Exploratory Data Analysis (EDA) Assignment

Sarang Zendehrooh

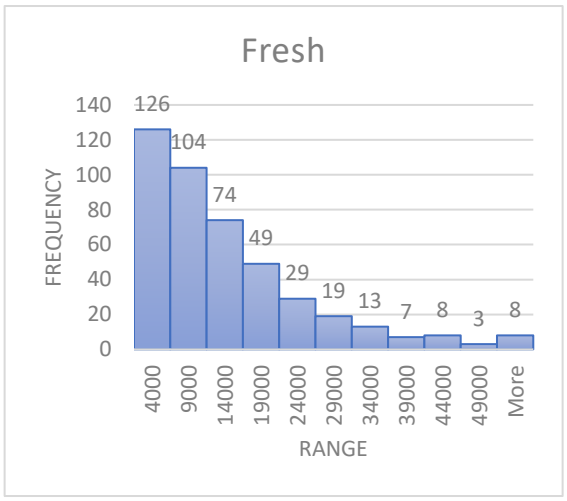
Q1- Preprocess you data, by removing missing values if needed. No missing data detected. In this data set we have 2 columns of categorical data which have been converted into numerical data type (channel and region). The process of converting categorical data type to numeric type can be very important for many processes in data science, particularly in machine learning models. The other 6 columns are numerical continues data type.

Q2- Present a table with the descriptive statistics of all products spending (min, max, mean, median, quartiles). Descriptive analysis of all products spending generated by utilising data analysis add-on in excel plus Quartile formula. The descriptive statistic enables users to get a better understanding of the dataset.

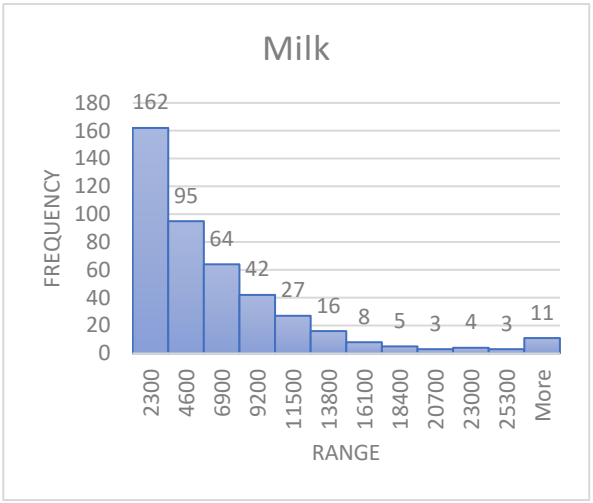
I have applied conditional formatting on mean, median, Quartile 1 to 4 and sum rows to help users visually explore and analyse data. We can see the spending on fresh products seems to be higher and Delicassen lower compared to all the products on average and total spending. Fresh product has a larger standard deviation which indicates values are spread out in a wider range whereas the Delicassen has smaller standard deviation which indicates the values are much closer to its mean. We have equal number of data points for each product as we can see in the count row.

Description	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
Mean	12000	5796	7951	3072	2881	1525
Median	8504	3627	4756	1526	817	966
Mode	9670	1012	2062	937	284	834
Range	112148	73443	92777	60844	40824	47940
Minimum	3	55	3	25	3	3
Quartile 1	3128	1533	2153	742	257	408
Median(Quartile 2)	8504	3627	4756	1526	817	966
Quartile 3	16934	7190	10656	3554	3922	1820
Maximum(Quartile 4)	112151	73498	92780	60869	40827	47943
Standard Deviation	12647	7380	9503	4855	4768	2820
Sample Variance	159954927	54469967	90310104	23567853	22732436	7952997
Kurtosis	12	25	21	55	19	171
Skewness	3	4	4	6	4	11
Sum	5280131	2550357	3498562	1351650	1267857	670943
Count	440	440	440	440	440	440
Standard Error	603	352	453	231	227	134

