데이타솔루션

SPSS Modeler Plus Pack

(주)데이타솔루션 | 서울시 강남구 언주로 620 현대인텔렉스빌딩 10층(4,5층)

www.datasolution.kr

목차

데이타솔루션
Modeler Plus Pack

데이타솔루션

Modeler Plus Pack

I. SPSS Modeler

(1) Introduction
(2) SPSS Modeler 방법론
(3) SPSS Modeler 기능구성

II. SPSS Modeler 특장점

(1) 다변화 작업에 적합한 구성과 Interface
(2) 개방형 시스템(Open Architecture)
(3) SQL Optimization
(4) Bulk Loading
(5) Learning Cost 감소

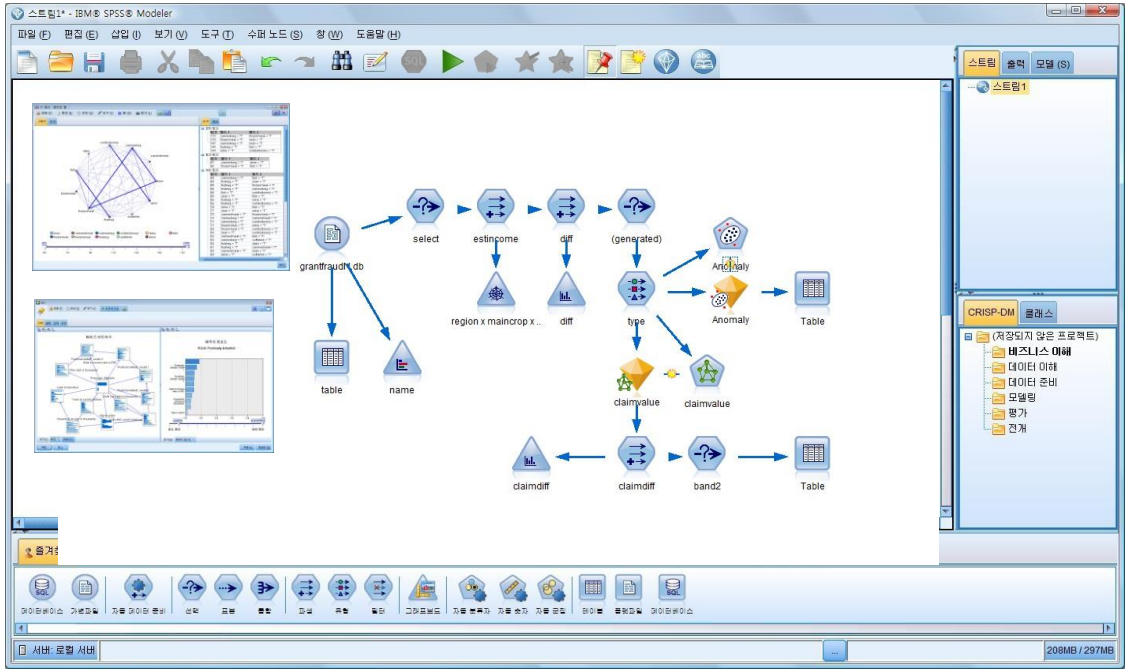
III. SPSS Modeler Plus Pack

(1) Score Card
(2) Data Visualization
(3) Statistical Utility
(4) Model Extraction
(5) TA Korean

SPSS Modeler Plus Pack

I. Introduction

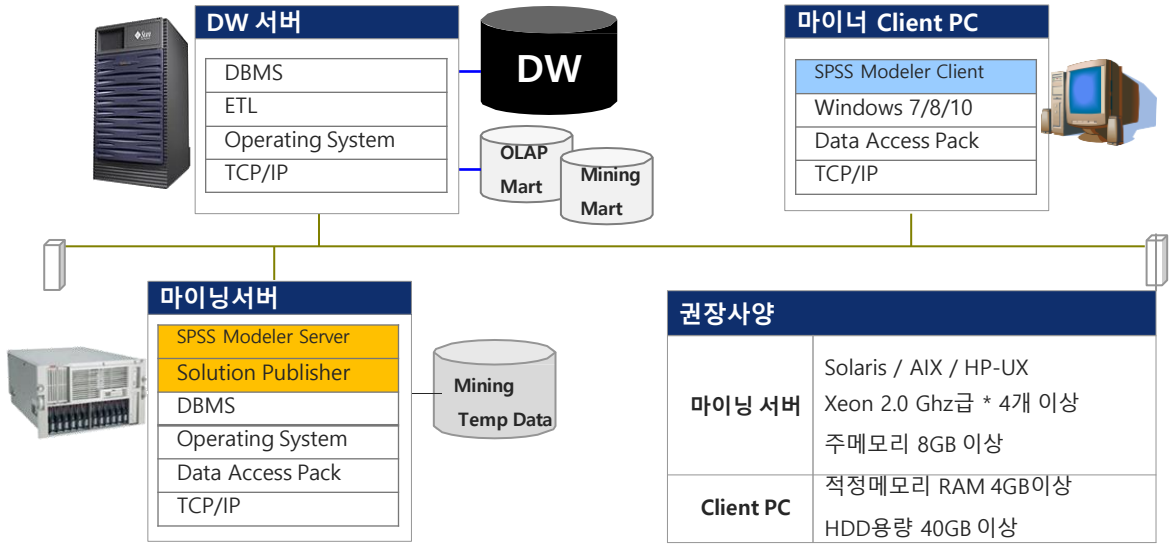
SPSS Modeler는 데이터를 로딩, 변환, 정제, 모델링, 그래프, 결과의 출력까지 하나의 소프트웨어 내에서 가능하며, 이러한 모든 기능을 대화상자와 아이콘, 메뉴를 이용하여 완벽한 GUI를 구현한 소프트웨어입니다.



- **Visual한 작업방식을 통한 손쉬운 접근**
- **다양한 모델 생성 가능**
- **작업의 유연성을 보장하는 인터페이스**
- **개방형 구조 (Open Architecture)**

I. Introduction – 시스템 구성

구 분	세부 기능
SPSS Modeler Server	▪ 대량의 데이터를 처리할 수 있는 마이닝서버에 탑재 ▪ Client의 요구사항을 받아서 마이닝 작업을 수행하는 Multi-thread Backend Engine ▪ 데이터 전처리, 모델 평가에 있어 최상의 퍼포먼스 제공
SPSS Modeler Client	▪ 마이너 Client PC에 탑재 사용자의 마이닝 작업 수행을 위한 비주얼 프로그래밍과 GUI 환경 제공을 제공하는 Front End

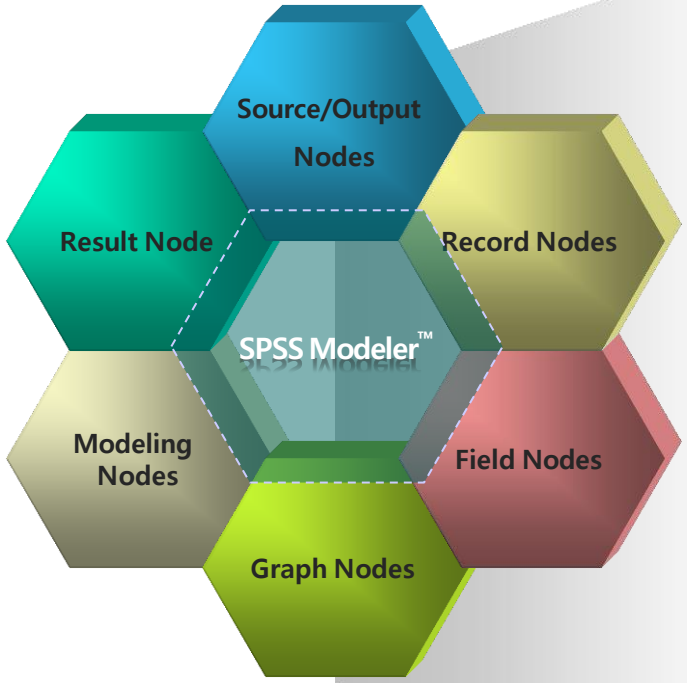


I. SPSS Modeler 방법론

SPSS Modeler은 대용량 데이터로부터 유용한 정보를 찾아내기 위한 CRISP-DM (Cross Industry Standard Process for Data Mining) 방법론의 Data Mining 전용 Tool입니다. CRISP-DM 방법론은 어떤 산업분야에서도 적용 될 수 있습니다.

