Predictive Analytics –Exploratory
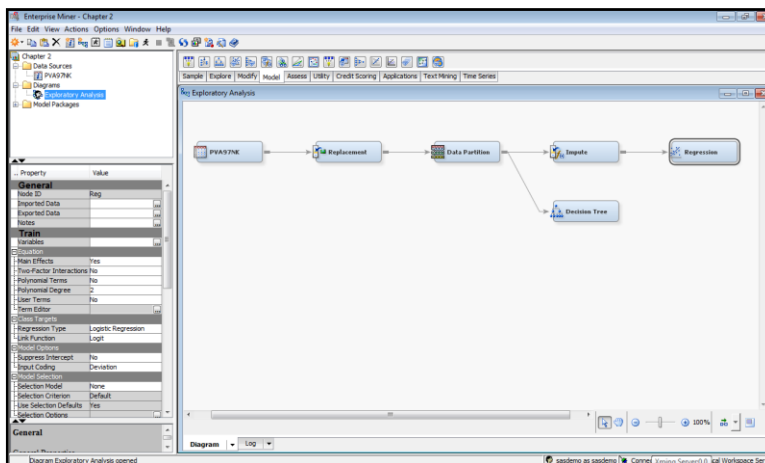
SAS Enterprise Miner: A Primer

Objectives

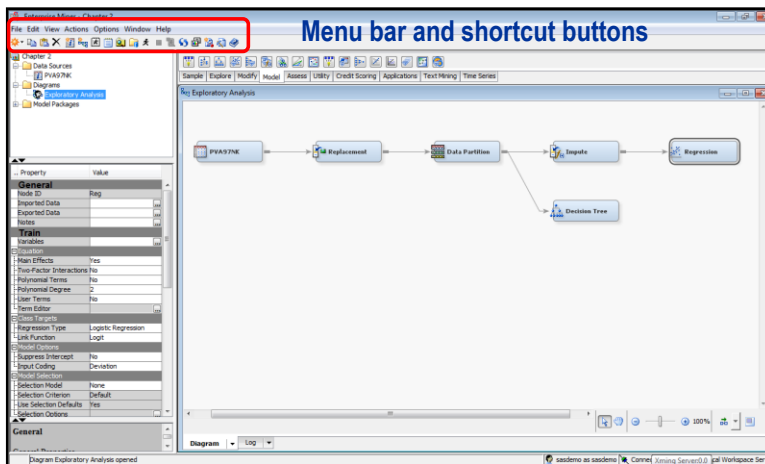- Describe the basic navigation of SAS Enterprise Miner.

88

SAS Enterprise Miner
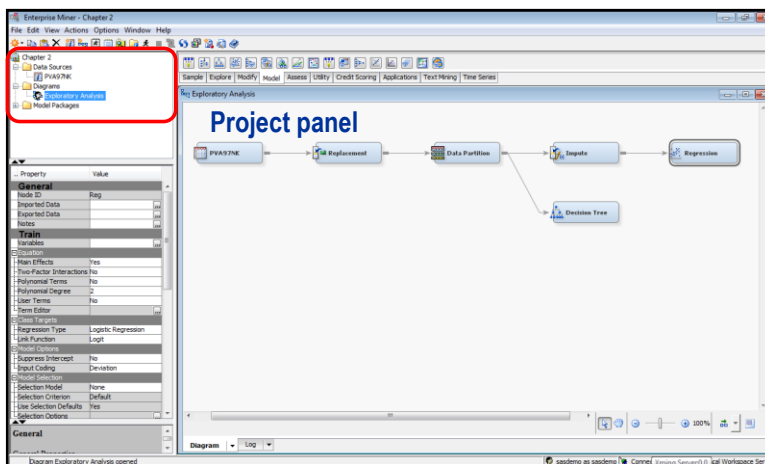
89

Earlier in the course, you learned about many of the features of SAS Enterprise Miner.

SAS Enterprise Miner – Interface Tour

Menu bar and shortcut buttons

90

SAS Enterprise Miner – Interface Tour

Project panel

91

SAS Enterprise Miner – Interface Tour
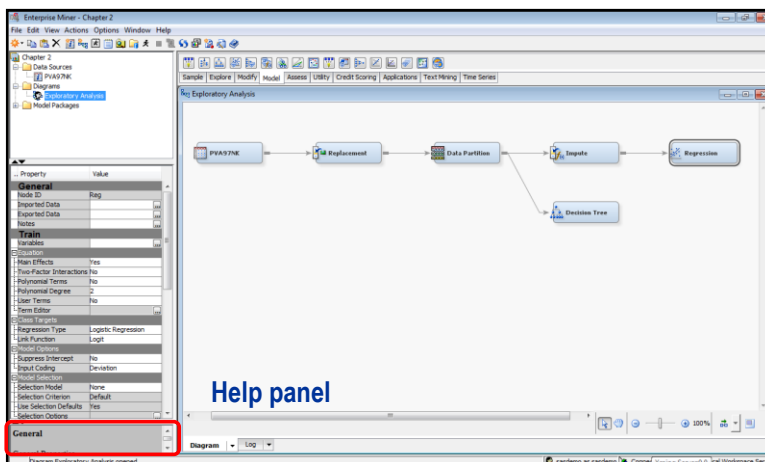
Properties panel

92

SAS Enterprise Miner – Interface Tour
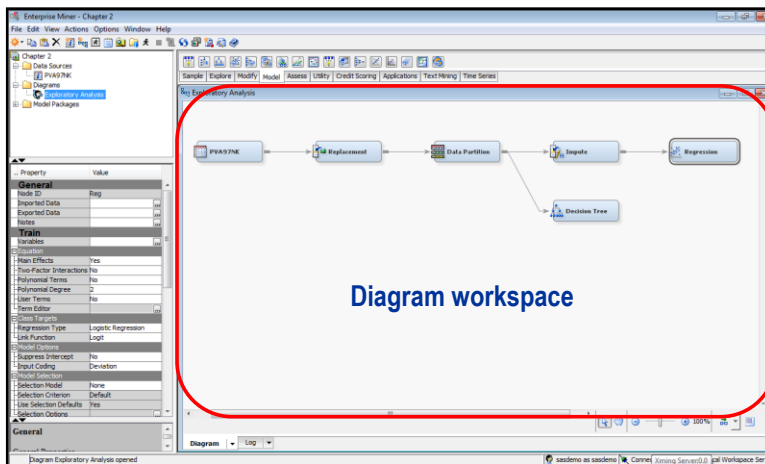
Help panel

93

SAS Enterprise Miner – Interface Tour

Diagram workspace

SAS Enterprise Miner – Interface Tour

Process flow

SAS Enterprise Miner – Interface Tour

Node
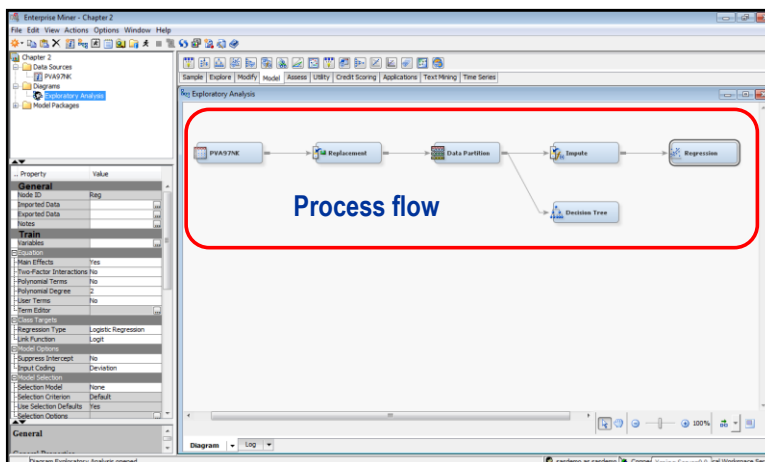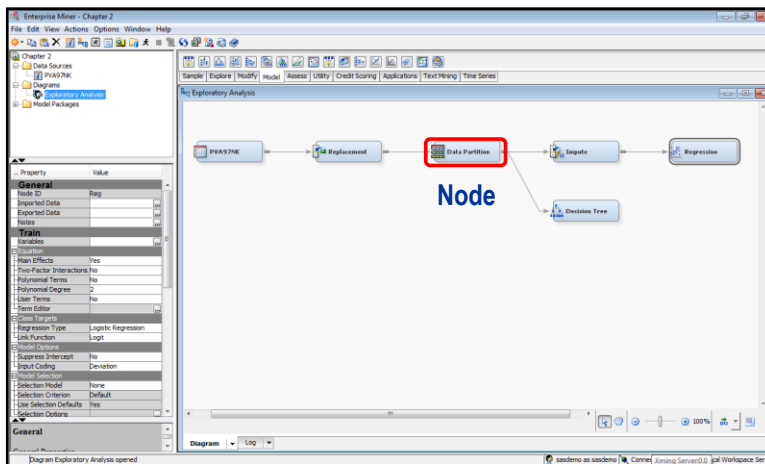
96

