INFS 5102 – Unsupervised Methods in Analytics

Practical #2: Data

Objectives:

1. Know how to preprocess data with SAS EM
2. Know how to export processed data from SAS EM to a local directory
3. Know how to use the Import node to load an external data set to SA EM

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 for this practical.
- *Deadline of the submission*: 11:59PM (Adelaide Time), Tuesday of Week 4.
- *Submission link*: "**Submission Link of Prac #2"** in **Week 3 section** on Learnonline course site.
- *Marks*: Prac#2 (part of the ongoing assessment of the course) is worth 2% of the total marks of the course.

Instructions:

Preprocess the PVA97NK data set using SAS EM

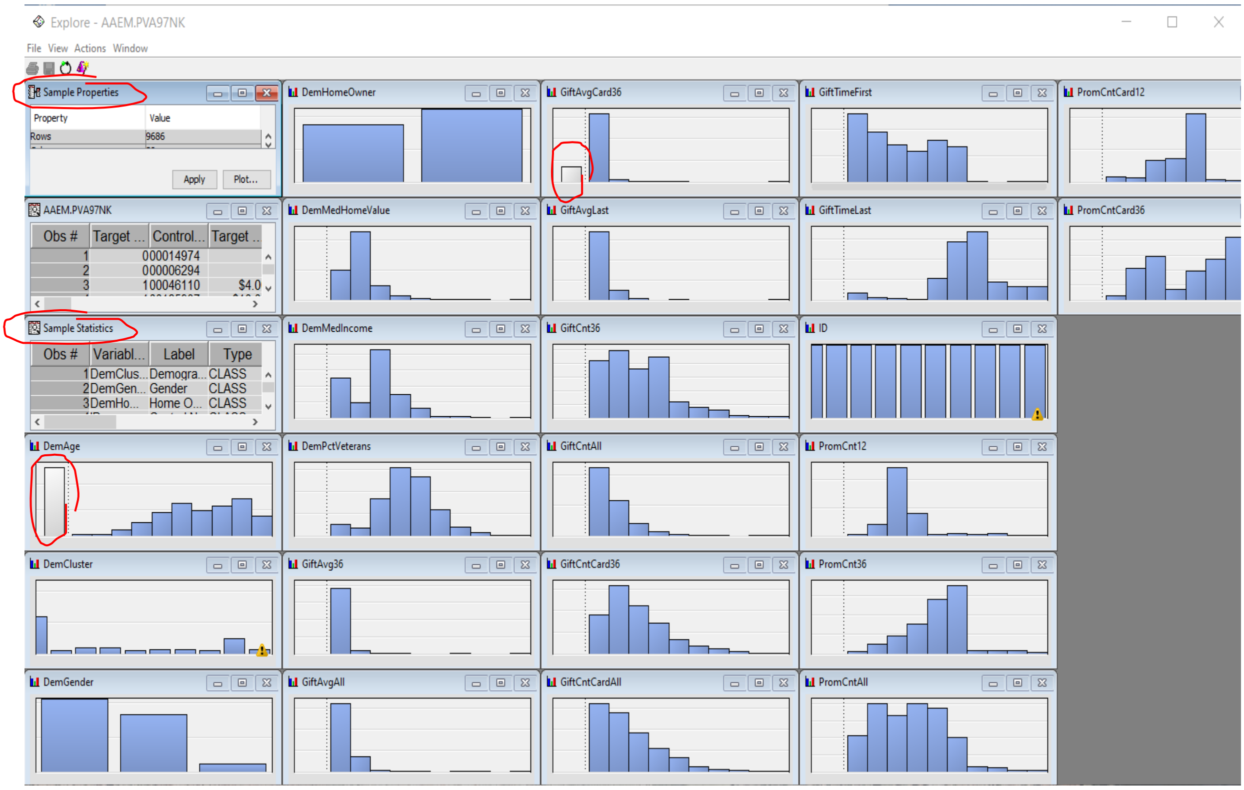
In Practical #1, you created a project, a diagram workspace within the project, and a data source for the project using the PVA97NK data set. For this practical, you will continue to work with this data set and try a few data preprocessing functions provided by SAS EM. You will also learn how to export SAS data and import your own data file to SAS EM.

Following the instruction given below to complete the practical.

1. Create a project named **Prac#2**, and in the project, create a diagram named **myPrac2** and a data source with the PVA97NK data set. Then drag the PVA97NK data source to the diagram workspace to create an Input Data node for a process flow. (Refer to the instruction in Tasks 1 to 4 of Practical #1 on how to create a project, diagram and data source.)

Note: you might wish to reuse the project and data source created for Practical #1. However, it appears that SAS EM (or the SAS OnDemand setting) does not allow us to save a project with a different name, which means you will lose the project created for Practical#1 if you simply update and save it for Practical#2. It is recommended that you create a new project for Practical #2. In this way, you can also get more familiar with the process of creating a project, diagram and data source.

2. Explore the PVA97NK data set, note down the number of samples and number of variables, and check if there are missing values. (Refer to the instruction given in Step 2 of Task 5 of Practical#1 on how to explore a data set.)



Property	Value
Rows	9686

Obs #	Target ...	Control...	Target ...
1		000014974	
2		000006294	
3		100046110	$4.0

Obs #	Variabl...	Label	Type
1	DemClus...	Demogra...	CLASS
2	DemGen...	Gender	CLASS
3	DemHo...	Home O...	CLASS

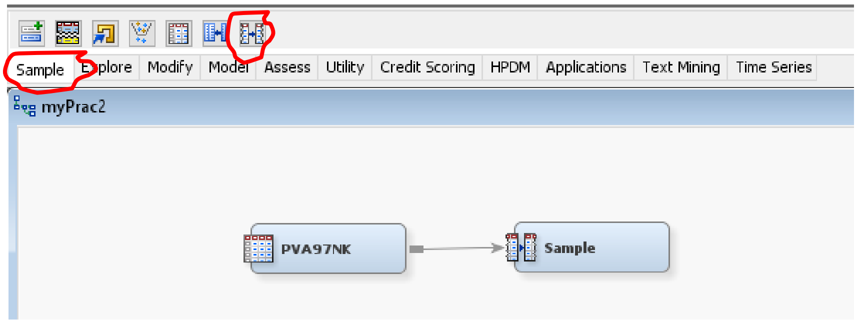
To see the number of samples and number of variables in the data set, in the **Explore** window, double click on the title bar of the **Sample Properties** window to enlarge this window. The number of samples and number of variables in the data set are shown as the **Value** of **Rows** and **Columns** respectively in the **Sample Properties** window. To restore this window to its original (small) size, double click on its title bar again.