I. SPSS Modeler 기능구성

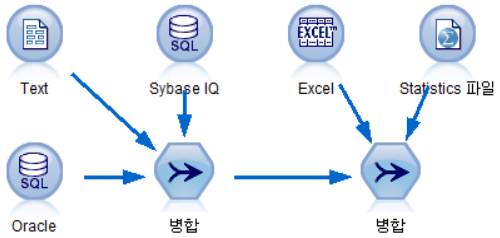
SPSS Modeler은 CRISP-DM (Cross Industry Standard Process for Data Mining) 방법론을 수행하기 위한 다양한 기능을 보유하고 있습니다.



Sources Node	• 데이터 연결 노드 • 데이터베이스 연결 또는 가변형식, 고정형식 파일의 데이터, SPSS Statistics, SAS 파일 등의 다양한 파일들을 데이터로 읽음
Operations Node	• 데이터 변환 작업 노드 • 샘플링, 레코드 또는 필드단위의 데이터 병합, 필터, 변수파생, 모형평가를 위한 파티션 작업포함
Graphs Node	• 데이터 도식화 노드 • 크게 데이터 탐색으로 이용되는 히스토그램, 2차원 및 3차원 도표와 ROI Chart 등과 같은 평가 도표로 이용
Modeling Node	• 데이터 모형화 노드 • Decision Tree, Regression, Neural Network, Clustering, Association 등 다양한 종류 이용가능
Output Node	• 마이닝 결과 출력 노드 • 최종 결과를 테이블, 외부파일로 출력하는 기능, 생성된 모델들 간의 예측력을 평가하는 기능 포함

I. SPSS Modeler 기능구성

다양한 소스 시스템에 동시 접근 및 Import



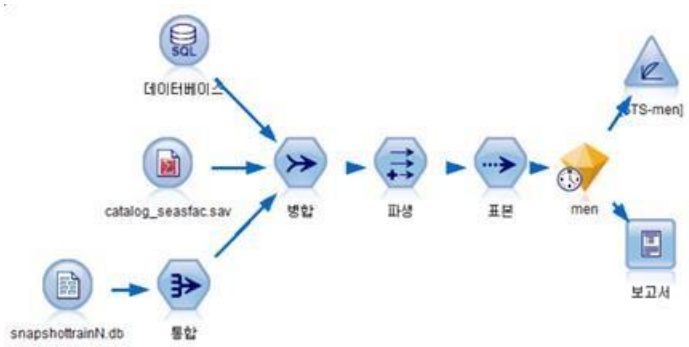
- SPSS Statistics 파일, 엑셀 파일, 가변 파일, 고정 파일, SAS 파일, Database 등 다양한 소스에 동시에 접근하여, 데이터 핸들링, 모델링, 전개 등을 할 수 있습니다.
- 다양한 소스 시스템을 한 작업 파일(스트림)에서 동시에 활용 가능하므로 데이터 유형을 통일 시키기 위한 변환 작업을 할 필요가 없기 때문에 업무의 효율을 향상 시킬 수 있습니다.

Database에서 직접 데이터 추출

- Database로부터 데이터를 불러올 때, 테이블 단위로 전부 불러올지 또는 분석자가 직접 해당 Database의 SQL 질의를 작성하여 불러올지를 지정하는 옵션을 사용할 수 있습니다.

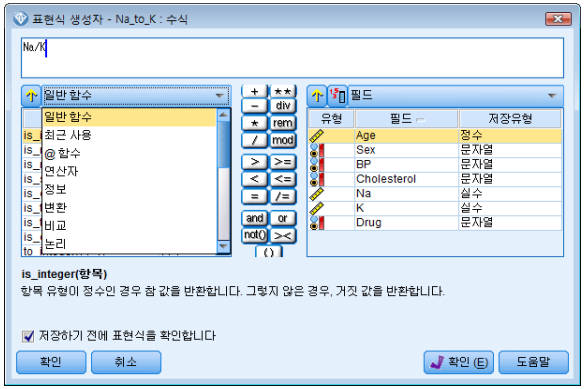
I. SPSS Modeler 기능구성

편리한 데이터 핸들링 기능



- SQL이나 ETL을 사용하지 않아도 데이터 선택, 파생, 병합 등의 다양한 핸들링 기능을 제공합니다.
- 병합, 추가 등의 노드를 이용하여 여러 개의 데이터 파일에서 하나의 데이터 파일로 통합 기능을 지원합니다.
- 다양하고 쉬운 레코드 처리 기능(선택, 통합, 표본추출, 병합, 정렬 등) 및 필드 처리 기능(파생, 구간화 등) 을 지원합니다.

자체 함수 기능



- SPSS Modeler에서는 자체 함수 기능을 이용하여 일반 함수, 연산자 함수, 변환 함수, 비교 함수, 논리 함수, 날짜 및 시간 함수, Null 값 선택 함수 등 문자형/숫자형/날짜형 등 다양한 데이터 유형에 대해 편리하게 함수를 작성할 수 있는 기능을 제공합니다.

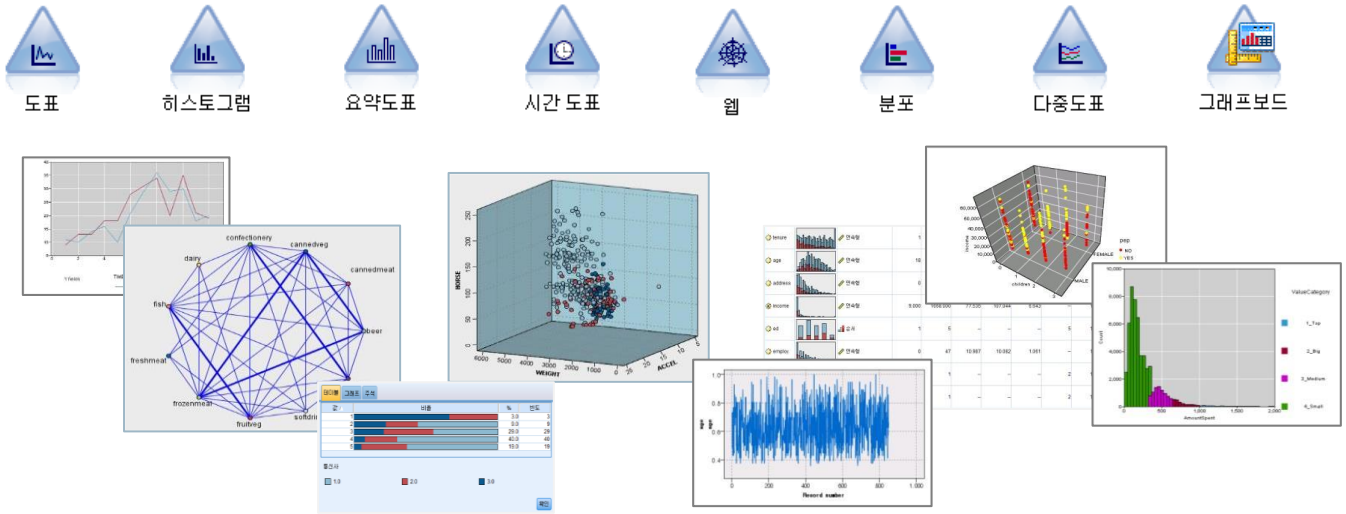
I. SPSS Modeler 기능구성

다양한 분석 알고리즘

- SPSS Modeler는 최신의 다양한 분석기법을 제공하고, 비즈니스 목적에 맞게 이를 적용하여 예측력이 높고 적합한 모델을 선택할 수 있습니다.
- 자동화 모형을 이용하게 되면, 다양한 분석 기법을 함께 고려하여 자동으로 탐색해 주므로 초보자 / 미경험자들에게 매우 편리한 기능입니다.
- 스코어링 모델을 개발할 때에도 Supervised Learning의 통계 분석 기법(의사결정나무 분석 등)을 이용한 룰을 도출하는 등 다양한 모델을 고려해 볼 수 있습니다.

