Master of Data Science (Semester 1 – 2022/2023)

Faculty of Computer Science & Information Technology

WQD7005 Data Mining

Group Assignment

Bike Sales Forecasting by Data Mining Methodology

Prof. Dr. Teh Ying Wah

Group Members	Student ID
Salah AlKafrawi	S2108437
Usman Ali	S2155904
Chai Kang Sheng	S2156317
Ahmed Abdalla	S2030177
Sadman Chowdhury	S2199546

Table of Contents

1 Introduction

Back in 2009, the European Union adopted an Urban Mobility Action plan. The main goal of this plan is for European cities to adopt a more environmentally friendly future by reducing the usage of carbon-emission vehicles. In order to run this plan, citizens have to resort to using highly efficient and low-emission transport (Waldykowski et al., 2021). The bicycle fits into this category as a means of urban transport. Additionally, many other countries gathered at the Paris Climate Conference back in 2015, and they all agreed to put in more effort to improve the environment (International Energy Agency Report, 2015). Most of the discussions were related to road transport. In 2015, countries like Netherlands and China had a 1% increase in sales for electric vehicles and bicycles (Campinez-Romero et al., 2018).

Cities have been actively trying to push for a more environmentally friendly area by adopting different plans such as banning cars from city centers (Doll & Vetter, 2017) and automobile traffic restrictions have been adopted in Germany for more than 20 years (Pucher & Buehler, 2008). All these studies imply that with the reduced usage of automobiles, there is an increase in sales of urban transport, mainly the bicycle. Given that there is an increase usage of bicycle globally, our group is interested to predict bike sales and look further in terms of which country has the higher sales, which gender is more likely to purchase bikes, as well as if there is an increase in sales of other bike merchandise as wel1. By further understanding this information, bike companies can make a better marketing strategy to increase their revenue in the future.

Furthermore, the data mining phase is a significant step in any business-oriented activity as it is the backbone of successful decision-making. Data scientists and managers apply data mining techniques to improve their decisions to enhance their business outcomes in terms of quality, revenue, and consistency. Hence, in order to create a successful business, we need to have a data mining phase focused on improving business outcomes. to have an impactful mining process, the goal of analysis must be well-defined because asking the right questions solves half of the problem. In this project**,** we want to find hidden yet valuable information about the market for bikes. Mining hidden information from bike sales data can lead to impactful results such as increasing profits in the short term or even becoming market prone in the long time. Therefore, we aim to use SAS Enterprise Miner software to understand the nature of data via visualization and inspect patterns via machine learning techniques.

2 Dataset

2.1 Dataset Information

Bikes Sales data is used to analyze the sales of bikes and its accessories in different countries to understand what are the major factors that affect the sales, it contains demographic information about the customers, such as their gender, age and geographical location. Furthermore, the data provides insight into three main things, customer demographics, customer behavior and sales. Using these three factors we can predict the sales in the future accordingly.

Table 1: Table Information

Property	Value
Data Source	https://www.kaggle.com/code/alaaalghmdi/bike-sales-eda-prediction-step-by-step/data?select=Sales.csv
Data Name	Bikes Sales
Data Size	15.24 MB
Year	2021
Dimension	113,036 Rows & 18 Columns

2.2 Column Metadata

Metadata explains the attributes of the datasets, their types, ranges and their descriptions. Moreover, it is useful to have a Metadata in order to facilitate the process of understanding the data, especially when it's handed over to other machine learning engineers or data scientists.

Table 2: Metadata

Attribute	Category	Description
Date	object	Date of product purchase
Day	int64	Day of product purchase
Month	object	Month of product purchase
Year	int64	Year of product purchase Range: 2011 - 2016
Customer Age	float64	Age of customer Range: 17 - 87
Age Group	object	Age group of customers Adults (35-64), Young Adults (25-34), Youth (<25)
Customer Gender	object	Gender of the customer ● F- Female ● M- Male

Country	object	The country in which the purchase has been made. Available countries: United States – Australia – Canada – United Kingdom – Germany - France
State	object	The state in which a purchase has been made. Available states: California - British Columbia – England – Washington - New South Wales
Product Category	object	The category of the product purchased. Available categories: ● Accessories ● Bikes ● Clothing
Subcategory	object	The subcategory of the purchased product. Available subcategories: Tires and Tubes - Bottles and Cages - Road Bikes – Helmets - Mountain Bikes
Product	object	The purchased product. Available Products: ● Water Bottle - 30 oz. ● Patch Kit/8 Patches ● Mountain Tire Tube ● AWC Logo Cap ● Sport-100 Helmet, Red
Order Quantity	int64	The purchased quantity of a product. Range: 1 to 32
Unit Cost	int64	The cost of producing one product unit. Range: 1 to 2171
Unit Price	int64	Unit price of the product
Profit	int64	The profit of selling one product unit. Profit = Revenue - Cost Range: 2 to 3578
Cost	int64	The total cost of a purchase. Range: 1 to 43K
Revenue	int64	The revenue from a purchase. Range: 2 to 58.1K

3 Business Understanding

3.1 Analysis Goal

A bike manufacturing company is interested in sales growth of its bikes and accessories. Use bikes sales data from earlier years to forecast profit. This goal can be achieved via

1 - Find most relevant features for profit forecasting

2 - Build models that forecast profit

3 - Evaluate which model is better to forecast profit

3.2 Analysis Data

- Extracted from bike sales transactional data between the years 2011 to 2016
- Located in America, Canada, Australia and Europe
- Contains customers' demographics like age, city and gender etc.
- Contains bike sales revenue and profit information

4 Methodology

In this project, we are following the systematic methodology SEMMA that was developed by SAS institute inc. SEMMA stands for Sample, Explore, Modify, Model, and Asses. The table below demonstrates each step in our project based on SEMMA methodology.

Table 3 Project SEMMA methodology

No.	Step	Action
1	Sample	In this step, we created a project in SAS. Then we imported a sample of data that represents the population. After that, we identified the role of each variable and its level. Lastly, we saved the data file in sas7bdat format.
2	Explore	In this step, we explored features and their patterns using univariate, bivariate and multivariate analysis. These analyses are conducted using the plotting tools in SAS Enterprise Miner. We used histogram and box plot to identify features distribution. We used pie charts for categorical variables, and we used scatter plot to quantitively identify the correlation between variables such as age group vs profits.
3	Modify	In this step, we will check data quality by inspecting missing values and replace them by the suitable values. We will also check for other errors such as duplicates, inconsistent, or outlier data.
4	Model	In this step, we will build a regression model that will predict the profit. The prediction can be achieved via regression algorithms.
5	Assess	In this step, we will assess the prediction performance in terms of R^2 and root mean squared error (RMSE) metrics.

5 Results

Bike Sales dataset excel file is uploaded on SAS Studio of SAS® OnDemand for Academics. Bike Data directory has been created in Server Files and Folders and file was uploaded and imported successfully in SAS format as shown in **Figure 1.**

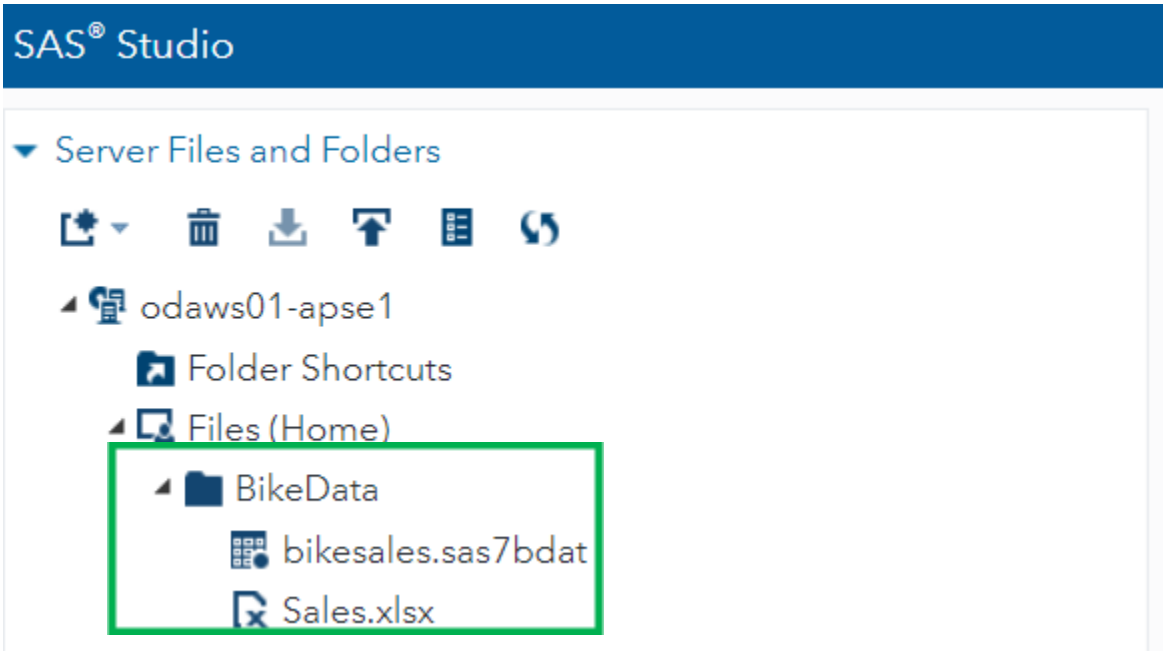


Figure 1 - Uploaded Bike Sales Dataset on SAS Studio

A new library 'Bike Sale' has been created in SAS Studio as shown in **Figure 2**.

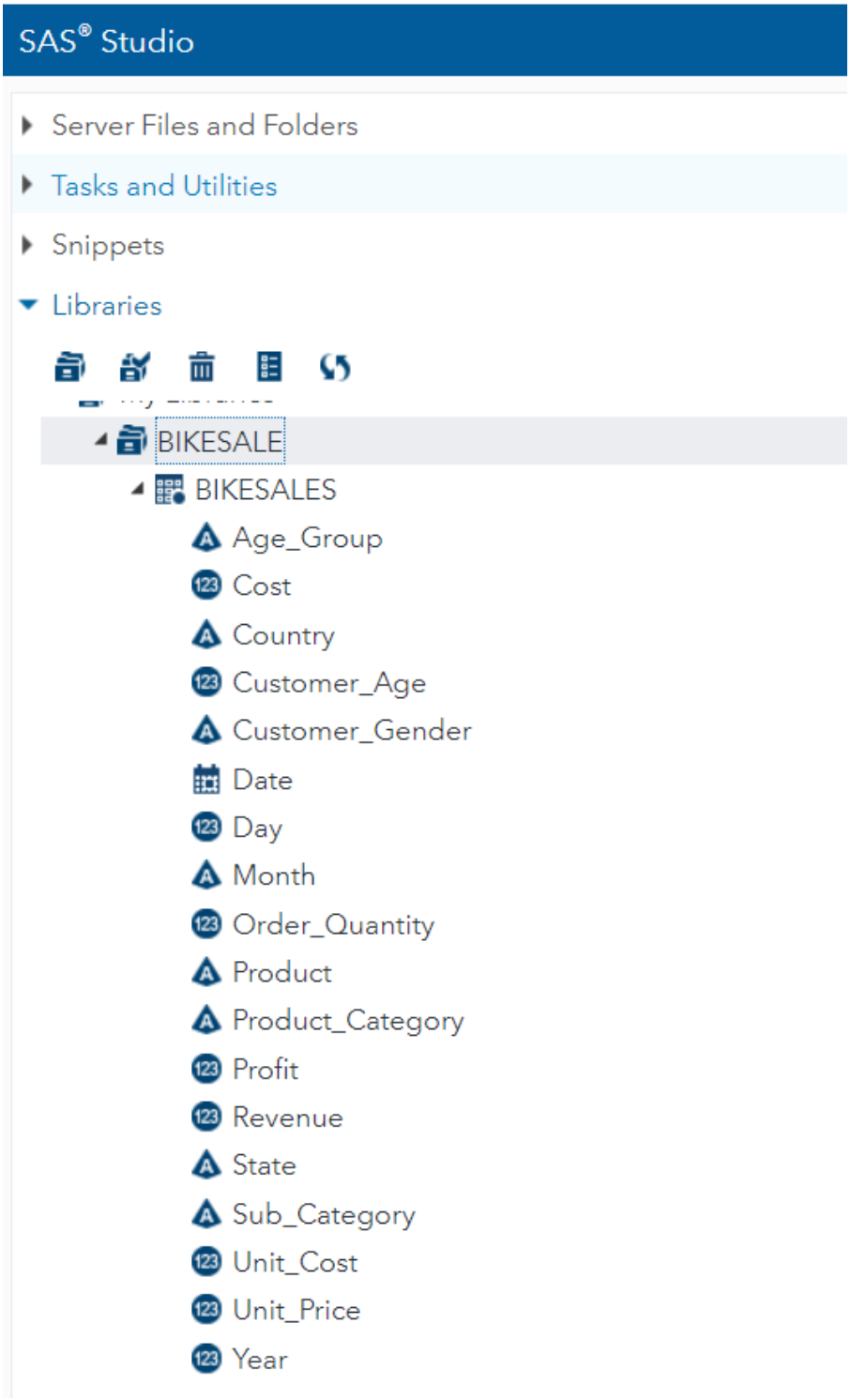


Figure 2 - Bike Sale Library Created

Properties of created library shown in **Figure 3.**

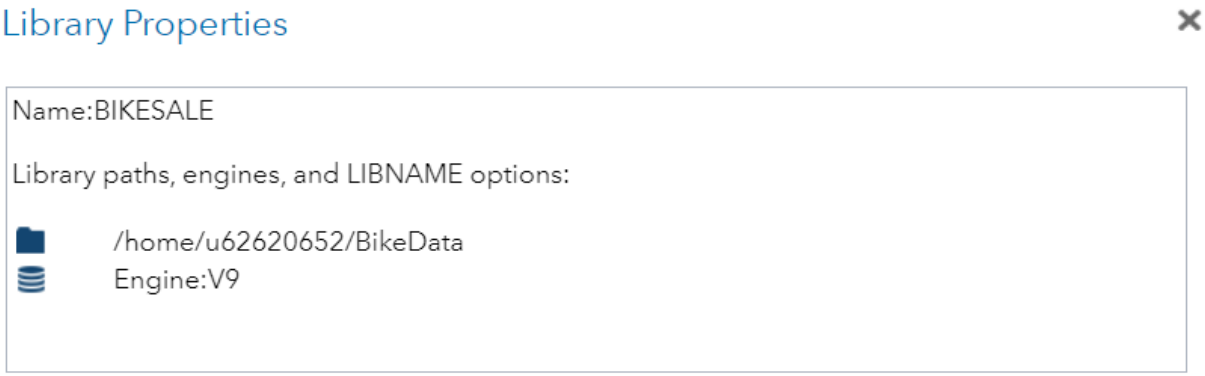


Figure 3 - Bike Sale Library Properties

Bike Sales project has been created in SAS Enterprise Miner of SAS® OnDemand for Academics using following steps:

- Created New Project 'Gourp3-Assignment'
- Created New Library 'Group3A' in the project
- Created New Diagram 'Predictive Analysis' in the project
- Create New Data Source SAS table 'Group3A.EM_SAVE_TRAIN'
- Opened 'Predictive Analysis' Diagram pane and performed Sample and Explore steps of SEMMA

The results of Sample and Explore steps of SEMMA are explained below:

5.1 Sample – Accessing and Assaying

Open Sample tab in diagram window and drag File Import node to the pane. Run File Import node and Bike Sales dataset csv file will be imported into SAS Enterprise Miner.

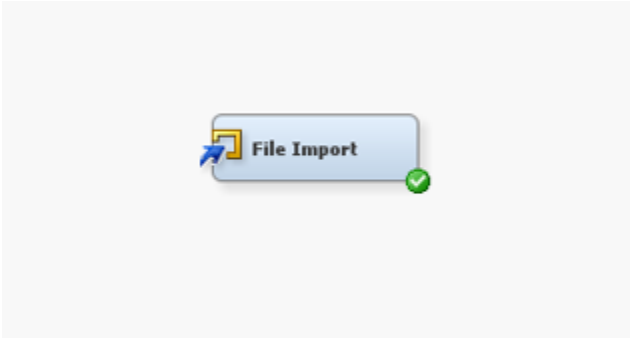


Figure 4 - Import Bike Sales CSV File

Dataset is imported successfully as shown in **Figure 4**, and output is shown in **Figure 5**.

Output

```
Access Permission           rw-r--r--
Owner Name                  u62620652
File Size                   20MB
File Size (bytes)           21102592


          Alphabetic List of Variables and Attributes

 #    Variable            Type    Len    Format       Informat

 6    Age_Group           Char     20    $20.         $20.
17    Cost                Num       8    BEST12.      BEST32.
 8    Country             Char     14    $14.         $14.
 5    Customer_Age        Num       8    BEST12.      BEST32.
 7    Customer_Gender     Char      1    $1.          $1.
 1    Date                Num       8    YYMMDD10.    YYMMDD10.
 2    Day                 Num       8    BEST12.      BEST32.
 3    Month               Char      9    $9.          $9.
13    Order_Quantity      Num       8    BEST12.      BEST32.
12    Product             Char     19    $19.         $19.
10    Product_Category    Char     11    $11.         $11.
16    Profit              Num       8    BEST12.      BEST32.
18    Revenue             Num       8    BEST12.      BEST32.
 9    State               Char     19    $19.         $19.
11    Sub_Category        Char     10    $10.         $10.
14    Unit_Cost           Num       8    BEST12.      BEST32.
15    Unit_Price          Num       8    BEST12.      BEST32.
 4    Year                Num       8    BEST12.      BEST32.


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


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




Exported Attributes for TRAIN Port

          Measurement    Frequency
 Role        Level         Count

COST       INTERVAL          1
INPUT      INTERVAL          8
INPUT      NOMINAL           8
TIMEID     INTERVAL          1
```

Figure 5 - Dataset Import Result

Now, drop Save Data node from utility tab into diagram pane, connect File Import node to Save Data node (**Figure 6**) and run it. It will save imported data into SAS format and output is shown in **Figure 7**.

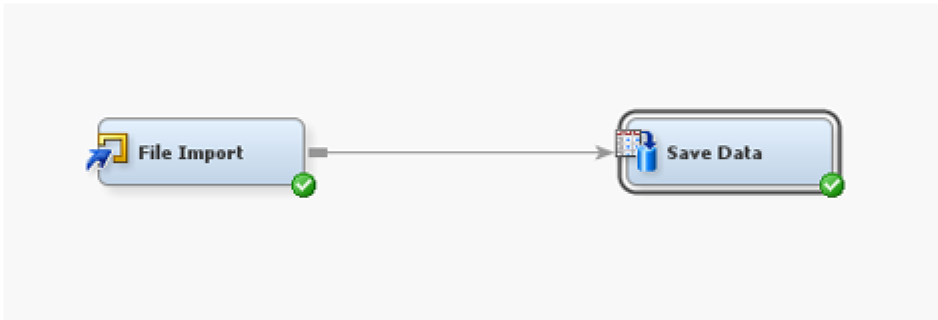


Figure 6 - Save data in SAS Format

Output

```


Variable Summary

           Measurement    Frequency
 Role         Level         Count

COST        INTERVAL          1
INPUT       INTERVAL          8
INPUT       NOMINAL           8
TIMEID      INTERVAL          1



Saved Data Properties

  Data                                                                                    Total        Saved      Number of
Library                                 Output Location                               Observations  Observations  Variables

_em_save    /home/u62620652/Group3-Assignment/Workspaces/EMWS1/EMSave/em_save_TRAIN.sas7bdat     113036        MAX           18


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

Figure 7 - Save Data Result

Bike Sales dataset contains 18 variables and 113,036 observations.

Basic view of dataset column metadata is shown in **Figure 8**.

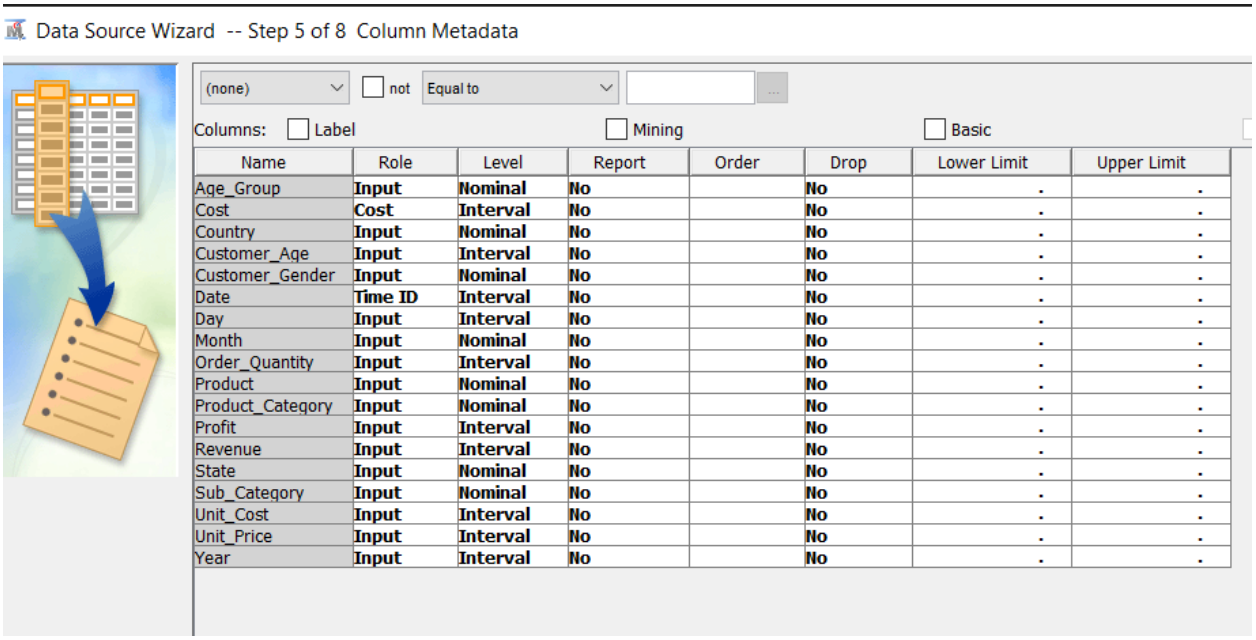


Figure 8 - Basic View of Column Metadata

The following issues were resolved in the advanced column metadata in **Table 5**.

Table 4 - Role and Level of Columns

Column Metadata Basic View			**Column Metadata Advanced View**		
Name	**Role**	**Level**	**Name**	**Role**	**Level**
Profit	Input	Interval	Profit	Target	Interval
Date	Time ID	Interval	Date	Input	Interval
Day	Input	Interval	Day	Input	Interval
Year	Input	Interval	Year	Input	Interval
Unit Cost	Input	Interval	Unit Cost	Input	Interval
Unit Price	Input	Interval	Unit Price	Input	Interval
Order Quantity	Input	Interval	Order Quantity	Input	Interval
Cost	Cost	Interval	Cost	Input	Interval
Revenue	Input	Interval	Revenue	Input	Interval
Customer Age	Input	Interval	Customer Age	Input	Interval
Product	Input	Nominal	Product	Input	Nominal
Product Category	Input	Nominal	Product Category	Input	Nominal
Subcategory	Input	Nominal	Subcategory	Input	Nominal

Age Group	Input	Nominal	Age Group	Input	Nominal
Month	Input	Nominal	Month	Input	Nominal
Country	Input	Nominal	Country	Input	Nominal
State	Input	Nominal	State	Input	Nominal
Customer Gender	Input	Nominal	Customer Gender	Input	Binary

Role and level summary of Columns is provided below in **Table 6** and **Figure 9.**

Table 5 – Role and Level Summary

Role	Level	Count
Input	Nominal	7
Input	Interval	9
Input	Binary	1
Target	Interval	1
		18

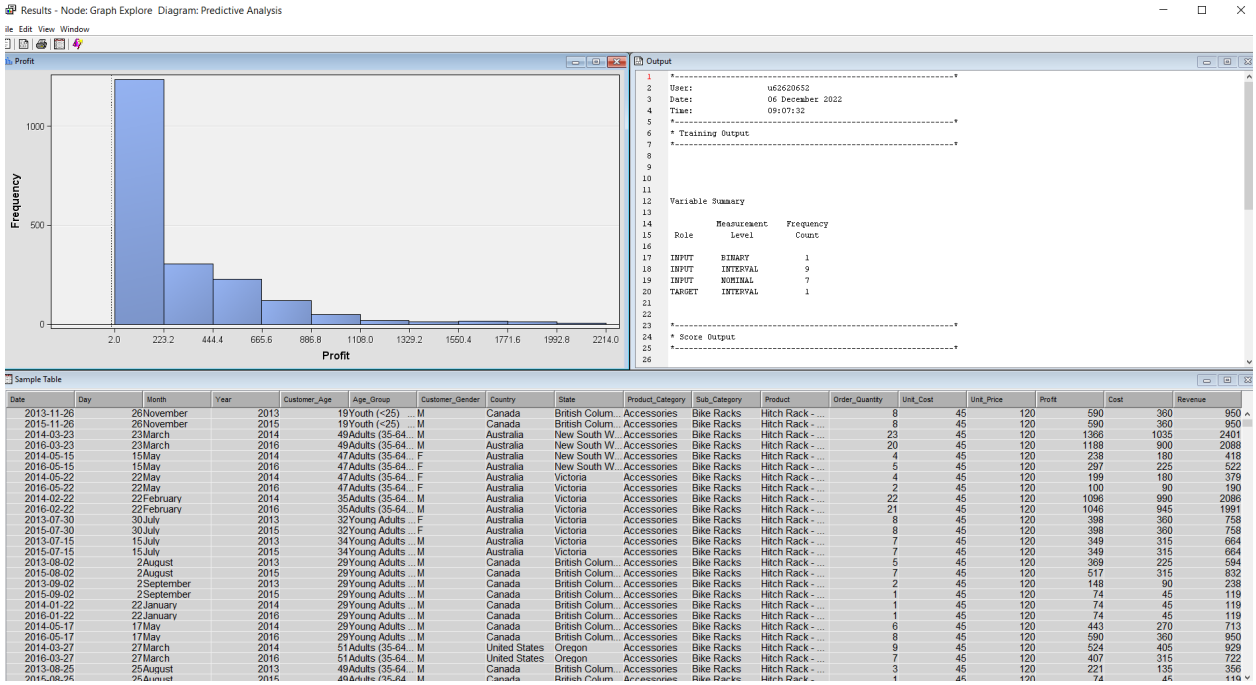
Output

```
*------------------------------------------------------------*
User:                  u62620652
Date:                  06 December 2022
Time:                  09:05:15
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*