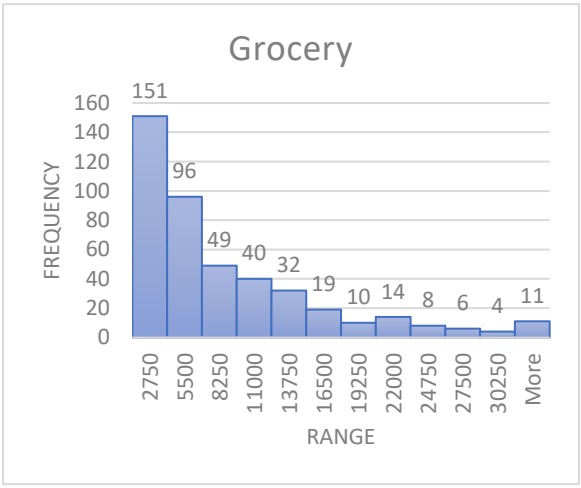
***Q3 - Present plots of histograms of the various product spending's, respecting good practice in graph presentation*.** Histogram is a frequency distribution chart that highlights how often each different value in a set of data occurs. Below I have plotted Histogram graph for each spending utilising excel data analysis add-on.



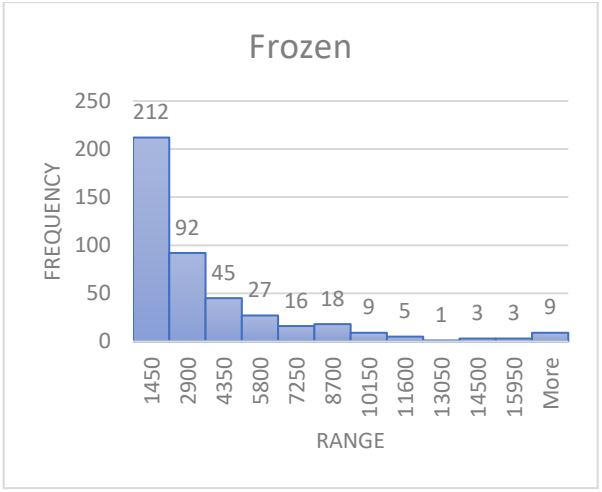
Histogram 1 : Fresh Spending



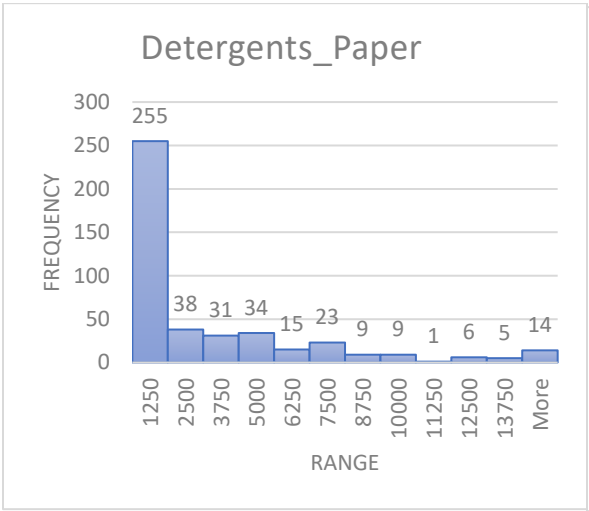
Histogram 2 :Milk Spending



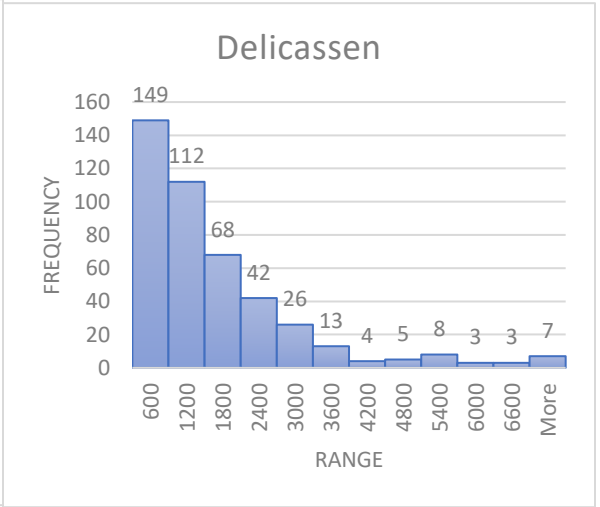
