The University of Texas at Dallas

School of Management

BUAN 6324 Business Analytics with SAS

Instructor: Syam Menon

Term Paper: Final report

Team	Name	Net ID
Group 7	Kavinayaa Mahesh	KXM220108
	Pradhiksha Suresh	PXS230024
	Srimahima Suresh	SXS220571
	Rohith Kumar Anandamohan	RXA220097
	Vishnu Ajay	VXA230006

We have performed our end-to-end analysis for our final report on SAS Enterprise miner. The learnings at each stage of the project have been recorded in this report.

Data Description: US Supermarket Data

We have selected a dataset from Kaggle.com called the “US Supermarket Data”

This dataset contains information about 50 branches of the supermarket XYZ throughout US. Supermarket XYZ's declining revenues since 2016 signalling a critical need for the organization to harness its vast database through data analytics.

The “US Supermarket Data” contains 4 datasets each of which can be used to derive different business insights.

1. **50_Supermarket_Branches.csv** contains the information of 50 supermarket branches. It contains information regarding the amount spend by 50 branches on Advertisement, Promotions & Administration. It also has the branch State information & the profit made.
2. **Ads_CTR_Optimisation.csv** is based on the Click-Through Rates 10K users in 10 different ads. It is a binary file having 0’s & 1’s representing the ad-click information of 10 different ads in 10000 instances.
3. **Market_Basket_Optimisation.csv**. This dataset contains 7500 sales transactions in a week.
4. **Supermarket_CustomerMembers.csv**. This dataset contains the profile information of 200 customers. It has information regarding the customer’s Gender, Age, Annual Income in thousands and Spending Score on a scale of 100.

Project Motivation and Background

Our focus is on developing an integrated data analytics approach that will cater for the unique business requirements of our organization. Because of the recent downward earnings, there is a pressing need to employ progressive ways that can help us reverse this situation and ensure that the company attains sustainable growth. Through using advanced data analysis techniques strategically, we aim at stopping revenue decline and improve operational efficiency.

Our goal is also to identify areas which are open for improvement and introduce effective solutions based on data trends. With this, not only will we be able to cut costs but increase productivity; furthermore, it places us where we can react quickly and efficiently to significant market volatility.

In addition, our data analytics initiative presents an incredible opportunity to improve the way we identify and retain customers. In this regard, by using the intelligence that can be derived from comprehensive data analysis, we can have a better feel for how customer(s) behave, what they prefer or hate most. Such knowledge will enable us to effectively customize marketing campaigns and personalizing customer experiences based on their needs and desires. As well as that, early detection of customers who are likely to leave allows the company to introduce initiatives aimed at improving loyalty among customers.

To put it simply, our data analytics strategy is more than just a tactical response to current challenges; it represents a long-term commitment in which we treat data as central in our decision-making process within our business setup. Our goal is therefore not only to use complete potential of data analytics to reverse declining revenue but also ensure the organization maintains its competitiveness in an increasingly crowded market.

Analysis Performed

We have applied two major techniques that we have applied through the term of this course.

1. **Market Basket Analysis:** Applied Association Rule Mining on **Market_Basket_Optimisation.csv** dataset to discover patterns and associations among products purchased together. This is performed to identify bundling opportunities to boost sales.
2. **Customer Segmentation:** Applied clustering techniques on the **Supermarket_CustomerMembers.csv** dataset creating customer segments based

on the customer purchase behaviours. This is performed to take data driven targeted marketing campaigns for different customer groups.

3. **Decision Tree Analysis:** Applied Decision Tree technique on the **Supermarket_CustomerMembers.csv** dataset for analysing customer purchase behaviours pattern.
4. **Regression Analysis:** Applied Linear Regression on the **50_Supermarket_Branches.csv** dataset to get insights into profit drivers.
5. **Exploratory Data Analysis:** Applied the capabilities of graphical interpretation on **Supermarket_CustomerMembers.csv, 50_Supermarket_Branches.csv and Ads_CTR_Optimisation.csv** datasets to extract insights and see any correlation amongst customer behaviour, spending patterns & profits.

Executive Summary

The comprehensive analysis conducted using SAS Enterprise Miner has provided XYZ Supermarket with actionable insights that can significantly influence both strategic decision-making and day-to-day operations.

The Analysis performed offer a multifaceted view of the supermarket's operations, customer behaviours, and market dynamics. Here are the key managerial implications drawn from the insights across these analyses:

1. **Enhancing Marketing Strategies:**

   Targeted Promotions and Product Bundling: The association rule mining revealed patterns that suggest potential product bundling which could appeal to customers' purchasing habits.

   Refined Customer Targeting: Customer segmentation has identified distinct groups such as high spenders and income-driven shoppers. Tailored marketing campaigns designed to meet the unique needs and preferences of these segments can enhance customer engagement and retention.

2. **Cost Management and Operational Efficiency:**

   Streamlined Operations: The analysis indicates branches where operational performance can be improved, especially in management and promotion spending. Streamlining those expenses can reduce overheads at the same time while preserving effectiveness.

   Advertising Spend Allocation: Linear regression analysis pointed out that advertising spend has a strong correlation with profitability. Management should consider increasing the budget allocation to advertising, focusing on high ROI channels and techniques.

3. **Predictive Analytics for Future Trends:**

   Utilizing Decision Trees for Predictive Insights: Decision tree analysis has provided a method to predict customer spending behaviour based on demographic data. These predictive insights can be used to forecast future trends and prepare strategies that align with anticipated market developments.

   Dynamic Adaptation to Market Changes: Continuous monitoring of key performance indicators as suggested by the regression model results can help the supermarket quickly adapt to changes in consumer behaviour and market conditions.

4. **Enhancing Customer Experience:**

   Using the detailed customer profiles generated from segmentation analysis, XYZ Supermarket can offer personalized shopping experiences, rewards, and loyalty programs that are aligned with individual customer preferences and purchasing history.

5. **Community Engagement and Localized Offers:**

   By understanding the geographic distribution of branches and local customer demographics, supermarkets can engage with local communities through tailored offers, events, and community-driven promotions.

The analyses performed provide XYZ Supermarket with a roadmap for data-driven decision-making that not only aims to rectify the current revenue declines but also positions the company for future growth and sustainability in a competitive market. By implementing the strategies based on these managerial implications, XYZ Supermarket can enhance operational efficiency, customer satisfaction, and ultimately, profitability.

The following sections dives deeper into each of the analysis performed by us detailing the steps done in SAS Enterprise Miner, providing an end-to-end project understanding.

Analysis 1: Association Rule Mining

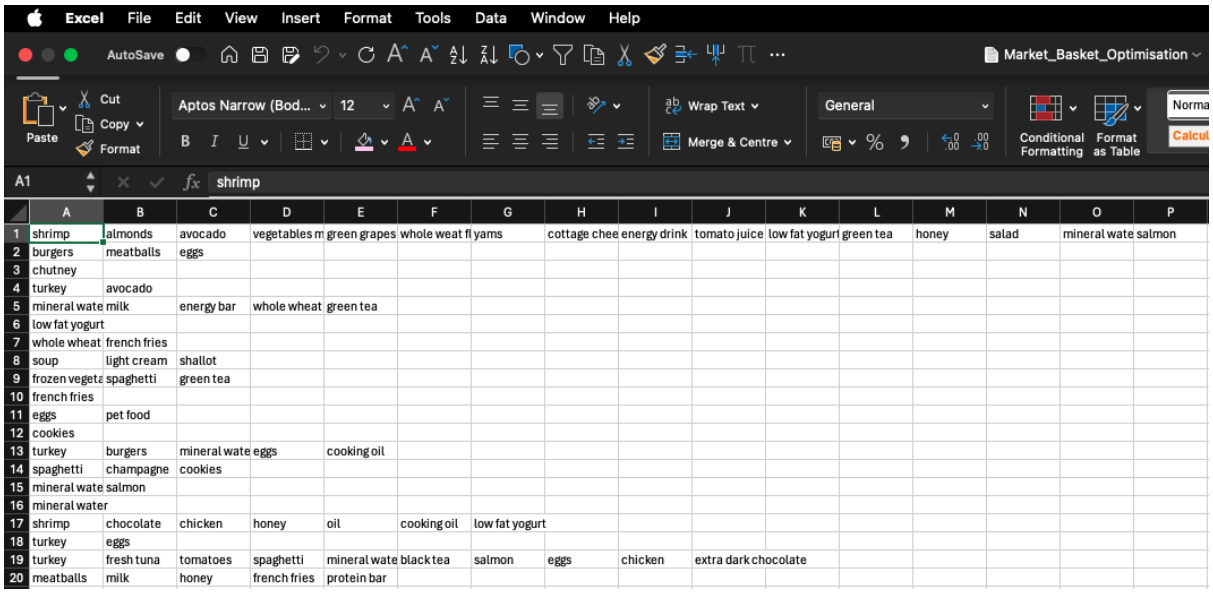
Dataset: Utilized the Market_Basket_Optimisation.csv dataset for performing the Association Rule Mining.

Procedure

To perform the analysis, we had to perform a few data pre-processing procedures to format the data in such a way that SAS enterprise can consume it to perform Association Rule Mining. Followed by this we have setup a process diagram to run the analysis.

Step 1: Exploring the original dataset

Below is the screenshot of a sample of the original data. The data is in .csv format, when visualised in excel, each row represents a transaction & each column of a row holds a product name that belongs to the transaction.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	shrimp	almonds	avocado	vegetables m	green grapes	whole weat fl	yams	cottage chee	energy drink	tomato juice	low fat yogurt	green tea	honey	salad	mineral wate	salmon
2	burgers	meatballs	eggs													
3	chutney															
4	turkey	avocado														
5	mineral wate	milk	energy bar	whole wheat	green tea											
6	low fat yogurt															
7	whole wheat	french fries														
8	soup	light cream	shallot													
9	frozen vegeta	spaghetti	green tea													
10	french fries															
11	eggs	pet food														
12	cookies															
13	turkey	burgers	mineral wate	eggs	cooking oil											
14	spaghetti	champagne	cookies													
15	mineral wate	salmon														
16	mineral water															
17	shrimp	chocolate	chicken	honey	oil	cooking oil	low fat yogurt									
18	turkey	eggs														
19	turkey	fresh tuna	tomatoes	spaghetti	mineral wate	black tea	salmon	eggs	chicken	extra dark chocolate						
20	meatballs	milk	honey	french fries	protein bar											

Figure 1

Step 2: Data pre-processing using python

For performing ARM in SAS miner, we need a ID column & a Target Column. Here ID should correspond to a transaction ID & Target should correspond to the Product.

In order to create this 2 column structured data table, we created a python code segment that can automate this data pre-processing procedure. The code was executed using Google Colab. Google Colab is a free Jupyter notebook environment that runs on Google's cloud servers. It allows us to write and execute Python code in a browser environment.

Below is the screenshot of the activity performed in Google Colab for the data transformation

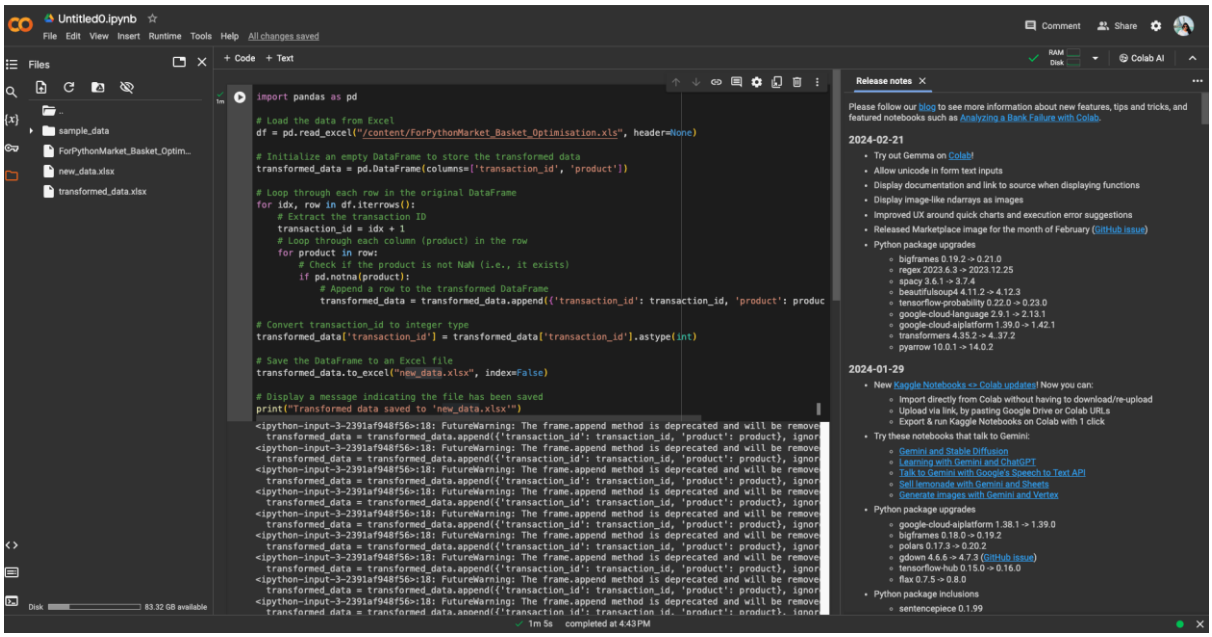
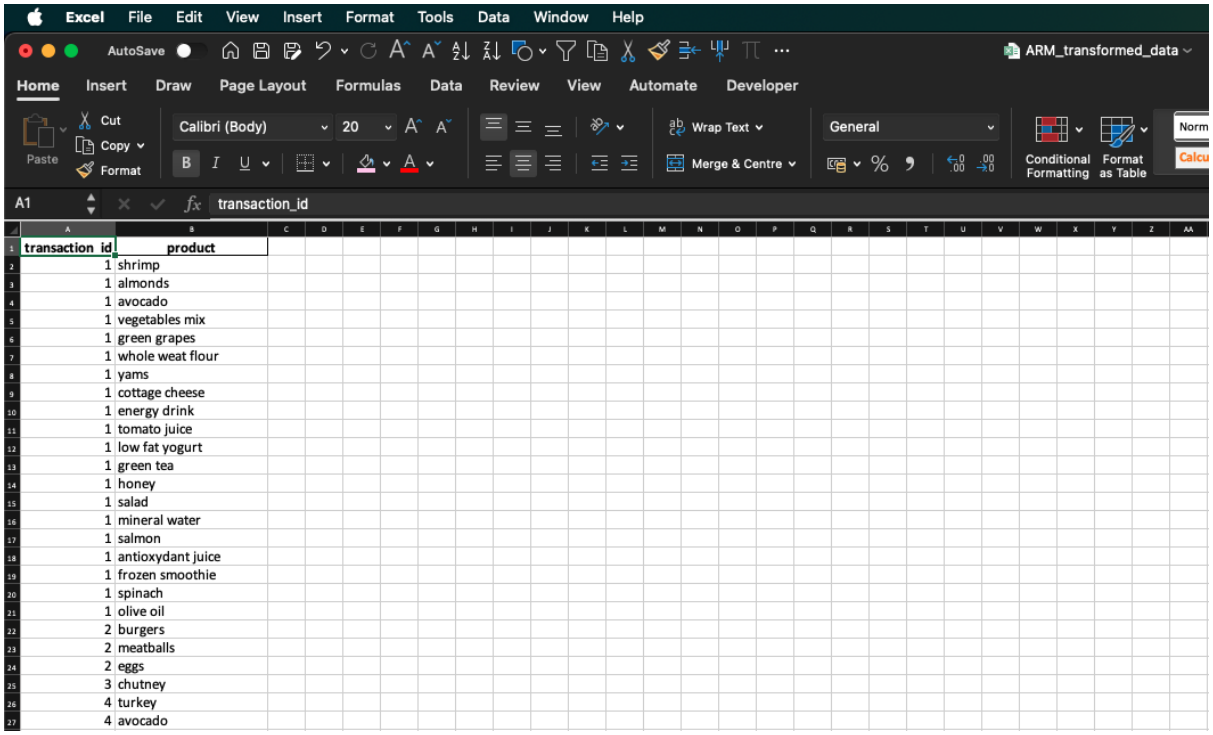


Figure 2

Once this transformation was completed, we exported the output into a .xlsx file. This file is in the desired format with 2 columns transaction ID & product corresponding to ID & Target respectively.

Below is the screenshot of a sample of the transformed data.



transaction_id	product
1	shrimp
1	almonds
1	avocado
1	vegetables mix
1	green grapes
1	whole weat flour
1	yams
1	cottage cheese
1	energy drink
1	tomato juice
1	low fat yogurt
1	green tea
1	honey
1	salad
1	mineral water
1	salmon
1	antioxydant juice
1	frozen smoothie
1	spinach
1	olive oil
2	burgers
2	meatballs
2	eggs
3	chutney
4	turkey
4	avocado

Figure 3

Step 3: File import

A new project is set for this step named as TermpaperARM. Once this is done the library is setup & a diagram named “Termpaper_ARM” is created.

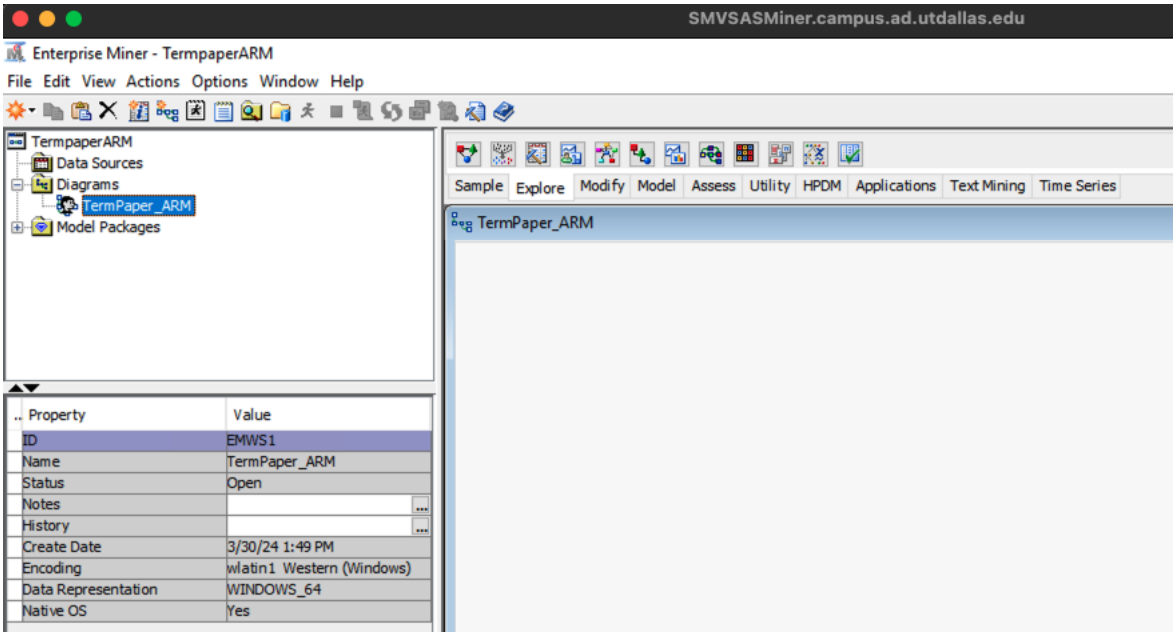


Figure 4

To import an excel file into SAS Enterprise miner we performed the below steps,

a. Click Sample tab and drag the File import node into the diagram.

b. Go to the Properties box of the File Import node and click on the ellipsis against the Import File option under Train.

c. Select My Computer and enter the complete path to the location where the file ARM_transformed_data.xlsx is saved, including the file name.

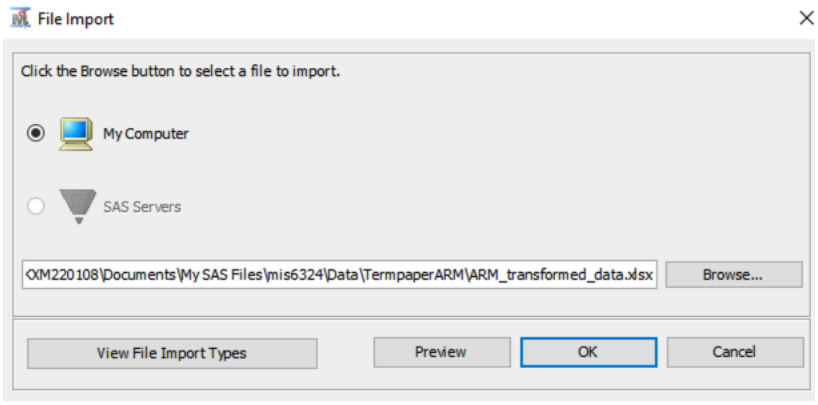


Figure 5

d. Right-click the File Import node and select Edit Variables. Set product to be the target and transaction_id to be the ID.

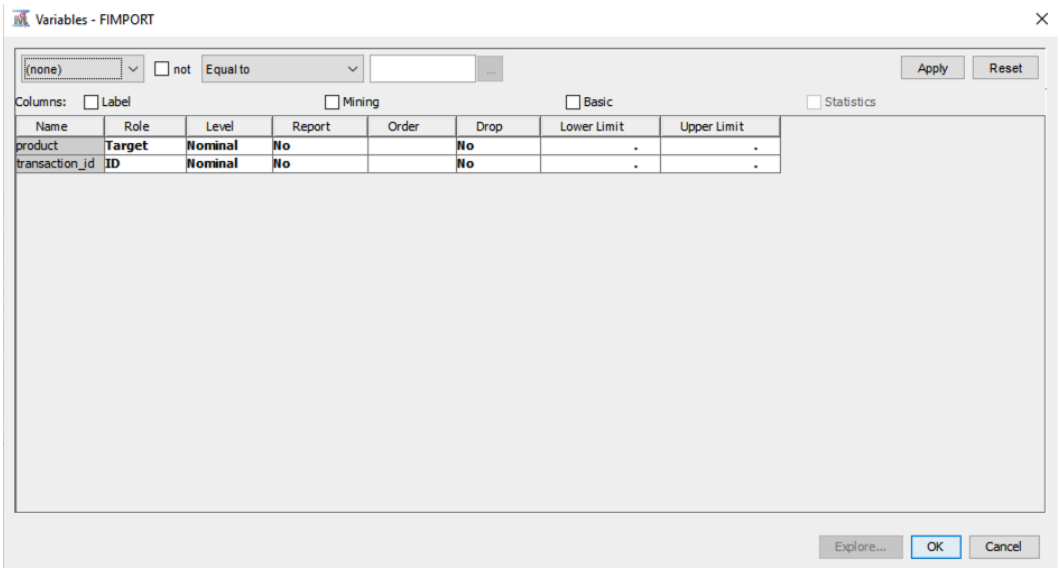


Figure 6

e. Go to the Properties box of the File Import node and set the Role as Transaction under Score, since this is a transaction data.

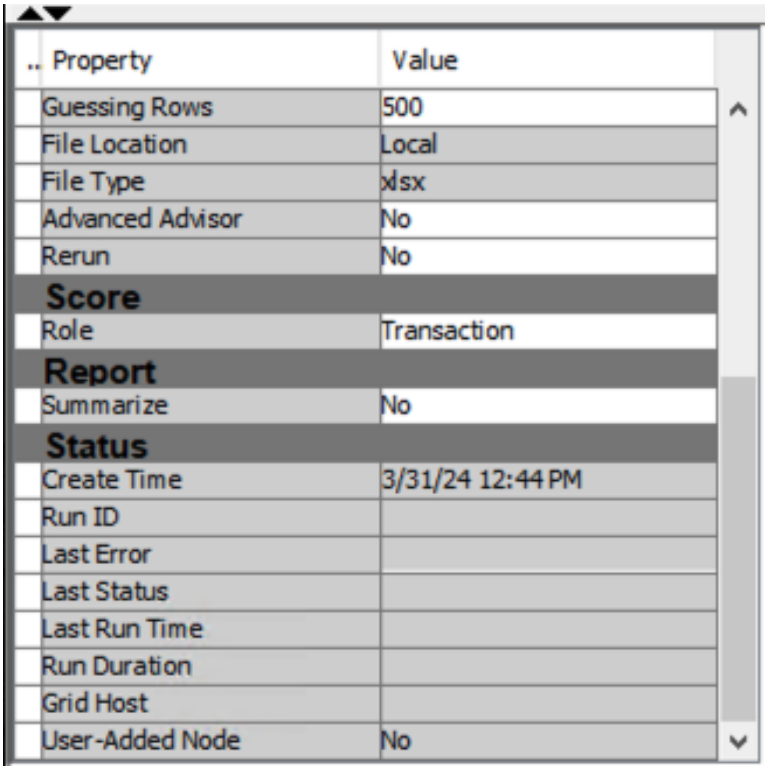


Figure 7

f. At this step, the File import is complete.

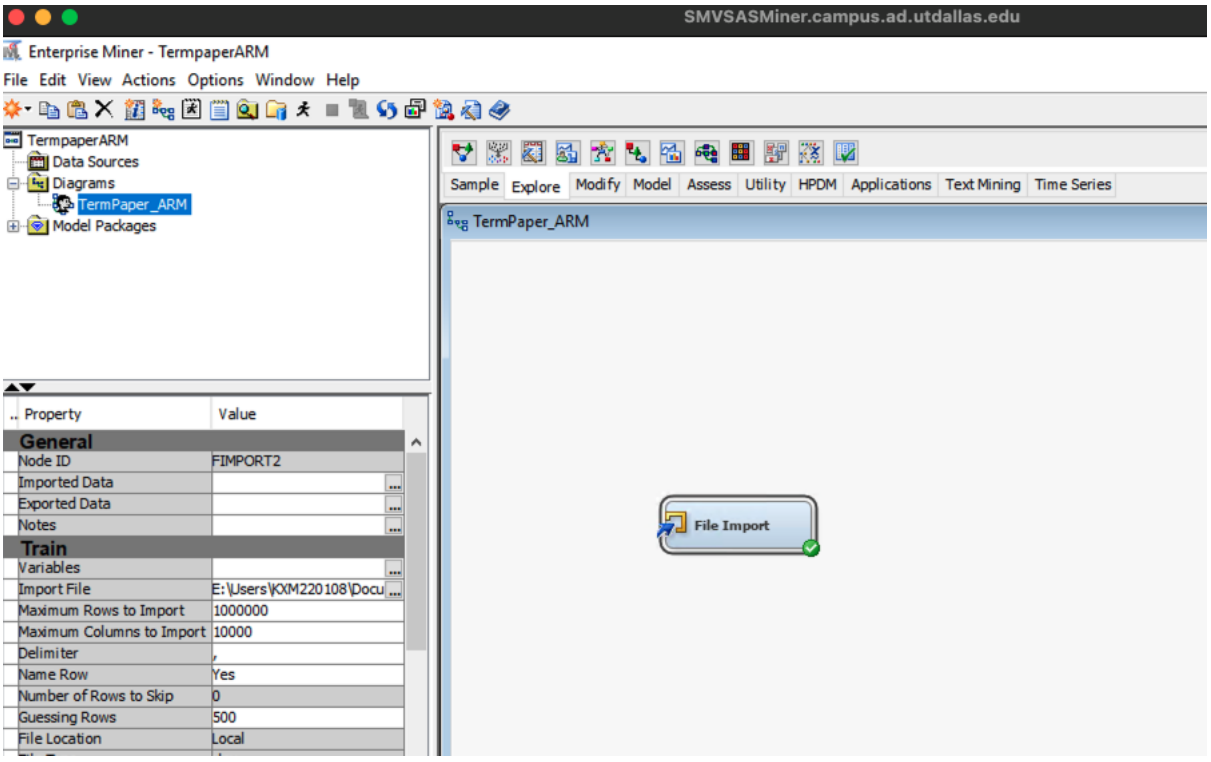


Figure 8

Step 4: Add the Association node

a. Click the Explore tab and drag the Association node to the diagram

b. Connect the File Import node & Association node to each other.

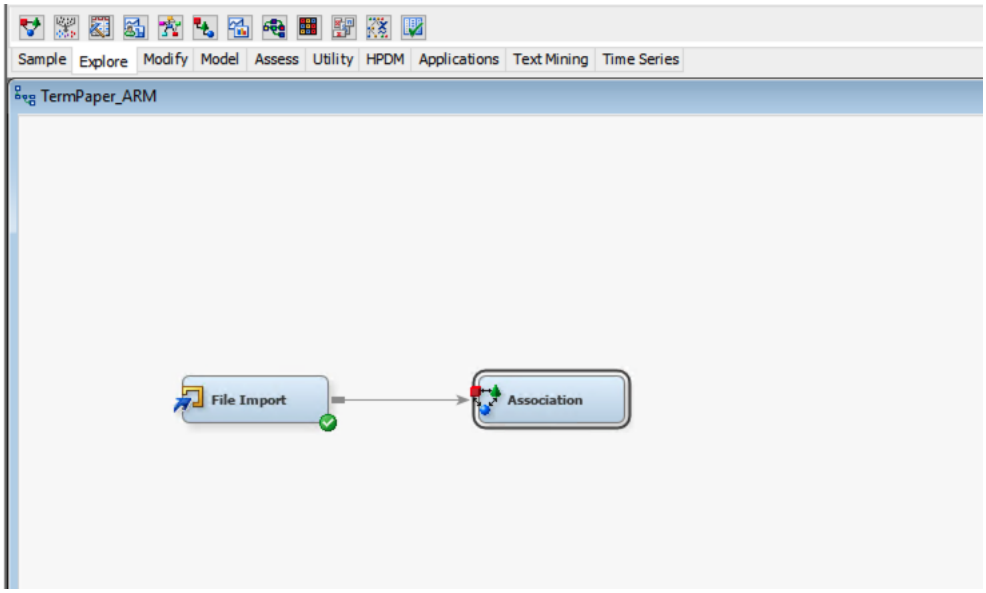


Figure 9

c. The properties of the Association node were set to default settings, where the Maximum Items is set at 4 and the Maximum Confidence level is set at 10.

Property	Value
General	
Node ID	Assoc
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Maximum Number of Items to	100000
Rules	
Association	
Maximum Items	4
Minimum Confidence Level	10
Support Type	Percent
Support Count	.
Support Percentage	5.0

Figure 10

d. Change the setting for Export Rule by ID to Yes.

Rules	
Number to Keep	200
Sort Criterion	Default
Number to Transpose	200
Export Rule by ID	Yes
Recommendation	No

Figure 11

Step 5: Run the process

At the end of step 4, we have set up the whole process and we are ready run the analysis.

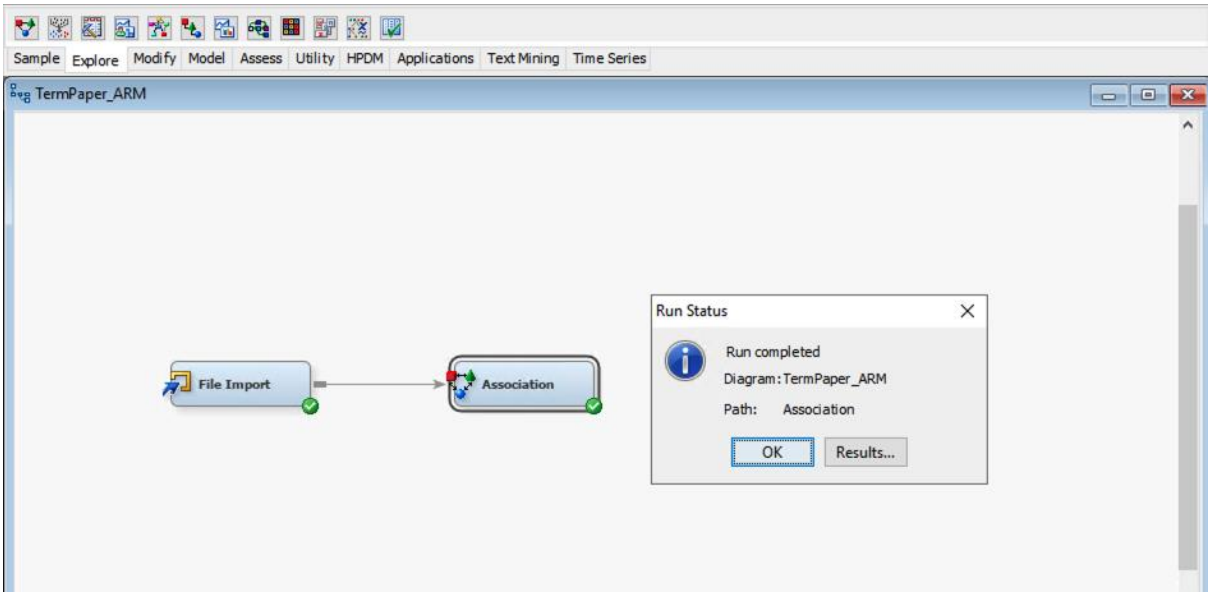


Figure 12

Results Interpretation

```
*-------------------------------------------------------------*
User:                KXM220108
Date:                March 31, 2024
Time:                12:50:03
*-------------------------------------------------------------*
* Training Output
*-------------------------------------------------------------*
```

Variable Summary

Role	Measurement Level	Frequency Count
ID	NOMINAL	1
TARGET	NOMINAL	1

Association Report

Relations	Expected Confidence (%)	Confidence (%)	Support (%)	Lift	Transaction Count	Rule	Left Hand of Rule	Right Hand of Rule	Rule Item 1	Rule Item 2	Rule Item 3	Rule Item 4	Rule Item 5	Rule Index
2	9.83	32.35	1.60	3.29	120.00	herb & pepper ==> ground beef	herb & pepper	ground beef	herb & pepper	===================>	ground beef			1
2	4.95	16.28	1.60	3.29	120.00	ground beef ==> herb & pepper	ground beef	herb & pepper	ground beef	===================>	herb & pepper			2
3	9.83	28.57	1.71	2.91	128.00	spaghetti & mineral water ==> ground beef	spaghetti & mineral water	ground beef	spaghetti	mineral water	===================>	ground beef		3
3	5.97	17.37	1.71	2.91	128.00	ground beef ==> spaghetti & mineral water	ground beef	spaghetti & mineral water	ground beef	===================>	spaghetti	mineral water		4
2	9.53	23.59	1.61	2.47	121.00	tomatoes ==> frozen vegetables	tomatoes	frozen vegetables	tomatoes	===================>	frozen vegetables			5
2	6.84	16.92	1.61	2.47	121.00	frozen vegetables ==> tomatoes	frozen vegetables	tomatoes	frozen vegetables	===================>	tomatoes			6
2	9.53	23.32	1.67	2.45	125.00	shrimp ==> frozen vegetables	shrimp	frozen vegetables	shrimp	===================>	frozen vegetables			7
2	7.15	17.48	1.67	2.45	125.00	frozen vegetables ==> shrimp	frozen vegetables	shrimp	frozen vegetables	===================>	shrimp			8
3	17.41	41.69	1.71	2.39	128.00	mineral water & ground beef ==> spaghetti	mineral water & ground beef	spaghetti	mineral water	ground beef	===================>	spaghetti		9
2	12.96	30.08	1.52	2.32	114.00	soup ==> milk	soup	milk	soup	===================>	milk			10
2	5.05	11.73	1.52	2.32	114.00	milk ==> soup	milk	soup	milk	===================>	soup			11
2	9.83	22.51	3.92	2.29	294.00	spaghetti ==> ground beef	spaghetti	ground beef	spaghetti	===================>	ground beef			12
2	17.41	39.89	3.92	2.29	294.00	ground beef ==> spaghetti	ground beef	spaghetti	ground beef	===================>	spaghetti			13
2	9.83	21.46	1.41	2.18	106.00	olive oil ==> ground beef	olive oil	ground beef	olive oil	===================>	ground beef			14
2	6.59	14.38	1.41	2.18	106.00	ground beef ==> olive oil	ground beef	olive oil	ground beef	===================>	olive oil			15
3	9.53	20.09	1.20	2.11	90.00	spaghetti & mineral water ==> frozen vegetables	spaghetti & mineral water	frozen vegetables	spaghetti	mineral water	===================>	frozen vegetables		16
3	5.97	12.59	1.20	2.11	90.00	frozen vegetables ==> spaghetti & mineral water	frozen vegetables	spaghetti & mineral water	frozen vegetables	===================>	spaghetti	mineral water		17
3	12.96	26.58	1.40	2.05	105.00	mineral water & chocolate ==> milk	mineral water & chocolate	milk	mineral water	chocolate	===================>	milk		18
3	5.27	10.80	1.40	2.05	105.00	milk ==> mineral water & chocolate	milk	mineral water & chocolate	milk	===================>	mineral water	chocolate		19
3	12.96	26.34	1.57	2.03	118.00	spaghetti & mineral water ==> milk	spaghetti & mineral water	milk	spaghetti	mineral water	===================>	milk		20
3	5.97	12.14	1.57	2.03	118.00	milk ==> spaghetti & mineral water	milk	spaghetti & mineral water	milk	===================>	spaghetti	mineral water		21
2	6.59	13.17	2.29	2.00	172.00	spaghetti ==> olive oil	spaghetti	olive oil	spaghetti	===================>	olive oil			22
2	17.41	34.82	2.29	2.00	172.00	olive oil ==> spaghetti	olive oil	spaghetti	olive oil	===================>	spaghetti			23
2	6.59	13.17	1.71	2.00	128.00	milk ==> olive oil	milk	olive oil	milk	===================>	olive oil			24
2	12.96	25.91	1.71	2.00	128.00	olive oil ==> milk	olive oil	milk	olive oil	===================>	milk			25

Figure 13

Frequent Item sets

Our analysis identified frequent item sets of varying sizes. Majority i.e. 85.50% of the frequent item sets consists of two items and the remaining 14.50% item sets consists of three items.

