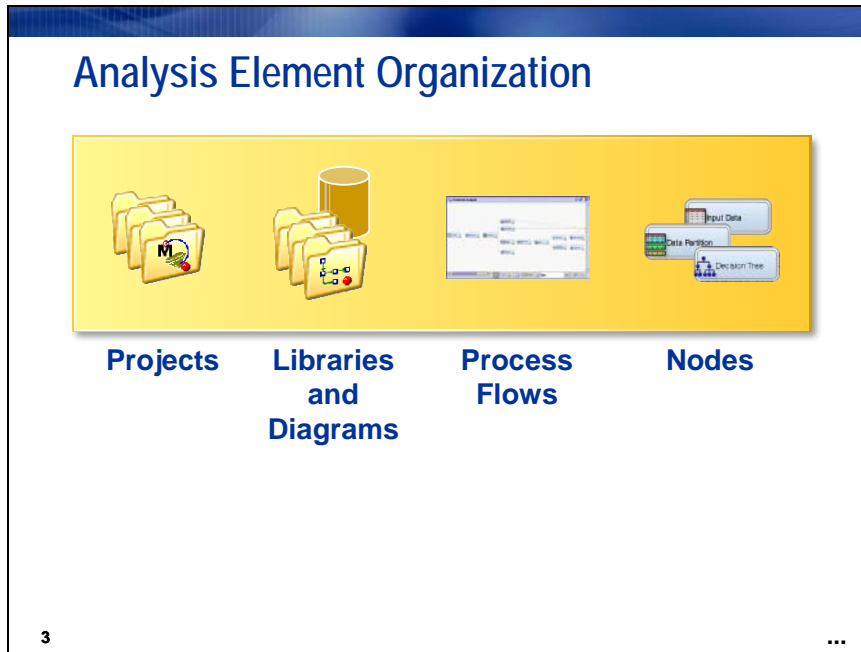


Chapter 2 Accessing and Assaying Prepared Data

0.1	Introduction.....	Error! Bookmark not defined.
0.2	A Section Title.....	Error! Bookmark not defined.
	Demonstration: <Type title of demo here.>.....	Error! Bookmark not defined.
	Exercises	Error! Bookmark not defined.
0.3	Chapter Summary.....	Error! Bookmark not defined.
0.4	Solutions	Error! Bookmark not defined.
	Solutions to Exercises	Error! Bookmark not defined.
	Solutions to Student Activities (Polls/Quizzes)	Error! Bookmark not defined.

2.1 Introduction



Analyses in SAS Enterprise Miner start by defining a project. A *project* is a container for a set of related analyses. After a project is defined, you define libraries and diagrams. A *library* is a collection of data sources accessible by the SAS Foundation Server. A *diagram* holds analyses for one or more data sources. Process flows are the specific steps you use in your analysis of a data source. Each step in a process flow is indicated by a node. The *nodes* represent the SAS Enterprise Miner tools that are available to the user.



A core set of tools is available to all SAS Enterprise Miner users. Additional tools can be added by licensing additional SAS products or by creating SAS Enterprise Miner extensions.

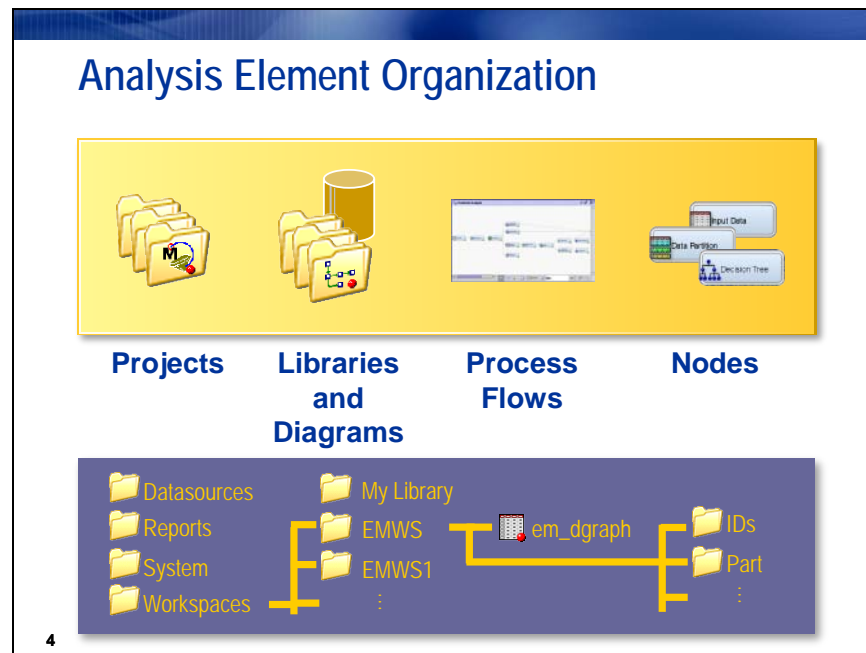
Before you can create a process flow, you need to define the following items:

- a SAS Enterprise Miner project to contain the data sources, diagrams, and model packages
- one or more SAS libraries inside your project, linking SAS Enterprise Miner to analysis data
- a diagram workspace to display the steps of your analysis

A SAS administrator can define these for you in advance of using SAS Enterprise Miner or you can define them yourself.



The demonstrations show you how to create the items in the list above.



Behind the scenes (on the SAS Foundation Server where the project resides), the organization is more complicated. Fortunately, the SAS Enterprise Miner client shields you from this complexity.

At the top of the hierarchy is the SAS Enterprise Miner project, corresponding to a directory on (or accessible by) the SAS Foundation. When a project is defined, four subdirectories are automatically created within the project directory: Datasources, Reports, System, and Workspaces. The SAS Enterprise Miner client via the SAS Foundation handles the writing, reading, and maintenance of these directories.



In both the personal workstation and SAS Enterprise Miner client configurations, all references to computing resources such as LIBNAME statements and access to SAS Foundation technologies must be made from the selected SAS Foundation Server's perspective. This is important because data and directories that are visible to the client machine might not be visible to the server.

Projects are defined to contain diagrams, the next level of the SAS Enterprise Miner organizational hierarchy. Diagrams usually pertain to a single analysis theme or project. When a diagram is defined, a new subdirectory is created in the Workspaces directory of the corresponding project. Each diagram is independent, and no information can be passed from one diagram to another.

Libraries are defined using a LIBNAME statement. You can create a library using the SAS Enterprise Miner Library Wizard, using SAS Management Console, or by using the Start-Up Code window when you define a project. Any data source compatible with a LIBNAME statement can provide data for SAS Enterprise Miner.

Specific analyses in SAS Enterprise Miner occur in process flows. A *process flow* is a sequence of tasks or nodes connected by arrows in the user interface, and it defines the order of analysis. The organization of a process flow is contained in a file, EM_DGRAPH, which is stored in the diagram directory. Each node in the diagram corresponds to a separate subdirectory in the diagram directory. Information in one process flow can be sent to another by connecting the two process flows.

2.2 Creating a SAS Enterprise Miner Project, Library, and Diagram

Your first task when you start an analysis is creating a SAS Enterprise Miner project, data library, and diagram. Often these are set up by a SAS administrator (or reused from a previous analysis), but knowing how to set up these items yourself is fundamental to learning about SAS Enterprise Miner.

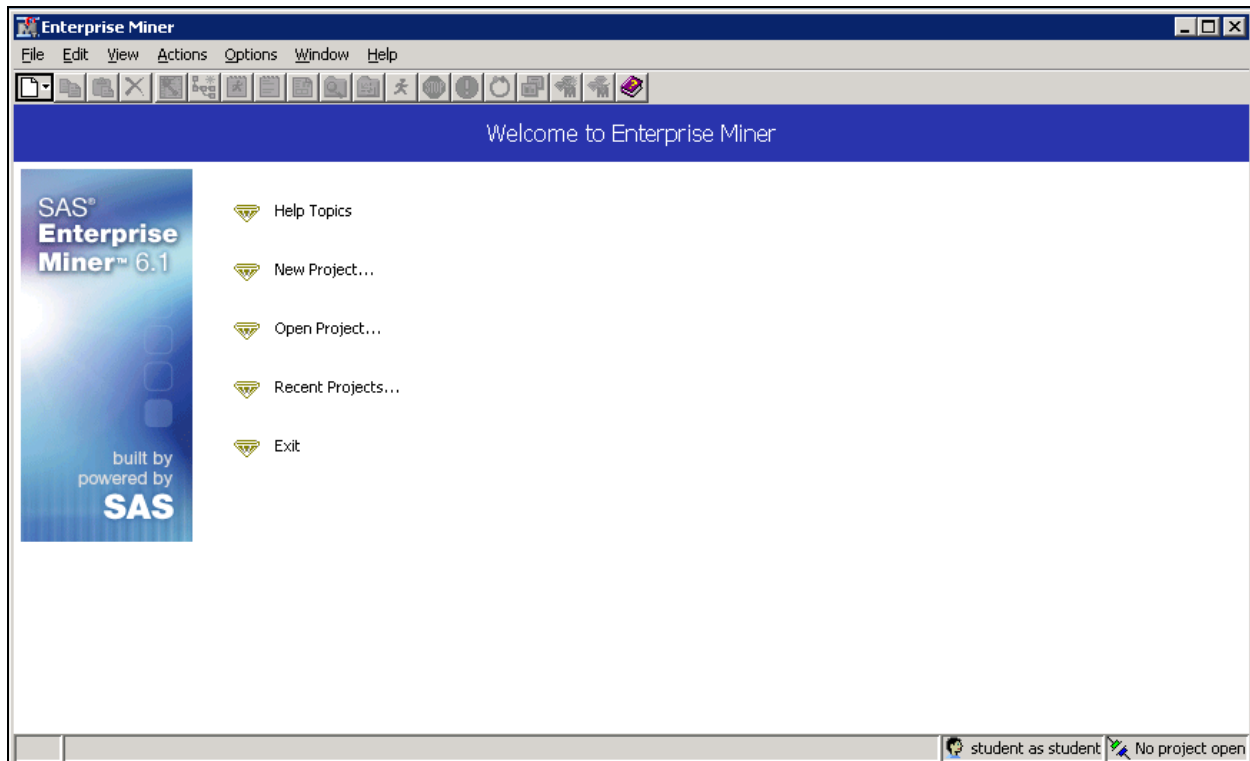
Before attempting the processes described in the following demonstrations, you need to log on to your computer, and start and log on to the SAS Enterprise Miner client program.



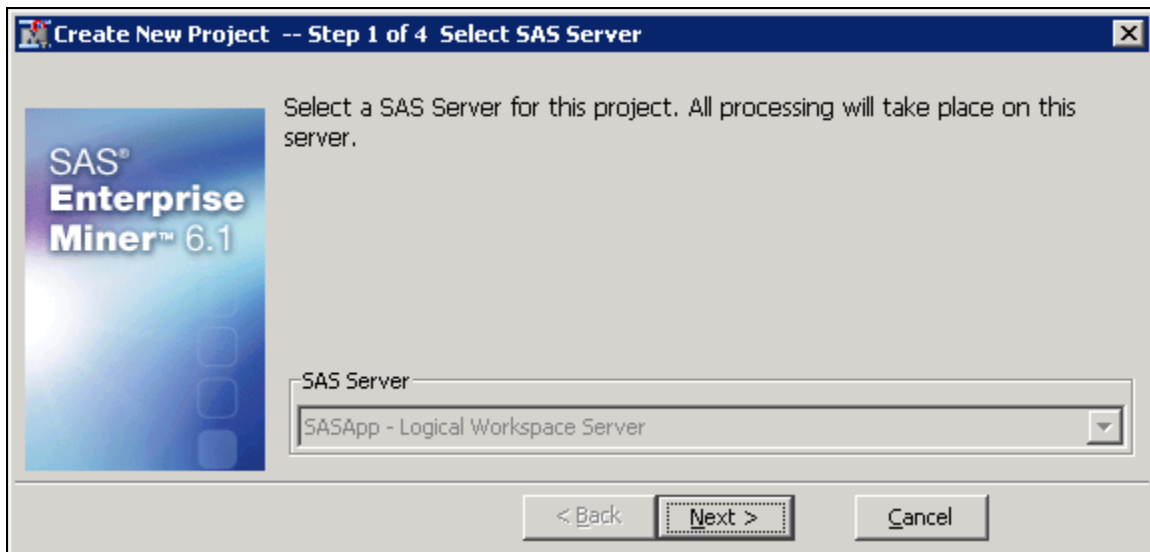
Creating a SAS Enterprise Miner Project

A SAS Enterprise Miner project contains materials related to a particular analysis task. These materials include analysis process flows, intermediate analysis data sets, and analysis results.

To define a project, you must specify a project name and the location of the project on the SAS Foundation Server. Follow the steps below to create a new SAS Enterprise Miner project.

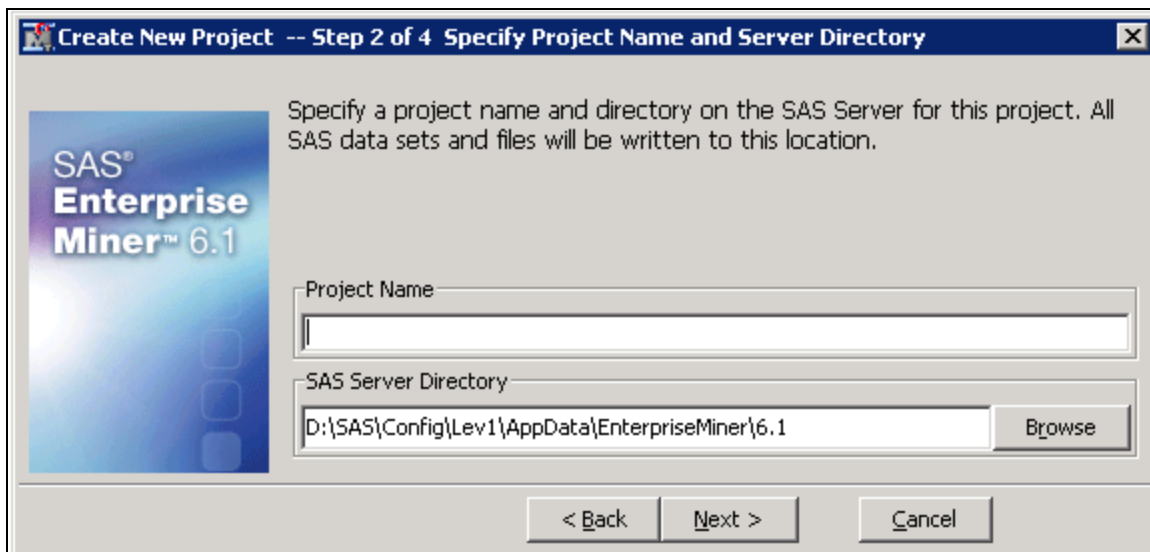


1. Select **File** ⇒ **New** ⇒ **Project** from the main menu. The Create New Project wizard opens at Step 1.



In this configuration of SAS Enterprise Miner, the only server available for processing is the host server listed above.

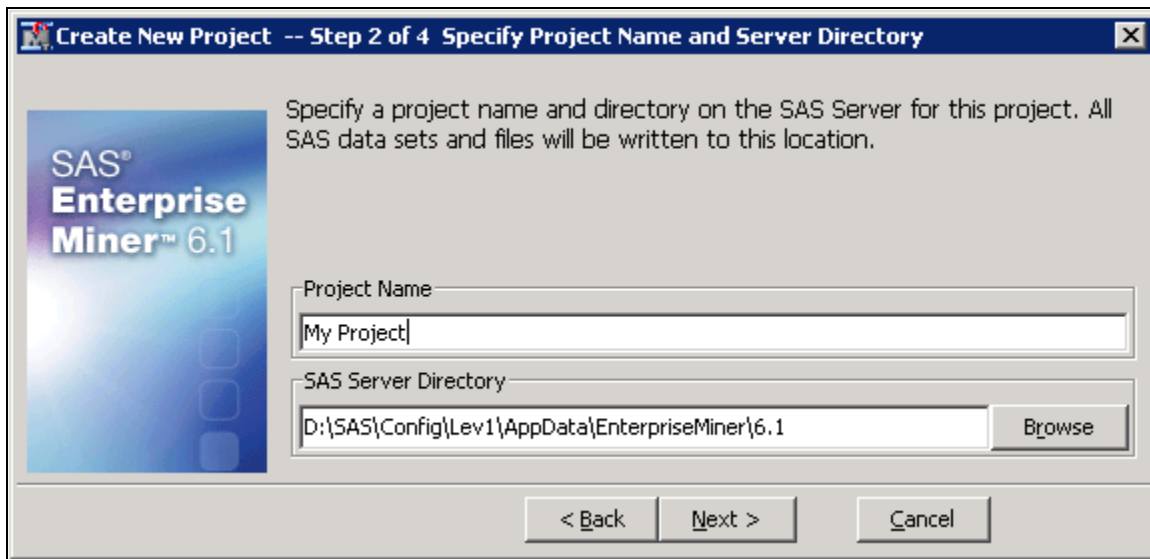
2. Select **Next >**.
3. Name the project.



Step 2 of the Create New Project wizard is used to specify the following information:

- the name of the project you are creating
- the location of the project

4. Type a project name, for example, **My Project**, in the Name field.



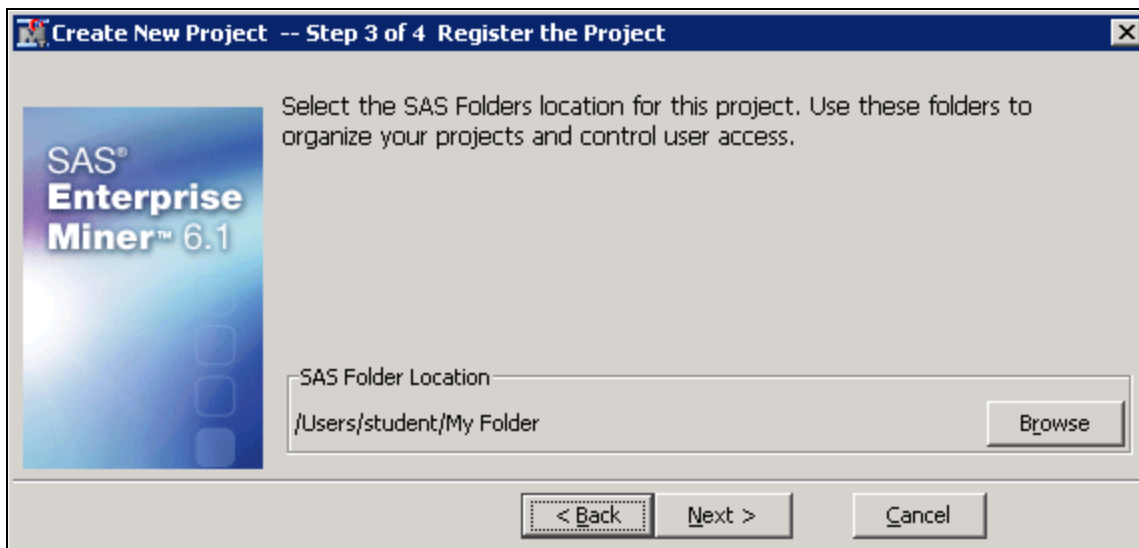
The path specified by the SAS Server Directory field is the physical location where the project folder will be created.

5. Select **Next >**.



If you have an existing project directory with the same name and location as specified, this project will be added to the list of available projects in SAS Enterprise Miner. This technique can be used to import a project created by another installation of SAS Enterprise Miner.

6. Select a location for the project's metadata.

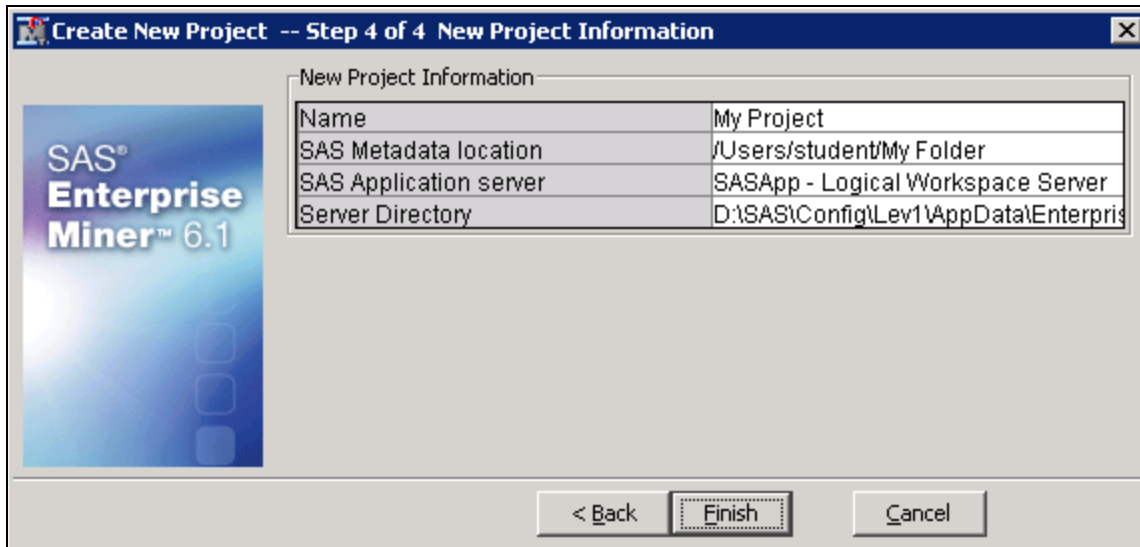


The SAS folder, My Folder, is in a WebDAV directory. This is where the metadata associated with the project is stored. This folder can be accessed and modified using SAS Management Console.

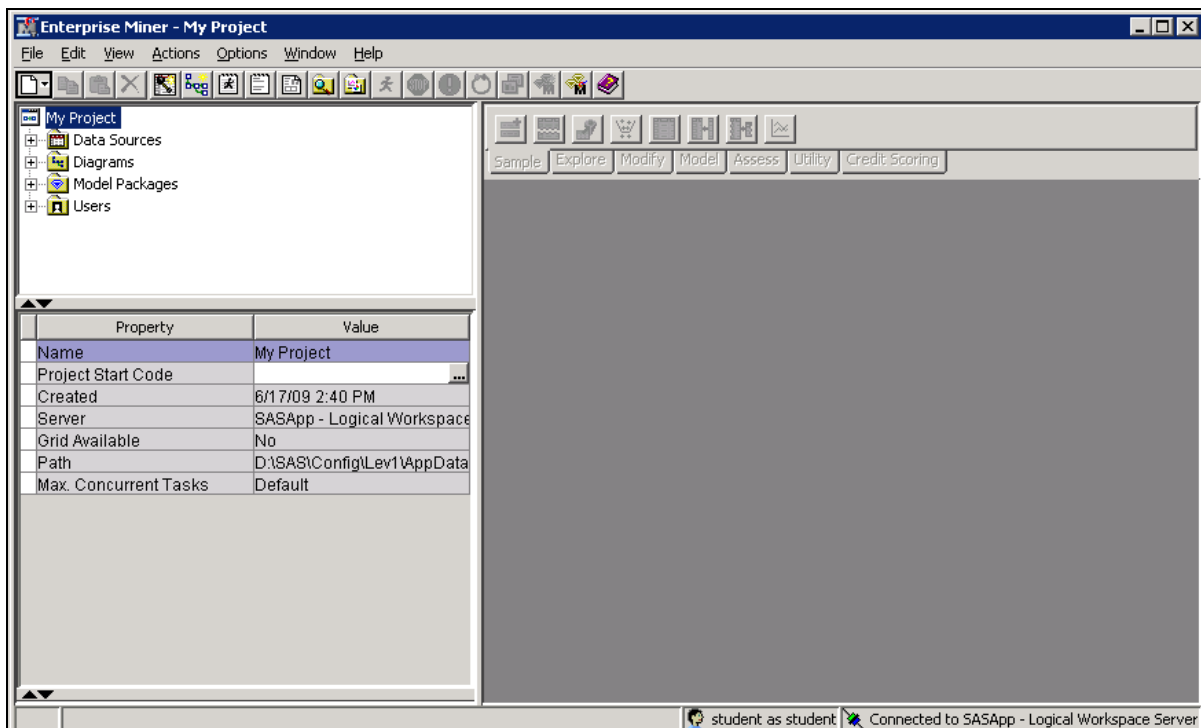
7. Select **Next >**.

Information about your project is summarized in Step 4.

8. To finish defining the project, select **Finish**.



The SAS Enterprise Miner client application opens the project that you created.





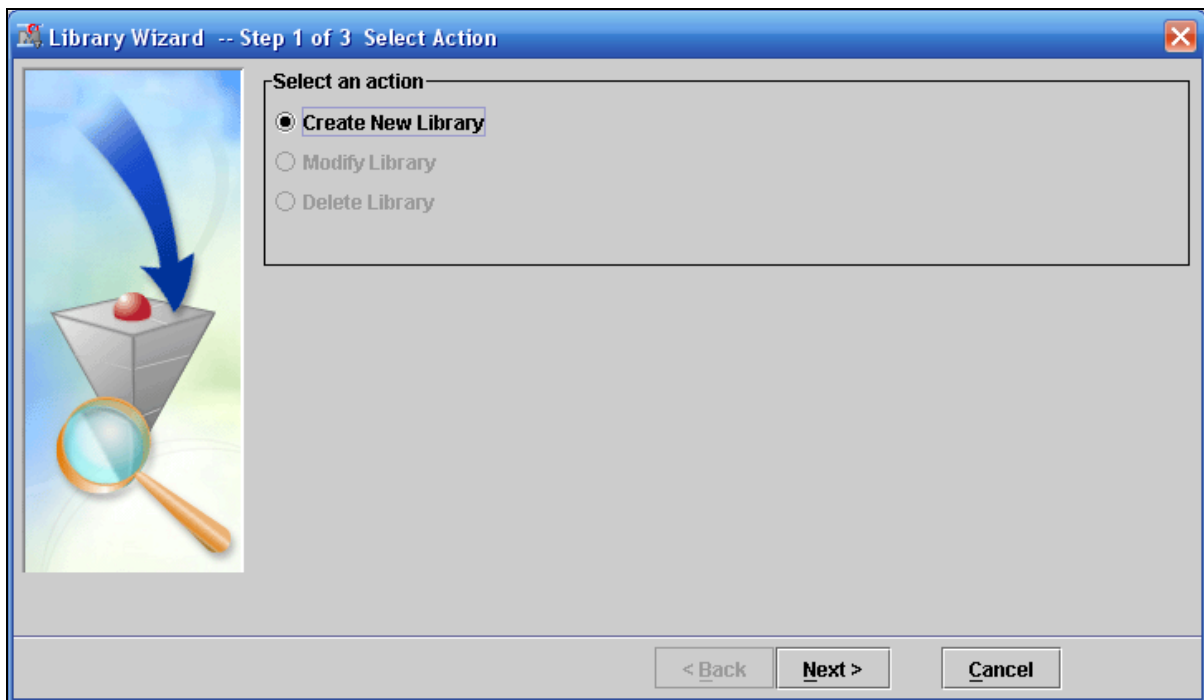
Creating a SAS Library

A SAS library connects SAS Enterprise Miner with the raw data sources, which are the basis of your analysis. A library can link to a directory on the SAS Foundation server, a relational database, or even an Excel workbook.

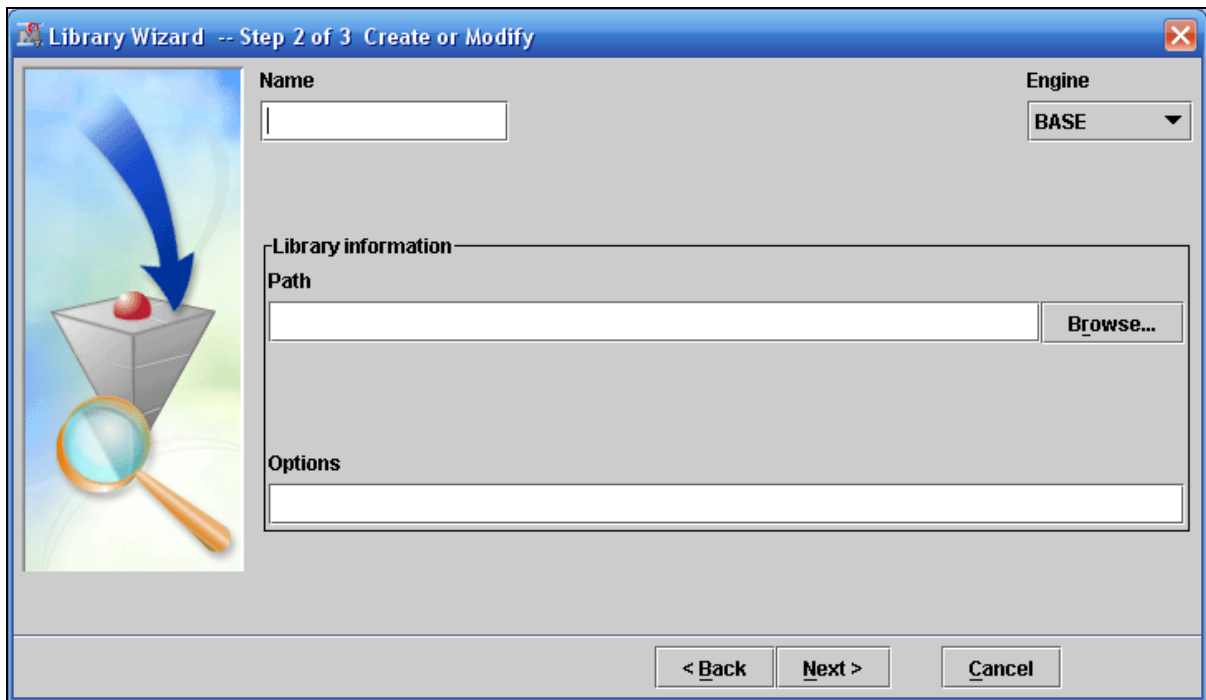
To define a library, you need to know the name and location of the data structure that you want to link with SAS Enterprise Miner, in addition to any associated options, such as user names and passwords.

Follow the steps below to create a new SAS library.