To check if there are missing values, in the **Explore** window, double click on the title bar of the **Sample Statistics** window. In the enlarged **Sample Statistics** window, the **Percent Missing** column shows the % of missing values for each variable in the data set. Additionally, if a variable has missing values, you will see a grey bar in its histogram in the **Explore** window too.

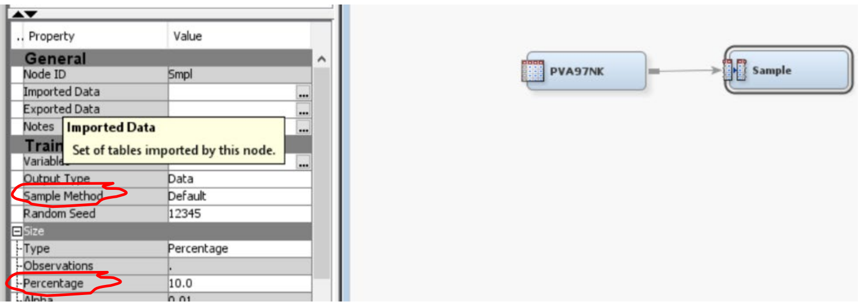
3. Sample the PVA97NK data set using the Sample tool of SAS EM.

   To create a Sample node, click on the **Sample** tab in the tools palette. Then you can see 7 tools/nodes shown above the tabs. Drag the Sample node and drop it onto the workspace.

   Connect the Sample node with the Input Data node.



4. Click on the Sample node created, the properties of the Sample node are shown in the **Properties** panel (on the left of the workspace). Look through the properties to know what options are available. Put mouse over a property item, you may see details of the property.



Property	Value
General	
Node ID	Smpl
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Sample Method	Default
Random Seed	12345
Size	
Type	Percentage
Observations	.
Percentage	10.0
Alpha	0.01

Imported Data

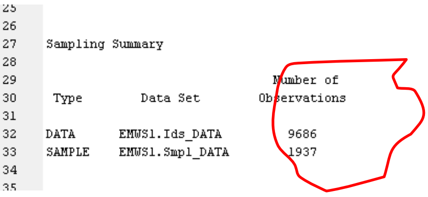
Set of tables imported by this node.

5. Use the random sampling method to sample the PVA97NK data set.

In the **Properties** panel while the Sample node is highlighted (if not, click on the node), click on the white field next to **Sample Method**, a dropdown menu will appear. Change sample method from **Default** to **Random**. Then click on the white field next to **Percentage** and change the default percentage (10.0) to 20 by typing 20 in the field.

Now right click on the Sample node, and choose **Run**, to conduct the random sampling.

In the **Run Status** window popped up, click on **Results**, the Result window will show, and you can see that the sampled data set contains 1937 observations (which is 20% of observations of the whole data set).

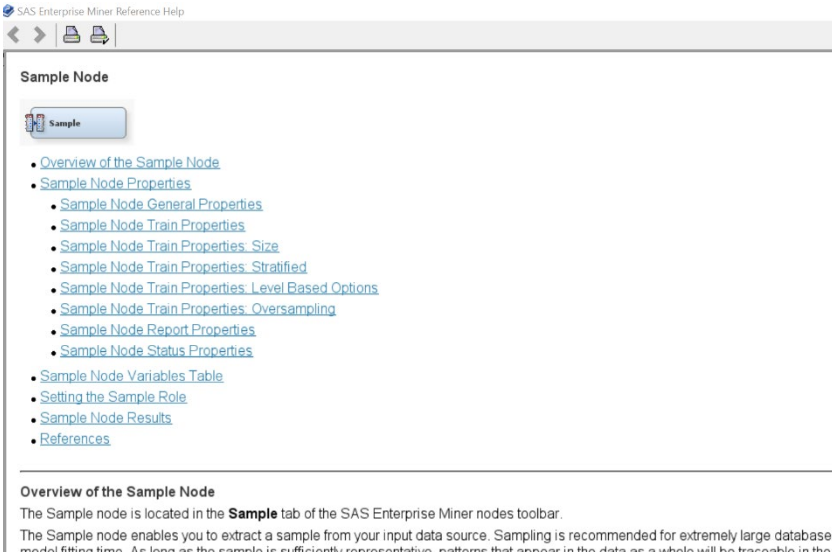


```
Sampling Summary

                                  Number of
 Type       Data Set             Observations

DATA      EMWS1.Ids_DATA           9686
SAMPLE    EMWS1.Smpl_DATA          1937
```

6. Now open the **Help** reference of the Sample node by selecting **File → Help → Content**, then in the navigate pane of the Help page, double click on **Sample**, then click on **Sample Node**. Read the overview of the Sample node, and the content related to random sampling, and the other sampling methods provided by SAS EM.



SAS Enterprise Miner Reference Help

Sample Node

- Overview of the Sample Node
- Sample Node Properties
  - Sample Node General Properties
  - Sample Node Train Properties
  - Sample Node Train Properties: Size
  - Sample Node Train Properties: Stratified
  - Sample Node Train Properties: Level Based Options
  - Sample Node Train Properties: Oversampling
  - Sample Node Report Properties
  - Sample Node Status Properties
- Sample Node Variables Table
- Setting the Sample Role
- Sample Node Results
- References

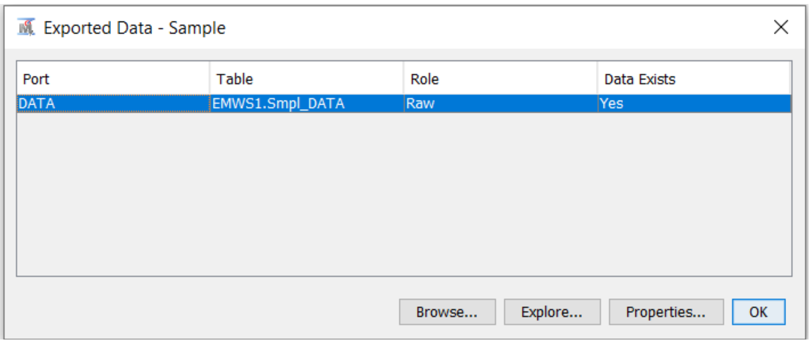
Overview of the Sample Node

The Sample node is located in the **Sample** tab of the SAS Enterprise Miner nodes toolbar.

