Practical #1: Getting started with SAS Enterprise Miner

Objectives:

1. Start SAS Enterprise Miner (EM)
2. Understand the organisation of the analysis elements of SAS EM
3. Know how to create a SAS EM project, a diagram and define a data source
4. Know how to build a process flow for data mining in SAS EM and use the Reporter node to generate a report of your process flow

Submission:

- *What to submit*: a PDF report generated by the Reporter node/tool of SAS EM, containing the information (diagram, results etc.) about the exercise done by you by following the instructions given below.

- *Deadline of the submission*: 11:59PM (Adelaide Time), **Tuesday** of Week 3.

- *Submission link*: "**Submission Link of Prac #1"** in **Week 2 section** on Learnonline course site.

- *Marks*: Prac#1 (part of the ongoing assessment of the course) is worth 2% of the total marks of the course.

Instructions:

Before start, review Week 1 slides no. 30-33 to know the basic information about SAS EM.

Task 1: Register and access SAS EM server via SAS OnDemand

SAS EM is a data analytics software package by SAS Institute. SAS EM operates with a client-server architecture. Users need to run a SAS EM client program to access a SAS EM server where the data analytics jobs are done.

For this course, we use SAS EM by connecting to a SAS server provided by the SAS OnDemand for Academics program, and you first need to enroll in the SAS OnDemand course by following these steps:

1. Go to https://welcome.oda.sas.com/, create and activate your SAS profile by clicking on the SAS Profile link near the bottom of the page or the "Don't have a SAS Profile?" link on the Sign In popup window and following the instructions on screen. **Use your UniSA email address when creating your profile.**
2. Sign in at https://welcome.oda.sas.com/ and register for SAS OnDemand for Academics.
3. Put mouse over the icons on the right-hand side of the screen, find and click on the **Enrollments** icon. Click on the + sign when the Enrolments section appears.
4. Enter the course code: b911c7be-5272-4d87-80a6-5b5f3ad388a9
5. Confirm that this is the correct course (course name: **UMA-UniSA-SP2-2023**), and then enroll in the course.

Task 2: Start SAS EM

1. After you have enrolled in the course, more icons will appear on the right-hand side of the screen. Put mouse over the icons, find and click on the Enterprise Miner icon.
2. The SAS Enterprise Miner section will show to the right, with a link for downloading the SAS EM client installation program and the SAS Environments URL needed during the installation.
3. Download and run the SAS EM installation program to install SAS EM client on your PC. During the installation, you will need to provide the SAS Environments URL (copy and paste the above mentioned SAS Environment URL when prompted).

Note that SAS only provides SAS EM client for Windows. If you use a Mac, please see the following SAS Communities post about How to use SAS on a Mac.

SAS support may be reached to get help with accessing SAS EM:
   a. FAQ of SAS OnDemand for Academics: https://support.sas.com/ondemand/caq_new.html
   b. https://www.sas.com/en_us/contact.html (live chat may be available)

Once SAS EM is started, you should see the following SAS EM Welcome window:
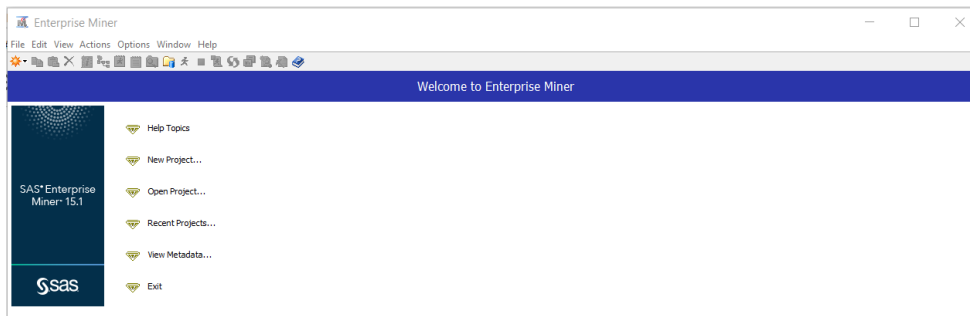
Task 3: Create a new Project

Analyses in SAS EM start by defining a **project**. After you have launched your SAS EM client session, you need to create a new project (or open an existing project) to proceed.

Follow the instruction below to create a new SAS EM project.

1. On Welcome page shown, click **New Project**, or from the main menu at the top of the page, select **File → New → Project**. The Create New Project wizard opens at Step 1.
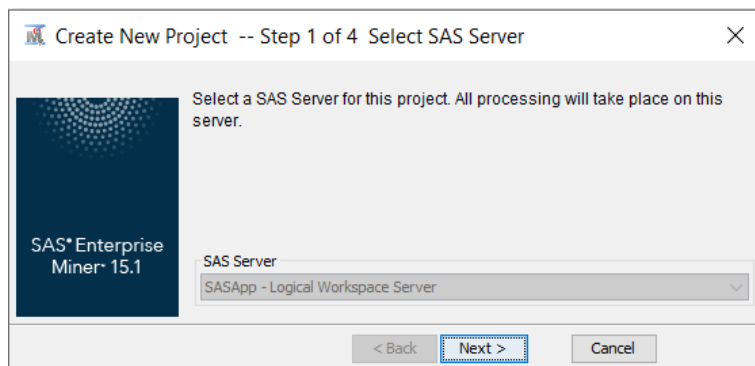
In this configuration of SAS EM, the only server available for processing is the host server listed.
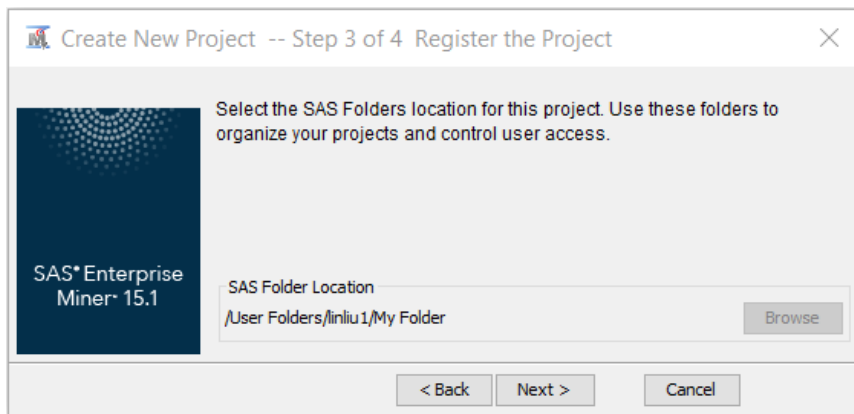
Click **Next**.

2. Step 2 of the Create New Project wizard is used to specify the name of the project you are creating and the location of the project. Type a project name, for example, Prac#1.
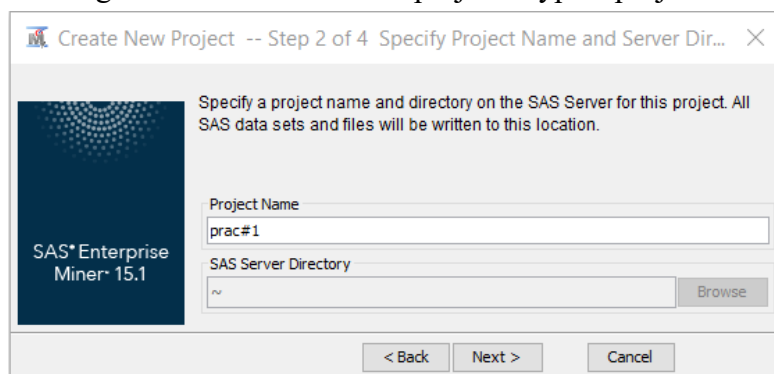
The path specified by the SAS Server Directory field is the physical location where the project folder will be created. Note that on SAS OnDemand server, normal user would not be able to change the directory.

Click **Next**.

