STSCI 5010 Module 2

SAS EM and Cluster Analysis

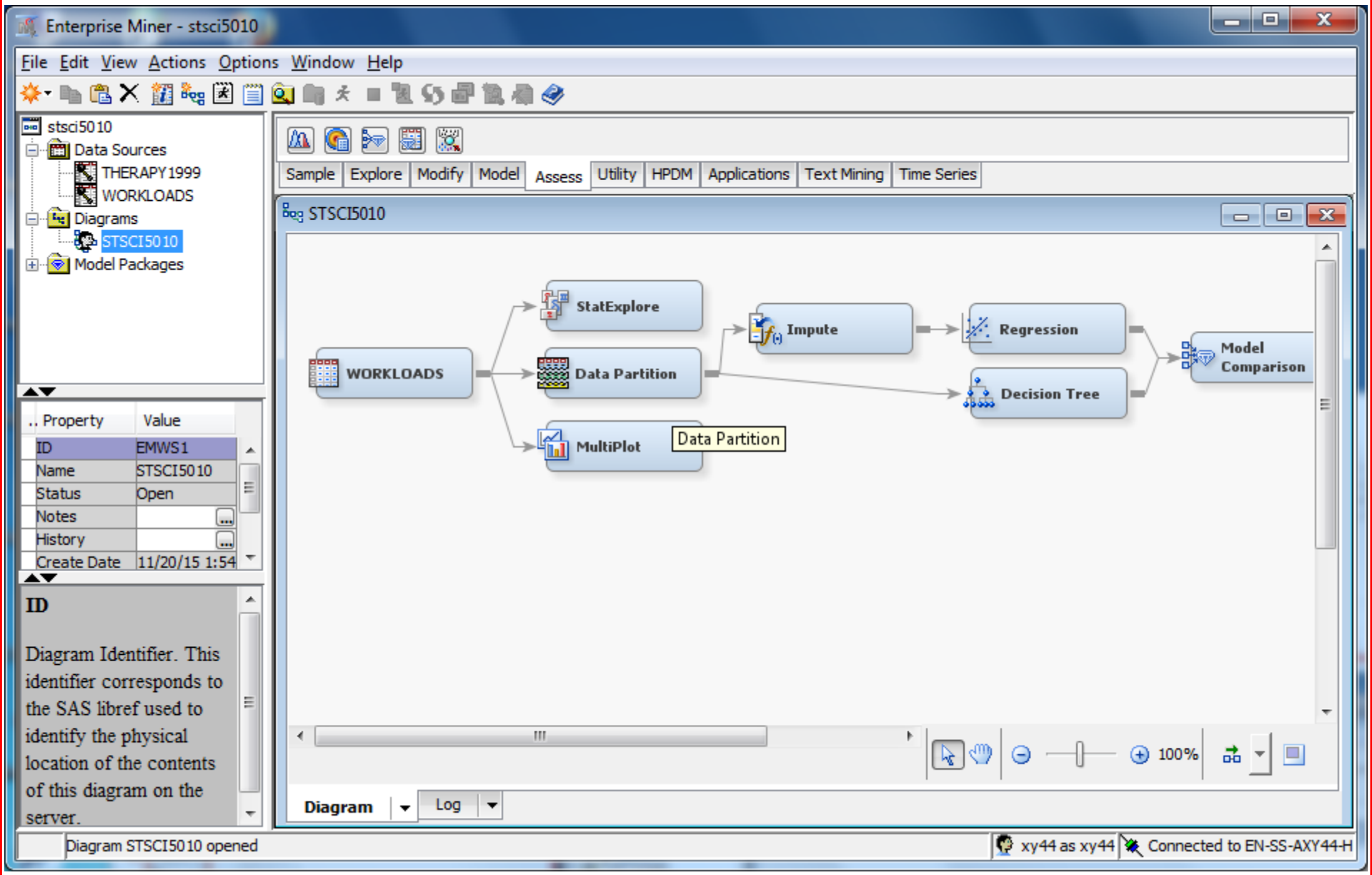
Xiaolong Yang • Statistical Science • Cornell University

Introduction to SAS Enterprise Miner (EM) and Data Visualization

What is SAS Enterprise Miner (EM)?

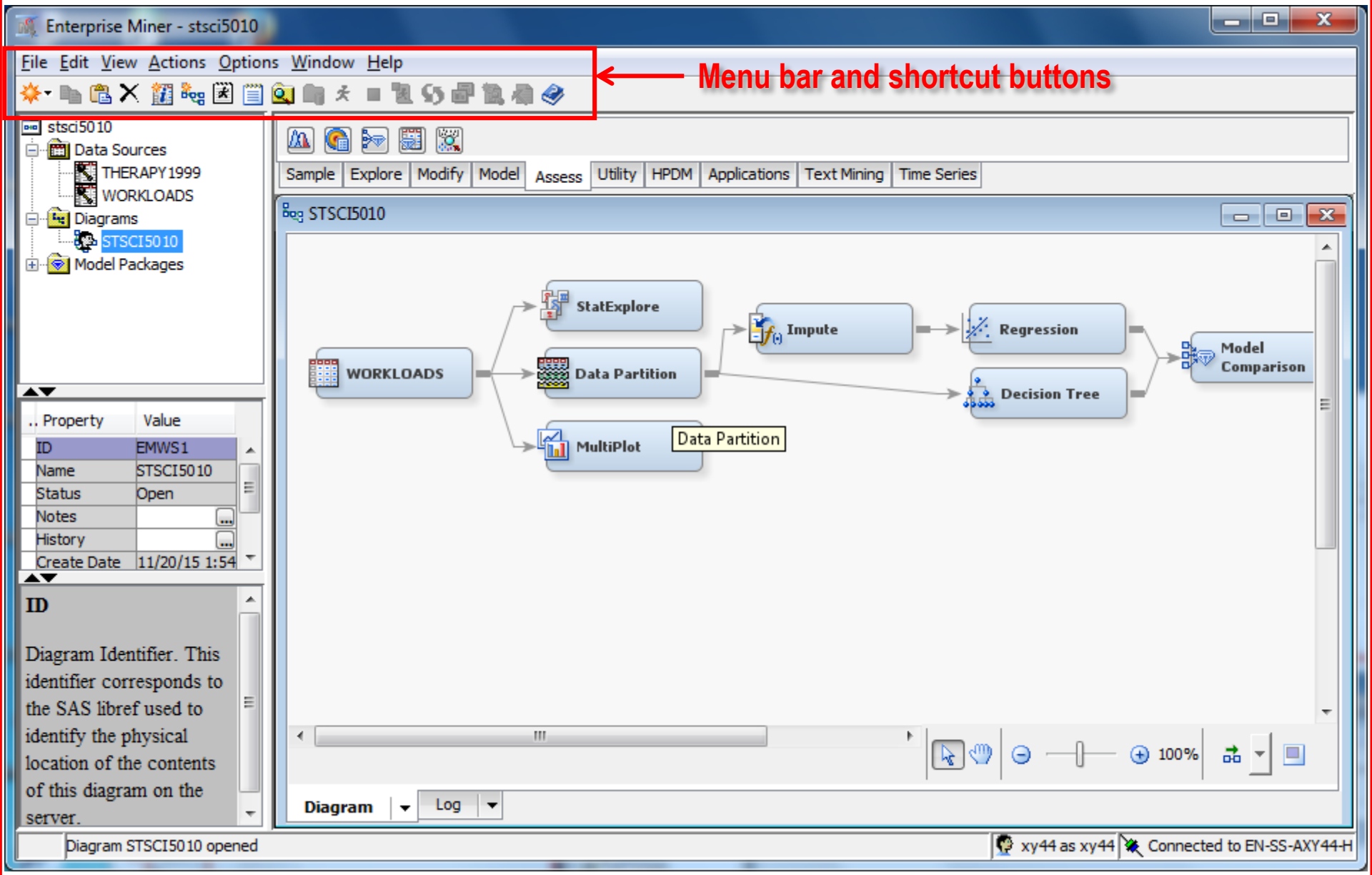
- A part of SAS software package
- A JAVA-based visual tool for data mining
- Simplifies many common tasks associated with applied analysis
- Provides a wide variety of tools with a consistent graphical interface
- Incorporates user-defined methods
- For investigating large complex data sets (thousands to millions of observations, hundreds to thousands of variables)

AS Enterprise Miner 14.2 Interface*

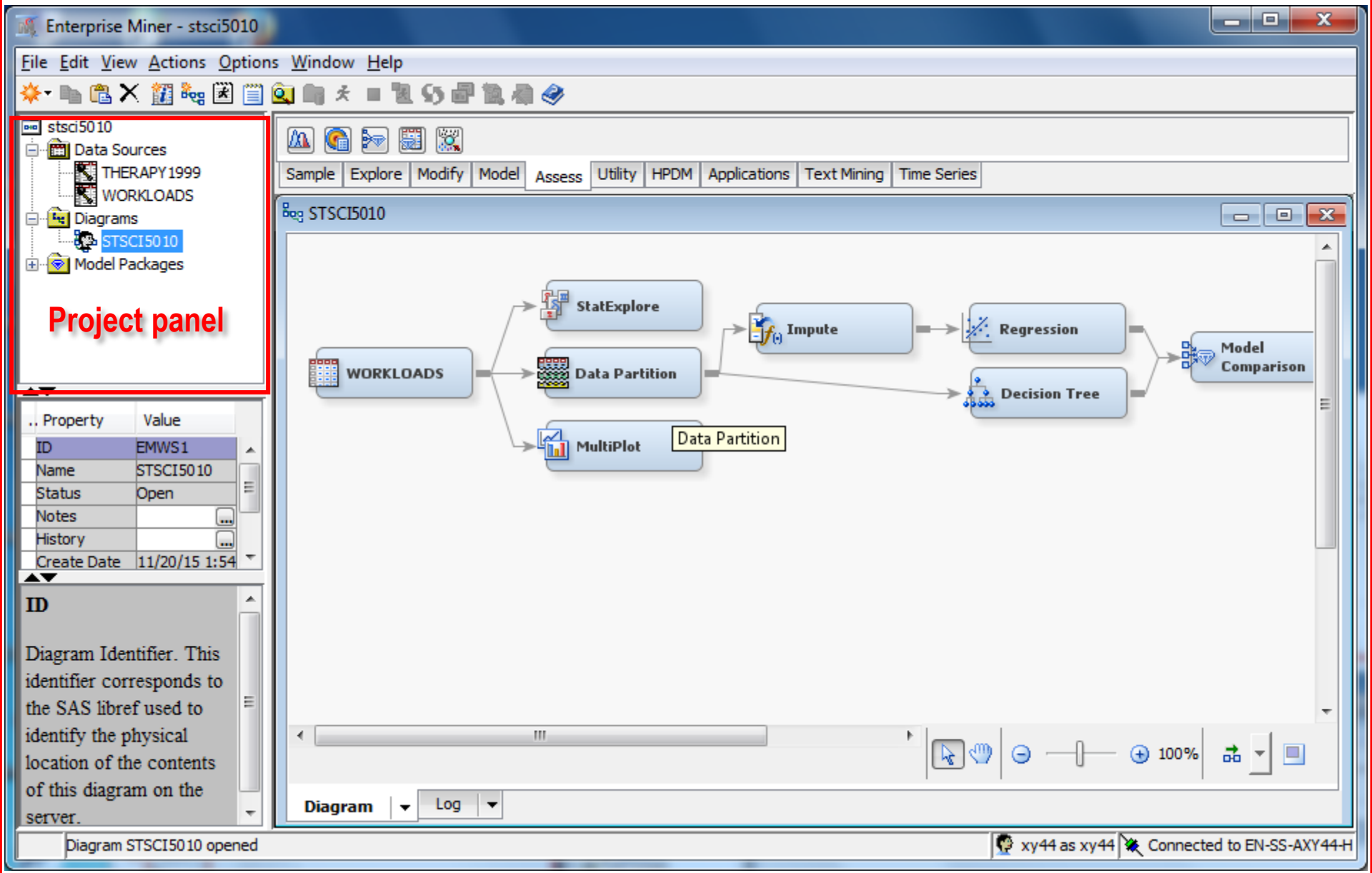


* The current version is 14.12?

