Background

In this assessment, I adapt a dataset named "E-Commerce Customer." The assessment will apply three tools: Talend Preparation, Talend Data Integration, and SAS Enterprise Miner. These three tools will be used to carry out tasks such as data import and preprocessing, conduct decision tree analysis, and apply various ensemble methods.

Dataset description

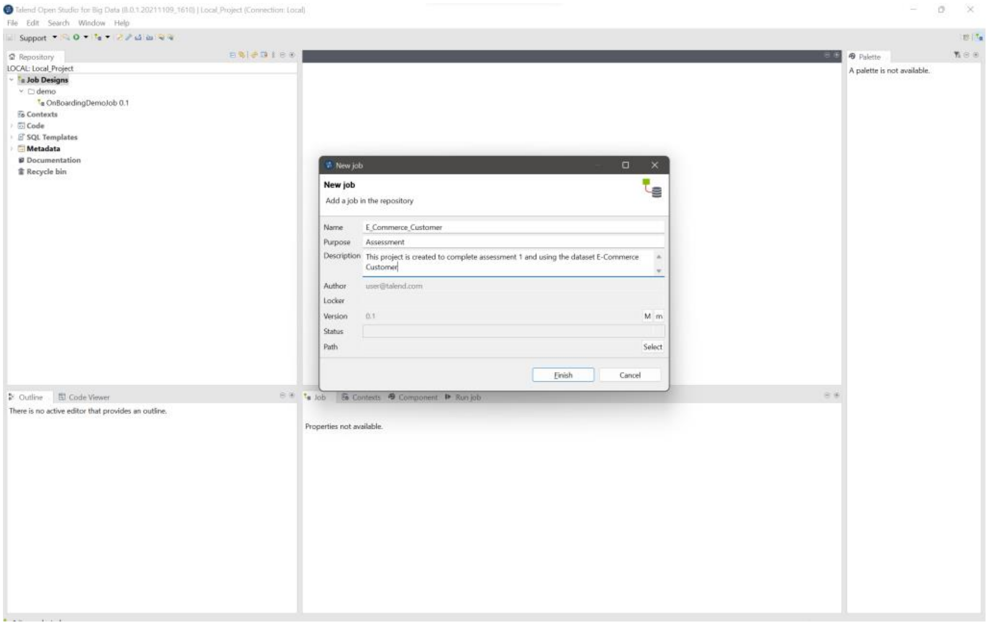
The dataset "E-Commerce Customer" records customer transactions on an e-commerce platform, including various customer attributes and purchase history in 2023. It has 12 attributes that include Customer ID, Age, Gender, Location, Member ship Level, Total Purchases, Total Spent, Favorite Category, Last Purchase Date, Occupation, Website Visits Frequency, and Churn. The dataset comprises 570 rows.

Variable	Data Type	Description
Customer ID	Numeric	Unique identifier for each customer.
Age	Numeric	Age of the customer.
Gender	String	Gender of the customer.
Location	String	State of the customer base in Malaysia
Membership Level	String	Membership level label in Bronze, Silver, Gold, Platinum.
Total Purchases	Numeric	Total number of purchases made by the customer in a year.
Total Spent	Numeric	Total amount spent by the customer in a year.
Favorite Category	String	The category in which the customer most frequently shops labelled in Electronics, Clothing, Home Goods.
Last Purchase Date	Date	The date of the last purchase.
Occupation	String	Customer's occupation.
Website Visits Frequency	String	Frequency of customer visits the website.
Churn	Numeric	Indicates whether the customer has stopped purchasing (1 for churned, 0 for active).

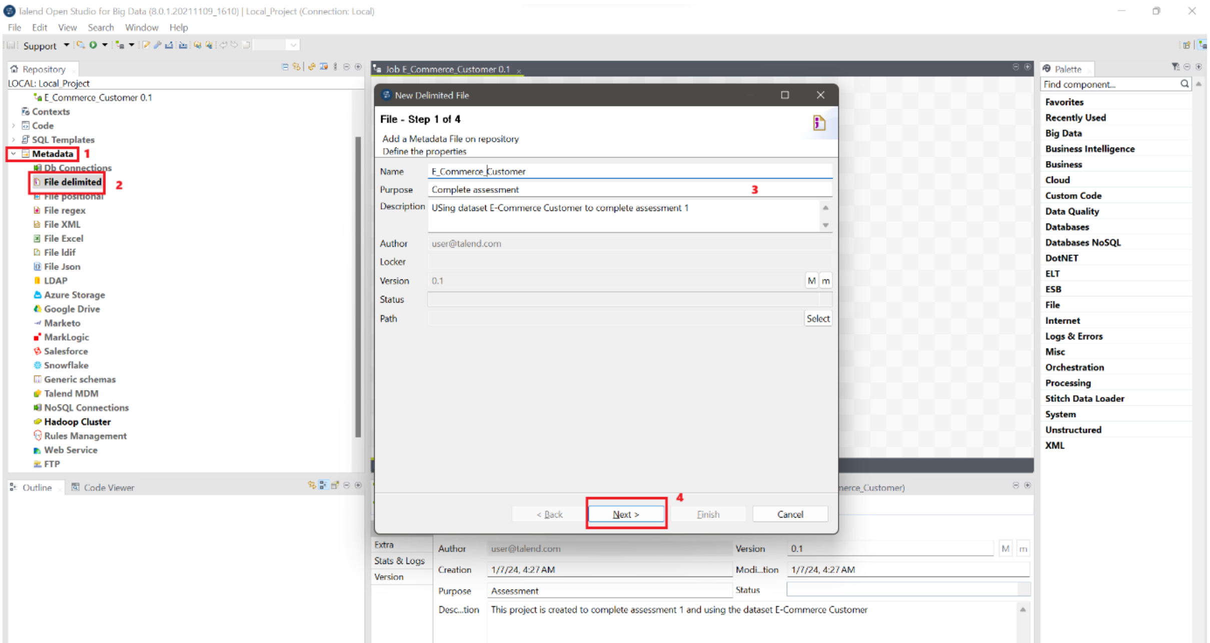
Steps

The assessment start with the help of Talend data Integration for data preprocessing.

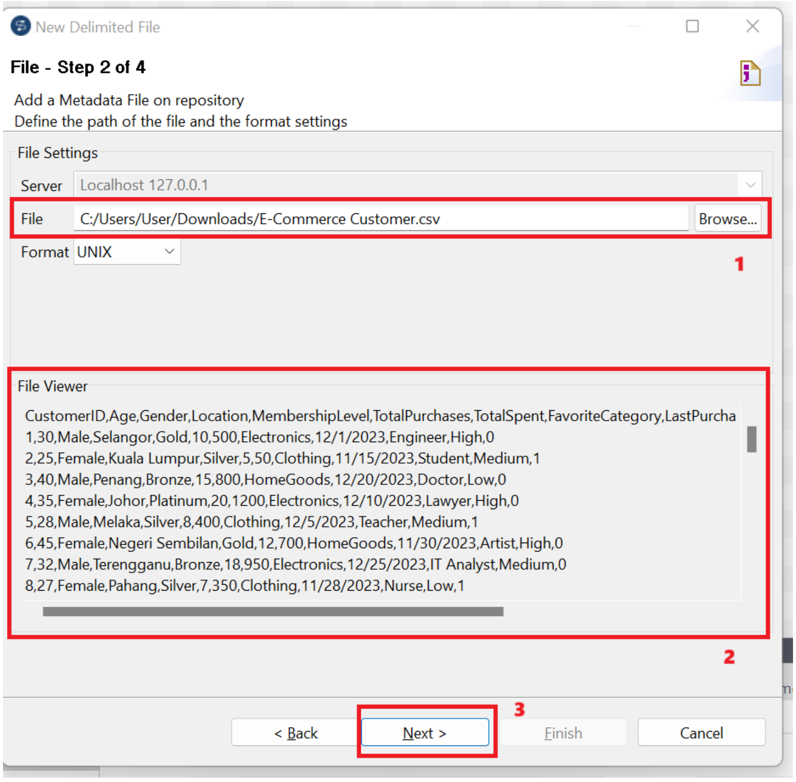
a. To start up, I created a project in Talend data integration.



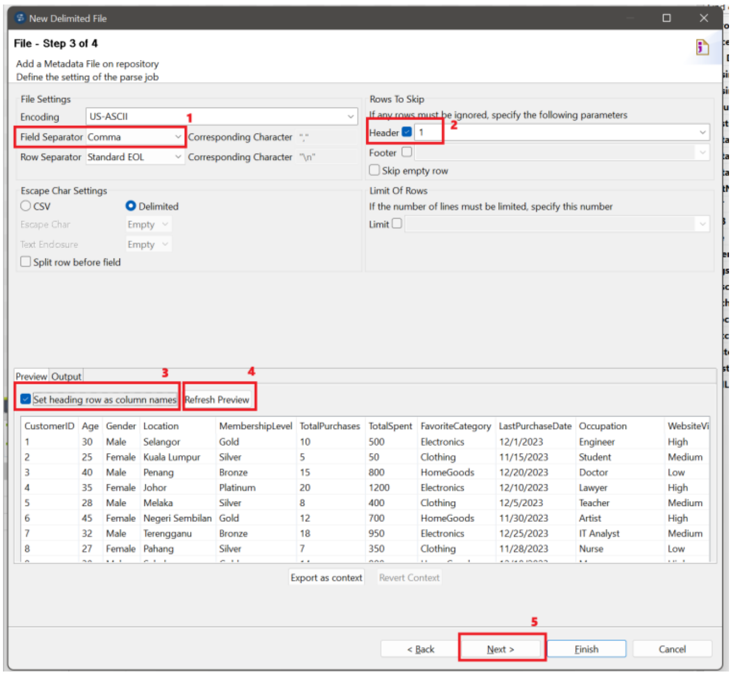
b. To extract data from CSV file, I went to Metadata, then right click File delimited. It will pop out a new window for me to fill up the 'Name', 'Purpose' and 'Description'. After that click 'Next'.



