



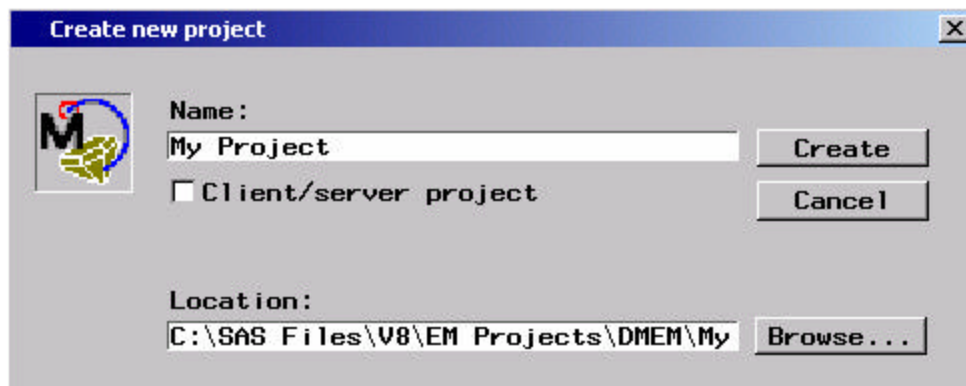
Introduction to Enterprise Miner

Opening The Enterprise Miner

1. Start a SAS session. Double-click on the SAS icon on your desktop or select **Start** ⇒ **Programs** ⇒ **The SAS System** ⇒ **The SAS System for Windows V8**.
2. To start Enterprise Miner, type **miner** in the command box or select **Solutions** ⇒ **Analysis** ⇒ **Enterprise Miner**.

Setting Up the Initial Project and Diagram

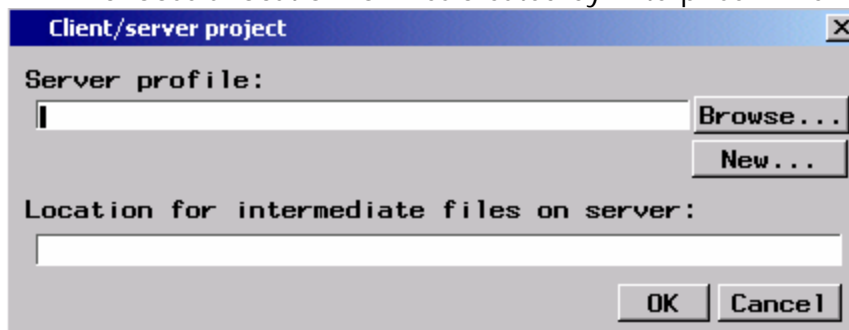
1. Select **File** ⇒ **New** ⇒ **Project...**
2. Modify the location of the project folder if desired by selecting **Browse...**
3. Type the name of the project (for example, **My Project**).



4. Check the box for Client/server project if needed. Do not check this box unless instructed to do so by the instructor.

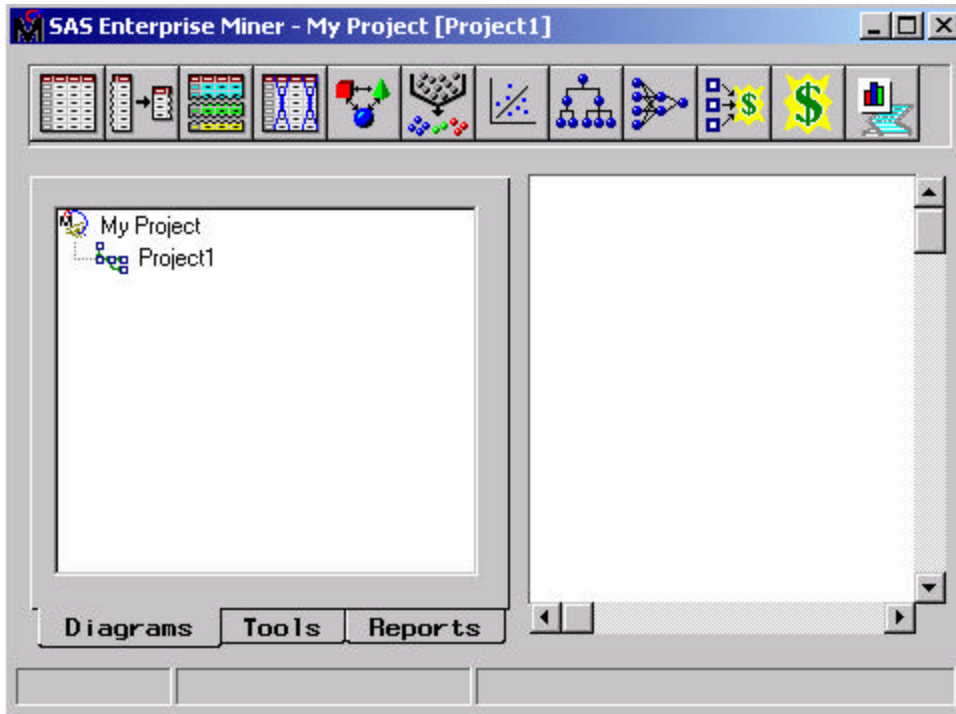


You must have access to a server running the same version of Enterprise Miner. This allows you to access databases on a remote host or distribute the data-intensive processing to a more powerful remote host. If you create a client/server project, you will be prompted to provide a server profile and to choose a location for files created by Enterprise Miner.