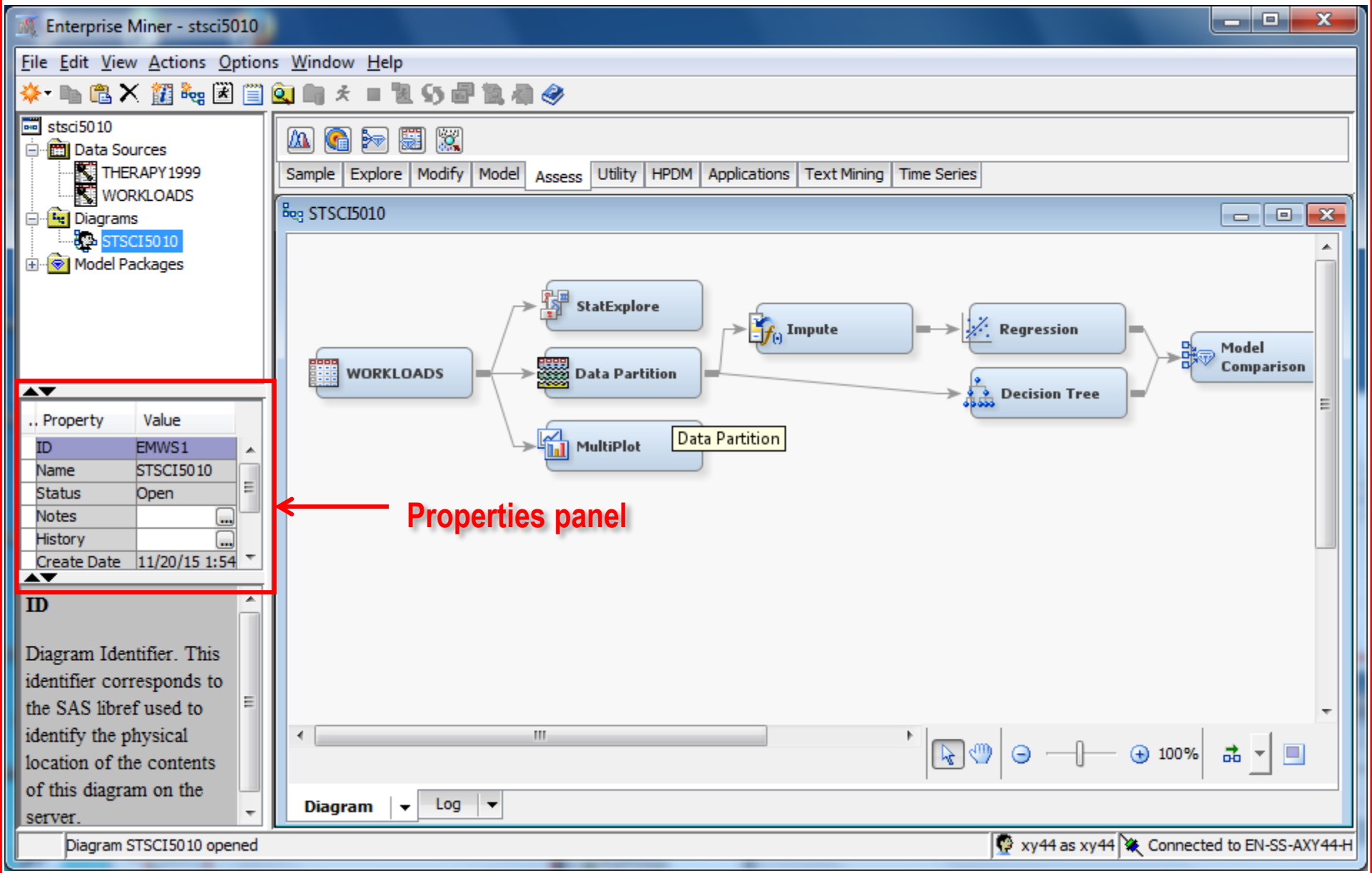
SAS Enterprise Miner – Interface Tour



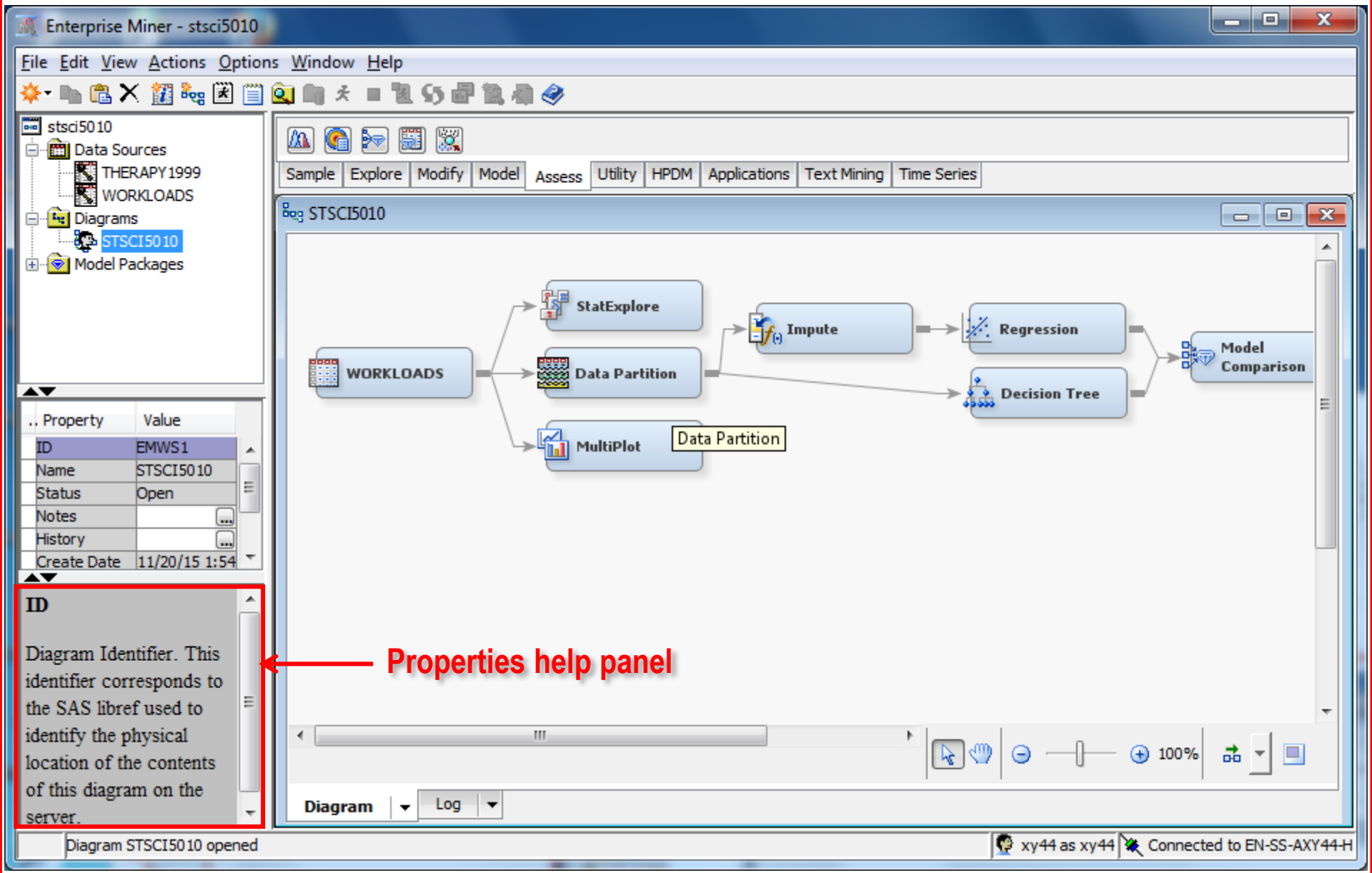
SAS Enterprise Miner – Interface Tour



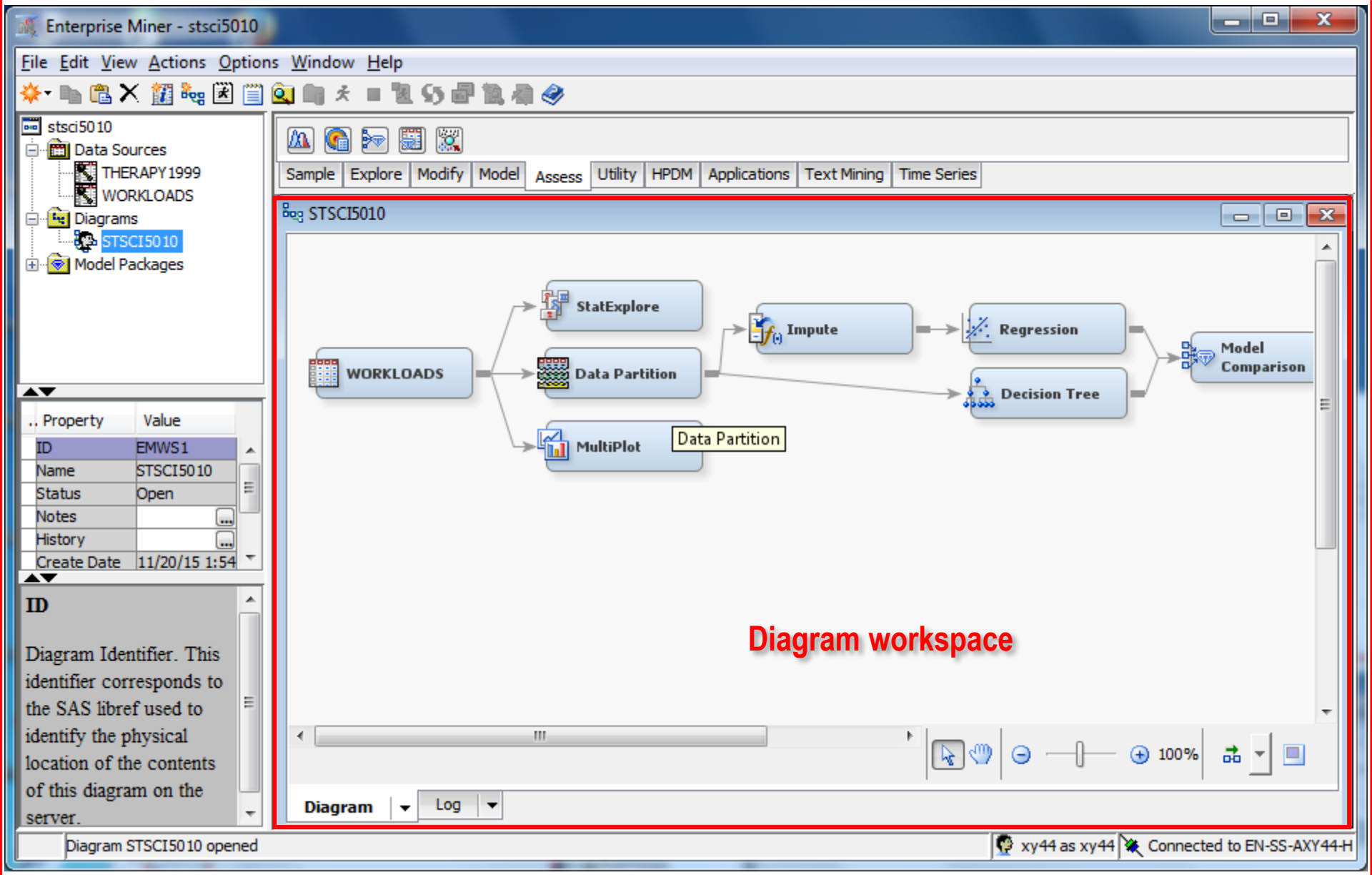
SAS Enterprise Miner – Interface Tour



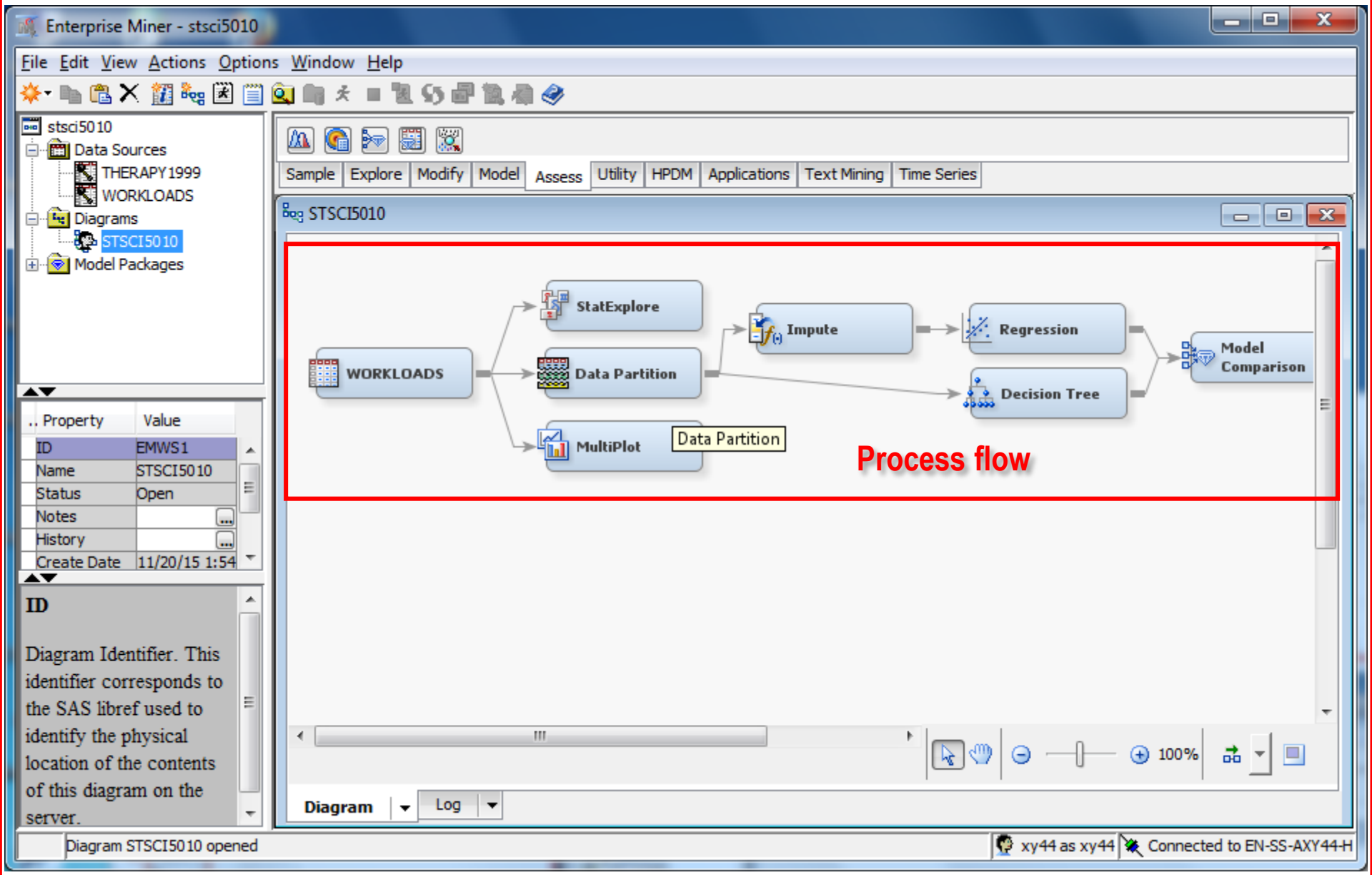
SAS Enterprise Miner – Interface Tour



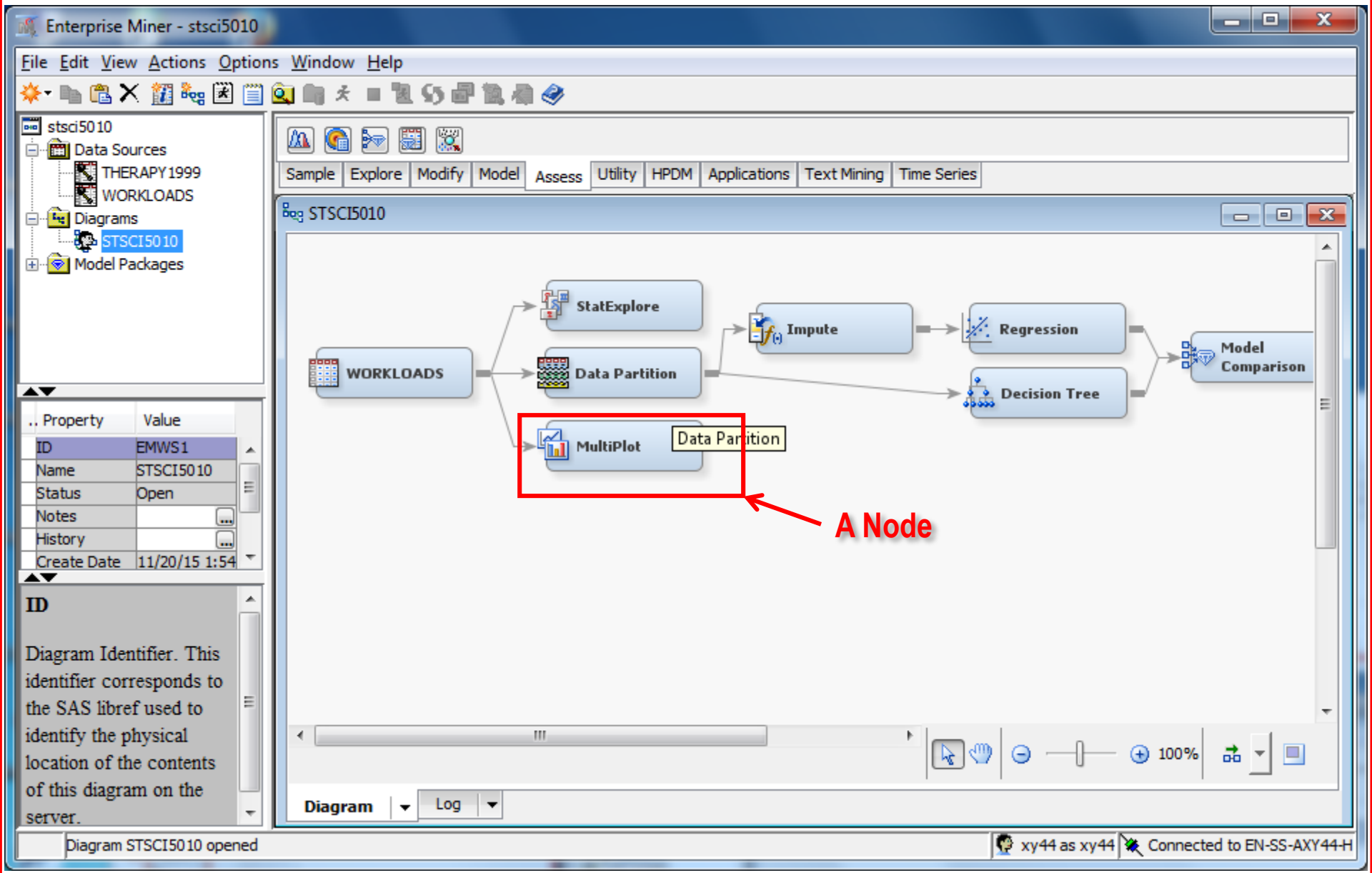
SAS Enterprise Miner – Interface Tour



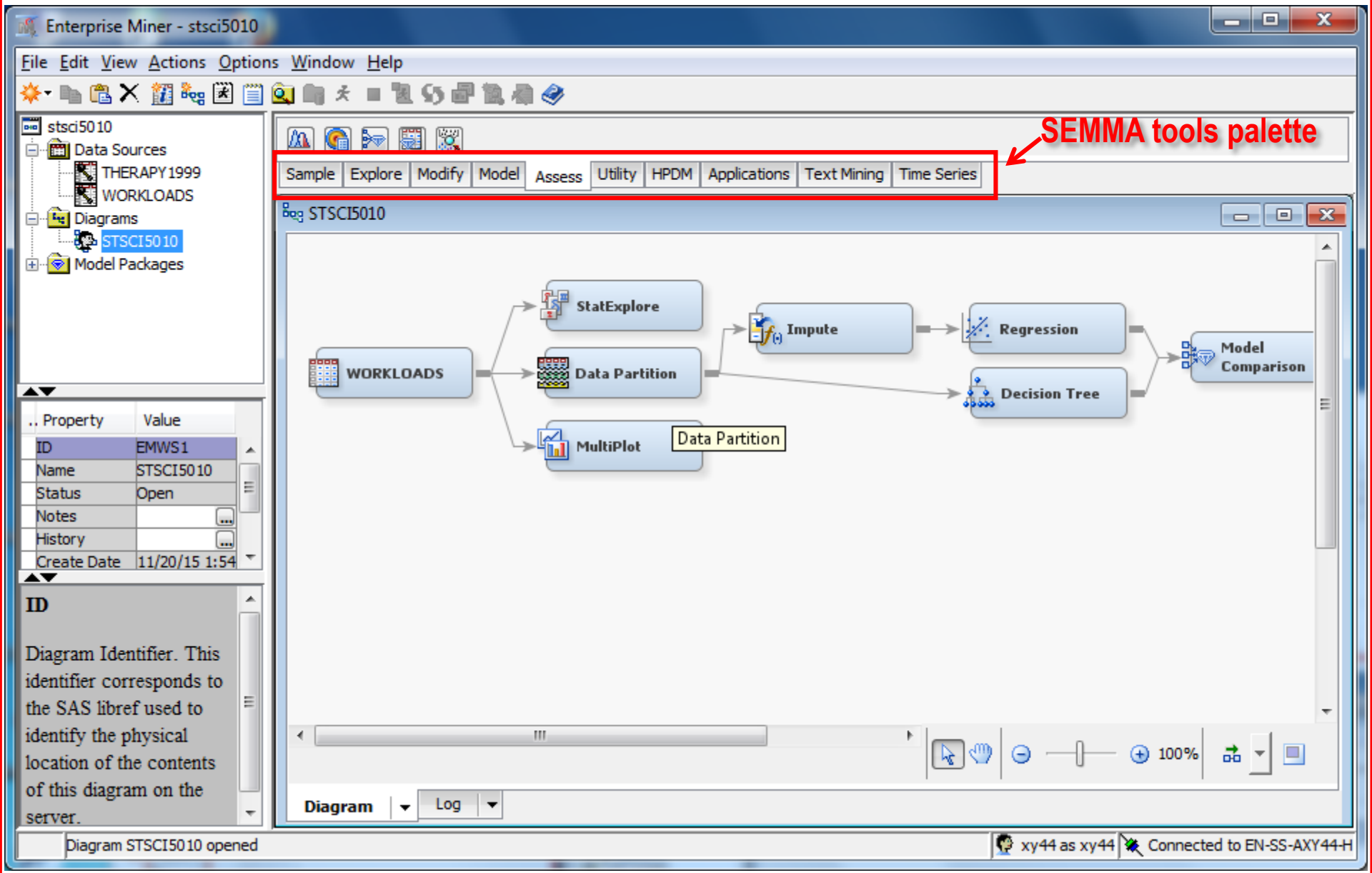
SAS Enterprise Miner – Interface Tour



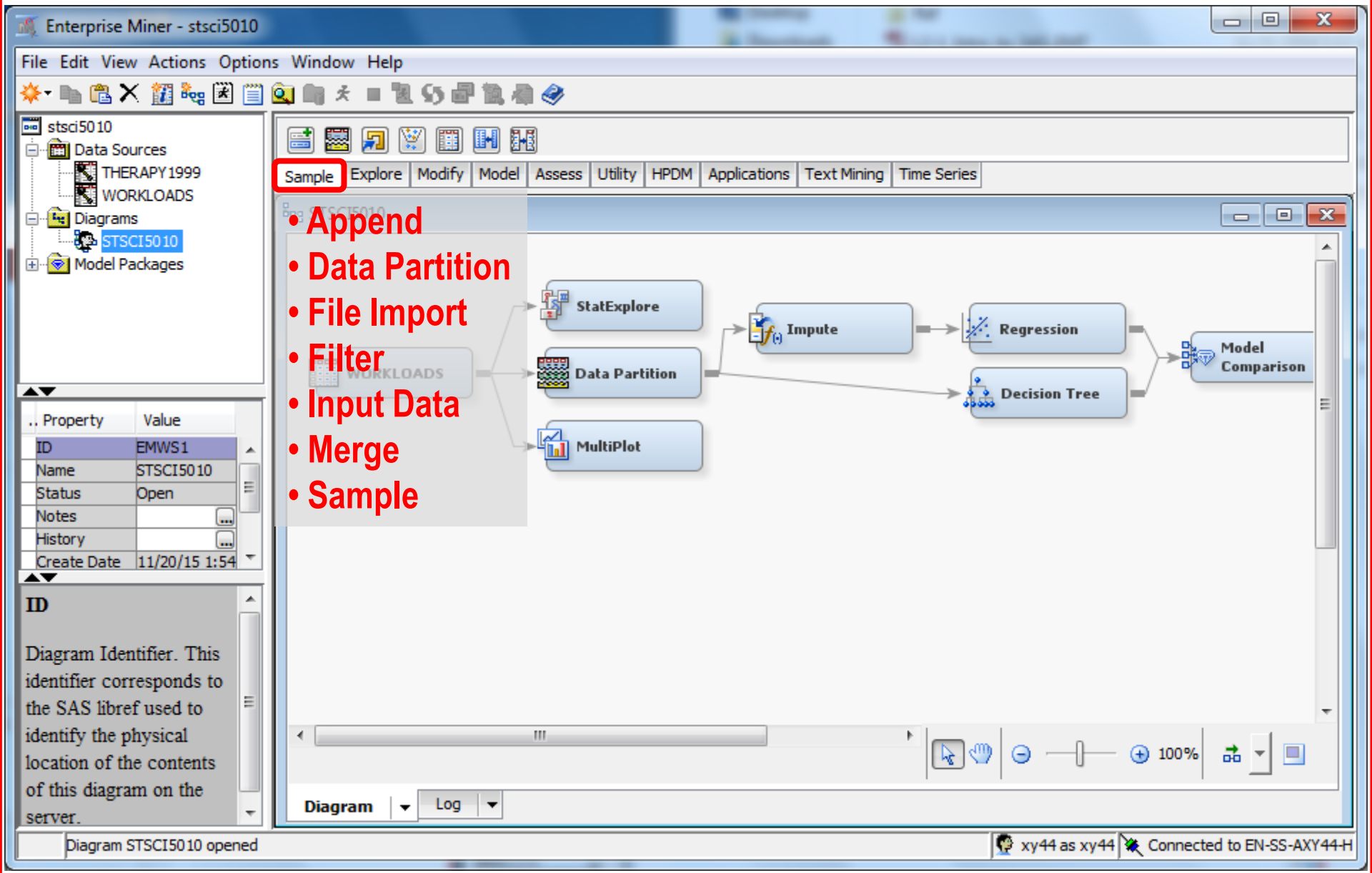
SAS Enterprise Miner – Interface Tour



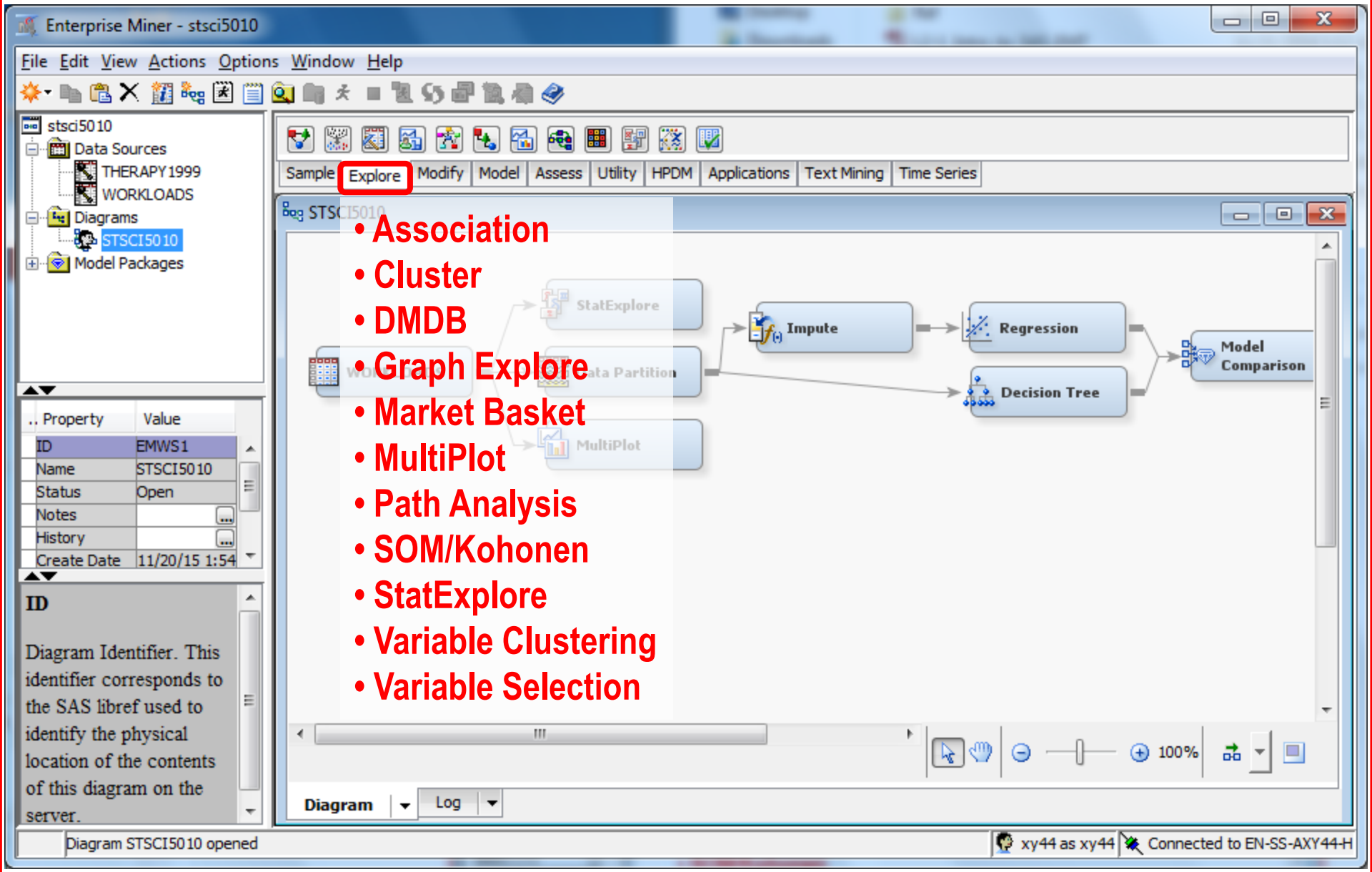
SAS Enterprise Miner – Interface Tour



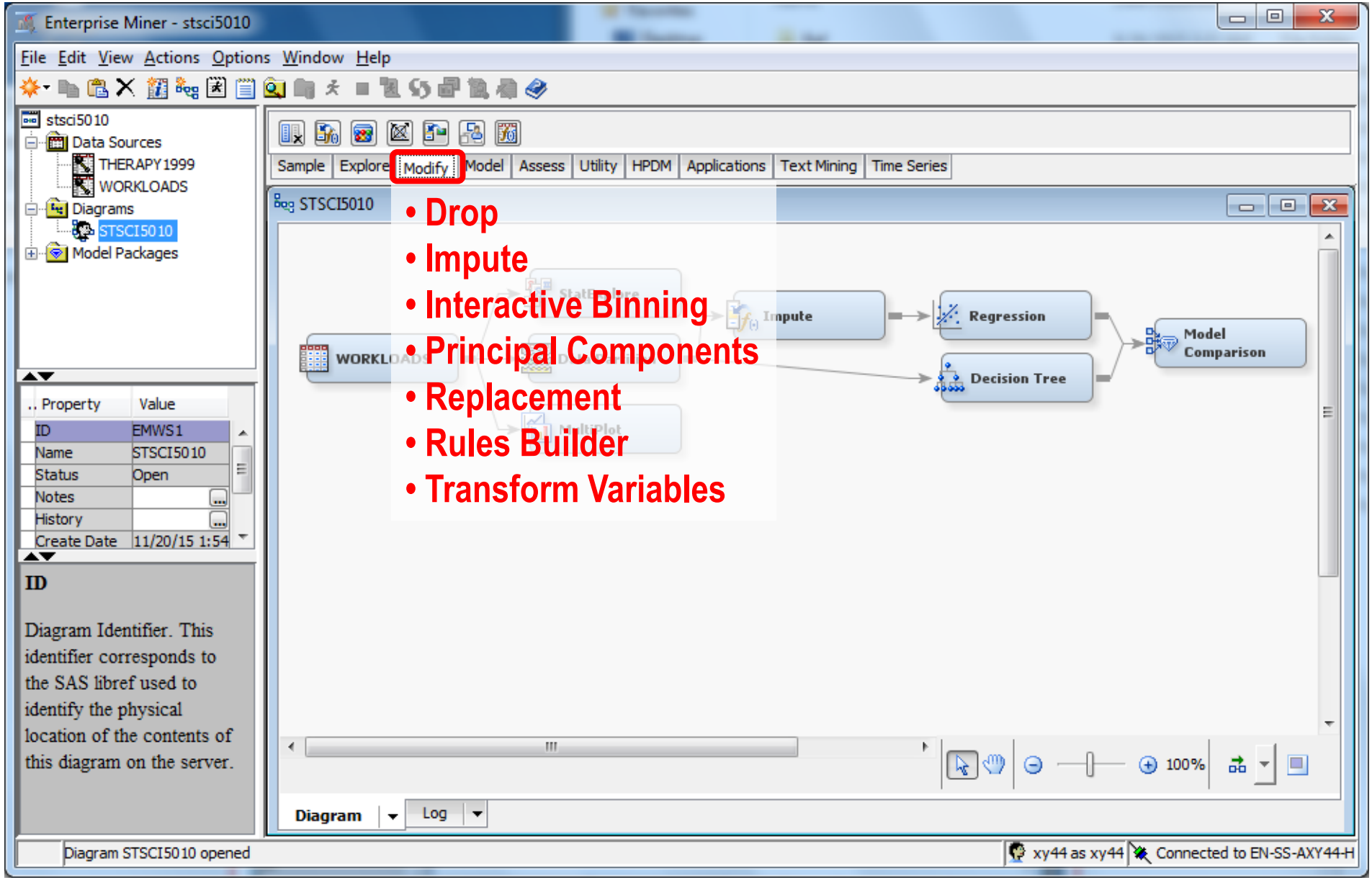
SAS Enterprise Miner – Sample Tab



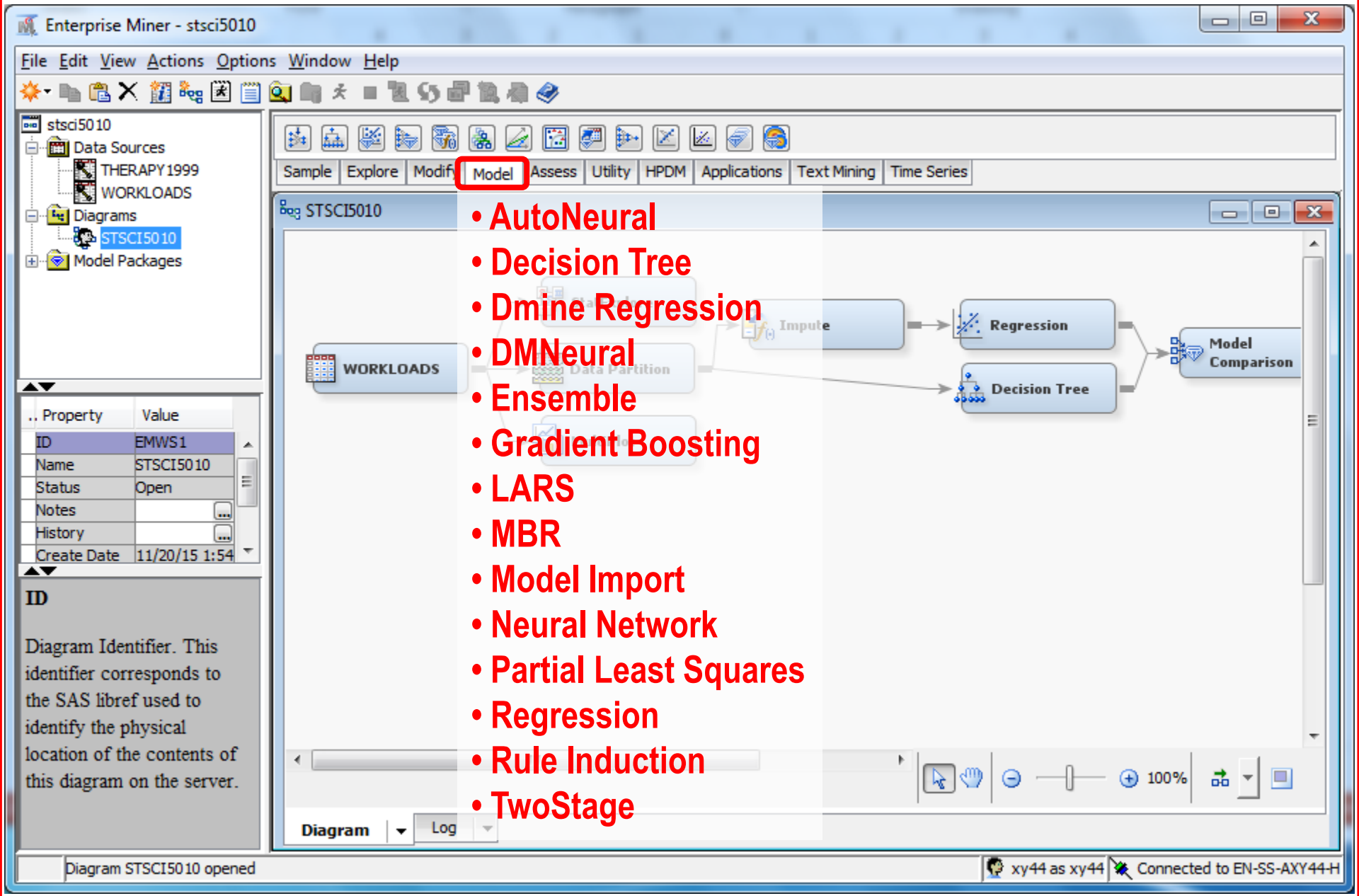
SEMMA – Explore Tab



SEMMA – Modify Tab

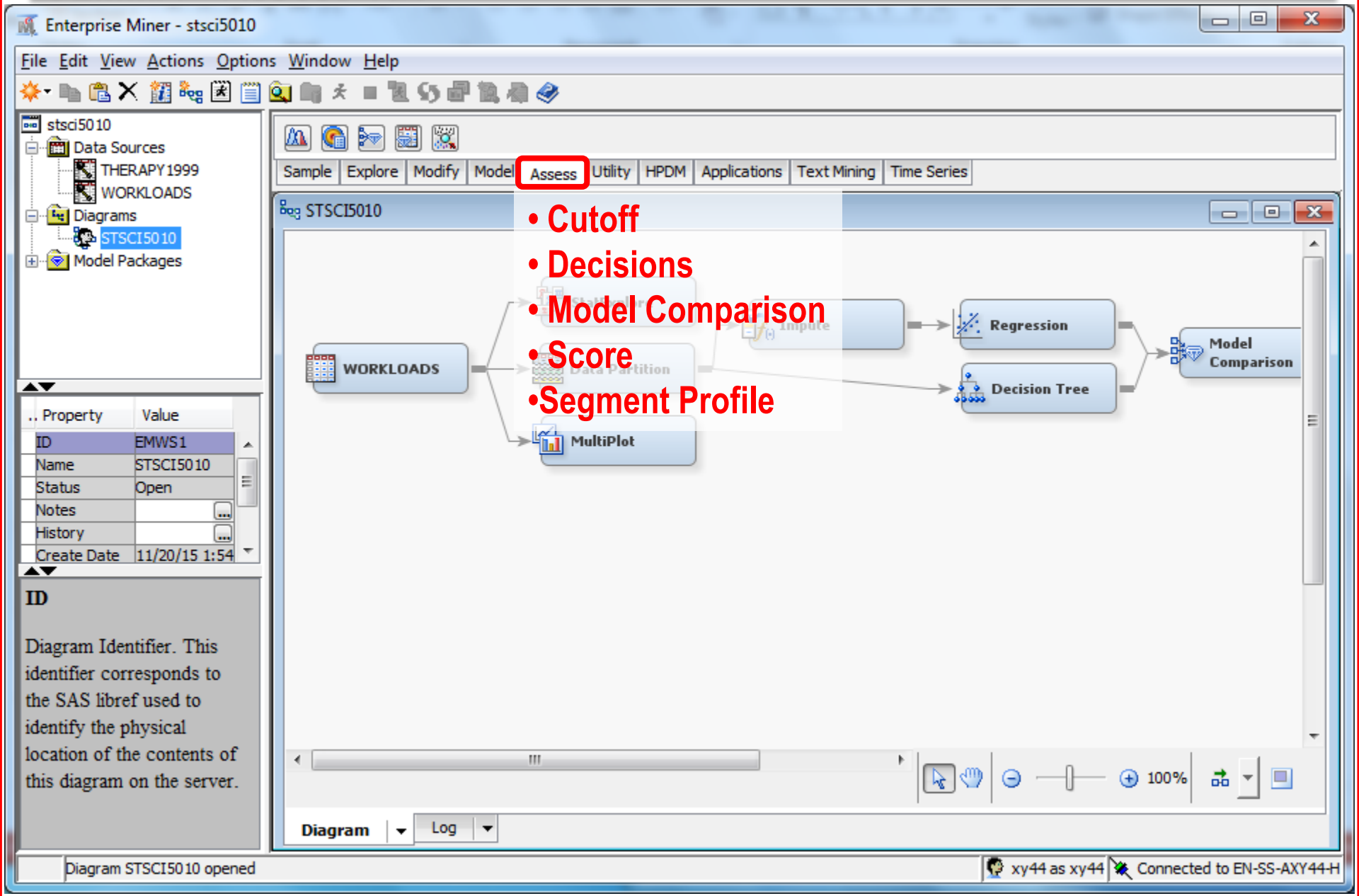


SEMMA – Model Tab

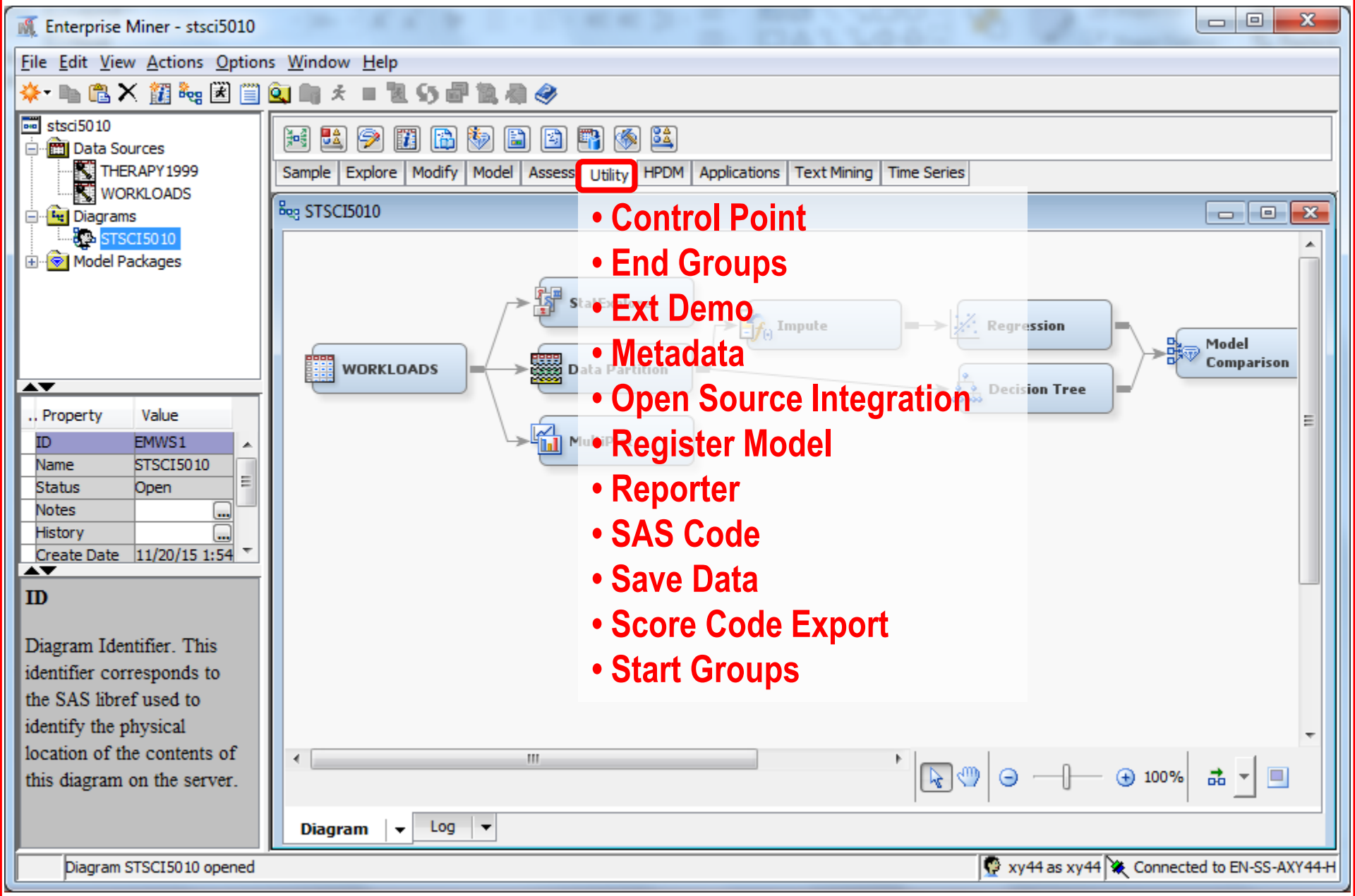


- AutoNeural
- Decision Tree
- Dmine Regression
- DMNeural
- Ensemble
- Gradient Boosting
- LARS
- MBR
- Model Import
- Neural Network
- Partial Least Squares
- Regression
- Rule Induction
- TwoStage

SEMMA – Assess Tab

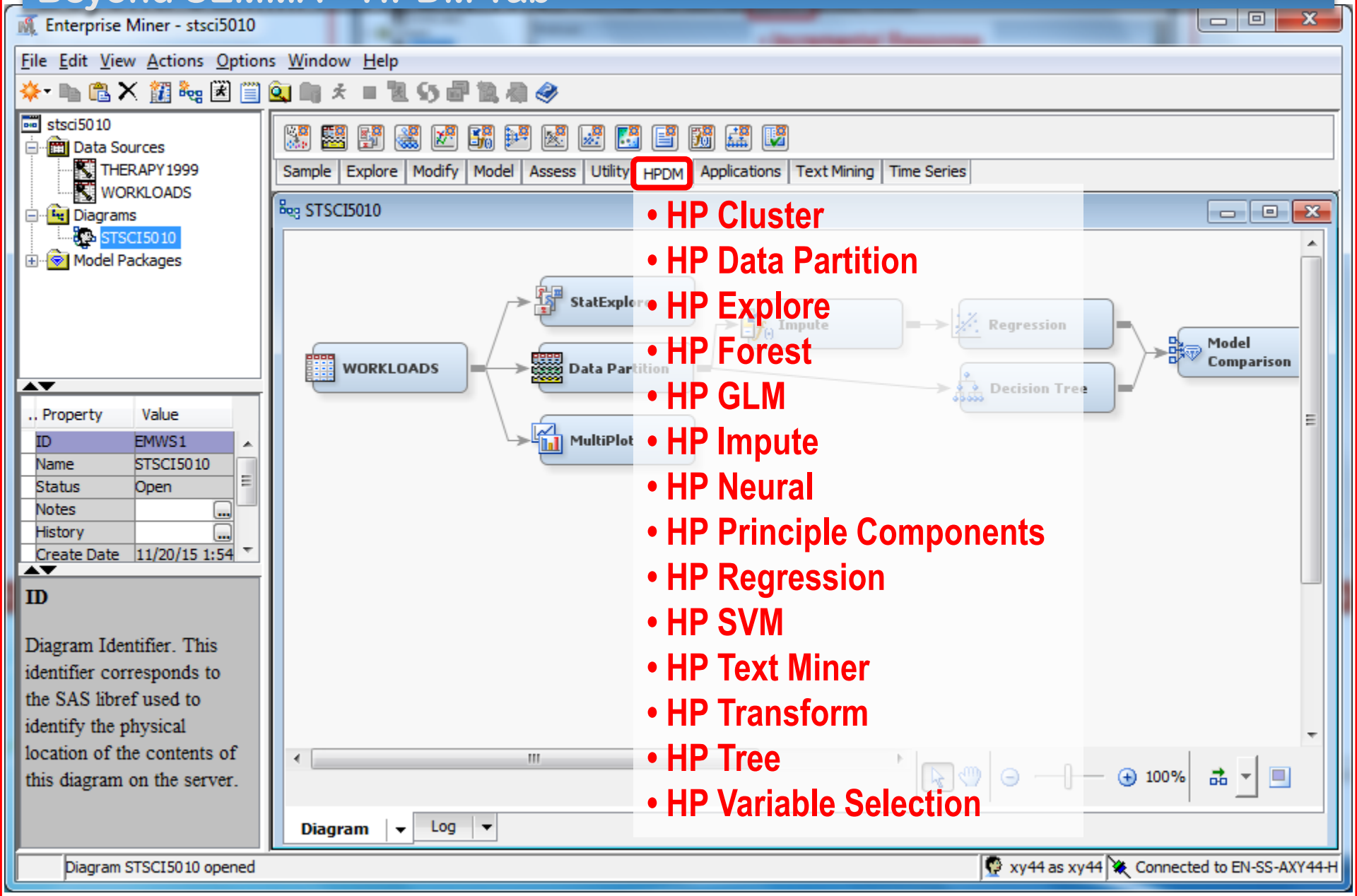


Beyond SEMMA – Utility Tab

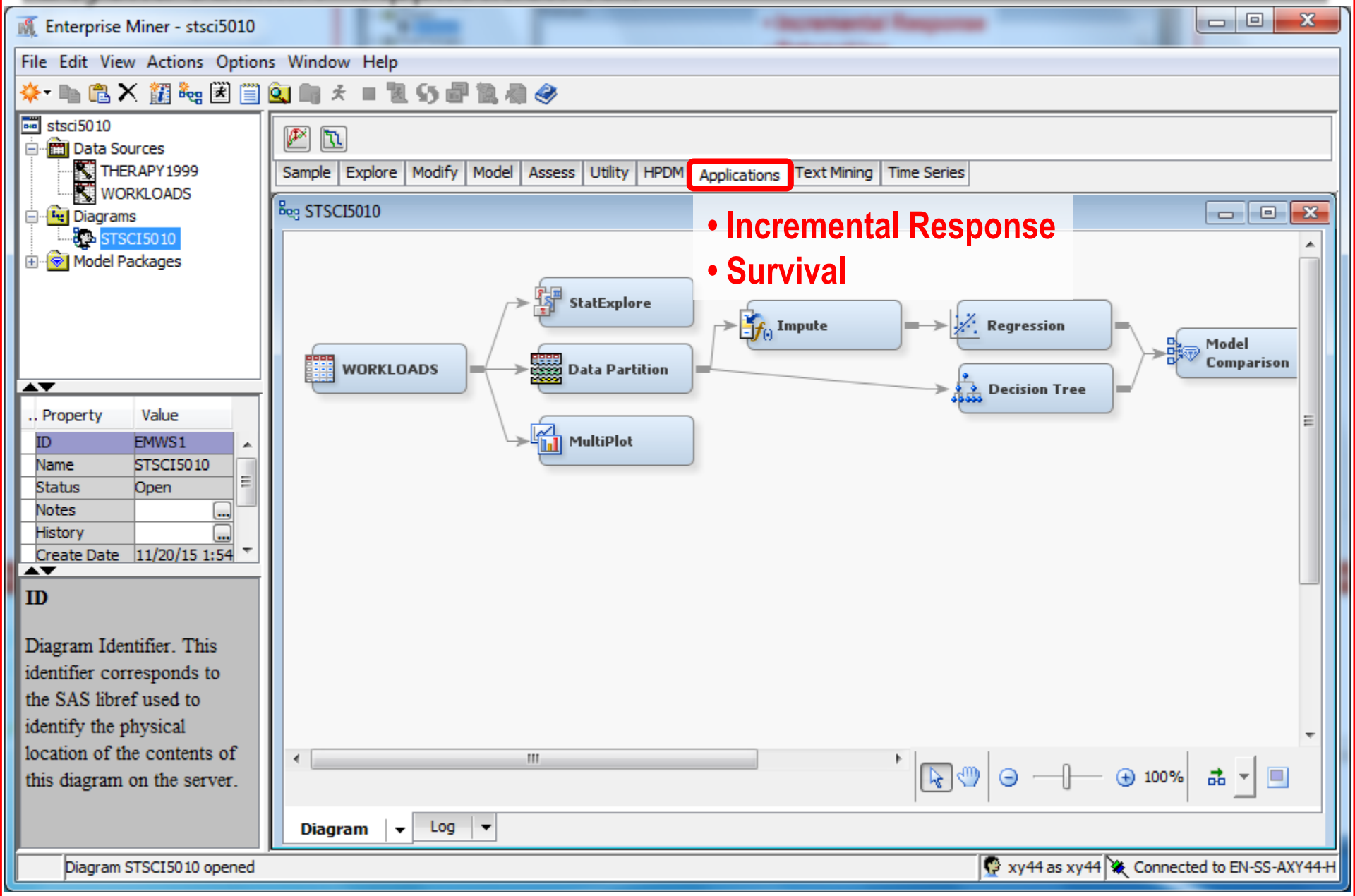


- Control Point
- End Groups
- Ext Demo
- Metadata
- Open Source Integration
- Register Model
- Reporter
- SAS Code
- Save Data
- Score Code Export
- Start Groups

