**SMDM Extended Project**

**Part 1**

**Statement:**

A wholesale distributor operating in different regions of Portugal has information on the 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 channels (Hotel, Retail).

**1.1** **Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?**

**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.**

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

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

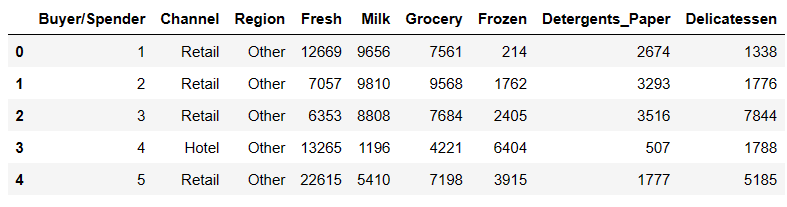
**1.5 On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective**

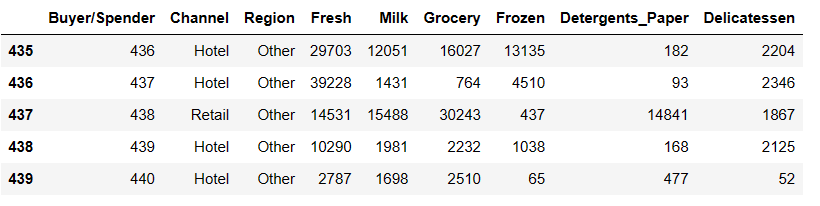
**1.1** **Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?**

**Ans:-** We have loaded the all the required packages and loaded Wholesale Customer Ads Data file using Pandas.

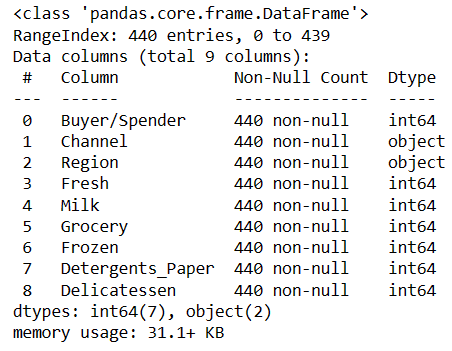
Dataset has 440 rows and 9 columns.

We have viewed first and last few rows using head() and tail() functions respectively.





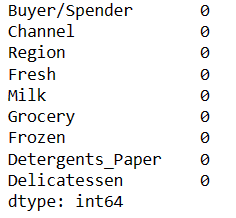
We can view dataset information using info()



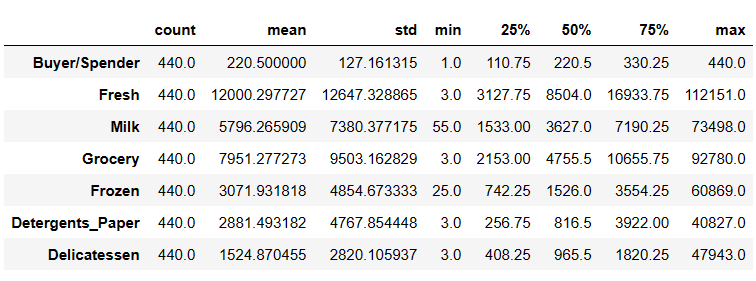
There are 7 columns with interger datatype, 2 with object datatype.

There are no duplicate values in this dataset.

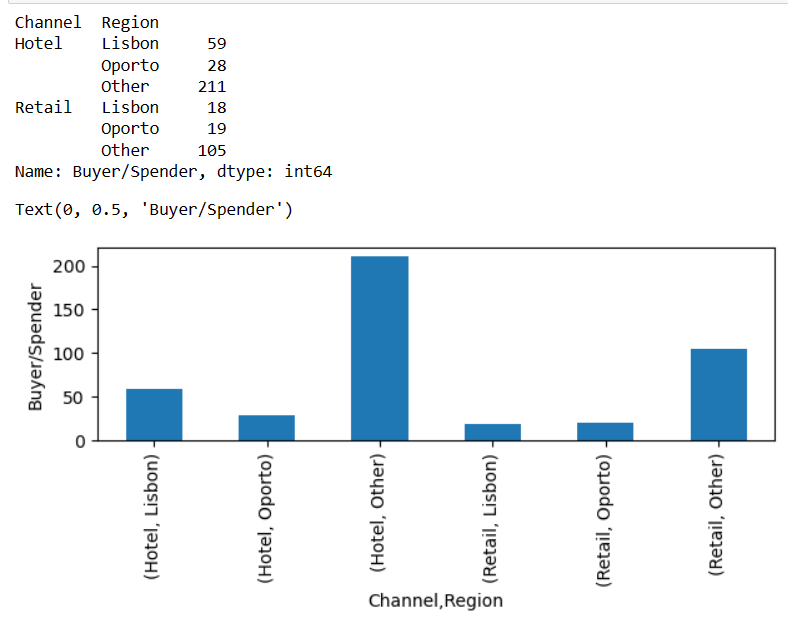
There are no null values.