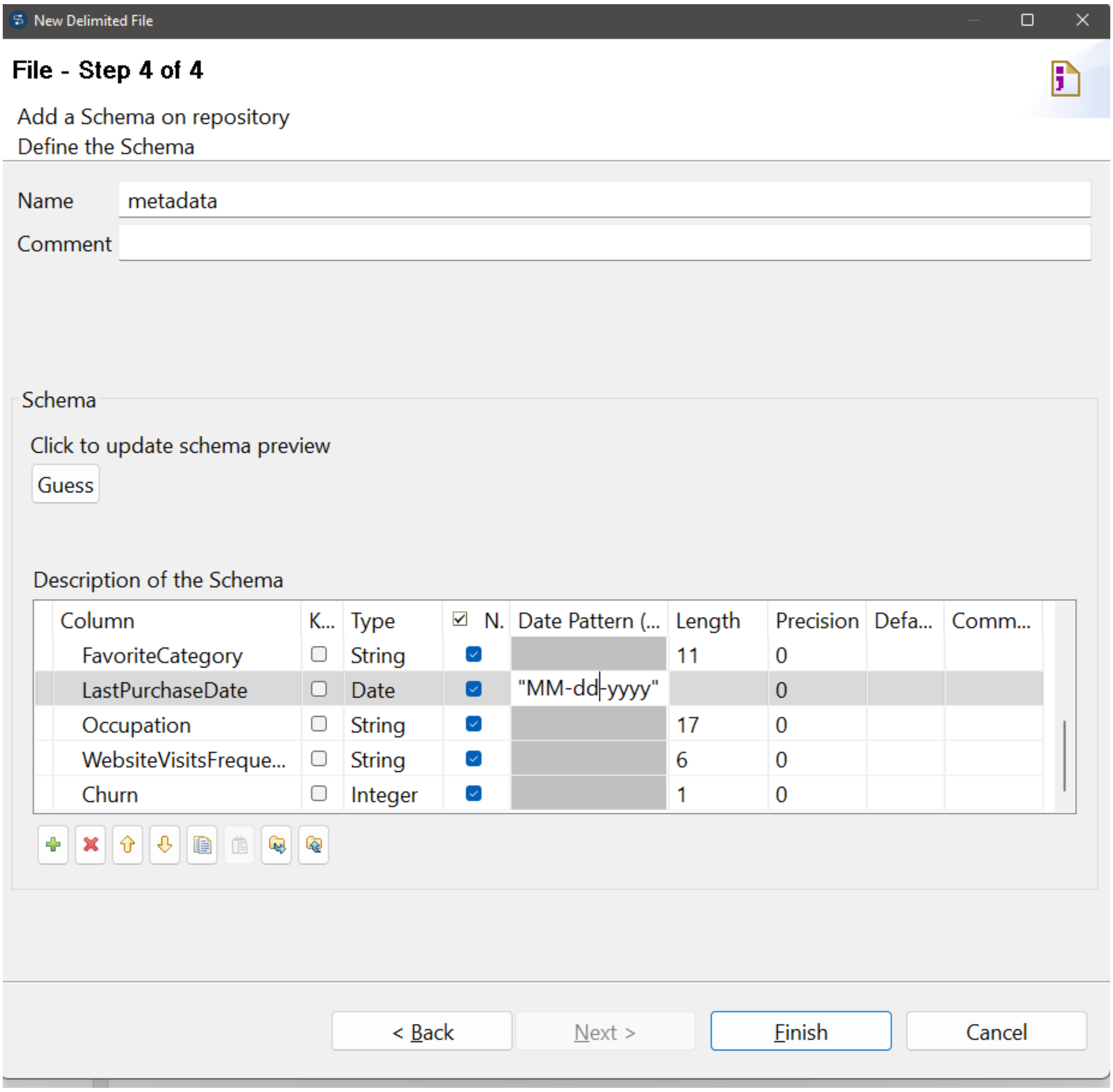
c. I browse for my data set. The File viewer give me a preview of my import data. Then click 'Next'



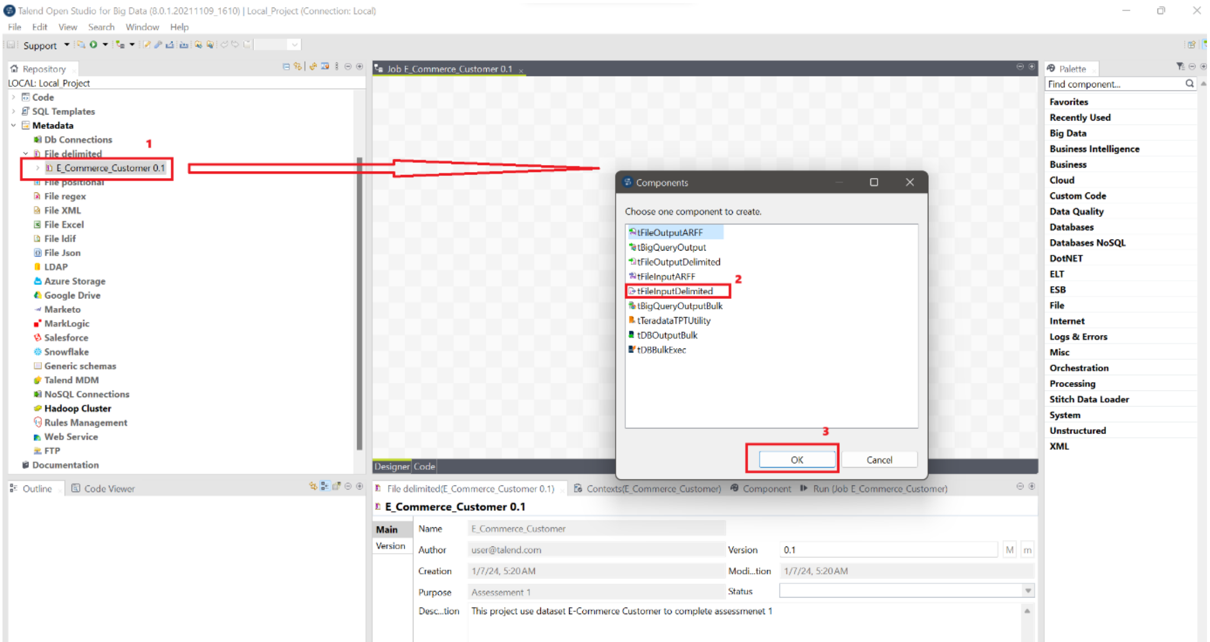
d. In next page, I change the 'Field Separator' to 'comma', tick 'header' then insert 1, and tick 'set heading row as column names'. Lastly, click 'Refresh Preview' to see the preview of my data. Once confirm I click 'Next'.



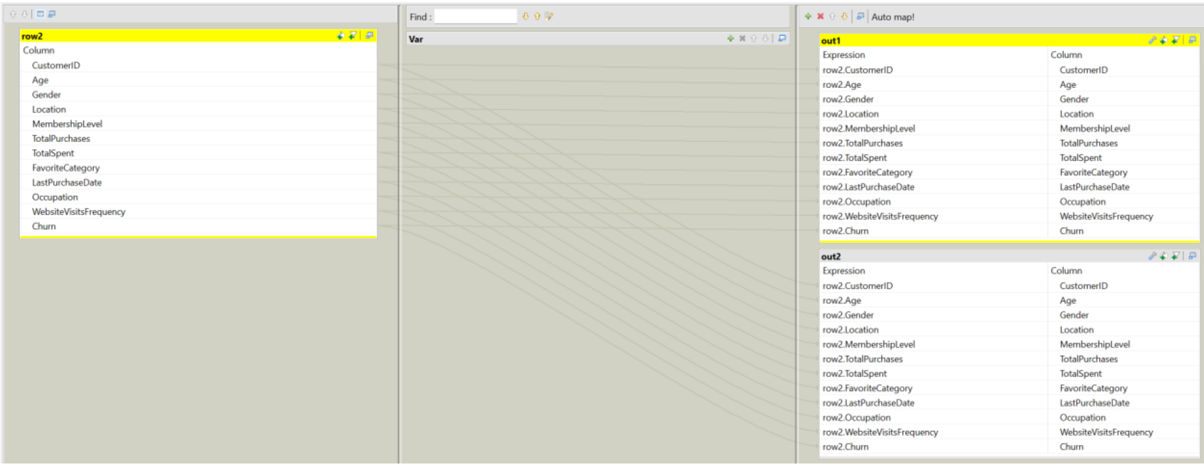
e. In the next page I change the data type for attribute LastPurchaseDate to date and adjust the pattern. Once done I proceed click finish.



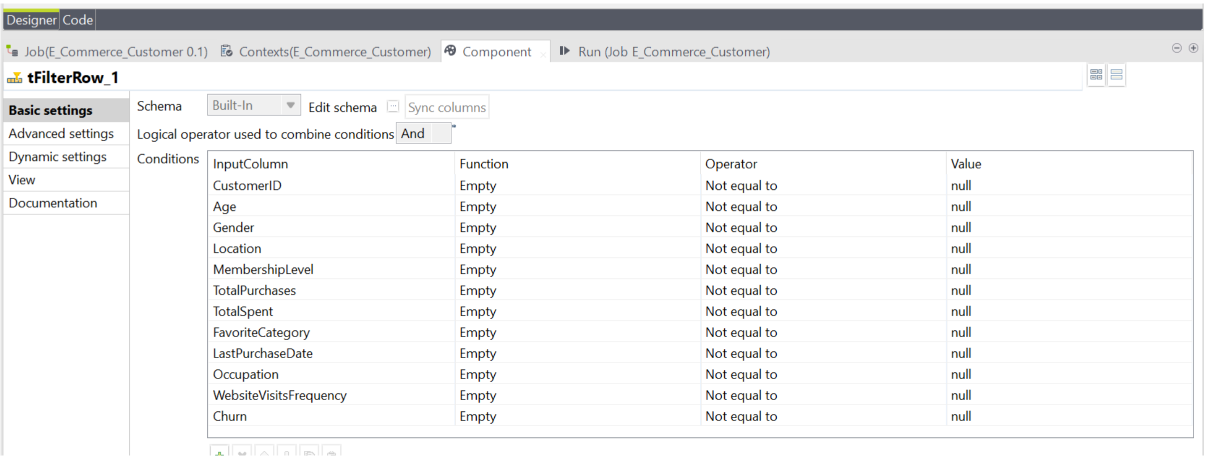
f. Drag the delimited file to the canvas, a window pop out, I select ‘tFileInputDelimited’ and click OK.



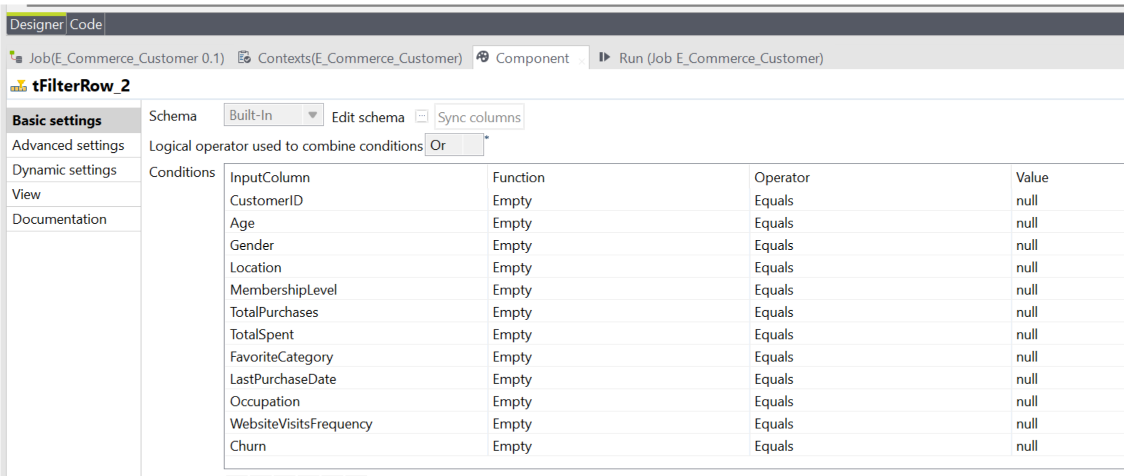
g. Then I add on component tLogRow to extract data and link it to tMap. In tmap I generate two output section for filter data in two scenario.