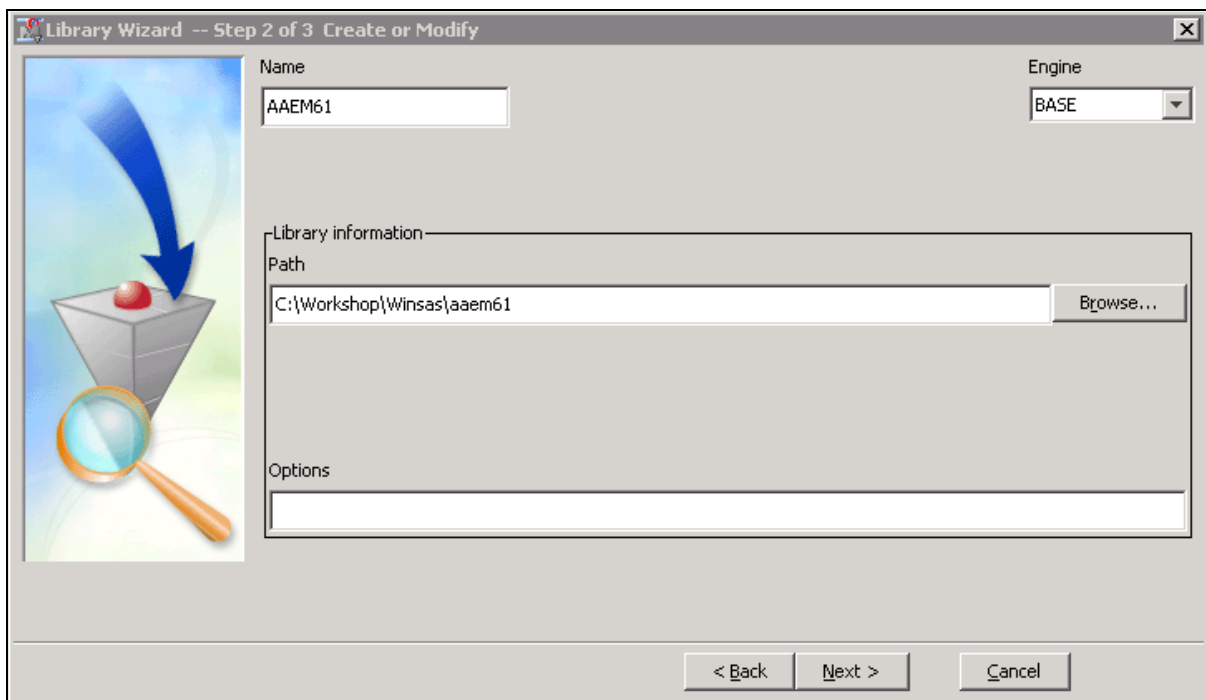
1. Select **File** ⇒ **New** ⇒ **Library...** from the main menu. The Library Wizard – Step 1 of 3 Select Action window opens.



2. Select **Next >**. The Library Wizard is updated to show Step 2 of 3 Create or Modify.

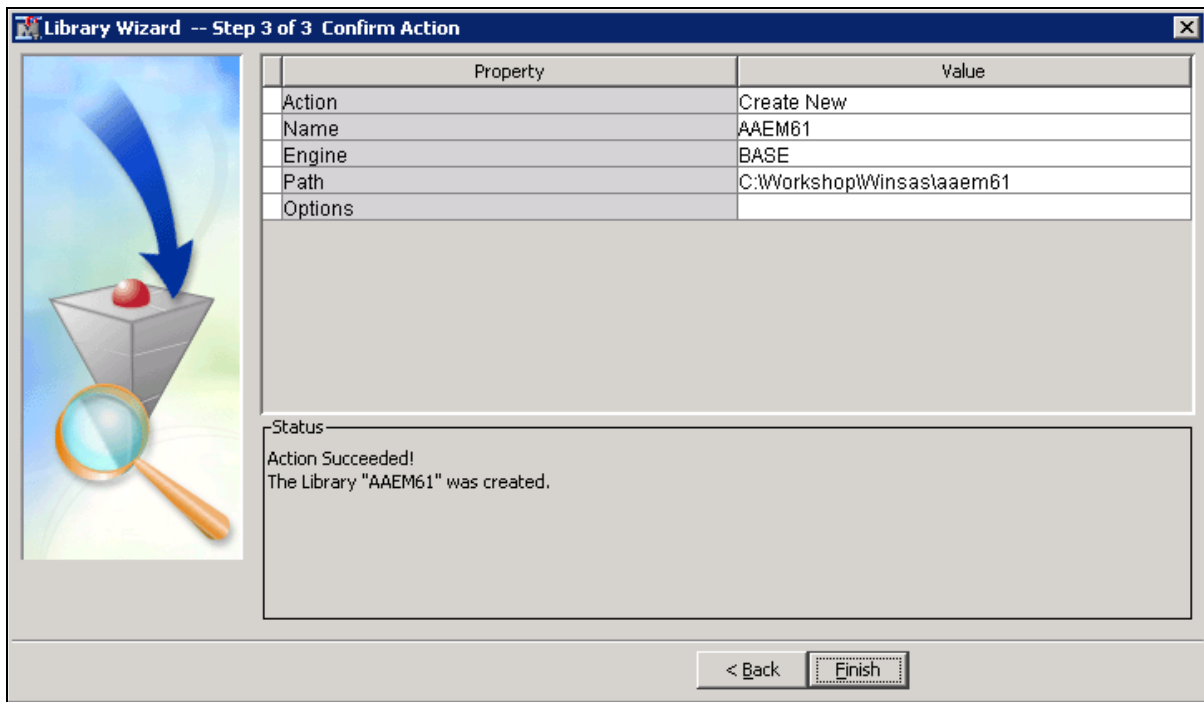


3. Type **AAEM61** in the Name field.
4. Type the path **C:\workshop\winsas\aaem61** in the Path field.



 If your data is installed in a different directory, specify this directory in the Path field.

5. Select **Next >**. The Library Wizard window is updated to show Step 3 of 3 Confirm Action.



The Confirm Action window shows the name, type, and path of the created SAS library.

6. Select **Finish**.

All data available in the SAS library can now be used by SAS Enterprise Miner.

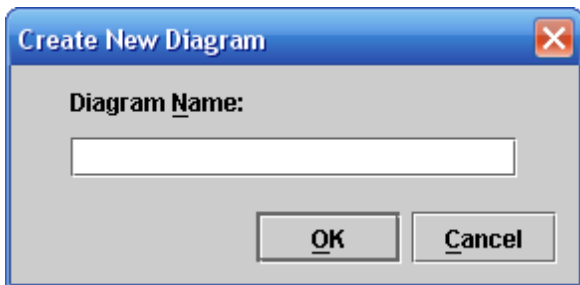


Creating a SAS Enterprise Miner Diagram

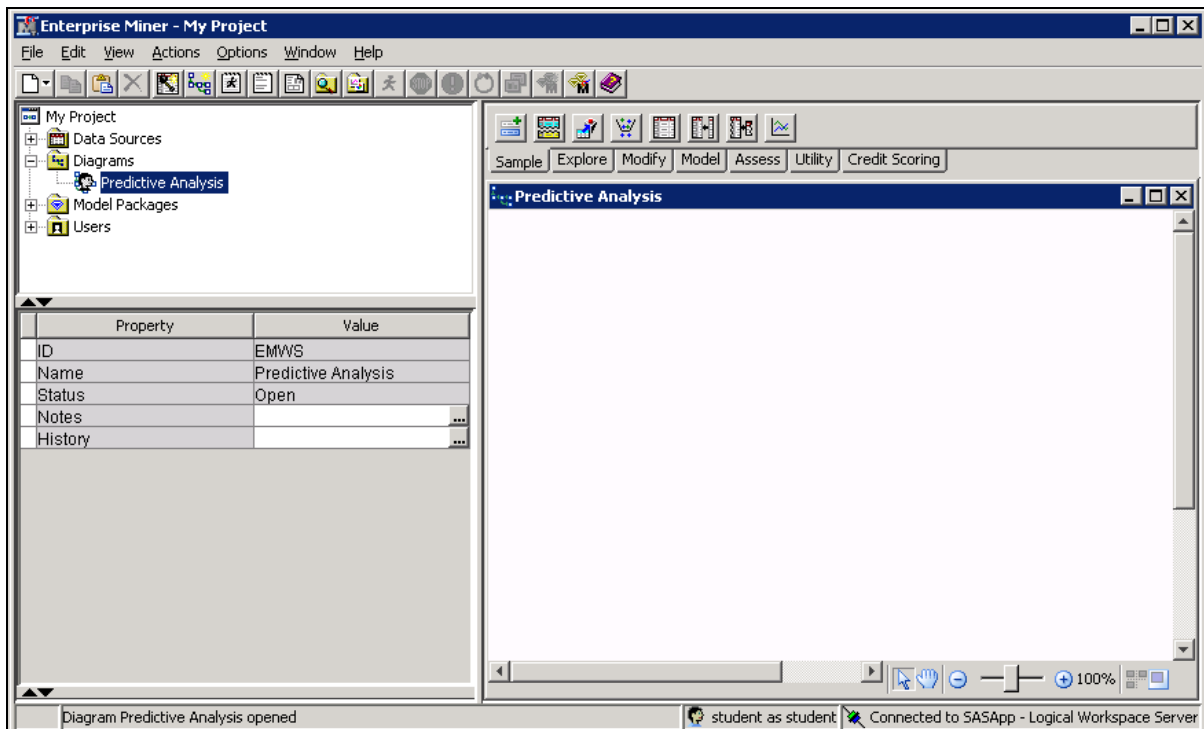
A SAS Enterprise Miner diagram workspace contains and displays the steps involved in your analysis. To define a diagram, you need only specify its name.

Follow the steps below to create a new SAS Enterprise Miner diagram workspace.

1. Select **File** ⇒ **New** ⇒ **Diagram...** from the main menu.



2. Type the name **Predictive Analysis** in the Diagram Name field and select **OK**. SAS Enterprise Miner creates an analysis workspace window labeled Predictive Analysis.



You use the Predictive Analysis window to create process flow diagrams.



Exercises

1. Creating a Project, Defining a Library, and Creating an Analysis Diagram

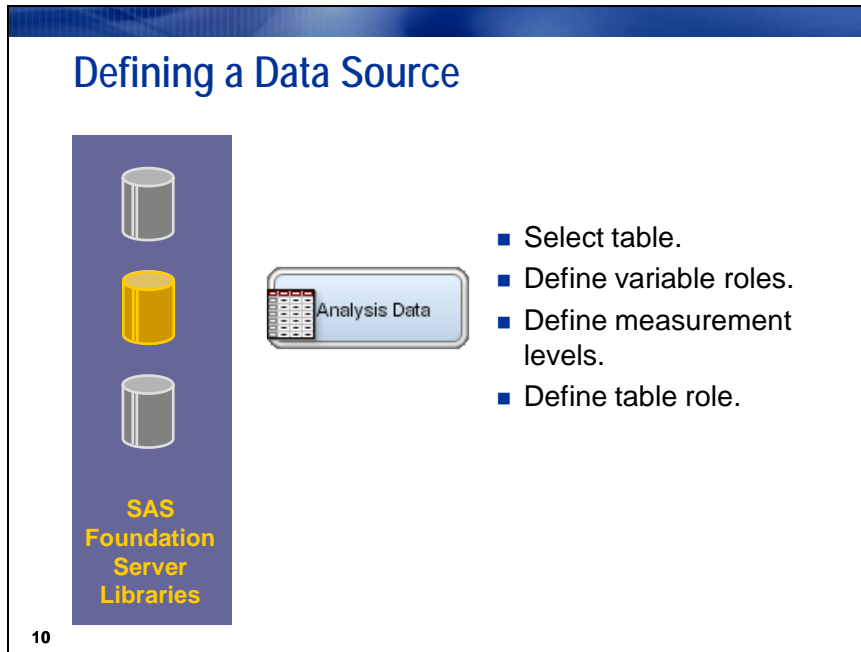
- a. Use the steps demonstrated on pages 2-6 to 2-8 to create a project for your SAS Enterprise Miner analyses on your computer.
- b. Use the steps demonstrated on pages 2-10 to 2-12 to define a SAS library to access the course data from your project.



Your instructor will provide the path to the data if it is different than that specified in these notes.

- c. Use the steps demonstrated on page 2-13 to create an analysis diagram in your project.

2.3 Defining a Data Source



After you define a new project and diagram, your next analysis task in SAS Enterprise Miner is usually to create an analysis data source. A *data source* is a link between an existing SAS table and SAS Enterprise Miner. To define a data source, you need to select the analysis table and define metadata appropriate to your analysis task.


The selected table must be visible to the SAS Foundation Server via a predefined SAS library. Any table found in a SAS library can be used, including those stored in formats external to SAS, such as tables from a relational database. (To define a SAS library to such tables might require SAS/ACCESS product licenses.)

The metadata definition serves three primary purposes for a selected data set. It informs SAS Enterprise Miner of the following:

- the analysis role of each variable
- the measurement level of each variable
- the analysis role of the data set

The analysis role of each variable tells SAS Enterprise Miner the purpose of the variable in the current analysis. The measurement level of each variable distinguishes continuous numeric variables from categorical variables. The analysis role of the data set tells SAS Enterprise Miner how to use the selected data set in the analysis. All of this information must be defined in the context of the analysis at hand.

Other (optional) metadata definitions are specific to certain types of analyses. These are discussed in the context of these analyses.

 Any data source that is defined for use in SAS Enterprise Miner should be largely ready for analysis. Usually, most of the data preparation work is completed before a data source is defined. It is possible (and useful for documentation purposes) to write the completed data preparation process in a SAS Code node embedded in SAS Enterprise Miner. Those details are outside the scope of this discussion.

Charity Direct Mail Demonstration

Analysis goal:

A veterans' organization seeks continued contributions from lapsing donors. Use lapsing-donor responses from an earlier campaign to predict future lapsing-donor responses.

11

...

To demonstrate defining a data source, and later, using SAS Enterprise Miner's predictive modeling tools, consider the following specific analysis example:

A national veterans' organization seeks to better target its solicitations for donation. By only soliciting the most likely donors, less money will be spent on solicitation efforts and more money will be available for charitable concerns. Solicitations involve sending a small gift to an individual and include a request for a donation. Gifts to donors include mailing labels and greeting cards.

The organization has more than 3.5 million individuals in its mailing database. These individuals are classified by their response behaviors to previous solicitation efforts. Of particular interest is the class of individuals identified as *lapsing donors*. These individuals made their most recent donation between 12 and 24 months ago. The organization seeks to rank its lapsing donors based on their responses to a greeting card mailing sent in June of 1997. (The charity calls this the 97NK Campaign.) With this ranking, a decision can be made to either solicit or ignore a lapsing individual in the June 1998 campaign.

Charity Direct Mail Demonstration

Analysis goal:

A veterans' organization seeks continued contributions from lapsing donors. Use lapsing-donor responses from an earlier campaign to predict future lapsing-donor responses.

Analysis data:

- Extracted from previous year's campaign
- Sample balances response/non-response rate
- Actual response rate approximately 5%

12

The source of this data is the Association for Computing Machinery's (ACM) 1998 KDD-Cup competition. The data set and other details of the competition are publicly available at the UCI KDD Archive at kdd.ics.uci.edu.

For model development, the data were sampled to balance the response and non-response rates. (The reason and consequences of this action are discussed in later chapters.) In the original campaign, the response rate was approximately 5%.

The following demonstrations show how to access the 97NK campaign data in SAS Enterprise Miner. The process is divided into three parts:

- specifying the source data
- setting the columns metadata
- finalizing the data source specification



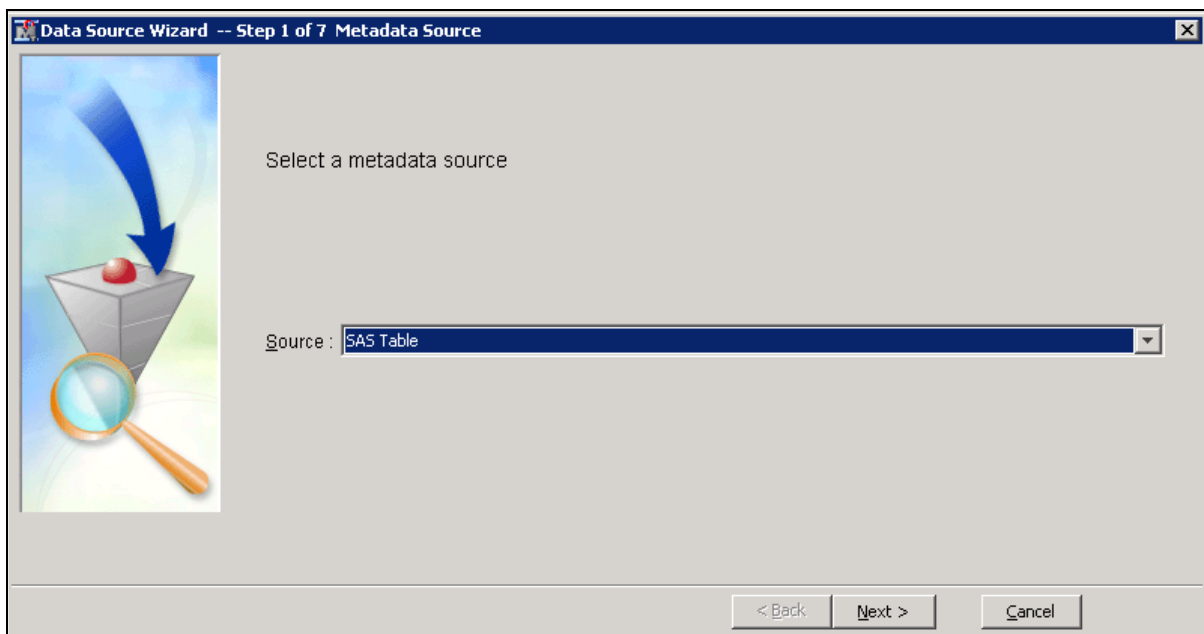
Defining a Data Source

Specifying Source Data

A data source links SAS Enterprise Miner to an existing analysis table. To specify a data source, you need to define a SAS library and know the name of the table that you will link to SAS Enterprise Miner.

Follow these steps to specify a data source.

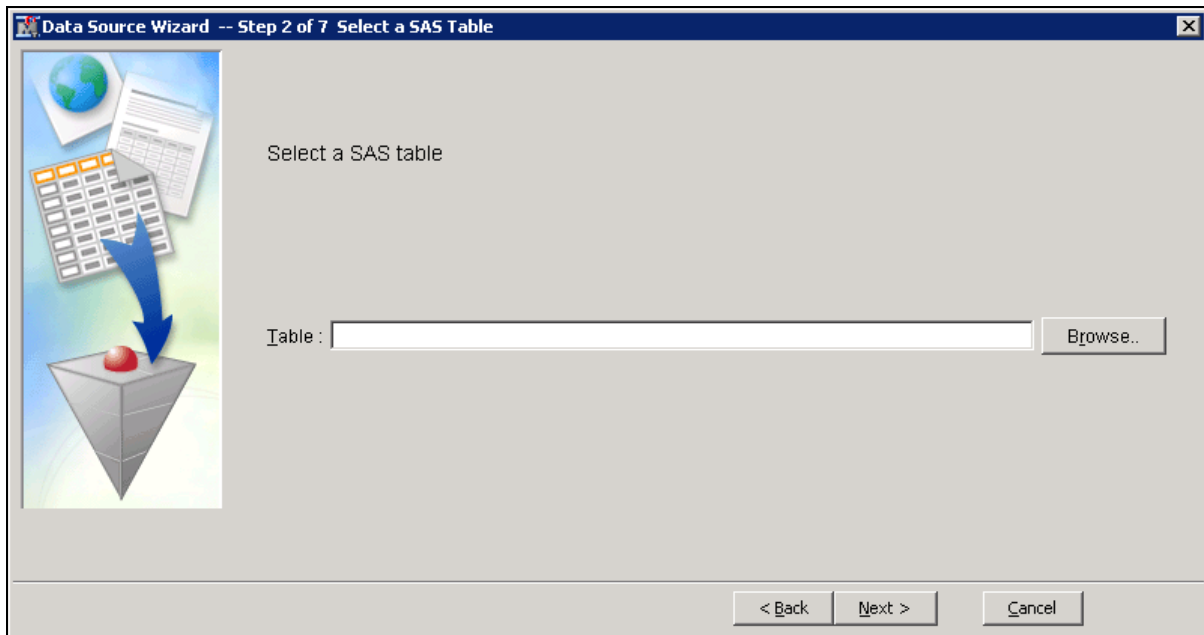
1. Select **File** ⇒ **New** ⇒ **Data Source...** from the main menu. The Data Source Wizard – Step 1 of 7 Metadata Source opens.



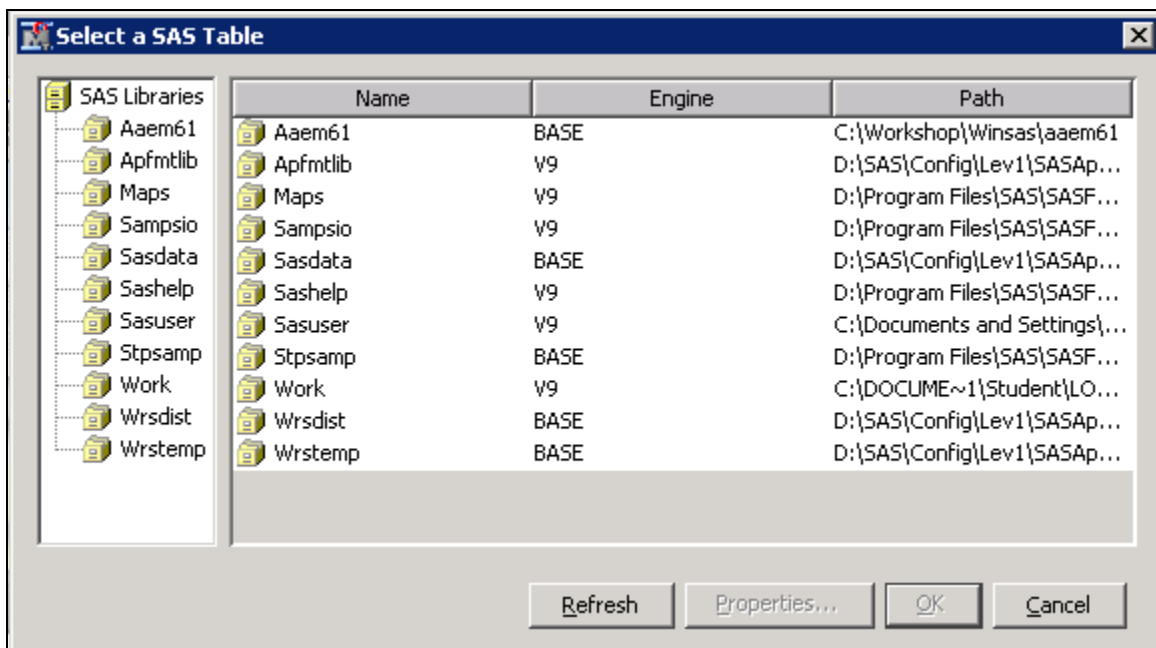
The Data Source Wizard guides you through a seven-step process to create a SAS Enterprise Miner data source. Step 1 tells SAS Enterprise Miner where to look for initial metadata values. The default and typical choice is the SAS Table that you will link to in the next step.

2. Select **Next >** to use a SAS table (the common choice) as the source for the metadata.

The Data Source Wizard continues to Step 2 of 7 Select a SAS Table.

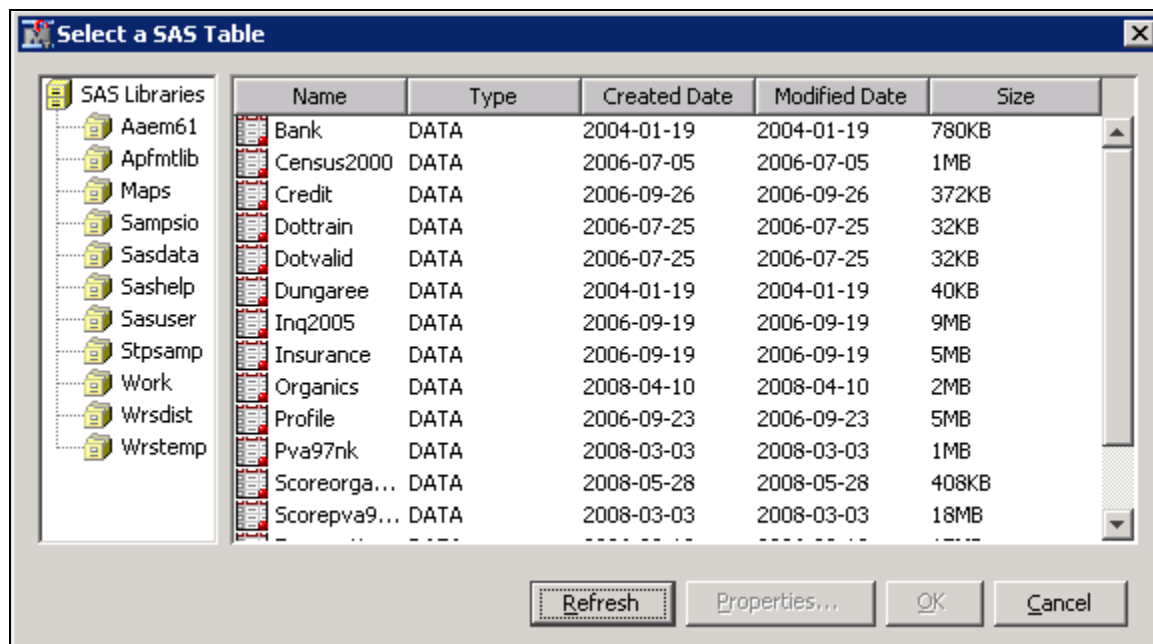


3. In this step, select the SAS table that you want to make available to SAS Enterprise Miner. You can either type the library name and SAS table name as **libname.tablename** or select the SAS table from a list.
4. Select **Browse...** to choose a SAS table from the libraries that are visible to the SAS Foundation Server. The Select a SAS Table window opens.

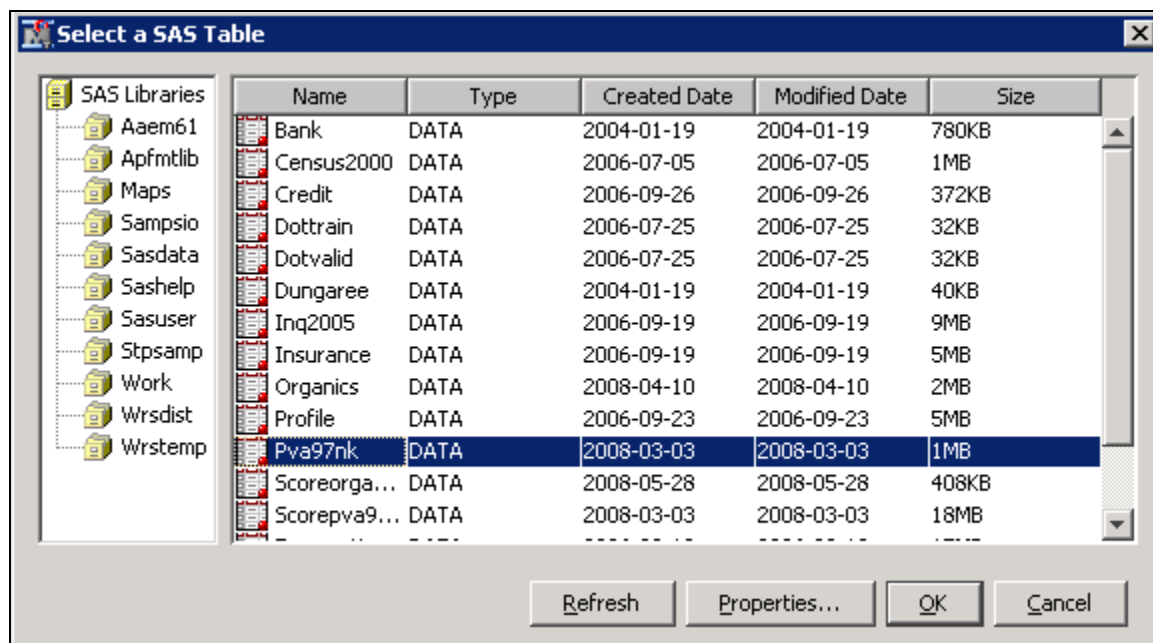


One of the libraries listed is named AAEM61, which is the library name defined in the Library Wizard.

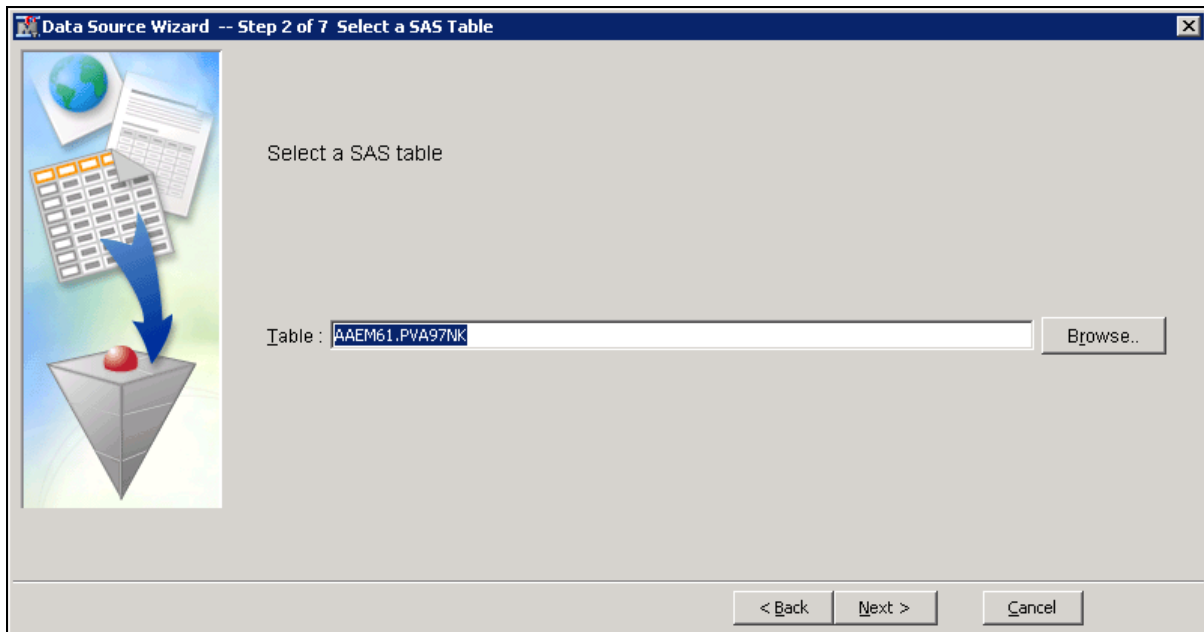
- Double-click the **Aaem61** library. The panel on the right is updated to show the contents of the AAEM61 library.



