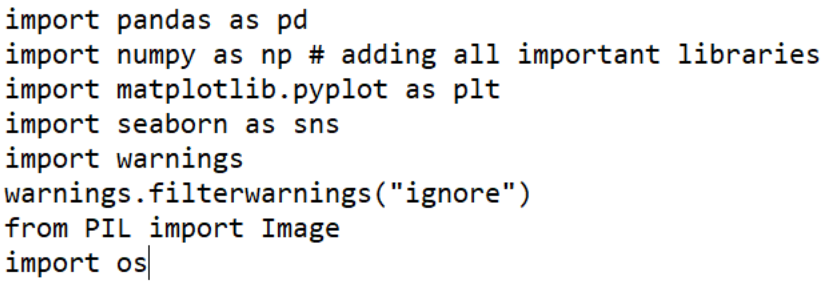
# BUISNESS REPORT

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. You have been given a dataset of 440 distribution outlets of a wholesale store in Boston.

QUESTION 1:

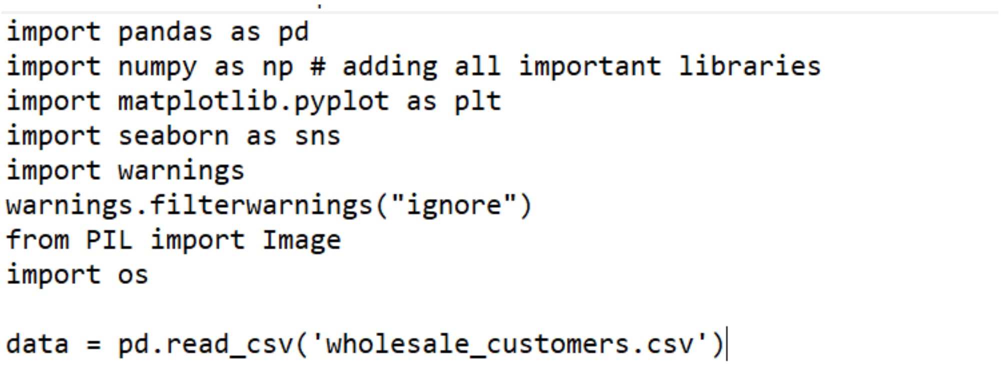
From the question we must import all the libraries required for us to analyse the data like pandas, seaborn, numpy, matplotlib.



```
import pandas as pd
import numpy as np # adding all important libraries
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
from PIL import Image
import os
```

QUESTION 2:

We can add the below code to get sample data.



```
import pandas as pd
import numpy as np # adding all important libraries
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
from PIL import Image
import os

data = pd.read_csv('wholesale_customers.csv')
```

| 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 |

QUESTION 3:

To find the shape we can use data.shape() we get rows and columns as (440, 9)

QUESTION 4:

To find the number of missing values we use data.isnull().sum() for which get a table that shows:

```
Buyer/Spender      0
Channel            0
Region             0
Fresh              0
Milk               0
Grocery            0
Frozen             0
Detergents_Paper   0
Delicatessen       0
```

From the above table we get that all the variables have no missing values.

QUESTION 5:

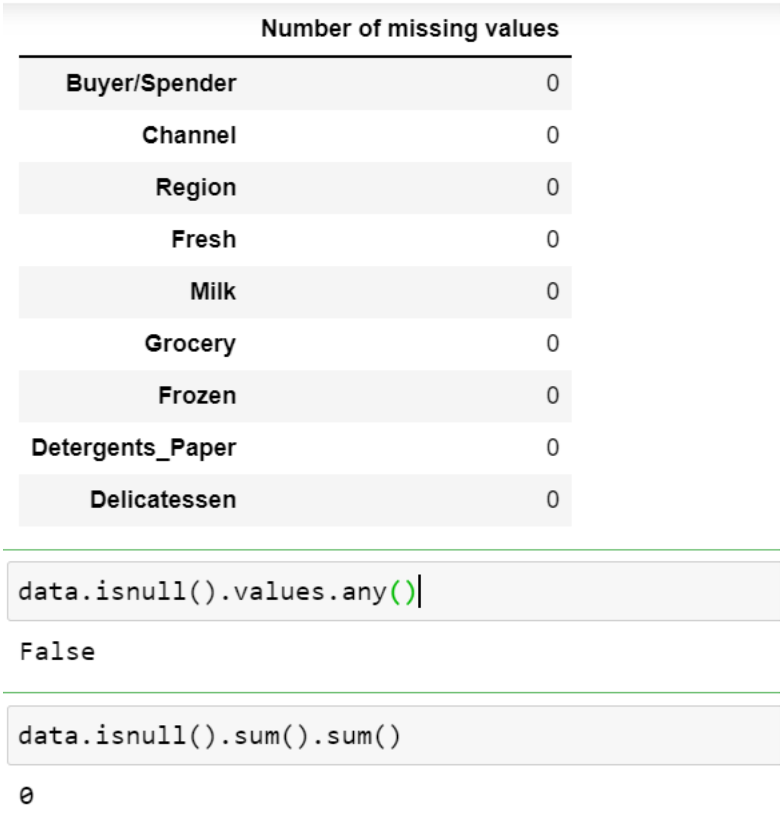
We have to find duplicates for that we use

```
dupes = data.duplicated()
sum(dupes)
```

