WHOLESALE CUSTOMER 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 seems to spend more? Which Region and which Channel seems to spend less?

Problem understanding:

We have to find the region and channel which has spent more and less.

Before that, below snippet helps us understand the data in a broader way that is mean we can find the count, unique values, frequency, mean, standard deviation max and min and also IQR values of each variables in the dataset.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Buyer/Spender	440	NaN	NaN	NaN	220.5	127.161	1	110.75	220.5	330.25	440
Channel	440	2	Hotel	298	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Region	440	3	Other	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Fresh	440	NaN	NaN	NaN	12000.3	12647.3	3	3127.75	8504	16933.8	112151
Milk	440	NaN	NaN	NaN	5796.27	7380.38	55	1533	3627	7190.25	73498
Grocery	440	NaN	NaN	NaN	7951.28	9503.16	3	2153	4755.5	10655.8	92780
Frozen	440	NaN	NaN	NaN	3071.93	4854.67	25	742.25	1526	3554.25	60869
Detergents_Paper	440	NaN	NaN	NaN	2881.49	4767.85	3	256.75	816.5	3922	40827
Delicatessen	440	NaN	NaN	NaN	1524.87	2820.11	3	408.25	965.5	1820.25	47943

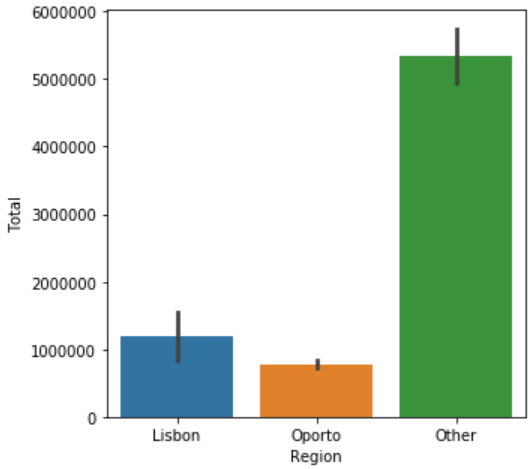
We can conclude that the dataset has,

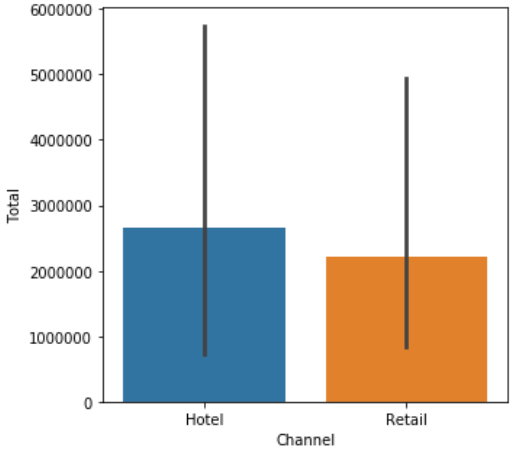
- 440 counts in all the variables
- Two unique values in channel variable
- Three unique values in Region variable
- Mean values of the variable are different
- The min value seems to be 3 for Fresh, Grocery, Detergents Paper and Delicatessen
- From the IQR values we understand the range of data lies in the 25%, 50%, 75%
- The max value seems to be 112151 holding by Fresh

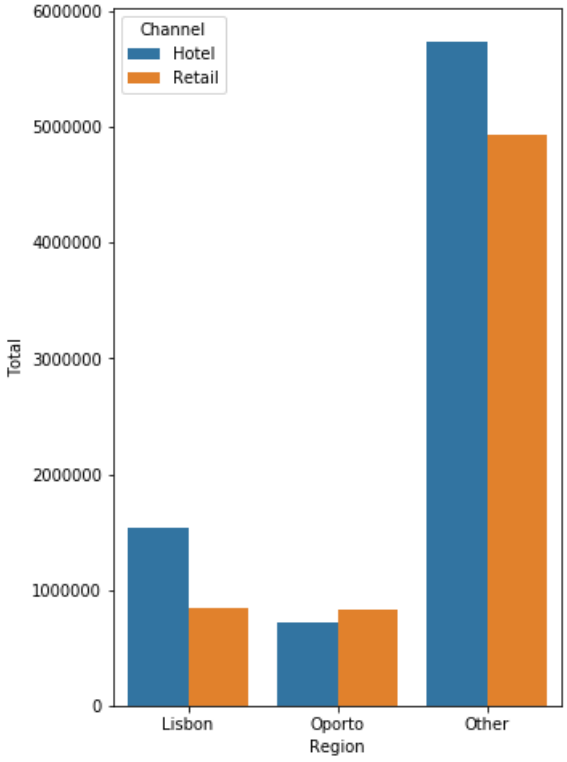
	Region	Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
0	Lisbon	Hotel	761233	228342	237542	184512	56081	70632	1538342
1	Lisbon	Retail	93600	194112	332495	46514	148055	33695	848471
2	Oporto	Hotel	326215	64519	123074	160861	13516	30965	719150
3	Oporto	Retail	138506	174625	310200	29271	159795	23541	835938
4	Other	Hotel	2928269	735753	820101	771606	165990	320358	5742077
5	Other	Retail	1032308	1153006	1675150	158886	724420	191752	4935522

From the below graph we can conclude that the

Region-Other spends more
Channel-Hotel spends more
Region-Oporto spends less
Channel –retail spends less