Variable Summary

             Measurement    Frequency
 Role          Level          Count

INPUT       BINARY            1
INPUT       INTERVAL          9
INPUT       NOMINAL           7
TARGET      INTERVAL          1
```

Figure 9 - Variable Summary

5.2 Explore – Exploring Data Source

Open ‘Explore’ tab of ‘Predictive Analysis’ diagram. Drag and drop Bike Sales data source into diagram pane. Now, drag and drop ‘Graph Explore’ node into diagram pane, connect to the data source and run it. The results are shown in **Figure 10**.



Date	Day	Month	Year	Customer_Age	Age_Group	Customer_Gender	Country	State	Product_Category	Sub_Category	Product	Order_Quantity	Unit_Cost	Unit_Price	Profit	Cost	Revenue
2013-11-26	26	November	2013	19	Youth (<25)	M	Canada	British Colum...	Accessories	Bike Racks	Hitch Rack - ...	8	45	120	590	360	950
2015-11-26	26	November	2015	19	Youth (<25)	M	Canada	British Colum...	Accessories	Bike Racks	Hitch Rack - ...	8	45	120	590	360	950
2014-03-23	23	March	2014	49	Adults (35-64)	M	Australia	New South W...	Accessories	Bike Racks	Hitch Rack - ...	23	45	120	1366	1035	2401
2016-03-23	23	March	2016	49	Adults (35-64)	M	Australia	New South W...	Accessories	Bike Racks	Hitch Rack - ...	20	45	120	1188	900	2088
2014-05-15	15	May	2014	47	Adults (35-64)	F	Australia	New South W...	Accessories	Bike Racks	Hitch Rack - ...	4	45	120	238	180	418
2016-05-15	15	May	2016	47	Adults (35-64)	F	Australia	New South W...	Accessories	Bike Racks	Hitch Rack - ...	5	45	120	297	225	522
2014-05-22	22	May	2014	47	Adults (35-64)	F	Australia	Victoria	Accessories	Bike Racks	Hitch Rack - ...	4	45	120	199	180	379
2016-05-22	22	May	2016	47	Adults (35-64)	F	Australia	Victoria	Accessories	Bike Racks	Hitch Rack - ...	2	45	120	100	90	190
2014-02-22	22	February	2014	35	Adults (35-64)	M	Australia	Victoria	Accessories	Bike Racks	Hitch Rack - ...	22	45	120	1096	990	2086
2016-02-22	22	February	2016	35	Adults (35-64)	M	Australia	Victoria	Accessories	Bike Racks	Hitch Rack - ...	21	45	120	1046	945	1991
2013-07-30	30	July	2013	32	Young Adults ...	F	Australia	Victoria	Accessories	Bike Racks	Hitch Rack - ...	8	45	120	398	360	758
2015-07-30	30	July	2015	32	Young Adults ...	F	Australia	Victoria	Accessories	Bike Racks	Hitch Rack - ...	8	45	120	398	360	758
2013-07-15	15	July	2013	34	Young Adults ...	M	Australia	Victoria	Accessories	Bike Racks	Hitch Rack - ...	7	45	120	349	315	664
2015-07-15	15	July	2015	34	Young Adults ...	M	Australia	Victoria	Accessories	Bike Racks	Hitch Rack - ...	7	45	120	349	315	664
2013-08-02	2	August	2013	29	Young Adults ...	M	Canada	British Colum...	Accessories	Bike Racks	Hitch Rack - ...	5	45	120	369	225	594
2015-08-02	2	August	2015	29	Young Adults ...	M	Canada	British Colum...	Accessories	Bike Racks	Hitch Rack - ...	7	45	120	517	315	832
2013-09-02	2	September	2013	29	Young Adults ...	M	Canada	British Colum...	Accessories	Bike Racks	Hitch Rack - ...	2	45	120	148	90	238
2015-09-02	2	September	2015	29	Young Adults ...	M	Canada	British Colum...	Accessories	Bike Racks	Hitch Rack - ...	1	45	120	74	45	119
2014-01-22	22	January	2014	29	Young Adults ...	M	Canada	British Colum...	Accessories	Bike Racks	Hitch Rack - ...	1	45	120	74	45	119
2016-01-22	22	January	2016	29	Young Adults ...	M	Canada	British Colum...	Accessories	Bike Racks	Hitch Rack - ...	1	45	120	74	45	119
2014-05-17	17	May	2014	29	Young Adults ...	M	Canada	British Colum...	Accessories	Bike Racks	Hitch Rack - ...	6	45	120	443	270	713
2016-05-17	17	May	2016	29	Young Adults ...	M	Canada	British Colum...	Accessories	Bike Racks	Hitch Rack - ...	8	45	120	590	360	950
2014-03-27	27	March	2014	51	Adults (35-64)	M	United States	Oregon	Accessories	Bike Racks	Hitch Rack - ...	9	45	120	524	405	929
2016-03-27	27	March	2016	51	Adults (35-64)	M	United States	Oregon	Accessories	Bike Racks	Hitch Rack - ...	7	45	120	407	315	722
2013-08-25	25	August	2013	49	Adults (35-64)	M	Canada	British Colum...	Accessories	Bike Racks	Hitch Rack - ...	3	45	120	221	135	356
2015-08-25	25	August	2015	49	Adults (35-64)	M	Canada	British Colum...	Accessories	Bike Racks	Hitch Rack - ...	1	45	120	74	45	119

Figure 10 - Explore Graph Results

5.2.1 Univariate Analysis

Descriptive or inferential analysis of one variable is called univariate analysis. The purpose of univariate analysis is to have a first glimpse of each variable and perform initial data quality assessment. Histogram of each interval variable is plotted below.

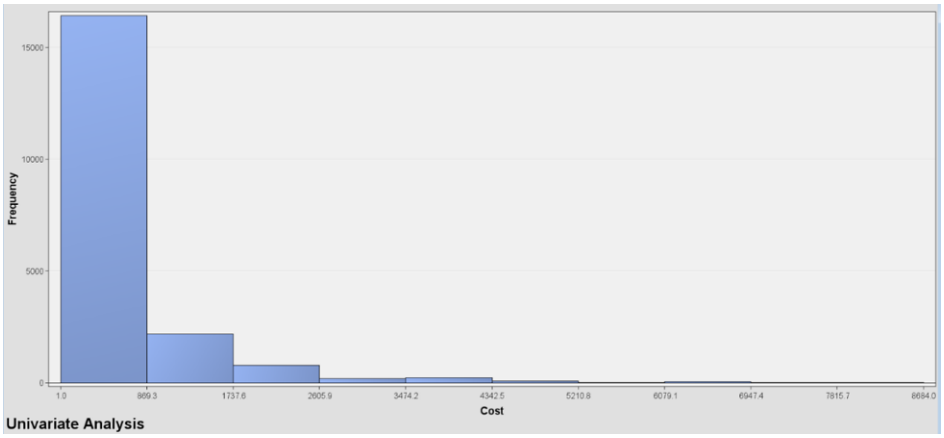


Figure 11 - Univariate Analysis of Cost

According to **Figure 11**, most products' costs are in the range of 1$ - 870$ range.

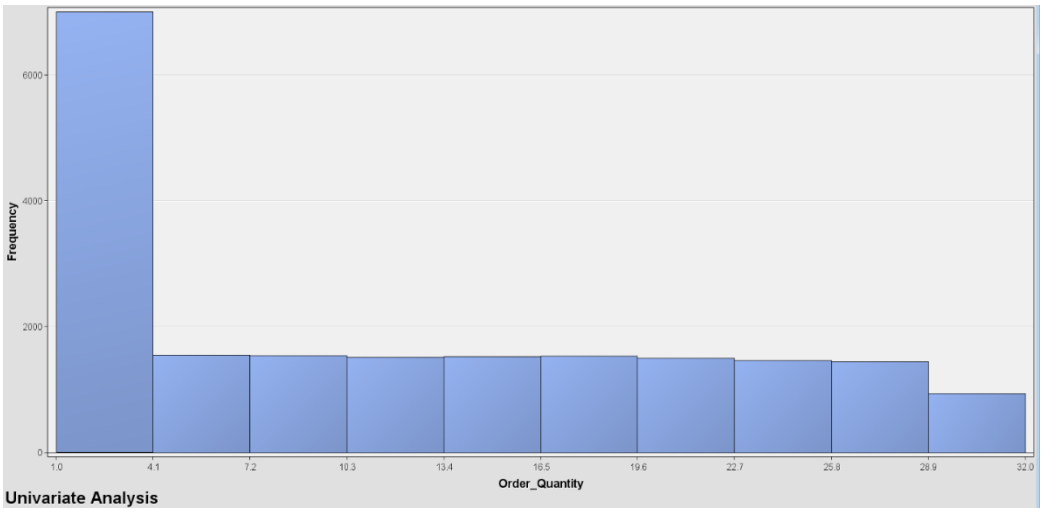


Figure 12 - Univariate Analysis of Order Quantity

Figure 12 shows that the most frequent order quantity is usually 1-4 items. On the other hand, the frequency of 5 and above items is around 1800 count.

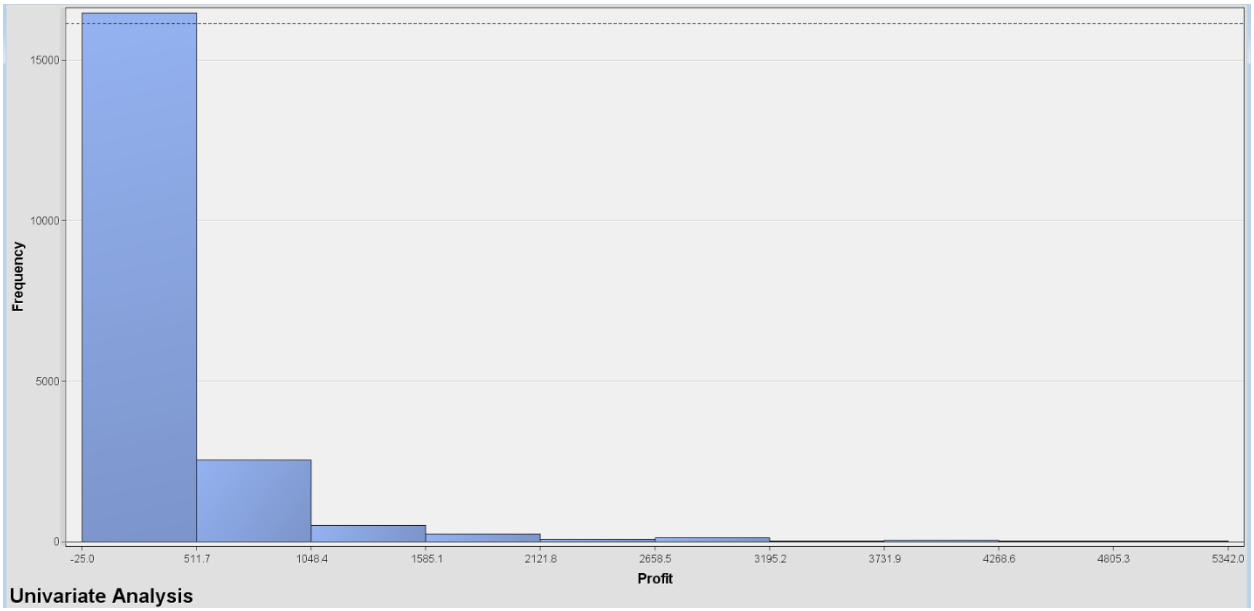


Figure 13 - Univariate Analysis of Profit

Figure 13 shows the profits distribution in our analysis. It is positively skewed to the right, and most transactions have a profit lower than 511$

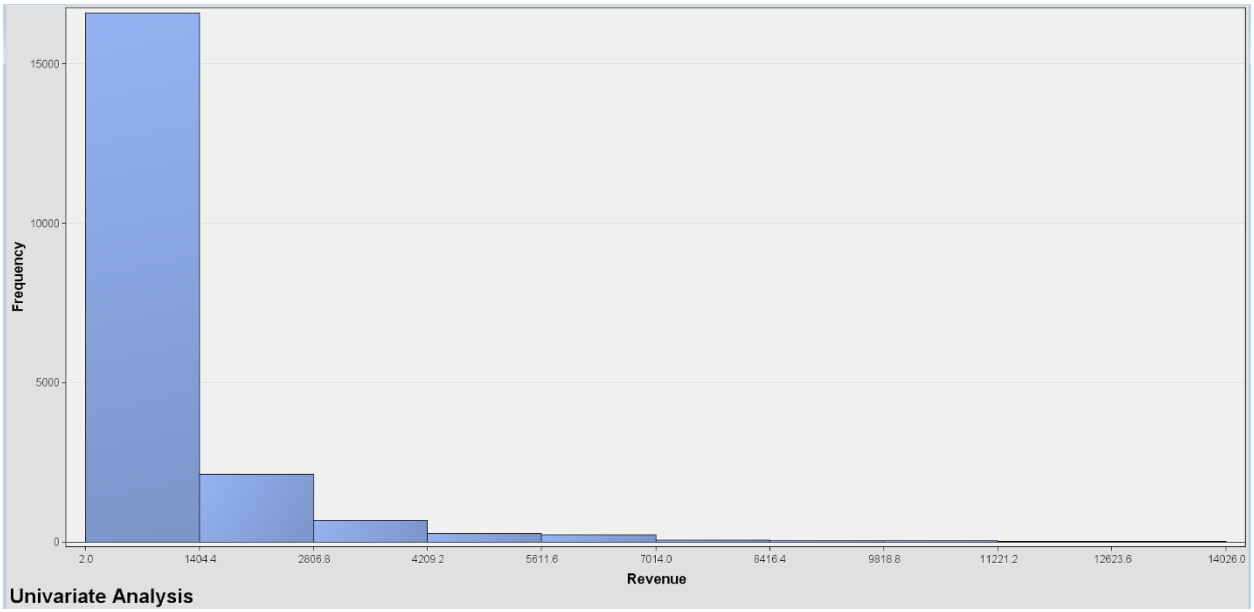


Figure 14 - Univariate Analysis of Revenue

Figure 14 shows revenue, we can observe that revenue is highly correlated with profits as profits is revenue – cost.

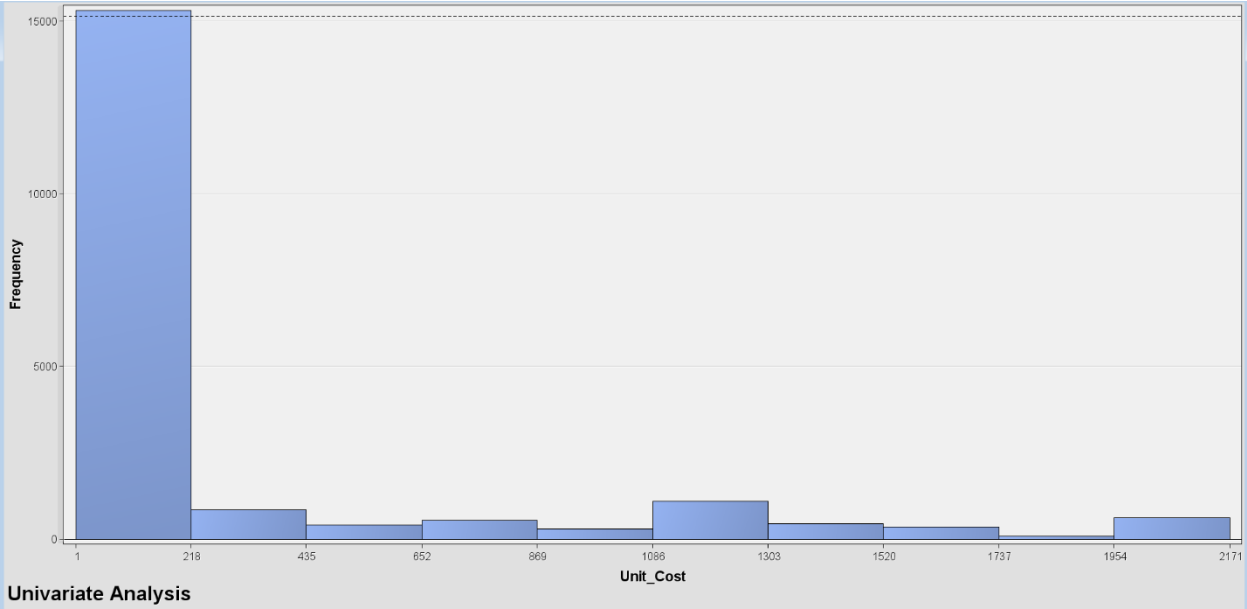


Figure 15 - Univariate Analysis of Unit Cost

As shown in **Figure 15**, the unit cost column is also positively skewed to the right, in the same pattern as revenue and profits.

There are missing bins in unit price variable in **Figure 16.**

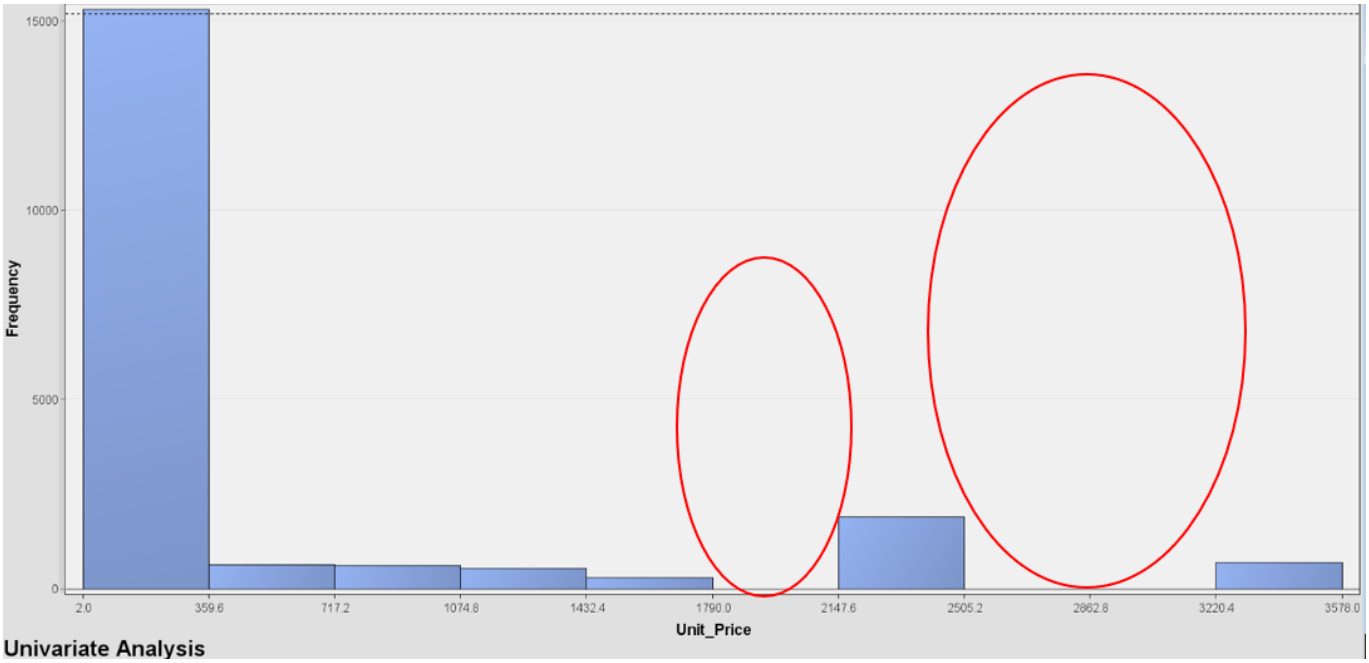


Figure 16 - Univariate Analysis of Unit Price

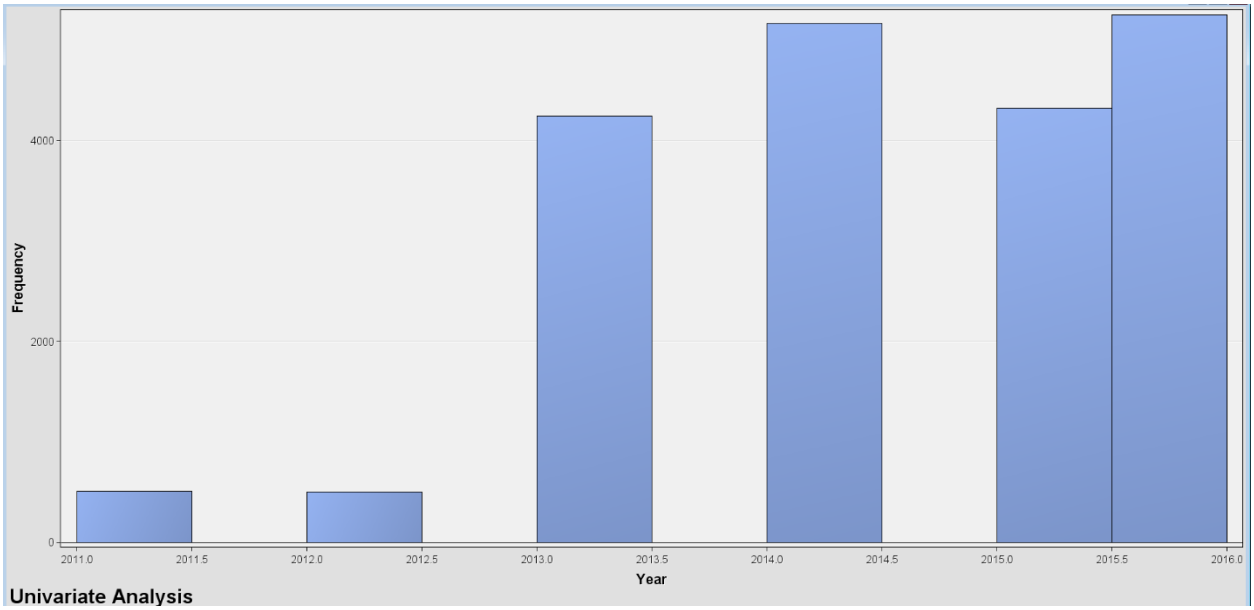


Figure 17 - Univariate Analysis of Year (Default View)

Figure 17 presents the frequency of transactions based on the year; we clearly observe that the market demand has increased with the years.

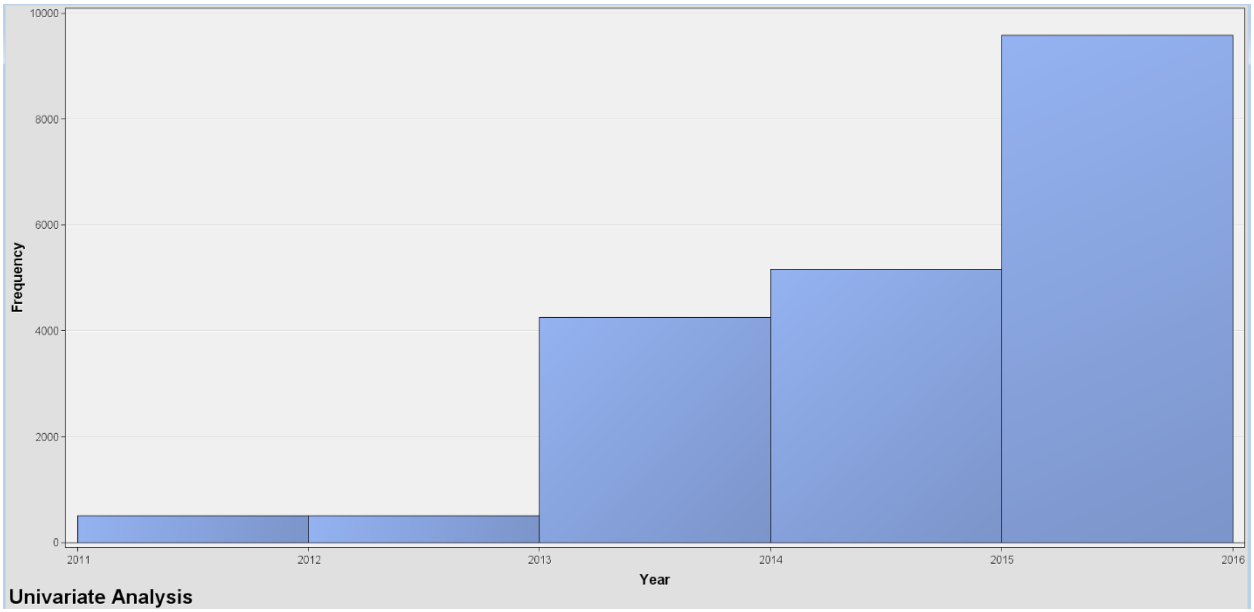


Figure 18 - Univariate Analysis of Year after Changing Graph Properties

Figure 18 presents the frequency of transactions based on the year after changing number of bins for the year variable; so that we can clearly observe that the market demand has increased with the years.

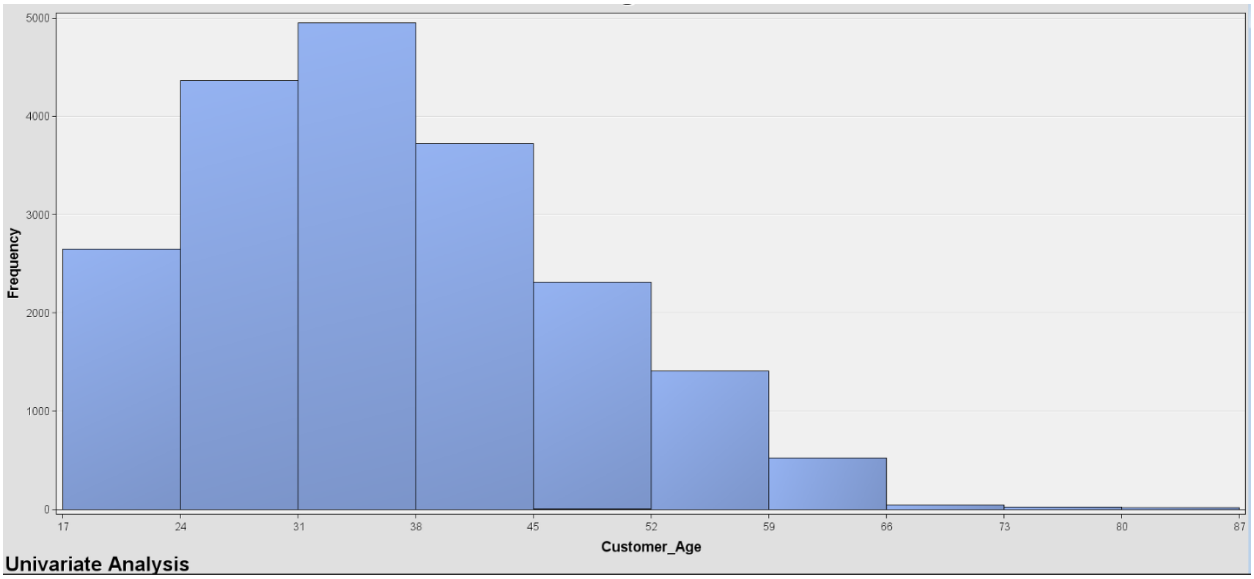


Figure 19 – Univariate Analysis of Customer Age

Figure 19 presenting the distribution of costumers' ages. Most costumers are from the range 24 to 45 years old.

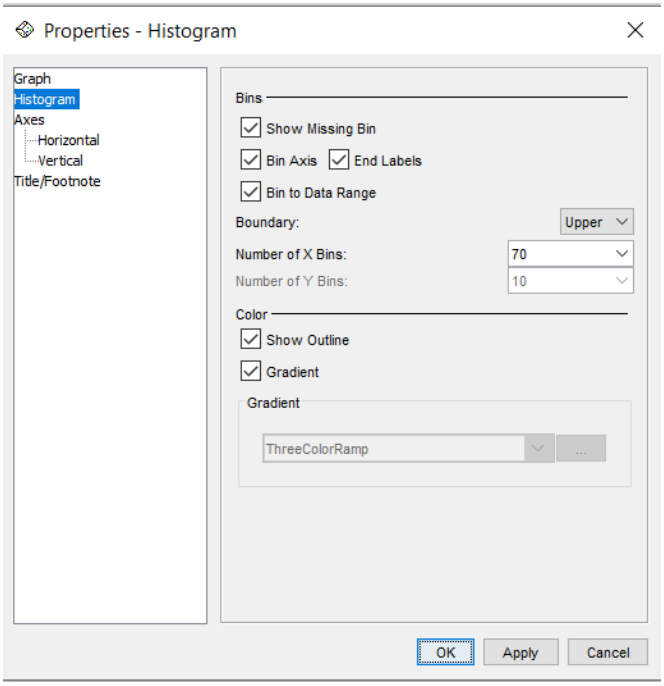


Figure 20 – Histogram Properties

Figure 20 shows the way to change graph properties.

There are missing bins in Customer Age variable in **Figure 21**.

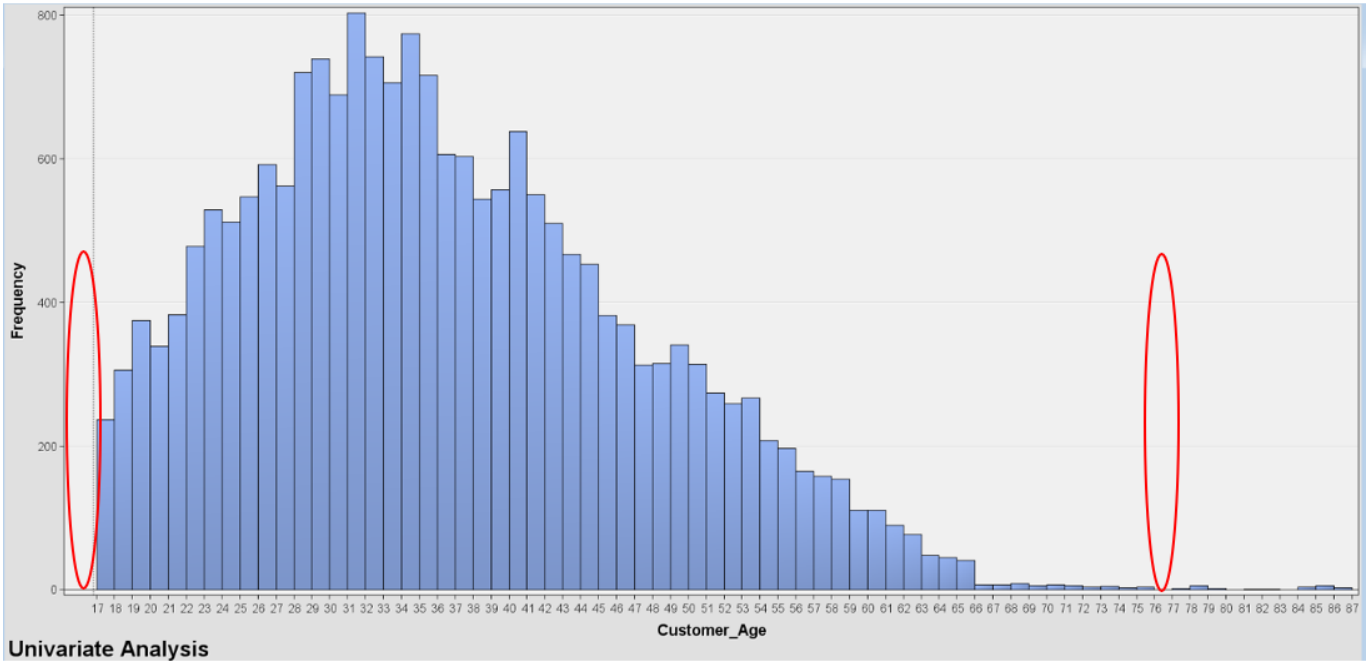


Figure 21 - Univariate Analysis of Customer Age after Changing Graph Properties

There are many values of profit variable at extreme right and outside of IQR fences which shows presence of outliers as shown in **Figure 22.**

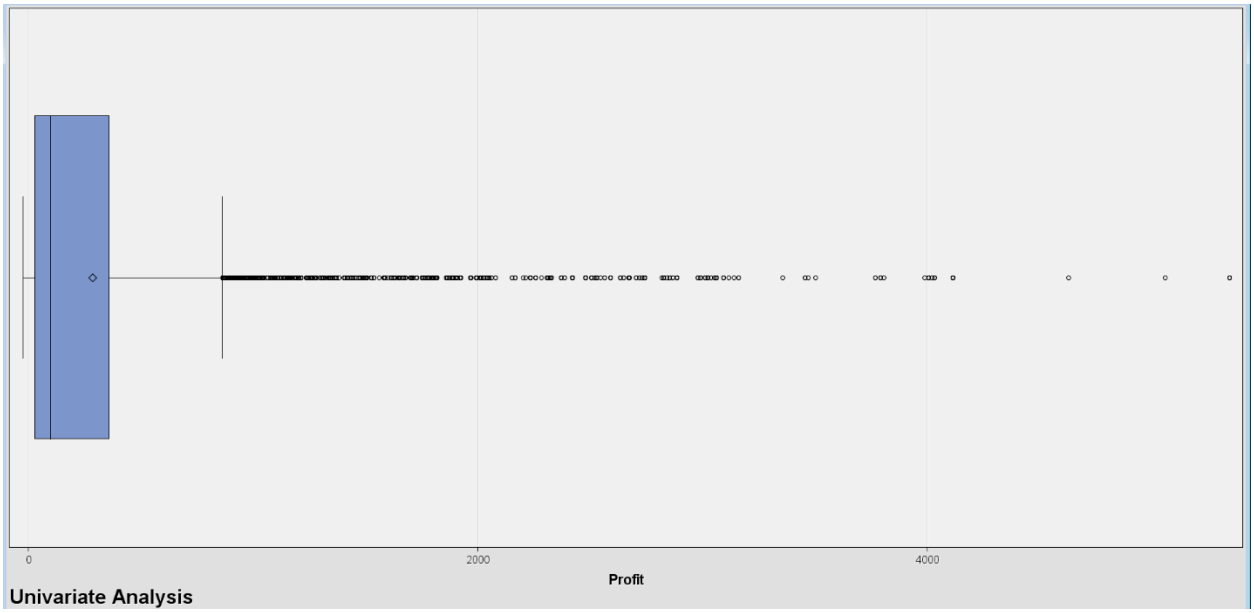


Figure 22 - Box Plot of Profit

There are many values of revenue variable at extreme right and outside of IQR fences which shows presence of outliers as shown in **Figure 23.**

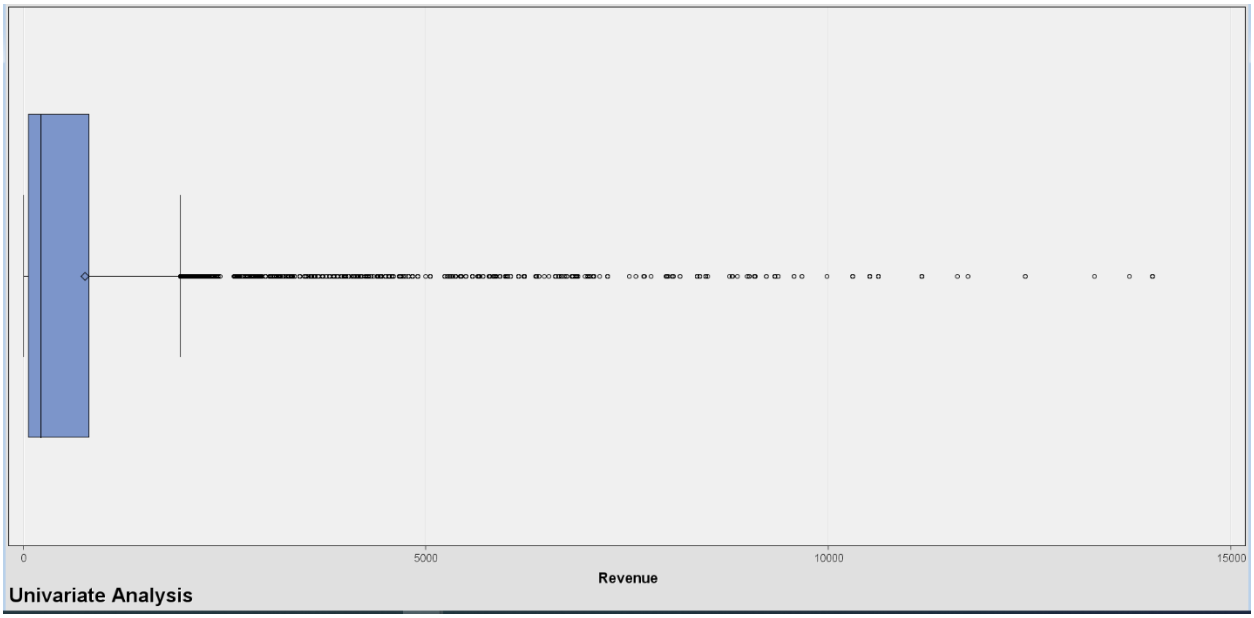


Figure 23 - Box Plot of Revenue

Figure 23 demonstrates box plot distribution of revenue; it is similar to **Figure 14** that shows the skewness of the column.

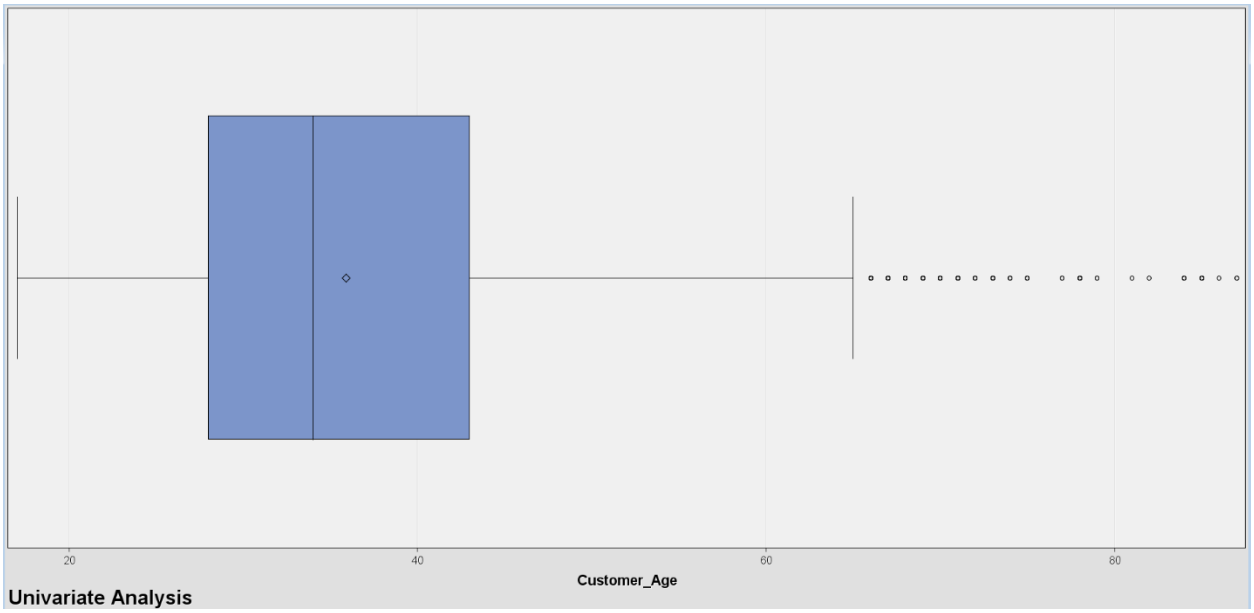


Figure 24 - Box Plot of Customer Age

Figure 24 shows the customer age distribution using the box plot method. The line in the middle of the blue square represents the median age, i.e., 36 years old.

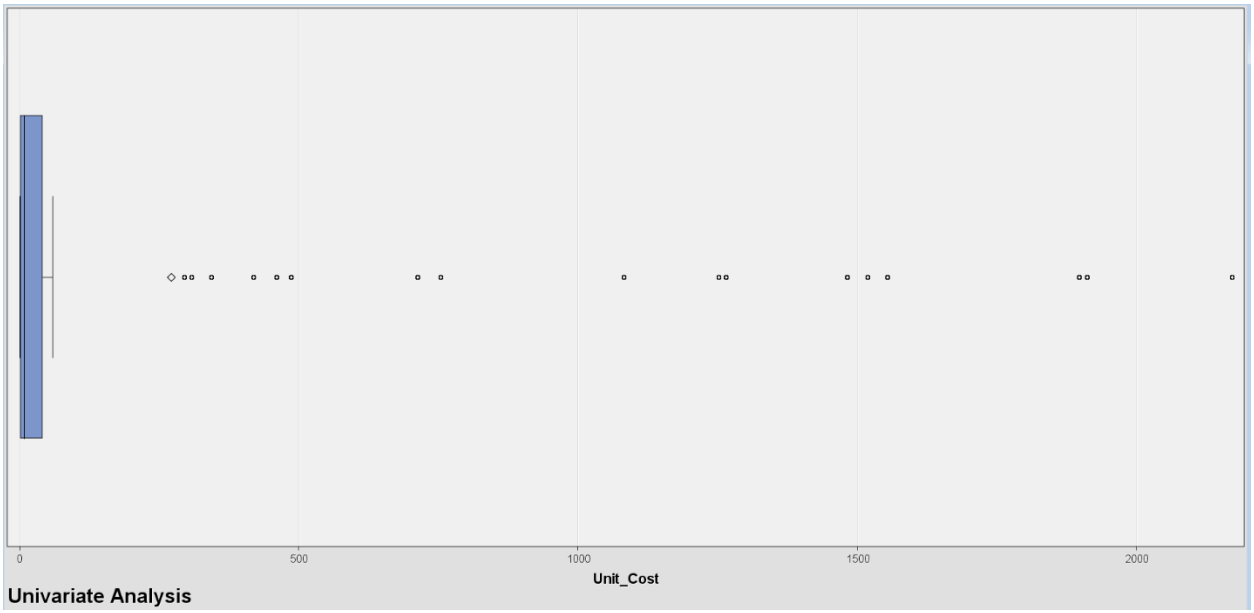


Figure 25 - Box Plot of Unit Cost

Figure 25 shows box plot of unit cost variable, there are some products which have high cost.

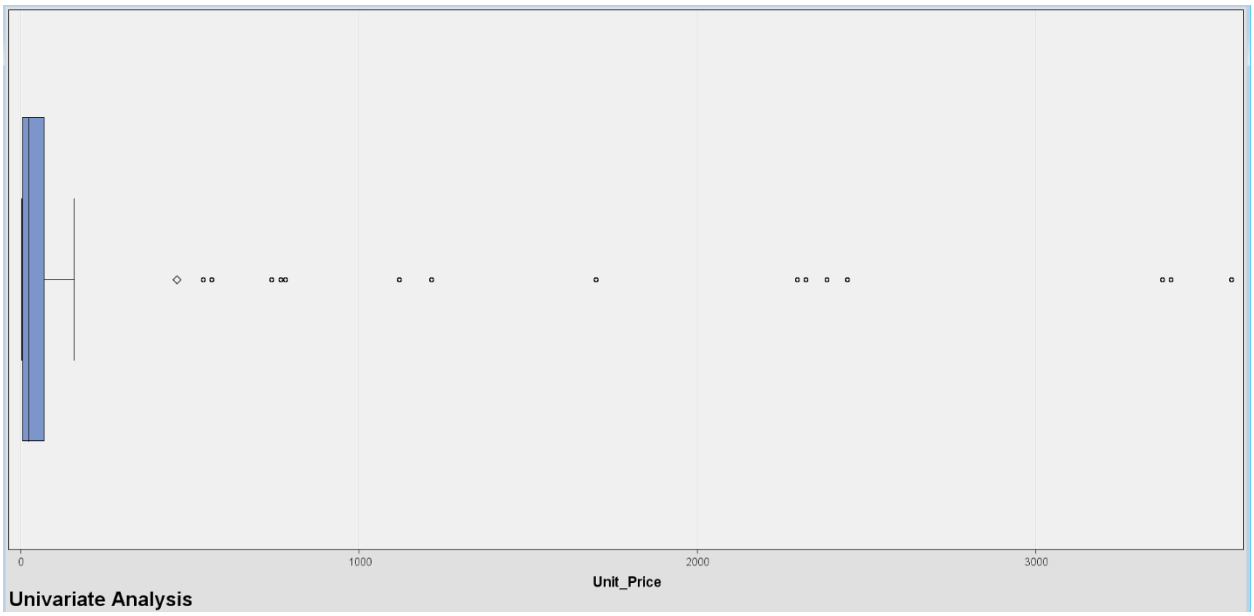


Figure 26 - Box Plot of Unit Price

Figure 26 shows box plot of unit price variable, there are some products which have high price. These products are the same as which have high unit cost.

Pie charts of categorical variables including year, gender, product category, country, subcategory and age group are shown in **Figure 27**.

Figure 27 - Categorical Variables Pie Charts

Profit Outliers Association with categorical variables is shown in **Figure 28**.

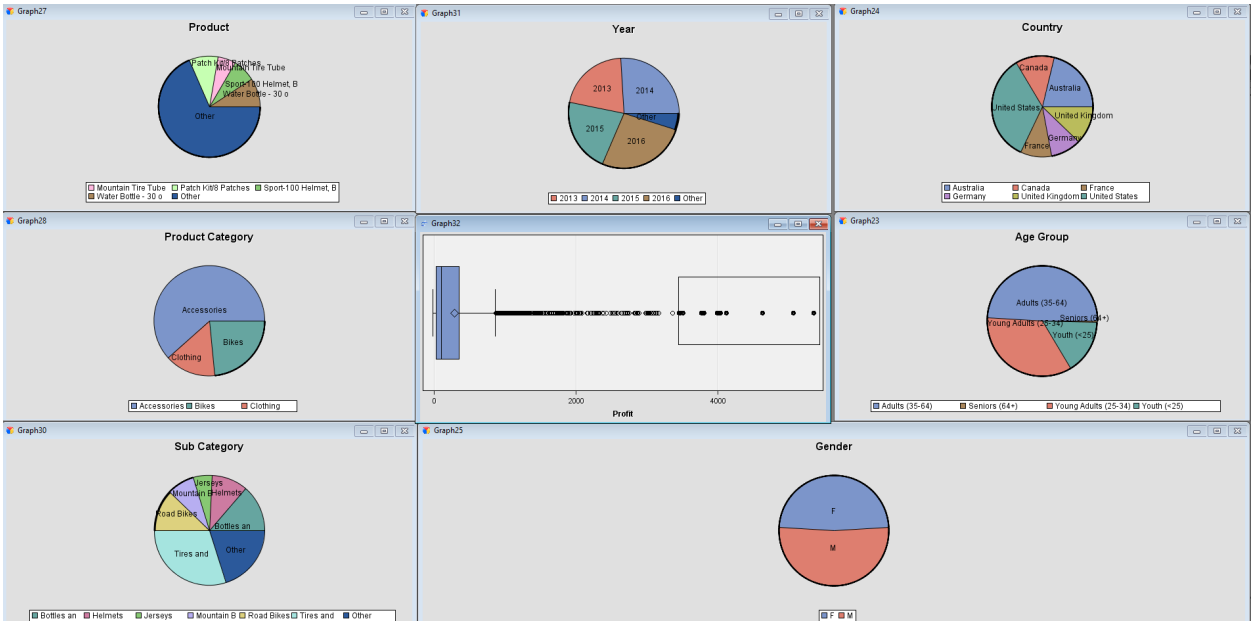


Figure 28 - Profit Outliers Association

Profit Outliers Association with Product Category is more clearly visible and shown in **Figure 29**. It represents that Bikes profit is more than Clothing and Accessories.

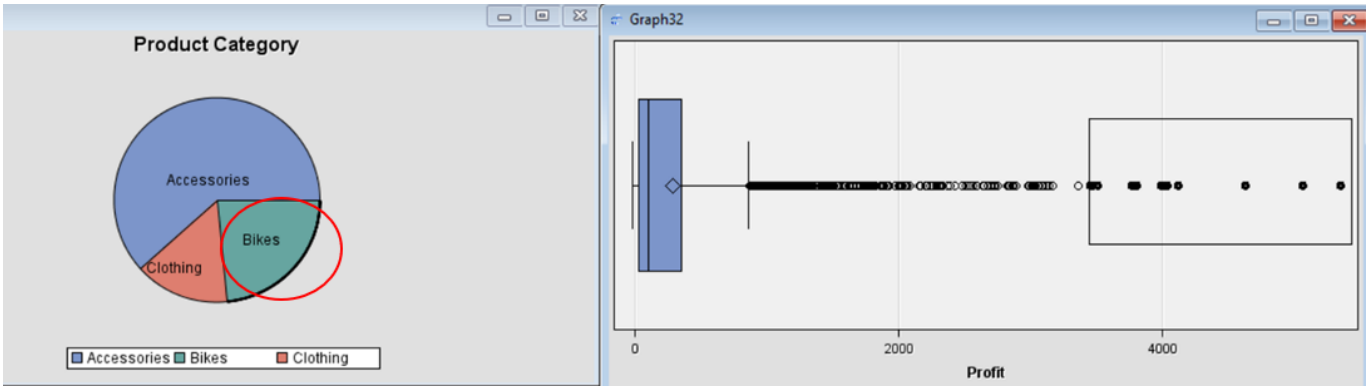


Figure 29 - Profit Outliers Association Product Category

