

**The University of Texas at Dallas**  
School of Management

**BUAN 6324 Business Analytics with SAS**  
**Instructor: Syam Menon**

**Term Paper: Final report**

---

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

---



**THE UNIVERSITY OF TEXAS AT DALLAS**  
**Naveen Jindal School of Management**

**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 spent 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.

The screenshot shows a Microsoft Excel spreadsheet titled "Market\_Basket\_Optimisation". The data is presented in a table with columns labeled A through P. Row 1 contains the column headers: A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P. Rows 2 through 20 contain transaction data. For example, Transaction 1 (row 2) includes items like "shrimp", "almonds", "avocado", "vegetables", "green grapes", "whole wheat", "yams", "cottage", "chee", "energy", "drink", "tomato", "juice", "low fat", "yogurt", "green tea", "honey", "salad", "mineral", "water", and "salmon". Transaction 20 (row 20) includes "shrimp", "chocolate", "chicken", "honey", "oil", "cooking oil", and "low fat yogurt". The Excel interface shows various toolbars and ribbon tabs at the top, and the formula bar at the bottom displays "fx shrimp".

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	shrimp	almonds	avocado	vegetables	green grapes	whole wheat	yams	cottage	chee	energy	drink	tomato	juice	low fat	yogurt
2	burgers	meatballs	eggs											green tea	honey
3	chutney													salad	mineral water
4	turkey	avocado												salmon	
5	mineral water	milk	energy bar	whole wheat	green tea										
6	low fat yogurt														
7	whole wheat	french fries													
8	soup	light cream	shallot												
9	frozen veget	spaghetti	green tea												
10	french fries														
11	eggs	pet food													
12	cookies														
13	turkey	burgers	mineral water	eggs	cooking oil										
14	spaghetti	champagne	cookies												
15	mineral water	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 water	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

```
import pandas as pd

# Load the data from Excel
df = pd.read_excel("content/ForPythonMarket_Basket_Optimisation.xls", header=None)

# Initialize an empty DataFrame to store the transformed data
transformed_data = pd.DataFrame(columns=['transaction_id', 'product'])

# Loop through each row in the original DataFrame
for idx, row in df.iterrows():
    # Extract the transaction ID
    transaction_id = idx + 1
    # Look for the first column (product) in the row
    for product in row:
        # Check if the product is not NaN (i.e., it exists)
        if pd.notna(product):
            # Append a row to the transformed DataFrame
            transformed_data = transformed_data.append({'transaction_id': transaction_id, 'product': product})

# Convert transaction_id to integer type
transformed_data['transaction_id'] = transformed_data['transaction_id'].astype(int)

# Save the DataFrame to an Excel file
transformed_data.to_excel('new_data.xlsx', index=False)

# Display a message indicating the file has been saved
print("Transformed data saved to 'new_data.xlsx'")
```

Please follow our [blog](#) to see more information about new features, tips and tricks, and featured notebooks such as [Analyzing a Bank Failure with Colab](#).

**2024-02-21**

- Try out Gemini on [Colab](#)!
- Allow unicode in form text inputs
- Display documentation and link to source when displaying functions
- Display image-like ndarrays as images
- Improved UX around quick charts and execution error suggestions
- Released Marketplace image for the month of February ([GitHub issue](#))
- Python package upgrade
  - bigframes 0.19.2->0.21.0
  - regeval 2.0.0->2.0.2
  - spacy 4.6.1->3.7.4
  - beautifulsoup4 4.11.2->4.12.3
  - tensorflow-probability 0.22.0->0.23.0
  - google-cloud-language 2.9.1->2.13.1
  - polars 0.20.0->1.39.0->1.42.1
  - transformers 4.9.2->4.37.2
  - pyarrow 10.0.1->14.0.2

**2024-01-29**

- New [Kaggle Notebooks <-> Colab](#) updates! Now you can:
  - Import directly from Colab without having to download/re-upload
  - Upload via link, by pasting Google Drive or Colab URLs
  - Export & run Kaggle Notebooks on Colab with 1 click
- Try them out with [this Colab notebook](#).
- Gemini and Stable Diffusion
- Learning with Gemini and ChatGPT
- Talk to Gemini with Google's Speech-to-Text API
- Sell lemonade with Gemini and Sheets
- Generate images with Gemini and Vertex
- Python package upgrade
  - google-cloud-storage 1.38.1->1.39.0
  - bigframes 0.18.0->0.19.2
  - polars 0.17.3->0.20.2
  - gdown 4.6.6->4.7.3 ([GitHub issue](#))
  - tensorflow-hub 0.15.0->0.16.0
  - flax 0.7.5->0.8.0
- Python package inclusions
  - sentencepiece 0.1.99

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.

The screenshot shows a Microsoft Excel spreadsheet titled "ARM\_transformed\_data". The table has two columns: "transaction\_id" and "product". The data includes various grocery items like shrimp, almonds, avocado, vegetables mix, green grapes, whole wheat flour, yams, cottage cheese, energy drink, tomato juice, low fat yogurt, green tea, honey, salad, mineral water, salmon, antioxidant juice, frozen smoothie, spinach, olive oil, burgers, meatballs, eggs, chutney, turkey, and avocado.

transaction_id	product
1	shrimp
1	almonds
1	avocado
1	vegetables mix
1	green grapes
1	whole wheat 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	antioxidant 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.

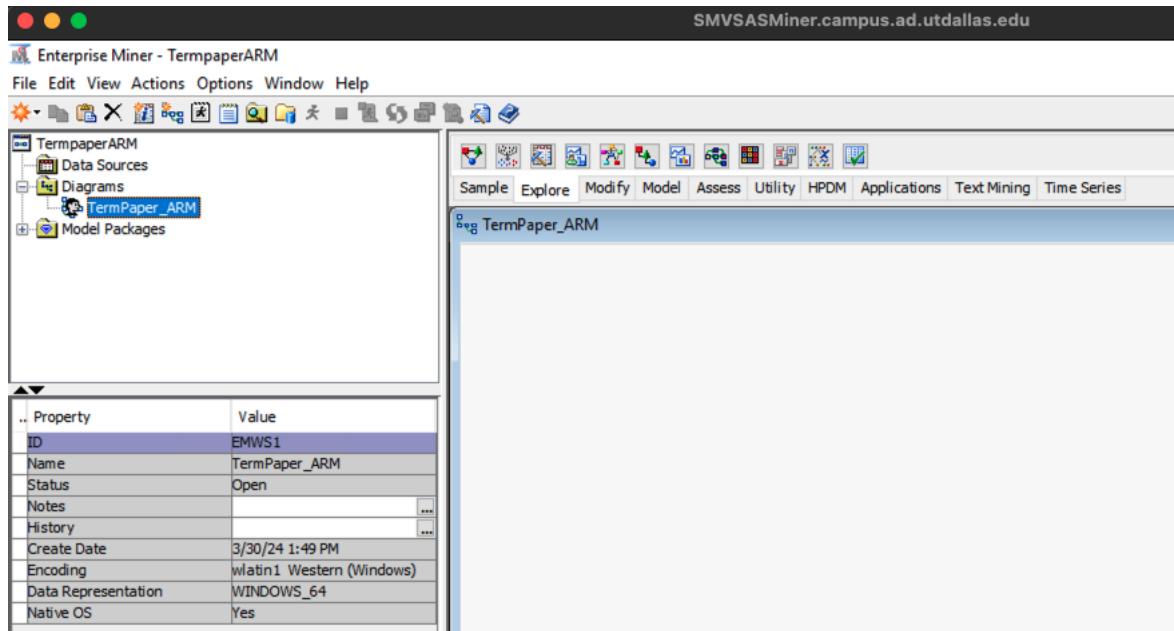


Figure 4

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.

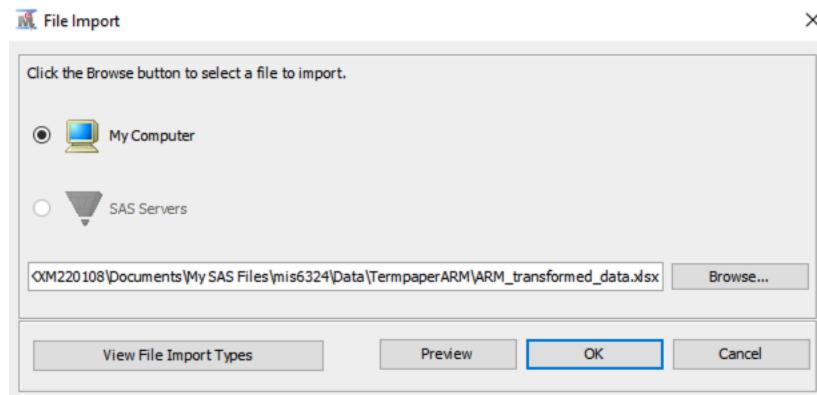


Figure 5

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

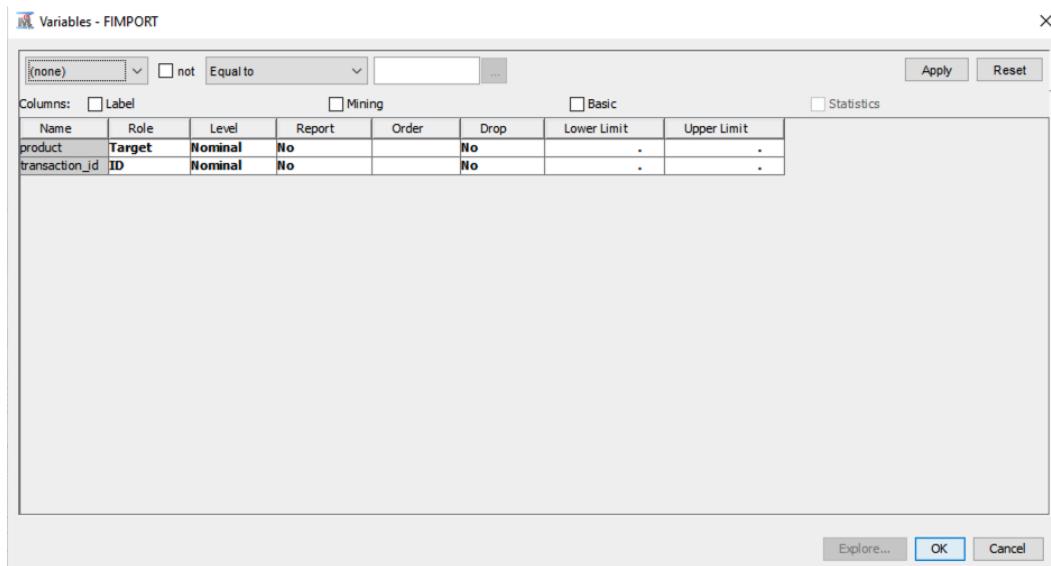


Figure 6

- 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.

.. Property	Value
Guessing Rows	500
File Location	Local
File Type	xlsx
Advanced Advisor	No
Rerun	No
<b>Score</b>	
Role	Transaction
<b>Report</b>	
Summarize	No
<b>Status</b>	
Create Time	3/31/24 12:44 PM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

Figure 7

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

The screenshot shows the Enterprise Miner interface with the title bar "Enterprise Miner - TermpaperARM" and the URL "SMVSASMiner.campus.ad.utdallas.edu". The menu bar includes File, Edit, View, Actions, Options, Window, and Help. The toolbar has various icons for file operations like Open, Save, and Print. The left sidebar displays a project structure under "TermpaperARM": Data Sources, Diagrams, TermPaper\_ARM (selected), and Model Packages. The main workspace contains a "File Import" node with a green checkmark icon. To the right of the workspace is a properties panel titled "General" with the following settings:

.. Property	Value
<b>General</b>	
Node ID	FIMPORT2
Imported Data	[...]
Exported Data	[...]
Notes	[...]
<b>Train</b>	
Variables	[...]
Import File	E:\Users\XJM220108\Docu...
Maximum Rows to Import	1000000
Maximum Columns to Import	10000
Delimiter	,
Name Row	Yes
Number of Rows to Skip	0
Guessing Rows	500
File Location	Local

Figure 8

#### Step 4: Add the Association node

- Click the Explore tab and drag the Association node to the diagram
- Connect the File Import node & Association node to each other.

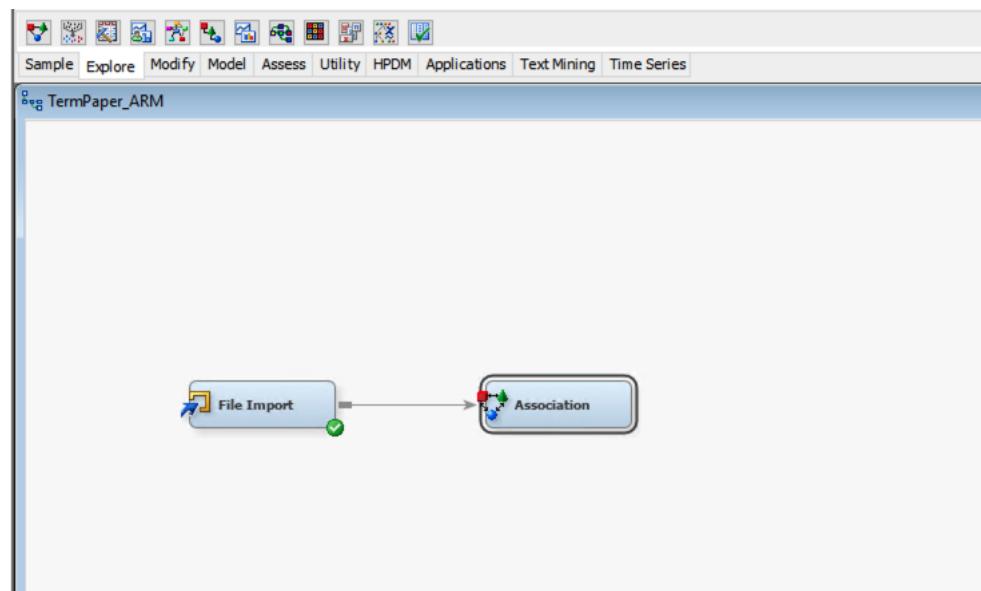


Figure 9

- 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
<b>General</b>	
Node ID	Assoc
Imported Data	[...]
Exported Data	[...]
Notes	[...]
<b>Train</b>	
Variables	[...]
Maximum Number of Items to	100000
Rules	[...]
<b>Association</b>	
Maximum Items	4
Minimum Confidence Level	10
Support Type	Percent
Support Count	.
Support Percentage	5.0

Figure 10

- 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.

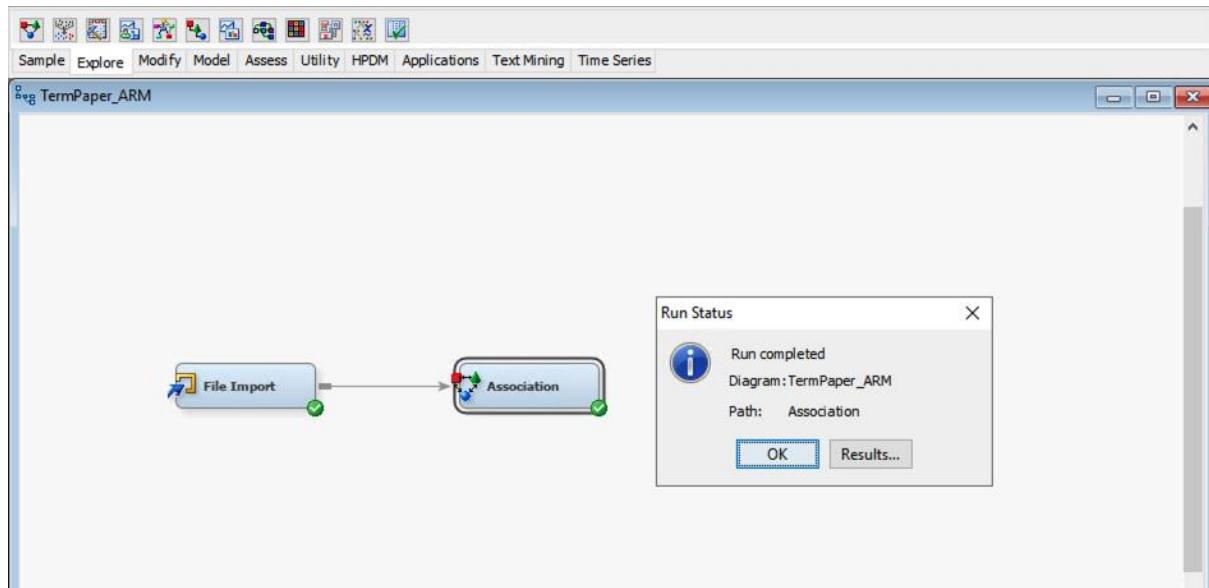


Figure 12

### Results Interpretation

```

Output
1 User: KCM230108
2 Date: March 31, 2024
3 Time: 12:50:03
4 -----
5 * Training Output
6 -----
7 -----
8 -----
9 -----
10 -----
11 -----
12 Variable Summary
13 -----
14 Measurement Frequency
15 Role Level Count
16 -----
17 ID NOMINAL 1
18 TARGET NOMINAL 1
19 -----
20 -----
21 -----
22 -----
23 Association Report
24 -----
25 Expected Confidence Support Transaction
26 Relations (%) (%) (%) Lift Count Rule
27 -----
28
29 2 9.83 32.35 1.60 3.29 120.00 herb & pepper ==> ground beef
30 2 4.95 16.28 1.69 3.29 120.00 ground beef ==> herb & pepper
31 3 9.83 28.57 1.71 2.91 128.00 spaghetti & mineral water ==> ground beef
32 3 5.97 17.37 1.71 2.91 128.00 ground beef ==> spaghetti & mineral water
33 2 9.53 23.59 1.46 2.47 120.00 tomatoes ==> frozen vegetables
34 2 6.94 18.92 1.61 2.47 121.00 frozen vegetables ==> tomatoes
35 2 9.53 23.32 1.67 2.45 125.00 shrimp ==> frozen vegetables
36 2 7.15 17.48 1.67 2.45 125.00 frozen vegetables ==> shrimp
37 3 17.41 41.69 1.71 2.39 128.00 mineral water & ground beef ==> spaghetti
38 2 12.96 30.08 1.52 2.32 114.00 soup ==> milk
39 2 5.05 11.73 1.52 2.32 114.00 milk ==> soup
40 2 9.83 32.35 1.52 2.32 114.00 ground beef ==> soup
41 2 17.41 39.89 1.52 2.32 154.00 ground beef ==> spaghetti
42 2 9.83 21.46 1.41 2.18 106.00 olive oil ==> ground beef
43 2 6.59 14.38 1.41 2.18 106.00 ground beef ==> olive oil
44 3 9.53 20.09 1.20 2.11 90.00 spaghetti & mineral water ==> frozen vegetables
45 2 5.97 12.59 1.20 2.11 90.00 spaghetti & mineral water ==> spaghetti & mineral water
46 3 12.96 26.58 1.40 2.05 106.00 mineral water & chocolate ==> milk
47 3 12.96 10.80 1.40 2.05 106.00 milk ==> mineral water & chocolate
48 3 12.96 26.34 1.57 2.03 118.00 spaghetti & mineral water ==> milk
49 3 5.97 12.14 1.57 2.03 118.00 milk ==> spaghetti & mineral water
50 2 6.59 13.17 2.29 2.00 172.00 spaghetti ==> olive oil
51 2 17.41 34.82 2.29 2.00 172.00 olive oil ==> spaghetti
52 2 6.59 13.17 1.71 2.00 128.00 milk ==> olive oil
53 2 12.96 25.91 1.71 2.00 128.00 olive oil ==> milk
54
55

```

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 valuable insights about the customer behaviour, preferences & purchase pattern. The strong associations indicate 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 relationships and patterns about customer purchase behaviour. We were able to meet our objective of discovering actionable insights that would assist 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).

