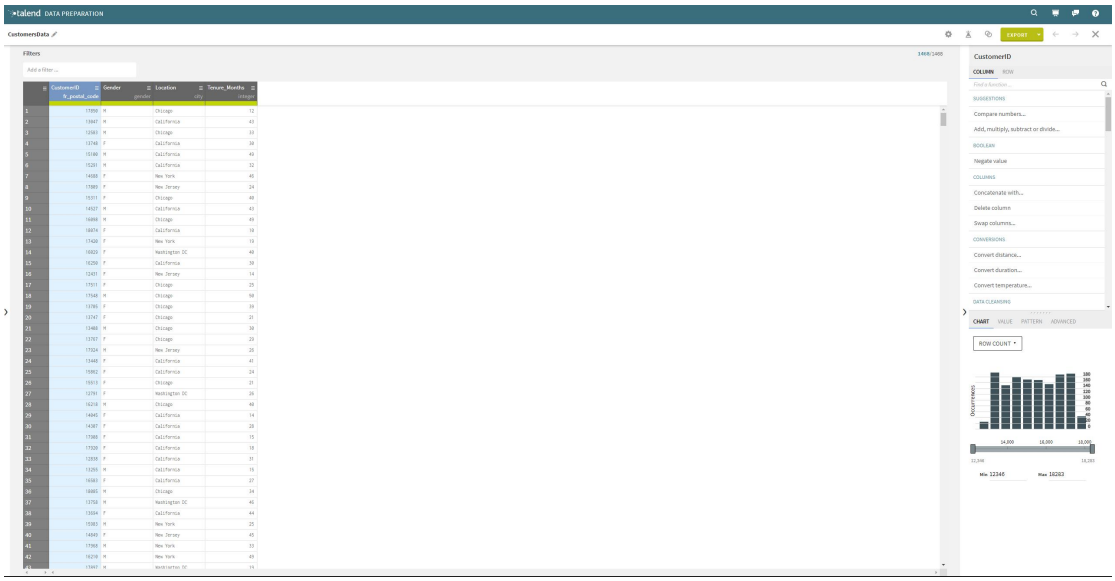
ALTERNATIVE ASSESSMENT 1
Lau Chyuan Jinn 22096017

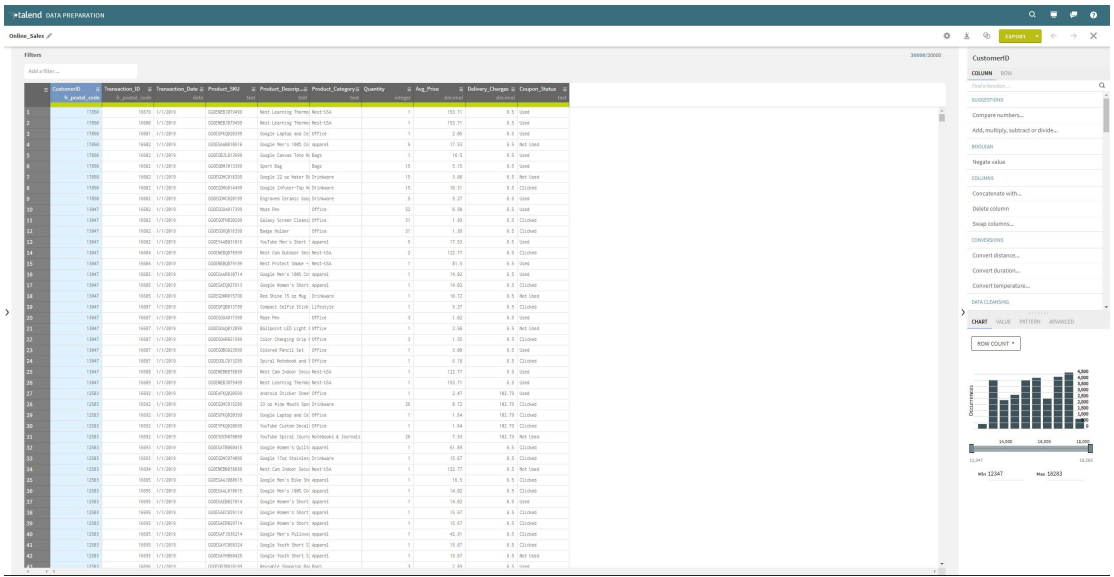
The dataset is found from Kaggle, link:
https://www.kaggle.com/datasets/rishikumarrajvansh/marketing-insights-for-e-commerce-company/data

Download the CustomersData.csv and Online_Sales.csv from Kaggle.

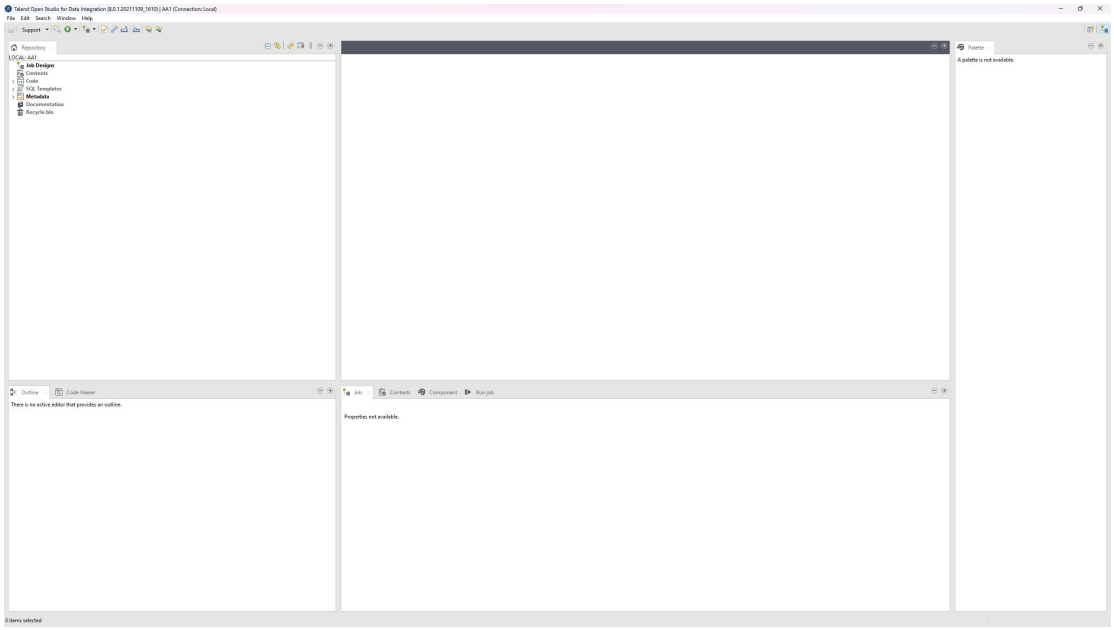
1. Using Talend Preparation to filter the CustomersData.csv, remove the null value row, it is the process for data cleaning.