SAS Enterprise Miner – Interface Tour

SEMMA tools palette
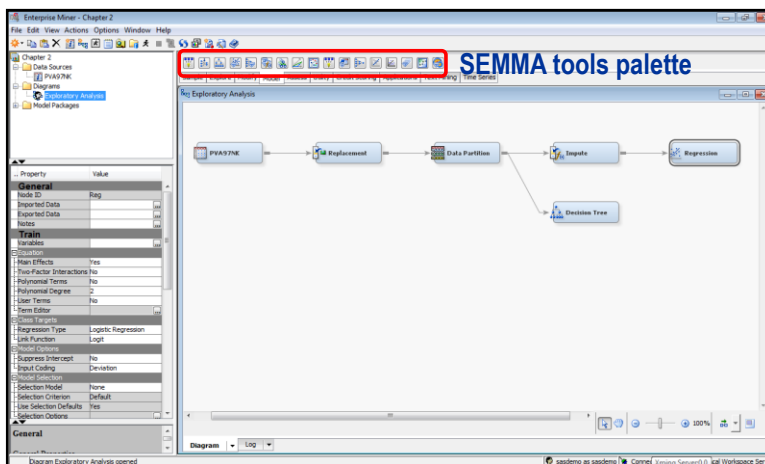
97

Catalog Case Study

Analysis Goal:

A mail-order catalog retailer wants to save money on mailing and increase revenue by targeting mailed catalogs to customers who are most likely to purchase in the future.

Data set: **CATALOG2010**

Number of rows: **48,356**

Number of columns: **98**

Contents: **sales figures summarized across departments and quarterly totals for 5.5 years of sales**

Targets: **RESPOND** (binary)

ORDERSIZE (continuous)

98

The source for the data set used in this chapter is Berry and Linoff (2000). The data represents a mail-order catalog that sells unusual, practical goods. The goal is to increase revenue per mailed catalog and to select customers to mail the 2010 catalog to, based on their past purchasing behaviors. This means selectively mailing to customers who are most likely to purchase and most likely to make a *large* purchase.

The catalog example serves a dual purpose: to illustrate predictive modeling and also to highlight the importance of customer relationship management for this type of application. According to Berry and Linoff (2000, p. 255):

> "Why would readers not involved with catalog retailing be interested in such a case study? … Perhaps the most important reason is that the catalog industry is not like the rest of the brick-and-mortar retail world. Where most retailers work with anonymous transactions, catalogers know who their customers are – and their purchasing patterns over time. In this respect, catalogers are more similar to e-commerce…."

Detailed Description of the CATALOG2010 Data

The columns contain information for two target variables, **RESPOND** and **ORDERSIZE**, and for potential input variables such as a customer's geographic location and methods of payment, as well as his or her historical purchase patterns and volume. Also included is information about how many catalogs the customer has been sent in the past. This data is used to develop a predictive model to identify likely and valuable responders to the next catalog mailing.

Target Variables

RESPOND	1/0 indicator variable identifies customers who made a purchase from a catalog sent to them in Q1 of 2009. Any purchase during Q1 or Q2 2009 is considered a response.
ORDERSIZE	dollar value of the catalog purchases during Q1 and Q2 of 2009.

Geographic Information

COUNTY	county code
STATE	state code
ZIP	ZIP code

Time since Purchases (based on purchases before Q4 2008)

MONLAST	number of months since the last purchase
TENURE	number of months since customer's first purchase

Payment Information

METHPAYM	method of payment
CCPAYM	1/0 indicator variable created from **METHPAYM** that indicates whether a customer has paid using a credit card only
PCPAYM	1/0 indicator variable created from **METHPAYM** that indicates whether a customer has paid using a personal check only
BOTHPAYM	1/0 indicator variable created from **METHPAYM** that indicates whether a customer has paid using both a check and a credit card

Cases where all three variables are 0 are designated as `Unknown`.

Historical Purchase Patterns and Behavior (before Q4 2008)

UNITSIDD	total number of items purchased by the customer
UNITSLAP	average price per unit
UNTLANPO	average units per order
FREQPRCH	number of times a customer placed an order for at least one item
DOLINDET	total dollar demand, gross
DOLNETDT	total demand, net
DOLINDEA	average dollar demand, gross
DOLNETDA	average dollar demand, net
DOLL24	dollar value of purchases in the last 24 months
DEPT01 to **DEPT27**	number of items purchased by the customer in the different departments
TOTORDQ01 to **TOTORDQ22**	total number of orders by calendar quarter starting in 2003 Q1 and ending in 2008 Q2
DOLLARQ01 to **DOLLARQ22**	total dollar value by calendar quarter starting in 2003 Q1 and ending in 2008 Q2

Catalog Case Study: Creating Projects and Diagrams in SAS Enterprise Miner

In SAS Enterprise Miner, you perform all aspects of data mining through a project interface. A project contains materials related to a particular analysis task. These materials include analysis process flows, intermediate analysis data sets, and analysis results.

This demonstration illustrates how to define a project in SAS Enterprise Miner and how to create a diagram.

1. Start SAS Enterprise Miner according to your instructor's instructions.

2. Log on to SAS Enterprise Miner. A welcome window appears.

3. Start a new project by selecting **New Project** or by selecting **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.

4. Click **Next**.

5. Name the project by typing a name (for example, **Chapter 2**) in the **Project Name** field. The **SAS Server Directory** field lists the physical location where the project folder will be created.

   Step 2 of the Create New Project Wizard is used to specify the following information:
   - the name of the project that you are creating
   - the location of the project (***Your instructor specifies the location for your project, if necessary.***)

6. Click **Next**.

7. The next window specifies a location for the project's metadata. The default location is your student directory. Retain this location.

8. Click **Next**.

   Information about your project is summarized in step 4.

9. To finish defining the project, click **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 a Microsoft Excel workbook.

All the data sets referenced in this book are in a predefined library named **ABA1**, so there is no need to create a SAS library.

🖉   Your instructor can load additional data sets to the SAS server for use in your class. To access these data sets, you need to define a library. 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, only if told to do so by your instructor.

1. Select **File ⇨ New ⇨ Library** from the main menu. The Library Wizard - Select Action window appears.

2. The **Create New Library** button is selected. Click **Next**.

3. Type a library name in the **Name** field.

4. Copy and paste the path from the sample LIBNAME statement that your instructor specifies.

5. Select **Next**. The Library Wizard is updated to show the Confirm Action window.

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

6. Click **Finish**.

   All data that is available in the new 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 that are involved in your analysis.

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

1. Select **File ⇨ New ⇨ Diagram** from the main menu to access the Create New Diagram dialog box.

2. Type the name **Exploratory Analysis** in the **Diagram Name** field and click **OK**.

   SAS Enterprise Miner creates an analysis workspace window labeled Exploratory Analysis.

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

Catalog Case Study: Defining a Data Source

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 the steps below to specify the **CATALOG2010** data source.

1. Select **File ⇨ New ⇨ Data Source** from the main menu. The Metadata Source window in the Data Source Wizard appears.

2. ~~The metadata source is a SAS Table, so do not make any changes~~ Select Metadata Repository from the drop down list and click **Next**.

The Data Source Wizard guides you through a multi-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 choice is the SAS Table source. The data for this class is registered in the metadata repository.

3. Select **Browse** in step 2.

4. Double-click the **ABA1** library.

5. Select the **CATALOG2010** table. Click **OK**.

6. Click **Next**.

    The Data Source Wizard proceeds to the Table Information step.

