*Chapter 4*

# Explore the Data and Replace Input Values



## About the Tasks That You Will Perform

You have already set up the project and defined the input data source that you will use in this example. Now, you will import the data and perform the following tasks, which help you learn properties of the input data and prepare it for subsequent modeling:

1. You will explore the statistical properties of the variables in the input data set. The results that are generated in this step will give you an idea of which variables are most useful in predicting the target response (whether a person donates or not) in this data set.
2. You will partition the data into two data sets, a training data set and a validation data set. Such partitioning is common practice in data mining and enables you to develop a complete model that is not overfitted to a particular set of data.
3. You will specify how SAS Enterprise Miner should handle missing values of predictor variables.

TIP It is always a good idea to plot the input data and to check it for missing values before you proceed to model building. Knowing the statistical properties of your input data is essential for building an accurate and robust predictive model.

## Generate Descriptive Statistics

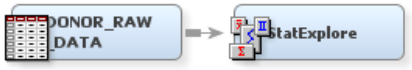
To use the StatExplore node to produce a statistical summary of the input data:

1. Select the **Explore** tab on the Toolbar.
2. Select the **StatExplore** node icon. Drag the node into the Diagram Workspace.

TIP To determine which node an icon represents, position the mouse pointer over the icon and read the tooltip.

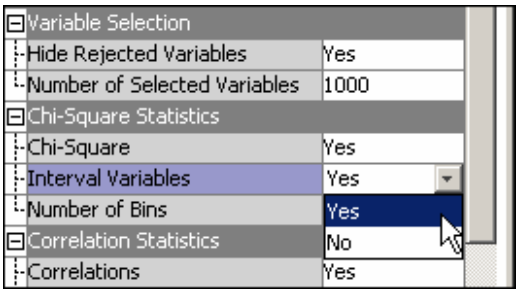
3. Connect the DONOR_RAW_DATA input data source node to the StatExplore node.

   To connect the two nodes, position the mouse pointer over the right edge of the input data source node until the pointer becomes a pencil. With the left mouse button held down, drag the pencil to the left edge of the **StatExplore** node. Then, release the mouse button. An arrow between the two nodes indicates a successful connection.




4. Select the **StatExplore** node. In the Properties Panel, scroll down to view the **Chi-Square Statistics** properties group. Click on the value of **Interval Variables** and select **Yes** from the drop-down menu that appears.

   Chi-square statistics are always computed for categorical variables. Changing the selection for interval variables causes SAS Enterprise Miner to distribute interval variables into five (by default) bins and compute chi-square statistics for the binned variables when you run the node.



