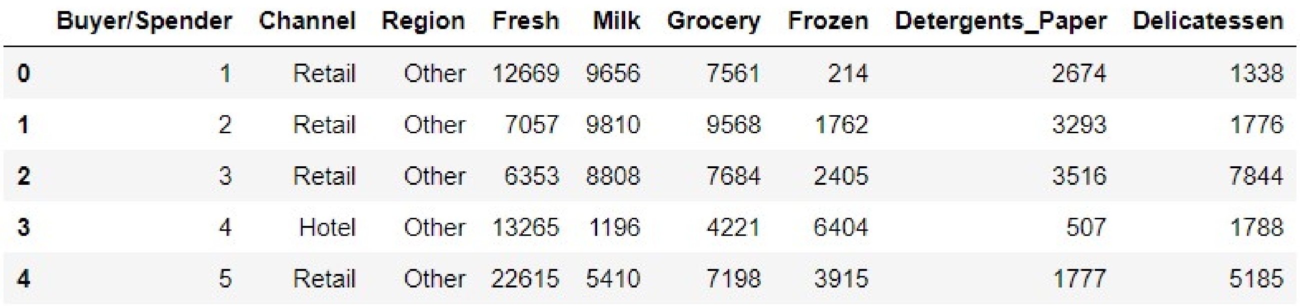
Problem Statement:

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).

Exploratory Data Analysis:



	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	1	Retail	Other	12669	9656	7561	214	2674	1338
1	2	Retail	Other	7057	9810	9568	1762	3293	1776
2	3	Retail	Other	6353	8808	7684	2405	3516	7844
3	4	Hotel	Other	13265	1196	4221	6404	507	1788
4	5	Retail	Other	22615	5410	7198	3915	1777	5185

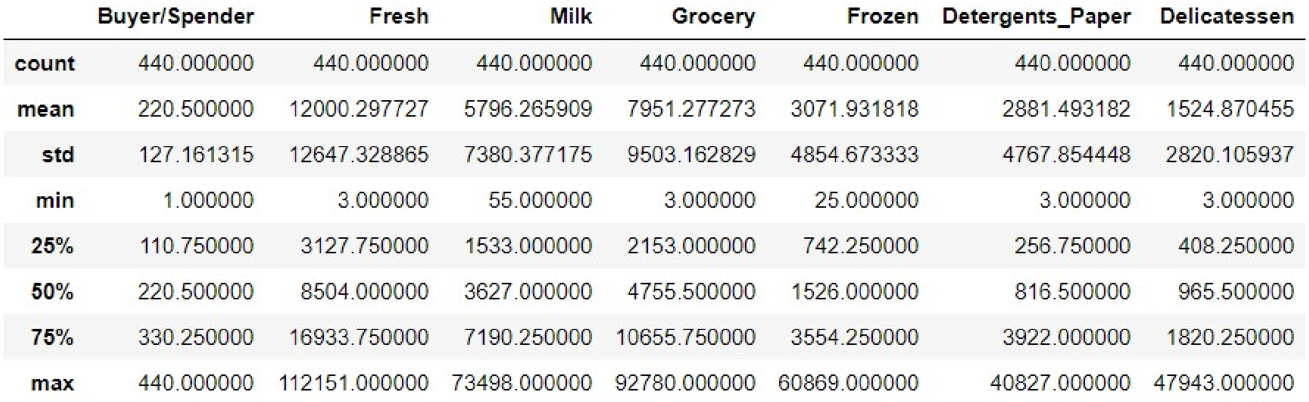
The Dataset has 9 variables:

Channel and Region are categorical variables

Buyer/Spender, Fresh, Milk, Grocery, Frozen, Detergents_Paper, Delicatessen are integer type.

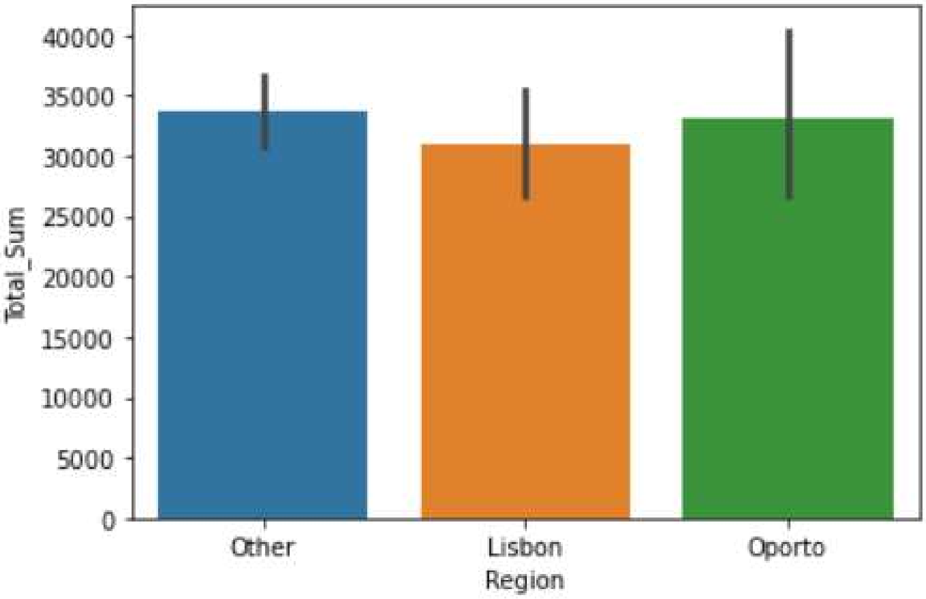
1.1 Use methods of descriptive statistics to summarize data. Which Region and which Channel seems to spend more? Which Region and which Channel seems to spend less?

To summarize the data, we have used the describe function:

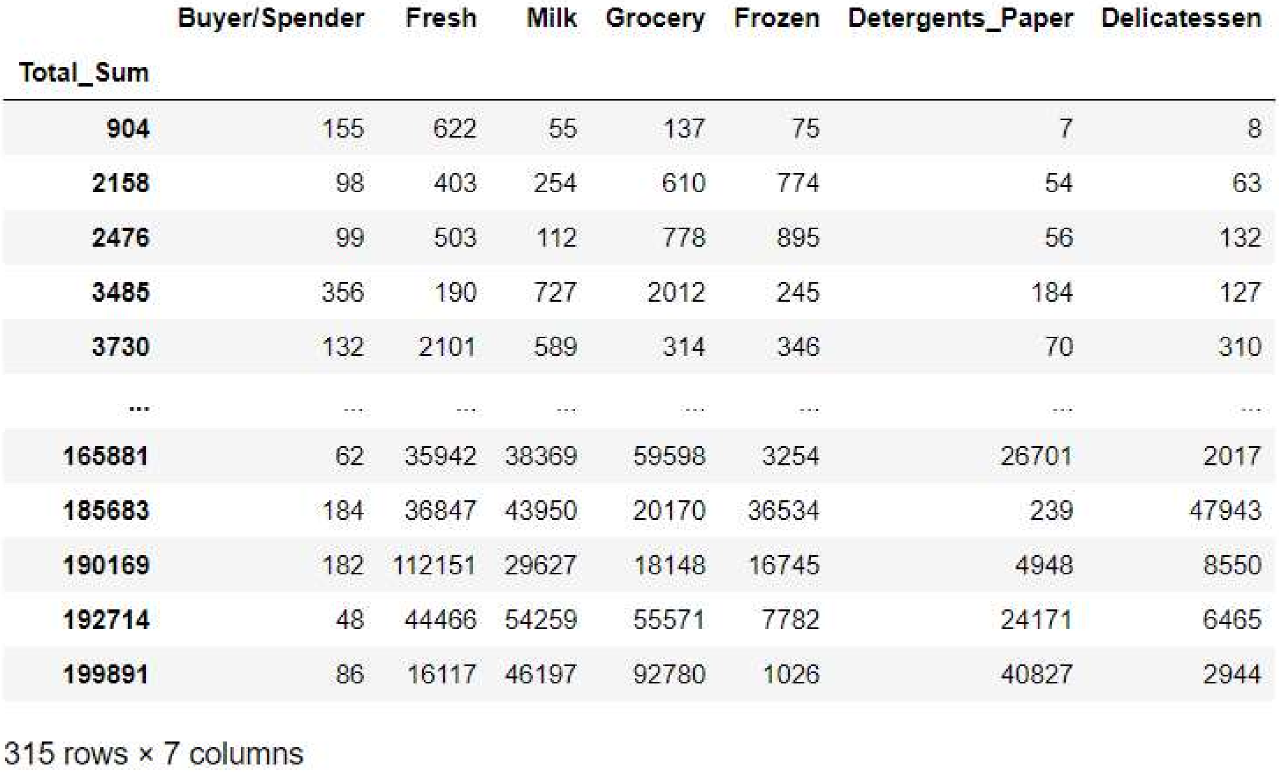


	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	220.500000	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	127.161315	12647.328865	7380.377175	9503.162829	4854.673333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

The region: Other spends more with 199891, region: Lisbon spends least with 107155 and the region: Oporto spends 130877



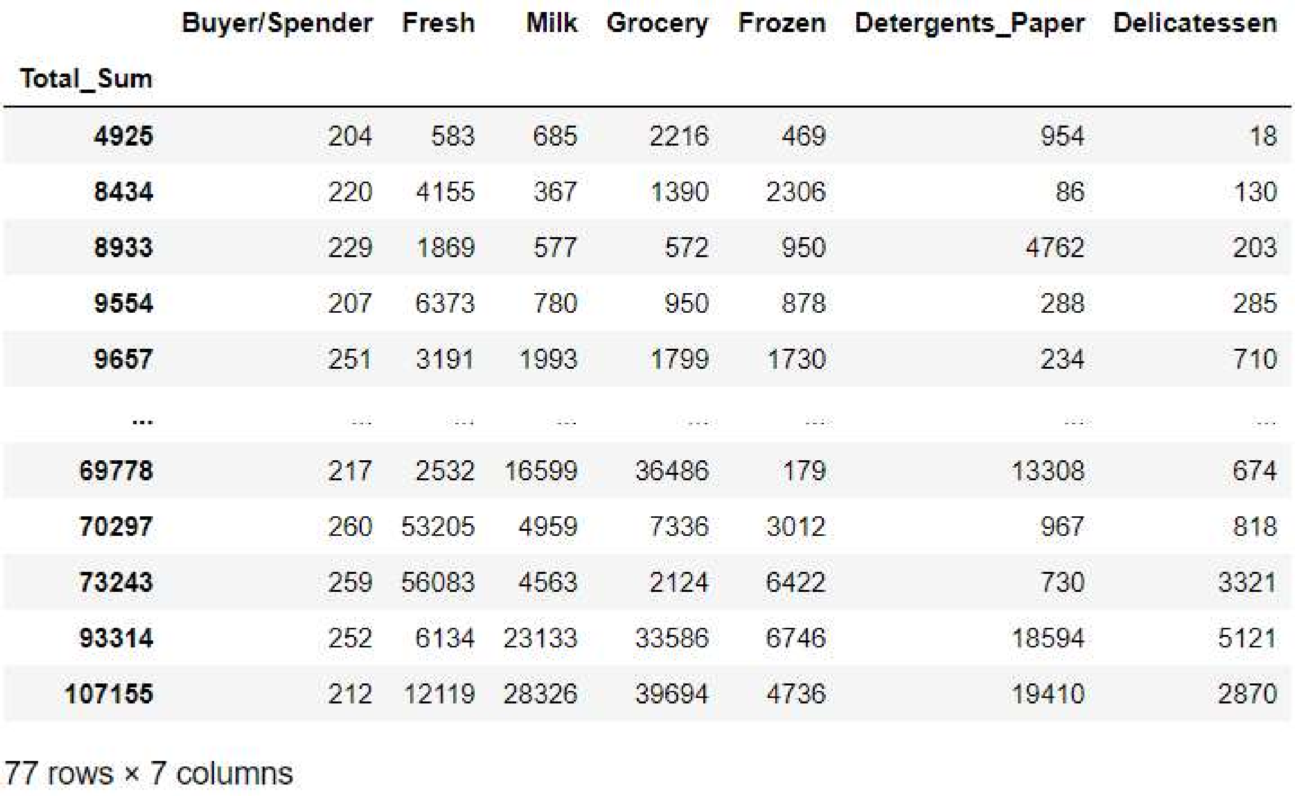
Total amount spent for the region ‘Other’ in the descending order of retailers:



	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Total_Sum							
904	155	622	55	137	75	7	8
2158	98	403	254	610	774	54	63
2476	99	503	112	778	895	56	132
3485	356	190	727	2012	245	184	127
3730	132	2101	589	314	346	70	310
...	...	...	...	...	...	...	...
165881	62	35942	38369	59598	3254	26701	2017
185683	184	36847	43950	20170	36534	239	47943
190169	182	112151	29627	18148	16745	4948	8550
192714	48	44466	54259	55571	7782	24171	6465
199891	86	16117	46197	92780	1026	40827	2944

315 rows × 7 columns

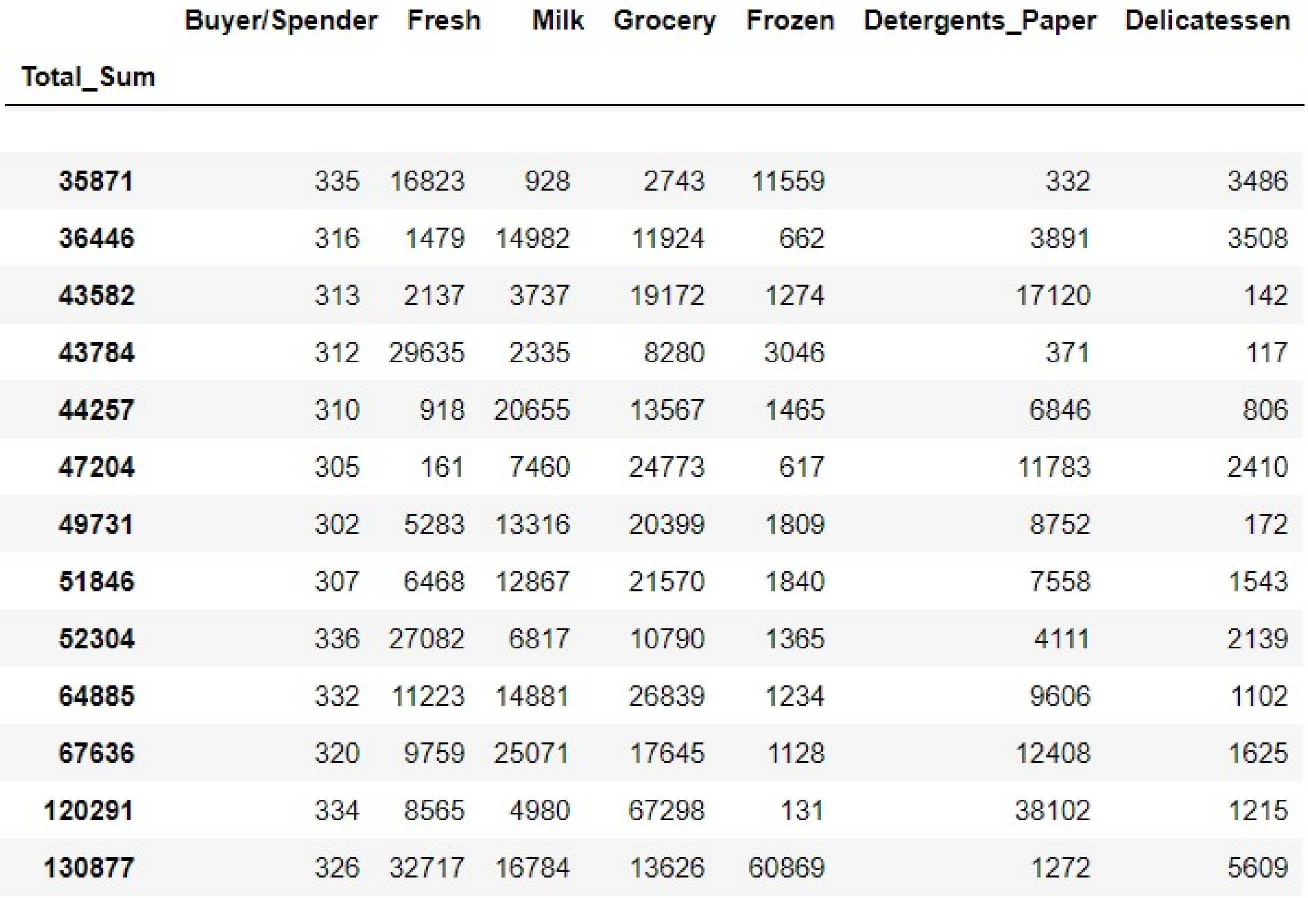
Total amount spent for the region ‘Lisbon’ in the descending order of retailers:



	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Total_Sum							
4925	204	583	685	2216	469	954	18
8434	220	4155	367	1390	2306	86	130
8933	229	1869	577	572	950	4762	203
9554	207	6373	780	950	878	288	285
9657	251	3191	1993	1799	1730	234	710
...	...	...	...	...	...	...	...
69778	217	2532	16599	36486	179	13308	674
70297	260	53205	4959	7336	3012	967	818
73243	259	56083	4563	2124	6422	730	3321
93314	252	6134	23133	33586	6746	18594	5121
107155	212	12119	28326	39694	4736	19410	2870

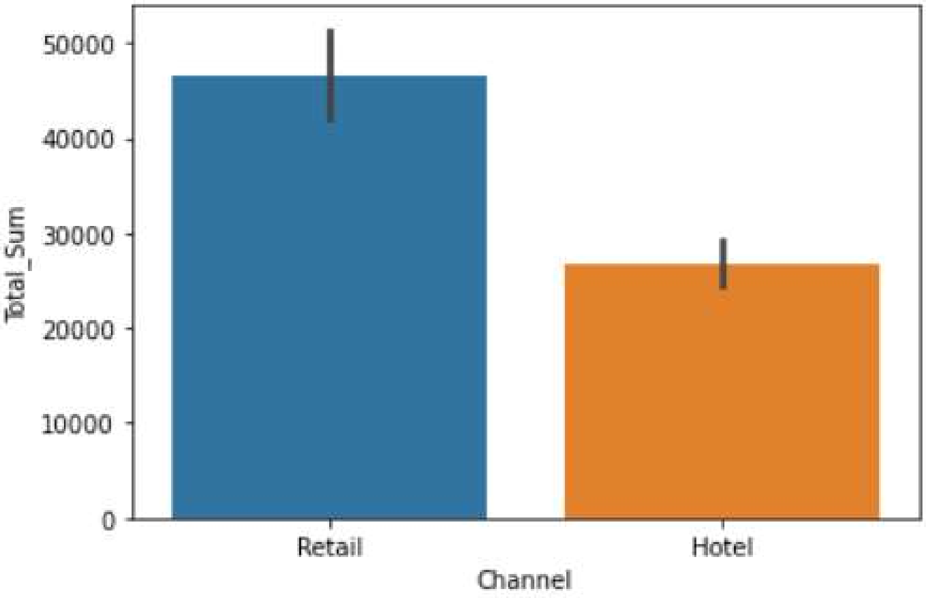
77 rows × 7 columns

Total amount spent for the region ‘Oporto’ in the descending order of retailers:

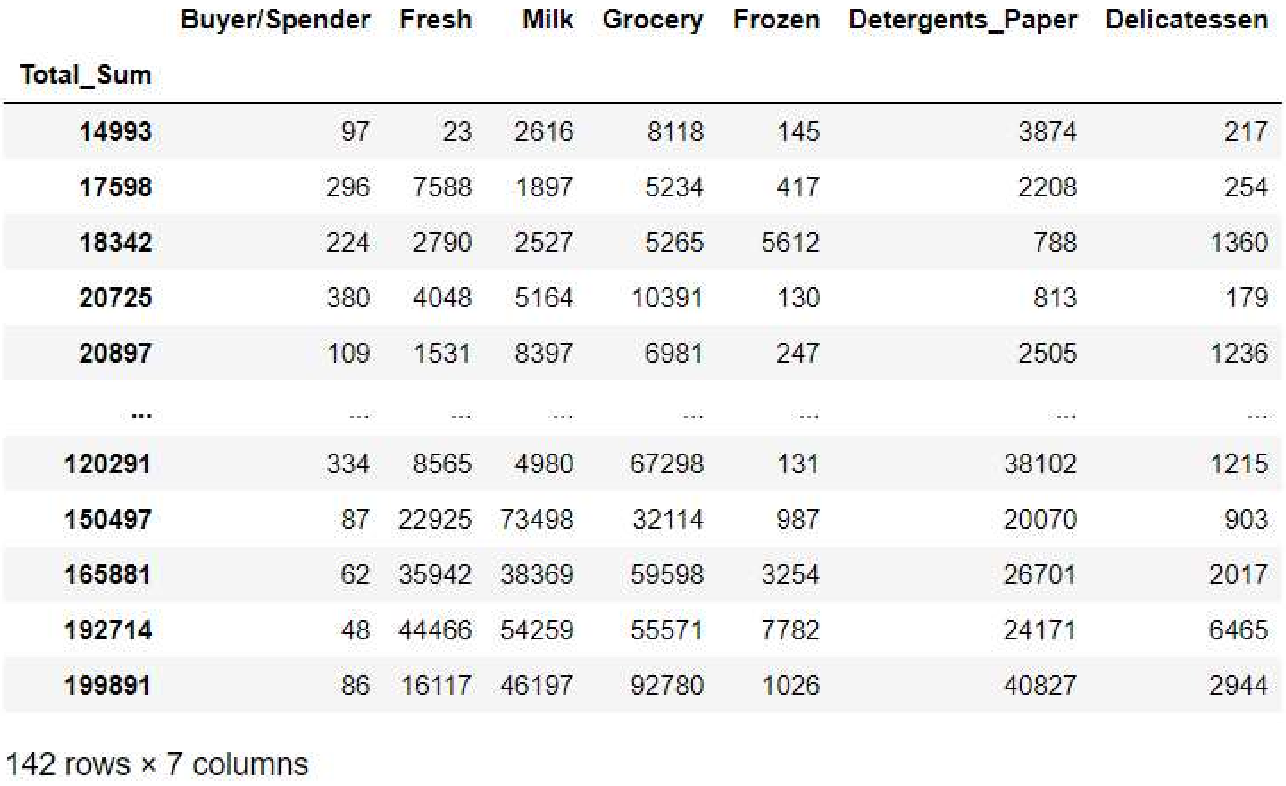


	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Total_Sum							
35871	335	16823	928	2743	11559	332	3486
36446	316	1479	14982	11924	662	3891	3508
43582	313	2137	3737	19172	1274	17120	142
43784	312	29635	2335	8280	3046	371	117
44257	310	918	20655	13567	1465	6846	806
47204	305	161	7460	24773	617	11783	2410
49731	302	5283	13316	20399	1809	8752	172
51846	307	6468	12867	21570	1840	7558	1543
52304	336	27082	6817	10790	1365	4111	2139
64885	332	11223	14881	26839	1234	9606	1102
67636	320	9759	25071	17645	1128	12408	1625
120291	334	8565	4980	67298	131	38102	1215
130877	326	32717	16784	13626	60869	1272	5609

