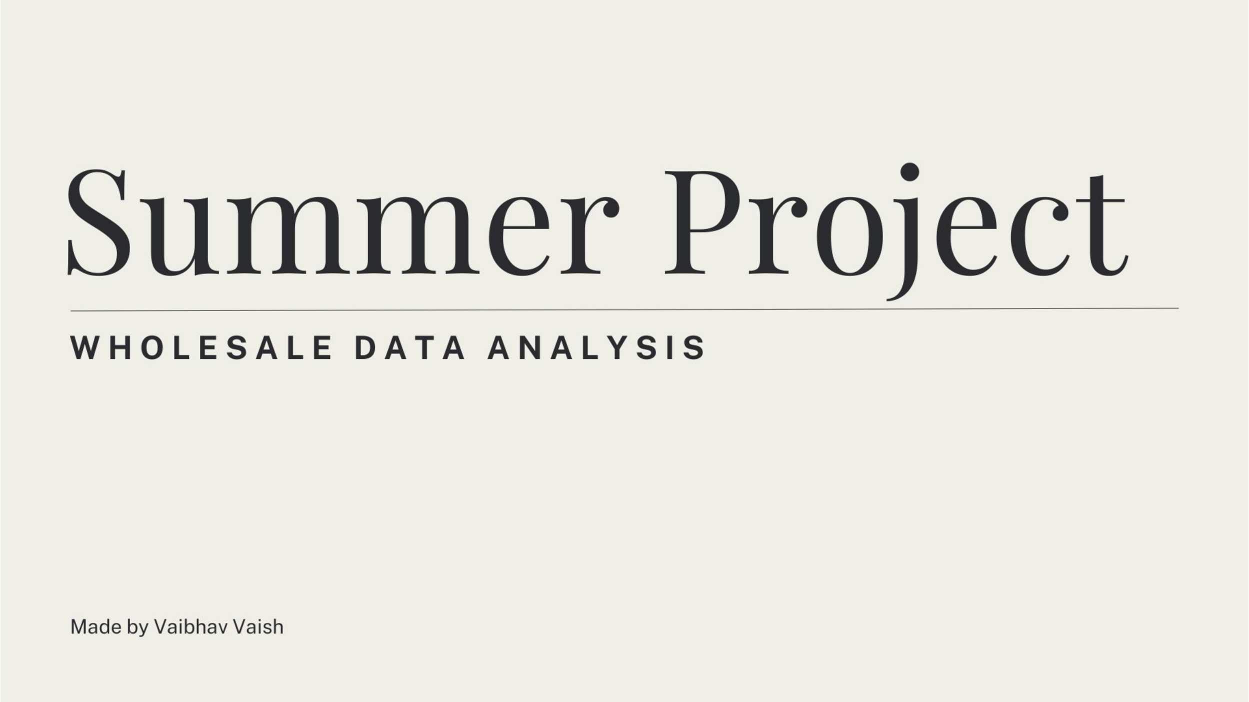
WHOLESALE DATA ANALYSIS



Summer Project

WHOLESALE DATA ANALYSIS

Made by Vaibhav Vaish

INDEX

List of Tables

List of Figures

Problem Statement / Objective

A wholesale distributor operating in different regions of Portugal has information about the annual spending of 440 large retailers on six different product varieties across three regions in Portugal (Lisbon, Oporto, and Other) and various sales channels (Hotel, Retail).

The dataset is provided by a wholesale distributor operating in different regions of Portugal. Here are the key features of the dataset:

- Number of Retailers: 440 large retailers
- Product Varieties: 6 different types

- Regions: Lisbon, Oporto, and Other
- Sales Channels: Hotel, Retail

This data can be used for various analyses, such as understanding spending patterns, identifying regional differences in product demand, and optimizing supply chain management for different sales channels.

Data

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 food

Importing the necessary Libraries

The code imports necessary libraries for data analysis and visualization in Python:

- pandas as pd: Used for data manipulation and analysis.
- numpy as np: Essential for numerical operations on arrays and matrices.
- matplotlib.pyplot as plt: Enables plotting graphs and charts.
- seaborn as sns: Enhances the visual appeal of plots.
- It also imports the KMeans algorithm from sklearn.cluster for clustering data points.

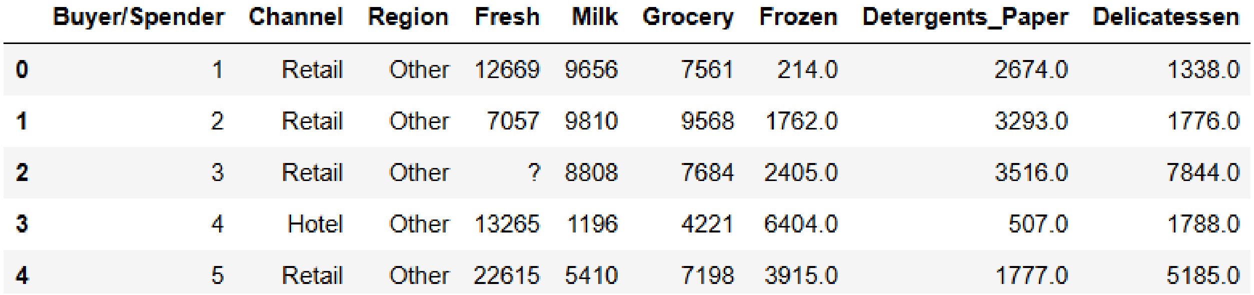
Load the dataset

- to load csv file- df=pd.read_csv- to load excel file- df=pd.read_excel

Basic Steps:

1. Display the top 5 rows

df.head() displays the first 5 rows of the DataFrame, providing a quick overview of the data.



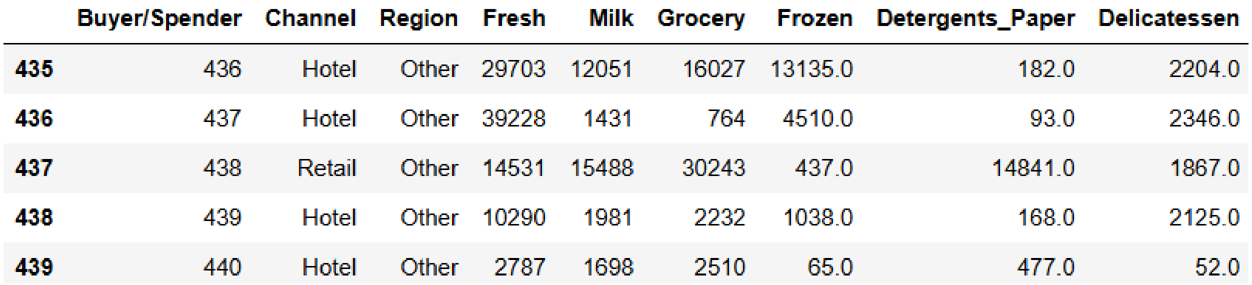
	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	1	Retail	Other	12669	9656	7561	214.0	2674.0	1338.0
1	2	Retail	Other	7057	9810	9568	1762.0	3293.0	1776.0
2	3	Retail	Other	?	8808	7684	2405.0	3516.0	7844.0
3	4	Hotel	Other	13265	1196	4221	6404.0	507.0	1788.0
4	5	Retail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

Observations

On Column Fresh row number 2 there is a null value

2. Display the last 5 rows

By default, df.tail() shows the last 5 rows of the DataFrame, but you can specify the number of rows to display by passing a number inside the parentheses.



