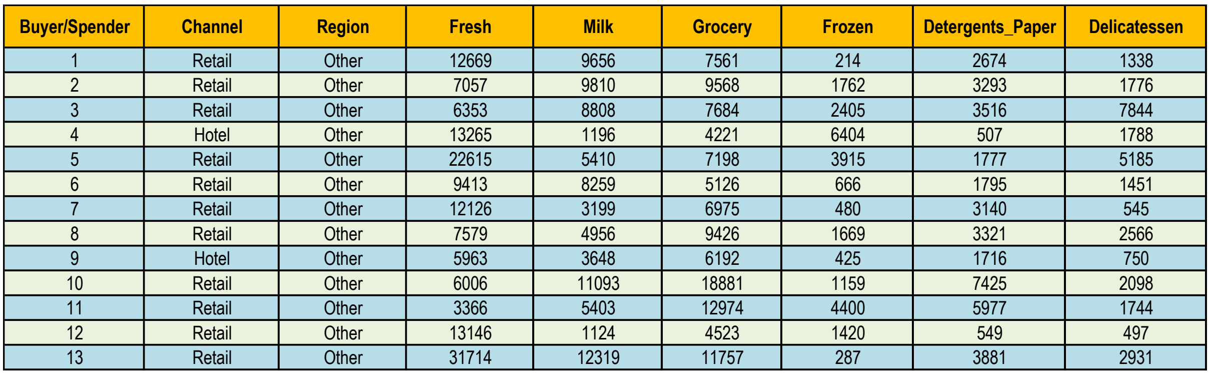
Problem 1 Statement

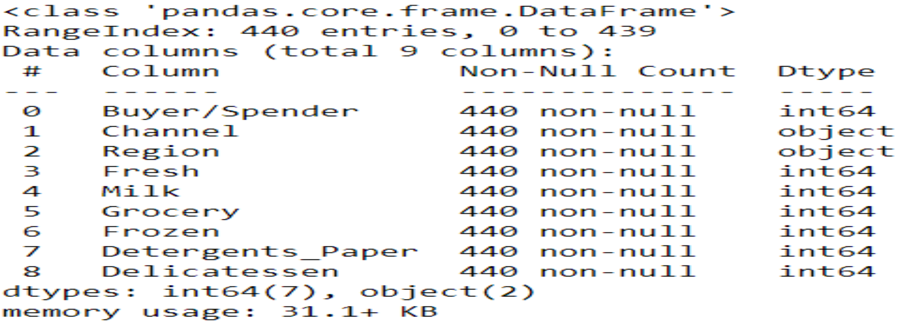
A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data (Wholesale Customer.csv) consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel, Retail).

Exploratory Data Analysis :



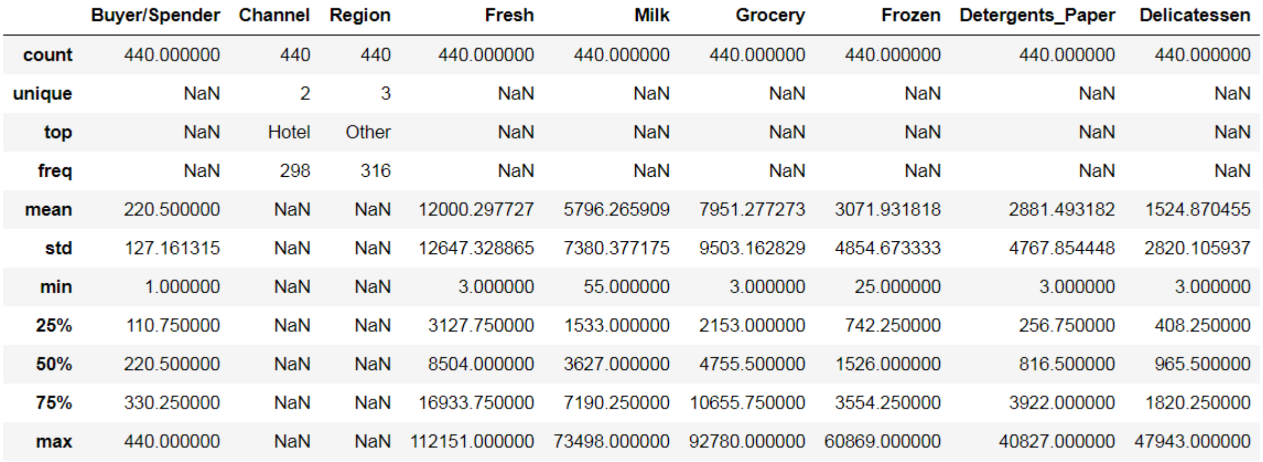
Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
1	Retail	Other	12669	9656	7561	214	2674	1338
2	Retail	Other	7057	9810	9568	1762	3293	1776
3	Retail	Other	6353	8808	7684	2405	3516	7844
4	Hotel	Other	13265	1196	4221	6404	507	1788
5	Retail	Other	22615	5410	7198	3915	1777	5185
6	Retail	Other	9413	8259	5126	666	1795	1451
7	Retail	Other	12126	3199	6975	480	3140	545
8	Retail	Other	7579	4956	9426	1669	3321	2566
9	Hotel	Other	5963	3648	6192	425	1716	750
10	Retail	Other	6006	11093	18881	1159	7425	2098
11	Retail	Other	3366	5403	12974	4400	5977	1744
12	Retail	Other	13146	1124	4523	1420	549	497
13	Retail	Other	31714	12319	11757	287	3881	2931

We are provided with the above data set of 440 rows and 9 columns. Among the available columns 2 are categorical type and 7 are integer type. The data has no Null values.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440 entries, 0 to 439
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Buyer/Spender     440 non-null    int64 
 1   Channel           440 non-null    object
 2   Region            440 non-null    object
 3   Fresh             440 non-null    int64 
