**IMAT3613 Data Mining**

References will be made to the text: **Data Mining using SAS Enterprise Miner: A Case Study Approach (SAS Publishing).**

**The pdf can be obtained from Blackboard.**

**Week 2 (Lab Lesson 1): A brief overview of SAS. Creating a Process Flow Diagram in SAS Enterprise Miner. Exploratory data analysis of SAMPSIO.HMEQ dataset.**

**Introductory Task: On-line SAS tutorial**

*(Time Scale in lab: max 20mins)*

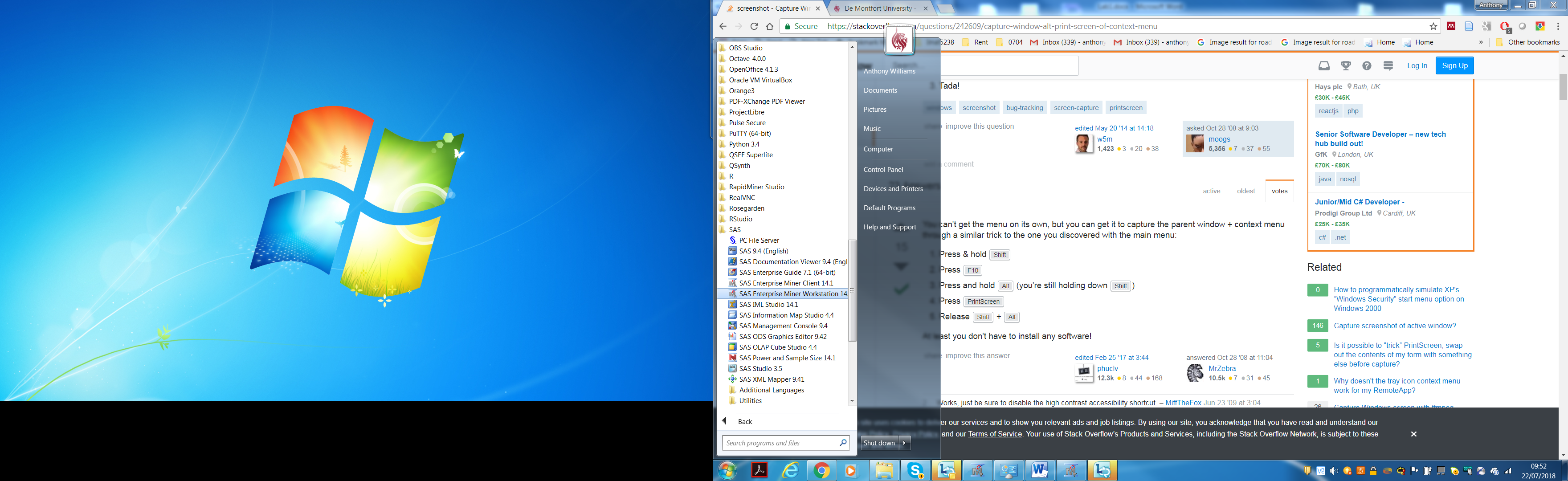
**Objectives:**

The aim of this practical is to give you an overview of the SAS system.

**By the end of this practical you will:**

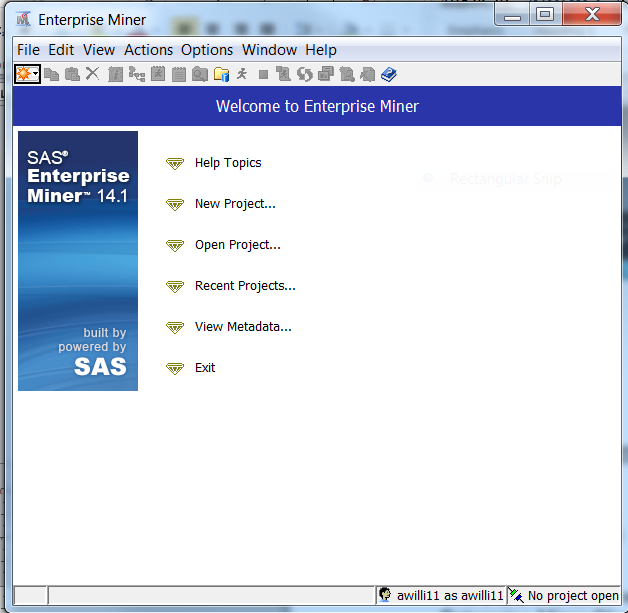
* Have an overview of the SAS system.
* Be familiar with the main components of the screen layout in SAS.

**Enter the SAS System** via the desktop by: **Start ->** **Programs ->** **SAS v9.4 -> Enterprise Miner Workstation Client 14.1**

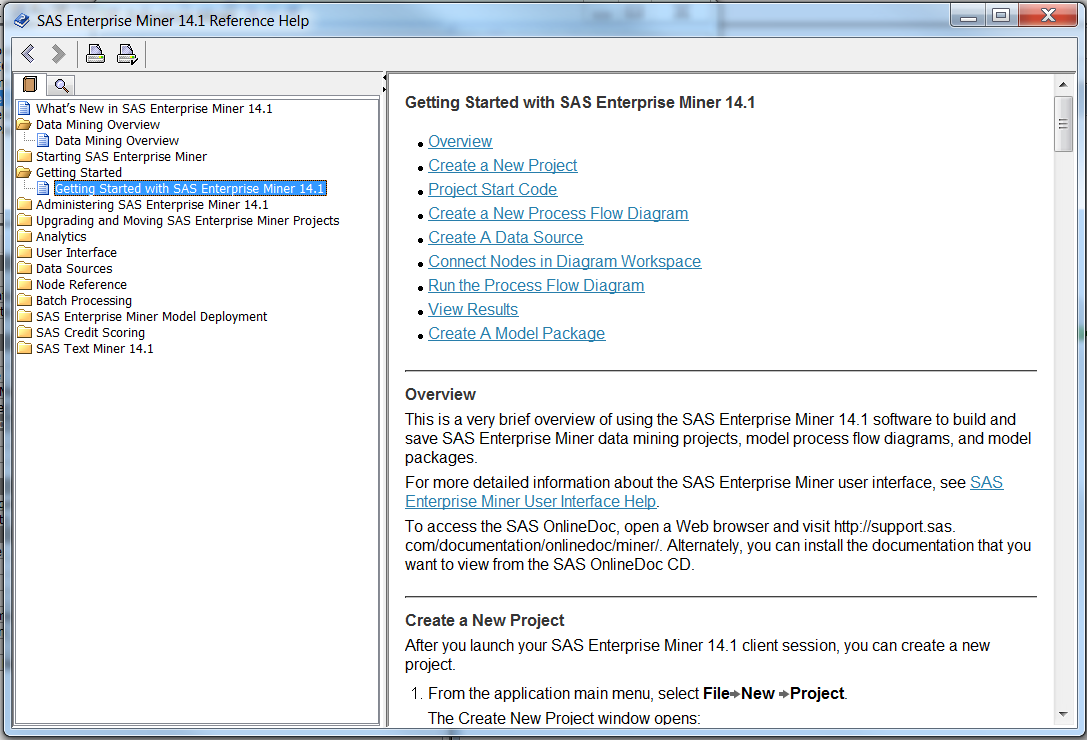


There are two icons very similar to each other do not, I repeat do not select the Enterprise Miner Client, this is the client server application which we will not be using this term. If you accidently click on this icon, you will be asked for a user id and an password. Close down the application and pick the workstation client.