	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

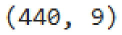
Observation

Here in the last five rows we can see that Buyer 437 has the highest spending rate on Fresh

3. Check the shape

- df.shape is a command used in Python with libraries like Pandas to display the dimensions of a DataFrame.

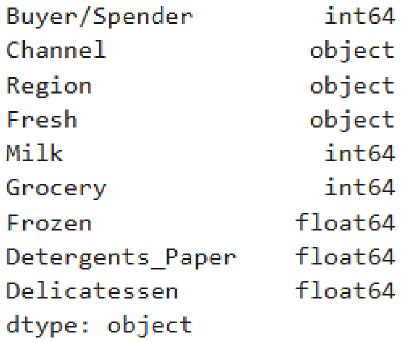
- It returns a tuple with two values: the number of rows and columns in the DataFrame.



```
(440, 9)
```

4. Check the datatypes of each feature

- df refers to a DataFrame, a table-like data structure in pandas library.- dtypes is a method used to display the data types of each column in the DataFrame
- This command helps in understanding the data types of the columns (e.g., integer, float, string) which is important for data manipulation and analysis.



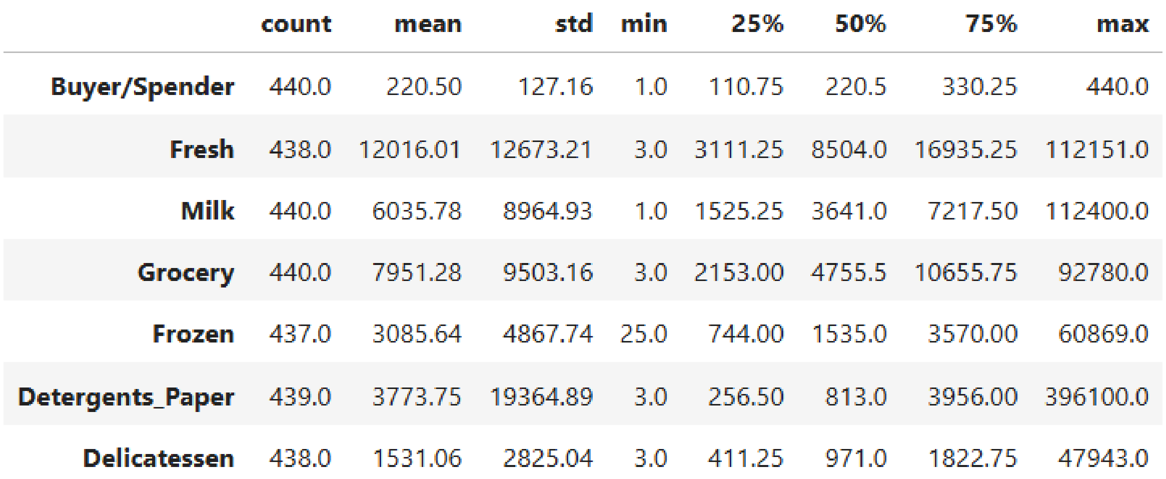
```
Buyer/Spender          int64
Channel               object
Region                object
Fresh                 object
Milk                   int64
Grocery                int64
Frozen               float64
Detergents_Paper     float64
Delicatessen         float64
dtype: object
```

Observations

- Fresh should be in float as observed from the table.
- The data set contains 440 observations of data and 9 variables. Only Channel and Region are categorical while rest is numeric data.

5. Check the statistical summary

- The df.describe() function is used in Python with pandas library to generate descriptive statistics of a DataFrame.- It provides statistical information such as count, mean, standard deviation, minimum, maximum, and quartile values for numerical columns in the DataFrame.



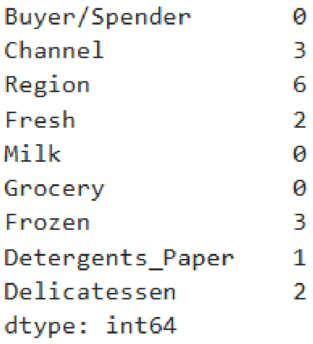
	count	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.0	220.50	127.16	1.0	110.75	220.5	330.25	440.0
Fresh	438.0	12016.01	12673.21	3.0	3111.25	8504.0	16935.25	112151.0
Milk	440.0	6035.78	8964.93	1.0	1525.25	3641.0	7217.50	112400.0
Grocery	440.0	7951.28	9503.16	3.0	2153.00	4755.5	10655.75	92780.0
Frozen	437.0	3085.64	4867.74	25.0	744.00	1535.0	3570.00	60869.0
Detergents_Paper	439.0	3773.75	19364.89	3.0	256.50	813.0	3956.00	396100.0
Delicatessen	438.0	1531.06	2825.04	3.0	411.25	971.0	1822.75	47943.0

Observations

- Spending on the Detergents Paper is the highest
- On checking the median values (50%), it appears that retailers spend more on Fresh products and grocery as compared to others.
- 75% of 440 retailers spend only 1820 or less on Delicatessen. So annual spend of Delicatessen appears to be least among all.

6. Check the null values

- df.isnull() is a method used in pandas, a Python library for data manipulation, to check for missing values in a DataFrame.- sum() is then applied to count the number of missing values in each column of the DataFrame
- The code df.isnull().sum() returns a Series showing the count of missing values in each column of the DataFrame df.



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

7. Check the duplicate values

- df.duplicated() checks for duplicated rows in the DataFrame df.- .sum() then sums up the boolean values returned by df.duplicated(), where True is counted as 1 and False as 0
- The result is the total count of duplicated rows in the DataFrame d

