Problem 1

Wholesale Customers Analysis

Problem Statement:

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).

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

The shape of the data is (440,10) meaning there are 440 rows and 10 columns.

There are no null values in this dataframe.

The summary of the data is given in the table below.

Summary of the data:

	count	mean	std	min	25%	50%	75%	max
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0
Delicatessen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0

There are 6 different varieties of items in this dataframe and all of them have 440 values. By looking at the above table, we are able to deduce that 'Fresh' has the highest mean value while 'Delicatessen' has the lowest mean value. 'Fresh' has the highest standard deviation value while 'Delicatessen' has the lowest standard deviation value

Adding a column 'Total_Spending':

'Total_Spending' is the total amount spent in 'Fresh', 'Milk', "Grocery', 'Frozen', 'Detergents_Paper' and 'Delicatessen'. By this way, the total amount spent can be calculated easily.

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_Spending
0	1	Retail	Other	12669	9656	7561	214	2674	1338	34112
1	2	Retail	Other	7057	9810	9568	1762	3293	1776	33266
2	3	Retail	Other	6353	8808	7684	2405	3516	7844	36610
3	4	Hotel	Other	13265	1196	4221	6404	507	1788	27381
4	5	Retail	Other	22615	5410	7198	3915	1777	5185	46100
...	...	...	...	...	...	...	...	...	...	...
435	436	Hotel	Other	29703	12051	16027	13135	182	2204	73302
436	437	Hotel	Other	39228	1431	764	4510	93	2346	48372
437	438	Retail	Other	14531	15488	30243	437	14841	1867	77407
438	439	Hotel	Other	10290	1981	2232	1038	168	2125	17834
439	440	Hotel	Other	2787	1698	2510	65	477	52	7589

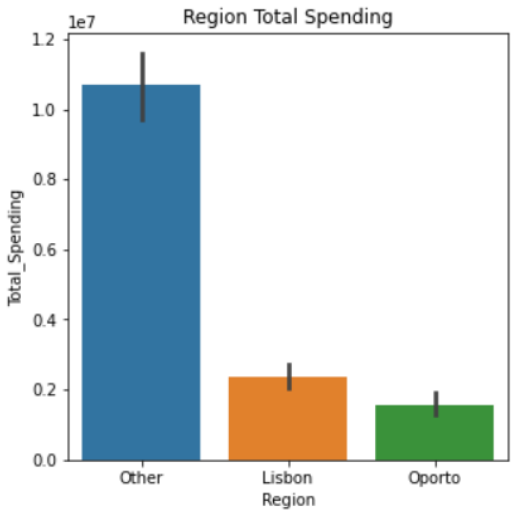
440 rows × 10 columns

From the column ‘Total_Spending that was created above, we can now calculate the highest and lowest amount spent in ‘Region’ and ‘Channel’.

Calculating highest and lowest amount spent in Region:

```
Region
Lisbon     2386813
Oporto     1555088
Other     10677599
Name: Spending, dtype: int64
```

From the above table, it can be seen that the **highest amount spent in the Region is from ‘Other’** with a value of 10677599 and **lowest amount spent in the Region is from ‘Oporto’** with a value of 1555088.

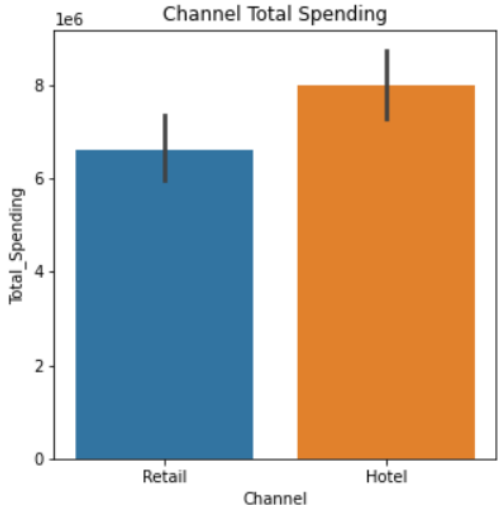


We can also see from the above graph that **highest amount spent in the Region is from ‘Other’** and **lowest amount spent in the Region is from ‘Oporto’**.