Property	Value
Table Name	ABA1.CATALOG2010
Description	
Member Type	DATA
Data Set Type	DATA
Engine	V9
Number of Variables	98
Number of Observations	48356
Created Date	April 30, 2012 3:16:50 PM EDT
Modified Date	April 30, 2012 3:16:50 PM EDT

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

Catalog Case Study: 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 the proper measurement level of each variable in the source data set.

Follow the steps below to define the column metadata.

1. Click **Next**. The Data Source Wizard proceeds to 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. Click **Next** to use the Basic setting.

   The Data Source Wizard proceeds to the Column Metadata step.

3. Select **Label** to view the column labels.

   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.

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

4. Click **Back** in the Data Source Wizard. This returns you to Metadata Advisor Options.

5. Select the **Advanced** option.

   You can customize the settings for variable rejection and inclusion. To see the settings, select **Customiz**e.

   🖊 This table is read from the bottom up. For example, do you want the advisor to pass through the database and try to detect measurement levels and compute summary statistics? If so, then do you want to reject variables with excessive class values? If so, then how many levels is considered excessive? Do you want to detect class levels? (In other words, do you want to automatically determine whether a numeric variable should be considered nominal based on its number of levels?) If so, what is the threshold for treating a numeric variable as nominal?

   You can overwrite any of these values to suit your needs in the Advanced Advisor Options window.

6. Accept the defaults. Click **OK**.

7. With **Advanced** selected, click **Next** to proceed to the next step.

   A few of the default metadata settings have changed based on the distribution of the variables. For example, a number of variables are now listed as Binary level variables. Several variables are also listed with a Rejected role. The reasons for rejection are from the thresholds set in the Advanced Advisor Options window. You learn more about the reasons for rejections on these specific variables by exploring the data.

8. Select the column heading **Role** to sort by role. Scroll to the bottom of the list. Highlight the variables that are rejected and click **Explore**.
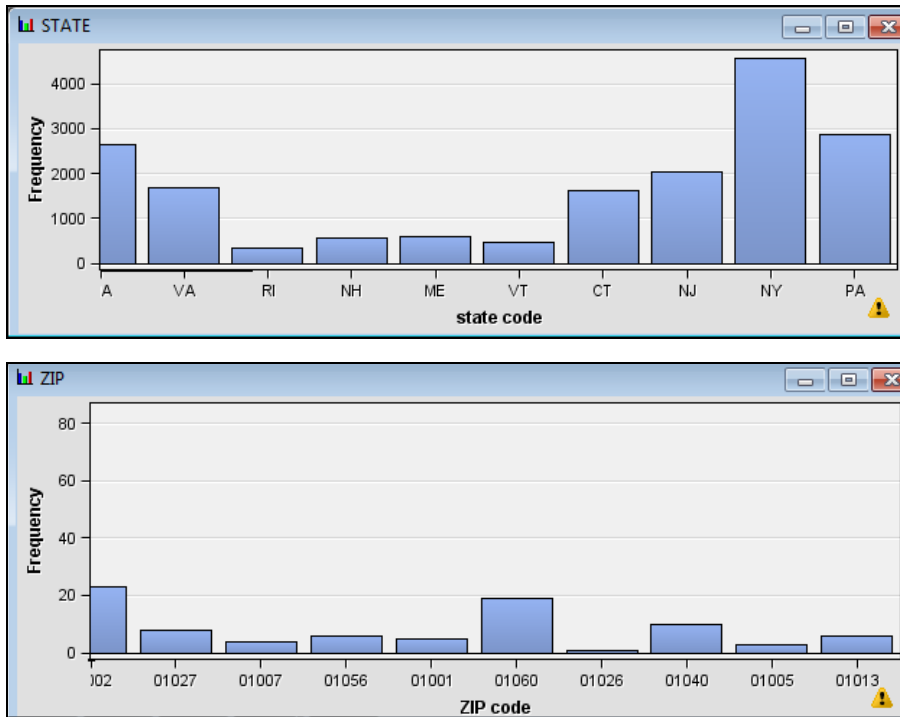
9. Notice the yellow exclamation icon in the lower right corner of each plot. This indicates that there are more values to see. Click on the icons for each plot to see the full range of data.

   ✎   The distributions show that there are many distinct values for each variable. SAS Enterprise Miner has a default rejection threshold for the maximum number of levels for nominal variables. You can change the defaults by selecting **Customize** in the Metadata Advisor Options window in the Data Source Wizard.

10. Close the Explore window.

There are several variables that should be modified. In the steps that follow, you change the variables with a level of **Nominal** to **Interval** except for **METHPAYM, COUNTY, STATE**, and **ZIP**. In addition, you change the **RESPOND** and **ORDERSIZE** variables, currently listed as having a role of **Input**, to target variables. Set the role for the variable **COUNTY** to **Rejected**. Inspect the two date variables: **DTBUYLST** and **DTBUYORG.** Set the roles for these variables to **Rejected.**

11. Select **Input** next to the variable **ORDERSIZE**. From the menu, select **Target**. Repeat the process for the variable **RESPOND**.

12. Use the same process to change all **Nominal** variables to **Interval** except for **METHPAYM**, **COUNTY, ZIP**, and **STATE**. Change **COUNTY**, **DTBUYLST**, and **DTBUYORG** from **Input** to **Rejected.**

13. Select the column heading **Role** to sort by role.

TOTORDQ19	Input	Interval	No		No	.
COUNTY	Rejected	Nominal	No		No	.
DTBUYLST	Rejected	Interval	No		No	.
DTBUYORG	Rejected	Interval	No		No	.
STATE	Rejected	Nominal	No		No	.
ZIP	Rejected	Nominal	No		No	.
ORDERSIZE	Target	Interval	No		No	.
RESPOND	Target	Binary	No		No	.

    The order of the variables might be different, because it is dependent on the order that the role and level updates were done, and how the columns were sorted.

14. Click **Next** to proceed to the Decision Processing step.

    The Data Source Wizard gained extra steps 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. You can learn more about decision processing in the SAS Enterprise Miner Help.

15. Click **Next**.

By skipping the Decision Processing step, you reach the next step of the Data Source Wizard.

This step enables you to create a sample data set to work from in the project. This is particularly useful for very large data sets where operating on the full data would be prohibitively time consuming.

16. Click **Next.**

This 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.

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

The **CATALOG2010** data source is added to the Data Sources entry in the Project panel.
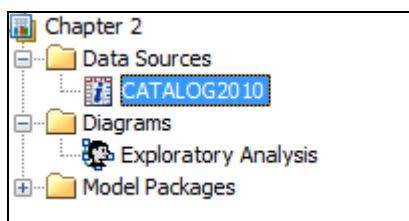
Catalog Case Study: Changing the Sampling Defaults in the Explore Window and Exploring a Data Source

The task of data assembly typically occurs outside SAS Enterprise Miner. Nonetheless, it is worthwhile to explore and validate your data's content. By evaluating 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.

Before performing exploratory analysis, you might want to change the default settings for data exploration. By default, SAS Enterprise Miner explores only the first 2000 rows of the data source.

Follow the steps below 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 appears.

2. Select **Sample Method** ⇨ **Random**.

If **Random** has already been selected, do not change it. 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**. Change the Property Sheet Tooltips property from **Off** to **On**. Click **OK**

The Max fetch enables a larger sample of data (up to 20,000 records) to be extracted for use in the Explore window.

Using these settings, the Explore window uses the entire data set or a random sample of up to 20,000 observations (whichever is smaller).

Exploring the Data with Graphs

Right-click **CATALOG2010** under Data Sources and select **Explore**.

The Explore – ABA1.CATALOG2010 window appears.

The Explore window shows sample properties and sample statistics in the top half of the window, and a preview of the data table is shown in the bottom half.

Creating a Histogram for a Single Variable

You can use the Explore window to browse a data set, but 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. Click **Next**. The Chart Wizard proceeds to the next step, Select Chart Roles.

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

4. Select **Role ⇨ X** for the **CATALOGCNT** variable.

The Chart Wizard is ready to make a histogram of the **CATALOGCNT** variable.

5. Click **Finish**. The Explore window is filled with a histogram of the **CATALOGCNT** 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 **number of catalogs received** has a minimum value of 1 and a maximum value of 23. The mode occurs in the first bin, which ranges between 1 and 3.2. The frequency tells you that there are approximately 12,000 observations in this range. Position the cursor over the tallest bar to see an exact count of 11,866 observations receiving between 1 and 3.2 catalogs.