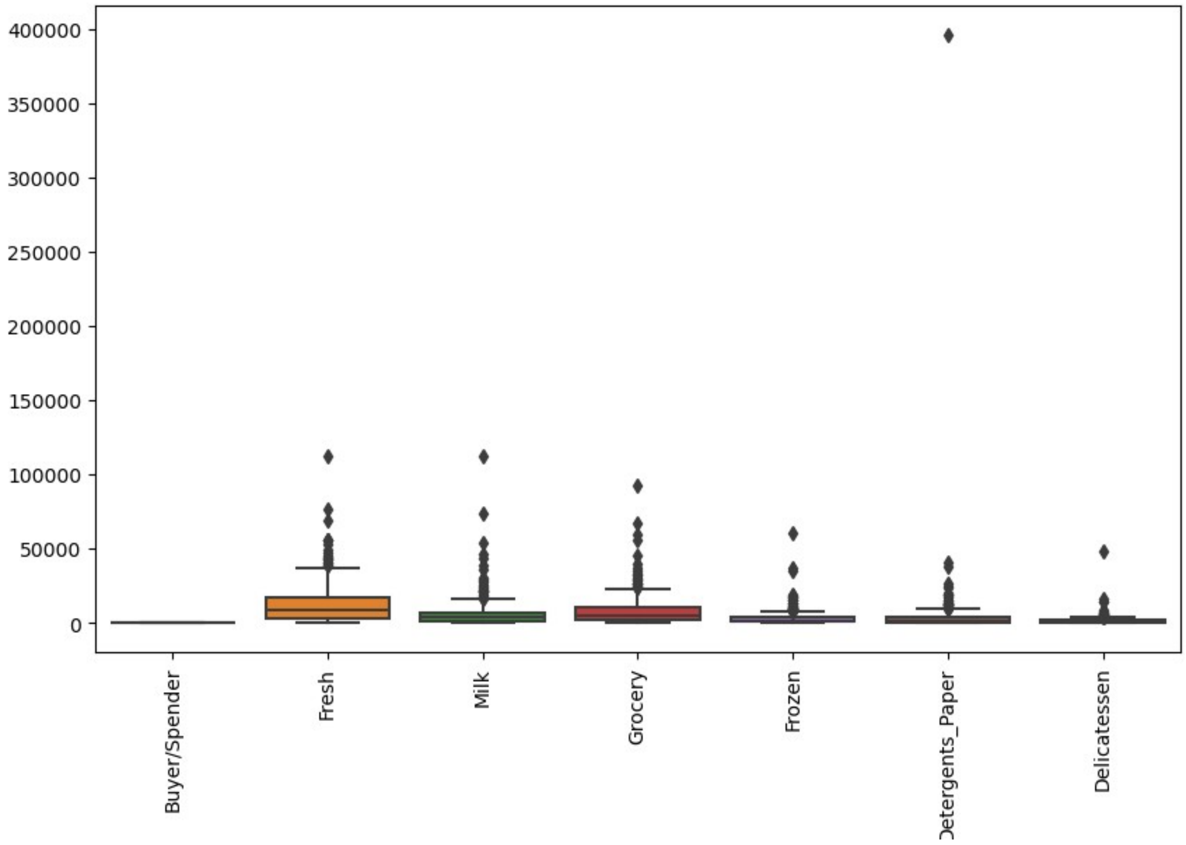
- . This code snippet is commonly used in data analysis to identify and count duplicate rows in a dataset.

```
0
```

8. Checking for anomalies or wrong entries

- You started by replacing any placeholder values ('?') with NaN (Not a Number), which is a standard way to represent missing data in Python.- You then either imputed these missing values (filling them in with estimated values like the mean) or removed rows with missing data to prepare the data for the clustering algorithm.

9. Check the outliers and their authenticity



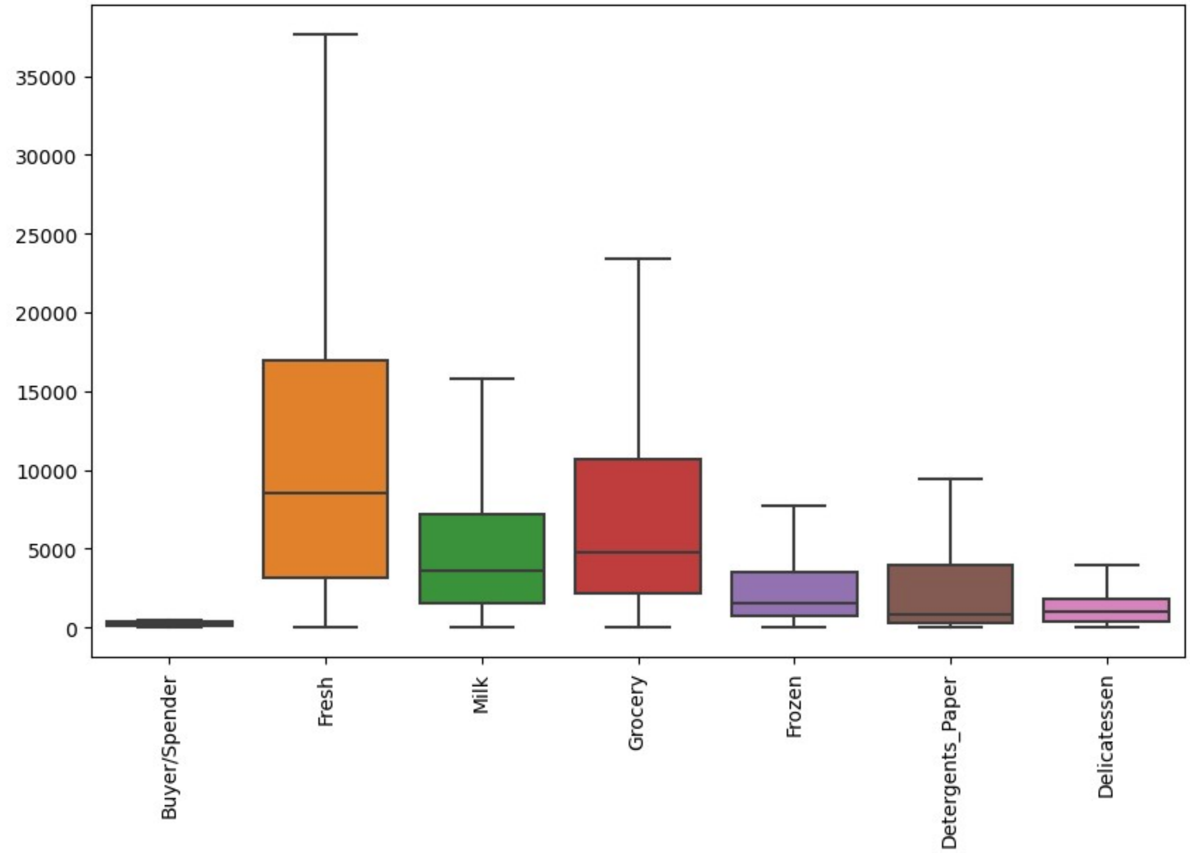
10. Do the necessary data cleaning steps like dropping duplicates, unnecessary columns, null value imputation, outliers treatment etc.

Data cleaning steps

Dropping duplicate rows- here we basically drop the duplicate row.

Handling Null Values

- Separate numeric and non-numeric columns
- Convert all numeric columns to numeric and coerce errors
- Fill numeric columns with the median
- Fill non-numeric columns with the most frequent value
- Treating outliers

