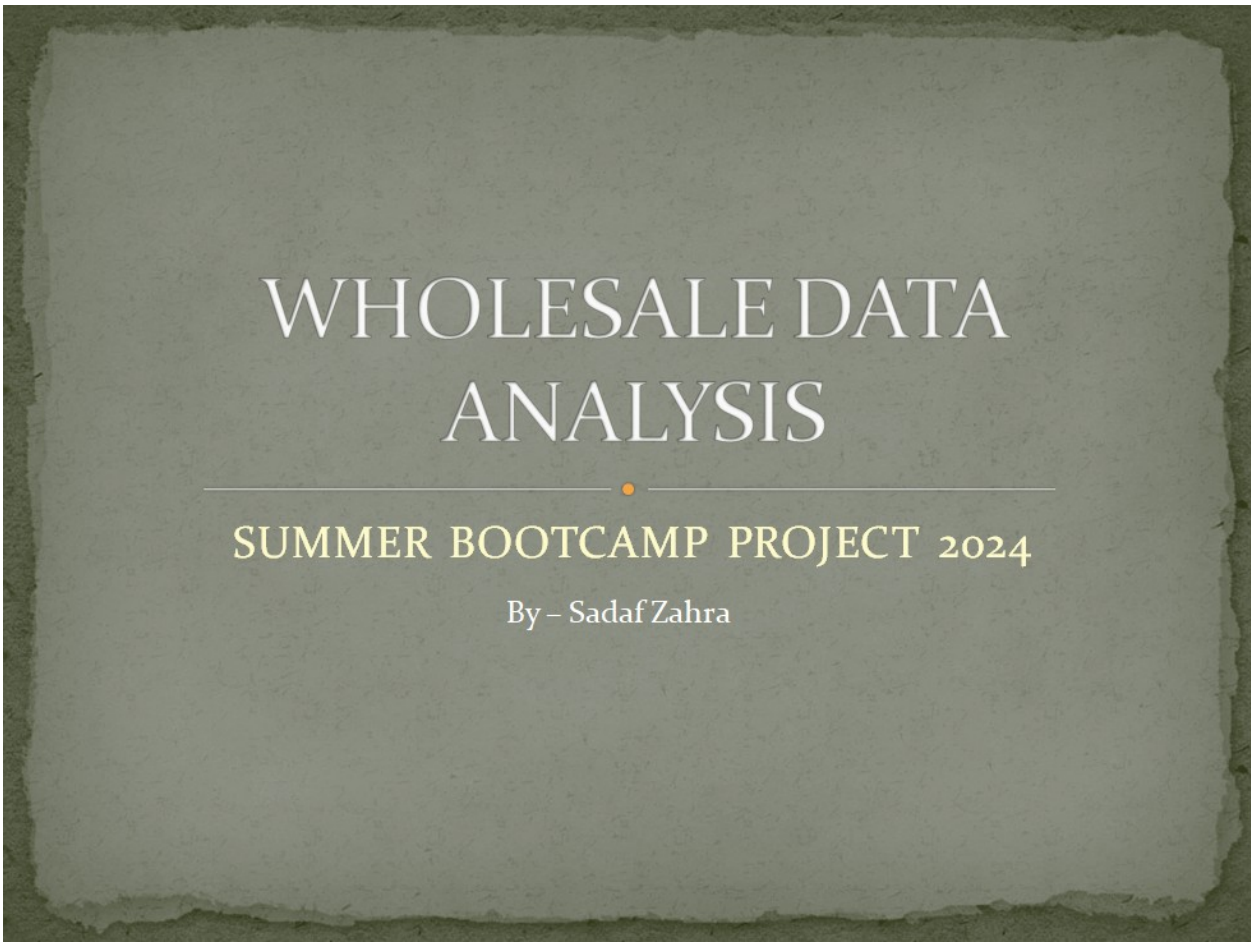

WHOLESALE DATA ANALYSIS
SUMMER BOOTCAMP PROJECT 2024
By – Sadaf Zahra

##Index

##List of Tables

- Table 8: Difference between high spenders and low spenders
- Table 9: Combined average spending on Fresh and Milk for each region

##List of Figures

- Figure 1 : Boxplot
- Figure 2 : Total spending region wise
- Figure 3 : Region having maximum spending on Milk
- Figure 4 : Average spending on Grocery by Region
- Figure 5 : Average spending on Detergents_Paper by Region
- Figure 6 : Correlation between Fresh and Frozen
- Figure 7 : Cluster Formation
- Figure 8: Barplot to represent the top 10% spenders in each category
- Figure 9: Scatter plot displaying the spending on Milk vs Grocery

##Problem Statement/ Objective A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers' annual spending on 6 different varieties of productsin 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel, Retail).

