코드에 대해 잠시 생각해 보는 것이 도움이 됩니다. 다양한 결합이 있으며 이러한 결합에 대해서는 아래에서 자세하게 다룹니다.

결합의 두 가지 기본적인 유형을 내부 결합과 외부 결합이라고 합니다. 이 방법은 *고객 ID* 등의 키 필드의 공통 값을 기반으로 관련 데이터 세트의 테이블을 병합하는 데 자주 사용됩니다. 내부 결합에서는 완전한 레코드만 포함하는 출력 데이터 세트와 정리 병합을 고려합니다. 외부 결합은 병합된 데이터의 완전한 데이터도 포함하지만 하나 이상의 입력 테이블의 고유 데이터를 포함할 수도 있게 합니다.

허용되는 결합의 유형에 대해서는 아래에 더 자세히 설명되어 있습니다.

	**내부 결합**은 키 필드의 값이 모든 입력 테이블에 대해 공통인 레코드만 포함합니다. 즉, 일치하지 않는 레코드는 출력 데이터 세트에 포함되지 않습니다.

	**전체 외부 결합**은 입력 테이블의 모든 레코드(일치하는 레코드와 일치하지 않는 레코드 모두)를 포함합니다. 왼쪽 및 오른쪽 외부 결합은 부분 외부 결합이라고 하며 아래에 설명되어 있습니다.
	**부분 외부 결합**은 키 필드를 사용하는 일치된 모든 레코드와 지정된 테이블의 일치하지 않는 레코드를 포함합니다. (즉, 일부 테이블의 모든 레코드와 다른 테이블의 일치하는 레코드만). 병합 탭의 선택 단추를 사용하여 외부 결합에 포함할 테이블(예: 여기에 표시된 A 및 B)을 선택할 수 있습니다. 부분 결합은 두 개의 테이블만 병합 중인 경우 왼쪽 또는 오른쪽 외부 결합이라고도 합니다. IBM SPSS Modeler는 셋 이상의 테이블의 병합을 허용하므로 이를 부분 외부 결합이라고 합니다.
	**안티 결합**은 첫 번째 입력 테이블(여기에 표시된 테이블 A)에 대해 일치하지 않는 레코드만 포함합니다. 이 유형의 결합은 내부 결합과 반대이며 출력 데이터 세트에 완전한 레코드를 포함하지 않습니다.

예를 들어, 한 데이터 세트에 농장에 대한 정보가 있고 다른 데이터 세트에 농장 관련 보험 클레임에 대한 정보가 있는 경우 병합 옵션을 사용하여 첫 번째 소스의 레코드를 두 번째 소스에 일치시킬 수 있습니다.

농장 표본에 있는 고객이 보험 클레임을 제기했는지 판별하려면 내부 결합 옵션을 사용하여 두 표본에서 모든 ID가 일치하는 위치를 표시하는 목록을 리턴하십시오.

	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...

그림 2. 내부 결합 병합의 표본 출력

전체 외부 결합 옵션을 사용하면 입력 테이블에서 일치하는 레코드와 일치하지 않는 레코드가 모두 리턴됩니다. 불완전한 값에는 시스템 결측값($null$)이 사용됩니다.

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimtype	claimvalu
1	id601	$null$	$null$	$null$	$null$	$null$	$null$	$null$	decomm...	74703.10
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

그림 3. 전체 외부 결합 병합의 표본 출력

부분 외부 결합은 키 필드를 사용하는 일치된 모든 레코드와 지정된 테이블의 일치하지 않는 레코드를 포함합니다. 테이블에는 ID 필드에서 일치된 모든 레코드와 첫 번째 데이터 세트에서 일치된 레코드가 표시됩니다.

	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

그림 4. 부분 외부 결합 병합의 표본 출력

안티 결합 옵션을 사용하는 경우 테이블은 첫 번째 입력 테이블에 대해 일치하지 않는 레코드만 리턴합니다.

	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

그림 5. 안티 결합 병합의 표본 출력

병합 방법 및 키 지정

병합 노드의 병합 탭에는 다음과 같은 필드가 포함되어 있습니다.

병합 방법 레코드 병합에 사용할 방법을 선택하십시오. **키** 또는 **조건**을 선택하면 대화 상자의 아래쪽 절반이 활성화됩니다.

- **순서** 각 입력의 *n*번째 레코드가 병합되어 *n*번째 출력 레코드를 생성하도록 순서별로 레코드를 병합합니다. 레코드에 일치하는 입력 레코드가 부족해지면 더 이상 출력 레코드가 생성되지 않습니다. 이는 작성되는 레코드 수는 가장 작은 데이터 세트의 레코드 수임을 의미합니다.
- **키** *트랜잭션 ID* 등의 키 필드를 사용하여 레코드를 키 필드의 동일한 값과 병합합니다. 이는 데이터베이스 "일치 결합"과 동등합니다. 키 값이 두 번 이상 발생하는 경우에는 가능한 모든 조합이 리턴됩니다. 예를 들어, 동일한 키 필드 값 *A*를 가진 레코드의 다른 필드에 다른 값 *B*, *C* 및 *D*가 포함되어 있는 경우 병합된 필드는 *B* 값을 가진 *A*, *C* 값을 가진 *A* 및 *D* 값을 가진 *A*라는 각각의 조합에 대해 별도의 레코드를 생성합니다.

  참고: 널값은 키별 병합 방법에서 동일한 것으로 간주되지 않으므로 결합되지 않습니다.
- **조건** 병합에 대한 조건을 지정하려면 이 옵션을 사용하십시오. 자세한 정보는 90 페이지의 『병합을 위한 조건 지정』의 내용을 참조하십시오.

- **순위화된 조건** 1차 및 모든 2차 데이터 세트에 있는 각각의 행 쌍을 병합해야 하는지 여부를 지정하려면 이 옵션을 사용하십시오. 순위화 표현식을 사용하여 다중 일치를 오름차순으로 정렬하십시오. 자세한 정보는 91 페이지의 『병합을 위한 순위화된 조건 지정』의 내용을 참조하십시오.

가능한 키 모든 입력 데이터 소스에서 정확하게 일치하는 필드 이름을 가진 필드만 나열합니다. 이 목록에서 필드를 선택한 후 화살표 단추를 사용하여 레코드 병합에 사용되는 키 필드로 추가하십시오. 둘 이상의 키 필드를 사용할 수 있습니다. 필터 노드 또는 소스 노드의 필터 탭을 사용하여 비일치 입력 필드의 이름을 바꿀 수 있습니다.

병합을 위한 키 키 필드의 값을 기반으로 모든 입력 데이터 소스의 레코드를 병합하는 데 사용되는 모든 필드를 나열합니다. 목록에서 키를 제거하려면 하나의 키를 선택한 후 화살표 단추를 사용하여 가능한 키 목록에 리턴하십시오. 둘 이상의 키 필드가 선택되면 아래의 옵션을 사용할 수 있습니다.

중복 키 필드 결합 위에서 둘 이상의 키 필드가 선택되는 경우 이 옵션을 사용하면 해당 이름의 출력 필드가 하나만 있게 됩니다. 이 옵션은 IBM SPSS Modeler의 이전 버전에서 스트림을 가져온 경우를 제외하고 기본적으로 사용으로 설정됩니다. 이 옵션이 사용 안함으로 설정되는 경우에는 합치기 노드 대화 상자의 필터 탭을 사용하여 중복 키 필드의 이름을 바꾸거나 중복 키 필드를 제외해야 합니다.

일치 레코드만 포함(내부 결합) 완전한 레코드만 병합하려면 선택하십시오.

일치 및 비일치 레코드 포함(전체 외부 결합) "전체 외부 결합"을 수행하려면 선택하십시오. 이는 키 필드의 값이 모든 입력 테이블에 없는 경우 불완전한 레코드가 계속 보존됨을 의미합니다. 정의되지 않은 값($null$)이 키 필드에 추가되고 출력 레코드에 포함됩니다.

일치 및 선택된 비일치 레코드 포함(부분 외부 결합) 하위 대화 상자에서 선택하는 테이블의 "부분 외부 결합"을 수행하려면 선택하십시오. **선택**을 클릭하여 병합 시 불완전한 레코드를 보존할 테이블을 지정하십시오.

다른 레코와 일치하지 않는 첫 번째 데이터 세트의 레코드 포함(안티 결합) 첫 번째 데이터 세트의 일치하지 않는 레코드만 다운스트림으로 전달되는 "안티 결합"의 한 유형을 수행하려면 선택하십시오. 입력 탭의 화살표를 사용하여 입력 데이터 세트의 순서를 지정할 수 있습니다. 이 유형의 결합에서는 완전한 레코드를 출력 데이터 세트에 포함하지 않습니다. 자세한 정보는 87 페이지의 『결합의 유형』의 내용을 참조하십시오.

부분 결합에 대한 데이터 선택

부분 외부 결합의 경우 불완전한 레코드를 보유할 테이블을 선택해야 합니다. 예를 들어, 주택 담보 대출 테이블의 일치된 레코드만 보유하면서 고객 테이블의 모든 레코드를 보유하길 원할 수 있습니다.

외부 결합 열. *외부 결합* 열에서 포함할 데이터 세트를 전부 선택하십시오. 부분 결합의 경우 겹치는 레코드뿐만 아니라 여기서 선택한 데이터 세트에 대한 불완전한 레코드도 보유됩니다. 자세한 정보는 87 페이지의 『결합의 유형』의 내용을 참조하십시오.

병합을 위한 조건 지정

병합 방법을 **조건**으로 설정하여 병합을 수행하기 위해 충족해야 하는 하나 이상의 조건을 지정할 수 있습니다.

조건 필드에 직접 조건을 입력하거나 필드 오른쪽의 계산기 기호를 클릭하여 표현식 작성기를 통해 조건을 작성할 수 있습니다.

병합 충돌을 피하기 위해 중복 필드 이름 태그 추가 병합할 둘 이상의 데이터 세트에 동일한 필드 이름이 포함되어 있는 경우에는 이 선택란을 선택하여 필드 열 헤더의 시작 부분에 다른 접두부 태그를 추가하십시오. 예를 들어, *Name*이라는 필드가 두 개 있으면 병합 결과에는 *1_Name*과 *2_Name*이 포함됩니다. 데이터 소스에서 태그의 이름이 바뀐 경우에는 번호가 지정된 접두부 태그 대신 새 이름이 사용됩니다. 이 선택란을 선택하지 않은 경우 데이터에 중복 이름이 있으면 선택란 오른쪽에 경고가 표시됩니다.

병합을 위한 순위화된 조건 지정

순위화된 조건 병합은 조건별 왼쪽 외부 결합 병합으로 간주될 수 있습니다. 병합의 왼쪽은 각각의 레코드가 이벤트인 1차 데이터 세트입니다. 예를 들어, 범죄 데이터에서 패턴을 찾는 데 사용되는 모델에서 1차 데이터 세트의 각 레코드는 범죄 및 해당 연관된 정보(위치, 유형 등)가 됩니다. 이 예제에서 오른쪽에는 관련 지리 공간적 데이터 세트가 포함되어 있습니다.

이 병합에서는 병합 조건과 순위화 표현식을 모두 사용합니다. 병합 조건은 *within* 또는 *close_to* 등의 지리 공간적 함수를 사용할 수 있습니다. 병합을 수행하는 동안 오른쪽 데이터 세트의 모든 필드가 왼쪽 데이터 세트에 추가되지만 다중 일치가 있으면 목록 필드가 생성됩니다. 예:

- 왼쪽: Crime data
- 오른쪽: Counties 데이터 세트 및 roads 데이터 세트
- 병합 조건: Crime data *within* counties and *close_to* roads, along with a definition of what counts as *close_to*.

이 예제에서 범죄가 3개 도로의 요구된 *close_to* 거리 내에서 발생한 경우(또한 리턴될 일치 수가 3 이상으로 설정된 경우) 3개 도로 모두 목록 항목으로 리턴됩니다.

병합 방법을 **순위화된 조건**으로 설정하여 병합을 수행하기 위해 충족할 하나 이상의 조건을 지정할 수 있습니다.

1차 데이터 세트 병할할 1차 데이터 세트를 선택하십시오. 선택하는 데이터 세트에 다른 모든 데이터 세트의 필드가 추가됩니다. 이것은 외부 결합 병합의 왼쪽으로 간주될 수 있습니다.

1차 데이터 세트를 선택하면 병합 노드에 연결되는 다른 모든 입력 데이터 세트가 자동으로 **병합** 테이블에 나열됩니다.

병합 충돌을 피하기 위해 중복 필드 이름에 태그 추가 병합할 둘 이상의 데이터 세트에 동일한 필드 이름이 포함되어 있는 경우에는 이 선택란을 선택하여 필드 열 헤더의 시작 부분에 다른 접두부 태그를 추가하십시오. 예를 들어, *Name*이라는 필드가 두 개 있으면 병합 결과에는 *1_Name*과 *2_Name*이 포함됩니다. 데이터 소스에서 태그의 이름이 바뀌는 경우에는 번호가 지정된 접두부 태그 대신 새 이름이 사용됩니다. 이 선택란을 선택하지 않은 경우 데이터에 중복 이름이 있으면 선택란 오른쪽에 경고가 표시됩니다.

병합

데이터 세트

병합 노드에 대해 입력으로 연결되는 2차 데이터 세트의 이름을 표시합니다. 기본적으로 둘 이상의 2차 데이터 세트가 있으면 병합 노드에 연결된 순서대로 해당 데이터 세트가 나열됩니다.

병합 조건

1차 데이터 세트를 가진 테이블의 각 데이터 세트를 병합하는 데 필요한 고유 조건을 입력하십시오. 셀에 직접 조건을 입력하거나 셀 오른쪽에 있는 계산기 기호를 클릭하여 표현식 작성기를 통해 조건을 작성할 수 있습니다. 예를 들어, 지리 공간적 술어를 사용하여 다른 데이터 세트의 구/군 데이터 내에 한 데이터 세트의 범죄 데이터를 배치하는 병합 조건을 작성할 수 있습니다. 기본 병합 조건은 아래 목록에 표시된 대로 지리 공간적 측정 수준에 따라 다릅니다.

- 점, 선 스트링, 다중 점, 다중 선 스트링 - *close_to*의 기본 조건입니다.
- 다각형, 다중 다각형 - *within*의 기본 조건입니다.

이 수준에 대한 자세한 정보는 142 페이지의 『지리 공간적 측정 수준』의 내용을 참조하십시오.

데이터 세트에 다양한 유형의 여러 지리 공간적 필드가 포함되어 있으면 사용되는 기본 조건은 다음 내림차순으로 데이터에서 발견되는 첫 번째 측정 수준에 의해 결정됩니다.

- 점
- 선 스트링
- 다각형

참고: 기본값은 2차 데이터베이스에 지리 공간적 데이터 필드가 있는 경우에만 사용할 수 있습니다.

순위화 표현식

데이터 세트의 병합을 순위화하는 표현식을 지정하십시오. 이 표현식은 순위화 기준을 기반으로 하는 순서로 다중 일치를 정렬하는 데 사용됩니다. 셀에 직접 조건을 입력하거나 셀 오른쪽에 있는 계산기 기호를 클릭하여 표현식 작성기를 통해 조건을 작성할 수 있습니다.

거리 및 영역의 기본 순위화 표현식이 표현식 작성기에서 제공되며 둘 다 오름차순으로 순위화됩니다 (예를 들어, 거리에 대한 맨 위 일치가 가장 작은 값임을 의미함). 거리별 순위화의 예제는 1차 데이터 세트에 범죄 및 해당 연관된 위치가 포함되어 있고 각각의 다른 데이터 세트에 위치를 가진 범죄 대상이 포함되어 있는 경우입니다. 이 경우 범죄와 범죄 대상 사이의 거리를 순위화 기준으로 사용할 수 있습니다. 기본 순위화 표현식은 아래 목록에 표시된 대로 지리 공간적 측정 수준에 따라 다릅니다.

- 점, 선 스트링, 다중 점, 다중 선 스트링 - 기본 표현식은 *distance*입니다.
- 다각형, 다중 다각형 - 기본 표현식은 *area*입니다.

참고: 기본값은 2차 데이터베이스에 지리 공간적 데이터 필드가 있는 경우에만 사용할 수 있습니다.

일치 수

조건 및 순위화 표현식을 기반으로 리턴되는 일치의 수를 지정하십시오. 기본 일치 수는 아래 목록에 표시된 대로 2차 데이터 세트의 지리 공간적 측정 수준에 따라 다릅니다. 하지만 셀을 두 번 클릭하여 최대 100까지 자체 값을 입력할 수 있습니다.

- 점, 선 스트링, 다중 점, 다중 선 스트링 - 기본값 3
- 다각형, 다중 다각형 - 기본값 1
- 지리 공간적 필드가 포함되어 있지 않은 데이터 세트 - 기본값 1

예를 들어, **병합 조건** *close_to* 및 **순위화 표현식** *distance*를 기반으로 하는 병합을 설정하면 1차 데이터 세트의 각 레코드에 대한 2차 데이터 세트의 상위 3개(가장 가까운) 일치가 결과 목록 필드에 값으로 리턴됩니다.

병합 노드의 필드 필터링

병합 노드에는 여러 데이터 소스를 병합한 결과로 중복 필드를 필터링하거나 이름을 바꾸는 편리한 방법이 포함되어 있습니다. 대화 상자의 **필터** 탭을 클릭하여 필터링 옵션을 선택하십시오.

여기서 제공되는 옵션은 필터 노드의 옵션과 거의 동일합니다. 하지만 필터 메뉴에서는 여기서 다루지 않은 추가적인 옵션을 사용할 수 있습니다. 자세한 정보는 153 페이지의 『필드 필터링 또는 이름 바꾸기』의 내용을 참조하십시오.

필드. 현재 연결된 데이터 소스의 입력 필드를 표시합니다.

태그. 데이터 소스 링크와 연관된 태그 이름(또는 번호)을 나열합니다. **입력** 탭을 클릭하여 이 병합 노드에 대한 활성 링크를 변경하십시오.

소스 노드. 데이터를 병합 중인 소스 노드를 표시합니다.

연결된 노드. 병합 노드에 연결된 노드의 노드 이름을 표시합니다. 복잡한 데이터 마이닝을 사용하려면 동일한 소스 노드를 포함할 수 있는 여러 병합 또는 추가 조작이 필요할 수 있습니다. 연결된 노드 이름은 이를 구별하는 방법을 제공합니다.

필터. 입력 필드와 출력 필드 간 현재 연결을 표시합니다. 활성 연결은 중단되지 않은 화살표를 표시합니다. 빨간색 X가 있는 연결은 필터링된 필드를 표시합니다.

필드. 병합 또는 추가 후 출력 필드를 나열합니다. 중복 필드는 빨간색으로 표시됩니다. 위의 필터 필드를 클릭하여 중복 필드를 사용 안함으로 설정하십시오.

현재 필드 보기. 키 필드로 사용하기 위해 선택한 필드에 대한 정보를 보려면 선택하십시오.

사용하지 않은 필드 설정 보기. 현재 사용되고 있지 않은 필드에 대한 정보를 보려면 선택하십시오.

입력 순서 및 태그 지정 설정

합치기 및 붙여쓰기 노드 대화 상자에서 입력 탭을 사용하면 입력 데이터 소스의 순서를 지정하고 각 소스에 대한 태그 이름을 변경할 수 있습니다.

태그 및 입력 데이터 세트의 순서. 완전한 레코드만 합치거나 붙여쓰려면 선택하십시오.

- **태그.** 각 입력 데이터 소스에 대한 현재 태그 이름을 나열합니다. 태그 이름 또는 **tags**는 합치기 또는 붙여쓰기 작업에 대한 데이터 링크를 고유하게 식별하는 방법입니다. 예를 들어, 여러 파이프의 물이 한 지점에

서 합쳐져서 단일 파이프를 통해 흘러가는 것을 상상해 보십시오. IBM SPSS Modeler의 데이터도 유사하게 흘러가고 합치는 포인트는 종종 다양한 데이터 소스 사이의 복합 상호작용이 됩니다. 태그는 노드가 저장되거나 연결 해제될 때 링크를 유지하고 쉽게 인식할 수 있도록 합치기 또는 붙여쓰기 노드에 대한 입력("파이프")을 관리하는 방법을 제공합니다.

추가 데이터 소스를 합치기 또는 붙여쓰기 노드에 연결할 때 사용자가 노드를 연결한 순서를 표시하기 위해 자동으로 숫자를 사용하여 기본 태그가 작성됩니다. 이 순서는 입력 또는 출력 데이터 세트의 필드 순서와는 연관이 없습니다. *태그* 열에 새 이름을 입력하여 기본값을 변경할 수 있습니다.

- **소스 노드.** 데이터가 결합되는 소스 노드를 표시합니다.
- **연결된 노드.** 합치기 또는 붙여쓰기 노드에 연결된 노드의 노드 이름을 표시합니다. 복합 데이터 마이닝의 경우, 동일한 소스 노드를 포함하는 여러 합치기 작업이 필요한 경우가 종종 있습니다. 연결된 노드 이름은 이를 구분하는 방법을 제공합니다.
- **필드.** 각 데이터 소스 내의 필드 수를 나열합니다.

현재 태그 보기. 합치기 또는 붙여쓰기 노드에 의해 활성 상태로 사용되는 태그를 보려면 선택하십시오. 즉, 현재 태그는 데이터가 플로우되는 노드에 대한 링크를 식별합니다. 파이프 비유를 사용하자면, 현재 태그는 기존 물이 흘러가고 있는 파이프와 유사합니다.

사용되지 않은 태그 설정 보기. 이전에는 합치기 또는 붙여쓰기 노드에 연결되는 데 사용되었으나 현재는 데이터 소스와 연결되어 있지 않은 태그 또는 링크를 보려면 선택하십시오. 이는 배수 시스템 내에 아직 그대로 있는 비어 있는 파이프와 유사합니다. 이러한 "파이프"를 새 소스에 연결하거나 제거할 수 있습니다. 노드에서 사용되지 않은 태그를 제거하려면 **지우기**를 클릭하십시오. 그러면 사용되지 않은 모든 태그가 한 번에 선택 취소됩니다.

병합 최적화 설정

시스템은 특정 상황에서 더 효율적으로 데이터를 병합할 수 있는 두 가지 옵션을 제공합니다. 이 옵션을 사용하면 한 입력 데이터 세트가 다른 데이터 세트보다 상당히 *크거나* 병합에 사용하는 키 필드 중 일부 또는 전부를 기준으로 데이터가 이미 정렬되어 있는 경우 병합을 최적화할 수 있습니다.

참고: 이 탭에서의 최적화는 IBM SPSS Modeler 원시 노드 실행에만 적용됩니다. 즉, 병합 노드가 SQL로 푸시백되지 않습니다. 최적화 설정은 SQL 생성에 영향을 미치지 않습니다.

한 입력 데이터 세트가 상대적으로 큼. 입력 데이터 세트 중 하나가 다른 입력 데이터 세트보다 훨씬 큼을 나타내려면 선택하십시오. 시스템은 메모리에서 더 작은 데이터 세트를 캐싱한 후 큰 데이터 세트는 캐싱 또는 정렬하지 않고 처리하여 병합을 수행합니다. 공유 데이터의 큰 중심 테이블이 있는 경우(예: 트랜잭션 데이터에) 스타 스키마 또는 비슷한 디자인을 사용하여 설계된 이 유형의 데이터 결합을 일반적으로 사용합니다. 이 옵션을 선택하는 경우에는 **선택**을 클릭하여 큰 데이터 세트를 지정하십시오. 큰 데이터 세트는 *하나*만 선택할 수 있습니다. 다음 표에는 이 방법을 사용하여 최적화할 수 있는 결합이 요약되어 있습니다.

표 15. 결합 최적화 요약.

결합 유형	큰 입력 데이터 세트에 대해 최적화할 수 있는지 여부
내부	예

표 15. 결합 최적화 요약 (계속).

결합 유형	큰 입력 데이터 세트에 대해 최적화할 수 있는지 여부
부분	예(큰 데이터 세트에 불완전한 레코드가 없는 경우)
전체	아니오
안티 결합	예(큰 데이터 세트가 첫 번째 입력인 경우)

모든 입력이 이미 키 필드별로 정렬되어 있음. 병합을 위해 사용하는 키 필드 중 하나 이상을 기준으로 입력 데이터가 이미 정렬되어 있음을 나타내려면 선택하십시오. 모든 입력 데이터 세트가 정렬되어 있는지 확인하십시오.

기존 정렬 순서 지정. 이미 정렬된 필드를 지정하십시오. 필드 선택 대화 상자를 사용하여 필드를 목록에 추가하십시오. 병합 탭에서 지정된 병합에 사용 중인 키 필드 중에서만 선택할 수 있습니다. *순서* 열에서 각 필드가 오름차순 또는 내림차순으로 정렬되는지 여부를 지정하십시오. 여러 필드를 지정하는 경우에는 올바른 정렬 순서로 필드를 나열하는지 확인하십시오. 목록 오른쪽의 화살표를 사용하여 올바른 순서로 필드를 배열하십시오. 올바른 기존 정렬 순서 지정 시 실수가 있는 경우에는 스트림을 실행할 때 오류가 표시되어 정렬이 지정된 사항과 불일치하는 레코드 번호를 표시합니다.

데이터베이스에서 사용하는 데이터 정렬 방법의 대소문자 구분 여부에 따라 하나 이상의 입력이 데이터베이스에 의해 정렬된 경우 최적화가 올바르게 작동하지 않을 수 있습니다. 예를 들어, 두 개의 입력이 있는데 하나는 대소문자를 구분하고 다른 하나는 대소문자를 구분하지 않는 경우에는 정렬 결과가 다를 수 있습니다. 병합 최적화를 수행하면 정렬된 순서를 사용하여 레코드가 처리됩니다. 결과적으로 서로 다른 데이터 정렬 방법을 사용하여 입력을 정렬하면 병합 노드에서 오류를 보고하고 정렬이 불일치하는 레코드 번호를 표시합니다. 모든 입력이 한 소스에서 제공되거나 서로 포함하는 데이터 배열을 사용하여 정렬되는 경우에는 레코드가 성공적으로 병합될 수 있습니다.

참고: 병렬 처리를 사용하면 병합 속도가 향상될 수 있습니다.

붙여쓰기 노드

붙여쓰기 노드를 사용하면 레코드 세트를 연결할 수 있습니다. 다른 소스의 레코드를 함께 결합하는 병합 노드와 달리 붙여쓰기 노드는 한 소스에서 모든 레코드를 더 이상 없을 때까지 다운스트림으로 읽고 전달합니다. 이후에 다음 소스의 레코드는 첫 번째 또는 기본 입력과 동일한 데이터 구조(레코드 수, 필드 수 등)를 사용하여 읽습니다. 기본 소스에 다른 입력 소스보다 많은 필드가 있는 경우 시스템 널 문자열(`$null$`)이 불완전한 값에 사용됩니다.

붙여쓰기 노드는 구조가 비슷하지만 다른 데이터가 있는 데이터 세트를 결합하는 데 유용합니다. 예를 들어, 3월의 판매 데이터 파일 및 4월의 개별 파일과 같이 다른 시간 중 다른 파일에 트랜잭션 데이터를 저장했을 수 있습니다. 구조가 동일하다고 가정할 때(동일한 순서의 동일한 필드) 붙여쓰기 노드는 이러한 파일을 분석할 수 있도록 하나의 큰 파일에 결합할 수 있습니다.

참고: 파일을 붙여쓰려면 필드 측정 수준이 비슷해야 합니다. 예를 들어, *명목* 필드는 측정 수준이 *연속형*인 필드와 함께 붙여쓸 수 없습니다.

붙여쓰기 옵션 설정

필드 일치 기준. 매치하는 필드를 붙여쓸 때 사용할 방법을 선택하십시오.

- **위치.** 기본 데이터 소스에서 필드의 위치에 기반하여 데이터 세트를 추가하려면 이를 선택하십시오. 이 방법을 사용할 때 적절하게 붙여쓰기 위해 데이터를 정렬해야 합니다.
- **이름.** 입력 데이터 세트에서 필드의 이름에 기반하여 데이터 세트를 붙여쓰려면 이를 선택하십시오. 필드 이름을 일치시킬 때 대소문자를 감지하도록 하려면 **대소문자 구분**도 선택하십시오.

출력 필드. 붙여쓰기 노드에 연결되는 소스 노드를 나열합니다. 목록의 첫 번째 노드는 기본 입력 소스입니다. 열 표제를 클릭하여 화면에서 필드를 정렬할 수 있습니다. 이 정렬은 데이터 세트에서 필드를 실제로 다시 정렬하지 않습니다.

필드 포함. 기본 데이터 세트의 필드에 기반하여 출력 필드를 생성하려면 **기본 데이터 세트만**을 선택하십시오. 기본 데이터 세트는 입력 탭에 지정된 첫 번째 입력입니다. 모든 입력 데이터 세트에서 일치하는 필드가 있는지 여부에 관계없이 모든 데이터 세트에서 모든 필드의 출력 필드를 생성하려면 **모든 데이터 세트**를 선택하십시오.

필드에 소스 데이터 세트를 포함하여 레코드 태그 지정. 해당 값이 각 레코드의 소스 데이터 세트를 표시하는 출력 파일에 추가 필드를 추가하려면 선택하십시오. 텍스트 필드에 이름을 지정하십시오. 기본 필드 이름은 *Input* 입니다.

고유 노드

데이터 마이닝을 시작하려면 먼저 데이터 세트의 중복 레코드를 제거해야 합니다. 예를 들어, 마케팅 데이터베이스에서 주소 또는 회사 정보가 다른 개인이 여러 번 표시될 수 있습니다. 고유 노드를 사용하여 데이터에서 중복 레코드를 찾거나 제거하거나 중복 레코드 그룹에서 하나의 복합 레코드를 작성할 수 있습니다.

고유 노드를 사용하려면 먼저 두 레코드가 중복으로 간주되는 경우를 판별하는 키 필드 세트를 정의해야 합니다.

모든 필드를 키 필드로 선택하지 않는 경우 두 개의 "중복" 레코드는 나머지 필드의 값에 여전히 차이가 있을 수 있기 때문에 진정으로 동일할 수 없습니다. 이 경우에는 각각의 중복 레코드 그룹 내에서 적용되는 정렬 순서도 정의할 수 있습니다. 이 정렬 순서를 통해 그룹에서 첫 번째로 처리되는 레코드를 미세 제어할 수 있습니다. 그렇지 않으면 모든 중복이 교환 가능한 것으로 간주되어 모든 레코드를 선택할 수 있습니다. 레코드의 수신 순서는 고려되지 않으므로 업스트림 정렬 노드를 사용해도 도움이 되지 않습니다(아래의 "고유 노드 내에서 레코드 정렬" 참조).

모드. 복합 레코드를 작성할지 아니면 첫 번째 레코드를 포함 또는 제외(삭제)할지 지정하십시오.

- **각 그룹에 대해 복합 레코드 작성.** 숫자가 아닌 필드를 통합하는 방법을 제공합니다. 이 옵션을 선택하면 복합 레코드 작성 방법을 지정하는 복합 탭을 사용할 수 있습니다. 자세한 정보는 99 페이지의 『고유 복합 설정』을 참조하십시오.

- **각 그룹의 첫 번째 레코드만 포함.** 중복 레코드 그룹 각각의 첫 번째 레코드를 선택하고 나머지는 삭제합니다. *첫 번째* 레코드는 레코드의 수신 순서가 아니라 아래에 정의된 정렬 순서에 의해 결정됩니다.
- **각 그룹의 첫 번째 레코드만 삭제.** 중복 레코드 그룹 각각의 첫 번째 레코드를 삭제하는 대신 나머지는 선택합니다. *첫 번째* 레코드는 레코드의 수신 순서가 아니라 아래에 정의된 정렬 순서에 의해 결정됩니다. 이 옵션은 스트림에서 나중에 중복을 검사할 수 있도록 데이터에서 중복을 찾는 경우 유용합니다.

그룹화를 위한 키 필드. 레코드가 동일한지 판별하는 데 사용되는 필드를 나열합니다. 다음을 수행할 수 있습니다.

- 오른쪽의 필드 선택도구 단추를 사용하여 이 목록에 필드를 추가하십시오.
- 빨간색 X(제거) 단추를 사용하여 목록에서 필드를 삭제하십시오.

그룹 내에서 레코드 정렬 기준. 각 중복 그룹 내에서 레코드가 정렬되는 방식과 해당 레코드가 오름차순과 내림차순 중 어느 순서로 정렬되는지를 결정하는 데 사용되는 필드를 나열합니다. 다음을 수행할 수 있습니다.

- 오른쪽의 필드 선택도구 단추를 사용하여 이 목록에 필드를 추가하십시오.
- 빨간색 X(제거) 단추를 사용하여 목록에서 필드를 삭제하십시오.
- 위로 또는 아래로 단추를 사용하여 필드를 이동하십시오(둘 이상의 필드를 기준으로 정렬하는 경우).

각 그룹의 첫 번째 레코드를 포함하거나 제외하도록 선택한 경우 첫 번째로 처리되는 레코드가 사용자에게 중요하면 정렬 순서를 지정해야 합니다.

복합 탭의 특정 옵션에 대해 복합 레코드를 작성하도록 선택한 경우에도 정렬 순서를 지정할 수 있습니다. 자세한 정보는 99 페이지의 『고유 복합 설정』을 참조하십시오.

기본 정렬 순서. 기본적으로 레코드가 정렬 키 값의 **오름차순**과 **내림차순** 중 어느 순서로 정렬되는지를 지정하십시오.

고유 노드 내에서 레코드 정렬

중복 그룹 내 레코드의 순서가 사용자에게 중요한 경우에는 고유 노드에서 **그룹 내, 레코드 정렬 기준** 옵션을 사용하여 순서를 지정해야 합니다. 업스트림 정렬 노드에 의존하지 마십시오. 레코드의 수신 순서는 고려되지 않으며 노드 내에서 지정된 순서만 고려된다는 점을 기억하십시오.

정렬 필드를 지정하지 않는 경우(또는 충분하지 않은 정렬 필드를 지정하는 경우) 각 중복 그룹 내 레코드는 정렬되지 않거나 불완전하게 정렬되므로 결과를 예측할 수 없습니다.

예를 들어, 다수의 머신에 관한 매우 큰 로그 레코드 세트가 있다고 가정해 봅니다. 이 로그에는 다음과 같은 데이터가 포함되어 있습니다.

표 16. 머신 로그 데이터

Timestamp	Machine	Temperature
17:00:22	Machine A	31
13:11:30	Machine B	26
16:49:59	Machine A	30

표 16. 머신 로그 데이터 (계속)

Timestamp	Machine	Temperature
18:06:30	Machine X	32
16:17:33	Machine A	29
19:59:04	Machine C	35
19:20:55	Machine Y	34
15:36:14	Machine X	28
12:30:41	Machine Y	25
14:45:49	Machine C	27
19:42:00	Machine B	34
20:51:09	Machine Y	36
19:07:23	Machine X	33

레코드 수를 각 머신에 대한 최신 레코드까지 줄이려면 Machine을 키 필드로 사용하고 Timestamp를 정렬 필드(내림차순)로 사용하십시오. 정렬 선택사항은 지정된 머신에 대한 다수의 행 중 리턴될 행을 지정하므로 입력 순서는 결과에 영향을 미치지 않으며 최종 데이터 출력은 다음과 같습니다.

표 17. 정렬된 머신 로그 데이터

Timestamp	Machine	Temperature
17:00:22	Machine A	31
19:42:00	Machine B	34
19:59:04	Machine C	35
19:07:23	Machine X	33
20:51:09	Machine Y	36

고유 최적화 설정

작업 중인 데이터가 적은 수의 레코드만 포함하고 있거나 이미 정렬된 경우에는 IBM SPSS Modeler가 데이터를 더 효율적으로 처리할 수 있도록 처리 방식을 최적화할 수 있습니다.

참고: **입력 데이터 세트에 적은 수의 고유 키가 있음**을 선택하거나 노드에 대해 SQL 생성을 사용하는 경우 고유 키 값 내의 행이 리턴될 수 있습니다. 고유 키 내부에 리턴되는 행을 제어하려면 설정 탭의 **그룹 내, 레코드 정렬 기준** 필드를 사용하여 정렬 순서를 지정해야 합니다. 설정 탭에서 정렬 순서를 지정한 경우에는 최적화 옵션이 고유 노드에 의한 결과 출력에 영향을 미치지 않습니다.

입력 데이터 세트에 적은 수의 고유 키가 있음. 적은 수의 레코드, 적은 수의 키 필드 고유 값 또는 둘 다를 가진 경우 이 옵션을 선택하십시오. 그렇게 하면 성능이 개선될 수 있습니다.

설정 탭에서 필드를 정렬하거나 필드를 그룹화하여 입력 데이터 세트가 이미 정렬되어 있음. 설정 탭의 **그룹 내, 레코드 정렬 기준**에 나열되는 모든 필드를 기준으로 데이터가 이미 정렬되어 있거나 데이터의 오름차순 정렬 순서와 내림차순 정렬 순서가 동일한 경우에만 이 옵션을 선택하십시오. 그렇게 하면 성능이 개선될 수 있습니다.

SQL 생성 사용 안함. 노드에 대한 SQL 생성을 사용 안함으로 설정하려면 이 옵션을 선택하십시오.

고유 복합 설정

작업 중인 데이터에 예를 들어, 동일한 사용자에 대한 다중 레코드가 포함되어 있는 경우 처리할 하나의 복합(또는 통합) 레코드를 작성하여 데이터 처리 방식을 최적화할 수 있습니다. IBM SPSS Modeler Entity Analytics가 설치되어 있는 경우에는 이를 사용하여 SPSS Entity Analytics에서 출력되는 중복 레코드를 결합(또는 단일 배열로 변환)할 수도 있습니다.

참고: 이 탭은 설정 탭에서 **각 그룹에 대해 복합 레코드 작성**을 선택하는 경우에만 사용할 수 있습니다.

예를 들어, SPSS Entity Analytics가 다음 표와 같이 3개의 레코드를 동일한 엔티티인 것으로 표시한다고 가정해 봅니다.

표 18. 동일한 엔티티에 대한 다중 레코드의 예.

$EA-ID	이름	연령	은행	최종 학력	총 부채
0003	Bob Jones	27	K	School	27000
0003	Robert Jones	35	N	Degree	42000
0003	Robbie Jones	27	D	PhD	7000

우리가 원하는 것은 이 3개의 레코드를 다운스트림을 사용하는 단일 레코드로 통합하는 것입니다. 통합 노드를 사용하여 총 부채를 합산하고 평균 연령을 계산할 수 있지만 이름, 은행 등의 세부사항에서 평균을 계산할 수는 없습니다. 복합 레코드를 작성하는 데 사용할 세부사항을 지정하면 단일 레코드를 파생시킬 수 있습니다.

이 표에서 다음과 같은 세부사항을 선택하여 복합 레코드를 작성할 수 있습니다.

- **이름**에 대해 첫 번째 레코드 사용
- **연령**에 대해 최고값 사용
- **은행**에 대해 구분자 없이 모든 값 연결
- **최종 학력**에 대해 목록에서 처음 발견되는 항목 사용(PhD Degree School)
- **부채**에 대해 총계 사용

이 세부사항을 결합(또는 통합)하면 다음과 같은 세부사항이 포함된 단일 복합 레코드가 생성됩니다.

- 이름: Bob Jones
- 연령: 35
- 은행: KND
- 최종 학력: PhD
- 부채: 76000

여기서는 셋 이상의 알려진 은행 계좌를 보유하고 있고 총 부채가 많은 35세 이상의 PhD 학위를 가진 Bob Jones라는 최적의 결과를 제공합니다.

복합 탭에 대한 옵션 설정

필드. 이 열에는 데이터 모델의 키 필드를 제외한 모든 필드가 기본 정렬 순서로 표시됩니다(노드가 연결되어 있지 않으면 필드가 표시되지 않음). 필드 이름을 기준으로 알파벳순으로 행을 정렬하려면 열 헤더를 클릭하십시오. Shift+클릭 또는 Ctrl+클릭을 사용하여 둘 이상의 행을 선택할 수 있습니다. 또한 필드를 마우스 오른쪽 단추로 클릭하면 모든 행을 표시하거나 오름차순 또는 내림차순 필드 이름 또는 값을 기준으로 행을 정렬하거나 측도 또는 저장 유형별로 필드를 선택하거나 값을 선택하여 동일한 **다음을 기반으로 값 채우기** 항목을 선택된 모든 행에 자동으로 추가하도록 선택할 수 있는 메뉴가 표시됩니다.

다음을 기반으로 값 채우기. 필드에 대한 복합 레코드에 사용할 값 유형을 선택하십시오. 사용 가능한 옵션은 필드 유형에 따라 다릅니다.

- 숫자 범위 필드의 경우 다음 중에서 선택할 수 있습니다.
  - 그룹에서 첫 번째 레코드
  - 그룹에서 마지막 레코드
  - 총계
  - 평균
  - 최소값
  - 최대
  - 사용자 정의
- 시간 또는 날짜 필드의 경우 다음 중에서 선택할 수 있습니다.
  - 그룹에서 첫 번째 레코드
  - 그룹에서 마지막 레코드
  - 최초
  - 최근
  - 사용자 정의
- 문자열 또는 유형 없는 필드의 경우 다음 중에서 선택할 수 있습니다.
  - 그룹에서 첫 번째 레코드
  - 그룹에서 마지막 레코드
  - 첫 번째 영숫자
  - 마지막 영숫자
  - 사용자 정의

각각의 경우에 **사용자 정의** 옵션을 사용하여 복합 레코드를 채우는 데 사용되는 값에 대한 제어를 향상시킬 수 있습니다. 자세한 정보는 101 페이지의 『고유 복합 - 사용자 정의 탭』을 참조하십시오.

필드에 레코드 개수 포함. 각각의 출력 레코드에 추가 필드를 포함하려면 이 옵션을 선택하십시오(기본적으로 Record_Count라고 함). 이 필드는 각각의 통합 레코드를 형성하기 위해 통합된 입력 레코드 수를 표시합니다. 이 필드에 대해 사용자 정의 이름을 작성하려면 편집 필드에서 항목을 입력하십시오.

고유 복합 - 사용자 정의 탭

사용자 정의 채우기 대화 상자는 새 복합 레코드를 완료하는 데 사용되는 값에 대한 추가적인 제어를 제공합니다. 복합 탭에서 단일 필드 행만 사용자 정의하는 경우에는 이 옵션을 사용하기 전에 먼저 데이터를 인스턴스화해야 합니다.

참고: 이 대화 상자는 복합 탭의 **다음을 기반으로 값 채우기** 열에서 사용자 정의 값을 선택하는 경우에만 사용할 수 있습니다.

필드 유형에 따라 다음 옵션 중 하나에서 선택할 수 있습니다.

- **빈도별 선택.** 데이터 레코드에서 발생하는 빈도를 기반으로 값을 선택하십시오.

  참고: 연속형, 유형 없음 또는 날짜/시간 유형을 가진 필드에는 사용할 수 없습니다.

  - **사용.** 최대 빈도와 최소 빈도 중에서 선택하십시오.
  - **동률.** 발생 빈도가 동일한 레코드가 둘 이상 있는 경우에는 필요한 레코드를 선택하는 방법을 지정하십시오. 첫 번째 사용, 마지막 사용, 최저 사용 또는 최고 사용이라는 네 가지 옵션 중 하나에서 선택할 수 있습니다.
- **값(T/F) 포함.** 필드를 그룹의 레코드가 지정된 값을 가지고 있는지 식별하는 플래그로 변환하려면 선택하십시오. 그런 다음 선택된 필드에 대한 목록에서 **값**을 선택할 수 있습니다.

  참고: 복합 탭에서 둘 이상의 필드 행을 선택하는 경우에는 사용할 수 없습니다.
- **목록에서 첫 번째 일치.** 복합 레코드에 제공할 값의 우선 순위를 지정하려면 선택하십시오. 그런 다음 선택된 필드에 대한 목록에서 **항목** 중 하나를 선택할 수 있습니다.

  참고: 복합 탭에서 둘 이상의 필드 행을 선택하는 경우에는 사용할 수 없습니다.
- **값 연결.** 그룹의 모든 값을 문자열로 연결하여 유지하려면 선택하십시오. 각각의 값 사이에서 사용할 구분자를 지정해야 합니다.

  참고: 이는 연속형, 유형 없음 또는 날짜/시간 유형을 가진 하나 이상의 행을 선택하는 경우 사용 가능한 유일한 옵션입니다.
- **구분자 사용. 공백** 또는 **쉼표**를 연결된 문자열의 구분자 값으로 사용하도록 선택할 수 있습니다. 또는 **기타** 필드에서 자체 구분자 값 문자를 입력할 수 있습니다.

  참고: 값 연결 옵션을 선택하는 경우에만 사용할 수 있습니다.

스트리밍 시계열 노드

스트리밍 시계열 노드를 사용하여 한 단계로 시계열 모델을 작성하고 스코어링합니다. 목표 필드마다 개별 시계열 모델이 작성되지만, 생성된 모델 팔레트에 모델 너깃이 추가되지 않으며 모델 정보를 찾아볼 수 없습니다.

시계열 데이터를 모델링하는 방법에서는 각 측정 사이에 균일한 구간이 필요합니다(빈 행으로 결측값을 표시함). 데이터가 이미 이 요구사항을 만족하지 않으면 필요에 따라 값을 변환해야 합니다.

시계열 노드와 관련하여 또 다른 주의 사항은 다음과 같습니다.

- 필드는 숫자여야 합니다.
- 날짜 필드를 입력으로 사용할 수 없습니다.
- 파티션은 무시됩니다.

스트리밍 시계열 노드는 시계열에 대한 지수평활, 일변량 ARIMA(Autoregressive Integrated Moving Average), 다변량 ARIMA(또는 전이 함수) 모델을 추정하고 시계열 데이터에 기반하여 예측을 생성합니다. 하나 이상의 목표 필드에 가장 적합한 ARIMA 또는 지수평활 모델을 자동으로 식별하고 추정하는 자동 모델 생성기도 사용 가능합니다.

시계열 모델링에 대한 자세한 정보는 SPSS Modeler 모델링 노드 안내서의 시계열 모델 섹션을 참조하십시오.

IBM SPSS Collaboration and Deployment Services 스코어링 서비스 또는 IBM InfoSphere Warehouse를 사용하여 IBM SPSS Modeler Solution Publisher를 통해 스트리밍 배포 환경에서 스트리밍 시계열 노드를 사용할 수 있습니다.

스트리밍 시계열 노드 - 필드 옵션

필드 탭에서는 업스트림 노드에 이미 정의된 필드 역할 설정을 사용하거나 수동으로 필드를 지정할 수 있습니다.

사전 정의된 역할 사용 이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)에서 역할 설정(목표, 예측변수 등)을 사용합니다.

사용자 정의 필드 할당 사용 목표, 예측변수, 기타 역할을 수동으로 할당하려면 이 옵션을 선택하십시오.

필드 화살표 단추를 사용하여 이 목록에서 화면 오른쪽의 다양한 역할 필드에 항목을 수동으로 할당합니다. 아이콘은 각 역할 필드에 대한 유효한 측정 수준을 나타냅니다.

목록의 모든 필드를 선택하려면 **모두** 단추를 클릭하거나 개별 측정 수준 단추를 클릭하여 이 측정 수준의 모든 필드를 선택하십시오.

목표 한 필드를 예측 목표로 선택합니다.

후보 입력 예측에 대한 입력으로 하나 이상의 필드를 선택합니다.

이벤트 및 개입 - 이 영역을 사용하여 특정 입력 필드를 이벤트 또는 개입 필드로 지정할 수 있습니다. 이와 같이 지정하면 이벤트(판매 프로모션과 같은 예측 가능한 반복 상황) 또는 개입(정전 또는 직원 파업과 같은 일회성 사건)의 영향을 받을 수 있는 시계열 데이터를 포함하는 항목으로 필드를 식별합니다.

스트리밍 시계열 노드 - 데이터 지정 사항 옵션

데이터 지정 사항 탭에서는 모델에 포함될 데이터에 대한 모든 옵션을 설정할 수 있습니다. 물론, **실행** 단추를 클릭하여 모든 기본 옵션으로 모델을 작성할 수도 있지만, 보통 사용자는 고유한 목적을 위해 작성을 사용자 정의하려고 합니다.

탭은 모델에 특정한 사용자 정의를 설정하는 여러 창을 포함합니다.

스트리밍 시계열 노드 - 관측값

이 분할창의 설정을 사용하여 관측값을 정의하는 필드를 지정할 수 있습니다.

날짜/시간 필드로 지정된 관측값

관측값이 날짜, 시간 또는 시간소인 필드를 통해 정의되도록 지정할 수 있습니다. 관측값을 정의하는 필드 외에, 관측값을 설명하는 적절한 시간 간격을 선택하십시오. 지정된 시간 간격에 따라, 관측값(증분) 사이의 구간이나 주당 일 수와 같은 다른 설정을 지정할 수도 있습니다. 다음 고려사항은 시간 간격에 적용됩니다.

- 관측값이 시간에서 비정규적으로 간격이 있는 경우(판매 순서가 처리되는 시간과 같이), **비정규** 값을 사용하십시오. **비정규**가 선택될 때, 데이터 지정 사항 탭의 **시간 간격** 설정에서 분석에 사용되는 시간 간격을 지정해야 합니다.
- 관측값이 날짜와 시간을 나타내고 시간 간격이 시, 분 또는 초인 경우 **하루 중 시간(시)**, **하루 중 시간(분)** 또는 **하루 중 시간(초)**을 사용하십시오. 관측값이 날짜에 대한 참조 없이 시간(기간)을 나타내고 시간 간격이 시, 분 또는 초일 경우, **시(비주기적)**, **분(비주기적)** 또는 **초(비주기적)**를 사용하십시오.
- 선택된 시간 간격을 기초로, 프로시저는 결측 관측값을 발견할 수 있습니다. 프로시저에서는 모든 관측값이 시간에서 동일하게 간격을 두고 결측 관측값이 없다고 가정하므로, 결측 관측값을 발견해야 합니다. 예를 들어, 시간 간격이 일(Days)이고 날짜 2015-10-27 뒤에 2015-10-29가 있는 경우, 2015-10-28에 대해 결측 관측값이 있습니다. 결측 관측값에 대해 값이 대체됩니다. 데이터 지정 사항 탭의 **결측값 처리** 영역에서 결측값 처리 설정을 지정하십시오.
- 지정된 시간 간격은 프로시저가 함께 통합해야 하는 동일한 시간 간격의 여러 관측값을 발견하고 관측값에 동일하게 간격이 있도록 월의 첫 번째와 같은 구간 경계에 관측값을 맞출 수 있도록 합니다. 예를 들어, 시간 간격이 월일 경우, 동일 월에 있는 여러 날짜가 함께 통합됩니다. 이 유형의 통합을 *그룹화*라고 합니다. 기본적으로, 관측값은 그룹화될 때 합산됩니다. 데이터 지정 사항 탭의 **통합 및 분포** 설정에서, 그룹화에 다른 방법(예: 관측값의 평균)을 지정할 수 있습니다.
- 일부 시간 간격의 경우, 추가 설정은 동일하게 간격이 있는 정규 구간에서 중단을 설정할 수 있습니다. 예를 들어, 시간 간격이 일(Days)이지만 평일만 유효한 경우, 주에 5일이 있고 주는 월요일에 시작함을 지정할 수 있습니다.

관측값이 주기 또는 순환 주기로 정의됨

관측값은 임의의 순환 수준 수까지, 주기 또는 반복 주기 순환을 나타내는 하나 이상의 정수 필드로 정의할 수 있습니다. 이 구조를 사용할 경우 표준 시간 간격 중 하나에 맞지 않은 관측값 계열을 기술

할 수 있습니다. 예를 들어, 10개월만 있는 회계연도는 연도를 나타내는 순환 필드와, 월을 나타내는 주기 필드로 설명할 수 있습니다. 여기서 하나의 주기 길이는 10입니다.

순환 주기를 지정하는 필드는 주기적 수준의 계층 구조를 정의합니다. 가장 낮은 수준은 **주기** 필드에 의해 정의됩니다. 다음 최상위 수준은 수준이 1인 순환 필드에 의해 지정되고, 그 다음은 수준 2의 순환 필드로 지정되며 뒤로도 마찬가지입니다. 가장 높은 수준을 제외하고, 각 수준의 필드 값은 다음 최상위 수준에 관하여 주기적이어야 합니다. 최상위 수준의 값은 주기적이 될 수 없습니다. 예를 들어, 10달 회계연도의 경우 월은 연도 내에서 주기적이며 연도는 주기적이 아닙니다.

- 특정 수준에 있는 순환의 길이는 다음으로 가장 낮은 수준의 주기성입니다. 회계연도 예의 경우, 단 하나 순환 수준이 있고 순환 길이는 10입니다. 다음으로 가장 낮은 수준이 월을 나타내고 지정된 회계 연도에 10달이 있기 때문입니다.
- 주기적 필드의 시작 값(1부터 시작하지 않음)을 지정하십시오. 이 설정은 결측값을 발견하는 데 필요합니다. 예를 들어, 주기적 필드는 2에서 시작하지만 시작 값은 1로 지정되는 경우, 프로시저는 해당 필드의 각 순환에 있는 첫 번째 주기에 대해 결측값이 있다고 가정합니다.

스트리밍 시계열 노드 - 분석 시간 간격

분석에 사용할 시간 간격은 관측값의 시간 간격과 다를 수 있습니다. 예를 들어, 관측값의 시간 간격이 일(Days)일 경우, 분석의 시간 간격으로는 월을 선택할 수 있습니다. 그런 다음 모델이 작성되기 전에 매일 데이터에서 매월 데이터까지 데이터가 통합됩니다. 또한 데이터를 장기 시간 간격에서 단기 시간 간격으로 분포할 것을 선택할 수도 있습니다. 예를 들어, 관측값이 분기별인 경우, 데이터를 분기별에서 월별 데이터로 분포할 수 있습니다.

이 분할창의 설정을 사용하여 분석 시간 간격을 지정할 수 있습니다. 데이터가 통합되거나 분포되는 방법은 데이터 지정 사항 탭의 **통합 및 분포** 설정에서 지정됩니다.

분석이 행해지는 시간 간격에 대해 사용 가능한 선택은 해당 관측값 정의 방법과 관측값의 시간 간격에 따라 다릅니다. 특히, 관측값이 순환 주기로 정의될 경우 통합만 지원됩니다. 그러한 경우, 분석의 시간 간격은 관측값의 시간 간격보다 크거나 같아야 합니다.

스트리밍 시계열 노드 - 통합 및 분포 옵션

이 분할창의 설정을 사용하여 관측값의 시간 간격과 관련한 입력 데이터 통합 또는 분포 설정을 지정할 수 있습니다.

통합 함수

분석에 사용되는 시간 간격이 관측에 사용되는 시간 간격보다 길 경우, 입력 데이터는 통합됩니다. 예를 들어, 관측값의 시간 간격이 일(Days)이고 분석의 시간 간격이 월일 경우 통합이 수행됩니다. mean, sum, mode, min 또는 max 통합 함수를 사용할 수 있습니다.

분포 함수

분석에 사용되는 시간 간격이 관측의 시간 간격보다 짧을 경우, 입력 데이터는 분포됩니다. 예를 들어, 관측값의 시간 간격이 분기이고 분석의 시간 간격이 월일 경우 분포가 수행됩니다. mean 또는 sum 분포 함수를 사용할 수 있습니다.

그룹화 함수

그룹화는 관측값이 날짜/시간에 의해 정의되고 여러 관측값이 동일 시간 간격에 발생하는 경우에 적용됩니다. 예를 들어, 관측값의 시간 간격이 월일 경우, 동일 월에 있는 여러 날짜가 그룹화되어 날짜가 발생하는 월과 연관됩니다. mean, sum, mode, min 또는 max와 같은 그룹화 함수를 사용할 수 있습니다. 그룹화는 항상 관측값이 날짜/시간에 의해 정의되고 관측값의 시간 간격이 비정규로 지정된 경우에 수행됩니다.

참고: 그룹화가 통합 양식이어도, 그룹화는 결측값 처리 이전에 수행됩니다(정상 통합은 결측값 처리 이후에 수행됩니다). 관측값의 시간 간격이 비정규로 지정되는 경우, 통합은 그룹화 함수로만 수행됩니다.

교차-일 관측값을 이전 일로 통합

1일 경계를 교차하는 시간을 사용하는 관측값이 전날의 값에 통합되는지 여부를 지정합니다. 예를 들어, 20:00시에 시작하는 8시간 노동의 시간별 관측값의 경우, 이 설정은 00:00시와 04:00시 사이의 관측값이 전날 통합 결과에 포함되는지 여부를 지정합니다. 이 설정은 관측값의 시간 간격이 하루 중 시간(시), 하루 중 시간(분) 또는 하루 중 시간(초)이고 분석의 시간 간격이 일(Days)인 경우에만 적용됩니다.

지정된 필드에 대한 사용자 정의 설정

필드 기준으로 필드에 통합, 분포 및 그룹화 함수를 지정할 수 있습니다. 이 설정은 통합, 분포 및 그룹화 함수에 대한 기본 설정을 대체합니다.

스트리밍 시계열 노드 - 결측값 옵션

이 분할창의 설정을 사용하여 입력 데이터의 결측값을 대체값으로 바꾸는 방법을 지정할 수 있습니다. 다음 방법으로 바꿀 수 있습니다.

선형 보간법

선형 보간법을 사용하여 결측값을 바꿉니다. 결측값 이전의 마지막 유효한 값과 결측값 이후의 첫 번째 유효한 값이 보간법에 사용됩니다. 계열에서 첫 번째 또는 마지막 관측값에 결측값이 있는 경우, 계열의 시작 또는 종료에서 두 개의 가장 근접한 비결측 값이 사용됩니다.

계열 평균

결측값을 전체 계열에 대한 평균으로 바꿉니다.

근접한 값들의 평균

결측값을 유효한 근접 값의 평균으로 바꿉니다. 근접한 값들의 계산너비는 평균을 계산하는데 사용되는 결측값 전후의 유효값 수입니다.

근접한 값들의 중앙값

결측값을 근접한 유효한 값의 중앙값으로 바꿉니다. 근접한 값들의 계산너비는 평균을 계산하는데 사용되는 결측값 전후의 유효값 수입니다.

선형 추세

이 옵션은 단순 선형 회귀 모델에 맞게 계열의 모든 비결측 관측값을 사용한 후 이 모델을 사용하여 결측값을 대체합니다.

기타 설정:

결측값의 최대 퍼센트(%)

모든 계열에 허용되는 최대 결측값 퍼센트를 지정합니다. 지정된 최대값보다 많은 결측값이 있는 계열은 분석에서 제외됩니다.

스트리밍 시계열 노드 - 추정 기간

추정 기간 분할창에서 모델 추정에 사용될 레코드의 범위를 지정할 수 있습니다. 기본적으로 추정 기간은 모든 계열에 걸쳐 최초 관측값 시간에 시작되고 최근 관측값 시간에 종료됩니다.

시작 및 종료 시간 기준

추정 기간의 시작 및 종료 둘 다를 지정하거나 시작 또는 종료만 지정할 수 있습니다. 추정 기간의 시작 또는 종료를 생략하는 경우, 기본값이 사용됩니다.

- 날짜/시간 필드에 의해 관측값이 정의된 경우, 날짜/시간 필드에 사용되는 것과 동일한 형식으로 시작 및 종료 값을 입력하십시오.
- 순환 주기에 의해 정의된 관측값의 경우, 순환 주기 필드마다 값을 지정하십시오. 각 필드는 별도의 열에 표시됩니다.

최근이거나 최초의 시간 간격(L)

선택적 오프셋으로, 데이터의 최초 시간 간격에 시작하거나 최근 시간 간격에 종료하는, 지정된 시간 간격 수로 추정 기간을 정의합니다. 이 컨텍스트에서, 시간 간격은 분석의 시간 간격을 가리킵니다. 예를 들어, 관측값이 매월 단위이지만 분석의 시간 간격은 분기일 수 있습니다. **최근**과 **시간 간격 수**로 24 값을 지정하면 최근 24개 분기를 의미합니다.

선택적으로, 지정된 시간 간격 수를 제외할 수 있습니다. 예를 들어, 최근 24 시간 간격을 지정하고 제외할 수로 1를 지정하면, 추정 기간은 마지막 구간 앞에 있는 24개 구간으로 구성됩니다.

스트리밍 시계열 노드 - 작성 옵션

작성 옵션 탭은 모델을 작성하기 위한 모든 옵션을 설정하는 위치입니다. 물론, **실행** 단추를 클릭하여 모든 기본 옵션으로 모델을 작성할 수도 있지만, 보통 사용자는 고유한 목적을 위해 작성을 사용자 정의하려고 합니다.

이 탭은 해당 모델에만 적용되는 사용자 정의를 설정하는 두 개의 분할창으로 구성됩니다.

스트리밍 시계열 노드 - 일반 작성 옵션

이 분할창에서 사용 가능한 옵션은 **방법** 목록에서 선택한 다음 세 가지 설정에 따라 다릅니다.

- **자동 모델 생성기** - 자동 모델 생성기를 사용하려면 이 옵션을 선택하십시오. 그러면 각 종속 계열에 대한 최적 적합 모델을 자동으로 찾습니다.
- **지수평활** - 이 옵션을 사용하여 사용자 정의 지수평활 모델을 지정할 수 있습니다.
- **ARIMA** - 이 옵션을 사용하여 사용자 정의 ARIMA 모델을 지정할 수 있습니다.

자동 모델 생성기

모델 유형 - 작성하려는 모델의 유형을 선택하십시오.

- **모든 모델** - 자동 모델 생성기에서 ARIMA 및 지수평활 모델을 모두 고려합니다.
- **지수평활 모델만** - 자동 모델 생성기에서 지수평활 모델만 고려합니다.
- **ARIMA 모델만** - 자동 모델 생성기에서 ARIMA 모델만 고려합니다.

자동 모델 생성기에서 계절 모델 고려 - 이 옵션은 활성 데이터 세트에 대해 주기성이 정의된 경우에만 사용할 수 있습니다. 이 옵션을 선택하면 자동 모델 생성기는 계절 및 비계절 모델을 모두 고려합니다. 이 옵션을 선택하지 않은 경우 자동 모델 생성기에서 비계절 모델만 고려합니다.

자동으로 이상치 검색 - 기본적으로 이상치 자동 검색을 수행하지 않습니다. 이상값 자동 발견을 수행하려면 이 옵션을 선택하고 원하는 이상값 유형을 선택합니다.

입력 필드는 이 목록에 포함되기 전에 *플래그*, *명목* 또는 *순서*와 같은 측정 수준이 있어야 하며 숫자여야 합니다(예: 플래그 필드의 경우 참/거짓이 아닌 1/0).

자동 모델 생성기는 **필드** 탭의 이벤트 또는 개입 필드로 식별된 입력에 대해 임의 전이 함수가 아닌 단순 회귀분석만 고려합니다.

지수평활

모델 유형 - 지수평활 모델은 계절 또는 비계절로 분류됩니다. [1] 계절 모델은 데이터 지정 사항 탭의 시간 간격 분할창을 사용하여 정의된 주기성이 계절인 경우에만 사용할 수 있습니다. 계절 주기성은 다음과 같습니다. 주기적 기간, 연도, 분기, 월, 한 주의 요일, 하루의 시간, 하루의 분, 하루의 초. 다음 모델 유형을 사용할 수 있습니다.

- **단순** - 이 모델은 추세나 계절성이 없는 계열에 적합합니다. 유일하게 관련된 평활 모수는 수준입니다. 단순 지수평활은 자동 선형회귀 차수가 0, 차이 차수가 1, 이동 평균 차수가 1, 및 상수 없음인 ARIMA와 가장 비슷합니다.
- **Holt의 선형 추세** - 이 모델은 선형 추세는 있고 계절성이 없는 계열에 적합합니다. 관련 평활 모수는 수준 및 추세이며, 이 모델에서는 서로의 값으로 제한되지 않습니다. Holt 모델은 Brown 모델보다 일반적이지만 대형 계열의 추정값 계산에는 더 오래 걸립니다. Holt 지수평활은 자동 선형회귀 차수가 0, 차이 차수가 2, 이동 평균 차수가 2인 ARIMA와 가장 비슷합니다.
- **Brown의 선형 추세** - 이 모델은 선형 추세는 있고 계절성이 없는 계열에 적합합니다. 관련 평활 모수는 수준 및 추세지만, 이 모델에서는 동일하다고 가정합니다. 따라서 Brown의 모형은 Holt 모형의 특별한 케이스입니다. Brown의 지수평활은 자동 선형회귀 차수가 0, 차이 차수가 2, 이동 평균 차수가 2이며, 이동 평균의 두 번째 차수에 대한 계수가 첫 번째 차수에 대한 계수의 절반과 같은 ARIMA와 가장 비슷합니다.

---

1. Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting,* 4, 1-28.

- **진폭감소 추세** - 이 모델은 점점 소멸되는 선형 추세는 있고 계절성이 없는 계열에 적합합니다. 관련된 평활 모수는 수준, 추세, 진폭감소 추세입니다. 진폭감소 지수평활은 자동 선형회귀 차수가 1, 차이 차수가 1, 이동 평균 차수가 2인 ARIMA와 가장 비슷합니다.
- **단순 계절모델** - 이 모델은 추세와 계절 효과가 없고 시간에 따라 일정한 계열에 적합합니다. 관련된 평활 모수는 수준과 계절입니다. 계절 지수평활은 자동 선형회귀 차수가 0, 차이 차수가 1, 계절 차이 차수가 1, 이동 평균의 경우 차수 1, *p*, *p*+1인 ARIMA와 가장 유사합니다. 여기서 *p*는 계절 간격에서 기간 수입니다. 월별 데이터의 경우 $p = 12$입니다.
- **Winters의 가법** - 이 모델은 선형 추세와 계절 효과가 있고 시간에 따라 일정한 계열에 적합합니다. 관련된 평활 모수는 수준, 추세, 계절입니다. Winters의 가법 지수평활은 자동 선형회귀 차수가 0, 차이 차수가 1, 계절 차이 차수가 1, 이동 평균의 경우 차수가 *p*+1인 ARIMA와 가장 유사합니다. 여기서 *p*는 계절 간격에서 기간 수입니다. 월별 데이터의 경우 $p = 12$입니다.
- **Winters의 승법** - 이 모델은 선형 추세와 계절 효과가 있고 계열 규모에 따라 바뀌는 계열에 적합합니다. 관련된 평활 모수는 수준, 추세, 계절입니다. Winters의 승법 지수평활은 ARIMA 모델과 다릅니다.

목표 변환 - 모델링하기 전에 각 종속 변수에 대해 수행할 변환을 지정할 수 있습니다.

- **없음** - 변환을 수행하지 않습니다.
- **제곱근** - 제곱근 변환을 수행합니다.
- **자연로그** - 자연로그 변환을 수행합니다.

ARIMA

사용자 정의 ARIMA 모델의 구조를 지정하십시오.

ARIMA 차수 - 눈금의 해당 셀에 모델의 다양한 ARIMA 구성요소 값을 입력하십시오. 모든 값은 0또는 음이 아닌 정수여야 합니다. 자기회귀 및 이동 평균 구성요소의 경우 값은 최대 차수를 나타냅니다. 더 낮은 양의 차수가 모두 모델에 포함됩니다. 예를 들어, 2를 지정하면 모델에 차수 2와 1이 포함됩니다. 계절 열의 셀은 활성 데이터 세트에 대해 주기성이 정의된 경우에만 사용할 수 있습니다.

- **자기회귀(p)** - 모델의 자기회귀 차수 수입니다. 자기회귀 차수는 계열의 이전 값 중 현재 값 예측에 사용될 값을 지정합니다. 예를 들어, 자기회귀 차수 2는 과거 2개 시간 주기의 계열 값을 현재 값 예측에 사용하도록 지정합니다.
- **차분(d)** - 모델을 추정하기 전 계열에 적용할 차이 차수를 지정합니다. 추세가 존재하며(추세가 있는 계열은 일반적으로 비정상이며 ARIMA 모델링은 정상성을 가정) 해당 효과 제거를 위해 사용되는 경우 차이가 필요합니다. 차이 차수는 계열 추세 수준에 해당합니다. 1차 차이는 선형 추세, 2차 차이는 2차 추세 등을 나타냅니다.
- **이동 평균(q)** - 모델의 이동 평균 차수 수입니다. 이동 평균 차수는 이전 값에 대한 계열 평균 편차를 사용하여 현재 값을 예측하는 방법을 지정합니다. 예를 들어, 이동 평균 차수 1과 2는 계열의 현재 값을 예측하는 경우 지난 2개 시간 주기 각각의 계열 평균값 편차를 고려하도록 지정합니다.

계절모델 - 계절 자기회귀, 이동 평균, 차이 구성요소는 해당 비계절 구성요소와 동일한 역할을 합니다. 그러나 계절 차수의 경우 현재 계열 값이 한 개 이상의 계절 주기에 의해 구분된 이전 계열 값의 영향을 받습니다. 예

를 들어, 월별 데이터(계절 주기 12)의 경우 계절 차수 1은 현재 계열 값이 현재 계열 이전의 계열 값 12 주기의 영향을 받는다는 것을 의미합니다. 월별 데이터의 경우 계절 차수 1은 비계절 차수 12를 지정하는 것과 동일합니다.

자동으로 이상치 검색 - 이상치 자동 검색을 수행하려면 이 옵션을 선택하고 사용 가능한 이상치 유형 중에서 한 개 이상을 선택하십시오.

탐지할 이상값 유형 - 발견할 이상값 유형을 선택하십시오. 지원되는 유형은 다음과 같습니다.

- 가법(기본값)
- 수준 이동(기본값)
- 혁신적
- 일시적
- 계절 가법
- 국소적 추세
- 가법 수정

전이 함수 차수 및 변환 - 변환을 지정하고 ARIMA 모델의 일부 또는 전체 입력 모델에 대해 전이 함수를 정의하려면 **설정**을 클릭하십시오. 전이 및 변환 세부사항을 입력할 수 있는 별도의 대화 상자가 표시됩니다.

모델에 상수 포함 - 전체 평균 계열 값이 확실히 0인 경우가 아니라면 상수를 포함시키는 것이 표준입니다. 차분을 적용하는 경우에는 상수를 제외시키는 것이 좋습니다.

전이 및 변환 함수: 전이 함수 차수 및 변환 대화 상자에서는 변환을 지정하고 ARIMA 모델의 일부 또는 전체 입력 모델에 대해 전이 함수를 정의할 수 있습니다.

목표 변환 - 이 분할창에서는 모델링하기 전에 각 목표변수에 대해 수행할 변환을 지정할 수 있습니다.

- **없음** - 변환을 수행하지 않습니다.
- **제곱근** - 제곱근 변환을 수행합니다.
- **자연로그** - 자연로그 변환을 수행합니다.

후보 입력 전이 함수 및 변환 - 전이 함수를 사용하여 목표 계열의 미래 값을 예측하는 데 입력 필드의 과거 값을 사용하는 방식을 지정할 수 있습니다. 분할창 왼쪽에 있는 목록에는 모든 입력 필드가 표시됩니다. 이 분할창의 나머지 정보는 선택한 입력 필드에 따라 다릅니다.

전이 함수 차수 - 구조 눈금의 해당 셀에 전이 함수의 다양한 구성요소 값을 입력하십시오. 모든 값은 0또는 음이 아닌 정수여야 합니다. 분자 및 분모 구성요소의 경우 값은 최대 차수를 나타냅니다. 더 낮은 양의 차수가 모두 모델에 포함됩니다. 또한 분자 구성요소에 대해 차수 0은 항상 포함됩니다. 예를 들어, 분자로 2를 지정하면 모델에 차수 2, 1, 0이 포함됩니다. 분모로 3을 지정하면 모델에 차수 3, 2, 1이 포함됩니다. 계절 열의 셀은 활성 데이터 세트에 대해 주기성이 정의된 경우에만 사용할 수 있습니다.

분자 - 전이 함수의 분자 차수는 종속 계열의 현재 값을 예측하는 데 사용되는 선택된 독립(예측변수) 계열의 이전 값을 지정합니다. 예를 들어, 분자 차수 1은 각 종속 계열의 현재 값 예측에 과거 1개 시간 주기의 독립 계열 값과 독립 계열의 현재 값을 사용하도록 지정합니다.

분모 - 전이 함수의 분모 차수는 선택된 독립(예측변수) 계열의 이전 값에 대해 계열 평균의 편차가 종속 계열의 현재 값 예측에 어떻게 사용되는지 지정합니다. 예를 들어 분모 차수 1은 각 종속 계열의 현재 값을 예측하는 경우 과거 1개 시간 주기의 독립 계열 평균 값 편차를 고려하도록 지정합니다.

차이 - 모델을 추정하기 전 선택된 독립(예측변수) 계열에 적용할 차이 차수를 지정합니다. 추세가 있으며 해당 효과 제거를 위해 사용되는 경우 차이가 필요합니다.

계절모델 - 계절 분자, 분모 및 차이 구성요소는 해당 비계절 구성요소와 동일한 역할을 합니다. 그러나 계절 차수의 경우 현재 계열 값이 한 개 이상의 계절 주기에 의해 구분된 이전 계열 값의 영향을 받습니다. 예를 들어, 월별 데이터(계절 주기 12)의 경우 계절 차수 1은 현재 계열 값이 현재 계열 이전의 계열 값 12 주기의 영향을 받는다는 것을 의미합니다. 월별 데이터의 경우 계절 차수 1은 비계절 차수 12를 지정하는 것과 동일합니다.

보류 - 보류를 설정하면 지정된 구간 수만큼 입력 필드의 영향력이 보류됩니다. 예를 들어, 보류가 5로 설정된 경우 시간 *t*의 입력 필드 값은 다섯 구간이 경과할 때까지 예측에 영향을 주지 않습니다(*t* + 5).

변환 - 독립변수 세트에 대해 전이 함수를 지정하면 해당 변수에 대해 수행할 선택적 변환도 포함됩니다.

- **없음** - 변환을 수행하지 않습니다.
- **제곱근** - 제곱근 변환을 수행합니다.
- **자연로그** - 자연로그 변환을 수행합니다.

스트리밍 시계열 노드 - 모델 옵션

신뢰구간 너비(%) - 모델 예측 및 잔차 자기상관에 대해 신뢰구간을 계산합니다. 100보다 작은 양수 값을 지정할 수 있습니다. 기본적으로, 95% 신뢰구간이 사용됩니다.

예측

- **레코드를 미래로 확장** 옵션은 추정 기간이 끝난 이후에 예측할 시간 간격 수를 설정합니다. 이 경우의 시간 간격은 데이터 지정 사항 탭에 지정한 분석의 시간 간격입니다. 예측이 요청되면 물론 목표가 아닌 입력 계열에서 자기회귀분석 모델이 자동으로 작성됩니다. 그런 다음, 이 모델을 사용하여 예측 기간에 해당 입력 계열의 값을 생성합니다. 이 설정에 최대 한계는 없습니다.
- **입력의 미래 값 계산** - 이 옵션을 선택하면 예측자의 예측 값, 잡음 예측, 분산 추정 및 미래 시간 값이 계산됩니다.

점수에 사용 가능 - 모델 너깃에 대한 대화 상자에 표시될 스코어링 옵션에 대한 기본값을 여기서 설정할 수 있습니다.

- **신뢰구간 상한 및 하한 계산** - 선택할 경우 이 옵션은 각 대상 필드에서 하한 및 하한 신뢰구간과 이러한 값의 총계에 대한 새 필드(기본 접두문자: $TSLCI- 및 $TSUCI-)를 작성합니다.

- **잡음 잔차 계산** - 선택할 경우 이 옵션은 각 대상 필드에서 모델 잔차와 이러한 값의 총계에 대한 새 필드(기본 접두문자: $TSNR-)를 작성합니다.

스트리밍 TCM 노드

스트리밍 TCM 노드는 임시 원인 모델을 한 번에 작성하고 스코어링하는 데 사용할 수 있습니다.

시간 인과 모델링에 대한 자세한 정보는 SPSS Modeler Modeling Nodes 안내서의 시계열 모델 섹션에서 시간 인과 모델 주제를 참조하십시오.

스트리밍 TCM 노드 - 시계열 옵션

필드 탭에서, 모델 시스템에서 포함할 계열을 지정하려면 **시계열** 설정을 사용하십시오.

데이터에 적용되는 데이터 구조에 맞는 옵션을 선택하십시오. 다차원 데이터의 경우, 차원 필드를 지정하기 위해 **차원 선택**을 클릭하십시오. 차원 필드의 지정된 순서는 모두 연속 대화 상자 및 출력에 차원이 나타나는 순서를 정의합니다. 차원 필드를 다시 정렬하려면 차원 선택 하위 대화 상자에서 위로 및 아래로 화살표 단추를 사용하십시오.

열 기반 데이터에 대해 *계열* 용어의 의미는 *필드* 용어 의미와 같습니다. 다차원 데이터의 경우, 시계열을 포함하는 필드는 *메트릭* 필드로 언급됩니다. 다차원 데이터에 대해 시계열은 차원 필드 각각에 대한 메트릭 필드 및 값으로 정의됩니다. 다음 고려사항은 열 기반 및 다차원 데이터에 적용됩니다.

- 후보 입력으로, 또는 목표 및 입력 둘 다로 지정되는 계열은 각 목표에 대해 모델에서 포함을 위해 고려됩니다. 각 목표에 대한 모델에는 항상 목표 자체의 시차 값이 포함됩니다.
- 강제 입력으로 지정되는 계열은 항상 각 목표에 대해 모델에 포함됩니다.
- 최소 하나의 계열을 목표로 또는 목표 및 입력 둘 다로 지정해야 합니다.
- **사전 정의된 역할 사용**이 선택되면, 입력 역할을 가지고 있는 필드가 후보 입력으로 설정됩니다. 사전 정의된 어떤 규칙도 강제 입력에 맵핑하지 않습니다.

다차원 데이터

다차원 데이터의 경우, 눈금에서 메트릭 필드 및 연관된 역할을 지정합니다. 눈금의 각 행은 단일 메트릭 및 역할을 지정합니다. 기본적으로, 모델 시스템에는 눈금의 각 행에 대한 차원 필드의 모든 조합에 대한 계열이 포함됩니다. 예를 들어, *region* 및 *brand*에 대한 차원이 있는 경우, 기본적으로 목표로 메트릭 *sales*를 지정하면, 이는 각각의 *region* 및 *brand* 조합에 대해 별도의 sales 목표 계열이 있음을 의미합니다.

눈금의 각 행에 대해, 차원의 생략 기호 단추를 클릭하여 차원 필드에 대한 값 세트를 사용자 정의할 수 있습니다. 이 동작은 차원 값 선택 하위 대화 상자를 엽니다. 또한 눈금 행을 추가, 삭제 또는 복사할 수도 있습니다.

계열 개수 열은 현재 연관 메트릭에 대해 지정된 차원 값 세트 수를 표시합니다. 표시된 값은 실제 계열 수(세트당 하나의 계열)보다 클 수 있습니다. 이 조건은 지정된 차원 값 조합 중 일부가 연관 메트릭에 의해 포함된 계열에 해당되지 않는 경우에 발생합니다.

스트리밍 TCM 노드 - 차원 값 선택

다차원 데이터의 경우, 특정 역할이 있는 특정 메트릭 필드에 적용되는 차원 값을 지정하여 분석을 사용자 정의할 수 있습니다. 예를 들어, *sales*가 메트릭 필드이고 *channel*이 'retail' 및 'web' 값을 가지고 있는 차원인 경우, 'web' 판매가 입력이고 'retail' 판매가 목표임을 지정할 수 있습니다.

모든 값

현재 차원 필드의 모든 값이 포함됨을 지정합니다. 이 옵션은 기본값입니다.

포함하거나 제외할 값 선택

현재 차원 필드의 값 세트를 지정하려면 이 옵션을 사용하십시오. **모드**에 대해 **포함**이 선택되는 경우, **선택된 값** 목록에 지정되는 값만 포함됩니다. **모드**에 대해 **제외**가 선택되는 경우, **선택된 값** 목록에 지정된 값이 아닌 다른 모든 값이 포함됩니다.

선택할 값 세트를 필터링할 수 있습니다. 필터 조건에 충족하는 값은 **매치됨** 탭에 나타나고, 필터 조건과 일치하지 않는 값은 **선택되지 않은 값** 목록의 **매치하지 않음** 탭에 나타납니다. **모두** 탭은 필터 조건과 관계없이 선택되지 않은 모든 값을 나열합니다.

- 필터를 지정할 때 와일드카드 문자를 표시하기 위해 별표(*)를 사용할 수 있습니다.
- 현재 필터를 지우려면, 표시된 값 필터링 대화 상자에 검색어로 비어 있는 값을 지정하십시오.

스트리밍 TCM 노드 - 관측 옵션

필드 탭에서, 관측값을 정의하는 필드를 지정하려면 **관측값** 설정을 사용하십시오.

날짜/시간에 의해 정의되는 관측값

관측값이 날짜, 시간 또는 시간소인 필드에 의해 정의됨을 지정할 수 있습니다. 관측값을 정의하는 필드 외에, 관측값을 설명하는 적절한 시간 간격을 선택하십시오. 지정된 시간 구간에 따라, 관측값(증분) 사이의 구간이나 주당 일 수와 같은 다른 설정을 지정할 수도 있습니다. 다음 고려사항은 시간 간격에 적용됩니다.

- 관측값이 시간에서 비정규적으로 간격이 있는 경우(판매 순서가 처리되는 시간과 같이), **비정규** 값을 사용하십시오. **비정규**가 선택될 때, 데이터 지정 사항 탭의 **시간 간격** 설정에서 분석에 사용되는 시간 간격을 지정해야 합니다.
- 관측값이 날짜와 시간을 나타내고 시간 간격이 시, 분 또는 초인 경우 **하루 중 시간(시)**, **하루 중 시간(분)** 또는 **하루 중 시간(초)**을 사용하십시오. 관측값이 날짜에 대한 참조 없이 시간(기간)을 나타내고 시간 간격이 시, 분 또는 초일 경우, **시(비주기적)**, **분(비주기적)** 또는 **초(비주기적)**를 사용하십시오.
- 선택된 시간 간격을 기초로, 프로시저는 결측 관측값을 발견할 수 있습니다. 프로시저에서는 모든 관측값이 시간에서 동일하게 간격을 두고 결측 관측값이 없다고 가정하므로, 결측 관측값을 발견해야 합니다. 예를 들어, 시간 간격이 일(Days)이고 날짜 2014-10-27 뒤에 2014-10-29가 있는 경우, 2014-10-28에 대해 결측 관측값이 있습니다. 값은 결측 관측값에 대해 대체됩니다. 결측값 처리에 대한 설정은 데이터 지정 사항 탭으로부터 지정할 수 있습니다.
- 지정된 시간 간격은 프로시저가 함께 통합해야 하는 동일한 시간 간격의 여러 관측값을 발견하고 관측값에 동일하게 간격이 있도록 월의 첫 번째와 같은 구간 경계에 관측값을 맞출 수 있도록 합니

다. 예를 들어, 시간 간격이 월일 경우, 동일 월에 있는 여러 날짜가 함께 통합됩니다. 이 유형의 통합을 *그룹화*라고 합니다. 기본적으로, 관측값은 그룹화될 때 합산됩니다. 데이터 지정 사항 탭의 **통합 및 분포** 설정에서, 그룹화에 다른 방법(예: 관측값의 평균)을 지정할 수 있습니다.

- 일부 시간 간격의 경우, 추가 설정은 동일하게 간격이 있는 정규 구간에서 중단을 설정할 수 있습니다. 예를 들어, 시간 간격이 일(Days)이지만 평일만 유효한 경우, 주에 5일이 있고 주는 월요일에 시작함을 지정할 수 있습니다.

관측값이 주기 또는 순환 주기로 정의됨

관측값은 임의의 순환 수준 수까지, 주기 또는 반복 주기 순환을 나타내는 하나 이상의 정수 필드로 정의할 수 있습니다. 이 구조에서, 표준 시간 간격 중 하나에 맞지 않은 관측값 계열을 설정할 수 있습니다. 예를 들어, 10개월만 있는 회계연도는 연도를 나타내는 순환 필드와, 월을 나타내는 주기 필드로 설명할 수 있습니다. 여기서 하나의 주기 길이는 10입니다.

순환 주기를 지정하는 필드는 주기적 수준의 계층 구조를 정의합니다. 가장 낮은 수준은 **주기** 필드에 의해 정의됩니다. 다음 최상위 수준은 수준이 1인 순환 필드에 의해 지정되고, 그 다음은 수준 2의 순환 필드로 지정되며 뒤로도 마찬가지입니다. 가장 높은 수준을 제외하고, 각 수준의 필드 값은 다음 최상위 수준에 관하여 주기적이어야 합니다. 최상위 수준의 값은 주기적이 될 수 없습니다. 예를 들어, 10달 회계연도의 경우 월은 연도 내에서 주기적이며 연도는 주기적이 아닙니다.

- 특정 수준에 있는 순환의 길이는 다음으로 가장 낮은 수준의 주기성입니다. 회계연도 예의 경우, 단 하나 순환 수준이 있고 순환 길이는 10입니다. 다음으로 가장 낮은 수준이 월을 나타내고 지정된 회계 연도에 10달이 있기 때문입니다.
- 1에서 시작하지 않은 주기적 필드의 시작 값을 지정하십시오. 이 설정은 결측값을 발견하는데 필요합니다. 예를 들어, 주기적 필드는 2에서 시작하지만 시작 값은 1로 지정되는 경우, 프로시저는 해당 필드의 각 순환에 있는 첫 번째 주기에 대해 결측값이 있다고 가정합니다.

스트리밍 TCM 노드 - 시간 간격 옵션

분석에 사용되는 시간 간격은 관측값의 시간 간격과 다를 수 있습니다. 예를 들어, 관측값의 시간 간격이 일(Days)일 경우, 분석의 시간 간격으로는 월을 선택할 수 있습니다. 데이터는 모델이 작성되기 전에 매일 데이터에서 매월 데이터까지 통합됩니다. 또한 데이터를 장기 시간 간격에서 단기 시간 간격으로 분포할 것을 선택할 수도 있습니다. 예를 들어, 관측값이 분기별인 경우, 데이터를 분기별에서 월별 데이터로 분포할 수 있습니다.

분석이 행해지는 시간 간격에 대해 사용 가능한 선택은 해당 관측값 정의 방법과 관측값의 시간 간격에 따라 다릅니다. 특히, 관측값이 순환 주기로 정의될 경우, 통합만 지원됩니다. 그러한 경우, 분석의 시간 간격은 관측값의 시간 간격보다 *크거나* 같아야 합니다.

분석 시간 간격은 데이터 지정 사항 탭의 **시간 간격** 설정에서 지정됩니다. 데이터가 통합되거나 분포되는 방법은 데이터 지정 사항 탭의 **통합 및 분포** 설정에서 지정됩니다.

스트리밍 TCM 노드 - 통합 및 분포 옵션

통합 함수

분석에 사용되는 시간 간격이 관측에 사용되는 시간 간격보다 길 경우, 입력 데이터는 통합됩니다. 예를 들어, 관측값의 시간 간격이 일(Days)이고 분석의 시간 간격이 월일 경우 통합이 수행됩니다. mean, sum, mode, min 또는 max 통합 함수를 사용할 수 있습니다.

분포 함수

분석에 사용되는 시간 간격이 관측의 시간 간격보다 짧을 경우, 입력 데이터는 분포됩니다. 예를 들어, 관측값의 시간 간격이 분기이고 분석의 시간 간격이 월일 경우 분포가 수행됩니다. mean 또는 sum 분포 함수를 사용할 수 있습니다.

그룹화 함수

그룹화는 관측값이 날짜/시간에 의해 정의되고 여러 관측값이 동일 시간 간격에 발생하는 경우에 적용됩니다. 예를 들어, 관측값의 시간 간격이 월일 경우, 동일 월에 있는 여러 날짜가 그룹화되어 날짜가 발생하는 월과 연관됩니다. mean, sum, mode, min 또는 max와 같은 그룹화 함수를 사용할 수 있습니다. 그룹화는 항상 관측값이 날짜/시간에 의해 정의되고 관측값의 시간 간격이 비정규로 지정된 경우에 수행됩니다.

참고: 그룹화가 통합 양식이어도, 그룹화는 결측값 처리 이전에 수행됩니다(정상 통합은 결측값 처리 이후에 수행됩니다). 관측값의 시간 간격이 비정규로 지정되는 경우, 통합은 그룹화 함수로만 수행됩니다.

교차-일 관측값을 이전 일로 통합

1일 경계를 교차하는 시간을 사용하는 관측값이 전날의 값에 통합되는지 여부를 지정합니다. 예를 들어, 20:00시에 시작하는 8시간 노동의 시간별 관측값의 경우, 이 설정은 00:00 및 04:00 사이의 관측값이 전날 통합 결과에 포함되는지 여부를 지정합니다. 이 설정은 관측값의 시간 간격이 하루 중 시간(시), 하루 중 시간(분) 또는 하루 중 시간(초)이고 분석의 시간 간격이 일(Days)인 경우에만 적용됩니다.

지정된 필드에 대한 사용자 정의 설정

필드 기준으로 필드에 통합, 분포 및 그룹화 함수를 지정할 수 있습니다. 이 설정은 통합, 분포 및 그룹화 함수에 대한 기본 설정을 대체합니다.

스트리밍 TCM 노드 - 결측값 옵션

입력 데이터의 결측값은 대치된 값으로 바뀝니다. 다음 방법으로 바꿀 수 있습니다.

선형 보간법

선형 보간법을 사용하여 결측값을 바꿉니다. 결측값 이전의 마지막 유효한 값과 결측값 이후의 첫 번째 유효한 값이 보간법에 사용됩니다. 계열에서 첫 번째 또는 마지막 관측값에 결측값이 있는 경우, 계열의 시작 또는 종료에서 두 개의 가장 근접한 비결측 값이 사용됩니다.

계열 평균

결측값을 전체 계열에 대한 평균으로 바꿉니다.

근접한 값들의 평균

결측값을 유효한 근접 값의 평균으로 바꿉니다. 근접한 값들의 계산너비는 평균을 계산하는데 사용되는 결측값 전후의 유효값 수입니다.

근접한 값들의 중앙값

결측값을 근접한 유효한 값의 중앙값으로 바꿉니다. 근접한 값들의 계산너비는 평균을 계산하는데 사용되는 결측값 전후의 유효값 수입니다.

선형 추세

이 옵션은 단순 선형 회귀 모형을 적합시키기 위해 계열에서 모든 비결측 관측값을 사용합니다. 이 모델은 결측값을 대치하기 위해 사용됩니다.

기타 설정:

결측값의 최대 퍼센트(%)

어떤 계열에 대해서도 허용되는 최대 결측값 퍼센트를 지정합니다. 지정된 최대값보다 많은 결측값이 있는 계열은 분석에서 제외됩니다.

스트리밍 TCM 노드 - 일반 데이터 옵션

차원 필드당 최대 고유 값 수

이 설정은 다차원 데이터에 적용되며 하나의 차원 필드에 대해 허용되는 최대 고유 값 수를 지정합니다. 기본적으로, 이 한계는 10000으로 설정되지만, 임의로 큰 숫자로 증가될 수 있습니다.

스트리밍 TCM 노드 - 일반 작성 옵션

신뢰구간 너비(%)

이 설정은 예측 및 모델 모수 둘 다의 신뢰구간을 제어합니다. 100보다 작은 양수 값을 지정할 수 있습니다. 기본적으로, 95% 신뢰구간이 사용됩니다.

각 목표에 대한 최대 입력 수

이 설정은 각 목표에 대한 모델에서 허용되는 최대 입력 수를 지정합니다. 1 - 20 범위의 정수를 지정할 수 있습니다. 각 목표에 대한 모델에는 항상 자체의 시차 값이 포함되므로, 이 값을 1로 설정하면 입력만 목표 자체가 됩니다.

모델 허용 한도

이 설정은 각 목표에 대한 최상의 입력 세트를 판별하기 위해 사용되는 반복 프로세스를 제어합니다. 0보다 큰 값을 지정할 수 있습니다. 기본값은 0.001입니다.

이상값 임계값(%)

모델에서 계산된 확률(이상값)이 이 임계값을 초과하는 경우 관측값은 이상값으로 플래그가 붙습니다. 50 - 100 범위의 값을 지정할 수 있습니다.

각 입력의 시차 수

이 설정은 각 목표에 대한 모델에서 각 입력의 시차 항 수를 지정합니다. 기본적으로, 시차 항 수는

분석에서 사용되는 시간 간격에서 자동으로 결정됩니다. 예를 들어, 시간 구간이 월(증분: 한 달)인 경우 시차 수는 12입니다. 선택적으로, 시차 수를 명시적으로 지정할 수 있습니다. 지정된 값은 1 - 20 범위의 정수여야 합니다.

기존 모델을 사용하여 추정 계속

이미 시간 인과 모델을 생성한 경우, 새 모델을 작성하기 보다는 해당 모델에 대해 지정된 기준 설정을 재사용하려면 이 옵션을 선택하십시오. 이 방식에서는 이전(그러나, 최근의 데이터)과 동일한 모델 설정을 기반으로 하는 새 예측을 다시 추정하고 생성하여 시간을 절약할 수 있습니다.

스트리밍 TCM 노드 - 추정 기간 옵션

기본적으로 추정 기간은 모든 계열에 걸쳐 최초 관측값 시간에 시작되고 최근 관측값 시간에 종료됩니다.

시작 및 종료 시간 기준

추정 기간의 시작 및 종료 둘 다를 지정하거나 시작 또는 종료만 지정할 수 있습니다. 추정 기간의 시작 또는 종료를 생략하는 경우, 기본값이 사용됩니다.

- 날짜/시간 필드에 의해 관측값이 정의된 경우, 날짜/시간 필드에 사용되는 동일한 형식으로 시작 및 종료 값을 입력하십시오.
- 순환 주기에 의해 정의된 관측값의 경우, 순환 주기 필드마다 값을 지정하십시오. 각 필드는 별도의 열에 표시됩니다.

최근이거나 최초의 시간 간격(L)

선택적 오프셋으로, 데이터의 최초 시간 간격에 시작하거나 최근 시간 간격에 종료하는, 지정된 시간 간격 수로 추정 기간을 정의합니다. 이 컨텍스트에서, 시간 간격은 분석의 시간 간격을 가리킵니다. 예를 들어, 관측값이 매월 단위이지만 분석의 시간 간격은 분기일 수 있습니다. **최근**과 **시간 간격 수**로 24 값을 지정하면 최근 24개 분기를 의미합니다.

선택적으로, 지정된 시간 간격 수를 제외할 수 있습니다. 예를 들어, 최근 24 시간 간격을 지정하고 제외할 수로 1를 지정하면, 추정 기간은 마지막 구간 앞에 있는 24개 구간으로 구성됩니다.

스트리밍 TCM 노드 - 모델 옵션

모델 이름

모델에 대한 사용자 정의 이름을 지정하거나 자동으로 생성되는 이름(*TCM*)을 승인할 수 있습니다.

예측 **레코드를 미래로 확장** 옵션은 추정 기간이 끝난 이후에 예측할 시간 간격 수를 설정합니다. 이 경우의 시간 간격은 데이터 지정 사항 탭에 지정된 분석의 시간 간격입니다. 예측이 요청되면 물론 목표가 아닌 입력 계열에서 자기회귀분석 모델이 자동으로 작성됩니다. 그런 다음, 이 모델을 사용하여 예측 기간에 해당 입력 계열의 값을 생성합니다. 이 설정에 최대 한계는 없습니다.

Space-Time-Box 노드

STB(Space-Time-Box)는 Geohash 공간 위치의 확장입니다. 보다 상세하게 설명하자면 STB는 공간 및 시간을 정기적으로 모양으로 표시하는 알파뉴메릭 문자열입니다.

예를 들어, STB **dr5ru7|2013-01-01 00:00:00|2013-01-01 00:15:00**은 다음과 같은 세 부분으로 구성됩니다.

- Geohash **dr5ru7**
- 시작 시간소인 **2013-01-01 00:00:00**
- 종료 시간소인 **2013-01-01 00:15:00**

예를 들어, 공간 및 시간 정보를 사용하면 두 개의 엔티티가 동시에 동일한 공간에 시각적으로 표시되므로 동일함에 대한 신뢰도를 높일 수 있습니다. 또는 두 엔티티가 공간 및 시간의 근접성으로 인해 관련됨을 표시함으로써 관계 식별의 정확도를 높일 수 있습니다.

사용자의 요구 사항에 맞게 **개별 레코드** 또는 **행아웃(Hangout)** 모드를 선택할 수 있습니다. 두 모드 모두 다음과 같은 동일한 기본 세부사항이 필요합니다.

위도 필드. WGS84 좌표계에서 위도를 식별하는 필드를 선택하십시오.

경도 필드. WGS84 좌표계에서 경도를 식별하는 필드를 선택하십시오.

시간소인 필드. 시간 또는 날짜를 식별하는 필드를 선택하십시오.

개별 레코드 옵션

지정된 시간의 위치를 식별하기 위해 레코드에 추가 필드를 추가하려면 이 모드를 사용하십시오.

파생. 새 필드를 파생시킬 공간 및 시간의 밀도를 하나 이상 선택하십시오. 자세한 정보는 119 페이지의 『Space-Time-Box 밀도 정의』를\ 참조하십시오.

필드 이름 확장. 새 필드 이름에 추가할 확장자를 입력하십시오. 이 확장자를 **접미문자** 또는 **접두문자**로 추가할 수 있습니다.

행아웃 옵션

행아웃은 엔티티가 지속적으로 또는 반복적으로 발견될 수 있는 위치 및/또는 시간으로 생각할 수 있습니다. 예를 들어, 정기적인 운송 작업을 하는 차량을 식별하고 규범으로부터의 일탈을 식별하기 위해 사용할 수 있습니다.

행아웃 감지기는 엔티티의 움직임을 모니터링하고 엔티티가 영역에서 "행아웃"으로 관찰되는 조건에 플래그를 지정합니다. 행아웃 감지기는 플래그가 지정된 각 행아웃을 하나 이상의 STB에 자동으로 지정하고 인메모리 엔티티 및 이벤트 추적을 사용하여 가장 효율적인 행아웃을 감지합니다.

STB 밀도. 새 필드를 파생시킬 공간 및 시간의 밀도를 선택하십시오. 예를 들어, **STB_GH4_10MINS** 값은 크기가 약 20km x 20km이며 시간이 10분인 4문자 Geohash 상자 창에 해당됩니다. 자세한 정보는 119 페이지의 『Space-Time-Box 밀도 정의』를 참조하십시오.

엔티티 ID 필드. 행아웃 식별자로 사용할 엔티티를 선택하십시오. 이 ID 필드는 이벤트를 식별합니다.

최소 이벤트 수. 이벤트는 데이터의 행입니다. 행아웃으로 간주하기 위한 엔티티에 대한 이벤트 최소 발생 수를 선택하십시오. 행아웃은 **최소 체류 시간** 필드 기준도 충족해야 합니다.

최소 체류 시간. 엔티티가 동일한 위치에 체류해야 하는 최소 기간을 지정합니다. 예를 들어, 차량이 신호등 때문에 대기하는 것이 행아웃으로 간주되는 경우를 제외할 수 있습니다. 행아웃은 앞에서 설명한 **최소 이벤트 수** 필드 기준도 충족해야 합니다.

다음은 행아웃을 규정하는 것에 대한 더 자세한 정보입니다.

e_1, ..., e_n은 지속 기간(t_1, t_n) 동안 지정된 이벤트 ID에서 수신된 시간 순으로 정렬된 모든 이벤트를 표시합니다. 이러한 이벤트는 다음 경우에 행아웃으로 규정됩니다.

- $n >=$ *최소 이벤트 수*
- $t_n - t_1 >=$ *최소 체류 시간*
- 모든 이벤트 e_1, ..., e_n이 동일한 STB에서 발생

행아웃 범위가 STB 경계를 포함하도록 허용. 이 옵션이 선택되면 행아웃의 정의가 덜 엄격해집니다. 예를 들어, 둘 이상의 Space-Time-Box에 행아웃하는 엔티티를 포함할 수 있습니다. STB가 전체 시간으로 정의된 경우, 이 옵션을 선택하면 한 시간이 자정 전 30분과 자정 후 30분으로 구성되더라도 유효한 한 시간 동안 행아웃하는 엔티티로 인식합니다. 이 옵션을 선택하지 않으면 100%의 행아웃 시간이 단일 Space-Time-Box 내에 있어야 합니다.

시간상자 규정의 최소 이벤트 비율(%). 행아웃 범위가 STB 경계를 포함하도록 허용이 선택된 경우에만 사용 가능합니다. 한 STB에서 보고되는 행아웃이 실제로 다른 STB에 겹칠 수 있는 정도를 제어하는 데 사용됩니다. 행아웃을 식별하기 위해 단일 STB 내에서 발생해야 하는 최소 이벤트 비율을 선택하십시오. 25%로 설정한 상태에서 이벤트의 비율이 26%이면 이는 행아웃으로 규정됩니다.

예를 들어, 4바이트 Geohash 공간 상자 및 10분 시간 상자(STB_NAME = STB_GH4_10MINS) 내에 두 개 이상의 이벤트(최소 이벤트 수 = 2) 및 2분 이상의 연속 체류 시간을 요구하도록 행아웃 감지기를 구성한 경우를 가정해 보십시오. 행아웃이 발견될 때 4:57pm 및 5:07pm 사이의 10분의 시간 범위 내에서 4:58pm, 5:01pm 및 5:03pm에 세 개의 규정 이벤트가 발생하는 동안 엔티티가 동일한 4바이트 Geohash 공간 상자에 체류한다고 말할 수 있습니다. 규정 시간 상자 퍼센트 값은 STB가 행아웃으로 간주되기 위한 퍼센트 값을 다음과 같이 지정합니다.

- **100%.** 행아웃이 5:00 - 5:10pm 시간 상자에서는 보고되고 4:50 - 5:00pm 시간 상자에서는 보고되지 않습니다(5:01pm 및 5:03pm의 이벤트는 행아웃을 규정하는 데 필요한 모든 조건을 충족하며 이러한 이벤트의 100%가 5:00 - 5:10 시간 상자에서 발생했습니다).
- **50%.** 두 시간 상자 모두에서 행아웃이 보고됩니다(5:01pm 및 5:03pm의 이벤트는 행아웃을 규정하는 데 필요한 모든 조건을 충족하며 이러한 이벤트 중 50% 이상이 4:50 - 5:00 시간 상자에서 발생하고 이러한 이벤트 중 50% 이상이 5:00 - 5:10 시간 상자에서 발생했습니다).
- **0%.** 두 시간 상자 모두에서 행아웃이 보고됩니다.

0%를 지정하면 행아웃 보고서에 규정 기간이 속한 모든 시간 상자를 나타내는 STB가 포함됩니다. 규정 기간은 STB 내의 시간 상자의 해당 기간 이하여야 합니다. 즉, 10분 STB가 20분의 규정 기간과 함께 구성될 수 없습니다.

행아웃은 규정 조건이 충족되자 마자 보고되며 STB당 두 번 이상 보고되지 않습니다. 세 개의 이벤트가 행아웃을 규정하고 총 열 개의 이벤트가 모두 동일한 STB 내의 규정 기간 동안 발생했다고 가정해 보십시오. 이 경우, 세 번째 규정 이벤트가 발생할 때 행아웃이 보고됩니다. 추가 일곱 개의 이벤트는 행아웃 보고서를 트리거하지 않습니다.

참고:

- 행아웃 감지기의 인메모리 이벤트 데이터는 프로세스에 걸쳐 공유되지 않습니다. 따라서 특정 엔티티가 특정 행아웃 감지기 노드와 연관관계를 갖습니다. 즉, 엔티티에 대해 수신되는 이동 데이터는 항상 해당 엔티티를 추적하는 행아웃 감지기 노드(일반적으로 실행 전체에서 동일 노드임)에 일관적으로 전달되어야 합니다.
- 행아웃 감지기의 인메모리 이벤트 데이터는 일시적입니다. 행아웃 감지기가 종료되고 다시 시작되면 모든 진행 중인 행아웃이 손실됩니다. 즉, 프로세스를 중지하고 다시 시작하면 시스템이 실제 행아웃을 손실하게 됩니다. 잠재적인 해결 방법은 히스토리 이동 데이터의 일부를 재생하는 것입니다. 예를 들어, 48시간 뒤로 이동하여 재시작된 임의의 노드에 적용 가능한 이동 레코드를 재생하는 것입니다.
- 시간 순서로 행아웃 감지기에 데이터를 공급해야 합니다.

Space-Time-Box 밀도 정의

각각 포함할 실제 영역 및 경과 시간을 지정하여 STB(Space-Time-Box)의 *크기*(밀도)를 지정할 수 있습니다.

지역 밀도. 각 STB에 포함할 영역의 *크기*를 선택하십시오.

시간 간격. 각 STB에 포함할 시간 수를 선택하십시오.

필드 이름. STB라는 접두문자가 사용되고 앞의 두 필드의 선택사항을 기준으로 자동으로 완료됩니다.

제 4 장 필드 작업 노드

필드 작업 개요

초기 데이터 탐색 후에는 분석을 준비하기 위해 데이터를 선택하거나 정리하거나 구성해야 할 수 있습니다. 필드 작업 팔레트에는 이 변환 및 준비를 위해 유용한 다수의 노드가 포함되어 있습니다.

예를 들어, 파생 노드를 사용하면 데이터에서 현재 표시되지 않는 속성을 작성할 수 있습니다. 또는 구간화 노드를 사용하여 목표로 지정된 분석을 위해 자동으로 필드 값의 코딩을 변경할 수 있습니다. 유형 노드를 자주 사용하고 있음을 알 수 있으며 이를 통해 데이터 세트의 각 필드에 대한 측정 수준 값 및 모델링 역할을 지정할 수 있습니다. 해당 조작은 결측값 및 다운스트림 모델링 처리를 위해 유용합니다.

필드 작업 팔레트에는 다음과 같은 노드가 포함되어 있습니다.

자동 데이터 준비(ADP) 노드는 데이터를 분석하고 수정사항을 식별하고, 문제가 있거나 유용할 것 같지 않은 필드를 제외시키고, 적절한 경우 새 속성을 파생시키고, 지능형 선별 및 표본추출 기법을 통해 성능을 개선할 수 있습니다. 완전 자동화된 방식으로 노드를 사용하여 노드가 수정사항을 선택하고 적용할 수 있게 하거나, 변경사항이 작성 및 승인되기 전에 변경을 미리보거나, 거부 또는 원하는 대로 개정할 수 있습니다.

유형 노드는 필드 메타데이터 및 특성을 지정합니다. 예를 들어 각 필드에 대한 측정 수준(연속형, 명목형, 순서형 또는 플래그)을 지정하고, 결측값 및 시스템 널 처리를 위한 옵션을 설정하고, 모델링 목적으로 필드의 역할을 설정하고, 필드와 값 레이블을 지정하고, 필드의 값을 지정할 수 있습니다.

필터 노드는 필드를 필터링(삭제)하고, 필드 이름을 변경하고 한 소스에서 다른 소스로 필드를 맵핑합니다.

파생 노드는 데이터 값을 수정하거나 하나 이상의 기존 필드로부터 새 필드를 작성합니다. 수식, 플래그, 명목형, 상태, 개수, 조건부 유형의 필드를 작성합니다.

앙상블 노드는 둘 이상의 모델 너깃을 결합하여 임의의 한 모델에서 얻을 수 있는 것보다 정확한 예측을 얻습니다.

채움 노드는 필드 값을 대체하고 저장 공간을 변경합니다. `@BLANK(@FIELD)` 같은 CLEM 조건을 기반으로 값을 대체할 수 있습니다. 또는 모든 공백 또는 널값을 특정 값으로 대체할 것을 선택할 수 있습니다. 채움 노드는 종종 유형 노드와 함께 사용하여 결측값을 대체합니다.

값 익명화 노드는 필드 이름 및 값이 다운스트림으로 표시되는 방법을 변환하여 원 데이터를 위장합니다. 이것은 다른 사용자가 고객 이름이나 기타 세부사항 같은 민감한 데이터를 사용하여 모델을 작성하도록 허용하려는 경우에 유용할 수 있습니다.

재분류 노드는 한 세트의 범주형 값을 다른 값으로 변환합니다. 재분류는 분석을 위해 범주를 접거나 데이터를 재그룹화하는 데 유용합니다.

구간화 노드는 하나 이상의 기존 연속형(숫자 범위) 필드의 값을 기반으로 새 명목형(세트) 필드를 자동으로 작성합니다. 예를 들어, 연속형 수입 필드를 평균값에서의 편차로서 수입 그룹을 포함하는 새 범주형 필드로 변환할 수 있습니다. 새 필드에 대한 구간을 작성한 후에는 절단점을 기반으로 파생 노드를 생성할 수 있습니다.

RFM(Recency, Frequency, Monetary) 분석 노드를 사용하면 얼마나 최근에 사용자로부터 구매했는지(최근성), 얼마나 자주 구매했는지(빈도) 및 모든 트랜잭션에서 얼마나 소비했는지(구매총액)를 조사하여 최고의 고객이 될 수 있는 고객을 정량적으로 판별할 수 있습니다.

파티션 노드는 파티션 필드를 생성하는데, 이 필드는 모델 작성의 학습, 검정, 검증 단계를 위한 별개의 서브세트로 데이터를 분할합니다.

플래그로 설정 노드는 하나 이상의 명목 필드에 대해 정의된 범주형 값을 바탕으로 다중 플래그 필드를 파생시킵니다.

구조변환 노드는 명목 또는 플래그 필드를 아직 또 다른 필드의 값으로 채워질 수 있는 필드 그룹으로 변환합니다. 예를 들어, *payment type*이라는 이름의 필드와 *credit*, *cash*, *debit*의 값이 주어진 경우, 각각이 실제 이루어진 지불의 값을 포함할 수 있는 세 개의 새 필드(*credit*, *cash*, *debit*)가 작성됩니다.

전치 노드는 행 및 열의 데이터를 바꿔서 레코드가 필드가 되고 필드가 레코드가 되게 합니다.

구간을 지정하고 추정 또는 시계열 분석을 위한 새 시간 필드를 파생하려면 시간 간격 노드를 사용하십시오. 초부터 년까지, 모든 범위의 시간 간격이 지원됩니다.

히스토리 노드는 이전 레코드의 필드에 있는 데이터를 포함하는 새 필드를 작성합니다. 히스토리 노드는 시계열 데이터 같은 순차 데이터에 가장 자주 사용합니다. 히스토리 노드를 사용하기 전에 정렬 노드를 사용하여 데이터를 정렬할 수 있습니다.

필드 다시 정렬 노드는 필드를 다운스트림으로 표시하는 데 사용하는 기본 순서를 정의합니다. 이 순서는 테이블, 목록 및 필드 선택기 같은 다양한 장소에서 필드의 표시에 영향을 줍니다. 이 작업은 관심있는 필드를 더 잘 보이게 만들기 위해 넓은 데이터 세트에 대해 작업할 때 유용합니다.

SPSS Modeler 내에서 표현식 작성기 공간 함수, STP(Spatio-Temporal Prediction) 노드, 맵 시각화 노드 같은 항목이 투영된 좌표계를 사용합니다. 지리적 좌표계를 사용하고 사용자가 가져오는 임의의 데이터의 좌표계를 변경하려면 재투영 노드를 사용하십시오.

이 노드 중 몇몇 노드는 데이터 검토 노드에 의해 작성된 감사 보고서에서 직접 생성될 수 있습니다. 자세한 정보는 324 페이지의 『데이터 준비를 위해 기타 노드 생성』의 내용을 참조하십시오.

자동 데이터 준비

분석을 위한 데이터 준비는 모든 프로젝트에서 가장 중요한 단계 중 하나이며 일반적으로 가장 많은 시간이 소요되는 단계 중 하나입니다. 자동 데이터 준비(ADP)는 데이터 분석, 수정사항 식별, 문제가 있거나 유용할 것 같지 않은 필드 필터링, 적절한 경우 새 속성 파생 및 지능형 선별 기술을 통한 성능 개선 등의 작업을 자동으로 처리합니다. 완전히 **자동화된** 방식으로 알고리즘을 사용하여 알고리즘이 수정사항을 선택하고 적용할 수 있도록 하거나, **대화식** 방식으로 알고리즘을 사용하여 변경을 수행하기 전에 변경사항을 미리보고 원하는 바에 따라 변경사항을 수락하거나 거부할 수 있습니다.

ADP를 사용하면 관련된 통계 개념에 대한 사전 지식 없이 모델 작성을 위한 데이터를 쉽고 빠르게 준비할 수 있습니다. 모델은 더 빨리 작성되고 스코어링되는 경향이 있습니다. 이에 더하여 ADP를 사용하면 자동화된 모델링 프로세스(예: 모델 새로 고치기 및 챔피언/도전자)의 강력함이 더욱 향상됩니다.

참고: ADP는 분석을 위해 필드를 준비할 때 이전 필드의 기존 값 및 특성을 대체하는 대신 조정 또는 변환을 포함하는 새 필드를 작성합니다. 이전 필드는 추가 분석에서 사용되지 않습니다(해당 역할이 없음으로 설정됨).

예. 자택 소유자의 보험 청구를 조사하기 위한 제한된 자원을 가진 보험 회사가 사기일 가능성이 높은 의심스러운 보험 청구를 플래그 지정하는 모델을 작성하려 합니다. 모델을 작성하기 전에 보험 회사는 자동 데이터 준비를 사용하여 모델링을 위한 데이터를 준비합니다. 변환이 적용되기 전에 제안된 변환을 검토할 수 있기를 원하므로 대화식 모드에서 자동 데이터 준비를 사용합니다.

자동차 산업 그룹은 다양한 개인용 자동차의 판매량을 추적합니다. 과성능 모델과 성능 미달 모델을 식별하기 위한 노력으로 자동차 판매량과 자동차 특성 사이의 관계를 설정하려 합니다. 자동 데이터 준비를 사용하여 분석을 위한 데이터를 준비하고 결과가 어떻게 다른지 확인하기 위해 준비 "전"과 "후"의 데이터를 사용하여 모델을 작성합니다.

원하는 목적 자동 데이터 준비는 다른 알고리즘이 모델을 작성하는 속도에 영향을 주고 그러한 모델의 예측력을 향상시키는 데이터 준비 단계를 권장합니다. 여기에는 기능 변환, 생성 및 선택이 포함됩니다. 목표도 변환할 수 있습니다. 데이터 준비 프로세스가 집중해야 하는 모델 작성 우선순위를 지정할 수 있습니다.

- **속도와 정확도의 균형.** 이 옵션은 모델 작성 알고리즘이 데이터를 처리하는 속도와 예측의 정확도 둘 다에 동일한 우선순위를 부여하도록 데이터를 준비합니다.

- **속도 최적화.** 이 옵션은 모델 작성 알고리즘이 데이터를 처리하는 속도에 우선순위를 부여하도록 데이터를 준비합니다. 매우 큰 데이터 세트를 사용하여 작업하거나 빠른 해답을 찾으려면 이 옵션을 선택하십시오.
- **정확도 최적화.** 이 옵션은 모델 작성 알고리즘이 생성하는 예측의 정확도에 우선순위를 부여하도록 데이터를 준비합니다.
- **사용자 정의 분석.** 설정 탭에서 알고리즘을 수동으로 변경하려면 이 옵션을 선택하십시오. 이후 설정 탭에서 다른 목적 중 하나와 호환되지 않는 옵션을 변경하면 이 설정이 자동으로 선택됩니다.

노드 훈련

ADP 노드는 프로세스 노드로 구현되고 유형 노드와 유사한 방식으로 작동합니다. ADP 노드를 **훈련**하는 것은 유형 노드를 인스턴스화하는 것에 해당합니다. 일단 분석이 수행되면, 업스트림 데이터 모델이 변경되지 않는 한 추가 분석 없이 지정된 변환이 데이터에 적용됩니다. 유형 및 필터 노드와 마찬가지로, ADP 노드는 다시 연결될 때 다시 훈련하지 않아도 되도록 연결이 끊어질 때 데이터 모델 및 변환을 기억합니다. 이로 인해 일반 데이터 서브세트를 대상으로 ADP 노드를 훈련한 후 ADP 노드를 복사하거나 배포하여 필요한 만큼 자주 실시간 데이터를 대상으로 ADP 노드를 사용할 수 있습니다.

도구 모음 사용

도구 모음을 사용하여 데이터 분석을 실행하고 표시를 업데이트하며 원래 데이터와 관련하여 사용할 수 있는 노드를 생성할 수 있습니다.

- **생성** 이 메뉴에서 필터 또는 파생 노드를 생성할 수 있습니다. 이 메뉴는 분석 탭에 분석이 표시된 경우에만 사용할 수 있습니다.

  필터 노드는 변환된 입력 필드를 제거합니다. 원래 입력 필드를 데이터 세트에 남겨 두도록 ADP 노드를 구성하면 원래 입력 세트가 복원되고, 이로 인해 입력의 관점에서 스코어 필드를 해석할 수 있습니다. 예를 들어, 이는 다양한 입력에 대한 스코어 필드 그래프를 생성하려는 경우에 유용할 수 있습니다.

  파생 노드는 원래 데이터 세트 및 목표 단위를 복원할 수 있습니다. ADP 노드에 범위 목표의 척도를 조정하는 분석이 있는 경우(즉, 입력 및 출력 준비 패널에서 Box-Cox 척도 조정을 선택한 경우)에만 파생 노드를 생성할 수 있습니다. 목표가 범위가 아니거나 Box-Cox 척도 조정을 선택하지 않은 경우에는 파생 노드를 생성할 수 없습니다. 자세한 정보는 137 페이지의 『파생 노드 생성』의 내용을 참조하십시오.
- **보기**에는 분석 탭에 표시되는 내용을 제어하는 옵션이 있습니다. 여기에는 그래프 편집 제어와 기본 패널 및 링크된 보기 둘 다에 대한 표시 선택사항이 포함됩니다.
- **미리보기**: 입력 데이터에 적용될 변환 표본을 표시합니다.
- **데이터 분석**: 현재 설정을 사용하여 분석을 시작하고 분석 탭에 결과를 표시합니다.
- **분석 지우기**: 기존 분석을 지웁니다(현재 분석이 있는 경우에만 사용 가능).

노드 상태

IBM SPSS Modeler 캔버스에서 ADP 노드의 상태는 분석이 수행되었는지 여부를 나타내는 아이콘 위의 화살표 또는 체크로 표시됩니다.

필드 탭

모델을 작성하려면 먼저 목표 및 입력으로 사용할 필드를 지정해야 합니다. 몇 가지 예외가 있지만, 모든 모델링 노드는 업스트림 유형 노드에서 필드 정보를 사용합니다. 유형 노드를 사용하여 입력 및 목표 필드를 선택하는 경우 이 탭의 내용을 변경하지 않아도 됩니다.

유형 노드 설정 사용. 이 옵션은 업스트림 유형 노드의 필드 정보를 사용하도록 노드에 지시합니다. 기본값입니다.

사용자 정의 설정 사용. 이 옵션은 업스트림 유형 노드에 제공된 정보 대신 여기에 지정된 필드 정보를 사용하도록 노드에 알립니다. 이 옵션을 선택한 후 필요하면 아래 필드를 지정합니다.

목표. 하나 이상의 목표 필드가 필요한 모델의 경우 하나 이상의 목표 필드를 선택합니다. 유형 노드에서 필드 역할을 *목표*로 설정하는 것과 유사합니다.

입력. 하나 이상의 입력 필드를 선택합니다. 유형 노드에서 필드 역할을 *입력*으로 설정하는 것과 유사합니다.

설정 탭

설정 탭에는 서로 다른 여러 설정 그룹이 포함되며, 이러한 설정을 수정하여 알고리즘이 데이터를 처리하는 방법을 미세 조정할 수 있습니다. 다른 목적과 호환되지 않는 기본 설정을 변경하면 목적 탭이 자동으로 업데이트되어 **분석 사용자 정의** 옵션을 선택합니다.

필드 설정

빈도 필드 사용. 이 옵션에서는 빈도 가중치로 필드를 선택할 수 있습니다. 훈련 데이터의 레코드가 각각 둘 이상의 단위를 나타내는 경우(예: 통합 데이터를 사용하는 경우) 이 옵션을 사용하십시오. 필드 값은 각 레코드로 나타낸 노드 수여야 합니다.

가중 필드 사용. 이 옵션에서는 케이스 가중치로 필드를 선택할 수 있습니다. 케이스 가중치는 출력 필드의 수준에서 분산의 차이를 설명하는 데 사용됩니다.

모델링에서 제외된 필드 처리 방법. 제외된 필드에 적용되는 조치를 지정하십시오. 제외된 필드를 데이터에서 필터링하거나 단순히 해당 *역할*을 **없음**으로 설정할 수 있습니다.

참고: 이 조치는 목표가 변환되는 경우 목표에도 적용됩니다. 예를 들어, 목표의 새 파생된 버전이 **목표** 필드로 사용되는 경우 원래 목표는 필터링되거나 **없음**으로 설정됩니다.

수신 필드가 기존 분석과 일치하지 않는 경우. 훈련된 ADP 노드를 실행할 때 수신 데이터 세트에서 하나 이상의 필수 입력 필드가 누락된 경우에 발생하는 상황을 지정하십시오.

- **실행을 중지하고 기존 분석 유지.** 실행 프로세스를 중지하고 현재 분석 정보를 유지하며 오류를 표시합니다.
- **기존 분석을 지우고 새 데이터 분석.** 기존 분석을 지우고 수신 데이터를 분석하며 해당 데이터에 권장되는 변환을 적용합니다.

날짜 및 시간 준비

다수의 모델링 알고리즘은 날짜 및 시간 세부사항을 직접 처리할 수 없습니다. 이러한 설정을 사용하면 기존 데이터의 날짜 및 시간에서 모델 입력으로 사용할 수 있는 새 지속 기간 데이터를 파생시킬 수 있습니다. 날짜 및 시간을 포함하는 필드는 날짜 또는 시간 저장 유형으로 사전 정의해야 합니다. 원래 날짜 및 시간 필드는 자동 데이터 준비를 따르는 모델 입력으로 권장되지 않습니다.

모델링을 위한 날짜 및 시간 준비. 이 옵션을 선택 취소하면 기타 모든 날짜 및 시간 준비 제어가 사용되지 않습니다(단, 선택항목은 유지보수됨).

참조 날짜까지의 경과 시간 계산. 날짜를 포함하는 각 변수의 참조 날짜 이후 경과한 년/월/일 수를 생성합니다.

- **참조 날짜.** 입력 데이터의 날짜 정보와 관련하여 지속 기간을 계산할 시작 날짜를 지정하십시오. **오늘의 날짜**를 선택하면 ADP를 실행할 때 현재 시스템 날짜가 항상 사용됩니다. 특정 날짜를 선택하려면 **고정 날짜**를 선택하고 필요한 날짜를 입력하십시오. 노드를 처음 작성할 때는 **고정 날짜** 필드에 현재 날짜가 자동으로 입력됩니다.
- **날짜 지속 기간 단위.** ADP가 날짜 지속 기간 단위를 자동으로 결정하는지 여부를 지정하거나 **고정 단위** 년, 월 및 일 중에서 선택하십시오.

참조 시간까지의 경과 시간 계산. 시간을 포함하는 각 변수의 참조 시간 이후 경과한 시/분/초 수를 생성합니다.

- **참조 시간.** 입력 데이터의 시간 정보와 관련하여 지속 기간을 계산할 시작 시간을 지정하십시오. **현재 시간**을 선택하면 ADP를 실행할 때 현재 시스템 시간이 항상 사용됩니다. 특정 시간을 선택하려면 **고정 시간**을 선택하고 필요한 세부사항을 입력하십시오. 노드를 처음 작성할 때는 **고정 시간** 필드에 현재 시간이 자동으로 입력됩니다.
- **시간 지속 기간 단위.** ADP가 시간 지속 기간 단위를 자동으로 결정하는지 여부를 지정하거나 **고정 단위** 시, 분 및 초 중에서 선택하십시오.

순환 시간 요소 추출. 이러한 설정을 사용하여 단일 날짜 또는 시간 필드를 하나 이상의 필드로 분할하십시오. 예를 들어, 날짜 선택란을 세 개 모두 선택하는 경우, 입력 날짜 필드 "1954-05-23"은 세 개의 필드 1954, 5 및 23으로 분할되고 각 필드는 **필드 이름** 패널에 정의된 접미부를 사용하며 원래 날짜 필드는 무시됩니다.

- **날짜에서 추출.** 모든 날짜 입력에 대해 년, 월, 일 또는 임의 조합을 추출할지 여부를 지정하십시오.
- **시간에서 추출.** 모든 시간 입력에 대해 시, 분, 초 또는 임의 조합을 추출할지 여부를 지정하십시오.

필드 제외

품질이 낮은 데이터는 예측 정확도에 영향을 미칠 수 있습니다. 따라서 입력 기능에 대해 허용 가능한 품질 수준을 지정할 수 있습니다. 상수이거나 100% 결측값을 갖는 모든 필드는 자동으로 제외됩니다.

저품질 입력 필드 제외. 이 옵션을 선택 취소하면 기타 모든 필드 제외 제어가 사용되지 않습니다(단, 선택항목은 유지보수됨).

결측값이 너무 많은 필드 제외. 결측값이 지정된 백분율보다 더 많은 필드는 추가 분석에서 제거됩니다. 0(이 옵션을 선택 취소하는 것과 동등함)보다 크거나 같고 100(결측값이 100%인 필드는 자동으로 제외됨)보다 작거나 같은 값을 지정하십시오. 기본값은 50입니다.

고유 범주가 너무 많은 명목 필드 제외. 범주가 지정된 수보다 더 많은 명목 필드는 추가 분석에서 제거됩니다. 양수를 지정하십시오. 기본은 100입니다. 이 옵션은 모델링에서 ID, 주소, 이름 등의 레코드 고유 정보를 포함하는 필드를 자동으로 제거하는 데 유용합니다.

단일 범주에 너무 많은 값이 있는 범주형 필드 제외. 범주에 지정된 백분율보다 더 많은 레코드가 있는 순서 및 명목 필드는 추가 분석에서 제거됩니다. 0(이 옵션을 선택 취소하는 것과 동등함)보다 크거나 같고 100(상수 필드는 자동으로 제외됨)보다 작거나 같은 값을 지정하십시오. 기본값은 95입니다.