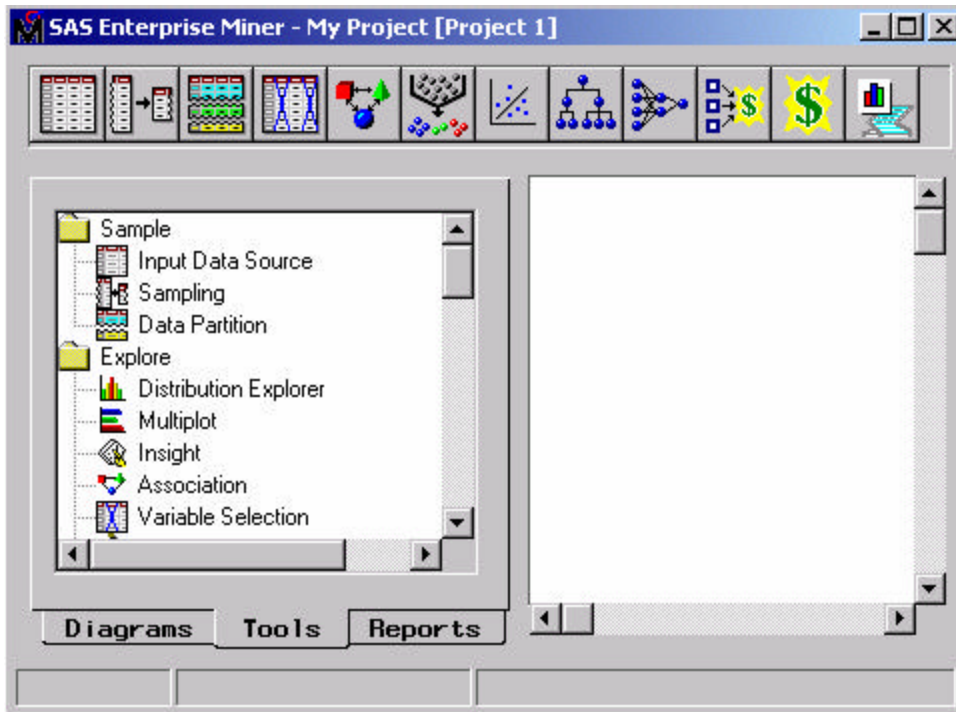
5. Select **Create**. The project opens with an initial untitled diagram.

6. Click on the diagram title and type a new title if desired (for example, **Project1**).



Identifying the Workspace Components

1. Observe that the project window opens with the Diagrams tab activated. Select the **Tools** tab located to the right of the Diagrams tab in the lower-left portion of the project window. This tab enables you to see all of the tools (or nodes) that are available in Enterprise Miner.



Many of the commonly used tools are shown on the toolbar at the top of the window. If you want additional tools on this toolbar, you can drag them from the window above onto the toolbar. In addition, you can rearrange the tools on the toolbar by dragging each tool to the desired location on the bar.

2. Select the **Reports** tab located to the right of the Tools tab. This tab reveals any reports that have been generated for this project. This is a new project, so no reports are currently available.

The open space on the right is your diagram workspace. This is where you graphically build, order, and sequence the nodes you use to mine your data and generate reports.

The Scenario

- Determine who should be approved for a home equity loan.
- The target variable is a binary variable that indicates whether an applicant eventually defaulted on the loan.
- The input variables are variables such as the amount of the loan, amount due on the existing mortgage, the value of the property, and the number of recent credit inquiries.

4

The consumer credit department of a bank wants to automate the decision-making process for approval of home equity lines of credit. To do this, they will follow the recommendations of the Equal Credit Opportunity Act to create an empirically derived and statistically sound credit scoring model. The model will be based on data collected from recent applicants granted credit through the current process of loan underwriting. The model will be built from predictive modeling tools, but the created model must be sufficiently interpretable so as to provide a reason for any adverse actions (rejections).

The HMEQ data set contains baseline and loan performance information for 5,960 recent home equity loans. The target (BAD) is a binary variable that indicates if an applicant eventually defaulted or was seriously delinquent. This adverse outcome occurred in 1,189 cases (20%). For each applicant, 12 input variables were recorded.