3. Step 3 of the wizard is used to specify the SAS folder location to store the project information. On SAS OnDemand server, the folder location has been given.
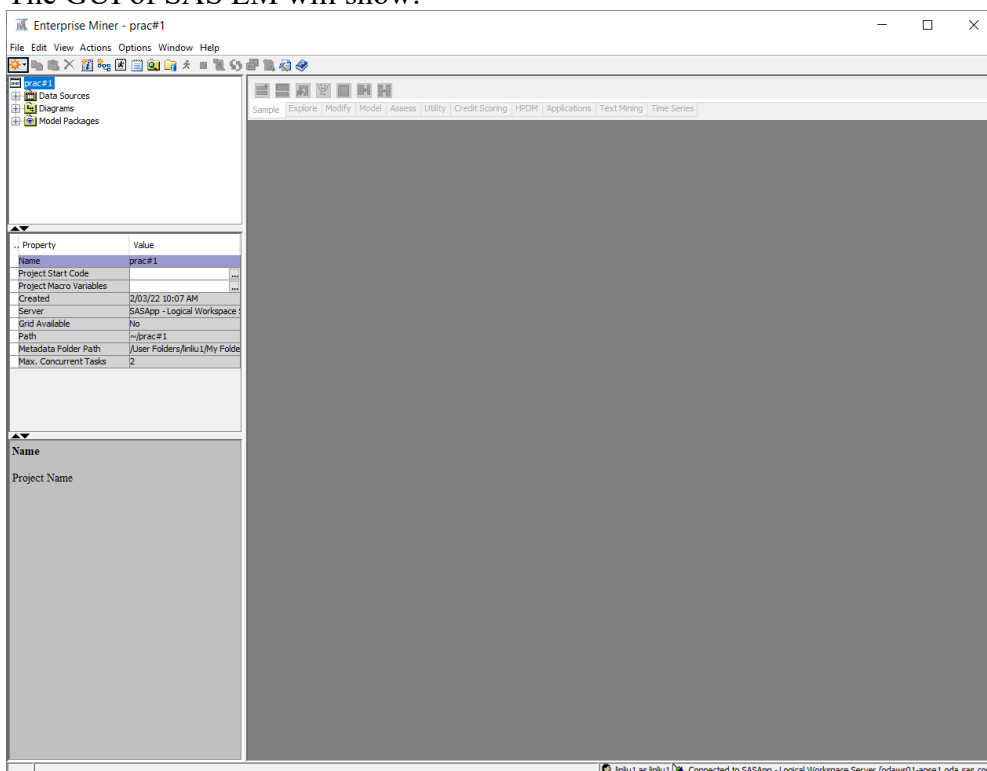

Click **Next**.

4. Review the project information that you have just specified in the **New Project Information** table. Use the **Back** button if you need to make changes. When you are ready to create your SAS EM project, click **Finish**.
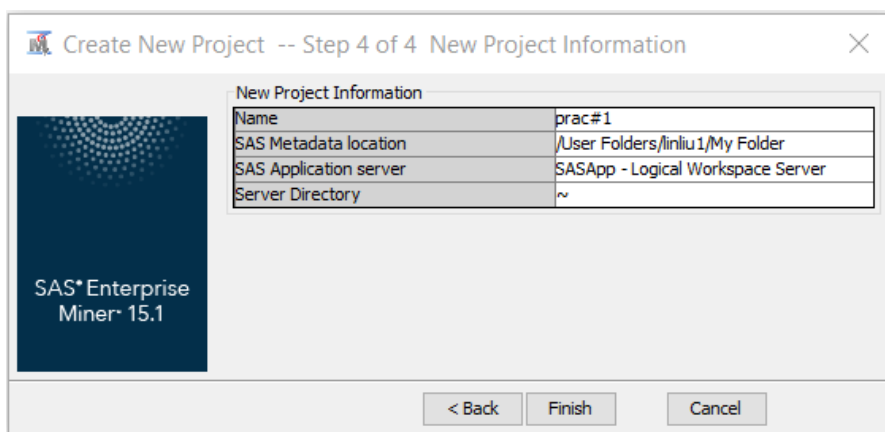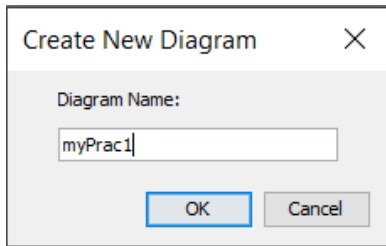

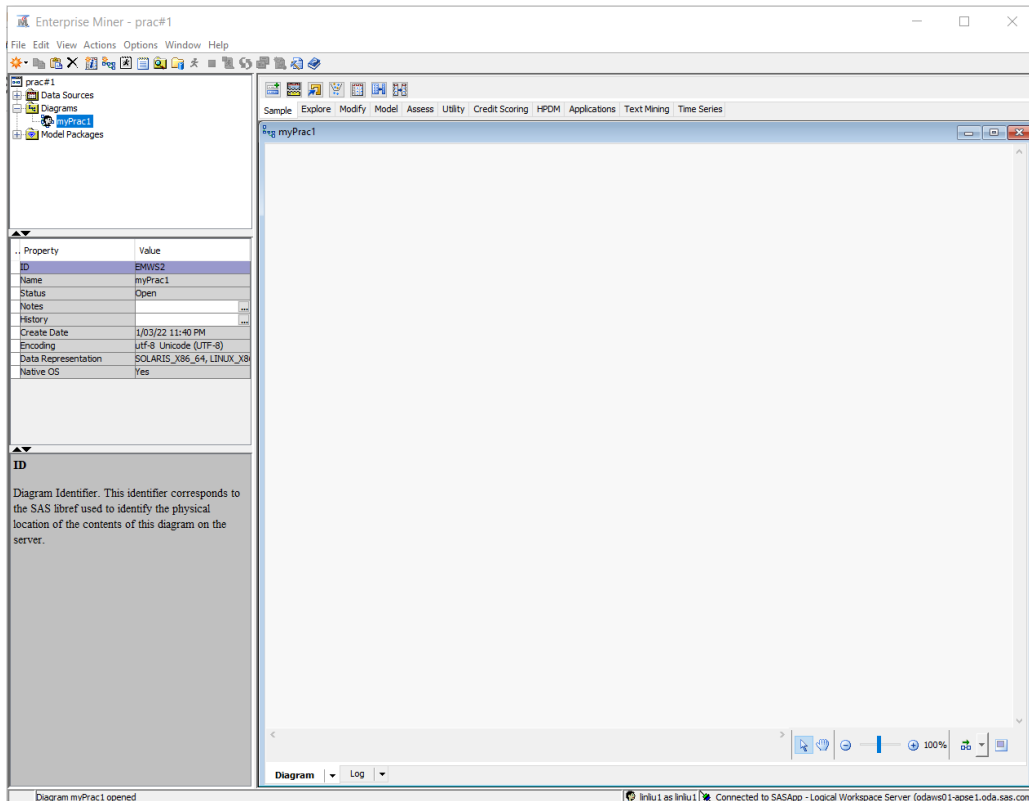The GUI of SAS EM will show:

Task 4: Create a new diagram

To create a new diagram, select **File → New → Diagram** from the main menu. Enter a name, e.g. myPrac1, for your diagram in the Create New Diagram window and click **OK**.

The new diagram opens in the Diagram Workspace.

The diagram workspace currently is empty. To conduct analysis, you need to create a process flow. This normally starts with defining a data source for the analysis.

Task 4: Define a data source

A data source links SAS EM to an existing table (data set). 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 EM.

On SAS OnDemand server, some libraries have been created and they contain information of existing data sets. So we skip the step of creating SAS library, and use the existing libraries.

### Part 1: Specifying Data Source

Follow these steps to specify a data source.