Association Rules

Many association rules were discovered, each of them indicating a strong relationship between items. Below are a few notable rules:

Rule 1: Herb & Pepper ==> Ground beef

If a customer purchases Herb & Pepper, there is a strong tendency (with a confidence of 32.35%) that they will also purchase Ground beef. Additionally, this association is 3.29 times more likely to occur than if Herb, Pepper and Ground beef were purchased independently.

This explain the meal & recipe choices of the customers.

Rule 2: Ground beef ==> Herb & Pepper

If a customer purchases Ground beef, there is a strong tendency (with a confidence of 16.28%) that they will also purchase Herb & Pepper. Additionally, this association is 3.29 times more likely to occur than if Ground beef, Herb and Pepper were purchased independently.

Rule 3: Spaghetti & mineral water ==> Ground beef

If a customer purchases Spaghetti & mineral water, there is a moderate probability (with a confidence of 28.57%) that they will also purchase Ground beef. The occurrence of this association is 2.91 times more likely to occur than if Herb & Pepper and Ground beef were purchased independently.

This also suggests a potential meal combination preferred by the customers.

Rule 5: Tomatoes ==> Frozen Vegetables

If a customer purchases tomatoes, there is a high probability (with a confidence of 23.59%) that they will also purchase Frozen vegetables. The occurrence of this association is 2.47 times more likely compared to if Tomatoes and Frozen vegetables were purchased independently.

This also shows customer purchase pattern that they prefer fresh tomatoes, however with respect to other vegetables they prefer frozen packs. This will help the stores maintain their inventory management well

Lift values

Higher lift values indicate stronger associations between items. Rules with higher lift values are generally more useful as they indicate stronger relationships between items. For example, rules 1 and 2 have a lift value of 3.29, indicating a strong association between Herb, Pepper and Ground beef.

Similarity

Rules that are similar in terms of antecedents and consequents may indicate redundancy. For example, rules 1 and 2 ("Herb & Pepper ==> Ground beef" and " Ground beef ==> Herb & Pepper") are essentially mirror images of each other. Such redundancy can be minimized by considering only one direction of association.

Practical Relevance

Assessing the practical relevance of rules is essential. Some rules may make intuitive sense from a business perspective and can be considered more useful. For instance, rule 9 ("Mineral water & Ground beef ==> Spaghetti") has a high confidence level (41.69%) and a moderate lift value (1.71), indicating a practical association that could inform marketing or product placement strategies.

Managerial Implications

The association role mining uncovered valuables insights about the customer behaviour, preferences & purchase patter. The strong associations indicates potential opportunities for cross-selling, marketing strategies, product placements in store & inventory management

a. **Personalized Recommendations**

   By leveraging the association rules, the business can develop personalised marketing strategies like bundle offers based on the customers purchase history & value contribution

b. **Product Placement**

   The analysis provides information about what items are purchased together informing consumer preferences, this can be used to optimise product placements within the stores. Placing complimentary items in closer proximity promotes sales.

c. **Inventory Management**

   Identify the associating between various item sets helps the business for inventory management, predicting demand for related products. This also assists to take decisions about supply-chain and meet customer demands efficiently.

Analysis 1 Conclusion

Association Rule Mining served as a powerful method for uncovering meaningful relationship and patterns about customer purchase behaviour. We were able to meet our objective of discovering actionable insights that would assists in driving revenue growth, improve operation efficiency and enhanced marketing strategies.

Analysis 2: Customer Segmentation

Dataset: Utilized the **Supermarket_CustomerMembers.csv** dataset for performing customer segmentation.

Procedure

To perform segmentation, we must first ensure that the dataset contains appropriately labelled columns that will be used for our segmentation tasks. This dataset contains appropriately labelled columns (CustomerID, Gender, Age, Annual Income, Spending Score).

CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3

Figure 14

Step 1: Create Project, Library, Diagram and data source

a. To start doing any data mining process, we first create the project, library and define the data source. Since our data source is a csv file, we use the file import block to load the data into the SAS Enterprise Miner (as done in Association Rule Mining).

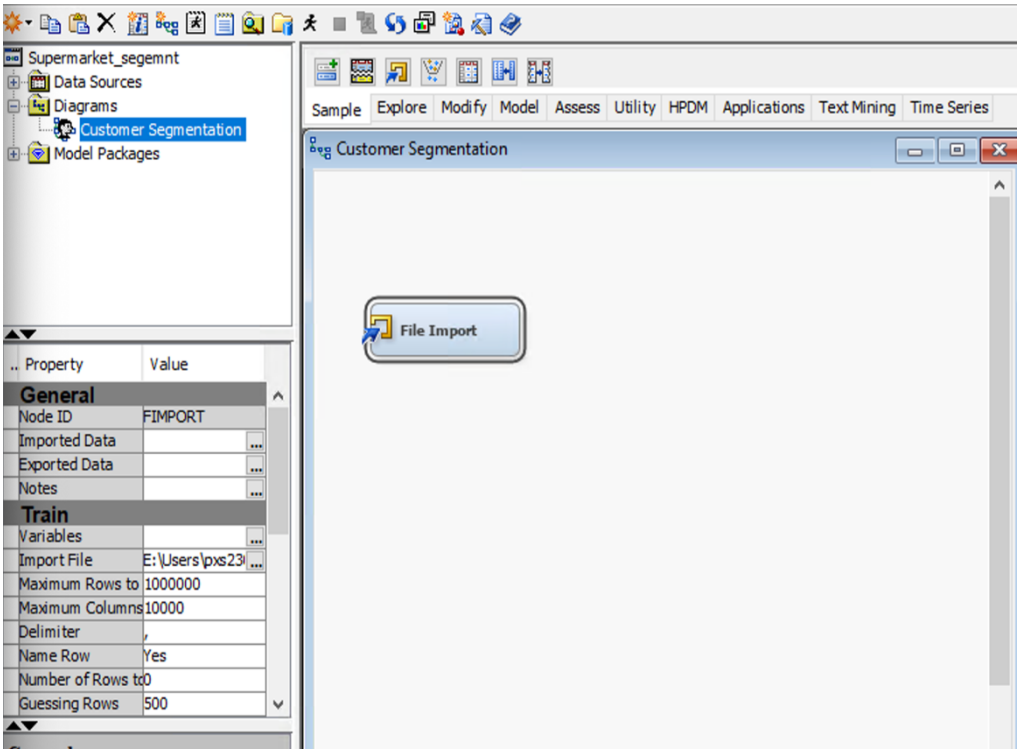
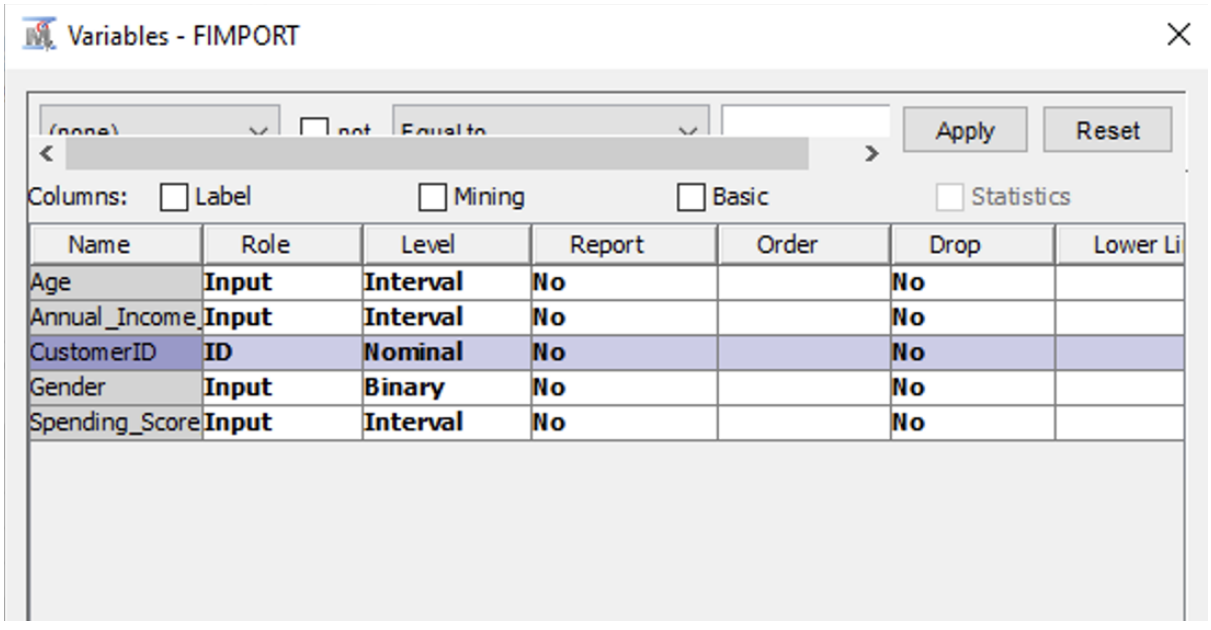


Figure 15

b. Once we load the data, we check the role of the variables and make the following changes to help SAS Miner understand the data better to give optimal results

- Set the role of CustomerID to ID
- Set the level of Gender to Binary as it has only 2 inputs (M/F)



Name	Role	Level	Report	Order	Drop	Lower Li
Age	Input	Interval	No		No	
Annual_Income	Input	Interval	No		No	
CustomerID	ID	Nominal	No		No	
Gender	Input	Binary	No		No	
Spending_Score	Input	Interval	No		No	

Figure 16

Step 2: Analyse missing values using stat explore node

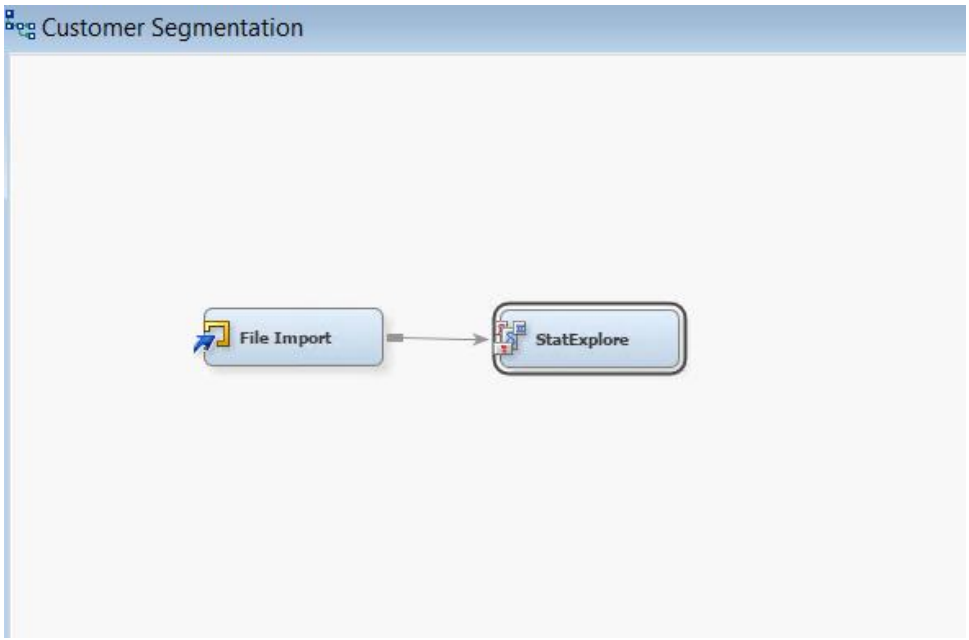


Figure 17

To make sure that the data set obtained is free of error or missing values, we use the stat explore node in SAS EM and check for the same.

Once you connect the File Import node with the StatExplore node and run it, you obtain the following results:

Output

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Gender	INPUT	2	0	Female	56.00	Male	44.00

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Age	INPUT	38.85	13.96901	200	0	18	36	70	0.485569	-0.67157
Annual_Income__k__	INPUT	60.56	26.26472	200	0	15	61	137	0.321843	-0.09849
CustomerID	INPUT	100.5	57.87918	200	0	1	100	200	0	-1.2
Spending_Score__1_100_	INPUT	50.2	25.82352	200	0	1	50	99	-0.04722	-0.82663

Figure 18

The Missing column values are zero, indicating that the dataset contains no missing values and is appropriate for use for further analysis.

Step 3: Clustering using default SAS EM settings

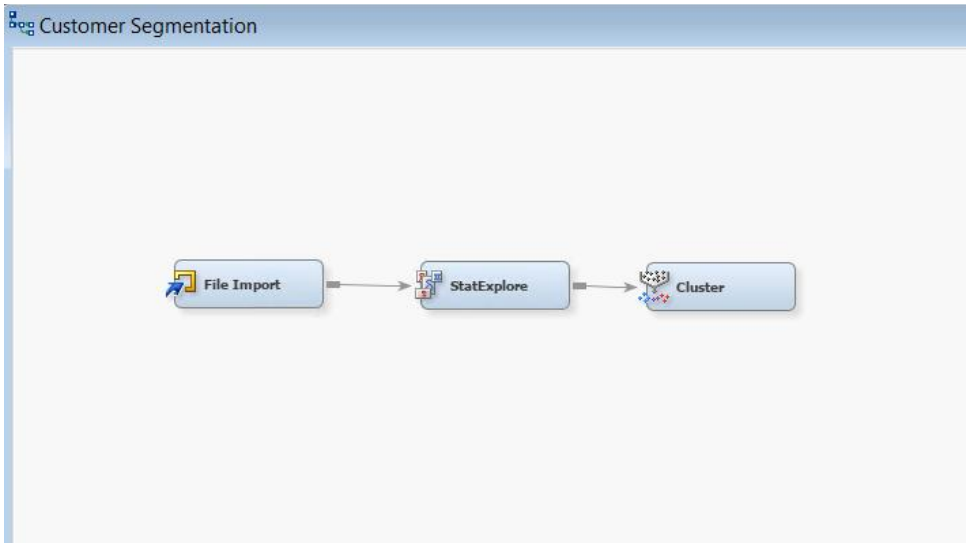


Figure 19

Once we confirm that the data is free of any error, we perform segmentation by connecting the StatExplore node to the Cluster node and by using the default SAS EM Settings. We also make sure to set internal standardisation to standardization to standardize all columns in the same range to get meaningful results.

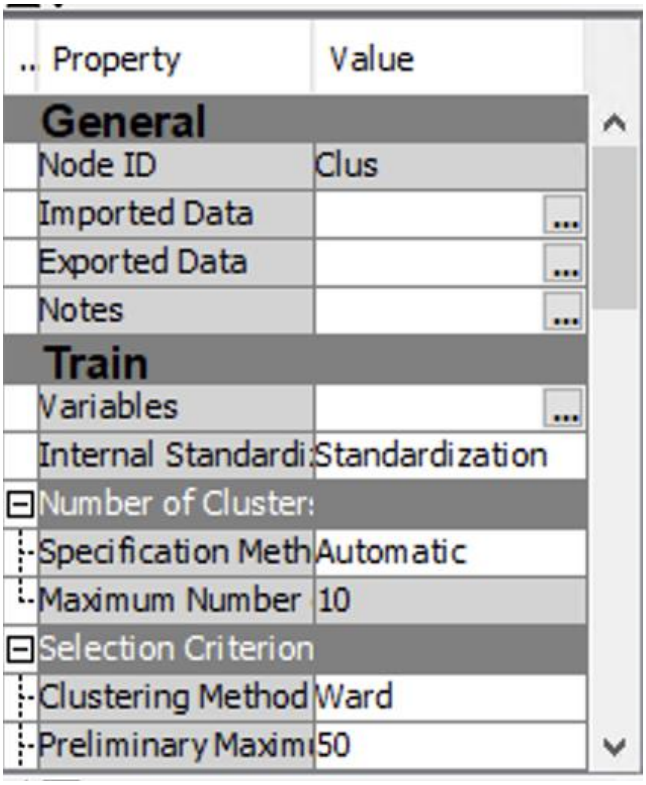


Figure 20

The data set was partitioned into 20 clusters when we used the default standardization settings (Automatic) for segmentation. Because this number appears to be excessive, implying that the data may have been over-segmented, we must manually lower the number of clusters to create a more acceptable segmentation, which will provide us with more insights into the dataset under consideration.

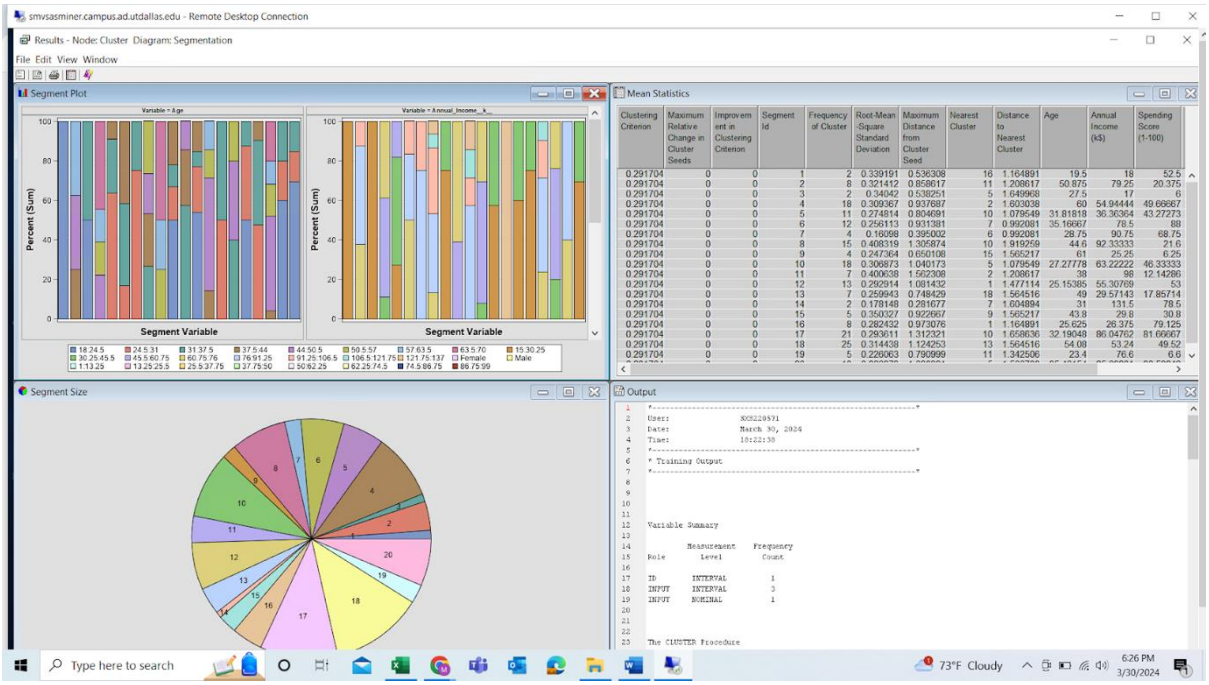


Figure 21

Step 4: User Defined Segmentation

In this step, manually select the number of clusters by changing the specification method to user specified and entering the desired value.

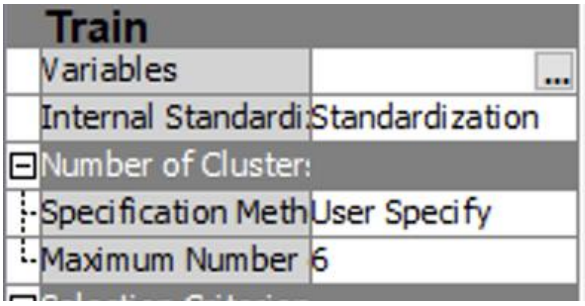


Figure 22

In this case, we manually specify the number of segments to be 6 and run the nodes. In the segment size output, we can notice that all the clusters are well populated unlike the earlier segmentation which had uneven distribution of data. This indicates that lesser number of segmentations will improve the quality of it.

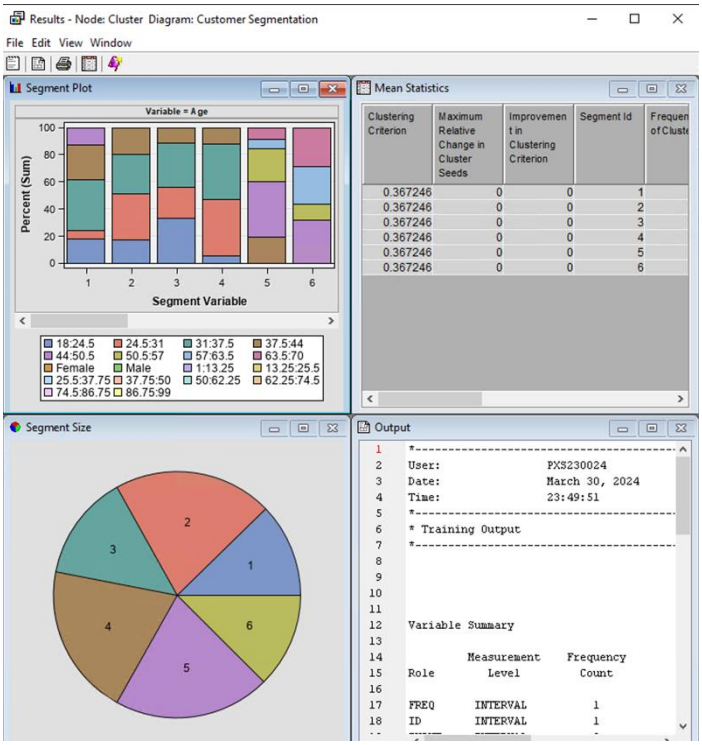


Figure 23

Step 5: Add a segment profile

We can analyse the results of the segmentation using the segment profile node. For that we add the segment profile node and connect the cluster node.

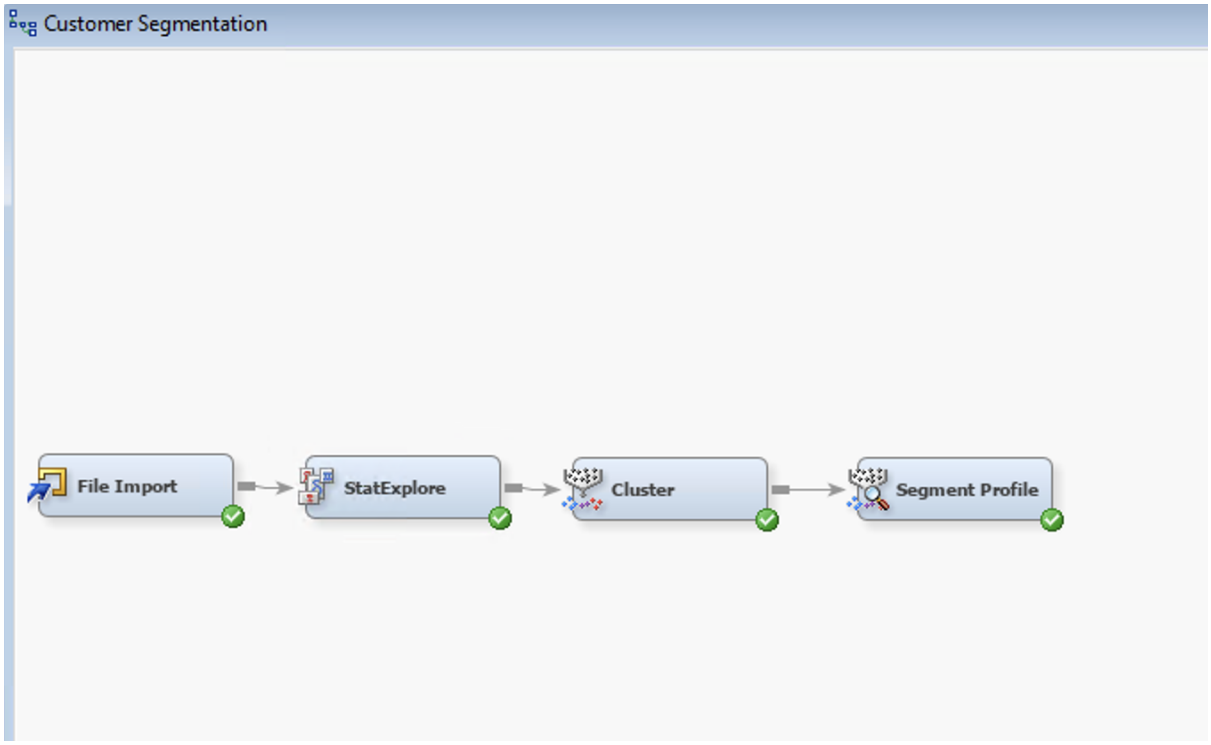


Figure 24

Once we run the nodes, we open the results and check the segment profile

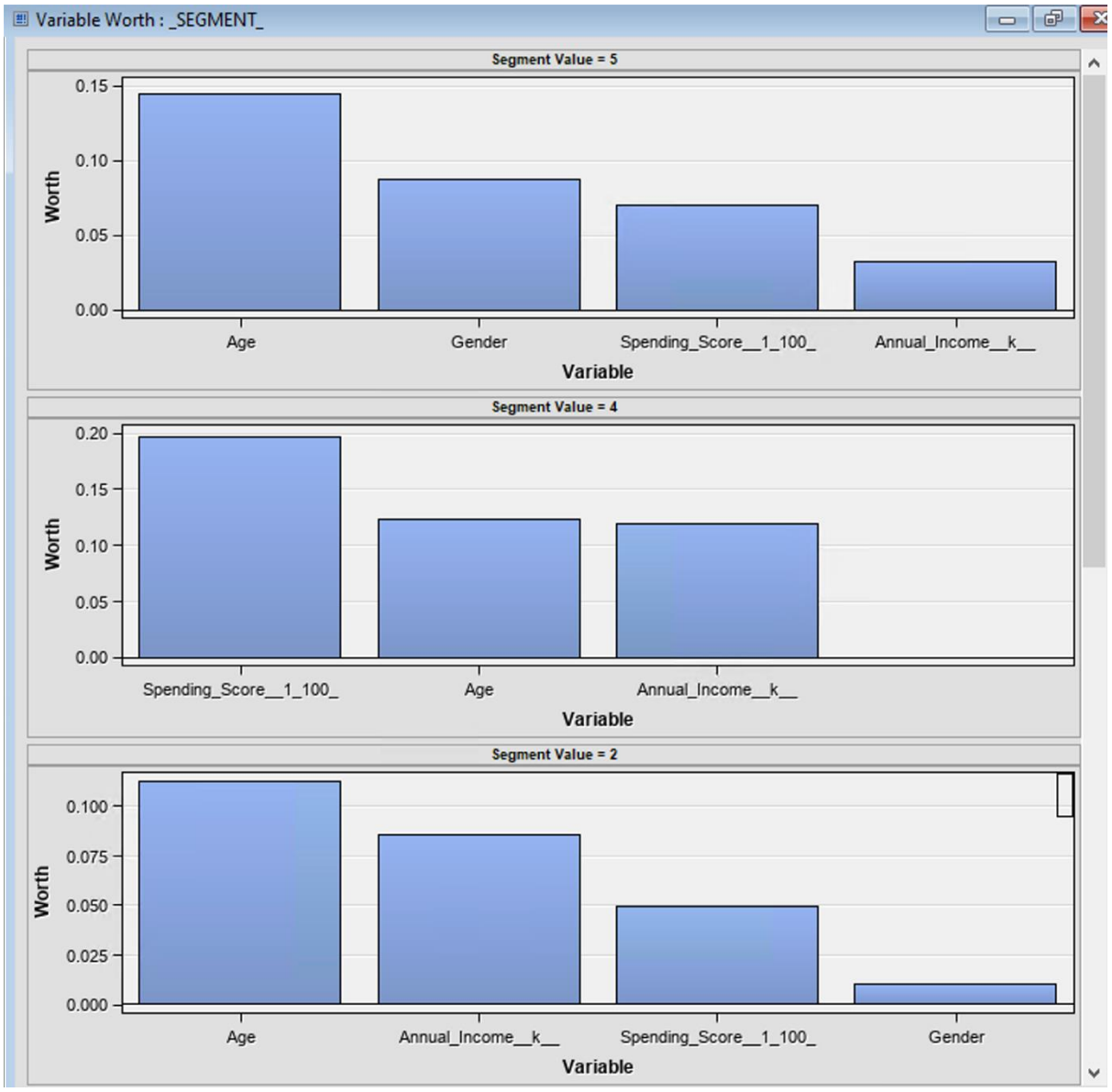


Figure 25

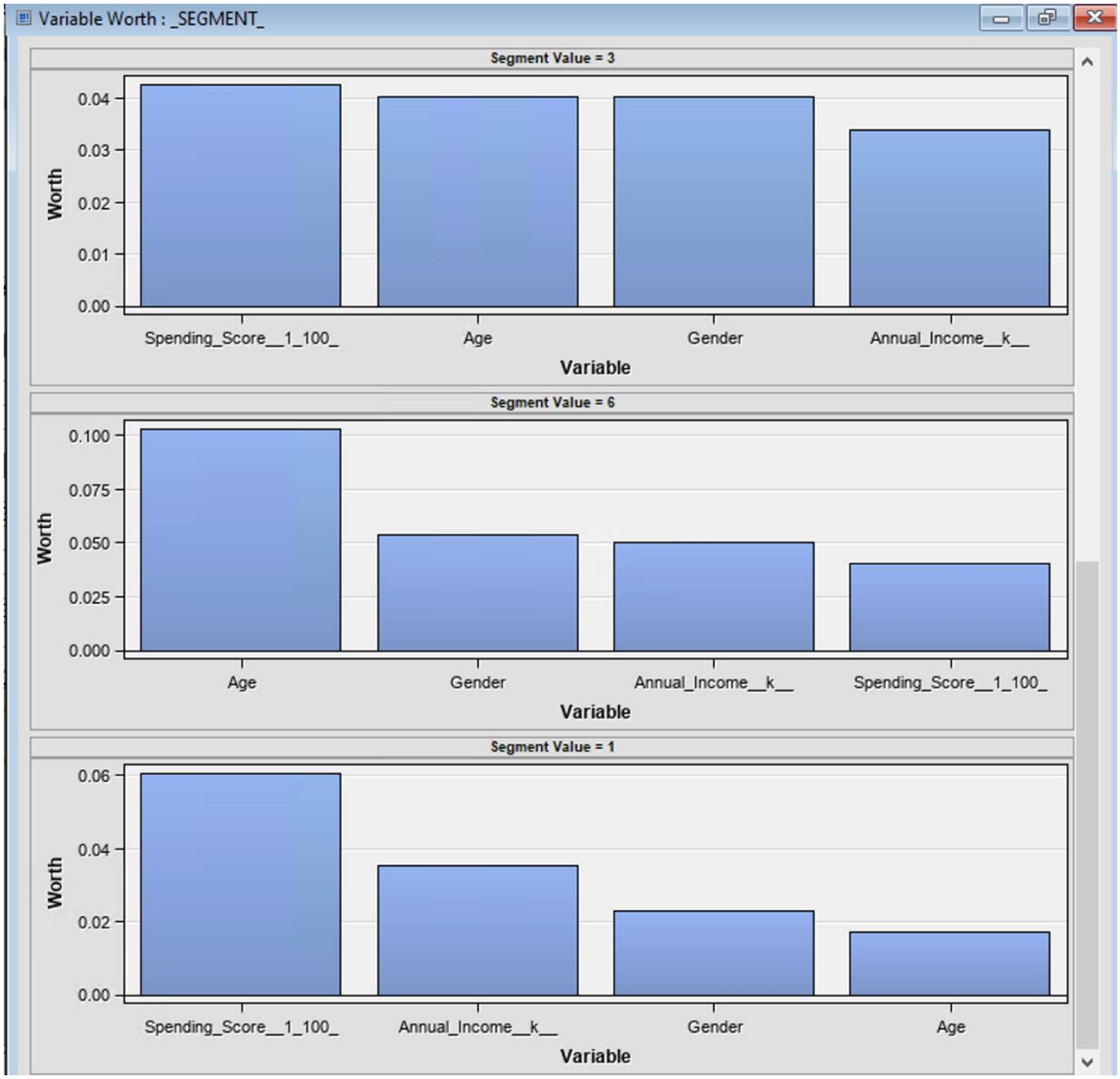


Figure 26

From this we notice that even though the clusters are evenly distributed, there is no clear distinguishing factors in between clusters, making it difficult to interpret the segments and customer profile of each segment. So, we try and reduce the number of segments even more.

