Exploratory Data Analysis (EDA) Assignment

Introduction

I completed the exploratory data analysis for the WholeSaleCustomer dataset, or data of clients of a wholesale distributor.

Preprocess

During the preprocessing phase, there was no need to remove missing values because there were no missing values. Each product had a "Count" of 440, as shown by the Descriptive Statistics via the Data Analysis Tools. Below is an example of the Descriptive Statistics for Fresh products:

Fresh	
Mean	12,000
Standard Error	603
Median	8,504
Mode	9,670
Standard Deviation	12,647
Sample Variance	159,954,927
Kurtosis	12
Skewness	3
Range	112,148
Minimum	3
Maximum	112,151
Sum	5,280,131
Count	440

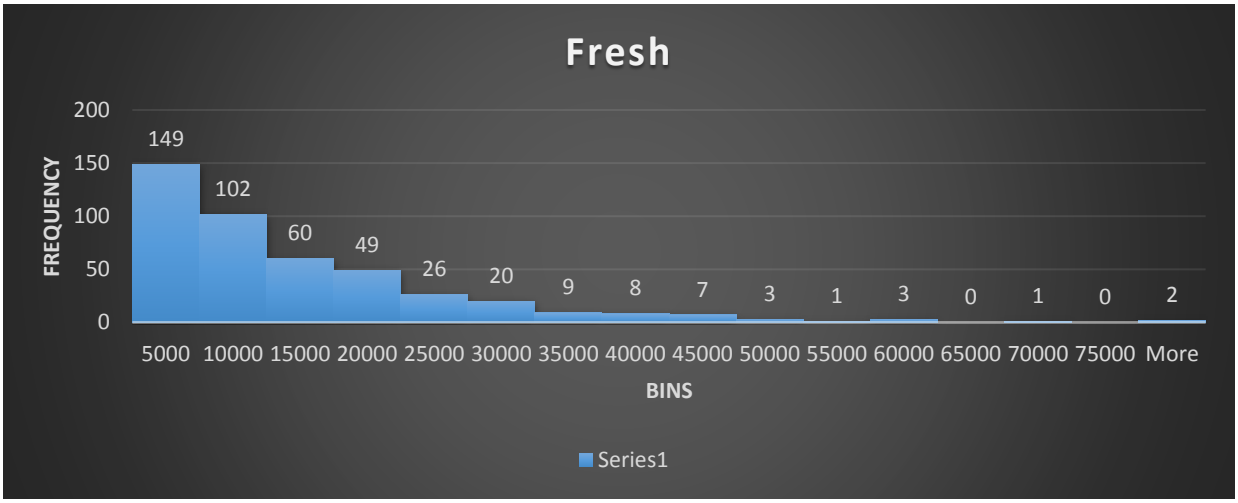
Summary: Descriptive Statistics

Descriptive statistics is a graphical and numerical procedure to summarize and process data. Below is a table with the descriptive statistics (min, max, mean, median, and quartiles) of each product:

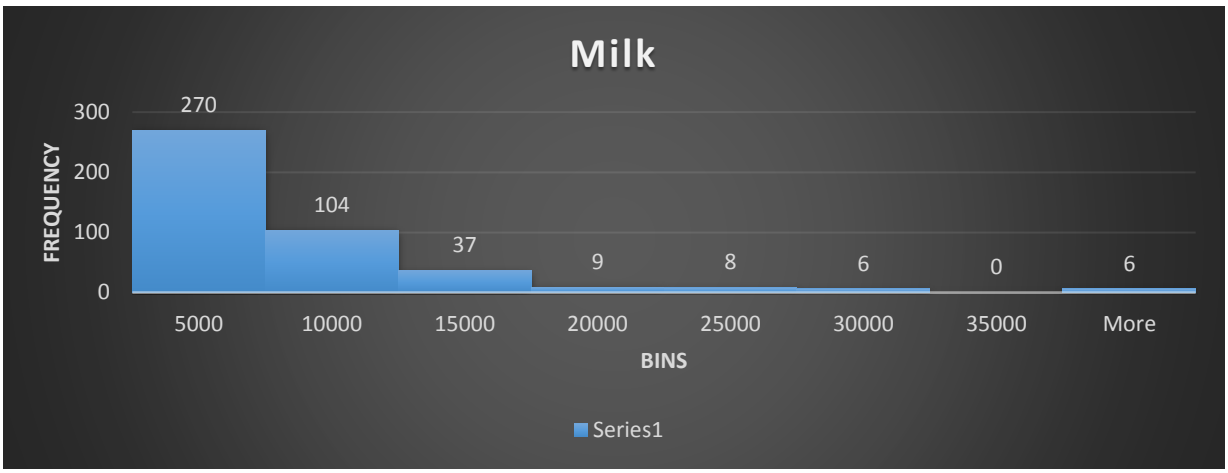
Descriptive Statistics						
Product	**Fresh**	**Milk**	**Grocery**	**Frozen**	**Detergents_Paper**	**Delicatessen**
Minimum	3	55	3	25	3	3
Q1	3,128	1,533	2,153	742	257	408
Median	8,504	3,627	4,756	1,526	817	966
Q3	16,934	7,190	10,656	3,554	3,922	1,820
Maximum	112,151	73,498	92,780	60,869	40,827	47,943
Mean	12,000	5,796	7,951	3,072	2,881	1,525
Range	112,148	73,443	92,777	60,844	40,824	47,940