The channel: Retail spends more with 199891 and the channel: Hotel spends less with 190169



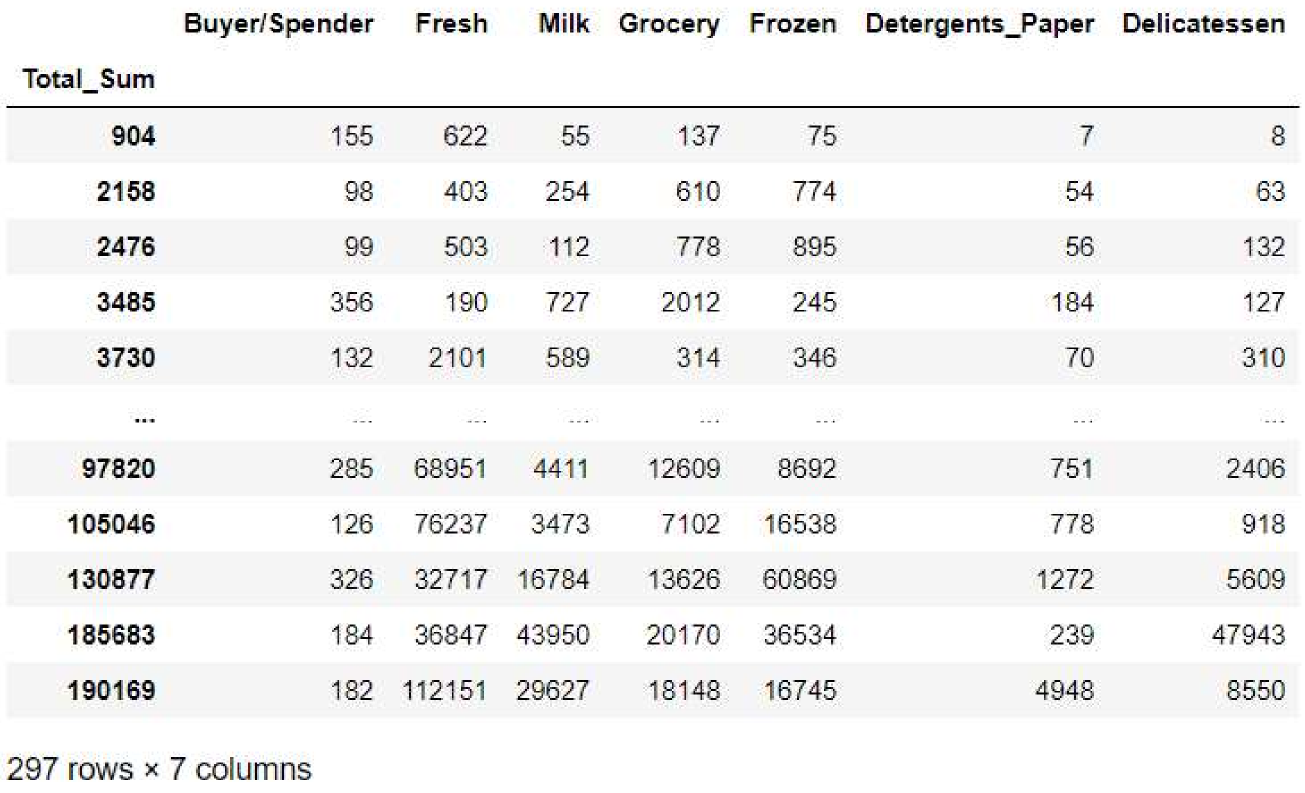
Shows the listing of the total amount spent for the products across the Retail Channel in the descending order:



	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Total_Sum							
14993	97	23	2616	8118	145	3874	217
17598	296	7588	1897	5234	417	2208	254
18342	224	2790	2527	5265	5612	788	1360
20725	380	4048	5164	10391	130	813	179
20897	109	1531	8397	6981	247	2505	1236
...	...	...	...	...	...	...	...
120291	334	8565	4980	67298	131	38102	1215
150497	87	22925	73498	32114	987	20070	903
165881	62	35942	38369	59598	3254	26701	2017
192714	48	44466	54259	55571	7782	24171	6465
199891	86	16117	46197	92780	1026	40827	2944

142 rows × 7 columns

Shows the listing of the total amount spent for the products across the Hotel Channel in the descending order:



	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Total_Sum							
904	155	622	55	137	75	7	8
2158	98	403	254	610	774	54	63
2476	99	503	112	778	895	56	132
3485	356	190	727	2012	245	184	127
3730	132	2101	589	314	346	70	310
...	...	...	...	...	...	...	...
97820	285	68951	4411	12609	8692	751	2406
105046	126	76237	3473	7102	16538	778	918
130877	326	32717	16784	13626	60869	1272	5609
185683	184	36847	43950	20170	36534	239	47943
190169	182	112151	29627	18148	16745	4948	8550

297 rows × 7 columns

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

The answer can be derived out of the describe function used earlier.

The money spent on Fresh Milk is the maximum at 112151.

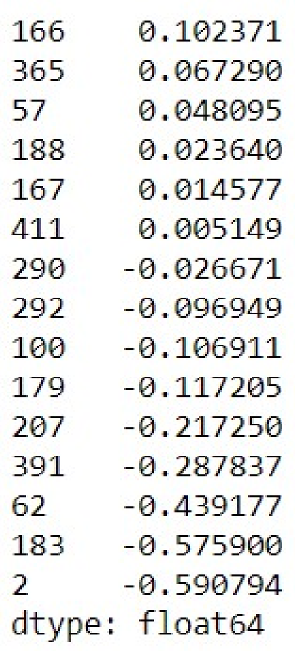
The Standard Deviation is minimum for Delicatessen with 2820.105937

The minimum amount has been spent on Fresh, Grocery, Detergents_Paper, Delicatessen with 3.

The highest IQR (75th percentile - 25th percentile) is for the product: Fresh with (16933.7500 - 3127.7500) = 13806

The lowest IQR is for the product: Delicatessen with (1820.2500 - 408.2500) = 1412

We have listed the skew value using the function: mydata.skew(axis = 1) and then sorted the skew value in the descending order between row 425 and 440 because this is where the values are separated between positive and negative.

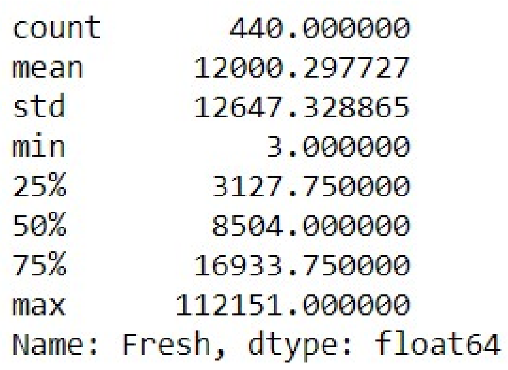


```
166    0.102371
365    0.067290
57     0.048095
188    0.023640
167    0.014577
411    0.005149
290   -0.026671
292   -0.096949
100   -0.106911
179   -0.117205
207   -0.217250
391   -0.287837
62    -0.439177
183   -0.575900
2     -0.590794
dtype: float64
```

We can see that most of the values in skewness are greater than 0 which means that there is more weight in the Left tail of the distribution.

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?

First, we use the describe function to get the standard deviation and mean for the product Fresh.

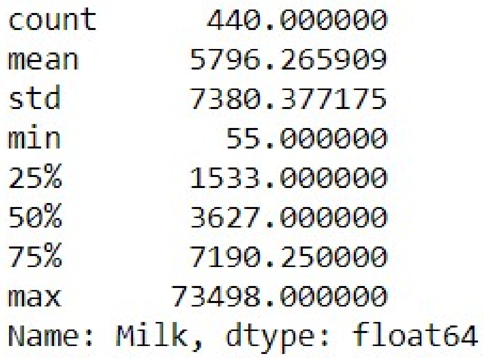


```
count       440.000000
mean      12000.297727
std       12647.328865
min           3.000000
25%        3127.750000
50%        8504.000000
75%       16933.750000
max      112151.000000
Name: Fresh, dtype: float64
```

Coefficient of Variation for the product: Fresh is defined by the formula: coeff_var_fresh = Standard deviation/Mean)*100

coeff_var_fresh = (12647.328865/12000.297727)*100 = 105.391

Then we use the describe function to get the standard deviation and mean for the product Milk:

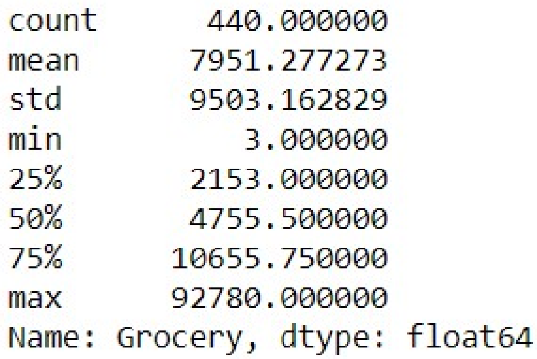


```
count      440.000000
mean      5796.265909
std       7380.377175
min         55.000000
25%       1533.000000
50%       3627.000000
75%       7190.250000
max      73498.000000
Name: Milk, dtype: float64
```