If all is well you will be greeted by a splash screen and the following project area.



Select the menu option Help, then select Contents. Read the data mining overview, then select *“Getting Started”*.



Read the overview. The parts of particular interest are the following links:

a) see an overview of the SAS Enterprise Miner User Interface.

b) explore the SAS workspace

c) working with SAS datasets.

Make sure you are clear about the following:

1. The name and purpose of each of the 6 main SAS GUI components. The two main windows are diagram and properties. Tool ribbon bar and a node ribbon bar.
2. What a SAS library is. Note libname is an internal alias for a dataset, follows a library path akin to SAS internal file structure. Look up libname or libref in the help.
3. Understand the distinction between, project, library, data table, diagram, node. One project can contain many data objects and many diagram objects, one diagram can contain one or more workflow, and a workflow is composed of nodes.

# Data Exploratory Analysis Task

*(Time scale in lab: 1 hour + own time)*

**Objectives:**

The aim of this practical is to use SAS Enterprise Miner to start conducting an exploratory data analysis on the SAMPSIO.HMEQ dataset.

**By the end of this practical you will be able to use SAS Enterprise Miner to:**

* Set up the initial project and diagram.
* Access the SAMPSIO.HMEQ data through a SAS library.
* Start creating a Process Flow Diagram by adding the nodes, using the cursor, identifying the input data.
* Start to understand the data by:

1. Identifying target variables
2. Inspecting the distribution
3. Modifying variable information
4. Investigating descriptive statistics
5. Inspecting Default Settings in the Data Partition Node.

# Setting Up the Initial Project and Diagram

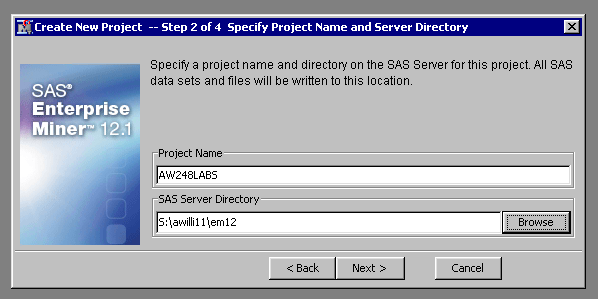
SAS Enterprise Miner organises data analyses into projects and diagrams.

Each project may have several process flow diagrams, and each diagram may contain several analyses. Typically each diagram contains an analysis of one data set.

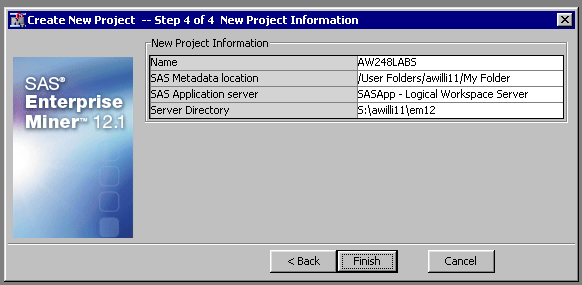
Follow these steps to create a project:

**1** Note if using the server version projects must have a unique name, prefix directory with your initials and last three digits from your student ID. From the SAS menu bar, select **File -> New -> Project**

**2** In the create project dialog box type a name for the project, such as **USERNAMEDDESCRIPTION.** An error will occur if the name is not unique.BUT Modify the location of the project folder by clicking Browse and locating a suitable drive. This will more likely be the H:\ drive and your home directory.



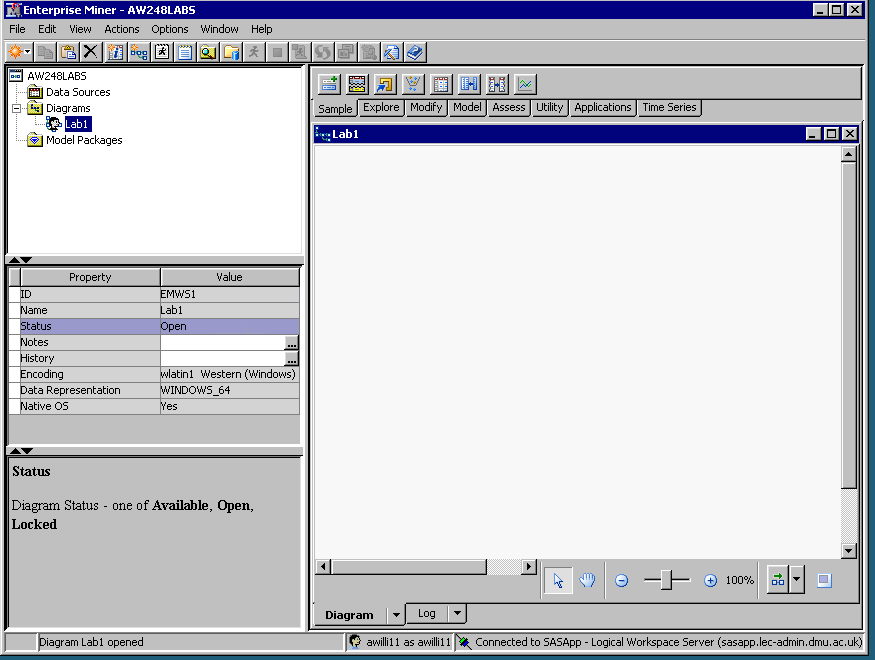
**3** Select next twice, a summary of the project workspace appears. Then click finish.



A blank project is created.

You must create a “diagram” this is a container which contains the data mining workflow. A project can contain many diagrams a diagram can contain one or more workflows.

**4** Select menu option new diagram call it lab1.



SEMMA Toolbar Tab

Blank Diagram

Properties Box

Project Navigation Box

A new object appears under the diagram in the project navigation panel and a blank canvas appears. A properties panel on the left confirms the name of the diagram.