Histograms

A graph of the data in a frequency distribution is called a histogram. Below is a histogram of each product:

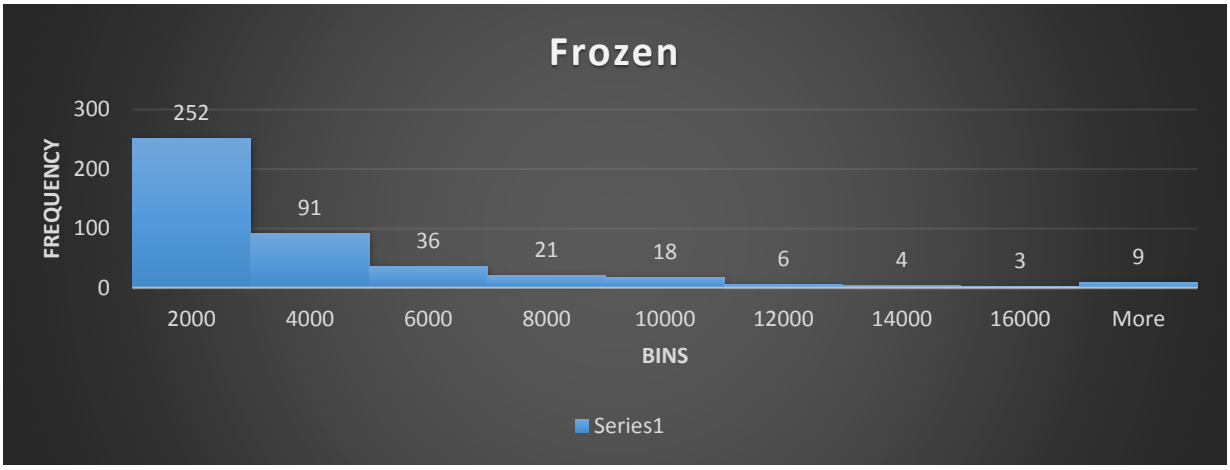


Histogram 1: Fresh

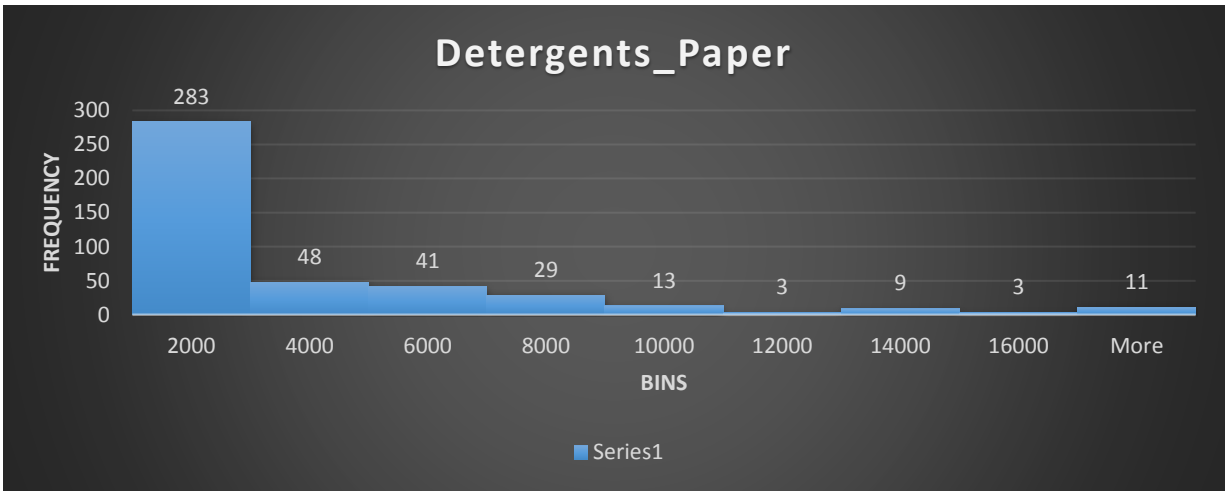


