Statistics Project

Submitted by

Aruneema

Index

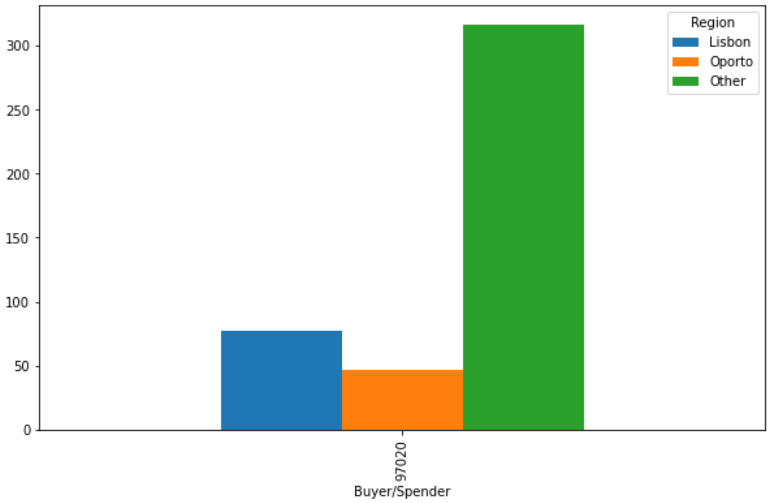
Problem Statement 1:

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 products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel, Retail).

1.1a Use methods of descriptive statistics to summarize data.

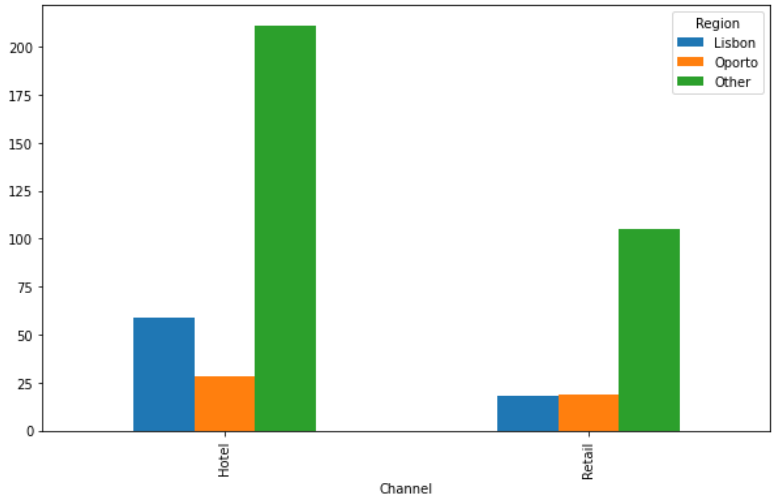
On loading the data, it was found that-

- The data consists of 6 continuous quantitative variables ('Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents_Paper', 'Delicatessen') and 3 categorical variable (Buyer/Spender, Region and Channels).
- There are in total 440 entries with no null value present in the data
- There are more 'Buyer/Spender' in 'Other', followed by Lisbon and Oporto



Graph 1

- 'Other' places spend more Hotels and Retail channels than Lisbon and Oporto. Between Lisbon and Oporto, Lisbon spends more on Hotels and Oporto pends more on Retail channel