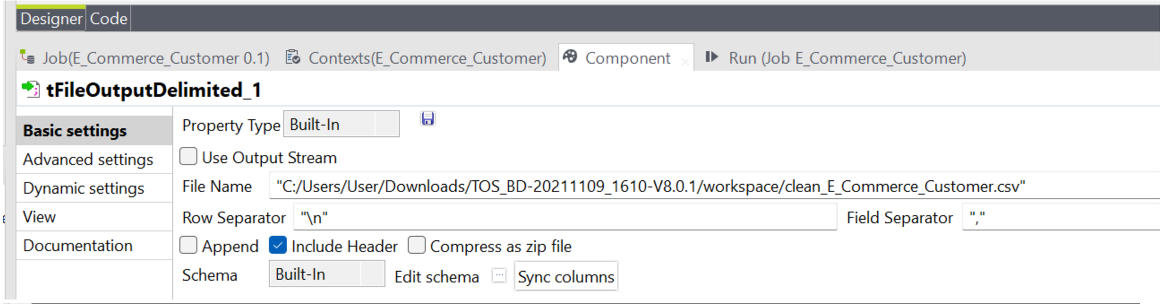
h. In filter 1 I set filter for each attribute for getting dataset with no missing value.

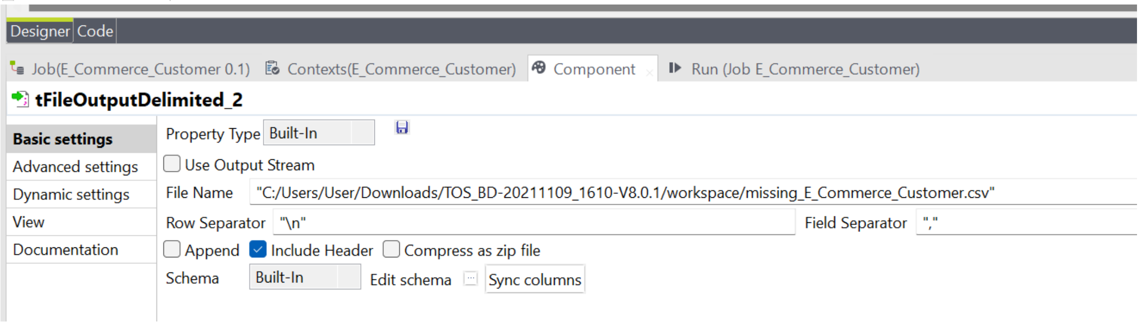


i. In filter 2 I set filter for each attribute for getting dataset with missing value.

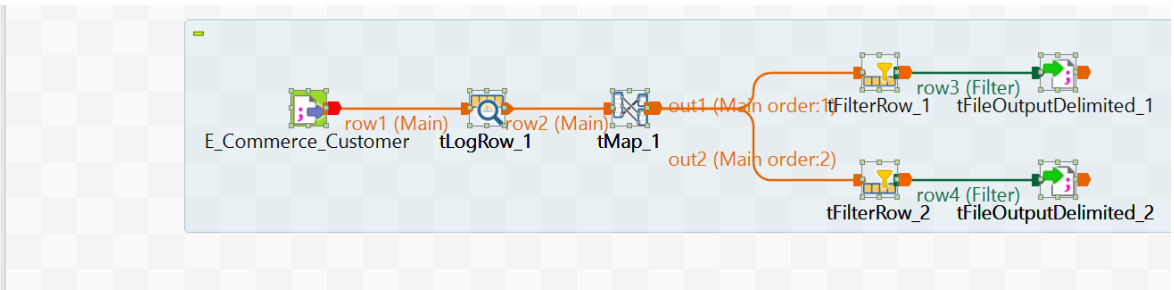


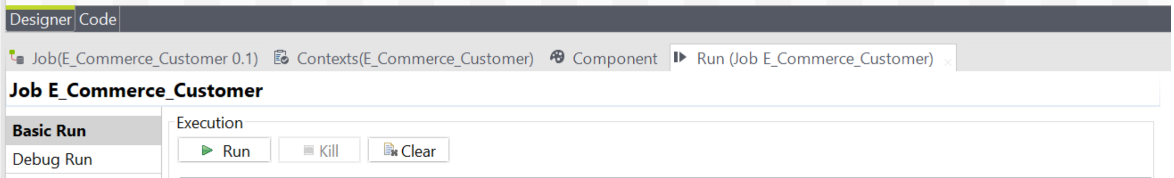
j. After that I link both filter to tFileOutputDelimited respectively. The file name and setting for each output file as below.



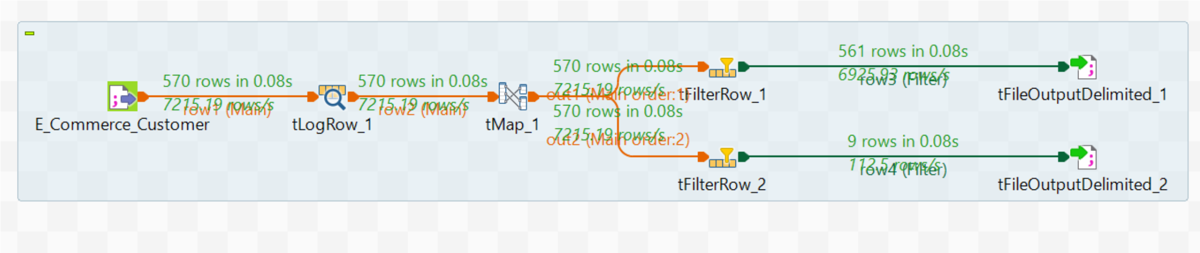


k. The overall design as below. Once confirm I click 'Run'



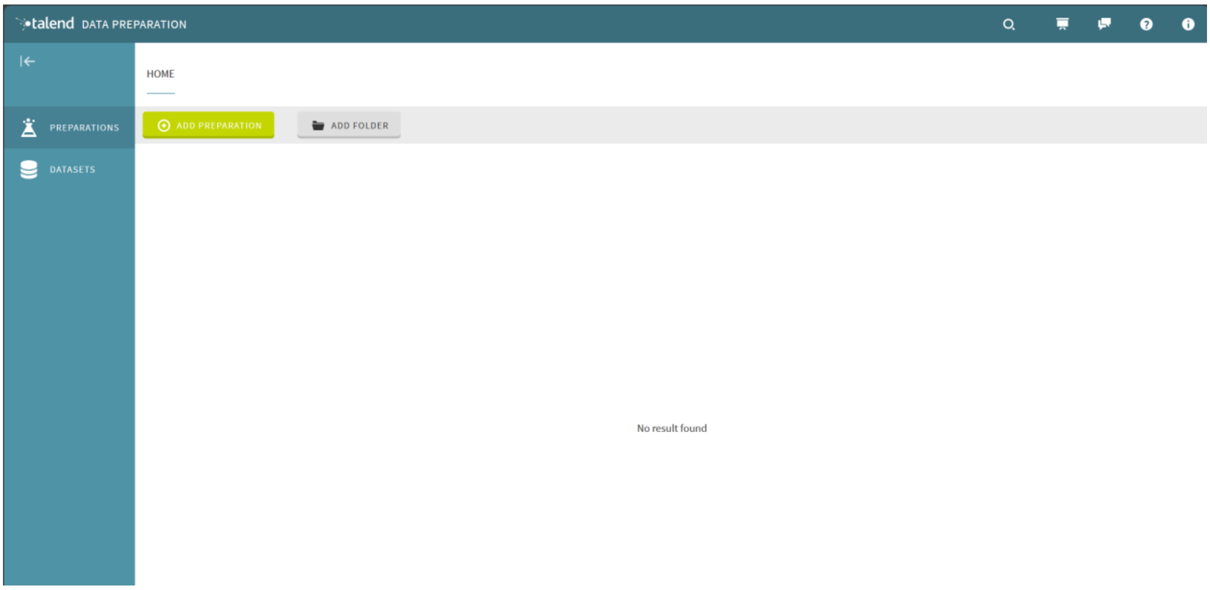


I. The outcome as below. From here I found there 9 three rows being filter due to missing data.



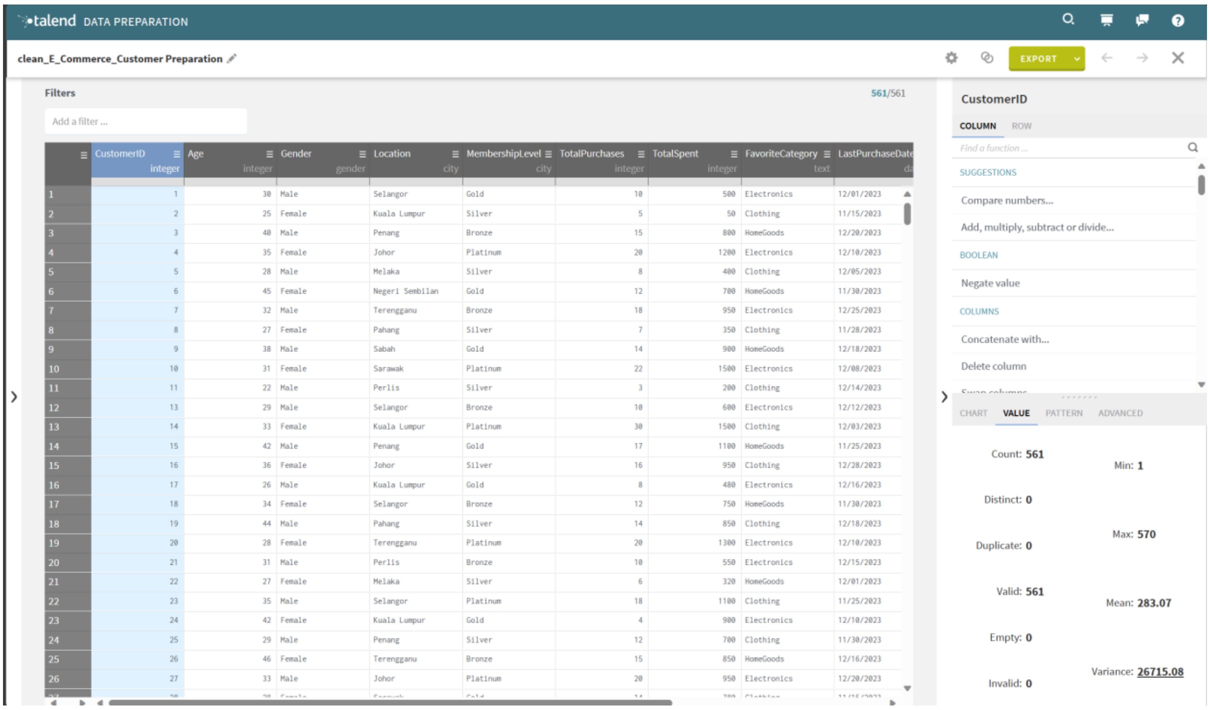
The assessment continue with the Talend preparation for deeper data preprocessing and understanding.

a) Import Data



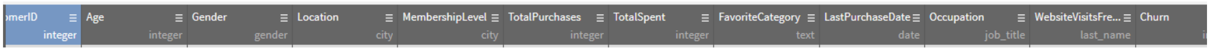
b) After successfully import data prio clean in Talend data integration , I start study the data

- From the value section, can see that there are 561 counts which means I have 561 rows left after the first round cleaning.