Step 6: Manually specifying number of clusters as 3

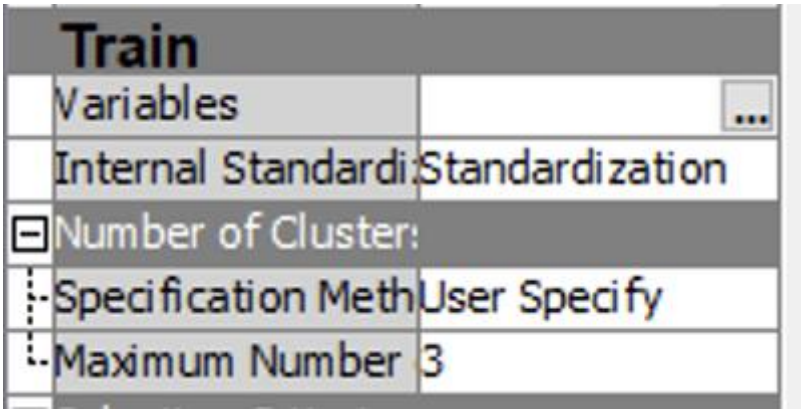


Figure 27

When we use 3 clusters, we get an evenly distributed segmentation, enabling us to do a more refined segmentation, which can benefit the supermarket to get a better understanding of their customers and on what basis they need to fix their targets for the betterment of their profits and sales.

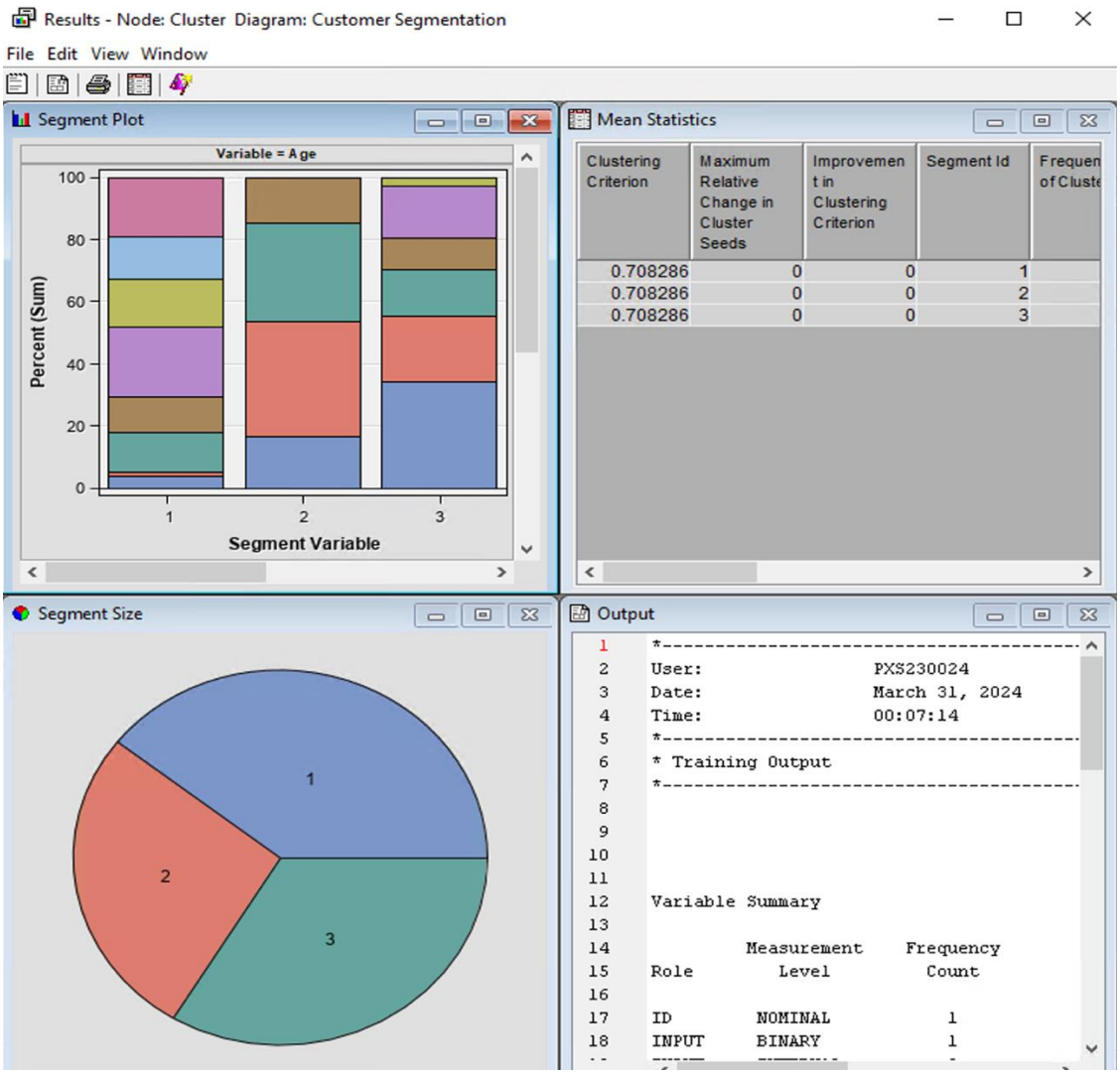


Figure 28

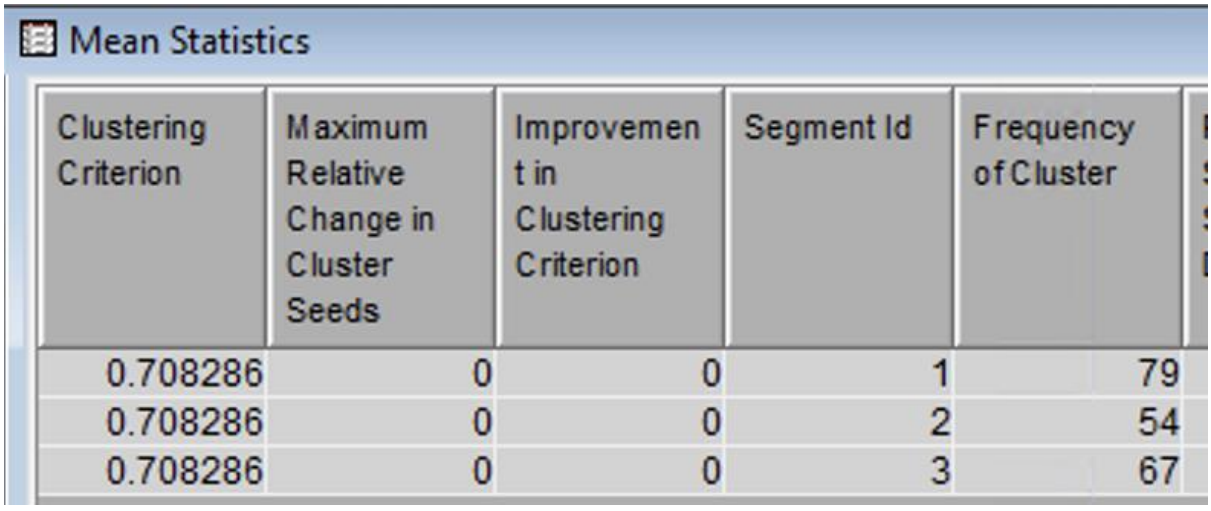


Figure 29

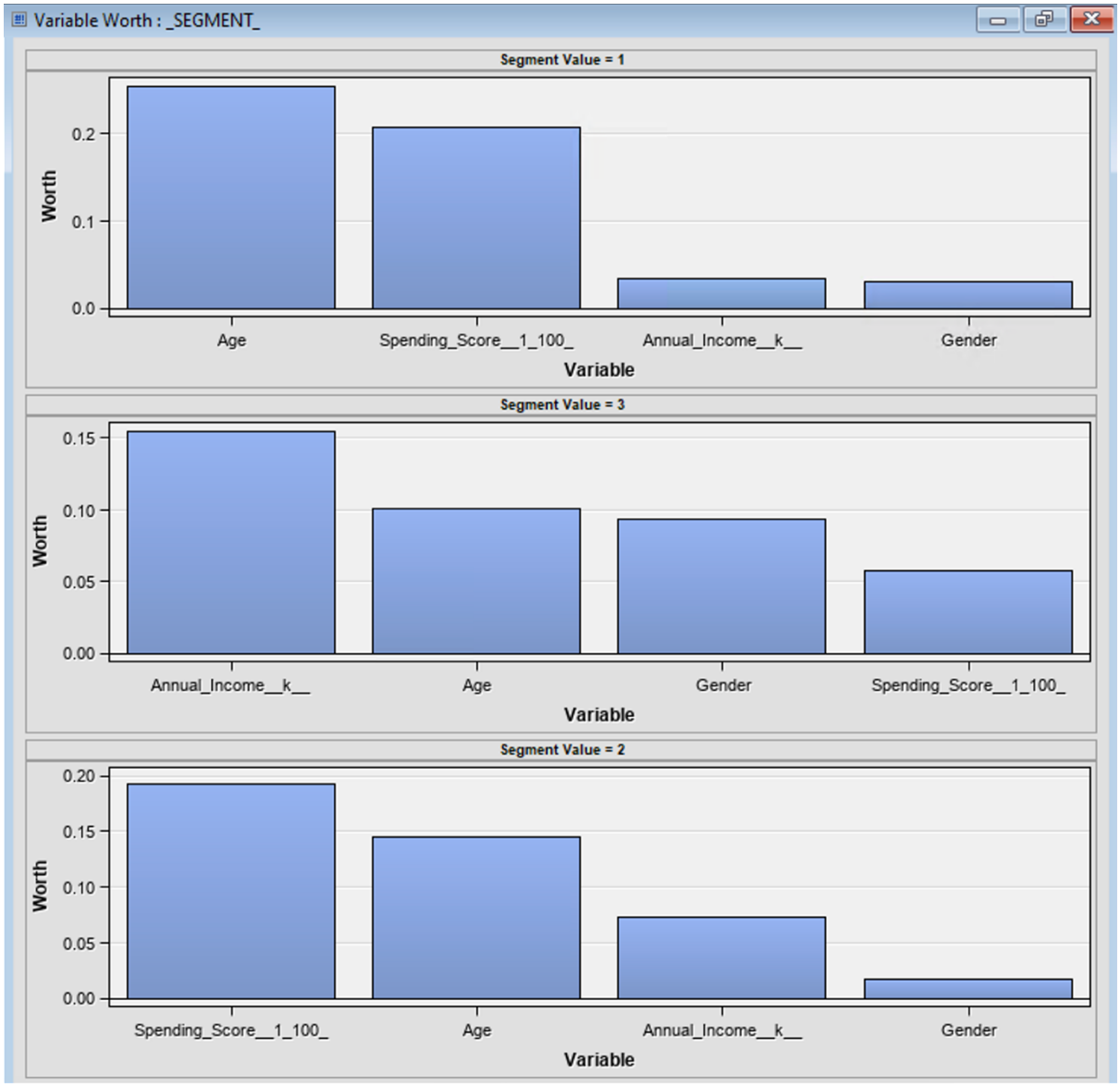

Variable Worth : _SEGMENT_
Segment Value = 1
Worth
0.2
0.1
0.0
Age
Spending_Score__1_100_
Annual_Income__k__
Gender
Variable
Segment Value = 3
Worth
0.15
0.10
0.05
0.00
Annual_Income__k__
Age
Gender
Spending_Score__1_100_
Variable
Segment Value = 2
Worth
0.20
0.15
0.10
0.05
0.00
Spending_Score__1_100_
Age
Annual_Income__k__
Gender
Variable

Figure 30

Figure 31

The following is our interpretation of each segment:

Segment 1: Middle-aged Shoppers

- This segment has the highest count of 79 customers.
- The most important variables for this segment, according to clustering, are Age, Spending Score, Annual Income, and Gender (in descending order of importance).
- Based on the importance of variables, we can infer that this segment might consist of middle-aged to older customers with varying income levels and spending habits.

Segment 2: High Spenders

- This segment has a count of 54 customers.
- The most important variables for this segment are Spending Score, Age, Annual Income, and Gender.
- This segment appears to be characterized by customers with a strong emphasis on their spending scores, followed by age and income.

Segment 3: Income-driven Shoppers

- This segment has a count of 67 customers.
- The most important variables for this segment are Annual Income, Age, Gender, and Spending Score.
- It seems that for this segment, annual income is the primary distinguishing factor, followed by age and gender.

Managerial Implications

As business analysts, we would suggest the management of the Superstore XYZ to use the results of the segmentation we performed in the following ways:

1. **Targeted Marketing Campaigns:**

   Tailor marketing campaigns to each segment's characteristics and preferences. For example, promotions and advertisements could focus on different aspects such as age, income level, or spending habits that are particularly relevant to each segment.

2. **Product Assortment and Placement:**

   Adjust product assortment and placement within the store based on the preferences of each segment. For instance, products appealing to high-spending customers could be prominently displayed, while items catering to budget-conscious shoppers may be placed strategically elsewhere.

3. **Customer Experience Enhancement:**

   Customize the customer experience based on segment characteristics. This could involve personalized recommendations, special discounts, or loyalty programs tailored to the preferences and behaviours of each segment.

4. **Market Expansion:**

   Use insights from this segmentation analysis to identify opportunities for market expansion. Understanding the characteristics of different segments can guide decisions regarding new product launches or expansion into new market segments.

Analysis 2 Conclusion

Segmentation may assist any business in narrowing down their consumers' wants and needs, allowing them to provide a more personalized shopping experience. It can also help businesses identify new target categories as well as those to avoid, allowing them to improve their company strategies overall.

Analysis 3: Decision Tree Analysis

Dataset: Utilized the **Supermarket_CustomerMembers.csv** dataset for performing Decision Tree.

Procedure

To perform regression, we must first ensure that the dataset contains appropriately labelled columns that will be used for our segmentation tasks. This dataset contains appropriately labelled columns (CustomerID, Gender, Age, Annual Income, Spending Score).

CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3

Figure 32

Step 1: Create Project, Library, Diagram, and data source

To start doing any data mining process, we first create the project, library and define the data source. Since our data source is a csv file, we use the file import block to load the data into the SAS Enterprise Miner (as done in Association Rule Mining).

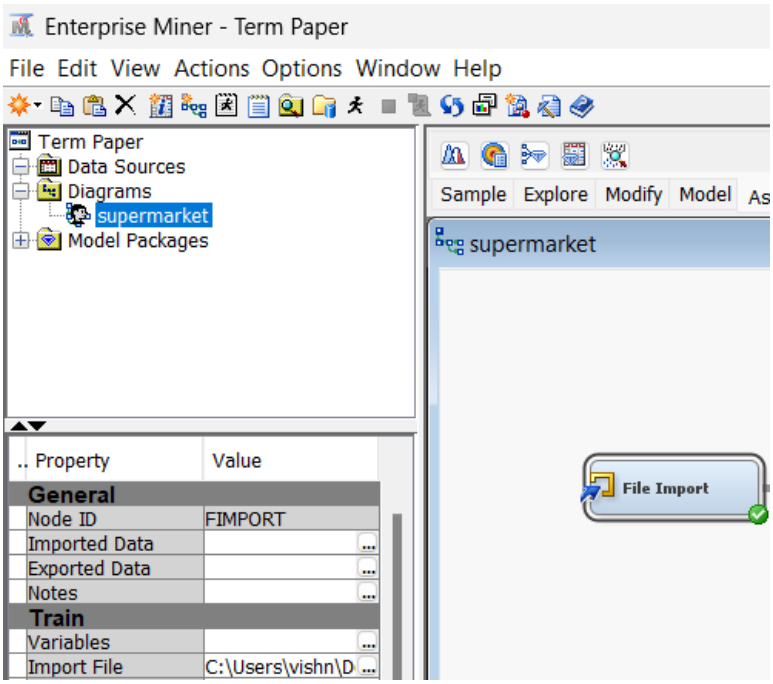


Figure 33

Step 2: We complete the diagram by adding the required nodes to perform decision tree analysis.

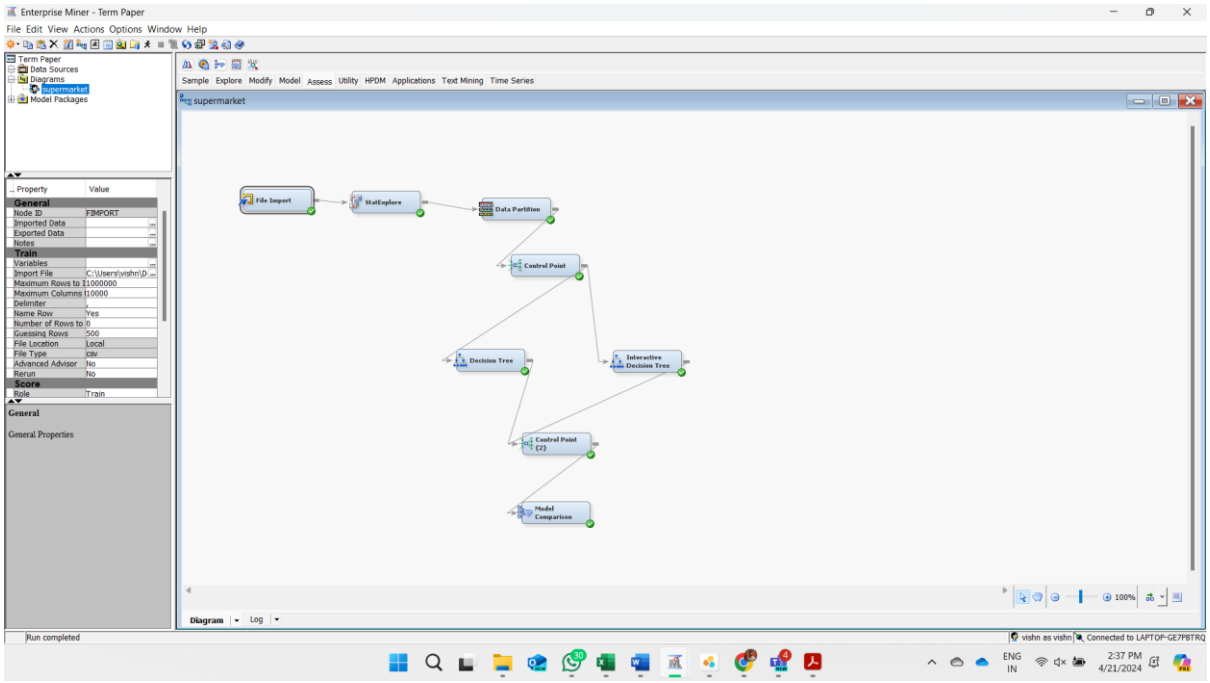


Figure 34

We look at the results:

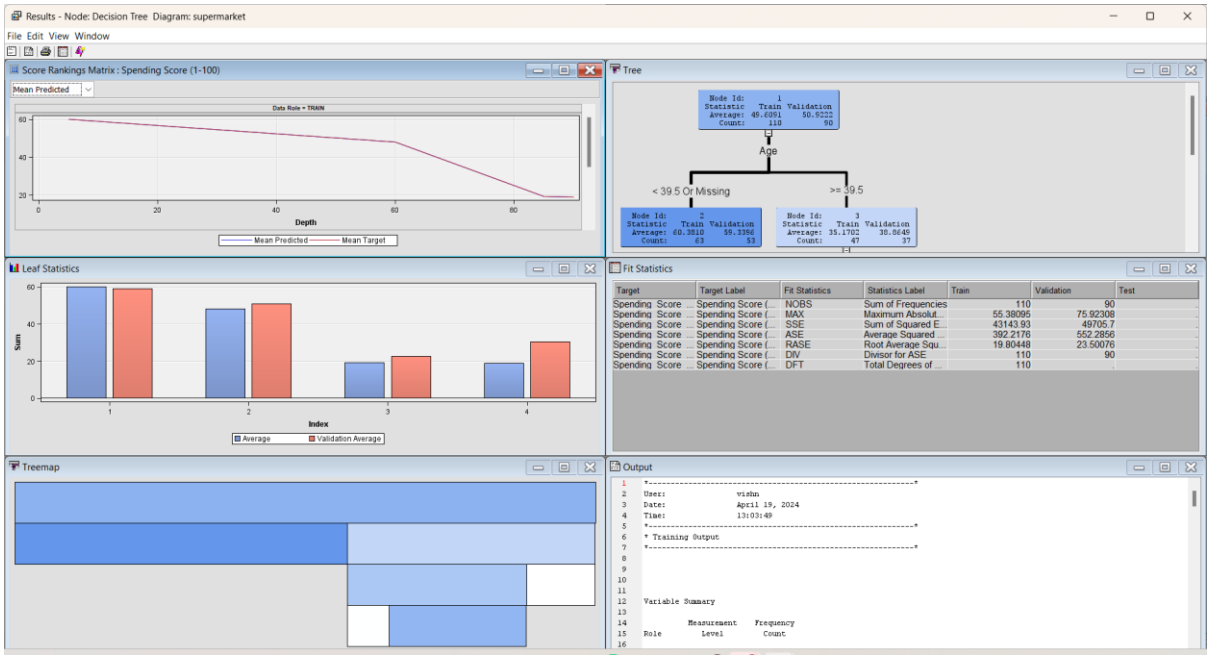


Figure 35

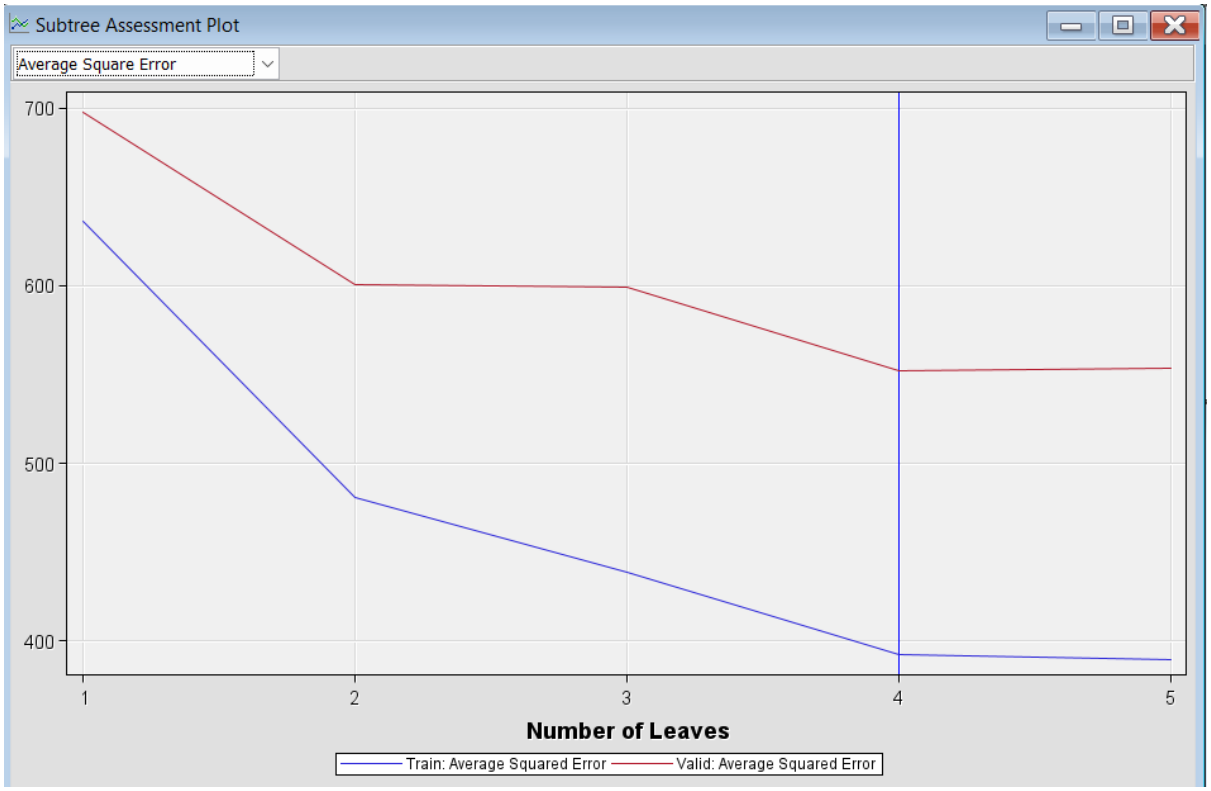


Figure 36

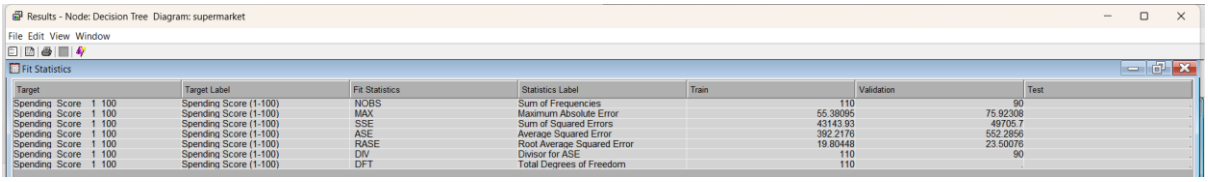
From this Subtree Assessment Plot, we can make the following inferences:

The plot shows the Average Square Error for both the training data (solid blue line) and validation data (dashed red line) as the number of leaves increases from 1 to 5. Both lines

exhibit a decreasing trend, indicating that as the number of leaves increases, the Average Square Error decreases for both the training and validation data.

However, the Average Square Error for the validation data is consistently higher than that of the training data across all values of the number of leaves. This suggests that while increasing the number of leaves improves the model's performance on both the training and validation data, there is still some degree of overfitting, as the model performs better on the training data it was trained on compared to the unseen validation data.

Overall, the plot suggests that increasing the number of leaves in the model can lead to better performance, but there is a trade-off between model complexity and generalization to new data, as evidenced by the gap between the training and validation error curves.



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Spending Score 1 100	Spending Score (1-100)	NOBS	Sum of Frequencies	110	90	
Spending Score 1 100	Spending Score (1-100)	MAX	Maximum Absolute Error	55.38095	75.92308	
Spending Score 1 100	Spending Score (1-100)	SSE	Sum of Squared Errors	43143.93	49705.7	
Spending Score 1 100	Spending Score (1-100)	ASE	Average Squared Error	392.2176	552.2856	
Spending Score 1 100	Spending Score (1-100)	RASE	Root Average Squared Error	19.80448	23.50076	
Spending Score 1 100	Spending Score (1-100)	DIV	Divisor for ASE	110	90	
Spending Score 1 100	Spending Score (1-100)	DFT	Total Degrees of Freedom	110		

Figure 37

Based on the fit statistics table, we can say that:

The target variable being predicted is the "Spending Score", which appears to be a numerical score ranging from 1 to 100. This score represents some measure of customer spending or sales in the supermarket context.

The fit statistics provided include the Node (1.0), Sum of Squared Error (SSE), Root Average Squared Error (RASE), Degrees of Freedom (DFT), and Total Degrees of Freedom (110). The RASE value of 19.80448 for the training data indicates the average magnitude of the prediction errors made by the decision tree model.

The validation data has a slightly higher RASE value of 23.50076 compared to the training data, suggesting that the model's performance is slightly worse on unseen data, which is expected due to overfitting.

Other relevant metrics for the training data include the Sum of Frequencies (552.2856), Maximum Absolute Error (53.38095), Sum of Absolute Errors (43143.93), and Average

Squared Error (19.80448). These provide insights into the overall magnitude and distribution of errors made by the model on the training data.

For the validation data, the key metric provided is the Sum of Frequencies (90), which is lower than the training data, indicating a smaller sample size for validation.

Overall, the decision tree model is reasonably accurate on the training data but exhibits some degradation in performance on the unseen validation data, as evidenced by the higher RASE value. The provided metrics give a comprehensive view of the model's fit and predictive performance on both datasets.

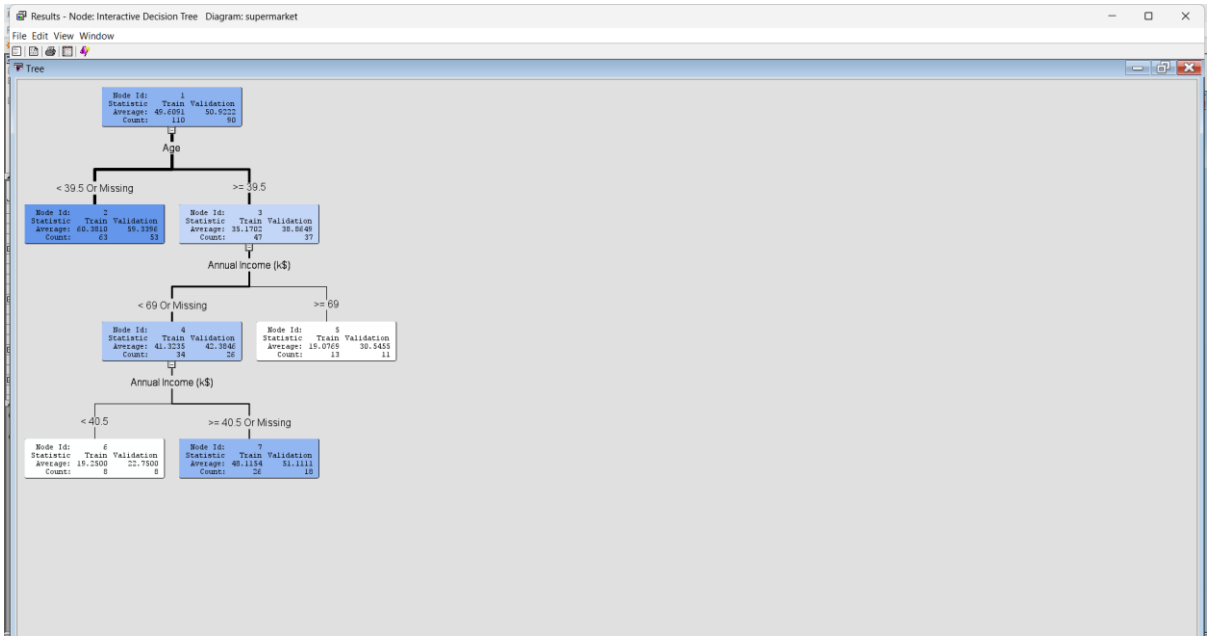


Figure 38

For the decision tree model, interactive training was performed to visually build the tree structure by selectively splitting nodes based on variable importance and logworth values. There were no significant difficulties encountered or rework required during this process. However, it provided a valuable learning experience in understanding how to construct decision trees interactively, specify split points, and interpret the resulting tree visualization.

The interactive decision tree training process begins by allowing SAS Enterprise Miner to automatically grow an initial tree. Then nodes are manually selected for splitting based on the logworth metric, which measures how effective each input variable is at

separating the target classes. Variables with higher logworth values are stronger candidates for splitting. At each node, the top variable is chosen and users can customize the split point value or let the software determine the optimal binary split. This iterative splitting adds new child nodes to the tree.

Thick branches in the tree visualization indicate the node have a higher percentage of correctly classified observations compared to thin branches. Interactively growing the tree highlights which variables are most important at each level for discriminating the target. Users can also selectively stop splitting nodes once the logworth values become very small, avoiding overfitting. This approach allows intuitive understanding of how the decision rules are constructed.

The logworth metric is used because it is an easily interpretable value indicating the significance of a variable split compared to the null model. Higher positive logworth indicates the split is more important for separating the target classes. Variables are ranked by logworth at each node to guide the interactive splitting decisions. Specifying split point thresholds gives flexibility in defining the decision rules. Overall, the interactive process makes the tree model more transparent and overcomes potential issues with fully automatic growth. We stopped the interactive tree model when the logworth value became zero.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Age		1	1.0000	1.0000	1.0000
Annual_Income__k__	Annual Income (k$)	2	0.7550	0.7050	0.9338

Figure 39

Based on the screenshot, which shows variable importance information, we can derive the following insights:

1. Variable Importance: The table displays the importance of two variables, "Age" and "Annual_Income_k", in the decision tree model. Variable importance measures how much each variable contributes to the prediction or decision-making process.

2. Number of Splitting Rules: The "Number of Splitting Rules" column indicates the number of times each variable was used to split the data in the decision tree. Age was used for splitting once, while Annual_Income_k was used twice.

3. Importance Scores: The "Importance" column shows the numerical scores representing the relative importance of each variable in the model. Age has an importance score of 1.0000, which is the highest, indicating it is the most important predictor variable. Annual_Income_k has a lower importance score of 0.7750, suggesting it is less important than Age for predicting the target variable.

4. Validation Importance: The "Validation Importance" column displays the importance scores for the validation dataset. These scores are typically used to assess the generalization performance of the model on unseen data. The validation importance scores for Age (1.0000) and Annual_Income_k (0.7050) are similar to their training counterparts, indicating consistent variable importance across the training and validation datasets.

5. Ratio of Validation to Training Importance: The final column shows the ratio of validation importance to training importance. A ratio close to 1 suggests that the variable's importance is consistent between the training and validation datasets. For Age, the ratio is exactly 1.0000, indicating perfect consistency. For Annual_Income_k, the ratio is 0.9338, which is slightly lower but still reasonably consistent.

Overall, the insights from this screenshot suggest that Age is the most important predictor variable in the decision tree model, followed by Annual_Income_k. The importance scores are relatively consistent between the training and validation datasets, indicating good generalization performance for these variables.