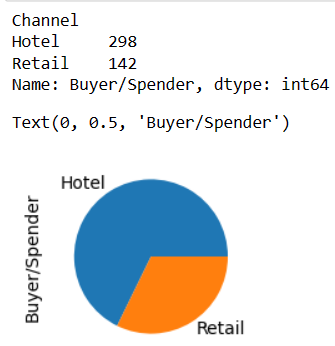


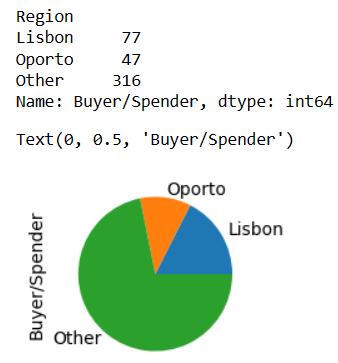
Dataset can be described using describe() function.

Observation:-

1. There are total 440 records for each dataset’s field.
2. Buyer/Spender has minimum of 1 and maximum of 440 buyer/spender.
3. Fresh items,Grocery,Detergents\_Paper and Delicatessen have 3 as a minimum spendings.
4. Spendings on Milk ranges from 7k to 73k.







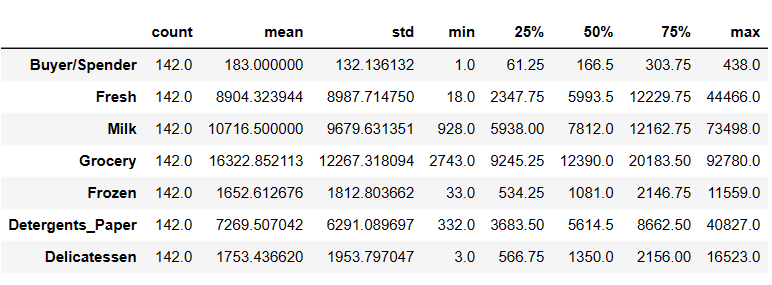
Channel = Hotel and Region = Other spent the most as both Hotel and Other Region including has 177 buyer/spender.

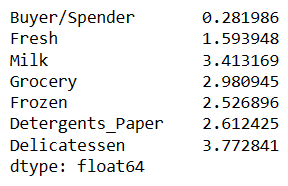
Channel = Retail and Region = Lisbon spent the least as they have only 18 buyer/spender.

**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.**

**Ans:-**

1. For Channel = Retail





A) Spending on Grocery in Retail Channel is the highest and spending on Frozen item is minimum.

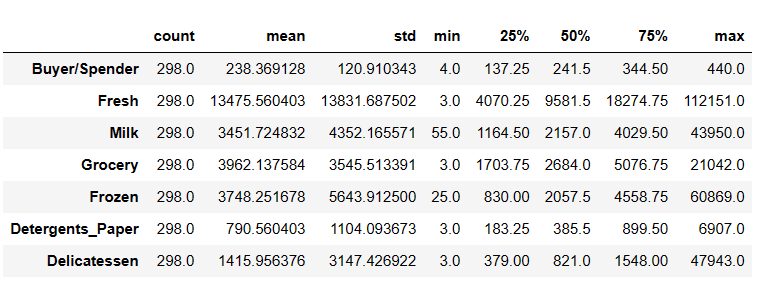
B) Spending on Fresh Items in Retail Channel ranges between 18 to 44k.

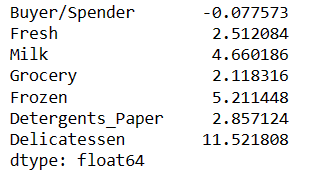
C) On an average 7k is spent on Detergent Paper in Retail Channel.