This gives us 0 which means there are no duplicate variables

QUESTION 6:

Here we need to find any missing values in case we must replace them with:
pd.DataFrame(data.isnull().sum(), columns=['Number of missing values'])



| | Number of missing values |
|---|---|
| Buyer/Spender | 0 |
| Channel | 0 |
| Region | 0 |
| Fresh | 0 |
| Milk | 0 |
| Grocery | 0 |
| Frozen | 0 |
| Detergents_Paper | 0 |
| Delicatessen | 0 |

```
data.isnull().values.any()
```

```
False
```

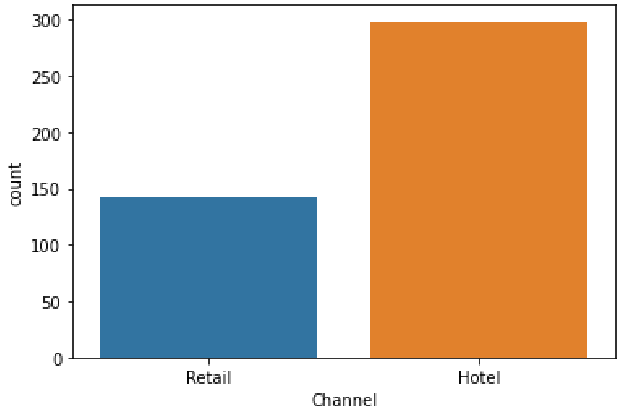
```
data.isnull().sum().sum()
```

```
0
```

# Univariate Analysis:

QUESTION 7:

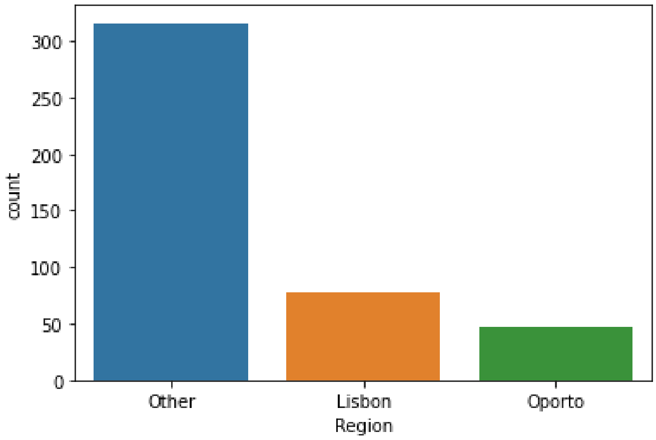
a.) ax = sns.countplot(x='Channel', data=data)



We use the sns plot where we plot channel against the customer count.

There are two categories of customers in channel that is hotel and Retail. This count plot shows the number of customers for both categorical variables. The customer count is highest for Hotel.

ax = sns.countplot(x='Region', data=data)

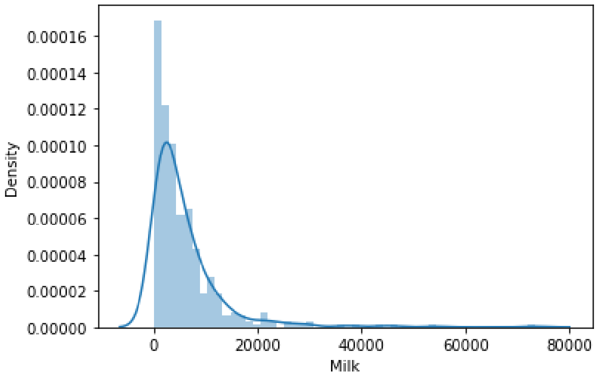


We use sns plot  where plot region against customer count.

There are three geographical region that are involved. From the above plot we see the customer distribution area wise in Portugal. The highest count is in all the other regions in Portugal, 2nd highest in Lisbon and the least in Oporto.