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 **CATALOGCNT** histogram and select **Graph Properties** from the Option menu. The Properties - Histogram window appears.

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 **22** in the **Number of X Bins** field.

Because **CATALOGCNT** is integer valued and the original distribution plot had a maximum of 23, there is one bin per possible **CATALOGCNT** value.

3. Click **OK**. The Explore window reappears and shows many more bins in the **CATALOGCNT** histogram.
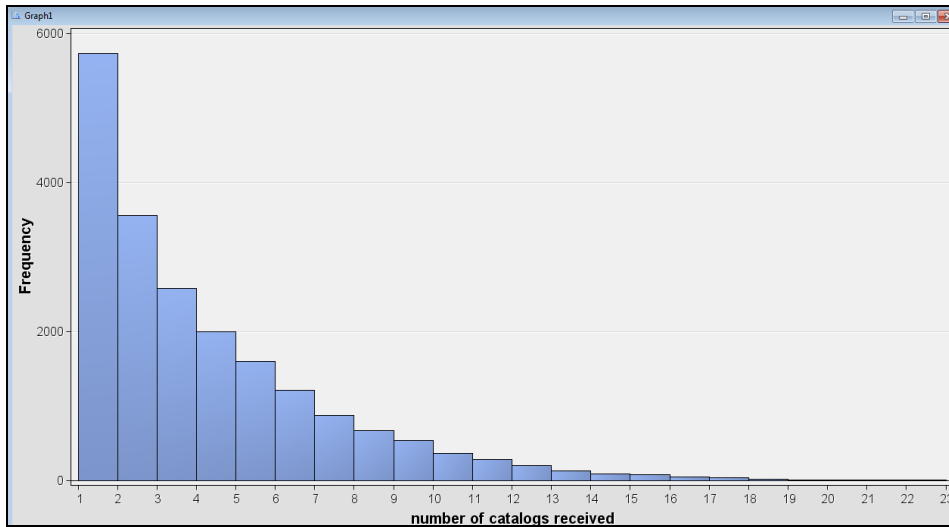
With the increase in resolution, it is clear that most customers have received only one catalog. Very few have received more than 10 catalogs.

Adding Other Graphs

You can add other plots to the Explore window. Perhaps it would be useful to look at the distribution of the target variable and how it relates to the number of catalogs that a customer has received. Are customers who have received more catalogs more likely to respond to a particular catalog mailing? Follow the steps below 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 the chart list and select **Pie** for the type of chart.

3. Click **Next**. The Chart Wizard continues to the Select Chart Roles step.

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

4. Scroll in the variable list and select **Role** ⇨ **Category** for the **RESPOND** variable.

5. Click **Finish** to create the pie chart for **RESPOND**.

The chart shows substantially more cases for **RESPOND**=0 (red, which is the much larger portion here) than for **RESPOND**=1 (blue).

6. Select **Window** ⇨ **Tile** to view all sub-windows of the Explore window simultaneously.
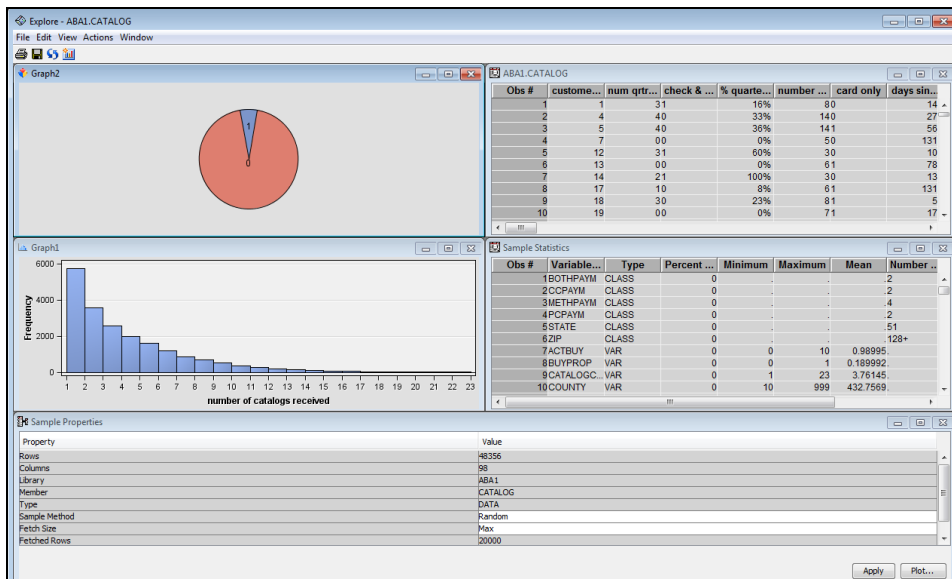
Exploring Variable Associations

All elements of the Explore window are connected. For example, when you select a bar in one histogram, corresponding observations in the data table and other plots are also selected. Follow the steps below to use this feature to explore variable associations.

1. Double-click the **CATALOGCNT** histogram title bar to make it fill the Explore window.

2. Click and drag a rectangle in the **CATALOGCNT** histogram to select cases where the **CATALOGCNT** is greater than or equal to 8.
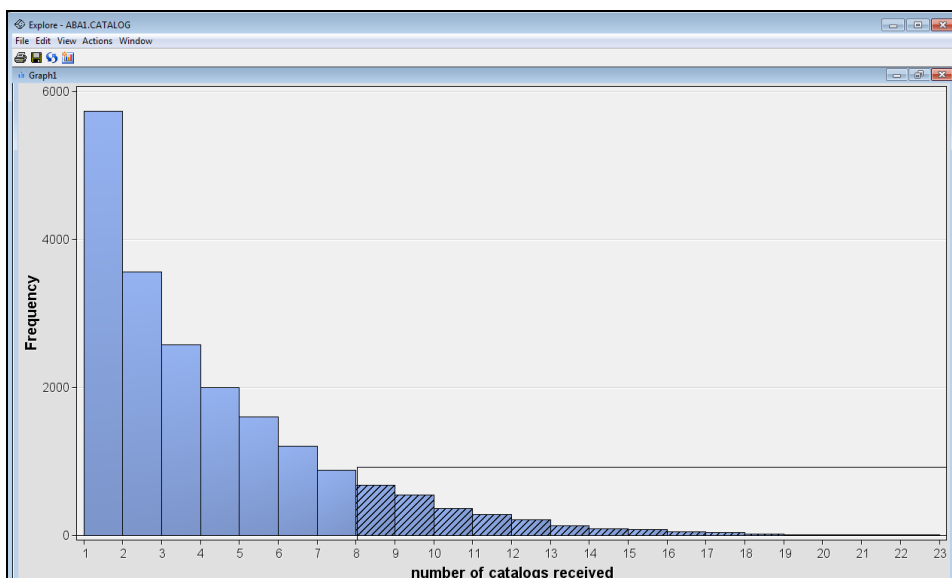

The selected cases are crosshatched. (The vertical axis is rescaled to better show the selection.)

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

Notice that part of the **RESPOND** pie chart is selected. This selection shows the relative proportion of observations with the value of **CATALOGCNT** greater than 8 that do and do not respond. Because the arc on the **RESPOND**=1 segment is thicker, it appears that there is a higher number of responders than non-responders in this **CATALOGCNT** selection. People who receive eight or more catalogs are more likely to respond than people who receive 7 or fewer.

4. Double-click the **RESPOND** pie chart title bar to confirm this observation.
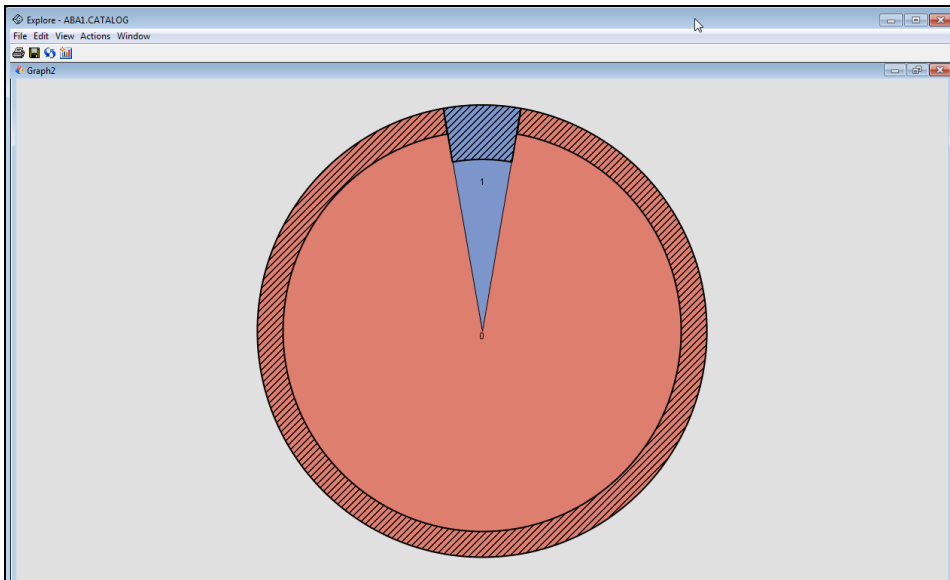
5. Close the Explore window to return to the SAS Enterprise Miner client interface screen.