D) Delicatessan ranges from 3 to 16k and has skewness of +3.77

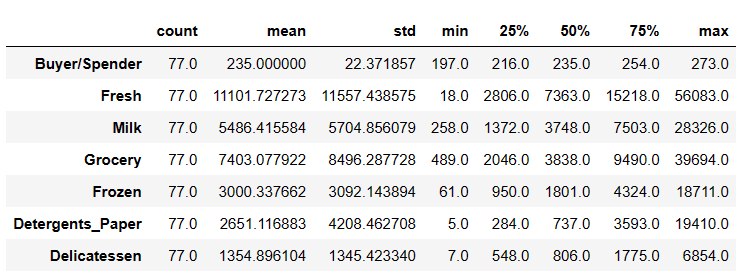
E) Mean and Median value for every field are not very close so we can say that can be skewness in the data.

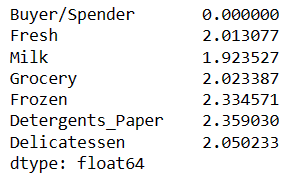
1. For Channel = Hotel

****

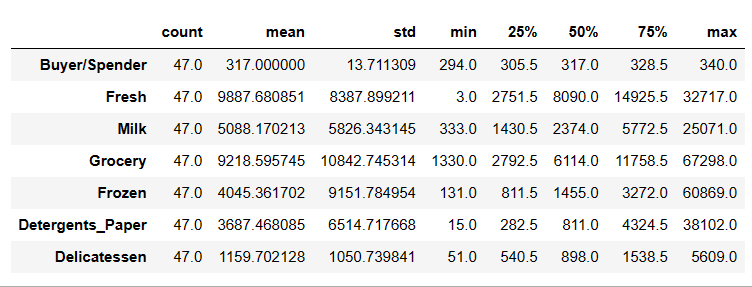
****

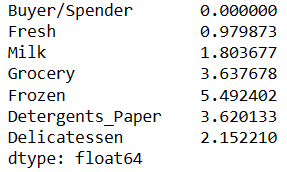
1. Spending on Fresh items in Hotel Channel is the highest and lowest for Detergents\_Paper. So Customers in Hotel Channel tend to buy more Fresh items than Customers in Retail Channel
2. None of the Varieties of items have same mean and median so there is skewness on every dataset.
3. Spending on Milk ranges between 55 to 43k. Milk items has skewness of + 4.66 in Hotel Channel.
4. Grocery has mean of 3962.
5. Maximum spendings on Delicatessan in Hotel Channel ranges between 821 to 47943.
6. For Region = Lisbon



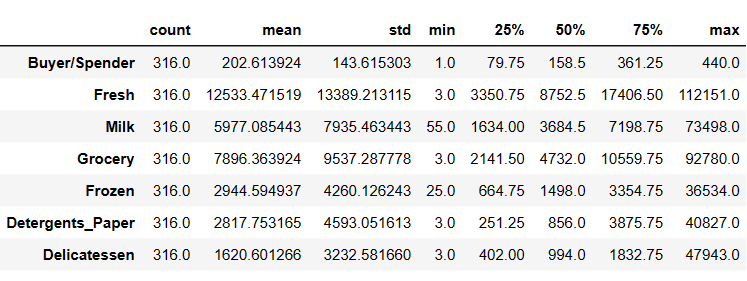


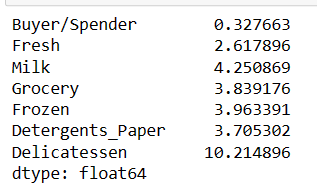
1. Spending on Fresh items in Lisbon Region is the highest and lowest for Delicatessen. So Customers in Lisbon Region tend to buy more Fresh items than any other items..
2. None of the Varieties of items have same mean and median so there is skewness on every dataset.
3. Spending on Milk ranges between 258 to 28k. Milk items has skewness of + 1.92 in Lisbon Region.
4. Grocery has mean of 7403.
5. Maximum spendings on Delicatessan in Lisbon Region ranges between 7 and 6854
6. For Region = Oporto