 4   Milk              440 non-null    int64 
 5   Grocery           440 non-null    int64 
 6   Frozen            440 non-null    int64 
 7   Detergents_Paper  440 non-null    int64 
 8   Delicatessen      440 non-null    int64 
dtypes: int64(7), object(2)
memory usage: 31.1+ KB
```

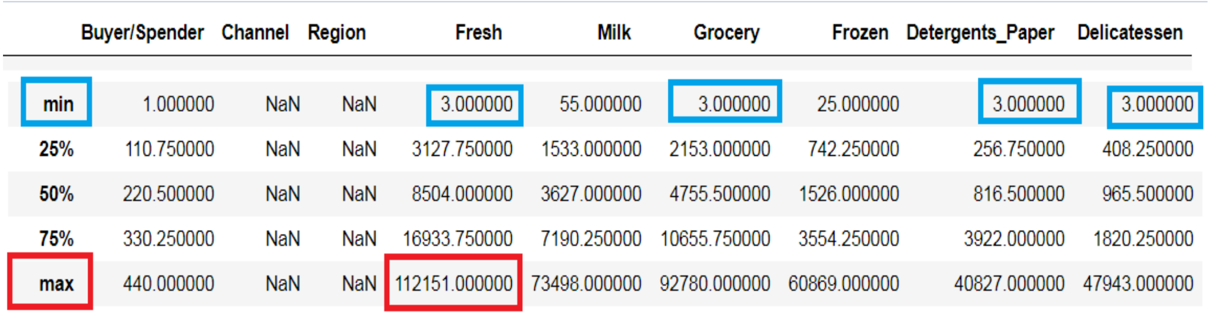
Descriptive statistics for the dataset:



	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440	440	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
unique	NaN	2	3	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Hotel	Other	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	298	316	NaN	NaN	NaN	NaN	NaN	NaN
mean	220.500000	NaN	NaN	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	127.161315	NaN	NaN	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	1.000000	NaN	NaN	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	110.750000	NaN	NaN	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	220.500000	NaN	NaN	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	330.250000	NaN	NaN	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	440.000000	NaN	NaN	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

There are 2 unique values in column **Channel**, out of which '**Hotel**' appears mostly with frequency of 298 times and also 3 unique values in column **Region** ,out of which '**Other**' has highest repetition with frequency of 316 times.