Profit Range Association with categorical variables is shown in **Figure 30**.

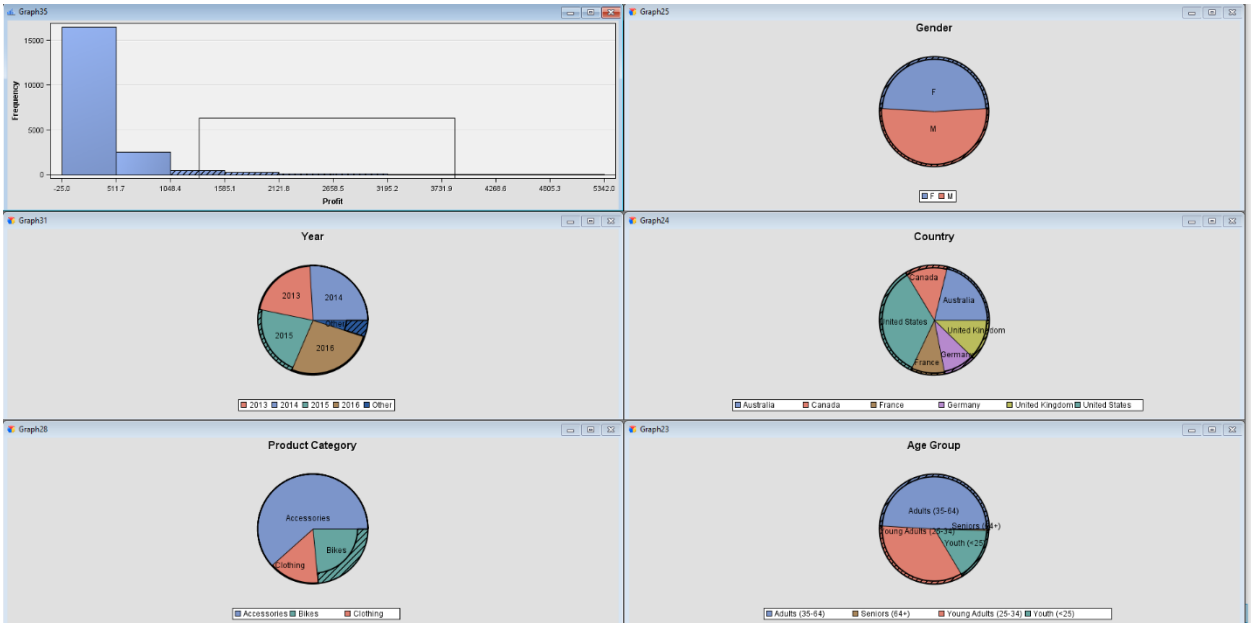


Figure 30 - Profit Range Association

Revenue Range Association with categorical variables is shown in **Figure 31**.

Figure 31 - Revenue Range Association

Unit Price Range Association with categorical variables is shown in **Figure 32**.

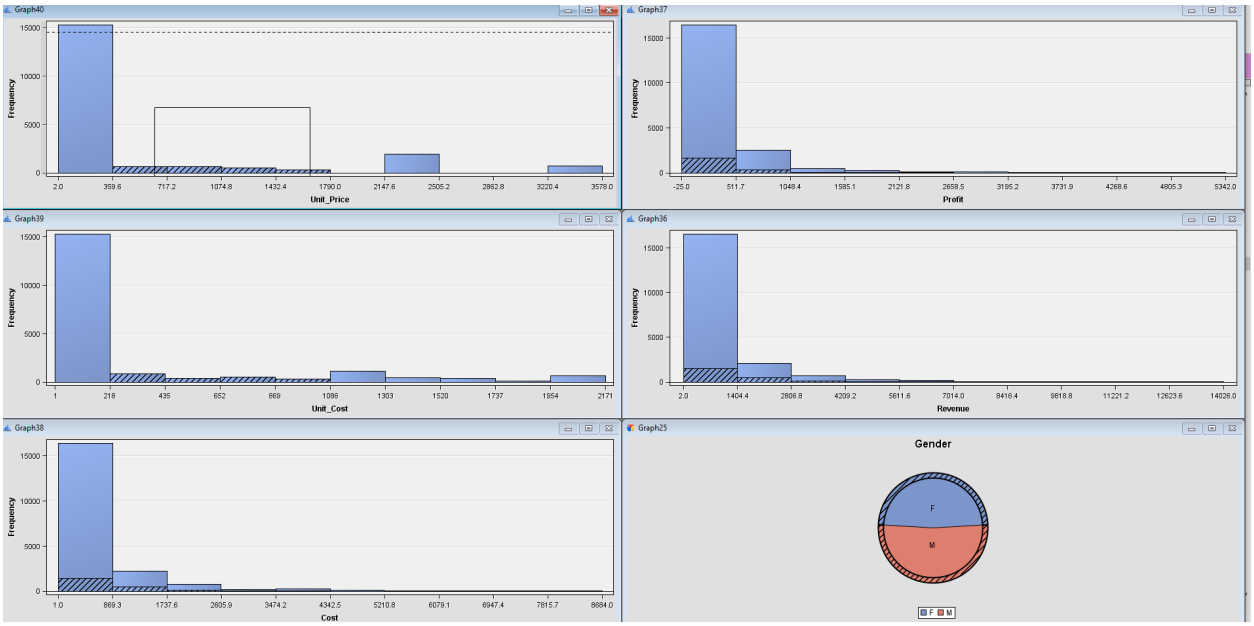


Figure 32 - Unit Price Range Association

We have found from the univariate analysis that most of the attributes do not have normal distribution of data. Data scaling and normalization techniques can be applied to improve the data quality.

5.2.2 Bivariate Analysis

The purpose of bivariate analysis to identify relationship between two variables.

Redundant Variable

The Variable Clustering Algorithm of SAS Enterprise Miner has been used to identify redundancy in variables.

Drag and drop a ‘Sample’ node and connect it to data source. Drag and drop a ‘Variable Clustering’ node, connect it to Sample node and run it. The results are shown from **Figure 33** to **Figure 36**.

One cluster includes year, date and customer age.

Second cluster includes cost, revenue, order quantity, unit price and day.

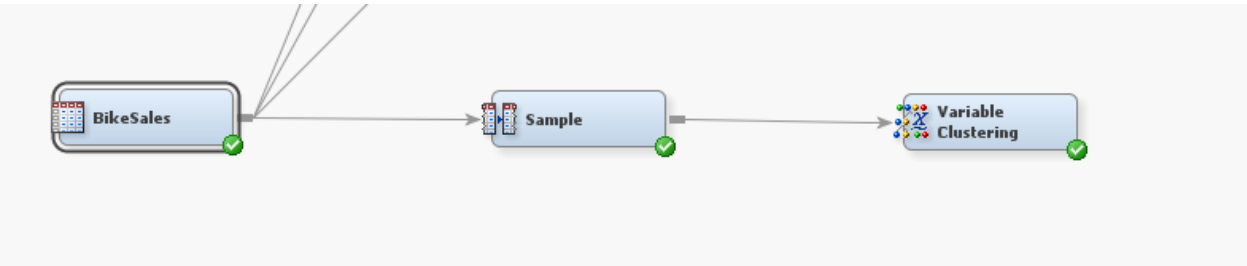


Figure 33 - Sample and Variable Clustering

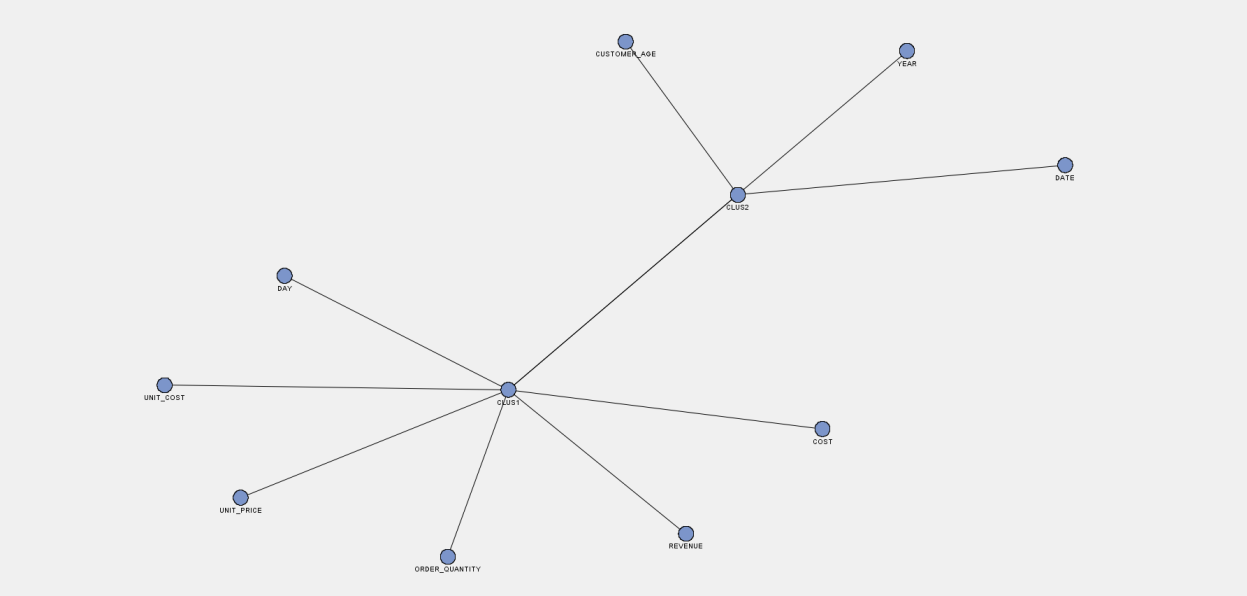


Figure 34 - Variable Cluster Graph

Variable Frequency Table

Cluster	Frequency Count	Percent of Total Frequency
CLUS1	6	66.66667
CLUS2	3	33.33333

Selected Variables

Cluster	Variable	R-Square With Own Cluster Component	Next Closest Cluster	R-Square with Next Cluster Component	Type	Label	1-R2 Ratio	Variable Selected
CLUS1	CLUS1	1	CLUS2	0.061224	ClusterComp	Cluster 1	0	YES
CLUS2	CLUS2	1	CLUS1	0.061224	ClusterComp	Cluster 2	0	YES

Figure 35 - Clusters Detail

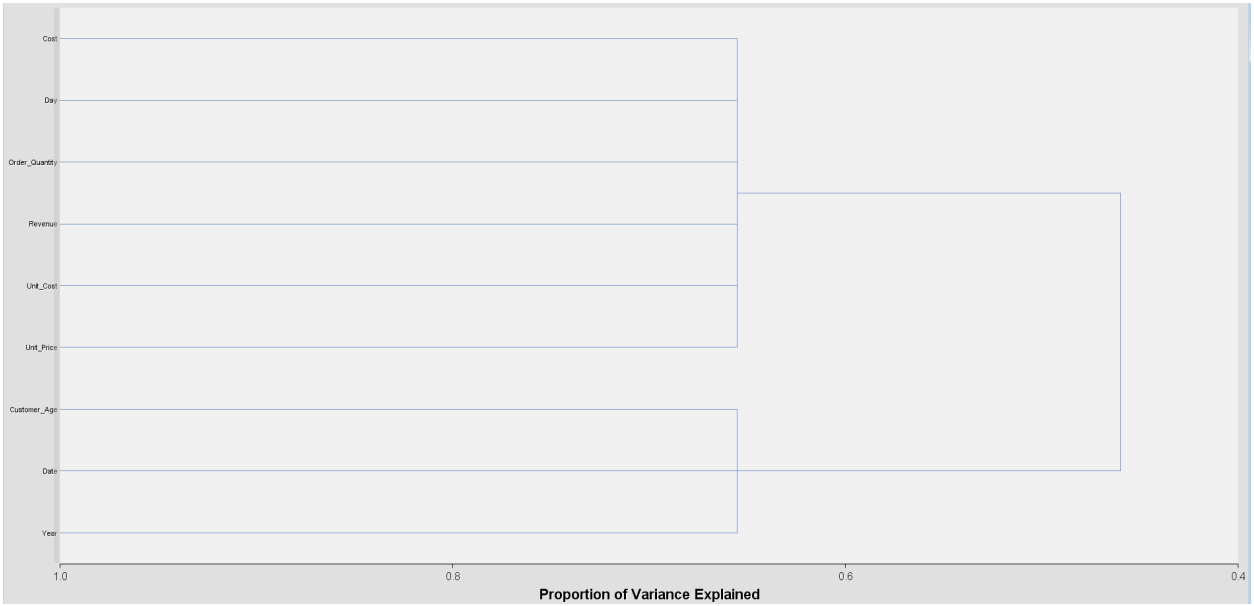


Figure 36 – Dendrogram (Variable Association)

Relevant Variable

Variable Selection Algorithm of SAS Enterprise Miner has been used to identify the relevant variables. From **Figure 37**, we can see that the variable revenue and cost are displayed as relevant variables for the target variable profit. More details are displayed in **Figure 39**.

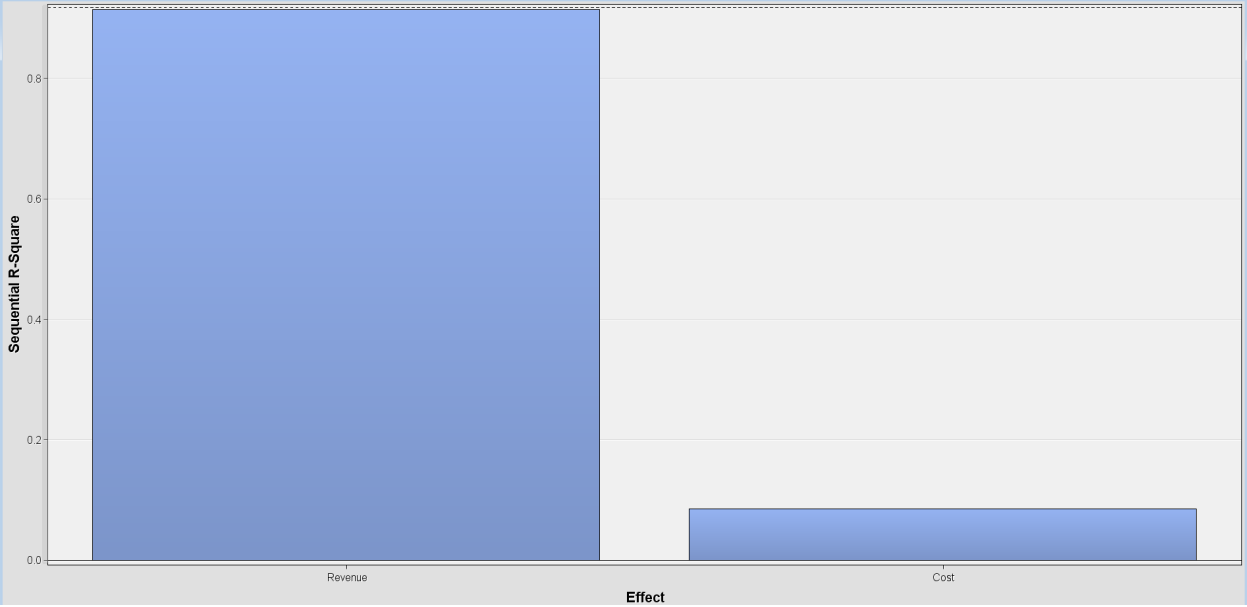


Figure 37 - Variable Selection (Effects in the Model)

The DMINE Procedure

Effects Chosen for Target: Profit

Effect	DF	R-Square	F Value	p-Value	Sum of Squares	Error Mean Square
Var: Revenue	1	0.915029	1217237	<.0001	21308079009	17505
Var: Cost	1	0.084971	.	.	1978692113	.

The DMINE Procedure

The Final ANOVA Table for Target: Profit

Effect	DF	R-Square	Sum of Squares
Model	2	1.000000	23286771122
Error	113033	.	0
Total	113035	.	23286771122

The DMINE Procedure

Effects Not Chosen for Target: Profit

Effect	DF	R-Square	F Value	p-Value	Sum of Squares
Group: Product	5	1.663903E-30	0	1.0000	3.874693E-20
Var: Unit_Price	1	2.595041E-32	0	1.0000	6.043013E-22
Var: Unit_Cost	1	1.012918E-31	0	1.0000	2.358759E-21
Group: Sub_Category	4	1.74309E-30	0	1.0000	4.059095E-20
Group: Product_Category	1	1.811933E-31	0	1.0000	4.219406E-21
Var: Order_Quantity	1	3.877006E-31	0	1.0000	9.028296E-21
Var: Date	1	4.91123E-33	0	1.0000	1.143667E-22
Var: Year	1	1.434892E-26	0	1.0000	3.341401E-16
Group: State	6	5.310804E-31	0	1.0000	1.236715E-20

Figure 38 - Variable Selection (R2 Values)

Variable Selection

Variable Name	Role	Measurement Level	Type
Age Group	Rejected	Nominal	Character
Cost	Input	Interval	Numeric
Country	Rejected	Nominal	Character
Customer Age	Rejected	Interval	Numeric
Customer Gender	Rejected	Binary	Character
Date	Rejected	Interval	Numeric
Day	Rejected	Interval	Numeric
Month	Rejected	Nominal	Character
Order Quantity	Rejected	Interval	Numeric
Product	Rejected	Nominal	Character
Product Category	Rejected	Nominal	Character
Revenue	Input	Interval	Numeric
State	Rejected	Nominal	Character
Sub Category	Rejected	Nominal	Character
Unit Cost	Rejected	Interval	Numeric
Unit Price	Rejected	Interval	Numeric
Year	Rejected	Interval	Numeric

Figure 39 - Variable Selection Result

Customer Age and Profit scatter plot shows relationship between the variables in **Figure 40**. The highest profit is earned from customers of age range between 30 to 60.

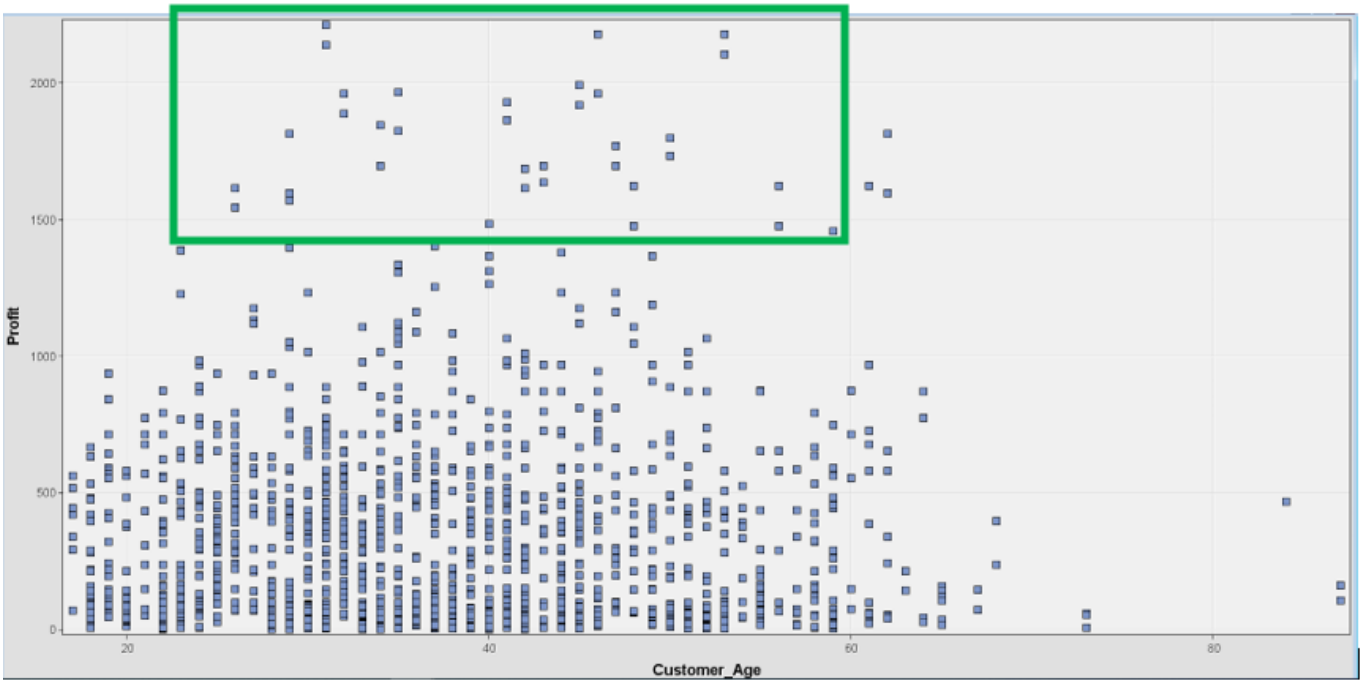


Figure 40 - Customer Age and Profit

Customer Age and Order Quantity scatter plot shows relationship between the variables in **Figure 41**. The highest number of orders were received from customers of the age range between 30 to 60.

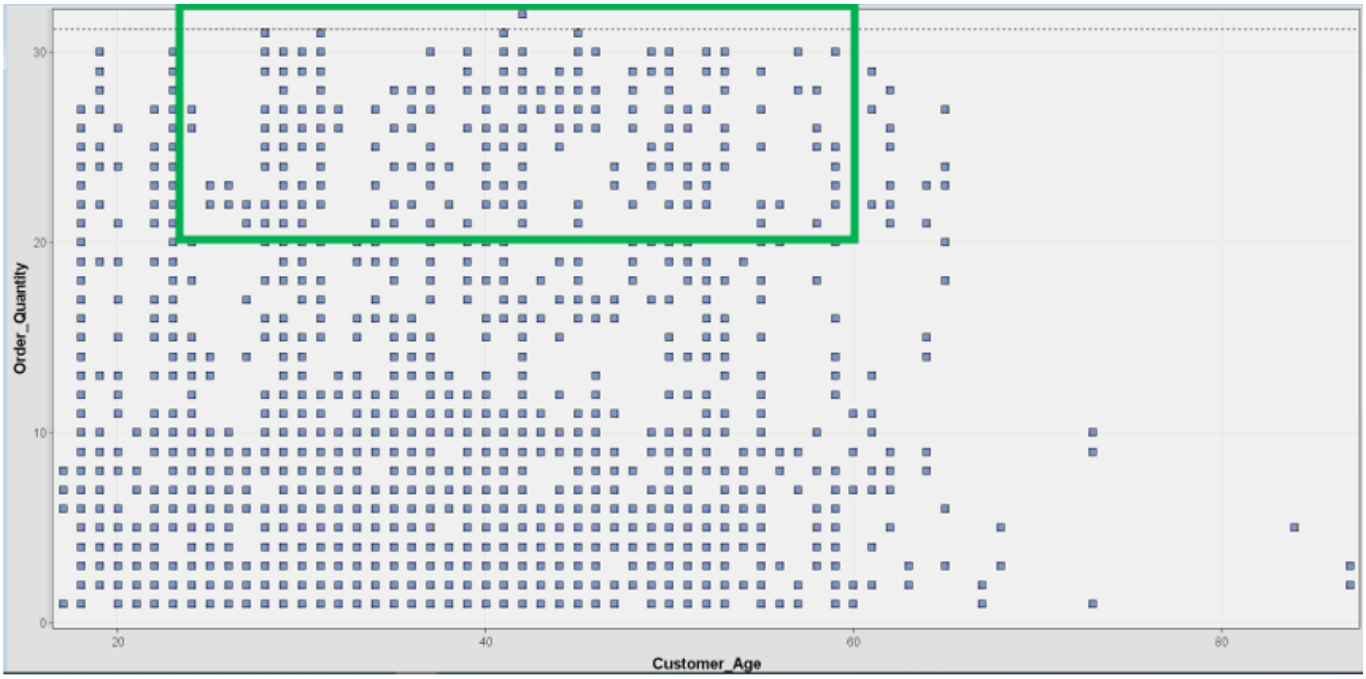


Figure 41 - Order Quantity and Customer Age

Customer Age and Revenue scatter plot shows relationship between the variables in **Figure 42**. The highest revenue received from customers of age range between 30 to 60.

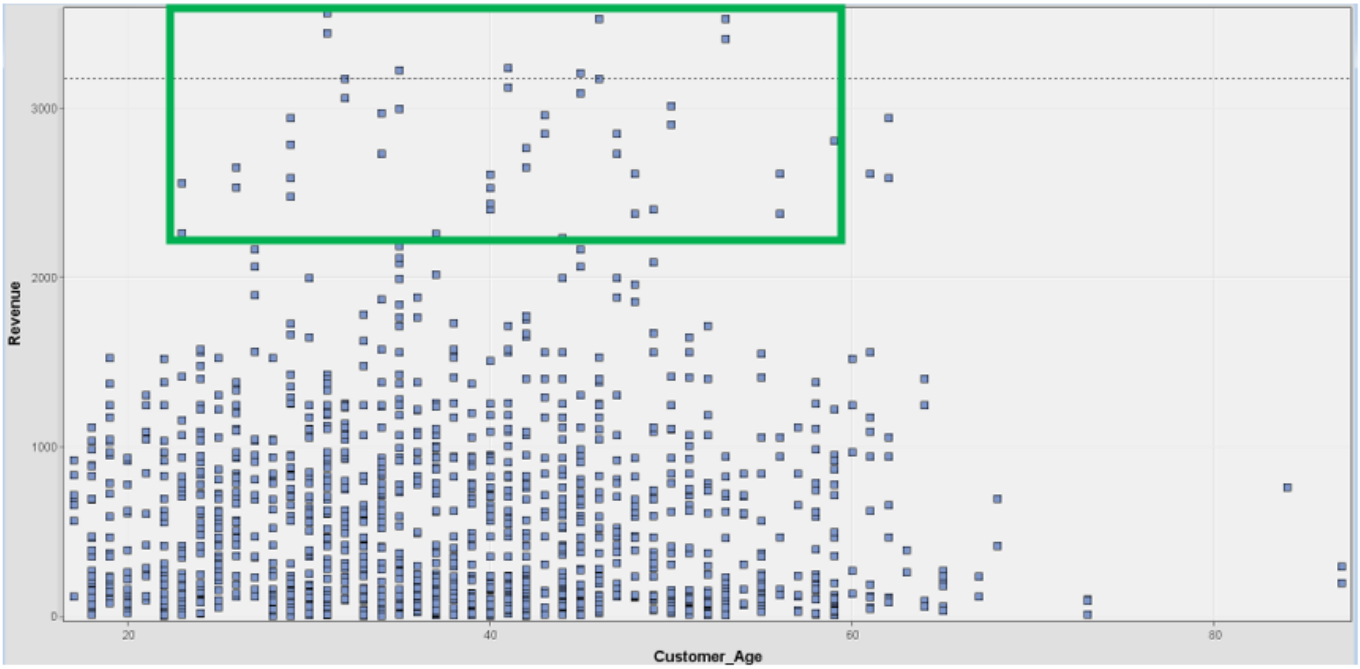


Figure 42 - Customer Age and Revenue

The result of bivariate analysis shows that customers of age between 30 to 60 are the most significant as sales have highest profit, revenue and order quantity from this age group. This is something we can highly focus on for future marketing strategies.

5.2.3 Multivariate Analysis

In this next step, Multivariate Analysis was conducted so more than one variable can be observed and analyzed at a time. Drag and drop 'StatExplore' node, connect it to data source and run it. The results are shown.

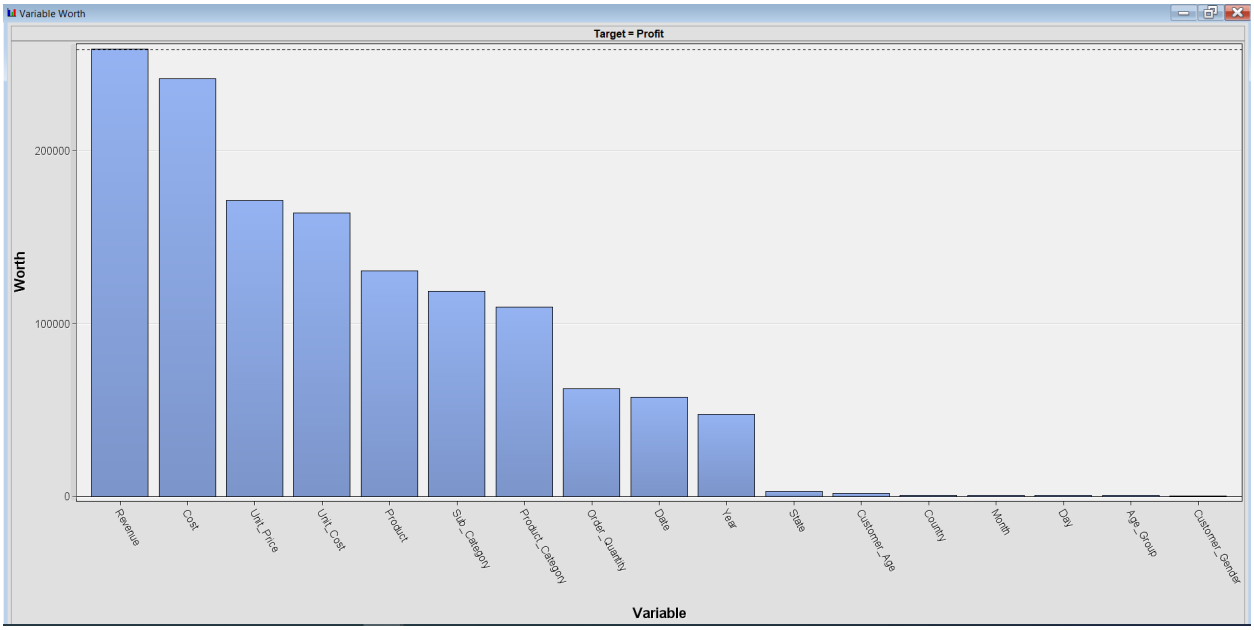


Figure 43 - Variable Worth

In **Figure 43 above**, Variable worth shows that Revenue and cost have the most worth in the dataset. Customer gender and age group have the least worth in the dataset.

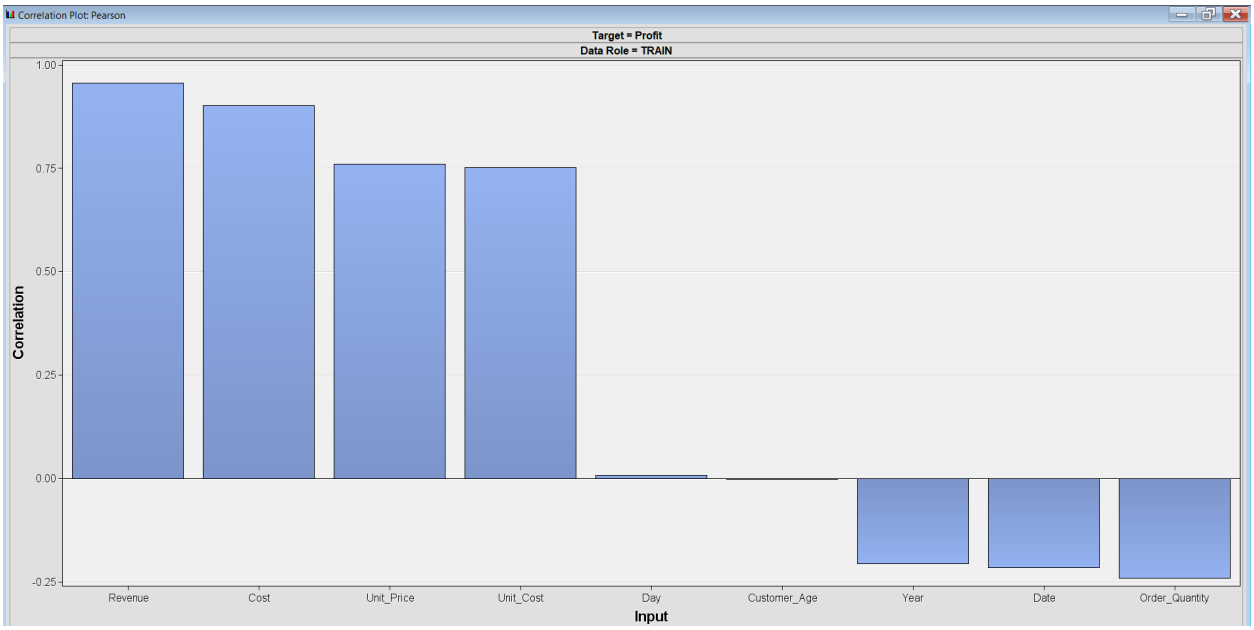


Figure 44 - Pearson Correlation Plot

In **Figure 44 above**, Pearson correlation plot shows that Revenue, cost, unit price and unit cost variables have positive correlation. Customer age variable has almost zero correlation. Year, date and order quantity variables have negative correlation.

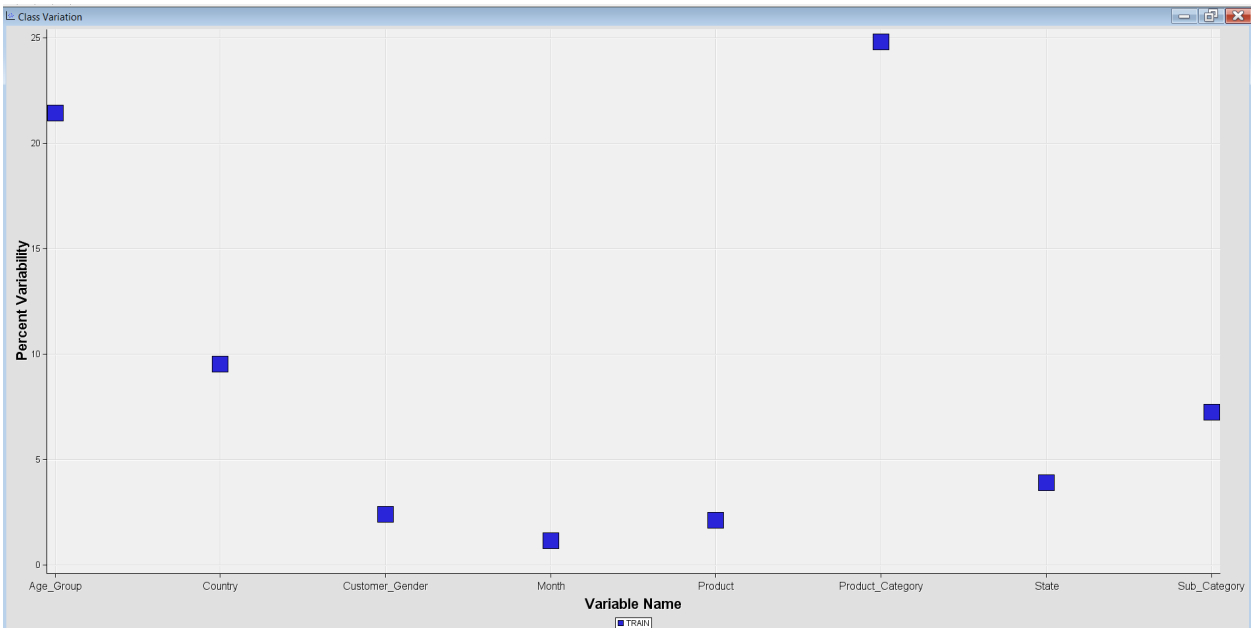


Figure 45 - Class Variation

Based on **Figure 45** above, Product Category variable has the highest percentage of variation.

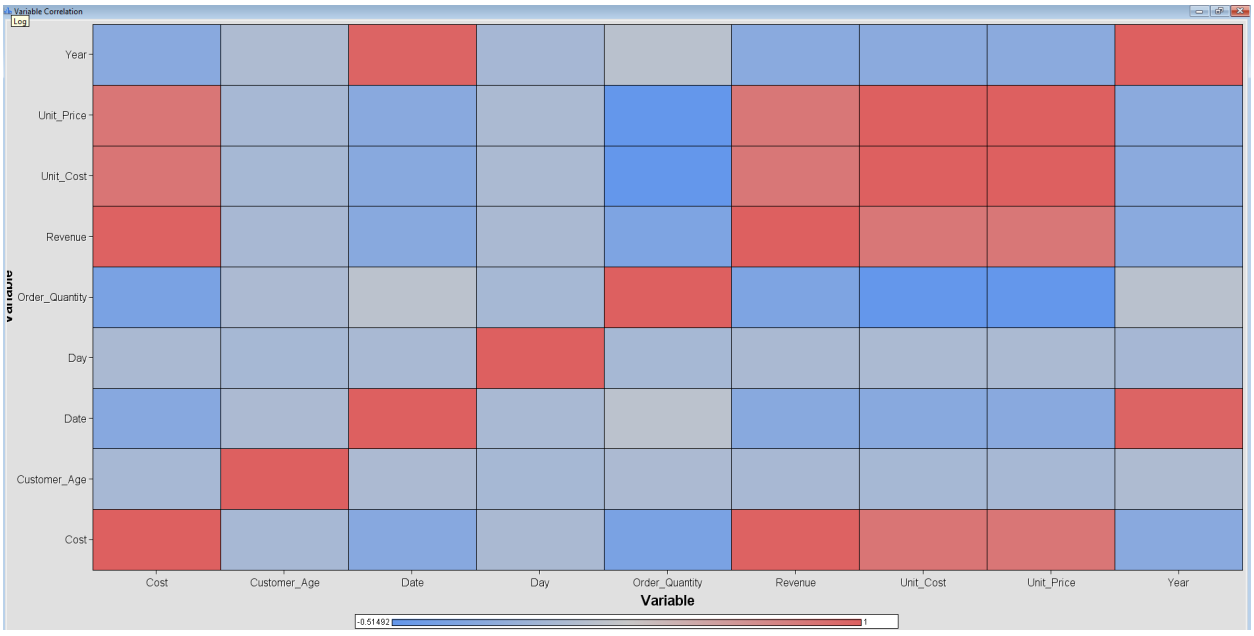


Figure 46 - Variable Correlation

Based on **Figure 46,** revenue and cost variables have a very strong positive correlation while Unit Price and Order Quantity has a strong negative correlation.

5.2.4 Interesting Visualization

Which is the highest and lowest profit earning products in all countries?

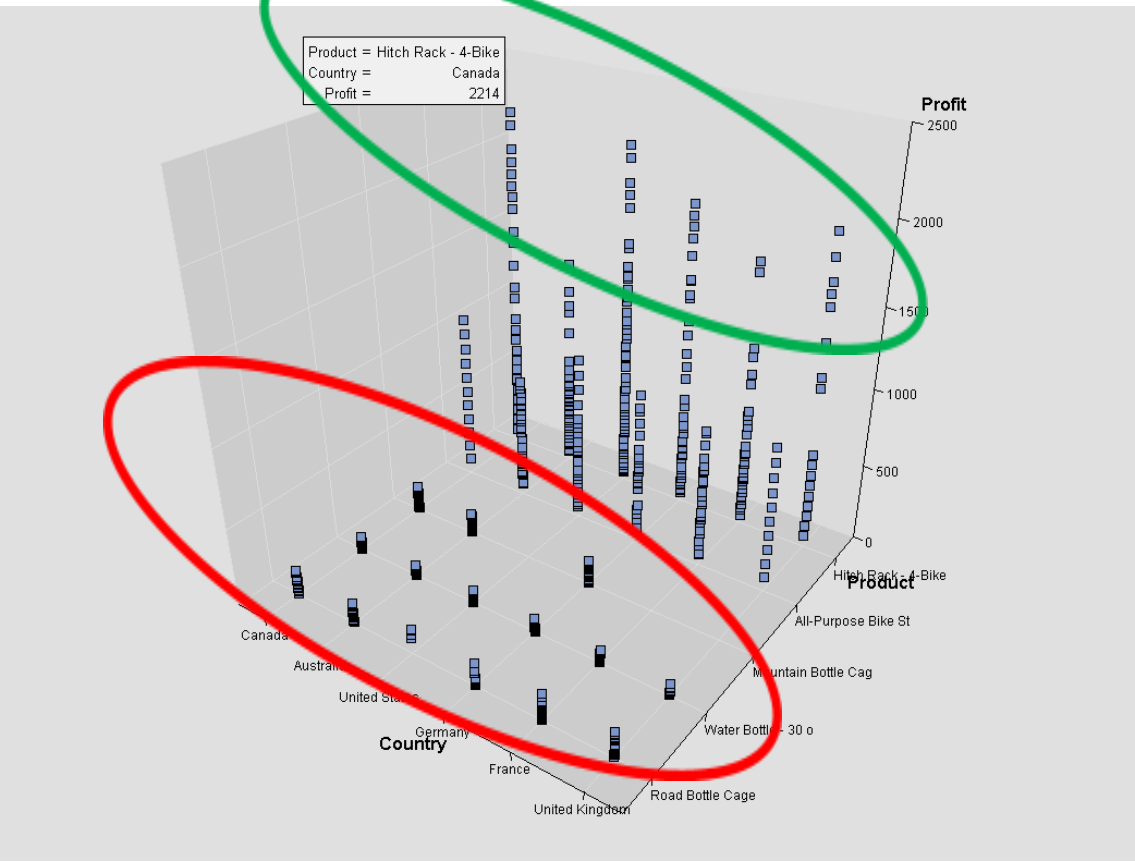


Figure 47 - Profit by Country and Product

According to **Figure 47**, it is clear that Hitch Rack – 4 Bike is the highest profit earning product in all countries. Mountain Bottle cag, Road Bottle Cage and Water Bottle are the lowest profit earning products in all countries.

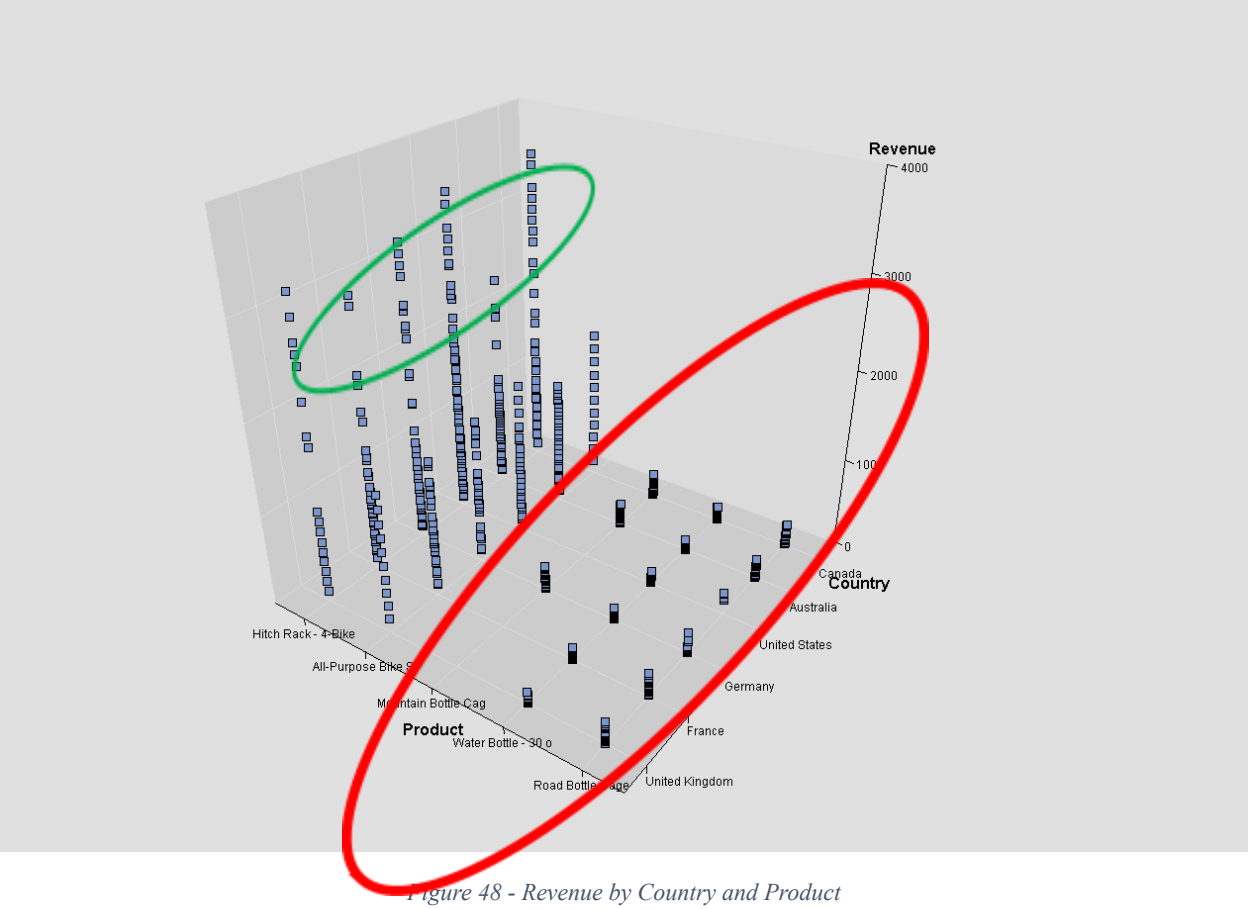


Figure 48 - Revenue by Country and Product

According to **Figure 48,** Canada generated the highest revenue with the product Hitch Rack – 4 Bike. Just like in **Figure 47**, Mountain Bottle Cag, Water Bottle and Road Bottle Cage generated the lowest revenue consistently throughout the countries listed. Interestingly enough, the United Kingdom did not generate any revenue for Mountain Bottle Cag.

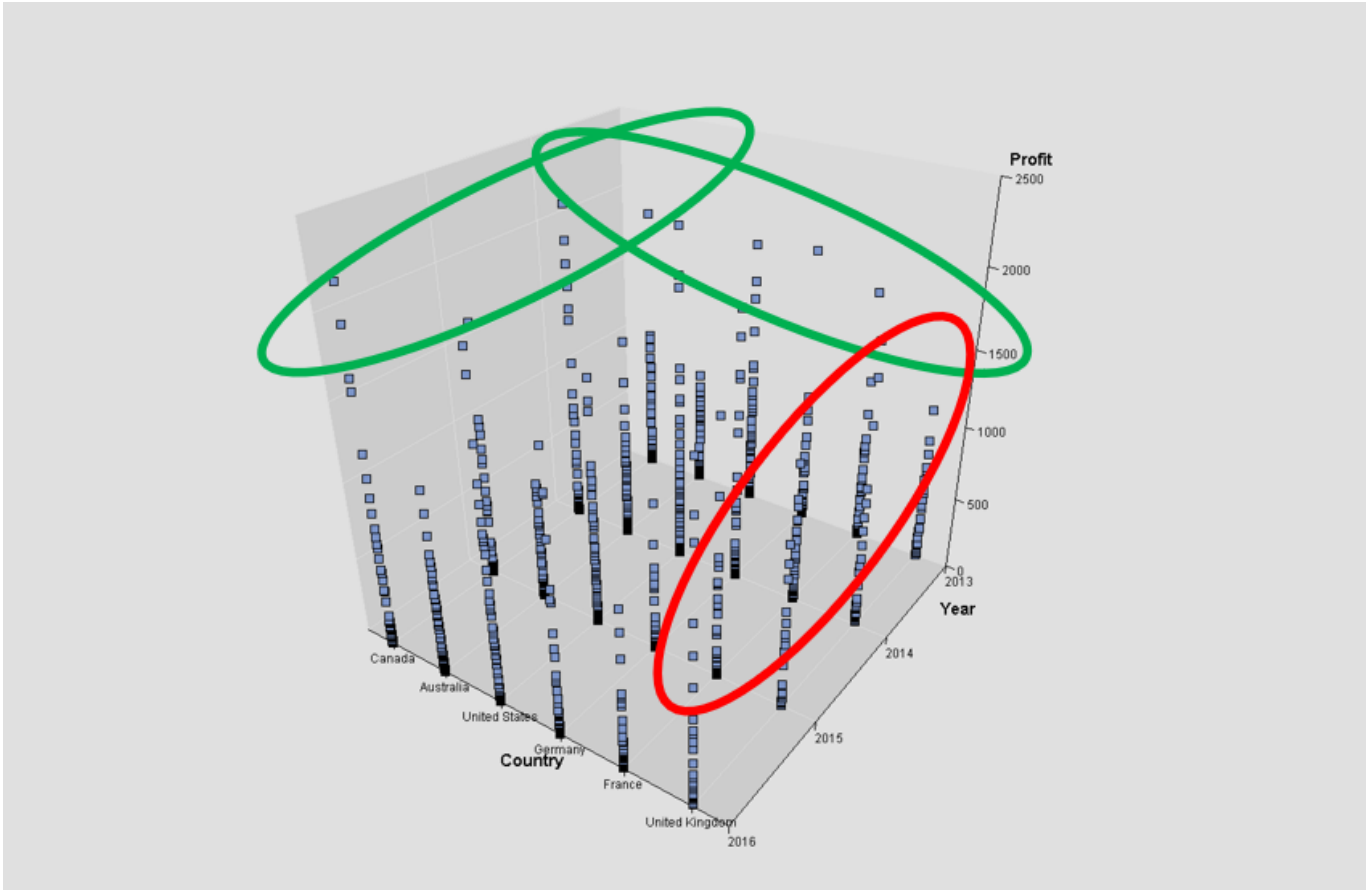


Figure 49 - Profit by Country and Year

Based on **Figure 49**, Canada generated the highest profit in the year 2014 and 2016. An interesting trend to look at is there was very high profit generated in the year 2014 throughout the countries except for the United Kingdom. The United Kingdom generated the least profit throughout the years.

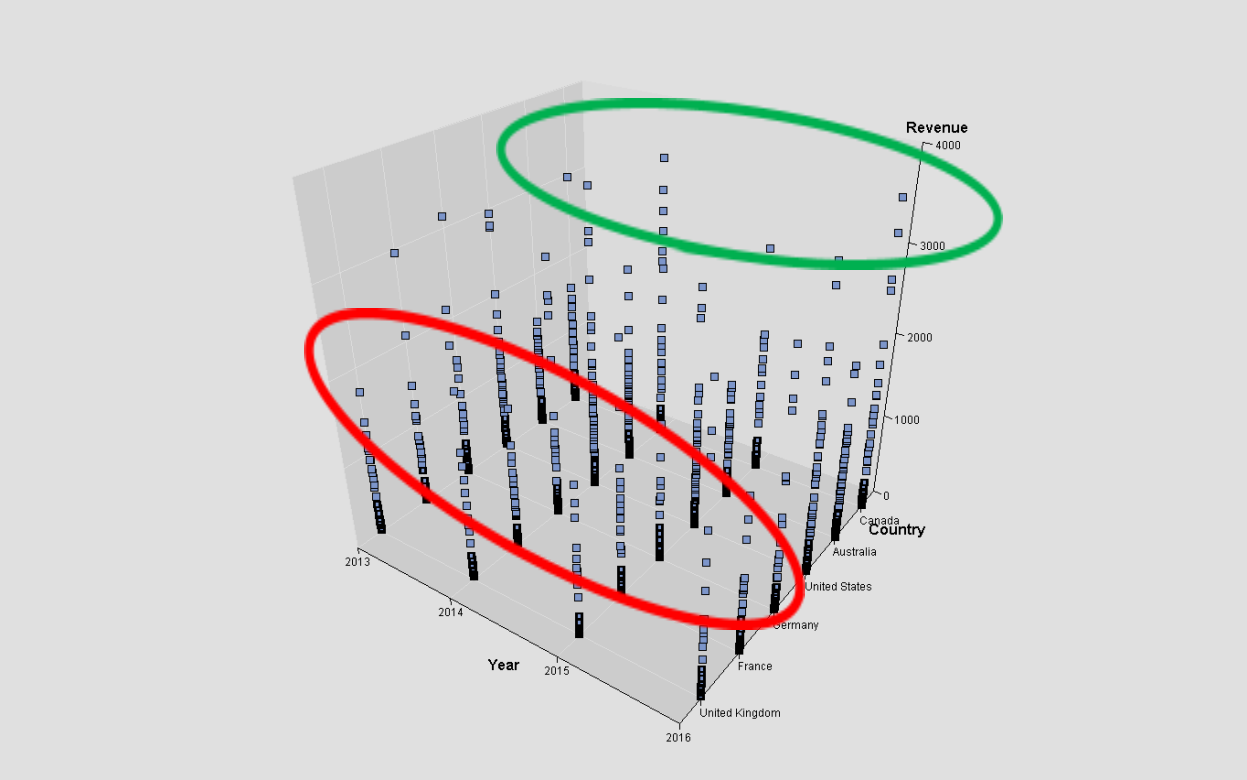


Figure 50 - Revenue by Country and Year

Based on **Figure 50**, Canada generated the highest revenue among other countries with the year 2014 and 2016 being the highest revenue while the United Kingdom generated the lowest revenue throughout the years with 2013 and 2016 being one of the lowest revenues generated.

5.3 Modify – Data Modification

5.3.1 Modify Dataset using Talend Data Preparation

The third stage in SEMMA methodology is the modification stage. The importance of this stage comes from the necessity of ensuring that the data we use in building the mining model is consistent, accurate, and reliable. Without modifying the dataset, the data might be inconsistent, incorrect, or completely wrong, which leads to poor conclusions and wrong decisions.

By cleaning the dataset, we achieve the following goals:
Save time and resources: As the chance of getting faulty results after data cleaning is minimized.
Improve data quality: Because data cleaning makes data more consistent, reliable, and complete.
