

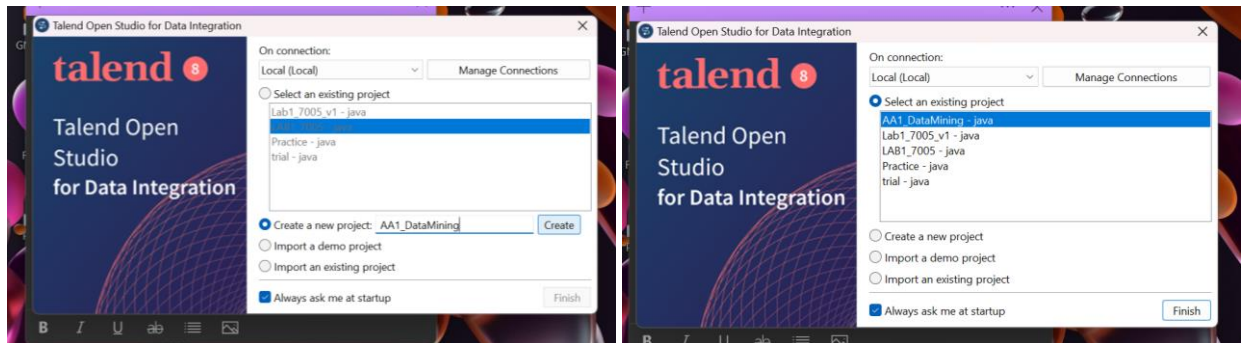
WQD7005

Alternative Assessment 1 – Case Study

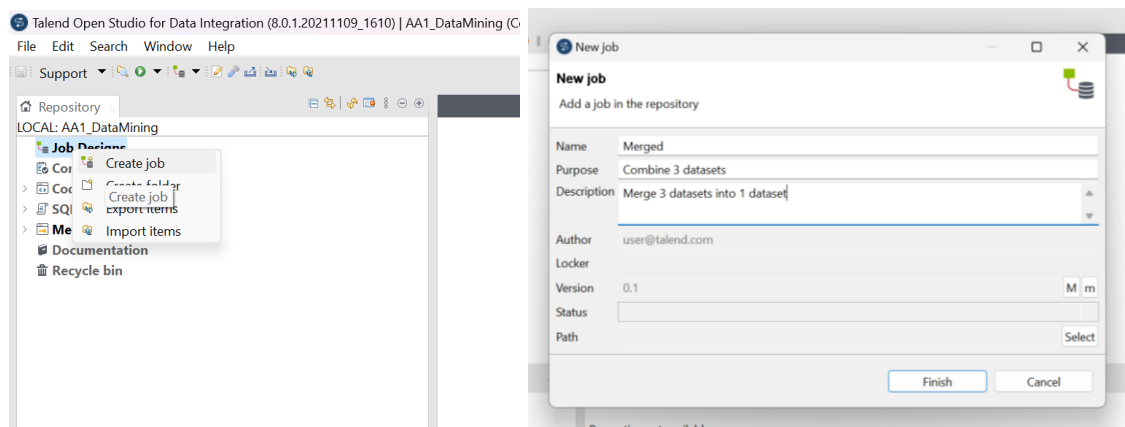
Farzana Syakira binti Bahari (22058163)

## MERGING THREE DATASETS INTO ONE DATASET USING TALEND OPEN STUDIO FOR DATA INTEGRATION

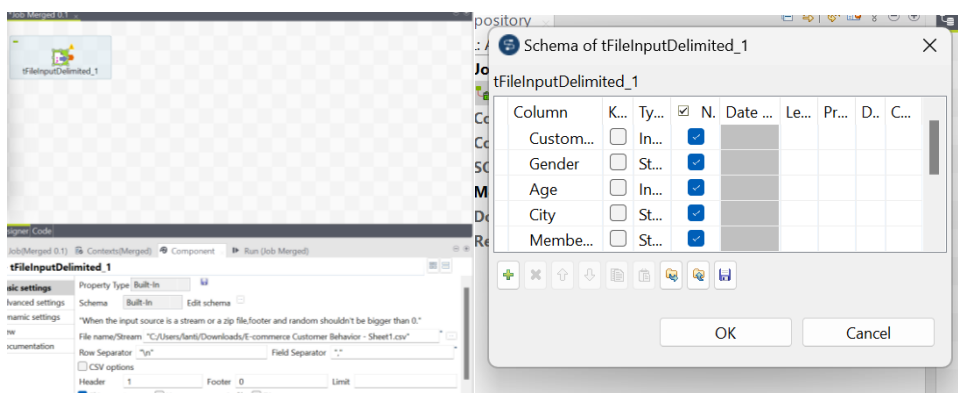
a. Create a new project in Talend Open Studio for Data Integration