Calculating highest and lowest amount spent in Channel:

```
Channel
Hotel     7999569
Retail    6619931
Name: Spending, dtype: int64
```

From the above table, it can be seen that the **highest amount spent in the Channel is from 'Hotel'** with a value of 7999569 and **lowest amount spent in the Channel is from 'Retail'** with a value of 6619931.



We can also see from the above graph that **highest amount spent in the Channel is from 'Hotel'** and **lowest amount spent in the Channel is from 'Retail'**.

Calculating highest and lowest amount spent within 'Region':

Region	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_Spending
Lisbon	235.00	11101.73	5486.42	7403.08	3000.34	2651.12	1354.9	30997.57
Oporto	317.00	9887.68	5088.17	9218.60	4045.36	3687.47	1159.7	33086.98
Other	202.61	12533.47	5977.09	7896.36	2944.59	2817.75	1620.6	33789.87

In Region 'Lisbon', the highest amount spent (11101.73) is on 'Fresh' and lowest amount spent (1354.9) is on 'Delicatessen' items.

In Region 'Oporto', the highest amount spent (9887.68) is on 'Fresh' and lowest amount spent (1159.7) is on 'Delicatessen' items.

In Region ‘Other’, the highest amount spent (12533.47) is on ‘Fresh’ and lowest amount spent (1620.6) is on ‘Delicatessen’ items.

Calculating highest and lowest amount spent within ‘Channel’:

Channel	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_Spending
Hotel	238.37	13475.56	3451.72	3962.14	3748.25	790.56	1415.96	26844.19
Retail	183.00	8904.32	10716.50	16322.85	1652.61	7269.51	1753.44	46619.23

In Channel, highest amount ‘Hotel’ has spent on is ‘Fresh items’ with a value of 13475.56 and lowest amount ‘Hotel’ has spent on is ‘Detergents_Paper’ with a value of 790.56.

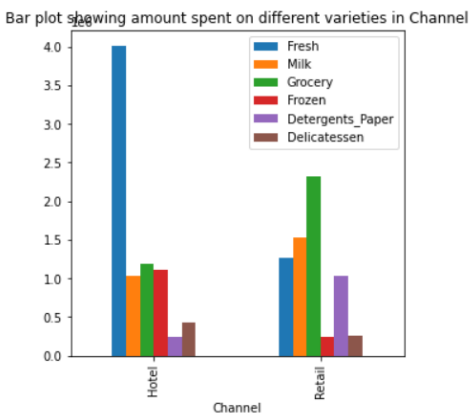
In Channel, highest amount ‘Retail’ has spent on is ‘Grocery’ items with a value of 16322.85 and lowest amount ‘Retail’ has spent on is ‘Frozen’ items with a value of 1652.61.

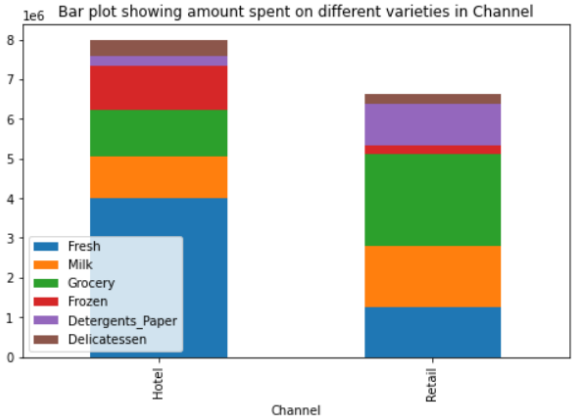
1.2 There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

Channel:

The amount spent in 6 different varieties of items in Channel is given below and the same is plotted using a bar graph.

	Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	Hotel	4015717	1028614	1180717	1116979	235587	421955
1	Retail	1264414	1521743	2317845	234671	1032270	248988





From the graph of ‘Channel’ it is clear that, ‘Hotel’ spends more on ‘Fresh’ items and least on ‘Detergents_Paper’ while ‘Retail’ spends more on ‘Grocery’ and the least on ‘Frozen’ items. This might be the case because Retail market focuses more on fresh goods and grocery more than frozen foods whereas ‘Hotel’ would require a lot of ingredients since it will have restaurants serving a lot of people. Hence they would have to stock up on both all food items – fresh, milk, grocery and frozen.