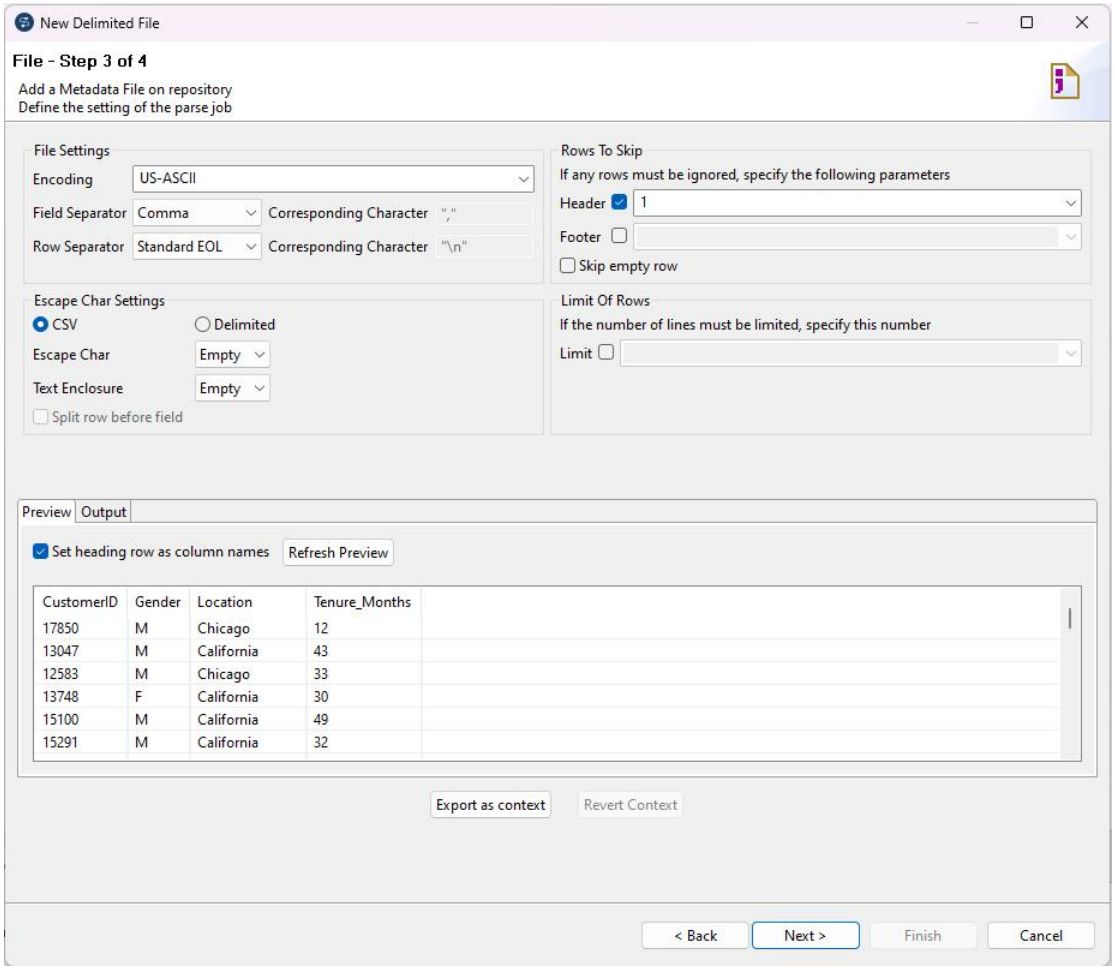
2. Also same to Online_Sales.csv. After that, Export both csv data.



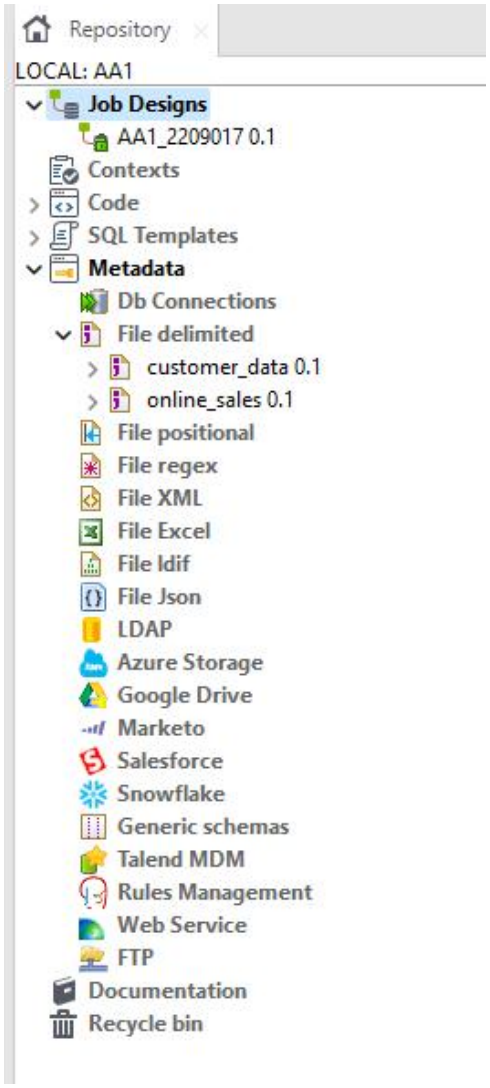
3. Using Talend Open Studio for Data Integration for next step.



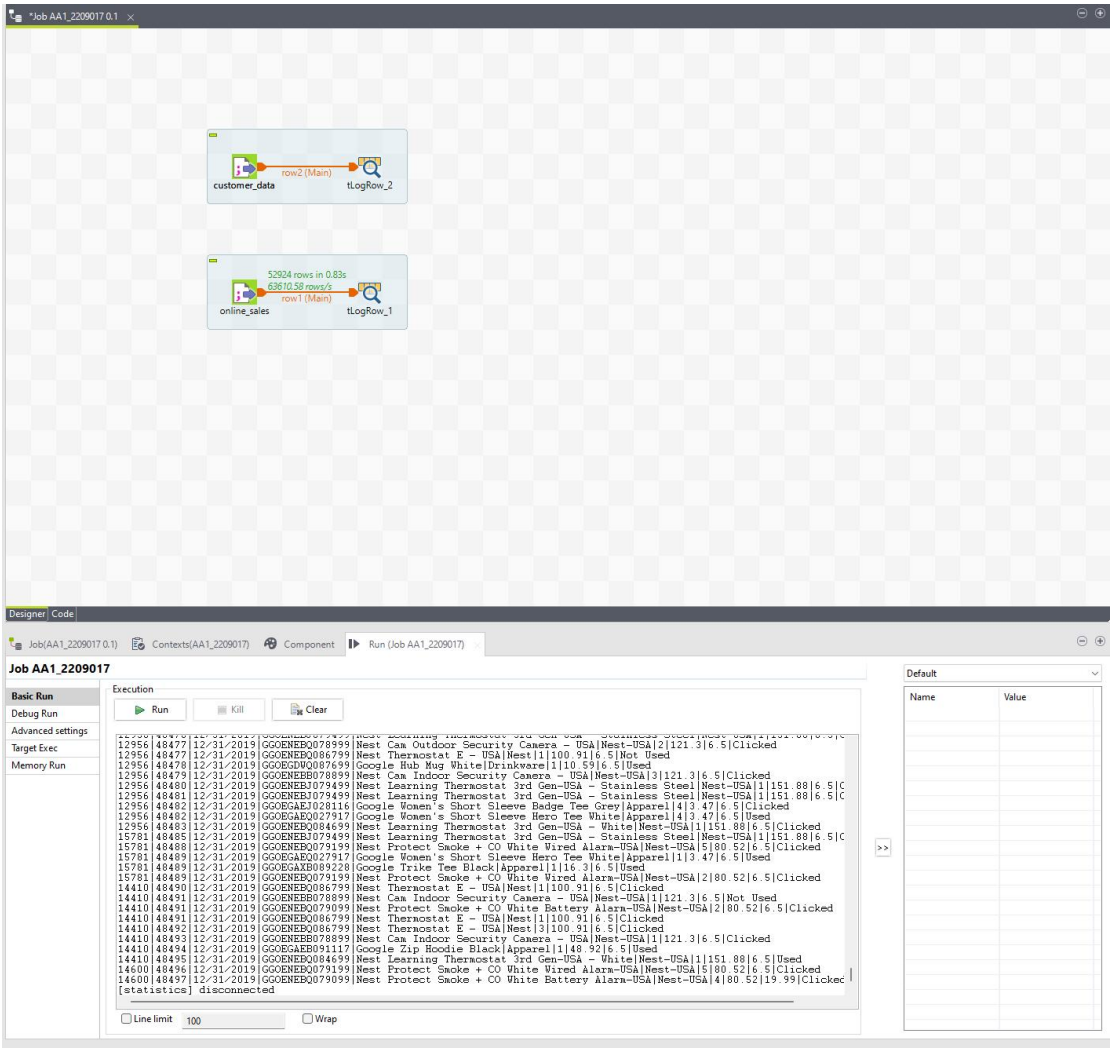
4. Import both csv files in Delimited File.