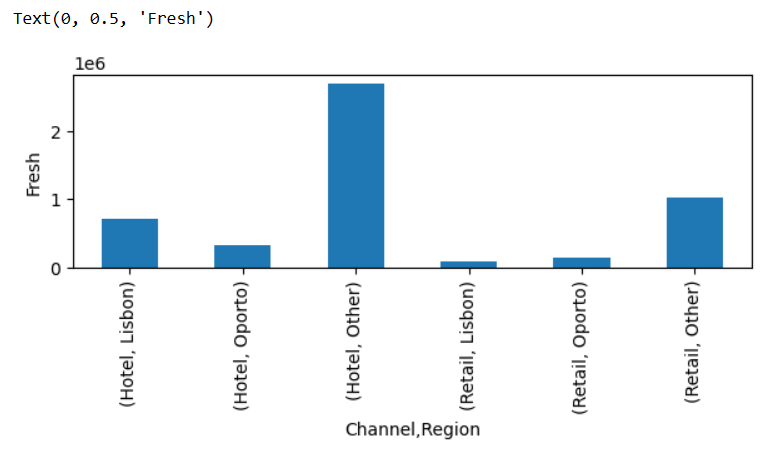


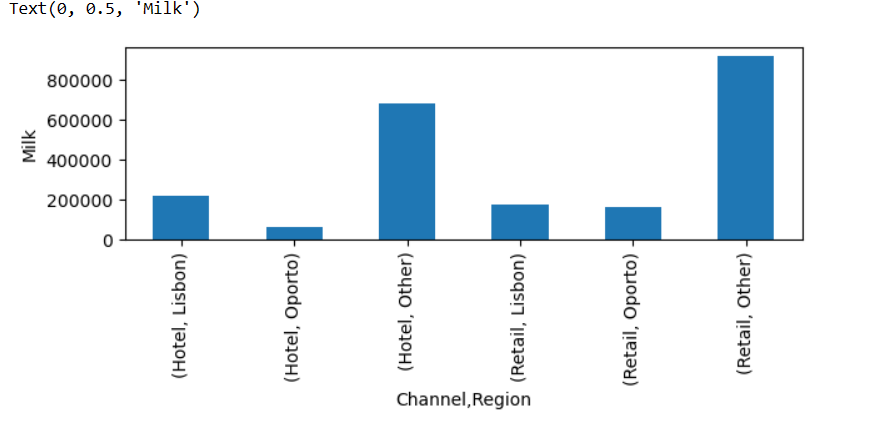
1. In Oporto Region, Customers have spent highest on Grocery and lowest on Delicatessen items.
2. Spending on Fresh items ranges between 3 and 32717 and has skewness of +0.979873.
3. Milk spending on an average is around 5k.
4. There is no comparision between mean and median so we can say that there is skewness on every plot of each items.
5. For Other Region

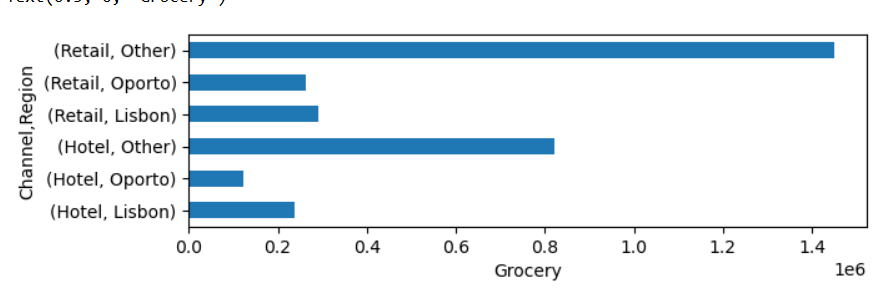


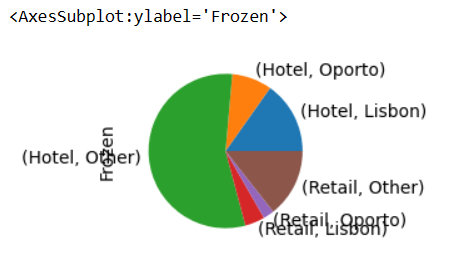


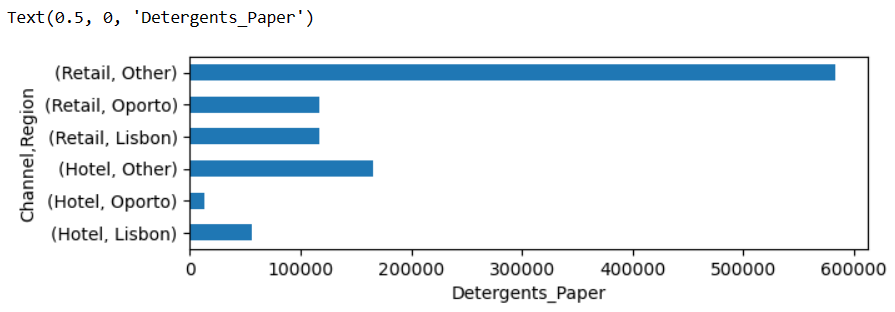
1. Other Regions have spent highest on Fresh Items with skewness of +2.6 and lowest on Frozen Items with skewness of +3.963. It means that Other Regions prefer Fresh goods than Frozen.
2. Milk spending ranges between 55 to 74k. It has average spending of around 60k.
3. Mean and median have no similarity to there could be skewness in the plots.
4. Delicatessen shows high skewness.