입력 및 목표 준비

처리하기에 완벽한 상태인 데이터는 없으므로 분석을 실행하기 전에 일부 설정을 조정하려 할 수 있습니다. 예를 들어, 이러한 조정에 이상값 제거, 결측값 처리 방법 지정 또는 유형 조정 등이 포함될 수 있습니다.

참고: 이 패널의 값을 변경하는 경우, **목적** 탭이 자동으로 업데이트되어 **사용자 정의 분석** 옵션이 선택됩니다.

모델링을 위한 입력 및 목표 필드 준비. 패널의 모든 필드를 켜거나 끕니다.

유형 조정 및 데이터 품질 개선. 입력 및 목표에 대해 몇 가지 데이터 변환을 개별적으로 지정할 수 있습니다(목표의 값을 변경하지 않으려는 이유에서). 예를 들어, 달러 단위의 수입 예측이 log(dollars)로 측정된 예측보다 더 의미가 있습니다. 또한, 목표에 결측값이 있는 경우 결측값을 채워도 예측에 아무런 이점이 없지만, 입력의 결측값을 채우는 경우에는 일부 알고리즘이 결측값을 채우지 않은 경우에 유실될 수 있는 정보를 처리할 수도 있습니다.

이러한 변환을 위한 추가 설정(예: 이상값 절사 값)은 목표 및 입력 모두에 공통됩니다.

입력, 목표 또는 둘 다에 대해 다음 설정을 선택할 수 있습니다.

- **숫자 필드의 유형 조정.** 측정 수준이 *순서*인 숫자 필드를 *연속형*으로 또는 그 반대로 변환할 수 있는지 결정하려면 이를 선택하십시오. 변환을 제어하는 최소 및 최대 임계값을 지정할 수 있습니다.
- **명목 필드 다시 정렬.** 명목(세트) 필드를 가장 작은 범주에서 가장 큰 범주로 순서대로 정렬하려면 이를 선택하십시오.
- **연속형 필드의 이상값 대체.** 이상값을 대체할지 여부를 지정하십시오. 이 옵션을 아래의 **이상값 대체 방법** 옵션과 함께 사용하십시오.
- **연속형 필드: 결측값을 평균으로 대체.** 연속형(범위) 기능의 결측값을 대체하려면 이를 선택하십시오.
- **명목 필드: 결측값을 모드로 대체.** 명목(세트) 기능의 결측값을 대체하려면 이를 선택하십시오.
- **순서 필드: 결측값을 중앙값으로 대체.** 순서(순서형 세트) 기능의 결측값을 대체하려면 이를 선택하십시오.

순서 필드의 값 최대 수. 순서(순서형 세트) 필드를 연속형(범위)으로 재정의하기 위한 임계값을 지정하십시오. 기본값은 10입니다. 따라서, 순서 필드에 10개가 넘는 범주가 있는 경우 이 필드는 연속형(범위)으로 재정의됩니다.

연속형 필드의 값 최소 수. 척도 또는 연속형(범위) 필드를 순서(순서형 세트)로 재정의하기 위한 임계값을 지정하십시오. 기본값은 5입니다. 따라서, 연속형 필드에 5개 미만의 값이 있으면 이 필드는 순서(순서형 세트)로 재정의됩니다.

이상값 절사 값. 표준 편차에서 측정되는 이상값 절사 기준을 지정하십시오. 기본값은 3입니다.

이상값 대체 방법. 이상값을 자르기(강제 적용)를 통해 절사 값으로 대체할지 또는 이상값을 삭제하고 결측값으로 설정할지 여부를 선택하십시오. 결측값으로 설정된 이상값은 위에서 선택한 결측값 처리 설정을 따릅니다.

모든 연속형 입력 필드를 일반 척도에 둠. 연속형 입력 필드를 정규화하려면 이 선택란을 선택하고 정규화 방법을 선택하십시오. 기본값은 **z-스코어 변환**입니다. 이 방법에서는 기본값이 0인 **최종 평균**과 기본값이 1인 **최종 표준 편차**를 지정할 수 있습니다. 또는 **최소/최대 변환**을 사용하도록 선택하고 최소 및 최대 값(기본값은 각각 0과 100임)을 지정할 수 있습니다.

이 필드는 특히 기능 생성 및 선택 패널에서 **기능 생성 수행**을 선택할 때 유용합니다.

Box-Cox 변환으로 연속형 대상 척도 조정. 연속형(척도 또는 범위) 목표 필드를 정규화하려면 이 선택란을 선택하십시오. Box-Cox 변환에서 **최종 평균**의 기본값은 0이고 **최종 표준 편차**의 기본값은 1입니다.

참고: 목표를 정규화하기로 선택하면 목표의 차원이 변환됩니다. 이 경우, 추가 처리를 위해 변환된 단위를 다시 인식 가능한 형식으로 되돌리기 위해 역변환을 적용할 파생 노드를 생성해야 할 수도 있습니다. 자세한 정보는 137 페이지의 『파생 노드 생성』의 내용을 참조하십시오.

생성 및 필드선택

데이터의 예측력을 향상시키기 위해 입력 필드를 변환하거나 기존 필드를 기반으로 새 입력 필드를 생성할 수 있습니다.

참고: 이 패널의 값을 변경하는 경우, **목적** 탭이 자동으로 업데이트되어 **사용자 정의 분석** 옵션이 선택됩니다.

예측력 향상을 위해 입력 필드 변환, 생성 및 선택. 패널의 모든 필드를 켜거나 끕니다.

성긴 범주를 병합하여 목표와의 연관 최대화. 목표와 연관하여 처리할 변수의 수를 줄여 보다 경제적인 모델을 작성하려면 이 옵션을 선택하십시오. 필요한 경우, 확률값을 기본값 0.05에서 변경하십시오.

모든 범주가 하나로 병합되는 경우, 필드의 원래 버전과 파생된 버전은 예측변수로서 값이 없기 때문에 제외됩니다.

목표가 없는 경우 개수를 기반으로 성긴 범주 병합. 목표가 없는 데이터를 처리하는 경우, 순서(순서형 세트) 기능, 명목(세트) 기능 또는 둘 다의 성긴 범주를 병합하도록 선택할 수 있습니다. 병합할 범주를 식별하는 데이터에 케이스(또는 레코드)의 최소 백분율을 지정하십시오. 기본값은 10입니다.

다음 규칙에 따라 범주가 병합됩니다.

- 2진 필드에 대해서는 병합이 수행되지 않습니다.
- 병합 동안 범주가 두 개뿐인 경우에는 병합이 중지합니다.

- 원래 범주가 없고 병합 중에 범주가 작성되지 않으며 지정된 케이스 최소 백분율보다 적으면 병합이 중지합니다.

예측력을 유지하면서 연속형 필드 구간화. 범주형 대상이 포함된 데이터가 있는 경우, 처리 성능을 향상시키기 위해 강력한 연관이 있는 연속형 입력을 구간화할 수 있습니다. 필요한 경우, 동종 서브세트의 확률값을 기본값 0.05에서 변경하십시오.

구간화 조작으로 특정 필드의 단일 구간이 생성되는 경우, 필드의 원래 버전과 구간화된 버전은 예측변수로서 값이 없기 때문에 제외됩니다.

참고: ADP에서의 구간화는 IBM SPSS Modeler의 다른 파트에서 사용되는 최적 구간화와 다릅니다. 최적 구간화는 엔트로피 정보를 사용하여 연속형 변수를 범주형 변수로 변환합니다. 이 구간화는 데이터를 정렬하고 모두 메모리에 저장해야 합니다. ADP는 동종 서브세트를 사용하여 연속형 변수를 구간화합니다. 이는 ADP 구간화는 데이터를 정렬하지 않아도 되며 모든 데이터를 메모리에 저장하지 않음을 의미합니다. 동종 서브세트를 사용하여 연속형 변수를 구간화하는 방법은 구간화 후의 범주 수가 항상 목표 범주 수 이하임을 의미합니다.

변수 선택 수행. 상관 계수가 낮은 기능을 제거하려면 이 옵션을 선택하십시오. 필요한 경우, 확률값을 기본값 0.05에서 변경하십시오.

이 옵션은 목표가 연속형인 연속형 입력 기능과 범주형 입력 기능에만 적용됩니다.

기능 생성 수행. 기존의 여러 기능의 조합(이후 모델링에서 삭제됨)에서 새 기능을 파생시키려면 이 옵션을 선택하십시오.

이 옵션은 목표가 연속형이거나 목표가 없는 연속형 입력 기능에만 적용됩니다.

필드 이름

변환된 새 기능을 쉽게 식별하기 위해 ADP는 기본 새 이름, 접두부 또는 접미부를 작성하고 적용합니다. 이러한 이름을 사용자 요구 및 데이터와 연관성이 더 큰 이름으로 수정할 수 있습니다. 다른 레이블을 지정하려면 다운스트림 유형 노드에서 이를 수행해야 합니다.

변환 및 생성 필드. 변환된 목표 및 입력 필드에 적용할 이름 확장자를 지정하십시오.

ADP 노드에서 문자열 필드를 아무것도 포함하지 않도록 설정하면 사용하지 않는 필드 처리를 위해 선택한 방법에 따라 오류가 발생할 수 있습니다. 설정 탭의 필드 설정 패널에서 **모델링에서 제외된 필드 처리 방법**을 **사용하지 않는 필드 필터링**으로 설정한 경우, 입력 및 목표의 이름 확장자를 빈 값으로 설정할 수 있습니다. 원래 필드는 필터링되고 변환된 필드가 그 대신 저장됩니다. 이 경우 변환된 새 필드는 원래 필드와 동일한 이름을 가집니다.

사용하지 않는 필드의 방향을 '없음'으로 설정을 선택한 경우, 목표 및 입력에 빈(또는 널) 이름 확장자를 사용하면 중복 필드 이름을 작성하게 되므로 오류가 발생합니다.

또한, 선택 및 생성 설정을 통해 생성되는 기능에 적용할 접두부 이름을 지정하십시오. 이 접두부 루트 이름에 숫자 접미부를 첨부하여 새 이름이 작성됩니다. 숫자의 형식은 파생되는 새 기능의 수에 따라 다릅니다. 예를 들면, 다음과 같습니다.

- 1-9개의 생성된 기능은 feature1 - feature9으로 이름이 지정됩니다.
- 10-99개의 생성된 기능은 feature01 - feature99으로 이름이 지정됩니다.
- 100-999개의 생성된 기능은 feature001 - feature999으로 이름이 지정되며, 계속해서 이와 같이 반복됩니다.

이 방식을 사용하면 생성된 기능을 기능의 수와 상관없이 합리적인 순서로 정렬할 수 있습니다.

날짜 및 시간에서 계산된 지속 기간. 날짜 및 시간 둘 다에서 계산된 지속 기간에 적용할 이름 확장자를 지정하십시오.

날짜 및 시간에서 추출된 순환 요소. 날짜 및 시간 둘 다에서 추출된 순환 요소에 적용할 이름 확장자를 지정하십시오.

분석 탭

1. 목적, 필드 및 설정 탭에서 수행된 변경을 포함하여 ADP 설정에 만족하는 경우 **데이터 분석**을 클릭하십시오. 알고리즘이 해당 설정을 데이터 입력에 적용하고 분석 탭에 결과를 표시합니다.

분석 탭은 데이터 처리를 요약하는 테이블 및 그래프 형식 출력을 모두 포함하며 스코어링을 위한 데이터 수정 또는 개선 방법과 관련하여 권장사항을 표시합니다. 그러면 그러한 권장사항을 검토하여 수락하거나 거부할 수 있습니다.

분석 탭은 두 개의 패널(왼쪽에 있는 기본 보기와 오른쪽에 있는 링크된(또는 보조) 보기)로 구성됩니다. 세 개의 기본 보기가 있습니다.

- 필드 처리 요약(기본값). 자세한 정보는 131 페이지의 『필드 처리 요약』의 내용을 참조하십시오.
- 필드. 자세한 정보는 131 페이지의 『필드』의 내용을 참조하십시오.
- 조치 요약. 자세한 정보는 133 페이지의 『조치 요약』의 내용을 참조하십시오.

4개의 링크/보조 보기가 있습니다.

- 예측력(기본값). 자세한 정보는 133 페이지의 『예측력』의 내용을 참조하십시오.
- 필드 테이블. 자세한 정보는 133 페이지의 『필드 테이블』의 내용을 참조하십시오.
- 필드 세부사항. 자세한 정보는 134 페이지의 『필드 세부사항』의 내용을 참조하십시오.
- 조치 세부사항. 자세한 정보는 135 페이지의 『조치 세부사항』의 내용을 참조하십시오.

보기 사이의 링크

기본 보기 내에서 테이블의 밑줄 친 텍스트는 링크된 보기에서의 표시를 제어합니다. 텍스트를 클릭하면 특정 필드, 필드 세트 또는 처리 단계에 대한 세부사항을 볼 수 있습니다. 마지막으로 선택한 링크가 더 어두운 색으로 표시됩니다. 이는 두 보기 패널에 있는 컨텐츠 사이의 연결을 식별하는 데 도움이 됩니다.

보기 재설정

원래의 분석 권장사항을 다시 표시하고 분석 보기에 대해 수행한 모든 변경을 중단하려면 기본 보기 패널 맨 아래에 있는 **재설정**을 클릭하십시오.

필드 처리 요약

필드 처리 요약 테이블은 생성된 기능의 수 및 기능의 상태 변경을 포함하여 처리가 미치는 전반적인 영향에 대한 추정 스냅샷을 제공합니다.

모델이 실제로 작성되지는 않으므로 데이터 준비 전과 후의 전체 예측력 변화에 대한 측도나 그래프가 없습니다.

이 테이블에는 다음 정보가 표시됩니다.

- 목표 필드 수
- 원래(입력) 예측변수 수
- 분석 및 모델링에서의 사용이 권장되는 예측변수. 여기에는 권장되는 총 필드 수, 권장되는 변환되지 않은 원래 필드 수, 권장되는 변환된 필드 수(필드의 중간 버전, 날짜/시간 예측변수에서 파생된 필드 및 생성된 예측변수 제외), 날짜/시간 필드에서 파생된 권장되는 필드 수 및 권장되는 생성된 예측변수 수가 포함됩니다.
- 원래 양식이든, 파생된 필드로서든 또는 생성된 예측변수에 대한 입력으로서든 어떤 양식으로도 사용이 권장되지 않는 입력 예측변수 수

필드 정보에 밑줄이 있는 경우, 클릭하면 링크된 보기에 자세한 내용이 표시됩니다. **목표**, **입력 기능** 및 **사용되지 않는 입력 기능**의 세부사항이 필드 테이블 링크된 보기에 표시됩니다. 자세한 정보는 133 페이지의 『필드 테이블』 주제를 참조하십시오. **분석에서 사용이 권장되는 기능**은 예측력 링크된 보기에 표시됩니다. 자세한 정보는 133 페이지의 『예측력』의 내용을 참조하십시오.

필드

필드 기본 보기는 처리된 필드를 표시하고 또한 ADP가 다운스트림 모델에서 처리된 필드의 사용을 권장하는지 여부를 표시합니다. 필드에 대한 권장을 대체할 수 있습니다(예를 들어, 생성된 기능을 제외시키거나 ADP가 제외시키도록 권장하는 기능을 포함시키기 위해). 필드가 변환된 경우, 제안된 변환을 수락할지 또는 원래의 버전을 사용할지 여부를 결정할 수 있습니다.

필드 보기는 두 개의 테이블(목표에 대한 테이블과 처리되었거나 작성된 예측변수에 대한 테이블)로 구성됩니다.

목표 테이블

목표 테이블은 데이터에 목표가 정의된 경우에만 표시됩니다.

이 테이블에는 두 개의 열이 있습니다.

- **이름.** 목표 필드의 이름 또는 레이블입니다. 필드가 변환된 경우에도 원래 이름이 항상 사용됩니다.

- **측정 수준.** 측정 수준을 나타내는 아이콘을 표시합니다. 아이콘 위에 마우스를 놓으면 데이터를 설명하는 레이블(연속형, 순서, 명목 등)이 표시됩니다.

  목표가 변환된 경우 **측정 수준** 열은 변환된 최종 버전을 반영합니다. *참고*: 목표에 대해 변환을 끌 수 없습니다.

예측변수 테이블

예측변수 테이블은 항상 표시됩니다. 테이블의 각 행은 필드를 나타냅니다. 기본적으로 행은 예측력의 내림차순으로 정렬됩니다.

일반 기능의 경우 항상 원래 이름이 행 이름으로 사용됩니다. 날짜/시간 필드의 원래 및 파생 버전 모두 테이블에 표시됩니다(개별 행에). 테이블에는 생성된 예측변수도 포함됩니다.

테이블에 표시된 필드의 변환된 버전은 항상 최종 버전을 나타냅니다.

기본적으로 권장되는 필드만 예측변수 테이블에 표시됩니다. 나머지 필드를 표시하려면 테이블 위에 있는 **테이블에 권장되지 않는 필드 포함** 선택란을 선택하십시오. 그러면 나머지 필드가 테이블 맨 아래에 표시됩니다.

이 테이블에는 다음 열이 포함됩니다.

- **사용할 버전.** 필드가 다운스트림에서 사용될지 여부와 제안된 변환을 사용할지 여부를 제어하는 드롭 다운 목록을 표시합니다. 기본적으로, 이 드롭 다운 목록은 권장사항을 반영합니다.

  변환된 일반 예측변수의 경우 드롭 다운 목록에는 세 가지 선택사항인 **변환**, **원래** 및 **사용하지 않음**이 있습니다.

  변환되지 않은 일반 예측변수의 경우 선택사항은 **원래** 및 **사용하지 않음**입니다.

  파생된 날짜/시간 필드 및 생성된 예측변수의 경우 선택사항은 **변환** 및 **사용하지 않음**입니다.

  원래 날짜 필드의 경우 드롭 다운 목록이 사용되지 않고 **사용하지 않음**으로 설정됩니다.

  참고: 원래 및 변환된 버전이 모두 있는 예측변수의 경우, **원래** 버전과 **변환**된 버전 간에 변경하면 해당 기능의 **측정 수준**과 **예측력** 설정이 자동으로 업데이트됩니다.
- **이름.** 각 필드의 이름은 하나의 링크입니다. 이름을 클릭하면 링크된 보기에 해당 필드에 대한 자세한 정보가 표시됩니다. 자세한 정보는 134 페이지의 『필드 세부사항』의 내용을 참조하십시오.
- **측정 수준.** 데이터 유형을 나타내는 아이콘을 표시합니다. 아이콘 위에 마우스를 놓으면 데이터를 설명하는 레이블(연속형, 순서, 명목 등)이 표시됩니다.
- **예측력.** 예측력은 ADP가 권장하는 필드에 대해서만 표시됩니다. 목표가 정의되지 않은 경우에는 이 열이 표시되지 않습니다. 예측력 범위는 0 - 1이며 값이 클수록 "더 좋은" 예측변수입니다. 일반적으로 예측력은 ADP 분석 내에서 예측변수를 비교하는 데 유용하지만, 분석들 간에 예측력 값을 비교해서는 안 됩니다.

조치 요약

자동 데이터 준비가 조치를 수행할 때마다 입력 예측변수가 변환 및/또는 필터링되며, 하나의 조치 이후에 남은 필드는 그 다음 조치에서 사용됩니다. 마지막 단계까지 남아 있는 필드는 모델링에서 사용하도록 권장되는 반면, 변환되고 생성된 예측변수에 대한 입력은 필터링됩니다.

조치 요약은 ADP가 수행한 처리 조치를 나열하는 단순한 테이블입니다. **조치**에 밑줄이 있는 경우, 클릭하면 링크된 보기에 수행된 조치에 대한 자세한 내용이 표시됩니다. 자세한 정보는 135 페이지의 『조치 세부사항』의 내용을 참조하십시오.

참고: 각 필드의 원래 버전과 최종 변환 버전만 표시되고 분석 중에 사용된 중간 버전은 표시되지 않습니다.

예측력

분석을 처음 실행할 때 기본적으로 표시되거나 필드 처리 요약 기본 보기에서 **분석에서 사용이 권장되는 예측변수**를 선택할 때 표시되는 차트에는 권장되는 예측변수의 예측력이 표시됩니다. 필드는 예측력을 기준으로 정렬되며 가장 높은 값을 갖는 필드가 맨 위에 표시됩니다.

일반 예측변수의 변환된 버전인 경우, 필드 이름은 설정 탭의 필드 이름 패널에서 선택한 접미부를 반영합니다(예: *_transformed*).

개별 필드 이름 뒤에는 측정 수준이 표시됩니다.

권장되는 각 예측변수의 예측력은 목표가 연속형인지 또는 범주형인지에 따라 선형 회귀 또는 naïve Bayes 모델에서 계산됩니다.

필드 테이블

필드 처리 요약 기본 보기에서 **목표**, **예측변수** 또는 **사용되지 않는 예측변수**를 클릭하면 표시되는 필드 테이블 보기에는 관련 기능을 나열하는 단순한 테이블이 표시됩니다.

이 테이블에는 두 개의 열이 있습니다.

- **이름.** 예측변수 이름입니다.

  목표의 경우, 목표가 변환된 경우에도 필드의 원래 이름 또는 레이블이 사용됩니다.

  일반 예측변수의 변환된 버전인 경우, 이름은 설정 탭의 필드 이름 패널에서 선택한 접미부를 반영합니다(예: *_transformed*).

  날짜 및 시간에서 파생된 필드의 경우, 변환된 최종 버전의 이름이 사용됩니다(예: *bdate_years*).

  생성된 예측변수의 경우, 생성된 예측변수의 이름이 사용됩니다(예: *Predictor1*).
- **측정 수준.** 데이터 유형을 나타내는 아이콘을 표시합니다.

  목표의 경우, **측정 수준**은 항상 변환된 버전을 반영합니다(목표가 변환된 경우). 순서(순서형 세트)에서 연속형(범위, 척도) 또는 그 반대로의 변경을 예로 들 수 있습니다.

필드 세부사항

필드 기본 보기에서 아무 **이름**을 클릭하면 표시되는 필드 세부사항 보기에는 선택된 필드의 분포, 결측값 및 예측력 차트(적용 가능한 경우)가 있습니다. 또한 필드의 처리 히스토리와 변환된 필드의 이름도 표시됩니다(적용 가능한 경우).

차트 세트마다 두 개의 버전이 나란히 표시되어 변환이 적용된 필드와 변환이 적용되지 않은 필드를 비교할 수 있습니다. 필드의 변환된 버전이 없는 경우에는 원래 버전에 대한 차트만 표시됩니다. 파생된 날짜 또는 시간 필드와 생성된 예측변수의 경우 새 예측변수에 대한 차트만 표시됩니다.

참고: 너무 많은 범주를 포함해서 필드가 제외되는 경우에는 처리 히스토리만 표시됩니다.

분포 차트

연속형 필드 분포는 평균값에 대한 수직 참조선이 있고 정상 곡선이 오버레이된 히스토그램으로 표시됩니다. 범주형 필드는 막대형 차트로 표시됩니다.

히스토그램은 표준 편차 및 왜도를 표시하는 레이블이 지정됩니다. 그러나, 값 수가 2보다 적거나 최초 필드의 분산이 10-20 미만인 경우에는 왜도가 표시되지 않습니다.

히스토그램의 평균을 표시하거나 막대형 차트에서 범주의 수 및 총 레코드 수의 백분율을 표시하려면 차트 위에 마우스를 놓으십시오.

결측값 차트

원형 차트는 변환이 적용된 결측값 백분율과 변환이 적용되지 않은 결측값 백분율을 비교합니다. 차트 레이블은 백분율을 표시합니다.

ADP가 결측값 처리를 수행한 경우, 변환 후 원형 차트에는 대체값(즉, 결측값 대신 사용된 값)도 레이블로 포함됩니다.

차트 위에 마우스를 놓으면 결측값 수와 총 레코드 수의 백분율이 표시됩니다.

예측력 차트

권장되는 필드의 경우, 막대형 차트는 변환 전후의 예측력을 표시합니다. 목표가 변환된 경우, 계산된 예측력은 변환된 목표와 관련이 있습니다.

참고: 목표가 정의되지 않았거나 기본 보기 패널에서 목표를 클릭한 경우에는 예측력 차트가 표시되지 않습니다.

차트 위에 마우스를 놓으면 예측력 값이 표시됩니다.

처리 히스토리 테이블

이 테이블은 필드의 변환된 버전이 파생된 방법을 보여줍니다. ADP가 수행한 조치가 수행된 순서대로 나열됩니다. 그러나 특정 단계에서는 특정 필드에 대해 여러 조치가 수행되었을 수도 있습니다.

참고: 변환되지 않은 필드의 경우 이 테이블이 표시되지 않습니다.

테이블의 정보는 두 개 또는 세 개의 열로 분류됩니다.

- **조치.** 조치의 이름입니다. 예를 들어, 연속형 예측변수입니다. 자세한 정보는 『조치 세부사항』의 내용을 참조하십시오.
- **세부사항.** 수행된 처리 목록입니다. 예를 들어, 표준 단위로의 변환입니다.
- **함수.** 생성된 예측변수에 대해서만 표시되며, 입력 필드의 선형 조합(예: .06*age + 1.21*height)을 표시합니다.

조치 세부사항

조치 요약 기본 보기에서 밑줄이 있는 **조치**를 선택할 때 표시되는 조치 세부사항 링크된 보기에는 수행된 각 처리 단계에 대한 조치별 정보와 공통 정보가 표시됩니다. 조치별 세부사항이 먼저 표시됩니다.

조치마다 링크된 보기의 맨 위에 설명이 제목으로 사용됩니다. 조치별 세부사항은 제목 아래에 표시되며 여기에는 생성되거나 제외된 예측변수, 병합되거나 다시 정렬된 범주, 목표 변환, 리캐스트된 필드 및 파생된 예측변수의 수 세부사항이 포함될 수 있습니다.

각 조치가 처리됨에 따라 처리에서 사용되는 예측변수의 수가 변경될 수도 있습니다(예를 들어, 예측변수가 제외되거나 병합됨에 따라).

참고: 조치가 꺼져 있거나 목표를 지정하지 않은 경우에 조치 요약 기본 보기에서 조치를 클릭하면 조치 세부사항 위치에 오류 메시지가 표시됩니다.

9개의 가능한 조치가 있지만 분석할 때마다 9개 조치 모두 활성 상태일 필요는 없습니다.

텍스트 필드 테이블

이 테이블에는 다음이 표시됩니다.

- 잘린 후미 공백 값 수
- 분석에서 제외된 예측변수 수

날짜 및 시간 예측변수 테이블

이 테이블에는 다음이 표시됩니다.

- 날짜 및 시간 예측변수에서 파생된 지속 기간 수
- 날짜 및 시간 요소 수
- 파생된 날짜 및 시간 예측변수 총계

날짜 지속 기간이 계산된 경우 참조 날짜 또는 시간이 각주로 표시됩니다.

예측변수 선별 테이블

이 테이블에는 처리에서 제외된 다음 예측변수의 수를 표시합니다.

- 상수
- 결측값이 너무 많은 예측변수
- 단일 범주에 케이스가 너무 많은 예측변수
- 범주가 너무 많은 명목 필드(세트)
- 제외된 예측변수 총계

측정 수준 확인 테이블

이 테이블은 리캐스된 필드 수를 표시하며 다음과 같이 분류됩니다.

- 연속형 필드로 리캐스트된 순서 필드(순서형 세트)
- 순서 필드로 리캐스트된 연속형 필드
- 리캐스트된 총 수

연속형 또는 순서 입력 필드(목표 또는 예측변수)가 없는 경우에는 각주가 표시됩니다.

이상값 테이블

이 테이블은 예상값이 처리된 필드 수를 표시합니다.

- 설정 탭의 입력 및 목표 준비 패널에 지정된 설정에 따라, 이상값이 발견되고 잘린 연속형 필드 수 또는 이상값이 발견되고 결측으로 설정된 연속형 필드 수
- 이상값 처리 후 상수여서 제외된 연속형 필드 수

하나의 각주는 이상값 절사 값을 표시하는 반면, 다른 각주는 연속형 입력 필드(목표 또는 예측변수)가 없는 경우에 표시됩니다.

결측값 테이블

이 테이블은 결측값이 대체된 필드의 수를 표시하며 다음과 같이 분류됩니다.

- 목표. 목표를 지정하지 않은 경우에는 이 행이 표시되지 않습니다.
- 예측변수. 다시 명목(세트), 순서(순서형 세트) 및 연속형의 수로 분류됩니다.
- 대체된 결측값 총계

목표 테이블

이 테이블은 목표가 변환되었는지 여부를 다음과 같이 표시합니다.

- 정규로의 Box-Cox 변환. 이는 다시 지정된 기준(평균 및 표준 편차) 및 람다를 표시하는 열로 분류됩니다.
- 목표 범주가 안정성 향상을 위해 다시 정렬됨

범주형 예측변수 테이블

테이블에 다음과 같은 범주형 예측변수의 수가 표시됩니다.

- 안정성을 향상시키기 위해 해당 범주가 가장 낮은 값에서 가장 높은 값 순으로 다시 정렬된 범주형 예측변수
- 목표와의 연관을 최대화하기 위해 해당 범주가 병합된 범주형 예측변수
- 성긴 범주를 처리하기 위해 해당 범주가 병합된 범주형 예측변수
- 목표와의 연관이 낮아 제외된 범주형 예측변수
- 병합 후 상수여서 제외된 범주형 예측변수

범주형 예측변수가 없는 경우에는 각주가 표시됩니다.

연속형 예측변수 테이블

두 개의 테이블이 있습니다. 첫 번째 테이블에는 다음과 같은 변환의 수 중 하나가 표시됩니다.
- 표준 단위로 변환된 예측변수 값. 또한, 변환된 예측변수 수, 지정된 평균값 및 표준 편차를 표시합니다.
- 공통 범위로 맵핑된 예측변수 값. 또한, 지정된 최소 및 최대 값과 함께 최소-최대 변환을 사용하여 변환된 예측변수 수를 표시합니다.
- 구간화된 예측변수 값 및 구간화된 예측변수 수

두 번째 테이블에는 다음과 같은 예측변수의 수로 표시되는 예측변수 공간 생성 세부사항이 표시됩니다.
- 생성된 예측변수
- 목표와의 연관이 낮아 제외된 예측변수
- 구간화 후 상수여서 제외된 예측변수
- 생성 후 상수여서 제외된 예측변수

연속형 예측변수가 입력되지 않은 경우에는 각주가 표시됩니다.

파생 노드 생성

파생 노드를 생성하는 경우, 파생 노드는 스코어 필드에 역 목표 변환을 적용합니다. 기본적으로, 이 노드는 자동 모델 작성기(예: Auto Classifier 또는 Auto Numeric) 또는 앙상블 노드에 의해 생성될 스코어 필드의 이름을 입력합니다. 척도(범위) 목표가 변환된 경우 스코어 필드는 변환된 단위로 표시됩니다(예: $ 대신 log($)). 결과를 해석하고 사용하려면 예측값을 다시 원래 척도로 변환해야 합니다.

참고: ADP 노드에 범위 목표의 척도를 조정하는 분석이 포함되는 경우(즉, 입력 및 목표 준비 패널에서 Box-Cox 척도 조정을 선택한 경우)에만 파생 노드를 생성할 수 있습니다. 목표가 범위가 아니거나 Box-Cox 척도 조정을 선택하지 않은 경우에는 파생 노드를 생성할 수 없습니다.

파생 노드는 다중 모드에서 작성되고 표현식에서 @FIELD를 사용하므로 필요한 경우 변환된 목표를 추가할 수 있습니다. 예를 들어, 다음과 같은 세부사항을 사용합니다.
- 목표 필드 이름: response
- 변환된 목표 필드 이름: response_transformed
- 스코어 필드 이름: $XR-response_transformed

파생 노드는 새 필드 $XR-response_transformed_inverse를 작성합니다.

참고: 자동 모델 작성기 또는 앙상블 노드를 사용하지 않는 경우에는 모델에 올바른 스코어 필드를 변환하도록 파생 노드를 편집해야 합니다.

정규화된 연속형 대상

기본적으로, 입력 및 목표 준비 패널에서 **Box-Cox 변환으로 연속형 대상 척도 조정** 선택란을 선택하면 목표가 변환되고 사용자는 모델 작성의 목표가 될 새 필드를 작성합니다. 예를 들어, 원래의 목표가 *response*였던 경우 새 목표는 *response_transformed*가 됩니다.

그러나 이 방법은 원래 목표에 따라 여러 가지 문제점이 발생할 수 있습니다. 예를 들어, 목표가 *Age*였던 경우, 새 목표의 값은 *Years*가 아니라 *Years*의 변환된 버전이 됩니다. 이는 스코어가 인식 가능한 단위가 아니기 때문에 스코어를 보고 해석할 수 없음을 의미합니다. 이 경우, 변환된 단위를 원래 의도한 것으로 되돌리는 역변환을 적용할 수 있습니다. 이를 위해서는 다음을 수행하십시오.

1. **데이터 분석**을 클릭하여 ADP 분석을 실행한 후 *생성* 메뉴에서 *파생* 노드를 선택하십시오.
2. 모델 캔버스에서 사용하는 너깃 다음에 파생 노드를 배치하십시오.

파생 노드는 예측값이 원래의 *Years* 값이 되도록 스코어 필드를 원래 차원으로 복원합니다.

기본적으로 파생 노드는 자동 모델 작성기 또는 앙상블 모형으로 생성된 스코어 필드를 변환합니다. 개별 모델을 작성하는 경우 실제 스코어 필드에서 파생되도록 파생 노드를 편집해야 합니다. 모델을 평가하려면 파생 노드에서 **파생 소스** 필드에 변환된 목표를 추가해야 합니다. 그러면 목표에 동일한 역변환이 적용되며, 다운스트림 평가 또는 분석 노드는 메타데이터 대신 필드 이름을 사용하도록 전환할 경우 변환된 데이터를 올바르게 사용합니다.

원래 이름을 복원하려면 필터 노드를 사용하여 원래 목표 필드(아직 거기에 있는 경우)를 제거하고 목표 및 스코어 필드의 이름을 바꾸십시오.

유형 노드

필드 특성은 소스 노드 또는 별도의 유형 노드에서 지정할 수 있습니다. 두 노드 모두에서 기능은 비슷합니다. 사용 가능한 특성은 다음과 같습니다.

- **필드** 필드 이름을 두 번 클릭하여 IBM SPSS Modeler에서 데이터에 대한 값 및 필드 레이블을 지정하십시오. 예를 들어, IBM SPSS Statistics에서 가져온 필드 메타데이터를 여기서 보거나 수정할 수 있습니다. 마찬가지로 필드 및 해당 값에 대한 새 레이블을 작성할 수 있습니다. 여기서 지정하는 레이블은 스트림 특성 대화 상자에서 작성하는 선택사항에 따라 IBM SPSS Modeler 전체에 표시됩니다.
- **측정** 지정된 필드의 데이터 특성을 설명하는 데 사용되는 측정 수준입니다. 필드의 모든 세부사항이 알려져 있는 경우 이를 **완전히 인스턴스화**되어 있다고 합니다. 자세한 정보는 140 페이지의 『측정 수준』의 내용을 참조하십시오.

  참고: 필드의 측정 수준은 데이터가 문자열, 정수, 실수, 날짜, 시간소인, 목록 중 어느 것으로 저장되는지를 표시하는 해당 저장 유형과 다릅니다.

- **값** 이 열에서는 데이터 세트에서 데이터 값을 읽어오기 위한 옵션을 지정하거나 **지정** 옵션을 사용하여 별도의 대화 상자에서 측정 수준 및 값을 지정할 수 있습니다. 해당 값을 읽지 않고 필드를 통과하도록 선택할 수도 있습니다. 자세한 정보는 145 페이지의 『데이터 값』의 내용을 참조하십시오.

  참고: 해당 **필드** 항목에 목록이 포함되어 있으면 이 열에서 셀을 수정할 수 없습니다.
- **결측** 필드의 결측값이 처리되는 방식을 지정하는 데 사용됩니다. 자세한 정보는 149 페이지의 『결측값 정의』의 내용을 참조하십시오.

  참고: 해당 **필드** 항목에 목록이 포함되어 있으면 이 열에서 셀을 수정할 수 없습니다.
- **확인** 이 열에서는 필드 값이 지정된 값 또는 범위를 준수하는지 확인하는 옵션을 설정할 수 있습니다. 자세한 정보는 150 페이지의 『유형 값 검사』의 내용을 참조하십시오.

  참고: 해당 **필드** 항목에 목록이 포함되어 있으면 이 열에서 셀을 수정할 수 없습니다.
- **역할** 필드가 시스템 학습 프로세스에 대해 **입력**(예측변수 필드)인지 아니면 **목표**(예측 필드)인지를 모델링 노드에 알리는 데 사용됩니다. 레코드를 훈련, 검정 및 검증을 위한 별도의 표본으로 파티셔닝하는 데 사용되는 필드를 표시하는 **파티션**과 함께 **모두** 및 **없음** 역할도 사용할 수 있습니다. **분할** 값은 필드의 가능한 각각의 값에 대해 별도의 모델이 작성되도록 지정합니다. 자세한 정보는 150 페이지의 『필드 역할 설정』의 내용을 참조하십시오.

유형 노드 창을 사용하여 기타 여러 옵션을 지정할 수 있습니다.

- 도구 메뉴 단추를 사용하여 유형 노드가 인스턴스화(지정, 값 읽기 또는 스트림 실행을 통해)된 후 **고유 필드를 무시**하도록 선택할 수 있습니다. 고유 필드를 무시하는 경우 값이 하나뿐인 필드가 자동으로 무시됩니다.
- 도구 메뉴 단추를 사용하여 유형 노드가 인스턴스화된 후 **대형 세트를 무시**하도록 선택할 수 있습니다. 대형 세트를 무시하는 경우 다수의 멤버를 포함하는 세트가 자동으로 무시됩니다.
- 도구 메뉴 단추를 사용하여 유형 노드가 인스턴스화된 후 **연속형 정수를 순서로 변환**하도록 선택할 수 있습니다. 자세한 정보는 143 페이지의 『연속형 데이터 변환』의 내용을 참조하십시오.
- 도구 메뉴 단추를 사용하여 선택된 필드를 버리는 필터 노드를 생성할 수 있습니다.
- 선글라스 토글 단추를 사용하여 모든 필드의 기본값을 읽기 또는 패스로 설정할 수 있습니다. 소스 노드의 유형 탭은 기본적으로 필드를 패스하는 반면, 유형 노드 자체는 기본적으로 값을 읽습니다.
- **값 지우기** 단추를 사용하여 이 노드에서 작성된 필드 값 변경사항(비상속 값)을 지우고 업스트림 조작에서 값을 다시 읽을 수 있습니다. 이 옵션은 특정 업스트림 필드에 대해 수행한 변경사항을 재설정하는 데 유용합니다.
- **모든 값 지우기** 단추를 사용하여 노드로 읽어온 **모든** 필드의 값을 재설정할 수 있습니다. 이 옵션은 모든 필드에 대해 효과적으로 **값** 열을 **Read**로 설정합니다. 이 옵션은 모든 필드의 값을 재설정하고 업스트림 조작에서 값 및 유형을 다시 읽는 데 유용합니다.
- 컨텍스트 메뉴를 사용하여 하나의 필드에서 다른 필드로 속성을 **복사**하도록 선택할 수 있습니다. 자세한 정보는 151 페이지의 『유형 속성 복사』의 내용을 참조하십시오.

- **사용하지 않는 필드 설정 보기** 옵션을 사용하여 데이터에 더 이상 존재하지 않거나 이 유형 노드에 연결된 적이 있는 필드의 유형 설정을 볼 수 있습니다. 이 옵션은 변경된 데이터 세트에 유형 노드를 재사용할 때 유용합니다.

측정 수준

측정 수준(이전의 "데이터 유형" 또는 "사용 유형")은 IBM SPSS Modeler에서 데이터 필드가 어떻게 사용되는지 설명합니다. 소스 또는 유형 노드의 유형 탭에서 측정 수준을 지정할 수 있습니다. 예를 들어, 1 및 0 값을 갖는 정수 필드의 측정 수준을 *플래그*로 설정할 수 있습니다. 이는 보통 1 = *True*이고 0 = *False*임을 표시합니다.

저장 공간 대 측정. 필드의 측정 수준은 해당 저장 유형과 다릅니다. 저장 유형은 데이터가 문자열, 정수, 실수, 날짜, 시간 또는 시간소인으로 저장되는지 여부를 표시합니다. 데이터 유형은 유형 노드를 사용하여 스트림의 어느 위치에서든 수정할 수 있는 반면, 저장 공간은 IBM SPSS Modeler로 데이터를 읽을 때 소스에서 결정해야 합니다(단, 이후에 변환 함수를 사용하여 변경할 수 있음). 자세한 정보는 9 페이지의 『필드 저장 공간 및 형식화 설정』의 내용을 참조하십시오.

일부 모델링 노드는 해당 필드 탭에서 아이콘으로 해당 입력 및 목표 필드에 허용되는 측정 수준 유형을 표시합니다.

측정 수준 아이콘

표 19. 측정 수준 아이콘

아이콘	측정 수준
	기본값
	연속형
	범주형
	플래그
	명목
	순서
	유형 없음
[ ]	요약도표
	지리 공간적

사용 가능한 측정 수준은 다음과 같습니다.

- **기본값** 저장 유형 및 값을 알 수 없는 데이터(예를 들어, 아직 읽지 않았기 때문)는 **<기본값>**으로 표시됩니다.
- **연속형** 0–100 또는 0.75–1.25 범위 등의 숫자 값을 설명하는 데 사용됩니다. 연속형 값은 정수, 실수 또는 날짜/시간일 수 있습니다.
- **범주형** 고유 값의 정확한 숫자를 알 수 없는 경우 문자열 값에 사용됩니다. 이는 **인스턴스화되지 않은** 데이터 유형입니다(데이터의 사용 및 저장에 대해 가능한 모든 정보가 아직 알려져 있지 않음을 의미함). 데이터를 읽고 나면 스트림 특성 대화 상자에서 지정된 명목 필드의 최대 멤버 수에 따라 측정 수준은 *플래그*, *명목* 또는 *유형 없음*이 됩니다.
- **플래그** 특성의 존재 여부를 표시하는 두 개의 고유 값(예: `true` 및 `false`, `Yes` 및 `No` 또는 `0` 및 `1`)을 가진 데이터에 사용됩니다. 사용된 값은 다를 수 있지만 항상 하나의 값은 "참" 값으로 지정하고 다른 하나의 값은 "거짓" 값으로 지정해야 합니다. 데이터는 텍스트, 정수, 실수, 날짜, 시간 또는 시간소인으로 표시될 수 있습니다.
- **명목** 각각 세트의 멤버로 처리되는 다중 고유 값을 가진 데이터(예: `small/medium/large`)를 설명하는 데 사용됩니다. 명목 데이터는 모든 저장 공간(숫자, 문자열 또는 날짜/시간)을 가질 수 있습니다. 측정 수준을 명목으로 설정해도 값이 문자열 저장 공간으로 자동으로 변경되지는 않습니다.
- **순서** 내재된 순서가 있는 다중 고유 값을 가진 데이터를 설명하는 데 사용됩니다. 예를 들어, 급여 범주 또는 만족도 순위를 순서 데이터로 유형화할 수 있습니다. 순서는 데이터 요소의 자연 정렬 순서에 의해 정의됩니다. 예를 들어, `1, 3, 5`는 정수 세트에 대한 기본 정렬 순서이고 `HIGH, LOW, NORMAL`(알파벳 오름차순)은 문자열 세트에 대한 순서입니다. 순서 측정 수준을 사용하면 시각화, 모델 작성 및 순서 데이터를 고유 유형으로 인식하는 다른 애플리케이션(예: IBM SPSS Statistics)에 내보내기를 위해 범주형 데이터 세트를 순서 데이터로 정의할 수 있습니다. 명목 필드를 사용할 수 있는 모든 위치에서 순서 필드를 사용할 수 있습니다. 또한 모든 유형(실수, 정수, 문자열, 날짜, 시간 등)의 필드를 순서로 정의할 수 있습니다.
- **유형 없음** 위 유형을 준수하지 않는 데이터, 단일 값을 가진 필드 또는 세트에 정의된 최대값보다 많은 수의 멤버가 포함된 명목 데이터에 사용됩니다. 이는 그렇지 않으면 측정 수준이 다수의 멤버를 가진 세트(예: 계정 번호)가 되는 경우에도 유용합니다. 필드에 대해 **유형 없음**을 선택하면 역할이 **없음**으로 자동으로 설정되고 **레코드 ID**가 유일한 대안입니다. 세트의 기본 최대 *크기*는 250개의 고유 값입니다. 이 숫자는 도구 메뉴에서 액세스할 수 있는 스트림 특성 대화 상자의 옵션 탭에서 조정되거나 사용 안함으로 설정될 수 있습니다.
- **콜렉션** 목록에서 기록되는 비지리 공간적 데이터를 식별하는 데 사용됩니다. 콜렉션은 깊이가 0(영)인 목록 필드이며 이 목록에 있는 요소는 다른 측정 수준 중 하나를 가집니다.

  목록에 대한 자세한 정보는 SPSS Modeler 소스, 프로세스 및 출력 노드 안내서의 소스 노드 절에 있는 목록 저장 공간 및 연관된 측정 수준 주제를 참조하십시오.
- **지리 공간적** 지리 공간적 데이터를 식별하기 위해 목록 저장 유형과 함께 사용됩니다. 목록은 깊이가 0과 2 사이(경계값 포함)인 목록을 가진 정수 목록 또는 실수 목록 필드가 될 수 있습니다.

  자세한 정보는 SPSS Modeler 소스, 프로세스 및 출력 노드 안내서의 유형 노드 절에 있는 지리 공간적 측정 하위 수준 주제를 참조하십시오.

측정 수준을 수동으로 지정하거나, 소프트웨어가 데이터를 읽고 읽은 값을 기반으로 측정 수준을 결정하도록 할 수 있습니다.

선택적으로, 범주형 데이터로 처리해야 하는 연속형 데이터 필드가 있는 경우 이를 변환하는 옵션을 선택할 수 있습니다. 자세한 정보는 143 페이지의 『연속형 데이터 변환』의 내용을 참조하십시오.

자동 입력을 사용하려면 다음을 수행하십시오.

1. 유형 노드 또는 소스 노드의 유형 탭에서 원하는 필드의 *값* 열을 **<Read>**로 설정하십시오. 그러면 모든 다운스트림 노드에서 메타데이터를 사용할 수 있습니다. 대화 상자의 선글라스 단추를 사용하여 모든 필드를 신속하게 **<Read>** 또는 **<Pass>**로 설정할 수 있습니다.
2. 데이터 소스에서 바로 값을 읽으려면 **값 읽기**를 클릭하십시오.

필드의 측정 수준을 수동으로 설정하려면 다음을 수행하십시오.

1. 테이블에서 필드를 선택하십시오.
2. *측정* 열의 드롭 다운 목록에서 필드의 측정 수준을 선택하십시오.
3. 또는, Ctrl-A 또는 Ctrl-클릭을 사용하여 여러 필드를 선택한 후 드롭 다운 목록을 사용하여 측정 수준을 선택할 수 있습니다.

지리 공간적 측정 수준

목록 저장 유형과 함께 사용되는 지리 공간적 측정 수준에는 다양한 유형의 지리 공간적 데이터를 식별하는 데 사용되는 6개의 하위 수준이 있습니다.

- **점** - 특정 지점(예: 도시의 중심)을 식별합니다.
- **다각형** - 지역의 단일 경계와 해당 위치를 식별하는 일련의 점입니다(예: County).
- **LineString** - 폴리라인 또는 라인이라고도 하는 LineString은 선의 경로를 식별하는 일련의 점입니다. 예를 들어, LineString은 도로, 강, 철도 등의 고정 항목이거나 비행기의 비행 경로 또는 배의 항로와 같은 움직이는 물체의 이동 경로일 수 있습니다.
- **다중 점** - 데이터의 각 행에 지역당 여러 지점이 포함될 때 사용됩니다. 예를 들어, 각 행이 도시 거리를 나타내는 경우, 각 거리의 여러 지점이 모든 가로등을 식별하는 데 사용될 수 있습니다.
- **다중 다각형** - 데이터의 각 행에 여러 다각형이 포함될 때 사용됩니다. 예를 들어, 각 행이 국가의 윤곽선을 나타내는 경우, 미국은 본토, 알래스카, 하와이 등의 다양한 영역을 식별하는 여러 다각형으로 기록될 수 있습니다.
- **다중 LineString** - 데이터의 각 행에 여러 선이 포함될 때 사용됩니다. 선은 갈라질 수 없기 때문에 다중 LineString을 사용하여 선 그룹을 식별할 수 있습니다. 예를 들어, 각 국가의 철도망 또는 가항 수로와 같은 데이터입니다.

이러한 측정 하위 수준은 목록 저장 유형과 함께 사용됩니다. 자세한 정보는 11 페이지의 『목록 저장 공간 및 연관된 측정 수준』의 내용을 참조하십시오.

제한사항

지리 공간적 데이터를 사용할 때는 몇 가지 제한사항에 주의해야 합니다.

- 좌표계는 데이터의 형식에 영향을 줄 수 있습니다. 예를 들어, 평면직각 좌표계는 좌표값 x, y 및 z(필요한 경우)를 사용하는 반면, 지리 좌표계는 좌표값 경도 및 위도와 (필요한 경우) 고도 또는 깊이 값을 사용합니다.

  좌표계에 대한 자세한 정보는 SPSS Modeler 사용자 안내서의 스트림에 대한 작업 섹션에서 스트림의 지리 공간적 옵션 설정 주제를 참조하십시오.
- LineString은 자체를 가로지를 수 없습니다.
- 다각형은 자동으로 닫히지 않습니다. 다각형마다 첫 번째 및 마지막 점을 동일한 점으로 정의해야 합니다.
- 다중 다각형의 데이터 방향이 중요합니다. 시계방향은 단단한 구조를 표시하고 반시계 방향은 비어 있는 구조를 표시합니다. 예를 들어, 호수를 포함하는 국가의 영역을 기록하는 경우, 본토 영역 경계는 시계방향으로 기록하고 각 호수의 형태는 반시계 방향으로 기록할 수 있습니다.
- 다각형은 그 자체와 교차할 수 없습니다. 이러한 교차의 예는 다각형의 경계를 숫자 8 양식의 연속 선으로 구성하려는 경우입니다.
- 다중 다각형은 서로 겹칠 수 없습니다.
- 지리 공간적 필드의 경우, 관련된 유일한 저장 유형은 **실수** 및 **정수**입니다(기본 설정은 **실수**임).

지리 공간적 측정 하위 수준 아이콘

표 20. 지리 공간적 측정 하위 수준 아이콘

아이콘	측정 수준
	점
	다각형
	LineString
	다중 점
	다중 다각형
	다중 LineString

연속형 데이터 변환

범주형 데이터를 연속형으로 처리하면 모델(특히, 목표 필드인 경우)의 품질에 심각한 영향을 미칠 수 있습니다(예를 들어, 2진 모형이 아닌 회귀 모형을 생성함). 이를 방지하기 위해 정수 범위를 범주형 유형(예: *순서* 또는 *플래그*)으로 변환할 수 있습니다.

1. 조작 및 메뉴 생성 단추(도구 기호가 있음)에서 **연속형 정수를 순서로 변환**을 선택하십시오. 변환 값 대화 상자가 표시됩니다.
2. 자동으로 변환되는 범위 *크기*를 지정하십시오. 이는 입력하는 *크기*까지의 모든 범위에 적용됩니다.
3. **확인**을 클릭하십시오.적용되는 범위가 *플래그* 또는 *순서*로 변환되고 유형 노드의 유형 탭에 표시됩니다.

변환 결과

- 정수 저장 공간을 포함하는 *연속형* 필드가 *순서*로 변경되는 경우, 최소값과 최대값이 펼쳐져 최소값에서 최대값까지의 모든 정수 값이 포함됩니다. 예를 들어, 범위가 1, 5이면 값 세트는 1, 2, 3, 4, 5입니다.
- *연속형* 필드가 *플래그*로 변경되는 경우, 최소값 및 최대값은 플래그 필드의 false 및 true 값이 됩니다.

인스턴스화 개념

인스턴스화는 데이터 필드의 저장 유형 및 값과 같은 정보를 읽거나 지정하는 프로세스입니다. 인스턴스화는 시스템 자원을 최적화하기 위한 사용자 지시 프로세스입니다. 소스 노드의 유형 탭에서 옵션을 지정하거나 유형 노드를 통해 데이터를 실행하여 소프트웨어에 값을 읽도록 지시합니다.

- 유형을 알 수 없는 데이터도 **인스턴스화되지 않음**으로 참조됩니다. 해당 저장 유형 및 값을 알 수 없는 데이터는 유형 탭의 *측정* 열에 **<기본값>**으로 표시됩니다.
- 필드의 저장 공간(예: 문자열 또는 숫자)에 대한 일부 정보가 있는 경우 해당 데이터는 **부분적으로 인스턴스화됨**이라고 합니다. **범주형** 또는 **연속형**은 부분적으로 인스턴스화된 측정 수준입니다. 예를 들어, **범주형**은 필드가 기호 필드임을 지정하지만 명목, 순서 및 플래그 필드 중 어느 것인지 알 수 없습니다.
- 값을 포함하여 유형에 대한 모든 세부사항을 알고 있는 경우, **완전히 인스턴스화됨** 측정 수준(명목, 순서, 플래그 또는 연속형)이 이 열에 표시됩니다. *참고: 연속형* 유형은 부분적으로 인스턴스화된 데이터 필드 및 완전히 인스턴스화된 데이터 필드 모두에 사용됩니다. 연속형 데이터는 정수 또는 실수일 수 있습니다.

유형 노드가 있는 데이터 스트림의 실행 동안, 인스턴스화되지 않은 유형은 초기 데이터 값을 기반으로 즉시 부분적으로 인스턴스화됩니다. 모든 데이터가 노드를 통과하면, 값을 **<Pass>**로 설정하지 않은 한, 모든 데이터가 완전히 인스턴스화됩니다. 실행이 중단되면 데이터는 부분적으로 인스턴스화된 상태로 유지됩니다. 유형 탭이 인스턴스화되면 필드의 값은 스트림의 해당 위치에서 정적입니다. 이는 스트림을 다시 실행해도 업스트림 변경사항은 특정 필드의 값에 영향을 주지 않음을 의미합니다. 새 데이터 또는 추가된 조작을 기반으로 값을 변경하거나 업데이트하려면 유형 탭 자체에서 값을 편집하거나 필드의 값을 **<Read>** 또는 **<Read +>**로 설정해야 합니다.

인스턴스화 시점

일반적으로, 데이터 세트가 매우 크지 않고 나중에 스트림에 필드를 추가할 계획이 아닌 경우에는 소스 노드에서 인스턴스화하는 것이 가장 편리한 방법입니다. 그러나 다음 경우에는 별도의 유형 노드에서 인스턴스화하는 것이 유용합니다.

- 데이터 세트가 크고 스트림이 해당 유형 노드 이전에 서브세트를 필터링합니다.
- 스트림에서 데이터가 필터링되었습니다.
- 스트림에서 데이터가 병합되었거나 추가되었습니다.

• 처리 중에 새 데이터 필드가 파생됩니다.

데이터 값

유형 탭의 **값** 열을 사용하여 데이터에서 자동으로 값을 읽거나 별도의 대화 상자에서 측정 수준 및 값을 지정할 수 있습니다.

값 드롭 다운 목록에서 사용 가능한 옵션은 다음 표에 표시된 바와 같이 자동 입력을 위한 명령어를 제공합니다.

표 21. 자동 입력을 위한 명령어

옵션	기능
<읽기>	노드가 실행될 때 데이터를 읽습니다.
<읽기+>	데이터를 읽어 현재 데이터(있는 경우)에 추가합니다.
<통과>	데이터를 읽지 않습니다.
<현재>	현재 데이터 값을 유지합니다.
지정...	값 및 측정 수준 옵션을 지정하는 별도의 대화 상자가 열립니다.

유형 노드를 실행하거나 **값 읽기**를 클릭하면 자동 입력이 수행되고 선택항목을 기반으로 데이터 소스에서 값을 읽습니다. 이러한 값은 지정 옵션을 사용하거나 **필드** 열의 셀을 두 번 클릭하여 수동으로 지정할 수도 있습니다.

유형 노드에서 필드에 대한 변경을 수행한 후에는 대화 상자 도구 모음의 다음 단추를 사용하여 값 정보를 재설정할 수 있습니다.

- **값 지우기** 단추를 사용하여 이 노드에서 작성된 필드 값 변경사항(비상속 값)을 지우고 업스트림 조작에서 값을 다시 읽을 수 있습니다. 이 옵션은 특정 업스트림 필드에 대해 수행한 변경사항을 재설정하는 데 유용합니다.
- **모든 값 지우기** 단추를 사용하여 노드로 읽어온 **모든** 필드의 값을 재설정할 수 있습니다. 이 옵션은 모든 필드에 대해 효과적으로 *값* 열을 **읽기**로 설정합니다. 이 옵션은 모든 필드의 값을 재설정하고 업스트림 조작에서 값 및 측정 수준을 다시 읽는 데 유용합니다.

값 열의 회색 텍스트

유형 노드 또는 소스 노드 내에서 **값** 열의 데이터가 검은색 텍스트로 표시된다면 필드 값을 읽어 해당 노드에 저장된 것입니다. 이 필드에 검은색 텍스트가 표시되지 않는다면 해당 필드의 값을 읽지 않아 이후 업스트림으로 판별된 것입니다.

데이터가 회색 텍스트로 표시되는 경우가 있습니다. SPSS Modeler가 실제로 데이터를 읽고 저장하지 않아도 유효한 값을 식별하거나 유추할 수 있는 경우에 이와 같이 표시됩니다. 다음 노드 중 하나를 사용하는 경우에 이러한 상황이 발생할 수 있습니다.

- 사용자 입력 노드. 데이터는 노드 내에 정의되므로 값이 노드에 저장되지 않았더라도 필드의 값 범위는 항상 알려져 있습니다.

- 통계 파일 소스 노드. 데이터 유형에 대한 메타데이터가 있을 경우 데이터를 읽거나 저장하지 않아도 SPSS Modeler가 가능한 값 범위를 유추할 수 있습니다.

이 노드에서는 **값 읽기**를 클릭할 때까지 값이 회색 텍스트로 표시됩니다.

참고: 스트림의 데이터를 인스턴스화하지 않았는데 데이터 값이 회색으로 표시될 경우 **확인** 열에서 설정한 유형 값 확인이 적용되지 않습니다.

값 대화 상자 사용

유형 탭의 **값** 또는 **결측** 열을 클릭하면 사전 정의된 값의 드롭 다운 목록이 표시됩니다. 이 목록에서 **지정...** 옵션을 선택하면 선택된 필드에 대한 값을 읽고 지정하고 레이블 지정하고 처리하기 위한 옵션을 설정할 수 있는 별도의 대화 상자가 열립니다.

다수의 제어는 모든 유형의 데이터에 대해 공통입니다. 이 공통 제어에 대해 여기서 설명합니다.

측정 현재 선택된 측정 수준을 표시합니다. 데이터를 사용할 방식을 반영하도록 설정을 변경할 수 있습니다. 예를 들어, day_of_week라는 필드에 개별 날짜를 나타내는 숫자가 포함되어 있는 경우에는 각각의 범주를 개별적으로 조사하는 분포 노드를 작성하기 위해 이를 명목 데이터로 변경할 수 있습니다.

저장 공간 저장 유형을 표시합니다(알려진 경우). 저장 유형은 선택하는 측정 수준의 영향을 받지 않습니다. 저장 유형을 변경하기 위해 고정 파일 및 가변파일 소스 노드에서 데이터 탭을 사용하거나 채움 노드에서 변환 함수를 사용할 수 있습니다.

모델 필드 모델 너깃 스코어링의 결과로 생성되는 필드의 경우 모델 필드 세부사항도 볼 수 있습니다. 이 세부사항에는 목표 필드의 이름과 모델링에서 필드의 역할(예측값, 확률, 성향 등)이 포함됩니다.

값 선택된 필드에 대한 값을 판별할 방법을 선택하십시오. 여기서 작성하는 선택사항은 유형 노드 대화 상자의 **값** 열에서 이전에 작성한 선택사항을 대체합니다. 값을 읽기 위한 선택사항은 다음과 같습니다.

- **데이터에서 읽기** 노드가 실행될 때 값을 읽으려면 선택하십시오. 이 옵션은 **<읽기>**와 동일합니다.
- **통과** 현재 필드에 대한 데이터를 읽지 않으려면 선택하십시오. 이 옵션은 **<통과>**와 동일합니다.
- **값 및 레이블 지정** 여기의 옵션은 선택된 필드에 대한 값 및 레이블을 지정하는 데 사용됩니다. 값 확인과 함께 사용하여 이 옵션을 통해 현재 필드에 대한 사용자의 지식을 기반으로 하는 값을 지정하십시오. 이 옵션은 각 유형의 필드에 대한 고유 제어를 활성화합니다. 값 및 레이블에 대한 옵션은 후속 주제에서 개별적으로 다룹니다.

  참고: 측정 수준이 유형 없음 또는 <기본값>인 필드에 대해서는 값 또는 레이블을 지정할 수 없습니다.
- **데이터의 값 확장** 여기서 입력하는 값을 사용하여 현재 데이터를 추가하려면 선택하십시오. 예를 들어, field_1의 범위가 (0,10)에서 시작하는 경우 (8,16)에서 시작하는 값 범위를 입력하면 원래 최소값은 제거하지 않고 16을 추가하여 범위가 확장됩니다. 새 범위는 (0,16)입니다. 이 옵션을 선택하면 자동 입력 옵션이 **<읽기+>**로 자동으로 설정됩니다.

최대 목록 길이 측정 수준이 지리 공간 또는 콜렉션인 데이터에만 사용할 수 있습니다. 목록이 포함할 수 있는 요소 수를 지정하여 목록의 최대 길이를 설정하십시오.

값 확인 지정된 연속형, 플래그 또는 명목 값을 준수하도록 값을 값을 강제하는 방법을 선택하십시오. 이 옵션은 유형 노드 대화 상자의 **확인** 열에 해당하며 여기서 작성되는 설정은 해당 대화 상자의 설정을 대체합니다. **값 및 레이블 지정** 옵션과 함께 사용되면 값 확인을 통해 예측값을 가진 데이터에서 값을 준수할 수 있습니다. 예를 들어, 값을 1, 0으로 지정한 후 **삭제** 옵션을 사용하는 경우 값이 1 또는 0이 아닌 모든 레코드를 삭제할 수 있습니다.

공백 정의 데이터에서 결측값 또는 공백을 선언하는 데 사용하는 다음과 같은 제어를 활성화하려면 선택하십시오.

- **결측값** 특정 값(예: 99 또는 0)을 공백으로 정의하려면 이 테이블을 사용하십시오. 값은 필드의 저장 유형에 대해 적합해야 합니다.
- **범위** 결측값의 범위(예: 연령 1–17 또는 65 초과)를 지정하는 데 사용됩니다. 경계값이 비어 있으면 범위는 한정되지 않습니다. 예를 들어, 상한 없이 하한으로 100이 지정되면 100 이상의 값은 모두 결측값으로 정의됩니다. 경계값도 포함됩니다. 예를 들어, 하한이 5이고 상한이 10인 범위의 범위 정의에는 5와 10이 포함됩니다. 결측값 범위는 날짜/시간 및 문자열(이 경우 값이 범위 내에 있는지 판별하기 위해 알파벳 정렬 순서가 사용됨)을 포함한 모든 저장 유형에 대해 정의할 수 있습니다.
- **널/공백** 시스템 널(데이터에서 $null$로 표시됨) 및 공백(표시되는 문자가 없는 문자열 값)을 공백으로 지정할 수도 있습니다.

  참고: 유형 노드에서는 분석을 위해 빈 문자열도 공백으로 처리하지만 빈 문자열은 내부적으로 다르게 저장되며 특정 케이스에서 다르게 처리될 수 있습니다.

참고: 공백을 정의되지 않음 또는 $null$로 코딩하려면 채움 노드를 사용하십시오.

설명 필드 레이블을 지정하려면 이 텍스트 상자를 사용하십시오. 이 레이블은 스트림 특성 대화 상자에서 작성하는 선택사항에 따라 그래프, 테이블, 출력, 모델 브라우저 등의 다양한 위치에 표시됩니다.

연속형 데이터의 값 및 레이블 지정

숫자 필드에는 *연속형* 측정 수준이 사용됩니다. 연속형 데이터의 경우 세 가지 저장 유형이 있습니다.

- 실수
- 정수
- 날짜/시간

모든 연속형 필드를 편집하는 데 동일한 대화 상자가 사용됩니다. 저장 유형은 참조용으로만 표시됩니다.

값 지정

다음 제어는 연속형 필드에 고유하고 값 범위를 지정하는 데 사용됩니다.

하한. 값 범위의 하한을 지정하십시오.

상한. 값 범위의 상한을 지정하십시오.

레이블 지정

범위 필드의 값에 레이블을 지정할 수 있습니다. **레이블** 단추를 클릭하여 값 레이블을 지정하기 위한 별도의 대화 상자를 여십시오.

값 및 레이블 하위 대화 상자: 범위 필드에 대한 값 대화 상자에서 **레이블**을 클릭하면 범위의 값에 대한 레이블을 지정할 수 있는 새 대화 상자가 열립니다.

이 테이블의 *값* 및 *레이블* 열을 사용하여 값 및 레이블 쌍을 정의할 수 있습니다. 현재 정의된 쌍이 여기에 표시됩니다. 비어 있는 셀을 클릭한 후 값 및 해당 레이블을 입력하여 새 레이블 쌍을 추가할 수 있습니다. 참고 : 이 테이블에 값/값-레이블 쌍을 추가해도 새 값이 필드에 추가되지는 않습니다. 대신 필드 값에 대한 메타데이터만 작성됩니다.

유형 노드에서 지정하는 레이블은 스트림 특성 대화 상자에서 작성하는 선택사항에 따라 많은 위치에(도구 팁, 출력 레이블 등으로) 표시됩니다.

명목 및 순서 데이터의 값 및 레이블 지정

명목(세트) 및 순서(순서 지정된 세트) 측정 수준은 데이터 값이 세트의 멤버로서 따로따로 사용됨을 표시합니다. 세트의 저장 유형은 문자열, 정수, 실수 또는 날짜/시간입니다.

다음 제어는 명목 및 순서 필드에 고유하고 값 및 레이블을 지정하는 데 사용됩니다.

값. 테이블의 *값* 열을 사용하여 현재 필드에 대한 지식을 기반으로 값을 지정할 수 있습니다. 이 테이블을 사용하여 필드에 예상되는 값을 입력하고 값 검사 드롭 다운 목록을 사용하여 데이터 세트가 이러한 값과 일치하는지 검사할 수 있습니다. 화살표와 삭제 단추를 사용하면 기존 값을 수정할 뿐만 아니라 값을 다시 정렬하고 삭제할 수 있습니다.

레이블. *레이블* 열을 사용하여 세트에 있는 각 값에 대해 레이블을 지정할 수 있습니다. 이러한 레이블은 스트림 특성 대화 상자에서 선택하는 항목에 따라 그래프, 테이블, 출력 및 모델 브라우저 등의 다양한 위치에 표시됩니다.

플래그의 값 지정

플래그 필드는 두 개의 고유 값을 갖는 데이터를 표시하는 데 사용됩니다. 플래그의 저장 유형은 문자열, 정수, 실수 또는 날짜/시간입니다.

참. 조건이 충족될 때 필드의 플래그 값을 지정합니다.

거짓. 조건이 충족되지 않을 때 필드의 플래그 값을 지정합니다.

레이블. 플래그 필드의 값마다 레이블을 지정합니다. 이러한 레이블은 스트림 특성 대화 상자에서 선택하는 항목에 따라 그래프, 테이블, 출력 및 모델 브라우저 등의 다양한 위치에 표시됩니다.

콜렉션 데이터의 값 지정

콜렉션 필드는 목록에 있는 비지리 공간적 데이터를 표시하는 데 사용됩니다.

콜렉션 **측정** 수준에 대해 설정할 수 있는 유일한 항목은 **목록 측도**입니다. 기본적으로 이 측도는 유형 없음으로 설정되지만, 다른 값을 선택하여 목록 내에 있는 요소의 측정 수준을 설정할 수 있습니다. 다음 옵션 중 하나를 선택할 수 있습니다.

- 유형 없음
- 연속형
- 명목
- 순서
- 플래그

지리 공간적 데이터의 값 지정

지리 공간적 필드는 목록에 있는 지리 공간적 데이터를 표시하는 데 사용됩니다.

지리 공간적 **측정** 수준의 경우, 다음 옵션을 설정하여 목록에 있는 요소의 측정 수준을 설정할 수 있습니다.

유형 지리 공간적 필드의 측정 하위 수준을 선택하십시오. 사용 가능한 하위 수준은 목록 필드의 깊이로 결정됩니다. 기본값은 Point(깊이 0), LineString(깊이 1) 및 Polygon(깊이 1)입니다.

하위 수준에 대한 자세한 정보는 142 페이지의 『지리 공간적 측정 수준』의 내용을 참조하십시오.

목록 깊이에 대한 자세한 정보는 11 페이지의 『목록 저장 공간 및 연관된 측정 수준』의 내용을 참조하십시오.

좌표계 이 옵션은 측정 수준을 비지리 공간적 수준에서 지리 공간적 수준으로 변경한 경우에만 사용할 수 있습니다. 지리 공간적 데이터에 좌표계를 적용하려면 이 선택란을 선택하십시오. 기본적으로, **도구 > 스트림 특성 > 옵션 > 지리 공간적** 분할창에서 설정된 좌표계가 표시됩니다. 다른 좌표계를 사용하려면, **변경** 단추를 클릭하여 좌표계 선택 대화 상자를 표시하고 필요한 좌표계를 선택하십시오.

좌표계에 대한 자세한 정보는 SPSS Modeler 사용자 안내서의 스트림에 대한 작업 섹션에서 스트림의 지리 공간적 옵션 설정 주제를 참조하십시오.

결측값 정의

유형 탭의 **결측** 열은 필드에 대해 결측값 처리가 정의되었는지 여부를 표시합니다. 가능한 설정은 다음과 같습니다.

On(*). 이 필드에 대해 결측값 처리가 정의되어 있음을 표시합니다. 지정 옵션(아래 참조)을 사용하여 명시적으로 지정하거나 다운스트림 채움 노드를 통해 결측값 처리를 정의할 수 있습니다.

Off. 필드에 대해 결측값 처리가 정의되어 있지 않습니다.

지정. 이 필드에 대해 결측값으로 간주할 명시적 값을 선언할 수 있는 대화 상자를 표시하려면 이 옵션을 선택하십시오.

유형 값 검사

각 필드에 대해 검사 옵션을 사용하면 해당 필드의 모든 값을 조사하여 값이 현재 유형 설정을 따르는지 또는 값 지정 대화 상자에 지정한 값을 따르는지 판별합니다. 이 옵션은 단일 조작 내에서 데이터 세트를 정리하고 데이터 세트의 *크기*를 줄이는 데 유용합니다.

유형 노드 대화 상자에 있는 *검사* 열의 설정은 유형 한계를 벗어난 값이 발견될 때 발생하는 상황을 결정합니다. 필드에 대한 검사 설정을 변경하려면 *검사* 열에서 해당 필드에 대한 드롭 다운 목록을 사용하십시오. 모든 필드에 대한 검사 설정을 설정하려면 *필드* 열에서 클릭하고 Ctrl-A를 누르십시오. 그런 다음 *검사* 열에서 임의 필드에 대한 드롭 다운 목록을 사용하십시오.

다음 검사 설정을 사용할 수 있습니다.

없음. 값이 검사 없이 전달됩니다. 기본 설정입니다.

무효화. 한계를 벗어난 값을 시스템 널($null$)로 변경합니다.

강제 적용. 해당 측정 수준이 완전히 인스턴스화된 필드에 지정된 범위를 벗어난 값이 있는지 조사합니다. 지정되지 않은 값은 다음 규칙에 따라 해당 측정 수준에 유효한 값으로 변환됩니다.

- 플래그의 경우, true 및 false 값 이외의 값은 false 값으로 변환됩니다.
- 세트(명목 또는 순서)의 경우, 알 수 없는 값은 세트 값의 첫 번째 멤버로 변환됩니다.
- 범위의 상한보다 큰 숫자는 상한으로 대체됩니다.
- 범위의 하한보다 작은 숫자는 하한으로 대체됩니다.
- 범위 내의 널값에는 해당 범위의 중간 값이 제공됩니다.

삭제. 유효하지 않은 값이 있으면 전체 레코드가 삭제됩니다.

경고. 모든 데이터를 읽은 후에는 스트림 특성 대화 상자에서 유효하지 않은 항목 수가 계산되고 보고됩니다.

중단. 유효하지 않은 값이 처음 발견될 때 스트림 실행이 종료됩니다. 스트림 특성 대화 상자에서 오류가 보고됩니다.

필드 역할 설정

필드의 역할은 모델 작성에 필드가 사용되는 방법을 지정합니다. 예를 들어, 필드가 입력인지 또는 목표(예측 목표)인지 여부입니다.

참고: 파티션, 빈도 및 레코드 ID 역할은 각각 단일 필드에만 적용할 수 있습니다.

다음 역할을 사용할 수 있습니다.

입력. 필드가 시스템 학습에 대한 입력으로 사용됩니다(예측변수 필드).

목표. 필드가 시스템 학습의 출력 또는 목표로 사용됩니다(모델이 예측하려는 필드 중 하나).

둘 다. 필드가 Apriori 노드에서 입력 및 출력 둘 다로 사용됩니다. 기타 모든 모델링 노드에서는 해당 필드를 무시합니다.

없음. 필드가 시스템 학습에서 무시됩니다. 해당 측정 수준이 **유형 없음**으로 설정된 필드는 역할 열에서 자동으로 **없음**으로 설정됩니다.

파티션. 훈련, 테스트 및 검증을 위해 데이터를 개별 표본으로 파티셔닝하는 데 사용되는 필드를 표시합니다. 필드는 두 개 또는 세 개의 가능한 값(필드 값 대화 상자에 정의된)을 갖는 인스턴스화된 세트 유형이어야 합니다. 첫 번째 값은 훈련 표본을 나타내고 두 번째 값은 테스트 표본을 나타내며 세 번째 값(있는 경우)은 검증 표본을 나타냅니다. 추가 값은 무시되며 플래그 필드를 사용할 수 없습니다. 분석에서 파티션을 사용하려면, 해당 모델 작성 또는 분석 노드의 모델 옵션 탭에서 파티셔닝을 사용 가능하게 설정해야 합니다. 파티셔닝이 사용될 때, 파티션 필드에 대해 널값을 갖는 레코드는 분석에서 제외됩니다. 스트림에 여러 파티션 필드를 정의한 경우, 적용 가능한 각 모델링 노드의 필드 탭에서 단일 파티션 필드를 지정해야 합니다. 데이터에 아직 적합한 필드가 없으면 파티션 노드 또는 파생 노드를 사용하여 필드를 작성할 수 있습니다. 자세한 정보는 181 페이지의 『파티션 노드』의 내용을 참조하십시오.

분할. (명목, 순서 및 플래그 필드만 해당) 필드의 가능한 값마다 모델이 작성되도록 지정합니다.

빈도. (숫자 필드만 해당) 이 역할을 설정하면 필드 값을 레코드의 빈도 가중치로 사용할 수 있습니다. 이 기능은 C&R 트리, CHAID, QUEST 및 선형 모델에서만 지원됩니다. 기타 모든 노드에서는 이 역할을 무시합니다. 빈도 가중치는 이 기능을 지원하는 모델링 노드의 필드 탭에서 **빈도 가중치 사용** 옵션을 사용하여 사용합니다.

레코드 ID. 필드가 고유 레코드 식별자로 사용됩니다. 이 기능은 대부분의 노드에서 무시되지만 선형 모형에서 지원되고 IBM Netezza In-Database 마이닝 노드에 필요합니다.

유형 속성 복사

하나의 필드에서 다른 필드로 유형의 속성(예: 값, 검사 옵션 및 결측값)을 쉽게 복사할 수 있습니다.

1. 해당 속성을 복사할 필드에서 마우스 오른쪽 단추를 클릭하십시오.
2. 컨텍스트 메뉴에서 **복사**를 선택하십시오.
3. 해당 속성을 변경할 필드에서 마우스 오른쪽 단추를 클릭하십시오.
4. 컨텍스트 메뉴에서 **특수 속성 붙여넣기**를 선택하십시오. *참고*: Ctrl-클릭 방법을 사용하거나 컨텍스트 메뉴에서 **필드 선택** 옵션을 사용하여 여러 필드를 선택할 수 있습니다.

새 대화 상자가 열리고 여기서 붙여넣을 특정 속성을 선택할 수 있습니다. 여러 필드에 붙여넣는 경우, 여기서 선택하는 옵션은 모든 목표 필드에 적용됩니다.

다음 속성을 붙여넣으십시오. 아래의 목록에서 선택하여 하나의 필드에서 다른 필드로 속성을 붙여넣으십시오.

- **유형.** 측정 수준을 붙여넣으려면 선택하십시오.
- **값.** 필드 값을 붙여넣으려면 선택하십시오.
- **결측.** 결측값 설정을 붙여넣으려면 선택하십시오.

- **검사.** 값 검사 옵션을 붙여넣으려면 선택하십시오.
- **역할.** 필드의 역할을 붙여넣으려면 선택하십시오.

필드 형식 설정 탭

테이블 및 유형 노드의 형식 탭에는 현재 또는 사용하지 않은 필드와 각 필드에 대한 형식 옵션의 목록이 표시됩니다. 필드 형식 필드의 각 열에 대한 설명은 다음과 같습니다.

필드. 선택된 필드의 이름을 표시합니다.

형식. 이 열의 셀을 두 번 클릭하면 열리는 대화 상자를 사용하여 필드에 대한 형식을 개별적으로 지정할 수 있습니다. 자세한 정보는 153 페이지의 『필드 형식 옵션 설정』 주제를 참조하십시오. 여기서 지정된 형식은 전체 스트림 특성에서 지정된 형식을 대체합니다.

참고: Statistics 내보내기 및 Statistics 출력 노드는 필드별 형식을 해당 메타데이터에 포함하는 *.sav* 파일을 내보냅니다. IBM SPSS Statistics *.sav* 파일 형식에서 지원하지 않는 필드별 형식이 지정된 경우 노드는 IBM SPSS Statistics 기본 형식을 사용합니다.

맞춤. 테이블 열 내에서 값을 맞추는 방식을 지정하려면 이 열을 사용하십시오. 기본 설정은 **자동**이며 이는 기호 값은 왼쪽으로 맞추고 숫자 값은 오른쪽으로 맞춥니다. **왼쪽**, **오른쪽** 또는 **가운데**를 선택하여 기본값을 대체할 수 있습니다.

열 너비. 기본적으로 열 너비는 필드의 값을 기반으로 자동으로 계산됩니다. 자동 너비 계산을 대체하려면 테이블 셀을 클릭한 후 드롭 다운 목록을 사용하여 새 너비를 선택하십시오. 여기에 나열되지 않는 사용자 정의 너비를 입력하려면 필드 또는 형식 열에서 테이블 셀을 두 번 클릭하여 필드 형식 하위 대화 상자를 여십시오. 또는 셀을 마우스 오른쪽 단추로 클릭한 후 **형식 설정**을 선택할 수 있습니다.

현재 필드 보기. 기본적으로 대화 상자에는 현재 활성 필드의 목록이 표시됩니다. 사용하지 않은 필드의 목록을 보려면 **사용하지 않은 필드 설정 보기**를 선택하십시오.

컨텍스트 메뉴. 이 탭에 대한 컨텍스트 메뉴는 다양한 선택사항 및 설정 업데이터 옵션을 제공합니다. 열을 마우스 오른쪽 단추로 클릭하여 이 메뉴를 표시하십시오.

- **모두 선택.** 모든 필드를 선택합니다.
- **선택 안함.** 선택사항을 선택 취소합니다.
- **필드 선택.** 유형 또는 저장 공간 특성을 기반으로 필드를 선택합니다. 옵션은 **범주형 선택**, **연속형 선택**(숫자), **유형 없음 선택**, **문자열 선택**, **숫자 선택** 또는 **날짜/시간 선택**입니다. 자세한 정보는 140 페이지의 『측정 수준』의 내용을 참조하십시오.
- **형식 설정.** 필드별로 날짜. 시간 및 소수점 옵션을 지정하는 데 필요한 하위 대화 상자를 엽니다.
- **맞춤 설정.** 선택된 필드에 대한 맞춤을 설정합니다. 옵션은 **자동**, **가운데**, **왼쪽** 또는 **오른쪽**입니다.
- **열 너비 설정.** 선택된 필드의 필드 너비를 설정합니다. 데이터에서 너비를 읽어오려면 **자동**을 지정하십시오. 또는 필드 너비를 5, 10, 20, 30, 50, 100 또는 200으로 설정할 수 있습니다.

필드 형식 옵션 설정

필드 형식은 유형 및 테이블 노드의 형식 탭에서 사용할 수 있는 하위 대화 상자에서 지정됩니다. 이 대화 상자를 열기 전에 둘 이상의 필드를 선택한 경우에는 선택사항 첫 번째 필드의 설정이 모두에 대해 사용됩니다. 여기서 지정한 후 **확인**을 클릭하면 이 설정이 형식 탭에서 선택된 모든 필드에 적용됩니다.

필드별로 다음과 같은 옵션을 사용할 수 있습니다. 이들 옵션 중 다수는 스트림 특성 대화 상자에서도 지정할 수 있습니다. 필드 수준에서 작성된 설정은 스트림에 대해 지정된 기본값을 대체합니다.

날짜 형식. 날짜 저장 공간 필드에 사용하거나 문자열이 CLEM 날짜 함수에 의해 날짜로 해석될 때 사용할 날짜 형식을 선택합니다.

시간 형식. 시간 저장 공간 필드에 사용하거나 문자열이 CLEM 시간 함수에 의해 시간으로 해석될 때 사용할 시간 형식을 선택합니다.

숫자 표시 형식. 표준(**####.###**), 지수표기(**#.###E+##**) 또는 통화 표시 형식(**$###.##**)에서 선택할 수 있습니다.

소수점 기호. 쉼표(,) 또는 마침표(.)를 소수점 구분자로 선택하십시오.

그룹 기호. 숫자 표시 형식에 대해 값을 그룹화하는 데 사용된 기호를 선택하십시오(예: *3,000.00*의 쉼표). 옵션에는 없음, 마침표, 쉼표, 공백 및 정의된 로케일이 포함됩니다(이 경우 현재 로케일의 기본값을 사용함).

소수점 이하 자릿수(표준, 지수, 통화, 내보내기). 숫자 표시 형식에 대해 실수를 표시할 때 사용할 소수점 이하 자릿수를 지정합니다. 이 옵션은 각 표시 형식마다 별도로 지정됩니다.

맞춤. 열에서 값을 맞추는 방식을 지정합니다. 기본 설정은 **자동**이며 이는 기호 값은 왼쪽으로 맞추고 숫자 값은 오른쪽으로 맞춥니다. 왼쪽, 오른쪽 또는 가운데를 선택하여 기본값을 대체할 수 있습니다.

열 너비. 기본적으로 열 너비는 필드의 값을 기반으로 자동으로 계산됩니다. 목록 상자 오른쪽의 화살표를 사용하여 5구간의 사용자 너비를 지정할 수 있습니다.

필드 필터링 또는 이름 바꾸기

스트림의 어느 지점에서나 필드의 이름을 바꾸고 필드를 제외할 수 있습니다. 예를 들어, 의학 연구자로서 환자(레코드 수준 데이터)의 칼륨 수준(필드 수준 데이터)에 대해 관심이 없을 수 있으므로 *K*(칼륨) 필드를 필터링할 수 있습니다. 이는 소스 또는 출력 노드의 필터 탭을 사용하거나 별도의 필터 노드를 사용하여 수행할 수 있습니다. 액세스되는 노드에 관계없이 기능은 동일합니다.

- 가변파일, 고정 파일, Statistics 파일, XML 등의 소스 노드에서 데이터를 IBM SPSS Modeler로 읽어올 때 필드의 이름을 바꾸거나 필드를 필터링할 수 있습니다.
- 필터 노드를 사용하면 스트림의 어느 지점에서나 필드의 이름을 바꾸거나 필드를 필터링할 수 있습니다.
- Statistics 내보내기, Statistics 변환, Statistics 모델 및 Statistics 출력 노드에서 IBM SPSS Statistics 이름 지정 표준을 준수하도록 필드를 필터링하거나 필드의 이름을 바꿀 수 있습니다. 자세한 정보는 387 페이지의 『IBM SPSS Statistics에 대한 필드 이름 변경 또는 필터링』의 내용을 참조하십시오.

- 위 노드의 필터 탭을 사용하여 다중 응답 세트를 정의하거나 편집할 수 있습니다. 자세한 정보는 156 페이지의 『다중 응답 세트 편집』의 내용을 참조하십시오.
- 최종적으로 필터 노드를 사용하여 한 소스 노드에서 다른 소스 노드로 필드를 맵핑할 수 있습니다.

필터링 옵션 설정

필터 탭에서 사용되는 테이블은 노드에 들어갈 때와 노드에서 나갈 때 각 필드의 이름을 표시합니다. 이 테이블의 옵션을 사용하여 중복이거나 다운스트림 조작에 불필요한 필드의 이름을 바꾸거나 해당 필드를 필터링할 수 있습니다.

- **필드.** 현재 연결된 데이터 소스의 입력 필드를 표시합니다.
- **필터.** 모든 입력 필드의 필터 상태를 표시합니다. 필터링된 필드에는 이 필드가 다운스트림으로 전달되지 않음을 나타내는 빨간색 X가 이 열에 포함되어 있습니다. 선택된 필드에 대해 *필터* 열을 클릭하여 필터링을 켜고 끄십시오. Shift+클릭 선택 방법을 사용하여 동시에 여러 필드에 대한 옵션을 선택할 수도 있습니다.
- **필드.** 필드가 필터 노드를 나갈 때 필드를 표시합니다. 중복 이름은 빨간색으로 표시됩니다. 이 열을 클릭한 후 새 이름을 입력하여 필드 이름을 편집할 수 있습니다. 또는 *필터* 열을 클릭하여 중복 필드를 사용 안함으로 설정하여 필드를 제거하십시오.

테이블의 모든 열은 열 헤더를 클릭하여 정렬할 수 있습니다.

현재 필드 보기. 필터 노드에 활성 상태로 연결된 데이터 세트에 대한 필드를 보려면 선택하십시오. 이 옵션은 기본적으로 선택되며 필터 노드를 사용하는 가장 일반적인 방법입니다.

사용하지 않은 필드 설정 보기. 필터 노드에 한 번 연결되었지만 더 이상 연결되지 않는 데이터 세트에 대한 필드를 보려면 선택하십시오. 이 옵션은 한 스트림에서 다른 스트림으로 필터 노드를 복사하거나 필터 노드를 저장하고 재로드할 때 유용합니다.

필터 단추 메뉴

대화 상자의 왼쪽 상단에 있는 필터 단추를 클릭하여 다수의 단축키 및 기타 옵션을 제공하는 메뉴에 액세스하십시오.

다음과 같은 작업을 수행하도록 선택할 수 있습니다.

- 모든 필드 제거
- 모든 필드 포함
- 모든 필드 토글
- 중복 제거. *참고*: 이 옵션을 선택하면 중복 이름의 모든 발생(첫 번째 발생 포함)이 제거됩니다.
- 다른 애플리케이션과 부합하도록 필드 및 다중 응답 세트의 이름 바꾸기. 자세한 정보는 387 페이지의 『IBM SPSS Statistics에 대한 필드 이름 변경 또는 필터링』의 내용을 참조하십시오.
- 필드 이름 자르기
- 필드 및 다중 응답 세트 이름 값 익명화
- 입력 필드 이름 사용

- 다중 응답 세트 편집. 자세한 정보는 156 페이지의 『다중 응답 세트 편집』의 내용을 참조하십시오.
- 기본 필터 상태 설정

대화 상자의 맨 위에 있는 화살표 토글 단추를 사용하여 기본적으로 필드를 포함할지 아니면 삭제할지 지정할 수도 있습니다. 이는 몇몇 필드만 다운스트림으로 포함되는 큰 데이터 세트의 경우 유용합니다. 예를 들어, 삭제할 모든 필드를 개별적으로 선택하는 대신 보존할 필드만 선택한 후 다른 모든 필드는 삭제하도록 지정할 수 있습니다.

필드 이름 자르기

필터 단추 메뉴(필터 탭의 상단 왼쪽 구석)에서 필드 이름을 자르도록 선택할 수 있습니다.

최대 길이. 필드 이름 길이를 제한하기 위한 문자 수를 지정하십시오.

숫자 수. 필드 이름을 잘랐을 때 이름이 더 이상 고유하지 않는 경우, 필드 이름을 더 자르고 이름에 숫자를 추가하여 차별화할 수 있습니다. 사용되는 숫자 수를 지정할 수 있습니다. 화살표 단추를 사용하여 숫자를 조정하십시오.

예를 들어, 다음 표에서는 기본 설정(최대 길이=8 및 숫자 수=2)을 사용하여 의학 데이터 세트의 필드 이름을 자르는 방법을 보여줍니다.

표 22. 필드 이름 자르기

필드 이름	잘린 필드 이름
Patient Input 1	Patien01
Patient Input 2	Patien02
Heart Rate	HeartRat
BP	BP

필드 이름 익명화

왼쪽 맨 위에서 필터 단추 메뉴를 클릭하고 **필드 이름 익명화**를 선택하여 필터 탭이 포함된 노드에서 필드 이름을 익명화할 수 있습니다. 익명화된 필드 이름은 문자열 접두부와 고유 숫자 기반 값으로 구성되어 있습니다.

해당 이름 익명화. 필터 탭에서 이미 선택된 필드의 이름만 익명화하려면 **선택된 필드만**을 선택하십시오. 기본값은 **모든 필드**이며 모든 필드 이름을 익명화합니다.

필드 이름 접두부. 익명화된 필드 이름의 기본 접두부는 **anon_**입니다. 다른 접두부를 원하면 **사용자 정의**를 선택하고 직접 접두부를 입력하십시오.

다중 응답 세트 익명화. 여러 응답 세트의 이름을 필드와 동일한 방식으로 익명화합니다. 자세한 정보는 156 페이지의 『다중 응답 세트 편집』의 내용을 참조하십시오.

원래의 필드 이름을 복원하려면 필터 단추 메뉴에서 **입력 필드 이름 사용**을 선택하십시오.

다중 응답 세트 편집

왼쪽 상단의 필터 단추 메뉴를 클릭한 후 **다중 응답 세트 편집**을 선택하여 필터를 포함하는 노드에서 다중 응답 세트를 추가하거나 편집할 수 있습니다.

다중 응답 세트는 각 케이스(예: 설문조사 응답자에게 방문한 박물관 또는 읽은 잡지를 묻는 경우)에 대해 둘 이상의 값을 가질 수 있는 데이터를 기록하는 데 사용됩니다. 다중 응답 세트는 Data Collection 소스 노드 또는 통계 파일 소스 노드를 사용하여 IBM SPSS Modeler로 가져오거나 필터 노드를 사용하여 IBM SPSS Modeler에서 정의할 수 있습니다.

새로 작성을 클릭하여 새 다중 응답 세트를 작성하거나 **편집**을 클릭하여 기존 세트를 수정하십시오.

이름 및 레이블. 세트에 대한 이름 및 설명을 지정합니다.

유형. 다중 응답 질문은 두 방법 중 하나로 처리할 수 있습니다.

- **다중 이분형 세트.** 각각의 가능한 응답에 대해 별도의 플래그 필드가 작성되므로 10개의 잡지가 있으면 10개의 플래그 필드가 있으며 각각의 플래그 필드는 *참* 또는 *거짓*에 대해 0 또는 1과 같은 값을 가질 수 있습니다. 계수된 값을 사용하면 참으로 계수되는 값을 지정할 수 있습니다. 이 방법은 응답자가 적용되는 모든 옵션을 선택할 수 있게 하려는 경우에 유용합니다.
- **다중 범주 세트.** 지정된 응답자의 최대 응답 수까지 각 응답에 대해 명목 필드가 작성됩니다. 각각의 명목 필드는 *Time*에 대한 *1*, *Newsweek*에 대한 *2*, *PC Week*에 대한*3* 등의 가능한 응답을 나타내는 값을 가지고 있습니다. 이 방법은 응답 수를 제한하려는 경우(예: 응답자에게 가장 자주 읽은 세 개의 잡지를 선택하도록 요구하는 경우)에 가장 유용합니다.

세트의 필드. 오른쪽의 아이콘을 사용하여 필드를 추가하거나 제거하십시오.

설명

- 다중 응답 세트에 포함된 모든 필드는 동일한 저장 공간을 가지고 있어야 합니다.
- 세트는 포함하는 필드와 구별됩니다. 예를 들어, 세트를 삭제해도 세트에 포함된 필드는 삭제되지 않고 해당 필드 간 링크만 삭제됩니다. 세트는 여전히 삭제 지점에서 업스트림으로 표시되지만 다운스트림으로는 표시되지 않습니다.
- 필터 메뉴의 IBM SPSS Statistics에 대해 이름 바꾸기, **자르기** 또는 **값 익명화** 옵션을 선택하거나 탭에서 직접 필터 노드를 사용하여 필드의 이름을 바꾸는 경우 다중 응답 세트에서 사용되는 이 필드에 대한 참조도 업데이트됩니다. 하지만 필터 노드에 의해 삭제되는 다중 응답 세트의 필드는 다중 응답 세트에서 제거되지 않습니다. 이러한 필드는 더 이상 스트림에 표시되지 않지만 여전이 다중 응답 세트에 의해 참조되므로 예를 들어, 내보내기 수행 시 이를 고려할 수 있습니다.

파생 노드

IBM SPSS Modeler에서 가장 강력한 기능 중 하나는 데이터 값을 수정하고 기존 데이터에서 새 필드를 파생시키는 기능입니다. 긴 데이터 마이닝 프로젝트 동안에는 웹 로그 데이터의 문자열에서 고객 ID를 추출하거나 트랜잭션 및 인구 통계 데이터를 기반으로 고객 생애 가치를 작성하는 등의 여러 파생 작업을 수행하는 것이 일반적입니다. 이 변환은 모두 다양한 필드 작업 노드를 사용하여 수행할 수 있습니다.

몇몇 노드는 새 필드를 파생시키는 기능을 제공합니다.

파생 노드는 데이터 값을 수정하거나 하나 이상의 기존 필드로부터 새 필드를 작성합니다. 수식, 플래그, 명목형, 상태, 개수, 조건부 유형의 필드를 작성합니다.

재분류 노드는 한 세트의 범주형 값을 다른 값으로 변환합니다. 재분류는 분석을 위해 범주를 접거나 데이터를 재그룹화하는 데 유용합니다.

구간화 노드는 하나 이상의 기존 연속형(숫자 범위) 필드의 값을 기반으로 새 명목형(세트) 필드를 자동으로 작성합니다. 예를 들어, 연속형 수입 필드를 평균값에서의 편차로서 수입 그룹을 포함하는 새 범주형 필드로 변환할 수 있습니다. 새 필드에 대한 구간을 작성한 후에는 절단점을 기반으로 파생 노드를 생성할 수 있습니다.

플래그로 설정 노드는 하나 이상의 명목 필드에 대해 정의된 범주형 값을 바탕으로 다중 플래그 필드를 파생시킵니다.

구조변환 노드는 명목 또는 플래그 필드를 아직 또 다른 필드의 값으로 채워질 수 있는 필드 그룹으로 변환합니다. 예를 들어, *payment type*이라는 이름의 필드와 *credit*, *cash*, *debit*의 값이 주어진 경우, 각각이 실제 이루어진 지불의 값을 포함할 수 있는 세 개의 새 필드(*credit*, *cash*, *debit*)가 작성됩니다.

히스토리 노드는 이전 레코드의 필드에 있는 데이터를 포함하는 새 필드를 작성합니다. 히스토리 노드는 시계열 데이터 같은 순차 데이터에 가장 자주 사용합니다. 히스토리 노드를 사용하기 전에 정렬 노드를 사용하여 데이터를 정렬할 수 있습니다.

파생 노드 사용

파생 노드를 사용하면 하나 이상의 기존 필드에서 여섯 가지 유형의 새 필드를 작성할 수 있습니다.

- **수식.** 새 필드는 임의의 CLEM 표현식의 결과입니다.
- **플래그.** 새 필드는 지정된 조건을 나타내는 플래그입니다.
- **명목.** 새 필드는 명목이며 이는 해당 멤버가 지정된 값의 그룹임을 의미합니다.
- **상태.** 새 필드는 두 상태 중 하나입니다. 이 상태 사이에서의 전환은 지정된 조건에 의해 트리거됩니다.
- **개수.** 새 필드는 조건이 참인 횟수를 기반으로 합니다.

- **조건부.** 새 필드는 조건의 값에 따라 두 표현식 중 하나의 값입니다.

이 노드 각각의 파생 노드 대화 상자에는 특별한 옵션 세트가 포함되어 있습니다. 이 옵션은 후속 주제에서 다룹니다.

파생 노드에 대한 기본 옵션 설정

파생 노드에 대한 대화 상자의 맨 위에는 필요한 파생 노드의 유형을 선택할 수 있는 다수의 옵션이 있습니다.

모드. 다중 필드를 파생할지 여부에 따라 **단일** 또는 **다중**을 선택하십시오. **다중**을 선택하면 다중 파생 필드에 대한 옵션을 포함하도록 대화 상자가 변경됩니다.

파생 필드. 단순 파생 노드의 경우 파생하여 각 레코드에 추가할 필드의 이름을 지정하십시오. 기본 이름은 Derive*N*입니다. 여기서 *N*은 현재 세션 동안 지금까지 작성한 파생 노드의 수입니다.

파생 유형. 드롭 다운 목록에서 파생 노드의 유형(예: 수식 또는 명목)을 선택하십시오. 각각의 유형에 대해 유형별 대화 상자에서 사용자가 지정하는 조건을 기반으로 새 필드가 작성됩니다.

드롭 다운 목록에서 옵션을 선택하면 각 파생 노드 유형의 특성에 따라 새 제어 세트가 기본 대화 상자에 추가됩니다.

필드 유형. 새로 파생된 노드에 대해 측정 수준(예: 연속형, 범주형 또는 플래그)을 선택하십시오. 이 옵션은 파생 노드의 모든 양식에 공통입니다.

참고: 새 필드를 파생시키려면 특별한 함수 또는 수학 표현식을 사용해야 할 수 있습니다. 이 표현식 작성을 지원하기 위해 모든 유형의 파생 노드에 대해 대화 상자에서 표현식 작성기를 사용할 수 있으며 이 표현식 작성기는 CLEM 표현식의 전체 목록과 규칙 검사를 제공합니다.

다중 필드 파생

파생 노드 내에서 모드를 **다중**으로 설정하면 동일한 노드 내 동일한 조건을 기반으로 다중 필드를 파생시킬 수 있습니다. 이 기능은 데이터 세트의 여러 필드에서 동일한 변환을 작성하길 원할 때 시간을 절약합니다. 예를 들어, 최초 급여와 이전의 경험을 기반으로 현재 급여를 예측하는 회귀 모형을 작성하려는 경우 3개의 비대칭 변수 모두에 로그 변환을 적용하면 도움이 될 수 있습니다. 각각의 변환에 대해 새 파생 노드를 추가하는 대신 한 번에 모든 필드에 동일한 함수를 적용할 수 있습니다. 단순히 새 필드를 파생시킬 모든 필드를 선택한 후 필드 소괄호 내 `@FIELD` 함수를 사용하여 파생 표현식을 입력하십시오.

참고: `@FIELD` 함수는 동시에 여러 필드를 파생시키는 데 필요한 중요한 도구입니다. 이 함수를 사용하면 정확한 필드 이름을 지정하지 않고 현재 필드의 내용을 참조할 수 있습니다. 예를 들어, 다중 필드에 로그 변환을 적용하는 데 사용되는 CLEM 표현식은 `log(@FIELD)`입니다.

다중 모드를 선택하면 다음과 같은 옵션이 대화 상자에 추가됩니다.

파생 위치. 필드 선택기를 사용하여 새 필드를 파생시킬 필드를 선택하십시오. 각각의 선택된 필드에 대해 하나의 출력 필드가 생성됩니다. 참고: 선택된 필드의 저장 유형은 동일하지 않아도 됩니다. 하지만 모든 필드에 대해 조건이 유효하지 않으면 파생 조작이 실패합니다.

필드 이름 확장. 새 필드 이름에 추가할 확장자를 입력하십시오. 예를 들어, *Current Salary*의 로그가 포함된 새 필드에 대해 *log_* 확장자를 필드 이름에 추가하여 *log_Current Salary*를 생성할 수 있습니다. 단일 선택 단추를 사용하여 확장자를 필드 이름의 접두부(시작 부분에)와 접미부(끝 부분에) 중 어느 것으로 추가할지 선택하십시오. 기본 이름은 Derive*N*입니다. 여기서 *N*은 현재 세션 동안 지금까지 작성한 파생 노드의 수입니다.

단일 모드 파생 노드에서와 같이 이제 새 필드 파생에 사용할 표현식을 작성해야 합니다. 선택된 파생 조작의 유형에 따라 조건을 작성하는 다수의 옵션이 있습니다. 이 옵션은 후속 주제에서 다룹니다. 표현식을 작성하기 위해 간단하게 수식 필드를 입력하거나 계산기 단추를 클릭하여 표현식 작성기를 사용할 수 있습니다. 다중 필드에 대한 조작을 참조할 때 `@FIELD` 함수를 사용하는 것을 기억하십시오.

다중 필드 선택

다중 입력 필드에 대해 조작을 수행하는 모든 노드(예: 파생(다중 모드), 통합, 정렬, 다중 도표, 시간 구성)의 경우 필드 선택 대화 상자를 사용하여 쉽게 다중 필드를 선택할 수 있습니다.

정렬 기준. 다음 옵션 중 하나를 선택하여 볼 수 있는 필드를 정렬할 수 있습니다.

- **기본.** 필드가 데이터 스트림에서 현재 노드에 전달된 대로 필드의 순서를 보십시오.
- **이름.** 알파벳순으로 볼 필드를 정렬하십시오.
- **유형.** 측정 수준별로 정렬된 필드를 보십시오. 이 옵션은 특정 측정 수준을 갖는 필드를 선택할 때 유용합니다.

한 번에 하나씩 목록에서 필드를 선택하거나 Shift-클릭 및 Ctrl-클릭 방법을 사용하여 다중 필드를 선택하십시오. 또한 목록 아래의 단추를 사용하여 측정 수준을 기반으로 필드 그룹을 선택하거나 테이블의 모든 필드를 선택 또는 선택 취소할 수 있습니다.

수식 파생 옵션 설정

수식 파생 노드는 CLEM 표현식의 결과를 기반으로 데이터 세트에서 각 레코드에 대한 새 필드를 작성합니다. 이 표현식은 조건부 표현식이 되어서는 안 됩니다. 조건부 표현식을 기반으로 하는 값을 파생시키려면 파생 노드의 플래그 또는 조건부 유형을 사용하십시오.

수식 CLEM 언어를 사용하여 새 필드에 대한 값을 파생시켜 수식을 지정하십시오.

참고: SPSS Modeler는 파생된 목록 필드에 사용될 하위 측정 수준을 알 수 없으므로 콜렉션 및 지리 공간적 측정 수준에 대해 **지정...**을 클릭하여 값 대화 상자를 열고 필요한 하위 측정 수준을 설정할 수 있습니다. 자세한 정보는 160 페이지의 『파생된 목록 값 설정』의 내용을 참조하십시오.

지리 공간적 필드의 경우, 관련된 유일한 저장 유형은 **실수** 및 **정수**입니다(기본 설정은 **실수**임).

파생된 목록 값 설정

파생 노드 수식 **필드 유형** 드롭 다운 목록에서 **지정...**을 선택하면 값 대화 상자가 표시됩니다. 이 대화 상자에서는 수식 **필드 유형** 측정 수준 콜렉션 또는 지리 공간적에 사용될 하위 측정 수준 값을 설정합니다.

측정 콜렉션 또는 **지리 공간적**을 선택하십시오. 기타 측정 수준을 선택하면 편집 가능한 값이 없다는 메시지가 대화 상자에 표시됩니다.

콜렉션

콜렉션 **측정** 수준에 대해 설정할 수 있는 유일한 항목은 **목록 측도**입니다. 기본적으로 이 측도는 유형 없음으로 설정되지만 다른 값을 선택하여 목록 내 요소의 측정 수준을 설정할 수 있습니다. 다음 옵션 중 하나를 선택할 수 있습니다.

- 유형 없음
- 범주형
- 연속형
- 명목
- 순서
- 플래그

지리 공간적

지리 공간적 **측정** 수준의 경우 다음과 같은 옵션을 선택하여 목록 내 요소의 측정 수준을 설정할 수 있습니다.

유형 지리 공간적 필드의 측정 하위 수준을 선택하십시오. 사용 가능한 하위 수준은 목록 필드의 깊이에 의해 결정되며 기본값은 다음과 같습니다.

- 점(깊이 0)
- 선 스트링(깊이 1)
- 다각형(깊이 1)
- 다중 점(깊이 1)
- 다중 선 스트링(깊이 2)
- 다중 다각형(깊이 2)

하위 수준에 대한 자세한 정보는 SPSS Modeler 소스, 프로세스 및 출력 노드 안내서의 유형 노드 절에서 지리 공간적 측정 하위 수준 주제를 참조하십시오.

목록 깊이에 대한 자세한 정보는 SPSS Modeler 소스, 프로세스 및 출력 노드 안내서의 소스 노드 절에서 목록 저장 공간 및 연관된 측정 수준 주제를 참조하십시오.

좌표계 이 옵션은 측정 수준을 지리 공간적 수준에서 지리 공간적이 아닌 수준으로 변경한 경우에만 사용할 수 있습니다. 좌표계를 지리 공간적 데이터에 적용하려면 이 선택란을 선택하십시오. 기본적으로 **도구 > 스트**

림 특성 > 옵션 > 지리 공간적 분할창에서 설정된 좌표계가 표시됩니다. 다른 좌표계를 사용하려면 **변경** 단추를 클릭하여 좌표계 선택 대화 상자를 표시하고 데이터와 일치하는 좌표계를 선택하십시오.

좌표계에 대한 자세한 정보는 SPSS Modeler 사용자 안내서의 스트림에 대한 작업 절에서 스트림에 대한 지리 공간적 옵션 설정 주제를 참조하십시오.

목록 또는 지리 공간적 필드 파생

목록 항목으로 기록되어야 하는 데이터를 잘못된 속성을 가진 SPSS Modeler로 가져오는 경우가 있습니다. 예를 들어, 별도의 지리 공간적 필드(예: x 좌표 및 y 좌표 또는 위도 및 경도)로 또는 `.csv` 파일의 개별 행으로 가져오는 경우가 있습니다. 이 상황에서는 개별 필드를 단일 목록 필드로 결합해야 합니다. 이를 수행하는 한 가지 방법은 파생 노드를 사용하는 것입니다.

참고: 지리 공간적 데이터를 결합할 때는 x(또는 경도) 필드와 y(또는 위도) 필드를 알고 있어야 합니다. 결과 목록 필드가 지리 공간적 좌표의 표준 형식인 [x, y] 또는 [경도, 위도] 요소 순서를 가지도록 데이터를 결합해야 합니다.

다음의 단계에서는 목록 필드를 파생시키는 단순한 예를 보여줍니다.

1. 스트림에서 파생 노드를 소스 노드에 연결하십시오.
2. 파생 노드의 설정 탭에 있는 **파생 유형** 목록에서 **수식**을 선택하십시오.
3. **필드 유형**에서 **콜렉션**(지리 공간적 목록이 아닌 경우) 또는 **지리 공간적**을 선택하십시오. 기본적으로 SPSS Modeler는 "최선의 추측" 접근 방식을 사용하여 올바른 목록 세부사항을 설정합니다. **지정...**을 선택하여 값 대화 상자를 열 수 있습니다. 이 대화 상자에서 목록의 데이터에 대한 추가 정보를 입력할 수 있습니다. 예를 들어, 지리 공간적 목록에 대해 측정 수준을 변경할 수 있습니다.
4. **수식** 분할창에서 데이터를 올바른 목록 형식으로 결합하는 수식을 입력하십시오. 또는 계산기 단추를 클릭하여 표현식 작성기를 여십시오.

   목록을 파생시키는 수식의 단순한 예는 [x, y]입니다. 여기서 x 및 y는 데이터 소스에 있는 별도의 필드입니다. 작성되는 새 파생된 필드는 각 레코드의 값이 해당 레코드에 대해 연결된 x 및 y 값인 목록입니다.

   참고: 이 방식으로 목록에 결합되는 필드는 동일한 저장 유형을 가지고 있어야 합니다.

목록과 목록의 깊이에 대한 자세한 정보는 11 페이지의 『목록 저장 공간 및 연관된 측정 수준』의 내용을 참조하십시오.

파생 플래그 옵션 설정

파생 플래그 노드는 고혈압 또는 고객 계정 비활성화 등의 특정 조건을 표시하는 데 사용됩니다. 각각의 레코드에 대해 플래그 필드가 작성되며 참 조건이 충족되면 참에 대한 플래그 값이 필드에서 추가됩니다.

참 값. 아래에 지정된 조건과 일치하는 레코드에 대해 플래그 필드에 포함할 값을 지정하십시오. 기본값은 T입니다.

거짓 값. 아래에 지정된 조건과 일치하지 않는 레코드에 대해 플래그 필드에 포함할 값을 지정하십시오. 기본 값은 F입니다.

참일 조건. CLEM 조건을 지정하여 각 레코드의 특정 값을 평가하고 해당 레코드에 위에서 정의된 참 값 또는 거짓 값을 제공하십시오. 거짓이 아닌 숫자 값인 경우 참 값이 레코드에 제공됩니다.

참고: 빈 문자열을 리턴하려면 ""와 같이 사이에 아무것도 없는 여는 따옴표 및 닫는 따옴표를 입력해야 합니다. 빈 문자열은 예를 들어, 테이블에서 참 값을 더 명확하게 나타내기 위해 거짓 값으로 사용될 수 있습니다. 마찬가지로 따옴표로 묶지 않으면 숫자로 처리되는 문자열 값을 원하는 경우 따옴표를 사용해야 합니다.

예제

12.0 이전의 IBM SPSS Modeler 릴리스에서는 값이 쉼표로 구분된 다중 응답을 단일 필드로 가져왔습니다. 예:

```
museum_of_design,institute_of_textiles_and_fashion
museum_of_design
archeological_museum
$null$national_art_gallery,national_museum_of_science,other
```

이 데이터의 분석을 준비하기 위해 hassubstring 함수를 사용하여 다음과 같은 표현식으로 각 반응에 대해 별도의 플래그 필드를 생성할 수 있습니다.

```
hassubstring(museums,"museum_of_design")
```

파생 명목 옵션 설정

파생 명목 노드는 각각의 레코드가 충족하는 조건을 판별하기 위해 CLEM 조건 세트를 실행하는 데 사용됩니다. 각 레코드에 대해 조건이 충족되면 충족된 조건 세트를 표시하는 값이 파생된 새 필드에 추가됩니다.

기본값. 충족되는 조건이 없는 경우 새 필드에서 사용할 값을 지정하십시오.

필드 설정. 특정 조건이 충족되는 경우 새 필드에서 입력할 값을 지정하십시오. 목록에 있는 각각의 값은 인접 열에서 지정되는 연관된 조건을 가지고 있습니다.

이 조건이 참인 경우. 나열할 세트 필드에서 각 멤버에 대한 조건을 지정하십시오. 표현식 작성기를 사용하여 사용 가능한 함수 및 필드를 선택하십시오. 화살표 및 삭제 단추를 사용하여 조건을 다시 정렬하거나 제거할 수 있습니다.

조건은 데이터 세트에서 특정 필드의 값을 검정하여 작동합니다. 각각의 조건이 검정되면 위에서 지정된 값이 충족된 조건(있는 경우)을 표시하는 새 필드에 할당됩니다. 충족되는 조건이 없으면 기본값이 사용됩니다.

파생 상태 옵션 설정

파생 상태 노드는 파생 플래그 노드와 비슷합니다. 플래그 노드는 현재 레코드에 대한 *단일* 조건의 이행에 따라 값을 설정하지만 파생 상태 노드는 *두개의 독립* 조건을 이행하는 방식에 따라 필드의 값을 변경할 수 있습니다. 이는 각각의 조건이 충족되면 값이 변경됨(켜짐 또는 꺼짐)을 의미합니다.

초기 상태. 초기에 새 필드의 각 레코드에 **On** 값과 **Off** 값 중 어느 값을 지정할지 선택하십시오. 이 값은 각각의 조건이 충족되면 변경될 수 있습니다.

"On" 값. On 조건이 충족될 때 새 필드의 값을 지정하십시오.

"On"으로 전환 조건. 조건이 참일 때 상태를 On으로 변경할 CLEM 조건을 지정하십시오. 계산기 단추를 클릭하여 표현식 작성기를 여십시오.

"Off" 값. Off 조건이 충족될 때 새 필드에 대한 값을 지정하십시오.

"Off"로 전환 조건. 조건이 거짓일 때 상태를 Off로 변경할 CLEM 조건을 지정하십시오. 계산기 단추를 클릭하여 표현식 작성기를 여십시오.

참고: 빈 문자열을 지정하려면 ""와 같이 사이에 아무것도 없는 여는 따옴표 및 닫는 따옴표를 입력해야 합니다. 마찬가지로 따옴표로 묶지 않으면 숫자로 처리되는 문자열 값을 원하는 경우 따옴표를 사용해야 합니다.

파생 개수 옵션 설정

파생 개수 노드는 일련의 조건을 데이터 세트의 숫자 필드 값에 적용하는 데 사용됩니다. 각각의 조건이 충족되면 파생된 개수 필드의 값이 설정된 증분만큼 증가합니다. 이 유형의 파생 노드는 시계열 데이터의 경우 유용합니다.

초기값. 새 필드의 실행에 사용되는 값을 설정합니다. 초기값은 숫자 상수여야 합니다. 화살표 단추를 사용하여 값을 늘리거나 줄이십시오.

증분 조건. 충족되면 증분량에서 지정된 수를 기반으로 파생된 값을 변경할 CLEM 조건을 지정하십시오. 계산기 단추를 클릭하여 표현식 작성기를 여십시오.

증분량. 개수를 증분시키는 데 사용되는 값을 설정하십시오. CLEM 표현식의 결과 또는 숫자 상수를 사용할 수 있습니다.

재설정 조건. 충족되면 파생된 값을 초기값으로 재설정할 조건을 지정하십시오. 계산기 단추를 클릭하여 표현식 작성기를 여십시오.

파생 조건부 옵션 설정

파생 조건부 노드에서는 일련의 If-Then-Else문을 사용하여 새 필드의 값을 파생시킵니다.

If. 실행 시 각 레코드에 대해 평가될 CLEM 조건을 지정하십시오. 조건이 true인 경우(숫자의 경우 false가 아닌 경우) Then 표현식에 의해 아래에서 지정된 값이 새 필드에 제공됩니다. 계산기 단추를 클릭하여 표현식 작성기를 여십시오.

Then. 위의 If문이 true인 경우(또는 false가 아닌 경우) 새 필드에 대한 CLEM 표현식 또는 값을 지정하십시오. 계산기 단추를 클릭하여 표현식 작성기를 여십시오.

Else. 위의 If문이 false인 경우 새 필드에 대한 CLEM 표현식 또는 값을 지정하십시오. 계산기 단추를 클릭하여 표현식 작성기를 여십시오.

파생 노드를 사용하여 값 코딩변경

파생 노드는 예를 들어, 범주형 값을 가진 문자열 필드를 숫자 명목(세트) 필드로 변환하여 값의 코딩을 변경하는 데도 사용할 수 있습니다.

1. 파생 유형에 대해 필드의 유형(명목, 플래그 등)을 적절하게 선택하십시오.
2. 값의 코딩변경에 대한 조건을 지정하십시오. 예를 들어, `Drug='drugA'`인 경우 값을 1로 설정하고 `Drug='drugB'`인 경우 값을 2로 설정하는 방식으로 값을 설정할 수 있습니다.

채움 노드

채움 노드는 필드 값을 바꾸고 저장 공간을 변경하는 데 사용됩니다. 지정된 CLEM 조건(예: `@BLANK(FIELD)`)을 기반으로 값을 바꾸도록 선택할 수 있습니다. 또는 모든 공백 또는 널값을 특정 값으로 대체할 것을 선택할 수 있습니다. 채움 노드는 종종 유형 노드와 함께 사용되어 결측값을 바꿉니다. 예를 들어, `@GLOBAL_MEAN`과 같은 표현식을 지정하여 필드의 평균값으로 공백을 채울 수 있습니다. 이 표현식은 전역값 설정 노드에 의해 계산된 대로 평균값으로 모든 공백을 채웁니다.

필드 채우기. 텍스트 필드의 오른쪽에 있는 필드 선택기 단추를 사용하여 값을 검토하여 바꿀 필드를 데이터 세트에서 선택하십시오. 기본 작동은 아래에 지정된 조건 및 바꾸기 표현식에 따라 값을 바꾸는 것입니다. 아래의 바꾸기 옵션을 사용하여 대체 바꾸기 방법을 선택할 수도 있습니다.

참고: 사용자 정의 값으로 바꿀 필드를 여럿 선택하는 경우에는 필드 유형이 비슷한 것이 중요합니다(모두 숫자 또는 모두 기호).

바꾸기. 다음 방법 중 하나를 사용하여 선택된 필드의 값을 바꾸려면 선택하십시오.

- **조건 기반.** 이 옵션은 지정된 값으로 바꾸기 위한 조건으로 사용되는 표현식을 작성할 수 있도록 조건 필드 및 표현식 작성기를 활성화합니다.
- **항상.** 선택된 필드의 모든 값을 바꿉니다. 예를 들어, 이 옵션을 통해 CLEM 표현식 `(to_string (income))`을 사용하여 수입의 저장 공간을 문자열로 변환할 수 있습니다.
- **공백값.** 선택된 필드에서 사용자가 지정한 공백값을 모두 바꿉니다. 표준 조건 `@BLANK(@FIELD)`를 사용하여 공백을 선택합니다. *참고*: 유형 노드 또는 소스 노드의 유형 탭을 사용하여 공백을 정의할 수 있습니다.
- **널값.** 선택된 필드에서 모든 시스템 널값을 바꿉니다. 표준 조건 `@NULL(@FIELD)`을 사용하여 널을 선택합니다.
- **공백값 및 널값.** 선택된 필드에서 공백값과 시스템 널을 모두 바꿉니다. 이 옵션은 널이 결측값으로 정의되었는지 확신할 수 없는 경우에 유용합니다.

조건. 이 옵션은 **조건 기반** 옵션을 선택한 경우에만 사용할 수 있습니다. 선택된 필드를 평가하기 위해 CLEM 표현식을 지정하려면 이 텍스트 상자를 사용하십시오. 계산기 단추를 클릭하여 표현식 작성기를 여십시오.

바꾸기. 선택된 필드에 새 값을 제공하려면 CLEM 표현식을 지정하십시오. 텍스트 상자에 `undef`를 입력하여 값을 널값으로 바꿀 수도 있습니다. 계산기 단추를 클릭하여 표현식 작성기를 여십시오.

참고: 선택된 필드가 문자열인 경우에는 해당 필드를 문자열 값으로 바꿔야 합니다. 기본값인 0 또는 다른 숫자 값을 문자열 필드의 대체값으로 사용하면 오류가 발생합니다.

채움 노드를 사용한 저장 공간 변환

채움 노드의 바꾸기 조건을 사용하면 단일 또는 다중 필드에 대한 필드 저장 공간을 쉽게 변환할 수 있습니다. 예를 들어, 변환 함수 `to_integer`를 사용하면 CLEM 표현식 `to_integer(income)`를 사용하여 *income*을 문자열에서 정수로 변환할 수 있습니다.

사용 가능한 변환 함수를 보고 표현식 작성기를 사용하여 CLEM 표현식을 자동으로 작성할 수 있습니다. 함수 드롭 다운 목록에서 **변환**을 선택하여 저장 공간 변환 함수의 목록을 보십시오. 사용 가능한 변환 함수는 다음과 같습니다.

- `to_integer(ITEM)`
- `to_real(ITEM)`
- `to_number(ITEM)`
- `to_string(ITEM)`
- `to_time(ITEM)`
- `to_timestamp(ITEM)`
- `to_date(ITEM)`
- `to_datetime(ITEM)`

날짜 및 시간 값 변환. 변환 함수(및 날짜 또는 시간 값 등의 특정 입력 유형이 필요한 기타 함수)는 스트림 옵션 대화 상자에서 지정된 현재 형식에 따라 결정됩니다. 예를 들어, 값이 *Jan 2003*, *Feb 2003* 등의 값을 가진 문자열 필드를 날짜 저장 공간으로 변환하려면 **MON YYYY**를 스트림의 기본 날짜 형식으로 선택하십시오.

파생 계산 중 임시 변환의 경우 파생 노드에서도 변환 함수를 사용할 수 있습니다. 파생 노드를 사용하여 범주형 값을 가진 문자열 필드의 코딩 변경 등의 기타 조작을 수행할 수도 있습니다. 자세한 정보는 164 페이지의 『파생 노드를 사용하여 값 코딩변경』의 내용을 참조하십시오.

재분류 노드

재분류 노드는 한 세트의 범주형 값을 다른 값으로 변환합니다. 재분류는 분석을 위해 범주를 접거나 데이터를 재그룹화하는 데 유용합니다. 예를 들어, *제품*에 대한 값을 *부엌 제품*, *욕실 및 침구*, *가전 제품*이라는 세 그룹으로 다시 분류할 수 있습니다. 종종 이 작업은 값을 분류하고 재분류 노드를 생성하여 분포 노드에서 직접 수행됩니다. 자세한 정보는 245 페이지의 『분포 노드 사용』의 내용을 참조하십시오.

재분류는 하나 이상의 기호 필드에 대해 수행될 수 있습니다. 또한 기존 필드를 새로운 값으로 대체하거나 새 필드를 생성할 수 있습니다.

재분류 노드를 사용하는 시기

재분류 노드를 사용하기 전에 또 다른 필드 작업 노드가 현재 작업에 더 적절하지 않은지 고려하십시오.

- 자동 방법을 사용하여 순자 범위를 변수군(순위 또는 백분위수 등)으로 변환하려면 구간화 노드를 사용해야 합니다. 자세한 정보는 170 페이지의 『구간화 노드』 주제를 참조하십시오.
- 수치 범위를 수동으로 변수군으로 분류하려면 파생 노드를 사용해야 합니다. 예를 들어, 월급 값을 특정 월급 범위 범주로 축소하려면 파생 노드를 사용하여 각 범주를 수동으로 정의해야 합니다.
- 범주형 필드의 값을 기반으로 하여 *Mortgage_type*과 같은 하나 이상의 플래그 필드를 생성하려면 플래그로 설정 노드를 사용해야 합니다.
- 범주형 필드를 수치 저장 공간으로 변환하기 위해 파생 노드를 사용할 수 있습니다. 예를 들어, *아니오* 및 *예* 값을 각각 0 및 1로 변환할 수 있습니다. 자세한 정보는 164 페이지의 『파생 노드를 사용하여 값 코딩 변경』의 내용을 참조하십시오.

재분류 노드에 대한 옵션 설정

재분류 노드를 사용하기 위한 세 단계가 있습니다.

1. 먼저 다중 필드 또는 단일 필드 재분류를 선택해야 합니다.
2. 기존 필드로 코딩변경을 수행할 것인지 새 필드를 작성할 것인지 선택하십시오.
3. 그런 다음 재분류 노드 대화 상자의 동적 옵션을 사용하여 변수군을 원하는 대로 맵핑하십시오.

모드. 한 필드에 대한 범주를 재분류하려면 **단일**을 선택하십시오. 한 번에 둘 이상의 필드의 변환을 가능하게 하는 옵션을 활성화하려면 **다중**을 선택하십시오.

재분류. 원래 명목 필드를 유지하고 재분류된 값을 포함하는 추가 필드를 파생시키려면 **새 필드**를 선택하십시오. 원래 필드의 값을 새 분류로 덮어쓰려면 **기존 필드**를 선택하십시오. 이는 본질적으로 "채우기" 작업입니다.

일단 모드 및 대체 옵션을 지정했으면 변환 필드를 선택하고 대화 상자의 아래쪽에 있는 동적 옵션을 사용하여 새 분류 값을 지정해야 합니다. 이러한 옵션은 사용자가 위에서 선택한 모드에 따라 다릅니다.

필드 분류. 오른쪽의 필드 선택기 단추를 사용하여 하나(단일 모드) 또는 다중(다중 모드) 범주형 필드를 선택하십시오.

새 필드 이름. 코딩변경된 값을 포함하는 새 명목 필드의 이름을 지정하십시오. 이 옵션은 위에서 **새 필드**가 선택된 경우에 단일 모드에서만 사용 가능합니다. **기존 필드**가 선택된 경우에는 원래 필드 이름이 유지됩니다. 다중 모드에서 작업하는 경우, 각 새 필드에 추가된 확장자를 지정하는 제어가 이 옵션을 대체합니다. 자세한 정보는 167 페이지의 『다중 필드 재분류』의 내용을 참조하십시오.

값 재분류. 이 테이블을 사용하면 이전 변수군 값에서 사용자가 여기서 지정하는 변수군 값으로 맵핑을 지울 수 있습니다.