Graph 2

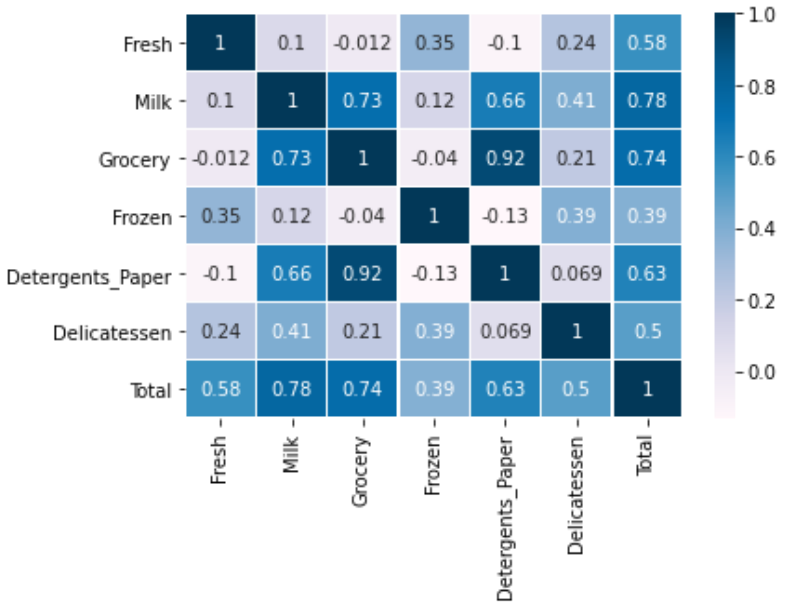
EDA

Exploratory Data analysis gives us further information about the given dataset. Table 1 gives us insight into the descriptive statistics of the 6 continuous quantitative variables. (Note: the categorical variables were removed from the table while generating descriptive statistics)

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
count	440.00	440.00	440.0	440.00	440.00	440.00	440.00
mean	12000.30	5796.30	7951.3	3071.90	2881.50	1524.90	33226.10
std	12647.30	7380.40	9503.2	4854.70	4767.90	2820.10	26356.30
min	3.00	55.00	3.0	25.00	3.00	3.00	904.00
25%	3127.80	1533.00	2153.0	742.20	256.80	408.20	17448.80
50%	8504.00	3627.00	4755.5	1526.00	816.50	965.50	27492.00
75%	16933.80	7190.20	10655.8	3554.20	3922.00	1820.20	41307.50
max	112151.00	73498.00	92780.0	60869.00	40827.00	47943.00	199891.00
CV	1.05	1.27	1.2	1.58	1.65	1.85	0.79
Range	112148.00	73443.00	92777.0	60844.00	40824.00	47940.00	198987.00

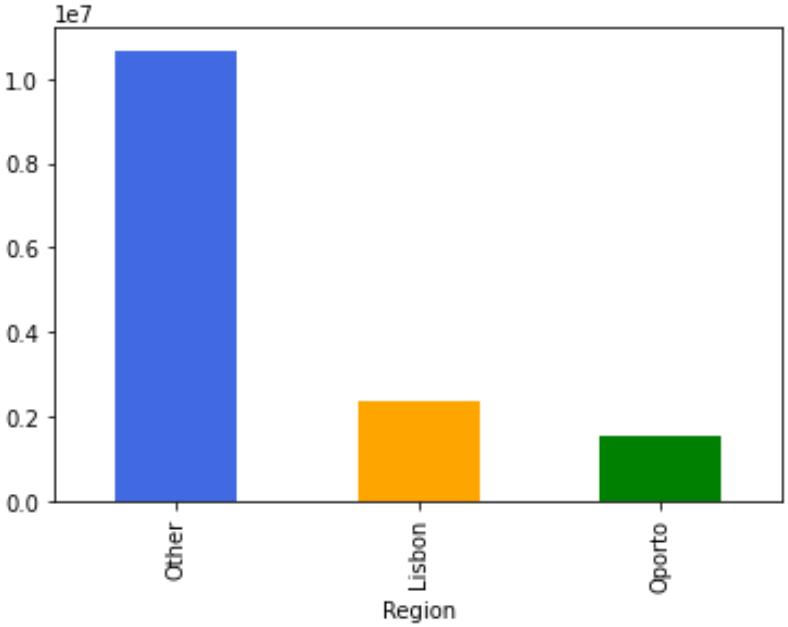
Table 1

The correlation table below also gives us good idea about relationship between different items.