Name	Model Role	Measurement Level	Description
BAD	Target	Binary	1 = defaulted on loan, 0 = paid back loan
REASON	Input	Binary	HomeImp = home improvement, DebtCon = debt consolidation
JOB	Input	Nominal	Six occupational categories
LOAN	Input	Interval	Amount of loan request
MORTDUE	Input	Interval	Amount due on existing mortgage
VALUE	Input	Interval	Value of current property
DEBTINC	Input	Interval	Debt-to-income ratio
YOJ	Input	Interval	Years at present job
DEROG	Input	Interval	Number of major derogatory reports
CLNO	Input	Interval	Number of trade lines
DELINQ	Input	Interval	Number of delinquent trade lines
CLAGE	Input	Interval	Age of oldest trade line in months
NINQ	Input	Interval	Number of recent credit inquiries

The credit scoring model computes a probability of a given loan applicant defaulting on loan repayment. A threshold is selected such that all applicants whose probability of default is in excess of the threshold are recommended for rejection.

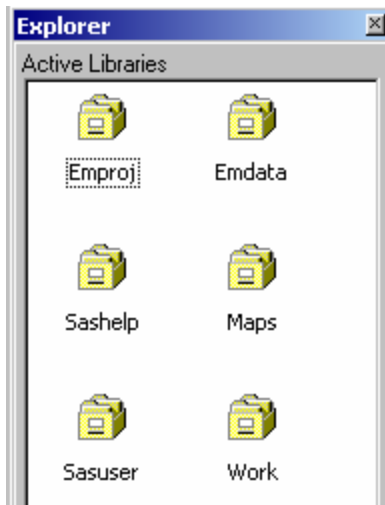


Project Setup and Initial Data Exploration

Using SAS Libraries

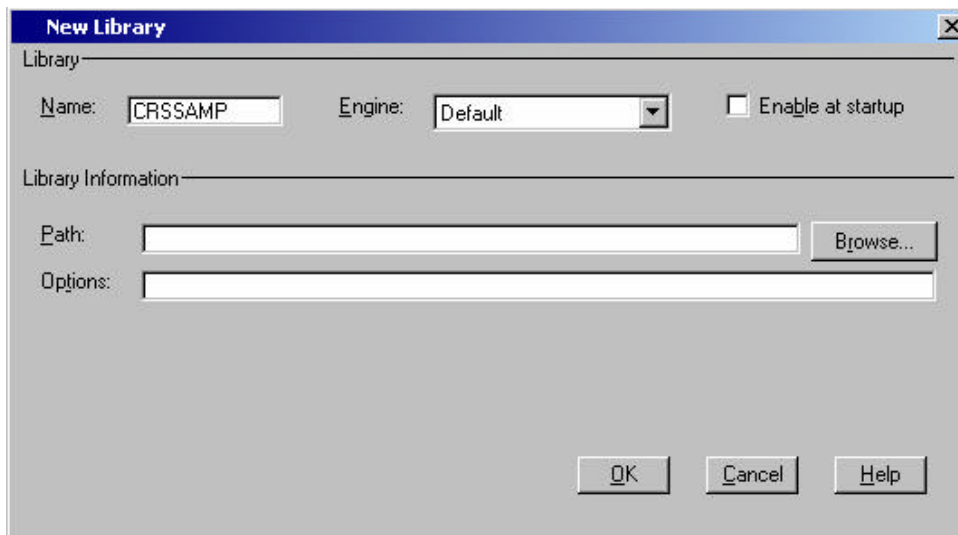
To identify a SAS data library, you assign it a library reference name, or *libref*. When you open Enterprise Miner, several libraries are automatically assigned and can be seen in the Explorer window.

1. Double-click on the Libraries icon in the Explorer window.



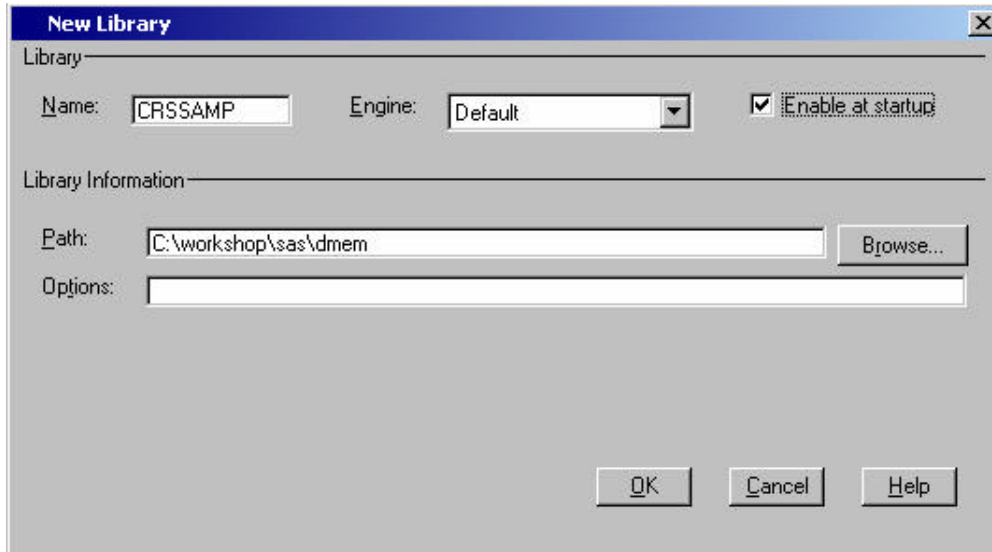
To define a new library:

2. Right-click in the Explorer window and select **New**.

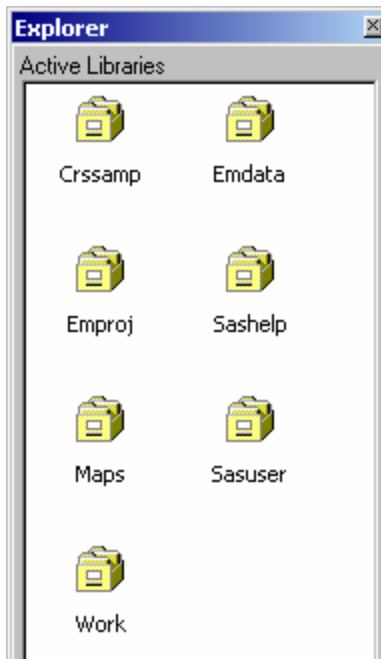


3. In the New Library window, type a name for the new library. For example, type CRSSAMP.

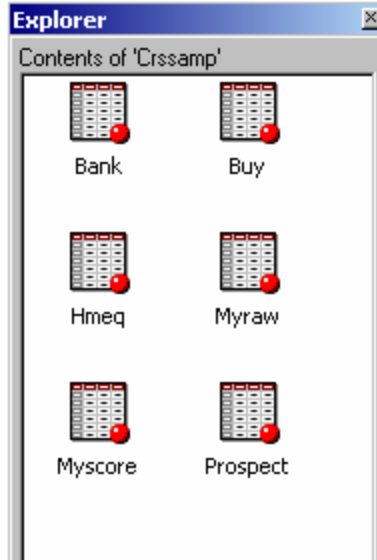
4. Type in the path name or select **Browse** to choose the folder to be connected with the new library name. For example, the chosen folder might be located at C:\workshop\sas\dmem.
5. If you want this library name to be connected with this folder every time you open SAS, select **Enable at startup**.



6. Select **OK**. The new library is now assigned and can be seen in the Explorer window.

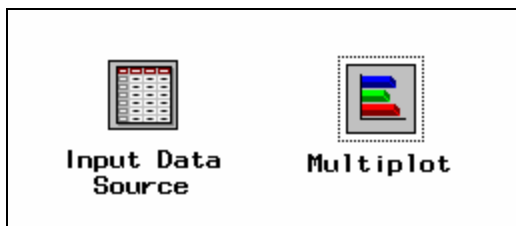


7. To view the data sets that are included in the new library, double-click on the icon for Crssamp.



Building the Initial Flow

1. Presuming that the diagram Project1 in the project named My Project is open, add an Input Data Source node by dragging the node from the toolbar or from the Tools tab to the diagram workspace.
2. Add a Multiplot node to the workspace to the right of the Input Data Source node. Your diagram should appear as shown below.



Observe that the Multiplot node is selected (as indicated by the dotted line around it), but the Input Data Source node is not selected. If you click in any open space on the workspace, all nodes become deselected.

In addition to dragging a node onto the workspace, there are two other ways to add a node to the flow. You can right-click in the workspace where you want the node to be placed and select **Add node** from the pop-up menu, or you can double-click where you want the node to be placed. In either case, a list of nodes appears, enabling you to select the desired node.

The shape of the cursor changes depending on where it is positioned. The behavior of the mouse commands depends on the shape as well as the selection state of the node over which the cursor is positioned. Right-click in an open area to see the menu. The last three menu items (Connect items, Move items, Move and Connect) enable you to modify the ways in which the cursor can be used. Move and Connect is selected by default, and it is highly recommended that you do not change this setting. If your cursor is not performing a desired task, check this menu to make sure

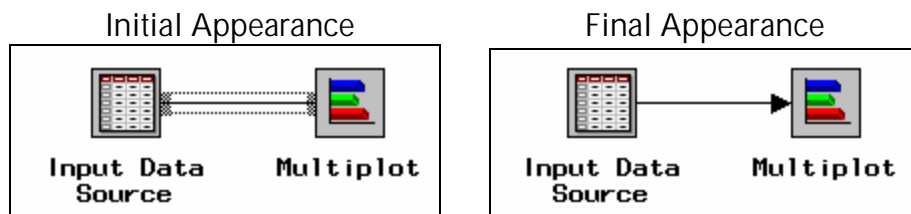
that Move and Connect is selected. This selection allows you to move the nodes around the workspace as well as connect them.

Observe that when you put your cursor in the middle of a node, the cursor appears as a hand. To move the nodes around the workspace:

1. Position the cursor in the middle of the node until the hand appears.
2. Press the left mouse button and drag the node to the desired location.
3. Release the left mouse button.