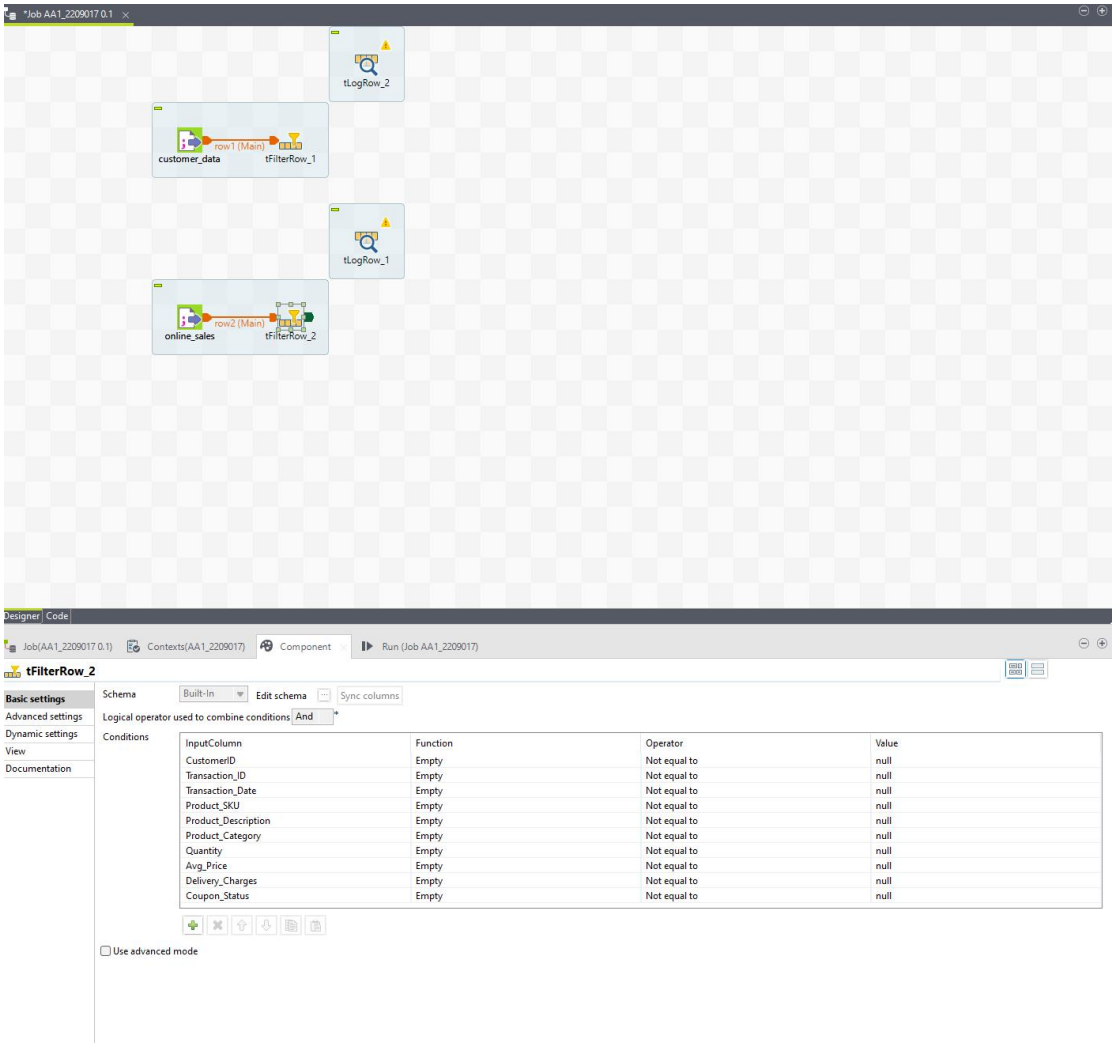
5. The left hand side, Repository will look like that.



6. Then drag customer_data 0.1 and online_sales0.1 to Job diagram. Next connect with tLogRow to view the data whether success display.