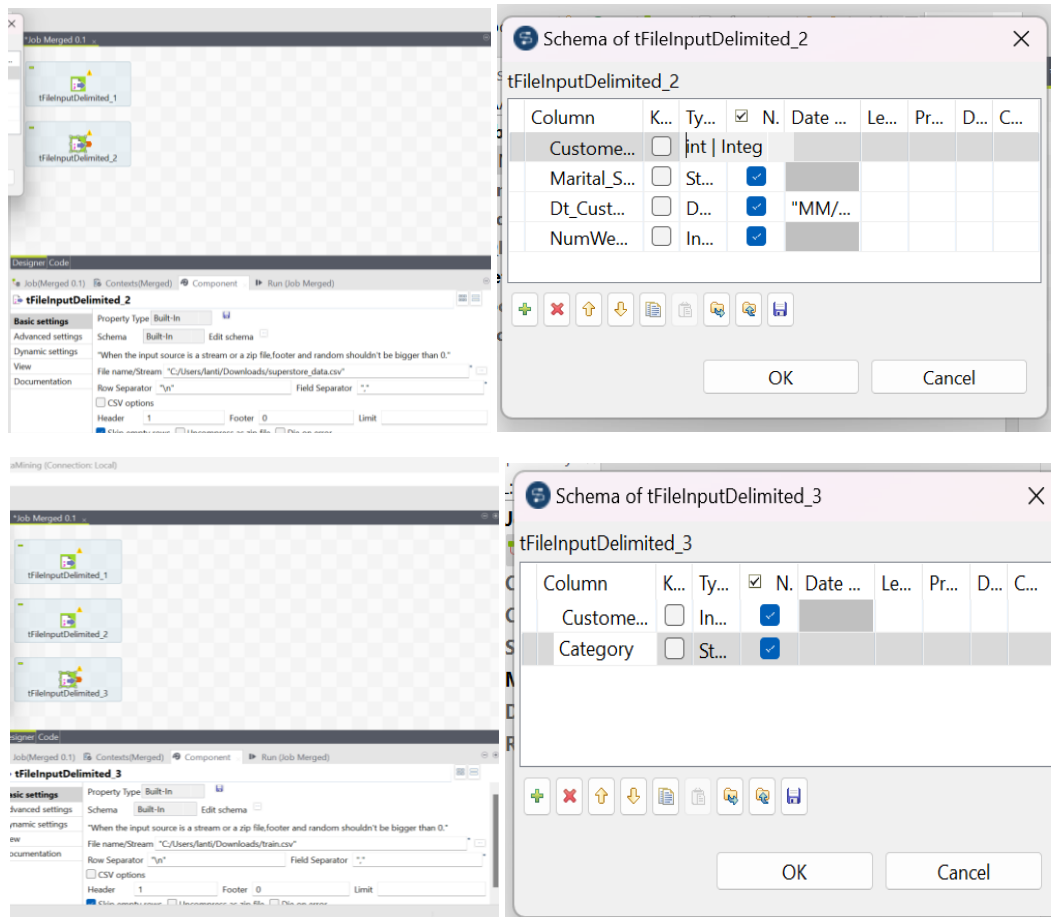


b. Create new job

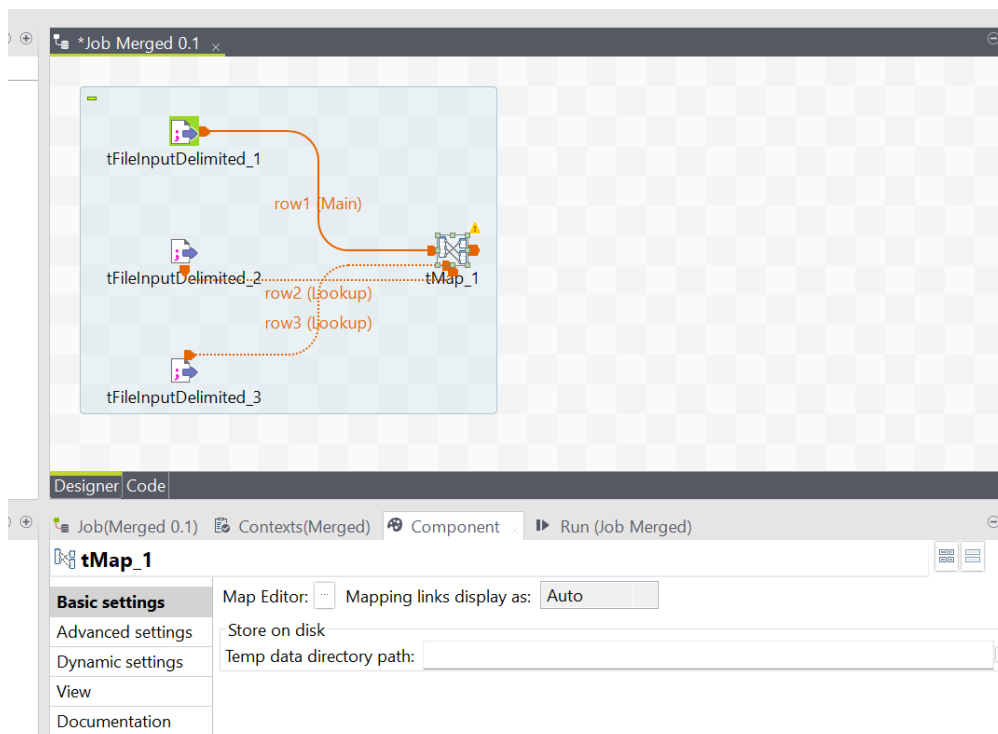


c. Ingesting the data and setting up the schema by dragging the “tFileInputDelimited” into the job space

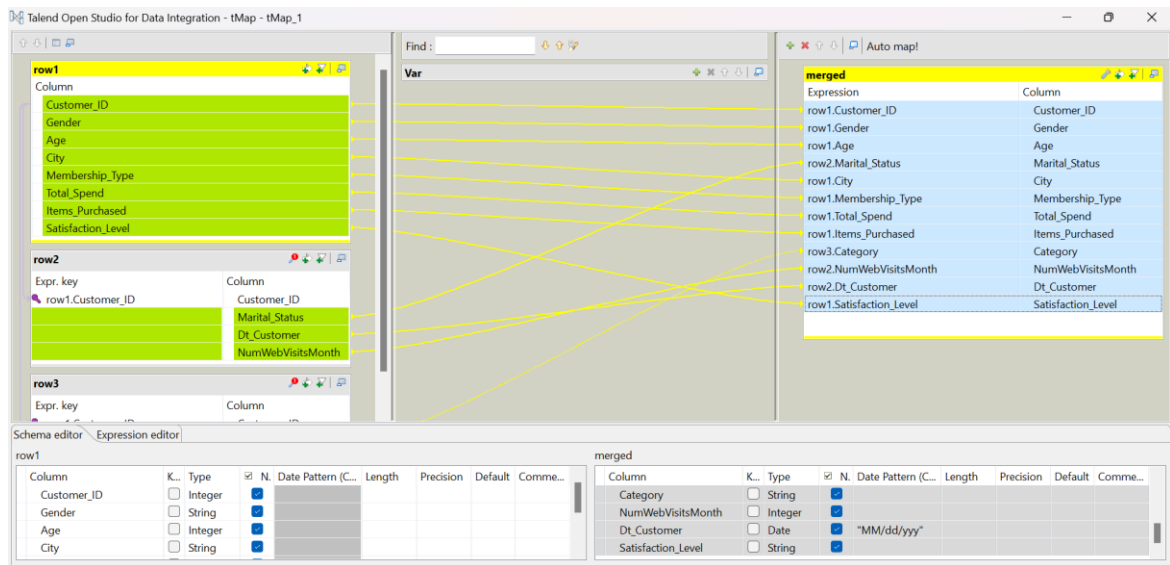




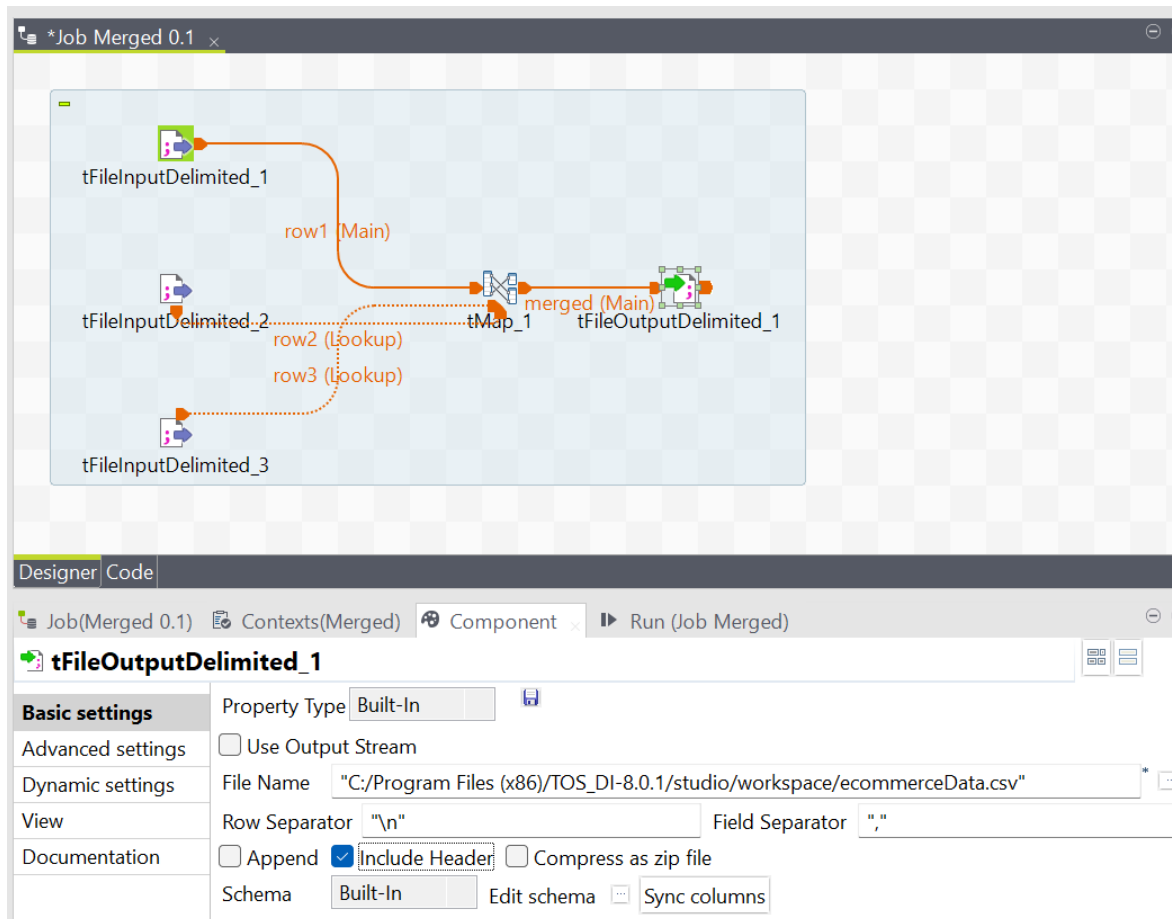
- d. Combine all the datasets using “tMap\_1” node by dragging it into the job space



- e. Connect all of the datasets using “Customer\_ID” column then map all the datasets into one new dataset.