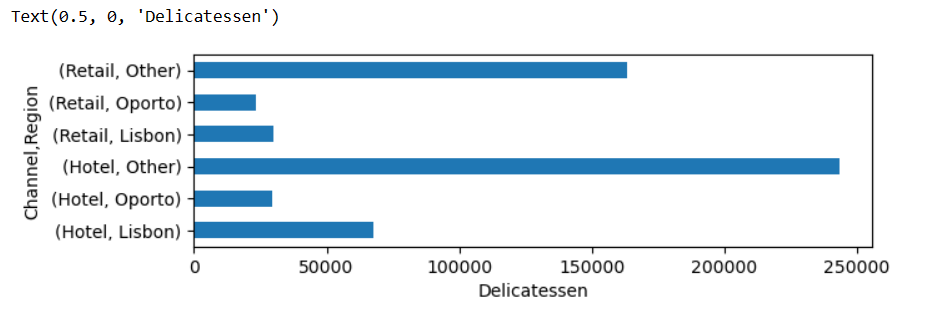




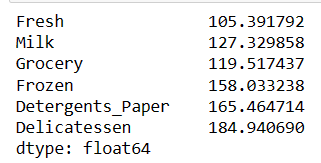








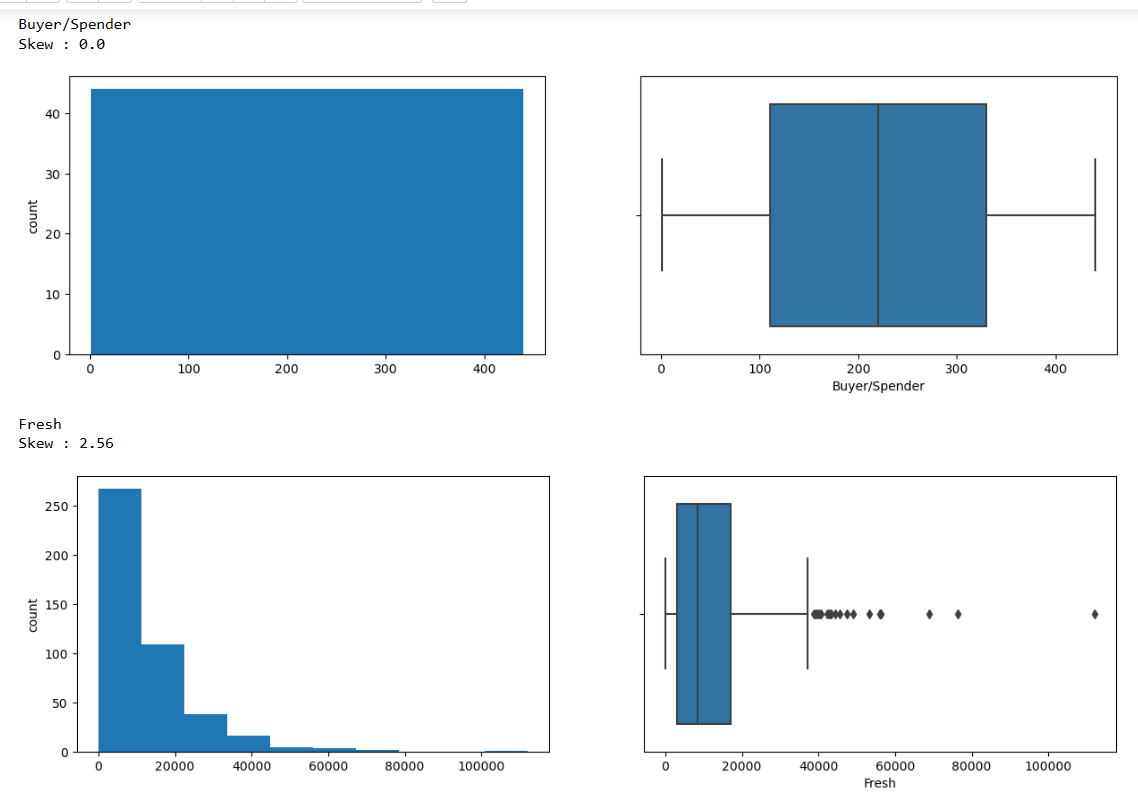
**1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?**