Improve decision making: The quality of conclusions relies heavily on the quality of the used data, and data cleaning improves the quality of it.

Table 6 Error Types

Error type	Column
Incomplete data	Customer_Age
Incomplete data	Age_Group
Inconsistent data	Month
Inconsistent data	Country
Noise data	Product
Intentional error	Age_Column
Duplicates	Entire Table

Bike Sales dataset is imported in Talend Data Preparation.

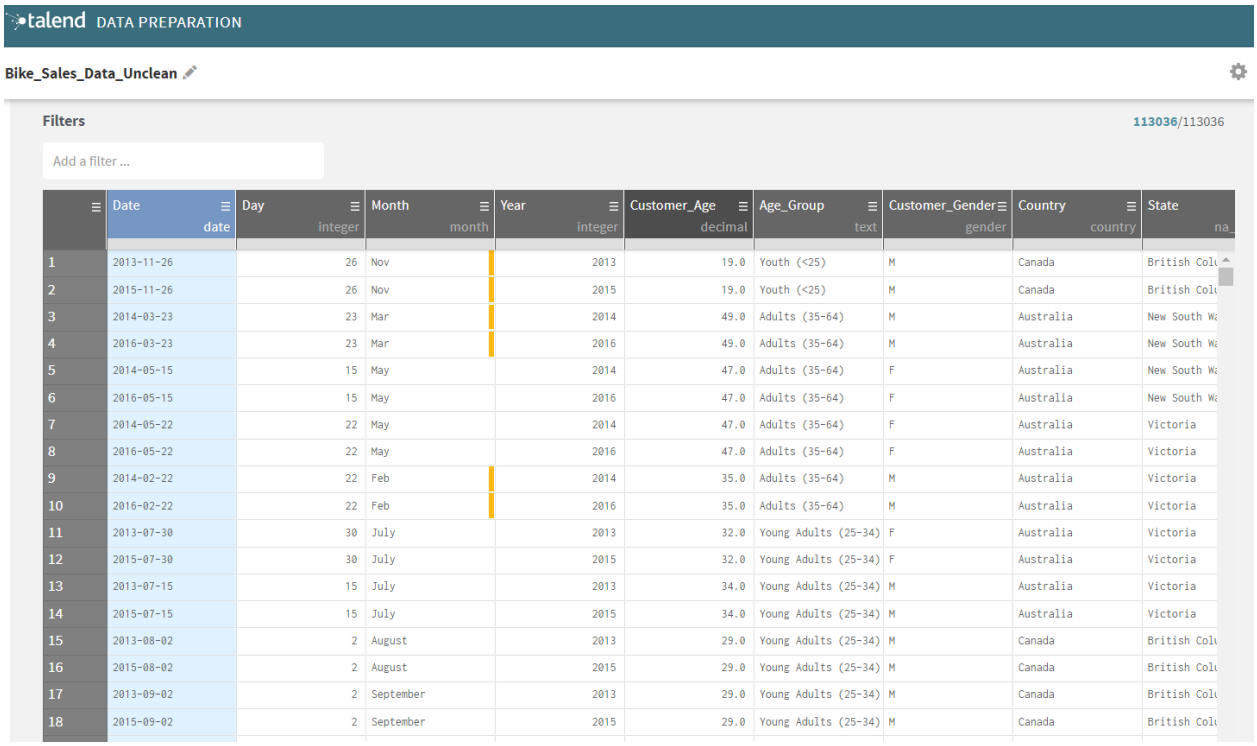


Figure 51 Talend Interface

5.3.2 Incomplete Data

The customer age column has 529 empty values, and the age group column has 529 empty values. The median value for age is 34.

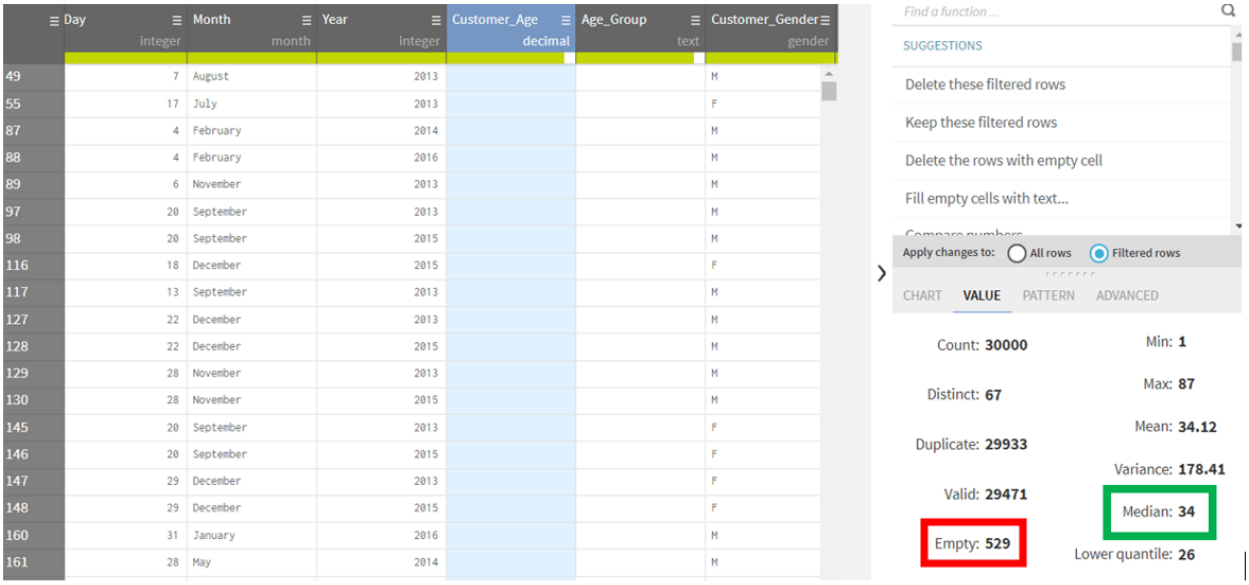


Figure 52 Incomplete data in Customer_Age column

Replace empty value of customer age with median value 34 and apply it to all cells with empty values.

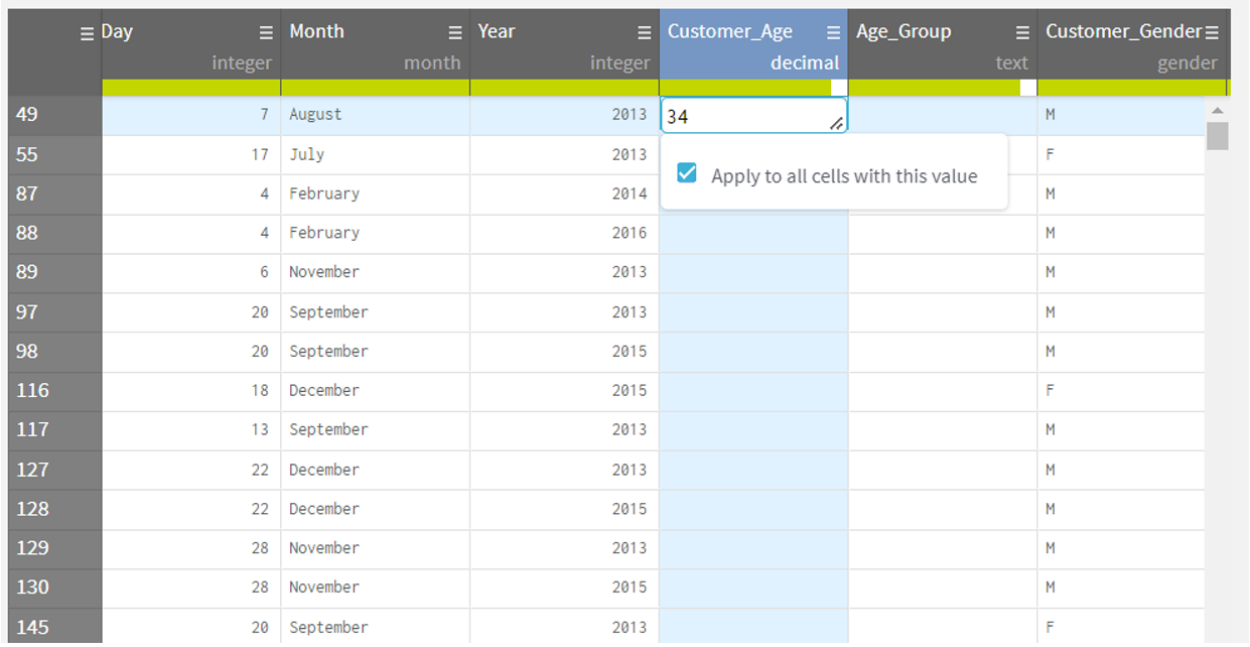


Figure 53 Replacing missing values with the median

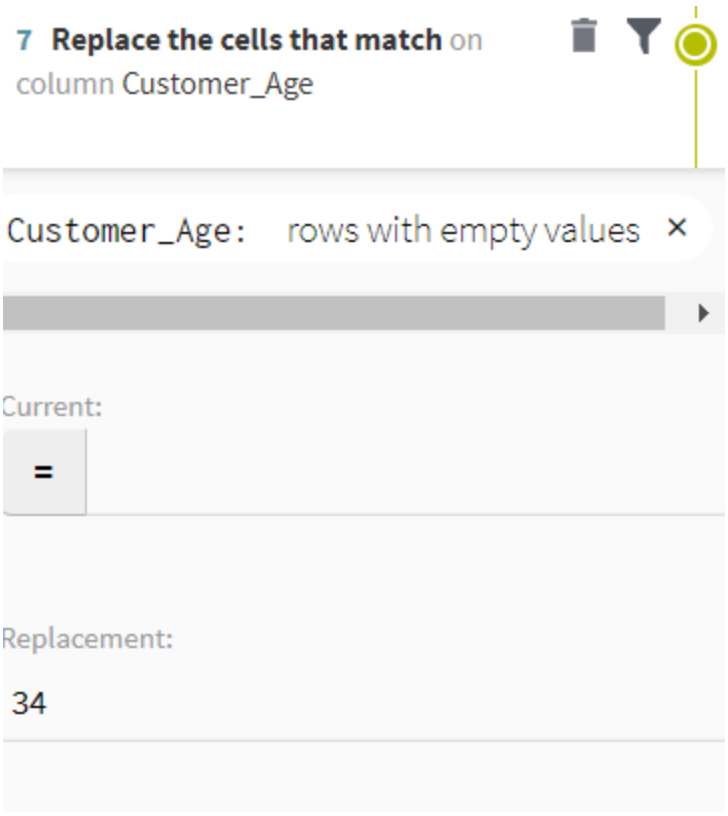


Figure 54 Customer_Age missing values replacement

Now, the customer age column has no empty values.

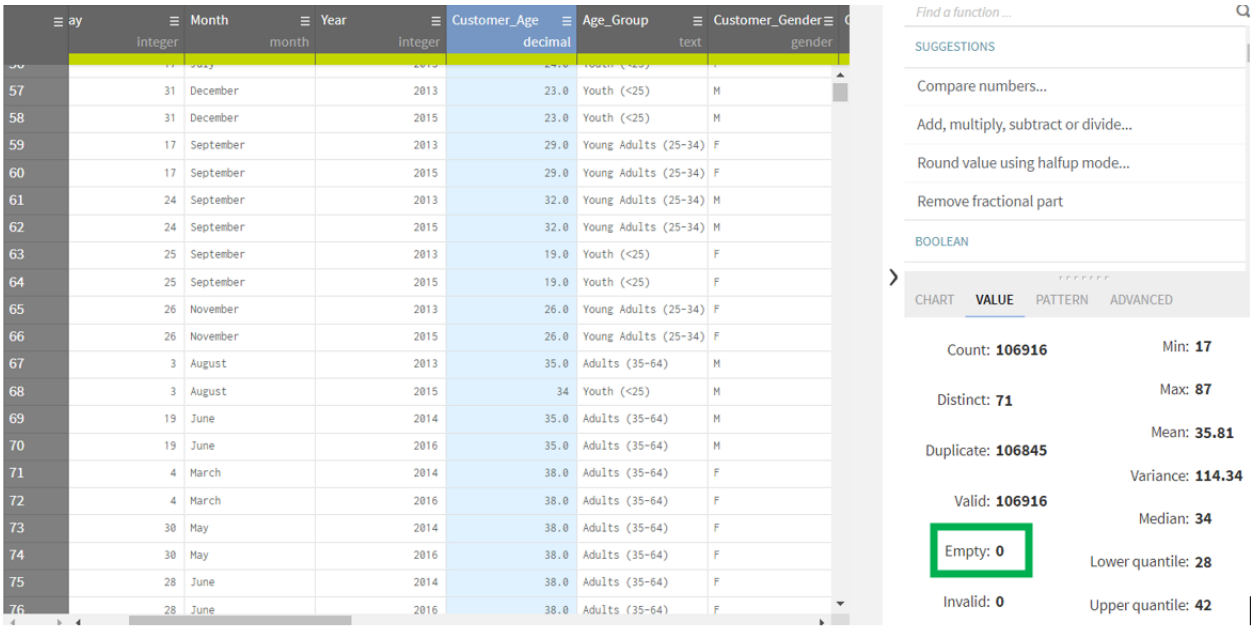


Figure 55 Customer_Age after replacement

Replacing empty value of age group column with ‘Young Adults (25-34)’ according to age value and applied it to all cells with empty values.

	Day integer	Month month	Year integer	Customer_Age decimal	Age_Group text	Customer_Gender gender
49	7	August	2013	34	Young Adults (25-34)	M
55	17	July	2013	34		F
87	4	February	2014	34		
88	4	February	2016	34		
89	6	November	2013	34		M
97	20	September	2013	34		M
98	20	September	2015	34		M
116	18	December	2015	34		F
117	13	September	2013	34		M
127	22	December	2013	34		M
128	22	December	2015	34		M
129	28	November	2013	34		M
130	28	November	2015	34		M
145	20	September	2013	34		F
146	20	September	2015	34		F

Apply to all cells with this value

Figure 56 Age_Group replacement

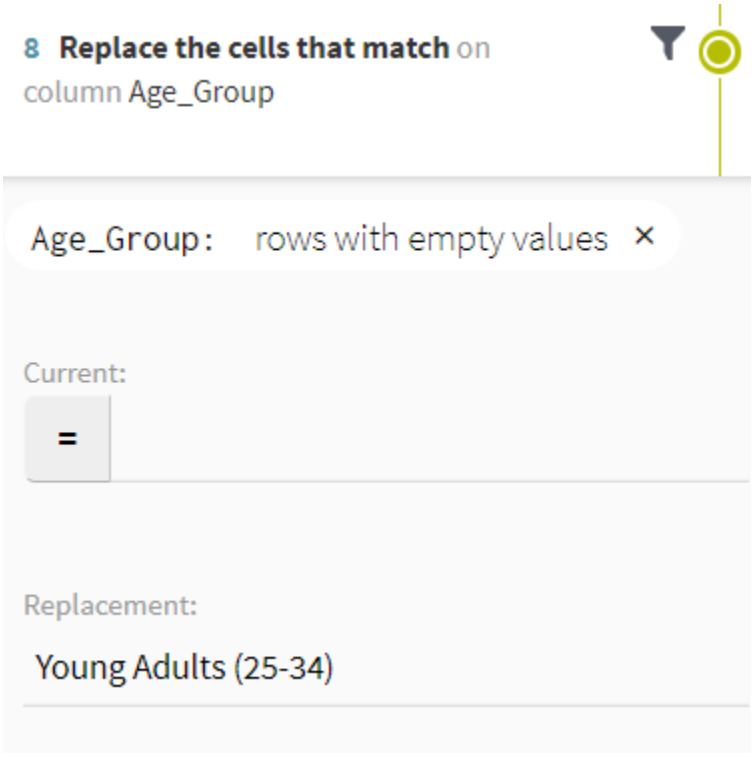
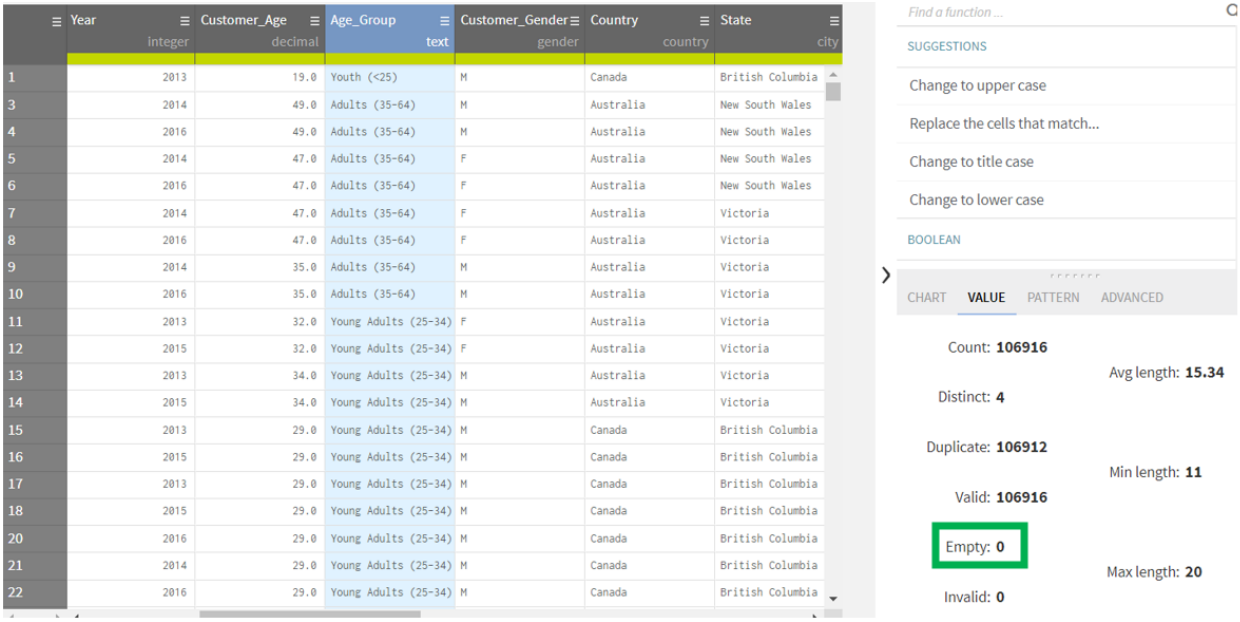


Figure 57 Age_Group replacement results

Now, the age group column has no empty values.

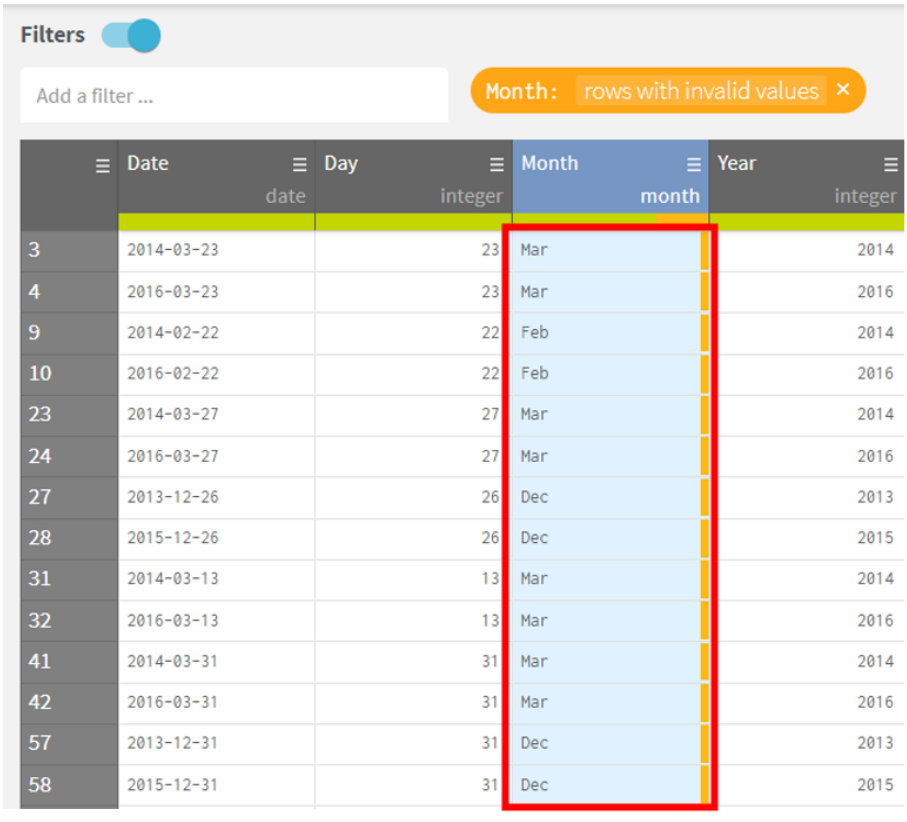


	Year integer	Customer_Age decimal	Age_Group text	Customer_Gender gender	Country country	State city
1	2013	19.0	Youth (<25)	M	Canada	British Columbia
3	2014	49.0	Adults (35-64)	M	Australia	New South Wales
4	2016	49.0	Adults (35-64)	M	Australia	New South Wales
5	2014	47.0	Adults (35-64)	F	Australia	New South Wales
6	2016	47.0	Adults (35-64)	F	Australia	New South Wales
7	2014	47.0	Adults (35-64)	F	Australia	Victoria
8	2016	47.0	Adults (35-64)	F	Australia	Victoria
9	2014	35.0	Adults (35-64)	M	Australia	Victoria
10	2016	35.0	Adults (35-64)	M	Australia	Victoria
11	2013	32.0	Young Adults (25-34)	F	Australia	Victoria
12	2015	32.0	Young Adults (25-34)	F	Australia	Victoria
13	2013	34.0	Young Adults (25-34)	M	Australia	Victoria
14	2015	34.0	Young Adults (25-34)	M	Australia	Victoria
15	2013	29.0	Young Adults (25-34)	M	Canada	British Columbia
16	2015	29.0	Young Adults (25-34)	M	Canada	British Columbia
17	2013	29.0	Young Adults (25-34)	M	Canada	British Columbia
18	2015	29.0	Young Adults (25-34)	M	Canada	British Columbia
20	2016	29.0	Young Adults (25-34)	M	Canada	British Columbia
21	2014	29.0	Young Adults (25-34)	M	Canada	British Columbia
22	2016	29.0	Young Adults (25-34)	M	Canada	British Columbia

Figure 58 Age_Group after replacement

5.3.3 Inconsistent Data

The Month column has inconsistent values for Feb, Mar, Nov and Dec as shown.



Filters

Add a filter ...

Month: rows with invalid values ×

	Date date	Day integer	Month month	Year integer
3	2014-03-23	23	Mar	2014
4	2016-03-23	23	Mar	2016
9	2014-02-22	22	Feb	2014
10	2016-02-22	22	Feb	2016
23	2014-03-27	27	Mar	2014
24	2016-03-27	27	Mar	2016
27	2013-12-26	26	Dec	2013
28	2015-12-26	26	Dec	2015
31	2014-03-13	13	Mar	2014
32	2016-03-13	13	Mar	2016
41	2014-03-31	31	Mar	2014
42	2016-03-31	31	Mar	2016
57	2013-12-31	31	Dec	2013
58	2015-12-31	31	Dec	2015

Replaced ‘Feb’ to ‘February’, ‘Mar’ to ‘March’, ‘Nov’ to ‘November’ and ‘Dec’ to ‘December’ to remove inconsistency and applied it to all rows with same value.

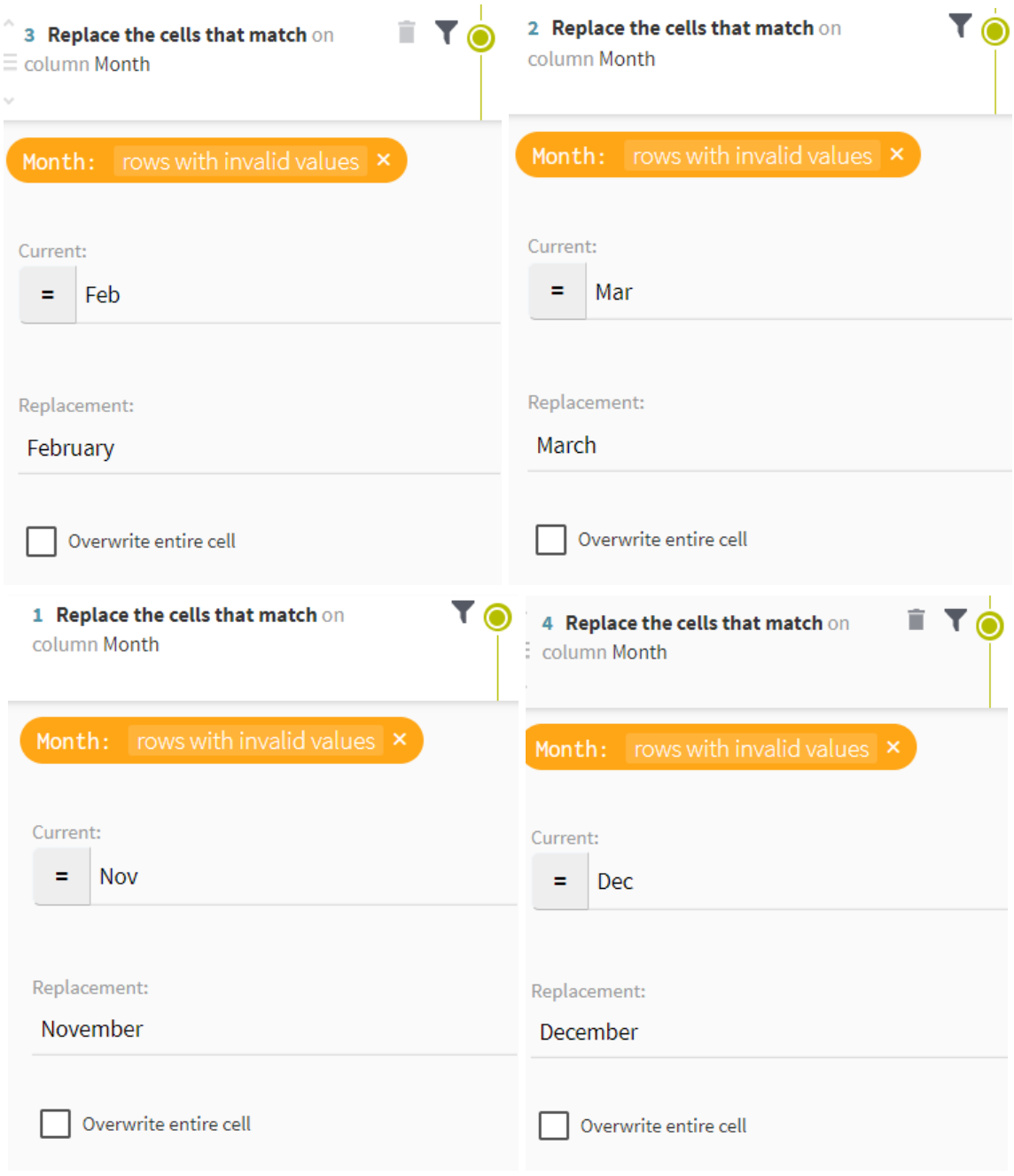
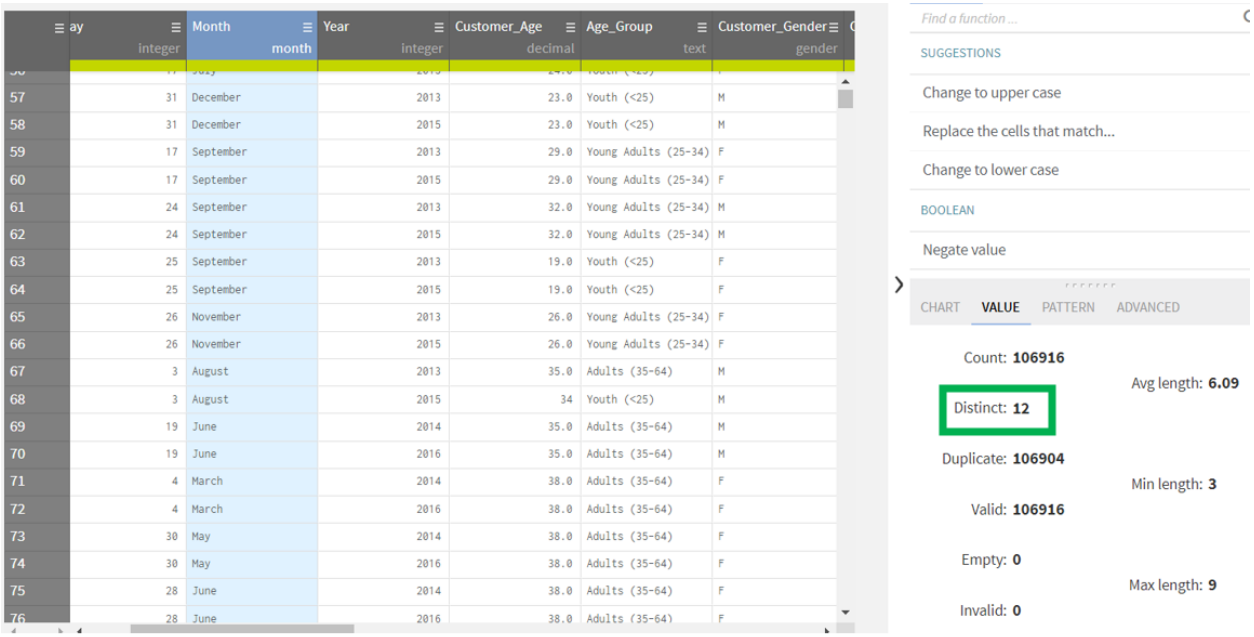


Figure 60 Replacement of month names

The Month column has no missing, invalid or inconsistent values.



	ay	Month	Year	Customer_Age	Age_Group	Customer_Gender
	integer	month	integer	decimal	text	gender
57	31	December	2013	23.0	Youth (<25)	M
58	31	December	2015	23.0	Youth (<25)	M
59	17	September	2013	29.0	Young Adults (25-34)	F
60	17	September	2015	29.0	Young Adults (25-34)	F
61	24	September	2013	32.0	Young Adults (25-34)	M
62	24	September	2015	32.0	Young Adults (25-34)	M
63	25	September	2013	19.0	Youth (<25)	F
64	25	September	2015	19.0	Youth (<25)	F
65	26	November	2013	26.0	Young Adults (25-34)	F
66	26	November	2015	26.0	Young Adults (25-34)	F
67	3	August	2013	35.0	Adults (35-64)	M
68	3	August	2015	34	Youth (<25)	M
69	19	June	2014	35.0	Adults (35-64)	M
70	19	June	2016	35.0	Adults (35-64)	M
71	4	March	2014	38.0	Adults (35-64)	F
72	4	March	2016	38.0	Adults (35-64)	F
73	30	May	2014	38.0	Adults (35-64)	F
74	30	May	2016	38.0	Adults (35-64)	F
75	28	June	2014	38.0	Adults (35-64)	F
76	28	June	2016	38.0	Adults (35-64)	F

Figure 61 Month column after replacement

The country column has inconsistent values for United States and United Kingdom.

	Customer_Gender gender	Country country	State city	Product_Category text	Sub_Category text
24	M	USA	Oregon	Accessories	Bike Racks
35	F	USA	Oregon	Accessories	Bike Racks
59	F	UK	England	Accessories	Bike Racks
82	F	USA	Washington	Accessories	Bike Racks
84	F	USA	California	Accessories	Bike Racks
88	M	USA	Washington	Accessories	Bike Racks
102	M	USA	California	Accessories	Bike Racks
108	M	USA	California	Accessories	Bike Racks
110	M	USA	Oregon	Accessories	Bike Racks
113	F	USA	Washington	Accessories	Bike Racks
114	F	USA	Washington	Accessories	Bike Racks
116	F	USA	Washington	Accessories	Bike Racks
118	M	UK	England	Accessories	Bike Racks
123	F	USA	California	Accessories	Bike Racks
126	F	USA	California	Accessories	Bike Racks
128	M	USA	Washington	Accessories	Bike Racks
129	M	USA	Oregon	Accessories	Bike Racks

Figure 62 Inconsistency in Country column

Replaced 'USA' to 'United States' to remove inconsistency and applied it to all rows with same value.
Replaced 'UK' to 'United Kingdom' to remove inconsistency and applied it to all rows with same value.

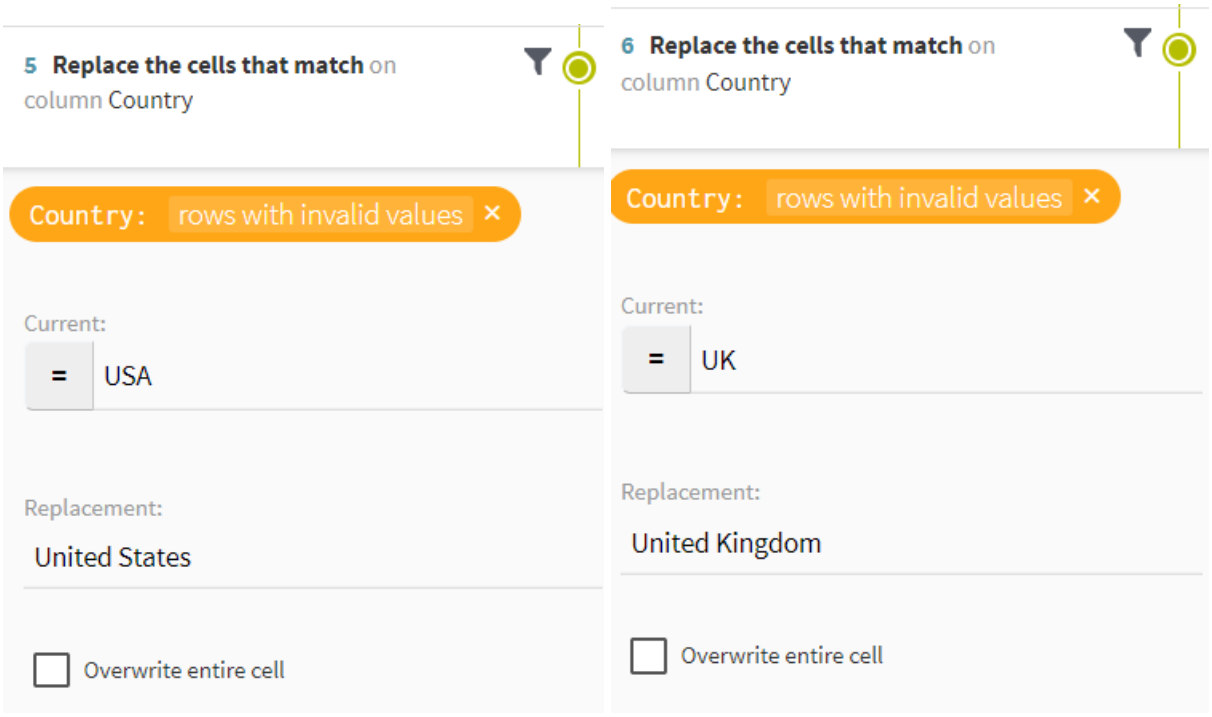


Figure 63 Replacing the values in Country column

The Country column has no missing, invalid or inconsistent values.

	nth	Year integer	Customer_Age decimal	Age_Group text	Customer_Gender gender	Country country	State
55		2013	34	Young Adults (25-34)	F	Australia	Queensland
56		2015	24.0	Youth (<25)	F	Australia	Queensland
57		2013	23.0	Youth (<25)	M	United States	California
58		2015	23.0	Youth (<25)	M	United States	California
59		2013	29.0	Young Adults (25-34)	F	United Kingdom	England
60		2015	29.0	Young Adults (25-34)	F	United Kingdom	England
61		2013	32.0	Young Adults (25-34)	M	France	Nord
62		2015	32.0	Young Adults (25-34)	M	France	Nord
63		2013	19.0	Youth (<25)	F	Germany	Saarland
64		2015	19.0	Youth (<25)	F	Germany	Saarland
65		2013	26.0	Young Adults (25-34)	F	Canada	British Colu
66		2015	26.0	Young Adults (25-34)	F	Canada	British Colu

Figure 64 Country column after replacement

5.3.4 Noisy Data

The Product column has '-----' values which is type of data noise.
6,120 Products have '-----' value out of 113,036 total values which is 0.05% of total data. So, we delete these filtered rows.

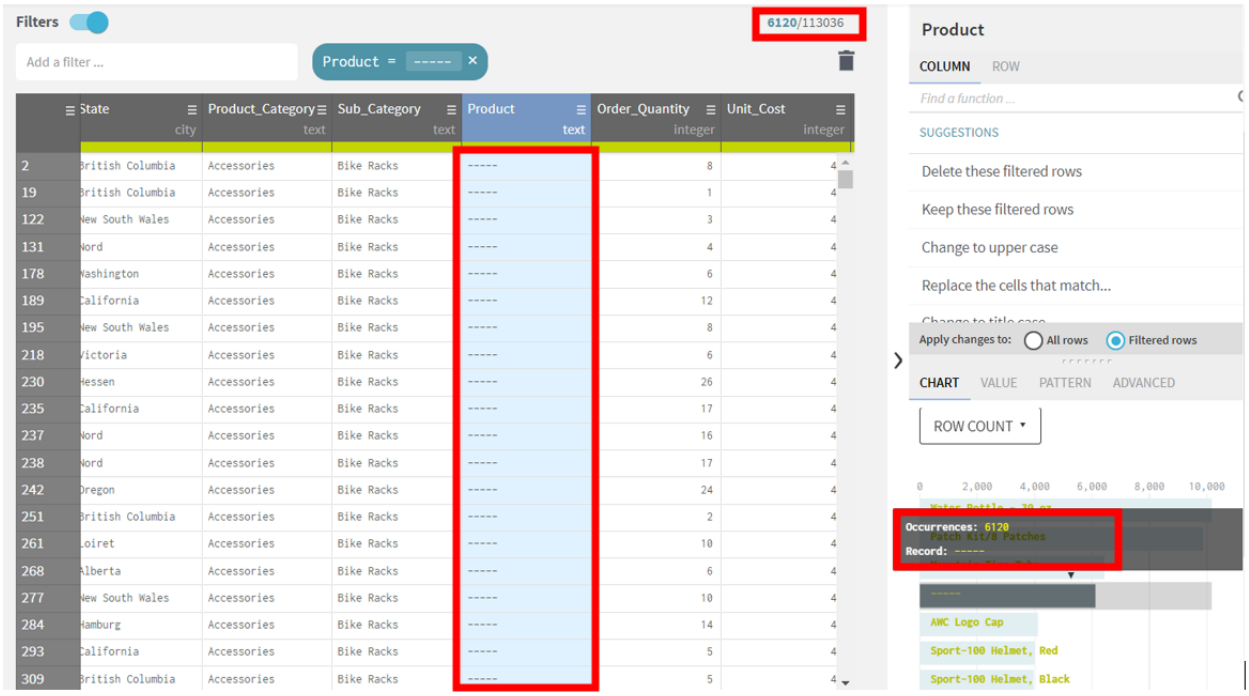


Figure 65 Noise data in Product column

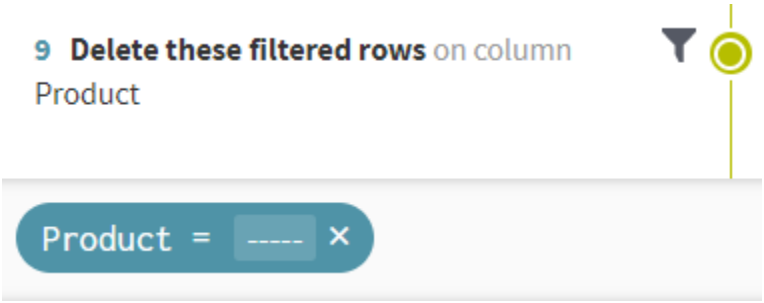


Figure 66 Removing noisy rows

Now, Product column has no noisy data.

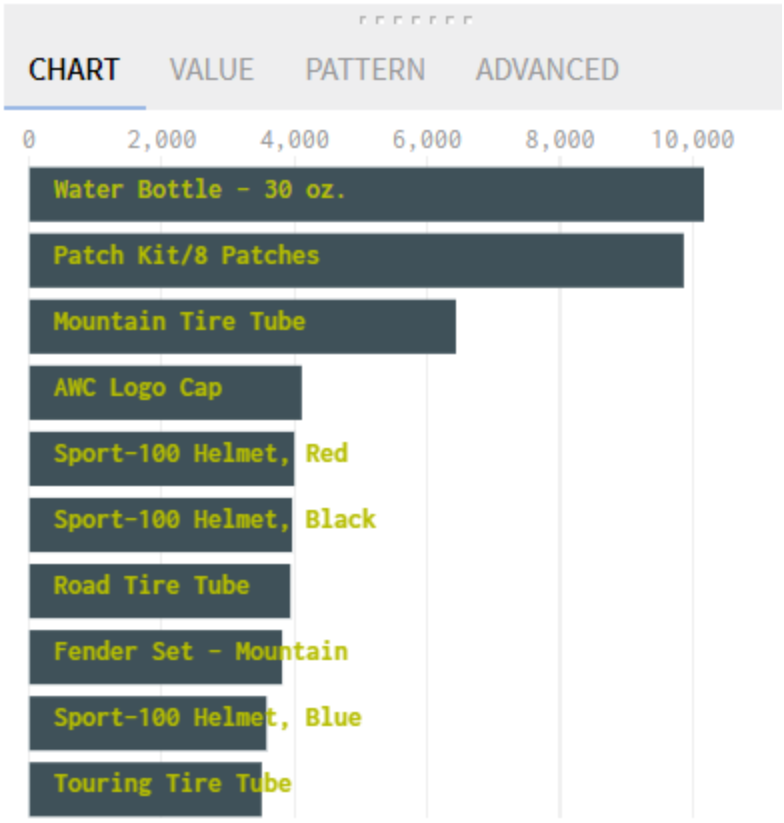


Figure 67 Product column after replacement

5.3.5 Intentional Error

The age column shows presence of '1.0' value which is an intentional error to hide the individual's age.

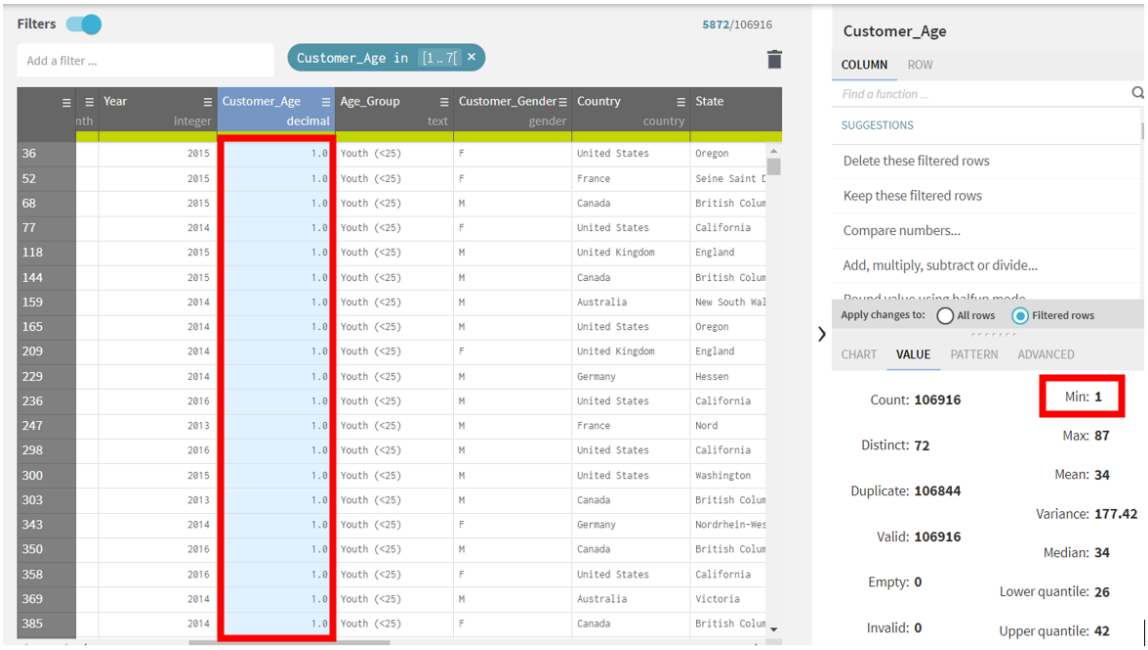


Figure 68 Intentional error in Age_Column

Replaced the customer age '1.0' to value 17.

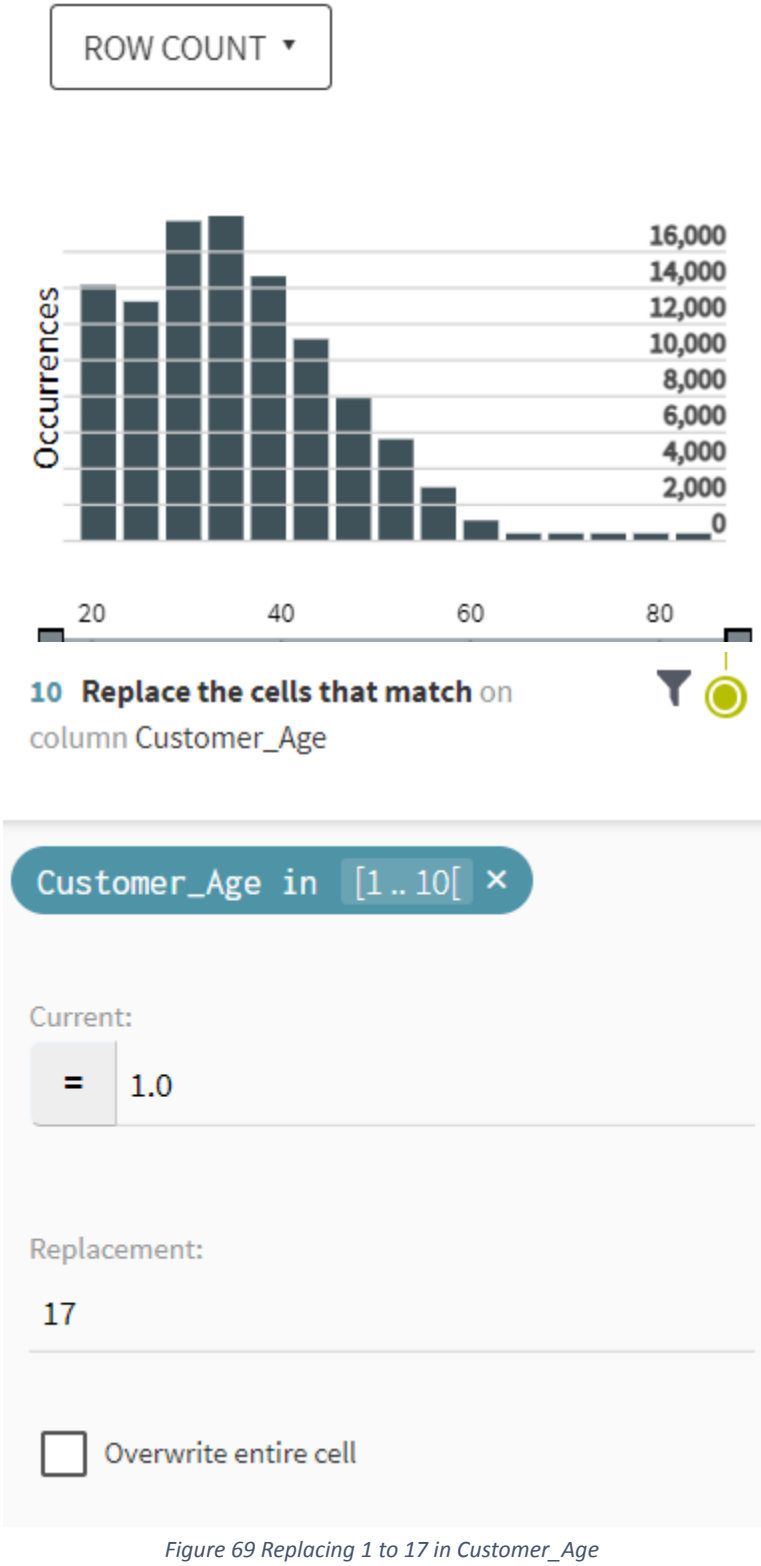


Figure 69 Replacing 1 to 17 in Customer_Age

So, it solved the intentional errors of the dataset as shown.

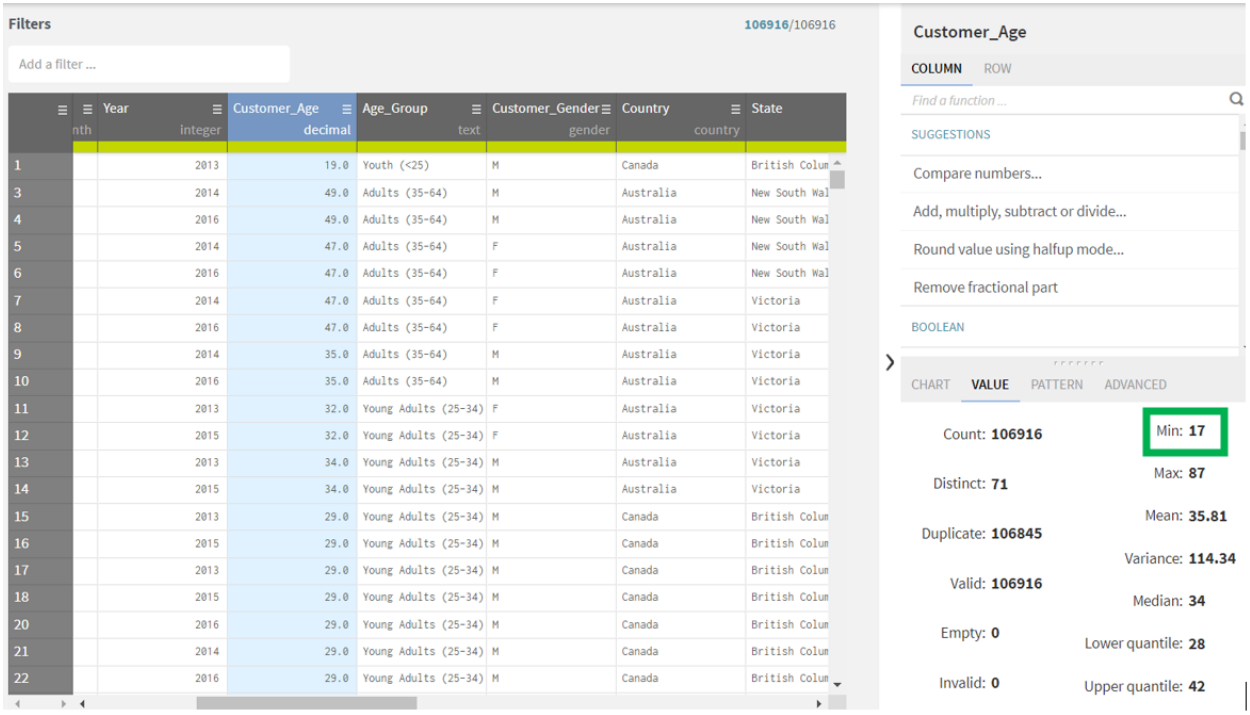


Figure 70 Customer_Age after cleaning

The following rules are applied for data cleaning and clean dataset is exported.

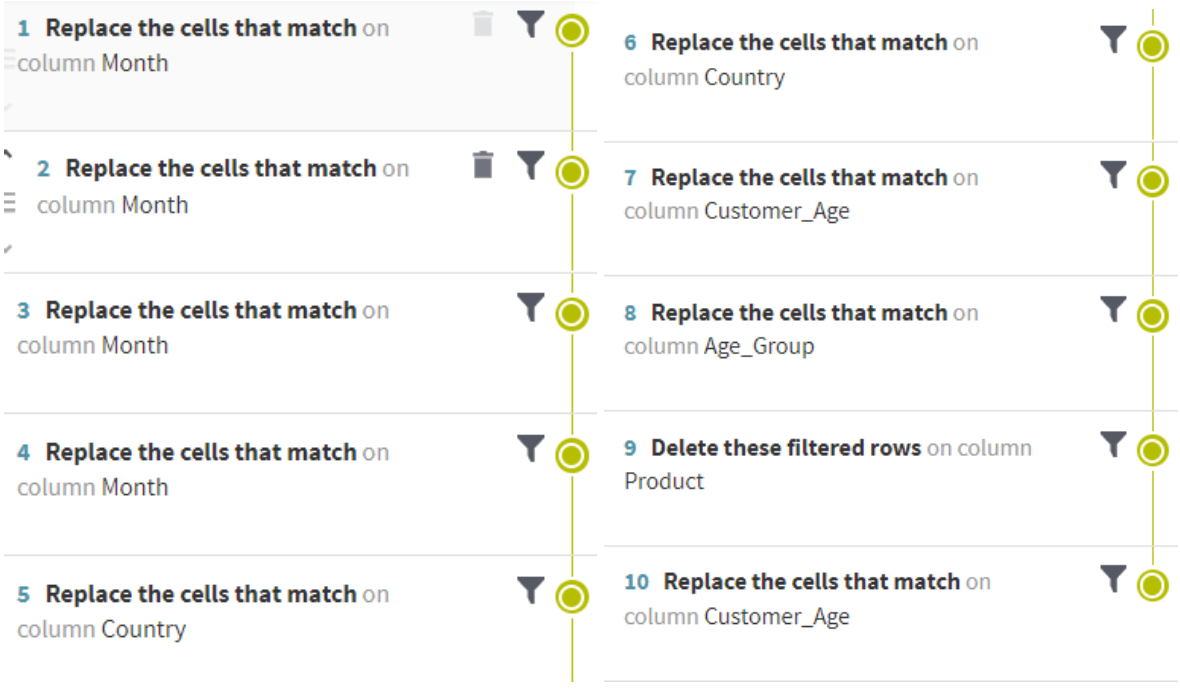


Figure 71 Data Cleaning Rules in Project

5.3.6 Duplicates

The following code is used in SAS Studio to find duplicate records, then delete these records.

CODE LOG RESULTS OUTPUT DATA