As per the details resulted from the descriptive statistics of the dataset , we can find that :



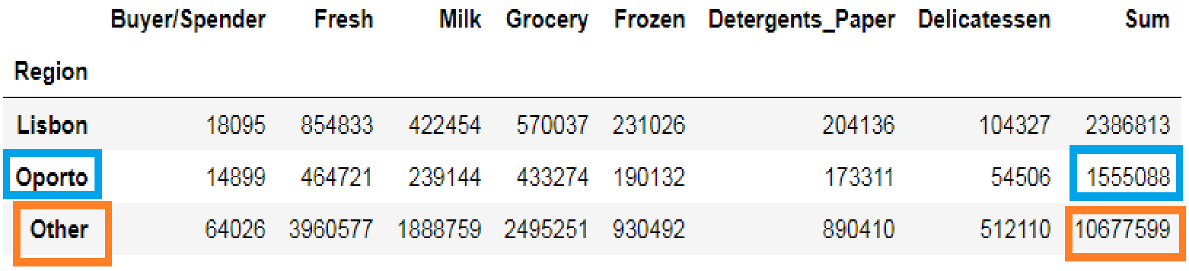
	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
min	1.000000	NaN	NaN	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	110.750000	NaN	NaN	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	220.500000	NaN	NaN	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	330.250000	NaN	NaN	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	440.000000	NaN	NaN	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

Column 'Fresh' has the highest annual spending with value = 112151

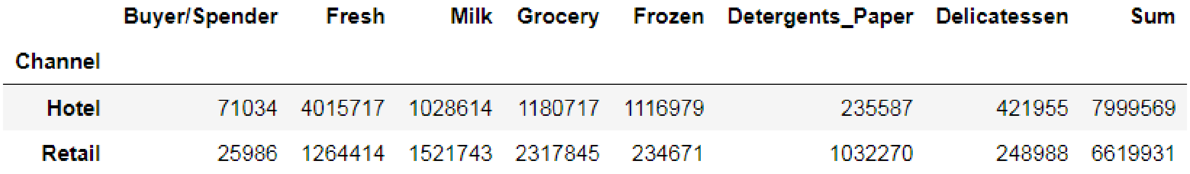
Columns 'Fresh' , 'Grocery', 'Detergents_Paper', 'Delicatessen' have the least annual spending with value = 3

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?



	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Sum
Region								
Lisbon	18095	854833	422454	570037	231026	204136	104327	2386813
Oporto	14899	464721	239144	433274	190132	173311	54506	1555088
Other	64026	3960577	1888759	2495251	930492	890410	512110	10677599



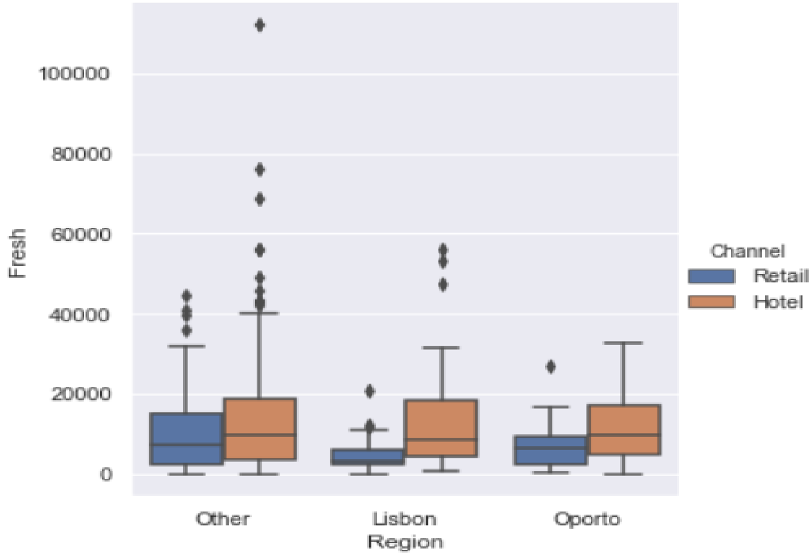
	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Sum
Channel								
Hotel	71034	4015717	1028614	1180717	1116979	235587	421955	7999569
Retail	25986	1264414	1521743	2317845	234671	1032270	248988	6619931

Region – **'Other'** , has **highest** annual spending
Region – '**Oporto'** has **lowest** annual spending
Channel – **'Hotel'** has **higher** annual spending
Channel – **'Retail'** has **lower** annual spending