Coefficient of Variation for the product: Milk is defined by the formula: coeff_var_Milk = (Standard deviation/Mean)*100

coeff_var_Milk = (7380.377175/5796.265909)*100 = 127.329

Then we use the describe function to get the standard deviation and mean for the product Grocery:

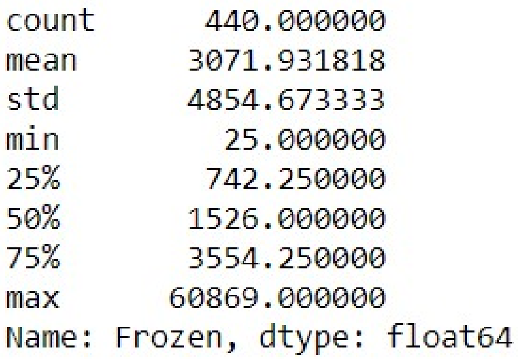


```
count      440.000000
mean      7951.277273
std       9503.162829
min          3.000000
25%       2153.000000
50%       4755.500000
75%      10655.750000
max      92780.000000
Name: Grocery, dtype: float64
```

Coefficient of Variation for the product: Grocery is defined by the formula: coeff_var_Grocery = (Standard deviation/Mean)*100

coeff_var_Grocery = (9503.162829/7951.277273)*100 = 119.517

Then we use the describe function to get the standard deviation and mean for the product Frozen:

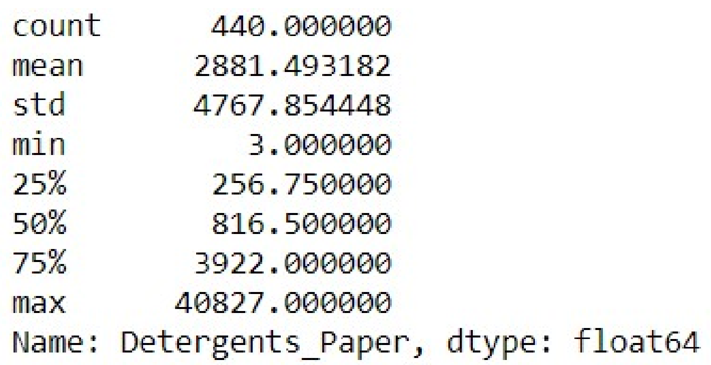


```
count      440.000000
mean      3071.931818
std       4854.673333
min         25.000000
25%        742.250000
50%       1526.000000
75%       3554.250000
max      60869.000000
Name: Frozen, dtype: float64
```

Coefficient of Variation for the product: Frozen is defined by the formula: coeff_var_Frozen = (Standard deviation/Mean)*100

coeff_var_Frozen = (4854.673333/3071.931818)*100 = 158.033

Then we use the describe function to get the standard deviation and mean for the product Detergents_Paper:

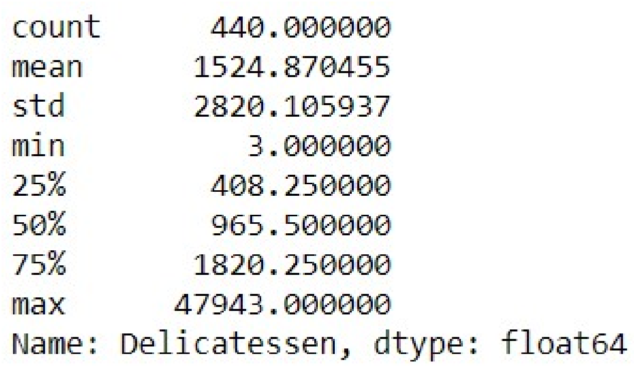


```
count      440.000000
mean      2881.493182
std       4767.854448
min          3.000000
25%        256.750000
50%        816.500000
75%       3922.000000
max      40827.000000
Name: Detergents_Paper, dtype: float64
```

Coefficient of Variation for the product: Detergents_Paper is defined by the formula:
coeff_var_Detergents_Paper = (Standard deviation/Mean)*100

coeff_var_Detergents_Paper = (4767.854448/2881.493182)*100 = 165.464

Then we use the describe function to get the standard deviation and mean for the product Delicatessen:



```
count      440.000000
mean      1524.870455
std       2820.105937
min          3.000000
25%        408.250000
50%        965.500000
75%       1820.250000
max      47943.000000
Name: Delicatessen, dtype: float64
```

Coefficient of Variation for the product: Delicatessen is defined by the formula: coeff_var_Delicatessen = (Standard deviation/Mean)*100

coeff_var_Delicatessen = (2820.105937/1524.870455)*100 = 184.940

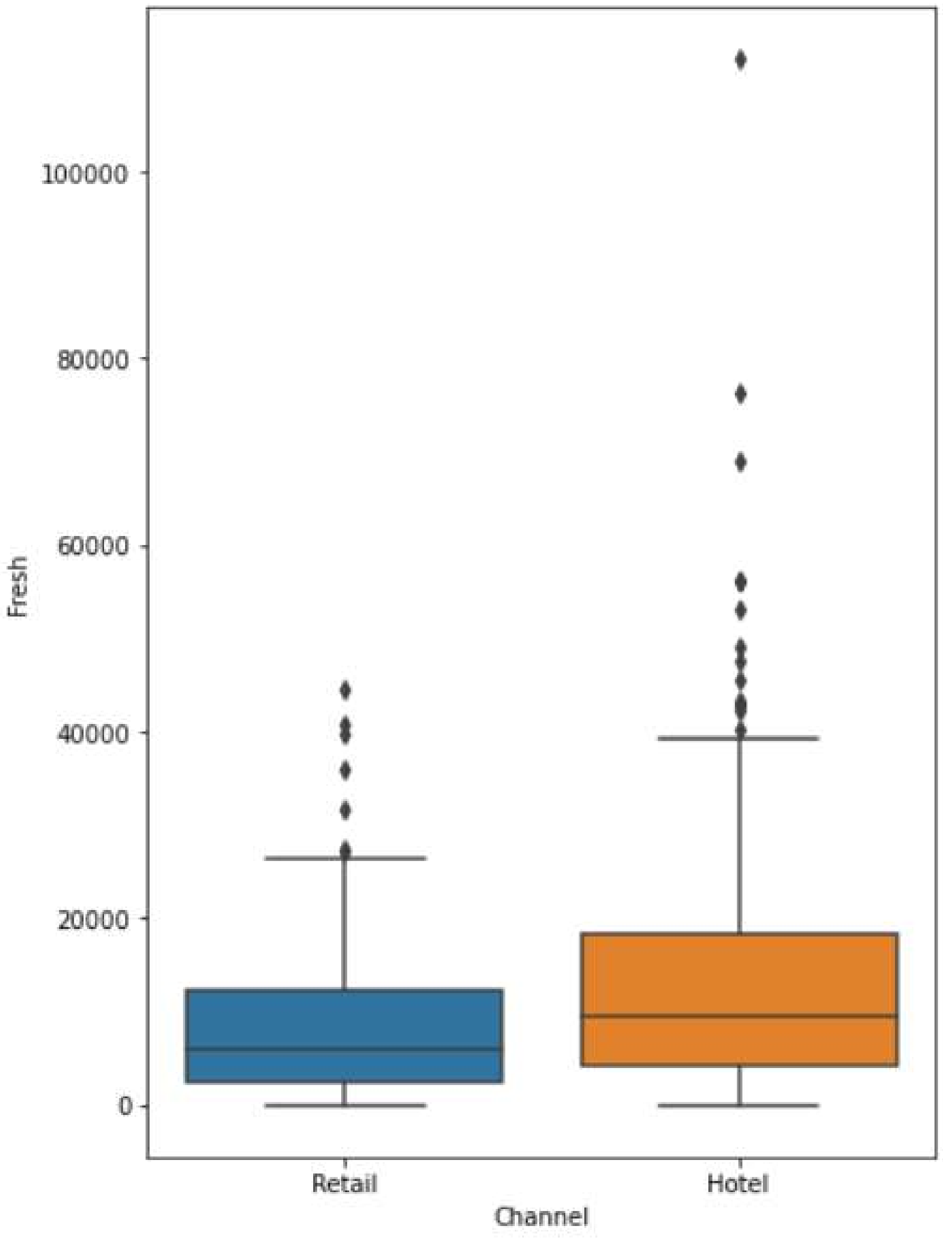
Conclusion: Delicatessen shows the most inconsistent behaviour with coefficient of variation = 184.9406897322304

Fresh shows the least inconsistent behaviour with coefficient of variation = 105.39179237648592

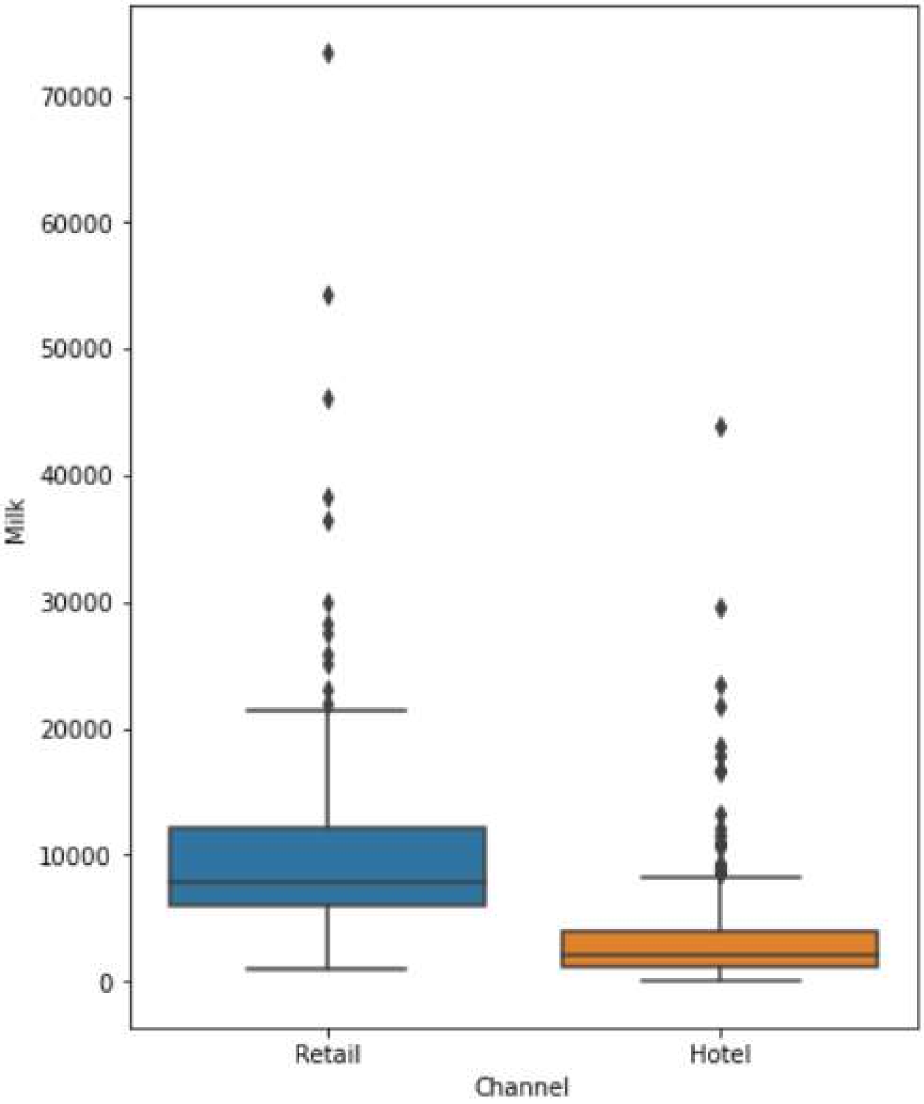
1.4: Are there any outliers in the data?