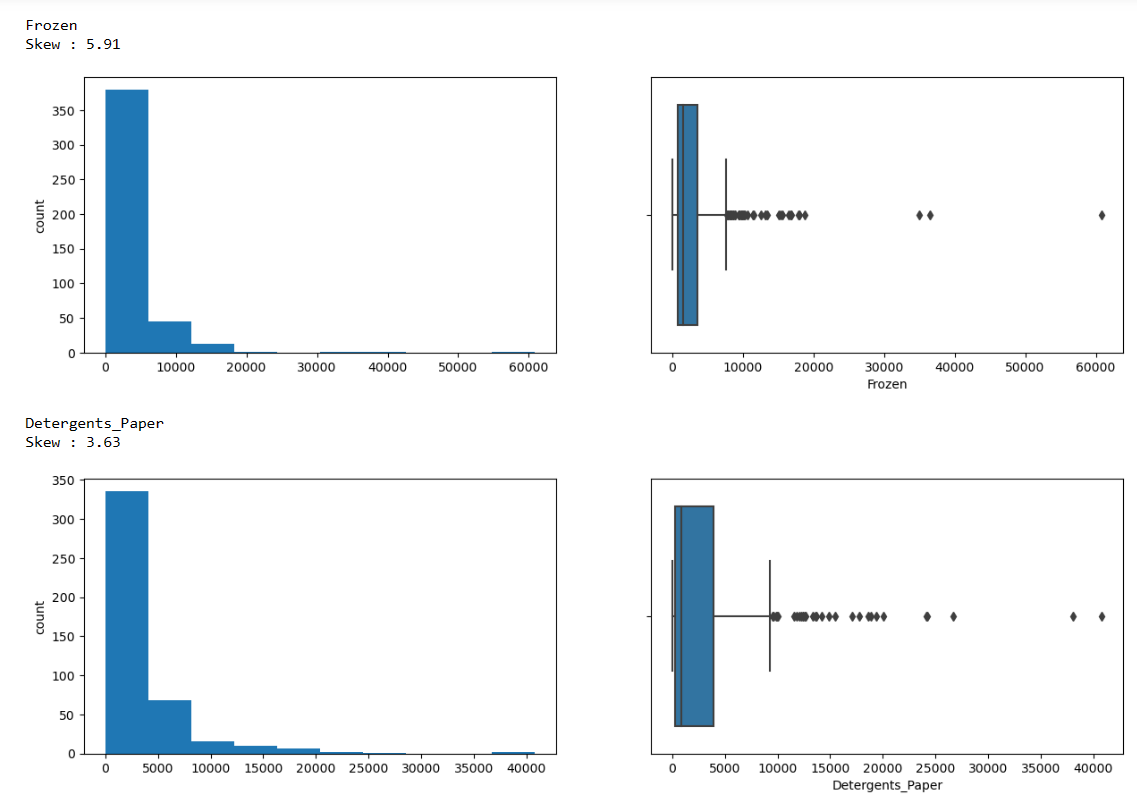
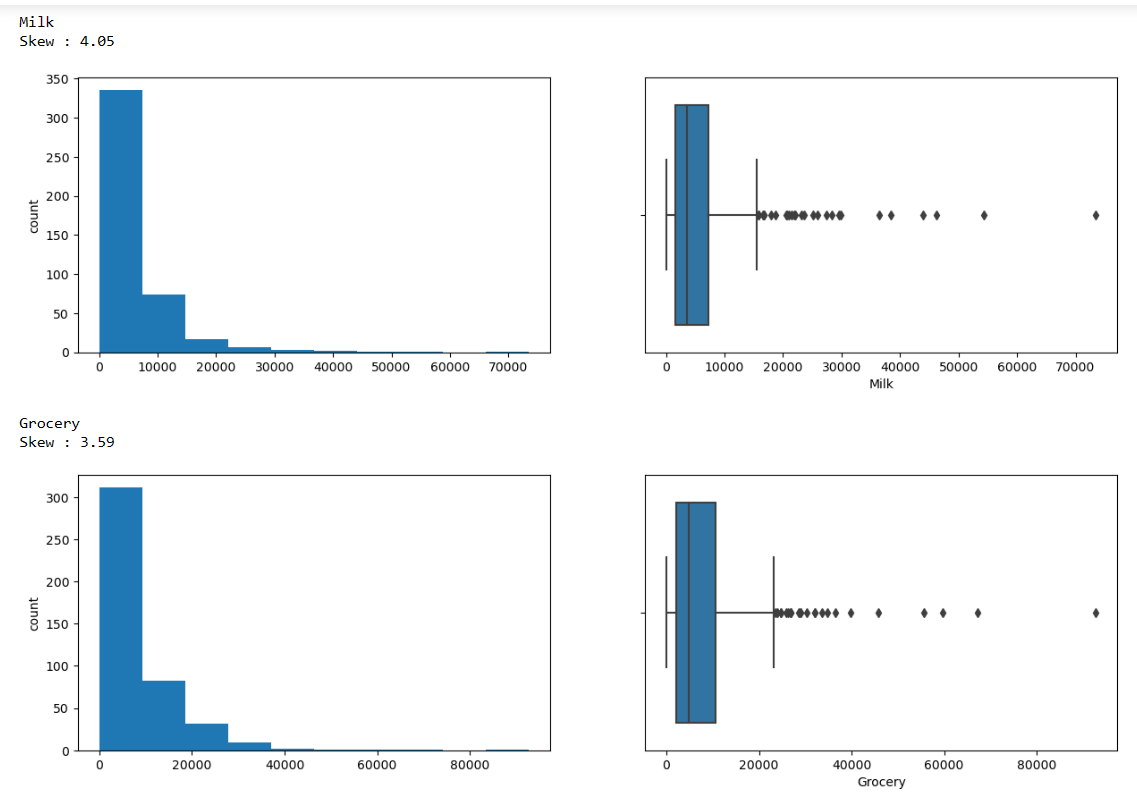
Ans:- 