The Sample node enables you to extract a sample from your input data source. Sampling is recommended for extremely large database

7. Explore the sampled data.

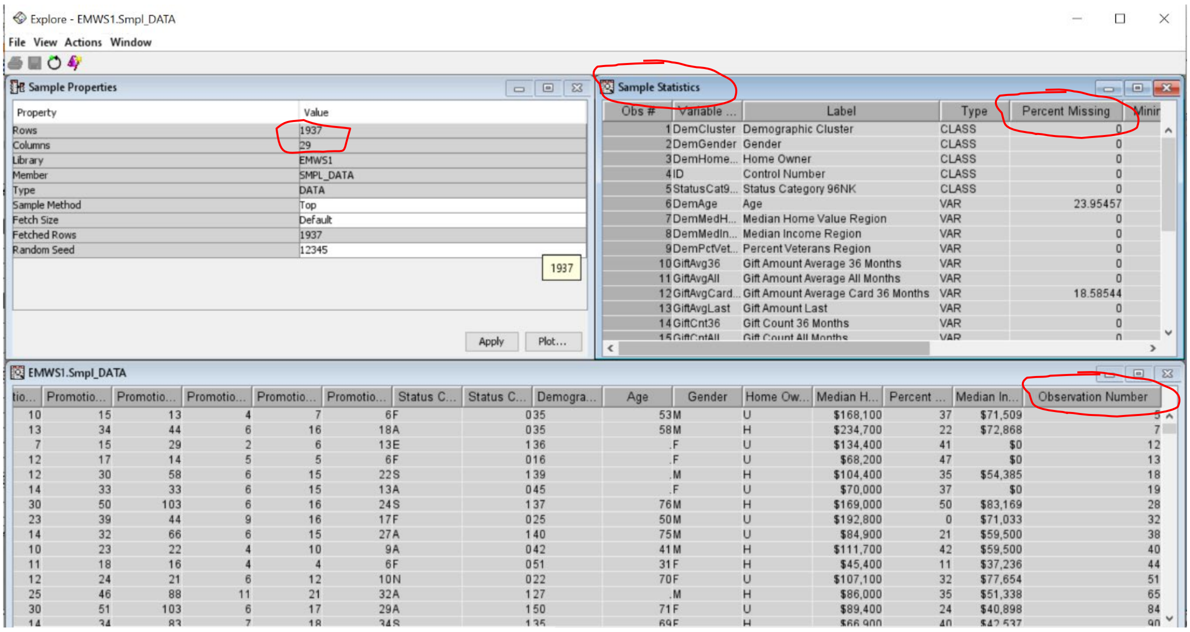
In the **Properties** panel of the Sample node (i.e. when the Sample node in the diagram is highlighted), click on the **…** (ellipse) button shown next to **Exported Data**, the **Exported Data – Sample** window will appear. In this window, click on the first row of the table (there is only 1 row in the table, in addition to the column headers), as indicated in the figure below, then click on the **Explore** button.



Exported Data - Sample

Port	Table	Role	Data Exists
DATA	EMWS1.Smpl_DATA	Raw	Yes

The **Explore** window shows:



Property	Value
Rows	1937
Columns	29
Library	EMWS1
Member	SMPL_DATA
Type	DATA
Sample Method	Top
Fetch Size	Default
Fetched Rows	1937
Random Seed	12345

Obs #	Variable ...	Label	Type	Percent Missing	Minir
1	DemCluster	Demographic Cluster	CLASS	0	
2	DemGender	Gender	CLASS	0	
3	DemHome...	Home Owner	CLASS	0	
4	ID	Control Number	CLASS	0	
5	StatusCat9...	Status Category 96NK	CLASS	0	
6	DemAge	Age	VAR	23.95457	
7	DemMedH...	Median Home Value Region	VAR	0	
8	DemMedIn...	Median Income Region	VAR	0	
9	DemPctVet...	Percent Veterans Region	VAR	0	
10	GiftAvg36	Gift Amount Average 36 Months	VAR	0	
11	GiftAvgAll	Gift Amount Average All Months	VAR	0	
12	GiftAvgCard...	Gift Amount Average Card 36 Months	VAR	18.58544	
13	GiftAvgLast	Gift Amount Last	VAR	0	
14	GiftCnt36	Gift Count 36 Months	VAR	0	
15	GiftCntAll	Gift Count All Months	VAR	0	

tio...	Promotio...	Promotio...	Promotio...	Promotio...	Promotio...	Status C...	Status C...	Demogra...	Age	Gender	Home Ow...	Median H...	Percent ...	Median In...	Observation Number
10	15	13	4	7	6	F	0	35	53	M	U	$168,100	37	$71,509	5
13	34	44	6	16	18	A	0	35	58	M	H	$234,700	22	$72,868	7
7	15	29	2	6	13	E	1	36	.	F	U	$134,400	41	$0	12
12	17	14	5	5	6	F	0	16	.	F	U	$68,200	47	$0	13
12	30	58	6	15	22	S	1	39	.	M	H	$104,400	35	$54,385	18
14	33	33	6	15	13	A	0	45	.	F	U	$70,000	37	$0	19
30	50	103	6	16	24	S	1	37	76	M	H	$169,000	50	$83,169	28
23	39	44	9	16	17	F	0	25	50	M	U	$192,800	0	$71,033	32
14	32	66	6	15	27	A	1	40	75	M	U	$84,900	21	$59,500	38
10	23	22	4	10	9	A	0	42	41	M	H	$111,700	42	$59,500	40
11	18	16	4	4	6	F	0	51	31	F	H	$45,400	11	$37,236	44
12	24	21	6	12	10	N	0	22	70	F	U	$107,100	32	$77,654	51
25	46	88	11	21	32	A	1	27	.	M	H	$86,000	35	$51,338	65
30	51	103	6	17	29	A	1	50	71	F	U	$89,400	24	$40,898	84
14	34	83	7	18	34	S	1	35	69	F	H	$66,900	40	$42,537	90

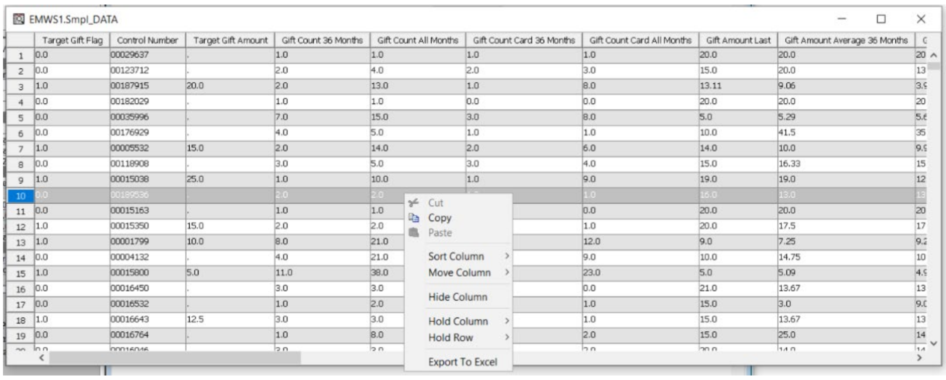
Note that the number of samples is now shown as 1937, 20% of the samples in the original data set. Interestingly, the number of variables is 29, while the original dataset only contains 28 variables. Drag the scroll bar at the bottom, you will find that an extra column (variable) named **Observation Number** is added, which shows the record ID of each sample in the original data set. (You can expand a column to see the full column header name.)