To check for outliers, we will plot Box plots across various Channels and each product and across various Regions and each product.

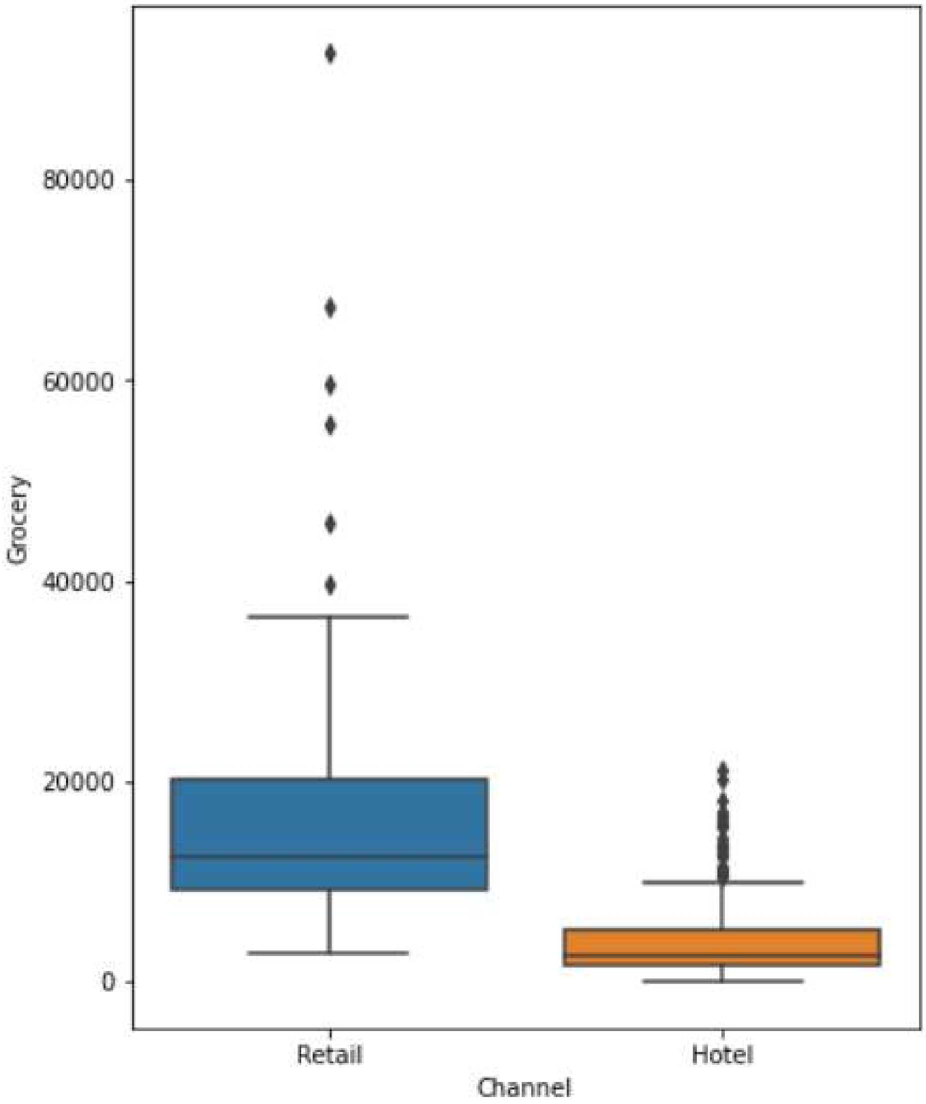
Boxplot across Channel and Fresh:



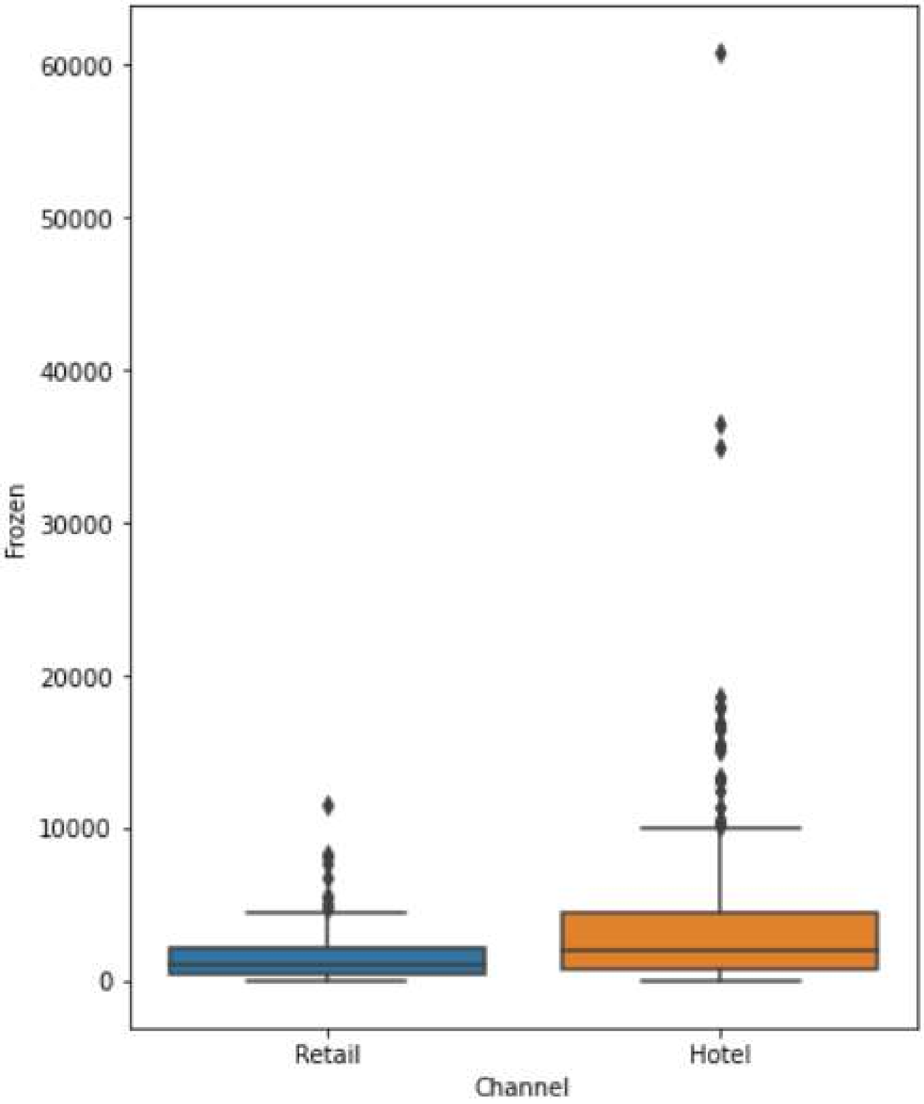
Boxplot across Channel and Milk:



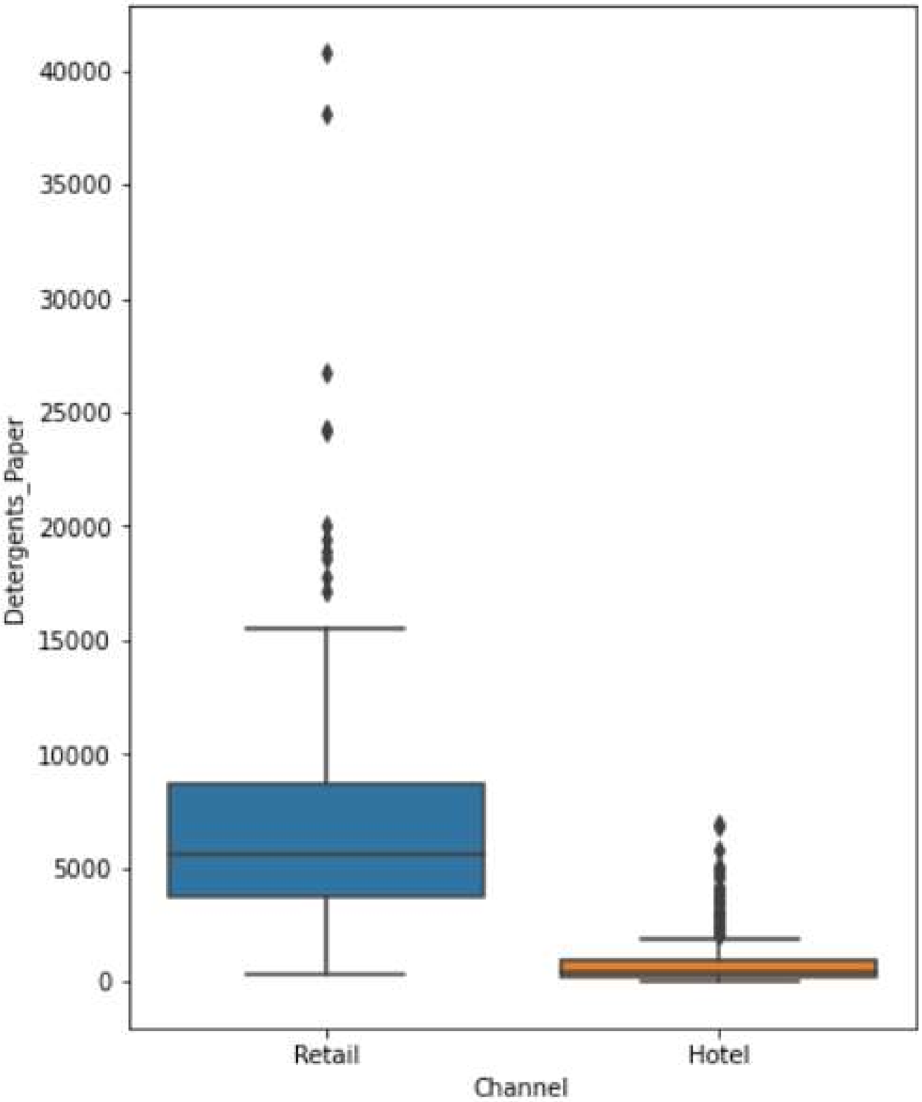
Boxplot across Channel and Grocery:



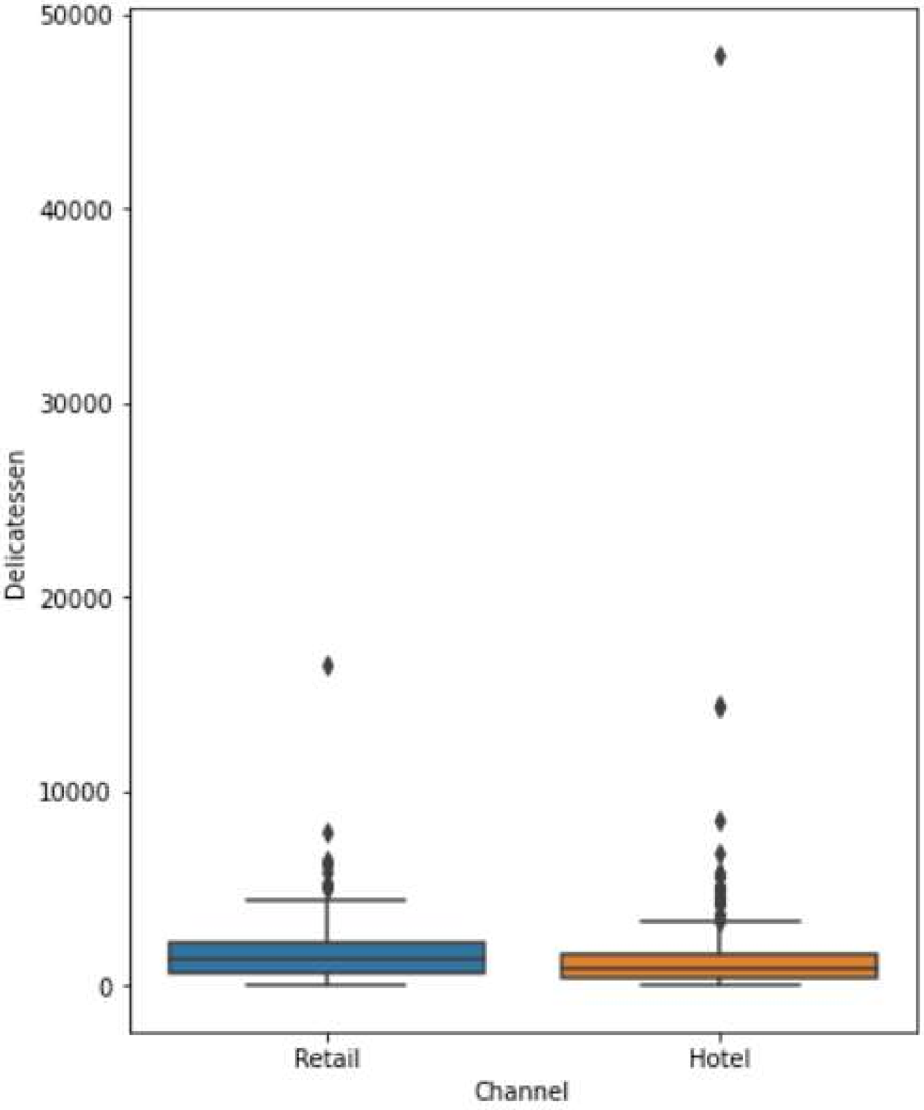
<u>Boxplot across Channel and Frozen:</u>



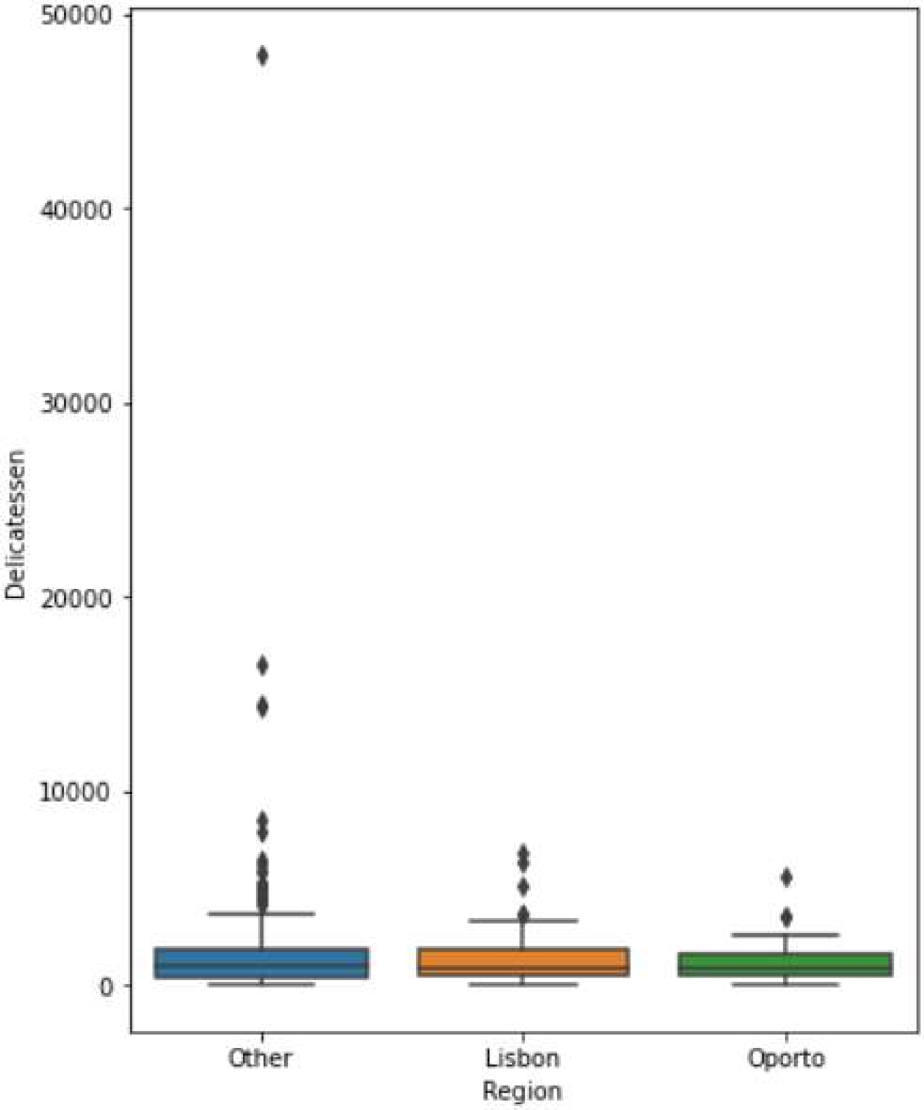
Boxplot across Channel and Detergents_Paper:



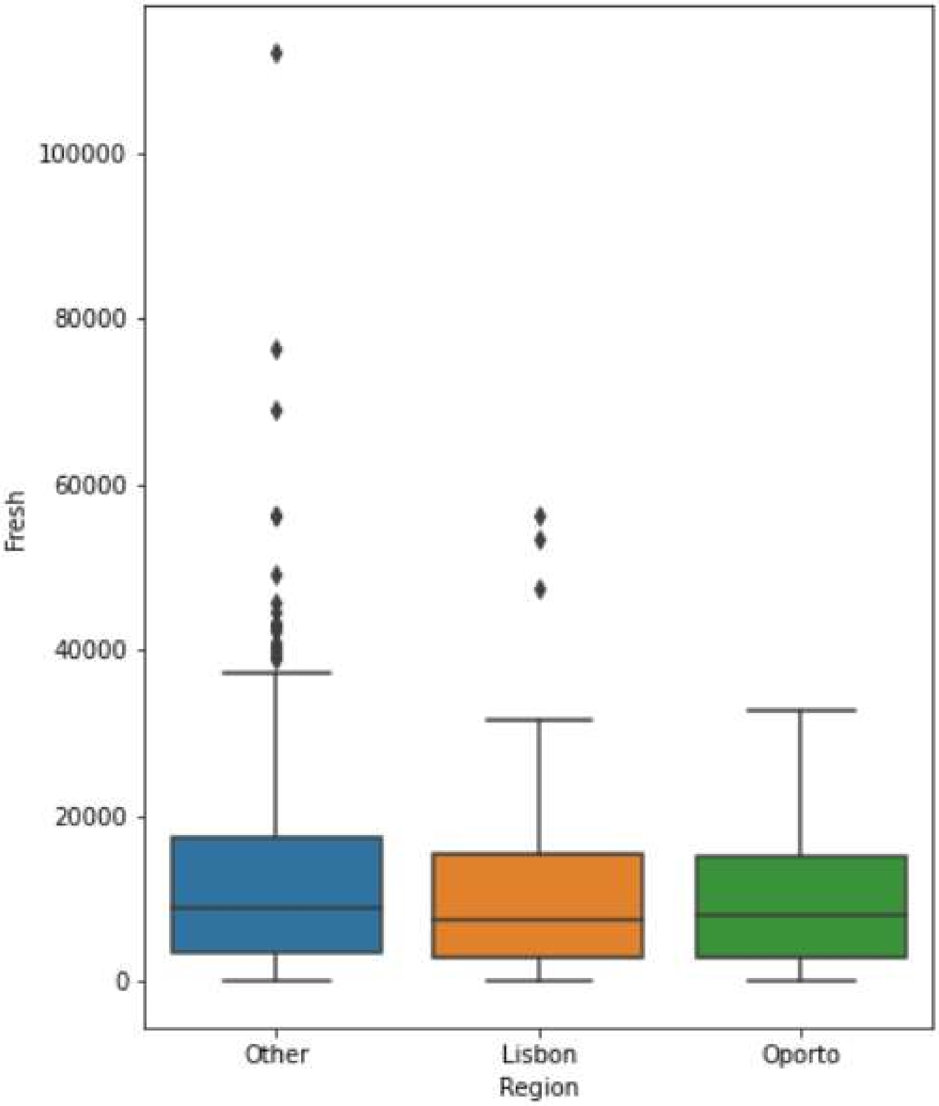
Boxplot across Channel and Delicatessen:



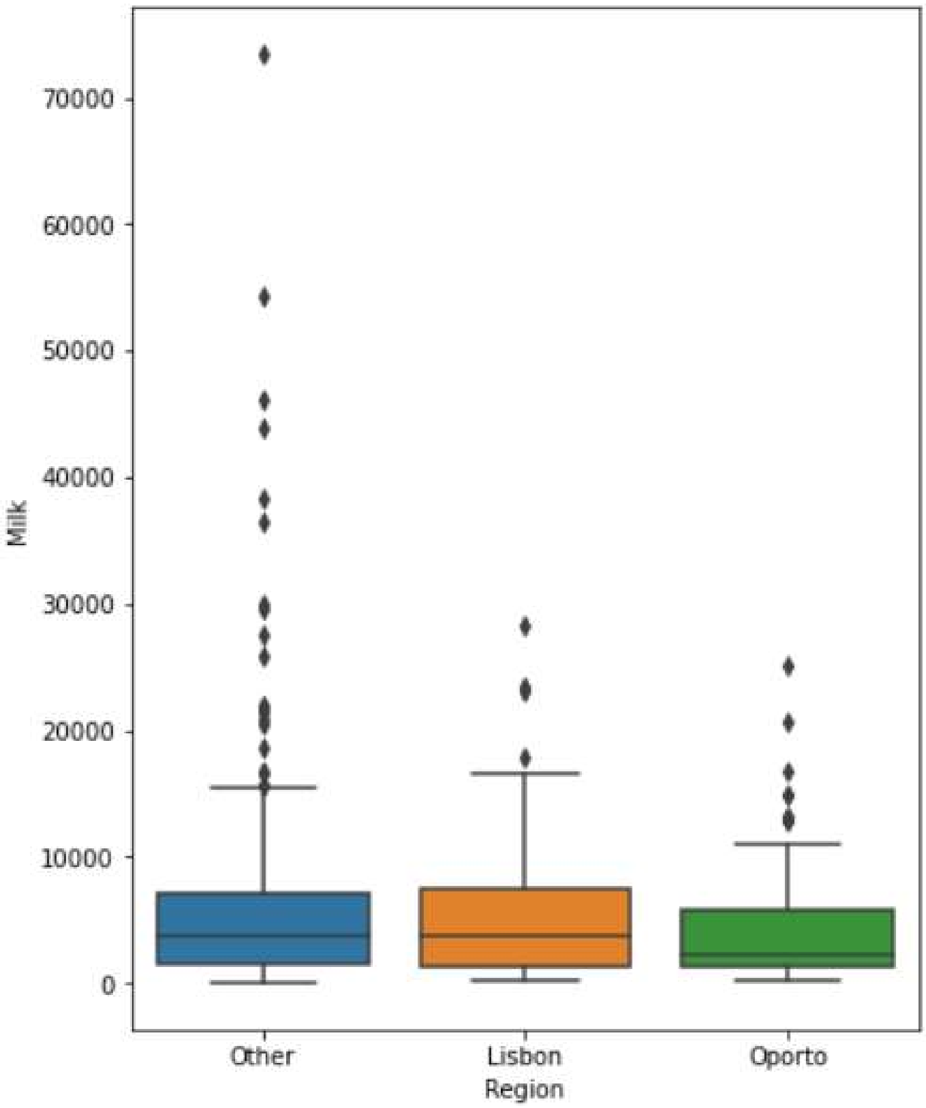
Boxplot across Region and Delicatessen:



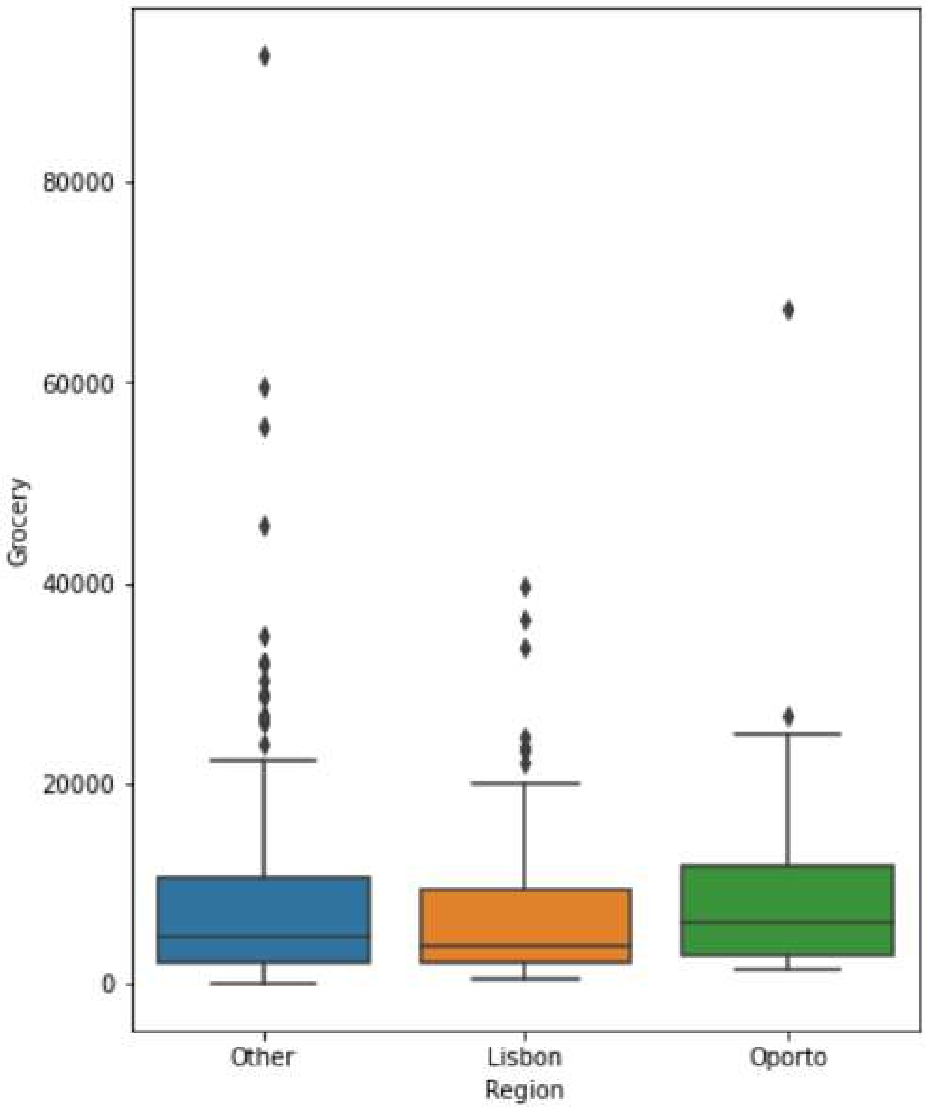
Boxplot across Region and Fresh:



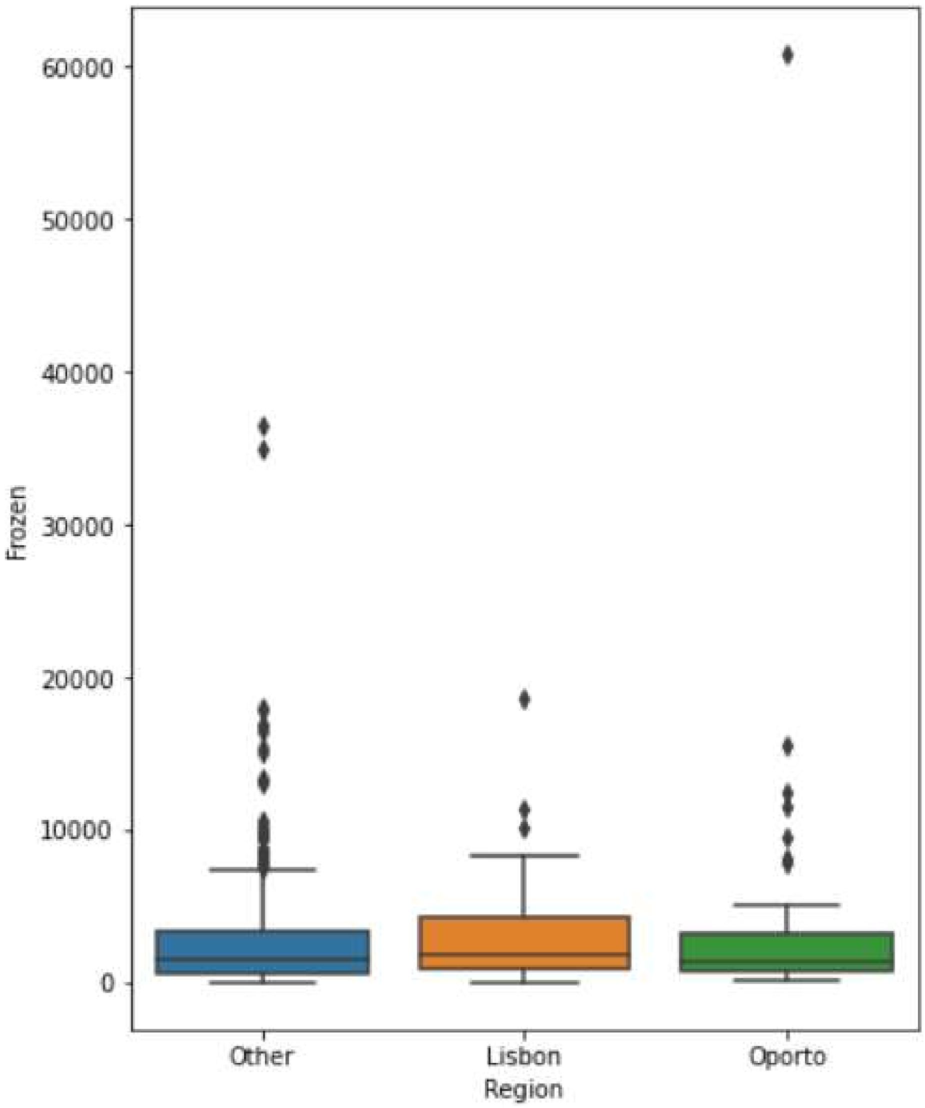
Boxplot across Region and Milk:



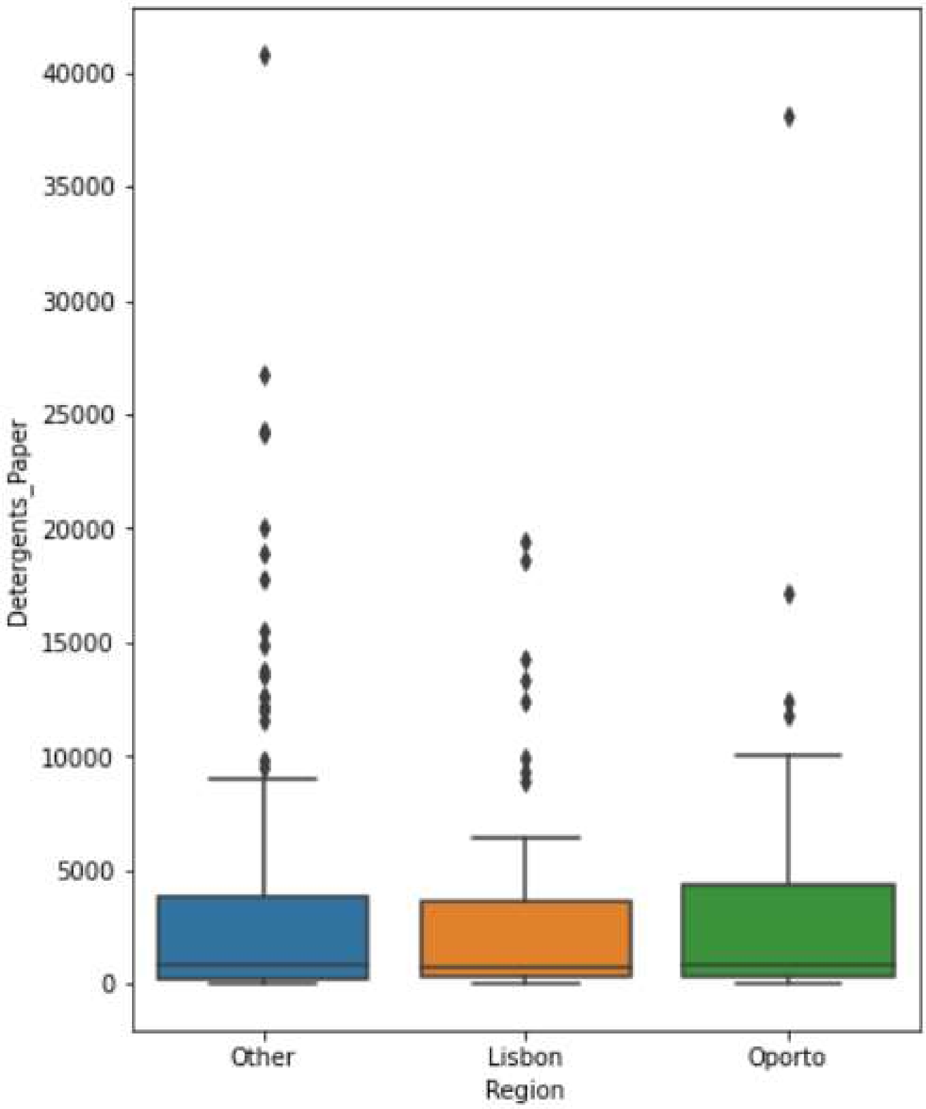
<u>Boxplot across Region and Grocery:</u>