Histogram 3 :Grocery Spending



Histogram 4 :Frozen Spending



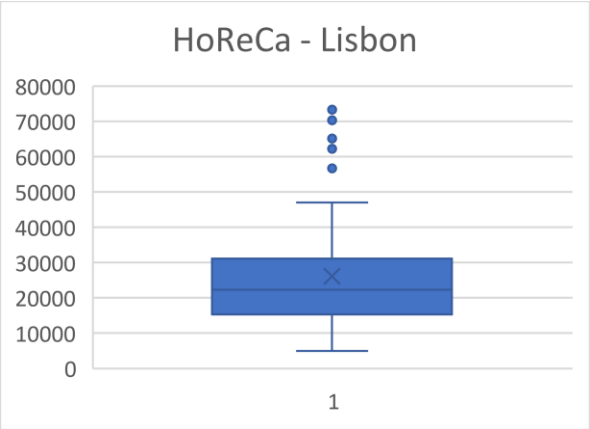
Histogram 5 : Detergents_Paper Spending



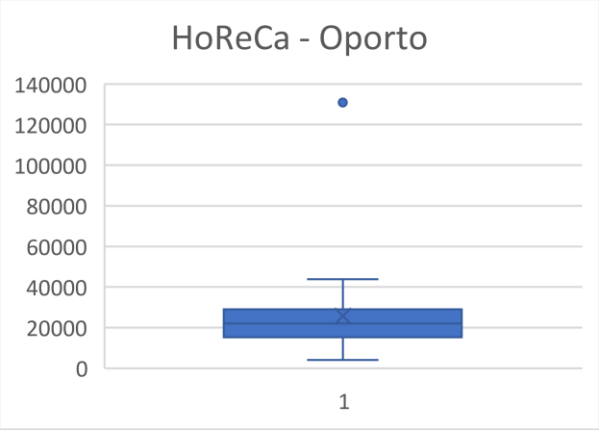
Histogram 6 : Delicassen Spending

Q4 - Present boxplots of the product spending for each channel and region (6 plots in total)?

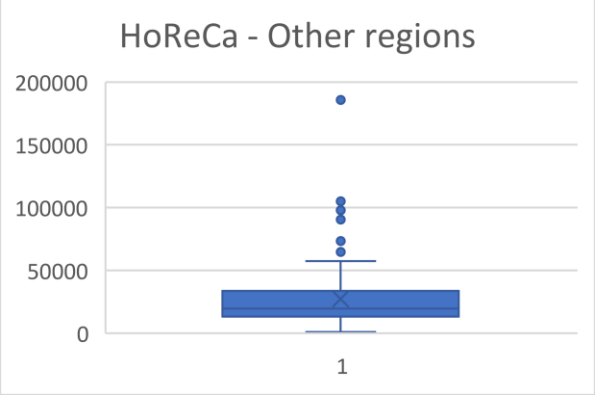
A boxplot displays the minimum, first quartile, median, third quartile, and maximum of the given data. Below I have plotted a boxplot for combination of each channel and region total product spending:



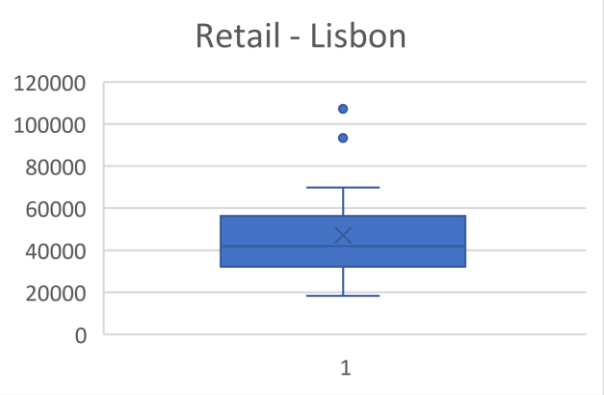
Boxplot 1 : Channel: HoReCa -Region: Lisbon



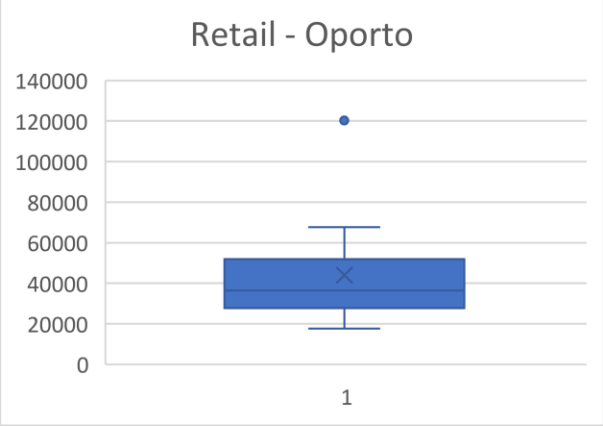
Boxplot 2 : Channel: HoReCa -Region: Oporto



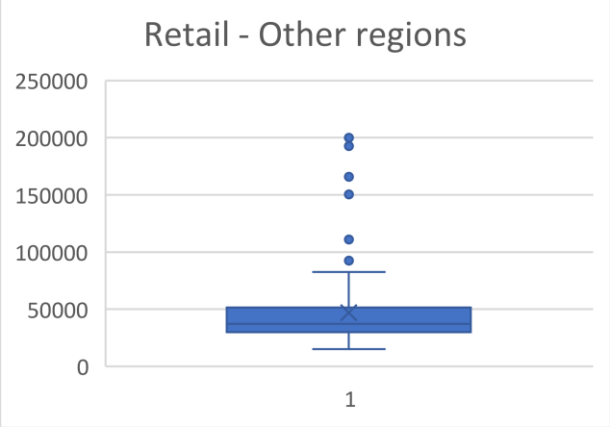
Boxplot 3 : Channel: HoReCa -Region: Other regions



Boxplot 4 : Channel: Retail -Region: Lisbon



Boxplot 5 : Channel: Retail -Region: Oporto



Boxplot 6: Channel: Retail -Region: Other regions

Q5-Perform a filtering of outliers of the Delicaessen column discussing the methods used to obtain them. To filter out the outliers in the Delicaessen, I have utilised the inter quartile range rule. To do this I first calculated the first(Q1) and third(Q3) quartiles.
Q1=408.25 Q3=1820.25

To calculate the inter quartile range (IQR), we need to calculate the difference between Q3 and Q1.

Q3-Q1 = IQR (1412)

Now we can use the IQR to calculate the higher and lower limits for our dataset, which we can use to filter out the outliers.