- **원래 값.** 이 열에는 선택 필드에 대한 기존 값이 나열됩니다.
- **새로운 값.** 이 열을 사용하면 새 범주 값을 입력하거나 드롭 다운 목록에서 하나를 선택할 수 있습니다. 분포 차트의 값을 사용하여 재분류 노드를 자동으로 생성하는 경우, 해당 값이 드롭 다운 목록에 포

함됩니다. 이로 인해 기존 값을 알려진 값 변수군에 신속하게 맵핑할 수 있습니다. 예를 들어, 의료 조직은 네트워크 또는 로케일에 따라 진단을 다르게 그룹화하는 경우가 있습니다. 합병 후에 새 데이터를 재분류해야 하며 심지어 기존 데이터를 일관성 있게 재분류해야 합니다. 기존 데이터에서 수동으로 각 목표 값을 입력하지 않고 여러 마스터 값을 IBM SPSS Modeler로 읽어온 다음 *진단* 필드에 대한 분포 차트를 실행하여 차트에서 직접 이 필드에 대한 재분류(값) 노드를 생성할 수 있습니다. 이 프로세스를 통해 새 값 드롭 다운 목록에서 모든 목표 진단 값을 사용할 수 있습니다.

4. **가져오기**를 클릭하여 위에서 선택한 하나 이상의 필드에 대한 원래 값을 읽을 수 있습니다.
5. **복사**를 클릭하여 아직 맵핑되지 않은 필드에 대한 *새 값* 열로 원래 값을 붙여넣을 수 있습니다. 맵핑되지 않은 원래 값은 드롭 다운 목록에 추가됩니다.
6. **새로 지우기**를 클릭하여 *새로운 값* 열에서 모든 지정 사항을 지우십시오. *참고*: 이 옵션은 드롭 다운 목록의 값은 지우지 않습니다.
7. 각 원래 값에 대한 연속 정수를 자동으로 생성하려면 **자동**을 클릭하십시오. 1.5, 2.5 등의 실수 값이 아니라 정수만 생성될 수 있습니다.

예를 들어, 제품 이름에 대해 자동으로 연속 제품 ID 번호를 생성하거나 대학 강좌에 대한 강좌 번호를 생성할 수 있습니다. 이 기능은 IBM SPSS Statistics의 집합에 대한 자동 코딩변경 변환에 해당됩니다.

지정되지 않은 값 사용. 이 옵션은 새 필드에서 지정되지 않은 값을 채우는 데 유용합니다. **원래 값**을 선택하여 원래 값을 유지하거나 기본값을 지정하십시오.

다중 필드 재분류

한 번에 둘 이상의 필드에 대한 범주 값을 맵핑하려면 모드를 **다중**으로 설정하십시오. 그러면 아래에서 버려지는 재분류 대화 상자 내의 새 설정을 사용할 수 있습니다.

필드 재분류. 오른쪽의 필드 선택기 단추를 사용하여 변환할 필드를 선택하십시오. 필드 선택기를 사용하여 한 번에 모든 필드 또는 유사한 유형의 필드(명목 또는 플래그 등)를 선택할 수 있습니다.

필드 이름 확장. 다중 필드의 코딩을 동시에 변경하는 경우, 개별 필드 이름이 아니라 모든 새 필드에 추가된 공통 확장자를 지정하는 것이 더 효율적입니다. `_recode` 등의 확장자를 지정하고 이 확장자를 원래 필드 이름에 첨부할 것인지 여부를 지정하십시오.

재분류 필드에 대한 저장 공간 및 측정 수준

재분류 노드는 항상 코딩변경 작업에서 명목 필드를 생성합니다. 일부 경우에는 이로 인해 **기존 필드** 재분류 모드를 사용할 때 필드의 측정 수준이 변경될 수 있습니다.

새 필드의 저장 공간(데이터가 *사용되는* 방법이 아니라 *저장되는* 방법)은 다음과 같은 설정 탭 옵션을 기준으로 하여 계산됩니다.

- 지정되지 않은 값은 기본값을 사용하도록 설정된 경우, 기본값 및 새로운 값을 둘 다 검사하고 적절한 저장 공간을 판별하여 저장 유형이 결정됩니다. 예를 들어, 모든 값이 정수로 구문 분석되는 경우에는 필드가 정수 저장 유형을 사용하게 됩니다.

- 지정되지 않은 값은 원래 값을 사용하도록 사용하도록 설정된 경우, 저장 유형은 원래 필드의 저장 공간을 기반으로 합니다. 모든 값이 원래 필드의 저장 공간으로 구문 분석될 수 있는 경우, 해당 저장 공간이 유지됩니다. 그렇지 않으면 기존값 및 새로운 값을 둘 다 포함하여 가장 적절한 저장 유형을 발견하여 저장 공간이 판별됩니다. 예를 들어, 정수 집합 { 1, 2, 3, 4, 5 }를 재분류 4 => 0, 5 => 0을 사용하여 재분류하면 새 정수 집합 { 1, 2, 3, 0 }이 생성되며 재분류 4 => "Over 3", 5 => "Over 3"을 사용하여 재분류하면 { "1", "2", "3", "Over 3" } 문자열 집합이 생성됩니다.

참고: 원래 유형이 인스턴스화되지 않으면 새 유형도 인스턴스화되지 않습니다.

값 익명화 노드

익명화 노드를 사용하면 노드의 모델 다운스트림에 포함되는 데이터에 대해 작업할 때 필드 이름, 필드 값 또는 둘 다를 위장할 수 있습니다. 이러한 방식으로, 권한없는 사용자가 직원 기록 또는 환자의 의료 기록과 같은 기밀 데이터를 볼 수 있는 위험없이 생성된 모델을 예를 들어, 기술 지원에 자유롭게 분배할 수 있습니다.

스트림에서 익명화 노드를 배치하는 위치에 따라 다른 노드를 변경해야 할 수도 있습니다. 예를 들어, 선택 노드에서 업스트림으로 익명화 노드를 삽입하는 경우 선택 노드의 선택 기준이 현재 익명화된 값에 적용되면 이 기준을 변경해야 합니다.

익명화하는 데 사용하는 방법은 다양한 요인에 따라 다릅니다. 필드 이름 및 연속형 측정 수준을 제외한 모든 필드 값의 경우 데이터는 다음과 같은 형식의 문자열로 대체됩니다.

prefix_Sn

여기서 *prefix_* is는 사용자 지정 문자열 또는 기본 문자열 anon_이고 *n*은 0부터 시작하는 정수 값이며 각 고유 값에 대해 증분됩니다(예: anon_S0, anon_S1 등).

숫자 범위가 문자열이 아닌 정수 또는 실수 값을 처리하므로 유형이 연속형인 필드 값을 변환해야 합니다. 이러한 방식으로 범위를 다른 범위로 변환하므로 원래 데이터를 위장해야만 이를 익명화할 수 있습니다. 범위에서 값 *x*의 변환은 다음과 같은 방식으로 수행됩니다.

A*(*x* + B)

여기서,

*A*는 0보다 커야 하는 환산 계수입니다.

*B*는 값에 추가할 변환 오프셋입니다.

예제

환산 계수 *A*가 7로 설정되고 변환 오프셋 *B*가 3으로 설정되는 필드 *AGE*의 경우 *AGE*의 값이 다음으로 변환됩니다.

7*(AGE + 3)

익명화 노드의 옵션 설정

여기에서 다운스트림으로 값을 위장할 필드를 선택할 수 있습니다.

익명화 작업을 수행하기 전에 익명화 노드에서 업스트림으로 데이터 필드를 인스턴스화해야 한다는 점을 참고하십시오. 유형 노드 또는 소스 노드의 유형 탭에서 **값 읽기** 단추를 클릭하여 데이터를 인스턴스화할 수 있습니다.

필드. 현재 데이터 세트의 필드를 나열합니다. 필드 이름이 이미 익명화된 경우 익명화된 이름이 여기에 표시됩니다.

측정. 필드의 측정 수준입니다.

값 익명화. 하나 이상의 필드를 선택하고 이 열을 클릭한 후 **예**를 선택하여 기본 접두부 **anon_**으로 필드 값을 익명화하십시오. 사용자가 직접 접두부를 입력하거나 유형이 *연속형*인 필드 값의 경우 필드 값의 변환이 무작위 또는 사용자 지정 값을 사용할 것인지를 지정할 수 있는 대화 상자를 표시하려면 **지정**을 선택하십시오. *연속형* 및 *비연속형* 필드 유형은 동일한 작업에서 지정할 수 없다는 점을 참고하십시오. 필드의 각 유형에 대해 개별적으로 지정해야 합니다.

현재 필드 보기. 익명화 노드에 활동적으로 연결된 데이터 세트의 필드를 보려면 이를 선택하십시오. 이 옵션은 기본적으로 선택됩니다.

사용하지 않은 필드 설정 보기. 노드에 한 번 연결되었지만 더 이상 연결되어 있지 않은 데이터 세트의 필드를 보려면 이를 선택하십시오. 이 옵션은 한 스트림에서 다른 스트림으로 노드를 복사하거나 노드를 저장하고 다시 로드하는 경우에 유용합니다.

필드 값이 익명화되는 방식 지정

값 대체 대화 상자를 사용하면 익명화된 필드 값의 기본 접두부를 사용하거나 사용자 정의 접두부를 사용할 것인지 선택할 수 있습니다. 이 대화 상자에서 **확인**을 클릭하면 설정 탭에서 값 익명화의 설정이 선택된 필드에 대해 **예**로 변경됩니다.

필드 값 접두부. 익명화된 필드 값의 기본 접두부는 **anon_**입니다. 다른 접두부를 원하면 **사용자 정의**를 선택하고 직접 접두부를 입력하십시오.

값 변환 대화 상자는 유형이 *연속형*인 필드의 경우에만 표시되며 필드 값의 변환이 무작위 또는 사용자 지정 값을 사용하기 위한 것인지를 지정할 수 있습니다.

변량. 변환에 무작위 값을 사용하려면 이 옵션을 선택하십시오. **난수 시드 설정**이 기본적으로 선택되어 있습니다. **시드** 필드에 값을 지정하거나 기본값을 사용하십시오.

고정됨. 변환에 사용자의 값을 지정하려면 이 옵션을 선택하십시오.

- **척도 기준.** 변환에서 필드 값에 곱하는 수입니다. 최소값은 1이고 최대값은 일반적으로 10이지만 오버플로우를 방지하기 위해 이 값을 낮출 수 있습니다.

- **변환 기준.** 변환에서 필드 값에 추가되는 수입니다. 최소값은 0이고 최대값은 일반적으로 1000이지만 오버플로우를 방지하기 위해 이 값을 낮출 수 있습니다.

필드 값 익명화

설정 탭에서 익명화하기 위해 선택된 필드에는 익명화된 값이 있습니다.

- 익명화 노드가 포함된 스트림을 실행할 때
- 값을 미리 볼 때

값을 미리 보려면 익명화된 값 탭에서 **값 익명화** 단추를 클릭하십시오. 그런 다음, 드롭 다운 목록에서 필드 이름을 선택하십시오.

측정 수준이 연속형인 경우 다음과 같이 표시됩니다.

- 원래 범위의 최소값 및 최대값
- 값을 변환하는 데 사용되는 등식

측정 수준이 연속형 외의 수준인 경우 화면에는 해당 필드의 원래 및 익명화된 값이 표시됩니다.

화면이 노란색 배경으로 표시되는 경우 마지막으로 값이 익명화된 이후에 선택된 필드의 설정이 변경되었거나 익명화된 값이 더 이상 올바르지 않은 경우와 같이 익명화 노드의 데이터 업스트림이 변경되었음을 표시합니다. 현재 값 세트가 표시됩니다. **값 익명화** 단추를 다시 클릭하여 현재 설정에 따라 새 값 세트를 생성하십시오.

값 익명화. 선택된 필드의 익명화된 값을 작성하고 테이블에 이 값을 표시합니다. 연속형 유형의 필드에 대해 난수 시드를 사용하는 경우 이 단추를 반복적으로 클릭하면 매번 다른 값 세트가 작성됩니다.

값 지우기. 테이블에서 원래 및 익명화된 값을 지웁니다.

구간화 노드

구간화 노드를 사용하면 하나 이상의 기존 연속(숫자 범위) 필드 값에 기반하여 새 명목 필드를 자동으로 작성할 수 있습니다. 예를 들어, 연속 수입 필드를 동일한 너비의 수입 그룹이 포함된 새 범주형 필드로 변환하거나 평균의 편차로 변환할 수 있습니다. 또는 두 필드 간 원래 연관의 강도를 유지하기 위해 범주형 "수퍼바이저" 필드를 선택할 수 있습니다.

다음과 같은 여러 가지 이유로 구간화가 유용할 수 있습니다.

- **알고리즘 요구사항.** Naive Bayes, 로지스틱 회귀분석과 같은 특정 알고리즘에는 범주형 입력이 필요합니다.
- **성능.** 입력 필드의 고유 값 수가 감소되면 다항 로지스틱과 같은 알고리즘의 성능이 향상될 수 있습니다. 예를 들어, 원래 값을 사용하지 않고 각 구간에 대해 중앙값 또는 평균 값을 사용하십시오.
- **데이터 개인정보 보호정책.** 급여와 같은 민감한 개인 정보는 개인정보를 보호하기 위해 실제 급여 수치가 아닌 범위로 보고될 수 있습니다.

여러 구간화 방법을 사용할 수 있습니다. 새 필드에 대한 구간을 작성한 후에는 절단점을 기반으로 파생 노드를 생성할 수 있습니다.

구간화 노드 사용 시기

구간화 노드를 사용하기 전에 다른 기술이 즉시 사용되는 작업에 보다 적절한지 여부를 고려하십시오.

- 사전정의된 특정 급여 범위와 같은 범주의 절단점을 수동으로 지정하려면 파생 노드를 사용하십시오. 자세한 정보는 157 페이지의 『파생 노드』의 내용을 참조하십시오.
- 기존 세트의 새 범주를 작성하려면 재분류 노드를 사용하십시오. 자세한 정보는 165 페이지의 『재분류 노드』의 내용을 참조하십시오.

결측값 처리

구간화 노드는 다음과 같은 방법으로 결측값을 처리합니다.

- **사용자 지정 공백.** 공백으로 지정된 결측값은 변환 중 포함됩니다. 예를 들어, 유형 노드를 사용하여 공백 값을 표시하기 위해 –99를 지정한 경우 이 값이 구간화 프로세스에 포함됩니다. 구간화 중 공백을 무시하려면 채움 노드를 사용하여 공백 값을 시스템 널값으로 대체해야 합니다.
- **시스템 결측값($null$).** 널값은 구간화 변환 중 무시되며 변환 이후에 널을 유지합니다.

설정 탭은 사용 가능한 기술에 대한 옵션을 제공합니다. 보기 탭은 노드를 통해 이전에 실행된 데이터에 대해 설정된 절단점을 표시합니다.

구간화 노드의 옵션 설정

구간화 노드를 사용하면 다음 기술을 사용하여 구간(범주)을 자동으로 생성할 수 있습니다.

- 고정 너비 구간화
- 분위수(동일 개수 또는 합계)
- 평균 및 표준 편차
- 순위
- 범주형 "수퍼바이저" 필드와 관련하여 최적화됨

대화 상자의 하단 부분은 사용자가 선택하는 구간화 방법에 따라 동적으로 변경됩니다.

구간화 필드. 변환 보류 중인 연속(숫자 범위) 필드가 여기에 표시됩니다. 구간화 노드를 사용하면 여러 필드를 동시에 구간화할 수 있습니다. 오른쪽의 단추를 사용하여 필드를 추가하거나 제거하십시오.

구간화 방법. 새 필드 구간(범주)의 절단점을 판별하는 데 사용되는 방법을 선택하십시오. 이후 주제에서는 각 케이스에서 사용할 수 있는 옵션에 대해 설명합니다.

구간 임계값. 구간 임계값을 계산하는 방법을 지정하십시오.

- **항상 재계산.** 절단점 및 구간 할당은 항상 노드가 실행될 때 재계산됩니다.

- **사용 가능한 경우 구간 값 탭에서 읽기.** 절단점 및 구간 할당은 필요한 경우(예: 새 데이터가 추가된 경우)에만 계산됩니다.

다음 주제에서는 사용 가능한 구간화 방법의 옵션에 대해 설명합니다.

고정 너비 구간

구간화 방법으로 **고정 너비**를 선택하는 경우 새 옵션 세트가 대화 상자에 표시됩니다.

이름 확장자. 생성 필드에 사용할 확장자를 지정하십시오. *_BIN*이 기본 확장자입니다. 확장자가 필드 이름의 처음(**접두부**) 또는 끝(**접미부**)에 추가되는지 여부를 지정할 수도 있습니다. 예를 들어, *income_BIN*이라는 새 필드를 생성할 수 있습니다.

구간 너비. 구간의 "너비"를 계산하는 데 사용되는 값(정수 또는 실수)을 지정하십시오. 예를 들어, 필드 *연령*을 구간화하기 위해 기본값 10을 사용할 수 있습니다. *연령*에 18–65의 범위가 있으므로 생성되는 구간은 다음 표에 표시된 바와 같습니다.

표 23. 범위 18–65인 연령의 구간

구간 1	구간 2	구간 3	구간 4	구간 5	구간 6
>=13부터 <23까지	>=23부터 <33까지	>=33부터 <43까지	>=43부터 <53까지	>=53부터 <63까지	>=63부터 <73까지

구간 간격의 시작은 *스캐닝된 최저 값*에서 구간 너비(지정됨)의 절반을 뺀 값을 사용하여 계산됩니다. 예를 들어, 위에 표시된 구간에서 13이 다음 계산에 따라 간격을 시작하는 데 사용됩니다. 18 *[최저 데이터 값]* – 5 *[0.5 × (구간 너비 10)]* = 13.

구간 수. 새 필드의 고정 너비 구간(범주) 수를 판별하는 데 사용되는 정수를 지정하려면 이 옵션을 사용하십시오.

스트림에서 구간화 노드를 실행한 경우 구간화 노드 대화 상자에서 **미리보기** 탭을 클릭하여 생성되는 구간 임계값을 볼 수 있습니다. 자세한 정보는 176 페이지의 『생성된 구간 미리보기』의 내용을 참조하십시오.

분위수(동일 개수 또는 합계)

분위수 구간화 방법은 각 그룹에 동일한 수의 레코드가 있거나 각 그룹에 있는 값의 합계가 동일하도록 *스캐닝된* 레코드를 백분위수 그룹(또는 사분위수, 십분위수 등)으로 분할하는 데 사용할 수 있는 명목 필드를 작성합니다. 레코드는 지정된 구간 필드의 값에 기반하여 오름차순으로 순위가 지정되므로 선택된 구간 변수의 최저 값이 있는 레코드에 1순위가 지정되고 다음 레코드 세트에 2순위가 지정되는 등과 같습니다. 각 구간의 임계값은 사용된 데이터 및 분위수 지정 방법에 기반하여 자동으로 생성됩니다.

분위수 이름 확장자. 표준 p-분위수를 사용하여 생성되는 필드에 사용되는 확장자를 지정하십시오. 기본 확장자는 *_TILE* 더하기 *N*으로, *N*은 분위수 번호입니다. 확장자가 필드 이름의 처음(**접두부**) 또는 끝(**접미부**)에 추가되는지 여부를 지정할 수도 있습니다. 예를 들어, *income_BIN4*라는 새 필드를 생성할 수 있습니다.

사용자 정의 분위수 확장자. 사용자 정의 분위수 범위에 사용되는 확장자를 지정하십시오. 기본값은 *_TILEN*입니다. 이 케이스의 *N*은 사용자 정의 수로 대체되지 *않는다*는 점을 참고하십시오.

사용 가능한 p-tile은 다음과 같습니다.

- **사분위수.** 각각 케이스의 25%를 포함하는 네 개의 구간을 생성하십시오.
- **5분위수.** 각각 케이스의 20%를 포함하는 다섯 개의 구간을 생성하십시오.
- **십분위수.** 각각 케이스의 10%를 포함하는 10개의 구간을 생성하십시오.
- **20분위수.** 각각 케이스의 5%를 포함하는 20개의 구간을 생성하십시오.
- **백분위수.** 각각 케이스의 1%를 포함하는 100개의 구간을 생성하십시오.
- **사용자 정의 N.** 구간 수를 지정하려면 이를 선택하십시오. 예를 들어, 값 3은 각각 케이스의 33.3%를 포함하는 세 개의 연결된 범주(두 개의 절단점)를 생성합니다.

데이터에 지정된 분위수보다 적은 이산 값이 있는 경우 모든 분위수가 사용되지 않는다는 점을 참고하십시오. 이러한 경우 새 분포에 데이터의 원래 분포가 반영될 가능성이 높습니다.

분위수 지정 방법. 구간에 레코드를 지정하는 데 사용되는 방법을 지정합니다.

- **레코드 개수.** 각 구간에 동일한 수의 레코드를 지정합니다.
- **값 합계.** 각 구간의 값의 합계가 동일하도록 구간에 레코드를 지정합니다. 예를 들어, 영업 성과를 목표로 할 때 이 방법을 사용하면 최고 값 가능성이 맨 위 구간에 있을 때 레코드 당 값에 기반하여 십분위수 그룹에 가능성을 지정할 수 있습니다. 예를 들어, 제약 회사는 작성하는 처방전 수에 기반하여 십분위수 그룹으로 내과 의사의 순위를 지정할 수 있습니다. 각 십분위수에 대략적으로 동일한 수의 스크립트가 포함되지만 이러한 스크립트에 기여하는 개인 수는 동일하지 않으며 개인은 십분위수 10에 집중된 대부분의 스크립트를 작성합니다. 이러한 방법에서는 모든 값이 0보다 크다고 가정하고 이 경우에 해당하지 않는 경우 예상하지 않은 결과가 나타날 수 있다는 점을 참고하십시오.

경계값. 절단점의 한 쪽에 있는 값이 동일한 경우 경계값 조건 결과입니다. 예를 들어, 십분위수를 지정하며 레코드의 10% 이상에 구간 필드에 대해 동일한 값이 있는 경우 임계값을 한 방향 또는 다른 방향으로 강제 실행하지 않고 이러한 모든 레코드를 동일한 구간에 넣을 수 없습니다. 경계값이 위로 다음 구간으로 이동되거나 현재 구간에서 유지될 수 있지만 동일한 값의 모든 레코드가 동일한 구간에 속하도록 분석해야 합니다(이로 인해 일부 구간에 예상보다 많은 레코드가 있는 경우에도). 따라서 이후 구간의 임계값도 조정할 수 있으므로 경계값을 분석하는 데 사용되는 방법에 기반하여 동일한 번호 세트에 대해 다르게 값이 지정됩니다.

- **다음에 추가.** 위로 다음 구간으로 경계값을 이동하려면 이를 선택하십시오.
- **현재에서 유지.** 현재(낮은) 구간에서 경계값을 유지합니다. 이 방법으로 인해 작성 중인 총 구간 수가 줄어들 수 있습니다.
- **무작위 지정.** 한 구간에 경계값을 무작위로 할당하려면 이를 선택하십시오. 각 구간에서 레코드 수를 동일한 양으로 유지하려고 합니다.

예: 레코드 수별 분위수 지정

다음 표는 레코드 수별로 분위수를 지정할 때 간단해진 필드 값이 분위수로 순위가 지정되는 방식을 나타냅니다. 결과는 선택된 경계값 옵션에 따라 다를 수 있다는 점을 참고하십시오.

표 24. 레코드 수별 분위수 지정 예.

값	다음에 추가	현재에서 유지
10	1	1
13	2	1
15	3	2
15	3	2
20	4	3

구간 당 항목 수는 다음과 같이 계산됩니다.

총 값 수 / 분위수

위의 간단한 예에서 구간 당 원하는 항목 수는 1.25(5개의 값 / 4 사분위수)입니다. 값 13(값 번호 2)은 1.25 원하는 개수 임계값을 스트래들하므로 선택된 경계값 옵션에 따라 다르게 처리됩니다. **다음에 추가** 모드에서는 구간 2에 추가됩니다. **현재에서 유지** 모드에서는 구간 1에 계속 남아 있어서 구간 4에 대한 값의 범위를 기존 데이터 값의 범위 외부에 넣습니다. 따라서 세 개의 구간만 작성되며 각 구간의 임계값이 적절하게 조정됩니다 (다음 표 참조).

표 25. 구간화 예 결과.

구간	하한	상한
1	>=10	<15
2	>=15	<20
3	>=20	<=20

참고: 분위수별로 구간화하는 속도는 병렬 처리를 사용할 경우 빨라질 수 있습니다.

케이스 순위 지정

순위를 구간화 방법으로 선택하면 새 옵션 세트가 대화 상자에 표시됩니다.

순위화는 아래에 지정된 옵션에 따라 숫자 필드의 순위, 분수 순위, 백분위수 값이 포함된 새 필드를 작성합니다.

순위 순서. 오름차순(최저값이 1로 표시됨) 또는 **내림차순**(최고값이 1로 표시됨)을 선택하십시오.

순위. 위에 지정된 바와 같이 오름차순 또는 내림차순으로 케이스의 순위를 지정하려면 이를 선택하십시오. 새 필드에 있는 값의 범위는 1–*N*이며 여기서 *N*은 원래 필드에 있는 이산 값의 수입니다. 경계값에는 해당 순위의 평균이 제공됩니다.

분수 순위. 새 필드의 값이 순위를 비결측 케이스의 가중치 합계로 나눈 값인 케이스의 순위를 지정하려면 이를 선택하십시오. 분수순위는 0 - 1의 범위에 들어갑니다.

백분율 분수 순위. 각 순위를 유효한 값을 갖는 레코드 수로 나누고 100을 곱합니다. 퍼센트 분수순위는 1 - 100의 범위에 들어갑니다.

확장자. 모든 순위 옵션에 대해 사용자 정의 확장자를 작성하고 필드 이름의 처음(**접두부**) 또는 끝(**접미부**)에 추가되는지 여부를 지정할 수 있습니다. 예를 들어, *income_P_RANK*라는 새 필드를 생성할 수 있습니다.

평균/표준 편차

구간화 방법으로 **평균/표준 편차**를 선택하는 경우 새 옵션 세트가 대화 상자에 표시됩니다.

이 방법은 지정된 필드의 분포에 대한 평균 및 표준 편차의 값에 기반하여 연결된 범주가 있는 하나 이상의 새 필드를 생성합니다. 아래에서 사용할 편차 수를 선택하십시오.

이름 확장자. 생성 필드에 사용할 확장자를 지정하십시오. *_SDBIN*이 기본 확장자입니다. 확장자가 필드 이름의 처음(**접두부**) 또는 끝(**접미부**)에 추가되는지 여부를 지정할 수도 있습니다. 예를 들어, *income_SDBIN*이라는 새 필드를 생성할 수 있습니다.

- **+/– 1 표준 편차.** 세 개의 구간을 생성하려면 이를 선택하십시오.
- **+/– 2 표준 편차.** 다섯 개의 구간을 생성하려면 이를 선택하십시오.
- **+/– 3 표준 편차.** 일곱 개의 구간을 생성하려면 이를 선택하십시오.

예를 들어, +/–1 표준 편차를 선택하면 세 개의 구간이 다음 표와 같이 계산되고 표시됩니다.

표 26. 표준 편차 구간 예.

구간 1	구간 2	구간 3
x < (평균 – 표준 편차)	(평균 – 표준 편차) <= x <= (평균 + 표준 편차)	x > (평균 + 표준 편차)

정규 분포에서 케이스의 68%는 평균의 표준 편차 1에 속하고 95%는 표준 편차 2에 속하고 99%는 표준 편차 3에 속합니다. 그러나 표준 편차에 기반하여 연결된 범주를 작성하면 실제 데이터 범위 외부 및 가능한 데이터 값의 범위 외부(예: 음수 급여 범위)에도 일부 구간이 정의될 수 있습니다.

최적 구간화

구간화할 필드가 다른 범주형 필드와 강력하게 연관되어 있는 경우 두 필드 간 원래 연관의 강도를 유지하기 위한 방식으로 구간을 작성하기 위해 범주형 필드를 "수퍼바이저" 필드로 선택할 수 있습니다.

예를 들어, 군집 분석을 사용하여 주택 자금 융자의 연체율에 기반하여 상태를 그룹화했다고 가정하십시오(최고 비율이 첫 번째 군집에 있음). 이 경우 구간 필드로 *기일 경과 후 퍼센트* 및 *압류 퍼센트*를 선택하고 수퍼바이저 필드로 모델에서 생성한 소속군집 필드를 선택할 수 있습니다.

이름 확장자 생성 필드에 사용할 확장자와 필드의 처음(**접두부**) 또는 끝(**접미부**)에 이를 추가할 것인지 여부를 지정하십시오. 예를 들어, pastdue_OPTIMAL이라는 새 필드와 inforeclosure_OPTIMAL이라는 다른 필드를 생성할 수 있습니다.

수퍼바이저 필드 구간을 작성하는 데 사용되는 범주형 필드입니다.

큰 데이터 세트로 성능을 향상시키는 사전 구간화 필드 사전 처리를 사용하여 최적 구간화를 간소화해야 하는지 여부를 표시합니다. 이는 간단한 무감독 구간화 방법을 사용하여 매우 많은 구간으로 척도 값을 그룹화하

고 각 구간의 값을 평균으로 표시하고 감독 구간화를 계록하기 전에 케이스 가중치를 적절하게 조정합니다. 실제로 이 방법은 속도에 대한 정밀도를 교환하며 대형 데이터 세트의 경우에 권장됩니다. 이 옵션이 사용되는 경우 사전 처리한 후 변수가 사용하게 되는 최대 구간 수도 지정할 수 있습니다.

큰 이웃 항목이 있는 상대적으로 작은 케이스 개수가 있는 구간 병합. 사용되는 경우 구간의 크기(케이스 수)와 인접 구간의 크기 비율이 지정된 임계값보다 작은 경우 구간이 병합됨을 표시합니다. 임계값이 크면 추가 병합이 발생할 수 있다는 점을 참고하십시오.

절단점 설정

절단점 설정 대화 상자를 사용하면 최적 구간화 알고리즘의 고급 옵션을 지정할 수 있습니다. 이러한 옵션은 대상 필드를 사용하여 구간을 계산하는 방법을 알고리즘에 지시합니다.

구간 끝점. 낮거나 높은 끝점을 포함하거나(lower <= *x*) 제외하는지(lower < *x*) 여부를 지정할 수 있습니다.

첫 번째 및 마지막 구간. 첫 번째 및 마지막 구간 모두의 경우 구간이 최저 또는 최고 데이터 점으로 무제한(양수 또는 음수 무한대로 확장) 또는 제한되어야 하는지 여부를 지정할 수 있습니다.

생성된 구간 미리보기

구간화 노드의 구간 값 탭을 사용하면 생성된 구간의 임계값을 볼 수 있습니다. 생성 메뉴를 사용하면 하나의 데이터 세트에서 다른 데이터 세트로 이러한 임계값을 적용하는 데 사용할 수 있는 파생 노드를 생성할 수도 있습니다.

구간화된 필드. 드롭 다운 목록을 사용하여 보고자 하는 필드를 선택할 수 있습니다. 표시되는 필드 이름은 명확히 하기 위해 원래 필드 이름을 사용합니다.

분위수. 드롭 다운 목록을 사용하여 보고자 하는 분위수(예: 10 또는 100)를 선택할 수 있습니다. 이 옵션은 분위수 방법(동일한 개수 또는 합계)을 사용하여 구간이 생성된 경우에만 사용할 수 있습니다.

구간 임계값. 각 구간에 속하는 레코드 수와 함께 생성된 각 구간에 대한 임계값이 여기에 표시됩니다. 최적 구간화 방법의 경우에만 각 구간의 레코드 수가 전체의 백분율로 표시됩니다. 순위 구간화 방법이 사용되는 경우 임계값을 적용할 수 없다는 점을 참고하십시오.

값 읽기. 데이터 세트에서 구간화된 값을 읽습니다. 새 데이터가 스트림을 통해 실행되는 경우 임계값도 겹쳐 쓴다는 점을 참고하십시오.

파생 노드 생성

생성 메뉴를 사용하면 현재 임계값에 기반하여 파생 노드를 작성할 수 있습니다. 설정된 구간 임계값을 하나의 데이터 세트에서 다른 데이터 세트로 적용하는 데 유용합니다. 또한 이러한 분할 지점이 알려진 경우 대형 데이터 세트에 대해 작업할 때 파생 작업이 구간화 작업보다 효율적입니다(속도가 빠름을 의미함).

RFM 분석 노드

RFM(Recency, Frequency, Monetary) 분석 노드를 사용하면 얼마나 최근에 사용자로부터 구매했는지(최근성), 얼마나 자주 구매했는지(빈도) 및 모든 트랜잭션에서 얼마나 소비했는지(구매총액)를 조사하여 최고의 고객이 될 수 있는 고객을 정량적으로 판별할 수 있습니다.

RFM 분석의 추론은 제품 또는 서비스를 한 번 구매한 고객이 다시 구매하는 경향이 있다는 것입니다. 분류된 고객 데이터는 사용자가 요구하는 대로 조정된 구간화 기준이 있는 여러 개의 구간으로 분할됩니다. 각 구간에서 고객에게 점수가 지정되고 이러한 점수가 결합되어 전체 RFM 점수를 제공합니다. 이 점수는 각 RFM 모수에 대해 작성된 구간에 대한 고객의 소속을 나타냅니다. 이 구간화된 데이터는 사용자의 요구를 충분히 충족할 수 있습니다. 예를 들어, 가장 빈번한 높은 값의 고객을 식별하거나 추가 모델링 및 분석에 필요한 스트림에서 전달될 수 있습니다.

단, RFM 점수를 분석하고 순위를 매기는 기능이 유용한 도구이기는 하지만 이를 사용할 때 특정 요인을 인식하고 있어야 합니다. 가장 높은 순위의 고객을 대상으로 하려는 유혹이 있을 수 있으나 해당 고객에게 지나치게 구매를 요구하는 경우 분노를 표출하거나 실제로는 반복 비즈니스가 실패할 수 있습니다. 또한 낮은 점수의 고객을 무시하는 대신 더 나은 고객으로 만들 수 있도록 노력해야 합니다. 반대로 높은 점수 단독으로는 시장에 따라 반드시 판매 전망이 높음을 나타내지 않을 수 있습니다. 예를 들어, 최근 범주에서 매우 최근에 구매했음을 의미하는 구간 5에 속하는 고객은 자동차 또는 텔레비전과 같은 비싸고 오래 사용하는 제품을 판매하려는 사람에게는 최선의 목표 고객이 아닐 수 있습니다.

참고: 데이터가 저장되는 방법에 따라 데이터를 유용한 형식으로 변환하기 위해 RFM 통합 노드를 RFM 분석 노드에 앞서 사용해야 합니다. 예를 들어, 입력 데이터는 고객당 한 행인 고객 형식이어야 합니다. 고객의 데이터가 트랜잭션 양식인 경우, RFM 통합 노드 업스트림을 사용하여 최근, 빈도 및 구매총액 필드를 파생시키십시오. 자세한 정보는 84 페이지의 『RFM 통합 노드』의 내용을 참조하십시오.

IBM SPSS Modeler의 RFM 통합 및 RFM 분석 노드는 독립적 구간화를 사용하기 위해 설정됩니다. 즉, 해당 값 또는 다른 두 측도에 관계없이 RFM(Recency, Frequency, Monetary) 값의 각 측도에 대한 데이터를 순위화하고 구간화합니다.

RFM 분석 노드 설정

최근. 필드 선택기(텍스트 상자의 오른쪽에 있는 단추)를 사용하여 최근 필드를 선택하십시오. 이것은 날짜, 시간소인 또는 단순 숫자일 수 있습니다. 날짜 또는 시간소인이 가장 최근 트랜잭션의 날짜를 나타내는 경우 최고값은 가장 최근으로 간주됩니다. 숫자가 지정되면 가장 최근 트랜잭션 이후의 경과 시간을 나타내며 최저값이 가장 최근으로 간주됩니다.

참고: RFM 분석 노드에 앞서 RFM 통합 노드의 스트림이 사용되면 RFM 통합 노드에 의해 생성된 최근, 빈도 및 구매총액 필드가 RFM 분석 노드에서 입력으로 선택되어야 합니다.

빈도. 필드 선택기를 사용하여 사용할 빈도 필드를 선택하십시오.

구매총액. 필드 선택기를 사용하여 사용할 구매총액 필드를 선택하십시오.

구간 수. 세 개의 출력 유형 각각에 대해 작성할 구간 수를 선택하십시오. 기본값은 5입니다.

참고: 최소 구간 수는 2이며 최대 구간 수는 9입니다.

가중치. 기본적으로 점수를 계산할 때 최고 중요도는 최근 데이터에 지정되며 빈도가 그 다음, 구매총액에 마지막 중요도가 지정됩니다. 필요한 경우 이중 하나 또는 여러 개에 대해 가중치 영향을 수정하여 최고 중요도 지정을 변경할 수 있습니다.

RFM 점수는 (최근성 점수 x 최근성 가중치) + (구매빈도 점수 x 빈도 가중치) + (구매총액 점수 x 구매총액 가중치)로 계산됩니다.

등순위. 동일한(등순위) 점수를 구간화하는 방법을 지정합니다. 옵션은 다음과 같습니다.

- **다음에 추가.** 등순위 값을 다음 구간으로 이동하도록 선택합니다.
- **현재구간에 유지.** 등순위 값을 현재(더 낮은) 구간에 유지합니다. 이 방법을 사용하면 더 적은 수의 총 구간이 작성될 수 있습니다. 이 옵션이 기본값입니다.

구간 임계값. 노드가 실행될 때 RFM 점수 및 구간 할당이 항상 다시 계산되는지 또는 필요한 경우(새 데이터가 추가되는 경우 등)에만 계산되는지 지정하십시오. **사용 가능한 경우 구간 값 탭에서 읽기**를 선택한 경우에는 구간 값 탭에서 다른 구간에 대한 상한 및 하한 절단점을 편집할 수 있습니다.

실행될 때 RFM 분석 노드가 원래 최근, 빈도 및 구매총액 필드를 구간화하고 다음 새 필드를 데이터 세트에 추가합니다.

- 최근성 점수. 최근성에 대한 순위(구간 값)
- 구매빈도 점수. 빈도에 대한 순위(구간 값)
- 구매총액 점수. 구매총액에 대한 순위(구간 값)
- RFM 점수. 최근성, 빈도 및 구매총액 점수의 가중된 합계입니다.

이상값을 마지막 구간에 추가. 이 확인 상자를 선택하면 하한 구간 아래에 위치한 레코드는 하한 구간에 추가되고 최고 구간 위의 레코드는 최고 구간에 추가되며 널 값이 지정됩니다. 이 선택란은 **사용 가능한 경우 구간 값 탭에서 읽기**를 선택한 경우에만 사용 가능합니다.

RFM 분석 노드 구간화

구간 값 탭을 사용하면 생성된 구간에 대한 임계값을 볼 수 있으며 특정 케이스에 이를 수정할 수 있습니다.

참고: 설정 탭에서 **사용 가능한 경우 구간 값 탭에서 읽기**를 선택한 경우에만 이 탭에서 값을 수정할 수 있습니다.

구간화된 필드. 드롭 다운 목록을 사용하여 구간으로 나눌 필드를 선택하십시오. 사용 가능한 값은 설정 탭에서 선택된 값입니다.

구간 값 테이블. 생성된 각 구간에 대한 임계값이 여기에 표시됩니다. 설정 탭에서 **사용 가능한 경우 구간 값 탭에서 읽기**를 선택한 경우에는 관련 셀을 두 번 클릭하여 각 구간에 대한 상한 및 하한 절단점을 수정할 수 있습니다.

값 읽기. 데이터 세트에서 구간화된 값을 읽고 구간 값 테이블을 채웁니다. 설정 탭에서 **항상 다시 계산**을 선택한 경우에는 새 데이터가 스트림을 통해 실행될 때 구간 임계값을 덮어씁니다.

앙상블 노드

앙상블 노드는 둘 이상의 모델 너깃을 결합하여 개별 모델에서 얻을 수 있는 것보다 정확한 예측을 얻습니다. 여러 모델의 예측을 결합하면 개별 모델의 제한을 피할 수 있어 전반적인 정확도가 향상됩니다. 이 방식으로 결합된 모델은 일반적으로 개별 모델 이상의 성능을 보여줄 수 있습니다.

이러한 노드의 결합은 자동 분류자, 자동 숫자 및 자동 군집 자동화된 모델링 노드에서 자동으로 발생합니다.

앙상블 노드를 사용한 후에는 분석 노드 또는 평가 노드를 사용하여 결합된 결과의 정확도를 각각의 입력 모델과 비교할 수 있습니다. 이를 수행하려면 앙상블 노드의 설정 탭에서 **앙상블 모델이 생성한 필드 필터링** 옵션이 선택되지 않았는지 확인하십시오.

출력 필드

각각의 앙상블 노드는 결합된 스코어가 포함된 필드를 생성합니다. 이름은 지정된 목표 필드를 기반으로 하며 필드 측정 수준(플래그, 명목(세트) 또는 연속형(범위))에 따라 각각 *$XF_*, *$XS_* 또는 *$XR_* 접두부가 지정됩니다. 예를 들어, 목표가 *response*라는 플래그 필드인 경우 출력 필드는 *$XF_response*입니다.

신뢰도 또는 성향 필드. 플래그 및 명목 필드의 경우 다음 표에 자세히 설명된 대로 앙상블 방법을 기반으로 추가적인 신뢰도 또는 성향 필드가 작성됩니다.

표 27. 앙상블 방법 필드 작성.

앙상블 방법	필드 이름
투표 신뢰 가중 투표 원시 성향 가중 투표 수정된 성향 가중 투표 가장 높은 신뢰도 승리	*$XFC_<field>*
평균 원시 성향	*$XFRP_<field>*
평균 수정된 원시 성향	*$XFAP_<field>*

앙상블 노드 설정

앙상블의 목표 필드. 둘 이상의 업스트림 모델에서 목표로 사용하는 단일 필드를 선택하십시오. 업스트림 모델은 플래그, 명목 또는 연속형 대상을 사용할 수 있지만 스코어를 결합하려면 둘 이상의 모델이 동일한 목표를 공유해야 합니다.

앙상블 모델이 생성한 필드 필터링. 앙상블 노드에 반영되는 개별 모델이 생성한 모든 추가 필드의 출력에서 제거합니다. 모든 입력 모델에서 결합된 스코어에만 관심이 있는 경우 이 확인 상자를 선택합니다. 예를 들어, 분석 노드 또는 평가 노드를 사용하여 각 개별 입력 모델의 정확도와 결합된 스코어의 정확도를 비교하려는 경우에는 이 옵션을 선택 취소해야 합니다.

사용 가능한 설정은 목표로 선택된 필드의 측정 수준에 따라 다릅니다.

연속형 대상

연속형 대상의 경우 스코어는 평균값을 구합니다. 이는 스코어 결합을 위해 사용할 수 있는 유일한 방법입니다.

스코어 또는 추정값의 평균을 구하는 경우 앙상블 노드는 표준 오차 계산을 사용하여 측정되거나 추정된 값과 참 값 사이의 차이를 계산하고 해당 추정값이 얼마나 일치하는지 표시합니다. 표준 오차 계산은 새 모델에 대해 기본적으로 생성되지만 기존 모델에 대해 선택란을 선택 취소할 수 있습니다(예: 재생성해야 하는 경우).

범주형 대상

범주형 대상의 경우 각각의 가능한 예측값이 선택되는 횟수를 기록하고 총계가 가장 큰 값을 선택하여 작동하는 **투표**를 포함한 다수의 방법이 지원됩니다. 예를 들어, 5개 모델 중 3개가 *yes*를 예측하고 다른 두 모델은 *no*를 예측하는 경우 *yes*가 3 대 2 투표로 이깁니다. 또는 각 예측에 대한 신뢰도 및 성향 값을 기반으로 투표에 **가중치**가 부여될 수 있습니다. 그런 다음 가중치가 합계되고 총계가 가장 높은 값이 다시 선택됩니다. 최종 예측에 대한 신뢰도는 앙상블에 포함된 모델의 수로 나눈 승리 값에 대한 가중치의 합계입니다.

모든 범주형 필드. 플래그 필드와 명목 필드 모두에 대해 다음과 같은 방법이 지원됩니다.

- 투표
- 신뢰 가중 투표
- 가장 높은 신뢰도 승리

플래그 필드만. 플래그 필드만의 경우 성향을 기반으로 한 다수의 방법도 사용할 수 있습니다.

- 원시 성향 가중 투표
- 수정된 성향 가중 투표
- 평균 원시 성향
- 평균 수정된 성향

투표 동률. 투표 방법에 대해 동률을 해결하는 방법을 지정할 수 있습니다.

- **무작위 선택.** 동률 값 중 하나가 무작위로 선택됩니다.
- **가장 높은 신뢰도.** 가장 높은 신뢰도로 예측된 동률 값이 승리합니다. 이것이 반드시 모든 예측값의 가장 높은 신뢰도와 동일하지는 않습니다.
- **원시 또는 수정된 성향(플래그 필드만).** 가장 큰 절대 성향으로 예측된 동률 값입니다. 여기서 절대 성향은 다음과 같이 계산됩니다.

```
abs(0.5 - propensity) *
2
```

수정된 성향의 경우 다음과 같이 계산됩니다.

```
abs(0.5 - adjusted propensity) * 2
```

파티션 노드

파티션 노드는 파티션 필드를 생성하는데 사용되며 이 필드는 모델 작성의 학습, 검정, 검증 단계를 위한 별개의 서브세트 또는 표본으로 데이터를 분할합니다. 일표본을 사용하여 모델을 생성하고, 별도의 표본으로 검정하면, 현재 데이터와 유사한 더 큰 데이터 세트에 대해 모델을 일반화할 때 효율성을 효과적으로 표시할 수 있습니다.

파티션 노드는 **파티션**으로 설정된 역할이 있는 명목 필드를 생성합니다. 또는 적절한 필드가 사용자의 데이터에 이미 있는 경우에는 유형 노드를 사용하여 파티션으로 지정될 수 있습니다. 이 케이스에서 별도의 파티션 노드가 필요하지 않습니다. 두 개 또는 세 개의 값이 있는 인스턴스화된 모든 명목 필드는 파티션으로 사용될 수 있으나 플래그 필드는 사용될 수 없습니다. 자세한 정보는 150 페이지의 『필드 역할 설정』의 내용을 참조하십시오.

다중 파티션 필드가 스트림에서 정의될 수 있으나 이 경우, 분할을 사용하는 각 모델링 노드의 필드 탭에서 단일 파티션 필드를 선택해야 합니다. (하나의 파티션만 존재하는 경우 파티션이 사용될 때마다 자동으로 사용됩니다.)

파티션 사용. 분석에서 파티션을 사용하려면 적절한 모델 작성 또는 분석 노드의 모형 옵션 탭에서 파티션을 사용할 수 있어야 합니다. 이 옵션을 선택 취소하면 필드를 제거하지 않은 채로 분할을 사용할 수 없게 됩니다.

날짜 범위 또는 위치와 같이 몇 가지 기타 기준을 기반으로 하여 파티션 필드를 작성하려면 파생 노드를 사용할 수 있습니다. 자세한 정보는 157 페이지의 『파생 노드』의 내용을 참조하십시오.

예. 이전 마케팅 캠페인에 긍정적으로 반응한 최근 고객을 식별하기 위해 RFM 스트림을 작성하는 경우, 판매 회사의 마케팅 부서가 파티션 노드를 사용하여 데이터를 학습 및 검정 파티션으로 분할합니다.

파티션 노드 옵션

파티션 필드. 노드에 의해 생성된 필드의 이름을 지정합니다.

파티션. 데이터를 두 표본(학습 및 검정) 또는 세 표본(학습, 검정 및 검증)으로 파티셔닝할 수 있습니다.

- **학습 및 검정.** 데이터를 두 개의 표본으로 파티셔닝하여 한 표본으로 모델을 학습하고 다른 표본으로 모델을 검정할 수 있습니다.
- **학습, 검정 및 검증.** 데이터를 세 개의 표본으로 파티셔닝하여 한 표본으로 모델을 학습하고 두 번째 표본으로 모델을 검정하고 세분화하며 세 번째 표본으로 결과를 검증할 수 있습니다. 따라서 각 파티션의 *크기*는 축소되나 매우 큰 데이터 세트를 사용하여 작업할 때에 비해 더 적합할 수 있습니다.

파티션 *크기*. 각 파티션의 상대적인 *크기*를 지정합니다. 파티션 *크기*의 합계가 100% 미만이면 파티션에 포함되지 않은 레코드가 삭제됩니다. 예를 들어, 사용자에게 천만 레코드가 있으며 파티션 *크기*를 학습 5%, 검정 10%로 지정한 경우 노드를 실행한 후에 나머지는 삭제되고 대략 500,000개의 학습 레코드 및 백만 개의 검정 레코드가 있어야 합니다.

값. 데이터에서 각 파티션 표본을 나타내기 위해 사용하는 값을 지정합니다.

- **시스템 정의 값 사용("1," "2," 및 "3").** 각 파티션을 나타내기 위해 정수를 사용합니다. 예를 들어, 학습 표본에 포함되는 모든 레코드는 파티션 필드의 값으로 1을 사용합니다. 그러면 데이터를 로케일 사이에서 이동할 수 있으며 파티션 필드가 다른 위치에서 다시 인스턴스화되는 경우(데이터베이스에서 다시 데이터를 읽는 경우 등), 정렬 순서가 유지됩니다. 따라서 1이 여전히 학습 파티션을 나타냅니다. 그러나 값에 몇 가지 해석이 필요합니다.
- **시스템 정의 값에 레이블 추가.** 정수와 레이블을 조합합니다. 예를 들어, 학습 파티션 레코드는 `1_Training` 값을 사용합니다. 그러면 정렬 순서를 유지하면서도 데이터를 보는 사용자가 값을 식별할 수 있습니다. 그러나 값은 지정된 로케일에 한정됩니다.
- **값으로 레이블 사용.** 정수 없이 레이블을 사용합니다. 예를 들어, **Training**과 같습니다. 이 경우, 레이블을 편집하여 값을 지정할 수 있습니다. 그러나 이로 인해 데이터가 로케일에 한정되며 파티션 열을 다시 인스턴스화하면 값이 기본 정렬 순서로 배열되어 "시맨틱" 순서와 일치하지 않을 수 있습니다.

시드. 반복 가능한 파티션 할당이 선택된 경우에만 사용 가능합니다. 임의 퍼센트를 기반으로 사용하여 표본추출 레코드 또는 파티셔닝을 수행하는 경우에 이 옵션을 사용하면 다른 세션에서 동일한 결과를 복제할 수 있습니다. 난수 생성기에서 사용하는 시작값을 지정하면 노드를 실행할 때마다 동일한 레코드를 지정하도록 보장할 수 있습니다. 원하는 시드 값을 입력하거나 **생성** 단추를 클릭하여 자동으로 임의 값을 생성하십시오. 이 옵션을 선택하지 않으면 노드를 실행할 때마다 다른 표본이 생성됩니다.

참고: 데이터베이스에서 읽은 레코드에서 **시드** 옵션을 사용하는 경우 노드를 실행할 때마다 동일한 결과를 보장하려면 표본 추출 전에 정렬 노드가 필요할 수도 있습니다. 난수 시드는 레코드 순서에 의존하여 관계형 데이터베이스에서는 동일하게 보장되지 않기 때문입니다. 자세한 정보는 86 페이지의 『정렬 노드』의 내용을 참조하십시오.

고유 필드를 사용하여 파티션 지정. 반복 가능한 파티션 할당이 선택된 경우에만 사용 가능합니다. (계층 1 데이터베이스 전용) SQL 푸시백을 사용하여 레코드를 파티션에 지정하려면 이 선택란을 선택하십시오. 레코드가 무작위이지만 반복 가능한 방식으로 지정되도록 하려면 드롭 다운 목록에서 고유 값이 있는 필드(ID 필드 등)를 선택하십시오.

데이터베이스 계층에 대서는 데이터베이스 소스 노드에 대한 설명을 참조하십시오. 자세한 정보는 18 페이지의 『데이터베이스 소스 노드』의 내용을 참조하십시오.

선택 노드 생성

파티션 노드에서 일반 메뉴를 사용하여 각 파티션에 대한 선택 노드를 자동으로 생성할 수 있습니다. 예를 들어, 학습 파티션에서 모든 레코드를 선택하여 이 파티션만 사용하여 추가 평가 또는 분석을 작성할 수 있습니다.

플래그로 설정 노드

플래그로 설정 노드는 하나 이상의 명목 필드에 대해 정의된 범주형 값을 바탕으로 다중 플래그 필드를 파생시키는 데 사용됩니다. 예를 들어, 사용자의 데이터 세트에 값이 높음, 보통 및 낮음인 *BP*(혈압) 명목 필드가 포함될 수 있습니다. 데이터를 더 쉽게 조작하기 위해 환자의 혈압이 높은지 여부를 표시하는 높은 혈압에 대한 플래그를 작성할 수 있습니다.

플래그로 설정 노드에 대한 옵션 설정

세트 필드. 측정 수준이 명목(변수군)인 모든 데이터 필드를 나열합니다. 목록에서 하나를 선택하여 변수군 내의 값을 표시하십시오. 해당 값에서 선택하여 플래그 필드를 작성할 수 있습니다. 사용 가능한 명목 필드 및 해당 값을 보려면 데이터가 업스트림 소스 또는 유형 노드를 사용하여 완전히 인스턴스화되어야 합니다. 자세한 정보는 138 페이지의 『유형 노드』의 내용을 참조하십시오.

필드 이름 확장. 새 플래그 필드에 대한 접미문자 또는 접두문자로 추가될 확장자를 지정하기 위한 제어를 선택하십시오. 기본적으로 새 필드 이름은 원래 필드 이름과 필드 값을 레이블에 조합하여 자동으로 생성됩니다. 예를 들어, *Fieldname_fieldvalue*와 같습니다.

사용가능 변수군 값. 위에서 선택된 변수군 내의 값이 여기에 표시됩니다. 플래그를 생성할 값을 하나 이상 선택하십시오. 예를 들어, *blood_pressure* 필드의 값이 높음, *중간* 및 낮음이면 높음을 선택하고 이를 오른쪽 목록에 추가할 수 있습니다. 그러면 높은 혈압을 표시하는 값이 있는 레코드에 대한 플래그가 있는 필드가 생성됩니다.

플래그 필드 생성. 새로 생성된 플래그 필드가 여기에 나열됩니다. 필드 이름 확장자 제어를 사용하여 새 필드의 이름을 지정하기 위한 옵션을 지정할 수 있습니다.

참 값. 플래그를 설정할 때 노드가 사용하는 참 값을 지정합니다. 기본적으로 이 값은 **T**입니다.

거짓 값. 플래그를 설정할 때 노드가 사용하는 거짓 값을 지정합니다. 기본적으로 이 값은 **F**입니다.

통합 키. 아래에 지정된 키 필드를 기준으로 하여 레코드를 함께 그룹화하려면 선택하십시오. **통합 키**를 선택하면 *임의의* 레코드가 참으로 설정된 경우에 그룹 내의 모든 플래그 필드가 "켜집니다". 레코드를 통합하는 데 사용할 키 필드를 지정하려면 필드 선택기를 사용하십시오.

구조변환 노드

명목형 또는 플래그 필드의 값을 기준으로 하여 다중 필드를 생성하는 데 구조변환 노드가 사용될 수 있습니다. 새로 생성된 필드는 또 다른 필드 또는 숫자 플래그(0 및 1)의 값을 포함할 수 있습니다. 이 노드의 기능은 플래그로 설정 노드의 기능과 유사합니다. 단, 더 많은 융통성을 제공합니다. 이 노드를 사용하면 또 다른 필드의 값을 사용하여 숫자 플래그를 포함하여 모든 유형의 필드를 작성할 수 있습니다. 그런 다음 기타 노드 다운스트림을 사용하여 통합 또는 기타 조작을 수행할 수 있습니다. 플래그로 설정 노드를 사용하면 한 단계로 필드를 통합할 수 있으며 플래그 필드를 생성할 때 편리합니다.

예를 들어, 다음 데이터 세트에는 값이 *Savings* 및 *Draft*인 *Account*라는 명목 필드가 포함됩니다. 각 계정에 대해 시작 잔액 및 현재 잔액이 기록되고 일부 고객은 각 유형의 계정이 여러 개 있을 수 있습니다. 각 고객이 특정 계정 유형을 갖고 있는지 파악하고 해당되는 경우 각 계정 유형의 금액이 얼마인지 알고자 하는 경우를 가정해 보십시오. 구조변환 노드를 사용하면 각 *Account* 값에 대한 필드를 생성하고 값으로 *Current_Balance*를 선택할 수 있습니다. 각 새 필드는 지정된 레코드에 대한 현재 잔액으로 채워집니다.

표 28. 구조변환 전의 표본 데이터.

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

표 29. 구조변환 후의 표본 데이터.

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	$null$

통합 노드와 함께 구조변환 노드 사용

구조변환 노드와 통합 노드를 한 쌍으로 만들고자 하는 경우가 많습니다. 이전 예에서 ID가 12703인 한 고객은 세 개의 계정을 갖고 있습니다. 통합 노드를 사용하여 각 계정 유형에 대한 총 잔액을 계산할 수 있습니다. 키 필드는 *CustID*이며 통합 필드는 새로 구조변환된 필드인 *Account_Draft_Current_Bal* 및 *Account_Savings_Current_Bal*입니다. 다음은 결과를 표시하는 표입니다.

표 30. 구조변환 및 통합 후의 표본 데이터.

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

구조변환 노드에 대한 옵션 설정

사용 가능 필드. 측정 수준이 *명목형*(변수군) 또는 *플래그*인 모든 데이터 필드를 나열합니다. 목록에서 하나를 선택하여 변수군 또는 플래그 내의 값을 표시한 다음 해당 값에서 선택하여 구조변환 필드를 작성하십시오. 사용 가능한 필드 및 해당 값을 보려면 데이터가 업스트림 소스 또는 유형 노드를 사용하여 완전히 인스턴스화되어야 합니다. 자세한 정보는 138 페이지의 『유형 노드』의 내용을 참조하십시오.

사용가능 값. 위에서 선택된 변수군 내의 값이 여기에 표시됩니다. 구조변환 필드를 생성할 값을 하나 이상 선택하십시오. 예를 들어, *혈압* 필드의 값이 *높음*, *중간* 및 *낮음*이면 *높음*을 선택하고 이를 오른쪽 목록에 추가할 수 있습니다. 그러면 값이 *높음*인 레코드에 대해 지정된 값(아래 참조)이 있는 필드가 생성됩니다.

구조변환 필드 생성. 새로 생성된 구조변환 필드가 여기에 나열됩니다. 기본적으로 새 필드 이름은 원래 필드 이름과 필드 값을 레이블에 조합하여 자동으로 생성됩니다. 예를 들어, *Fieldname_fieldvalue*와 같습니다.

필드 이름 포함. 새 필드 이름에서 접두문자로 원래 필드 이름을 제거하려면 선택 취소하십시오.

기타 필드의 값 사용. 구조변환 필드를 채우는 데 값을 사용할 필드를 한 개 이상 지정하십시오. 한 개 이상의 필드를 선택하려면 필드 선택기를 사용하십시오. 선택된 각 필드에 대해 하나의 새 필드가 생성됩니다. 값 필드 이름이 구조변환된 필드 이름에 추가됩니다. 예를 들어, *BP_High_Age* 또는 *BP_Low_Age*와 같습니다. 각 새 필드는 원래 값 필드의 유형을 상속합니다.

숫자 값 플래그 생성. 다른 필드의 값을 사용하지 않고 숫자 값 플래그(거짓의 경우 0, 참의 경우 1)로 새 필드를 채우려면 선택하십시오.

전치 노드

기본적으로 열은 필드이고 행은 레코드 또는 관측입니다. 필요한 경우, 필드가 레코드가 되고 레코드가 필드가 되도록 전치 노드를 사용하여 행과 열의 데이터를 스왑할 수 있습니다. 예를 들어, 각 계열이 열이 아니고 행인 시계열 데이터가 있는 경우, 분석 전에 데이터를 전치시킬 수 있습니다.

전치 노드의 옵션 설정

새 필드 이름

새 필드 이름은 지정된 접두부를 기반으로 자동으로 생성하거나 데이터의 기존 필드에서 읽어올 수 있습니다.

접두부 사용. 이 옵션은 지정된 접두부(*Field1*, *Field2* 등)를 기반으로 자동으로 새 필드 이름을 생성합니다. 필요에 따라 접두부를 사용자 정의할 수 있습니다. 이 옵션을 사용하는 경우에는 원래 데이터의 행 수와 상관없이 작성할 필드 수를 지정해야 합니다. 예를 들어, **새 필드 수**를 100으로 설정하면 처음 100행 이후의 모든 데이터는 삭제됩니다. 원래 데이터에 100개 미만의 행이 있는 경우 일부 필드가 널이 됩니다. (필요에 따라 필드 수를 늘릴 수 있지만, 이 설정의 목적은 백만 개의 레코드를 백만 개의 필드로 전치하는 것을 피하기 위한 것입니다. 이러한 전치는 감당할 수 없는 결과를 초래할 수 있습니다.

예를 들어, 행에 계열이 있고 각 월에 대한 개별 필드(열)가 있는 데이터가 있다고 가정하십시오. 각 계열이 개별 필드에 있고 각 월이 하나의 행에 있도록 이 데이터를 전치할 수 있습니다.

필드에서 읽기. 기존 필드에서 필드 이름을 읽습니다. 이 옵션을 사용하는 경우, 새 필드 수는 데이터에 의해 지정된 최대값까지 결정됩니다. 선택된 필드의 각 값은 출력 데이터에서 새 필드가 됩니다. 선택된 필드의 저장 유형(정수, 문자열, 날짜 등)에는 제한이 없지만, 중복된 필드 이름을 피하려면 선택된 필드의 각 값이 고유해야 합니다(즉, 값 수가 행 수와 일치해야 함). 중복된 필드 이름이 발견되면 경고가 표시됩니다.

- **값 읽기.** 선택된 필드가 인스턴스화되지 않은 경우 이 옵션을 선택하여 새 필드 이름 목록을 채우십시오. 필드가 이미 인스턴스화된 경우에는 이 단계가 필요하지 않습니다.
- **읽을 값 최대 수.** 데이터에서 필드 이름을 읽을 때 지나치게 많은 수의 필드가 작성되지 않도록 하기 위해 상한을 지정합니다. (앞에서 언급한 바와 같이, 백만 개의 레코드를 백만 개의 필드로 전치하는 경우 감당할 수 없는 결과를 초래합니다.)

예를 들어, 데이터의 첫 번째 열이 각 계열의 이름을 지정하는 경우, 전치된 데이터에서 이러한 값을 필드 이름으로 사용할 수 있습니다.

전치. 기본적으로 연속형(숫자 범위) 필드만 전치됩니다(정수 또는 실제 저장 공간). 선택적으로 숫자 필드의 서브세트를 선택하거나 문자열 필드를 대신 전치할 수 있습니다. 그러나, 전치되는 모든 필드는 동일한 저장 유형(숫자 또는 문자열이어야 하고 둘 다가 아니어야 함)이어야 합니다. 그 이유는 입력 필드를 혼합할 경우 각 출력 열 내에 혼합된 값이 생성되며, 이는 필드의 모든 값이 동일한 저장 공간을 가져야 하는 규칙을 위반하기 때문입니다. 다른 저장 유형(날짜, 시간, 시간소인)은 전치할 수 없습니다.

- **모든 숫자.** 모든 숫자 필드(정수 또는 실제 저장 공간)를 전치합니다. 출력의 행 수가 원래 데이터의 숫자 필드 수와 일치합니다.
- **모든 문자열.** 모든 문자열 필드를 전치합니다.
- **사용자 정의.** 숫자 필드의 서브세트를 선택할 수 있습니다. 출력의 행 수가 선택된 필드 수와 일치합니다. *참고*: 이 옵션은 숫자 필드에만 사용할 수 있습니다.

행 ID 이름. 노드에 의해 작성되는 행 ID 필드의 이름을 지정합니다. 이 필드의 값은 원래 데이터에 있는 필드의 이름으로 결정됩니다.

팁: 행의 시계열 데이터를 열로 전치할 때, 원래의 데이터에 각 측정 주기의 레이블을 지정하는 행(예: 날짜, 월, 연도 등)이 포함되는 경우, 데이터의 첫 번째 행에 해당 레이블을 포함시키기 보다는 이러한 레이블을 IBM SPSS Modeler에 필드 이름으로 읽어와야 합니다(위의 예에 설명된 바와 같이, 월 또는 날짜를 각각 원래 데이터의 필드 이름으로 표시함). 그러면 각 열에서 레이블과 값이 혼합되지 않습니다(열 내에서 저장 유형을 혼합할 수 없기 때문에 숫자가 문자열로 읽혀지도록 함).

히스토리 노드

히스토리 노드는 시계열 데이터 같은 순차 데이터에 가장 자주 사용합니다. 이전 레코드에 있는 필드의 데이터가 포함된 새 필드를 작성하는 데 사용됩니다. 히스토리 노드를 사용하는 경우 특정 필드별로 사전 정렬되는 데이터를 가지길 원할 수 있습니다. 정렬 노드를 사용하여 이를 수행할 수 있습니다.

히스토리 노드에 대한 옵션 설정

선택된 필드. 필드 선택기(텍스트 상자 오른쪽의 단추)를 사용하여 히스토리를 원하는 필드를 선택하십시오. 각각의 선택된 필드는 데이터 세트의 모든 레코드에 대한 새 필드를 작성하는 데 사용됩니다.

오프셋. 히스토리 필드 값을 추출할 현재 레코드 이전의 마지막 레코드를 지정하십시오. 예를 들어, 오프셋이 3으로 설정된 경우에는 각각의 레코드가 이 노드를 통과할 때 이전 세 번째 레코드에 대한 필드 값이 현재 레코드에 포함됩니다. 범위 설정을 사용하여 이전 몇 번째 레코드까지 추출할지 지정하십시오. 화살표를 사용하여 오프셋 값을 조정하십시오.

범위. 값을 추출할 이전 레코드의 수를 지정하십시오. 예를 들어, 오프셋이 3으로 설정되고 범위가 5로 설정되면 노드를 통과하는 각각의 레코드에는 선택된 필드 목록에서 지정된 각각의 필드에 대해 5개의 필드가 추가됩니다. 이는 노드가 레코드 10을 처리하면 레코드 7부터 레코드 3까지의 필드가 추가됨을 의미합니다. 화살표를 사용하여 범위 값을 조정하십시오.

히스토리가 사용 불가능한 경우. 히스토리 값이 없는 레코드를 처리하기 위해 다음 옵션 중 하나를 선택하십시오. 이는 일반적으로 히스토리로 사용할 이전 레코드가 없는 데이터 세트 맨 위의 처음 몇몇 레코드를 나타냅니다.

- **레코드 삭제.** 선택된 필드에 사용 가능한 히스토리 값이 없는 레코드를 삭제하려면 선택하십시오.
- **히스토리를 정의하지 않은 상태로 두기.** 히스토리 값을 사용할 수 없는 레코드를 보존하려면 선택하십시오. 히스토리 필드는 `$null$`로 표시된 정의되지 않은 값으로 채워집니다.
- **값 채우기.** 히스토리 값을 사용할 수 없는 레코드에 사용할 값 또는 문자열을 지정하십시오. 기본 대체값은 시스템 널인 *undef*입니다. 널값은 문자열 `$null$`를 사용하여 표시됩니다.

대체값을 선택할 때는 적절한 실행이 수행되도록 하기 위해 다음과 같은 규칙에 유의하십시오.

- 선택된 필드의 저장 유형은 동일해야 합니다.
- 선택된 필드가 모두 숫자 저장 공간을 가진 경우 대체값은 정수로 구문 분석되어야 합니다.
- 선택된 필드가 모두 실수 저장 공간을 가진 경우 대체값은 실수로 구문 분석되어야 합니다.
- 선택된 필드가 모두 기호 저장 공간을 가진 경우 대체값은 문자열로 구문 분석되어야 합니다.
- 선택된 필드가 모두 날짜/시간 저장 공간을 가진 경우 대체값은 날짜/시간 필드로 구문 분석되어야 합니다.

위 조건 중에 충족되는 조건이 없는 경우에는 히스토리 노드 실행 시 오류가 수신됩니다.

필드 다시 정렬 노드

필드 다시 정렬 노드를 사용하면 필드를 다운스트림으로 표시하는 데 사용하는 기본 순서를 정의할 수 있습니다. 이 순서는 테이블, 목록 및 필드 선택기 같은 다양한 장소에서 필드의 표시에 영향을 줍니다. 이 작업은 관심 있는 필드를 더 잘 보이게 만들기 위해 넓은 데이터 세트에 대해 작업하는 등의 경우에 유용합니다.

필드 다시 정렬 설정 옵션

필드를 다시 정렬하는 방법은 사용자 정의 순서 및 자동 정렬 두 가지입니다.

사용자 정의 순서

사용자 정의 순서를 선택하면 모든 필드를 볼 수 있으며 화살표 단추를 사용할 수 있는 필드 이름 및 유형 테이블을 사용하여 사용자 정의 순서를 작성할 수 있습니다.

필드를 다시 정렬하려면 다음을 수행하십시오.

1. 테이블에서 필드를 선택하십시오. Ctrl-클릭 방법을 사용하여 다중 필드를 선택하십시오.
2. 단순 화살표 단추를 사용하여 필드를 한 행 위 또는 아래로 이동하십시오.
3. 선 화살표 단추를 사용하여 필드를 목록의 맨 아래나 맨 위로 이동하십시오.
4. [기타 필드]로 표시되는 구분선 행을 위 또는 아래로 이동하여 여기에 포함하지 않을 필드 정렬을 지정하십시오.

[기타 필드]에 대한 자세한 정보

기타 필드. [기타 필드] 구분선 행의 목적은 테이블을 두 개의 절반으로 나누는 것입니다.

- 구분선 행 위에 표시되는 필드는 테이블에 표시될 때 이 노드의 필드 다운스트림을 표시하는 데 사용된 모든 자연 순서의 맨 위에 정렬됩니다.
- 구분선 행 아래에 표시되는 필드는 테이블에 표시될 때 이 노드의 필드 다운스트림을 표시하는 데 사용된 모든 자연 순서의 맨 아래에 정렬됩니다.

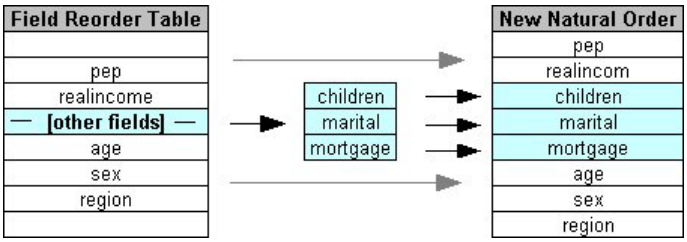


그림 6. "기타 필드"가 새 필드 순서에 통합되는 방법을 설명하는 다이어그램

- 필드 다시 정렬 테이블에 표시되지 않는 모든 기타 필드는 구분선 행의 위치에 의해 표시되는 대로 "위쪽" 및 "아래쪽" 필드 사이에 표시됩니다.

추가적인 사용자 정의 정렬 옵션은 다음과 같습니다.

- 각 열 헤더(**유형**, **이름**, **저장 공간**) 위에 있는 화살표를 클릭하여 오름차순 또는 내림차순으로 필드를 정렬하십시오. 열 기준으로 정렬하는 경우, 여기서 지정되지 않은 필드([기타 필드] 행으로 표시됨)는 자연 순서 마지막에 정렬됩니다.
- **미사용 항목 지우기**를 클릭하여 필드 다시 정렬 노드에서 사용되지 않은 필드를 모두 삭제하십시오. 사용되지 않은 필드는 테이블에 빨강 글꼴로 표시됩니다. 이는 필드가 업스트림 조작에서 삭제되었음을 표시합니다.
- (새 필드 또는 지정되지 않은 필드를 표시하기 위한 빛나는 아이콘으로 표시되는) 새 필드에 대한 순서를 지정하십시오. **확인** 또는 **적용**을 클릭하면 아이콘이 없어집니다.

참고: 사용자 정의 순서가 적용된 후에 필드가 업스트림에 추가되면 새 필드가 사용자 정의 목록의 아래쪽에 추가됩니다.

자동 정렬

정렬 모수를 지정하려면 **자동 정렬**을 선택하십시오. 자동 정렬에 대한 옵션을 제공하기 위해 대화 상자 옵션이 동적으로 변경됩니다.

정렬기준. 필드 읽기를 다시 정렬 노드로 정렬하는 세 가지 방법 중 하나를 선택하십시오. 화살표 단추는 순서가 오름차순 또는 내림차순인지 표시합니다. 변경하려면 하나를 선택하십시오.

- 이름
- 유형
- 저장 공간

자동 정렬이 적용된 후에 필드 다시 정렬 노드의 업스트림에 추가된 필드는 선택된 정렬 유형을 기준으로 하여 자동으로 적절한 위치에 배치됩니다.

시간 간격 노드

SPSS Modeler에 있는 원래의 시간 간격 노드는 Analytic Server(AS)와 호환되지 않으며 SPSS Modeler 릴리스 18.0에서 더 이상 사용되지 않습니다.

대체 시간 간격 노드(SPSS Modeler 릴리스 17.0의 새로운 사항)에는 Analytic Server와 함께 사용할 수 있는 원래 시간 간격 노드의 함수 서브세트가 있습니다.

시간 간격 노드를 사용하여 간격을 지정하고 추정하거나 예측하기 위한 새 시간 필드를 파생시키십시오. 초부터 년까지, 모든 범위의 시간 간격이 지원됩니다.

노드를 사용하여 새 시간 필드를 파생시키십시오. 새 필드에 사용자가 선택한 입력 시간 필드와 동일한 저장 유형이 있습니다. 노드는 다음 항목을 생성합니다.

- 선택한 접두부/접미부와 함께 필드 탭에 **시간 필드**로 지정된 필드. 기본적으로 접두부는 `$TI_`입니다.
- 필드 탭에 **차원 필드**로 지정된 필드.
- 필드 탭에 **통합할 필드**로 지정된 필드.

선택한 간격 또는 기간(예: 측정이 속하는 분 또는 초)에 따라 여러 추가 필드가 생성될 수도 있습니다.

시간 간격 - 필드 옵션

시간 간격 노드의 필드 탭을 사용하면 새 시간 간격이 파생되는 데이터를 선택할 수 있습니다.

필드 노드에 대한 모든 입력 필드와 해당 측정 유형 아이콘을 표시합니다. 모든 시간 필드에는 '연속형' 측정 유형이 있습니다. 입력으로 사용할 필드를 선택하십시오.

시간 필드 새 시간 간격이 파생되는 입력 필드를 표시합니다. 단일 연속형 필드만 허용됩니다. 이 필드는 시간 간격 노드에서 간격을 변환하기 위한 통합 키로 사용합니다. 새 필드에는 선택한 입력 시간 필드와 동일한 저장 유형이 있습니다. 정수 필드를 선택하는 경우 시간 지수로 간주됩니다.

차원 필드 선택적으로 여기에 필드를 추가하여 필드 값에 기반한 개별 시계열을 작성할 수 있습니다. 한 예로 특정 지역과 관련된 데이터를 사용하는 경우 차원으로 위치 필드를 사용할 수 있습니다. 이 예에서 시간 간격 노드의 데이터 출력이 위치 필드에 있는 각 위치 값의 시계열로 정렬됩니다.

통합할 필드 시간 필드의 기간을 변경하기 위한 일부분으로 통합할 필드를 선택하십시오. 여기에서 선택하는 필드만 **지정된 필드의 사용자 정의 설정** 테이블의 작성 탭에서 사용할 수 있습니다. 포함되지 않은 필드는 노드에서 나가는 데이터에서 필터링되어 제거됩니다. 즉, **필드** 목록에 남아 있는 필드가 데이터에서 필터링되어 제거됩니다.

시간 간격 - 작성 옵션

작성 탭을 사용하면 시간 간격을 변경하기 위한 옵션과 해당 측정 유형에 기반하여 데이터의 필드를 통합하는 방식을 지정할 수 있습니다.

데이터를 통합할 때 기존 날짜, 시간 또는 시간소인 필드는 생성 필드로 대체되고 출력에서 삭제됩니다. 기타 필드는 사용자가 이 탭에 지정하는 옵션에 기반하여 통합됩니다.

시간 간격 계열을 작성하기 위한 간격 및 주기성을 선택하십시오. 자세한 정보는 지원되는 구간의 내용을 참조하십시오.

기본 설정 다른 유형의 데이터에 적용할 기본 통합을 선택하십시오. 기본값은 측정 수준에 기반하여 적용됩니다. 예를 들어, 명목 필드는 모드를 사용하지만 연속형 필드는 합계로 통합됩니다. 세 가지 다른 측정 수준의 기본값을 설정할 수 있습니다.

- **연속형** 연속형 필드의 사용 가능한 함수로는 **Sum**, **Mean**, **Min**, **Max**, **Median**, **1st Quartile**, **3rd Quartile**이 있습니다.
- **명목** 옵션으로는 **Mode**, **Min**, **Max**가 있습니다.
- **플래그** 옵션은 **True if any true** 또는 **False if any false**입니다.

지정된 필드의 사용자 정의 설정 개별 필드의 기본 통합 설정에 대한 예외를 지정할 수 있습니다. 오른쪽의 아이콘을 사용하여 테이블에서 필드를 추가 또는 제거하거나 적절한 열의 셀을 클릭하여 해당 필드의 통합 함수를 변경하십시오. 유형이 없는 필드는 목록에서 제외되고 테이블에 추가될 수 없습니다.

새 필드 이름 확장자 노드에 의해 생성되는 모든 필드에 적용되는 **접두부** 또는 **접미부**를 지정하십시오.

재투영 노드

지리 공간적 데이터 또는 맵 데이터를 사용하는 경우, 좌표를 식별하기 위해 사용되는 가장 일반적인 두 가지 방법은 투영된 좌표 및 지리적 좌표계입니다. IBM SPSS Modeler에서 표현식 작성기 공간 기능, STP(Spatio-Temporal Prediction) 노드 및 맵 시각화 노드와 같은 항목은 투영된 좌표계를 사용하며, 따라서 지리적 좌표계로 기록된 가져온 데이터를 다시 투영해야 합니다. 가능하면 지리 공간적 필드(지리 공간적 측정 수준이 있는 모든 필드)가 가져올 때가 아니라 사용될 때 자동으로 재투영됩니다. 자동으로 재투영할 수 있는 필드가 없으면 재투영 노드를 사용하여 좌표계를 변경할 수 있습니다. 이 방법으로 재투영하면 잘못된 좌표계를 사용하여 오류가 발생하는 상황을 정정할 수 있습니다.

좌표계를 변경하기 위해 재투영해야 하는 상황의 예는 다음과 같습니다.

- *붙여쓰기* 지리 공간적 필드에 대한 좌표계가 다른 두 데이터 세트를 붙여쓰려고 시도하면 SPSS Modeler에서 다음과 같은 오류 메시지를 표시합니다. `Coordinate systems of <Field1> and <Field2> are not compatible. Reproject one or both fields to the same coordinate system.`

  <*Field1*> 및 <*Field2*>는 오류를 발생시킨 지리 공간적 필드의 이름입니다.
- *If/else 표현식* 표현식의 두 파트 모두에 지리 공간적 필드 또는 리턴 유형이 있으나 좌표계가 다른 if/else 명령문을 포함하는 표현식을 사용하면 SPSS Modeler에서 다음과 같은 오류 메시지를 표시합니다. `The conditional expression contains incompatible coordinate systems: <arg1> and <arg2>.`

  <*arg1*> 및 <*arg2*>는 좌표계가 다른 지리 공간적 유형을 리턴하는 then 또는 else 인수입니다.
- *지리 공간적 필드 목록 생성* 수많은 지리 공간적 필드로 구성된 목록 필드를 생성하려면 목록 표현식에 제공되는 모든 지리 공간적 필드 인수가 도일한 좌표계 내에 있어야 합니다. 그렇지 않으면 다음과 같은 오류 메시지가 표시됩니다. `Coordinate systems of <Field1> and <Field2> are not compatible. Reproject one or both fields to the same coordinate system.`

좌표계에 대한 자세한 정보는 SPSS Modeler 사용자 안내서의 스트림 작업 절에서 스트림에 대한 지리 공간적 옵션 설정 주제를 참조하십시오.

재투영 노드에 대한 설정 옵션

필드

지리 필드

기본적으로 이 목록은 비어있습니다. **재투영할 필드** 목록에서 이 목록으로 지리 공간적 필드를 이동하여 해당 필드가 재투영되지 않도록 할 수 있습니다.

재투영할 필드

기본적으로 이 목록에는 이 노드에 대한 입력이 되는 모든 지리 공간적 필드가 포함됩니다. 이 목록 내의 모든 필드는 사용자가 **좌표계** 영역에서 설정한 좌표계로 재투영됩니다.

좌표계

스트림 기본값

기본 좌표계를 사용하려면 이 옵션을 선택하십시오.

지정 이 옵션을 선택하면 **변경** 단추를 사용하여 좌표계 선택 대화 상자를 표시하고 재투영에 사용할 좌표계를 선택할 수 있습니다.

좌표계에 대한 자세한 정보는 SPSS Modeler 사용자 안내서의 스트림 작업 절에서 스트림에 대한 지리 공간적 옵션 설정 주제를 참조하십시오.

제 5 장 그래프 노드

공통 그래프 노드 기능

데이터 마이닝 프로세스의 여러 단계에서는 그래프 및 차트를 사용하여 IBM SPSS Modeler로 가져온 데이터를 탐색합니다. 예를 들어, 도표 또는 분포 노드를 데이터 소스에 연결하여 데이터 유형 및 분포를 살펴볼 수 있습니다. 그런 다음 레코드 및 필드 조작을 수행하여 다운스트림 모델링 조작을 위해 데이터를 준비할 수 있습니다. 그래프의 또다른 공통 사용법은 새로 파생된 필드 간 분포 및 관계를 확인하는 것입니다.

그래프 팔레트에는 다음과 같은 노드가 포함되어 있습니다.

그래프보드 노드는 하나의 단일 노드에 있는 여러 가지 유형의 많은 그래프를 제공합니다. 이 노드를 사용하여 탐색하려는 데이터 필드를 선택하고 선택된 데이터에 대해 사용 가능한 것 중에서 그래프를 선택할 수 있습니다. 이 노드는 필드 선택사항에 대해 작업하지 않는 모든 그래프 유형을 자동으로 필터링합니다.

구성 노드는 수치 필드 사이의 관계를 보여줍니다. 포인트(산점도) 또는 선을 사용하여 도표를 작성할 수 있습니다.

분포 노드는 대출 유형이나 성별 같은 기호적(범주형) 값의 발생을 보여줍니다. 일반적으로, 데이터의 불균형을 표시하기 위해 분포 노드를 사용하는 경우 모델을 작성하기 전에 균형 노드를 사용하여 교정할 수 있습니다.

히스토그램 노드는 수치 필드에 대한 값의 발생을 표시합니다. 보통 조작 및 모델 작성 전에 데이터를 탐색하는 데 사용합니다. 분포 노드와 비슷하게, 히스토그램 노드는 자주 데이터의 불균형을 드러내 보입니다.

요약도표 노드는 다른 필드의 값에 상대적으로 하나의 숫자 필드의 값의 분포를 표시합니다. (히스토그램과 유사한 그래프를 작성합니다.) 값이 시간에 따라 변하는 변수 또는 필드를 설명하는 데 유용합니다. 3-D 그래프를 사용하여 범주별 분포를 표시하는 기호 축을 포함할 수도 있습니다.

다중 도표 노드는 단일 *X* 필드 위에 다중 *Y* 필드를 표시하는 도표를 작성합니다. *Y* 필드는 색상이 지정된 선으로 도표됩니다. 각각은 스타일이 **Line**으로 설정되고 X 모드가 **Sort**로 설정된 구성 노드와 동등합니다. 다중 도표는 시간에 따라서 여러 변수의 변동을 탐색하기 원할 때 유용합니다.

웹 노드는 둘 이상의 기호(범주형) 필드의 값 사이의 관계의 강도를 설명합니다. 그래프는 다양한 너비의 선을 사용하여 연결 강도를 표시합니다. 예를 들어 웹 노드를 사용하여 전자상거래 사이트에 있는 항목 세트의 구매 사이의 관계를 탐색할 수 있습니다.

시간 구성 노드는 하나 이상의 시계열 데이터 세트를 표시합니다. 일반적으로, 먼저 시간 간격 노드를 사용하여 *TimeLabel* 필드를 작성하는데, 이것이 *x*축을 레이블하는 데 사용합니다.

평가 노드는 예측 모델을 평가하고 비교하는 데 도움이 됩니다. 평가 차트는 모델이 특정 결과를 얼마나 잘 예측하는지를 보여줍니다. 예측값과 예측의 신뢰도를 바탕으로 레코드를 정렬합니다. 레코드를 동일한 크기의 그룹(**분위수**)으로 분할한 후 각 분위수에 대한 비즈니스 기준의 값을 가장 높은 값부터 가장 낮은 값으로 도표를 그립니다. 다중 모델이 도표에 선구분 변수로 표시됩니다.

맵 시각화 노드는 다중 입력 연결을 승인하고 지리 공간적 데이터를 맵에 일련의 레이어로 표시할 수 있습니다. 각각의 레이어는 하나의 지리 공간적 필드입니다. 예를 들어, 기준 레이어가 한 국가의 맵이고 그 위에 도로에 대한 레이어 하나, 강에 대한 레이어 하나, 도시에 대한 레이어 하나가 있을 수 있습니다.

그래프 노드를 스트림에 추가한 경우에는 해당 노드를 두 번 클릭하여 옵션을 지정하기 위한 대화 상자를 열 수 있습니다. 대부분의 그래프에는 하나 이상의 탭에 제공된 다수의 고유 옵션이 포함되어 있습니다. 모든 그래프에 공통인 몇몇 탭 옵션도 있습니다. 다음의 절에는 이 공통 옵션에 대한 자세한 정보가 포함되어 있습니다.

그래프 노드에 대한 옵션을 구성한 경우에는 대화 상자 내에서 또는 스트림의 일부로 해당 옵션을 실행할 수 있습니다. 생성된 그래프 창에서는 데이터의 영역 또는 선택사항을 기반으로 파생(세트 및 플래그) 및 선택 노드를 생성하여 사실상 데이터의 "서브세트를 작성"할 수 있습니다. 예를 들어, 이 강력한 기능을 사용하여 이상치를 식별하여 제외할 수 있습니다.

모양, 오버레이, 패널 및 애니메이션

오버레이 및 모양

모양(및 오버레이)은 시각화에 차원성을 추가합니다. 모양(그룹화, 군집화 또는 누적)의 효과는 시각화 유형, 필드(변수) 유형, 그래픽 요소 유형 및 통계에 따라 다릅니다. 예를 들어, 색상에 대한 범주형 필드는 산점도에서 점을 그룹화하거나 누적 막대형 차트에서 누적을 작성하는 데 사용할 수 있습니다. 색상에 대한 연속형 숫자 범위는 산점도에 있는 각 점의 범위 값을 표시하는 데 사용할 수 있습니다.

요구에 맞는 모양과 오버레이를 찾으려면 여러 모양과 오버레이를 사용하여 실험해 보아야 합니다. 다음 설명은 적합한 모양과 오버레이를 선택하는 데 도움이 될 수 있습니다.

참고: 모든 모양 또는 오버레이를 모든 시각화 유형에 사용할 수 있는 것은 아닙니다.

- **색상.** 색상이 범주형 필드에 의해 정의되면 개별 범주를 기준으로 각 범주에 한 색상씩 시각화를 분할합니다. 색상이 연속형 숫자 범위일 때는 범위 필드의 값에 따라 색상이 달라집니다. 그래픽 요소(예: 막대 또는 선택란)가 둘 이상의 레코드/케이스를 나타내고 범위 필드가 색상에 사용되는 경우 범위 필드의 *평균*에 따라 색상이 달라집니다.
- **형태.** 형태는 시각화를 각 범주에 하나씩 서로 다른 여러 형태의 요소로 분할하는 범주형 필드에 의해 정의됩니다.

- **투명도.** 투명도가 범주형 필드에 의해 정의되는 경우 개별 범주를 기준으로 각 범주에 한 투명도 수준씩 시각화를 분할합니다. 투명도가 연속형 숫자 범위일 때는 범위 필드의 값에 따라 투명도가 달라집니다. 그래픽 요소(예: 막대 또는 선택란)가 둘 이상의 레코드/케이스를 나타내고 범위 필드가 투명도에 사용되는 경우 범위 필드의 *평균*에 따라 색상이 달라집니다. 가장 큰 값에서 그래픽 요소는 완전 투명합니다. 가장 작은 값에서는 완전 불투명합니다.
- **데이터 레이블.** 데이터 레이블은 해당 값이 그래픽 요소에 연결되는 레이블을 작성하는 데 사용되는 필드 유형에 의해 정의됩니다.
- ***크기.*** *크기*가 범주형 필드에 의해 정의되면 개별 범주를 기준으로 각 범주에 한 *크기*씩 시각화를 분할합니다. *크기*가 연속형 숫자 범위일 때는 범위 필드의 값에 따라 *크기*가 달라집니다. 그래픽 요소(예: 막대 또는 선택란)가 둘 이상의 레코드/케이스를 나타내고 범위 필드가 *크기*에 사용되는 경우 범위 필드의 *평균*에 따라 *크기*가 달라집니다.

패널링 및 애니메이션

패널링. 패널링(면 작성이라고도 함)은 그래프 테이블을 작성합니다. 패널링 필드에 범주당 하나의 그래프가 생성되지만 모든 패널이 동시에 표시됩니다. 패널링은 패널링 필드의 조건이 시각화에 영향을 미치는지 여부를 확인하는 데 유용합니다. 예를 들어, 빈도 분포가 남성과 여성 간에 동일한지 판별하기 위해 성별로 히스토그램을 패널링할 수 있습니다. 즉, 성별 차이가 급여에 영향을 미치는지 여부를 확인할 수 있습니다. 패널링할 범주형 필드를 선택하십시오.

애니메이션. 애니메이션은 애니메이션 필드의 값으로 여러 그래프가 작성된다는 점에서 패널링과 비슷하지만 이러한 그래프는 함께 표시되지 않습니다. 오히려 사용자가 탐색 모드의 제어를 사용하여 출력을 애니메이션하고 개별 그래프를 시퀀스대로 표시합니다. 또한 패널링과 달리 애니메이션은 범주형 필드를 필요로 하지 않습니다. 값이 자동으로 범위로 분할되는 연속형 필드를 지정할 수 있습니다. 탐색 모드에서 애니메이션 제어로 범위의 *크기*를 바꿀 수 있습니다. 모든 시각화에서 애니메이션을 제공하는 것은 아닙니다.

출력 탭 사용

모든 그래프 유형에 대해 생성된 그래프의 표시 및 파일 이름에 대해 다음과 같은 옵션을 지정할 수 있습니다.

참고: 분포 노드 그래프에는 추가적인 설정이 있습니다.

출력 이름. 노드가 실행될 때 생성되는 그래프의 이름을 지정합니다. **자동**은 출력을 생성하는 노드를 기반으로 이름을 선택합니다. 선택적으로 **사용자 정의**를 선택하여 다른 이름을 지정할 수 있습니다.

화면으로 출력. 새 창에서 그래프를 생성하고 표시하려면 선택하십시오.

파일로 출력. 출력을 파일로 저장하려면 선택하십시오.

- **그래프 출력.** 그래프 형식으로 출력을 생성하려면 선택하십시오. 분포 노드에서만 사용할 수 있습니다.
- **테이블 출력.** 테이블 형식으로 출력을 생성하려면 선택하십시오. 분포 노드에서만 사용할 수 있습니다.
- **파일 이름.** 생성되는 그래프 또는 테이블에 사용되는 파일 이름을 지정하십시오. 생략 기호 단추(**...**)를 사용하여 특정 파일 및 위치를 지정하십시오.

- **파일 유형.** 드롭 다운 목록에서 파일 유형을 지정하십시오. **테이블 출력** 옵션을 가진 분포 노드를 제외한 모든 그래프 노드에 대해 사용 가능한 그래프 파일 유형은 다음과 같습니다.

  - 비트맵(*.bmp*)
  - PNG(*.png*)
  - 출력 오브젝트(*.cou*)
  - JPEG(*.jpg*)
  - HTML(*.html*)
  - 다른 IBM SPSS Statistics 애플리케이션에서 사용하기 위한 ViZml 문서(*.xml*)

  분포 노드의 **테이블 출력** 옵션에 대해 사용 가능한 파일 유형은 다음과 같습니다.

  - 탭으로 구분된 데이터(*.tab*)
  - 쉼표로 구분된 데이터(*.csv*)
  - HTML(*.html*)
  - 출력 오브젝트(*.cou*)

출력 페이지 번호 매기기. 출력을 HTML로 저장하면 이 옵션이 사용으로 설정되어 각 HTML 페이지의 *크기*를 제어할 수 있습니다. (분포 노드에만 적용됩니다.)

페이지당 행 수. **출력 페이지 번호 매기기**가 선택되면 이 옵션이 사용으로 설정되어 각 HTML 페이지의 길이를 판별할 수 있습니다. 기본 설정은 400개의 행입니다. (분포 노드에만 적용됩니다.)

주석(Annotation) 탭 사용

이 탭은 모든 노드에 사용되며 노드의 이름을 바꾸고 사용자 맞춤 도구팁을 제공하며 긴 주석(Annotation)을 저장하기 위한 옵션을 제공합니다.

3차원 그래프

IBM SPSS Modeler의 도표 및 콜렉션 그래프는 세 번째 축에 정보를 표시할 수 있습니다. 이를 통해 서브세트를 선택하기 위해 데이터를 시각화하고 모델링을 위해 새 필드를 파생시킬 때 추가적인 유연성이 제공됩니다.

3차원 그래프를 작성한 후에는 해당 그래프를 클릭한 후 마우스를 끌어서 회전시켜 임의의 각에서 볼 수 있습니다.

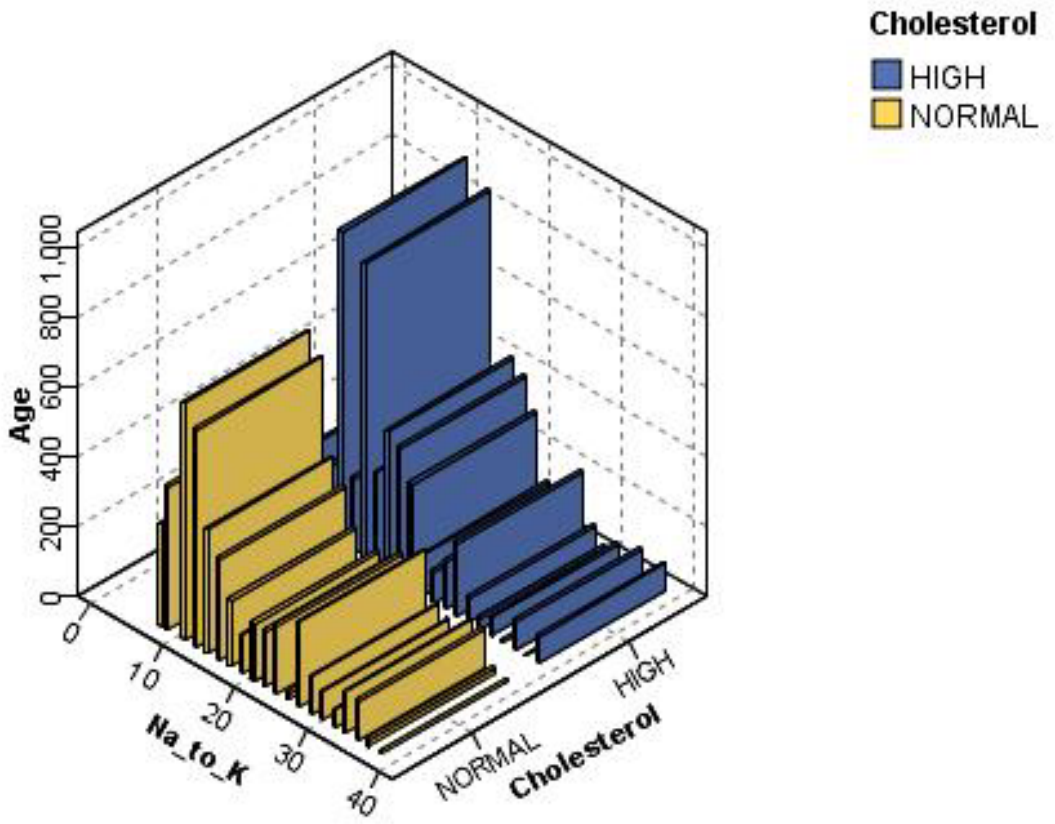


그림 7. x, y 및 z축이 있는 콜렉션 그래프

IBM SPSS Modeler에서는 세 번째 축에서 정보를 도표화(진정한 3차원 그래프)하고 3차원 효과로 그래프를 표시하는 두 가지 방식으로 3차원 그래프를 작성할 수 있습니다. 두 방법 모두 도표 및 콜렉션에 사용할 수 있습니다.

세 번째 축에서 정보를 도표화하려면 다음을 수행하십시오.

1. 그래프 노드 대화 상자에서 **도표** 탭을 클릭하십시오.
2. 3차원 단추를 클릭하여 *z*축에 대한 옵션을 사용으로 설정하십시오.
3. 필드 선택기 단추를 사용하여 *z*축에 대한 필드를 선택하십시오. 일부 경우에는 기호 필드만 여기서 허용됩니다. 필드 선택기는 적절한 필드를 표시합니다.

3차원 효과를 그래프에 추가하려면 다음을 수행하십시오.

1. 그래프를 작성하고 나면 출력 창에서 **그래프** 탭을 클릭하십시오.
2. 3차원 단추를 클릭하여 보기를 3차원 그래프로 전환하십시오.

그래프보드 노드

그래프보드 노드를 사용하면 단일 노드에서 다양한 그래프 출력(막대형 차트, 원형 차트, 히스토그램, 산점도, 히트 맵 등) 중에서 선택할 수 있습니다. 첫 번째 탭에서 탐색할 데이터 필드를 선택하여 시작하면 노드에서 데이터에 대해 작동하는 그래프 유형을 선택할 수 있습니다. 이 노드는 필드 선택사항에 대해 작업하지 않는 모든 그래프 유형을 자동으로 필터링합니다. 세부사항 탭에서 상세 또는 고급 그래프 옵션을 정의할 수 있습니다.

참고: 노드를 편집하거나 그래프 유형을 선택하기 위해 그래프보드 노드를 데이터가 있는 스트림에 연결해야 합니다.

사용 가능한 시각화 템플리트(및 스타일 시트 및 맵)를 제어할 수 있게 하는 두 개의 단추가 있습니다.

관리. 컴퓨터에서 시각화 템플리트 , 스타일시트 및 맵을 관리합니다. 로컬 시스템에서 시각화 템플리트 , 스타일시트 및 맵을 가져오고, 내보내고, 이름을 바꾸고, 삭제할 수 있습니다. 자세한 정보는 223 페이지의 『템플리트, 스타일시트 및 맵 파일 관리』 주제를 참조하십시오.

위치. 시각화 템플리트 , 스타일시트 및 맵이 저장된 위치를 변경합니다. 현재 위치는 단추의 오른쪽에 표시됩니다. 자세한 정보는 222 페이지의 『템플리트, 스타일시트 및 맵 위치 설정』의 내용을 참조하십시오.

그래프보드 기본 탭

어느 시각화 유형이 데이터를 가장 잘 표현하는지 확신할 수 없는 경우에는 기본 탭을 사용하십시오. 데이터를 선택하면 해당 데이터에 적합한 시각화 유형 서브세트가 표시됩니다. 예를 들어, 210 페이지의 『그래프보드 예제』의 내용을 참조하십시오.

1. 목록에서 하나 이상의 필드(변수)를 선택하십시오. 여러 개의 필드를 선택하려면 Ctrl+클릭을 사용하십시오.

   필드의 측정 수준은 사용 가능한 시각화 유형을 결정합니다. 목록에서 필드를 마우스 오른쪽 단추로 클릭하고 옵션을 선택하여 측정 수준을 변경할 수 있습니다. 사용 가능한 측정 수준 유형에 대한 자세한 정보는 200 페이지의 『필드(변수) 유형』의 내용을 참조하십시오.
2. 시각화 유형을 선택하십시오. 사용 가능한 유형에 대한 설명은 204 페이지의 『사용 가능한 내장 그래프보드 시각화 유형』의 내용을 참조하십시오.
3. 특정 시각화의 경우 요약 통계를 선택할 수 있습니다. 통계가 개수 기반 통계인지 또는 연속형 필드에서 계산되는지 여부에 따라 사용 가능한 통계 서브세트가 다릅니다. 또한 템플리트 자체에 따라서도 사용 가능한 통계가 다릅니다. 다음 단계 뒤에 사용 가능한 전체 통계 목록을 제공합니다.
4. 추가 옵션(예: 선택적 모양 및 패널 필드)을 정의하려면 **세부사항**을 클릭하십시오.자세한 정보는 202 페이지의 『그래프보드 세부사항 탭』의 내용을 참조하십시오.

연속형 필드에서 계산한 요약 통계

- *평균(Mean)*. 중심 경향에 대한 측도입니다. 합계를 케이스 수로 나눈 산술 평균 값입니다.

- *중앙값(Median)*. 전체 케이스의 절반이 위 아래에 해당되는 값으로 제50 백분위수입니다. 케이스 수가 짝수인 경우 중앙값은 케이스를 오름차순이나 내림차순으로 정렬했을 때 중간에 있는 두 개의 케이스의 평균입니다. 중앙값은 평균과 달리 중심을 벗어난 값에는 영향을 받지 않는 중심 경향 측도이며, 상한 극단값 또는 하한 극단값에 따라 달라질 수 있습니다.
- *최빈값(Mode)*. 가장 자주 발생하는 값입니다. 여러 값에서 최대 발생 빈도를 공유하는 경우 각각을 최빈값이라고 합니다.
- *최소값(Minimum)*. 숫자변수의 가장 작은 값입니다.
- *최대값(Maximum)*. 숫자변수의 가장 큰 값입니다.
- *범위*. 최소값과 최대값의 차이입니다.
- *중간 범위*. 범위의 중간 값으로 최소값과의 차이와 최대값과의 차이가 같은 값입니다.
- *합계(Sum)*. 비결측값을 갖는 전체 케이스 값의 총계입니다.
- *누적 합계*. 값의 누적 합계입니다. 각 그래픽 요소는 하나의 하위 그룹 합계와 이전의 모든 그룹의 총 합계를 더한 값을 표시합니다.
- *퍼센트 합계*. 모든 그룹의 합계에 대비되는 합산 필드 기준의 각 하위 그룹 내 백분율입니다.
- *누적 퍼센트 합계*. 모든 그룹의 합계에 대비되는 합산 필드 기준의 각 하위 그룹 내 누적 백분율입니다. 각 그래픽 요소는 하나의 하위 그룹의 백분율과 이전의 모든 그룹의 전체 백분율을 더한 값을 표시합니다.
- *분산(Variance)*. 평균에 대한 산포 측도로, 평균으로부터의 제곱합 편차를 케이스 수에서 1을 뺀 값으로 나눈 값과 같습니다. 분산은 변수 자체의 제곱 단위로 측정됩니다.
- *표준 편차(Standard Deviation)*. 평균에 대한 산포 측도입니다. 정규 분포에서 케이스의 68%는 평균의 표준 편차 내에 있으며 케이스의 95%는 2배 표준 편차 내에 있습니다. 예를 들어, 평균 연령이 45세이고 표준 편차가 10인 경우 정규 분포 내에서 95% 케이스는 25세와 65세 사이에 있습니다.
- *표준 오차(Standard Error)*. 검정 통계량 값이 표본마다 얼마나 달라지는지에 대한 측도입니다. 이 항목은 통계에 대한 표본 분포의 표준 편차가 됩니다. 예를 들어, 평균의 표준 오차는 표본 평균의 표준 편차입니다.
- *첨도(Kurtosis)*. 관측값이 중심에 군집하는 정도에 대한 측도입니다. 정규 분포의 경우 첨도 통계 값은 0입니다. 양의 첨도는 정규 분포에 비해 관측값이 분포 중심에 더 많이 군집되어 있고 분포 극단값까지의 꼬리가 더 얇다는 의미입니다. 즉, 정규 분포에 비해 급첨 분포의 꼬리가 더 두껍습니다. 음의 첨도는 정규 분포에 비해 관측값이 분포 중심에 덜 군집되어 있고 분포 극단값까지의 꼬리가 더 두껍다는 의미입니다. 즉, 정규 분포에 비해 평첨 분포의 꼬리가 더 얇습니다.
- *왜도(Skewness)*. 분포의 비대칭성에 대한 측도입니다. 정규 분포는 대칭이므로 왜도 값이 0입니다. 양의 왜도가 많은 분포는 오른쪽이 깁니다. 유의한 음의 왜도를 가지는 분포에는 왼쪽으로 긴 꼬리가 나타납니다. 왜도값이 표준 오차의 두 배를 넘는 것은 대칭에서 벗어난 정도를 나타냅니다.

다음과 같은 지역 통계는 하위 그룹당 둘 이상의 그래픽 요소를 생성할 수 있습니다. 구간, 영역 또는 가장자리 그래픽 요소를 사용하는 경우 지역 통계는 범위를 표시하는 하나의 그래픽 요소를 생성합니다. 다른 모든 그래픽 요소는 두 개의 개별 요소, 즉 범위의 시작을 표시하는 요소와 범위의 끝을 표시하는 요소를 생성합니다.

- **지역: 범위**. 최소값과 최대값 사이의 값 범위입니다.
- **지역: 평균의 95% 신뢰구간**. 모집단 평균을 포함할 가능성이 95%인 값 범위입니다.
- **지역: 개별의 95% 신뢰구간**. 주어진 개별 케이스의 예측값을 포함할 가능성이 95%인 값 범위입니다.
- **지역: 평균 이상/이하의 1배 표준 편차**. *평균*의 이상 및 이하로 1배 *표준 편차*만큼 떨어져 있는 값 사이의 범위입니다.
- **지역: 평균 이상/이하의 1배 표준 오차**. *평균*의 이상 및 이하로 1배 *표준 오차*만큼 떨어져 있는 값 사이의 범위입니다.

개수 기준 요약 통계

- **개수**. 행/케이스의 수입니다.
- **누적 개수**. 행/케이스의 수입니다. 각 그래픽 요소는 하나의 하위 그룹의 개수와 이전의 모든 그룹의 총계를 더한 값을 표시합니다.
- **개수 퍼센트**. 행/케이스의 총 수에 대비되는 각 하위 그룹의 행/케이스 백분율입니다.
- **누적 개수 퍼센트**. 행/케이스의 총 수에 대비되는 각 하위 그룹의 행/케이스 누적 백분율입니다. 각 그래픽 요소는 하나의 하위 그룹의 백분율과 이전의 모든 그룹의 전체 백분율을 더한 값을 표시합니다.

필드(변수) 유형

아이콘은 필드 목록에서 필드 옆에 표시되며 필드 유형과 데이터 유형을 나타냅니다. 아이콘은 또한 다중 응답 세트를 식별합니다.

표 31. 측정 수준 아이콘.

측정 수준	숫자	문자열	날짜	시간
연속형		해당사항 없음		
순서형 세트				
세트				

표 32. 다중 응답 세트 아이콘.

다중 반응 세트 유형	아이콘
다중 응답 세트, 다중 범주	
다중 응답 세트, 다중 이분형	

측정 수준

필드의 측정 수준은 시각화를 만들 때 중요한 역할을 합니다. 다음은 측정 수준에 대한 설명입니다. 필드 목록에서 필드를 마우스 오른쪽 단추로 클릭하고 옵션을 선택하여 측정 수준을 임시로 변경할 수 있습니다. 대부분의 경우 가장 광범위한 필드 분류인 범주형과 연속형 두 가지만 고려해야 합니다.

범주형. 고유한 값 또는 범주의 수가 제한된 데이터(예: 성별 또는 종교)입니다. 문자열(영숫자) 필드 또는 숫자 코드를 사용하여 범주를 나타내는 숫자 필드(예: 0 = *남성*, 1 = *여성*)가 범주형 필드에 해당될 수 있습니다. 질적 데이터라고도 합니다. 세트, 순서형 세트 및 플래그는 모두 범주형 필드입니다.

- *세트*. 해당 값이 고유한 순위가 없는 범주를 나타내는 필드/변수입니다(예: 직원이 근무하는 회사의 부서). 명목 변수의 예는 지역, 우편번호 및 종교입니다. 명목 변수라고도 합니다.
- *순서형 세트*. 해당 값이 고유한 순위가 있는 범주를 나타내는 필드/변수입니다(예: 매우 불만족에서 매우 만족에 이르는 서비스 만족도 수준). 순서형 세트의 예로는 만족도나 신뢰도를 나타내는 태도 스코어 및 선호도 등급 스코어가 있습니다. 순서 변수라고도 합니다.
- *플래그*. 두 개의 고유한 값(예: 예와 아니오 또는 1과 2)이 있는 필드/변수입니다. 이분형 변수라고도 합니다.

연속형. 구간 또는 비율 척도로 측정된 데이터이며 데이터 값은 값의 순서와 값 사이의 차이를 모두 나타냅니다. 예를 들어, 급여 $72,195는 급여 $52,398보다 높으며 두 값 사이의 차이는 $19,797입니다. 양적, 척도 또는 숫자 범위 데이터라고도 합니다.

범주형 필드는 일반적으로 별도의 그래픽 요소를 그리거나 그래픽 요소를 그룹화하기 위해 시각화에서 범주를 정의합니다. 연속형 필드는 대개 범주형 필드의 범주 내에 요약됩니다. 예를 들어, 성별 범주에 대한 수입의 기본 시각화에는 남자의 평균 수입과 여자의 평균 수입이 표시됩니다. 산점도에서와 마찬가지로 연속형 필드의 원래 값을 도표화할 수도 있습니다. 예를 들어, 산점도는 각 케이스의 현재 급여와 시작 급여를 표시할 수 있습니다. 범주형 필드를 사용하여 케이스를 성별로 그룹화할 수 있습니다.

데이터 유형

측정 수준이 필드 유형을 결정하는 필드의 유일한 특성인 것은 아닙니다. 필드는 특정 데이터 유형으로도 저장됩니다. 가능한 데이터 유형은 문자열(문자와 같이 숫자가 아닌 데이터), 숫자 값(실수) 및 날짜입니다. 필드의 데이터 유형은 측정 수준과 달리 임시로 변경할 수 없습니다. 데이터가 원래 데이터 세트에 저장되는 방식을 변경해야 합니다.

다중 응답 세트

일부 데이터 파일에서는 **다중 응답 세트**라는 특수한 종류의 "필드"도 지원합니다. 다중 응답 세트는 일반적인 의미에서는 실제로 "필드"가 아닙니다. 다중 응답 세트는 반응자가 둘 이상의 응답을 제공할 수 있는 질문에 대해 다중 필드를 사용하여 응답을 기록합니다. 다중 응답 세트는 범주형 필드처럼 처리되며 범주형 필드로 수행할 수 있는 대부분의 작업은 다중 응답 세트로도 수행할 수 있습니다.

다중 응답 세트는 다중 이분형 세트 또는 다중 범주 세트일 수 있습니다.

다중 이분형 세트. 다중 이분형 세트는 일반적으로 예/아니오, 유/무, 선택함/선택 안 함 등과 같이 값을 두 개만 가질 수 있는 다중 이분형 필드로 구성됩니다. 필드가 엄격하게는 이분형이 아닐 수도 있지만 변수 세트의 모든 필드가 같은 방식으로 코딩됩니다.

예를 들어, 설문조사에서는 "다음 중 주로 어디에서 뉴스를 보십니까?"라는 질문에 대해 다섯 가지의 선택 가능한 응답을 제공할 수 있습니다. 반응자는 각 선택사항 옆에 있는 선택란을 선택하여 복수의 선택사항을 표시할 수 있습니다. 다섯 가지 응답은 데이터 파일에서 다섯 개의 필드가 되며 *아니오*(선택 안 함)는 0으로, *예*(선택함)는 1로 코딩됩니다.

다중 범주 세트. 다중 범주 세트는 대개 선택 가능한 다수의 반응 범주를 포함하고 모두 같은 방법으로 코딩되는 다중 필드로 구성됩니다. 예를 들어, "여러분의 민족 전통을 가장 잘 설명하는 민족성을 세 개까지 나열해보십시오."라는 설문조사 항목이 있습니다. 수백 개의 선택 가능한 응답이 있지만 코딩을 위해 목록에는 가장 일반적인 민족성 40개만 나열하고 그 외의 나머지는 "기타" 범주로 분류합니다. 데이터 파일에서 세 개의 선택사항은 세 개의 필드가 되고 각 필드에는 41개의 범주(40개의 코딩된 민족성과 하나의 "기타" 범주)가 포함됩니다.

그래프보드 세부사항 탭

작성할 시각화 유형을 알고 있거나 시각화에 선택적 모양, 패널 및/또는 애니메이션을 추가하려는 경우에는 세부사항 탭을 사용하십시오. 예를 들어, 210 페이지의 『그래프보드 예제』의 내용을 참조하십시오.

1. 기본 탭에서 시각화 유형을 선택했으면 해당 유형이 표시됩니다. 그렇지 않은 경우에는 드롭 다운 목록에서 시각화 유형을 선택하십시오. 시각화 유형에 대한 정보는 204 페이지의 『사용 가능한 내장 그래프보드 시각화 유형』의 내용을 참조하십시오.
2. 시각화 썸네일 이미지 바로 오른쪽에는 시각화 유형에 필요한 필드(변수)를 지정하는 제어가 있습니다. 이러한 필드를 모두 지정해야 합니다.
3. 특정 시각화의 경우 요약 통계를 선택할 수 있습니다. 일부의 경우(예: 막대형 차트) 투명도 모양에 이러한 요약 옵션 중 하나를 사용할 수 있습니다. 요약 통계에 대한 설명은 198 페이지의 『그래프보드 기본 탭』의 내용을 참조하십시오.
4. 선택적 모양을 하나 이상 선택할 수 있습니다. 이 경우 시각화에 다른 필드를 포함시킬 수 있으므로 차원을 추가할 수 있습니다. 예를 들어, 필드를 사용하여 산점도에 있는 점의 *크기*에 변화를 줄 수 있습니다. 선택적 모양에 대한 자세한 정보는 194 페이지의 『모양, 오버레이, 패널 및 애니메이션』의 내용을 참조하십시오. 스크립팅을 통해서는 투명도 모양이 지원되지 않습니다.
5. 맵 시각화를 작성하는 경우 **맵 파일** 그룹은 사용할 맵 파일을 표시합니다. 기본 맵 파일이 있으면 이 파일이 표시됩니다. 맵 파일을 변경하려면 **맵 파일 선택**을 클릭하여 맵 선택 대화 상자를 표시하십시오. 이 대화 상자에서 기본 맵 파일을 지정할 수도 있습니다. 자세한 정보는 203 페이지의 『맵 시각화를 위한 맵 파일 선택』의 내용을 참조하십시오.
6. 패널링 또는 애니메이션 옵션 중 하나 이상 선택할 수 있습니다. 패널링 및 애니메이션 옵션에 대한 자세한 정보는 194 페이지의 『모양, 오버레이, 패널 및 애니메이션』의 내용을 참조하십시오.

맵 시각화를 위한 맵 파일 선택

맵 시각화 템플리트를 선택하는 경우 맵을 그리기 위한 지리적 정보를 정의하는 맵 파일이 필요합니다. 기본 맵 파일이 있으면 이 파일이 맵 시각화에 사용됩니다. 다른 맵 파일을 선택하려면 세부사항 탭에서 **맵 파일 선택**을 클릭하여 맵 선택 대화 상자를 표시하십시오.

맵 선택 대화 상자에서는 기본 맵 파일과 참조 맵 파일을 선택할 수 있습니다. 맵 파일은 맵을 그리기 위한 지리적 정보를 정의합니다. 애플리케이션은 표준 맵 파일과 함께 설치됩니다. 사용하려는 다른 ESRI 형태 파일이 있는 경우 먼저 형태 파일을 SMZ 파일로 변환해야 합니다. 자세한 정보는 224 페이지의 『맵 형태 파일 변환 및 배포』의 내용을 참조하십시오. 맵을 변환한 후에는 템플리트 선택기 대화 상자에서 **관리...**를 클릭하여 맵을 관리 시스템으로 가져오십시오. 그러면 맵 선택 대화 상자에서 해당 맵을 사용할 수 있습니다.

다음은 맵 파일을 지정할 때 고려해야 할 사항입니다.

- 모든 맵 템플리트는 하나 이상의 맵 파일이 필요합니다.
- 맵 파일은 일반적으로 맵 키 속성을 데이터 키에 연결합니다.
- 템플리트에 데이터 키에 연결하는 맵 키가 필요 없는 경우에는 참조 맵 파일과 참조 맵에 요소를 그리기 위한 좌표(예: 경도 및 위도)를 지정하는 필드가 필요합니다.
- 오버레이 맵 템플리트는 두 개의 맵, 즉 기본 맵 파일과 참조 맵 파일이 필요합니다. 참조 맵이 기본 맵 파일 뒤에 있도록 참조 맵이 먼저 그려집니다.

속성 및 지형과 같은 맵 용어에 대한 정보는 225 페이지의 『맵의 핵심 개념』의 내용을 참조하십시오.

맵 파일. 관리 시스템에 있는 어떤 맵 파일이든 선택할 수 있습니다. 여기에는 사전 설치된 맵 파일 및 가져온 맵 파일도 포함됩니다. 맵 파일 관리에 대한 자세한 정보는 223 페이지의 『템플리트, 스타일시트 및 맵 파일 관리』의 내용을 참조하십시오.

맵 키. 맵 파일을 데이터 키에 연결하는 키로 사용할 속성을 지정하십시오.

이 맵 파일 및 설정을 기본값으로 저장. 선택한 맵 파일을 기본값으로 사용하려면 이 선택란을 선택하십시오. 기본 맵 파일을 지정한 경우에는 맵 시각화를 작성할 때마다 맵 파일을 지정하지 않아도 됩니다.

데이터 키. 이 제어는 템플리트 선택기 세부사항 탭에 표시되는 것과 동일한 값을 나열합니다. 여기서는 선택하는 특정 맵 파일로 인해 키를 변경해야 하는 경우 편의를 위해 제공됩니다.

시각화에 모든 맵 지형 표시. 이 옵션을 선택하면 일치하는 데이터 키 값이 없는 경우에도 시각화에 맵의 모든 지형이 렌더링됩니다. 데이터가 있는 지형만 보려면 이 옵션을 선택 취소하십시오. **일치하지 않는 맵 키** 목록에 표시된 맵 키로 식별되는 지형이 시각화에 렌더링되지 않습니다.

맵과 데이터 값 비교. 맵 키와 데이터 키는 서로 연결되어 맵 시각화를 작성합니다. 맵 키 및 데이터 키는 동일한 도메인(예: 국가 및 지역)에서 그려야 합니다. 데이터 키 및 맵 키 값이 일치하는지 테스트하려면 **비교**를 클릭하십시오. 표시되는 아이콘은 비교 상태를 알려줍니다. 아래에 이러한 아이콘에 대한 설명이 있습니다. 비교를 수행한 후 일치하는 맵 키 값이 없는 데이터 키 값이 있으면 해당 데이터 키 값이 **일치하지 않는 데이터**

키 목록에 표시됩니다. **일치하지 않는 맵 키** 목록에는 일치하는 데이터 키 값이 없는 맵 키 값이 표시됩니다. **시각화에 모든 맵 지형 표시**를 선택하지 않은 경우, 이러한 맵 키 값으로 식별되는 지형은 렌더링되지 않습니다.

표 33. 비교 아이콘.

아이콘	설명
	비교를 수행하지 않았습니다. **비교**를 클릭하기 전의 기본 상태입니다. 데이터 키와 맵 키의 값이 일치하는지 모르므로 주의해서 진행해야 합니다.
	비교를 수행했으며 데이터 키와 맵 키의 값이 완전히 일치합니다. 데이터 키 값마다 맵 키로 식별되는 일치하는 지형이 있습니다.
	비교를 수행했으며 일부 데이터 키와 맵 키의 값이 일치하지 않습니다. 일부 데이터 키 값의 경우 맵 키로 식별되는 일치하는 지형이 없습니다. 주의해서 진행해야 합니다. 진행하는 경우 맵 시각화에 일부 데이터 값이 포함되지 않습니다.
	비교를 수행했으며 데이터 키 값과 맵 키 값이 일치하지 않습니다. 진행할 경우 맵이 렌더링되지 않으므로 다른 데이터 키 또는 맵 키를 선택해야 합니다.

사용 가능한 내장 그래프보드 시각화 유형

다양한 여러 시각화 유형을 작성할 수 있습니다. 기본 및 세부사항 탭에서 다음에 나열된 내장 유형을 모두 사용할 수 있습니다. 템플리트(특히 맵 템플리트)에 대한 일부 설명은 **특수 텍스트**를 사용하여 세부사항 탭에 지정된 필드(변수)를 식별합니다.

표 34. 사용 가능한 그래프 유형.

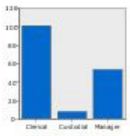
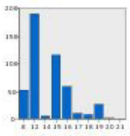
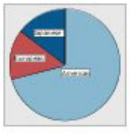
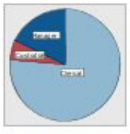
차트 아이콘	설명	차트 아이콘	설명
	**막대** 연속형 숫자 필드에 대한 요약 통계를 계산하고 범주형 필드의 각 범주에 대한 결과를 막대로 표시합니다. *필수*: 범주형 필드 및 연속형 필드.		**개수 막대형 차트** 범주형 필드의 각 범주에 있는 행/케이스의 비율을 막대로 표시합니다. 분포 그래프 노드를 사용하여 이 그래프를 생성할 수도 있습니다. 이 노드는 일부 추가 옵션을 제공합니다. 자세한 정보는 244 페이지의 『분포 노드』 주제를 참조하십시오. *필수*: 하나의 범주형 필드.
	**원** 연속형 숫자 필드의 합계를 계산하고 범주형 필드의 각 범주에 분포된 연속형 숫자 필드 합의 비율을 원의 조각으로 표시합니다. *필수*: 범주형 필드 및 연속형 필드.		**개수 원형 차트** 범주형 필드의 각 범주에 있는 행/케이스의 비율을 원의 조각으로 표시합니다. *필수*: 하나의 범주형 필드.

표 34. 사용 가능한 그래프 유형 (계속).

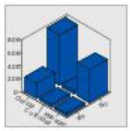
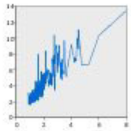
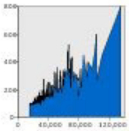
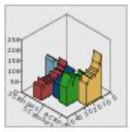
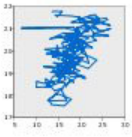
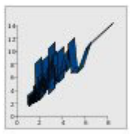
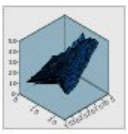
차트 아이콘	설명	차트 아이콘	설명
	**3차원 막대형 차트** 연속형 숫자 필드에 대한 요약 통계를 계산하고 두 범주형 필드의 범주가 교차하는 지점의 결과를 표시합니다. *필수*: 범주형 필드 쌍 및 연속형 필드.		**3차원 원형 차트** 이 차트는 추가된 3차원 효과를 제외하면 원형 차트와 동일합니다. *필수*: 범주형 필드 및 연속형 필드.
	**선** 한 필드의 각 값에 대한 다른 필드의 요약 통계를 계산하고 값을 연결하는 선을 그립니다. 도표 그래프 노드를 사용하여 선 도표 그래프를 생성할 수도 있습니다. 이 노드는 일부 추가 옵션을 제공합니다. 자세한 정보는 232 페이지의 『구성 노드』 주제를 참조하십시오. *필수*: 임의 유형의 필드 쌍.		**영역** 한 필드의 각 값에 대한 다른 필드의 요약 통계를 계산하고 값을 연결하는 영역을 그립니다. 영역은 아래 공간이 채색된 선과 유사하므로 선과 영역 간의 차이는 매우 작습니다. 그러나 색상 모양을 사용할 경우 선이 단순하게 분할되고 영역이 누적됩니다. *필수*: 임의 유형의 필드 쌍.
	**3차원 영역** 한 필드의 값에 대해 구성되고 범주형 필드에 의해 분할된 다른 필드의 값을 표시합니다. 범주마다 영역 요소가 그려집니다. *필수*: 범주형 필드 및 임의 유형의 필드 쌍.		**경로** 한 필드의 값에 대해 구성된 다른 필드의 값을 표시하고 이러한 값을 원래 데이터 세트에 표시된 순서대로 하나의 선으로 연결합니다. 순서 지정이 경로와 선의 주된 차이점입니다. *필수*: 임의 유형의 필드 쌍.
	**리본도표** 한 필드의 각 값에 대한 다른 필드의 요약 통계를 계산하고 값을 연결하는 리본을 그립니다. 리본은 본질적으로 3차원 효과를 갖는 선입니다. 진정한 3차원 그래프는 아닙니다. *필수*: 임의 유형의 필드 쌍.		**표면** 서로의 값에 대해 구성된 세 필드의 값을 표시하고 이러한 값을 하나의 표면으로 연결합니다. *필수*: 임의 유형의 세 개의 필드.

표 34. 사용 가능한 그래프 유형 (계속).