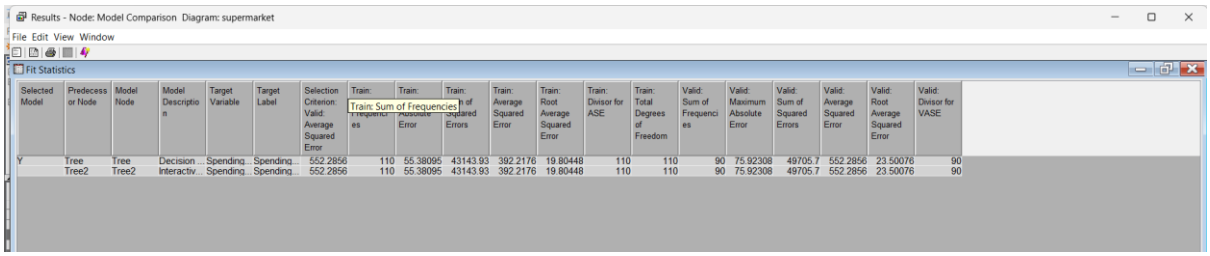


Figure 40

Based on the screenshot displaying the "Model Comparison Diagram" for the "supermarket" dataset, we can make the following inferences:

The two models being compared are "Tree" and "Decision Interactive Tree". Both models are targeting the "Spending" variable, which represents customer spending or sales data in a supermarket setting.

The "Selection Criterion" used for both models is "Average Squared Error", indicating that the models are evaluated based on their ability to minimize the average squared difference between predicted and actual values.

Several metrics are provided for both the training and validation datasets. For the training data, both models have the same "Sum of Frequencies" (552.2856), "Sum of Absolute Errors" (43143.93), "Sum of Squared Errors" (392.2176), "Average Squared Error" (19.80448), "Degrees of Freedom" (110), and "Root Average Squared Error" (49705.7). This suggests that the models perform identically on the training data.

However, for the validation data, there are slight differences in the "Average Squared Error" (Tree: 23.50076, Decision Interactive Tree: 23.50076) and "Root Average Squared Error" (Tree: 90, Decision Interactive Tree: 90). These differences, although small, indicate that the models may have slightly different generalization performance on unseen data.

Overall, the comparison suggests that both models perform similarly on the training data but may have minor differences in their ability to generalize to new data, as evidenced by the slightly different validation metrics.

Managerial Implications

The decision tree analysis performed on the supermarket customer dataset provides valuable insights that can inform business strategies and decision-making for the supermarket management. Here are the key managerial implications:

1. Customer Segmentation: The decision tree model identifies the most important variables for predicting customer spending scores, which are Age and Annual Income. This information can be leveraged to segment customers based on these demographic

characteristics, enabling more targeted marketing campaigns, and tailored promotional offers.

2. Resource Allocation: By understanding the spending patterns of different customer segments, the supermarket can allocate resources more effectively. For instance, they may choose to focus more marketing efforts and inventory management strategies on the segments with higher spending potential, such as older customers or those with higher annual incomes.

3. Personalized Customer Experience: The insights from the decision tree model can be used to personalize the customer experience within the supermarket. For example, the layout, product placement, and in-store promotions could be optimized based on the preferences and spending patterns of specific customer segments identified by the model.

4. Predictive Analytics: While the decision tree model provides valuable insights, it can also be used for predictive analytics. By incorporating new customer data into the model, the supermarket can estimate the potential spending scores of new or existing customers, enabling more effective targeting and resource allocation.

5. Continuous Improvement: The decision tree analysis should be considered an iterative process. As new customer data becomes available or market conditions change, the model can be updated and refined to maintain its predictive accuracy and relevance.

Analysis 3 Conclusion

It is important to note that we could not add a score node and perform scoring in this analysis due to the absence of a dedicated scoring dataset. Scoring is typically performed on a separate dataset to evaluate the model's performance on unseen data and generate predictions for new instances. Without a designated scoring dataset, we cannot assess the model's predictive capabilities on new customer data or generate spending score predictions for potential customers. However, the insights derived from the decision tree analysis can still inform business strategies and decision-making processes within the supermarket.

Analysis 4: Linear Regression Analysis

Dataset: Utilized the 50_Supermarket_Branches.csv dataset for performing Linear Regression.

Procedure

To perform Linear Regression, we first ensured that the dataset contains appropriately labelled columns that will be used for the model. This dataset contains appropriately labelled columns (Branch_Id, Advertisement Spend, Promotion Spend, Administration Spend, State and Profit)

A	B	C	D	E	F
Branch_Id	Advertisement Spend	Promotion Spend	Administration Spend	State	Profit
1	165349.2	136897.8	471784.1	New York	192261.83
2	162597.7	151377.59	443898.53	California	191792.06
3	153441.51	101145.55	407934.54	Florida	191050.39
4	144372.41	118671.85	383199.62	New York	182901.99
5	142107.34	91391.77	366168.42	Florida	166187.94
6	131876.9	99814.71	362861.36	New York	156991.12
7	134615.46	147198.87	127716.82	California	156122.51
8	130298.13	145530.06	323876.68	Florida	155752.6
9	120542.52	148718.95	311613.29	New York	152211.77
10	123334.88	108679.17	304981.62	California	149759.96

Figure 41

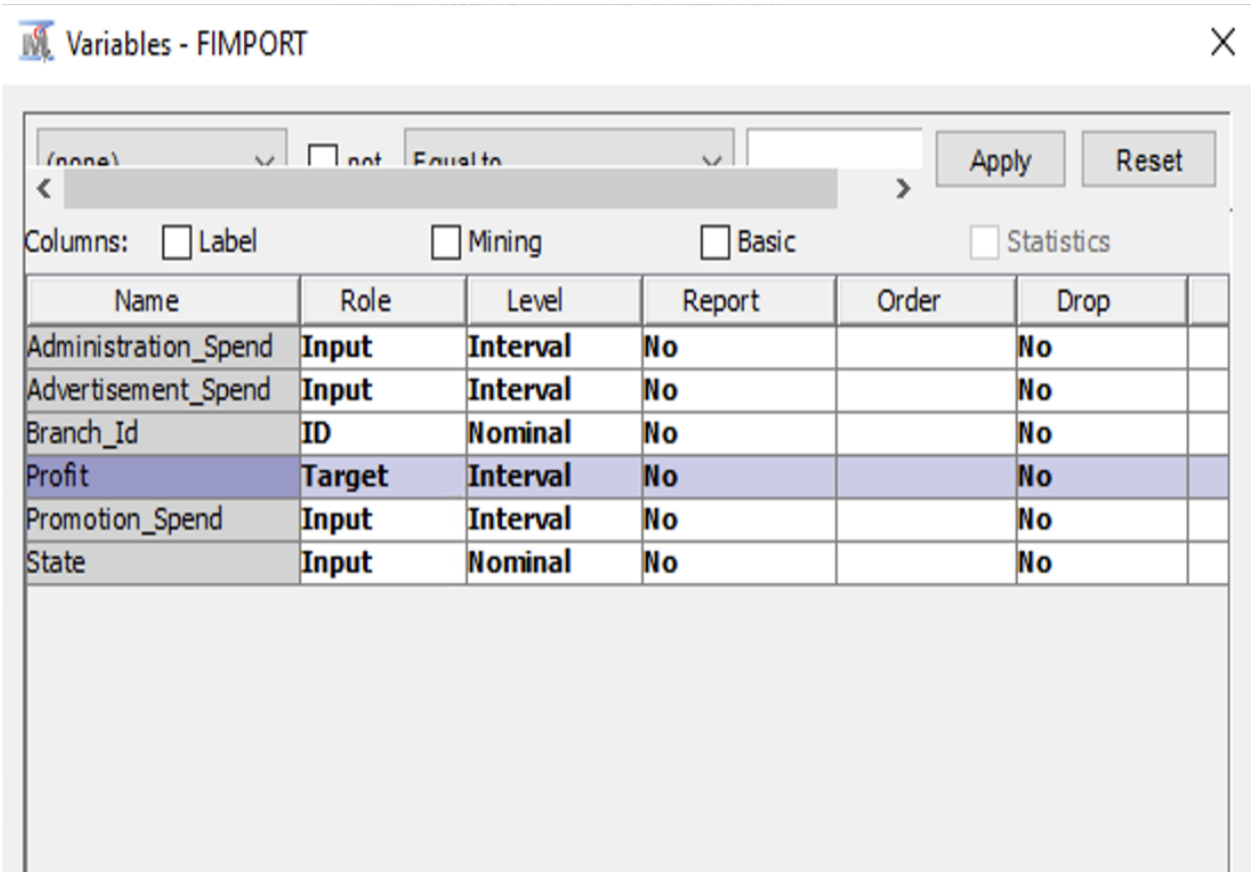
Step 1: Start a New Project, Library and Diagram

a. We first create the project, library, Diagram and data source.

To start doing any data mining process, we first create the project, library and define the data source. Since our data source is a csv file, we use the file import block to load the data into the SAS Enterprise Miner (as done before)

Figure 42

b. Define Role and Level: After importing, we define the role and measurement level of each variable.
    - Advertisement Spend, Promotion Spend, Administration Spend, and Profit is set to 'input' and 'interval'.
    - State is set to 'input' and 'nominal'.
    - Profit is set as 'target'.



Name	Role	Level	Report	Order	Drop
Administration_Spend	Input	Interval	No		No
Advertisement_Spend	Input	Interval	No		No
Branch_Id	ID	Nominal	No		No
Profit	Target	Interval	No		No
Promotion_Spend	Input	Interval	No		No
State	Input	Nominal	No		No

Figure 43

Step 2: Explore and Prepare Data

Data Exploration: To make sure that the data set obtained is free of error or missing values, we use the stat explore node in SAS EM and check for the same.

Once you connect the File Import node with the StatExplore node and run it, you obtain the following results:

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median
Administration_Spend	INPUT	211025.1	122290.3	50	0	0	210797.7
Advertisement_Spend	INPUT	73721.62	45902.26	50	0	0	72107.6
Promotion_Spend	INPUT	121344.6	28017.8	50	0	51283.14	122616.8
Profit	TARGET	112012.6	40306.18	50	0	14681.4	107404.3

Figure 44

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	State	INPUT	3	0	California	34.00	New York	34.00

Figure 45

The Missing column values are zero, indicating that the dataset contains no missing values and is appropriate for use for further analysis.

a. Correlation Plot:

In the output from the statistical exploration node, we have the option to observe the correlation plot (Pearson coefficient).

The correlation plot reveals that "Profit" is most strongly correlated with "Advertisement_spend," indicating that higher advertising expenditure is associated with increased profitability. "Administration_spend" shows a moderately high correlation with profit, suggesting its influence on profitability, albeit slightly less than advertising spend. Conversely, "Promotion_spend" exhibits the weakest correlation with profit, implying a comparatively smaller impact on profitability.

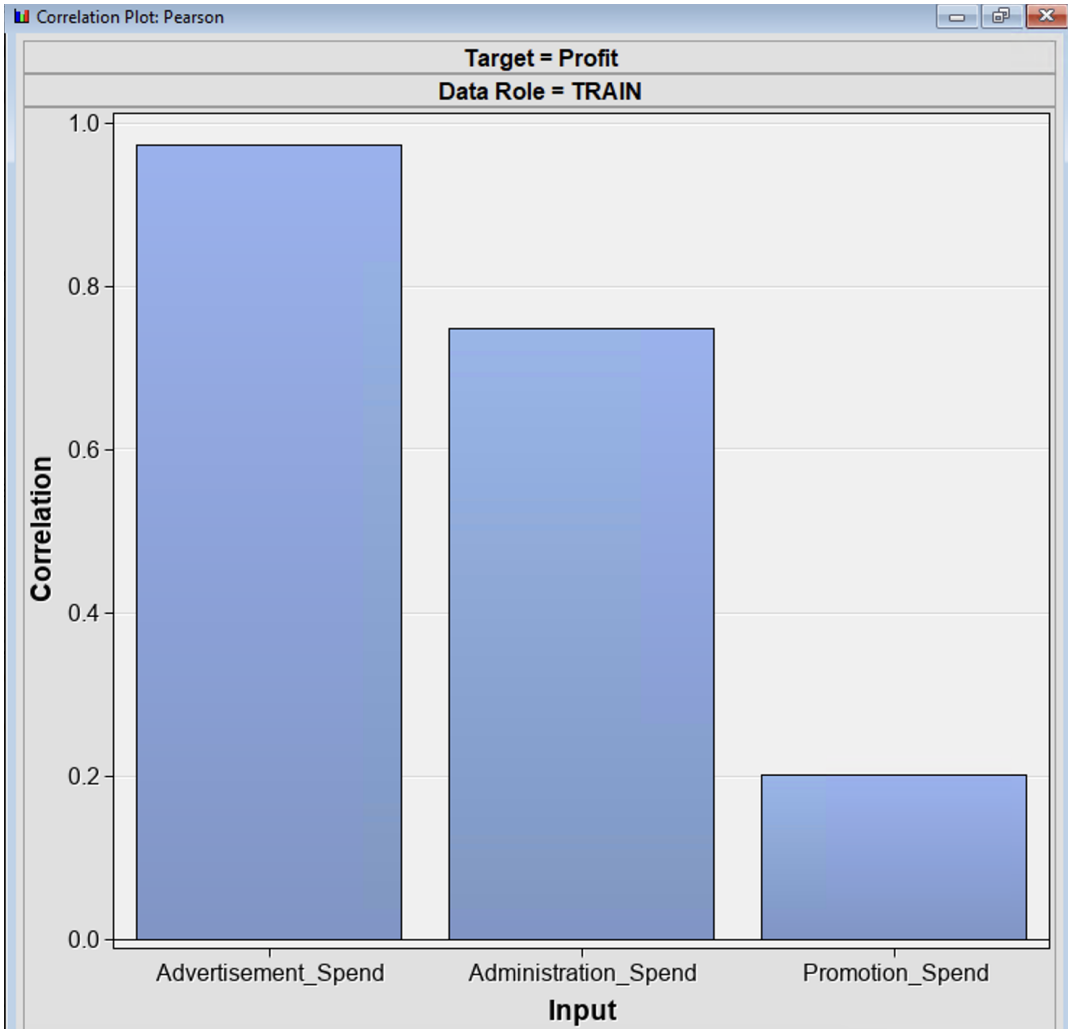


Figure 46

b. Variable Worth Plot

Following that, we examine the “Variable Worth” plot within the outcomes of the StatExplore node. The variable worth plot displays the importance or worth of each variable in predicting the target variable. The height of the bars represents the relative significance of each predictor variable, with taller bars indicating greater importance. In our case, where the target variable is profit, the variable worth plot suggests that 'Advertisement_spend' has the highest predictive value, followed by 'Administration_spend', then 'Promotion_spend', and 'State' being the least influential. This implies that among the factors considered, allocating resources to advertising expenditures appears to have the most substantial impact on profit, followed by administrative spending and promotional activities, while the geographical factor represented by 'State' has comparatively lesser influence.

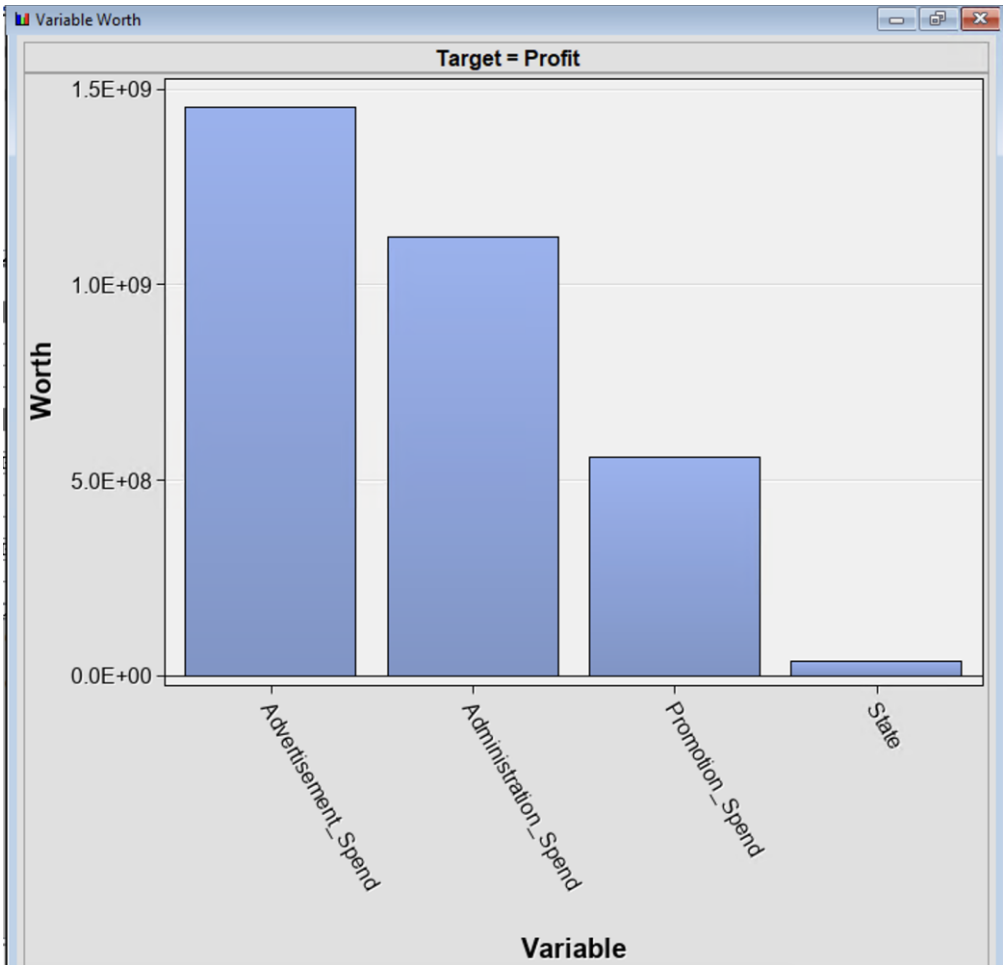


Figure 47

Step 3: Partition Data

We add the 'Data Partition' node to the diagram and link it to the 'StatExplore' node. In the properties of the Data Partition node, we specify that 55% of the data will be allocated for training and 45% for validation. Subsequently, we execute the node.

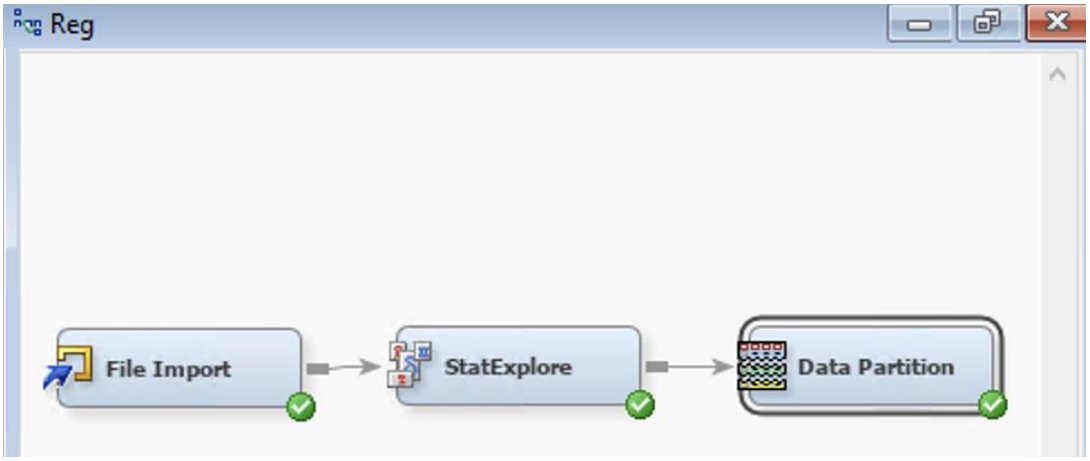


Figure 48

Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS1.Stat_TRAIN	50
TRAIN	EMWS1.Part_TRAIN	28
VALIDATE	EMWS1.Part_VALIDATE	22

Figure 49

Step 4: Setup Linear Regression

We include the 'Regression' node from the 'Model' tab in our diagram and link it to the data partition node.

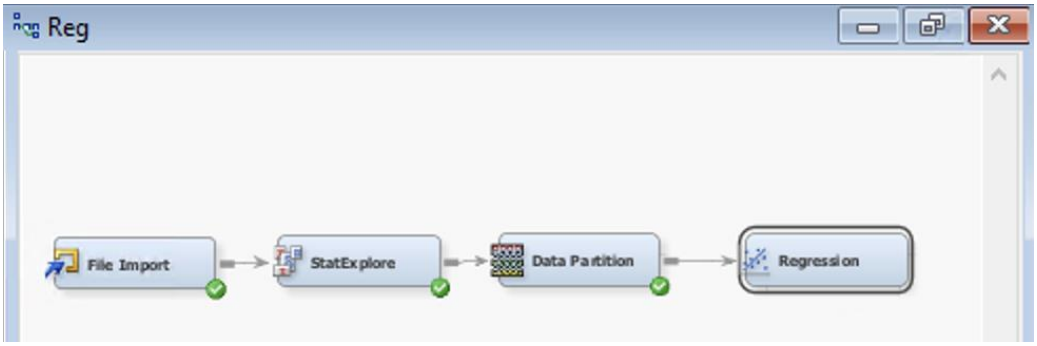


Figure 50

a. **Configure the Regression Node:**

We configure the Regression node by setting the ‘Regression type’ as ‘Linear Regression’ in the properties box. Then, we ensure that the target variable is set to ‘Profit’.

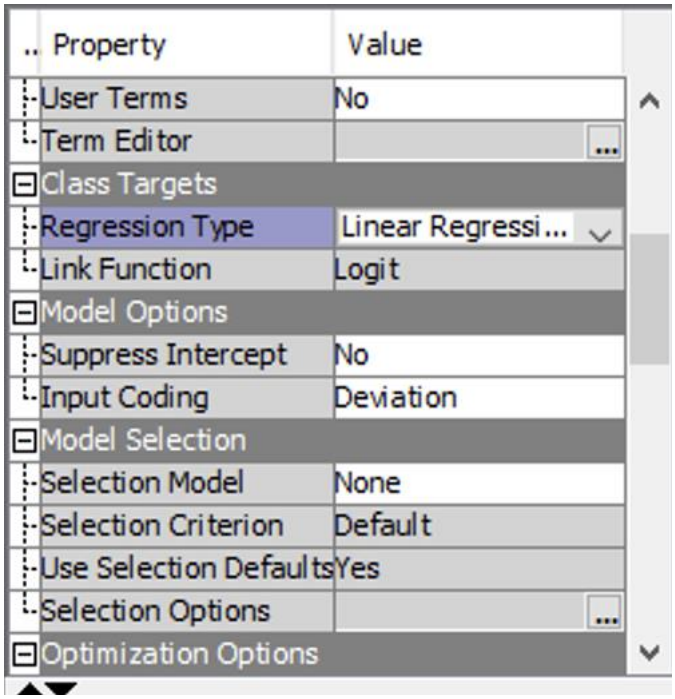


Figure 51

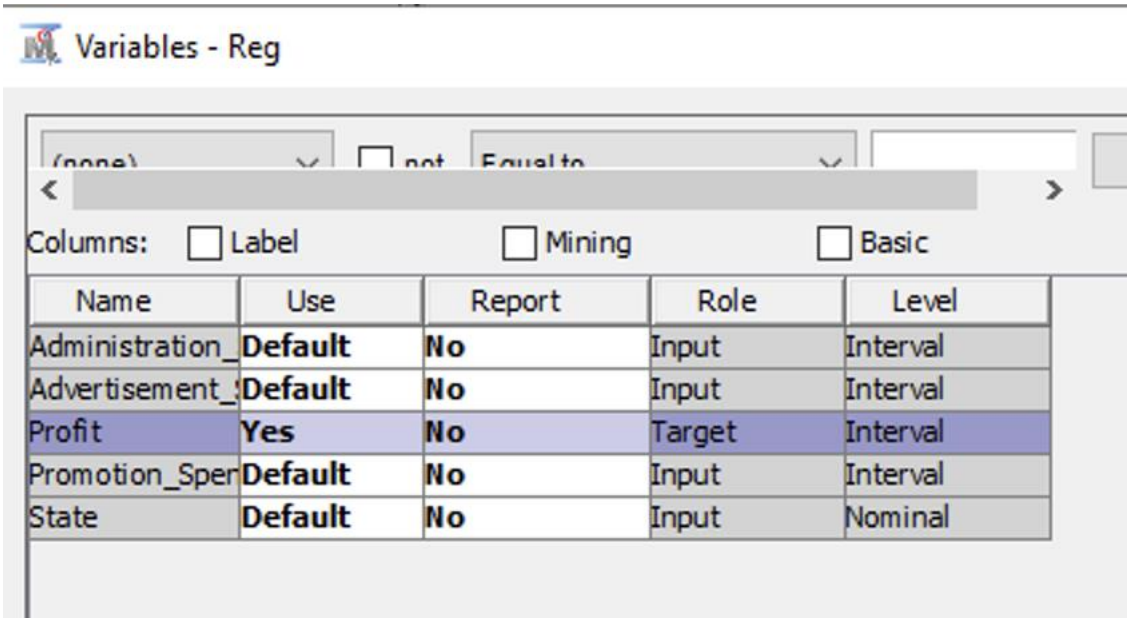


Figure 52

Step 5: Run the Model

We right click the 'Regression' node and click run to execute the model. This will perform the data partitioning and fit the linear regression model using the training data.

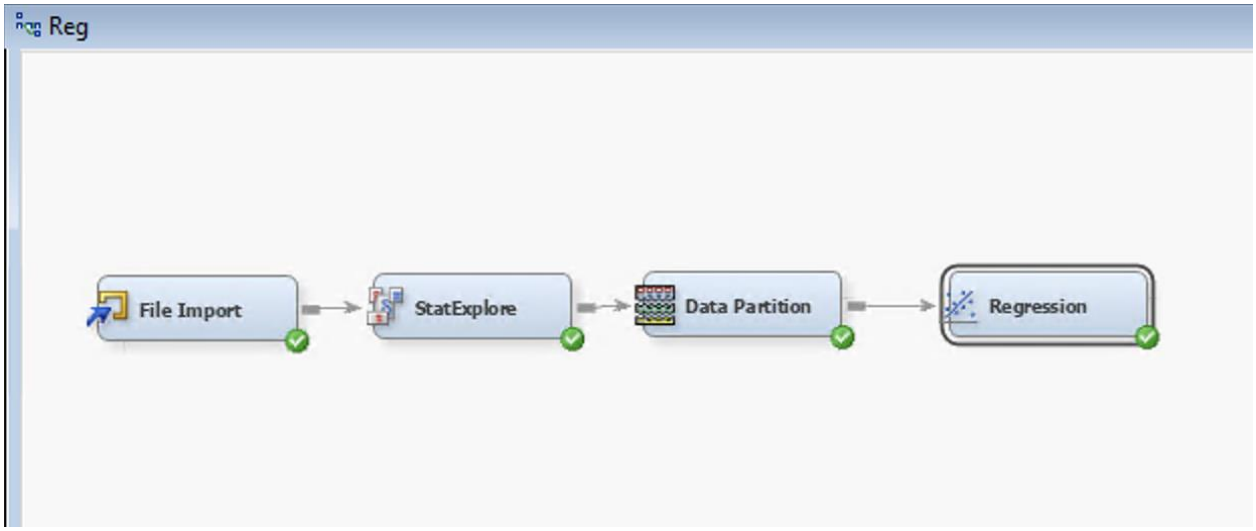


Figure 53

Step 6: Review Model Results

Once the run is complete. we could view various outputs such as model statistics, coefficients, and diagnostic plots.

a. Score Rankings Matrix Profit

In SAS Enterprise Miner, the score rankings matrix profit plot, provides insights into the predictive performance of the model. This plot displays the predicted profit values against the actual profit values for each observation in the dataset. Each point on the plot represents an observation, where the x-axis corresponds to the actual profit values, and the y-axis corresponds to the predicted profit values generated by the regression model. The plot helps evaluate the accuracy of the model's predictions by examining how closely the predicted values align with the actual values.

As observed in both the training and validation plots, the predicted line closely aligns with the target line, suggesting that the model fits exceptionally well.

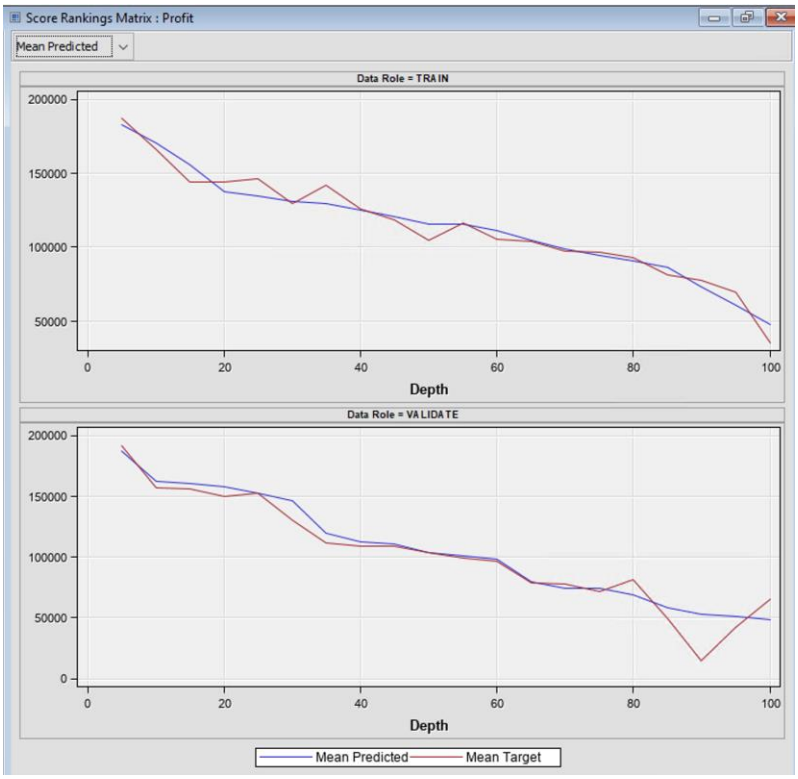


Figure 54

b. Fit Statistics

Upon examining the fit statistics section of the regression node results, we observe that the ASE (Average Squared Error) of our model's validation set is minimal, indicating highly accurate predictions.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Profit		_AIC_	Akaike's Information Criterion	512.3357	.	.
Profit		_ASE_	Average Squared Error	57605687	1.1841E8	.
Profit		_AVERR_	Average Error Function	57605687	1.1841E8	.
Profit		_DFE_	Degrees of Freedom for Error	22	.	.
Profit		_DFM_	Model Degrees of Freedom	6	.	.
Profit		_DFT_	Total Degrees of Freedom	28	.	.
Profit		_DIV_	Divisor for ASE	28	22	.
Profit		_ERR_	Error Function	1.613E9	2.605E9	.
Profit		_FPE_	Final Prediction Error	89026971	.	.
Profit		_MAX_	Maximum Absolute Error	17597.96	37988.59	.
Profit		_MSE_	Mean Square Error	73316329	1.1841E8	.
Profit		_NOBS_	Sum of Frequencies	28	22	.
Profit		_NW_	Number of Estimate Weights	6	.	.
Profit		_RASE_	Root Average Sum of Squares	7589.841	10881.54	.
Profit		_RFPE_	Root Final Prediction Error	9435.411	.	.
Profit		_RMSE_	Root Mean Squared Error	8562.496	10881.54	.
Profit		_SBC_	Schwarz's Bayesian Criterion	520.3289	.	.
Profit		_SSE_	Sum of Squared Errors	1.613E9	2.605E9	.
Profit		_SUMW_	Sum of Case Weights Times Freq	28	22	.

Figure 55

Based on the output of our model, we obtain an adjusted R-Squared value of 94.02%. This indicates that the model explains approximately 94.02% of the variance in the target variable, suggesting a strong relationship between the predictors and the target.

59				
60		Model Fit Statistics		
61				
62	R-Square	0.9512	Adj R-Sq	0.9402
63	AIC	512.3357	BIC	517.4597
64	SBC	520.3289	C(p)	6.0000
65				

Figure 56

c. Residual Analysis

The residual analysis box plot in the SAS EM regression node results offers a visual representation of the distribution of residuals. In this plot, the box represents the interquartile range (IQR) of the residuals, with the median line inside the box. Whiskers extend to the minimum and maximum values within a specified range. A well-balanced and symmetrical box plot with no outliers suggests that the residuals are evenly distributed and follow a normal pattern, indicating the model's adequacy in capturing the variability in the data. In our case, the residual analysis box plot of our model demonstrates such characteristics, affirming the effectiveness of our model in handling the dataset.