I. SPSS Modeler 기능구성

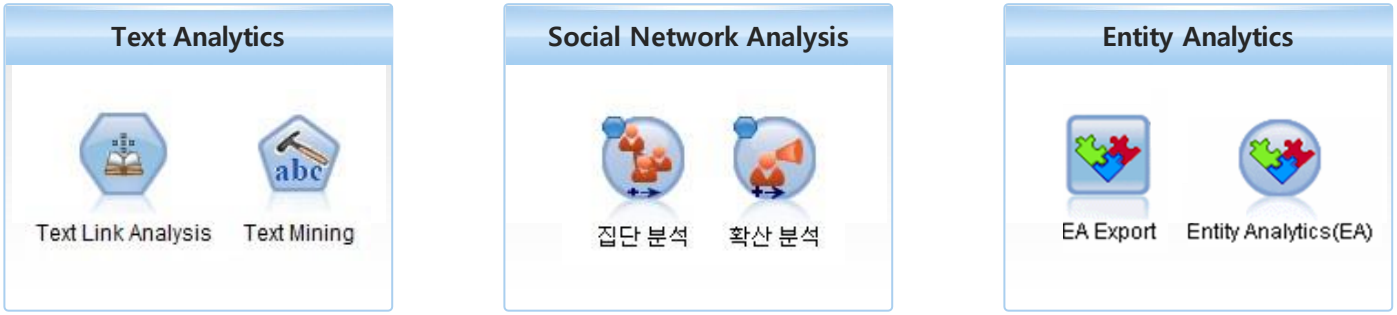
화려한 그래프 기능



- SPSS Modeler의 그래프 기능을 이용하면 시각적으로 데이터의 특성을 파악할 수 있기 때문에 비전문가들도 그 결과가 나타내는 의미에 대한 해석이 매우 용이하므로 전반적인 데이터의 특성 및 분석 결과를 효율적으로 파악하고 공유할 수 있습니다.
- 현존하는 모든 Data Mining S/W 중 가장 수려한 Visual을 보여 줍니다.
- 현재 대다수의 S/W의 경우, 수행 후 해당 결과 그래프와 편집표를 보고서용으로 만들 때 MS-Office를 다시 사용합니다. SPSS Modeler는 이런 2중 작업이 필요 없는 Visual 및 편리성을 제공합니다.

I. SPSS Modeler 기능구성

특화된 추가 모듈

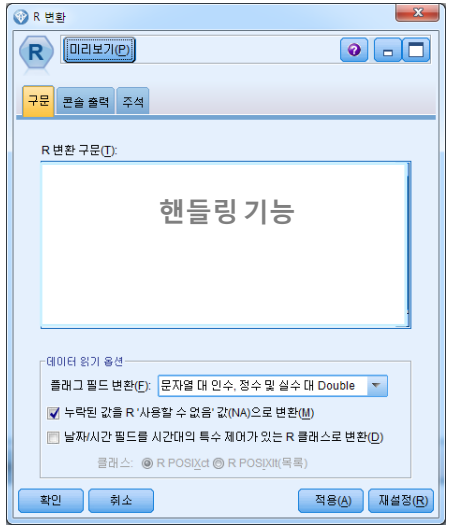


- Text Analytics는 비정형의 텍스트를 분석하기 핵심개념/용어로 추출 및 추출된 개념/용어를 분석하기 쉬운 범주로 그룹핑 하고, Text 안에 포함한 의미를 수치화하여 마이닝 모델의 적중률 향상 가능을 제공합니다.
- Social Network Analysis는 네트워크 안의 개개인의 상호작용 패턴을 사용하여 유사한 개인들의 그룹을 식별함으로써 특성 파악 가능하고, 분석을 통해 식별된 그룹에 대한 지표와 네트워크 안의 개개인에 대한 지표를 제공합니다.
- Entity Analytics는 레코드 내에 식별 문제를 해결하여 현재 데이터의 일관성을 향상시키는데 초점을 맞추는 분석 기법으로, 식별 해결은 고객 관계 관리, 사기 탐지, 자금세탁방지, 국내 및 국제 안보를 포함한 많은 분야에서 사용되고 있습니다.

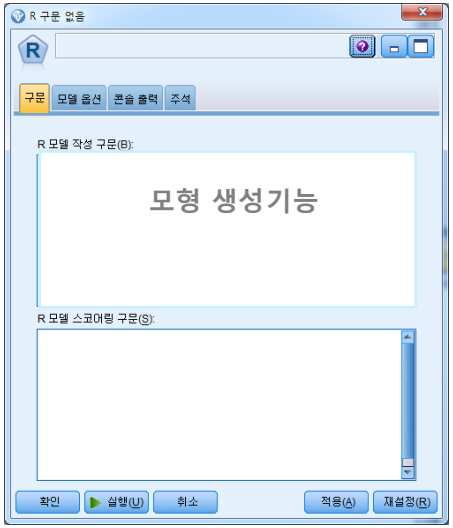
I. SPSS Modeler 기능구성

R을 이용한 확장 분석 기능

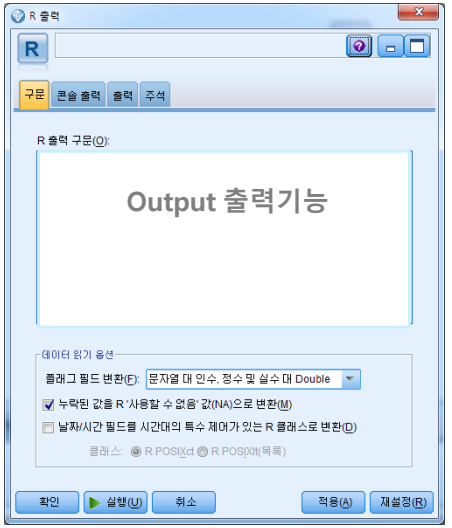
R 변환



R



R 출력



- Open Source 인 R을 사용하여 데이터 핸들링, 모델링, 출력을 할 수 있는 인터페이스를 제공합니다.
- R의 최신 분석 알고리즘 및 화려한 그래프, 자유로운 데이터 핸들링을 Modeler에서도 손쉽게 사용할 수 있으며, 이를 통해 상용프로그램의 한계를 뛰어 넘을 수 있습니다.