1. Select **File → New → Data Source** from the main menu. The Data Source Wizard window appears.
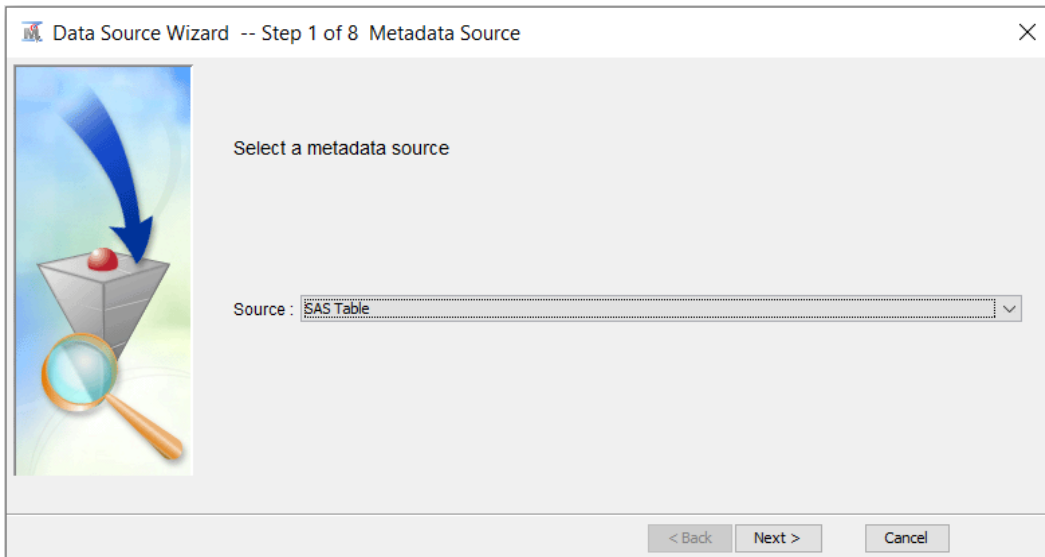
The Data Source Wizard guides you through creating a SAS EM data source. Step 1 of the Wizard tells SAS EM where to look for initial metadata values. The default and typical choice is **SAS table**. For this practical, however, we choose to use **Metadata Repository**.

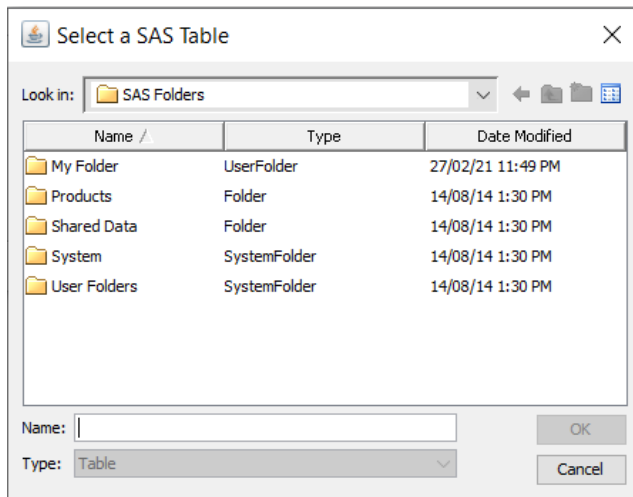From the dropdown menu, select **Metadata repository**, and click **Next**.
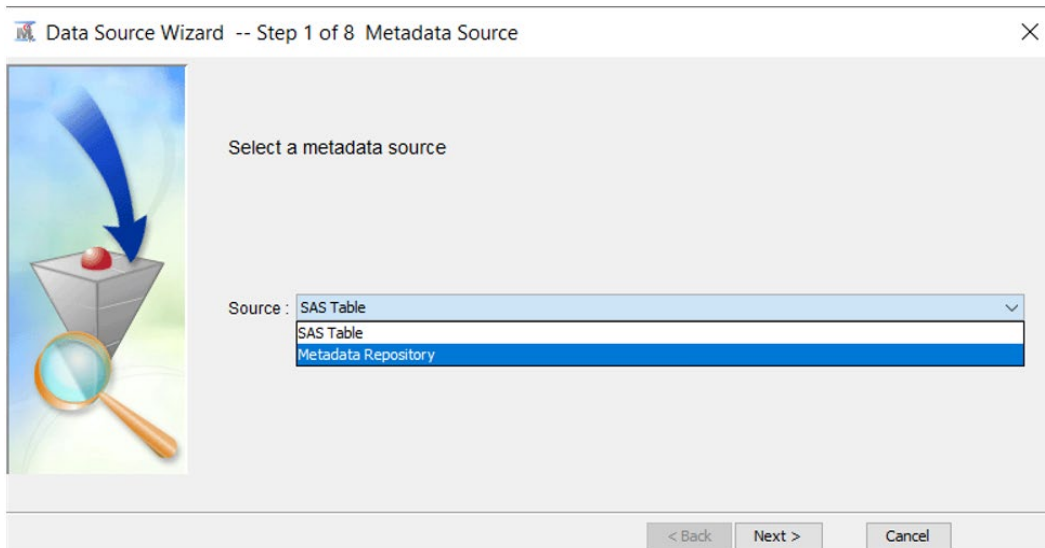
**2.** The Data Source Wizard continues to Step 2 - Select a SAS Table.

In this step, select the SAS table that you want to make available to SAS EM.

For this practical, select **Browse** to choose a SAS table from the libraries that are visible to the SAS Server. The Select a SAS Table window appears.
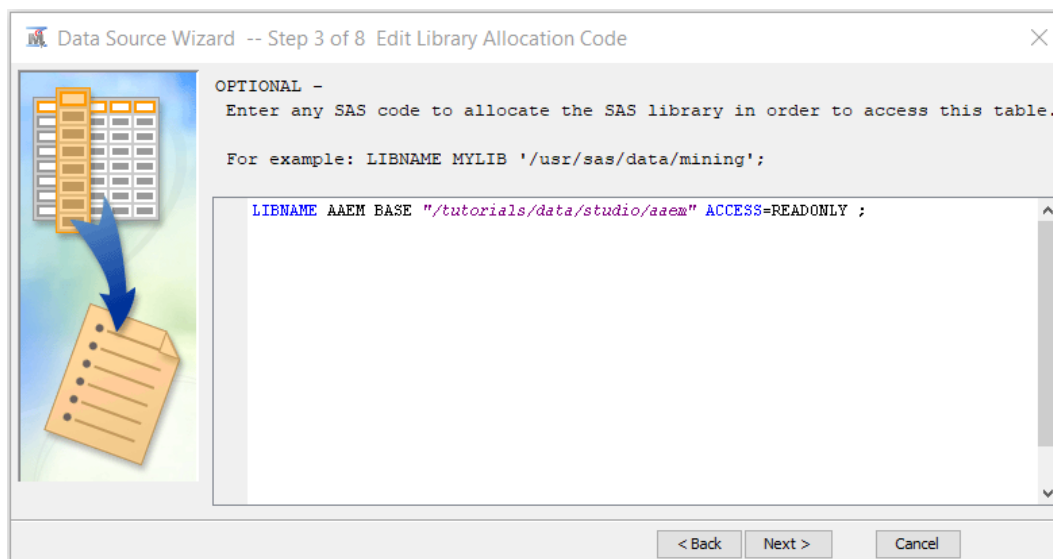

Then select **Shared Data → Libraries →AAEM**, and choose the table named **PVA97NK**, click **OK**.

The Select a SAS Table window closes and the selected table appears in the **Table** field of the Wizard window.
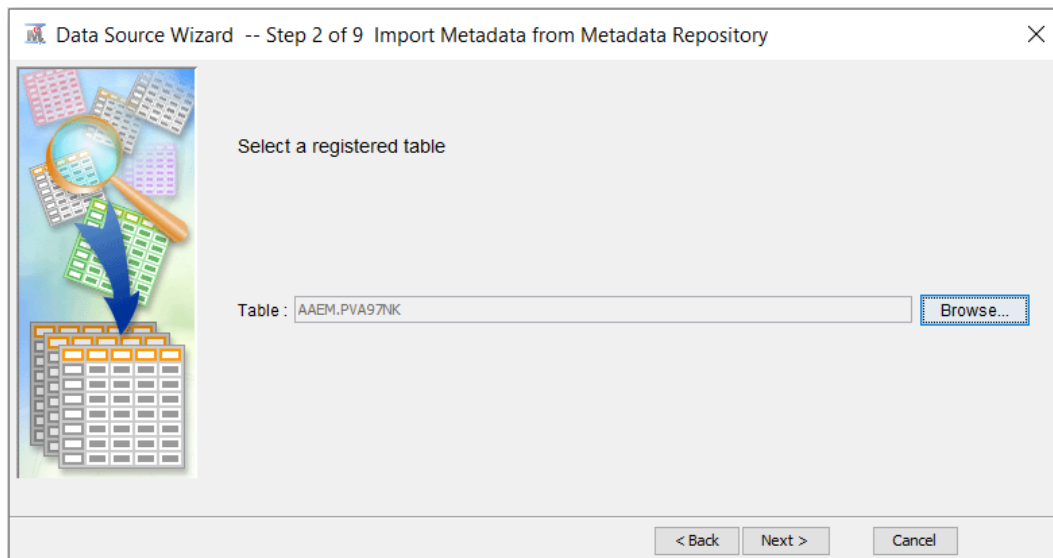
3. Click **Next**. The Data Source Wizard proceeds to Step 3 - Edit Library Allocation Code.