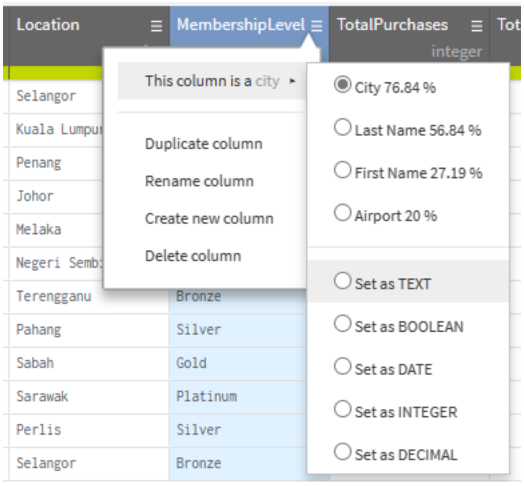
c) Clean the data

- Notice there are some attributes wrongly label and missing value (white box)



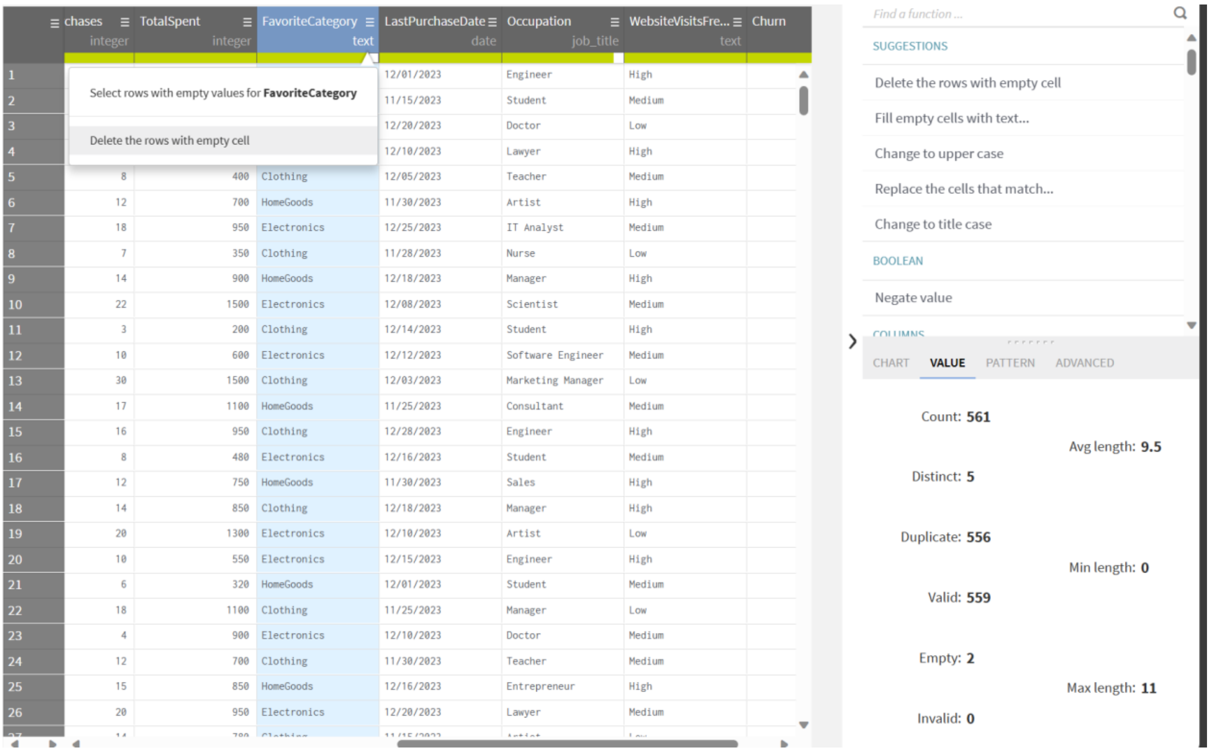
1. Change the label

- For example change the label for attribute Member ship level from city to text and the attribute website visit frequency level from last_name to text



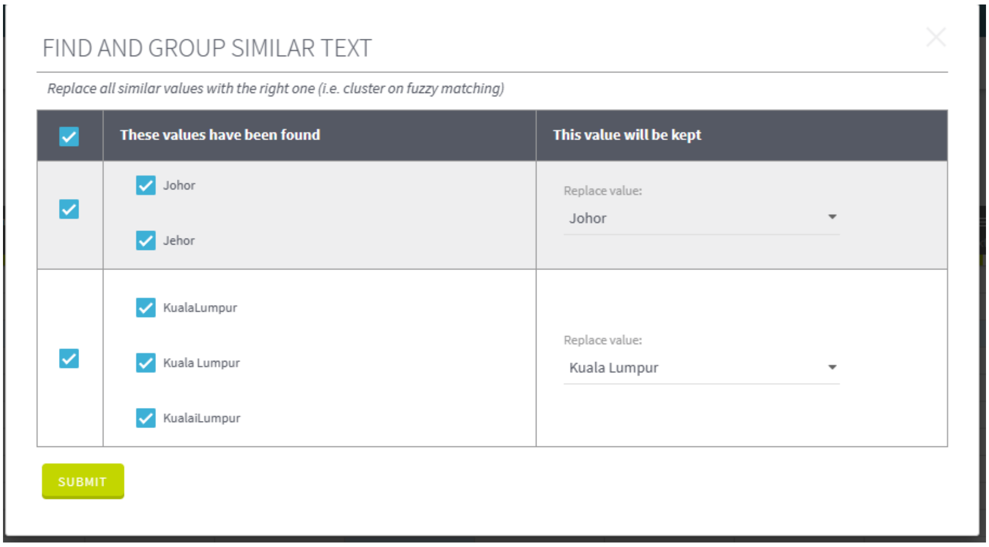
2. Remove missing data

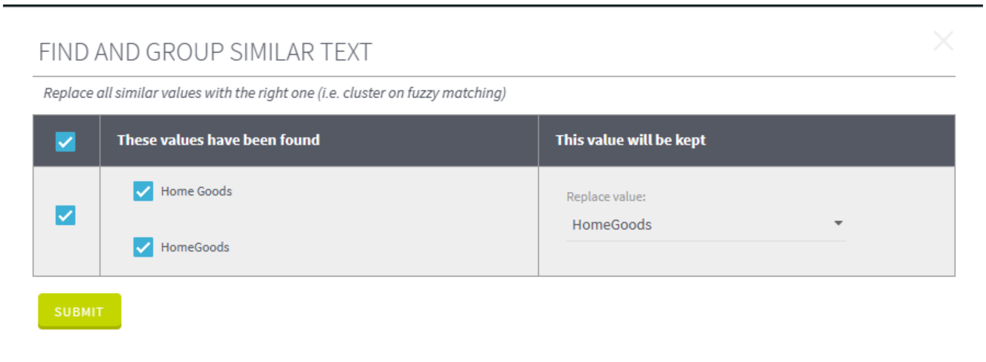
- Notice there are white box at the corner of certain attribute. In the values section it show that out of 561 instance there are 2 empty cell in Favourite category and 6 missing in job_title. I decide to remove the row that missing data as my data consider big therefore deleted few rows will not effect my analysis.



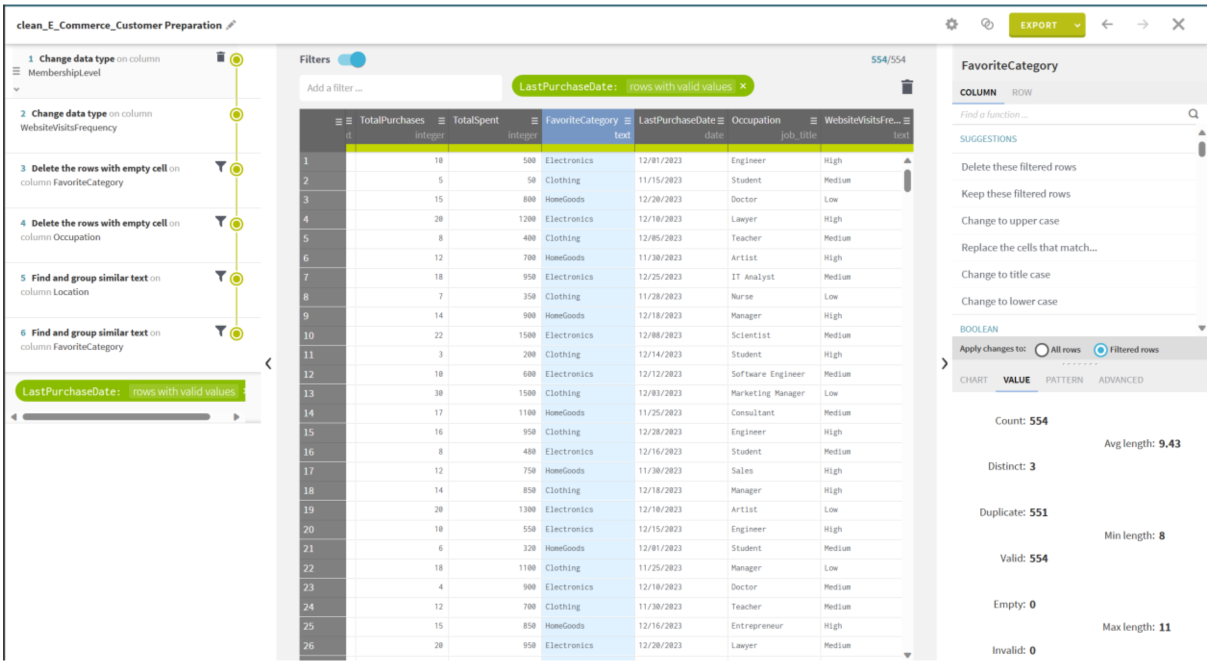
3. Find and group similar text data

- Notice there are data with similar text that might be due to typo during data entry. This were found in the attribute of location and Favourite category.