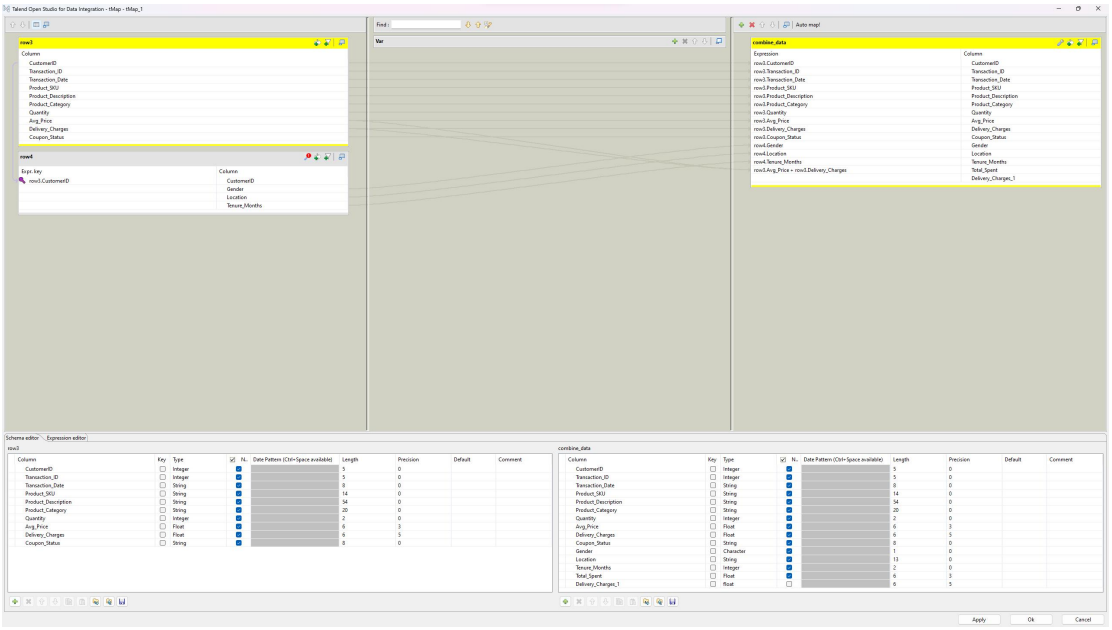


7. Using tFilterRow to filter the null value for both tables.

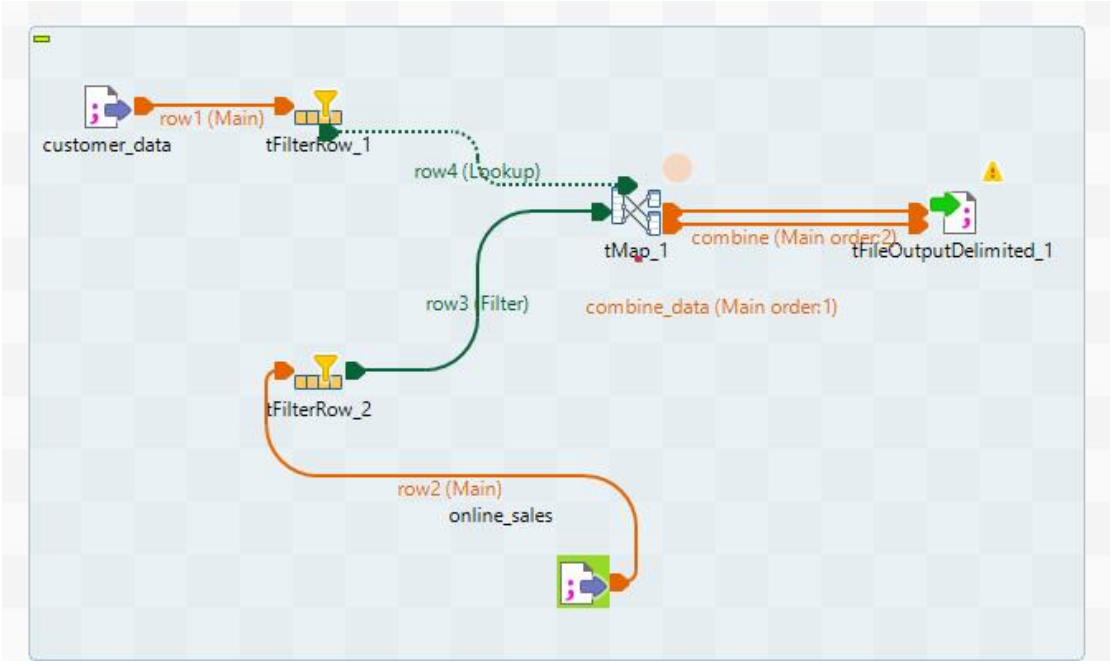


8. Using tMap to join both data, result to combine_data.csv

9. Set CustomerID as expression key.

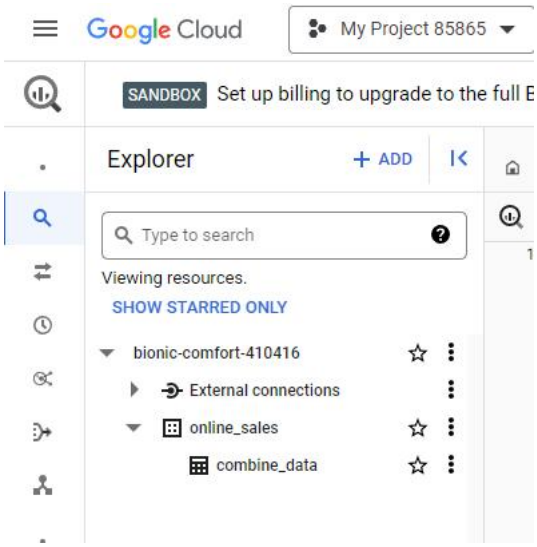


10. Finally export the data using tFileOutputDelimited. The diagram shown as below.

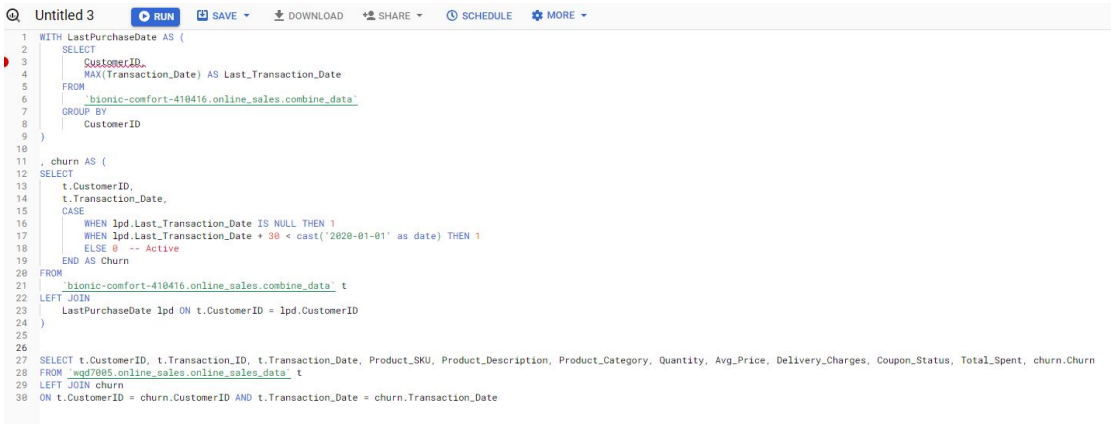


11. Since combine_data.csv still haven't churn variable, which indicates whether the customer has stopped purchasing. Will process to next step to create churn variable.

12. Import combine_data.csv in Google Cloud, then using tool in Google Cloud called BigQuery to apply SQL.

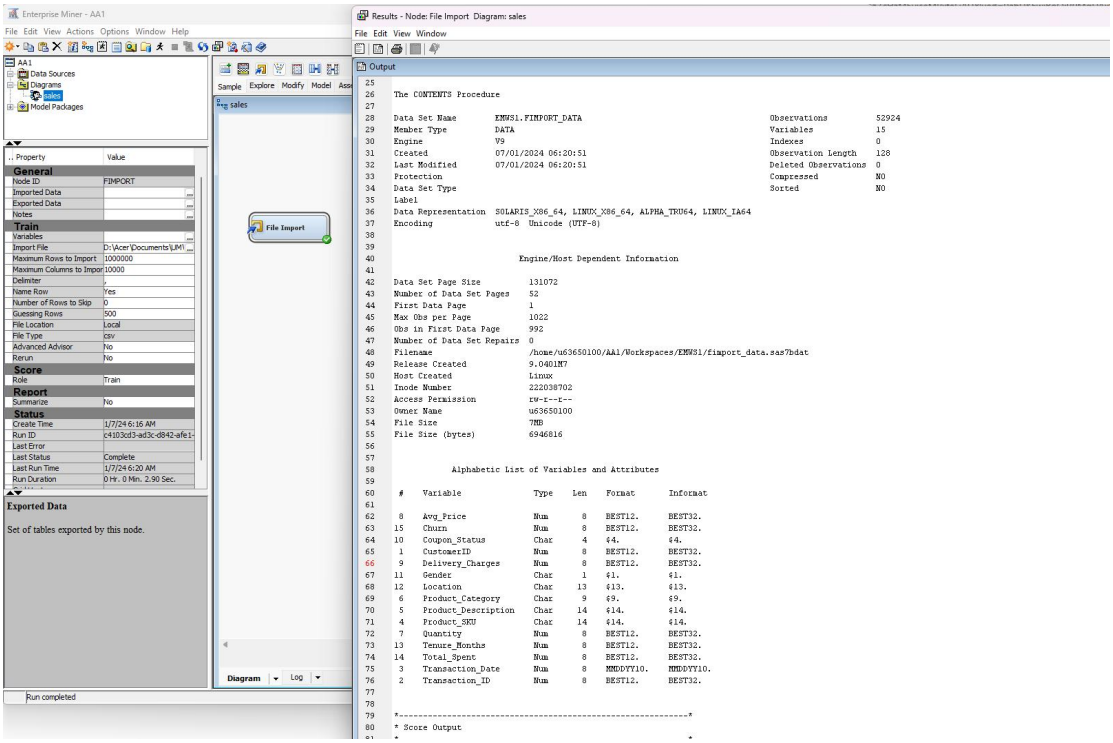


13. With the query below, will add a variable named churn which indicates whether the customer has stopped purchasing more than 30 days before the next year.