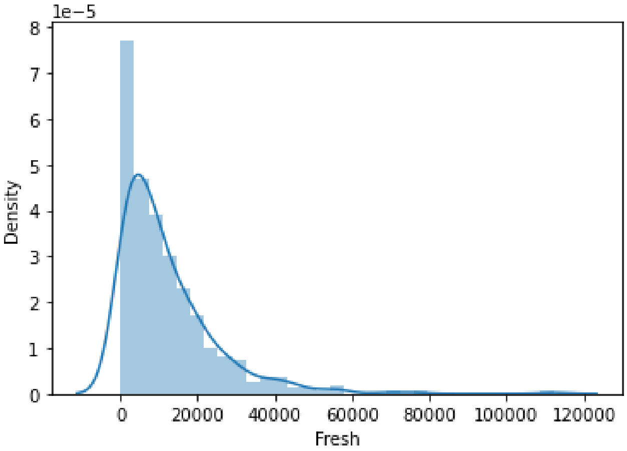
b.) Plot distribution of each continuous variable

- sns.distplot(data['Milk'])



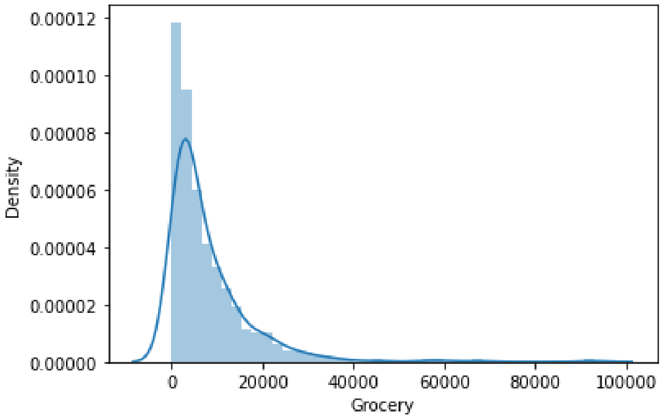
This is a distribution plot for milk(continuous Variable).This positively skewed and left tailed. This plot shows that the left side has large number of data values is clustered and lower data on the right side. The large spends amount on milk id between 0 to 20000.

- sns.distplot(data['Fresh'])



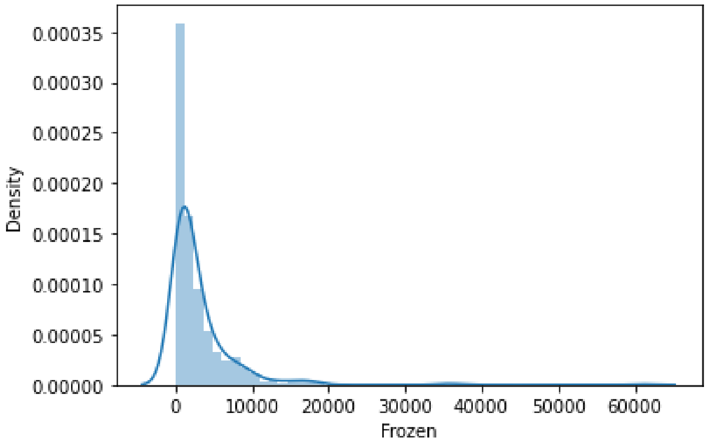
This is a distribution plot for fresh products ordered. This plot is positively skewed. The amount of fresh products ordered are between 0 to almost 60000.Maximum data is clustered in this region.

- sns.distplot(data['Grocery'])



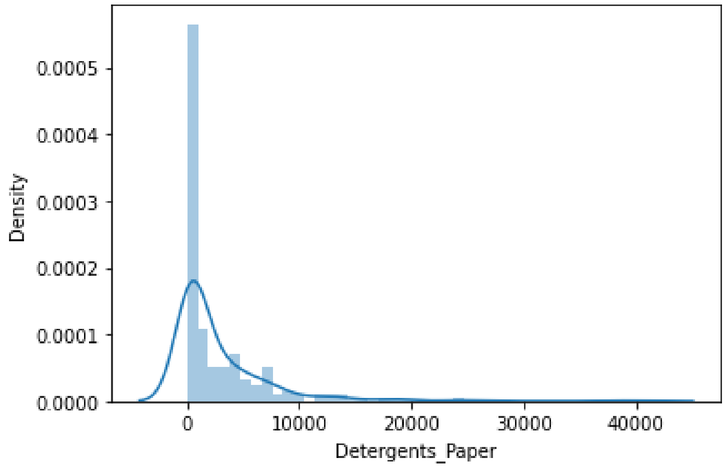
This a distribution plot for Grocery. This distribution shows the amount spent on the grocery which is positively skewed. The count of customer might not be very large but a lot people seem to buy within the range of 40000.

- sns.distplot(data['Frozen'])



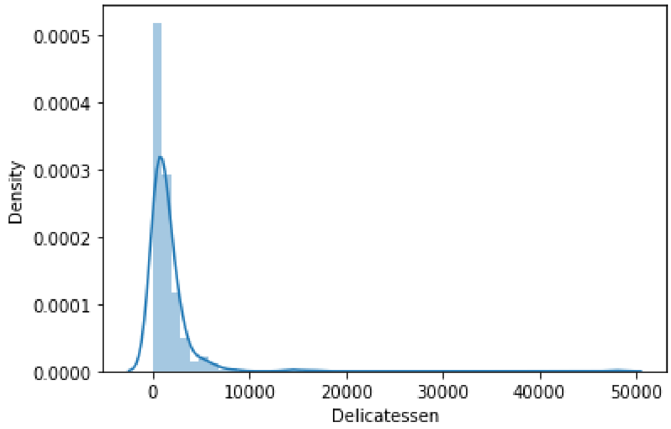
This is a distribution plot for frozen products. From the above plot we can see that not many customer are going for frozen items. This plot is positively skewed and maximum customers are clustered between 0 to 20000.