Now have a look at the **Sample Statistics** shown in the top right window. You will see that there are missing values in the sampled data set too. Move the scroll bar on the right of the **Sample Statistics** window, you will notice that the three variables, **Age**, **Gift Amount Average Card 36 Month**s, and **Target Gift Amount** have 23.95457%, 18.58544% and 49.6128% of missing values respectively.

8. Export the sampled data set to your local computer as a **.csv** file.

Close the **Explore** window by clicking on the **x** button at the top right corner of the **Explore** window (not the **x** in the **Sample Statistics** window). The **Exported Data – Sample** window will show again. Click on the **Browse** button (while the first row is still selected).

A data browse window appears with the sampled data records, as indicated below. Move the scroll bars on the right and at the bottom to browse the data records. Then right click mouse anywhere on the data browse window and choose **Export to Excel**.



EMWS1.Smpl_DATA

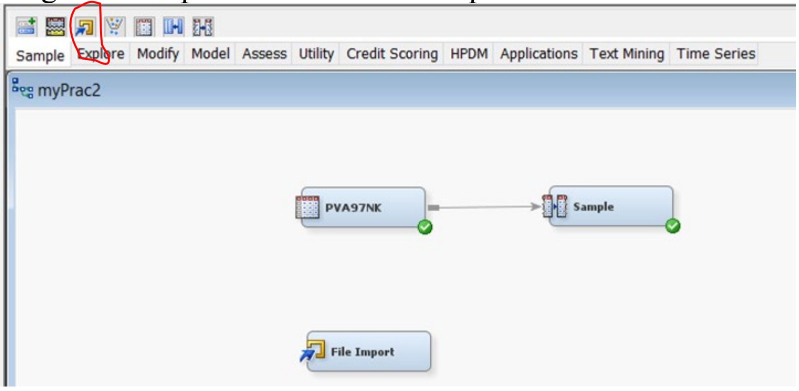
	Target Gift Flag	Control Number	Target Gift Amount	Gift Count 36 Months	Gift Count All Months	Gift Count Card 36 Months	Gift Count Card All Months	Gift Amount Last	Gift Amount Average 36 Months	G
1	0.0	00029637	.	1.0	1.0	1.0	1.0	20.0	20.0	20
2	0.0	00123712	.	2.0	4.0	2.0	3.0	15.0	20.0	13
3	1.0	00187915	20.0	2.0	13.0	1.0	8.0	13.11	9.06	3.9
4	0.0	00182029	.	1.0	1.0	0.0	0.0	20.0	20.0	20
5	0.0	00035996	.	7.0	15.0	3.0	8.0	5.0	5.29	5.6
6	0.0	00176929	.	4.0	5.0	1.0	1.0	10.0	41.5	35
7	1.0	00005532	15.0	2.0	14.0	2.0	6.0	14.0	10.0	9.9
8	0.0	00118908	.	3.0	5.0	3.0	4.0	15.0	16.33	15
9	1.0	00015038	25.0	1.0	10.0	1.0	9.0	19.0	19.0	12
10	0.0	00189536	.	2.0	2.0		1.0	16.0	13.0	13
11	0.0	00015163	.	1.0	1.0		0.0	20.0	20.0	20
12	1.0	00015350	15.0	2.0	2.0		1.0	20.0	17.5	17
13	1.0	00001799	10.0	8.0	21.0		12.0	9.0	7.25	9.2
14	0.0	00004132	.	4.0	21.0		9.0	10.0	14.75	10
15	1.0	00015800	5.0	11.0	38.0		23.0	5.0	5.09	4.9
16	0.0	00016450	.	3.0	3.0		0.0	21.0	13.67	13
17	0.0	00016532	.	1.0	2.0		1.0	15.0	3.0	9.0
18	1.0	00016643	12.5	3.0	3.0		1.0	15.0	13.67	13
19	0.0	00016764	.	1.0	8.0		2.0	15.0	25.0	14

The Excel app on your computer will be opened and choose **YES** in the popup window to continue. The sampled data is now shown in the Excel app. Use the **Save As** menu of Excel to save the data set as a **.csv** file on your computer, and name this .csv file as **prac#2-sample.csv**. Remember the folder where you save this file.

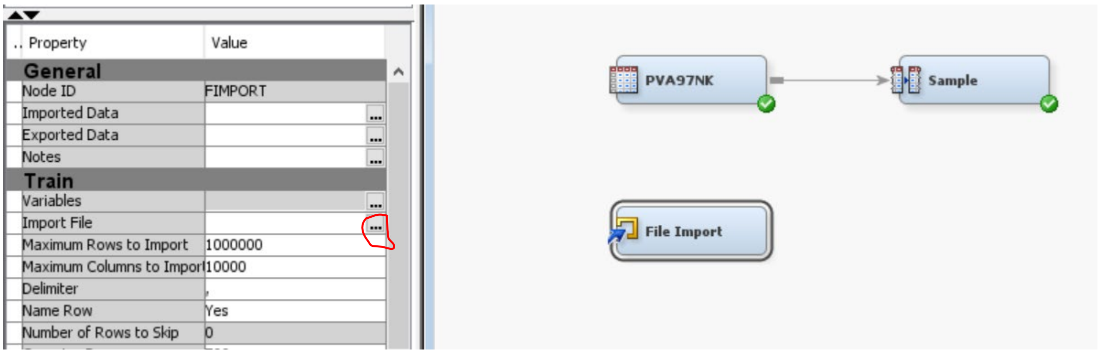
Click **OK** in the **Exported Data – Sample** window to close it.

9. Now let us create another process flow to import the saved .csv file and preprocess the data. (Note that you could connect nodes to the Sample node to preprocess the sampled data, but here we learn how to use the File Import node to load your own data, so we will create a new process flow which starts with a File Import node.)

In the same diagram workspace, make sure that the **Sample** tab in the tool palette is clicked, and drag a File Import node to the workspace.