Lower limit = Q1-(IQR*1.5) = -1709.75

Higher limit = Q3 + (IQR*1.5) = 3938.25

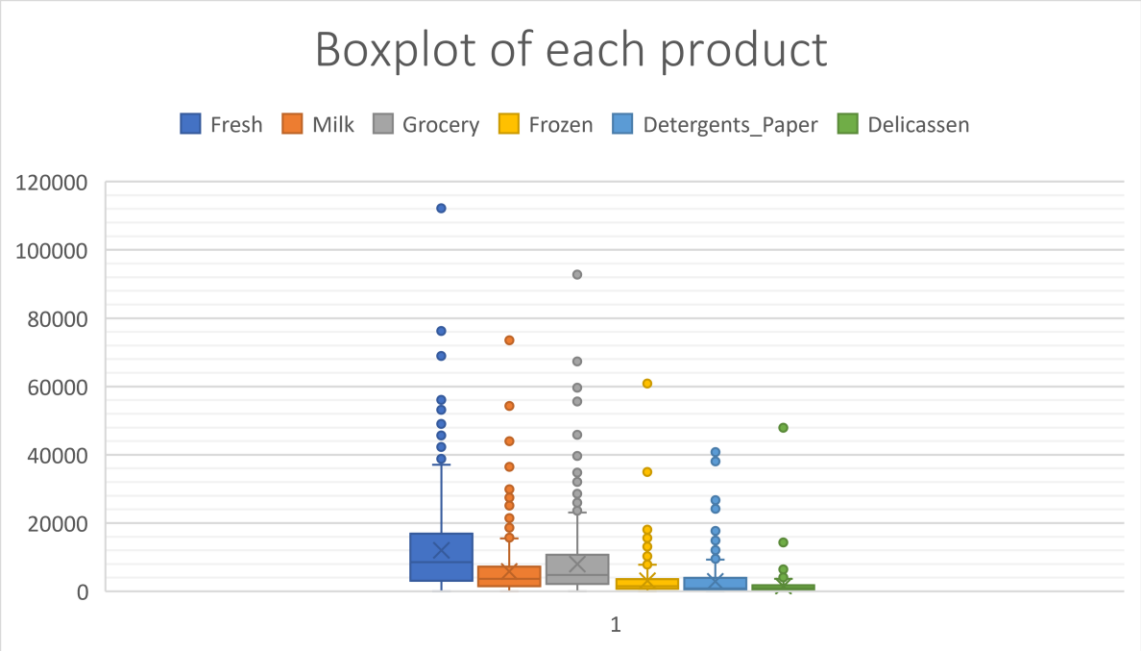
To filter the outliers, I will exclude any values smaller than Lower limit (-1709.75) and any value larger than the higher limit (3938.25). In this data set as we don't have any value the higher limit would eliminate 27 outliers. The new descriptive analysis of Delicassen values excluding the outliers can be seen below:

Delicassen	
Mean	1,092
Standard Error	43
Median	838
Mode	834
Standard Deviation	867
Sample Variance	751,614
Kurtosis	0
Skewness	1
Range	3,634
Minimum	3
Maximum	3,637
Sum	451,022
Count	413

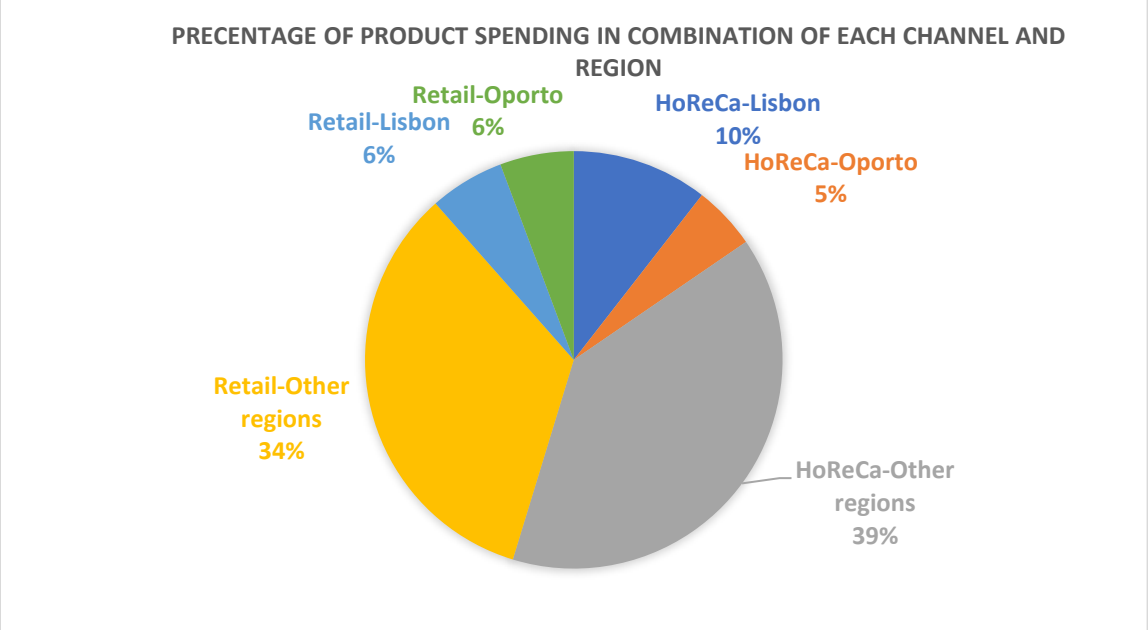
Q6 --Draw conclusions regarding the data set:the conclusions must be pertinent, and give a good synthesis of the data set to a non data-savy person.