```
proc sql;
   title 'Duplicate Records';
   select count(*) from
   (select *, count(*) as Count
      from BIKESALE.BIKESALES
      group by Age_Group, Cost, Country, Customer_Age, Customer_Gender, Date, Day, Month, Order_Quantity,
      Product, Product_Category, Profit, Revenue, State, Sub_Category, Unit_Cost, Unit_Price, Year
      having count(*) > 1);


proc sql;
   title 'All Records';
   select count(*) from BIKESALE.BIKESALES;


proc sql;
    title 'Records after deleting duplicates';
    delete from
    (select *, count(*) as Count
      from BIKESALE.BIKESALES
      group by Age_Group, Cost, Country, Customer_Age, Customer_Gender, Date, Day, Month, Order_Quantity,
      Product, Product_Category, Profit, Revenue, State, Sub_Category, Unit_Cost, Unit_Price, Year
      having count(*) > 1);
```

Figure 72 detecting and replacing duplicates

The output shows the count of all records, duplicated records, and remaining records after deleting duplicated records.

Figure 73 Outputs of removing duplicates

5.3.7 Creating Training and Validation Data

The modified dataset is exported from Talend Data Preparation and imported into SAS Enterprise Miner. The imported dataset includes a target variable “Profit Label” which has High and Low values, and it can be used for classification. We used the data partition node to split the data randomly to 50% training and 50% validation.

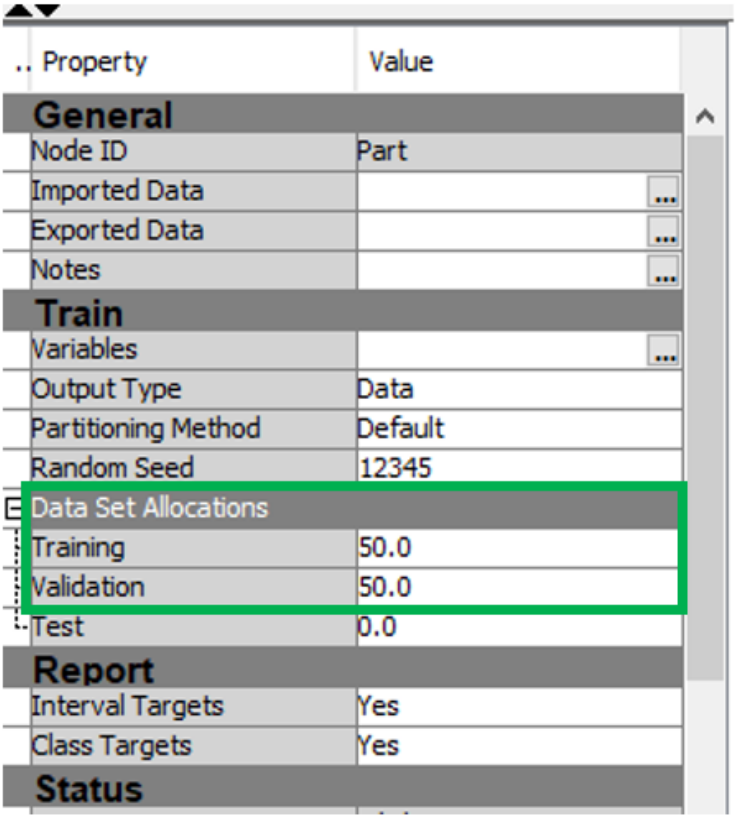


Figure 74 Data Partition

The output for data partition is shown below.

Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS2.FIMPORT_train	106131
TRAIN	EMWS2.Part_TRAIN	53065
VALIDATE	EMWS2.Part_VALIDATE	53066

--
* Score Output
--

--
* Report Output
--

Summary Statistics for Class Targets

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Profit_Label	.	High_Profit	53444	50.3566	
Profit_Label	.	Low_Profit	52687	49.6434	

Data=TRAIN

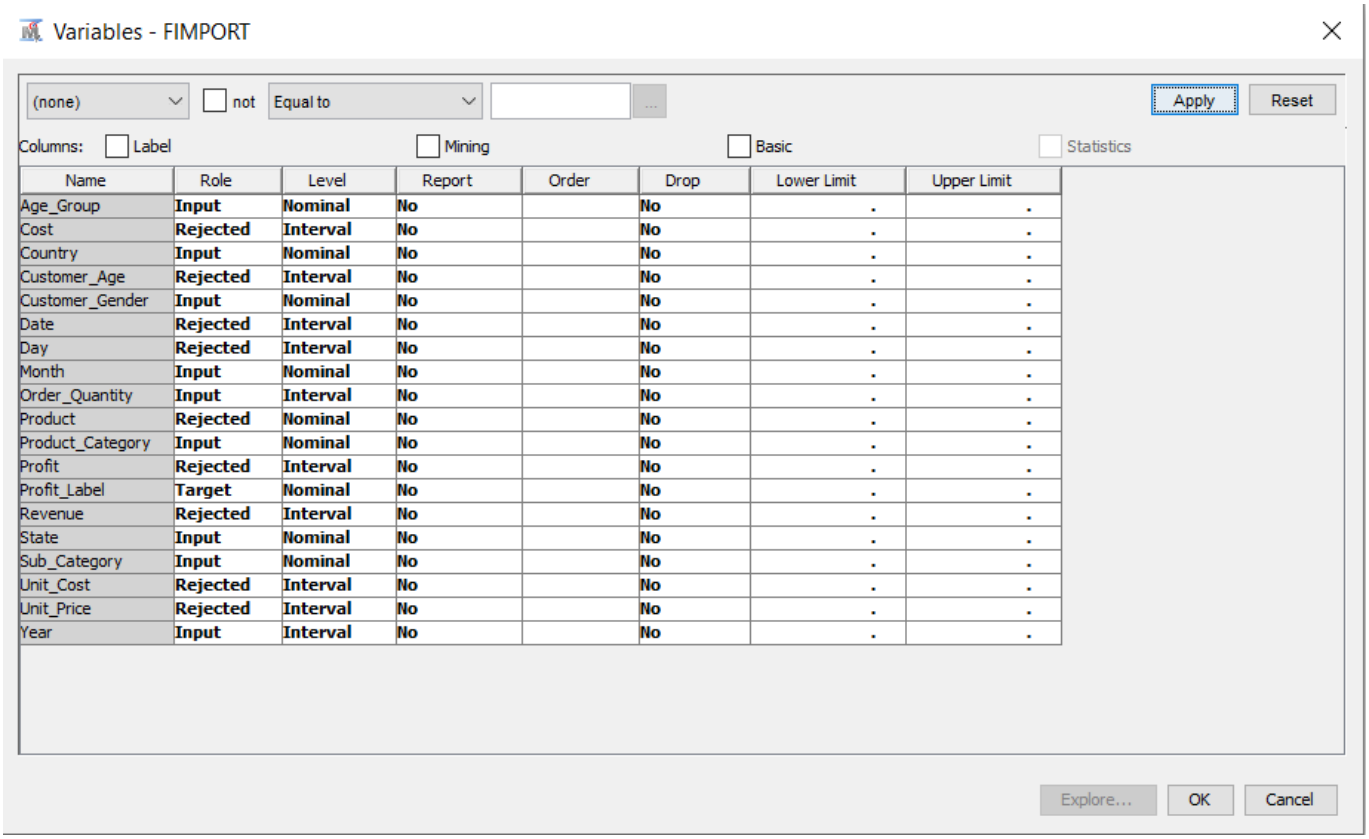
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Profit_Label	.	High_Profit	26722	50.3571	
Profit_Label	.	Low_Profit	26343	49.6429	

Data=VALIDATE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Profit_Label	.	High_Profit	26722	50.3562	
Profit_Label	.	Low_Profit	26344	49.6438	

Figure 75 Data Partition Summary

The selected columns to train the models are [Age_Group, Customer_Gender, Country, State, Year, Month, Order_Quantity, Product Category, Sub_Cateogry] and [Profit_Label] as the target feature with two possible outcomes [High – Low]. These attributes were selected based on their correlation with profit values. For example, customers' demographics such as age group and country, have direct association with total profit as found previously in Explore stage. The target feature is [Profit_Label] is derived from cost and revenue columns, so rejected them to avoid multicellularity.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age_Group	Input	Nominal	No		No	.	.
Cost	Rejected	Interval	No		No	.	.
Country	Input	Nominal	No		No	.	.
Customer_Age	Rejected	Interval	No		No	.	.
Customer_Gender	Input	Nominal	No		No	.	.
Date	Rejected	Interval	No		No	.	.
Day	Rejected	Interval	No		No	.	.
Month	Input	Nominal	No		No	.	.
Order_Quantity	Input	Interval	No		No	.	.
Product	Rejected	Nominal	No		No	.	.
Product_Category	Input	Nominal	No		No	.	.
Profit	Rejected	Interval	No		No	.	.
Profit_Label	Target	Nominal	No		No	.	.
Revenue	Rejected	Interval	No		No	.	.
State	Input	Nominal	No		No	.	.
Sub_Category	Input	Nominal	No		No	.	.
Unit_Cost	Rejected	Interval	No		No	.	.
Unit_Price	Rejected	Interval	No		No	.	.
Year	Input	Interval	No		No	.	.

Figure 76 Variable Roles

Variable Summary is shown below.

Variable Summary

Role	Measurement Level	Frequency Count
INPUT	INTERVAL	2
INPUT	NOMINAL	7
REJECTED	INTERVAL	8
REJECTED	NOMINAL	1
TARGET	NOMINAL	1

Figure 77 Data Partition node

5.4 Model – Data Modeling

Data is partitioned in training and validation. Now, we train different models including decision tree, gradient boosting and neural network.

5.4.1 Decision Tree

Decision Tree classifier is used to develop prediction model for the dataset. We developed the model with 2 branches maximum and a maximum depth of 6 so we can avoid overfitting.

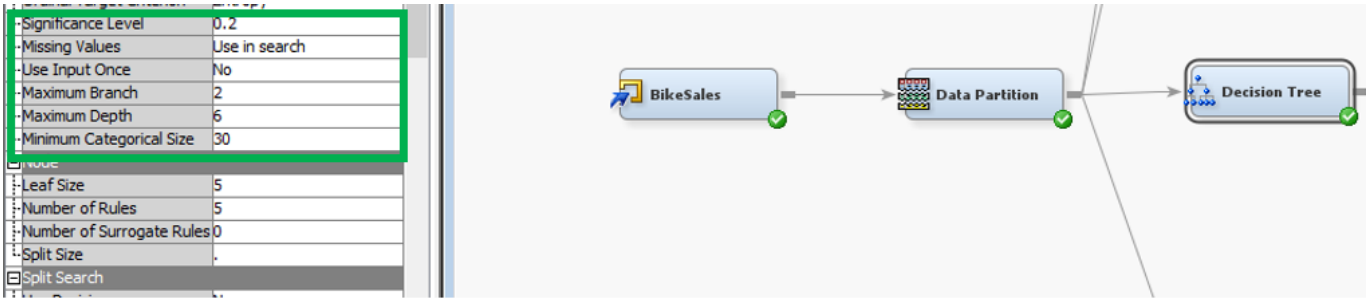


Figure 78 Training the Decision Tree Model.

We can see below the tree resulted from the decision tree model.

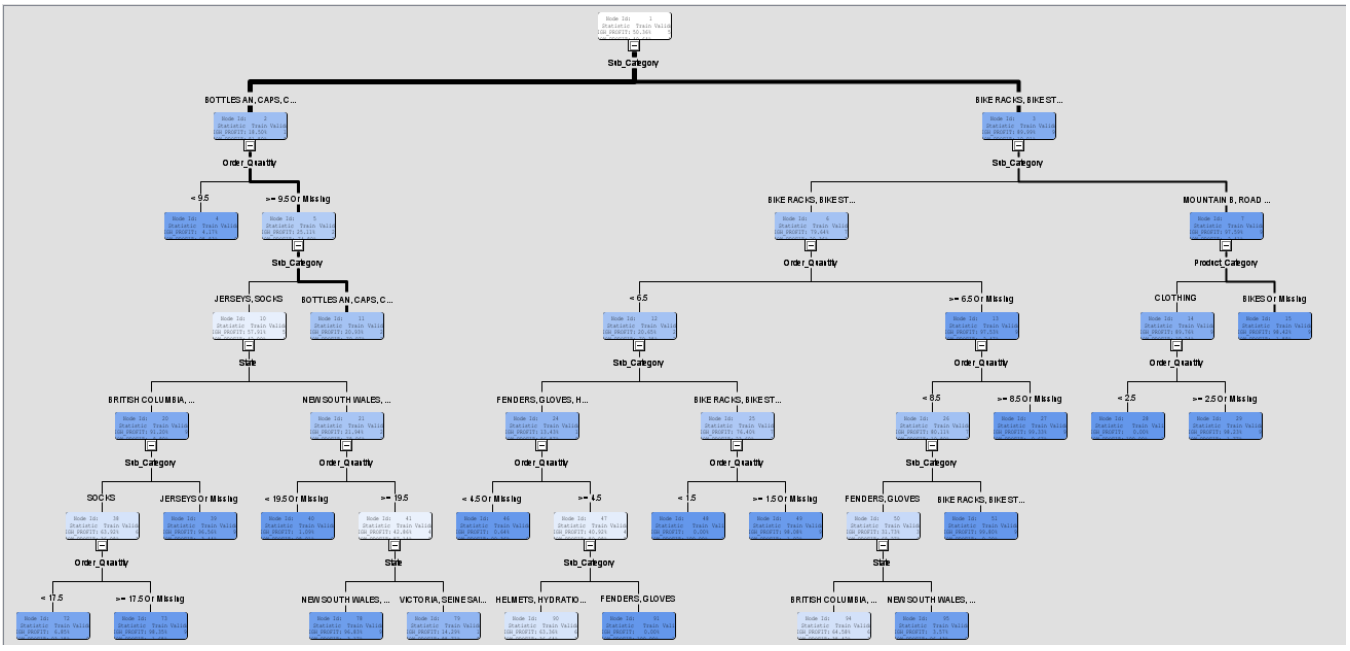


Figure 79 Project decision tree model

The hidden patterns observed from the decision model are:

- If a customer buys more than 18 socks in an order, then it will be profitable for the business.

- If bottles, caps and cleaners are sold less than 10 in an order, then it will not be a profitable sale.
- If more than 3 clothing items are sold in a sale transaction, then it will be a profitable sale for the business

Following are the rules based on which the decision tree is built:

Each node number refers to a node in the decision tree map.

Node = 4
if Sub_Category IS ONE OF: BOTTLES AN, CAPS, CLEANERS, JERSEYS, SOCKS, TIRES AND or MISSING
AND Order_Quantity < 9.5
then
Tree Node Identifier = 4
Number of Observations = 9281
Predicted: Profit_Label=Low_Profit = 0.96
Predicted: Profit_Label=High_Profit = 0.04

Node = 11
if Sub_Category IS ONE OF: BOTTLES AN, CAPS, CLEANERS, TIRES AND or MISSING
AND Order_Quantity >= 9.5 or MISSING
then
Tree Node Identifier = 11
Number of Observations = 17863
Predicted: Profit_Label=Low_Profit = 0.79
Predicted: Profit_Label=High_Profit = 0.21

Node = 15
if Sub_Category IS ONE OF: MOUNTAIN B, ROAD BIKES, SHORTS, TOURING BI, VESTS or MISSING
AND Product_Category IS ONE OF: BIKES or MISSING
then
Tree Node Identifier = 15
Number of Observations = 12335
Predicted: Profit_Label=Low_Profit = 0.02
Predicted: Profit_Label=High_Profit = 0.98

Node = 27
if Sub_Category IS ONE OF: BIKE RACKS, BIKE STAND, FENDERS, GLOVES, HELMETS, HYDRATION
AND Order_Quantity >= 8.5 or MISSING
then
Tree Node Identifier = 27
Number of Observations = 6963
Predicted: Profit_Label=Low_Profit = 0.01
Predicted: Profit_Label=High_Profit = 0.99

Node = 28
if Sub_Category IS ONE OF: MOUNTAIN B, ROAD BIKES, SHORTS, TOURING BI, VESTS or MISSING
AND Product_Category IS ONE OF: CLOTHING
AND Order_Quantity < 2.5
then
Tree Node Identifier = 28
Number of Observations = 112
Predicted: Profit_Label=Low_Profit = 1.00
Predicted: Profit_Label=High_Profit = 0.00

Node = 29
if Sub_Category IS ONE OF: MOUNTAIN B, ROAD BIKES, SHORTS, TOURING BI, VESTS or MISSING
AND Product_Category IS ONE OF: CLOTHING
AND Order_Quantity >= 2.5 or MISSING
then
Tree Node Identifier = 29
Number of Observations = 1187
Predicted: Profit_Label=Low_Profit = 0.02
Predicted: Profit_Label=High_Profit = 0.98

Node = 39
if Sub_Category IS ONE OF: JERSEYS or MISSING
AND State IS ONE OF: BRITISH COLUMBIA, CALIFORNIA, ENGLAND, HESSEN, HAMBURG, NORDRHEIN-WESTFALEN, BAYERN, HAUTS DE SEINE, ESSONNE, BRANDENBURG
AND Order_Quantity >= 9.5 or MISSING
then
Tree Node Identifier = 39
Number of Observations = 988
Predicted: Profit_Label=Low_Profit = 0.03
Predicted: Profit_Label=High_Profit = 0.97

Node = 40
if Sub_Category IS ONE OF: JERSEYS, SOCKS
AND State IS ONE OF: NEW SOUTH WALES, VICTORIA, OREGON, SEINE SAINT DENIS, MOSELLE, QUEENSLAND, NORD, WASHINGTON, SAARLAND, SEINE (PARIS), SOUTH AUSTRALIA, TASMANIA, LOIRET, YVELINE, SEINE ET MARNE or MISSING
AND Order_Quantity < 19.5 AND Order_Quantity >= 9.5 or MISSING
then
Tree Node Identifier = 40
Number of Observations = 548
Predicted: Profit_Label=Low_Profit = 0.99
Predicted: Profit_Label=High_Profit = 0.01

Node = 46
if Sub_Category IS ONE OF: FENDERS, GLOVES, HELMETS, HYDRATION or MISSING
AND Order_Quantity < 4.5 or MISSING
then
Tree Node Identifier = 46
Number of Observations = 1407
Predicted: Profit_Label=Low_Profit = 0.99
Predicted: Profit_Label=High_Profit = 0.01

Node = 48
if Sub_Category IS ONE OF: BIKE RACKS, BIKE STAND
AND Order_Quantity < 1.5
then
Tree Node Identifier = 48
Number of Observations = 59
Predicted: Profit_Label=Low_Profit = 1.00
Predicted: Profit_Label=High_Profit = 0.00

Node = 49
if Sub_Category IS ONE OF: BIKE RACKS, BIKE STAND
AND Order_Quantity < 6.5 AND Order_Quantity >= 1.5 or MISSING
then

Tree Node Identifier = 49
Number of Observations = 208
Predicted: Profit_Label=Low_Profit = 0.02
Predicted: Profit_Label=High_Profit = 0.98

Node = 51
if Sub_Category IS ONE OF: BIKE RACKS, BIKE STAND, HELMETS, HYDRATION or MISSING
AND Order_Quantity < 8.5 AND Order_Quantity >= 6.5
then
Tree Node Identifier = 51
Number of Observations = 511
Predicted: Profit_Label=Low_Profit = 0.00
Predicted: Profit_Label=High_Profit = 1.00

Node = 72
if Sub_Category IS ONE OF: SOCKS
AND State IS ONE OF: BRITISH COLUMBIA, CALIFORNIA, ENGLAND, HESSEN, HAMBURG, NORDRHEIN-WESTFALEN, BAYERN, HAUTS DE SEINE, ESSONNE, BRANDENBURG
AND Order_Quantity < 17.5 AND Order_Quantity >= 9.5
then
Tree Node Identifier = 72
Number of Observations = 73
Predicted: Profit_Label=Low_Profit = 0.93
Predicted: Profit_Label=High_Profit = 0.07

Node = 73
if Sub_Category IS ONE OF: SOCKS
AND State IS ONE OF: BRITISH COLUMBIA, CALIFORNIA, ENGLAND, HESSEN, HAMBURG, NORDRHEIN-WESTFALEN, BAYERN, HAUTS DE SEINE, ESSONNE, BRANDENBURG
AND Order_Quantity >= 17.5 or MISSING
then
Tree Node Identifier = 73
Number of Observations = 121
Predicted: Profit_Label=Low_Profit = 0.02
Predicted: Profit_Label=High_Profit = 0.98

Node = 78
if Sub_Category IS ONE OF: JERSEYS, SOCKS
AND State IS ONE OF: NEW SOUTH WALES, OREGON, NORD
AND Order_Quantity >= 19.5
then
Tree Node Identifier = 78
Number of Observations = 189
Predicted: Profit_Label=Low_Profit = 0.03
Predicted: Profit_Label=High_Profit = 0.97

Node = 79
if Sub_Category IS ONE OF: JERSEYS, SOCKS
AND State IS ONE OF: VICTORIA, SEINE SAINT DENIS, QUEENSLAND, WASHINGTON, SAARLAND, SEINE (PARIS), SOUTH AUSTRALIA, TASMANIA, YVELINE or MISSING
AND Order_Quantity >= 19.5
then
Tree Node Identifier = 79
Number of Observations = 357

Predicted: Profit_Label=Low_Profit = 0.86
Predicted: Profit_Label=High_Profit = 0.14

Node = 90
if Sub_Category IS ONE OF: HELMETS, HYDRATION or MISSING
AND Order_Quantity < 6.5 AND Order_Quantity >= 4.5
then
Tree Node Identifier = 90
Number of Observations = 423
Predicted: Profit_Label=Low_Profit = 0.37
Predicted: Profit_Label=High_Profit = 0.63

Node = 91
if Sub_Category IS ONE OF: FENDERS, GLOVES
AND Order_Quantity < 6.5 AND Order_Quantity >= 4.5
then
Tree Node Identifier = 91
Number of Observations = 232
Predicted: Profit_Label=Low_Profit = 1.00
Predicted: Profit_Label=High_Profit = 0.00

Node = 94
if Sub_Category IS ONE OF: FENDERS, GLOVES
AND State IS ONE OF: BRITISH COLUMBIA, CALIFORNIA, ENGLAND
AND Order_Quantity < 8.5 AND Order_Quantity >= 6.5
then
Tree Node Identifier = 94
Number of Observations = 96
Predicted: Profit_Label=Low_Profit = 0.35
Predicted: Profit_Label=High_Profit = 0.65

Node = 95
if Sub_Category IS ONE OF: FENDERS, GLOVES
AND State IS ONE OF: NEW SOUTH WALES, VICTORIA, OREGON, QUEENSLAND, WASHINGTON, NORDRHEIN-WESTFALEN or MISSING
AND Order_Quantity < 8.5 AND Order_Quantity >= 6.5
then
Tree Node Identifier = 95
Number of Observations = 112
Predicted: Profit_Label=Low_Profit = 0.96
Predicted: Profit_Label=High_Profit = 0.04

The general statistics about the performance of the decision tree model are shown below.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Profit_Label		NOBS	Sum of Frequencies	53065	53066	.
Profit_Label		MISC	Misclassification ...	0.088552	0.090717	.
Profit_Label		MAX	Maximum Absolut...	0.998043	0.998043	.
Profit_Label		SSE	Sum of Squared E...	7637.35	7786.237	.
Profit_Label		ASE	Average Squared ...	0.071962	0.073364	.
Profit_Label		RASE	Root Average Squ...	0.268258	0.270857	.
Profit_Label		DIV	Divisor for ASE	106130	106132	.
Profit_Label		DFT	Total Degrees of ...	53065	.	.

Figure 80 Decision Tree Model Statistics

As we can see above, the model scored a misclassification rate of 0.091 approximately, which translates to an accuracy of 90.9%, which means that the model is quite accurate. Moreover, we can see below the features that most impacted the performance of the model which are Subcategory, Order Quantity, State and Product Category.

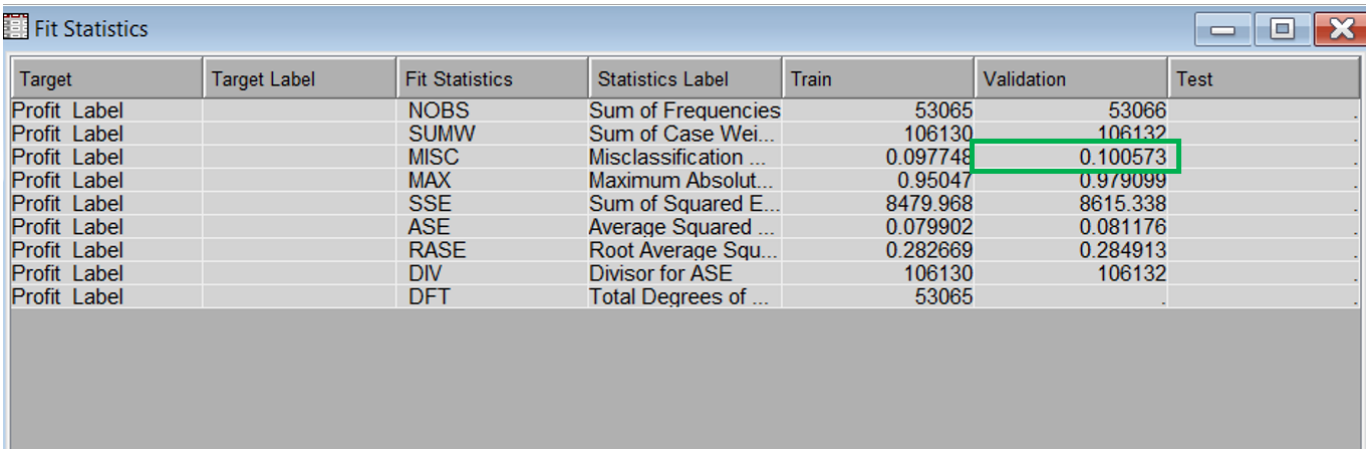
Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Sub_Category		7	1.0000	1.0000	1.0000
Order_Quantity		8	0.4736	0.4719	0.9964
State		3	0.2254	0.2241	0.9945
Product_Category		1	0.0345	0.0272	0.7888

Figure 81 Feature Importance for the Decision Tree Model

5.4.2 Gradient Boosting

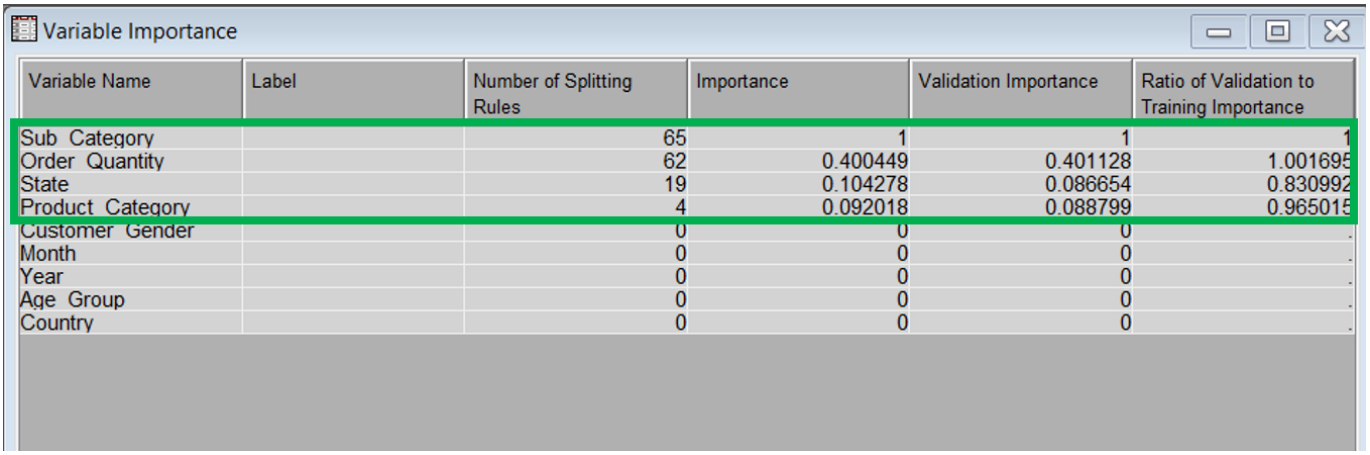
Gradient Boosting is an ensemble technique and can be used for classification problems. It creates multiple weak models and combines the output to get better performance. Gradient Boosting node is created in the diagram and connected to the data partition node. The statistics of the Gradient Boosting model show that the misclassification rate is 0.1001 which translate into an accuracy of 89.99% as shown below.



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Profit Label		NOBS	Sum of Frequencies	53065	53066	.
Profit Label		SUMW	Sum of Case Wei...	106130	106132	.
Profit Label		MISC	Misclassification ...	0.097748	0.100573	.
Profit Label		MAX	Maximum Absolut...	0.95047	0.979099	.
Profit Label		SSE	Sum of Squared E...	8479.968	8615.338	.
Profit Label		ASE	Average Squared ...	0.079902	0.081176	.
Profit Label		RASE	Root Average Squ...	0.282669	0.284913	.
Profit Label		DIV	Divisor for ASE	106130	106132	.
Profit Label		DFT	Total Degrees of ...	53065	.	.

Figure 82 - Gradient Boosting Model Statistics

The variable importance of the Gradient Boosting model is similar to Decision Tree model which include Subcategory, Order Quantity, State and Product Category as shown below.

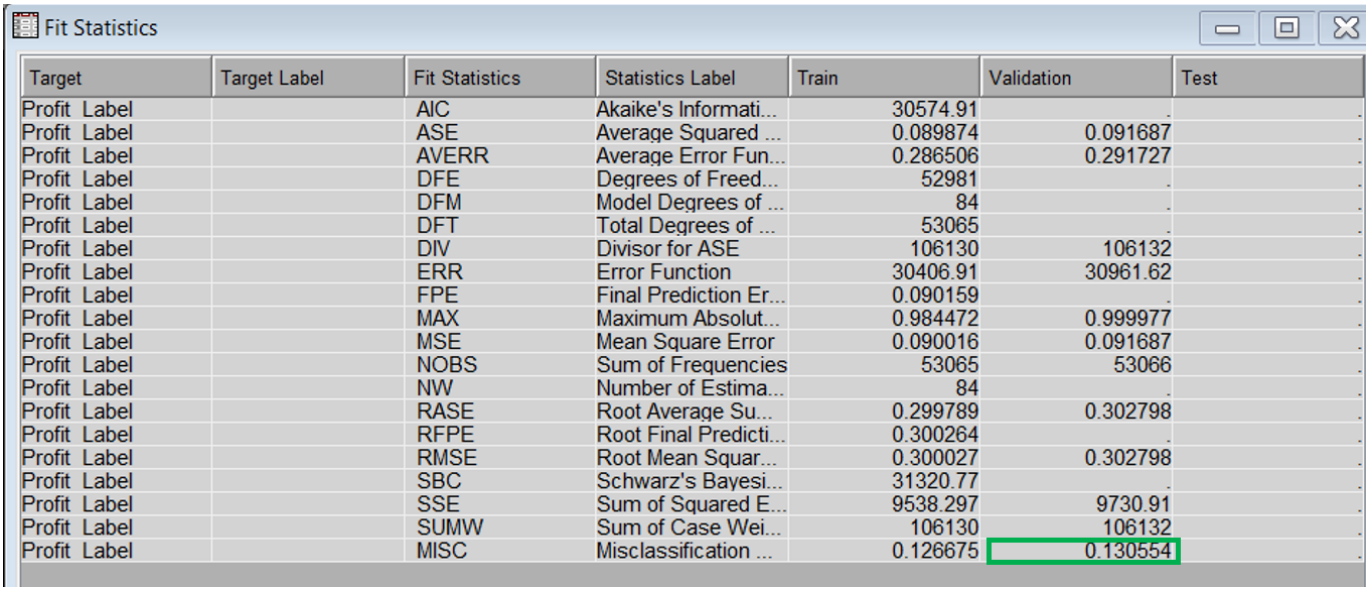


Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Sub Category		65	1	1	1
Order Quantity		62	0.400449	0.401128	1.001695
State		19	0.104278	0.086654	0.830992
Product Category		4	0.092018	0.088799	0.965015
Customer Gender		0	0	0	.
Month		0	0	0	.
Year		0	0	0	.
Age Group		0	0	0	.
Country		0	0	0	.

Figure 83 - Variable Importance of the Gradient Boosting

5.4.3 Logistic Regression

Logistic Regression is commonly used for classification and predictive analysis. It estimates the probability of an event occurring based on the independent variables. It uses logistic function to perform regression analysis to find out the binary dependent variable. Regression node is added in the diagram and connected to the data partition node. The target variable is a binary variable so logistic regression will be applied and the statistics of logistic regression model show that the misclassification rate is 0.131 which translates into an accuracy of 86.9% as shown below.



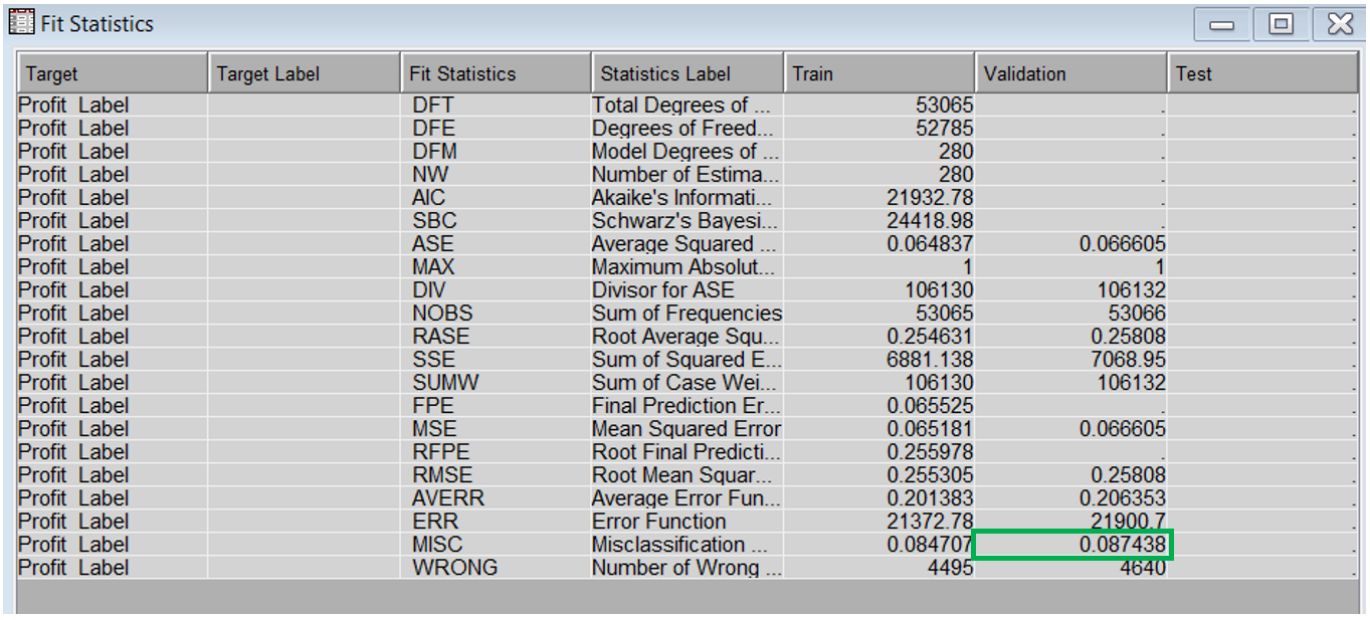
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Profit	Label	AIC	Akaike's Informati...	30574.91	.	.
Profit	Label	ASE	Average Squared ...	0.089874	0.091687	.
Profit	Label	AVERR	Average Error Fun...	0.286506	0.291727	.
Profit	Label	DFE	Degrees of Freed...	52981	.	.
Profit	Label	DFM	Model Degrees of ...	84	.	.
Profit	Label	DFT	Total Degrees of ...	53065	.	.
Profit	Label	DIV	Divisor for ASE	106130	106132	.
Profit	Label	ERR	Error Function	30406.91	30961.62	.
Profit	Label	FPE	Final Prediction Er...	0.090159	.	.
Profit	Label	MAX	Maximum Absolut...	0.984472	0.999977	.
Profit	Label	MSE	Mean Square Error	0.090016	0.091687	.
Profit	Label	NOBS	Sum of Frequencies	53065	53066	.
Profit	Label	NW	Number of Estima...	84	.	.
Profit	Label	RASE	Root Average Su...	0.299789	0.302798	.
Profit	Label	RFPE	Root Final Predicti...	0.300264	.	.
Profit	Label	RMSE	Root Mean Squar...	0.300027	0.302798	.
Profit	Label	SBC	Schwarz's Bayesi...	31320.77	.	.
Profit	Label	SSE	Sum of Squared E...	9538.297	9730.91	.
Profit	Label	SUMW	Sum of Case Wei...	106130	106132	.
Profit	Label	MISC	Misclassification ...	0.126675	0.130554	.

Figure 84 - Logistic Regression Model Statistics

5.4.4 Neural Network

Neural Network algorithms simulate the working of human brains, and it is similar to connection of neurons. It can be used for classification tasks. Neutral network node is added in the diagram and connected to the data partition node. The statistics for the neural network model show that the misclassification rate is 0.088 which translates to an accuracy of 91.2% as shown below.



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Profit	Label	DFT	Total Degrees of ...	53065	.	.
Profit	Label	DFE	Degrees of Freed...	52785	.	.
Profit	Label	DFM	Model Degrees of ...	280	.	.
Profit	Label	NW	Number of Estima...	280	.	.
Profit	Label	AIC	Akaike's Informati...	21932.78	.	.
Profit	Label	SBC	Schwarz's Bayesi...	24418.98	.	.
Profit	Label	ASE	Average Squared ...	0.064837	0.066605	.
Profit	Label	MAX	Maximum Absolut...	1	1	.
Profit	Label	DIV	Divisor for ASE	106130	106132	.
Profit	Label	NOBS	Sum of Frequencies	53065	53066	.
Profit	Label	RASE	Root Average Squ...	0.254631	0.25808	.
Profit	Label	SSE	Sum of Squared E...	6881.138	7068.95	.
Profit	Label	SUMW	Sum of Case Wei...	106130	106132	.
Profit	Label	FPE	Final Prediction Er...	0.065525	.	.
Profit	Label	MSE	Mean Squared Error	0.065181	0.066605	.
Profit	Label	RFPE	Root Final Predicti...	0.255978	.	.
Profit	Label	RMSE	Root Mean Squar...	0.255305	0.25808	.
Profit	Label	AVERR	Average Error Fun...	0.201383	0.206353	.
Profit	Label	ERR	Error Function	21372.78	21900.7	.
Profit	Label	MISC	Misclassification ...	0.084707	0.087438	.
Profit	Label	WRONG	Number of Wrong ...	4495	4640	.

Figure 85 - Neural Network Model Statistics

5.5 Assess – Models Assessment

The developed models are assessed in the Assess phase of SEMMA methodology. So, the Decision Tree, Gradient Boosting, Logistic Regression and Neural Network model nodes are connected to a control point node and control point node is connected to the Model Comparison node. The model comparison node is used to compare the performance of each model. The result of the model comparison is shown below.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate ▲
Y	Neural	Neural	Neural Network	Profit Label		0.087438
	Tree	Tree	Decision Tree	Profit Label		0.090717
	Boost	Boost	Gradient Boosting	Profit Label		0.100573
	Reg	Reg	Regression	Profit Label		0.130554

Figure 86 Models Comparison

It is concluded from the model comparison that the neural network had the best performance with a misclassification rate of 0.088 , i.e., 91.2 % accuracy. Then, the decision tree model with a misclassification rate of 0.091, i.e., 90.9% accuracy, the gradient boosting model with a misclassification rate of 0.1001, i.e., 89.9% accuracy and the logistic regression model with misclassification rate is 0.131, i.e., 86.9% accuracy.

6 Conclusion

In this project, we applied SEMMA methodology to mine impactful insights from Bike Sales Data set. We started mining process by Sampling the data into SAS Enterprise Miner software and storing it as a data source. Then, we Explored the dataset features via the visualization tools available in SAS software, which included univariate, bivariate, and correlation analysis. It was found during the data exploration that the data had incomplete, noisy, inconsistent and intentional errors.

Table 7 - Variables and Found Error Types

Error type	Variable
Incomplete data	Customer_Age
Incomplete data	Age_Group
Inconsistent data	Month
Inconsistent data	Country
Noise data	Product
Intentional error	Age_Column
Duplicates	Entire Table

After that, we modified the dataset from noisy data, i.e., we removed duplicated rows, imputed missing values with median in Customer_Age column, unified naming convention in Country and Months columns, replaced intentional error in Customer_Age with correct values, and removed noisy data from Product Column. It is found that the following variable are most relevant and other variables are rejected for the data modelling phase.

Table 8 - Role of Variables for Model

No.	Attribute	Description	Role
1	Date	Date of product purchase	Rejected
2	Day	Day of product purchase	Rejected
3	Month	Month of product purchase	Input
4	Year	Year of product purchase Range: 2011 - 2016	Input
5	Customer Age	Age of customer Range: 17 - 87	Rejected
6	Age Group	Age group of customers Adults (35-64), Young Adults (25-34), Youth (<25)	Input
7	Customer Gender	Gender of the customer ● F- Female ● M- Male	Input
8	Country	The country in which the purchase has been made. Available countries: United States – Australia – Canada – United Kingdom – Germany - France	Input
9	State	The state in which a purchase has been made. Available states: California - British Columbia – England – Washington - New South Wales	Input
10	Product Category	The category of the product purchased. Available categories: ● Accessories ● Bikes ● Clothing	Input
11	Subcategory	The subcategory of the purchased product. Available subcategories: Tires and Tubes - Bottles and Cages - Road Bikes – Helmets - Mountain Bikes	Input
12	Product	The purchased product. Available Products: ● Water Bottle - 30 oz. ● Patch Kit/8 Patches ● Mountain Tire Tube ● AWC Logo Cap ● Sport-100 Helmet, Red	Rejected

13	Order Quantity	The purchased quantity of a product. Range: 1 to 32	Input
14	Unit Cost	The cost of producing one product unit. Range: 1 to 2171	Rejected
15	Unit Price	Unit price of the product	Rejected
16	Profit	The profit of selling one product unit. Profit = Revenue - Cost Range: 2 to 3578	Rejected
17	Cost	The total cost of a purchase. Range: 1 to 43K	Rejected
18	Revenue	The revenue from a purchase. Range: 2 to 58.1K	Rejected
19	Profit Label	It shows profit is high or low ● High ● Low	Target

Then, we split the data into 50% training and 50% validation, and built three machine learning models, i.e., Decision Tree, Gradient Boosting, and Neural Network. Lastly, we assessed the performance of these models based on accuracy and Neural Network achieved the highest accuracy with 91.2%. The created model can predict whether the profits will be high or low based on the customers demographics such as gender, county, age, and other correlated attributes such as product categories and months.

These are the hidden patterns found during the implementation of SEMMA methdology

Table 9 - Hidden Patterns

No.	Hidden Pattern	Phase
1	June and December are highest sales month.	Explore
2	Hitch Rack – 4 Bike is the highest profit earning product in all countries.	Explore
3	The most profitable customers are between age 30 to 60.	Explore
4	If a customer buys more than 18 socks in an order, then it will be profitable for the business.	Model
5	If bottles, caps and cleaners are sold less than 10 in an order, then it will not be a profitable sale.	Model
6	If more than 3 clothing items are sold in a sale transaction, then it will be a profitable sale for the business	Model

7 References

Campíñez-Romero, S., Colmenar-Santos, A., Pérez-Molina, C., & Mur-Pérez, F. (2018). A hydrogen refuelling stations infrastructure deployment for cities supported on fuel cell taxi roll-out. *Energy*, *148*, 1018-1031. Doll, N., & Vetter, P. (2017). Fahrverbot Für Diesel-Pkw rückt näher (WWW Doc-ument). *Die Welt*.

International Energy Agency. (2009). *World energy outlook* (p. 17). Paris: OECD/IEA. Pucher, J., & Buehler, R. (2008). Making cycling irresistible: lessons from the Netherlands, Denmark and Germany. *Transport reviews*, *28*(4), 495-528.

Wałdykowski, P., Adamczyk, J., & Dorotkiewicz, M. (2021). Sustainable Urban Transport—Why a Fast Investment in a Complete Cycling Network Is Most Profitable for a City. *Sustainability*, *14*(1), 119.

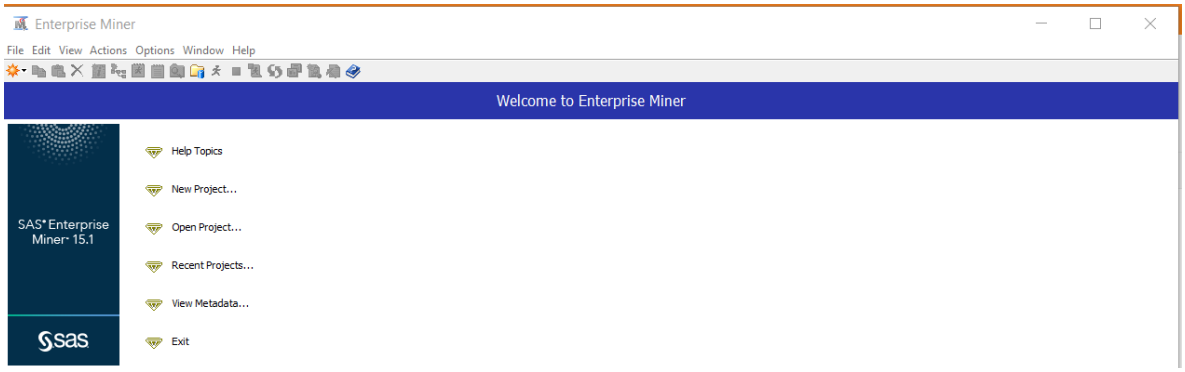
8 Appendix

SAS Enterprise Miner project Steps of Sample and Explore part of SEMMA are described below:

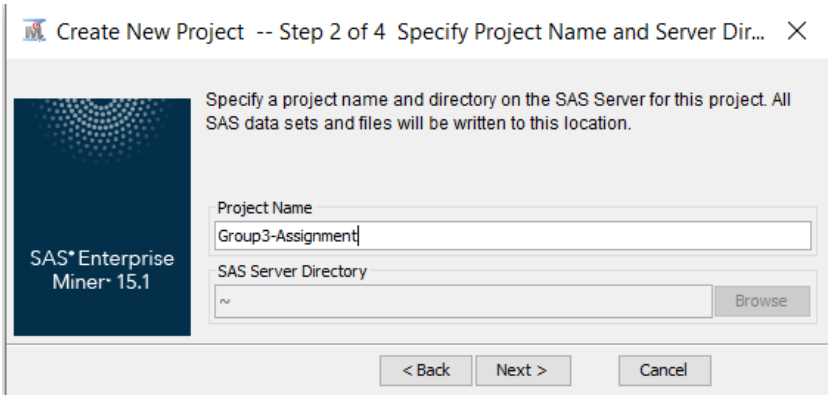
SAMPLE

1. Steps for Creating SAS Enterprise Miner Project

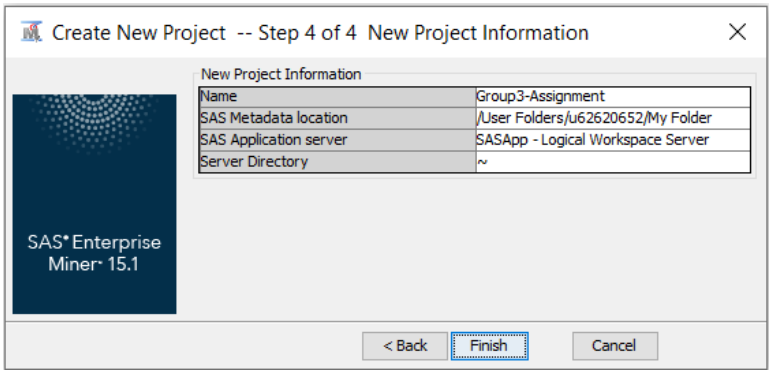
- Go to File menu and click ‘New’ option. Choose Project and click New Project



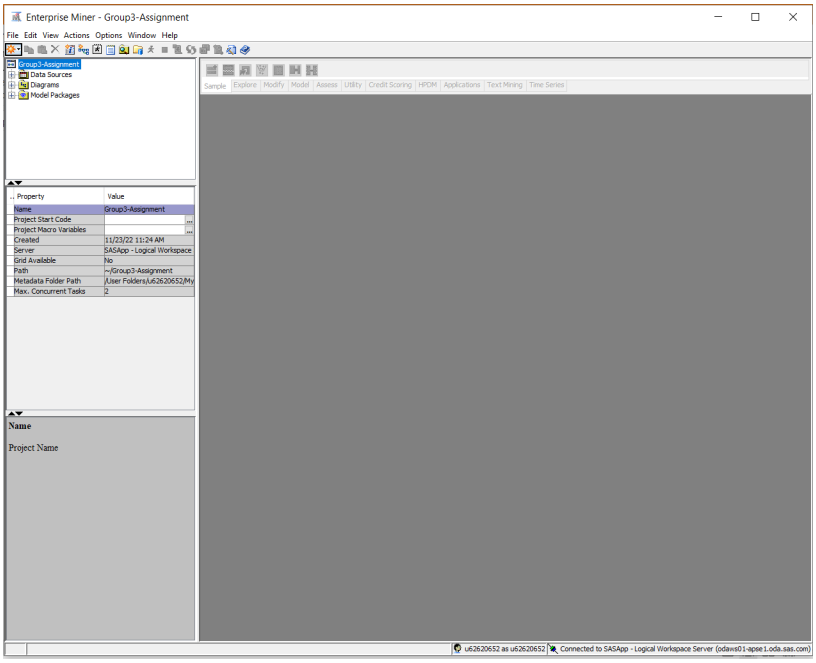
- Project Name is **Group3-Assignment**
- Click Next



- Click Finish

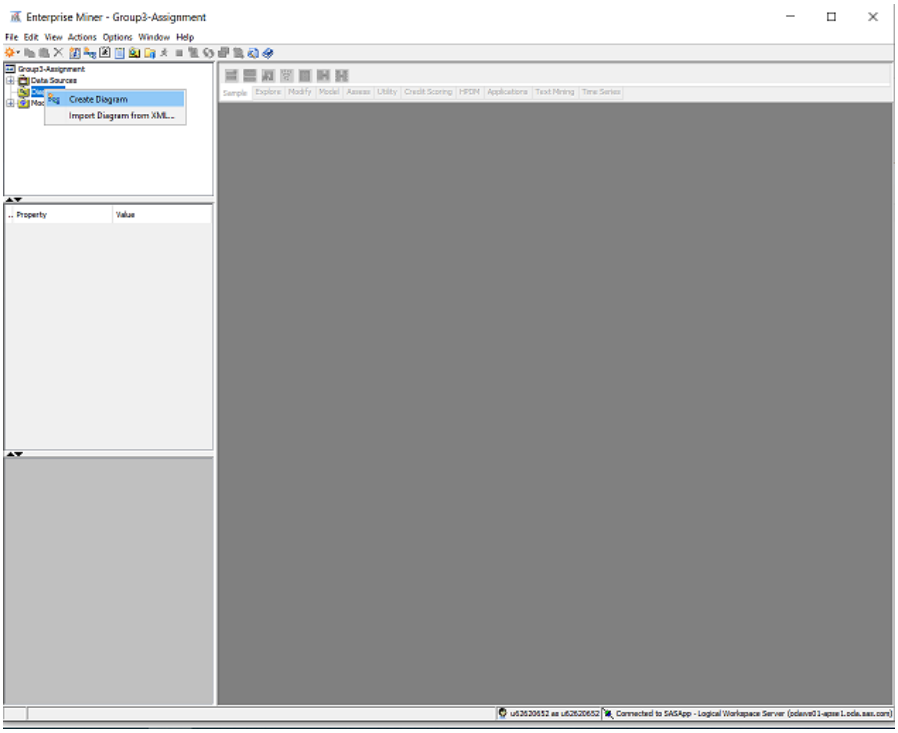


- Project created and opened default view.

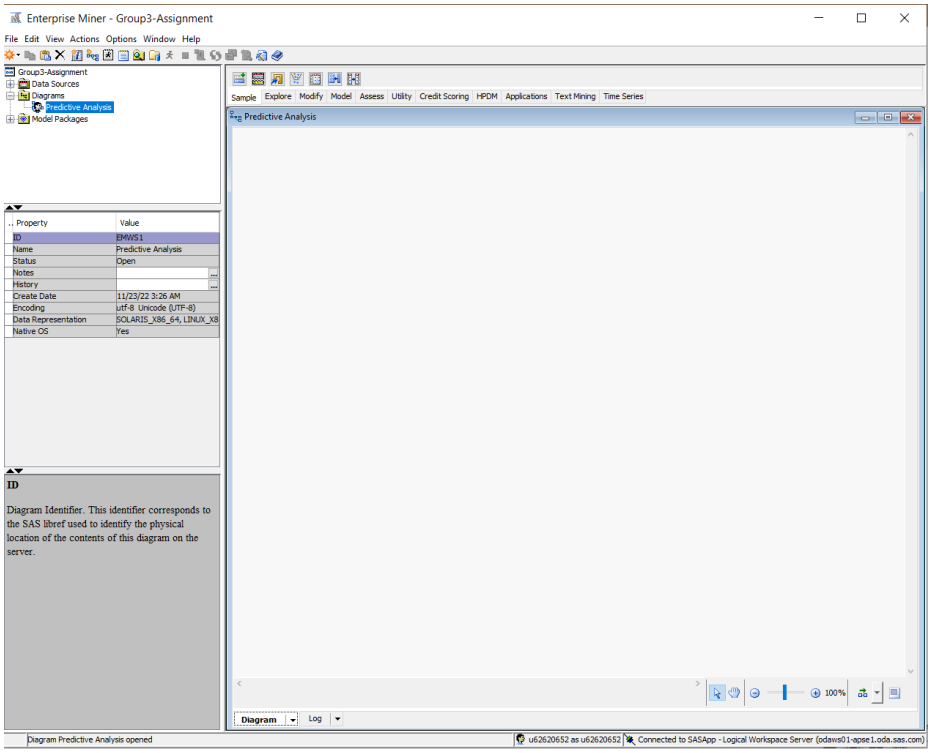


2. Create SAS Enterprise Miner Diagram

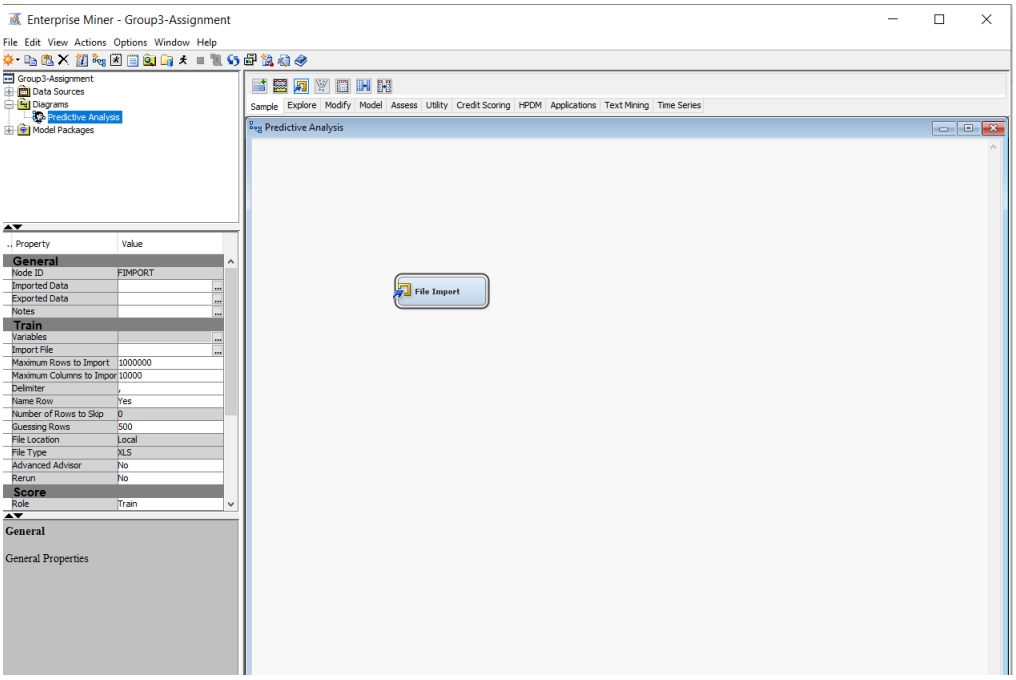
- Right click on Diagrams directory of project and choose Create Diagram option



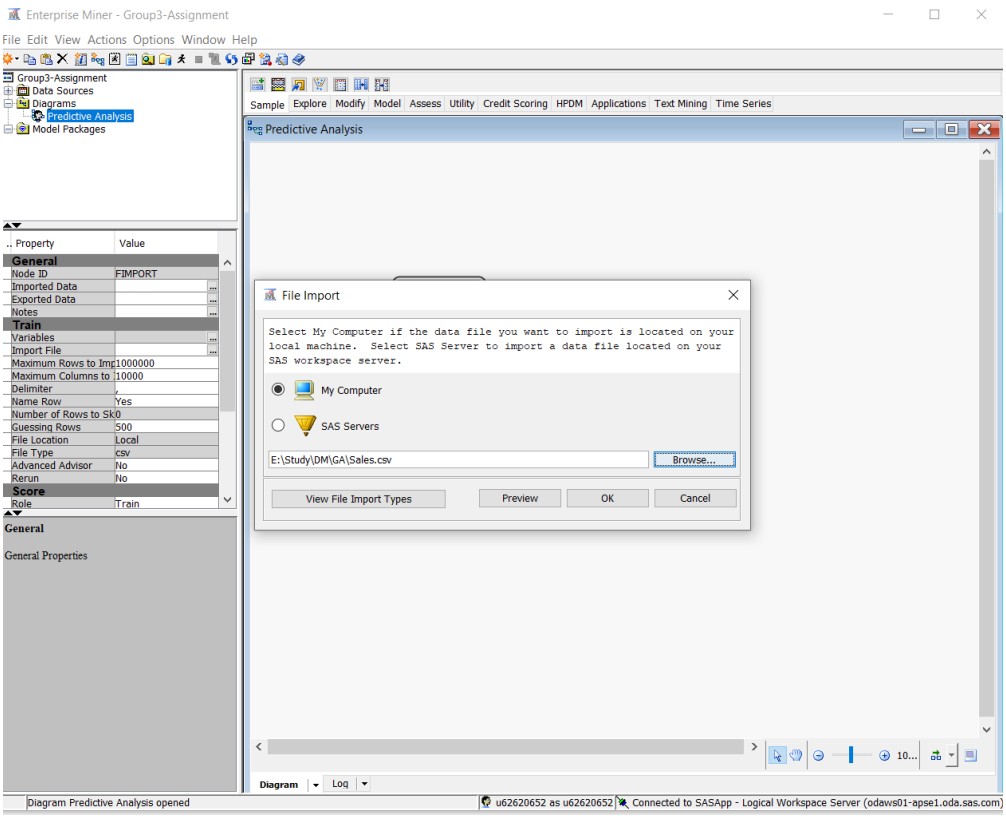
- Diagram Name is typed as ‘Predictive Analysis’, click OK
- Diagram created and opened default empty window



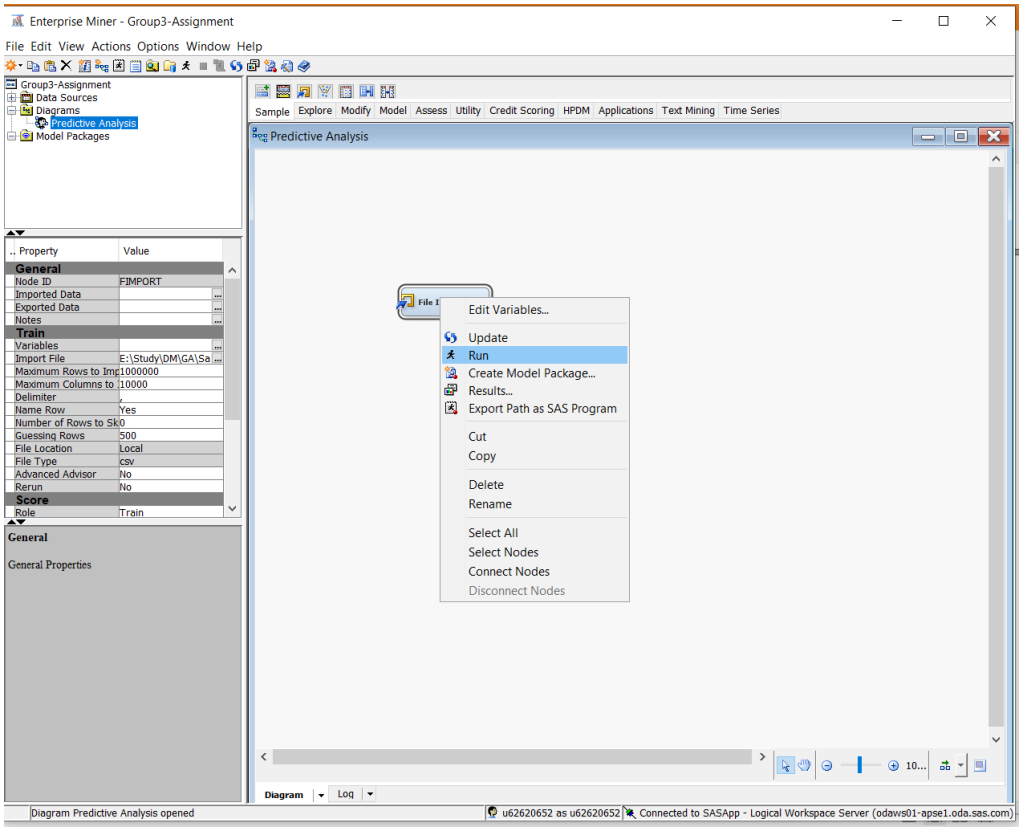
- File Input node is dragged from Sample tab to project workspace



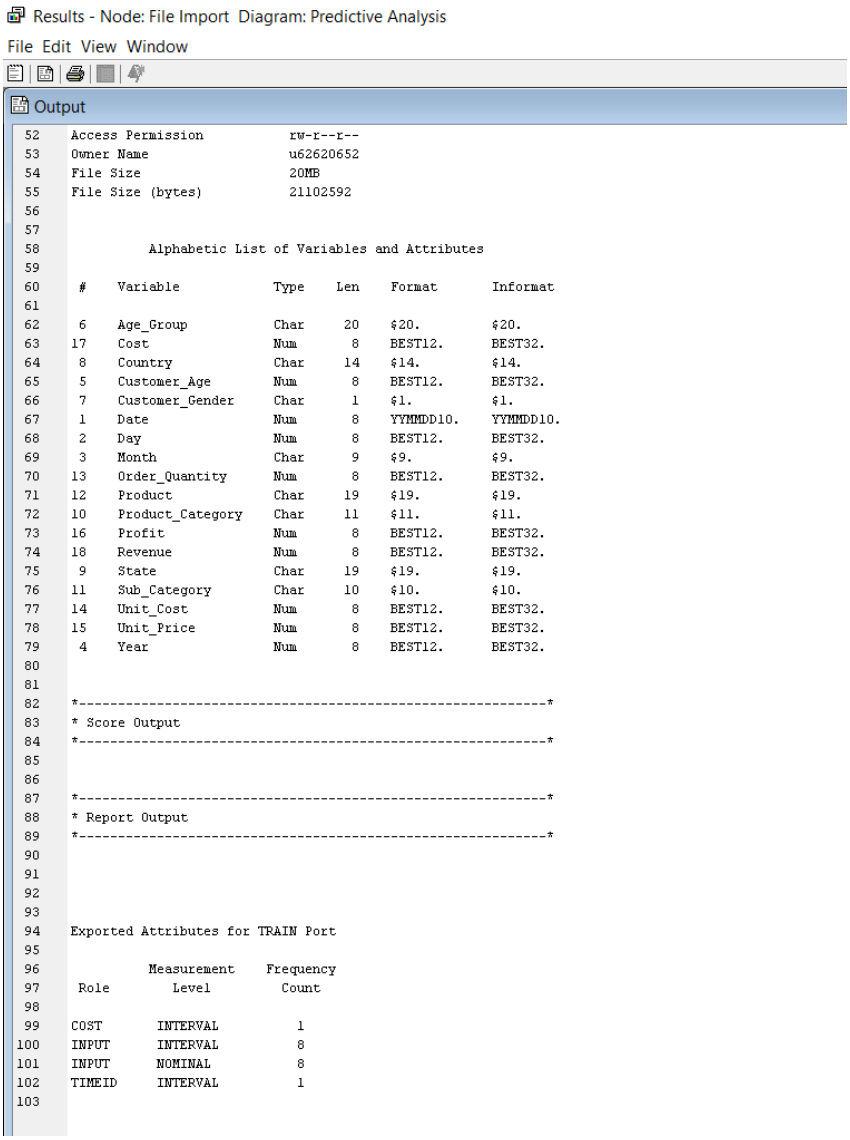
- File import option is selected and browse dataset csv local file. Click OK



- Run the File Import node by right clicking and selecting the option

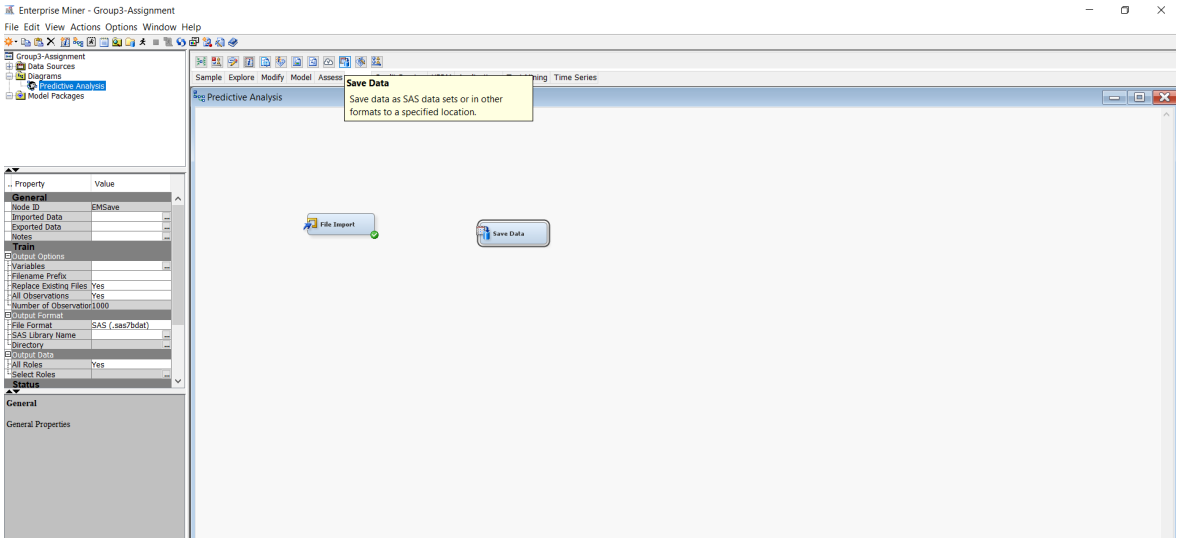


- Running the File Import node is successful and shown by green tick on right bottom of the node. Click Results from the success message dialog.

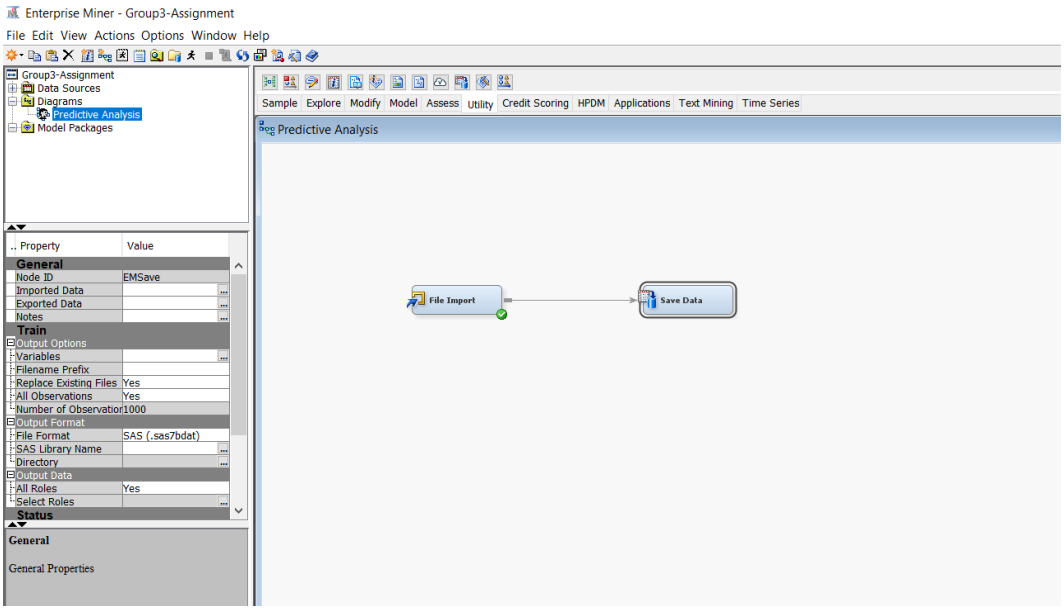