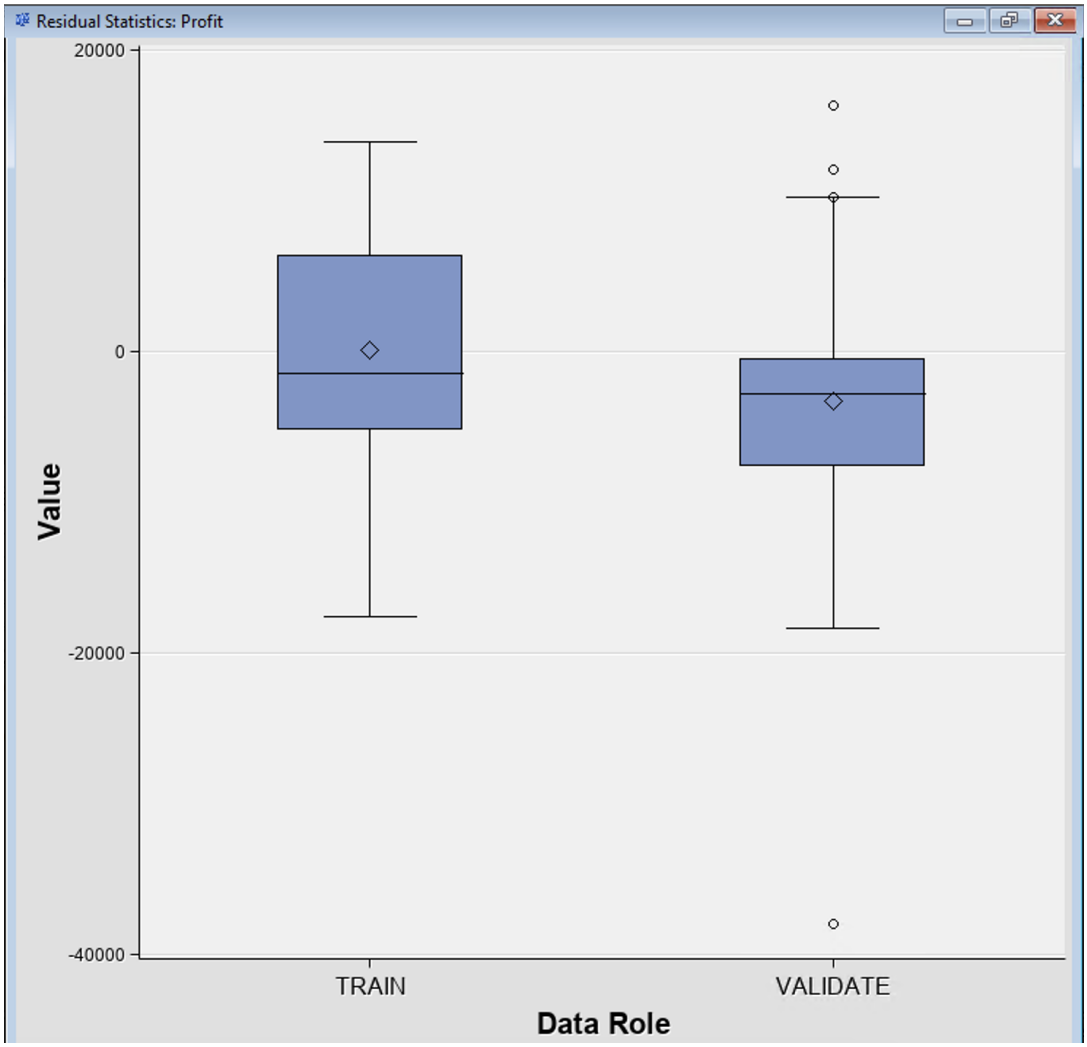


Figure 57

Managerial Implications

1. Utilize Predictive Analytics: Leveraging the regression model as a predictive tool, the management can forecast future profitability based on varying levels of advertising, administration, and promotion spending. This can inform strategic decision-making and resource allocation to optimize profits.
2. Allocate More Resources to Advertising: Given that higher advertising expenditure is strongly correlated with increased profitability, we recommend that the management should consider allocating more resources towards advertising campaigns. Investing in targeted advertising strategies will help attract more customers and boost sales.
3. Optimize Administration Spending: While administration spending also shows a significant correlation with profit, it is slightly less influential than advertising spend.

The management should review administrative expenses to ensure efficiency and identify areas where cost-saving measures can be implemented without compromising operations.

4. Re-evaluate Promotion Strategies: Although promotion spend exhibits the weakest correlation with profit, it still contributes to overall profitability. The management should assess the effectiveness of current promotional activities and consider adjusting strategies to maximize returns on investment.
5. Continuous Monitoring and Adjustment: Profitability is dynamic and influenced by various internal and external factors. The management should continuously monitor key performance indicators, revisit the regression model periodically, and adjust strategies as needed to adapt to changing market conditions and consumer preferences.

Final Diagram

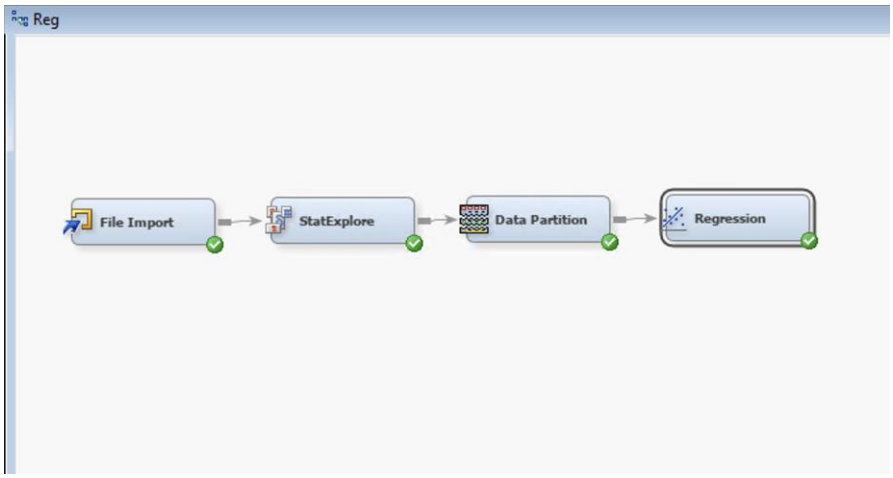


Figure 58

Analysis 4 Conclusion

In summary, our linear regression analysis of the XYZ Supermarket dataset revealed crucial insights into profit drivers. Key findings highlight the significant impact of advertisement spend, followed by administration and promotion spend. Leveraging predictive analytics, management can forecast profitability and allocate resources effectively. Recommendations include increasing advertising investments, optimizing administration costs, and refining promotion strategies. Continuous monitoring and adaptation are essential for maintaining profitability amid evolving market conditions.

Overall, our analysis provides actionable insights for driving growth and maximizing profitability at XYZ supermarket.

Analysis 5(a): Exploratory Data Analysis

(Supermarket_CustomerMembers)

Dataset: We performed our first EDA on the Supermarket_CustomerMembers.csv data.

Objective: To explore the diversity of the customer pool in terms of their gender, age, income and spending score and identify any high level correlation.

Procedure

We observed that the **Supermarket_CustomerMembers.csv** data has concise information about a customer, which when visualized can share better insights about the customers of the supermarket. The below figure is sample clipping of the data. We used the Graph Explore feature of SAS EM to visualize the data from several dimensions.

	A	B	C	D	E
1	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
2	1	Male	19	15	39
3	2	Male	21	15	81
4	3	Female	20	16	6
5	4	Female	23	16	77
6	5	Female	31	17	40
7	6	Female	22	17	76
8	7	Female	35	18	6
9	8	Female	23	18	94
10	9	Male	64	19	3
11	10	Female	30	19	72
12	11	Male	67	19	14
13	12	Female	35	19	99
14	13	Female	58	20	15
15	14	Female	24	20	77
16	15	Male	37	20	13
17	16	Male	22	20	79
18	17	Female	35	21	35
19	18	Male	20	21	66
20	19	Male	52	23	29
21	20	Female	35	23	98
22	21	Male	35	24	35
23	22	Male	25	24	73
24	23	Female	46	25	5
25	24	Male	31	25	73

Figure 59

Step 1: Import the file & edit variables.

To import the excel file, we performed the same steps performed in the previous analysis and entered the path to the location where the file **Supermarket_CustomerMembers.csv** is located.

Rename the File Import node as **Supermarket_CustomerMembers**. We performed the edit variable step to set the below roles as shown in the diagram to the columns of the dataset. By this step the file import is complete

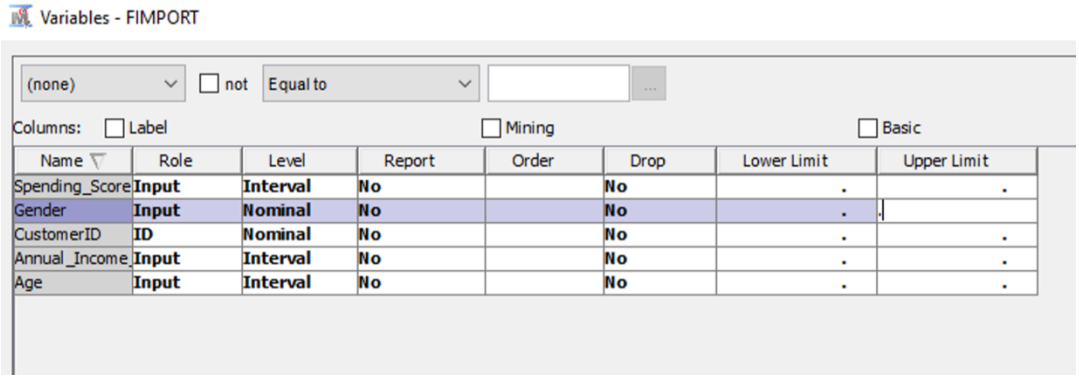


Figure 60

Step 2: Add the Graph Explore node

a. Click the Explore tab and find Graph Explore node

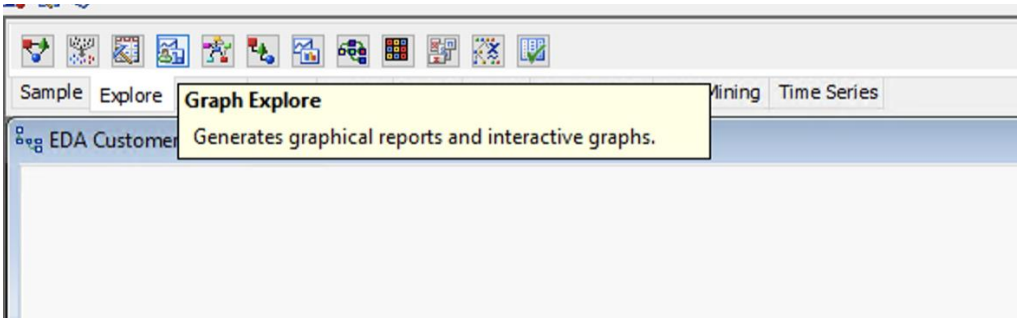


Figure 61

b. Drag the Association node to the diagram and then connect the File Import node & Graph Explore node to each other

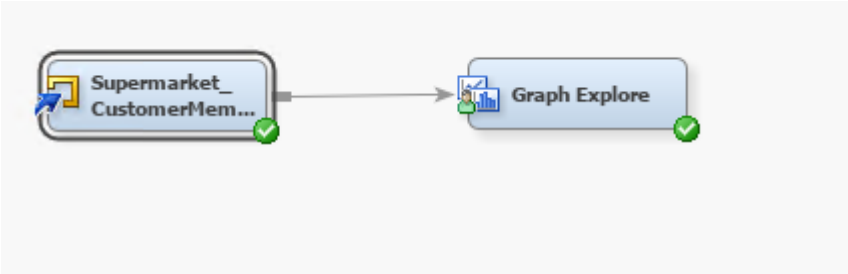


Figure 62

Step 3: Run the diagram.

a. Run the diagram & right click on Graph Explore node to see the results.

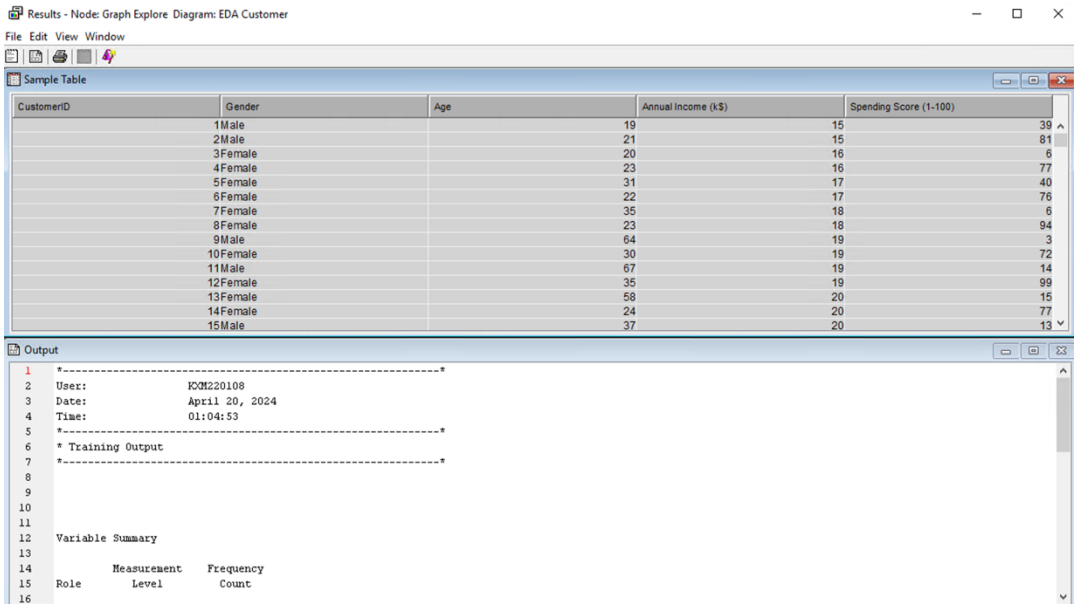


Figure 63

b. Click on View -> Plot

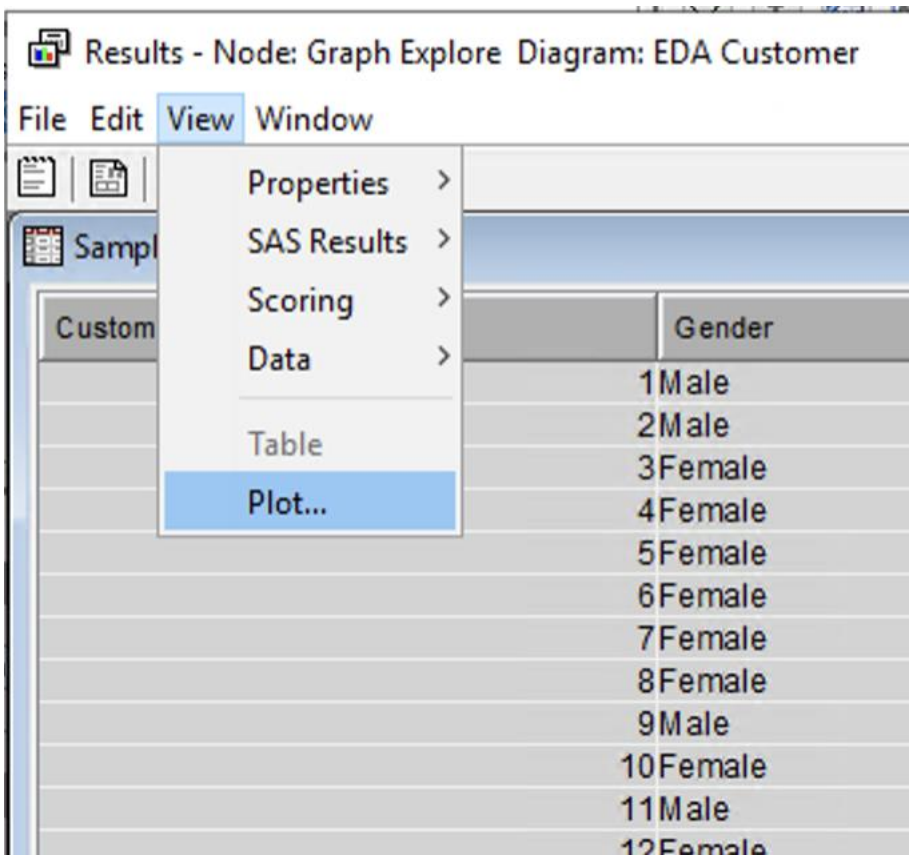


Figure 64

Step 4: Explore the Gender distribution

a. Exploring the Gender distribution using pie chart. Select the chart type as Pie and click Next

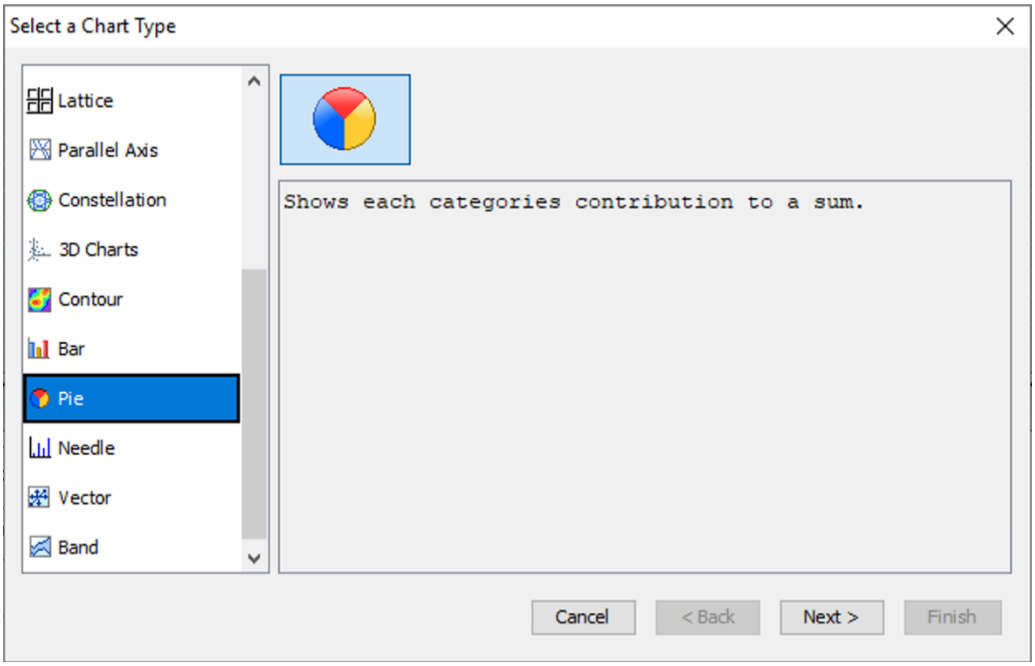


Figure 65

b. Set the chart roles by setting a category variable. We set gender as category. Click Next.

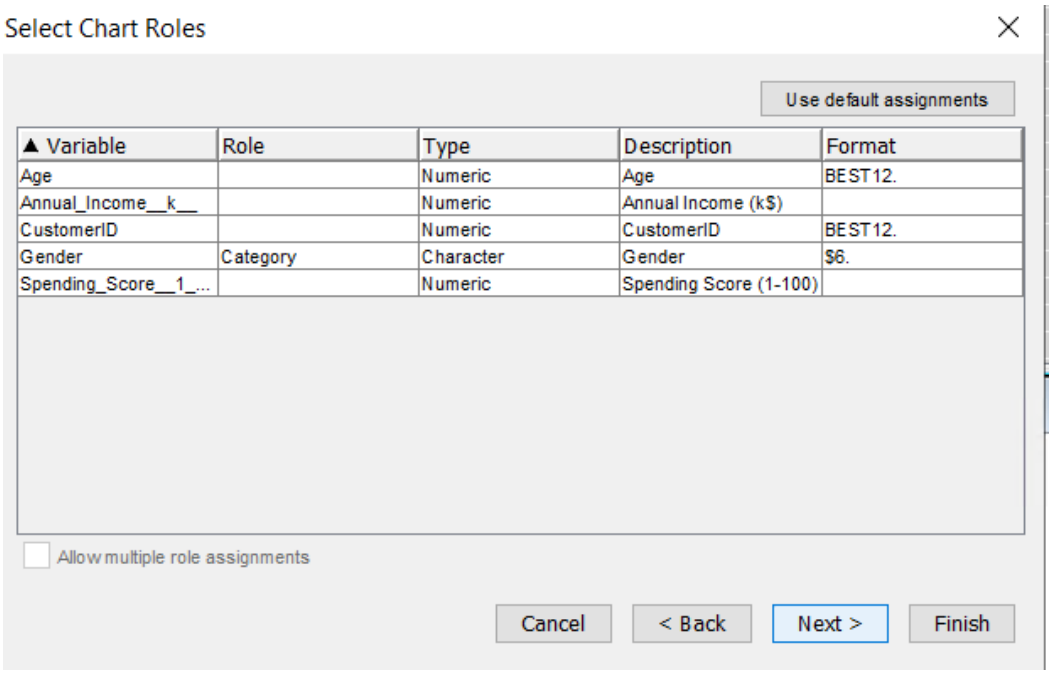


Figure 66

c. There is no where clause, click Next

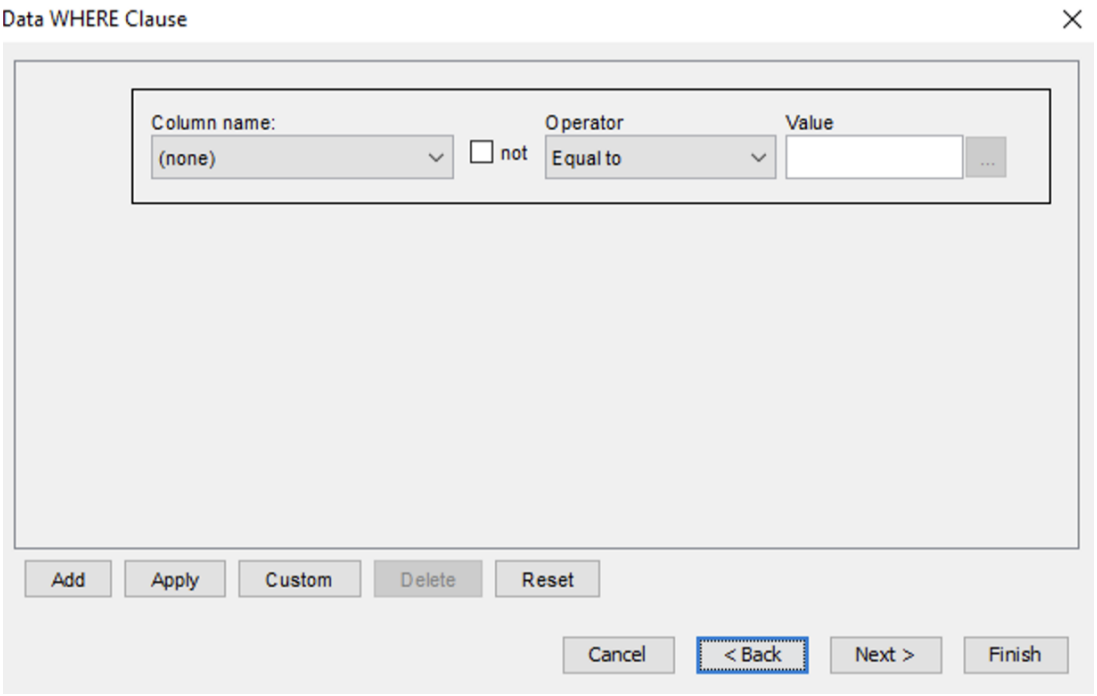


Figure 67

d. Set the chart title as “Gender distribution”. Click Finish

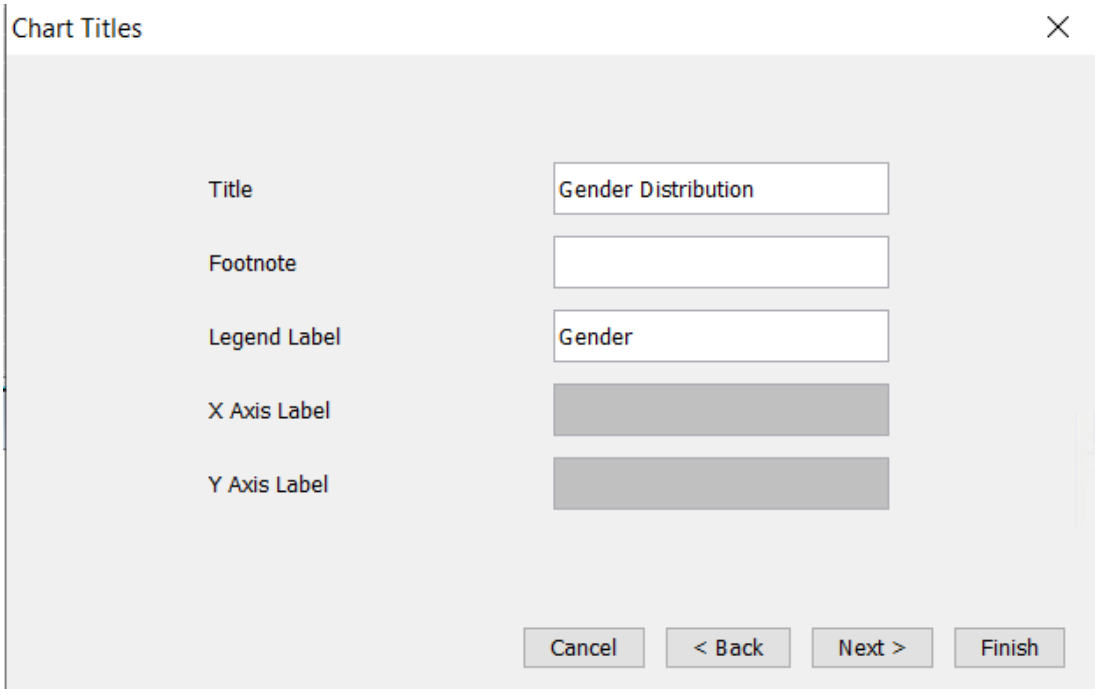


Figure 68

e. Set the graph properties to show Category Name & Percentage

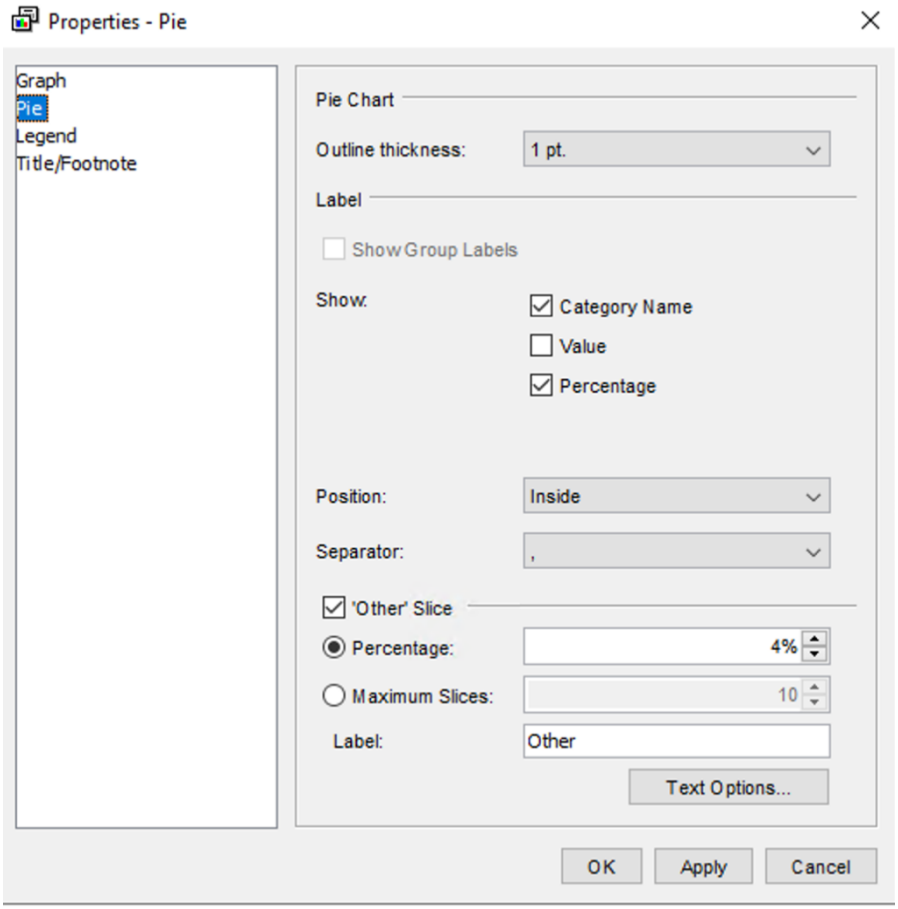


Figure 69

f. The Below graph is developed by SAS EM

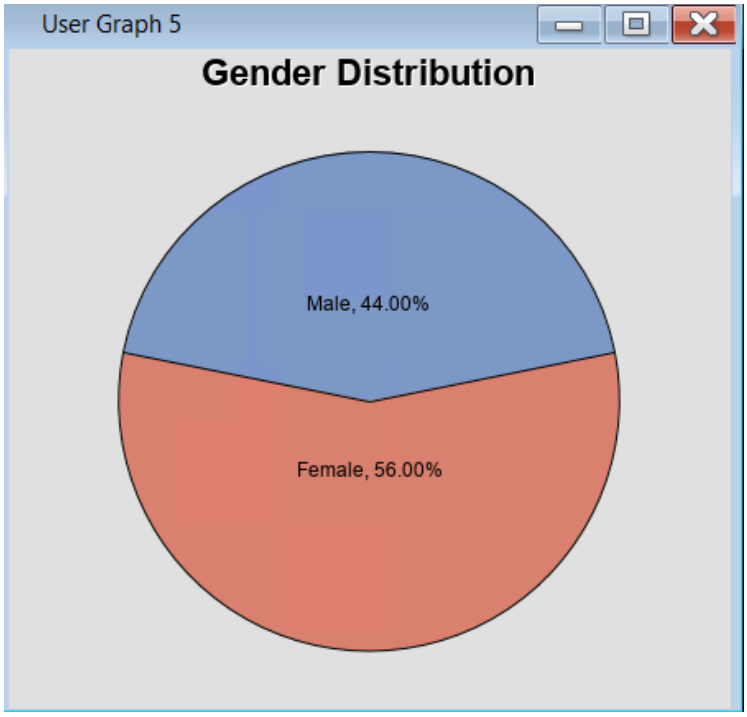


Figure 70

Interpretation & Managerial Implications

It was observed that 44% of the supermarket customers are Males and 56% of the supermarket customers are Females. This gender composition can help the supermarket grow their business in the following ways:

Targeted Market sales: This ratio can guide the supermarket to do targeted sales as some products appeal more to one specific gender than the other.

Store Layout and Product Placement: The layout of the store can be planned in such a way that would attract that particular group. For example, stitching the results of market basket analysis to gender and improvising the product placement in a store.

Step 5: Explore the Age distribution.

a. Exploring the Age distribution using pie chart. Select the chart type as Bar and click Next.

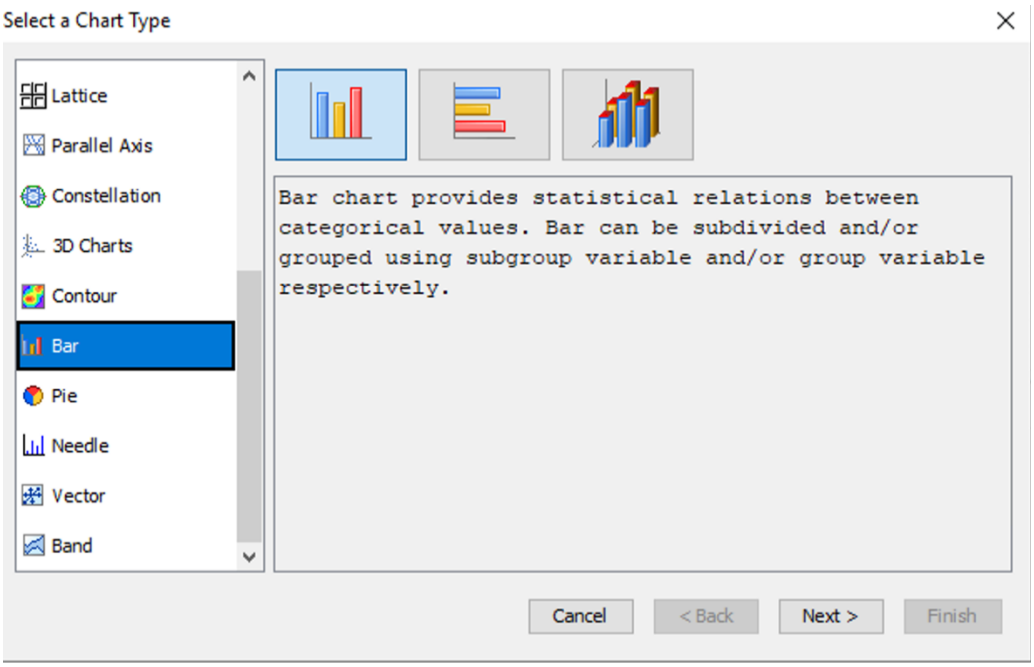


Figure 71

b. Set the chart roles by setting a category variable. We set Age as category. Click Next.

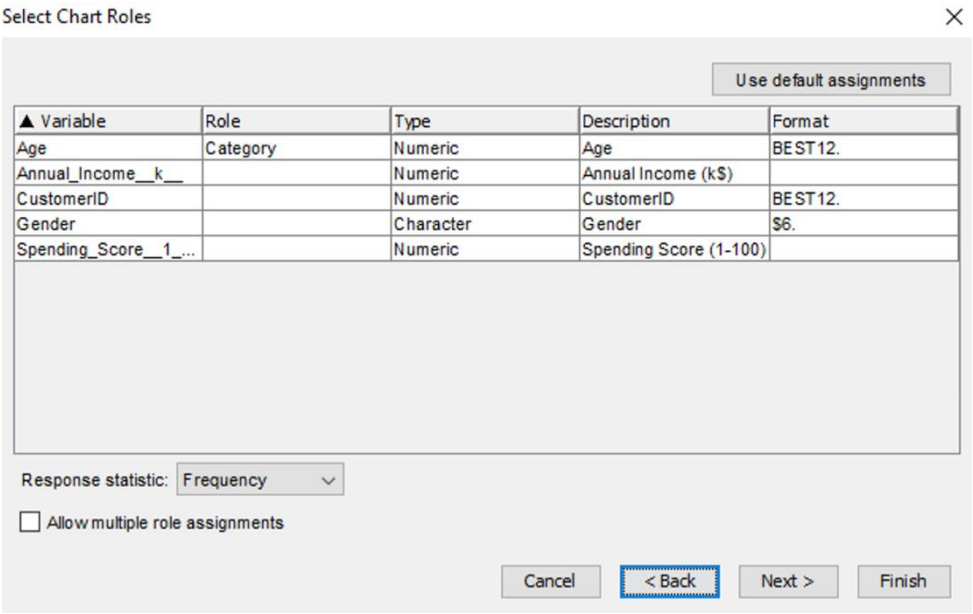


Figure 72

c. There is no where clause, click Next.