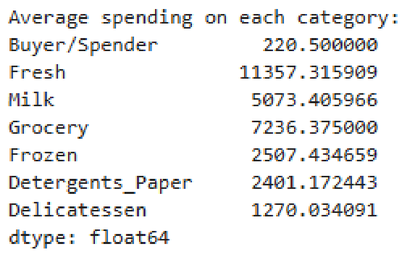


1. Spending Analysis

• 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:
Buyer/Spender         220.500000
Fresh               11357.315909
Milk                 5073.405966
Grocery              7236.375000
Frozen               2507.434659
Detergents_Paper     2401.172443
Delicatessen         1270.034091
dtype: float64
```

• Which category has the highest average spending?

Category with the highest average spending: Fresh

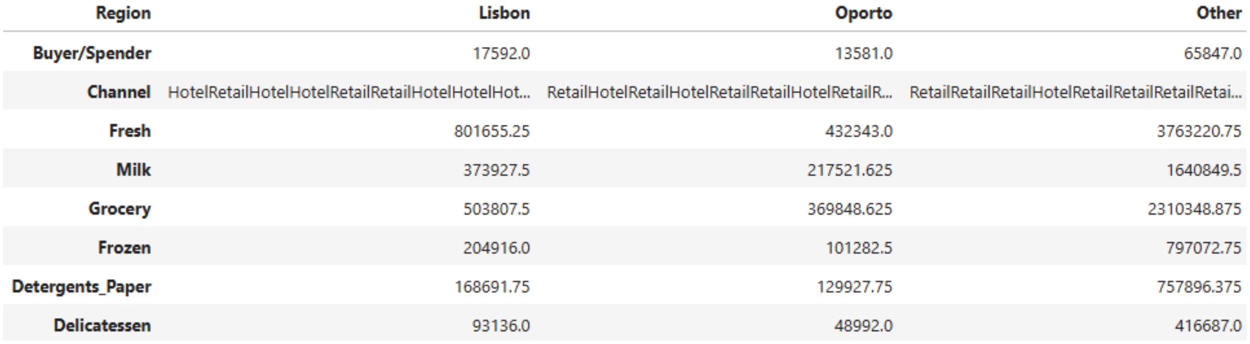
• How many buyers spend above the average on Fresh Vegetables?

- Total number of buyers: 440- Number of buyers who spend above the average on Fresh Vegetables: 170

2. Regional Demand

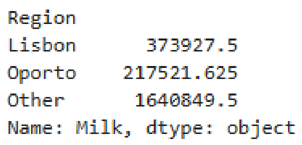
• What is the total spending in each region?

Total spending in each region:



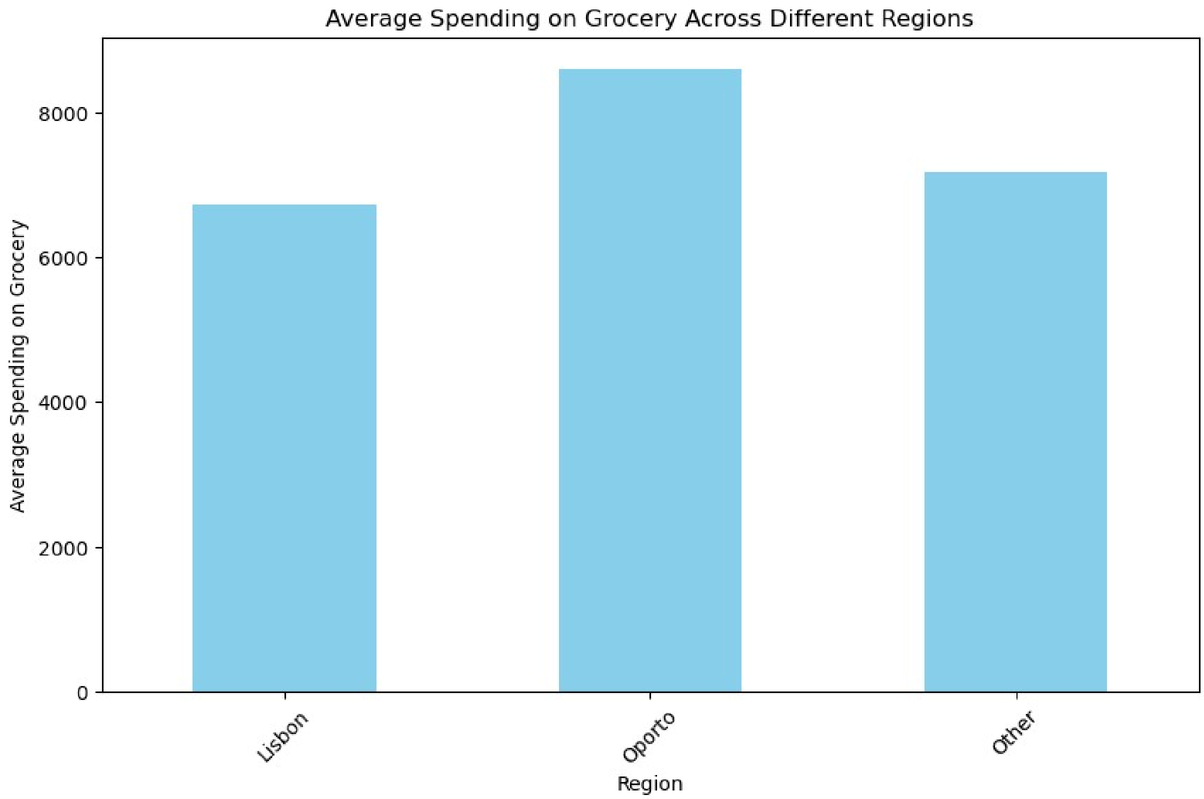
Region	Lisbon	Oporto	Other
Buyer/Spender	17592.0	13581.0	65847.0
Channel	HotelRetailHotelHotelRetailRetailHotelHotelHot...	RetailHotelRetailHotelRetailRetailHotelRetailR...	RetailRetailRetailHotelRetailRetailRetailRetai...
Fresh	801655.25	432343.0	3763220.75
Milk	373927.5	217521.625	1640849.5
Grocery	503807.5	369848.625	2310348.875
Frozen	204916.0	101282.5	797072.75
Detergents_Paper	168691.75	129927.75	757896.375
Delicatessen	93136.0	48992.0	416687.0

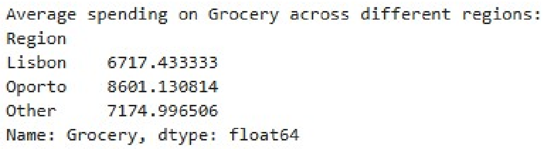
• Which region has the highest spending on Milk?



```
Region
Lisbon      373927.5
Oporto    217521.625
Other      1640849.5
Name: Milk, dtype: object
```

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