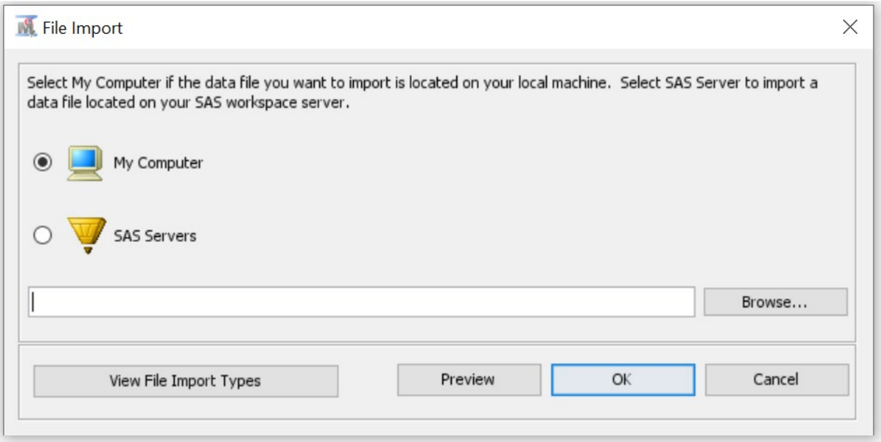


10. While the File Important node is highlighted, in the **Properties** panel, click on the **…** button next to **Import File**.



Property	Value
General	
Node ID	FIMPORT
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Import File	...
Maximum Rows to Import	1000000
Maximum Columns to Import	10000
Delimiter	,
Name Row	Yes
Number of Rows to Skip	0

The **File Import** window shows. In the window, choose **My Computer**, and then **Browse**, and select the prac#2-sample.csv file you saved in Step 8 above, and finally click **OK** in the window to import the .csv file (the sampled data set).



File Import

Select My Computer if the data file you want to import is located on your local machine. Select SAS Server to import a data file located on your SAS workspace server.

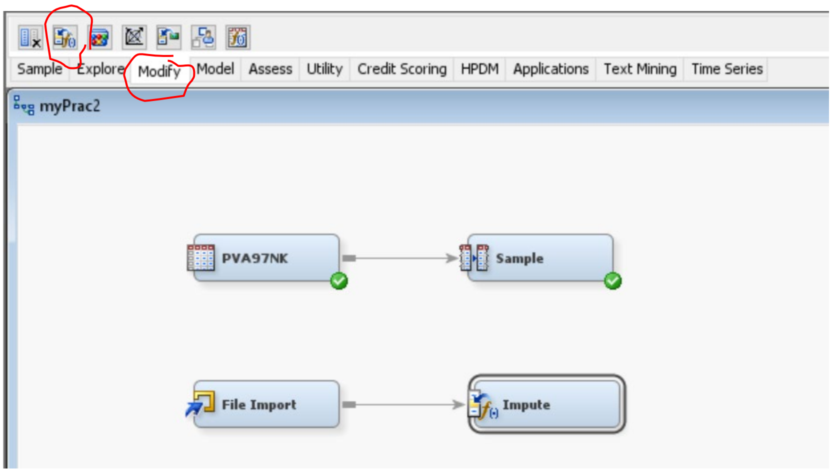
● My Computer

○ SAS Servers

11. Impute missing values using the Impute tool.

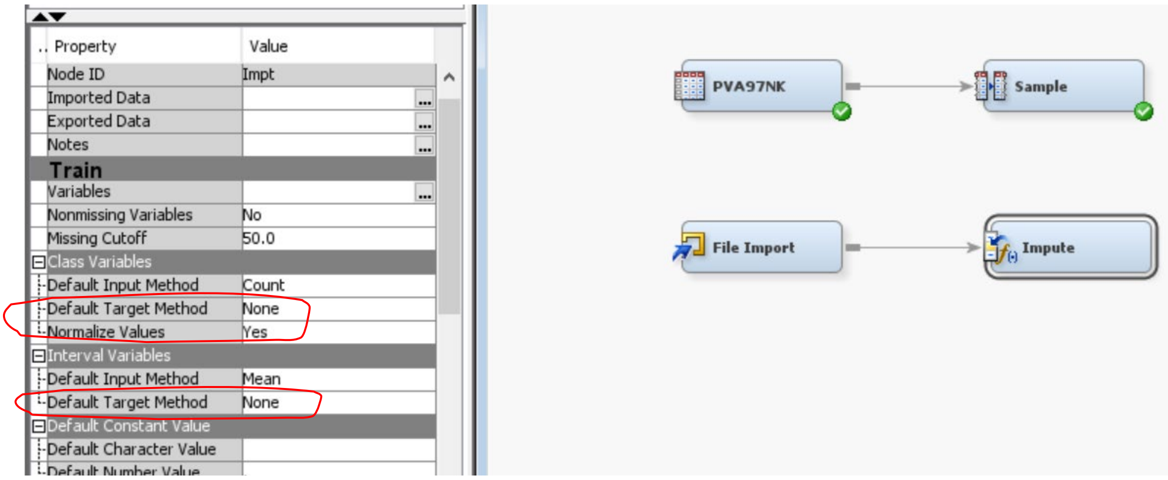
Recall that in **Step 7** above, we noted that in the sampled data set, three variables had missing values. In this step, we use the Impute tool to impute the missing values.

Click on the **Modify** tab in the tools palette, then drag the Impute node and drop it onto the workspace, and connect it with the File Import node.



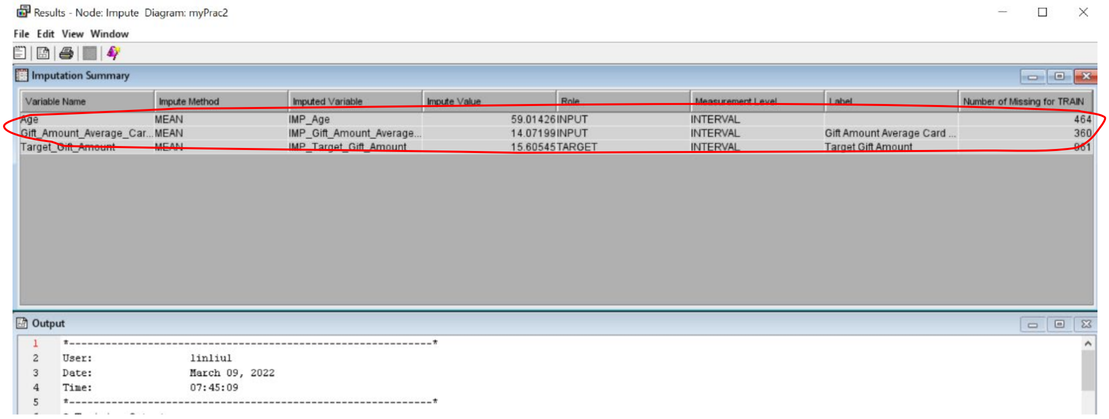
In the **Properties** panel, put your mouse over each item (or click on an item), and read the text shown in the yellow box (or in the **Property help** panel when you click on an item), to understand the meanings/purposes of these items. Also click on the white fields to see the options listed in the dropdown menu shown.
In the **Properties** panel of the Impute node, in the **Class Variables** section, change the **Default Target Method** from **None** to **Count**, and in the **Interval Variables** section, change the **Default Target Method** from **None** to **Mean.** (Note that in the PVA97NK dataset, variable **Targe Gift Amount** is set as a target variable for training a predictive model. By default, the Impute node does not imput missing values for a target variable, but only as a pratice of using the Impute node in this practical, we want to impute all missing values, so we change the default settings for target variables.)