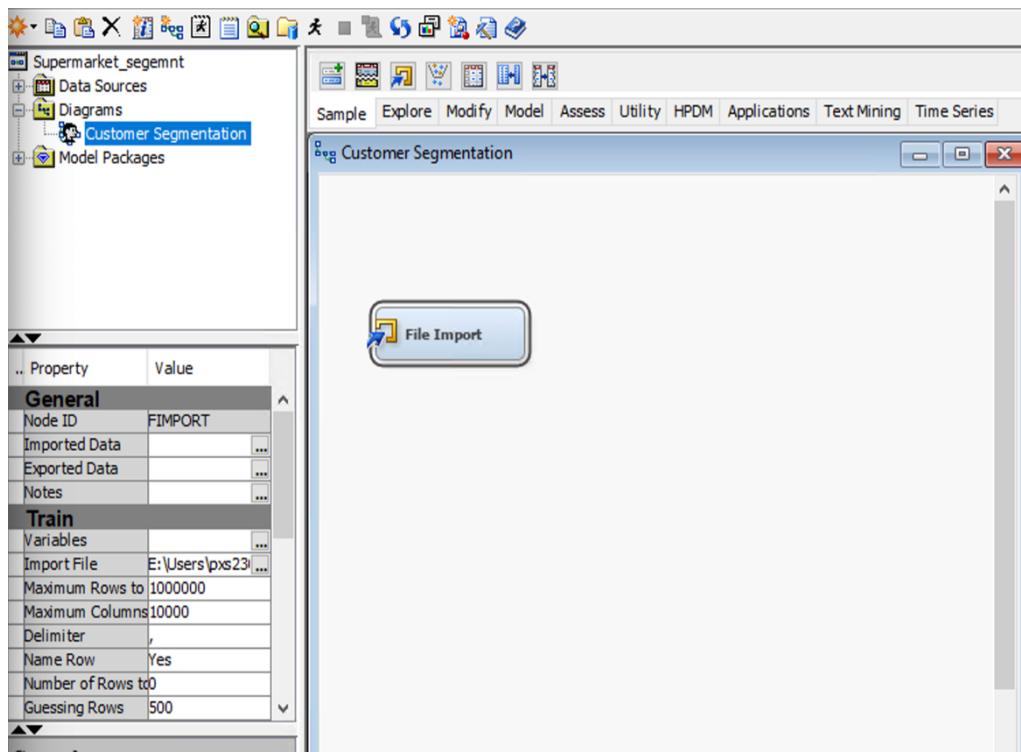


Figure 15

- 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)

Variables - FIMPORT						
<input type="checkbox"/> (none)		<input type="checkbox"/> not	<input type="checkbox"/> Equals		<input type="checkbox"/> Apply	<input type="checkbox"/> Reset
Columns:		<input type="checkbox"/> Label	<input type="checkbox"/> Mining	<input type="checkbox"/> Basic	<input type="checkbox"/> Statistics	
Name	<input type="checkbox"/> Role	<input type="checkbox"/> Level	<input type="checkbox"/> Report	<input type="checkbox"/> Order	<input type="checkbox"/> Drop	<input type="checkbox"/> Lower Li
Age	<input checked="" type="checkbox"/> Input	<input checked="" type="checkbox"/> Interval	<input checked="" type="checkbox"/> No		<input checked="" type="checkbox"/> No	
Annual_Income	<input checked="" type="checkbox"/> Input	<input checked="" type="checkbox"/> Interval	<input checked="" type="checkbox"/> No		<input checked="" type="checkbox"/> No	
CustomerID	<input checked="" type="checkbox"/> ID	<input checked="" type="checkbox"/> Nominal	<input checked="" type="checkbox"/> No		<input checked="" type="checkbox"/> No	
Gender	<input checked="" type="checkbox"/> Input	<input checked="" type="checkbox"/> Binary	<input checked="" type="checkbox"/> No		<input checked="" type="checkbox"/> No	
Spending_Score	<input checked="" type="checkbox"/> Input	<input checked="" type="checkbox"/> Interval	<input checked="" type="checkbox"/> No		<input checked="" type="checkbox"/> No	

Figure 16

## Step 2: Analyse missing values using stat explore node

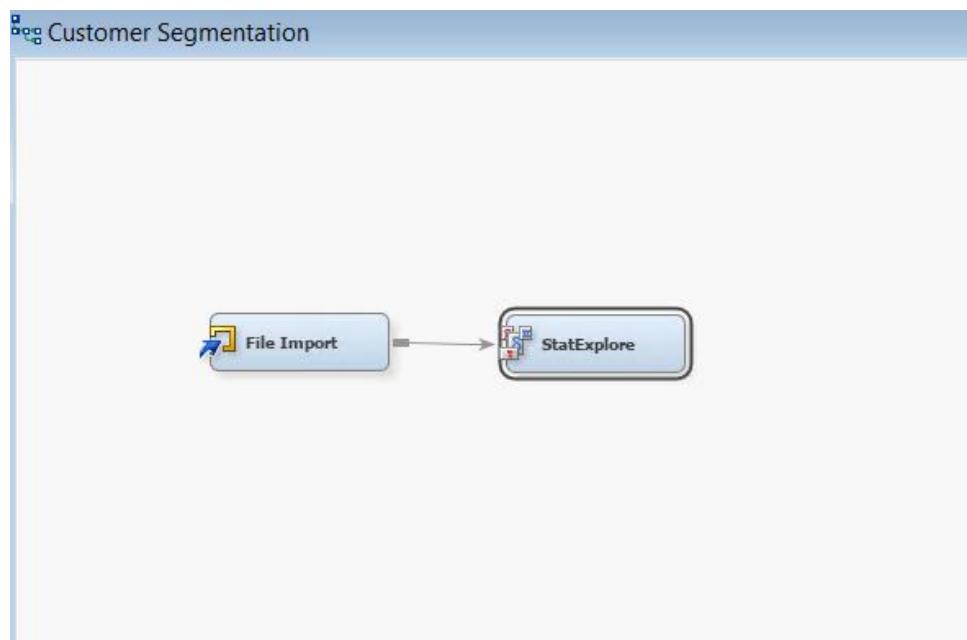


Figure 17

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											
22 Class Variable Summary Statistics 23 (maximum 500 observations printed)											
24											
25 Data Role=TRAIN											
26											
27											
28	Role	Variable	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage			
29	Role	Name	Role								
30	31	TRAIN	Gender	INPUT	2	0	Female	56.00	Male	44.00	
32	33										
34	35	Interval Variable Summary Statistics (maximum 500 observations printed)									
36	37	Data Role=TRAIN									
38	39										
40	41	Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness
42	43	Age	INPUT	38.85	13.96901	200	0	18	36	70	0.485569
44	45	Annual_Income_k_	INPUT	60.56	26.26472	200	0	15	61	137	0.321843
46	47	CustomerID	INPUT	100.5	57.87918	200	0	1	100	200	-0.09849
48	Spending_Score__1_100_	INPUT		50.2	25.82352	200	0	1	50	99	-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

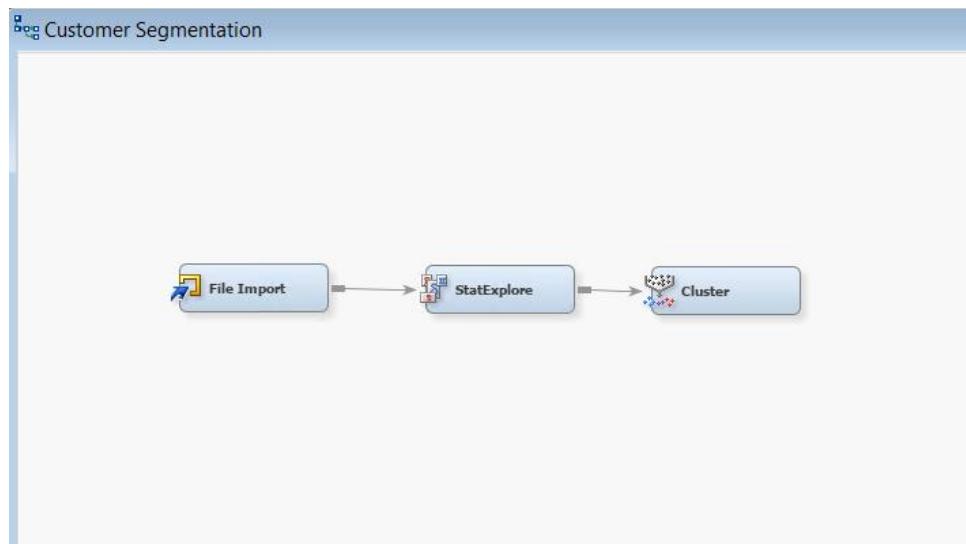


Figure 19

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.

.. Property	Value
<b>General</b>	
Node ID	Clus
Imported Data	[...]
Exported Data	[...]
Notes	[...]
<b>Train</b>	
Variables	[...]
Internal Standardization	Standardization
Number of Clusters:	
Specification Method	Automatic
Maximum Number	10
Selection Criterion	
Clustering Method	Ward
Preliminary Maximum	50

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.

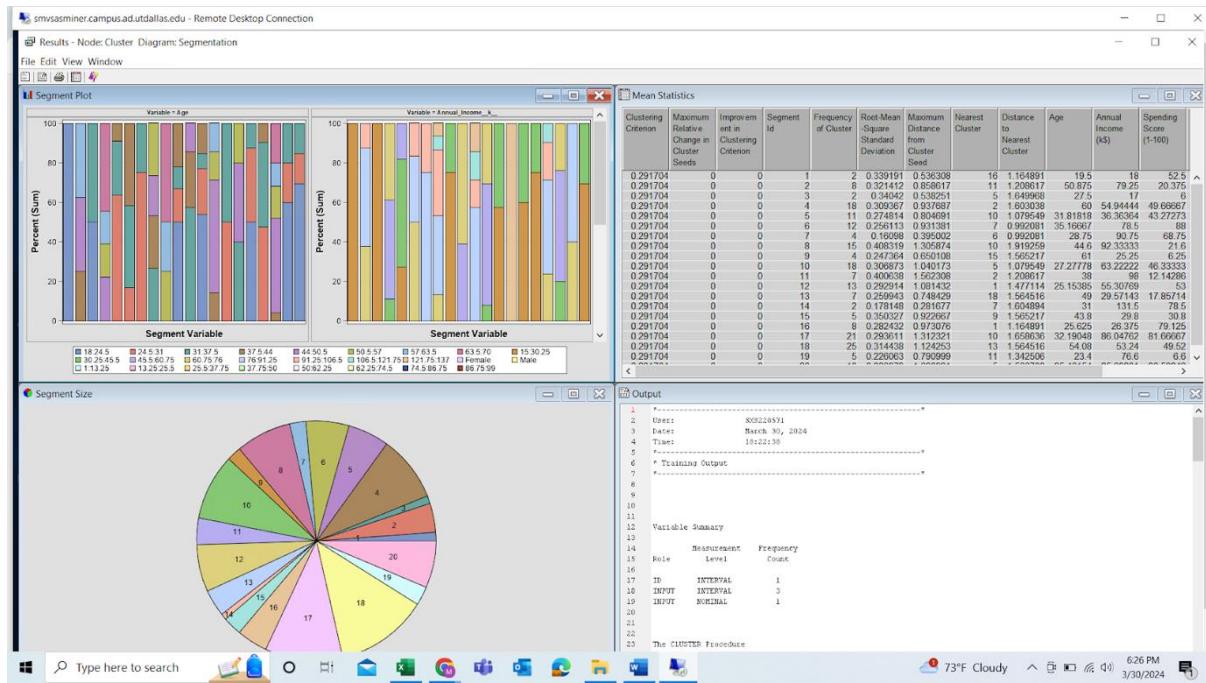


Figure 21

#### 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.

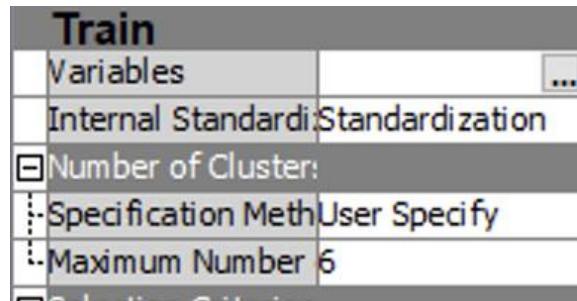


Figure 22

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.

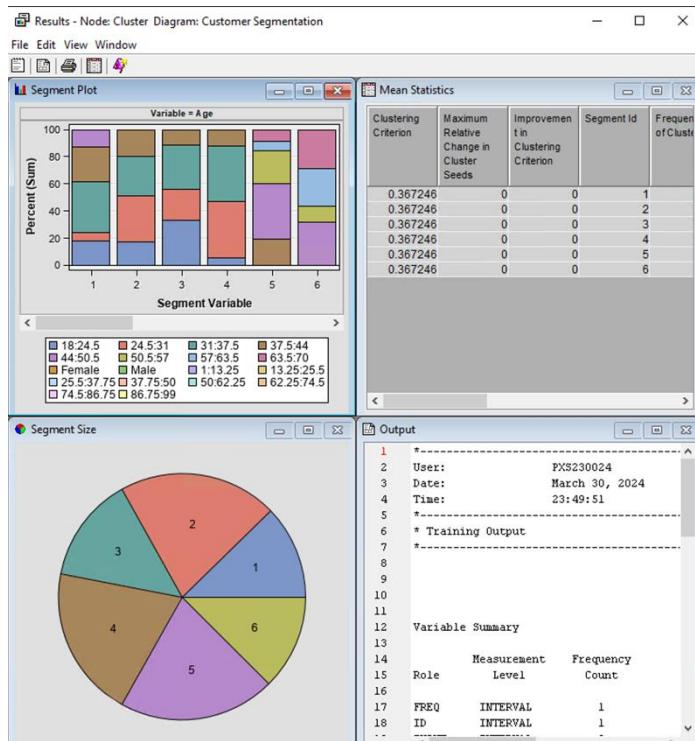


Figure 23

### 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.

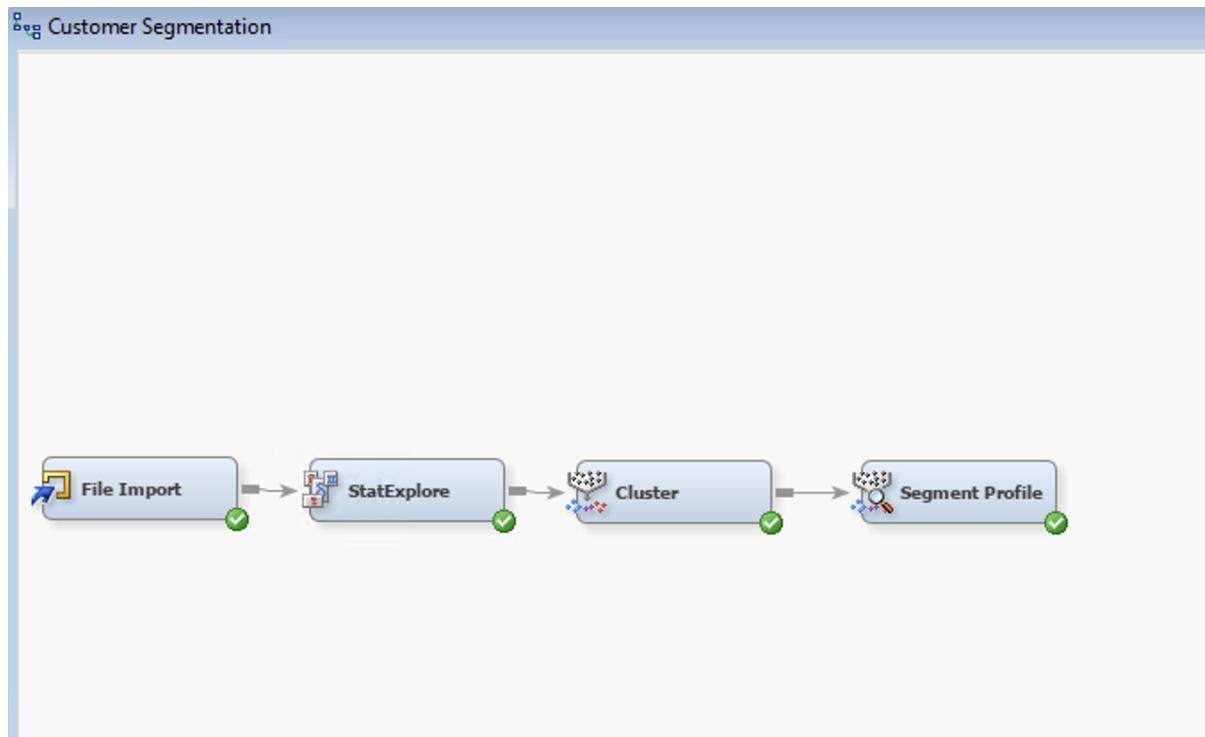


Figure 24

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

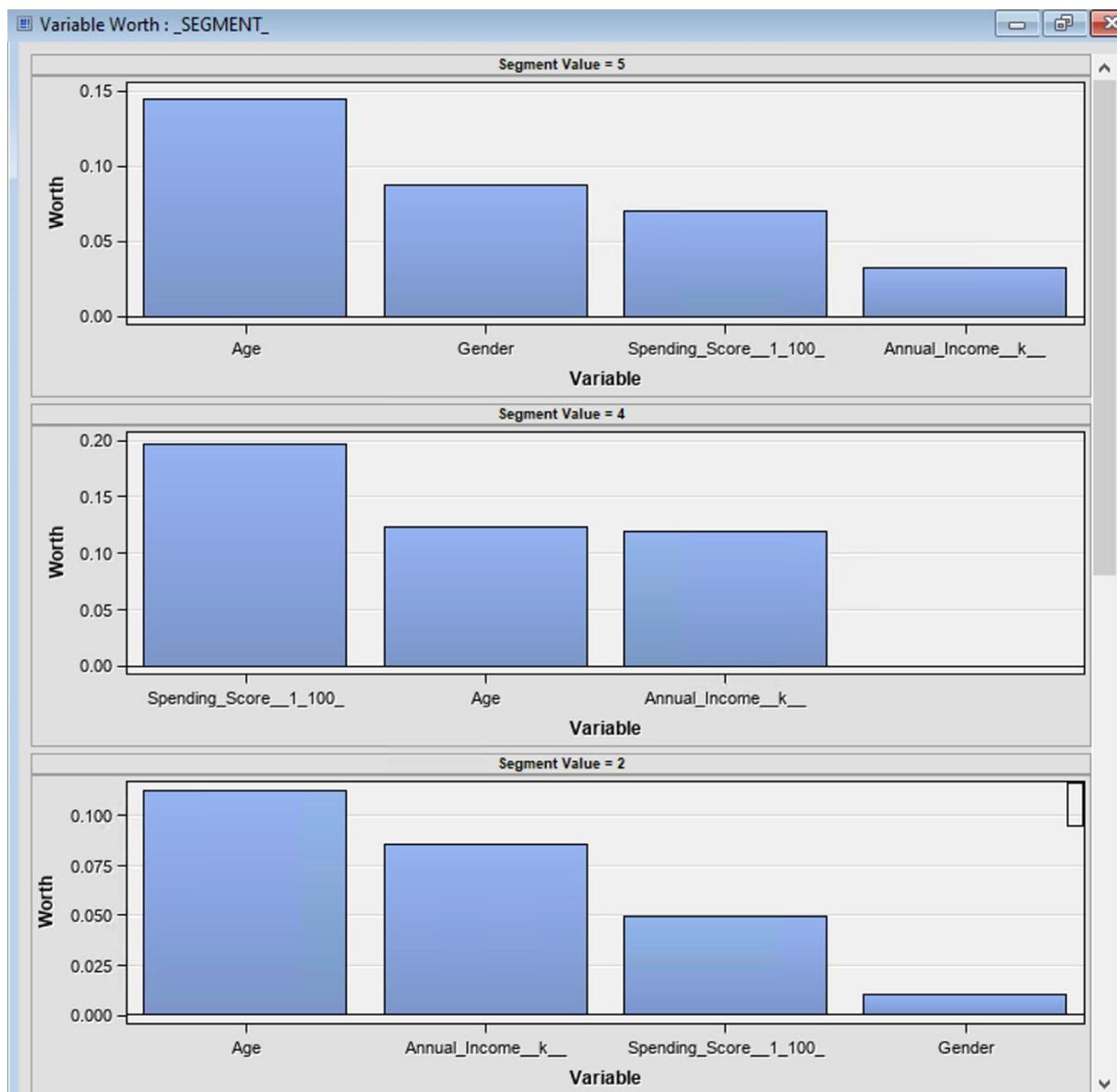


Figure 25

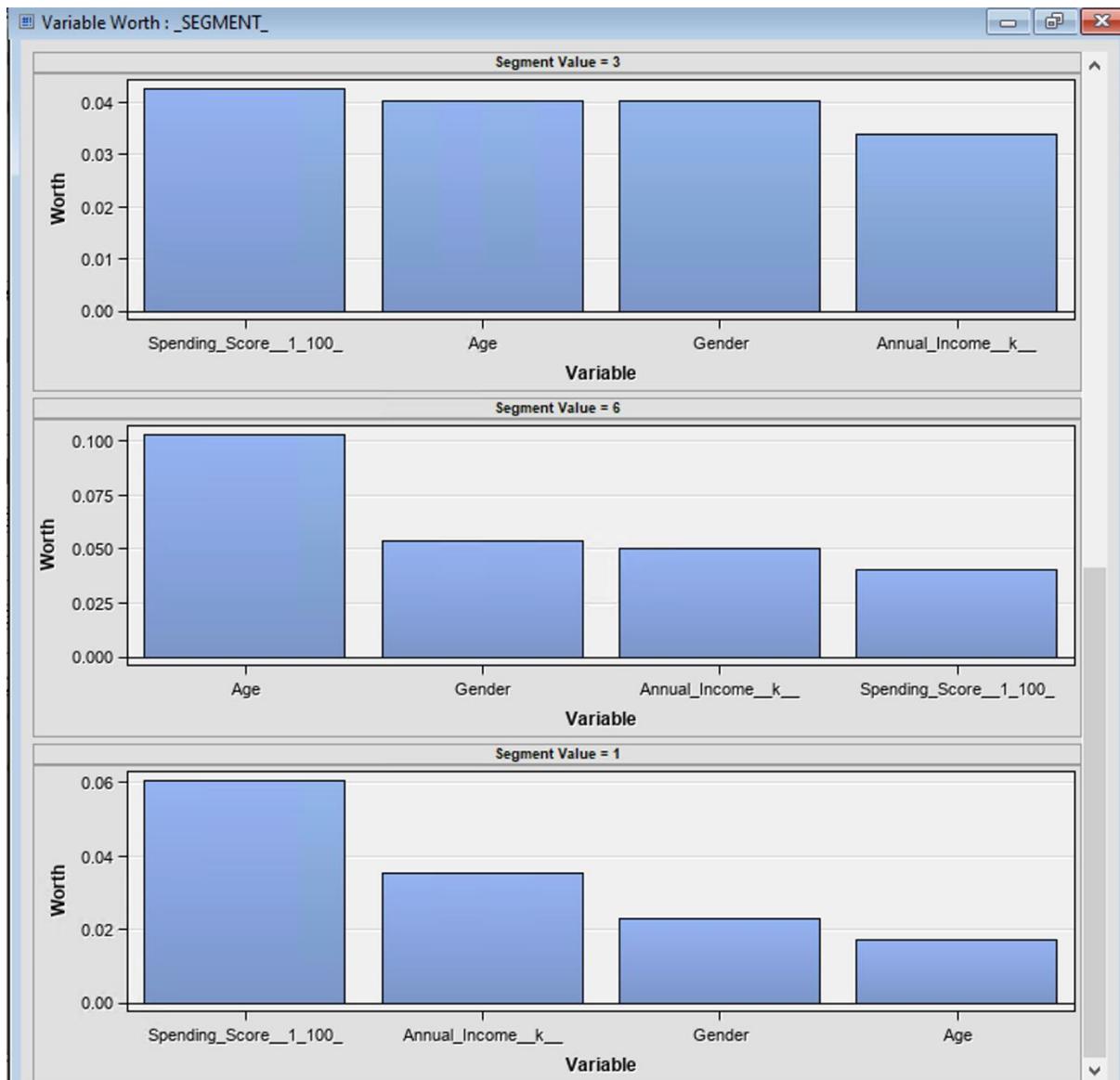


Figure 26

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**

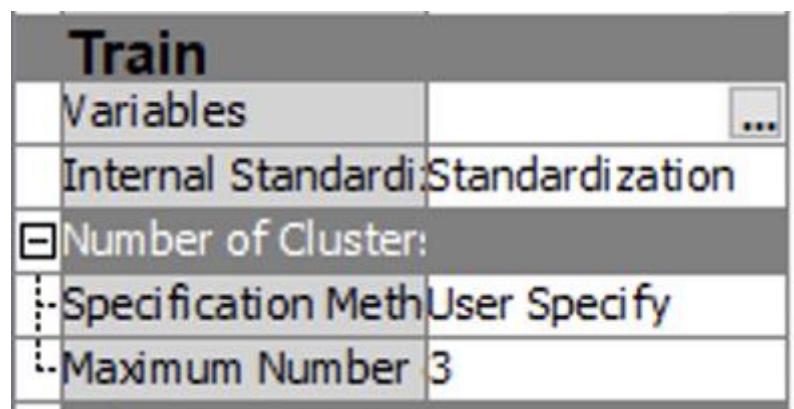


Figure 27

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.