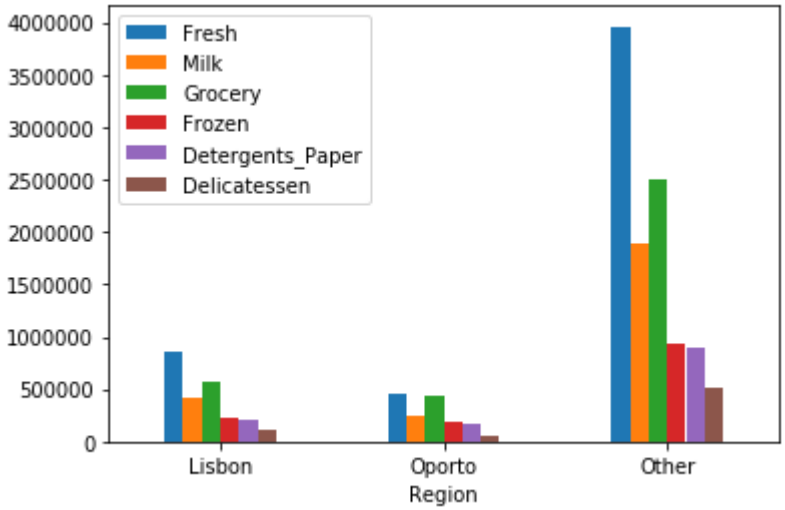
To conclude if we see, we understand in other region hotel spends more and in Oporto region retail spends less.

1.2 There are 6 different varieties of items are considered. Do all varieties show similar behaviour across Region and Channel?

Problem understanding:

We have 6 varieties so, if we see across Region we can find insights on each of 6 varieties.

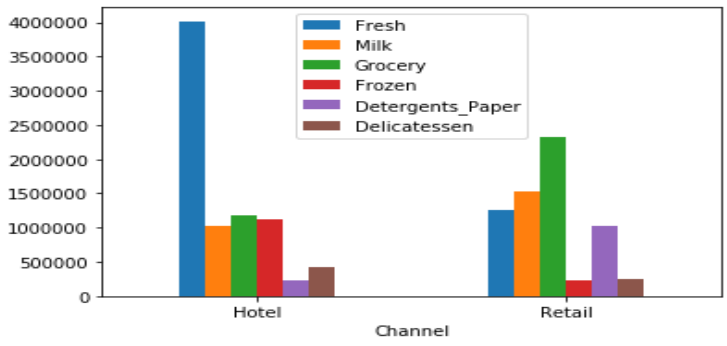
Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Lisbon	854833	422454	570037	231026	204136	104327
Oporto	464721	239144	433274	190132	173311	54506
Other	3960577	1888759	2495251	930492	890410	512110



In OTHER REGION we can the spending is maximum on all varieties, and in the OPORTO REGION we can find the spending is less on all varieties.

We have 6 varieties so, if we see across Channel we can find insights on each of 6 varieties.

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



Across channel if we the spending on varieties is different for each products, we can Fresh variety spends large in Hotel channel where it is less in Retail channel.

1.3 based on a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?

Problem understanding,

We can use IQR method or STD to find the MOST and LEAST inconsistent item.

IQR for all 6 varieties

Fresh	**13806.00**
Milk	**5657.25**
Grocery	**8502.75**
Frozen	**2812.00**
Detergents Paper	**3665.25**
Delicatessen	**1412.00**

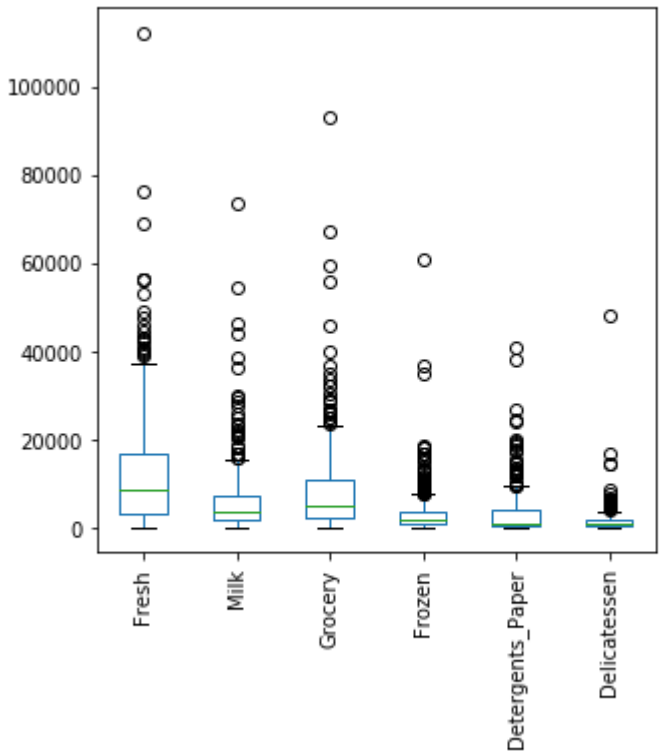
Fresh varieties seems to be the most inconsistent in terms of spending by the buyer

Delicatessen varieties seems to be least inconsistent in terms of spending by the buyer

1.4 Are there any outliers in the data?

YES all 6 varieties have outliers in the dataset.

We can use IQR or box plot to find the outliers in the dataset



By IQR method,

	Channel	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk	Region
0	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False
2	False	True	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False
4	False	True	False	False	False	False	False	False

All the true have indicate the presence of outliers in the dataset. I have used head to display only top 5 rows. But all varieties have outliers in the dataset.

1.5 based on **this report, what are the recommendations?**
Based on the analysis done so far it is evident that there are many buyers in other region so we could improve our retailer base in Lisbon and Oporto region as well
We can find evidence Fresh items are considerably used by many retailers among all the region. We find that varieties like Frozen, Detergents paper and Delicatessen are not Popular among the retailers so we can try maximize these products where the demand Is more. As of now there is only two modes of sales either by Retail or Hotel. Considerin g the base of the region we can use multiple sale method to reach our customers.