

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*

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series

STSCI5010

WORKLOADS

STSCI5010

Model Comparison

Regression

Decision Tree

Impute

StatExplore

Data Partition

MultiPlot

WORKLOADS

Diagram | Log

Diagram STSCI5010 opened

xy44 as xy44 Connected to EN-SS-AXY44-H

* The current version is 14.12?

SAS Enterprise Miner – Interface Tour

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

Menu bar and shortcut buttons

Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series

STSCI5010

WORKLOADS

Diagrams

STSCI5010

Model Packages

Property Value

ID	EMWS1
Name	STSCI5010
Status	Open
Notes	
History	
Create Date	11/20/15 1:54

ID

Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.

Diagram STSCI5010 opened

xy44 as xy44 Connected to EN-SS-AXY44-H

Diagram Log

```
graph LR; WORKLOADS[WORKLOADS] --> DataPartition1[Data Partition]; WORKLOADS --> MultiPlot[MultiPlot]; DataPartition1 --> StatExplore[StatExplore]; DataPartition1 --> Impute[Impute]; Impute --> Regression[Regression]; Impute --> DecisionTree[Decision Tree]; Regression --> ModelComparison[Model Comparison]; DecisionTree --> ModelComparison;
```

SAS Enterprise Miner – Interface Tour

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

Project panel

stsci5010

- Data Sources
 - THERAPY1999
 - WORKLOADS
- Diagrams
 - STSCI5010
- Model Packages

WORKLOADS

StatExplore

Impute

Regression

Decision Tree

Model Comparison

MultiPlot

Data Partition

STSCI5010

Diagram | Log

Diagram STSCI5010 opened

xy44 as xy44 Connected to EN-SS-AXY44-H

The screenshot shows the SAS Enterprise Miner interface. On the left, a 'Project panel' is highlighted with a red border, displaying a tree view of the project structure under 'stsci5010'. The 'Diagrams' section contains a selected item 'STSCI5010'. Below the project panel is a 'Properties' table with fields like ID, Name, Status, Notes, History, and Create Date. To the right of the project panel is the main workspace titled 'STSCI5010', which contains a flowchart of mining steps. The flow starts with 'WORKLOADS', which branches into 'StatExplore', 'Data Partition', and 'MultiPlot'. 'Data Partition' then connects to 'Impute', which further connects to 'Regression' and 'Decision Tree'. Both 'Regression' and 'Decision Tree' connect to 'Model Comparison'. At the bottom of the workspace are zoom controls (100%) and navigation buttons (back, forward, search). The bottom status bar indicates 'Diagram STSCI5010 opened' and 'Connected to EN-SS-AXY44-H'. The top menu bar includes File, Edit, View, Actions, Options, Window, and Help.

SAS Enterprise Miner – Interface Tour

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series

STSCI5010

WORKLOADS

Diagrams

Model Packages

Properties panel

ID

Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.

Diagram STSCI5010 opened

xy44 as xy44 Connected to EN-SS-AXY44-H

Diagram | Log

The screenshot shows the SAS Enterprise Miner software interface. On the left, there's a navigation pane with a tree view of projects and a properties panel for a selected item named 'STSCI5010'. A red arrow points from the text 'Properties panel' to this properties panel. The main workspace displays a workflow diagram titled 'STSCI5010'. The diagram consists of several nodes connected by arrows: 'WORKLOADS' connects to 'Data Partition', which then connects to 'StatExplore', 'Impute', 'Regression', and 'Decision Tree'. 'StatExplore', 'Impute', 'Regression', and 'Decision Tree' all connect to a final node 'Model Comparison'. There are also two additional 'Data Partition' nodes shown near the bottom of the diagram. The top menu bar includes File, Edit, View, Actions, Options, Window, and Help. The toolbar above the menu has various icons for file operations like Open, Save, and Print. The bottom of the screen shows a status bar with 'Diagram STSCI5010 opened' and a connection message 'xy44 as xy44 Connected to EN-SS-AXY44-H'.

SAS Enterprise Miner – Interface Tour

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series

STSCI5010

WORKLOADS

Diagrams

STSCI5010

Model Packages

Property Value

ID	EMWS1
Name	STSCI5010
Status	Open
Notes	
History	
Create Date	11/20/15 1:54

ID

Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.

Properties help panel

Diagram STSCI5010 opened

xy44 as xy44 Connected to EN-SS-AXY44-H

SAS Enterprise Miner – Interface Tour

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series

STSCI5010

WORKLOADS

STSCI5010

Model Packages

Property Value

ID	EMWS1
Name	STSCI5010
Status	Open
Notes	
History	
Create Date	11/20/15 1:54

ID

Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.

Diagram STSCI5010 opened

xy44 as xy44 Connected to EN-SS-AXY44-H

Diagram workspace

```
graph LR; WORKLOADS[WORKLOADS] --> StatExplore[StatExplore]; WORKLOADS --> DataPartition1[Data Partition]; WORKLOADS --> MultiPlot[MultiPlot]; DataPartition1 --> Impute[Impute]; Impute --> Regression[Regression]; Impute --> DecisionTree[Decision Tree]; Regression --> ModelComparison[Model Comparison]; DecisionTree --> ModelComparison;
```

The diagram workspace displays a workflow starting from a 'WORKLOADS' node. It branches into three parallel paths: one leading to 'StatExplore', another to 'Data Partition', and a third to 'MultiPlot'. The 'Data Partition' and 'MultiPlot' nodes merge into a single 'Data Partition' node. This then connects to 'Impute', which further connects to 'Regression' and 'Decision Tree'. Both 'Regression' and 'Decision Tree' connect to a final 'Model Comparison' node.

SAS Enterprise Miner – Interface Tour

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series

STSCI5010

Process flow

```
graph LR; WORKLOADS[WORKLOADS] --> StatExplore[StatExplore]; WORKLOADS --> DataPartition1[Data Partition]; WORKLOADS --> MultiPlot[MultiPlot]; DataPartition1 --> Impute[Impute]; DataPartition1 --> DecisionTree[Decision Tree]; Impute --> Regression[Regression]; DecisionTree --> Regression; Regression --> ModelComparison[Model Comparison]; ModelComparison --> ModelComparison
```

Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.

Diagram STSCI5010 opened

xy44 as xy44 Connected to EN-SS-AXY44-H

SAS Enterprise Miner – Interface Tour

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series

STSCI5010

WORKLOADS

Diagrams

Model Packages

Property Value

ID	EMWS1
Name	STSCI5010
Status	Open
Notes	
History	
Create Date	11/20/15 1:54

ID

Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.

Diagram STSCI5010 opened

xy44 as xy44 Connected to EN-SS-AXY44-H

```
graph LR; WORKLOADS[WORKLOADS] --> StatExplore[StatExplore]; WORKLOADS --> DataPartition[Data Partition]; WORKLOADS --> MultiPlot[MultiPlot]; DataPartition --> Impute[Impute]; MultiPlot --> Impute; Impute --> Regression[Regression]; Impute --> DecisionTree[Decision Tree]; Regression --> ModelComparison[Model Comparison]; DecisionTree --> ModelComparison;
```

SAS Enterprise Miner – Interface Tour

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

SEMMA tools palette

Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series

STSCI5010

WORKLOADS

WORKLOADS

StatExplore

Impute

Regression

Decision Tree

Model Comparison

MultiPlot

Data Partition

Diagram Log

Diagram STSCI5010 opened

xy44 as xy44 Connected to EN-SS-AXY44-H

The screenshot shows the SAS Enterprise Miner interface. On the left, there's a navigation pane with 'stsci5010' selected, containing 'Data Sources' (THERAPY1999, WORKLOADS), 'Diagrams' (STSCI5010), and 'Model Packages'. Below it is a property grid for 'STSCI5010' with fields like ID, Name, Status, Notes, History, and Create Date. To the right is the main workspace titled 'STSCI5010' showing a workflow diagram. The diagram starts with a 'WORKLOADS' node, which branches into 'StatExplore', 'Data Partition', and 'MultiPlot'. The 'Data Partition' node further branches into 'Impute' and 'Decision Tree'. 'Impute' leads to 'Regression', which then leads to 'Model Comparison'. A red box highlights the top menu bar, and a red arrow points to the 'Assess' tab, labeled 'SEMMA tools palette'. The bottom status bar indicates 'Diagram STSCI5010 opened' and 'Connected to EN-SS-AXY44-H'.

SAS Enterprise Miner – Sample Tab

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series

- Append
- Data Partition
- File Import
- Filter
- Input Data
- Merge
- Sample

```
graph LR; WORKLOADS[WORKLOADS] --> StatExplore[StatExplore]; WORKLOADS --> DataPartition[Data Partition]; WORKLOADS --> MultiPlot[MultiPlot]; DataPartition --> Impute[Impute]; Impute --> Regression[Regression]; Impute --> DecisionTree[Decision Tree]; Regression --> ModelComparison[Model Comparison]; DecisionTree --> ModelComparison;
```

Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.

Diagram STSCI5010 opened

xy44 as xy44 Connected to EN-SS-AXY44-H

SEMMA – Explore Tab

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series

STSCI5010

- Association
- Cluster
- DMDB
- Graph Explore
- Market Basket
- MultiPlot
- Path Analysis
- SOM/Kohonen
- StatExplore
- Variable Clustering
- Variable Selection

Diagram

Log

Diagram STSCI5010 opened

xy44 as xy44 Connected to EN-SS-AXY44-H

The screenshot shows the SEMMA software interface with the 'Explore' tab highlighted. A list of analysis methods is overlaid in red text. In the background, a process flow diagram is visible, showing nodes like 'StatExplore', 'Impute', 'Regression', 'Decision Tree', and 'Model Comparison' connected by arrows. The left sidebar shows a tree view of 'stsci5010' with 'THERAPY1999' and 'WORKLOADS' expanded. The bottom status bar indicates 'Diagram STSCI5010 opened'.

SEMMA – Modify Tab

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

Sample Explore **Modify** Model Assess Utility HPDM Applications Text Mining Time Series

STSCI5010

- Drop
- Impute
- Interactive Binning
- Principal Components
- Replacement
- Rules Builder
- Transform Variables

WORKLOADS

.. Property Value

ID	EMWS1
Name	STSCI5010
Status	Open
Notes	[...]
History	[...]
Create Date	11/20/15 1:54

ID

Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.

Diagram STSCI5010 opened

xy44 as xy44 Connected to EN-SS-AXY44H

The screenshot shows the SEMMA application window with the 'Modify' tab selected. On the left, there's a tree view of project components like Data Sources, Diagrams, and Model Packages, with 'STSCI5010' selected. A properties panel shows details for 'STSCI5010'. Below that is a section for 'ID' with a description. The main workspace displays a data mining workflow: 'StateSpace' feeds into 'Impute', which then splits into 'Regression' and 'Decision Tree', both of which feed into 'Model Comparison'. A red box highlights the 'Modify' tab in the menu bar, and a list of modification tools is overlaid on the workflow diagram.

SEMMA – Model Tab

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

Sample Explore Modify **Model** Assess Utility HPDM Applications Text Mining Time Series

STSCI5010

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

WORKLOADS

Impute → Regression → Model Comparison

WORKLOADS

Diagram | Log

Diagram STSCI5010 opened

xy44 as xy44 Connected to EN-SS-AXY44H

The screenshot shows the SEMMA interface with the 'Model' tab selected. On the right, a workflow diagram for 'STSCI5010' is displayed, featuring nodes for 'Impute', 'Regression', and 'Decision Tree' connected to a 'Model Comparison' node. A 'WORKLOADS' node is also present. On the left, there's a tree view of project structure ('stsci5010'), a properties panel with fields like ID, Name, Status, Notes, History, and Create Date, and a detailed 'ID' section. A large list of modeling techniques is overlaid on the right side of the screen.

SEMMA – Assess Tab

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

Sample Explore Modify Model **Assess** Utility HPDM Applications Text Mining Time Series

STSCI5010

- Cutoff
- Decisions
- Model Comparison
- Score
- Segment Profile

WORKLOADS

WORKLOADS → Data Partition → Impute → Regression → Model Comparison

WORKLOADS → MultiPlot

Diagram | Log | 100%

Diagram STSCI5010 opened

xy44 as xy44 Connected to EN-SS-AXY44H

```
graph LR; WORKLOADS[WORKLOADS] --> DP[Data Partition]; DP --> I[Impute]; I --> R[Regression]; I --> DT[Decision Tree]; R --> MC[Model Comparison]; DT --> MC; MP[MultiPlot] --> WORKLOADS;
```

Beyond SEMMA – Utility Tab

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

Utility tab highlighted

Diagram STSCI5010 opened

Diagram Identifier: This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.

WORKLOADS

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

STSCI5010

Impute

Regression

Decision Tree

Model Comparison

```
graph LR; WORKLOADS -->|Impute| Impute -->|Regression| Regression -->|Decision Tree| Decision Tree -->|Model Comparison| ModelComparison
```

Beyond SEMMA – HPDM Tab

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

HPDM Applications Text Mining Time Series

HP Cluster
HP Data Partition
HP Explore
HP Forest
HP GLM
HP Impute
HP Neural
HP Principle Components
HP Regression
HP SVM
HP Text Miner
HP Transform
HP Tree
HP Variable Selection

Diagram STSCI5010 opened

Connected to EN-SS-AXY44H

The screenshot shows the SAS Enterprise Miner application window. The title bar reads "Enterprise Miner - stsci5010". The menu bar includes "File", "Edit", "View", "Actions", "Options", "Window", and "Help". Below the menu is a toolbar with various icons. On the left, there's a project tree for "stsci5010" with nodes like "Data Sources", "Diagrams", and "Model Packages", and a properties panel for a selected "WORKLOADS" item. The main workspace displays a diagram titled "STSCI5010" showing a workflow from "WORKLOADS" through "StatExplorer", "Data Partition", "MultiPlot", "Impute", "Regression", "Decision Tree", and finally "Model Comparison". A red box highlights the "HPDM" tab in the top navigation bar. To the right of the diagram are zoom controls and a status bar indicating "Connected to EN-SS-AXY44H".

Beyond SEMMA – Application Tab

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series

STSCI5010

- Incremental Response
- Survival

WORKLOADS

StatExplore

Data Partition

MultiPlot

Impute

Regression

Decision Tree

Model Comparison

ID

Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.