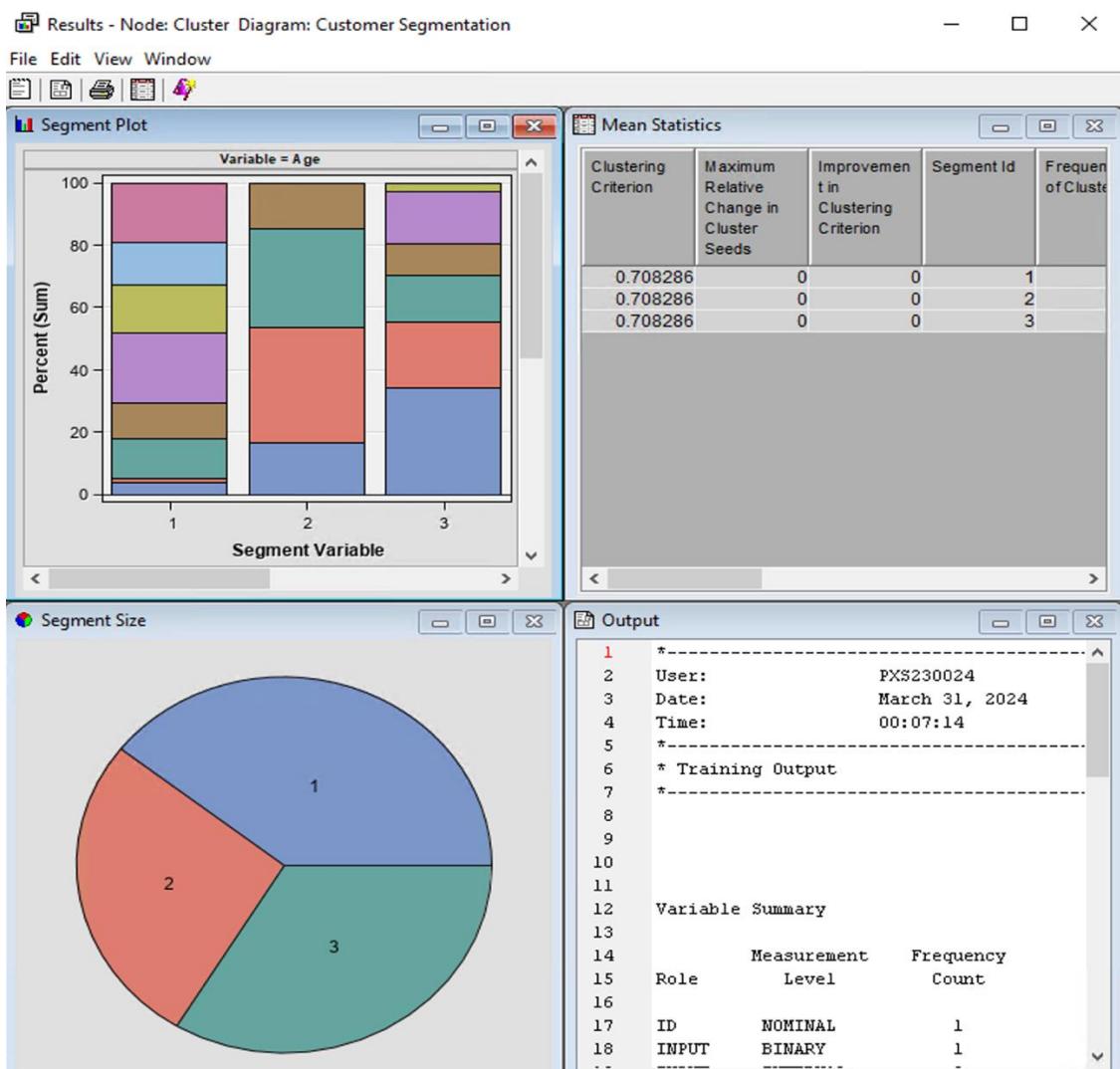


Figure 28

**Mean Statistics**

Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster
0.708286	0	0	1	79
0.708286	0	0	2	54
0.708286	0	0	3	67

Figure 29

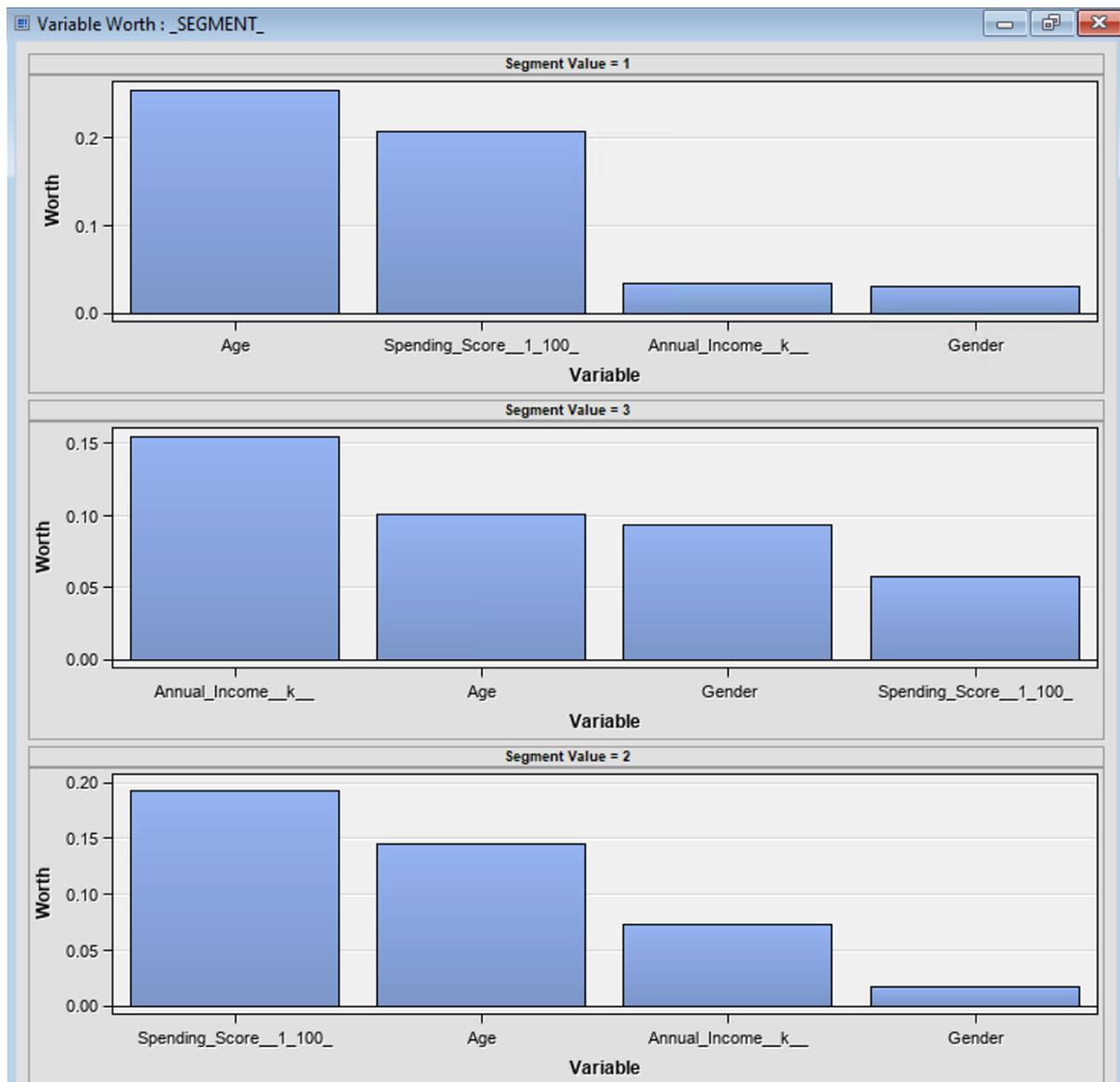


Figure 30



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).

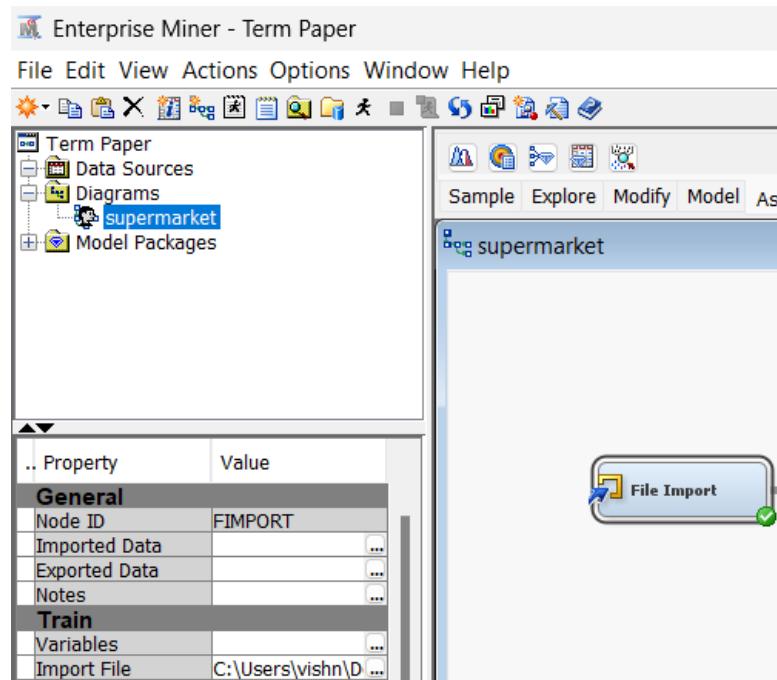


Figure 33

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

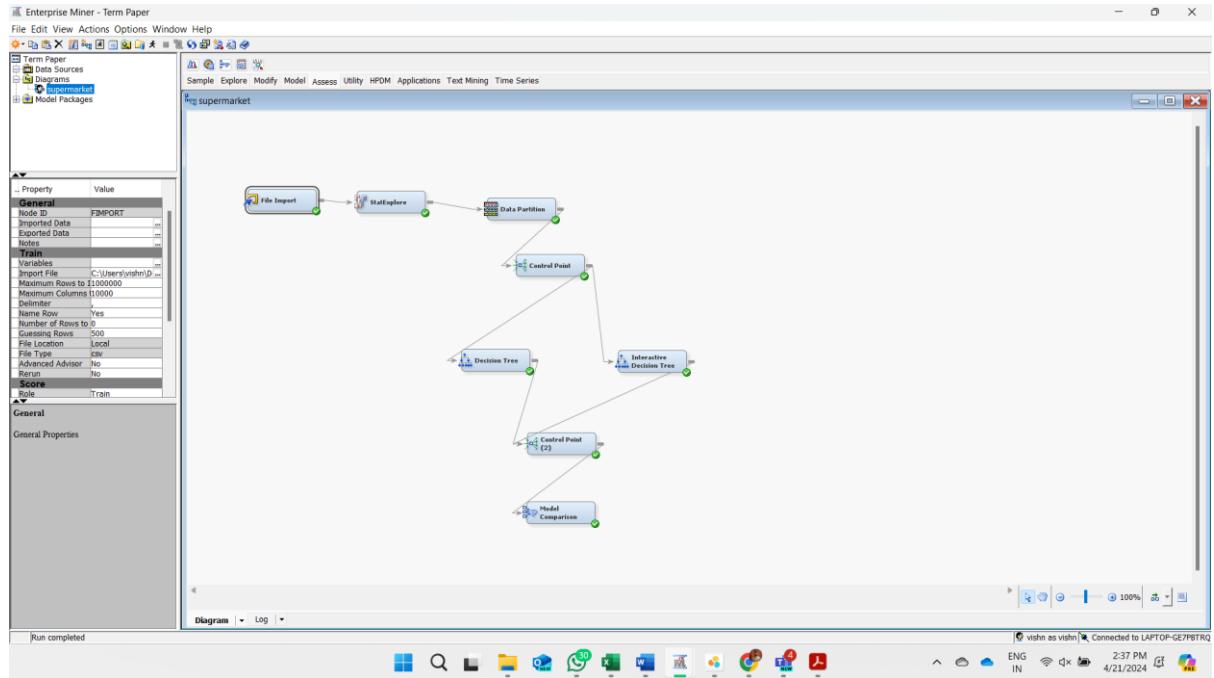


Figure 34

We look at the results:

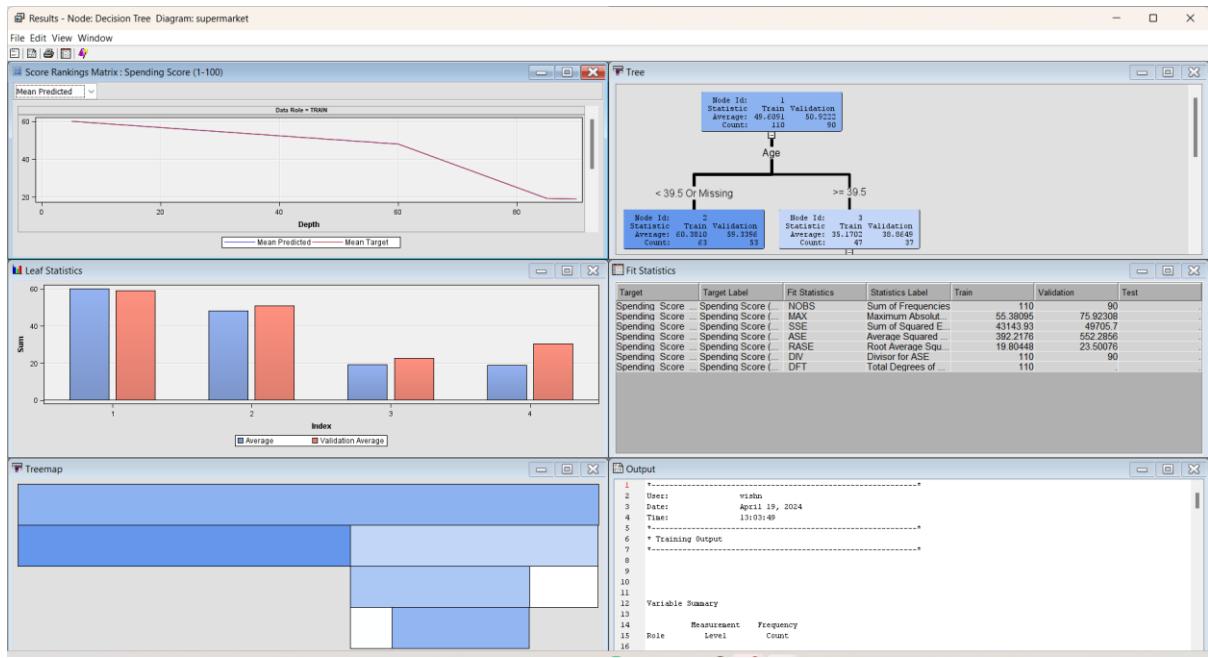


Figure 35

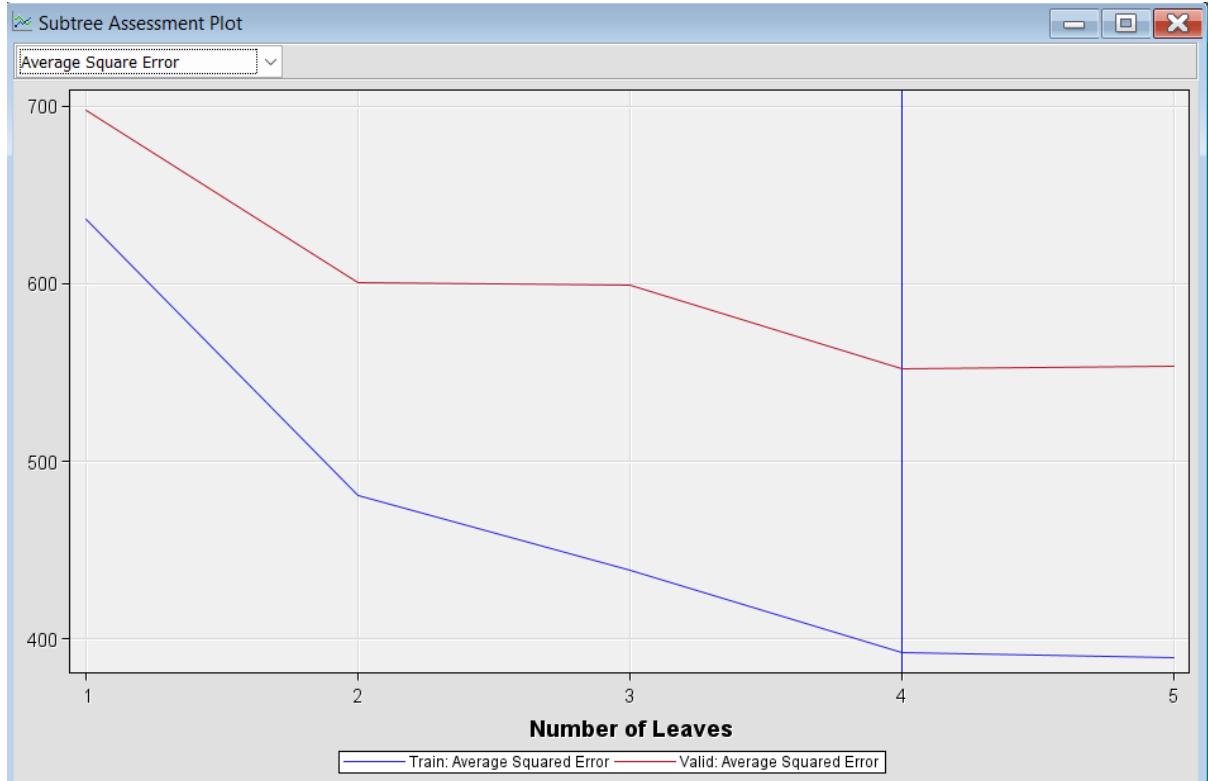


Figure 36

**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)	NodeS	Sum of Frequencies	110	90	
Spending Score 1 100	Spending Score (1-100)	MAV	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)	DFT	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.

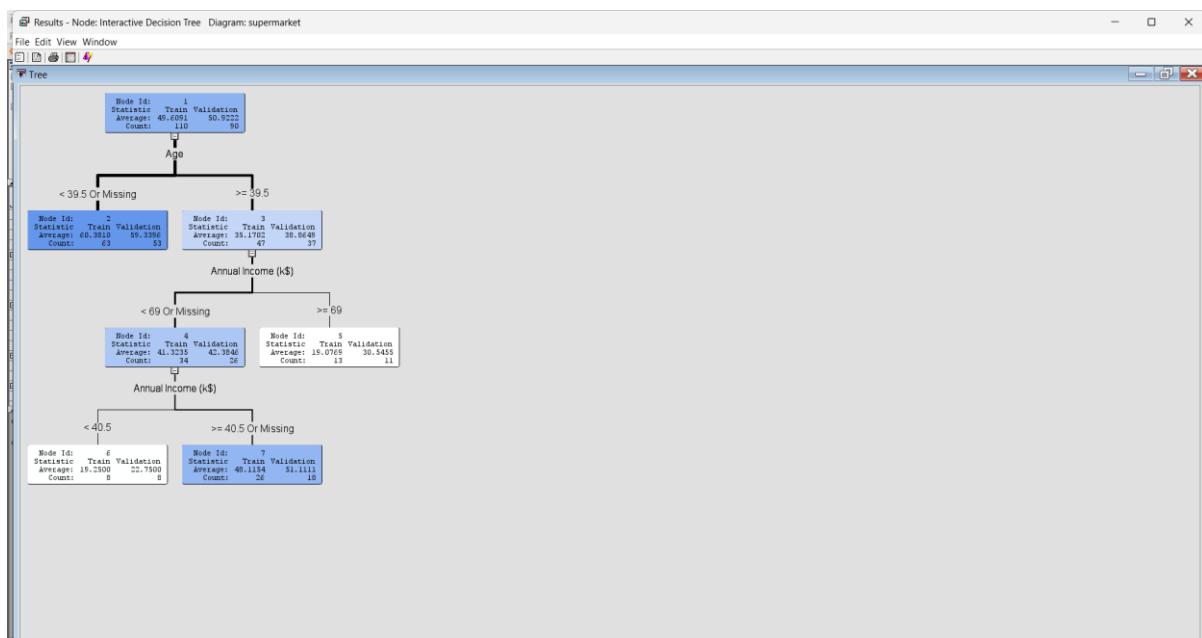


Figure 38

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.

44	Variable Importance					
45	Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
51	Age		1	1.0000	1.0000	1.0000
52	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.