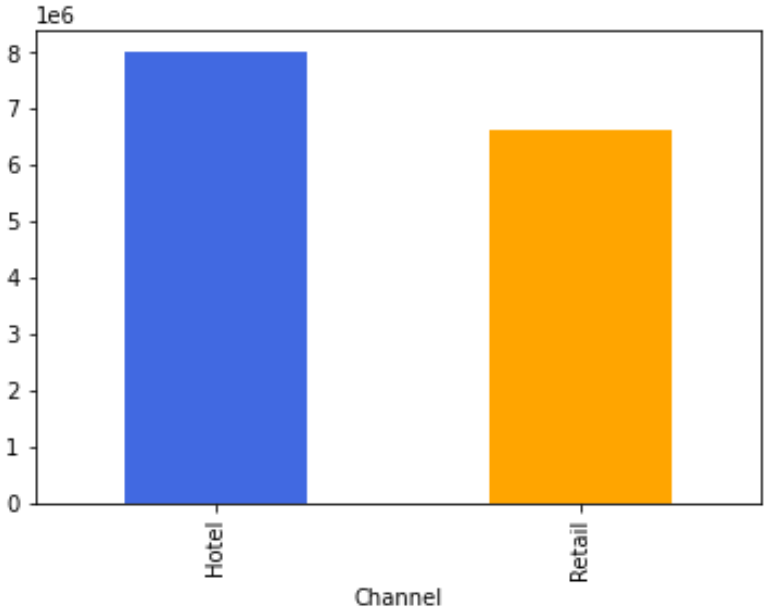


Milk and Grocery and Detergent Paper and Grocery have strong correlation, which means buyers spend more money on these two combination of products.

1.1b Which Region and which Channel spent the most? Which Region and which Channel spent the least?
Overall, 'Others' and 'Hotel' to spends more, while Oporto and Retail channel spends less on all the 6 items.



Graph 3



Graph 4

1.2 There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

The tables below show some basic descriptive statistics for 6 different items, region wise and channel wise. These tables help us to infer that, region wise-

	Region	Lisbon	Oporto	Other
Delicatessen	max	6854.0	5609.0	47943.0
	mean	1354.9	1159.7	1620.6
	min	7.0	51.0	3.0
	std	1345.4	1050.7	3232.6
Detergents_Paper	max	19410.0	38102.0	40827.0
	mean	2651.1	3687.5	2817.8
	min	5.0	15.0	3.0
	std	4208.5	6514.7	4593.1
Fresh	max	56083.0	32717.0	112151.0
	mean	11101.7	9887.7	12533.5
	min	18.0	3.0	3.0
	std	11557.4	8387.9	13389.2
Frozen	max	18711.0	60869.0	36534.0
	mean	3000.3	4045.4	2944.6
	min	61.0	131.0	25.0
	std	3092.1	9151.8	4260.1
Grocery	max	39694.0	67298.0	92780.0
	mean	7403.1	9218.6	7896.4
	min	489.0	1330.0	3.0
	std	8496.3	10842.7	9537.3
Milk	max	28326.0	25071.0	73498.0
	mean	5486.4	5088.2	5977.1
	min	258.0	333.0	55.0
	std	5704.9	5826.3	7935.5

Table 2: Region wise crosstab

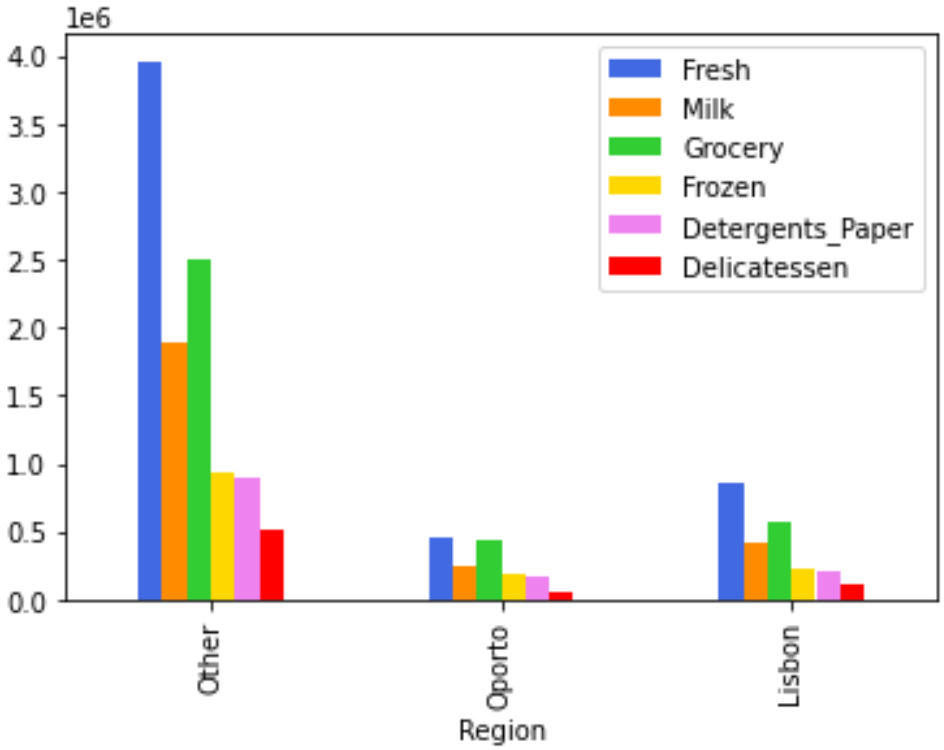
	Channel	Hotel	Retail
Delicatessen	max	47943.0	16523.0
	mean	1416.0	1753.4
	min	3.0	3.0
	std	3147.4	1953.8
Detergents_Paper	max	6907.0	40827.0
	mean	790.6	7269.5
	min	3.0	332.0
	std	1104.1	6291.1
Fresh	max	112151.0	44466.0
	mean	13475.6	8904.3
	min	3.0	18.0
	std	13831.7	8987.7
Frozen	max	60869.0	11559.0
	mean	3748.3	1652.6
	min	25.0	33.0
	std	5643.9	1812.8
Grocery	max	21042.0	92780.0
	mean	3962.1	16322.9
	min	3.0	2743.0
	std	3545.5	12267.3
Milk	max	43950.0	73498.0
	mean	3451.7	10716.5
	min	55.0	928.0
	std	4352.2	9679.6