NOTE It is strongly recommended that in the module you only create ONE project and then add diagrams to this one project that have a descriptive meaning either the dataset or lab number.

The SAS Enterprise Miner window contains the following interface components:

* The Project Navigator — enables you to manage projects and diagrams and add tools to the Diagram Workspace.

Note When a tool is added to the Diagram Workspace, the tool is called a node.

* Diagram Workspace — enables you to build, edit, run, and save process flow diagrams.
* Tools Bar — contains a customisable subset of Enterprise Miner tools that are commonly used to build process flow diagrams in the Diagram Workspace.
* Each object on the diagram canvas can be configured by the property box on the left-hand side of the diagram.

# The SAMPSIO.HMEQ dataset

Ref pp20-21 *Predictive Modelling: Problem formulation.*

Background.

A financial services company offers a home equity line of credit to its clients. The company has extended several thousand lines of credit in the past, and many of these accepted applicants (approximately 20%) have defaulted on their loans.

By using geographic, demographic and financial variables, the company wants to build a model to predict whether an applicant will default.

The data is housed in the SAMPSIO library and refers to the credit rating of loan

applicants and is entitled HMEQ (Home Equity). It is a large dataset of 5960

records and consists of 13 variables: BAD, CLAGE, CLNO, DEBTINC, etc.

**The SAMPSIO.HMEQ dataset continued**

BAD is the response (or target) variable and indicates whether an applicant

defaulted on the home equity line of credit. For example, BAD takes the value 1 if

a loan applicant has ever defaulted on loan payments and 0 if not.

There are 12 input variables to model whether each applicant defaulted. These variables, along with their model role, measurement level, and description are shown in Table 1.

**Table 1** Variables in the SAMPSIO.HMEQ Data Set

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

# Task 3: Creating a Process Flow Diagram

Ref pp21-24 *Creating a Process Flow Diagram: Adding the Nodes, Using the Cursor, Identifying the input data*

**1** The first step is to add a data source to the project panel. Select file menu **New add data source**. Data can be imported from text files and Excel spreadsheets. On this occasion you are going to import the data from a SAS table.

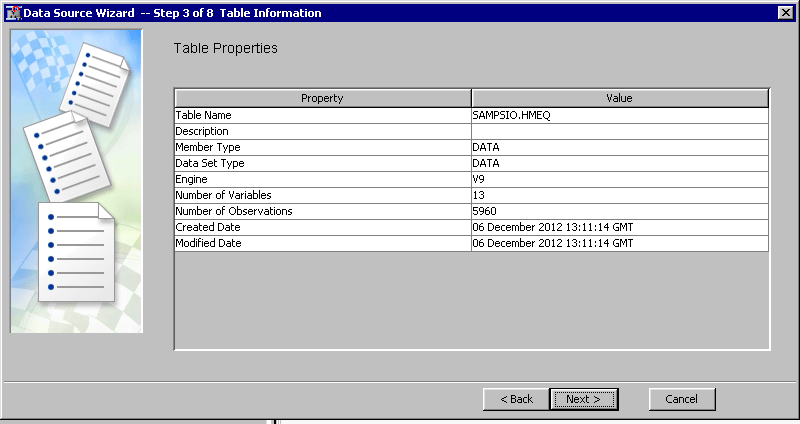
The following dialog box appears



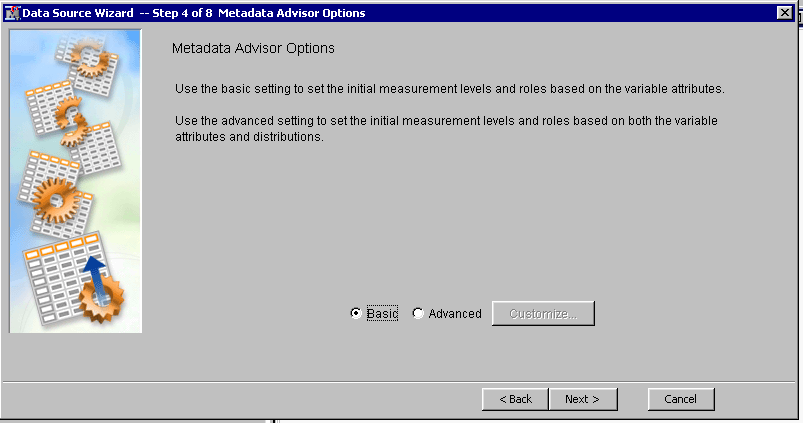
Click on the **Browse** button, which provides you access to SAS libraries. SAS libraries are SAS’s internal storage of data files, alias that point to a directory structure on disk.

**Select** the library **SAMPSIO**, and then **select** the table **HMEQ** and click the **OK** button.

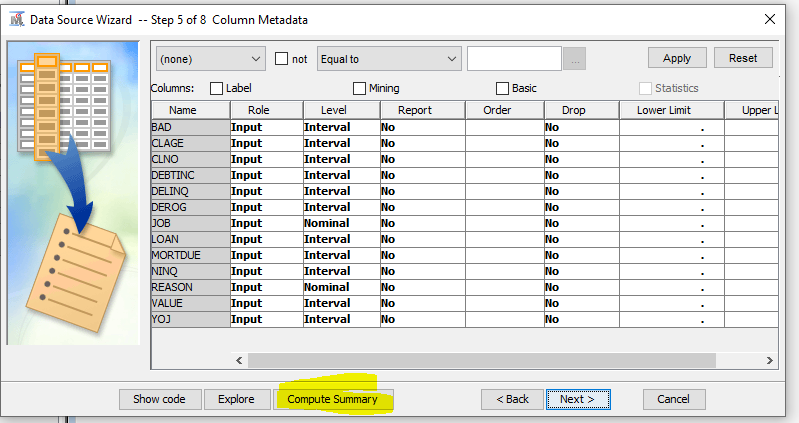
Select **Next**, you are then presented with a list of table properties number of rows and number of variables.