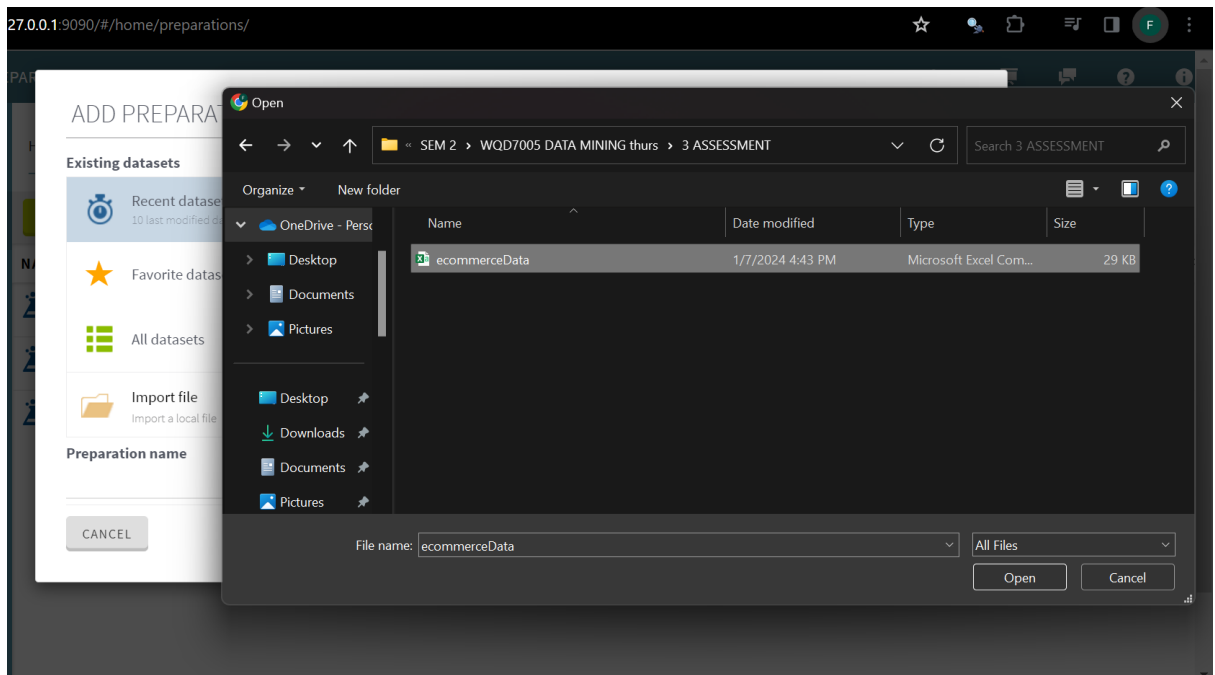
- f. Export the data using “tFileOutputDelimited\_1” node





## STANDARDIZING DAN CLEANING THE DATA USING TALEND DATA PREPARATION

### a. Ingesting the data



### b. Pre-process the data. This includes the standardization of the format of the date and values in the Marital\_Status column.

- All values in "Dt\_Customer" are changed from "MM/dd/yyyy" to "dd-MM-yyyy".
- Values in Marital\_Status such as "YOLO" and "Alone" are replaced by "Single".
- Values in Marital\_Status such as "Widow" is replaced by "Divorced".
- Values in Marital\_Status such as "Together" is replaced by "Married".
- The column name "Dt\_Customer" is renamed to "CustomerEnrollmentDate"

The screenshot shows the Talend Data Preparation interface with the 'ecommerceData Preparation' workflow. The workflow consists of six steps:

- 1 Change date format on column Dt\_Customer
- 2 Replace the cells that match on column Marital\_Status
- 3 Replace the cells that match on column Marital\_Status
- 4 Replace the cells that match on column Marital\_Status
- 5 Replace the cells that match on column Marital\_Status
- 6 Rename column on column Dt\_Customer

The 'Filters' section shows a table with columns: ory, NumWebVisitsM..., CustomerEnroll..., and Satisfaction\_Level. The table contains 13 rows of data.

The 'CustomerEnrollmentDate' section shows a table with columns: COLUMN and ROW. The table contains 13 rows of data.

The 'CHART' section shows a bar chart with the title 'Occurrences' and the x-axis labeled 'Occurrences'. The chart shows the frequency of values in the 'CustomerEnrollmentDate' column.

c. Exporting the file as ecommerceDataPrep.xlsx file

The screenshot displays the Talend Data Preparation interface. On the left, a workflow is shown with six steps: 1. Change date format on column Dt\_Customer, 2. Replace the cells that match on column Marital\_Status, 3. Replace the cells that match on column Marital\_Status, 4. Replace the cells that match on column Marital\_Status, 5. Replace the cells that match on column Marital\_Status, and 6. Rename column on column Dt\_Customer. The central pane shows a data table with columns: Customer, NumWebVisitsMonth, CustomerEnrollmentDate, and Satisfaction\_Level. The right pane shows the 'EXPORT' button and a dropdown menu with options: Local CSV file, Local XLSX file, and Local TABLEAU file. Below the dropdown, there are suggestions for functions like 'Calculate time unit...', 'Extract date parts...', and 'Change date format...'. A 'CHART' section is also visible, showing a bar chart with 'Occurrences' on the y-axis and 'ROW COUNT' on the x-axis.

talend DATA PREPARATION

ecommerceData Preparation

1 Change date format on column Dt\_Customer

2 Replace the cells that match on column Marital\_Status

3 Replace the cells that match on column Marital\_Status

4 Replace the cells that match on column Marital\_Status

5 Replace the cells that match on column Marital\_Status

6 Rename column on column Dt\_Customer

New name: 127.0.0.1:9090/rollmentDate

Filters

Add a filter ...

Customer	NumWebVisitsMonth	CustomerEnrollmentDate	Satisfaction_Level
ure	4	16-06-2014	Satisfied
ure	7	15-06-2014	Neutral
Supplies	3	13-05-2014	Unsatisfied
ure	1	05-11-2014	Satisfied
Supplies	3	04-08-2014	Unsatisfied
ure	4	17-03-2014	Neutral
Supplies	10	29-01-2014	Satisfied
logy	2	18-01-2014	Neutral
Supplies	6	01-11-2014	Unsatisfied
Supplies	6	01-11-2014	Satisfied
ure	5	27-12-2013	Unsatisfied
logy	3	12-09-2013	Neutral
Supplies	3	12-07-2013	Satisfied

EXPORT

Customize

Local CSV file

Local XLSX file

Local TABLEAU file

SUGGESTION

Calculate time unit...

Extract date parts...

Change date format...

CHART

VALUE

PATTERN

ADVANCED

ROW COUNT

Occurrences

50

40

30

20

10

0

talend DATA PREPARATION

ecommerceData Preparation

1 Change date format on column Dt\_Customer

2 Replace the cells that match on column Marital\_Status

3 Replace the cells that match on column Marital\_Status

4 Replace the cells that match on column Marital\_Status

5 Replace the cells that match on column Marital\_Status

6 Rename column on column Dt\_Customer

New name: CustomerEnrollmentDate

EXPORT TO XLSX

Filename:

ecommerceDataPrep

CANCEL

EXPORT

Extract date parts...

Change date format...