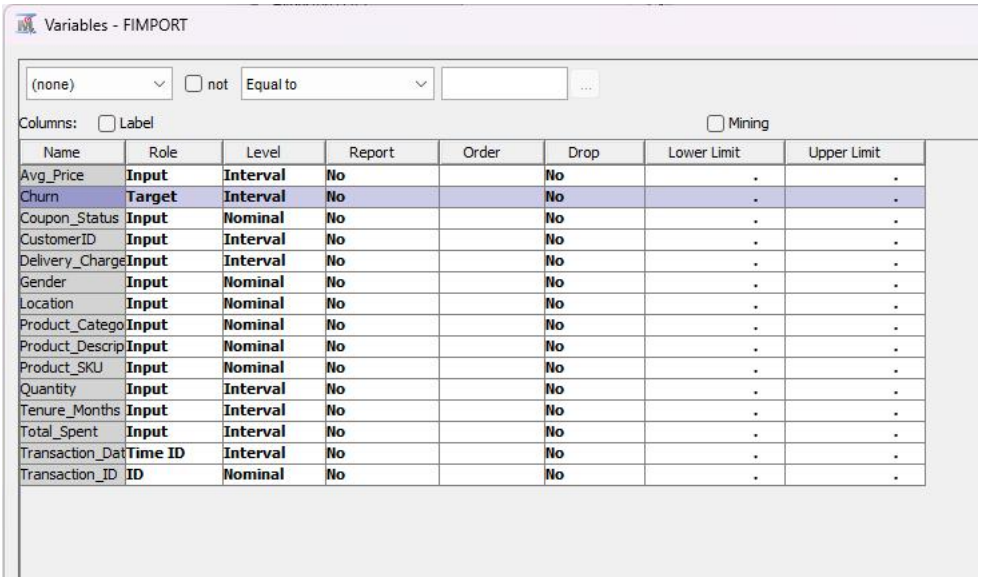


14. The dataset is prepare enough and proceed to next step, which is analyse the dataset in SAS Enterprise Miner.

15. First, import the dataset combine_data.csv into SAS Enterprise Miner.

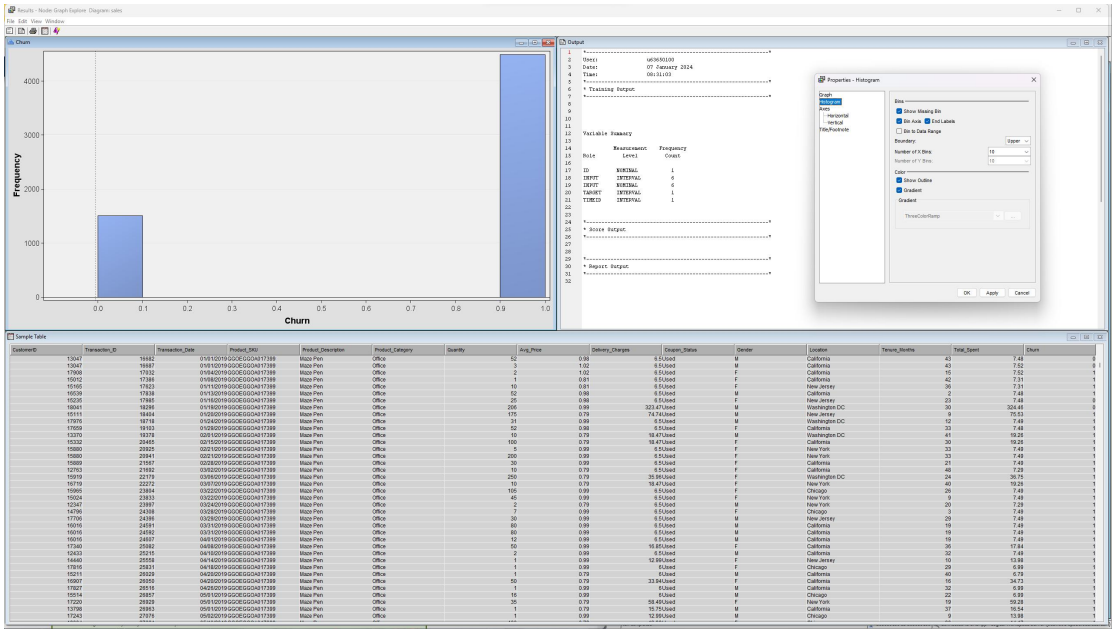


16. Next, set the churn as target variable.

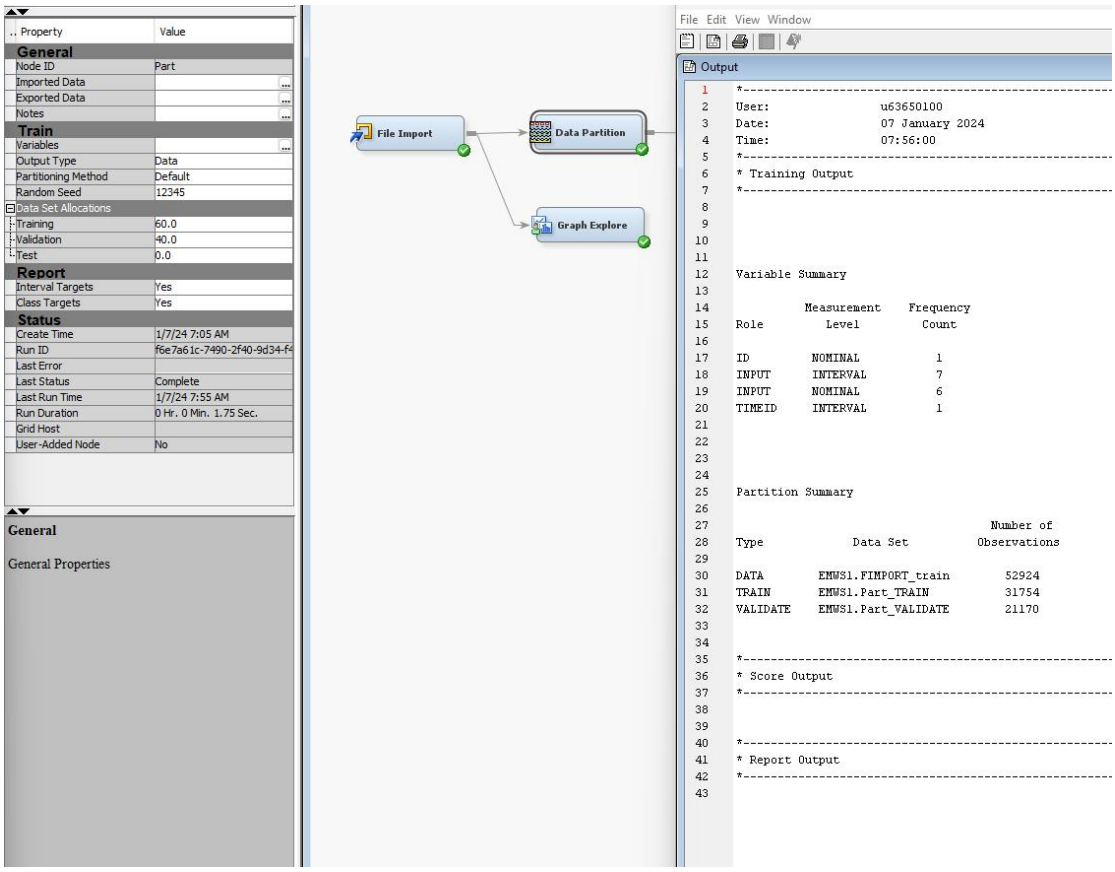


Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Avg_Price	Input	Interval	No		No	.	.
Churn	Target	Interval	No		No	.	.
Coupon_Status	Input	Nominal	No		No	.	.
CustomerID	Input	Interval	No		No	.	.
Delivery_Charge	Input	Interval	No		No	.	.
Gender	Input	Nominal	No		No	.	.
Location	Input	Nominal	No		No	.	.
Product_Catego	Input	Nominal	No		No	.	.
Product_Descrip	Input	Nominal	No		No	.	.
Product_SKU	Input	Nominal	No		No	.	.
Quantity	Input	Interval	No		No	.	.
Tenure_Months	Input	Interval	No		No	.	.
Total_Spent	Input	Interval	No		No	.	.
Transaction_Dat	Time ID	Interval	No		No	.	.
Transaction_ID	ID	Nominal	No		No	.	.