Fit Statistics																			
Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion:	Train: Sum of Frequencies of Responses	Train: Average Error	Train: Squared Error	Train: Root Average Squared Error	Train: Degrees of Freedom	Valid: Sum of Frequencies	Valid: Maximum Absolute Error	Valid: Sum of Squared Errors	Valid: Root Average Squared Error	Valid: Degrees of Freedom	Valid: VASE		
Y	Tree	Tree2	Decision Tree	Spending ..	Spending ..	552 2856	110	55.38095	43143.93	392.2176	19.80448	110	110	90	75.92308	49705.7	552 2856	23.50076	90
			Interactive	Spending ..	Spending ..	552 2856	110	55.38095	43143.93	392.2176	19.80448	110	110	90	75.92308	49705.7	552 2856	23.50076	90

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)

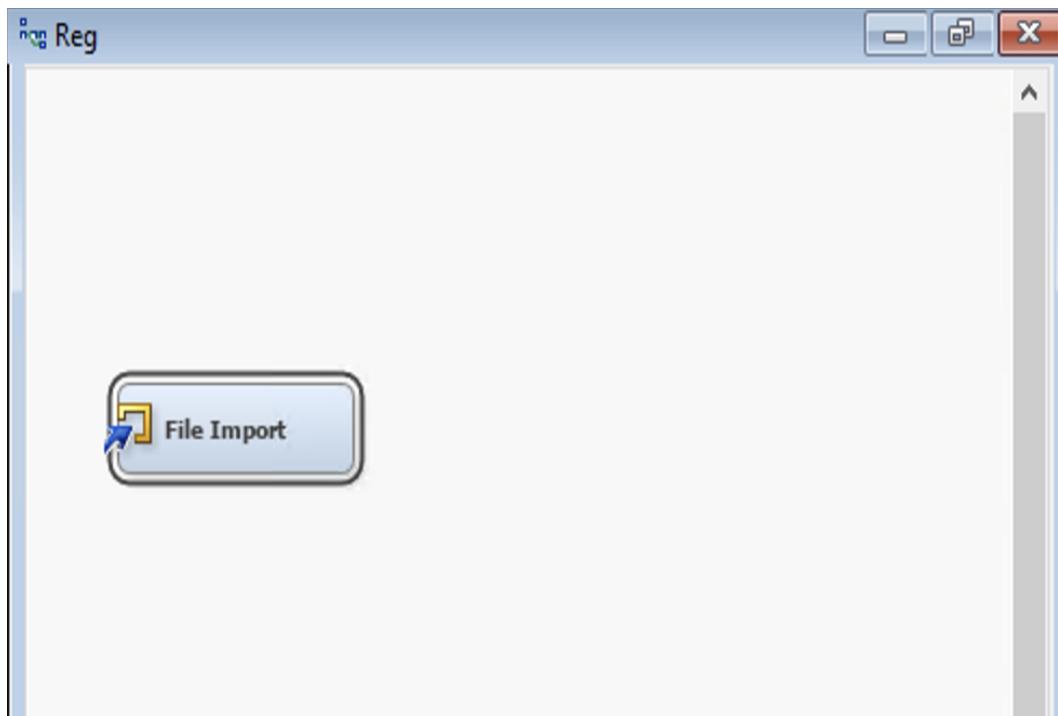


Figure 42

- 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.

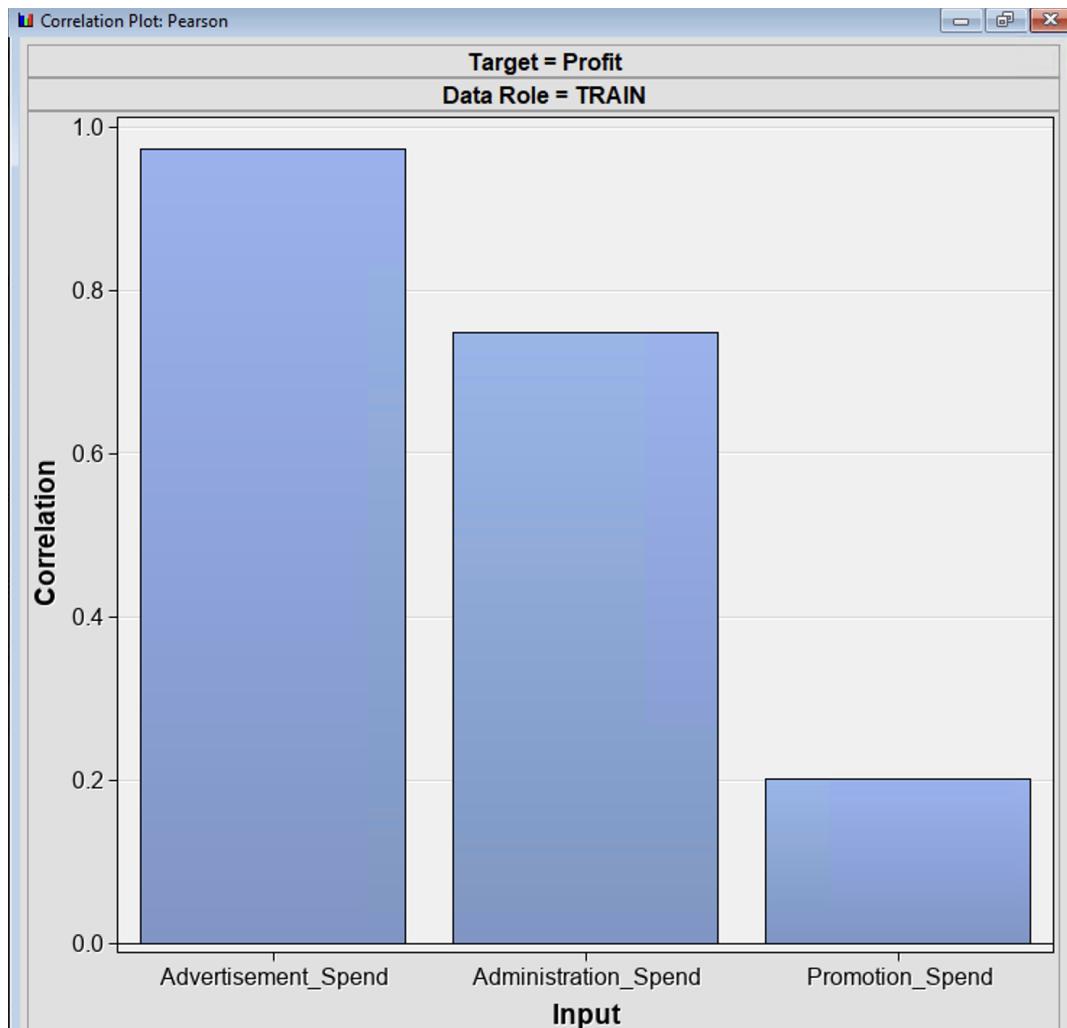


Figure 46

### 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.

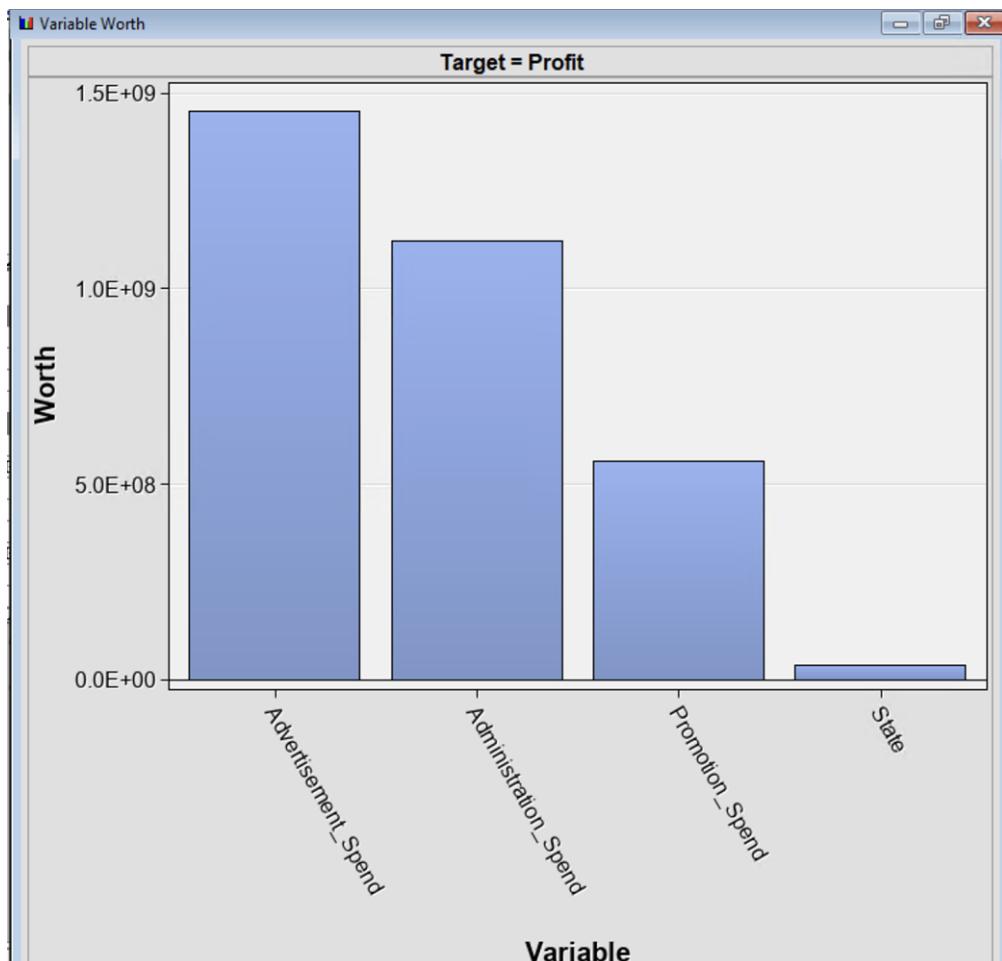


Figure 47

### 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.

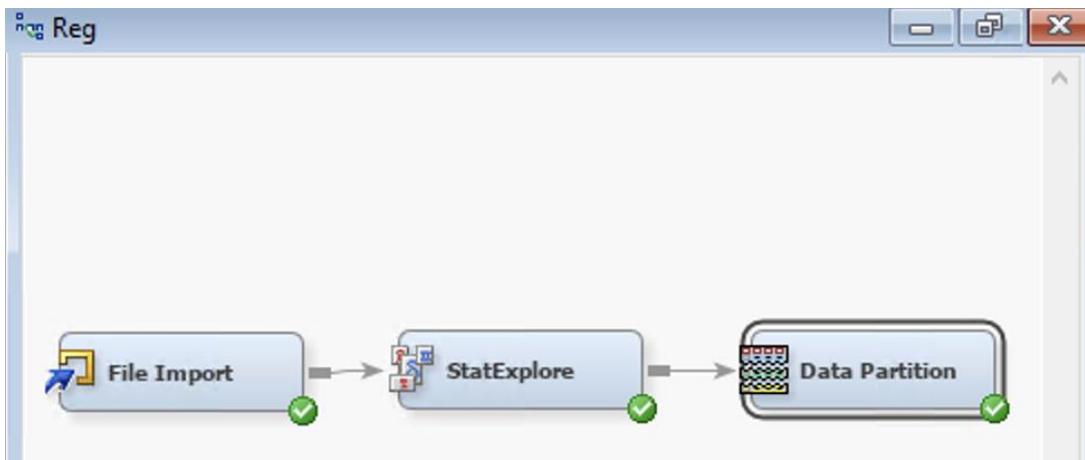


Figure 48

#### 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.

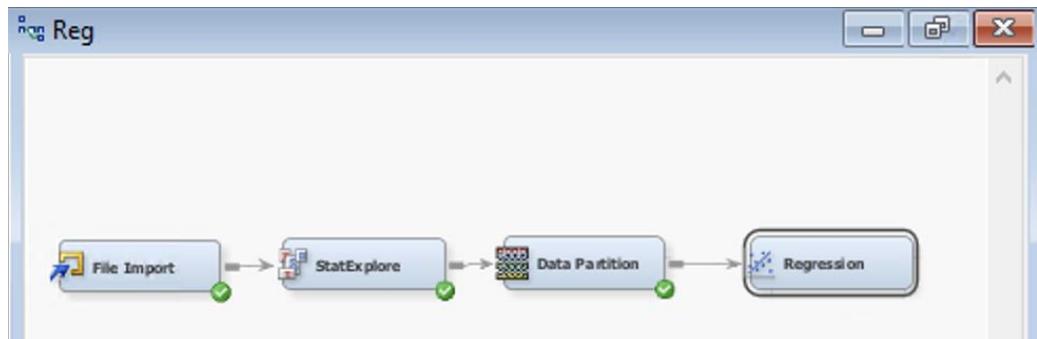


Figure 50

##### 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’.

.. Property	Value
-User Terms	No
-Term Editor	...
Class Targets	
-Regression Type	Linear Regressi...
-Link Function	Logit
Model Options	
-Suppress Intercept	No
-Input Coding	Deviation
Model Selection	
-Selection Model	None
-Selection Criterion	Default
-Use Selection Defaults	Yes
-Selection Options	...
Optimization Options	

Figure 51

Variables - Reg				
<input type="button" value="("/> <input )"="" type="button" value="none"/> <input type="checkbox"/> not <input type="button" value="Equal to"/> <input type="button" value="&lt;"/> <input type="button" value="&gt;"/>				
<input type="checkbox"/> Label <input type="checkbox"/> Mining <input type="checkbox"/> Basic				
Name	Use	Report	Role	Level
Administration_Default	<b>Default</b>	<b>No</b>	Input	Interval
Advertisement_Default	<b>Default</b>	<b>No</b>	Input	Interval
Profit	<b>Yes</b>	<b>No</b>	Target	Interval
Promotion_Spen	<b>Default</b>	<b>No</b>	Input	Interval
State	<b>Default</b>	<b>No</b>	Input	Nominal

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.

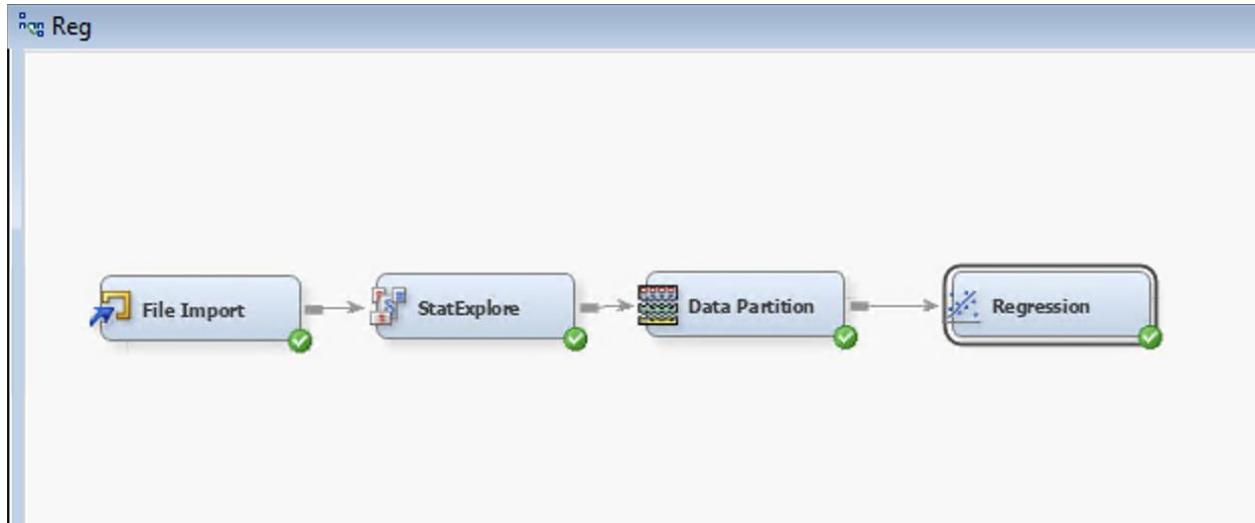


Figure 53

## 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.

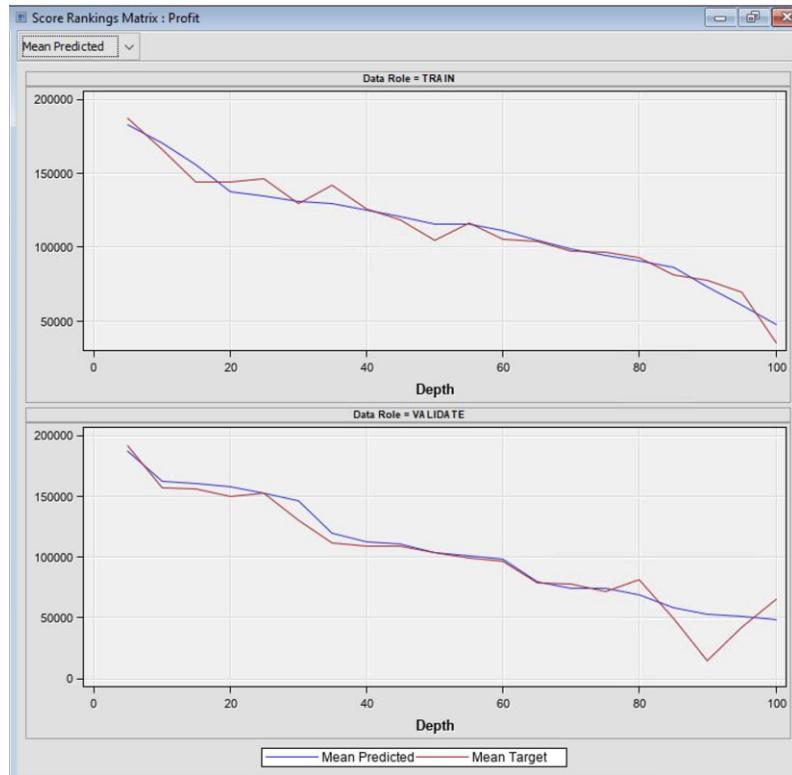


Figure 54

### 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.

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.

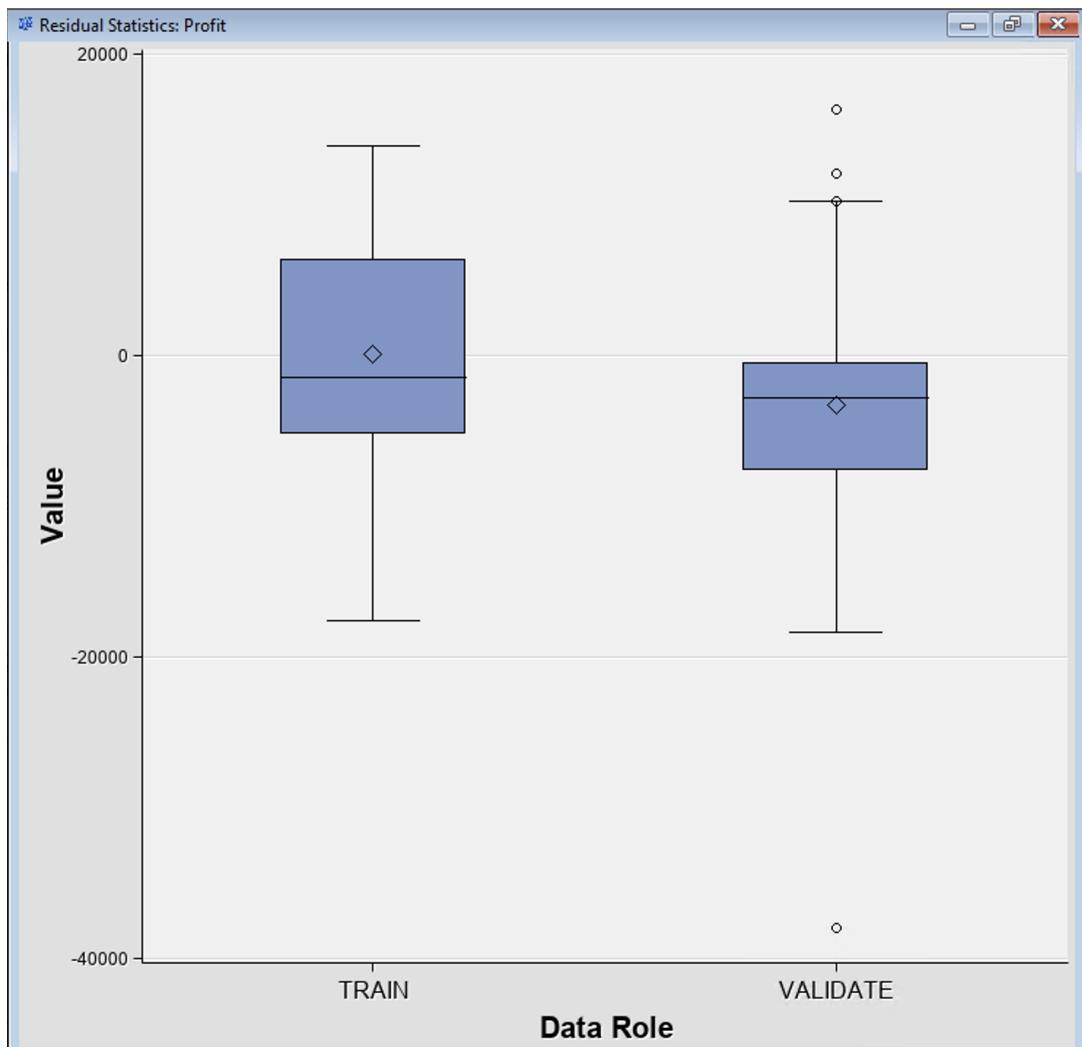


Figure 57

### 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

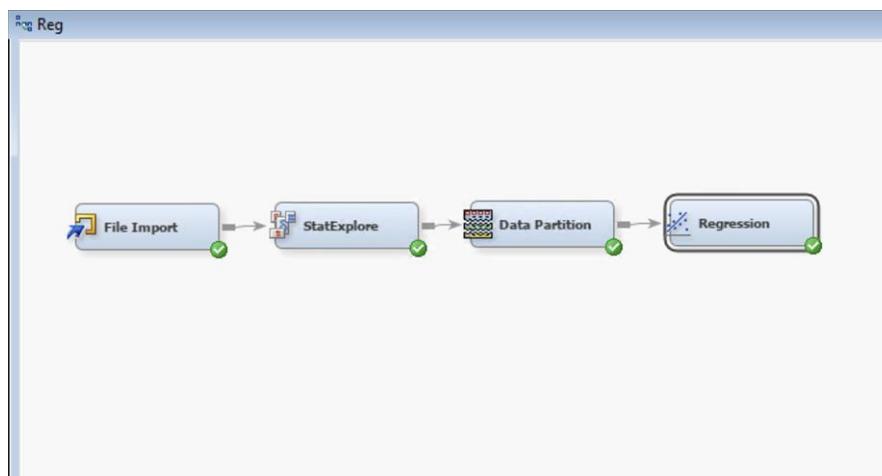


Figure 58

#### 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

Variables - FIMPORT

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Spending_Score	Input	Interval	No	No	No	.	.
Gender	Input	Nominal	No	No	No	.	.
CustomerID	ID	Nominal	No	No	No	.	.
Annual_Income	Input	Interval	No	No	No	.	.
Age	Input	Interval	No	No	No	.	.