CHART

VALUE

PATTERN

ADVANCED

ROW COUNT

Occurrences

50

40

30

20

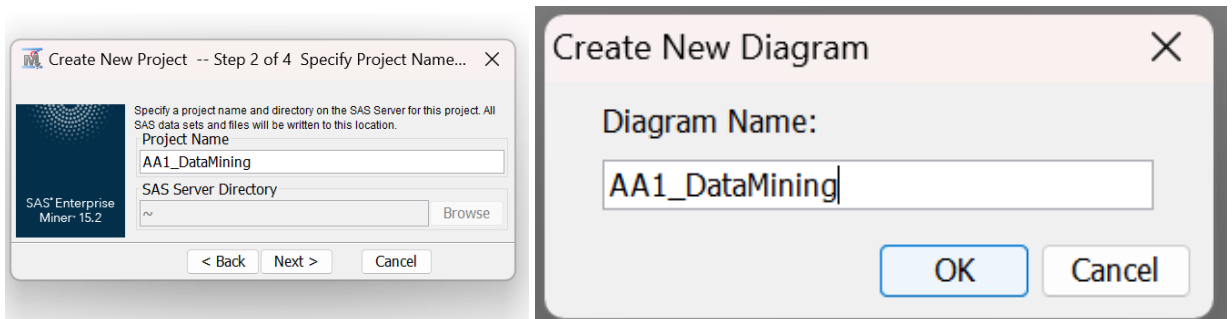
10

0

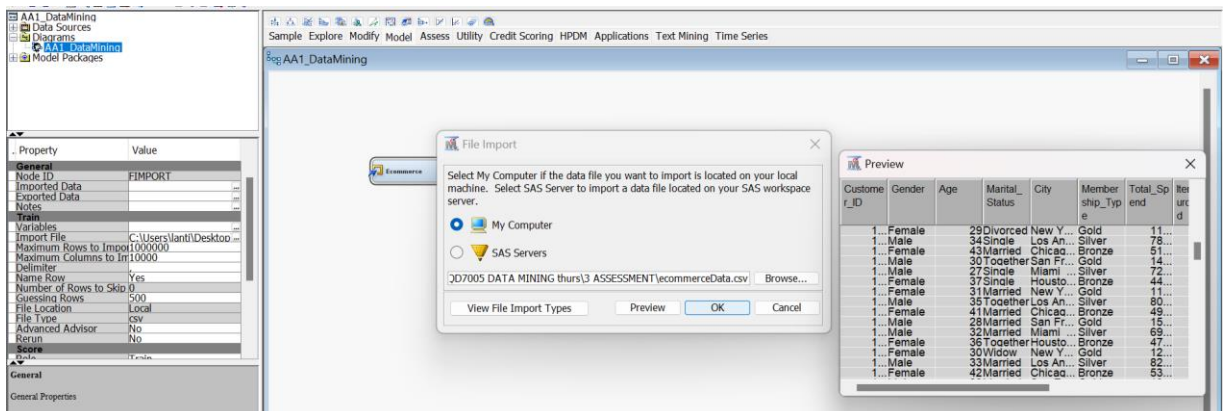
## ANALYSING THE DATA USING SAS EM

### Data Import and pre-processing

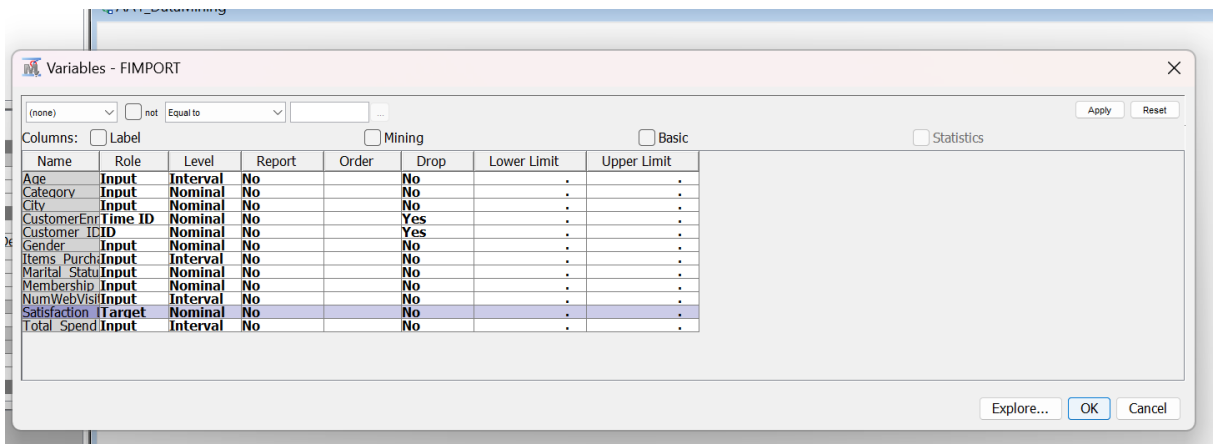
- Creating a new project and a new diagram

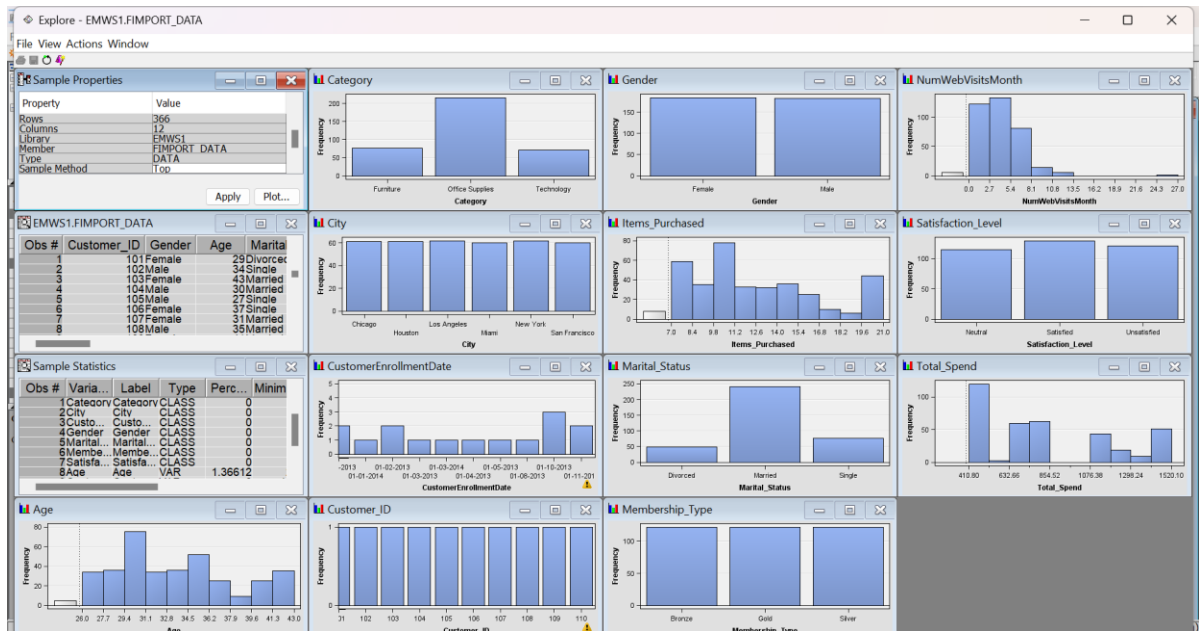


- Ingesting the data and reviewing it to ensure that the correct data is ingested