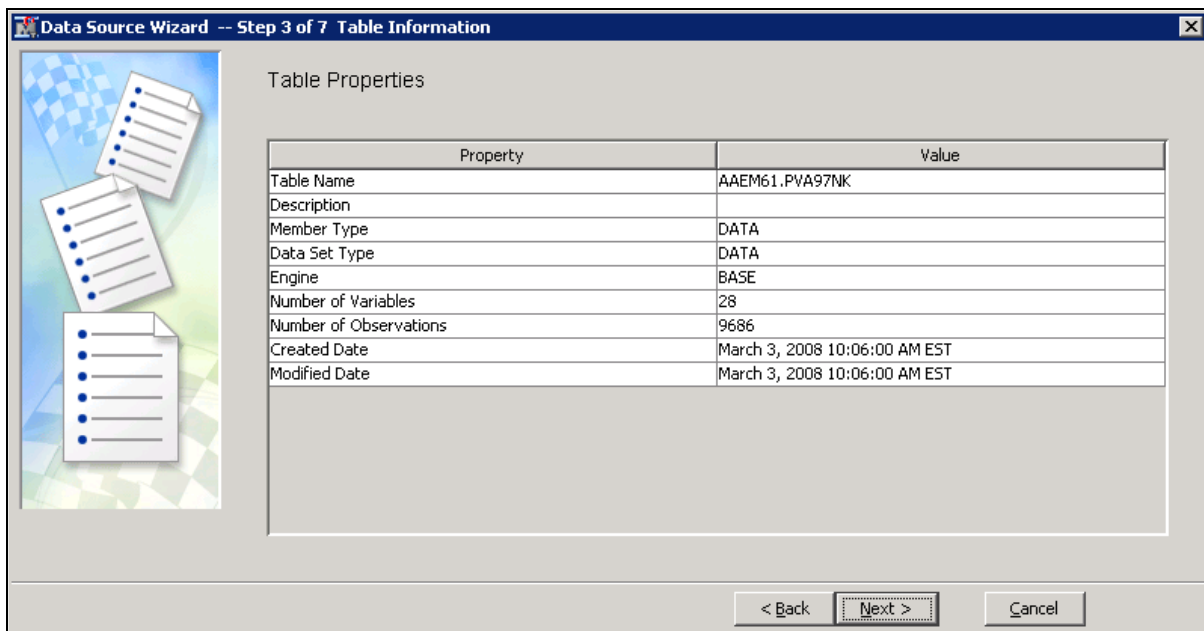
- Select the **Pva97nk** SAS table.



7. Select **OK**. The Select a SAS Table window closes and the selected table appears in the Table field.



8. Select **Next >**. The Data Source Wizard proceeds to Step 3 of 7 Table Information.



This step of the Data Source Wizard provides basic information about the selected table.



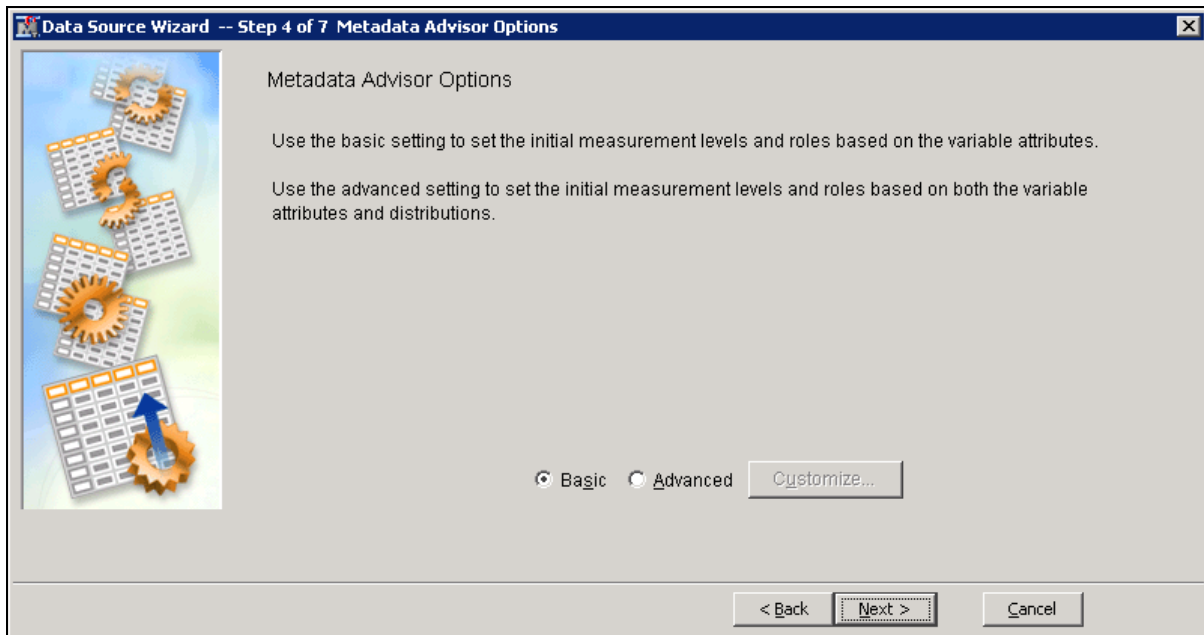
The SAS table **PVA97NK** is used in this chapter and subsequent chapters to demonstrate the predictive modeling tools of SAS Enterprise Miner. As seen in the Data Source Wizard – Step 3 of 7 Table Information window, the table contains 9,686 cases and 28 variables.

Defining Column Metadata

With a data set specified, your next task is to set the column metadata. To do this, you need to know the modeling role and proper measurement level of each variable in the source data set.

Follow these steps to define the column metadata:

1. Select **Next >**. The Data Source Wizard proceeds to Step 4 of 7 Metadata Advisor Options.



This step of the Data Source Wizard starts the metadata definition process. SAS Enterprise Miner assigns initial values to the metadata based on characteristics of the selected SAS table. The Basic setting assigns initial values to the metadata based on variable attributes such as the variable name, data type, and assigned SAS format. The Advanced setting assigns initial values to the metadata in the same way as the Basic setting, but it also assesses the distribution of each variable to better determine the appropriate measurement level.

2. Select **Next >** to use the Basic setting.

The Data Source Wizard proceeds to Step 5 of 7, Column Metadata.

Data Source Wizard -- Step 5 of 7 Column Metadata

(none) ☐ not Equal to

Columns: ☐ Label ☐ Mining ☐ Basic ☐ Statistics

Name	Role	Level	Report	Order	Dr
DemAge	Input	Interval	No		No
DemCluster	Input	Nominal	No		No
DemGender	Input	Nominal	No		No
DemHomeOwner	Input	Nominal	No		No
DemMedHomeValue	Input	Interval	No		No
DemMedIncome	Input	Interval	No		No
DemPctVeterans	Input	Interval	No		No
GiftAvg36	Input	Interval	No		No
GiftAvgAll	Input	Interval	No		No
GiftAvgCard36	Input	Interval	No		No
GiftAvgLast	Input	Interval	No		No
GiftCnt36	Input	Interval	No		No
GiftCntAll	Input	Interval	No		No
GiftCntCard36	Input	Interval	No		No
GiftCntCardAll	Input	Interval	No		No
GiftTimeFirst	Input	Interval	No		No
GiftTimeLast	Input	Interval	No		No
ID	ID	Nominal	No		No
PromCnt12	Input	Interval	No		No
PromCnt36	Input	Interval	No		No
PromCntAll	Input	Interval	No		No
PromCntCard12	Input	Interval	No		No
PromCntCard36	Input	Interval	No		No
PromCntCardAll	Input	Interval	No		No
StatusCat96NK	Input	Nominal	No		No
StatusCatStarAll	Input	Interval	No		No
TargetB	Target	Interval	No		No
TargetD	Target	Interval	No		No

The Data Source Wizard displays its best guess for the metadata assignments. This guess is based on the name and data type of each variable. The correct values for model role and measurement level are found in the **PVA97NK** metadata table on the next page.

A comparison of the currently assigned metadata to that in the **PVA97NK** metadata table shows several discrepancies. While the assigned modeling roles are mostly correct, the assigned measurement levels for several variables are in error.

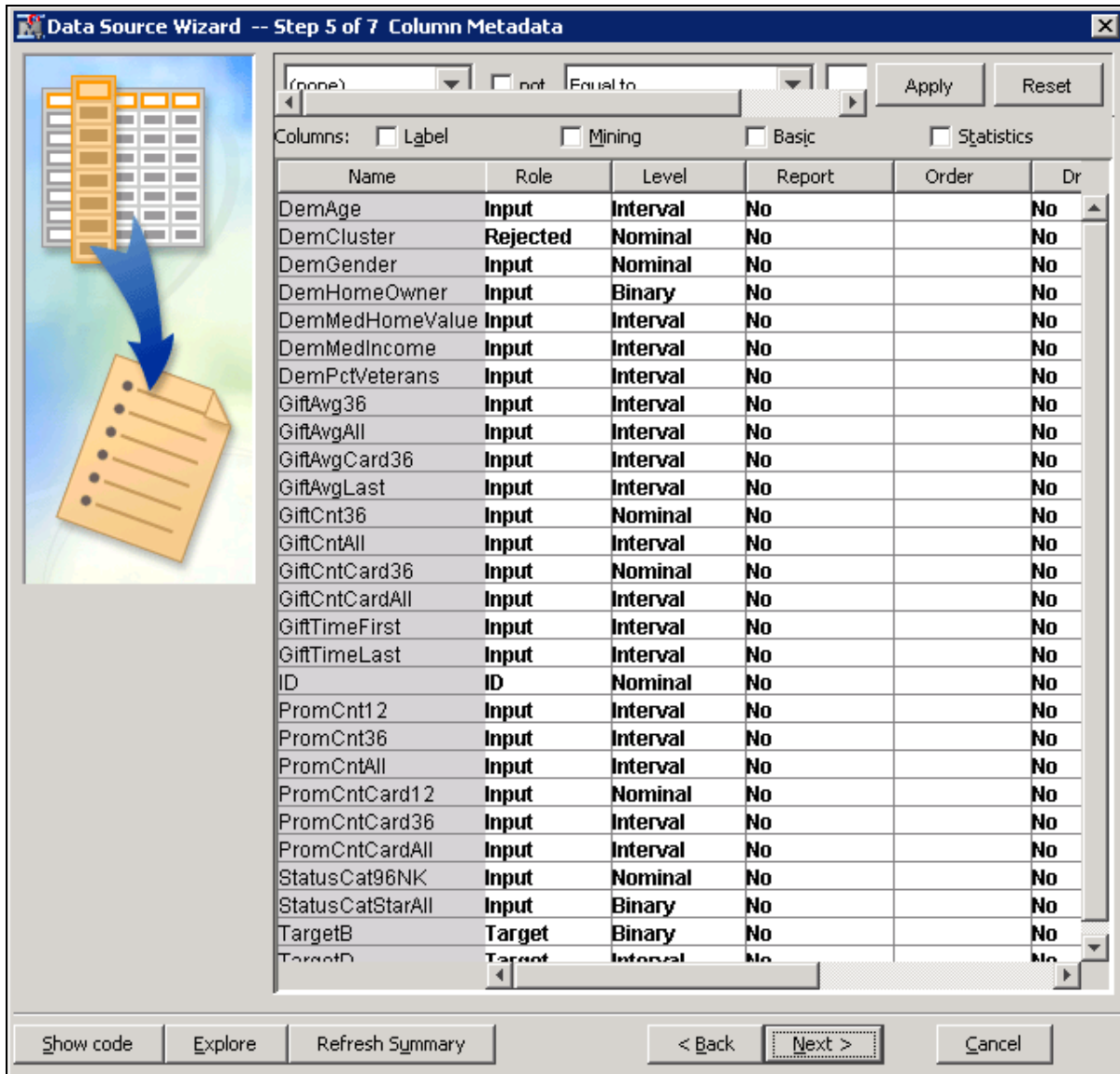
It is possible to improve the default metadata assignments by using the Advanced option in the Metadata Advisor.

3. Select **< Back** in the Data Source Wizard. This returns you to Step 4 of 6 Metadata Advisor Options.
4. Select the **Advanced** option.

PVA97NK Metadata Table

Name	Model Role	Measurement Level	Description
DemAge	Input	Interval	Age
DemCluster	Input	Nominal	Demographic Cluster
DemGender	Input	Nominal	Gender
DemHomeOwner	Input	Binary	Home Owner
DemMedHomeValue	Input	Interval	Median Home Value Region
DemMedIncome	Input	Interval	Median Income Region
DemPctVeterans	Input	Interval	Percent Veterans Region
GiftAvg36	Input	Interval	Gift Amount Average 36 Months
GiftAvgAll	Input	Interval	Gift Amount Average All Months
GiftAvgCard36	Input	Interval	Gift Amount Average Card 36 Months
GiftAvgLast	Input	Interval	Gift Amount Last
GiftCnt36	Input	Interval	Gift Count 36 Months
GiftCntAll	Input	Interval	Gift Count All Months
GiftCntCard36	Input	Interval	Gift Count Card 36 Months
GiftCntCardAll	Input	Interval	Gift Count Card All Months
GiftTimeFirst	Input	Interval	Time Since First Gift
GiftTimeLast	Input	Interval	Time Since Last Gift
ID	ID	Nominal	Control Number
PromCnt12	Input	Interval	Promotion Count 12 Months
PromCnt36	Input	Interval	Promotion Count 36 Months
PromCntAll	Input	Interval	Promotion Count All Months
PromCntCard12	Input	Interval	Promotion Count Card 12 Months
PromCntCard36	Input	Interval	Promotion Count Card 36 Months
PromCntCardAll	Input	Interval	Promotion Count Card All Months
StatusCat96NK	Input	Nominal	Status Category 96NK
StatusCatStarAll	Input	Binary	Status Category Star All Months
TargetB	Target	Binary	Target Gift Flag
TargetD	Rejected	Interval	Target Gift Amount

5. Select **Next >** to use the Advanced setting. The Data Source Wizard again proceeds to Step 5 of 7 Column Metadata.



Data Source Wizard -- Step 5 of 7 Column Metadata

(none) not Equal to Apply Reset

Columns: ☐ Label ☐ Mining ☐ Basic ☐ Statistics

Name	Role	Level	Report	Order	Dr
DemAge	Input	Interval	No		No
DemCluster	Rejected	Nominal	No		No
DemGender	Input	Nominal	No		No
DemHomeOwner	Input	Binary	No		No
DemMedHomeValue	Input	Interval	No		No
DemMedIncome	Input	Interval	No		No
DemPctVeterans	Input	Interval	No		No
GiftAvg36	Input	Interval	No		No
GiftAvgAll	Input	Interval	No		No
GiftAvgCard36	Input	Interval	No		No
GiftAvgLast	Input	Interval	No		No
GiftCnt36	Input	Nominal	No		No
GiftCntAll	Input	Interval	No		No
GiftCntCard36	Input	Nominal	No		No
GiftCntCardAll	Input	Interval	No		No
GiftTimeFirst	Input	Interval	No		No
GiftTimeLast	Input	Interval	No		No
ID	ID	Nominal	No		No
PromCnt12	Input	Interval	No		No
PromCnt36	Input	Interval	No		No
PromCntAll	Input	Interval	No		No
PromCntCard12	Input	Nominal	No		No
PromCntCard36	Input	Interval	No		No
PromCntCardAll	Input	Interval	No		No
StatusCat96NK	Input	Nominal	No		No
StatusCatStarAll	Input	Binary	No		No
TargetB	Target	Binary	No		No
TargetD	Target	Interval	No		No

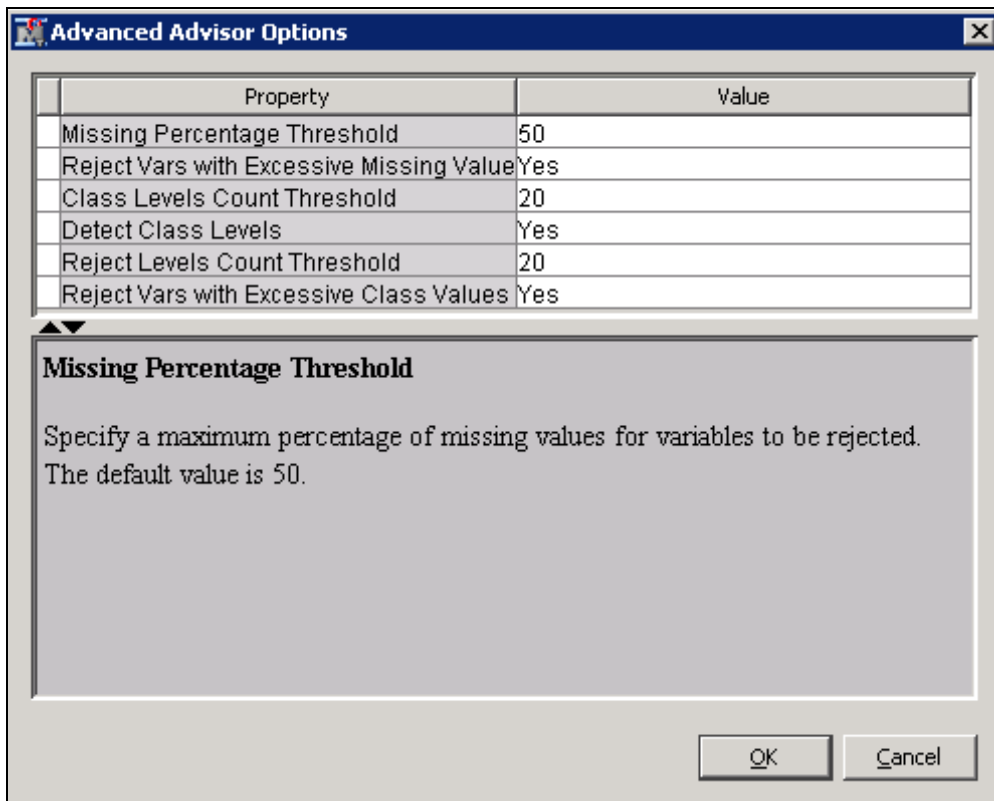
Show code Explore Refresh Summary < Back **Next >** Cancel

While many of the default metadata settings are correct, there are several items that need to be changed. For example, the **DemCluster** variable is rejected (for having too many distinct values), and several numeric inputs have their measurement level set to Nominal instead of Interval (for having too few distinct values).

To avoid the time-consuming task of making metadata adjustments, go back to the previous Data Source Wizard step and customize the Metadata Advisor.

6. Select **< Back**. You return to the Metadata Advisor Options window.

7. Select **Customize...**. The Advanced Advisor Options dialog box opens.



Using the default Advanced options, the Metadata Advisor can do the following:

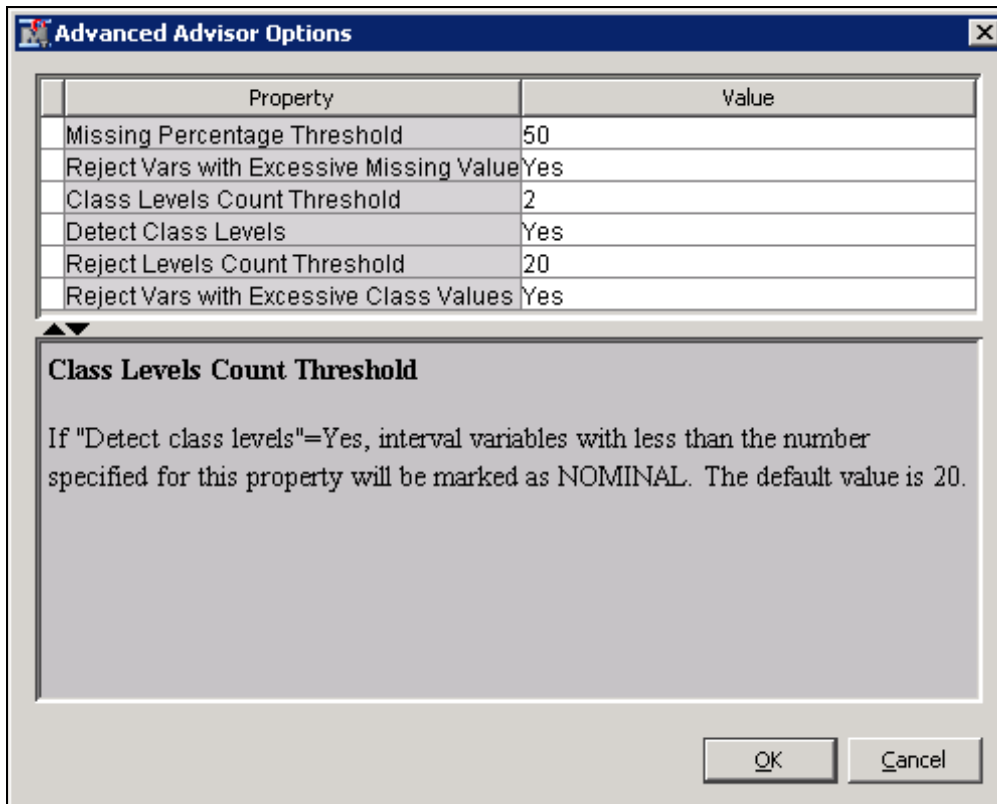
- reject variables with an excessive number of missing values (default=50%)
- detect the number class levels of **numeric** variables and assign a role of Nominal to those with class counts below the selected threshold (default=20)
- detect the number class levels of **character** variables and assign a role of Rejected to those with class counts above the selected threshold (default=20)



In the **PVA97NK** table, there are several numeric variables with fewer than 20 distinct values that should **not** be treated as nominal. Similarly, there is one class variable with more than 20 levels that should **not** be rejected.

To avoid changing many metadata values in the next step of the Data Source Wizard, you should alter these defaults.

8. Type **2** as the Class Levels Count Threshold value so that only binary numeric variables are treated as categorical variables.



The image shows a dialog box titled "Advanced Advisor Options". It contains a table with two columns: "Property" and "Value". The table lists several properties and their corresponding values. Below the table, there is a section titled "Class Levels Count Threshold" with a descriptive text. At the bottom right, there are "OK" and "Cancel" buttons.

Property	Value
Missing Percentage Threshold	50
Reject Vars with Excessive Missing Value	Yes
Class Levels Count Threshold	2
Detect Class Levels	Yes
Reject Levels Count Threshold	20
Reject Vars with Excessive Class Values	Yes

Class Levels Count Threshold

If "Detect class levels"=Yes, interval variables with less than the number specified for this property will be marked as NOMINAL. The default value is 20.

OK Cancel

9. Type **100** as the Reject Levels Count Threshold value, so that only character variables with more than 100 distinct values are rejected.



Be sure to press ENTER after you type the number **100**. Otherwise, the value might not be registered in the field.

Property	Value
Missing Percentage Threshold	50
Reject Vars with Excessive Missing Value	Yes
Class Levels Count Threshold	2
Detect Class Levels	Yes
Reject Levels Count Threshold	100
Reject Vars with Excessive Class Values	Yes

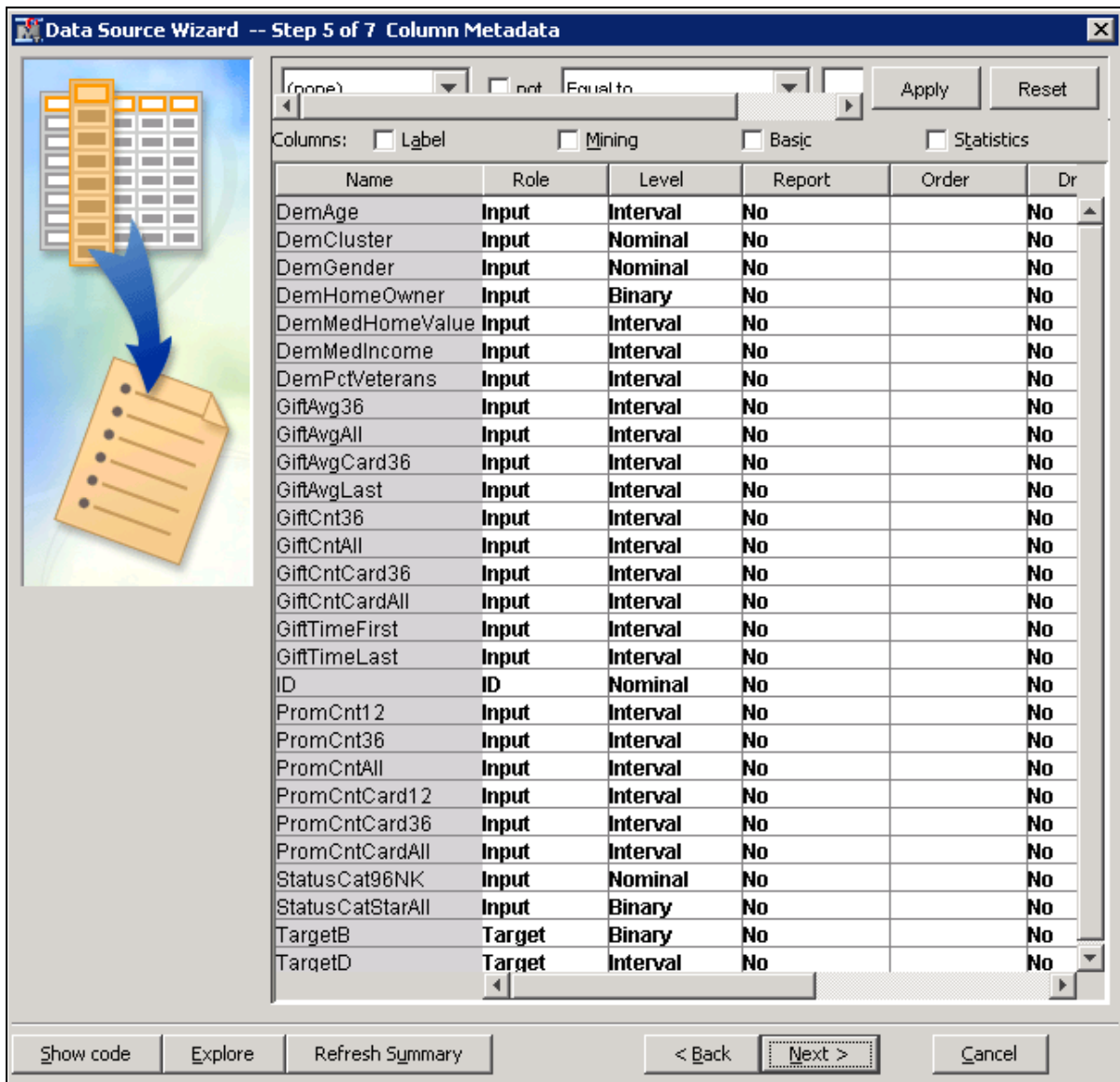
Reject Levels Count Threshold

Specify a maximum number of levels for a class variable before being marked REJECTED. The default value is 20.

OK Cancel

10. Select **OK** to close the Advanced Advisor Options dialog box.

11. Select **Next >** to proceed to Step 5 of the Data Source Wizard.



Data Source Wizard -- Step 5 of 7 Column Metadata

(none) ☐ not Equal to

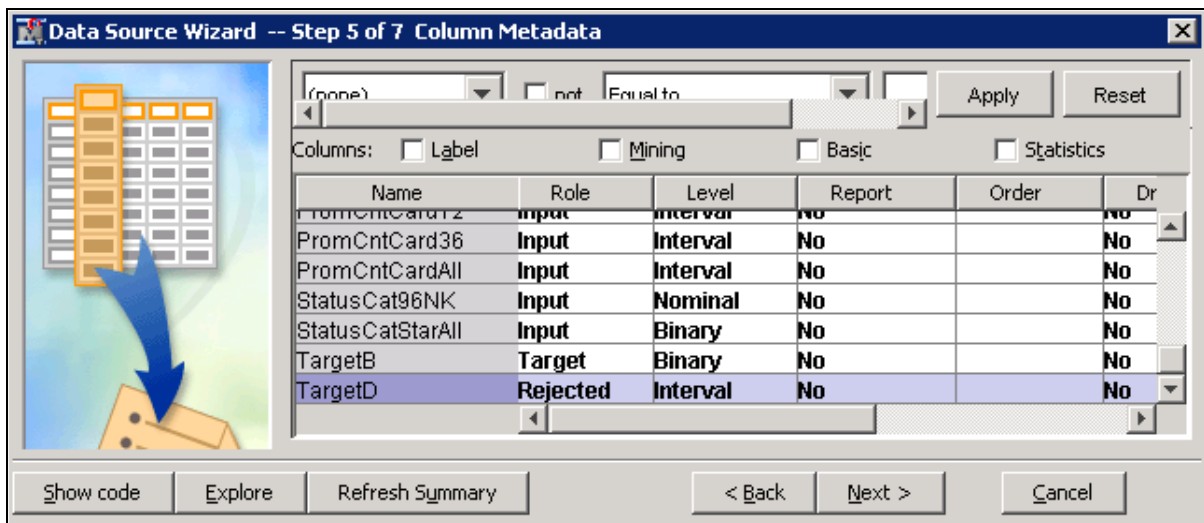
Columns: ☐ Label ☐ Mining ☐ Basic ☐ Statistics

Name	Role	Level	Report	Order	Dr
DemAge	Input	Interval	No		No
DemCluster	Input	Nominal	No		No
DemGender	Input	Nominal	No		No
DemHomeOwner	Input	Binary	No		No
DemMedHomeValue	Input	Interval	No		No
DemMedIncome	Input	Interval	No		No
DemPctVeterans	Input	Interval	No		No
GiftAvg36	Input	Interval	No		No
GiftAvgAll	Input	Interval	No		No
GiftAvgCard36	Input	Interval	No		No
GiftAvgLast	Input	Interval	No		No
GiftCnt36	Input	Interval	No		No
GiftCntAll	Input	Interval	No		No
GiftCntCard36	Input	Interval	No		No
GiftCntCardAll	Input	Interval	No		No
GiftTimeFirst	Input	Interval	No		No
GiftTimeLast	Input	Interval	No		No
ID	ID	Nominal	No		No
PromCnt12	Input	Interval	No		No
PromCnt36	Input	Interval	No		No
PromCntAll	Input	Interval	No		No
PromCntCard12	Input	Interval	No		No
PromCntCard36	Input	Interval	No		No
PromCntCardAll	Input	Interval	No		No
StatusCat96NK	Input	Nominal	No		No
StatusCatStarAll	Input	Binary	No		No
TargetB	Target	Binary	No		No
TargetD	Target	Interval	No		No

A comparison of the Column Metadata table to the table at the beginning of the demonstration shows that most of the metadata is correctly defined. SAS Enterprise Miner correctly inferred the model roles for the non-input variables by their names. The measurement levels are correctly defined by using the Advanced Metadata Advisor.