Diagram STSCI5010 opened

xy44 as xy44 Connected to EN-SS-AXY44H

```
graph LR; WORKLOADS[WORKLOADS] --> StatExplore[StatExplore]; WORKLOADS --> DataPartition[Data Partition]; WORKLOADS --> MultiPlot[MultiPlot]; DataPartition --> Impute[Impute]; Impute --> Regression[Regression]; Impute --> DecisionTree[Decision Tree]; Regression --> ModelComparison[Model Comparison]; DecisionTree --> ModelComparison;
```

Beyond SEMMA – Text Mining Tab

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility HPDM Applications **Text Mining** Time Series

STSCI5010

WORKLOADS

WORKLOADS

StatExplore

Data Partition

MultiPlot

Impute

Regression

Decision Tree

Model Comparison

• Text Cluster
• Text Filter
• Text Import
• Text Parsing
• Text Profile
• Text Rule Builder
• Text Topic

Diagram STSCI5010 opened

xy44 as xy44 Connected to EN-SS-AXY44H

Beyond SEMMA –Time Series Tab

Enterprise Miner - stsci5010

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility HPDM Applications Text Mining **Time Series**

STSCI5010

- TS Correlation
- TS Data Preparation
- TS Decomposition
- TS Dimension Reduction
- TS Exponential Smoothing
- TS Similarity

WORKLOADS

StatExplore

Data Partition

Impute

MultiPlot

Decision Tree

Model

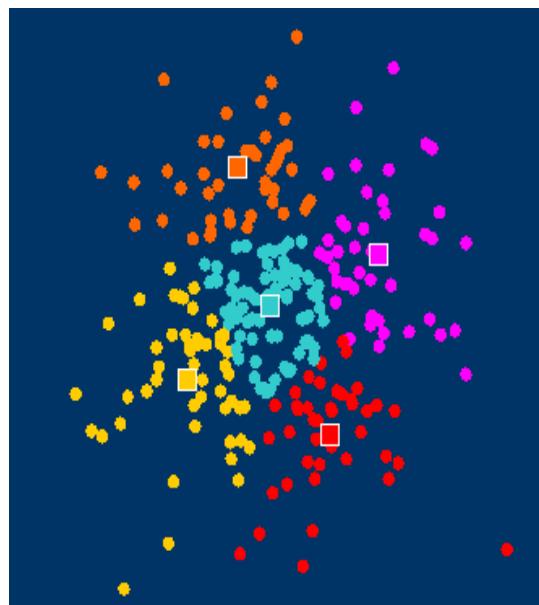
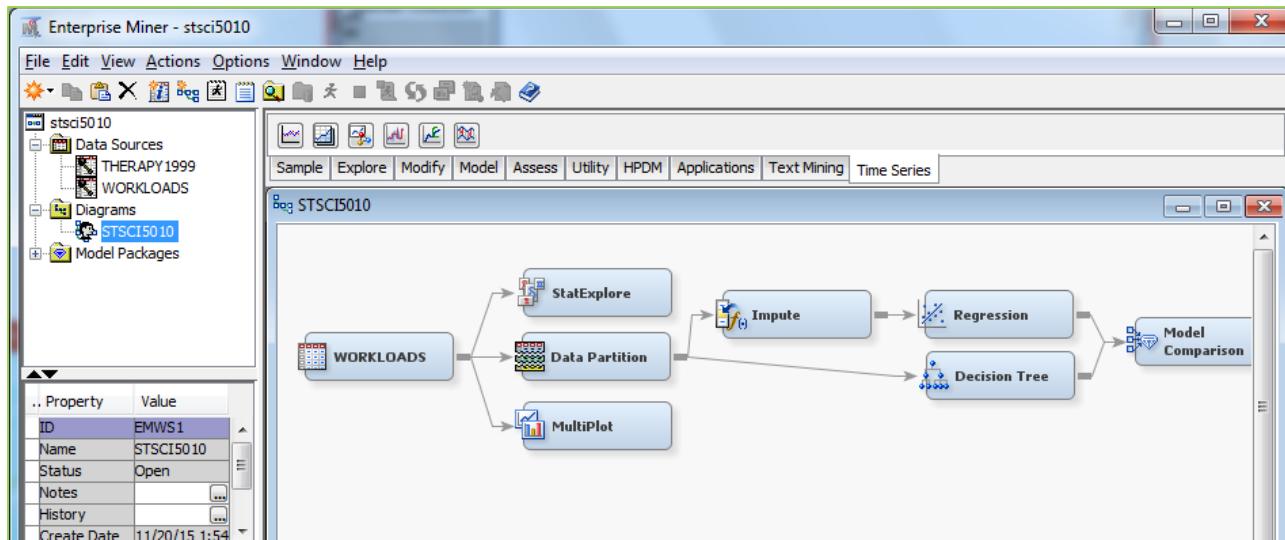
WORKLOADS

ID

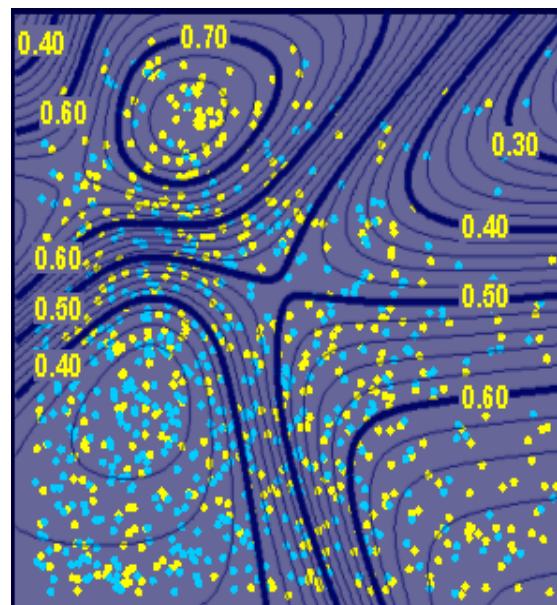
Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical

```
graph LR; WORKLOADS[WORKLOADS] --> StatExplore[StatExplore]; WORKLOADS --> DataPartition[Data Partition]; WORKLOADS --> MultiPlot[MultiPlot]; DataPartition --> WORKLOADS; Impute[Impute] --> DataPartition; MultiPlot --> WORKLOADS;
```

SAS Enterprise Miner Analytic Strengths



Pattern
Discovery



Predictive
Modeling

Setting up an EM Project

Start SAS Enterprise Miner 14.2

Start



All Programs



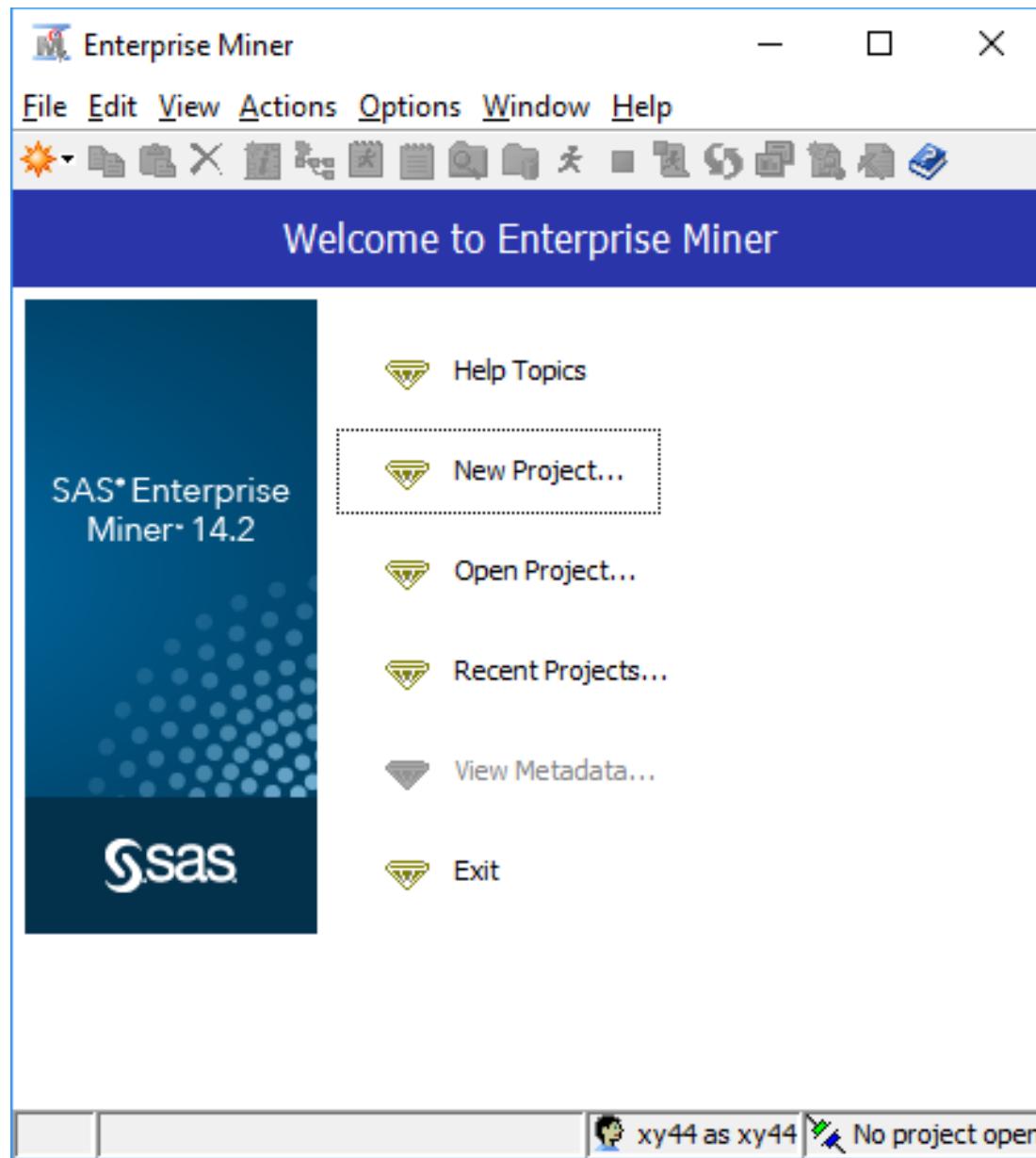
SAS



SAS Enterprise Miner Workstation 14.2 →



Click “New Project...”



Create a new project

 Create New Project -- Step 1 of 2 Specify Project Name and Server Directory X

Specify a project name and directory on the SAS Server for this project. All SAS data sets and files will be written to this location.

SAS* Enterprise Miner® 14.2

Project Name

SAS Server Directory Browse

[< Back](#) [Next >](#) [Cancel](#)

 Create New Project -- Step 2 of 2 New Project Information X

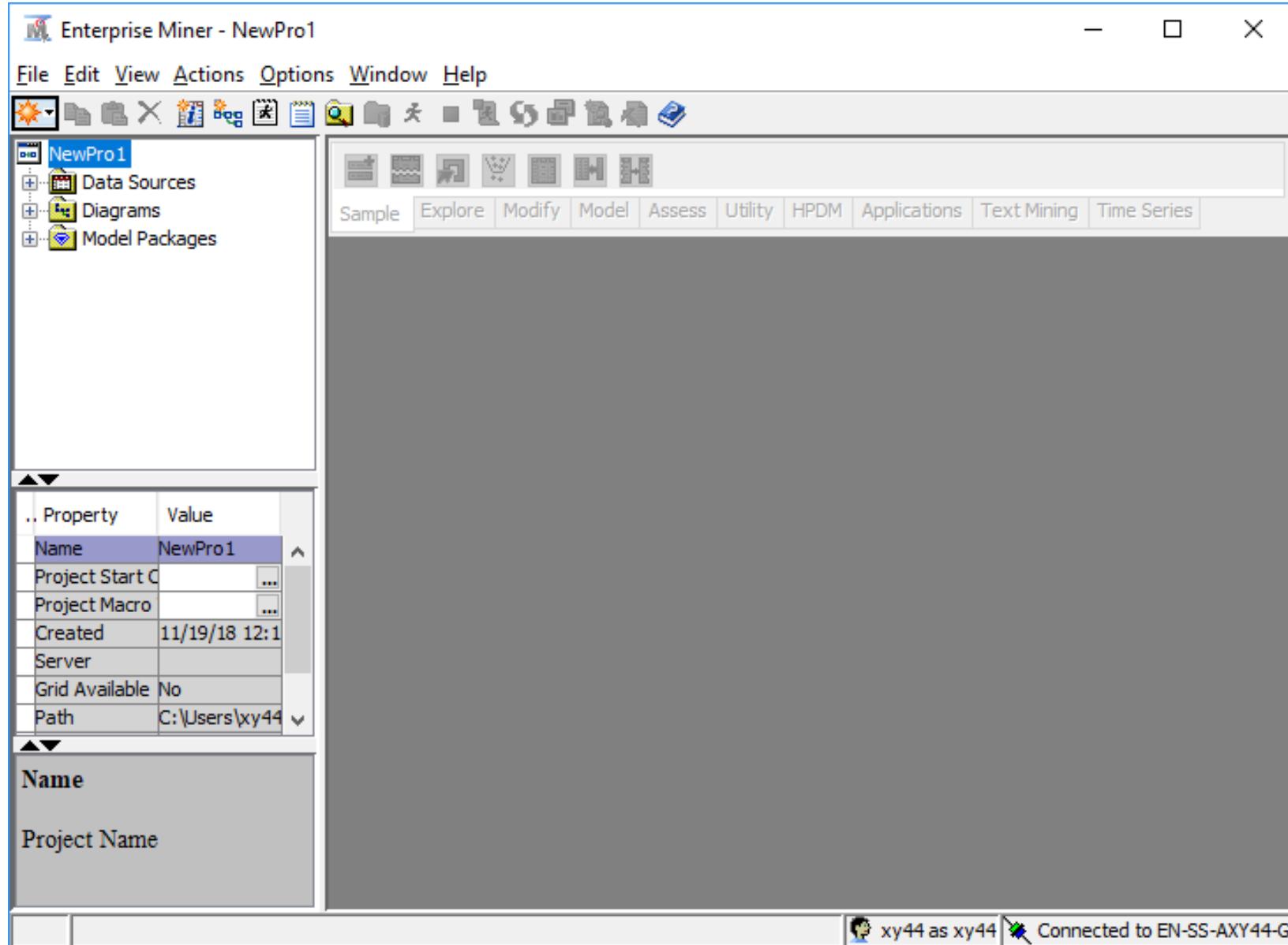
New Project Information

Name	NewPro1
Server Directory	C:\Users\xy44\Documents\SASEM

SAS* Enterprise Miner® 14.2

[< Back](#) Finish [Cancel](#)