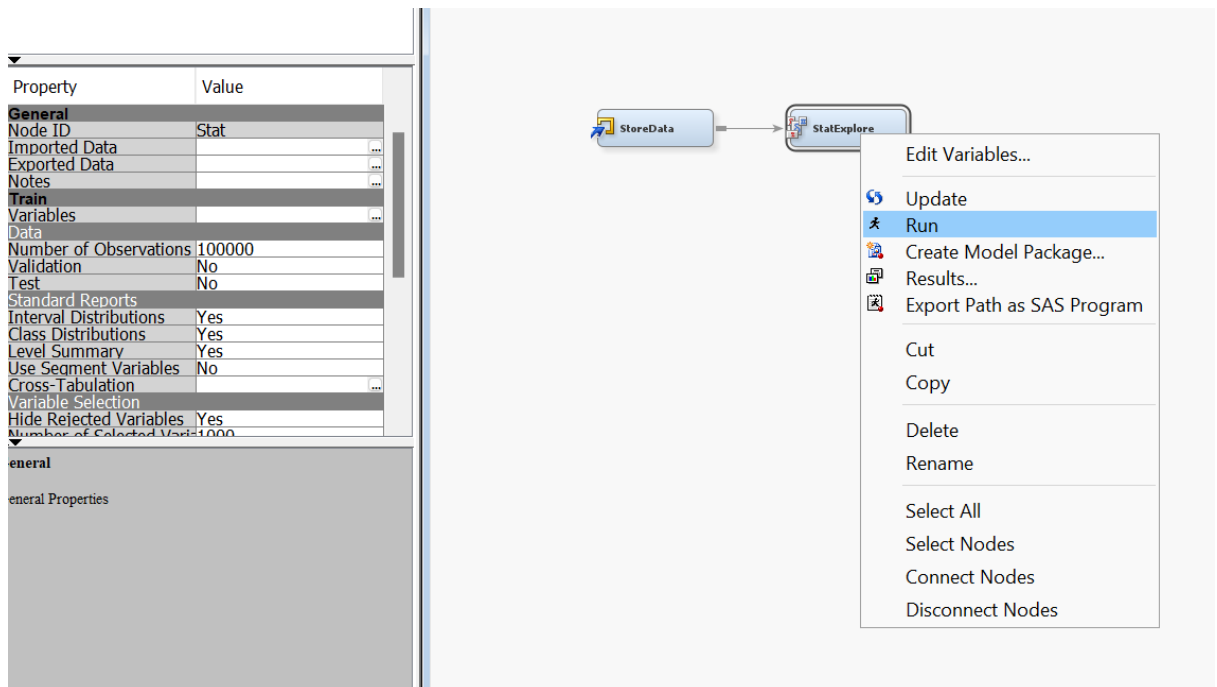


- Edit the variables to assign the roles for each of the variables. The target variable will be the Satisfaction\_Level. Customer\_ID is assigned as the ID and the CustomerEnrollmentDate is assigned as Time ID. The data is explored to see every variable's distributions.





- d. The StatExplore is attached to the diagram to see if there are missing values in the data. The categorical variables do not have any missing values whereas the interval variables such as Age, Items\_purchased and NumWebVisitMonth have missing values each.





Class Variable Summary Statistics  
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Category	INPUT	3	0	Office Supplies	59.29	Furniture	21.04
TRAIN	City	INPUT	6	0	Los Angeles	16.94	New York	16.94
TRAIN	Gender	INPUT	2	0	Female	50.27	Male	49.73
TRAIN	Marital_Status	INPUT	3	0	Married	65.85	Single	21.04
TRAIN	Membership_Type	INPUT	3	0	Bronze	33.33	Gold	33.33
TRAIN	Satisfaction_Level	TARGET	3	0	Satisfied	35.52	Unsatisfied	33.06

Distribution of Class Target and Segment Variables  
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	Satisfaction_Level	TARGET	Satisfied	130	35.5191
TRAIN	Satisfaction_Level	TARGET	Unsatisfied	121	33.0601
TRAIN	Satisfaction_Level	TARGET	Neutral	115	31.4208

Interval Variable Summary Statistics  
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Age	INPUT	33.62881	4.858058	361	5	26	33	43	0.463465	-0.76948
Items_Purchased	INPUT	12.60056	4.154536	358	8	7	12	21	0.653491	-0.61776
NumWebVisitsMonth	INPUT	4.141667	3.027163	360	6	0	4	27	2.410881	13.94905
Total_Spend	INPUT	844.1173	361.6808	366	0	410.8	770.2	1520.1	0.566578	-1.07491

## e. Imputing the missing values using Mean

The screenshot displays the SAS Enterprise Miner interface. On the left, the 'Properties' pane shows the configuration for the 'Impute' node. The 'General' tab is active, showing 'Node ID' as 'Impt'. Under the 'Train' section, 'Variables' are set to 'All', 'Nonmissing Variables' to 'No', and 'Missing Cutoff' to '50.0'. Under the 'Class Variables' section, 'Default Input Method' is 'Count', 'Default Target Method' is 'None', and 'Normalize Values' is 'Yes'. Under the 'Interval Variables' section, 'Default Input Method' is 'Mean', 'Default Target Method' is 'None', 'Default Constant Value' is '0', and 'Default Number Value' is '0'. The 'General' tab also shows 'General Properties'.

On the right, the 'Diagram' pane shows a workflow with three nodes: 'StoreData', 'StatExplore', and 'Impute'. The 'Impute' node is selected, and a context menu is open, showing options such as 'Update', 'Run', 'Create Model Package...', 'Results...', 'Export Path as SAS Program', 'Cut', 'Copy', 'Delete', 'Rename', 'Select All', 'Select Nodes', 'Connect Nodes', and 'Disconnect Nodes'.

At the bottom of the interface, the status bar indicates 'Run completed' and the user ID '22058163@siswa.ur'.

Variables - Impt2

Columns: ☐ Label ☐ Mining ☐ Basic ☐ Statistics

Name	Use	Method	Use Tree	Role	Level
Age	Default	Mean	Default	Input	Interval
Category	Default	Default	Default	Input	Nominal
City	Default	Default	Default	Input	Nominal
Gender	Default	Default	Default	Input	Nominal
Items_Purch	Default	Mean	Default	Input	Interval
Marital_Status	Default	Default	Default	Input	Nominal
Membership	Default	Default	Default	Input	Nominal
NumWebVisits	Default	Mean	Default	Input	Interval
Satisfaction	Default	Default	Default	Target	Nominal
Total_Spend	Default	Default	Default	Input	Interval

Explore... Update Path OK Cancel

Property Value

**General**

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

**General**

General Properties

Run completed

Diagram Log

22058163@siswa.ur

StoreData StatExplore Impute

Edit Variables...  
Update  
Run  
Create Model Package...  
Results...  
Export Path as SAS Program  
Cut  
Copy  
Delete  
Rename  
Select All  
Select Nodes  
Connect Nodes  
Disconnect Nodes

f. Check if the missing values have been imputed using the “StatExplore” node.

Property Value

**General**

Node ID Stat2

Imported Data

Exported Data

Notes

**Train**

Variables

**Data**

Number of Observations 100000

Validation No

Test No

**Standard Reports**

Interval Distributions Yes

Class Distributions Yes

Level Summary Yes

Use Segment Variables No

Cross-Tabulation

**Model Selection**

Hide Rejected Variables Yes

Number of Collected Variables 1000

**General**

General Properties

Run completed

Diagram Log

22058163@siswa.um.edu.my as u63454901

StoreData StatExplore Impute StatExplore after imputation

Edit Variables...  
Update  
Run  
Create Model Package...  
Results...  
Export Path as SAS Program  
Cut  
Copy  
Delete  
Rename  
Select All  
Select Nodes  
Connect Nodes  
Disconnect Nodes

(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Category	INPUT	3	0	Office Supplies	59.29	Furniture	21.04
TRAIN	City	INPUT	6	0	Los Angeles	16.94	New York	16.94
TRAIN	Gender	INPUT	2	0	Female	50.27	Male	49.73
TRAIN	Marital_Status	INPUT	3	0	Married	65.85	Single	21.04
TRAIN	Membership_Type	INPUT	3	0	Bronze	33.33	Gold	33.33
TRAIN	Satisfaction_Level	TARGET	3	0	Satisfied	35.52	Unsatisfied	33.06

Distribution of Class Target and Segment Variables  
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	Satisfaction_Level	TARGET	Satisfied	130	35.5191
TRAIN	Satisfaction_Level	TARGET	Unsatisfied	121	33.0601
TRAIN	Satisfaction_Level	TARGET	Neutral	115	31.4208