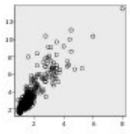
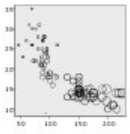
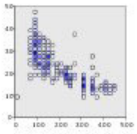
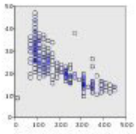
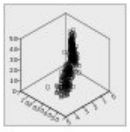
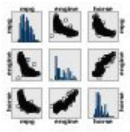
차트 아이콘	설명	차트 아이콘	설명
	**산점도** 한 필드의 값에 대해 구성된 다른 필드의 값을 표시합니다. 이 그래프는 필드 사이의 관계를 강조표시할 수 있습니다(관계가 있는 경우). 도표 그래프 노드를 사용하여 산점도를 생성할 수도 있습니다. 이 노드는 일부 추가 옵션을 제공합니다. 자세한 정보는 232 페이지의 『구성 노드』 주제를 참조하십시오. *필수*: 임의 유형의 필드 쌍.		**거품 도표** 기본 산점도와 마찬가지로 한 필드의 값에 대해 구성된 다른 필드의 값을 표시합니다. 차이점은 세 번째 필드의 값이 개별 점의 *크기*에 변화를 주는 데 사용된다는 점입니다. *필수*: 임의 유형의 세 개의 필드.
	**구간화된 산점도** 기본 산점도와 마찬가지로 한 필드의 값에 대해 구성된 다른 필드의 값을 표시합니다. 차이점은 유사한 값이 그룹으로 구간화되고 색상 또는 *크기* 모양이 각 구간의 케이스 수를 표시하는 데 사용된다는 점입니다. *필수* : 연속형 필드 쌍.		**육각형 구간화된 산점도** 구간화된 산점도 설명을 참조하십시오. 차이점은 기본 구간의 형태이며 구간이 원이 아니라 육각형과 비슷합니다. 결과로 생성되는 육각형 구간화된 산점도는 구간화된 산점도와 유사합니다. 그러나 기본 구간의 형태가 다르기 때문에 두 그래프 간에 각 구간의 값 수가 다릅니다. *필수* : 연속형 필드 쌍.
	**3차원 산점도** 서로에 대해 구성된 세 필드의 값을 표시합니다. 이 그래프는 필드 사이의 관계를 강조표시할 수 있습니다(관계가 있는 경우). 도표 그래프 노드를 사용하여 3차원 산점도를 생성할 수도 있습니다. 이 노드는 일부 추가 옵션을 제공합니다. 자세한 정보는 232 페이지의 『구성 노드』 주제를 참조하십시오. *필수*: 임의 유형의 세 개의 필드.		**산점도 행렬(SPLOM)** 필드마다 한 필드의 값에 대해 구성된 다른 필드의 값을 표시합니다. SPLOM은 산점도 테이블과 유사합니다. SPLOM에도 각 필드의 히스토그램이 포함됩니다. *필수* : 두 개 이상의 연속형 필드.

표 34. 사용 가능한 그래프 유형 (계속).

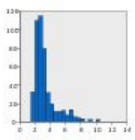
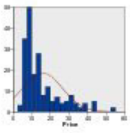
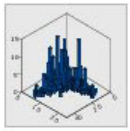
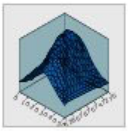
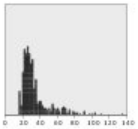
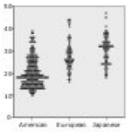
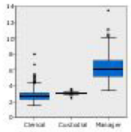
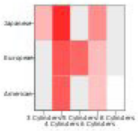
차트 아이콘	설명	차트 아이콘	설명
	**히스토그램** 필드의 빈도 분포를 표시합니다. 히스토그램은 분포 유형을 판별하고 분포가 비대칭인지 여부를 확인하는 데 도움이 될 수 있습니다. 히스토그램 그래프 노드를 사용하여 이 그래프를 생성할 수도 있습니다. 이 노드는 일부 추가 옵션을 제공합니다. 자세한 정보는 248 페이지의 『히스토그램 도표 탭』 주제를 참조하십시오. *필수*: 임의 유형의 하나의 필드.		**정규 분포 히스토그램** 정규 분포 곡선이 겹쳐진 연속형 필드의 빈도 분포를 표시합니다. *필수*: 하나의 연속성 필드.
	**3차원 히스토그램** 연속형 필드 쌍의 빈도 분포를 표시합니다. *필수* : 연속형 필드 쌍.		**3차원 밀도** 연속형 필드 쌍의 빈도 분포를 표시합니다. 3차원 히스토그램과 유사하며 유일한 차이점은 분포를 표시하는 데 막대 대신 표면이 사용된다는 점입니다. *필수* : 연속형 필드 쌍.
	**점도표** 개별 케이스/행을 표시하고 *x*축의 고유 데이터 점에서 케이스/행을 누적시킵니다. 이 그래프는 데이터의 분포를 표시하는 점에서 히스토그램과 유사하지만 특정 구간(값 범위)의 집계된 개수가 아니라 각 케이스/행을 표시합니다. *필수*: 임의 유형의 하나의 필드.		**2차원 점도표** 범주형 필드의 범주마다 개별 케이스/행을 표시하고 *y*축의 고유 데이터 점에서 케이스/행을 누적시킵니다. *필수*: 범주형 필드 및 연속형 필드.
	**상자도표** 범주형 필드의 범주마다 연속형 필드의 5가지 통계(최소값, 첫 번째 사분위수, 중앙값, 세 번째 사분위수 및 최대값)를 계산합니다. 결과가 상자도표/스키마 요소로 표시됩니다. 상자도표를 통해 범주 간에 연속형 데이터의 분포가 얼마나 다른지 알 수 있습니다. *필수*: 범주형 필드 및 연속형 필드.		**히트 맵** 두 범주형 필드 사이에서 범주가 교차하는 지점의 연속형 필드 평균을 계산합니다. *필수*: 범주형 필드 쌍 및 연속형 필드.

표 34. 사용 가능한 그래프 유형 (계속).

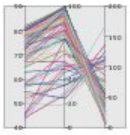
차트 아이콘	설명	차트 아이콘	설명
	**동형** 각 필드에 대한 평행 축을 만들고 데이터의 케이스/행마다 필드 값을 지나는 선을 그립니다. *필수* : 두 개 이상의 연속형 필드.		**개수의 코로플레스** 범주형 필드(**데이터 키**)의 각 범주에 대한 개수를 계산하고 범주에 해당하는 맵 지형에서 색포화도를 사용하여 개수를 나타내는 맵을 그립니다. *필수* : 범주형 필드. 해당 키가 **데이터 키** 범주와 일치하는 맵 파일.
	**평균/중앙값/합계의 코로플레스** 범주형 필드(**데이터 키**)의 각 범주에 대한 연속형 필드(**색상**)의 평균, 중앙값 또는 합계를 계산하고 범주에 해당하는 맵 지형에서 색포화도를 사용하여 계산된 통계를 나타내는 맵을 그립니다. *필수*: 범주형 필드 및 연속형 필드. 해당 키가 **데이터 키** 범주와 일치하는 맵 파일.		**값의 코로플레스** 하나의 범주형 필드(**데이터 키**)로 정의된 값에 해당하는 맵 지형에 대한 다른 범주형 필드(**색상**)의 값을 색상을 사용하여 나타내는 맵을 그립니다. 각 지형에 대한 색상 필드의 범주형 값이 여러 개인 경우 모달 값이 사용됩니다. *필수*: 범주형 필드 쌍. 해당 키가 **데이터 키** 범주와 일치하는 맵 파일.
	**개수의 코로플레스 위의 좌표** 코로플레스 맵에 점을 그리기 위한 좌표를 식별하는 두 개의 추가적 연속형 필드(**경도** 및 **위도**)가 있다는 점을 제외하곤 개수의 코로플레스와 유사합니다. *필수*: 범주형 필드와 연속형 필드의 쌍. 해당 키가 **데이터 키** 범주와 일치하는 맵 파일.		**평균/중앙값/합계의 코로플레스 위의 좌표** 코로플레스 맵에 점을 그리기 위한 좌표를 식별하는 두 개의 추가적 연속형 필드(**경도** 및 **위도**)가 있다는 점을 제외하곤 평균/중앙값/합계의 코로플레스와 유사합니다. *필수* : 범주형 필드 및 세 개의 연속형 필드. 해당 키가 **데이터 키** 범주와 일치하는 맵 파일.
	**값의 코로플레스 위의 좌표** 코로플레스 맵에 점을 그리기 위한 좌표를 식별하는 두 개의 추가적 연속형 필드(**경도** 및 **위도**)가 있다는 점을 제외하곤 값의 코로플레스와 유사합니다. *필수*: 범주형 필드 쌍과 연속형 필드 쌍. 해당 키가 **데이터 키** 범주와 일치하는 맵 파일.		**맵 위의 개수 막대형 차트** 각 맵 지형(**데이터 키**)에 대해 범주형 필드(**범주**)의 각 범주에 있는 행/케이스의 비율을 계산하여 맵과 함께 각 맵 지형의 중앙에 막대형 차트를 그립니다. *필수*: 범주형 필드 쌍. 해당 키가 **데이터 키** 범주와 일치하는 맵 파일.

표 34. 사용 가능한 그래프 유형 (계속).

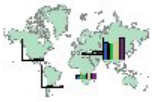
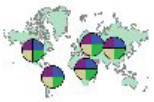
차트 아이콘	설명	차트 아이콘	설명
	**맵 위의 막대형 차트** 연속형 필드(**값**)의 요약 통계를 계산하고 각 맵 지형(**데이터 키**)에 대한 범주형 필드(**범주**)의 각 범주 결과를 각 맵 지형의 중앙에 위치한 막대형 차트로 표시합니다. *필수*: 범주형 필드 쌍 및 연속형 필드. 해당 키가 **데이터 키** 범주와 일치하는 맵 파일.		**맵 위의 개수 원형 차트** 각 맵 지형(**데이터 키**)에 대해 범주형 필드(**범주**)의 각 범주에 있는 행/케이스의 비율을 표시하고 맵과 함께 각 맵 지형의 중앙에 비율을 원형 차트의 조각으로 그립니다. *필수*: 범주형 필드 쌍. 해당 키가 **데이터 키** 범주와 일치하는 맵 파일.
	**맵 위의 원형 차트** 각 맵 지형(**데이터 키**)에 대해 범주형 필드(**범주**)의 각 범주에 있는 연속형 필드(**값**)의 합계를 계산하고 맵과 함께 각 맵 지형의 중앙에 합계를 원형 차트의 조각으로 그립니다. *필수*: 범주형 필드 쌍 및 연속형 필드. 해당 키가 **데이터 키** 범주와 일치하는 맵 파일.		**맵 위의 선형 차트** 각 맵 지형(**데이터 키**)에 대해 한 필드(**X**)의 각 값에 대한 다른 연속형 필드(**Y**)의 요약 통계를 계산하고 맵과 함께 각 맵 지형의 중앙에 값을 연결하는 선형 차트를 그립니다. *필수*: 범주형 필드 및 임의 유형의 필드 쌍. 해당 키가 **데이터 키** 범주와 일치하는 맵 파일.
	**참조 맵 위의 좌표** 점의 좌표를 식별하는 연속형 필드(**경고** 및 **위도**)를 사용하여 맵과 점을 그립니다. *필수*: 범위 필드 쌍. 맵 파일.		**참조 맵 위의 화살표** 각 화살표의 시작점(**시작 경도** 및 **시작 위도**)과 종료점(**종료 경도** 및 **종료 위도**)을 식별하는 연속형 필드를 사용하여 맵과 화살표를 그립니다. 데이터의 각 레코드/케이스가 맵에서 하나의 화살표를 생성합니다. *필수* : 네 개의 연속형 필드. 맵 파일.
	**점 오버레이 맵** 참조 맵을 그리고 그 위에 점 지형이 범주형 필드(**색상**)로 채색된 다른 점 맵을 겹칩니다. *필수*: 범주형 필드 쌍. 해당 키가 **데이터 키** 범주와 일치하는 점 맵 파일. 참조 맵 파일.		**다각형 오버레이 맵** 참조 맵을 그리고 그 위에 다각형 지형이 범주형 필드(**색상**)로 채색된 다른 다각형 맵을 겹칩니다. *필수*: 범주형 필드 쌍. 해당 키가 **데이터 키** 범주와 일치하는 다각형 맵 파일. 참조 맵 파일.

표 34. 사용 가능한 그래프 유형 (계속).

차트 아이콘	설명	차트 아이콘	설명
	**선 오버레이 맵** 참조 맵을 그리고 그 위에 선 지형이 범주형 필드(**색상**)로 채색된 다른 선 맵을 겹칩니다. *필수*: 범주형 필드 쌍. 해당 키가 **데이터 키** 범주와 일치하는 선 맵 파일. 참조 맵 파일.		

맵 시각화 작성

다수의 시각화는 관심 필드(변수)와 이러한 필드를 시각화할 템플리트 이 두 가지만 선택하면 됩니다. 추가적 선택이나 조치가 필요하지 않습니다. 그러나 맵 시각화를 작성하려면 최소한 하나의 추가 단계가 필요합니다. 즉, 맵 시각화를 위한 지리적 정보를 정의하는 맵 파일을 선택해야 합니다.

단순한 맵을 작성하는 기본 단계는 다음과 같습니다.

1. 기본 탭에서 관심 필드를 선택하십시오. 다양한 맵 시각화에 필요한 필드 유형 및 수에 대한 정보는 204 페이지의 『사용 가능한 내장 그래프보드 시각화 유형』의 내용을 참조하십시오.
2. 맵 템플리트를 선택하십시오.
3. 세부사항 탭을 클릭하십시오.
4. **데이터 키** 및 기타 필요한 드롭 다운 목록이 올바른 필드로 설정되었는지 확인하십시오.
5. 맵 파일 그룹에서 **맵 파일 선택**을 클릭하십시오.
6. 맵 선택 대화 상자를 사용하여 맵 파일 및 맵 키를 선택하십시오. 맵 키의 값은 **데이터 키**로 지정한 필드의 값과 일치해야 합니다. **비교** 단추를 사용하여 이러한 값을 비교할 수 있습니다. 오버레이 맵 템플리트를 선택하는 경우에는 참조 맵도 선택해야 합니다. 참조 맵은 데이터에 맞추어져 있지 않습니다. 참조 맵은 주요 맵의 배경으로 사용됩니다. 맵 선택 대화 상자에 대한 자세한 정보는 203 페이지의 『맵 시각화를 위한 맵 파일 선택』의 내용을 참조하십시오.
7. **확인**을 클릭하여 맵 선택 대화 상자를 닫으십시오.
8. 그래프보드 템플리트 선택기에서 **실행**을 클릭하여 맵 시각화를 작성하십시오.

그래프보드 예제

이 절에는 사용 가능한 옵션을 설명하는 서로 다른 여러 예제가 있습니다. 이러한 예제에서는 또한 시각화 결과물에 대한 해석 정보를 제공합니다.

이 예제에서는 *graphboard.str*이라는 스트림을 사용하고 이 스트림은 *employee_data.sav*, *customer_subset.sav* 및 *worldsales.sav*라는 데이터 파일을 참조합니다. 이러한 파일은 IBM SPSS Modeler 클라이언트 설치의 *데모* 폴더에 있습니다. Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 데모 폴더에 액세스할 수 있습니다. *graphboard.str* 파일은 *스트림* 폴더에 있습니다.

표시된 순서대로 예제를 읽는 것이 좋습니다. 이어지는 예제는 이전 예제를 기반으로 작성됩니다.

예제: 요약 통계가 포함된 막대형 차트

세트/범주형 변수의 각 범주에 대해 연속형 숫자 필드/변수를 요약한 막대형 차트를 작성합니다. 구체적으로 남여 평균 급여를 보여주는 막대형 차트를 작성합니다.

이 예제와 다음 예제 중 일부에서는 회사의 직원에 대한 정보가 포함된 가설 데이터 세트인 *직원 데이터*를 사용합니다.

1. *employee_data.sav*를 가리키는 통계 파일 소스 노드를 추가하십시오.
2. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
3. 기본 탭에서 *성별* 및 *현재 급여* 를 선택하십시오. (여러 필드/변수를 선택하려면 Ctrl+클릭을 사용하십시오.)
4. **막대**를 선택하십시오.
5. 요약 드롭 다운 목록에서 **평균**을 선택하십시오.
6. **실행**을 클릭하십시오.
7. 결과로 표시되는 화면에서 "필드 및 값 레이블 표시" 도구 모음 단추(도구 모음 가운데 있는 두 개의 단추 중 두 번째)를 클릭하십시오.

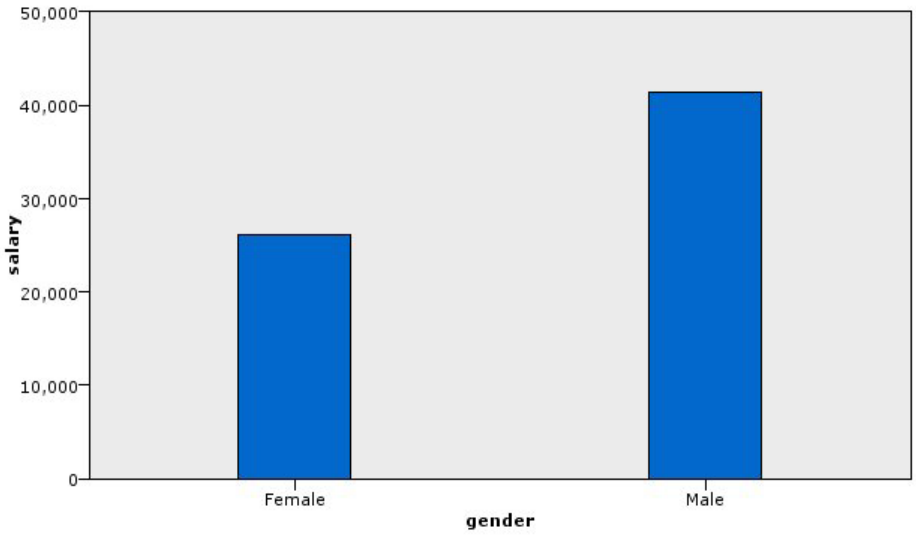


그림 8. 요약 통계가 포함된 막대형 차트

다음을 관측할 수 있습니다.

- 막대의 높이를 볼 때 남자의 평균 급여가 여자의 평균 급여보다 높습니다.

예제: 요약 통계가 포함된 누적 막대형 차트

이제 누적 막대형 차트를 만들어 남녀 평균 급여의 차이가 직업 유형과 관련이 있는지 확인할 수 있습니다. 특정 직업 유형에서는 여자의 평균 급여가 남자보다 높을 수 있습니다.

참고: 이 예제에서는 *직원 데이터*를 사용합니다.

1. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
2. 기본 탭에서 *직원 범주* 및 *현재 급여*를 선택하십시오. (여러 필드/변수를 선택하려면 Ctrl+클릭을 사용하십시오.)
3. **막대**를 선택하십시오.
4. 요약 목록에서 **평균**을 선택하십시오.
5. 세부사항 탭을 클릭하십시오. 이전 탭에서의 선택이 여기에 반영됩니다.
6. 선택적 모양 그룹의 색상 드롭 다운 목록에서 *성별*을 선택하십시오.
7. **실행**을 클릭하십시오.

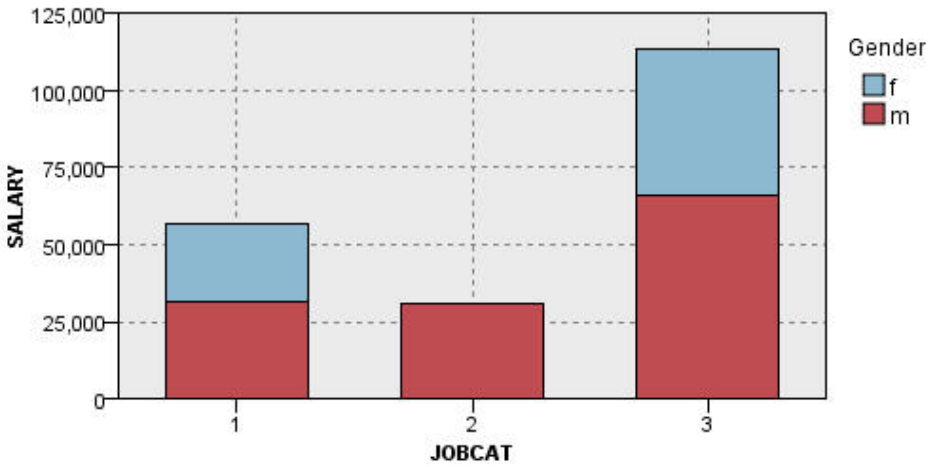


그림 9. 누적 막대형 차트

다음을 관측할 수 있습니다.

- 각 직업 유형에 따른 평균 급여 차이는 모든 남녀의 평균 급여를 비교한 막대형 차트만큼 커 보이지 않습니다. 그룹에 따라 남녀 수가 다를 수 있습니다. 개수 막대형 차트를 만들어 이를 확인할 수 있습니다.
- 직업 유형에 관계없이 남자의 평균 급여가 항상 여자의 평균 급여보다 높습니다.

예제: 패널링된 히스토그램

남녀 급여의 빈도 분포를 비교할 수 있도록 성별로 패널링된 히스토그램을 만듭니다. 빈도 분포는 특정 급여 범위 내에 얼마나 많은 케이스/행이 포함되는지 보여줍니다. 패널링된 히스토그램을 사용하여 성별에 따른 급여 차이를 더 자세히 분석할 수 있습니다.

참고: 이 예제는 *직원 데이터*를 사용합니다.

1. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
2. 기본 탭에서 *현재 급여* 를 선택하십시오.
3. **히스토그램**을 선택하십시오.

4. 세부사항 탭을 클릭하십시오.
5. 패널 및 애니메이션 그룹의 패널 전체 드롭 다운 목록에서 *성별*을 선택하십시오.
6. **실행**을 클릭하십시오.

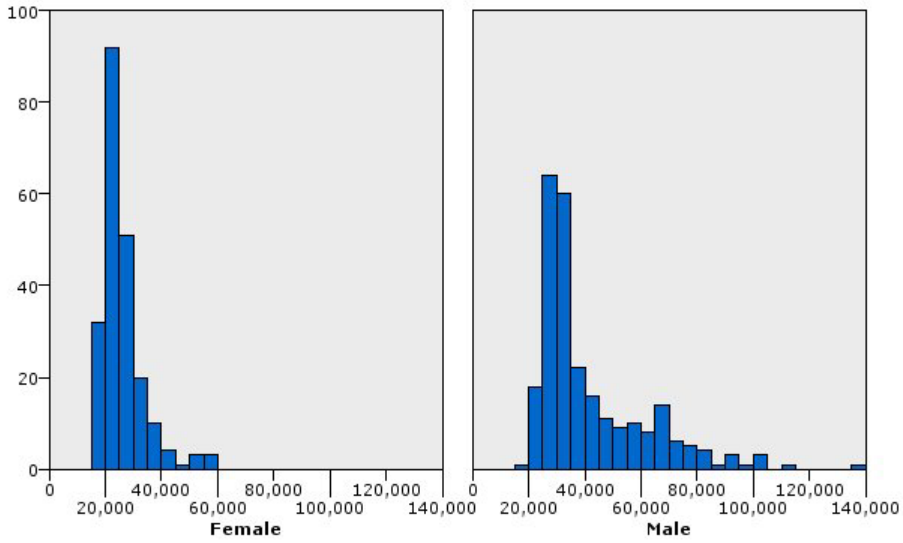


그림 10. 패널링된 히스토그램

다음을 관측할 수 있습니다.

- 두 빈도 분포 모두 정규 분포가 아닙니다. 즉, 이 두 히스토그램은 데이터가 정규 분포일 때 보이는 종 곡선과 유사하지 않습니다.
- 더 높은 막대가 각 그래프의 왼쪽에 있습니다. 따라서 남녀 모두 더 높은 급여가 아니라 더 낮은 급여를 더 작성해야 합니다.
- 남자와 여자의 급여 빈도 분포가 서로 같지 않습니다. 히스토그램의 모양에 주의하십시오. 높은 급여를 받는 사람은 남자가 여자보다 많습니다.

예제: 패널링된 점도표

히스토그램처럼 점도표는 연속형 숫자 범위의 분포를 보여줍니다. 구간화된 데이터 범위에 대한 개수를 보여주는 히스토그램과는 달리 점도표는 데이터에 있는 모든 행/케이스를 보여줍니다. 따라서 점도표는 히스토그램에 비해 높은 세분성(granularity)을 제공합니다. 실제로 빈도 분포를 분석할 때 시작점으로 점도표를 더 선호할 수 있습니다.

참고: 이 예제는 *직원 데이터*를 사용합니다.

1. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
2. 기본 탭에서 *현재 급여* 를 선택하십시오.
3. **점도표**를 선택하십시오.
4. 세부사항 탭을 클릭하십시오.

5. 패널 및 애니메이션 그룹의 패널 전체 드롭 다운 목록에서 *성별*을 선택하십시오.
6. **실행**을 클릭하십시오.
7. 결과로 표시되는 출력 창을 최대화하여 도표를 더욱 분명하게 볼 수 있습니다.

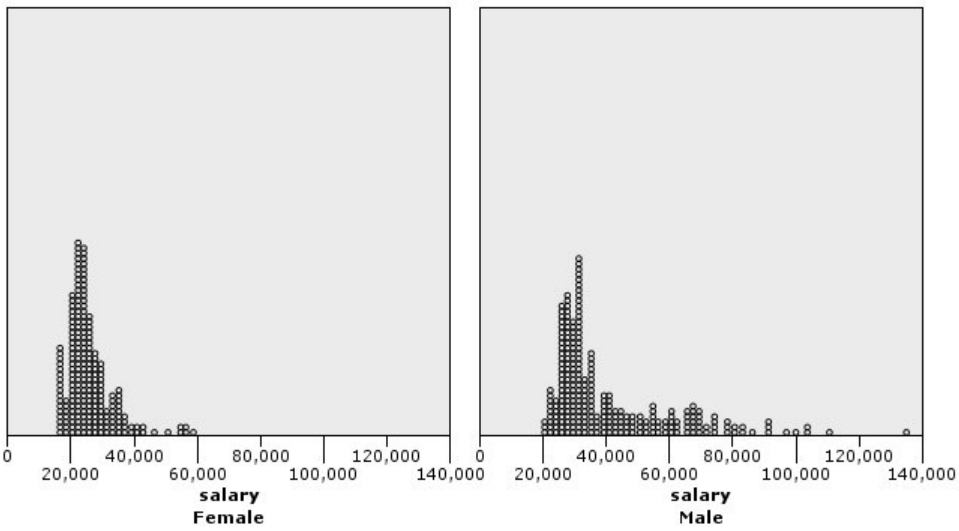


그림 11. 패널링된 점도표

히스토그램(212 페이지의 『예제: 패널링된 히스토그램』 참조)과 비교하여 다음을 관측할 수 있습니다.

- 여자 히스토그램에 표시된 피크 20,000이 점도표에서는 그만큼 급격한 증가로 표시되지 않습니다. 다수의 케이스/행이 그 값 주위에 집중되어 있지만 대부분의 값은 25,000에 더 가깝습니다. 이러한 단위 수준은 히스토그램에서 표시되지 않습니다.
- 남자 히스토그램에서는 남자 평균 급여가 40,000에 이르러 점차 하강하지만 점도표에서는 이 값 이후부터 80,000까지 꽤 일정한 분포를 보여줍니다. 해당 범위 내의 특정 급여 값에서 세 명 이상의 남자가 특별한 급여를 받고 있습니다.

예제: 상자도표

상자도표는 데이터의 분포 상태를 표시하는 또하나의 유용한 시각화입니다. 상자도표에는 시각화 작성 후 탐색하는 여러 통계 측도가 포함됩니다.

참고: 이 예제는 *직원 데이터*를 사용합니다.

1. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
2. 기본 탭에서 *성별* 및 *현재 급여* 를 선택하십시오. (여러 필드/변수를 선택하려면 Ctrl+클릭을 사용하십시오.)
3. **상자도표**를 선택하십시오.
4. **실행**을 클릭하십시오.

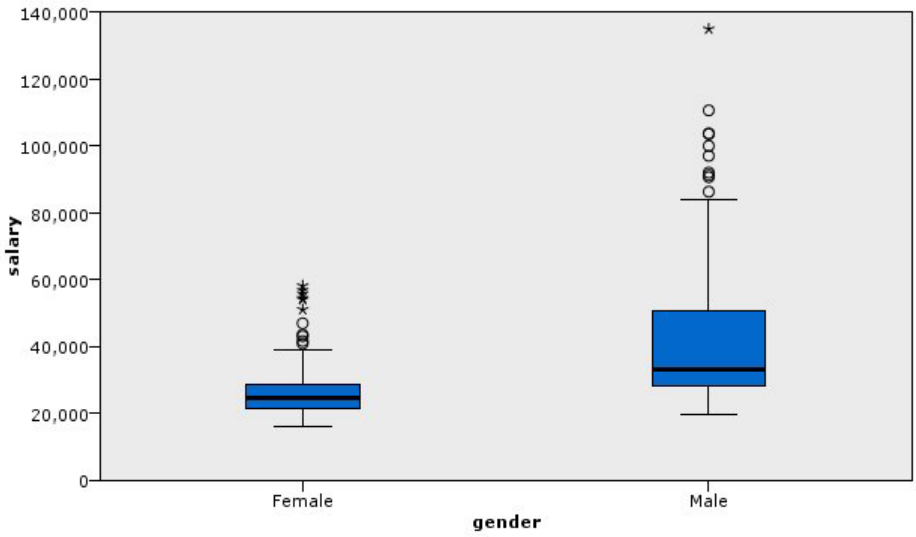


그림 12. 상자도표

다음은 상자도표의 다양한 부분에 대한 설명입니다.

- 상자 가운데 있는 진한 선은 *급여*의 중앙값입니다. 케이스/행의 반은 중앙값보다 큰 값을 가지고 나머지 반은 낮은 값을 가집니다. 평균과 마찬가지로 중앙값은 중심 경향의 측도입니다. 평균과 달리 중앙값은 극단값을 갖는 케이스/행에 덜 영향을 받습니다. 이 예제에서는 중앙값이 평균보다 낮습니다(211 페이지의 『예제: 요약 통계가 포함된 막대형 차트』와 비교할 때). 평균과 중앙값의 차이는 일부 케이스/행이 평균을 높이는 극단값을 가짐을 의미합니다. 즉, 일부 직원이 많은 급여를 받습니다.
- 상자의 맨 아래는 25번째 백분위수를 표시합니다. 케이스/행의 25%는 25번째 백분위수보다 낮은 값을 가집니다. 상자의 맨 위는 75번째 백분위수를 표시합니다. 케이스/행의 25%는 75번째 백분위수보다 높은 값을 가집니다. 이는 케이스/행의 50%가 상자 내에 있음을 의미합니다. 여자의 상자가 남자의 상자보다 훨씬 짧습니다. 이로써 *급여* 범위가 남자보다 여자가 더 작다는 결론이 도출됩니다. 상자의 맨 위와 맨 아래를 종종 **힌지**라고 합니다.
- 상자에서 확장된 T형 막대는 **내부 펜스** 또는 **수염 도표**라고 합니다. T형 막대는 상자 높이의 1.5배까지 확장되거나, 그러한 범위 내의 값을 가진 케이스/행이 없는 경우, 최소값 또는 최대값까지 확장됩니다. 데이터가 정규 분포되어 있는 경우 데이터의 약 95%가 내부 펜스 사이에 있을 것으로 예상됩니다. 이 예제에서는 여자의 내부 펜스가 남자보다 더 적게 확장됩니다. 이는 *급여* 범위가 남자보다 여자가 더 작다는 의미도 내포하고 있습니다.
- 점은 **이상값**입니다. 이상값은 내부 펜스에 해당하지 않는 값으로 정의됩니다. 이상값은 극단값입니다. 별표는 **극단적인 이상값**입니다. 극단적인 이상값은 상자 높이의 세 배보다 많은 값을 갖는 케이스/행을 나타냅니다. 남녀 모두 여러 개의 이상값이 있습니다. 평균이 중앙값보다 더 큽니다. 평균이 더 큰 이유는 이러한 이상값 때문입니다.

예제: 원형 차트

이제 다른 데이터 세트를 사용하여 일부 다른 시각화 유형을 탐색합니다. 데이터 세트는 고객에 대한 정보가 포함된 가설 데이터 파일인 *customer_subset*입니다.

먼저 원형 차트를 만들어 서로 다른 지역의 고객 비율을 확인합니다.

1. *customer_subset.sav*를 가리키는 통계 파일 소스 노드를 추가하십시오.
2. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
3. 기본 탭에서 *지리 표시기*를 선택하십시오.
4. **개수 원형 차트**를 선택하십시오.
5. **실행**을 클릭하십시오.

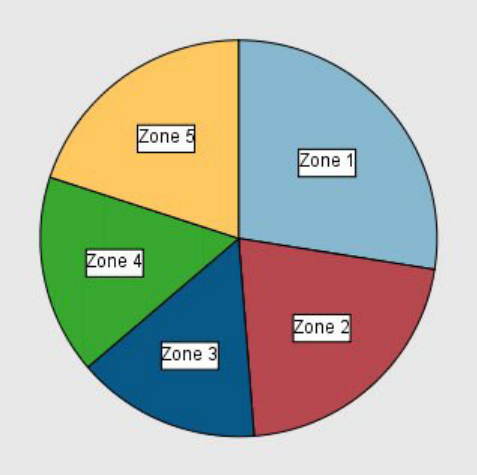


그림 13. 원형 차트

다음을 관측할 수 있습니다.

- 다른 지역보다 Zone 1에 더 많은 고객이 있습니다.
- 나머지 지역에서는 고객이 균등하게 분포되어 있습니다.

예제: 히트 맵

이제 서로 다른 지역 및 연령 그룹에 속하는 고객의 평균 소득을 확인하기 위한 범주형 히트 맵을 만듭니다.

참고: 이 예제에서는 *customer_subset*를 사용합니다.

1. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
2. 기본 탭에서 *지형 표시기*, *연령 범주* 및 *가구소득(천단위)* 순으로 선택하십시오. (여러 필드/변수를 선택하려면 Ctrl+클릭을 사용하십시오.)
3. **히트 맵**을 선택하십시오.
4. **실행**을 클릭하십시오.
5. 결과로 표시되는 출력 화면에서 "필드 및 값 레이블 표시" 도구 모음 단추(도구 모음 가운데 있는 두 개 중 오른쪽 단추)를 클릭하십시오.

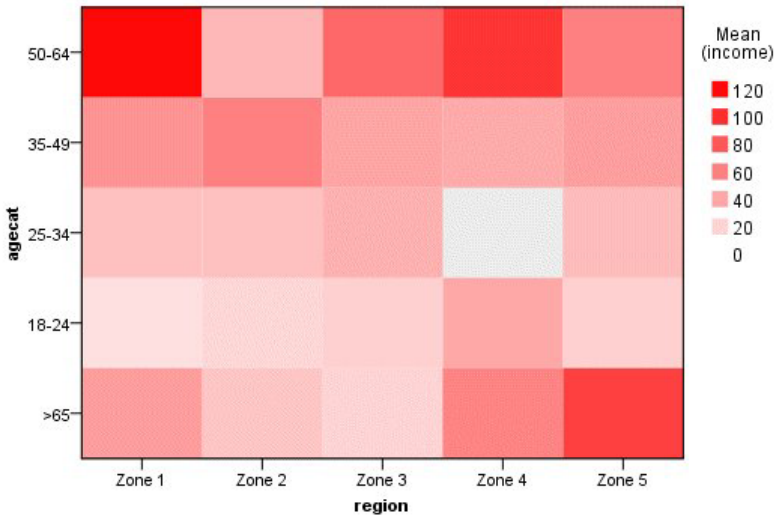


그림 14. 범주형 히트 맵

다음을 관측할 수 있습니다.

- 히트 맵은 숫자 대신 색상을 사용하여 셀 값을 표시하는 테이블과 비슷합니다. 밝고 짙은 빨강은 가장 높은 값을 표시하고 회색은 낮은 값을 표시합니다. 각 셀의 값은 각 범주 쌍에 대한 연속형 필드/변수의 평균입니다.
- Zone 2 및 Zone 5를 제외하고, 연령이 50세에서 64세 사이인 고객 그룹이 다른 그룹보다 평균 가구소득이 더 많습니다.
- Zone 4에는 25세에서 34세 사이의 고객이 없습니다.

예제: 산점도 행렬(SPLOM)

서로 다른 여러 변수에 대한 산점도 행렬을 작성하여 데이터 세트의 변수 간에 관계가 있는지 판별합니다.

참고: 이 예제에서는 *customer_subset*를 사용합니다.

1. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
2. 기본 탭에서 *연령*, *가구소득(천단위)* 및 *신용카드 대출(천단위)* 을 선택하십시오. (여러 필드/변수를 선택하려면 Ctrl+클릭을 사용하십시오.)
3. **SPLOM**을 선택하십시오.
4. **실행**을 클릭하십시오.
5. 출력 창을 최대화하여 행렬을 더욱 분명하게 볼 수 있습니다.

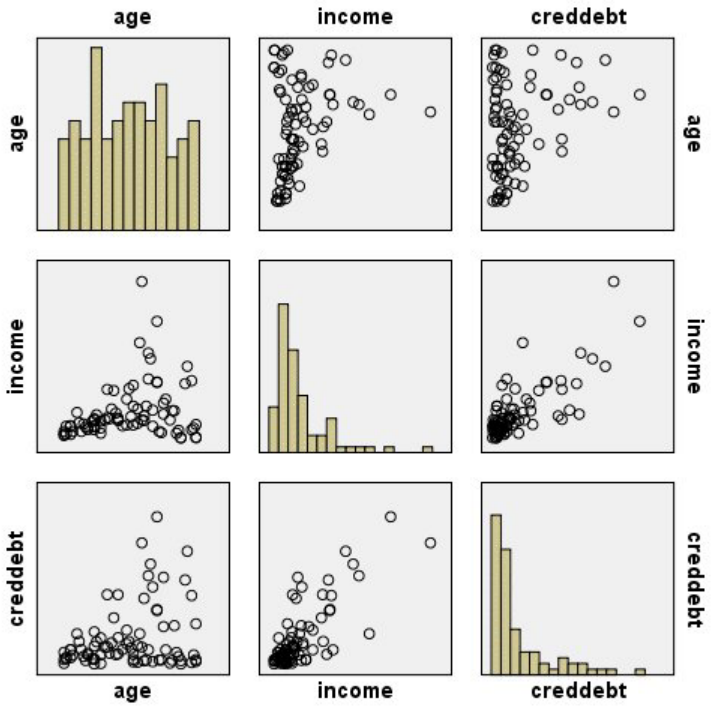


그림 15. 산점도 행렬(SPLOM)

다음을 관측할 수 있습니다.

- 대각선으로 표시되는 히스토그램은 SPLOM에서 각 변수의 분포를 보여줍니다. *연령*에 대한 히스토그램은 상단 왼쪽 셀에 표시되고 *소득*에 대한 히스토그램은 중앙 셀에 표시되며 *신용대출*에 대한 히스토그램은 하단 오른쪽 셀에 표시됩니다. 정규 분포로 표시되는 변수가 없습니다. 즉, 종 곡선과 유사한 히스토그램이 없습니다. 또한 *소득* 및 *신용대출*에 대한 히스토그램은 정적으로 비대칭됩니다.
- *연령*과 다른 변수 사이에 아무런 관계가 없어 보입니다.
- *소득*과 *신용대출* 사이에는 선형 관계가 있습니다. 즉, *소득*이 증가할수록 *신용대출*이 증가합니다. 이러한 변수와 다른 관련 변수에 대한 개별 산점도를 만들어 관계를 더욱 자세히 탐색할 수 있습니다.

예제: 합계의 코로플레스(색상 맵)

이제 맵 시각화를 작성합니다. 그런 후에 다음 예제에서 이 시각화의 변형을 작성할 것입니다. 데이터 세트는 대륙 및 제품별 판매 수입이 포함된 가설 데이터 파일인 *worldsales*입니다.

1. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
2. 기본 탭에서 *대륙* 및 *수입*을 선택하십시오. (여러 필드/변수를 선택하려면 Ctrl+클릭을 사용하십시오.)
3. **합계의 코로플레스**를 선택하십시오.
4. 세부사항 탭을 클릭하십시오.
5. 선택적 모양 그룹의 데이터 레이블 드롭 다운 목록에서 *대륙*을 선택하십시오.
6. 맵 파일 그룹에서 **맵 파일 선택**을 클릭하십시오.
7. 맵 선택 대화 상자에서 **맵**이 *대륙*으로 설정되고 **맵 키**가 *CONTINENT*로 설정되었는지 확인하십시오.

8. 맵과 데이터 값 비교 그룹에서 **비교**를 클릭하여 맵 키가 데이터 키와 일치하는지 확인하십시오. 이 예제에서는 모든 데이터 키 값이 맵 키 및 지형과 일치합니다. 오세아니아에 대한 데이터가 없음도 알 수 있습니다.
9. 맵 선택 대화 상자에서 **확인**을 클릭하십시오.
10. **실행**을 클릭하십시오.

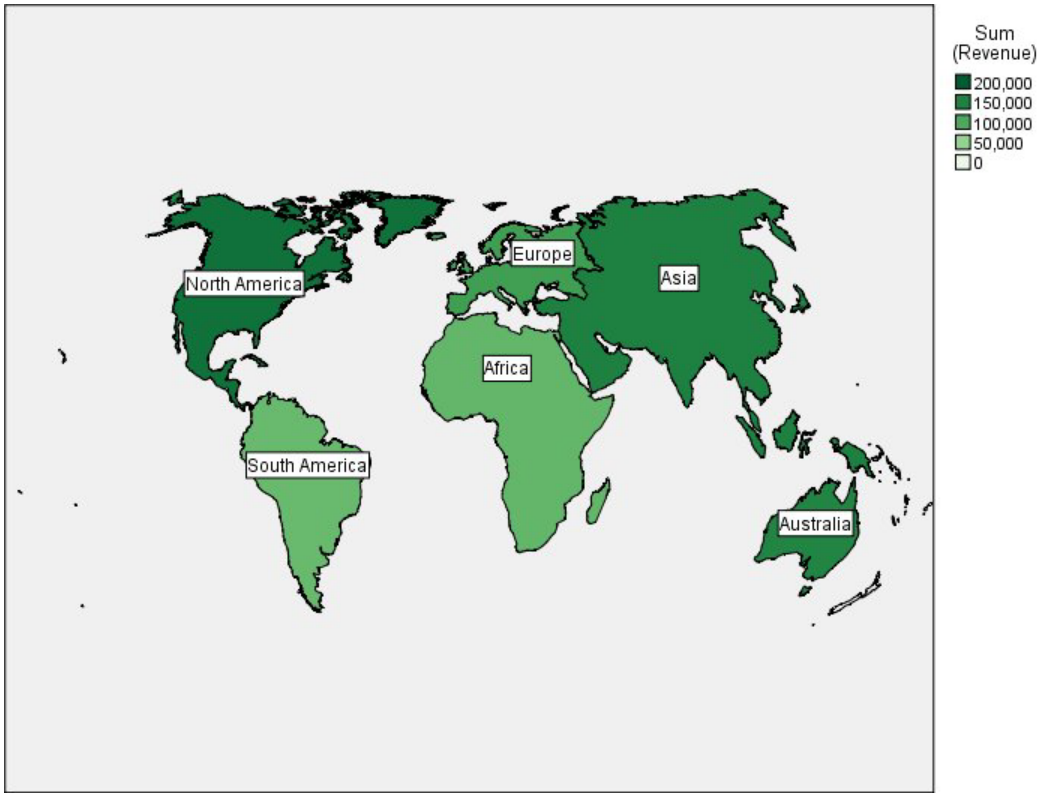


그림 16. 합계의 코로플레스

이 맵 시각화에서는 북미의 수입이 가장 높고 남미와 아프리카의 수입이 가장 낮다는 것을 쉽게 알 수 있습니다. 데이터 레이블 모양에 *대륙*을 사용했기 때문에 각 대륙의 레이블이 지정되어 있습니다.

예제: 맵 위의 막대형 차트

이 예제는 각 대륙에서 제품별로 수입이 어떻게 나뉘는지 보여줍니다.

참고: 이 예제에서는 *worldsales*를 사용합니다.

1. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
2. 기본 탭에서 *대륙*, *제품* 및 *수입*을 선택하십시오. (여러 필드/변수를 선택하려면 Ctrl+클릭을 사용하십시오.)
3. **맵 위의 막대형 차트**를 선택하십시오.
4. 세부사항 탭을 클릭하십시오.

특정 유형의 필드를 둘 이상 사용하는 경우 각 필드가 올바른 슬롯에 지정되었는지 확인하는 것이 중요합니다.

5. 범주 드롭 다운 목록에서 *제품*을 선택하십시오.
6. 값 드롭 다운 목록에서 *수입*을 선택하십시오.
7. 데이터 키 드롭 다운 목록에서 *대륙*을 선택하십시오.
8. 요약 드롭 다운 목록에서 *합계*를 선택하십시오.
9. 맵 파일 그룹에서 **맵 파일 선택**을 클릭하십시오.
10. 맵 선택 대화 상자에서 **맵**이 *대륙*으로 설정되고 **맵 키**가 *CONTINENT*로 설정되었는지 확인하십시오.
11. 맵과 데이터 값 비교 그룹에서 **비교**를 클릭하여 맵 키가 데이터 키와 일치하는지 확인하십시오. 이 예제에서는 모든 데이터 키 값이 맵 키 및 지형과 일치합니다. 오세아니아에 대한 데이터가 없음도 알 수 있습니다.
12. 맵 선택 대화 상자에서 **확인**을 클릭하십시오.
13. **실행**을 클릭하십시오.
14. 결과로 표시되는 출력 창을 최대화하여 화면을 더욱 분명하게 볼 수 있습니다.

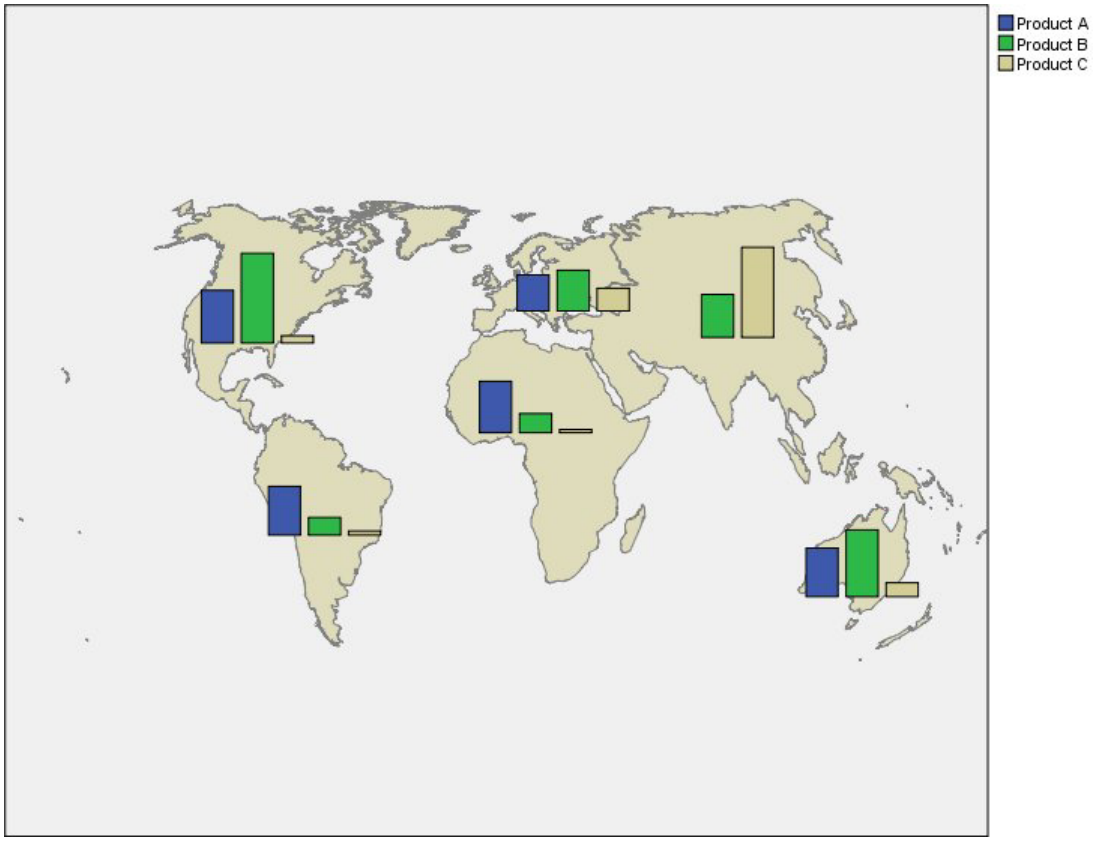


그림 17. 맵 위의 막대형 차트

다음을 관찰할 수 있습니다.

- 남미와 아프리카에서 전체 제품의 총 수입 분포가 유사합니다.

- *제품 C*는 아시아를 제외한 모든 곳에서 수입이 가장 적습니다.
- 아시아에서는 *제품 A*로 인한 수입이 없거나 최소입니다.

그래프보드 모양 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

일반적 모양 옵션

제목. 그래프 제목으로 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

캡션. 그래프 캡션에 사용할 텍스트를 입력하십시오.

표본추출. 큰 데이터 세트에 대한 방법을 지정하십시오. 최대 데이터 세트 *크기*를 지정하거나 기본 레코드 수를 사용할 수 있습니다. **표본** 옵션을 선택하면 큰 데이터 세트에 대한 성능이 개선됩니다. 또는 **모든 데이터 사용**을 선택하여 모든 데이터 점을 도표화하도록 선택할 수 있지만 소프트웨어의 성능이 급격히 저하될 수 있다는 점에 유의해야 합니다.

스타일시트 모양 옵션

사용 가능한 시각화 템플리트(및 스타일시트 및 맵)를 제어할 수 있게 하는 두 개의 단추가 있습니다.

관리. 컴퓨터에서 시각화 템플리트 , 스타일시트 및 맵을 관리합니다. 로컬 시스템에서 시각화 템플리트 , 스타일시트 및 맵을 가져오고, 내보내고, 이름을 바꾸고, 삭제할 수 있습니다. 자세한 정보는 223 페이지의 『템플리트, 스타일시트 및 맵 파일 관리』 주제를 참조하십시오.

위치. 시각화 템플리트 , 스타일시트 및 맵이 저장된 위치를 변경합니다. 현재 위치는 단추의 오른쪽에 표시됩니다. 자세한 정보는 222 페이지의 『템플리트, 스타일시트 및 맵 위치 설정』의 내용을 참조하십시오.

다음 예제에서는 그래프에서 모양 옵션이 배치되는 위치를 보여줍니다. (*참고*: 일부 그래프는 이 모든 옵션을 사용하지 않습니다.)

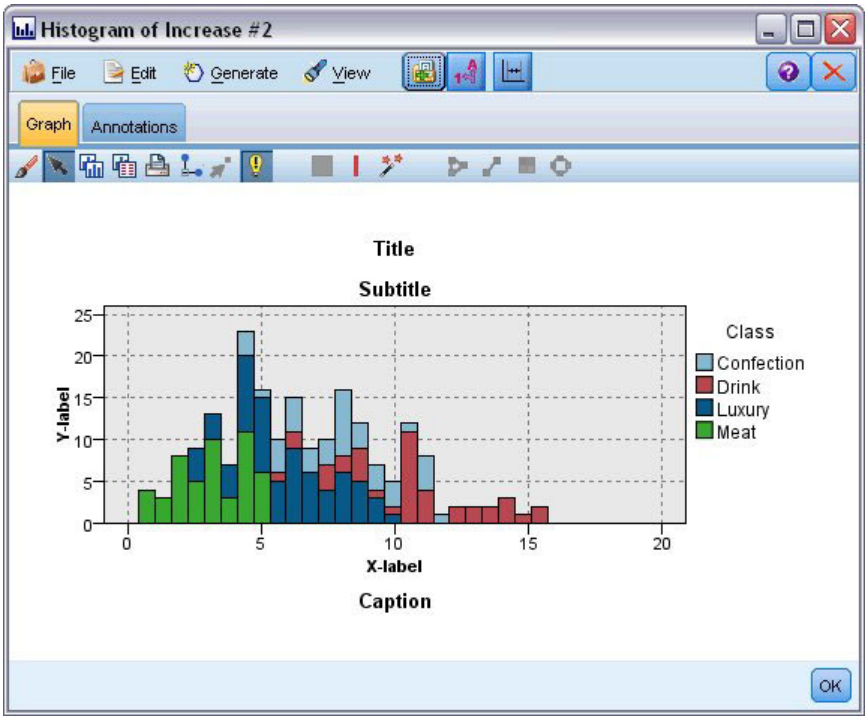


그림 18. 다양한 그래프 모양 옵션의 위치

템플리트, 스타일시트 및 맵 위치 설정

시각화 템플리트, 시각화 스타일시트 및 맵 파일은 특정 로컬 폴더 또는 IBM SPSS Collaboration and Deployment Services Repository에 저장됩니다. 템플리트, 스타일시트 및 맵을 선택하면 이 위치에 내장된 것만 표시됩니다. 모든 템플리트, 스타일시트 및 맵 파일을 한 곳에 보관하면 IBM SPSS 애플리케이션이 이러한 템플리트, 스타일시트 및 맵 파일에 쉽게 액세스할 수 있습니다. 이 위치에 템플리트, 스타일시트 및 맵 파일을 추가하는 방법에 대해서는 223 페이지의 『템플리트, 스타일시트 및 맵 파일 관리』의 내용을 참조하십시오.

템플리트, 스타일시트 및 맵 파일 위치 설정 방법

1. 템플리트 또는 스타일시트 대화 상자에서 **위치...**를 클릭하여 템플리트, 스타일시트 및 맵 대화 상자를 표시하십시오.
2. 템플리트, 스타일시트 및 맵 파일의 기본 위치 옵션을 선택하십시오.

   로컬 시스템. 템플리트, 스타일시트 및 맵 파일이 로컬 컴퓨터의 특정 폴더에 있습니다. Windows XP에서 이 폴더의 위치는 *C:\Documents and Settings\<user>\Application Data\SPSSInc\Graphboard*입니다. 폴더를 변경할 수 없습니다.

   IBM SPSS Collaboration and Deployment Services Repository. 템플리트, 스타일시트 및 맵 파일이 IBM SPSS Collaboration and Deployment Services Repository의 사용자 지정 폴더에 있습니다. 특

정 폴더를 지정하려면 **폴더**를 클릭하십시오. 자세한 정보는 『IBM SPSS Collaboration and Deployment Services Repository를 템플리트, 스타일시트 및 맵 파일 위치로 사용』의 내용을 참조하십시오.

3. **확인**을 클릭하십시오.

IBM SPSS Collaboration and Deployment Services Repository를 템플리트, 스타일시트 및 맵 파일 위치로 사용

시각화 템플리트 및 스타일시트는 IBM SPSS Collaboration and Deployment Services Repository에 저장할 수 있습니다. 이 위치는 IBM SPSS Collaboration and Deployment Services Repository에 있는 특정 폴더입니다. 이 위치를 기본 위치로 설정하면 이 위치에 있는 모든 템플리트, 스타일시트 및 맵 파일을 선택해 사용할 수 있습니다.

IBM SPSS Collaboration and Deployment Services Repository의 폴더를 템플리트, 스타일시트 및 맵 파일 위치로 설정하는 방법

1. 위치 단추가 있는 대화 상자에서 **위치...**를 클릭하십시오.
2. IBM SPSS Collaboration and Deployment Services Repository를 선택하십시오.
3. **폴더**를 클릭하십시오.

   참고: IBM SPSS Collaboration and Deployment Services Repository에 아직 연결되지 않은 경우 연결 정보 입력을 요구하는 프롬프트가 표시됩니다.
4. 폴더 선택 대화 상자에서 템플리트, 스타일시트 및 맵 파일이 저장된 폴더를 선택하십시오.
5. 선택적으로 **레이블 검색**에서 레이블을 선택할 수 있습니다. 해당 레이블을 가진 템플리트, 스타일시트 및 맵 파일만 표시됩니다.
6. 특정 템플리트 또는 스타일시트가 포함된 폴더를 찾으려면 검색 탭에서 템플리트, 스타일시트 또는 맵 파일을 검색할 수 있습니다. 폴더 선택 대화 상자는 찾은 템플리트, 스타일시트 또는 맵 파일이 있는 폴더를 자동으로 선택합니다.
7. **폴더 선택**을 클릭하십시오.

템플리트, 스타일시트 및 맵 파일 관리

템플리트, 스타일시트 및 맵 파일 대화 상자를 사용하여 컴퓨터의 로컬 위치에 있는 템플리트, 스타일시트 및 맵 파일을 관리할 수 있습니다. 이 대화 상자를 사용하여 컴퓨터의 로컬 위치에 있는 시각화 템플리트, 스타일시트 및 맵 파일에 대해 가져오기, 내보내기, 이름 바꾸기 및 삭제를 수행할 수 있습니다.

템플리트, 스타일시트 또는 맵을 관리하는 대화 상자 중 하나에서 **관리...**를 클릭하십시오.

템플리트, 스타일시트 또는 맵 관리 대화 상자

템플리트 탭은 모든 로컬 템플리트를 나열합니다. 스타일시트 탭은 로컬 스타일시트를 나열하고 표본 데이터가 포함된 표본 시각화를 표시합니다. 스타일시트 중 하나를 선택하여 해당 스타일을 시각화 예에 적용할 수 있습

니다. 자세한 정보는 300 페이지의 『스타일시트 적용』 주제를 참조하십시오. 맵 탭은 모든 로컬 맵 파일을 나열합니다. 이 탭은 또한 맵의 미리보기, 주석(맵을 작성할 때 주석을 제공한 경우) 및 표본 값이 포함된 맵 키도 표시합니다.

현재 활성화된 탭에서 작동하는 단추는 다음과 같습니다.

가져오기. 파일 시스템에서 시각화 템플리트, 스타일시트 또는 맵 파일을 가져옵니다. 템플리트, 스타일시트 또는 맵 파일을 가져오면 IBM SPSS 애플리케이션에서 이러한 파일을 사용할 수 있습니다. 다른 사용자가 템플리트, 스타일시트 또는 맵 파일을 보내온 경우에는 파일을 애플리케이션으로 가져온 후에 사용합니다.

내보내기. 파일 시스템으로 시각화 템플리트, 스타일시트 또는 맵 파일을 내보냅니다. 다른 사용자에게 템플리트, 스타일시트 또는 맵 파일을 보내려면 해당 파일을 내보내십시오.

이름 바꾸기. 선택한 시각화 템플리트, 스타일시트 또는 맵 파일의 이름을 바꿉니다. 이름을 이미 사용되는 이름으로 변경할 수 없습니다.

맵 키 내보내기. 맵 키를 CSV(쉼표로 구분된 값) 파일로 내보냅니다. 이 단추는 맵 탭에서만 사용됩니다.

삭제. 선택한 시각화 템플리트, 스타일시트 또는 맵 파일을 삭제합니다. Ctrl-클릭을 사용하여 여러 템플리트, 스타일시트 또는 맵 파일을 선택할 수 있습니다. 삭제에 대한 실행 취소 조치가 없으므로 주의해서 사용하십시오.

맵 형태 파일 변환 및 배포

그래프보드 템플리트 선택기에서는 시각화 템플리트와 SMZ 파일을 조합하여 맵 시각화를 작성할 수 있습니다. SMZ 파일은 맵을 그리기 위한 지리적 정보(예: 국경)를 포함하는 점에서 ESRI 형태 파일(SHP 파일 형식)과 유사하지만 맵 시각화를 위해 최적화되어 있습니다. 그래프보드 템플리트 선택기는 선택한 수의 SMZ 파일과 함께 사전 설치되어 있습니다. 맵 시각화에 사용하려는 기존 ESRI 형태 파일이 있는 경우, 먼저 맵 변환 유틸리티를 사용하여 형태 파일을 SMZ 파일로 변환해야 합니다. 맵 변환 유틸리티는 점, 폴리라인 또는 단일 레이어를 포함한 다각형(형태 유형 1, 3 및 5) ESRI 형태 파일을 지원합니다.

맵 변환 유틸리티에서는 ESRI 형태 파일을 변환할 수 있을 뿐만 아니라 맵의 세부사항 수준을 수정하고 지형 레이블을 변경하며 지형을 합치고 지형을 이동시킬 수 있습니다. 기존 SMZ 파일(사전 설치된 SMZ 파일 포함)을 수정하는 데에도 맵 변환 유틸리티를 사용할 수 있습니다.

사전 설치된 SMZ 파일 편집

1. 관리 시스템에서 SMZ 파일을 내보내십시오. 자세한 정보는 223 페이지의 『템플리트, 스타일시트 및 맵 파일 관리』 주제를 참조하십시오.
2. 맵 변환 유틸리티를 사용하여 내보낸 SMZ 파일을 열고 편집하십시오. 파일을 다른 이름으로 저장하는 것이 좋습니다. 자세한 정보는 226 페이지의 『맵 변환 유틸리티 사용』 주제를 참조하십시오.
3. 수정된 SMZ 파일을 관리 시스템으로 가져오십시오. 자세한 정보는 223 페이지의 『템플리트, 스타일시트 및 맵 파일 관리』 주제를 참조하십시오.

맵 파일에 대한 추가 자원

맵핑 요구사항을 지원하는 데 사용할 수 있는 SHP 파일 형식의 지리 공간적 데이터는 다양한 개인용 및 공용 소스로부터 얻을 수 있습니다. 무료 데이터를 찾는 경우에는 지역 정부 웹 사이트를 확인하십시오. 이 제품에 포함된 다수의 템플리트는 GeoCommons( ) 및 미국 통계국(http://www.census.gov)에서 제공하는 공개적으로 사용 가능한 데이터를 기반으로 합니다.

중요 주의사항: 비IBM 제품에 관한 정보는 해당 제품의 공급업체, 공개 자료 또는 기타 범용 소스로부터 얻은 것입니다. IBM에서는 이러한 제품들을 테스트하지 않았으므로, 비IBM 제품과 관련된 성능의 정확성, 호환성 또는 기타 청구에 대해서는 확신할 수 없습니다. 비IBM 제품의 성능에 대한 의문사항은 해당 제품의 공급업체에 문의하십시오. 이 정보에서 언급되는 비IBM의 웹 사이트는 단지 편의상 제공된 것으로, 어떤 방식으로든 이들 웹 사이트를 옹호하고자 하는 것은 아닙니다. 이 IBM 프로그램과 함께 제공되는 고지 파일에 표시하지 않는 한, 해당 웹 사이트의 자료는 본 IBM 프로그램 자료의 일부가 아니므로 해당 웹 사이트 사용으로 인한 위험은 사용자 본인이 감수해야 합니다.

맵의 핵심 개념

형태 파일과 관련된 몇 가지 핵심 개념을 이해하면 맵 변환 유틸리티를 효과적으로 사용하는 데 도움이 됩니다.

형태 파일은 맵을 그리기 위한 지리적 정보를 제공합니다. 맵 변환 유틸리티는 다음의 세 가지 유형의 형태 파일을 지원합니다.

- **점.** 이 형태 파일은 점의 위치(예: 도시)를 식별합니다.
- **폴리라인.** 이 형태 파일은 경로 및 경로의 위치(예: 강)를 식별합니다.
- **다각형.** 이 형태 파일은 경계가 있는 지역 및 이러한 지역의 위치(예: 국가)를 식별합니다.

대부분 다각형 형태 파일을 가장 많이 사용하게 됩니다. 코로플레스 맵은 다각형 형태 파일에서 작성됩니다. 코로플레스 맵은 색상을 사용하여 개별 다각형(지역) 내의 값을 나타냅니다. 점 및 폴리라인 형태 파일은 일반적으로 다각형 형태 파일 위에 오버레이됩니다. 미국 주의 다각형 형태 파일에 오버레이된 미국 도시의 점 형태 파일이 한 예입니다.

형태 파일은 **지형**으로 구성됩니다. 지형은 개별 지리적 엔티티입니다. 예를 들어, 지형은 국가, 주, 도시 등일 수 있습니다. 형태 파일은 지형에 대한 데이터도 포함합니다. 이러한 데이터는 **속성**에 저장됩니다. 속성은 데이터 파일의 필드 또는 변수와 유사합니다. 지형의 **맵 키**인 속성이 하나 이상 있습니다. 맵 키는 국가 이름 또는 주 이름과 같은 레이블일 수 있습니다. 맵 키는 맵 시각화를 작성하기 위해 데이터 파일의 변수/필드에 연결하는 것입니다.

SMZ 파일에서는 키 속성만 유지할 수 있습니다. 맵 변환 유틸리티는 추가 속성 저장을 지원하지 않습니다. 따라서 서로 다른 수준에서 통합하려면 여러 개의 SMZ 파일을 작성해야 합니다. 예를 들어, 미국 주와 지역을 통합하려면 주를 식별하는 키가 있는 SMZ 파일과 지역을 식별하는 키가 있는 SMZ 파일이 개별적으로 필요합니다.

맵 변환 유틸리티 사용

맵 변환 유틸리티 시작 방법

메뉴에서 다음을 선택하십시오.

도구 > 맵 변환 유틸리티

맵 변환 유틸리티에는 네 개의 주 화면(단계)이 있습니다. 단계 중 하나에는 맵 파일 편집과 관련된 보다 세부적인 제어를 위한 하위 단계도 포함됩니다.

1단계 - 대상 및 소스 파일 선택

먼저 변환되는 맵 파일의 소스 맵 파일과 대상을 선택해야 합니다. 형태 파일에는 *.shp* 및 *.dbf* 파일이 모두 필요합니다.

변환을 위해 .shp(ESRI) 또는 .smz 파일 선택. 찾아보기로 컴퓨터에 있는 기존 맵 파일을 찾으십시오. 이 파일은 SMZ 파일로 변환하고 저장할 파일입니다. 형태 파일을 위한 *.dbf* 파일은 *반드시 .shp* 파일과 일치하는 기준 파일 이름과 동일한 위치에 저장해야 합니다. *.dbf* 파일이 필요한 이유는 이 파일에 *.shp* 파일의 속성 정보가 있기 때문입니다.

변환되는 맵 파일의 대상 및 파일 이름 설정. 원래 맵 소스에서 작성할 SMZ 파일의 경로 및 파일 이름을 입력하십시오.

- **템플리트 선택기로 가져오기.** 파일 시스템에 파일을 저장할 수 있을 뿐만 아니라 선택적으로 템플리트 선택기의 관리 목록에 맵을 추가할 수 있습니다. 이 옵션을 선택하면 컴퓨터에 설치된 IBM SPSS 제품의 템플리트 선택기에서 자동으로 해당 맵을 사용할 수 있습니다. 지금 템플리트 선택기로 가져오지 않는 경우 나중에 수동으로 가져와야 합니다. 템플리트 선택기 관리 시스템에 맵을 가져오는 방법에 대한 자세한 정보는 223 페이지의 『템플리트, 스타일시트 및 맵 파일 관리』의 내용을 참조하십시오.

2단계 - 맵 키 선택

이제 SMZ 파일에 포함시킬 맵 키를 선택합니다. 그런 다음 맵 렌더링에 영향을 주는 일부 옵션을 변경할 수 있습니다. 맵 변환 유틸리티에서의 이후 단계에는 맵의 미리보기가 포함됩니다. 선택하는 렌더링 옵션은 맵 미리보기를 생성하는 데 사용됩니다.

기본 맵 키 선택. 맵의 지형을 식별하고 레이블을 지정하는 기본 키인 속성을 선택하십시오. 예를 들어, 세계 맵의 기본 키는 국가 이름을 식별하는 속성일 수 있습니다. 기본 키는 또한 데이터를 맵 지형에 연결하므로 선택하는 속성의 값(레이블)이 데이터의 값과 일치해야 합니다. 속성을 선택하면 레이블 예가 표시됩니다. 이러한 레이블을 변경해야 하는 경우 나중 단계에서 이를 수행할 수 있습니다.

포함시킬 추가 키 선택. 기본 맵 키 이외에, 생성된 SMZ 파일에 포함시킬 다른 키 속성을 선택하십시오. 예를 들어, 일부 속성에 변환된 레이블이 있을 수 있습니다. 다른 언어로 코딩된 데이터를 예상하는 경우 이러한 속성을 유지하려 할 수 있습니다. 기본 키와 동일한 지형을 나타내는 추가 키만 선택할 수 있습니다. 예를 들어, 기본 키가 미국 주의 전체 이름인 경우 미국 주를 나타내는 대체 키(예: 주 약어)만 선택할 수 있습니다.

자동으로 맵 평활화. 다각형을 포함하는 형태 파일에는 일반적으로 통계 맵 시각화를 위한 너무 많은 데이터 점과 너무 많은 세부사항이 있습니다. 세부사항이 지나치게 많으면 산만하고 성능에 부정적인 영향을 미칠 수 있습니다. 세부사항 수준을 낮추고 평활화로 맵을 일반화할 수 있습니다. 이렇게 하면 맵이 더 단정해 보이고 더 빨리 렌더링됩니다. 맵이 자동으로 평활화되는 경우 최대 각은 15도이고 유지할 백분율은 99입니다. 이러한 설정에 대한 정보는 『맵 평활화』의 내용을 참조하십시오. 나중에 다른 단계에서 평활화를 추가로 적용할 기회가 있습니다.

같은 지형에서 맞닿은 다각형 사이의 경계 제거. 일부 지형은 주 관심 지형 내부에 경계가 있는 하위 지형을 포함할 수 있습니다. 예를 들어, 세계 대륙 맵에는 각 대륙에 포함된 국가의 내부 경계가 포함될 수 있습니다. 이 옵션을 선택하면 맵에 내부 경계가 표시되지 않습니다. 세계 대륙 맵 예에서 이 옵션을 선택하면 대륙 경계는 유지되지만 국가 경계가 제거됩니다.

3단계 - 맵 편집

이제 맵에 대한 기본 옵션을 지정했으므로 보다 구체적인 옵션을 편집할 수 있습니다. 이러한 수정은 선택사항입니다. 맵 변환 유틸리티의 이 단계에서는 연관된 작업의 수행 과정을 안내하고 변경사항을 확인할 수 있도록 맵의 미리보기를 표시합니다. 형태 파일 유형(점, 폴리라인 또는 다각형) 및 좌표계에 따라 일부 작업이 사용불가능할 수도 있습니다.

모든 작업은 맵 변환 유틸리티의 왼쪽에 다음과 같은 공통 제어가 있습니다.

맵에 레이블 표시. 기본적으로 미리보기에는 지형 레이블이 표시되지 않습니다. 이러한 레이블을 표시하도록 선택할 수 있습니다. 지형 레이블은 지형을 식별하는 데 도움이 될 수 있지만 미리보기 맵에서 직접 선택하는 데 방해가 될 수 있습니다. 필요한 경우, 예를 들어, 지형 레이블을 편집하는 경우에는 이 옵션을 끄십시오.

맵 미리보기 채색. 기본적으로 맵 미리보기는 영역을 단색으로 표시합니다. 모든 지형의 색상이 동일합니다. 개별 맵 지형에 다양한 색상이 지정되도록 선택할 수 있습니다. 이 옵션은 맵에서 서로 다른 지형을 구별하는 데 유용할 수 있습니다. 특히 지형을 합칠 때 미리보기에서 새 지형이 어떻게 표시되는지 확인하려는 경우에 유용합니다.

모든 작업은 또한 맵 변환 유틸리티의 오른쪽에 다음과 같은 공통 제어가 있습니다.

실행 취소. 이전 상태로 되돌아가려면 **실행 취소**를 클릭하십시오. 최대 100번의 변경을 실행 취소할 수 있습니다.

맵 평활화: 다각형을 포함하는 형태 파일에는 일반적으로 통계 맵 시각화를 위한 너무 많은 데이터 점과 너무 많은 세부사항이 있습니다. 세부사항이 지나치게 많으면 산만하고 성능에 부정적인 영향을 미칠 수 있습니다. 세부사항 수준을 낮추고 평활화로 맵을 일반화할 수 있습니다. 이렇게 하면 맵이 더 단정해 보이고 더 빨리 렌더링됩니다. 점 및 폴리라인 맵에는 이 옵션을 사용할 수 없습니다.

최대 각. 최대 각은 1과 20 사이의 값이어야 하며 거의 선형인 일련의 점을 평활화하기 위한 허용 오차를 지정합니다. 값이 클수록 선형 평활화에 대한 허용 오차가 커지고 그에 따라 더 많은 점이 삭제되어 좀 더 일반화된 맵이 됩니다. 선형 평활화를 적용하기 위해 맵 변환 유틸리티는 맵에서 세 개의 점으로 구성된 각각의 세트가 이루는 내각을 확인합니다. 180에서 내각을 뺀 값이 지정한 값보다 작으면 맵 변환 유틸리티는 가운데

점을 삭제합니다. 즉, 맵 변환 유틸리티는 세 개의 점으로 형성된 선이 거의 직선인지 여부를 확인합니다. 그러한 경우 맵 변환 유틸리티는 해당 선을 엔드포인트 사이의 직선으로 처리하여 중간 점을 삭제합니다.

유지할 퍼센트. 유지할 백분율은 90과 100 사이의 값이어야 하며 맵을 평활화할 때 유지할 땅 영역의 양을 결정합니다. 이 옵션은 여러 다각형을 포함하는 지형(예: 지형에 섬이 포함된 경우)에만 영향을 줍니다. 지형의 총 영역에서 다각형을 뺀 값이 원래 영역의 지정된 백분율보다 큰 경우 맵 변환 유틸리티는 맵에서 해당 다각형을 삭제합니다. 맵 변환 유틸리티는 지형의 모든 다각형을 제거하지는 않습니다. 즉, 적용되는 평활화 양과 무관하게 지형에는 항상 하나 이상의 다각형이 있습니다.

최대 각과 유지할 백분율을 선택한 후에는 **적용**을 클릭하십시오. 미리보기가 평활화 변경사항으로 업데이트됩니다. 맵을 다시 평활화해야 하는 경우 원하는 평활 수준이 될 때까지 반복하십시오. 평활화에는 한계가 있습니다. 반복해서 평활화하면 맵에 추가 평활화를 적용할 수 없는 지점에 이르게 됩니다.

지형 레이블 편집: 필요에 따라(예: 예상 데이터와 일치시키기 위해) 지형 레이블을 편집하고 맵에서 레이블의 위치를 바꿀 수 있습니다. 레이블을 변경할 필요가 없다고 생각하는 경우에도 맵에서 시각화를 작성하기 전에 레이블을 검토해야 합니다. 미리보기에는 기본적으로 레이블이 표시되지 않으므로 **맵에 레이블 표시**를 선택하여 레이블을 표시할 수도 있습니다.

키. 검토하거나 편집할 지형 레이블이 포함된 키를 선택하십시오.

변수. 이 목록은 선택한 키에 포함된 지형 레이블을 표시합니다. 레이블을 편집하려면 목록에서 레이블을 두 번 클릭하십시오. 맵에 레이블이 표시되는 경우 맵 미리보기에서 직접 지형 레이블을 두 번 클릭할 수도 있습니다. 레이블을 실제 데이터 파일과 비교하려면 **비교**를 클릭하십시오.

X/Y. 이 텍스트 상자는 맵에서 선택한 지형 레이블의 현재 중심점을 나열합니다. 단위는 맵의 좌표에 표시됩니다. 좌표는 로컬 데카르트 좌표(예: 미국 평면 좌표계) 또는 지리적 좌표(**X**는 경도이고 **Y**는 위도인 좌표)일 수 있습니다. 레이블의 새 위치에 대한 좌표를 입력하십시오. 레이블이 표시되는 경우 맵에서 레이블을 클릭하여 끌기 조작으로 이동시킬 수 있습니다. 텍스트 상자가 새 위치로 업데이트됩니다.

비교. 특정 키의 지형 레이블과 비교할 데이터 값이 포함된 데이터 파일이 있는 경우 **비교**를 클릭하여 외부 데이터 소스와 비교 대화 상자를 표시하십시오. 이 대화 상자에서 데이터 파일을 열고 해당 값을 맵 키의 지형 레이블에 있는 값과 직접 비교할 수 있습니다.

외부 데이터 소스와 비교 대화 상자: 외부 데이터 소스와 비교 대화 상자에서는 탭으로 구분된 값 파일(*.txt* 확장자를 가짐), 쉼표로 구분된 값 파일(*.csv* 확장자를 가짐) 또는 IBM SPSS Statistics에 맞게 형식화된 데이터 파일(*.sav* 확장자를 가짐)을 열 수 있습니다. 파일이 열리면 데이터 파일에서 필드를 선택하여 특정 맵 키의 지형 레이블과 비교할 수 있습니다. 그런 다음 맵 파일에서 일치하지 않는 부분을 정정할 수 있습니다.

데이터 파일의 필드. 지형 레이블과 값을 비교할 필드를 선택하십시오. *.txt* 또는 *.csv* 파일의 첫 번째 행에 각 필드의 설명 레이블이 있으면 **첫 번째 행을 열 레이블로 사용**을 선택하십시오. 그렇지 않은 경우 각 필드는 데이터 파일에서 해당 위치로 식별됩니다(예: "열 1", "열 2" 등).

비교할 키. 데이터 파일 필드 값과 지형 레이블을 비교할 맵 키를 선택하십시오.

비교. 값을 비교할 준비가 되었을 때 클릭하십시오.

비교 결과. 기본적으로 비교 결과 테이블은 데이터 파일에서 일치하지 않는 필드 값만 나열합니다. 애플리케이션은 주로 삽입되었거나 누락된 공간이 있는지 확인함으로써 관련된 지형 레이블을 찾으려고 합니다. *맵 레이블* 열에서 드롭 다운 목록을 클릭하여 맵 파일의 지형 레이블을 표시된 필드 값과 일치시키십시오. 맵 파일에 일치하는 지형 레이블이 없으면 *일치하지 않은 상태로 두기*를 선택하십시오. 이미 지형 레이블과 일치하는 필드 값을 포함하여 모든 필드 값을 보려면 **일치하지 않는 케이스만 표시**를 선택 취소하십시오. 하나 이상의 일치를 대체하려는 경우 이를 수행할 수 있습니다.

각 지형을 한 번만 사용하여 필드 값에 일치시킬 수 있습니다. 여러 지형을 하나의 필드 값에 일치시키려는 경우 지형을 합친 후 합쳐진 새 지형을 필드 값에 일치시킬 수 있습니다. 지형 합치기에 대한 자세한 정보는 『지형 합치기』의 내용을 참조하십시오.

지형 합치기: 지형 합치기는 맵에서 더 큰 지역을 작성하는 데 유용합니다. 예를 들어, 주 맵을 변환하는 경우 주(이 예제의 지형)를 더 큰 북부, 남부, 동부 및 서부 지역으로 합칠 수 있습니다.

키. 합칠 지형을 식별하는 데 도움이 될 지형 레이블이 포함된 맵 키를 선택하십시오.

변수. 합칠 첫 번째 지형을 클릭하십시오. Ctrl-클릭을 사용하여 합칠 다른 지형을 선택하십시오. 지형은 맵 미리보기에서도 선택됩니다. 목록에서 지형을 선택할 수 있을 뿐만 아니라 맵 미리보기에서 직접 지형을 클릭 및 Ctrl-클릭할 수 있습니다.

합칠 지형을 선택한 후에는 **합치기**를 클릭하여 합친 지형의 이름 지정 대화 상자를 표시하십시오. 여기서 새 지형에 레이블을 적용할 수 있습니다. 지형을 합친 후 결과가 예상과 일치하는지 확인하기 위해 **맵 미리보기 채색**을 선택할 수 있습니다.

지형을 합친 후 새 지형의 레이블을 이동시킬 수도 있습니다. *지형 레이블 편집* 작업에서 이를 수행할 수 있습니다. 자세한 정보는 228 페이지의 『지형 레이블 편집』의 내용을 참조하십시오.

합친 지형의 이름 지정 대화 상자: 합친 지형의 이름 지정 대화 상자에서는 합친 새 지형에 레이블을 지정할 수 있습니다.

레이블 테이블은 맵 파일에 있는 각 키에 대한 정보를 표시하며 이 테이블에서 각 키에 레이블을 지정할 수 있습니다.

새 레이블. 특정 맵 키에 지정할 합친 지형의 새 레이블을 입력하십시오.

키. 새 레이블을 지정할 맵 키입니다.

이전 레이블. 새 지형으로 합칠 지형의 레이블입니다.

맞닿은 다각형 사이의 경계 제거. 합쳐진 지형에서 경계를 제거하려면 이 옵션을 선택하십시오. 예를 들어, 주를 지리적 지역으로 합친 경우 이 옵션은 개별 주의 경계를 제거합니다.

지형 이동: 맵에서 지형을 이동시킬 수 있습니다. 이는 본토와 딸린 섬처럼 여러 지형을 한 데 모으려는 경우에 유용할 수 있습니다.

키. 이동시킬 지형을 식별하는 데 도움이 될 지형 레이블이 포함된 맵 키를 선택하십시오.

변수. 이동시킬 지형을 클릭하십시오. 지형은 맵 미리보기에서 선택됩니다. 맵 미리보기에서 직접 지형을 클릭할 수도 있습니다.

X/Y. 이 텍스트 상자는 맵에서 지형의 현재 중심점을 나열합니다. 단위는 맵의 좌표에 표시됩니다. 좌표는 로컬 데카르트 좌표(예: 미국 평면 좌표계) 또는 지리적 좌표(**X**는 경도이고 **Y**는 위도인 좌표)일 수 있습니다. 지형의 새 위치에 대한 좌표를 입력하십시오. 맵에서 지형을 클릭하여 끌기 조작으로 이동시킬 수도 있습니다. 텍스트 상자가 새 위치로 업데이트됩니다.

지형 삭제: 맵에서 원하지 않는 지형을 삭제할 수 있습니다. 이는 맵 시각화에서 관련이 없는 지형을 삭제하여 복잡성을 제거하려는 경우에 유용할 수 있습니다.

키. 삭제할 지형을 식별하는 데 도움이 될 지형 레이블이 포함된 맵 키를 선택하십시오.

변수. 삭제할 지형을 클릭하십시오. 동시에 여러 지형을 삭제하려면 Ctrl-클릭을 사용하여 지형을 추가로 선택하십시오. 지형은 맵 미리보기에서도 선택됩니다. 목록에서 지형을 선택할 수 있을 뿐만 아니라 맵 미리보기에서 직접 지형을 클릭 및 Ctrl-클릭할 수 있습니다.

개별 요소 삭제: 전체 지형을 삭제할 수 있을 뿐만 아니라 지형을 구성하는 일부 개별 요소(예: 호수 및 작은 섬)를 삭제할 수도 있습니다. 점 맵에는 이 옵션을 사용할 수 없습니다.

요소. 삭제할 요소를 클릭하십시오. 동시에 여러 요소를 삭제하려면 Ctrl-클릭을 사용하여 요소를 추가로 선택하십시오. 요소는 맵 미리보기에서도 선택됩니다. 목록에서 요소를 선택할 수 있을 뿐만 아니라 맵 미리보기에서 직접 요소를 클릭 및 Ctrl-클릭할 수 있습니다. 요소 이름 목록은 설명적이지 않으므로(각 요소는 지형 내에서 번호로 지정됨) 맵 미리보기에서 원하는 요소를 선택했는지 확인해야 합니다.

투영법 설정:

맵 투영법은 3차원의 지구를 2차원으로 나타내는 방법을 지정합니다. 모든 투영법은 왜곡을 야기시킵니다. 그러나 구형 맵을 보는지 또는 좀 더 국지적인 맵을 보는지에 따라 더 적합한 투영법이 있습니다. 또한 일부 투영법은 원래 지형의 형태를 유지합니다. 형태를 유지하는 투영법은 정각 투영법입니다. 이 옵션은 지리적 좌표(경도 및 위도)가 있는 맵에만 사용할 수 있습니다.

맵 변환 유틸티리의 다른 옵션과 달리 투영법은 맵 시각화 작성 후에 변경할 수 있습니다.

투영법. 맵 투영법을 선택하십시오. 구형 또는 반구형 맵을 작성하는 경우에는 *국지*, *메르카토르* 또는 *빈켈 트리펠* 투영법을 사용하십시오. 더 작은 영역에는 *국지*, *람베르트 정각원추* 또는 *횡축 메르카토르* 투영법을 사용하십시오. 모든 투영법은 데이텀에 WGS83 타원체를 사용합니다.

- **국지** 투영법은 항상 맵이 국지 좌표계(예: 미국 평면 좌표계)로 작성된 경우에 사용됩니다. 이러한 좌표계는 지리적 좌표(경도 및 위도)가 아닌 데카르트 좌표에 의해 정의됩니다. 국지 투영법에서는 데카르트 좌표계의 수평선과 수직선의 간격이 동일합니다. 국지 투영법은 정각 투영법이 아닙니다.

- **메르카토르** 투영법은 구형 맵을 위한 정각 투영법입니다. 수평선과 수직선이 일직선이며 항상 서로 수직을 이룹니다. 메르카토르 투영법은 북극과 남극에 접근함에 따라 무한대로 확장되므로 북극 또는 남극을 포함하는 맵에는 사용할 수 없습니다. 맵이 이러한 한계에 접근할 때 가장 크게 왜곡됩니다.
- **빈켈 트리펠** 투영법은 구형 맵을 위한 비정각 투영법입니다. 정각 투영법은 아니지만 형태와 크기 사이에 적절한 균형을 제공합니다. 적도와 본초자오선을 제외한 모든 선이 곡선입니다. 구형 맵에 북극 또는 남극이 포함되는 경우 이 투영법을 선택하는 것이 좋습니다.
- 이름에서 알 수 있듯이 **람베르트 정각원추** 투영법은 정각 투영법이며 북쪽과 남쪽에 비해 동쪽과 서쪽이 더 긴 대륙 또는 그보다 작은 육지의 맵에 사용됩니다.
- **횡축 메르카토르**는 대륙 또는 그보다 작은 육지의 맵을 위한 또하나의 정각 투영법입니다. 동쪽과 서쪽에 비해 북쪽과 남쪽이 더 긴 육지에 이 투영법을 사용하십시오.

4단계 - 완료

이 시점에서 맵 파일을 설명하는 주석을 추가하고 맵 키에서 표본 데이터 파일을 작성할 수 있습니다.

맵 키. 맵 파일에 여러 키가 있으면 미리보기에 지형 레이블을 표시할 맵 키를 선택하십시오. 맵에서 데이터 파일을 작성하는 경우 이러한 레이블이 데이터 값에 사용됩니다.

주석. 맵을 설명하거나 사용자와 관련이 있을 수 있는 추가 정보(예: 원래 형태 파일의 소스)를 제공하는 주석을 입력하십시오. 주석은 그래프보드 템플리트 선택기의 관리 시스템에 표시됩니다.

지형 레이블에서 데이터 세트 작성. 표시된 지형 레이블에서 데이터 파일을 작성하려는 경우 이 옵션을 선택하십시오. **찾아보기...**를 클릭하면 위치 및 파일 이름을 지정할 수 있습니다. *.txt* 확장자를 추가하면 파일이 탭으로 구분된 값 파일로 저장됩니다. *.csv* 확장자를 추가하면 파일이 쉼표로 구분된 값 파일로 저장됩니다. *.sav* 확장자를 추가하면 파일이 IBM SPSS Statistics 형식으로 저장됩니다. 확장자를 지정하지 않으면 SAV가 기본값입니다.

맵 파일 배포

맵 변환 유틸리티의 첫 번째 단계에서 변환된 SMZ 파일을 저장할 위치를 선택했습니다. 또한 그래프보드 템플리트 선택기의 관리 시스템에 맵을 추가하도록 선택했을 수도 있습니다. 관리 시스템에 저장하도록 선택한 경우, 동일한 컴퓨터에서 실행하는 모든 IBM SPSS 제품에서 해당 맵을 사용할 수 있습니다.

맵을 다른 사용자에게 배포하려면 맵을 배포할 사용자에게 SMZ를 보내야 합니다. 그러면 해당 사용자가 관리 시스템을 사용하여 맵을 가져올 수 있습니다. 1단계에서 위치를 지정한 파일은 보내기만 하면 됩니다. 관리 시스템에 있는 파일을 보내려면 먼저 파일을 내보내야 합니다.

1. 템플리트 선택기에서 **관리...**를 클릭하십시오.
2. 맵 탭을 클릭하십시오.
3. 배포할 맵을 선택하십시오.
4. **내보내기...**를 클릭하고 파일을 저장할 위치를 선택하십시오.

이제 실제 맵 파일을 다른 사용자에게 보낼 수 있습니다. 사용자는 이 프로세스를 역으로 수행하여 맵을 관리 시스템으로 가져와야 합니다.

구성 노드

구성 노드는 수치 필드 사이의 관계를 보여줍니다. 포인트(산점도) 또는 선을 사용하여 도표를 작성할 수 있습니다. 대화 상자에서 X 모드를 지정하여 세 가지 유형의 선 도표를 작성할 수 있습니다.

X 모드 = 정렬

X 모드를 **정렬**로 설정하면 데이터가 *x* 축에 구성된 필드에 대한 값으로 정렬됩니다. 그러면 그래프의 왼쪽에서 오른쪽으로 진행되는 단일 선이 생성됩니다. 명목 필드를 오버레이로 사용하면 그래프에서 왼쪽에서 오른쪽으로 진행되는 다양한 색상의 다중 선이 생성됩니다.

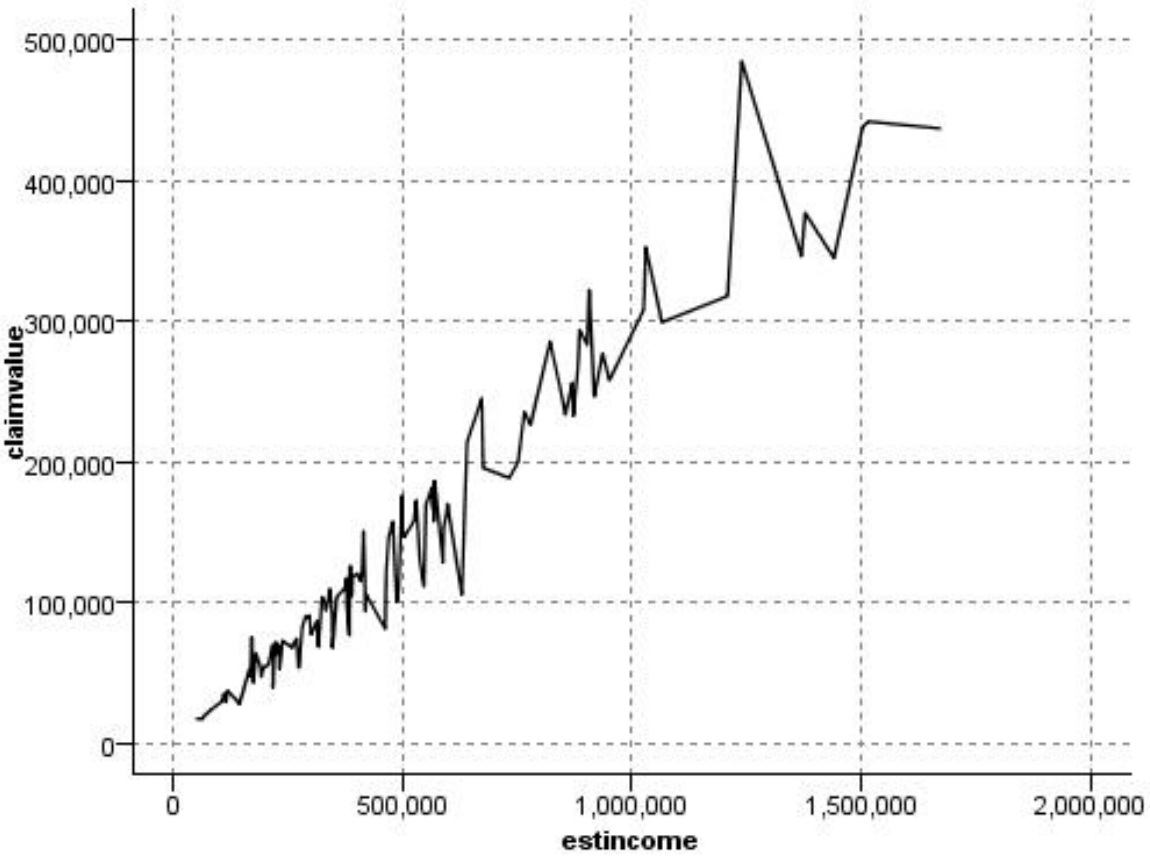


그림 19. X 모드가 정렬로 설정된 선 도표

X 모드 = 오버레이

X 모드를 **오버레이**로 설정하면 동일한 그래프에서 다중 선 도표가 작성됩니다. *x* 축의 값이 증가하는 한 데이터가 단일 선에 구성되며 데이터가 오버레이 도표에 대해 정렬되지 않습니다. 값이 감소하면 새 선이 시작됩니다. 예를 들어, *x*가 0에서 100으로 이동하면 *y*값이 단일 선에 구성될 것입니다. *x*가 100 아래가 되면 첫 번째 선 외에 새 선이 구성될 것입니다. 완료된 도표는 일련의 *y*값을 비교하는 데 유용한 수많은 도표를 갖게 될 것

입니다. 이 유형의 도표는 정기적인 시간 구성요소(연속되는 24시간 동안의 전기 요구량 등)가 있는 데이터에 유용합니다.

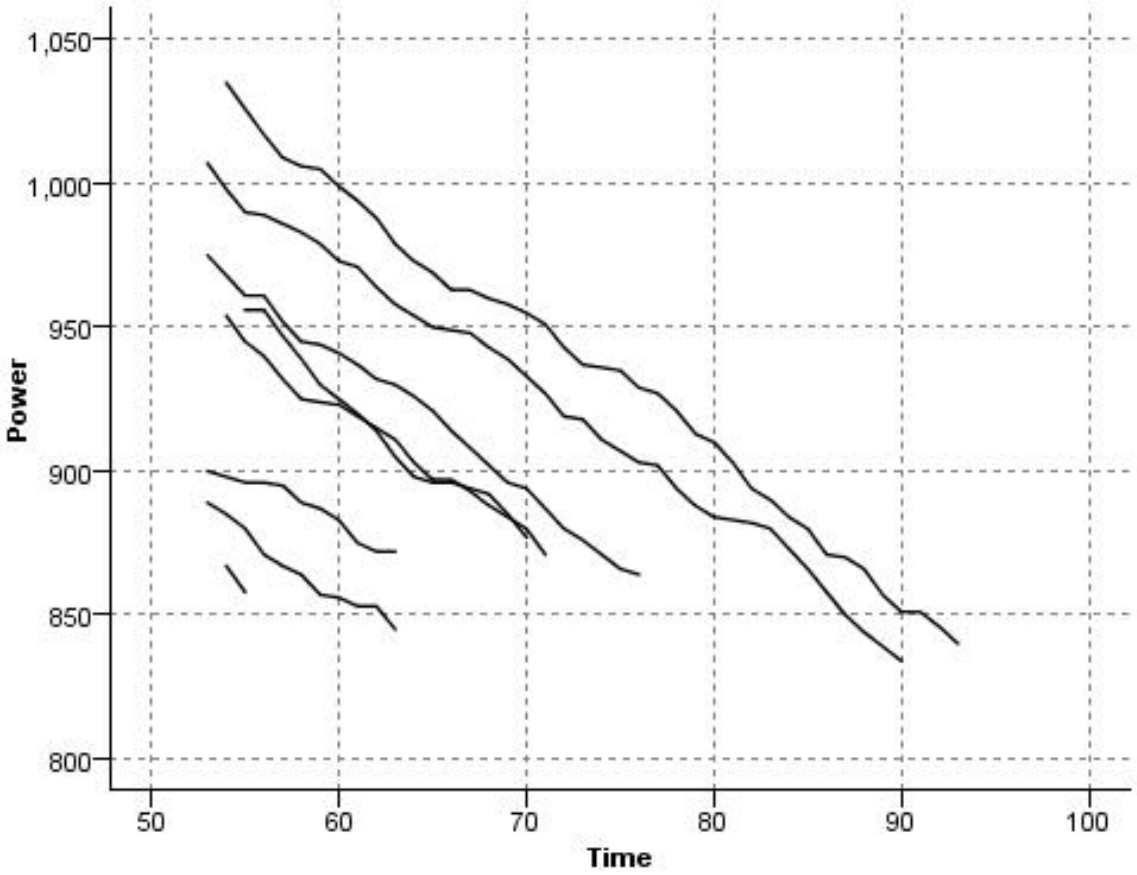


그림 20. X 모드가 오버레이로 설정된 선 도표

X 모드 = 읽은 대로

X 모드를 **읽은 대로** 도표로 설정하면 *x* 및 *y* 값을 데이터 소스에서 읽은 대로 구성합니다. 이 옵션은 사용자가 데이터의 순서에 따른 추세 또는 패턴에 관심이 있는 경우에 시계열 구성요소가 있는 데이터에 유용합니다. 이 유형의 도표를 작성하기 전에 데이터를 정렬해야 합니다. 또한 패턴이 정렬에 의존하는 정도를 판별하기 위해 X 모드가 **정렬** 및 **읽은 대로**로 설정된 두 개의 유사한 도표를 비교하는 데 유용합니다.

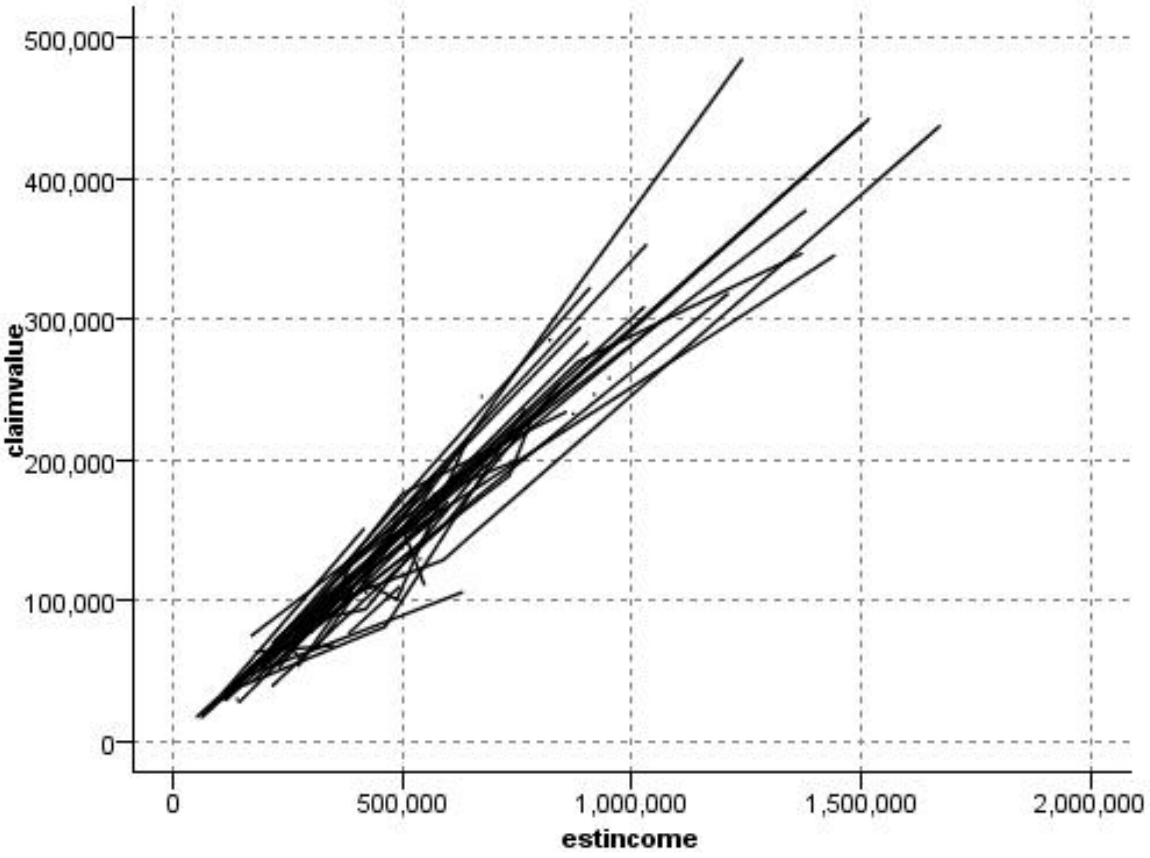


그림 21. 이전에 정렬로 표시된 선 도표, X 모드를 읽은 대로로 설정하고 다시 실행

또한 그래프보드 노드를 사용하는 방법으로도 산점도 및 선 도표를 생성할 수 있습니다. 그러나 이 노드에서 선택할 수 있는 옵션이 더 많습니다. 자세한 정보는 204 페이지의 『사용 가능한 내장 그래프보드 시각화 유형』의 내용을 참조하십시오.

구성 노드 탭

도표는 *Y* 필드의 값 대 *X* 필드의 값을 표시합니다. 종종 이러한 필드는 각각 종속변수 및 독립변수에 해당됩니다.

X 필드. 목록에서 수평 *x*축으로 표시할 필드를 선택합니다.

Y 필드. 목록에서 수직 *y*축으로 표시할 필드를 선택합니다.

Z 필드. 3D 차트 단추를 클릭하면 목록에서 *z*축을 표시할 필드를 선택할 수 있습니다.

오버레이. 여러 가지 방식으로 데이터 값에 대한 범주를 표시할 수 있습니다. 예를 들어, *maincrop*를 색상 오버레이로 사용하여 클레임 지원자가 키운 주요 작물에 대한 *estincome* 및 *claimvalue* 값을 표시할 수 있습니다. 자세한 정보는 194 페이지의 『모양, 오버레이, 패널 및 애니메이션』의 내용을 참조하십시오.

오버레이 유형. 오버레이 함수나 다듬기가 표시되는지 여부를 지정합니다. 다듬기 및 오버레이 함수는 항상 y 함수로 계산됩니다.

- **없음.** 오버레이가 표시되지 않습니다.
- **다듬기.** LOESS(locally weighted iterative robust least squares regression)를 사용하여 계산된 다듬은 회귀선 적합을 표시합니다. 이 방법은 각각 도표 내의 작은 영역에 집중하여 일련의 회귀분석을 효과적으로 계산합니다. 이로 인해 평활 곡선을 작성하기 위해 결합된 일련의 "로컬" 회귀선이 생성됩니다.

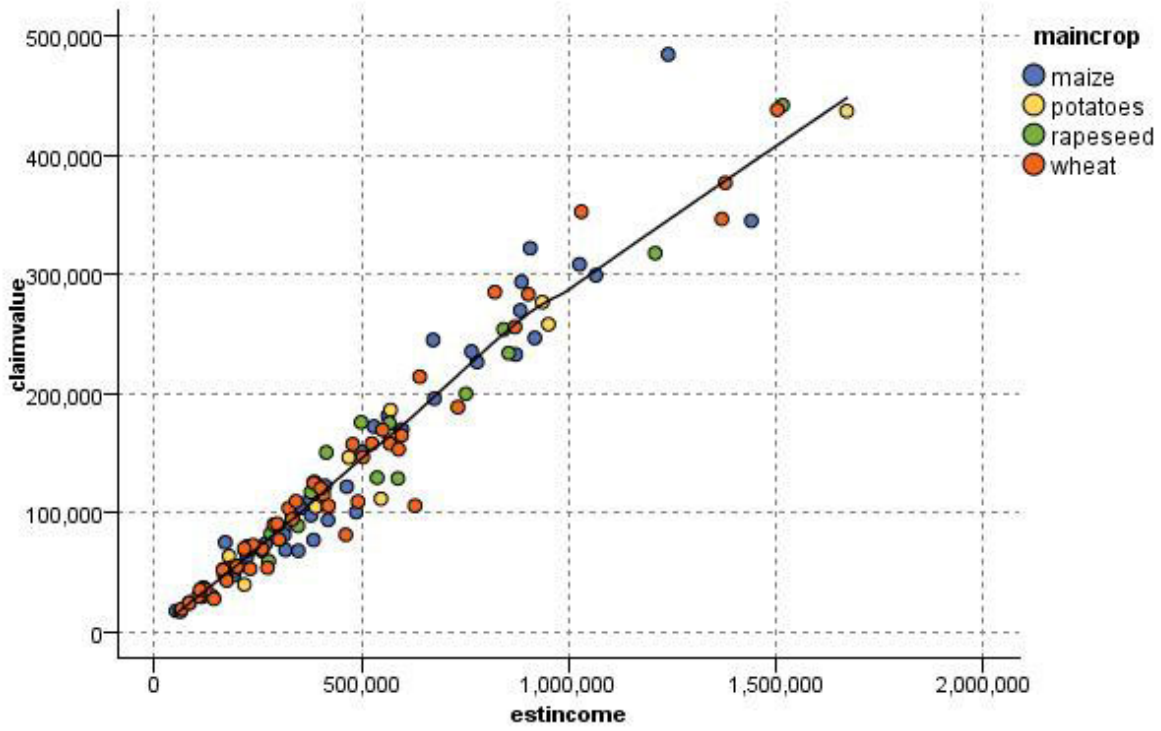


그림 22. LOESS 다듬기 오버레이로 구성

- **함수.** 실제 값과 비교하기 위해 알려진 함수를 지정하려면 선택하십시오. 예를 들어, 실제값 대 예측값을 비교하려면 오버레이로 y = x 함수를 구성할 수 있습니다. 텍스트 상자에서 **y =** 에 대한 함수를 지정하십시오. 기본 함수는 y = x이나 x 측에서 임의의 함수(이차 함수 또는 임의 표현식 등)를 지정할 수 있습니다.

  참고: 오버레이 함수는 패널 또는 애니메이션 그래프에 대해 사용할 수 없습니다.

일단 도표에 대한 옵션을 설정한 후에는 **실행**을 클릭하여 대화 상자에서 직접 도표를 실행할 수 있습니다. 단, 구간화, X 모드 및 유형 등의 추가 지정 사항에 대한 옵션을 사용해야 하는 경우도 있습니다.

도표 옵션 탭

스타일. 도표 스타일에 대해 **점** 또는 **선**을 선택하십시오. **선**을 선택하면 **X 모드** 제어가 활성화됩니다. **점**을 선택하면 더하기 기호(+)가 기본 점 모양으로 사용됩니다. 그래프가 작성되고 나면 점 모양을 변경하고 해당 크기를 변경할 수 있습니다.

X 모드. 선 도표의 경우 X 모드를 선택하여 선 도표의 스타일을 정의해야 합니다. **정렬**, **오버레이** 또는 **읽은 대로**를 선택하십시오. **오버레이** 또는 **읽은 대로**의 경우에는 처음 *n*개 레코드의 표본을 추출하는 데 사용되는 최대 데이터 세트 *크기*를 지정해야 합니다. 그렇지 않으면 기본값인 2,000개의 레코드가 사용됩니다.

자동 X 범위. 이 축을 따르는 데이터의 전체 값 범위를 사용하려면 선택하십시오. 지정된 **최소** 및 **최대** 값을 기반으로 값의 명시적 서브세트를 사용하려면 선택 취소하십시오. 값을 입력하거나 화살표를 사용하십시오. 빠르게 그래프를 작성할 수 있도록 기본적으로 자동 범위가 선택됩니다.

자동 Y 범위. 이 축을 따르는 데이터의 전체 값 범위를 사용하려면 선택하십시오. 지정된 **최소** 및 **최대** 값을 기반으로 값의 명시적 서브세트를 사용하려면 선택 취소하십시오. 값을 입력하거나 화살표를 사용하십시오. 빠르게 그래프를 작성할 수 있도록 기본적으로 자동 범위가 선택됩니다.

자동 Z 범위. 도표 탭에서 3차원 그래프가 지정되는 경우에만 사용됩니다. 이 축을 따르는 데이터의 전체 값 범위를 사용하려면 선택하십시오. 지정된 **최소** 및 **최대** 값을 기반으로 값의 명시적 서브세트를 사용하려면 선택 취소하십시오. 값을 입력하거나 화살표를 사용하십시오. 빠르게 그래프를 작성할 수 있도록 기본적으로 자동 범위가 선택됩니다.

지터. 변동으로도 알려져 있는 지터는 다수의 값이 반복되는 데이터 세트의 포인트 도표에 유용합니다. 더 명확한 값 분포를 보기 위해 지터를 사용하여 실제 값 주위에 무작위로 점을 분포시킬 수 있습니다.

*IBM SPSS Modeler*의 이전 버전 사용자에 대한 참고: 도표에서 사용되는 지터 값은 이 IBM SPSS Modeler 릴리스에서 다른 메트릭을 사용합니다. 이전 버전에서는 값이 실제 숫자였지만 이제는 프레임 *크기*의 비율입니다. 이는 이전 스트림의 변동 값이 지나치게 커질 수 있음을 의미합니다. 이 릴리스의 경우 0(영)이 아닌 변동 값은 값 0.2로 변환됩니다.

도표화할 최대 레코드 수. 큰 데이터 세트를 도표화하는 방법을 지정하십시오. 최대 데이터 세트 *크기*를 지정하거나 기본값인 2,000개의 레코드를 사용할 수 있습니다. **구간** 또는 **표본** 옵션을 선택하면 큰 데이터 세트에 대해 성능이 향상됩니다. 또는 **모든 데이터 사용**을 선택하여 모든 데이터 점을 도표화하도록 선택할 수 있지만 소프트웨어의 성능이 급격하게 저하될 수 있다는 점에 유의해야 합니다.

참고: X 모드가 **오버레이** 또는 **읽은 대로**로 설정된 경우 이 옵션은 사용 안함으로 설정되고 처음 *n*개 레코드만 사용됩니다.

- **구간.** 데이터 세트에 지정된 수의 레코드보다 많은 레코드가 포함되어 있는 경우 구간화를 사용으로 설정하려면 선택하십시오. 구간화는 실제로 도표화하기 전에 그래프를 세분화된 눈금으로 나누고 각각의 눈금 셀에 표시되는 점의 수를 계수합니다. 최종 그래프에서는 구간 중심값(구간에 있는 모든 점 위치의 평균)에서 셀당 하나의 점이 도표화됩니다. 도표화된 기호의 *크기*는 해당 영역에 있는 점의 수를 표시합니다(*크기*를 오버레이로 사용하지 않은 경우). 중심값 및 *크기*를 사용하여 점의 수를 나타내면 밀집된 영역(구별되지 않는

색상 덩어리)에서 도표가 겹치는 걸 방지하고 기호 아티팩트(밀도의 인위적인 패턴)를 줄이므로 구간화된 도표는 큰 데이터 세트를 나타내는 우수한 방법이 됩니다. 기호 아티팩트는 원시 데이터에 없는 밀집된 영역을 생성하는 방식으로 특정 기호(특히, 더하기 기호 [+])가 충돌할 때 발생합니다.

- **표본.** 텍스트 필드에서 입력한 레코드 수까지 무작위로 데이터의 표본을 추출하려면 선택하십시오. 기본값은 2,000입니다.

도표 모양 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

제목. 그래프 제목으로 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

캡션. 그래프 캡션에 사용할 텍스트를 입력하십시오.

X 레이블. 자동으로 생성된 *x*축(가로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

Y 레이블. 자동으로 생성된 *y*축(세로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

Z 레이블. 3-D 그래프에만 사용할 수 있습니다(자동으로 생성된 *z*축 레이블의 경우 또는 **사용자 정의**를 선택하여 사용자 정의 레이블을 지정하는 경우는 제외).

눈금선 표시. 기본적으로 선택되는 이 옵션은 더 쉽게 영역 및 밴드 절사 지점을 결정할 수 있게 하는 눈금선을 도표 또는 그래프 뒤에 표시합니다. 그래프 배경이 흰색인 경우가 아니면 눈금선은 항상 흰색으로 표시됩니다. 그래프 배경이 흰색이면 눈금선은 회색으로 표시됩니다.

도표 그래프 사용

도표 및 다중 도표는 본질적으로 *Y*에 대한 *X*의 도표입니다. 예를 들어, 농업 허가 신청에서 잠재적 사기를 조사하는 경우에는 신경망에 의해 추정된 수입에 대해 해당 신청에서 청구된 수입을 도표화할 수 있습니다. 작물 유형 등의 오버레이를 사용하면 청구(값 또는 번호)와 작물 유형 사이에 관계가 있는 여부가 표시됩니다.

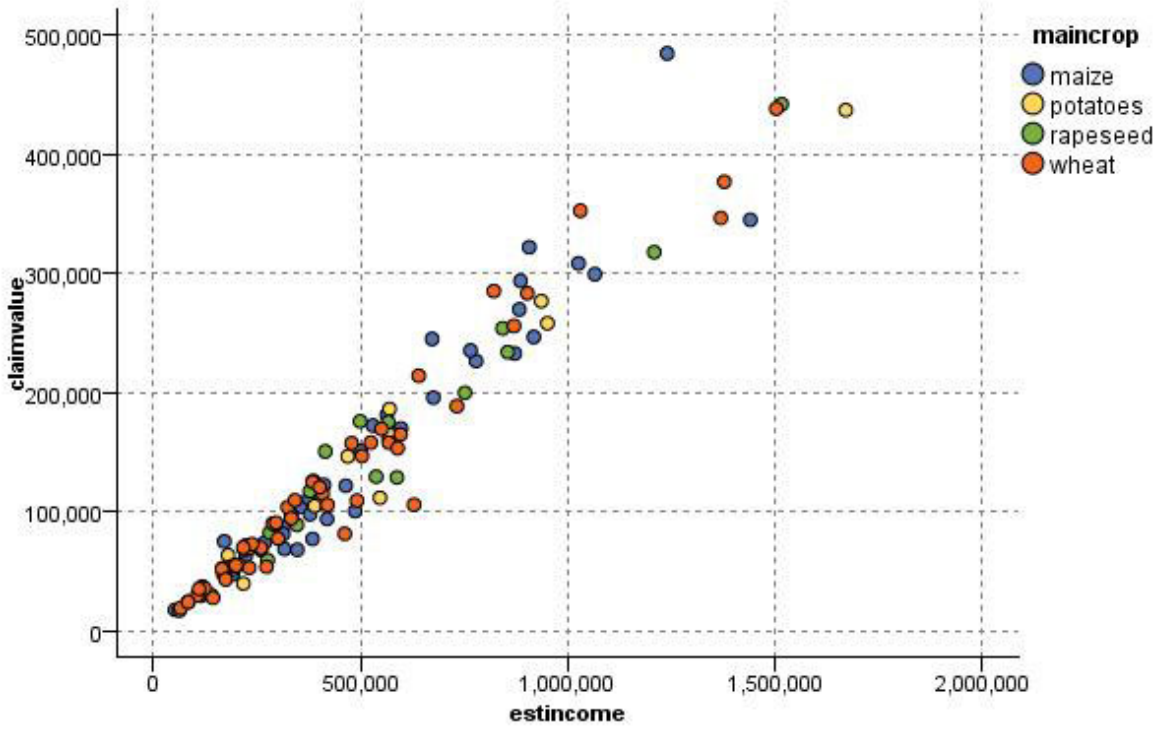


그림 23. 기본 작물 유형을 오버레이로 가진 추정된 수입과 청구 값 간 관계의 도표

도표, 다중 도표 및 평가 차트는 *X*에 대한 *Y*의 2차원 표시이므로 영역을 정의하거나 요소를 표시하거나 밴드를 그려서 이들과 쉽게 상호작용할 수 있습니다. 해당 영역, 밴드 또는 요소로 표시된 데이터에 대한 노드도 생성할 수 있습니다. 자세한 정보는 276 페이지의 『그래프 탐색』의 내용을 참조하십시오.

다중 도표 노드

다중 도표는 단일 *X* 필드 위에 다중 *Y* 필드를 표시하는 특수 유형의 도표입니다. *Y* 필드는 색상 지정된 선으로 도표화되며 각각 스타일이 **선**으로 설정되고 **X** 모드가 **정렬**로 설정된 구성 노드와 동등합니다. 다중 도표는 시간 시퀀스 데이터를 가지고 있을 때 시간 경과에 따른 여러 변수의 변동을 탐색하려는 경우에 유용합니다.

다중 도표 도표 탭

X 필드. 목록에서 수평 *x*축에 표시할 필드를 선택하십시오.

Y 필드. *X* 필드 값의 범위에 대해 표시할 하나 이상의 필드를 목록에서 선택하십시오. 다중 필드를 선택하려면 필드 선택기 단추를 사용하십시오. 목록에서 필드를 제거하려면 삭제 단추를 클릭하십시오.

오버레이. 여러 가지 방식으로 데이터 값에 대한 범주를 표시할 수 있습니다. 예를 들어, 애니메이션 오버레이를 사용하여 데이터의 값 각각에 대한 다중 도표를 표시할 수 있습니다. 이는 10개보다 많은 범주가 포함된 세트에 대해 유용합니다. 15개보다 많은 범주를 가진 세트에 사용된 경우에는 성능이 저하될 수 있습니다. 자세한 정보는 194 페이지의 『모양, 오버레이, 패널 및 애니메이션』의 내용을 참조하십시오.

정규화. 그래프에 표시할 모든 *Y* 값의 척도를 범위 0–1로 지정하려면 선택하십시오. 정규화는 각 시리즈에 대한 값의 범위에 있는 차이로 인해 모호할 수 있는 선 사이의 관계를 탐색하는 데 도움이 되며 동일한 그래프에서 여러 선을 도표화하거나 패널에서 나란히 도표를 비교할 때 권장됩니다. (모든 데이터 값이 비슷한 범위 내에 속하는 경우에는 정규화가 필요하지 않습니다.)

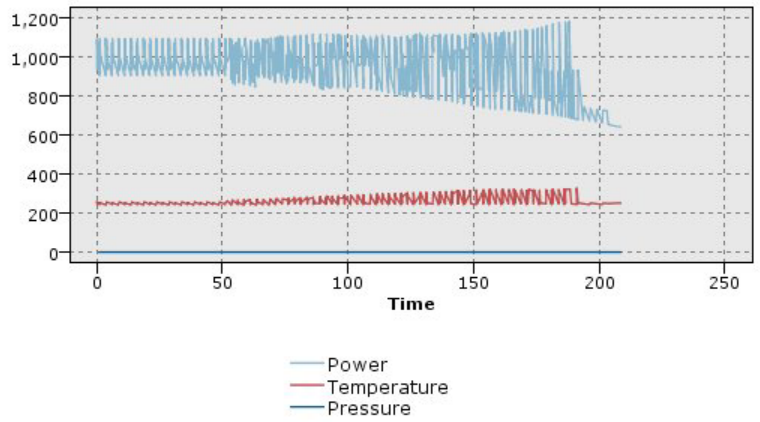


그림 24. 시간 경과에 따른 발전소 변동을 표시하는 표준 다중 도표(정규화를 수행하지 않으면 전압에 대한 도표를 볼 수 없음)

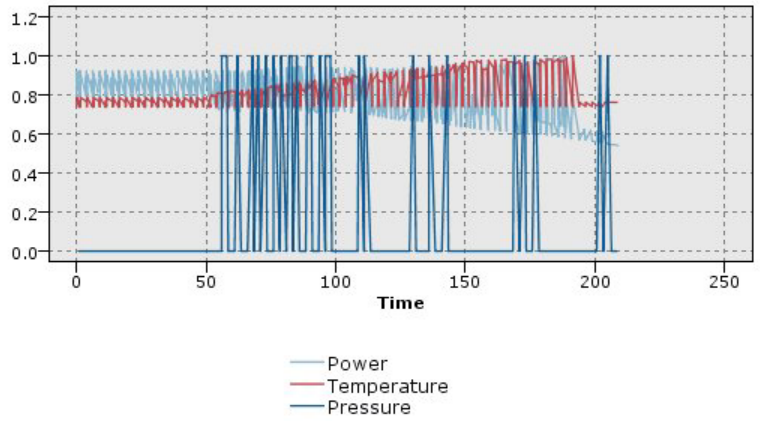


그림 25. 전압에 대한 도표를 표시하는 정규화된 다중 도표

오버레이 함수. 실제 값과 비교할 알려진 함수를 지정하려면 선택하십시오. 예를 들어, 실제 값과 예측값을 비교하려면 y = x 함수를 오버레이로 도표화하십시오. 텍스트 상자에서 **y =**에 대한 함수를 지정하십시오. 기본 함수는 y = x이지만 x와 관련하여 모든 유형의 함수(예: 2차 함수 또는 임의의 표현식)를 지정할 수 있습니다.

참고: 오버레이 함수는 패널 또는 애니메이션 그래프에는 사용할 수 없습니다.

레코드 수가 다음보다 많은 경우. 큰 데이터 세트를 도표화하는 방법을 지정하십시오. 최대 데이터 세트 *크기*를 지정하거나 기본값인 2,000개의 점을 사용할 수 있습니다. **구간** 또는 **표본** 옵션을 선택하면 큰 데이터 세트에 대해 성능이 개선됩니다. 또는 **모든 데이터 사용**을 선택하여 모든 데이터 점을 도표화하도록 선택할 수 있지만 소프트웨어의 성능이 급격하게 저하될 수 있다는 점에 유의해야 합니다.

참고: X 모드가 **오버레이** 또는 **읽은 대로**로 설정된 경우 이 옵션은 사용 안함으로 설정되고 처음 *n*개 레코드만 사용됩니다.

- **구간.** 데이터 세트에 지정된 수의 레코드보다 많은 레코드가 포함되어 있는 경우 구간화를 사용으로 설정하려면 선택하십시오. 구간화는 실제로 도표화하기 전에 그래프를 세분화된 눈금으로 나누고 각각의 눈금 셀에 표시되는 연결의 수를 계수합니다. 최종 그래프에서는 구간 중심값(구간에 있는 모든 연결 점의 평균)에서 셀당 하나의 연결이 사용됩니다.
- **표본.** 지정된 수의 레코드로 데이터에서 무작위로 표본을 추출하려면 선택하십시오.

다중 도표 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

제목. 그래프 제목으로 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

캡션. 그래프 캡션에 사용할 텍스트를 입력하십시오.

X 레이블. 자동으로 생성된 *x*축(가로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

Y 레이블. 자동으로 생성된 *y*축(세로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

눈금선 표시. 기본적으로 선택되는 이 옵션은 더 쉽게 영역 및 밴드 절사 지점을 결정할 수 있게 하는 눈금선을 도표 또는 그래프 뒤에 표시합니다. 그래프 배경이 흰색인 경우가 아니면 눈금선은 항상 흰색으로 표시됩니다. 그래프 배경이 흰색이면 눈금선은 회색으로 표시됩니다.

다중 도표 그래프 사용

도표 및 다중 도표는 본질적으로 *Y*에 대한 *X*의 도표입니다. 예를 들어, 농업 허가 신청에서 잠재적 사기를 조사하는 경우에는 신경망에 의해 추정된 수입에 대해 해당 신청에서 청구된 수입을 도표화할 수 있습니다. 작물 유형 등의 오버레이를 사용하면 청구(값 또는 번호)와 작물 유형 사이에 관계가 있는 여부가 표시됩니다.

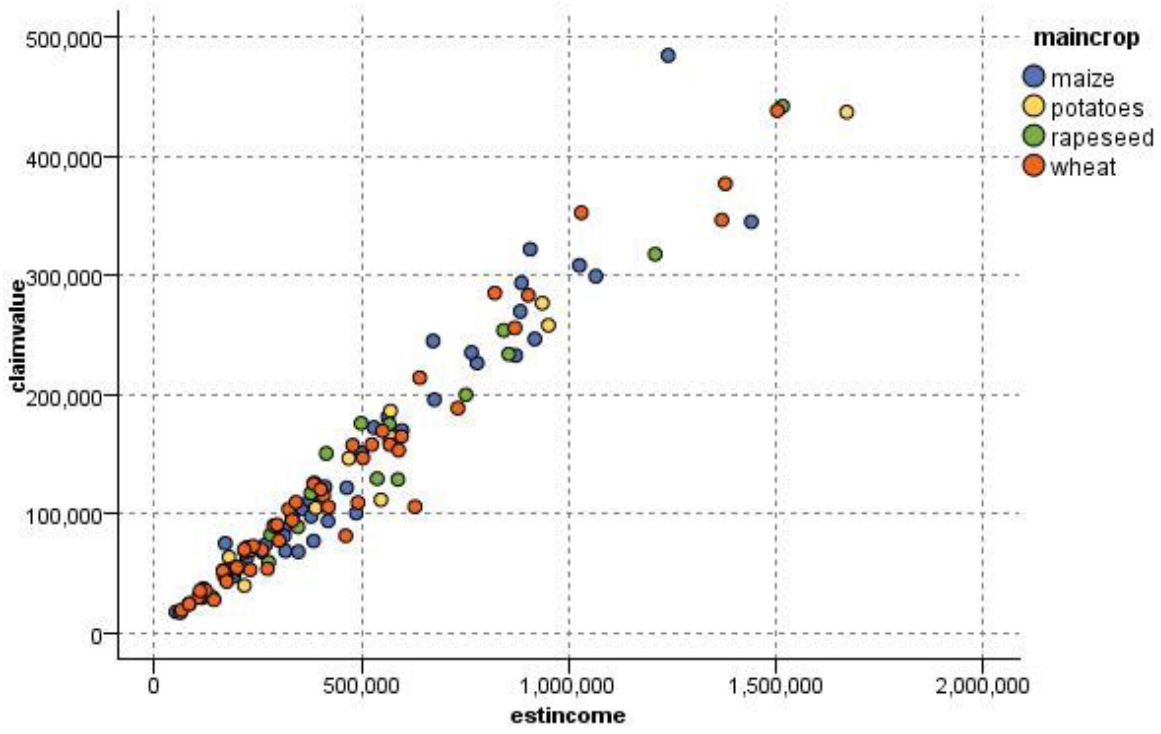


그림 26. 기본 작물 유형을 오버레이로 가진 추정된 수입과 청구 값 간 관계의 도표

도표, 다중 도표 및 평가 차트는 *X*에 대한 *Y*의 2차원 표시이므로 영역을 정의하거나 요소를 표시하거나 밴드를 그려서 이들과 쉽게 상호작용할 수 있습니다. 해당 영역, 밴드 또는 요소로 표시된 데이터에 대한 노드도 생성할 수 있습니다. 자세한 정보는 276 페이지의 『그래프 탐색』의 내용을 참조하십시오.

시간 구성 노드

시간 구성 노드에서는 시간에 따라 구성된 하나 이상의 시계열을 볼 수 있습니다. 구성하는 계열은 숫자 값을 포함해야 하며 주기가 균일한 시간 범위에서 발생한다고 가정합니다. 일반적으로, 시간 구성 노드 전에 시간 간격 노드를 사용하여 *TimeLabel* 필드를 작성합니다. 이 필드는 그래프에 있는 *x*축의 레이블을 지정하는 데 기본적으로 사용됩니다. 자세한 정보는 시간 간격 노드(더 이상 사용되지 않음)의 내용을 참조하십시오.