Histogram 2: Milk

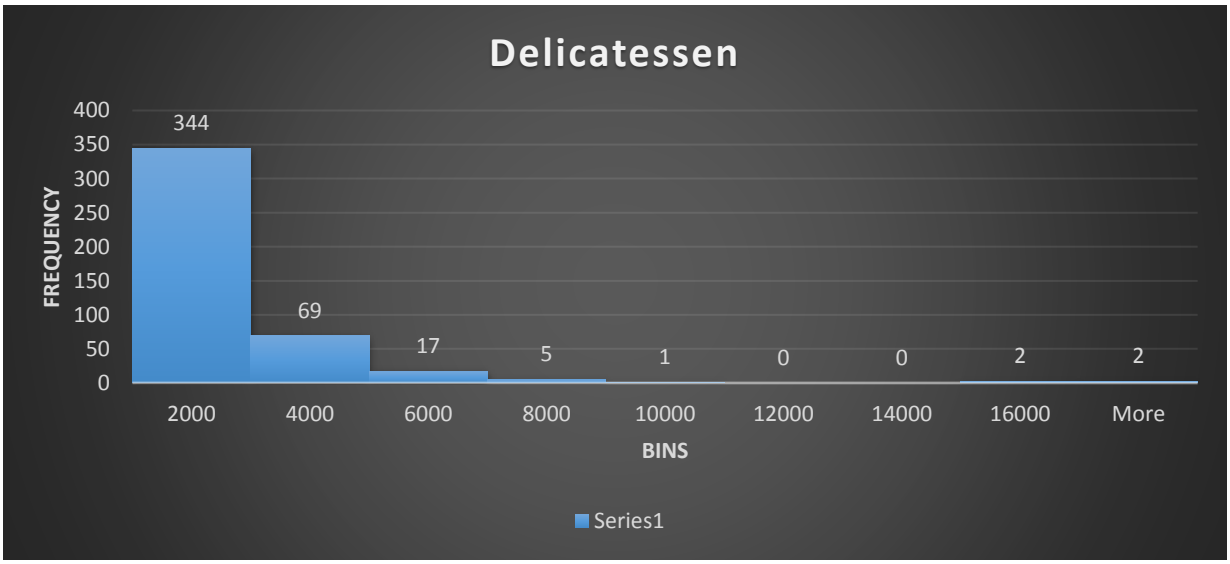
Histogram 3: Grocery



Histogram 4: Frozen



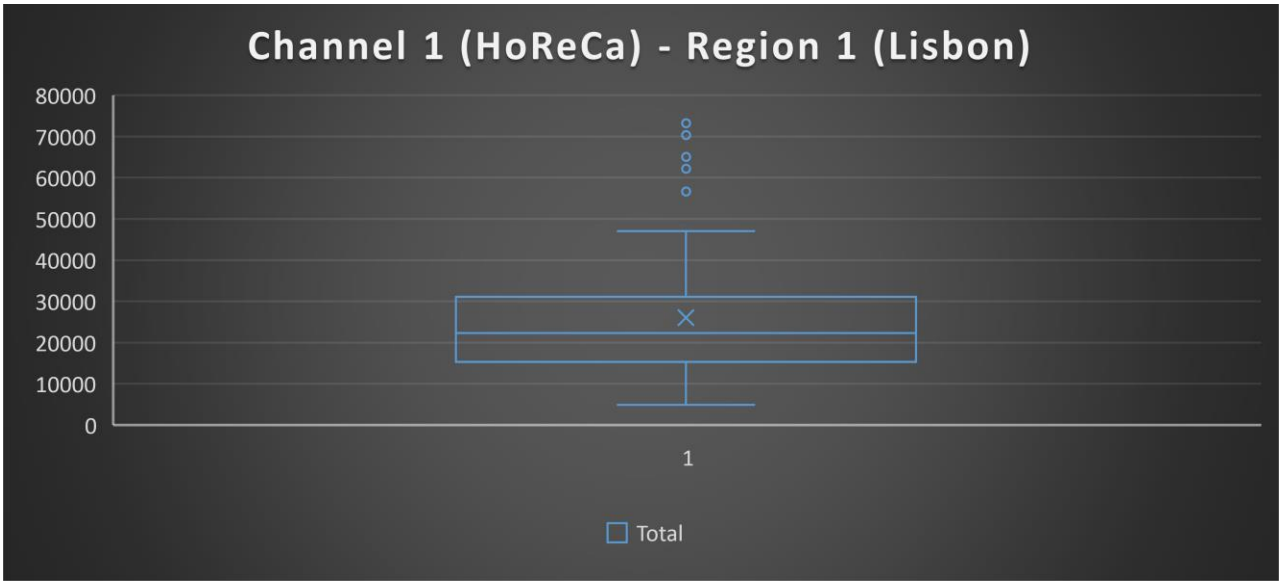
Histogram 5: Detergents_Paper



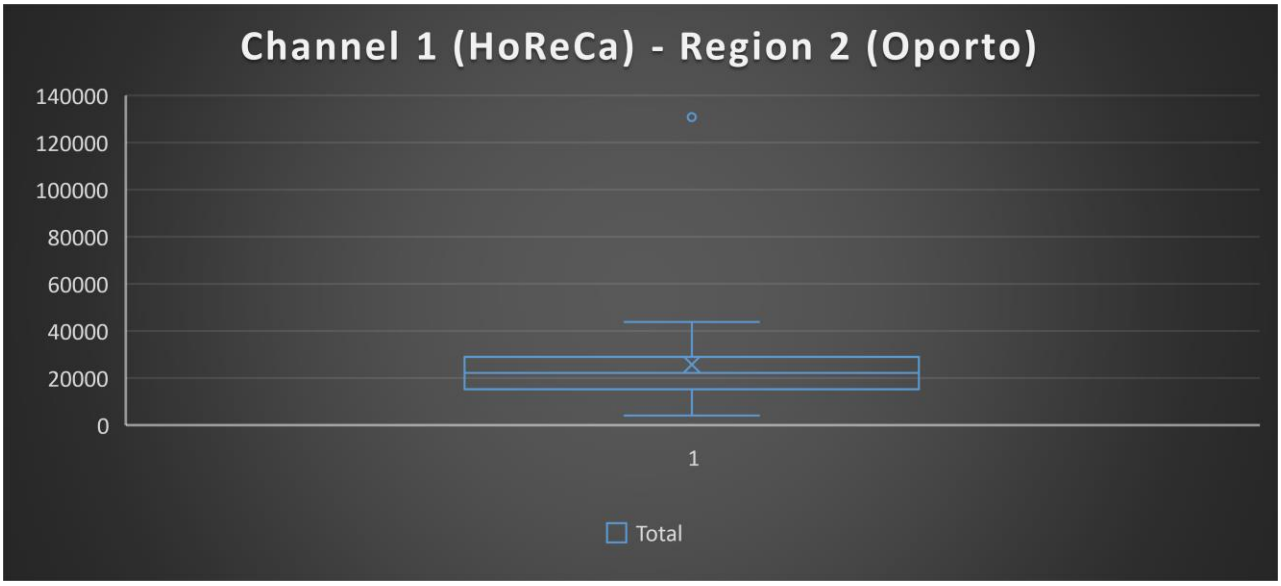