1.2. There are 6 different varieties of items are considered.

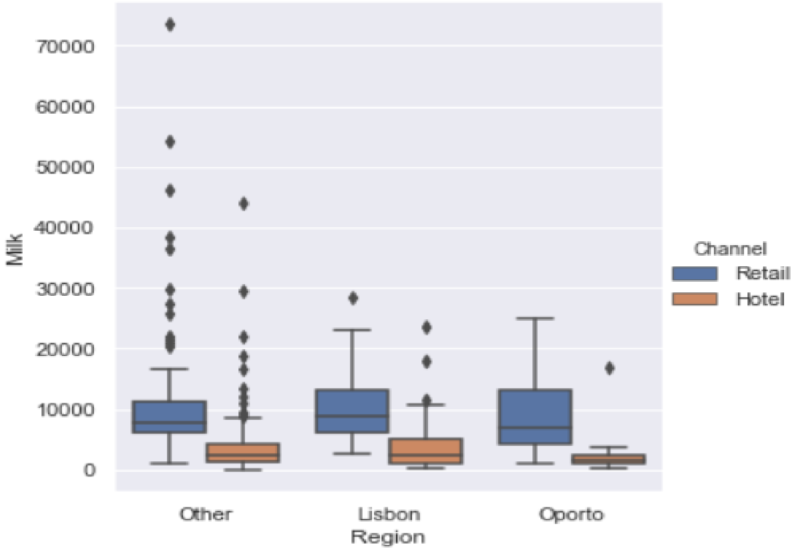
Do all varieties show similar behaviour across Region and Channel

Item – Fresh



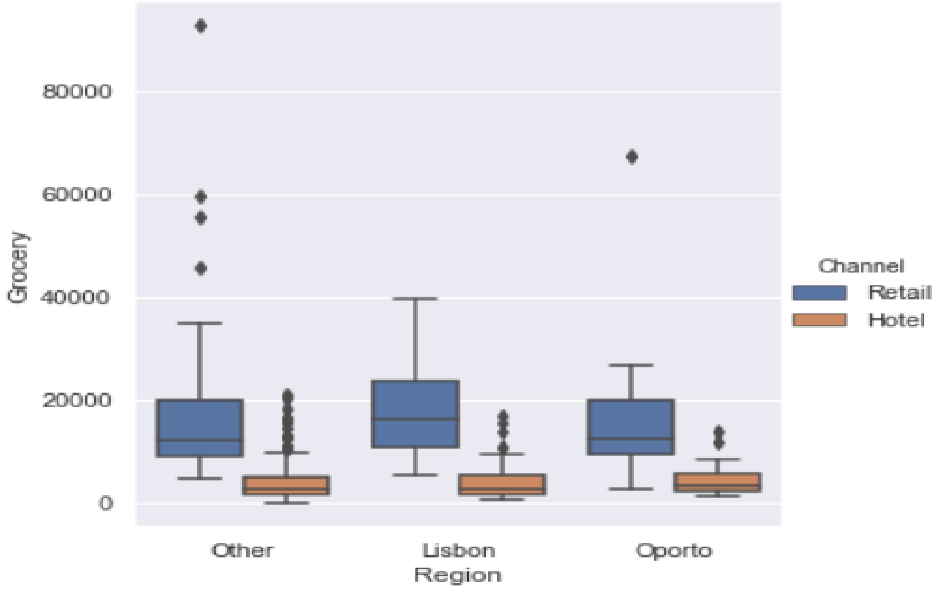
From above plot , we can find that Item ‘Fresh’ has more annual spending in Channel – ‘Hotel’ & Region -‘Other’

Item – Milk



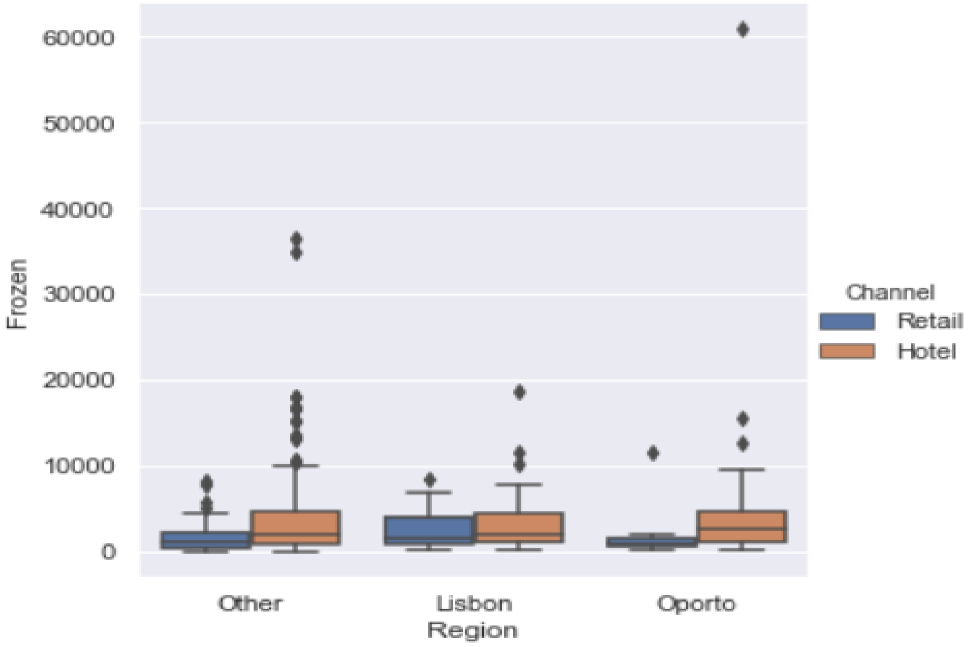
From above plot , we can find that Item ‘Milk’ has more annual spending in Channel – ‘Retail’ & Region – ‘Lisbon’

Item – Grocery



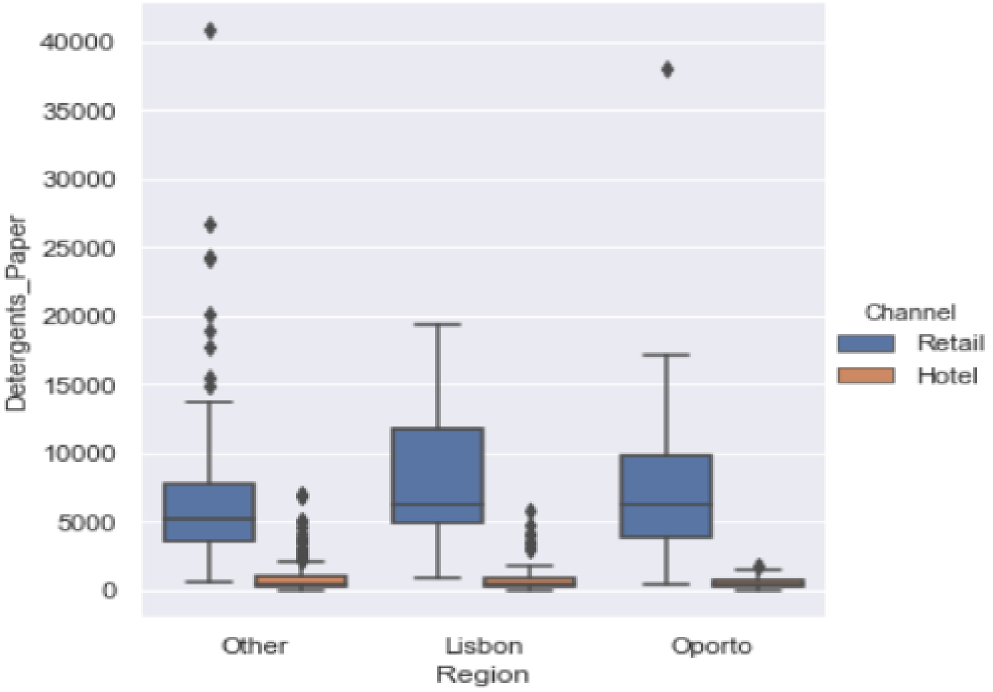
From above plot , we can find that Item ‘Grocery’ has more annual spending in Channel – ‘Retail’ & Region – ‘Lisbon’

Item – Frozen



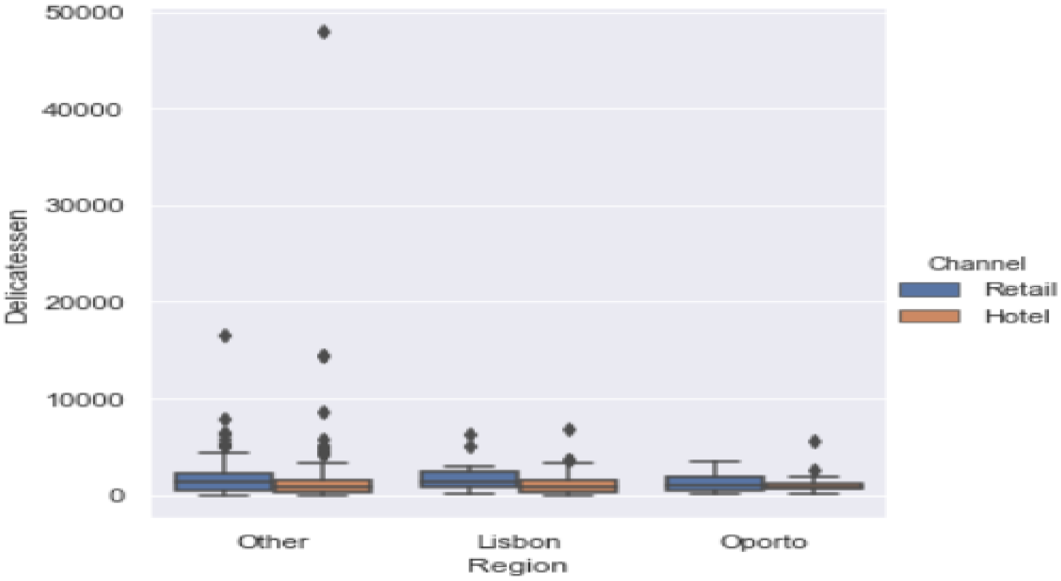