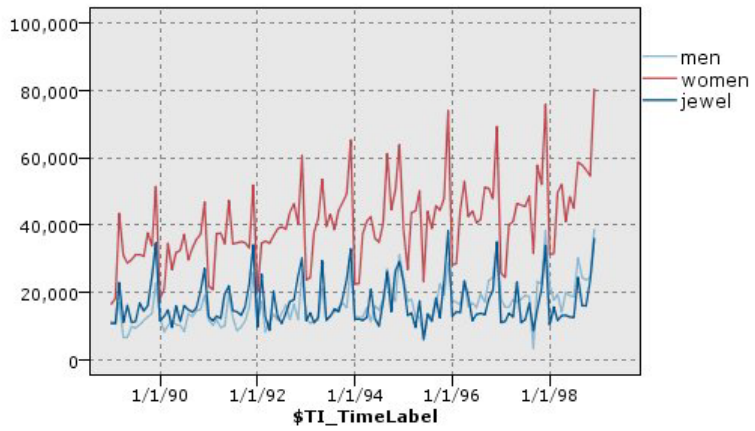


그림 27. 시간에 따른 남성 및 여성용 의류와 장신구 판매량 구성

개입 및 이벤트 작성

컨텍스트 메뉴에서 파생(플래그 또는 명목) 노드를 생성하여 시간 구성에서 이벤트 및 개입 필드를 작성할 수 있습니다. 예를 들어, 철도 파업의 경우 이벤트 필드를 작성할 수 있으며, 이벤트가 발생했으면 드라이브 상태는 True이고 그렇지 않으면 False입니다. 개입 필드의 경우, 가격 상승을 예로 들면, 파생 개수를 사용하여 상승 날짜를 식별할 수 있습니다(이전 가격에는 0, 새 가격에는 1 사용). 자세한 정보는 157 페이지의 『파생 노드』의 내용을 참조하십시오.

시간 구성 탭

도표. 시계열 데이터를 구성하는 방법을 선택할 수 있습니다.

- **선택된 계열.** 선택된 시계열의 값을 구성합니다. 신뢰구간을 구성할 때 이 옵션을 선택하는 경우 **정규화** 선택란을 선택 취소하십시오.
- **선택된 시계열 모델.** 시계열 모델과 함께 사용되는 경우, 이 옵션은 하나 이상의 선택된 시계열에 대한 모든 관련 필드(실제 및 예측 값과 신뢰구간)를 구성합니다. 이 옵션을 사용하면 대화 상자의 다른 옵션 중 일부가 사용되지 않습니다. 이 옵션은 신뢰구간을 구성하는 경우에 선호되는 옵션입니다.

계열. 구성할 시계열 데이터를 포함하는 하나 이상의 필드를 선택하십시오. 데이터는 숫자여야 합니다.

X축 레이블. 도표에서 *x*축의 레이블로 사용할 단일 필드 또는 기본 레이블을 선택하십시오. 기본값을 선택하는 경우, 시스템은 업스트림 시간 간격 노드에서 작성된 TimeLabel 필드 또는 순차 정수(시간 간격 노드가 없는 경우)를 사용합니다. 자세한 정보는 시간 간격 노드(더 이상 사용되지 않음)의 내용을 참조하십시오.

개별 패널에 계열 표시. 각 계열을 개별 패널에 표시할지 여부를 지정합니다. 각 계열을 개별 패널에 표시하지 않는 경우, 모든 시계열이 동일한 그래프에서 구성되고 평활기를 사용할 수 없습니다. 동일한 그래프에서 모든 시계열을 구성하면 각 계열이 다른 색상으로 표시됩니다.

정규화. 그래프에 표시되도록 모든 *Y* 값을 범위 0–1로 스케일링하려면 이 옵션을 선택하십시오. 정규화는 각 시리즈에 대한 값의 범위에 있는 차이로 인해 모호할 수 있는 선 사이의 관계를 탐색하는 데 도움이 되며 동

일한 그래프에서 여러 선을 도표화하거나 패널에서 나란히 도표를 비교할 때 권장됩니다. (모든 데이터 값이 유사한 범위에 속하는 경우에는 정규화가 필요 없습니다.)

표시. 도표에 표시할 하나 이상의 요소를 선택하십시오. 선, 점 및 (LOESS) 평활기 중에서 선택할 수 있습니다. 평활기는 계열을 개별 패널에 표시하는 경우에만 사용할 수 있습니다. 기본적으로 선 요소가 선택됩니다. 그래프 노드를 실행하기 전에 하나 이상의 도표 요소를 선택해야 합니다. 그러지 않을 경우, 시스템에서 구성할 요소를 선택하지 않았음을 알리는 오류를 리턴합니다.

레코드 제한. 구성할 레코드 수를 제한하려면 이 옵션을 선택하십시오. **구성할 최대 레코드 수** 옵션에서 구성할 레코드 수(데이터 파일의 시작 부분에서 읽혀지는)를 지정하십시오. 기본적으로 이 수는 2,000으로 설정됩니다. 데이터 파일에 마지막 *n*개의 레코드를 구성하려면 이 노드 전에 정렬 노드를 사용하여 레코드를 시간을 기준으로 내림차순으로 배열할 수 있습니다.

시간 구성 모양 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

제목. 그래프 제목으로 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

캡션. 그래프 캡션에 사용할 텍스트를 입력하십시오.

X 레이블. 자동으로 생성된 *x*축(가로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

Y 레이블. 자동으로 생성된 *y*축(세로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

눈금선 표시. 기본적으로 선택되는 이 옵션은 더 쉽게 영역 및 밴드 절사 지점을 결정할 수 있게 하는 눈금선을 도표 또는 그래프 뒤에 표시합니다. 그래프 배경이 흰색인 경우가 아니면 눈금선은 항상 흰색으로 표시됩니다. 그래프 배경이 흰색이면 눈금선은 회색으로 표시됩니다.

레이아웃. 시간 도표의 경우에만 시간 값이 가로 축과 세로 축 중 어느 축을 따라 도표화되는지를 지정할 수 있습니다.

시간 도표 그래프 사용

시간 도표 그래프를 작성하면 그래프 표시를 조정하고 추가 분석을 위해 노드를 생성하는 여러 옵션을 사용할 수 있습니다. 자세한 정보는 276 페이지의 『그래프 탐색』의 내용을 참조하십시오.

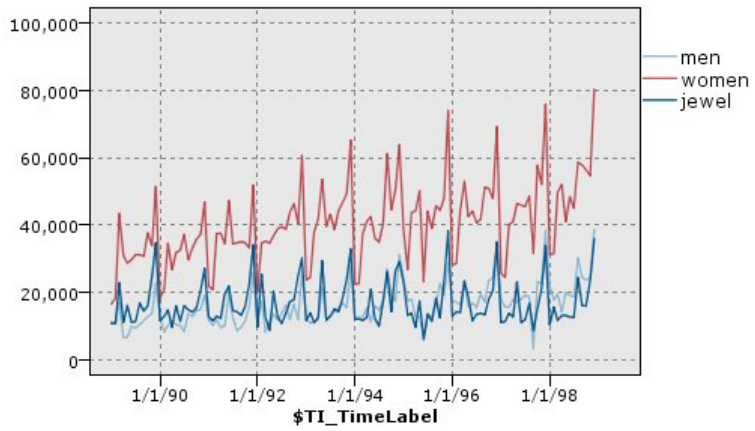


그림 28. 시간에 따른 남성 및 여성용 의류와 장신구 판매량 구성

시간 구성을 작성하고 밴드를 정의하며 결과를 검토한 후에는 메뉴 생성 및 컨텍스트 메뉴의 옵션을 사용하여 선택 또는 파생 노드를 작성할 수 있습니다. 자세한 정보는 284 페이지의 『그래프에서 노드 생성』의 내용을 참조하십시오.

분포 노드

분포 그래프 또는 테이블은 데이터 세트에서 담보 유형 또는 성별 등의 숫자가 아닌 기호 값의 발생을 표시합니다. 분포 노드는 일반적으로 모델을 작성하기 전에 균형 노드를 사용하여 수정할 수 있는 데이터의 불균형을 표시하는 데 사용됩니다. 분포 그래프 또는 테이블 창의 생성 메뉴를 사용하여 균형 노드를 자동으로 생성할 수 있습니다.

그래프보드 노드를 사용하여 개수 그래프의 막대를 생성할 수도 있습니다. 하지만 이 노드에서 더 많은 옵션을 선택할 수 있습니다. 자세한 정보는 204 페이지의 『사용 가능한 내장 그래프보드 시각화 유형』의 내용을 참조하십시오.

참고: 숫자 값의 발생을 표시하려면 히스토그램 노드를 사용해야 합니다.

분포 도표 탭

도표. 분포 유형을 선택하십시오. 선택된 필드의 분포를 표시하려면 **선택된 필드**를 선택하십시오. 데이터 세트에서 플래그 필드에 대한 true 값의 분포를 표시하려면 **모든 플래그(true 값)**를 선택하십시오.

필드. 값의 분포를 표시할 명목 또는 플래그 필드를 선택하십시오. 명시적으로 숫자로 설정되지 않은 필드만 목록에 표시됩니다.

오버레이. 지정된 필드의 각 값에서 해당 값의 분포를 보여주는 색상 오버레이로 사용할 명목 또는 플래그 필드를 선택하십시오. 예를 들어, 마케팅 캠페인 반응(*pep*)을 하위의 수에 대한 오버레이(*children*)로 사용하여 패밀리 크기별 응답성을 보여줄 수 있습니다. 자세한 정보는 194 페이지의 『모양, 오버레이, 패널 및 애니메이션』의 내용을 참조하십시오.

색상별 정규화. 모든 막대가 그래프의 전체 너비를 차지하도록 막대의 축척을 지정하려면 선택하십시오. 오버레이 값은 각 막대의 비율과 동일하므로 더 쉽게 범주에서 비교를 수행할 수 있습니다.

정렬. 분포 그래프에서 값을 표시하는 데 사용되는 방법을 선택하십시오. 알파벳순을 사용하려면 **알파벳**을 선택하고 발생의 내림차순으로 값을 나열하려면 **개수별**을 선택하십시오.

비례 척도. 개수가 가장 큰 값이 도표의 전체 너비를 채우도록 값 분포의 축척을 지정하려면 선택하십시오. 다른 모든 막대는 이 값에 대해 축척이 지정됩니다. 이 옵션을 선택 취소하면 각 값의 총 수에 따라 막대의 축척이 지정됩니다.

분포 모양 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

제목. 그래프 제목으로 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

캡션. 그래프 캡션에 사용할 텍스트를 입력하십시오.

X 레이블. 자동으로 생성된 *x*축(가로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

Y 레이블. 자동으로 생성된 *y*축(세로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

눈금선 표시. 기본적으로 선택되는 이 옵션은 더 쉽게 영역 및 밴드 절사 지점을 결정할 수 있게 하는 눈금선을 도표 또는 그래프 뒤에 표시합니다. 그래프 배경이 흰색인 경우가 아니면 눈금선은 항상 흰색으로 표시됩니다. 그래프 배경이 흰색이면 눈금선은 회색으로 표시됩니다.

분포 노드 사용

분포 노드는 데이터 세트에서 기호 값의 분포를 표시하는 데 사용됩니다. 분포 노드는 데이터를 탐색하고 불균형을 정정하기 위해 조작 노드 전에 자주 사용됩니다. 예를 들어, 자녀가 없는 응답자의 인스턴스가 다른 유형의 응답자보다 훨씬 자주 발생하는 경우에는 향후 데이터 마이닝 조작에서 더 유용한 규칙이 생성될 수 있도록 이 인스턴스를 줄이길 원할 수 있습니다. 분포 노드를 사용하면 이러한 불균형에 대한 의사결정을 검토하고 작성하는 데 도움이 됩니다.

분포 노드는 데이터를 분석하는 데 필요한 그래프와 테이블을 모두 생성한다는 점에서 특이합니다.

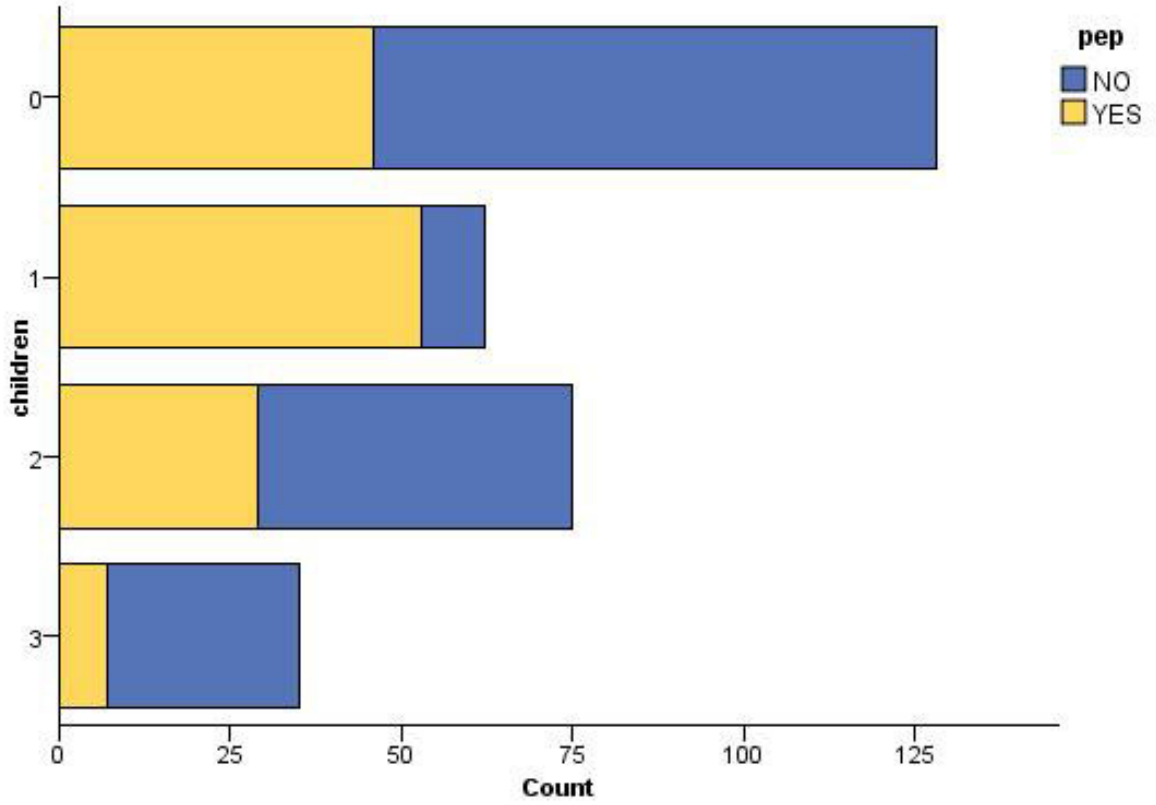


그림 29. 마케팅 캠페인에 응답한 자녀가 있거나 없는 사람의 수를 표시하는 분포 그래프

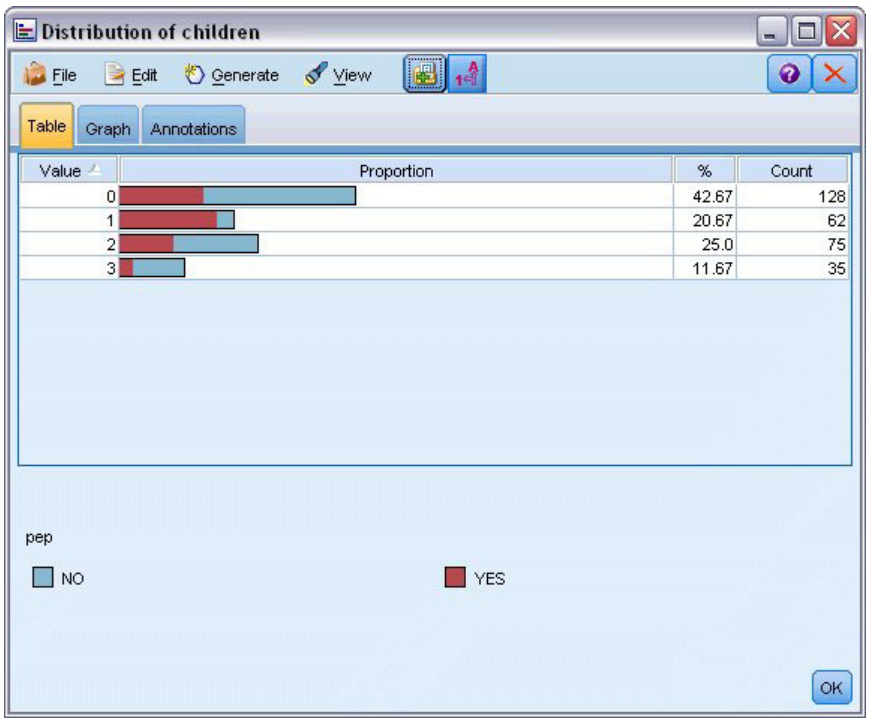


그림 30. 마케팅 캠페인에 응답한 자녀가 있거나 없는 사람의 비율을 표시하는 분포 테이블

분포 테이블 및 그래프를 작성하고 결과를 검토한 후에는 메뉴의 옵션을 사용하여 값을 그룹화하고 값을 복사하고 데이터 준비를 위해 다수의 노드를 생성할 수 있습니다. 또한 MS Word 또는 MS PowerPoint 등의 기타 애플리케이션에서 사용하기 위해 그래프 및 테이블 정보를 복사하거나 내보낼 수 있습니다. 자세한 정보는 301 페이지의 『그래프 인쇄, 저장, 복사 및 내보내기』의 내용을 참조하십시오.

분포 테이블에서 값을 선택하고 복사하려면 다음을 수행하십시오.

1. 마우스 단추를 클릭한 상태로 행 위로 끌어서 값 세트를 선택하십시오. 편집 메뉴를 사용하여 값을 **모두 선택**할 수도 있습니다.
2. 편집 메뉴에서 **테이블 복사** 또는 **테이블 복사(필드 이름 포함)**를 선택하십시오.
3. 클립보드 또는 원하는 애플리케이션에 붙여넣으십시오.

   참고: 막대는 직접 복사되지 않습니다. 대신 테이블 값이 복사됩니다. 이는 오버레이된 값은 복사된 테이블에 표시되지 않음을 의미합니다.

분포 테이블에서 값을 그룹화하려면 다음을 수행하십시오.

1. Ctrl+클릭 방법을 사용하여 그룹화할 값을 선택하십시오.
2. 편집 메뉴에서 **그룹**을 선택하십시오.

참고: 값을 그룹화하고 그룹 해제하면 그래프 탭의 그래프가 자동으로 다시 그려져 변경사항이 표시됩니다.

다음과 같은 작업도 수행할 수 있습니다.

- 분포 목록에서 그룹 이름을 선택한 후 편집 메뉴에서 **그룹 해제**를 선택하여 그룹 해제

- 분포 목록에서 그룹 이름을 선택한 후 편집 메뉴에서 **그룹 편집**을 선택하여 그룹 편집. 이 작업을 수행하면 값을 그룹으로 전환하거나 그룹에서 전환할 수 있는 대화 상자가 열립니다.

메뉴 생성 옵션

생성 메뉴의 옵션을 사용하여 데이터의 서브세트를 선택하거나 플래그 필드를 파생시키거나 값을 재그룹화하거나 값을 재분류하거나 그래프 또는 테이블에서 데이터의 균형을 맞출 수 있습니다. 이 조작은 데이터 준비 노드를 생성하여 스트림 캔버스에 배치합니다. 생성된 노드를 사용하려면 해당 노드를 기존 스트림에 연결하십시오. 자세한 정보는 284 페이지의 『그래프에서 노드 생성』의 내용을 참조하십시오.

히스토그램 노드

히스토그램 노드는 숫자 필드에 대한 값의 발생을 표시합니다. 히스토그램 노드는 조작 및 모델 작성 전에 데이터를 탐색하는 데 사용될 수도 있습니다. 분포 노드와 마찬가지로 히스토그램 노드는 데이터의 불균형을 표시하는 데 자주 사용됩니다. 그래프보드 노드를 사용하여 히스토그램을 생성할 수도 있지만 이 노드에서 더 많은 옵션을 선택할 수 있습니다. 자세한 정보는 204 페이지의 『사용 가능한 내장 그래프보드 시각화 유형』의 내용을 참조하십시오.

참고: 기호 필드에 대한 값의 발생을 표시하려면 분포 노드를 사용해야 합니다.

히스토그램 도표 탭

필드. 값의 분포를 표시할 숫자 필드를 선택하십시오. 명시적으로 기호(범주형)로 정의되지 않은 필드만 나열됩니다.

오버레이. 지정된 필드에 대한 값의 범주를 표시할 기호 필드를 선택하십시오. 오버레이 필드를 선택하면 히스토그램이 오버레이 필드의 다양한 범주를 나타내는 데 사용되는 색상을 가진 누적 차트로 변환됩니다. 히스토그램 노드를 사용하는 경우에는 세 가지 유형의 오버레이(색상, 패널, 애니메이션)가 있습니다. 자세한 정보는 194 페이지의 『모양, 오버레이, 패널 및 애니메이션』의 내용을 참조하십시오.

히스토그램 옵션 탭

자동 X 범위. 이 축을 따르는 데이터의 전체 값 범위를 사용하려면 선택하십시오. 지정된 **최소** 및 **최대** 값을 기반으로 값의 명시적 서브세트를 사용하려면 선택 취소하십시오. 값을 입력하거나 화살표를 사용하십시오. 빠르게 그래프를 작성할 수 있도록 기본적으로 자동 범위가 선택됩니다.

구간. 숫자별 또는 **너비별**을 선택하십시오.

- 지정된 구간의 수 및 범위에 따라 너비가 결정되는 고정된 수의 막대를 표시하려면 **숫자별**을 선택하십시오. **구간 수** 옵션에서 그래프에 사용할 구간 수를 표시하십시오. 화살표를 사용하여 수를 조정하십시오.
- 고정된 너비의 구간을 사용하여 그래프를 작성하려면 **너비별**을 선택하십시오. 구간 수는 값의 범위 및 지정된 너비에 따라 다릅니다. **구간 너비** 옵션에서 막대의 너비를 표시하십시오.

색상별 정규화. 모든 막대를 동일한 높이로 조정하여 오버레이된 값을 각 막대에서 전체 케이스의 백분율로 표시하려면 선택하십시오.

정규 곡선 표시. 데이터의 평균 및 분산을 표시하는 정규 곡선을 그래프에 추가하려면 선택하십시오.

각 색상별로 밴드 분리. 각각의 오버레이된 값을 그래프에서 별도의 밴드로 표시하려면 선택하십시오.

히스토그램 모양 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

제목. 그래프 제목으로 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

캡션. 그래프 캡션에 사용할 텍스트를 입력하십시오.

X 레이블. 자동으로 생성된 *x*축(가로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

Y 레이블. 자동으로 생성된 *y*축(세로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

눈금선 표시. 기본적으로 선택되는 이 옵션은 더 쉽게 영역 및 밴드 절사 지점을 결정할 수 있게 하는 눈금선을 도표 또는 그래프 뒤에 표시합니다. 그래프 배경이 흰색인 경우가 아니면 눈금선은 항상 흰색으로 표시됩니다. 그래프 배경이 흰색이면 눈금선은 회색으로 표시됩니다.

히스토그램 사용

히스토그램은 값 범위가 *x*축을 따라 분포하는 숫자 필드에서 값의 분포를 보여줍니다. 히스토그램은 콜렉션 그래프와 비슷하게 작동합니다. 콜렉션은 단일 필드에 대한 값의 발생 대신 *다른 숫자 필드의 값에 대해 상대적인* 하나의 숫자 필드의 값 분포를 표시합니다.

그래프를 작성하고 나면 결과를 검토하고 밴드를 정의하여 *x*축을 따라 값을 분할하거나 영역을 정의할 수 있습니다. 그래프 내에서 요소를 표시할 수도 있습니다. 자세한 정보는 276 페이지의 『그래프 탐색』의 내용을 참조하십시오.

생성 메뉴의 옵션을 통해 그래프의 데이터 또는 더 구체적으로 밴드, 영역 또는 표시된 요소 내의 데이터를 사용하여 균형, 선택 또는 파생 노드를 작성할 수 있습니다. 이 유형의 그래프는 스트림에서 사용할 그래프에서 균형 노드를 생성하여 불균형을 정정하고 데이터를 탐색하기 위해 조작 노드 앞에서 자주 사용됩니다. 파생 플래그 노드를 생성하여 각 레코드가 속하는 밴드를 표시하는 필드를 추가하거나 선택 노드를 생성하여 특정 값 범위 또는 세트 내 모든 레코드를 선택할 수도 있습니다. 이러한 조작을 통해 데이터의 특정 서브세트에 초점을 두고 추가적으로 탐색할 수 있습니다. 자세한 정보는 284 페이지의 『그래프에서 노드 생성』의 내용을 참조하십시오.

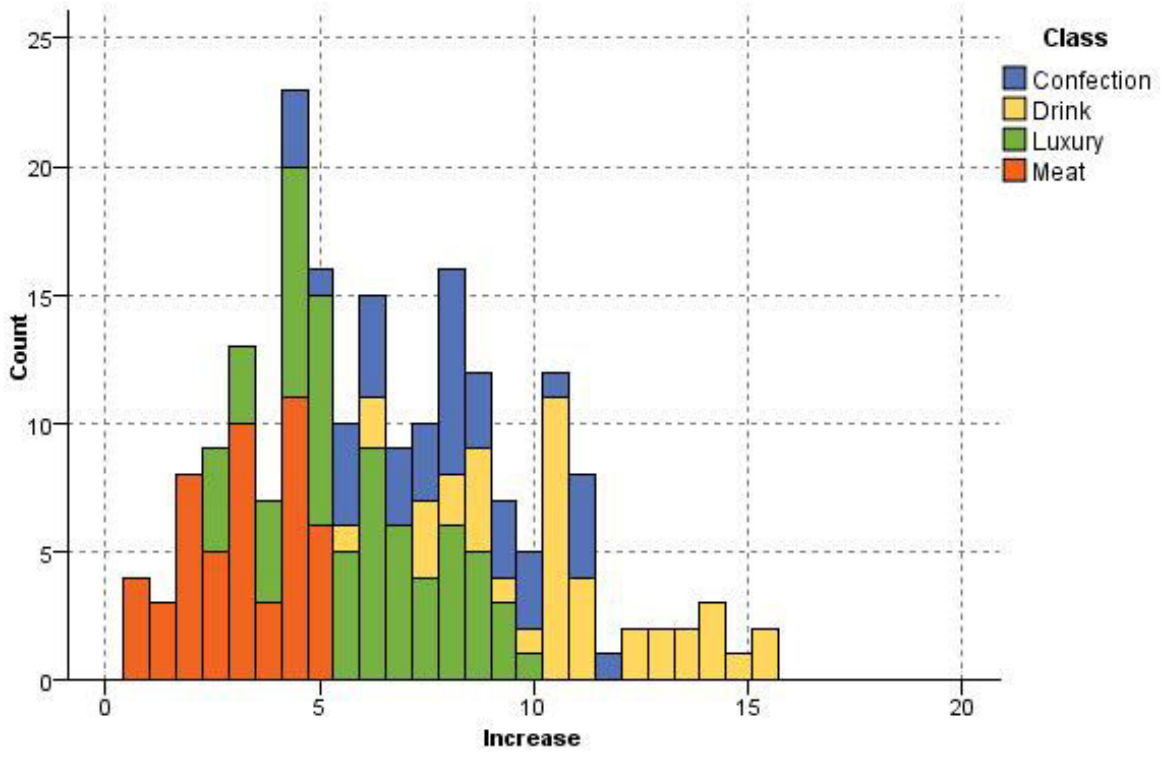


그림 31. 프로모션으로 인한 범주별 구매 증가 분포를 보여주는 히스토그램

요약도표 노드

요약도표는 단일 필드에 대한 값의 발생 대신 다른 필드의 값에 대해 상대적인 하나의 숫자 필드에 대한 값의 분포를 표시한다는 점을 제외하고 히스토그램과 비슷합니다. 요약도표는 시간 경과에 따라 값이 변경되는 변수 또는 필드를 보여주는 데 유용합니다. 3-D 그래프를 사용하여 범주별 분포를 표시하는 기호 축을 포함할 수도 있습니다. 2차원 요약도표는 사용된 오버레이와 함께 누적 막대형 차트로 표시됩니다. 자세한 정보는 194 페이지의 『모양, 오버레이, 패널 및 애니메이션』의 내용을 참조하십시오.

콜렉션 도표 탭

수집. 값을 **기간**에서 지정된 필드에 대한 값의 범위 동안 수집하고 표시할 필드를 선택하십시오. 기호로 정의되지 않은 필드만 나열됩니다.

기간. 값을 **수집**에서 지정된 필드를 표시하는 데 사용할 필드를 선택하십시오.

기준. 3차원 그래프 작성 시 사용으로 설정되면 이 옵션을 사용하여 범주별로 콜렉션 필드를 표시하는 데 사용되는 명목 또는 플래그 필드를 선택할 수 있습니다.

연산. 콜렉션 그래프의 각 막대가 표시하는 사항을 선택하십시오. 옵션으로는 **합계**, **평균**, **최대값**, **최소값**, **표준 편차**가 있습니다.

오버레이. 선택한 필드에 대한 값의 범주를 표시할 기호 필드를 선택하십시오. 오버레이 필드를 선택하면 콜렉션이 변환되고 각 범주에 대해 다양한 색상의 여러 막대가 작성됩니다. 이 노드에는 색상, 패널, 애티메이션이라는 세 가지 유형의 오버레이가 있습니다. 자세한 정보는 194 페이지의 『모양, 오버레이, 패널 및 애니메이션』의 내용을 참조하십시오.

콜렉션 옵션 탭

자동 X 범위. 이 축을 따르는 데이터의 전체 값 범위를 사용하려면 선택하십시오. 지정된 **최소** 및 **최대** 값을 기반으로 값의 명시적 서브세트를 사용하려면 선택 취소하십시오. 값을 입력하거나 화살표를 사용하십시오. 빠르게 그래프를 작성할 수 있도록 기본적으로 자동 범위가 선택됩니다.

구간. 숫자별 또는 **너비별**을 선택하십시오.

- 지정된 구간의 수 및 범위에 따라 너비가 결정되는 고정된 수의 막대를 표시하려면 **숫자별**을 선택하십시오. **구간 수** 옵션에서 그래프에 사용할 구간 수를 표시하십시오. 화살표를 사용하여 수를 조정하십시오.
- 고정된 너비의 구간을 사용하여 그래프를 작성하려면 **너비별**을 선택하십시오. 구간 수는 값의 범위 및 지정된 너비에 따라 다릅니다. **구간 너비** 옵션에서 막대의 너비를 표시하십시오.

콜렉션 모양 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

제목. 그래프 제목으로 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

캡션. 그래프 캡션에 사용할 텍스트를 입력하십시오.

기간 레이블. 자동으로 생성된 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

수집 레이블. 자동으로 생성된 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

기준 레이블. 자동으로 생성된 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

눈금선 표시. 기본적으로 선택되는 이 옵션은 더 쉽게 영역 및 밴드 절사 지점을 결정할 수 있게 하는 눈금선을 도표 또는 그래프 뒤에 표시합니다. 그래프 배경이 흰색인 경우가 아니면 눈금선은 항상 흰색으로 표시됩니다. 그래프 배경이 흰색이면 눈금선은 회색으로 표시됩니다.

다음 예제에서는 3차원 버전의 그래프에서 모양 옵션의 위치를 보여줍니다.

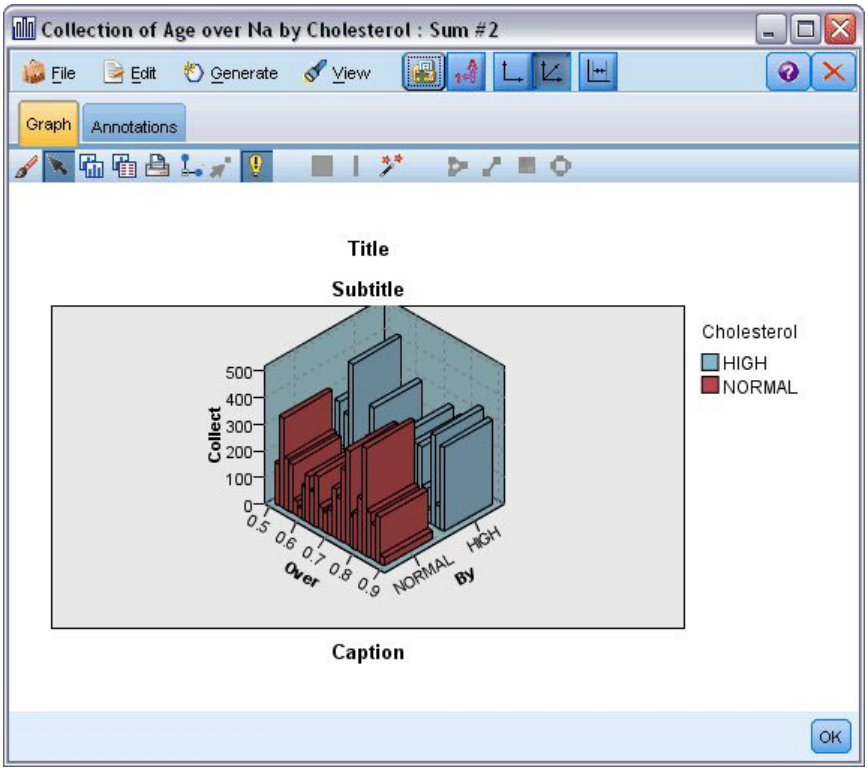


그림 32. 3차원 콜렉션 그래프에서 그래프 모양 옵션의 위치

콜렉션 그래프 사용

콜렉션은 단일 필드에 대한 값의 발생 대신 *다른 숫자 필드의 값에 대해 상대적인* 하나의 숫자 필드의 값 분포를 표시합니다. 히스토그램은 콜렉션 그래프와 비슷하게 작동합니다. 히스토그램은 값 범위가 *x*축을 따라 분포하는 숫자 필드에서 값의 분포를 보여줍니다.

그래프를 작성하고 나면 결과를 검토하고 밴드를 정의하여 *x*축을 따라 값을 분할하거나 영역을 정의할 수 있습니다. 그래프 내에서 요소를 표시할 수도 있습니다. 자세한 정보는 276 페이지의 『그래프 탐색』의 내용을 참조하십시오.

생성 메뉴의 옵션을 통해 그래프의 데이터 또는 더 구체적으로 밴드, 영역 또는 표시된 요소 내의 데이터를 사용하여 균형, 선택 또는 파생 노드를 작성할 수 있습니다. 이 유형의 그래프는 스트림에서 사용할 그래프에서 균형 노드를 생성하여 불균형을 정정하고 데이터를 탐색하기 위해 조작 노드 앞에서 자주 사용됩니다. 파생 플래그 노드를 생성하여 각 레코드가 속하는 밴드를 표시하는 필드를 추가하거나 선택 노드를 생성하여 특정 값 범위 또는 세트 내 모든 레코드를 선택할 수도 있습니다. 이러한 조작을 통해 데이터의 특정 서브세트에 초점을 두고 추가적으로 탐색할 수 있습니다. 자세한 정보는 284 페이지의 『그래프에서 노드 생성』의 내용을 참조하십시오.

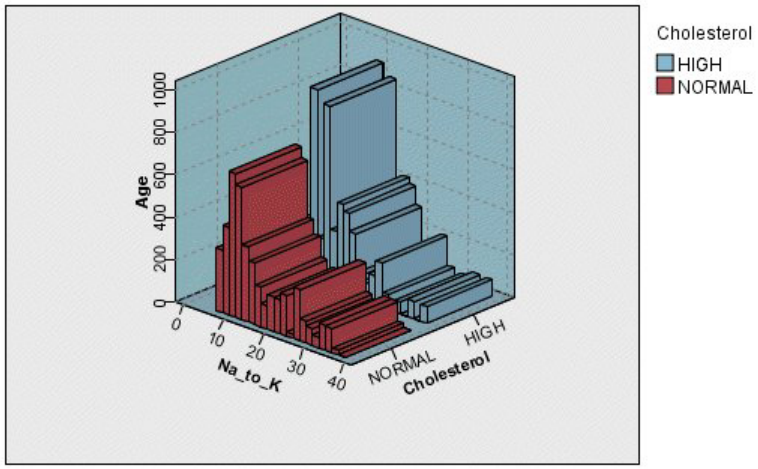


그림 33. 콜레스테롤 수준 높음 및 정상에 대해 연령에 대한 Na_to_K의 합계를 보여주는 3차원 콜렉션 그래프

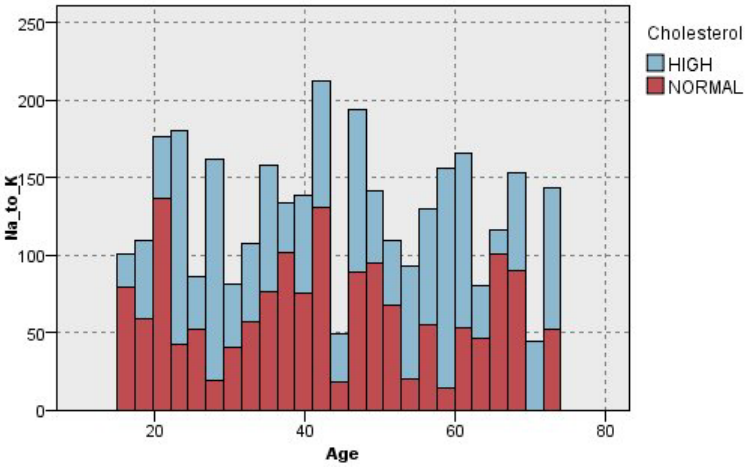


그림 34. z축은 표시되지 않지만 콜레스테롤을 색상 오버레이로 가진 콜렉션 그래프

웹 노드

웹 노드는 둘 이상의 기호 필드의 값 사이의 관계 강도를 표시합니다. 그래프는 연결 강도를 나타내는 다양한 유형의 선을 사용하여 연결을 표시합니다. 예를 들어, 웹 노드를 사용하여 전자상거래 사이트 또는 일반 소매 판매점에서의 다양한 항목 구매 간의 관계를 탐색할 수 있습니다.

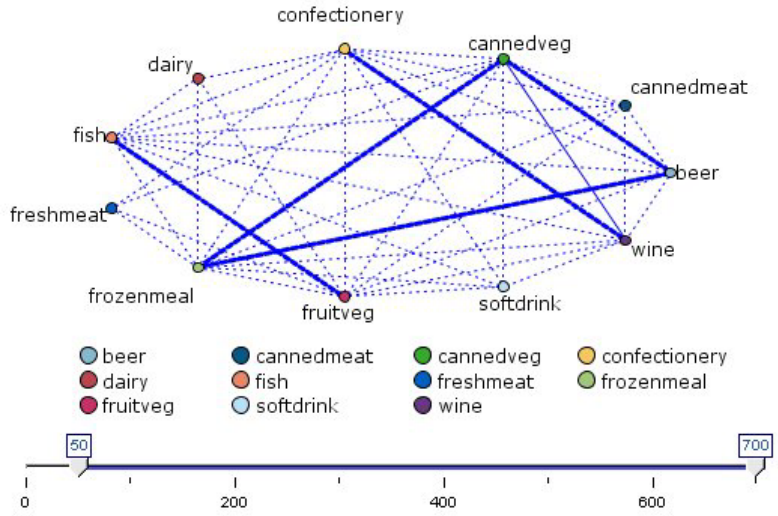


그림 35. 식료품 항목 구매 간의 관계를 보여주는 웹 그래프

방향이 있는 웹

방향이 있는 웹 노드는 기호 필드 사이의 관계 강도를 표시한다는 점에서 웹 노드와 유사합니다. 그러나, 방향이 있는 웹 그래프는 하나 이상의 시작 필드에서 하나의 대상 필드로의 연결만 표시합니다. 한 방향으로만 연결된다는 의미에서 연결은 단방향 연결입니다.

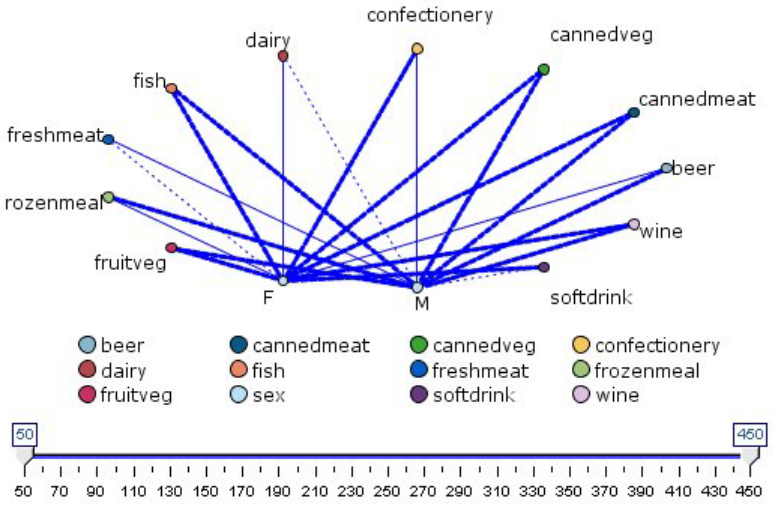


그림 36. 식료품 항목의 구매와 성별 간의 관계를 보여주는 방향이 있는 웹 그래프

웹 노드와 마찬가지로, 그래프는 연결 강도를 나타내는 다양한 유형의 선을 사용하여 연결을 표시합니다. 예를 들어, 방향이 있는 웹 노드를 사용하여 특정 구매 항목에 대한 성향과 성별 간의 관계를 탐색할 수 있습니다.

웹 구성 탭

웹. 지정된 모든 필드 간의 관계 강도를 보여주는 웹 그래프를 작성하려면 선택하십시오.

방향이 있는 웹. 하나의 필드(예: 성별 또는 종교)의 값과 여러 필드 간의 관계 강도를 보여주는 방향이 있는 웹 그래프를 작성하려면 선택하십시오. 이 옵션을 선택하면 대상 필드가 활성화되고 아래의 필드 제어가 보다 명확하게 알 수 있도록 시작 필드로 이름이 바뀝니다.

대상 필드(방향이 있는 웹에만 해당). 방향이 있는 웹에 사용되는 플래그 또는 명목 필드를 선택하십시오. 명시적으로 숫자로 설정되지 않은 필드만 나열됩니다.

필드/시작 필드. 웹 그래프를 작성할 필드를 선택하십시외. 명시적으로 숫자로 설정되지 않은 필드만 나열됩니다. 필드 선택기 단추를 사용하여 여러 필드를 선택하거나 유형별로 필드를 선택하십시오.

참고: 방향이 있는 웹의 경우, 이 제어는 시작 필드를 선택하는 데 사용됩니다.

참 플래그만 표시. 플래그 필드의 참 플래그만 표시하려면 선택하십시오. 이 옵션은 웹 디스플레이를 단순화하며 가끔 양의 값의 발생이 특히 중요한 데이터에 사용됩니다.

선 값. 드롭 다운 목록에서 임계값 유형을 선택하십시오.

- **절대값**: 각 값 쌍을 갖는 레코드 수를 기반으로 임계값을 설정합니다.
- **전체 백분율**: 링크가 웹 그래프에서 제공되는 각 값 쌍의 모든 발생의 비율로서 제공하는 절대 케이스 수를 표시합니다.
- **더 작은 필드/값의 백분율** 및 **더 큰 필드/값의 백분율**: 백분율을 평가하는 데 사용할 필드/값을 나타냅니다. 예를 들어, 100개의 레코드가 *약제* 필드에 *drugY* 값을 갖고 10개만 *BP* 필드에 *낮음* 값을 갖는다고 가정하십시오. 7개의 레코드가 *drugY* 및 *낮음* 값을 모두 갖는 경우, 이 백분율은 참조하는 필드(더 작은 필드(*BP*) 또는 더 큰 필드(*약제*))에 따라 70% 또는 7%입니다.

참고: 방향이 있는 웹 그래프의 경우, 위의 세 번째 및 네 번째 옵션을 사용할 수 없습니다. 대신, **"대상" 필드/값의 백분율** 및 **"시작" 필드/값의 백분율**을 선택할 수 있습니다.

강한 링크가 더 굵음. 기본적으로 선택됩니다. 필드 사이의 링크를 표시하는 표준 방법입니다.

약한 링크가 더 굵음. 굵은 선으로 표시되는 링크의 의미를 정반대로 바꾸려면 선택하십시오. 이 옵션은 부정 행위를 발견하거나 이상값을 조사하는 데 자주 사용됩니다.

웹 옵션 탭

웹 노드의 옵션 탭에는 출력 그래프를 사용자 정의하는 다수의 추가 옵션이 있습니다.

링크 수. 다음 옵션은 출력 그래프에 표시되는 링크 수를 제어하는 데 사용됩니다. 이 옵션 중 일부(예: **약한 링크 상한** 및 **강한 링크 하한**)는 출력 그래프 창에서도 사용 가능합니다. 또한 최종 그래프에 있는 슬라이더 제어를 사용하여 표시되는 링크 수를 조정할 수도 있습니다.

- **표시할 최대 링크 수.** 출력 그래프에 표시할 최대 링크 수를 나타내는 숫자를 지정하십시오. 화살표를 사용하여 값을 조정하십시오.
- **다음 이상의 링크만 표시.** 웹의 연결을 표시할 최소값을 나타내는 숫자를 지정하십시오. 화살표를 사용하여 값을 조정하십시오.

- **모든 링크 표시.** 최소 또는 최대 값과 상관없이 모든 링크를 표시하려면 지정하십시오. 이 옵션을 선택하면 많은 수의 필드가 있는 경우 처리 시간이 늘어날 수 있습니다.

레코드 수가 매우 적은 경우 삭제. 지원하는 레코드 수가 너무 적은 연결을 무시하려면 선택하십시오. **Min. records/line**에 숫자를 입력하여 이 옵션의 임계값을 설정하십시오.

레코드 수가 매우 많은 경우 삭제. 강력하게 지원되는 연결을 무시하려면 선택하십시오. **Max. records/line**에 숫자를 입력하십시오.

약한 링크 상한. 약한 연결(점선)과 보통 연결(실선)의 임계값을 나타내는 숫자를 지정하십시오. 이 값 아래의 모든 연결은 약한 연결로 간주됩니다.

강한 링크 하한. 강한 연결(굵은 선)과 보통 연결(실선)의 임계값을 지정하십시오. 이 값 위의 모든 연결은 강한 연결로 간주됩니다.

링크 *크기*. 링크 *크기*를 제어하는 옵션을 지정하십시오.

- **링크 *크기*가 계속 변화.** 실제 데이터 값을 기반으로 연결 강도의 변화를 반영하는 링크 *크기* 범위를 표시하려면 선택하십시오.
- **링크 *크기*가 강한/보통/약한 범주 표시.** 연결의 세 가지 강도(강함, 보통 및 약함)를 표시하려면 선택하십시오. 최종 그래프에서뿐만 아니라 위에서도 이러한 범주의 분별점을 지정할 수 있습니다.

웹 디스플레이. 웹 디스플레이의 유형을 선택하십시오.

- **원 레이아웃.** 표준 웹 디스플레이를 사용하려면 선택하십시오.
- **네트워크 레이아웃.** 가장 강한 링크를 함께 모으는 알고리즘을 사용하려면 선택하십시오. 굵은 선뿐만 아니라 공간 분화를 사용하여 강한 링크를 강조표시하는 데 사용됩니다.
- **방향이 있는 레이아웃.** 방향이 있는 웹 디스플레이를 작성하려면 선택하십시오. 방향이 있는 웹 디스플레이는 방향의 초점으로 구성 탭의 **대상 필드** 선택항목을 사용합니다.
- **눈금 레이아웃.** 간격이 일정한 눈금 패턴으로 레이아웃된 웹 디스플레이를 작성하려면 선택하십시오.

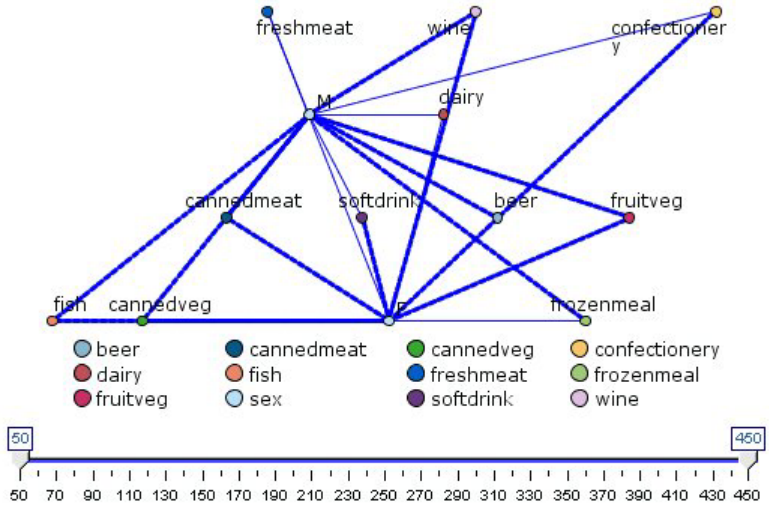


그림 37. 냉동식품 및 통조림 야채에서 다른 식료품 항목으로의 강한 연결을 보여주는 웹 그래프

웹 모양 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

제목. 그래프 제목으로 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

캡션. 그래프 캡션에 사용할 텍스트를 입력하십시오.

범례 표시. 범례가 표시되는지 여부를 지정할 수 있습니다. 많은 수의 필드를 가진 도표의 경우 범례를 숨기면 도표의 모양이 개선될 수 있습니다.

레이블을 노드로 사용. 인접 레이블을 표시하는 대신 각 노드 내부에 레이블 텍스트를 포함할 수 있습니다. 적은 수의 필드를 가진 도표의 경우에는 이를 통해 차트의 가독성이 향상될 수 있습니다.

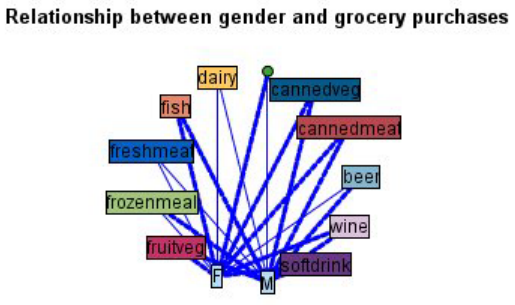


그림 38. 레이블을 노드로 표시하는 웹 그래프

웹 그래프 사용

웹 노드는 둘 이상의 기호 필드의 값 사이의 관계 강도를 표시하는 데 사용됩니다. 연결은 그래프에서 연결 강도를 나타내는 다양한 유형의 선으로 표시됩니다. 웹 노드를 사용하여, 예를 들어, 콜레스테롤 수준과 혈압 그리고 환자의 질병을 치료하는 데 효과적이었던 약제 사이의 관계를 탐색할 수 있습니다.

- 강한 연결은 굵은 선으로 표시됩니다. 이는 두 값이 강하게 관련되어 있고 추가 탐색이 필요함을 표시합니다.
- 중간 연결은 보통 굵기의 선으로 표시됩니다.
- 약한 연결은 점선으로 표시됩니다.
- 두 값 사이에 선이 표시되지 않는 경우, 이는 두 값이 결코 동일한 레코드에서 발생하지 않거나 이러한 조합이 웹 노드 대화 상자에 지정된 임계값 아래의 다수의 레코드에서 발생함을 의미합니다.

웹 노드를 작성하면 그래프 표시를 조정하고 추가 분석을 위해 노드를 생성하는 여러 옵션을 사용할 수 있습니다.

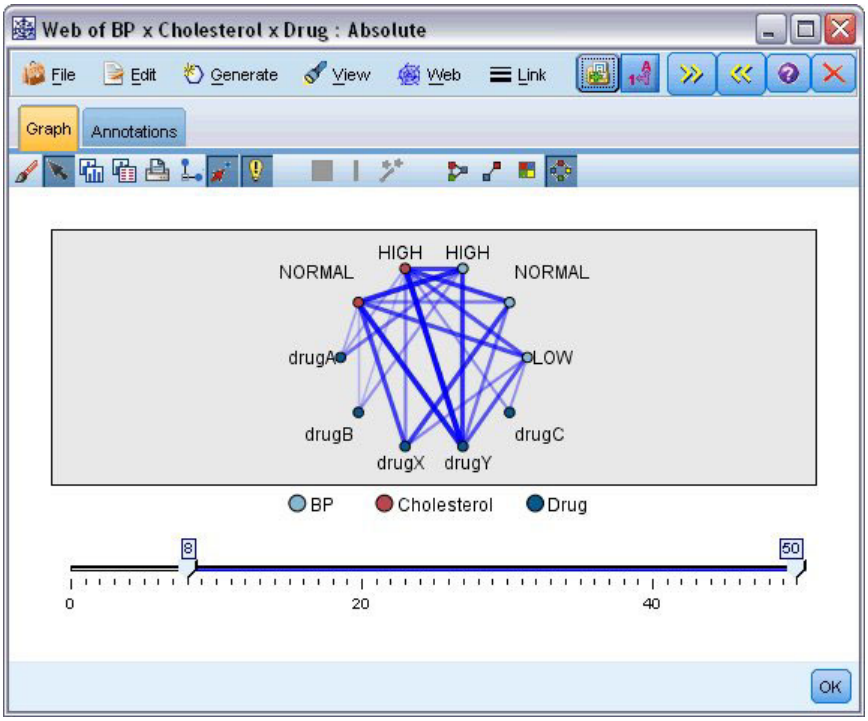


그림 39. 다수의 강한 관계(예: 정상 혈압과 DrugX 및 고콜레스테롤과 DrugY)를 나타내는 웹 그래프

웹 노드 및 방향이 있는 웹 노드 둘 다에 대해 다음을 수행할 수 있습니다.

- 웹 디스플레이의 레이아웃을 변경합니다.
- 점을 숨겨 디스플레이를 단순화합니다.
- 선 스타일을 제어하는 임계값을 변경합니다.

- 값 사이의 선을 강조표시하여 "선택된" 관계를 나타냅니다.
- 하나 이상의 "선택된" 레코드에 대한 선택 노드를 생성하거나 웹에 있는 하나 이상의 관계와 연관된 파생 플래그 노드를 생성합니다.

점을 조정하려면 다음을 수행하십시오.

- **이동**: 점에서 마우스를 클릭하여 새 위치로 끌어와 점을 이동시킵니다. 새 위치를 반영하도록 웹이 다시 그려집니다.
- **숨기기**: 웹의 점을 마우스 오른쪽 단추로 클릭하고 컨텍스트 메뉴에서 **숨기기** 또는 **숨기기 및 다시 계획**을 선택하여 점을 숨깁니다. **숨기기**는 단지 선택된 점 및 그와 연관된 선을 숨기기만 합니다. **숨기기 및 다시 계획**은 수행한 변경에 맞게 웹을 다시 그립니다. 수동 이동이 수행되지 않습니다.
- **표시**: 그래프 창의 웹 메뉴에서 **모두 표시** 또는 **모두 표시 및 다시 계획**을 선택하여 숨겨진 모든 점을 표시합니다. **모두 표시 및 다시 계획**을 선택하면 이전에 숨겨진 모든 점과 해당 연결을 포함하도록 웹이 다시 그려집니다.

선을 선택하거나 "강조표시"하려면 다음을 수행하십시오.

선택된 선은 빨간색으로 강조표시됩니다.

1. 한 개의 선을 선택하려면 선을 마우스 왼쪽 단추로 클릭하십시오.
2. 여러 개의 선을 선택하려면 다음 중 하나를 수행하십시오.

- 커서를 사용하여 해당 선을 선택할 점 주위에 원을 그리십시오.
- Ctrl 키를 누른 상태에서 선택할 개별 선을 마우스 왼쪽 단추로 클릭하십시오.

그래프 배경을 클릭하거나 그래프 창의 웹 메뉴에서 **선택항목 지우기**를 선택하여 선택된 모든 행을 선택 취소할 수 있습니다.

다른 레이아웃을 사용하여 웹을 보려면 다음을 수행하십시오.

웹 메뉴에서 **원 레이아웃**, **네트워크 레이아웃**, **방향이 있는 레이아웃** 또는 **눈금 레이아웃**을 선택하여 그래프의 레이아웃을 변경하십시오.

링크 슬라이더를 켜거나 끄려면 다음을 수행하십시오.

보기 메뉴에서 **링크 슬라이더**를 선택하십시오.

단일 관계의 레코드를 선택하거나 플래그 지정하려면 다음을 수행하십시오.

1. 관심이 있는 관계를 나타내는 선을 마우스 오른쪽 단추로 클릭하십시오.
2. 컨텍스트 메뉴에서 **링크에 대한 선택 노드 생성** 또는 **링크에 대한 파생 노드 생성**을 선택하십시오.

선택 노드 또는 파생 노드는 해당 옵션 및 조건이 지정된 상태로 스트림 캔버스에 자동으로 추가됩니다.

- 선택 노드는 지정된 관계의 모든 레코드를 선택합니다.

- 파생 노드는 선택된 관계가 전체 데이터 세트의 레코드에 대해 참인지 여부를 나타내는 플래그를 생성합니다. 플래그 필드의 이름은 관계를 갖는 두 값을 밑줄로 결합하여 지정합니다(예: *LOW_drugC* 또는 *drugC_LOW*).

관계 그룹의 레코드를 선택하고 플래그를 지정하려면 다음을 수행하십시오.

1. 웹 디스플레이에서 관심이 있는 관계를 나타내는 선을 선택하십시오.
2. 그래프 창의 생성 메뉴에서 **Select Node ("And")**, **Select Node ("Or")**, **Derive Node ("And")** 또는 **Derive Node ("Or")**를 선택하십시오.

- "Or" 노드는 조건의 이접성을 제공합니다. 이는 선택된 관계 중 임의 관계가 참인 레코드에 노드가 적용됨을 의미합니다.
- "And" 노드는 조건의 연접성을 제공합니다. 이는 선택된 모든 관계가 참인 레코드에만 노드가 적용됨을 의미합니다. 선택된 관계 중 상호 배타적인 관계가 있으면 오류가 발생합니다.

선택을 완료하면 선택 노드 또는 파생 노드가 해당 옵션 및 조건이 지정된 상태로 스트림 캔버스에 자동으로 추가됩니다.

웹 임계값 조정

웹 그래프를 작성한 후에는 최소 가시선을 변경하는 도구 모음 슬라이더를 사용하여 선 스타일을 제어하는 임계값을 조정할 수 있습니다. 또한 도구 모음에서 노란색 이중 화살표 단추를 클릭하여 웹 그래프 창을 펼친 후 추가 임계값 옵션을 볼 수 있습니다. **제어** 탭을 클릭하면 추가 옵션이 표시됩니다.

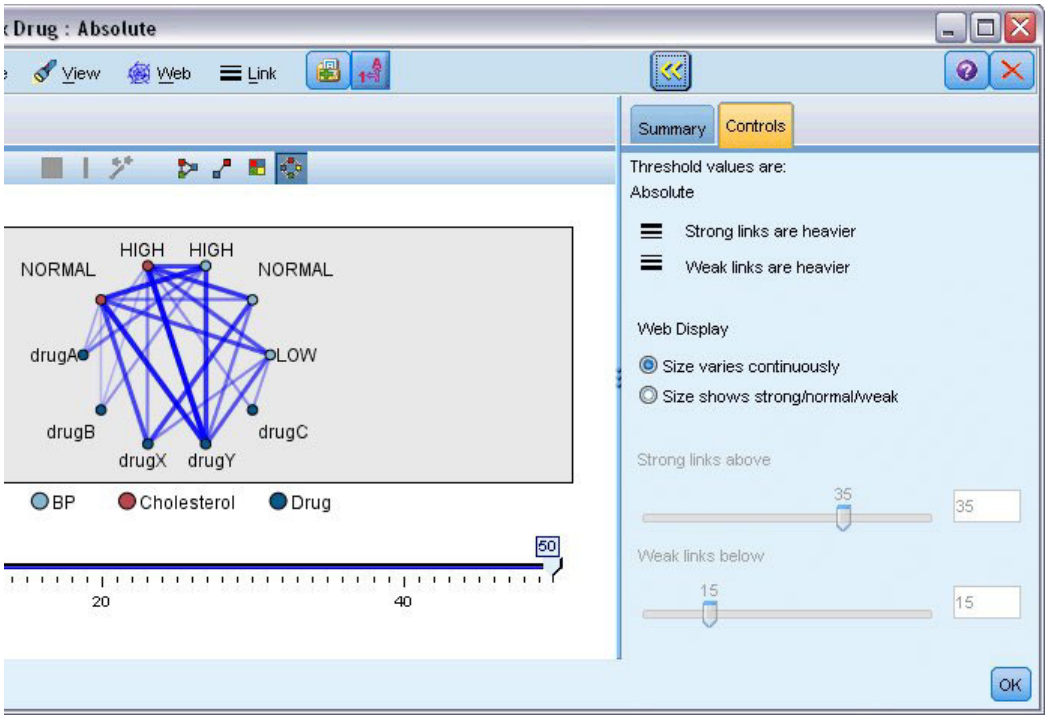


그림 40. 디스플레이 및 임계값 옵션이 있는 펼친 창

임계값. 웹 노드 대화 상자에서 작성 중에 선택된 임계값 유형을 표시합니다.

강한 링크가 더 굵음. 기본적으로 선택됩니다. 필드 사이의 링크를 표시하는 표준 방법입니다.

약한 링크가 더 굵음. 굵은 선으로 표시되는 링크의 의미를 정반대로 바꾸려면 선택하십시오. 이 옵션은 부정 행위를 발견하거나 이상값을 조사하는 데 자주 사용됩니다.

웹 디스플레이. 출력 그래프에서 링크 크기를 제어하는 옵션을 지정하십시오.

- **크기가 계속 변화.** 실제 데이터 값을 기반으로 연결 강도의 변화를 반영하는 링크 크기 범위를 표시하려면 선택하십시오.
- **크기가 강함/보통/약함 표시.** 연결의 세 가지 강도(강함, 보통 및 약함)를 표시하려면 선택하십시오. 최종 그래프에서뿐만 아니라 위에서도 이러한 범주의 분별점을 지정할 수 있습니다.

강한 링크 하한. 강한 연결(굵은 선)과 보통 연결(실선)의 임계값을 지정하십시오. 이 값 위의 모든 연결은 강한 연결로 간주됩니다. 슬라이더를 사용하여 값을 조정하거나 필드에 숫자를 입력하십시오.

약한 링크 상한. 약한 연결(점선)과 보통 연결(실선)의 임계값을 나타내는 숫자를 지정하십시오. 이 값 아래의 모든 연결은 약한 연결로 간주됩니다. 슬라이더를 사용하여 값을 조정하거나 필드에 숫자를 입력하십시오.

웹의 임계값을 조정한 후에는 웹 그래프 도구 모음에 있는 웹 메뉴를 통해 새 임계값으로 웹 디스플레이를 다시 계획하거나 다시 그릴 수 있습니다. 가장 의미있는 패턴을 표시하는 설정을 알았으면, 그래프 창의 웹 메뉴에서 **상위 노드 업데이트**를 선택하여 웹 노드(상위 웹 노드라고도 함)의 원래 설정을 업데이트할 수 있습니다.

웹 요약 작성

도구 모음에서 노란색 이중 화살표 단추를 클릭하여 웹 그래프 창을 펼친 후 강한 링크, 중간 링크 및 약한 링크를 나열하는 웹 요약 문서를 작성할 수 있습니다. **요약** 탭을 클릭하면 각 링크 유형의 테이블이 표시됩니다. 각각의 토글 단추를 사용하여 테이블을 펼치고 접을 수 있습니다.

요약을 인쇄하려면 웹 그래프 창의 메뉴에서 다음을 선택하십시오.

파일 > 요약 인쇄

평가 노드

평가 노드는 애플리케이션에 대해 최적 모델을 선택하기 위해 예측 모형을 평가하고 비교하는 쉬운 방법을 제공합니다. 평가 차트는 특정 결과 예측 시 모델이 작동하는 방식을 보여줍니다. 평가 차트는 예측의 신뢰도 및 예측값을 기반으로 레코드를 정렬하고 레코드를 동일한 크기의 그룹(**분위수**)으로 분할한 후 각 분위수에 대한 비즈니스 기준의 값을 내림차순으로 도표로 작성하여 작동합니다. 다중 모델이 도표에 선구분 변수로 표시됩니다.

결과는 특정 값 또는 값 범위를 **적중**으로 정의하여 처리됩니다. 적중은 일반적으로 관심 있는 이벤트(예: 특정 의료 진단) 또는 일부 정렬(예: 고객에 대한 판매)의 성공을 표시합니다. 대화 상자의 옵션 탭에서 적중 기준을 정의하거나 다음과 같이 기본 적중 기준을 사용할 수 있습니다.

- **플래그** 출력 필드는 직설적이어서 적중은 *참* 값에 해당합니다.
- **명목** 출력 필드의 경우 세트의 첫 번째 값이 적중을 정의합니다.

- **연속형** 출력 필드의 경우 적중은 필드 범위의 중심점보다 큰 값과 동일합니다.

여섯 가지 유형의 평가 차트가 있으며 각각의 차트는 다른 평가 기준을 강조합니다.

Gains 차트

Gains는 각 분위수에서 발생하는 적중 총계의 비율로 정의됩니다. Gains는 (분위수의 적중 수 / 적중 수 총계) × 100%로 계산됩니다.

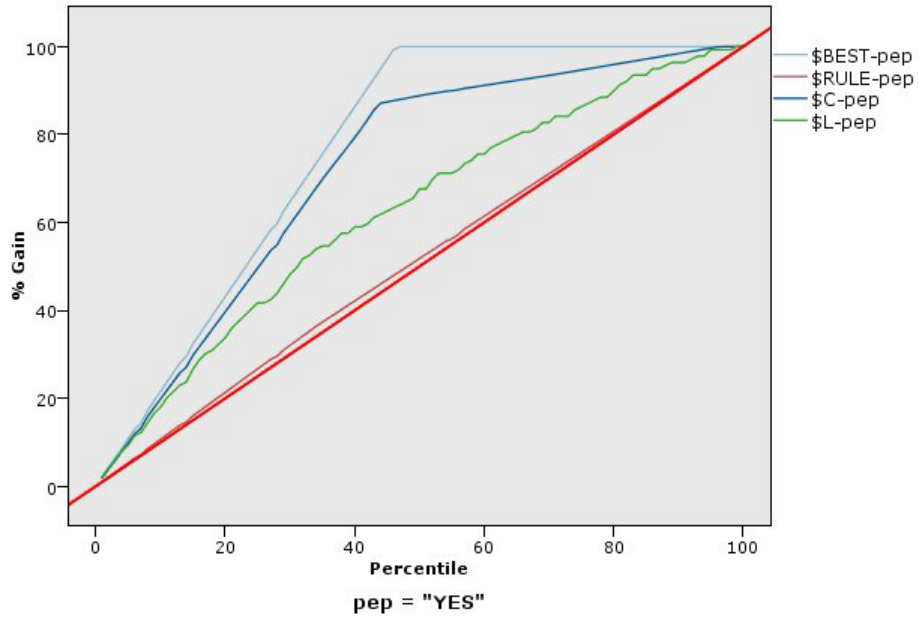


그림 41. 기준선, 최적 예측선 및 비즈니스 규칙이 표시된 Gains 차트(누적)

리프트 도표

리프트는 적중인 각 분위수의 레코드 백분율을 훈련 데이터의 전체 적중 백분율과 비교합니다. 리프트는 (분위수의 적중 수 / 분위수의 레코드 수) / (적중 총계 / 레코드 총계)로 계산됩니다.

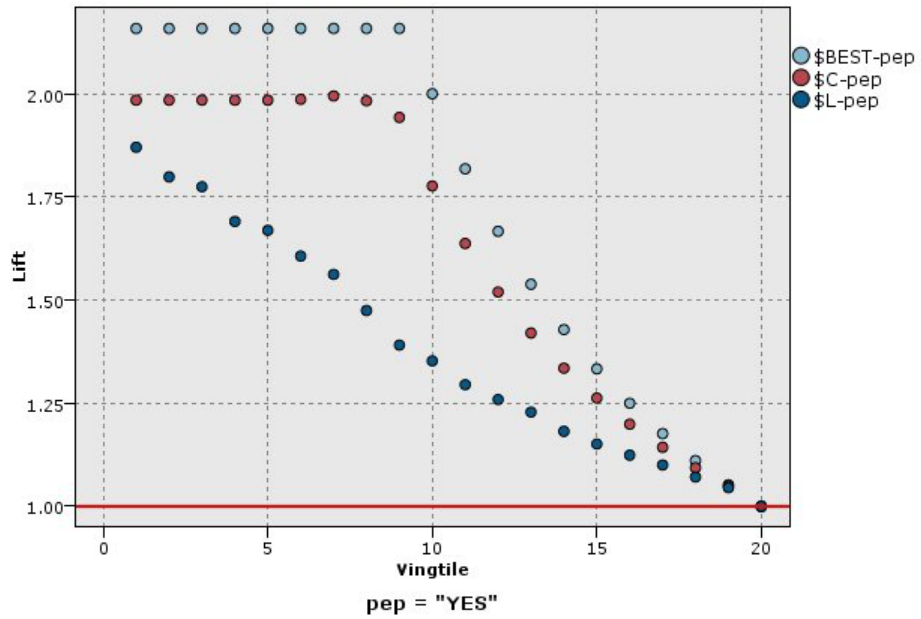


그림 42. 점 및 최적 예측선을 사용하는 리프트 도표(누적)

반응 차트

반응은 단순히 적중인 분위수의 레코드 백분율입니다. 반응은 (분위수의 적중 수 / 분위수의 레코드 수) × 100%로 계산됩니다.

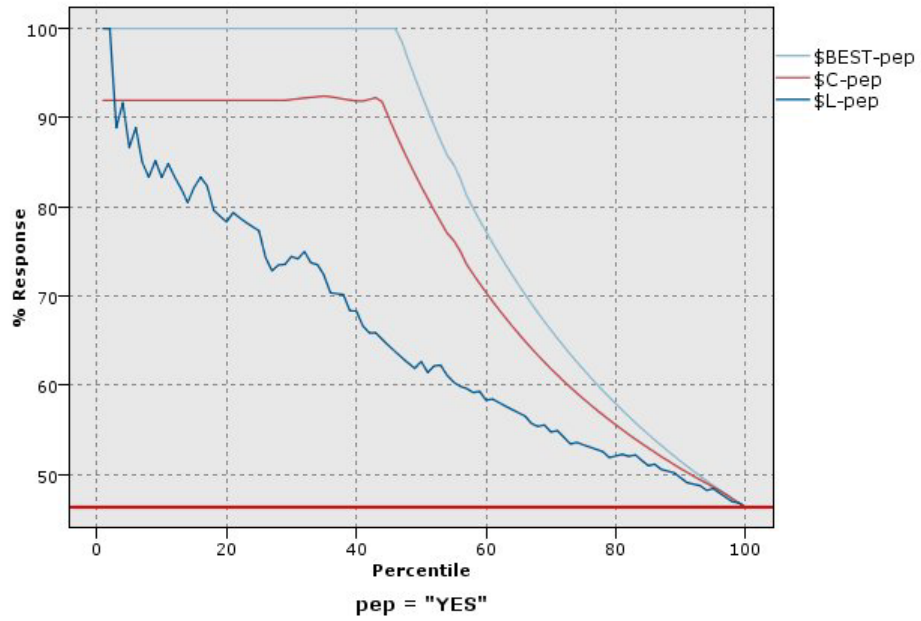


그림 43. 최적 예측선을 사용하는 반응 차트(누적)

이익 차트

이익은 각 레코드에 대한 **수입**에서 해당 레코드에 대한 **비용**을 뺀 값입니다. 분위수의 이익은 단순히 분위수의 전체 레코드 이익 합계입니다. 수입은 적중에만 적용되는 것으로 가정되지만 비용은 모든 레코드에 적용됩니다. 이익 및 비용은 고정이거나 데이터의 필드에 의해 정의될 수 있습니다. 이익은 (분위수의 레코드에 대한 수입의 합계 - 분위수의 레코드에 대한 비용의 합계)로 계산됩니다.

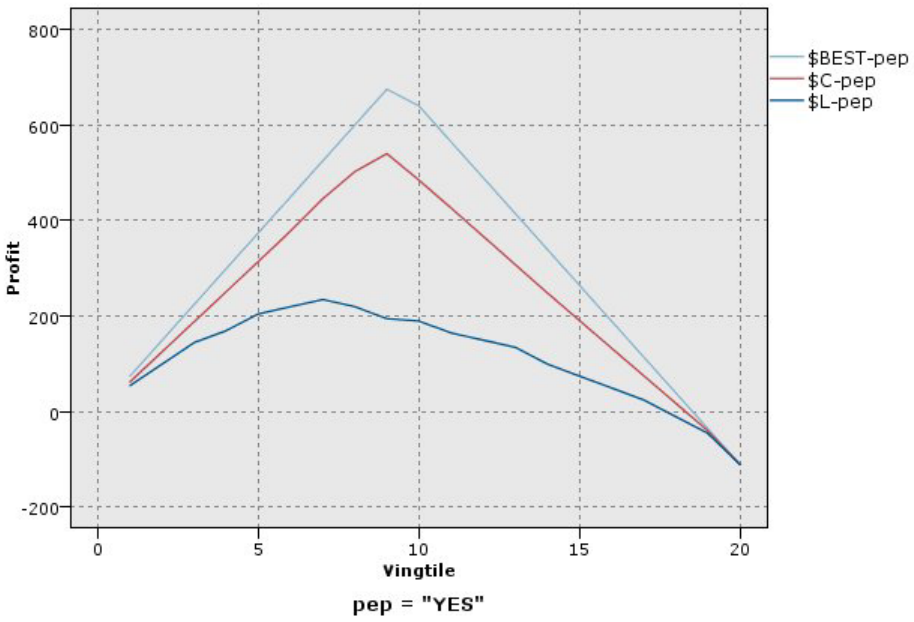


그림 44. 최적 예측선을 사용하는 이익 차트(누적)

ROI 차트

ROI(Return On Investment)는 수입 및 비용 정의를 포함한다는 점에서 이익과 비슷합니다. ROI는 분위수에 대한 비용과 이익을 비교합니다. ROI는 (분위수에 대한 이익 / 분위수에 대한 비용) × 100%로 계산됩니다.

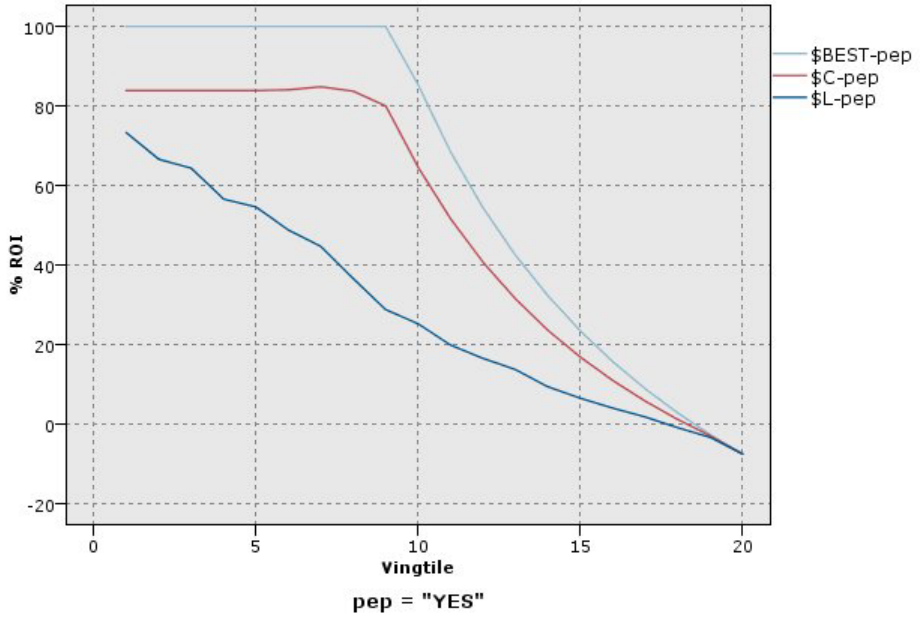


그림 45. 최적 예측선을 사용하는 ROI 차트(누적

ROC 차트

ROC(Receiver Operator Characteristic)는 이분형 분류자와 함께만 사용할 수 있습니다. ROC는 분류자의 성능을 기반으로 분류자를 시각화하고 구성하고 선택하는 데 사용할 수 있습니다. ROC 차트는 분류자의 거짓 긍정 비율에 대해 참 긍정 비율(민감도)을 도표화합니다. ROC 차트는 이익(참 긍정)과 비용(거짓 긍정) 간 상대적인 균형을 보여줍니다. 참 긍정은 적중인 인스턴스이며 적중으로 분류됩니다. 따라서 참 긍정 비율은 참 긍정 수를 실제로 적중인 인스턴스 수로 나눠서 계산됩니다. 거짓 긍정은 빗나감인 인스턴스이며 적중으로 분류됩니다. 따라서 거짓 긍정 비율은 거짓 긍정 수를 실제로는 빗나감인 인스턴스 수로 나눠서 계산됩니다.

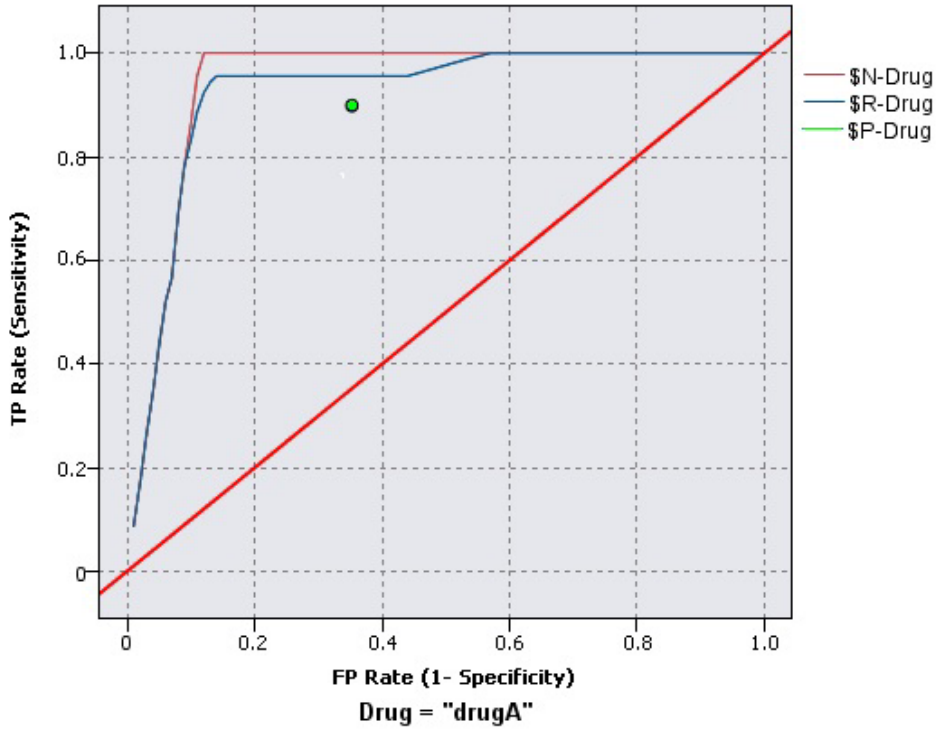


그림 46. 최적 예측선을 사용하는 ROC 차트

각각의 점이 해당 분위수에 대한 값에 더 높은 모든 분위수를 더한 값과 동일하도록 평가 차트는 누적일 수도 있습니다. 누적 차트가 일반적으로 모델의 전체 성능을 더 잘 전달하지만 비누적 차트가 모델에 대한 특정 문제점 영역 표시에서 뛰어날 수도 있습니다.

평가 도표 탭

차트 유형. 이익(Gains), **반응**, **리프트**, **이익(Profit)**, **투자수익률**(ROI) 또는 **ROC**(Receiver Operator Characteristic) 유형 중 하나를 선택하십시오.

누적 도표. 누적 차트를 작성하려면 선택하십시오. 누적 차트에 각 분위수 및 더 높은 분위수에 대해 값이 표시됩니다. (ROC 차트에는 **누적 도표**를 사용할 수 없습니다.)

기준선 포함. 도표에 기준선을 포함하려면 선택하십시오. 이 기준선은 신뢰도가 관련이 없어지는, 적중 수에 대한 완전한 임의 분포를 표시합니다. (이익 및 ROI 차트에는 **기준선 포함**을 사용할 수 없습니다.)

최적 예측선 포함. 도표에 최적 예측선을 포함하려면 선택하십시오. 이 최적 예측선은 완벽한 신뢰도(적중 수 = 케이스 중 100%)를 표시합니다. (ROC 차트에는 **최적 예측선**을 사용할 수 없습니다.)

모든 차트 유형에 이익 기준 사용. 정규 적중 수 대신 평가 측도를 계산할 때 이익 기준(비용, 수입, 가중치)을 사용하려면 선택하십시오. 특정 숫자 대상을 포함한 모델의 경우(예: 제안에 응해 고객으로부터 얻은 수입을 예측하는 모델), 대상 필드 값은 적중 수보다 나은 모델 성능 측도를 제공합니다. 이 옵션을 선택하면 이익, 응답, 리프트 차트에 **비용**, **수익**, **가중치** 필드를 사용할 수 있습니다. 이 세 가지 차트 유형에 이익 기준을 사

용하려면 **수입**을 대상 필드로, **비용**을 0.0으로 설정하여 이익이 수입과 같도록 해야 하고 모든 레코드가 적중 수로 계수되도록 사용자 정의된 적중 조건을 "참"으로 지정해야 합니다. (ROC 차트에는 **모든 차트 유형에 이익 기준 사용**을 사용할 수 없습니다.)

예측/예측변수 필드를 찾을 때 사용. 해당 메타데이터를 사용하여 그래프에서 예측 필드를 검색하려면 **모델 출력 필드 메타데이터**를 선택하고, 이름을 기준으로 검색하려면 **필드 이름 형식**을 선택하십시오.

도표 스코어 필드. 스코어 필드 선택기를 사용하려면 이 확인 상자를 선택하십시오. 그런 다음 하나 이상의 범위 또는 연속 스코어 필드, 즉 엄격한 예측 모형은 아니지만 적중 성향이라는 점에서 레코드의 순위를 매기는 데 유용한 필드를 선택하십시오. 평가 노드는 하나 이상의 스코어 필드의 조합을 하나 이상의 예측 모형과 비교할 수 있습니다. 일반 예에서는 여러 RFM 필드를 최적 예측 모형과 비교합니다.

목표. 필드 선택기를 사용하여 대상 필드를 선택하십시오. 둘 이상의 값을 가진 명목 필드 또는 인스턴스화된 플래그를 선택하십시오.

참고: 이 대상 필드는 스코어 필드에만 적용 가능하며(예측 모형이 고유 대상 정의) 사용자 정의 적중 기준이 옵션 탭에 설정되어 있는 경우 무시됩니다.

파티션별 분할. 파티션 필드를 사용하여 레코드를 학습, 검정, 검증 표본으로 분할하는 경우, 이 옵션을 선택하여 각 파티션에 대해 개별 평가 차트를 표시하십시오. 자세한 정보는 181 페이지의 『파티션 노드』의 내용을 참조하십시오.

참고: 파티션별로 분할하는 경우 파티션 필드에 널값이 있는 레코드가 평가에서 제외됩니다. 파티션 노드는 널값을 생성하지 않으므로 파티션 노드가 사용되는 경우 이는 문제가 되지 않습니다.

도표. 드롭 다운 목록에서 차트에 표시할 분위수의 *크기*를 선택하십시오. 옵션에는 **사분위수**, **오분위수**, **십분위수**, **이십분위수**, **백분위수**, **천분위수**가 있습니다. (ROC 차트에는 **도표**를 사용할 수 없습니다.)

스타일. **선** 또는 **점**을 선택하십시오.

ROC 차트를 제외한 모든 차트 유형의 경우 추가 제어를 사용하면 비용, 수입, 가중치를 지정할 수 있습니다.

- **비용.** 각 레코드와 연관된 비용을 지정합니다. **고정** 또는 **가변** 비용을 선택할 수 있습니다. 고정 비용의 경우 비용 값을 지정하십시오. 가변 비용의 경우에는 필드 선택기 단추를 클릭하여 한 필드를 비용 필드로 선택하십시오. (ROC 차트에는 **비용**을 사용할 수 없습니다.)
- **수입.** 적중을 나타내는 각 레코드와 연관된 수입을 지정합니다. **고정** 또는 **가변** 비용을 선택할 수 있습니다. 고정 수입의 경우 수입 값을 지정하십시오. 가변 수입의 경우에는 필드 선택기 단추를 클릭하여 한 필드를 수입 필드로 선택하십시오. (ROC 차트에는 **수입**을 사용할 수 없습니다.)
- **가중치.** 데이터의 레코드가 둘 이상의 단위를 표시하는 경우 빈도 가중치를 사용하여 결과를 조정할 수 있습니다. **고정** 또는 **가변** 가중치를 사용하여 각 레코드와 연관된 가중치를 지정하십시오. 고정 가중치의 경우 가중값(레코드별 노드 수)을 지정하십시오. 가변 가중치의 경우에는 필드 선택기 단추를 클릭하여 한 필드를 가중 필드로 선택하십시오. (ROC 차트에는 **가중치**를 사용할 수 없습니다.)

평가 옵션 탭

평가 차트에 대한 옵션 탭은 차트에 표시되는 적중, 스코어링 기준 및 비즈니스 규칙 정의 시 유연성을 제공합니다. 모델 평가 결과를 내보내기 위한 옵션도 설정할 수 있습니다.

사용자 정의 적중. 적중을 표시하는 데 사용되는 사용자 정의 조건을 지정하려면 선택하십시오. 이 옵션은 값의 순서 및 목표 필드의 유형에서 관심 있는 결과를 추론할 때보다 관심 있는 결과를 정의하는 경우에 유용합니다.

- **조건.** 위에서 **사용자 정의 적중**이 선택되면 적중 조건에 대해 CLEM 표현식을 지정해야 합니다. 예를 들어, `@TARGET = "YES"`는 목표 필드에 대한 *Yes* 값이 평가에서 적중으로 계수됨을 나타내는 유효한 조건입니다. 지정된 조건은 모든 목표 필드에 사용됩니다. 조건을 작성하려면 필드를 입력하거나 표현식 작성기를 사용하여 조건식을 생성하십시오. 데이터가 인스턴스화되는 경우에는 표현식 작성기에서 직접 값을 삽입할 수 있습니다.

사용자 정의 스코어. 스코어링 케이스를 분위수에 지정하기 전에 스코어링 케이스에 사용되는 조건을 지정하려면 선택하십시오. 기본 스코어는 예측값 및 신뢰도로부터 계산됩니다. 표현식 필드를 사용하여 사용자 정의 스코어링 표현식을 작성하십시오.

- **표현식.** 스코어링에 사용되는 CLEM 표현식을 지정하십시오. 예를 들어, 0–1 범위의 숫자 출력이 낮은 값이 높은 값보다 나은 것으로 정렬되는 경우 적중을 `@TARGET < 0.5`로 정의하고 연관된 스코어를 `1 - @PREDICTED`로 정의할 수 있습니다. 스코어 표현식에서는 숫자 값을 생성해야 합니다. 조건을 작성하려면 필드를 입력하거나 표현식 작성기를 사용하여 조건식을 생성하십시오.

비즈니스 규칙 포함. 관심 있는 기준을 반영하는 규칙 조건을 지정하려면 선택하십시오. 예를 들어, `mortgage = "Y" and income >= 33000`인 모든 케이스에 대해 규칙을 표시하길 원할 수 있습니다. 비즈니스 규칙이 차트에서 그려지고 키에서 *규칙*으로 레이블 지정됩니다. (**비즈니스 규칙 포함**은 REC 차트에 대해서는 지원되지 않습니다.)

- **조건.** 출력 차트에서 비즈니스 규칙을 정의하는 데 사용되는 CLEM 표현식을 지정하십시오. 단순히 필드를 입력하거나 표현식 작성기를 사용하여 조건식을 생성하십시오. 데이터가 인스턴스화되는 경우에는 표현식 작성기에서 직접 값을 삽입할 수 있습니다.

파일로 결과 내보내기. 모델 평가 결과를 구분된 텍스트 파일로 내보내려면 선택하십시오. 이 파일을 읽고 계산된 값에 대해 특수 분석을 수행할 수 있습니다. 내보내기를 위해 다음과 같은 옵션을 설정하십시오.

- **파일 이름.** 출력 파일의 파일 이름을 입력하십시오. 생략 기호 단추(**...**)를 사용하여 원하는 폴더를 찾아보십시오.
- **구분자.** 필드 구분자로 사용할 문자(예: 쉼표 또는 공백)를 입력하십시오.

필드 이름 포함. 필드 이름을 출력 파일의 첫 번째 행으로 포함하려면 이 옵션을 선택하십시오.

각 레코드 다음에 줄 바꾸기. 새로운 행에서 각각의 레코드를 시작하려면 이 옵션을 선택하십시오.

평가 모양 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

제목. 그래프 제목으로 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

텍스트. 자동으로 생성된 텍스트 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

X 레이블. 자동으로 생성된 *x*축(가로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

Y 레이블. 자동으로 생성된 *y*축(세로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

눈금선 표시. 기본적으로 선택되는 이 옵션은 더 쉽게 영역 및 밴드 절사 지점을 결정할 수 있게 하는 눈금선을 도표 또는 그래프 뒤에 표시합니다. 그래프 배경이 흰색인 경우가 아니면 눈금선은 항상 흰색으로 표시됩니다. 그래프 배경이 흰색이면 눈금선은 회색으로 표시됩니다.

모델 평가의 결과 읽기

평가 차트의 해석은 차트 유형에 대한 특정 범위에 따라 다르지만 모든 평가 차트에 공통인 몇몇 특성이 있습니다. 누적 차트의 경우 더 높이 있는 선은 더 나은 모델을 표시합니다(특히 차트의 왼쪽에서). 많은 경우 여러 모델을 비교하면 선이 겹쳐 한 모델이 차트의 한 부분에서 더 높고 다른 모델이 차트의 다른 부분에서 더 높습니다. 이 경우에는 선택할 모델을 결정할 때 원하는 표본의 부분(*x*축에서의 위치를 정의함)을 고려해야 합니다.

대부분의 비누적 차트는 매우 비슷합니다. 양호한 모델의 경우 비누적 차트는 차트의 왼쪽에서 더 높고 차트의 오른쪽에서 낮아야 합니다. (비누적 차트에 톱니 패턴이 표시되는 경우에는 분위수의 수를 줄여 그래프를 도표화하고 재실행하여 평탄하게 할 수 있습니다.) 차트 왼쪽의 내려간 부분 또는 오른쪽의 올라간 부분은 모델의 예측이 양호하지 않은 영역을 표시할 수 있습니다. 전체 그래프에서 평평한 선은 본질적으로 정보를 제공하지 않는 모델을 표시합니다.

Gains 차트. 누적 Gains 차트는 항상 0%에서 시작하여 왼쪽에서 오른쪽으로 이동하면서 100%에서 끝납니다. 양호한 모델의 경우 Gains 차트는 100%를 향해 가파르게 상승한 후 수평을 유지합니다. 정보를 제공하지 않는 모델은 왼쪽 하단에서 오른쪽 상단으로 대각선으로 진행합니다(**기준선 포함**이 선택된 경우 차트에 표시됨).

리프트 도표. 누적 리프트 도표는 1.0 이상에서 시작하여 왼쪽에서 오른쪽으로 이동함에 따라 1.0에 도달할 때까지 점진적으로 내려갑니다. 차트의 오른쪽 가장자리는 전체 데이터 세트를 나타내므로 데이터의 적중 수에 대한 누적 분위수의 적중 수 비율은 1.0입니다. 양호한 모델의 경우 리프트는 왼쪽에서 1.0보다 훨씬 위에서 시작하여 오른쪽으로 이동할 때 높은 위치에서 안정 상태를 유지한 후 차트 오른쪽에서 1.0을 향해 급격하게 하강해야 합니다. 정보를 제공하지 않는 모델의 경우에는 전체 그래프에 대해 선이 1.0 주위를 맴돕니다. (**기준선 포함**이 선택되면 참조를 위해 1.0에서 가로 선이 차트에 표시됩니다.)

반응 차트. 누적 응답 차트는 척도화를 제외하고 리프트 도표와 매우 비슷합니다. 반응 차트는 일반적으로 100% 근처에서 시작하여 차트의 오른쪽 가장자리에서 전체 응답률(적중 총계/레코드 총계)에 도달할 때까지 점진적으로 내려갑니다. 양호한 모델의 경우 선은 왼쪽에서 100% 또는 이에 근접한 값에서 시작하여 오른쪽으로 이동함에 따라 높은 위치에서 안정 상태를 유지한 후 차트 오른쪽에서 전체 반응률을 향해 급격하게 하강합니다. 정보를 제공하지 않는 모델의 경우에는 전체 그래프에 대해 선이 전체 반응률 주위를 맴돕니다. (**기준선 포함**이 선택되면 참조를 위해 전체 반응률에서 가로 선이 차트에 표시됩니다.)

이익 차트. 누적 이익 차트는 선택된 표본의 크기를 늘리면서 왼쪽에서 오른쪽으로 이동할 때 이익의 합계를 표시합니다. 이익 차트는 일반적으로 0 근처에서 시작하고 가운데에서 최대치 또는 높은 위치의 안정 상태에 도달할 때까지 오른쪽으로 이동하면서 점진적으로 증가한 후 차트의 오른쪽 가장자리를 향해 감소합니다. 양호한 모델의 경우 이익은 차트의 가운데 쪽에 잘 정의된 최대치를 표시합니다. 정보를 제공하지 않는 모델의 경우에는 선이 상대적으로 직선이며 적용되는 비용/수입 구조에 따라 증가하거나 감소하거나 수평을 유지할 수 있습니다.

ROI 차트. 누적 ROI(Return On Investment) 차트는 척도화를 제외하고 반응 차트 및 리프트 도표와 비슷합니다. ROI 차트는 일반적으로 0% 이상에서 시작한 후 전체 데이터 세트에 대한 전체 ROI(음수가 될 수 있음)에 도달할 때까지 점진적으로 내려갑니다. 양호한 모델의 경우 선은 0%보다 훨씬 위에서 시작하고 오른쪽으로 이동함에 따라 높은 위치에서 안정 상태를 유지한 후 차트 오른쪽의 전체 ROI를 향해 급격하게 하강해야 합니다. 정보를 제공하지 않는 모델의 경우에는 선이 전체 ROI 값 주위를 맴돌아야 합니다.

ROC 차트. ROC 곡선은 일반적으로 누적 Gains 차트 모양을 가지고 있습니다. 곡선은 (0,0) 좌표에서 시작한 후 왼쪽에서 오른쪽으로 이동하면서 (1,1) 좌표에서 끝납니다. (0,1) 좌표를 향해 급격하게 상승한 후 수평을 유지하는 차트는 양호한 분류자를 표시합니다. 인스턴스를 무작위로 적중 또는 빗나감으로 분류하는 모델은 왼쪽 하단에서 오른쪽 상단으로 대각선으로 진행합니다(**기준선 포함**이 선택된 경우 차트에 표시됨). 모델에 대해 신뢰도 필드가 제공되지 않으면 해당 모델은 단일 점으로 도표화됩니다. 분류의 최적 임계값을 가진 분류자는 차트의 (0,1) 좌표(또는 왼쪽 상단)에 가장 가까운 위치에 있습니다. 이 위치는 적중으로 올바르게 분류되는 많은 수의 인스턴스와 적중으로 잘못 분류되는 적은 수의 인스턴스를 나타냅니다. 대각선 위의 점은 양호한 분류 결과를 나타냅니다. 대각선 아래의 점은 인스턴스가 무작위로 분류된 경우보다 나쁜 양호하지 않은 분류 결과를 나타냅니다.

평가 차트 사용

마우스를 사용하여 평가 차트를 탐색하는 것은 히스토그램 또는 콜렉션 그래프를 사용하는 것과 비슷합니다. *x* 축은 이십분위수 또는 십분위수 등의 지정된 분위수에서 모델 스코어를 나타냅니다.

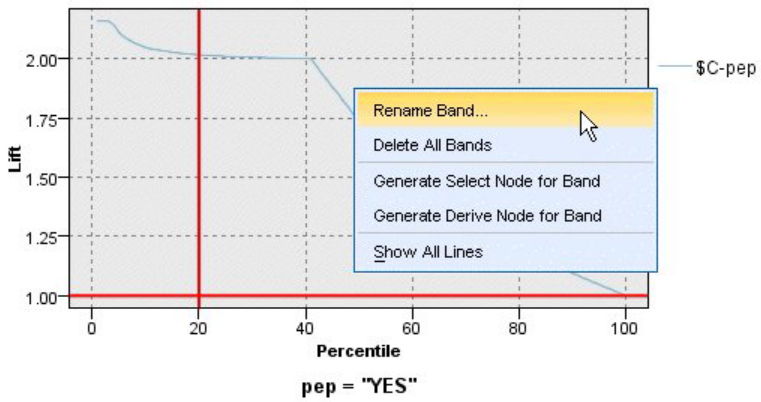


그림 47. 평가 차트에 대한 작업

분할자 아이콘을 사용하여 *x*축을 동등한 밴드로 자동으로 분할하는 옵션을 표시하여 히스토그램의 경우와 마찬가지로 x축을 밴드로 파티셔닝할 수 있습니다. 자세한 정보는 276 페이지의 『그래프 탐색』의 내용을 참조하십시오. 편집 메뉴에서 **그래프 밴드**를 선택하여 밴드의 경계를 수동으로 편집할 수 있습니다.

평가 차트를 작성하고 밴드를 정의하고 결과를 검토한 후에는 컨텍스트 메뉴 및 생성 메뉴의 옵션을 사용하여 그래프의 선택사항을 기반으로 자동으로 노드를 작성할 수 있습니다. 자세한 정보는 284 페이지의 『그래프에서 노드 생성』의 내용을 참조하십시오.

평가 차트에서 노드를 생성할 때 차트에서 사용 가능한 모든 모델 중에서 하나의 모델을 선택하라는 프롬프트가 표시됩니다.

모델을 선택한 후 **확인**을 클릭하여 새 노드를 스트림 캔버스에 생성하십시오.

맵 시각화 노드

맵 시각화 노드는 다중 입력 연결을 승인하고 지리 공간적 데이터를 맵에 일련의 레이어로 표시할 수 있습니다. 각각의 레이어는 하나의 지리 공간적 필드입니다. 예를 들어, 기준 레이어가 한 국가의 맵이고 그 위에 도로에 대한 레이어 하나, 강에 대한 레이어 하나, 도시에 대한 레이어 하나가 있을 수 있습니다.

대부분의 지리 공간적 데이터 세트는 일반적으로 하나의 지리 공간적 필드를 포함하고 있지만 하나의 입력에 여러 지리 공간적 필드가 있으면 표시할 필드를 선택할 수 있습니다. 동일한 입력 연결의 두 필드는 동시에 표시할 수 없습니다. 하지만 수신 연결을 복사하여 붙여넣고 각각으로부터 다른 필드를 표시할 수 있습니다.

맵 시각화 도표 탭

레이어

이 테이블에는 맵 노드에 대한 입력에 관한 정보가 표시됩니다. 레이어의 순서는 노드가 실행될 때 맵 미리보기와 시각적 출력 모두에서 레이어가 표시되는 순서를 지시합니다. 테이블의 맨 위 행이 '맨 위' 레이어이고 맨 아래 행이 '맨 아래' 레이어입니다. 즉, 각각의 레이어는 맵의 테이블에서 바로 아래에 있는 레이어 앞에 표시됩니다.

참고: 테이블의 레이어에 3차원 지리 공간적 필드가 포함되어 있으면 x축 및 y축만 도표화됩니다. z축은 무시됩니다.

이름 이름은 각 레이어에 대해 자동으로 작성되며 `tag[source node:connected node]` 형식을 사용하여 구성됩니다. 기본적으로 태그는 숫자로 표시되며 1은 연결되는 첫 번째 입력을 나타내고 2는 두 번째 입력을 나타내는 방식으로 표시됩니다. 필요한 경우 **레이어 편집** 단추를 눌러 맵 레이어 옵션 변경 대화 상자에서 태그를 변경하십시오. 예를 들어, 태그가 "도로" 또는 "구/군/시"가 되도록 변경하여 데이터 입력을 반영할 수 있습니다.

유형 레이어로 선택되는 지리 공간적 필드의 측정 유형 아이콘을 표시합니다. 입력 데이터에 지리 공간적 측정 유형을 가진 여러 필드가 포함되어 있는 경우 기본 선택사항에서는 다음 정렬 순서를 사용합니다.

1. 점
2. 선 스트링
3. 다각형
4. 다중 점
5. 다중 선 스트링
6. 다중 다각형

참고: 동일한 측정 유형을 가진 두 개의 필드가 있으면 첫 번째 필드(이름별 알파벳순)가 기본적으로 선택됩니다.

기호

참고: 이 열은 점 및 다중 점 필드의 경우에만 완료됩니다.
점 또는 다중 점 필드에 사용되는 기호를 표시합니다. 필요한 경우 **레이어 편집** 단추를 눌러 맵 레이어 옵션 변경 대화 상자에서 기호를 변경하십시오.

색상 맵에서 레이어를 나타내는 데 사용되는 색상을 표시합니다. 필요한 경우 **레이어 편집** 단추를 눌러 맵 레이어 옵션 변경 대화 상자에서 색상을 변경하십시오. 색상은 측정 유형에 따라 다양한 항목에 적용됩니다.

- 점 또는 다중 점의 경우 색상은 레이어에 대한 기호에 적용됩니다.
- 선 스트링 및 다각형의 경우 색상은 전체 모양에 적용됩니다. 다각형은 항상 검은색 윤곽선을 가지고 있습니다. 열에 표시되는 색상은 모양을 채우는 데 사용되는 색상입니다.

미리보기

이 분할창에는 **레이어** 테이블에서의 현재 입력 선택사항에 대한 미리보기가 표시됩니다. 미리보기는 레이어의 순서, 기호, 색상 및 레이어와 연관된 기타 표시 설정을 고려하며 가능한 경우 설정이 변경될 때마다 표시를 업데이트합니다. 스트림의 다른 위치(예: 레이어로 사용할 지리 공간적 필드)에서 세부사항을 변경하거나 연관된 통합 함수 등의 세부사항을 수정하는 경우에는 **데이터 새로 고치기** 단추를 클릭하여 미리보기를 업데이트해야 할 수 있습니다.

스트림을 실행하기 전에 **미리보기**를 사용하여 표시 설정을 설정하십시오. 큰 데이터 세트를 사용할 때 발생할 수 있는 시간 보류를 방지하기 위해 미리보기에서는 각각의 레이어에 대한 표본을 추출하고 처음 100개 레코드로부터 표시를 작성합니다.

맵 레이어 변경

맵 레이어 옵션 변경 대화 상자를 사용하여 시각화 노드의 **도표** 탭에 표시되는 레이어의 다양한 세부사항을 수정할 수 있습니다.

입력 세부사항

태그 기본적으로 태그는 숫자입니다. 이 숫자를 더 의미 있는 태그로 바꿔 맵에서 레이어 식별을 지원할 수 있습니다. 예를 들어, 태그는 데이터 입력의 이름일 수 있습니다(예: "구/군/시").

레이어 필드

입력 데이터에 둘 이상의 지리 공간적 필드가 있는 경우 이 옵션을 사용하여 맵에 레이어로 표시할 필드를 선택하십시오.

기본적으로 선택할 수 있는 레이어는 다음과 같은 순서로 정렬되어 있습니다.

- 점
- 선 스트링
- 다각형
- 다중 점
- 다중 선 스트링
- 다중 다각형

표시 설정

육각형 구간화 사용

참고: 이 옵션은 점 및 다중 점 필드에만 영향을 미칩니다.

육각형 구간화에서는 x 및 y 좌표를 기반으로 인접한 점을 단일 점으로 결합하여 맵에 표시합니다. 단일 점은 육각형으로 표시되지만 사실상 다각형으로 렌더링됩니다.

육각형은 다각형으로 렌더링되므로 육각형 구간화가 켜진 점 필드는 모두 다각형으로 처리됩니다. 이는 맵 노드 대화 상자에서 **유형별 정렬**을 선택하면 육각형 구간화가 적용된 점 레이어는 모두 다각형 레이어 위와 선 스트링 및 점 레이어 아래에 렌더링됨을 의미합니다.

다중 점 필드에 대해 육각형 구간화를 사용하는 경우에는 먼저 중심 점을 계산하기 위해 다중 점 값을 구간화하여 해당 필드가 점 필드로 변환됩니다. 중심 점은 육각형 구간을 계산하는 데 사용됩니다.

통합

참고: 이 열은 **육각형 구간화 사용** 선택란을 선택하고 **오버레이**도 선택하는 경우에만 사용할 수 있습니다.

육각형 구간화를 사용하는 점 레이어에 대해 **오버레이** 필드를 선택하는 경우에는 육각형 내 모든 점에 대해 해당 필드에 있는 모든 값을 통합해야 합니다. 맵에 적용할 오버레이 필드에 대한 통합 함수를 지정하십시오. 사용 가능한 통합 함수는 측정 유형에 따라 다릅니다.

- 실수 또는 정수 저장 공간을 가진 연속형 측정 유형에 대한 통합 함수:
  - 합계
  - 평균
  - 최소값
  - 최대값
  - 중앙값
  - 첫 번째 사분위수
  - 세 번째 사분위수
- 시간, 날짜 또는 시간소인 저장 공간을 가진 연속형 측정 유형에 대한 통합 함수:
  - 평균
  - 최소값
  - 최대값
- 명목 또는 범주형 측정 유형에 대한 통합 함수:
  - 모드
  - 최소값
  - 최대값
- 플래그 측정 유형에 대한 통합 함수:
  - 참(참인 항목이 있는 경우)
  - 거짓(거짓인 항목이 있는 경우)

색상

데이터에 있는 다른 필드의 값을 기반으로 기능에 색상을 지정하는 오버레이 필드 또는 지리 공간적 필드의 모든 기능에 적용할 표준 색상을 선택하려면 이 옵션을 사용하십시오.

표준을 선택하는 경우에는 사용자 옵션 대화 상자의 표시 탭에 있는 **차트 범주 색상 순서** 분할창에 표시되는 색상의 팔레트에서 색상을 선택할 수 있습니다.

오버레이를 선택하는 경우에는 **레이어 필드**로 선택된 지리 공간적 필드가 포함된 데이터 소스에서 필드를 선택할 수 있습니다.

- 명목 또는 범주형 오버레이 필드의 경우 선택할 수 있는 색상 팔레트는 **표준** 색상 옵션에 대해 표시되는 것과 동일합니다.
- 연속형 및 순서 오버레이 필드의 경우에는 두 번째 드롭 다운 목록이 표시되고 여기서 색상을 선택합니다. 색상을 선택하면 연속형 또는 순서 필드의 값에 따라 해당 색상의 채도가 변경되어 오버레이가 적용됩니다. 가장 높은 값은 드롭 다운 목록에서 선택된 색상을 사용하고 더 낮은 값은 더 낮은 채도로 표시됩니다.

기호

참고: 점 및 다중 점 측정 유형에 대해서만 사용으로 설정됨.

지리 공간적 필드의 모든 레코드에 적용되는 **표준** 기호를 사용할지 아니면 데이터에 있는 다른 필드의 값을 기반으로 점에 대한 기호 아이콘을 변경하는 **오버레이** 기호를 사용할지 선택하려면 이 옵션을 사용하십시오.

표준을 선택하는 경우에는 드롭 다운 목록에서 기본 기호 중 하나를 선택하여 맵에 점 데이터를 나타낼 수 있습니다.

오버레이를 선택하는 경우에는 **레이어 필드**로 선택된 지리 공간적 필드가 포함된 데이터 소스에서 명목, 순서 또는 범주형 필드를 선택할 수 있습니다. 오버레이 필드의 각각의 값에 대해 다른 기호가 맵에 표시됩니다.

예를 들어, 데이터에 상점의 위치를 나타내는 점 필드가 포함되어 있고 오버레이가 상점 유형 필드일 수 있습니다. 이 예제에서 모든 식품 상점은 맵에서 십자 기호로 식별되고 모든 전자제품 상점은 사각형 기호로 식별될 수 있습니다.

Size

참고: 점, 다중 점, 선 스트링 및 다중 선 스트링 측정 유형에 대해서만 사용으로 설정됨.

지리 공간적 필드의 모든 레코드에 적용되는 **표준** *크기*를 사용할지 아니면 데이터에 있는 다른 필드의 값을 기반으로 선 굵기 또는 기호 아이콘의 *크기*를 변경하는 **오버레이** *크기*를 사용할지 선택하려면 이 옵션을 사용하십시오.

표준을 선택하는 경우에는 픽셀 너비 값을 선택할 수 있습니다. 사용 가능한 옵션은 1, 2, 3, 4, 5, 10, 20 또는 30입니다.

오버레이를 선택하는 경우에는 **레이어 필드**로 선택된 지리 공간적 필드가 포함된 데이터 소스에서 필드를 선택할 수 있습니다. 점 또는 선의 굵기는 선택한 필드의 값에 따라 다릅니다.

투명도

지리 공간적 필드의 모든 레코드에 적용되는 **표준** 투명도를 사용할지 아니면 데이터에 있는 다른 필드의 값을 기반으로 기호, 선 또는 다각형의 투명도를 변경하는 **오버레이** 투명도를 사용할지 선택하려면 이 옵션을 사용하십시오.

표준을 선택하는 경우에는 0%(불투명)에서 시작하여 10%씩 증분되어 100%(투명)까지 증가하는 투명도 수준 선택사항 중에서 선택할 수 있습니다.

오버레이를 선택하는 경우에는 **레이어 필드**로 선택된 지리 공간적 필드가 포함된 데이터 소스에서 필드를 선택할 수 있습니다. 오버레이 필드의 각각의 값에 대해 다른 투명도 수준이 맵에 표시됩니다. 투명도는 점, 선 또는 다각형에 대해 색상 드롭 다운 목록에서 선택한 색상에 적용됩니다.

데이터 레이블

참고: 육각형 구간화 사용 선택란을 선택하는 경우 이 옵션은 사용할 수 없습니다.

맵에서 데이터 레이블로 사용할 필드를 선택하려면 이 옵션을 사용하십시오. 예를 들어, 다각형 레이어에 적용된 경우 데이터 레이블은 각 다각형의 이름이 포함된 이름 필드일 수 있습니다. 이름 필드를 선택하면 해당 이름이 맵에 표시됩니다.

맵 시각화 모양 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

제목. 그래프 제목으로 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

캡션. 그래프 캡션에 사용할 텍스트를 입력하십시오.

그래프 탐색

편집 모드를 사용하면 그래프의 레이아웃 및 모양을 편집할 수 있지만 탐색 모드를 사용하면 그래프로 표시되는 데이터 및 값을 분석적으로 탐색할 수 있습니다. 탐색의 주요 목표는 데이터를 분석한 후 밴드, 영역 및 표시를 통해 값을 식별하여 선택, 파생 또는 균형 노드를 생성하는 것입니다. 이 모드를 선택하려면 도구 모음 아이콘을 클릭하거나 메뉴에서 **보기 > 탐색 모드**를 선택하십시오.

일부 그래프는 모든 탐색 도구를 사용할 수 있지만 다른 그래프는 하나만 승인합니다. 탐색 모드는 다음을 포함합니다.

- 척도 *x*축을 따라 값을 분할하는 데 사용되는 밴드 정의 및 편집. 자세한 정보는 277 페이지의 『밴드 사용』의 내용을 참조하십시오.
- 직사각형 영역에서 값 그룹을 식별하는 데 사용되는 영역 정의 및 편집. 자세한 정보는 281 페이지의 『영역 사용』의 내용을 참조하십시오.
- 선택 또는 파생 노드를 생성하는 데 사용할 수 있는 값을 직접 선택하기 위해 요소 표시 또는 표시 해제. 자세한 정보는 283 페이지의 『표시된 요소 사용』의 내용을 참조하십시오.
- 스트림에서 사용할 밴드, 영역, 표시된 요소 및 웹 링크에 의해 식별된 값을 사용하여 노드 생성. 자세한 정보는 284 페이지의 『그래프에서 노드 생성』의 내용을 참조하십시오.

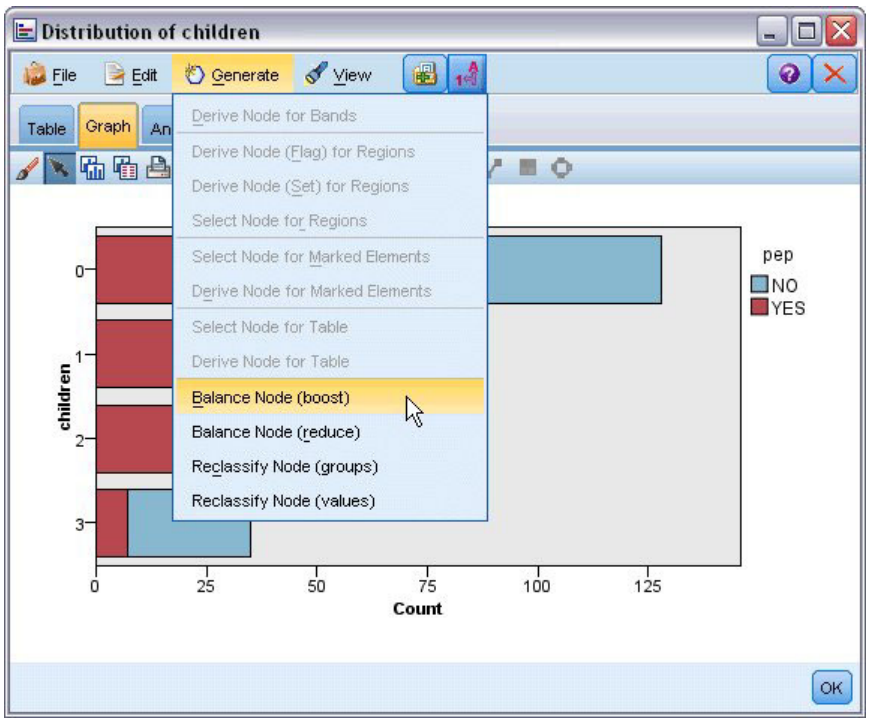


그림 48. 생성 메뉴가 표시되는 그래프

밴드 사용

*x*축에 척도 필드가 있는 그래프에서는 세로 밴드 라인을 그려 *x*축에서 값의 범위를 분할할 수 있습니다. 그래프에 패널이 여럿 있는 경우에는 한 패널에서 그려진 밴드 라인이 다른 패널에도 표시됩니다.

일부 그래프는 밴드를 승인하지 않습니다. 밴드를 가질 수 있는 그래프로는 히스토그램, 막대형 차트 및 분포, 도표(선, 산점도, 시간 등), 콜렉션 및 평가 차트 등이 있습니다. 패널이 있는 그래프에서는 밴드가 모든 패널에 표시됩니다. SPLOM에서는 필드/변수 밴드가 그려진 축이 플립되었기 때문에 가로 밴드 라인이 표시됩니다.

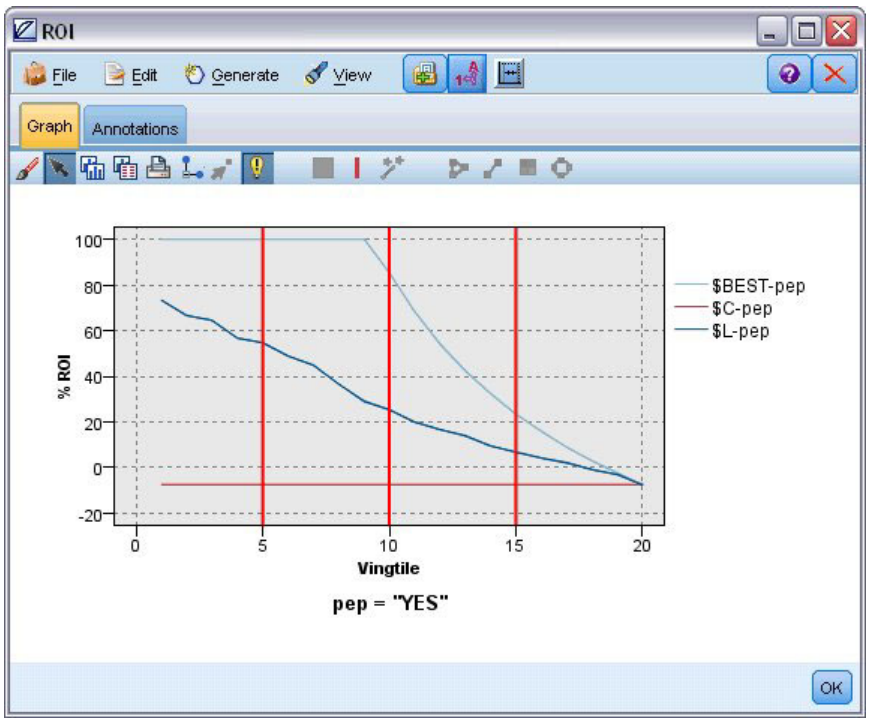


그림 49. 3개의 밴드가 있는 그래프

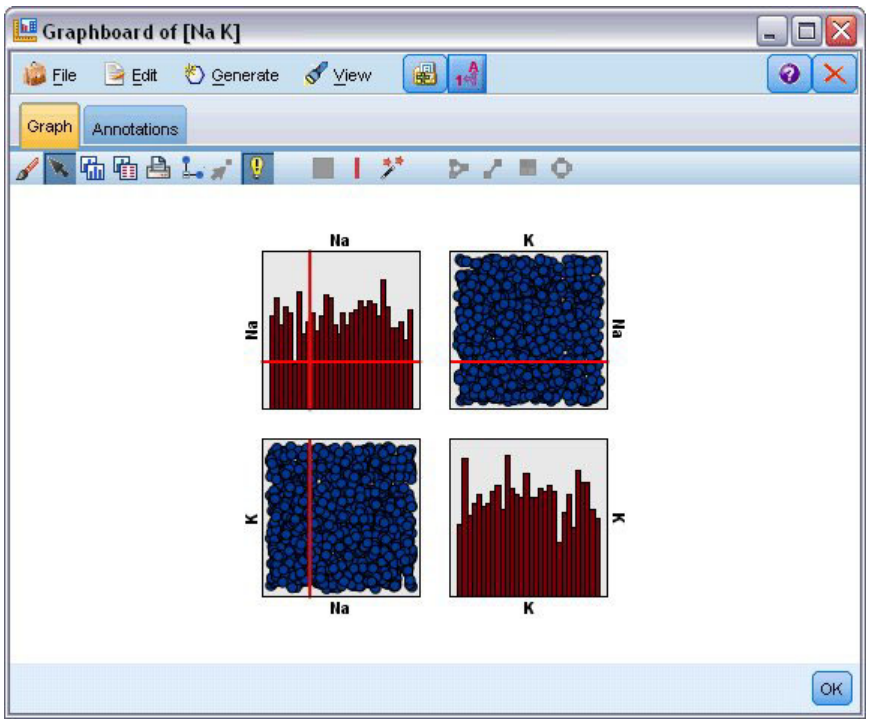


그림 50. 밴드가 있는 SPLOM

밴드 정의

밴드가 없는 그래프에서 밴드 라인을 추가하면 그래프가 2개의 밴드로 분할됩니다. 밴드 라인 값은 왼쪽에서 오른쪽으로 그래프를 읽을 때 두 번째 밴드의 시작점(하한이라고도 함)을 나타냅니다. 마찬가지로 2개의 밴드가 있는 그래프에서 밴드 라인을 추가하면 두 밴드 중 하나가 둘로 분할되어 3개의 밴드가 있게 됩니다. 기본적으로 밴드는 *bandN*으로 이름이 지정됩니다. 여기서 *N*은 *x*축에서 왼쪽부터 오른쪽까지의 밴드 수와 동일합니다.

밴드를 정의하고 나면 밴드를 끌어서 놓아 *x*축에서 밴드의 위치를 재설정할 수 있습니다. 밴드 내부를 마우스 오른쪽 단추로 클릭하여 해당 특정 밴드에 대한 노드 이름 바꾸기, 삭제 또는 생성 등의 작업에 대한 더 많은 단축키를 볼 수 있습니다.

밴드를 정의하려면 다음을 수행하십시오.

1. 탐색 모드에 있는지 확인하십시오. 메뉴에서 **보기 > 탐색 모드**를 선택하십시오.
2. 탐색 모드 도구 모음에서 밴드 그리기 단추를 클릭하십시오.

그림 51. 밴드 그리기 도구 모음 단추

3. 밴드를 승인하는 그래프에서 밴드 라인을 정의할 *x*축 값 지점을 클릭하십시오.

참고: **그래프를 밴드로 분할** 도구 모음 아이콘을 클릭하고 원하는 동등한 밴드의 수를 입력한 후 **분할**을 클릭할 수도 있습니다.

그림 52. 밴드로 분할하기 위한 옵션이 포함된 도구 모음을 펼치는 데 사용되는 분할자 아이콘

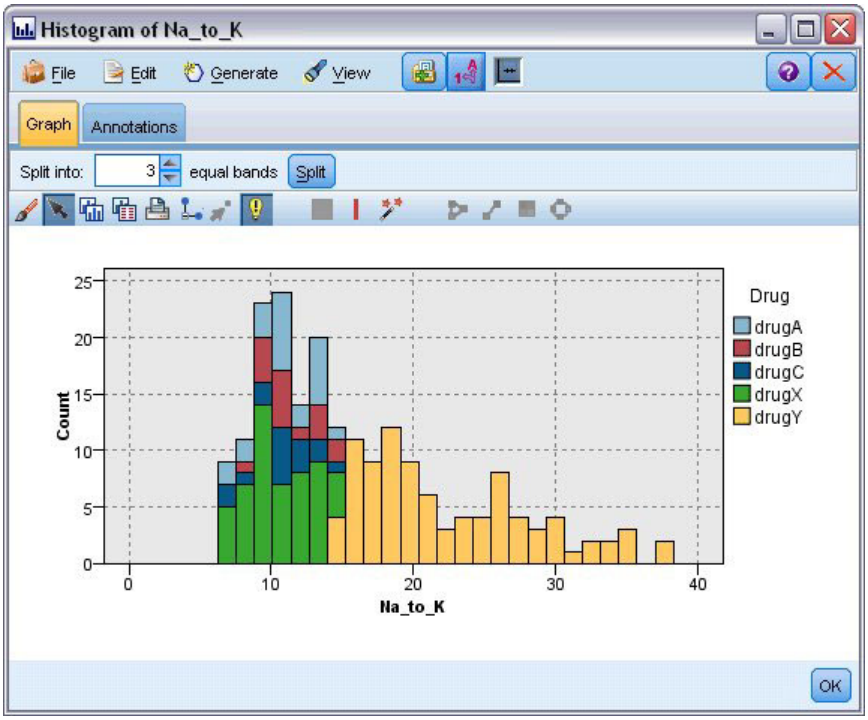


그림 53. 밴드가 사용으로 설정된 동등한 밴드 도구 모음 작성

밴드 편집, 이름 바꾸기 및 삭제

그래프 밴드 편집 대화 상자에서 또는 그래프 자체의 컨텍스트 메뉴를 통해 기존 밴드의 특성을 편집할 수 있습니다.

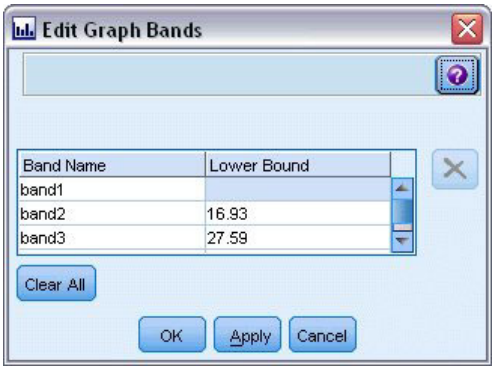


그림 54. 그래프 밴드 편집 대화 상자

밴드를 편집하려면 다음을 수행하십시오.

1. 탐색 모드에 있는지 확인하십시오. 메뉴에서 **보기 > 탐색 모드**를 선택하십시오.
2. 탐색 모드 도구 모음에서 밴드 그리기 단추를 클릭하십시오.
3. 메뉴에서 **편집 > 그래프 밴드**를 선택하십시오. 그래프 밴드 편집 대화 상자가 열립니다.

4. 그래프에 필드가 여럿 있는 경우(예: SPLOM 그래프)에는 드롭 다운 목록에서 원하는 필드를 선택할 수 있습니다.
5. 이름 및 하한을 입력하여 새 밴드를 추가하십시오. Enter 키를 눌러 새 행을 시작하십시오.
6. **하한** 값을 조정하여 밴드의 경계를 편집하십시오.
7. 새 밴드 이름을 입력하여 밴드의 이름을 바꾸십시오.
8. 테이블에서 라인을 선택한 후 삭제 단추를 클릭하여 밴드를 삭제하십시오.
9. **확인**을 클릭하여 변경사항을 적용하고 대화 상자를 닫으십시오.

참고: 밴드의 라인을 마우스 오른쪽 단추로 클릭한 후 컨텍스트 메뉴에서 원하는 옵션을 선택하여 그래프에서 직접 밴드를 삭제하고 밴드의 이름을 바꿀 수도 있습니다.

영역 사용

두 개의 척도(또는 범위) 축이 있는 그래프에서는 영역을 그려서 그리는 직사각형 영역(영역이라고 함) 내에서 값을 그룹화할 수 있습니다. **영역**은 최소 및 최대 *X* 및 *Y* 값으로 설명되는 그래프의 영역입니다. 그래프에 분할창이 여럿 있으면 한 패널에서 그려지는 영역이 다른 패널에도 표시됩니다.

일부 그래프는 영역을 승인하지 않습니다. 영역을 승인하는 그래프로는 도표(선, 산점도, 버블, 시간 등), SPLOM, 콜렉션 등이 있습니다. 이 영역은 X,Y 공간에서 그려지므로 1차원, 3차원 또는 애니메이션 도표에서 정의할 수 없습니다. 패널이 있는 그래프에서는 영역이 모든 패널에 표시됩니다. 산점도 교차표(SPLOM)를 사용하는 경우 대각선 도표는 하나의 척도 필드만 표시하기 때문에 대각선 도표가 아니라 해당 상단 도표에 해당 영역이 표시됩니다.