Boxplot across Region and Frozen:



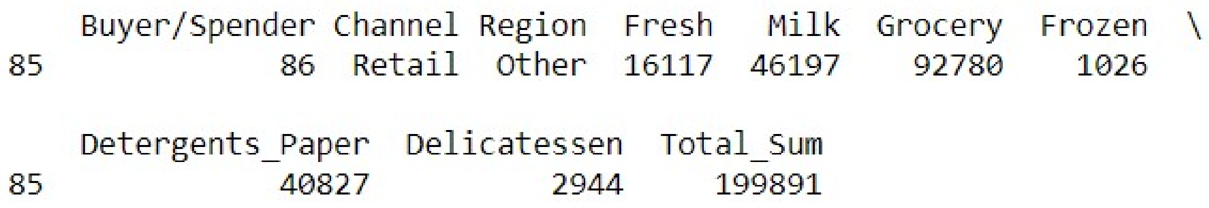
Boxplot across Region and Detergents_Paper:



Conclusion: As we Can see, there are many outliers in the data for each product when plotted against Channel and Region

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

Using the max function in the Total_Sum column,



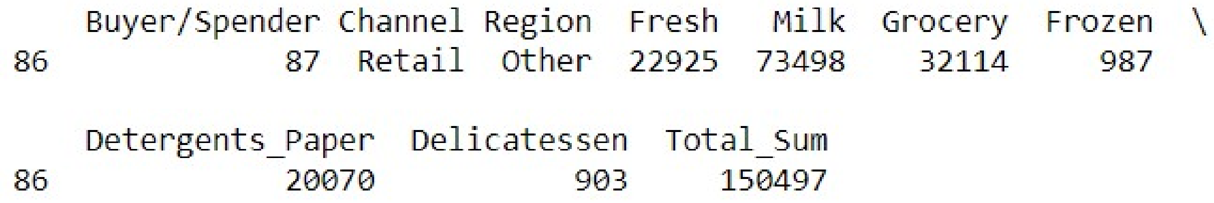
```
    Buyer/Spender Channel Region  Fresh   Milk  Grocery  Frozen  \
85             86  Retail  Other  16117  46197    92780    1026   

    Detergents_Paper  Delicatessen  Total_Sum  
85             40827          2944     199891  
```

The Buyer/Spender: 86 has spent the maximum with the majority towards the product: Fresh.

Since the data has a lot of outliers, it is wise to consider the median or the 50th percentile rather than the mean to compute the average.

Using the max function for the product Milk column:

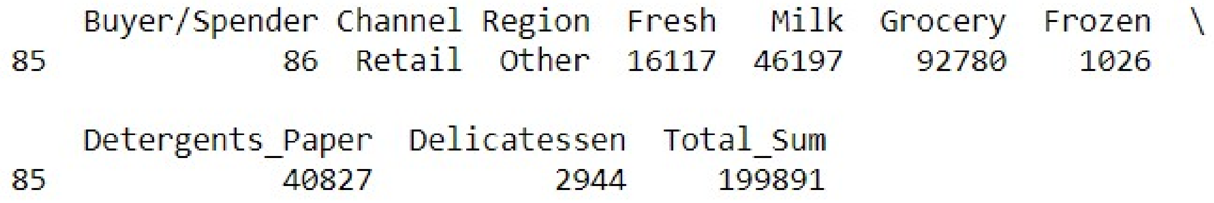


```
    Buyer/Spender Channel Region  Fresh   Milk  Grocery  Frozen  \
86             87  Retail  Other  22925  73498    32114     987   

    Detergents_Paper  Delicatessen  Total_Sum  
86             20070           903     150497  
```

The Buyer/Spender # 87 has spent the maximum on Milk with 73498.

Using the max function on the product Grocery column:

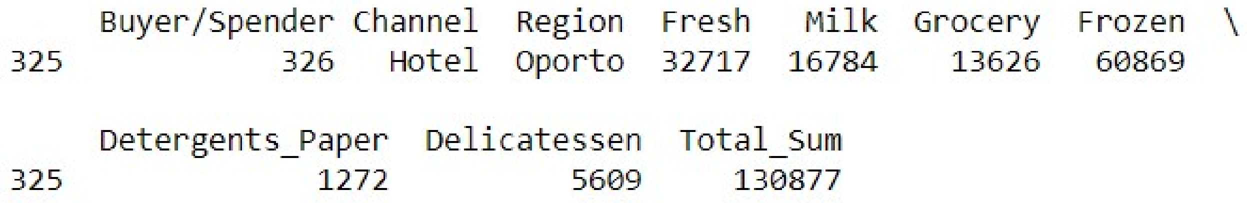


```
    Buyer/Spender Channel Region  Fresh   Milk  Grocery  Frozen  \
85             86  Retail  Other  16117  46197    92780    1026   

    Detergents_Paper  Delicatessen  Total_Sum  
85             40827          2944     199891  
```

The Buyer/Spender # 86 has spent the maximum on Grocery with 92780.

Using the max function on the product Frozen column:

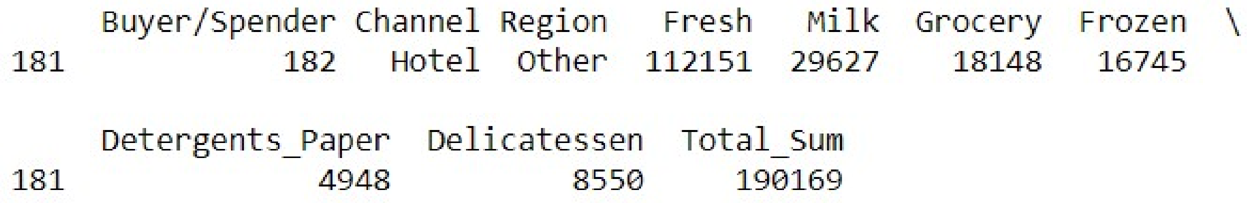


```
     Buyer/Spender Channel   Region  Fresh   Milk  Grocery  Frozen  \
325            326   Hotel  Oporto  32717  16784    13626   60869   

     Detergents_Paper  Delicatessen  Total_Sum  
325              1272          5609     130877  
```

The Buyer / Spender # 326 has spent the maximum on Frozen with 60869.

Using the max function on the product Fresh column:

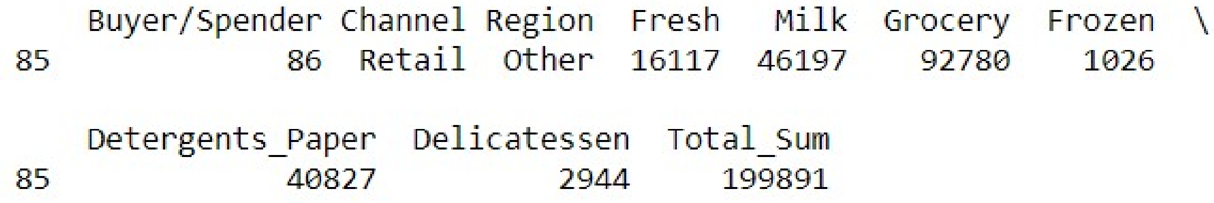


```
     Buyer/Spender Channel Region   Fresh   Milk  Grocery  Frozen  \
181            182   Hotel  Other  112151  29627    18148   16745   

     Detergents_Paper  Delicatessen  Total_Sum  
181              4948          8550     190169  
```

The Buyer/Spender # 182 has spent the maximum on Fresh with 112151.

Using the max function on the product Detergents_Paper:

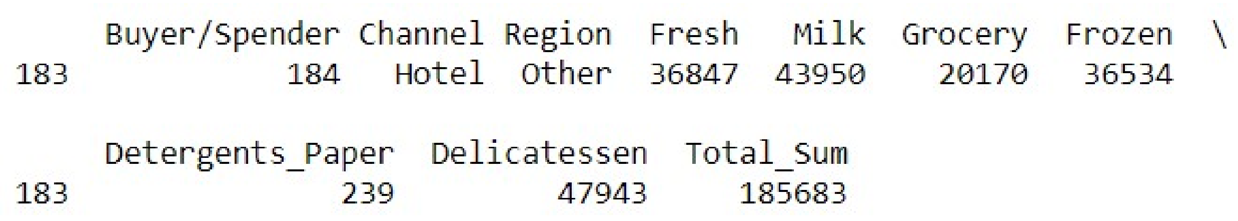


```
    Buyer/Spender Channel Region  Fresh   Milk  Grocery  Frozen  \
85             86  Retail  Other  16117  46197    92780    1026   

    Detergents_Paper  Delicatessen  Total_Sum  
85             40827          2944     199891  
```

The Buyer/ Spender # 86 has spent the maximum on Detergents_Paper with 40827.

Using the max function on the product Delicatessen column:



```
     Buyer/Spender Channel Region  Fresh   Milk  Grocery  Frozen  \
183            184   Hotel  Other  36847  43950    20170   36534   

     Detergents_Paper  Delicatessen  Total_Sum  
183               239         47943     185683  
```

The Buyer / Spender # 184 has spent the maximum on Delicatessen with 47943.

Conclusion: The Buyer / Spender # 184 has spent the maximum on Delicatessen with 47943.

More focus should be placed on the Buyer / Spender#: 86, 87, 326, 182, 184 as they have purchased more.

More discounts/offers need to be provided to them so that we can drive them to purchase more.