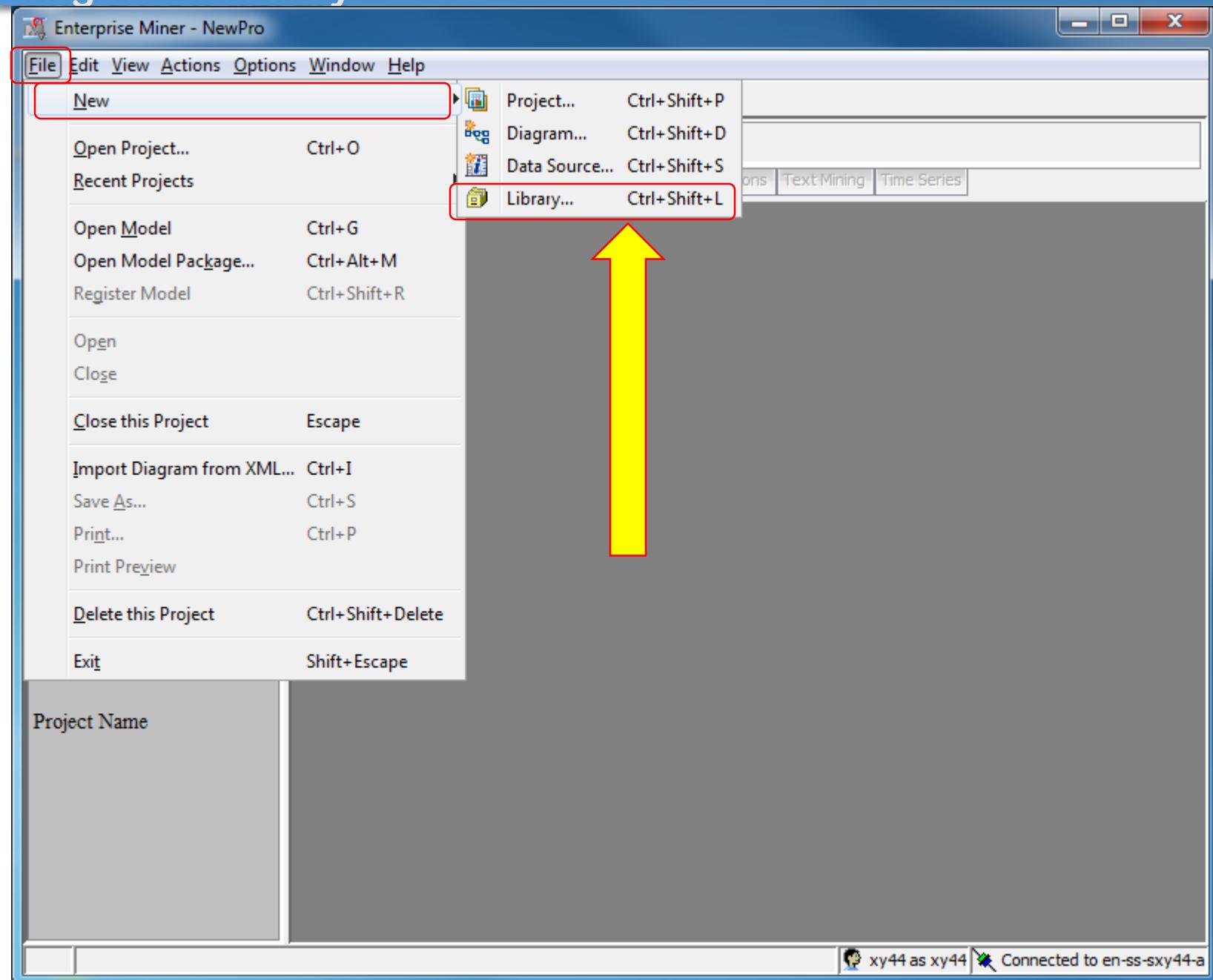
A new project window



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

Library Wizard -- Step 1 of 3 Select Action

Select an action —

Create New Library
 Modify Library
 Delete Library

Name	Engine	Path
------	--------	------

< Back Next > Cancel

Library Wizard -- Step 3 of 3 Confirm Action

... Property	Value
Action	Create New
Name	Teaching
Engine	BASE
Path	C:\Users\xy44\Documents\My SAS Files\9.4
Options	

Status —
Action Succeeded!
The Library "Teaching" was created.

< Back Finish

Library Wizard -- Step 2 of 3 Create or Modify

Name —
Teaching

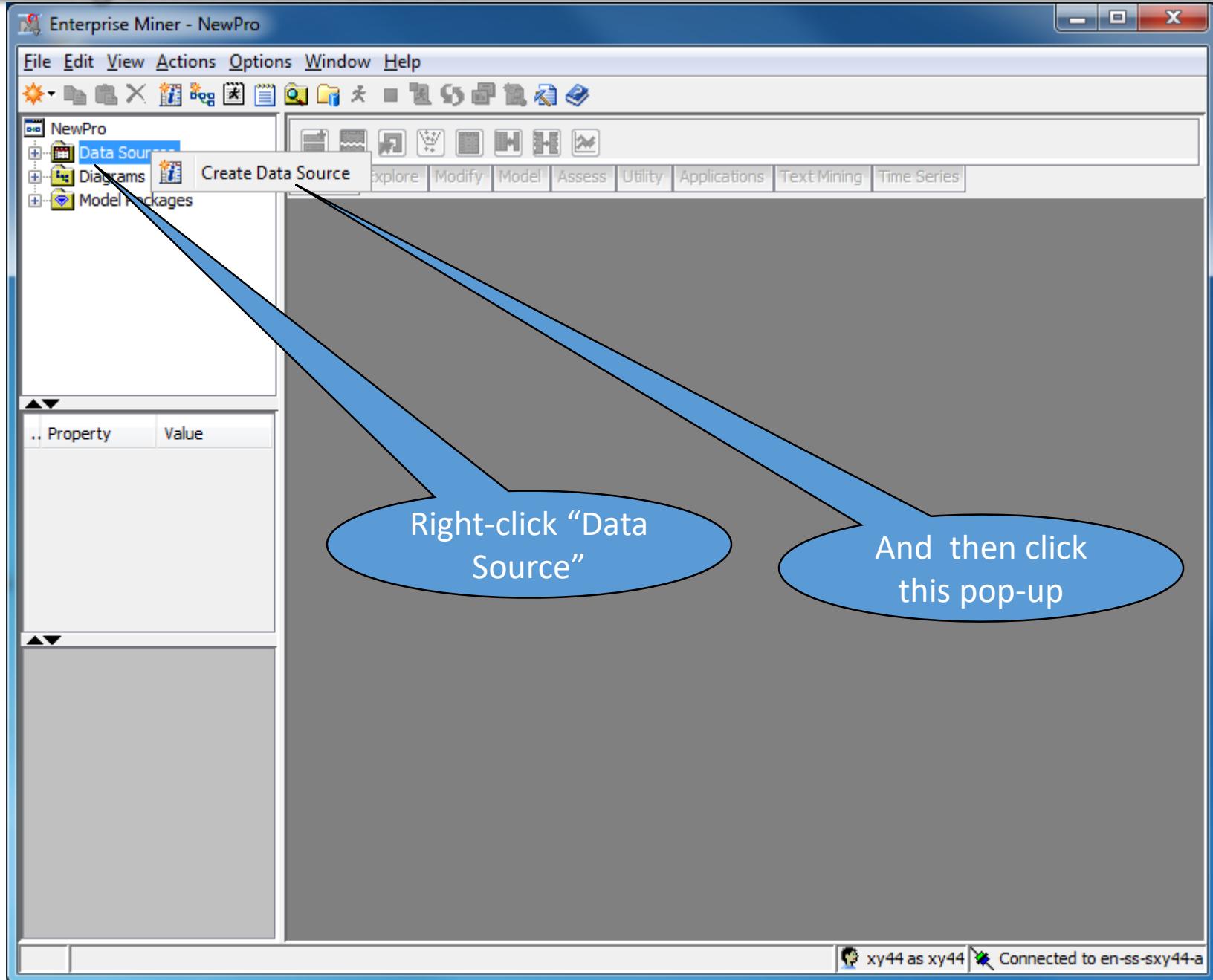
Engine —
BASE

Library information —
Path
C:\Users\xy44\Documents\My SAS Files\9.4 Browse...

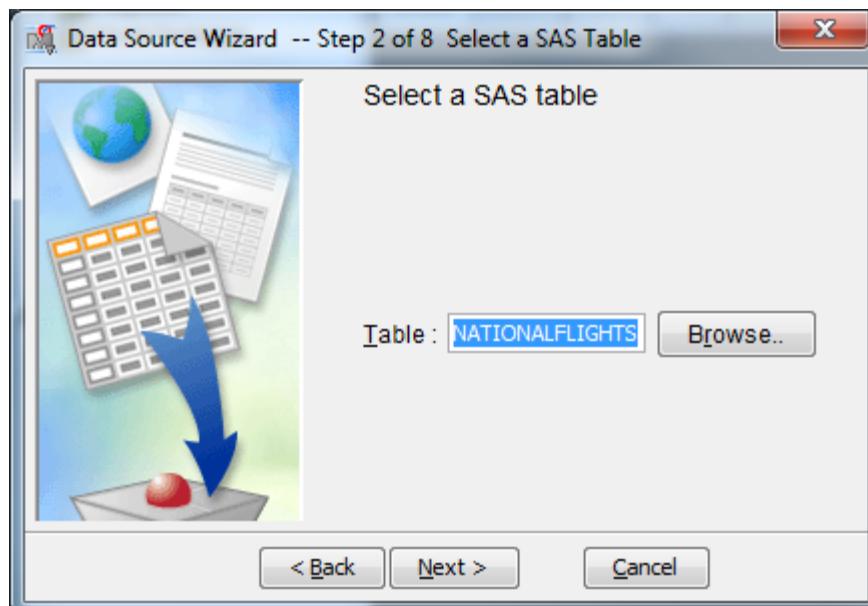
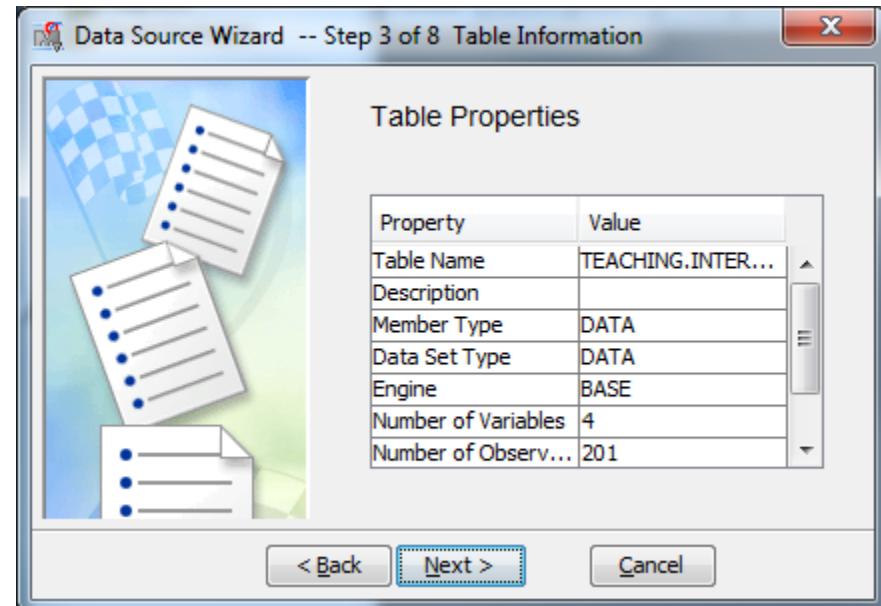
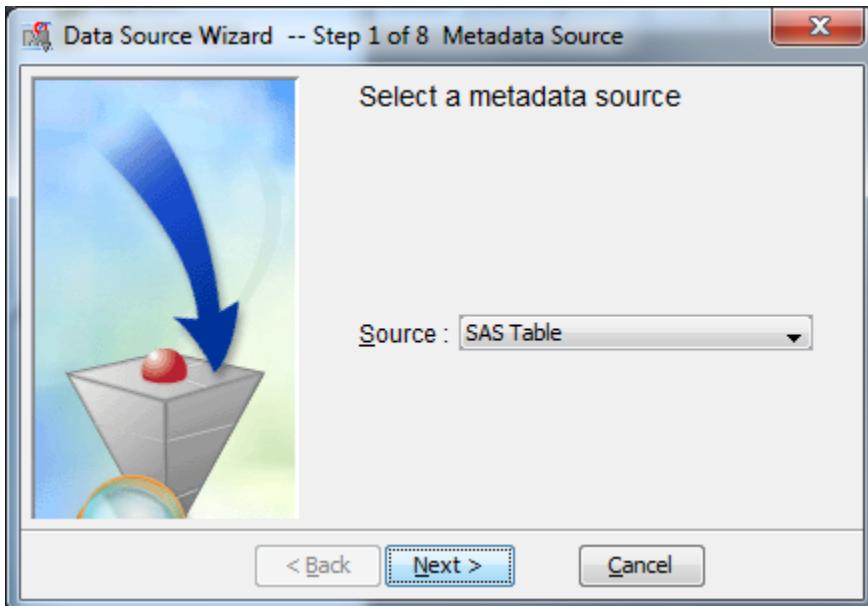
Options

< Back Next > Cancel

Creating a new data source

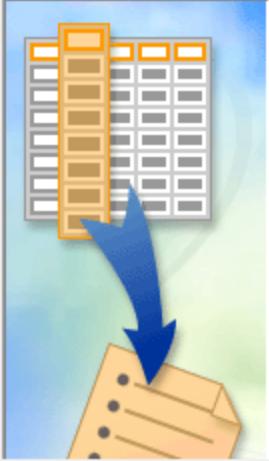


8 steps to create a new data source



8 steps to create a new data source

Data Source Wizard -- Step 5 of 8 Column Metadata



Name	Role	Level	Report
Boarded	Input	Interval	No
Date	Time ID	Interval	No
Destination	Input	Nominal	No
FlightNumber	Input	Nominal	No

Show code Explore Compute Summary < Back

Data Source Wizard -- Step 7 of 8 Data Source Attributes



You may change the name and the role, and can specify a population segment identifier for the data source to be created.

Name : INTERNATIONALFLIGHTS
Role : Raw
Segment :
Notes :

< Back Next > Cancel

Data Source Wizard -- Step 6 of 8 Create Sample



Do you wish to create a sample data set?

No Yes

Table Info

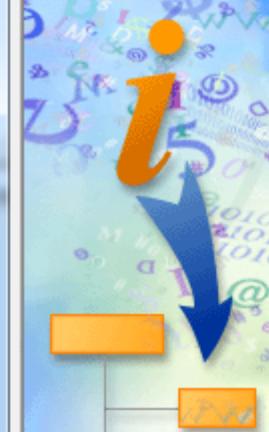
Columns 4
Rows 201

Sample Size

Type Percent
Percent 10

< Back Next > Cancel

Data Source Wizard -- Step 8 of 8 Summary



Metadata Completed.

Library:	TEACHING
Data Source:	INTERNATIONALFLIGHTS
Role:	Raw
Input	Level Interval
Input	Level Nominal
Time ID	Level Interval

< Back Finish Cancel

Now the data source is created

The screenshot shows the Enterprise Miner interface with a yellow arrow pointing to the 'INTERNATIONALFLIGHTS' entry in the 'Data Sources' tree view.

Data Sources Tree View:

- NewPro
- Data Sources
 - INTERNATIONALFLIGHTS**
- Diagrams
- Model Packages

Toolbar:

- Sample
- Explore
- Modify
- Model
- Assess
- Utility
- Applications
- Text Mining
- Time Series

Properties View (Left Panel):

Property	Value
ID	internationalflights
Name	INTERNATION
Variables	[...]
Decisions	[...]
Role	Raw
Notes	[...]
Library	TEACHING
Table	INTERNATION

Description View (Bottom Left Panel):

ID

Data Source identifier. The metadata tables associated with the data source are stored in the EMDS SAS library and use this identifier as the prefix for naming these tables.