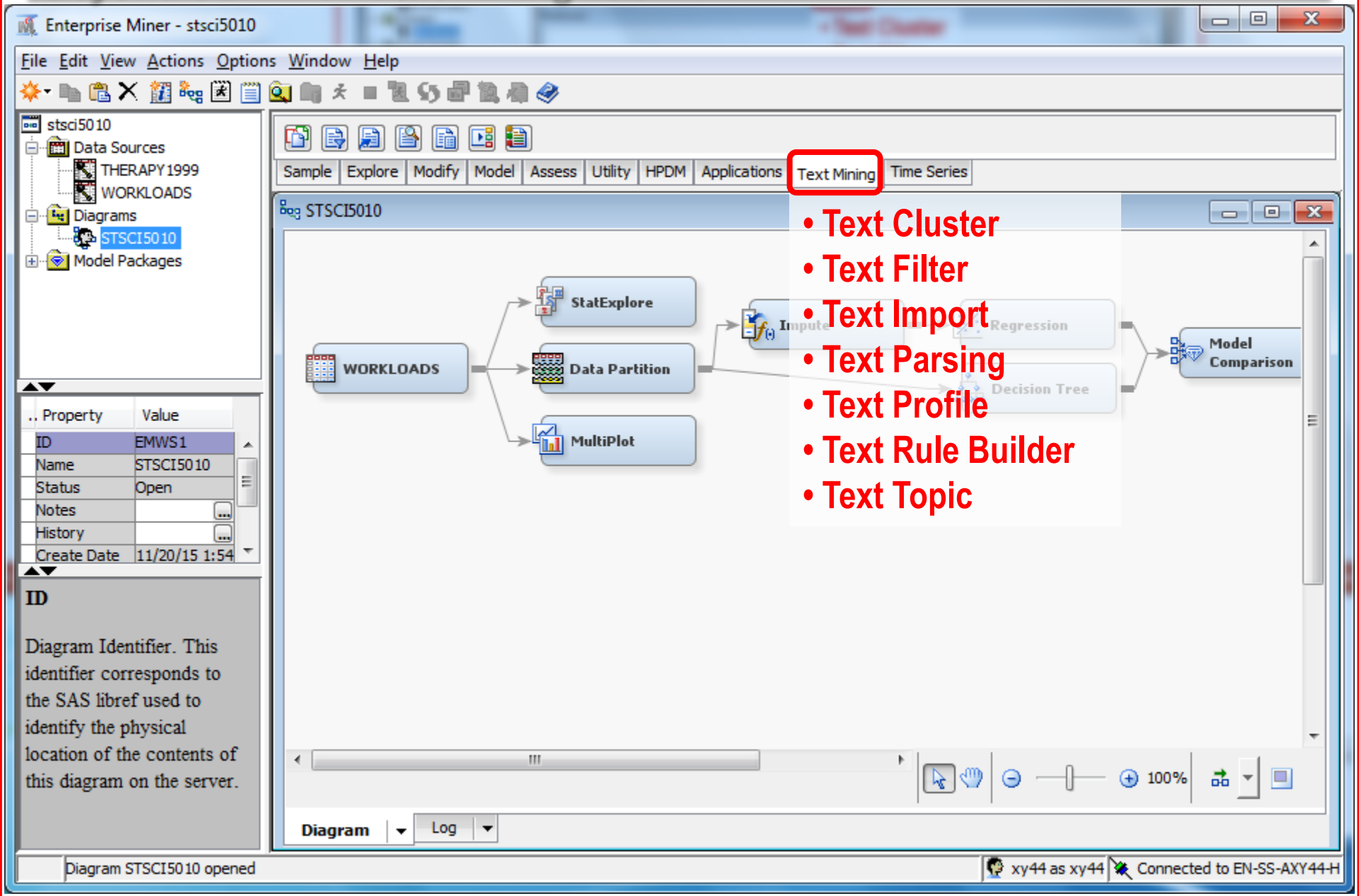
Beyond SEMMA – HPDM Tab



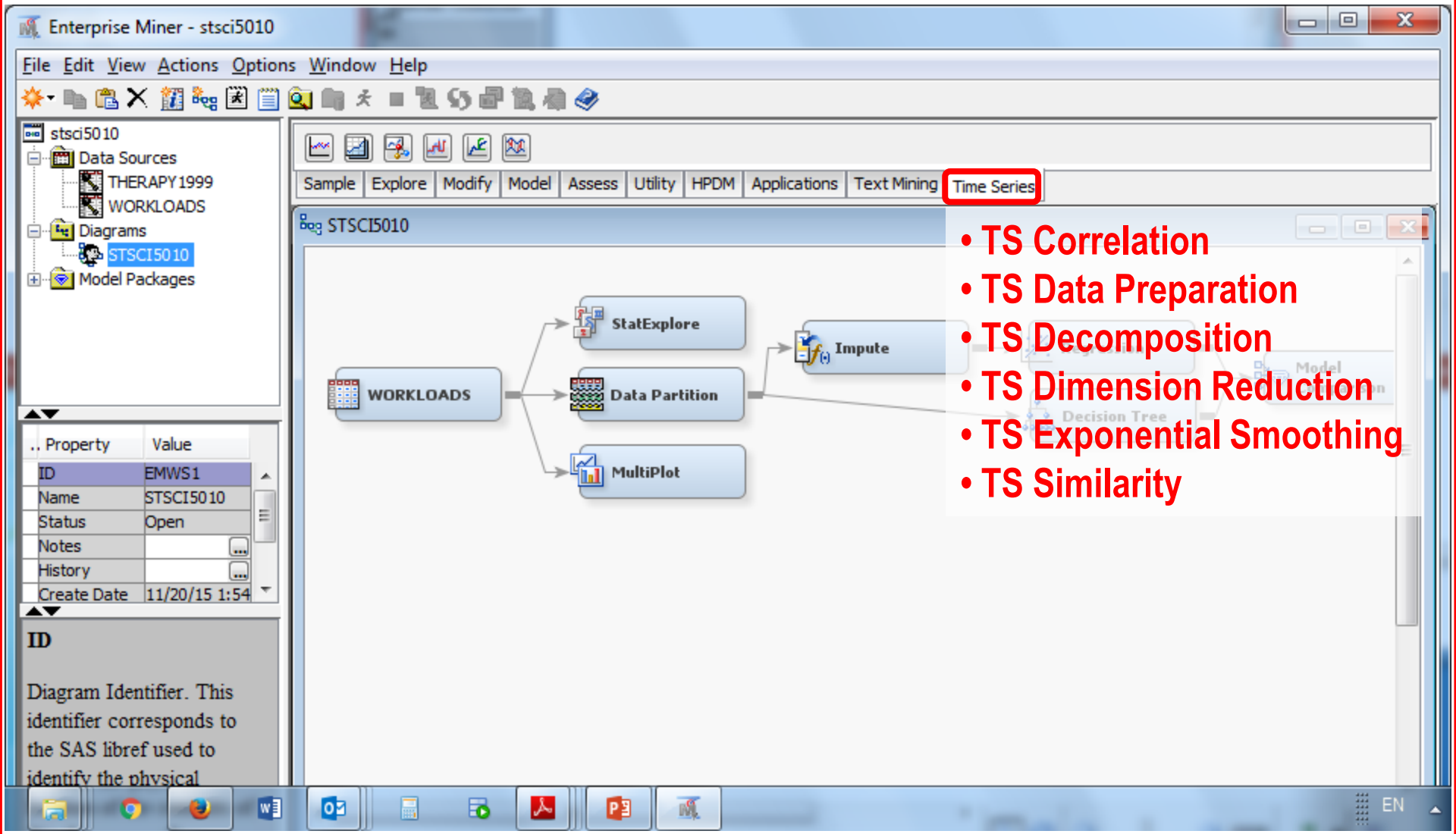
Beyond SEMMA – Application Tab



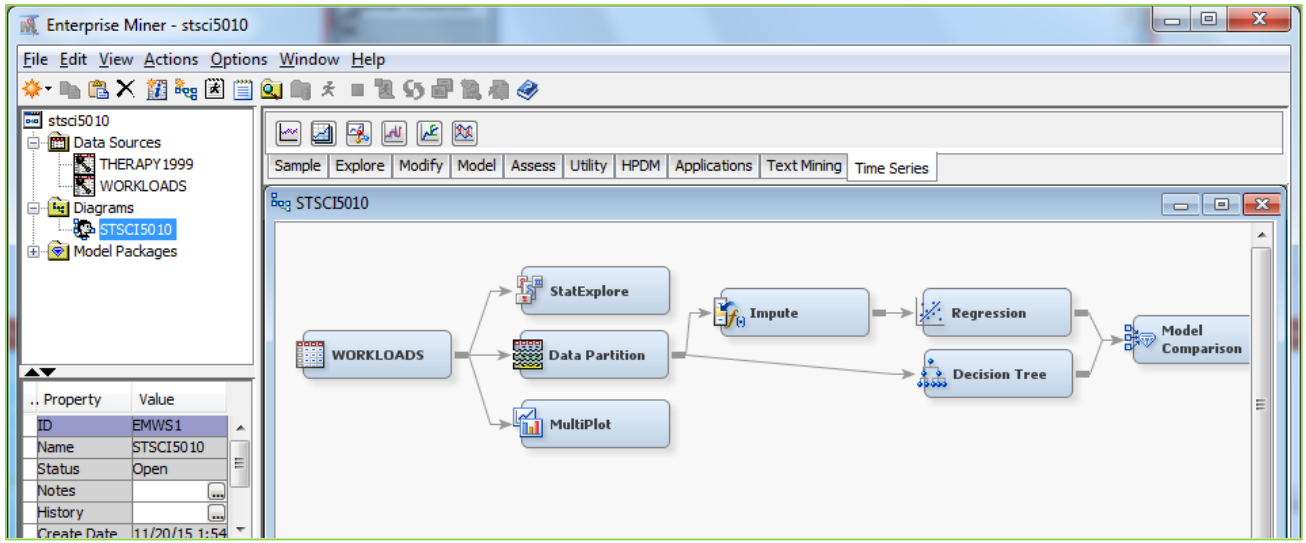
Beyond SEMMA – Text Mining Tab

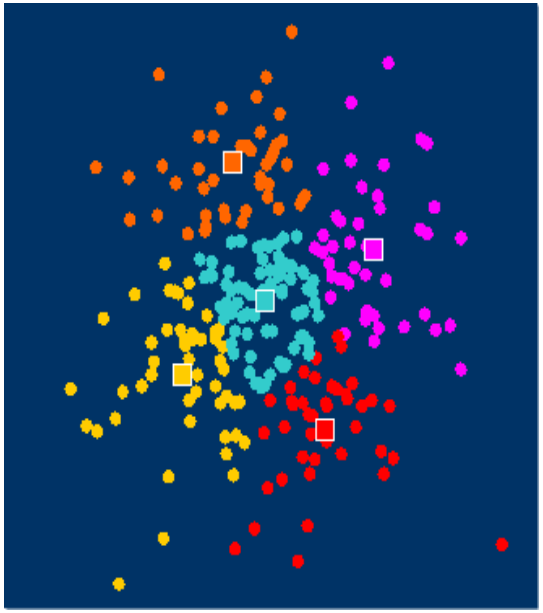


Beyond SEMMA –Time Series Tab

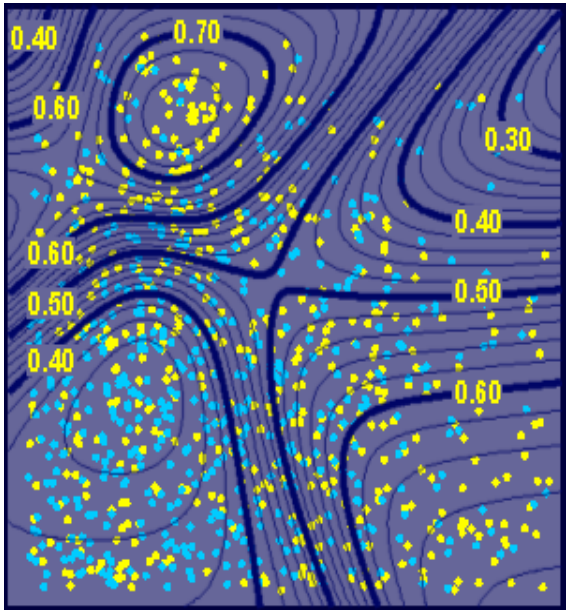


SAS Enterprise Miner Analytic Strengths





Pattern Discovery



Predictive Modeling

Setting up an EM Project

Start SAS Enterprise Miner 14.2

Start

↓

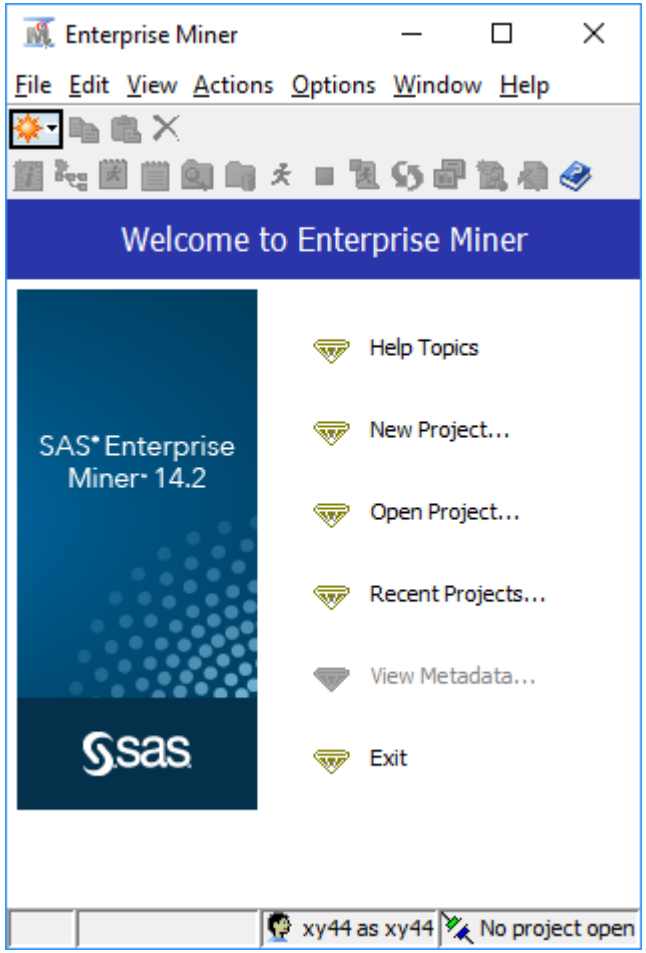
All Programs

↓

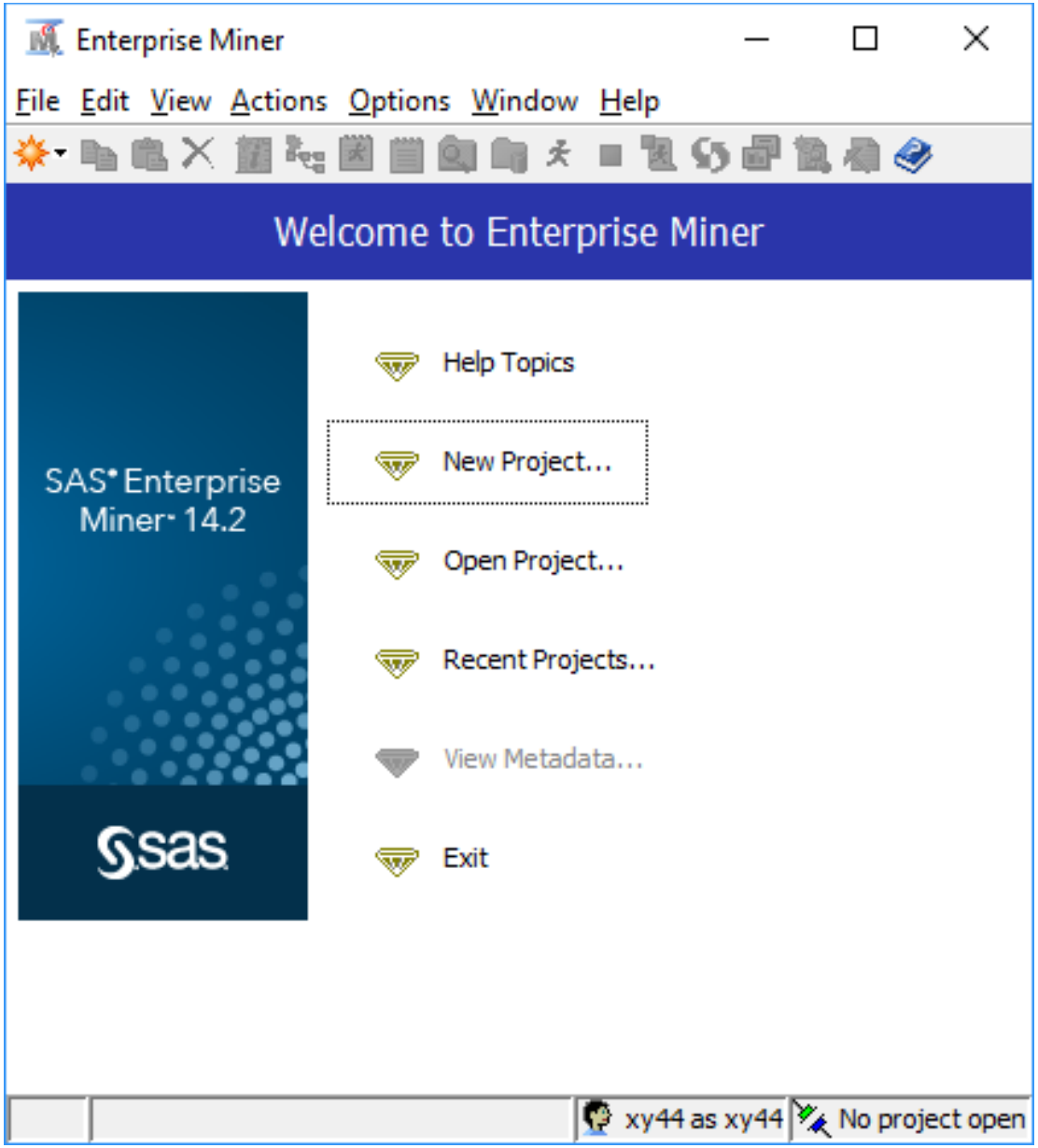
SAS

↓

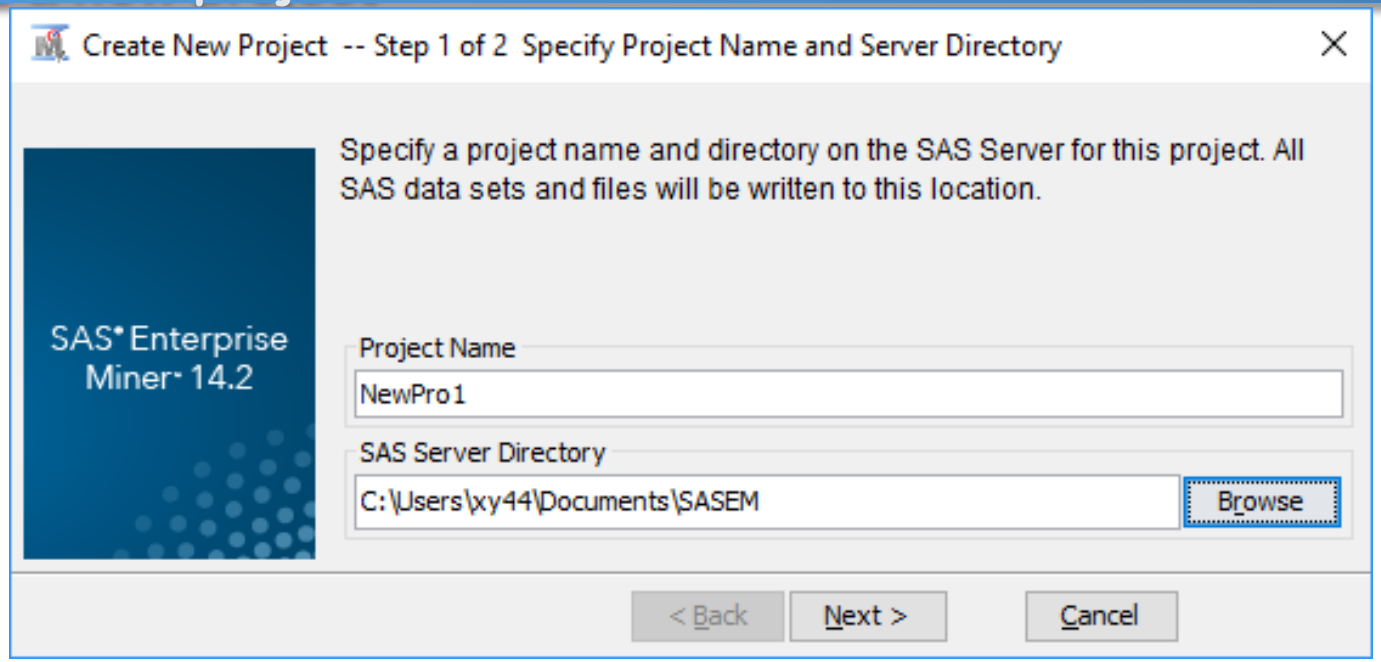
SAS Enterprise Miner Workstation 14.2 →

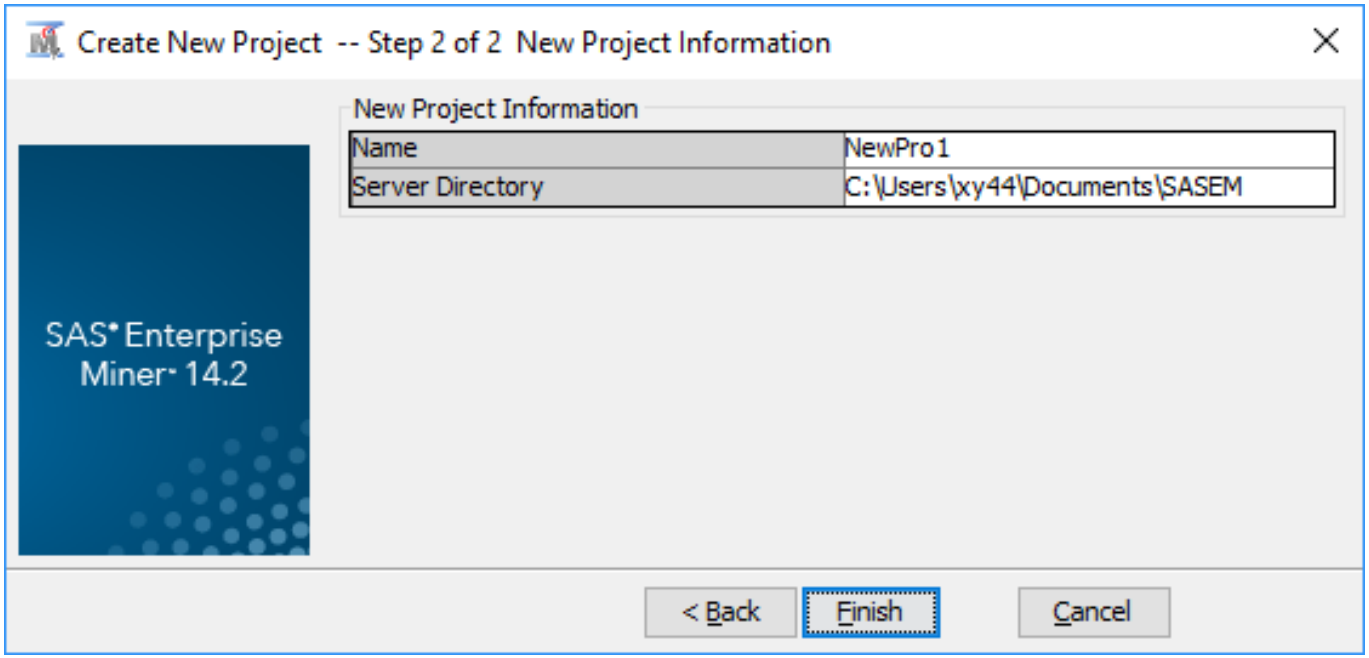


Click “New Project…”