##Data Description

1. Buyer/Spender- ID's of customers
2. Region- Region of the distributor
3. Fresh- spending on Fresh Vegetables
4. Milk- spending on milk
5. Grocery- spending on grocery
6. Frozen- spending on frozen food
7. Detergents_paper- spending on detergents and toilet paper
8. Delicatessen- spending on instant foods

##Loading the necessary libraries

1- Display the top 5 rows.

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	1	Retail	Other	12669	9656	7561	214.0	2674.0	1338.0
1	2	Retail	Other	7057	9810	9568	1762.0	3293.0	1776.0
2	3	Retail	Other	?	8808	7684	2405.0	3516.0	7844.0
3	4	Hotel	Other	13265	1196	4221	6404.0	507.0	1788.0
4	5	Retail	Other	22615	5410	7198	3915.0	1777.0	5185.0

Based on the above result we can observe that:

- Fresh have values as '?'.

2- Display the last 5 rows

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
435	436	Hotel	Other	29703	12051	16027	13135.0	182.0	2204.0
436	437	Hotel	Other	39228	1431	764	4510.0	93.0	2346.0
437	438	Retail	Other	14531	15488	30243	437.0	14841.0	1867.0
438	439	Hotel	Other	10290	1981	2232	1038.0	168.0	2125.0
439	440	Hotel	Other	2787	1698	2510	65.0	477.0	52.0

3- Check the shape of dataset.

It shows that our dataset have 440 rows and 9 columns.

4- Check the datatypes of each feature.

The datasets have following types of data in given column

- Buyer/Spender has int type data which is correct
- Channel column has object type data; also correct
- Region column has object type data; correct
- Fresh column has object type data, which needs to be corrected to int type
- Milk has int type data
- Grocery has int type data
- Frozen has float type data, I'll change it to int
- Detergents_Paper has float type data, I'll change it to int
- Delicatessen has float type data, I'll change it to int

5- Check the Statistical summary

-- Results of checking the statistical summary is:

- Buyer/Spender have count of the 440.000000 ,average of 220.500000,Standard Deviation of 127.161315,minimum of 1.000000 ,25 percentile of 110.750000,50 percentile of 220.500000,75 percentile of 330.250000,maximum of 440.000000.
- Milk have count of the 440.000000,average of 6035.779545,Standard Deviation of 8964.929649,minimum of 1.000000,25 percent of 1525.250000,50 percent of 3641.000000,75 percent of 7217.500000,maximum of 112400.000000.
- Grocery have count of the 440.000000,average of 7951.277273,Standard Deviation of 9503.162829,minimum of 3.000000,25 percent of 2153.000000,50 percent of 4755.500000,75 percent of 10655.750000,maximum of 92780.000000.
- Frozen have count of the 437.000000,average of 3085.638444,Standard Deviation of 4867.744145,minimum of 25.000000,25 percent of 744.000000,50 percent of 1535.000000,75 percent of 3570.000000,maximum of 60869.000000.
- Detergents_Paper have count of the 439.000000,average of 3773.747153,Standard Deviation of 19364.886053,minimum of 3.000000,25 percent of 256.500000,50 percent of 813.000000,75 percent of 3956.000000,maximum of 396100.000000.
- Delicatessen have count of the 438.000000,average of 1531.057078,Standard Deviation of 2825.044262,minimum of 3.000000,25 percent of 411.250000,50 percent of 971.000000,75 percent of 1822.750000,maximum of 47943.000000

6- Check the null values

```
Buyer/Spender       0
Channel             3
Region              6
Fresh               0
Milk                0
Grocery             0
Frozen              3
Detergents_Paper    1
Delicatessen        2
dtype: int64
```

In the given dataset we have null values in following columns only:

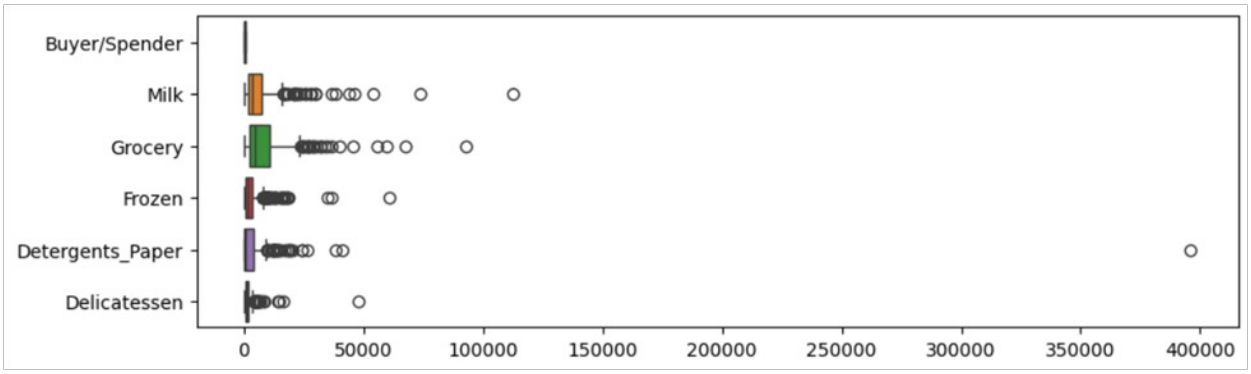
- Channel has 3 missing values.
- Region has 6 missing values.
- Frozen has 3 missing values.
- Detergents_Paper has 1 missing values.
- Delicatessen has 2 missing values. **Notice that the Fresh column indicates that there are 0 null values. But there are 2 rows that contain '?' instead of a number. So after removing that, we'll have 2 null values in the Fresh column**

7- Check the duplicate values

In the given dataset there are 0 duplicate values.

9- Check the outliers and their authenticity.

We will use IQR(Interquartile Range) for detecting outlier in our dataset.



10- Do the necessary data cleaning steps like dropping duplicates, unnecessary columns, null value imputation, outliers treatment etc.

Observation: Replaced the '?' with null values which are later handled

- Fresh have object type of data so we need to convert it to int type to correctly analyse the data(which we have done)
- Dropping the duplicate values in case if there is any: There are 0 dupliccate values.
- Unecessary columns: All the columns are required in this datasets.
- Null Value imputation

- Fresh has value as '?' on two indexes 2 and 78.

Here we can see that the there are many unique values in this column but '?' makes it a object type so we will replace it with nan and the with median. We will fill null values of numeric columns with median of that column and non numeric columns with mode here so that it won't affect our datasets.

Observation: All the null values of the numeric columns are filled with the median of that column. 'Channel' and 'Region' are two non numeric columns that have 3 and 6 null values respectively, which will be filled with mode of that column

Observation: All the null values are handled

```
Buyer/Spender       0
Channel             0
Region              0
Fresh               0
Milk                0
Grocery             0
Frozen              0
Detergents_Paper    0
Delicatessen        0
dtype: int64
```

Based on the above operations we can now see that the this datasets have 0 null-values.

#1. Spending Analysis

- What is the total number of buyers in the dataset?

Total number of buyers: 440

- What is the average spending on each category (Fresh, Milk, Grocery, Frozen, Detergents_paper, Delicatessen)?

```
Average spending on each category:
Fresh              12000.045455
Milk                5073.352273
Grocery             7236.327273
Frozen              2507.361364
Detergents_Paper    2401.163636
Delicatessen        1270.034091
dtype: float64
```

- Which category has the highest average spending?

Fresh 12000.045454545454 The category with the highest average spending is 'Fresh' with an average spending of 12000.05.

- How many buyers spend above the average on Fresh Vegetables?

Number of buyers spending above the average on Fresh Vegetables: 158

###2. Regional Demand

- What is the total spending in each region?

Total spending in each region for numeric columns:

Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Lisbon	845508	397328	534652	222550	577447	98496
Oporto	432343	232764	418529	173172	172339	50668
Other	4002169	2025651	2545381	957307	907702	523381