Status Bar:

xy44 as xy44 Connected to en-ss-sxy44-a

You can add more data sources as needed

The screenshot shows the Enterprise Miner - NewPro application window. On the left, the 'Data Sources' pane lists 'INTERNATIONALFLIGHTS' and 'WORKLOADS'. A yellow arrow points from the text 'Here the WORKLOADS dataset is added' to the 'WORKLOADS' entry. The central workspace has a toolbar with icons for Sample, Explore, Modify, Model, Assess, Utility, Applications, Text Mining, and Time Series. Below the toolbar is a large, dark gray workspace area. The bottom right corner shows a status bar with a user icon, the text 'xy44 as xy44', and the message 'Connected to en-ss-sxy44-a'.

File Edit View Actions Options Window Help

NewPro

Data Sources
INTERNATIONALFLIGHTS
WORKLOADS

Diagrams
Model Packages

Sample Explore Modify Model Assess Utility Applications Text Mining Time Series

.. Property Value

ID	workloads
Name	WORKLOADS
Variables	[...]
Decisions	[...]
Role	Raw
Notes	[...]
Library	TEACHING
Table	WORKLOADS

ID

Data Source identifier. The metadata tables associated with the data source are stored in the EMDS SAS library and use this identifier as the prefix for naming these tables.

xy44 as xy44 Connected to en-ss-sxy44-a

Here the WORKLOADS dataset is added

Edit variables if needed

The screenshot shows the Enterprise Miner - NewPro interface. In the left sidebar, under 'Data Sources', the 'WORKLOADS' item is selected. A context menu is open over this item, listing options: Rename, Duplicate, Delete, Edit Variables..., Refresh Metadata, Edit Decisions..., and Explore... The 'Edit Variables...' option is highlighted with a red rectangular box. A large yellow arrow points from the bottom-left towards this red box. On the right side of the interface, there is a detailed description of the 'ID' field, which is the Data Source identifier.

ID

Data Source identifier. The metadata tables associated with the data source are stored in the EMDS SAS library and use this identifier as the prefix for naming these tables.

xy44 as xy44 Connected to en-ss-sxy44-a

Edit variables if needed

Variables - WORKLOADS

(none) not Equal to ... Apply Reset

Columns: Label Mining Basic Statistics

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	.	.

Explore... Edit Using SAS Code OK Cancel

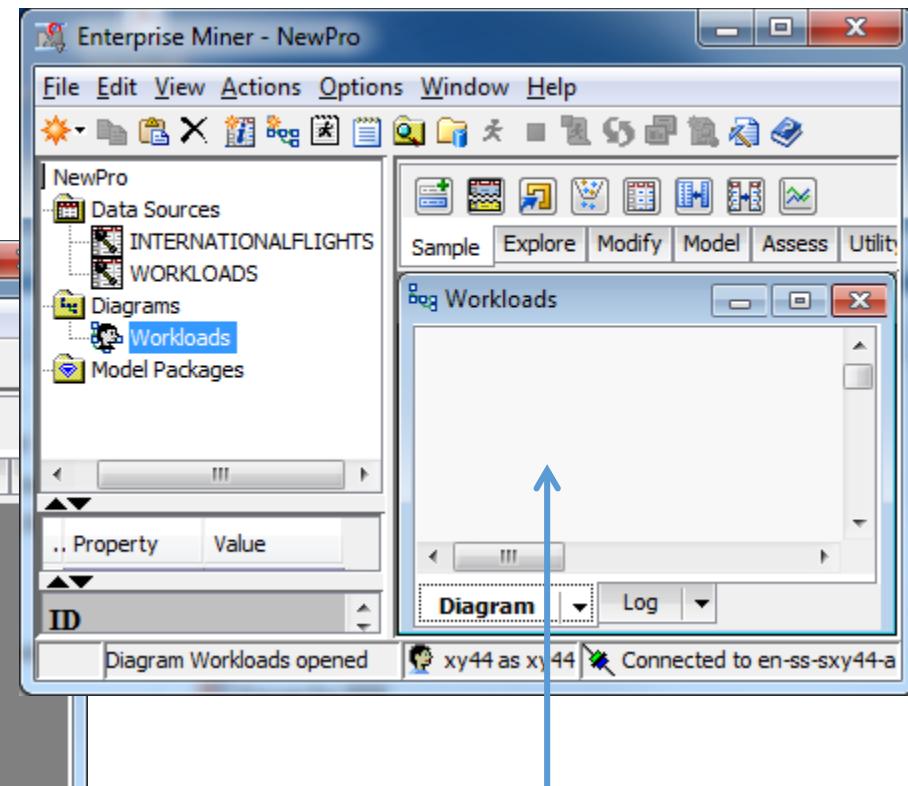
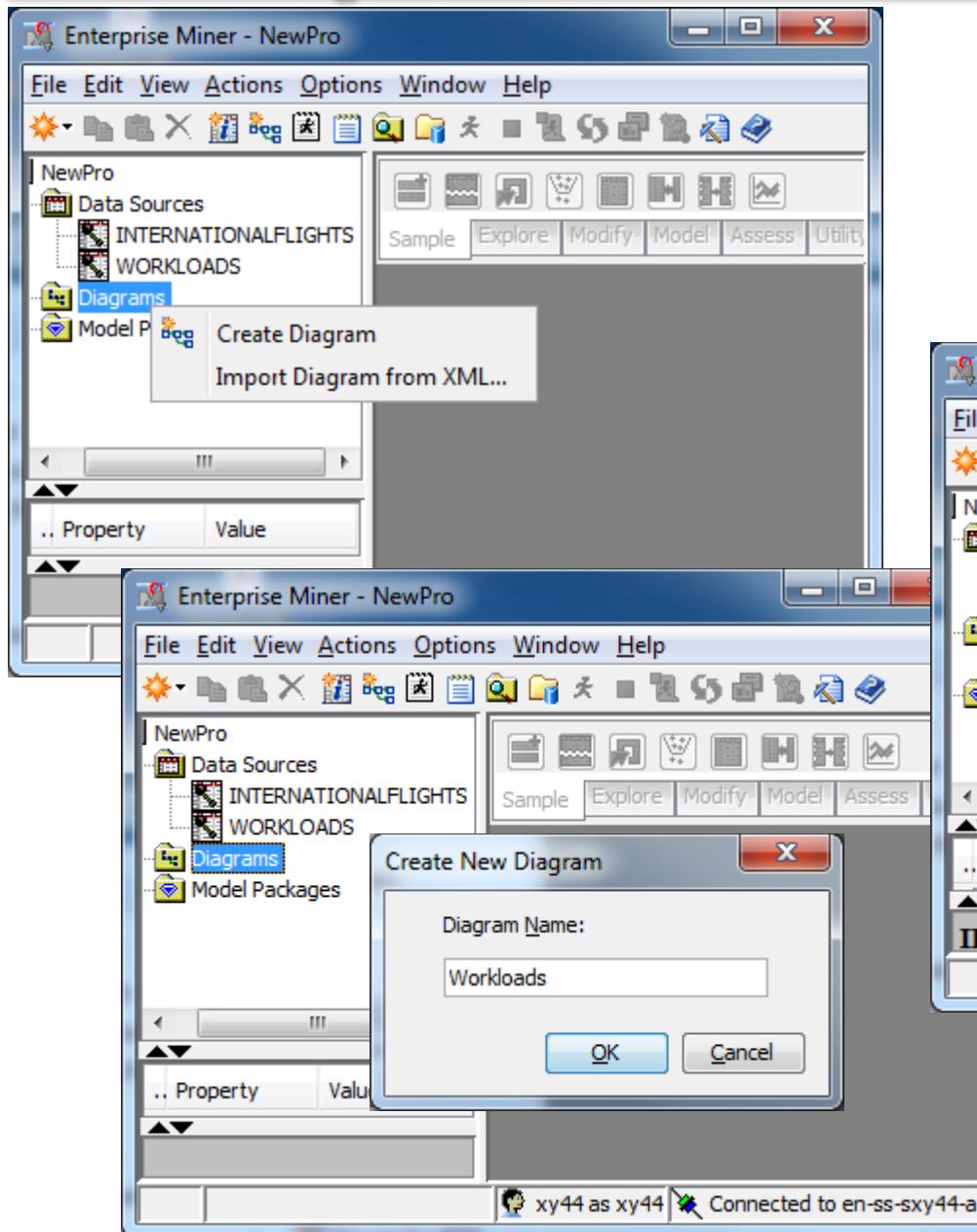
Edit variables if needed

The screenshot shows the 'Variables - WORKLOADS' dialog box. At the top, there are filter options: '(none)', 'not', 'Equal to', and a search field. Below these are checkboxes for 'Label', 'Mining', 'Basic', and 'Statistics'. The main area is a grid with columns: Name, Role, Level, Report, Order, Drop, Lower Limit, and Upper Limit. Rows represent various variables: Administration, Faculty_, Large_Lecture, Number_of_Cou, Preparations, Rank, Sabbatical, Student_Superv, Year, _Calculus_Cou, _Courses, _Instruction, Prof_Activity, and Service. The 'Role' column for '_Courses' is highlighted with a red box and a red arrow points to it from the explanatory text below.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Administration	Input	Interval	No		No	.	.
Faculty_	Input	Interval	No		No	.	.
Large_Lecture	Rejected	Interval	No		No	.	.
Number_of_Cou	Residual	Interval	No		No	.	.
Preparations	Segment	Interval	No		No	.	.
Rank	Sequence	Nominal	No		No	.	.
Sabbatical	Target	Nominal	No		No	.	.
Student_Superv	Text	Interval	No		No	.	.
Year	Text Locati	Nominal	No		No	.	.
_Calculus_Cou	Time ID	Interval	No		No	.	.
_Courses	Input	Interval	No		No	.	.
_Instruction	Input	Interval	No		No	.	.
Prof_Activity	Input	Interval	No		No	.	.
Service	Input	Interval	No		No	.	.

Click any cell to make a change

Create a diagram



This area becomes white (active)

Create a diagram: Input data node and StatExplore

Enterprise Miner - NewPro

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series

Workloads

WORKLOADS → StatExplore

Property Value

ID	EMWS2
Name	Workloads
Status	Open
Notes	
History	
Create Date	11/16/17 4:58 PM
Encoding	wlatin1 Western (Windows)
Data Representation	WINDOWS_64
Native OS	Yes

ID

Diagram Workloads opened

xy44 as xy44 Connected to EN-SS-AXY44-G

Run the program/procedure

Enterprise Miner - NewPro

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series

Workloads

WORKLOADS → StatExplore

Context menu (Run highlighted):

- Edit Variables...
- Update
- Run**
- Create Model Package...
- Results...
- Export Path as SAS Program
- Cut
- Copy
- Delete
- Rename
- Select All
- Select Nodes
- Connect Nodes
- Disconnect Nodes

Diagram Log

Diagram Workloads opened

xy44 as xy44 Connected to EN-SS-AXY44-G

```
graph LR; WORKLOADS[WORKLOADS] --> StatExplore[StatExplore]
```

Run the program/procedure

Enterprise Miner - NewPro

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series

Workloads

```
graph LR; WORKLOADS[WORKLOADS] --> StatExplore[StatExplore]
```

Property Value

General

Node ID	Stat
Imported Data	[...]
Exported Data	[...]
Notes	[...]

Train

Variables	[...]
-----------	-------

Data

Number of Observation	100000
Validation	No
Test	No

Standard Reports

Interval Distributions	Yes
Class Distributions	Yes
Level Summary	Yes

General

Running StatExplore

xy44 as xy44 Connected to EN-SS-AXY44-G

The screenshot shows the SAS Enterprise Miner interface with a project named 'NewPro'. The left sidebar contains 'Data Sources' (INTERNATIONALFLIGHTS, WORKLOADS), 'Diagrams' (Workloads selected), and 'Model Packages'. The main workspace displays a 'Workloads' diagram with a 'WORKLOADS' source node connected to a 'StatExplore' target node. The 'Properties' pane on the left provides detailed settings for the 'StatExplore' node, including training variables, data sizes, and reporting preferences. The bottom status bar indicates the process is 'Running StatExplore'.

Run the program/procedure → get the Results...

Enterprise Miner - NewPro

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series

Workloads

WORKLOADS StatExplore

Run Status

Run completed
Diagram: Workloads
Path: StatExplore

OK Results... Click it

Number of Observation: 100000

Validation: No

Test: No

Interval Distributions: Yes

Class Distributions: Yes

Level Summary: Yes

General

Running StatExplore

xy44 as xy44 Connected to EN-SS-AXY44-G

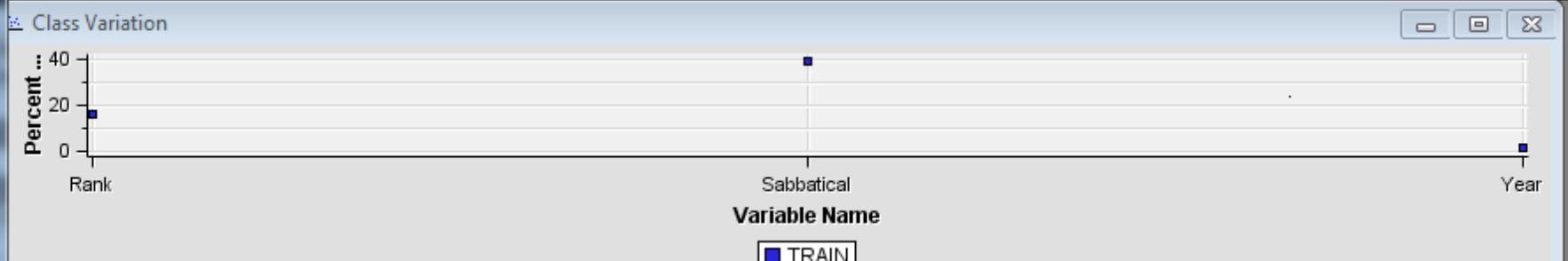
The screenshot shows the SAS Enterprise Miner interface. On the left, there's a tree view of the project structure under 'NewPro' with nodes like 'INTERNATIONALFLIGHTS', 'WORKLOADS', 'Diagrams', and 'Model Packages'. Below the tree is a 'Property' panel with sections for 'General', 'Train', 'Data', 'Standard Reports', and 'General'. The 'Data' section shows 'Number of Observation' as 100000, 'Validation' as 'No', and 'Test' as 'No'. The 'Standard Reports' section has 'Interval Distributions', 'Class Distributions', and 'Level Summary' all set to 'Yes'. A 'Diagram' tab is selected at the bottom.

The main workspace shows a workflow diagram with a 'WORKLOADS' node connected to a 'StatExplore' node. A 'Run Status' dialog box is open, indicating a successful run completion for the 'StatExplore' path. The 'Results...' button in the dialog is circled in red with a callout pointing to it, accompanied by the text 'Click it'.

At the bottom right, there's a status bar with the text 'xy44 as xy44 Connected to EN-SS-AXY44-G'.

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

SAS Results

Scoring

Summary Statistics

Plots

Table

Plot...