The analysis of the **PVA97NK** data in this course focuses on the **TargetB** variable, so the **TargetD** variable should be rejected.

12. Select **Role** ⇒ **Rejected** for **TargetD**.



In summary, Step 5 of 7 Column Metadata is usually the most time-consuming of the Data Source Wizard steps. You can use the following tips to reduce the amount of time required to define metadata for SAS Enterprise Miner predictive modeling data sets:

- Only include variables that you intend to use in the modeling process in your raw data source.
- For variables that are not inputs, use variable names that start with the intended role. For example, an ID variable should start with **ID** and a target variable should start with **Target**.
- Inputs that are to have a nominal measurement level should have a **character** data type.
- Inputs that are to be interval must have a **numeric** data type.
- Customize the Metadata Advisor to have a Class Level Count set equal to **2** and a Reject Levels Count set equal to a number greater than the maximum cardinality (level count) of your nominal inputs.

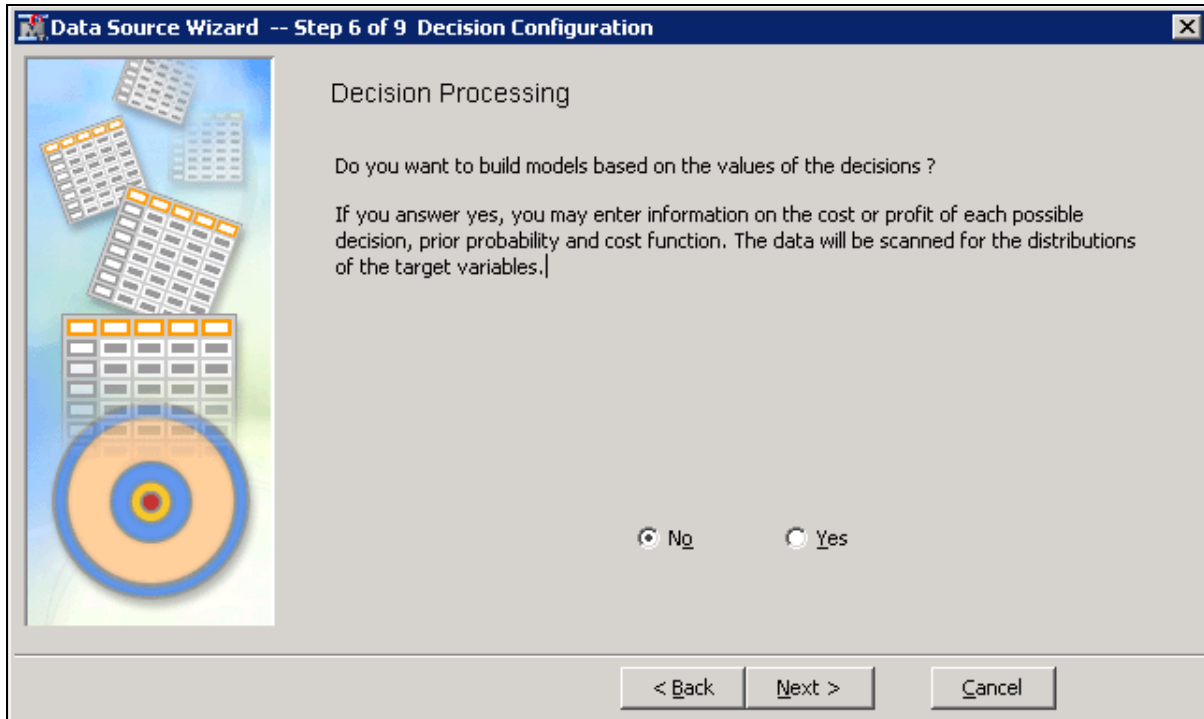
Finalizing the Data Source Specification

Follow these steps to complete the data source specification process:

1. Select **Next >** to proceed to Decision Configuration.



The Data Source Wizard gained an extra step due to the presence of a categorical (binary, ordinal, or nominal) target variable.

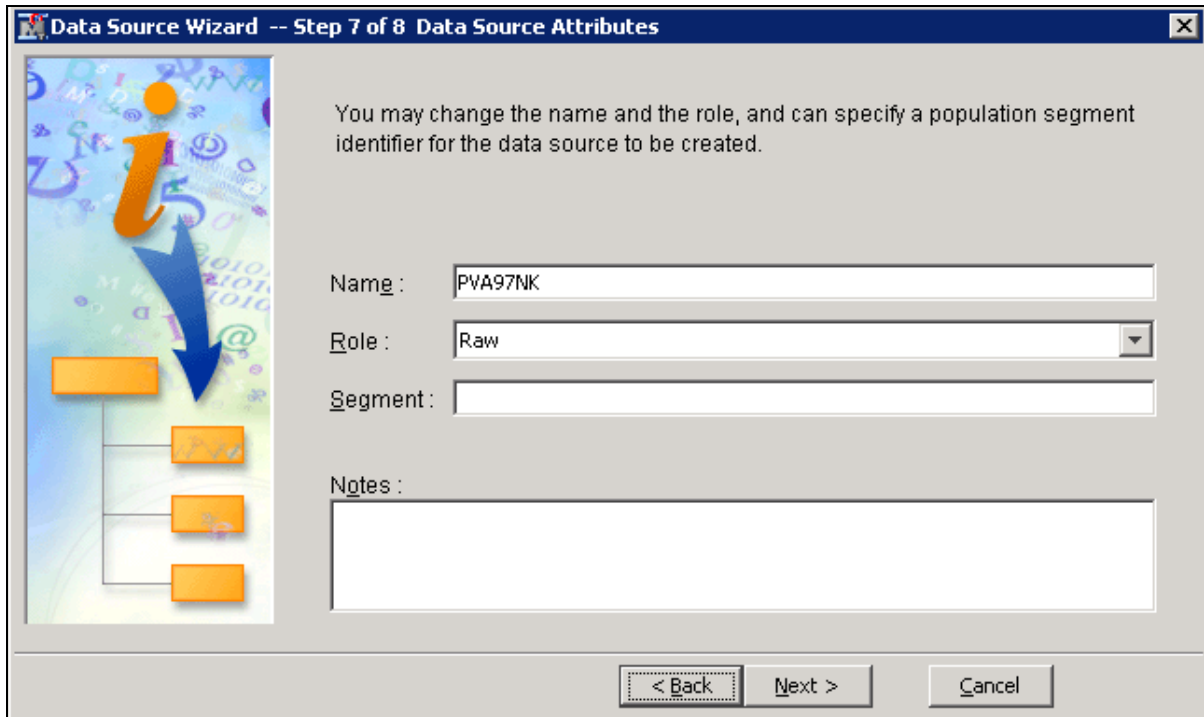


When you define a predictive modeling data set, it is important to properly configure decision processing. In fact, obtaining meaningful models often requires using these options. The **PVA97NK** table was structured so that reasonable models are produced **without** specifying decision processing. However, this might not be the case for data sources that you will encounter outside this course. Because you need to understand how to set these options, a detailed discussion of decision processing is provided in Chapter 6, “Model Assessment.”



Do **not** select **Yes** here because that changes the default settings for subsequent analysis steps and yields results that diverge from those in the course notes.

2. Select **Next >**. By skipping the decision processing step, you reach the next step of the Data Source Wizard.



The image shows a screenshot of the 'Data Source Wizard' window, specifically 'Step 7 of 8 Data Source Attributes'. The window has a title bar with the text 'Data Source Wizard -- Step 7 of 8 Data Source Attributes' and a close button. On the left side, there is a vertical navigation pane with a large orange 'i' icon and a blue arrow pointing down. Below the arrow are three orange rectangular buttons. The main area of the window contains the following text and controls:

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

Name :

Role :

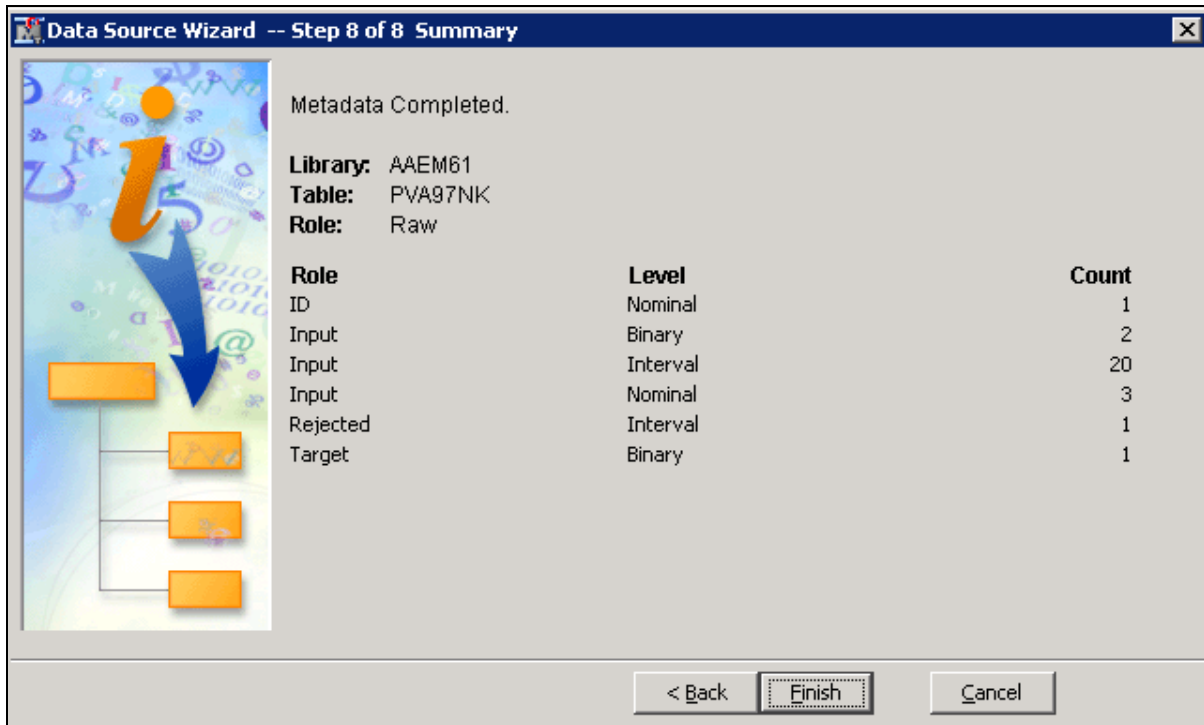
Segment :

Notes :

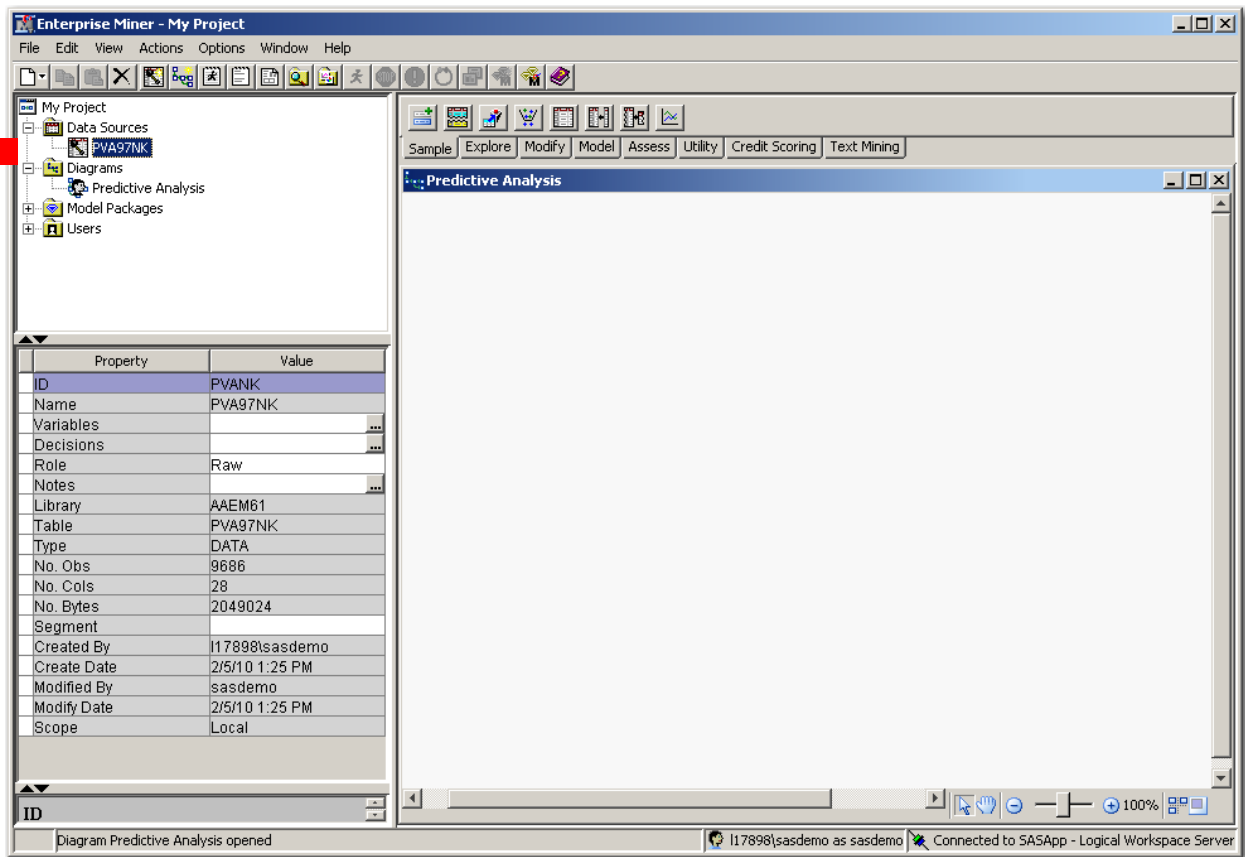
At the bottom of the window, there are three buttons: '< Back', 'Next >', and 'Cancel'.

This penultimate step enables you to set a role for the data source and add descriptive comments about the data source definition. For the upcoming analysis, a table role of Raw is acceptable.

3. The final step in the Data Source Wizard provides summary details about the data table that you created. Select **Finish**.



The **PVA97NK** data source is added to the Data Sources entry in the Project panel.



4. Select the **PVA97NK** data source to obtain table properties in the SAS Enterprise Miner Properties panel.



Exercises

2. Creating a SAS Enterprise Miner Data Source

Use the steps demonstrated on pages 2-18 to 2-34 to create a SAS Enterprise Miner data source from the **PVA97NK** data.

2.4 Exploring a Data Source

As stated in Chapter 1 and noted in Section 2.3, the task of data assembly largely occurs outside of SAS Enterprise Miner. However, it is quite worthwhile to explore and validate your data's content. By assaying the prepared data, you substantially reduce the chances of erroneous results in your analysis, and you can gain insights graphically into associations between variables.

In this exploration, you should look for sampling errors, unexpected or unusual data values, and interesting variable associations.

The next demonstrations illustrate SAS Enterprise Miner tools that are useful for data validation.



Exploring Source Data

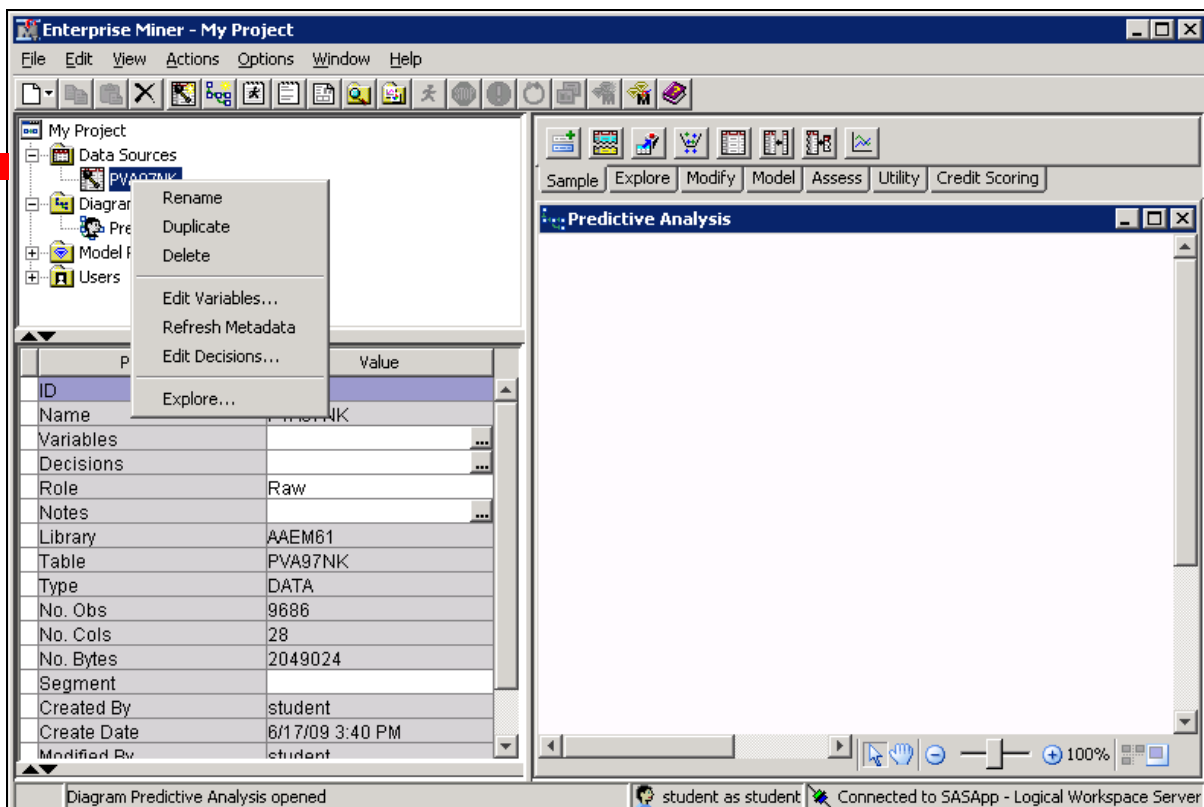
SAS Enterprise Miner can construct interactive plots to help you explore your data. This demonstration shows the basic features of the Explore window. These include the following:

- opening the Explore window
- changing the Explore window sample size
- creating a histogram for a single variable
- changing graph properties for a histogram
- changing chart axes
- adding a missing bin to a histogram
- adding plots to the Explore window
- exploring variable associations

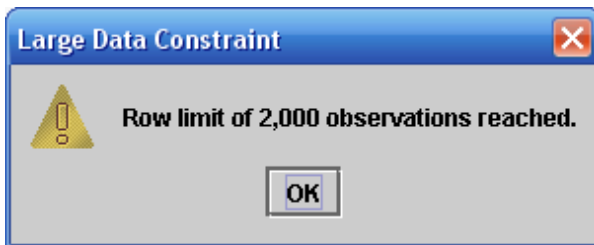
Opening the Explore Window

There are several ways to access the Explore window. Use these steps to open the Explore window through the Project panel.

1. Open the Data Sources folder in the Project panel and right-click the data source of interest. The Data Source Option menu appears.

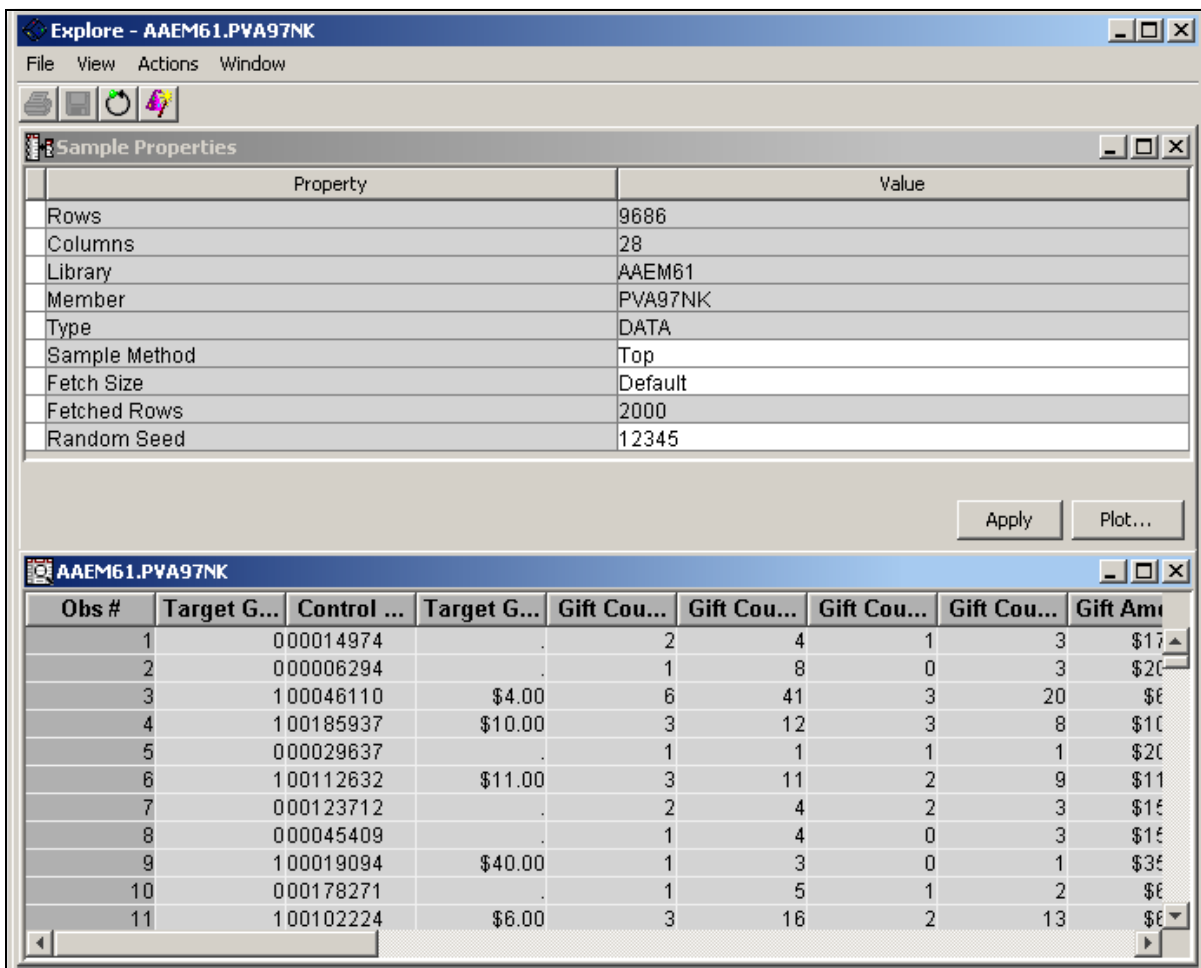


2. Select **Explore...** from the Data Source Option menu. If you are using an installation of SAS Enterprise Miner with factory default settings, the Large Data Constraint alert window opens.



By default, a maximum of 2000 observations are transferred from the SAS Foundation Server to the SAS Enterprise Miner client. This represents about a fifth of the **PVA97NK** table.


3. Select **OK** to close the Large Data Constraint alert window. The Explore – AAEM61.PVA97NK window opens.



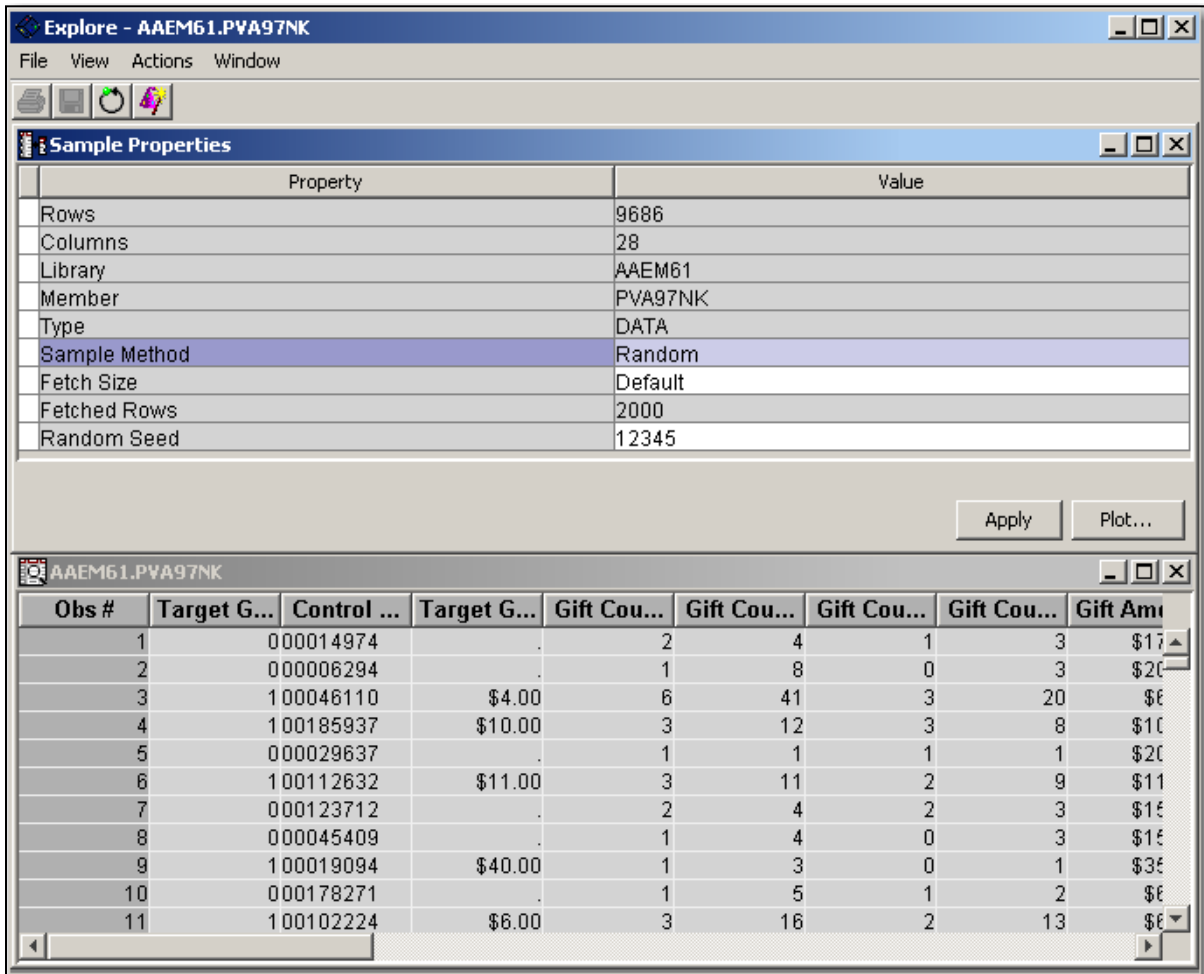
The Explore window features a 2000-observation sample from the **PVA97NK** data source. Sample properties are shown in the top half of the window and a data table is shown in the bottom half.

Changing the Explore Sample Size

The Sample Method property indicates that the sample is drawn from the **top** (first 2000 rows) of the data set. Use these steps to change the sampling properties in the Explore window.

 Although selecting a sample through this method is quick to execute, fetching the top rows of a table might not produce a representative sample of the table.


1. Left-click the **Sample Method** value field. The Option menu lists two choices: Top (the current setting) and Random.
2. Select **Random** from the Option menu.



Property	Value
Rows	9686
Columns	28
Library	AAEM61
Member	PVA97NK
Type	DATA
Sample Method	Random
Fetch Size	Default
Fetch Size	2000
Random Seed	12345

Obs #	Target G...	Control ...	Target G...	Gift Cou...	Gift Cou...	Gift Cou...	Gift Cou...	Gift Am...
1	000014974			2	4	1	3	\$17
2	000006294			1	8	0	3	\$20
3	100046110		\$4.00	6	41	3	20	\$8
4	100185937		\$10.00	3	12	3	8	\$10
5	000029637			1	1	1	1	\$20
6	100112632		\$11.00	3	11	2	9	\$11
7	000123712			2	4	2	3	\$15
8	000045409			1	4	0	3	\$15
9	100019094		\$40.00	1	3	0	1	\$35
10	000178271			1	5	1	2	\$8
11	100102224		\$6.00	3	16	2	13	\$8

3. Select **Actions** ⇒ **Apply Sample Properties** from the Explore window menu. A new, random sample of 2000 observations is made. This 2000-row sample now has distributional properties that are similar to the original 9686 observation table. This gives you an idea about the general characteristics of the variables. If your goal is to examine the data for potential problems, it is wise to examine the entire data set.

 SAS Enterprise Miner enables you to increase the sample transferred to the client (up to a maximum of 30,000 observations). See the SAS Enterprise Miner Help file to learn how to increase this maximum value.

4. Select the **Fetch Size** property and select **Max** from the Option menu.
5. Select **Actions** ⇒ **Apply Sample Properties**. Because there are fewer than 30,000 observations, the entire **PVA97NK** table is transferred to the SAS Enterprise Miner client machine, as indicated by the Fetched Rows field.

The screenshot shows the SAS Enterprise Miner interface. The main window is titled 'Explore - AAEM61.PVA97NK'. Below the menu bar (File, View, Actions, Window) is a toolbar with icons for file operations. The 'Sample Properties' dialog box is open, displaying a table of properties and their values. The 'Fetch Size' property is set to 'Max', and the 'Fetched Rows' field shows 9686. Below the dialog box, the data table 'AAEM61.PVA97NK' is displayed, showing 11 observations with columns for 'Obs #', 'Target G...', 'Control ...', 'Target G...', 'Gift Cou...', 'Gift Cou...', 'Gift Cou...', 'Gift Cou...', and 'Gift Am'.