Conclusion:

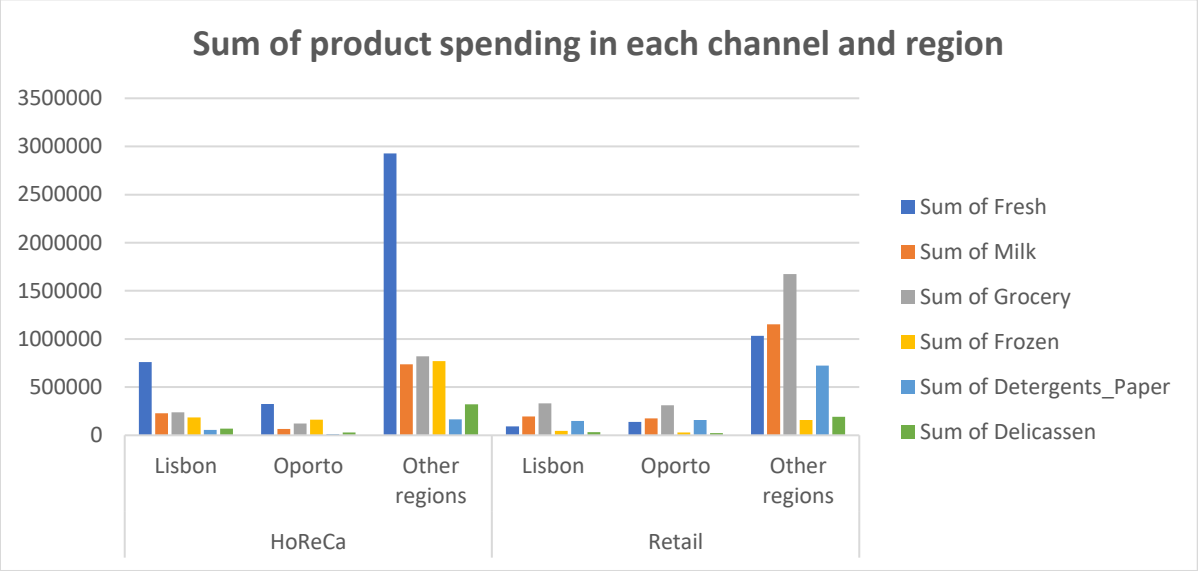
Looking at the comparison graph we can visualise that the values for Delicatessen are nearer to the median and it has a smaller IQR range and the outliers seem to be nearer to the median relative to other products. The data for fresh on the other hand indicates a larger IQR range and the outlier seem to be further away from the median.



In the pie chart below we have split the data into different channels and then split the channels into different regions to see the spending in each subgroup. It is clear “other regions” have significantly larger spending compared to Lisbon and Oporto with a combined sum of 73% of all spending.



To get a better understanding of the results attributes, I have created a clustered column chart. With this chart we can see, Fresh spending is scientifically higher in the other regions within HoReCa channel and followed by grocery spending in other regions within Retail channel.



In general, visualisation of data plays a crucial role in enabling users to understand the data better.