Interval Variables

Class Variables

Cell Chi-Squares

Variable Summary

Measurement Frequency

Role Level Count

INPUT INTERVAL 11

INPUT NOMINAL 3

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Variable Number of

Class Variation

Percent .. 40

Rank

Sabbatical

Variable Name

TRAIN

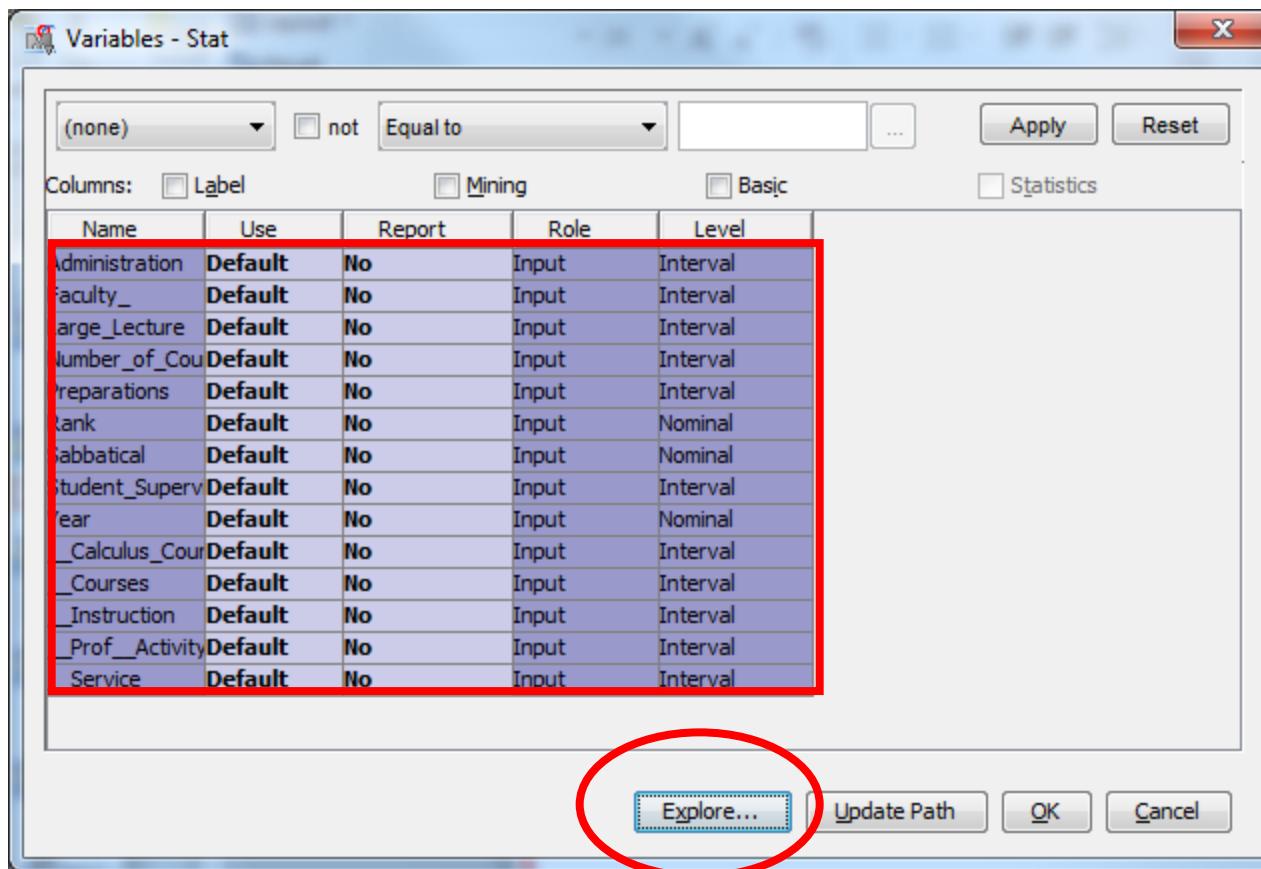
Interval Variables

Ordered Inputs	Data Role	Variable	Median	Missing
1TRAIN		Preparations	0	0
2TRAIN		Student_Su...	0	0
3TRAIN		Administrati...	0	0
4TRAIN		Large_Lect...	0	0
5TRAIN		_Calculus...	1	0
6TRAIN		_Service	15	0
7TRAIN		Faculty_	14	0
8TRAIN		_Prof_Ac...	32	0
9TRAIN		_Courses	45	0
10TRAIN		_Instruction	50	0
11TRAIN		Number_of...	4	0

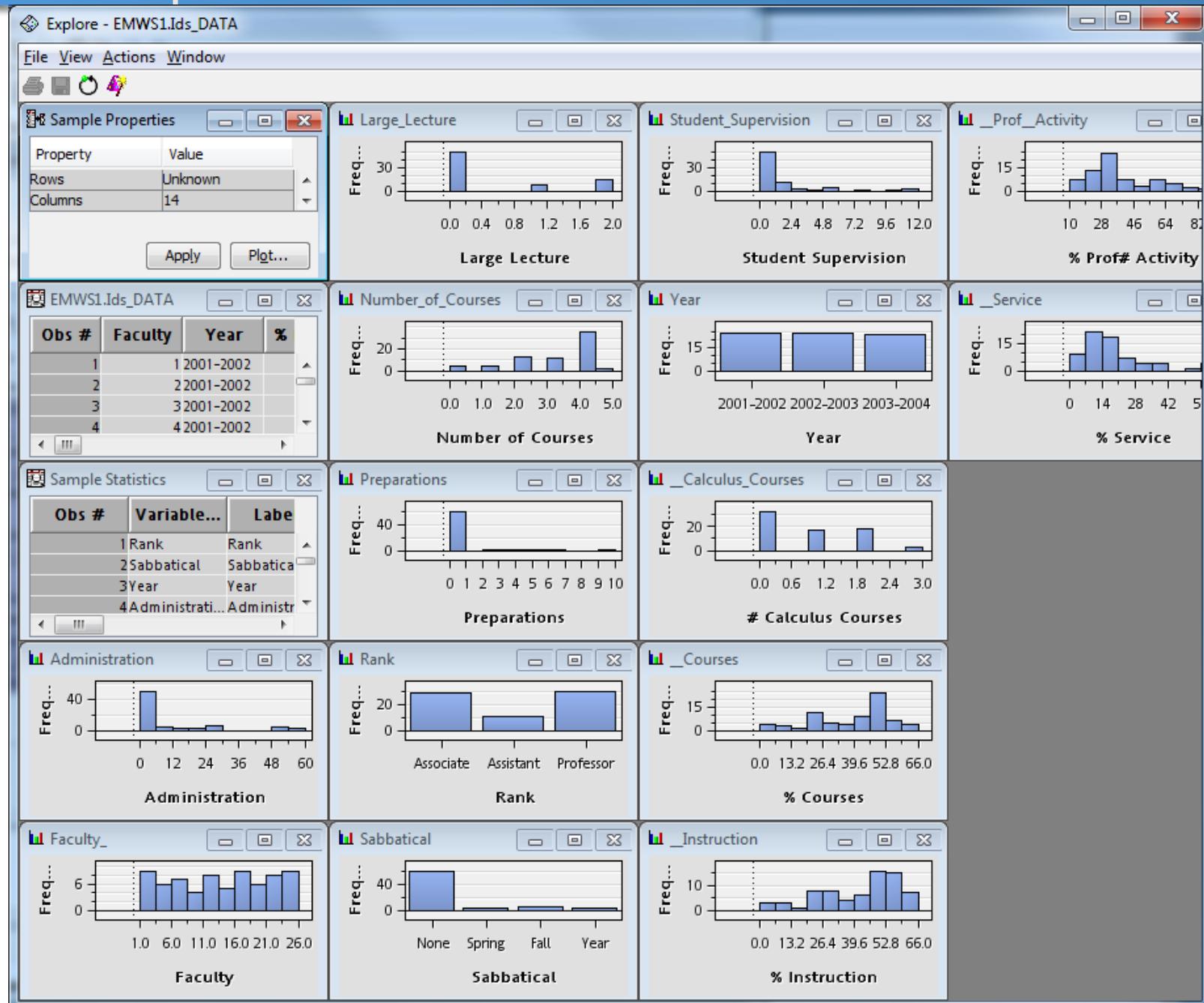
Edit/explore variables from the node

The screenshot shows the SAS Enterprise Miner interface with the title "Enterprise Miner - NewPro". The left sidebar displays project structure under "NewPro": Data Sources (INTERNATIONALFLIGHTS, WORKLOADS), Diagrams (Workloads), and Model Packages. The main workspace shows a flow diagram with a "WORKLOADS" node connected to a "StatExpl..." node. A context menu is open on the "StatExpl..." node, with the "Edit Variables..." option highlighted by a red box and arrow. The menu also includes options like Update, Run, Create Model Package..., Results..., Export Path as SAS Program, Cut, Copy, Delete, Rename, Select All, Select Nodes, Connect Nodes, and Disconnect Nodes. The bottom status bar indicates "Run completed".