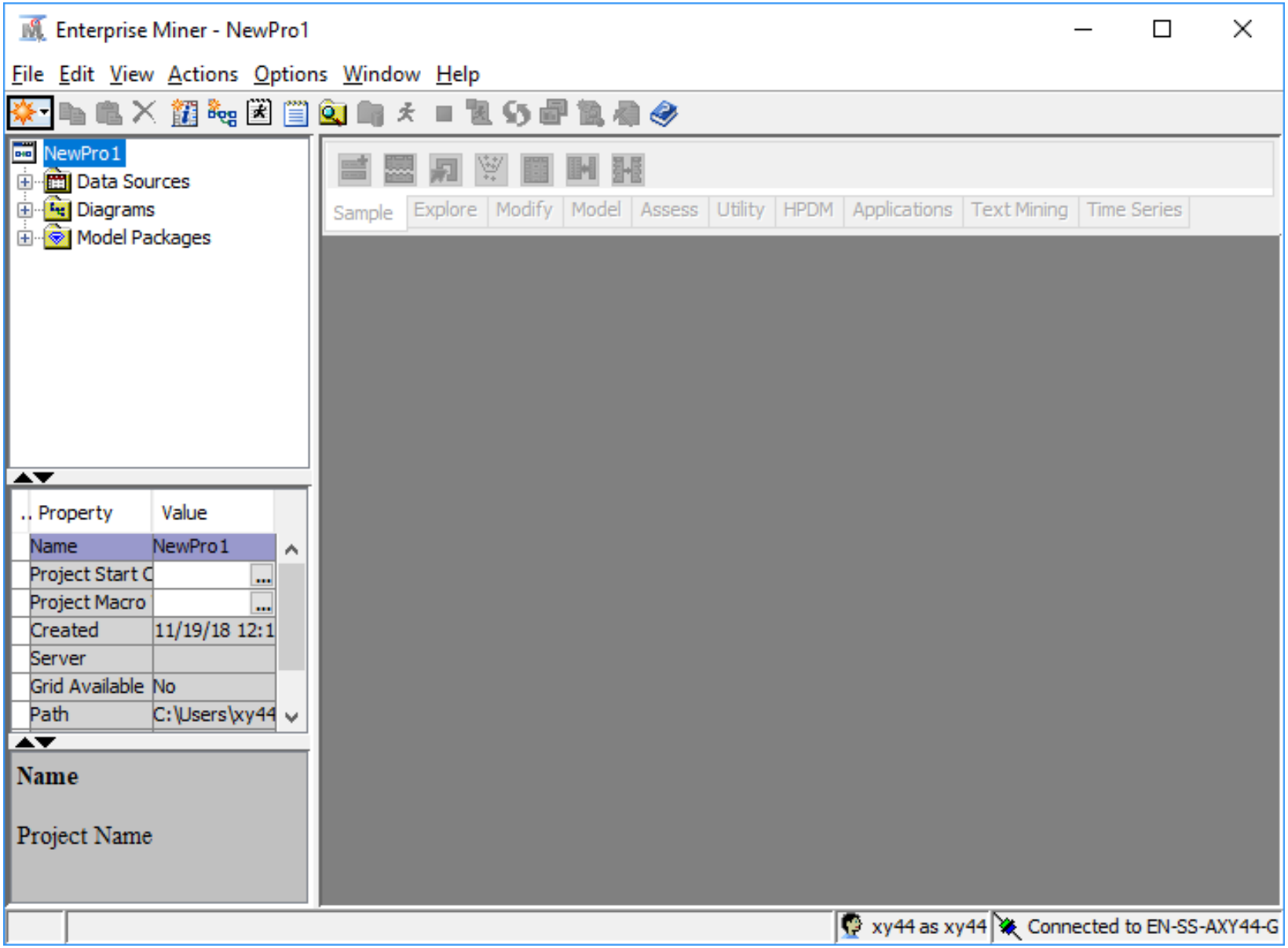


Create a new project





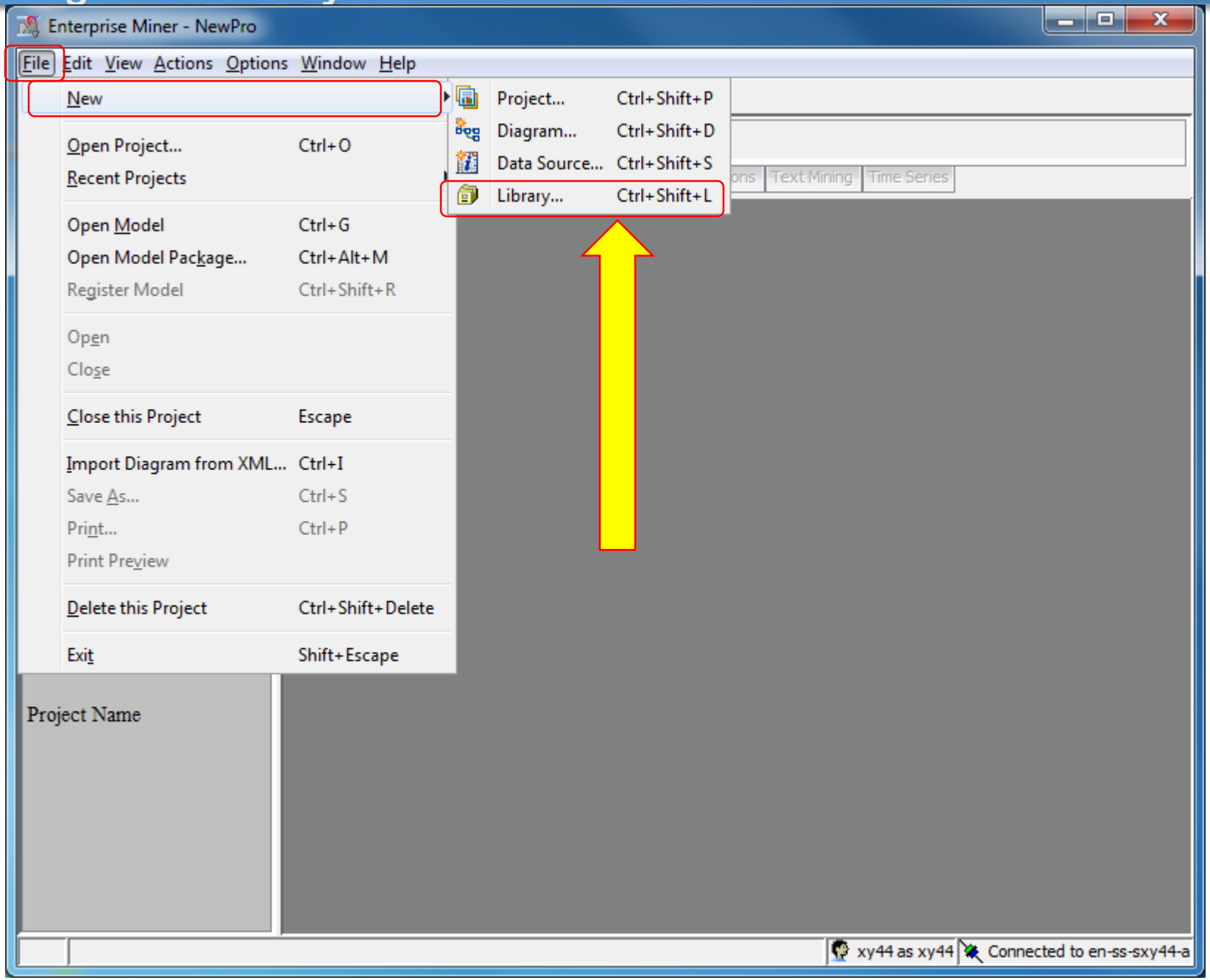
A new project window



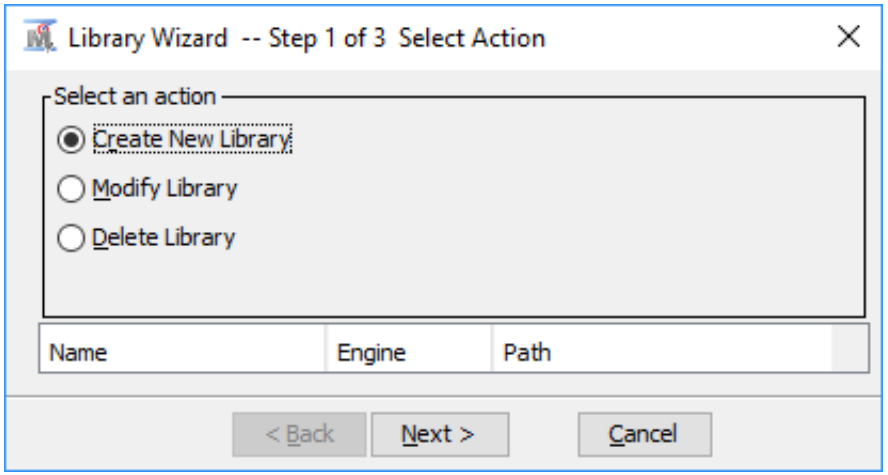
Six new folders are created

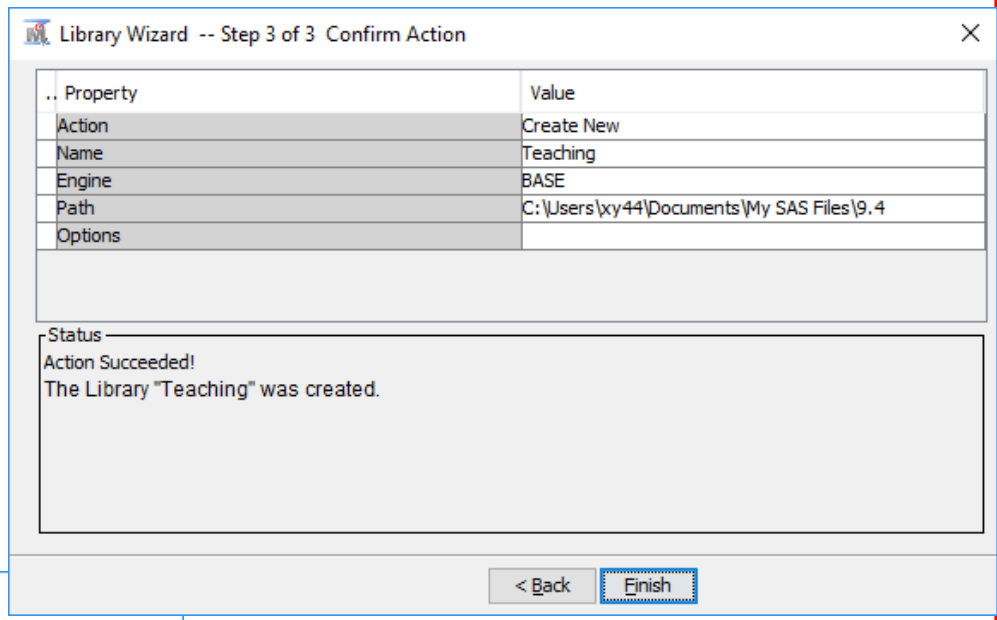
- DataSources
- HPDM (high-performance data mining)
- Meta
- Reports
- System
- Workspaces
  - EMWS1
    - Filter
    - Drop
    - Trans
    - …