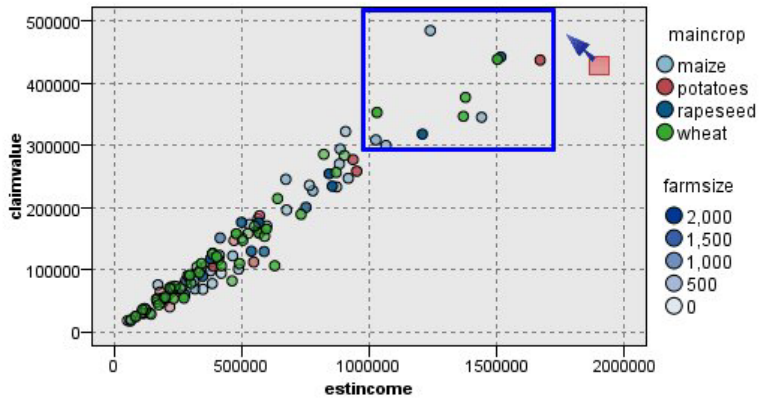


그림 55. 높은 클레임 값의 영역 정의

영역 정의

영역을 정의하는 모든 위치에서 값 그룹을 작성합니다. 기본적으로 각각의 새 영역을 *Region<N>*이라고 합니다. 여기서 *N*은 이미 작성된 영역의 수에 해당합니다.

정의된 영역이 있으면 영역 라인을 마우스 오른쪽 단추로 클릭하여 일부 기본 단축키를 가져올 수 있습니다. 하지만 라인 위가 아니라 영역 내부를 마우스 오른쪽 단추로 클릭하여 해당 특정 영역을 위한 이름 바꾸기, 삭제, 선택 및 파생 노드 생성 등의 작업에 대한 많은 다른 단축키를 볼 수 있습니다.

특정 영역 또는 여러 영역 중 하나에 포함되어 있으면 레코드의 서브세트를 선택할 수 있습니다. 영역에 포함되어 있는지 여부에 따라 플래그 레코드에 대한 파생 노드를 생성하여 레코드에 대한 영역 정보도 통합할 수 있습니다. 자세한 정보는 284 페이지의 『그래프에서 노드 생성』의 내용을 참조하십시오.

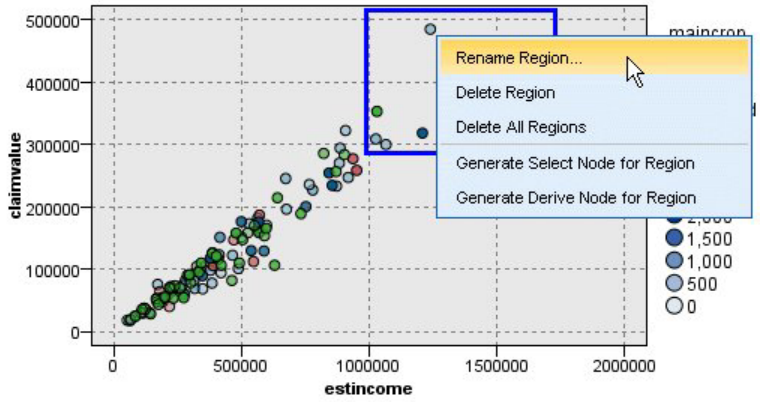


그림 56. 높은 클레임 값의 영역 탐색

영역을 정의하려면 다음을 수행하십시오.

1. 탐색 모드에 있는지 확인하십시오. 메뉴에서 **보기 > 탐색 모드**를 선택하십시오.
2. 탐색 모드 도구 모음에서 영역 그리기 단추를 클릭하십시오.

그림 57. 영역 그리기 도구 모음 단추

3. 영역을 승인하는 그래프에서 마우스를 클릭한 후 끌어서 직사각형 영역을 그리십시오.

영역 편집, 이름 바꾸기 및 삭제

그래프 영역 편집 대화 상자에서 또는 그래프 자체의 컨텍스트 메뉴를 통해 기존 영역의 특성을 편집할 수 있습니다.

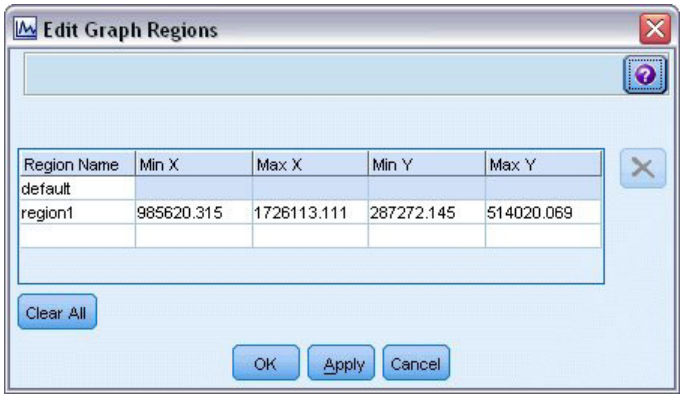


그림 58. 정의된 영역에 대한 특성 지정

영역을 편집하려면 다음을 수행하십시오.

1. 탐색 모드에 있는지 확인하십시오. 메뉴에서 **보기 > 탐색 모드**를 선택하십시오.
2. 탐색 모드 도구 모음에서 영역 그리기 단추를 클릭하십시오.
3. 메뉴에서 **편집 > 그래프 영역**을 선택하십시오. 그래프 영역 편집 대화 상자가 열립니다.
4. 그래프에 필드가 여럿 있는 경우(예: SPLOM 그래프)에는 *필드 A* 및 *필드 B* 열에서 영역에 대한 필드를 정의해야 합니다.
5. 이름을 입력하고 필드 이름을 선택(해당되는 경우)하고 각 필드의 최대 및 최소 경계를 정의하여 새 라인에서 새 영역을 추가하십시오. Enter 키를 눌러 새 행을 시작하십시오.
6. *A* 및 *B*에 대한 **최소** 및 **최대** 값을 조정하여 기존 영역 경계를 편집하십시오.
7. 테이블에서 영역 이름을 변경하여 영역의 이름을 바꾸십시오.
8. 테이블의 라인을 선택한 후 삭제 단추를 클릭하여 영역을 삭제하십시오.
9. **확인**을 클릭하여 변경사항을 적용하고 대화 상자를 닫으십시오.

참고: 또는 영역의 라인을 마우스 오른쪽 단추로 클릭한 후 컨텍스트 메뉴에서 원하는 옵션을 선택하여 그래프에서 직접 영역을 삭제하고 영역의 이름을 바꿀 수 있습니다.

표시된 요소 사용

모든 그래프에서 막대, 조각 및 점 등의 요소를 표시할 수 있습니다. 선은 해당 케이스의 필드를 나타내므로 시간 도표, 다중 도표 및 평가 그래프 이외의 그래프에서는 선, 영역 및 면을 표시할 수 없습니다. 요소를 표시할 때마다 반드시 해당 요소가 나타내는 모든 데이터를 강조표시합니다. 동일한 케이스가 둘 이상의 위치에서 표시되는 그래프(예: SPLOM)에서는 표시가 브러싱과 동의어입니다. 그래프에서 요소를 표시할 수 있으며 밴드 및 영역 내에서도 표시할 수 있습니다. 요소를 표시한 후 편집 모드로 돌아갈 때마다 표시는 계속 표시됩니다.

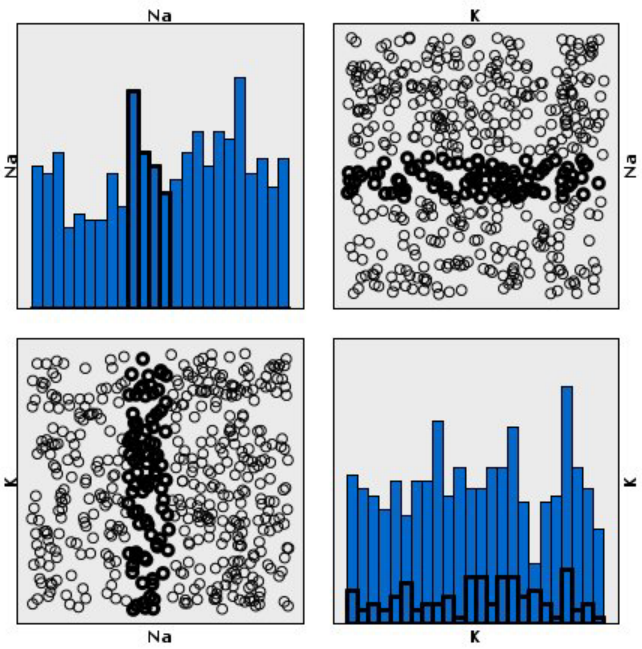


그림 59. SPLOM에서 요소 표시

그래프에서 요소를 클릭하여 요소를 표시하고 표시 해제할 수 있습니다. 처음으로 요소를 클릭하여 표시하면 표시되었음을 나타내는 굵은 테두리 색상과 함께 요소가 표시됩니다. 요소를 다시 클릭하면 테두리가 사라지고 요소가 더 이상 표시되지 않습니다. 여러 요소를 표시하려면 요소를 클릭하는 동안 Ctrl 키를 누르고 있거나 "마술 지팡이"를 사용하여 표시할 각 요소 주위에서 마우스를 끄십시오. Ctrl 키를 누르지 않고 다른 영역 또는 요소를 클릭하면 이전에 표시된 모든 요소가 선택 취소된다는 점을 기억하십시오.

그래프의 표시된 요소에서 선택 및 파생 노드를 생성할 수 있습니다. 자세한 정보는 『그래프에서 노드 생성』의 내용을 참조하십시오.

요소를 표시하려면 다음을 수행하십시오.

1. 탐색 모드에 있는지 확인하십시오. 메뉴에서 **보기 > 탐색 모드**를 선택하십시오.
2. 탐색 모드 도구 모음에서 요소 표시 단추를 클릭하십시오.
3. 필요한 요소를 클릭하거나 마우스를 클릭한 후 끌어서 여러 요소가 포함된 영역 주위에 선을 그리십시오.

그래프에서 노드 생성

IBM SPSS Modeler 그래프가 제공하는 가장 강력한 기능 중 하나는 그래프 또는 그래프 내 선택사항으로부터 노드를 생성하는 기능입니다. 예를 들어, 시간 도표 그래프에서 데이터의 영역 또는 선택사항을 기반으로 파생 및 선택 노드를 생성하여 사실상 데이터의 "서브세트를 작성"할 수 있습니다. 예를 들어, 이 강력한 기능을 사용하여 이상치를 식별하고 제외할 수 있습니다.

밴드를 그릴 수 있을 때마다 파생 노드도 생성할 수 있습니다. 두 개의 척도 축을 가진 그래프에서는 그래프에서 그려진 영역에서 파생 또는 선택 노드를 생성할 수 있습니다. 표시된 요소를 가진 그래프에서는 파생 노드 및 선택 노드를 생성할 수 있으며 일부 경우에는 이 요소에서 필터 노드를 생성할 수 있습니다. 균형 노드

생성은 개수 분포를 표시하는 그래프에 대해 사용으로 설정됩니다.

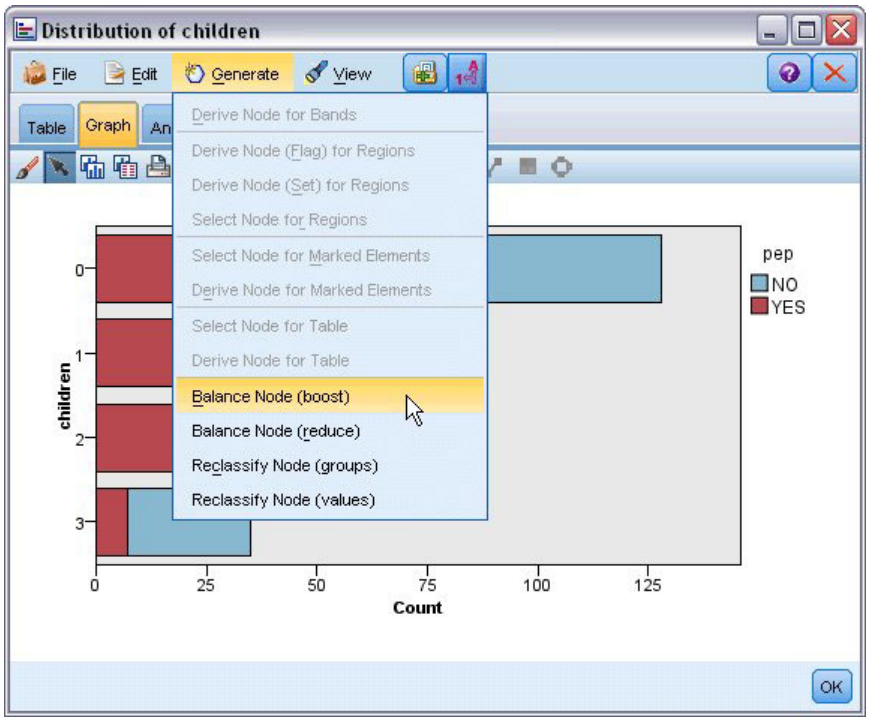


그림 60. 생성 메뉴가 표시되는 그래프

노드를 생성할 때마다 노드를 기존 스트림에 연결할 수 있도록 노드가 스트림 캔버스에 직접 배치됩니다. 그래프에서 선택, 파생, 균형, 필터 및 재분류 노드를 생성할 수 있습니다.

선택 노드

선택 노드는 영역 내 레코드 포함 및 영역 외부의 모든 레코드 제외(다운스트림 처리의 경우 그 반대)에 대해 검증하기 위해 선택 노드를 생성할 수 있습니다.

- **밴드의 경우**. 해당 밴드 내 레코드를 포함하거나 제외하는 선택 노드를 생성할 수 있습니다. 선택 노드에서 사용할 밴드를 선택해야 하므로 **밴드에 대한 선택 노드만**은 컨텍스트 메뉴를 통해서만 사용 가능합니다.
- **영역의 경우**. 영역 내 레코드를 포함하거나 제외하는 선택 노드를 생성할 수 있습니다.
- **표시된 요소의 경우**. 표시된 요소 또는 웹 그래프 링크에 해당하는 레코드를 캡처하는 선택 노드를 생성할 수 있습니다.

파생 노드

파생 노드는 영역, 밴드 및 표시된 요소에서 생성될 수 있습니다. 모든 그래프는 파생 노드를 생성할 수 있습니다. 평가 차트의 경우에는 모델 선택을 위한 대화 상자가 표시됩니다. 웹 그래프의 경우에는 **파생 노드("And")** 및 **파생 노드("Or")**를 사용할 수 있습니다.

- **밴드의 경우**. 밴드 편집 대화 상자에 나열되는 밴드 이름을 범주 이름으로 사용하여 축에 표시된 각각의 간격에 대해 하나의 범주를 생성하는 파생 노드를 생성할 수 있습니다.

- **영역의 경우**. 플래그가 영역 내 레코드에 대해 *T*로 설정되고 모든 영역 외부의 레코드에 대해 *F*로 설정되는 *in_region*이라는 플래그 필드를 작성하는 파생 노드(**플래그로 파생**)를 생성할 수 있습니다. 레코드가 속하는 영역의 이름을 값으로 사용하는 각 레코드에 대해 *region*이라는 새 필드를 가진 각각의 영역에 대한 값을 가진 세트를 생성하는 파생 노드(**세트로 파생**)도 생성할 수 있습니다. 모든 영역 외부의 레코드는 기본 영역의 이름을 수신합니다. 값 이름은 영역 편집 대화 상자에 나열되는 영역 이름이 됩니다.
- **표시된 요소의 경우**. 모든 표시된 요소에 대해 *참*이고 모든 기타 레코드에 대해 *거짓*인 플래그를 계산하는 파생 노드를 생성할 수 있습니다.

균형 노드

균형 노드는 데이터에서 불균형을 정정하기 위해 생성될 수 있습니다(예: 공통 값의 빈도 감소(**균형 노드(감소)** 메뉴 옵션 사용) 또는 빈도가 낮은 값의 발생 부스팅(**균형 노드(부스트)** 메뉴 옵션 사용)). 균형 노드 생성은 개수의 분포를 표시하는 그래프에 대해 사용으로 설정됩니다(예: 히스토그램, 점, 콜렉션, 개수의 막대형, 개수의 원형, 다중 도표).

필터 노드

필터 노드는 그래프에서 표시된 노드 또는 선을 기반으로 필드의 이름을 바꾸고 필드를 필터링하기 위해 생성될 수 있습니다. 평가 차트의 경우 최적 맞춤 선이 필터 노드를 생성하지 않습니다.

재분류 노드

재분류 노드는 값의 코딩을 변경하기 위해 생성될 수 있습니다. 이 옵션은 분포 그래프에 사용됩니다. 그룹에 포함되는지 여부에 따라 표시된 필드의 특정 값의 코딩을 변경하기 위해 **그룹**에 대해 재분류 노드를 생성할 수 있습니다(**테이블** 탭에서 Ctrl+클릭을 사용하여 그룹 선택). 수많은 값의 기존 세트로 데이터의 코딩을 변경하기 위해 **값**에 대해 재분류 노드를 생성할 수도 있습니다(예: 분석을 위해 다양한 회사의 재무 데이터를 병합하기 위해 데이터를 표준 값 세트로 재분류).

참고: 값이 사전 정의되어 있는 경우에는 해당 값을 플랫 파일로서 IBM SPSS Modeler로 읽어오고 분포를 사용하여 모든 값을 표시할 수 있습니다. 그런 다음 차트에서 직접 이 필드에 대한 재분류(값) 노드를 생성하십시오. 그러면 재분류 노드의 *새 값* 열(드롭 다운 목록)에 모든 목표 값이 배치됩니다.

재분류 노드에 대한 옵션을 설정할 때 테이블을 사용하면 이전 세트 값으로부터 사용자가 지정하는 새 값에 대한 명확한 맵핑을 사용할 수 있습니다.

- **원래 값.** 이 열에는 선택 필드의 기존 값이 나열됩니다.
- **새 값.** 이 열을 사용하여 새 범주 값을 입력하거나 드롭 다운 목록에서 값을 선택하십시오. 분포 차트의 값을 사용하여 재분류 노드를 자동으로 생성하는 경우 이 값은 드롭 다운 목록에 포함되어 있습니다. 이를 통해 기존 값을 알려진 값 세트에 신속하게 맵핑할 수 있습니다. 예를 들어, 의료 기관에서 네트워크 또는 로케일을 기반으로 진단을 다르게 그룹화하는 경우가 있습니다. 합병 또는 인수 이후 모든 당사자는 일관된 방식으로 새 데이터 또는 기존 데이터를 재분류해야 합니다. 긴 목록으로부터 각각의 목표 값을 수동으로 입력하는 대신 값의 마스터 목록을 IBM SPSS Modeler로 읽어오고 *진단* 필드에 대한 분포 차트를 실행하

고 이 차트에서 직접 이 필드에 대한 재분류(값) 노드를 생성할 수 있습니다. 이 프로세스를 수행하면 모든 목표 진단 값을 새 값 드롭 다운 목록에서 사용할 수 있습니다.

재분류 노드에 대한 자세한 정보는 166 페이지의 『재분류 노드에 대한 옵션 설정』의 내용을 참조하십시오.

그래프에서 노드 생성

그래프 출력 창의 생성 메뉴를 사용하여 노드를 생성할 수 있습니다. 생성된 노드는 스트림 캔버스에 배치됩니다. 해당 노드를 사용하려면 해당 노드를 기존 스트림에 연결하십시오.

그래프에서 노드를 생성하려면 다음을 수행하십시오.

1. 탐색 모드에 있는지 확인하십시오. 메뉴에서 **보기 > 탐색 모드**를 선택하십시오.
2. 탐색 모드 도구 모음에서 영역 단추를 클릭하십시오.
3. 노드를 생성하기 위해 필요한 밴드, 영역 또는 표시된 요소를 정의하십시오.
4. 생성 메뉴에서 생성할 노드의 유형을 선택하십시오. 가능한 유형만 사용으로 설정됩니다.

참고: 마우스 오른쪽 단추를 클릭한 후 컨텍스트 메뉴에서 원하는 생성 옵션을 선택하여 그래프에서 직접 노드를 생성할 수도 있습니다.

시각화 편집

탐색 모드에서는 시각화로 표시되는 데이터 및 값을 분석적으로 탐색할 수 있는 반면, 편집 모드에서는 시각화의 레이아웃 및 모양을 변경할 수 있습니다. 예를 들어, 조직의 스타일 가이드에 맞게 글꼴 및 색상을 변경할 수 있습니다. 이 모드를 선택하려면 메뉴에서 **보기 > 편집 모드**를 선택하십시오(또는 도구 모음 아이콘 클릭).

편집 모드에서는 시각화 레이아웃의 다양한 측면에 영향을 주는 여러 도구 모음이 제공됩니다. 사용하지 않는 도구 모음이 있는 경우 이러한 도구 모음을 숨겨 대화 상자에서 그래프가 표시되는 공간을 늘릴 수 있습니다. 도구 모음을 선택하거나 선택 취소하려면 보기 메뉴에서 관련 도구 모음 이름을 클릭하십시오.

참고: 시각화에 세부사항을 추가하기 위해 제목, 각주 및 축 레이블을 적용할 수 있습니다. 자세한 정보는 299 페이지의 『제목 및 꼬리말 추가』의 내용을 참조하십시오.

편집 모드에서는 시각화를 편집하는 몇 가지 옵션이 제공됩니다. 다음을 수행할 수 있습니다.

- 텍스트를 편집하고 형식화합니다.
- 프레임 및 그래픽 요소의 채움 색상, 투명도 및 패턴을 변경합니다.
- 경계와 선의 색상 및 대시를 변경합니다.
- 점 요소의 형태와 가로 세로 비율을 회전시키고 변경합니다.
- 그래픽 요소(예: 막대 및 점)의 크기를 변경합니다.
- 여백 및 패딩을 사용하여 항목 주위의 공간을 조정합니다.
- 숫자에 대한 형식화를 지정합니다.
- 축 및 척도 설정을 변경합니다.

- 범주 축에서 범주를 정렬하고 제외시키며 합칩니다.
- 패널의 방향을 설정합니다.
- 좌표계에 변환을 적용합니다.
- 통계, 그래픽 요소 유형 및 충돌 한정자를 변경합니다.
- 범례의 위치를 변경합니다.
- 시각화 스타일시트를 적용합니다.

다음 주제에서는 이러한 다양한 작업을 수행하는 방법에 대해 설명합니다. 그래프를 편집하는 일반 규칙도 읽을 것을 권장합니다.

편집 모드로 전환하는 방법

메뉴에서 다음을 선택하십시오.

보기 > 편집 모드

시각화 편집 일반 규칙

편집 모드

모든 편집은 편집 모드에서 수행됩니다. 편집 모드를 사용하려면 메뉴에서 다음을 선택하십시오.

보기 > 편집 모드

선택

편집에 사용 가능한 옵션은 선택에 따라 다릅니다. 선택하는 항목에 따라 다른 도구 모음 및 특성 팔레트 옵션이 사용됩니다. 사용되는 항목만 현재 선택에 적용됩니다. 예를 들어, 축을 선택하는 경우 특성 팔레트에서 척도, 주 눈금 및 보조 눈금 탭이 사용 가능합니다.

다음은 시각화에서 항목을 선택하는 데 유용한 몇 가지 팁입니다.

- 항목을 클릭하면 항목이 선택됩니다.
- 그래픽 요소(예: 산점도의 점 또는 막대형 차트의 막대)는 한 번 클릭하여 선택합니다. 첫 번째 선택 후 다시 클릭하면 선택 범위가 그래픽 요소 그룹 또는 하나의 그래픽 요소로 좁혀집니다.
- 모든 것을 선택 취소하려면 Esc를 누르십시오.

팔레트

시각화에서 항목을 선택하면 다양한 팔레트가 업데이트되어 선택을 반영합니다. 팔레트에는 선택을 편집할 수 있는 제어가 있습니다. 팔레트는 여러 제어 및 탭이 포함된 패널이거나 도구 모음일 수 있습니다. 팔레트가 숨겨져 있을 수 있으므로 편집에 필요한 팔레트가 표시되었는지 확인하십시오. 보기 메뉴에 현재 표시된 팔레트가 있는지 확인하십시오.

도구 모음 팔레트의 빈 공간 또는 다른 팔레트의 왼쪽을 클릭한 후 끌어와 팔레트의 위치를 바꿀 수 있습니다. 시각적 피드백을 통해 팔레트를 고정할 위치를 알 수 있습니다. 도구 모음이 아닌 팔레트의 경우, 닫기 단추를 클릭하여 팔레트를 숨기고 고정 해제 단추를 클릭하여 팔레트를 별도의 창에 표시할 수도 있습니다. 특정 팔레트에 대한 도움말을 표시하려면 도움말 단추를 클릭하십시오.

자동 설정

일부 설정은 **-자동-** 옵션을 제공합니다. 이는 자동 값이 적용됨을 의미합니다. 사용되는 자동 설정은 고유한 시각화 및 데이터 값에 따라 다릅니다. 값을 입력하여 자동 설정을 대체할 수 있습니다. 자동 설정을 복원하려면 현재 값을 삭제하고 Enter를 누르십시오. 설정이 **-자동-**을 다시 표시합니다.

항목 제거/숨기기

시각화에서 다양한 항목을 제거하거나 숨길 수 있습니다. 예를 들어, 범례 또는 축 레이블을 숨길 수 있습니다. 항목을 삭제하려면 항목을 선택하고 삭제를 누르십시오. 항목이 삭제를 허용하지 않는 경우에는 항목이 삭제되지 않습니다. 실수로 항목을 삭제한 경우에는 Ctrl+Z를 눌러 삭제를 실행 취소하십시오.

상태

일부 도구 모음은 현재 선택 상태를 반영하지만 일부는 그렇지 않습니다. 특성 팔레트는 항상 상태를 반영합니다. 도구 모음이 상태를 반영하지 않는 경우 해당 도구 모음을 설명하는 주제에서 이에 대해 설명합니다.

텍스트 편집 및 형식화

기존 텍스트를 편집하고 전체 텍스트 블록의 형식화를 변경할 수 있습니다. 데이터 값에 직접 연결된 텍스트는 편집할 수 없습니다. 예를 들어, 눈금 레이블은 해당 컨텐츠가 기본 데이터에서 파생되기 때문에 편집할 수 없습니다. 그러나 시각화의 모든 텍스트를 형식화할 수 있습니다.

기존 텍스트 편집 방법

1. 텍스트 블록을 두 번 클릭하십시오. 이 조치는 모든 텍스트를 선택합니다. 텍스트를 편집하는 동안에는 시각화의 다른 부분을 변경할 수 없으므로 모든 도구 모음이 사용되지 않습니다.
2. 텍스트를 입력하여 기존 텍스트를 대체하십시오. 텍스트를 다시 클릭하면 커서가 표시됩니다. 원하는 위치에 커서를 놓고 추가 텍스트를 입력하십시오.

텍스트 형식화 방법

1. 텍스트를 포함하는 프레임을 선택하십시오. 텍스트를 두 번 클릭하지 마십시오.
2. 글꼴 도구 모음을 사용하여 텍스트를 형식화하십시오. 이 도구 모음이 사용되지 않는 경우 텍스트를 포함하는 *프레임*만 선택되었는지 확인하십시오. 텍스트 자체가 선택된 경우에는 이 도구 모음이 사용되지 않습니다.

글꼴과 관련하여 다음을 변경할 수 있습니다.

- 색상
- 글자체(예: Arial 또는 Verdana)

- 크기(다른 단위(예: pc)를 표시하지 않는 한 단위는 pt임)
- 두께
- 텍스트 프레임과 관련한 맞추기

형식화는 프레임 내의 모든 텍스트에 적용됩니다. 텍스트의 특정 블록에 있는 개별 문자 또는 단어의 형식화는 변경할 수 없습니다.

색상, 패턴, 대시 및 투명도 변경

시각화에 있는 다양한 항목에는 채움과 경계가 있습니다. 가장 명확한 예는 막대형 차트의 막대입니다. 막대의 색상은 채움 색상입니다. 또한 막대 주위에 검은색 실선 경계가 있을 수도 있습니다.

시각화에는 좀 덜 명확하지만 역시 채움 색상이 있는 다른 항목들이 있습니다. 채움 색상이 투명하면 채움이 있는지 모를 수도 있습니다. 예를 들어, 축 레이블에 있는 텍스트를 고려하십시오. 이 텍스트가 "떠 있는" 텍스트처럼 보이지만 실제로는 투명한 채움 색상을 갖는 프레임 안에 표시됩니다. 축 레이블을 선택하여 프레임을 볼 수 있습니다.

전체 시각화를 두르는 프레임을 비롯하여 시각화 내의 모든 프레임은 채움 및 경계 스타일을 가질 수 있습니다. 또한 모든 채움은 조정 가능한 불투명도/투명도 수준과 연관됩니다.

색상, 패턴, 대시 및 투명도 변경 방법

1. 형식화할 항목을 선택하십시오. 예를 들어, 막대형 차트의 막대 또는 텍스트를 포함하는 프레임을 선택하십시오. 시각화가 범주형 변수 또는 필드로 분할되는 경우에는 개별 범주에 해당되는 그룹을 선택할 수도 있습니다. 그러면 해당 그룹에 지정된 기본 모양을 변경할 수 있습니다. 예를 들어, 누적 막대형 차트에서 누적 그룹 중 하나의 색상을 변경할 수 있습니다.
2. 채움 색상, 경계 색상 또는 채움 패턴을 변경하려면 색상 도구 모음을 사용하십시오.

   참고: 이 도구 모음은 현재 선택 상태를 반영하지 않습니다.

   색상 또는 채움을 변경하려면 단추를 클릭하여 표시된 옵션을 선택하거나 드롭 다운 화살표를 클릭하여 다른 옵션을 선택하십시오. 색상의 경우, 빨간색 대각선이 그어진 흰색처럼 보이는 하나의 색상이 있습니다. 이것은 투명한 색상입니다. 예를 들어, 막대의 경계를 숨기는 데 이 색상을 사용할 수 있습니다.
   - 첫 번째 단추는 채움 색상을 제어합니다. 색상이 연속형 또는 순서 필드와 연관되어 있는 경우, 이 단추는 데이터의 가장 높은 값과 연관된 색상의 채움 색상을 변경합니다. 특성 팔레트의 색상 탭을 사용하여 가장 낮은 값 및 결측 데이터와 연관된 색상을 변경할 수 있습니다. 요소의 색상은 기본 데이터의 값이 증가함에 따라 낮은 색상에서 높은 색상으로 증분식으로 변경됩니다.
   - 두 번째 단추는 경계 색상을 제어합니다.
   - 세 번째 단추는 채움 패턴을 제어합니다. 채움 패턴은 경계 색상을 사용합니다. 따라서 채움 패턴은 표시되는 경계 색상이 있는 경우에만 표시됩니다.
   - 네 번째 제어는 채움 색상 및 패턴의 불투명도를 제어하는 슬라이더 및 텍스트 상자입니다. 백분율이 낮을수록 불투명도가 낮고 투명도가 높습니다. 100%는 완전히 불투명한 상태입니다(투명도 없음).

3. 경계 또는 선의 대시를 변경하려면 선 도구 모음을 사용하십시오.

   참고: 이 도구 모음은 현재 선택 상태를 반영하지 않습니다.

   다른 도구 모음과 마찬가지로 단추를 클릭하여 표시된 옵션을 선택하거나 드롭다운 화살표를 클릭하여 다른 옵션을 선택하십시오.

점 요소의 형태 및 가로 세로 비율 회전과 변경

점 요소를 회전시키거나 사전 정의된 다른 형태를 지정하거나 가로 세로 비율(너비 대 높이 비율)을 변경할 수 있습니다.

점 요소 수정 방법

1. 점 요소를 선택하십시오. 개별 점 요소의 형태 및 가로 세로 비율은 회전시키거나 변경할 수 없습니다.
2. 기호 도구 모음을 사용하여 점을 수정하십시오.
   - 첫 번째 단추를 사용하여 점의 형태를 변경할 수 있습니다. 드롭 다운 화살표를 클릭하고 사전 정의된 형태를 선택하십시오.
   - 두 번째 단추를 사용하여 점을 특정 컴퍼스 위치로 회전시킬 수 있습니다. 드롭 다운 화살표를 클릭한 후 바늘을 원하는 위치로 끌어오십시오.
   - 세 번째 단추를 사용하여 가로 세로 비율을 변경할 수 있습니다. 드롭 다운 화살표를 클릭한 후 표시되는 직사각형을 끌어오십시오. 직사각형의 형태는 가로 세로 비율을 나타냅니다.

그래픽 요소의 크기 변경

시각화에 있는 그래픽 요소의 *크기*를 변경할 수 있습니다. 여기에는 막대, 선 및 점이 포함됩니다. 변수 또는 필드로 그래픽 요소의 *크기*가 지정되는 경우 지정된 *크기*는 *최소 크기*입니다.

그래픽 요소의 *크기* 변경 방법

1. *크기* 조정할 그래픽 요소를 선택하십시오.
2. 슬라이더를 사용하거나 기호 도구 모음에서 제공되는 해당 옵션에 고유한 *크기*를 입력하십시오. 단위는 다른 단위(아래의 단위 약어 전체 목록 참조)를 표시하지 않는 한 픽셀입니다. 또한 백분율(예: 30%)을 지정할 수도 있으며 이는 그래픽 요소가 사용 가능한 공간을 지정된 백분율만큼 사용하고 있음을 의미합니다. 사용 가능한 공간은 그래픽 요소 유형 및 특정 시각화에 따라 다릅니다.

표 35. 유효한 단위 약어

약어	단위
cm	센티미터
in	인치
mm	밀리미터
pc	파이카
pt	포인트
px	픽셀

여백 및 패딩 지정

시각화에서 프레임 주위 또는 내부에 공간이 너무 많거나 너무 적은 경우 여백 및 패딩 설정을 변경할 수 있습니다. **여백**은 프레임과 프레임 주변에 있는 다른 항목들 사이에 있는 공간의 양입니다. **패딩**은 프레임의 경계와 프레임의 *컨텐츠* 사이에 있는 공간의 양입니다.

여백 및 패딩 지정 방법

1. 여백 및 패딩을 지정할 프레임을 선택하십시오. 이 프레임은 텍스트 프레임, 범례를 두르는 프레임 또는 그래픽 요소(예: 막대 및 점)를 표시하는 데이터 프레임일 수 있습니다.
2. 특성 팔레트의 여백 탭을 사용하여 설정을 지정하십시오. 다른 단위(예: cm 또는 in)를 표시하지 않는 한 모든 *크기*의 단위는 픽셀입니다.

숫자 형식 지정

연속형 축의 눈금 레이블 또는 숫자를 표시하는 데이터 값 레이블에 표시되는 숫자의 형식을 지정할 수 있습니다. 예를 들어, 눈금 레이블에 표시되는 숫자가 천단위로 표시되도록 지정할 수 있습니다.

숫자 형식 지정 방법

1. 연속형 축 눈금 레이블 또는 데이터 값 레이블(숫자를 포함하는 경우)을 선택하십시오.
2. 특성 팔레트에서 **형식** 탭을 클릭하십시오.
3. 원하는 숫자 형식 지정 옵션을 선택하십시오.

   접두부. 숫자 시작 부분에 표시할 문자입니다. 예를 들어, 숫자가 미국 달러 단위의 급여이면 달러 부호($)를 입력하십시오.

   접미부. 숫자 끝 부분에 표시할 문자입니다. 예를 들어, 숫자가 백분율이면 백분율 부호(%)를 입력하십시오.

   최소 정수 자릿수. 십진 표시의 정수 부분에 표시할 최소 자릿수입니다. 실제값에 최소 자릿수가 포함되지 않는 경우 값의 정수 부분은 0으로 채워집니다.

   최대 정수 자릿수. 십진 표시의 정수 부분에 표시할 최대 자릿수입니다. 실제값이 최대 자릿수를 초과하는 경우 값의 정수 부분은 별표로 대체됩니다.

   최소 소수 자릿수. 십진 또는 지수 표시의 소수 부분에 표시할 최소 자릿수입니다. 실제값에 최소 자릿수가 포함되지 않는 경우 값의 소수 부분은 0으로 채워집니다.

   최대 소수 자릿수. 십진 또는 지수 표시의 소수 부분에 표시할 최대 자릿수입니다. 실제값이 최대 자릿수를 초과하는 경우 소수는 해당 자릿수로 반올림됩니다.

   지수 표기법. 숫자를 지수 표기법으로 표시할지 여부입니다. 지수 표기법은 매우 *크거나* 작은 숫자에 유용합니다. **-자동-**을 선택하면 애플리케이션에서 지수 표기법이 적합한 시점을 결정합니다.

스케일링. 척도 요인이며 원래 값을 나누는 하나의 숫자입니다. 숫자가 크지만 숫자를 수용하기 위해 레이블이 너무 많이 확장되는 것을 원하지 않는 경우에는 척도 요인을 사용하십시오. 눈금 레이블의 숫자 형식을 변경하는 경우 축 제목을 편집하여 숫자를 해석하는 방법을 표시하십시오. 예를 들어, 척도 축이 급여를 표시하고 레이블이 30,000, 50,000 및 70,000이라고 가정하십시오. 30, 50 및 70을 표시하기 위해 척도 요인 1000을 입력할 수 있습니다. 그런 다음에는 천단위 텍스트를 포함하도록 척도 축 제목을 편집해야 합니다.

괄호(음의 값). 음의 값에 괄호를 사용하는지 여부를 지정합니다.

숫자 분리. 숫자 그룹 사이에 문자를 표시하는지 여부입니다. 사용하는 컴퓨터의 현재 로케일이 숫자 분리에 사용되는 문자를 결정합니다.

축 및 척도 설정 변경

축과 척도를 수정하는 데 사용하는 몇 가지 옵션이 있습니다.

축 및 척도 설정 변경 방법

1. 축의 특정 부분을 선택하십시오(예: 축 레이블 또는 눈금 레이블).
2. 특성 팔레트에 있는 척도, 주 눈금 및 보조 눈금 탭을 사용하여 축 및 척도 설정을 변경하십시오.

척도 탭

참고: 데이터가 미리 수집되는 그래프(예: 히스토그램)의 경우 척도 탭이 표시되지 않습니다.

유형. 척도가 선형 척도인지 또는 변환 척도인지 여부를 지정합니다. 척도 변환을 통해 더 쉽게 데이터를 이해하고 통계 추론에 필요한 가정을 세울 수 있습니다. 산점도에서는 독립변수(또는 필드)와 종속변수(또는 필드) 간의 관계가 비선형인 경우 변환된 척도를 사용할 수 있습니다. 비대칭 히스토그램을 정규 분포와 유사하게 보이도록 대칭적으로 만드는 데에도 척도 변환을 사용할 수 있습니다. 데이터가 표시되는 척도만 변환하고 실제 데이터는 변환하지 않습니다.

- **선형.** 변환되지 않은 선형 척도를 지정합니다.
- **로그.** 기본-10로그 변환 척도를 지정합니다. 0과 음의 값을 수용하기 위해 이 변환에서는 수정된 로그 함수 버전을 사용합니다. 이 "안전 로그" 함수는 `sign(x) * log(1 + abs(x))`로 정의됩니다. 따라서 `safeLog(-99)`는 다음과 같습니다.

  ```
  sign(-99) * log(1 + abs(-99)) = -1 * log(1 + 99) = -1 * 2 = -2
  ```

- **거듭제곱.** 지수 0.5를 사용하여 거듭제곱 변환 척도를 지정합니다. 음의 값을 수용하기 위해 이 변환에서는 수정된 거듭제곱 함수 버전을 사용합니다. 이 "안전 거듭제곱" 함수는 `sign(x) * pow(abs(x), 0.5)`로 정의됩니다. 따라서 `safePower(-100)`은 다음과 같습니다.

  ```
  sign(-100) * pow(abs(-100), 0.5) = -1* pow(100, 0.5) = -1 * 10 = -10
  ```

최소값/최대값/적당히 낮음/적당히 높음. 척도의 범위를 지정합니다. **적당히 낮음**과 **적당히 높음**을 선택하면 애플리케이션이 데이터를 기반으로 적절한 척도를 선택합니다. 일반적으로 최소값 및 최대값은 최대 및 최소 데

이터 값보다 *크거나* 작은 전체 값이므로 "적당한" 값입니다. 예를 들어, 데이터 범위가 4-92인 경우 척도의 적당히 낮은 값과 높은 값은 실제 데이터 최소값 및 최대값이 아닌 0과 100이 될 수 있습니다. 너무 작은 범위를 설정해 중요한 항목이 숨겨지지 않도록 주의해야 합니다. 또한 **0 포함** 옵션이 선택된 경우 명시적 최소값과 최대값을 설정할 수 없습니다.

낮은 여백/높은 여백. 낮거나 높은 축 끝에 여백을 만듭니다. 여백은 선택된 축과 직각으로 표시됩니다. 단위는 다른 단위(cm 또는 in)를 표시하지 않는 한 픽셀이 됩니다. 예를 들어, 수직축에 대해 **높은 여백**을 5로 설정하면 데이터 프레임 맨 위를 따라 5px의 수평 여백이 만들어집니다.

반전. 척도가 반전되는지 여부를 지정합니다.

0 포함. 척도에 0이 포함됨을 표시합니다. 이 옵션은 일반적으로 막대형 차트에서 가장 작은 막대의 높이에 가까운 값이 아닌 0에서 막대가 시작되도록 하는 데 사용됩니다. 이 옵션을 선택하면 척도 범위에 대해 사용자 정의 최소값 및 최대값을 설정할 수 없으므로 **최소값** 및 **최대값**이 사용되지 않습니다.

주 눈금/보조 눈금 탭

눈금 또는 **눈금 표시**는 축 위에 표시되는 선입니다. 이러한 눈금은 특정 구간 또는 범주에서 값을 표시합니다. **주 눈금**은 레이블이 있는 눈금 표시입니다. 주 눈금은 또한 다른 눈금 표시보다 깁니다. **보조 눈금**은 주 눈금 표시 사이에 표시되는 눈금 표시입니다. 눈금 유형에 고유한 옵션도 있지만 대부분의 옵션은 주 눈금 및 보조 눈금에 사용할 수 있습니다.

눈금 표시. 그래프에 주 눈금을 표시할지 보조 눈금을 표시할지 여부를 지정합니다.

눈금선 표시. 눈금선을 주 눈금에서 표시할지 보조 눈금에서 표시할지 여부를 지정합니다. **눈금선**은 축에서 축까지 전체 그래프를 가로지르는 선입니다.

위치. 축과 관련된 눈금 표시의 위치를 지정합니다.

길이. 눈금 표시의 길이를 지정합니다. 단위는 다른 단위(cm 또는 in)를 표시하지 않는 한 픽셀이 됩니다.

기준. *주 눈금에만 적용합니다.* 첫 번째 주 눈금이 표시되는 지점의 값을 지정합니다.

델타. *주 눈금에만 적용합니다.* 주 눈금 사이의 간격을 지정합니다. 즉, 주 눈금은 *n*번째 값마다 표시되며 여기서 *n*은 델타 값입니다.

구획. *보조 눈금에만 적용합니다.* 주 눈금 사이의 보조 눈금 구획 수를 지정합니다. 보조 눈금 수는 구획 수보다 하나가 적습니다. 예를 들어, 0과 100에 주 눈금이 있다고 가정하십시오. 보조 눈금 구획 수로 2를 입력하면 50에 *한 개*의 보조 눈금이 생기고 0–100 범위를 나누어 *두 개*의 구획이 작성됩니다.

범주 편집

범주 축에서 다음과 같은 방법으로 범주를 편집할 수 있습니다.

- 범주를 표시하는 정렬 순서를 변경합니다.
- 특정 범주를 제외시킵니다.

- 데이터 세트에 나타나지 않는 범주를 추가합니다.
- 작은 범주들을 하나의 범주로 합칩니다.

범주 정렬 순서 변경 방법

1. 범주 축을 선택하십시오. 범주 팔레트가 축의 범주를 표시합니다.

   참고: 팔레트가 표시되지 않으면 팔레트를 사용하도록 설정했는지 확인하십시오. IBM SPSS Modeler의 보기 메뉴에서 **범주**를 선택하십시오.
2. 범주 팔레트의 드롭 다운 목록에서 정렬 옵션을 선택하십시오.

   사용자 정의. 팔레트에 표시되는 순서를 기준으로 범주를 정렬합니다. 화살표 단추를 사용하여 범주를 목록 맨 위, 위, 아래 또는 맨 아래로 이동시키십시오.

   데이터. 데이터 세트의 범주 순서를 기반으로 범주를 정렬합니다.

   이름. 팔레트에 표시되는 이름을 사용하여 알파벳순으로 범주를 정렬합니다. 값 및 레이블을 표시하는 도구 모음 단추가 선택되었는지 여부에 따라 값이거나 레이블일 수 있습니다.

   값. 팔레트에서 괄호 안에 표시되는 값을 사용하여 기본 데이터 값을 기준으로 범주를 정렬합니다. 메타데이터가 있는 데이터 소스(예: IBM SPSS Statistics 데이터 파일)만 이 옵션을 지원합니다.

   통계. 각 범주에 대해 계산된 통계를 기준으로 범주를 정렬합니다. 통계의 예로는 개수, 퍼센트, 평균 등을 들 수 있습니다. 이 옵션은 그래프에서 통계가 사용되는 경우에만 사용할 수 있습니다.

범주 추가 방법

기본적으로 데이터 세트에 나타나는 범주만 사용할 수 있습니다. 필요한 경우 범주를 시각화에 추가할 수 있습니다.

1. 범주 축을 선택하십시오. 범주 팔레트가 축의 범주를 표시합니다.

   참고: 팔레트가 표시되지 않으면 팔레트를 사용하도록 설정했는지 확인하십시오. IBM SPSS Modeler의 보기 메뉴에서 **범주**를 선택하십시오.
2. 범주 팔레트에서 범주 추가 단추를 클릭하십시오.

그림 61. 범주 추가 단추

3. 새 범주 추가 대화 상자에서 범주 이름을 입력하십시오.
4. **확인**을 클릭하십시오.

특정 범주 제외 방법

1. 범주 축을 선택하십시오. 범주 팔레트가 축의 범주를 표시합니다.

참고: 팔레트가 표시되지 않으면 팔레트를 사용하도록 설정했는지 확인하십시오. IBM SPSS Modeler의 보기 메뉴에서 **범주**를 선택하십시오.

2. 범주 팔레트에서 포함 목록에 있는 범주 이름을 선택한 다음 X 단추를 클릭하십시오. 범주를 다시 옮기려면 제외 목록에서 범주 이름을 선택한 다음 목록의 오른쪽에 있는 화살표를 클릭하십시오.

작은 범주를 합치는 방법

너무 작아 개별적으로 표시할 필요가 없는 범주를 결합할 수 있습니다. 예를 들어, 범주가 많은 원형 차트가 있는 경우 백분율이 10 미만인 범주를 합칠 수 있습니다. 가산적 통계인 경우에만 합칠 수 있습니다. 예를 들어, 평균은 가산적이지 않으므로 합칠 수 없습니다. 따라서 평균을 사용한 범주 합치기는 사용할 수 없습니다.

1. 범주 축을 선택하십시오. 범주 팔레트가 축의 범주를 표시합니다.

참고: 팔레트가 표시되지 않으면 팔레트를 사용하도록 설정했는지 확인하십시오. IBM SPSS Modeler의 보기 메뉴에서 **범주**를 선택하십시오.

2. 범주 팔레트에서 **합치기**를 선택하고 백분율을 지정하십시오. 총 백분율이 지정된 수보다 작은 범주가 하나의 범주로 결합됩니다. 백분율은 차트에 표시된 통계를 기반으로 합니다. 합치기는 개수 기반 및 합계(합) 통계에만 사용할 수 있습니다.

패널 방향 변경

시각화에서 패널을 사용하는 경우 패널 방향을 변경할 수 있습니다.

패널 방향 변경 방법

1. 시각화의 임의 부분을 선택하십시오.
2. 특성 팔레트에서 **패널** 탭을 클릭하십시오.
3. **레이아웃**에서 옵션을 선택하십시오.

테이블. 각 개별 값에 지정된 행 또는 열이 있다는 점에서 패널의 레이아웃이 테이블과 유사합니다.

전치. 패널의 레이아웃이 테이블과 유사할 뿐만 아니라 원래 행과 열이 스왑됩니다. 이 옵션은 그래프 자체를 전치시키는 것과는 다릅니다. 이 옵션을 선택할 때 *x*축 및 *y*축은 변경되지 않습니다.

목록. 각 셀이 값 조합을 나타낸다는 점에서 패널의 레이아웃이 목록과 유사합니다. 열 및 행이 더 이상 개별 값에 지정되지 않습니다. 이 옵션을 사용하면 필요한 경우 패널이 랩핑됩니다.

좌표계 변환

다수의 시각화가 평면 직교 좌표계에 표시됩니다. 필요에 따라 좌표계를 변환할 수 있습니다. 예를 들어, 좌표계에 극 변환을 적용하고 기울기 아래 그림자 효과를 추가하며 축을 전치시킬 수 있습니다. 또한 현재 시각화에 변환이 이미 적용된 경우 이러한 변환을 취소할 수 있습니다. 예를 들어, 극 좌표계에 원형 차트가 그려집니다. 극 변환을 실행 취소하고 원형 차트를 직교 좌표계에서 하나의 누적 막대형 차트로 표시할 수 있습니다.

좌표계 변환 방법

1. 변환할 좌표계를 선택하십시오. 개별 그래프의 프레임을 선택하여 좌표계를 선택합니다.

2. 특성 팔레트에서 **좌표** 탭을 클릭하십시오.
3. 좌표계에 적용할 변환을 선택하십시오. 변환을 선택 취소하여 실행을 취소할 수도 있습니다.

   전치. 축 방향을 변경하는 것을 **전치**라고 합니다. 전치는 2차원 시각화에서 수직 축과 수평 축을 스와핑하는 것과 유사합니다.

   극. 극 변환은 그래프 중심으로부터의 특정 거리와 특정 각도에서 그래픽 요소를 그립니다. 원형 차트는 특정 각도에서 개별 막대를 그리는 극 변환이 적용된 1차원 시각화입니다. 방사형 차트는 그래프 중심으로부터의 특정 거리와 특정 각도에서 그래픽 요소를 그리는 극 변환이 적용된 2차원 시각화입니다. 3차원 시각화에는 깊이 차원이 추가로 포함됩니다.

   기울기. 기울기 변환은 그래픽 요소에 3차원 효과를 추가합니다. 이 변환은 그래픽 요소에 깊이를 추가하지만 깊이는 단순한 장식 차원에 불과합니다. 깊이는 특정 테이터 값에 영향을 받지 않습니다.

   동일한 비율. 동일한 비율을 적용하는 경우 각 척도에서 동일한 거리는 데이터 값에서의 차이가 동일함을 나타냅니다. 예를 들어, 두 척도 모두 2cm는 차이값 1000을 나타냅니다.

   변환 전 여백 %. 변환 후 축이 클리핑되는 경우 변환을 적용하기 전에 그래프에 여백을 추가하려 할 수 있습니다. 여백은 좌표계에 변환이 적용되기 전에 차원을 특정 백분율만큼 축소시킵니다. 아래쪽 *x*, 위쪽 *x*, 아래쪽 *y* 및 위쪽 *y* 차원을 나열된 순서대로 제어할 수 있습니다.

   변환 후 여백 %. 그래프의 가로 세로 비율을 변경하려는 경우 변환을 적용한 후에 그래프에 여백을 추가할 수 있습니다. 여백은 좌표계에 변환이 적용된 후 차원을 특정 백분율만큼 축소시킵니다. 그래프에 변환이 적용되지 않는 경우에도 이러한 여백을 적용할 수 있습니다. 아래쪽 *x*, 위쪽 *x*, 아래쪽 *y* 및 위쪽 *y* 차원을 나열된 순서대로 제어할 수 있습니다.

통계 및 그래픽 요소 변경

그래픽 요소를 다른 유형으로 변환하고 그래픽 요소를 그리는 데 사용하는 통계를 변경하며 그래픽 요소가 겹쳐질 때 발생하는 상황을 결정하는 충돌 한정자를 지정할 수 있습니다.

그래픽 요소 변환 방법

1. 변환할 그래픽 요소를 선택하십시오.
2. 특성 팔레트에서 **요소** 탭을 클릭하십시오.
3. 유형 목록에서 새 그래픽 요소 유형을 선택하십시오.

표 36. 그래픽 요소 유형

그래픽 요소 유형	설명
점	특정 데이터 점을 식별하는 마커입니다. 점 요소는 산점도 및 다른 관련 시각화에서 사용됩니다.
구간	특정 데이터 값에서 그려지고 원점과 다른 데이터 값 사이의 공간을 채우는 직사각형 형태입니다. 구간 요소는 막대형 차트와 히스토그램에서 사용됩니다.
선	데이터 값을 연결하는 선입니다.
경로	데이터 세트에 나타나는 순서로 데이터 값을 연결하는 선입니다.

표 36. 그래픽 요소 유형 (계속)

그래픽 요소 유형	설명
영역	데이터 요소를 연결하는 선이며 선과 원점 사이의 영역이 채워집니다.
다각형	데이터 영역을 둘러싼 여러 면으로 구성된 도형입니다. 다각형 요소는 구간화된 산점도 또는 맵에서 사용할 수 있습니다.
스키마	이상값을 나타내는 수염 도표 및 마커가 있는 하나의 상자로 구성된 요소입니다. 스키마 요소는 상자 도표에 사용됩니다.

통계 변경 방법

1. 통계를 변경할 그래픽 요소를 선택하십시오.
2. 특성 팔레트에서 **요소** 탭을 클릭하십시오.

충돌 한정자 지정 방법

충돌 한정자는 그래픽 요소가 겹쳐질 때 발생하는 상황을 결정합니다.

1. 충돌 한정자를 지정할 그래픽 요소를 선택하십시오.
2. 특성 팔레트에서 **요소** 탭을 클릭하십시오.
3. 수정자 드롭 다운 목록에서 충돌 한정자를 선택하십시오. **-자동-**을 선택하면 애플리케이션이 그래픽 요소 유형 및 통계에 적합한 충돌 한정자를 결정합니다.

   오버레이. 값이 동일하면 서로의 위에 그래픽 요소를 그립니다.

   누적. 데이터 값이 동일한 경우 일반적으로 겹쳐지는 그래픽 요소를 누적시킵니다.

   닷지. 같은 값에서 표시되는 다른 그래픽 요소 위에 그래픽 요소를 겹치는 대신 그 옆으로 이동시킵니다. 그래픽 요소가 대칭적으로 배열됩니다. 즉, 그래픽 요소가 중앙 위치의 반대편으로 이동합니다. 닷지는 군집화와 유사합니다.

   적재. 같은 값에서 표시되는 다른 그래픽 요소 위에 그래픽 요소를 겹치는 대신 그 옆으로 이동시킵니다. 그래픽 요소가 비대칭적으로 배열됩니다. 즉, 맨 아래의 그래픽 요소가 척도의 특정 값에 위치하고 그래픽 요소가 서로의 위에 적재됩니다.

   지터(정규). 정규 분포를 사용하여 동일한 데이터 값에 있는 그래픽 요소의 위치를 무작위로 바꿉니다.

   지터(균등). 균등 분포를 사용하여 동일한 데이터 값에 있는 그래픽 요소의 위치를 무작위로 바꿉니다.

범례 위치 변경

그래프에 범례가 포함되는 경우 범례는 일반적으로 그래프의 오른쪽에 표시됩니다. 필요한 경우 이 위치를 변경할 수 있습니다.

범례 위치 변경 방법

1. 범례를 선택하십시오.
2. 특성 팔레트에서 **범례** 탭을 클릭하십시오.

3. 위치를 선택하십시오.

시각화 및 시각화 데이터 복사

일반 팔레트에는 시각화와 해당 데이터를 복사하는 단추가 있습니다.

그림 62. 시각화 복사 단추

시각화 복사. 이 조치는 시각화를 클립보드에 이미지로 복사합니다. 여러 이미지 형식을 사용할 수 있습니다. 이미지를 다른 애플리케이션에 붙여넣을 때는 "선택하여 붙여넣기" 옵션을 선택하여 사용 가능한 이미지 형식 중 하나를 선택할 수 있습니다.

그림 63. 시각화 데이터 복사 단추

시각화 데이터 복사. 이 조치는 시각화를 작성하는 데 사용되는 기본 데이터를 복사합니다. 데이터를 일반 텍스트 또는 HTML 형식의 텍스트로 클립보드에 복사합니다. 데이터를 다른 애플리케이션에 붙여넣을 때는 "선택하여 붙여넣기" 옵션을 선택하여 이러한 형식 중 하나를 선택할 수 있습니다.

그래프보드 편집기 키보드 단축키

표 37. 키보드 단축키

단축키	기능
Ctrl+Space	탐색 모드와 편집 모드 간 전환
Delete	시각화 항목 삭제
Ctrl+Z	실행 취소
Ctrl+Y	다시 실행
F2	그래프에서 항목을 선택하기 위한 아웃라인 표시

제목 및 꼬리말 추가

모든 그래프 유형에 대해 그래프에 표시되는 항목의 식별을 돕기 위해 고유 제목, 꼬리말 또는 축 레이블을 추가할 수 있습니다.

그래프에 제목 추가

1. 메뉴에서 **편집 > 그래프 제목 추가**를 선택하십시오. **<TITLE>**이 포함된 텍스트 상자가 그래프 위에 표시됩니다.
2. 편집 모드에 있는지 확인하십시오. 메뉴에서 **보기 > 편집 모드**를 선택하십시오.

3. **<TITLE>** 텍스트를 두 번 클릭하십시오.
4. 필요한 제목을 입력한 후 Return을 누르십시오.

그래프에 꼬리말 추가

1. 메뉴에서 **편집 > 그래프 꼬리말 추가**를 선택하십시오. **<FOOTNOTE>**가 포함된 텍스트 상자가 그래프 아래에 표시됩니다.
2. 편집 모드에 있는지 확인하십시오. 메뉴에서 **보기 > 편집 모드**를 선택하십시오.
3. **<FOOTNOTE>** 텍스트를 두 번 클릭하십시오.
4. 필요한 제목을 입력한 후 Return을 누르십시오.

그래프 스타일시트 사용

색상, 글꼴, 기호, 선 굵기 등의 기본 그래프 표시 정보는 스타일시트에 의해 제어됩니다. IBM SPSS Modeler와 함께 제공되는 기본 스타일시트가 있지만 필요한 경우 변경할 수 있습니다. 예를 들어, 그래프에서 사용할 프리젠테이션에 대한 공동 색상 구성표를 가질 수 있습니다. 자세한 정보는 287 페이지의 『시각화 편집』의 내용을 참조하십시오.

그래프 노드에서 편집 모드를 사용하여 그래프 모양에 대해 스타일 변경사항을 작성할 수 있습니다. 그런 다음 **편집 > 스타일** 메뉴를 사용하여 변경사항을 현재 그래프 노드에서 이후에 생성되는 모든 그래프에 적용할 스타일시트로 저장하거나 IBM SPSS Modeler를 사용하여 생성하는 모든 그래프에 대한 새로운 기본 스타일시트로 저장할 수 있습니다.

편집 메뉴의 **스타일** 옵션에서는 다섯 가지 스타일시트 옵션을 사용할 수 있습니다.

- **스타일시트 전환.** 그래프 모양을 변경하기 위해 선택할 수 있는 저장된 다양한 스타일시트의 목록을 표시합니다. 자세한 정보는 『스타일시트 적용』의 내용을 참조하십시오.
- **노드에서 스타일 저장.** 현재 스트림의 동일한 그래프 노드에서 작성되는 향후 그래프에 적용되도록 수정사항을 선택된 그래프의 스타일에 저장합니다.
- **스타일을 기본값으로 저장.** 모든 스트림의 모든 그래프 노드에서 작성되는 모든 향후 그래프에 적용되도록 수정사항을 선택된 그래프의 스타일에 저장합니다. 이 옵션을 선택한 후에는 **기본 스타일 적용**을 사용하여 동일한 스타일을 사용하도록 다른 기존 그래프를 변경할 수 있습니다.
- **기본 스타일 적용.** 선택된 그래프의 스타일을 현재 기본 스타일로 저장되는 스타일로 변경합니다.
- **원본 스타일 적용.** 그래프의 스타일을 다시 원래 기본값으로 제공된 스타일로 변경합니다.

스타일시트 적용

시각화의 스타일 특성을 지정하는 시각화 스타일시트를 적용할 수 있습니다. 예를 들어, 스타일시트는 글꼴, 대시 및 색상을 정의할 수 있습니다. 스타일시트를 사용하면 어느 정도까지는 수동으로 수행해야 할 편집을 쉽게 수행할 수 있습니다. 그러나 스타일시트는 *스타일* 변경에 한정됩니다. 범례 위치 또는 척도 범위와 같은 다른 변경은 스타일시트에 저장되지 않습니다.

스타일시트 적용 방법

1. 메뉴에서 다음을 선택하십시오.

편집 > 스타일 > 스타일시트 전환

2. 스타일시트 전환 대화 상자를 사용하여 스타일시트를 선택하십시오.
3. 대화 상자를 닫지 않고 시각화에 스타일시트를 적용하려면 **적용**을 클릭하십시오. 스타일시트를 적용하고 대화 상자를 닫으려면 **확인**을 클릭하십시오.

스타일시트 전환/선택 대화 상자

대화 상자의 맨 위에 있는 테이블은 현재 사용 가능한 모든 시각화 스타일시트를 나열합니다. 일부 스타일시트는 사전 설치되었고 나머지 스타일시트는 IBM SPSS Visualization Designer (별매품)에서 작성되었을 수 있습니다.

대화 상자 맨 아래에서는 표본 데이터를 사용한 시각화의 예를 표시합니다. 스타일시트 중 하나를 선택하여 해당 스타일을 시각화 예에 적용하십시오. 이러한 예는 스타일시트가 어떻게 실제 시각화에 영향을 주는지 판별하는 데 도움이 됩니다.

대화 상자는 또한 다음과 같은 옵션을 제공합니다.

기존 스타일. 기본적으로 스타일시트는 시각화의 모든 스타일을 겹쳐쓸 수 있습니다. 이 작동을 변경할 수 있습니다.

- **모든 스타일 겹쳐쓰기**. 스타일시트를 적용할 때 현재 편집 세션 동안 시각화에서 수정된 스타일을 포함하여 시각화의 모든 스타일을 겹쳐씁니다.
- **수정된 스타일 유지**. 스타일시트를 적용할 때 현재 편집 세션 동안 시각화에서 수정되지 않은 스타일만 겹쳐씁니다. 현재 편집 세션 동안 수정된 스타일은 유지됩니다.

관리. 컴퓨터에서 시각화 템플리트 , 스타일시트 및 맵을 관리합니다. 로컬 시스템에서 시각화 템플리트 , 스타일시트 및 맵을 가져오고, 내보내고, 이름을 바꾸고, 삭제할 수 있습니다. 자세한 정보는 223 페이지의 『템플리트, 스타일시트 및 맵 파일 관리』 주제를 참조하십시오.

위치. 시각화 템플리트 , 스타일시트 및 맵이 저장된 위치를 변경합니다. 현재 위치는 단추의 오른쪽에 표시됩니다. 자세한 정보는 222 페이지의 『템플리트, 스타일시트 및 맵 위치 설정』의 내용을 참조하십시오.

그래프 인쇄, 저장, 복사 및 내보내기

각각의 그래프에는 그래프를 저장하거나 인쇄하거나 다른 형식으로 내보낼 수 있게 하는 다수의 옵션이 있습니다. 이 옵션은 대부분 파일 메뉴에서 사용할 수 있습니다. 또한 편집 메뉴에서 다른 애플리케이션에서 사용하기 위해 그래프 또는 그래프 내 데이터를 복사하도록 선택할 수 있습니다.

인쇄

그래프를 인쇄하려면 **인쇄** 메뉴 항목 또는 단추를 사용하십시오. 인쇄하기 전에 **페이지 설정** 및 **인쇄 미리보기**를 사용하여 인쇄 옵션을 설정하고 출력을 미리 볼 수 있습니다.

그래프 저장

그래프를 IBM SPSS Modeler 출력 파일(*.*cou*)에 저장하려면 메뉴에서 **파일 > 저장** 또는 **파일 > 다른 이름으로 저장**을 선택하십시오.

또는

그래프를 리포지토리에 저장하려면 메뉴에서 **파일 > 출력 저장**을 선택하십시오.

그래프 복사

MS Word 또는 MS PowerPoint 등의 다른 애플리케이션에서 사용하기 위해 그래프를 복사하려면 메뉴에서 **편집 > 그래프 복사**를 선택하십시오.

데이터 복사

MS Excel 또는 MS Word 등의 다른 애플리케이션에서 사용하기 위해 데이터를 복사하려면 메뉴에서 **편집 > 데이터 복사**를 선택하십시오. 기본적으로 데이터의 형식은 HTML로 지정됩니다. 붙여넣을 때 다른 형식 옵션을 보려면 다른 애플리케이션에서 **선택하여 붙여넣기**를 사용하십시오.

그래프 내보내기

그래프 내보내기 옵션을 사용하면 다른 IBM SPSS Statistics 애플리케이션에서 사용하기 위해 비트맵(*.bmp*), JPEG(*.jpg*), PNG(*.png*), HTML(*.html*) 또는 ViZml 문서(*.xml*) 형식 중 하나로 그래프를 내보낼 수 있습니다.

그래프를 내보내려면 메뉴에서 **파일 > 그래프 내보내기**를 선택한 후 형식을 선택하십시오.

테이블 내보내기

테이블 내보내기 옵션을 사용하면 탭 구분(*.tab*), 쉼표 구분(*.csv*) 또는 HTML(*.html*) 형식 중 하나로 테이블을 내보낼 수 있습니다.

테이블을 내보내려면 메뉴에서 **파일 > 테이블 내보내기**를 선택한 후 형식을 선택하십시오.

제 6 장 출력 노드

출력 노드 개요

출력 노드는 데이터 및 모형에 대한 정보를 얻을 수 있는 방법을 제공합니다. 또한 기타 소프트웨어 도구와 접속할 수 있도록 데이터를 다양한 형식으로 내보낼 수 있는 메커니즘을 제공합니다.

다음 출력 노드를 사용할 수 있습니다.

테이블 노드는 데이터를 표 형식으로 표시하는데, 이것을 파일에 쓸 수도 있습니다. 이것은 쉽게 읽을 수 있는 양식으로 데이터 값을 조사하거나 내보내야 할 때 유용합니다.

행렬 노드는 필드 사이의 관계를 표시하는 테이블을 작성합니다. 두 기호 필드 사이의 관계를 표시하기 위해 가장 일반적으로 사용하지만, 플래그 필드나 수치 필드 사이의 관계도 표시할 수 있습니다.

분석 노드는 정확한 예측을 생성하기 위한 예측 모델의 능력을 평가합니다. 분석 노드는 하나 이상의 모델 너깃에 대해 예측값과 실제 값 사이의 다양한 비교를 수행합니다. 또한 예측 모델을 서로 비교할 수도 있습니다.

데이터 검토 노드는 요약 통계량, 각 필드에 대한 히스토그램과 분포뿐 아니라 이상값, 결측값, 극단값에 대한 정보를 포함하여 데이터에 대한 포괄적인 정보를 간략하게 제공합니다. 결과는 전체 크기 그래프 및 데이터 준비 노드를 생성하기 위해 정렬하고 사용할 수 있는 읽기 쉬운 행렬로 표시됩니다.

변환 노드를 사용하면 선택된 필드에 적용하기 전에 변환 결과를 선택하고 시각적으로 미리볼 수 있습니다.

통계량 노드는 수치 필드에 관한 기본 요약 정보를 제공합니다. 개별 필드에 대한 요약 통계량 및 필드 사이의 상관계수를 계산합니다.

평균 노드는 독립 집단 사이 또는 관련된 필드의 쌍 사이의 평균을 비교하여 상당한 차이가 존재하는지 여부를 검정합니다. 예를 들어, 프로모션을 실행하기 전후의 평균 수익을 비교하거나 프로모션을 받지 않은 고객과 받은 고객으로부터의 수익을 비교할 수 있습니다.

보고서 노드는 고정 텍스트뿐 아니라 데이터 및 데이터로부터 파생된 기타 표현식을 포함한 형식화된 보고서를 작성합니다. 텍스트 템플리트를 사용하여 보고서의 형식을 지정하여 고정 텍스트 및 데이터 출력 생성을 정의합니다. 템플리트에서 HTML 태그를 사용하고 출력 탭에서 옵션을 설정하여 사용자 정의 텍스트 형식화를 제공할 수 있습니다. 템플리트에서 CLEM 표현식을 사용하여 데이터 값과 기타 조건부 출력을 포함할 수 있습니다.

전역값 설정 노드는 데이터를 스캔하고 CLEM 표현식에서 사용할 수 있는 요약 값을 계산합니다. 예를 들어, 이 노드를 사용하여 *age*라는 필드에 대한 통계량을 계산한 후 `@GLOBAL_MEAN(age)` 함수를 삽입하여 CLEM 표현식에서 *age*의 전체 평균을 사용할 수 있습니다.

시뮬레이션 적합 노드는 각 필드에 있는 데이터의 통계 분포를 분석하고 각 필드에 최상의 적합 분포가 지정된 시뮬레이션 생성 노드를 생성(또는 업데이트)합니다. 시뮬레이션 생성 노드를 사용하여 시뮬레이션된 데이터를 생성할 수 있습니다.

시뮬레이션 평가 노드는 지정된 예측 목표 필드를 평가하고 목표 필드에 관한 분포 및 상관관계 정보를 제공합니다.

출력 관리

출력 관리자는 IBM SPSS Modeler 세션 동안 생성된 차트, 그래프 및 테이블을 표시합니다. 출력 관리자에서 출력을 두 번 클릭하여 항상 출력을 다시 열 수 있습니다. 해당 스트림 또는 노드를 재실행하지 않아도 됩니다.

출력 관리자를 보려면 다음을 수행하십시오.

보기 메뉴를 열고 **관리자**를 선택하십시오. **출력** 탭을 클릭하십시오.

출력 관리자에서 다음을 수행할 수 있습니다.

- 히스토그램, 평가 차트, 테이블 등의 기존 출력 오브젝트 표시
- 출력 오브젝트의 이름 바꾸기
- 디스크 또는 IBM SPSS Collaboration and Deployment Services Repository에 출력 오브젝트 저장(사용 가능한 경우).
- 현재 프로젝트에 출력 파일 추가
- 현재 세션에서 저장되지 않은 출력 오브젝트 삭제
- 저장된 출력 오브젝트 열기 또는 IBM SPSS Collaboration and Deployment Services Repository에서 저장된 출력 오브젝트 검색(사용 가능한 경우)

이 옵션에 액세스하려면 출력 탭을 마우스 오른쪽 단추로 클릭하십시오.

출력 보기

화면 출력은 출력 브라우저 창에 표시됩니다. 출력 브라우저 창에는 출력을 인쇄 또는 저장하거나 다른 형식으로 내보낼 수 있는 메뉴 세트가 있습니다. 출력 유형에 따라 특정 옵션은 다를 수 있습니다.

데이터 인쇄, 저장 및 내보내기. 자세한 정보는 다음과 같이 사용 가능합니다.

- 출력을 인쇄하려면 **인쇄** 메뉴 옵션 또는 단추를 사용하십시오. 인쇄하기 전에 **페이지 설정** 및 **인쇄 미리보기**를 사용하여 인쇄 옵션을 설정하고 출력을 미리 볼 수 있습니다.
- 출력을 IBM SPSS Modeler 출력 파일(*.cou*)에 저장하려면 파일 메뉴에서 **저장** 또는 **다른 이름으로 저장**을 선택하십시오.
- 텍스트 또는 HTML 등의 다른 형식으로 출력을 저장하려면 파일 메뉴에서 **내보내기**를 선택하십시오. 자세한 정보는 307 페이지의 『출력 내보내기』의 내용을 참조하십시오.

  출력에 해당 형식으로 내보내기에 적합한 데이터가 포함된 경우에만 해당 형식을 선택할 수 있습니다. 예를 들어, 의사결정 트리의 내용은 텍스트로 내보낼 수 있으나 K-평균 모델의 내용은 텍스트로는 의미가 전달되지 않습니다.
- 다른 사용자가 IBM SPSS Collaboration and Deployment Services Deployment Portal을 사용하여 출력을 볼 수 있도록 공유 리포지토리에 출력을 저장하려면 파일 메뉴에서 **웹에 출판**을 선택하십시오. 이 옵션을 사용하려면 IBM SPSS Collaboration and Deployment Services에 대한 별도의 라이센스가 필요합니다.

셀 및 열 선택. 편집 메뉴에는 현재 출력 유형에 맞게 셀 및 열을 선택, 선택 취소 및 복사하기 위한 다양한 옵션이 포함되어 있습니다. 자세한 정보는 307 페이지의 『셀 및 열 선택』의 내용을 참조하십시오.

새 노드 생성. 생성 메뉴를 사용하면 출력 브라우저의 내용을 기반으로 하여 새 노드를 생성할 수 있습니다. 옵션은 출력 유형 및 현재 선택된 출력 내의 항목에 따라 다릅니다. 특정 유형의 출력에 대한 노드 생성 옵션에 대한 세부사항은 해당 출력에 대한 문서를 참조하십시오.

웹에 출판

웹에 출판 기능을 사용하면 특정 유형의 스트림 출력을 IBM SPSS Collaboration and Deployment Services의 기초가 되는 중앙 공유 IBM SPSS Collaboration and Deployment Services Repository에 출판할 수 있습니다. 이 옵션을 사용하면 이 출력을 볼 필요가 있는 다른 사용자가 IBM SPSS Modeler를 설치할 필요 없이 인터넷 액세스 및 IBM SPSS Collaboration and Deployment Services 계정을 사용하여 출력을 볼 수 있습니다.

다음은 웹에 출판 기능을 지원하는 IBM SPSS Modeler 노드를 나열한 표입니다. 이러한 노드의 출력은 출력 오브젝트(*.cou*) 형식으로 IBM SPSS Collaboration and Deployment Services Repository에 저장되며 IBM SPSS Collaboration and Deployment Services Deployment Portal에서 직접 볼 수 있습니다.

기타 유형의 출력은 사용자의 시스템에 관련 애플리케이션(예를 들어, 스트림 오브젝트의 경우 IBM SPSS Modeler)이 설치된 경우에만 볼 수 있습니다.

표 38. 웹에 출판을 지원하는 노드.

노드 유형	노드
그래프	모두
출력	테이블
	교차표
	데이터 검토
	변환
	평균
	분석
	통계량
	보고서(HTML)
IBM SPSS Statistics	통계량 출력

웹에 출력 출판

웹에 출력을 출판하려면 다음을 수행하십시오.

1. IBM SPSS Modeler 스트림에서 표에 나열된 노드 중 하나를 실행하십시오. 그러면 새 창에 출력 오브젝트(표, 교차표 또는 보고서 오브젝트 등)가 작성됩니다.
2. 출력 오브젝트 창에서 다음을 선택하십시오.

   파일 > 웹에 출판

   참고: 표준 웹 브라우저와 함께 사용하도록 단순 HTML 파일만 내보내려면 파일 메뉴에서 **내보내기**를 선택하고 **HTML**을 선택하십시오.
3. IBM SPSS Collaboration and Deployment Services Repository에 연결하십시오.

   연결되면 여러 가지 저장 공간 옵션을 제공하는 리포지토리: 저장 대화 상자가 표시됩니다.
4. 원하는 저장 공간 옵션을 선택하였으면 **저장**을 클릭하십시오.

웹을 통해 출판된 출력 보기

이 기능을 사용하려면 IBM SPSS Collaboration and Deployment Services 계정이 설정되어 있어야 합니다. 볼 오브젝트 유형과 관련된 애플리케이션(IBM SPSS Modeler 또는 IBM SPSS Statistics)이 설치되어 있는 경우에는 브라우저가 아니라 애플리케이션 자체에 출력이 표시됩니다.

웹을 통해 출판된 출력을 보려면 다음을 수행하십시오.

1. 브라우저에서 *http://<repos_host>:<repos_port>/peb*으로 이동하십시오.

   여기서, *repos_host* 및 *repos_port*는 IBM SPSS Collaboration and Deployment Services 호스트의 호스트 이름 및 포트 번호입니다.
2. IBM SPSS Collaboration and Deployment Services 계정에 대한 로그인 세부사항을 입력하십시오.
3. **내용 리포지토리**를 클릭하십시오.
4. 볼 오브젝트로 이동하거나 검색하십시오.

5. 오브젝트 이름을 클릭하십시오. 그래프 등의 일부 오브젝트 유형에 대해서는 오브젝트가 브라우저에서 렌더링되는 동안 보류될 수 있습니다.

HTML 브라우저에서 출력 보기

선형, 로지스틱 및 PCA/Factor 모델 너깃의 고급 탭에서 Internet Explorer와 같은 별도의 브라우저에서 표시되는 정보를 볼 수 있습니다. 정보는 회사 인트라넷 또는 인터넷 사이트와 같이 어디서나 저장하고 재사용할 수 있는 HTML 등의 출력입니다.

브라우저에 정보를 표시하려면 모델 너깃의 고급 탭의 왼쪽 상단에 있는 모델 아이콘 아래의 시작 단추를 클릭하십시오.

출력 내보내기

출력 브라우저 창에서 출력을 텍스트 또는 HTML과 같은 다른 형식으로 내보내도록 선택할 수 있습니다. 내보내기 형식은 출력의 유형에 따라 다르지만 일반적으로 출력을 생성하기 위해 사용한 노드에서 **파일에 저장**을 선택할 때 사용할 수 있는 파일 유형 옵션과 유사합니다.

참고: 출력에 해당 형식으로 내보내기에 적합한 데이터가 포함된 경우에만 해당 형식을 선택할 수 있습니다. 예를 들어, 의사결정 트리의 내용은 텍스트로 내보낼 수 있으나 K-평균 모델의 내용은 텍스트로는 의미가 전달되지 않습니다.

출력을 내보내려면 다음을 수행하십시오.

1. 출력 브라우저에서 파일 메뉴를 열고 **내보내기**를 선택하십시오. 그런 다음 생성할 파일 유형을 선택하십시오.
   - **탭 구분 데이터(*.tab).** 이 옵션은 데이터 값을 포함하는 형식화된 텍스트 파일을 생성합니다. 이 스타일은 정보를 다른 애플리케이션으로 가져올 수 있는 일반 텍스트 표시를 생성하는 경우에 유용할 때가 많습니다. 이 옵션은 표, 교차표 및 평균 노드에 대해 사용할 수 있습니다.
   - **콤마 구분 데이터(*.dat).** 이 옵션은 데이터 값을 포함하는 콤마 구분 텍스트 파일을 생성합니다. 이 유형은 스프레드시트 또는 기타 데이터 분석 애플리케이션으로 가져올 수 있는 데이터 파일을 생성하는 빠른 방법으로 유용한 경우가 많습니다. 이 옵션은 표, 교차표 및 평균 노드에 대해 사용할 수 있습니다.
   - **전치된 탭 구분 데이터(*.tab).** 이 옵션은 탭 구분 데이터와 동일하나 데이터가 전치되어 행이 필드를 나타내고 열이 레코드를 나타낸다는 점만 다릅니다.
   - **전치된 콤마 구분 데이터(*.dat).** 이 옵션은 콤마 구분 데이터와 동일하나 데이터가 전치되어 행이 필드를 나타내고 열이 레코드를 나타낸다는 점만 다릅니다.
   - **HTML (*.html).** 이 옵션은 파일에 HTML 형식의 출력을 씁니다.

셀 및 열 선택

표 노드, 교차표 노드 및 평균 노드를 포함하여 많은 노드가 표 형식의 출력을 생성합니다. 이러한 출력 표는 셀 선택, 클립보드에 표의 전부 또는 일부 복사, 현재 선택을 기반으로 하여 새 노드 생성, 표 저장 및 인쇄 등을 포함하여 유사한 방법으로 보고 조작할 수 있습니다.

셀 선택. 셀을 선택하려면 해당 셀을 클릭하십시오. 직사각형 범위의 셀을 선택하려면 원하는 범위의 한 코너를 클릭하고 마우스를 해당 범위의 다른 코너로 끈 다음 마우스 단추를 놓으십시오. 전체 열을 선택하려면 열 머리말을 클릭하십시오. 다중 열을 선택하려면 열 머리말에서 Shift-클릭 또는 Ctrl-클릭을 사용하십시오.

새로 선택하면 이전 선택이 지워집니다. 선택하는 동안 Ctrl 키를 아래로 누르고 있으면 이전 선택을 지우지 않고 새 선택을 기존 선택에 추가할 수 있습니다. 이 방법을 사용하여 연속되지 않은 다중 표 영역을 선택할 수 있습니다. 편집 메뉴에는 **모두 선택** 및 **선택 지우기** 옵션도 포함됩니다.

열 다시 정렬. 테이블 노드 및 평균 노드 출력 브라우저를 사용하면 열 머리말을 클릭하고 이를 원하는 위치에 끌어다 놓음으로써 표에서 열을 이동할 수 있습니다. 한 번에 한 열만 이동할 수 있습니다.

테이블 노드

테이블 노드는 데이터의 값을 나열하는 테이블을 작성합니다. 스트림의 모든 필드 및 모든 값이 포함되므로 읽기 쉬운 양식으로 데이터 값을 조사하거나 내보내는 데 유용합니다. 선택적으로, 특정 조건을 충족시키는 레코드를 강조표시할 수 있습니다.

참고: 작업 중인 데이터 세트가 작지 않은 경우에는 테이블 노드에 전달할 데이터의 서브세트를 선택하는 것이 좋습니다. 레코드 수가 표시 구조에 포함될 수 있는 *크기*(예: 1억 행)를 초과하면 테이블 노드는 데이터를 적절히 표시할 수 없습니다.

테이블 노드 설정 탭

레코드 강조표시 조건. 레코드 강조표시 조건을 충족시키는 CLEM 표현식을 입력하여 테이블에서 레코드를 강조표시할 수 있습니다. 이 옵션은 **화면에 출력**을 선택한 경우에만 사용됩니다.

테이블 노드 형식 탭

형식 탭에는 필드마다 형식을 지정하는 데 사용되는 옵션이 있습니다. 이 탭은 유형 노드에서도 사용됩니다. 자세한 정보는 152 페이지의 『필드 형식 설정 탭』의 내용을 참조하십시오.

출력 노드 출력 탭

표 유형의 출력을 생성하는 노드의 경우, 출력 탭을 사용하면 결과의 형식 및 위치를 지정할 수 있습니다.

출력 이름. 노드가 실행될 때 생성되는 출력의 이름을 지정합니다. **자동**은 출력을 생성하는 노드를 기반으로 이름을 선택합니다. 필요에 따라 **사용자 정의**를 선택하여 다른 이름을 지정할 수 있습니다.

화면에 출력(기본값)**.** 온라인으로 볼 출력 오브젝트를 생성합니다. 출력 오브젝트는 출력 노드가 실행될 때 관리자 창의 출력 탭에 표시됩니다.

파일로 출력. 노드가 실행될 때 출력을 파일에 저장합니다. 이 옵션을 선택하는 경우, 파일 이름을 입력하거나 디렉토리로 이동하여 파일 선택자 단추를 사용하여 파일 이름을 지정한 다음 파일 유형을 선택하십시오. 일부 파일 유형은 특정 유형의 출력에 대해 사용 불가능합니다.

데이터는 시스템 기본값 인코딩 형식의 출력이며 Windows 제어판에서 지정되며 분산 모드로 실행 중인 경우에는 서버 컴퓨터에서 지정됩니다.

- **데이터(탭 구분 데이터)(*.tab).** 이 옵션은 데이터 값을 포함하는 형식화된 텍스트 파일을 생성합니다. 이 스타일은 정보를 다른 애플리케이션으로 가져올 수 있는 일반 텍스트 표시를 생성하는 경우에 유용할 때가 많습니다. 이 옵션은 표, 교차표 및 평균 노드에 대해 사용할 수 있습니다.
- **데이터(콤마 구문 데이터)(*.dat).** 이 옵션은 데이터 값을 포함하는 콤마 구분 텍스트 파일을 생성합니다. 이 유형은 스프레드시트 또는 기타 데이터 분석 애플리케이션으로 가져올 수 있는 데이터 파일을 생성하는 빠른 방법으로 유용한 경우가 많습니다. 이 옵션은 표, 교차표 및 평균 노드에 대해 사용할 수 있습니다.
- **HTML (*.html).** 이 옵션은 파일에 HTML 형식의 출력을 씁니다. (표, 교차표 또는 평균 노드의) 표 형식의 출력인 경우, HTML 파일 세트에 HTML 표 내에 필드 이름 및 데이터를 나열하는 내용 패널이 포함됩니다. 표의 행의 수가 **페이지당 선** 지정 사항을 초과하면 표가 다중 HTML 파일로 분할될 수 있습니다. 이 경우, 내용 패널에 모든 표 페이지에 대한 링크가 포함되며 표를 탐색할 수 있는 방법이 제공됩니다. 표가 아닌 출력의 경우, 노드 결과를 포함하는 단일 HTML 파일이 생성됩니다.

  참고: HTML 출력에 첫 번째 페이지에 대한 형식화만 포함된 경우, **출력 페이지 번호 매기기**를 선택하고 모든 출력이 단일 페이지에 포함되도록 **페이지당 선** 지정 사항을 조정하십시오. 또는 보고서 노드와 같이 노드에 대한 출력 템플리트가 사용자 정의 HTML 태그를 포함하는 경우에는 형식 유형으로 **사용자 정의**를 지정해야 합니다.
- **텍스트 파일(*.txt).** 이 옵션은 출력을 포함하는 텍스트 파일을 생성합니다. 이 스타일은 워드프로세서 또는 프리젠테이션 소프트웨어 등의 다른 애플리케이션으로 가져올 수 있는 출력을 생성하는 경우에 유용할 때가 많습니다. 이 옵션은 일부 노드에는 사용할 수 없습니다.
- **출력 오브젝트(*.cou).** 이 형식으로 저장되는 출력 오브젝트는 IBM SPSS Modeler에서 열고 볼 수 있으며 프로젝트에 추가할 수 있으며 IBM SPSS Collaboration and Deployment Services Repository를 사용하여 공개하고 추적할 수 있습니다.

출력 보기. 평균 노드에 대해서 기본적으로 단순 또는 고급 출력이 표시되도록 지정할 수 있습니다. 또한 생성된 출력을 찾아볼 때 이러한 보기 사이를 토글할 수 있습니다. 자세한 정보는 331 페이지의 『평균 노드 출력 브라우저』의 내용을 참조하십시오.

형식. 보고서 노드의 경우, 출력이 자동으로 형식화되거나 템플리트에 포함된 HTML을 사용하여 형식화되도록 선택할 수 있습니다. 템플리트에서 HTML 형식화를 허용하도록 **사용자 정의**를 선택하십시오.

제목. 보고서 노드의 경우, 보고서 출력의 맨 위에 표시될 선택적 제목 텍스트를 지정할 수 있습니다.

삽입된 텍스트 강조표시. 보고서 노드의 경우, 보고서 템플리트의 CLEM 표현식을 사용하여 생성된 텍스트를 강조표시하려면 이 옵션을 선택하십시오. 자세한 정보는 333 페이지의 『보고서 노드 템플리트 탭』 주제를 참조하십시오. **사용자 정의** 형식화를 사용하는 경우에는 이 옵션을 권장하지 않습니다.

페이지당 선. 보고서 노드의 경우, 출력 보고서의 **자동** 형식화 동안 각 페이지에 포함할 선 수를 지정합니다.

데이터 전치. 이 옵션은 행이 필드를 나타내고 열이 레코드를 나타내도록 데이터를 내보내기 전에 전치합니다.

참고: 큰 표의 경우, 특히 원격 서버로 작업하는 경우 위 옵션이 비능률적일 수 있습니다. 이런 경우, 파일 출력 노드를 사용하면 성능이 훨씬 개선됩니다. 자세한 정보는 366 페이지의 『플랫 파일 내보내기 노드』의 내용을 참조하십시오.

테이블 브라우저

테이블 브라우저는 표 형식 데이터를 표시하며 여기서 셀 선택 및 복사, 열 다시 정렬, 테이블 저장 및 인쇄를 포함한 표준 조작을 수행할 수 있습니다. 자세한 정보는 307 페이지의 『셀 및 열 선택』의 내용을 참조하십시오. 이러한 조작은 노드에서 데이터를 미리볼 때 수행할 수 있는 조작과 동일합니다.

테이블 데이터 내보내기. 다음을 선택하여 테이블 브라우저에서 데이터를 내보낼 수 있습니다.

파일 > 내보내기

자세한 정보는 307 페이지의 『출력 내보내기』의 내용을 참조하십시오.

Windows 제어판에 지정되어 있거나 분산 모드에서 실행 중인 경우 서버 컴퓨터에 지정된 시스템 기본 인코딩 형식으로 데이터를 내보냅니다.

테이블 검색. 주 도구 모음의 검색 단추(쌍안경 아이콘 포함)가 검색 도구 모음을 활성화하며 이 검색 도구를 사용하여 테이블에서 특정 값을 검색할 수 있습니다. 테이블에서 앞 또는 뒤로 검색할 수 있고 대소문자 구분 검색(**Aa** 단추)을 지정할 수 있으며 검색 중단 단추로 진행 중인 검색을 중단할 수 있습니다.

새 노드 생성. 생성 메뉴에는 노드 생성 작업이 포함됩니다.

- **Select Node ("Records").** 테이블의 셀이 선택된 레코드를 선택하는 선택 노드를 생성합니다.
- **Select ("And").** 테이블에서 선택된 모든 값을 포함하는 레코드를 선택하는 선택 노드를 생성합니다.
- **Select ("Or").** 테이블에서 선택된 값 중 *임의* 값을 포함하는 레코드를 선택하는 선택 노드를 생성합니다.
- **Derive ("Records").** 새 플래그 필드를 작성하는 파생 노드를 생성합니다. 플래그 필드는 테이블의 임의 셀이 선택된 레코드의 경우 *T*를 포함하고 나머지 레코드의 경우 *F*를 포함합니다.
- **Derive ("And").** 새 플래그 필드를 작성하는 파생 노드를 생성합니다. 플래그 필드는 테이블에서 선택된 모든 값을 포함하는 레코드의 경우 *T*를 포함하고 나머지 레코드의 경우 *F*를 포함합니다.
- **Derive ("Or").** 새 플래그 필드를 작성하는 파생 노드를 생성합니다. 플래그 필드는 테이블에서 선택된 값 중 *임의* 값을 포함하는 레코드의 경우 *T*를 포함하고 나머지 레코드의 경우 *F*를 포함합니다.

교차표 노드

교차표 노드를 사용하면 필드 간 관계를 표시하는 테이블을 작성할 수 있습니다. 이는 두 범주형 필드(플래그, 명목 또는 순서) 간 관계를 표시하는 데 가장 일반적으로 사용되지만 연속형(숫자 범위) 필드 간 관계를 표시하는 데도 사용될 수 있습니다.

교차표 노드 설정 탭

설정 탭에서는 교차표의 구조에 대한 옵션을 지정할 수 있습니다.

필드. 다음 옵션에서 필드 선택 유형을 선택하십시오.

- **선택.** 이 옵션을 사용하면 행에 대한 범주형 필드 하나와 교차표의 열에 대한 범주형 필드 하나를 선택할 수 있습니다. 교차표의 행 및 열은 선택된 범주형 필드에 대한 값의 목록에 의해 정의됩니다. 교차표의 셀에는 아래에서 선택된 요약 통계가 포함되어 있습니다.
- **모든 플래그(참 값).** 이 옵션은 데이터의 각 플래그 필드에 대해 하나의 행 및 열을 가진 교차표를 요청합니다. 교차표의 셀에는 각 플래그 조합에 대한 이중 긍정의 개수가 포함되어 있습니다. 즉, *빵 구입*에 해당하는 행과 *치즈 구입*에 해당하는 열의 경우 해당 행 및 열의 교차점에 있는 셀에는 *빵 구입*과 *치즈 구입*이 모두 참인 레코드의 수가 포함되어 있습니다.
- **모든 숫자.** 이 옵션은 각 숫자 필드에 대해 하나의 행 및 열을 가진 교차표를 요청합니다. 교차표의 셀은 해당 필드 쌍에 대한 교차곱의 합계를 나타냅니다. 즉, 교차표의 각 셀에 대해 행 필드 및 열 필드의 값이 각 레코드에 대해 곱해진 후 레코드에 대해 합계가 계산됩니다.

결측값 포함. 사용자 결측(공백) 및 시스템 결측($null$) 값을 행 및 열 출력에 포함합니다. 예를 들어, *N/A* 값이 선택된 열 필드에 대해 사용자 결측으로 정의된 경우에는 다른 범주와 마찬가지로 *N/A*라는 별도의 열이 테이블에 포함됩니다(이 값이 실제로 데이터에서 발생한다고 가정함). 이 옵션이 선택 취소되면 발생 빈도에 관계없이 *N/A* 열은 제외됩니다.

참고: 결측값을 포함하는 옵션은 선택된 필드가 교차 분석표인 경우에만 적용됩니다. 공백값은 $null$에 맵핑되며 모드가 **선택됨**이고 컨텐츠가 **함수**로 설정된 경우 함수 필드에 대한 통합에서 제외되고 모드가 **모든 숫자**로 설정된 경우 모든 숫자 필드에 대한 통합에서 제외됩니다.

셀 내용. 위에서 **선택** 필드를 선택한 경우에는 교차표의 셀에서 사용할 통계를 지정할 수 있습니다. 개수 기반 통계를 선택하거나 오버레이 필드를 선택하여 행 및 열 필드의 값을 기반으로 숫자 필드의 값을 요약하십시오.

- **교차 분석표.** 셀 값은 해당 값 조합을 가진 레코드 수의 백분율 및/또는 개수입니다. 모양 탭의 옵션을 사용하여 원하는 교차 분석표 요약을 지정할 수 있습니다. 전역값 카이제곱 값도 유의성과 함께 표시됩니다. 자세한 정보는 312 페이지의 『교차표 노드 출력 브라우저』의 내용을 참조하십시오.
- **함수.** 요약 함수를 선택하는 경우 셀 값은 적절한 행 및 열 값을 가진 케이스에 대해 선택된 오버레이 필드 값의 함수입니다. 예를 들어, 행 필드가 *지역*이고 열 필드가 *제품*인 경우 오버레이 필드가 *수입*이면 *북동 지역* 행의 셀과 *위젯* 열은 복동 지역에서 판매된 위젯에 대한 수입의 합계(또는 평균, 최소값, 최대값)를 포함합니다. 기본 요약 함수는 **평균**입니다. 함수 필드를 요약하기 위해 다른 함수를 선택할 수 있습니다. 옵션으로는 **평균**, **합계**, **SDev**(표준 편차), **최대값** 및 **최소값**이 있습니다.

교차표 노드 모양 탭

모양 탭에서는 교차 분석표 교차표에 대해 제공되는 통계와 교차표에 대한 정렬 및 강조표시 옵션을 제거할 수 있습니다.

행 및 열. 교차표에서 행 및 열 표제의 정렬을 제어합니다. 기본값은 **정렬되지 않음**입니다. **오름차순** 또는 **내림차순**을 선택하여 지정된 방향으로 행 및 열 표제를 정렬하십시오.

오버레이. 교차표에서 극단값을 강조표시할 수 있게 합니다. 값은 셀 개수(교차 분석표 교차표의 경우) 또는 계산된 값(함수 교차표의 경우)을 기반으로 강조표시됩니다.

- **맨 위 강조표시.** 교차표에서 가장 높은 값을 빨간색으로 강조표시하도록 요청할 수 있습니다. 강조표시할 값 수를 지정하십시오.
- **맨 아래 강조표시.** 교차표에서 가장 낮은 값을 녹색으로 강조표시하도록 요청할 수도 있습니다. 강조표시할 값 수를 지정하십시오.

참고: 두 강조표시 옵션에 대해 동률을 사용하면 요청한 것보다 많은 값을 강조표시할 수 있습니다. 예를 들어, 셀 사이에 6개의 0(영)이 있는 교차표가 있을 때 **맨 아래 5개 강조표시**를 요청하면 6개의 0(영)이 모두 강조표시됩니다.

교차 분석표 셀 내용. 교차 분석표의 경우 교차 분석표 교차표에 대해 교차표에 포함된 요약 통계를 지정할 수 있습니다. 이 옵션은 설정 탭에서 **모든 숫자** 또는 **함수** 옵션을 선택하는 경우 사용할 수 없습니다.

- **개수.** 셀에는 해당 열 값을 가진 행 값이 포함된 레코드의 수가 포함되어 있습니다. 이는 유일한 기본 셀 내용입니다.
- **기대값.** 행과 열 사이에 관계가 없다고 가정했을 때 셀에 있는 레코드의 수에 대한 기대값입니다. 기대값은 다음 수식을 기반으로 합니다.

$$p(\textit{row value}) * p(\textit{column value}) * \textit{total number of records}$$

- **잔차.** 관측값과 기대값 사이의 차이입니다.
- **행의 백분율.** 해당 열 값을 가진 행 값이 포함된 모든 레코드의 백분율입니다. 행 내에서 백분율 합계는 100입니다.
- **열의 백분율.** 해당 행 값을 가진 열 값이 포함된 모든 레코드의 백분율입니다. 열 내에서 백분율 합계는 100입니다.
- **총계의 백분율.** 열 값과 행 값의 조합을 가진 모든 레코드의 백분율입니다. 전체 교차표에서 백분율 합계는 100입니다.
- **행 및 열 총계 포함.** 열 및 행 총계에 대한 교차표에 행 및 열을 추가합니다.
- **설정 적용.** (출력 브라우저 전용) 출력 브라우저를 닫은 후 다시 열지 않고 교차표 노드 출력의 모양을 변경할 수 있게 합니다. 출력 브라우저의 이 탭에서 변경사항을 작성하고 이 단추를 클릭한 후 교차표 탭을 선택하여 변경사항의 영향을 확인하십시오.

교차표 노드 출력 브라우저

교차표 브라우저는 교차 분석표 데이터를 표시하며 교차표에 대해 셀 선택, 교차표의 전부 또는 일부를 클립보드에 복사, 교차표 선택사항을 기반으로 새 노드 생성, 교차표 저장 및 인쇄를 포함한 조작을 수행할 수 있게 합니다. 교차표 브라우저는 Oracle의 Naive Bayes 모델 등의 특정 모델의 출력을 표시하는 데도 사용할 수 있습니다.

파일 및 편집 메뉴는 출력 인쇄, 저장 및 내보내기와 데이터 선택 및 복사를 위한 일반적인 옵션을 제공합니다. 자세한 정보는 305 페이지의 『출력 보기』의 내용을 참조하십시오.

카이제곱 두 범주형 필드의 교차 분석표에 대해 전역값 Pearson의 카이제곱도 테이블 아래에 표시됩니다. 이 검정은 관계가 존재하지 않는 경우 예상하는 개수와 관측개수 간 차이를 기반으로 두 필드가 관련되지 않을 확률을 표시합니다. 예를 들어, 고객 만족도와 상점 위치 사이에 관계가 없으면 모든 상점에 대해 비슷한 만족도를 예상합니다. 하지만 특정 상점의 고객이 다른 고객보다 높은 비율을 지속적으로 보고하는 경우에는 우연의 일치가 아니라고 의심할 수 있습니다. 차이가 클수록 우연 표본추출 오류만의 결과였을 확률이 더 낮습니다.

- 카이제곱 검정은 두 필드가 관계가 없을 확률을 표시하며 이 경우 관측빈도와 기대빈도 간 차이는 우연만의 결과입니다. 이 확률이 매우 낮은 경우(일반적으로 5% 미만) 두 필드 간 관계를 유의하다고 합니다.
- 하나의 열 또는 하나의 행만 있는 경우(일원 카이제곱 검정) 자유도는 셀의 수에서 1을 뺀 값입니다. 이원 카이제곱의 경우 자유도는 행의 수에서 1을 뺀 값에 열의 수에서 1을 뺀 값을 곱한 값입니다.
- 셀 기대빈도가 5 미만인 경우에는 카이제곱 통계량 해석 시 주의하십시오.
- 카이제곱 검정은 두 필드의 교차 분석표에만 사용할 수 있습니다. (설정 탭에서 **모든 플래그** 또는 **모든 숫자**가 선택되면 이 검정이 표시되지 않습니다.)

생성 메뉴. 생성 메뉴에는 노드 생성 작업이 포함됩니다. 이 조작은 교차 분석표 교차표에만 사용할 수 있으므로 교차표에서 하나 이상의 셀이 선택되어 있어야 합니다.

- **선택 노드.** 교차표에서 선택된 셀과 일치하는 레코드를 선택하는 선택 노드를 생성합니다.
- **파생 노드(플래그).** 새 플래그 필드를 작성하는 파생 노드를 생성합니다. 플래그 필드에는 *T*(교차표에서 선택된 셀과 일치하는 레코드의 경우) 및 *F*(나머지 레코드의 경우)가 포함되어 있습니다.
- **파생 노드(세트).** 새 명목 필드를 작성하기 위해 파생 노드를 생성합니다. 명목 필드에는 교차표에서 선택된 셀의 연속 세트 각각에 대해 하나의 범주가 포함되어 있습니다.

분석 노드

분석 노드를 통해 정확한 예측을 생성하기 위한 모델의 능력을 평가할 수 있습니다. 분석 노드는 하나 이상의 모델 너깃에 대한 예측 값과 실제 값(대상 필드)의 다양한 비교를 수행합니다. 분석 노드를 사용하여 예측 모형을 다른 예측 모형과 비교할 수도 있습니다.

분석 노드를 실행하는 경우 분석 결과의 요약이 실행된 스트림의 각 모델 너깃에 대한 요약 탭의 분석 섹션에 자동으로 추가됩니다. 자세한 분석 결과는 관리자 창의 출력 탭에 표시되거나 파일에 직접 기록될 수 있습니다.

참고: 분석 노드가 예측 값을 실제 값과 비교하므로 감독 모델(대상 필드가 필요함)에서만 유용합니다. 클러스터링 알고리즘과 같은 무감독 모델의 경우 비교의 기준으로 사용할 수 있는 실제 결과가 없습니다.

분석 노드 분석 탭

분석 탭을 사용하면 분석의 세부사항을 지정할 수 있습니다.

일치 교차표(기호 또는 범주형 대상의 경우). 생성(예측)된 각 필드와 범주형 대상의 대상 필드(플래그, 명목 또는 순서) 간 일치 패턴을 표시합니다. 실제 값으로 정의된 행과 예측 값으로 정의된 열이 있으며 각 셀에 해

당 패턴이 있는 레코드 수가 있는 테이블이 표시됩니다. 예측에서 계통 오류를 식별하는 데 유용합니다. 두 개 이상의 생성 필드가 동일한 출력 필드와 관련되어 있지만 다른 모델에서 생성한 경우 이러한 필드가 동의하고 거부하는 케이스가 계산되고 총계가 표시됩니다. 동의하는 케이스의 경우 다른 올바른/잘못된 통계 세트가 표시됩니다.

성능 평가. 범주형 출력이 있는 모델의 성능 평가 통계를 표시합니다. 출력 필드의 각 범주에 대해 보고되는 이 통계는 해당 범주에 속하는 레코드를 예측하기 위한 모델의 평균 정보 컨텐츠의 측도(비트 단위)입니다. 분류 문제의 어려움을 고려하므로 드문 범주의 정확한 예측은 공통 범주의 정확한 예측보다 높은 성능 평가 지수를 얻습니다. 모델이 범주 추측에 지나지 않는 경우 해당 범주의 성능 평가 지수는 0이 됩니다.

평가 메트릭(AUC & Gini, 2진 분류자만). 2진 분류자의 경우 이 옵션은 AUC(Area Under Curve) 및 Gini 계수 평가 메트릭을 보고합니다. 각 2진 모델에 대해 이러한 두 개의 평가 메트릭을 함께 계산합니다. 메트릭의 값은 분석 출력 브라우저에 테이블로 보고됩니다.

AUC 평가 메트릭은 ROC(Receiver Operator Characteristic) 곡선 아래의 면적으로 계산되며 분류자의 예상 성능에 대한 스칼라 표시입니다. AUC는 항상 0과 1 사이이며 높은 수는 좋은 분류자를 나타냅니다. 좌표 (0,0)과 (1,1) 사이의 대각선 ROC 곡선은 무작위 분류자를 나타내며 AUC가 0.5입니다. 따라서 실제 분류자에는 0.5 미만의 AUC가 없습니다.

Gini 계수 평가 메트릭은 AUC에 대한 대체 평가 메트릭으로 사용되는 경우가 있으며 두 개의 측도는 밀접하게 관련되어 있습니다. Gini 계수는 ROC 곡선과 대각선 간 면적의 2배로 계산되거나 Gini = 2AUC - 1로 계산됩니다. Gini 계수는 항상 0과 1 사이이며 높은 수는 좋은 분류자를 나타냅니다. Gini 계수는 가능성은 없지만 ROC 곡선이 대각선 아래에 있는 경우에 음수입니다.

신뢰도 수치(사용 가능한 경우). 신뢰도 필드를 생성하는 모델의 경우 이 옵션은 신뢰도 값의 통계와 예측에 대한 관계를 보고합니다. 이 옵션에 대해 두 개의 설정이 있습니다.

- **해당 임계값.** 정확도가 지정된 백분율이 되는 신뢰수준 하한을 보고합니다.
- **정확도 향상.** 정확도가 지정된 요인만큼 향상되는 신뢰수준 하한을 보고합니다. 예를 들어, 전체 정확도가 90%이며 이 옵션이 2.0으로 설정된 경우 보고된 값은 95% 정확도에 필요한 신뢰도입니다.

예측/예측변수 필드를 찾을 때 사용. 예측 필드가 원래의 대상 필드와 일치하는 정도를 판별합니다.

- **모델 출력 필드 메타데이터.** 예측 필드를 모델 필드 정보에 기반하여 대상과 일치시키므로 예측 필드의 이름이 변경된 경우에도 일치가 허용됩니다. 유형 노드를 사용하여 값 대화 상자에서 예측 필드의 모델 필드 정보에도 액세스할 수 있습니다. 자세한 정보는 146 페이지의 『값 대화 상자 사용』의 내용을 참조하십시오.
- **필드 이름 형식.** 이름 지정 규칙에 기반하여 필드를 일치시킵니다. 예를 들어, *response*라는 대상의 C5.0 모델 너깃에서 생성한 예측 값은 *$C-response*라는 필드에 있어야 합니다.

파티션별 구분. 파티션 필드를 사용하여 레코드를 학습, 테스트, 검증 샘플로 분할하는 경우 이 옵션을 선택하여 각 파티션에 대해 개별적으로 결과를 표시하십시오. 자세한 정보는 181 페이지의 『파티션 노드』의 내용을 참조하십시오.

참고: 파티션별로 구분하는 경우 파티션 필드에 널값이 있는 레코드가 분석에서 제외됩니다. 파티션 노드는 널 값을 생성하지 않으므로 파티션 노드가 사용되는 경우 이는 문제가 되지 않습니다.

사용자 정의 분석. 모델을 평가하는 데 사용할 사용자의 분석 계산을 지정할 수 있습니다. CLEM 표현식을 사용하여 각 레코드에 대해 계산해야 하는 값과 레코드 수준 스코어를 전체 스코어로 결합하는 방법을 지정하십시오. 함수 **@TARGET** 및 **@PREDICTED**를 사용하여 각각 대상(실제 출력) 값과 예측 값을 참조하십시오.

- **If.** 일부 조건에 따라 다른 계산을 사용해야 하는 경우 조건식을 지정하십시오.
- **Then.** If 조건이 true인 경우 계산을 지정하십시오.
- **Else.** If 조건이 false인 경우 계산을 지정하십시오.
- **Use.** 개별 스코어에서 전체 스코어를 계산하기 위한 통계를 선택하십시오.

분석을 필드별로 구분. 분석을 구분하는 데 사용 가능한 범주형 필드를 표시합니다. 전체 분석 외에 각 구분 필드의 각 범주에 대한 개별 분석이 보고됩니다.

분석 출력 브라우저

분석 출력 브라우저는 분석 노드를 실행한 결과를 표시합니다. 일반 저장, 내보내기, 인쇄 옵션은 파일 메뉴에서 사용할 수 있습니다. 자세한 정보는 305 페이지의 『출력 보기』의 내용을 참조하십시오.

분석 출력을 처음 찾아볼 때 결과가 펼쳐집니다. 결과를 본 후에 숨기려면 항목 왼쪽의 펼치기 제어를 사용하여 숨길 특정 결과를 접거나 **모두 접기** 단추를 클릭하여 모든 결과를 접으십시오. 결과를 접은 후에 다시 보려면 항목 왼쪽의 펼치기 제어를 사용하여 결과를 표시하거나 **모두 펼치기** 단추를 클릭하여 모든 결과를 표시하십시오.

출력 필드의 결과. 분석 출력에는 생성된 모델에 의해 작성되는 해당 예측 필드가 있는 각 출력 필드의 섹션이 있습니다.

비교. 출력 필드 섹션에는 해당 출력 필드와 연관된 각 예측 필드의 서브섹션이 있습니다. 범주형 출력 필드의 경우 이 섹션의 최상위 수준에는 올바른 예측과 잘못된 예측의 수 및 백분율과 스트림에 있는 총 레코드 수를 표시하는 테이블이 있습니다. 숫자 출력 필드의 경우 이 섹션은 다음 정보를 표시합니다.

- **최소 오차.** 최소 오차(관측 값과 예측 값의 차이)를 표시합니다.
- **최대 오차.** 최대 오차를 표시합니다.
- **평균 오차.** 모든 레코드에서 오차의 평균을 표시합니다. 모델에 계통 **편향**이 있는지 여부(과소평가보다 과대평가 경향이 강하거나 반대의 경우)를 표시합니다.
- **평균 절대 오차.** 모든 레코드에서 오차의 절대값 평균을 표시합니다. 방향에 관계없이 오차의 평균 크기를 표시합니다.
- **표준 편차.** 오차의 표준 편차를 표시합니다.
- **선형 상관.** 예측 값과 실제 값의 선형 상관을 표시합니다. 이 통계는 –1.0과 1.0 사이입니다. +1.0에 가까운 값은 강하게 긍정적인 연관을 표시하며 높은 예측 값이 높은 실제 값과 연관되어 있고 낮은 예측 값이 낮은 실제 값과 연관되어 있습니다. –1.0에 가까운 값은 강하게 부정적인 연관을 표시하며 높은 예측 값이

낮은 실제 값과 연관되어 있고 반대의 경우도 같습니다. 0.0에 가까운 값은 약한 연관을 표시하며 예측 값이 실제 값에 관계없이 크거나 작습니다. *참고*: 여기에서 공백 항목은 실제 또는 예측 값이 상수이므로 이 경우 선형 상관을 계산할 수 없음을 표시합니다.

- **발생.** 분석에서 사용되는 레코드 수를 표시합니다.

일치 교차표. 범주형 출력 필드의 경우 분석 옵션에서 일치 교차표를 요청하면 교차표가 포함된 서브섹션이 여기에 표시됩니다. 행은 실제 관측 값을 표시하며 열은 예측 값을 표시합니다. 테이블의 셀은 예측 값과 실제 값의 각 조합에 대한 레코드 수를 표시합니다.

성능 평가. 범주형 출력 필드의 경우 분석 옵션에서 성능 평가 통계를 요청하면 성능 평가 결과가 여기에 표시됩니다. 각 출력 범주가 해당 성능 평가 통계와 함께 나열됩니다.

신뢰도 보고서. 범주형 출력 필드의 경우 분석 옵션에서 신뢰도를 요청하면 값이 여기에 표시됩니다. 모델 신뢰도에 대해 다음 통계가 보고됩니다.

- **범위.** 스트림 데이터에 있는 레코드에 대한 신뢰도의 범위(최소값 및 최대값)를 표시합니다.
- **올바른 평균.** 올바르게 분류된 레코드의 평균 신뢰도를 표시합니다.
- **잘못된 평균.** 잘못 분류된 레코드의 평균 신뢰도를 표시합니다.
- **항상 올바른 하한.** 예측이 항상 올바른 신뢰도 임계값 하한을 표시하고 이 기준을 충족시키는 케이스의 백분율을 표시합니다.
- **항상 잘못된 상한.** 예측이 항상 잘못된 신뢰도 임계값 상한을 표시하고 이 기준을 충족시키는 경우의 백분율을 표시합니다.
- **X% 정확도 하한.** 정확도가 *X*%인 신뢰수준을 표시합니다. *X*는 분석 옵션에서 **해당 임계값**에 지정된 대략적인 값입니다. 일부 모델 및 데이터 세트의 경우 옵션에 지정된 정확한 임계값을 제공하는 신뢰도를 선택할 수 없습니다(일반적으로 임계값에 가까운 동일한 신뢰도가 있는 비슷한 케이스의 군집으로 인해). 보고된 임계값은 단일 신뢰도 임계값과 함께 얻을 수 있는 지정된 정확도 기준에 가장 가까운 값입니다.
- **올바른 X 중첩 하한.** 정확도가 전체 데이터 세트의 경우보다 *X*배 양호한 신뢰도를 표시합니다. *X*는 분석 옵션에서 **정확도 향상**에 지정된 값입니다.

상호 동의. 동일한 출력 필드를 예측하는 두 개 이상의 생성된 모델이 스트림에 포함된 경우 모델에서 생성한 예측 간 **동의**에서 통계를 확인할 수도 있습니다. 여기에는 예측이 동의하는 레코드의 수와 백분율(범주형 출력 필드의 경우) 또는 오차 요약 통계(연속형 출력 필드의 경우)가 포함됩니다. 범주형 필드의 경우 여기에는 모델이 동의하는(동일한 예측 값을 생성함) 레코드의 서브세트에 대한 실제 값과 비교한 예측의 분석이 포함됩니다.

평가 메트릭. 2진 분류자의 경우 분석 옵션에서 평가 메트릭을 요청하면 AUC 및 Gini 계수 평가 메트릭의 값이 이 섹션의 테이블에 표시됩니다. 테이블에는 각 2진 분류자 모델에 대한 하나의 행이 있습니다. 각 모델이 아닌 각 출력 필드에 대한 평가 메트릭 테이블이 표시됩니다.

데이터 검토 노드

데이터 검토 노드는 전체 크기 그래프 및 다양한 데이터 준비 노드를 생성하기 위해 정렬하고 사용할 수 있는 읽기 쉬운 교차표에 제공되는 IBM SPSS Modeler로 가져오는 데이터에 대한 포괄적인 정보를 간략하게 제공합니다.

- 감사 탭은 데이터에 대한 사전 이해를 얻는 데 유용할 수 있는 요약 통계, 히스토그램 및 분포 그래프를 제공하는 보고서를 표시합니다. 보고서는 필드 이름 앞에 저장 공간 아이콘도 표시합니다.
- 감사 보고서의 품질 탭은 이상치, 극단값 및 결측값에 대한 정보를 표시하고 이 값을 처리하는 데 필요한 도구를 제공합니다.

데이터 검토 노드 사용

데이터 검토 노드는 소스 노드에 직접 연결되거나 인스턴스화된 유형 노드로부터 다운스트림으로 연결될 수 있습니다. 결과를 기반으로 다수의 데이터 준비 노드도 생성할 수 있습니다. 예를 들어, 모델링 시 유용하도록 결측값이 지나치게 많이 포함된 필드를 제외하는 필터 노드를 생성하고 나머지 모든 필드에 대한 결측값을 대치하는 SuperNode를 생성할 수 있습니다. 여기서 감사의 실제 효과가 적용되어 데이터의 현재 상태를 평가할 수 있을 뿐만 아니라 평가를 기반으로 조치를 취할 수도 있습니다.

데이터 선별 또는 표본추출. 초기 감사는 "큰 데이터"를 처리할 때 특히 효과적이므로 표본 노드를 사용하면 레코드의 서브세트만 선택하여 초기 탐색 중 처리 시간을 줄일 수 있습니다. 데이터 검토 노드는 분석의 탐색 단계에서 필드선택 및 이상 항목 발견 등의 노드와 조합하여 사용할 수도 있습니다.

데이터 검토 노드 설정 탭

설정 탭에서는 감사에 대한 기본 매개변수를 지정할 수 있습니다.

기본값. 다음과 같이 단순히 노드를 스트림에 연결하고 **실행**을 클릭하여 기본 설정을 기반으로 모든 필드에 대한 감사 보고서를 생성할 수 있습니다.

- 유형 노드 설정이 없으면 모든 필드가 보고서에 포함됩니다.
- 유형 설정(인스턴스화되었는지 여부는 관계없음)이 있으면 모든 *입력*, *목표* 및 *둘 다* 필드가 표시에 포함됩니다. 하나의 목표 필드가 있는 경우에는 해당 필드를 오버레이 필드로 사용하십시오. 둘 이상의 목표 필드가 지정된 경우에는 기본 오버레이가 지정되지 않습니다.

사용자 정의 필드 사용. 수동으로 필드를 선택하려면 이 옵션을 선택하십시오. 오른쪽의 필드 선택기 단추를 사용하여 개별적으로 또는 유형별로 필드를 선택하십시오.

오버레이 필드. 오버레이 필드는 감사 보고서에 표시된 썸네일 그래프를 그리는 데 사용됩니다. 연속형(숫자 범위) 필드의 경우 이변량 통계(공분산 및 상관)도 계산됩니다. 유형 노드 설정을 기반으로 단일 목표 필드가 있는 경우 해당 필드는 위에 설명된 대로 기본 오버레이 필드로 사용됩니다. 또는 오버레이를 지정하기 위해 **사용자 정의 필드 사용**을 선택할 수 있습니다.

표시. 출력에서 그래프를 사용할 수 있는지 여부를 지정하고 기본적으로 표시되는 통계를 선택할 수 있게 합니다.

- **그래프.** 각각의 선택된 필드에 대한 그래프를 표시합니다(데이터에 적합한 대로 분산(막대형) 그래프, 히스토그램 또는 산점도). 그래프는 초기 보고서에서 썸네일로 표시되지만 전체 크기 그래프 및 그래프 노드도 생성될 수 있습니다. 자세한 정보는 319 페이지의 『데이터 검토 출력 브라우저』의 내용을 참조하십시오.
- **기본/고급 통계.** 기본적으로 출력에 표시되는 통계의 수준을 지정합니다. 이 설정이 초기 표시를 결정하는 동안 이 설정과 관계없이 출력에서 모든 통계를 사용할 수 있습니다. 자세한 정보는 320 페이지의 『통계 표시』의 내용을 참조하십시오.

중앙값 및 모드. 보고서에서 모든 필드에 대한 중앙값 및 모드를 계산합니다. 큰 데이터 세트를 사용하는 경우 이 통계는 다른 통계보다 계산하는 데 시간이 오래 걸리므로 처리 시간이 늘어날 수 있습니다. 중앙값만 계산하는 경우 보고된 값은 일부 경우 전체 데이터 세트 대신 2000개의 레코드를 가진 표본을 기반으로 할 수 있습니다. 이 표본추출은 이를 수행하지 않으면 메모리 제한이 초과되는 경우에 필드별로 수행됩니다. 표본추출이 적용되는 경우 이에 따라 출력에서 결과에 레이블이 지정됩니다(단순히 중앙값보다는 표본 중앙값이 지정됨). 중앙값 이외의 모든 통계는 항상 전체 데이터 세트를 사용하여 계산됩니다.

비어 있거나 유형 없는 필드. 인스턴스화된 데이터와 함께 사용되는 경우 유형 없는 필드는 감사 보고서에 포함되지 않습니다. 유형 없는 필드(비어 있는 필드 포함)를 포함하려면 업스트림 유형 노드에서 **모든 값 지우기**를 선택하십시오. 그러면 데이터가 인스턴스화되지 않아 모든 필드가 보고서에 포함됩니다. 예를 들어, 모든 필드의 전체 목록을 얻으려고 하거나 비어 있는 필드를 제외할 필터 노드를 생성하려는 경우 이것이 유용할 수 있습니다. 자세한 정보는 323 페이지의 『결측 데이터로 필드 필터링』의 내용을 참조하십시오.

데이터 검토 품질 탭

데이터 검토 노드의 품질 탭은 결측값, 이상치 및 극단값을 처리하기 위한 옵션을 제공합니다.

결측값

- **유효한 값을 가진 레코드 수.** 각각의 평가된 필드에 대해 유효한 값을 가진 레코드의 수를 표시하려면 이 옵션을 선택하십시오. 널(정의되지 않음) 값, 공백값, 공백 및 빈 문자열은 항상 유효하지 않은 값으로 처리됩니다.
- **유효하지 않은 값을 가진 레코드 수 분석.** 각 필드에 대해 각 유형의 유효하지 않은 값을 가진 레코드의 수를 표시하려면 이 옵션을 선택하십시오.

이상치 및 극단값

이상치 및 극단값에 대한 발견 방법. 두 가지 방법이 지원됩니다.

평균으로부터의 표준 편차. 평균으로부터의 표준 편차 수를 기반으로 이상치 및 극단값을 발견합니다. 예를 들어, 평균이 100이고 표준 편차가 10인 필드가 있는 경우 3.0을 지정하여 70 미만 또는 130 이상의 값은 이상치로 처리됨을 나타낼 수 있습니다.

사분위수 범위. 두 개의 중심 사분위수가 속하는 범위(25번째 백분위수와 75번째 백분위수 사이)인 사분위수 범위를 기반으로 이상치 및 극단값을 발견합니다. 예를 들어, 기본 설정인 1.5를 기반으로 하면 이상치의 하한 임계값은 Q1 – 1.5 * IQR이고 상한 임계값은 Q3 + 1.5*IQR입니다. 이 옵션을 사용하면 큰 데이터 세트에서 성능이 저하될 수 있습니다.

데이터 검토 출력 브라우저

데이터 검토 브라우저는 데이터에 대한 개요를 얻기 위한 강력한 도구입니다. 감사 탭에는 모든 필드에 대한 썸네일 그래프, 저장 공간 아이콘 및 통계가 표시되지만 품질 탭에는 이상치, 극단값 및 결측값에 대한 정보가 표시됩니다. 초기 그래프 및 요약 통계를 기반으로 숫자 필드의 코딩을 변경하거나 새 필드를 파생시키거나 명목 필드의 값을 재분류하도록 결정할 수 있습니다. 또는 더 정교한 시각화를 사용하여 추가로 탐색할 수 있습니다. 데이터를 변환하거나 시각화하는 데 사용할 수 있는 임의의 수의 노드를 작성하기 위해 생성 메뉴를 사용하여 감사 보고서 브라우저에서 이를 수행할 수 있습니다.

- 열 헤더를 클릭하여 열을 정렬하거나 끌어서 놓기를 사용하여 열을 다시 정렬하십시오. 대부분의 표준 출력 조작도 지원됩니다. 자세한 정보는 305 페이지의 『출력 보기』 주제를 참조하십시오.
- 측정 또는 고유 열에서 필드를 두 번 클릭하여 필드에 대한 값 및 범위를 보십시오.
- 도구 모음 또는 편집 메뉴를 사용하여 값 레이블을 표시하거나 숨기거나 표시할 통계를 선택할 수 있습니다. 자세한 정보는 320 페이지의 『통계 표시』의 내용을 참조하십시오.
- 필드 이름 왼쪽의 저장 공간 아이콘을 확인하십시오. 저장 공간은 데이터를 필드에 저장하는 방식을 설명합니다. 예를 들어, 값이 1 및 0인 필드는 정수 데이터를 저장합니다. 이는 데이터 사용에 대해 설명하고 저장 공간에 영향을 미치지 않는 측정 수준과 구별됩니다. 자세한 정보는 9 페이지의 『필드 저장 공간 및 형식화 설정』의 내용을 참조하십시오.

그래프 보기 및 생성

오버레이가 선택되지 않은 경우 감사 탭에는 막대형 차트(명목 또는 플래그 필드용) 또는 히스토그램(연속형 필드)이 표시됩니다.

명목 또는 플래그 필드 오버레이의 경우 그래프는 오버레이의 값을 기준으로 색상이 지정됩니다.

연속형 필드 오버레이의 경우에는 1차원 막대 또는 히스토그램이 아니라 2차원 산점도가 생성됩니다. 이 경우 *x*축은 오버레이 필드에 맵핑되어 테이블을 아래로 읽을 때 모든 *x*축에서 동일한 척도를 볼 수 있습니다.

- 플래그 또는 명목 필드의 경우 막대 위에 마우스 커서를 두면 도구 팁에 기본 값 또는 레이블이 표시됩니다.
- 플래그 또는 명목 필드의 경우 도구 모음을 사용하여 썸네일 그래프의 방향을 가로에서 세로로 토글하십시오.
- 썸네일로부터 전체 *크기* 그래프를 생성하려면 썸네일을 두 번 클릭하거나 썸네일을 선택한 후 생성 메뉴에서 **그래프 출력**을 선택하십시오. *참고:* 썸네일 그래프가 표본 추출된 데이터를 기반으로 한 경우 생성되는 그래프는 원래 데이터 스트림이 계속 열려 있으면 모든 케이스를 포함합니다.

  출력을 작성한 데이터 검토 노드가 스트림에 연결되어 있는 경우에만 그래프를 생성할 수 있습니다.
- 일치하는 그래프 노드를 생성하려면 감사 탭에서 하나 이상의 필드를 선택한 후 생성 메뉴에서 **그래프 노드**를 선택하십시오. 결과 노드가 스트림 캔버스에 추가되며 스트림이 실행될 때마다 그래프를 다시 작성하는 데 사용될 수 있습니다.
- 오버레이 세트에 100개를 초과하는 값이 포함되어 있으면 경고가 발생하고 해당 오버레이는 포함되지 않습니다.

통계 표시

통계 표시 대화 상자에서는 감사 탭에 표시되는 통계를 선택할 수 있습니다. 초기 설정은 데이터 검토 노드에서 지정됩니다. 자세한 정보는 317 페이지의 『데이터 검토 노드 설정 탭』의 내용을 참조하십시오.

최소값(Minimum). 숫자변수의 가장 작은 값입니다.

최대값(Maximum). 숫자변수의 가장 큰 값입니다.

합계(Sum). 비결측값을 갖는 전체 케이스 값의 총계입니다.

범위(Range). 숫자변수의 가장 큰 값과 가장 작은 값의 차이로 최대값에서 최소값을 뺀 값을 의미합니다.

평균(Mean). 중심 경향에 대한 측도입니다. 합계를 케이스 수로 나눈 산술 평균 값입니다.

평균의 표준 오차(Standard Error of Mean). 동일 분포로부터 선택한 표본 간에 발생할 수 있는 평균값의 차이에 대한 측도입니다. 이 값을 사용하여 관측 평균과 가설 값을 간략하게 비교할 수 있습니다. 즉, 표준 오차에 대한 차이 비율이 ?2보다 작거나 +2보다 큰 경우 두 값이 다르다고 판단할 수 있습니다.