Table 3: Channel wise cross Tab

- Delicatessen, Fresh and Milk products, average spend by buyers in 'Others' is highest followed by Lisbon and Oporto
- for, Detergents, Frozen and Grocery, Oporto's average spend is more than Lisbon and Others
- it's quite interesting to note that the min value of amount spent on diffrent food categories is drastically low as compared to maximum amount spent

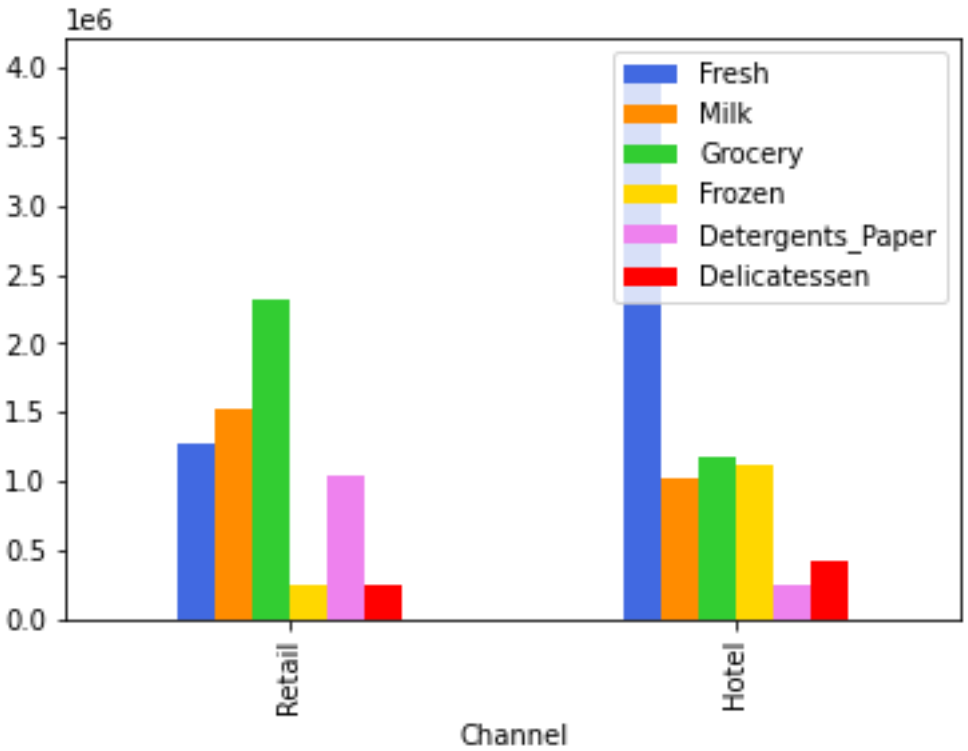
Channel wise-

- the average spend of buyers for Delicatessen, Detergents, Grocery and Milk products is higher in Retail channel than in Hotels
- for Fresh and Frozen products, Hotel channel average spend is more than Retail's
- it's quite interesting to note that the min value of amount spent on diffrent food categories is drastically low as compared to maximum amount spent