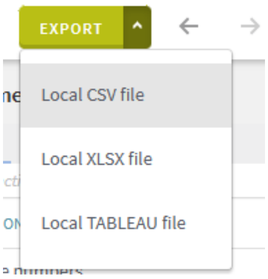




d) After done cleaning the data we can see that the dataset had been reduce to 554 rows.

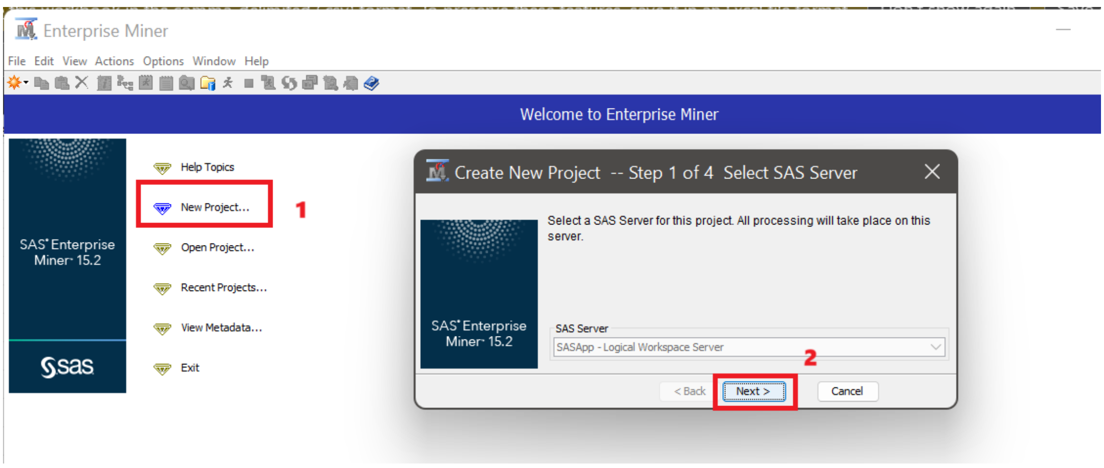


e) Once confirm, I export the dataset in CSV file.

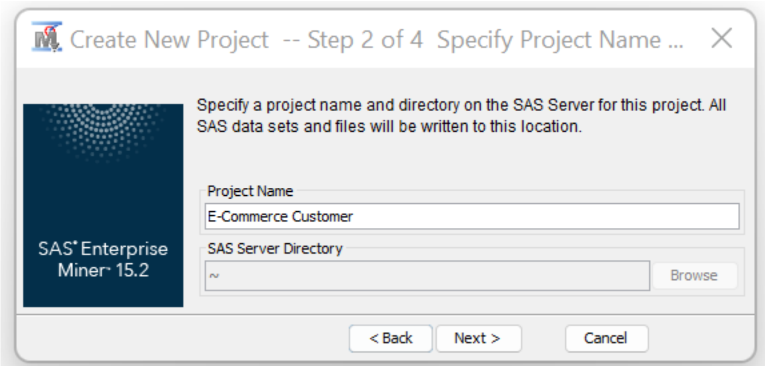


The project continued with tools SAS Enterprise Miner.

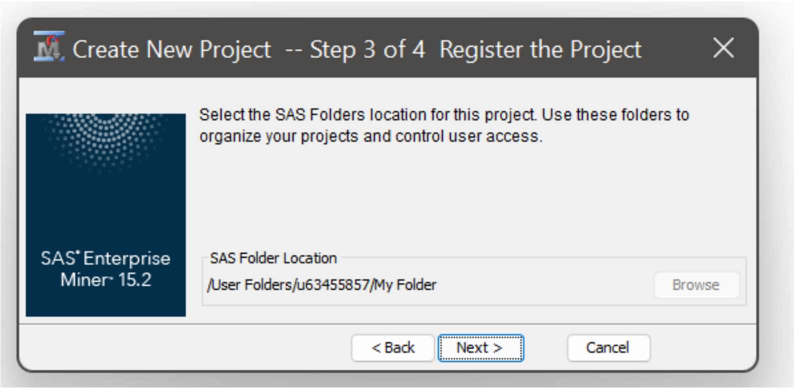
a. First, I create New project, a window will pop out and I click next.



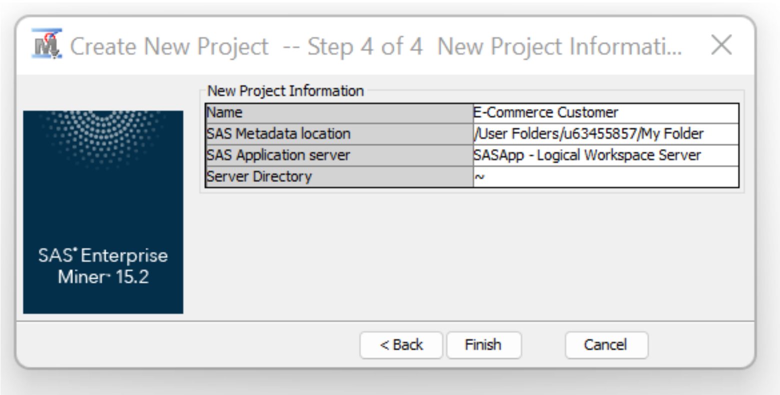
b. Insert Project name and click next.



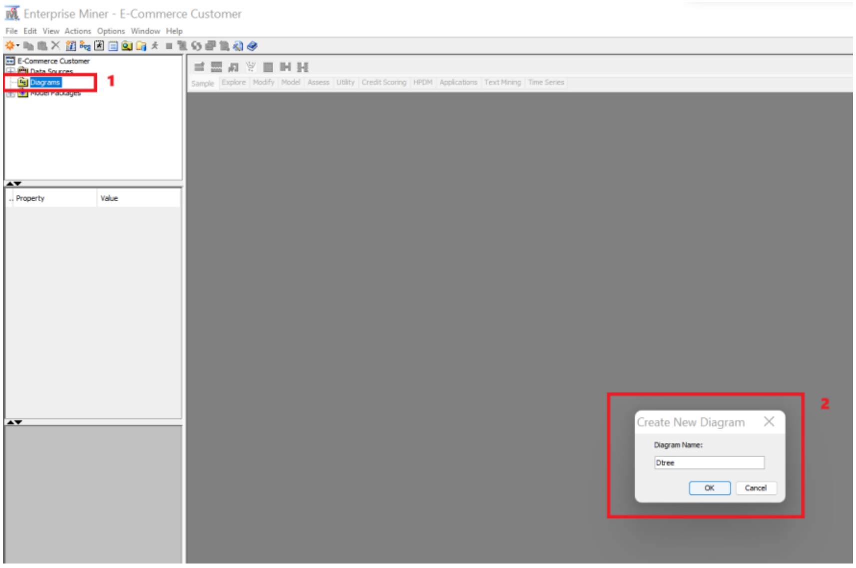
c. Click next 1 more time.



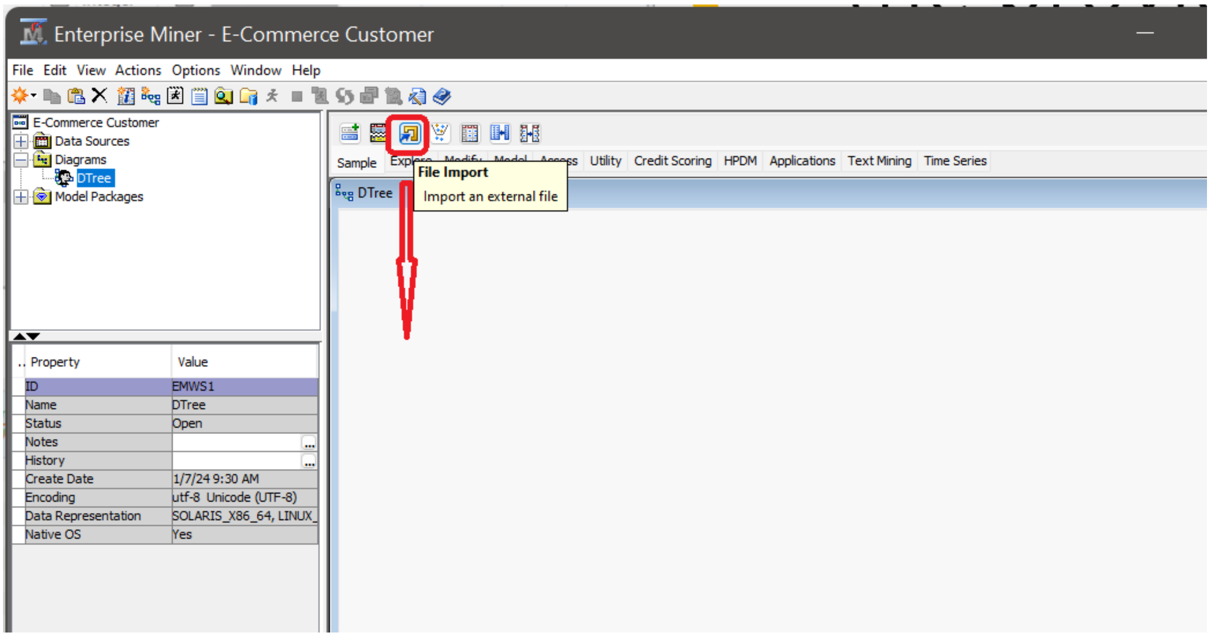
d. Then click finish.



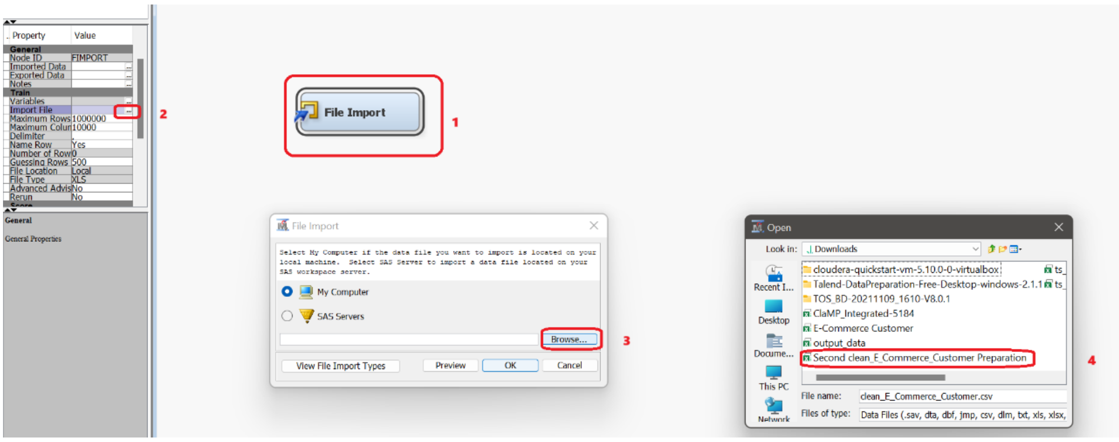
e. Through right click the ‘Diagrams’, I click create diagram and name it ‘DTree’.