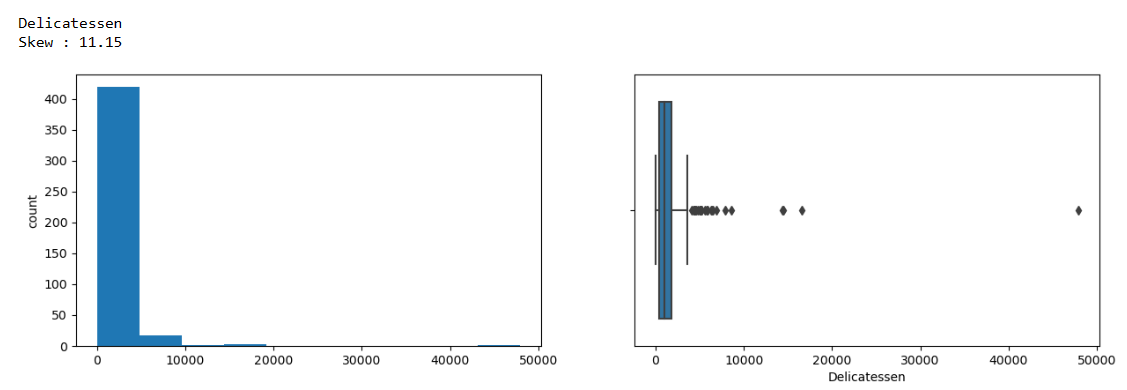
Fresh items have lowest standard deviation so are inconsistent and

Delicatessen items have highest standard deviation so shows least inconsistent behaviour and are consistent.