```
Average spending on Grocery across different regions:
Region
Lisbon     6717.433333
Oporto     8601.130814
Other      7174.996506
Name: Grocery, dtype: float64
```

• Which region has the highest average spending per buyer?

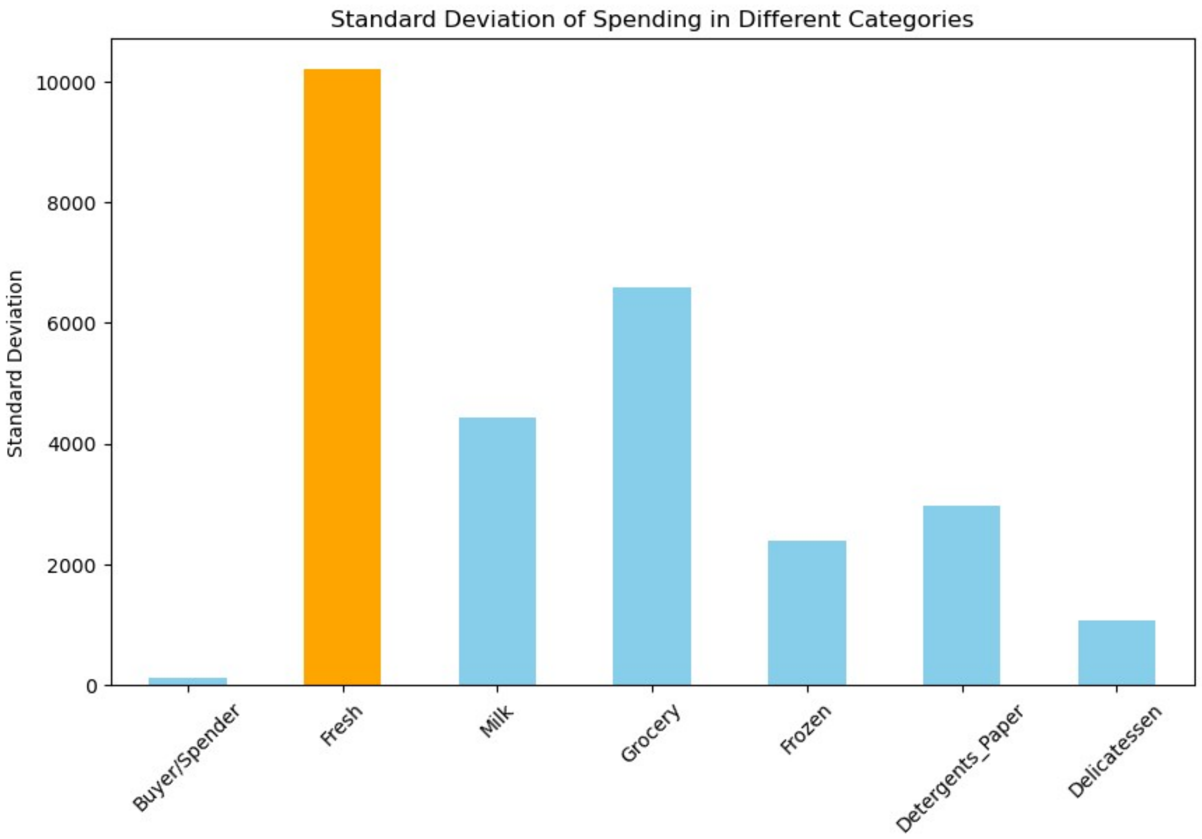
Region with the highest average spending per buyer: Oporto

3. Category Preferences

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

Percentage of buyers who spend more on Frozen food compared to Delicatessen: 66.14%

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

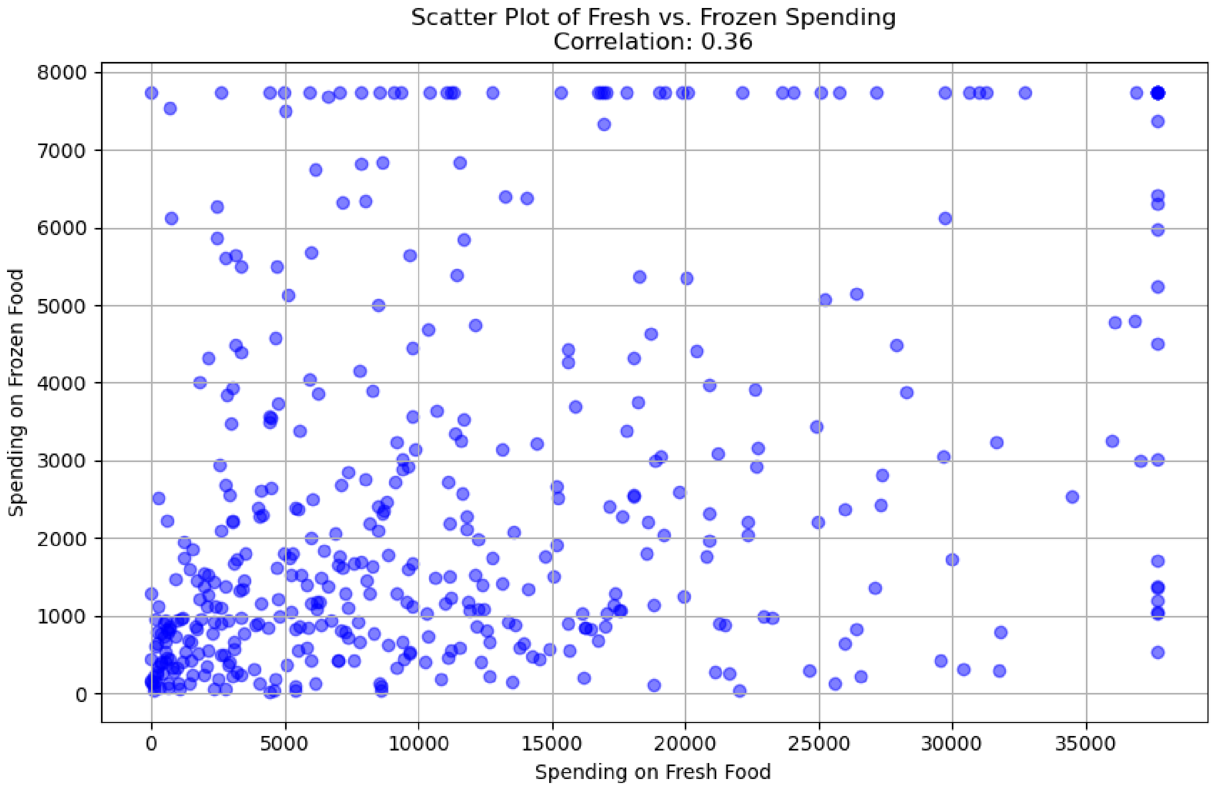