데이타솔루션

Modeler Plus Pack

I. SPSS Modeler

(1) Introduction
(2) SPSS Modeler 방법론
(3) SPSS Modeler 기능구성

II. SPSS Modeler 특장점

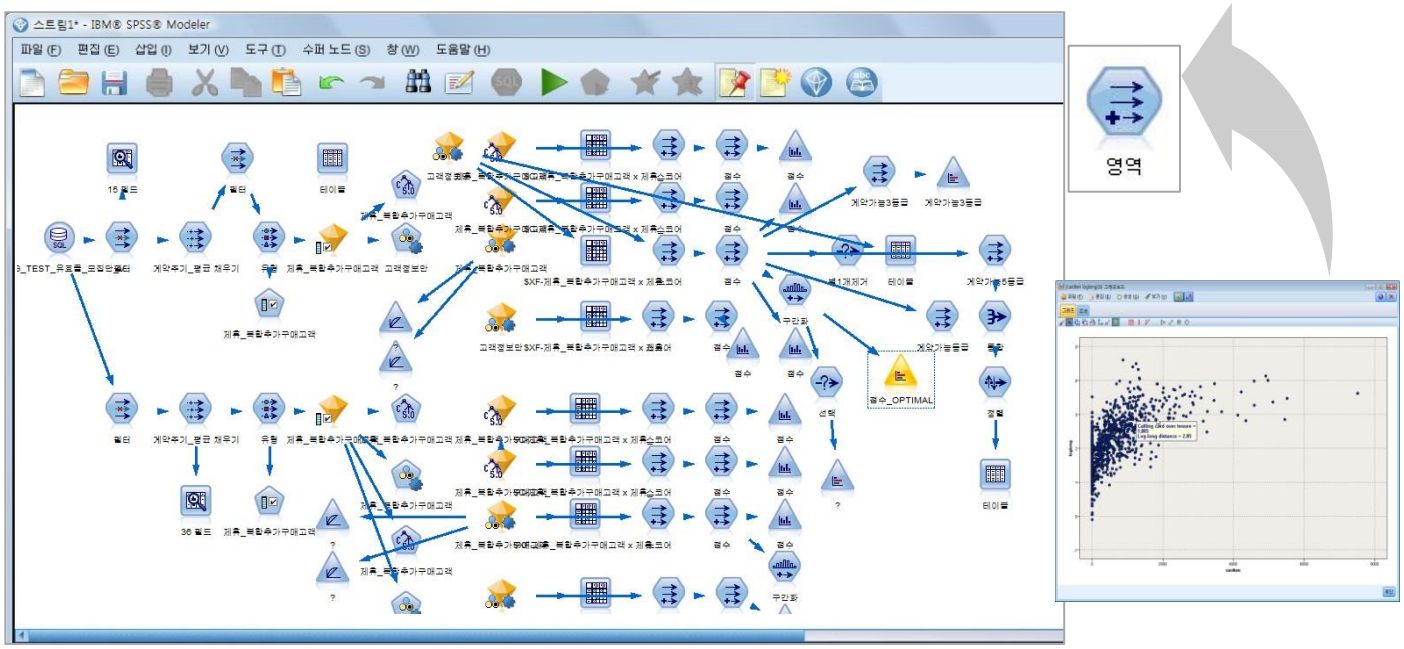
(1) 다변화 작업에 적합한 구성과 Interface
(2) 개방형 시스템(Open Architecture)
(3) SQL Optimization
(4) Bulk Loading
(5) Learning Cost 감소

III. SPSS Modeler Plus Pack

(1) Score Card
(2) Data Visualization
(3) Statistical Utility
(4) Model Extraction
(5) TA Korean

II. SPSS Modeler 특장점(1/5)

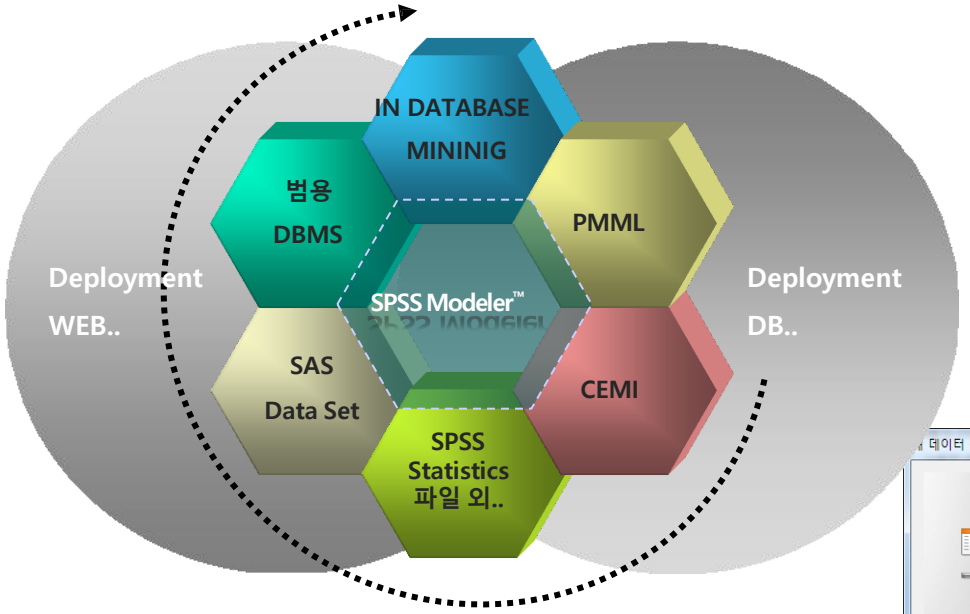
마이닝은 한 번에 이루어 지지 않는다.

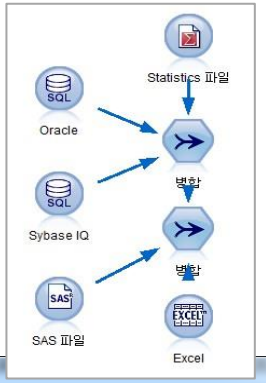


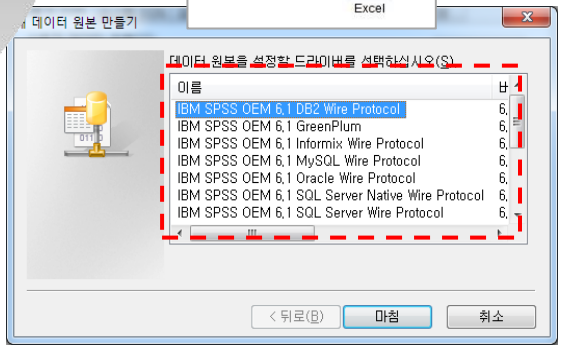
- 다양한 경우의 Mining 작업에서 수행과 변경이 가능
- 모든 Graph와 Output에 Interactive 기능 제공
- 직관적인 프로세스 제공 ▶ 구성한 Stream 자체가 작업 History이며, Process / Output 탐색 시간 단축