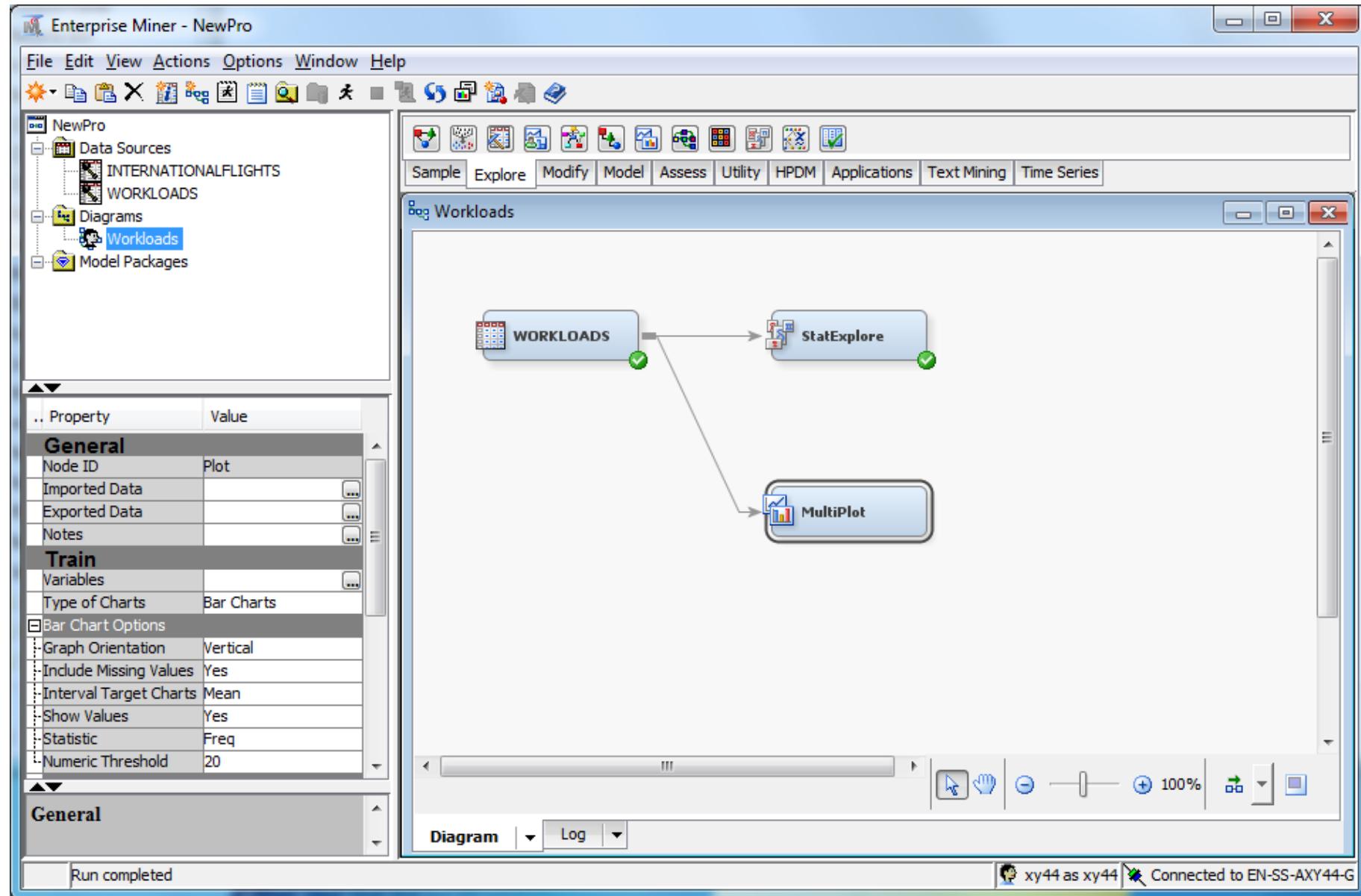
Highlight the variable(s) to explore



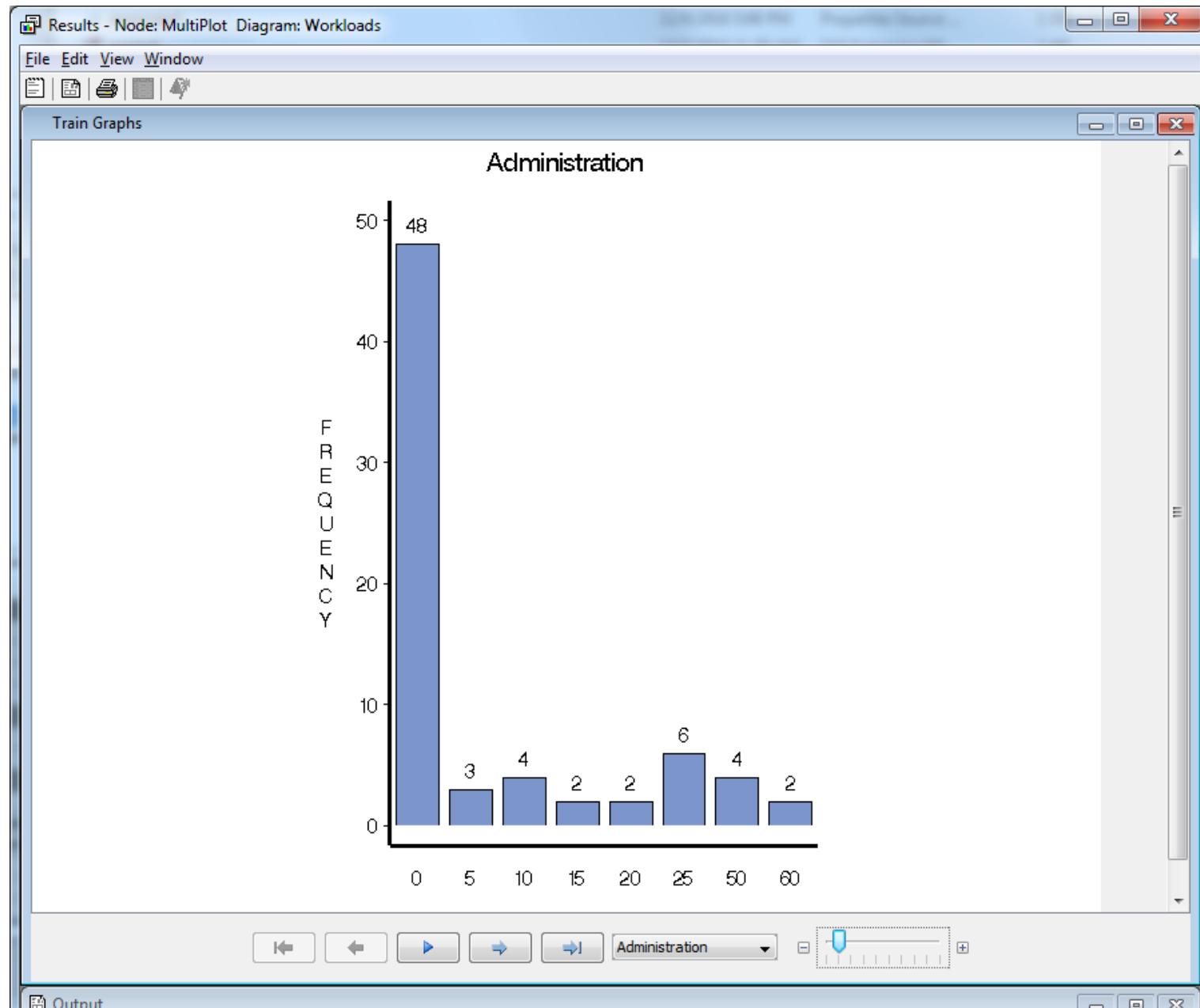
Explorer the plots



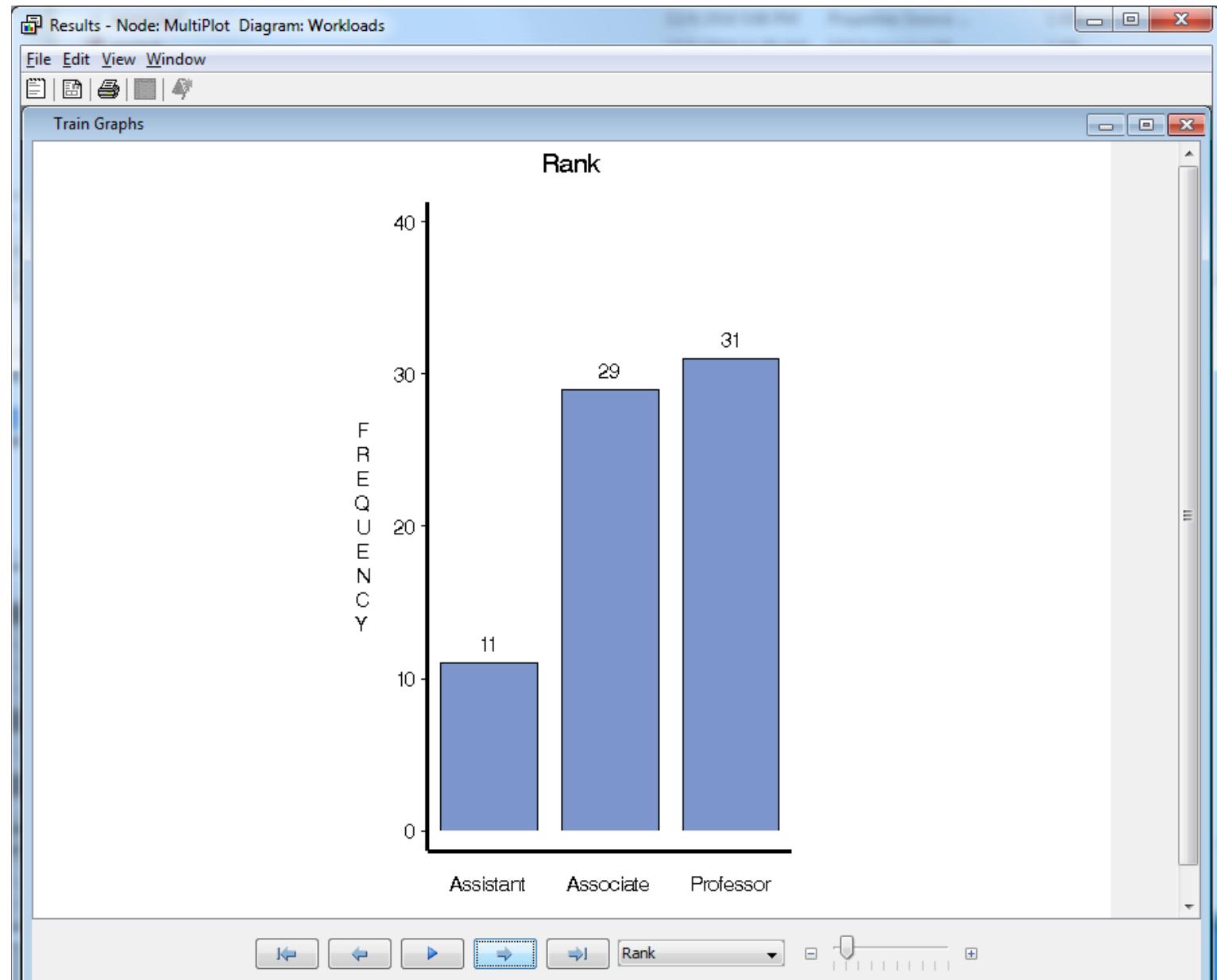
Add a multiplot node to examine the variables



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

Enterprise Miner - NewPro

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series

Workloads

1. Run the Node

2. Highlight “Imported Data”

3. Click the ellipsis Button

StatExplore

WORKLOADS

Data Partition

MultiPlot

General

Node ID	Part
Imported Data	EMWWS2.Ids_DATA
Exported Data	
Notes	

Train

Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345

Data Set Allocations

Training	40.0
----------	------

Imported Data - Data Partition

Port	Source	Table	Role	Data Exists
DATA		EMWWS2.Ids_DATA	Raw	Yes

Browse... Explore... Properties... OK

Exported Data - Data Partition

Port	Table	Role	Data Exists
TRAIN	EMWWS2.Part_TRAIN	Train	Yes
VALIDATE	EMWWS2.Part_VALI...	Validate	Yes
TEST	EMWWS2.Part_TEST	Test	Yes

Browse... Explore... Properties... OK

```
graph LR; WORKLOADS[WORKLOADS] --> DataPartition[Data Partition]; DataPartition --> StatExplore[StatExplore]; DataPartition --> MultiPlot1[MultiPlot]; DataPartition --> MultiPlot2[MultiPlot];
```

The partition results

The diagram illustrates the partitioning of a dataset into three parts: TRAIN, VALIDATE, and TEST. Blue arrows point from the main dataset to each of the three partitioned windows.

Explore - EMWS1.Ids_DATA

Property	Value
Rows	Unknown
Columns	14
Library	EMWS1
Member	IDS_DATA
Type	VIEW
Sample Method	Top
Fetch Size	Default
Fetched Rows	71
Random Seed	12345

Explore - EMWS1.Part_VALIDATE

Property	Value
Rows	21
Columns	15
Library	EMWS1
Member	PART_VALIDATE
Type	DATA
Sample Method	Top
Fetch Size	Default
Fetched Rows	21
Random Seed	12345

Explore - EMWS1.Part_TRAIN

Obs #	_dataobs_	Faculty	Year	% Instruction	% Courses
1	1	1	12001-2002	66	66
2	2	2	22001-2002	38	30

Explore - EMWS1.Part_TEST

Property	Value
Rows	22
Columns	15
Library	EMWS1
Member	PART_TEST
Type	DATA
Sample Method	Top
Fetch Size	Default
Fetched Rows	22
Random Seed	12345

Explore - EMWS1.Part_TEST

Obs #	_dataobs_	Faculty	Year	% Instruction	% Courses
1	9	9	92001-2002	66	66
2	11	11	12001-2002	56	54

Add more nodes: Drop, Filter & Transformation ...

Enterprise Miner - NewPro

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility Applications Text Mining Time Series

Beg Workloads

```
graph TD; WORKLOADS[WORKLOADS] --> DataPartition[Data Partition]; WORKLOADS --> StatExplore1[StatExplore]; WORKLOADS --> MultiPlot1[MultiPlot]; WORKLOADS --> Drop[Drop]; Drop --> Filter[Filter]; Drop --> Transform[Transform Variables]; Drop --> StatExplore2[StatExplore (2)]; Transform --> MultiPlot2[MultiPlot (2)]
```

.. Property Value

General

Node ID	Drop
Imported Data	
Exported Data	
Notes	

Train

Variables	
-----------	--

Drop Selection Options

Drop from Tables	No
Assess	No
Classification	No
Frequency	No
Hidden	Yes
Input	No
Predict	No
Rejected	Yes
Residual	No
Target	No
Other	No

Variables

Variable Properties

Run completed

yhm as yhm Connected to po yang

Drop: to remove variables not for further analysis
Filter: to remove outliers/extreme values (observ.)
Transform variables: to create new variables from the existing ones to stabilize variances, remove nonlinearity, ...

Impute: to replace missing values

Enterprise Miner - NewPro

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series

Workloads

Missing data: randomly scattered; $\leq 5\%$ of total number of values

```
graph LR; WORKLOADS[WORKLOADS] --> DataPartition[Data Partition]; WORKLOADS --> MultiPlot[MultiPlot]; WORKLOADS --> DecisionTree[Decision Tree]; DataPartition --> Impute[Impute]; StatExplore[StatExplore] --> DataPartition;
```

Diagram Log

Run completed

xy44 as xy44 Connected to EN-SS-AXY44-G

Integration of SAS codes into SAS EM 13.2.1

SAS Enterprise Miner - NewPro

File Edit View Actions Options Window Help

Utility tab selected (1)

Utility toolbar icon (2)

SAS Code node highlighted (3)

Code Editor button in Properties panel (4)

Diagram pane showing a workflow:

```
graph LR; WORKLOADS[WORKLOADS] --> DataPartition[Data Partition]; WORKLOADS --> MultiPlot[MultiPlot]; DataPartition --> Impute[Impute]; DataPartition --> DecisionTree[Decision Tree]; Impute --> SASCode[SAS Code];
```

Properties panel (left side) showing node properties:

Property	Value
General	
Node ID	EMCODE
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Code Editor	
Tool Type	Utility
Data Needed	No
Rerun	No
Use Priors	Yes
Score	
Advisor Type	Basic
General	

Run completed message at the bottom left.

Connected to EN-SS-AXY44-G at the bottom right.

Finding population distribution with kernal density estimation (kde)

Training Code - Code Node

File Edit Run View

Macro

Train

- Utility
- EM_REGISTER
- EM_REPORT
- EM_DATA2CODE
- EM_DECDATA
- EM_CHECKMACRO
- EM_CHECKSETINIT
- EM_ODSLISTON
- EM_ODSLISTOFF

Macros Macro Variables Variables

Training Code

```
proc kde data=EMWS1.Impt_TRAIN;
  univar Number_of_Courses/gridl=0 gridu=5 out=sasuser.kde_NOC bwm=1;
run;
```

Low grid limit Up grid limit Bandwidth multipliers

Output Log Result Log

1
2

yhm as yhm - NewPro - Workloads - EMCODE - COMPLETE

```
proc kde data=EMWS1.Impt_TRAIN;
  univar Number_of_Courses/gridl=0 gridu=5 out=sasuser.kde_NOC bwm=1;
run;
```

Enter the result into the project by creating a data source

Enterprise Miner - NewPro

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility Applications Text Mining T