Graph 4

From graph 4, it can also be inferred that buyers in 'Others' and Lisbon spend more on fresh items and least on 'Delicatessen'



Graph 5

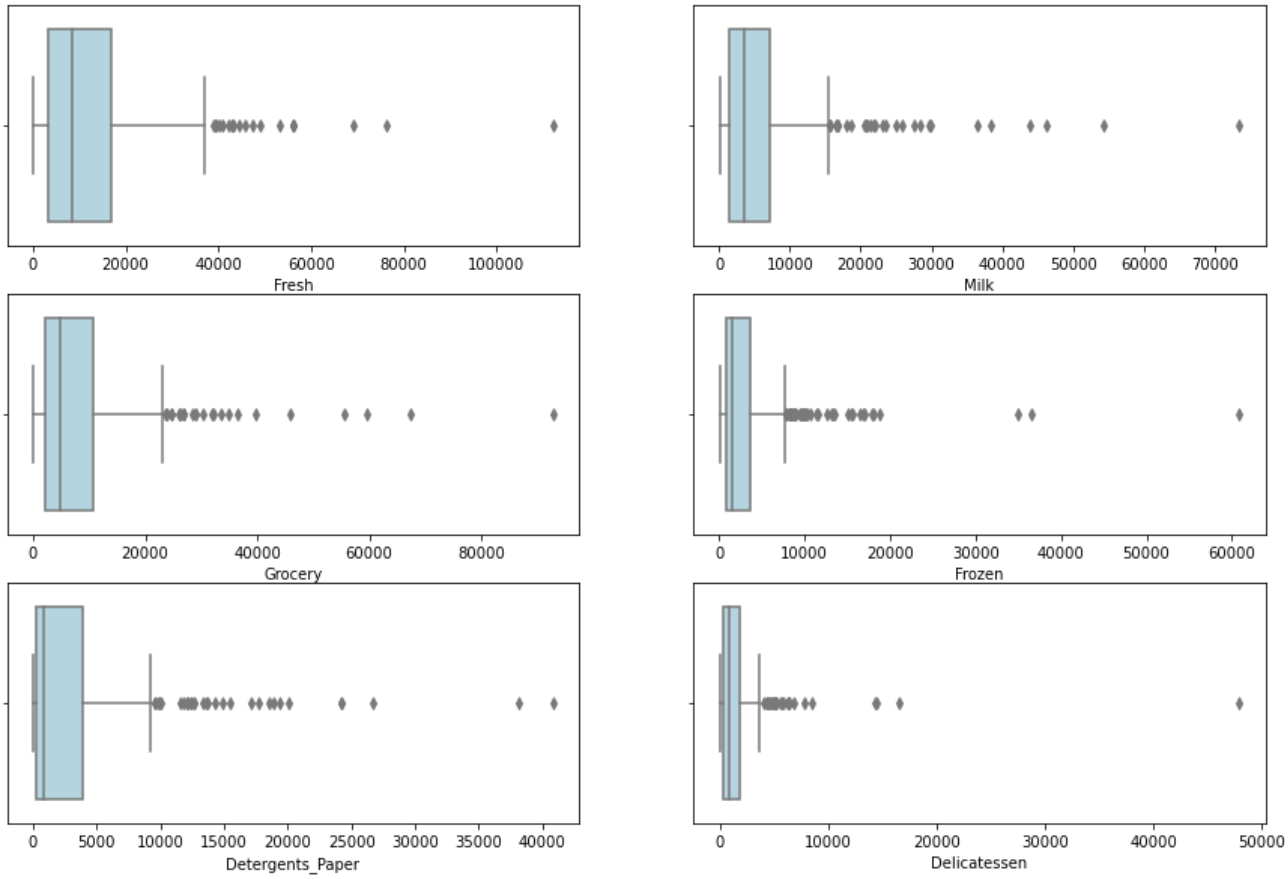
Channel wise, buyers spend on retail channel spend more on Grocery, while in Hotels, they spend more on fresh products.