Interval Variable Summary Statistics  
(maximum 500 observations printed)

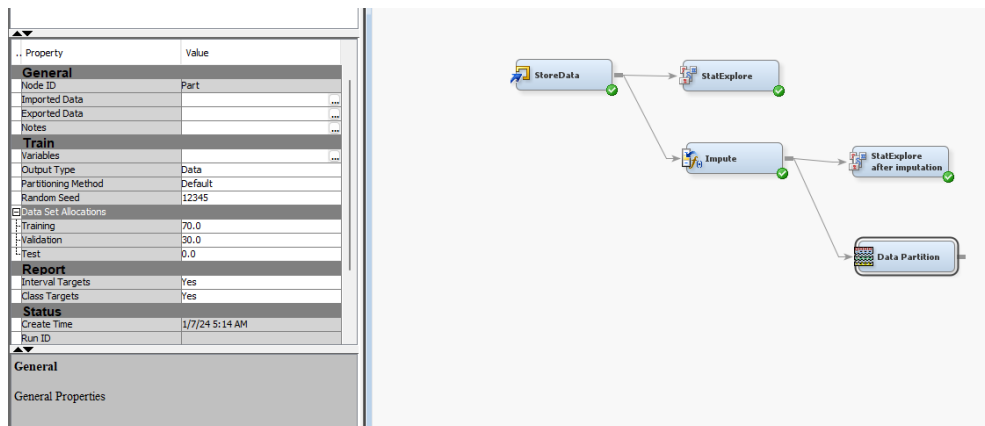
Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
IMP_Age	INPUT	33.62881	4.824668	366	0	26	33	43	0.466637	-0.73833
IMP_Items_Purchased	INPUT	12.60056	4.108754	366	0	7	12	21	0.660691	-0.56415
IMP_NumWebVisitsMonth	INPUT	4.141667	3.00218	366	0	0	4	27	2.430721	14.22842
Total_Spend	INPUT	844.1173	361.6808	366	0	410.8	770.2	1520.1	0.566578	-1.07491

Class Variable Summary Statistics by Class Target  
(maximum 500 observations printed)

Data Role=TRAIN Variable=Membership\_Type

- g. Attach “Data Partition” node to separate 70% of the data into testing set and the remaining into validation set. Run it and see the results to ensure that each value of the target variable has almost the same frequency.



# Summary Statistics for Class Targets

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Satisfaction_Level	.	Neutral	115	31.4208	Satisfaction_Level
Satisfaction_Level	.	Satisfied	130	35.5191	Satisfaction_Level
Satisfaction_Level	.	Unsatisfied	121	33.0601	Satisfaction_Level

Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Satisfaction_Level	.	Neutral	80	31.3725	Satisfaction_Level
Satisfaction_Level	.	Satisfied	90	35.2941	Satisfaction_Level
Satisfaction_Level	.	Unsatisfied	85	33.3333	Satisfaction_Level

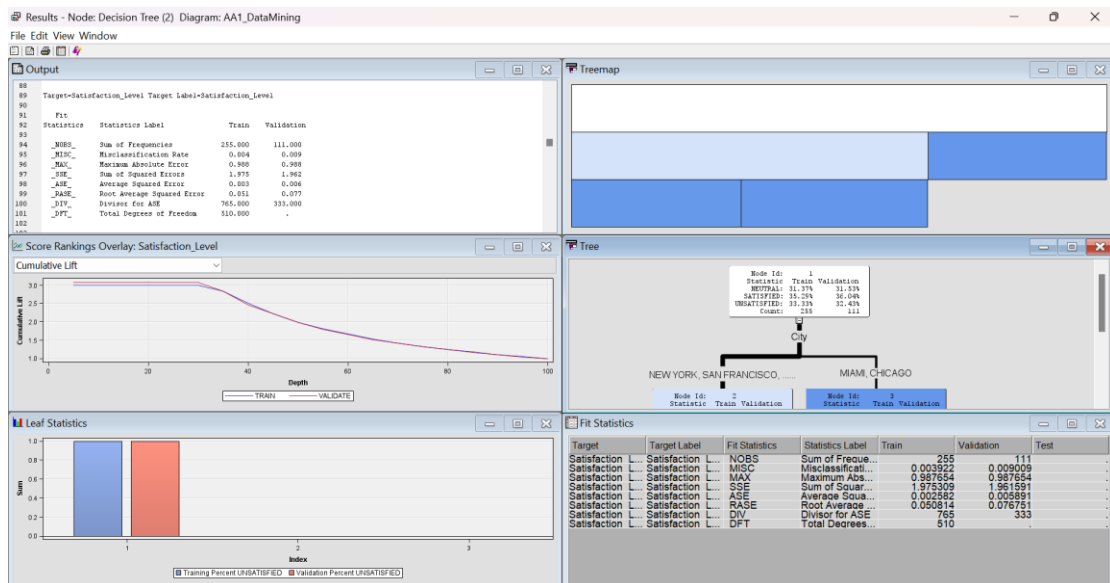
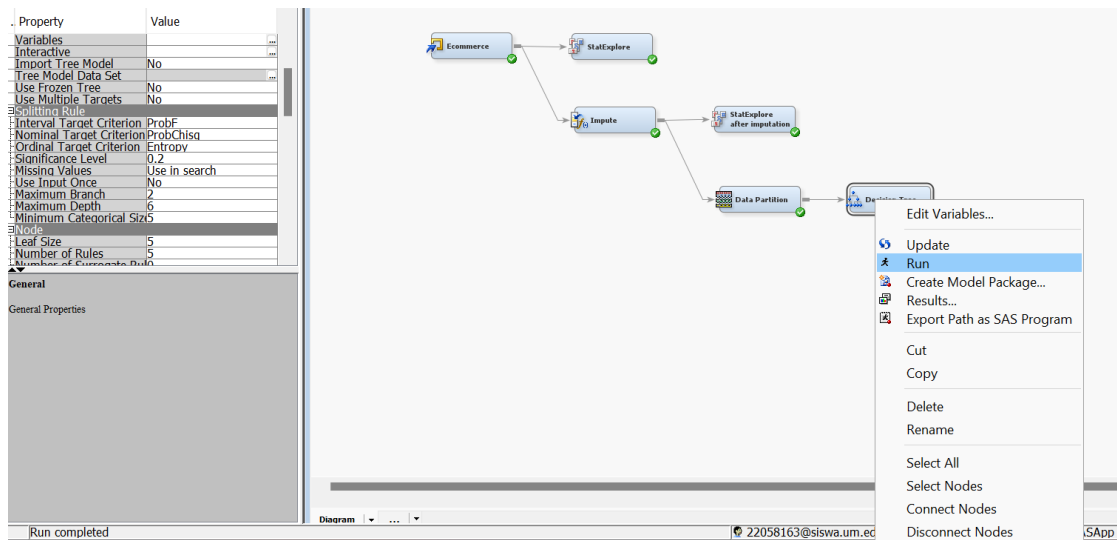
Data=VALIDATE