Histogram 6: Delicatessen

Boxplots

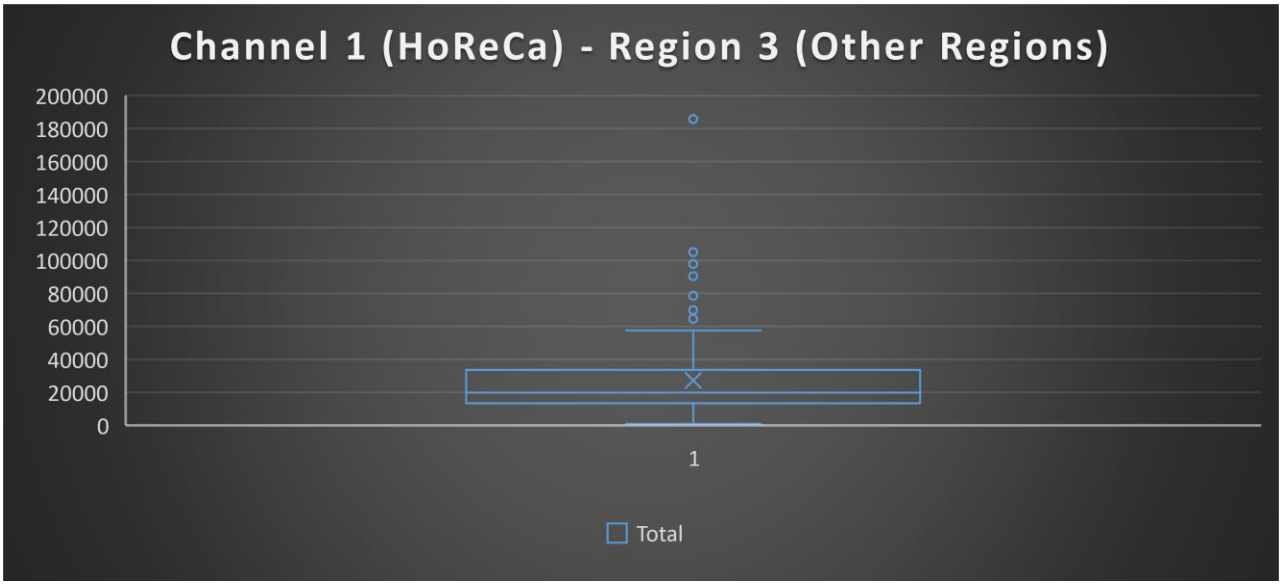
A boxplot is a graphical illustration of the minimum, first quartile, median, third quartile, and maximum of a dataset. Below is a boxplot of each channel