Table 6: Total spending in each region

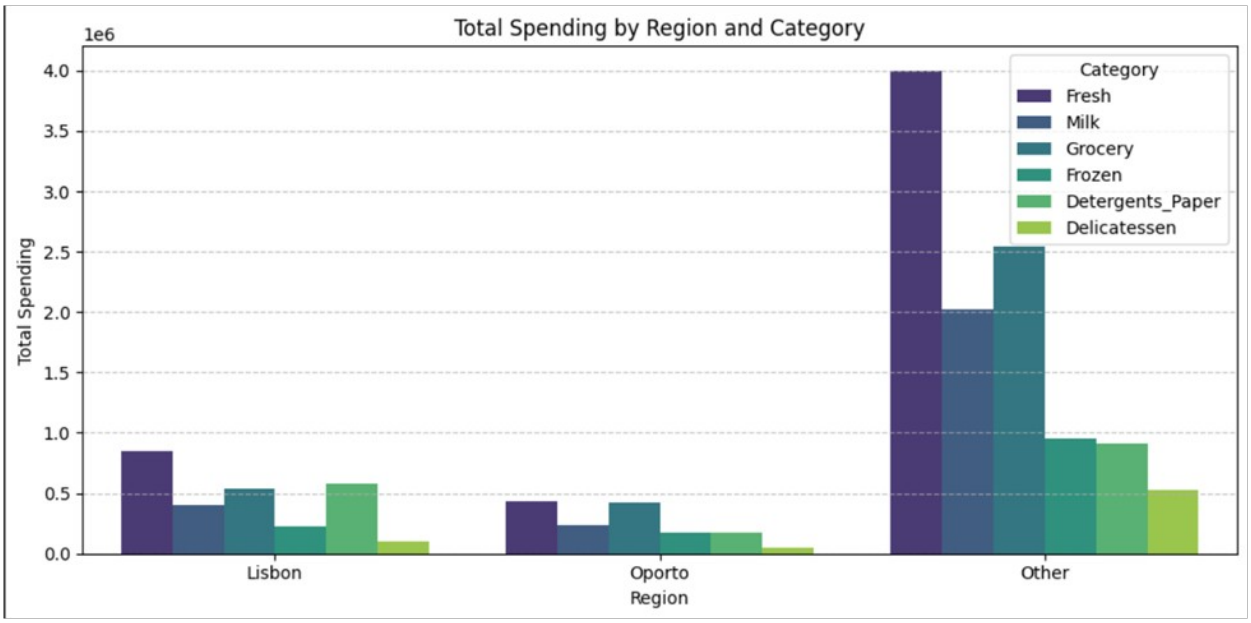
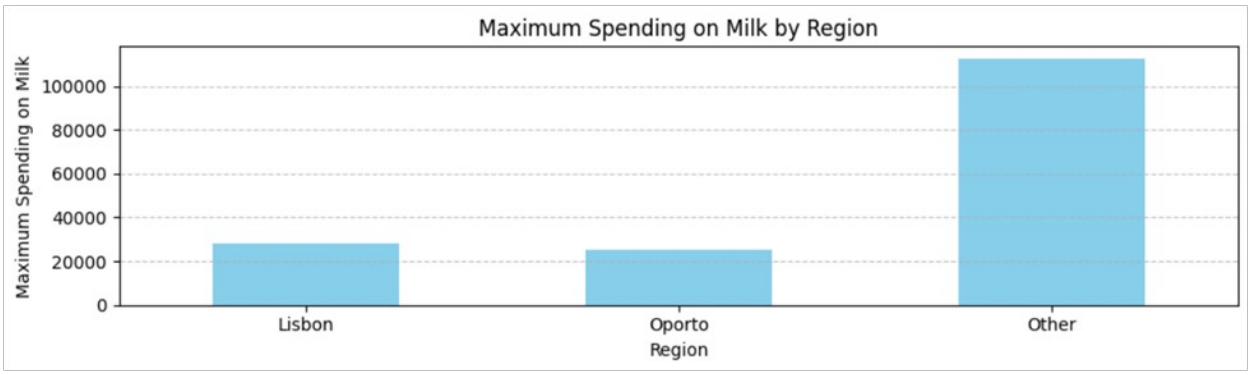