Figure 60

### Step 2: Add the Graph Explore node

- Click the Explore tab and find Graph Explore node

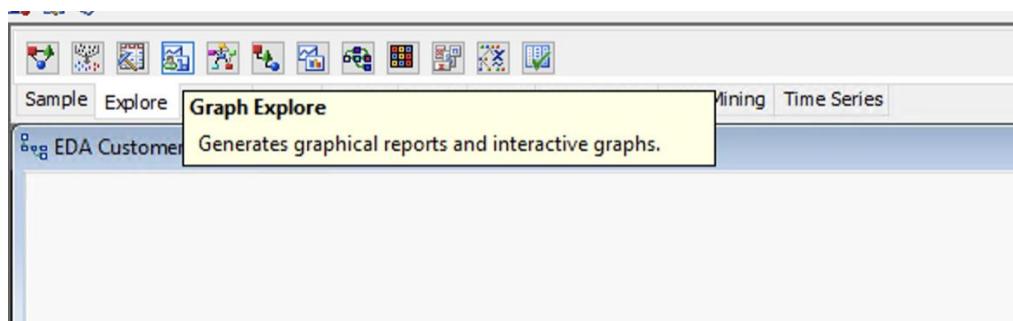


Figure 61

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

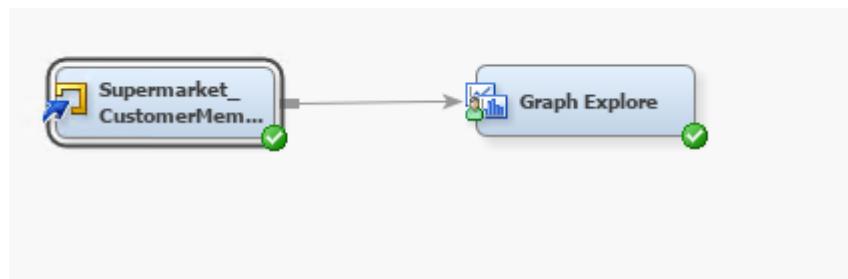


Figure 62

### Step 3: Run the diagram.

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

Results - Node: Graph Explore Diagram: EDA Customer

File Edit View Window

Sample Table

CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)
1Male		19	15	39
2Male		21	15	81
3Female		20	16	6
4Female		23	16	77
5Female		31	17	40
6Female		22	17	76
7Female		35	18	6
8Female		23	18	94
9Male		64	19	3
10Female		30	19	72
11Male		67	19	14
12Female		35	19	99
13Female		58	20	15
14Female		24	20	77
15Male		37	20	13

Output

```

1 -----
2 User: KGM220108
3 Date: April 20, 2024
4 Time: 01:04:53
5 -----
6 * Training Output
7 -----
8
9
10
11
12 Variable Summary
13
14      Measurement   Frequency
15 Role      Level      Count
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
279
280
281
282
283
284
285
286
287
288
289
289
290
291
292
293
294
295
296
297
298
299
299
300
301
302
303
304
305
306
307
308
309
309
310
311
312
313
314
315
316
317
318
319
319
320
321
322
323
324
325
326
327
328
329
329
330
331
332
333
334
335
336
337
338
339
339
340
341
342
343
344
345
346
347
348
349
349
350
351
352
353
354
355
356
357
358
359
359
360
361
362
363
364
365
366
367
368
369
369
370
371
372
373
374
375
376
377
378
378
379
380
381
382
383
384
385
386
387
388
389
389
390
391
392
393
394
395
396
397
398
399
399
400
401
402
403
404
405
406
407
408
409
409
410
411
412
413
414
415
416
417
418
419
419
420
421
422
423
424
425
426
427
428
429
429
430
431
432
433
434
435
436
437
438
439
439
440
441
442
443
444
445
446
447
448
449
449
450
451
452
453
454
455
456
457
458
459
459
460
461
462
463
464
465
466
467
468
469
469
470
471
472
473
474
475
476
477
478
478
479
480
481
482
483
484
485
486
487
488
489
489
490
491
492
493
494
495
496
497
498
498
499
499
500
501
502
503
504
505
506
507
508
509
509
510
511
512
513
514
515
516
517
518
519
519
520
521
522
523
524
525
526
527
528
529
529
530
531
532
533
534
535
536
537
538
539
539
540
541
542
543
544
545
546
547
548
549
549
550
551
552
553
554
555
556
557
558
559
559
560
561
562
563
564
565
566
567
568
569
569
570
571
572
573
574
575
576
577
578
578
579
580
581
582
583
584
585
586
587
588
589
589
590
591
592
593
594
595
596
597
598
598
599
599
600
601
602
603
604
605
606
607
608
609
609
610
611
612
613
614
615
616
617
618
619
619
620
621
622
623
624
625
626
627
628
629
629
630
631
632
633
634
635
636
637
638
639
639
640
641
642
643
644
645
646
647
648
649
649
650
651
652
653
654
655
656
657
658
659
659
660
661
662
663
664
665
666
667
668
669
669
670
671
672
673
674
675
676
677
678
678
679
680
681
682
683
684
685
686
687
688
688
689
689
690
691
692
693
694
695
696
697
697
698
698
699
699
700
701
702
703
704
705
706
707
708
709
709
710
711
712
713
714
715
716
717
718
719
719
720
721
722
723
724
725
726
727
728
729
729
730
731
732
733
734
735
736
737
738
739
739
740
741
742
743
744
745
746
747
748
749
749
750
751
752
753
754
755
756
757
758
759
759
760
761
762
763
764
765
766
767
768
769
769
770
771
772
773
774
775
776
777
778
778
779
779
780
781
782
783
784
785
786
787
787
788
788
789
789
790
791
792
793
794
795
796
796
797
797
798
798
799
799
800
801
802
803
804
805
806
807
808
809
809
810
811
812
813
814
815
816
817
818
818
819
819
820
821
822
823
824
825
826
827
828
829
829
830
831
832
833
834
835
836
837
838
839
839
840
841
842
843
844
845
846
847
848
849
849
850
851
852
853
854
855
856
857
858
859
859
860
861
862
863
864
865
866
867
868
869
869
870
871
872
873
874
875
876
877
878
878
879
879
880
881
882
883
884
885
886
887
888
888
889
889
890
891
892
893
894
895
896
897
897
898
898
899
899
900
901
902
903
904
905
906
907
908
909
909
910
911
912
913
914
915
916
917
918
918
919
919
920
921
922
923
924
925
926
927
928
929
929
930
931
932
933
934
935
936
937
938
939
939
940
941
942
943
944
945
946
947
948
948
949
949
950
951
952
953
954
955
956
957
958
959
959
960
961
962
963
964
965
966
967
968
968
969
969
970
971
972
973
974
975
976
977
978
978
979
979
980
981
982
983
984
985
986
987
987
988
988
989
989
990
991
992
993
994
995
996
997
997
998
998
999
999
1000
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1018
1019
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1028
1029
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1038
1039
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1048
1049
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1068
1069
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1078
1079
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1088
1089
1089
1090
1091
1092
1093
1094
1095
1096
1097
1097
1098
1098
1099
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1118
1119
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1128
1129
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1138
1139
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1148
1149
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1168
1169
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1178
1179
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1188
1189
1189
1190
1191
1192
1193
1194
1195
1196
1197
1197
1198
1198
1199
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1218
1219
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1228
1229
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1238
1239
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1248
1249
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1268
1269
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1278
1279
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1288
1289
1289
1290
1291
1292
1293
1294
1295
1296
1297
1297
1298
1298
1299
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1318
1319
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1328
1329
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1338
1339
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1348
1349
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1368
1369
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1378
1379
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1388
1389
1389
1390
1391
1392
1393
1394
1395
1396
1397
1397
1398
1398
1399
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1418
1419
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1428
1429
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1438
1439
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1448
1449
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1468
1469
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1478
1479
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1488
1489
1489
1490
1491
1492
1493
1494
1495
1496
1497
1497
1498
1498
1499
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1518
1519
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1528
1529
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1538
1539
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1548
1549
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1568
1569
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1578
1579
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1588
1589
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1598
1599
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1618
1619
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1628
1629
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1638
1639
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1648
1649
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1668
1669
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1678
1679
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1688
1689
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1698
1699
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1718
1719
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1728
1729
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1738
1739
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1748
1749
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1768
1769
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1778
1779
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1788
1789
1789
1790
1791
1792
1793
1794
1795
1796
1797
1797
1798
1798
1799
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1818
1819
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1828
1829
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1838
1839
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1848
1849
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1878
1879
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1888
1889
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1898
1899
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1918
1919
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1928
1929
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1938
1939
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1948
1949
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1978
1979
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1988
1989
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1998
1999
1999
2000
2001
2002
2003
2004
200
```

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

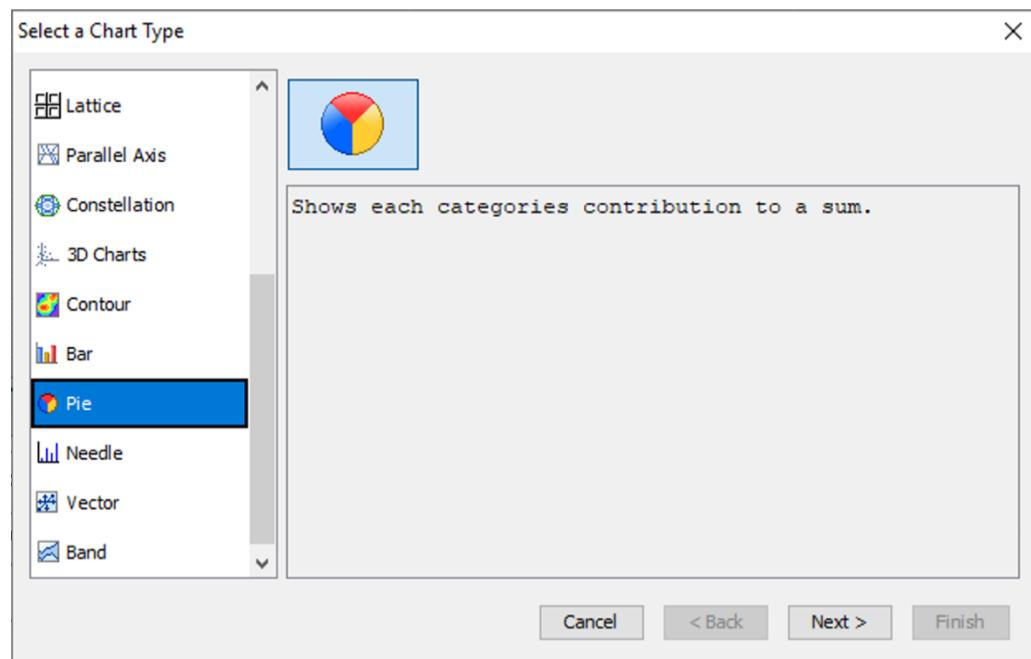


Figure 65

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

Select Chart Roles				
<input type="checkbox"/> Use default assignments				
▲ Variable	Role	Type	Description	Format
Age		Numeric	Age	BEST12.
Annual_Income_k_		Numeric	Annual Income (k\$)	
CustomerID		Numeric	CustomerID	BEST12.
Gender	Category	Character	Gender	\$6.
Spending_Score_1_...		Numeric	Spending Score (1-100)	

Allow multiple role assignments

Cancel < Back Next > Finish

Figure 66

- c. There is no where clause, click Next

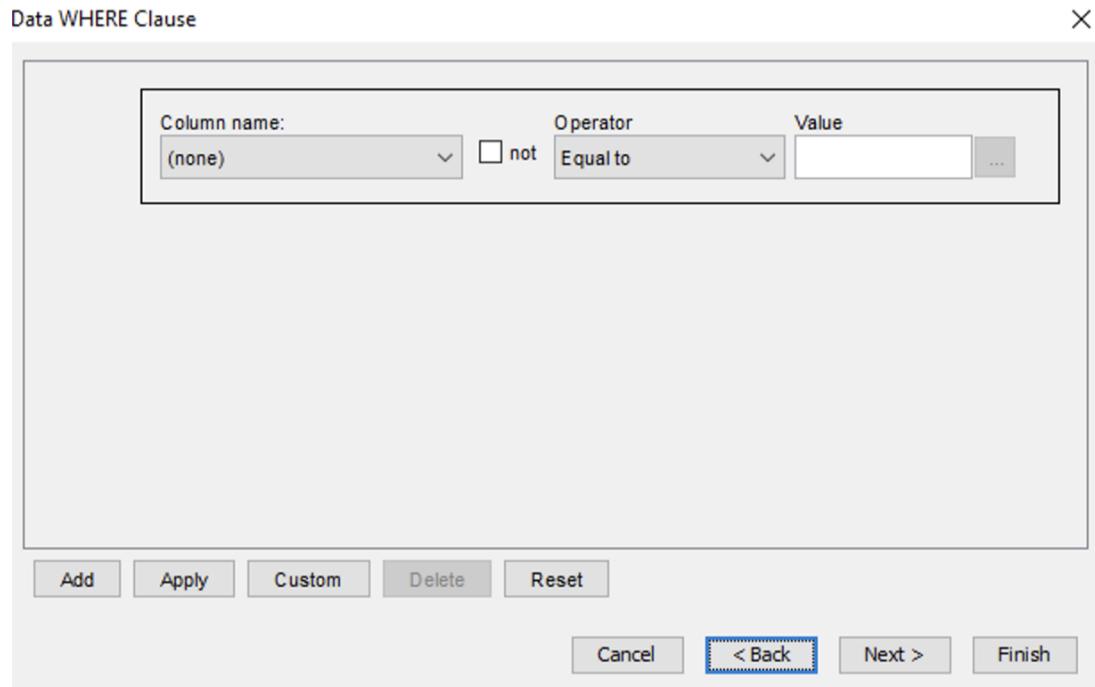


Figure 67

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

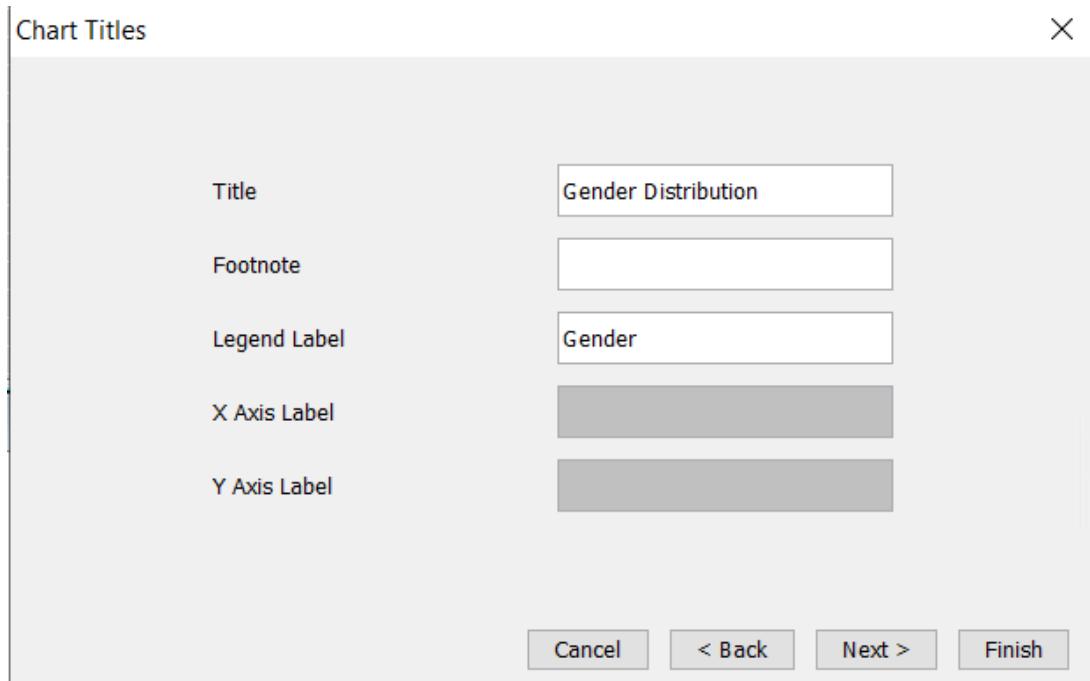


Figure 68

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

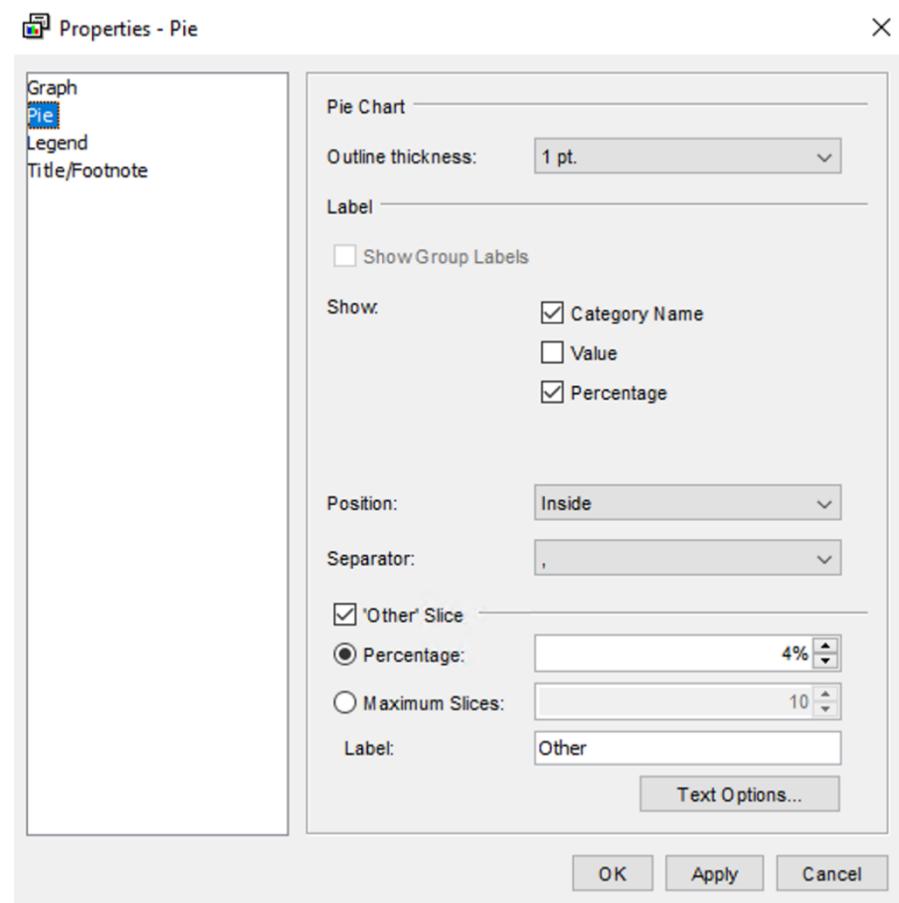


Figure 69

f. The Below graph is developed by SAS EM

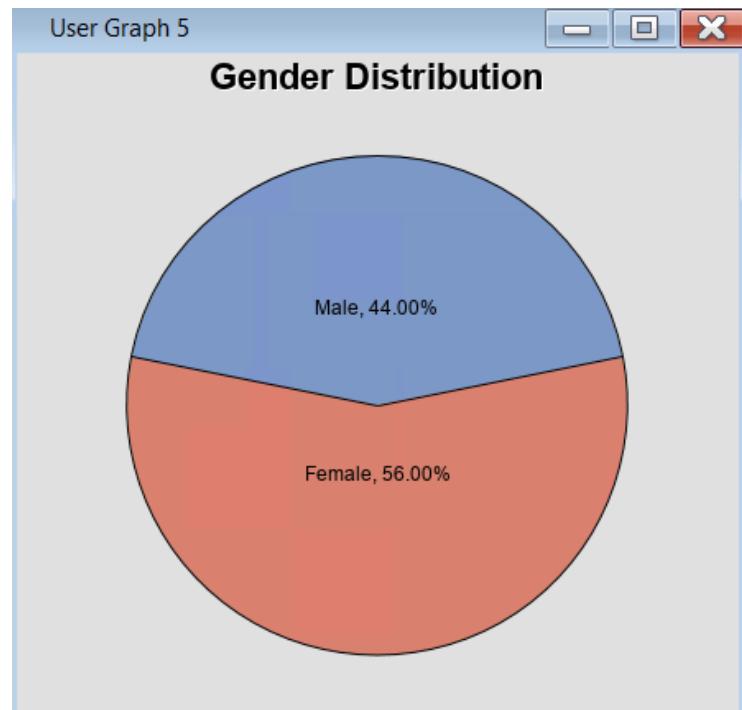


Figure 70

## 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.

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

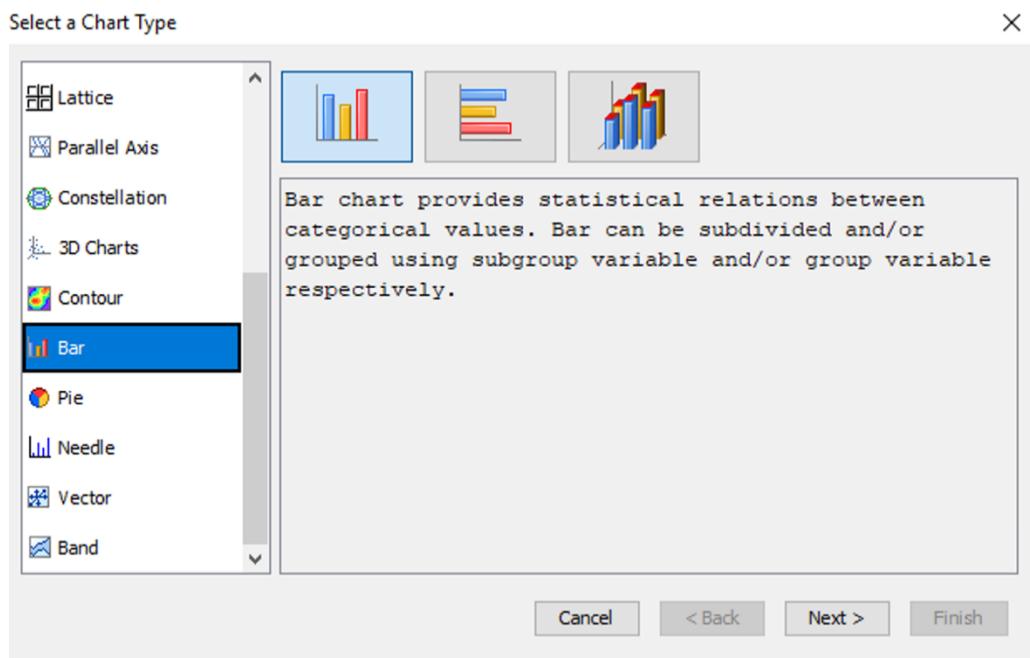


Figure 71

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

Select Chart Roles

▲ Variable	Role	Type	Description	Format
Age	Category	Numeric	Age	BEST12.
Annual_Income_k_		Numeric	Annual Income (k\$)	
CustomerID		Numeric	CustomerID	BEST12.
Gender		Character	Gender	\$6.
Spending_Score_1_...		Numeric	Spending Score (1-100)	

Use default assignments

Response statistic: Frequency

Allow multiple role assignments

Cancel < Back Next > Finish

Figure 72

- c. There is no where clause, click Next.