표준 편차(standard deviation). 평균 주위의 산포 측도이며 분산의 제곱근과 같습니다. 표준 편차는 원래 변수와 같은 단위로 측정됩니다.

분산(Variance). 평균에 대한 산포 측도로, 평균으로부터의 제곱합 편차를 케이스 수에서 1을 뺀 값으로 나눈 값과 같습니다. 분산은 변수 자체의 제곱 단위로 측정됩니다.

왜도(Skewness). 분포의 비대칭성에 대한 측도입니다. 정규 분포는 대칭이므로 왜도 값이 0입니다. 양의 왜도가 많은 분포는 오른쪽이 깁니다. 유의한 음의 왜도를 가지는 분포에는 왼쪽으로 긴 꼬리가 나타납니다. 왜도 값이 표준 오차의 두 배를 넘는 것은 대칭에서 벗어난 정도를 나타냅니다.

왜도의 표준 오차(Standard Error of Skewness). 표준 오차에 대한 왜도의 비율을 정규성 검정에 사용할 수 있습니다. 즉, 비율이 -2보다 작거나 +2보다 큰 경우 정규성을 거부할 수 있습니다. 왜도가 큰 양의 값인 경우 오른쪽이 길어지고 큰 음의 값인 경우 왼쪽이 길어집니다.

첨도(Kurtosis). 관측값이 중심에 군집하는 정도에 대한 측도입니다. 정규 분포의 경우 첨도 통계 값은 0입니다. 양의 첨도는 정규 분포에 비해 관측값이 분포 중심에 더 많이 군집되어 있고 분포 극단값까지의 꼬리가 더 얇다는 의미입니다. 즉, 정규 분포에 비해 급첨 분포의 꼬리가 더 두껍습니다. 음의 첨도는 정규 분포에 비해 관측값이 분포 중심에 덜 군집되어 있고 분포 극단값까지의 꼬리가 더 두껍다는 의미입니다. 즉, 정규 분포에 비해 평첨 분포의 꼬리가 더 얇습니다.

첨도의 표준 오차(Standard Error of Kurtosis). 표준 오차에 대한 첨도의 비율을 정규성 검정에 사용할 수 있습니다. 즉, 비율이 -2보다 작거나 +2보다 큰 경우 정규성을 거부할 수 있습니다. 첨도가 높은 양의 값인 경우 분포의 양끝이 정규 분포의 양끝보다 길어지고 음의 값인 경우 양끝이 짧아집니다(상자 형태 균일 분포와 유사).

고유(Unique). 모든 효과를 동시에 평가하고 유형에 관계없이 다른 모든 효과에 대해 각 효과를 조정합니다.

유효함(Valid). 시스템 결측값 또는 사용자 결측값이 지정되어 있지 않은 케이스가 유효 케이스입니다.널(정의되지 않은) 값, 공백값, 공백 및 빈 문자열은 항상 유효하지 않은 값으로 처리됩니다.

중앙값(Median). 전체 케이스의 절반이 위 아래에 해당되는 값으로 제50 백분위수입니다. 케이스 수가 짝수인 경우 중앙값은 케이스를 오름차순이나 내림차순으로 정렬했을 때 중간에 있는 두 개의 케이스의 평균입니다. 중앙값은 평균과 달리 중심을 벗어난 값에는 영향을 받지 않는 중심 경향 측도이며, 상한 극단값 또는 하한 극단값에 따라 달라질 수 있습니다.

최빈값(Mode). 가장 자주 발생하는 값입니다. 여러 값에서 최대 발생 빈도를 공유하는 경우 각각을 최빈값이라고 합니다.

중앙값 및 모드는 성능 향상을 위해 기본적으로 표시되지 않지만 데이터 검토 노드의 설정 탭에서 선택할 수 있습니다. 자세한 정보는 317 페이지의 『데이터 검토 노드 설정 탭』의 내용을 참조하십시오.

오버레이에 대한 통계

연속형(숫자 범위) 오버레이 필드가 사용 중인 경우에는 다음과 같은 통계도 사용할 수 있습니다.

공분산(Covariance). 두 변수 간 연관을 표준화하지 않은 측도로서, N-1로 나눈 교차곱 편차와 같습니다.

데이터 검토 브라우저 품질 탭

데이터 검토 브라우저의 품질 탭에는 데이터 품질 분석의 결과가 표시되며 사용자가 이상값, 극단값 및 결측값에 대한 처리를 지정할 수 있습니다.

결측값 대치: 감사 보고서에는 유효한 값, 널 값 및 공백 값의 수와 함께 각 필드에 대한 완료 레코드의 퍼센트가 나열됩니다. 특정 필드에 대한 결측값 대체를 선택하여 이러한 변환을 적용할 수퍼노드를 생성할 수 있습니다.

1. **결측값 대치** 열에서 대치할 값의 유형을 지정하십시오. 단, 있는 경우에 한합니다. 공백 또는 널 또는 둘 다 대치하도록 선택하거나 대치할 값을 선택하는 사용자 정의 조건 또는 표현식을 지정할 수 있습니다.

   IBM SPSS Modeler에 의해 인지되는 결측값에는 몇 가지 유형이 있습니다.

   - **널 또는 시스템 결측값.** 이들은 데이터베이스나 소스 파일에 공백으로 남겨졌고 소스 또는 유형 노드에서 "결측"으로 정의되지 않은 문자열이 아닌 값입니다. 시스템 결측값은 **$null$**로 표시됩니다. 빈 문자열은 특정 데이터베이스에 의해 널로 처리되더라도 IBM SPSS Modeler에서는 널로 간주되지 않음을 유의하십시오.
   - **빈 문자열 및 공백.** 빈 문자열 값과 공백(눈에 보이는 문자가 없는 문자열)은 널값과는 별개로 처리됩니다. 빈 문자열은 대부분의 경우에서 공백과 동일하게 처리됩니다. 예를 들어, 소스나 유형 노드에서 공백을 공란으로 처리하는 옵션을 선택한 경우 이 설정은 빈 문자열에도 적용됩니다.
   - **공백 또는 사용자 정의 결측값.** 이들은 소스 노드 또는 유형 노드에서 결측으로 명백하게 정의되어 있는 `unknown`, `99` 또는 `–1` 등과 같은 값입니다. 또는 널과 공백을 공란으로 처리하기로 선택할 수도 있는데 그러면 이들은 특수 처리용으로 플래그가 지정되고 대부분의 계산에서 제외됩니다. 예를 들어, `@BLANK` 함수를 사용하여 이들 값 및 다른 유형의 결측값을 공란으로 처리할 수 있습니다.

2. **방법** 열에서 사용할 방법을 지정하십시오.

   결측값을 대치하기 위해 다음 방법을 사용할 수 있습니다.

   고정됨. 고정된 값을 대체합니다(필드 평균, 범위의 중심점 또는 사용자가 지정하는 상수).

   변량. 보통 또는 균일 분포를 기반으로 변량 값을 대체합니다.

   표현식. 사용자 정의 표현식을 지정할 수 있습니다. 예를 들어, 전역값 설정 노드에 의해 생성된 글로벌 변수로 값을 대체할 수 있습니다.

   알고리즘. C&RT 알고리즘을 기반으로 모델에 의해 예측된 값을 대체합니다. 이 방법을 사용하여 대체된 각 필드의 경우, 공백과 널을 모델에 의해 예측된 값으로 대체하는 채움 노드와 함께 별도의 C&RT 모델이 있습니다. 그러면 필터 노드가 모델에 의해 생성된 예측 필드를 제거하는 데 사용합니다.
3. 결측값 수퍼노드를 생성하려면 메뉴에서 다음을 선택하십시오.

   생성 > 결측값 수퍼노드

   결측값 수퍼노드 대화 상자가 표시됩니다.
4. **모든 필드** 또는 **선택된 필드만**을 선택하고 필요에 따라 표본 *크기*를 지정하십시오. (지정된 표본은 퍼센트이며 기본적으로 모든 레코드의 10%가 표본화됩니다.)
5. 생성된 수퍼노드를 스트림 캔버스에 추가하려면 **확인**을 클릭하십시오.
6. 수퍼노드를 스트림에 첨부하여 변환을 적용하십시오.

수퍼노드 내에서 모델 너깃, 채움 및 필터 노드의 조합이 적절히 사용됩니다. 수퍼노드를 편집하고 **확대**를 클릭하여 수퍼노드 내의 특정 노드를 추가, 편집 또는 제거하여 작동을 세분화함으로써 작동 방법을 이해할 수 있습니다.

이상값 및 극단값 처리: 감사 보고서에는 이상값 수가 나열되고 데이터 검토 노드에서 지정된 발견 옵션을 기반으로 하여 각 필드에 대한 극단값이 나열됩니다. 자세한 정보는 318 페이지의 『데이터 검토 품질 탭』 주제를 참조하십시오. 특정 필드에 대해 이러한 값 강제 변환, 삭제 또는 무효화를 필요에 따라 선택한 다음 변환을 적용할 수퍼노드를 생성할 수 있습니다.

1. **동작** 열에서 특정 필드에 대한 이상값 및 극단값 처리를 지정하십시오.

   이상값 및 극단값에 대해 사용 가능한 동작은 다음과 같습니다.

   - **강제 변환.** 이상값 및 극단값을 극단값으로 간주되지 않는 가장 가까운 값으로 대체합니다. 예를 들어, 이상값이 세 표준편차 위 또는 아래의 값으로 정의된 경우, 모든 이상값이 해당 범위 내의 최대값 또는 최저값으로 대체됩니다.
   - **삭제.** 지정된 필드에 대한 이상값 및 극단값을 삭제합니다.
   - **무효화.** 널이거나 시스템 결측값인 이상값 및 극단값을 바꿉니다.
   - **이상값 강제 변환/극단값 삭제.** 극단값만 삭제합니다.
   - **이상값 강제 변환/극단값 무효화.** 극단값만 무효화합니다.

2. 수퍼노드를 생성하려면 메뉴에서 다음을 선택하십시오.

   생성 > 이상값 및 극단값 수퍼노드

   이상값 수퍼노드 대화 상자가 표시됩니다.
3. **모든 필드** 또는 **선택된 필드만**을 선택하고 **확인**을 클릭하여 생성된 수퍼노드를 스트림 캔버스에 추가하십시오.
4. 수퍼노드를 스트림에 첨부하여 변환을 적용하십시오.

선택적으로 수퍼노드를 편집하고 확대하여 찾아보거나 변경할 수 있습니다. 수퍼노드 내에서 필요에 따라 일련의 선택 및/또는 채움 노드를 사용하여 값이 삭제, 강제 변환 또는 무효화됩니다.

결측 데이터로 필드 필터링: 데이터 검토 브라우저를 통해, 품질에서 필터 생성 대화 상자를 사용하여 품질 분석의 결과를 기반으로 하여 새 필터 노드를 생성할 수 있습니다.

모드. 지정된 필드에 대해 원하는 작업, 즉, **포함** 또는 **제외**를 선택하십시오.

- **선택된 필드.** 필터 노드가 품질 탭에서 선택된 필드를 포함/제외합니다. 예를 들어, **% 완료** 열에서 테이블을 정렬할 수 있으며 Shift-클릭을 선택하여 가장 적게 완료된 필드를 선택한 다음 해당 필드를 제외하는 필터 노드를 생성할 수 있습니다.
- **다음보다 큰 품질 퍼센트 필드.** 필터 노드가 레코드 완료 퍼센트가 지정된 임계값보다 큰 필드를 포함/제외합니다. 기본 임계값은 50%입니다.

비어 있거나 유형이 없는 필드 필터링

데이터 값이 인스턴스화된 후에 유형이 없거나 비어 있는 필드가 감사 결과 및 IBM SPSS Modeler의 대부분의 기타 출력에서 제외됩니다. 이러한 필드는 모델링 목적으로는 무시되나 데이터를 과장하거나 산만하게 만들 수 있습니다. 그런 경우, 데이터 검토 브라우저를 사용하여 이러한 필드를 스트림에서 제거하는 필터 노드를 생성할 수 있습니다.

1. 비어 있거나 유형이 없는 필드를 포함하여 모든 필드가 감사에 포함되도록 하려면 업스트림 소스 또는 유형 노드에서 **모든 값 지우기**를 클릭하거나 모든 필드에 대해 *<Pass>*로 값을 설정하십시오.
2. 데이터 검토 브라우저에서 **% 완료** 열을 정렬하고 0 개의 유효한 값이 있는 필드 또는 기타 임계값이 적용된 필드를 선택하고 생성 메뉴를 사용하여 스트림에 추가될 수 있는 필터를 생성하십시오.

결측 데이터가 있는 레코드 선택: 데이터 검토 브라우저를 통해, 품질 분석의 결과를 기반으로 하여 새 선택 노드를 생성할 수 있습니다.

1. 데이터 검토 브라우저에서 품질 탭을 선택하십시오.
2. 메뉴에서 다음을 선택하십시오.

   생성 > 결측값 선택 노드

   선택 노드 생성 대화 상자가 표시됩니다.

레코드의 선택 조건. 레코드가 **유효** 또는 **유효하지 않음**일 때 유지하는지 여부를 지정합니다.

유효하지 않은 값 검색. 유효하지 않은 값 검사 여부를 지정합니다.

- **모든 필드.** 선택 노드가 모든 필드의 유효하지 않은 값을 검사합니다.
- **테이블에서 선택된 필드.** 선택 노드가 품질 출력 테이블에서 현재 선택된 필드만 검사합니다.
- **다음보다 큰 품질 퍼센트 필드.** 선택 노드가 레코드 완료 퍼센트가 지정된 임계값보다 큰 필드를 검사합니다. 기본 임계값은 50%입니다.

유효하지 않은 값이 다음 위치에서 발견되는 경우 레코드를 유효하지 않은 것으로 간주. 레코드를 유효하지 않은 것으로 식별하는 조건을 지정합니다.

- **위 필드 중 임의의 필드.** 지정된 위 필드 중 *임의의* 필드에 해당 레코드에 대해 유효하지 않은 값이 포함되면 선택 노드가 레코드를 유효하지 않은 것으로 간주합니다.
- **위 필드 중 모든 필드.** 지정된 위 필드 중 모든 필드에 해당 레코드에 대해 유효하지 않은 값이 포함되면 선택 노드가 레코드를 유효하지 않은 것으로 간주합니다.

데이터 준비를 위해 기타 노드 생성

데이터 준비에서 사용되는 다양한 노드를 데이터 검토 브라우저에서 직접 생성할 수 있습니다(재분류, 구간화 및 파생 노드 포함). 예:

- 감사 보고서에서 *claimvalue* 값과 *farmincome* 값을 모두 선택한 후 생성 메뉴에서 **파생**을 선택하여 이 두 값을 기반으로 새 필드를 파생시킬 수 있습니다. 새 노드는 스트림 캔버스에 추가됩니다.
- 마찬가지로 감사 결과에 따라 *farmincome*을 백분위수 기반 구간으로 코딩을 변경하면 더 집중된 분석이 제공되는지 판별할 수 있습니다. 구간화 노드를 생성하려면 표시에서 필드 행을 선택하고 생성 메뉴에서 **구간화**를 선택하십시오.

노드가 생성되어 스트림 캔버스에 추가된 후에는 해당 노드를 스트림에 연결하고 해당 노드를 열어서 선택된 필드에 대한 옵션을 지정해야 합니다.

변환 노드

회귀분석, 로지스틱 회귀분석 및 판별 분석과 같은 일반 스코어링 기술을 사용하기 전에 수행해야 하는 중요한 단계 중 하나는 입력 필드를 정규화하는 것입니다. 이러한 기술은 다수의 원시 데이터 파일에 적용되지 않을 수 있는 데이터의 정규 분포에 대한 가정을 수행합니다. 실세계 데이터를 처리하는 한 가지 방법은 원시 데이터 요소를 정규성이 더 높은 정규 분포 쪽으로 이동시키는 변환을 적용하는 것입니다. 또한 정규화된 필드는 서로 쉽게 비교할 수 있습니다. 예를 들어, 원시 데이터 파일에서 수입과 연령은 척도가 완전히 다르지만 정규화하는 경우 각각의 상대적 영향력을 쉽게 해석할 수 있습니다.

변환 노드에서는 사용할 가장 좋은 변환을 시각적으로 신속하게 평가할 수 있는 출력 뷰어를 제공합니다. 변수가 정상적으로 분포되는지 한 눈에 알 수 있고 필요한 경우 원하는 변환을 선택하여 적용할 수 있습니다. 여러 필드를 선택하여 필드당 하나의 변환을 수행할 수 있습니다.

필드에 대해 선호하는 변환을 선택한 후에는 변환을 수행하는 파생 또는 채움 노드를 생성하여 이들을 스트림에 연결할 수 있습니다. 파생 노드는 새 필드를 작성하고 채움 노드는 기존 필드를 변환합니다. 자세한 정보는 327 페이지의 『그래프 생성』의 내용을 참조하십시오.

변환 노드 필드 탭

필드 탭에서는 가능한 변환을 보고 적용하는 데 사용할 데이터 필드를 지정할 수 있습니다. 숫자 필드만 변환할 수 있습니다. 필드 선택기 단추를 클릭하고 표시된 목록에서 하나 이상의 숫자 필드를 선택하십시오.

변환 노드 옵션 탭

옵션 탭에서는 포함시키려는 변환의 유형을 지정할 수 있습니다. 사용 가능한 모든 변환을 포함시키거나 변환을 개별적으로 선택할 수 있습니다.

후자의 경우, 역변환 및 로그 변환을 위해 데이터를 오프셋하기 위한 숫자를 입력할 수도 있습니다. 이는 데이터에 있는 다수의 0으로 인해 평균 및 표준 편차 결과가 편향되는 경우에 유용합니다.

예를 들어, 몇 개의 0 값이 있는 *BALANCE* 필드가 있고 이 필드에서 역변환을 사용하려 합니다. 원하지 않는 편향을 피하기 위해 **역(1/x)**을 선택하고 **데이터 오프셋 사용** 필드에 1을 입력합니다. (이 오프셋은 IBM SPSS Modeler에서 `@OFFSET` 시퀀스 함수에 의해 수행된 오프셋과 관련이 없습니다.)

모든 공식. 사용 가능한 모든 변환이 계산되고 출력에 표시되어야 함을 나타냅니다.

공식 선택. 계산하고 출력에 표시할 다양한 변환을 선택할 수 있습니다.

- **역(1/x).** 역변환이 출력에 표시되어야 함을 나타냅니다.
- **로그(로그 n).** 로그$_n$ 변환이 출력에 표시되어야 함을 나타냅니다.
- **로그(로그 10).** 로그$_{10}$ 변환이 출력에 표시되어야 함을 나타냅니다.
- **지수.** 지수 변환(e^x)이 출력에 표시되어야 함을 나타냅니다.
- **제곱근.** 제곱근 변환이 출력에 표시되어야 함을 나타냅니다.

변환 노드 출력 탭

출력 탭을 사용하여 출력의 형식 및 위치를 지정할 수 있습니다. 결과를 화면에 표시하거나 표준 파일 유형 중 하나로 보낼 수 있습니다. 자세한 정보는 308 페이지의 『출력 노드 출력 탭』의 내용을 참조하십시오.

변환 노드 출력 뷰어

출력 뷰어를 사용하여 변환 노드의 실행 결과를 볼 수 있습니다. 뷰어는 변환의 썸네일 보기에 필드당 여러 변환을 표시하는 강력한 도구이므로 뷰어를 통해 필드를 신속하게 비교할 수 있습니다. 해당 파일 메뉴의 옵션을 사용하여 출력을 저장, 내보내기 또는 인쇄할 수 있습니다. 자세한 정보는 305 페이지의 『출력 보기』 주제를 참조하십시오.

변환마다(선택된 변환 제외) 아래에 다음 형식의 범례가 표시됩니다.

```
Mean (Standard deviation)
```

변환을 위한 노드 생성

출력 뷰어는 데이터 준비에 유용한 시작점을 제공합니다. 예를 들어, 정규 분포를 가정하는 스코어링 기술(예: 로지스틱 회귀분석 또는 판별 분석)을 사용할 수 있도록 *AGE* 필드를 정규화하려 할 수 있습니다. 초기 그래프

와 요약 통계를 기반으로, 특정 분포(예: 로그)에 따라 *AGE* 필드를 변환하기로 할 수 있습니다. 선호하는 분포를 선택한 후에는 스코어링에 사용할 표준화된 변환이 있는 파생 노드를 생성할 수 있습니다.

출력 뷰어에서 다음 필드 작업 노드를 생성할 수 있습니다.

- 파생
- 채움

파생 노드는 원하는 변환으로 새 필드를 작성하는 반면, 채움 노드는 기존 필드를 변환합니다. 노드는 수퍼노드 양식으로 캔버스에 배치됩니다.

서로 다른 필드에 대해 동일한 변환을 선택하는 경우, 파생 또는 채움 노드는 해당 변환이 적용되는 모든 필드에 대한 해당 변환 유형의 공식을 포함합니다. 예를 들어, 파생 노드를 생성하기 위해 다음 테이블에 표시된 필드 및 변환을 선택했다고 가정하십시오.

표 39. 파생 노드 생성 예.

필드	변환
AGE	현재 분포
INCOME	로그
OPEN_BAL	역
BALANCE	역

수퍼노드에는 다음 노드가 포함됩니다.

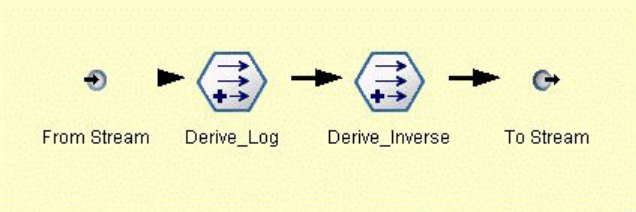


그림 64. 캔버스의 수퍼노드

이 예에서, Derive_Log 노드에는 *INCOME* 필드에 대한 로그 공식이 있고 Derive_Inverse 노드에는 *OPEN_BAL* 및 *BALANCE* 필드에 대한 역 공식이 있습니다.

노드를 생성하려면 다음을 수행하십시오.

1. 출력 뷰어의 필드마다 원하는 변환을 선택하십시오.
2. 생성 메뉴에서 **파생 노드** 또는 **채움 노드**를 선택하십시오.

   그러면 파생 노드 생성 또는 채움 노드 생성 대화 상자가 표시됩니다.

   비표준화 변환 또는 **표준화 변환(z-스코어)**을 선택하십시오. 두 번째 옵션은 변환에 *z* 스코어를 적용합니다. *z* 스코어는 값을 표준 편차에서 변수 평균으로부터의 거리 함수로 표현합니다. 예를 들어, *AGE* 필드에 로그 변환을 적용하고 표준 변환을 선택하는 경우, 생성되는 노드에 대한 최종 등식은 다음과 같습니다.

```
(log(AGE)-Mean)/SD
```

노드가 생성되고 스트림 캔버스에 표시되면 다음을 수행하십시오.

1. 노드를 스트림에 연결하십시오.
2. 수퍼노드의 경우, 노드를 두 번 클릭하여 해당 컨텐츠를 보십시오(선택사항).
3. 파생 또는 채움 노드를 두 번 클릭하여 선택된 필드의 옵션을 수정하십시오(선택사항).

그래프 생성: 출력 뷰어에서 썸네일 히스토그램으로부터 전체 *크기* 히스토그램 출력을 생성할 수 있습니다.

그래프를 생성하려면 다음을 수행하십시오.

1. 출력 뷰어에서 썸네일 그래프를 두 번 클릭하십시오.

   또는

   출력 뷰어에서 썸네일 그래프를 선택하십시오.
2. 생성 메뉴에서 **그래프 출력**을 선택하십시오.

그러면 정규 분포 곡선이 오버레이된 히스토그램이 표시됩니다. 이를 통해 사용 가능한 변환이 정규 분포와 얼마나 일치하는지 비교할 수 있습니다.

참고: 출력을 작성한 변환 노드가 스트림에 연결되어 있는 경우에만 그래프를 생성할 수 있습니다.

기타 조작: 출력 뷰어에서 다음 조작도 수행할 수 있습니다.

- 필드 열을 기준으로 출력 눈금을 정렬합니다.
- 출력을 HTML 파일로 내보냅니다. 자세한 정보는 307 페이지의 『출력 내보내기』의 내용을 참조하십시오.

통계량 노드

통계량 노드는 수치 필드에 관한 기본 요약 정보를 제공합니다. 개별 필드에 대한 요약 통계량 및 필드 사이의 상관계수를 얻을 수 있습니다.

통계량 노드 설정 탭

탐색적 데이터분석. 개별 요약 통계량이 필요한 필드를 선택하십시오. 다중 필드를 선택할 수 있습니다.

통계량. 보고할 통계량을 선택하십시오. 사용 가능한 옵션은 **개수**, **평균**, **합계**, **최소**, **최대**, **범위**, **분산**, **표준편차**, **평균의 표준오차**, **중앙값** 및 **최빈값**입니다.

상관분석. 상관분석할 필드를 선택하십시오. 다중 필드를 선택할 수 있습니다. 상관관계 필드가 선택되면 각 탐색적 데이터분석 필드 및 상관관계 필드 사이의 상관관계가 출력에 나열됩니다.

상관관계 설정. 출력에 상관관계의 강도를 표시하기 위한 옵션을 지정할 수 있습니다.

상관관계 설정

IBM SPSS Modeler는 중요한 관계를 강조표시하는 것을 돕기 위해 기술통계 레이블을 사용하여 상관관계 특성을 지정할 수 있습니다. **상관관계**는 두 연속형(수치 범위) 필드 사이의 관계의 강도를 측정합니다. –1.0에서 1.0 사이의 값을 사용합니다. +1.0에 가까운 값은 강력한 양의 연관을 표시하므로 한 필드의 높은 값은 다른 필드의 높은 값과 연관되며 낮은 값은 낮은 값과 연관됩니다. -1.0에 가까운 값은 강력한 음의 연관을 표시하므로 한 필드의 높은 값은 다른 필드의 낮은 값과 연관되며 반대의 경우도 마찬가지입니다. 0.0에 가까운 값은 약한 연관을 표시하므로 두 필드의 값이 다소 독립적입니다.

상관관계 설정 대화 상자를 사용하여 상관관계 레이블의 표시를 제어하고 범주를 정의하는 임계값을 변경하고 각 범위에 사용되는 레이블을 변경할 수 있습니다. 상관관계 값의 특성을 지정하는 방법은 문제점 도메인에 크게 종속되므로 특정 상황에 맞게 범위 및 레이블을 사용자 정의해야 합니다.

출력에 상관관계 강도 레이블 표시. 이 옵션은 기본적으로 선택됩니다. 이 옵션을 선택 취소하면 출력에서 기술통계 레이블이 생략됩니다.

상관관계 강도. 상관관계 강도를 정의하고 레이블을 지정하기 위한 두 개의 옵션이 있습니다.

- **중요도 기준으로 상관관계 강도 정의(1-p).** 평균과의 차분을 단독 가능성으로 설명할 수 있는 1 빼기 유의성 또는 1 빼기 확률로 정의되는 중요도를 기준으로 하여 상관관계 레이블을 지정합니다. 이 값이 1에 가까울수록 두 필드가 독립적이지 *않을* 가능성이 커집니다. 즉, 둘 사이에 어떠한 연관이 존재할 가능성이 커집니다. 일반적으로 절대값보다는 중요도를 기준으로 하여 상관관계 레이블을 지정하는 것을 권장합니다. 이 방법이 데이터의 변동을 고려하기 때문입니다. 예를 들어, 계수 0.6이 한 데이터 세트에서는 매우 유의적이나 다른 데이터 세트에서는 유의적이지 않을 수 있습니다. 기본적으로 0.0에서 0.9 사이의 중요도 값은 *약함*, 0.9에서 0.95 사이는 *중간*, 0.95에서 1.0 사이는 *강함*으로 레이블이 지정됩니다.
- **절대값 기준으로 상관관계 강도 정의.** 위에서 설명한 대로 –1.0에서 1.0 사이의 값을 사용하여 Pearson 상관 계수의 절대값을 기준으로 하여 상관관계 레이블을 지정합니다. 이 측도의 절대값이 1에 가까울수록 상관관계가 강한 것입니다. 기본적으로 (절대값에서) 0.0에서 0.3333 사이의 상관관계는 *약함*, 0.3333에서 0.6666 사이는 *중간*, 0.6666에서 1.0 사이는 *강함*으로 레이블이 지정됩니다. 단, 지정된 값의 유의성은 한 데이터 세트에서 다른 데이터 세트로 일반화하기 어렵습니다. 이런 이유로 인해 대부분의 경우에 절대값이 아니라 확률을 기준으로 하는 상관관계 정의가 사용됩니다.

통계량 출력 브라우저

통계량 노드 출력 브라우저는 통계 분석의 결과를 표시하며 필드 선택, 선택을 기반으로 하여 새 노드 생성 및 결과 저장 및 인쇄 등의 작업을 수행할 수 있도록 해줍니다. 파일 메뉴에서 일반적인 저장, 내보내기 및 인쇄 옵션을 사용할 수 있으며 일반적인 편집 옵션은 편집 메뉴에서 사용할 수 있습니다. 자세한 정보는 305 페이지의 『출력 보기』의 내용을 참조하십시오.

처음 통계량 출력을 찾으면 결과가 펼쳐집니다. 결과를 본 후에 숨기려면 항목의 왼쪽에 있는 펼치기 제어를 사용하여 숨길 특정 결과를 접거나 **모두 접기** 단추를 클릭하여 모든 결과를 접으십시오. 결과를 접은 후에 다시 보려면 항목 왼쪽에 있는 펼치기 제어를 사용하여 결과를 표시하거나 **모두 펼치기** 단추를 클릭하여 모든 결과를 표시하십시오.

출력에는 요청된 통계량 표를 포함하여 각 *탐색적 데이터분석* 필드에 대한 섹션이 포함됩니다.

- **개수.** 필드에 대한 유효한 값이 있는 레코드 수입니다.
- **평균.** 모든 레코드 전체에 걸친 필드의 평균 값입니다.
- **합계.** 모든 레코드 전체에 걸친 필드의 값의 합계입니다.
- **최소.** 필드의 최소값입니다.
- **최대.** 필드의 최대값입니다.
- **범위.** 최소값 및 최대값의 차이입니다.
- **분산.** 필드의 값에서 변동의 측도입니다. 각 값 및 전체 평균 사이의 차분을 구하고 이를 제곱하여 전체 값을 합한 다음 레코드 수로 나누는 방법으로 계산합니다.
- **표준 편차.** 필드 값에서의 또 다른 변동 측도이며 분산의 제곱근으로 계산됩니다.
- **평균의 표준오차.** 평균을 새 데이터에 적용하는 것으로 가정할 때 필드의 평균 추정값의 불확실성 측도입니다.
- **중앙값.** 필드의 "중간" 값입니다. 즉, 필드의 값을 기준으로 하여 데이터를 상한 절반과 하한 절반으로 나누는 값입니다.
- **최빈값.** 데이터에서 가장 흔한 단일값입니다.

상관계수. 상관분석 필드를 지정하면 출력에도 탐색적 데이터분석 필드와 각 상관분석 필드 사이의 Pearson 상관을 나열하는 섹션 및 상관관계 값에 대한 선택적 기술통계 레이블이 포함됩니다. 자세한 정보는 328 페이지의 『상관관계 설정』의 내용을 참조하십시오.

생성 메뉴. 생성 메뉴에는 노드 생성 작업이 포함됩니다.

- **필터.** 다른 필드와 상관관계가 없거나 약한 필드를 필터링하기 위해 필터 노드를 생성합니다.

통계량에서 필터 노드 생성

통계량 출력 브라우저에서 생성된 필터 노드는 다른 필드와의 상관관계를 기준으로 하여 필드를 필터링합니다. 절대값 순서로 상관관계를 정렬하고 (통계량 대화 상자의 생성 필터의 기준 설정에 따라) 가장 큰 상관관계를 구하고 이러한 큰 상관관계에 표시되는 모든 필드를 전달하는 필터를 작성하는 방법으로 작업합니다.

모드. 상관관계를 선택하는 방법을 결정합니다. **포함**을 선택하면 지정된 상관관계에 표시되는 필드가 유지됩니다. **제외**를 선택하면 필드가 필터링됩니다.

표시되는 포함/제외 필드. 상관관계 선택의 기준을 정의합니다.

- **최대 상관관계 수.** 지정된 수의 상관관계 및 해당 상관관계에 표시되는 포함/제외 필드 수를 선택합니다.
- **최대 상관관계 퍼센트(%).** 지정된 퍼센트(*n*%)의 상관관계 및 해당 상관관계에 표시되는 포함/제외 필드 수를 선택합니다.
- **다음보다 큰 상관관계.** 지정된 임계값보다 절대값이 큰 상관관계를 선택합니다.

평균 노드

평균 노드는 독립 집단 사이 또는 관련된 필드의 쌍 사이의 평균을 비교하여 상당한 차이가 존재하는지 여부를 검정합니다. 예를 들어, 프로모션 실행 전후의 평균 수입을 비교하거나 프로모션을 받지 않은 고객으로부터의 수입을 프로모션을 받은 고객으로부터의 수입과 비교할 수 있습니다.

데이터에 따라 두 가지 다른 방식으로 평균을 비교할 수 있습니다.

- **필드 내 그룹 사이.** 독립 그룹을 비교하려면 검정 필드 및 그룹화 필드를 선택하십시오. 예를 들어, 프로모션을 보낼 때 "검증용" 고객의 표본을 제외하고 검증용 그룹에 대한 평균 수입을 모든 다른 그룹과 비교할 수 있습니다. 이 경우 고객이 제안을 받았는지를 표시하는 플래그 또는 명목 필드를 사용하여 각 고객에 대한 수입을 표시하는 단일 검정 필드를 지정할 수 있습니다. 표본은 각 레코드가 하나의 그룹 또는 다른 그룹에 지정된다는 점에서 독립적이며 한 그룹의 특정 멤버를 다른 그룹의 특정 멤버에 링크할 수 있는 방법이 없습니다. 여러 그룹에 대한 평균을 비교하기 위해 셋 이상의 값을 가진 명목 필드를 지정할 수도 있습니다. 실행되면 노드는 선택된 필드에서 일원 ANOVA 검정을 계산합니다. 두 개의 필드 그룹만 있는 경우 일원 ANOVA 결과는 본질적으로 독립 표본 *t* 검정과 동일합니다. 자세한 정보는 『독립 그룹에 대한 평균 비교』의 내용을 참조하십시오.
- **필드 쌍 사이.** 두 관련 필드에 대한 평균을 비교하는 경우 결과가 의미를 가지려면 어떤 방식으로든 그룹이 쌍을 이루어야 합니다. 예를 들어, 프로모션 실행 전후 동일한 고객 그룹으로부터의 평균 수입을 비교하거나 남편-아내 쌍 간 서비스 사용 요금을 비교하여 차이가 있는지 확인할 수 있습니다. 각각의 레코드에는 의미 있게 비교할 수 있는 두 개의 독립되어 있지만 관련된 측도가 포함되어 있습니다. 실행되면 노드는 선택된 각각의 필드 쌍에서 대응 표본 *t* 검정을 계산합니다. 자세한 정보는 『대응 필드 간 평균 비교』의 내용을 참조하십시오.

독립 그룹에 대한 평균 비교

평균 노드에서 **필드 내 그룹 사이**를 선택하여 둘 이상의 독립 그룹에 대한 평균을 비교하십시오.

그룹화 필드. 레코드를 비교할 그룹(예: 제안을 받은 그룹과 제안을 받지 않은 그룹)으로 나누는 둘 이상의 고유 값을 가진 숫자 플래그 또는 명목 필드를 선택하십시오. 검정 필드 수에 관계없이 하나의 그룹화 필드만 선택할 수 있습니다.

검정 필드. 검정할 측도가 포함된 하나 이상의 숫자 필드를 선택하십시오. 선택하는 각각의 필드에 대해 별도의 검정이 수행됩니다. 예를 들어, 사용, 수입 및 이탈에 대한 지정된 프로모션의 영향을 검정할 수 있습니다.

대응 필드 간 평균 비교

평균 노드에서 **필드 쌍 사이**를 선택하여 별도의 필드 간 평균을 비교하십시오. 결과가 의미를 가지려면 이 필드가 어떤 방식으로든 관련되어 있어야 합니다(예: 프로모션 전후의 수입). 다중 필드 쌍도 선택할 수 있습니다.

필드 1. 비교할 첫 번째 측도가 포함된 숫자 필드를 선택하십시오. 전후 연구에서 이는 "전" 필드입니다.

필드 2. 비교할 두 번째 필드를 선택하십시오.

추가. 선택한 쌍을 검정 필드 쌍 목록에 추가합니다.

필요에 따라 필드 선택을 반복하여 여러 쌍을 목록에 추가하십시오.

상관관계 설정. 상관관계의 강도에 레이블을 지정하는 옵션을 지정할 수 있게 합니다. 자세한 정보는 328 페이지의 『상관관계 설정』의 내용을 참조하십시오.

평균 노드 옵션

옵션 탭에서는 결과에 레이블을 지정하는 데 사용된 임계값 *p* 값을 중요, 보통 또는 중요하지 않음으로 설정할 수 있습니다. 각 순위화에 대한 레이블도 편집할 수 있습니다. 중요도는 백분율 척도에 따라 측정되며 우연에 의해서만 관측된 결과만큼 또는 그 이상 극단적인 결과를 얻는 확률(예: 두 필드 간 평균 차이)을 1에서 뺀 것으로 광범위하게 정의될 수 있습니다. 예를 들어, *p* 값이 0.95보다 크면 우연에 의해서만 결과를 설명할 수 있는 가능성이 5% 미안임을 나타냅니다.

중요도 레이블. 출력의 각 필드 쌍 또는 그룹에 레이블을 지정하는 데 사용되는 레이블을 편집할 수 있습니다. 기본 레이블은 *중요*, *보통*, *중요하지 않음*입니다.

절사 값. 각 순위에 대한 임계값을 지정합니다. 일반적으로 *p* 값이 0.95보다 크면 중요로 순위화되고 이 값이 0.9보다 작으면 중요하지 않음으로 순위화되지만 이 임계값은 필요에 따라 조정할 수 있습니다.

참고: 중요도 측도는 다수의 노드에서 사용할 수 있습니다. 구체적인 계산은 노드와 사용된 목표 및 입력 필드의 유형에 따라 다르지만 값은 모두 백분율 척도로 측정되기 때문에 계속 비교할 수 있습니다.

평균 노드 출력 브라우저

평균 출력 브라우저는 교차 분석표 데이터를 표시하며 한 번에 한 행씩 테이블을 선택하여 복사하고 열별로 정렬하고 테이블을 저장 및 인쇄하는 등의 표준 조작을 수행할 수 있게 합니다. 자세한 정보는 305 페이지의 『출력 보기』의 내용을 참조하십시오.

테이블의 구체적인 정보는 비교 유형(별도의 필드 또는 하나의 필드에 있는 그룹)에 따라 다릅니다.

정렬 기준. 특정 열을 기준으로 출력을 정렬할 수 있게 합니다. 위로 또는 아래로 화살표를 클릭하여 정렬 방향을 변경하십시오. 또는 열 표제를 클릭하여 해당 열을 기준으로 정렬할 수 있습니다. (열에서 정렬 방향을 변경하려면 다시 클릭하십시오.)

보기. 단순 또는 **고급**을 선택하여 표시의 세부사항 수준을 제어할 수 있습니다. 고급 보기는 단순 보기의 모든 정보가 포함되어 있으며 추가적인 세부사항도 제공합니다.

필드 내 평균 출력 비교 그룹

필드 내 그룹을 비교하면 그룹화 필드의 이름이 출력 테이블 위에 표시되고 평균 및 관련 통계가 각 그룹에 대해 별도로 보고됩니다. 해당 테이블에는 각 검정 필드에 대한 별도의 행이 포함되어 있습니다.

다음의 열이 표시됩니다.

- **필드.** 선택된 검정 필드의 이름을 나열합니다.

- **그룹별 평균.** 그룹화 필드의 각 범주에 대한 평균을 표시합니다. 예를 들어, 특별 판매 제안(*새 프로모션*)을 받은 사용자를 해당 제안을 받지 않은 사용자(*표준*)와 비교할 수 있습니다. 고급 보기에는 표준 편차, 표준 오차 및 개수도 표시됩니다.
- **중요도.** 중요도 값 및 레이블을 표시합니다. 자세한 정보는 331 페이지의 『평균 노드 옵션』의 내용을 참조하십시오.

고급 출력

고급 보기에는 다음과 같은 추가적인 열이 표시됩니다.

- **F 검정.** 이 검정은 그룹과 각 그룹 내 분산 사이의 분산 비율을 기반으로 합니다. 모든 그룹에 대해 평균이 동일하면 둘 다 동일한 모집단 분산의 추정값이므로 *F*비가 1에 가까울 것으로 예상합니다. 이 비율이 클수록 그룹 간 변동이 더 크고 유의차가 존재할 가능성이 더 높습니다.
- **자유도.** 자유도를 표시합니다.

필드의 평균 출력 비교 쌍

개별 필드를 비교하면 출력 테이블에 선택된 각 필드 쌍에 대해 하나의 행이 포함됩니다.

- **필드 1/2.** 각 쌍에서 첫 번째 및 두 번째 필드의 이름을 표시합니다. 고급 보기에는 표준 편차, 표준 오차 및 개수도 표시됩니다.
- **평균 1/2.** 각 필드에 대한 평균을 각각 표시합니다.
- **상관관계.** 두 연속형(숫자 범위) 필드 간 관계의 강도를 측정합니다. +1.0에 가까운 값은 강한 긍정적인 연관을 표시하고 –1.0에 가까운 값은 강한 부정적인 연관을 표시합니다. 자세한 정보는 328 페이지의 『상관관계 설정』의 내용을 참조하십시오.
- **평균 차이.** 두 필드 평균 간 차이를 표시합니다.
- **중요도.** 중요도 값 및 레이블을 표시합니다. 자세한 정보는 331 페이지의 『평균 노드 옵션』의 내용을 참조하십시오.

고급 출력

고급 출력은 다음의 열을 추가합니다.

95% 신뢰구간. 참 평균이 이 모집단에서 이 크기의 가능한 모든 표본 중 95%에 속할 수 있는 범위의 상한 및 하한입니다.

T 검정. *t* 통계는 평균 차이를 해당 표준 오차로 나눠서 얻습니다. 이 통계의 절대값이 클수록 평균이 동일하지 않을 확률이 높습니다.

자유도. 통계의 자유도를 표시합니다.

보고서 노드

보고서 노드를 사용하면 고정 텍스트뿐 아니라 데이터 및 데이터로부터 파생된 기타 표현식을 포함한 형식화된 보고서를 작성할 수 있습니다. 텍스트 템플리트를 사용하여 보고서의 형식을 지정하여 고정 텍스트 및 데이터 출력 생성을 정의합니다. 템플리트에서 HTML 태그를 사용하고 출력 탭에서 옵션을 설정하여 사용자 정의 텍스트 형식화를 제공할 수 있습니다. 템플리트에서 CLEM 표현식을 사용하면 데이터 값과 기타 조건부 출력이 보고서에 포함됩니다.

보고서 노드에 대한 대안

보고서 노드는 특정 조건을 충족하는 모든 레코드와 같이 스트림의 레코드 또는 케이스 출력을 나열하는 데 가장 일반적으로 사용됩니다. 따라서 테이블 노드에 대한 덜 구조화된 대안으로 생각될 수 있습니다.

- 유형 노드에서 지정된 필드 정의와 같은 데이터 자체가 아니라 필드 정보 또는 스트림에서 정의된 사항 등이 나열된 보고서를 원하는 경우에는 스크립트를 대신 사용할 수 있습니다.
- 다중 출력 개체(하나 이상의 스트림에 의해 생성된 모델, 테이블 및 그래프 콜렉션 등)를 포함하며 다중 형식(텍스트, HTML 및 Microsoft Word/Office 등)의 출력으로 사용할 수 있는 보고서를 원하는 경우에는 IBM SPSS Modeler 프로젝트를 사용하십시오.
- 스크립트를 사용하지 않고 필드 이름 목록을 생성하려면 모든 레코드를 삭제하는 표본 노드를 먼저 사용하고 테이블 노드를 사용할 수 있습니다. 그러면 행이 없는 테이블이 생성되어 단일 열에 필드 이름 목록을 생성하도록 내보낼 때 전치될 수 있습니다. (수행하려면 테이블 노드의 출력 탭에서 **데이터 전치**를 선택하십시오.)

보고서 노드 템플리트 탭

템플리트 작성. 보고서의 내용을 정의하기 위해 보고서 노드 템플리트 탭에서 템플리트를 작성할 수 있습니다. 템플리트는 각 줄이 보고서의 내용을 지정하는 텍스트, 내용 줄의 범위를 표시하는 데 사용되는 특수 태그 줄로 구성됩니다. 각 내용 줄 내의 대괄호([])로 묶인 CLEM 표현식은 해당 줄을 보고서에 보내기 전에 평가됩니다. 템플리트 내의 줄에 대해 가능한 범위는 세 가지입니다.

고정됨. 달리 표시되지 않은 줄은 고정된 것으로 간주됩니다. 고정된 줄은 포함된 표현식이 모두 평가된 후에 보고서에 한 번만 복사됩니다. 예를 들어,

```
This is my report, printed on [@TODAY]
```

줄은 텍스트 및 현재 날짜를 포함하는 단일 줄을 보고서에 복사합니다.

글로벌(반복계산 ALL). 특수 태그 #ALL 및 # 사이에 포함된 줄은 입력 데이터의 각 레코드에 대해 한 번씩 보고서에 복사됩니다. 대괄호로 묶인 CLEM 표현식은 각 출력 줄의 현재 레코드를 기반으로 하여 평가됩니다. 예를 들어,

```
#ALL
For record [@INDEX], the value of AGE is [AGE]
#
```

줄은 각 레코드에 대해 레코드 번호 및 연령을 표시하는 한 줄을 포함합니다.

모든 레코드의 목록을 생성하려면 다음을 수행하십시오.

```
#ALL
[Age] [Sex] [Cholesterol]  [BP]
#
```

조건부(반복계산 WHERE). 특수 태그 #WHERE <*condition*> 및 # 사이에 포함된 줄은 지정된 조건이 참인 경우에 각 레코드에 대해 한 번씩 보고서에 복사됩니다. 조건은 CLEM 표현식입니다. (WHERE 조건에서 대괄호는 선택적입니다.) 예를 들어,

```
#WHERE [SEX = 'M']
Male at record no. [@INDEX] has age [AGE].
#
```

줄은 성별에 대해 *M* 값을 갖는 각 레코드에 대해 파일에 한 줄을 작성합니다. 완전한 보고서는 템플리트를 입력 데이터에 적용하여 정의된 고정, 글로벌 및 조건부 줄을 포함합니다.

출력 탭을 사용하여 결과를 표시하거나 저장하기 위해 다양한 유형의 출력 노드에 대해 공통적인 옵션을 지정할 수 있습니다. 자세한 정보는 308 페이지의 『출력 노드 출력 탭』의 내용을 참조하십시오.

HTML 또는 XML 형식의 출력 데이터

이러한 형식 중 하나로 보고서를 쓰기 위해 템플리트에 직접 HTML 또는 XML 태그를 포함할 수 있습니다. 예를 들어, 다음 템플리트는 HTML 테이블을 생성합니다.

```
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>
```

보고서 노드 출력 브라우저

보고서 브라우저는 사용자에게 생성된 보고서의 내용을 표시합니다. 파일 메뉴에서 일반적인 저장, 내보내기 및 인쇄 옵션을 사용할 수 있으며 일반적인 편집 옵션은 편집 메뉴에서 사용할 수 있습니다. 자세한 정보는 305 페이지의 『출력 보기』의 내용을 참조하십시오.

전역값 설정 노드

전역값 설정 노드는 데이터를 스캔하고 CLEM 표현식에서 사용할 수 있는 요약 값을 계산합니다. 예를 들어, 전역값 설정 노드를 사용하여 *age*라는 필드에 대한 통계량을 계산한 후 `@GLOBAL_MEAN(age)` 함수를 삽입하여 CLEM 표현식에서 *age*의 전체 평균을 사용할 수 있습니다.

전역값 설정 노드 설정 탭

전역값 작성. 전역값을 사용 가능하도록 만들 필드를 선택합니다. 다중 필드를 선택할 수 있습니다. 각 필드에 대해 필드 이름 옆의 열에 원하는 통계가 선택되었는지 확인하여 계산할 통계를 지정하십시오.

- **평균.** 모든 레코드 전체에 걸친 필드의 평균 값입니다.
- **합계.** 모든 레코드 전체에 걸친 필드의 값의 합계입니다.
- **최소.** 필드의 최소값입니다.
- **최대.** 필드의 최대값입니다.
- **표준편차.** 표준 편차입니다. 필드 값에서의 변동 측도이며 분산의 제곱근으로 계산됩니다.

기본 작업. 새 필드가 위의 전역값 목록에 추가될 때 여기서 선택되는 옵션이 사용됩니다. 기본 통계 집합을 변경하려면 적절히 통계량을 선택 또는 선택 취소하십시오. 또한 **적용** 단추를 사용하여 기본 작업을 목록 내의 모든 필드에 적용할 수 있습니다.

참고: 일부 작업은 숫자가 아닌 필드(날짜/시간 필드의 합계 등)에는 적용할 수 없습니다. 선택된 필드와 함께 사용할 수 없는 작업은 사용할 수 없도록 설정됩니다.

실행 전에 모든 전역값 선택 취소. 새로운 값을 계산하기 전에 모든 전역값을 제거하려면 이 옵션을 선택하십시오. 이 옵션을 선택하지 않으면 새로 계산된 값이 기존값을 대체하나 다시 계산되지 않은 전역값은 사용 가능한 상태로 유지됩니다.

실행 후에 작성된 전역값 미리보기 표시. 이 옵션을 선택하면 계산된 전역값을 표시하기 위해 실행 후에 스트림 특성 대화 상자의 전역값 탭이 표시됩니다.

시뮬레이션 적합 노드

시뮬레이션 적합 노드는 후보 통계 분포 집합을 데이터 내의 각 필드에 맞춥니다. 필드에 대한 각 분포의 적합도는 적합도 기준을 사용하여 평가됩니다. 시뮬레이션 적합 노드가 실행될 때 시뮬레이션 생성 노드가 작성되거나 기존 노드가 업데이트됩니다. 각 필드에 가장 적합한 분포가 지정됩니다. 시뮬레이션 생성 노드를 사용하여 각 필드에 대한 시뮬레이션된 데이터를 생성할 수 있습니다.

시뮬레이션 적합 노드가 터미널 노드이더라도 생성된 모형 팔레트에 모형을 추가하거나 출력 탭에 출력 또는 도표를 추가하거나 데이터를 내보내지 않습니다.

참고: 히스토리 데이터가 희박한 경우, 즉, 결측값이 많은 경우, 분포를 데이터에 맞추기에 충분한 유효한 값을 찾기 위해 구성요소를 맞추는 데 어려움이 있을 수 있습니다. 데이터가 희박한 경우, 맞춤 수행 전에 희박한 데이터가 필수가 아니면 제거하거나 결측값을 대체해야 합니다. 데이터 검토 노드의 **품질** 탭에서 옵션을 사용하여 완료된 레코드의 수를 보고 희박한 필드를 식별하고 대체 방법을 선택할 수 있습니다. 레코드 수가 분포 맞춤에 대해 충분하지 않으면 균형 노드를 사용하여 레코드 수를 늘릴 수 있습니다.

시뮬레이션 생성 노드를 자동으로 작성하기 위해 시뮬레이션 적합 노드 사용

시뮬레이션 적합 노드가 처음 실행될 때 시뮬레이션 적합 노드에 대한 업데이트 링크가 있는 시뮬레이션 생성 노드가 작성됩니다. 시뮬레이션 적합 노드가 다시 실행될 때 업데이트 링크가 제거된 경우에만 새 시뮬레이션 생성 노드가 작성됩니다. 시뮬레이션 적합 노드는 연결된 시뮬레이션 생성 노드를 업데이트하는 데에도 사용될 수 있습니다. 결과는 동일한 필드가 두 노드에 모두 존재하는지, 필드가 시뮬레이션 생성 노드에서 잠겨있지 않은지 여부에 따라 다릅니다. 자세한 정보는 54 페이지의 『시뮬레이션 생성 노드』의 내용을 참조하십시오.

시뮬레이션 적합 노드는 시뮬레이션 생성 노드에 대한 업데이트 링크만 가질 수 있습니다. 시뮬레이션 생성 노드에 대한 업데이트를 정의하려면 다음 단계를 따르십시오.

1. 시뮬레이션 적합 노드를 마우스 오른쪽 단추로 클릭하십시오.
2. 메뉴에서 **업데이트 링크 정의**를 선택하십시오.
3. 업데이트 링크를 정의할 시뮬레이션 생성 노드를 클릭하십시오.

시뮬레이션 적합 노드 및 시뮬레이션 생성 노드 사이의 업데이트 링크를 제거하려면 마우스 오른쪽 단추로 업데이트 링크를 클릭하고 **링크 제거**를 선택하십시오.

분포 적합

통계 분포는 변수가 취할 수 있는 발생 값의 이론적 빈도입니다. 시뮬레이션 적합 노드에서 이론적 통계 분포 집합은 데이터의 각 필드와 비교됩니다. 맞춤에 대해 사용 가능한 분포에 대해서는 65 페이지의 『분포』 제목에서 설명합니다. 이론적 분포의 모수는 적합도(Anderson-Darling 기준 또는 Kolmogorov-Smirnov 기준)에 따라 데이터에 최적 맞춤될 수 있도록 조정됩니다. 시뮬레이션 적합 노드에 의한 분포 맞춤의 결과는 어떤 분포가 적합한지, 각 분포에 대한 모수의 최적 추정값 및 각 분포가 데이터에 적합한 정도 등을 표시합니다. 분포 적합 작업 중에 수치 저장 유형이 있는 필드 사이의 상관계수 및 범주형 분포가 있는 필드 사이의 우연성도 계산됩니다. 분포 맞춤의 결과는 시뮬레이션 생성 노드를 작성하는 데 사용됩니다.

분포가 데이터에 맞춰지기 전에 첫 번째 1000 개의 레코드에서 결측값을 검사합니다. 결측값이 너무 많으면 분포 맞춤이 가능하지 않습니다. 그런 경우, 다음 옵션 중 적합한 옵션을 선택해야 합니다.

- 결측값이 있는 레코드를 제거하기 위해 업스트림 노드 사용.
- 결측값에 대한 레코드를 대체하기 위해 업스트림 노드 사용.

분포 맞춤은 사용자 결측값을 제외하지 않습니다. 데이터에 사용자 결측값이 있으며 이러한 값을 분포 맞춤에서 제외하려면 해당 값을 시스템 결측값으로 설정해야 합니다.

분포가 맞춰질 때 필드의 역할은 고려되지 않습니다. 예를 들어, 역할이 **목표**인 필드는 역할이 **입력, 없음, 모두, 파티션, 분할, 빈도** 및 **ID**인 필드와 동일하게 처리됩니다.

필드는 저장 유형 및 측정 수준에 따라 분포 적합 작업 중에 다르게 처리됩니다. 분포 적합 작업 중의 필드 처리에 대해서는 다음 표에서 설명합니다.

표 40. 필드의 저장 유형 및 측정 수준에 따른 분포 적합

<table>
<tr><th rowspan="2">저장 유형</th><th colspan="6">측정 수준</th></tr>
<tr><th>연속형</th><th>범주형</th><th>플래그</th><th>명목형</th><th>순서</th><th>유형 없음</th></tr>
<tr><td>문자열</td><td>불가능</td><td colspan="3">범주형, Dice 및 고정 분포가 맞춰집니다.</td><td></td><td rowspan="8">필드가 무시되어 시뮬레이션 생성 노드로 전달되지 않습니다.</td></tr>
<tr><td>정수</td><td rowspan="6">모두 분포가 맞춰집니다. 상관관계 및 우연성이 계산됩니다.</td><td colspan="3" rowspan="6">범주형 분포가 맞춰집니다. 상관관계는 계산되지 않습니다.</td><td rowspan="6">이항, 음이항 및 포아송 분포가 맞춰지고 상관관계가 계산됩니다.</td></tr>
<tr><td>실수</td></tr>
<tr><td>시간</td></tr>
<tr><td>날짜</td></tr>
<tr><td>시간소인</td></tr>
<tr><td>알 수 없음</td><td colspan="5">적절한 저장 유형이 데이터에서 결정됩니다.</td></tr>
</table>

측정 수준 순서가 있는 필드는 연속형 필드처럼 처리되고 시뮬레이션 생성 노드의 상관관계 표에 포함됩니다. 이항, 음이항 또는 포아송 외의 분포를 순서 필드에 맞추려면 필드의 측정 수준을 연속형으로 변경해야 합니다. 순서 필드의 각 값에 대해 앞에서 레이블을 정의한 경우, 측정 수준을 연속형으로 변경하면 레이블이 손실됩니다.

단일값을 가진 필드는 분포 적합 작업 중에 다중 값을 가진 필드와 다르게 처리되지 않습니다. 저장 유형 시간, 날짜 또는 시간소인을 가진 필드는 수치로 처리됩니다.

분할 필드에 분포 적합

데이터에 분할 필드가 포함되며 각 분할에 대해 분포 맞춤을 별도로 수행하려면 업스트림 구조변환 노드를 사용하여 데이터를 변환해야 합니다. 구조변환 노드를 사용하여 분할 필드의 각 값에 대해 새 필드를 생성하십시오. 그러면 이 구조변환된 데이터를 시뮬레이션 적합 노드에서 분포 적합 작업 중에 사용할 수 있습니다.

시뮬레이션 적합 노드 설정 탭

소스 노드 이름. 자동을 선택하여 자동으로 생성되거나 업데이트된 시뮬레이션 노드의 이름을 생성할 수 있습니다. 자동으로 생성되는 이름은 사용자 정의 이름이 지정된 경우에는 시뮬레이션 적합 노드에서 지정된 이름이며 시뮬레이션 적합 노드에서 사용자 정의 이름이 지정되지 않은 경우에는 Sim Gen입니다. 인접한 텍스트 필드에서 사용자 정의 이름을 지정하려면 **사용자 정의**를 선택하십시오. 텍스트 필드를 편집하지 않은 한 기본 사용자 정의 이름은 Sim Gen입니다.

맞춤 옵션 이러한 옵션을 사용하여 분포가 필드에 맞춰지는 방법 및 분포의 맞춤이 평가되는 방법을 지정할 수 있습니다.

- **표본추출할 케이스 수.** 데이터 세트 내의 필드로 분포를 맞출 때 사용할 케이스 수를 지정합니다. 데이터 내의 모든 레코드에 분포를 맞추려면 **모두**를 선택하십시오. 데이터 세트가 매우 크면 분포 맞춤에 사용할 케이스 수를 제한하는 방법을 고려할 수도 있습니다. 첫 N 케이스만 사용하려면 **첫 N 케이스로 제한**을 선택하십시오. 사용할 케이스 수를 지정하려면 화살표를 클릭하십시오. 또는 업스트림 노드를 사용하여 분포 맞춤에 대한 레코드의 임의 표본을 사용할 수 있습니다.
- **기준 적합도(연속형 필드 전용).** 연속형 필드의 경우, 필드에 대한 분포를 맞출 때 Anderson-Darling 검정 또는 Kolmogorov-Smirnoff 검정 적합도를 선택하여 분포 순위를 매기십시오. Anderson-Darling 검정이 기본적으로 선택되며 끝 영역에서 가장 적합한 맞춤을 사용하려면 특히 권장됩니다. 모든 후보 분포에 대해 모든 통계량이 계산되나 분포 정렬 및 가장 적합한 분포 적합 판별에는 선택된 통계만 사용됩니다.
- **구간(경험적 분포 전용).** 연속형 필드의 경우, 경험적 분포는 히스토리 데이터의 누적 분포 함수입니다. 각 값 또는 값 범위의 확률이며 데이터에서 직접 파생됩니다. 화살표를 클릭하여 연속형 필드에 대한 경험적 분포를 계산하는 데 사용되는 구간 수를 지정할 수 있습니다. 기본값은 100이며 최대값은 1000입니다.
- **가중 필드(선택적).** 데이터 세트에 가중 필드가 포함된 경우, 필드 선택도구 아이콘을 클릭하여 목록에서 가중 필드를 선택하십시오. 그러면 분포 적합 프로세스에서 가중 필드가 제외됩니다. 목록은 측정 수준이 연속적인 데이터 세트 내의 모든 필드를 표시합니다. 가중 필드는 한 개만 선택할 수 있습니다.

시뮬레이션 평가 노드

시뮬레이션 평가 노드는 지정된 필드를 평가하고 필드의 분포를 제공하고 분포 및 상관계수 도표를 생성하는 터미널 노드입니다. 이 노드는 주로 연속형 필드를 평가하는 데 사용됩니다. 따라서 평가 노드에 의해 생성되는 평가 차트를 보완하며 이산형 필드 평가에 유용합니다. 또 다른 차이는 시뮬레이션 평가 노드는 여러 반복에 걸친 단일 예측을 평가하는 반면 평가 노드는 단일 반복이 있는 다중 평가를 각각 평가한다는 점입니다. 시뮬레이션 생성 노드의 분포모수에 대해 둘 이상의 값이 지정된 경우에 반복이 생성됩니다. 자세한 정보는 64 페이지의 『반복』의 내용을 참조하십시오.

시뮬레이션 평가 노드는 시뮬레이션 적합 및 시뮬레이션 생성 노드에서 얻은 데이터를 사용하여 계획됩니다. 단, 노드는 다른 임의의 노드와 함께 사용될 수 있습니다. 시뮬레이션 생성 노드 및 시뮬레이션 평가 노드 사이에 수에 관계없이 처리 단계를 배치할 수 있습니다.

중요사항: 시뮬레이션 평가 노드에는 목표 필드에 대한 유효한 값이 있는 1000개 이상의 레코드가 필요합니다.

시뮬레이션 평가 노드 설정 탭

시뮬레이션 평가 노드의 설정 탭에서 데이터 세트 내의 각 필드의 역할을 지정하고 시뮬레이션에 의해 생성되는 출력을 사용자 정의할 수 있습니다.

항목 선택. 시뮬레이션 평가 노드의 세 가지 뷰(필드, 밀도 함수 및 출력) 사이에서 전환할 수 있습니다.

필드 보기

목표 필드. 필수 필드입니다. 드롭 다운 목록에서 데이터 세트의 목표 필드를 선택하려면 화살표를 클릭하십시오. 선택된 필드는 연속형, 순서 또는 명목 측정 수준은 가질 수 있으나 날짜 또는 지정되지 않은 측정 수준은 가질 수 없습니다.

반복 필드(선택적). 데이터에 데이터 내의 각 레코드가 속한 반복을 표시하는 반복 필드가 있으면 여기서 선택해야 합니다. 즉, 각 반복이 별도로 평가됩니다. 연속형, 순서 또는 명목 측정 수준의 필드만 선택할 수 있습니다.

입력 데이터가 이미 반복에 의해 정렬되어 있음. 반복 필드가 **반복 필드(선택적)** 필드에서 지정되는 경우에만 사용 가능합니다. 입력 데이터가 **반복 필드(선택적)**에서 지정된 반복 필드에 의해 이미 정렬되었음을 확신하는 경우에만 이 옵션을 선택하십시오.

구성할 최대 반복 수. 반복 필드가 **반복 필드(선택적)** 필드에서 지정되는 경우에만 사용 가능합니다. 구성할 반복계산 수를 지정하려면 화살표를 클릭하십시오. 이 번호를 지정하면 단일 도표에 너무 많은 반복을 구성하여 도표 해석을 어렵게 만드는 것을 방지할 수 있습니다. 설정 가능한 최저 수준의 최대반복수는 2입니다. 최고 수준은 50입니다. 구성할 최대반복수는 처음에는 10으로 설정됩니다.

상관관계 토네이도에 대한 입력 필드. 상관관계 토네이도 도표는 지정된 목표 및 지정된 각 입력 사이의 상관계수를 표시하는 막대형 차트입니다. 사용가능한 시뮬레이션한 입력 목록에서 필드 선택도구 아이콘을 클릭하여 토네이도 도표에 포함할 입력 필드를 선택하십시오. 연속형 및 순서형 측정 수준이 있는 입력 필드만 선택할 수 있습니다. 명목형, 유형 없음 및 날짜 입력 필드는 목록에 사용할 수 없으며 선택할 수 없습니다.

밀도 함수 보기

이 보기의 옵션을 사용하면 범주형 대상에 대한 예측값의 막대형 차트와 마찬가지로 연속형 대상에 대한 확률 밀도 함수 및 누적 분포 함수의 출력을 사용자 정의할 수 있습니다.

밀도함수. 밀도함수는 시뮬레이션에서 결과 세트를 조사하는 기본적인 방법입니다.

- **확률 밀도 함수(PDF).** 목표 필드에 대한 확률 밀도 함수를 생성하려면 이 옵션을 선택하십시오. 확률 밀도 함수는 목표 값의 분포를 표시합니다. 확률 밀도 함수를 사용하면 목표가 특정 영역 내에 있을 확률을 판별할 수 있습니다. 범주형 대상(측정 수준이 명목형 또는 순서인 경우)의 경우, 각 범주의 대상이 해당되는 케이스 퍼센트를 표시하는 막대형 차트가 생성됩니다.
- **누적 분포 함수(CDF).** 목표 필드에 대한 누적 분포 함수를 생성하려면 이 옵션을 선택하십시오. 누적 분포 함수는 대상의 값이 지정된 값 이하인 확률을 표시합니다. 연속형 대상에만 사용 가능합니다.

참조선(연속형). 이러한 옵션은 **확률 밀도 함수(PDF)** 또는 **누적 분포 함수(CDF)** 또는 둘 다 선택된 경우에만 사용 가능합니다. 해당 옵션을 사용하면 확률 밀도 함수 및 누적 분포 함수에 다양한 고정된 수직 참조선을 추가할 수 있습니다.

- **평균.** 목표 필드의 평균 값에 참조선을 추가하려면 이 옵션을 선택하십시오.
- **중앙값.** 목표 필드의 중앙값에 참조선을 추가하려면 이 옵션을 선택하십시오.

- **표준 편차.** 목표 필드의 평균 값에서부터 지정된 수의 표준편차 더하기 및 빼기에 참조선을 추가하려면 이 옵션을 선택하십시오. 이 옵션을 선택하면 인접한 **숫자** 필드를 사용할 수 있습니다. 표준편차의 수를 지정하려면 화살표를 클릭하십시오. 최소 표준편차 수는 1이며 최대수는 10입니다. 표준편차의 수는 처음에 3으로 설정됩니다.
- **백분위수.** 목표 필드의 분포의 두 개의 백분위수 값에 참조선을 추가하려면 이 옵션을 선택하십시오. 이 옵션을 선택하면 인접한 **아래쪽** 및 **위쪽** 텍스트 필드를 사용할 수 있습니다. 예를 들어, **위쪽** 텍스트 필드에 90 값을 입력하면 목표의 90번째 백분위수에 참조선이 추가됩니다. 이 값은 관측값의 90% 아래에 해당하는 값입니다. 이와 유사하게 **아래쪽** 텍스트 필드의 10 값은 목표의 열 번째 백분위수를 나타내며 관측값의 10% 아래에 해당되는 값입니다.
- **사용자 정의 참조선.** 수평축 변수와 함께 지정된 값에 참조선을 추가하려면 이 옵션을 선택하십시오. 이 옵션을 선택하면 인접한 **값** 테이블을 사용할 수 있습니다. 유효한 숫자를 **값** 테이블에 입력할 때마다 비어 있는 새 행이 테이블의 아래 쪽에 추가됩니다. *유효한* 수는 목표 필드의 값 범위 내의 수입니다.

참고: (다중 반복으로부터) 다중 밀도함수 또는 분포 함수가 단일 차트에 표시되는 경우, 사용자 정의 선이 아니라 참조선이 별도로 각 함수에 적용됩니다.

범주형 목표(PDF만 해당). 이러한 옵션은 **확률 밀도 함수(PDF)**가 선택된 경우에만 사용 가능합니다.

- **보고할 범주 값.** 범주형 대상 필드가 있는 모형의 경우, 모형의 결과는 목표 값이 각 범주에 속하는 각 범주에 대해 하나씩 존재하는 예측 확률의 집합입니다. 가장 높은 확률의 범주가 예측 범주가 되고 확률 밀도 함수에 대한 막대형 차트를 생성하는 데 사용됩니다. 막대형 차트를 생성하려면 **예측 범주**를 선택하십시오. 목표 필드의 각 범주에 대한 예측 확률 분포 히스토그램을 생성하려면 **예측 확률**을 선택하십시오. 또한 두 유형의 차트를 모두 생성하기 위해 **모두**를 선택할 수 있습니다.
- **민감도 분석의 집단화.** 민감도 분석 반복계산이 포함된 시뮬레이션은 분석에 의해 정의되는 각 반복에 대해 독립적 목표 필드(또는 모델의 예측 목표 필드)를 생성합니다. 변형되는 분포모수의 각 값에 대해 하나의 반복이 있습니다. 반복이 있으면 범주형 목표 필드의 예측 범주의 막대형 차트가 모든 반복의 결과를 포함하는 수평배열 막대도표로 표시됩니다. **범주를 함께 집단화** 또는 **반복계산을 함께 집단화**를 선택하십시오.

출력 보기

목표 분포의 백분위수 값. 이 옵션을 사용하면 목표 분포의 백분위수 값의 표를 만들고 표시할 백분위수를 지정할 수 있습니다.

백분위수 값의 표 만들기. 연속형 대상 필드의 경우, 목표 분포의 지정된 백분위수 표를 얻으려면 이 옵션을 선택하십시오. 다음 옵션 중 하나를 선택하여 백분위수를 지정하십시오.

- **사분위수.** 사분위수는 목표 필드 분포의 25번째, 50번째, 75번째 백분위수입니다. 관측값은 네 그룹의 동일한 *크기*로 나뉩니다.
- **구간.** 네 개가 아닌 동일한 수의 그룹이 필요하면 **구간**을 선택하십시오. 이 옵션을 선택하면 인접한 **숫자** 필드를 사용할 수 있습니다. 구간 수를 지정하려면 화살표를 클릭하십시오. 최소 구간 수는 2이며 최대수는 100입니다. 구간 수는 처음에 10으로 설정됩니다.

- **사용자 정의 백분위수.** 개별 백분위수(예를 들어, 99번째 백분위수)를 지정하려면 **사용자 정의 백분위수**를 선택하십시오. 이 옵션을 선택하면 인접한 **값** 테이블을 사용할 수 있습니다. 1에서 100 사이의 유효한 숫자를 **값** 테이블에 입력할 때마다 비어 있는 새 행이 테이블의 아래 쪽에 추가됩니다.

시뮬레이션 평가 노드 출력

시뮬레이션 평가 노드가 실행될 때 출력이 출력 관리자에 추가됩니다. 시뮬레이션 평가 출력 브라우저는 시뮬레이션 평가 노드의 실행 결과를 표시합니다. **파일** 메뉴에서 일반적인 저장, 내보내기 및 인쇄 옵션을 사용할 수 있으며 일반적인 편집 옵션은 **편집** 메뉴에서 사용할 수 있습니다. 자세한 정보는 305 페이지의 『출력 보기』 주제를 참조하십시오. 도표 중 하나를 선택하면 **보기** 메뉴만 사용 가능합니다. 분포 테이블 또는 정보 출력에는 사용할 수 없습니다. **보기** 메뉴에서 **편집 모드**를 선택하여 도표의 레이아웃 및 모양을 변경하거나 **탐색 모드**를 선택하여 도표에 표시되는 데이터 및 값을 탐색할 수 있습니다. 정적 모드는 도표 참조선 및 슬라이더를 이동할 수 없도록 현재 위치에서 고정합니다. 정적 모드는 참조선이 있는 도표를 복사, 내보내기 또는 인쇄할 수 있는 유일한 모드입니다. 이 모드를 선택하려면 **보기** 메뉴에서 **정적 모드**를 클릭하십시오.

시뮬레이션 평가 출력 브라우저 창은 두 개의 패널로 구성됩니다. 창의 왼쪽에는 시뮬레이션 평가 노드가 실행될 때 생성된 도표의 썸네일 표시가 표시되는 탐색 패널이 있습니다. 썸네일을 선택하면 창의 오른쪽에 있는 패널에 도표 출력이 표시됩니다.

탐색 패널

출력 브라우저의 탐색 패널에는 시뮬레이션에서 생성된 도표의 썸네일이 포함됩니다. 탐색 패널에 표시되는 썸네일은 목표 필드의 측정 수준 및 시뮬레이션 평가 노드 대화 상자에서 선택된 옵션에 따라 다릅니다. 썸네일에 대한 설명은 다음 표에서 제공합니다.

표 41. 탐색 패널 썸네일

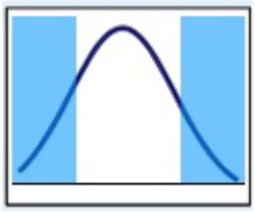
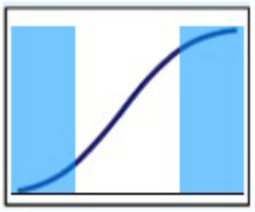
썸네일	설명	설명
	확률 밀도 함수	이 썸네일은 목표 필드의 측정 수준이 연속형이고 시뮬레이션 평가 노드 대화 상자의 밀도 함수 보기에서 **확률 밀도 함수(PDF)**가 선택된 경우에만 표시됩니다. 목표 필드의 측정 수준이 범주형이면 이 썸네일이 표시되지 않습니다.
	누적 분포 함수	이 썸네일은 목표 필드의 측정 수준이 연속형이고 시뮬레이션 평가 노드 대화 상자의 밀도 함수 보기에서 **누적 분포 함수(CDF)**가 선택된 경우에만 표시됩니다. 목표 필드의 측정 수준이 범주형이면 이 썸네일이 표시되지 않습니다.

표 41. 탐색 패널 썸네일 (계속)

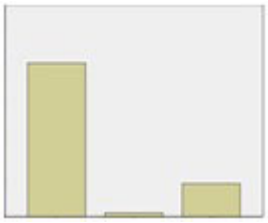
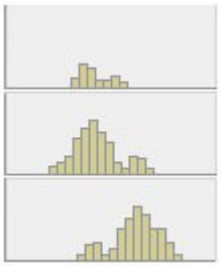
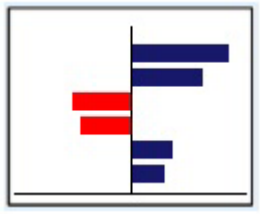
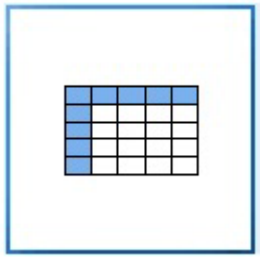
썸네일	설명	설명
	예측된 범주 값	이 썸네일은 목표 필드의 측정 수준이 범주형이고 시뮬레이션 평가 노드 대화 상자의 밀도 함수 보기에서 **확률 밀도 함수(PDF)**가 선택되고 **보고할 범주 값** 영역에서 **예측된 범주** 또는 **모두**가 선택된 경우에만 표시됩니다. 목표 필드의 측정 수준이 연속형이면 이 썸네일이 표시되지 않습니다.
	예측된 범주 확률	이 썸네일은 목표 필드의 측정 수준이 범주형이고 시뮬레이션 평가 노드 대화 상자의 밀도 함수 보기에서 **확률 밀도 함수(PDF)**가 선택되고 **보고할 범주 값** 영역에서 **예측된 확률** 또는 **모두**가 선택된 경우에만 표시됩니다. 목표 필드의 측정 수준이 연속형이면 이 썸네일이 표시되지 않습니다.
	토네이도 도표	이 썸네일은 시뮬레이션 평가 노드 대화 상자의 필드 보기의 **상관관계 토네이도의 입력 필드** 필드에서 하나 이상의 입력 필드가 선택된 경우에만 표시됩니다.
	분포 표	이 썸네일은 목표 필드의 측정 수준이 연속형이고 시뮬레이션 평가 노드 대화 상자의 출력 보기에서 **백분위수 값의 표 만들기**가 선택된 경우에만 표시됩니다. 이 도표에는 **보기** 메뉴를 사용할 수 없습니다. 목표 필드의 측정 수준이 범주형이면 이 썸네일이 표시되지 않습니다.
	정보	이 썸네일은 항상 표시됩니다. 이 출력에는 **보기** 메뉴를 사용할 수 없습니다.

도표 출력

사용 가능한 출력 도표의 유형은 목표 필드의 측정 수준, 반복 필드 사용 여부 및 시뮬레이션 평가 노드 대화 상자에서 선택된 옵션에 따라 다릅니다. 시뮬레이션에서 생성된 수많은 도표는 표시를 사용자 정의할 수 있는 대화형 기능을 갖고 있습니다. 대화형 기능은 **도표 옵션**을 클릭하여 사용 가능합니다. 모든 시뮬레이션 도표는 그래프 보드로 시각화됩니다.

연속형 대상의 확률 밀도 함수 차트. 이 도표는 확률 및 빈도를 모두 표시하며 왼쪽 수직 축에 확률 척도가 있으며 오른쪽 수직 축에 빈도 척도가 있습니다. 도표는 두 개의 슬라이딩 수직 참조선을 사용하여 별도의 영역으로 구분됩니다. 도표 아래 표는 각 영역 내의 분포 퍼센트를 표시합니다. 반복으로 인해 동일한 도표에 다중 밀도함수가 표시되는 경우, 표에 각 밀도 함수와 연관된 확률에 대한 별도의 행, 반복 이름을 포함하는 추가 열 및 각 밀도 함수와 연관된 색상이 있을 수 있습니다. 반복은 반복 레이블에 따라 표에 문자순으로 표시됩니다. 반복 레이블을 사용할 수 없으면 반복 값이 대신 사용됩니다. 표는 편집할 수 없습니다.

각 참조선에는 선을 쉽게 이동할 수 있는 슬라이더(역삼각형)가 있습니다. 각 슬라이더에는 현재 위치를 표시하는 레이블이 있습니다. 기본적으로 슬라이더는 분포의 5번째 및 95번째 백분위수에 위치합니다 다중 반복이 있는 경우, 표에 나열된 첫 번째 반복의 5번째 및 95번째 백분위수에 슬라이더가 위치합니다. 선을 서로 교차하여 이동할 수 없습니다.

수많은 추가 기능은 **도표 옵션**을 클릭하여 사용 가능합니다. 특히, 슬라이더의 위치를 명시적으로 설정하고 고정 참조선을 추가하고 도표 보기를 연속형 곡선에서 히스토그램으로 변경할 수 있습니다. 자세한 정보는 344 페이지의 『차트 옵션』 주제를 참조하십시오. 도표를 복사하거나 내보내려면 마우스 오른쪽 단추로 도표를 클릭하십시오.

연속형 대상의 누적 분포 함수. 이 도표에는 두 개의 동일한 이동 가능한 수직 참조선 및 확률 밀도 함수 도표에 대해 설명하는 연관된 표가 있습니다. 슬라이더 제어 및 표는 다중 반복이 있는 경우에 확률 밀도 함수와 동일하게 작동합니다. 각 반복에 속한 밀도함수를 식별하기 위해 사용되는 것과 동일한 색상이 분포 함수에 사용됩니다.

또한 이 도표는 슬라이더의 위치를 명시적으로 설정하고 고정 참조선을 추가하고 누적 분포 함수가 증가 함수(기본값) 또는 감소 함수로 표시되는지 여부를 지정하는 데 사용할 수 있는 도표 옵션 대화 상자에 대한 액세스를 제공합니다. 자세한 정보는 344 페이지의 『차트 옵션』 주제를 참조하십시오. 도표를 복사하거나 내보내거나 편집하려면 마우스 오른쪽 단추로 도표를 클릭하십시오. **편집**을 선택하면 Float 그래프보드 편집기 창에 도표가 열립니다.

범주형 대상의 예측 범주 값 도표. 범주형 대상 필드의 경우, 막대형 차트가 예측값을 표시합니다. 예측값은 각 범주에 해당될 것으로 예측되는 목표 필드의 퍼센트로 표시됩니다. 민감도 분석 반복계산이 있는 범주형 대상 필드의 경우,예측 목표 범주의 결과가 모든 반복의 결과를 포함하는 수평배열 막대도표로 표시됩니다. 도표는 시뮬레이션 평가 노드 대화 상자의 밀도 함수 보기의 **민감도 분석의 집단화** 영역에서 선택한 옵션에 따라 범주 또는 반복에 의해 수평배열됩니다. 도표를 복사하거나 내보내거나 편집하려면 마우스 오른쪽 단추로 도표를 클릭하십시오. **편집**을 선택하면 Float 그래프보드 편집기 창에 도표가 열립니다.

범주형 대상의 예측 범주 확률 도표. 범주형 대상 필드의 경우, 히스토그램은 대상의 각 범주에 대한 예측 확률의 분포를 표시합니다. 민감도 분석 반복계산이 있는 범주형 대상 필드의 경우, 시뮬레이션 평가 노드 대화 상자의 밀도 함수 보기의 **민감도 분석의 집단화** 영역에서 선택한 옵션에 따라 범주 또는 반복 기준으로 히스토그램이 표시됩니다. 이 히스토그램은 범주 기준으로 집단화되고 반복 레이블을 포함하는 드롭 다운 목록을 사용하면 표시할 반복을 선택할 수 있습니다. 또한 마우스 오른쪽 단추로 도표를 클릭하고 **반복** 하위 메뉴에서 반복을 선택함으로써 표시할 반복을 선택할 수 있습니다. 이 히스토그램은 반복 기준으로 집단화되고 범주 이름을 포함하는 드롭 다운 목록을 사용하면 표시할 범주를 선택할 수 있습니다. 또한 마우스 오른쪽 단추로 도표를 클릭하고 **범주** 하위 메뉴에서 범주를 선택함으로써 표시할 범주를 선택할 수 있습니다.

이 도표는 모델의 서브세트에서만 사용 가능하며 모델 너깃에서 모든 집단 확률을 생성하기 위한 옵션을 선택해야 합니다. 예를 들어, 로지스틱 모델 너깃에서 **모든 확률 추가**를 선택해야 합니다. 다음 모델 너깃은 이 옵션을 지원합니다.

- 로지스틱, SVM, Bayes, 신경망 및 KNN
- 로지스틱 회귀분석, 의사결정 트리 및 Naive Bayes에 대한 DB2/ISW In-Database 마이닝 모형

기본적으로 모든 집단 확률을 생성하기 위한 옵션은 이러한 모델 너깃에서 선택되지 않습니다.

토네이도 도표. 토네이도 도표는 각 지정된 입력에 대한 목표 필드의 민감도를 표시하는 막대형 차트입니다. 민감도는 목표와 각 입력의 상관관계에 의해 측정됩니다. 도표의 제목에는 목표 필드의 이름이 포함됩니다. 도표의 각 막대는 목표 필드 및 입력 필드 사이의 상관관계를 나타냅니다. 도표에 포함되는 시뮬레이션한 입력은 시뮬레이션 평가 노드 대화 상자의 필드 보기의 **상관관계 토네이도의 입력 필드** 필드에서 선택된 입력입니다. 각 막대는 상관관계 값으로 레이블이 붙여집니다. 막대는 가장 큰 값에서 가장 작은 값까지 상관계수의 절대값에 의해 순서가 지정됩니다. 반복이 있는 경우에는 각 반복에 대해 별도의 차트가 생성됩니다. 각 도표에는 반복의 이름을 포함하는 부제목이 있습니다.

분포 표. 이 표에는 목표 필드의 값이 포함되며 해당 값 아래에 관측값의 지정된 퍼센트가 포함됩니다. 표에는 시뮬레이션 평가 노드 대화 상자의 출력 보기에서 지정된 각 백분위수 값에 대한 행이 포함됩니다. 백분위수 값은 사분위수, 동등하게 간격이 지정된 수가 다른 백분위수, 개별적으로 지정된 백분위수 등이 될 수 있습니다. 분포 표에는 각 반복에 대한 열이 포함됩니다.

정보. 이 절에서는 평가에 사용되는 필드 및 레코드의 전체 요약을 제공합니다. 또한 각 반복으로 구분된 입력 필드 및 레코드 빈도를 표시합니다.

차트 옵션

도표 옵션 대화 상자에서는 시뮬레이션에서 생성된 확률 밀도 함수 및 누적 분포 함수의 활성화 도표 표시를 사용자 정의할 수 있습니다.

보기. 보기 드롭 다운 목록은 확률 밀도 함수 차트에만 적용됩니다. 이를 사용하여 연속형 곡선에서 히스토그램으로 차트 보기를 토글할 수 있습니다. 이 기능은 다중 반복의 다중 밀도함수가 동일한 차트에 표시되는 경우에는 사용할 수 없습니다. 다중 밀도함수가 있으면 다중 밀도함수를 연속형 곡선으로만 볼 수 있습니다.

순서. 순서 드롭 다운 목록은 누적 분포 함수 차트에만 적용됩니다. 이는 누적 분포 함수가 오름차순 함수(기본값) 또는 내림차순 함수로 표시되는지 지정합니다. 내림차순 함수로 표시되면 수평축 변수의 지정된 포인트에서 함수의 값이 해당 포인트의 오른쪽에 목표 필드가 놓이는 확률이 됩니다.

슬라이더 위치. 상한 텍스트 필드에는 오른쪽 슬라이딩 참조선의 현재 위치가 포함됩니다. **하한** 텍스트 필드에는 왼쪽 슬라이딩 참조선의 현재 위치가 포함됩니다. **상한** 및 **하한** 텍스트 필드에 값을 입력하여 슬라이더의 위치를 명시적으로 설정할 수 있습니다. **하한** 텍스트 필드의 값은 반드시 **상한** 텍스트 필드의 값 미만이어야 합니다. **-무한**을 선택하여 왼쪽 참조선을 제거하면 위치를 효과적으로 음의 무한대로 설정할 수 있습니다. 이 조치를 수행하면 **하한** 텍스트 필드를 사용할 수 없습니다. **-무한**을 선택하여 오른쪽 참조선을 제거하면 위치를 효과적으로 무한대로 설정할 수 있습니다. 이 조치를 수행하면 **상한** 텍스트 필드를 사용할 수 없습니다. 두 참조선을 모두 제거할 수는 없습니다. **-무한**을 선택하면 **무한대** 확인 상자를 선택할 수 없으며 반대의 경우도 마찬가지입니다.

참조선. 확률 밀도 함수 및 누적 분포 함수에 다양한 고정된 수직 참조선을 추가할 수 있습니다.

- **평균.** 목표 필드의 평균에 참조선을 추가할 수 있습니다.
- **중앙값.** 목표 필드의 중앙값에 참조선을 추가할 수 있습니다.
- **표준 편차.** 목표 필드의 평균 값에서부터 지정된 수의 표준편차 더하기 및 빼기에 참조선을 추가할 수 있습니다. 인접한 텍스트 필드에서 사용할 표준편차의 수를 입력할 수 있습니다. 최소 표준편차 수는 1이며 최대 수는 10입니다. 표준편차의 수는 처음에 3으로 설정됩니다.
- **백분위수. 아래쪽** 및 **위쪽** 텍스트 필드에 값을 입력하여 목표 필드에 대한 분포의 한 개 또는 두 개의 백분위수 값에 참조선을 추가할 수 있습니다. 예를 들어, **위쪽** 텍스트 필드의 95 값은 95번째 백분위수를 나타내며 관측값의 95% 아래에 해당되는 값입니다. 이와 유사하게 **아래쪽** 텍스트 필드의 5 값은 다섯 번째 백분위수를 나타내며 관측값의 5% 아래에 해당되는 값입니다. **아래쪽** 텍스트 필드의 경우, 최소 백분위수 값은 0이며 최대수는 49입니다. **위쪽** 텍스트 필드의 경우, 최소 백분위수 값은 50이며 최대수는 100입니다.
- **사용자 정의 위치.** 수평축 변수와 함께 지정된 값에 참조선을 추가할 수 있습니다. 눈금에서 항목을 삭제하여 사용자 정의 참조선을 제거할 수 있습니다.

확인을 클릭하면 도표 옵션 대화 상자에서 선택된 옵션을 반영하기 위해 슬라이더, 슬라이더 위의 레이블, 참조선 및 도표 아래의 표가 업데이트됩니다. 변경하지 않고 대화 상자를 닫으려면 **취소**를 클릭하십시오. 참조선은 도표 옵션 대화 상자에서 연관된 선택을 선택 취소하고 **확인**을 클릭하여 제거할 수 있습니다.

참고: 민감도 분석 반복계산의 결과로 인해 다중 밀도함수 또는 분포 함수가 단일 도표에 표시되는 경우, 사용자 정의 선이 아니라 참조선이 각 함수에 별도로 적용됩니다. 첫 번째 반복에 대한 참조선만 표시됩니다. 참조선 레이블에는 반복 레이블이 포함됩니다. 반복 레이블은 업스트림(일반적으로 시뮬레이션 생성 노드)에서 파생됩니다. 반복 레이블을 사용할 수 없으면 반복 값이 대신 사용됩니다. **평균**, **중앙값**, **표준 편차** 및 **백분위수** 옵션은 다중 반복이 있는 누적 분포 함수에는 사용할 수 없습니다.

IBM SPSS Statistics 헬퍼 애플리케이션

컴퓨터에 호환 가능한 IBM SPSS Statistics 버전이 설치되고 라이센스 부여되어 있는 경우Statistics 변환, Statistics 모델, Statistics 출력 또는 Statistics 내보내기 노드를 사용하여 IBM SPSS Statistics 기능을 통해 데이터를 처리하도록 IBM SPSS Modeler를 구성할 수 있습니다.

현재 IBM SPSS Modeler 버전과의 제품 호환성에 대한 정보는 회사 지원 사이트(*http://www.ibm.com/support*)를 참조하십시오.

IBM SPSS Modeler가 IBM SPSS Statistics 및 기타 애플리케이션과 함께 작동하도록 구성하려면 다음을 선택하십시오.

도구 > 옵션 > 헬퍼 애플리케이션

IBM SPSS Statistics Interactive. 통계량 내보내기 노드에 의해 생성된 데이터 파일에서 직접 IBM SPSS Statistics를 실행하여 사용할 명령의 전체 경로 및 이름(예: *C:\Program Files\IBM\SPSS\Statistics\<nn>\stats.exe*)을 입력하십시오. 자세한 정보는 385 페이지의 『통계량 내보내기 노드』의 내용을 참조하십시오.

연결. IBM SPSS Statistics 서버가 IBM SPSS Modeler Server와 동일한 호스트에 있으면 두 애플리케이션 간 연결을 사용으로 설정하여 분석 중에 서버에 데이터를 남겨 효율을 높일 수 있습니다. **서버**를 선택하여 아래의 **포트** 옵션을 사용으로 설정하십시오. 기본 설정은 **로컬**입니다.

포트. IBM SPSS Statistics 서버의 서버 포트를 지정하십시오.

IBM SPSS Statistics 위치 유틸리티. IBM SPSS Modeler가 통계량 변환, 통계량 모델 및 통계량 출력 노드를 사용할 수 있게 하려면 스트림이 실행되는 컴퓨터에 IBM SPSS Statistics의 사본이 설치되고 라이센스 부여되어 있어야 합니다.

- 로컬(독립형) 모드에서 IBM SPSS Modeler를 실행 중인 경우 IBM SPSS Statistics의 라이센스 부여된 사본이 로컬 컴퓨터에 있어야 합니다. 이 단추를 클릭하여 라이센스 부여에 사용할 로컬 IBM SPSS Statistics 설치의 위치를 지정하십시오.
- 또한 원격 IBM SPSS Modeler Server에 대해 분산 모드에서 실행 중인 경우에는 IBM SPSS Modeler Server 호스트에서 유틸리티를 실행하여 *statistics.ini* 파일을 작성하여 IBM SPSS Statistics에 IBM SPSS Modeler Server의 설치 경로를 표시해야 합니다. 이를 수행하려면 명령 프롬프트에서 IBM SPSS Modeler Server *bin* 디렉토리로 변경한 후 Windows의 경우 다음을 실행하십시오.

  ```
  statisticsutility -location=<IBM SPSS Statistics_installation_path>/
  ```

  UNIX의 경우 다음을 실행하십시오.

  ```
  ./statisticsutility -location=<IBM SPSS Statistics_installation_path>/bin
  ```

로컬 시스템에 IBM SPSS Statistics의 라이센스 부여된 사본이 없는 경우에는 여전히 IBM SPSS Statistics 서버에 대해 통계 파일 노드를 실행할 수 있지만 다른 IBM SPSS Statistics 노드를 실행하면 오류 메시지가 표시됩니다.

설명

IBM SPSS Statistics 프로시저 노드 실행에 문제가 있으면 다음과 같은 팁을 고려하십시오.

- IBM SPSS Modeler에서 사용된 필드 이름이 8자(IBM SPSS Statistics 12.0 이전 버전의 경우) 또는 64자(IBM SPSS Statistics 12.0 이상 버전의 경우)보다 길거나 유효하지 않은 문자를 포함하는 경우에는 해당 필드 이름을 IBM SPSS Statistics로 읽어오기 전에 해당 필드 이름을 바꾸거나 잘라야 합니다. 자세한 정보는 387 페이지의 『IBM SPSS Statistics에 대한 필드 이름 변경 또는 필터링』의 내용을 참조하십시오.
- IBM SPSS Statistics가 IBM SPSS Modeler 다음에 설치된 경우에는 위에서 설명한 대로 IBM SPSS Statistics 위치를 지정해야 할 수 있습니다.

제 7 장 내보내기 노드

내보내기 노드의 개요

내보내기 노드는 다양한 형식으로 데이터를 내보내 다른 소프트웨어 도구와 인터페이스로 접속하는 메커니즘을 제공합니다.

사용 가능한 내보내기 노드는 다음과 같습니다.

데이터베이스 내보내기 노드는 데이터를 ODBC 준수 관계형 데이터 소스에 기록합니다. ODBC 데이터 소스에 쓰기 위해 데이터 소스가 존재하고 사용자에게 쓰기 권한이 있어야 합니다.

플랫 파일 내보내기 노드는 데이터를 구분된 텍스트 파일로 출력합니다. 다른 분석 또는 스프레드시트 소프트웨어가 읽을 수 있는 데이터 내보내기에 유용합니다.

통계량 내보내기 노드는 IBM SPSS Statistics *.sav* 또는 *.zsav* 형식으로 데이터를 출력합니다. *.sav* 또는 *.zsav* 파일은 IBM SPSS Statistics Base 및 기타 제품에서 읽을 수 있습니다. 이것은 또한 IBM SPSS Modeler의 캐시 파일에 사용하는 형식입니다.

Data Collection 내보내기 노드는 Data Collection 시장 조사 소프트웨어에서 사용하는 형식으로 데이터를 출력합니다. 이 노드를 사용하려면 Data Collection 데이터 라이브러리가 설치되어야 합니다.

IBM Cognos BI 내보내기 노드는 데이터를 Cognos BI 데이터베이스가 읽을 수 있는 형식으로 내보냅니다.

IBM Cognos TM1 내보내기 노드는 Cognos TM1 데이터베이스가 읽을 수 있는 형식으로 데이터를 내보냅니다.

SAS 내보내기 노드는 SAS 또는 SAS 호환 가능한 소프트웨어 패키지로 읽어들이기 위해 데이터를 SAS 형식으로 출력합니다. SAS for Windows/OS2, SAS for UNIX 또는 SAS 버전 7/8의 세 가지 SAS 파일 형식이 사용 가능합니다.

Excel 내보내기 노드는 데이터를 Microsoft Excel .xlsx 파일 형식으로 출력합니다. (선택사항)노드가 실행될 때 Excel을 자동으로 시작하고 내보내진 파일을 열도록 선택할 수 있습니다.

XML 내보내기 노드는 데이터를 XML 형식의 파일로 출력합니다. 선택적으로 XML 소스 노드를 작성하여 내보내진 데이터를 다시 스트림으로 읽을 수 있습니다.

데이터베이스 내보내기 노드

데이터베이스 노드를 사용하여 ODBC 준수 관계형 데이터 소스에 데이터를 쓸 수 있습니다. 여기에 대해서는 데이터베이스 소스 노드에 대한 설명을 참조하십시오. 자세한 정보는 18 페이지의 『데이터베이스 소스 노드』의 내용을 참조하십시오.

데이터베이스에 데이터를 쓰려면 다음과 같은 일반 단계를 사용하십시오.

1. ODBC 드라이버를 설치하고 원하는 데이터베이스에 데이터 소스를 구성하십시오.
2. 데이터베이스 노드 내보내기 탭에서 쓸 데이터 소스 및 테이블을 지정하십시오. 새 테이블을 작성하거나 데이터를 기존 테이블에 삽입할 수 있습니다.
3. 필요에 따라 추가 옵션을 지정하십시오.

이러한 단계에 대해서는 다음 몇 가지 주제에서 더 자세히 설명합니다.

데이터베이스 노드 내보내기 탭

참고: 내보낼 수 있는 일부 데이터베이스는 길이가 30자를 초과하는 열 이름을 테이블에서 지원하지 않을 수 있습니다. 테이블에 올바르지 않은 열 이름이 있다는 오류 메시지가 표시되면 30자 미만으로 해당 이름의 크기를 줄이십시오.

데이터 소스. 선택된 데이터 소스를 표시합니다. 이름을 입력하거나 드롭 다운 목록에서 이름을 선택하십시오. 목록에 원하는 데이터베이스가 표시되지 않으면 **새 데이터베이스 연결 추가**를 선택하고 데이터베이스 연결 대화 상자에서 데이터베이스를 찾으십시오. 자세한 정보는 20 페이지의 『데이터베이스 연결 추가』의 내용을 참조하십시오.

테이블 이름. 데이터를 전송할 테이블의 이름을 입력하십시오. **테이블에 삽입** 옵션을 선택하는 경우에는 **선택** 단추를 클릭하여 데이터베이스에서 기존 테이블을 선택할 수 있습니다.

테이블 작성. 새 데이터베이스 테이블을 작성하거나 기존 데이터베이스 테이블을 겹쳐쓰려면 이 옵션을 선택하십시오.

테이블에 삽입. 기존 데이터베이스 테이블에서 새 행으로 데이터를 삽입하려면 이 옵션을 선택하십시오.

테이블 병합. (사용 가능한 경우) 선택된 데이터베이스 열을 해당 소스 데이터 필드의 값으로 업데이트하려면 이 옵션을 선택하십시오. 이 옵션을 선택하면 소스 데이터 필드를 데이터베이스 열에 맵핑할 수 있는 대화 상자를 표시하는 **병합** 단추를 사용할 수 있습니다.

기존 테이블 삭제. 새 테이블 작성 시 동일한 이름의 기존 테이블을 삭제하려면 이 옵션을 선택하십시오.

기존 행 삭제. 테이블에 삽입 시 내보내기 전에 테이블에서 기존 행을 삭제하려면 이 옵션을 선택하십시오.

참고: 위 두 옵션 중 하나가 선택되면 노드를 실행할 때 **겹쳐쓰기 경고** 메시지가 수신됩니다. 경고를 표시하지 않으려면 사용자 옵션 대화 상자의 알림 탭에서 **노드가 데이터베이스 테이블을 겹쳐쓸 때 경고**를 선택 취소하십시오.

기본 문자열 크기. 업스트림 유형 노드에서 유형 없음으로 표시한 필드는 데이터베이스에 문자열 필드로 작성됩니다. 유형 없는 필드에 사용할 문자열의 크기를 지정하십시오.

스키마를 클릭하여 다양한 내보내기 옵션을 설정(이 기능을 지원하는 데이터베이스의 경우)하고 필드에 대해 SQL 데이터 유형을 설정하고 데이터베이스 인덱싱을 위해 기본 키를 지정할 수 있는 대화 상자를 여십시오. 자세한 정보는 353 페이지의 『데이터베이스 내보내기 스키마 옵션』의 내용을 참조하십시오.

인덱스를 클릭하여 데이터베이스 성능을 향상시키기 위해 내보낸 테이블을 인덱싱하는 데 필요한 옵션을 지정하십시오. 자세한 정보는 355 페이지의 『데이터베이스 내보내기 인덱스 옵션』의 내용을 참조하십시오.

고급을 클릭하여 벌크 로드 및 데이터베이스 커미트 옵션을 지정하십시오. 자세한 정보는 357 페이지의 『데이터베이스 내보내기 고급 옵션』의 내용을 참조하십시오.

테이블 및 열 이름 따옴표로 묶기. CREATE TABLE문을 데이터베이스에 전송할 때 사용되는 옵션을 선택하십시오. 공백 또는 비표준 문자가 포함된 테이블 또는 열은 따옴표로 묶어야 합니다.

- **필요에 따라.** IBM SPSS Modeler가 개별적으로 따옴표가 필요한 시기를 자동으로 판별할 수 있게 하려면 선택하십시오.
- **항상.** 테이블 및 열 이름을 항상 따옴표로 묶으려면 선택하십시오.
- **사용 안 함.** 따옴표를 사용하지 않으려면 선택하십시오.

현재 데이터의 입력 노드 생성. 지정된 데이터 소스 및 테이블로 내보낸 대로 데이터에 대한 데이터베이스 소스 노드를 생성하려면 선택하십시오. 실행 시 이 노드는 스트림 캔버스에 추가됩니다.

데이터베이스 내보내기 병합 옵션

이 대화 상자에서는 소스 데이터의 필드를 목표 데이터베이스 테이블의 열에 맵핑할 수 있습니다. 소스 데이터 필드가 데이터베이스 열에 맵핑되는 경우에는 스트림이 실행될 때 열 값이 소스 데이터 값으로 바뀝니다. 맵핑되지 않은 소스 필드는 데이터베이스에서 변경되지 않고 유지됩니다.

맵 필드. 소스 데이터 필드와 데이터베이스 열 사이의 맵핑을 지정하는 위치입니다. 데이터베이스의 열과 동일한 이름을 가진 소스 데이터 필드는 자동으로 맵핑됩니다.

- **맵핑.** 단추 왼쪽의 필드 목록에서 선택된 소스 데이터 필드를 오른쪽 목록에서 선택된 데이터베이스 열에 맵핑합니다. 한 번에 둘 이상의 필드를 맵핑할 수 있지만 두 목록에서 선택된 항목 수는 동일해야 합니다.
- **맵핑 해제.** 하나 이상의 선택된 데이터베이스 열에 대한 맵핑을 제거합니다. 이 단추는 대화 상자 오른쪽의 테이블에서 필드 또는 데이터베이스 열을 선택하면 활성화됩니다.
- **추가.** 단추 왼쪽의 필드 목록에서 선택된 하나 이상의 소스 데이터 필드를 맵핑 준비가 된 오른쪽의 목록에 추가합니다. 이 단추는 왼쪽의 목록에서 필드를 선택했을 때 해당 이름을 가진 필드가 오른쪽의 목록에 없는 경우 활성화됩니다. 이 단추를 클릭하면 선택된 필드가 동일한 이름의 새 데이터베이스 열에 맵핑됩니다. **<NEW>**라는 단어가 데이터베이스 열 이름 뒤에 표시되어 이 필드가 새 필드임을 표시합니다.

행 병합. 키 필드(예: *트랜잭션 ID*)를 사용하여 키 필드에서 동일한 값을 가진 레코드를 병합합니다. 이는 데이터베이스 "일치 결합"과 동등합니다. 키 값은 기본 키의 값과 동일해야 합니다. 즉, 고유해야 하며 널값을 포함할 수 없습니다.

- **가능한 키.** 모든 입력 데이터 소스에서 발견된 모든 필드를 나열합니다. 이 목록에서 하나 이상의 필드를 선택한 후 화살표 단추를 사용하여 레코드 병합을 위해 키 필드로 추가하십시오. 해당 맵핑된 데이터베이스 열을 가진 맵 필드를 모두 키로 사용할 수 있습니다(이름 뒤에 **<NEW>**가 표시된 새 데이터베이스 열로 추가된 필드는 사용할 수 없음).
- **병합을 위한 키.** 키 필드의 값을 기반으로 모든 입력 데이터 소스의 레코드를 병합하는 데 사용되는 모든 필드를 나열합니다. 목록에서 키를 제거하려면 하나의 키를 선택한 후 화살표 단추를 사용하여 가능한 키 목록에 리턴하십시오. 둘 이상의 키 필드가 선택되면 아래의 옵션을 사용할 수 있습니다.
- **데이터베이스에 있는 레코드만 포함.** 부분 결합을 수행합니다. 레코드가 데이터베이스 및 스트림에 있는 경우에는 맵핑된 필드가 업데이트됩니다.
- **데이터베이스에 레코드 추가.** 외부 결합을 수행합니다. 스트림의 모든 레코드가 병합되거나(동일한 레코드가 데이터베이스에 있는 경우) 추가됩니다(레코드가 아직 데이터베이스에 없는 경우).

새 데이터베이스 열에 소스 데이터 필드를 맵핑하려면 다음을 수행하십시오.

1. 왼쪽 목록의 **맵 필드** 아래에서 소스 필드 이름을 클릭하십시오.
2. **추가** 단추를 클릭하여 맵핑을 완료하십시오.

기존 데이터베이스 열에 소스 데이터 필드를 맵핑하려면 다음을 수행하십시오.

1. 왼쪽 목록의 **맵 필드** 아래에서 소스 필드 이름을 클릭하십시오.
2. 오른쪽의 **데이터베이스 열** 아래에서 열 이름을 클릭하십시오.
3. **맵** 단추를 클릭하여 맵핑을 완료하십시오.

맵핑을 제거하려면 다음을 수행하십시오.

1. 오른쪽 목록의 필드 아래에서 맵핑을 제거할 필드의 이름을 클릭하십시오.
2. **맵핑 해제** 단추를 클릭하십시오.

목록에서 필드를 선택 취소하려면 다음을 수행하십시오.

CTRL 키를 누른 상태로 필드 이름을 클릭하십시오.

데이터베이스 내보내기 스키마 옵션

데이터베이스 내보내기 스키마 대화 상자에서는 데이터베이스 내보내기를 위한 옵션을 설정하고(이 옵션을 지원하는 데이터베이스의 경우) 필드에 대한 SQL 데이터 유형을 설정하고 기본 키인 필드를 지정하고 내보낼 때 생성되는 CREATE TABLE문을 사용자 정의할 수 있습니다.

이 대화 상자에는 여러 파트가 있습니다.

- 맨 위의 섹션(표시된 경우)에는 이 옵션을 지원하는 데이터베이스에 내보내기 위한 옵션이 포함되어 있습니다. 해당 데이터베이스에 연결되지 않은 경우에는 이 섹션이 표시되지 않습니다.
- 가운데의 텍스트 필드에는 기본적으로 다음 형식을 따르는 CREATE TABLE 명령을 생성하는 데 사용되는 템플리트가 표시됩니다.

  ```
  CREATE TABLE <table-name> <(table columns)>
  ```

- 아래쪽의 테이블에서는 각 필드에 대한 SQL 데이터 유형을 지정하고 아래에 설명된 대로 기본 키인 필드를 표시할 수 있습니다. 이 대화 상자는 테이블에서 지정하는 사항에 따라 <table-name> 및 <(table columns)> 매개변수의 값을 자동으로 생성합니다.

데이터베이스 내보내기 옵션 설정

이 섹션이 표시되는 경우에는 데이터베이스에 내보내는 데 필요한 다수의 설정을 지정할 수 있습니다. 이 기능을 지원하는 데이터베이스 유형은 다음과 같습니다.

- IBM InfoSphere Warehouse. 자세한 정보는 354 페이지의 『IBM DB2 InfoSphere Warehouse에 대한 옵션』의 내용을 참조하십시오.
- SQL Server Enterprise 및 Developer Edition. 자세한 정보는 354 페이지의 『SQL Server에 대한 옵션』의 내용을 참조하십시오.
- Oracle Enterprise 또는 Personal Edition. 자세한 정보는 354 페이지의 『Oracle에 대한 옵션』의 내용을 참조하십시오.

CREATE TABLE문 사용자 정의

이 대화 상자의 텍스트 필드 부분을 사용하여 CREATE TABLE문에 데이터베이스별 옵션을 추가할 수 있습니다.

1. **CREATE TABLE 명령 사용자 정의** 선택란을 선택하여 텍스트 창을 활성화하십시오.
2. 명령문에 데이터베이스별 옵션을 추가하십시오. 텍스트 <table-name> 및 (<table-columns>) 매개변수는 IBM SPSS Modeler에 의해 실제 테이블 이름 및 열 정의에 대해 대체되므로 이들 매개변수는 보존해야 합니다.

SQL 데이터 유형 설정

기본적으로 IBM SPSS Modeler를 사용하면 데이터베이스 서버가 SQL 데이터 유형을 자동으로 지정할 수 있습니다. 필드에 대한 자동 유형을 대체하려면 해당 필드에 해당하는 행을 찾은 후 스키마 테이블의 *유형* 열에 있는 드롭 다운 목록에서 원하는 유형을 선택하십시오. Shift+클릭을 사용하여 둘 이상의 행을 선택할 수 있습니다.

길이, 정밀도 또는 척도 인수(BINARY, VARBINARY, CHAR, VARCHAR, NUMERIC 및 NUMBER)를 사용하는 유형의 경우 데이터베이스 서버에 자동 길이 지정을 허용하는 대신 길이를 지정해야 합니다. 예를 들어, 길이에 대해 상당한 값(예: VARCHAR(25))을 지정하면 사용자가 의도한 경우 IBM SPSS Modeler에서의 저장 유형이 겹쳐써집니다. 자동 지정을 대체하려면 유형 드롭 다운 목록에서 **지정**을 선택하고 유형 정의를 원하는 SQL 유형 정의 명령문으로 바꾸십시오.

이를 수행하는 가장 쉬운 방법은 먼저 원하는 유형 정의에 가장 근접한 유형을 선택한 후 **지정**을 선택하여 해당 정의를 편집하는 것입니다. 예를 들어, SQL 데이터 유형을 VARCHAR(25)로 설정하려면 먼저 유형 드롭 다운 목록에서 유형을 **VARCHAR(길이)**로 설정한 후 **지정**을 선택하고 텍스트 길이를 값 25로 바꾸십시오.

기본 키

내보낸 테이블의 열 중 하나 이상이 모든 행에 고유 값 또는 값 조합을 가져야 하는 경우에는 적용되는 각 필드에 대해 **기본 키** 선택란을 선택하여 이를 표시할 수 있습니다. 대부분의 데이터베이스는 기본 키 제한조건을 무효화하는 방식으로 테이블을 수정할 수 없게 하며 이 제한을 적용하기 위해 기본 키에 대한 인덱스를 자동으로 작성합니다. (선택적으로 인덱스 대화 상자에서 기타 필드에 대한 인덱스를 작성할 수 있습니다. 자세한 정보는 355 페이지의 『데이터베이스 내보내기 인덱스 옵션』의 내용을 참조하십시오.)

IBM DB2 InfoSphere Warehouse에 대한 옵션

테이블스페이스. 내보내기에 사용할 테이블스페이스입니다. 데이터베이스 관리자는 테이블스페이스를 파티션되도록 작성하거나 구성할 수 있습니다. 기본 테이블스페이스 대신 내보내기에 사용할 이 테이블스페이스 중 하나를 선택하는 것이 좋습니다.

필드별 데이터 파티션. 파티셔닝에 사용할 입력 필드를 지정합니다.

압축 사용. 선택된 경우 압축을 사용하여 내보낼 테이블을 작성합니다(예: SQL의 CREATE TABLE MYTABLE(...) COMPRESS YES;와 동등).

SQL Server에 대한 옵션

압축 사용. 선택된 경우 압축을 사용하여 내보낼 테이블을 작성합니다.

압축. 압축의 수준을 선택하십시오.

- **행.** 행 수준 압축을 사용으로 설정합니다(예: SQL의 CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = ROW);와 동등).
- **페이지.** 페이지 수준 압축을 사용으로 설정합니다(예: SQL의 CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = PAGE);).

Oracle에 대한 옵션

Oracle 설정 - 기본 옵션

압축 사용. 선택된 경우 압축을 사용하여 내보낼 테이블을 작성합니다.

압축. 압축의 수준을 선택하십시오.

- **기본값.** 기본 압축을 사용으로 설정합니다(예: SQL의 CREATE TABLE MYTABLE(...) COMPRESS;). 이 케이스에서 이는 **기본** 옵션과 동일한 효과를 가집니다.
- **기본.** 기본 압축을 사용으로 설정합니다(예: SQL의 CREATE TABLE MYTABLE(...) COMPRESS BASIC;).

Oracle 설정 - 고급 옵션

압축 사용. 선택된 경우 압축을 사용하여 내보낼 테이블을 작성합니다.

압축. 압축의 수준을 선택하십시오.

- **기본값.** 기본 압축을 사용으로 설정합니다(예: SQL의 CREATE TABLE MYTABLE(...) COMPRESS;). 이 케이스에서 이는 **기본** 옵션과 동일한 효과를 가집니다.
- **기본.** 기본 압축을 사용으로 설정합니다(예: SQL의 CREATE TABLE MYTABLE(...) COMPRESS BASIC;).
- **OLTP.** OLTP 압축을 사용으로 설정합니다(예: SQL의 CREATE TABLE MYTABLE(...)COMPRESS FOR OLTP;).
- **쿼리 낮음/높음.** (Exadata 서버 전용) 쿼리에 대해 HCC(Hybrid Columnar Compression)를 사용으로 설정합니다(예: SQL의 CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY LOW; 또는 CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY HIGH;). 쿼리에 대한 압축은 데이터 웨어하우징 환경에서 유용합니다. HIGH는 LOW보다 높은 압축 비율을 제공합니다.
- **아카이브 낮음/높음.** (Exadata 서버 전용) 아카이브에 대해 HCC(Hybrid Columnar Compression)를 사용으로 설정합니다(예: SQL의 CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE LOW; 또는 CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE HIGH;). 아카이브에 대한 압축은 장기간 저장될 데이터를 압축하는 경우에 유용합니다. HIGH는 LOW보다 높은 압축 비율을 제공합니다.

데이터베이스 내보내기 인덱스 옵션

인덱스 대화 상자를 사용하면 IBM SPSS Modeler에서 내보낸 데이터베이스 테이블에서 인덱스를 작성할 수 있습니다. 필요에 따라 포함할 필드 세트를 지정하고 CREATE INDEX 명령을 사용자 정의할 수 있습니다.

이 대화 상자는 두 개의 파트로 구성됩니다.

- 위쪽 텍스트 필드에는 하나 이상의 CREATE INDEX 명령을 생성하는 데 사용할 수 있는 템플리트가 표시되며 기본적으로 형식은 다음과 같습니다.

  ```
  CREATE INDEX <index-name> ON <table-name>
  ```

- 대화 상자 아래쪽의 테이블에서는 작성할 각각의 인덱스에 대한 사양을 추가할 수 있습니다. 각각의 인덱스에 대해 포함할 필드 또는 열 및 인덱스 이름을 지정하십시오. 이 대화 상자에서는 자동으로 <index-name> 및 <table-name> 매개변수의 값을 적절하게 생성합니다.

  예를 들어, *empid* 및 *deptid* 필드의 단일 인덱스에 대해 생성된 SQL의 모양은 다음과 같습니다.

  ```
  CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID)
  ```

여러 행을 추가하여 여러 인덱스를 작성할 수 있습니다. 각각의 행에 대해 별도의 CREATE INDEX 명령이 생성됩니다.

CREATE INDEX 명령 사용자 정의

선택적으로 모든 인덱스 또는 특정 인덱스에 대해 CREATE INDEX 명령을 사용자 정의할 수 있습니다. 이를 통해 필요에 따라 특정 데이터베이스 요구사항 또는 옵션을 수용하고 모든 인덱스 또는 특정 인덱스에 사용자 정의를 적용할 수 있는 유연성이 제공됩니다.

- 위쪽 대화 상자에서 **CREATE INDEX 명령 사용자 정의**를 선택하여 이후에 추가된 모든 인덱스에 사용되는 템플리트를 수정하십시오. 테이블에 이미 추가된 인덱스에는 변경사항이 자동으로 적용되지 않습니다.
- 테이블에서 하나 이상의 행을 선택한 후 대화 상자 위쪽의 **선택된 인덱스 업데이트**를 클릭하여 선택된 모든 행에 현재 사용자 정의를 적용하십시오.
- 각각의 행에서 **사용자 정의** 선택란을 선택하여 해당 인덱스에 대한 명령 템플리트만 수정하십시오.

<index-name> 및 <table-name> 매개변수의 값은 테이블 사양을 기반으로 대화 상자에 의해 자동으로 생성되며 직접 편집할 수 없습니다.

BITMAP 키워드. Oracle 데이터베이스를 사용하는 경우에는 다음과 같이 표준 인덱스 대신 비트맵 인덱스를 작성하도록 템플리트를 사용자 정의할 수 있습니다.

```
CREATE BITMAP INDEX <index-name> ON <table-name>
```

비트맵 인덱스는 고유 값 수가 적은 열을 인덱싱하는 경우에 유용할 수 있습니다. 결과 SQL의 모양은 다음과 같습니다.

```
CREATE BITMAP INDEX MYTABLE_IDX1 ON MYTABLE(COLOR)
```

UNIQUE 키워드. 대부분의 데이터베이스는 CREATE INDEX 명령에서 UNIQUE 키워드를 지원합니다. 이 키워드는 기본 테이블에 대한 기본 키 제한조건과 비슷한 고유성 제한조건을 적용합니다.

```
CREATE UNIQUE INDEX <index-name> ON <table-name>
```

실제로 기본 키로 지정된 필드의 경우 이 사양은 필요하지 않습니다. 대부분의 데이터베이스는 CREATE TABLE 명령에서 기본 키 필드로 지정된 필드에 대해 자동으로 인덱스를 작성하므로 이 필드에서 명시적으로 인덱스를 작성하지 않아도 됩니다. 자세한 정보는 353 페이지의 『데이터베이스 내보내기 스키마 옵션』의 내용을 참조하십시오.

FILLFACTOR 키워드. 인덱스에 대한 일부 실제 매개변수를 미세 조정할 수 있습니다. 예를 들어, SQL Server를 사용하면 테이블에 대해 향후 변경사항이 작성될 때 사용자가 유지보수 비용과 인덱스 *크기*(초기 작성 후)의 균형을 맞출 수 있습니다.

```
CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID) WITH FILLFACTOR=20
```

기타 설명

- 지정된 이름을 가진 인덱스가 이미 존재하는 경우 인덱스 작성은 실패합니다. 모든 실패는 초기에 경고로 처리되어 후속 인덱스가 작성된 다음 모든 인덱스가 시도된 후 메시지 로그에 오류로 다시 보고될 수 있게 합니다.
- 최상의 성능을 위해 데이터가 테이블에 로드된 후 인덱스를 작성해야 합니다. 인덱스는 하나 이상의 열을 포함하고 있어야 합니다.
- 노드를 실행하기 전에 메시지 로그에서 생성된 SQL을 미리 볼 수 있습니다.
- 데이터베이스에 작성된 임시 테이블의 경우(즉, 노드 캐싱이 사용으로 설정된 경우) 기본 키 및 인덱스를 지정하는 옵션을 사용할 수 없습니다. 하지만 시스템에서는 다운스트림 노드에서 데이터가 사용되는 방식에 따라 적절하게 임시 테이블에서 인덱스를 작성할 수 있습니다. 예를 들어, 캐싱된 데이터가 이후에 *DEPT* 열에 의해 결합되는 경우에는 이 열에서 캐싱된 테이블을 인덱싱하는 것이 합리적입니다.

인덱스 및 쿼리 최적화

일부 데이터베이스 관리 시스템에서는 데이터베이스 테이블이 작성되고 로드되고 인덱싱되고 난 후 최적화 프로그램이 새 테이블에서 쿼리 실행의 속도를 높이기 위해 인덱스를 이용하려면 먼저 추가적인 단계가 필요합니다. 예를 들어, Oracle에서 비용 기반 쿼리 최적화 프로그램은 쿼리 최적화에서 인덱스를 사용하려면 먼저 테이블을 분석해야 합니다. Oracle에 대한 내부 ODBC 특성 파일(사용자에게 표시되지 않음)에는 다음과 같이 이를 수행하는 옵션이 포함되어 있습니다.

```
# 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'
```

기본 키와 인덱스 중 어느 것이 정의되는지 여부에 관계없이 Oracle에서 테이블이 작성될 때마다 이 단계가 실행됩니다. 필요한 경우 추가적인 데이터베이스에 대한 ODBC 특성 파일을 비슷한 방식으로 사용자 정의할 수 있습니다. 지원 부서에 문의하여 지원을 받으십시오.

데이터베이스 내보내기 고급 옵션

데이터베이스 내보내기 노드 대화 상자에서 고급 단추를 클릭하면 데이터베이스에 결과 내보내기에 대한 기술 세부사항을 지정할 수 있는 새 대화 상자가 표시됩니다.

일괄 커미트 사용. 데이터베이스에 대한 행별 커미트를 끄려면 선택하십시오.

일괄처리 크기. 메모리로 커미트하기 전에 데이터베이스로 보낼 레코드 수를 지정합니다. 이 숫자를 낮추면 전송 속도는 느려지지만 데이터 무결성이 향상됩니다. 데이터베이스의 최적 성능을 위해 이 숫자를 미세 조정할 수 있습니다.

InfoSphere Warehouse 옵션. InfoSphere Warehouse 데이터베이스(IBM DB2 9.7 이상)에 연결되는 경우에만 표시됩니다. **업데이트를 로그하지 않음**을 사용하면 테이블을 작성하고 데이터를 삽입할 때 이벤트 로깅을 방지할 수 있습니다.

벌크 로드 사용. IBM SPSS Modeler에서 직접 데이터베이스에 데이터를 벌크 로드하는 방법을 지정합니다. 특정 시나리오에 적합한 벌크 로드 옵션을 선택하기 위해 일부 실험이 필요할 수 있습니다.

- **ODBC를 통해.** 일반적인 데이터베이스에 내보내기보다 효율적으로 다중 행 삽입을 실행하기 위해 ODBC API를 사용하려면 선택하십시오. 아래 옵션에서 행 방식 바인딩과 열 방식 바인딩 중에서 선택하십시오.
- **외부 로더를 통해.** 데이터베이스에 고유한 사용자 정의 벌크 로더 프로그램을 사용하려면 선택하십시오. 이 옵션을 선택하면 아래의 다양한 옵션이 활성화됩니다.

고급 ODBC 옵션. 이 옵션은 **ODBC를 통해**가 선택된 경우에만 사용할 수 있습니다. 이 기능은 일부 ODBC 드라이버에서 지원하지 않을 수 있습니다.

- **행 방식.** 데이터베이스에 데이터를 로드하기 위해 SQLBulkOperations 호출을 사용하려면 행 방식 바인딩을 선택하십시오. 행 방식 바인딩은 레코드별로 데이터를 삽입하는 매개변수화된 삽입을 사용하는 경우보다 일반적으로 속도가 향상됩니다.
- **열 방식.** 데이터베이스에 데이터를 로드하기 위해 열 방식 바인딩을 사용하려면 선택하십시오. 열 방식 바인딩은 매개변수화된 INSERT문에서 각각의 데이터베이스 열을 *N*개 값의 배열에 바인딩하여 성능을 향상시킵니다. INSERT문을 한 번 실행하면 *N*개의 행이 데이터베이스에 삽입됩니다. 이 방법은 성능을 상당히 향상시킬 수 있습니다.

외부 로더 옵션. 외부 로더를 통해가 지정되면 파일에 데이터 세트를 내보내고 해당 파일의 데이터를 데이터베이스에 로드하기 위해 사용자 정의 로더 프로그램을 지정 및 실행하는 데 필요한 다양한 옵션이 표시됩니다. IBM SPSS Modeler는 다수의 인기 있는 데이터베이스 시스템에 대한 외부 로더와 인터페이스로 접속할 수 있습니다. 여러 스크립트가 소프트웨어와 함께 포함되었으며 *scripts* 서브디렉토리 아래의 기술 문서와 함께 사용 가능합니다. 이 기능을 사용하려면 Python 2.7이 IBM SPSS Modeler 또는 IBM SPSS Modeler Server와 동일한 시스템에 설치되어 있어야 하며 python_exe_path 매개변수가 *options.cfg* 파일에서 설정되어 있어야 합니다. 자세한 정보는 359 페이지의 『벌크 로더 프로그래밍』의 내용을 참조하십시오.

- **구분자 사용.** 내보낸 파일에서 사용해야 하는 구분 문자를 지정합니다. 탭으로 구분하려면 **탭**을 선택하고 공백으로 구분하려면 **공백**을 선택하십시오. 쉼표(,) 등의 기타 문자를 지정하려면 **기타**를 선택하십시오.
- **데이터 파일 지정.** 벌크 로드 수행 중에 작성된 데이터 파일에 사용할 경로를 입력하려면 선택하십시오. 기본적으로 서버의 temp 디렉토리에서 임시 파일이 작성됩니다.
- **로더 프로그램 지정.** 벌크 로드 프로그램을 지정하려면 선택하십시오. 기본적으로 소프트웨어는 IBM SPSS Modeler 설치의 *scripts* 서브디렉토리에서 지정된 데이터베이스에 대해 실행할 Python 스크립트를 검색합니다. 여러 스크립트가 소프트웨어와 함께 포함되었으며 *scripts* 서브디렉토리 아래의 기술 문서와 함께 사용 가능합니다.
- **로그 생성.** 지정된 디렉토리에 로그 파일을 생성하려면 선택하십시오. 로그 파일은 오류 정보를 포함하고 있으며 벌크 로드 조작이 실패하는 경우에 유용합니다.
- **테이블 크기 확인.** 테이블 크기 증가가 IBM SPSS Modeler에서 내보낸 행의 수와 일치하는지 확인하는 테이블 확인을 수행하려면 선택하십시오.
- **추가 로더 옵션.** 로더 프로그램에 대한 추가적인 인수를 지정합니다. 공백이 포함된 인수의 경우에는 큰따옴표를 사용하십시오.

큰따옴표는 백슬래시로 이스케이프하여 선택적 인수에 포함됩니다. 예를 들어, `-comment "This is a \"comment\""`로 지정된 옵션은 `-comment` 플래그와 `This is a "comment"`로 렌더링되는 주석 자체를 모두 포함합니다.