4. Click **Next**. The Data Source Wizard proceeds to Step 4 - Table Information.

This step of the Data Source Wizard provides basic information about the selected table. The SAS table **PVA97NK** contains 9,686 cases/records and 28 variables.

Part 2: Defining Column Metadata

With a data set specified, your next task is to set the column metadata, including modeling role (whether a variable is used in the analysis or not, and their roles in the analysis), and proper measurement level of each variable in the source data set.

5. Select **Next**. The Data Source Wizard proceeds to the Metadata Advisor Options step.

This step of the Data Source Wizard starts the metadata definition process. SAS EM 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.

6. Select **Next** to use the **Basic** setting.

The Data Source Wizard proceeds to the Column Metadata step. Select the box next to **Label**. Note that you may need to enlarge the Wizard window to see all the information.

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

A comparison of the currently assigned metadata to that in the **PVA97NK** metadata table on next page shows several discrepancies. While the assigned modeling roles are mostly correct, the assigned measurement levels for several variables are in error.
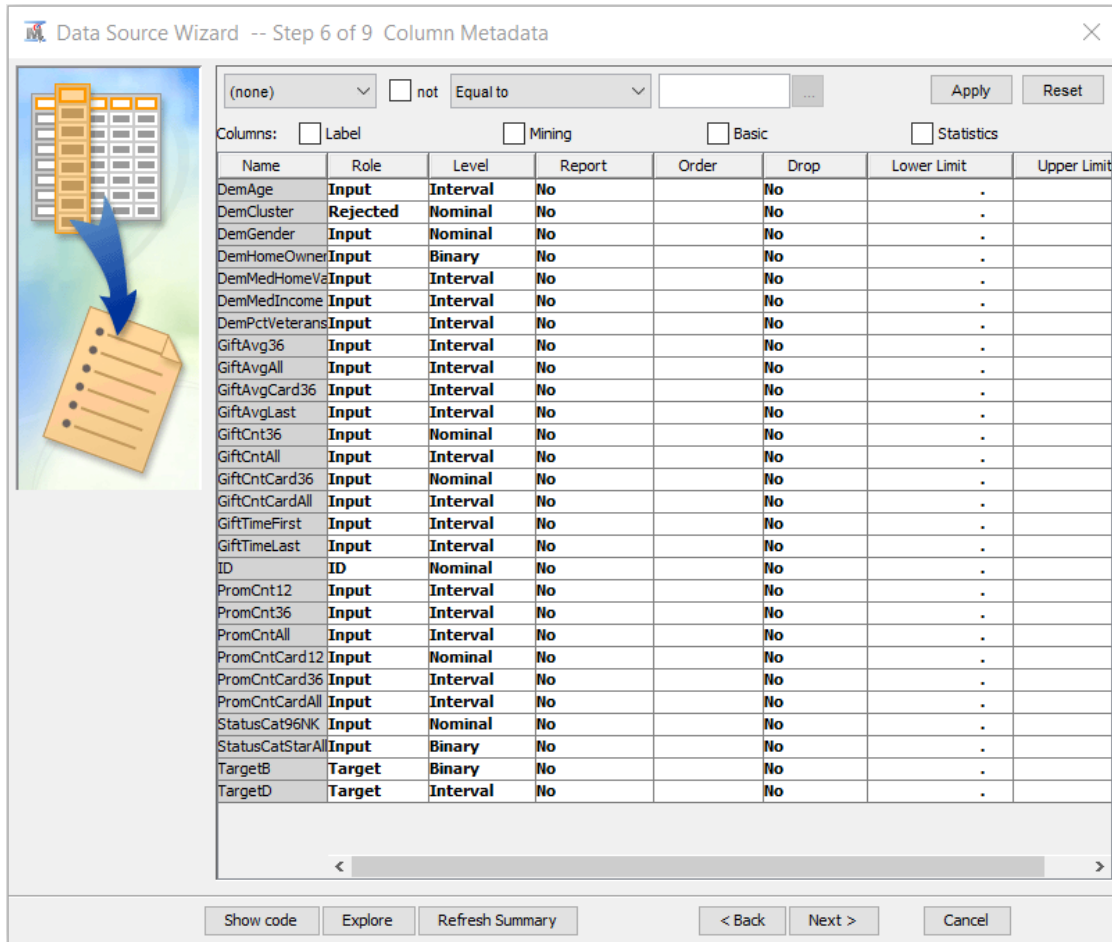
It is possible to improve the default metadata assignments by using the **Advanced** option in the Metadata Advisor.

Select **Back** in the Data Source Wizard. This returns you to the step of Metadata Advisor Options.

PVA97NK Metadata Table

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

7. Select the **Advanced** option and select **Next** to use the Advanced setting. The Data Source Wizard again proceeds to the Column Metadata step.
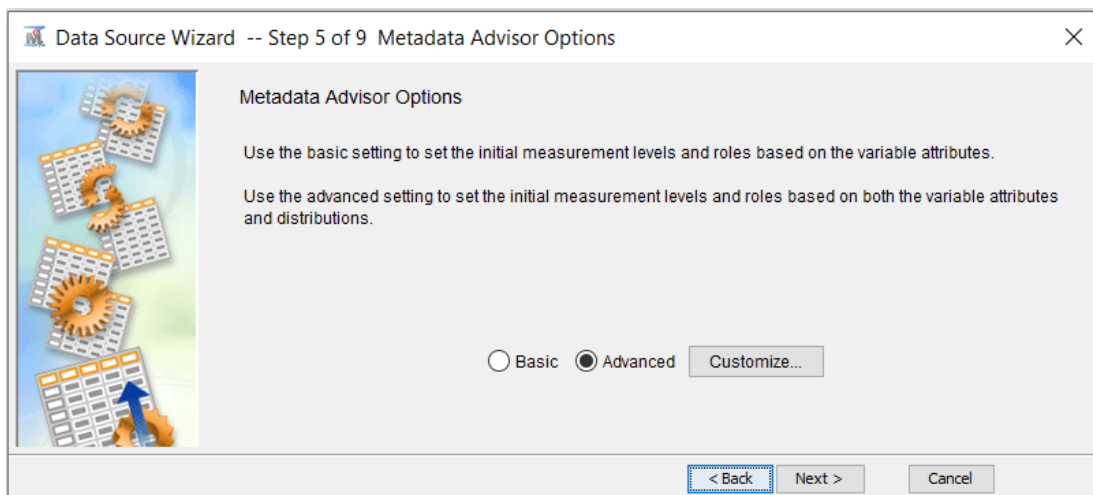
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
DemAge	Input	Interval	No		No	.	
DemCluster	Rejected	Nominal	No		No	.	
DemGender	Input	Nominal	No		No	.	
DemHomeOwner	Input	Binary	No		No	.	
DemMedHomeVa	Input	Interval	No		No	.	
DemMedIncome	Input	Interval	No		No	.	
DemPctVeterans	Input	Interval	No		No	.	
GiftAvg36	Input	Interval	No		No	.	
GiftAvgAll	Input	Interval	No		No	.	
GiftAvgCard36	Input	Interval	No		No	.	
GiftAvgLast	Input	Interval	No		No	.	
GiftCnt36	Input	Nominal	No		No	.	
GiftCntAll	Input	Interval	No		No	.	
GiftCntCard36	Input	Nominal	No		No	.	
GiftCntCardAll	Input	Interval	No		No	.	
GiftTimeFirst	Input	Interval	No		No	.	
GiftTimeLast	Input	Interval	No		No	.	
ID	ID	Nominal	No		No	.	
PromCnt12	Input	Interval	No		No	.	
PromCnt36	Input	Interval	No		No	.	
PromCntAll	Input	Interval	No		No	.	
PromCntCard12	Input	Nominal	No		No	.	
PromCntCard36	Input	Interval	No		No	.	
PromCntCardAll	Input	Interval	No		No	.	
StatusCat96NK	Input	Nominal	No		No	.	
StatusCatStarAll	Input	Binary	No		No	.	
TargetB	Target	Binary	No		No	.	
TargetD	Target	Interval	No		No	.	