17. With histogram, check the dataset whether have null value. (From diagram shown is 0).



18. Add a Data Partition node, in SAS Enterprise Miner, the Data Partition node is used to divide a dataset into training, validation, and testing sets. It helps in building, fine-tuning, and evaluating predictive models by ensuring proper separation of data for these purposes. The node supports randomization, stratified sampling, and may allow for k-fold cross-validation. Its primary role is in assessing model performance on unseen data and is a crucial step in the model development process.



Property	Value
General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	60.0
Validation	40.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	1/7/24 7:05 AM
Run ID	f6e7a61c-7490-2f40-9d34-f4
Last Error	
Last Status	Complete
Last Run Time	1/7/24 7:55 AM
Run Duration	0 Hr. 0 Min. 1.75 Sec.
Grid Host	
User-Added Node	No

General

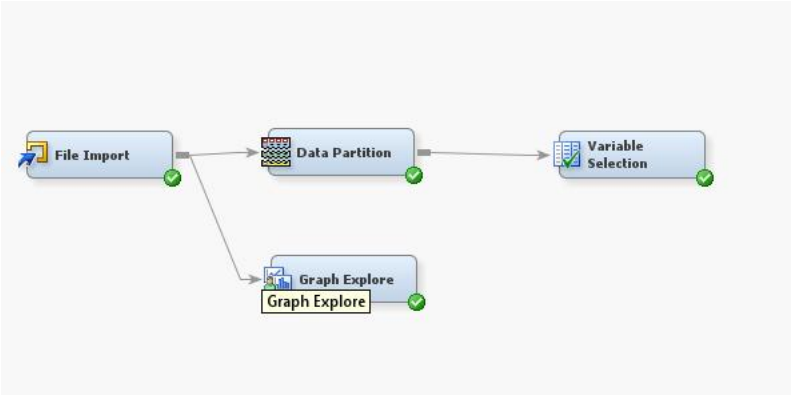
General Properties

File Edit View Window

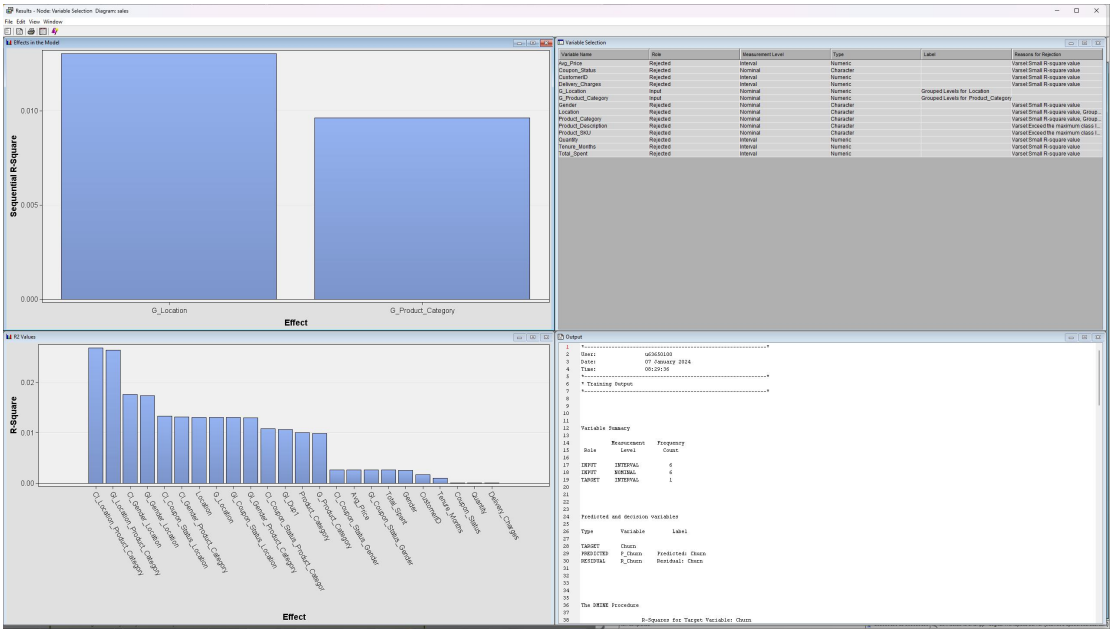
Output

```
*------------------------------------------------------------
User:                  u63650100
Date:                  07 January 2024
Time:                  07:56:00
*------------------------------------------------------------
* Training Output
*------------------------------------------------------------



Variable Summary

          Measurement    Frequency
Role         Level         Count

ID         NOMINAL           1
INPUT      INTERVAL          7
INPUT      NOMINAL           6
TIMEID     INTERVAL          1



Partition Summary

                                          Number of
Type             Data Set                 Observations

DATA       EMWS1.FIMPORT_train            52924
TRAIN      EMWS1.Part_TRAIN               31754
VALIDATE   EMWS1.Part_VALIDATE            21170


*------------------------------------------------------------
* Score Output
*------------------------------------------------------------


*------------------------------------------------------------
* Report Output
*------------------------------------------------------------
```

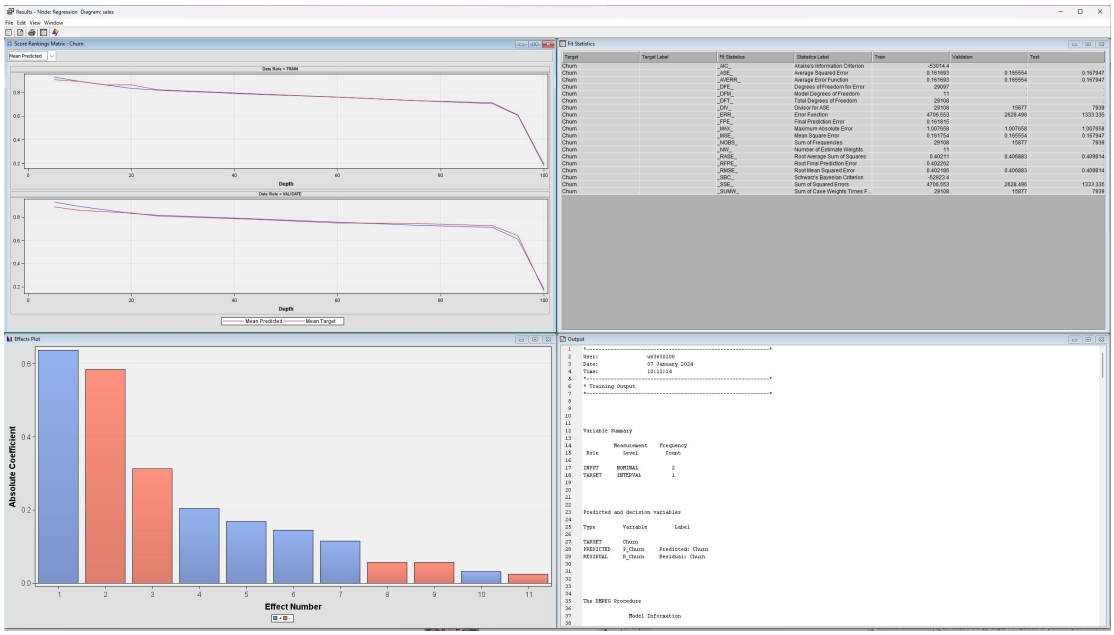
19. Then add a Variable Selection node, Variable selection is a crucial step in building predictive models as it helps identify the most relevant features or variables that contribute to the model's performance.