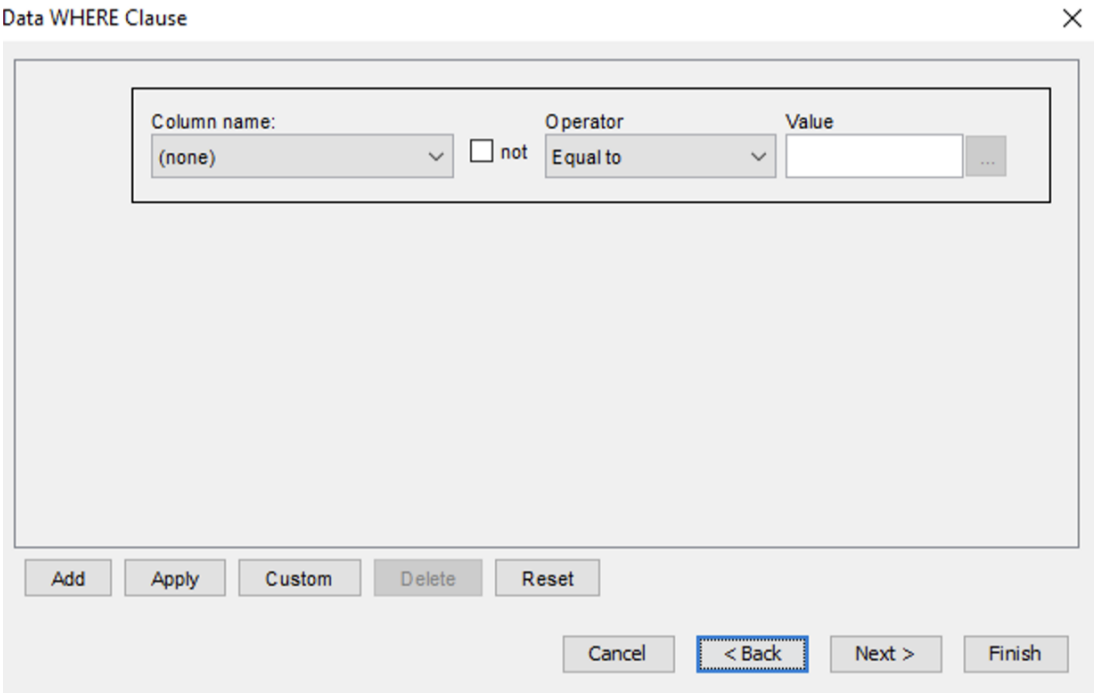


Figure 73

d. Set the chart title as “Age distribution”. X Axis Label as Age and Y Axis Label as Count. Click Finish

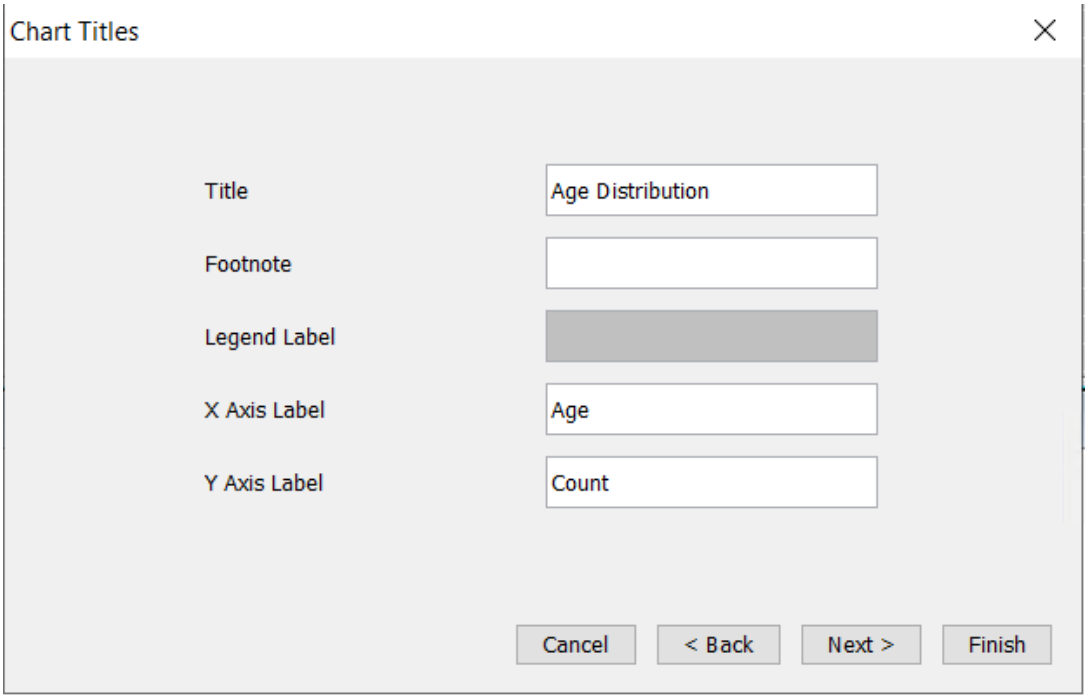


Figure 74

e. The Below graph is developed by SAS EM

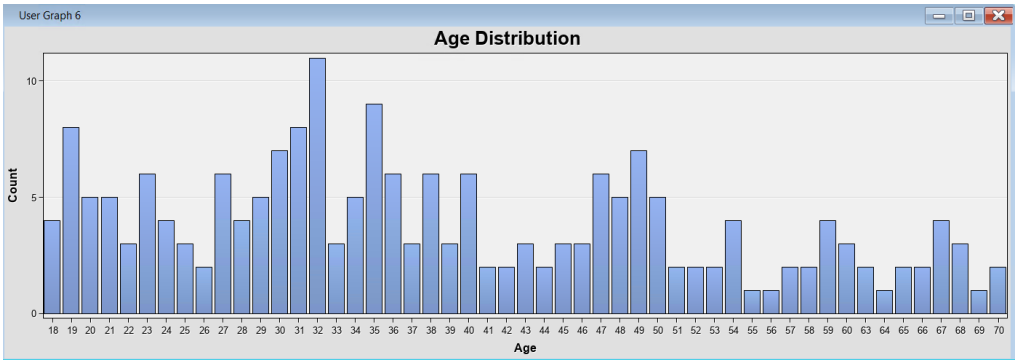


Figure 75

Interpretation & Managerial Implications

The age distribution chart shows the frequency of customers with respect to their age. The following are the interpretations:

Younger Customer Base: There is a higher concentration of customers in the younger age ranges, specifically from around 25 to 35 years old. This group seems to be the most frequent age demographic for this supermarket.

Variability in Middle Age Ranges: With customers in their late 30s to early 50s, the frequency of the customers is very varying.

In summary, the supermarket is very popular amongst the younger demographics.,this suggests a need for targeted marketing and product selection to cater them. Also, they need to investigate into the lesser customers in the older age & middle age brackets and attract them through tailored marketing strategies.

Step 6: Explore the relation between Age and Annual Income

a. Select the chart type as Bar and click Next.

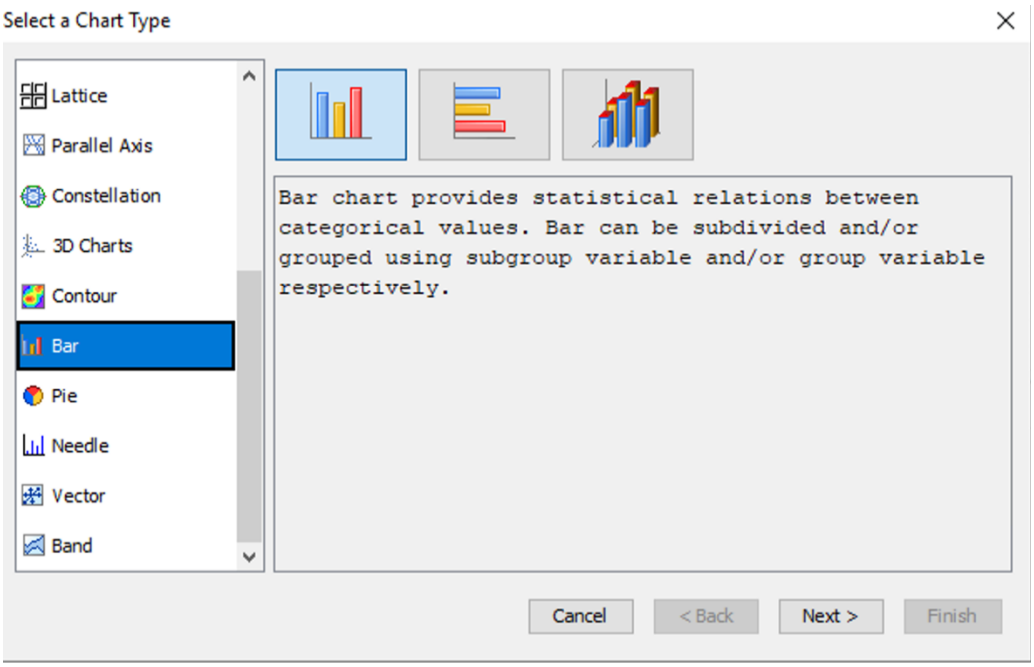


Figure 76

b. Set the chart roles by setting a category variable & the Frequency variable. We set Age as category & Annual_Income__k__ as Frequency. Click Next.

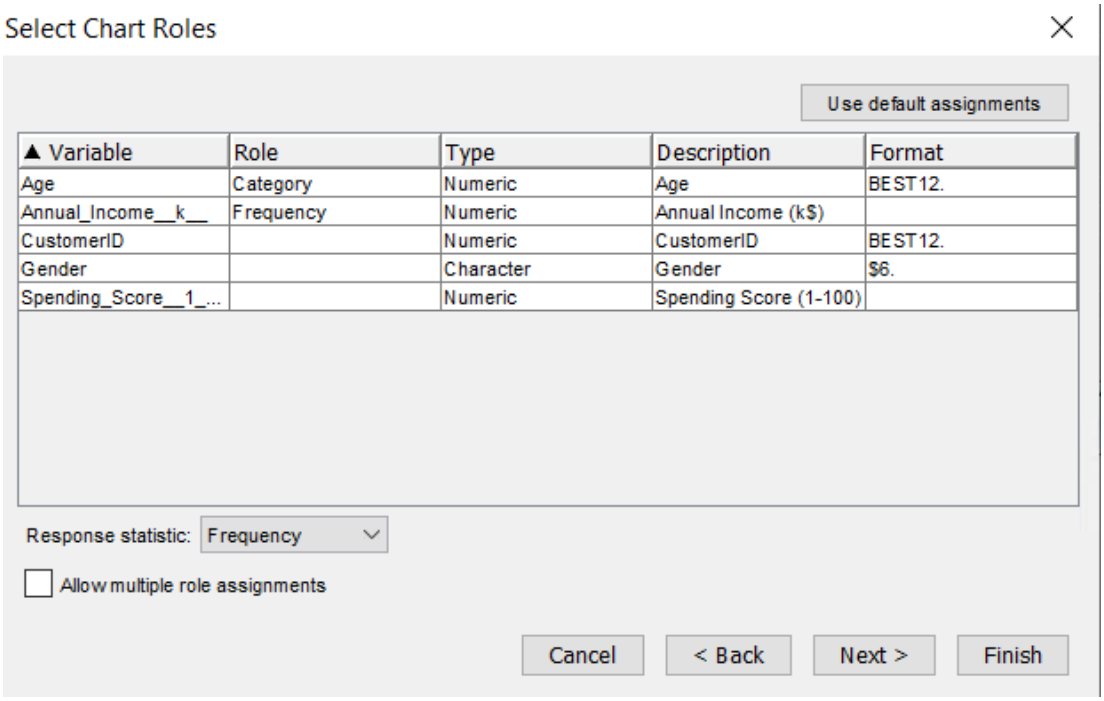


Figure 77

c. There is no where clause, click Next

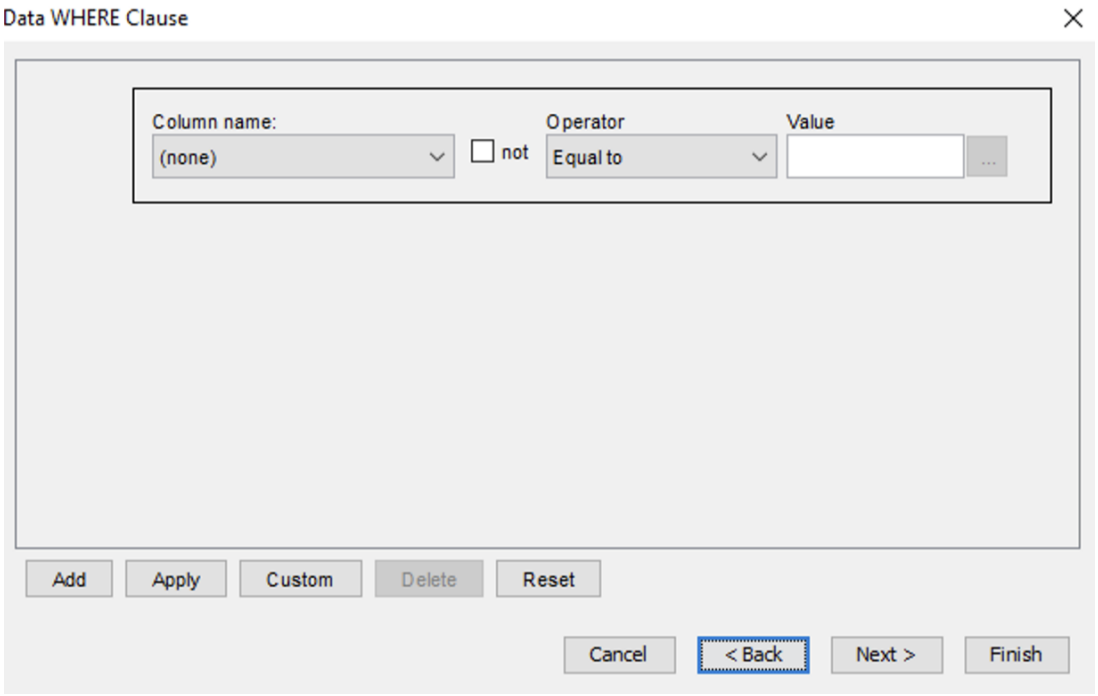


Figure 78

d. Set the chart title as “Age vs Annual Income Plot”. X Axis Label as Age and Y Axis Label as Annual income Sum (in K). Click Finish

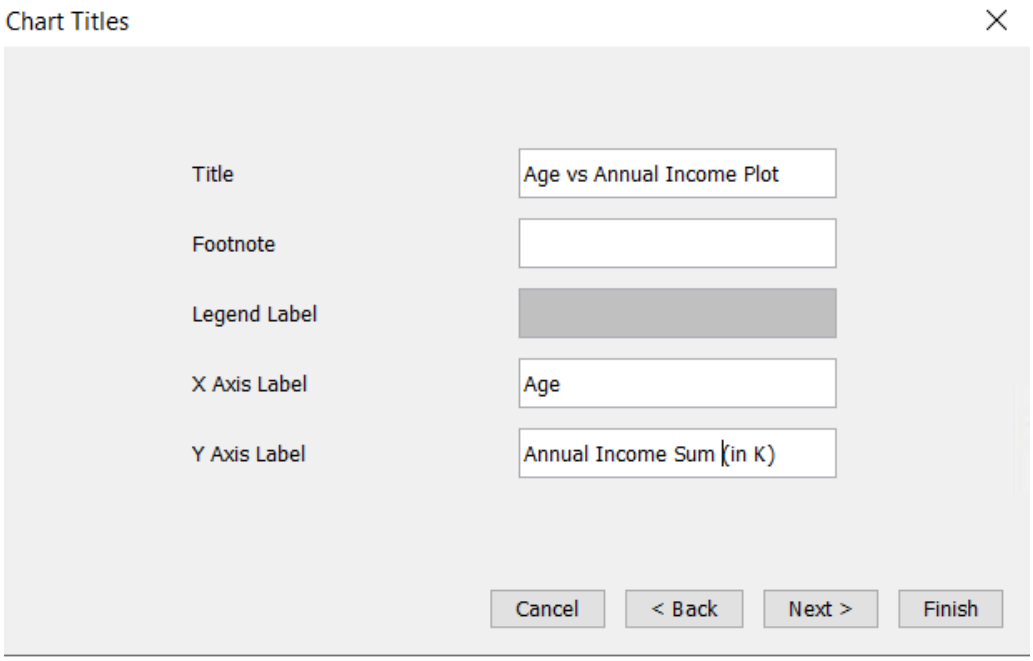


Figure 79

e. The Below graph is developed by SAS EM

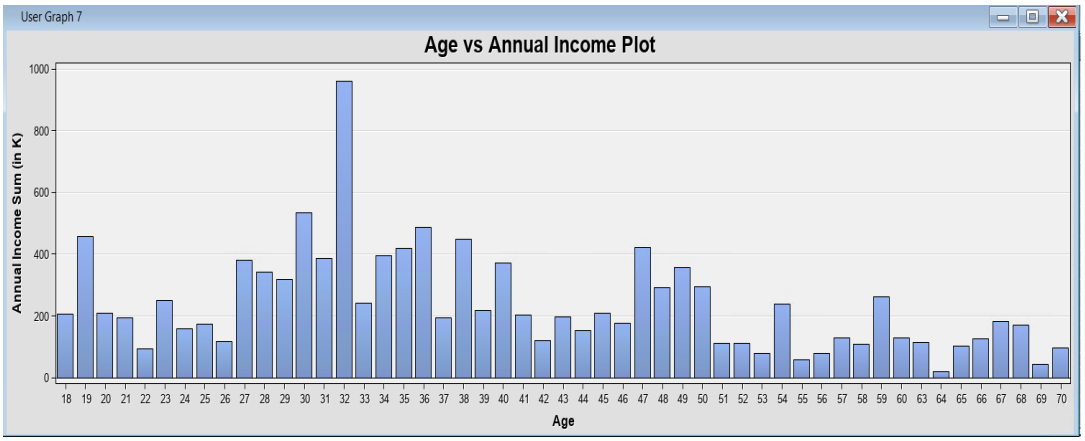


Figure 80

Interpretation & Managerial Implications

The graph presents the relationship between customers' ages and their annual income. Below are the major interpretation of the chart:

Lower Income groups: Customers in the ages between 18 and 30 typically a group often that includes students and individuals in the early stages of their careers., and hence the lower income. To attract this group & encourage more purchasing they can introduce deals that benefit their purchases.

High Income groups: A significant peak in income for customers in their mid-30s is observed in the graph.

Varied & Declining Incomes: Customers aged between 36-50 and After the mid-50s are exhibiting a variability in income and also considerable lower income respectively.

It is essential to this information to their spending score and take decision regarding targeted marketing, offer bundling & other long term strategic planning that can benefit both the supermarket and the customers.

Step 7: Explore the relation between Age and Spending Score

a. Select the chart type as Bar and click Next

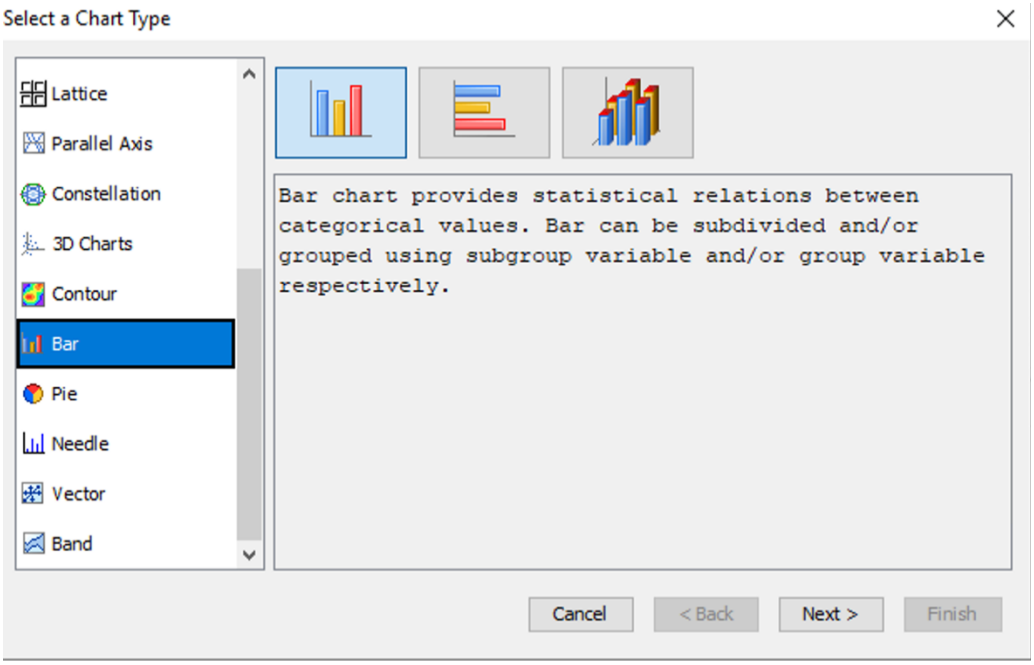


Figure 81

b. Set the chart roles by setting a category variable & the Frequency variable. We set Age as category & Spending_Score as Frequency. Click Next.

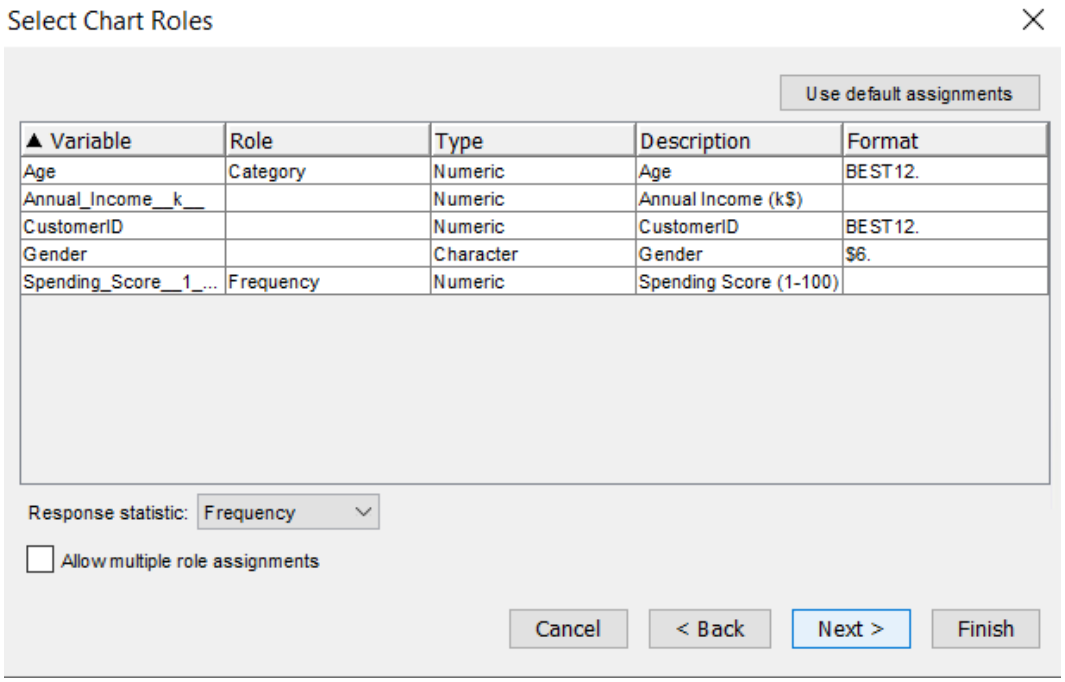


Figure 82

c. There is no where clause, click Next

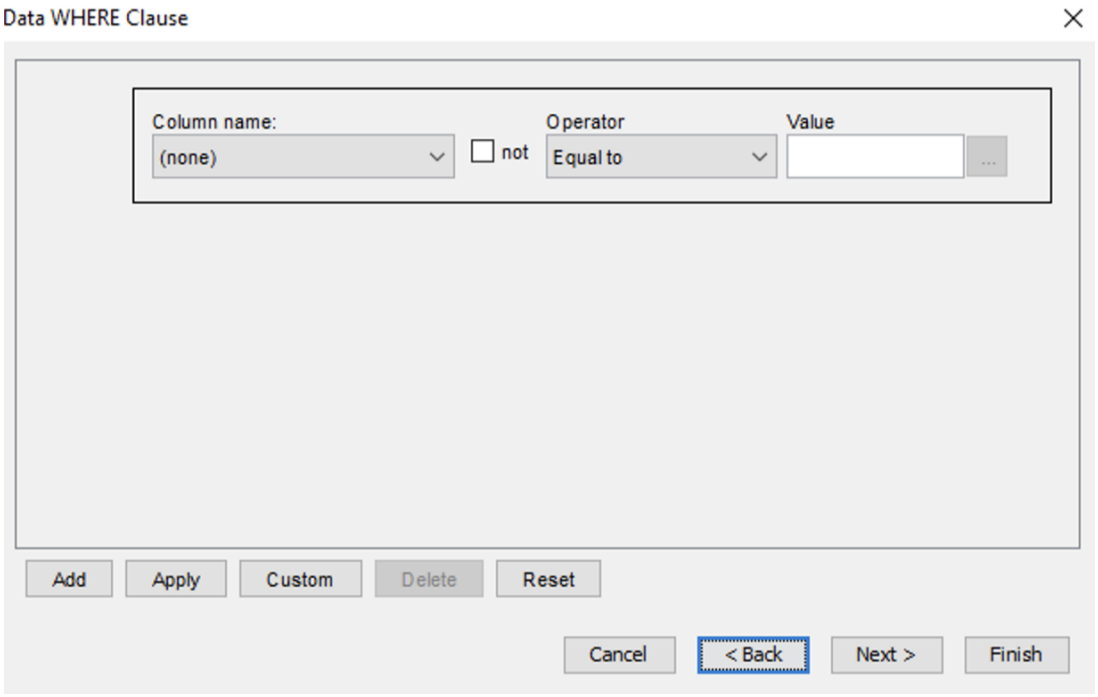


Figure 83

d. Set the chart title as “Age vs Spending Plot”. X Axis Label as Age and Y Axis Label as Sum of Spendings. Click Finish

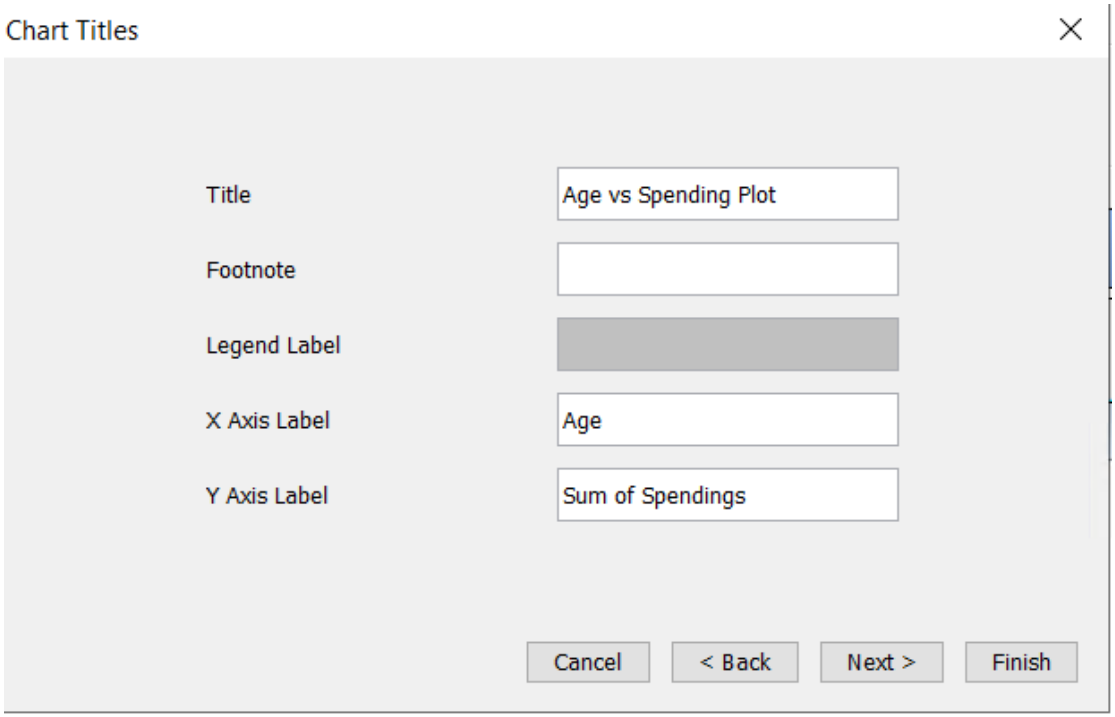


Figure 84

e. The Below graph is developed by SAS EM

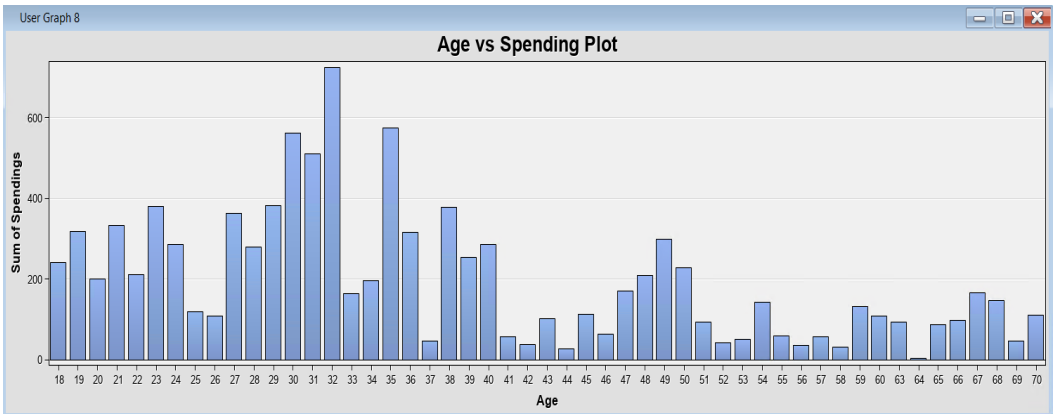


Figure 85

Interpretation & Managerial Implications

The graph presents the relationship between customers' ages and their spending scores in a supermarket. Below are the major interpretation of the chart:

Spending Score Trend: The spending score seems to increase with age until the early to mid-30s, relating to their annual income and also suggesting that, they tend to spend more up to a certain point.

Peak Spending Age: There is a noticeable peak in the spending score for customers in their early to mid-30s. This indicates that this age group is likely the most valuable in terms of spending power within the supermarket.

Declining Spending Score: After this peak, the spending score trends downward as the customer's age increases, indicating that older age groups may spend less on average.

Variability in Middle Age: The spending score among the middle-aged customers shows some variability, suggesting a high & low spending behaviors within this demographic.

The supermarket may consider the following strategies:

Customer Loyalty Programs: Implement loyalty programs that encourage increased spending & offer high value amongst the high-scoring demographics and incentivize repeat business.

Customer Retention: To focus at the age groups with declining spending scores, the supermarket needs to consider retention strategies such as special offers or services that cater to their specific needs and shopping preferences and encourage purchases

---

Analysis 5(b): Exploratory Data Analysis

(50_Supermarket_Branches)

Dataset: We performed our first EDA on the **50_Supermarket_Branches.csv** data.

Objective: To explore the diversity of the the branches spread across the there states, and also analyse in terms of the spendings on Advertisements and Promotions.

Procedure

We observed that the **50_Supermarket_Branches.csv** data has concise information about the money spent on advertisements, promotions, administration & profit for stored located in the there state of New York, California & Florida. The information which

when visualized can share better insights about branches. The below figure is sample clipping of the data. We used the Graph Explore feature of SAS EM to visualize the data from several dimensions.

	A	B	C	D	E
1	Advertisement Spend	Promotion Spend	Administration Spend	State	Profit
2	165349.2	136897.8	471784.1	New York	192261.83
3	162597.7	151377.59	443898.53	California	191792.06
4	153441.51	101145.55	407934.54	Florida	191050.39
5	144372.41	118671.85	383199.62	New York	182901.99
6	142107.34	91391.77	366168.42	Florida	166187.94
7	131876.9	99814.71	362861.36	New York	156991.12
8	134615.46	147198.87	127716.82	California	156122.51
9	130298.13	145530.06	323876.68	Florida	155752.6
10	120542.52	148718.95	311613.29	New York	152211.77
11	123334.88	108679.17	304981.62	California	149759.96
12	101913.08	110594.11	229160.95	Florida	146121.95
13	100671.96	91790.61	249744.55	California	144259.4
14	93863.75	127320.38	249839.44	Florida	141585.52
15	91992.39	135495.07	252664.93	California	134307.35
16	119943.24	156547.42	256512.92	Florida	132602.65
17	114523.61	122616.84	261776.23	New York	129917.04
18	78013.11	121597.55	264346.06	California	126992.93
19	94657.16	145077.58	282574.31	New York	125370.37
20	91749.16	114175.79	294919.57	Florida	124266.9
21	86419.7	153514.11	0	New York	122776.86
22	76253.86	113867.3	298664.47	California	118474.03
23	78389.47	153773.43	299737.29	New York	111313.02
24	73994.56	122782.75	303319.26	Florida	110352.25
25	67532.53	105751.03	304768.73	Florida	108733.99
26	77044.01	99281.34	140574.81	New York	108552.04
27	64664.71	139553.16	137962.62	California	107404.34
28	75328.87	144135.98	134050.07	Florida	105733.54
29	72107.6	127864.55	353183.81	New York	105008.31

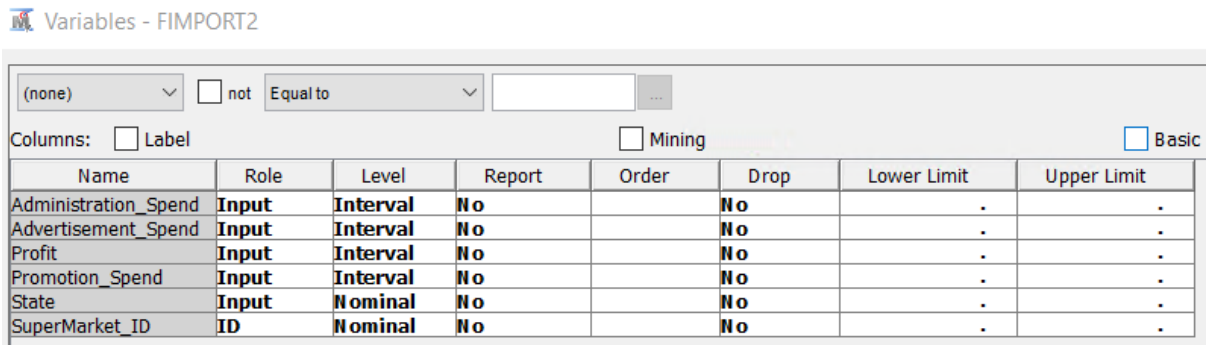
50_SupermarketBranches

Figure 86

Step 1: Import the file & edit variables

To import the excel file, we performed the same steps performed in the previous analysis and enter the path to the location using the import file option, where the file **50_SupermarketBranches.csv** is located.

Rename the File Import node as **50_SupermarketBranches**. We performed the edit variable step to set the below roles as shown in the diagram to the columns of the datset. By this step the file import is complete.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Administration_Spend	Input	Interval	No		No	.	.
Advertisement_Spend	Input	Interval	No		No	.	.
Profit	Input	Interval	No		No	.	.
Promotion_Spend	Input	Interval	No		No	.	.
State	Input	Nominal	No		No	.	.
SuperMarket_ID	ID	Nominal	No		No	.	.