```
Access Permission           rw-r--r--
Owner Name                  u62620652
File Size                   20MB
File Size (bytes)           21102592


          Alphabetic List of Variables and Attributes

 #    Variable            Type    Len    Format       Informat

 6    Age_Group           Char     20    $20.         $20.
17    Cost                Num       8    BEST12.      BEST32.
 8    Country             Char     14    $14.         $14.
 5    Customer_Age        Num       8    BEST12.      BEST32.
 7    Customer_Gender     Char      1    $1.          $1.
 1    Date                Num       8    YYMMDD10.    YYMMDD10.
 2    Day                 Num       8    BEST12.      BEST32.
 3    Month               Char      9    $9.          $9.
13    Order_Quantity      Num       8    BEST12.      BEST32.
12    Product             Char     19    $19.         $19.
10    Product_Category    Char     11    $11.         $11.
16    Profit              Num       8    BEST12.      BEST32.
18    Revenue             Num       8    BEST12.      BEST32.
 9    State               Char     19    $19.         $19.
11    Sub_Category        Char     10    $10.         $10.
14    Unit_Cost           Num       8    BEST12.      BEST32.
15    Unit_Price          Num       8    BEST12.      BEST32.
 4    Year                Num       8    BEST12.      BEST32.


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


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




Exported Attributes for TRAIN Port

           Measurement    Frequency
 Role         Level         Count

COST       INTERVAL          1
INPUT      INTERVAL          8
INPUT      NOMINAL           8
TIMEID     INTERVAL          1
```

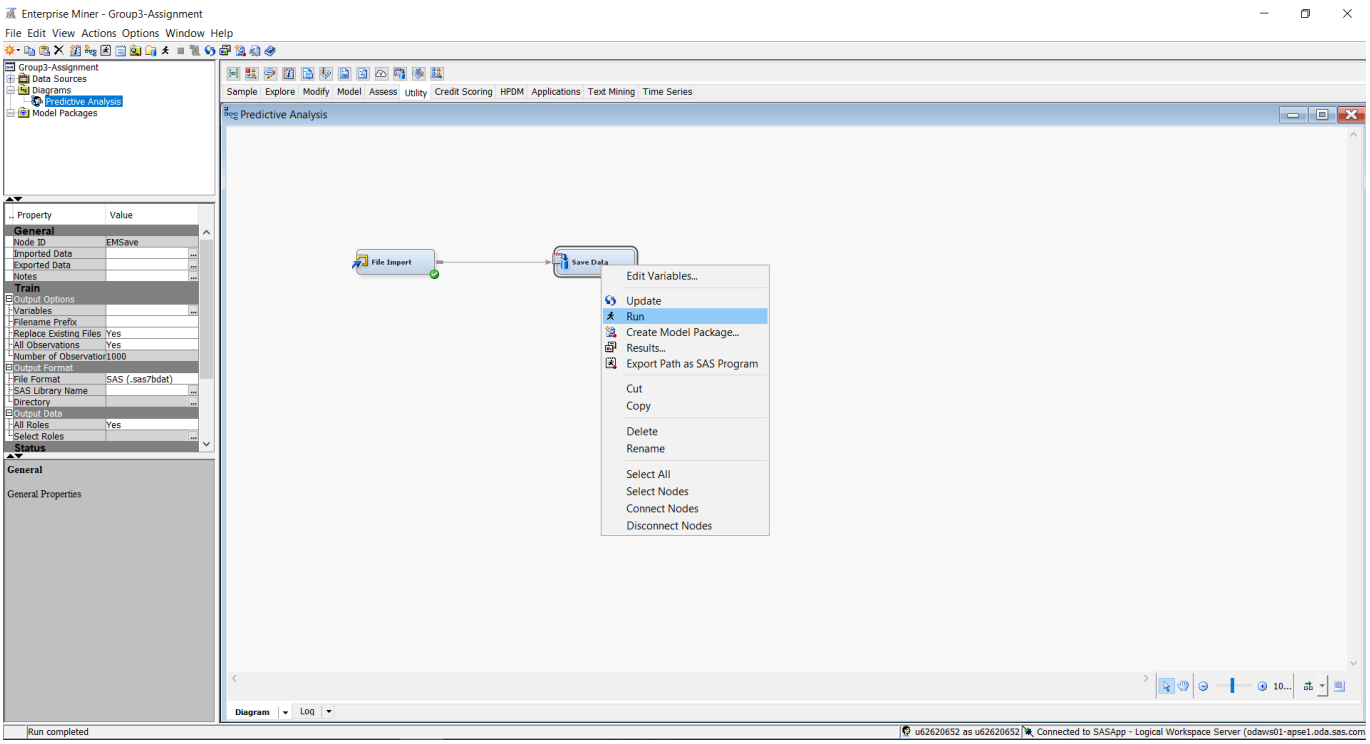
- Save Data node is dragged from utility tab to project workspace


Enterprise Miner - Group3-Assignment
File Edit View Actions Options Window Help
Group3-Assignment
Data Sources
Diagrams
Predictive Analysis
Model Packages
Sample Explore Modify Model Assess Time Series
Predictive Analysis
Save Data
Save data as SAS data sets or in other formats to a specified location.
File Import
Save Data
Property
Value
General
Node ID
EMSave
Imported Data
Exported Data
Notes
Train
Output Options
Variables
Filename Prefix
Replace Existing Files
Yes
All Observations
Yes
Number of Observation
1000
Output Format
File Format
SAS (.sas7bdat)
SAS Library Name
Directory
Output Data
All Roles
Yes
Select Roles
Status
General
General Properties

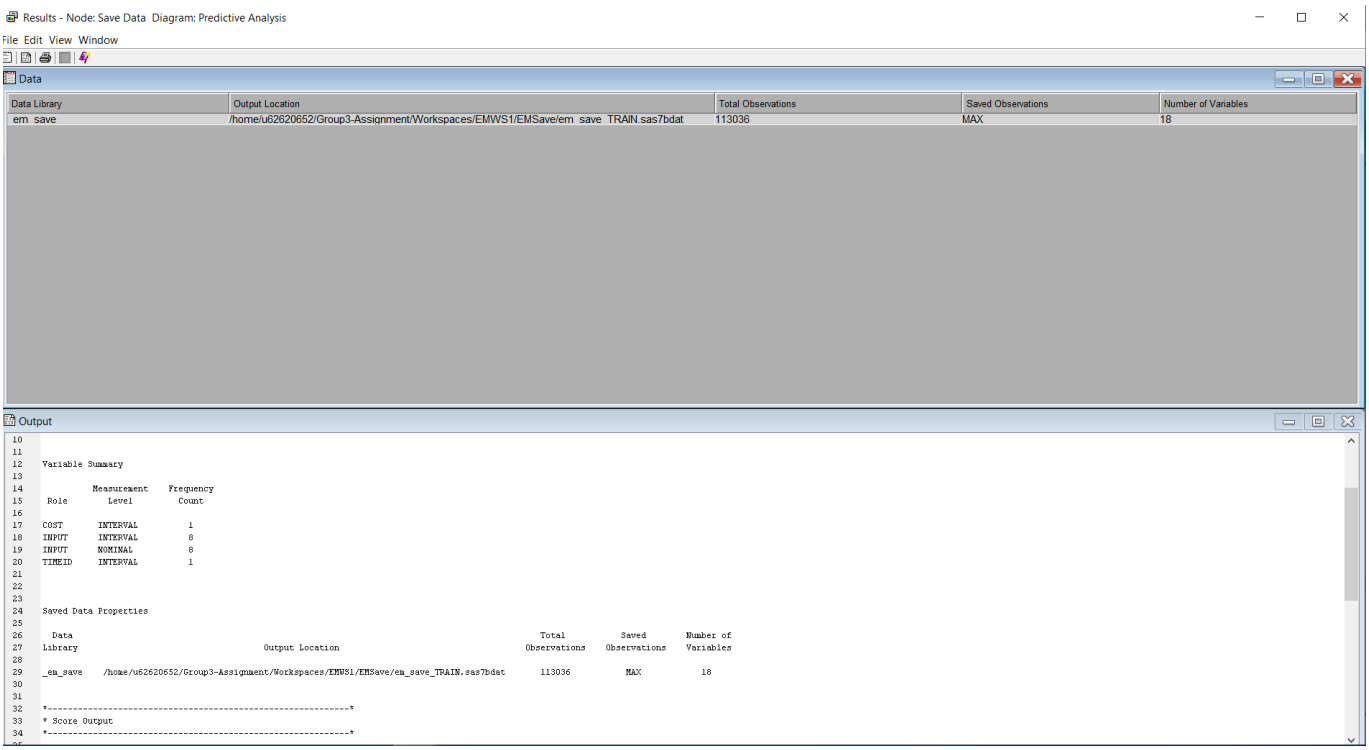
- Connect Save Data node to File Import node.



- Run the Save Data node by right clicking and selecting the option

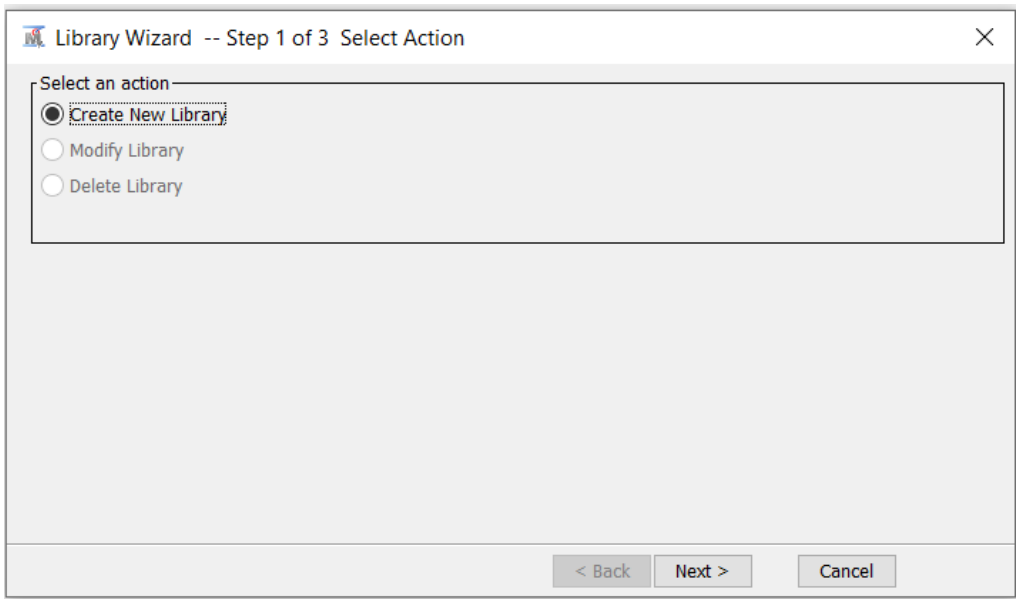


- Running the Save Data node is successful and shown by green tick on right bottom of the node. Click Results from the success message dialog.

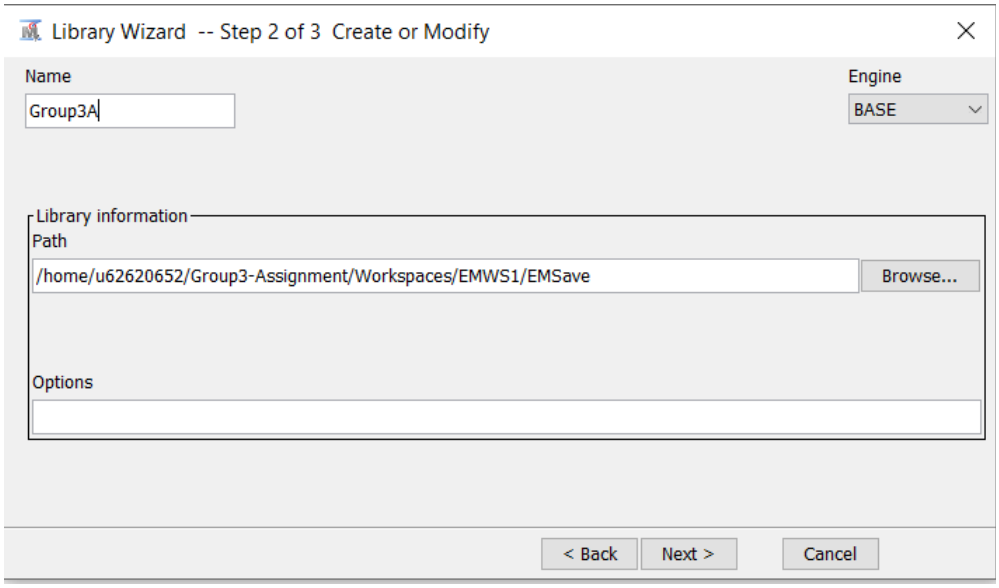


3. Steps for Creating SAS Enterprise Miner Library

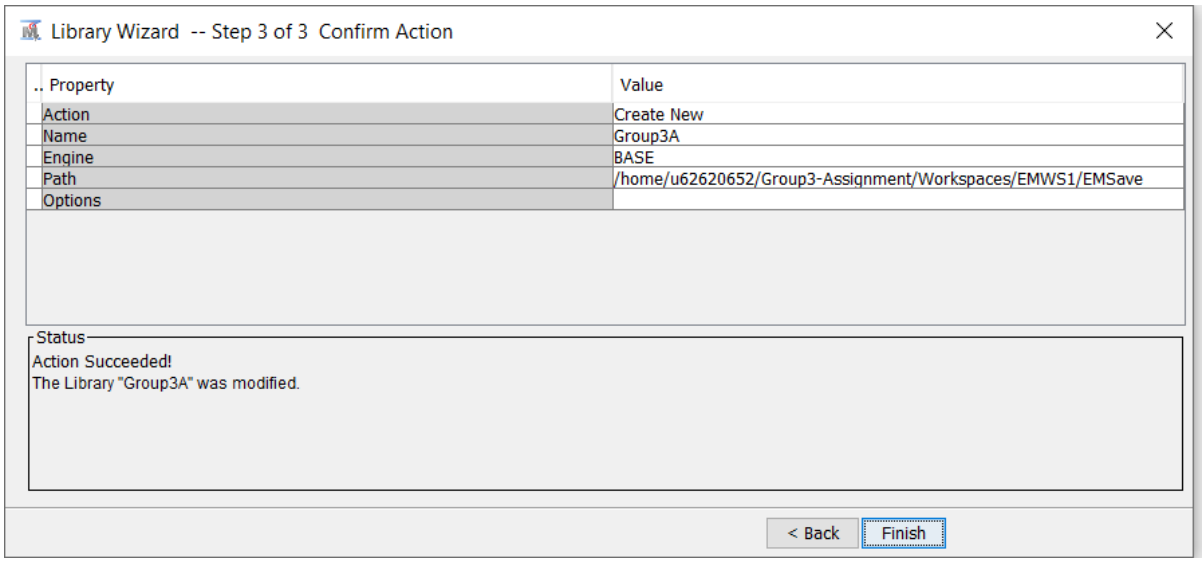
- Click File menu and select ‘New’ option. Click New Library
- Wizard is opened and choose Create New Library option. Click Next



- Library Name is typed **Group3A**
- Provide library path in the browse option and click Next

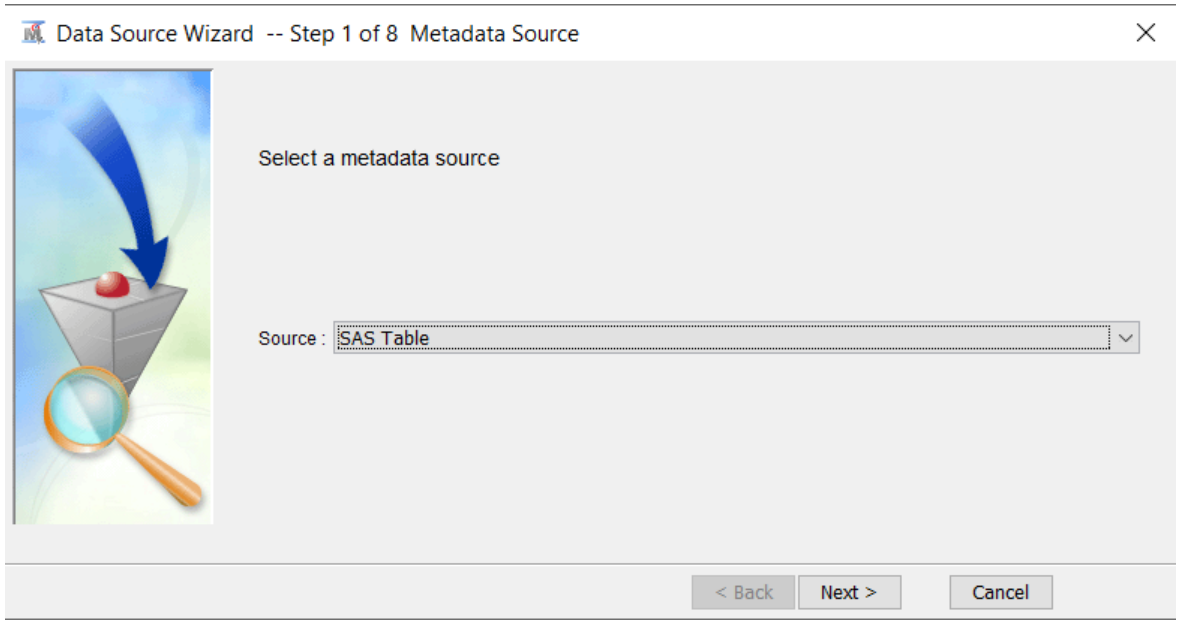

Library Wizard -- Step 2 of 3 Create or Modify
Name
Group3A
Engine
BASE
Library information
Path
/home/u62620652/Group3-Assignment/Workspaces/EMWS1/EMSave
Browse...
Options
< Back
Next >
Cancel

- Library created successfully and click Finish.

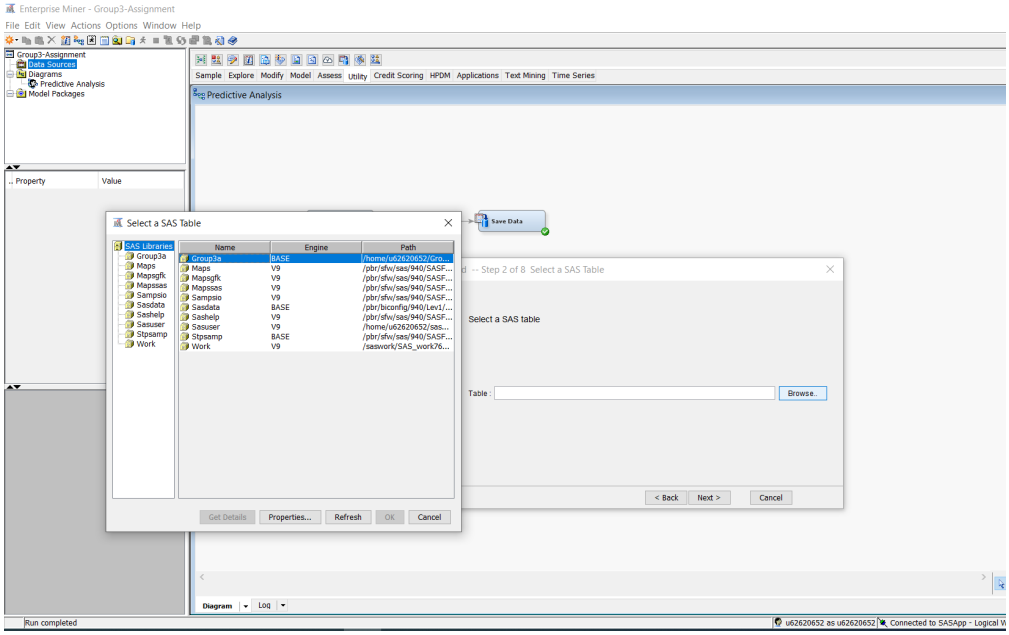


4. Steps for Creating Data Source

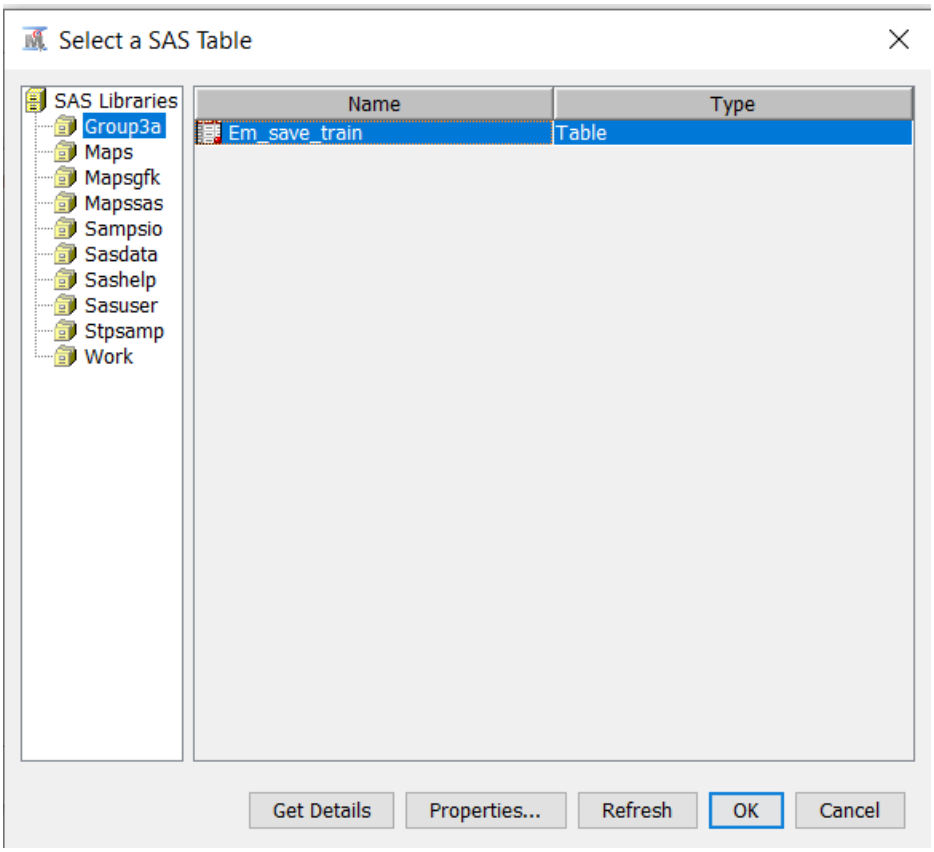
- Select Create Data Source option by right clicking the Data Sources directory of the project. Keep SAS Table option selected and click Next.


Data Source Wizard -- Step 1 of 8 Metadata Source
Select a metadata source
Source :
SAS Table
< Back
Next >
Cancel

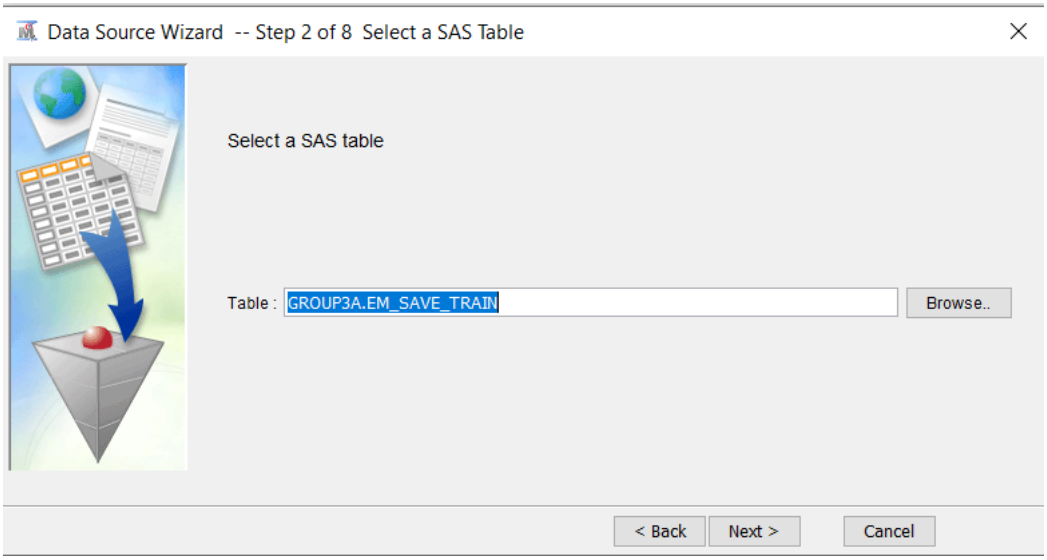
- Provide data source from browse option.



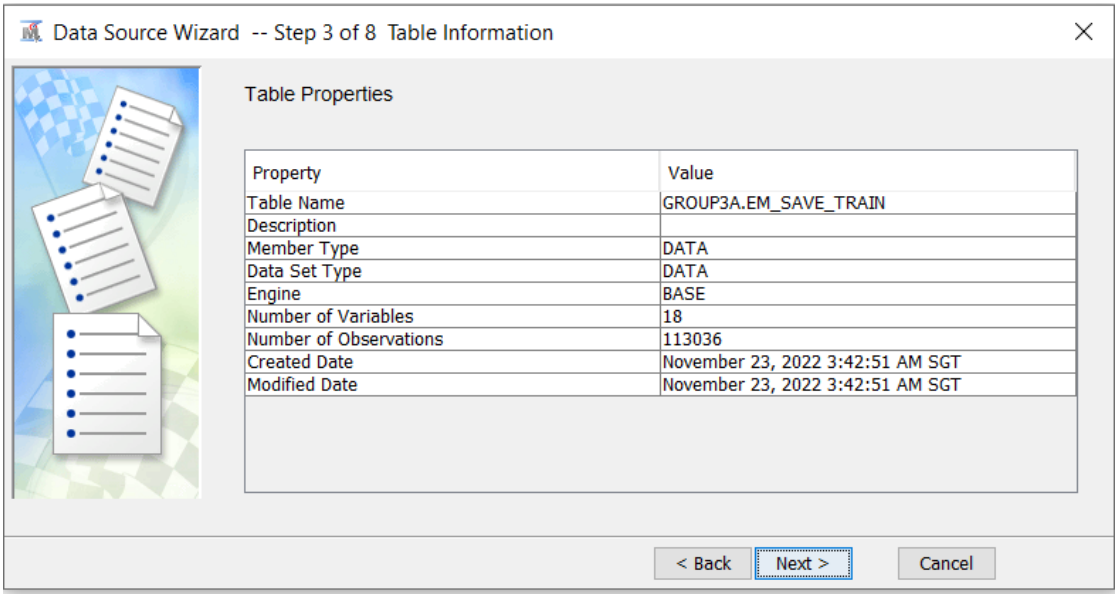
- Click OK



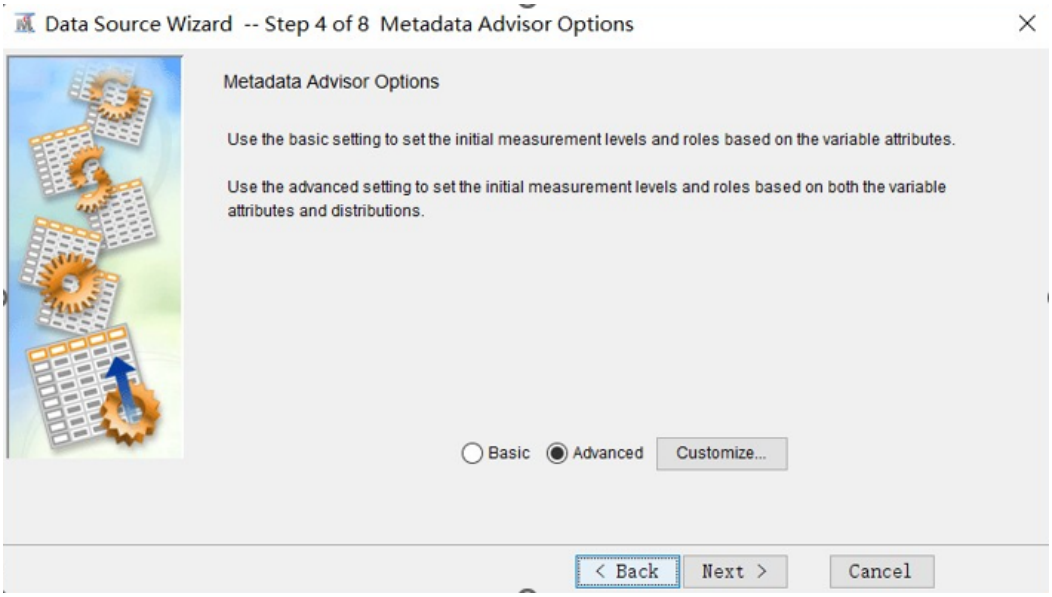
- Click Next



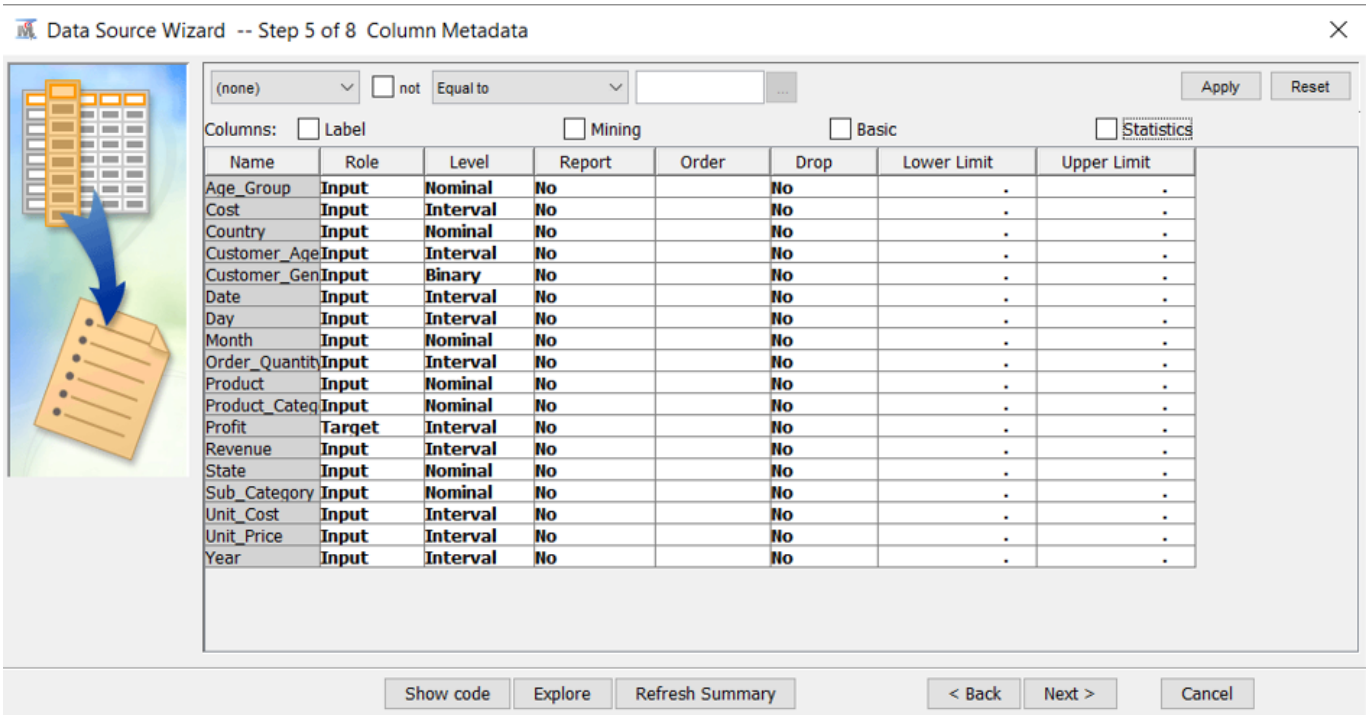
- Click Next


Data Source Wizard -- Step 3 of 8 Table Information
Table Properties
Property
Value
Table Name
GROUP3A.EM_SAVE_TRAIN
Description
Member Type
DATA
Data Set Type
DATA
Engine
BASE
Number of Variables
18
Number of Observations
113036
Created Date
November 23, 2022 3:42:51 AM SGT
Modified Date
November 23, 2022 3:42:51 AM SGT