From above plot , we can find that Item ‘Frozen’ has more annual spending in Channel – ‘Hotel’ & Region – ‘Lisbon’

Item – Detergents_Paper



From above plot , we can find that Item ‘Detergents_Paper’ has more annual spending in Channel – ‘Retail’ & Region – ‘Lisbon’

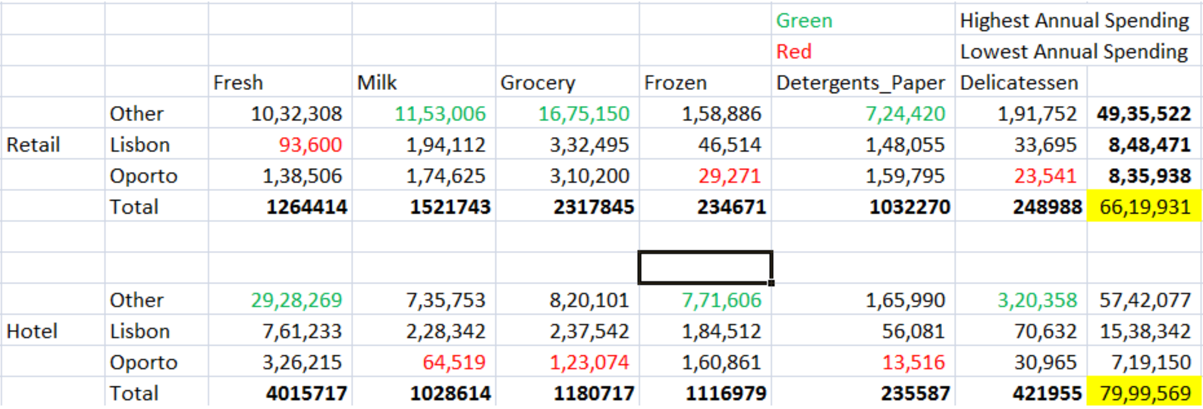
Item – Delicatessen



From above plot , we can find that Item ‘Delicatessen’ has more annual spending in Channel – ‘Retail’ & Region – ‘Other’

So, overall we can say that,

There is higher annual spending in channel ‘Hotel’ compared to channel ‘Retail’



						Green	Highest Annual Spending	
						Red	Lowest Annual Spending	
		Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	
	Other	10,32,308	11,53,006	16,75,150	1,58,886	7,24,420	1,91,752	**49,35,522**
Retail	Lisbon	93,600	1,94,112	3,32,495	46,514	1,48,055	33,695	**8,48,471**
	Oporto	1,38,506	1,74,625	3,10,200	29,271	1,59,795	23,541	**8,35,938**
	Total	**1264414**	**1521743**	**2317845**	**234671**	**1032270**	**248988**	66,19,931
	Other	29,28,269	7,35,753	8,20,101	7,71,606	1,65,990	3,20,358	57,42,077
Hotel	Lisbon	7,61,233	2,28,342	2,37,542	1,84,512	56,081	70,632	15,38,342
	Oporto	3,26,215	64,519	1,23,074	1,60,861	13,516	30,965	7,19,150
	Total	**4015717**	**1028614**	**1180717**	**1116979**	**235587**	**421955**	79,99,569

As per colour highlighted values we can see that most of the items have higher annual spending in Region – 'Other' and lower annual spending in Region – 'Oporto'.