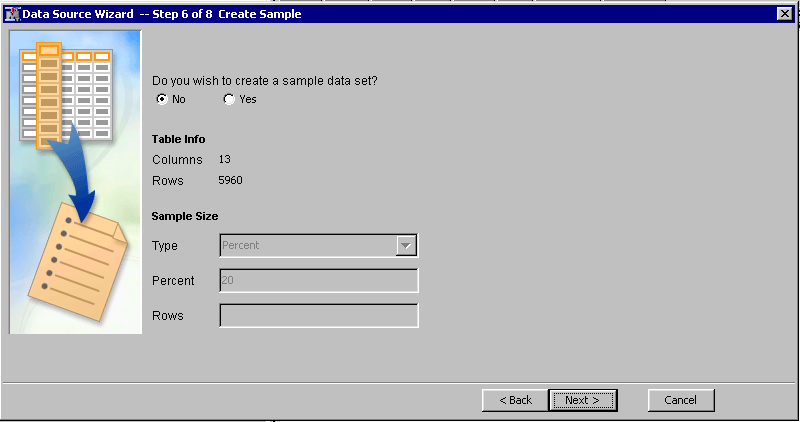
Select **Next**



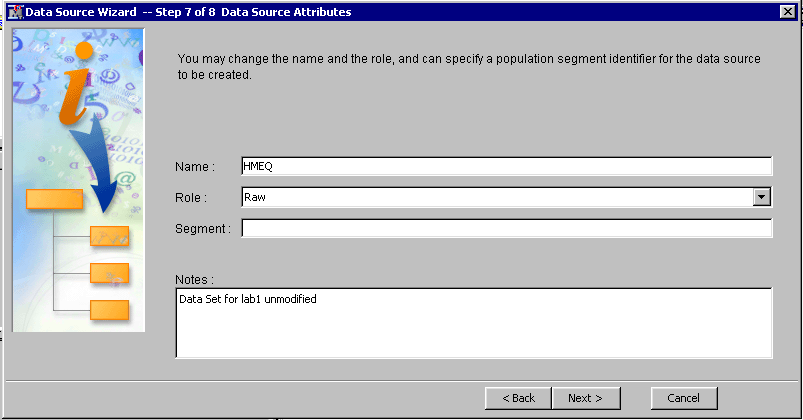
**Check** the radio button for the basic metadata advisor options, Select **Next**,



There are options to change the variable levels, explore data and calculate summary statistics leave these options for now and select **Next**.

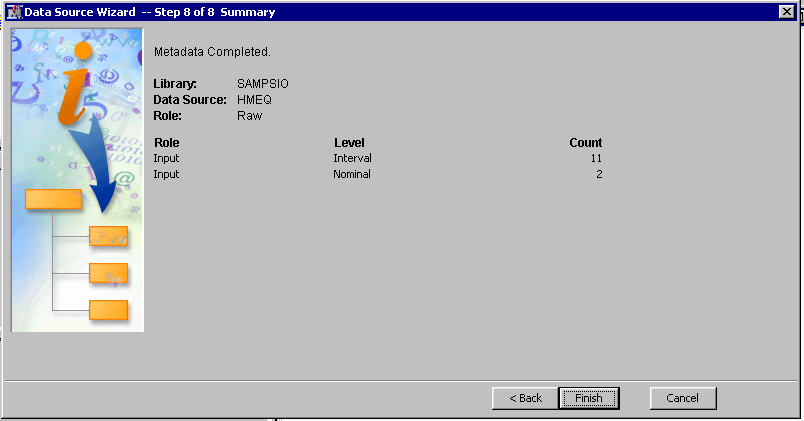


You are then presented with an option to create a sample data set, select **No** and then **Next**.

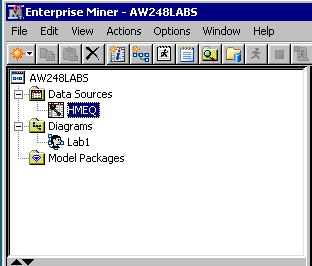


Add notes if you wish, make show Role is set to **“RAW**” through the drop down box. Select **Next**.

A summary card appears then select “**FINISH**”.



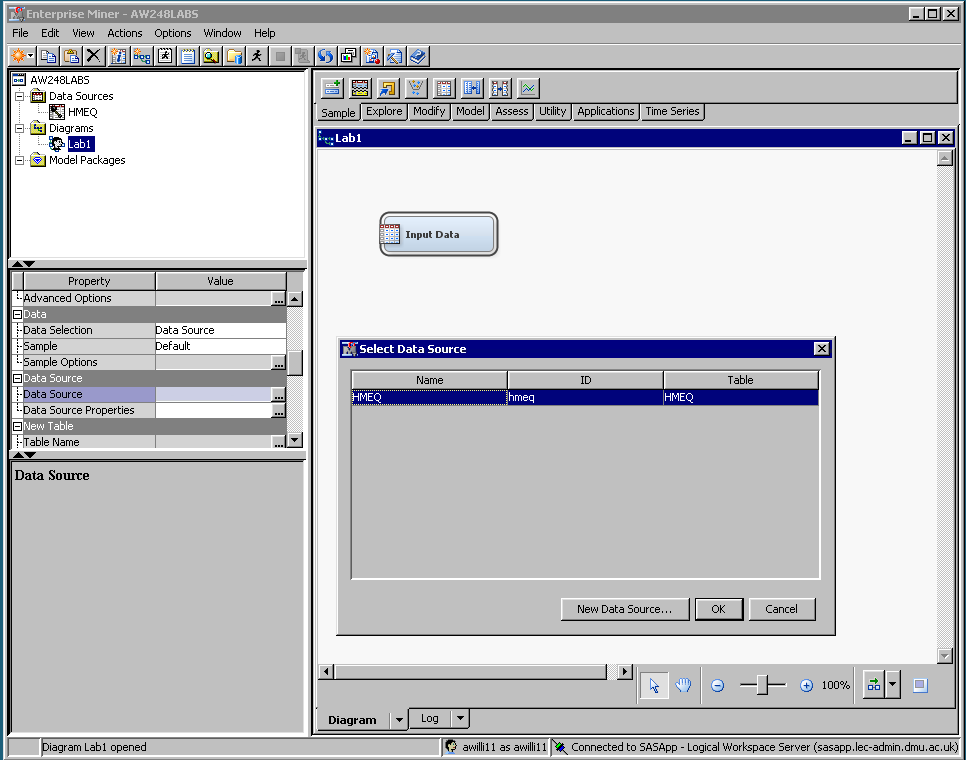
A data source appears called HMEQ in the project navigation box on the left.



# Importing data into the data input node

Add an Input Data Source node by dragging the node from the Tools Bar to the blank diagram canvas. Nodes are organised under the anagram SEMMA, earlier referred to the SEMMA toolbox bar tab. The data input node can be found under the sample tab fifth from the right. Add this node to the blank canvas.

**Select** the input node with the mouse. **View** properties for this node in the property box, located on the left-hand side of the diagram. Scroll down to the data section. Select data sources click on the **ellipsis “ …”**



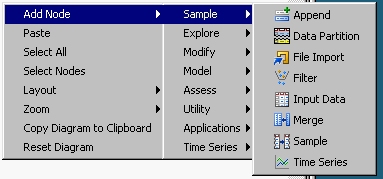
**Select** **HMEQ** data table select **OK**.