- sns.distplot(data['Detergents_Paper'])



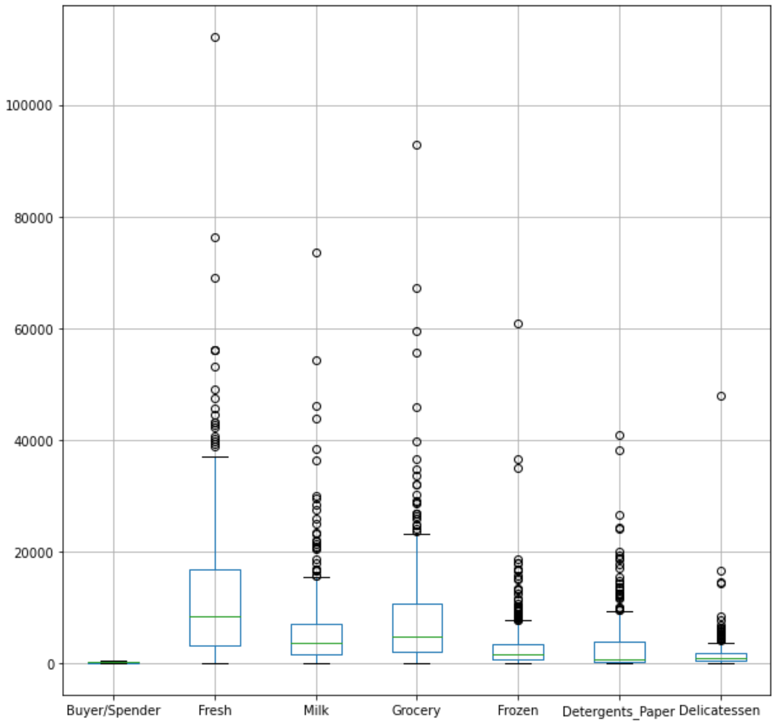
This is a positively skewed plot and from the graph we infer that very few customer spend on detergent-paper and distribution of these customers are between 0 to 10000.

- sns.distplot(data['Delicatessen'])



The customer count is high from the above graph, which means a lot of customer buy these products. This plot is positively skewed and left tailed. The amount spent on these items are between 0 to 10000.

c.) To create a boxplot:
data.boxplot(figsize=(10,10))

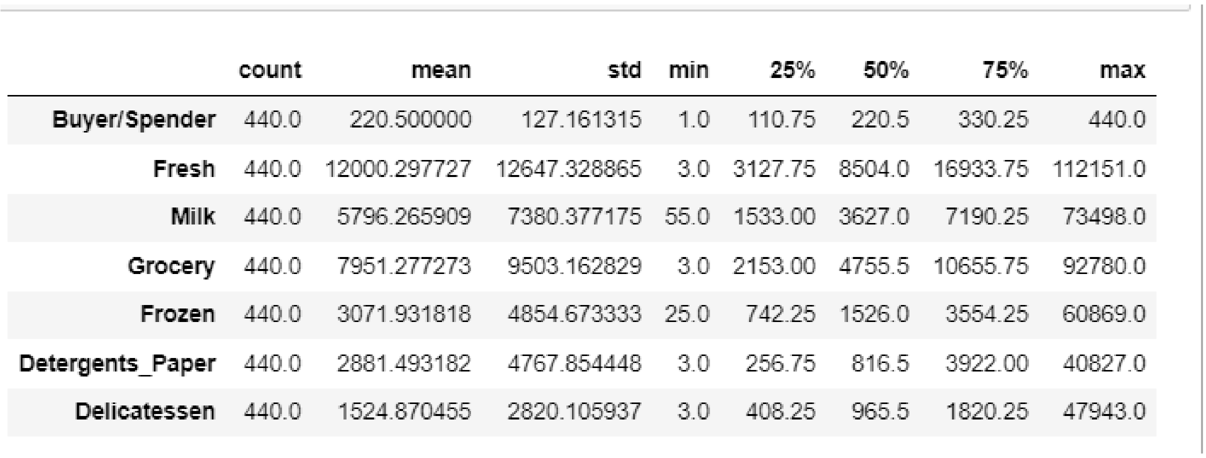


From the above box plot we get to understand that which variable is more inconsistent. The maximum outliers the more inconsistent the values are.

d.) Descriptive data gives us the mean, median, standard deviation:

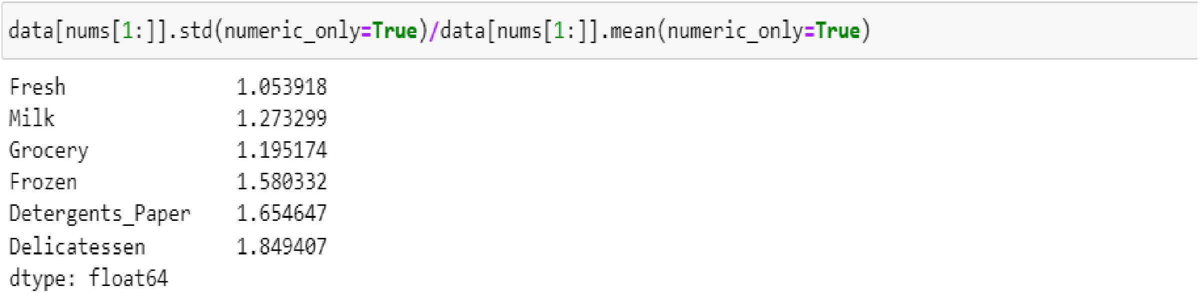
data.describe().T

-



| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Buyer/Spender** | 440.0 | 220.500000 | 127.161315 | 1.0 | 110.75 | 220.5 | 330.25 | 440.0 |
| **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 |

e.) Finding the coefficient of covariance:



```
data[nums[1:]].std(numeric_only=True)/data[nums[1:]].mean(numeric_only=True)
```

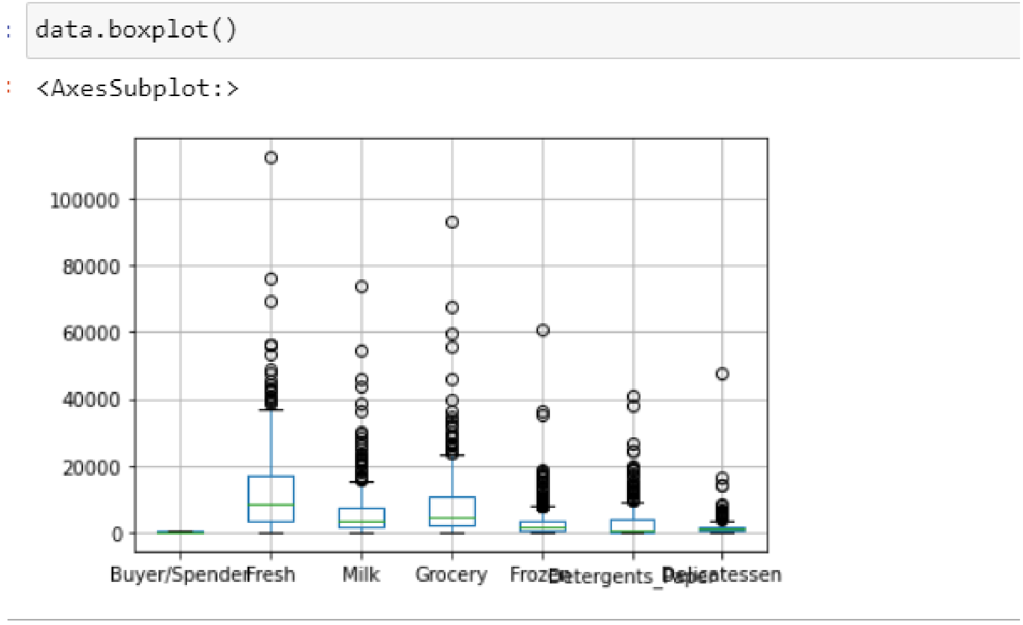
```
Fresh               1.053918
Milk                1.273299
Grocery             1.195174
Frozen              1.580332
Detergents_Paper    1.654647
Delicatessen        1.849407
dtype: float64
```

From the above covariance data, we can see that the data that has highest value has the highest variation and the lowest has the lowest variation. We see that delicatessen has highest variation(inconsistent) and fresh product has the least(consistent).

# Outlier Treatment & New Feature Creation:

QUESTION 8:

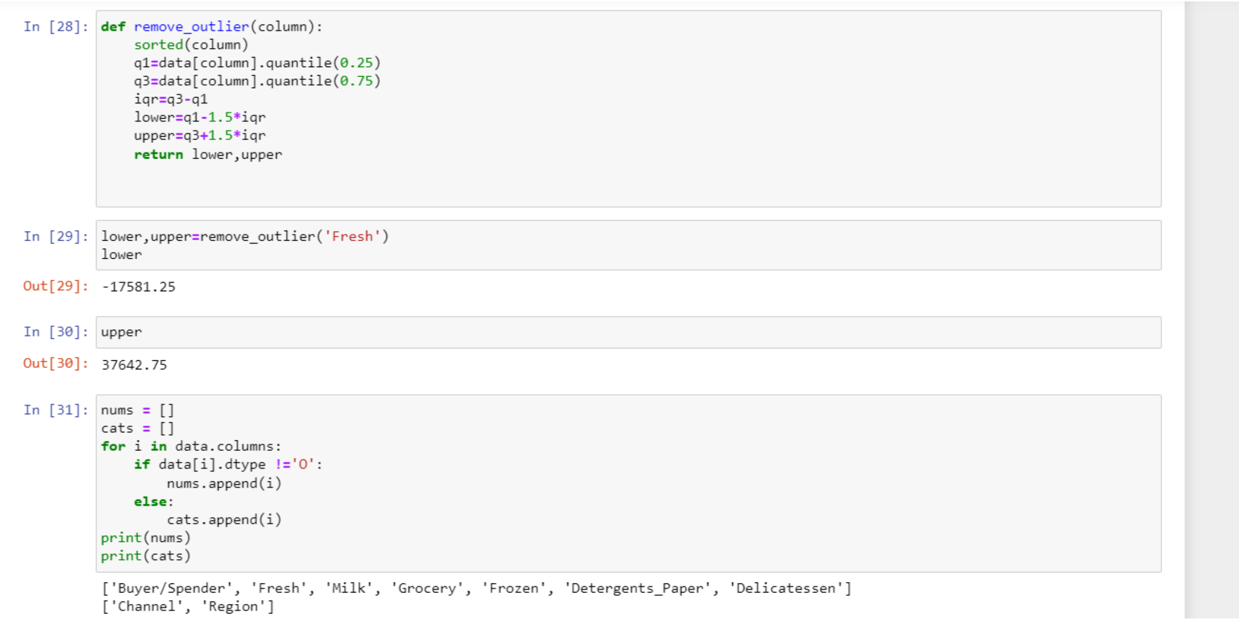
a.) There are outliers in the data.



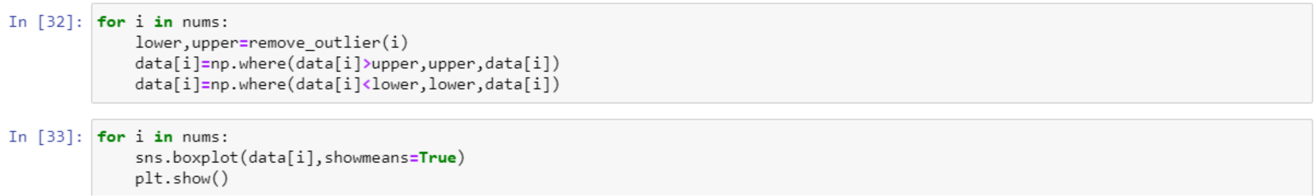
```
: data.boxplot()
```

```
: <AxesSubplot:>
```

b.) To treat the outlier data:



```
In [28]: def remove_outlier(column):
             sorted(column)
             q1=data[column].quantile(0.25)
             q3=data[column].quantile(0.75)
             iqr=q3-q1
             lower=q1-1.5*iqr
             upper=q3+1.5*iqr
             return lower,upper

In [29]: lower,upper=remove_outlier('Fresh')
         lower

Out[29]: -17581.25

In [30]: upper

Out[30]: 37642.75

In [31]: nums = []
         cats = []
         for i in data.columns:
             if data[i].dtype !='O':
                 nums.append(i)
             else:
                 cats.append(i)
         print(nums)
         print(cats)

         ['Buyer/Spender', 'Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents_Paper', 'Delicatessen']
         ['Channel', 'Region']
```



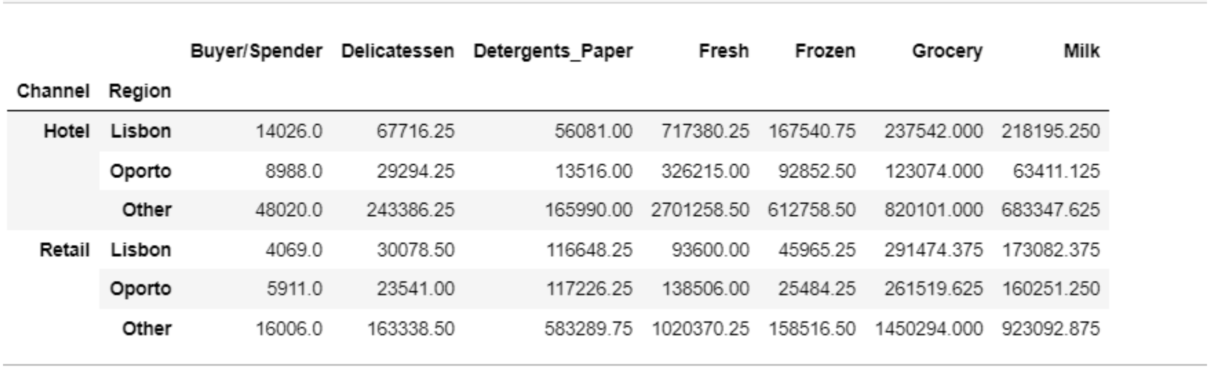
```
In [32]: for i in nums:
             lower,upper=remove_outlier(i)
             data[i]=np.where(data[i]>upper,upper,data[i])
             data[i]=np.where(data[i]<lower,lower,data[i])

In [33]: for i in nums:
             sns.boxplot(data[i],showmeans=True)
             plt.show()
```

This code gives us the result for each plot for all 6 variables and remove any outliers/ inconsistency. Here we create a separate definition and make list for both categorical and numerical list.From this we then create a box plot to get the final result .