II. SPSS Modeler 특장점(2/5)

다양한 Sources의 접근, 모델링, 전개가 가능한 개방형 구조





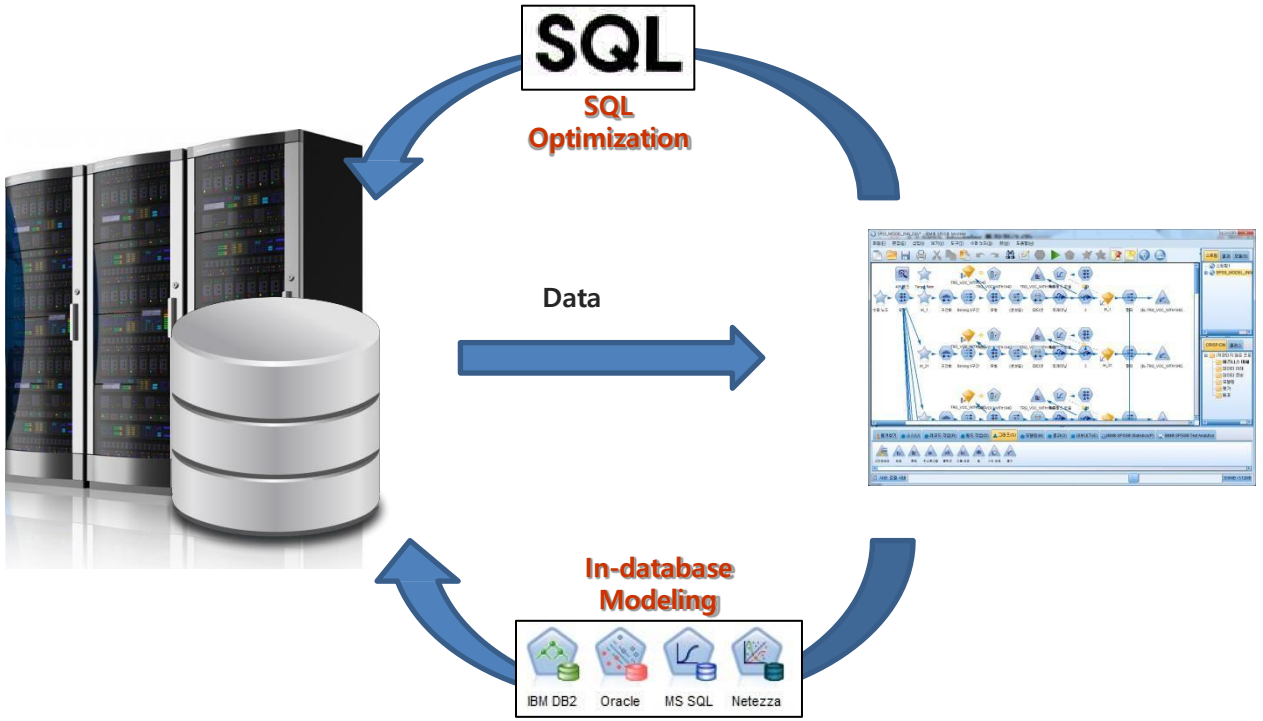


- 다양한 Source를 동시에 활용 가능
- 손쉬운 모델링 및 각종 기능 추가/변경
- 다른 제품의 Mining 기능 사용 (MS-SQL, Oracle, IBM DB2)
- 기존의 ODBC 외 SPSS Modeler에 최적화 된 SPSS Statistics OEM Wire Protocol 제공

II. SPSS Modeler 특장점(3/5)

다변화 작업
개방형 구조
SQL Optimization & In-DB Modeling
Bulk Loading
Learning Cost 감소

성능이 뛰어난 DB의 자원을 최대한 이용하라.

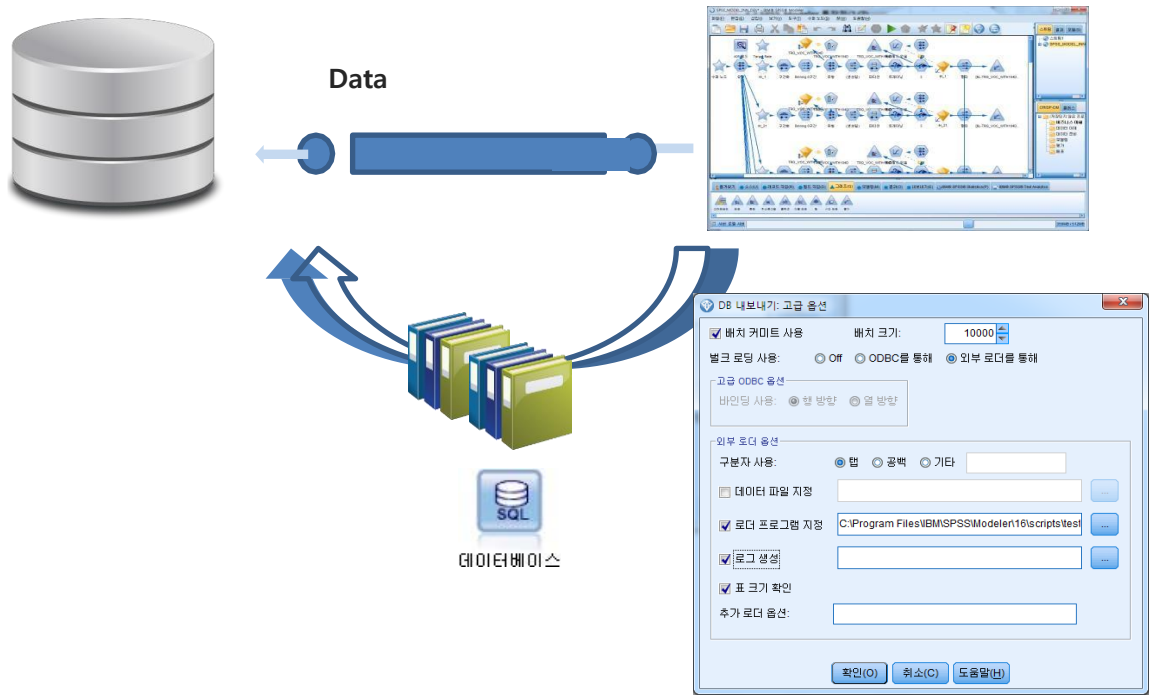


- SQL Optimization은 DB의 성능을 이용한 전처리 및 모델링 Performance(속도) 향상시킴 In-data base Modeling을 통해 DB내에서 모델링을 처리

II. SPSS Modeler 특장점(4/5)

다변화 작업
개방형 구조
SQL Optimization & In-DB Modeling
Bulk Loading
Learning Cost 감소

대용량 데이터의 효과적인 출력 기능

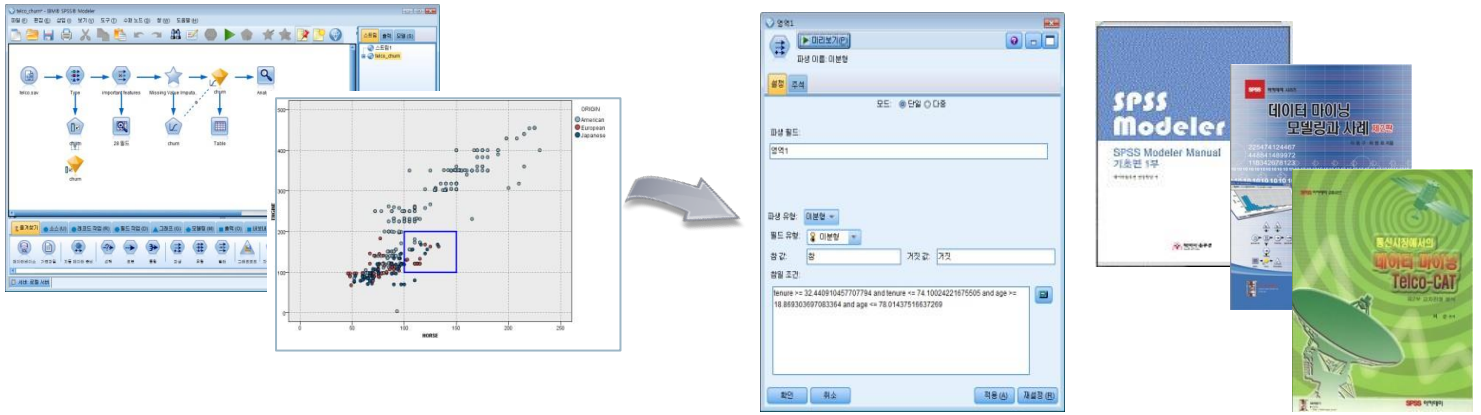


- DB의 외부 로더 프로그램을 이용한 Data의 DB Export 기능
- 현존하는 모든 방법 중 가장 빠른 DB Export 방법임
- 특히 대용량 처리시 Sybase IQ와 같은 DB는 본 방법 외에 없음