# Connecting Nodes

The shape of the cursor changes depending on where it is positioned.

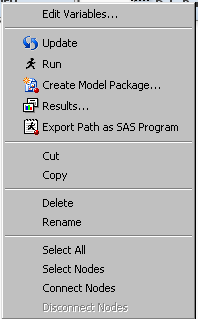
If the cursor is over the node it changes shape to a cross hair, whilst in this mode if you left click you can move the node to any position on the canvas.

Right-click in an open area to see the following pop-up menu:

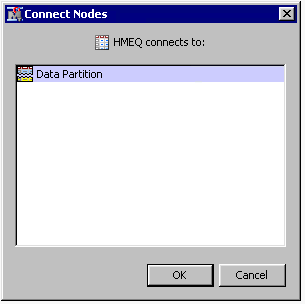


The hierarchical menu allows you to add further nodes to the workspace. Select the Data Partition node. The data partition node allows you to split the data into training, validation and test set. For now just add the node we will configure it later.

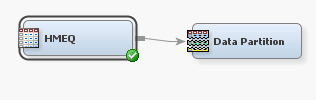
Right click on the data input node and a different set of hierarchical menus appears



Select connect node and a dialog box appears select the HMEQ data set connect to Data Partition node.



And an arrow connects the two nodes. The arrow signifies the movement of data from left to right, the direction is important. Each node processes the data or applies functions. Each node may have an input or an output. This is analogous to Unix scripting with a pipe command. However more complicated process for example forks and nested processes can be built up.

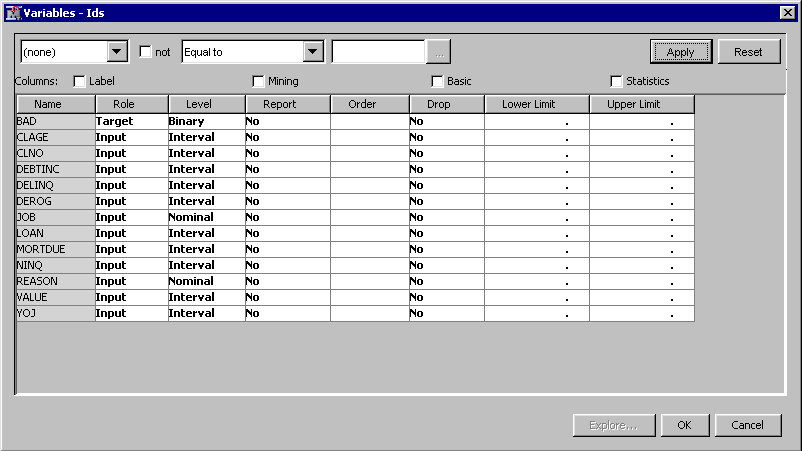


An alternative method of connecting nodes is to hover over the right-hand side of the node and the cursor changes to a pencil, draw a line between two nodes.

# Viewing Data and understanding meta data.

Ref pp24-26 *Creating a Process Flow Diagram: Modifying Variable Information,*

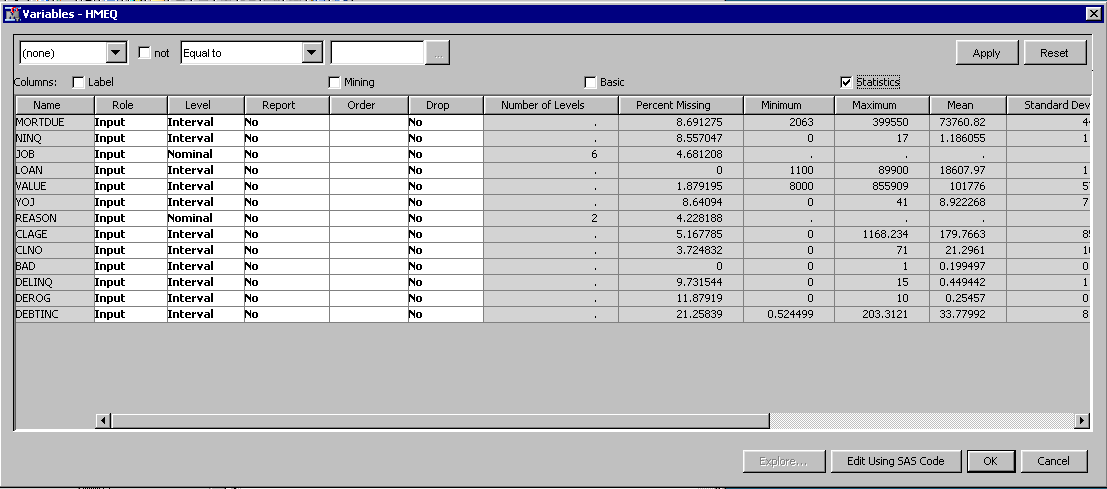
*Investigating Descriptive Statistics, Inspecting Default Settings in the Data Partition Node*



All analysis packages must determine how to use variables in the analysis.

SAS Enterprise Miner uses metadata in order to make a preliminary assessment of how to use each variable. By default, it takes a random sample of observations from the data set of interest, and uses this information to assign a model role and a measurement level to each variable. A measurement level is the number of different values a variable can obtain.

The following table shows a portion of the information for each of the 13 variables in this example:



* The first variable listed is BAD. Although BAD is a numeric variable in the data set, SAS Enterprise Miner identifies it as a binary variable since it has only two distinct non-missing levels in the metadata sample.
* *The next three variables (LOAN, MORTDUE, and VALUE) are assigned an interval measurement level since they are numeric variables in the SAS data set and have more than twenty distinct levels in the metadata sample. This default can be modified in the property dialog box, under columns, adviser property section, advanced options, class level count threshold. Previous versions had the default set to 10 levels.*
* The variables REASON and JOB are both character variables in the data set, yet they have different measurement levels. REASON is assigned a binary measurement level because it has only two distinct non-missing

levels in the metadata sample. JOB, however, is assigned a nominal measurement level since it is a character variable with more than two levels.

* DELINQ has been assigned an ordinal measurement level. The

assignment of an ordinal measurement level occurs when each is a numeric variable with more than two, but not more than ten, distinct non-missing levels in the metadata sample. This often occurs in variables that contain counts (such as number of children).Since this assignment depends on the number of levels for each variable in the metadata sample, ***the measurement level of DEROG or DELINQ for the analysis may be set to interval.***

**Note For the purpose of this analysis, treat the remaining variables (YOJ**

**through DEBTINC) as interval variables.**

* The model role for all binary, interval, ordinal and nominal variables is set to input by default.