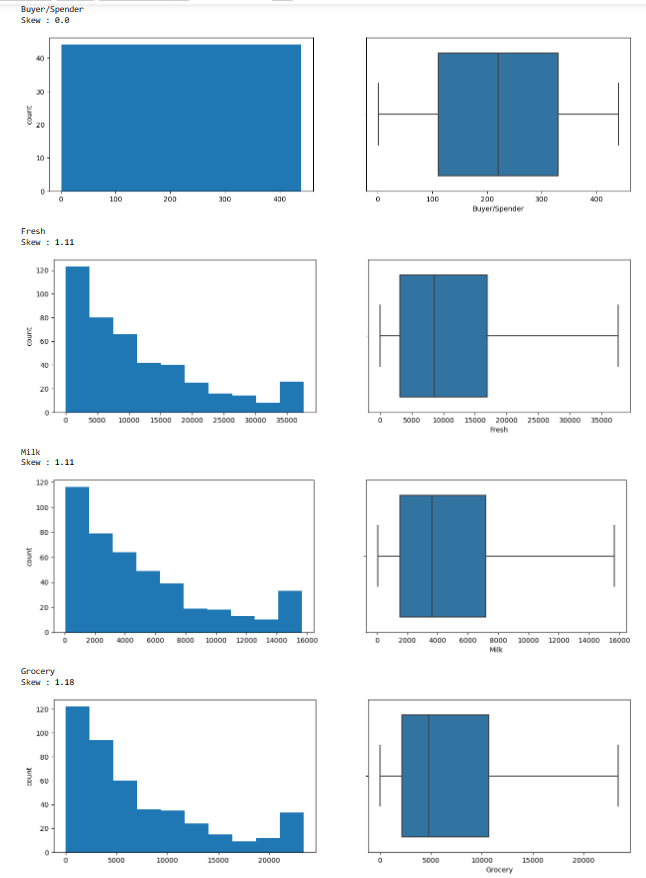
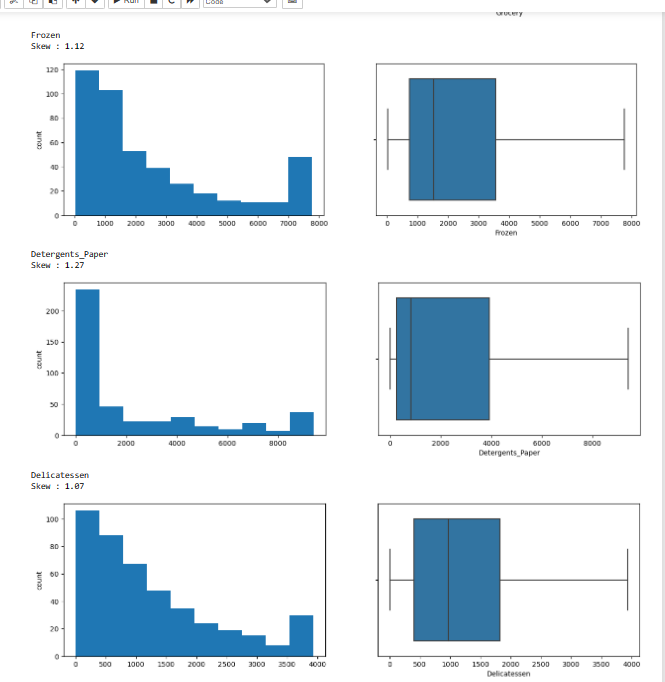
**1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.**

Ans:- 



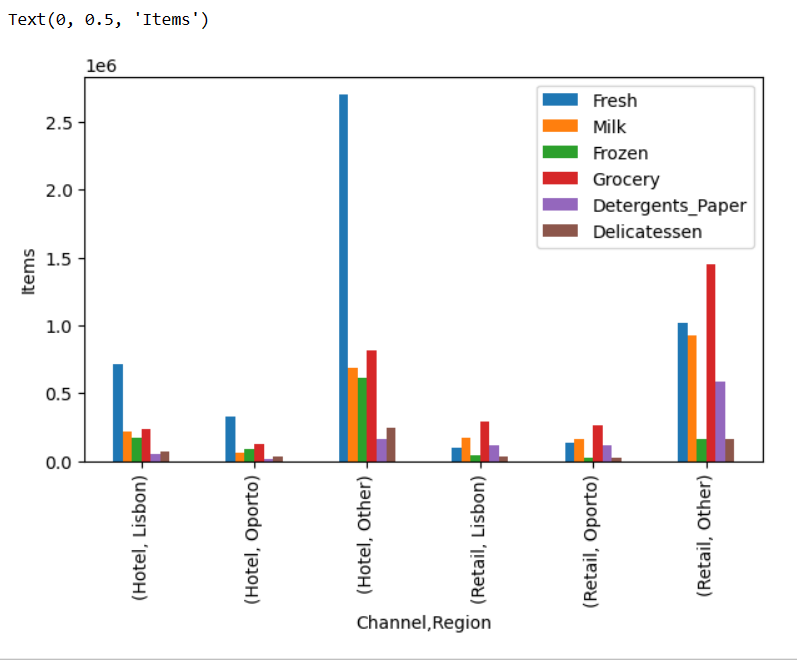


Observation:-

1. Yes there are Outliers in every variety of items in the data.
2. As Observed ,most of the variables have skewed distribution.
3. All the variables have highly right skewed distribution.
4. Delicatessen has highly right skewed graph.
5. We have treated the Outliers using (1.5 \* IQR) method.
6. After treatment there are no outliers in the plot.
7. 
8. 

**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? Answer from the business perspective**

Ans:-

1. There are 440 Rows and 9 columns.
2. There are no null or Duplicate Records.
3. 
4. From the graph we see that Customers have highly spent on Fresh Items and least on Delicatessan.
5. Hotel channel has spent the most than Retail Channel.
6. Other Regions have spent the most than the other two. Customer have spent least in Oporto Region.