QUESTION 9:

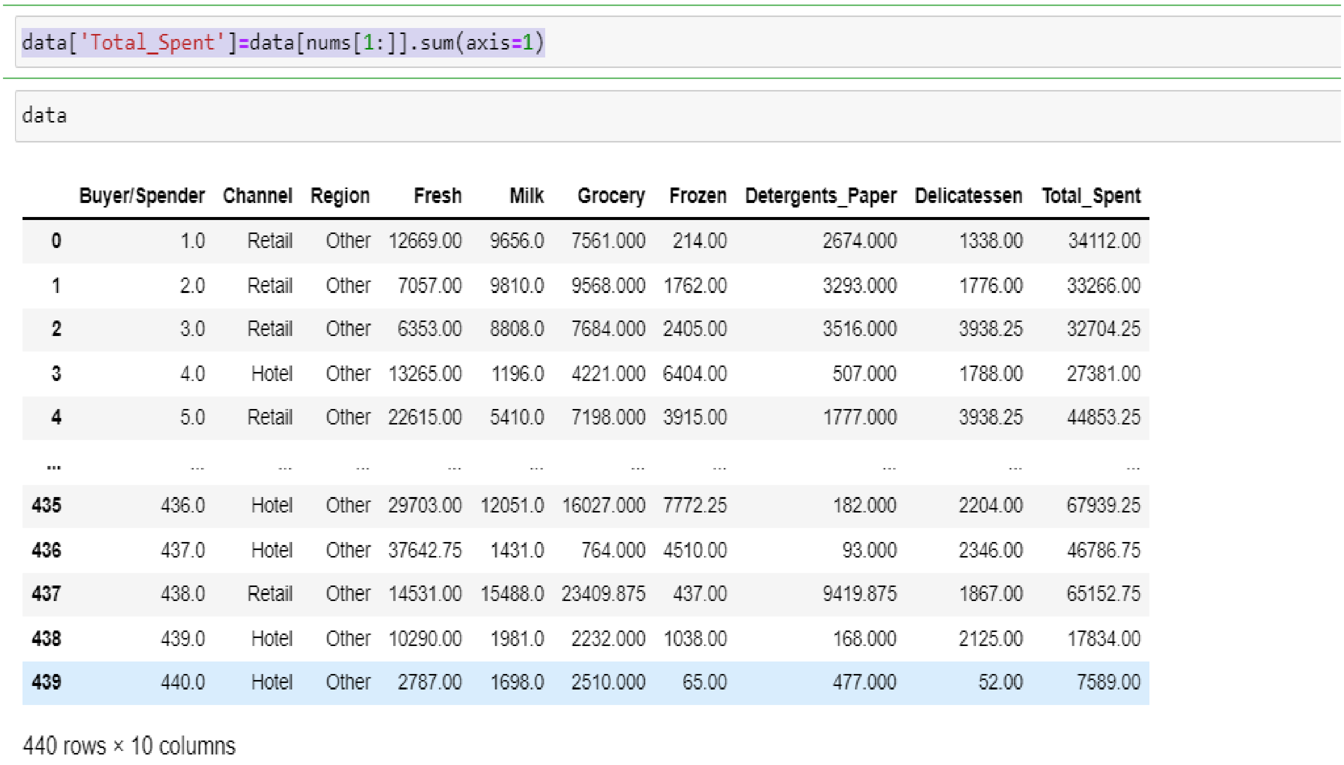
a.) We create pivot table to just understand aggregate sum and normal sums of all the variables.

pd.pivot_table(data, index=['Channel', 'Region'], aggfunc=np.sum)



| | | Buyer/Spender | Delicatessen | Detergents_Paper | Fresh | Frozen | Grocery | Milk |
|---|---|---|---|---|---|---|---|---|
| Channel | Region | | | | | | | |
| Hotel | Lisbon | 14026.0 | 67716.25 | 56081.00 | 717380.25 | 167540.75 | 237542.000 | 218195.250 |
| | Oporto | 8988.0 | 29294.25 | 13516.00 | 326215.00 | 92852.50 | 123074.000 | 63411.125 |
| | Other | 48020.0 | 243386.25 | 165990.00 | 2701258.50 | 612758.50 | 820101.000 | 683347.625 |
| Retail | Lisbon | 4069.0 | 30078.50 | 116648.25 | 93600.00 | 45965.25 | 291474.375 | 173082.375 |
| | Oporto | 5911.0 | 23541.00 | 117226.25 | 138506.00 | 25484.25 | 261519.625 | 160251.250 |
| | Other | 16006.0 | 163338.50 | 583289.75 | 1020370.25 | 158516.50 | 1450294.000 | 923092.875 |

Here we create a new feature named Total_spent as per the given question: We get the aggregate sum all the addition of the spent items/ values.