1.3 On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour?

As calculated from coefficient of variance in table.. , 'Delicatessen' shows greatest dispersion (CV= 1.85), proving to be most inconsistent, while 'Fresh' item shows relatively less dispersion (CV= 1.05), proving to be more consistent.

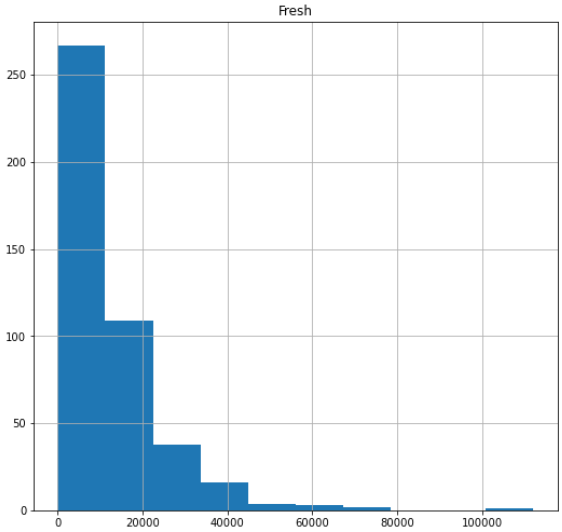
1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.

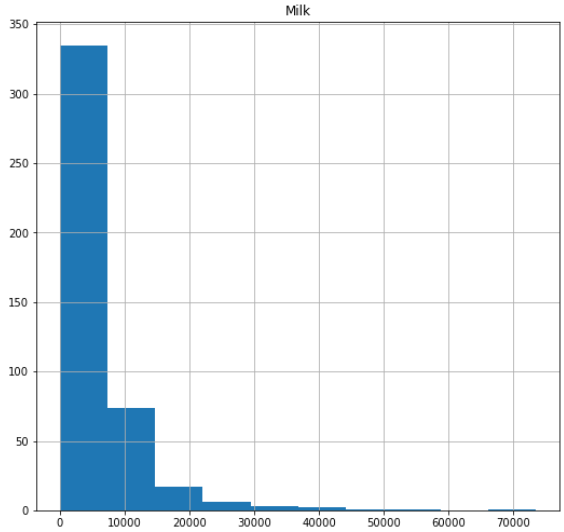
For all the 6 continuous variables on items, there are outliers, as depicted in the boxplot of each item-