and region combination of total product spending (aggregate of Fresh, Milk, Grocery, Frozen, Detergents_Paper, and Delicatessen):



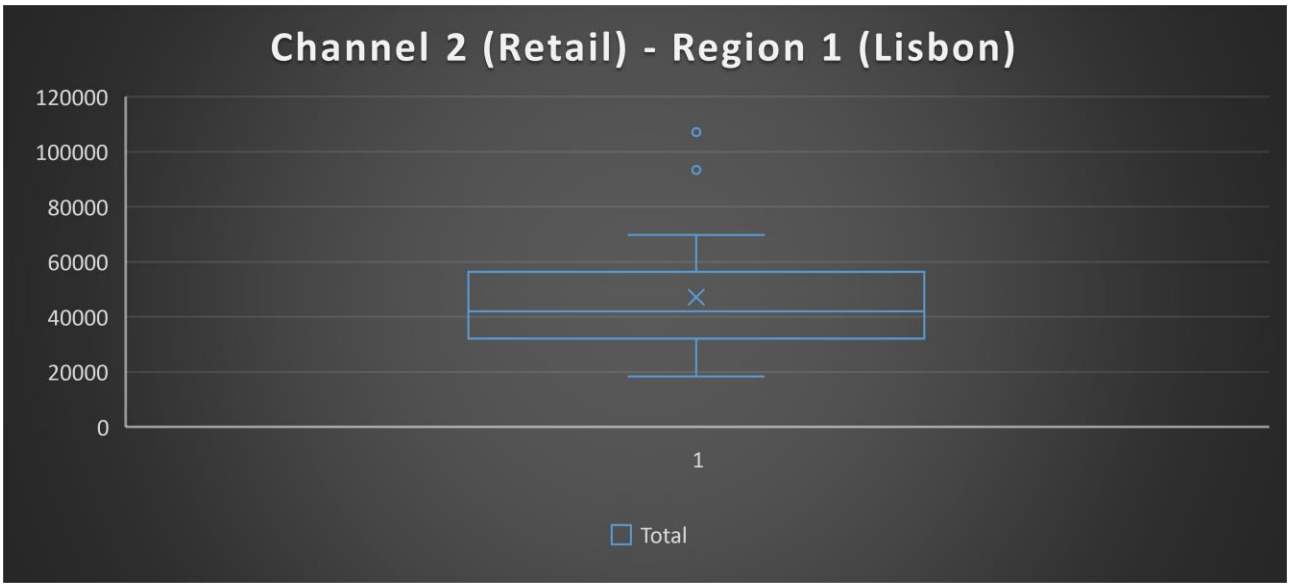
Boxplot 1: Channel 1 (HoReCa) – Region 1 (Lisbon)