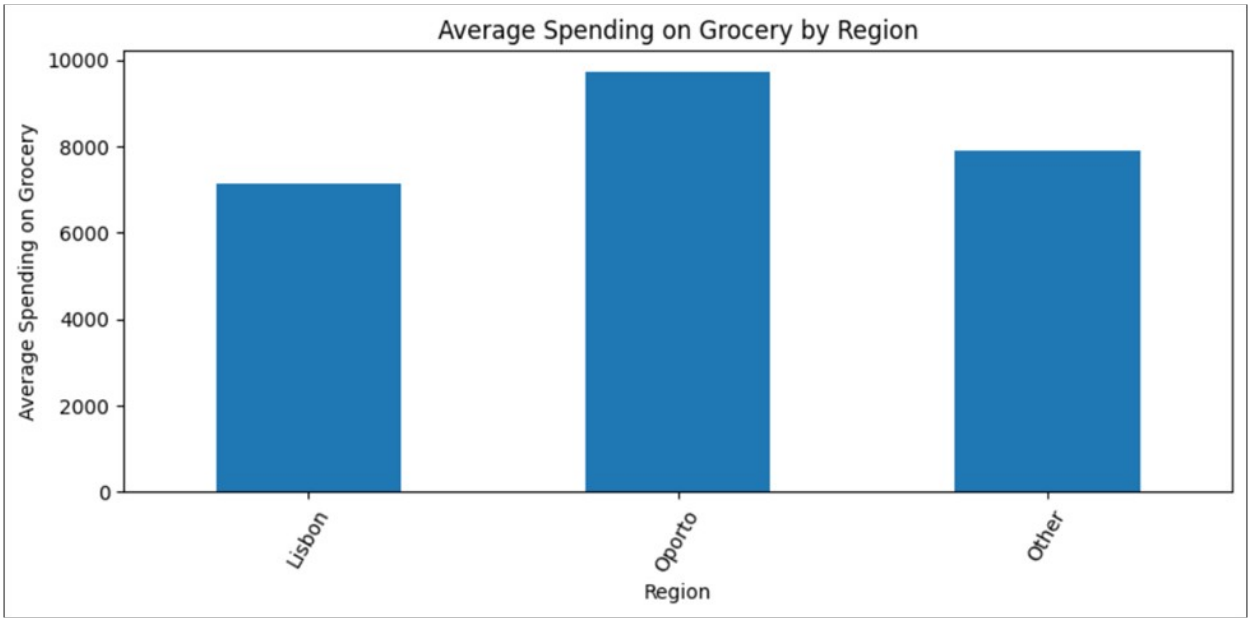
Figure 2: Total spending region wise

- Which region has the highest spending on Milk?

The region with the highest spending on Milk is 'Other' with a maximum spending of 112400.00.



- How does the average spending on Grocery vary across different regions?



- Which region has the highest average spending per buyer?

The region with the highest average spending per buyer is 'Lisbon' with an average spending of $35679.75 per buyer.

###3. Category Preferences

- What percentage of buyers spend more on Frozen food compared to Delicatessen?

The percentage of buyers who spend more on Frozen food compared to Delicatessen is 65.0000%.

- Which category shows the most variation in spending among buyers?

The category with the most variation in spending is 'Fresh' with a standard deviation of 12646.52.

- Are there any regions where spending on Detergents_paper is significantly higher than others?

```
Average spending on Detergents_Paper by Region:
Region
Lisbon    7699.293333
Oporto    4007.883721
Other     2818.950311
Name: Detergents_Paper, dtype: float64
```

Table 7: Average spending on Detergents_Paper by Region

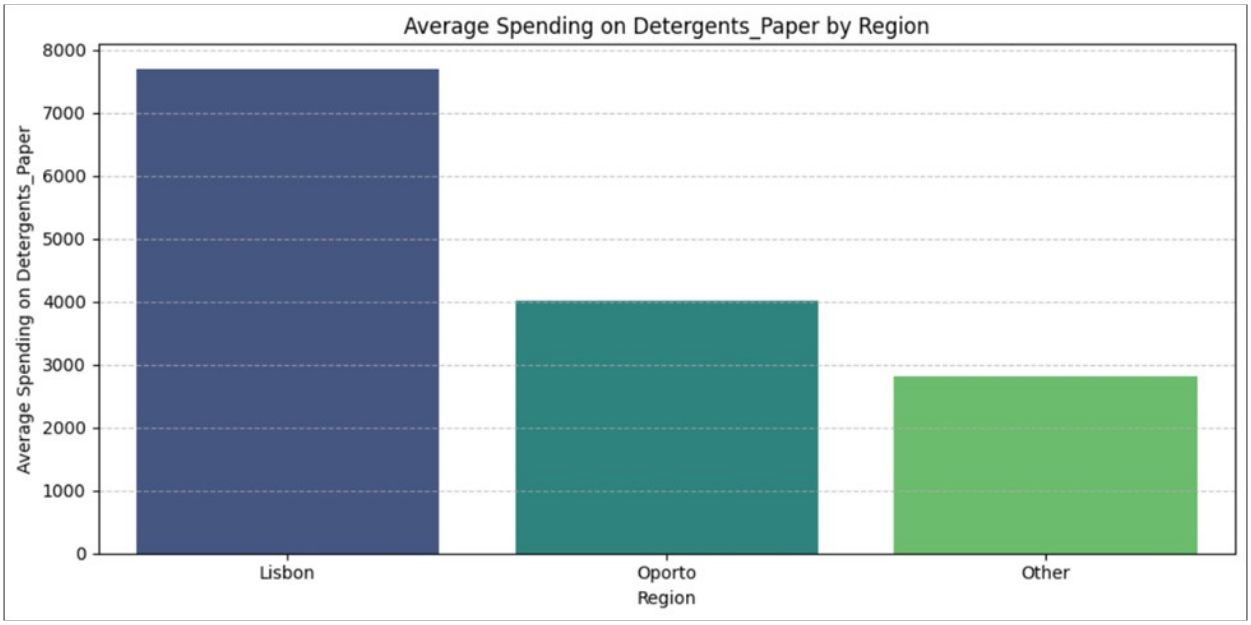


Figure 5: Average spending on Detergents_Paper by Region

- What is the correlation between spending on Fresh and Frozen food?