Also in the **Class Variables** section, change **Normalized Values** from **Yes** to **No**.



Property	Value
Node ID	Impt
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Nonmissing Variables	No
Missing Cutoff	50.0
Class Variables	
Default Input Method	Count
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Mean
Default Target Method	None
Default Constant Value	
Default Character Value	
Default Number Value	

Now right click the Impute node and choose **Run**. When the Run Status window pops up, select **Results** to view the Results. Note that as expected, missing values of all the three variables have been imputed using the specified method (i.e. imputing with mean).



Results - Node: Impute Diagram: myPrac2

Imputation Summary

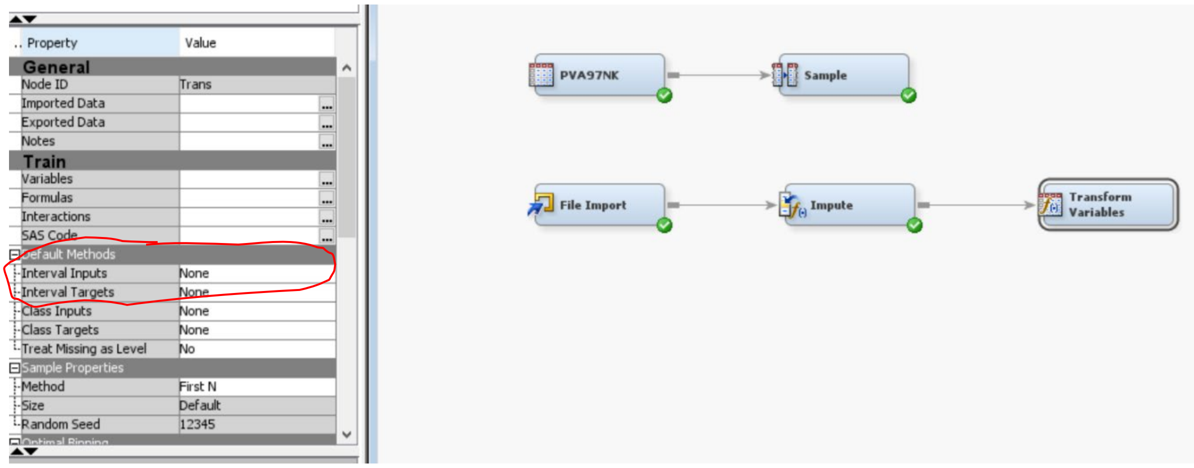
Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
Age	MEAN	IMP_Age	59.01426	INPUT	INTERVAL		464
Gift_Amount_Average_Car...	MEAN	IMP_Gift_Amount_Average...	14.07199	INPUT	INTERVAL	Gift Amount Average Card ...	360
Target_Gift_Amount	MEAN	IMP_Target_Gift_Amount	15.60545	TARGET	INTERVAL	Target Gift Amount	961

Output

```
*------------------------------------------------------------*
User:                 linliul
Date:                 March 09, 2022
Time:                 07:45:09
*------------------------------------------------------------*
```

12. Use the Transform Variables node to do data normalisation.

Drag a Transform node (when the **Modify** tab is selected), drop it onto the workspace, and connect it to the Impute node.



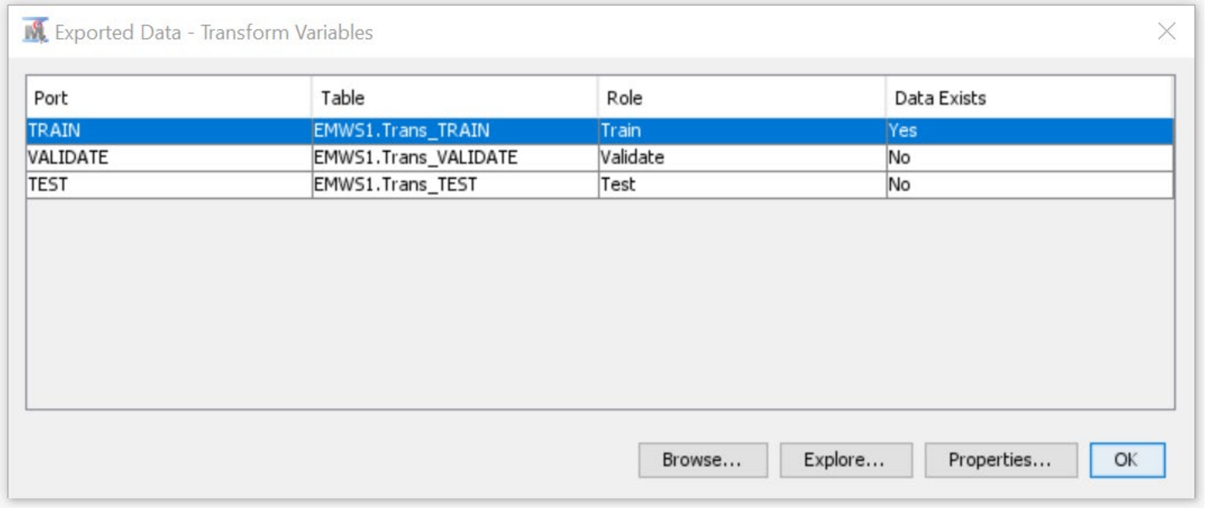
Property	Value
General	
Node ID	Trans
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Formulas	...
Interactions	...
SAS Code	...
Default Methods	
Interval Inputs	None
Interval Targets	None
Class Inputs	None
Class Targets	None
Treat Missing as Level	No
Sample Properties	
Method	First N
Size	Default
Random Seed	12345
Optimal Binning	

In the **Properties** panel, put your mouse over each item (or click on an item), and read the text shown in the yellow box (or in the **Property help** panel when you click on an item), to understand the meanings/purposes of these items. Also click on the white fields to see the options listed in the dropdown menu shown.