To connect the two nodes in the workspace:

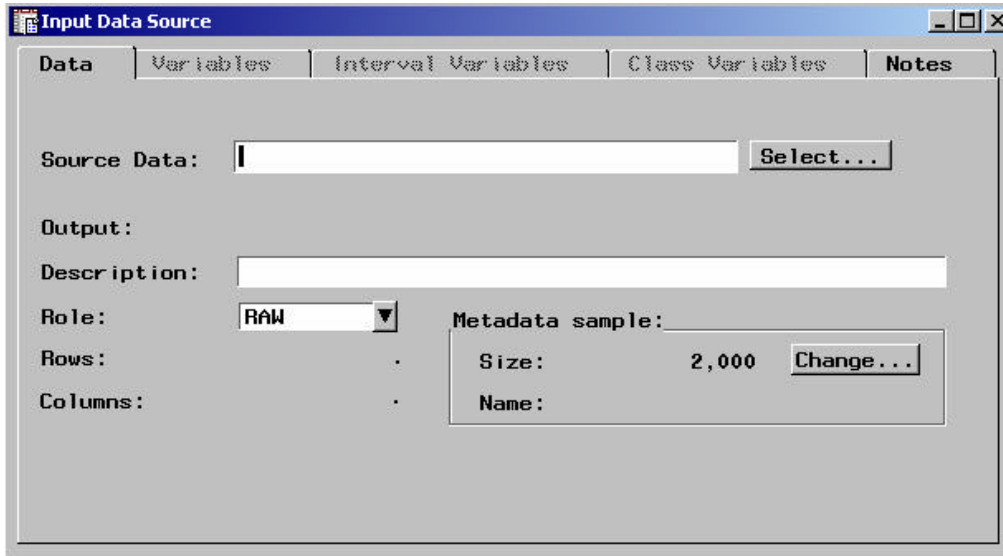
1. Ensure that the Input Data Source node is deselected. It is much easier to drag a line when the node is deselected. If the beginning node is selected, click in an open area of the workspace to deselect it.
2. Position the cursor on the edge of the icon representing the Input Data Source node (until the crosshair appears).
3. Press the left mouse button and immediately begin to drag in the direction of the Multiplot node. If you do not begin dragging immediately after pressing the left mouse button, you select only the node. Dragging a selected node generally results in moving the node; that is, no line forms.
4. Release the mouse button after reaching the edge of the icon that represents the ending node.
5. Click away from the line and the finished arrow forms as shown below.



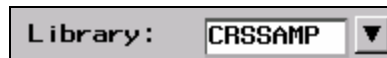
Identifying the Input Data

This example uses the HMEQ data set in the CRSSAMP library.

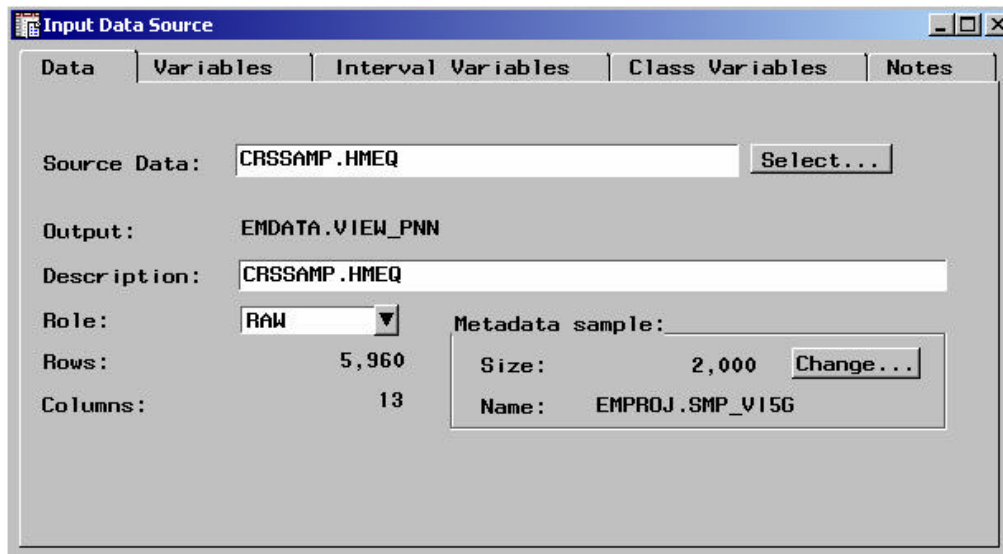
1. To specify the input data, double-click on the **Input Data Source** node or right-click on this node and select **Open...**. The Data tab is active. Your window should appear as follows:



2. Click on **Select...** to select the data set. Alternatively, you can enter the name of the data set.
3. The SASUSER library is selected by default. To view data sets in the CRSSAMP library, click on the ▼ and select **CRSSAMP** from the list of defined libraries.



4. Select **HMEQ** from the list of data sets in the CRSSAMP library and then select **OK**. The dialog shown below opens.