The correlation between spending on Fresh and Frozen food is 0.37

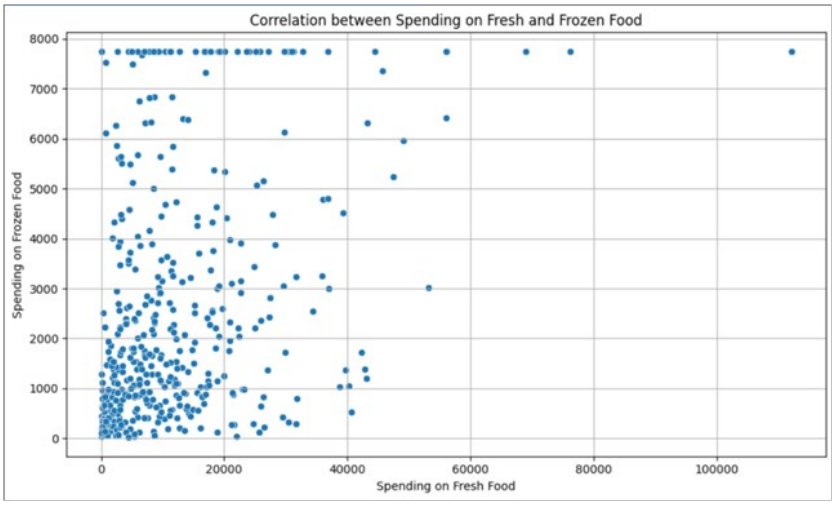
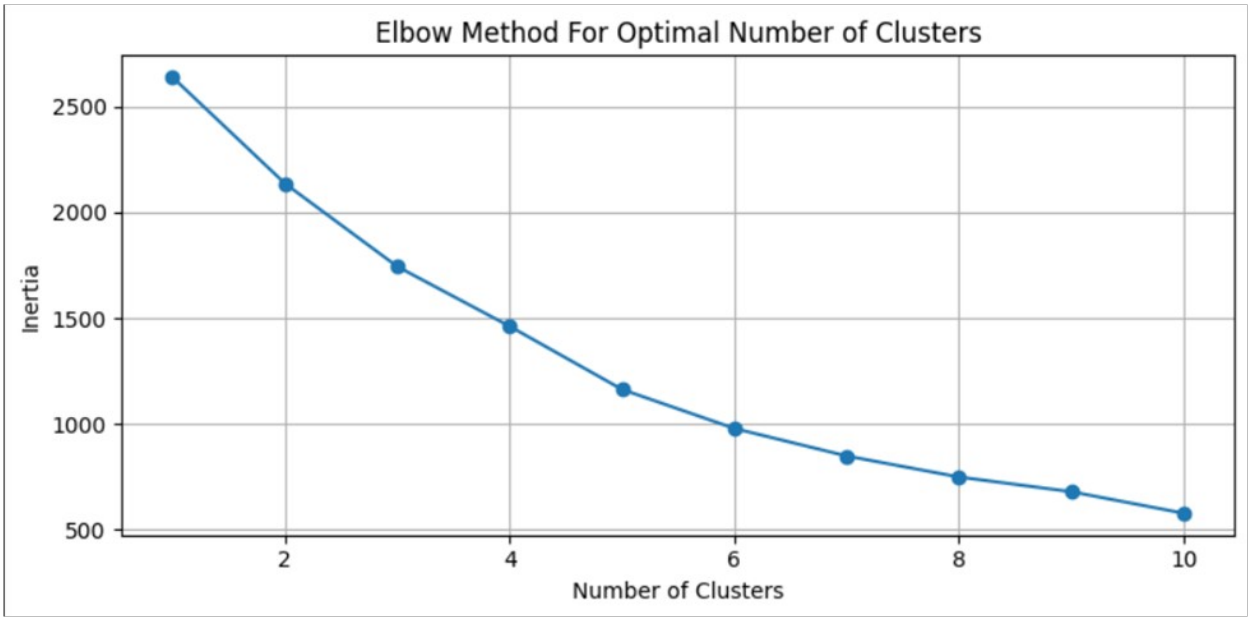


Figure 6: Correlation between Fresh and Frozen

A correlation coefficient of 0.37 between spending on Fresh and Frozen food indicates a positive correlation, meaning as spending on Fresh food increases, spending on Frozen food also tends to increase. However, the strength of this correlation is considered moderate rather than high.