While many of the default metadata settings are correct, there are several items that need to be changed. For example, the **DemCluster** variable is rejected (for having too many distinct values) and not to be included in the analysis, and several numeric inputs have their measurement level set to Nominal instead of Interval (for having too few distinct values).

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

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

9. Select **Customize**. The Advanced Advisor Options dialog box appears. (Need to select **Advanced** to enable **Customize**.)

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

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

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

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

10. Type **2** as the **Class Levels Count Threshold** value so that only binary numeric variables are treated as categorical variables, and type **100** as the **Reject Levels Count Threshold** value so that only character variables with more than 100 distinct values are rejected.

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

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

12. Select **Next** in the Wizard window to proceed to the next step, Column Metadata.

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

A comparison of the Column Metadata table in SAS EM to the given PVA97NK Metadata Table shows that most of the metadata is correctly defined. SAS EM correctly inferred the model roles for the non-input variables by their names. The measurement levels are correctly defined by using the Advanced Metadata Advisor.

13. Assuming that the analysis of the **PVA97NK** data is to focus on the **TARGET_B** variable (e.g. to create a prediction model for this variable), you can reject the other target variable **TARGET_D** variable by selecting **Role → Rejected → TARGET_D**.

|		
PromCntCardAll	Input	Interval
StatusCat96NK	Input	Nominal
StatusCatStarAll	Input	Binary
TargetB	Target	Binary
TargetD	Rejected	Interval

In summary, the Defining Column Metadata part is usually the most time-consuming part of the Data Source Wizard steps.

**Part 3: Finalizing the Data Source Specification**

14. Select **Next** to proceed to the Decision Configuration step.

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

When you define a predictive modeling data set, it is important to properly configure decision processing. In fact, obtaining meaningful models often requires using these options. The **PVA97NK** table was structured so that reasonable models are produced *without* specifying decision processing.
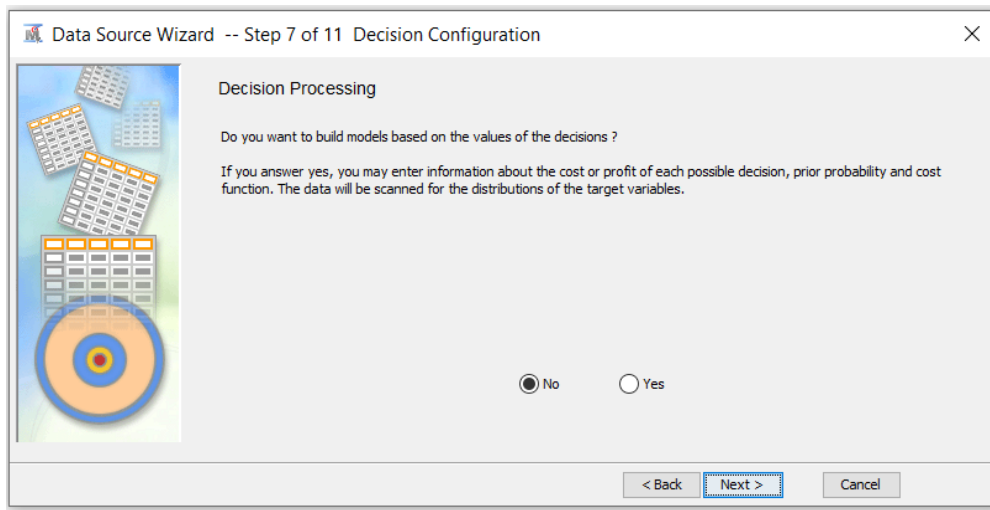
15. Select **Next** to proceed to Create Sample. No sample data set is created, so leave this step set to the default **No**.
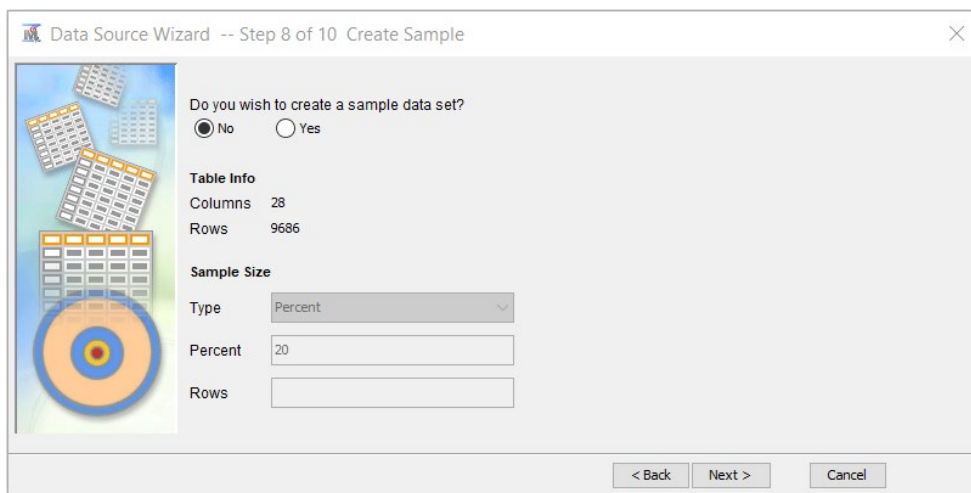

16. Select **Next** and you reach the second to the last step of the Data Source Wizard.
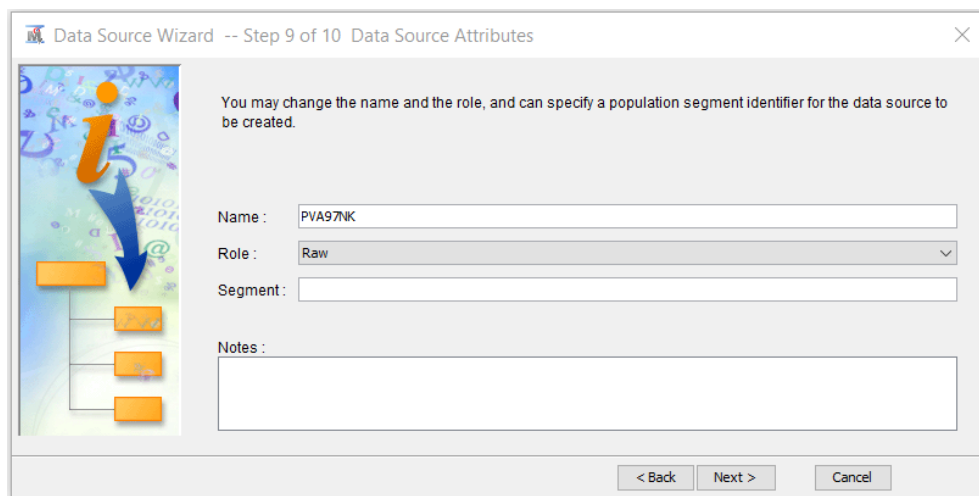

This step enables you to set a role for the data source and add descriptive comments about the data source definition. For the upcoming analysis, a table role of Raw is acceptable.