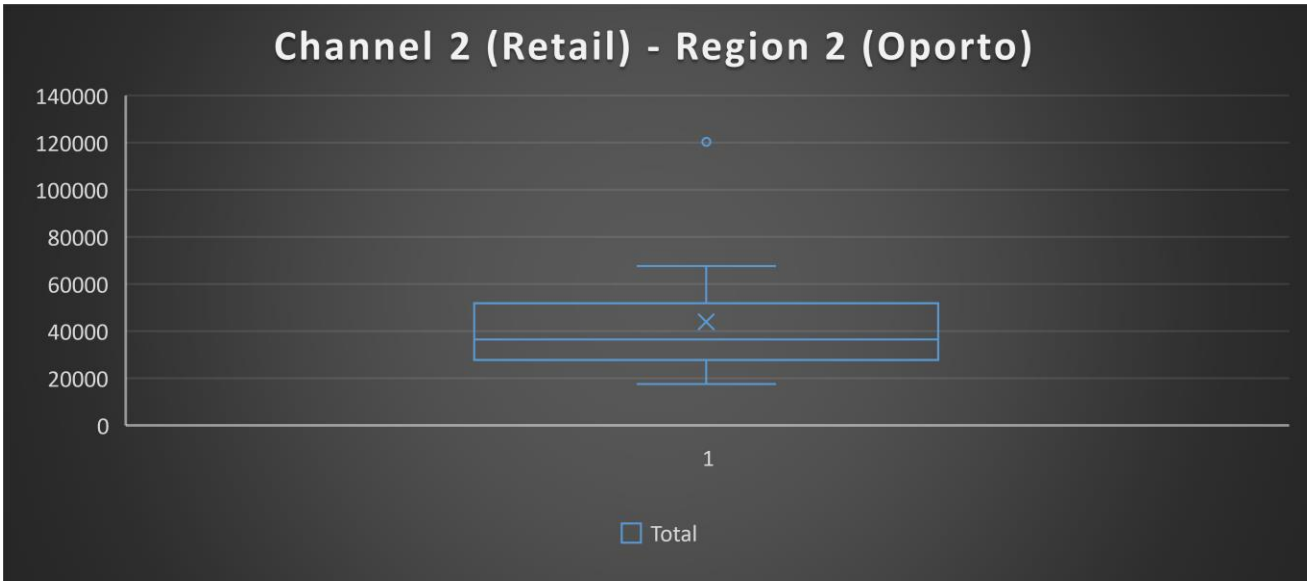
Boxplot 2: Channel 1 (HoReCa) – Region 2 (Oporto)