Now in the **Properties** panel, in the **Default Methods** section, change the method for **Interval Inputs** and **Interval Targets** from **None** to **Standardize**. This will normalise an interval variable (with SAS terminology, interval variable means continuous variables) by subtracting the mean and dividing by the standard deviation, which is the Z-score normalisation method we have learned in lecture. (Note that you can also specify different normalisation methods for individual variables by clicking on the **…** button next to **Variables**, then in the Variables window opened changing the method for an individual variable.)

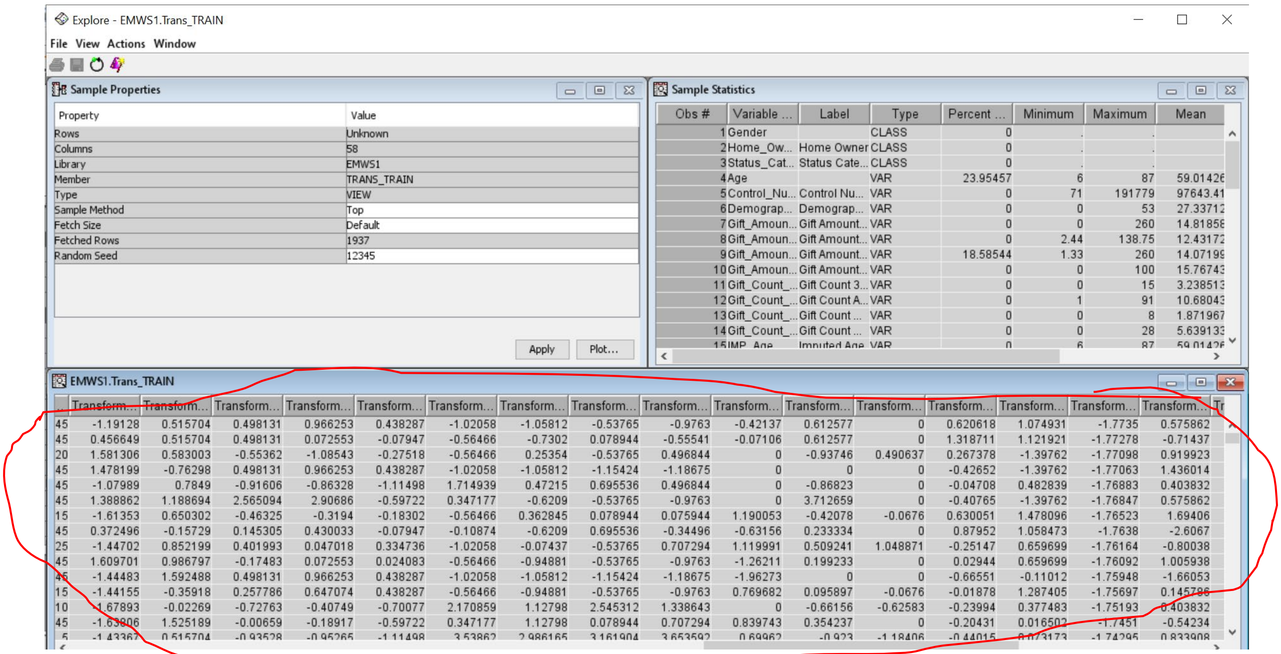
Now click on the Transform Variables node and choose **Run**.

After the Run has completed, in the **Properties** panel (when the Transform Variable node is highlighted), click on the **…** button next to **Exported Data**. When the **Exported Data – Transform Variables** window shows, click on the first row, and click on the **Explore** button.



Port	Table	Role	Data Exists
TRAIN	EMWS1.Trans_TRAIN	Train	Yes
VALIDATE	EMWS1.Trans_VALIDATE	Validate	No
TEST	EMWS1.Trans_TEST	Test	No

Now the **Explore** window shows. Move the scroll bar at bottom, you will see the Transformed values of the (interval) variables.

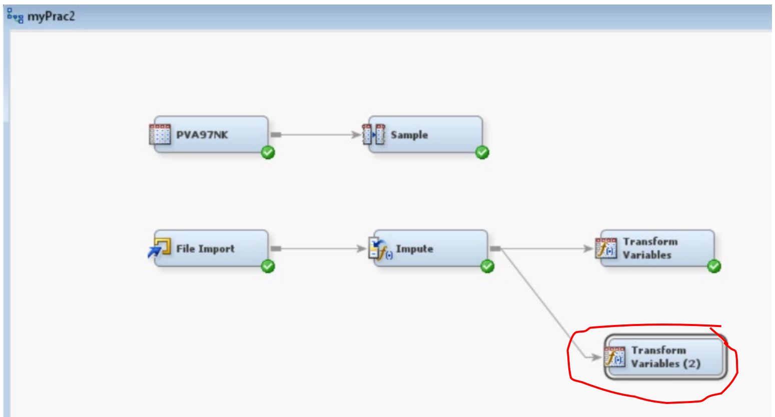


Property	Value
Rows	Unknown
Columns	58
Library	EMWS1
Member	TRANS_TRAIN
Type	VIEW
Sample Method	Top
Fetch Size	Default
Fetched Rows	1937
Random Seed	12345

Obs #	Variable ...	Label	Type	Percent ...	Minimum	Maximum	Mean
1	Gender		CLASS	0	.	.	
2	Home_Ow...	Home Owner	CLASS	0	.	.	
3	Status_Cat...	Status Cate...	CLASS	0	.	.	
4	Age		VAR	23.95457	6	87	59.01426
5	Control_Nu...	Control Nu...	VAR	0	71	191779	97643.41
6	Demograp...	Demograp...	VAR	0	0	53	27.33712
7	Gift_Amoun...	Gift Amount...	VAR	0	0	260	14.81858
8	Gift_Amoun...	Gift Amount...	VAR	0	2.44	138.75	12.43172
9	Gift_Amoun...	Gift Amount...	VAR	18.58544	1.33	260	14.07199
10	Gift_Amoun...	Gift Amount...	VAR	0	0	100	15.76743
11	Gift_Count_...	Gift Count 3...	VAR	0	0	15	3.238513
12	Gift_Count_...	Gift Count A...	VAR	0	1	91	10.68043
13	Gift_Count_...	Gift Count ...	VAR	0	0	8	1.871967
14	Gift_Count_...	Gift Count ...	VAR	0	0	28	5.639133
15	IMP_Age	Imputed Age	VAR	0	6	87	59.01426