< Back
Next >
Cancel

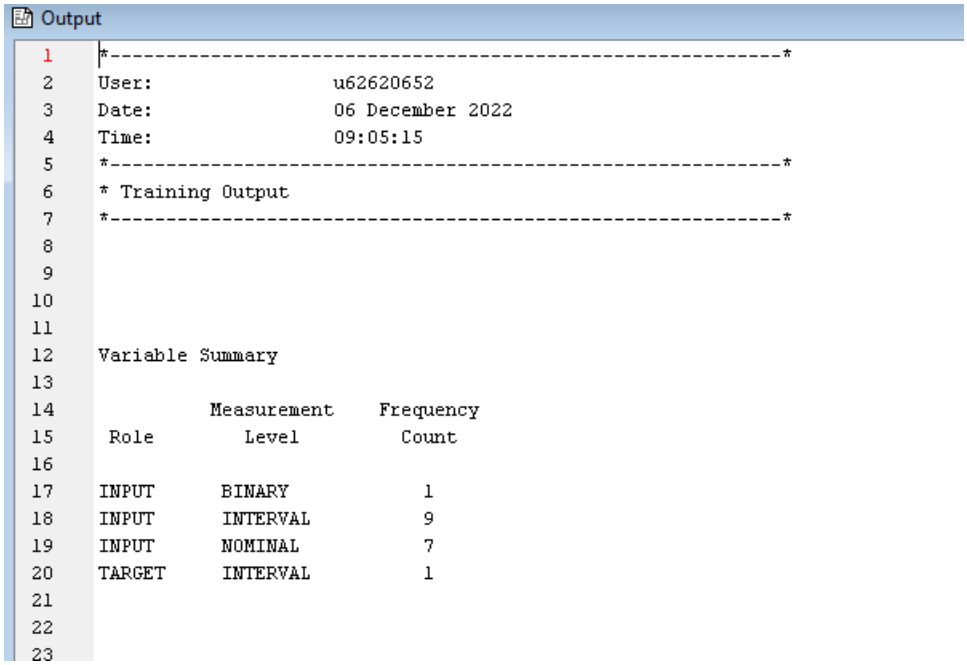
- Choose Advanced option and click Next



- Role and Level of the columns had been corrected in advanced option. Click Next.



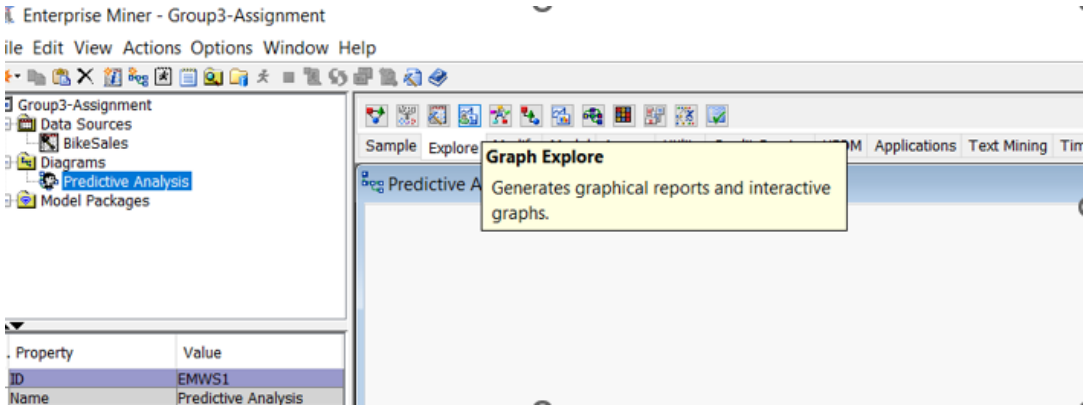
- Columns Metadata has been configured and ready for the Explore Phase.



EXPLORE

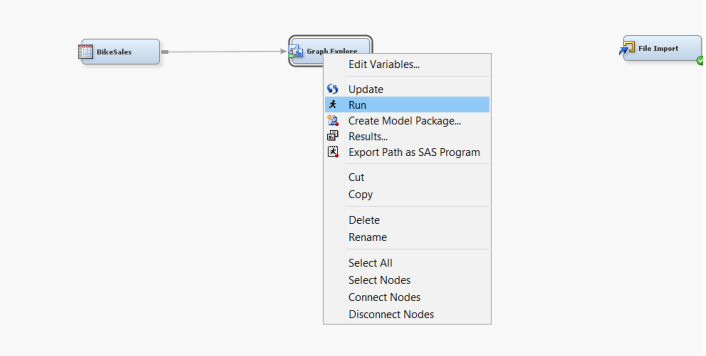
1. Steps for Creating Boxplot, Histogram and Pie Chart

1. Click Explore tab and drag Graph Explore node to the project workspace

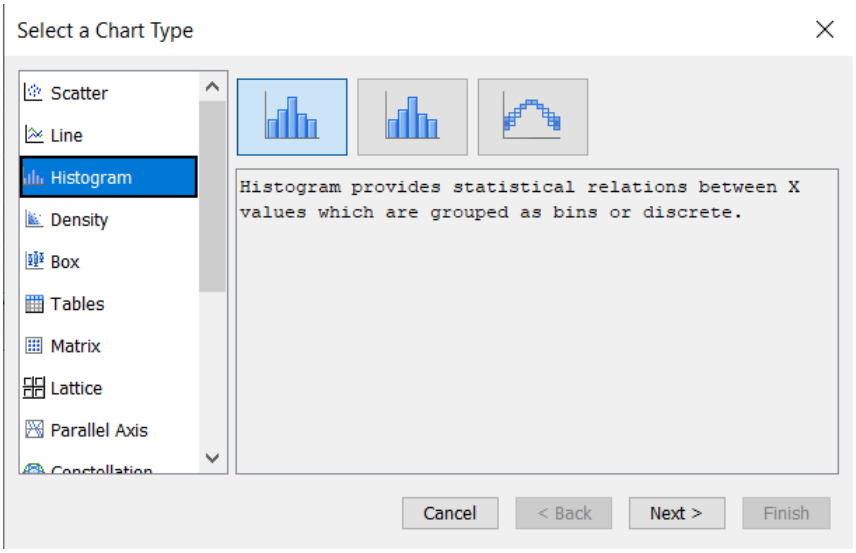


2. Data source and Graph Explore nodes are connected.

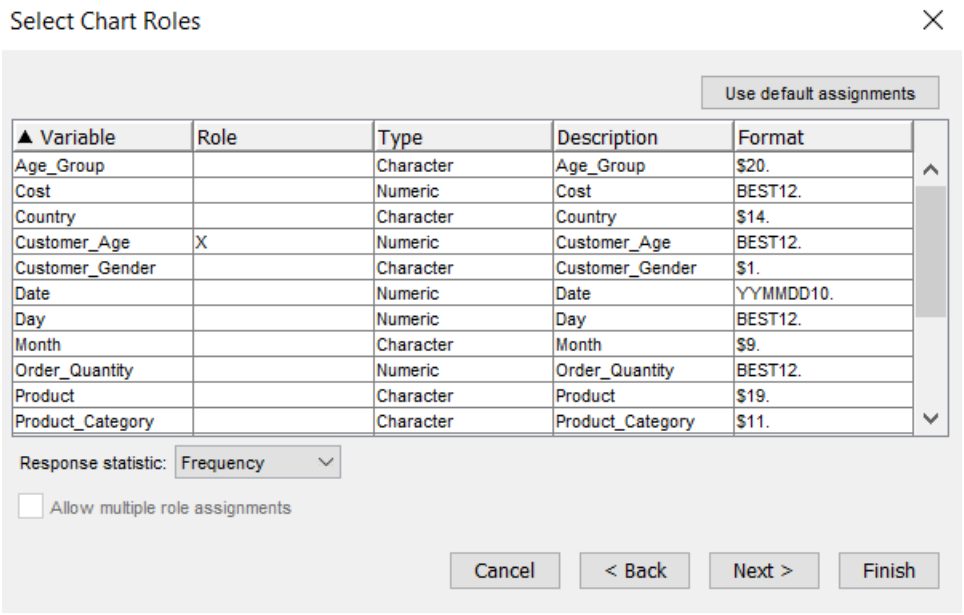
- Run the Graph Explore node



3. Results are shown and select plot option from View Menu. Select Histogram option

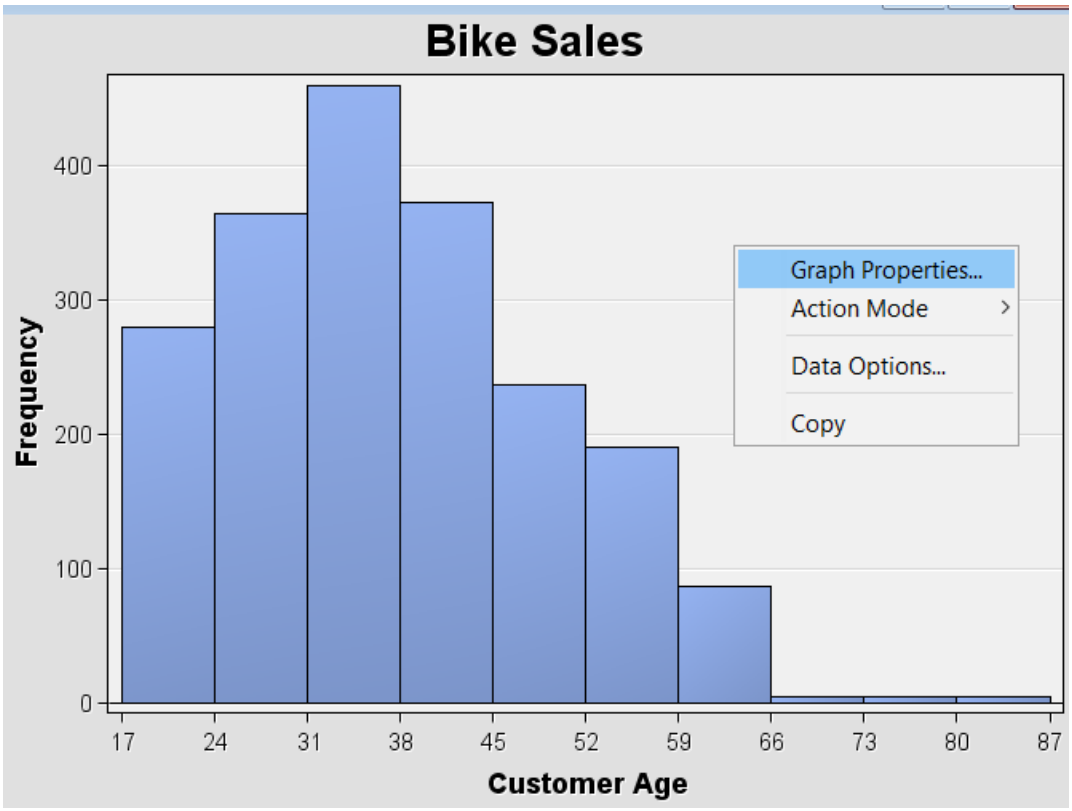


4. Select any variable from and click Next

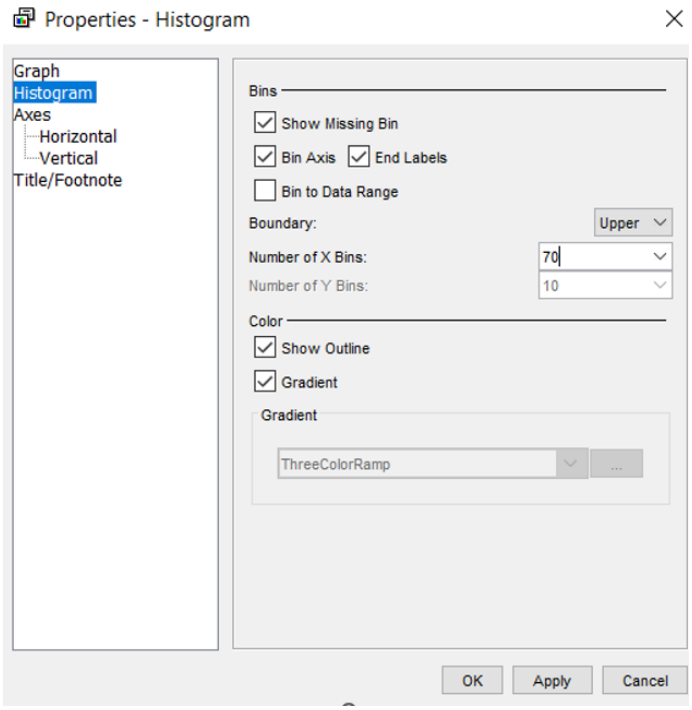


***2.* Steps for Changing graph properties including adding missing bin changing number of bins**

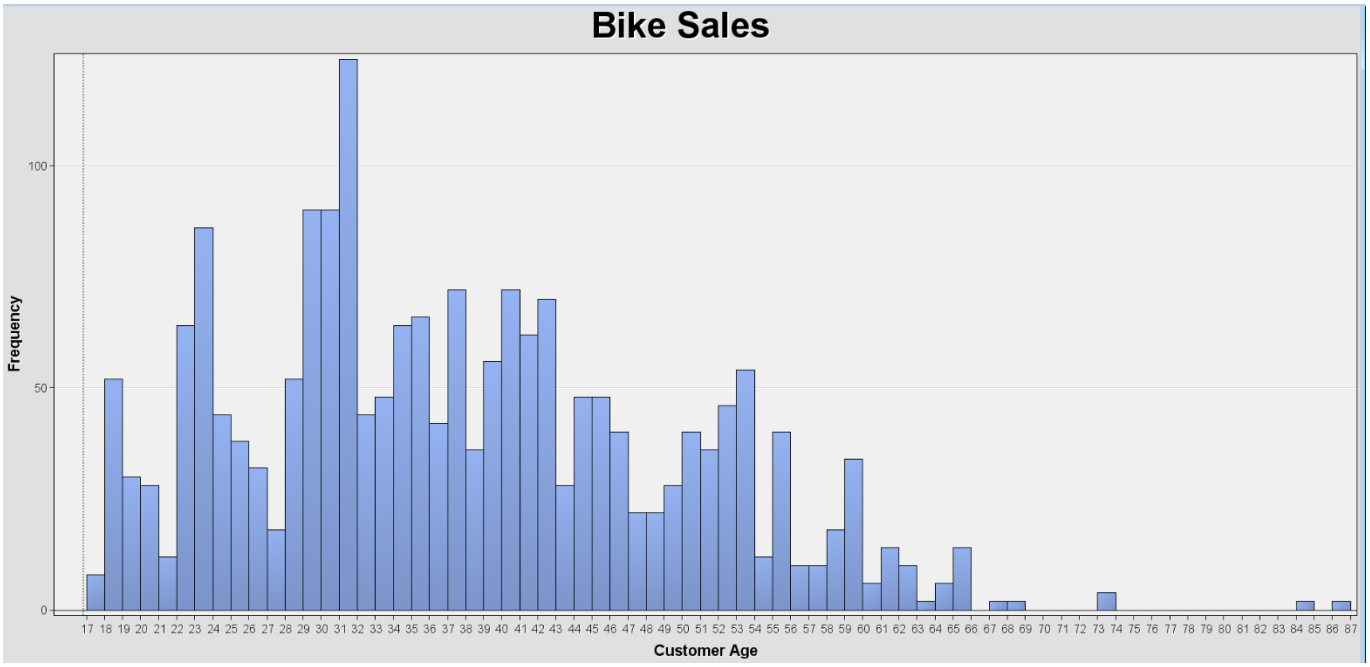
1. Graph properties can be opened by right clicking on the plotted histogram



2. Calculate the number of bins for the variable and enter the value in Number of X Bins. Check Show Missing Bin option to show in the histogram.

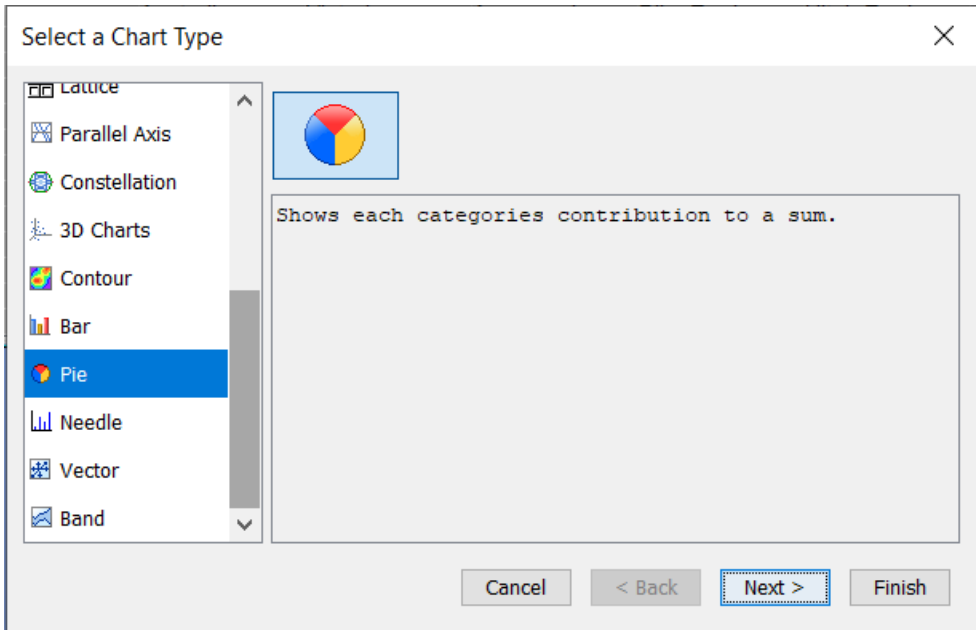


3. Histogram after changing properties is shown.

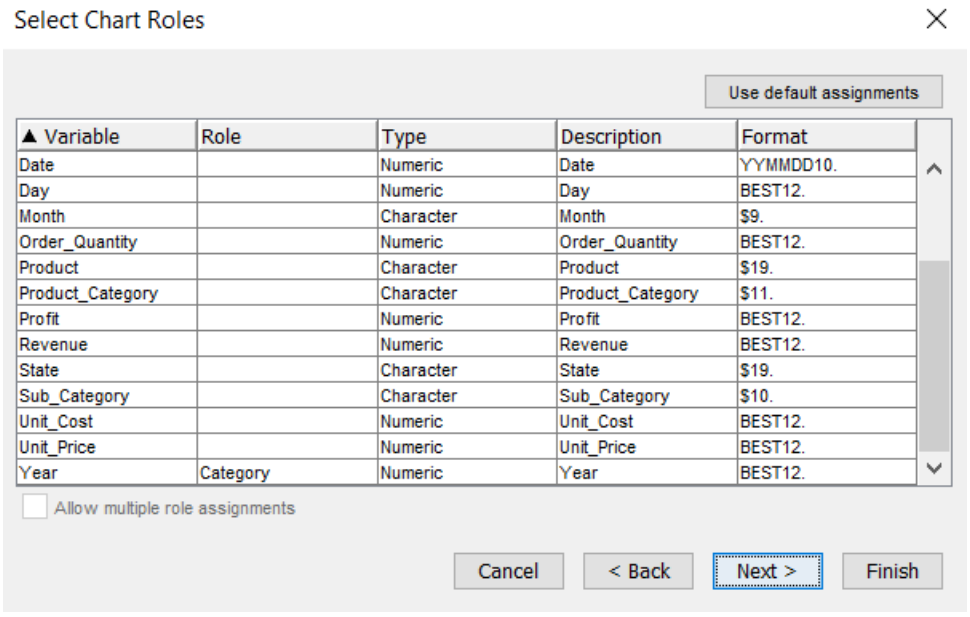


3. Steps for Variable Association, Variable Selection and Summary Statistics

1. Select Pie chart from Plot option



2. Select variable to show pie chart

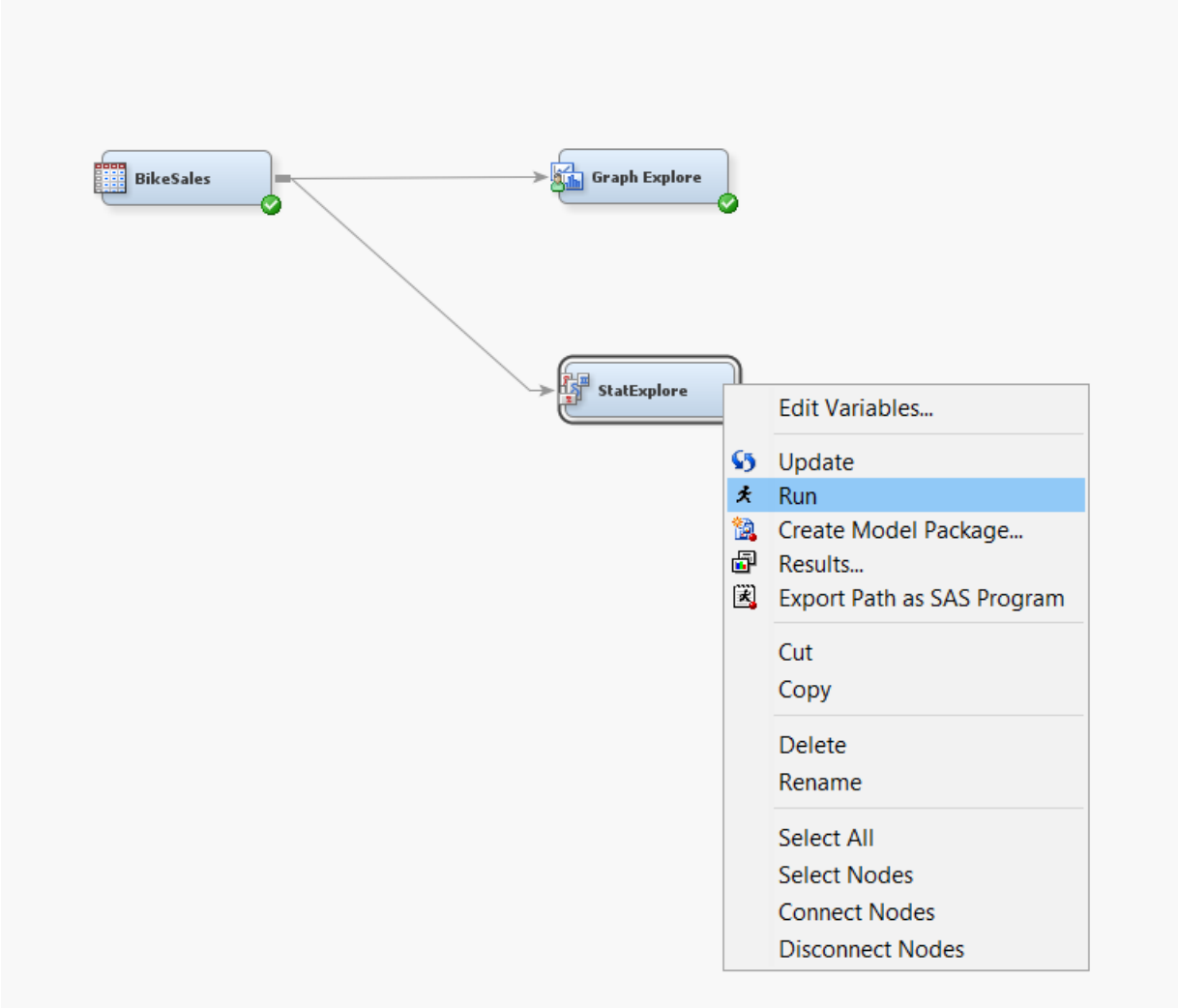


▲ Variable	Role	Type	Description	Format
Date		Numeric	Date	YYMMDD10.
Day		Numeric	Day	BEST12.
Month		Character	Month	$9.
Order_Quantity		Numeric	Order_Quantity	BEST12.
Product		Character	Product	$19.
Product_Category		Character	Product_Category	$11.
Profit		Numeric	Profit	BEST12.
Revenue		Numeric	Revenue	BEST12.
State		Character	State	$19.
Sub_Category		Character	Sub_Category	$10.
Unit_Cost		Numeric	Unit_Cost	BEST12.
Unit_Price		Numeric	Unit_Price	BEST12.
Year	Category	Numeric	Year	BEST12.

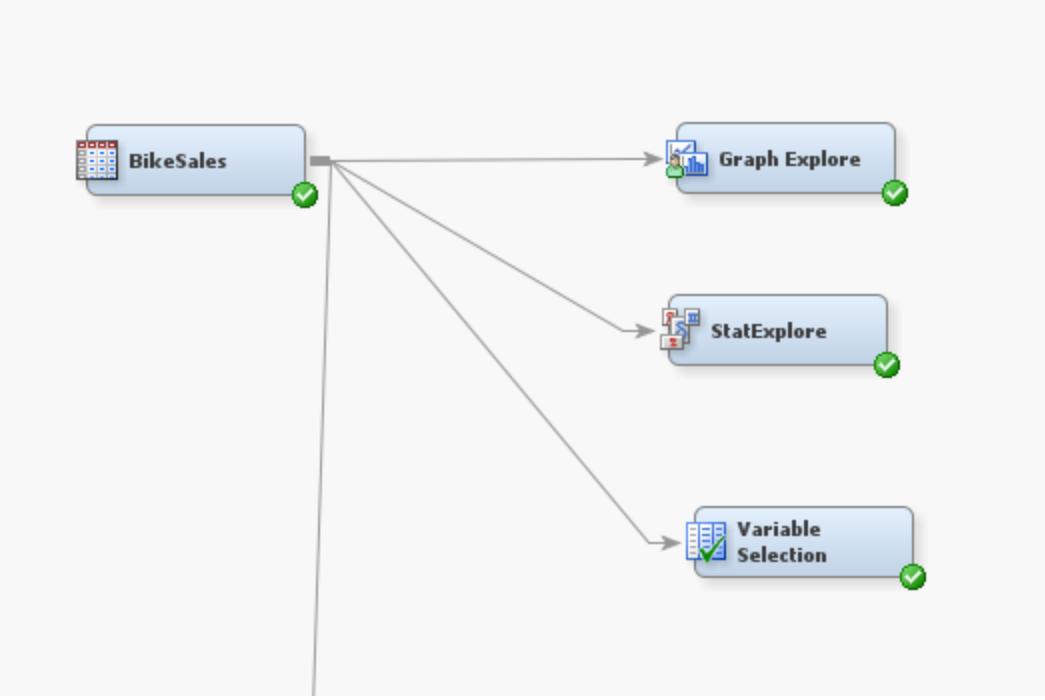
3. Select data on any plot and it will show same data on remaining plotted graphs. Product Category

and Customer Age association is shown.

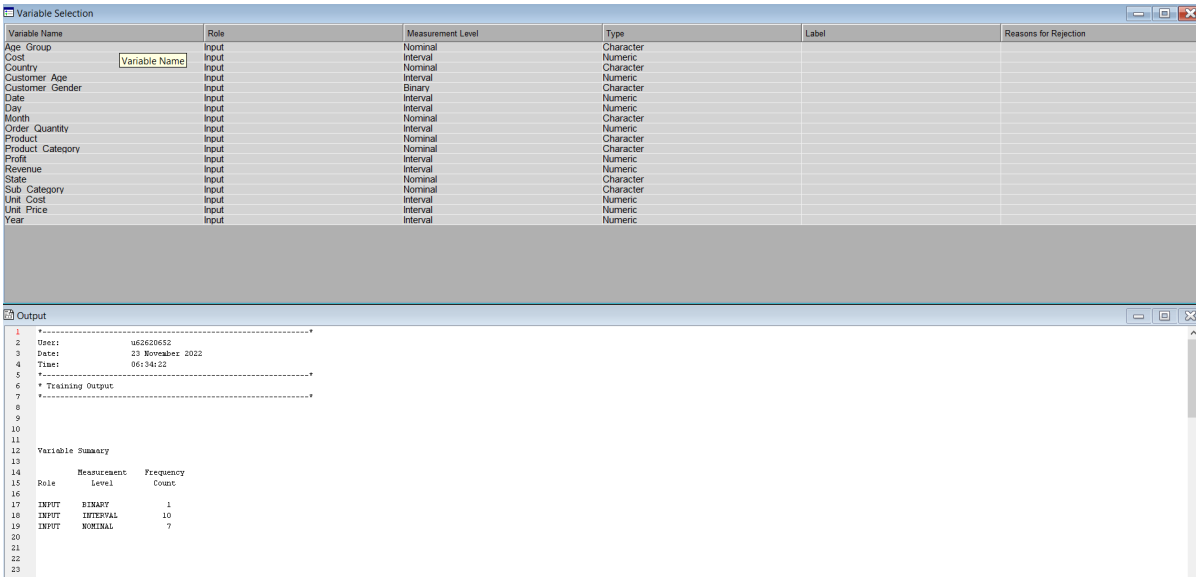
4. Drag StatExplorer node to project workspace and connect to data source. Run it by right clicking.



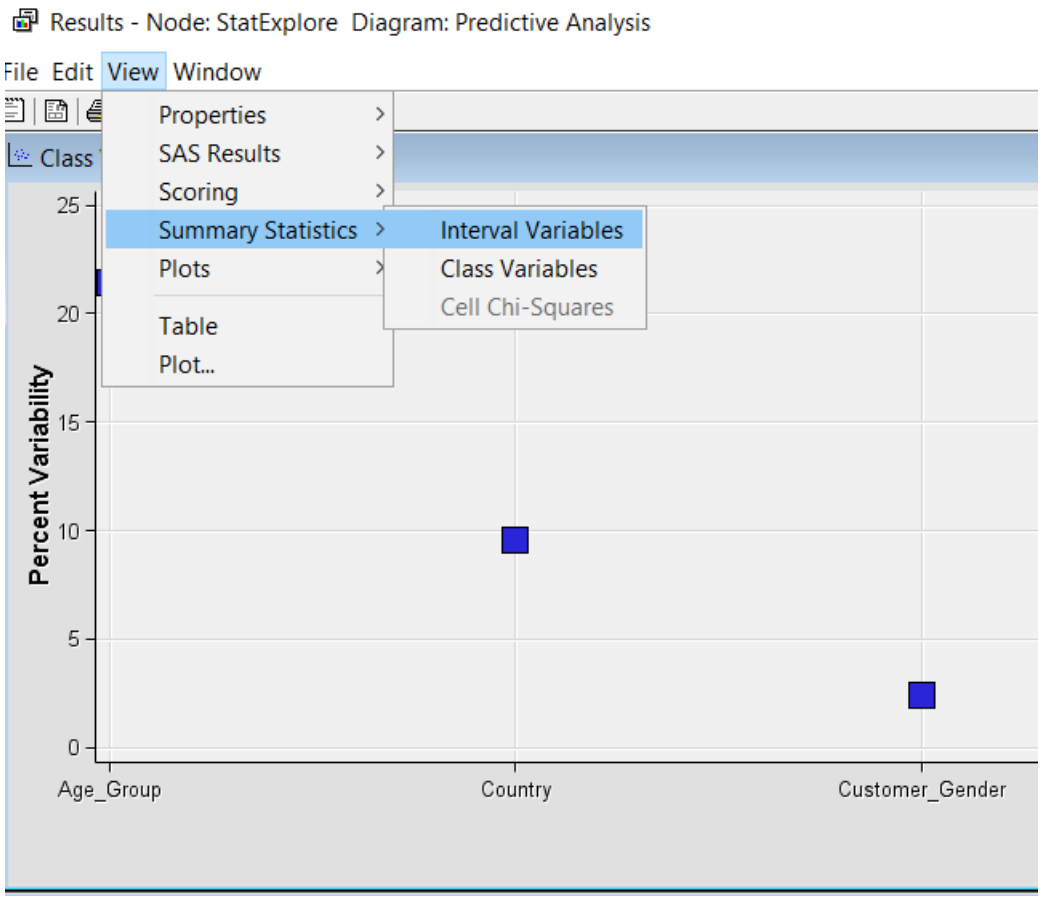
5. Drag Variable Selection node to project workspace and connect to data source. Run it by right clicking.



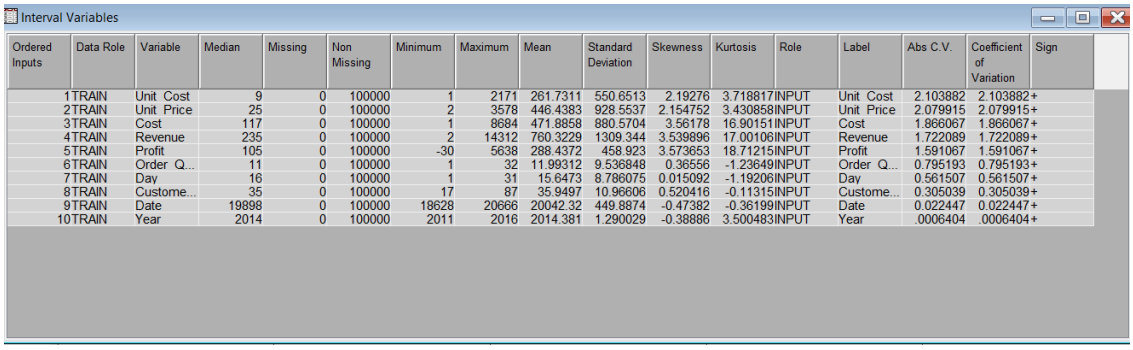
6. Results of Variable Selection are shown.



7. Select Interval Variables from Summary Statistics option of View Menu.



8. Results of Summary Statistics are shown.

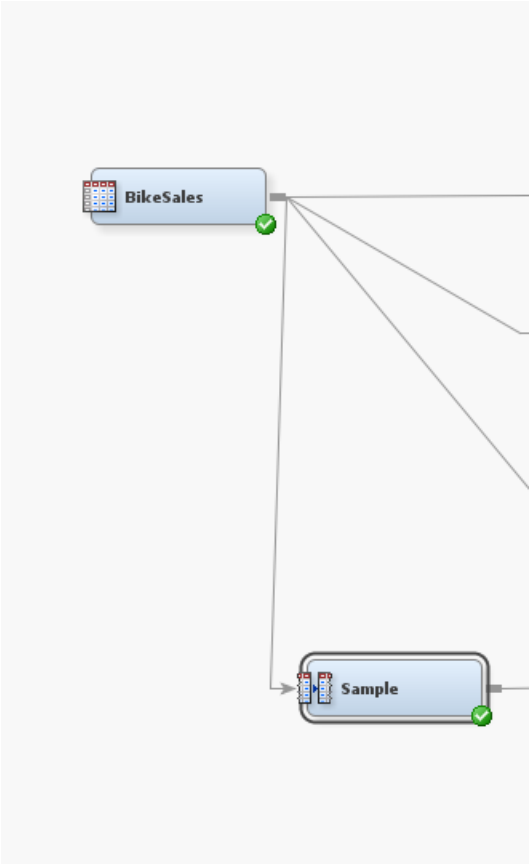


Interval Variables

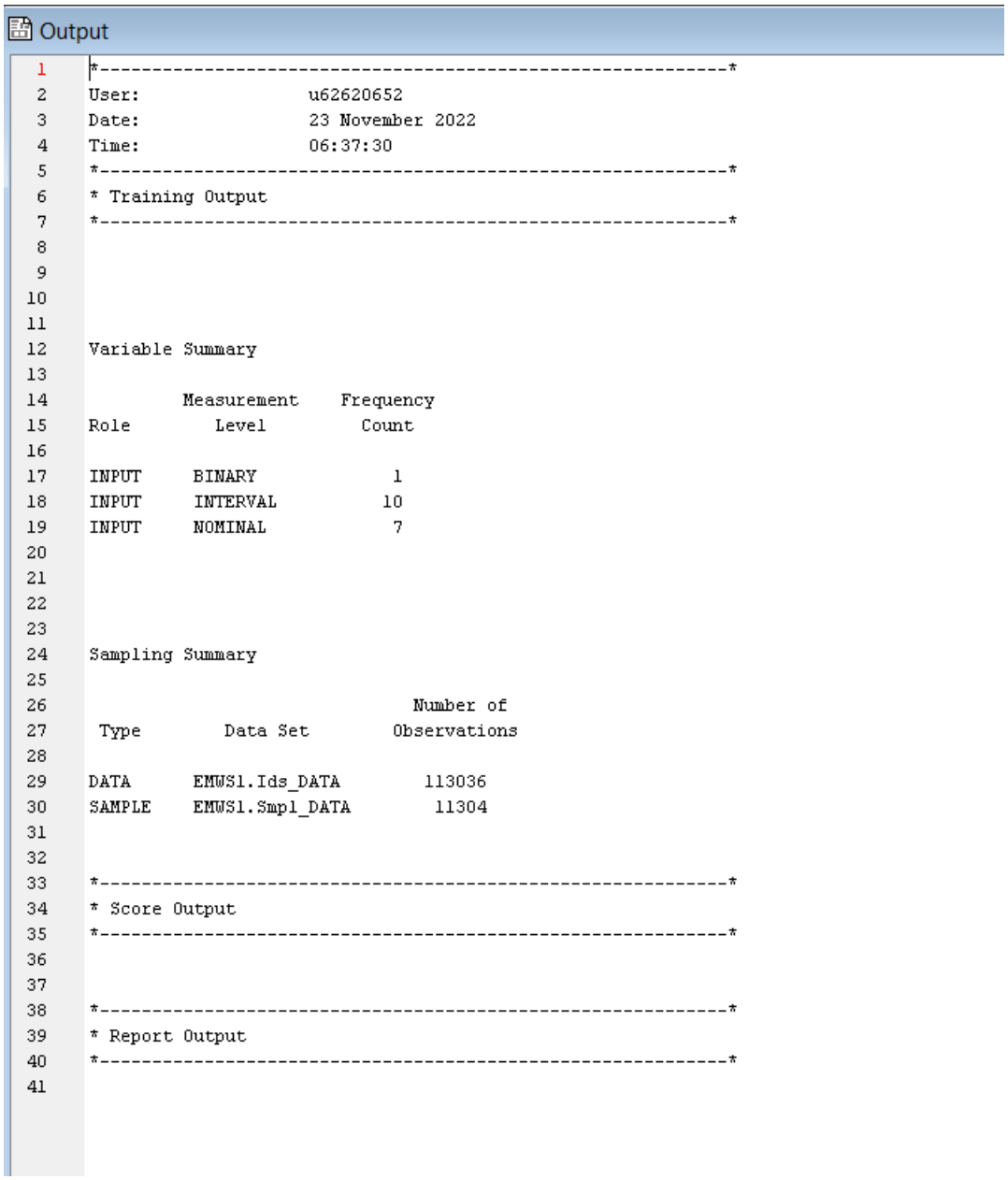
Ordered Inputs	Data Role	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Abs C.V.	Coefficient of Variation	Sign
1	TRAIN	Unit Cost	9	0	100000	1	2171	261.7311	550.6513	2.19276	3.718817	INPUT	Unit Cost	2.103882	2.103882	+
2	TRAIN	Unit Price	25	0	100000	2	3578	446.4383	928.5537	2.154752	3.430858	INPUT	Unit Price	2.079915	2.079915	+
3	TRAIN	Cost	117	0	100000	1	9684	471.8858	880.5704	3.56178	16.90151	INPUT	Cost	1.866067	1.866067	+
4	TRAIN	Revenue	235	0	100000	2	14312	760.3229	1309.344	3.539896	17.00106	INPUT	Revenue	1.722089	1.722089	+
5	TRAIN	Profit	105	0	100000	-30	5638	288.4372	458.923	3.573653	18.71215	INPUT	Profit	1.591067	1.591067	+
6	TRAIN	Order Q...	11	0	100000	1	32	11.99312	9.536848	0.36556	-1.23649	INPUT	Order Q...	0.795193	0.795193	+
7	TRAIN	Day	16	0	100000	1	31	15.6473	8.786075	0.015092	-1.19206	INPUT	Day	0.561507	0.561507	+
8	TRAIN	Custome...	35	0	100000	17	87	35.9497	10.96606	0.520416	-0.11315	INPUT	Custome...	0.305039	0.305039	+
9	TRAIN	Date	19898	0	100000	18628	20666	20042.32	449.8874	-0.47382	-0.36199	INPUT	Date	0.022447	0.022447	+
10	TRAIN	Year	2014	0	100000	2011	2016	2014.381	1.290029	-0.38886	3.500483	INPUT	Year	.0006404	.0006404	+

4. Steps for Variable Clustering, Variable Correlation and Interesting Visualizations

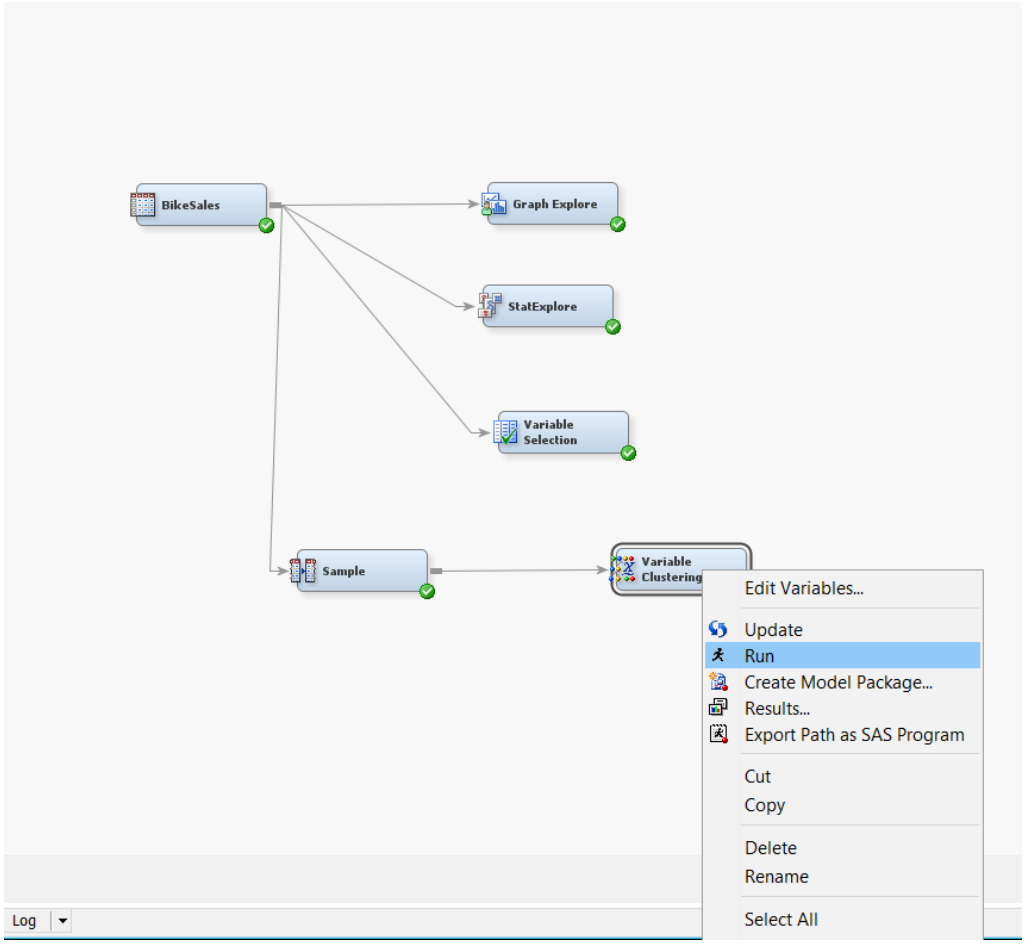
1. Drag Sample node to the project workspace and connect to data source. Run the sample node.



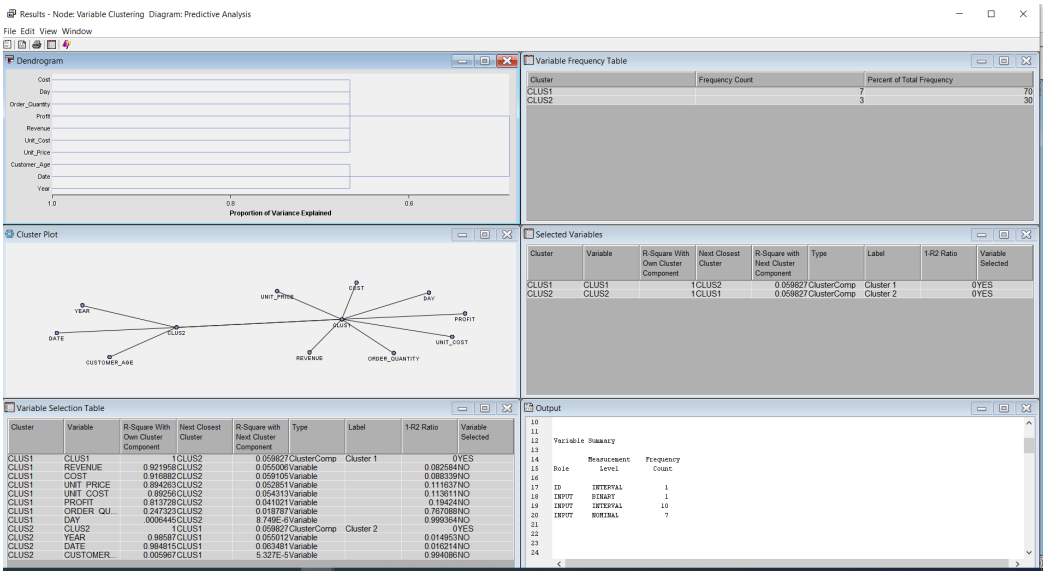
2. Sample results will be shown.



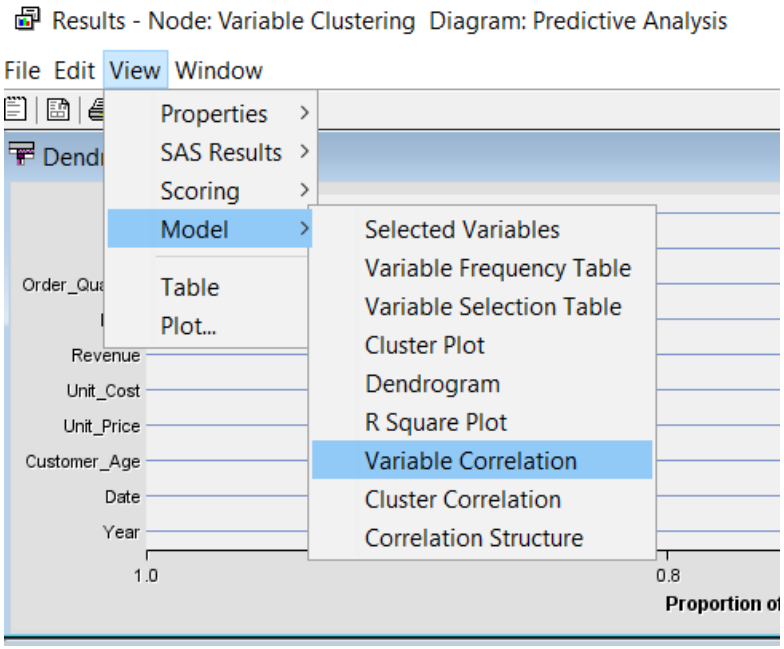