Figure 87

Step 2: Add the Graph Explore node

a. Click the Explore tab and find Graph Explore node

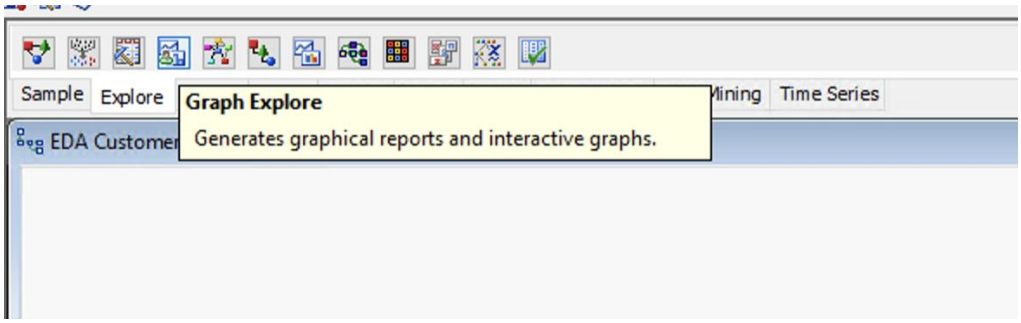


Figure 88

b. Drag the Association node to the diagram and then connect the File Import node & Graph Explore node to each other

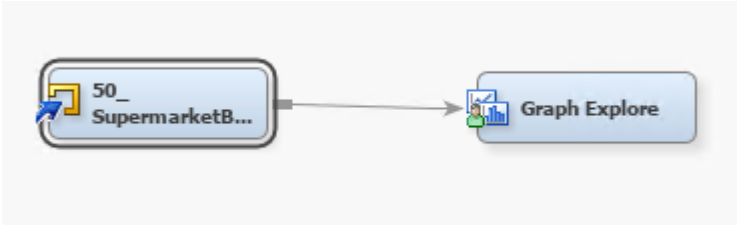


Figure 89

Step 3: Run the diagram

a. Run the diagram & right click on Graph Explore node to see the results.

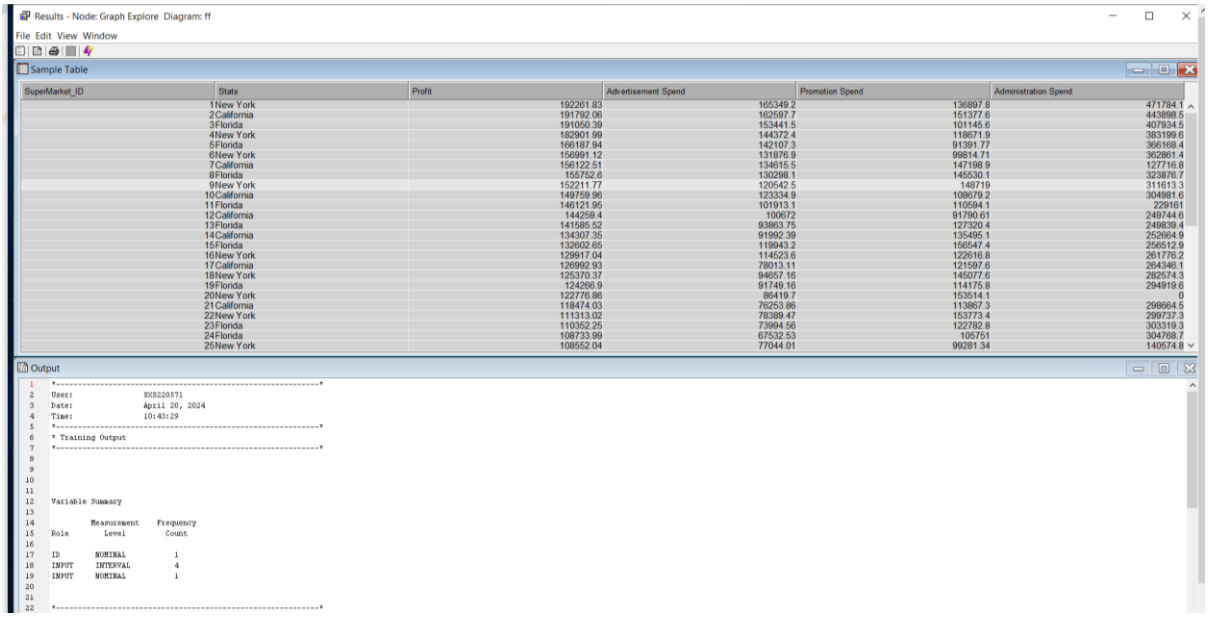


Figure 90

b. Click on View -> Plot

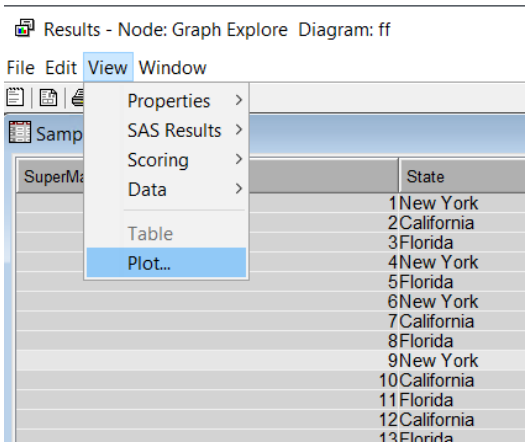


Figure 91

Step 4: Explore the Branch distribution.

a. Exploring the Branch distribution using bar chart. Select the chart type as Bar and click Next

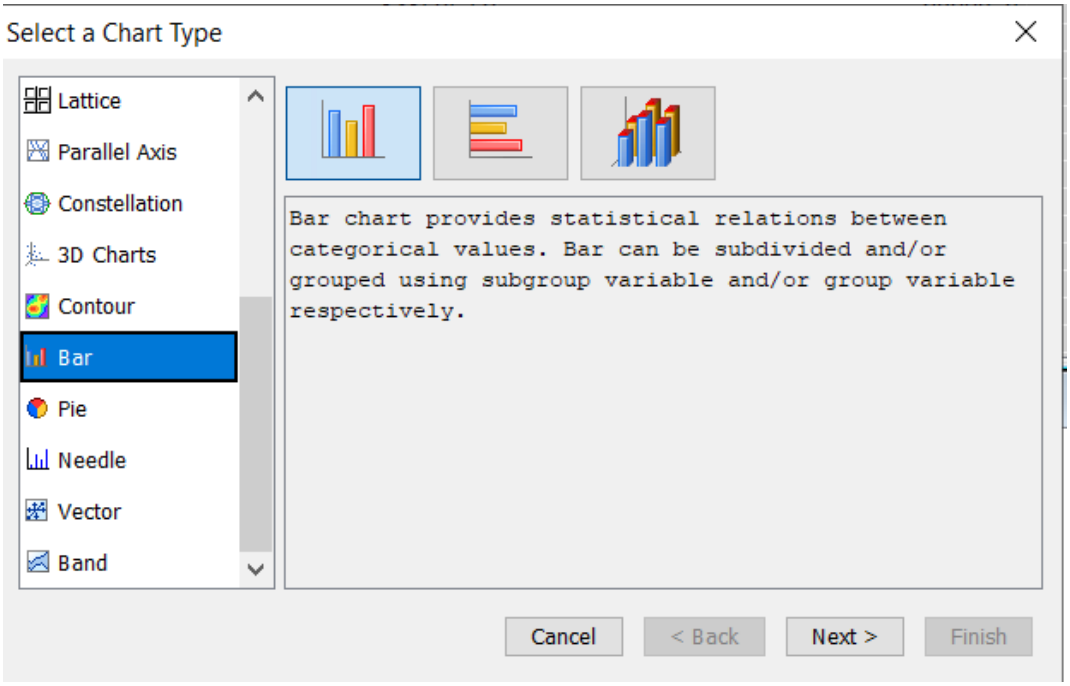


Figure 92

b. Select State as category and click next

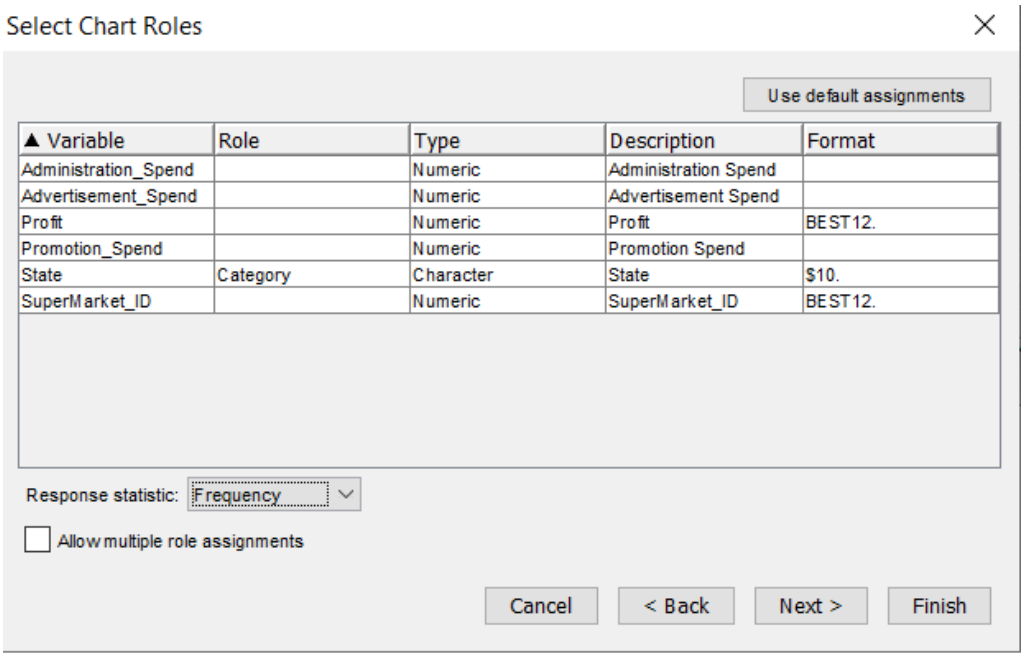


Figure 93

c. There is no where clause, click Next

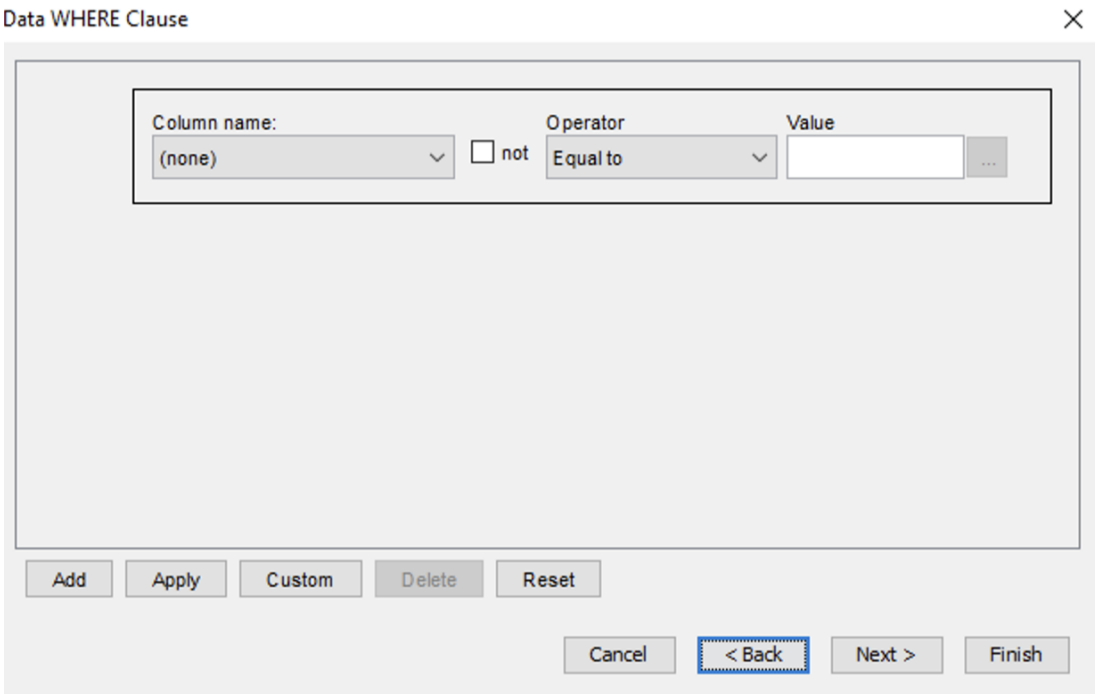


Figure 94

d. Set the chart title as “Count of Branches in each State”. Also set the X Axis Label as State and Y Axis Label as No. Of Branches. Click Finish

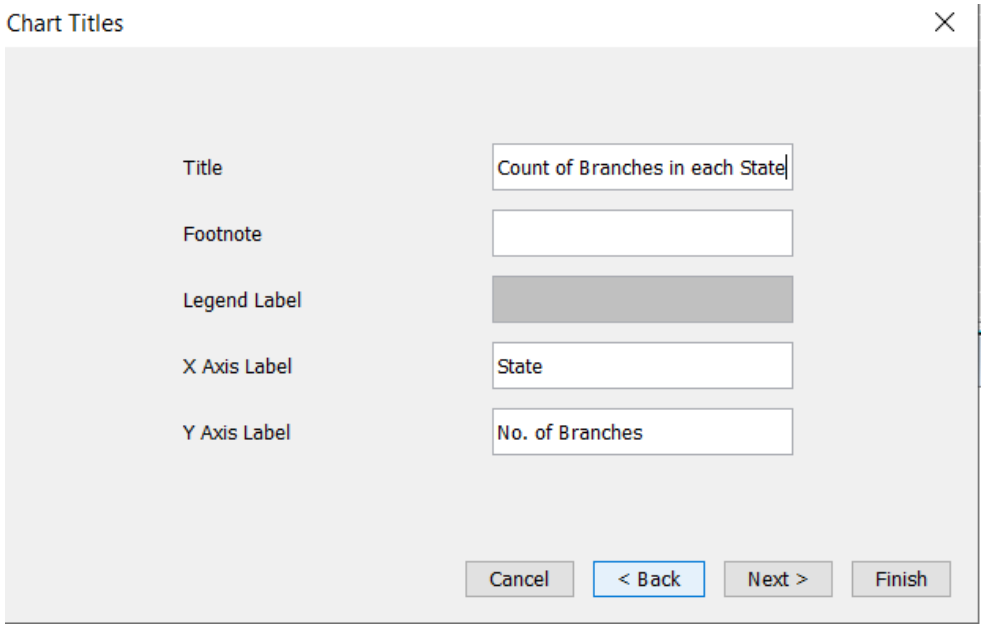


Figure 95

e. The Below graph is developed by SAS EM

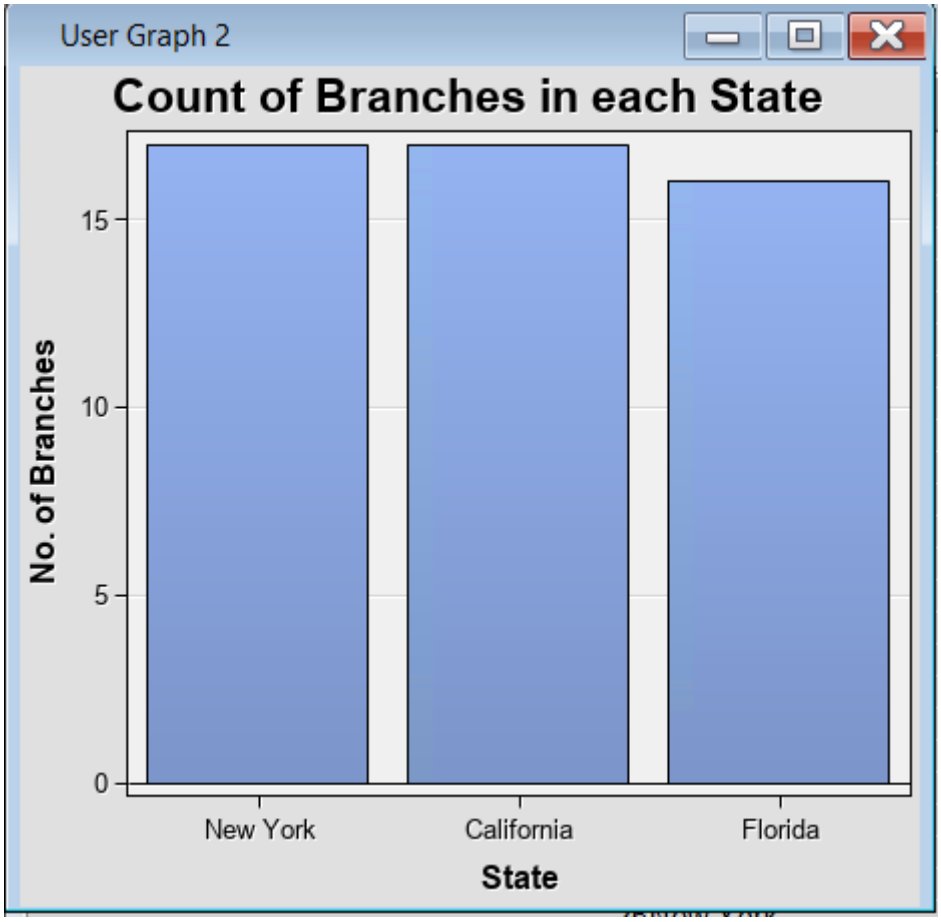


Figure 96

Interpretation & Managerial Implications

The bar graph shows the distribution of supermarket branches across three states: New York, California, and Florida. The following are the findings:

Even Distribution: The supermarket has an almost even distribution of branches across the states, with New York and California each having 17 branches, and Florida closely following with 16. This suggests a strategic decision to cater to large markets with significant potential customer bases.

Data-Driven Decision Making: The supermarket can use sales data from these branches to further refine its product assortment and marketing strategies. For instance, data analytics could reveal which products are popular in certain locations but not in others, leading to more targeted inventory control.

In summary, the graph indicates a well-distributed presence of the supermarket in three populous states, which presents opportunities for localized marketing and community engagement to drive sales and customer loyalty.

Step 5: Explore the Branch wise money spent on Advertisement

a. Exploring the Branch wise money spent on Advertisement distribution using bar chart. Select the chart type as Bar and click Next

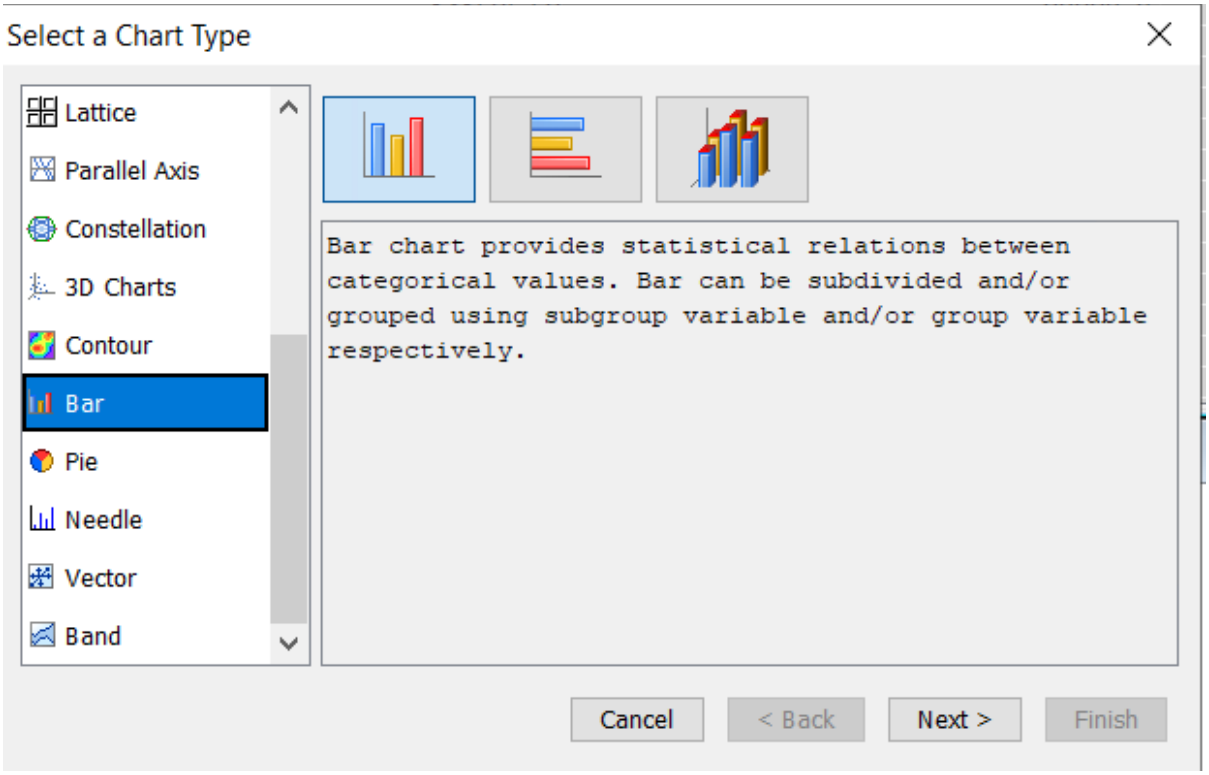


Figure 97

b. Select State as category and Advertisement_Spend as frequency. click Next.

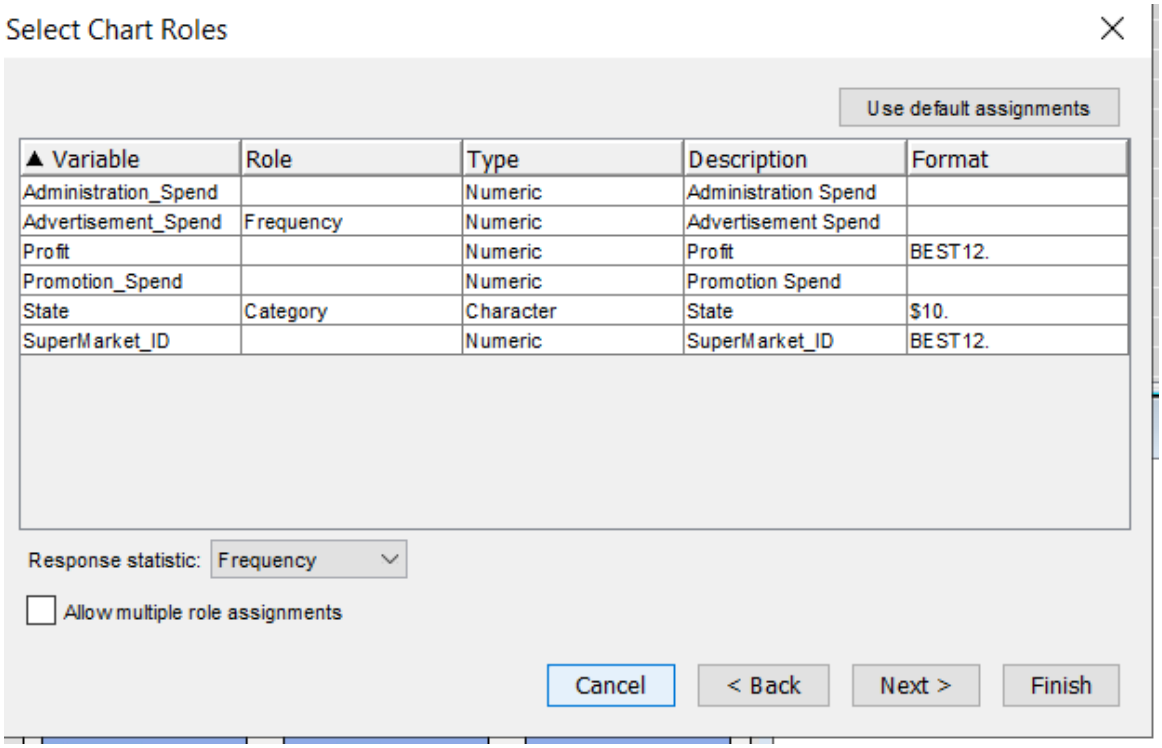


Figure 98

c. There is no where clause, click Next

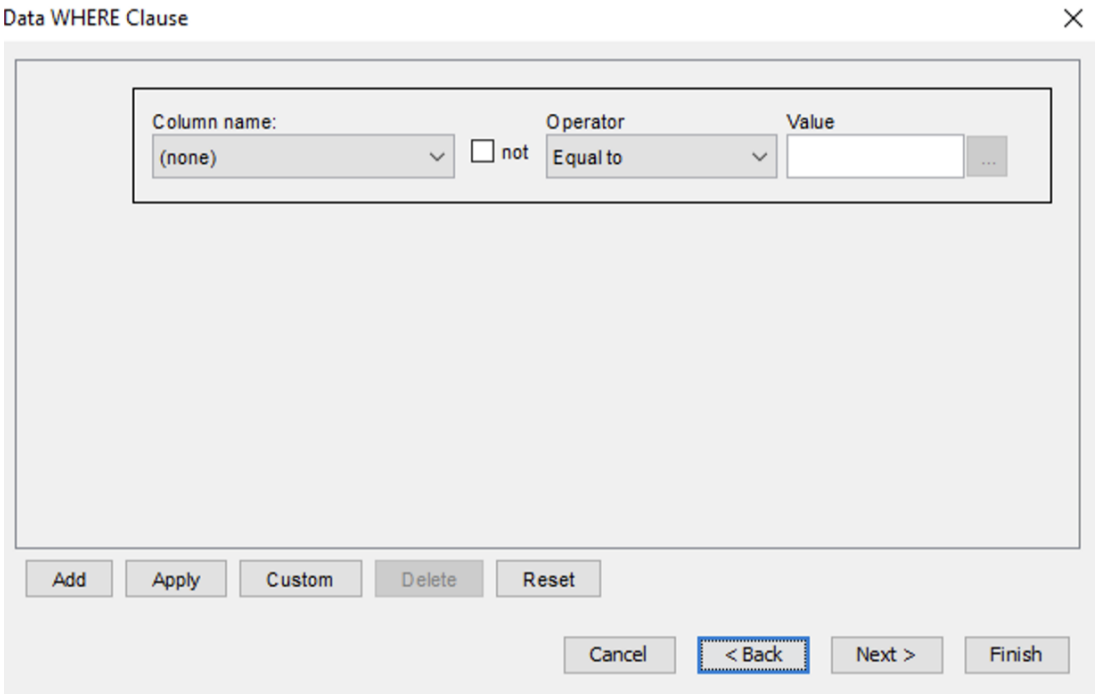


Figure 99

d. Set the chart title as “Branches vs Money Spent on Advertisement”. Also set the X Axis Label as Branch and Y Axis Label as Money Spent on Advertisement. Click Finish

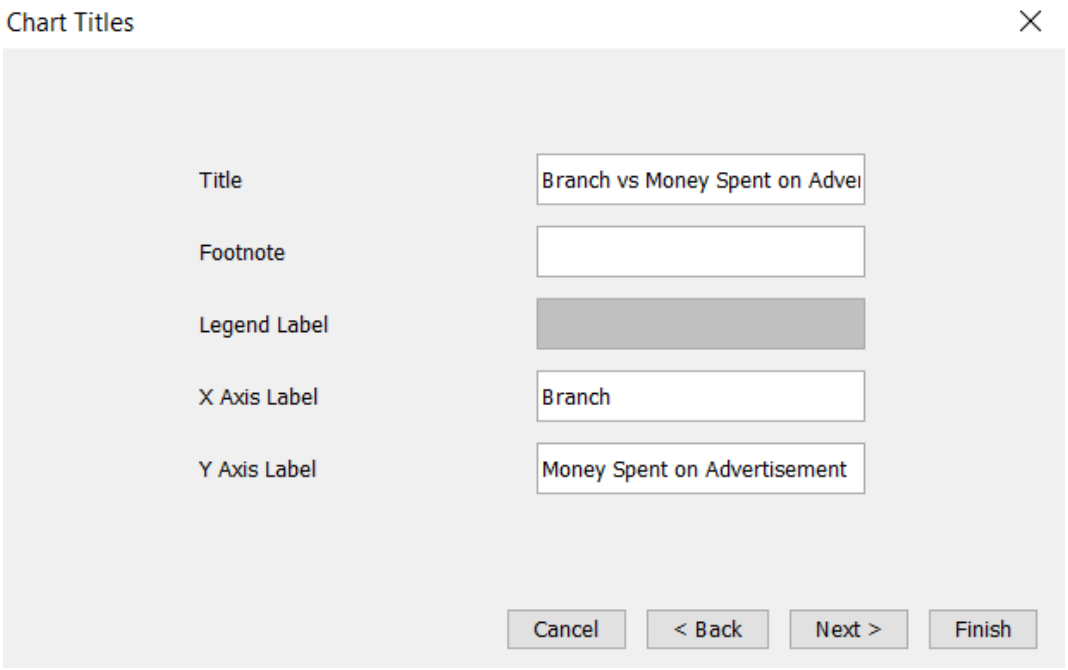


Figure 100

e. The Below graph is developed by SAS EM

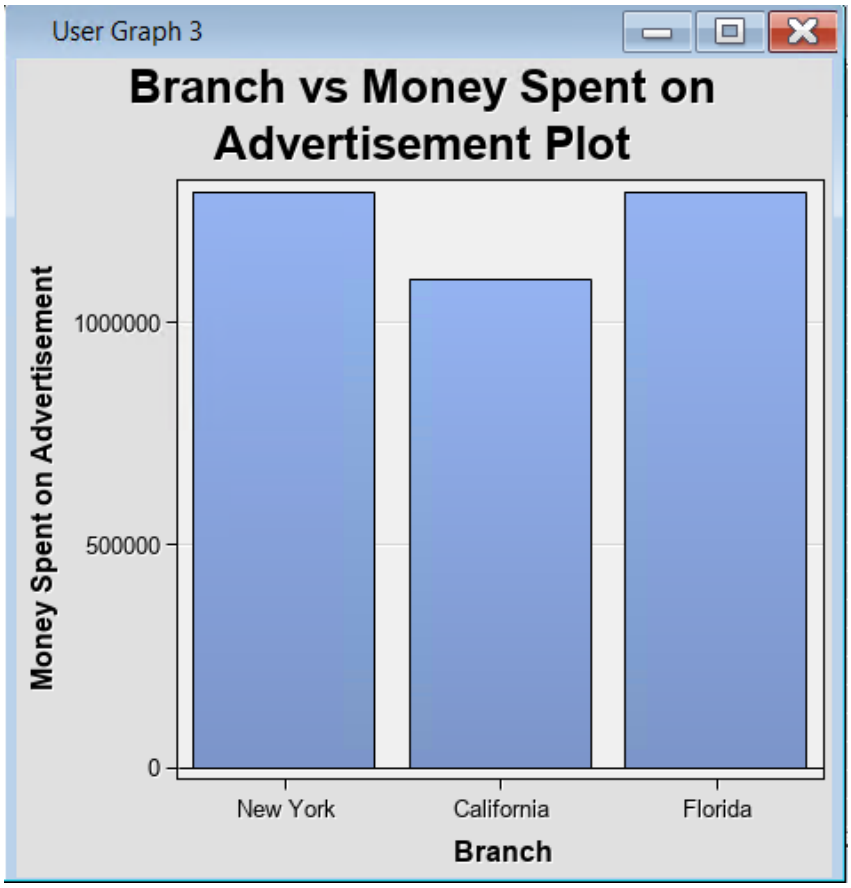


Figure 101

Interpretation & Managerial Implications

The bar graph represents the amount of money spent on advertising by supermarket branches in three different states: New York, California, and Florida.

The graph tells us that New York spent $1,295,309 which is the highest while California spent $1,099,171 being the lowest of all three. Florida spent $1,291,576 and from these figures, we can see that all three states have spent over $10 Million for Advertisement. The supermarket overall has spent around $36 million just for advertisement.

Market Competition: The high spend in New York could suggest a more competitive market, requiring more aggressive advertising efforts. It may also indicate higher advertising costs or a larger number of branches contributing to the total spend.

Advertising Efficiency: Florida's advertising spends being close to that of California, despite having one less branch, might indicate more efficient use of advertising dollars or different market conditions.

In conclusion, the graph suggests that the supermarket investing a huge amount in advertisement to make sure that they have a better reach.

Step 6: Explore the Branch wise money spent on Promotion.

a. Exploring the Branch wise money spent on Promotion distribution using bar chart. Select the chart type as Bar and click Next.

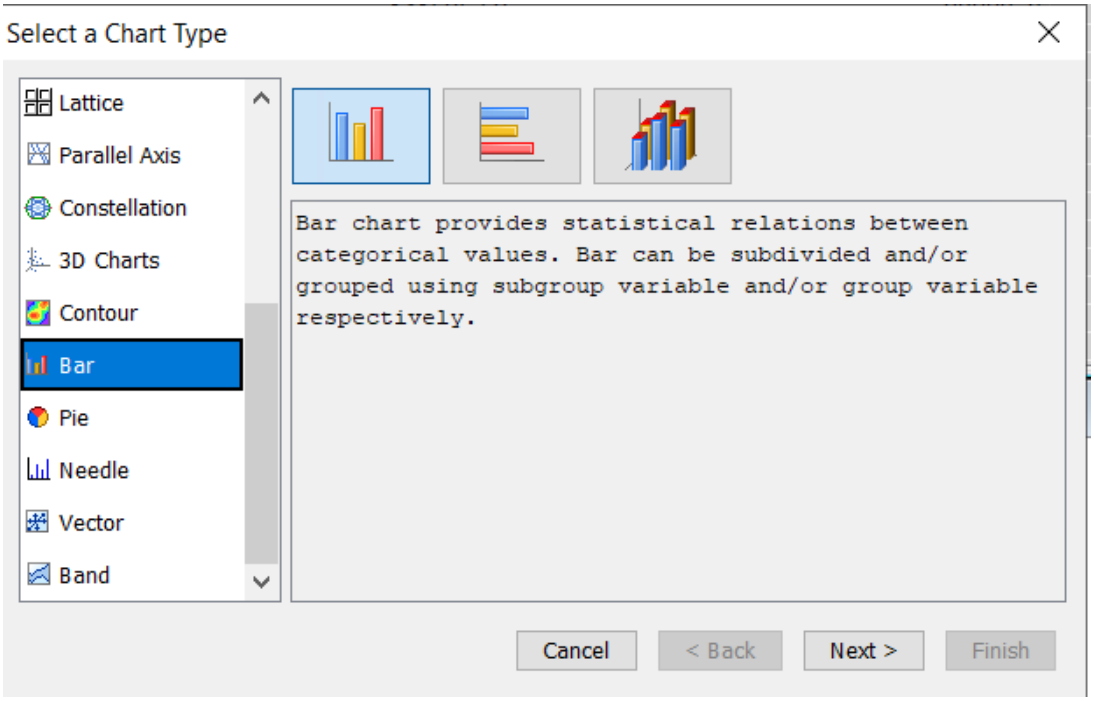


Figure 102

b. Select State as category and Promotion_Spend as frequency. click next.