II. SPSS Modeler 특장점(5/5)

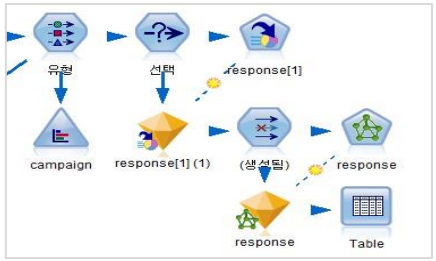
- 다변화 작업
- 개방형 구조
- SQL Optimization & In-DB Modeling
- Bulk Loading
- Learning Cost 감소

Easy to use



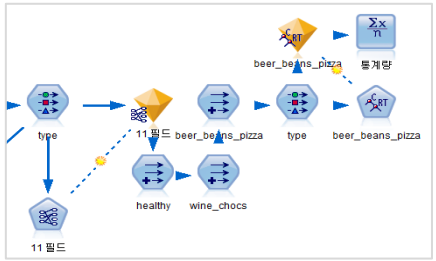
- 한글 버전을 지원합니다.
  - SPSS Modeler는 영문 버전뿐 아니라 한글버전까지 완벽히 지원하여 전반적인 사용에 있어 국내 사용자들에게 편리함을 제공합니다.
- 조직/인원 변동 시 손쉬운 학습으로 프로그램 사용이 지속됩니다.
  - 기업의 경우 잦은 인사변동이 발생합니다. 그 때마다 기업 내에 담당자가 바뀌었을 경우 새로운 S/W를 배우는데 쉽지 않다면 활용하기 매우 어렵습니다. SPSS Modeler는 간단한 학습만으로 사용법을 익힐 수 있어 지속적인 사용이 가능합니다.
- 연구 및 개발이 매우 편해집니다.
  - SPSS Modeler는 제품 뿐 아니라 관련 매뉴얼/도서 등도 완벽히 한글화 되어 연구 자료가 풍부합니다.

II. SPSS Modeler 모델링의 특장점



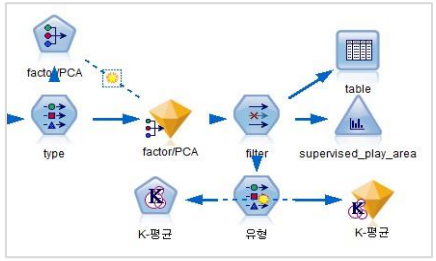
의사결정나무로 변수 선택 후 신경망 분석

1) 통계적인 검정(test)외에도 다양한 변수 선택 가능
2) 지도학습 중 의사결정나무, 회귀계열 분석 모두 간단하게 사용한 변 수만 추출 가능
   + 변수선택 자체 기능 비교
3) 변수 선택 후 2차 분석



연관성 분석과 지도학습 기법의 연결

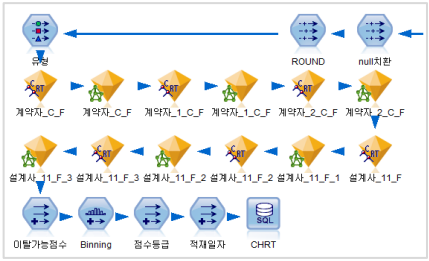
1) 1차적으로 연관성 분석을 수행하여, 특정 조건 또는 전혀 연관성이 없는 고객만을 선정하여, 이들만 가지고 특정 목표에 대하여 지도학습 수행



요인분석과 군집분석 등의 연결

1) 변수가 많은 경우 변수 축소 후 군집분석을 수행하여, 효율적으로 각 종 데이터를 Segment

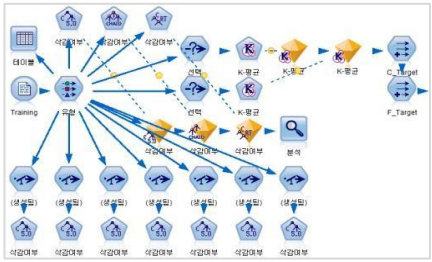
II. SPSS Modeler 모델링의 특장점



오차 패턴 모델링

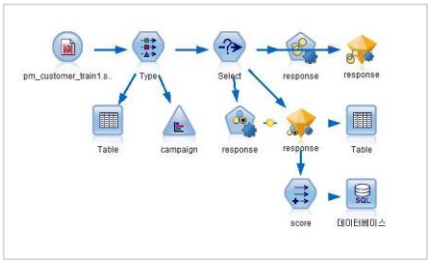
1) Hybrid 모델의 일종

2) 2개 이상의 서로 다른 모델 훈련시키고, 또한 레코드별로 더욱 잘 맞는 것이 어떤 모델 인지판별하는 모델을 별도로 만들어, 최적 모형 개발



K평균 군집분석을 이용한 Target 변수별 근접도 계산

1) 목적변수 범주 별로 별도 군집 모델링을 통한 군집 거리 계산 후 이 를 비교하여 레코드별로 목표 변수의 범주 별 근접성을 판별하는 방 법 (이상치 파악 모형)



최신의 알고리즘 추가

1) Binary classifier (다양한 이분형 분류 마이닝 모형을 자동으로 생성하고, 그 결과를 비교하여 주는 Node)

2) Numeric Predictor (binary classifier와 유사한 알고리즘이며, 연속형 숫자 범위의 결과값을 갖는 모델들을 추정하고 비교)

데이타솔루션

Modeler Plus Pack

I. SPSS Modeler

(1) Introduction
(2) SPSS Modeler 방법론
(3) SPSS Modeler 기능구성

II. SPSS Modeler 특장점

(1) 다변화 작업에 적합한 구성과 Interface
(2) 개방형 시스템(Open Architecture)
(3) SQL Optimization
(4) Bulk Loading
(5) Learning Cost 감소

III. SPSS Modeler Plus Pack

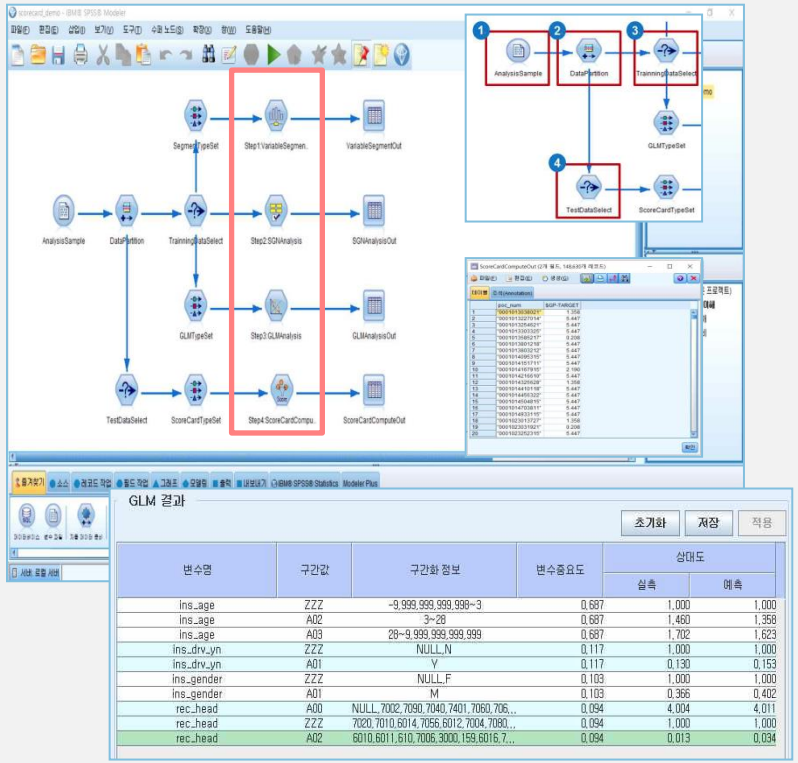
(1) Score Card
(2) Data Visualization
(3) Statistical Utility
(4) Model Extraction
(5) TA Korean

SPSS Modeler Plus Pack

III. SPSS Modeler Pus Pack - Score Card

금융기관 및 보험사에서 어떤 고객이 우량고객인지 불량고객인지 알고 선별해 내는 것은 모든 금융기관의 주된 관심사이며, 이를 예측하고 선별하기 위해서는 스코어카드 모델링 기법이 필요합니다.