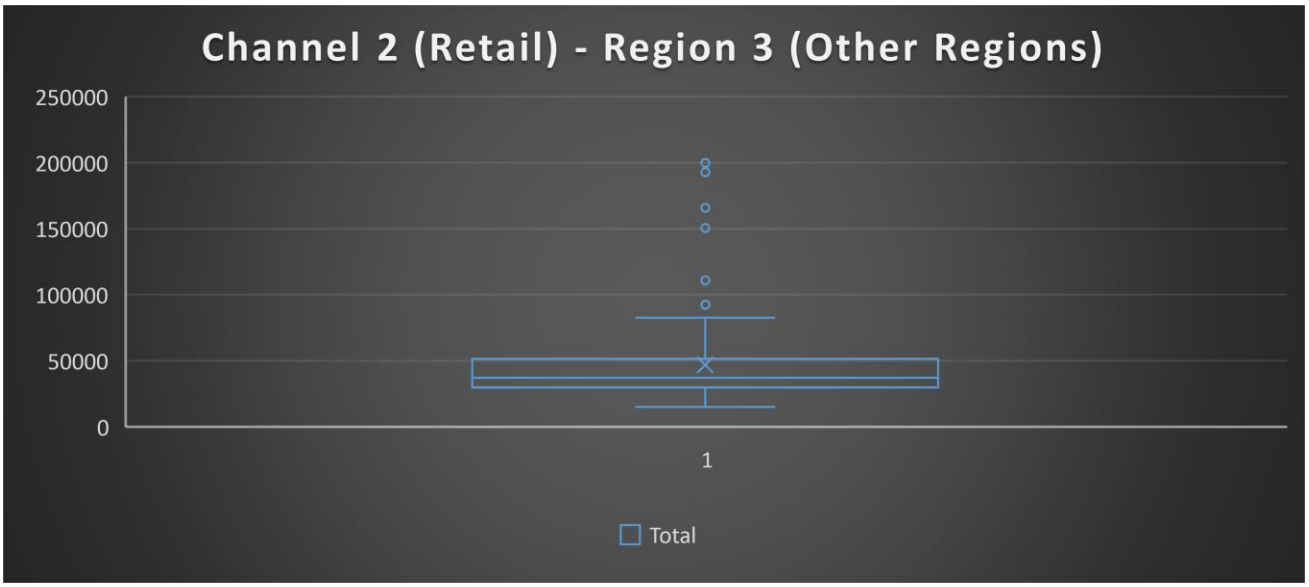
Boxplot 3: Channel 1 (HoReCa) – Region 3 (Other Regions)



Boxplot 4: Channel 2 (Retail) – Region 1 (Lisbon)



Boxplot 5: Channel 2 (Retail) – Region 2 (Oporto)



Boxplot 6: Channel 2 (Retail) – Region 3 (Other Regions)

Delicatessen

After calculating the inter quartile range (IQR), lower limit (LL), and higher limit (HL) for the Delicatessen product, I decided to exclude the values less than the lower level of -1,710 and exclude the values greater than the higher level of 3,938. Below is a table representing the calculation of the aforementioned metrics and a table representing the results:

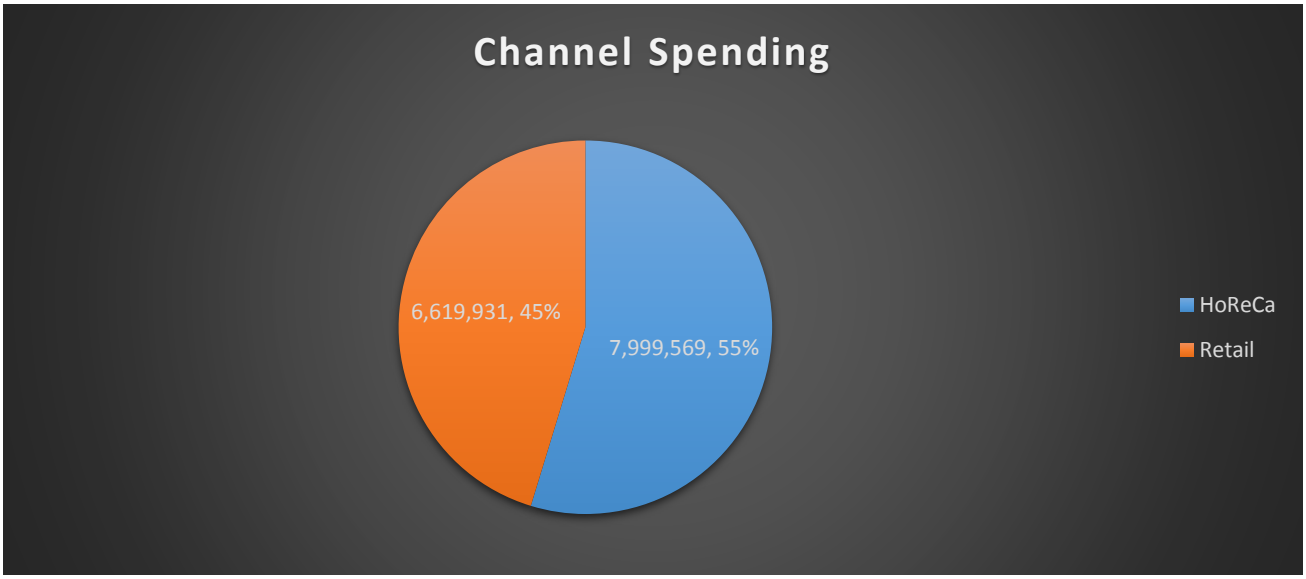
Calculations	
IQR	Q3-Q1
LL	Q1-(IQR*1.5)
HL	Q3+(IQR*1.5)