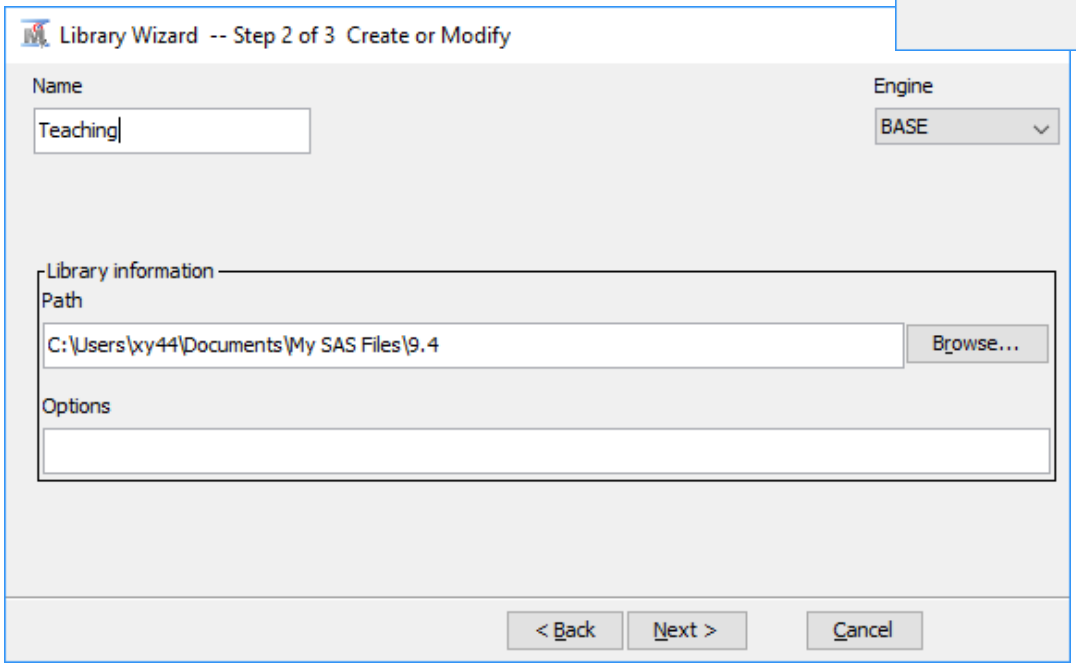
Creating a new library



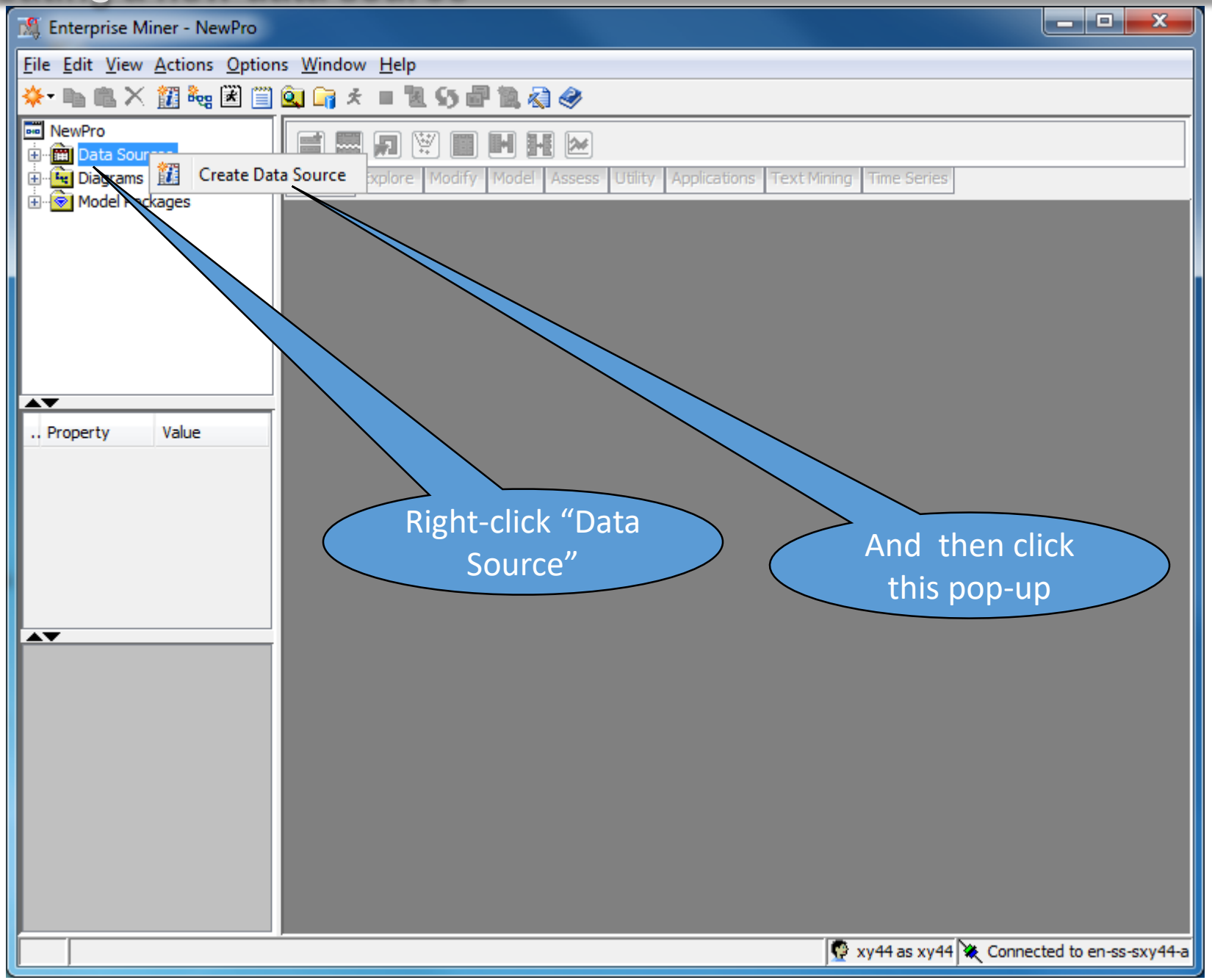
3 steps to create a new library



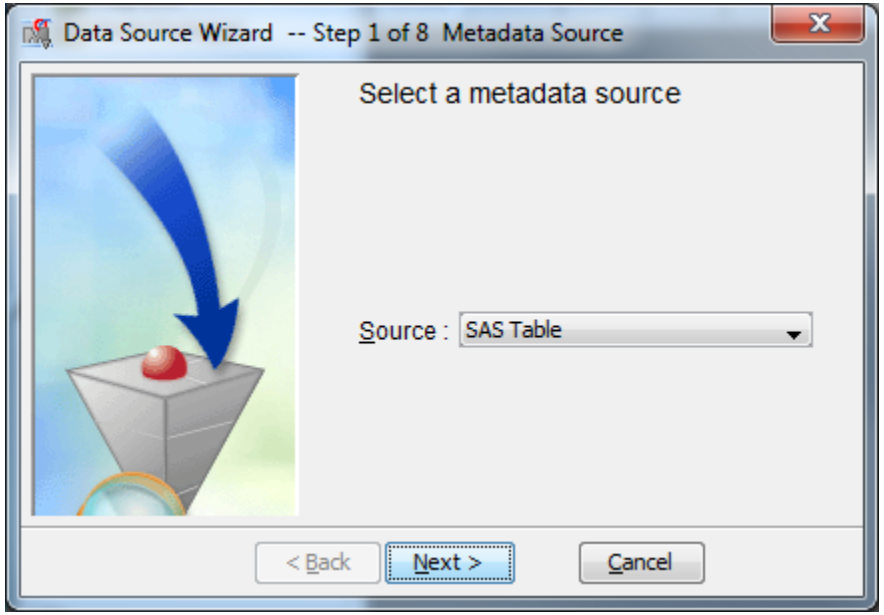


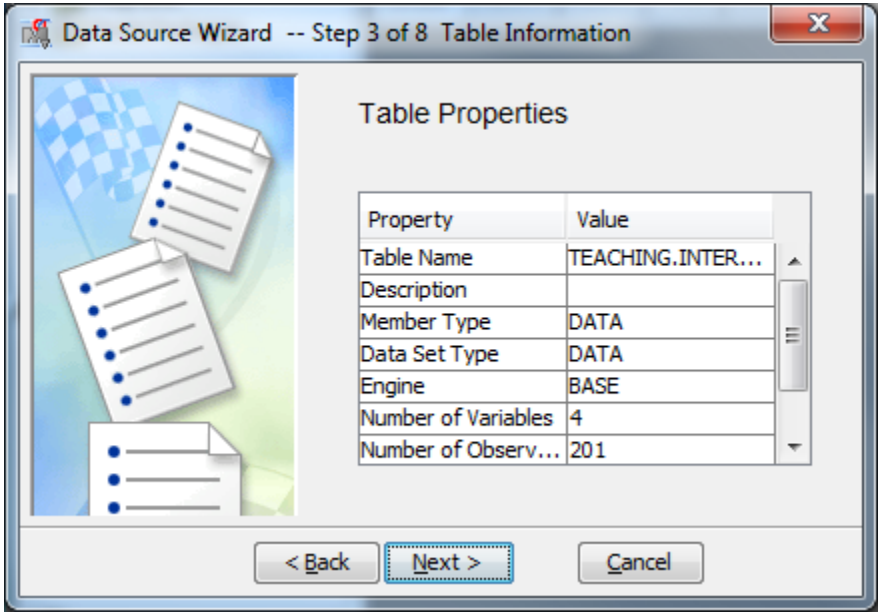


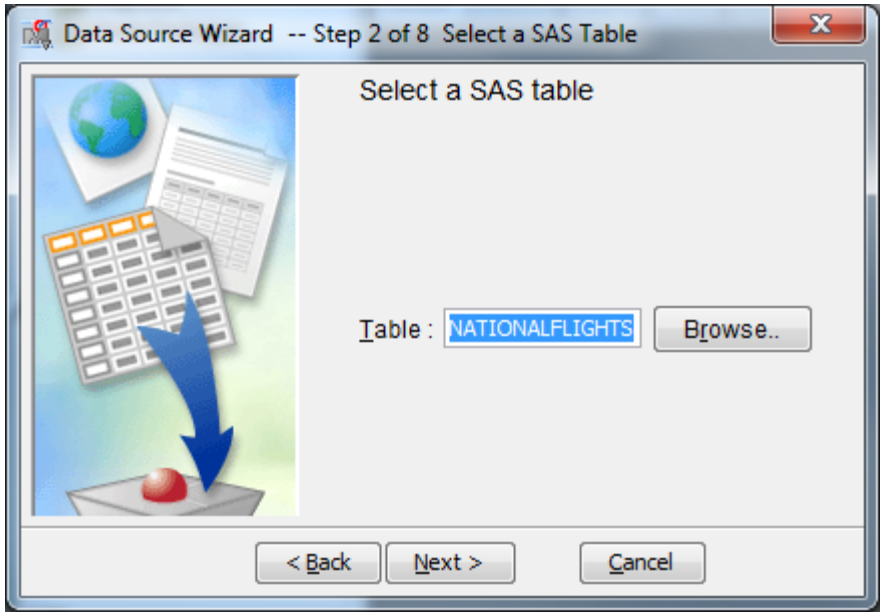
Creating a new data source



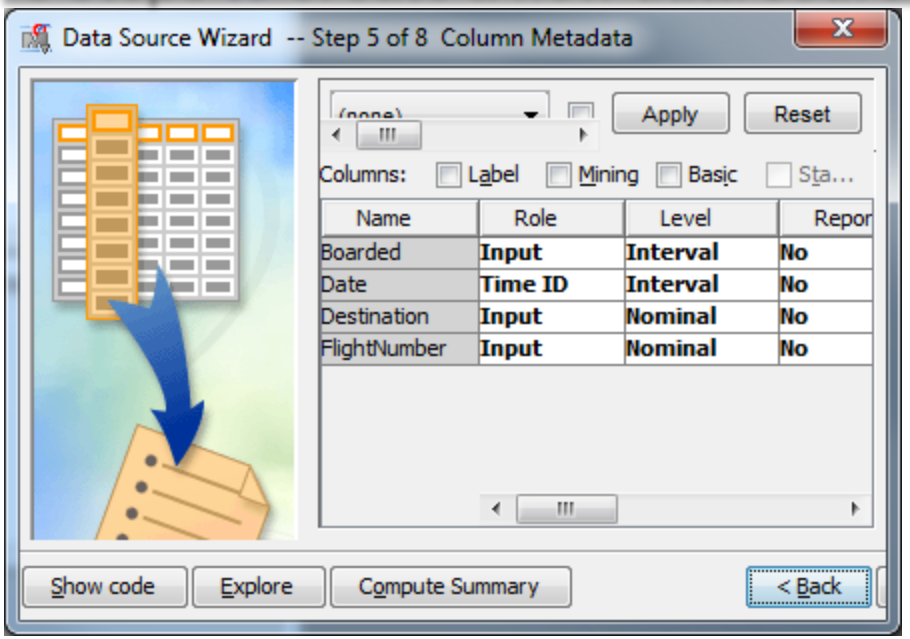
8 steps to create a new data source

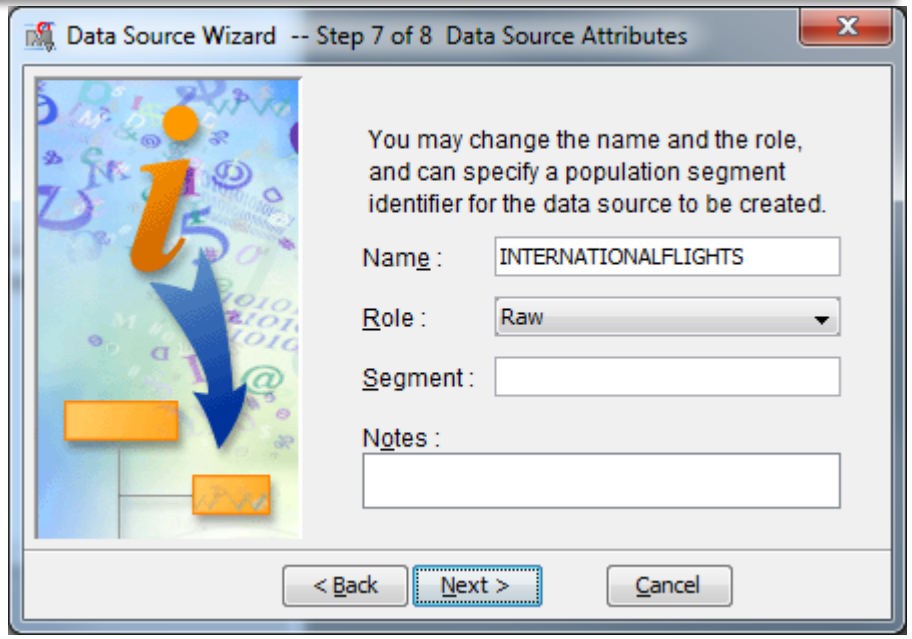


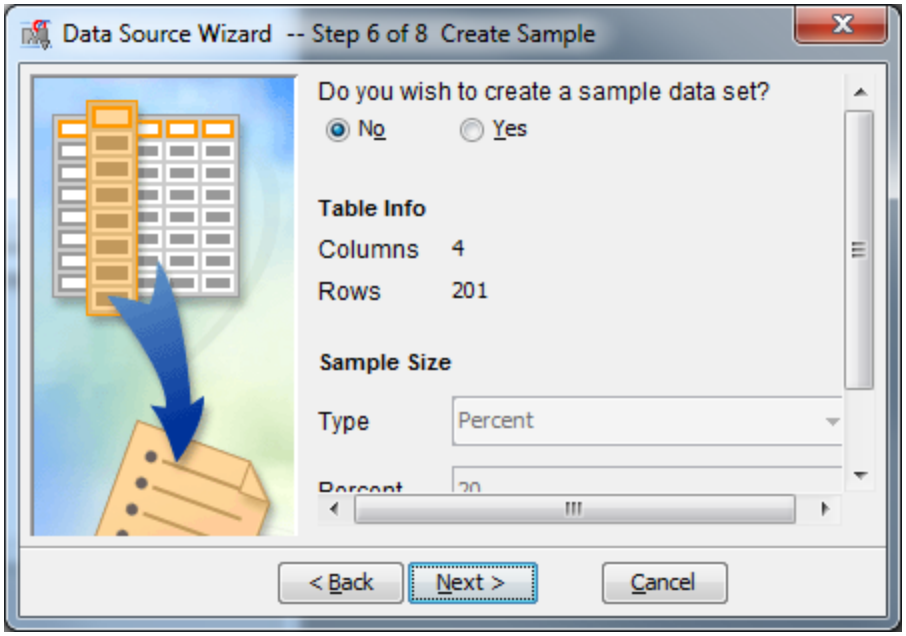


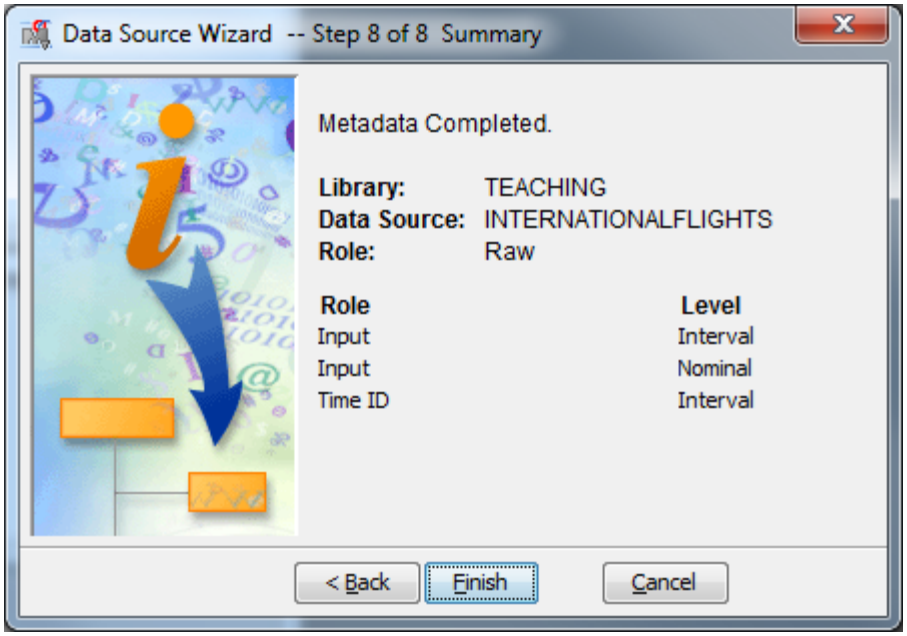


8 steps to create a new data source

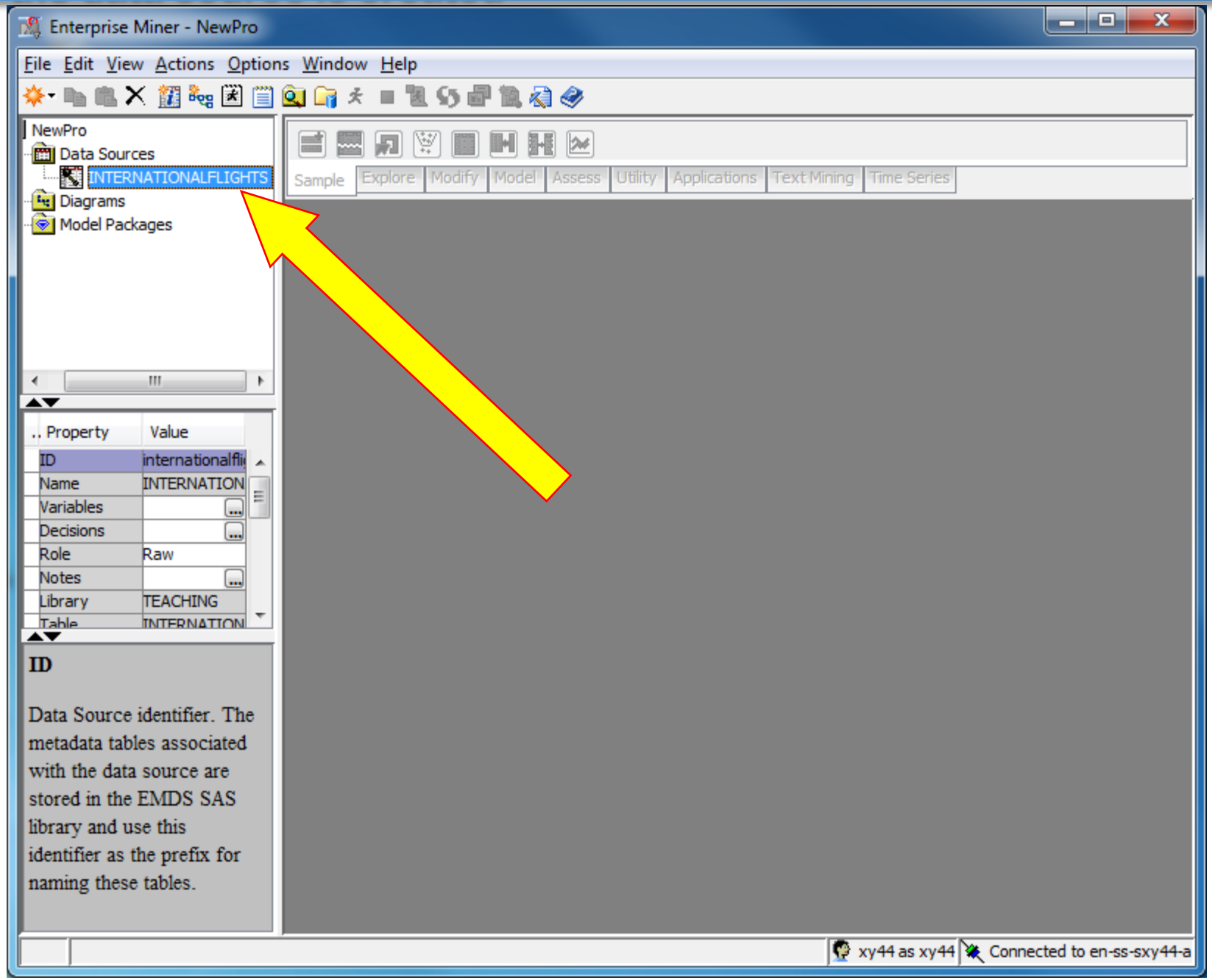




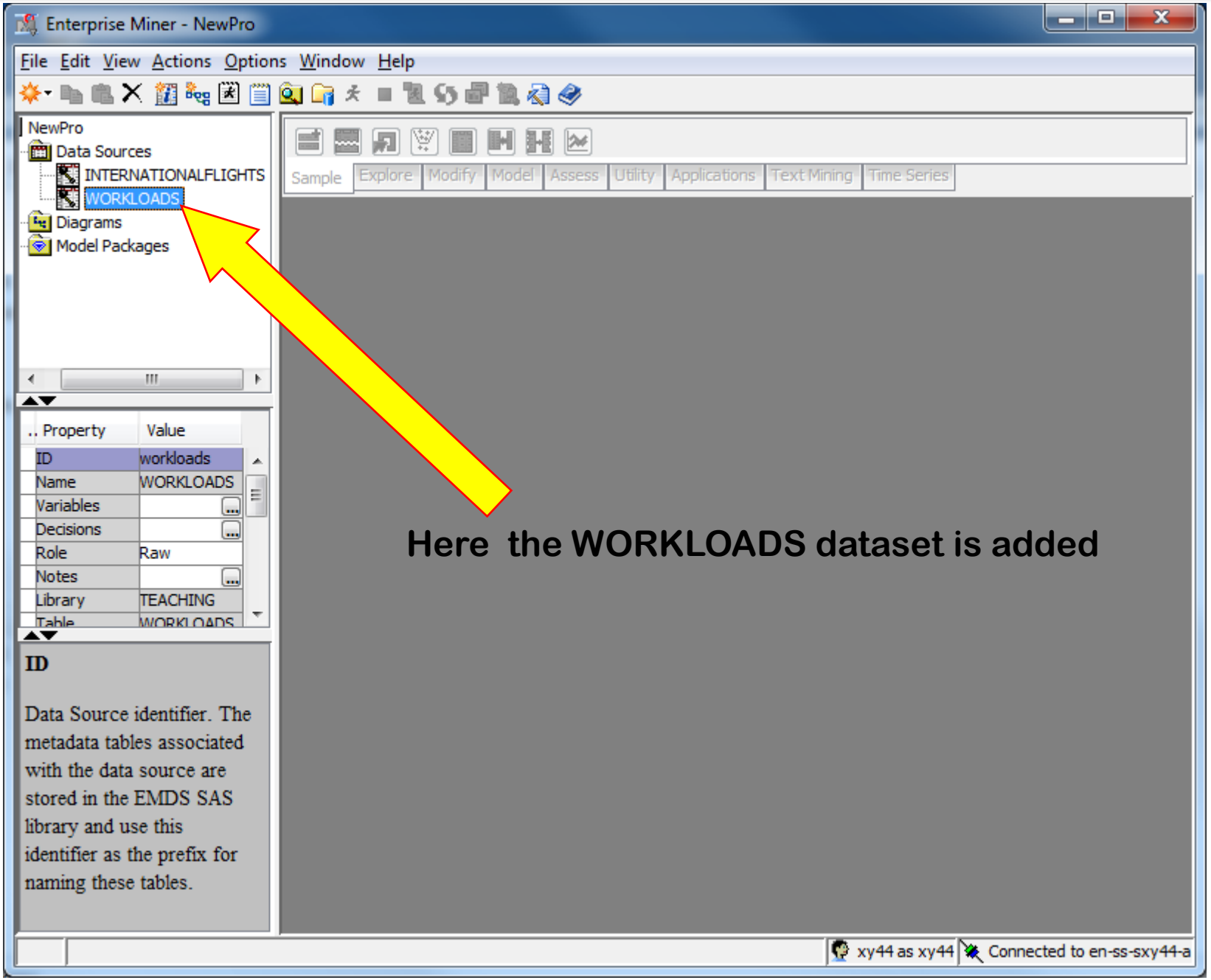




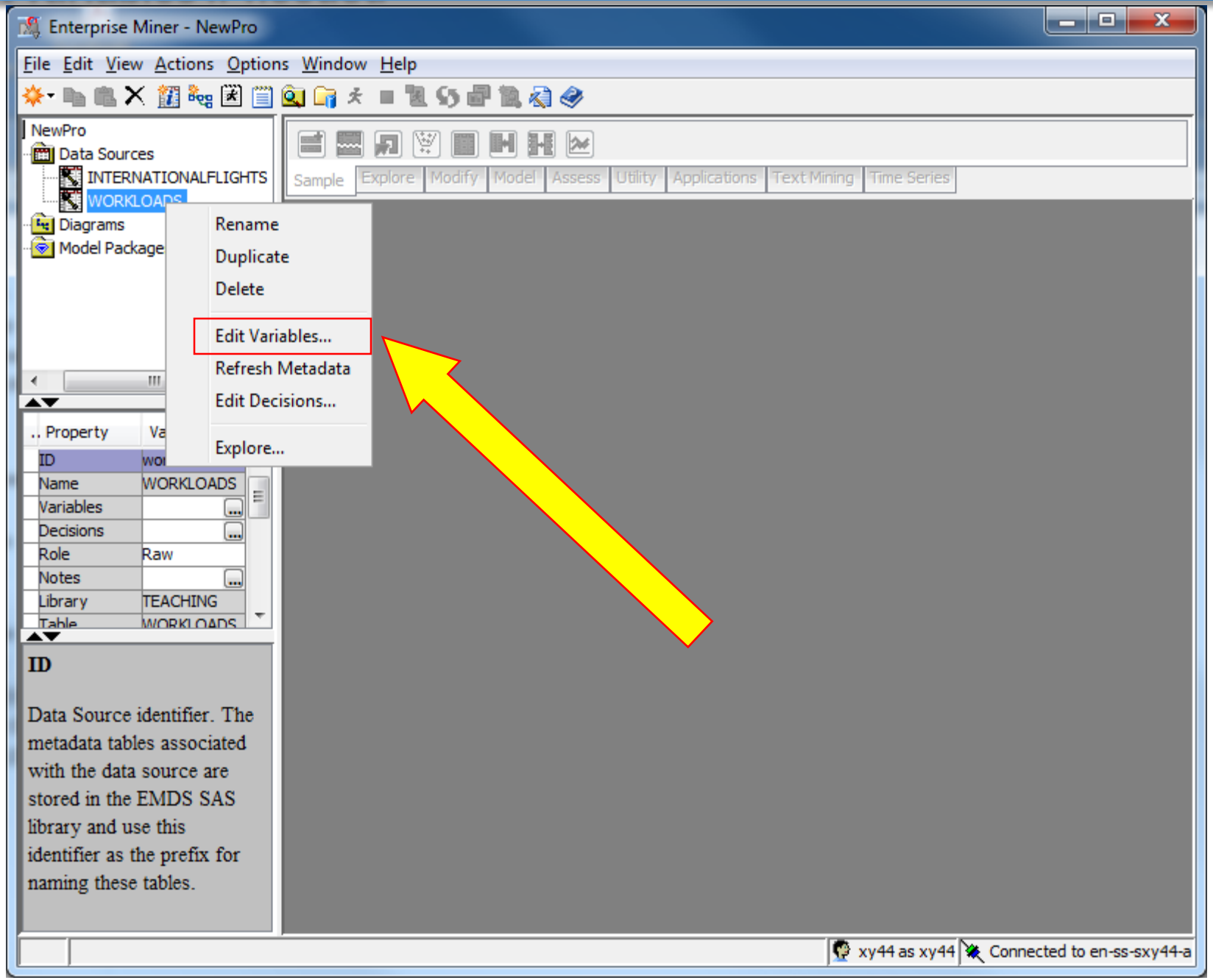
Now the data source is created



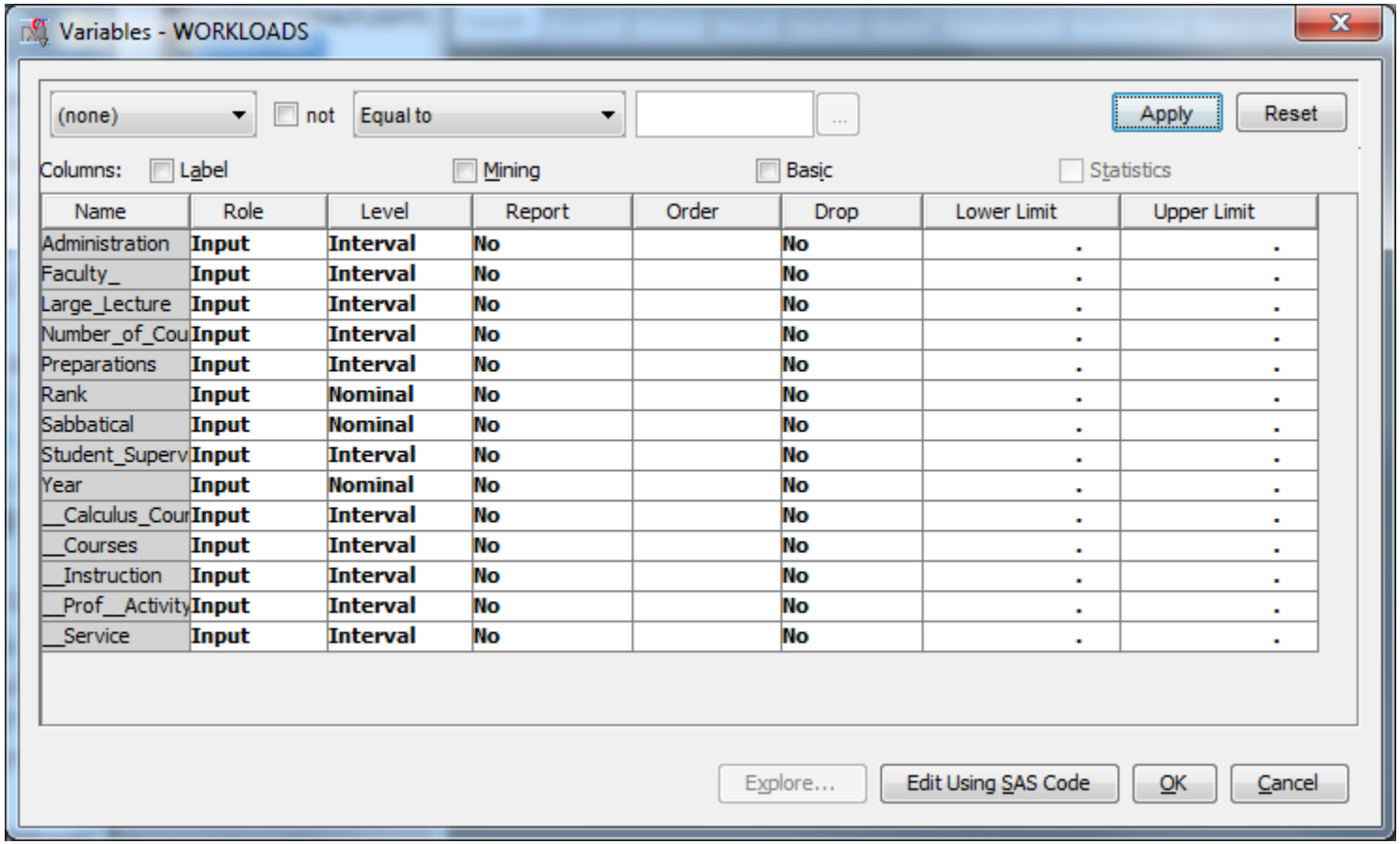
You can add more data sources as needed



Edit variables if needed

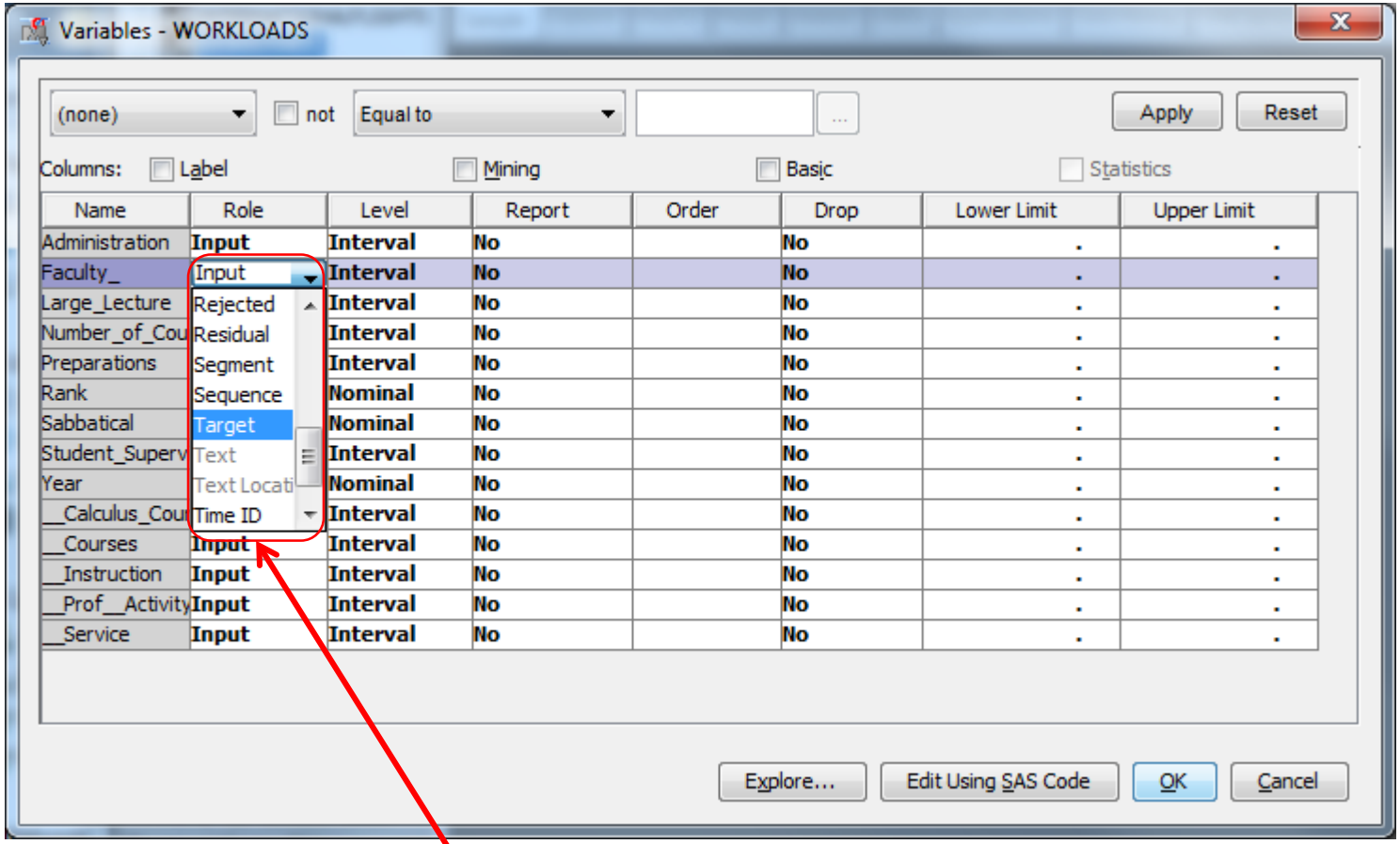


Edit variables if needed



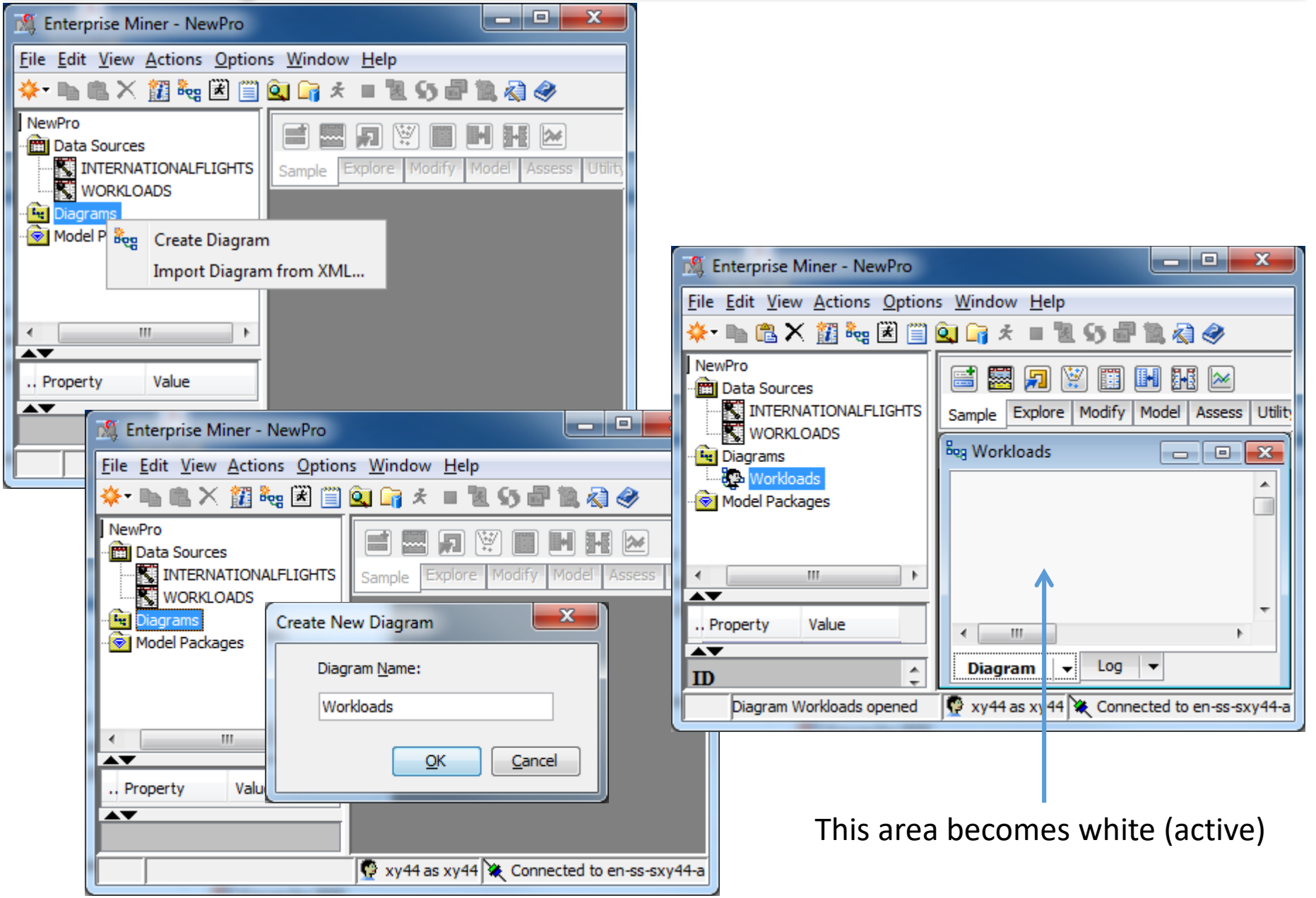
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Administration	Input	Interval	No		No	.	.
Faculty_	Input	Interval	No		No	.	.
Large_Lecture	Input	Interval	No		No	.	.
Number_of_Cou	Input	Interval	No		No	.	.
Preparations	Input	Interval	No		No	.	.
Rank	Input	Nominal	No		No	.	.
Sabbatical	Input	Nominal	No		No	.	.
Student_Superv	Input	Interval	No		No	.	.
Year	Input	Nominal	No		No	.	.
__Calculus_Cour	Input	Interval	No		No	.	.
__Courses	Input	Interval	No		No	.	.
__Instruction	Input	Interval	No		No	.	.
__Prof__Activity	Input	Interval	No		No	.	.
__Service	Input	Interval	No		No	.	.

Edit variables if needed

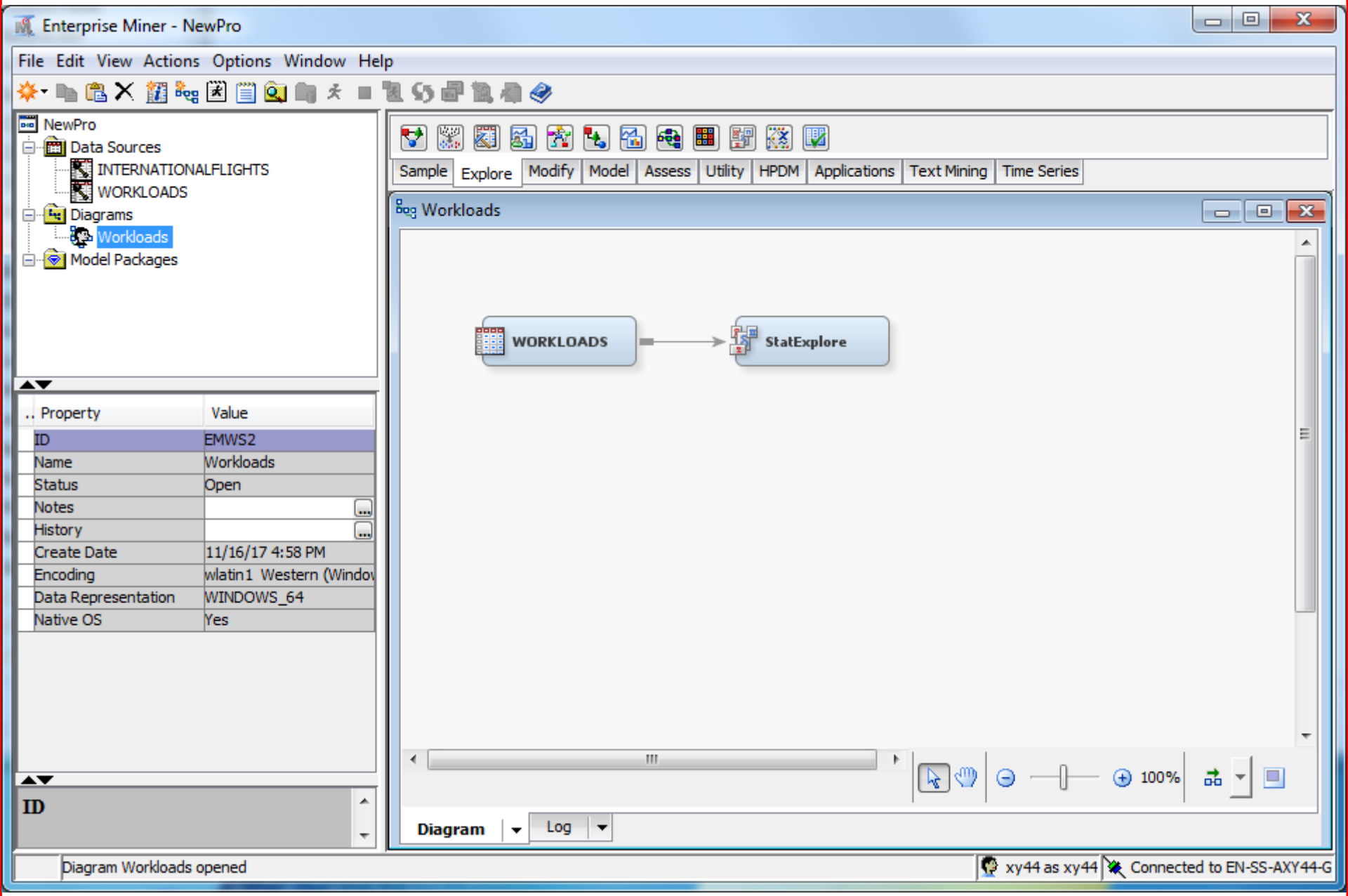


Click any cell to make a change

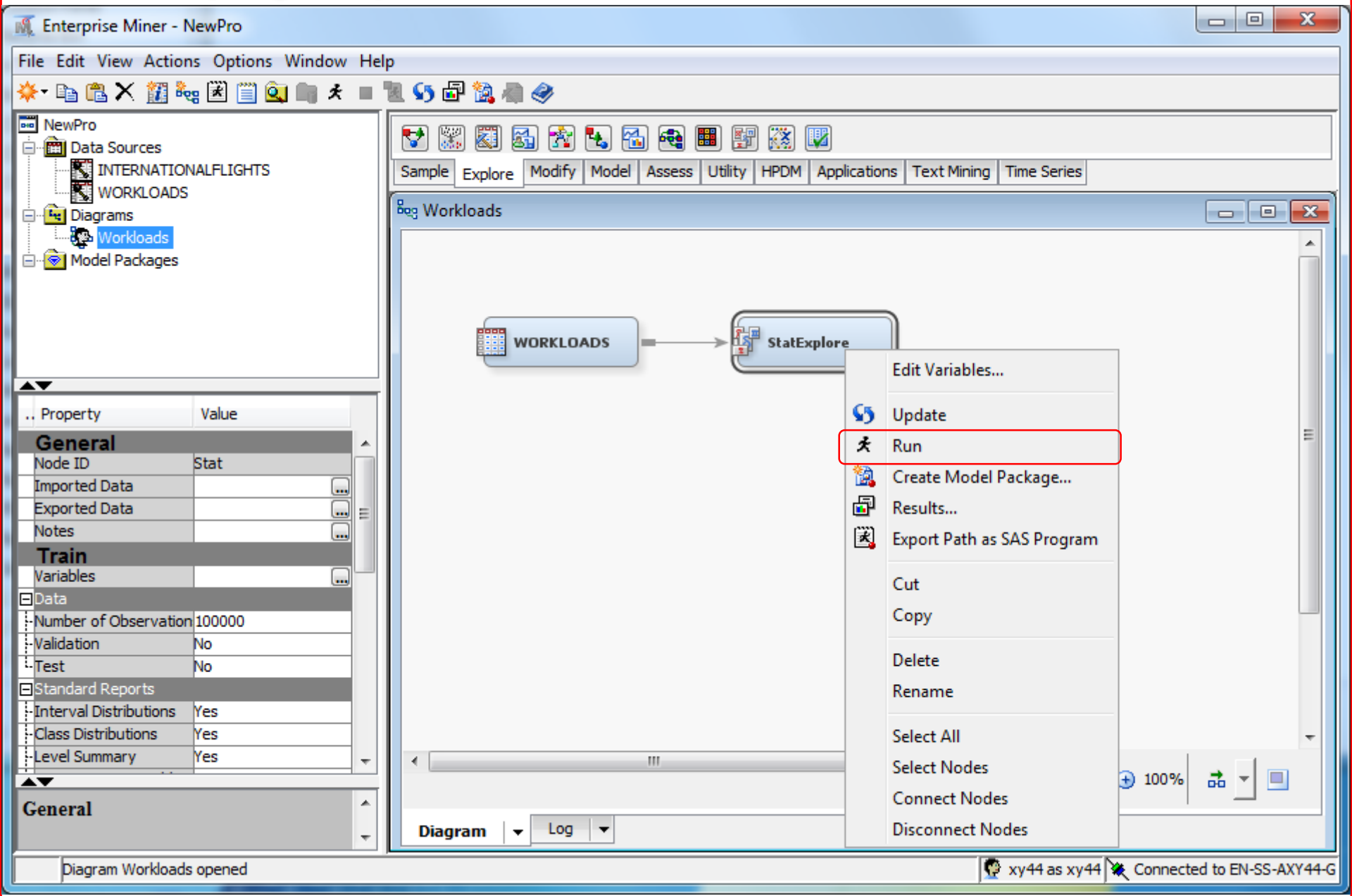
Create a diagram



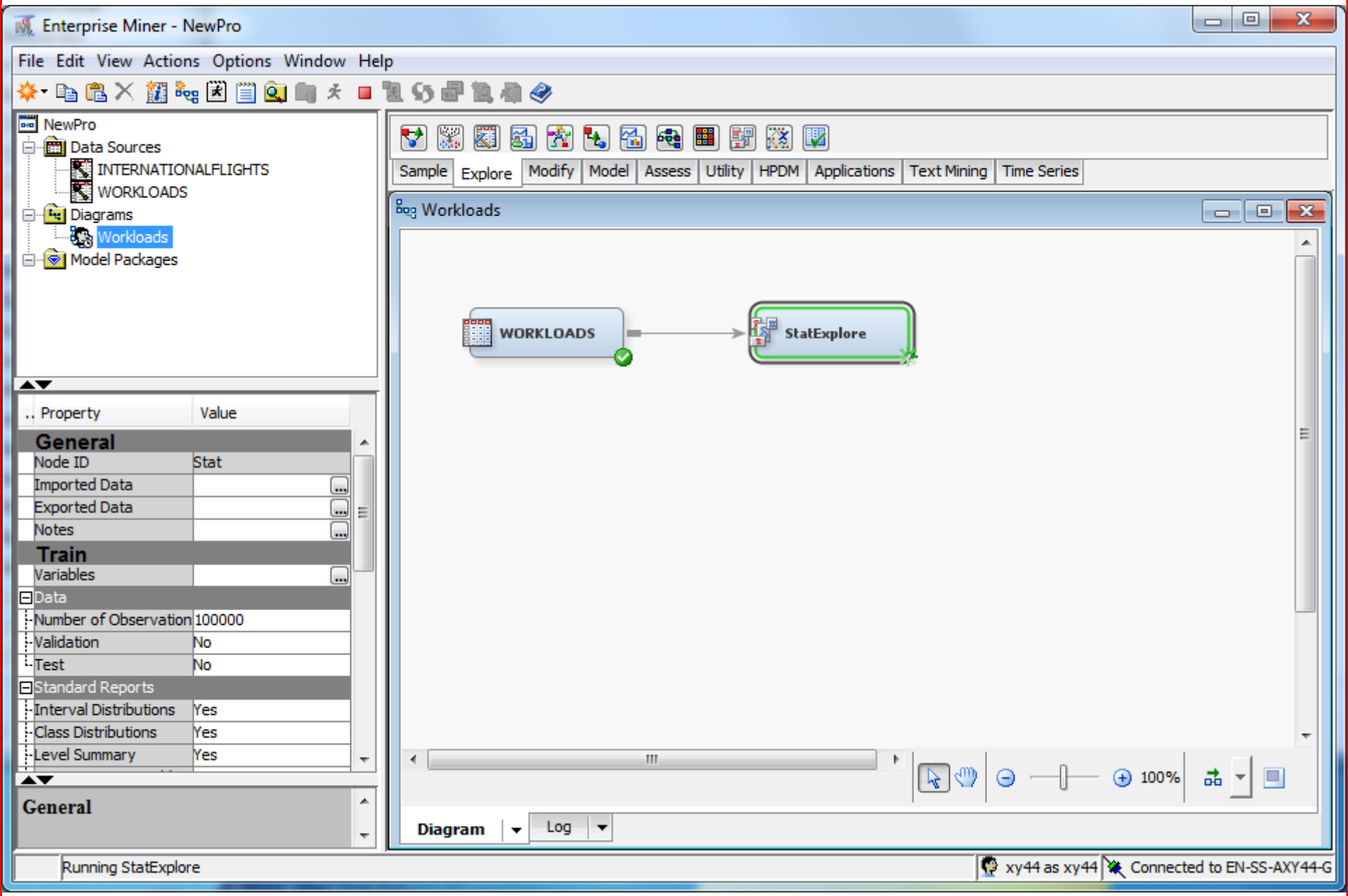
Create a diagram: Input data node and StatExplore