17. Select **Next**. The final step in the Data Source Wizard provides summary details about the data table that you created. Select **Finish**.
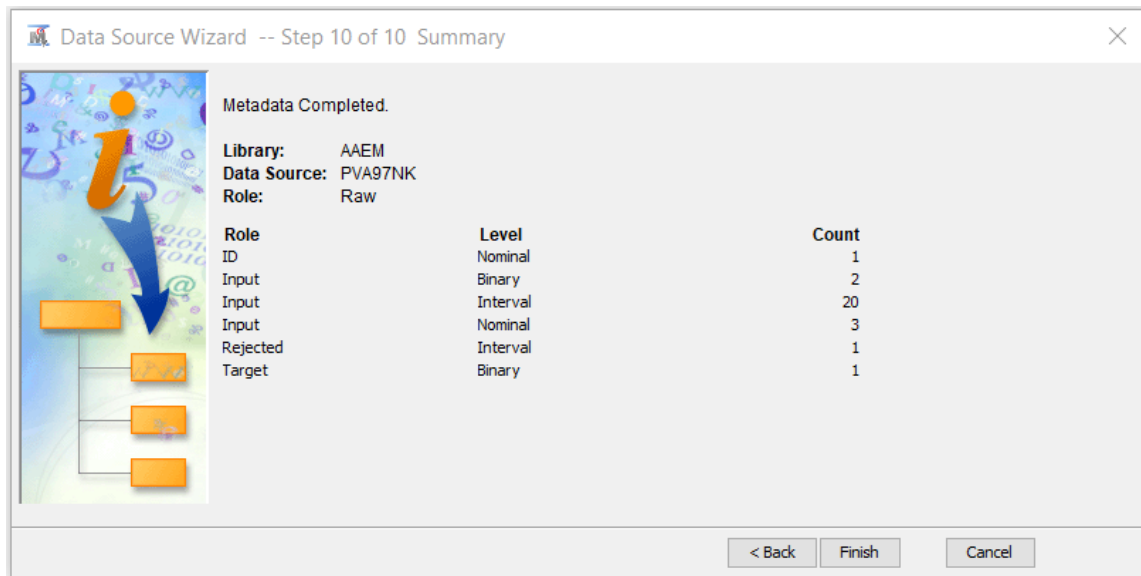
The **PVA97NK** data source is added to the Data Sources entry in the Project panel.
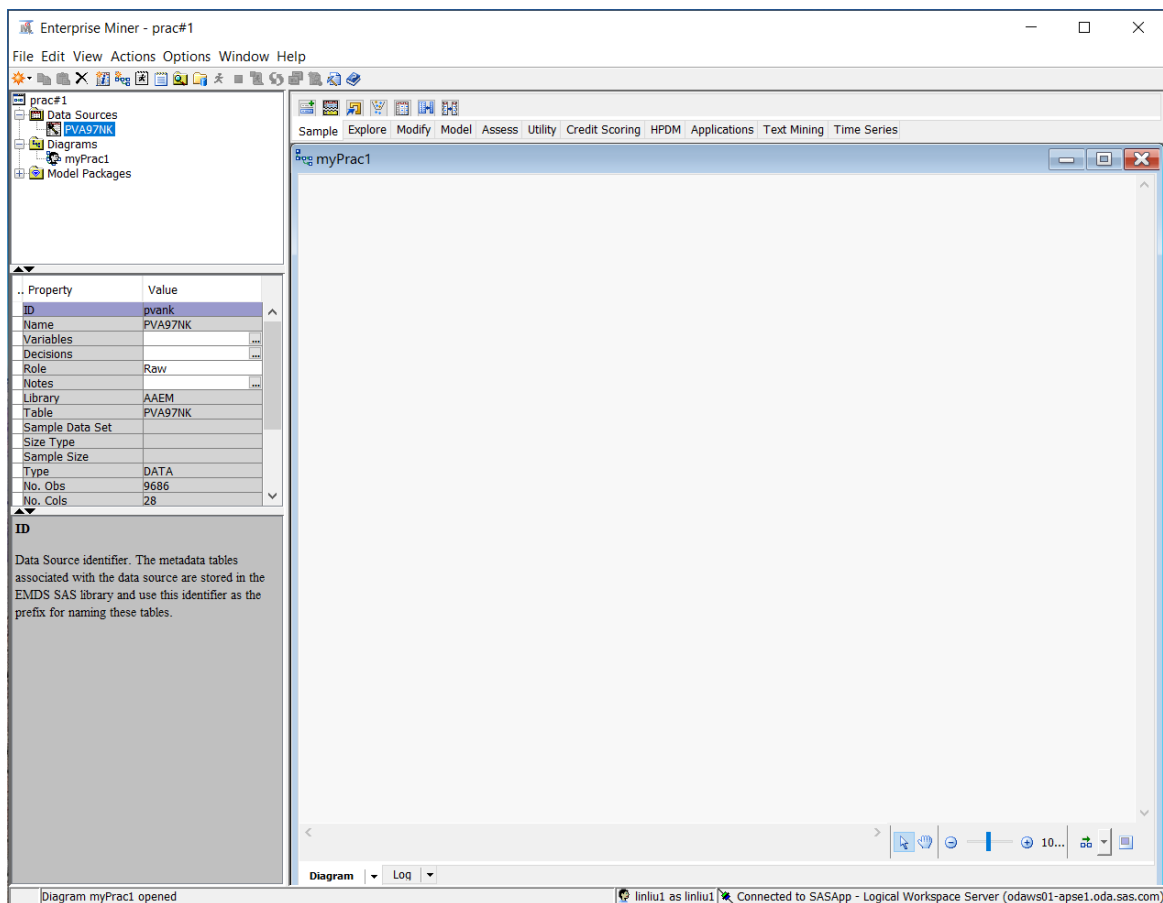
Task 5: Build a process flow diagram

Now we are ready to create a process flow in the diagram workspace to analyse the data source we have specified.

1. Create an **Input Data node** for your process flow.

   Drag the PVA97NK data source from the Data Sources folder in your Project Navigator/Panel onto your Diagram Workspace. This creates an Input Data node in the Diagram Workspace.
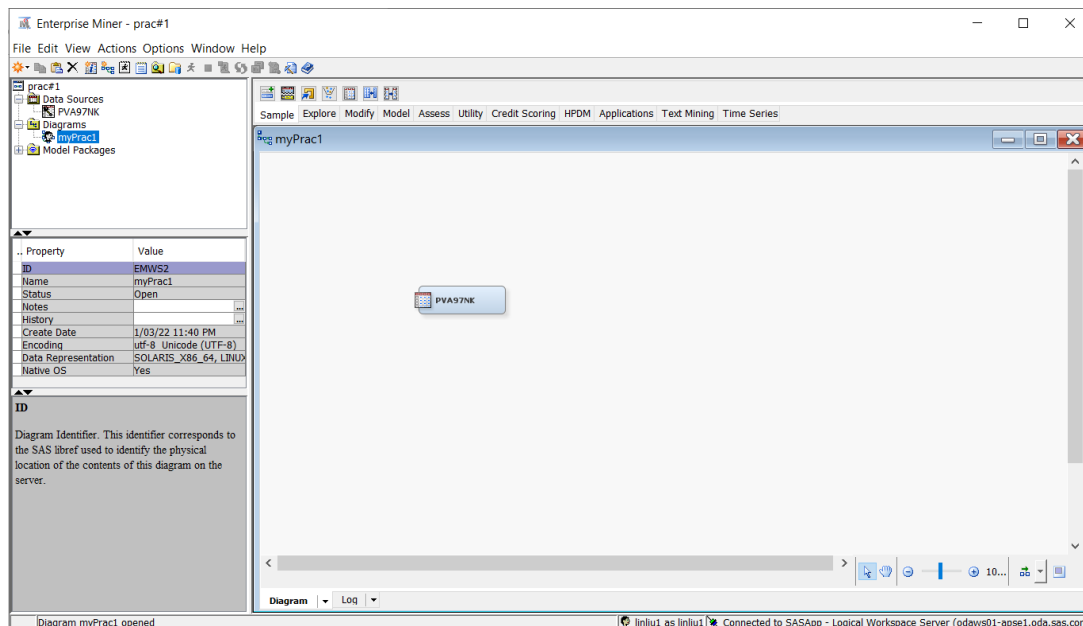
2. Explore the data set within the Input Data node