Additional Analysis

- **Task 1:** You need to perform additional analysis on various variables and make a report.

    o   You might want to study which variables are highly correlated. If you find such variables you can suggest dimension reduction by dropping one of the variables.

    o   You can study in if there are outliers in your variabes

    o   You can make 3-D plots to get a better sense of how independent variables affect the dependent variables.

The Methodology in Practice

Consider the **CATALOG2010** data set and the RFM analysis that you have discussed. Below is an example of applying the data mining methodology to identify profitable customers.

Define the business objective: Of the existing customers who currently receive catalog mailings, identify a group expected to be profitable enough that, on average, they generate more profit than the cost of producing and sending the catalog.

Translate into a data mining task: Perform RFM analysis to use information about how often, how recently, and how much they have purchased in the past to identify groups who had a greater than 6.7% rate of purchasing from the most recent catalog mailing.

Select data: Use customer-level purchasing history and response to the most recent catalog mailing. Define metadata: identify response, define levels of measurement, define roles.

Explore input data: Look at plots of the response and other key variables. (See previous example.)

Prepare and repair data: Check for missing values and so on.

Transform input data: Bin recency, frequency, and monetary value variables into quintiles; concatenate quintiles to produce RFM score.

Apply analysis: Identify the groups whose response rate exceeds 6.7%.

Deploy model: Mail the next catalog to all customers in the profitable customer groups, and to a random sample of existing, past, and candidate new customers. (This enables you to perform a case-control comparison.)

Assess Results: Evaluate the response rate for the targeted customers compared to the response rate of the random sample.

Catalog Case Study: Performing RFM Analysis of the CATALOG Data

To perform RFM analysis, you need to bin the R, F, and M variables into equal-sized quantiles and concatenate their resulting values. But before you can bin it, the recency variable must be recoded. In the **CATALOG2010** data set, **DAYLAST** is coded so that the lowest numbers are the most recent purchases. In order to obtain a code of 5 that is the most recent, you need to reverse-code **DAYLAST** by multiplying its values by -1.

1. Drag the **CATALOG2010** data source oSASCanto the Exploratory Analysis diagram. From the Modify tab, drag the **Transform** node onto the diagram.

2. Position the mouse pointer over the **CATALOG2010** data source until a gray dot appears on the right edge of the node. Click the gray dot and drag the arrow to the **Transform Variables** node.
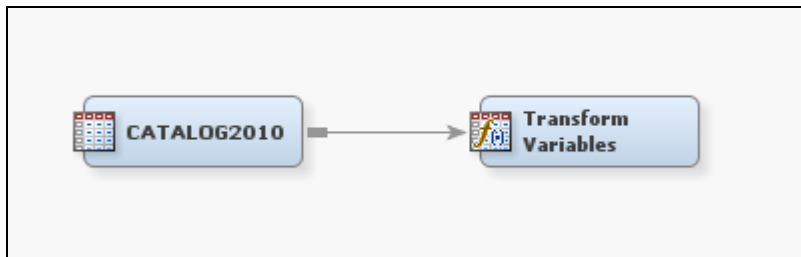
> ✎ Sometimes it is necessary to specify the sample properties within the Transform Variables node properties in addition to specifying them in the Preferences. To do so, go to the Sample Properties section of the Properties panel and change the settings as follows:

Sample Properties	
Method	Random
Size	Max
Random Seed	12345

3. In the Properties panel, click the ellipsis next to **Formulas**.

Train		
Variables		...
Formulas		...
Interactions		...
SAS Code		...
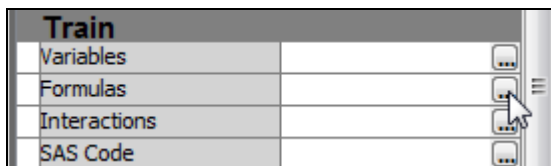
4. Select the **Create** icon at the top left of the toolbar in the Formulas window to create a new column.

5. Type **DAYLAST_REV** in the **Name** field.

6. In the Formula box, type the expression **-1*daylast**. Click **OK**.

    DAYLAST_REV now appears in the list of outputs.

7. Click **OK**.

8. In the diagram, right-click the **Transform Variables** node and select **Rename**. Rename the node **ReverseDaylast**. Click **OK**.

9. Right-click **ReverseDaylast** and select **Run**.

10. Select **Yes**. After the node runs, click **OK**.

Next you create quantiles for the R, F, and M variables.

11. Add another **Transform Variables** node to the diagram and connect the **ReverseDaylast** node to it.

12. In the Properties panel on the left, select the ellipsis next to **Variables**.

13. Select the word **Default** in the Method column next to **DAYLAST_REV** and change the value to **Quantile**. Change the Number of Bins value from **4** to **5**.

14. Make the same changes to the variables **DOLL24** and **FREQPRCH**.

15. Click **OK**.

16. Right-click **Transform Variables (2)** and rename the node **BinRFM**. Click **OK**.

17. Right-click and run the node. Select **Yes**.

18. After the node runs, click **OK**.

To create the final RFM variable, concatenate the values of the R, F, and M variables. To see how this is done, it is useful to investigate the binned values.

19. From the Properties panel on the left, select the ellipsis next to **Exported Data**.

20. Under Port, select **TRAIN**. Select **Explore**.

21. The data appears in a grid. Scroll to the far right to see the computed variables.

Transformed...	Transform...	Transform...
05:-256-high	04:24.75-71.6	05:6-high
04:-573--256	04:24.75-71.6	05:6-high
04:-573--256	05:71.6-high	05:6-high
02:-1826--1008	03:0-24.75	04:3-6
05:-256-high	05:71.6-high	05:6-high
03:-1008--573	03:0-24.75	02:1-2
05:-256-high	05:71.6-high	04:3-6
02:-1826--1008	03:0-24.75	04:3-6
05:-256-high	04:24.75-71.6	05:6-high
05:-256-high	05:71.6-high	05:6-high

The labels are displayed, although the underlying variables are named with the prefix **PCTL_**. Their values are the rank, 01 through 05, followed by the range of values in that bin.

The RFM variable includes only the bin numbers, 01 through 05. The approach that follows generalizes to situations with up to 99 bins per variable.

22. Close the Explore window and the Exported Data window. Add another **Transform Variables** node to the diagram and connect the **BinRFM** node to it.

23. Select the ellipsis next to **Formulas**. Click the **Create** icon. Change the name of the new variable to **RFM**. Fill in the Formula box as shown here:

.. Property	Value
Name	RFM
Type	Numeric
Length	8
Format	
Level	Interval
Label	
Role	Input
Report	No

Formula:

RFM =

substr(pctl_daylast_rev,1,2)||substr(pctl_freqprch,1,2)||substr(pctl_doll24,1,2)

Build...   OK   Cancel

The substring function takes the form substr(*x*, *start*, *continue*) to select a substring of the character variable *x*, starting at position *start*, and going forward *continue* spaces. In this example, the first two values are pulled from each variable.

The concatenation operator, ||, concatenates two values.

24. Click **OK**.

25. Click **OK**. Rename the node **RFM** and run the node. Select **Yes** ⇨ **OK**.

Catalog Case Study: Performing Graphical RFM Analysis

There are many ways to use RFM variables to investigate a target. Perhaps the most useful approach for exploratory purposes is a graph. This demonstration shows two graphical approaches to RFM analysis: a grouped pie chart and a grouped bar chart.