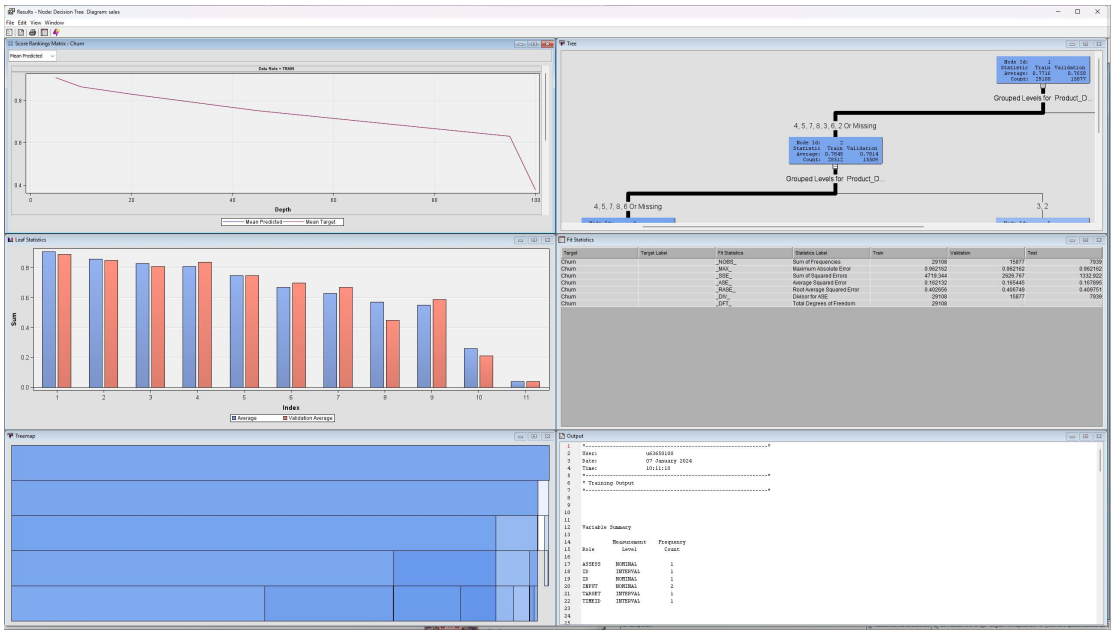
20. The result of Variable Selection shown as below.

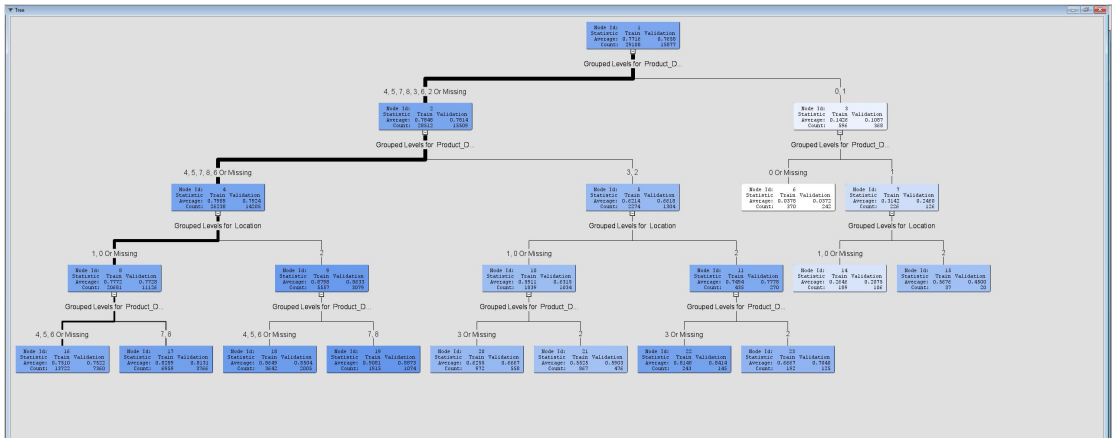


21. Add a Regression node, the result will shown as below. The Regression node is used for building regression models. The regression analysis helps predict a continuous target variable based on one or more predictor variables. The top left section displays two line graphs side by side, showing some trends or data over time. The X-axis is labeled “Days” and the Y-axis has numerical values, indicating it’s a plot of some variables over days. In the top right section, there’s a table with multiple columns including “Name,” “Mean,” “Std Dev,” “Minimum,” etc., suggesting statistical data of different variables. In the bottom left section, there’s a bar graph labeled “Effect Number” on the X-axis and “Value” on the Y-axis. It displays five bars with varying heights. The bottom right section contains text information about model summary including factors like ‘A’, ‘B’, ‘AB’, etc., and their levels. There are also details about random seed and total runs.

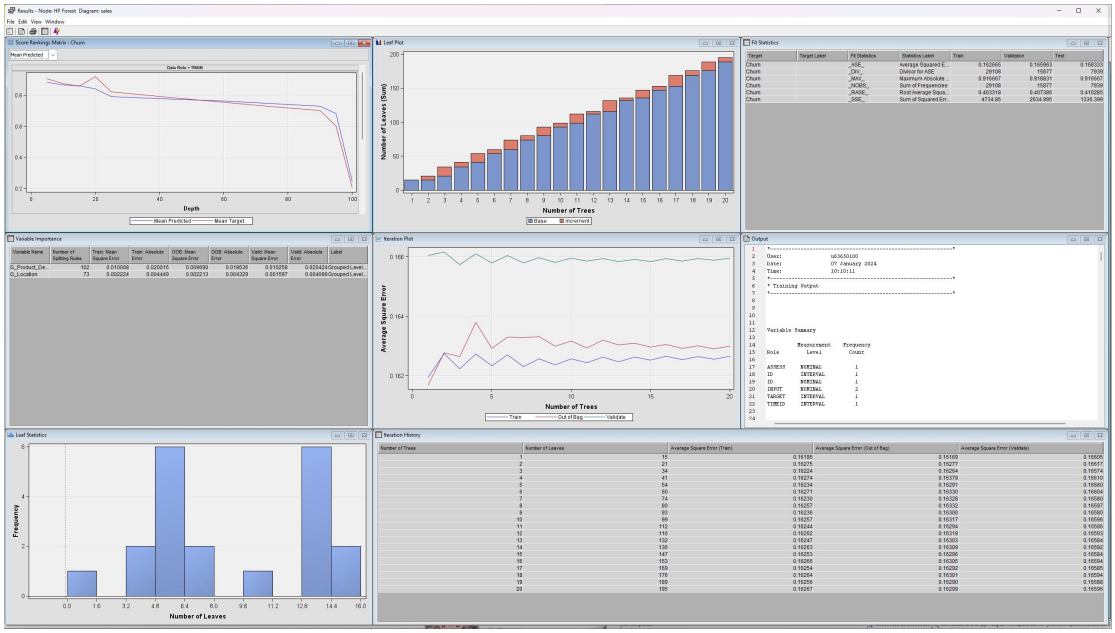


22. Next, add a Decision Tree node, the result will shown as below. The Decision Tree node is used to build decision tree models. Decision trees are a type of predictive modeling technique that recursively splits data into subsets based on the values of predictor variables, ultimately creating a tree-like structure. The image shows a complex flowchart with multiple blue boxes connected by arrows. Each box contains text and numbers, indicating steps or processes in data handling. There are different sections labeled as “0 Missing”, “0-3 Missing”, “4,5,7,8 < 0 Missing” etc., suggesting categories based on missing data. Inside the boxes are parameters like “Show Size”, “Hide Size”, “Validation” and numerical values indicating some form of measurement or evaluation criteria. Arrows connect these boxes indicating the flow of processes or steps from one to another. The background is plain white making the blue boxes and black text prominent.