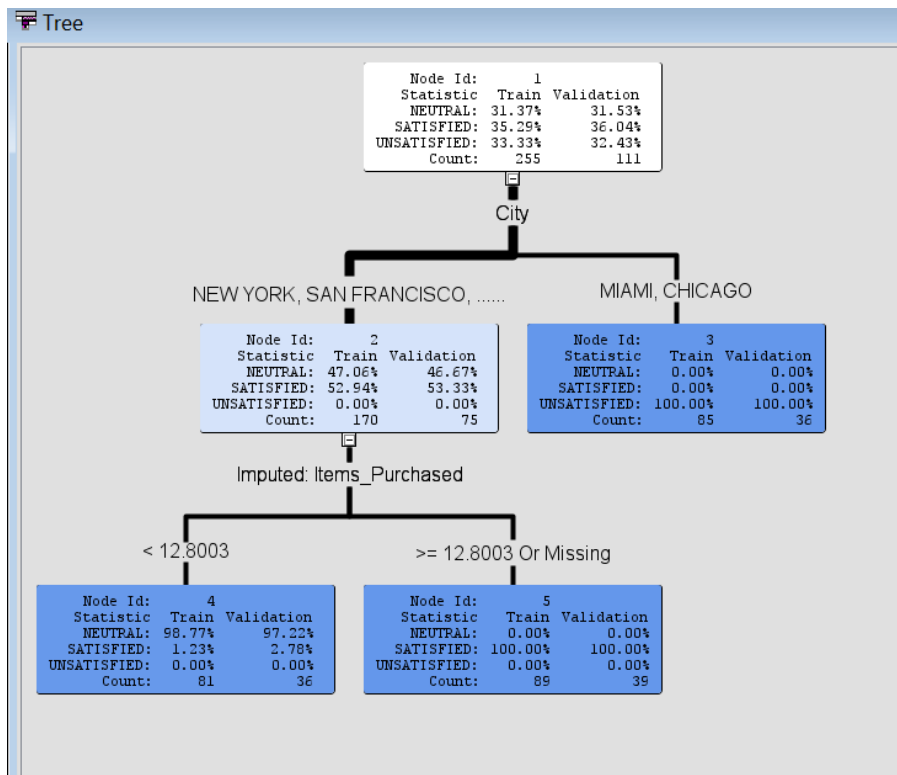
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Satisfaction_Level	.	Neutral	35	31.5315	Satisfaction_Level
Satisfaction_Level	.	Satisfied	40	36.0360	Satisfaction_Level
Satisfaction_Level	.	Unsatisfied	36	32.4324	Satisfaction_Level

In my dataset, each value of the target variable has almost the same frequency. “Neutral” values take up 31%, “Satisfied” values take up 35% whereas “Unsatisfied” values take up 32% of the target variable. The target variable must have the same amount of different values to ensure that it is balanced.

## Decision Tree Analysis

- Attach “Decision Tree” node into the diagram and configure the “Impute” node to the “decision Tree” node, then run it.





#### Event Classification Table

Data Role=TRAIN Target=Satisfaction\_Level Target Label=Satisfaction\_Level

False	True	False	True
Negative	Negative	Positive	Positive
0	170	0	85

Data Role=VALIDATE Target=Satisfaction\_Level Target Label=Satisfaction\_Level

False	True	False	True
Negative	Negative	Positive	Positive
0	75	0	36

Fit Statistics

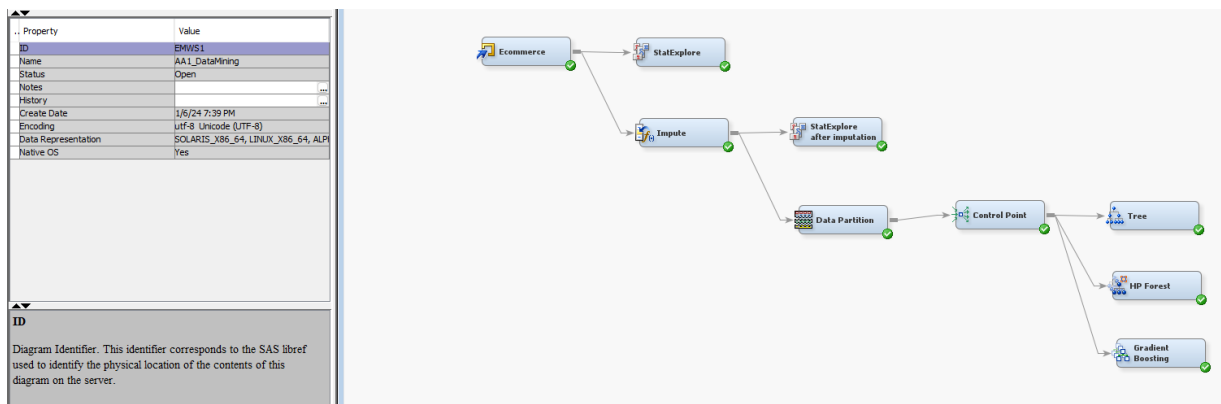
Target=Satisfaction\_Level Target Label=Satisfaction\_Level

Fit Statistics	Statistics Label	Train	Validation
_NOBS_	Sum of Frequencies	255.000	111.000
_MISC_	Misclassification Rate	0.004	0.009
_MAX_	Maximum Absolute Error	0.988	0.988
_SSE_	Sum of Squared Errors	1.975	1.962
_ASE_	Average Squared Error	0.003	0.006
_RASE_	Root Average Squared Error	0.051	0.077
_DIV_	Divisor for ASE	765.000	333.000
_DFT_	Total Degrees of Freedom	510.000	.

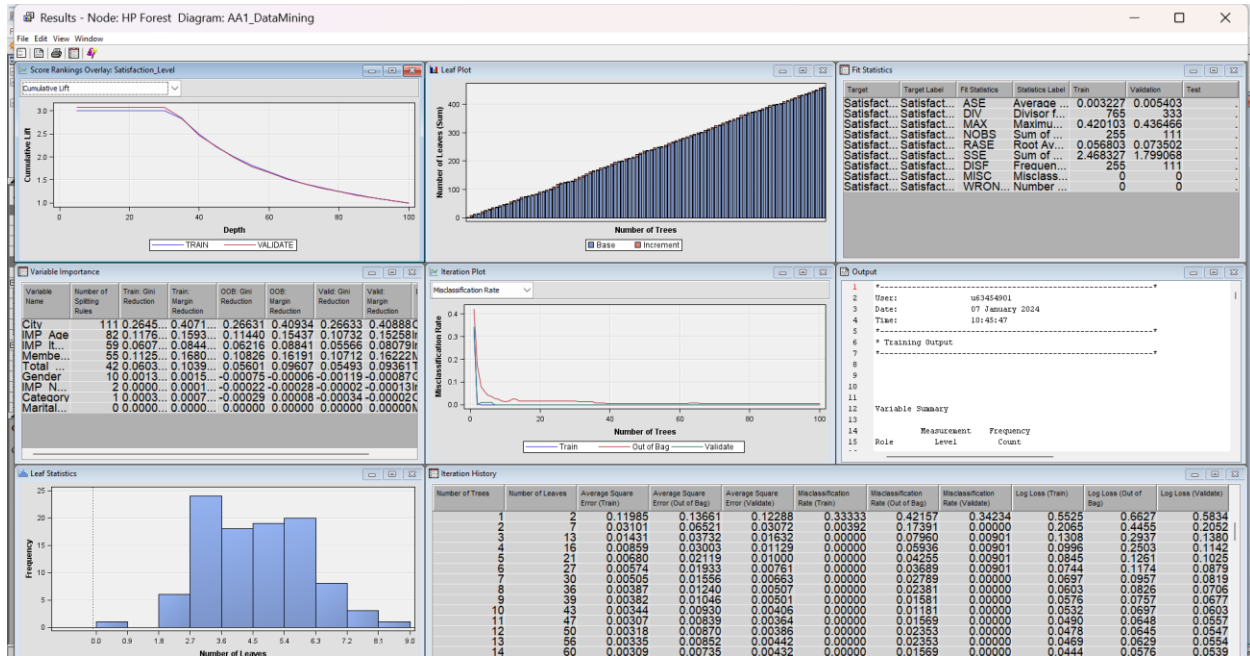
From the above results, it shows that the sum of squared errors in validation set is lower compared to training set whereas misclassification rate, average squared error and root average squared error in validation set is higher compared to training set. This shows that there is a slightly overfitting problem.