3. Drag Variable Clustering node to the project workspace and connect it to Sample node. Run Variable Clustering node.

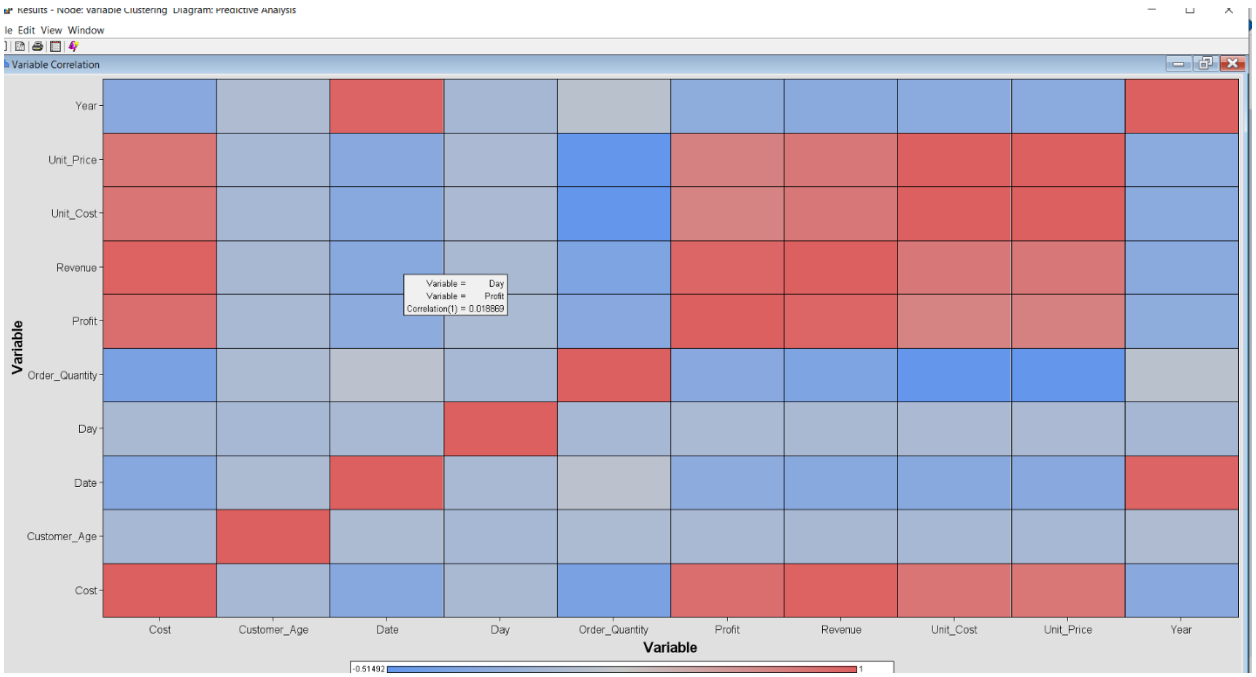


4. Variable Clustering results are shown.

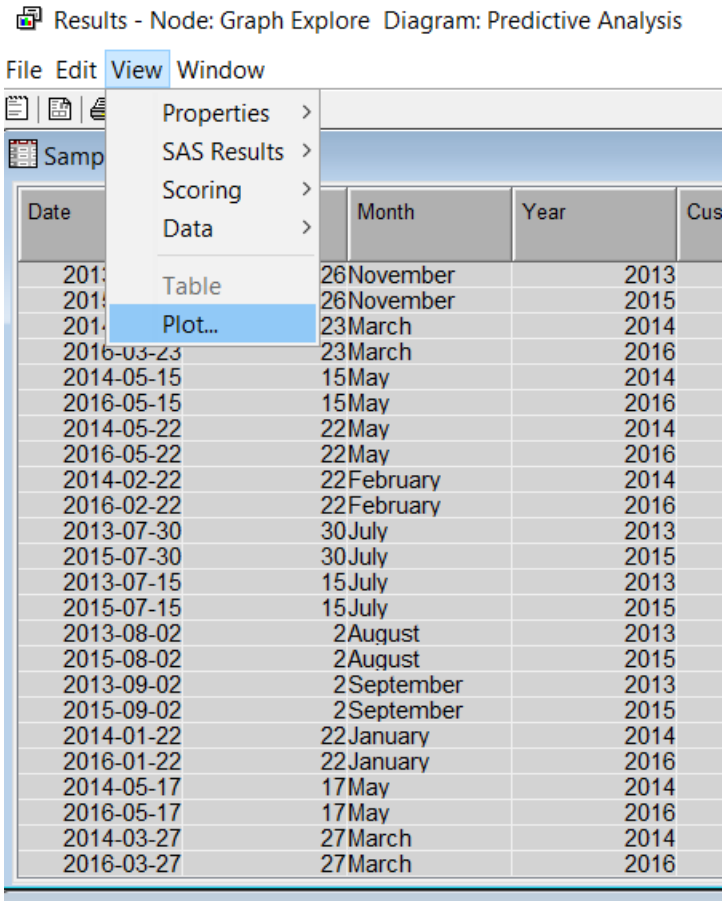


5. Run Variable Clustering node and show the results. Select Variable Correlation from Model option of View Menu.

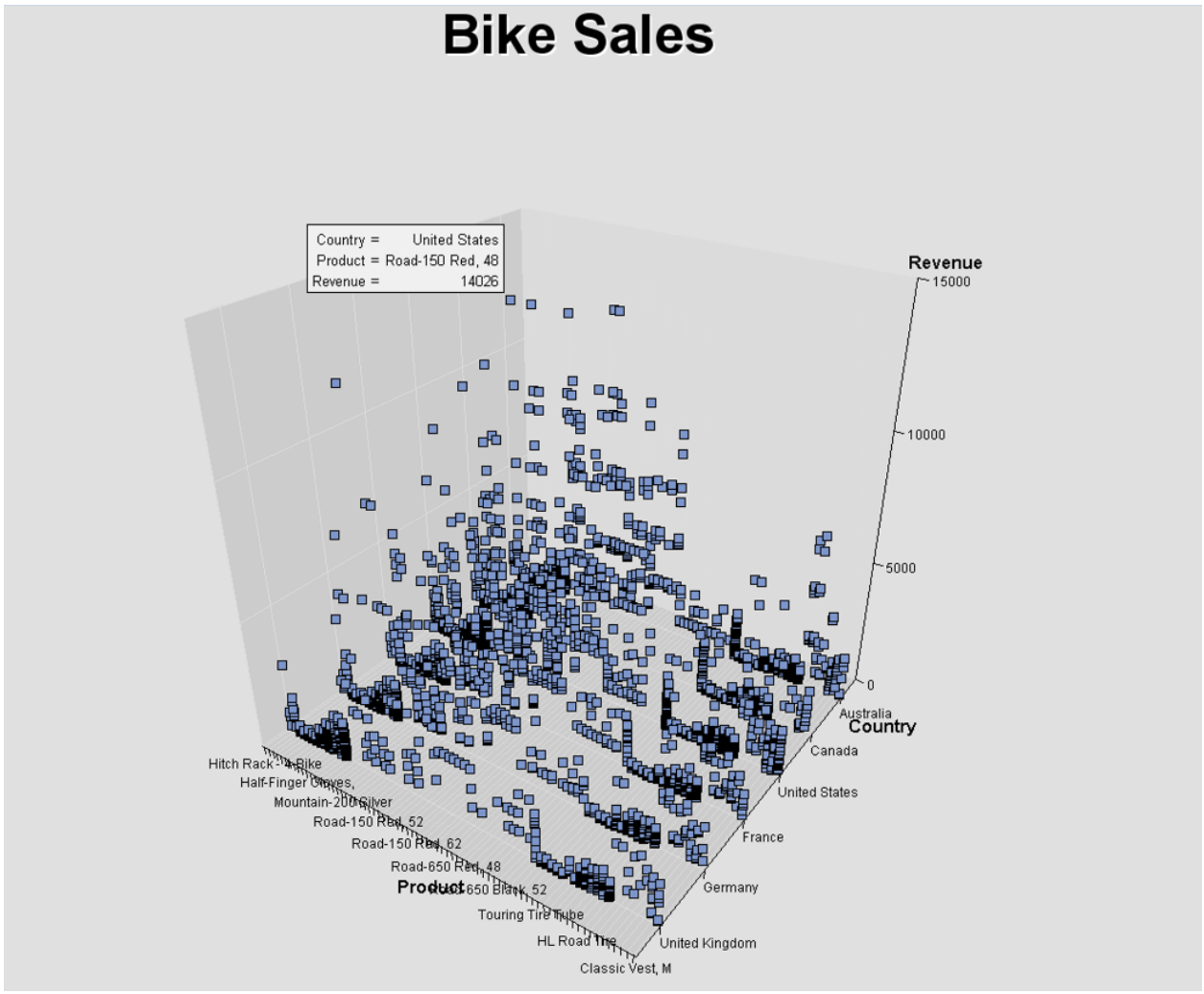




6. Run Graph Explore node, show the results and select plot option of View Menu. Select 3D Visualization options.

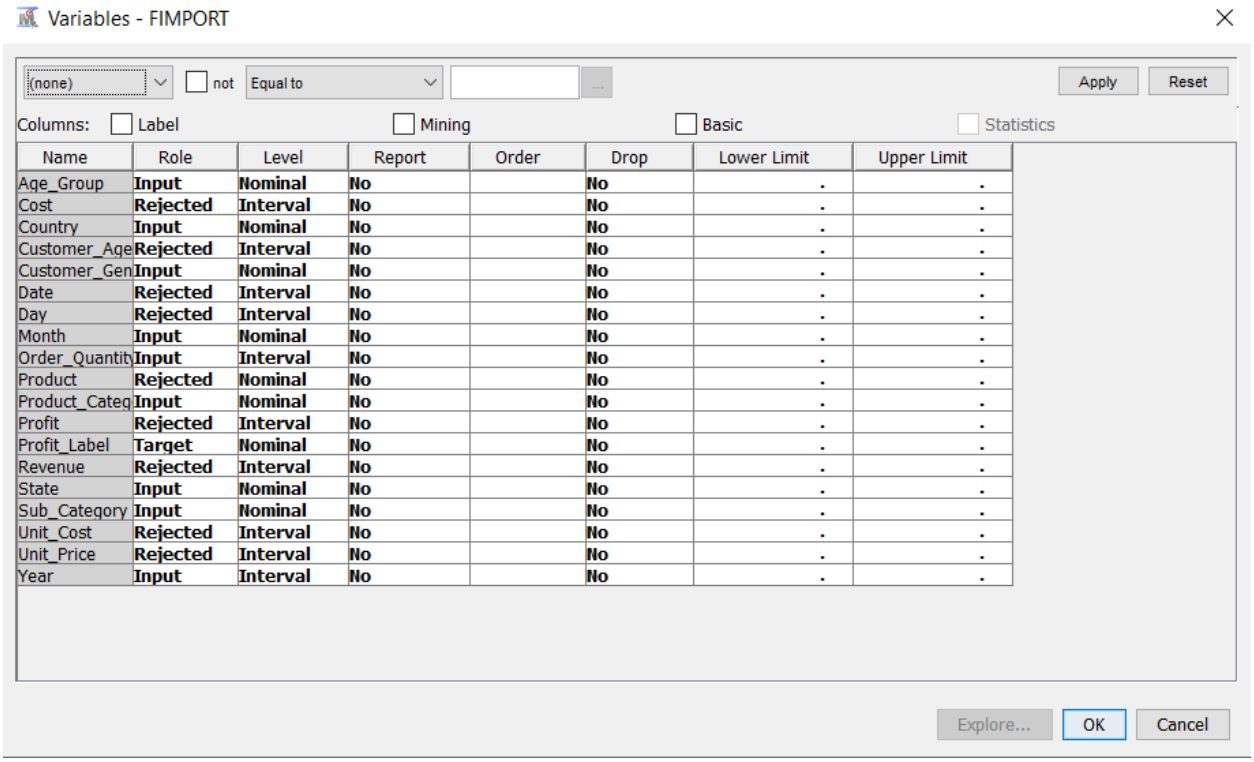


Bike Sales


Country = United States
Product = Road-150 Red, 48
Revenue = 14026
Revenue
15000
10000
5000
0
Australia
Canada
United States
France
Germany
United Kingdom
Country
Hitch Rack - 4-Bike
Half-Finger Gloves,
Mountain-200 Silver
Road-150 Red, 52
Road-150 Red, 62
Road-650 Red, 48
Road-650 Black, 52
Touring Tire Tube
HL Road Tire
Classic Vest, M
Product

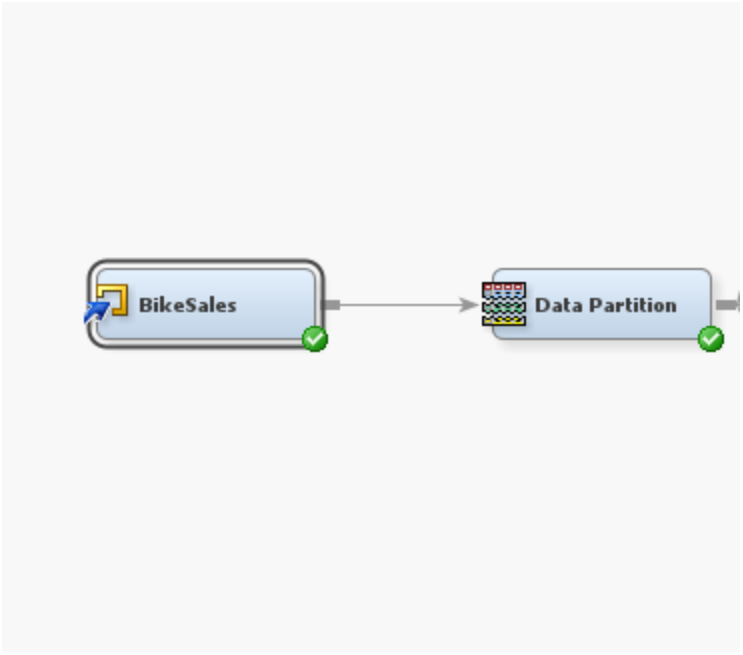
Modify, Model and Assess

1. Right click on data source node and click Edit Variables to modify the role of variables.

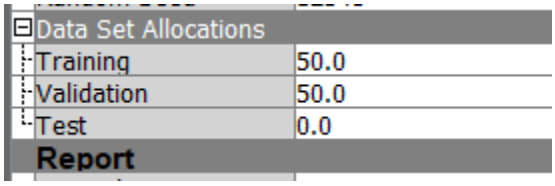


Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age_Group	Input	Nominal	No		No	.	.
Cost	Rejected	Interval	No		No	.	.
Country	Input	Nominal	No		No	.	.
Customer_Age	Rejected	Interval	No		No	.	.
Customer_Gen	Input	Nominal	No		No	.	.
Date	Rejected	Interval	No		No	.	.
Day	Rejected	Interval	No		No	.	.
Month	Input	Nominal	No		No	.	.
Order_Quantity	Input	Interval	No		No	.	.
Product	Rejected	Nominal	No		No	.	.
Product_Categ	Input	Nominal	No		No	.	.
Profit	Rejected	Interval	No		No	.	.
Profit_Label	Target	Nominal	No		No	.	.
Revenue	Rejected	Interval	No		No	.	.
State	Input	Nominal	No		No	.	.
Sub_Category	Input	Nominal	No		No	.	.
Unit_Cost	Rejected	Interval	No		No	.	.
Unit_Price	Rejected	Interval	No		No	.	.
Year	Input	Interval	No		No	.	.

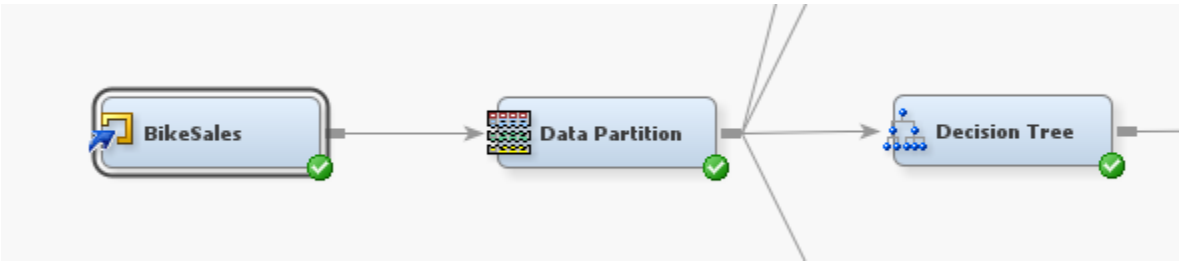
2. Click Sample tab and drag Data Partition node in the diagram. Connect it to the data source.



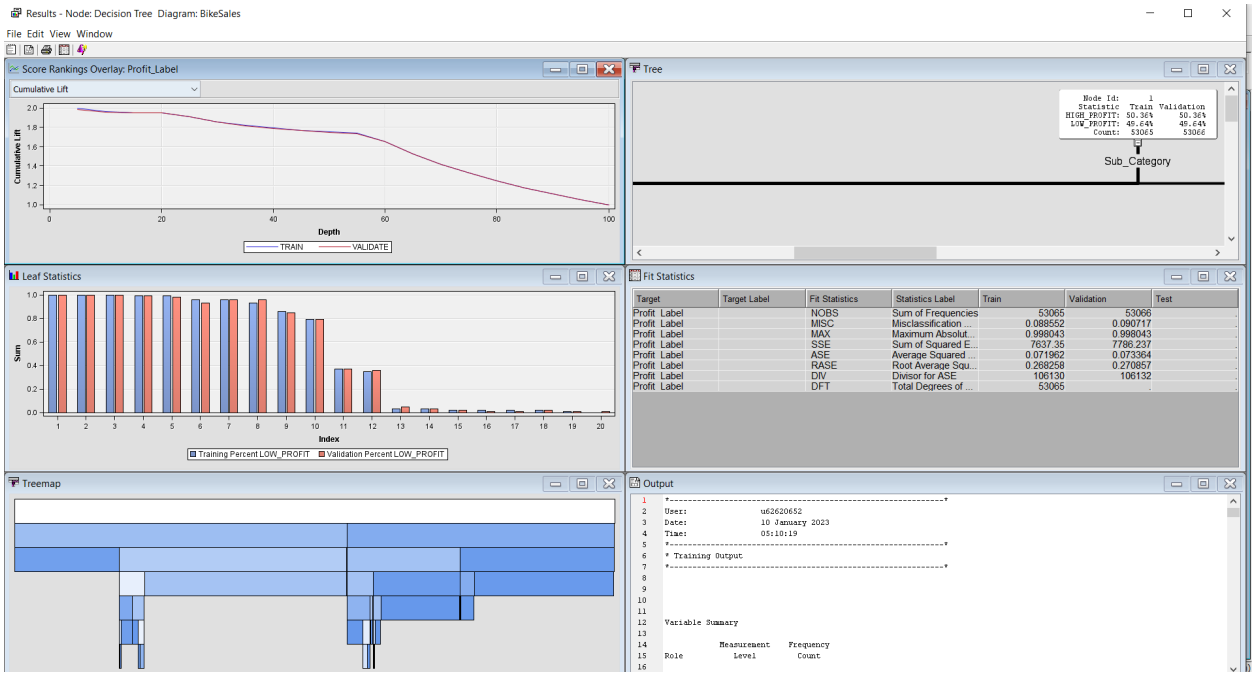
3. Set training and validation data to 50% each and run it to partition the data.



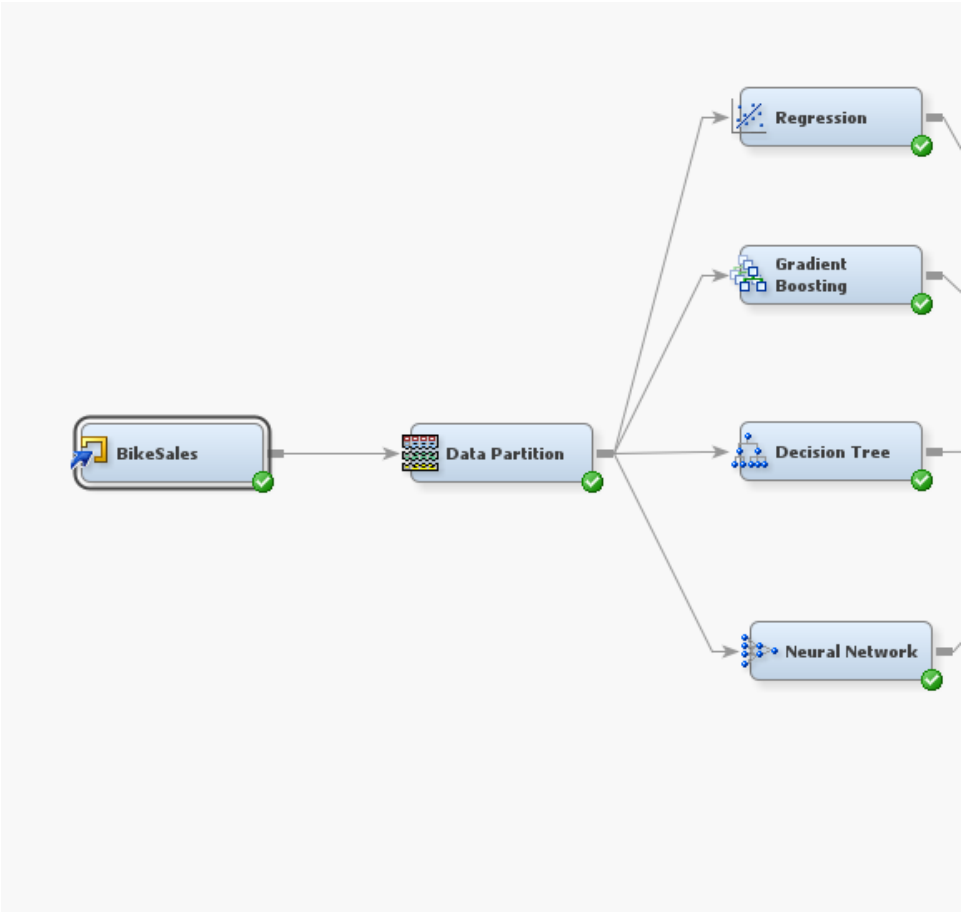
4. Click Model tab and drag Decision Tree node in the diagram. Connect it to the Data Partition node.



5. Set properties of decision tree node, maximum branch to 2 and maximum depth to 6 and run it to get results of decision tree model. The results of decision tree model are as below.



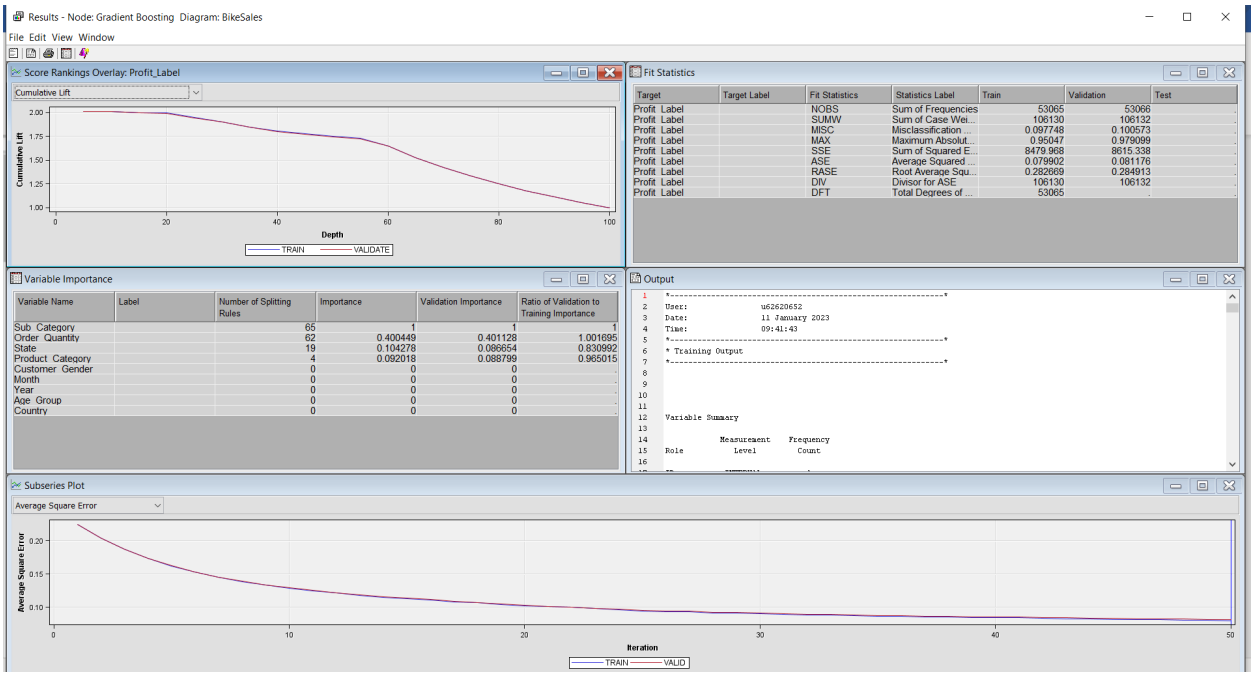
6. Drag Gradient Boosting, Logistic Regression and Neural Network nodes from the model tab to the diagram. Run these nodes and get result of each model.



7. The property values for Gradient Boosting are as below.

.. Property	Value
Series Options	
N Iterations	50
Seed	12345
Shrinkage	0.1
Train Proportion	60
Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	2
Minimum Categorical S	5
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Use in search
Performance	Disk
Node	
Leaf Fraction	0.001
Number of Surrogate R	0
Split Size	.
Split Search	

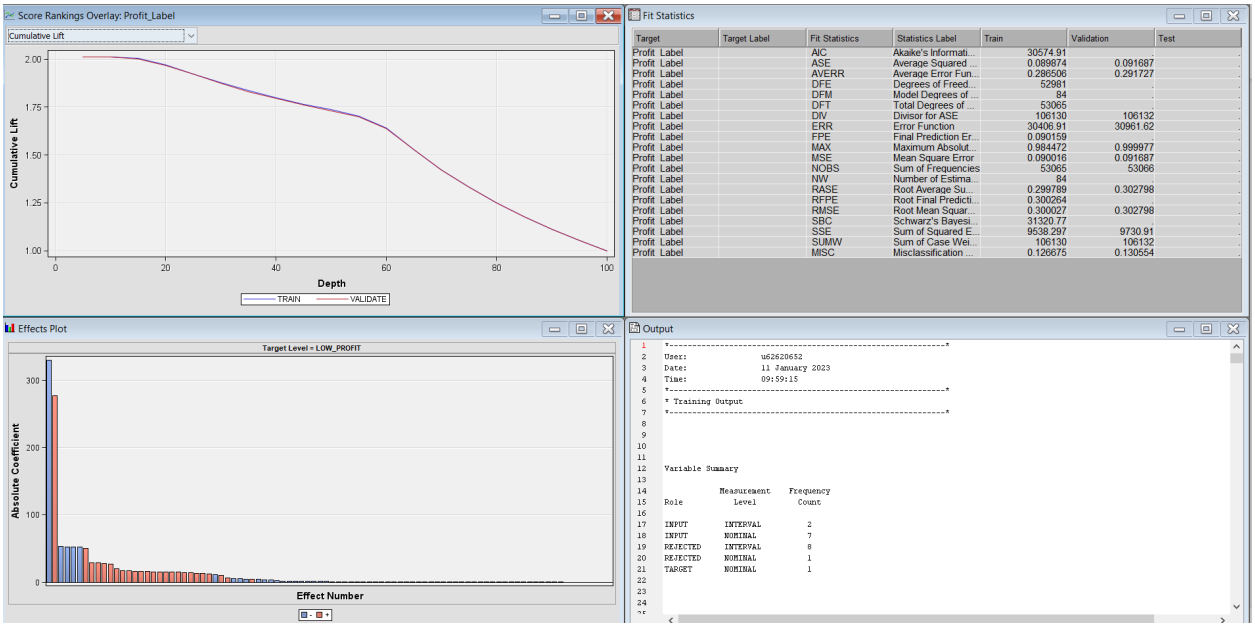
8. The results of Gradient Boosting are as below.



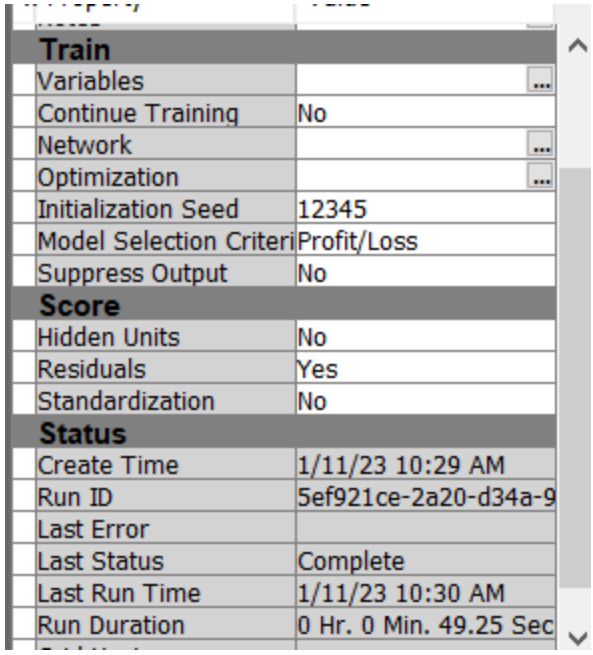
9. The property values for Logistic Regression are as below.

Property	Value
Equation	
Main Effects	Yes
Two-Factor Interaction	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	None
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	

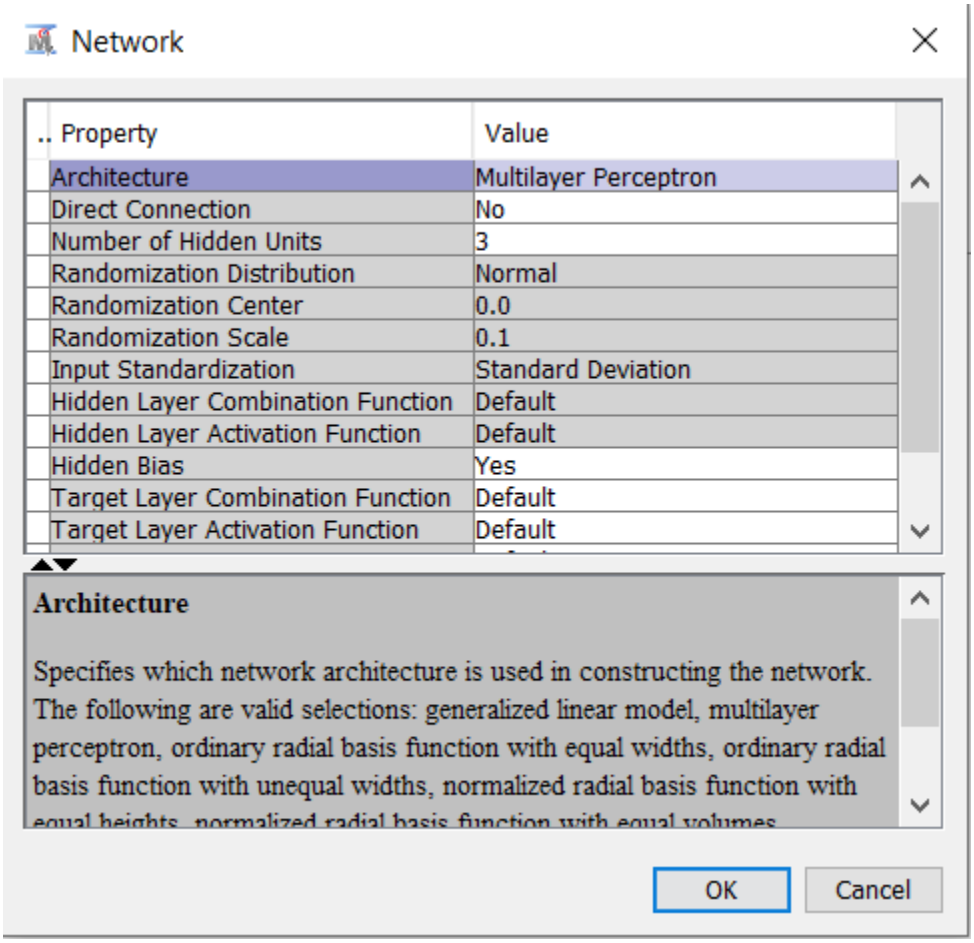
10. The results of Logistic Regression are as below.



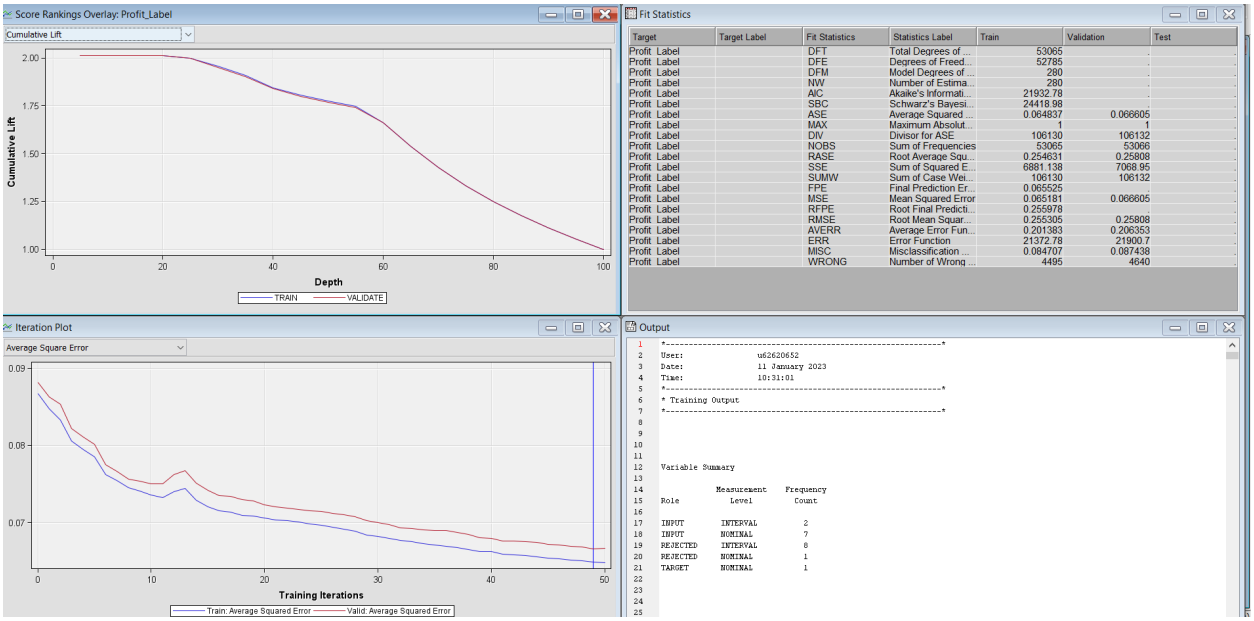
11. The property value for Neural Network is as below.



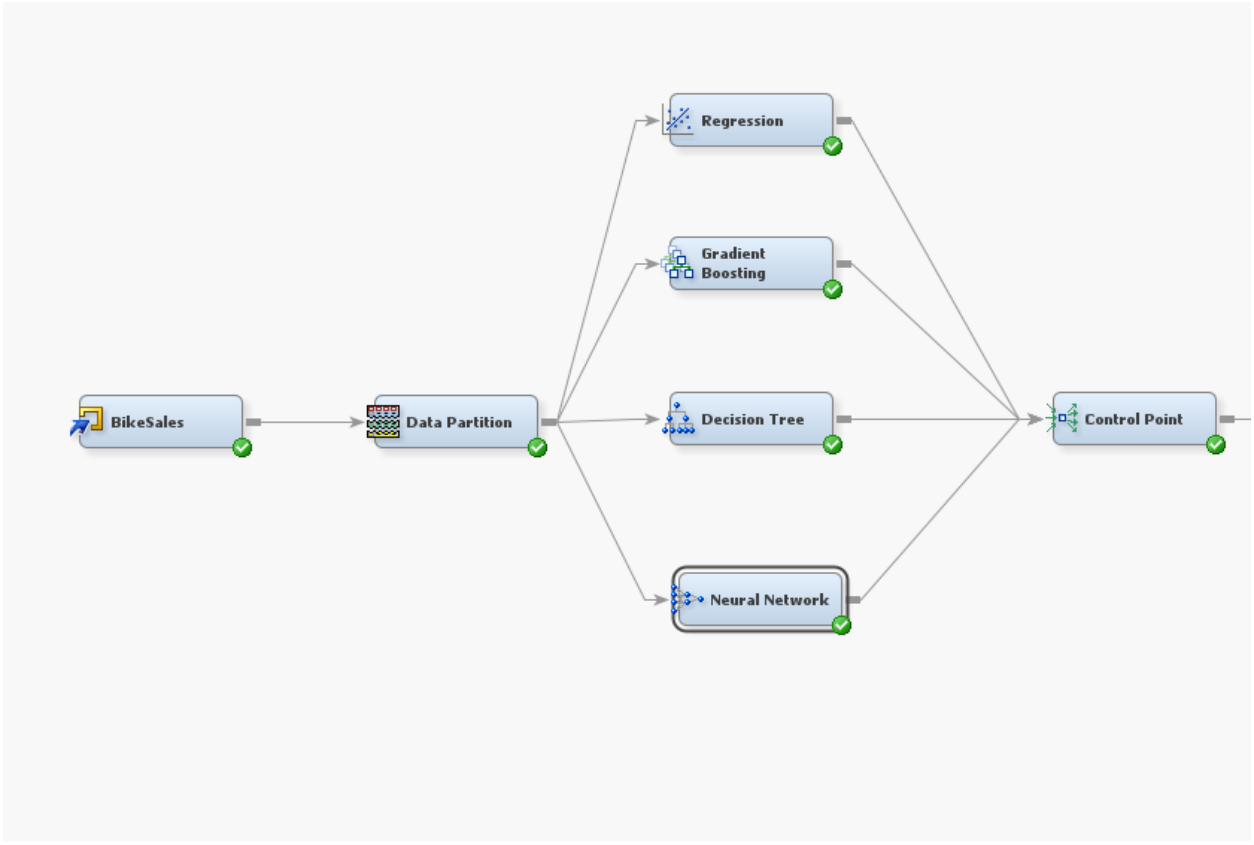
Train	
Variables	...
Continue Training	No
Network	...
Optimization	...
Initialization Seed	12345
Model Selection Criteri	Profit/Loss
Suppress Output	No
Score	
Hidden Units	No
Residuals	Yes
Standardization	No
Status	
Create Time	1/11/23 10:29 AM
Run ID	5ef921ce-2a20-d34a-9
Last Error	
Last Status	Complete
Last Run Time	1/11/23 10:30 AM
Run Duration	0 Hr. 0 Min. 49.25 Sec



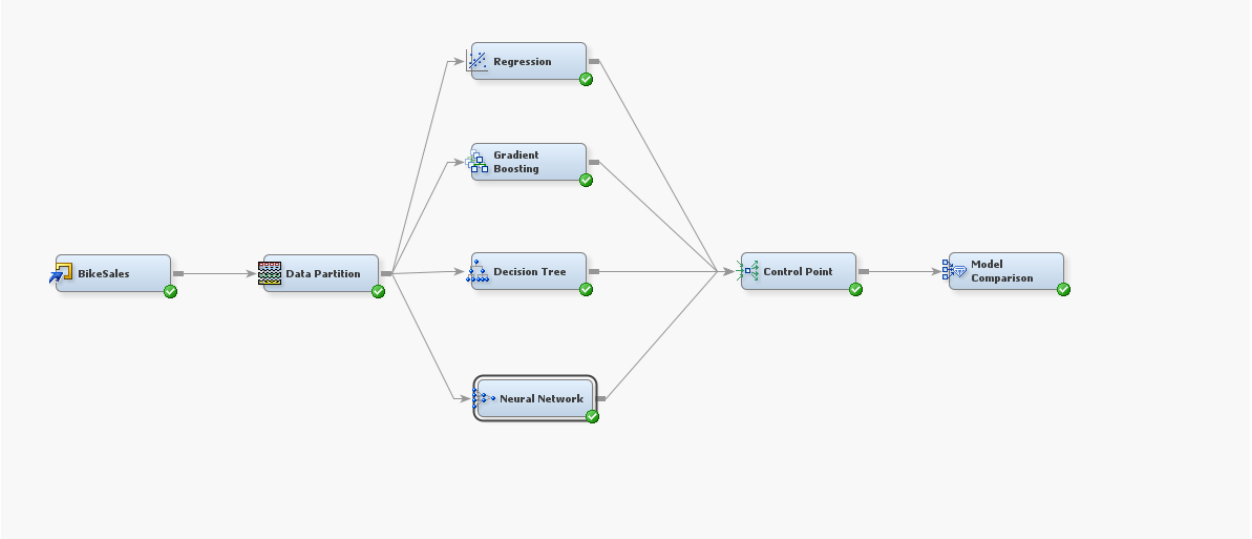
12. The results of neural network are as below.



13. Drag Control Point node from utility tab to the diagram. Connect all the models to the control point node.



14. Drag Model Comparison node from Assess tab to the diagram and connect control point node to the model comparison node.



15. Run the model comparison node to compare all the models and the results are as below.