##4. Customer Segmentation

- Can buyers be grouped into segments based on their spending patterns? (e.g., using clustering analysis)



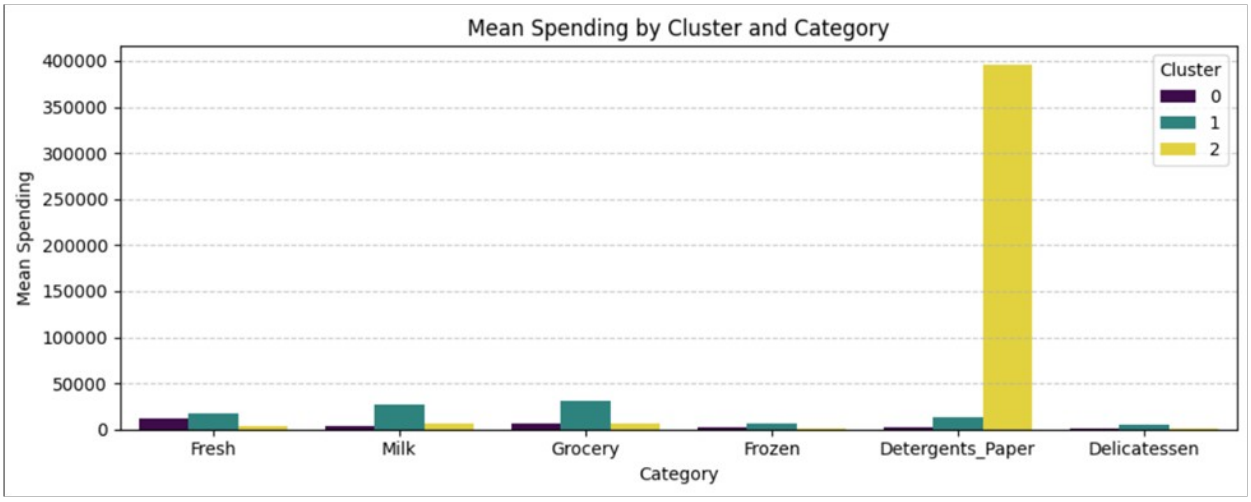


Figure 7: Cluster Formation

- What are the characteristics of the top 10% spenders in each category?

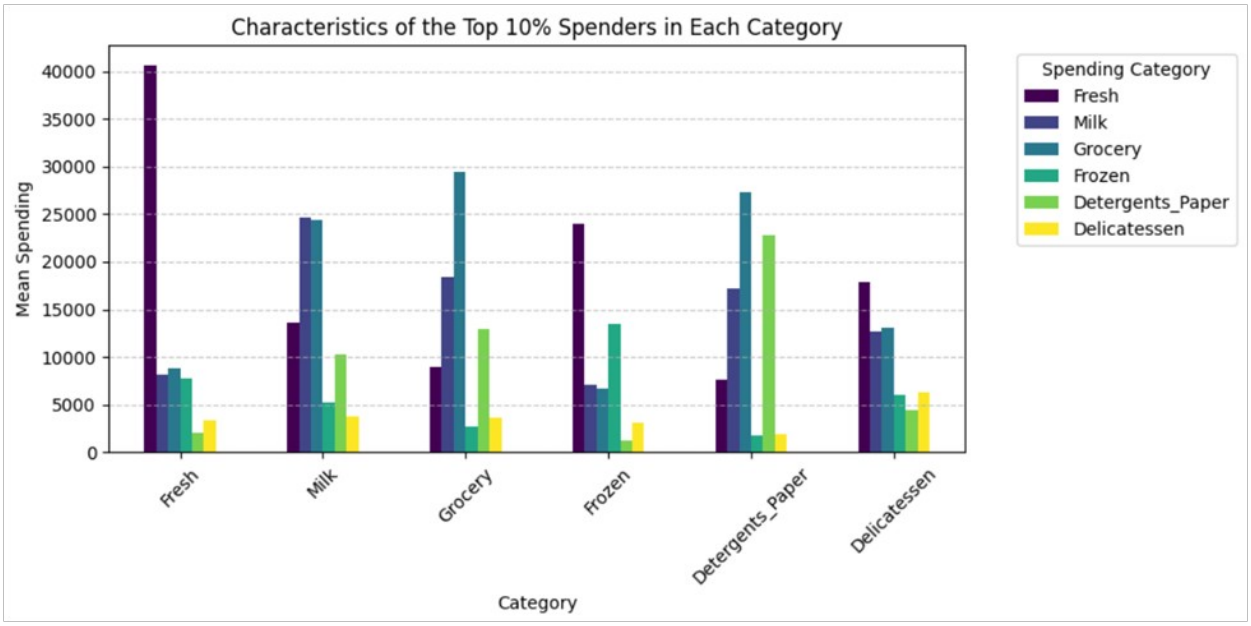


Figure 8: Bar plot to represent top 10% spenders in each category

- How do spending patterns differ between high spenders and low spenders?