## Ensemble Methods

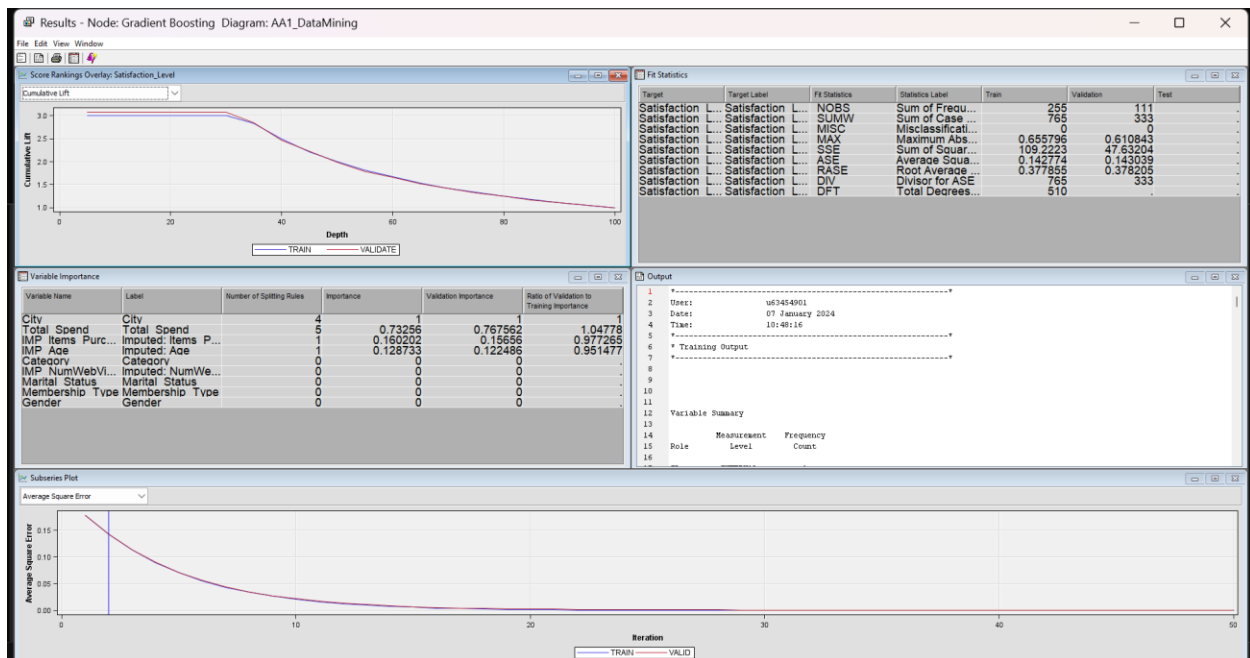
The ensemble methods used are random forest and gradient boosting. This is to compare which models will reveal the highest accuracy.



## Random Forests output

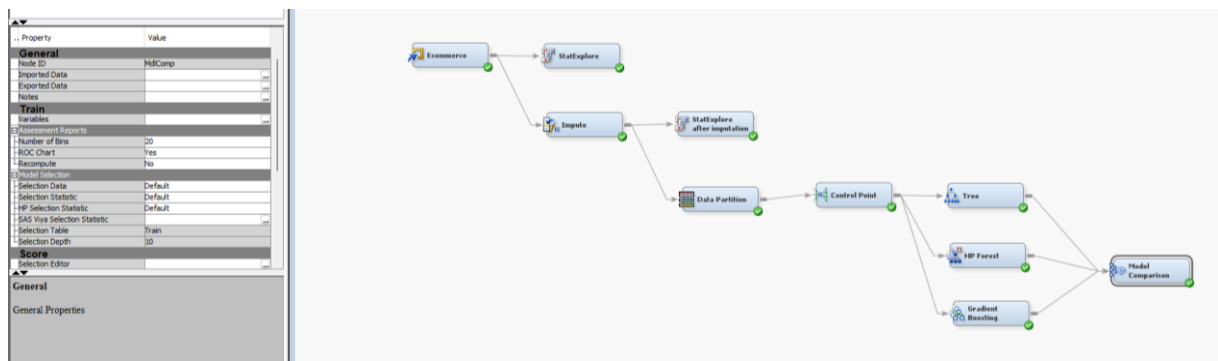


## Gradient boosting output





Model Comparison



Fit Statistics

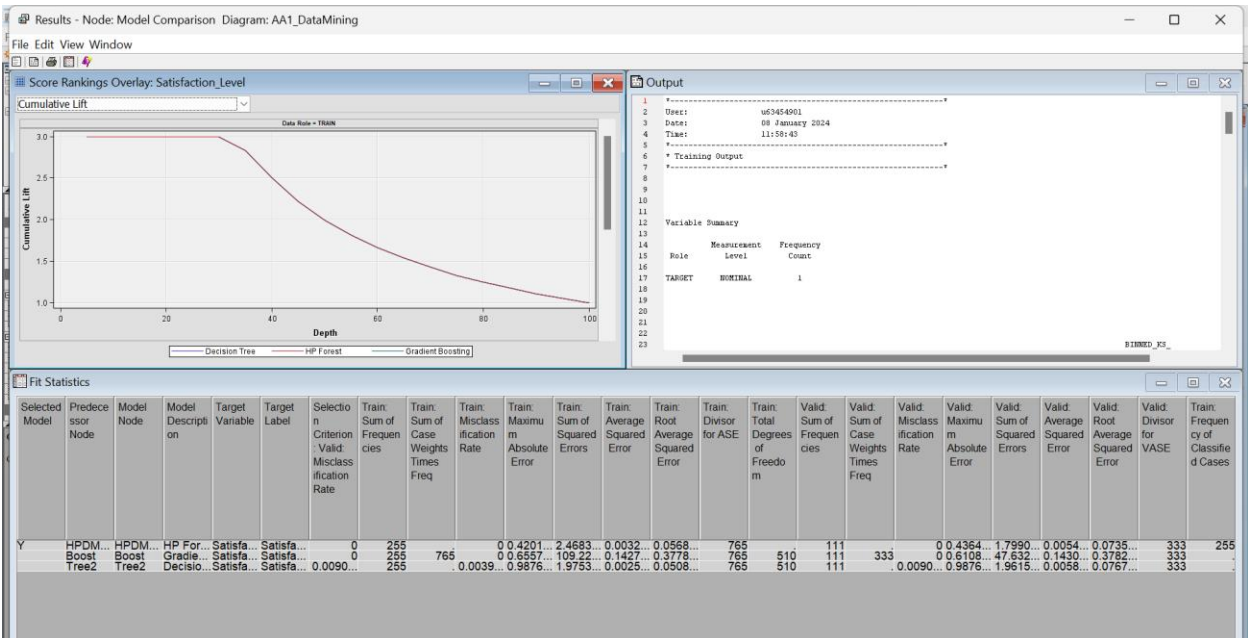
Model Selection based on Valid: Misclassification Rate (\_VMISC\_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	HPDMForest	HP Forest	.000000000	0.00323	.000000000	0.00540
	Boost	Gradient Boosting	.000000000	0.14277	.000000000	0.14304
	Tree2	Decision Tree (2)	.009009009	0.00258	.003921569	0.00589

Event Classification Table

Model Selection based on Valid: Misclassification Rate (\_VMISC\_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Boost	Gradient Boosting	TRAIN	Satisfaction_Level	Satisfaction_Level	0	170	0	85
Boost	Gradient Boosting	VALIDATE	Satisfaction_Level	Satisfaction_Level	0	75	0	36
HPDMForest	HP Forest	TRAIN	Satisfaction_Level	Satisfaction_Level	0	170	0	85
HPDMForest	HP Forest	VALIDATE	Satisfaction_Level	Satisfaction_Level	0	75	0	36
Tree2	Decision Tree (2)	TRAIN	Satisfaction_Level	Satisfaction_Level	0	170	0	85
Tree2	Decision Tree (2)	VALIDATE	Satisfaction_Level	Satisfaction_Level	0	75	0	36



From the above images, it seems that Decision tree shows the highest accuracy as it has the lowest average squared error and misclassification rate in both training and validation set. Though it displays an

overfitting behaviour yet, it is the most suitable model to be used as the other models show overfitting behaviour as well as providing a slightly higher misclassification rate, average squared error and root average squared error.