스코어카드 모델링이란 우량/불량으로 정의된 Target 변수와 이에 영향을 주는 독립변수(고객 특성)를 사용하여 어떤 특성을 가진 고객이 우량이고 불량인지를 예측하는 모델링입니다.



1단계 : 변수 구관화

→ 분석에 사용할 변수를 범주형으로 구간화 하는 단계

2단계 : 유의성 분석

→ 타겟으로 선택한 변수에 대하여 분석에 이용할 변수가 어느정도의 유의성을 가지는지 분석하는 단계

3단계 : 일반화선형모형(GLM)분석

→ 타겟에 대하여 분석을 선택한 변수를 이용하여 일반화선형모형(GLM) 분석을 실행하는 단계

4단계 : 스코어카드생성

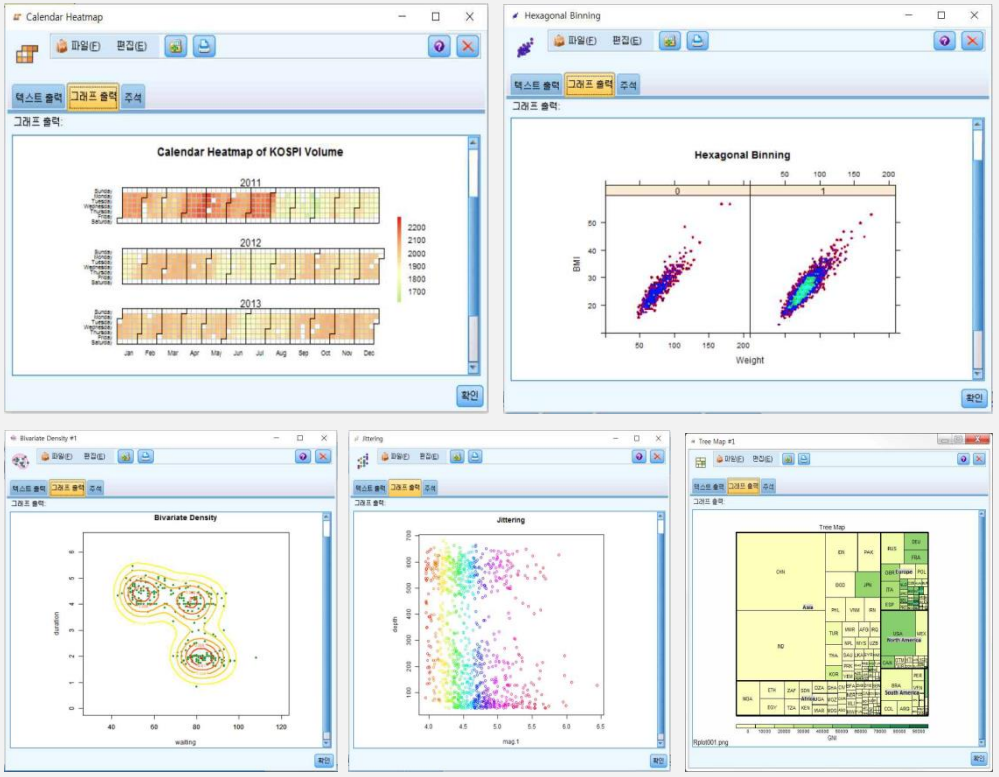
→ 3단계에서 산출된 일반화선형모형(GLM) 분석 결과를 바탕으로 신규 입력 데이터에 대하여 스코어를 산출하는 단계

SPSS Modeler Plus Pack

III. SPSS Modeler Pus Pack - Data Visualization

SPSS Modeler 의 그래프 노드에 데이터 특성을 표현하기 위한 다양한 데이터 시각화 기능을 추가할 수 있습니다.

그래프 기능 : Calendar Heatmap, Bivariate Density, Hexagonal Binning, Jittering, Tree Map.



1. Calendar Heatmap (채색달력)

→ 년도(year)별로 365일이 월(Month)과 주(Week)로 구분

전체적인 시계열의 흐름 외에 계절 효과와 요일 효과 확인

2. Bivariate Density

→ 산점도 : 밀도 등고선, 특이점 마킹, 이변량 히스토그램

3. Hexagonal Binning

→ 개체수 10,000 이상인 경우 hexagonal binning (육각형 격자 나누기) 로 표현

4. Jittering

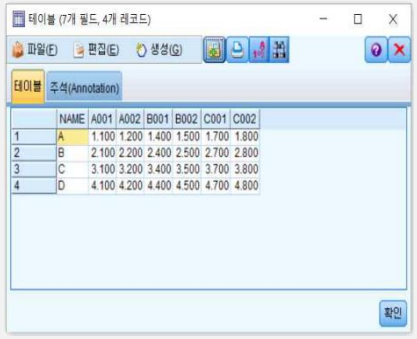
→ 일정간격의 연속형 변수, 정수로 코딩된 범주형 변수 표현

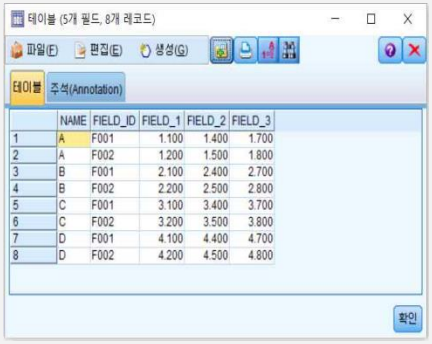
5. Tree Map

→ 계층적으로 타일(Tile)을 배열하여 붙인 통계그림

III. SPSS Modeler Pus Pack - Statistical Utility

Statistical Utility는 SPSS Modeler에 부족한 통계분석 기능을 개발하여 추가한 모듈 입니다.
키 필드별로 순위를 매기는 기능, 필드를 케이스별로 항목을 나누는 기능 등 사용자들이 많이 사용하지만 현재 Modeler에 없는 모듈을 제공합니다.
기능 : Rank by Key, Field to Case 변환, 다빈도 항목 추출





1. Rank by Key

→ 키 필드인 기준 값을 비교하여 키 필드별로 1,2,3,.. 혹은 1,1,3,3,... 등 원하는 모형으로 데이터 순위를 출력한다

2. Field to Case

→ 사용자가 지정한 그룹과 필드 수에 따라 필드 데이터를 행 데이터로 변환한다.