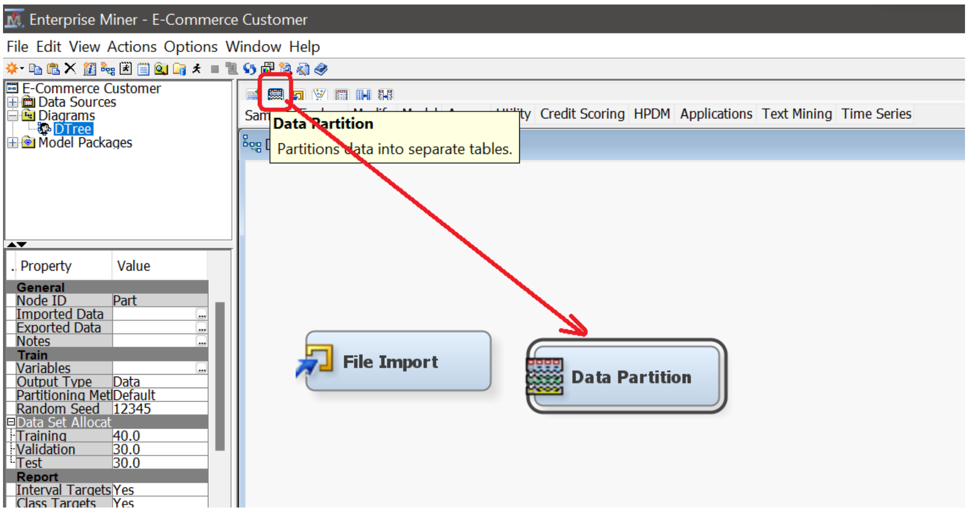
f. Drag the File import node down.



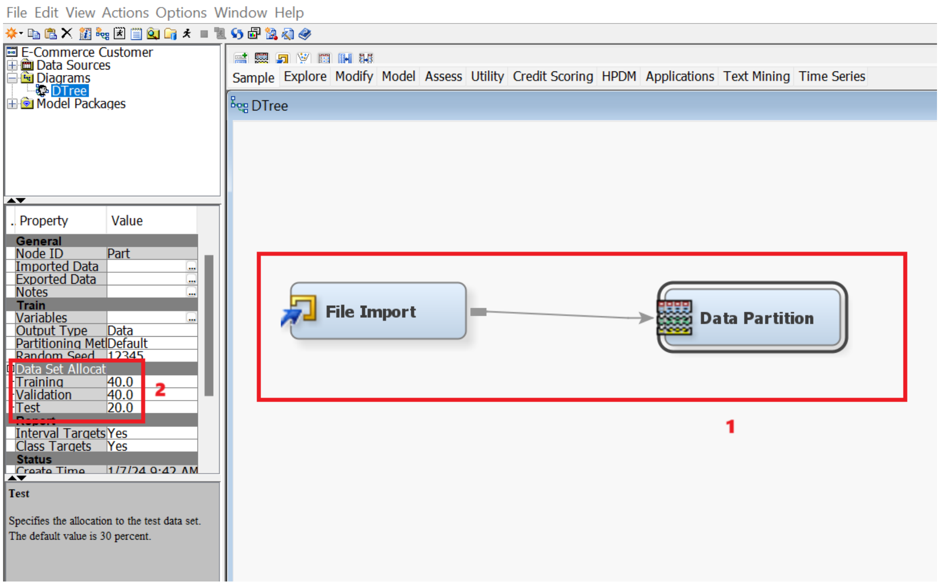
g. Click the ‘File Import’ node, then click the ‘…’ , click ‘Browse’ and select data to be use.



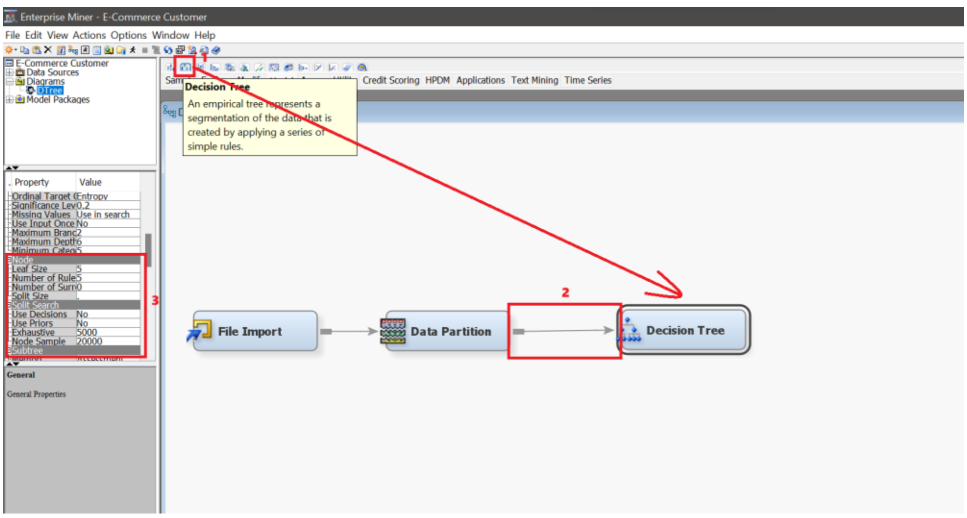
h. Drag the ‘Data Partition’ node down for splitting data into training, validating and testing.



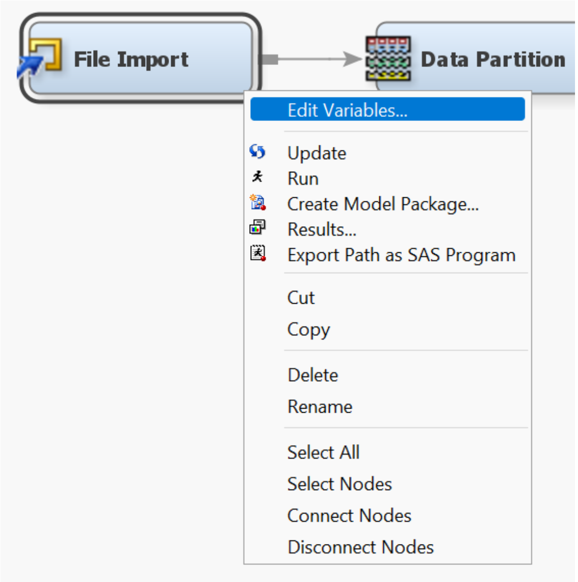
i. Link both nodes together. Change the Dataset Allocation to 4:4:2 for Training, Validation and Test.



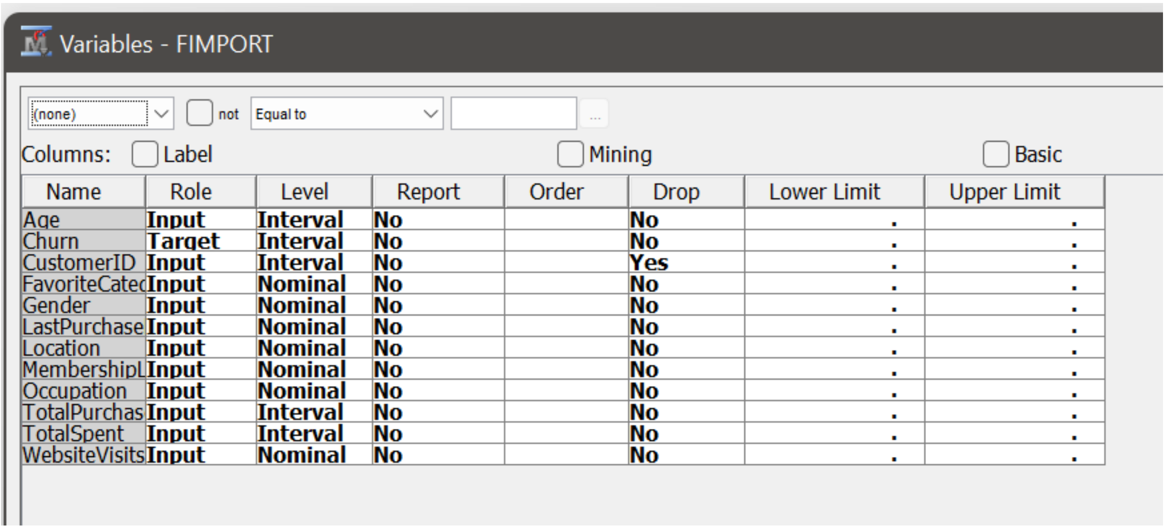
j. Drag the Decision Tress node down and connect to Data partition. At the right corner, I adjust the maximum branch, Depth and categorical size.



k. Right Click File Import node and select edit variable.

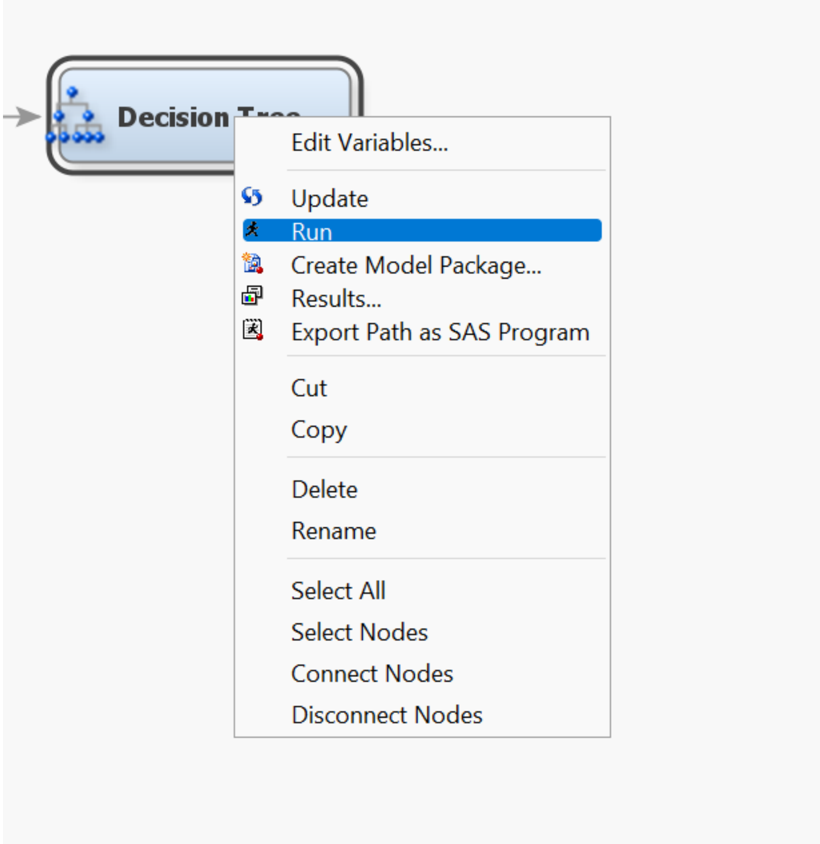


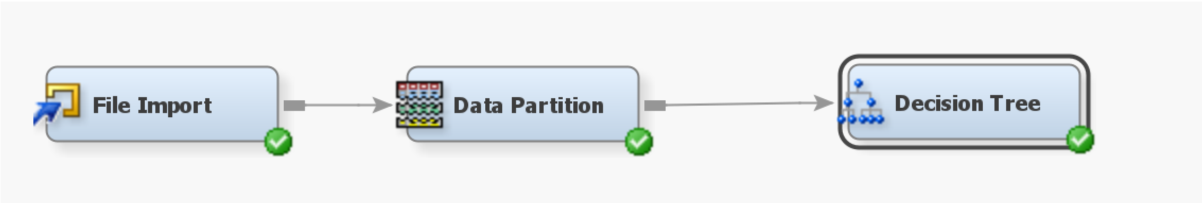
l. Below is my adjustment for my variables. Churn will be my target as I want to predict whether a customer will churn or not. I drop customer id as it is not important in this project. Once confirm, I Click ok.