Note summary statistics are only calculated on interval type variables.

**Task 6: Modifying Model Roles and Measurement levels**

Model Roles

BAD is the response variable for this analysis, so the model role for BAD needs to be changed from input to target by:

Select the cell adjacent to BAD varaible in the **Model Role** column of the row for BAD then select

**Set Model Role** **-> target** from the drop down menu.

Measurement levels

Change the measurement level for DELINQ to interval by:

Right-clicking the **Measurement Level** column of the row for DELINQ and then select **Set Measurement Level -> interval**

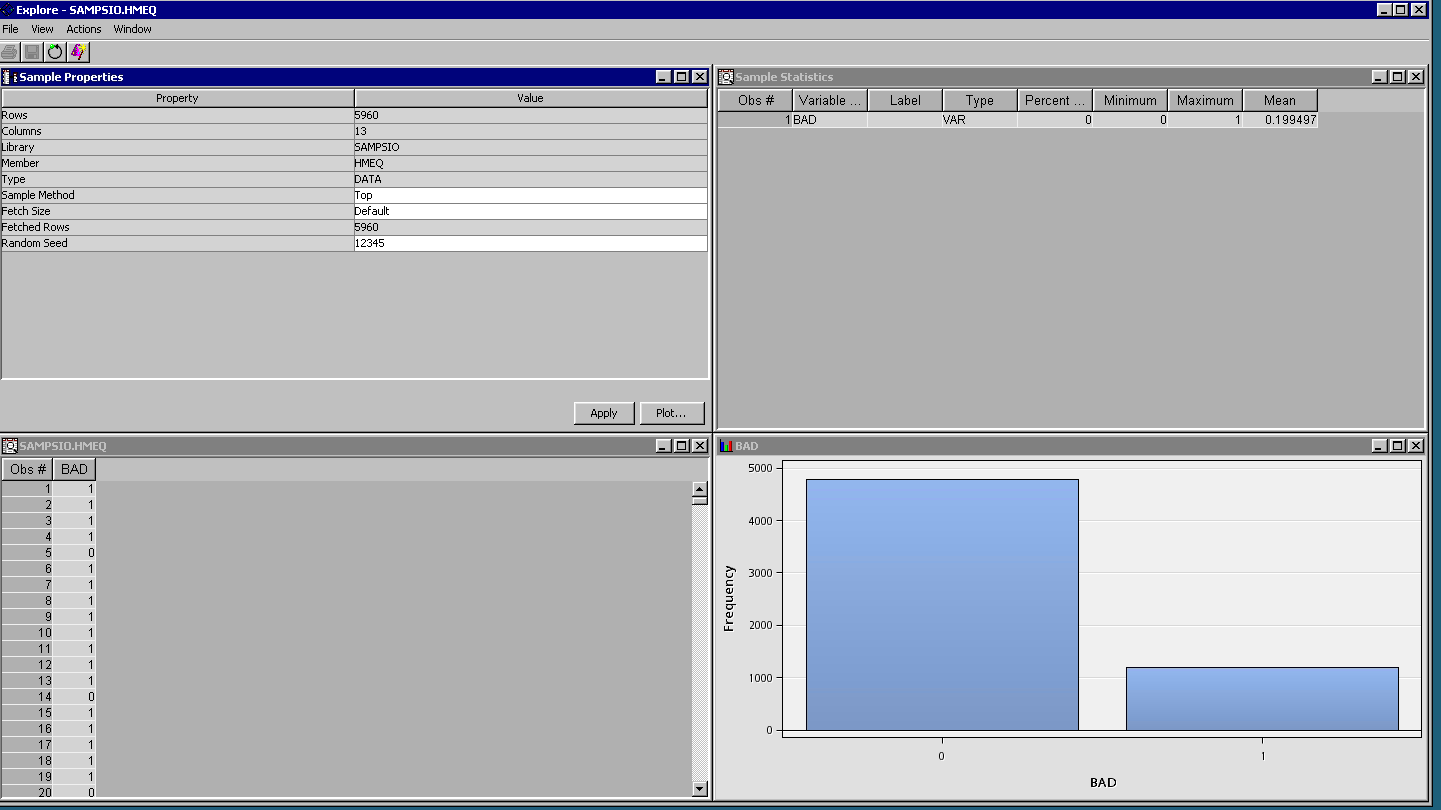
Repeat these steps to ensure that the remaining variables have the correct model role and measurement level information.

Note Assume DEROG and DELINQ are both interval measurements.

**Task 7: Viewing Distributions**

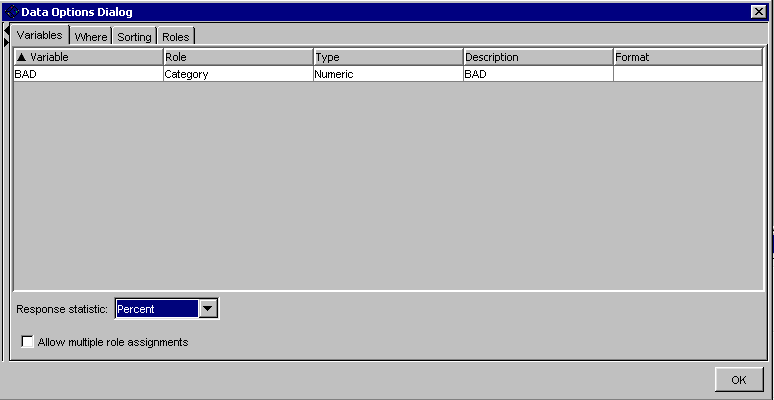
To view the distribution of BAD, for example:

Select in the **Name** column for the variable BAD then select **Explore button** to see the distribution of values for BAD in the metadata sample.

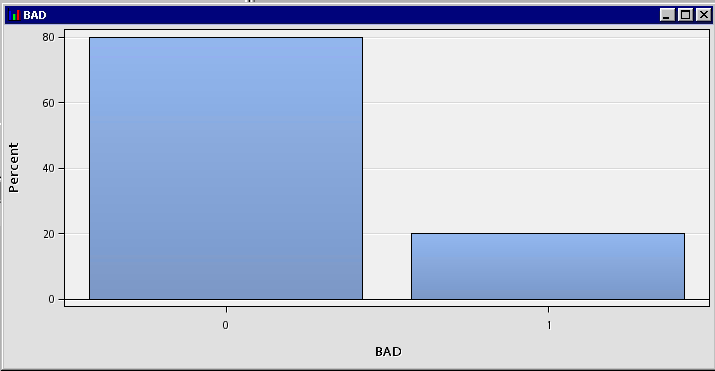


Four windows open in Explorer, Sample description window, Summary statistics, data, and a histogram window.