NewPro

Data Sources

INTER WORKLOADS Create Data Source

Diagrams

Workloads

Model Packages

Property Value

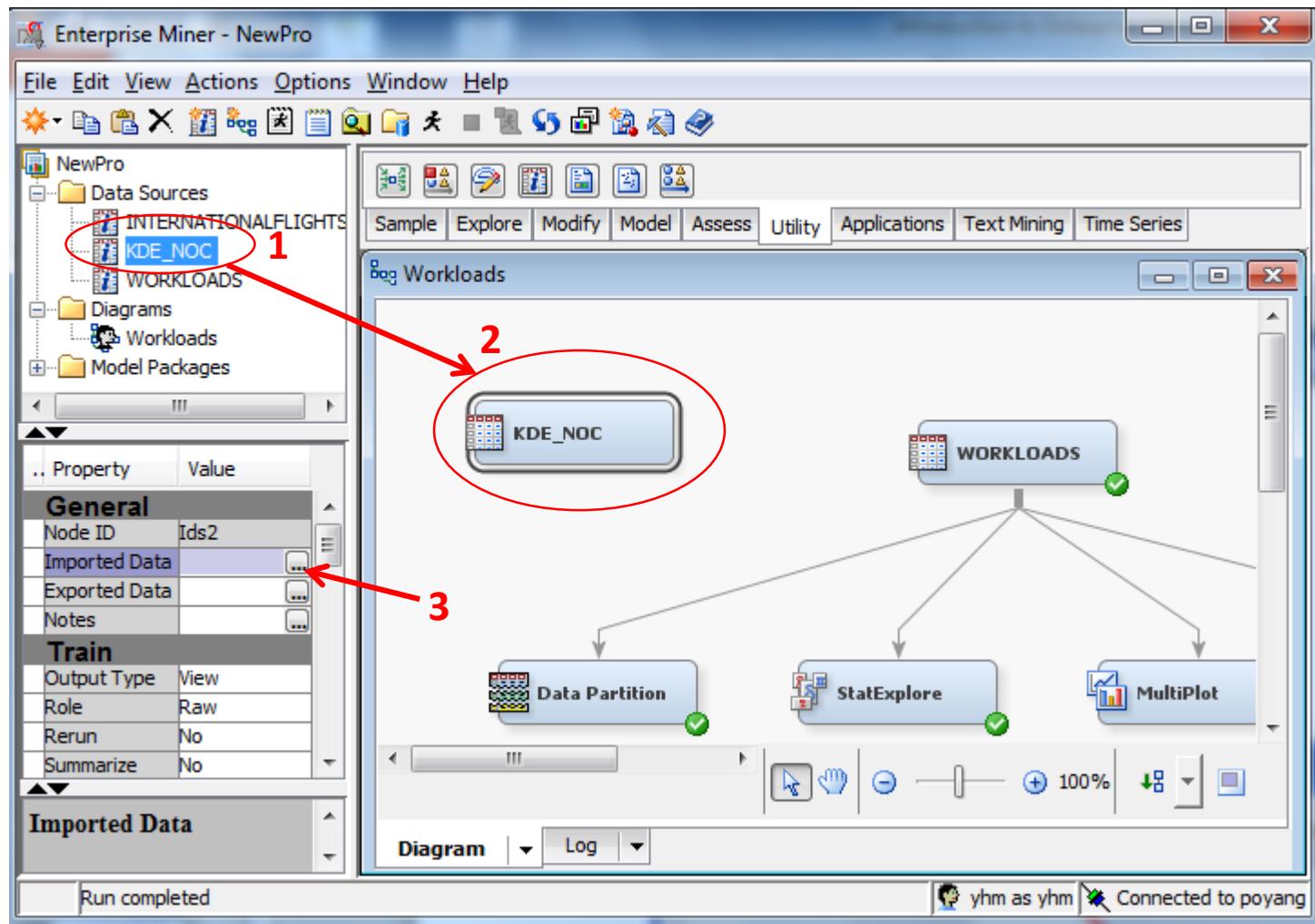
WORKLOADS

Data Partition

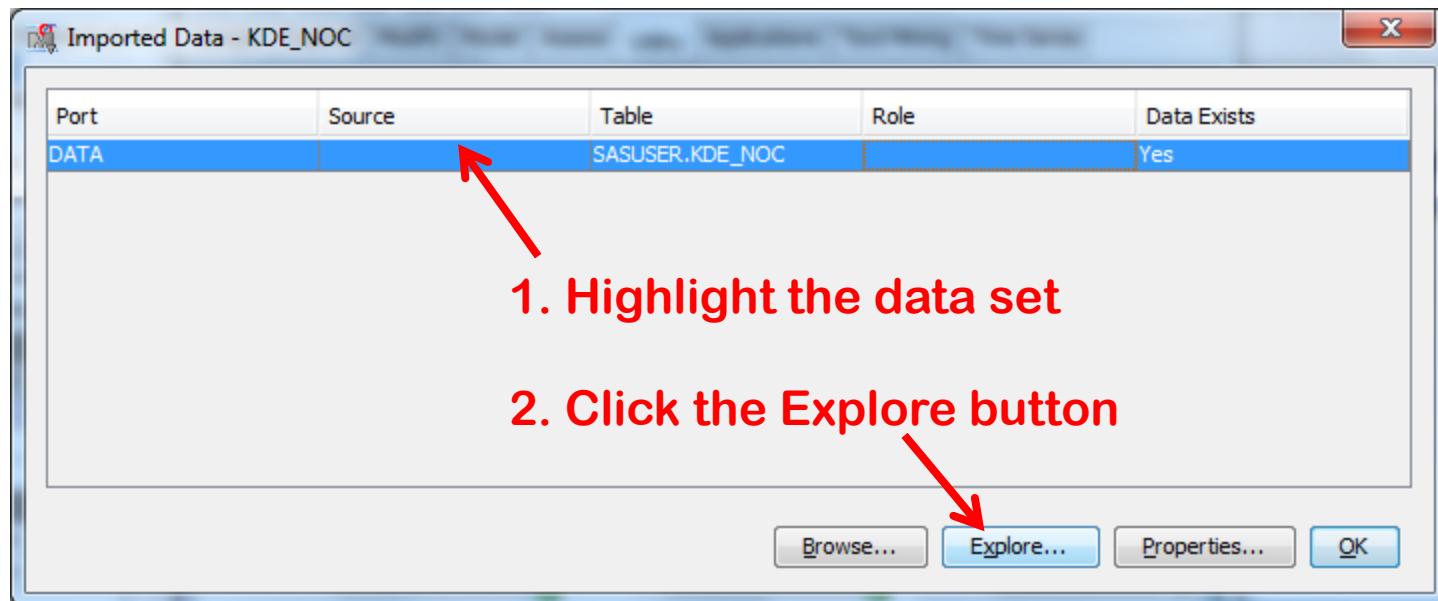
StatExplore

The screenshot shows the SAS Enterprise Miner interface with a blue header bar containing the title 'Enterprise Miner - NewPro' and various menu options. Below the header is a toolbar with icons for file operations like Open, Save, and Print, and analytical tools like Sample, Explore, and Modify. The main workspace is divided into two main sections. On the left, there's a tree view of the project structure under 'NewPro': 'Data Sources' (which is highlighted with a red box), 'INTER', 'WORKLOADS' (with a 'Create Data Source' button next to it), 'Diagrams' (containing 'Workloads'), and 'Model Packages'. At the bottom of this pane is a table with columns 'Property' and 'Value'. On the right, a workflow diagram is displayed. It starts with a node labeled 'WORKLOADS' at the top, which branches down to two other nodes: 'Data Partition' and 'StatExplore'. Both of these nodes have green checkmarks indicating they are active or successful. The entire workflow diagram is also enclosed in a red box.

Create a node with the new data source



Explore data through the Imported Data property



Plot a line-connected scatter plot

Explore - SASUSER.KDE_NOC

File View Actions Window

Sample Properties

Property	Value
Rows	401
Columns	4
Library	SASUSER
Member	KDE_NOC
Type	DATA
Sample Method	Top
Fetch Size	Default
Fetched Rows	401
Random Seed	12345

1

2

3

4

5

6

Select a Chart Type

Scatter

Line

Histogram

Density

Box

Tables

Matrix

Lattice

Parallel Axis

Constellation

Plots values of two variables against each other and connects data values with a line.

Cancel < Back Next > Finish

SASUSER.KDE_NOC

Select Chart Roles

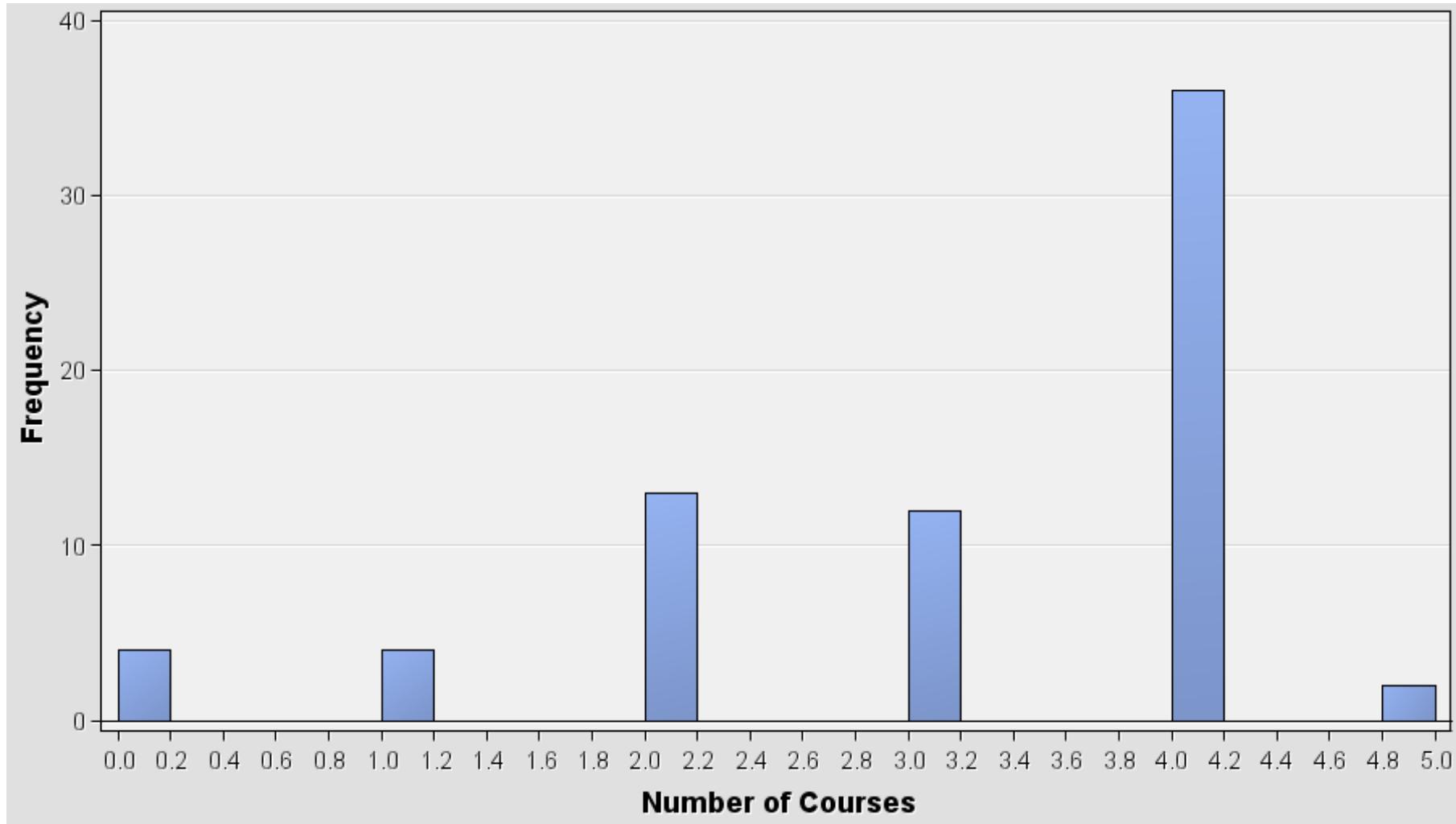
Variable	Role	Type	Description	Format
count		Numeric	Count	
density	Y	Numeric	Density	
value	X	Numeric	Value	
var		Character	Variable	