```
Category with the most variation in spending among buyers: Fresh
```

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

Region with significantly higher spending on Detergents_paper: Oporto

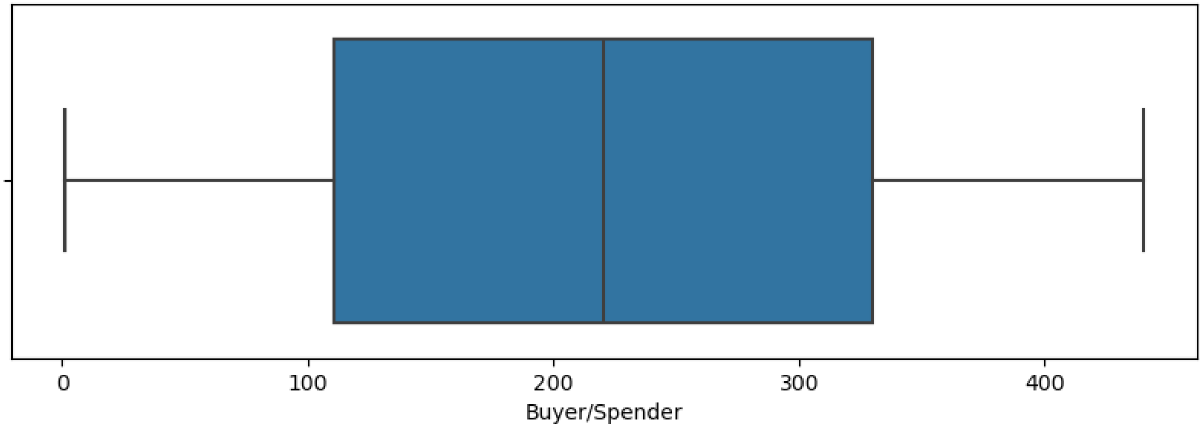
- What is the correlation between spending on Fresh and Frozen food?



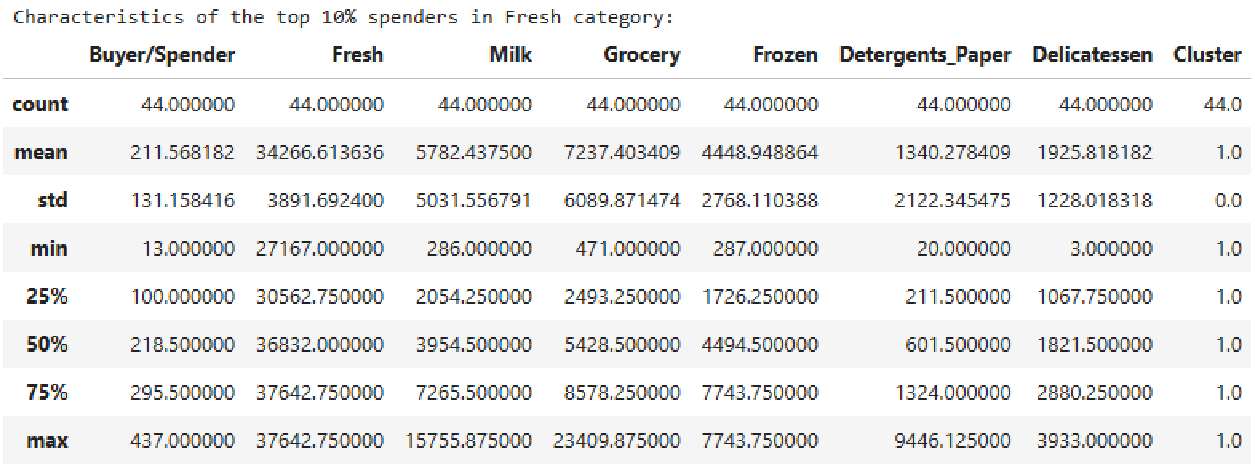
```
Correlation between spending on Fresh and Frozen food: 0.3559051601660422
```

4. Customer Segmentation

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



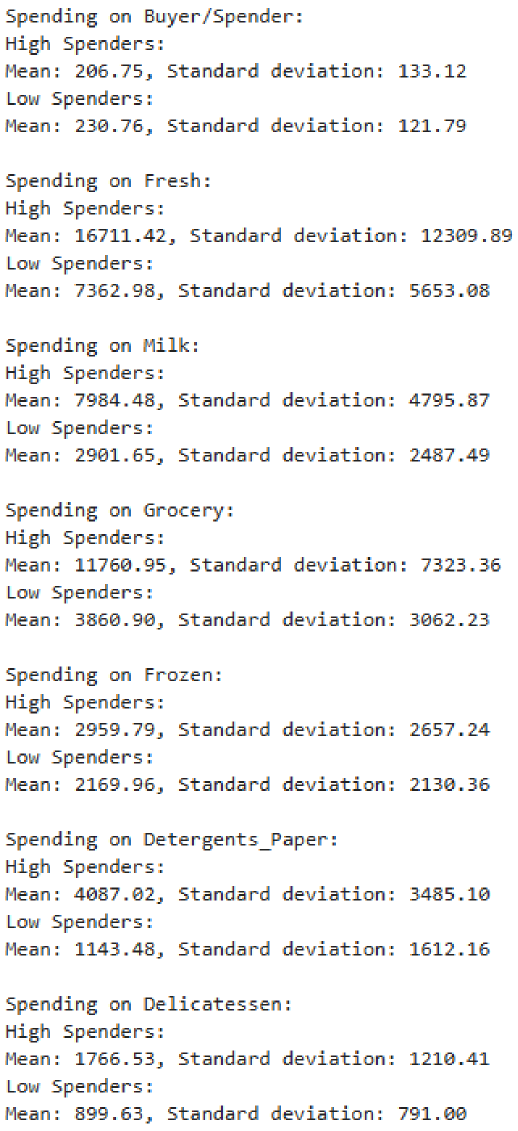
• What are the characteristics of the top 10% spenders in each category?



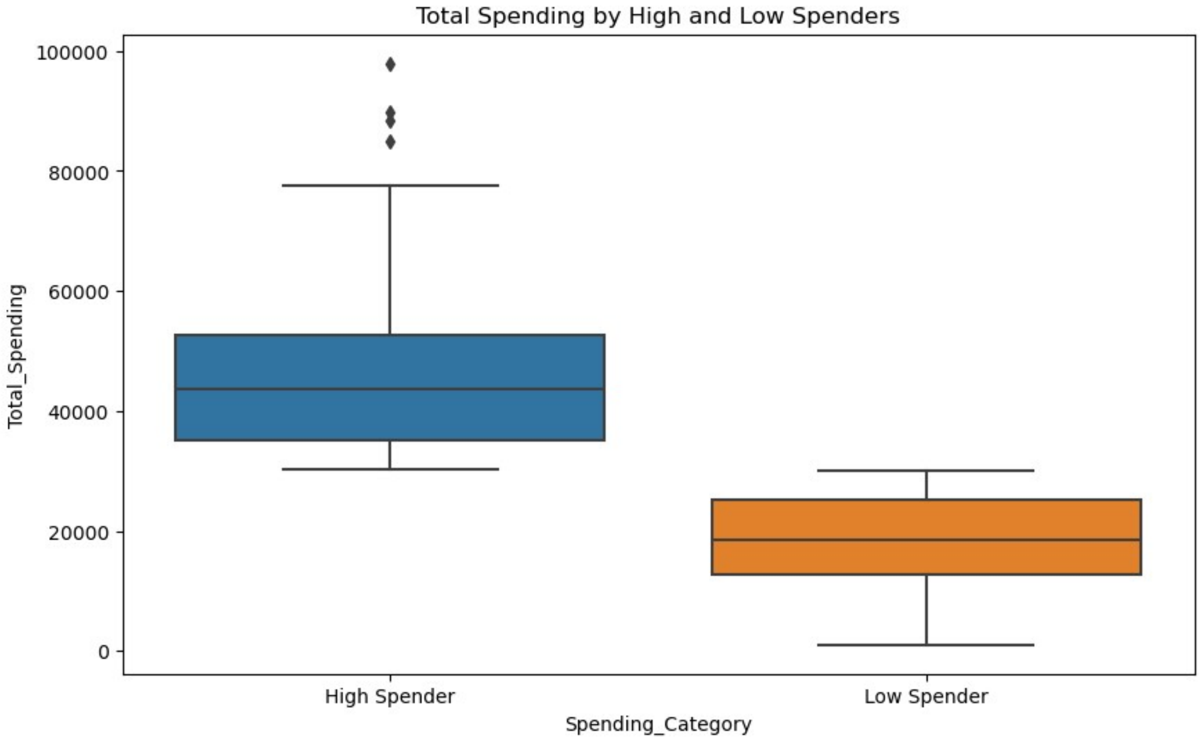
```
Characteristics of the top 10% spenders in Fresh category:
```

	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Cluster