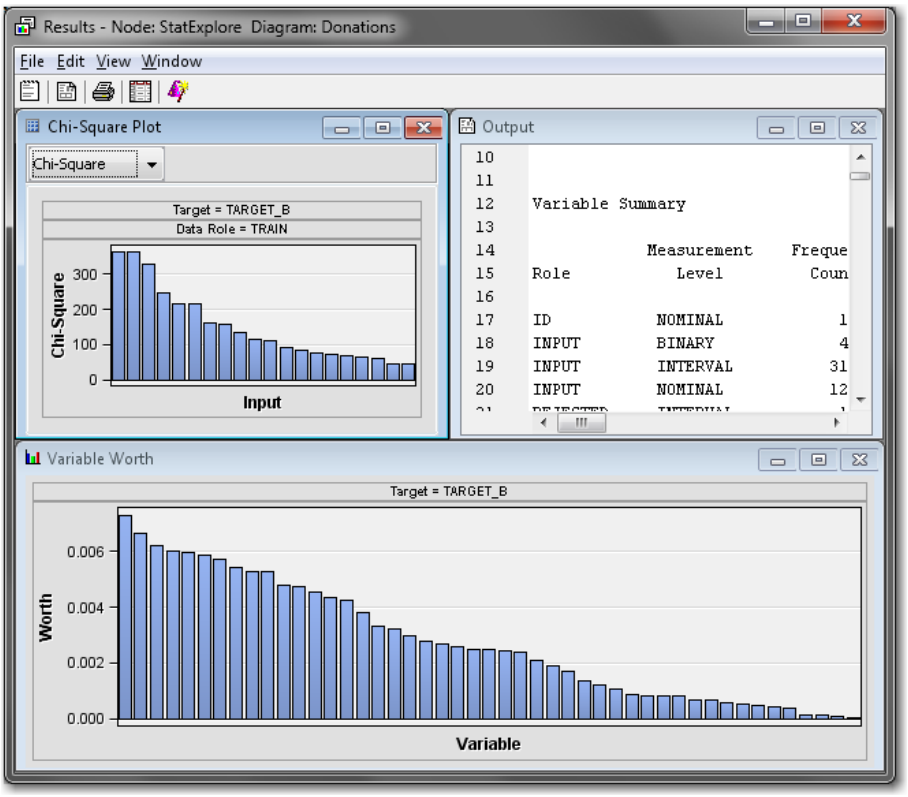

5. In the Diagram Workspace, right-click the **StatExplore** node, and select **Run** from the resulting menu. Click **Yes** in the Confirmation window that opens.

   When you run a node, all of the nodes preceding it in the process flow are also run in order, beginning with the first node that has changed since the flow was last run. If no nodes other than the one that you select have changed since the last run, then only the node that you select is run. You can watch the icons in the process flow diagram to monitor the status of execution.

   - Nodes that are outlined in green are currently running.
   - Nodes that are denoted with a check mark inside a green circle have successfully run.
   - Nodes that are outlined in red have failed to run due to errors.

   In this example, the DONOR_RAW_DATA input data node had not yet been run. Therefore, both nodes are run when you select to run the StatExplore node.

6. In the window that appears when processing completes, click **Results**. The Results window appears.




*Note:* Panels in Results windows might not have the same arrangement on your screen, due to window resizing. When the Results window is resized, SAS Enterprise Miner redistributes panels for optimal viewing.

The results window displays the following:

- a plot that orders the variables by their worth in predicting the target variable.

  *Note:* In the StatExplore node, SAS Enterprise Miner calculates variable worth using the Gini split worth statistic that would be generated by building a decision tree of depth 1. For detailed information about Gini split worth, see the SAS Enterprise Miner Help.

- the SAS output from the node.
- a plot that orders the top 20 variables by their chi-square statistics. You can also choose to view the top 20 variables ordered by their Cramer's V statistics on this plot.

**TIP** In SAS Enterprise Miner, you can select graphs, tables, and rows within tables and select **Copy** from the right-click pop-up menu to copy these items for subsequent pasting in other applications such as Microsoft Word and Microsoft Excel.

7. Expand the Output window, and then scroll to the **Class Variable Summary Statistics** and the **Interval Variable Summary Statistics** sections of the output.

   - Notice that there are two class variables and two interval variables for which there are missing values. Later in the example, you will impute values to use in the place of missing values for these variables.
   - Notice that several variables have relatively large standard deviations. Later in the example, you will plot the data and explore transformations that can reduce the variances of these variables.

8. Close the Results window.

## Partition the Data

In data mining, a strategy for assessing the quality of model generalization is to partition the data source. A portion of the data, called the *training data set*, is used for preliminary model fitting. The rest is reserved for empirical validation and is often split into two parts: validation data and test data. The *validation data set* is used to prevent a modeling node from overfitting the training data and to compare models. The *test data set* is used for a final assessment of the model.

*Note:* In SAS Enterprise Miner, the default data partitioning method for class target variables is to stratify on the target variable or variables. This method is appropriate for this sample data because there is a large number of non-donors in the input data relative to the number of donors. Stratifying ensures that both non-donors and donors are well-represented in the data partitions.

To use the Data Partition node to partition the input data into training and validation sets:

1. Select the **Sample** tab on the Toolbar.
2. Select the **Data Partition** node icon. Drag the node into the Diagram Workspace.
3. Connect the **StatExplore** node to the **Data Partition** node.


4. Select the **Data Partition** node. In the Properties Panel, scroll down to view the data set allocations in the Train properties.
   - Click on the value of **Training**, and enter `55.0`
   - Click on the value of **Validation**, and enter `45.0`
   - Click on the value of **Test**, and enter `0.0`

   These properties define the percentage of input data that is used in each type of mining data set. In this example, you use a training data set and a validation data set, but you do not use a test data set.
5. In the Diagram Workspace, right-click the **Data Partition** node, and select **Run** from the resulting menu. Click **Yes** in the Confirmation window that opens.
6. In the window that appears when processing completes, click **OK**.

# Replace Missing Values

In this example, the variables SES and URBANICITY are class variables for which the value `?` denotes a missing value. Because a question mark does not denote a missing value in the terms that SAS defines a missing value (that is, a blank or a period), SAS Enterprise Miner sees it as an additional level of a class variable. However, the knowledge that these values are missing will be useful later in the model-building process.