   Right click on the Input Data node (labelled with the data source name PVA97NK), select **Edit Variable**, a window with all the variables appears:
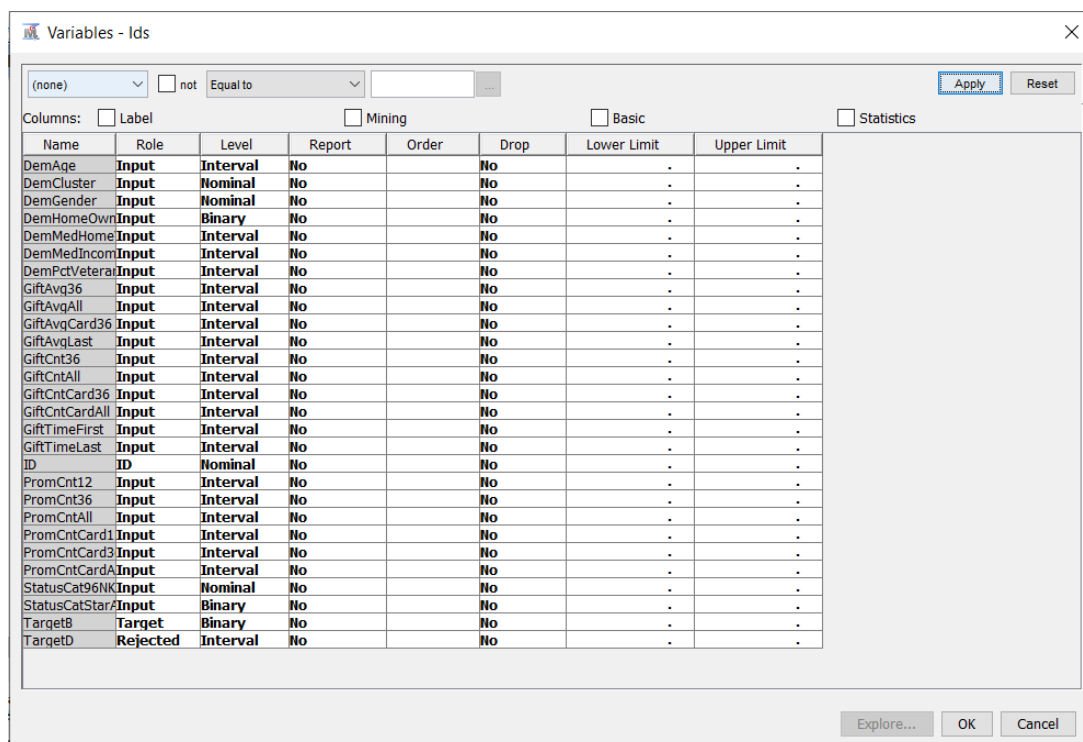
   Click/select the boxes next to **Label, Mining**, **Basic**, and **Statistic**, respectively, to see what extra information/columns are shown about the dataset.

Next Select all listed inputs by dragging the cursor across all of the input names or by holding down the **CTRL key** and typing **A.**

The **Explore** button now should change from grey to normal. Click on **Explore**.

Note: If the Explore button is still grey, try either of the following two methods:

Method 1: close and reopen the project.

Go to main menu, **File-> Close this Projec**t, then Open this project. When the project is opened, click on the "+" sign in the Project Panel (top left of screen), and double click on the diagram name. Now in the diagram workspace, right click on the Input Data node, choose Edit Variables, and in the Variables window shown, select the variables, the Explore button should be enabled.

Method 2: use the Edit Variable function of the data source

In the Project panel, right click on the PVA97NK data source, choose Edit Variables, and then in the Variables window shown to explore the variables.
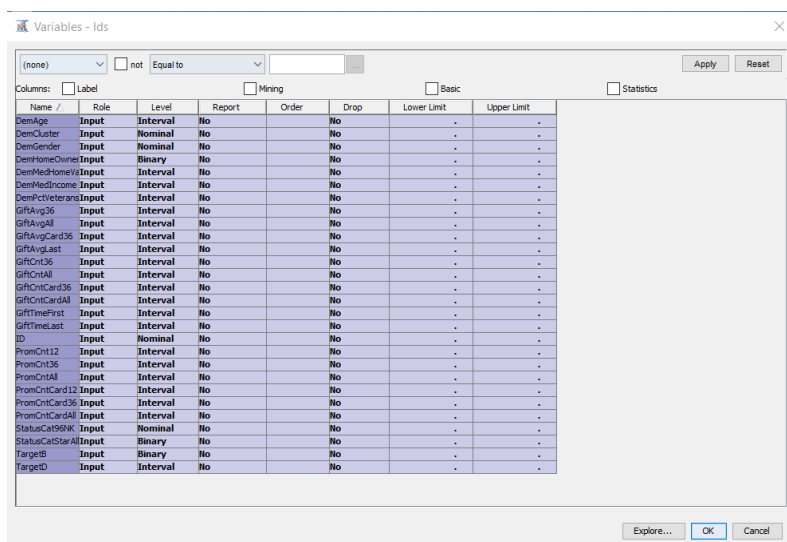

You may see a Note window popping up, click OK to proceed.

Then the Explore window shows, displaying the histograms for all the variables selected and other information of the data set. You can maximize a histogram or information window by double clicking on its title bar, then the histogram or information will fill the Explore window. To restore the size of a maximized window, double clicking the title bar, the window will return to its original size. Have a look at the information and histograms shown, and we will use this Explore function more later in this course.
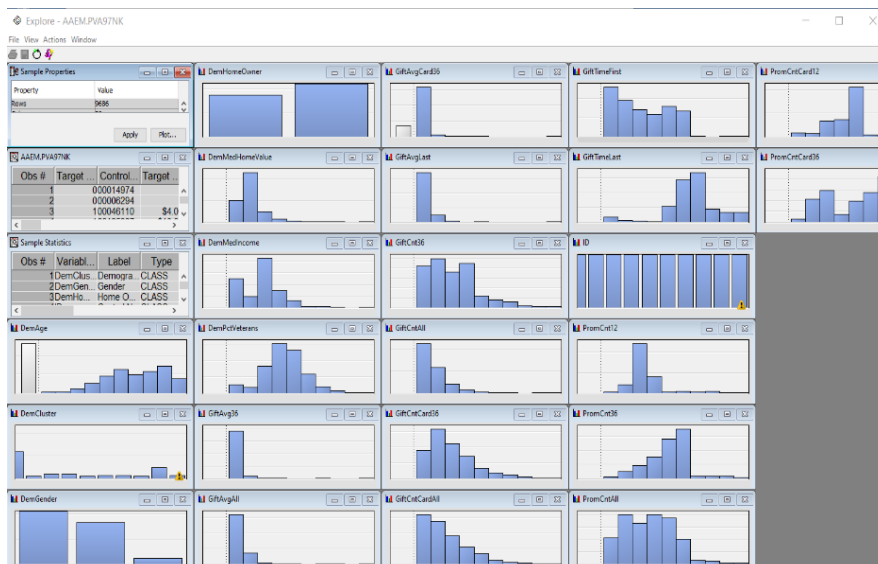
Task 6: Connect and run a reporter node to produce a report

Click on the Utility tab in the tools palette and then choose the **Reporter**:
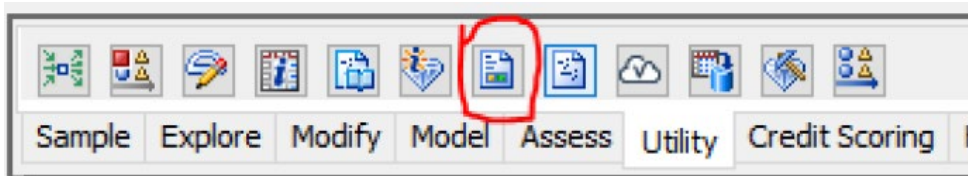

Drag a Reporter node to your diagram workspace, and connect it to the last node (which is the Input Data node labelled with PVA97NK) in your process flow diagram.