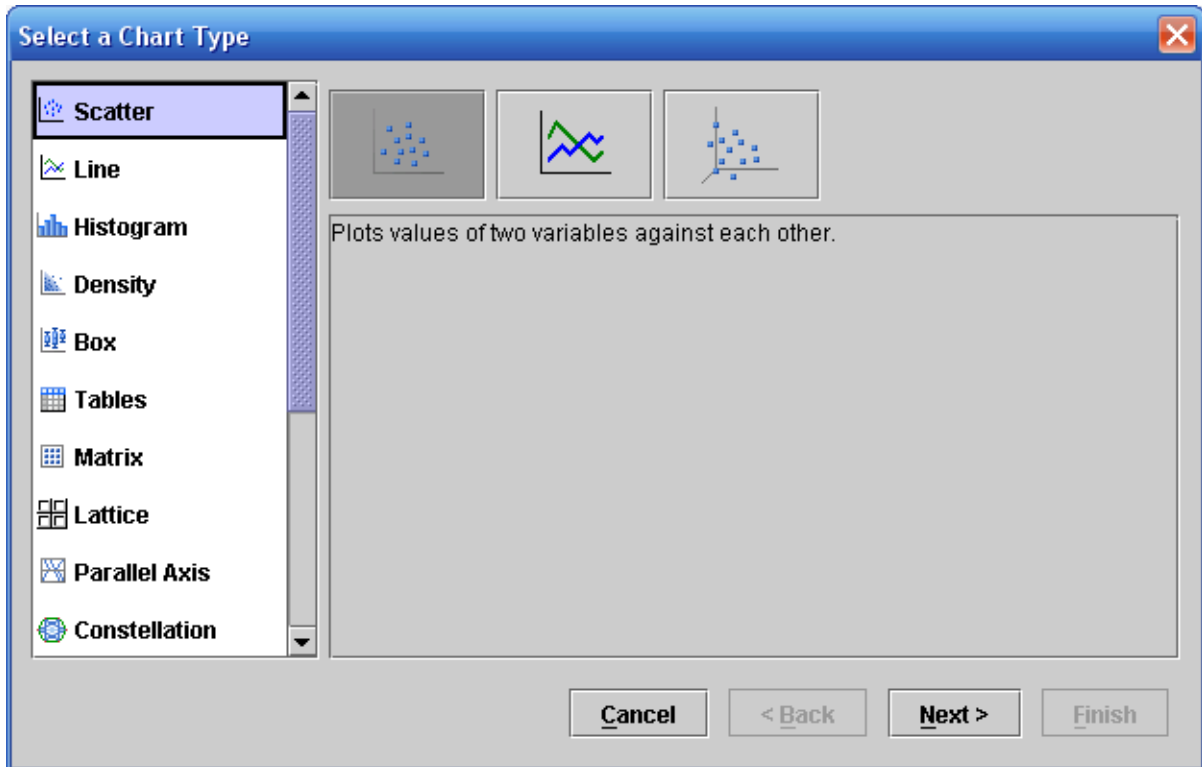
Property	Value
Rows	9686
Columns	28
Library	AAEM61
Member	PVA97NK
Type	DATA
Sample Method	Random
Fetch Size	Max
Fetched Rows	9686
Random Seed	12345

Obs #	Target G...	Control ...	Target G...	Gift Cou...	Gift Cou...	Gift Cou...	Gift Cou...	Gift Am
1	000014974			2	4	1	3	\$17
2	000006294			1	8	0	3	\$20
3	100046110		\$4.00	6	41	3	20	\$6
4	100185937		\$10.00	3	12	3	8	\$10
5	000029637			1	1	1	1	\$20
6	100112632		\$11.00	3	11	2	9	\$11
7	000123712			2	4	2	3	\$15
8	000045409			1	4	0	3	\$15
9	100019094		\$40.00	1	3	0	1	\$35
10	000178271			1	5	1	2	\$6
11	100102224		\$6.00	3	16	2	13	\$6

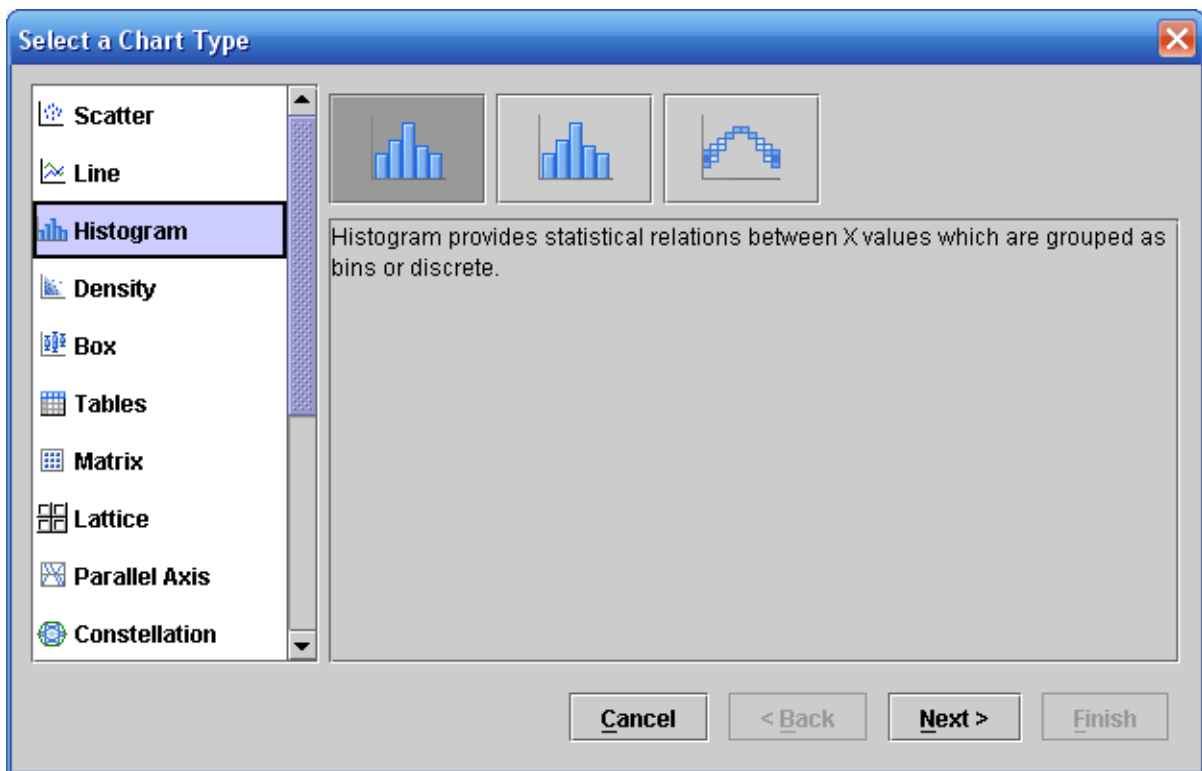
Creating a Histogram for a Single Variable

While you can use the Explore window to browse a data set, its primary purpose is to create statistical analysis plots. Use these steps to create a histogram in the Explore window.

1. Select **Actions** ⇒ **Plot** from the Explore window menu. The Chart wizard opens to the Select a Chart Type step.



The Chart wizard enables the construction of a multitude of analysis charts. This demonstration focuses on histograms.

2. Select **Histogram**.

Histograms are useful for exploring the distribution of values in a variable.

3. Select **Next >**. The Chart wizard proceeds to the next step, Select Chart Roles.

Select Chart Roles

Missing required roles: X.

Use default assignments

▲ Variable	Role	Type	Description	Format
DEMAGE		Numeric	Age	
DEMCLUSTER		Character	Demographic Cluster	
DEMGENDER		Character	Gender	
DEMHOMEOOWNER		Character	Home Owner	
DEMMEDHOMEVALUE		Numeric	Median Home Value ...	DOLLAR11
DEMMEDINCOME		Numeric	Median Income Region	DOLLAR11
DEMPCTVETERANS		Numeric	Percent Veterans Re...	
GIFTAVG36		Numeric	Gift Amount Average...	DOLLAR9.2
GIFTAVGALL		Numeric	Gift Amount Average...	DOLLAR9.2
GIFTAVGCARD36		Numeric	Gift Amount Average...	DOLLAR9.2
GIFTAVGLAST		Numeric	Gift Amount Last	DOLLAR9.2
GIFTCNT36		Numeric	Gift Count 36 Months	

Response statistic: Frequency ▼

☐ Allow multiple role assignments

Cancel < Back Next > Finish

To draw a histogram, one variable must be selected to have the role X.

4. Select **Role** ⇒ **X** for the **DEMAge** variable.

Select Chart Roles

Use default assignments

▲ Variable	Role	Type	Description	Format
DEMAge	X	Numeric	Age	
DEMCLUSTER		Character	Demographic Cluster	
DEMGENDER		Character	Gender	
DEMHOMEOOWNER		Character	Home Owner	
DEMMEDHOMEVALUE		Numeric	Median Home Value ...	DOLLAR11
DEMMEDINCOME		Numeric	Median Income Region	DOLLAR11
DEMPCTVETERANS		Numeric	Percent Veterans Re...	
GIFTAVG36		Numeric	Gift Amount Average...	DOLLAR9.2
GIFTAVGALL		Numeric	Gift Amount Average...	DOLLAR9.2
GIFTAVGCARD36		Numeric	Gift Amount Average...	DOLLAR9.2
GIFTAVGLAST		Numeric	Gift Amount Last	DOLLAR9.2
GIFTCNT36		Numeric	Gift Count 36 Months	

Response statistic: Frequency ▼

☐ Allow multiple role assignments

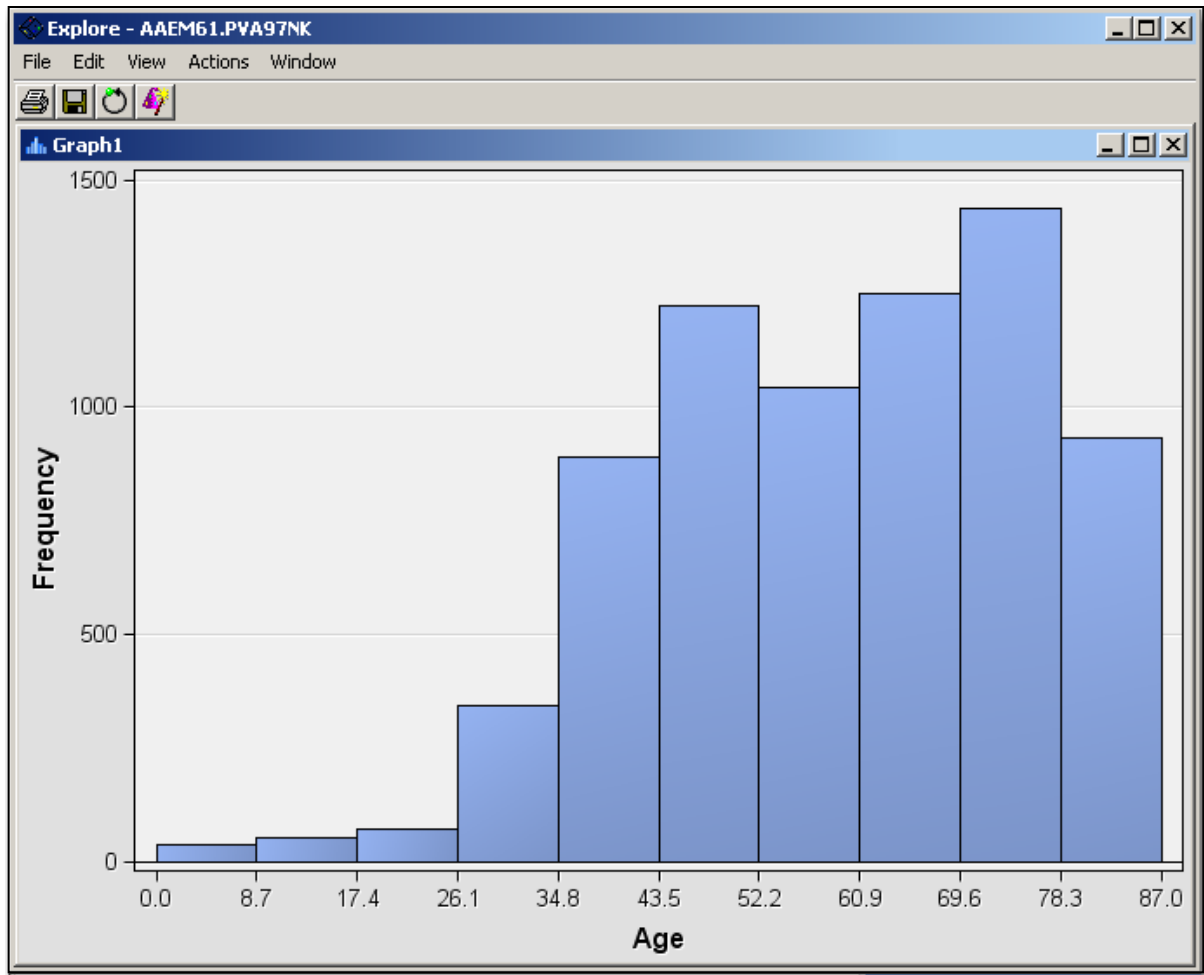
Cancel < Back Next > Finish

The Chart wizard is ready to make a histogram of the **DEMAge** variable.

5. Select **Finish**. The Explore window is filled with a histogram of the **DEMAge** variable.



Variable descriptions, rather than variable names, are used to label the axes of plots in the Explore window.



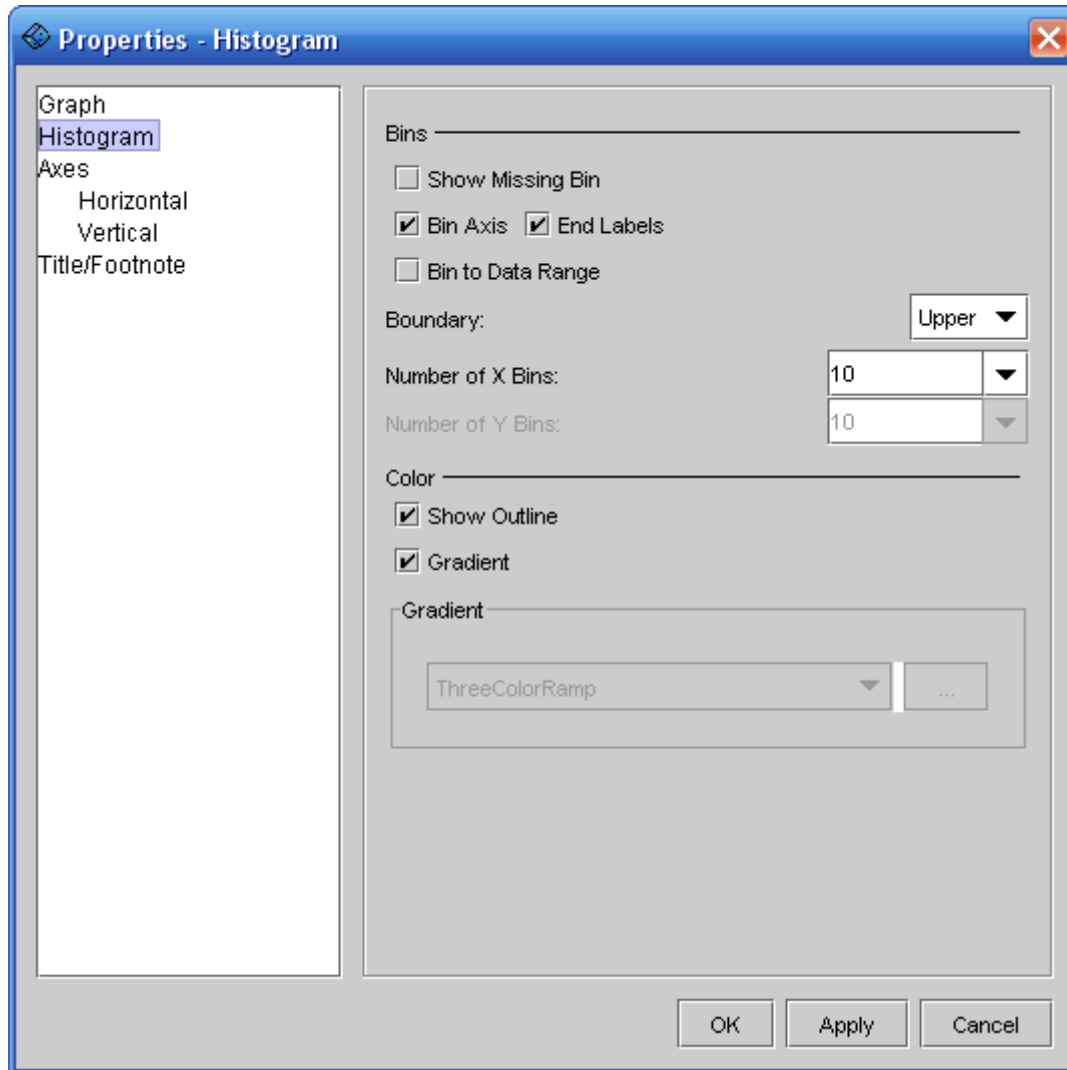
Axes in Explore window plots are chosen to range from the minimum to the maximum values of the plotted variable. Here you can see that **Age** has a minimum value of 0 and a maximum value of 87. The mode occurs in the ninth bin, which ranges between about 70 and 78. **Frequency** tells you that there are about 1400 observations in this range.

Changing the Graph Properties for a Histogram

By default, a histogram in SAS Enterprise Miner has 10 bins and is scaled to show the entire range of data. Use these steps to change the number of bins in a histogram and change the range of the axes.

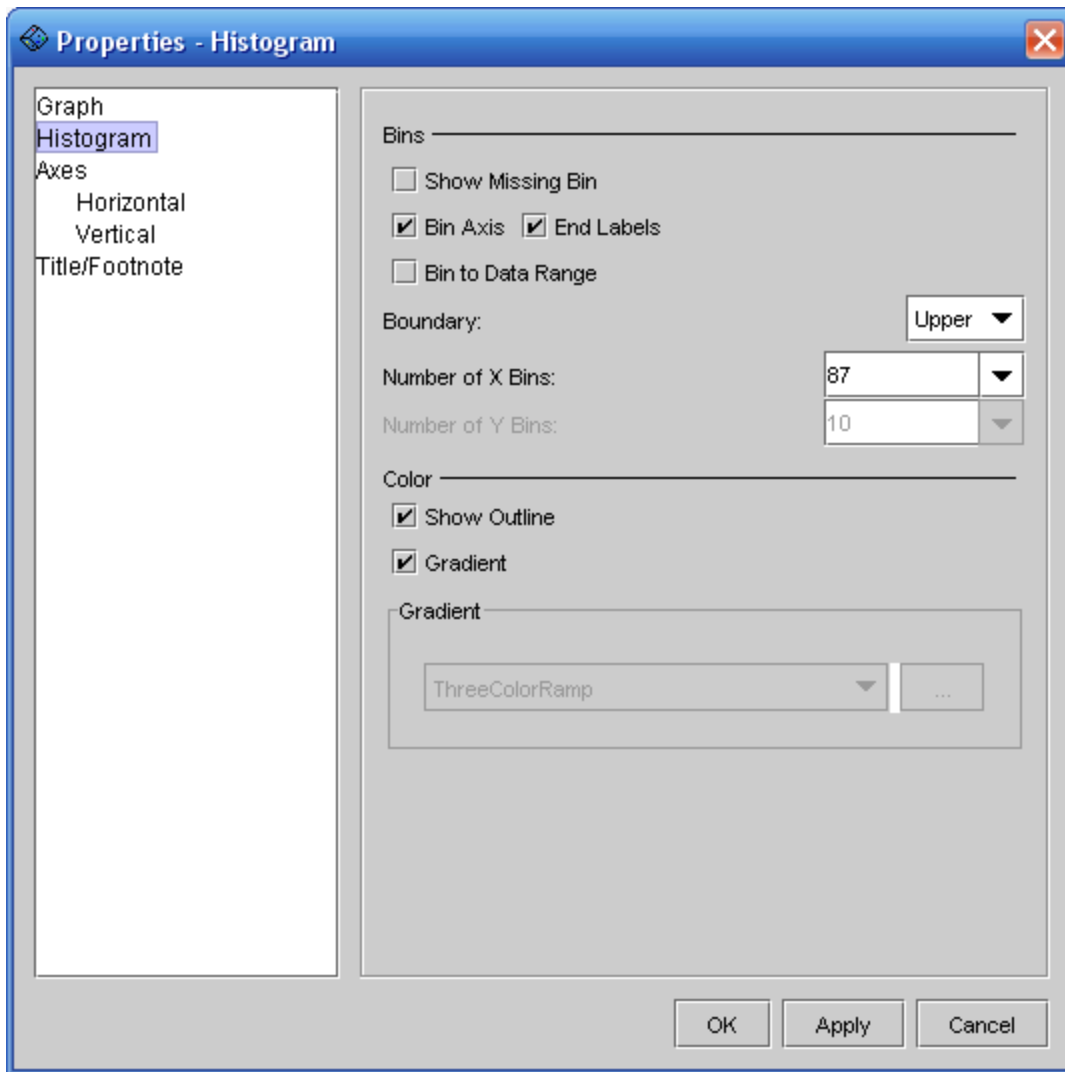
While the default bin size is sufficient to show the general shape of a variable's distribution, it is sometimes useful to increase the number of bins to improve the histogram's resolution.

1. Right-click in the data area of the **Age** histogram and select **Graph Properties...** from the Option menu. The Properties-Histogram window opens.



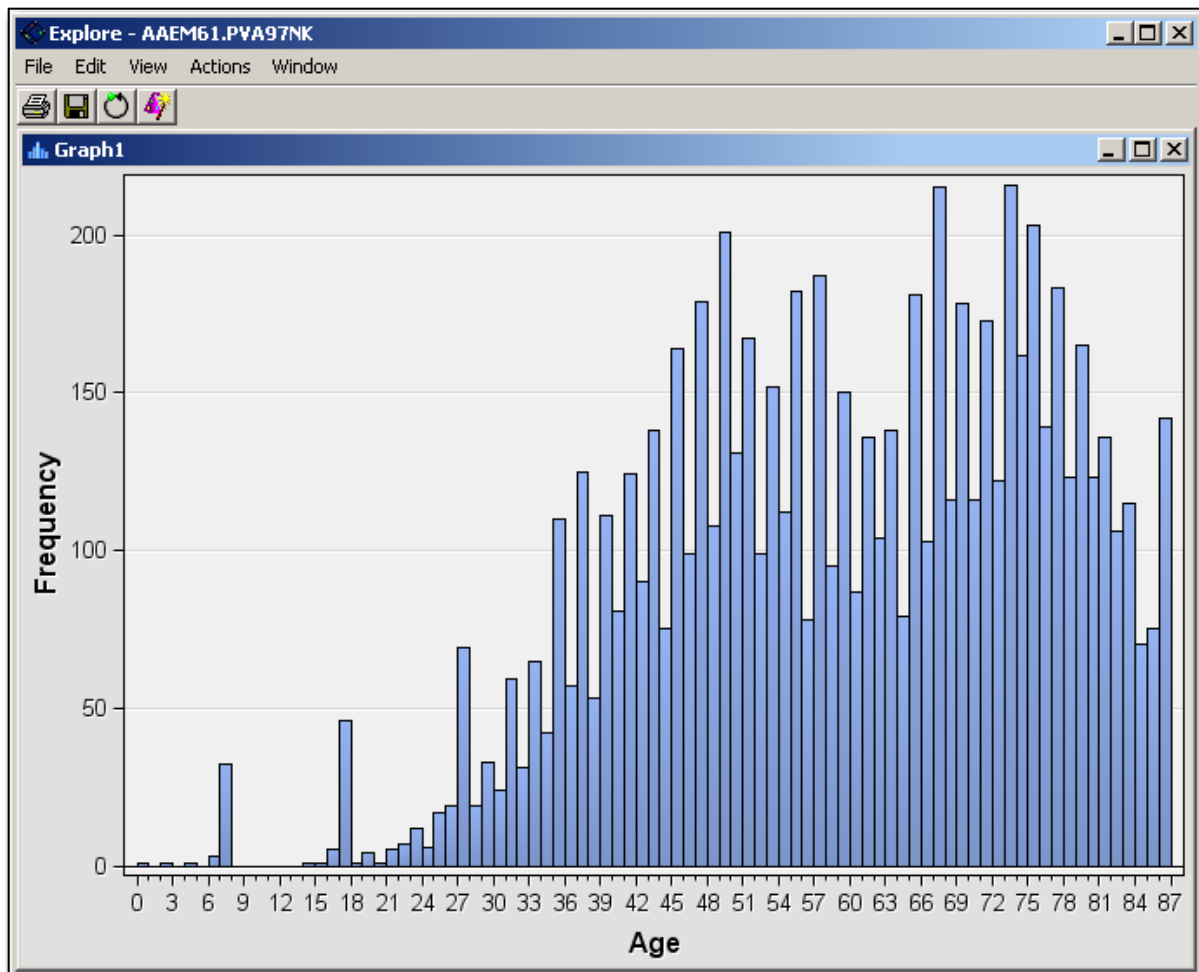
This window enables you to change the appearance of your charts. For histograms, the most important appearance property (at least in a statistical sense) is the number of bins.

2. Type **87** into the Number of X Bins field.



Because **Age** is integer-valued and the original distribution plot had a maximum of 87, there will be one bin per possible **Age** value.

3. Select **OK**. The Explore window reopens and shows many more bins in the **Age** histogram.

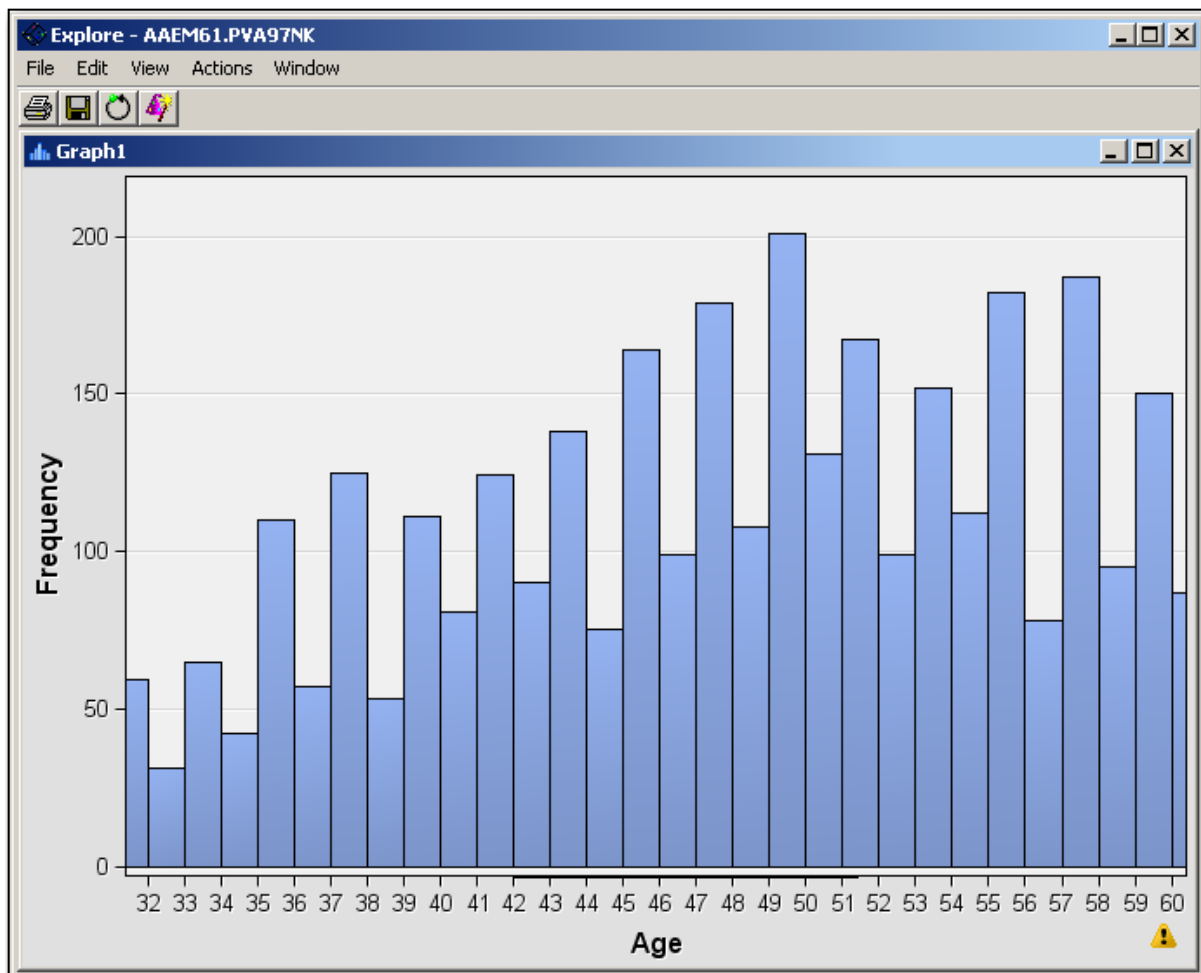


With the increase in resolution, unusual features become apparent in the **Age** variable. For example, there are unexpected spikes in the histogram at 10-year intervals, starting at **Age**=7. Also, you must question the veracity of ages below 18 for donors to the charity.