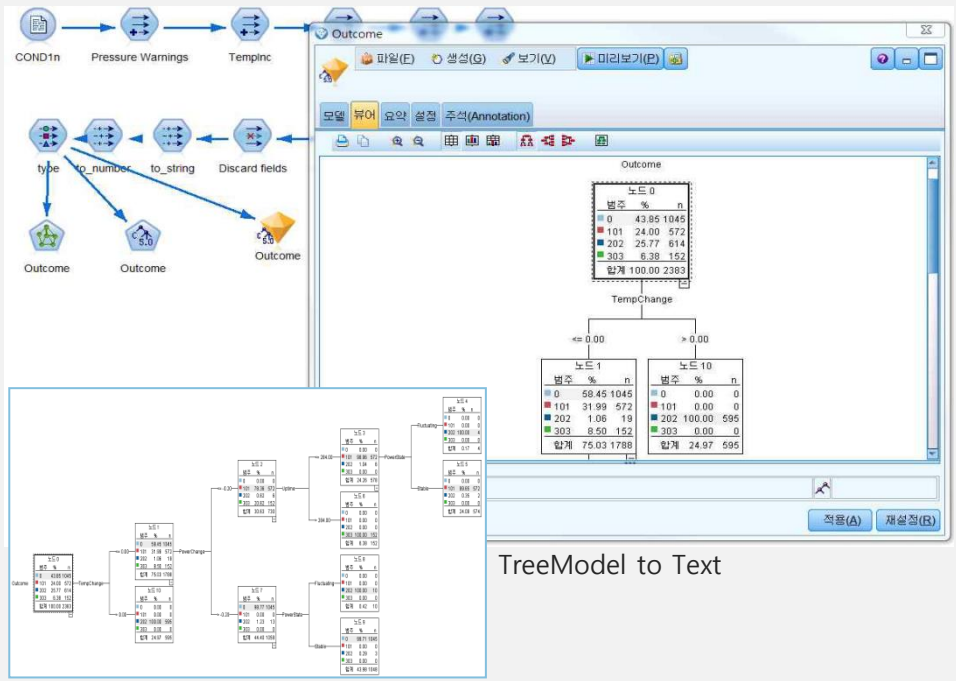
3. 다빈도 항목 추출

→ 다빈도 항목 조합만 출력

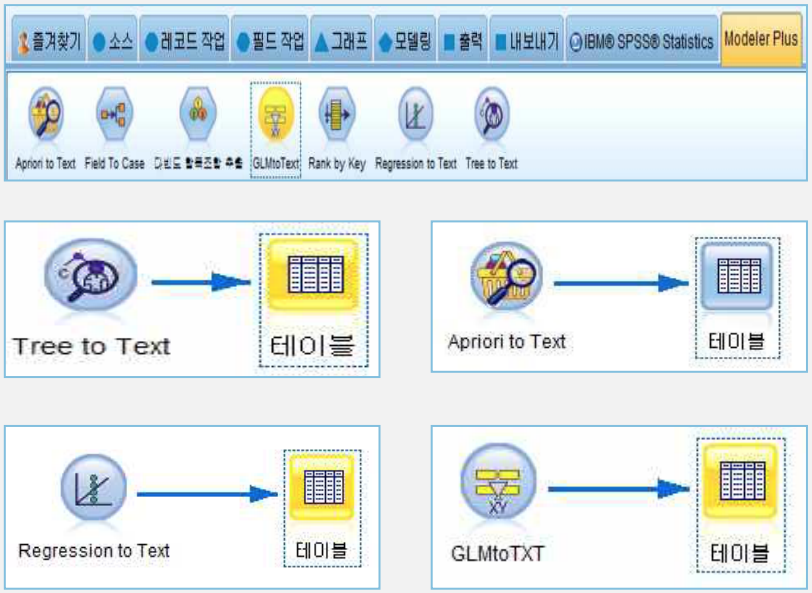
III. SPSS Modeler Pus Pack - Model Extraction

SPSS Modeler로 스트림 작성 후 모델링 결과를 다시 활용 해야 할 경우 ModelerPlus의 Model Extraction에서 제공하는 노드를 통해 활용 할 수 있습니다. Model Extraction 노드는 C5.0과 GLM, CHAID, 선형회귀, Apriori 모델링에 대해 지원합니다. 이 노드를 사용하여 좀 더 폭 넓은 데이터 활용 과 확장 된 스트림을 작성 할 수 있습니다.

기능 : TreeModel to Text, GLM to Text, Regression to Text, Apriori to Text

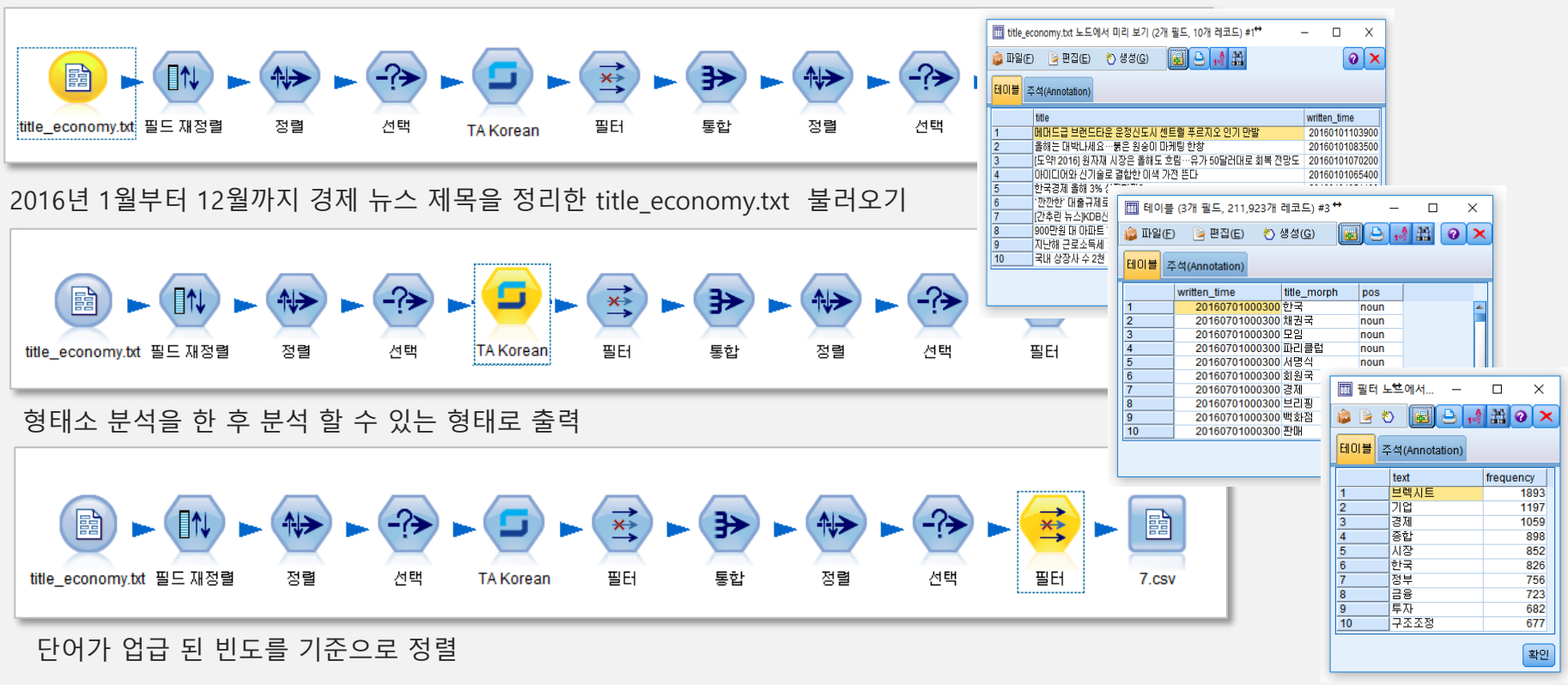


TreeModel to Text



III. SPSS Modeler Pus Pack - TA Korean

SPSS Modeler에 한글 형태소를 추출하여 분석을 할 수 있는 노드를 추가 하였습니다.
출력 시 두 가지 형태로 출력이 되며 (Record by record, Record by value) 품사 선택이 가능합니다.



데이타솔루션

Thank you