...	Transform...	Transform...	Transform...	Transform...	Transform...	Transform...	Transform...	Transform...	Transform...	Transform...	Transform...	Transform...	Transform...	Transform...	Transform...	Transform...	Tr
45	-1.19128	0.515704	0.498131	0.966253	0.438287	-1.02058	-1.05812	-0.53765	-0.9763	-0.42137	0.612577	0	0.620618	1.074931	-1.7735	0.575862	
45	0.456649	0.515704	0.498131	0.072553	-0.07947	-0.56466	-0.7302	0.078944	-0.55541	-0.07106	0.612577	0	1.318711	1.121921	-1.77278	-0.71437	
20	1.581306	0.583003	-0.55362	-1.08543	-0.27518	-0.56466	0.25354	-0.53765	0.496844	0	-0.93746	0.490637	0.267378	-1.39762	-1.77098	0.919923	
45	1.478199	-0.76298	0.498131	0.966253	0.438287	-1.02058	-1.05812	-1.15424	-1.18675	0	0	0	-0.42652	-1.39762	-1.77063	1.436014	
45	-1.07989	0.7849	-0.91606	-0.86328	-1.11498	1.714939	0.47215	0.695536	0.496844	0	-0.86823	0	-0.04708	0.482839	-1.76883	0.403832	
45	1.388862	1.188694	2.565094	2.90686	-0.59722	0.347177	-0.6209	-0.53765	-0.9763	0	3.712659	0	-0.40765	-1.39762	-1.76847	0.575862	
15	-1.61353	0.650302	-0.46325	-0.3194	-0.18302	-0.56466	0.362845	0.078944	0.075944	1.190053	-0.42078	-0.0676	0.630051	1.478096	-1.76523	1.69406	
45	0.372496	-0.15729	0.145305	0.430033	-0.07947	-0.10874	-0.6209	0.695536	-0.34496	-0.63156	0.233334	0	0.87952	1.058473	-1.7638	-2.6067	
25	-1.44702	0.852199	0.401993	0.047018	0.334736	-1.02058	-0.07437	-0.53765	0.707294	1.119991	0.509241	1.048871	-0.25147	0.659699	-1.76164	-0.80038	
45	1.609701	0.986797	-0.17483	0.072553	0.024083	-0.56466	-0.94881	-0.53765	-0.9763	-1.26211	0.199233	0	0.02944	0.659699	-1.76092	1.005938	
45	-1.44483	1.592488	0.498131	0.966253	0.438287	-1.02058	-1.05812	-1.15424	-1.18675	-1.96273	0	0	-0.66551	-0.11012	-1.75948	-1.66053	
15	-1.44155	-0.35918	0.257786	0.647074	0.438287	-0.56466	-0.94881	-0.53765	-0.9763	0.769682	0.095897	-0.0676	-0.01878	1.287405	-1.75697	0.145786	
10	-1.67893	-0.02269	-0.72763	-0.40749	-0.70077	2.170859	1.12798	2.545312	1.338643	0	-0.66156	-0.62583	-0.23994	0.377483	-1.75193	0.403832	
45	-1.63806	1.525189	-0.00659	-0.18917	-0.59722	0.347177	1.12798	0.078944	0.707294	0.839743	0.354237	0	-0.20431	0.016502	-1.7451	-0.54234	
5	-1.43367	0.515704	-0.93528	-0.95265	-1.11498	3.53862	2.986165	3.161904	3.653592	0.69962	-0.923	-1.18406	-0.44015	0.073173	-1.74295	0.833908	

Drag a Transform Variable node, drop it onto the workspace, and connect it to the Impute node.

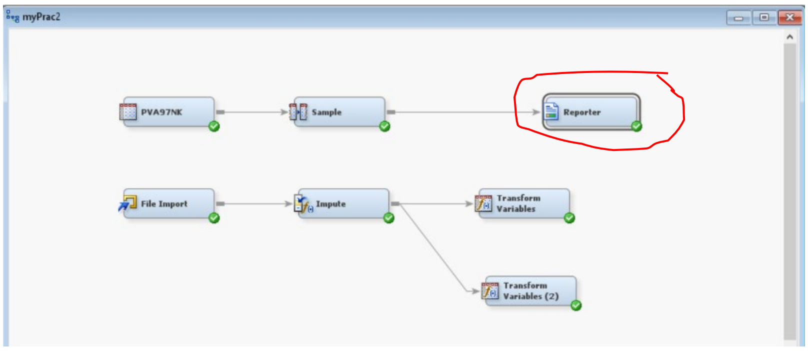


When the newly added Transform Variables node is highlighted, in the **Properties** panel, in the **Default Methods** section, change the transform methods for **Interval Inputs** and **Interval Targets** from **None** to **Range**. This normalisation method will transform a variable using the min-max method we have learned in lecture, i.e. the transformed value is equal to *(x - min) / (max - min)*, where *x* is current variable value, *min* is the minimum value for that variable, and *max* is the maximum value for that variable.

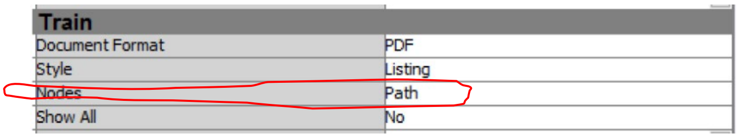
Now run this new Transform Variable node, and use the same way described above to see the transformed values, and observe the difference between the obtained transformed values from those obtained using the Z-score method.

13. Use the Reporter node to produce a report for Practical #2 submission.

Click on the **Utility** tab in the tools palette and then choose the Reporter node, drag it to the workspace, and connect it to the last node of any of the two process flows you have built.



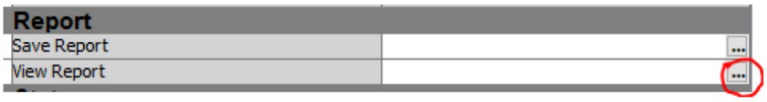
In the **Properties** panel of the Reporter node, in the **Train** section, **change Path to All** (**Note**: this change is crucial for Practical #2 submission, otherwise the report generated will only contain the information of the process flow/path to which the Reporter node is connected, instead of the information of all the process flows.)



Train	
Document Format	PDF
Style	Listing
Nodes	Path
Show All	No

Right-click the Reporter node and choose Run to generate the report. After the Run has completed (for all process flows), the Run Status window will show. Click **OK**.

In the **Properties** panel of the Reporter node, click the **…** button in the **View Report** field, the report will be opened in Adobe Reader. **Save the PDF report, and submit it by following the instruction given on page 1 of the practical document.**



Report	
Save Report	...
View Report	...