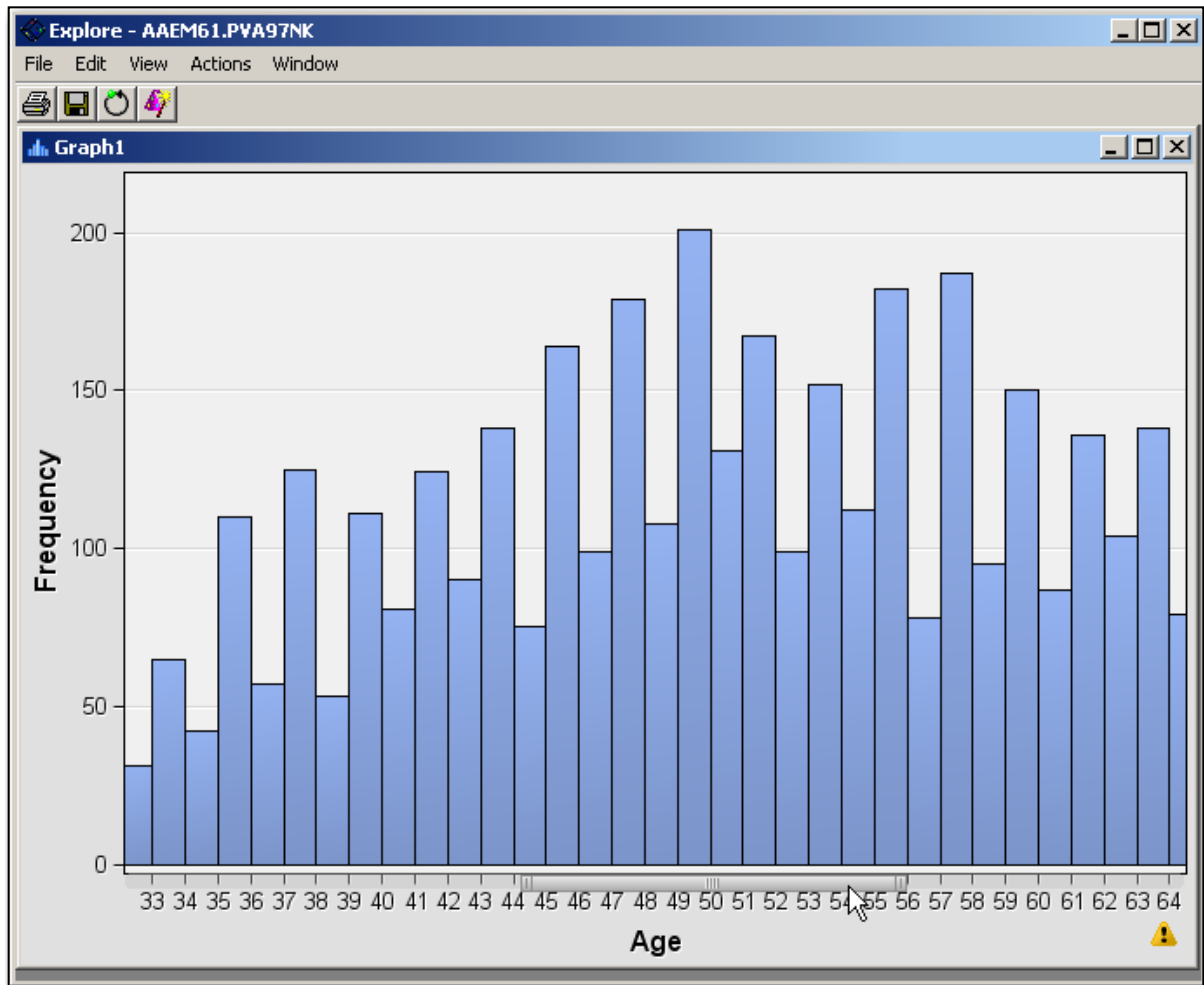
Changing Chart Axes

A very useful feature of the Explore window is the ability to zoom in on data of interest. Use the following steps to change chart axes in the Explore window:

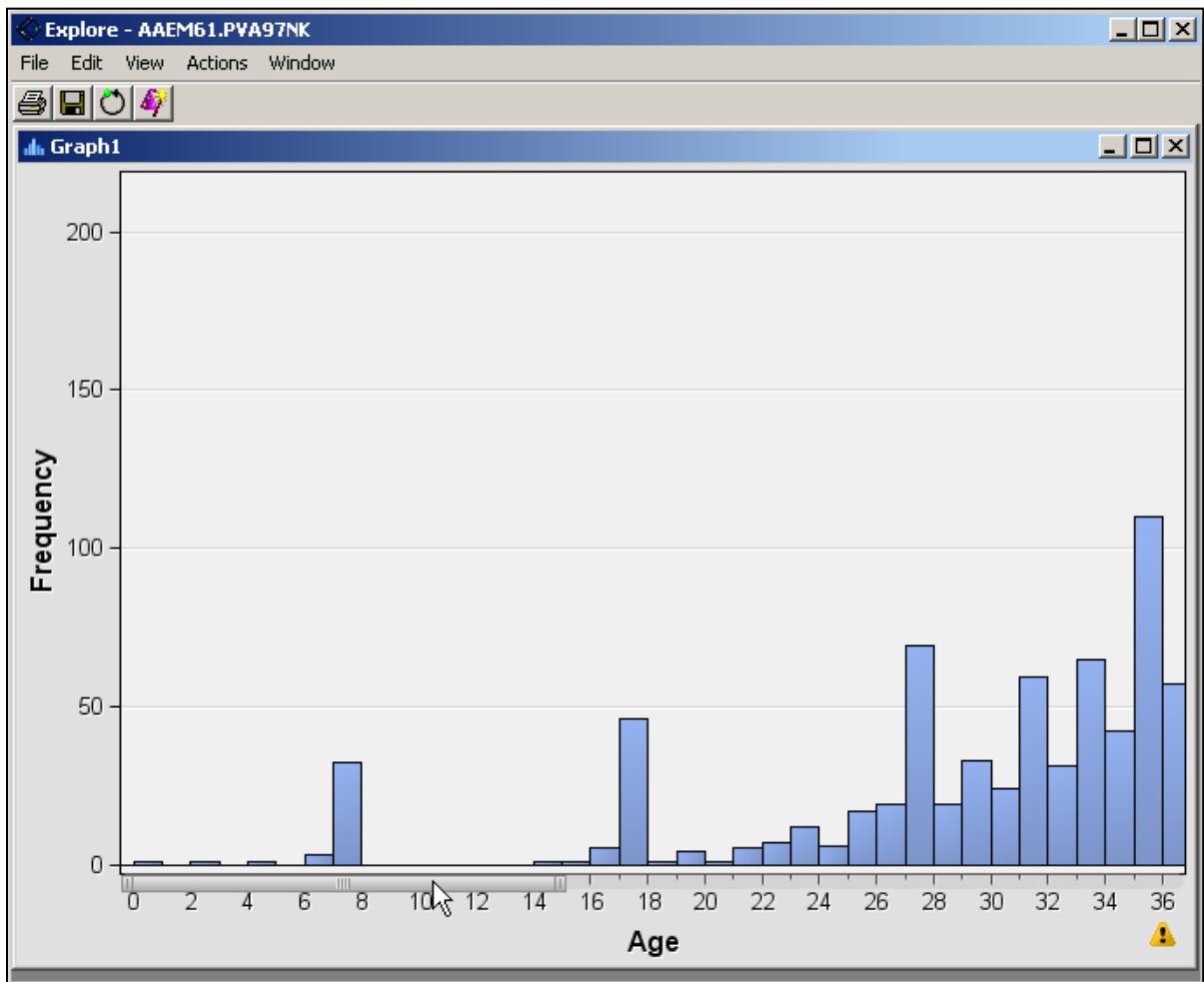
1. Position your cursor under the histogram until the cursor appears as a magnifying glass.
2. Clicking with your mouse and dragging the cursor to the right magnifies the horizontal axis of the histogram.



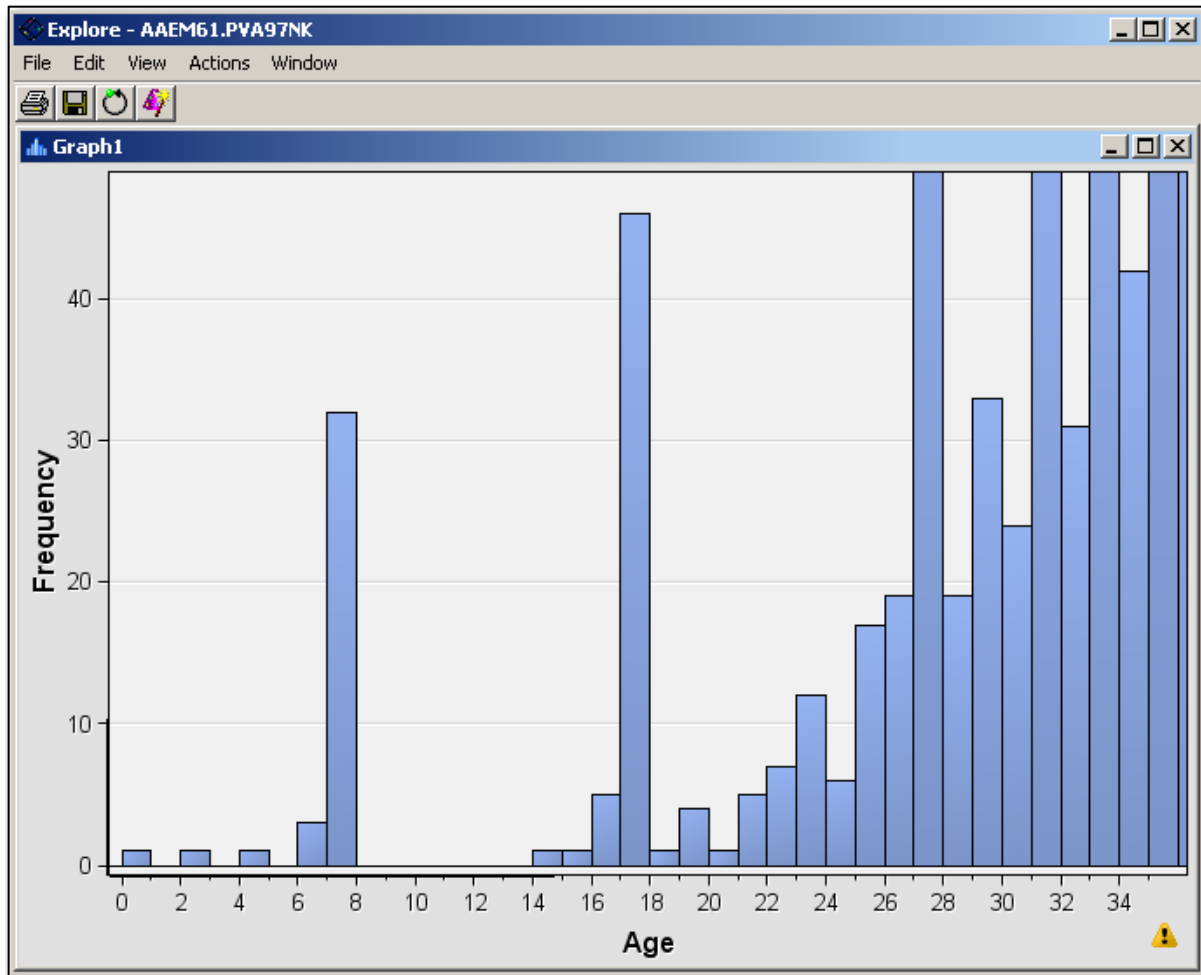
3. Position your cursor below the histogram but above the **Age** axis label. A horizontal scroll bar appears.




- Click and drag the scroll bar to the left to translate the horizontal axis to the left.

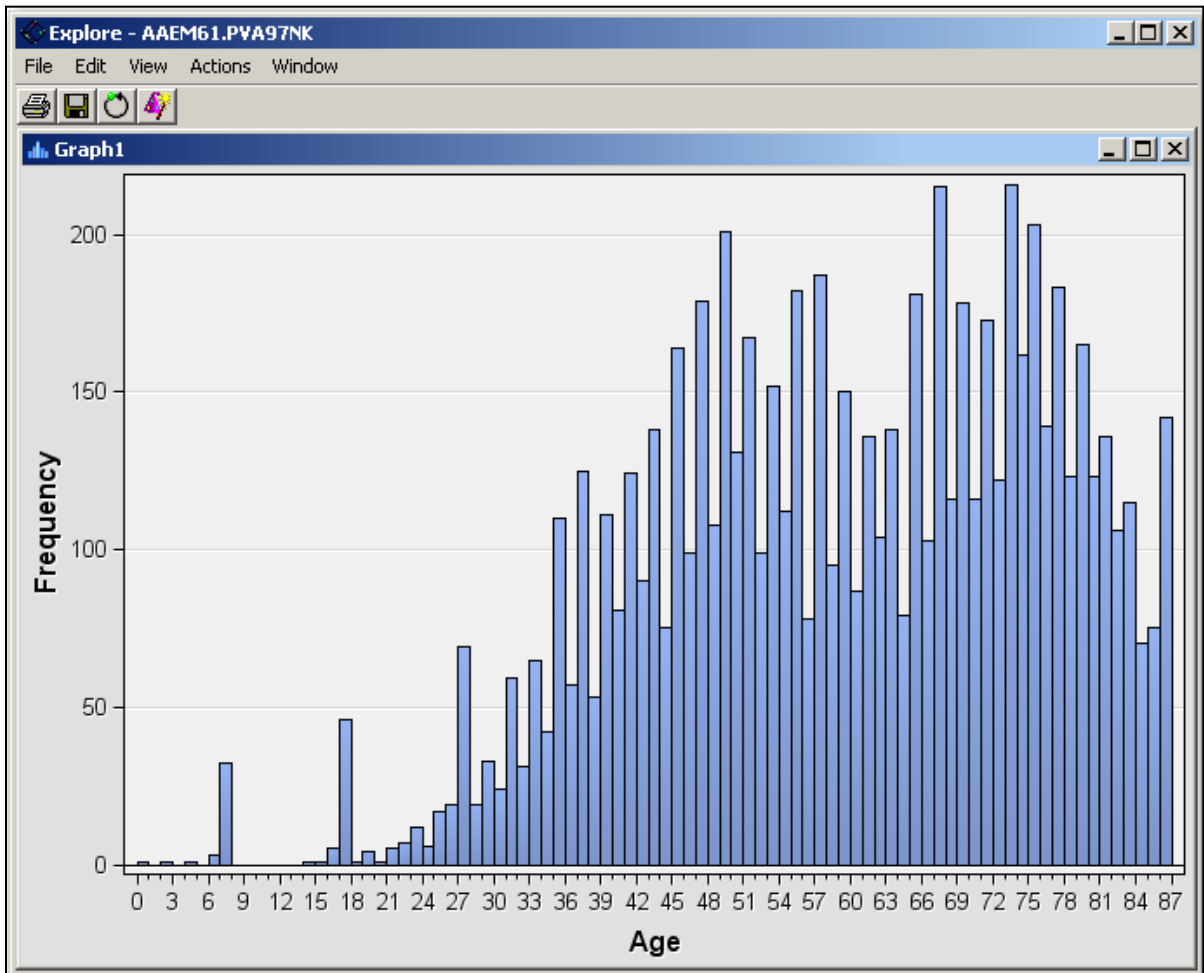


5. Position your cursor to the left of the histogram to make a similar adjustment to the vertical axis of the histogram.



At this resolution, you can see that there are approximately 40 observations with **Age** less than 8. A review of the data preparation process might help you determine whether these values are valid.

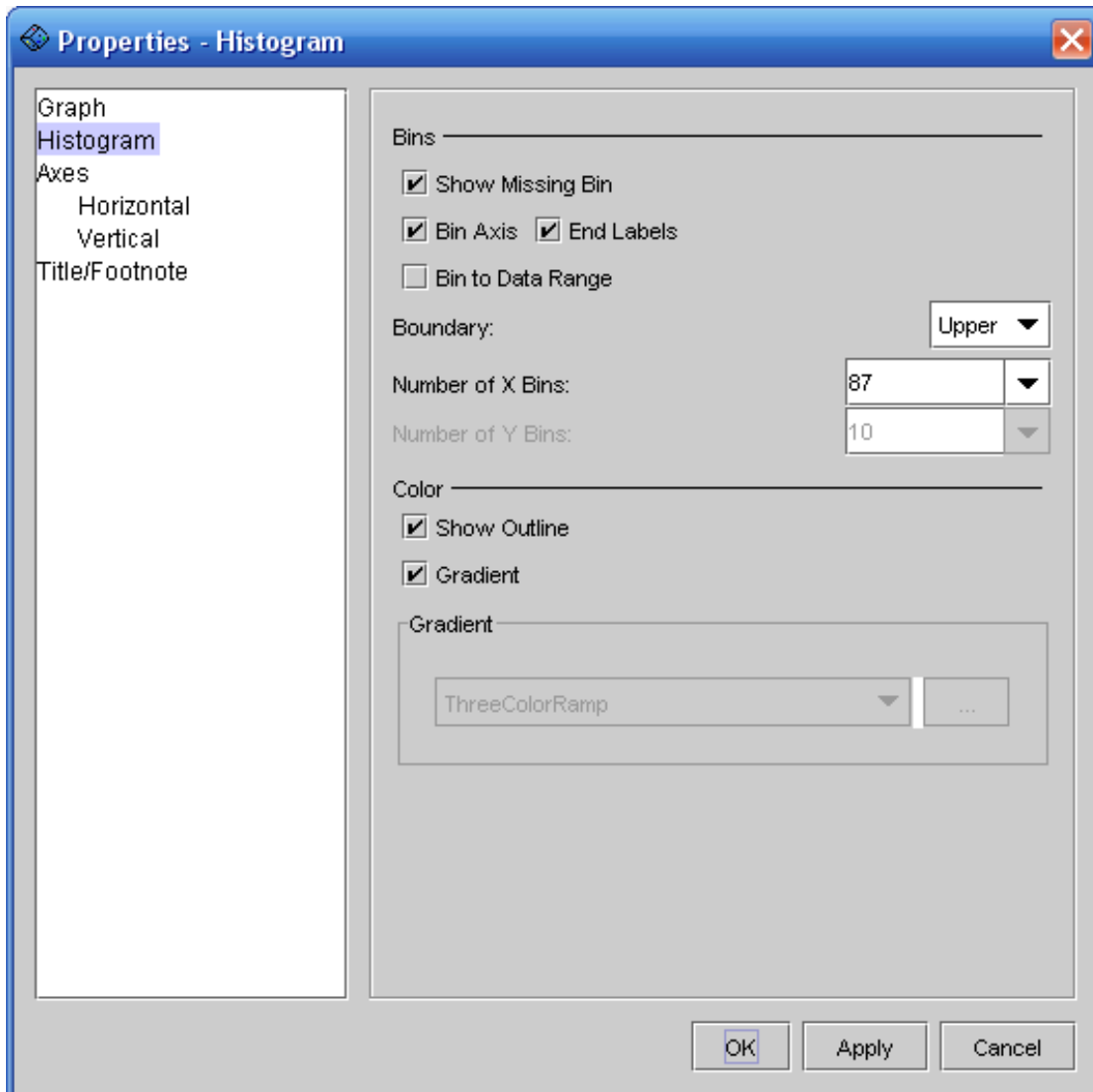
6. Select  (the yellow triangle) in the lower right corner of the Explore window to reset the axes to their original ranges.



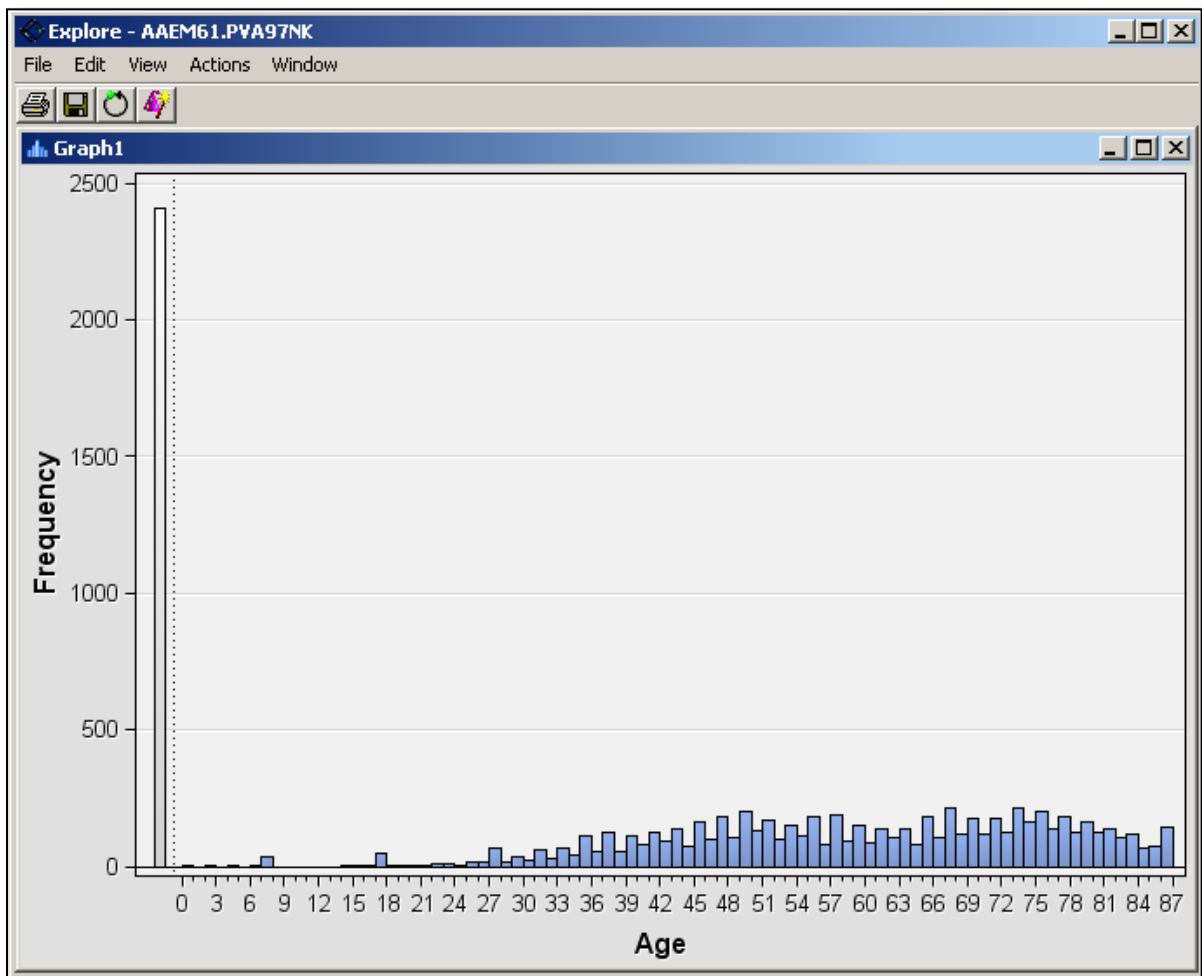
Adding a “Missing” Bin to a Histogram

Not all observations appear in the histogram for **Age**. There are many observations with missing values for this variable. Follow these steps to add a missing value bin to the **Age** histogram:

1. Right-click on the graph and select **Graph Properties...** from the Option menu.
2. Check the **Show Missing Bin** option.



3. Select **OK**. The **Age** histogram is modified to show a missing value bin.

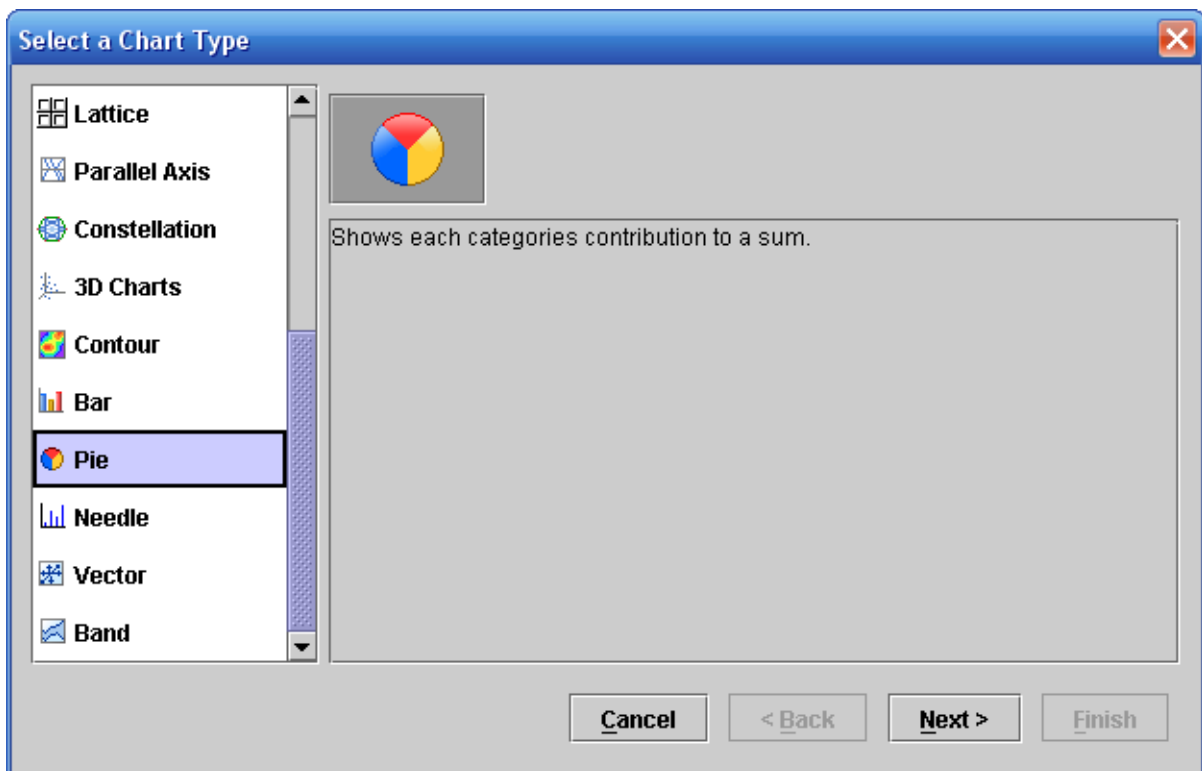


With the missing value bin added, it is easy to see that nearly a quarter of the observations are missing an **Age** value.

Adding Plots to the Explore Window

You can add other plots to the Explore window. Follow these steps to add a pie chart of the target variable.

1. Select **Actions** ⇒ **Plot** from the Explore window menu. The Chart wizard opens to the Select a Chart Type step.
2. Scroll down in the chart list and select a **Pie** chart.



3. Select **Next >**. The Chart wizard continues to the Select Chart Roles step.

Missing required roles: Category.

Use default assignments

Variable	Role	Type	Description	Format
DEMAGE		Numeric	Age	
DEMCLUSTER		Character	Demographic Cluster	
DEMGENDER		Character	Gender	
DEMHOMEOOWNER		Character	Home Owner	
DEMMEDHOMEVALUE		Numeric	Median Home Value ...	DOLLAR11
DEMMEDINCOME		Numeric	Median Income Region	DOLLAR11
DEMPCTVETERANS		Numeric	Percent Veterans Re...	
GIFTAVG36		Numeric	Gift Amount Average...	DOLLAR9.2
GIFTAVGALL		Numeric	Gift Amount Average...	DOLLAR9.2
GIFTAVGCARD36		Numeric	Gift Amount Average...	DOLLAR9.2
GIFTAVGLAST		Numeric	Gift Amount Last	DOLLAR9.2
GIFTCNT36		Numeric	Gift Count 36 Months	
GIFTCNTALL		Numeric	Gift Count All Months	
GIFTCNTCARD36		Numeric	Gift Count Card 36 M...	

☐ Allow multiple role assignments

Cancel < Back Next > Finish

The message at the top of the Select Chart Roles window states that a variable must be assigned the **Category** role.

4. Scroll the variable list and select **Role** ⇒ **Category** for the **TARGETB** variable.

Select Chart Roles ✕

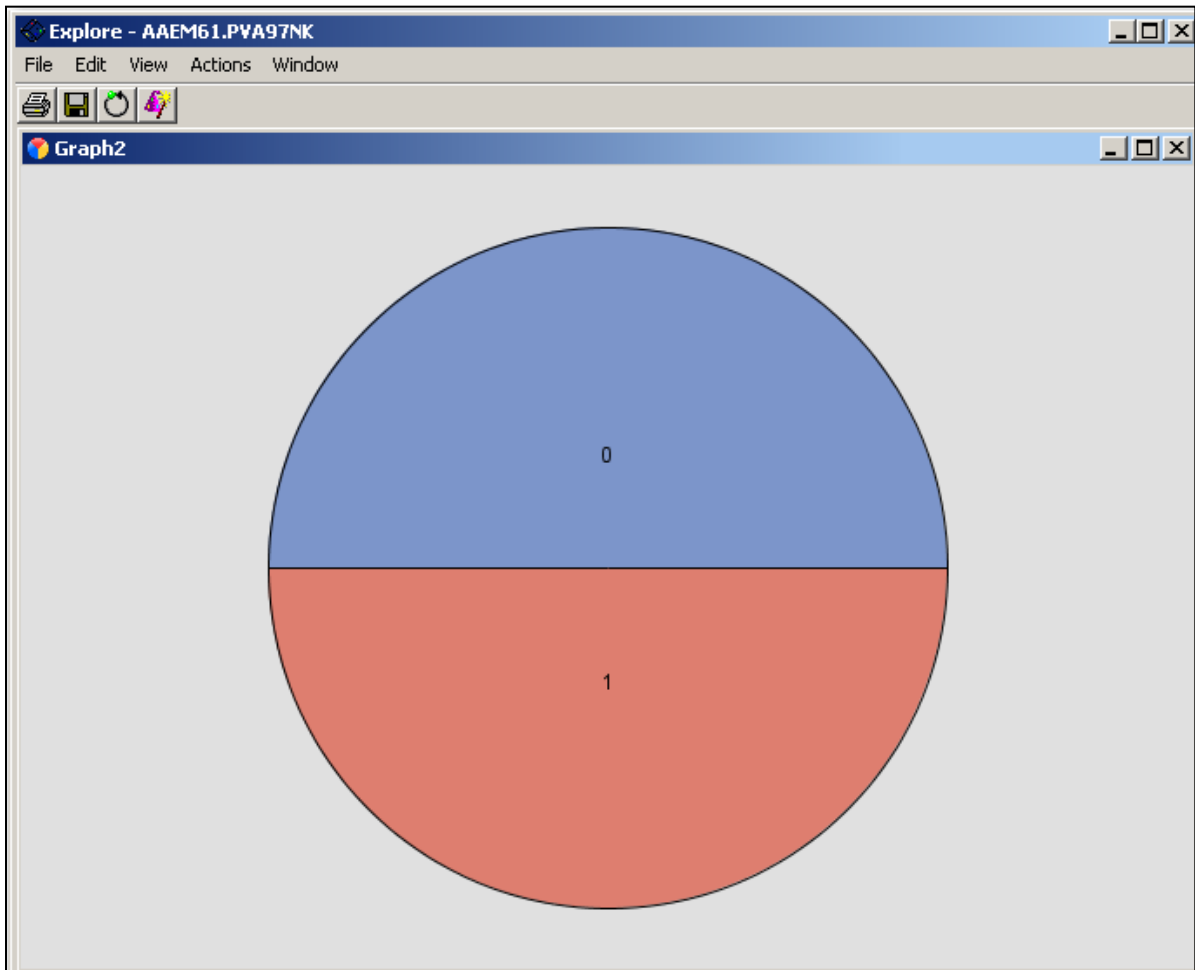
Use default assignments

▲ Variable	Role	Type	Description	Format
GIFTCNTCARDALL		Numeric	Gift Count Card All M...	
GIFTTIMEFIRST		Numeric	Times Since First Gift	
GIFTTIMELAST		Numeric	Time Since Last Gift	
ID		Character	Control Number	
PROMCNT12		Numeric	Promotion Count 12 M...	
PROMCNT36		Numeric	Promotion Count 36 M...	
PROMCNTALL		Numeric	Promotion Count All M...	
PROMCNTCARD12		Numeric	Promotion Count Card...	
PROMCNTCARD36		Numeric	Promotion Count Card...	
PROMCNTCARDALL		Numeric	Promotion Count Card...	
STATUSCAT96NK		Character	Status Category 96NK	
STATUSCATSTARALL		Numeric	Status Category Star ...	
TARGETB	Category	Numeric	Target Gift Flag	
TARGETD		Numeric	Target Gift Amount	DOLLAR9.2

☐ Allow multiple role assignments

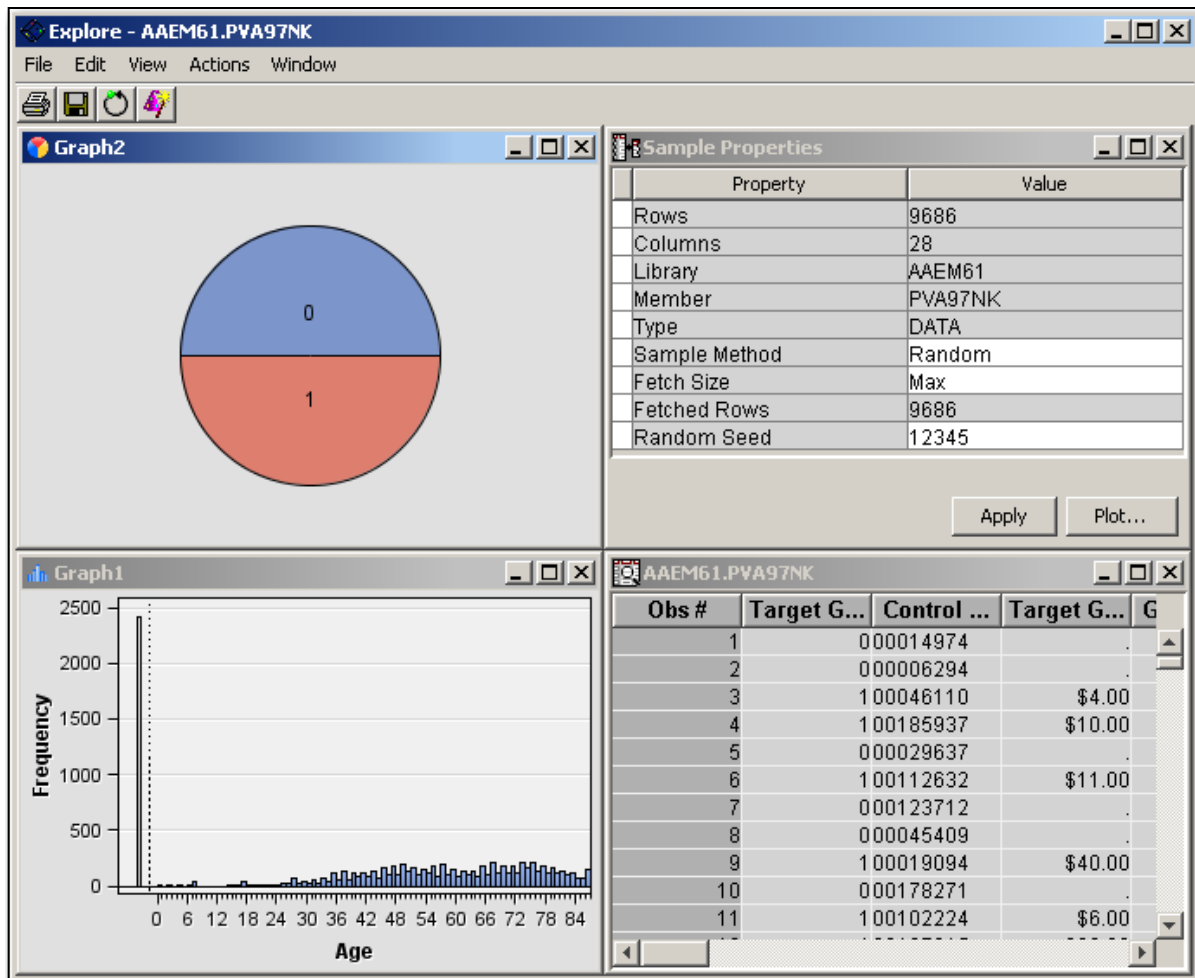
Cancel
< Back
Next >
Finish

5. Select **Finish** to create the pie chart for **TARGETB**.



The chart shows an equal number of cases for **TARGETB=0** (top) and **TARGETB=1** (bottom).