To connect the nodes, move the pointer to the right edge of the Input Data node icon. The pointer icon changes from an arrow to a pencil. Drag the pointer to the left edge of the Reporter node. Release the pointer, and the nodes are connected.
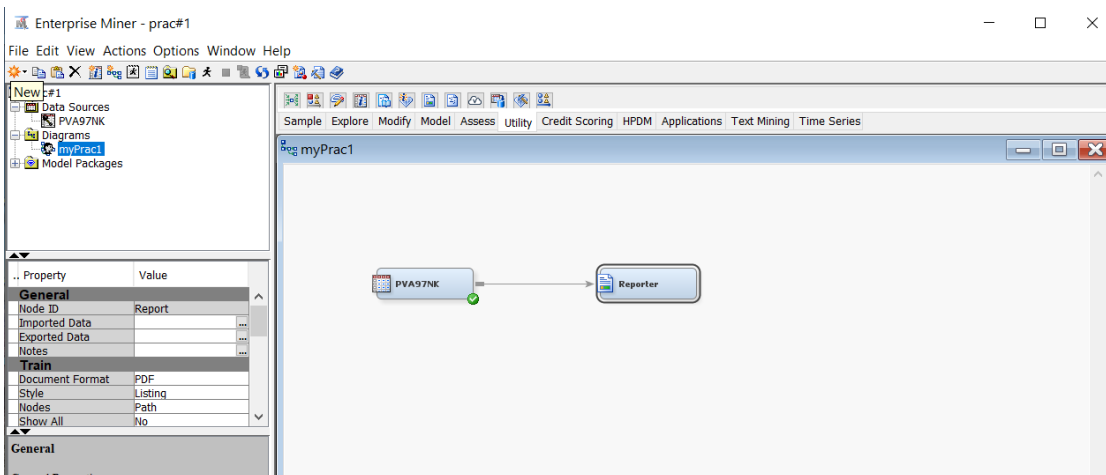
Now you have a (very simply) process flow diagram with only two nodes. To generate results from a process flow diagram in your Diagram Workspace, you must run the process flow path to execute the code that is associated with each node.

Right-click the node from which you want to run the flow (normally, a terminal node) and select **Run** from the pop-up menu.
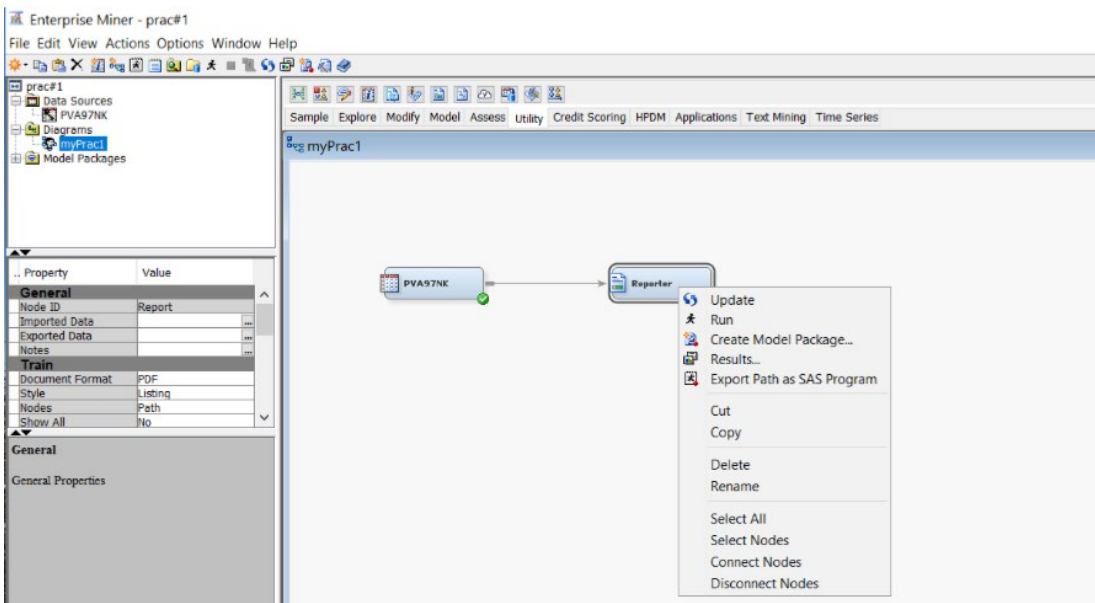
After the process flow has been run, the nodes in your process flow diagram will each have a green tick shown at its right bottom, and the Run Status window will show. Click **OK** if you don't need to see the results. In this example, Click **OK**.

Select the added Reporter node in your diagram (click on it) and explore its setting in the property panel. For our practicals, we will use the default settings as shown in following unless otherwise required:

| Train	
Document Format	PDF
Style	Listing
Nodes	Path
Show All	No

To view the generated report, click the ellipses button in the View Report field in property panel. (If the document format is set to PDF in step 3 above, the report will be opened in Adobe Reader.)

Report		
Save Report		...
View Report		...

Now save the PDF report, and submit it by following the instruction given on page 1 of the practical document.

Task 7: Get to know the GUI components of SAS Enterprise Miner

We have learned how to create and run a process flow. Let us now have a more detailed look at the GUI components of SAS EM and their functions.
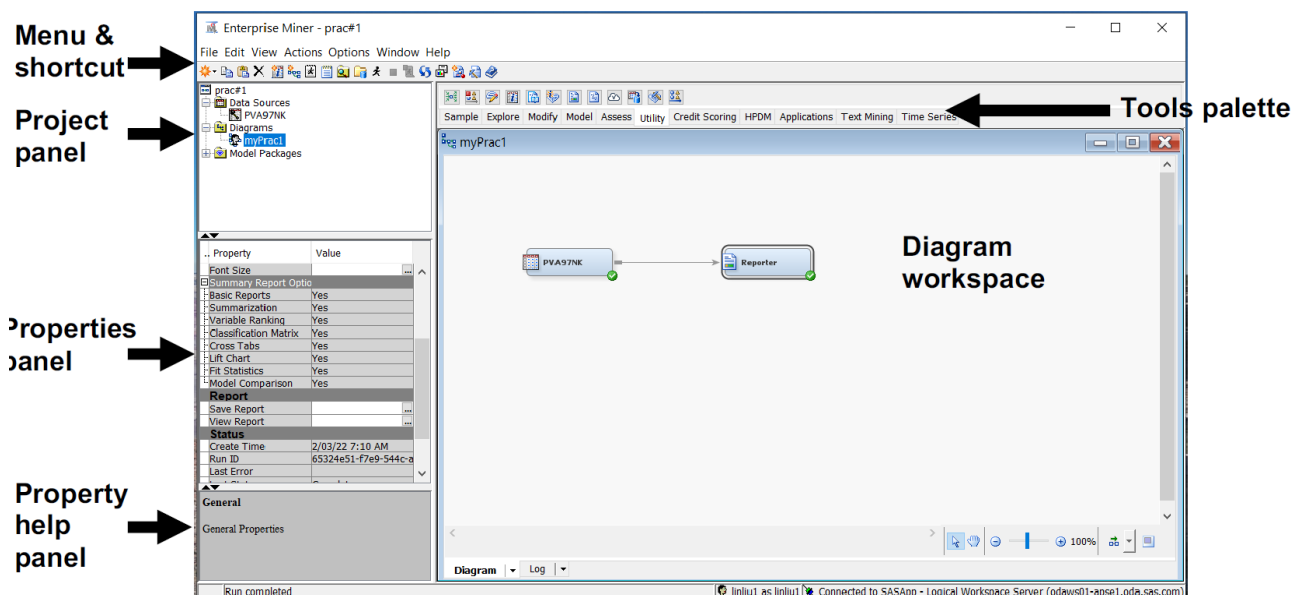
The SAS EM **menu** provides access to all functions provided by the software package. Please note that from the **Help** menu, you can access the full content of SAS EM reference booklet, which provides a comprehensive and detailed introduction to SAS EM. **You will need to refer to the Help contents throughout the semester** when you learn how to use SAS EM and completing practicals and assignments. The Reference Help is also available online at:

https://documentation.sas.com/doc/en/emref/15.2/bookinfo.htm. Bookmark this link for each reference.

The **shortcut** buttons are a graphic set of user interface tools that you use to perform common computer functions and frequently used SAS EM operations. The text name of any tool button is displayed when you position your mouse pointer over the button.

The SAS EM tools (a.k.a. nodes) available to your analysis are contained in the **tools palette**. The tools palette is arranged according to a process for data mining, SEMMA (Sample, Explore, Modify, Model, Assess). Click on a tab, e.g. Sample, you will have series of nodes shown above the tab.

The **Project panel** is used to manage and view data sources, diagrams, results, and project users.

The **Properties panel** is used to view and edit the settings of data sources, diagrams, nodes, results, and users.

The **Property help panel** displays a short description of the property that you select in the Properties panel. Extended help can be found in the Help main menu.

The **Diagram workspace** is where you build, edit, run, and save process flow diagrams. In this workspace, you graphically build, order, and sequence the nodes that you use to mine your data and generate reports.