A Grouped Pie Chart

1. With the **RFM** node selected, click the ellipsis next to **Exported Data**. Select **TRAIN** and then select **Explore**.

    🖊 You use data splitting later in the course for honest assessment. If the data has not been split, then the only data set is the train data.

2. Either select **Actions** ⇨ **Plot** from the main menu, or click the **Plot** button at the bottom of the Sample Properties window.

3. Select the **Pie** chart type and then click **Next**. Scroll to **RFM** and change its role to **Category**.

RESPOND		Numeric	response target	BEST12
RFM	None ▾	Character	RFM	
STATE	None	Character	state code	$2
TENURE	Category	Numeric	months since 1st	BEST12
TOTORDQ01	URL	Numeric	tot orders 93Q1	BEST12
TOTORDQ02	Group	Numeric	tot orders 93Q2	BEST12
TOTORDQ03		Numeric	tot orders 93Q3	BEST12
TOTORDQ04		Numeric	tot orders 93Q4	BEST12

4. Change the role for **RESPOND** to **Group**.

PCTL_FREQPRCH		Character	Transformed: lifetime ...	
RESPOND	None ▾	Numeric	response target	BEST12
RFM	None	Character	RFM	
STATE	Category	Character	state code	$2
TENURE	Response	Numeric	months since 1st	BEST12
TOTORDQ01	Color Index	Numeric	tot orders 93Q1	BEST12
TOTORDQ02	Opacity	Numeric	tot orders 93Q2	BEST12
TOTORDQ03	Transparency	Numeric	tot orders 93Q3	BEST12
TOTORDQ04	URL	Numeric	tot orders 93Q4	BEST12
TOTORDQ05	Group	Numeric	tot orders 94Q1	BEST12
TOTORDQ06		Numeric	tot orders 94Q2	BEST12
TOTORDQ07		Numeric	tot orders 94Q3	BEST12

5. Click **Finish**.
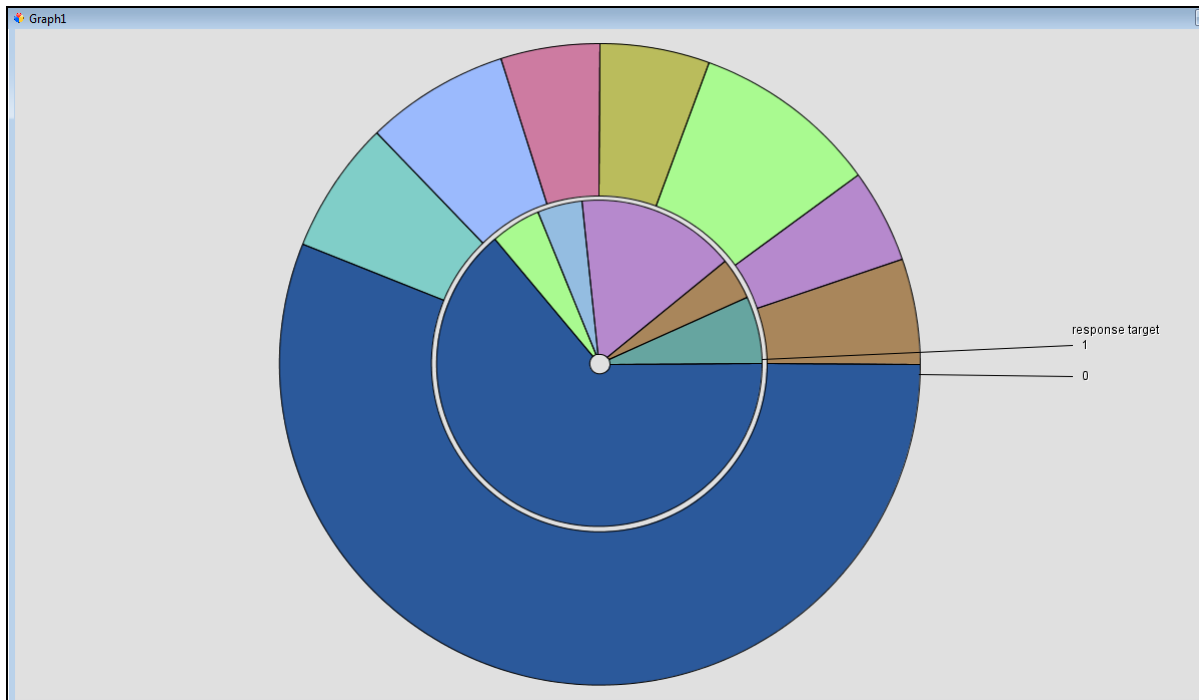
6. The pie chart is not very useful because many bins have been combined into a single slice.
    Right-click the plot and select **Graph Properties**.

7. Clear the box for the **'Other' Slice** option. Click **OK**.
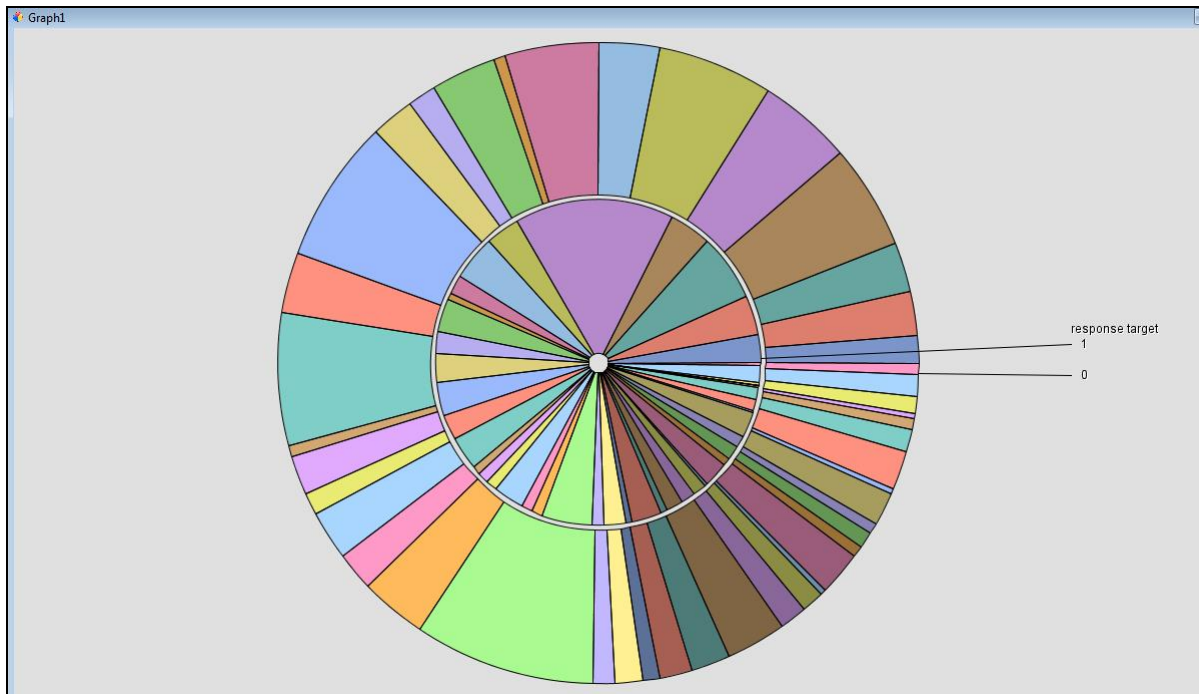
The inner ring of the pie shows responders, and the outer ring shows non-responders. The area of
each slice demonstrates the relative proportion of responders and non-responders in each RFM group.
You can position the mouse pointer over a slice to see its values, or select a slice to highlight those
rows of the data grid.