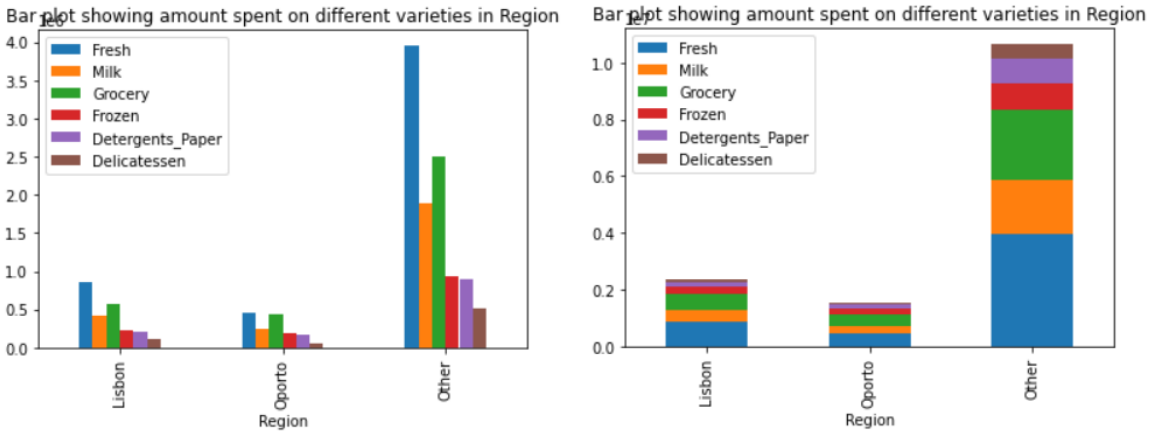
We also see that some categories like Milk, Grocery and Detergents_Paper have higher spend in the Retail channel versus Hotel.

We can also deduce that ‘Hotel’ spends more than ‘Retail’ from the graph on the right.

Region:

The amount spent in 6 different varieties of items in Region is given below and the same is plotted using a bar graph.

	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	Lisbon	854833	422454	570037	231026	204136	104327
1	Oporto	464721	239144	433274	190132	173311	54506
2	Other	3960577	1888759	2495251	930492	890410	512110



From the graph of ‘Region’ it is clear that, ‘Other’ cities spend more on ‘Fresh’ items and the least on ‘Delicatessen’ items. The same is the case in Lisbon and Oporto. Lisbon and Oporto spend more on ‘Fresh’ items and the least on ‘Delicatessen’ items. Therefore we can conclude that, ‘Fresh’ is the one most spent on followed by ‘Grocery’. This might be the case because the people are interested in cooking and purchasing fresh produce and grocery.

We can also deduce that ‘Other’ spends more than ‘Lisbon’ and ‘Oporto’ from the graph on the right.

The variety of each channel and each region is calculated by creating 5 dataframes and summary of the 5 dataframes.

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?

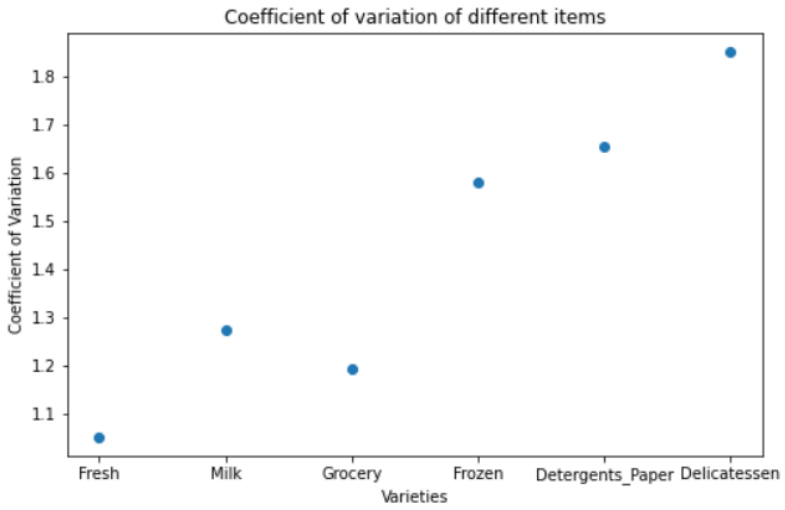
Based on Coefficient of Variation (CV):

Coefficient of variation is the population standard deviation divided by the population mean. The higher the coefficient of variation, the greater the level of dispersion around the mean. The lower the value of variation, the more precise the estimate.