Use default assignments

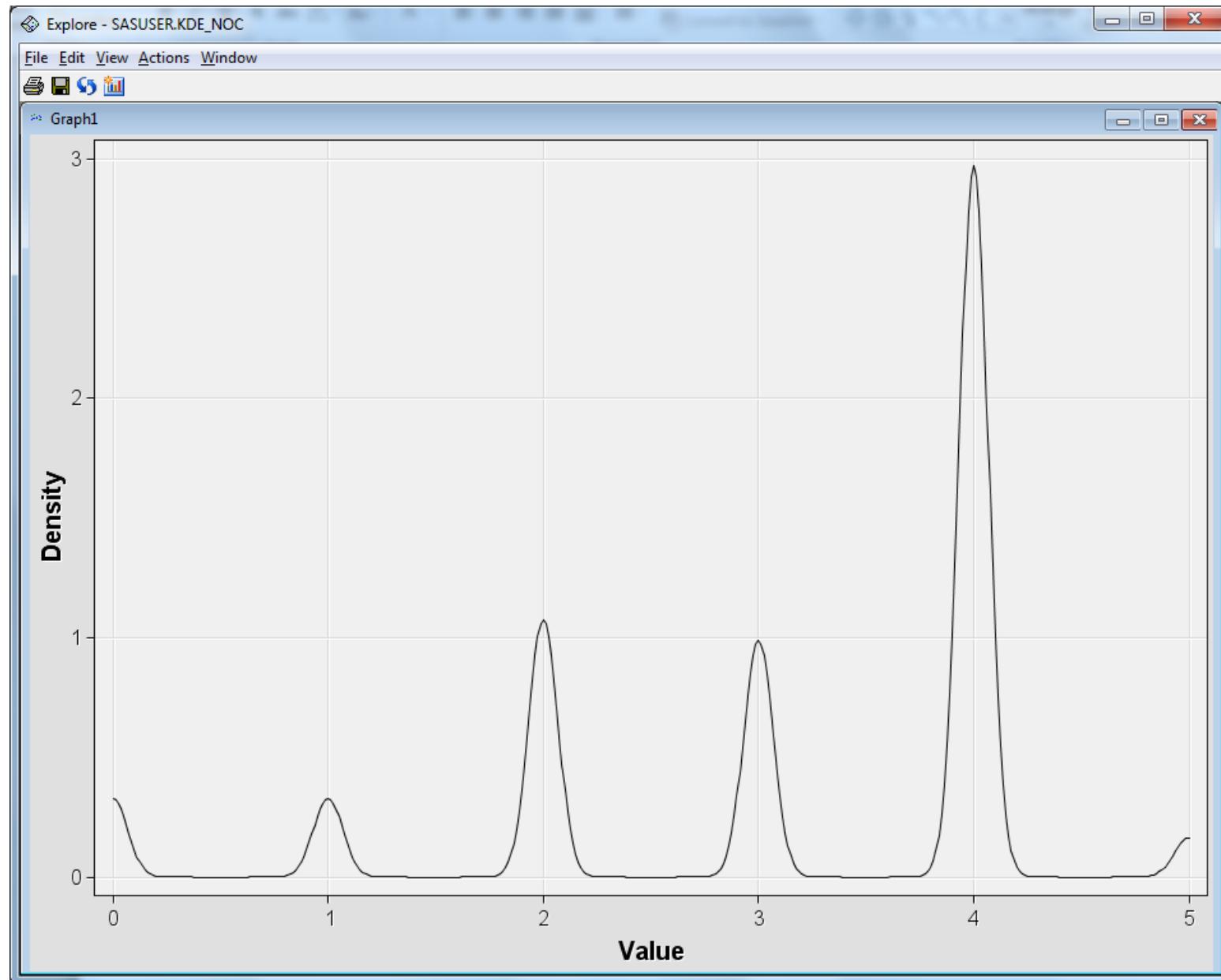
Allow multiple role assignments

Cancel < Back Next > Finish

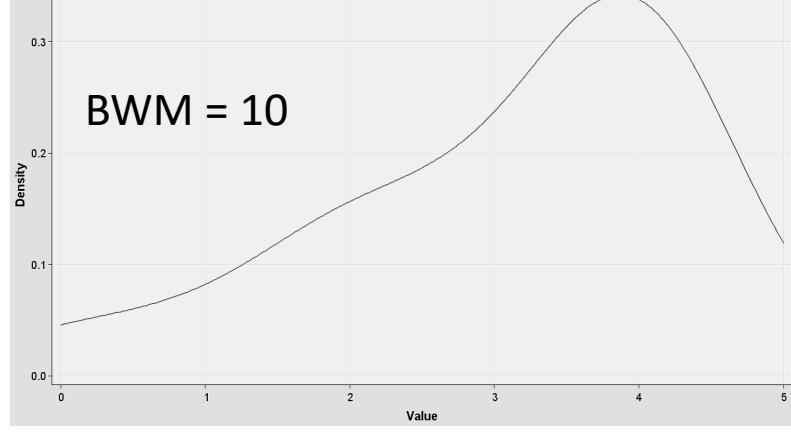
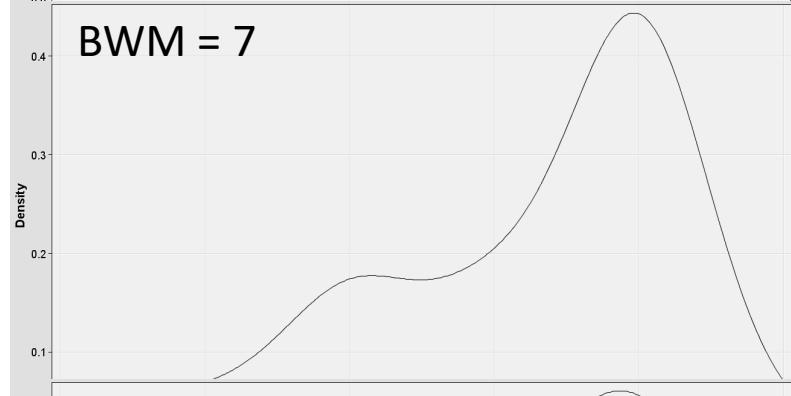
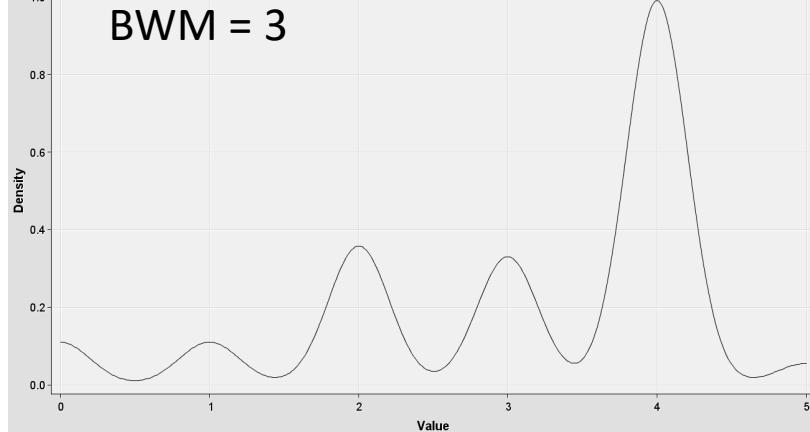
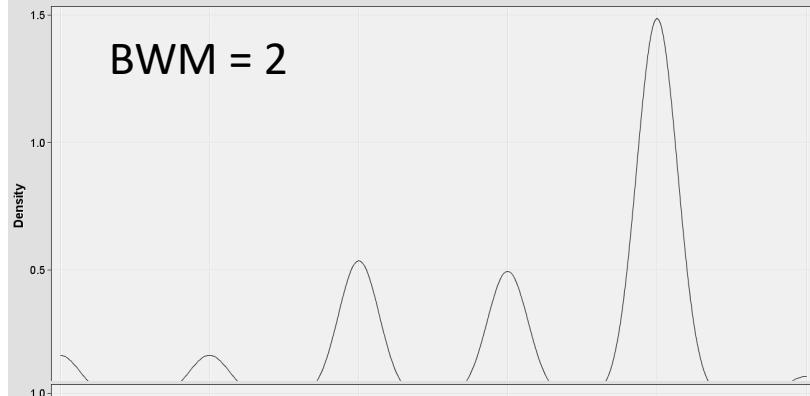
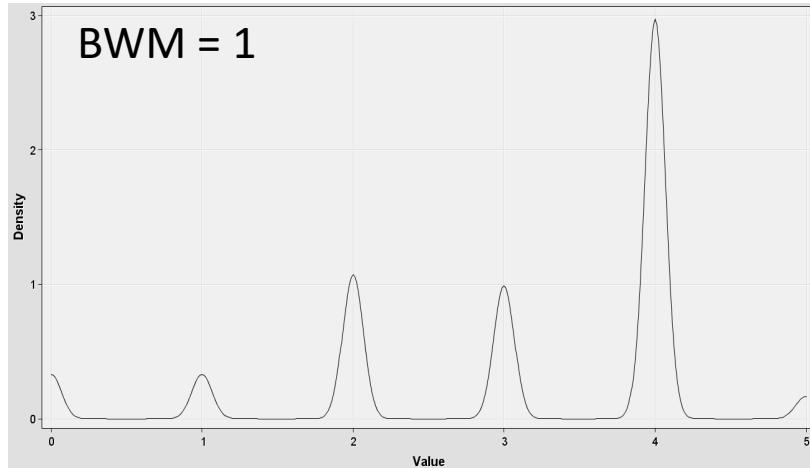
Bar chart of number of courses before kde



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;
```

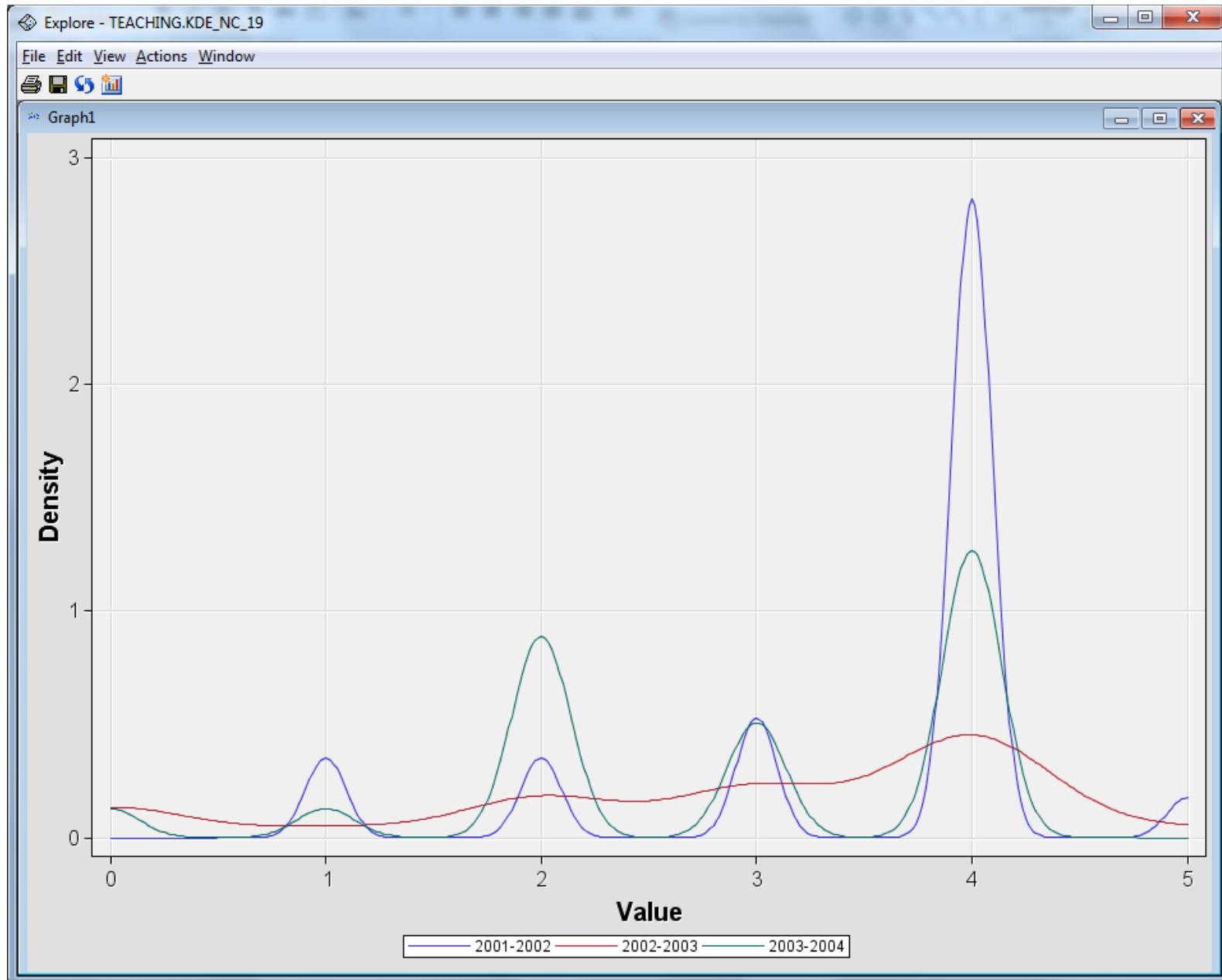
Determine the chart roles

Select Chart Roles X

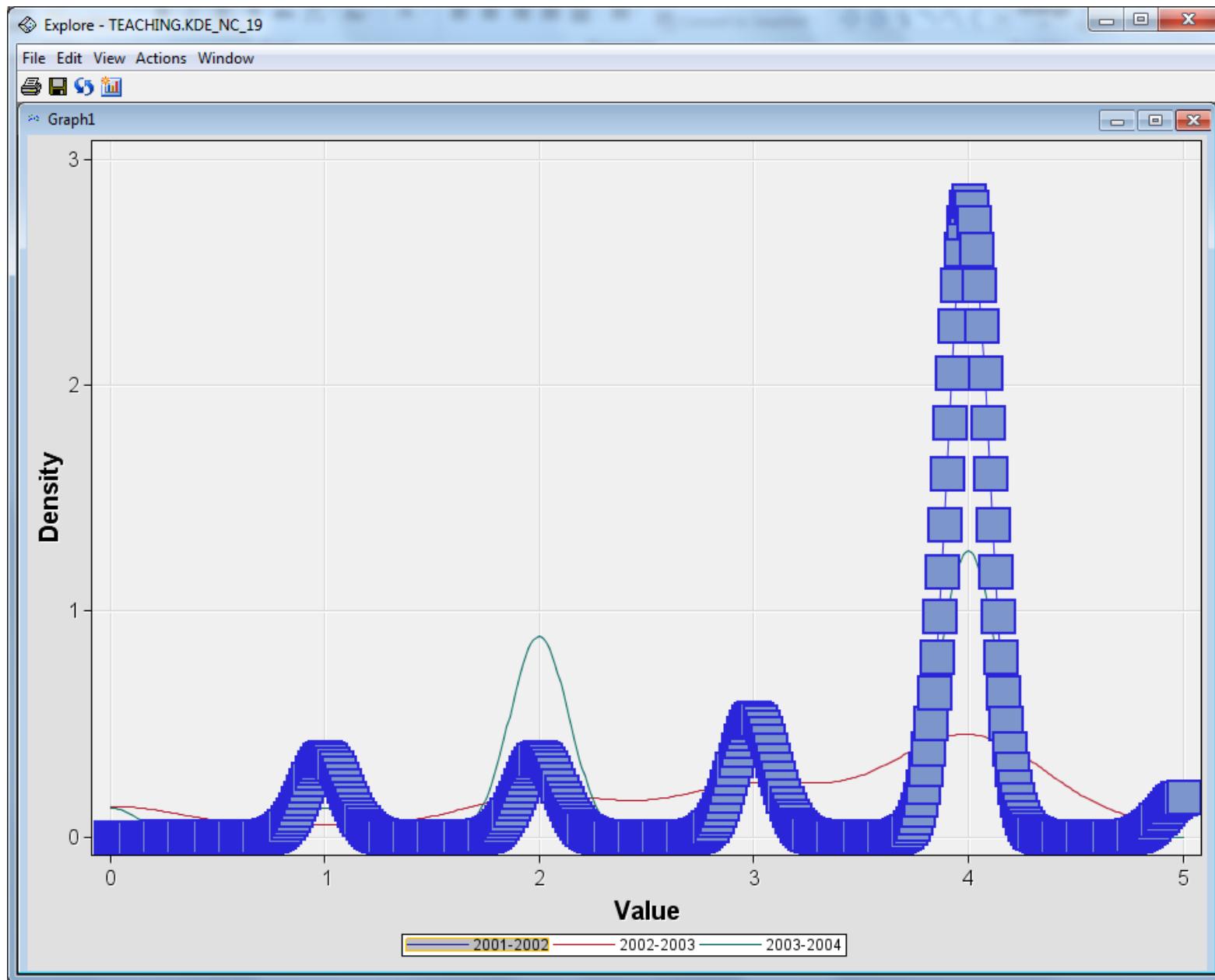
▲ Variable	Role	Type	Description	Format
count		Numeric	Count	
density	Y	Numeric	Density	
value	X	Numeric	Value	
var		Character	Variable	
Year	Group	Character	Year	\$9

Allow multiple role assignments

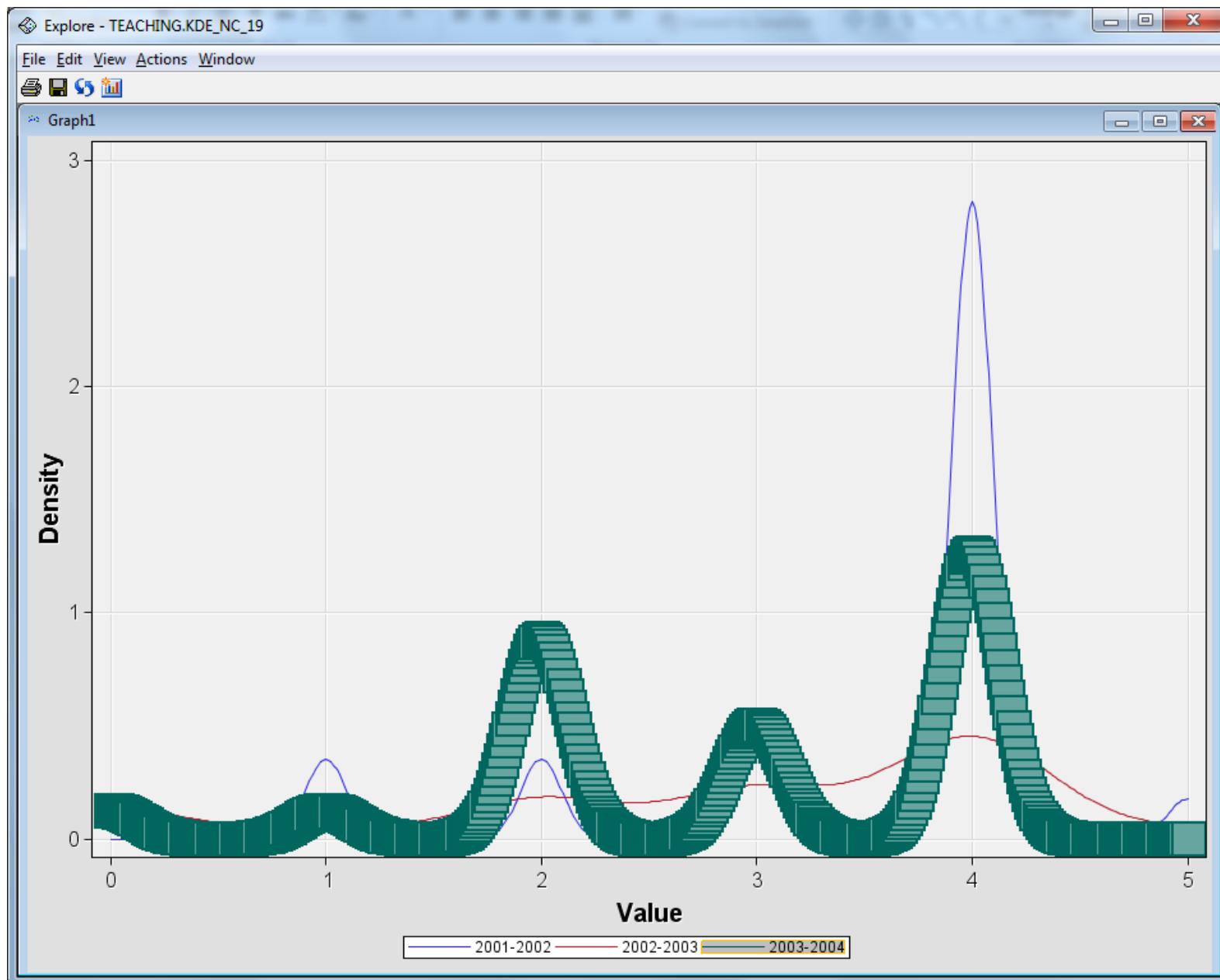
Overlay of kde graphs of number of courses by year



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...