✎   Although there are 125 groups possible with this RFM analysis, not all groups actually contain observations. Only 44 groups have observations in this data set. Therefore, there are 44 slices on the chart.
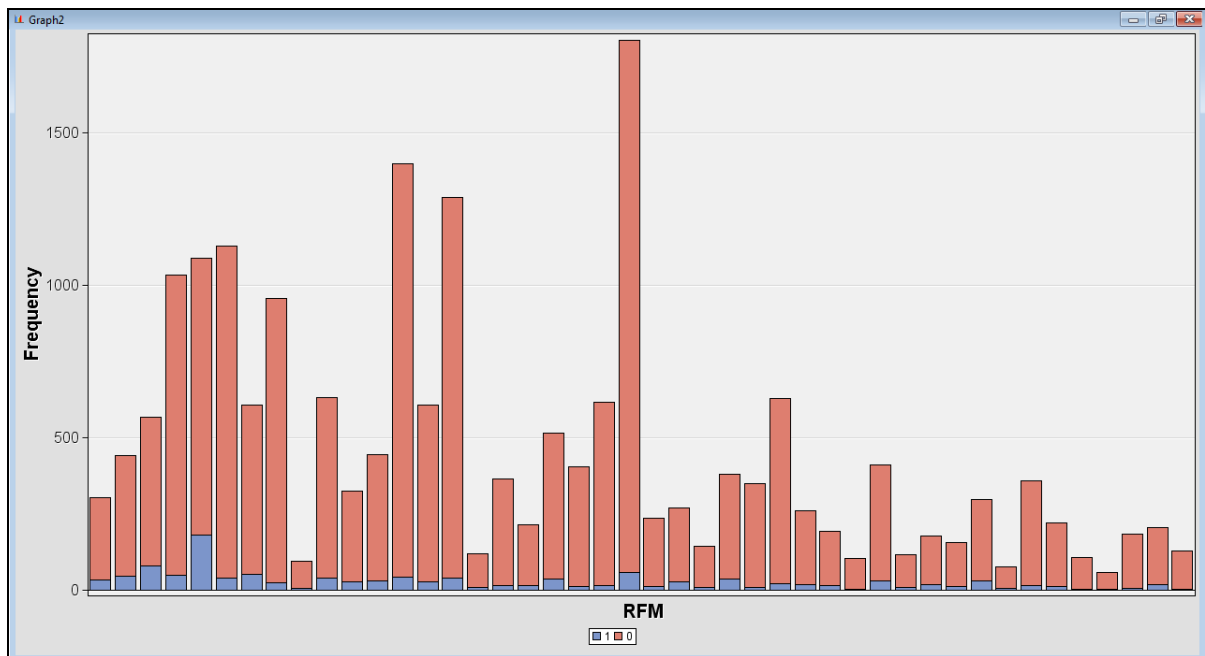
Pie charts with many slices tend to get very busy. Perhaps a grouped bar chart would be easier to read.

8. Select **Actions** ⇨ **Plot**.

9. Select **Bar** ⇨ **Next**.

10. Assign **RFM** as the category variable and **RESPOND** as the group variable.

11. Click **Finish**.

12. The grouped bar chart shows the number of observations in each RFM group as the height of the bar, with the upper portion representing the portion that did not respond. The lower portion of the bar is the proportion that did respond. The fifth bar from the left has a relatively high proportion of responders. Hold your cursor over this bar.
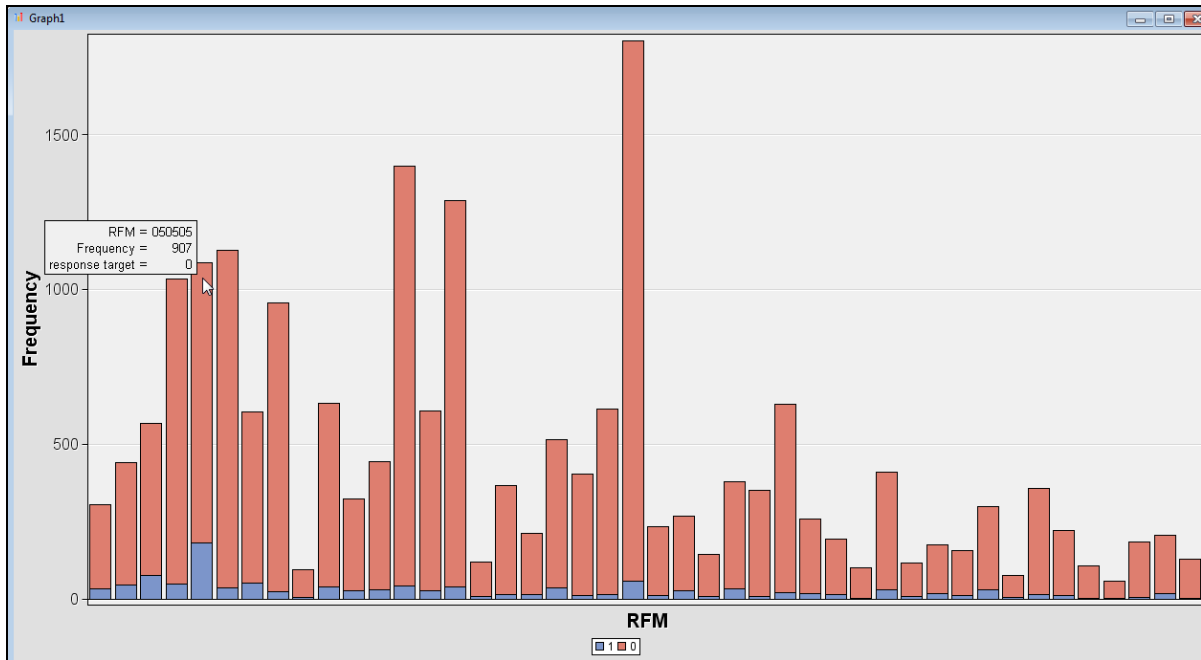
This is group 05(Recency)05(Frequency)05(Monetary Value). It is not surprising that this would be a high-response group.

13. Select **Window** ⇨ **Tile**.

14. You can select groups interactively. Select the **050505** group on the pie graph. This is the inner purple slice. Hold down the CTRL key and click to select both pieces of the purple slice (responders and non-responders).
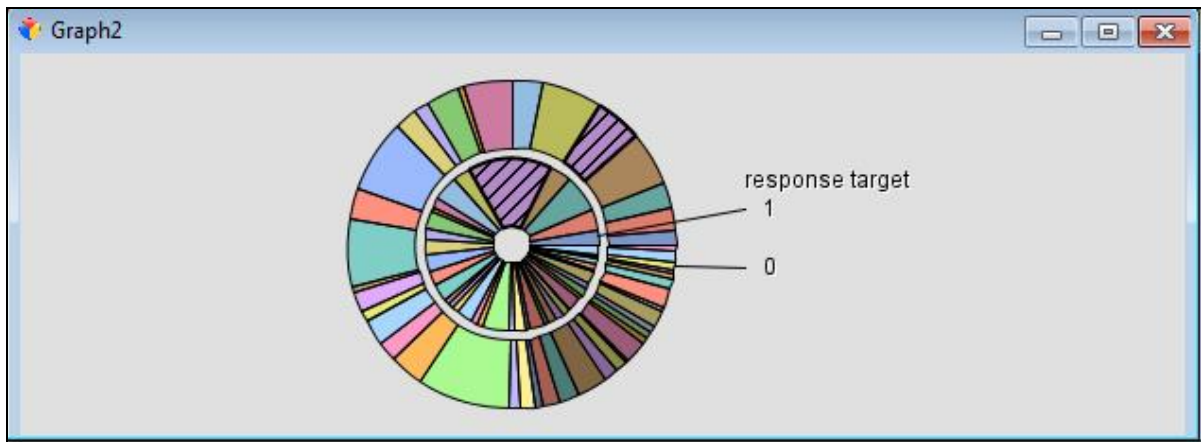
✏️ To select a group on the bar chart, click the small horizontal line dividing the two groups that make up a bar. This selects both categories that make up the bar.
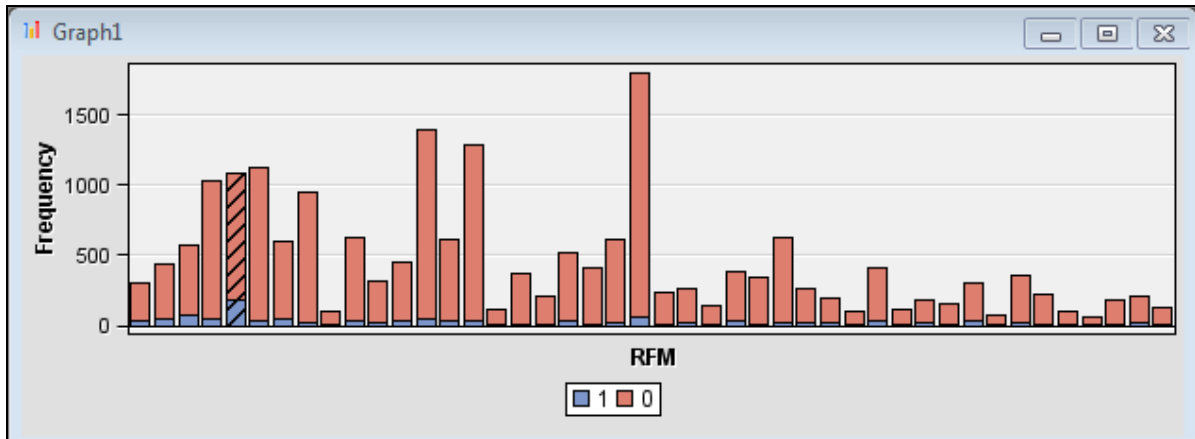