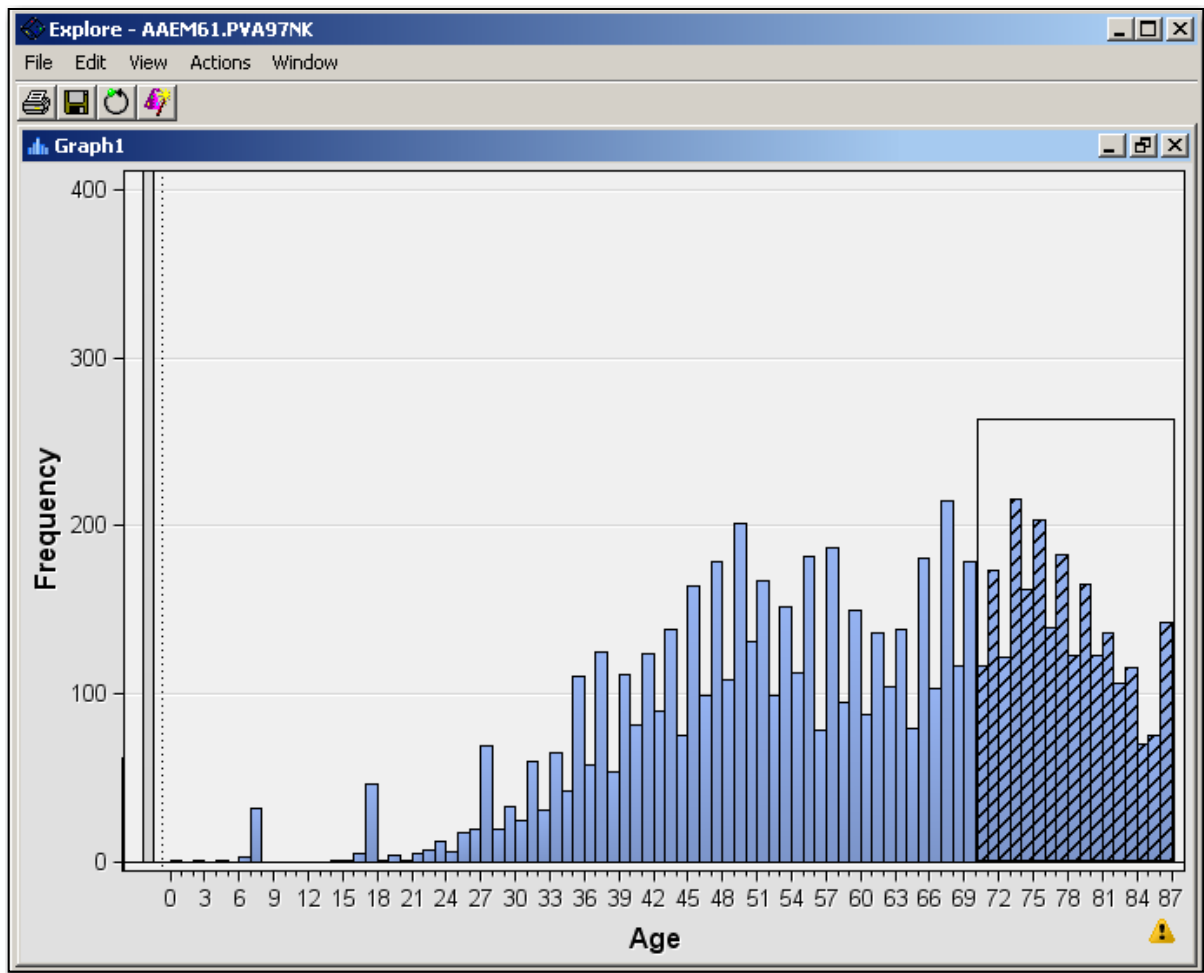
6. Select **Window** ⇒ **Tile** to simultaneously view all sub-windows of the Explore window.



Exploring Variable Associations

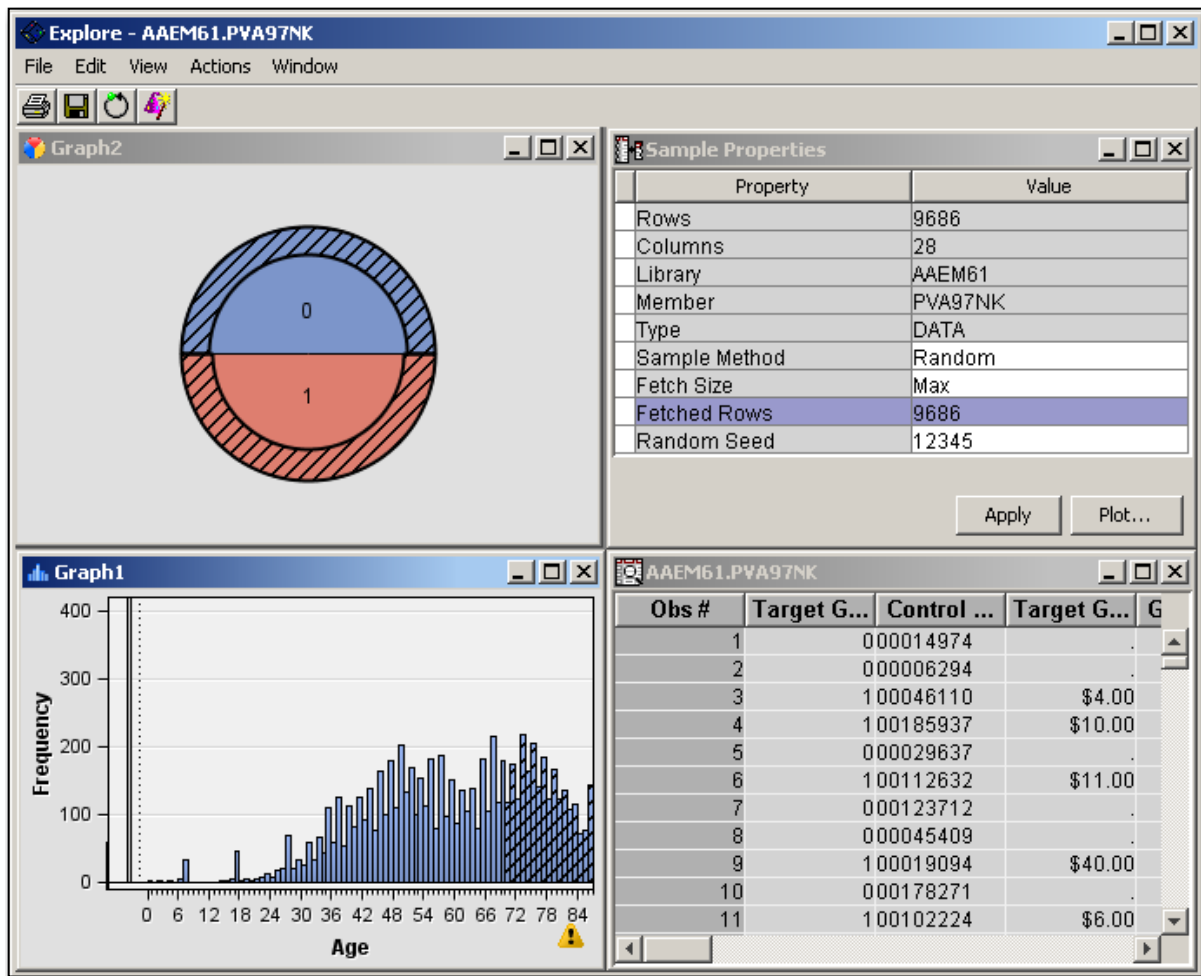
All elements of the Explore window are connected. By selecting a bar in one histogram, for example, corresponding observations in the data table and other plots are also selected. Follow these steps to use this feature to explore variable associations:

1. Double-click the **Age** histogram title bar so that it fills the Explore window.
2. Click and drag a rectangle in the Age histogram to select cases with **Age** in excess of 70 years.



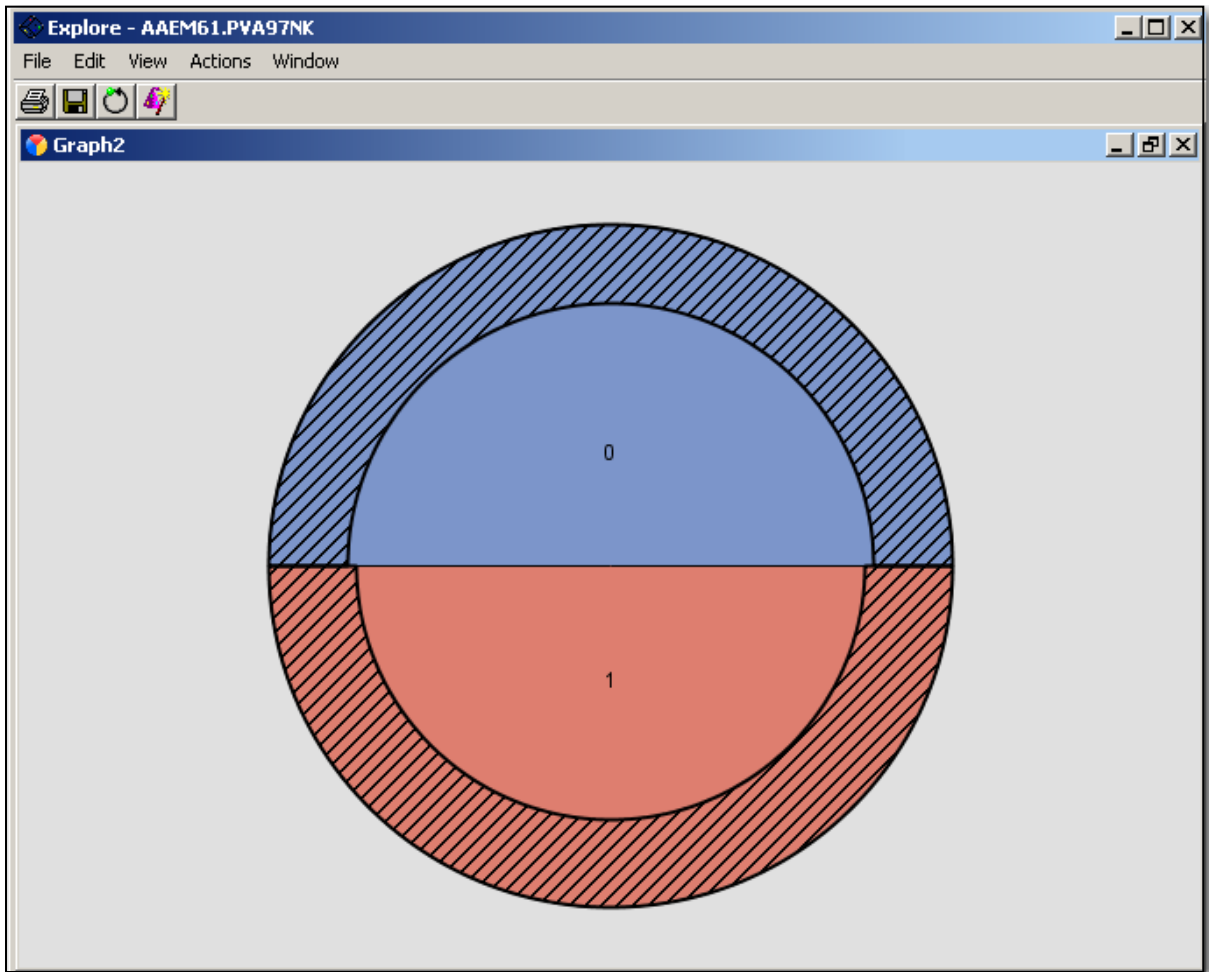
The selected cases are cross-hatched. (The vertical axis is rescaled to show the selection better.)

3. Double-click the **Age** histogram title bar. The tile display is restored.



Notice that part of the TargetB pie chart is selected. This selection shows the relative proportion of observations with **Age** greater than 70 that do and do not donate. Because the arc on the **TARGETB=1** segment is slightly thicker, it appears that there is a slightly higher number of donors than non-donors in this **Age** selection.

4. Double-click the **TargetB** pie chart title bar to confirm this observation.

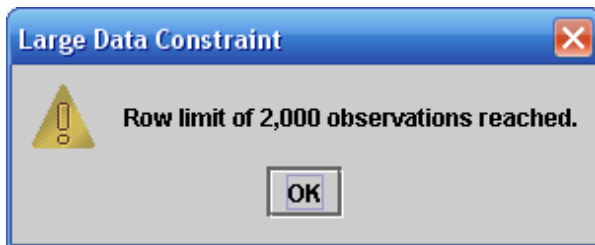


5. Close the Explore window to return to the SAS Enterprise Miner client interface screen.



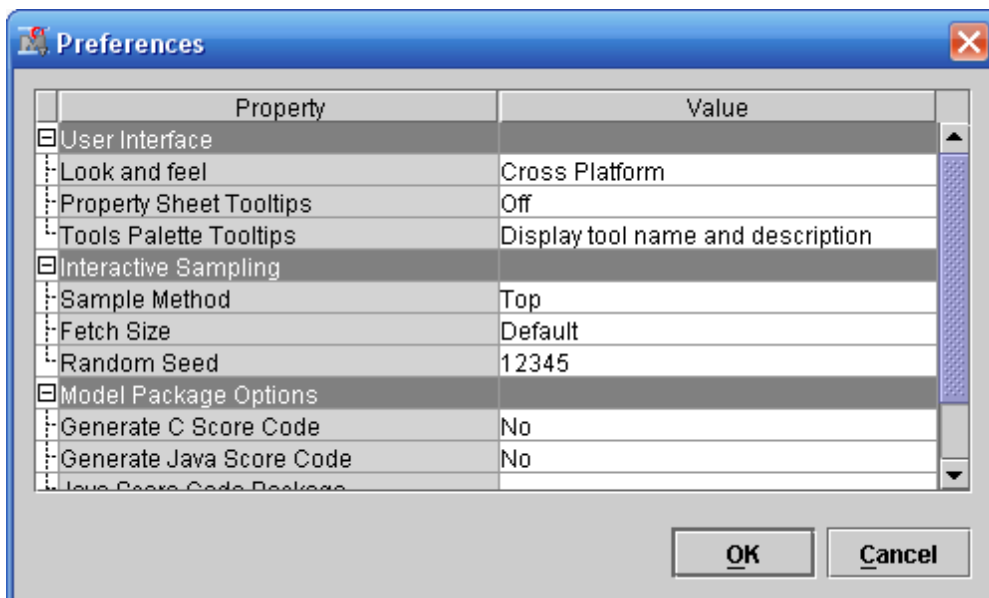
Changing the Explore Window Sampling Defaults

In the previous demonstration, you were alerted to the fact that only a sample of data was initially selected for use in the Explore window.

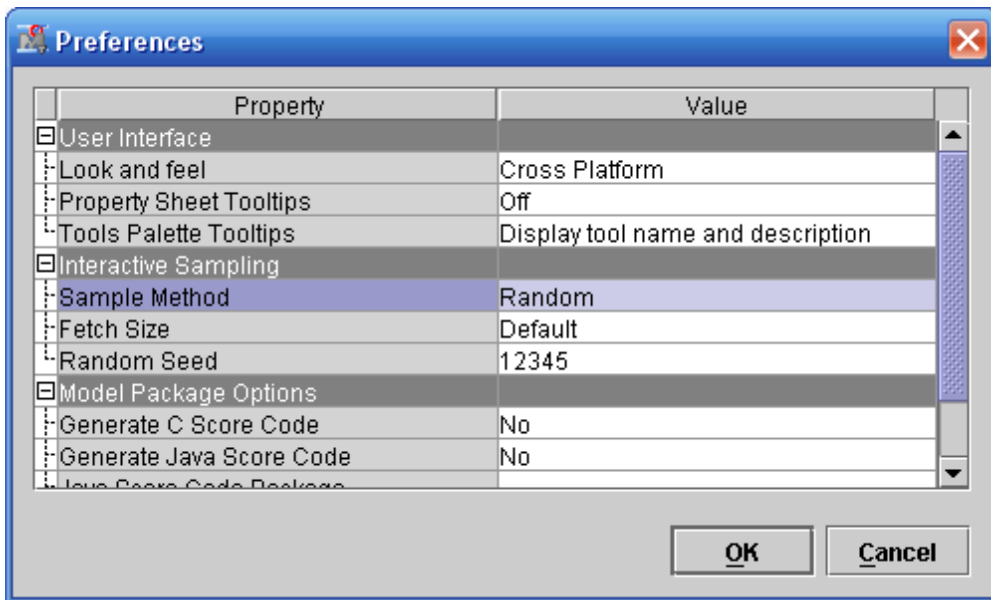


Follow these steps to change the preference settings of SAS Enterprise Miner to use a random sample or all of the data source data in the Explore window:

1. Select **Options** ⇒ **Preferences...** from the main menu. The Preferences window opens.

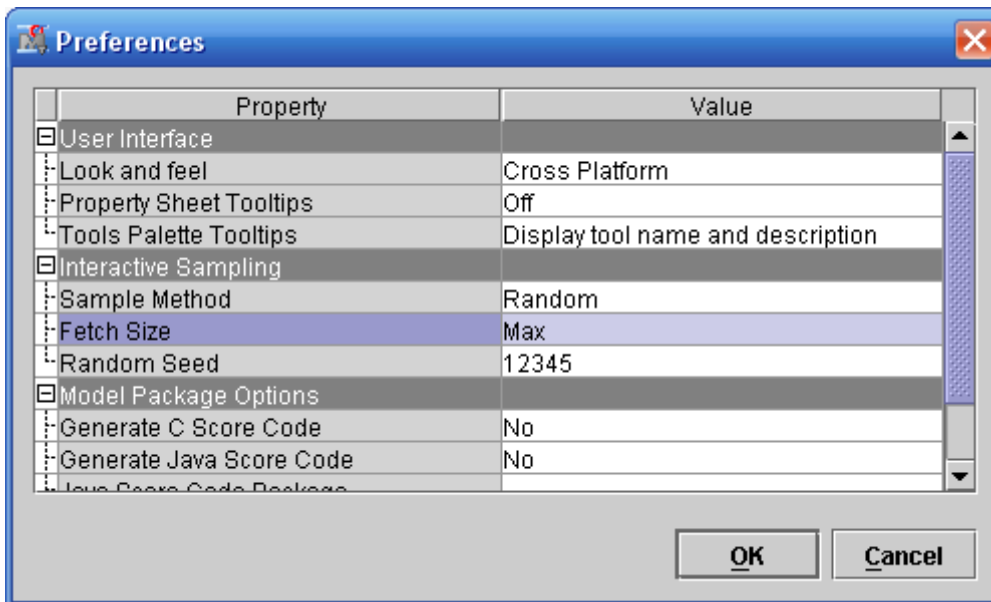


2. Select **Sample Method** ⇒ **Random**.



The random sampling method improves on the default method (at the top of the data set) by guaranteeing that the Explore window data is representative of the original data source. The only negative aspect is an increase in processing time for extremely large data sources.

3. Select **Fetch Size** ⇒ **Max**.



The Max fetch size enables a larger sample of data to be extracted for use in the Explore window.

If you use these settings, the Explore window uses the entire data set or a random sample of up to 30,000 observations (whichever is smaller).



Exercises

3. Using the Explore Window to Study Variable Distribution

Use the Explore window to study the distribution of the variable **Medium Income Region** (**DemMedIncome**). Answer the following questions:

- a. What is unusual about the distribution of this variable? _____

- b. What could cause this anomaly to occur? _____

- c. What do you think should be done to rectify the situation? _____



Modifying and Correcting Source Data

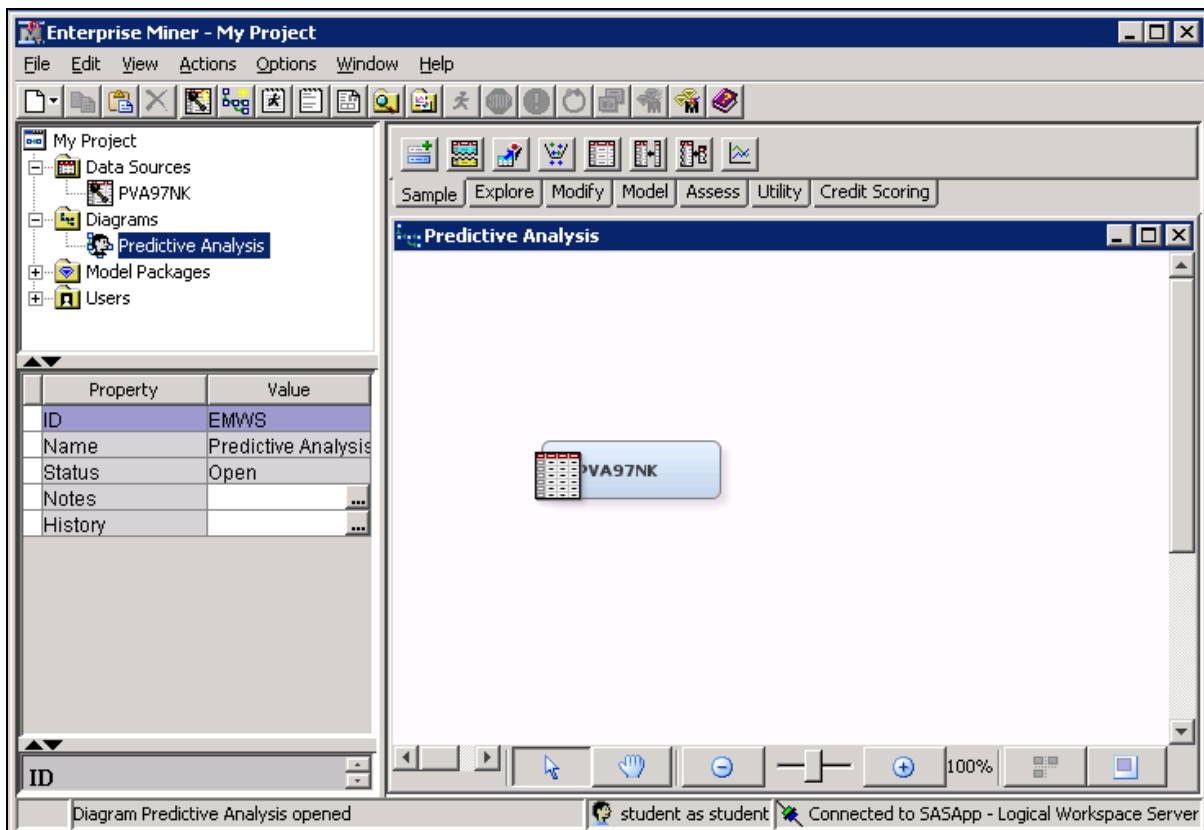
In the previous exercise, the **DemMedIncome** variable was seen to have an unusual spike at 0. This phenomenon often occurs in data extracted from a relational database table where 0 or another number is used as a substitute for the value missing or unknown. Clearly, having zero income is considerably different from having an unknown income. If you properly use the **income** variable in a predictive model, this discrepancy can be addressed.

This demonstration shows you how to replace a placeholder value for missing with a true missing value indicator. In this way, SAS Enterprise Miner tools can correctly respond to the true, but unknown, value. SAS Enterprise Miner includes several tools that you can use to modify the source data for your analysis. The following demonstrations show how to use the Replacement node to modify incorrect or improper values for a variable:

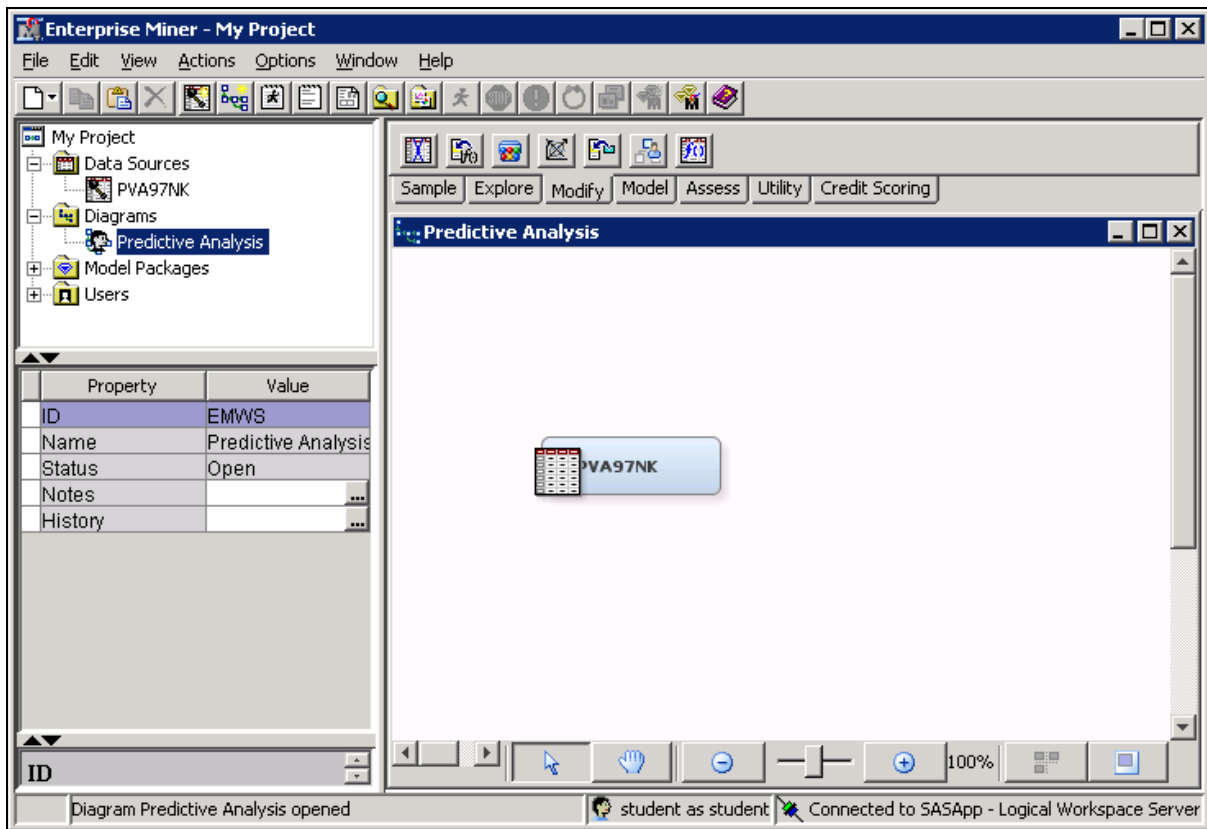
Process Flow Setup

Use the following steps to set up the process flow that will modify the **DemMedIncome** variable:

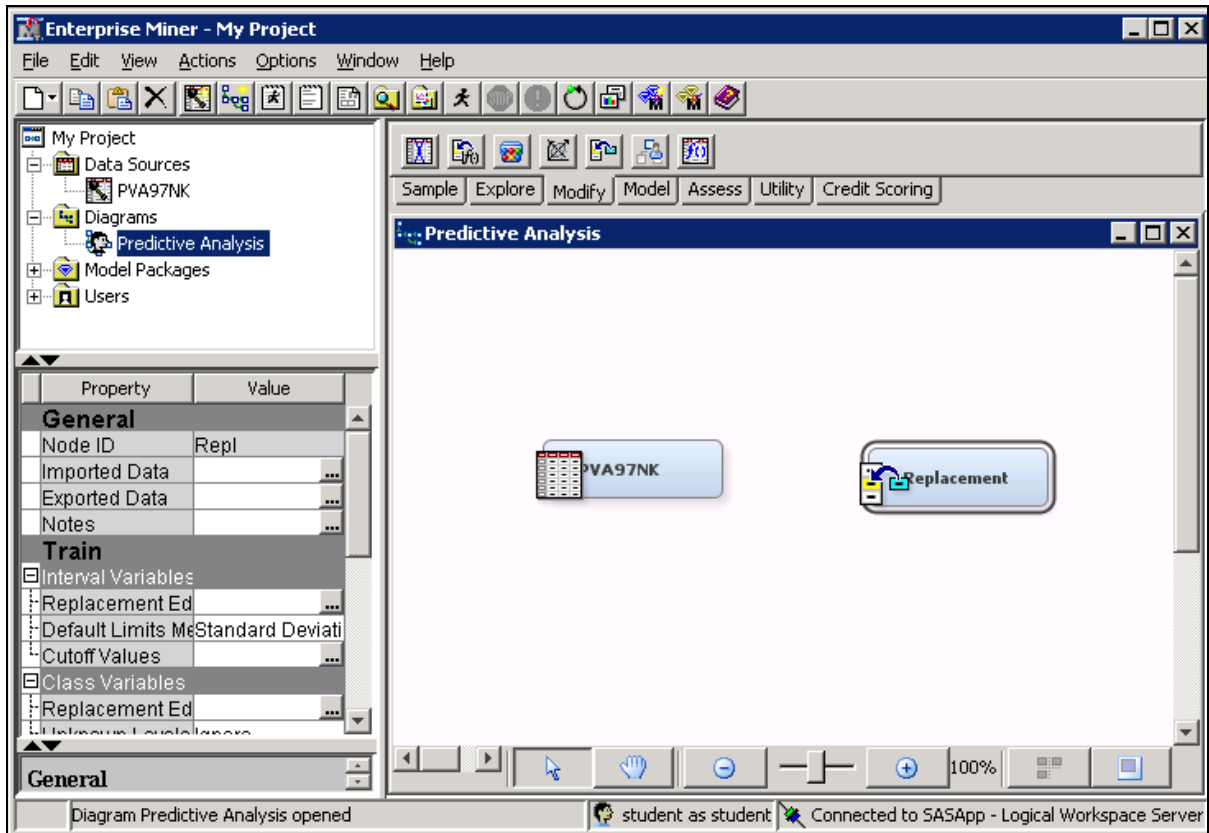
1. Drag the **PVA97NK** data source to the Predictive Analysis workspace window.