Data WHERE Clause

Column name:	(none)	Operator	Value
		not Equal to	

Add Apply Custom Delete Reset

Cancel < Back Next > Finish

Figure 73

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

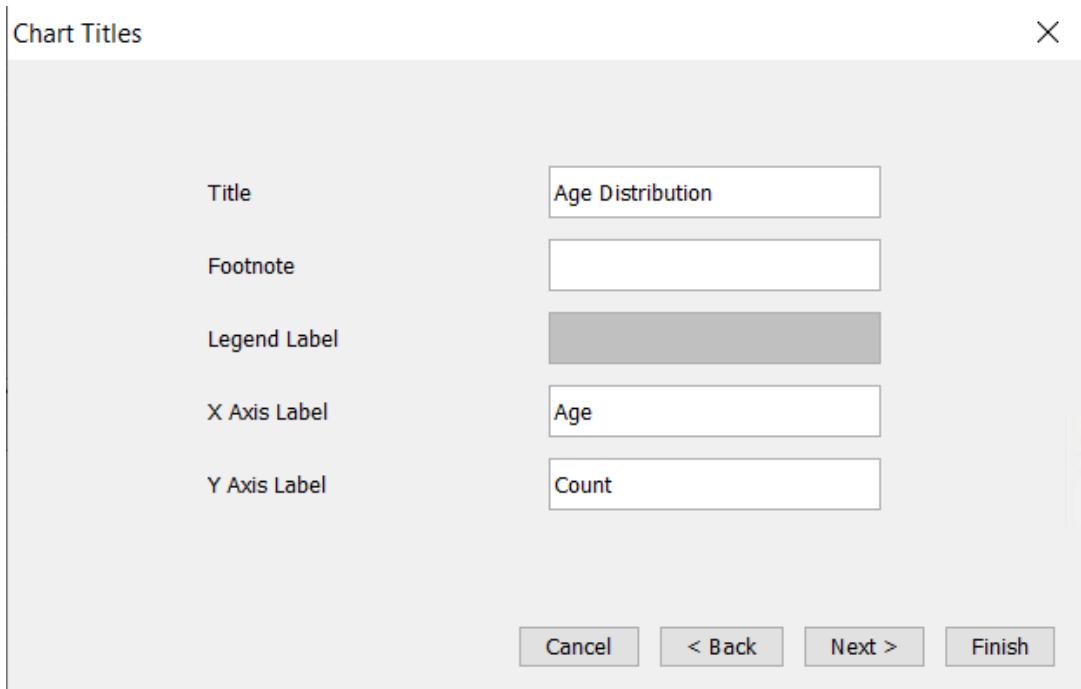


Figure 74

e. The Below graph is developed by SAS EM

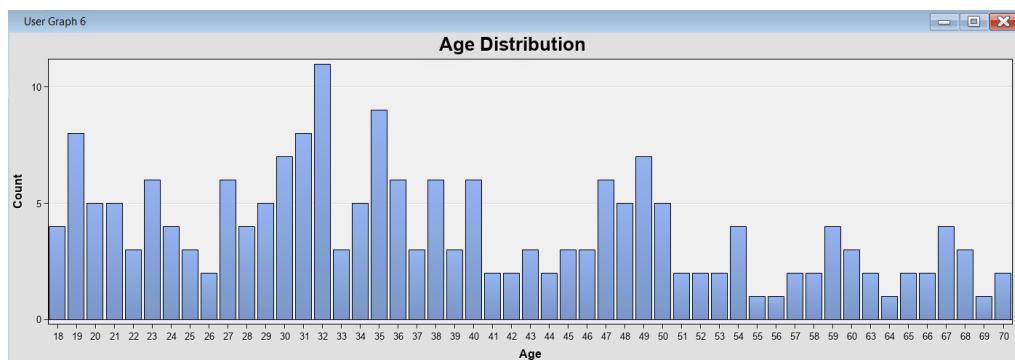


Figure 75

### 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

- Select the chart type as Bar and click Next.

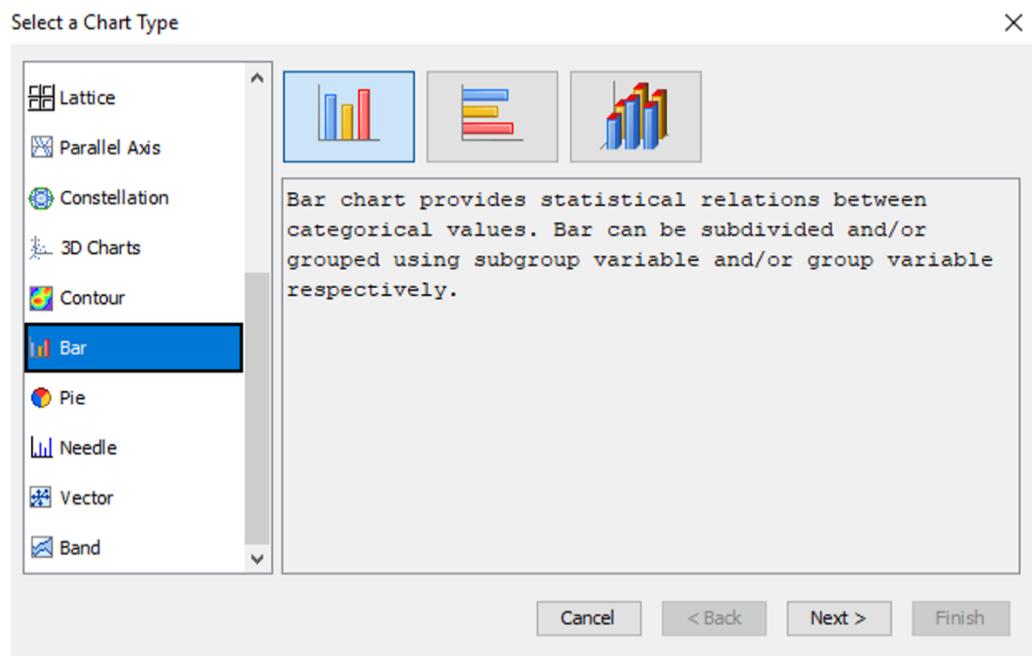


Figure 76

- Set the chart roles by setting a category variable & the Frequency variable. We set Age as category & Annual\_Income\_k as Frequency. Click Next.

Select Chart Roles

▲ Variable	Role	Type	Description	Format
Age	Category	Numeric	Age	BEST12.
Annual_Income_k	Frequency	Numeric	Annual Income (k\$)	
CustomerID		Numeric	CustomerID	BEST12.
Gender		Character	Gender	\$6.
Spending_Score_1...		Numeric	Spending Score (1-100)	

Response statistic: Frequency

Allow multiple role assignments

Cancel  < Back  Next >  Finish

Figure 77

- c. There is no where clause, click Next

Data WHERE Clause

Column name:	(none)	Operator	Value
		not	Equal to

Add  Apply  Custom  Delete  Reset

Cancel  < Back  Next >  Finish

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

Chart Titles

Title	Age vs Annual Income Plot
Footnote	
Legend Label	
X Axis Label	Age
Y Axis Label	Annual Income Sum (in K)

Cancel < Back Next > Finish

Figure 79

e. The Below graph is developed by SAS EM

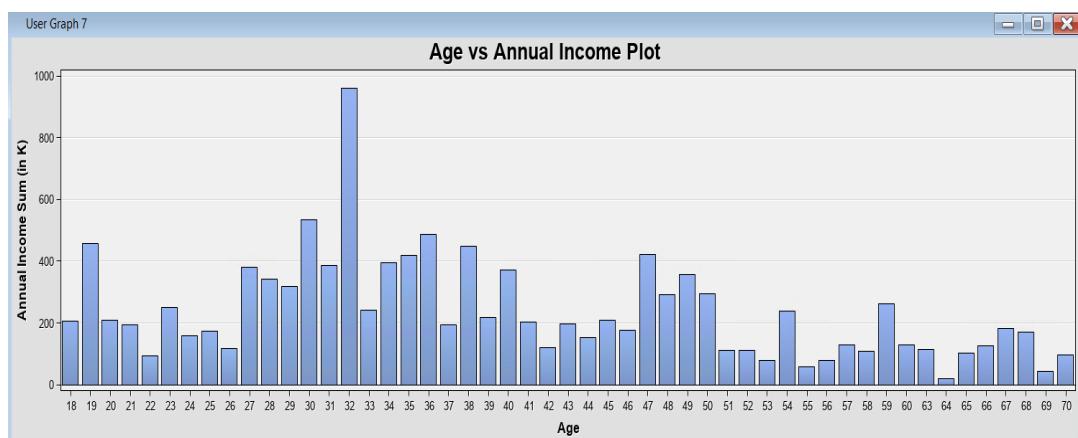


Figure 80

### 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

- Select the chart type as Bar and click Next

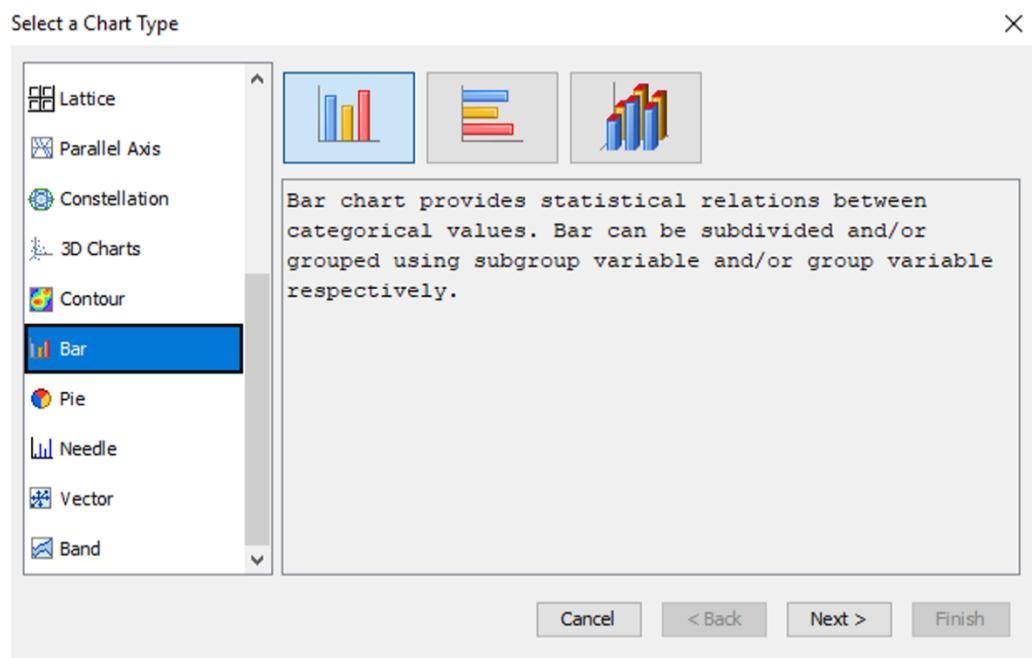


Figure 81

- Set the chart roles by setting a category variable & the Frequency variable. We set Age as category & Spending\_Score as Frequency. Click Next.

### Select Chart Roles

X

Variable	Role	Type	Description	Format
Age	Category	Numeric	Age	BEST12.
Annual_Income_k		Numeric	Annual Income (k\$)	
CustomerID		Numeric	CustomerID	BEST12.
Gender		Character	Gender	\$6.
Spending_Score_1...	Frequency	Numeric	Spending Score (1-100)	

Response statistic: Frequency

Allow multiple role assignments

Cancel < Back Next > Finish

Figure 82

- c. There is no where clause, click Next

Data WHERE Clause

X

Column name:	(none)	Operator	Value
		not Equal to	

Add Apply Custom Delete Reset Cancel < Back Next > Finish

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

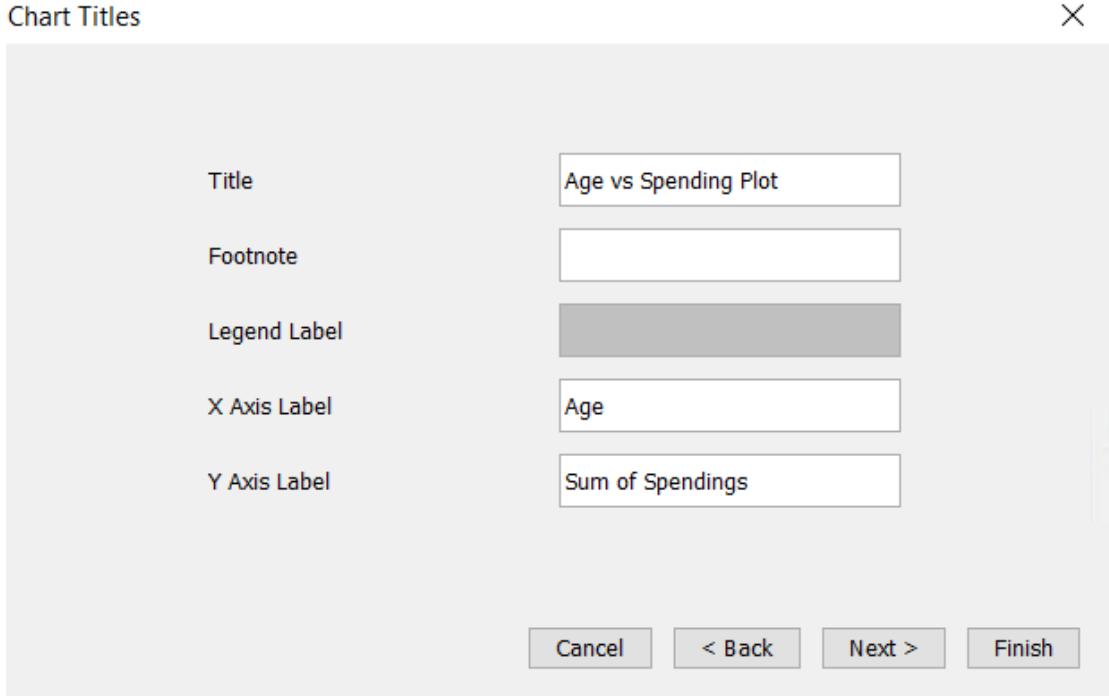


Figure 84

e. The Below graph is developed by SAS EM

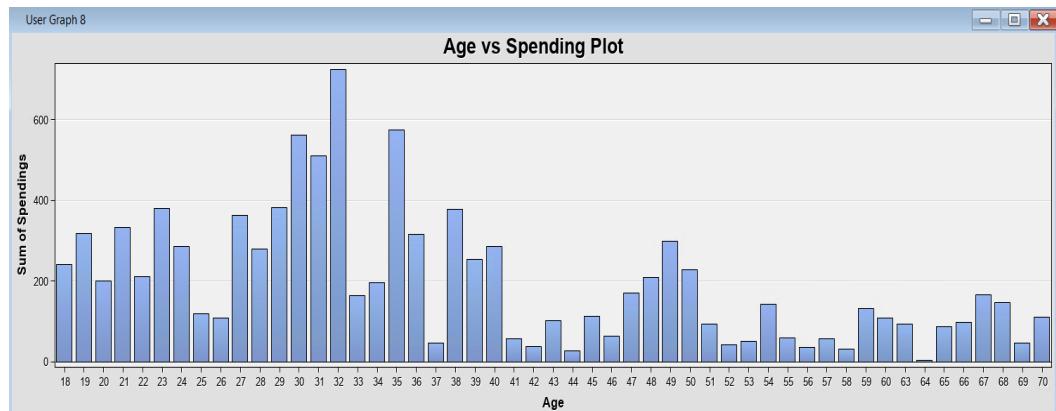


Figure 85

### 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 branches spread across the three 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 stores located in the three states 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
30	66651.60	100015.50	110110.00	Florida	100000.00

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 dataset. By this step the file import is complete.

The screenshot shows the 'Variables - FIMPORT2' dialog box. At the top, there is a search bar with dropdown menus for 'Label', 'not', 'Equal to', and a text input field. Below this is a 'Columns:' section with a 'Label' checkbox. To the right are 'Mining' and 'Basic' checkboxes. The main area is a table with columns: Name, Role, Level, Report, Order, Drop, Lower Limit, and Upper Limit. The table rows are:

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

- Click the Explore tab and find Graph Explore node

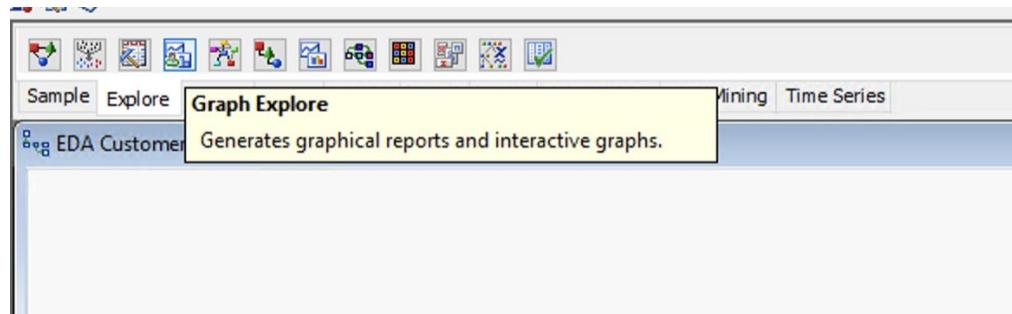


Figure 88

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

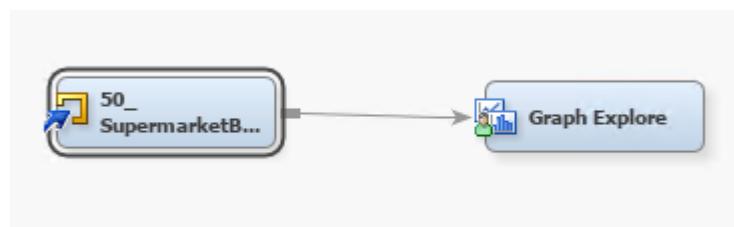


Figure 89

### Step 3: Run the diagram

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

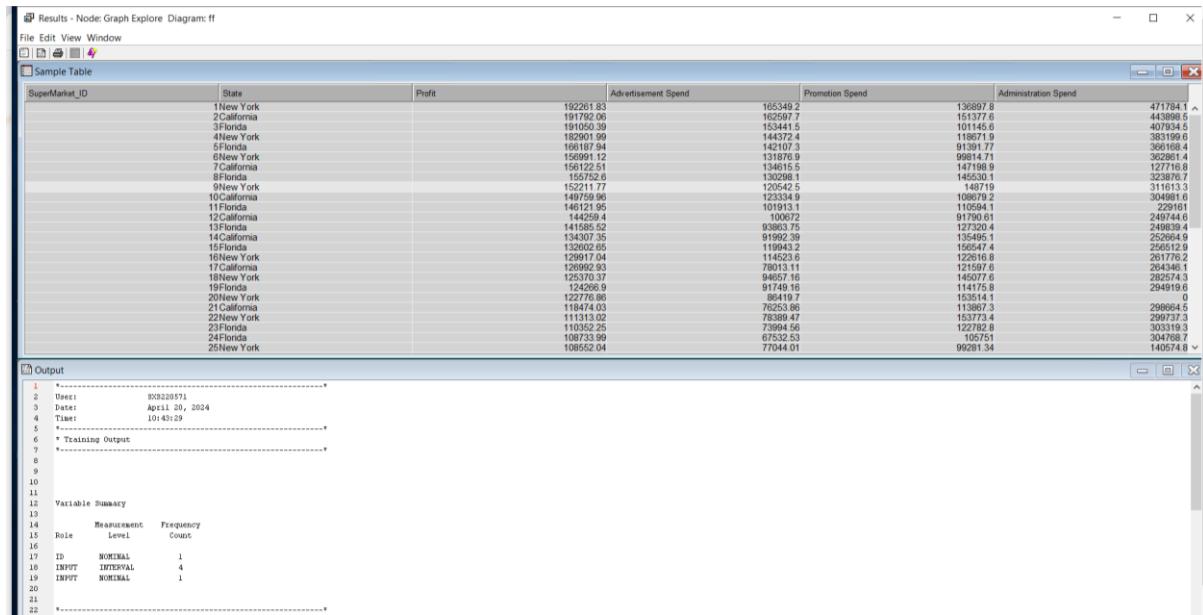


Figure 90

b. Click on View -> Plot

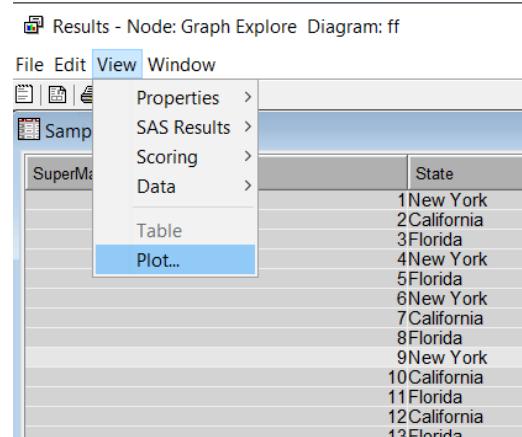


Figure 91

**Step 4: Explore the Branch distribution.**

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

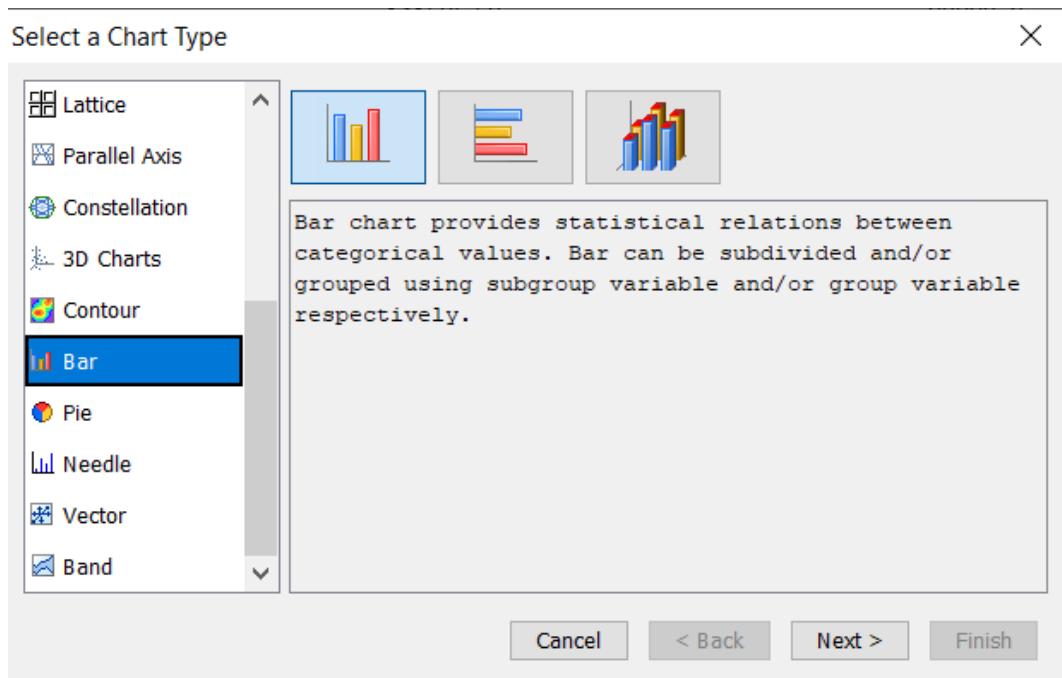


Figure 92

b. Select State as category and click next

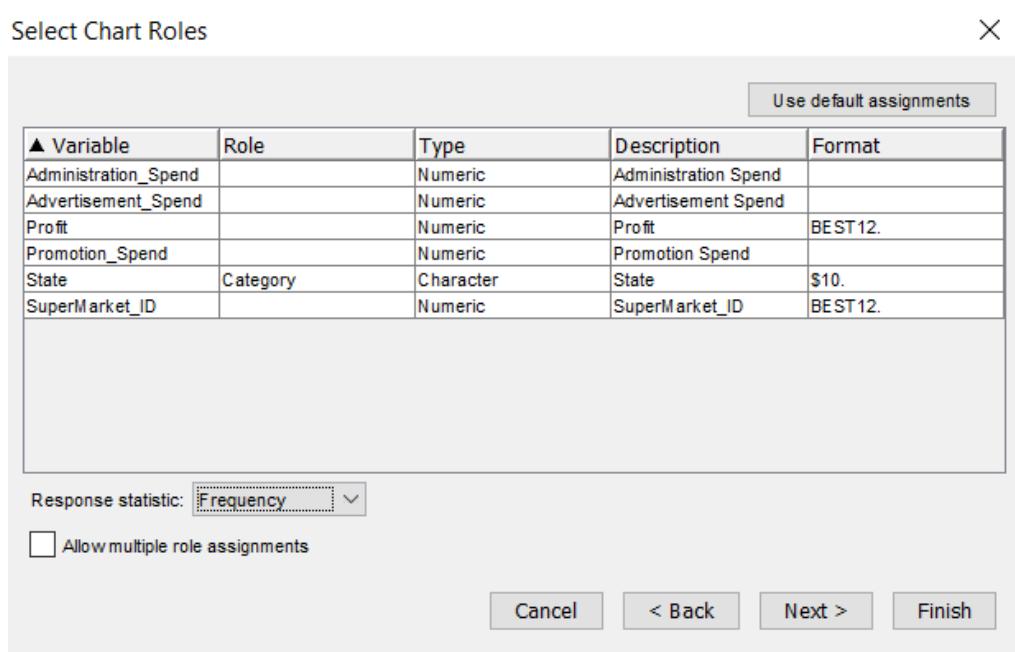


Figure 93

c. There is no where clause, click Next

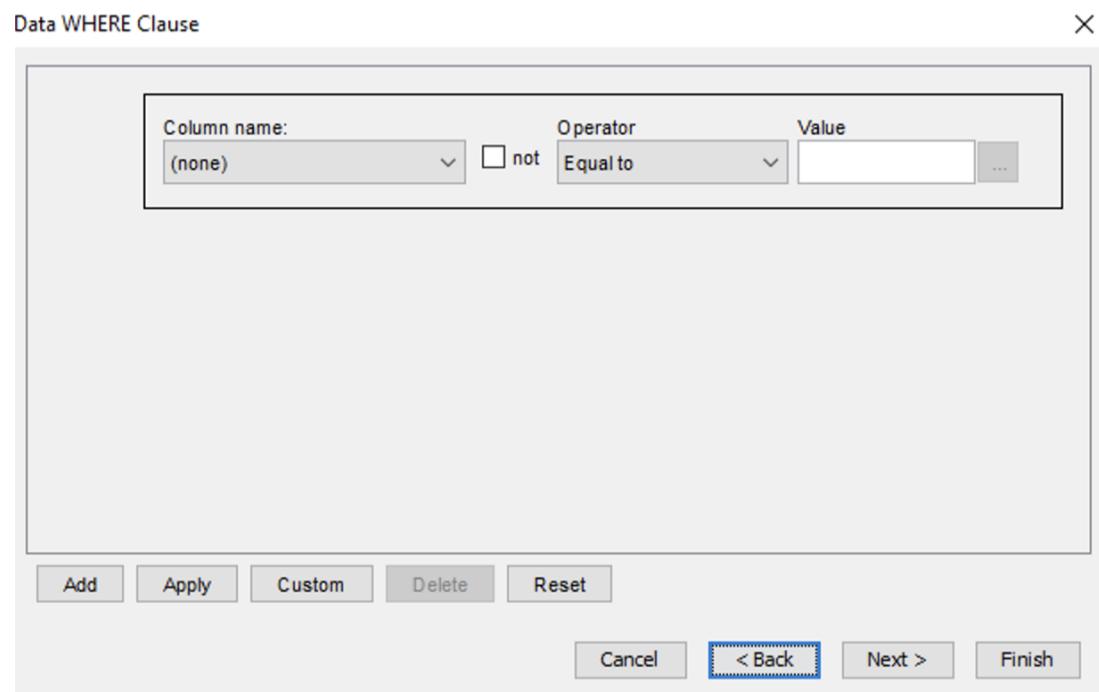


Figure 94

- 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

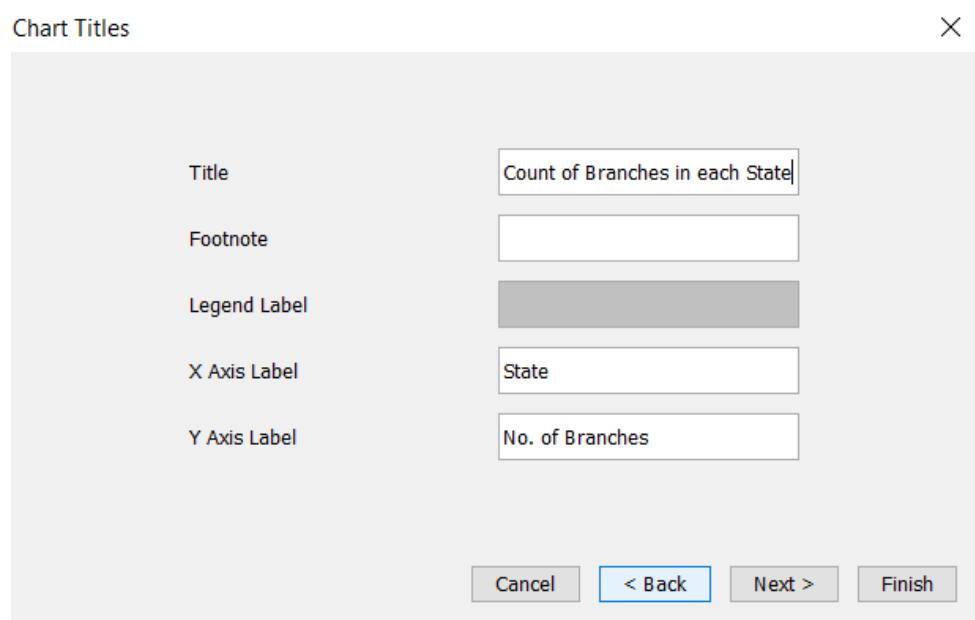


Figure 95

- e. The Below graph is developed by SAS EM

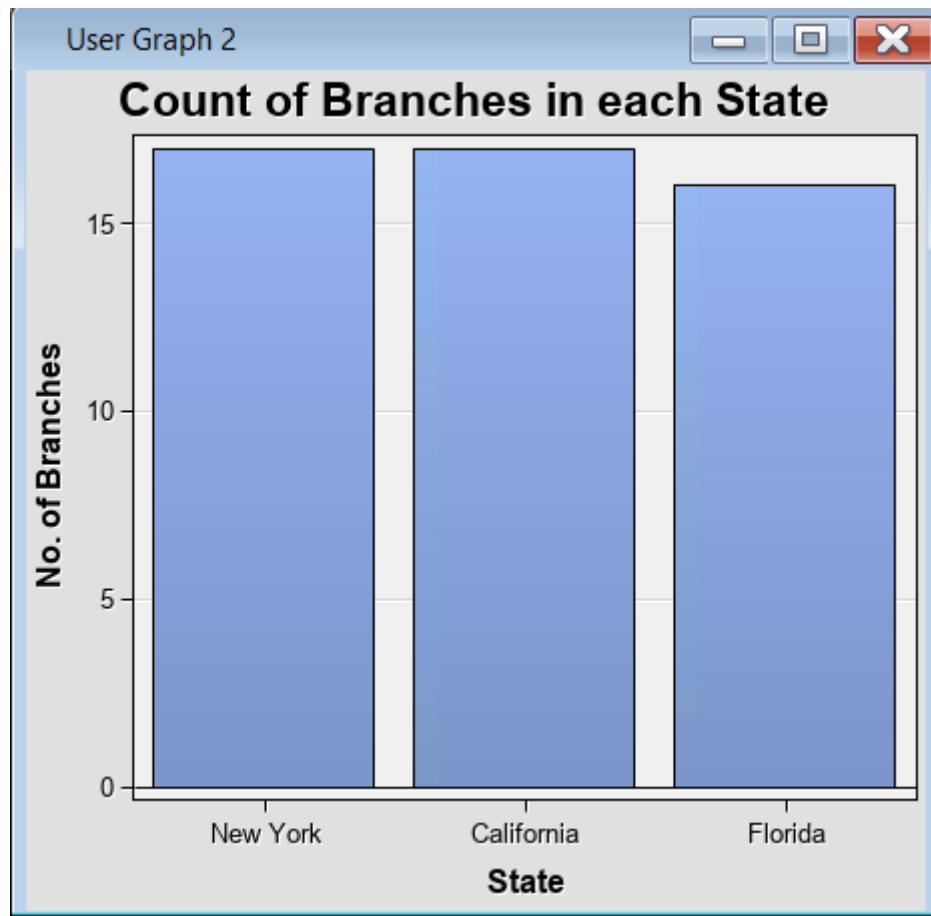


Figure 96

#### 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

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

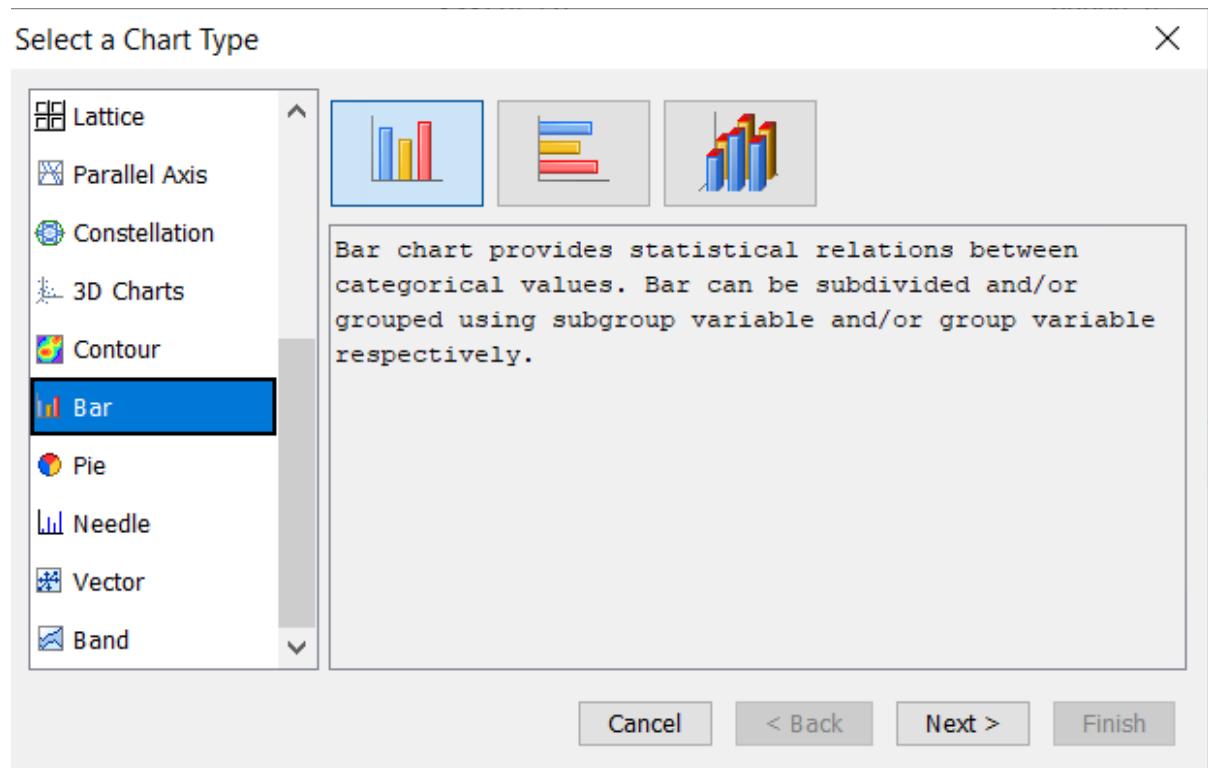


Figure 97

- Select State as category and Advertisement\_Spend as frequency. click Next.

Select Chart Roles

▲ Variable	Role	Type	Description	Format
Administration_Spend		Numeric	Administration Spend	
Advertisement_Spend	Frequency	Numeric	Advertisement Spend	
Profit		Numeric	Profit	BEST12.
Promotion_Spend		Numeric	Promotion Spend	
State	Category	Character	State	\$10.
SuperMarket_ID		Numeric	SuperMarket_ID	BEST12.

Response statistic: Frequency

Allow multiple role assignments

Figure 98

c. There is no where clause, click Next

Data WHERE Clause

Column name:	(none)	Operator	Equal to	Value
<input type="checkbox"/> not				

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

Chart Titles X

Title	Branch vs Money Spent on Advel
Footnote	
Legend Label	
X Axis Label	Branch
Y Axis Label	Money Spent on Advertisement

Cancel < Back Next > Finish

Figure 100

e. The Below graph is developed by SAS EM

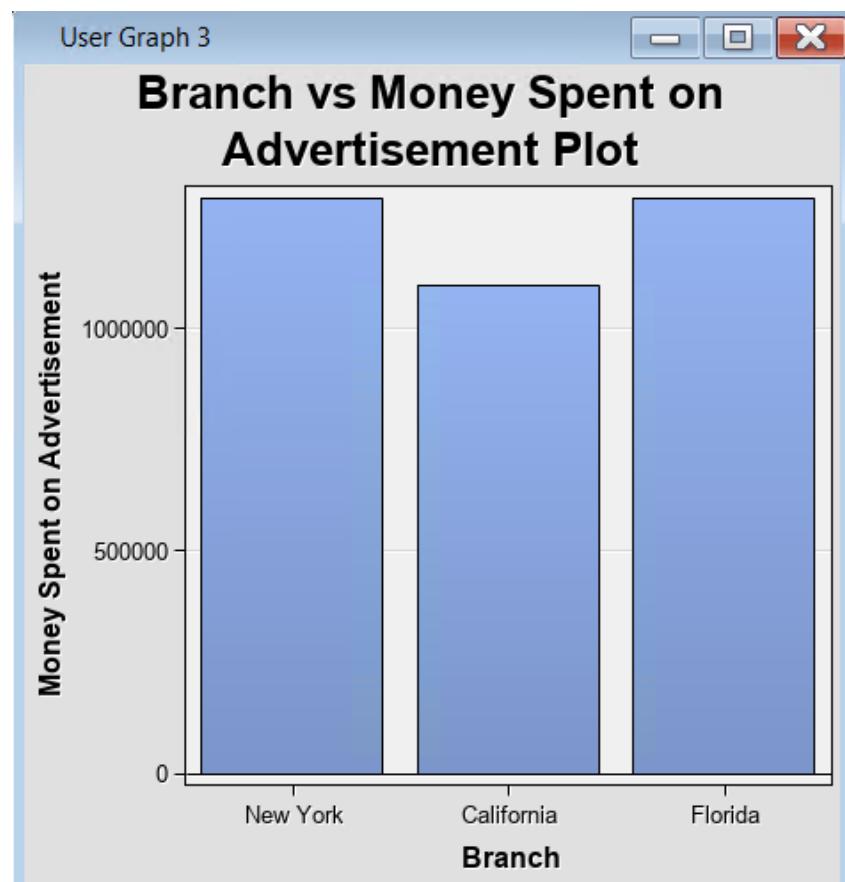


Figure 101

## **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.

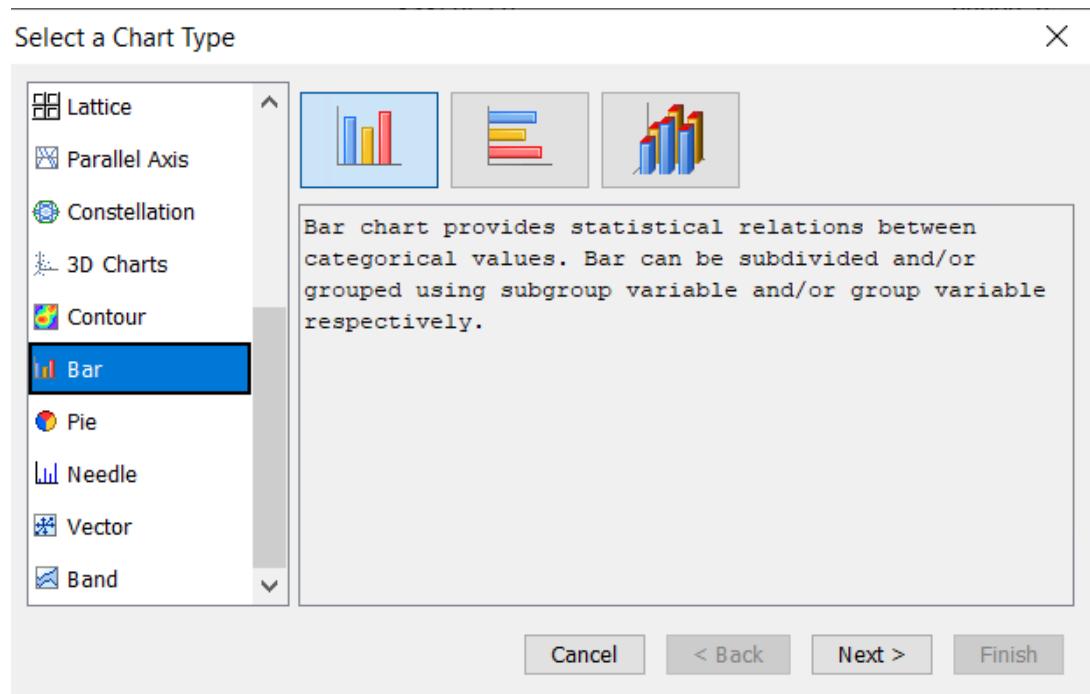


Figure 102

b. Select State as category and Promotion\_Spend as frequency. click next.

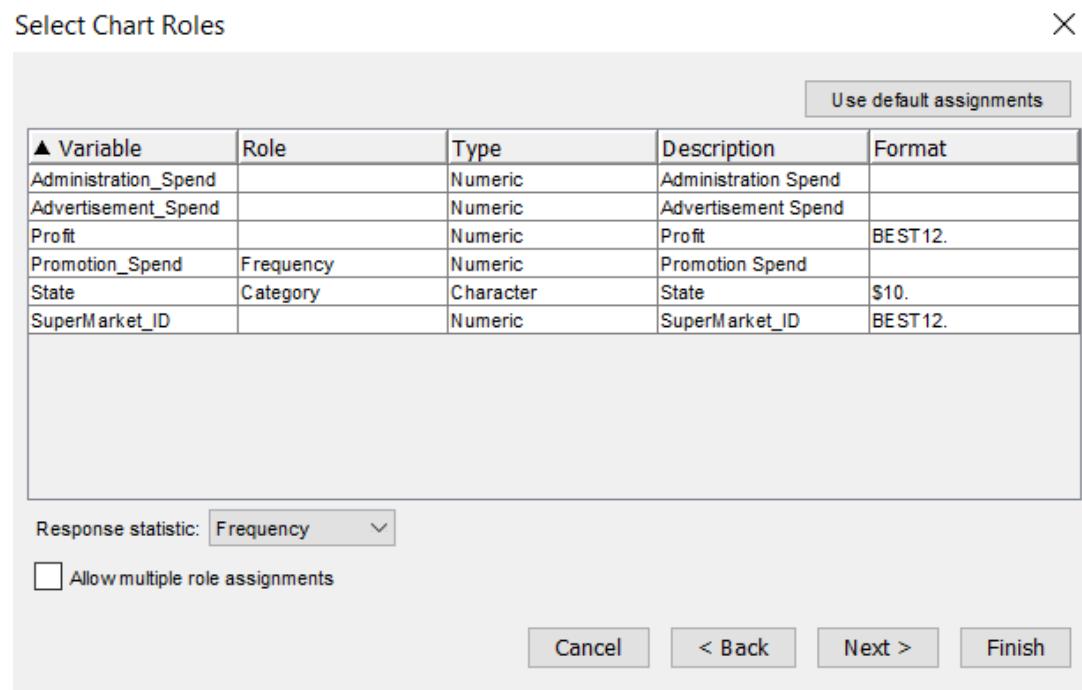


Figure 103

c. There is no where clause, click Next.