These cases are also selected in the data grid for easy exploration.

| Obs #	custome...	num qrtr...	check & ...	% quarte...	number ...	card only	days sin...
1	1	31		16%	80		146
2	4	40		33%	140		276
3	5	40		36%	141		564
4	7	00		0%	50		1312
5	12	31		60%	30		101
6	13	00		0%	61		782
7	14	21		100%	30		137
8	17	10		8%	61		1317
9	18	30		23%	81		50
10	19	00		0%	71		173

In order to target one of these groups, consider the break-even response rate that was described earlier. If it costs $2.00 to produce and mail a catalog and the average net profit per sale is $30, then you would do best to target cells with a response rate greater than 6.7%.

15. It is easy to determine the proportion of the bin that responded. Place your cursor over the bar as shown below.

RFM = 050505
Frequency =   907
response target =   0
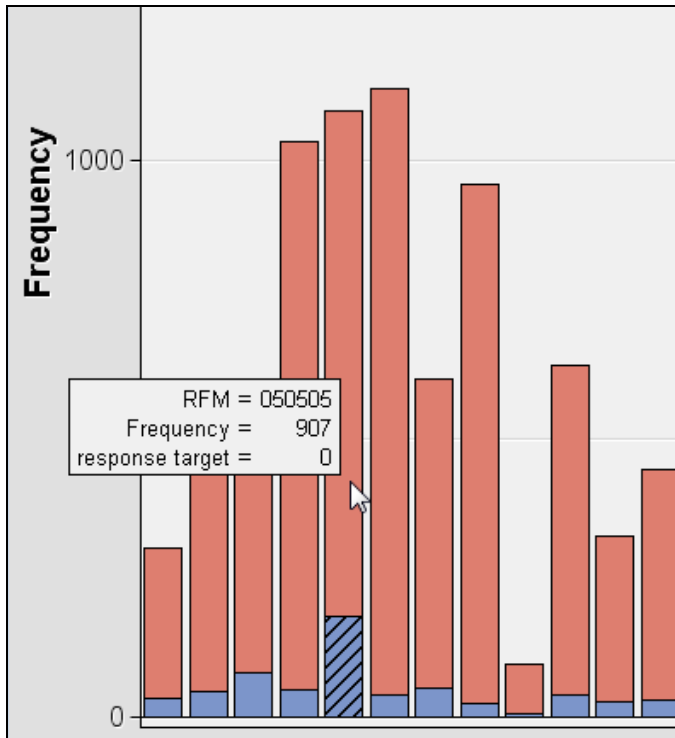
There are 907 non-responders in the group.

16. Place your cursor over the responders section of the bar.

RFM = 050505
Frequency =   181
response target =   1
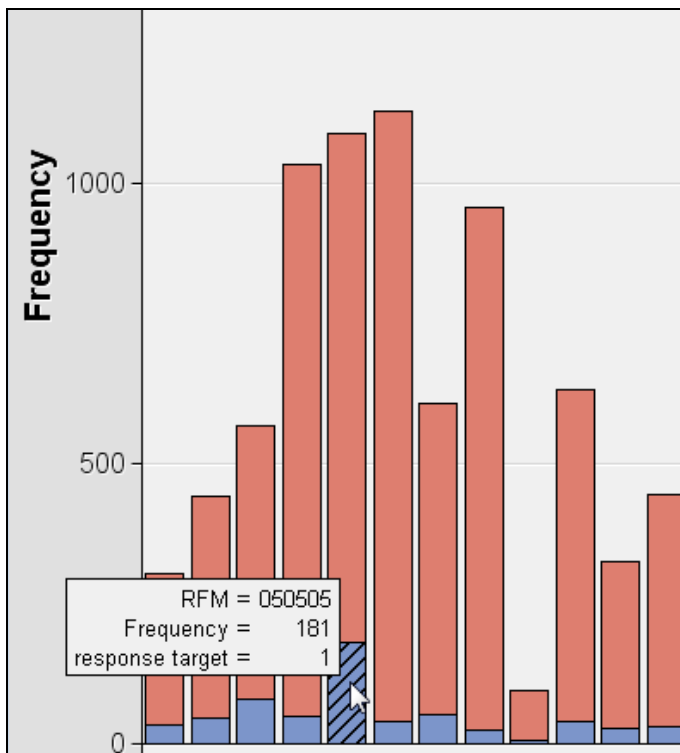
There are 181 responders in the group.

The response rate for the 050505 group is 181 / (181+907) = 16.63%. It would be profitable to mail to these 181 customers. However, this does not suggest how to score new customers, because the quantiles in a new sample might be different than the quantiles obtained here in terms of the original RFM variable values. Furthermore, you would like to send out catalogs to more than 181 customers.

Advanced statistical modeling techniques typically outperform RFM analysis for these purposes and are usually less tedious to implement than RFM analysis.

Limitations of RFM

Only uses three variables
- Modern data collection processes offer rich information about preferences, behaviors, attitudes, and demographics.

Scores are entirely categorical
- 515 and 551 and 155 are equally good, if RFM variables are of equal importance.
- Sorting by the RFM values is not informative and overemphasizes recency.

So many categories
- The simple example above results in 125 groups.

Not very useful for finding prospective customers
- Statistics are descriptive.

209

Task 2:

A national veterans' organization seeks to better target its solicitations for donation. By soliciting only the most likely donors, less money is spent on solicitation efforts and more money is available for charitable concerns. Solicitations involve sending a small gift to an individual and including 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.

Exercise Scenario

Practice with a charity direct mail example.

Analysis Goal:

A veteran's organization seeks continued contributions from lapsing donors. Use lapsing donor response from an earlier campaign to predict future lapsing donor response.

Exercise Data (PVA97NK):

- The data is extracted from previous year's campaign.
- The sample is balanced with regard to response/non-response rate.
- The actual response rate is approximately 5%.

212

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 was 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%.

R, F, M Variables in the Charity Data Set

In the data set **PVA97NK**, the following variables should be used for RFM analysis:

GiftTimeLast   Time since last gift (Recency)

GiftCntAll   Gift count over all months (Frequency)

Monetary value must be computed as follows:

GiftAvgAll*GiftCntAll   Average gift amount over lifetime * total gift count

Use SAS Enterprise Miner to create the RFM variables and bins, and then perform graphical RFM analysis.

213

In order to perform RFM analysis on the **PVA97NK** data set, you must perform some variable transformations. In order for **Recency** to take on higher values for more recent donations, you should multiply **GiftTimeLast** by -1. To derive monetary value, you must compute a new variable: the product of average gift amount and the total number of gifts.

Exercises

1. **RFM Analysis of Charity Direct Mail Data**

   The data set **PVA97NK** contains the following variables:

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

PromCntCardAl	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

a) Define **PVA97NK** as a data source in SAS Enterprise Miner. Use the Advanced Metadata Advisor options to customize the following:
   - Change the Class Levels Count Threshold from 20 to 5.
   - Change the Reject Levels Count Threshold from 20 to 80.
   - Reject the variable **TargetD**.
b) Create a new diagram and transform the R, F, and M variables as described previously to create four bins of each variable. Concatenate them to create an RFM variable.
c) Explore the data and perform graphical RFM analysis using a grouped pie chart and a stacked bar chart.
d) Calculate response rate for 040404 and 030303 group?
e) Each promotional mailing (request for a gift) costs $2.3, and the average donation is about $21 What is the break-even response rate for this promotion? Do any RFM cells exceed this response rate? Remember to account for the fact that in the population, 95% of mailings are not responded to, while this sample is oversampled to 50% responders and 50% non-responders.