Note SAS Enterprise Miner displays the level and the proportion of observations that are represented by the bar. By default these are presented as frequency distributions. Hover over each bar to display the exact value. Right click on the histogram and **select data options** a dialogue box appears change the response statistic from frequency to percent using the drop down box. Then select **OK**.



For this example, approximately 20% of the observations were loans for which the client defaulted (BAD=1).

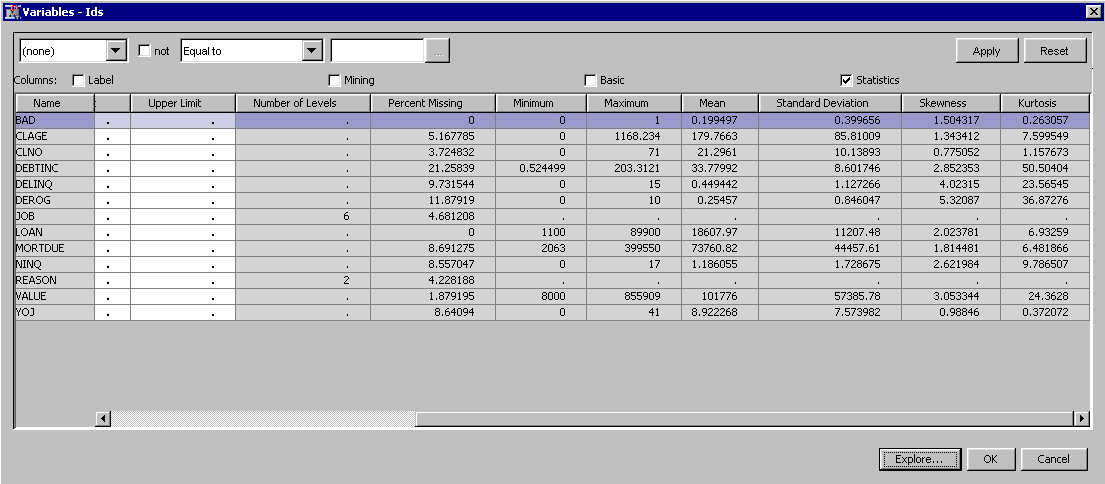


Recall that the plots and statistics in the Input Data Source node are based on the metadata sample, so the exact values in your window may differ slightly from those displayed here. These differences will not appear later in the modelling results since the modelling nodes use the entire training data set and not just a sample.

Select Close to return to the main dialog box when you have finished inspecting the plot.

**Investigating Descriptive Statistics**

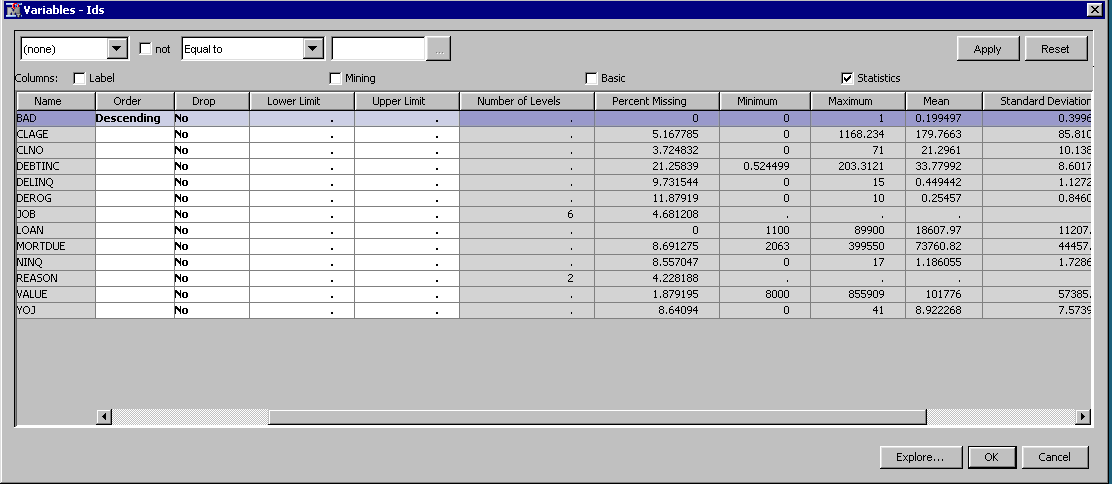
The metadata is used to compute descriptive statistics. Select the check box “Statistics”.



* ***Interval Variables*** display the minimum value, maximum value, mean, standard deviation,percentage of missing observations, skewness and kurtosis for interval variables. You might need to scroll to see all the columns.
* ***Class variables*** display the number of levels, percentage of missing values, and the sort order of the target variable.