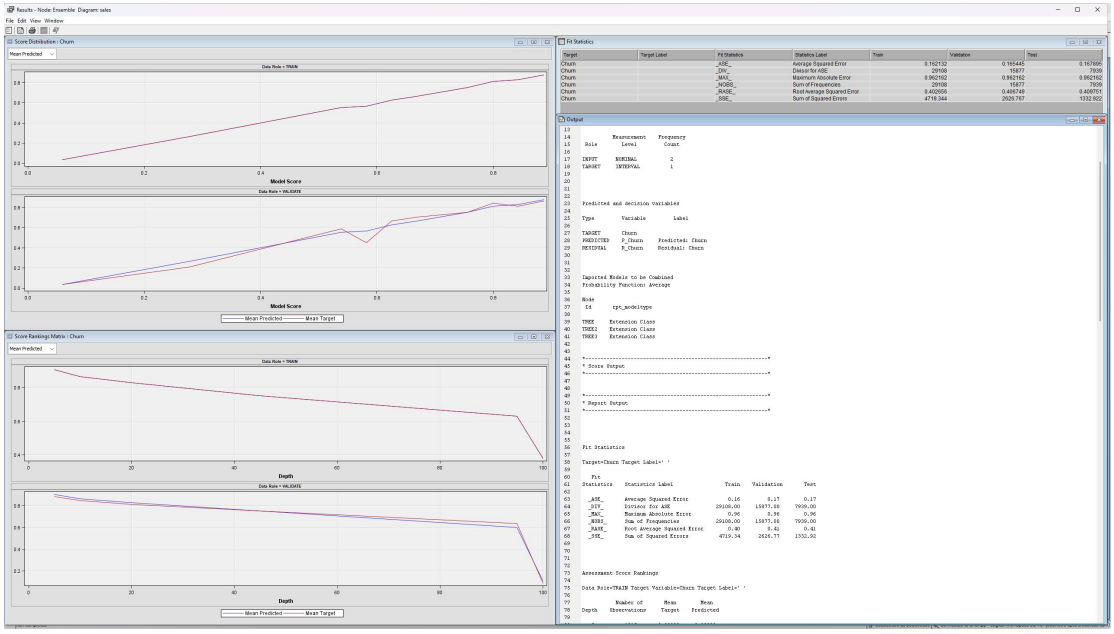




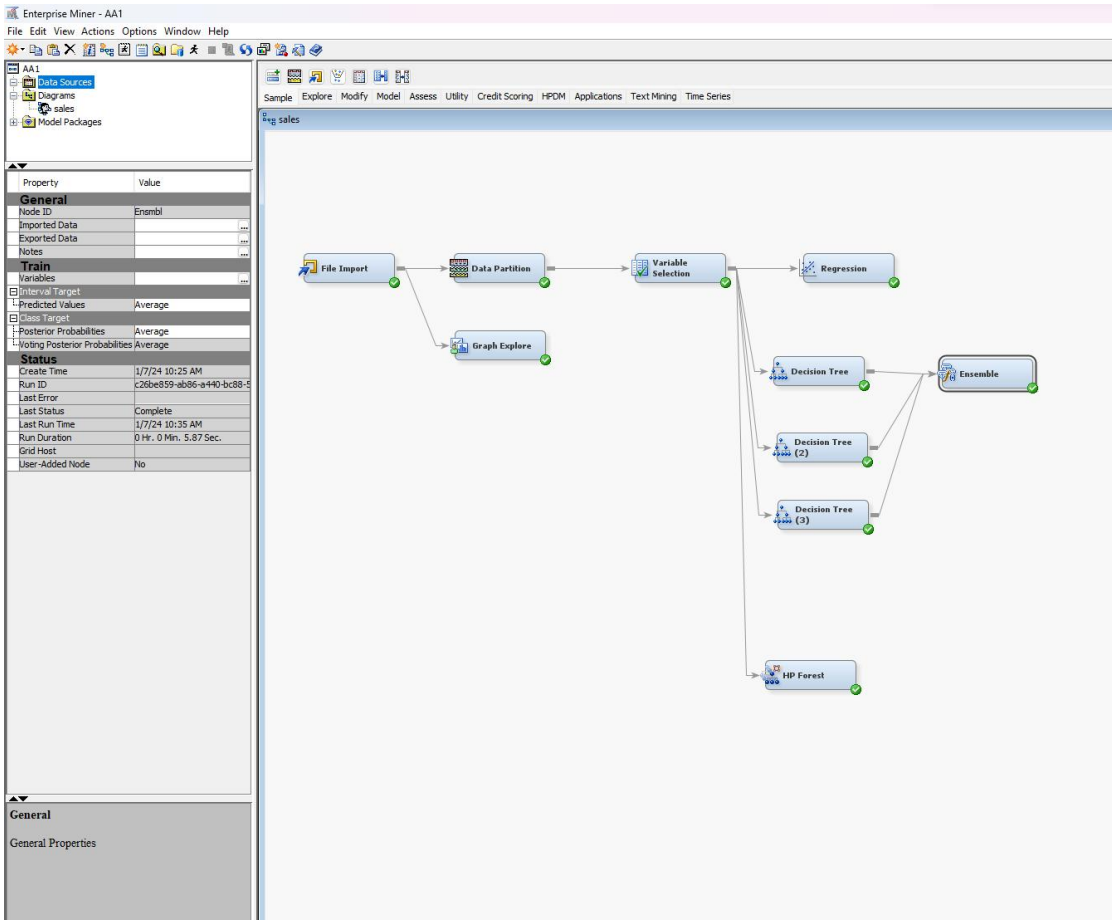
23. Next, add HP Forest, thee result as shown as below. The HP Forest Node is a predictive modeling tool in SAS Enterprise Miner that creates a forest predictive model, which is an ensemble of many decision trees, using the HPFOREST procedure. The top left panel displays a line graph labeled “Run Time vs Type Spectrum” showing two lines that appear to represent different datasets. Adjacent to this, on the top middle panel, there is a bar graph labeled “Number of Types” with bars in two shades representing different categories or datasets. On the top right, there are two small panels; one appears blank and the other displays textual information including values labeled “TOTAL TYPES”, “TOTAL METHODS”, etc. Below these panels in the middle row left side, there’s another line graph labeled “Number of Types”, showing three lines representing different datasets over time. In the middle row right side, there’s a text box containing specific numerical data and labels such as “TOTAL: RANGE” and individual numerical values associated with specific labels like ‘min’, ‘max’, etc. On the bottom left panel, there’s another bar graph labeled “Number of Classes” displaying blue bars at various heights representing different quantities or values. Adjacent to this on the bottom right are two panels displaying tabulated data with rows and columns containing numerical values and text.



24. Next, add 2 more Decision Tree nodes with different leaf (3 and 8). Then connect 3 Decision Tree nodes to a Ensemble node, result is shown as below. Ensemble models combine predictions from multiple individual models to improve overall predictive performance. The left side of the interface shows three graphs plotted, each in its own panel. The top graph is labeled “Big Arm Position (Degrees)” with X-axis labeled as “Time” and Y-axis ranging from 0 to 270. The middle graph is not clearly labeled but has the same X-axis label “Time” and Y-axis labels similar to the top graph. The bottom graph is labeled “Small Arm Velocity (Deg/Sec)” with X-axis labeled as “Time” and Y-axis ranging from 0 to 1000. Each graph displays two lines plotted over time, one in red and another in purple, representing different sets of data or conditions. On the right side, there’s a panel displaying numerical data and parameters under different categories like “Parameters,” “Display,” “Model & Physical Signals,” etc. It includes values for parameters like mass, length, inertia along with other settings and configurations for the displayed graphs or related computations. There are also buttons like “Load Config” indicating that users can load different configurations for analysis.



25. The SAS Enterprise Miner diagram will shown as below.



Conclusion:

In conclusion, the comprehensive analytical process undertaken, from data cleaning and integration using Talend to advanced modeling in SAS Enterprise Miner, provides a robust framework for understanding and predicting customer behavior. The decision tree and ensemble models offer valuable insights, particularly in identifying potential churn patterns. The strategic integration of Google Cloud for SQL queries adds depth to the dataset, enhancing the accuracy of predictive models. Moving forward, businesses can leverage these insights to implement targeted customer retention strategies and personalized marketing approaches, thereby improving overall customer satisfaction and optimizing marketing efforts.