Observe that this data set has 5,960 observations (rows) and 13 variables (columns). CRSSAMP.HMEQ is listed as the source data. You could have typed in this name in the field instead of selecting it through the dialog. Note that the lower-right corner indicates a metadata sample of size 2,000.

All analysis packages must determine how to use variables in the analysis. Enterprise Miner utilizes metadata in order to make a preliminary assessment of how to use each variable. By default, it takes a random sample of 2,000 observations from the data set of interest and uses this information to assign a model role and a measurement level to each variable. To take a larger sample, you can select the **Change...** button in the lower-right corner of the dialog. However, that is not shown here.

1. Click on the **Variables** tab to see all of the variables and their respective assignments.
2. Click on the first column heading, labeled Name, to sort the variables by their name. You can see all of the variables if you enlarge the window. The following table shows a portion of the information for each of the 13 variables.


Name	Model Role	Measurement	Type	Format
BAD	input	binary	num	BEST12.
CLAGE	input	interval	num	BEST12.
CLNO	input	interval	num	BEST12.
DEBT INC	input	interval	num	BEST12.
DEL INQ	input	interval	num	BEST12.
DEROG	input	interval	num	BEST12.
JOB	input	nominal	char	\$7.
LOAN	input	interval	num	BEST12.
MORTDUE	input	interval	num	BEST12.
N INQ	input	interval	num	BEST12.
REASON	input	binary	char	\$7.
VALUE	input	interval	num	BEST12.
YOJ	input	interval	num	BEST12.

Observe that two of the columns are grayed out. These columns represent information from the SAS data set that cannot be changed in this node. Type is either character (**char**) or numeric (**num**), and it affects how a variable can be used. The value for Type and the number of levels in the metadata sample of 2,000 is used to identify the model role and measurement level.

The first variable is BAD, which is the target variable. Although BAD is a numeric variable in the data set, Enterprise Miner identifies it as a **binary** variable because it has only two distinct nonmissing levels in the metadata sample. The model role for all **binary** variables is set to **input** by default. You need to change the model role for BAD to target before performing the analysis. The next five variables (CLAGE through DEROG) have the measurement level **interval** because they are numeric variables in the SAS data set and have more than 10 distinct levels in the

metadata sample. The model role for all **interval** variables is set to **input** by default.

The variables JOB and REASON are both character variables in the data set, but they have different measurement levels. REASON is binary because it has only two distinct nonmissing levels in the metadata sample. The model role for JOB, however, is nominal because it is a character variable with more than two levels. For the purpose of this analysis, treat the remaining variables as interval variables.

 At times, variables such as DEROG and DELINQ will be assigned the model role of **ordinal**. A variable is listed as ordinal when it is a numeric variable with more than two but no more than ten distinct nonmissing levels in the metadata sample. This often occurs with counting variables, such as a variable for the number of children. Because this assignment depends on the metadata sample, the measurement level of DEROG or DELINQ for your analysis might be set to **ordinal**. All ordinal variables are set to have the **input** model role; however, you treat these variables as interval inputs for the purpose of this analysis.

Identifying Target Variables

BAD is the response variables for this analysis. Change the model role for BAD to **target**.

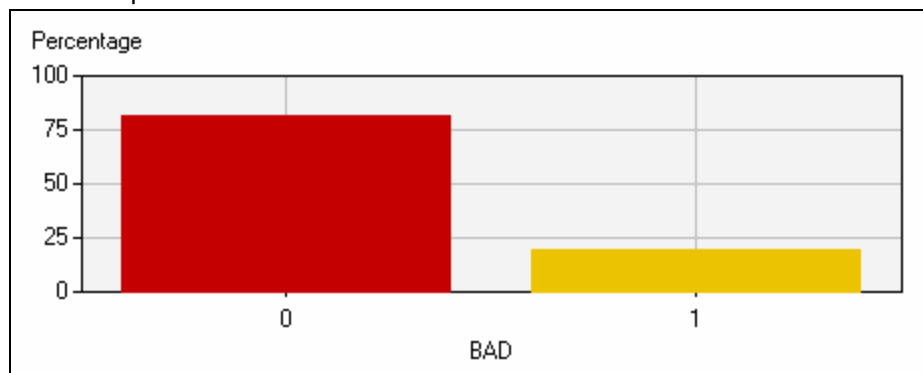
To modify the model role information, proceed as follows:


1. Position the tip of your cursor over the row for BAD in the Model Role column and right-click.
2. Select **Set Model Role** ⇒ **target** from the pop-up menu.

Inspecting Distributions

You can inspect the distribution of values in the metadata sample for each of the variables. To view the distribution of BAD:

1. Position the tip of your cursor over the variable BAD in the Name column.
2. Right-click and observe that you can sort by name, find a name, or view the distribution of BAD.
3. Select **View Distribution of BAD** to see the distribution of values for BAD in the metadata sample.