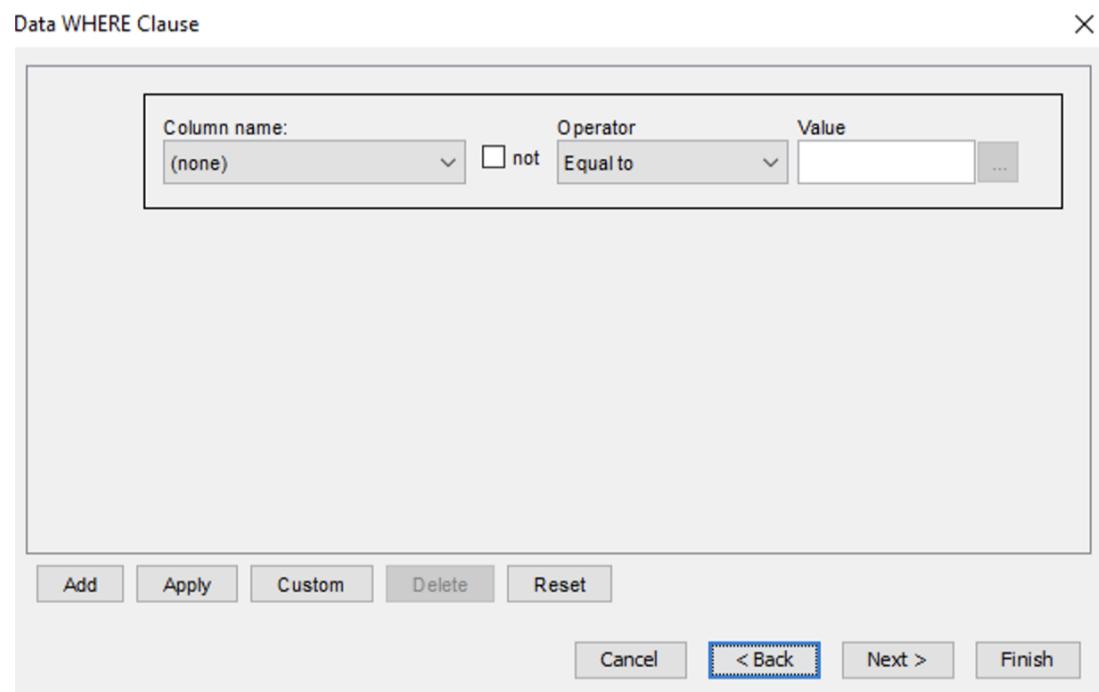


Figure 104

- 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

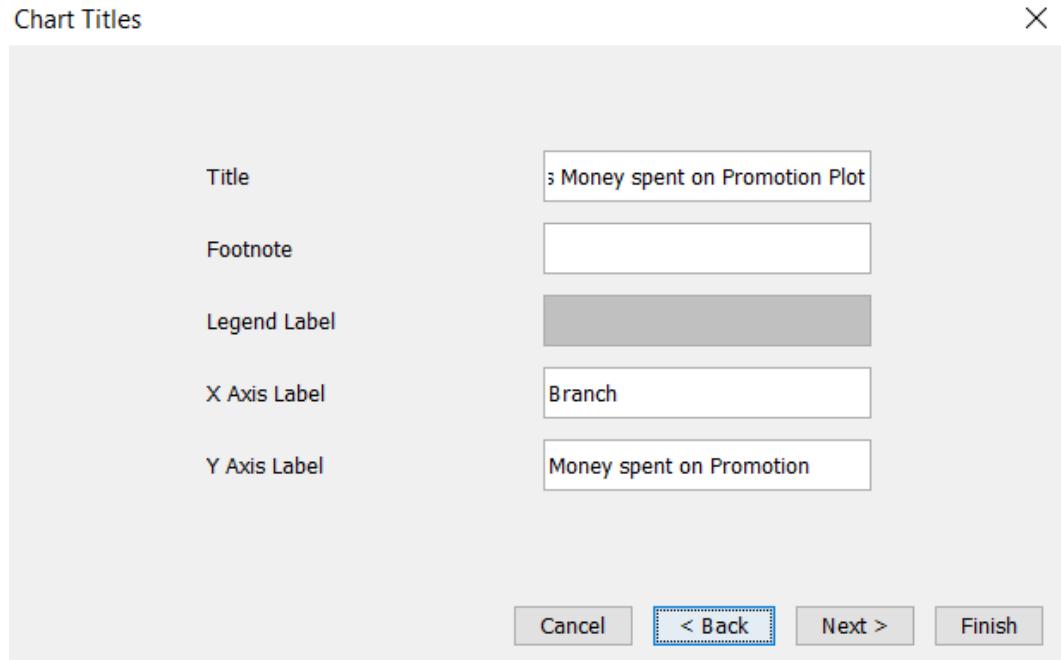


Figure 105

- e. The Below graph is developed by SAS EM

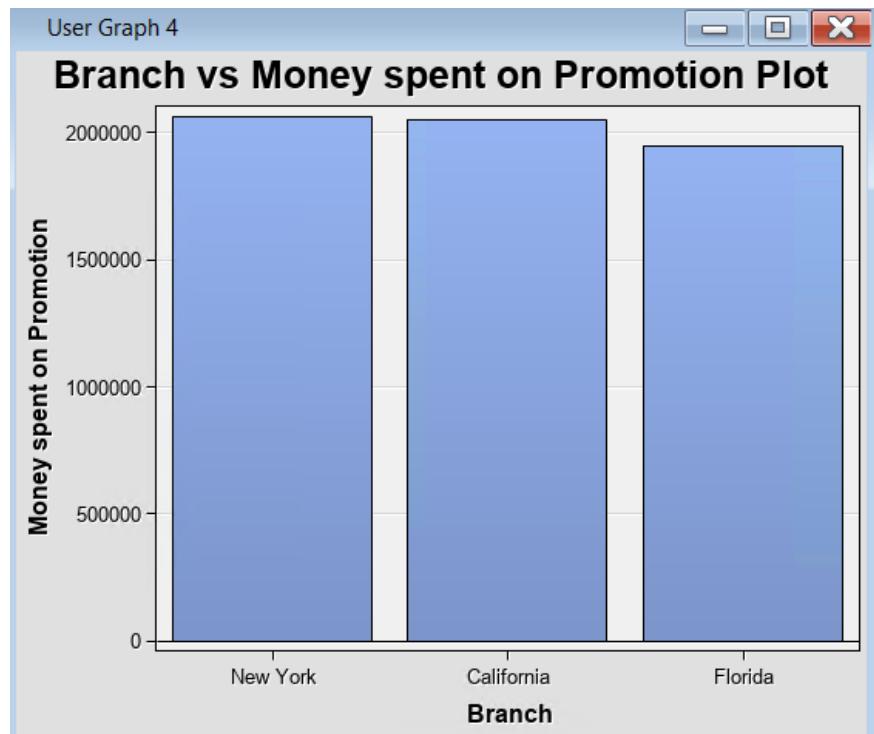


Figure 106

### **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

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**

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

- Click the Explore tab and find StatExplore node

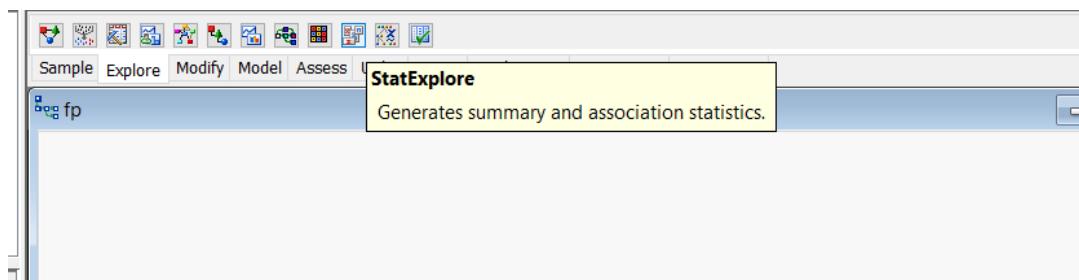


Figure 109

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

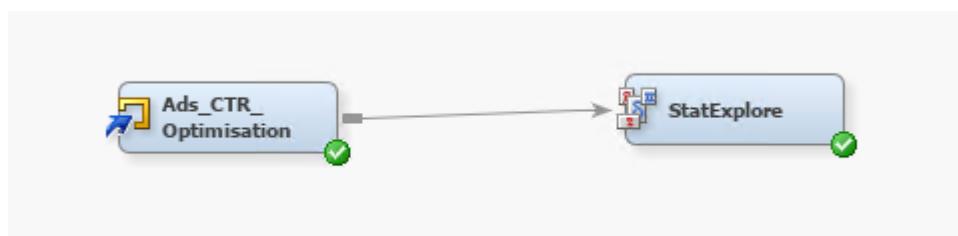


Figure 110

### Step 3: Run the diagram

- 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.

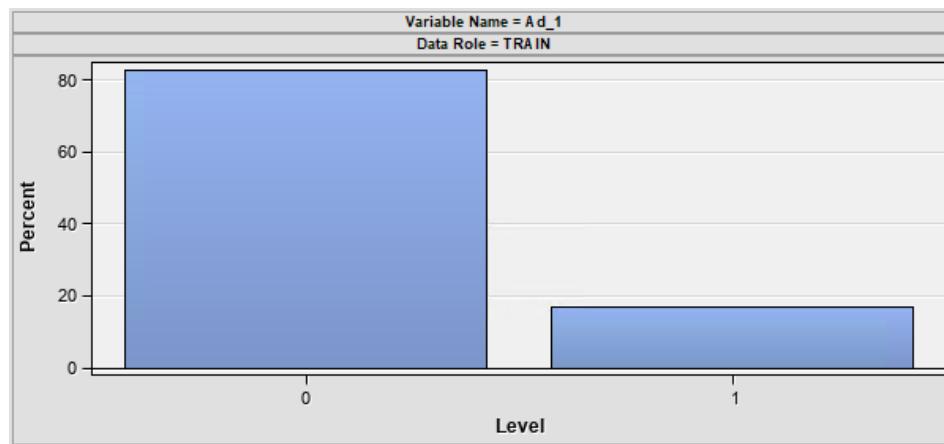


Figure 111

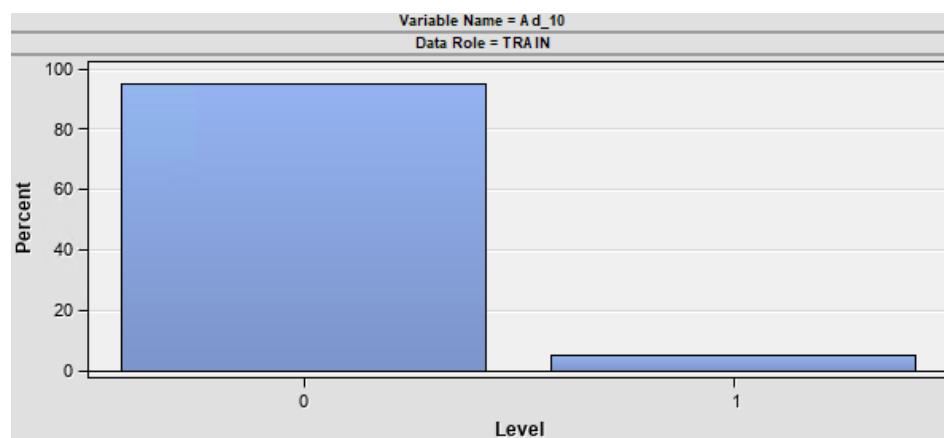


Figure 112

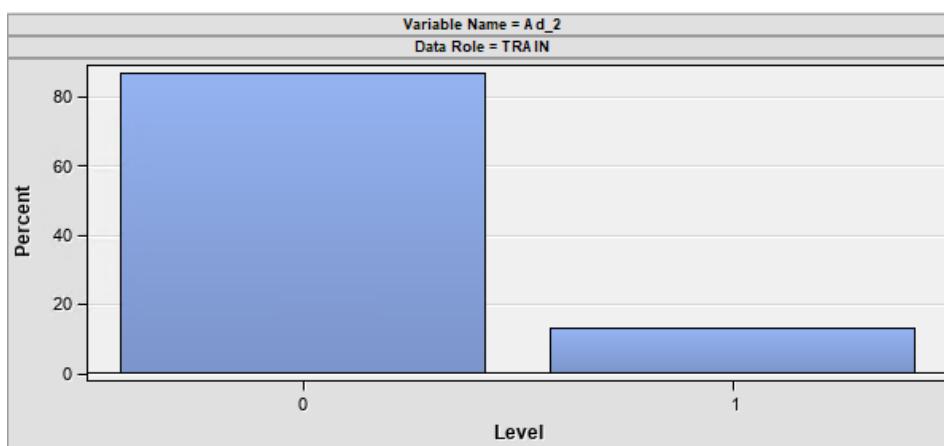


Figure 113

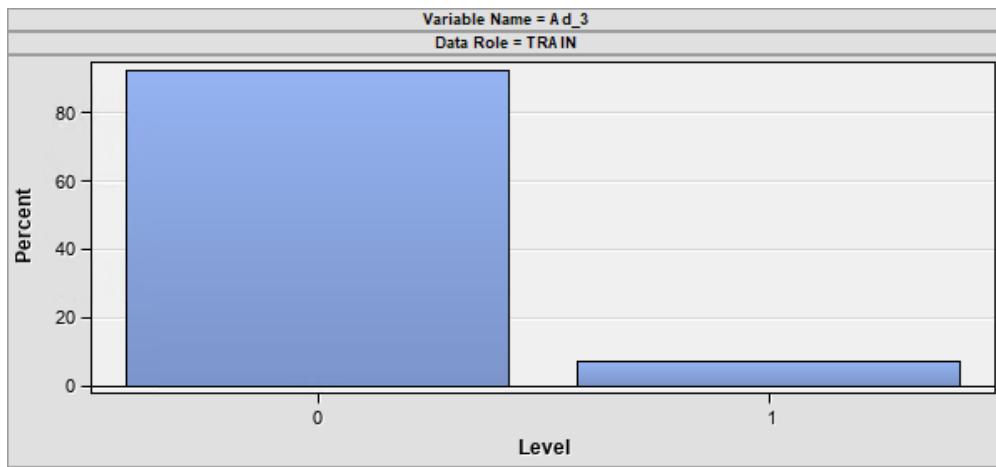


Figure 114

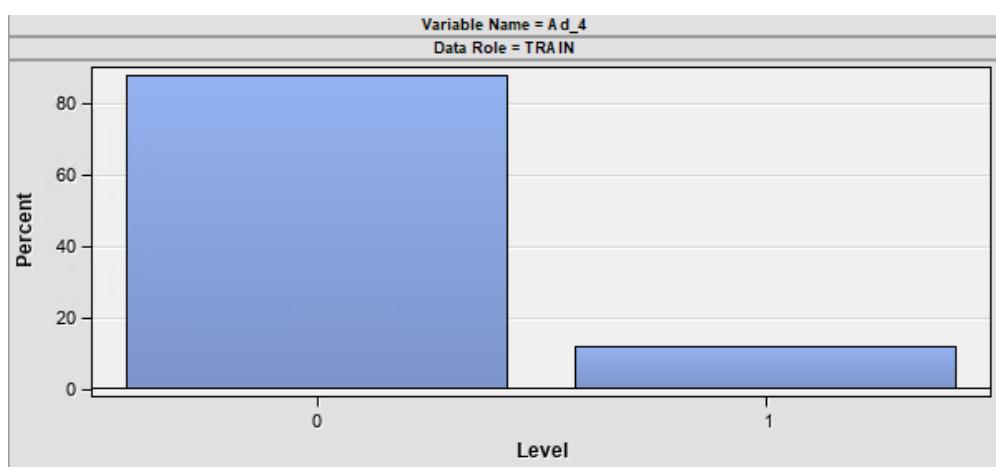


Figure 115

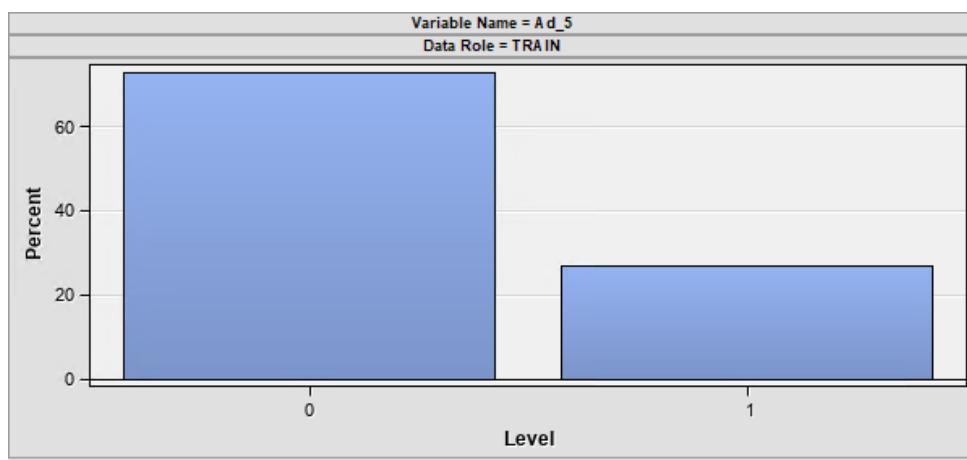


Figure 116

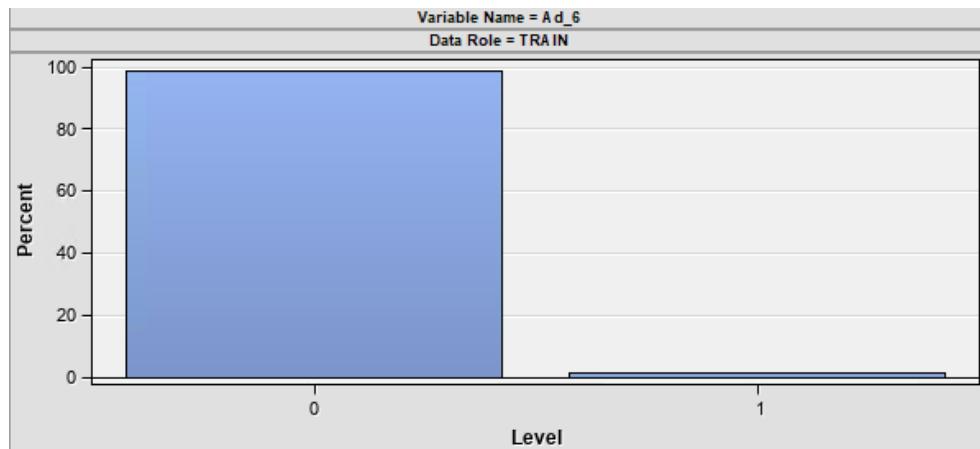


Figure 117

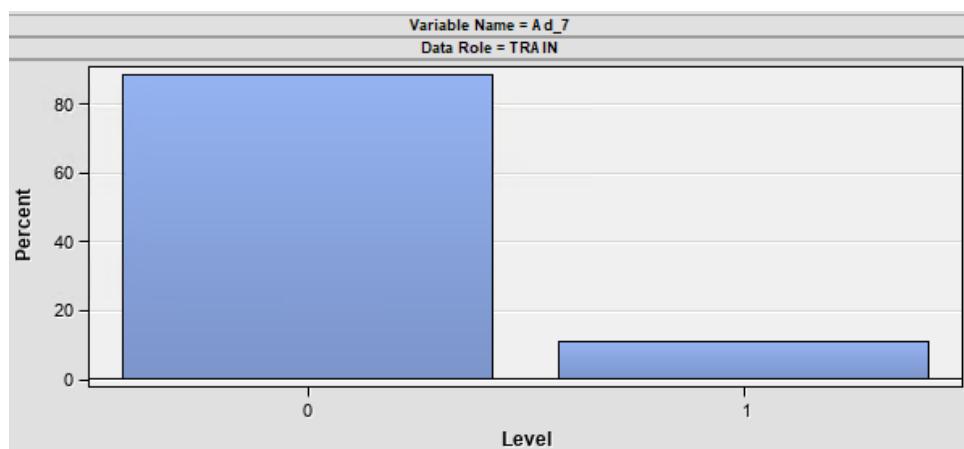


Figure 118

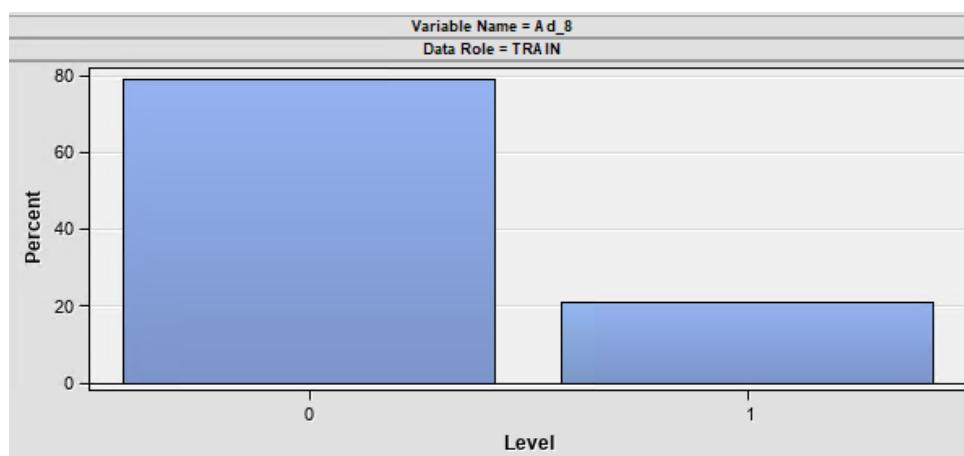


Figure 119

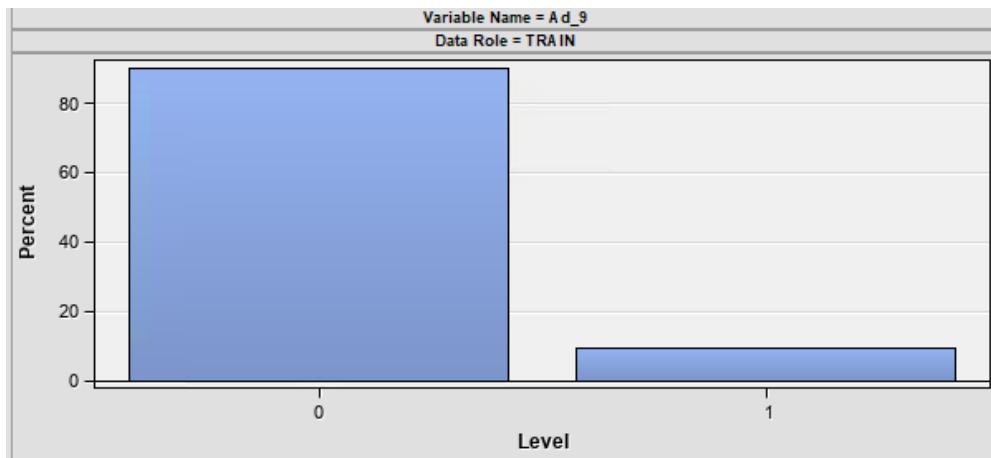


Figure 120

Data Role	Variable Name	Role	Number		Mode		Mode2	
			of Levels	Missing	Mode	Percentage	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.