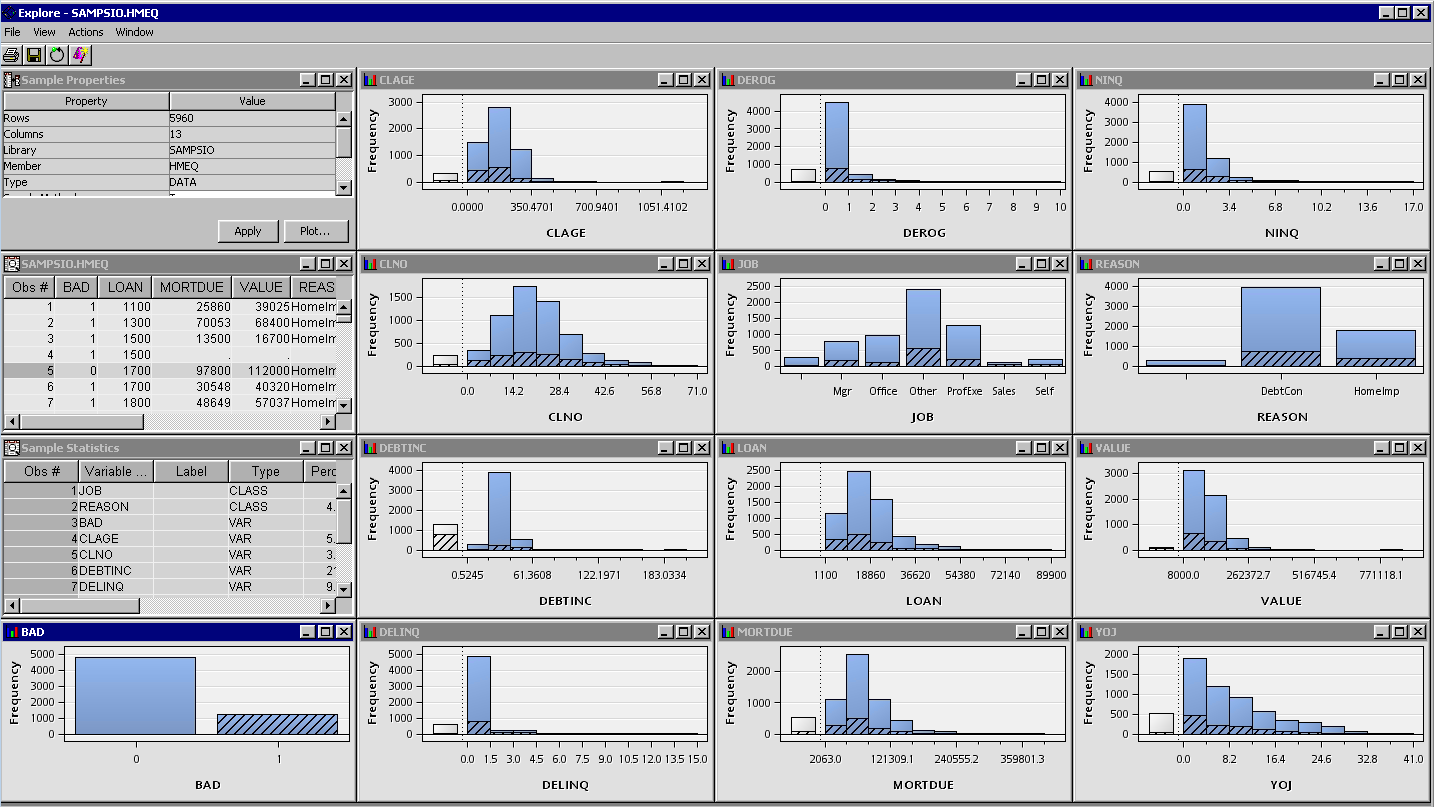
The sort order for BAD is descending. For a binary target such as BAD, the first sorted level is the target event. Since BAD has two levels (0 and 1) and BAD is sorted in descending order, BAD=1 is the target event.

Note In other datasets, you might need to change the sort order to get the target event that you want.



Close the Input Data Source node, and save changes when prompted.

It is possible to explore all the variables at once in order to get an overview of the data. If you select the BAD=1 bar, the brushing feature is enabled in the other histograms. This can provide you some insight on which variables might be important to predict the target class, DELINQ and DEBTINC for instance. White bars for interval variables represent missing values, missing labels for category data represent missing values.



The data exploration node is used to identify variables that have missing data, and which distributions are skewed or have outliers.

**Questions**

Which variable has the highest proportion of missing values and what percentage is it. Is it an important limitation and what would you do about it ?

Which variable has the most normal distribution ?

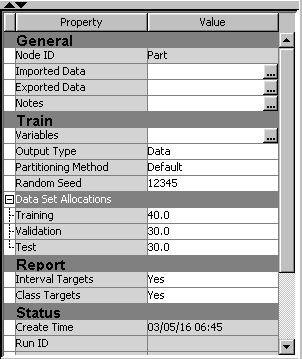
Which variable has the most skewed ?

Which variable has the fattest tails ?

**Task 8: Inspecting Default Settings in the Data Partition Node**

Select the Data Partition node and view the properties dialog box.

SAS Enterprise Miner takes a sample of the input data and divides it into training, validation and test data sets. By default if a target variable is identified then data is stratified on the target value, otherwise simple random sampling is used. These options can be modified in the properties dialogue box under **Partition Method.** Four choices, default, cluster, stratified and simple random.



**Simple Random -** Every observation in the data set has the same probability to be selected.

**Stratified** – ensures you get an even distribution of sub groups or specific values, every entry in a particular subgroup has an equal chance of being selected.

**Cluster -**  ensures you get an even distribution of observations from each cluster in property space, effectively stratifying on cluster. You must identify data partition role cluster for each variable you which to cluster on in property space.

But other sampling strategies are available, further customising can be achieved by altering the partition role in edit variables section of the partition node. To access these options right click on the data partition node.

The property dialog box allows you to chose a specific random seed to split the data, by default this is set to the same random seed 12345. If this is left blank then a different partitioning regime is generated each time the node is run.

The lower-left corner of the tab enables you to specify a random seed for initialising the sampling process. If you use the same data set with the same seed in different flows, you get the same partitioning or order of the training, test and validation sets. However a different seeds yields potentially different

results. Using the same number each time ensures you will get the same result.

The data allocation properties section allows you to customise the size of the split between the three partitions, training, validation and test set. By default this is set at 40%, 30% and 30%. It is not mandatory to have a validation or test set.

The percentages must add up to 100%.

**Task 9: Interpretation of Results**

Using the SAMPSIO.HMEQ data, answer the following questions.

*Understanding the Metadata Sample*

1. a) Explain why the variables REASON and JOB are defined as ‘character’ variables.

b) Why is REASON a ‘binary’ measurement and JOB a ‘nominal’ measurement?

2. All variables (except REASON and JOB) in the HMEQ dataset are numeric.

Explain why:

1. the variable BAD is a ‘binary’ measurement.
2. the other numeric variables are ‘interval’ measurements.

*Identifying Target Variables*

1. Explain why the variable BAD is the target (or response) variable in the dataset.

*Modifying Variable Information*

4. When might it be useful to change the measurement level of the variable DEROG to an ordinal measurement?

*Viewing the Distributions and Investigating descriptive statistics*

1. Explain why the minimum, maximum, mean and standard deviation are not shown for class variables.
2. Explain why the sort order for the variable BAD is descending.
3. View the distribution for each of the 13 variables and, for each one, consider the shape of the distribution and the descriptive statistics. Note your comments in the table on the following page:

8. From your results in Qu7 what generalisations can you make about the HMEQ dataset?

|  |  |  |
| --- | --- | --- |
| **Variable** | **Comments on the distribution** | **Comments on descriptive statistics** |
| BAD |  |  |
| LOAN |  |  |
| MORTDUE |  |  |
| VALUE |  |  |
| REASON |  |  |
| JOB |  |  |
| YOJ |  |  |
| DEROG |  |  |
| DELINQ |  |  |
| CLAGE |  |  |
| NINQ |  |  |
| CLNO |  |  |
| DEBTINC |  |  |

*Inspecting Default Settings in the Data Partition Node*

1. a) What type of sampling method is used?
2. Do you think the type of sampling method stated in 9a) is sensible for the SAMPSIO.HMEQ data?

c) What allocation of data is in the training, validation and test data sets?