To obtain additional information, select the the View Info tool, , from the toolbar at the top of the window and click on one of the bars. Enterprise Miner displays the level and the proportion of observations represented by the bar. These plots provide an initial overview of the data. For this example, approximately 20% of the observations were loans where the client defaulted. Because the plots are based on the metadata sample, they may vary slightly due to the differences in the sampled observations, but the bar for BAD=1 should represent approximately 20% of the data. Close the Variable Histogram window when you are finished inspecting the plot. You can evaluate the distribution of other variables as desired.

Modifying Variable Information

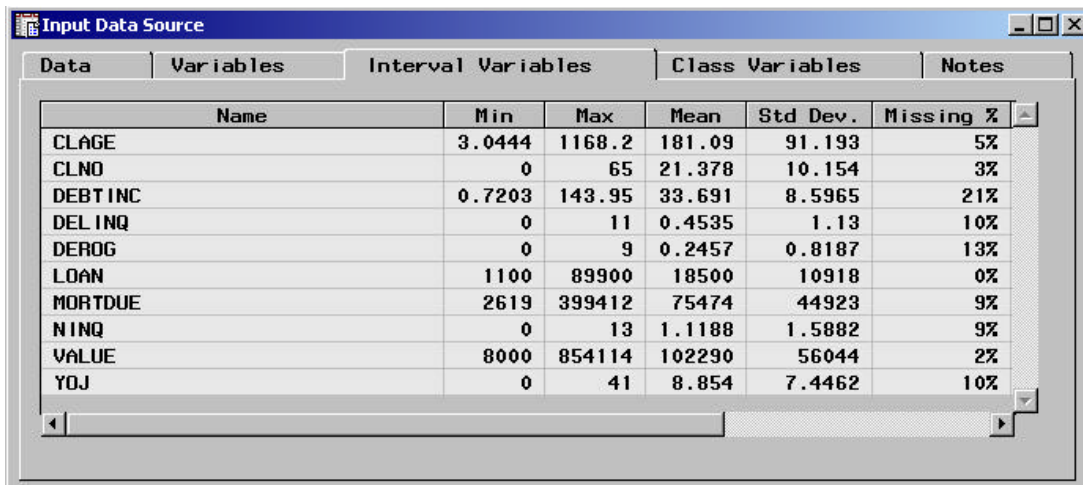
Ensure that the remaining variables have the correct model role and measurement level information. If necessary, change the measurement level for DEROG and DELINQ to **interval**. To modify the measurement level information:

1. Position the tip of your cursor over the row for DEROG in the measurement column and right-click.
2. Select **Set Measurement** ⇒ **interval** from the pop-up menu.
3. Repeat steps 1 and 2 for DELINQ.

Alternatively, you can update the measurement level information for both variables at the same time by highlighting the rows for DEROG and DELINQ simultaneously before following steps 1 and 2 above.

Investigating Descriptive Statistics

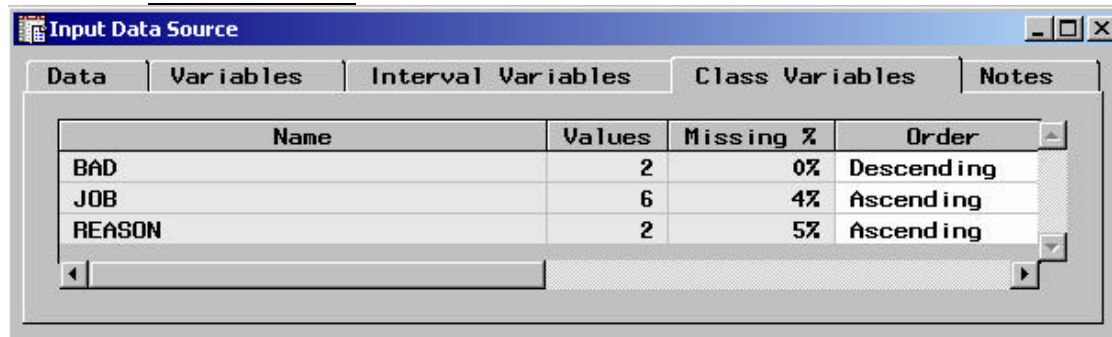
The metadata is used to compute descriptive statistics. Select the **Interval Variables** tab.



Name	Min	Max	Mean	Std Dev.	Missing %
CLAGE	3.0444	1168.2	181.09	91.193	5%
CLNO	0	65	21.378	10.154	3%
DEBTINC	0.7203	143.95	33.691	8.5965	21%
DELINQ	0	11	0.4535	1.13	10%
DEROG	0	9	0.2457	0.8187	13%
LOAN	1100	89900	18500	10918	0%
MORTDUE	2619	399412	75474	44923	9%
NINQ	0	13	1.1188	1.5882	9%
VALUE	8000	854114	102290	56044	2%
YOJ	0	41	8.854	7.4462	10%

Investigate the minimum value, maximum value, mean, standard deviation, percentage of missing observations, skewness, and kurtosis for interval variables. Based on business knowledge of the data, inspecting the minimum and maximum values indicates no unusual values. Observe that DEBTINC has a high percentage of missing values (21%).