count	44.000000	44.000000	44.000000	44.000000	44.000000	44.000000	44.000000	44.0
mean	211.568182	34266.613636	5782.437500	7237.403409	4448.948864	1340.278409	1925.818182	1.0
std	131.158416	3891.692400	5031.556791	6089.871474	2768.110388	2122.345475	1228.018318	0.0
min	13.000000	27167.000000	286.000000	471.000000	287.000000	20.000000	3.000000	1.0
25%	100.000000	30562.750000	2054.250000	2493.250000	1726.250000	211.500000	1067.750000	1.0
50%	218.500000	36832.000000	3954.500000	5428.500000	4494.500000	601.500000	1821.500000	1.0
75%	295.500000	37642.750000	7265.500000	8578.250000	7743.750000	1324.000000	2880.250000	1.0
max	437.000000	37642.750000	15755.875000	23409.875000	7743.750000	9446.125000	3933.000000	1.0

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

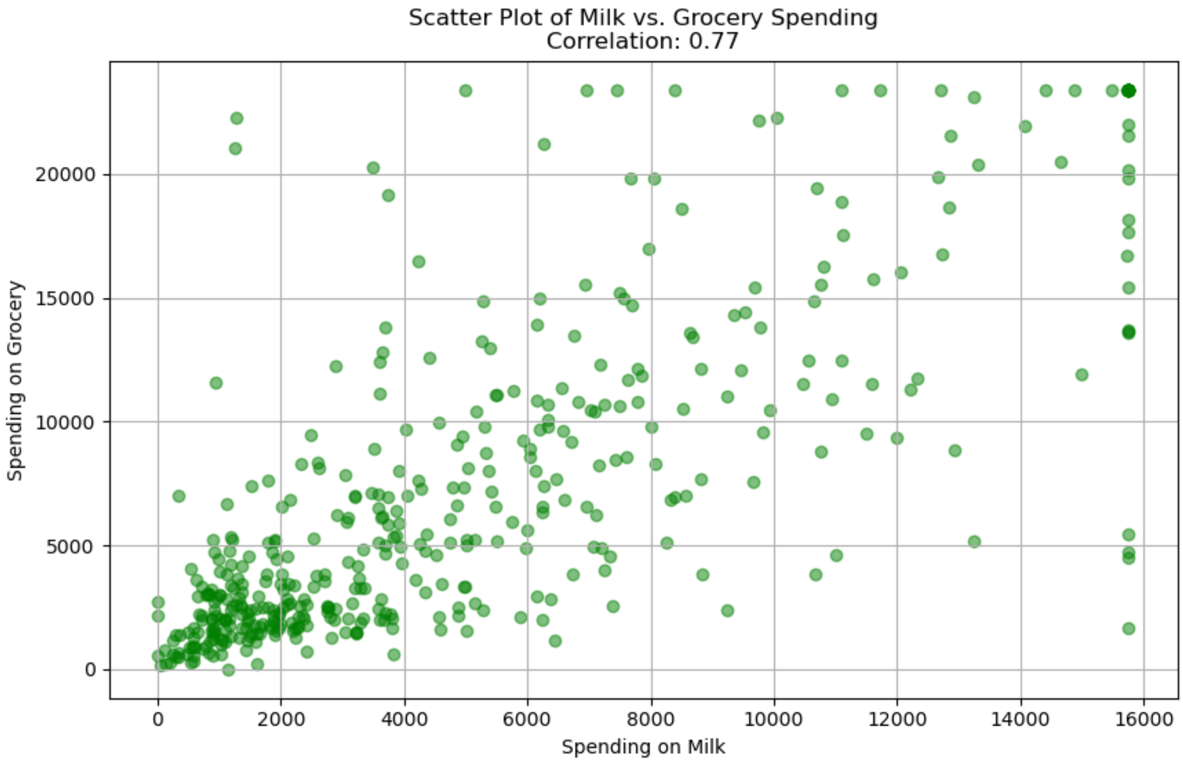