The coefficient of variation of all the 6 items are listed below.

```
Fresh              1.053918
Milk               1.273299
Grocery            1.195174
Frozen             1.580332
Detergents_Paper   1.654647
Delicatessen       1.849407
dtype: float64
```

These values are further plotted to visually understand the item with the smallest and highest CV value.



'Fresh' item has the smallest CV value (1.053918). Therefore that is the **least inconsistent**.

'Delicatessen' item has the highest CV value (1.849407). Therefore that is **most inconsistent**.

Based on Standard Deviation:

Standard deviation is one of the measures of variability. Standard deviation of each of the 6 items can be found under 'std' column of the table below.

Low standard deviation means that the data are clustered around the mean and high standard deviation means the data is more spread out.

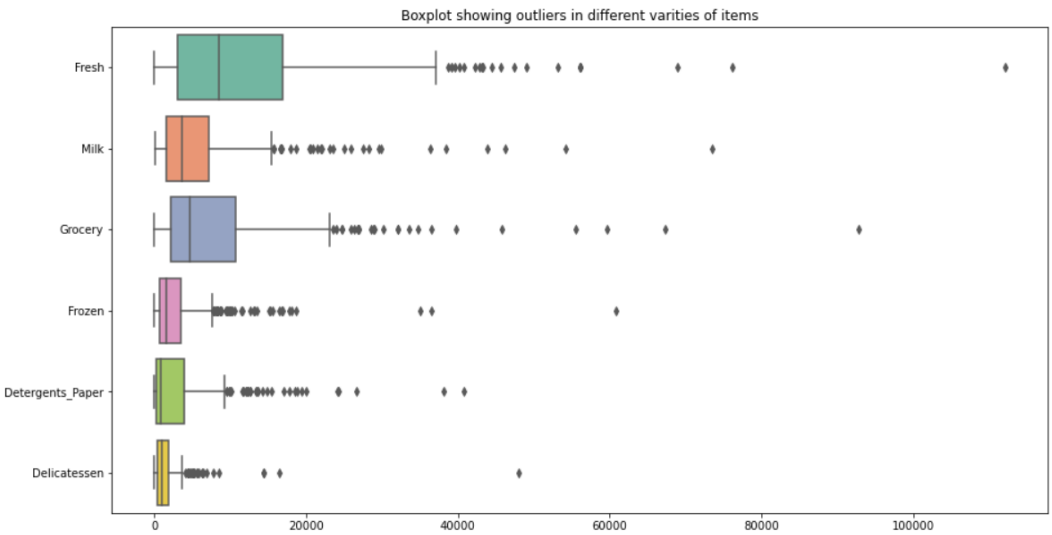
	count	mean	std	min	25%	50%	75%	max
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0
Delicatessen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0

'Fresh' item has the highest standard deviation value (12647.328865). Therefore it means that the data is more spread out. Therefore it is the **least inconsistent**.

'Delicatessen' item has the smallest standard deviation value (2820.105937). Therefore it means that the data is clustered around the mean. Therefore it is the **most inconsistent**.

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

Boxplot is used to plot the data to see if there are outliers in the data. The black points in the graph below represent the outliers.



A value is considered an outlier if it is more than 1.5 times the interquartile range below Q1 or above Q3.

Outliers is present in all the 6 different varieties of items – Fresh, Milk, Grocery, Frozen, Detergents_Paper and Delicatessen.

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

On the basis of the analysis done, it is found that a lot of revenue is got from 'Fresh'. Therefore we can stock up more on fresh items to generate more revenue. Detergents_Paper and Delicatessen gives the least revenue. Therefore that has to be stocked less. However the business can be improved in these items by giving offers, discounts or by advertising the uses, advantages of these items. The business should be improved such that the amount spent in all regions should be almost equal. By improving business across all regions, we can generate more revenue. There are inconsistencies in spending of different items (by calculating Coefficient of Variation), which has to be minimized.