To use the Replacement node to interactively specify that such observations of these variables are missing:

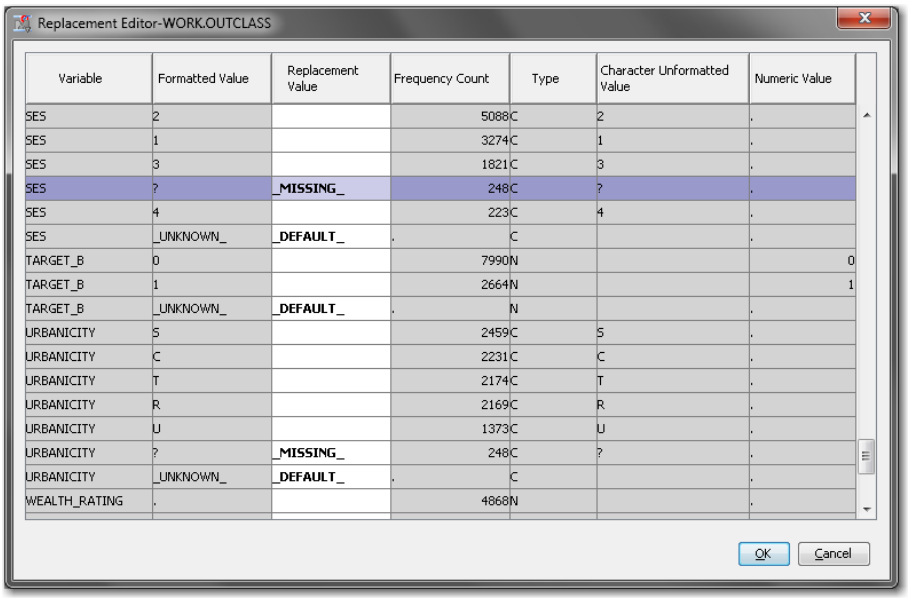
1. Select the **Modify** tab on the Toolbar.
2. Select the **Replacement** node icon. Drag the node into the Diagram Workspace.
3. Connect the **Data Partition** node to the **Replacement** node.


4. Select the **Replacement** node. In the Properties Panel, scroll down to view the Train properties.
   a. For interval variables, click on the value of **Default Limits Method**, and select **None** from the drop-down menu that appears. This selection indicates that no values of interval variables should be replaced. With the default selection, a particular range for the values of each interval variable would have been enforced. In this example, you do not want to enforce such a range.

      *Note:* In this data set, all missing interval variable values are correctly coded as SAS missing values (a blank or a period).

   b. For class variables, click on the ellipses that represent the value of **Replacement Editor**. The Replacement Editor opens.

      - Notice that SES and URBANICITY both have a level that contains observations with the value `?`. In the case of these two variables, this level represents observations with missing values. Enter **`_MISSING_`** as the **Replacement Value** for the two rows, as shown in the image below. This action enables SAS Enterprise Miner to see that the question marks indicate missing values for these two variables. Later, you will impute values for observations with missing values.




- Enter **_UNKNOWN_** as the **Replacement Value** for the level of DONOR_GENDER that has the value **A**. This value is the result of a data entry error, and you do not know whether the intention was to code it as an **F** or an **M**.

Click **OK**.

5. In the Diagram Workspace, right-click the Replacement node, and select **Run** from the resulting menu. Click **Yes** in the Confirmation window that opens.
6. In the window that appears when processing completes, click **OK**.

In the data that is exported from the Replacement node, a new variable is created for each variable that is replaced (in this example, SES, URBANICITY, and DONOR_GENDER). The original variable is not overwritten. Instead, the new variable has the same name as the original variable but is prefaced with REP_. The original version of each variable also exists in the exported data and has the role **Rejected**.

To view the data that is exported by a node, click the ellipsis button that represents the value of the General property **Exported Data** in the Properties Panel. To view the exported variables, click **Properties** in the window that opens, and then view the **Variables** tab. Similarly, you can view the data that is imported and used by a node by clicking the ellipsis button that represents the value of the General property **Imported Data** in the Properties Panel.

*TIP* "Predictive Modeling with SAS Enterprise Miner: Practical Solutions for Business Applications" provides examples of and options for the StatExplore and Replacement nodes. The book also discusses alternate configurations for the Data Partition node.