```
Spending on Buyer/Spender:
High Spenders:
Mean: 206.75, Standard deviation: 133.12
Low Spenders:
Mean: 230.76, Standard deviation: 121.79

Spending on Fresh:
High Spenders:
Mean: 16711.42, Standard deviation: 12309.89
Low Spenders:
Mean: 7362.98, Standard deviation: 5653.08

Spending on Milk:
High Spenders:
Mean: 7984.48, Standard deviation: 4795.87
Low Spenders:
Mean: 2901.65, Standard deviation: 2487.49

Spending on Grocery:
High Spenders:
Mean: 11760.95, Standard deviation: 7323.36
Low Spenders:
Mean: 3860.90, Standard deviation: 3062.23

Spending on Frozen:
High Spenders:
Mean: 2959.79, Standard deviation: 2657.24
Low Spenders:
Mean: 2169.96, Standard deviation: 2130.36

Spending on Detergents_Paper:
High Spenders:
Mean: 4087.02, Standard deviation: 3485.10
Low Spenders:
Mean: 1143.48, Standard deviation: 1612.16

Spending on Delicatessen:
High Spenders:
Mean: 1766.53, Standard deviation: 1210.41
Low Spenders:
Mean: 899.63, Standard deviation: 791.00
```



5. Cross-Category Analysis

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

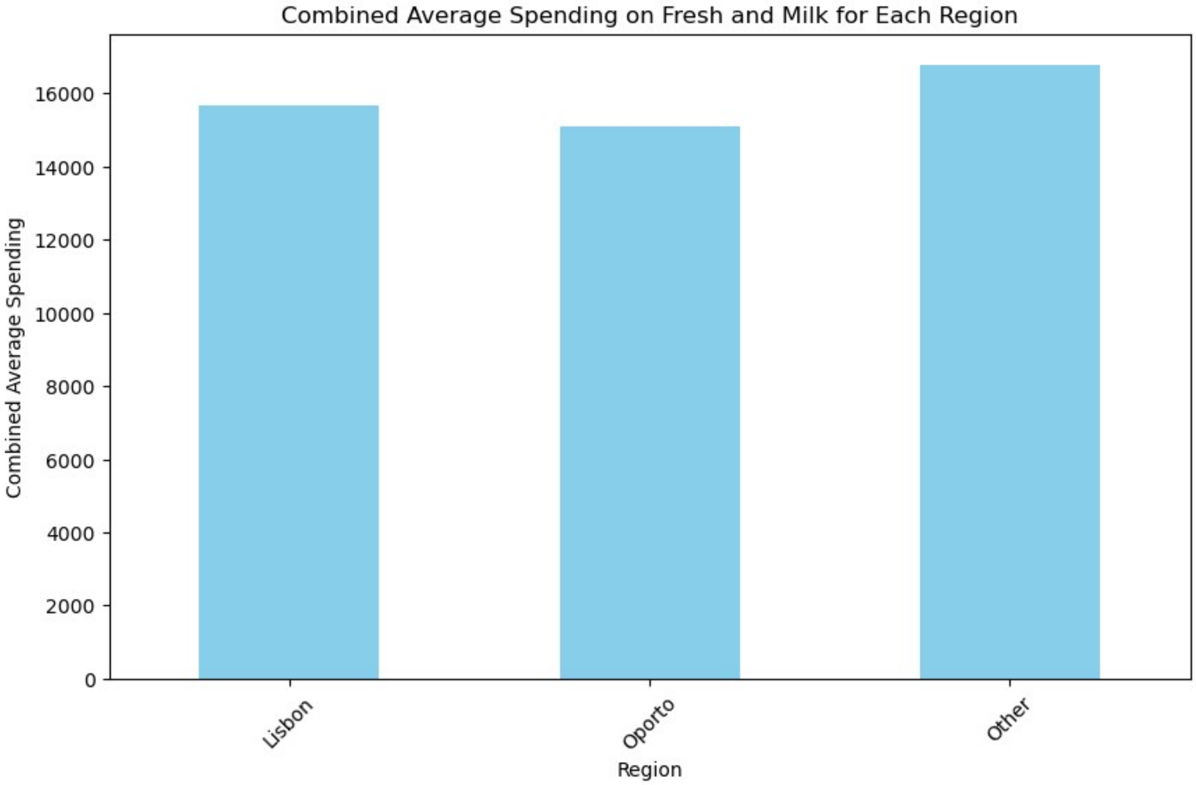


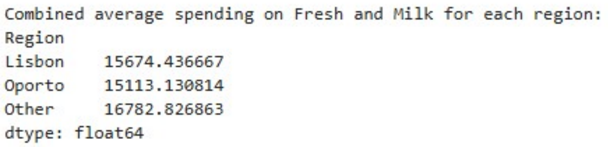
```
Correlation between spending on Milk and Grocery: 0.7737423597168179
```

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

Number of buyers who spend more on Delicatessen and Frozen food: 71

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

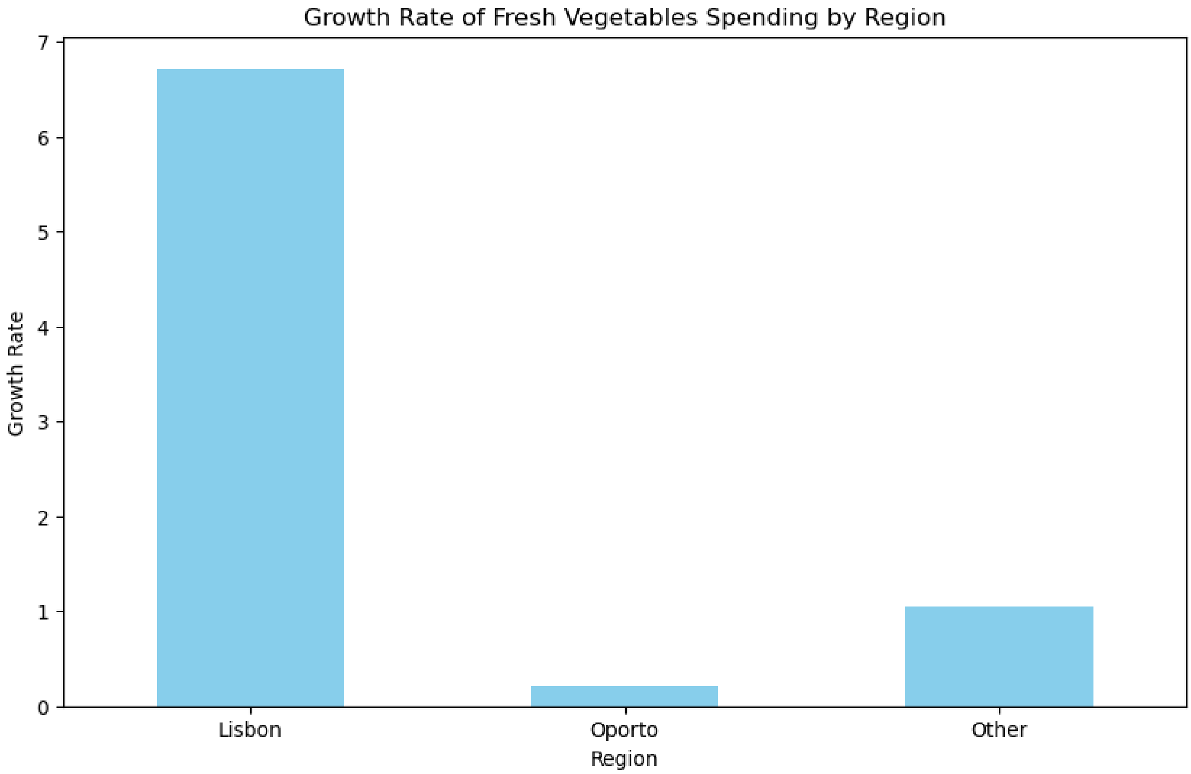




```
Combined average spending on Fresh and Milk for each region:
Region
Lisbon    15674.436667
Oporto    15113.130814
Other     16782.826863
dtype: float64
```

6. Demand Trends

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



```
Region with the fastest growing spending on Fresh Vegetables: Lisbon
```

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

Time data is 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

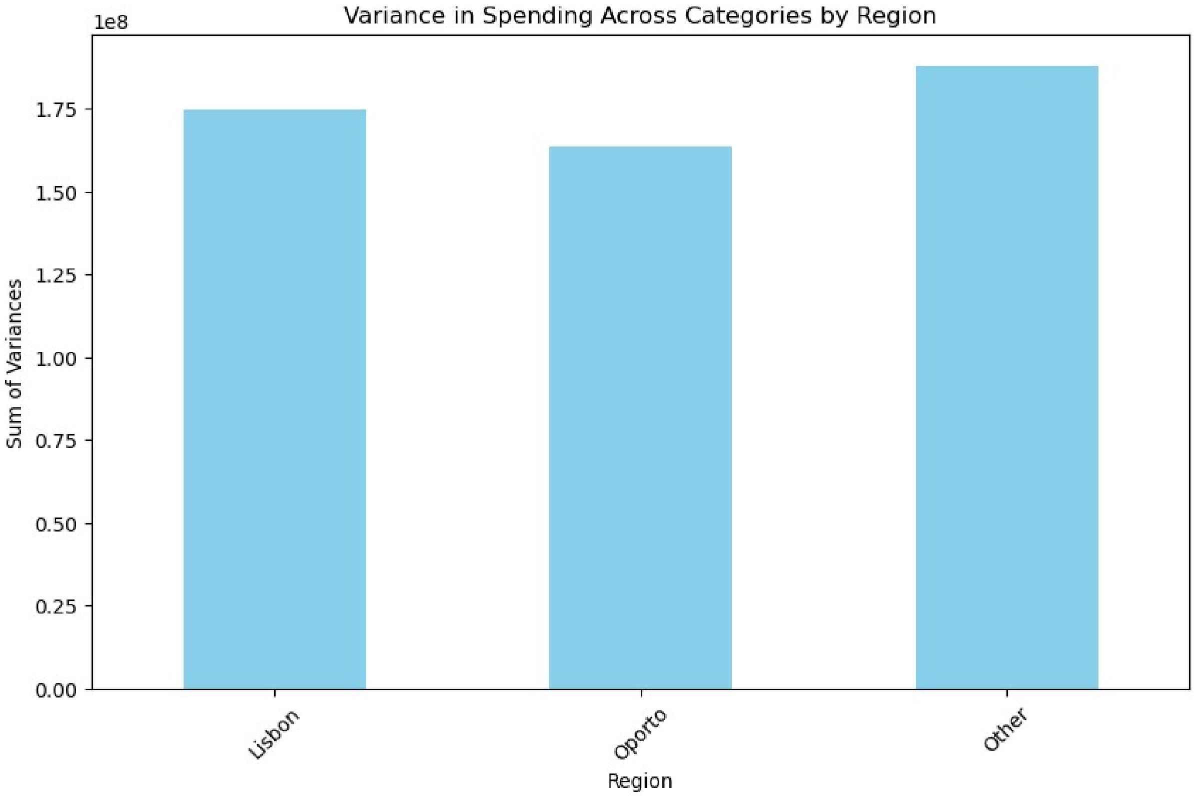
• What is the repeat purchase rate for buyers who spend above the average in at least three categories?

Number of buyers who spend above the average in at least three categories: 39

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

Number of buyers who spend consistently across all categories: 0

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



```
Region with the most diverse spending patterns: Other
```