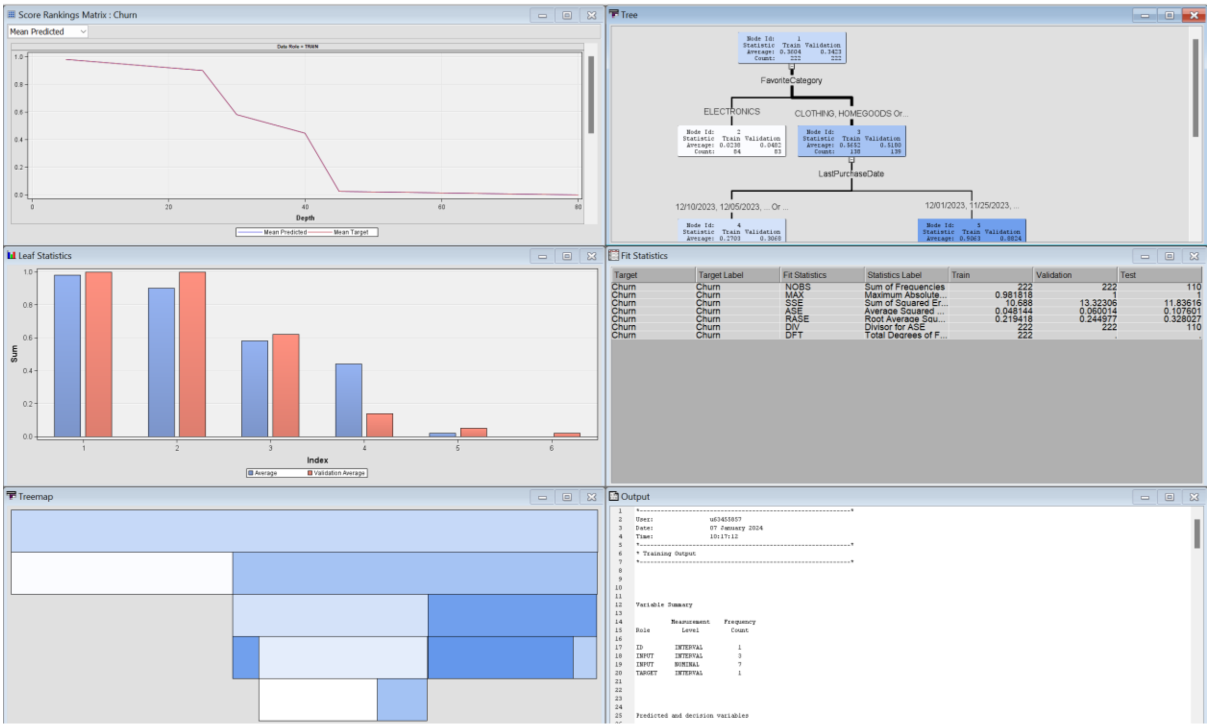
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Churn	Target	Interval	No		No	.	.
CustomerID	Input	Interval	No		Yes	.	.
FavoriteCateg	Input	Nominal	No		No	.	.
Gender	Input	Nominal	No		No	.	.
LastPurchase	Input	Nominal	No		No	.	.
Location	Input	Nominal	No		No	.	.
MembershipL	Input	Nominal	No		No	.	.
Occupation	Input	Nominal	No		No	.	.
TotalPurchas	Input	Interval	No		No	.	.
TotalSpent	Input	Interval	No		No	.	.
WebsiteVisits	Input	Nominal	No		No	.	.

m. Right click Decision Tree node and click Run.

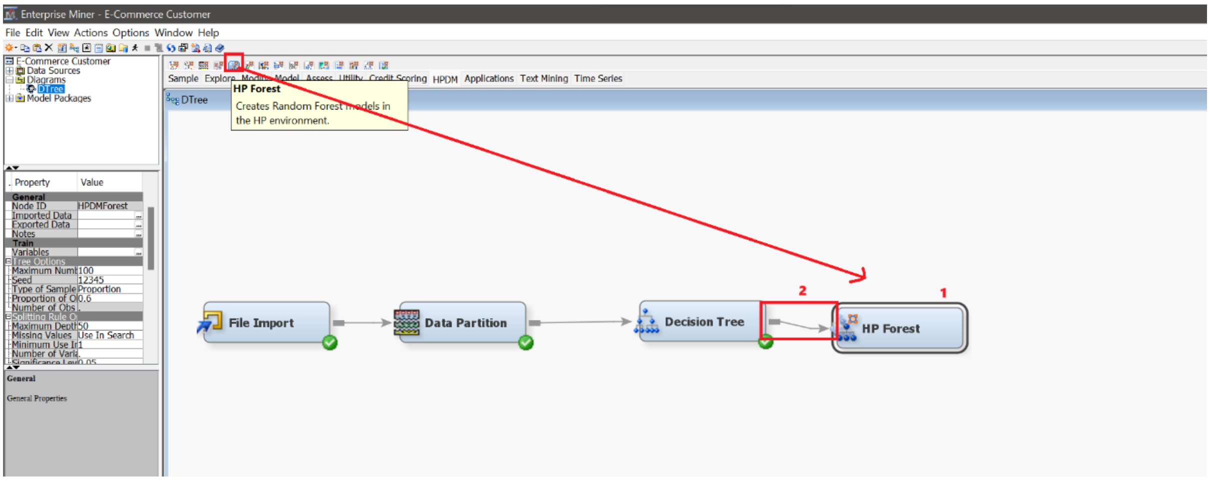




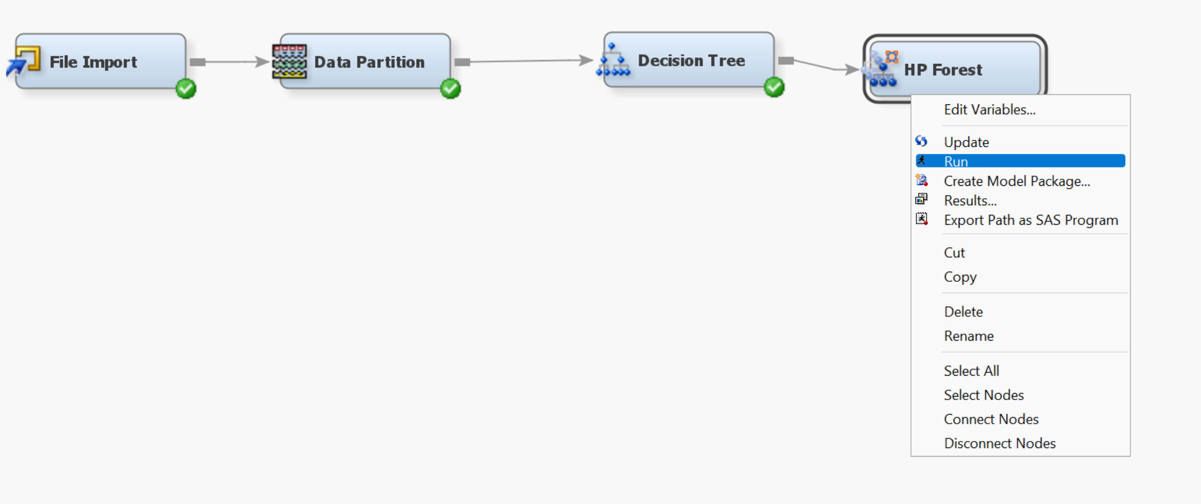
n. Click result to see the outcome. From The decision tree it seems to have identified "FavoriteCategory" as the most important variable for predicting churn, followed by "LastPurchaseDate," "Age," and "Occupation."



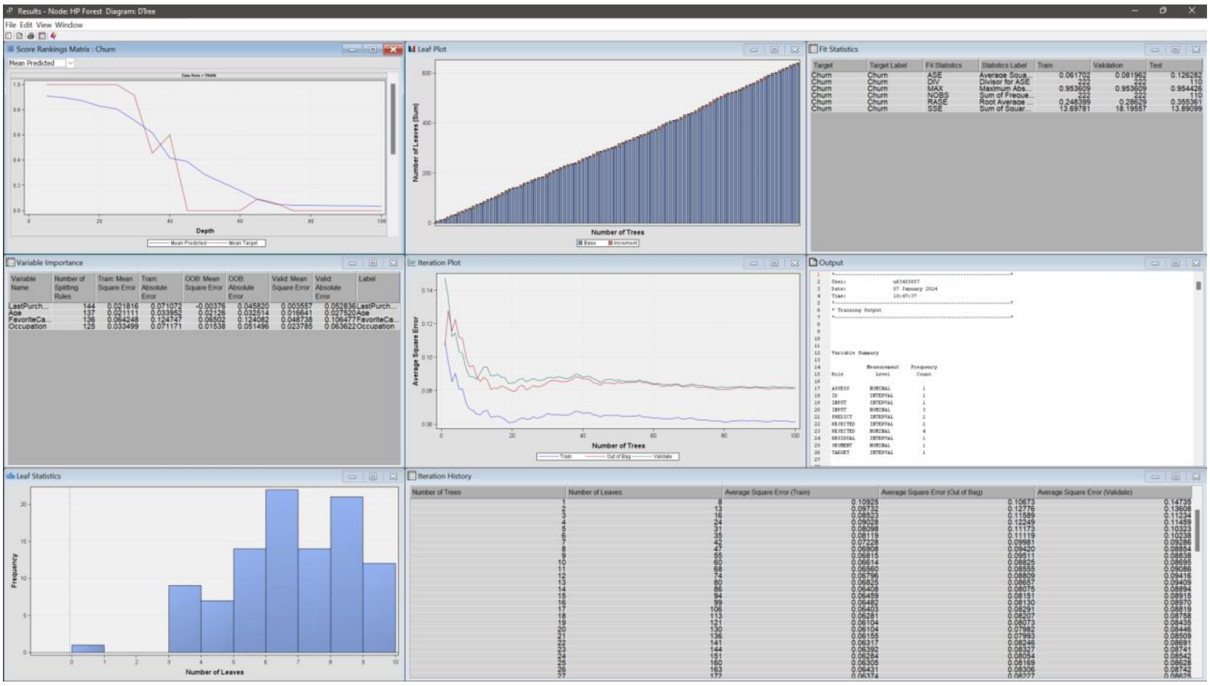
o. To apply Bagging Random Forest algorithm involves utilizing the corresponding nodes in the diagram. I drag 'HP Forest' which stand for Random forest in Sas and connect it to decision tree model node. This mean that the Random Forest will be built based on the individual decision trees I created earlier.