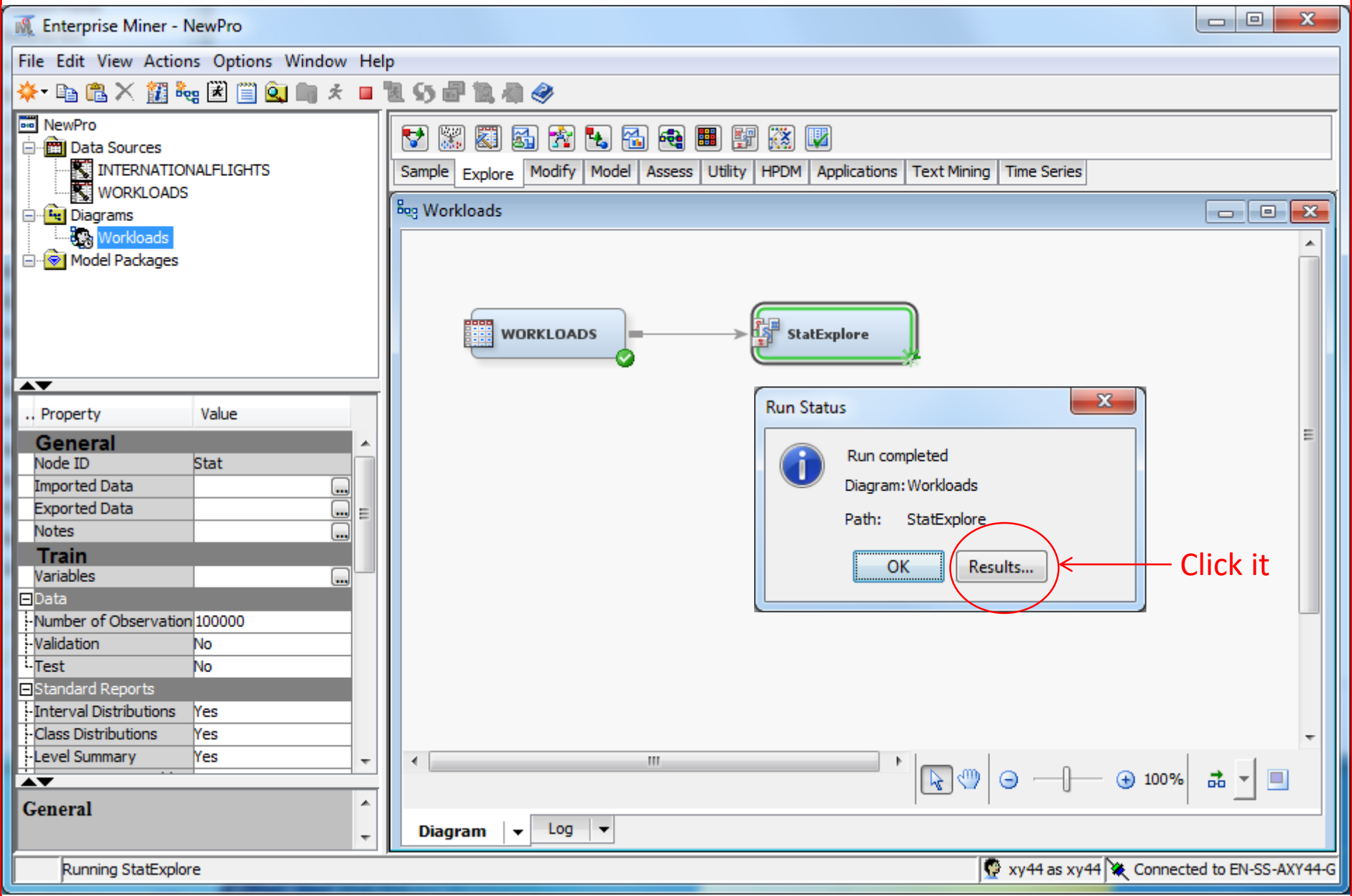
Run the program/procedure



Run the program/procedure

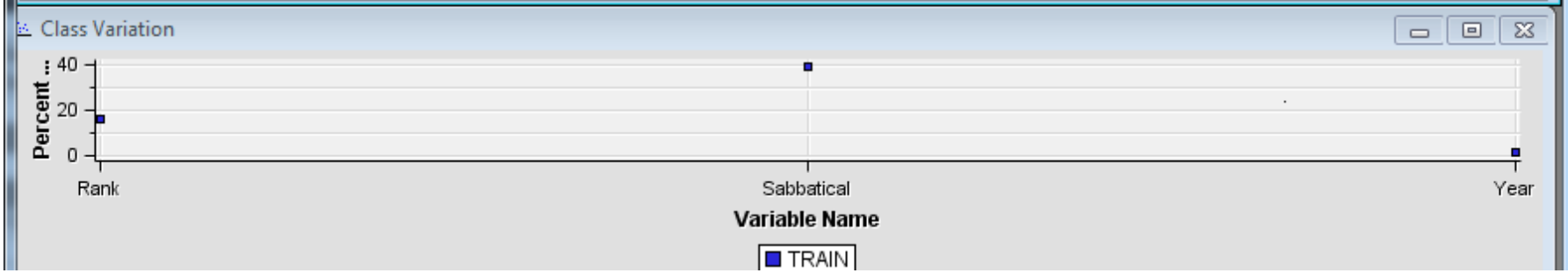


Run the program/procedure → get the Results…

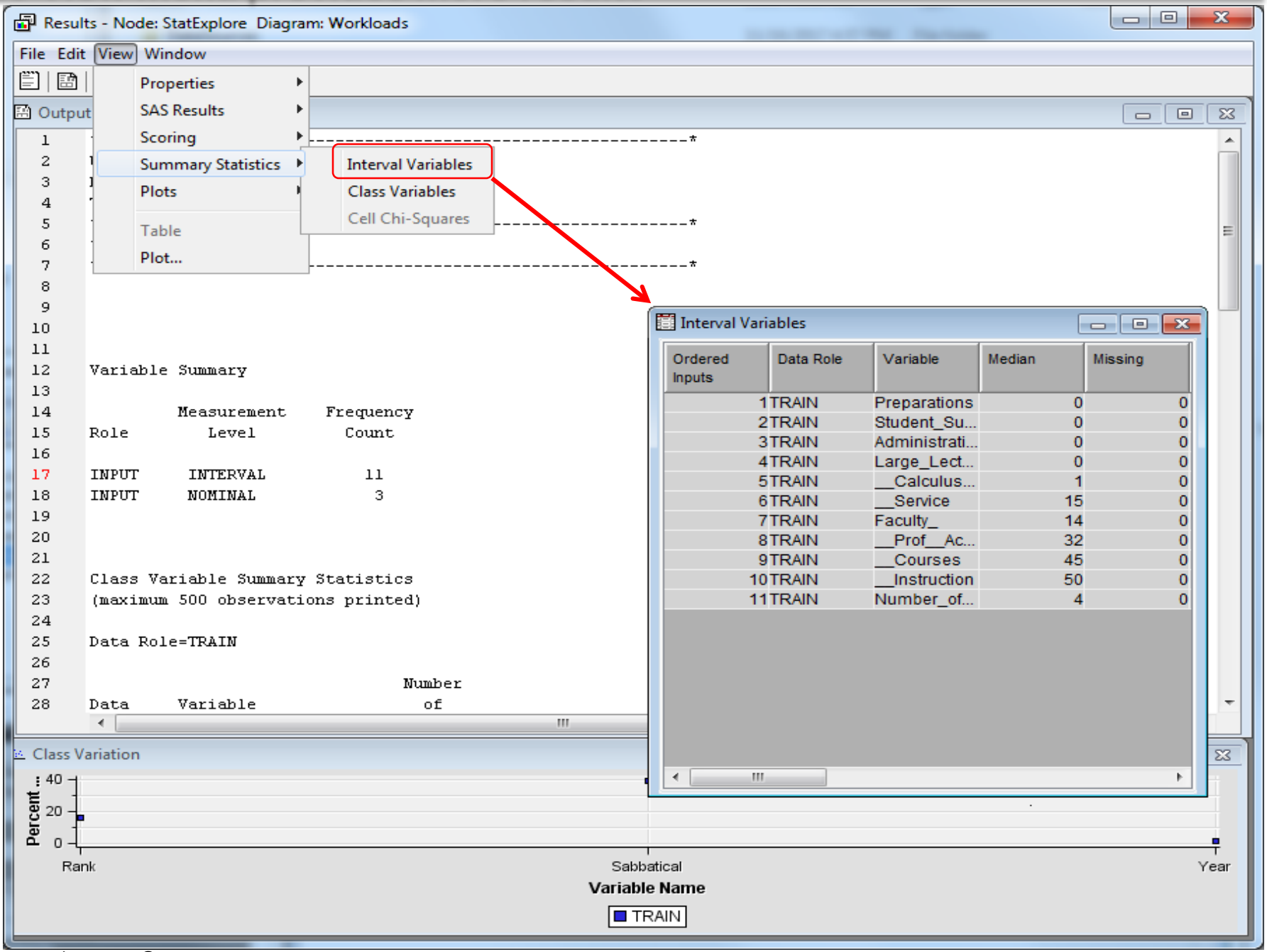


The result: Output window

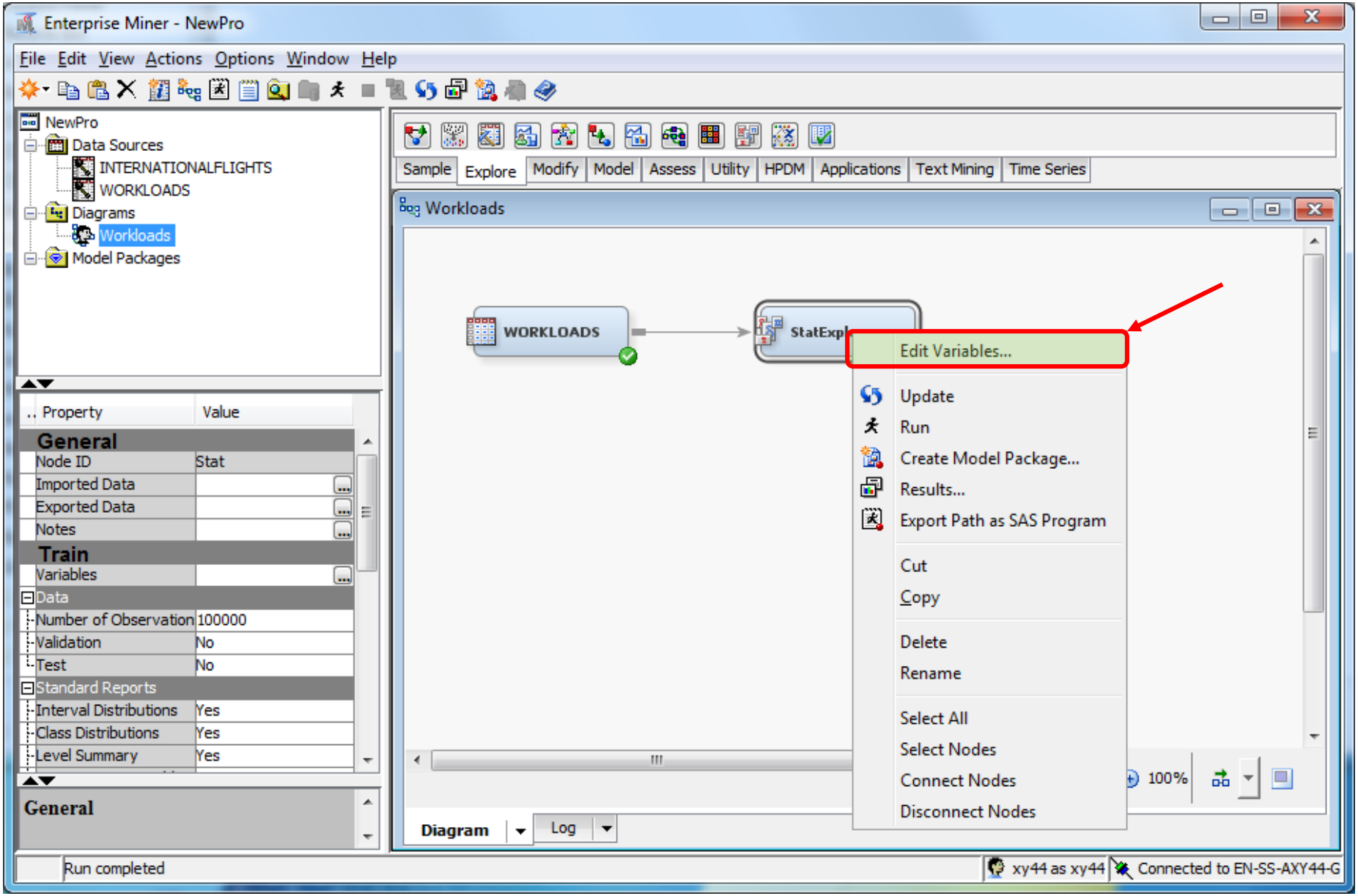
```
 5   *-------------------------------------------------------------*
 6   * Training Output
 7   *-------------------------------------------------------------*
 8
 9
10
11
12   Variable Summary
13
14            Measurement     Frequency
15   Role        Level          Count
16
17   INPUT      INTERVAL           11
18   INPUT      NOMINAL             3
19
20
21
22   Class Variable Summary Statistics
23   (maximum 500 observations printed)
24
25   Data Role=TRAIN
26
27                                 Number
28   Data      Variable                 of                    Mode                  Mode2
```



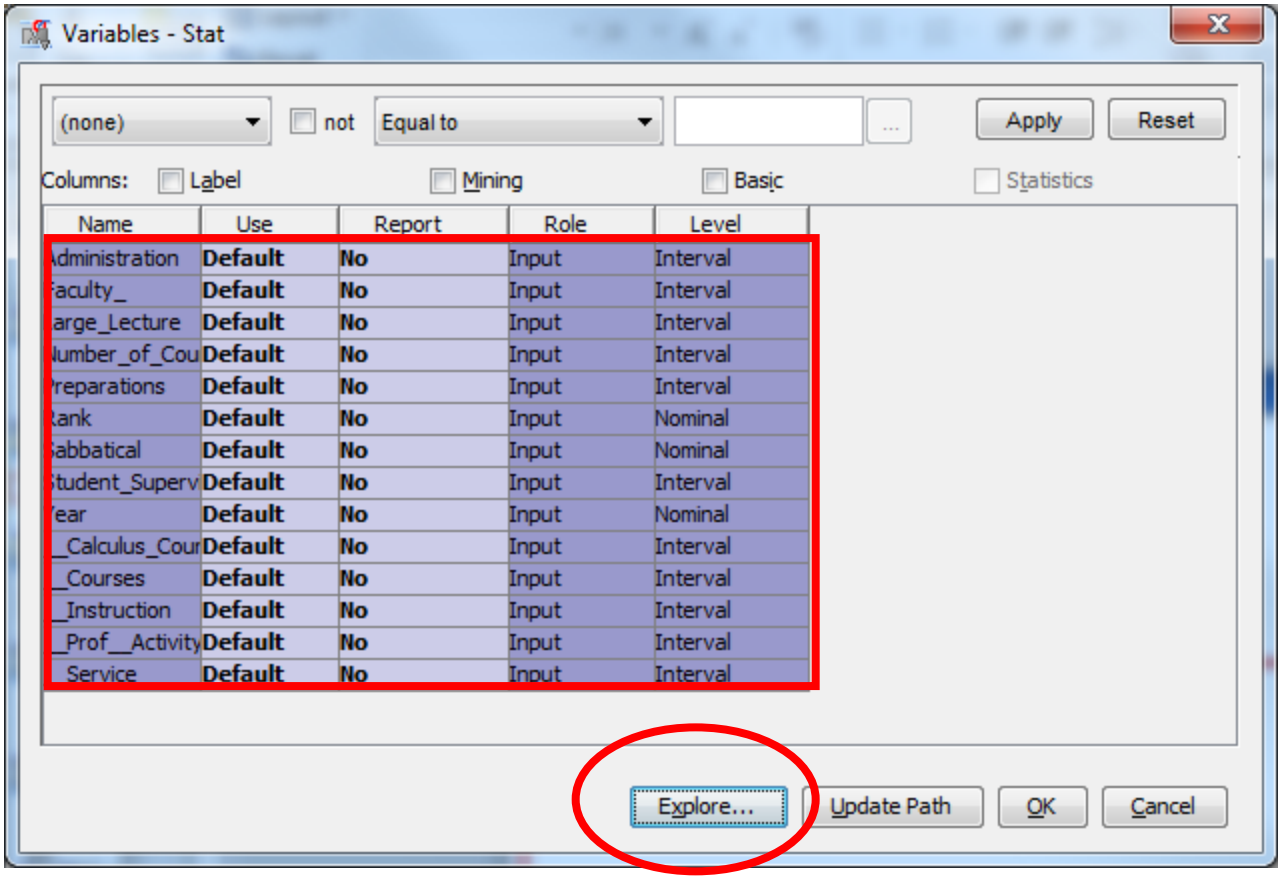
The View menu provides more results



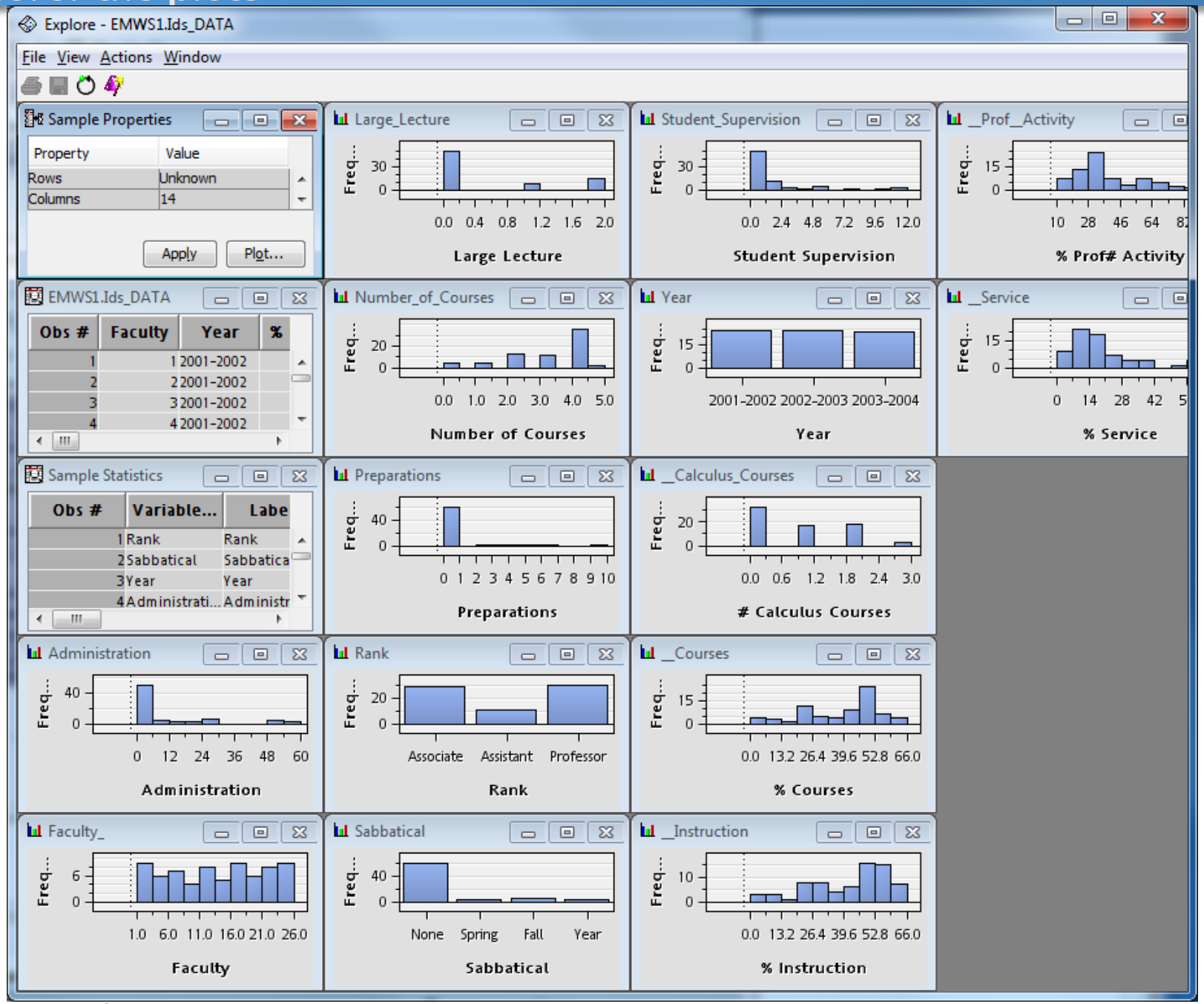
Edit/explore variables from the node



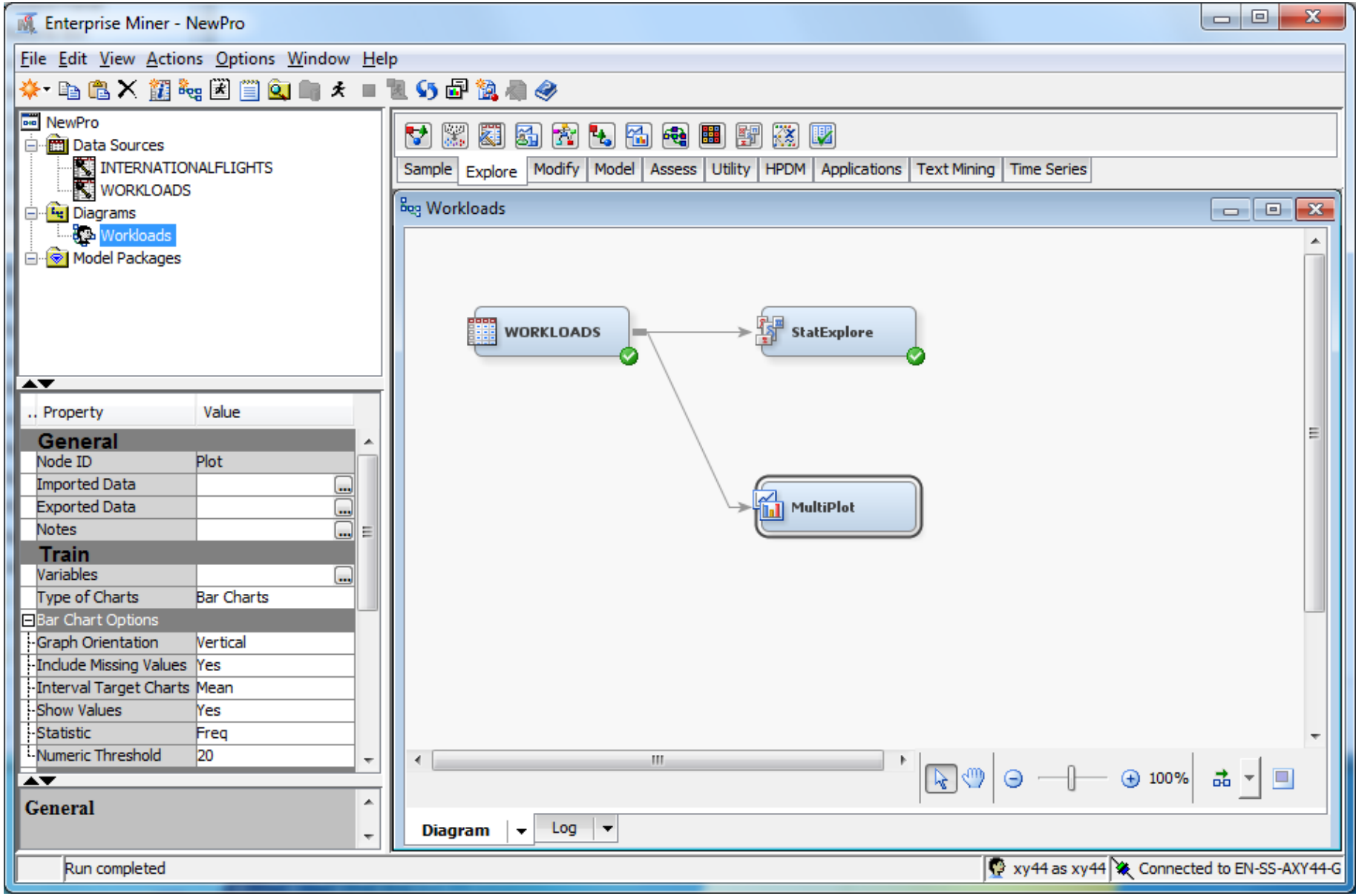
Highlight the variable(s) to explore



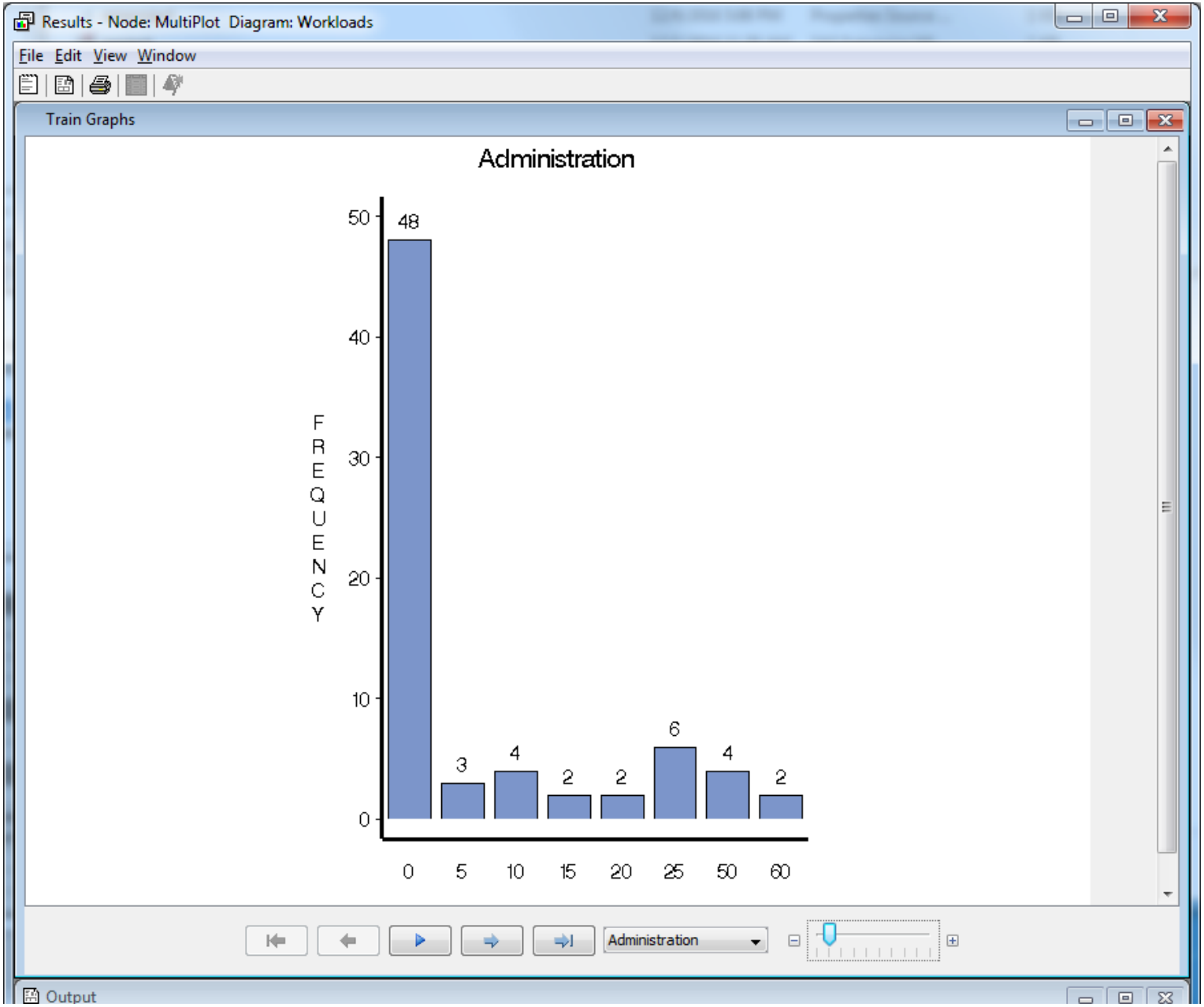
Explorer the plots



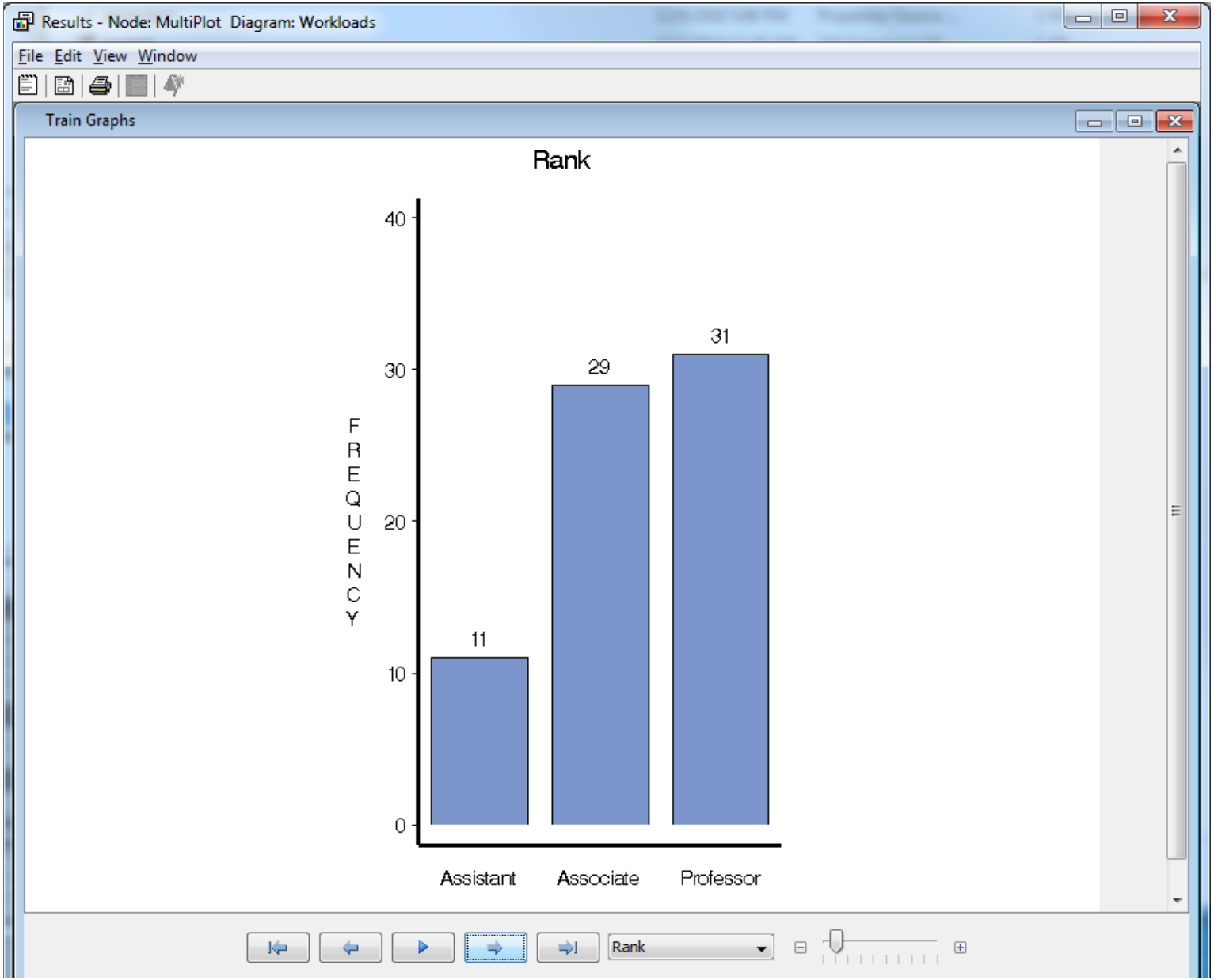
Add a multiplot node to examine the variables



An Output example: A bar chart for Administration



Example: A bar chart for Rank



Data partition

Training (40%) is used for preliminary model fitting. The analyst attempts to find the best model weights using this data set.

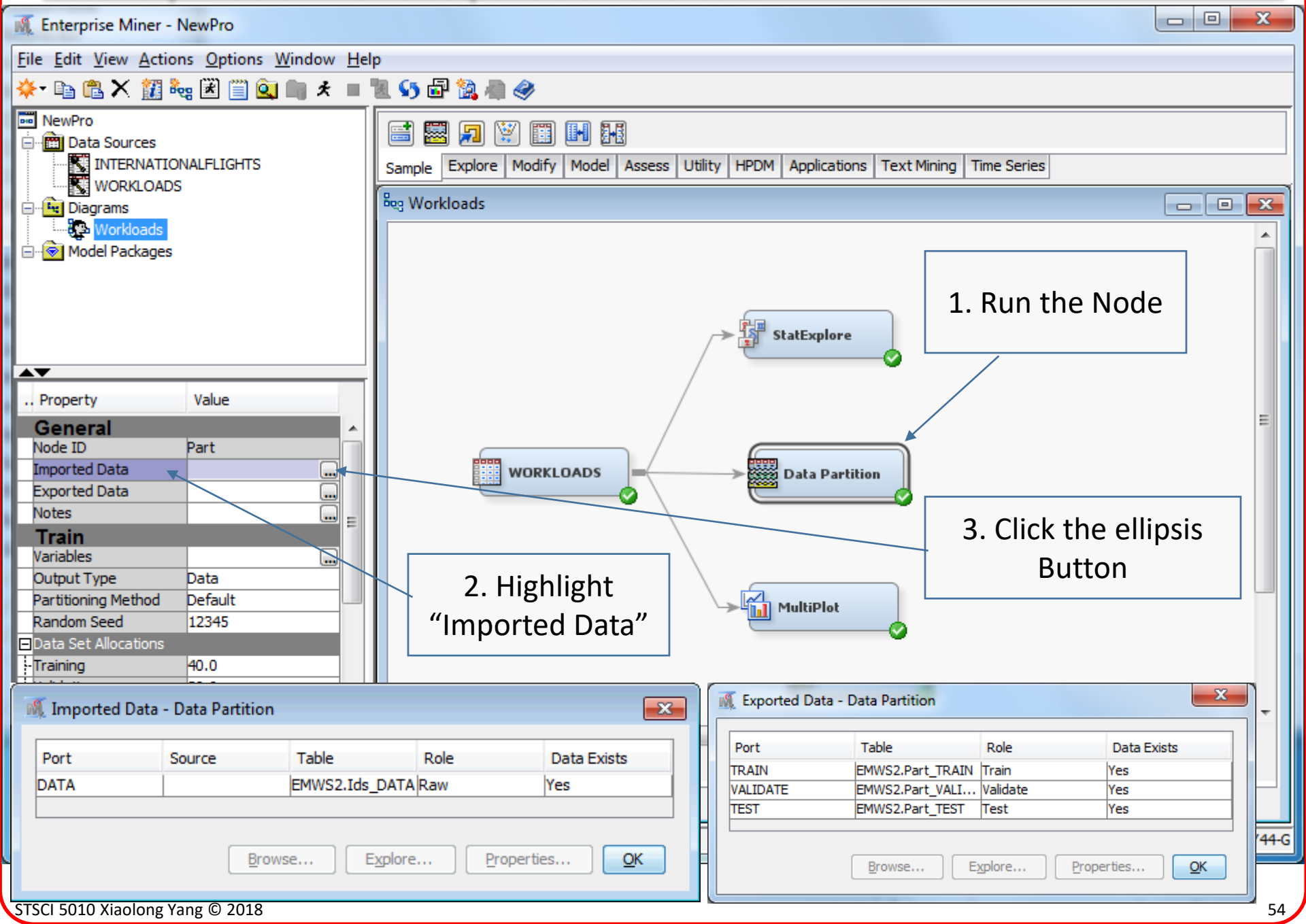
Validation (30%) is used to assess the adequacy of the model in the Model Comparison node.