To check the spending patterns differ between high spenders and low spenders:

- Calculate total spending: Sum the spending across all categories for each customer.
- Categorize spenders: Divide customers into high spenders and low spenders based on the median total spending.
- Calculate average spending: Compute the average spending in each category for both high spenders and low spenders.

Spender_Category	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
High Spender	17176.800000	9434.286364	12309.381818	4026.081818	6533.127273	2223.059091
Low Spender	6823.290909	2637.272727	3593.172727	2124.050000	1000.909091	833.963636

Table 8: Difference between the high spenders and low spenders

##5. Cross-Category Analysis

- Is there a correlation between spending on Milk and Grocery?

Correlation between spending on Milk and Grocery: 0.5902889029079396

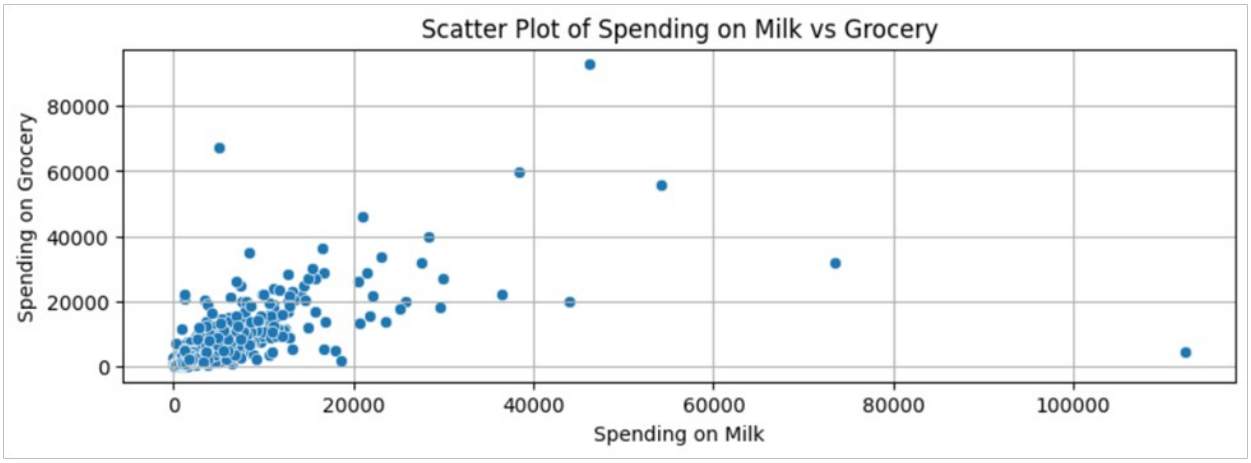


Figure 9: Scatter plot displaying the spending on Milk vs Grocery

- Do buyers who spend more on Delicatessen also spend more on Frozen food?

Correlation between spending on Delicatessen and Frozen food: 0.3904854847304075

- What is the combined average spending on Fresh and Milk for each region?

```
Region
Lisbon    16571.146667
Oporto    15467.604651
Other     18719.937888
Name: Fresh_Milk_Combined, dtype: float64
```

Table 9: combined average spending on Fresh and Milk for each region

##6. Demand Trends

- Which region has the fastest growing spending on Fresh Vegetables?

The region with the fastest growing spending on Fresh Vegetables is: Other

This code will calculate the average spending on Fresh Vegetables for each region and identifies which region has the highest average spending, implying faster growth in spending on Fresh Vegetables.

- How does the total spending on Grocery change across regions over time (if time data is available)?

Time data not available in the provided dataset.

- What is the average spending per buyer in each category over a specified time period (if time data is available)?

Time data is not available in the provided dataset.

##7. Buyer Insights

- What is the repeat purchase rate for buyers who spend above the average in at least three categories?

Repeat purchase rate for high spenders: 63.934426229508205

- How many buyers spend consistently (i.e., similar amounts) across all categories?

Number of buyers spending consistently across all categories: 17

- Which region has the most diverse spending patterns (i.e., high variance in spending across categories)?

Region with the most diverse spending patterns: Other