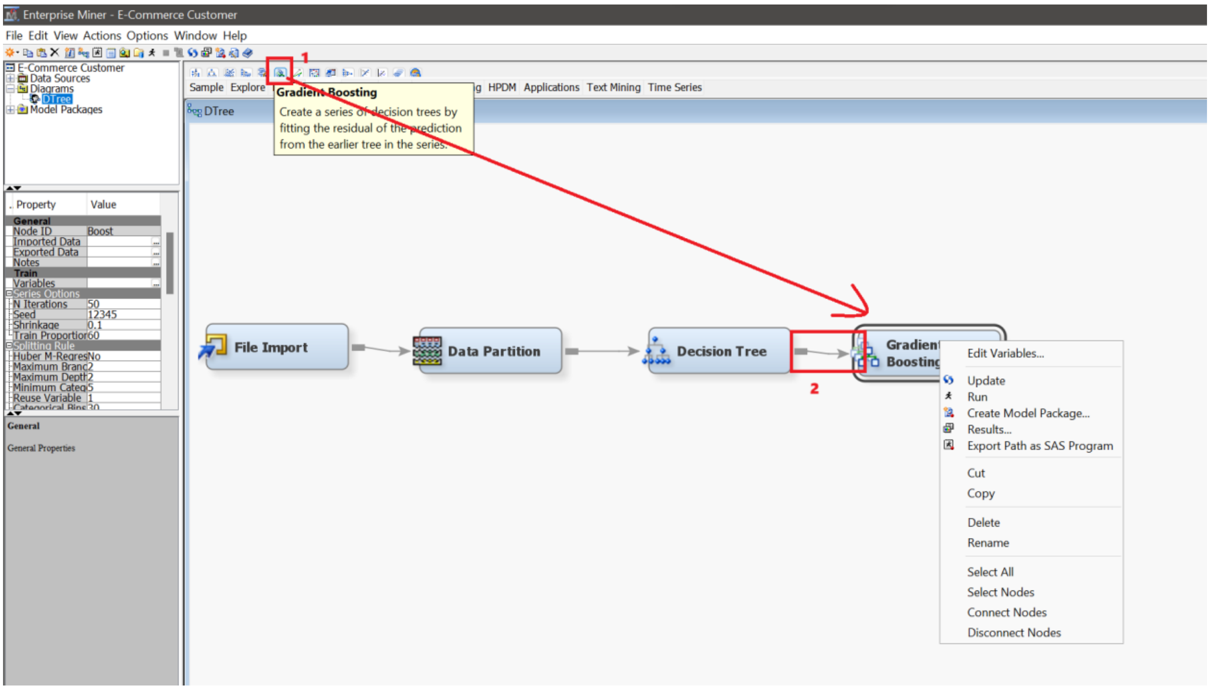
p. Right click HP Forest and run



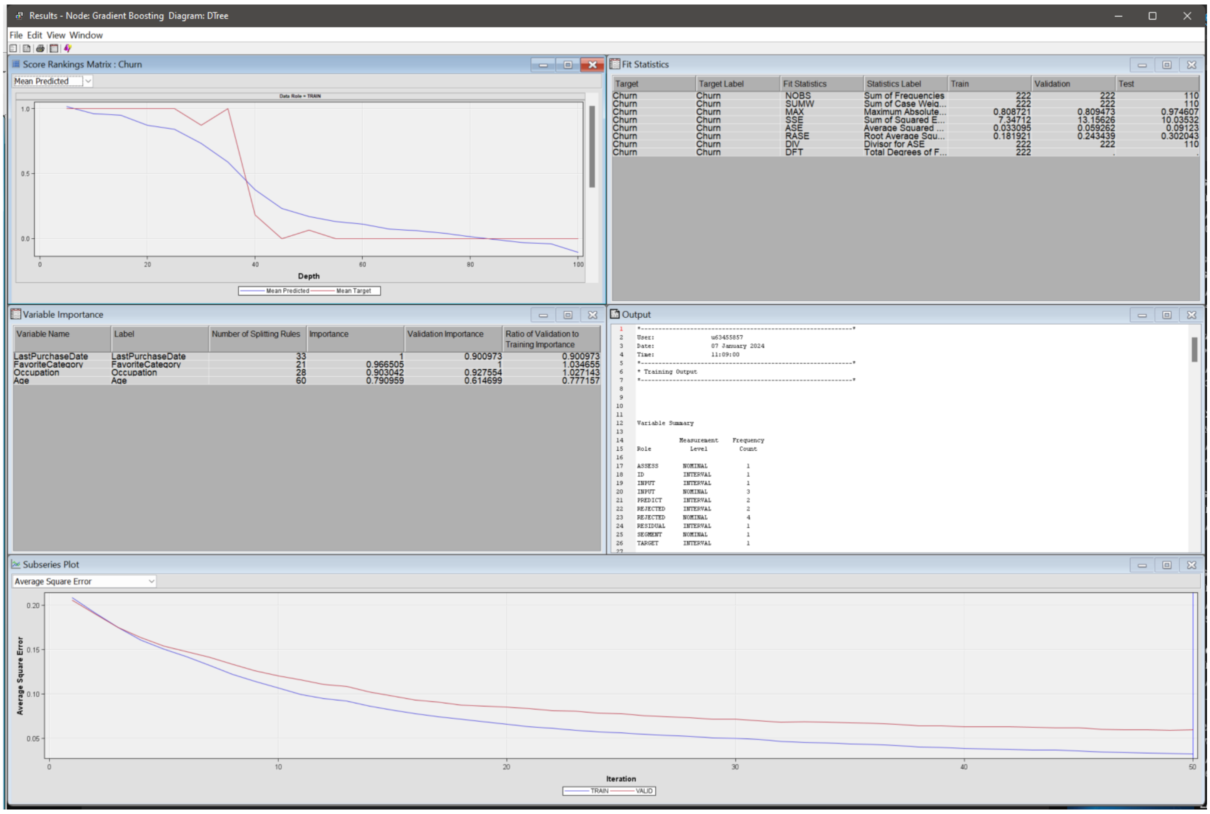
q. Once it done, click result to see the outcome. The Random Forest model, employing 100 trees, was trained on 222 observations, demonstrating competitive fit statistics. The average square error decreased with an increasing number of trees, stabilizing at 0.06170. Key predictors for predicting churn included "FavoriteCategory," "Age," "Occupation," and "LastPurchaseDate." The model's assessment score rankings showcase its efficacy across various depth levels. The scoring process was efficient, taking minimal time. The model's predictive accuracy was assessed on training, validation, and test datasets, yielding Root Average Squared Errors of 0.248, 0.286, and 0.355, respectively. Overall, the Random Forest demonstrated robust predictive performance in identifying potential churn instances.



r. To apply Boosting using Random Forest, I drag "Gradient Boosting" node from the onto the diagram then connect the Decision Tree model node to the Gradient Boosting node. This establishes the base learner for boosting. For the properties I follow the default. Once confirm right click ‘Gradient Boosting’ and click ‘run’.

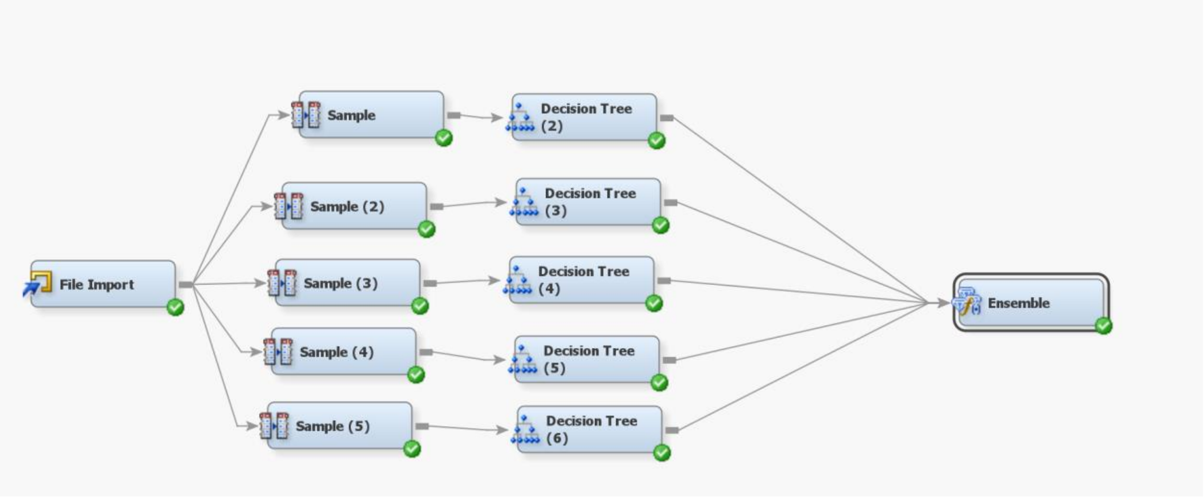


s. Once done, click result to see the outcome. The analysis predicting customer churn revealed significant predictors. The key variables influencing the model include "LastPurchaseDate," "FavoriteCategory," "Occupation," and "Age." The Fit Statistics show good performance on the training set, with an average squared error of 0.033. The Assessment Score Rankings illustrate the model's ability to differentiate between churn and active customers across different depths. Variable Importance indicates "LastPurchaseDate" as the most influential. Overall, the model demonstrates promising predictive power, emphasizing the importance of recent purchase behavior and customer demographics in identifying potential churn.



There another way in getting the bagging process which is below thohrugh esemble. We create few samples with different percentage then link decision tree for analysis. For this case I repeated for 5 times. Lastly I connect all using the ensemble node.

The analyst trained a model ensemble, combining models labeled TREE2 to TREE6 using the average probability function. The training fit statistics indicate an average squared error (ASE) of 0.0318 and a root average squared error (RASE) of 0.1782. Assessment score rankings show varying prediction accuracies across different depths, with observations at depth 5 achieving a perfect match. The assessment score distribution illustrates the model's ability to differentiate between high and low predicted churn probabilities.



Result show as below.