The validation data set is also used for model fine-tuning in the following nodes:

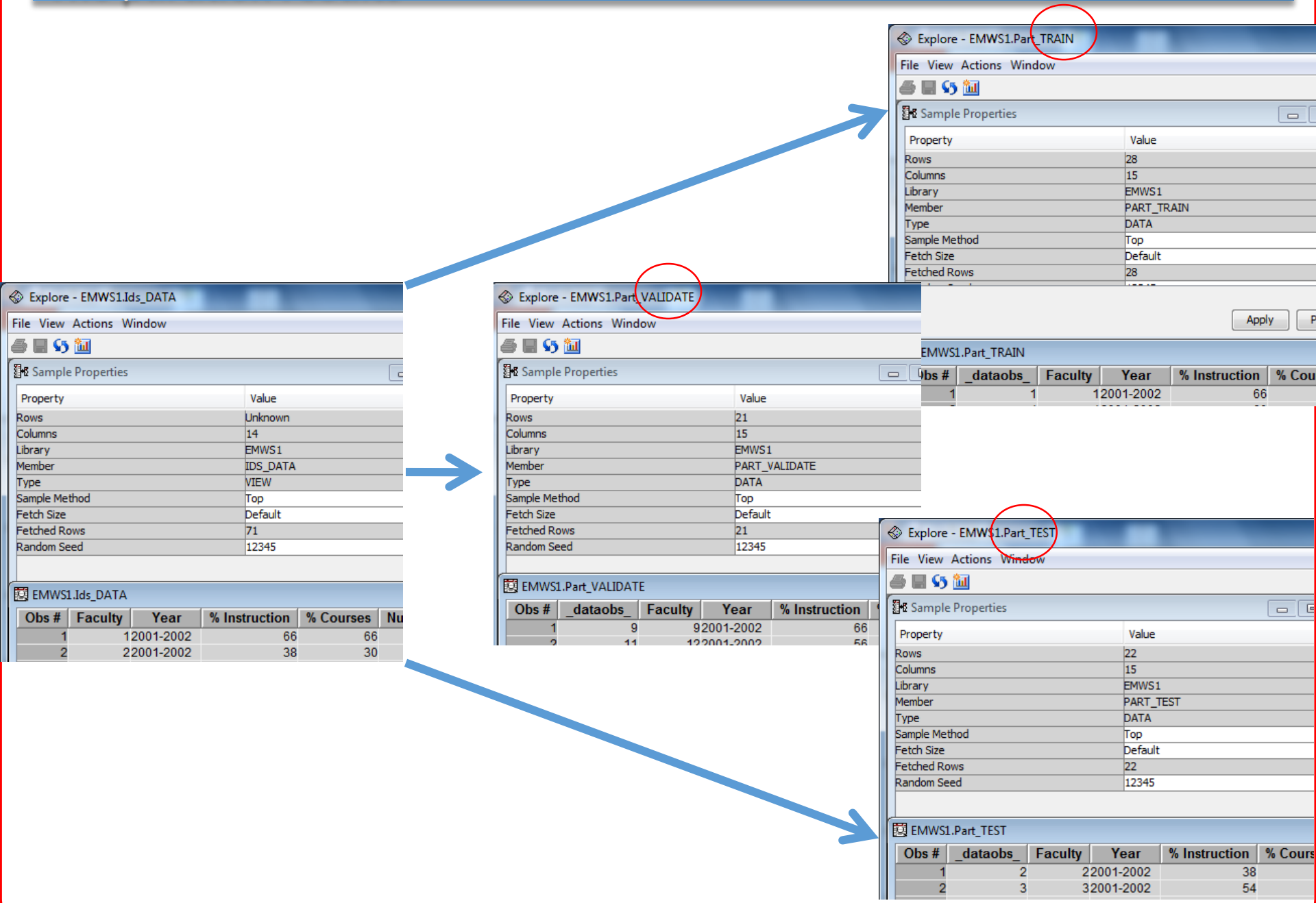
- **Decision Tree node** — to create the best subtree.
- **Neural Network node** — to choose among network architectures or for the early-stopping of the training algorithm.
- **Regression node** — to choose a final subset of predictors from all the subsets computed during stepwise regression.

Test (30%) is used to obtain a final, unbiased estimate of the generalization error of the model.

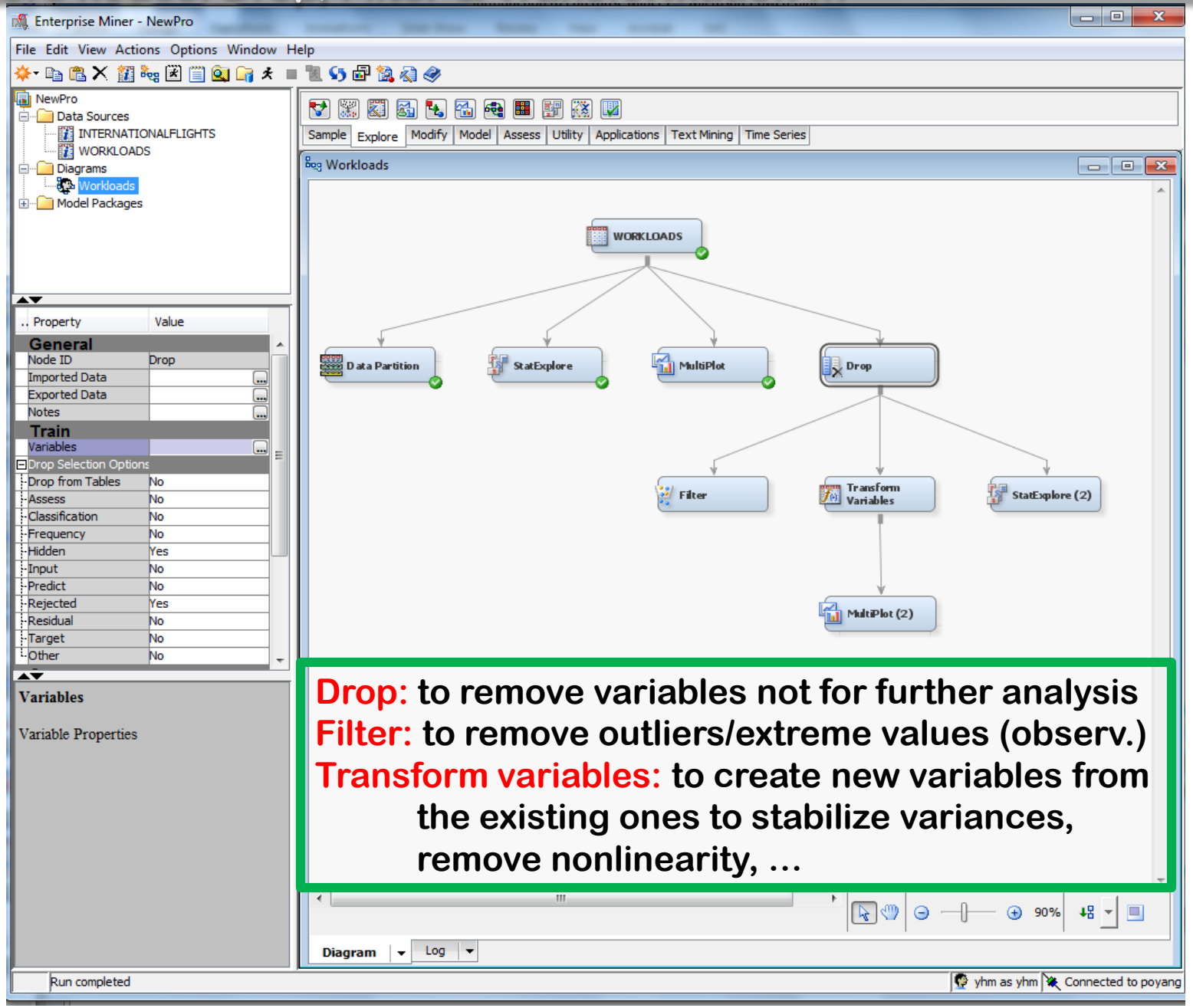
Add a partition node to partition the data



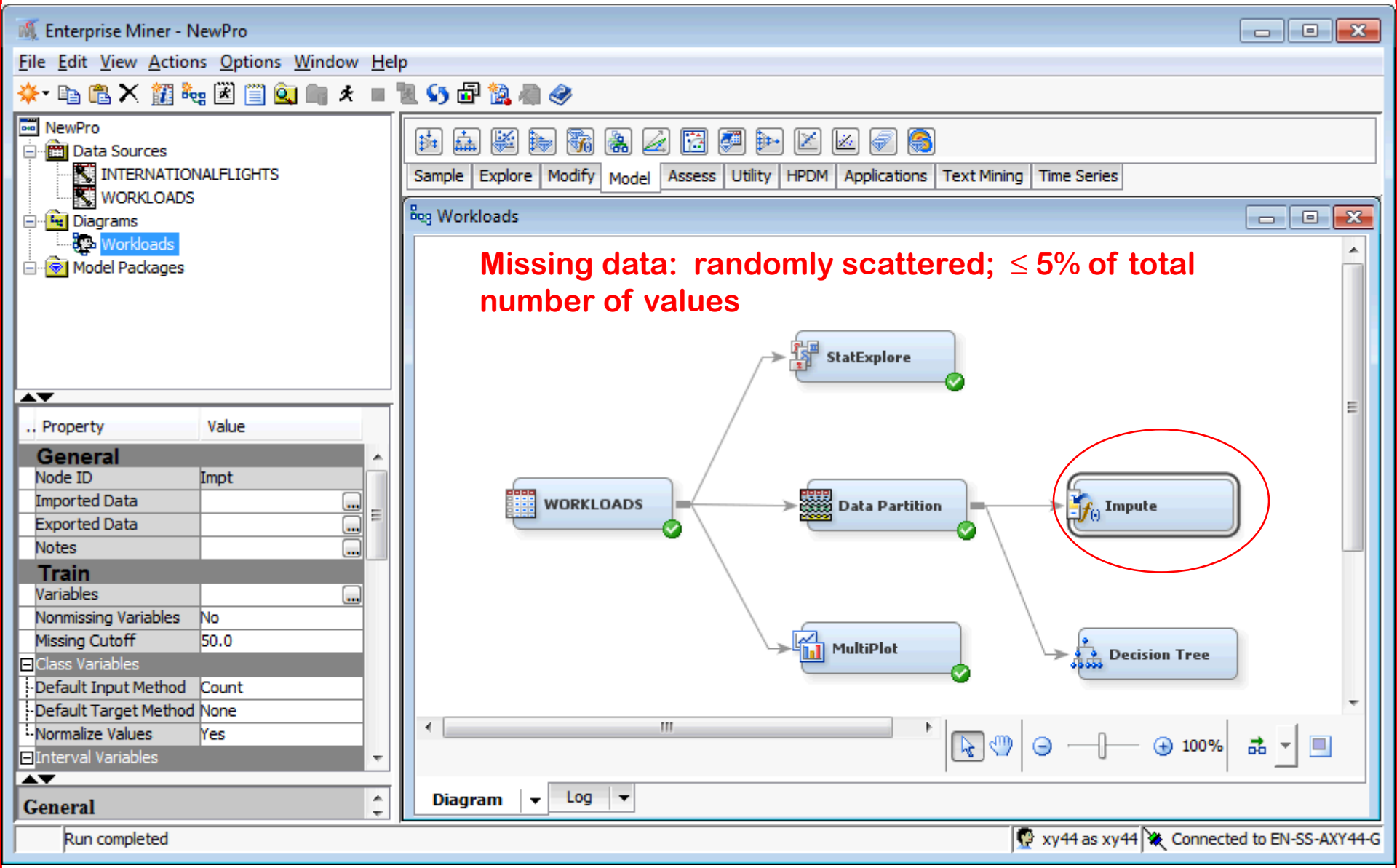
The partition results



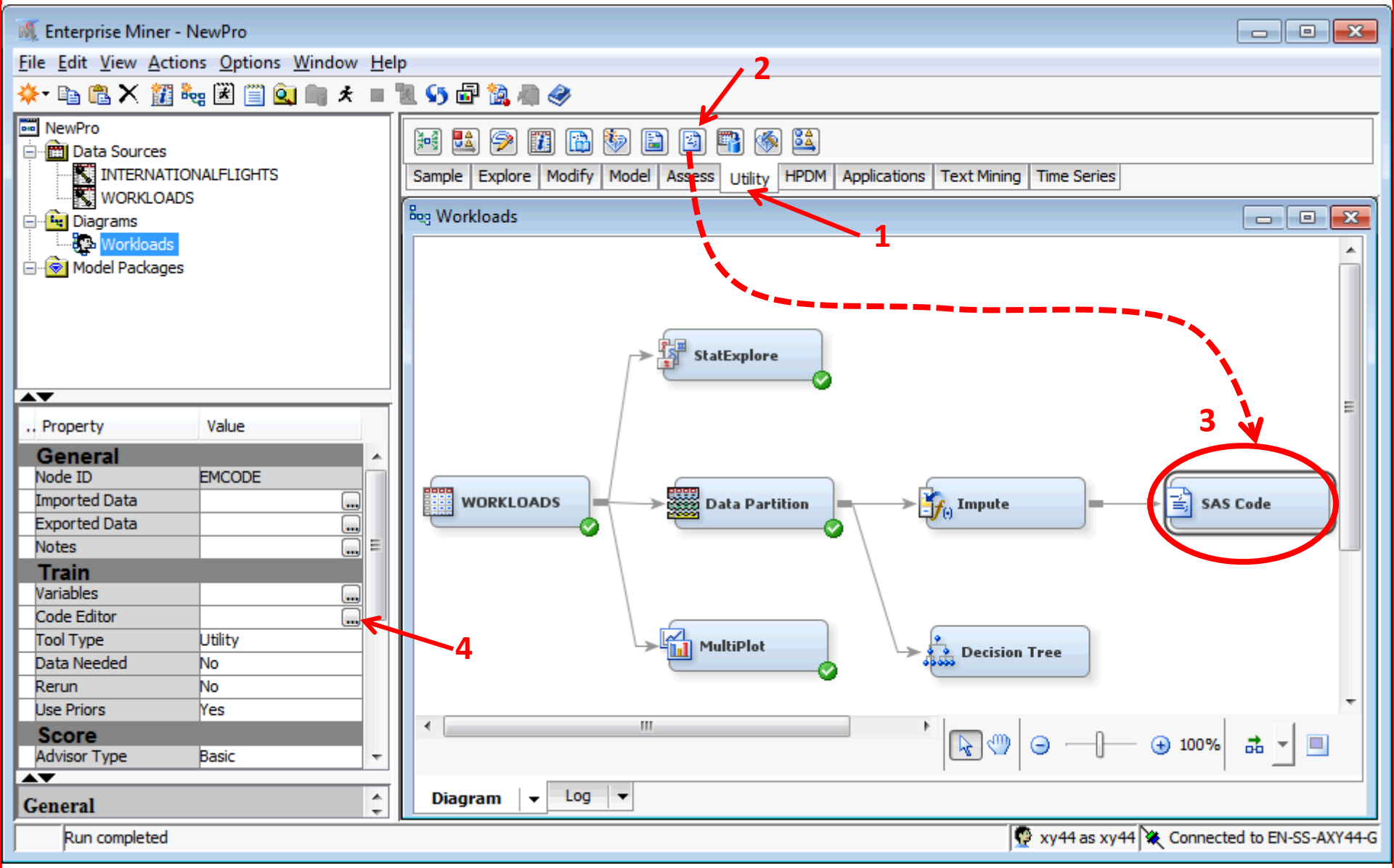
Add more nodes: Drop, Filter & Transformation …



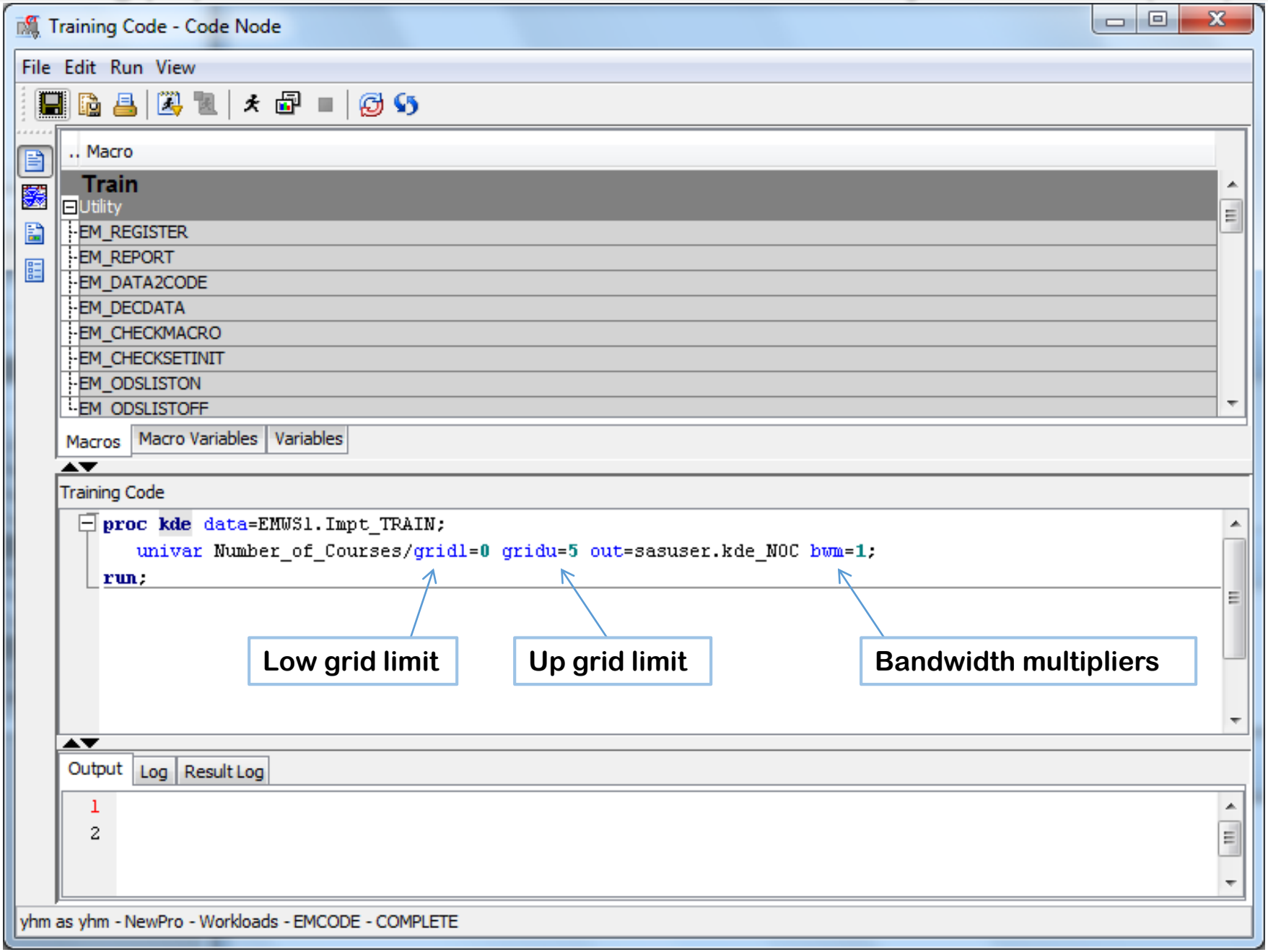
Impute: to replace missing values



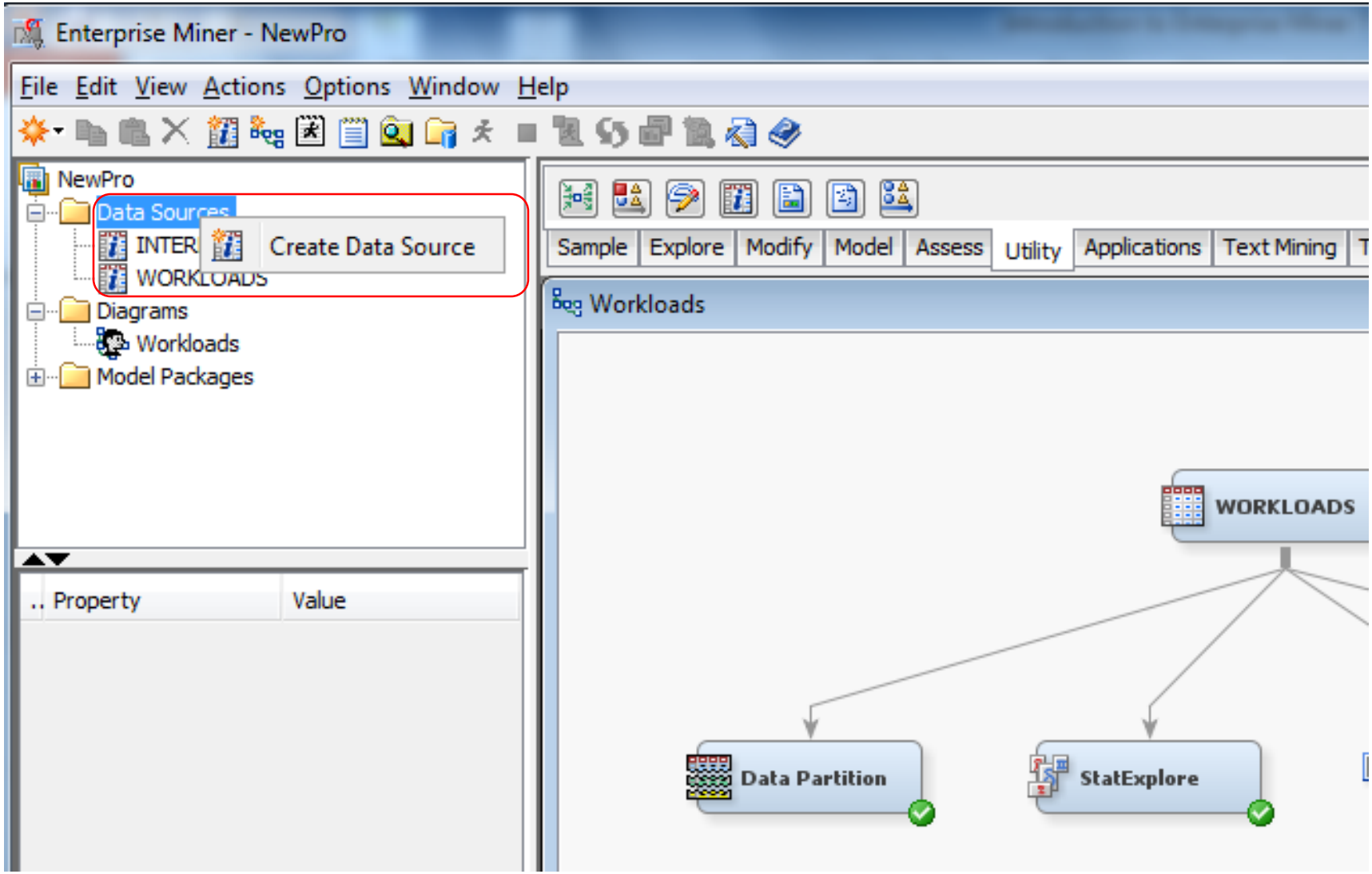
Integration of SAS codes into SAS EM 13.2.1



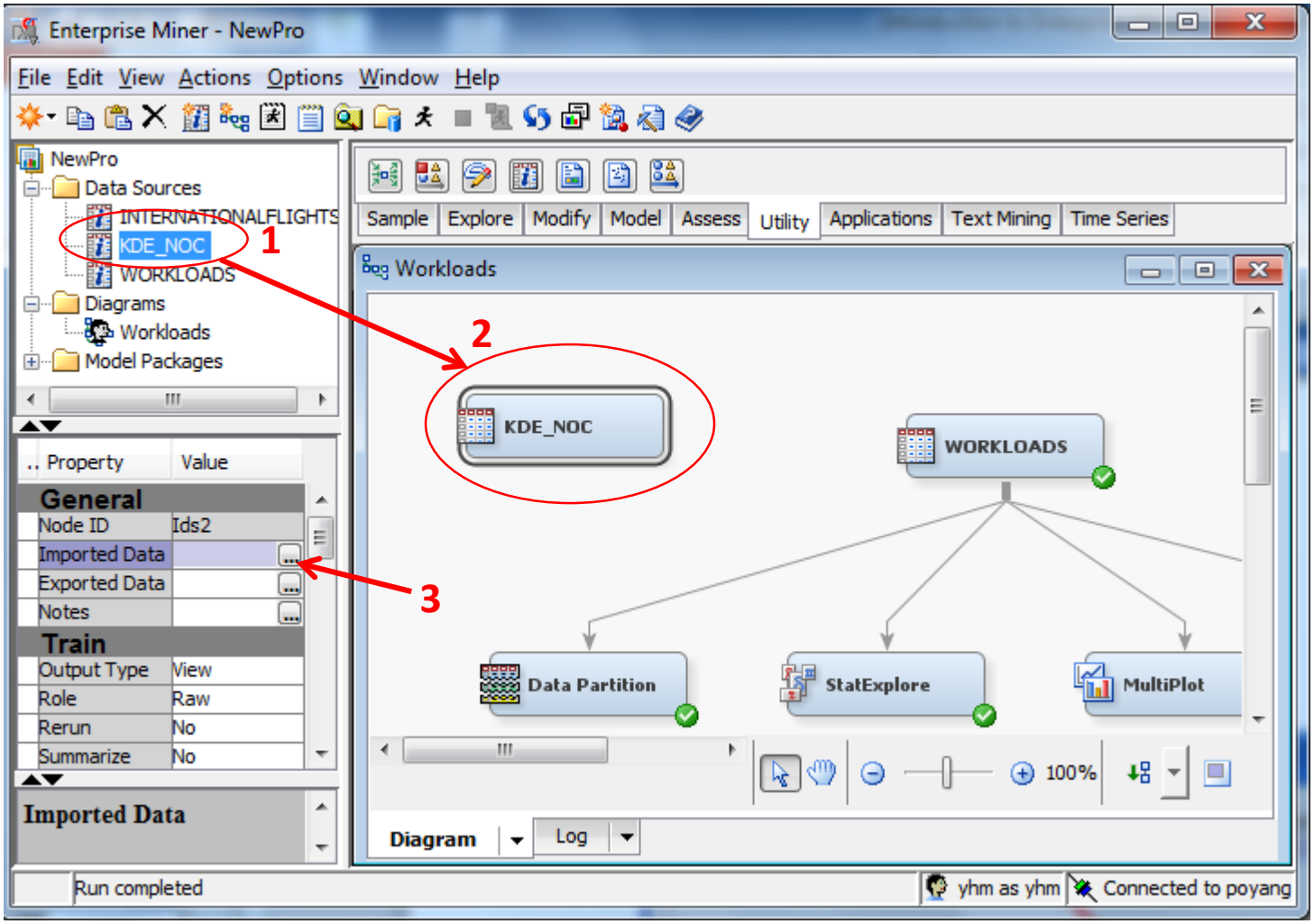
Finding population distribution with kernal density estimation (kde)



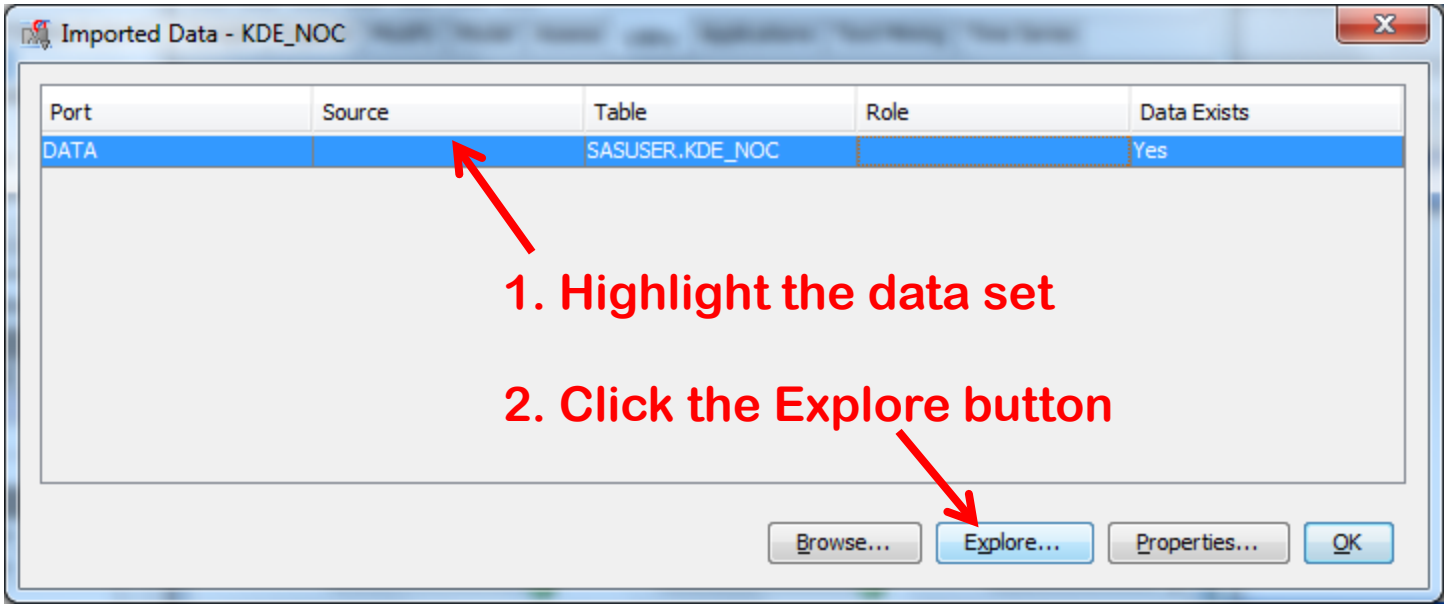
Enter the result into the project by creating a data source



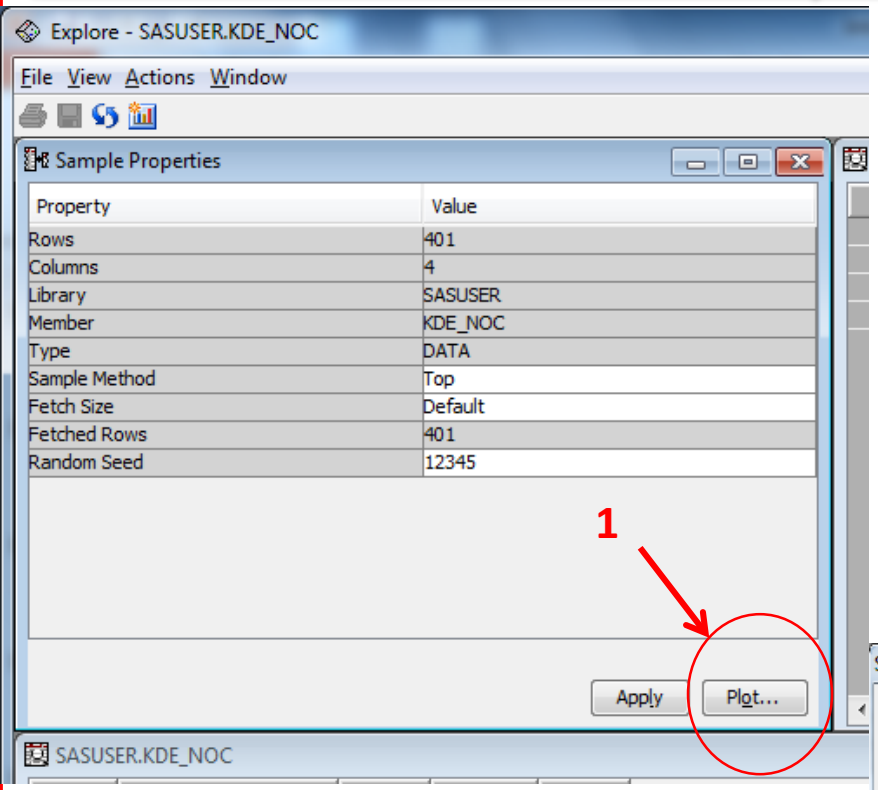
Create a node with the new data source

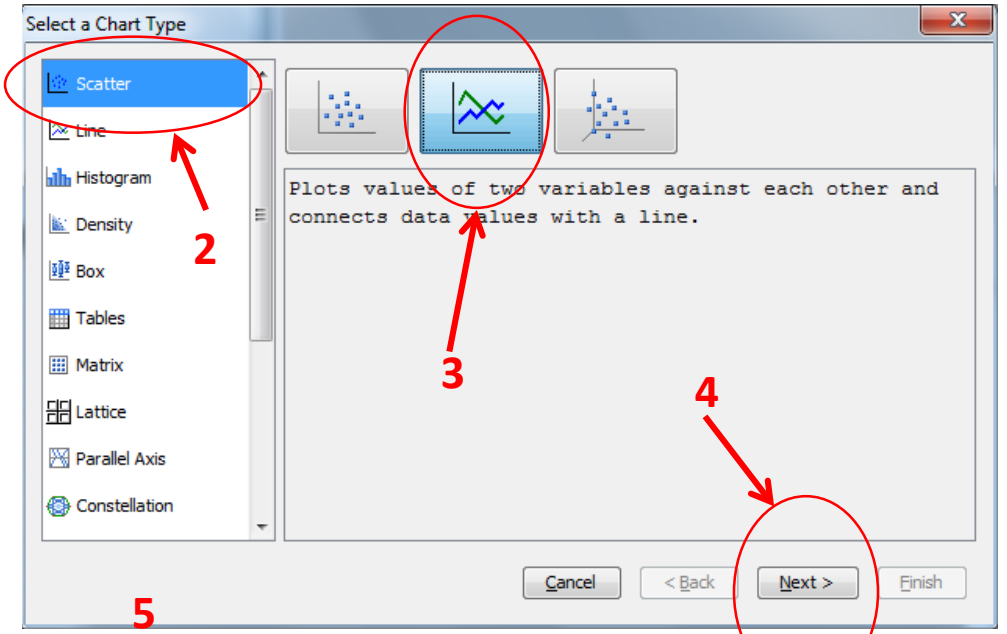


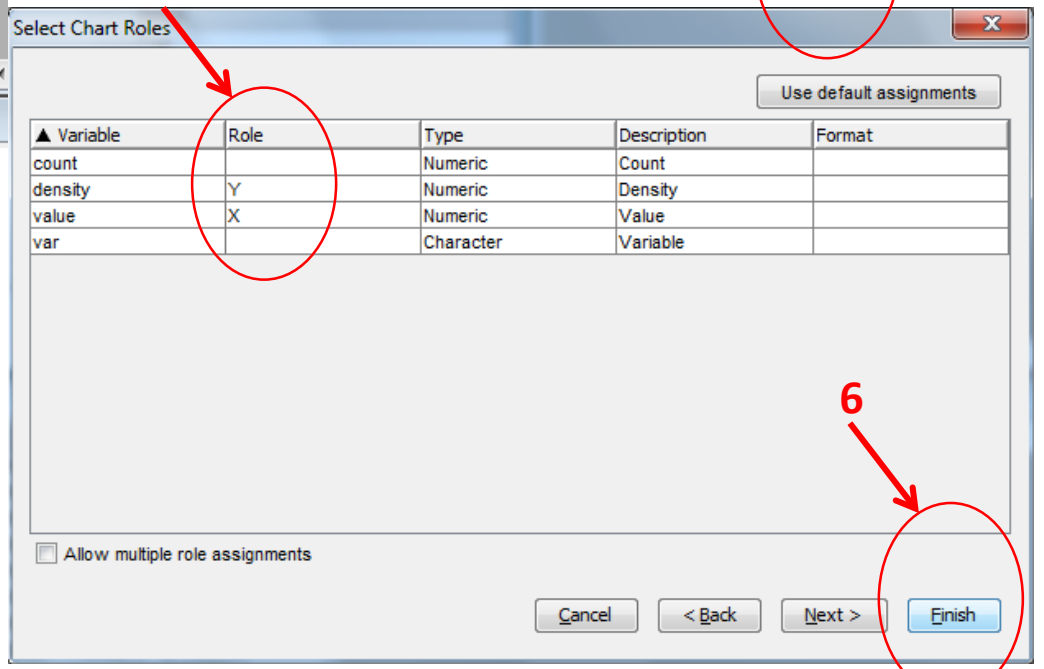
Explore data through the Imported Data property