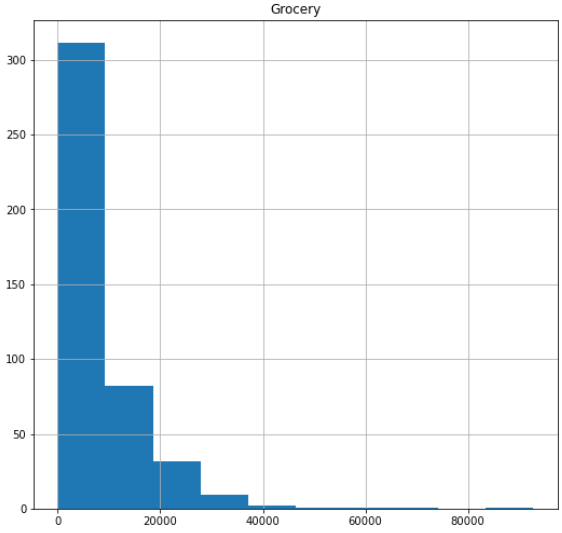


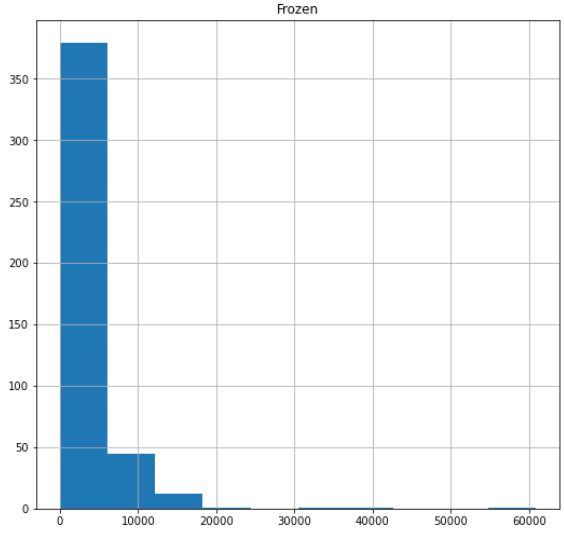
Graph 6

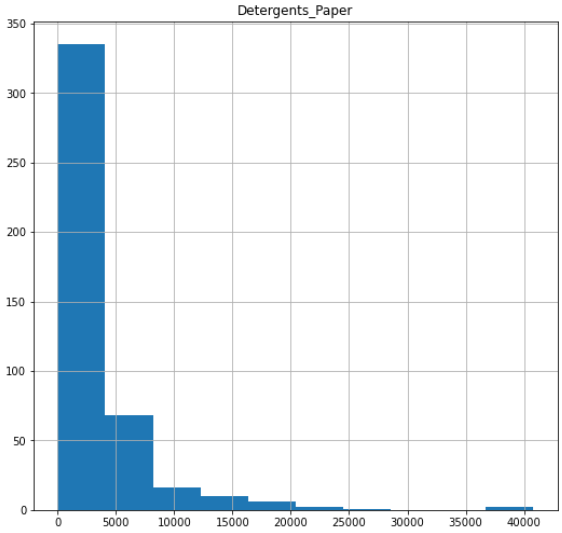
The histogram in graph 7 also proves the presence of outliers through the presence of skewness (right skewness)

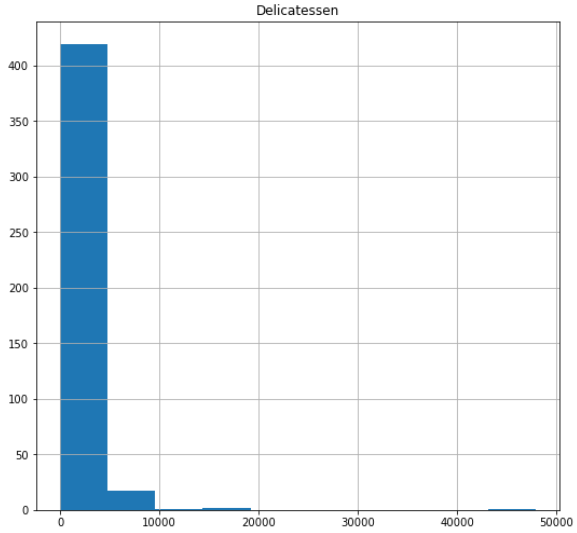

Fresh
0
50
100
150
200
250
0
20000
40000
60000
80000
100000


Milk
0
50
100
150
200
250
300
350
0
10000
20000
30000
40000
50000
60000
70000


Grocery
0
50
100
150
200
250
300
0
20000
40000
60000
80000


Frozen
0
50
100
150
200
250
300
350
0
10000
20000
30000
40000
50000
60000


Detergents_Paper
0
50
100
150
200
250
300
350
0
5000
10000
15000
20000
25000
30000
35000
40000


Delicatessen
0
50
100
150
200
250
300
350
400
0
10000
20000
30000
40000
50000

Graph 7

1.5 On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem?

Since buyers spend more in 'other' region, the wholesale company should plan to expand in 'other' region to increase sales.

The inconsistencies on the buying amount is quite high. This needs to be investigated and explored to understand why there are so many high value inconsistencies. It may be that there are specific customer profile purchasing items in high value.

Since there's a strong correlation between Milk and Grocery and Detergent paper and grocery, wholesalers should invent on schemes/promotions to further improves its sales.

There’s a need to focus on sales of Frozen, Detergents and Delicatessen.

It seems that buyers prefer to buy perishable and necessary items (Fresh, Milk, Grocery) from retail channels and non-perishable and less essential items (Frozen, detergents and Delicatessen) from hotels. This insight indicates that retail channels can be specialised for every day, perishable products while hotels can be customised for unique non-essential products.