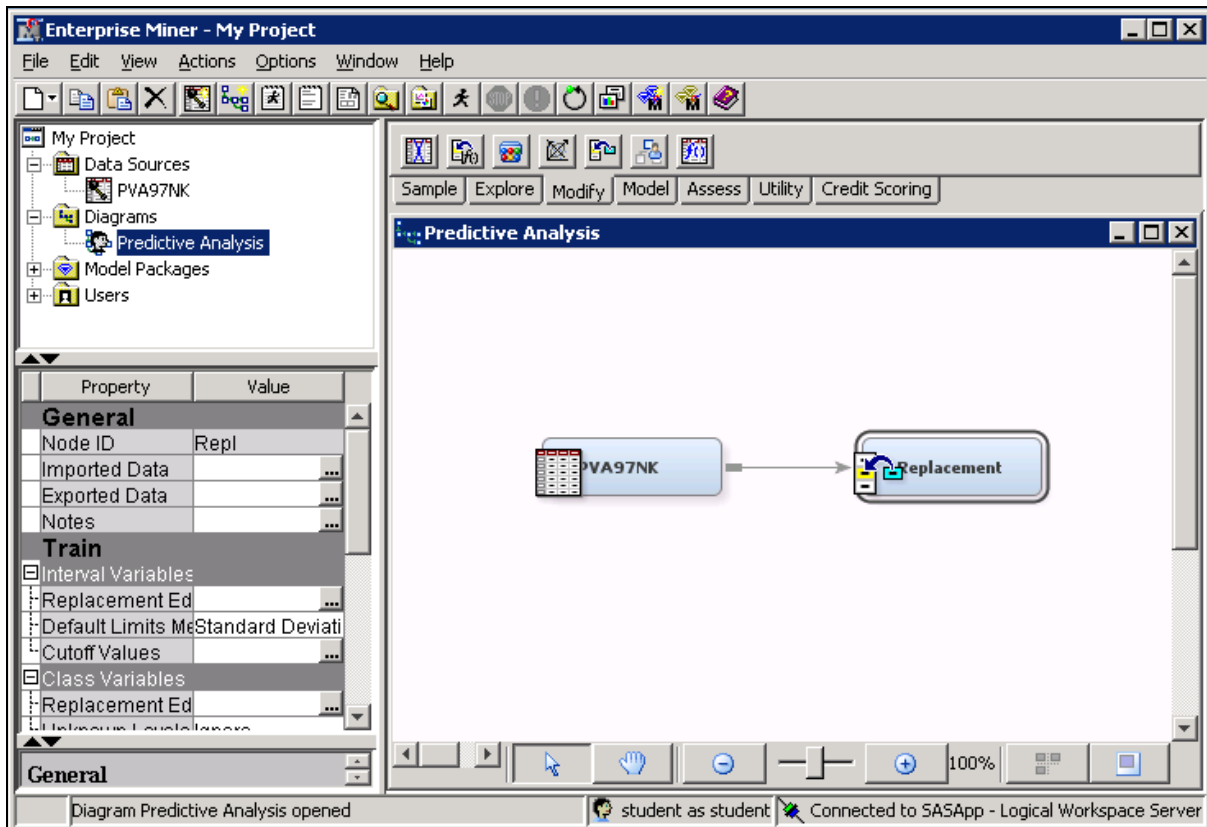
2. Select the **Modify** tab to access the Modify tool group.



3. Drag the **Replacement** tool (third from the right) from the Tools Palette into the Predictive Analysis workspace window.



4. Connect the **PVA97NK** data to the Replacement node by clicking near the right side of the **PVA97NK** node and dragging an arrow to the left side of the Replacement node.



You created a process flow, which is the method that SAS Enterprise Miner uses to carry out analyses. The process flow, at this point, reads the raw **PVA97NK** data and replaces the unwanted values of the observations. You must, however, specify which variables have unwanted values and what the correct values are. To do this, you must change the settings of the Replacement node.

Changing the Replacement Node Properties

Use the following steps to modify the default settings of the Replacement node:

1. Select the **Replacement** node and examine the Properties panel.

Property	Value
General	
Node ID	Repl
Imported Data	...
Exported Data	...
Notes	...
Train	
Interval Variables	
Replacement Editor	...
Default Limits Method	Standard Deviations
Cutoff Values	...
Class Variables	
Replacement Editor	...
Unknown Levels	Ignore
Score	
Replacement Value	Computed
Hide	No
Report	
Replacement Report	Yes

The Properties panel displays the analysis methods used by the node when it is run. By default, the node replaces all interval variables whose values are more than three standard deviations from the variable mean.



You can control the number of standard deviations by selecting the **Cutoff Values** property.

In this demonstration, you only want to replace the value for **DemMedIncome** when it equals zero. Thus, you need to change the default setting.

2. Select the **Default Limits Method** property and select **None** from the Options menu.

Property	Value
General	
Node ID	Repl
Imported Data	...
Exported Data	...
Notes	...
Train	
Interval Variables	
Replacement Editor	...
Default Limits Method	None
Cutoff Values	...
Class Variables	
Replacement Editor	...
Unknown Levels	Ignore
Score	
Replacement Value	Computed
Hide	No
Report	
Replacement Report	Yes

You want to replace improper values with missing values. To do this, you need to change the Replacement Value property.

3. Select the **Replacement Value** property and select **Missing** from the Options menu.

Property	Value
General	
Node ID	Repl
Imported Data	...
Exported Data	...
Notes	...
Train	
Interval Variables	
Replacement Editor	...
Default Limits Meth	None
Cutoff Values	...
Class Variables	
Replacement Editor	...
Unknown Levels	Ignore
Score	
Replacement Value	Missing
Hide	No
Report	
Replacement Repo	Yes

You are now ready to specify the variables that you want to replace.

4. Select  (Interval Variables: Replacement Editor ellipsis) from the Replacement node properties panel.

Property	Value
General	
Node ID	Repl
Imported Data	...
Exported Data	...
Notes	...
Train	
Interval Variables	
Replacement Editor	...
Default Limits Meth	None
Cutoff Values	...
Class Variables	
Replacement Editor	...
Unknown Levels	Ignore
Score	
Replacement Value	Missing
Hide	No
Report	
Replacement Repo	Yes



Be careful to open the Replacement Editor for Interval Variables, **not** for Class Variables.

The Interactive Replacement Interval Filter window opens.

Name	Use	Report	Limit Method	Lower Limit	Upper Limit
DemAge	Default	No	Default	.	.
DemMedHomeValue	Default	No	Default	.	.
DemMedIncome	Default	No	Default	.	.
DemPctVeterans	Default	No	Default	.	.
GiftAvg36	Default	No	Default	.	.
GiftAvgAll	Default	No	Default	.	.
GiftAvgCard36	Default	No	Default	.	.
GiftAvgLast	Default	No	Default	.	.
GiftCnt36	Default	No	Default	.	.
GiftCntAll	Default	No	Default	.	.
GiftCntCard36	Default	No	Default	.	.
GiftCntCardAll	Default	No	Default	.	.
GiftTimeFirst	Default	No	Default	.	.
GiftTimeLast	Default	No	Default	.	.
PromCnt12	Default	No	Default	.	.
PromCnt36	Default	No	Default	.	.
PromCntAll	Default	No	Default	.	.
PromCntCard12	Default	No	Default	.	.
PromCntCard36	Default	No	Default	.	.
PromCntCardAll	Default	No	Default	.	.
TargetD	Default	No	Default	.	.

Generate Summary OK Cancel

5. Select **User Specified** as the Limit Method value for **DemMedIncome**.

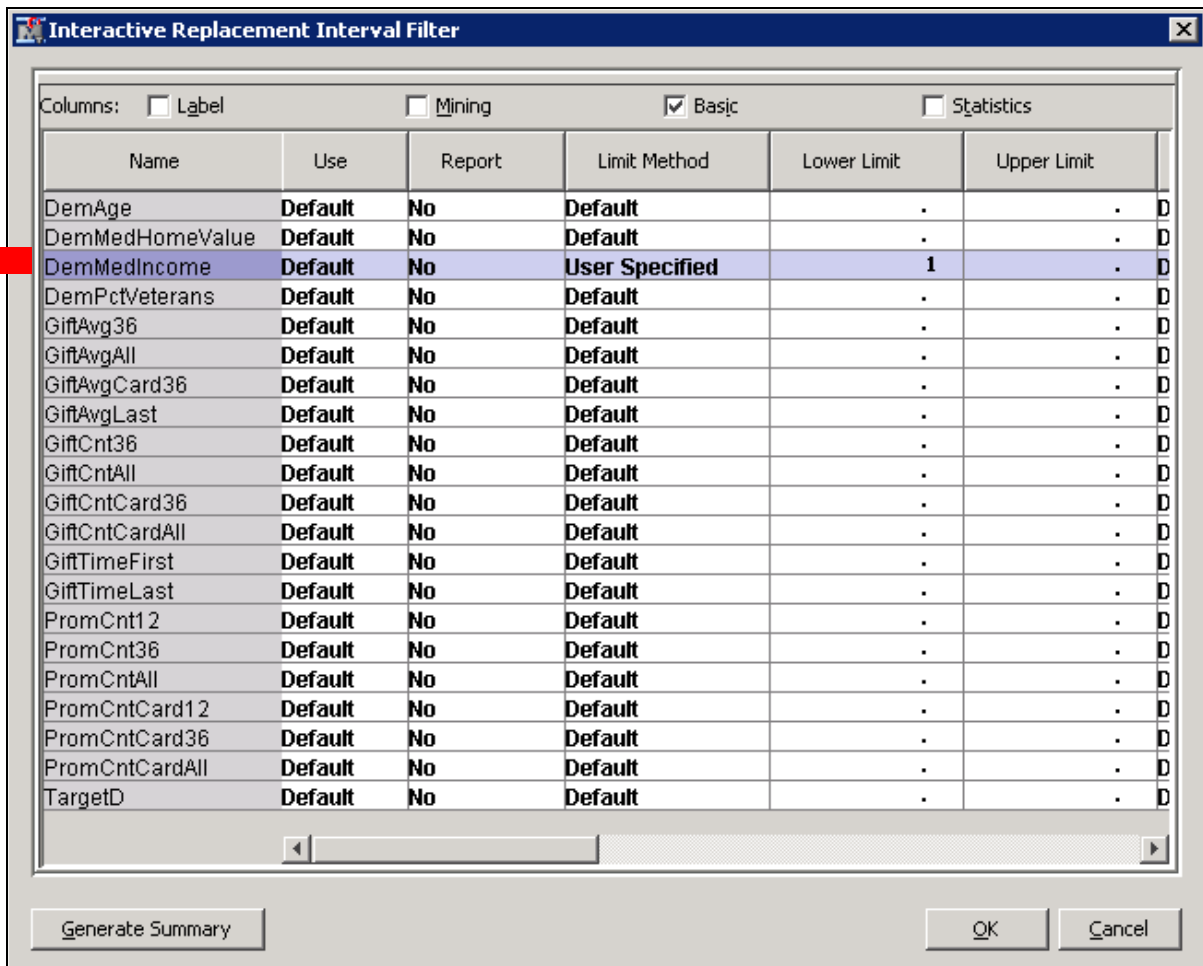
Interactive Replacement Interval Filter

Columns: ☐ Label ☐ Mining ☒ Basic ☐ Statistics

Name	Use	Report	Limit Method	Lower Limit	Upper Limit
DemAge	Default	No	Default	.	.
DemMedHomeValue	Default	No	Default	.	.
DemMedIncome	Default	No	User Specified	.	.
DemPctVeterans	Default	No	Default	.	.
GiftAvg36	Default	No	Default	.	.
GiftAvgAll	Default	No	Default	.	.
GiftAvgCard36	Default	No	Default	.	.
GiftAvgLast	Default	No	Default	.	.
GiftCnt36	Default	No	Default	.	.
GiftCntAll	Default	No	Default	.	.
GiftCntCard36	Default	No	Default	.	.
GiftCntCardAll	Default	No	Default	.	.
GiftTimeFirst	Default	No	Default	.	.
GiftTimeLast	Default	No	Default	.	.
PromCnt12	Default	No	Default	.	.
PromCnt36	Default	No	Default	.	.
PromCntAll	Default	No	Default	.	.
PromCntCard12	Default	No	Default	.	.
PromCntCard36	Default	No	Default	.	.
PromCntCardAll	Default	No	Default	.	.
TargetID	Default	No	Default	.	.

Generate Summary OK Cancel

6. Type **1** as the Lower Limit value for **DemMedIncome**.



The dialog box titled "Interactive Replacement Interval Filter" contains a table with columns: Name, Use, Report, Limit Method, Lower Limit, and Upper Limit. The "Basic" tab is selected. The row for "DemMedIncome" is highlighted, showing a "User Specified" limit method with a lower limit of 1. A red rectangle highlights the "DemMedIncome" row.

Name	Use	Report	Limit Method	Lower Limit	Upper Limit
DemAge	Default	No	Default	.	.
DemMedHomeValue	Default	No	Default	.	.
DemMedIncome	Default	No	User Specified	1	.
DemPctVeterans	Default	No	Default	.	.
GiftAvg36	Default	No	Default	.	.
GiftAvgAll	Default	No	Default	.	.
GiftAvgCard36	Default	No	Default	.	.
GiftAvgLast	Default	No	Default	.	.
GiftCnt36	Default	No	Default	.	.
GiftCntAll	Default	No	Default	.	.
GiftCntCard36	Default	No	Default	.	.
GiftCntCardAll	Default	No	Default	.	.
GiftTimeFirst	Default	No	Default	.	.
GiftTimeLast	Default	No	Default	.	.
PromCnt12	Default	No	Default	.	.
PromCnt36	Default	No	Default	.	.
PromCntAll	Default	No	Default	.	.
PromCntCard12	Default	No	Default	.	.
PromCntCard36	Default	No	Default	.	.
PromCntCardAll	Default	No	Default	.	.
TargetD	Default	No	Default	.	.

Buttons: Generate Summary, OK, Cancel

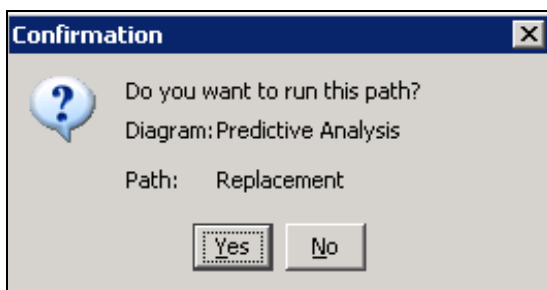
If you use this specification, any **DemMedIncome** values that fall below the lower limit of 1 are set to missing. All other values of this variable do not change.

7. Select **OK** to close the Interactive Replacement Interval Filter window.

Running the Analysis and Viewing the Results

Use these steps to run the process flow that you created.

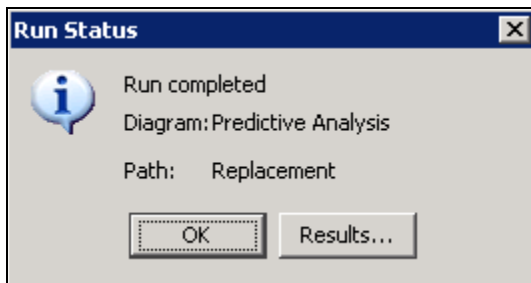
1. Right-click on the **Replacement** node and select **Run** from the Option menu. A Confirmation window appears, and requests that you verify the run action.



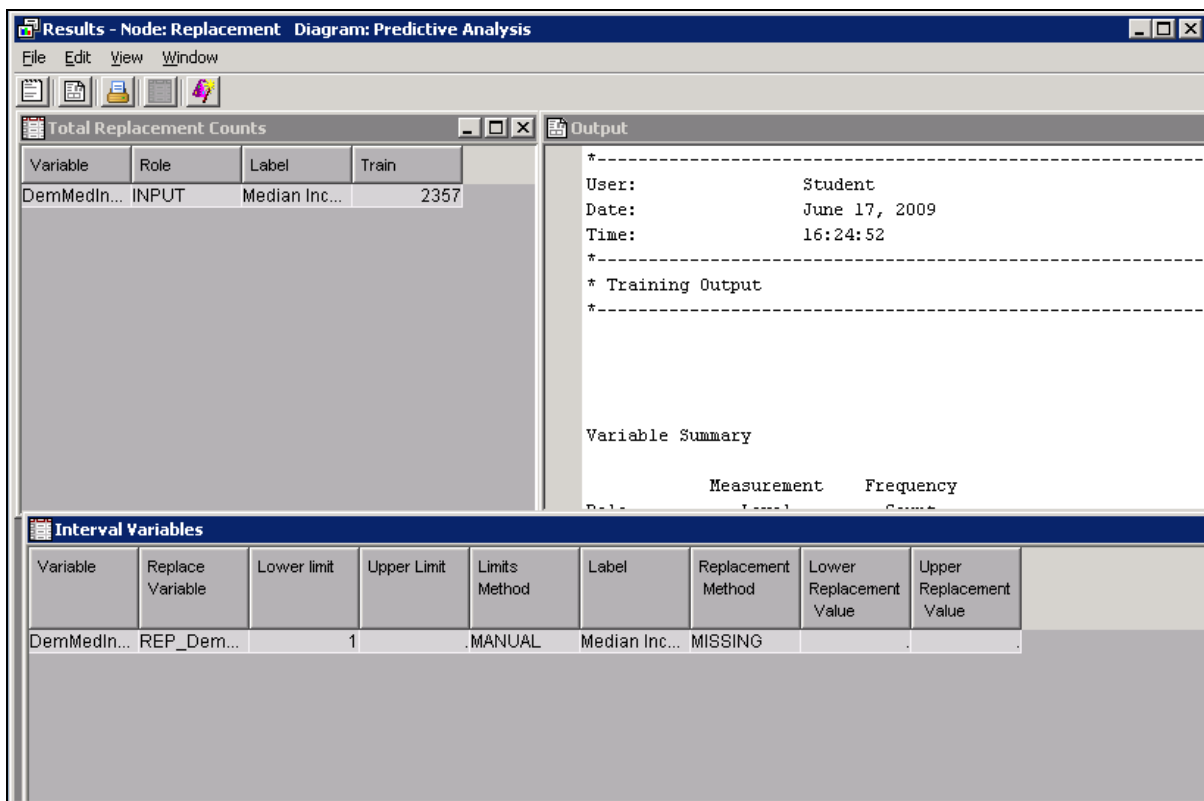
The "Confirmation" dialog box asks "Do you want to run this path?" and shows "Diagram: Predictive Analysis" and "Path: Replacement". It has "Yes" and "No" buttons.

2. Select **Yes** to close the Confirmation window. A small animation in the lower right corner of each node indicates analysis activity in the node.

The Run Status window opens when the process flow run is complete.



3. Select **Results...** to review the analysis outcome. The Results - Node: Replacement Diagram: Predictive Analysis window appears.




The Replacement Counts window shows that 2357 observations were modified by the Replacement node. The Interval Variables window summarizes the replacement that was conducted. The Output window provides more or less the same information as the Total Replacement Counts window and the Interval Variables window (but it is presented as a static text file).

4. Close the Results window.

Examining Exported Data

In a SAS Enterprise Miner process flow diagram, each node takes in data, analyses it, creates a result, and exports a possibly modified version of the imported data. While the report gives the analysis results in abstract, it is good practice to see the actual effects of an analysis step on the exported data. This enables you to validate your expectations at each step of the analysis.

Use these steps to examine the data exported from the Replacement node.

1. Select the **Replacement** node in your process flow diagram.
2. Select **Exported Data** ⇨ .

Property	Value
General	
Node ID	Repl
Imported Data	...
Exported Data	...
Notes	...
Train	
Interval Variables	
Replacement Editor	...
Default Limits Method	None
Cutoff Values	...
Class Variables	
Replacement Editor	...
Unknown Levels	Ignore
Score	
Replacement Value	Missing
Hide	No
Report	
Replacement Report	Yes

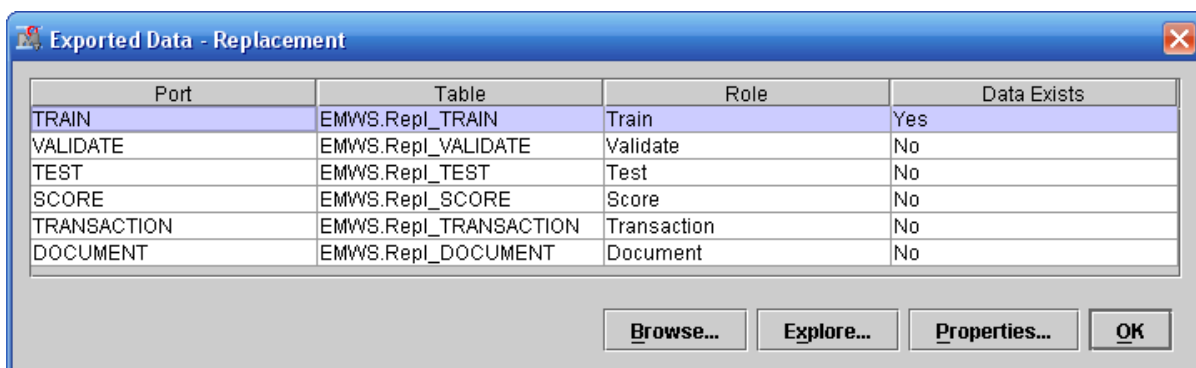
The Exported Data - Replacement window appears.

Exported Data - Replacement			
Port	Table	Role	Data Exists
TRAIN	EMWS.Repl_TRAIN	Train	Yes
VALIDATE	EMWS.Repl_VALIDATE	Validate	No
TEST	EMWS.Repl_TEST	Test	No
SCORE	EMWS.Repl_SCORE	Score	No
TRANSACTION	EMWS.Repl_TRANSACTION	Transaction	No
DOCUMENT	EMWS.Repl_DOCUMENT	Document	No

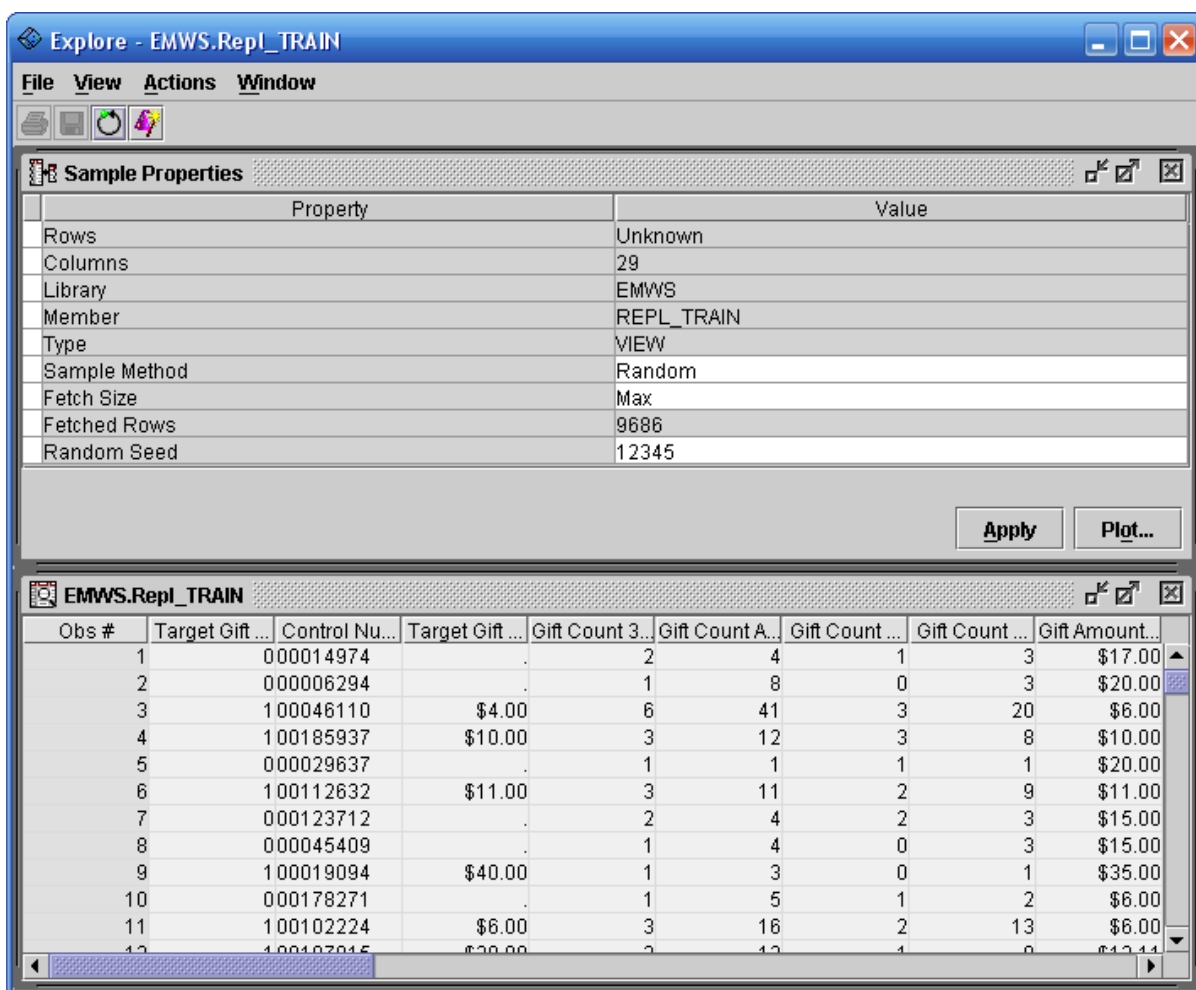
Browse... Explore... Properties... OK

This window lists the types of data sets that can be exported from a SAS Enterprise Miner process flow node. As indicated, only a **TRAIN** data set exists at this stage of the analysis.

3. Select the **Train** table from the Exported Data - Replacement window.



4. Select **Explore...** to open the Explore window again.



5. Scroll completely to the right in the data table.

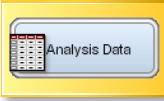
Age	Gender	Home Owner	Median Ho...	Percent Vet...	Median Income Region	Replacement: DemMedIncome
.	F	U	\$0	0	\$0	.
67	F	U	\$186,800	85	\$0	.
.	M	U	\$87,600	36	\$38,750	38750
.	M	U	\$139,200	27	\$38,942	38942
53	M	U	\$168,100	37	\$71,509	71509
47	M	H	\$253,100	0	\$92,514	92514
58	M	H	\$234,700	22	\$72,868	72868
.	U	U	\$207,000	44	\$0	.
.	F	U	\$137,300	32	\$0	.
.	U	U	\$180,700	37	\$0	.
.	U	U	\$143,900	30	\$0	.
.	F	U	\$134,400	44	\$0	.

A new column is added to the analysis data: **Replacement: DemMedIncome**. Notice that the values of this variable match the **Median Income Region** variable, except when the original variable's value equals zero. The replaced zero value is shown by a dot (.), which indicates a missing value.

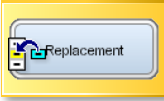
6. Close the Explore and Exported Data windows to complete this part of the analysis.

2.5 Chapter Summary

Data Access Tools Review



- Link existing analysis data sets to SAS Enterprise Miner.
- Set variable metadata.
- Explore variable distribution characteristics.



Remove unwanted cases from analysis data.

18

Analyses in SAS Enterprise Miner are organized hierarchically into projects, data libraries, diagrams, process flows, and analysis nodes. This chapter demonstrated the basics of creating each of these elements.

Most process flows begin with a data source node. To access a data source, you must define a SAS library. After a library is defined, a data source is created by linking a SAS table and associated metadata to SAS Enterprise Miner. After a data source is defined, you can assay the underlying cases using SAS Enterprise Miner Explore tools. Care should be taken to ensure that the sample of the data source that you explore is representative of the original data source.

You can use the Replacement node to modify variables that were incorrectly prepared. After all necessary modifications, the data is ready for subsequent analysis.