Select the **Class Variables** tab.



Name	Values	Missing %	Order
BAD	2	0%	Descending
JOB	6	4%	Ascending
REASON	2	5%	Ascending

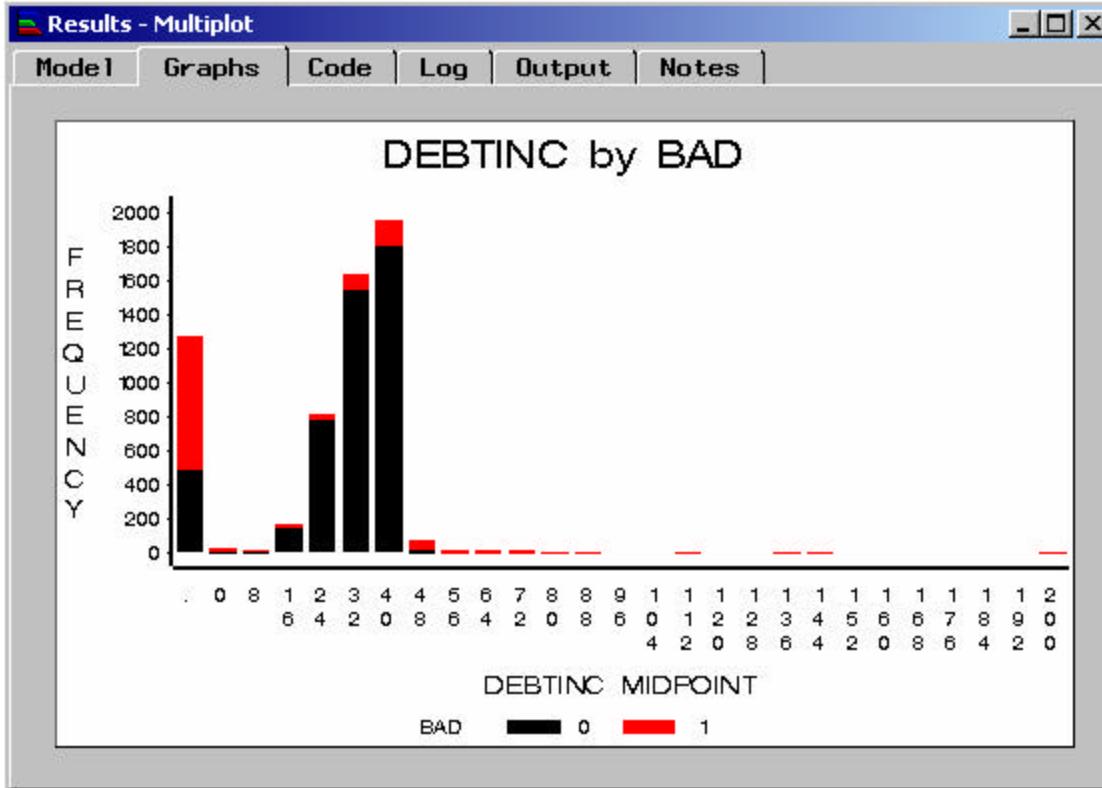
Investigate the number of levels, percentage of missing values, and the sort order of each variable. Observe that the sort order for BAD is descending, whereas the sort order for all the others is ascending. This occurs because you have a binary target event. It is common to code a binary target with a 1 when the event occurs and a 0 otherwise. Sorting in descending order makes level 1 the first level, which is the target event for a binary variable. It is useful to sort other similarly coded binary variables in descending order for interpreting parameter estimates in a regression model. Close the Input Data Source node, saving changes when prompted.

Additional Data Exploration

Other tools available in Enterprise Miner enable you to explore your data further. One such tool is the Multiplot node. The Multiplot node creates a series of histograms and bar charts that enable you to examine the relationships between the input variables and the binary target variable.

1. Right-click on the Multiplot node and select **Run**.
2. When prompted, select **Yes** to view the results.

By using the Page Down button on your keyboard, you can view the histograms generated for this data.


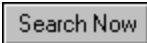


From this histogram, you can see that many of the defaulted loans were by homeowners with either a high debt-to-income ratio or an unknown debt-to-income ratio.



When you open a project diagram in Enterprise Miner, a lock is placed on the diagram to avoid the possibility of more than one person trying to change the diagram at the same time. If Enterprise Miner or SAS terminates abnormally, the lock files are not deleted and the lock remains on the diagram. If this occurs, you must delete the lock file to gain access to the diagram.

To delete a lock file:

1. Right-click on the project name in the diagrams tab of the workspace and select **Explore...**
2. In the toolbar of the explorer window that opens, click on .
3. In the Search for files or folders named field, type ***.lck**.
4. Select .
5. Once the lock file has been located, right-click on the filename and select **Delete**.

This deletes the lock file and makes the project accessible again.