단일 백슬래시는 또다른 백슬래시로 이스케이프하여 포함될 수 있습니다. 예를 들어, `-specialdir "C:\\Test Scripts\\"`로 지정된 옵션은 `-specialdir` 플래그와 `C:\Test Scripts\`로 렌더링된 디렉토리를 포함합니다.

벌크 로더 프로그래밍

데이터베이스 내보내기 노드에는 고급 옵션 대화 상자에서 벌크 로드를 위한 옵션이 있습니다. 벌크 로더 프로그램을 사용하면 텍스트 파일에서 데이터베이스로 데이터를 로드할 수 있습니다.

벌크 로드 사용 - 외부 로더를 통해 옵션은 다음 세 가지 작업을 수행하도록 IBM SPSS Modeler를 구성합니다.

- 필요한 데이터베이스 테이블 작성.
- 텍스트 파일에 데이터 내보내기.
- 이 파일에서 데이터베이스 테이블로 데이터를 로드하기 위해 벌크 로더 프로그램 호출.

일반적으로 벌크 로더 프로그램은 데이터베이스 로드 유틸리티 자체(예: Oracle의 sqlldr 유틸리티)가 아니지만 올바른 인수를 구성하는 작은 스크립트 또는 프로그램이 데이터베이스 특정 보조 파일(예: 제어 파일)을 작성한 후 데이터베이스 로드 유틸리티를 호출합니다. 다음 절의 정보는 기존 벌크 로더를 편집하는 데 유용합니다.

또는 벌크 로드를 위해 사용자의 프로그램을 작성할 수 있습니다. 자세한 정보는 364 페이지의 『벌크 로더 프로그램 개발』의 내용을 참조하십시오. 표준 기술 지원 계약에는 이 사항이 포함되어 있지 않으며 지원이 필요한 경우 IBM 서비스 담당자에게 문의해야 합니다.

벌크 로드를 위한 스크립트

IBM SPSS Modeler에는 Python 스크립트를 사용하여 구현되는 여러 가지 다른 데이터베이스를 위한 여러 개의 벌크 로더 프로그램이 제공됩니다. **외부 로더를 통해** 옵션을 선택하여 데이터베이스 내보내기 노드가 포함된 스트림을 실행하는 경우 IBM SPSS Modeler는 ODBC를 통해 데이터베이스 테이블을 작성하고(필요한 경우) IBM SPSS Modeler Server를 실행 중인 호스트에서 임시 파일에 데이터를 내보낸 후 벌크 로드 스크립트를 호출합니다. 그런 다음, 이 스크립트는 DBMS 벤더에서 제공하는 유틸리티를 실행하여 임시 파일에서 데이터베이스로 데이터를 업로드합니다.

참고: IBM SPSS Modeler 설치에는 Python 런타임 해석기가 포함되지 않으므로 Python을 별도로 설치해야 합니다. 자세한 정보는 357 페이지의 『데이터베이스 내보내기 고급 옵션』의 내용을 참조하십시오.

다음 표에 나열된 데이터베이스에 스크립트가 제공됩니다(IBM SPSS Modeler 설치 디렉토리의 *\scripts* 폴더에).

표 42. 제공되는 벌크 로더 스크립트

데이터베이스	스크립트 이름	추가 정보
IBM DB2	*db2_loader.py*	자세한 정보는 『IBM DB2 데이터베이스에 데이터 벌크 로드』의 내용을 참조하십시오.
IBM Netezza	*netezza_loader.py*	자세한 정보는 361 페이지의 『IBM Netezza 데이터베이스에 데이터 벌크 로드』의 내용을 참조하십시오.
Oracle	*oracle_loader.py*	자세한 정보는 361 페이지의 『Oracle 데이터베이스에 데이터 벌크 로드』의 내용을 참조하십시오.
SQL Server	*mssql_loader.py*	자세한 정보는 362 페이지의 『SQL Server 데이터베이스에 데이터 벌크 로드』의 내용을 참조하십시오.
Teradata	*teradata_loader.py*	자세한 정보는 363 페이지의 『Teradata 데이터베이스에 데이터 벌크 로드』의 내용을 참조하십시오.

IBM DB2 데이터베이스에 데이터 벌크 로드

다음 사항은 DB 내보내기 고급 옵션 대화 상자의 외부 로더 옵션을 사용하여 IBM SPSS Modeler에서 IBM DB2 데이터베이스로 벌크 로드하도록 구성하는 데 유용할 수 있습니다.

DB2 명령행 프로세서(CLP) 유틸리티가 설치되어 있는지 확인하십시오.

db2_loader.py 스크립트는 DB2 LOAD 명령을 호출합니다. 명령행 프로세서(UNIX의 *db2*, Windows의 *db2cmd*)가 *db2_loader.py*를 실행할 서버(일반적으로 IBM SPSS Modeler Server를 실행 중인 호스트)에 설치되어 있는지 확인하십시오.

로컬 데이터베이스 별명이 실제 데이터베이스 이름과 동일한지 확인하십시오.

DB2 로컬 데이터베이스 별명은 로컬 또는 원격 DB2 인스턴스에서 데이터베이스를 참조하기 위해 DB2 클라이언트 소프트웨어에서 사용하는 이름입니다. 로컬 데이터베이스 별명이 원격 데이터베이스의 이름과 다른 경우 추가 로더 옵션을 제공하십시오.

```
-alias <local_database_alias>
```

예를 들어, 원격 데이터베이스의 이름이 호스트 GALAXY에서 STARS로 지정되었지만 IBM SPSS Modeler Server를 실행 중인 호스트의 DB2 로컬 데이터베이스 별명이 STARS_GALAXY입니다. 추가 로더 옵션을 사용하십시오.

```
-alias STARS_GALAXY
```

비ASCII 문자 데이터 인코딩

ASCII 형식이 아닌 데이터를 벌크 로드하는 경우 *db2_loader.py*의 구성 섹션에 있는 코드 페이지 변수가 사용자의 시스템에서 올바르게 설정되었는지 확인하십시오.

공백 문자열

공백 문자열을 널값으로 데이터베이스에 내보냅니다.

IBM Netezza 데이터베이스에 데이터 벌크 로드

다음 사항은 DB 내보내기 고급 옵션 대화 상자의 외부 로더 옵션을 사용하여 IBM SPSS Modeler에서 IBM Netezza 데이터베이스로 벌크 로드하도록 구성하는 데 유용할 수 있습니다.

Netezza nzload 유틸리티가 설치되었는지 확인

netezza_loader.py 스크립트는 Netezza 유틸리티 *nzload*를 호출합니다. *netezza_loader.py*를 실행할 서버에 *nzload*가 설치되었으며 올바르게 구성되었는지 확인하십시오.

비ASCII 데이터 내보내기

내보내기에 ASCII 형식이 아닌 데이터가 포함된 경우 DB 내보내기 고급 옵션 대화 상자의 **추가 로더 옵션** 필드에 `-encoding UTF8`을 추가해야 할 수도 있습니다. 이 경우 비ASCII 데이터가 올바르게 업로드되었는지 확인해야 합니다.

날짜, 시간, 시간소인 형식 데이터

스트림 특성에서 데이터 형식을 **DD-MM-YYYY**로 설정하고 시간 형식을 **HH:MM:SS**로 설정하십시오.

공백 문자열

공백 문자열을 널값으로 데이터베이스에 내보냅니다.

기존 테이블에 데이터를 삽입할 때 스트림 및 대상 테이블에 있는 다른 순서의 열

스트림에 있는 열의 순서가 대상 테이블에 있는 열과 다른 경우 데이터 값이 잘못된 열에 삽입됩니다. 필드 재정렬 노드를 사용하여 스트림에 있는 열의 순서가 대상 테이블에 있는 순서와 일치하는지 확인하십시오. 자세한 정보는 187 페이지의 『필드 다시 정렬 노드』의 내용을 참조하십시오.

nzload 진행 상태 추적

로컬 모드에서 IBM SPSS Modeler를 실행하는 경우 DB 내보내기 고급 옵션 대화 상자의 **추가 로더 옵션** 필드에 `-sts`를 추가하여 *nzload* 유틸리티로 여는 명령 창에서 10000행마다 상태 메시지를 보십시오.

Oracle 데이터베이스에 데이터 벌크 로드

다음 사항은 DB 내보내기 고급 옵션 대화 상자의 외부 로더 옵션을 사용하여 IBM SPSS Modeler에서 Oracle 데이터베이스로 벌크 로드하도록 구성하는 데 유용할 수 있습니다.

Oracle sqlldr 유틸리티가 설치되었는지 확인

oracle_loader.py 스크립트는 Oracle 유틸리티 *sqlldr*을 호출합니다. *sqlldr*이 Oracle 클라이언트에 자동으로 포함되지 않는다는 점을 참고하십시오. *oracle_loader.py*를 실행할 서버에 *sqlldr*이 설치되었는지 확인하십시오.

데이터베이스 SID 또는 서비스 이름 지정

비로컬 Oracle 서버에 데이터를 내보내거나 로컬 Oracle 서버에 여러 데이터베이스가 있는 경우 SID 또는 서비스 이름을 전달하기 위해 DB 내보내기 고급 옵션 대화 상자의 **추가 로더 옵션** 필드에 다음을 지정해야 합니다.

```
-database <SID>
```

oracle_loader.py에서 구성 섹션 편집

UNIX(및 선택적으로 Windows) 시스템에서 *oracle_loader.py* 스크립트의 처음에 있는 구성 섹션을 편집하십시오. 여기에서 ORACLE_SID, NLS_LANG, TNS_ADMIN, ORACLE_HOME 환경 변수의 값을 적절하게 지정하고 *sqlldr* 유틸리티의 전체 경로를 지정할 수 있습니다.

날짜, 시간, 시간소인 형식 데이터

스트림 특성에서 일반적으로 날짜 형식을 **YYYY-MM-DD**로 설정하고 시간 형식을 **HH:MM:SS**로 설정해야 합니다.

위와 다른 날짜 및 시간 형식을 사용해야 하는 경우 Oracle 문서를 참조하고 *oracle_loader.py* 스크립트 파일을 편집하십시오.

비ASCII 문자 데이터 인코딩

ASCII 형식이 아닌 데이터를 벌크 로드하는 경우 시스템에서 환경 변수 NLS_LANG이 올바르게 설정되었는지 확인해야 합니다. 이는 Oracle 로더 유틸리티 *sqlldr*에서 읽습니다. 예를 들어, Windows에서 Shift-JIS의 NLS_LANG에 올바른 값은 Japanese_Japan.JA16SJIS입니다. NLS_LANG에 대한 자세한 정보는 Oracle 문서를 확인하십시오.

공백 문자열

공백 문자열을 널값으로 데이터베이스에 내보냅니다.

SQL Server 데이터베이스에 데이터 벌크 로드

다음 사항은 DB 내보내기 고급 옵션 대화 상자의 외부 로더 옵션을 사용하여 IBM SPSS Modeler에서 SQL Server 데이터베이스로 벌크 로드하도록 구성하는 데 유용할 수 있습니다.

SQL Server bcp.exe 유틸리티가 설치되었는지 확인

mssql_loader.py 스크립트는 SQL Server 유틸리티 *bcp.exe*를 호출합니다. *mssql_loader.py*를 실행할 서버에 *bcp.exe*가 설치되었는지 확인하십시오.

구분자로 공백 사용이 작동되지 않음

DB 내보내기 고급 옵션 대화 상자에서 구분자로 공백을 선택하지 마십시오.

테이블 크기 확인 옵션 권장

DB 내보내기 고급 옵션 대화 상자에서 **테이블 크기 확인** 옵션을 사용하도록 권장됩니다. 벌크 로드 프로세스의 실패는 항상 발견되지는 않으며 이 옵션을 사용하면 올바른 수의 행이 로드되었는지 추가 확인을 수행합니다.

공백 문자열

공백 문자열을 널값으로 데이터베이스에 내보냅니다.

완전한 SQL Server 이름 지정 인스턴스 지정

SPSS Modeler가 규정되지 않은 호스트 이름으로 인해 SQL Server에 액세스할 수 없는 경우가 있을 수 있으며 다음 오류를 표시합니다.

```
외부 벌크 로더를 실행하는 중 오류가 발생했습니다.
로그 파일이 자세한 정보를 제공할 수 있습니다.
```

이 오류를 정정하려면 **추가 로더 옵션** 필드에 큰따옴표를 포함하여 다음 문자열을 추가하십시오.

```
"-S mhreboot.spss.com\SQLEXPRESS"
```

Teradata 데이터베이스에 데이터 벌크 로드

다음 사항은 DB 내보내기 고급 옵션 대화 상자의 외부 로더 옵션을 사용하여 IBM SPSS Modeler에서 Teradata 데이터베이스로 벌크 로드하도록 구성하는 데 유용할 수 있습니다.

Teradata fastload 유틸리티가 설치되었는지 확인

teradata_loader.py 스크립트는 Teradata 유틸리티 *fastload*를 호출합니다. *teradata_loader.py*를 실행할 서버에 *fastload*가 설치되었으며 올바르게 구성되었는지 확인하십시오.

데이터를 비어 있는 테이블에만 벌크 로드할 수 있음

벌크 로드의 대상으로 비어 있는 테이블만을 사용할 수 있습니다. 대상 테이블에 벌크 로드 이전의 데이터가 있는 경우 작업이 실패합니다.

날짜, 시간, 시간소인 형식 데이터

스트림 특성에서 날짜 형식을 **YYYY-MM-DD**로 설정하고 시간 형식을 **HH:MM:SS**로 설정하십시오.

공백 문자열

공백 문자열을 널값으로 데이터베이스에 내보냅니다.

Teradata 프로세스 ID(tdpid)

기본적으로 *fastload*는 `tdpid=dbc`를 사용하여 Teradata 시스템에 데이터를 내보냅니다. 일반적으로 `dbccop1`을 Teradata 서버의 IP 주소와 연관시키는 HOSTS 파일의 항목이 있습니다. 다른 서버를 사용하려면 이 서버의 `tdpid`를 전달하기 위해 DB 내보내기 고급 옵션 대화 상자의 **추가 로더 옵션** 필드에 다음을 지정하십시오.

```
-tdpid <id>
```

테이블 및 열 이름의 공백

테이블 또는 열 이름에 공백이 포함된 경우 벌크 로드 작업이 실패합니다. 가능한 경우 공백을 제거하여 테이블 또는 열 이름을 변경하십시오.

벌크 로더 프로그램 개발

이 주제에서는 텍스트 파일에서 데이터베이스로 데이터를 로드하기 위해 IBM SPSS Modeler에서 실행할 수 있는 벌크 로더 프로그램을 개발하는 방법을 설명합니다. 표준 기술 지원 계약에는 이 사항이 포함되어 있지 않으며 지원이 필요한 경우 IBM 서비스 담당자에게 문의해야 합니다.

Python을 사용한 벌크 로더 프로그램 작성

기본적으로 IBM SPSS Modeler는 데이터베이스 유형에 기반하여 기본 벌크 로더 프로그램을 검색합니다. 360 페이지의 표 42의 내용을 참조하십시오.

일괄처리 로더 프로그램을 개발하는 데 도움이 되는 *test_loader.py* 스크립트를 사용할 수 있습니다. 자세한 정보는 366 페이지의 『벌크 로더 프로그램 테스트』의 내용을 참조하십시오.

벌크 로더 프로그램에 전달된 오브젝트

IBM SPSS Modeler는 벌크 로더 프로그램에 전달되는 두 개의 파일을 작성합니다.

- **데이터 파일.** 이 파일에는 텍스트 형식으로 로드할 데이터가 있습니다.
- **스키마 파일.** 이 파일은 열의 이름과 유형에 대해 설명하고 데이터 파일을 형식화하는 방법(예: 필드 간 구분자로 사용되는 문자)에 대한 정보를 제공하는 XML 파일입니다.

또한 IBM SPSS Modeler는 벌크 로드 프로그램을 호출할 때 테이블 이름, 사용자 이름 및 비밀번호와 같은 기타 정보를 인수로 전달합니다.

참고: IBM SPSS Modeler에 성공적으로 완료되었음을 알리기 위해 벌크 로더 프로그램은 스키마 파일을 삭제해야 합니다.

벌크 로더 프로그램에 전달되는 인수

프로그램에 전달되는 인수는 다음 표와 같습니다.

표 43. 벌크 로더에 전달되는 인수.

인수	설명
schemafile	스키마 파일의 경로입니다.
data file	데이터 파일의 경로입니다.
servername	DBMS 서버의 이름이며 공백일 수 있습니다.
databasename	DBMS 서버에 있는 데이터베이스의 이름이며 공백일 수 있습니다.
username	데이터베이스에 로그인하는 데 사용하는 사용자 이름입니다.
password	데이터베이스에 로그인하는 데 사용하는 비밀번호입니다.
tablename	로드할 테이블의 이름입니다.
ownername	테이블 소유자의 이름입니다(스키마 이름이라고도 함).
logfilename	로그 파일의 이름입니다(공백인 경우 로그 파일이 작성되지 않음).
rowcount	데이터 세트에 있는 행의 수입니다.

DB 내보내기 고급 옵션 대화 상자의 **추가 로더 옵션** 필드에 지정된 옵션은 이러한 표준 인수 이후에 벌크 로더 프로그램에 전달됩니다.

데이터 파일의 형식

데이터는 텍스트 형식으로 데이터 파일에 기록되며 각 필드는 DB 내보내기 고급 옵션 대화 상자에 지정된 구분 문자로 구분됩니다. 다음은 탭 구분 데이터 파일이 표시되는 방식의 예입니다.

```
48  F  HIGH    NORMAL  0.692623  0.055369  drugA
15  M  NORMAL  HIGH    0.678247  0.040851  drugY
37  M  HIGH    NORMAL  0.538192  0.069780  drugA
35  F  HIGH    HIGH    0.635680  0.068481  drugA
```

파일은 IBM SPSS Modeler Server(또는 IBM SPSS Modeler Server에 연결되지 않은 경우 IBM SPSS Modeler)에서 사용하는 로컬 인코딩으로 작성됩니다. 일부 형식화는 IBM SPSS Modeler 스트림 설정을 통해 제어됩니다.

스키마 파일의 형식

스키마 파일은 데이터 파일에 대해 설명하는 XML 파일입니다. 다음은 이전 데이터 파일에 함께 제공되는 예입니다.

```
<?xml version="1.0" encoding="UTF-8"?>
<DBSCHEMA version="1.0">
 <table delimiter="\t" commit_every="10000" date_format="YYYY-MM-DD" time_format="HH:MM:SS"
append_existing="false" delete_datafile="false">
  <column name="Age" encoded_name="416765" type="integer"/>
  <column name="Sex" encoded_name="536578" type="char" size="1"/>
  <column name="BP" encoded_name="4250" type="char" size="6"/>
  <column name="Cholesterol" encoded_name="43686F6C65737465726F6C" type="char" size="6"/>
  <column name="Na" encoded_name="4E61" type="real"/>
  <column name="K" encoded_name="4B" type="real"/>
  <column name="Drug" encoded_name="44727567" type="char" size="5"/>
 </table>
</DBSCHEMA>
```

다음 두 개의 표는 스키마 파일의 <table> 및 <column> 요소에 대한 속성을 나열합니다.

표 44. <table> 요소의 속성.

속성	설명
delimiter	필드 구분 문자입니다(TAB이 \t로 표시됨).
commit_every	일괄처리 크기 간격입니다(DB 내보내기 고급 옵션 대화 상자에 있음).
date_format	날짜를 표시하는 데 사용되는 형식입니다.
time_format	시간을 표시하는 데 사용되는 형식입니다.
append_existing	로드할 테이블에 데이터가 이미 있으면 true이고 그렇지 않으면 false입니다.
delete_datafile	벌크 로드 프로그램이 로드 완료 시 데이터 파일을 삭제해야 하는 경우 true입니다.

표 45. <column> 요소의 속성.

속성	설명
name	열 이름입니다.

표 45. <column> 요소의 속성 (계속).

속성	설명
encoded_name	데이터 파일과 동일한 인코딩으로 변환되고 일련의 2자리 16진수로 출력되는 열 이름입니다.
type	열의 데이터 유형이며 integer, real, char, time, date, datetime 중 하나입니다.
size	char 데이터 유형의 경우 문자 수로 나타내는 열의 최대 너비입니다.

벌크 로더 프로그램 테스트

IBM SPSS Modeler 설치 디렉토리의 *\scripts* 폴더에 포함된 테스트 스크립트 *test_loader.py*를 사용하여 벌크 로드를 테스트할 수 있습니다. 이와 같이 테스트하면 IBM SPSS Modeler에서 사용하도록 벌크 로드 프로그램 또는 스크립트를 개발, 디버그 또는 문제점 해결할 때 유용합니다.

테스트 스크립트를 사용하려면 다음과 같이 계속하십시오.

1. *test_loader.py* 스크립트를 실행하여 스키마 및 데이터 파일을 *schema.xml* 및 *data.txt* 파일에 복사하고 Windows 일괄처리 파일(*test.bat*)을 작성하십시오.
2. *test.bat* 파일을 편집하여 테스트할 벌크 로더 프로그램 또는 스크립트를 선택하십시오.
3. 명령 쉘에서 *test.bat*를 실행하여 선택한 벌크 로드 프로그램 또는 스크립트를 테스트하십시오.

참고: *test.bat*를 실행하면 데이터가 실제로 데이터베이스에 로드되지 않습니다.

플랫 파일 내보내기 노드

플랫 파일 내보내기 노드를 사용하면 데이터를 구분된 텍스트 파일로 쓸 수 있습니다. 이 방법은 다른 분석 또는 스프레드시트 소프트웨어가 읽을 수 있는 데이터 내보내기에 유용합니다.

데이터에 지리 공간적 정보가 포함되어 있으면 이를 플랫 파일로 내보낼 수 있으며 동일한 스트림 내에서 사용하기 위해 가변파일 소스 노드를 생성하는 경우에는 모든 저장 공간, 측정 및 지리 공간적 메타데이터가 새 소스 노드에서 세분화됩니다. 그러나 데이터를 내보낸 다음 이를 다른 스트림으로 가져오는 경우에는 새 소스 노드에서 지리 공간적 메타데이터를 설정하려면 몇 가지 추가 단계를 수행해야 합니다. 자세한 정보는 28 페이지의 『가변파일 노드』의 내용을 참조하십시오.

참고: IBM SPSS Modeler에서 더 이상 캐시 파일에 대해 이전 캐시 형식을 사용하지 않으므로 파일을 해당 형식으로 쓸 수 없습니다. IBM SPSS Modeler 캐시 파일은 이제 IBM SPSS Statistics *.sav* 형식으로 저장되며 이 형식은 Statistics 내보내기 노드를 사용하여 작성할 수 있습니다. 자세한 정보는 385 페이지의 『통계량 내보내기 노드』의 내용을 참조하십시오.

플랫 파일 내보내기 탭

파일 내보내기. 파일의 이름을 지정합니다. 파일 이름을 입력하거나 파일 선택기 단추를 클릭하여 파일의 위치로 이동하십시오.

쓰기 모드. 겹쳐쓰기가 선택되면 지정된 파일의 기존 데이터가 모두 겹쳐써집니다. **추가**가 선택되면 출력이 기존 파일의 끝에 추가되어 포함된 모든 데이터를 유지합니다.

- **필드 이름 포함.** 이 옵션이 선택되면 출력 파일의 첫 번째 행에 필드 이름이 기록됩니다. 이 옵션은 **겹쳐쓰기** 쓰기 모드에만 사용할 수 있습니다.

각 레코드 다음에 줄 바꾸기. 이 옵션이 선택된 경우에는 각각의 레코드가 출력 파일의 새로운 행에서 작성됩니다.

필드 구분 문자. 생성된 텍스트 파일의 필드 값 사이에 삽입할 문자를 지정합니다. 옵션은 **쉼표**, **탭**, **공백** 및 **기타**입니다. **기타**를 선택하는 경우에는 원하는 구분 문자를 텍스트 상자에 입력하십시오.

따옴표 기호. 기호 필드의 값에 사용할 따옴표의 유형을 지정합니다. 옵션은 **없음**(값을 따옴표로 묶지 않음), **작은따옴표(')**, **큰따옴표(")** 및 **기타**입니다. **기타**를 선택하는 경우에는 원하는 따옴표 문자를 텍스트 상자에 입력하십시오.

인코딩. 사용되는 텍스트 인코딩 방법을 지정합니다. 시스템 기본값, 스트림 기본값 또는 UTF-8 중에서 선택할 수 있습니다.

- 시스템 기본값은 Windows 제어판에 지정되어 있거나 분산 모드에서 실행 중인 경우 서버 컴퓨터에 지정되어 있습니다.
- 스트림 기본값은 스트림 특성 대화 상자에서 지정됩니다.

소수점 기호. 데이터에서 소수점 표시 방법을 지정합니다.

- **스트림 기본값.** 현재 스트림의 기본 설정에 의해 정의된 소수점 구분 문자가 사용됩니다. 이는 일반적으로 컴퓨터의 로케일 설정에 의해 정의된 소수점 구분 문자입니다.
- **마침표(.).** 마침표가 소수점 구분 문자로 사용됩니다.
- **쉼표(,).** 쉼표가 소수점 구분 문자로 사용됩니다.

현재 데이터의 입력 노드 생성. 내보낸 데이터 파일을 읽을 가변파일 소스 노드를 자동으로 생성하려면 이 옵션을 선택하십시오. 자세한 정보는 28 페이지의 『가변파일 노드』의 내용을 참조하십시오.

Data Collection 내보내기 노드

Data Collection 내보내기 노드는 Data Collection 데이터 모델을 기반으로 Data Collection 시장 조사 소프트웨어에서 사용하는 형식으로 데이터를 저장합니다. 이 형식은 케이스 데이터(설문조사 중에 수집된 질문에 대한 실제 응답)를 케이스 데이터의 수집 및 구성 방식에 대해 설명하는 메타데이터와 구별합니다. 메타데이터는 케이스 데이터의 구조 정의, 질문 텍스트, 변수 이름 및 설명, 다중 응답 세트, 다양한 텍스트의 변환 등의 정보로 구성됩니다. 자세한 정보는 34 페이지의 『Data Collection 노드』의 내용을 참조하십시오.

메타데이터 파일. 내보낸 메타데이터가 저장될 질문지 정의 파일(*.mdd*)의 이름을 지정합니다. 기본 질문지는 필드 유형 정보를 기반으로 작성됩니다. 예를 들어, 명목(세트) 필드는 각각의 정의된 값에 대해 별도의 선택란 및 질문 텍스트로 사용되는 필드 설명이 포함된 단일 질문으로 표시될 수 있습니다.

메타데이터 병합. 메타데이터가 기존 버전을 겹쳐쓸지 아니면 기존 메타데이터와 병합될지를 지정합니다. 병합 옵션이 선택되면 스트림이 실행될 때마다 새 버전이 작성됩니다. 이를 통해 질문지가 변경될 때 질문지의 버전을 추적할 수 있습니다. 각각의 버전은 특정 케이스 데이터 세트를 수집하는 데 사용되는 메타데이터의 스냅샷으로 간주될 수 있습니다.

시스템 변수 사용. 시스템 변수가 내보낸 *.mdd* 파일에 포함되는지 여부를 지정합니다. 여기에는 *Respondent.Serial*, *Respondent.Origin*, *DataCollection.StartTime* 등의 변수가 포함됩니다.

케이스 데이터 설정. 케이스 데이터를 내보내는 IBM SPSS Statistics 데이터(*.sav*) 파일을 지정합니다. 변수 및 값 이름에 대한 모든 제한사항이 여기서 적용되므로 예를 들어, 필터 탭으로 전환한 후 필터 옵션 메뉴의 "IBM SPSS Statistics에 대해 이름 바꾸기" 옵션을 사용하여 필드 이름에서 유효하지 않은 문자를 정정해야 할 수 있습니다.

현재 데이터의 입력 노드 생성. 내보낸 데이터 파일을 읽을 Data Collection 소스 노드를 자동으로 생성하려면 이 옵션을 선택하십시오.

다중 응답 세트. 스트림에서 정의된 다중 응답 세트는 파일을 내보낼 때 자동으로 유지됩니다. 필터 탭의 모든 노드에서 다중 응답 세트를 보고 편집할 수 있습니다. 자세한 정보는 156 페이지의 『다중 응답 세트 편집』의 내용을 참조하십시오.

Analytic Server 내보내기 노드

Analytic Server 내보내기 노드를 사용하면 분석의 데이터를 기존 Analytic Server 데이터 소스에 쓸 수 있습니다. 예를 들어, 이는 HDFS(Hadoop Distributed File System) 또는 데이터베이스의 텍스트 파일일 수 있습니다.

일반적으로 Analytic Server 내보내기 노드를 가진 스트림은 Analytic Server 소스 노드로도 시작하며 Analytic Server에 제출되고 HDFS에서 실행됩니다. 또는 "로컬" 데이터 소스를 가진 스트림은 Analytic Server와 함께 사용하도록 상대적으로 작은 데이터 세트(100,000개 이하의 레코드)를 업로드하기 위해 Analytic Server 내보내기 노드로 끝날 수 있습니다.

데이터 소스. 사용할 데이터가 포함된 데이터 소스를 선택하십시오. 데이터 소스에는 해당 소스와 연관된 파일 및 메타데이터가 포함되어 있습니다. **선택**을 클릭하여 사용 가능한 데이터 소스의 목록을 표시하십시오. 자세한 정보는 13 페이지의 『데이터 소스 선택』의 내용을 참조하십시오.

새 데이터 소스를 작성하거나 기존 데이터 소스를 편집해야 하는 경우에는 **데이터 소스 편집기 시작...**을 클릭하십시오.

모드. 기존 데이터 소스에 추가하려면 **추가**를 선택하고 데이터 소스의 컨텐츠를 바꾸려면 **겹쳐쓰기**를 선택하십시오.

현재 데이터의 입력 노드 생성. 지정된 데이터 소스에 내보낸 대로 데이터에 대한 소스 노드를 생성하려면 선택하십시오. 이 노드는 스트림 캔버스에 추가됩니다.

IBM Cognos BI 내보내기 노드

IBM Cognos BI 내보내기 노드를 사용하면 IBM SPSS Modeler 스트림의 데이터를 UTF-8 형식으로 Cognos BI에 내보낼 수 있습니다. 이 방식으로 Cognos BI는 IBM SPSS Modeler로부터 변환되거나 스코어링된 데이터를 사용할 수 있습니다. 예를 들어, Cognos BI Report Studio를 사용하여 예측 및 신뢰도를 포함한 내보낸 데이터를 기반으로 보고서를 작성할 수 있습니다. 그런 다음 보고서를 Cognos BI 서버에 저장하고 Cognos BI 사용자에게 분배할 수 있습니다.

참고: 관계형 데이터만 내보낼 수 있으며 OLAP 데이터는 내보낼 수 없습니다.

데이터를 Cognos BI에 내보내려면 다음을 지정해야 합니다.

- Cognos 연결 - Cognos BI 서버에 대한 연결
- ODBC 연결 - Cognos BI 서버가 사용하는 Cognos 데이터 서버에 대한 연결

Cognos 연결 내에서 사용할 Cognos 데이터 소스를 지정합니다. 이 데이터 소스는 ODBC 데이터 소스와 동일한 로그인을 사용해야 합니다.

실제 스트림 데이터를 데이터 서버에 내보내고 패키지 메타데이터를 Cognos BI 서버에 내보냅니다.

다른 내보내기 노드와 마찬가지로 노드 대화 상자의 출판 탭을 통해 IBM SPSS Modeler Solution Publisher를 사용하여 배포를 위한 스트림을 게시할 수도 있습니다.

Cognos 연결

여기서는 내보내기를 위해 사용할 Cognos BI 서버에 대한 연결을 지정합니다. 프로시저에는 Cognos BI 서버의 새 패키지에 메타데이터를 내보내는 것이 포함되지만 스트림 데이터는 Cognos 데이터 서버에 내보냅니다.

연결. 편집 단추를 클릭하여 데이터를 내보낼 Cognos BI 서버의 기타 세부사항 및 URL을 정의할 수 있는 대화 상자를 표시하십시오. IBM SPSS Modeler를 통해 이미 Cognos BI 서버에 로그인한 경우에는 현재 연결의 세부사항도 편집할 수 있습니다. 자세한 정보는 42 페이지의 『Cognos 연결』의 내용을 참조하십시오.

데이터 소스. 데이터를 내보내는 Cognos 데이터 소스(일반적으로 데이터베이스)의 이름입니다. 드롭 다운 목록에는 현재 연결에서 액세스할 수 있는 모든 Cognos 데이터 소스가 표시됩니다. **새로 고치기** 단추를 클릭하여 목록을 업데이트하십시오.

폴더. 내보내기 패키지를 작성할 Cognos BI 서버의 폴더 이름 및 경로입니다.

패키지 이름. 내보낸 메타데이터를 포함할 지정된 폴더의 패키지 이름입니다. 이는 단일 쿼리 제목을 가진 새 패키지여야 하며 기존 패키지에 내보낼 수 없습니다.

모드. 내보내기 수행 방법을 지정합니다.

- **지금 패키지 게시.** (기본값) **실행**을 클릭하는 즉시 내보내기 조작을 수행합니다.

- **조치 스크립트 내보내기.** 예를 들어, Framework Manager를 사용하여 나중에 실행할 수 있는 XML 스크립트를 작성하여 내보내기를 수행합니다. **파일** 필드에서 스크립트의 경로 및 파일 이름을 입력하거나 **편집** 단추를 사용하여 스크립트 파일의 이름 및 위치를 지정하십시오.

현재 데이터의 입력 노드 생성. 지정된 데이터 소스 및 테이블로 내보낸 대로 데이터에 대한 소스 노드를 생성하려면 선택하십시오. **실행**을 클릭하면 이 노드가 스트림 캔버스에 추가됩니다.

ODBC 연결

여기서는 스트림 데이터를 내보낼 Cognos 데이터 서버(즉, 데이터베이스)에 대한 연결을 지정합니다.

참고: 여기서 지정하는 데이터 소스가 **Cognos 연결** 패널에서 지정된 동일한 데이터 소스를 가리키는지 확인해야 합니다. Cognos 연결 데이터 소스가 ODBC 데이터 소스와 동일한 로그인을 사용하는지도 확인해야 합니다.

데이터 소스. 선택된 데이터 소스를 표시합니다. 이름을 입력하거나 드롭 다운 목록에서 이름을 선택하십시오. 목록에 원하는 데이터베이스가 표시되지 않으면 **새 데이터베이스 연결 추가**를 선택하고 데이터베이스 연결 대화 상자에서 데이터베이스를 찾으십시오. 자세한 정보는 20 페이지의 『데이터베이스 연결 추가』의 내용을 참조하십시오.

테이블 이름. 데이터를 전송할 테이블의 이름을 입력하십시오. **테이블에 삽입** 옵션을 선택하는 경우에는 **선택** 단추를 클릭하여 데이터베이스에서 기존 테이블을 선택할 수 있습니다.

테이블 작성. 새 데이터베이스 테이블을 작성하거나 기존 데이터베이스 테이블을 겹쳐쓰려면 이 옵션을 선택하십시오.

테이블에 삽입. 기존 데이터베이스 테이블에서 새 행으로 데이터를 삽입하려면 이 옵션을 선택하십시오.

테이블 병합. (사용 가능한 경우) 선택된 데이터베이스 열을 해당 소스 데이터 필드의 값으로 업데이트하려면 이 옵션을 선택하십시오. 이 옵션을 선택하면 소스 데이터 필드를 데이터베이스 열에 맵핑할 수 있는 대화 상자를 표시하는 **병합** 단추를 사용할 수 있습니다.

기존 테이블 삭제. 새 테이블 작성 시 동일한 이름의 기존 테이블을 삭제하려면 이 옵션을 선택하십시오.

기존 행 삭제. 테이블에 삽입 시 내보내기 전에 테이블에서 기존 행을 삭제하려면 이 옵션을 선택하십시오.

참고: 위 두 옵션 중 하나가 선택되면 노드를 실행할 때 **겹쳐쓰기 경고** 메시지가 수신됩니다. 경고를 표시하지 않으려면 사용자 옵션 대화 상자의 알림 탭에서 **노드가 데이터베이스 테이블을 겹쳐쓸 때 경고**를 선택 취소하십시오.

기본 문자열 크기. 업스트림 유형 노드에서 유형 없음으로 표시한 필드는 데이터베이스에 문자열 필드로 작성됩니다. 유형 없는 필드에 사용할 문자열의 크기를 지정하십시오.

스키마를 클릭하여 다양한 내보내기 옵션을 설정(이 기능을 지원하는 데이터베이스의 경우)하고 필드에 대해 SQL 데이터 유형을 설정하고 데이터베이스 인덱싱을 위해 기본 키를 지정할 수 있는 대화 상자를 여십시오. 자세한 정보는 353 페이지의 『데이터베이스 내보내기 스키마 옵션』의 내용을 참조하십시오.

인덱스를 클릭하여 데이터베이스 성능을 향상시키기 위해 내보낸 테이블을 인덱싱하는 데 필요한 옵션을 지정하십시오. 자세한 정보는 355 페이지의 『데이터베이스 내보내기 인덱스 옵션』의 내용을 참조하십시오.

고급을 클릭하여 벌크 로드 및 데이터베이스 커미트 옵션을 지정하십시오. 자세한 정보는 357 페이지의 『데이터베이스 내보내기 고급 옵션』의 내용을 참조하십시오.

테이블 및 열 이름 따옴표로 묶기. CREATE TABLE문을 데이터베이스에 전송할 때 사용되는 옵션을 선택하십시오. 공백 또는 비표준 문자가 포함된 테이블 또는 열은 따옴표로 묶어야 합니다.

- **필요에 따라.** IBM SPSS Modeler가 개별적으로 따옴표가 필요한 시기를 자동으로 판별할 수 있게 하려면 선택하십시오.
- **항상.** 테이블 및 열 이름을 항상 따옴표로 묶으려면 선택하십시오.
- **사용 안 함.** 따옴표를 사용하지 않으려면 선택하십시오.

현재 데이터의 입력 노드 생성. 지정된 데이터 소스 및 테이블로 내보낸 대로 데이터에 대한 소스 노드를 생성하려면 선택하십시오. **실행**을 클릭하면 이 노드가 스트림 캔버스에 추가됩니다.

IBM Cognos TM1 내보내기 노드

IBM Cognos BI 내보내기 노드에서는 IBM SPSS Modeler 스트림의 데이터를 Cognos TM1로 내보낼 수 있습니다. 이러한 방식으로 Cognos BI는 IBM SPSS Modeler의 변환된 데이터 또는 스코어링된 데이터를 이용할 수 있습니다.

참고: 측도만 내보낼 수 있습니다(컨텍스트 차원 데이터는 내보낼 수 없음). 또는 큐브에 새 요소를 추가할 수 있습니다.

Cognos BI로 데이터를 내보내려면 다음을 지정해야 합니다.

- Cognos TM1 서버로의 연결
- 데이터를 내보내는 큐브
- SPSS 데이터 이름에서 동등한 TM1 차원 및 측도로의 맵핑

참고: TM1 사용자에게는 큐브 쓰기 권한, 차원 읽기 권한 및 차원 요소 쓰기 권한이 필요합니다.

다른 내보내기 노드에서와 마찬가지로, 노드 대화 상자의 출판 탭을 사용하여 IBM SPSS Modeler Solution Publisher를 통해 배포할 스트림을 출판할 수도 있습니다.

참고: SPSS Modeler에서 TM1 소스 또는 내보내기 노드를 사용하려면 먼저 tm1s.cfg 파일에서 일부 설정을 유효화해야 합니다. 이 파일은 TM1 서버의 루트 디렉토리에 있는 TM1 서버 구성 파일입니다.

- HTTPPortNumber - 유효한 포트 번호를 설정합니다. 일반적으로 1 - 65535입니다.
- UseSSL - 참으로 설정하면 HTTPS가 전송 프로토콜로 사용됩니다. 이 경우 TM1 인증을 SPSS Modeler Server JRE로 가져와야 합니다.

데이터를 내보낼 IBM Cognos TM1 큐브에 연결

IBM Cognos TM1 데이터베이스로 데이터를 내보내는 첫 번째 단계는 IBM Cognos TM1 대화 상자의 **연결** 탭에서 관련 TM1 관리 호스트 및 연관된 서버와 큐브를 선택하는 것입니다.

참고: TM1에 데이터를 내보낼 때 실제 "널" 값만 삭제됩니다. 영(0) 값은 유효한 값으로 내보내집니다. 또한 맵핑 탭에서는 저장 유형이 *문자열*인 필드만 차원으로 맵핑할 수 있습니다. TM1로 내보내기 전, IBM SPSS Modeler 클라이언트를 사용하여 문자열이 아닌 데이터 유형을 문자열로 변환해야 합니다.

관리 호스트 연결할 TM1 서버가 설치된 관리 호스트의 URL을 입력하십시오. 관리 호스트는 모든 TM1 서버에 대한 단일 URL로 정의됩니다. 이 URL에서, 사용하는 환경에 설치되어 실행 중인 모든 IBM Cognos TM1 서버를 검색하고 액세스할 수 있습니다.

TM1 서버 관리 호스트에 연결한 경우, 가져올 데이터가 있는 서버를 선택하고 **로그인**을 클릭하십시오. 이전에 이 서버에 연결한 적이 없는 경우, **사용자 이름** 및 **비밀번호** 입력을 요구하는 프롬프트가 표시됩니다. 또는 이전에 입력하여 **저장된 신임 정보**로 저장한 로그인 세부사항을 검색할 수 있습니다.

내보낼 TM1 큐브 선택 데이터를 내보낼 수 있는 TM1 서버 내의 큐브 이름을 표시합니다.

내보낼 데이터를 선택하려면, 큐브를 선택하고 오른쪽 화살표를 클릭하여 큐브를 **큐브로 내보내기** 필드로 이동시키십시오. 큐브를 선택했으면 맵핑 탭을 사용하여 TM1 차원 및 측도를 관련 SPSS 필드 또는 고정값(*선택 조작*)으로 맵핑하십시오.

내보낼 IBM Cognos TM1 데이터 맵핑

TM1 관리 호스트 및 연관된 TM1 서버와 큐브를 선택한 후, IBM Cognos TM1 내보내기 대화 상자의 맵핑 탭을 사용하여 TM1 차원 및 측도를 SPSS 필드에 맵핑하거나 TM1 차원을 고정값으로 설정하십시오.

참고: 저장 유형이 *문자열*인 필드만 차원으로 맵핑할 수 있습니다. TM1로 내보내기 전, IBM SPSS Modeler 클라이언트를 사용하여 문자열이 아닌 데이터 유형을 문자열로 변환해야 합니다.

필드 내보내기에 사용할 수 있는 SPSS 데이터 파일의 데이터 필드 이름을 나열합니다.

TM1 차원 연결 탭에서 선택된 TM1 큐브를 해당 정규 차원, 측도 차원 및 선택된 측도 차원의 요소와 함께 표시합니다. SPSS 데이터 필드로 맵핑하려면 TM1 차원 또는 측도의 이름을 선택하십시오.

맵핑 탭에서는 다음 옵션을 사용할 수 있습니다.

측도 차원 선택 선택된 큐브의 차원 목록에서 측도 차원이 될 차원을 선택하십시오.

측도 차원을 제외한 차원을 선택하고 **선택**을 클릭하면 선택된 차원의 리프 요소를 표시하는 대화 상자가 표시됩니다. 리프 요소만 선택할 수 있습니다. 선택된 요소는 **S**로 레이블이 지정됩니다.

맵핑 선택된 SPSS 데이터 필드를 선택된 TM1 차원 또는 측도(정규 차원, 측도 차원의 특정 측도 또는 요소)로 맵핑합니다. 맵핑된 필드는 **M**으로 레이블 지정됩니다.

맵핑 해제 선택된 TM1 차원 또는 측도에서 선택된 SPSS 데이터 필드를 맵핑 해제합니다. 한 번에 하나의 맵핑만 맵핑 해제할 수 있습니다. 맵핑 해제된 SPSS 데이터 필드는 다시 왼쪽 열로 이동합니다.

새로 작성 TM1 측도 차원에서 측도를 새로 작성합니다. 새 **TM1 측도 이름**을 입력하는 대화 상자가 표시됩니다. 이 옵션은 측도 차원에만 사용할 수 있고 정규 차원에는 사용할 수 없습니다.

TM1에 대한 자세한 정보는 IBM Cognos TM1 문서(http://www-01.ibm.com/support/knowledgecenter/SS9RXT_10.2.2/com.ibm.swg.ba.cognos.ctm1.doc/welcome.html)를 참조하십시오.

SAS 내보내기 노드

참고: 이 기능은 SPSS Modeler Professional 및 SPSS Modeler Premium에서 사용 가능합니다.

SAS 내보내기 노드를 사용하면 SAS 또는 SAS 호환 가능한 소프트웨어 패키지로 읽어들이기 위해 데이터를 SAS 형식으로 쓸 수 있습니다. SAS for Windows/OS2, SAS for UNIX 또는 SAS 버전 7/8의 세 가지 SAS 파일 형식이 사용 가능합니다.

SAS 내보내기 노드 내보내기 탭

파일 내보내기. 파일의 이름을 지정합니다. 파일 이름을 입력하거나 파일 선택기 단추를 클릭하여 파일 위치를 찾아보십시오.

내보내기. 파일 내보내기 형식을 지정합니다. 옵션은 **SAS for Windows/OS2**, **SAS for UNIX** 또는 **SAS 버전 7/8**입니다.

필드 이름 내보내기. SAS와 함께 사용하기 위해 IBM SPSS Modeler에서 필드 이름 및 레이블을 내보내려면 이 옵션을 선택하십시오.

- **이름 및 변수 레이블.** IBM SPSS Modeler 필드 이름 및 레이블을 모두 내보내려면 선택하십시오. 이름은 SAS 변수이름으로 내보내는 반면 레이블은 SAS 변수 레이블로 내보냅니다.
- **이름을 변수 레이블로 사용.** SAS에서 변수 레이블로 IBM SPSS Modeler 필드 이름을 사용하려면 선택하십시오. IBM SPSS Modeler를 사용하면 SAS 변수 이름에서는 유효하지 않은 문자를 필드 이름에서는 허용합니다. 유효하지 않은 SAS 이름이 작성될 가능성을 방지하려면 대신 **이름 및 변수 레이블**을 선택하십시오.

현재 데이터의 입력 노드 생성. 내보낸 데이터 파일을 읽을 SAS 소스 노드를 자동으로 생성하려면 이 옵션을 선택하십시오. 자세한 정보는 45 페이지의 『SAS 소스 노드』의 내용을 참조하십시오.

Excel 내보내기 노드

Excel 내보내기 노드는 Microsoft Excel .xlsx 형식으로 데이터를 출력합니다. 선택적으로 자동으로 Excel을 시작한 후 노드가 실행될 때 내보낸 파일을 열도록 선택할 수 있습니다.

Excel 노드 내보내기 탭

파일 이름. 파일 이름을 입력하거나 파일 선택기 단추를 클릭하여 파일의 위치로 이동하십시오. 기본 파일 이름은 *excelexp.xlsx*입니다.

파일 유형. Excel .xlsx 파일 형식이 지원됩니다.

새 파일 작성. 새 Excel 파일을 작성합니다.

기존 파일에 삽입. 컨텐츠는 **셀에서 시작** 필드에 의해 지정된 셀에서 시작하여 대체됩니다. 스프레드시트의 기타 셀은 원래 컨텐츠로 남아 있습니다.

필드 이름 포함. 필드 이름이 워크시트의 첫 번째 행에 포함되는지 여부를 지정합니다.

셀에서 시작. 첫 번째 내보내기 레코드(**필드 이름 포함**이 선택된 경우에는 첫 번째 필드 이름)에 사용되는 셀 위치입니다. 데이터는 오른쪽까지 이 초기 셀에서 아래로 채워집니다.

워크시트 선택. 데이터를 내보낼 워크시트를 지정합니다. 인덱스별 또는 이름별로 워크시트를 식별할 수 있습니다.

- **인덱스별.** 새 파일을 작성하는 경우 0에서 9까지의 숫자를 지정하여 내보낼 워크시트를 식별하십시오(첫 번째 워크시트의 경우 0으로 시작하고 두 번째 워크시트의 경우 1로 시작하는 방식임). 워크시트가 이미 이 위치에 있는 경우에만 10 이상의 값을 사용할 수 있습니다.
- **이름별.** 새 파일을 작성하는 경우 워크시트에 사용되는 이름을 지정하십시오. 기존 파일에 삽입하는 경우 이 워크시트가 있으면 이 워크시트에 데이터가 삽입되고 이 워크시트가 없으면 이 이름을 가진 새 워크시트가 작성됩니다.

Excel 시작. 노드가 실행될 때 내보낸 파일에 대해 Excel이 자동으로 시작되는지를 지정합니다. IBM SPSS Modeler Server에 대해 분산 모드에서 실행 중인 경우 출력은 서버 파일 시스템에 저장되고 Excel은 내보낸 파일의 사본을 사용하여 클라이언트에서 시작됩니다.

현재 데이터의 입력 노드 생성. 내보낸 데이터 파일을 읽을 Excel 소스 노드를 자동으로 생성하려면 이 옵션을 선택하십시오. 자세한 정보는 46 페이지의 『Excel 소스 노드』의 내용을 참조하십시오.

XML 내보내기 노드

XML 내보내기 노드에서는 UTF-8 인코딩을 사용하여 XML 형식의 데이터를 출력할 수 있습니다. 선택적으로 XML 소스 노드를 작성하여 내보내진 데이터를 다시 스트림으로 읽을 수 있습니다.

XML 내보내기 파일. 데이터를 내보낼 XML 파일의 전체 경로 및 파일 이름입니다.

XML 스키마 사용. 스키마 또는 DTD를 사용하여 내보내는 데이터의 구조를 제어하려면 이 선택란을 선택하십시오. 그러면 아래에 설명된 **맵핑** 단추가 활성화됩니다.

스키마 또는 DTD를 사용하지 않는 경우에는 내보내는 데이터에 다음과 같은 기본 구조가 사용됩니다.

```
<records>      <record>
          <fieldname1>value</fieldname1>
          <fieldname2>value</fieldname2>
       :
          <fieldnameN>value</fieldnameN>
     </record>
     <record>
       :
       :
     </record>
       :
       :
</records>
```

필드 이름에 있는 공백은 밑줄로 대체됩니다. 예를 들어, "My Field"는 <My_Field>가 됩니다.

맵핑. XML 스키마를 사용하기로 선택한 경우, 이 단추는 각각의 새 레코드를 시작하는 데 사용할 XML 구조 파트를 지정할 수 있는 대화 상자를 엽니다. 자세한 정보는 『XML 레코드 맵핑 옵션』의 내용을 참조하십시오.

맵핑된 필드. 맵핑된 필드 수를 표시합니다.

현재 데이터의 입력 노드 생성. 내보낸 데이터 파일을 스트림으로 다시 읽어오는 XML 소스 노드를 자동으로 생성하려면 이 옵션을 선택하십시오. 자세한 정보는 47 페이지의 『XML 소스 노드』의 내용을 참조하십시오.

XML 데이터 쓰기

XML 요소를 지정하면 요소 태그 안에 해당 필드 값이 배치됩니다.

<element>value</element>

속성을 맵핑하면 해당 필드 값은 속성의 값으로서 배치됩니다.

<element attribute="value">

필드를 <records> 요소 위에 있는 요소에 맵핑하는 경우, 해당 필드는 한 번만 쓰여지고 모든 레코드에 대한 하나의 상수가 됩니다. 이 요소의 값은 첫 번째 레코드에서 비롯됩니다.

빈 컨텐츠를 지정하면 널값이 쓰여집니다. 요소의 경우 다음과 같습니다.

<element></element>

속성의 경우 다음과 같습니다.

<element attribute="">

XML 레코드 맵핑 옵션

레코드 탭에서 각각의 새 레코드를 시작하는 데 사용할 XML 구조 파트를 지정할 수 있습니다. 스키마로 올바르게 맵핑하려면 레코드 구분자를 지정해야 합니다.

XML 구조. 이전 화면에서 지정한 XML 스키마의 구조를 보여주는 계층 구조 트리입니다.

레코드(XPath 표현식). 레코드 구분자를 설정하려면 XML 구조에서 요소를 선택하고 오른쪽 화살표 단추를 클릭하십시오. 소스 데이터에서 이 요소가 발견될 때마다 출력 파일에 새 레코드가 작성됩니다.

참고: XML 구조의 루트 요소를 선택하는 경우, 하나의 레코드만 쓸 수 있고 기타 모든 레코드는 건너뜁니다.

XML 필드 맵핑 옵션

필드 탭은 스키마 파일이 사용될 때 데이터 세트의 필드를 XML 구조의 요소 또는 속성으로 맵핑하는 데 사용됩니다.

요소 또는 속성 이름과 일치하는 필드 이름은 해당 요소 또는 속성 이름이 고유한 경우 자동으로 맵핑됩니다. 따라서 이름이 `field1`인 요소 및 속성이 모두 있으면 자동 맵핑이 수행되지 않습니다. 구조에 이름이 `field1`인 항목이 하나만 있으면 스트림에서 해당 이름을 갖는 필드는 자동으로 맵핑됩니다.

필드. 모델의 필드 목록입니다. 맵핑의 소스 파트로 하나 이상의 필드를 선택하십시오. 목록 맨 아래에 있는 단추를 사용하여 모든 필드를 선택하거나 특정 측정 수준을 갖는 모든 필드를 선택할 수 있습니다.

XML 구조. 맵핑 대상으로 사용할 XML 구조의 요소를 선택하십시오. 맵핑을 작성하려면 맵핑을 클릭하십시오. 그러면 맵핑이 표시됩니다. 이러한 방식으로 맵핑된 필드의 수가 이 목록 아래에 표시됩니다.

맵핑을 제거하려면 XML 구조 목록에서 해당 항목을 선택하고 **맵핑 해제**를 클릭하십시오.

속성 표시. XML 구조에 있는 XML 요소의 속성(있는 경우)을 표시하거나 숨깁니다.

XML 맵핑 미리보기

미리보기 탭에서 **업데이트**를 클릭하면 작성될 XML의 미리보기를 볼 수 있습니다.

맵핑이 올바르지 않은 경우, 레코드 또는 필드 탭으로 돌아가 오류를 정정하고 다시 **업데이트**를 클릭하여 결과를 확인하십시오.

제 8 장 IBM SPSS Statistics 노드

IBM SPSS Statistics 노드 - 개요

IBM SPSS Statistics에서는 IBM SPSS Modeler 및 데이터 마이닝 기능을 보완하기 위해 추가 통계 분석 및 데이터 관리를 수행할 수 있는 기능을 제공합니다.

IBM SPSS Statistics의 호환 가능한 라이센스가 있는 사본이 설치된 경우, 이를 IBM SPSS Modeler에서 연결하여 복잡한 다중 단계의 데이터 조작 및 분석을 수행할 수 있습니다. 연결되지 않은 경우에는 IBM SPSS Modeler에서 지원되지 않습니다. 고급 사용자의 경우, 명령문을 사용하여 분석을 추가적으로 수정할 수 있는 옵션도 있습니다. 버전 호환성에 대한 정보는 릴리스 정보를 참조하십시오.

사용 가능하면 IBM SPSS Statistics 노드가 노드 팔레트의 전용 부분에 표시됩니다.

참고: IBM SPSS Statistics 변환, 모델 또는 출력 노드를 사용하기 전에 유형 노드에서 데이터를 인스턴스화하도록 권장합니다. AUTORECODE 명령문을 사용하는 경우에도 이 요구사항이 적용됩니다.

IBM SPSS Statistics 팔레트에는 다음 노드가 포함됩니다.

통계량 파일 노드는 IBM SPSS Statistics에서 사용하는 *.sav* 또는 *.zsav* 파일 형식뿐 아니라 동일한 형식을 사용하는 IBM SPSS Modeler에 저장된 캐시 파일로부터 데이터를 읽습니다.

통계량 변환 노드는 IBM SPSS Modeler의 데이터 소스에 대해 IBM SPSS Statistics 구문 명령문의 선택을 실행합니다. 이 노드는 IBM SPSS Statistics의 라이센스 사본이 필요합니다.

통계량 모델 노드를 사용하면 PMML을 생성하는 IBM SPSS Statistics 프로시저를 실행하여 데이터를 분석하고 작업할 수 있습니다. 이 노드는 IBM SPSS Statistics의 라이센스 사본이 필요합니다.

통계량 출력 노드를 사용하면 IBM SPSS Statistics 프로시저를 호출하여 IBM SPSS Modeler 데이터를 분석할 수 있습니다. 광범위한 IBM SPSS Statistics 분석 프로시저를 사용할 수 있습니다. 이 노드는 IBM SPSS Statistics의 라이센스 사본이 필요합니다.

통계량 내보내기 노드는 IBM SPSS Statistics *.sav* 또는 *.zsav* 형식으로 데이터를 출력합니다. *.sav* 또는 *.zsav* 파일은 IBM SPSS Statistics Base 및 기타 제품에서 읽을 수 있습니다. 이것은 또한 IBM SPSS Modeler의 캐시 파일에 사용하는 형식입니다.

참고: SPSS Statistics의 사본이 단일 사용자에게만 라이센스가 적용되며 둘 이상의 분기가 있는 스트림을 실행하는 경우, 각각 SPSS Statistics 노드를 포함하므로 라이센스 오류가 발생할 수 있습니다. 이 오류는 한 분기의 SPSS Statistics 세션이 종료되지 않은 상태에서 다른 분기의 세션이 시작되려고 시도되는 경우에 발생할 수 있습니다. 가능하면 SPSS Statistics 노드가 있는 다중 분기가 병렬 형식으로 실행되지 않도록 스트림을 다시 설계하십시오.

통계량 파일 노드

통계량 파일 노드를 사용하여 저장된 IBM SPSS Statistics 파일(*.sav* 또는 *.zsav*)에서 직접 데이터를 읽을 수 있습니다. 이제 이 형식이 IBM SPSS Modeler의 이전 버전의 캐시 파일을 바꾸는 데 사용됩니다. 저장된 캐시 파일을 가져오려면 IBM SPSS Statistics 파일 노드를 사용해야 합니다.

파일 가져오기. 파일의 이름을 지정합니다. 파일 이름을 입력하거나 생략 기호 단추(**...**) 를 클릭하여 파일을 선택할 수 있습니다. 파일을 선택하고 나면 경로가 표시됩니다.

파일이 비밀번호로 암호화됨. 파일이 비밀번호로 암호화되고 있음을 알고 있으면 이 선택란을 선택하십시오. **비밀번호**를 입력하도록 프롬프트됩니다. 파일이 비밀번호로 보호되고 있으나 비밀번호를 입력하지 않으면 다른 탭으로 변경하거나 데이터를 새로 고치거나 노드 내용을 미리 보거나 노드를 포함한 스트림을 실행하는 것과 같은 시도를 할 때 경고 메시지가 표시됩니다.

참고: 비밀번호로 보호된 파일은 IBM SPSS Modeler 버전 16 이상에서만 열 수 있습니다.

변수이름. IBM SPSS Statistics *.sav* 또는 *.zsav* 파일에서 가져올 때 변수 이름 및 레이블을 처리하는 방법을 선택합니다. 여기에 포함시키려고 선택하는 메타데이터는 IBM SPSS Modeler 내의 작업 전체에서 지속되며 IBM SPSS Statistics에서 사용하도록 다시 내보낼 수 있습니다.

- **이름 및 레이블 읽기.** 변수 이름 및 레이블을 둘 다 IBM SPSS Modeler에 읽어들일 수 있습니다. 기본적으로 이 옵션이 선택되고 변수 이름이 유형 노드에 표시됩니다. 레이블은 스트림 특성 대화 상자에서 지정된 옵션에 따라 도표, 모델 브라우저 및 기타 유형의 출력에 표시될 수 있습니다. 기본적으로 출력에 레이블 표시는 사용되지 않습니다.
- **이름으로 레이블 읽기.** IBM SPSS Statistics *.sav* 또는 *.zsav* 파일에서 짧은 필드 이름이 아니라 설명 변수 레이블을 읽고 해당 레이블을 IBM SPSS Modeler에서 변수 이름으로 사용하려면 선택하십시오.

값. IBM SPSS Statistics *.sav* 또는 *.zsav* 파일에서 가져올 때 값 및 레이블을 처리하는 방법을 선택합니다. 여기에 포함시키려고 선택하는 메타데이터는 IBM SPSS Modeler 내의 작업 전체에서 지속되며 IBM SPSS Statistics에서 사용하도록 다시 내보낼 수 있습니다.

- **데이터 및 레이블 읽기.** 실제 값 및 값 레이블을 둘 다 IBM SPSS Modeler에 읽어들일 수 있습니다. 기본적으로 이 옵션이 선택되고 값 자체가 유형 노드에 표시됩니다. 값 레이블은 스트림 특성 대화 상자에서 지정된 옵션에 따라 표현식 작성기, 도표, 모델 브라우저 및 기타 유형의 출력에 표시될 수 있습니다.
- **데이터로 레이블 읽기.** 값을 표시하는 데 수치 또는 기호 코드가 아니라 *.sav* 또는 *.zsav* 파일의 값 레이블을 사용하려면 선택하십시오. 예를 들어, 실제로는 *남성* 및 *여성*을 나타내는 1 및 2 값을 가진 성별 필드가 있는 데이터에 이 옵션을 선택하면 각 필드를 문자열로 변환하고 실제값으로 *남성* 및 *여성*을 가져옵니다.

이 옵션을 선택하기 전에 사용자의 IBM SPSS Statistics 데이터에서 결측값을 고려하는 것이 중요합니다. 예를 들어, 수치 필드가 결측값에 대해서만 레이블을 사용하는 경우(0 = *응답 없음*, –99 = *알 수 없음*), 위 옵션을 선택하면 *응답 없음* 및 *알 수 없음*이라는 값 레이블만 가져오고 필드를 문자열로 변환합니다. 이 경우, 값 자체를 가져오고 유형 노드에서 결측값을 설정해야 합니다.

필드 형식 정보를 사용하여 저장 공간 판별. 이 선택란을 선택 취소하면 *.sav* 파일에서 정수로 형식화된 필드 값(즉, F*n*으로 지정된 필드. IBM SPSS Statistics의 변수 보기의 0)을 정수 저장 공간을 사용하여 가져옵니다. 문자열을 제외한 모든 기타 필드 값은 실제 수로 가져옵니다.

이 상자를 선택하면(기본값), *.sav* 파일에서 정수로 형식화되었는지 여부에 상관없이 문자열을 제외한 모든 필드 값을 실제 수로 가져옵니다.

다중 응답 세트. 파일을 가져올 때 IBM SPSS Statistics 파일에서 정의한 모든 다중 응답 세트는 자동으로 유지됩니다. 필터 탭의 모든 노드에서 다중 응답 세트를 보고 편집할 수 있습니다. 자세한 정보는 156 페이지의 『다중 응답 세트 편집』의 내용을 참조하십시오.

통계량 변환 노드

통계량 변환 노드를 사용하면 IBM SPSS Statistics 명령문을 사용하여 데이터 변환을 완료할 수 있습니다. 그러면 IBM SPSS Modeler에서 지원되지 않는 수많은 변환을 완료할 수 있으며 단일 노드에서 수많은 필드를 작성하는 것을 포함하여 복합적인 다중 단계 변환을 자동화할 수 있습니다. 이 노드는 통계량 출력 노드와 유사합니다. 단, 이 노드에서는 데이터가 추가 분석을 위해 IBM SPSS Modeler로 돌아가나 출력 노드에서는 데이터가 그래프 또는 표 등의 요청된 출력 개체로 돌아갑니다.

이 노드를 사용하려면 컴퓨터에 호환 가능한 IBM SPSS Statistics 버전이 설치되고 라이센스가 부여되어 있어야 합니다. 자세한 정보는 346 페이지의 『IBM SPSS Statistics 헬퍼 애플리케이션』의 내용을 참조하십시오. 호환성 정보에 대해서는 릴리스 정보를 참조하십시오.

필요할 경우 필터 탭을 사용하여 IBM SPSS Statistics 이름 지정 표준을 준수하도록 필드를 필터링하거나 필드의 이름을 바꿀 수 있습니다. 자세한 정보는 387 페이지의 『IBM SPSS Statistics에 대한 필드 이름 변경 또는 필터링』의 내용을 참조하십시오.

구문 참조. 특정 IBM SPSS Statistics 프로시저에 대한 자세한 내용은 IBM SPSS Statistics 소프트웨어의 사본과 함께 포함된 *IBM SPSS Statistics 명령 구문 참조* 안내서를 참조하십시오. 구문 탭에서 안내서를 보려면 **구문 편집기** 옵션을 선택한 후 IBM SPSS Statistics 구문 도움말 시작 단추를 클릭하십시오.

참고: 모든 IBM SPSS Statistics 명령문이 이 노드에 의해 지원되는 것은 아닙니다. 자세한 정보는 380 페이지의 『허용 가능한 명령문』의 내용을 참조하십시오.

통계량 변환 노드 - 명령문 탭

IBM SPSS Statistics 대화 상자 옵션

프로시저에 대한 IBM SPSS Statistics 구문에 익숙하지 않은 경우 IBM SPSS Modeler에서 구문을 작성하는 가장 단순한 방법은 **IBM SPSS Statistics 대화 상자** 옵션을 선택하고 프로시저에 대한 대화 상자를 선택하고 대화 상자를 완료한 후 확인을 클릭하는 것입니다. 이를 수행하면 사용자가 IBM SPSS Modeler에서 사용 중인 IBM SPSS Statistics 노드의 구문 탭에 해당 구문이 배치됩니다. 그러면 스트림을 실행하여 프로시저로부터 출력을 얻을 수 있습니다.

IBM SPSS Statistics 명령문 편집기 옵션

검사. 대화 상자의 상단에 명령문을 입력한 후에 이 단추를 사용하여 항목을 검증할 수 있습니다. 올바르지 않은 모든 명령문이 대화 상자의 하단에서 식별됩니다.

검사 프로세스에 시간이 너무 오래 걸리지 않도록 하려면 명령문을 검증할 때 전체 데이터 세트가 아니라 데이터의 대표 표본을 검사하여 항목이 올바른지 확인하십시오.

허용 가능한 명령문

IBM SPSS Statistics의 수많은 레거시 명령문을 갖고 있거나 IBM SPSS Statistics의 데이터 준비 기능에 익숙한 경우, 통계량 변환 노드를 사용하여 여러 가지 기존 변환을 실행할 수 있습니다. 노드의 안내에 따라 데이터를 예상 가능한 방법으로 변환할 수 있습니다. 예를 들어, 루프 명령문을 실행하거나 데이터를 변경, 추가, 정렬, 필터링 또는 선택할 수 있습니다.

실행할 수 있는 명령의 예

- 이항 분포에 따라 난수 계산:

```
COMPUTE newvar = RV.BINOM(10000,0.1)
```

- 변수를 새 변수로 다시 코딩:

```
RECODE Age (Lowest thru 30=1) (30 thru 50=2) (50 thru Highest=3) INTO AgeRecoded
```

- 결측값 대체:

```
RMV Age_1=SMEAN(Age)
```

통계량 변환 노드에 의해 지원되는 IBM SPSS Statistics 명령문은 아래에 나열됩니다.

명령문 이름

```
ADD VALUE LABELS
APPLY DICTIONARY
AUTORECODE
BREAK
CD
CLEAR MODEL PROGRAMS
CLEAR TIME PROGRAM
CLEAR TRANSFORMATIONS
COMPUTE
COUNT
CREATE
```

명령문 이름

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

MATRIX-END MATRIX

MISSING VALUES

N OF CASES

NUMERIC

PERMISSIONS

PRESERVE

RANK

RECODE

RENAME VARIABLES

RESTORE

RMV

SAMPLE

SELECT IF

SET

SORT CASES

STRING

명령문 이름

SUBTITLE

TEMPORARY

TITLE

UPDATE

V2C

VALIDATEDATA

VALUE LABELS

VARIABLE ATTRIBUTE

VARSTOCASES

VECTOR

통계량 모델 노드

통계량 모델 노드를 사용하면 PMML을 생성하는 IBM SPSS Statistics 프로시저를 실행하여 데이터를 분석하고 작업할 수 있습니다. 작성하는 모델 너깃은 IBM SPSS Modeler 스트림에서 스코어링 등에 일반적인 방법으로 사용될 수 있습니다.

이 노드를 사용하려면 컴퓨터에 호환 가능한 IBM SPSS Statistics 버전이 설치되고 라이센스가 부여되어 있어야 합니다. 자세한 정보는 346 페이지의 『IBM SPSS Statistics 헬퍼 애플리케이션』의 내용을 참조하십시오. 호환성 정보에 대해서는 릴리스 정보를 참조하십시오.

사용 가능한 IBM SPSS Statistics 분석 프로시저는 사용자가 가진 라이센스의 유형에 따라 다릅니다.

통계량 모델 노드 - 모델 탭

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

대화 상자를 선택하십시오. 선택하고 실행할 수 있는 사용 가능한 IBM SPSS Statistics 프로시저의 목록을 표시하려면 클릭하십시오. 목록에는 PMML을 생성하며 사용자가 라이센스를 갖고 있는 프로시저만 표시되며 사용자 작성 프로시저는 포함되지 않습니다.

1. 필수 프로시저를 클릭하십시오. 관련 IBM SPSS Statistics 대화 상자가 표시됩니다.
2. IBM SPSS Statistics 대화 상자에 프로시저에 대한 세부사항을 입력하십시오.
3. 통계량 모델 노드로 돌아가려면 **확인**을 클릭하십시오. IBM SPSS Statistics 명령문이 모델 탭에 표시됩니다.
4. 언제든지 IBM SPSS Statistics 대화 상자로 돌아가려면, 예를 들어, 쿼리를 수정하려면 프로시저 선택 단추의 오른쪽에 있는 IBM SPSS Statistics 대화 상자 표시 단추를 클릭하십시오.

통계량 모델 노드 - 모델 너깃 요약

통계량 모델 노드를 실행하면 연관된 IBM SPSS Statistics 프로시저가 실행되고 IBM SPSS Modeler 스트림에서 스코어링에 사용할 수 있는 모델 너깃이 생성됩니다.

모델 너깃의 요약 탭에서는 필드, 작성 설정, 모들 추정 프로세스에 대한 정보를 표시합니다. 결과는 특정 항목을 클릭하여 펼치거나 접을 수 있는 트리 보기로 표시됩니다.

모델 보기 단추는 IBM SPSS Statistics 출력 뷰어의 수정된 양식으로 결과를 표시합니다. 이 뷰어에 대한 자세한 정보는 IBM SPSS Statistics 문서를 참조하십시오.

일반적인 내보내기 및 인쇄 옵션은 파일 메뉴에서 사용할 수 있습니다. 자세한 정보는 305 페이지의 『출력 보기』의 내용을 참조하십시오.

통계량 출력 노드

통계량 출력 노드를 사용하면 IBM SPSS Statistics 프로시저를 호출하여 IBM SPSS Modeler 데이터를 분석할 수 있습니다. 브라우저 창에서 결과를 보거나 IBM SPSS Statistics 출력 파일 형식으로 결과를 저장할 수 있습니다. IBM SPSS Modeler에서 광범위한 IBM SPSS Statistics 분석 프로시저에 액세스할 수 있습니다.

이 노드를 사용하려면 컴퓨터에 호환 가능한 IBM SPSS Statistics 버전이 설치되고 라이센스가 부여되어 있어야 합니다. 자세한 정보는 346 페이지의 『IBM SPSS Statistics 헬퍼 애플리케이션』의 내용을 참조하십시오. 호환성 정보에 대해서는 릴리스 정보를 참조하십시오.

필요할 경우 필터 탭을 사용하여 IBM SPSS Statistics 이름 지정 표준을 준수하도록 필드를 필터링하거나 필드의 이름을 바꿀 수 있습니다. 자세한 정보는 387 페이지의 『IBM SPSS Statistics에 대한 필드 이름 변경 또는 필터링』의 내용을 참조하십시오.

구문 참조. 특정 IBM SPSS Statistics 프로시저에 대한 자세한 내용은 IBM SPSS Statistics 소프트웨어의 사본과 함께 포함된 *IBM SPSS Statistics 명령 구문 참조* 안내서를 참조하십시오. 구문 탭에서 안내서를 보려면 **구문 편집기** 옵션을 선택한 후 IBM SPSS Statistics 구문 도움말 시작 단추를 클릭하십시오.

통계량 출력 노드 - 명령문 탭

데이터 분석에 사용할 IBM SPSS Statistics 프로시저에 대한 명령문을 작성하려면 이 탭을 사용하십시오. 명령문은 두 부분, 즉, **명령문** 및 연관된 **옵션**으로 구성됩니다. 명령문은 수행될 분석 또는 작업 및 사용될 필드를 지정합니다. 옵션은 표시할 통계량, 저장할 파생 필드 등을 포함하여 그 밖의 모든 것을 지정합니다.

IBM SPSS Statistics 대화 상자 옵션

프로시저에 대한 IBM SPSS Statistics 구문에 익숙하지 않은 경우 IBM SPSS Modeler에서 구문을 작성하는 가장 단순한 방법은 **IBM SPSS Statistics 대화 상자** 옵션을 선택하고 프로시저에 대한 대화 상자를 선택

하고 대화 상자를 완료한 후 확인을 클릭하는 것입니다. 이를 수행하면 사용자가 IBM SPSS Modeler에서 사용 중인 IBM SPSS Statistics 노드의 구문 탭에 해당 구문이 배치됩니다. 그러면 스트림을 실행하여 프로시저로부터 출력을 얻을 수 있습니다.

필요에 따라 결과 데이터를 가져오는 데 사용할 통계량 파일 소스 노드를 생성할 수 있습니다. 예를 들어, 프로시저가 출력을 표시하는 것뿐만 아니라 활성 데이터 세트에 점수 등의 필드를 작성하는 데 유용합니다.

명령문을 작성하려면 다음을 수행하십시오.

1. **대화 상자 선택** 단추를 클릭하십시오.
2. 옵션 중 하나를 선택하십시오.
   - **분석.** IBM SPSS Statistics 분석 메뉴의 내용을 나열합니다. 사용할 프로시저를 선택하십시오.
   - **기타.** 표시되는 경우, IBM SPSS Statistics의 사용자 정의 대화 상자 작성기에 의해 작성된 대화 상자를 나열합니다. 또한 분석 메뉴에 표시되지 않으며 사용자가 라이센스를 가진 기타 IBM SPSS Statistics 대화 상자도 나열합니다. 적용 가능한 대화 상자가 없으면 이 옵션이 표시되지 않습니다.

참고: 자동 데이터 준비 대화 상자가 표시되지 않습니다.

새 필드를 작성하는 IBM SPSS Statistics 사용자 정의 대화 상자가 있는 경우, 통계량 출력 노드가 터미널 노드이므로 해당 필드를 IBM SPSS Modeler에서 사용할 수 없습니다.

결과 데이터를 또 다른 스트림으로 가져오는 데 사용할 수 있는 통계량 파일 소스 노드를 작성하려면 **결과 데이터에 대한 가져오기 노드 생성** 선택란을 선택하십시오. 노드는 **파일** 필드에서 지정된 *.sav* 파일에 포함된 데이터와 함께 화면 캔버스에 배치됩니다. 기본 위치는 IBM SPSS Modeler 설치 디렉토리입니다.

명령문 편집기 옵션

자주 사용되는 프로시저에 대해 작성된 명령문을 저장하려면 다음을 수행하십시오.

1. 파일 옵션 단추(도구 모음의 첫 번째 단추)를 클릭하십시오.
2. 메뉴에서 **저장** 또는 **다른 이름으로 저장**을 선택하십시오.
3. 파일을 *.sps* 파일로 저장하십시오.

이전에 작성한 명령문 파일을 사용하려면 명령문 편집기의 현재 내용을 바꾸십시오.

1. 파일 옵션 단추(도구 모음의 첫 번째 단추)를 클릭하십시오.
2. 메뉴에서 **열기**를 선택하십시오.
3. 내용을 출력 노드 명령문 탭에 붙여넣으려면 *sps* 파일을 선택하십시오.

현재 내용을 바꾸지 않고 이전에 저장한 명령문을 삽입하려면 다음을 수행하십시오.

1. 파일 옵션 단추(도구 모음의 첫 번째 단추)를 클릭하십시오.
2. 메뉴에서 **삽입**을 선택하십시오.
3. 커서에 의해 지정된 위치에서 출력 노드에 내용을 붙여넣으려면 *sps* 파일을 선택하십시오.

결과 데이터를 또 다른 스트림으로 가져오는 데 사용할 수 있는 통계량 파일 소스 노드를 작성하려면 **결과 데이터에 대한 가져오기 노드 생성** 선택란을 선택하십시오. 노드는 **파일** 필드에서 지정된 *.sav* 파일에 포함된 데이터와 함께 화면 캔버스에 배치됩니다. 기본 위치는 IBM SPSS Modeler 설치 디렉토리입니다.

실행을 클릭하면 결과가 IBM SPSS Statistics 출력 뷰어에 표시됩니다. 뷰어에 대한 자세한 정보는 IBM SPSS Statistics 문서를 참조하십시오.

통계량 출력 노드 - 출력 탭

출력 탭을 사용하면 출력의 형식 및 위치를 지정할 수 있습니다. 화면에 결과를 표시하거나 결과를 사용 가능한 파일 유형 중 하나로 전송할 수 있습니다.

출력 이름. 노드가 실행될 때 생성되는 출력의 이름을 지정합니다. **자동**은 출력을 생성하는 노드를 기반으로 이름을 선택합니다. 필요에 따라 **사용자 정의**를 선택하여 다른 이름을 지정할 수 있습니다.

화면에 출력(기본값)**.** 온라인으로 볼 출력 오브젝트를 생성합니다. 출력 오브젝트는 출력 노드가 실행될 때 관리자 창의 출력 탭에 표시됩니다.

파일로 출력. 노드를 실행할 때 출력을 파일에 저장합니다. 이 옵션을 선택하는 경우, **파일 이름** 필드에 파일 이름을 입력하거나 디렉토리로 이동하여 파일 선택자 단추를 사용하여 파일 이름을 지정한 다음 파일 유형을 선택하십시오.

파일 유형. 출력을 전송할 파일 유형을 선택하십시오.

- **HTML 문서(*.html).** HTML 형식으로 출력을 작성합니다.
- **IBM SPSS Statistics 뷰어 파일(*.spv).** IBM SPSS Statistics 출력 뷰어가 읽을 수 있는 형식으로 출력을 작성합니다.
- **IBM SPSS Statistics 웹 보고서 파일(*.spw).** IBM SPSS Statistics 웹 보고서 형식으로 출력을 작성하며 IBM SPSS Collaboration and Deployment Services 리포지토리에 공개할 수 있으며 결과적으로 웹 브라우저에서 볼 수 있습니다. 자세한 정보는 305 페이지의 『웹에 출판』의 내용을 참조하십시오.

참고: **화면에 출력**을 선택하면 IBM SPSS Statistics OMS 지시문 `VIEWER=NO`이 적용되지 않습니다. 또한 API 스크립트(*Basic* 및 *Python SpssClient* 모듈)를 IBM SPSS Modeler에서 사용할 수 없습니다.

통계량 내보내기 노드

통계량 내보내기 노드를 사용하면 IBM SPSS Statistics *.sav* 형식으로 데이터를 내보낼 수 있습니다. IBM SPSS Statistics *.sav* 파일은 IBM SPSS Statistics 기본 및 기타 모듈에서 읽을 수 있습니다. 이 형식은 IBM SPSS Modeler 캐시 파일에도 사용되는 형식입니다.

IBM SPSS Statistics 변수 이름은 64자로 제한되어 있으며 특정 문자(공백, 달러 기호($), 대시(–) 등)를 포함할 수 없으므로 IBM SPSS Modeler 필드 이름을 IBM SPSS Statistics 변수 이름에 맵핑하면 오류가 발생할 수 있습니다. 이러한 제한에 맞게 조정하는 두 가지 방법이 있습니다.

- 필터 탭을 클릭하여 IBM SPSS Statistics 변수 이름에 요구 사항에 맞게 필드의 이름을 변경할 수 있습니다. 자세한 정보는 387 페이지의 『IBM SPSS Statistics에 대한 필드 이름 변경 또는 필터링』의 내용을 참조하십시오.
- IBM SPSS Modeler에서 필드 이름 및 레이블을 모두 내보내려면 선택하십시오.

참고: IBM SPSS Modeler는 유니코드 UTF-8 형식으로 *.sav* 파일을 작성합니다. IBM SPSS Statistics는 릴리스 16.0부터는 유니코드 UTF-8 형식의 파일만 지원합니다. 데이터 손상을 방지하기 위해 유니코드 인코딩으로 저장된 *.sav* 파일을 IBM SPSS Statistics 16.0 이전 버전에서 사용하지 마십시오. 자세한 정보는 IBM SPSS Statistics 도움말을 참조하십시오.

다중 응답 세트. 파일을 내보낼 때 스트림에서 정의한 모든 다중 응답 세트는 자동으로 유지됩니다. 필터 탭의 모든 노드에서 다중 응답 세트를 보고 편집할 수 있습니다. 자세한 정보는 156 페이지의 『다중 응답 세트 편집』의 내용을 참조하십시오.

통계량 내보내기 노드 - 내보내기 탭

파일 내보내기. 파일의 이름을 지정합니다. 파일 이름을 입력하거나 파일 선택기 단추를 클릭하여 파일 위치를 찾아보십시오.

파일 유형. 파일이 일반적인 *.sav* 또는 압축된 *.zsav* 형식으로 저장된 경우에 선택하십시오.

비밀번호로 파일 암호화. 비밀번호로 파일을 보호하려면 이 선택란을 선택하십시오. 별도의 대화 상자에 **비밀번호**를 입력하고 확인하도록 프롬프트됩니다.

참고: 비밀번호로 보호된 파일은 IBM SPSS Modeler 버전 16 이상 또는 IBM SPSS Statistics 버전 21 이상에서만 열 수 있습니다.

필드 이름 내보내기. IBM SPSS Modeler에서 IBM SPSS Statistics *.sav* 또는 *.zsav* 파일로 내보낼 때 변수 이름 및 레이블을 처리하는 방법을 지정합니다.

- **이름 및 변수 레이블.** IBM SPSS Modeler 필드 이름 및 레이블을 모두 내보내려면 선택하십시오. 이름은 IBM SPSS Statistics 변수이름으로 내보내는 반면 레이블은 IBM SPSS Statistics 변수 레이블로 내보냅니다.
- **이름을 변수 레이블로 사용.** IBM SPSS Statistics에서 변수 레이블로 IBM SPSS Modeler 필드 이름을 사용하려면 선택하십시오. IBM SPSS Modeler를 사용하면 IBM SPSS Statistics 변수 이름에서는 유효하지 않은 문자를 필드 이름에서는 허용합니다. 유효하지 않은 IBM SPSS Statistics 이름이 작성될 가능성을 방지하려면 대신 **이름을 변수 레이블로 사용**을 선택하십시오. 또는 필터 탭을 사용하여 필드 이름을 조정하십시오.

애플리케이션 시작. IBM SPSS Statistics가 사용자의 컴퓨터에 설치되어 있으면 이 옵션을 선택하여 저장된 데이터 파일에 대해 직접 애플리케이션을 호출할 수 있습니다. 애플리케이션 시작에 필요한 옵션이 헬퍼 애플리케이션 대화 상자에서 지정되어야 합니다. 자세한 정보는 346 페이지의 『IBM SPSS Statistics 헬퍼 애플리케이션』의 내용을 참조하십시오. 외부 프로그램을 열지 않고 간단하게 IBM SPSS Statistics *.sav* 또는 *.zsav* 파일을 작성하려면 이 옵션을 선택 취소하십시오.

현재 데이터의 입력 노드 생성. 내보낸 데이터 파일을 읽을 통계량 소스 노드를 자동으로 생성하려면 이 옵션을 선택하십시오. 자세한 정보는 378 페이지의 『통계량 파일 노드』의 내용을 참조하십시오.

IBM SPSS Statistics에 대한 필드 이름 변경 또는 필터링

IBM SPSS Modeler에서 IBM SPSS Statistics 등의 외부 애플리케이션으로 데이터를 내보내거나 배포하기 전에 필드 이름을 변경하거나 조정해야 하는 경우가 있습니다. 통계량 변환, 통계량 출력 및 통계량 내보내기 대화 상자에는 이 프로세스를 활용하는 데 필요한 필터 탭이 포함되어 있습니다.

필터 탭 기능에 대한 기본적인 설명은 다른 위치에서 설명합니다. 자세한 정보는 154 페이지의 『필터링 옵션 설정』 주제를 참조하십시오. 이 주제는 IBM SPSS Statistics로 데이터를 읽어들이는 데 관한 팁을 제공합니다.

IBM SPSS Statistics 이름 지정 규칙을 준수하도록 필드 이름을 조정하려면 다음을 수행하십시오.

1. 필터 탭에서 필터 옵션 메뉴 도구 모음 단추(도구 모음 중 첫 번째 도구)를 클릭하십시오.
2. IBM SPSS Statistics에 대한 이름 변경을 선택하십시오.
3. IBM SPSS Statistics에 대한 이름 변경 대화 상자에서 파일 이름의 유효하지 않은 문자를 **해시(#)** 문자 또는 **밑줄(_)**로 바꿀 수 있습니다.

다중 응답 세트 이름 변경. 통계량 파일 소스 노드를 사용하여 IBM SPSS Modeler로 가져올 수 있는 다중 응답 세트의 이름을 조정하려면 이 옵션을 선택하십시오. 설문조사 응답과 같이 각 케이스에 대해 둘 이상의 값을 가질 수 있는 데이터를 기록하는 데 사용됩니다.

제 9 장 수퍼노드

수퍼노드 개요

IBM SPSS Modeler 비주얼 프로그래밍 인터페이스가 배우기 쉬운 이유 중 하나는 각 노드가 명확하게 정의된 기능을 가진다는 것입니다. 그러나 복잡한 처리의 경우 긴 노드 시퀀스가 필요할 수도 있습니다. 이로 인해 결국 스트림 캔버스가 어수선해지고 스트림 다이어그램을 이해하기 어려워질 수 있습니다. 두 가지 방법을 사용하여 길고 복잡한 스트림으로 인한 혼잡을 피할 수 있습니다.

- 처리 시퀀스를 여러 스트림(하나를 다른 하나에 공급)으로 분할할 수 있습니다. 예를 들어, 첫 번째 스트림은 두 번째 스트림이 입력으로 사용하는 데이터 파일을 작성합니다. 두 번째 스트림은 세 번째 스트림이 입력으로 사용하는 파일을 작성하며, 계속해서 이와 같이 반복됩니다. 이러한 여러 스트림을 하나의 **프로젝트**에 저장하여 이들을 관리할 수 있습니다. 프로젝트는 여러 스트림과 해당 출력을 위한 조직을 제공합니다. 그러나 프로젝트 파일에는 포함되는 오브젝트에 대한 참조만 포함되며, 따라서 여전히 관리해야 할 다수의 스트림 파일이 있습니다.
- 복잡한 스트림 프로세스에 대해 작업할 때 보다 간소화된 대안으로서 **수퍼노드**를 작성할 수 있습니다.

수퍼노드는 데이터 스트림의 섹션을 캡슐화하여 여러 노드를 하나의 노드로 그룹화합니다. 이는 데이터 마이너에게 여러 가지 이점을 제공합니다.

- 스트림이 더 깔끔하고 관리하기가 더 쉽습니다.
- 노드를 하나의 비즈니스별 수퍼노드로 결합할 수 있습니다.
- 여러 데이터 마이닝 프로젝트에서 재사용할 수 있도록 수퍼노드를 라이브러리로 내보낼 수 있습니다.

수퍼노드 유형

수퍼노드는 데이터 스트림에서 별 아이콘으로 표시됩니다. 아이콘은 음영 처리되어 수퍼노드의 유형과 스트림이 흐르는 방향을 나타냅니다.

세 가지 유형의 수퍼노드가 있습니다.

- 소스 수퍼노드
- 프로세스 수퍼노드
- 터미널 수퍼노드

소스 수퍼노드

소스 수퍼노드는 보통 소스 노드처럼 데이터 소스를 포함하며, 보통 소스 노드를 사용할 수 있는 곳이면 어디서나 사용할 수 있습니다. 소스 수퍼노드의 왼쪽이 음영 처리되어 왼쪽이 "닫혀" 있고 수퍼노드*에서* 아래로 데이터가 흘러야 함을 표시합니다.

소스 수퍼노드는 연결점이 오른쪽에 하나만 있으며, 이는 데이터가 해당 수퍼노드에서 나와 스트림으로 흐름을 표시합니다.

프로세스 수퍼노드

프로세스 수퍼노드는 프로세스 노드만 포함하며, 이 유형의 수퍼노드로 데이터가 *들어올* 수 있을 뿐만 아니라 이 유형의 수퍼노드에서 데이터가 *나갈* 수 있음을 표시하기 위해 음영 처리되지 않습니다.

프로세스 수퍼노드는 왼쪽과 오른쪽 모두에 연결점이 있어서 데이터가 수퍼노드로 들어가고 여기서 나와 다시 스트림으로 흐름을 보여줍니다. 수퍼노드는 추가 스트림 단편과 추가 스트림까지 포함할 수 있지만, 양 연결점은 *시작 스트림*과 *끝 스트림* 지점을 연결하는 단일 경로를 통해 흘러야 합니다.

참고: 프로세스 수퍼노드는 *조작 수퍼*노드라고도 합니다.

터미널 수퍼노드

터미널 수퍼노드는 하나 이상의 터미널 노드(plot, 테이블 등)를 포함하며 터미널 노드와 동일한 방식으로 사용할 수 있습니다. 터미널 수퍼노드는 오른쪽이 음영 처리되어 오른쪽이 "닫혀" 있고 터미널 수퍼노드로만 데이터가 흐를 수 있음을 표시합니다.

터미널 수퍼노드는 연결점이 왼쪽에 하나만 있으며, 이는 데이터가 스트림에서 수퍼노드로 들어가 해당 수퍼노드 내에서 종료됨을 표시합니다.

터미널 수퍼노드에는 수퍼노드 내에 있는 모든 터미널 노드의 실행 순서를 지정하는 데 사용되는 스크립트도 포함될 수 있습니다. 자세한 정보는 396 페이지의 『수퍼노드 및 스크립팅』의 내용을 참조하십시오.

수퍼 노드 작성

수퍼노드를 작성하면 여러 노드가 하나의 노드로 캡슐화되므로 데이터 스트림이 "수축"됩니다. 캔버스에서 스트림을 작성하거나 로드한 후 여러 가지 방식으로 수퍼노드를 작성할 수 있습니다.

다중 선택

수퍼노드를 작성하는 가장 쉬운 방법은 캡슐화할 노드를 모두 선택하는 것입니다.

1. 마우스를 사용하여 스트림 캔버스에서 여러 노드를 선택하십시오. Shift-클릭을 사용하여 스트림 또는 스트림 섹션을 선택할 수도 있습니다. *참고*: 선택하는 노드는 연속 또는 갈라진 스트림의 노드여야 합니다. 인접하지 않거나 어떤 방식으로든 연결되지 않은 노드는 선택할 수 없습니다.
2. 다음 세 가지 방법 중 하나를 사용하여 선택된 노드를 캡슐화하십시오.
   - 도구 모음에서 수퍼노드 아이콘(별 모양과 유사)을 클릭하십시오.
   - 수퍼노드를 마우스 오른쪽 단추로 클릭하고 컨텍스트 메뉴에서 다음을 선택하십시오.

     수퍼노드 작성 > 선택항목에서
   - 수퍼노드 메뉴에서 다음을 선택하십시오.

수퍼노드 작성 > 선택항목에서

이 세 옵션 모두 노드를 하나의 수퍼노드로 캡슐화하며, 수퍼노드는 해당 컨텐츠를 기반으로 해당 유형(소스, 프로세스 또는 터미널)을 반영하도록 음영 처리됩니다.

단일 선택

단일 노드를 선택하고 메뉴 옵션으로 수퍼노드의 시작 및 끝을 정하거나 선택된 노드의 모든 다운스트림 노드를 캡슐화하여 수퍼노드를 작성할 수도 있습니다.

1. 캡슐화의 시작을 결정하는 노드를 클릭하십시오.
2. 수퍼노드 메뉴에서 다음을 선택하십시오.

수퍼노드 작성 > 여기에서

스트림 섹션의 시작 및 끝을 선택하여 노드를 캡슐화하면 보다 대화식으로 수퍼노드를 작성할 수 있습니다.

1. 수퍼노드에 포함시킬 첫 번째 또는 마지막 노드를 클릭하십시오.
2. 수퍼노드 메뉴에서 다음을 선택하십시오.

수퍼노드 작성 > ...선택

3. 또는 원하는 노드를 마우스 오른쪽 단추로 클릭하여 컨텍스트 메뉴 옵션을 사용할 수 있습니다.
4. 커서가 수퍼노드 아이콘으로 바뀌어 스트림의 다른 지점을 선택해야 함을 표시합니다. 위 또는 아래로 움직여 수퍼노드 단편의 "다른 쪽 끝"으로 이동한 후 노드를 클릭하십시오. 그러면 그 사이에 있는 모든 노드가 수퍼노드 별 아이콘으로 바뀝니다.

참고: 선택하는 노드는 연속 또는 갈라진 스트림의 노드여야 합니다. 인접하지 않거나 어떤 방식으로든 연결되지 않은 노드는 선택할 수 없습니다.

수퍼노드 중첩

수퍼노드는 다른 수퍼노드 내에 중첩시킬 수 있습니다. 중첩된 수퍼노드에는 각 수퍼노드 유형(소스, 프로세스 및 터미널)에 적용되는 것과 동일한 규칙이 적용됩니다. 예를 들어, 중첩을 포함하는 프로세스 수퍼노드가 프로세스 수퍼노드로 유지되려면 중첩된 모든 수퍼노드를 통과하는 연속된 데이터 플로우가 있어야 합니다. 중첩된 수퍼노드 중 하나가 터미널이면 데이터는 더 이상 해당 계층 구조를 통해 흐르지 않습니다.

터미널 및 소스 수퍼노드는 다른 유형의 중첩된 수퍼노드를 포함할 수 있지만, 수퍼노드 작성에 적용되는 기본 규칙과 동일한 규칙이 적용됩니다.

수퍼노드 잠금

수퍼노드를 작성한 후에는 수퍼노드가 수정되지 않도록 비밀번호를 사용하여 수퍼노드를 잠글 수 있습니다. 예를 들어, IBM SPSS Modeler 인콰이어리 설정 경험이 적은 조직의 다른 사용자가 사용할 수 있도록 고정값 템플리트로서 스트림 또는 스트림 파트를 작성하는 경우에 이를 수행할 수 있습니다.

수퍼노드가 잠긴 경우에도 사용자는 계속 매개변수 탭에서 정의된 매개변수의 값을 입력할 수 있으며, 비밀번호를 입력하지 않고 잠긴 수퍼노드를 실행할 수 있습니다.

참고: 스크립트를 사용하여 잠금 및 잠금 해제를 수행할 수 없습니다.

수퍼노드 잠금 및 잠금 해제

경고: 분실한 비밀번호는 복구할 수 없습니다.

세 탭 중 하나에서 수퍼노드를 잠그거나 잠금 해제할 수 있습니다.

1. **노드 잠금**을 클릭하십시오.
2. 비밀번호를 입력하고 확인하십시오.
3. **확인**을 클릭하십시오.

비밀번호가 보호되는 수퍼노드는 스트림 캔버스에서 수퍼노드 아이콘의 맨 위 왼쪽에 작은 자물쇠 기호로 식별됩니다.

수퍼노드 잠금 해제

1. 비밀번호 보호를 영구적으로 제거하려면 **노드 잠금 해제**를 클릭하십시오. 비밀번호 입력을 요구하는 프롬프트가 표시됩니다.
2. 비밀번호를 입력하고 **확인**을 클릭하십시오. 해당 수퍼노드의 비밀번호가 더 이상 보호되지 않고 스트림에서 해당 아이콘 옆에 자물쇠 기호가 더 이상 표시되지 않습니다.

잠긴 수퍼노드 편집

매개변수를 정의하거나 확대하여 잠긴 수퍼노드를 표시하려 하면 비밀번호 입력을 요구하는 프롬프트가 표시됩니다.

비밀번호를 입력하고 **확인**을 클릭하십시오.

이제 해당 수퍼노드가 있는 스트림을 닫을 때까지 필요할 때마다 매개변수 정의를 편집하고 수퍼노드를 확대/축소할 수 있습니다.

이 조치로 인해 비밀번호 보호가 제거되지는 않습니다. 단지 수퍼노드에 액세스하여 관련 작업을 수행할 수만 있습니다. 자세한 정보는 『수퍼노드 잠금 및 잠금 해제』의 내용을 참조하십시오.

수퍼노드 편집

수퍼노드를 작성한 후에는 수퍼노드를 확대하여 보다 자세하게 검토할 수 있습니다. 수퍼노드가 잠겨 있으면 비밀번호 입력을 요구하는 프롬프트가 표시됩니다. 자세한 정보는 『잠긴 수퍼노드 편집』의 내용을 참조하십시오.

수퍼노드의 컨텐츠를 보려면 IBM SPSS Modeler 도구 모음의 확대 아이콘을 사용하거나 다음 방법을 사용할 수 있습니다.

1. 수퍼노드를 마우스 오른쪽 단추로 클릭하십시오.
2. 컨텍스트 메뉴에서 **확대**를 선택하십시오.

조금 다른 IBM SPSS Modeler 환경에서 선택한 수퍼노드의 컨텐츠가 표시되며, 스트림 또는 스트림 단편을 통한 데이터의 흐름을 나타내는 커넥터도 함께 표시됩니다. 이 수준에서는 스트림 캔버스에서 다음 태스크를 수행할 수 있습니다.

- 수퍼노드 유형(소스, 프로세스 또는 터미널)을 수정할 수 있습니다.
- 매개변수를 작성하거나 매개변수의 값을 편집할 수 있습니다. 매개변수는 스크립팅과 CLEM 표현식에서 사용됩니다.
- 수퍼노드와 해당 하위 노드에 대해 캐싱 옵션을 지정할 수 있습니다.
- 수퍼노드 스크립트를 작성하거나 수정할 수 있습니다(터미널 수퍼노드에만 해당).

수퍼노드 유형 수정

일부 환경에서는 수퍼노드의 유형을 변경하는 것이 유용합니다. 이 옵션은 수퍼노드를 확대한 경우에만 사용할 수 있으며, 해당 수준에서는 이 옵션이 해당 수퍼노드에만 적용됩니다. 다음 표에서는 세 가지 유형의 수퍼노드를 설명합니다.

표 46. 수퍼노드 유형.

수퍼노드 유형	설명
소스 수퍼노드	나가는 하나의 연결
프로세스 수퍼노드	두 개의 연결: 들어오는 연결과 나가는 연결
터미널 수퍼노드	들어오는 하나의 연결

수퍼노드의 유형을 변경하려면 다음을 수행하십시오.

1. 수퍼노드를 확대해야 합니다.
2. 수퍼노드 메뉴에서 **수퍼노드 유형**을 선택한 후 유형을 선택하십시오.

수퍼노드 주석(Annotation) 작성 및 이름 바꾸기

스트림에서 표시되는 수퍼노드의 이름을 바꾸고 프로젝트 또는 보고서에서 사용되는 주석(Annotation)을 작성할 수 있습니다. 이러한 특성에 액세스하려면 다음을 수행하십시오.

- 수퍼노드를 마우스 오른쪽 단추로 클릭하고(축소) **이름 변경 및 주석달기**를 선택하십시오.
- 또는 수퍼노드 메뉴에서 **이름 변경 및 주석달기**를 선택하십시오. 이 옵션은 확대 및 축소 모드 둘 다에서 사용할 수 있습니다.

두 경우 모두 주석(Annotation) 탭이 선택된 대화 상자가 열립니다. 여기에 있는 옵션을 사용하여 스트림 캔버스에 표시되는 이름을 사용자 정의하고 수퍼노드 조작에 관한 문서를 제공하십시오.

수퍼노드에서 주석 사용

주석이 달린 노드 또는 너깃에서 수퍼노드를 작성하는 경우, 수퍼노드에 주석이 표시되도록 하려면 수퍼노드를 작성하기 위한 선택항목에 주석을 포함시켜야 합니다. 선택항목에서 주석을 생략하면 수퍼노드를 작성할 때 주석이 스트림에 남에 있습니다.

주석을 포함한 수퍼노드를 펼치면 주석은 수퍼노드를 작성하기 전에 있었던 위치로 복귀합니다.

주석이 달린 오브젝트가 포함된 수퍼노드를 펼치는 경우, 주석이 수퍼노드에 포함되지 않았으면 오브젝트는 원래 위치로 복귀하지만 주석은 다시 첨부되지 않습니다.

수퍼 노드 모수

IBM SPSS Modeler에서는 사용자 정의 변수(예: *Minvalue*)를 설정할 수 있으며, 스크립팅 또는 CLEM 표현식에서 사용할 때 해당 값을 지정할 수 있습니다. 이러한 변수를 **매개변수**라고 합니다. 스트림, 세션 및 수퍼노드의 매개변수를 설정할 수 있습니다. 수퍼노드에 대해 설정된 매개변수는 해당 수퍼노드 또는 중첩된 노드에서 CLEM 표현식을 작성할 때 사용할 수 있습니다. 중첩된 수퍼노드에 대해 설정된 매개변수는 해당 상위 수퍼노드에서 사용할 수 없습니다.

두 단계로 수퍼노드의 매개변수를 작성하고 설정할 수 있습니다.

1. 수퍼노드의 매개변수를 정의합니다.
2. 그런 다음, 수퍼노드의 각 매개변수의 값을 지정합니다.

그러면 캡슐화 노드의 CLEM 표현식에서 이러한 매개변수를 사용할 수 있습니다.

수퍼노드 매개변수 정의

확대 및 축소 모드 둘 다에서 수퍼노드의 매개변수를 정의할 수 있습니다. 정의된 매개변수는 모든 캡슐화 노드에 적용됩니다. 수퍼노드의 매개변수를 정의하려면 먼저 수퍼노드 대화 상자의 매개변수 탭에 액세스해야 합니다. 다음 방법 중 하나를 사용하여 대화 상자를 여십시오.

- 스트림에서 수퍼노드를 두 번 클릭하십시오.
- 수퍼노드 메뉴에서 **매개변수 설정**을 선택하십시오.
- 또는 수퍼노드를 확대한 경우 컨텍스트 메뉴에서 **매개변수 설정**을 선택하십시오.

대화 상자를 열면 매개변수 탭이 이전에 정의한 매개변수와 함께 표시됩니다.

새 매개변수 정의

매개변수 정의 단추를 클릭하여 대화 상자를 여십시오.

이름. 모수 이름이 여기 나열됩니다. 이 필드에 이름을 입력하여 새 모수를 작성할 수 있습니다. 예를 들어, 최저 기온에 대한 모수를 작성하려면 `minvalue`를 입력할 수 있습니다. CLEM 표현식에서 모수를 나타내는 `$P-` 접두문자를 포함시키지 마십시오. 이 이름은 CLEM 표현식 작성기에 표시하는 데도 사용합니다.

긴 이름. 작성된 각 모수에 대한 설명 이름을 나열합니다.

저장 공간. 목록에서 저장 유형을 선택하십시오. 저장 공간은 모수에 데이터 값이 저장되는 방법을 표시합니다. 예를 들어, 유지할 선행 0이 포함된 값(예: 008)에 대한 작업 시 저장 유형으로 **문자열**을 선택해야 합니다. 그렇지 않으면, 값에서 0이 제거됩니다. 사용 가능한 저장 유형은 문자열, 정수, 실수, 시간, 날짜, 시간소인입니다. 날짜 모수의 경우, 다음 단락에 표시된 대로 ISO 표준 표기법을 사용하여 값을 지정해야 합니다.

값. 각 모수의 현재 값을 나열합니다. 필요에 따라 모수를 조정하십시오. 데이터 모수의 경우, ISO 표준 표기법(즉, `YYYY-MM-DD`)으로 값을 지정해야 합니다. 다른 형식으로 지정된 날짜는 허용되지 않습니다.

유형(선택사항). 외부 애플리케이션에 스트림을 배포할 계획이면 목록에서 측정 수준을 선택하십시오. 그렇지 않으면 *유형* 열을 있는 그대로 두는 것이 바람직합니다. 모수의 값 제한조건(예: 숫자 범위의 상한 및 하한)을 지정하려면 목록에서 **지정**을 선택하십시오.

사용자 인터페이스를 통해서만 모수에 대해 긴 이름, 저장 공간, 유형 옵션을 설정할 수 있습니다. 이러한 옵션은 스크립트를 사용하여 설정할 수 없습니다.

선택된 모수를 사용 가능한 모수 목록 위, 아래로 추가로 이동하려면 오른쪽에 있는 화살표를 선택하십시오. 선택된 모수를 제거하려면 삭제 단추(*X*로 표시)를 사용하십시오.

수퍼노드 매개변수의 값 설정

수퍼노드의 매개변수를 정의한 후에는 CLEM 표현식 또는 스크립트에서 매개변수를 사용하여 값을 지정할 수 있습니다.

수퍼노드의 매개변수를 지정하려면 다음을 수행하십시오.

1. 수퍼노드 아이콘을 두 번 클릭하여 수퍼노드 대화 상자를 여십시오.
2. 또는 수퍼노드 메뉴에서 **매개변수 설정**을 선택하십시오.
3. **매개변수** 탭을 클릭하십시오. 참고: 이 대화 상자의 필드는 이 탭에서 **매개변수 정의** 단추를 클릭하여 정의한 필드입니다.
4. 작성한 각 매개변수의 텍스트 상자에 값을 입력하십시오. 예를 들어, *minvalue* 값을 관심이 있는 특정 임계값으로 설정할 수 있습니다. 그러면 다수의 조작(예: 향후 탐색을 위해 이 임계값보다 높거나 낮은 레코드 선택)에서 이 매개변수를 사용할 수 있습니다.

수퍼노드 매개변수를 사용하여 노드 특성 액세스

수퍼노드 매개변수를 사용하여 캡슐화 노드의 노드 특성(**슬롯 모수**라고도 함)을 정의할 수도 있습니다. 예를 들어, 수퍼노드가 사용 가능한 데이터의 무작위 표본을 사용하여 특정 시간 동안 캡슐화 신경망 노드를 훈련하도록 지정한다고 가정하십시오. 매개변수를 사용하여 시간 길이 및 백분율 표본에 대한 값을 지정할 수 있습니다.

예제 수퍼노드에 *표본*이라는 표본 노드와 *훈련*이라는 신경망 노드가 포함된다고 가정하십시오. 노드 대화 상자를 사용하여 표본 노드의 **표본** 설정을 **무작위 %**로 설정하고 신경망 노드의 **중지 시점** 설정을 **시간**으로 지정할 수 있습니다. 이러한 옵션을 지정하면 매개변수를 사용하여 노드 특성에 액세스하고 수퍼노드에 고유한 값을 지정할 수 있습니다. 수퍼노드 대화 상자에서 **매개변수 정의**를 클릭하고 다음 표에 표시된 매개변수를 작성하십시오.

표 47. 작성할 매개변수

매개변수	값	긴 이름
Train.time	5	훈련할 시간(분)
Sample.random	10	백분율 무작위 표본

참고: *Sample.random*과 같은 매개변수 이름은 노드 특성을 참조하는 데 올바른 구문을 사용하며, 여기서 *Sample*은 노드의 이름을 나타내고 *random*은 노드 특성입니다.

이러한 매개변수를 정의한 후에는 각 대화 상자를 다시 열지 않고 표본 및 신경망 노드 특성의 값을 쉽게 수정할 수 있습니다. 수퍼노드 메뉴에서 **매개변수 설정**을 선택하여 수퍼노드 대화 상자의 매개변수 탭에 액세스하고 여기서 **무작위 %** 및 **시간**에 대해 새 값을 지정할 수 있습니다. 이는 특히 모델 작성을 여러 번 반복할 때 데이터를 탐색하는 데 유용합니다.

수퍼노드 및 캐싱

수퍼노드 내에서 터미널 노드를 제외한 모든 노드를 캐싱할 수 있습니다. 노드를 마우스 오른쪽 단추로 클릭하고 캐시 컨텍스트 메뉴에서 여러 옵션 중 하나를 선택하여 캐싱을 제어합니다. 이 메뉴 옵션은 수퍼노드 외부에서 사용 가능하고 수퍼노드 내에 캡슐화된 노드에 사용할 수 있습니다.

수퍼노드 캐시에 대한 몇 가지 지침은 다음과 같습니다.

- 수퍼노드 내에 캡슐화된 노드 중 캐싱이 사용되는 노드가 있으면 해당 수퍼노드 또한 캐싱이 사용됩니다.
- 수퍼노드에서 캐시를 사용되지 않도록 설정하면 모든 캡슐화 노드에 대해서도 캐시가 사용되지 않습니다.
- 수퍼노드에서 캐싱을 사용하면 실제로 캐싱 가능한 마지막 하위 노드에서 캐시가 사용됩니다. 즉, 마지막 하위 노드가 선택 노드인 경우 해당 선택 노드에 캐시가 사용됩니다. 마지막 하위 노드가 터미널 노드(캐싱을 허용하지 않음)이면 캐싱을 지원하는 그 다음 업스트림 노드에 캐시가 사용됩니다.
- 수퍼노드의 하위 노드에 대해 캐시를 설정하면, 캐싱되는 노드에서 업스트림인 활동(예: 노드 추가 또는 편집)이 캐시를 비웁니다.

수퍼노드 및 스크립팅

SPSS Modeler 스크립팅 언어를 사용하여 터미널 수퍼노드의 컨텐츠를 조작하고 실행하는 단순 프로그램을 작성할 수 있습니다. 예를 들어, 복잡한 스트림의 실행 순서를 지정하려 할 수 있습니다. 수퍼노드에 구성 노드 전에 실행해야 하는 전역값 설정 노드가 포함되는 경우, 전역값 설정 노드를 먼저 실행하는 스크립트를 작성할 수 있습니다. 평균이나 표준 편차 같이 이 노드가 계산하는 값을 구성 노드가 실행될 때 사용할 수 있습니다.

수퍼노드 대화 상자의 스크립트 탭은 터미널 수퍼노드에만 사용할 수 있습니다.

터미널 수퍼노드에 대한 스크립팅 대화 상자를 열려면 다음을 수행하십시오.

- 수퍼노드 캔버스를 마우스 오른쪽 단추로 클릭하고 **수퍼노드 스크립트**를 선택하십시오.
- 또는 확대 및 축소 모드 둘 다에서 수퍼노드 메뉴로부터 **수퍼노드 스크립트**를 선택할 수 있습니다.

참고: 수퍼노드 스크립트는 대화 상자에서 **현재 스크립트 실행**을 선택한 경우에 해당 스트림 및 수퍼노드에서만 실행됩니다.

SPSS Modeler에서 스크립트를 작성하고 사용하는 데 필요한 고유 옵션에 대해서는 제품 다운로드에서 PDF 파일로 제공되는 *스크립팅 및 자동화 안내서*를 참조하십시오.

수퍼노드 저장 및 로드

수퍼노드의 장점 중 하나는 수퍼노드를 저장하여 다른 스트림에서 재사용할 수 있는 것입니다. 수퍼노드를 저장하고 로드할 때 수퍼노드는 *.slb* 확장자를 사용합니다.

수퍼노드를 저장하려면 다음을 수행하십시오.

1. 수퍼노드를 확대하십시오.
2. 수퍼노드 메뉴에서 **수퍼노드 저장**을 선택하십시오.
3. 대화 상자에서 파일 이름 및 디렉토리를 지정하십시오.
4. 저장된 수퍼노드를 현재 프로젝트에 추가할지 여부를 선택하십시오.
5. **저장**을 클릭합니다.

수퍼노드를 로드하려면 다음을 수행하십시오.

1. IBM SPSS Modeler 창의 삽입 메뉴에서 **수퍼노드**를 선택하십시오.
2. 현재 디렉토리의 수퍼노드 파일(*.slb*)을 선택하거나 찾아보기를 사용하여 다른 디렉토리의 수퍼노드 파일을 찾으십시오.
3. **로드**를 클릭하십시오.

참고: 가져온 수퍼노드의 매개변수는 모두 기본값을 가집니다. 매개변수를 변경하려면 스트림 캔버스에서 수퍼노드를 두 번 클릭하십시오.

주의사항

이 정보는 미국에서 제공되는 제품 및 서비스용으로 작성된 것입니다. 본 자료는 다른 언어로도 제공될 수 있습니다. 그러나 자료에 접근하기 위해서는 해당 언어로 된 제품 또는 제품 버전의 사본이 필요할 수 있습니다.

IBM은 다른 국가에서 이 책에 기술된 제품, 서비스 또는 기능을 제공하지 않을 수도 있습니다. 현재 사용할 수 있는 제품 및 서비스에 대한 정보는 한국 IBM 담당자에게 문의하십시오. 이 책에서 IBM 제품, 프로그램 또는 서비스를 언급했다고 해서 해당 IBM 제품, 프로그램 또는 서비스만을 사용할 수 있다는 것을 의미하지는 않습니다. IBM의 지적 재산권을 침해하지 않는 한, 기능상으로 동등한 제품, 프로그램 또는 서비스를 대신 사용할 수도 있습니다. 그러나 비IBM 제품, 프로그램 또는 서비스의 운영에 대한 평가 및 검증은 사용자의 책임입니다.

IBM은 이 책에서 다루고 있는 특정 내용에 대해 특허를 보유하고 있거나 현재 특허 출원 중일 수 있습니다. 이 책을 제공한다고 해서 특허에 대한 라이센스까지 부여하는 것은 아닙니다. 라이센스에 대한 의문사항은 다음으로 문의하십시오.

150-945
서울특별시 영등포구
국제금융로 10, 3IFC
한국 아이.비.엠 주식회사
대표전화서비스: 02-3781-7114

2바이트(DBCS) 정보에 관한 라이센스 문의는 한국 IBM에 문의하거나 다음 주소로 서면 문의하시기 바랍니다.

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

IBM은 타인의 권리 비침해, 상품성 및 특정 목적에의 적합성에 대한 묵시적 보증을 포함하여(단, 이에 한하지 않음) 묵시적이든 명시적이든 어떠한 종류의 보증 없이 이 책을 "현상태대로" 제공합니다. 일부 국가에서는 특정 거래에서 명시적 또는 묵시적 보증의 면책사항을 허용하지 않으므로, 이 사항이 적용되지 않을 수도 있습니다.

이 정보에는 기술적으로 부정확한 내용이나 인쇄상의 오류가 있을 수 있습니다. 이 정보는 주기적으로 변경되며, 변경된 사항은 최신판에 통합됩니다. IBM은 이 책에서 설명한 제품 및/또는 프로그램을 사전 통지 없이 언제든지 개선 및/또는 변경할 수 있습니다.

이 정보에서 언급되는 비IBM의 웹 사이트는 단지 편의상 제공된 것으로, 어떤 방식으로든 이들 웹 사이트를 옹호하고자 하는 것은 아닙니다. 해당 웹 사이트의 자료는 본 IBM 제품 자료의 일부가 아니므로 해당 웹 사이트 사용으로 인한 위험은 사용자 본인이 감수해야 합니다.

IBM은 귀하의 권리를 침해하지 않는 범위 내에서 적절하다고 생각하는 방식으로 귀하가 제공한 정보를 사용하거나 배포할 수 있습니다.

(i) 독립적으로 작성된 프로그램과 기타 프로그램(본 프로그램 포함) 간의 정보 교환 및 (ii) 교환된 정보의 상호 이용을 목적으로 본 프로그램에 관한 정보를 얻고자 하는 라이센스 사용자는 다음 주소로 문의하십시오.

150-945
서울특별시 영등포구
국제금융로 10, 3IFC
한국 아이.비.엠 주식회사
대표전화서비스: 02-3781-7114

이러한 정보는 해당 조건(예를 들면, 사용료 지불 등)하에서 사용될 수 있습니다.

이 정보에 기술된 라이센스가 부여된 프로그램 및 프로그램에 대해 사용 가능한 모든 라이센스가 부여된 자료는 IBM이 IBM 기본 계약, IBM 프로그램 라이센스 계약(IPLA) 또는 이와 동등한 계약에 따라 제공한 것입니다.

인용된 성능 데이터와 고객 예제는 예시 용도로만 제공됩니다. 실제 성능 결과는 특정 구성과 운영 조건에 따라 다를 수 있습니다.

비IBM 제품에 관한 정보는 해당 제품의 공급업체, 공개 자료 또는 기타 범용 소스로부터 얻은 것입니다. IBM에서는 이러한 제품들을 테스트하지 않았으므로, 비IBM 제품과 관련된 성능의 정확성, 호환성 또는 기타 청구에 대해서는 확신할 수 없습니다. 비IBM 제품의 성능에 대한 의문사항은 해당 제품의 공급업체에 문의하십시오.

IBM의 향후 제시 방향 또는 의도에 관한 언급은 특별한 통지 없이 변경 내지 철회될 수 있으며, 단순히 목표만을 의미합니다.

이 정보에는 일상의 비즈니스 운영에서 사용되는 자료 및 보고서에 대한 예제가 들어 있습니다. 이들 예제에는 개념을 가능한 완벽하게 설명하기 위하여 개인, 회사, 상표 및 제품의 이름이 사용될 수 있습니다. 이들 이름은 모두 가공의 것이며 실제 기업의 이름 및 주소와 유사하더라도 이는 전적으로 우연입니다.

상표

IBM, IBM 로고 및 ibm.com은 전세계 여러 국가에 등록된 International Business Machines Corp.의 상표 또는 등록상표입니다. 기타 제품 및 서비스 이름은 IBM 또는 타사의 상표입니다. 현재 IBM 상표 목록은 웹 "저작권 및 상표 정보"(www.ibm.com/legal/copytrade.shtml)에 있습니다.

Adobe, Adobe 로고, PostScript 및 PostScript 로고는 미국 및/또는 기타 국가에서 사용되는 Adobe Systems Incorporated의 등록상표 또는 상표입니다.

Intel, Intel 로고, Intel Inside, Intel Inside 로고, Intel Centrino, Intel Centrino 로고, Celeron, Intel Xeon, Intel SpeedStep, Itanium 및 Pentium은 미국 또는 기타 국가에서 사용되는 Intel Corporation 또는 그 계열사의 상표 또는 등록상표입니다.

Linux는 미국 또는 기타 국가에서 사용되는 Linus Torvalds의 등록상표입니다.

Microsoft, Windows, Windows NT 및 Windows 로고는 미국 또는 기타 국가에서 사용되는 Microsoft Corporation의 상표입니다.

UNIX는 미국 및 기타 국가에서 사용되는 The Open Group의 등록상표입니다.

Java 및 모든 Java 기반 상표와 로고는 Oracle 및/또는 그 계열사의 상표 또는 등록상표입니다.

제품 문서의 이용 약관

다음 이용 약관에 따라 이 책을 사용할 수 있습니다.

적용성

본 이용 약관은 IBM 웹 사이트의 모든 이용 약관에 추가됩니다.

개인적 사용

모든 소유권 사항을 표시하는 경우에 한하여 귀하는 이 책을 개인적, 비상업적 용도로 복제할 수 있습니다. 귀하는 IBM의 명시적 동의 없이 본 발행물 또는 그 일부를 배포 또는 전시하거나 2차적 저작물을 만들 수 없습니다.

상업적 사용

모든 소유권 사항을 표시하는 경우에 한하여 귀하는 이 책을 귀하 기업집단 내에서만 복제, 배포 및 전시할 수 있습니다. 귀하는 귀하의 기업집단 외에서는 IBM의 명시적 동의 없이 이 책의 2차적 저작물을 만들거나 이 책 또는 그 일부를 복제, 배포 또는 전시할 수 없습니다.

권한

본 허가에서 명시적으로 부여된 경우를 제외하고, 이 책이나 이 책에 포함된 정보, 데이터, 소프트웨어 또는 기타 지적 재산권에 대한 어떠한 허가나 라이센스 또는 권한도 명시적 또는 묵시적으로 부여되지 않습니다.

IBM은 이 책의 사용이 IBM의 이익을 해친다고 판단하거나 위에서 언급된 지시사항이 준수되지 않는다고 판단하는 경우 언제든지 부여한 허가를 철회할 수 있습니다.

귀하는 미국 수출법 및 관련 규정을 포함하여 모든 적용 가능한 법률 및 규정을 철저히 준수하는 경우에만 본 정보를 다운로드, 송신 또는 재송신할 수 있습니다.

IBM은 이 책의 내용과 관련하여 아무런 보장을 하지 않습니다. 타인의 권리 비침해, 상품성 및 특정 목적에의 적합성에 대한 묵시적 보증을 포함하여 (단 이에 한하지 않음) 묵시적이든 명시적이든 어떠한 종류의 보증 없이 현 상태대로 제공합니다.

용어

가

고유(Unique). 모든 효과를 동시에 평가하고 유형에 관계없이 다른 모든 효과에 대해 각 효과를 조정합니다.

공분산(Covariance). 두 변수 간 연관을 표준화하지 않은 측도로서, N-1로 나눈 교차곱 편차와 같습니다.

바

범위(Range). 숫자변수의 가장 큰 값과 가장 작은 값의 차이로 최대값에서 최소값을 뺀 값을 의미합니다.

분산(Variance). 평균에 대한 산포 측도로, 평균으로부터의 제곱합 편차를 케이스 수에서 1을 뺀 값으로 나눈 값과 같습니다. 분산은 변수 자체의 제곱 단위로 측정됩니다.

아

왜도(Skewness). 분포의 비대칭성에 대한 측도입니다. 정규 분포는 대칭이므로 왜도 값이 0입니다. 양의 왜도가 많은 분포는 오른쪽이 깁니다. 유의한 음의 왜도를 가지는 분포에는 왼쪽으로 긴 꼬리가 나타납니다. 왜도값이 표준 오차의 두 배를 넘는 것은 대칭에서 벗어난 정도를 나타냅니다.

왜도의 표준 오차(Standard Error of Skewness). 표준 오차에 대한 왜도의 비율을 정규성 검정에 사용할 수 있습니다. 즉, 비율이 -2보다 작거나 +2보다 큰 경우 정규성을 거부할 수 있습니다. 왜도가 큰 양의 값인 경우 오른쪽이 길어지고 큰 음의 값인 경우 왼쪽이 길어집니다.

유효함(Valid). 시스템 결측값 또는 사용자 결측값이 지정되어 있지 않은 케이스가 유효 케이스입니다.

자

중앙값(Median). 전체 케이스의 절반이 위 아래에 해당되는 값으로 제50 백분위수입니다. 케이스 수가 짝수인 경우 중앙값은 케이스를 오름차순이나 내림차순으로 정렬했을 때 중간에 있는 두 개의 케이스의 평균입니다. 중앙값은 평균과 달리 중심을 벗어난 값에는 영향을 받지 않는 중심 경향 측도이며, 상한 극단값 또는 하한 극단값에 따라 달라질 수 있습니다.

차

첨도(Kurtosis). 관측값이 중심에 군집하는 정도에 대한 측도입니다. 정규 분포의 경우 첨도 통계 값은 0입니다. 양의 첨도는 정규 분포에 비해 관측값이 분포 중심에 더 많이 군집되어 있고 분포 극단값까지의 꼬리가 더 얇다는 의미입니다. 즉, 정규 분포에 비해 급첨 분포의 꼬리가 더 두껍습니다. 음의 첨도는 정규 분포에 비해 관측값이 분포 중심에 덜 군집되어 있고 분포 극단값까지의 꼬리가 더 두껍다는 의미입니다. 즉, 정규 분포에 비해 평첨 분포의 꼬리가 더 얇습니다.

첨도의 표준 오차(Standard Error of Kurtosis). 표준 오차에 대한 첨도의 비율을 정규성 검정에 사용할 수 있습니다. 즉, 비율이 -2보다 작거나 +2보다 큰 경우 정규성을 거부할 수 있습니다. 첨도가 높은 양의 값인 경우 분포의 양끝이 정규 분포의 양끝보다 길어지고 음의 값인 경우 양끝이 짧아집니다(상자 형태 균일 분포와 유사).

최대값(Maximum). 숫자변수의 가장 큰 값입니다.

최빈값(Mode). 가장 자주 발생하는 값입니다. 여러 값에서 최대 발생 빈도를 공유하는 경우 각각을 최빈값이라고 합니다.

최소값(Minimum). 숫자변수의 가장 작은 값입니다.

파

평균(Mean). 중심 경향에 대한 측도입니다. 합계를 케이스 수로 나눈 산술 평균 값입니다.

평균의 표준 오차(Standard Error of Mean). 동일 분포로부터 선택한 표본 간에 발생할 수 있는 평균값의 차이에 대한 측도입니다. 이 값을 사용하여 관측 평균과 가설 값을 간략하게 비교할 수 있습니다. 즉, 표준 오차에 대한 차이 비율이 ?2보다 작거나 +2보다 큰 경우 두 값이 다르다고 판단할 수 있습니다.

표준 오차(Standard Error). 검정 통계량 값이 표본마다 얼마나 달라지는지에 대한 측도입니다. 이 항목은 통계에 대한 표본 분포의 표준 편차가 됩니다. 예를 들어, 평균의 표준 오차는 표본 평균의 표준 편차입니다.

표준 편차(standard deviation). 평균 주위의 산포 측도이며 분산의 제곱근과 같습니다. 표준 편차는 원래 변수와 같은 단위로 측정됩니다.

표준 편차(Standard Deviation). 평균에 대한 산포 측도입니다. 정규 분포에서 케이스의 68%는 평균의 표준 편차 내에 있으며 케이스의 95%는 2배 표준 편차 내에 있습니다. 예를 들어, 평균 연령이 45세이고 표준 편차가 10인 경우 정규 분포 내에서 95% 케이스는 25세와 65세 사이에 있습니다.

하

합계(Sum). 비결측값을 갖는 전체 케이스 값의 총계입니다.

색인

[가]

[나]

[다]

[라]

[마]

[바]

[사]

[아]

[자]

[차]

[카]

[타]

[파]

[하]

[숫자]

A

B

C

D

E

F

G

H

I

L

M

N

O

P

Q

R

S

T

U

V

X

[특수 문자]