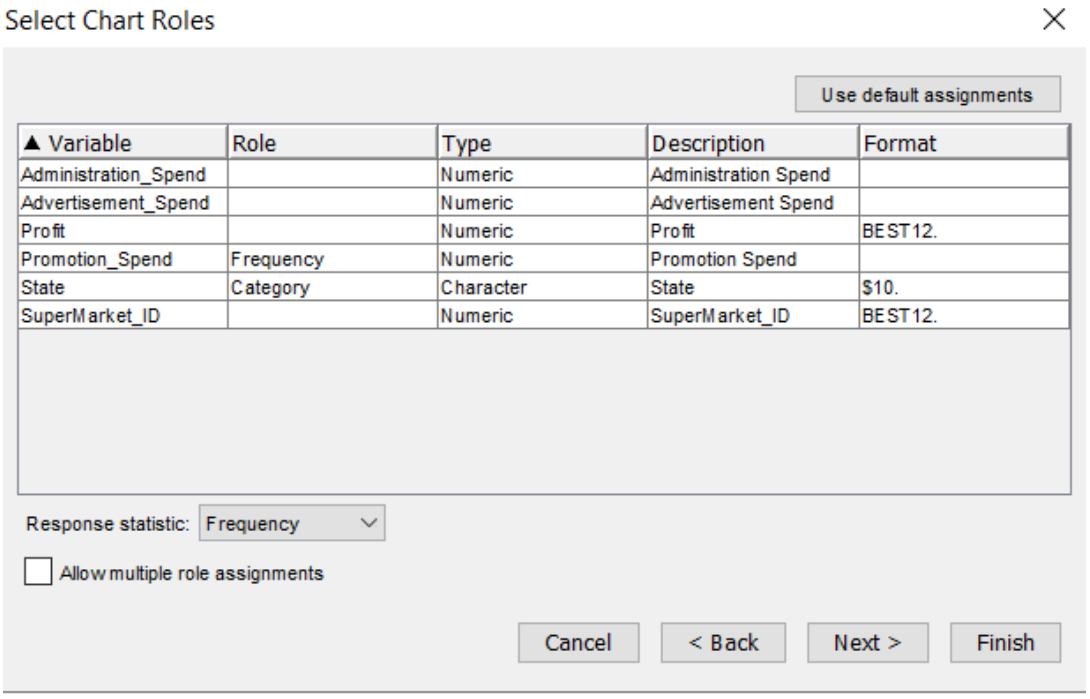


Figure 103

c. There is no where clause, click Next.

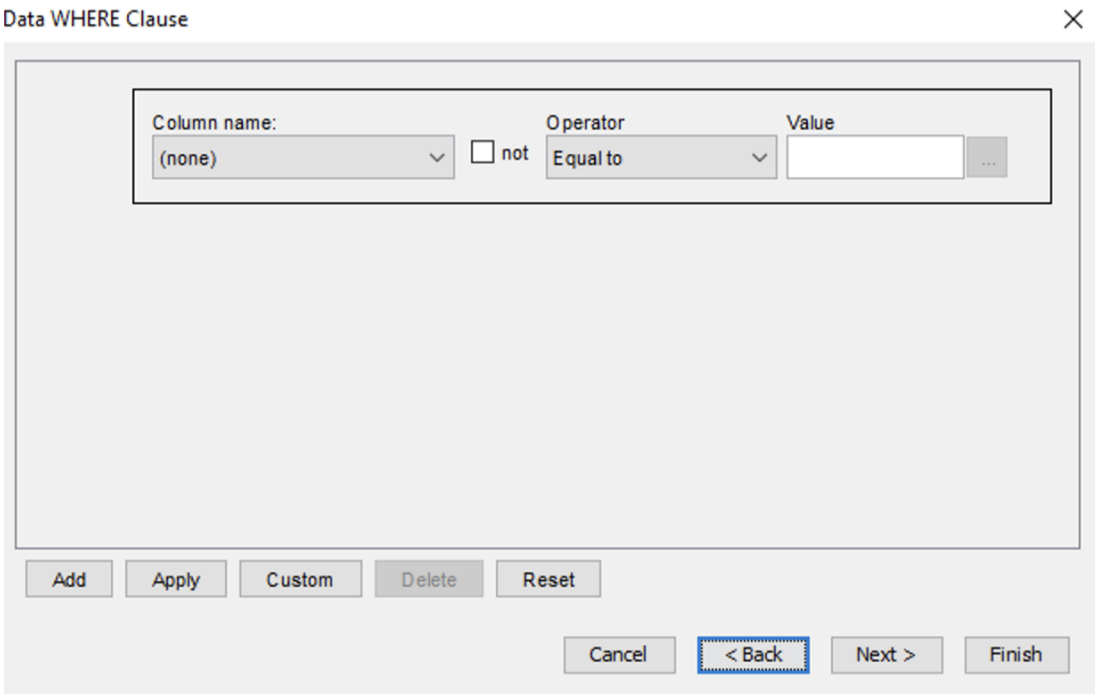


Figure 104

d. Set the chart title as “Branches vs Money Spent on Promotion Plot”. Also set the X Axis Label as Branch and Y Axis Label as Money Spent on Promotion. Click Finish

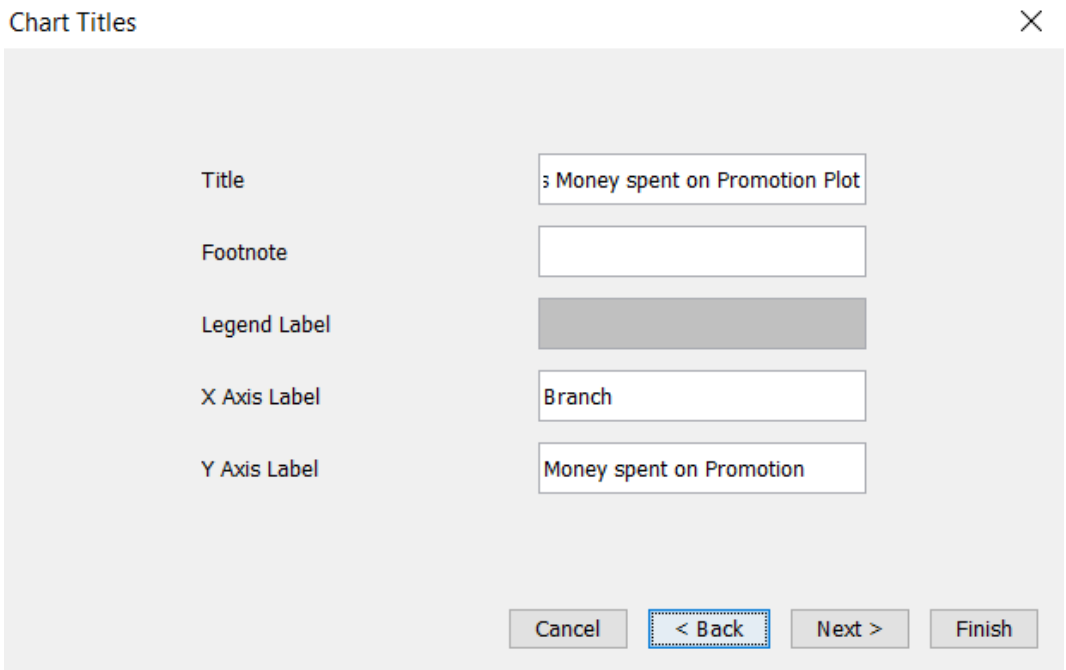


Figure 105

e. The Below graph is developed by SAS EM

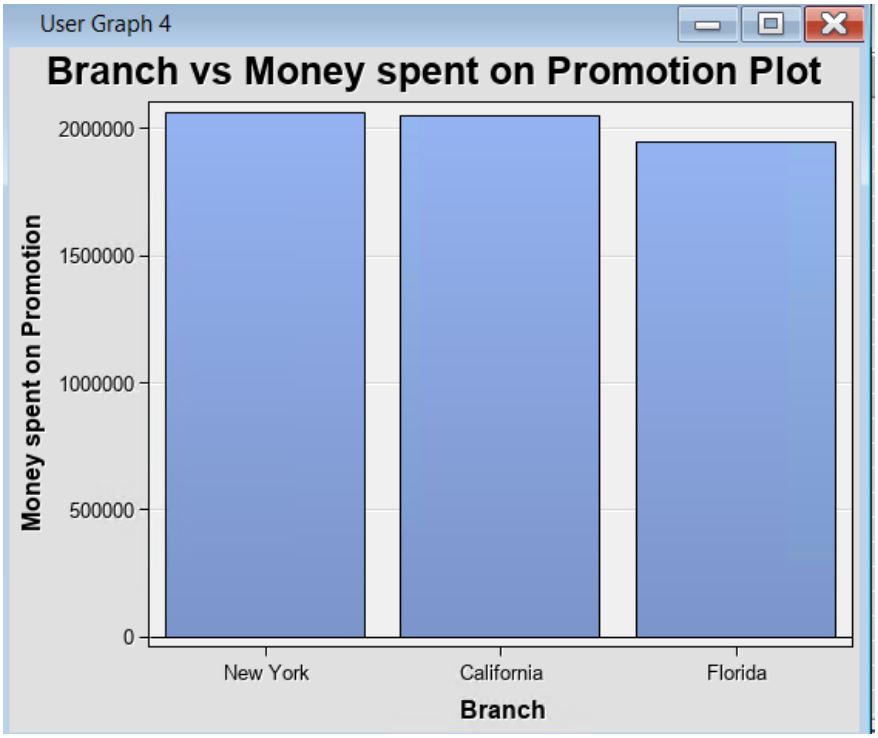


Figure 106

Interpretation & Managerial Implications

The bar graph illustrates the money spent on promotions by supermarket branches in three different states: New York, California, and Florida. From the chart, we can infer the following:

The promotional spend in New York and California is quite close, with New York spending slightly more at $2,066,230 compared to California's $2,052,683. The difference is minimal, suggesting a nearly equal level of investment in promotion in both states. Florida's spend is slightly lower at $1,948,295,but all three states are over the $15M mark.

Competitive Strategies: The high spend could reflect competitive market conditions in these states, necessitating substantial promotion to maintain or increase market share.

Budget Allocation: The slight differences in promotional spending could be due to various factors including market size, competition intensity, or historical sales data indicating where promotional efforts have been more successful.

In summary, the graph shows a strong commitment to promotional activities in all three states, and this suggests a balanced regional marketing approach.

Analysis 5(c): Exploratory Data Analysis

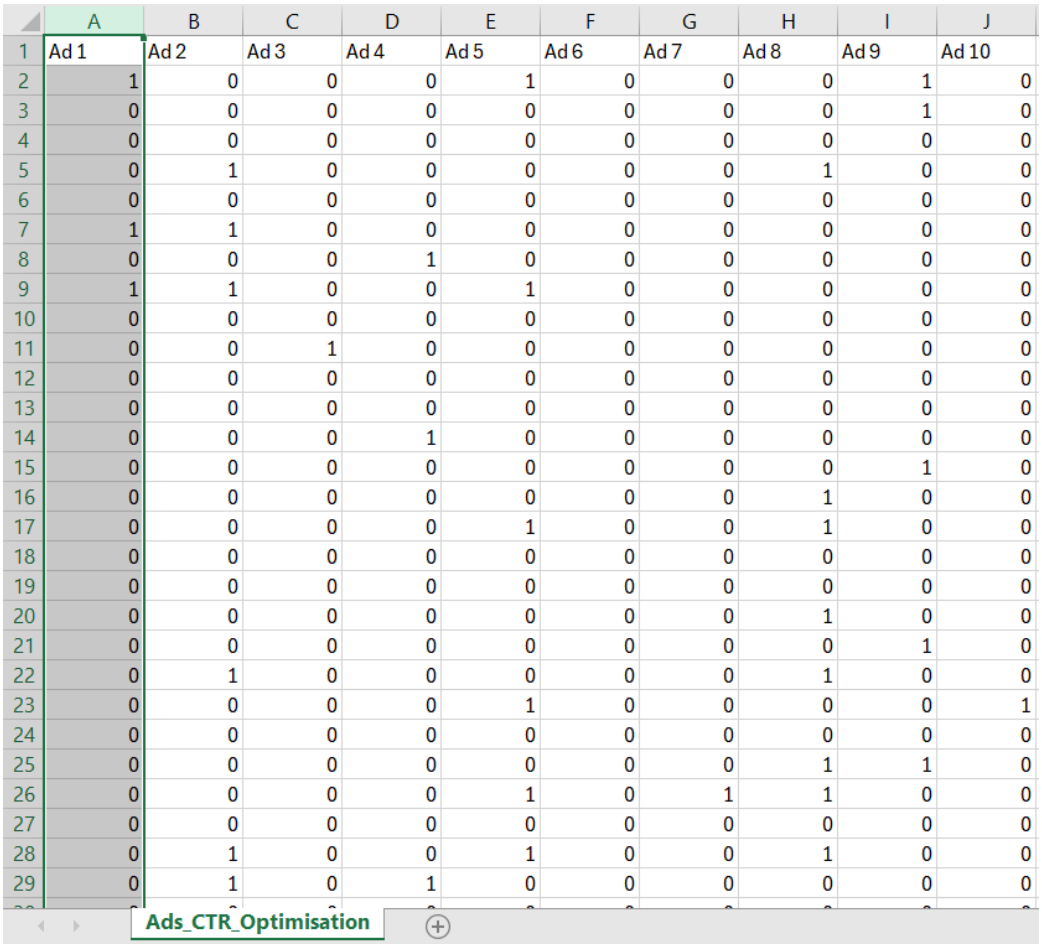
(Ads_CTR_Optimisation)

Dataset: We performed our EDA on the **Ads_CTR_Optimisation.csv** data.

Objective: To analyze the dataset that contains information about click through rates based out of 10000 users and 10 different advertisements that the supermarket created

Procedure

The dataset comprises information on 10,000 users and their interactions with various advertisements, encoded as binary data where '0' indicates the ad was not clicked and watched, and '1' signifies that the ad was engaged with. The following image is a select excerpt from the dataset. To analyze and visualize the dataset from multiple perspectives, we utilized the StatExplore functionality within SAS Enterprise Miner.



	A	B	C	D	E	F	G	H	I	J
1	Ad 1	Ad 2	Ad 3	Ad 4	Ad 5	Ad 6	Ad 7	Ad 8	Ad 9	Ad 10
2	1	0	0	0	1	0	0	0	1	0
3	0	0	0	0	0	0	0	0	1	0
4	0	0	0	0	0	0	0	0	0	0
5	0	1	0	0	0	0	0	1	0	0
6	0	0	0	0	0	0	0	0	0	0
7	1	1	0	0	0	0	0	0	0	0
8	0	0	0	1	0	0	0	0	0	0
9	1	1	0	0	1	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0
11	0	0	1	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0
14	0	0	0	1	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	1	0
16	0	0	0	0	0	0	0	1	0	0
17	0	0	0	0	1	0	0	1	0	0
18	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	1	0	0
21	0	0	0	0	0	0	0	0	1	0
22	0	1	0	0	0	0	0	1	0	0
23	0	0	0	0	1	0	0	0	0	1
24	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	1	1	0
26	0	0	0	0	1	0	1	1	0	0
27	0	0	0	0	0	0	0	0	0	0
28	0	1	0	0	1	0	0	1	0	0
29	0	1	0	1	0	0	0	0	0	0

Ads_CTR_Optimisation

Figure 107

Step 1: Import the file & edit variables

To import the excel file, we performed the same steps performed in the previous analysis and entered the path to the location where the file **Ads_CTR_Optimisation.csv** is located.

Rename the File Import node as **Ads_CTR_Optimisation**. We performed the edit variable step to set the below roles as shown in the diagram to the columns of the dataset. By this step the file import is complete.

In the edit variable section, we make sure that all the roles are set as per the image below and complete the file import.

Variables - FIMPORT

(none) not Equal to

Columns: Label Mining

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Ad_1	Input	Binary	No		No	.	.
Ad_10	Input	Binary	No		No	.	.
Ad_2	Input	Binary	No		No	.	.
Ad_3	Input	Binary	No		No	.	.
Ad_4	Input	Binary	No		No	.	.
Ad_5	Input	Binary	No		No	.	.
Ad_6	Input	Binary	No		No	.	.
Ad_7	Input	Binary	No		No	.	.
Ad_8	Input	Binary	No		No	.	.
Ad_9	Input	Binary	No		No	.	.

Figure 108

Step 2: Add the StatExplore node

a. Click the Explore tab and find StatExplore node

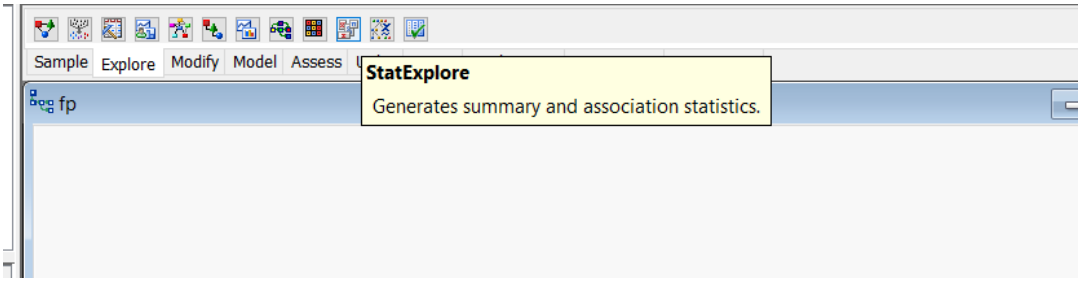


Figure 109

b. Drag the Association node to the diagram and then connect the File Import node & StatExplore node to each other

Figure 110

Step 3: Run the diagram

a. Run the diagram & right click on StatExplore node to see the results.

The graphs show us the percentage of users who have clicked through and who have not for the ads from 1 to 10 that were produced by the supermarket.

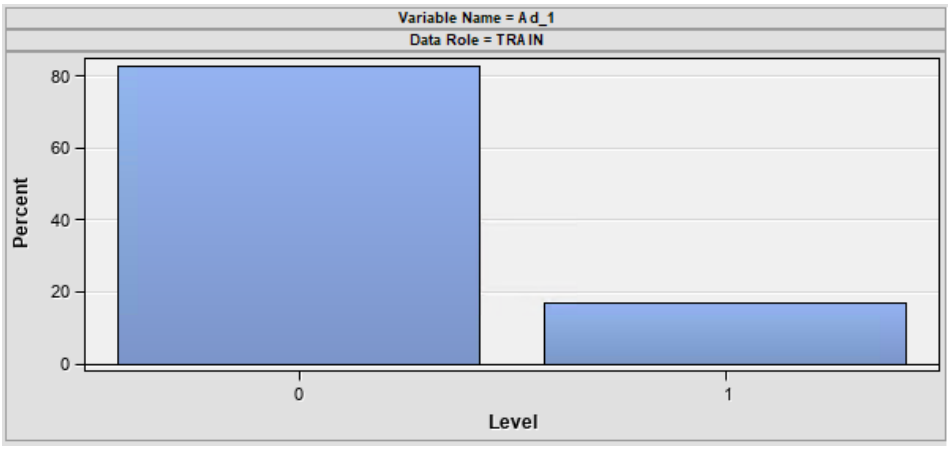


Figure 111

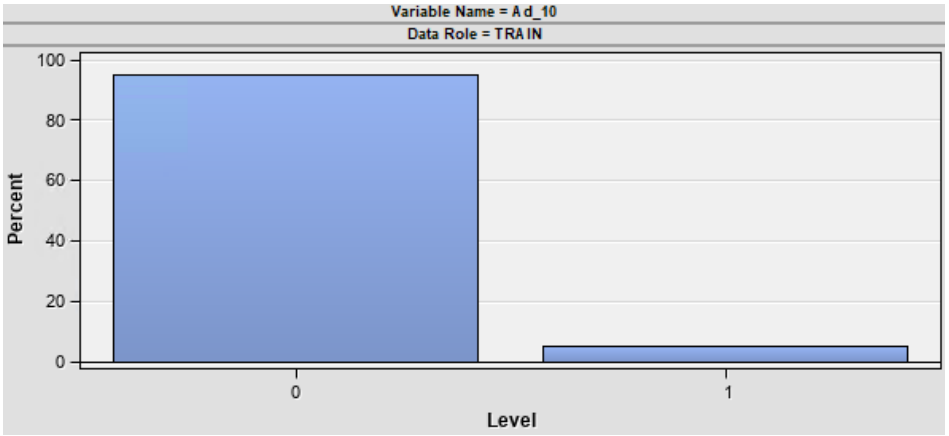


Figure 112

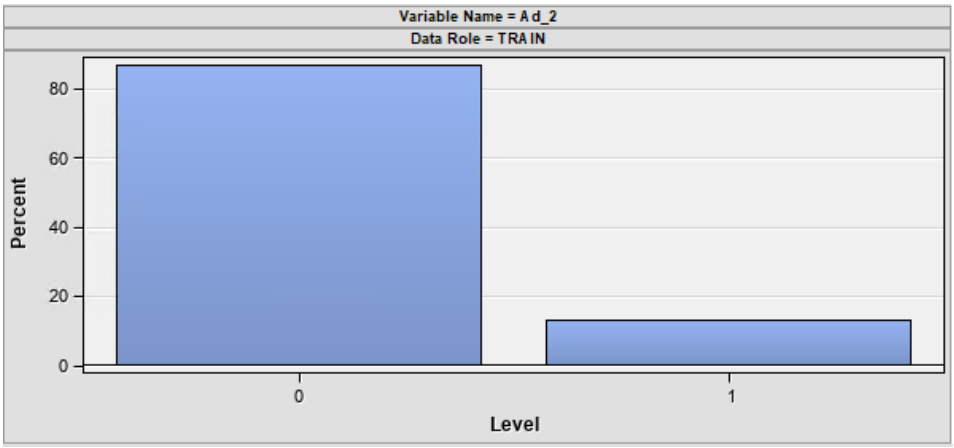


Figure 113

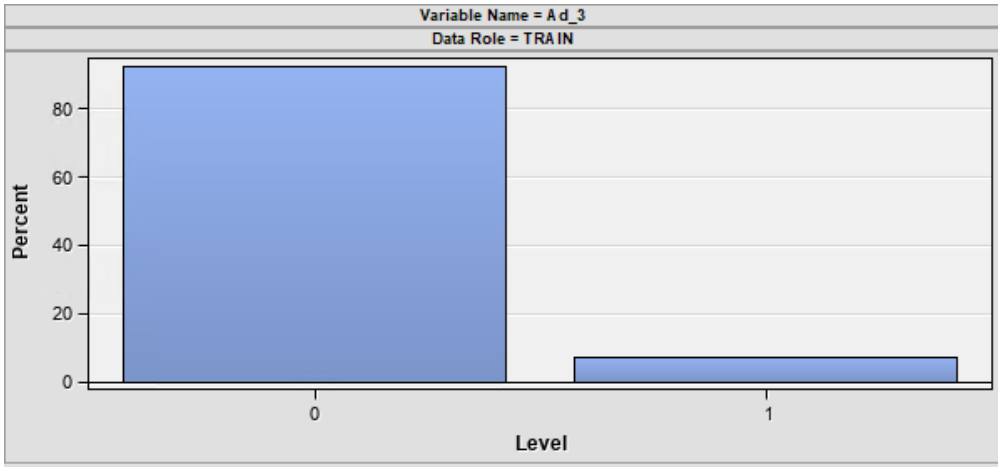


Figure 114

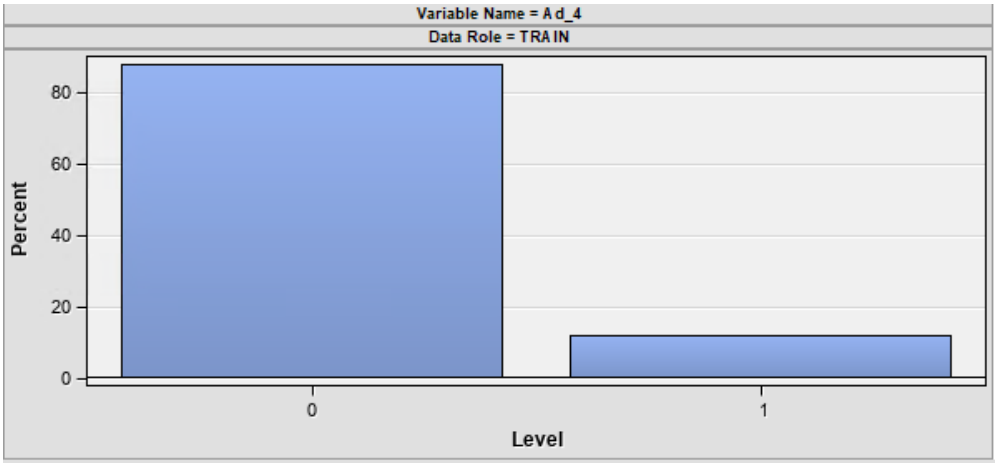


Figure 115

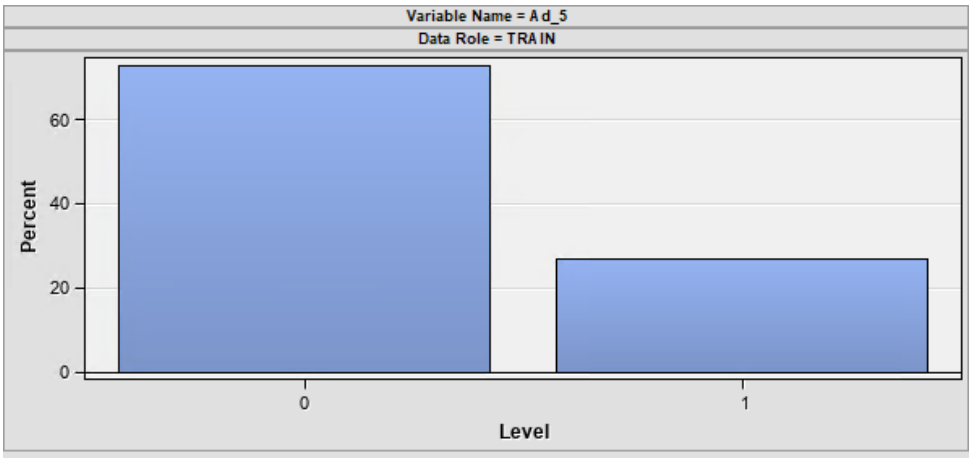


Figure 116

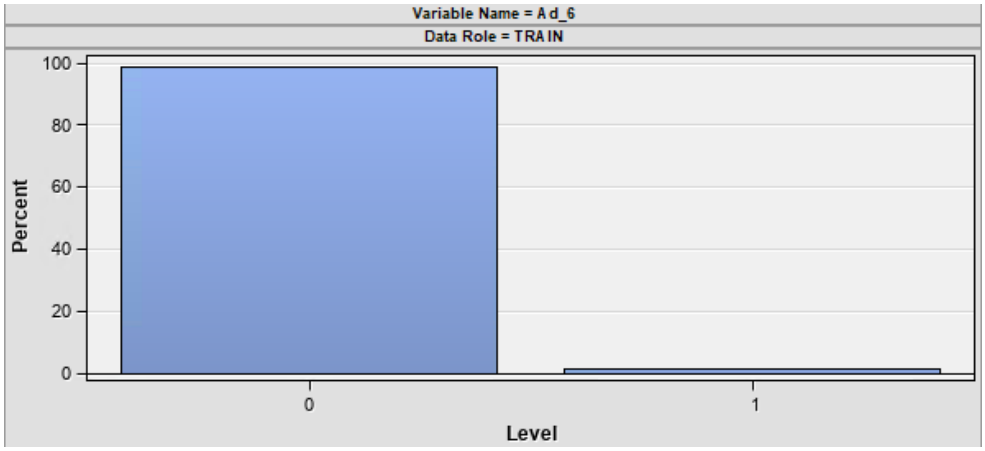


Figure 117

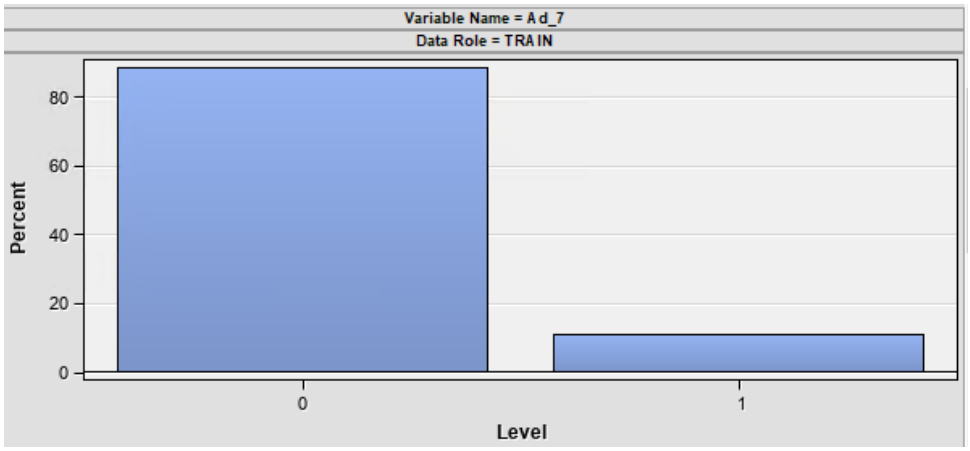


Figure 118

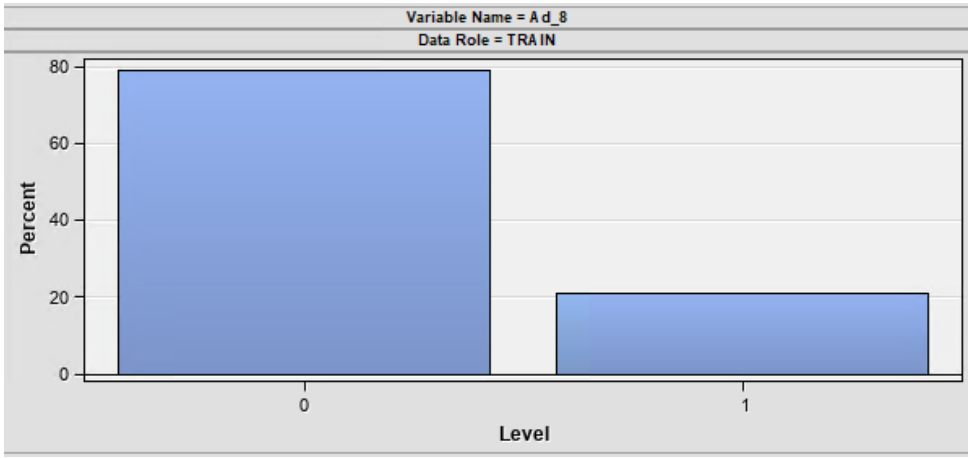


Figure 119

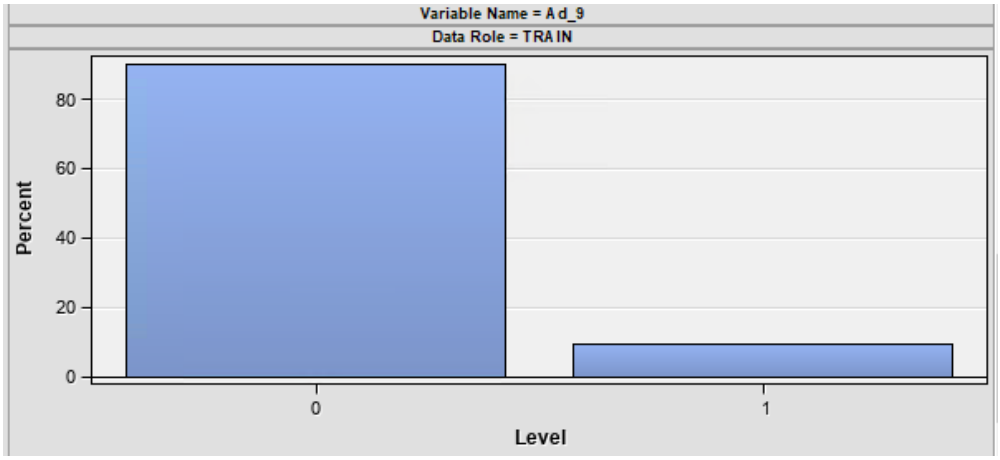


Figure 120

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Ad_1	INPUT	2	0	0	82.97	1	17.03
TRAIN	Ad_10	INPUT	2	0	0	95.11	1	4.89
TRAIN	Ad_2	INPUT	2	0	0	87.05	1	12.95
TRAIN	Ad_3	INPUT	2	0	0	92.72	1	7.28
TRAIN	Ad_4	INPUT	2	0	0	88.04	1	11.96
TRAIN	Ad_5	INPUT	2	0	0	73.05	1	26.95
TRAIN	Ad_6	INPUT	2	0	0	98.74	1	1.26
TRAIN	Ad_7	INPUT	2	0	0	88.88	1	11.12
TRAIN	Ad_8	INPUT	2	0	0	79.09	1	20.91
TRAIN	Ad_9	INPUT	2	0	0	90.48	1	9.52

Figure 121

Interpretation & Managerial Implications

The output contains data on the click-through rates (CTR) for various advertisements, labeled Ad_1 through Ad_10. Each ad variable has two levels, representing clicked (1) or not clicked (0), and the table includes the percentage of ads clicked (mode) and not clicked (mode2). The click rates for the advertisements vary from a low of 1.26% for Ad_6 to a high of 26.95% for Ad_5. This wide range reflects varying levels of engagement that each advertisement receives.

The table shows a significant imbalance in the click rates. Mode having the binary value 0 percentages are high, suggesting that most users did not click on the ads, which is common regarding advertisements. This interpretation tells us that all the advertisement needs to be reviewed and measures must be taken to increase engagement activities.

Learning Conclusion

These processes have provided us with a comprehensive understanding of SAS Enterprise Miner and the SEMMA methodology, demonstrating how various models can be applied to evaluate real-world issues through data-driven insights.

With respect to this project, by implementing the strategies based on these managerial implications inferred through each analysis and holistically considering the insights, XYZ Supermarket can enhance operational efficiency, customer satisfaction, and profitability.