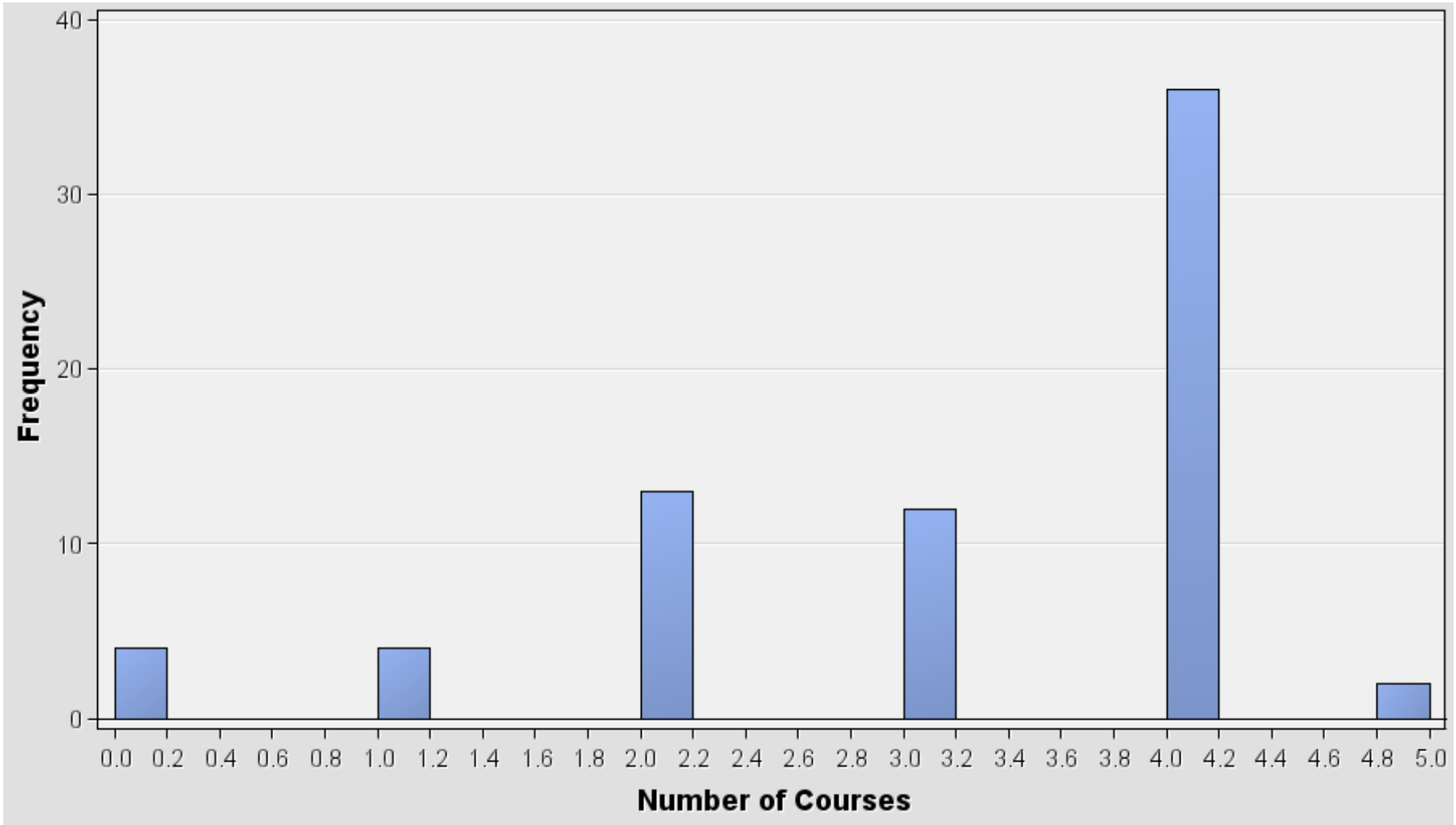
Plot a line-connected scatter plot



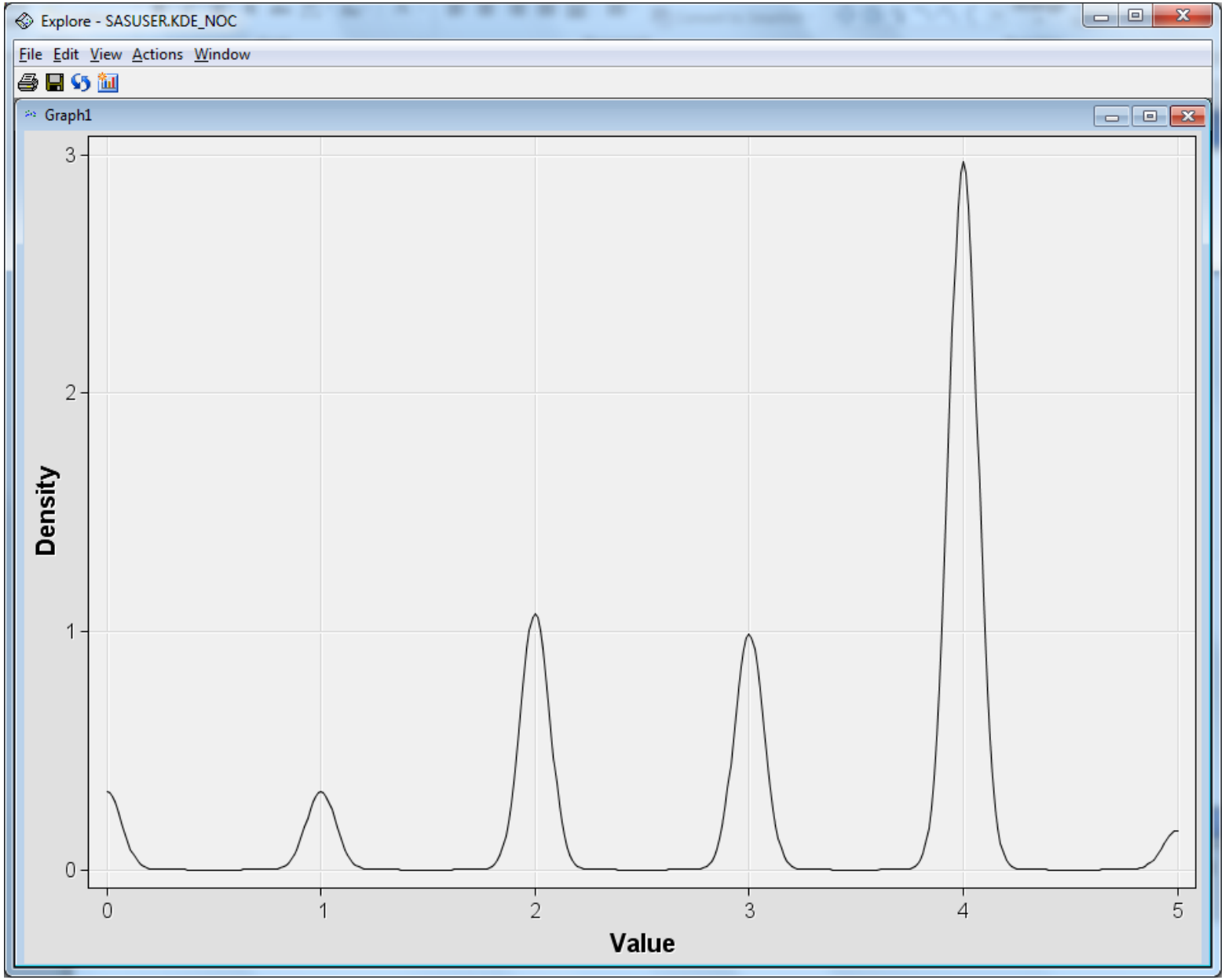




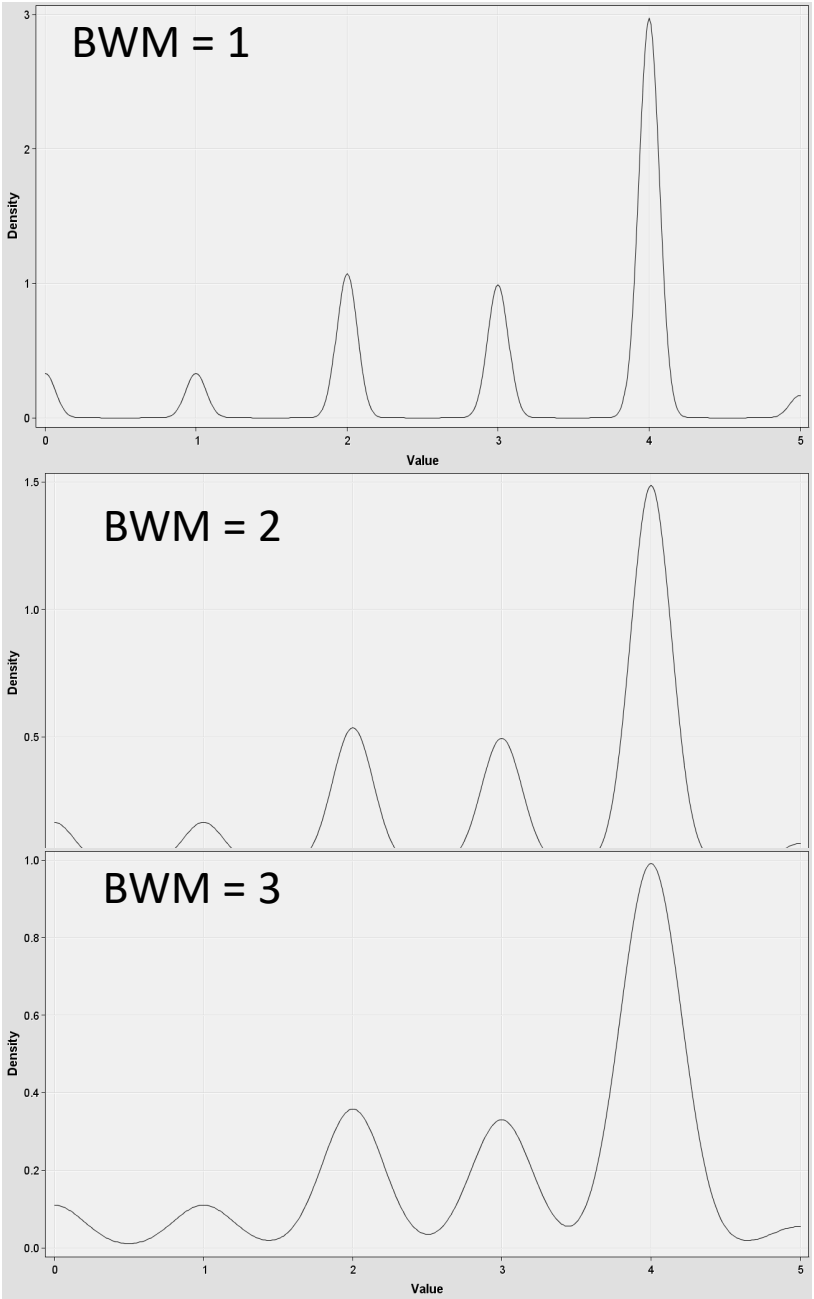
Bar chart of number of courses before kde

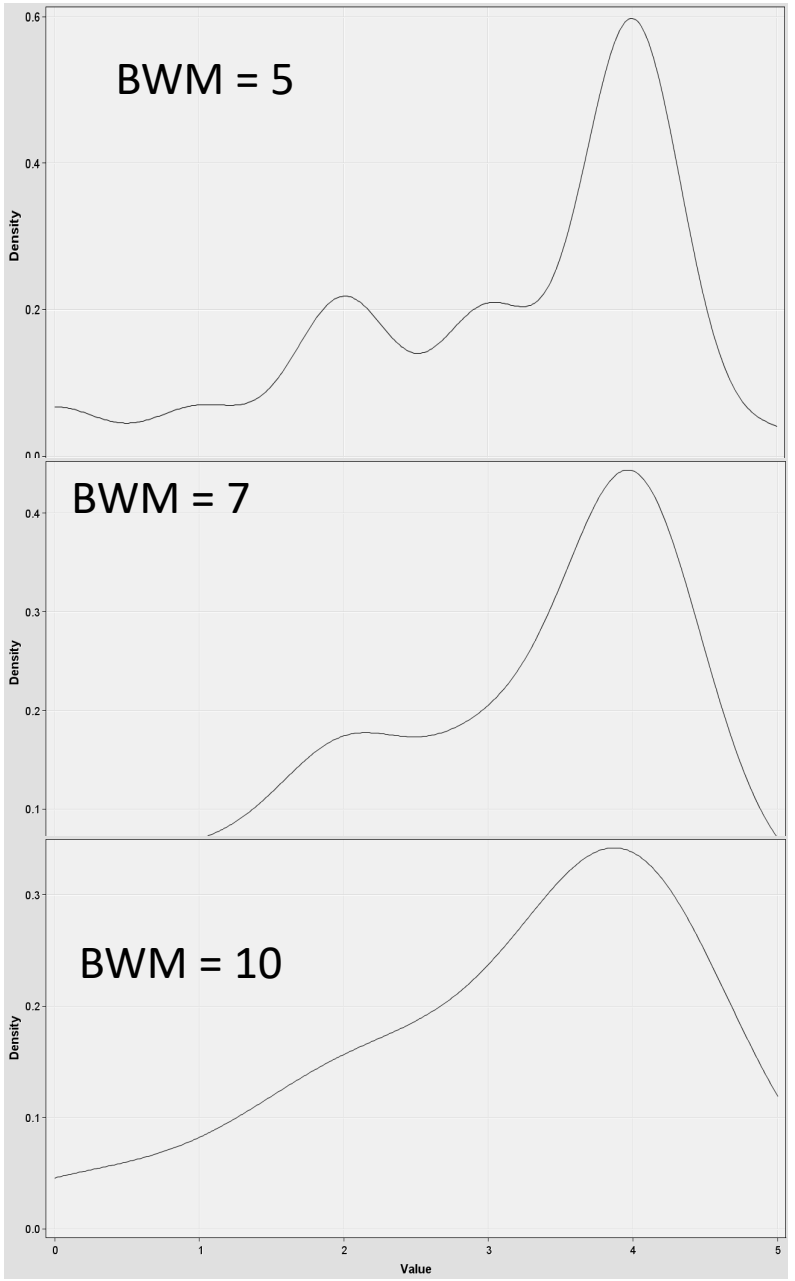


A plot after kde



Distributions smoothed with kde using different BWM values

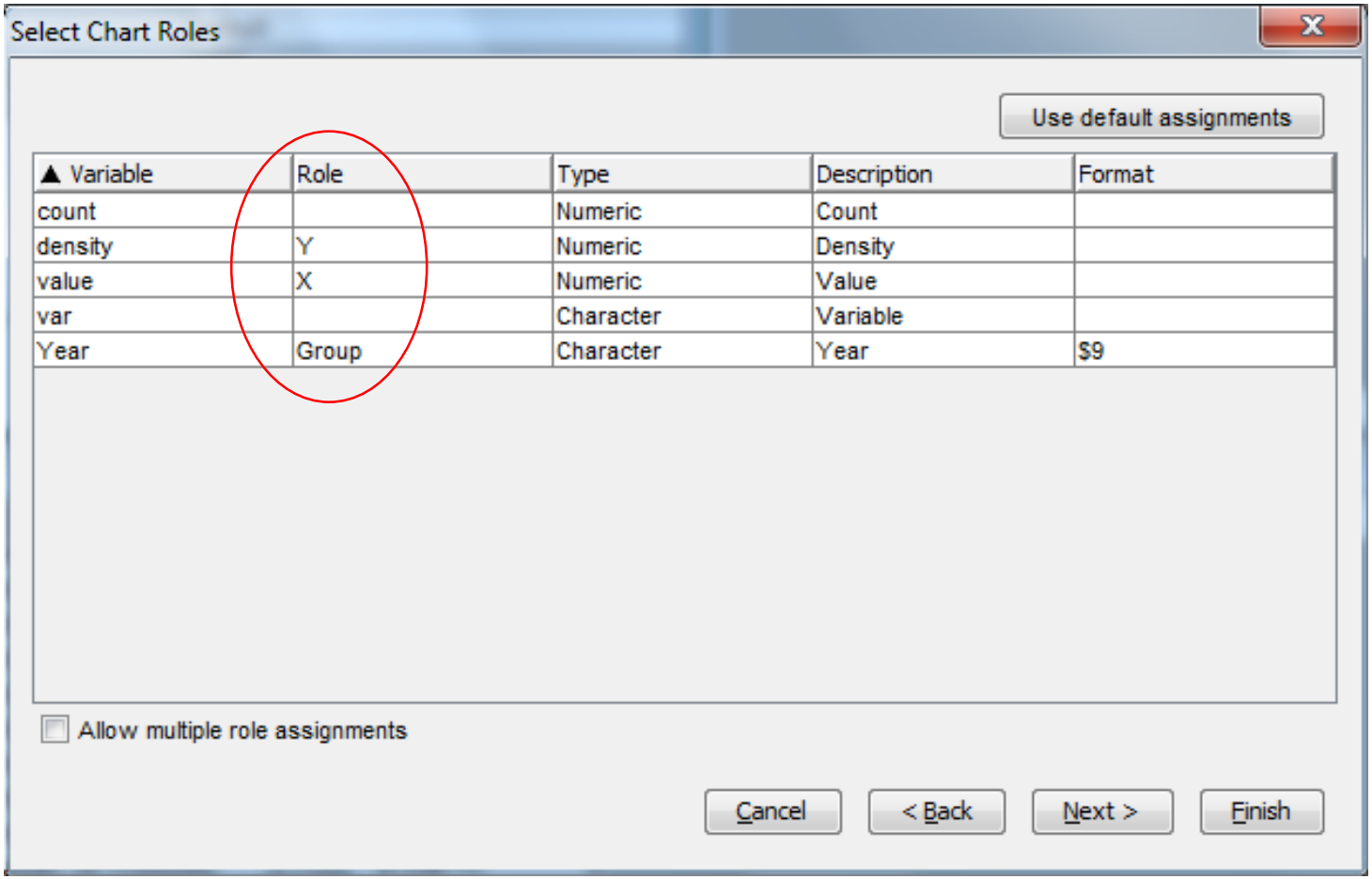




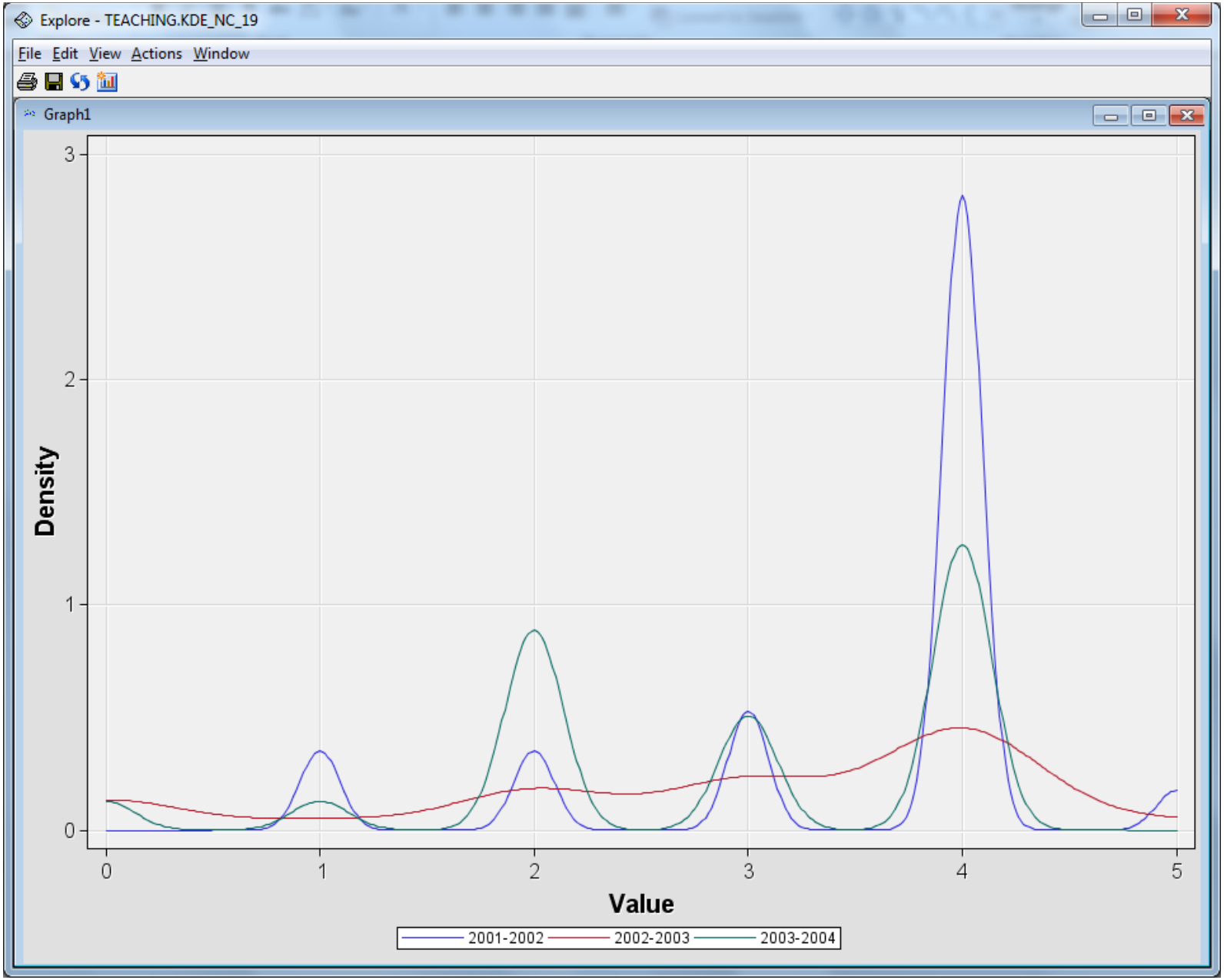
Code to produce grouped kde graphs

```
proc kde data=teaching.workloads;
  univar Number_of_Courses/gridl=0 gridu=5
  out=teaching.kde_NC_19;
  by year;
run;
```

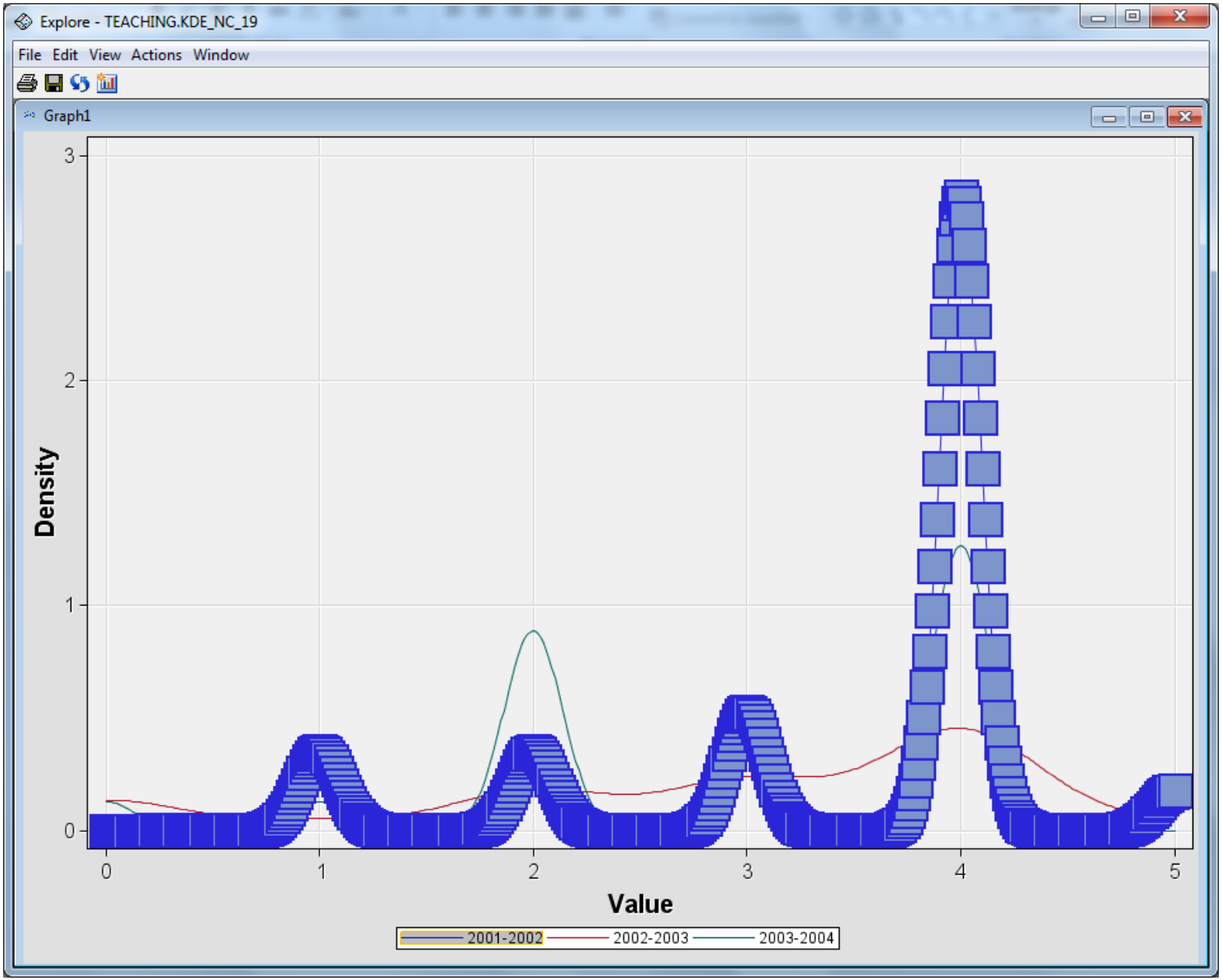
Determine the chart roles



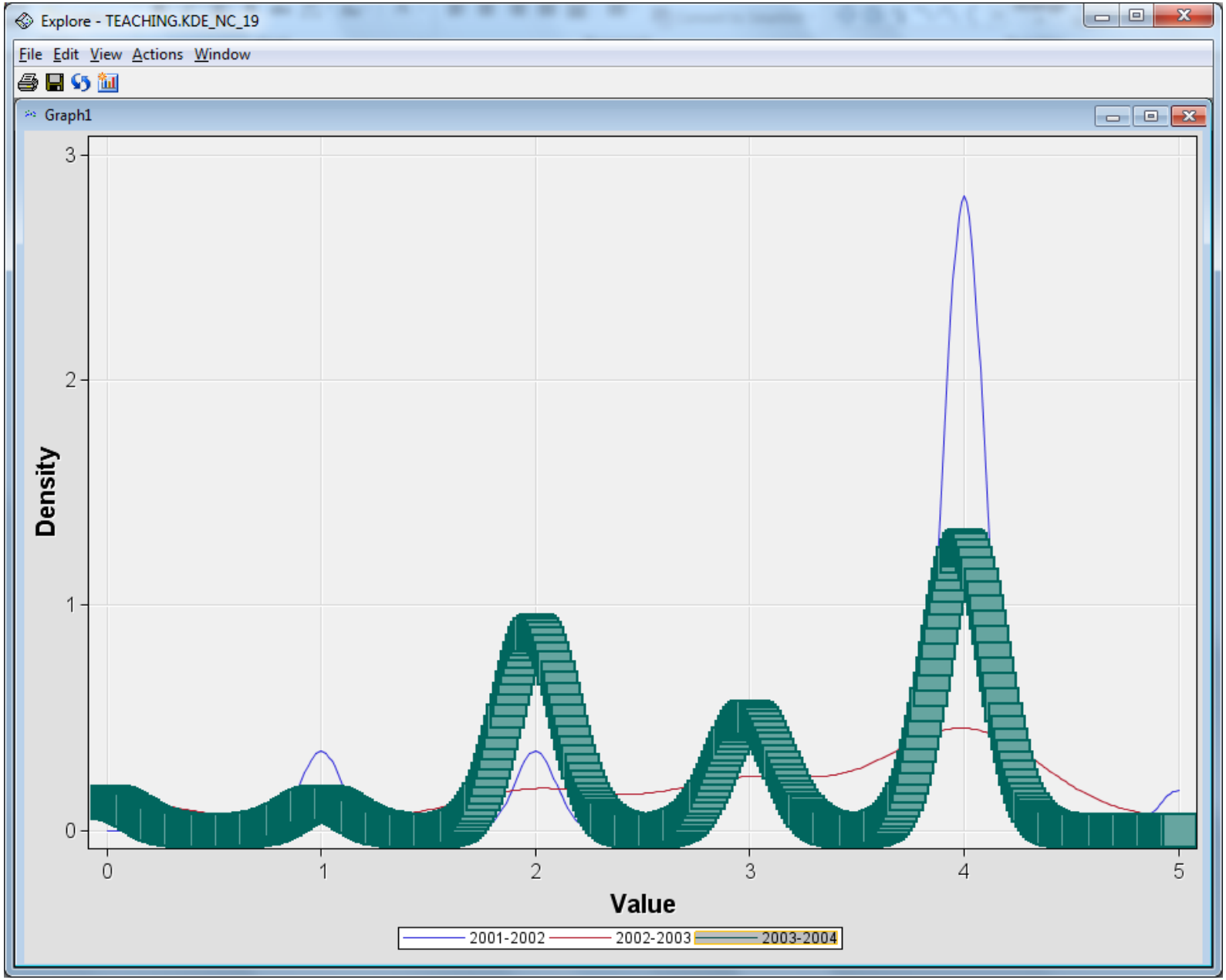
Overlay of kde graphs of number of courses by year



Overlay of kde graphs of number of courses by year



Overlay of kde graphs of number of courses by year



Program Demonstration

If time permits…