Milk ,Grocery and Detergents_Paper items are showing similar behaviour having highest spending in region –Other and least spending in region-'Oporto'

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?

We have the formula for coefficient of variation as

$$\mathrm{CV} = \frac{\sigma}{\mu}$$

$\textbf{where:}$

$\sigma = \text{standard deviation}$

$\mu = \text{mean}$

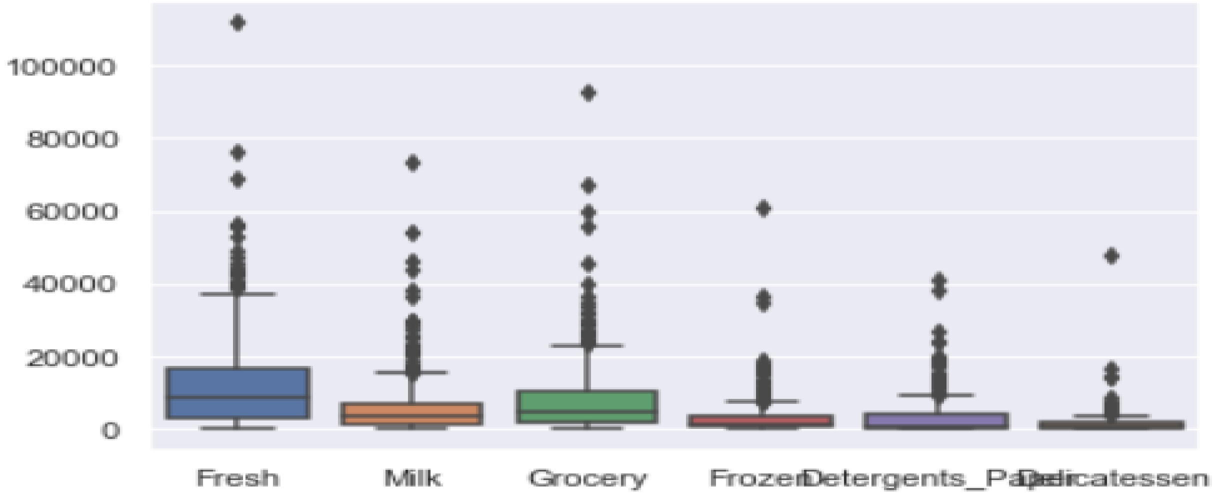
By calculating CV for all the items

	Fresh	**Milk**	**Grocery**	**Frozen**	**Detergents_Paper**	**Delicatessen**
Mean	12000.298	5796.266	7951.277	3071.932	2881.493	1524.870
std	12647.329	7380.377	9503.163	4854.673	4767.854	2820.106
CV	1.054	1.273	1.195	1.580	1.655	1.849

Item – '**Delicatessen**' shows the **most** inconsistent behaviour

Item – **'Fresh'** shows the **least** inconsistent behaviour

1.4. Are there any outliers in the data?



By looking at the box plot for all the items , we can conclude that all the items have outliers.

1.5. On the basis of this report, what are the recommendations?

We can summarize that overall annual spending is less in Retail channel compared to Hotel Channel. Therefore certain measures can be taken to increase annual spending in Retail channel.

Even comparing regional wise , we have more spending in Region- ‘Other’ and less spending in Region- ‘Oporto’. We have to explore options to increase in this region.

Also item ‘Fresh’ has highest spending and item ‘Delicatessen’ has least spending. Stocks have to be maintained accordingly to meet the demand and to increase the spending values.