Delicatessen Outliers Method	
Q1	408
Q3	1,820
IQR	1,412
LL	-1,710
HL	3,938

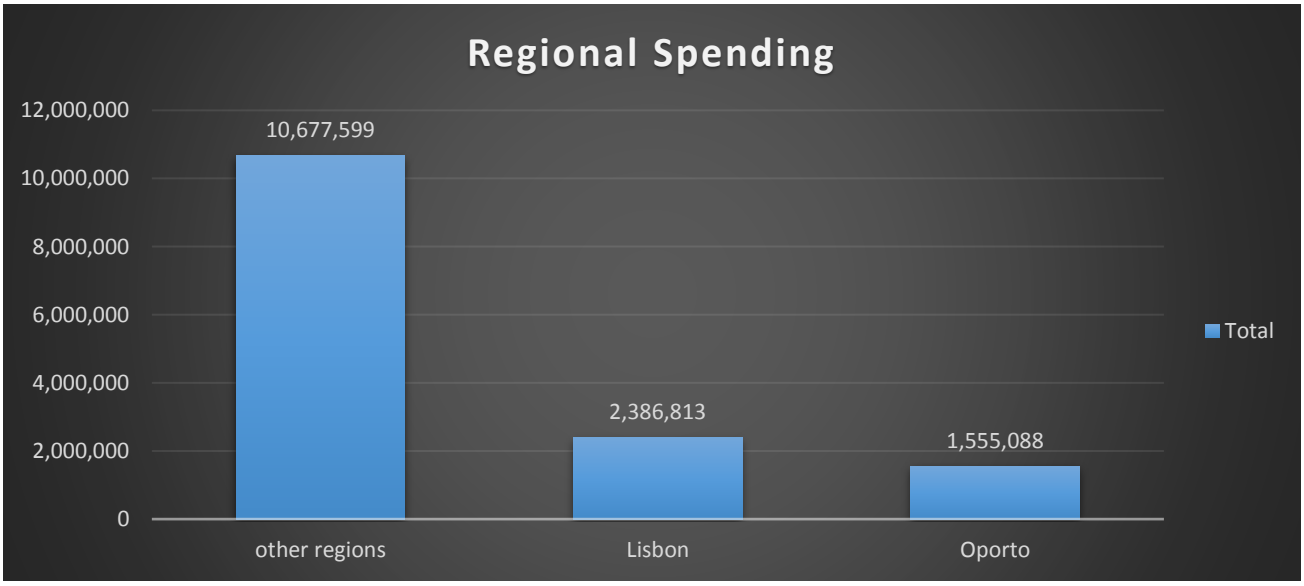
Conclusion

After completing the exploratory data analysis for the "WholeSaleCustomers" dataset, I found a few pertinent conclusions:

1) Aggregate spending was roughly 14.6 million. 55% of aggregate spending was through the HoReCa channel and the remaining 45% was through the Retail channel



2) Aggregate spending of both Lisbon and Oporto, or roughly 3.9 million, was significantly less than the spending for other regions, which was roughly 10.7 million



3) Whereas the product with the most annual spending is Fresh products, the product with the least annual spending is Delicatessen products