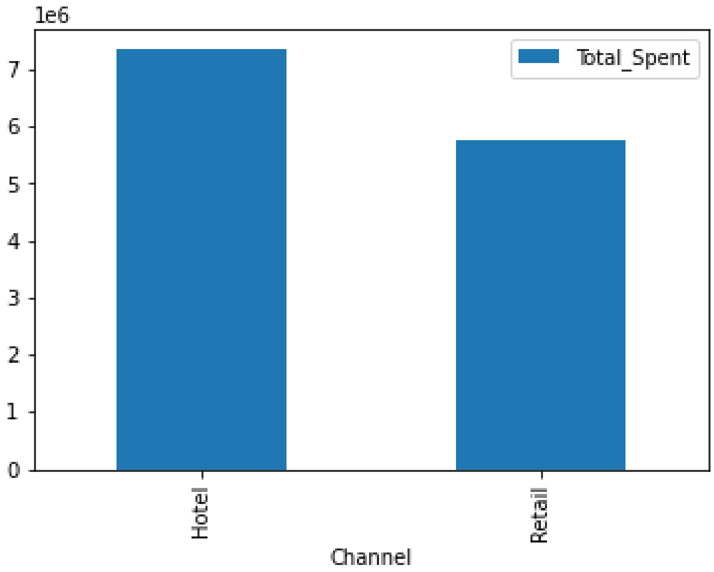
```
data['Total_Spent']=data[nums[1:]].sum(axis=1)
```

```
data
```

| | Buyer/Spender | Channel | Region | Fresh | Milk | Grocery | Frozen | Detergents_Paper | Delicatessen | Total_Spent |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1.0 | Retail | Other | 12669.00 | 9656.0 | 7561.000 | 214.00 | 2674.000 | 1338.00 | 34112.00 |
| **1** | 2.0 | Retail | Other | 7057.00 | 9810.0 | 9568.000 | 1762.00 | 3293.000 | 1776.00 | 33266.00 |
| **2** | 3.0 | Retail | Other | 6353.00 | 8808.0 | 7684.000 | 2405.00 | 3516.000 | 3938.25 | 32704.25 |
| **3** | 4.0 | Hotel | Other | 13265.00 | 1196.0 | 4221.000 | 6404.00 | 507.000 | 1788.00 | 27381.00 |
| **4** | 5.0 | Retail | Other | 22615.00 | 5410.0 | 7198.000 | 3915.00 | 1777.000 | 3938.25 | 44853.25 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **435** | 436.0 | Hotel | Other | 29703.00 | 12051.0 | 16027.000 | 7772.25 | 182.000 | 2204.00 | 67939.25 |
| **436** | 437.0 | Hotel | Other | 37642.75 | 1431.0 | 764.000 | 4510.00 | 93.000 | 2346.00 | 46786.75 |
| **437** | 438.0 | Retail | Other | 14531.00 | 15488.0 | 23409.875 | 437.00 | 9419.875 | 1867.00 | 65152.75 |
| **438** | 439.0 | Hotel | Other | 10290.00 | 1981.0 | 2232.000 | 1038.00 | 168.000 | 2125.00 | 17834.00 |
| **439** | 440.0 | Hotel | Other | 2787.00 | 1698.0 | 2510.000 | 65.00 | 477.000 | 52.00 | 7589.00 |

440 rows × 10 columns

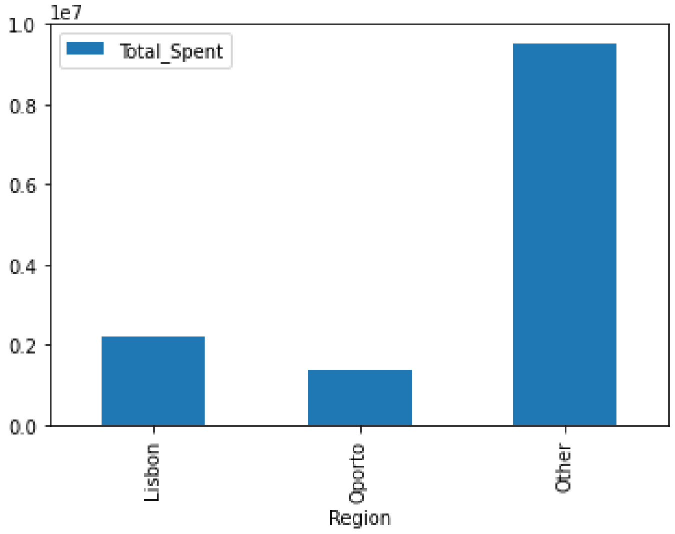
- pd.pivot_table(data, index=['Channel'], values=['Total Spent'],aggfunc=np.sum).plot. Bar()
- graph between spends and channel
- Total spends by customer in hotels than in the retail.



b.)

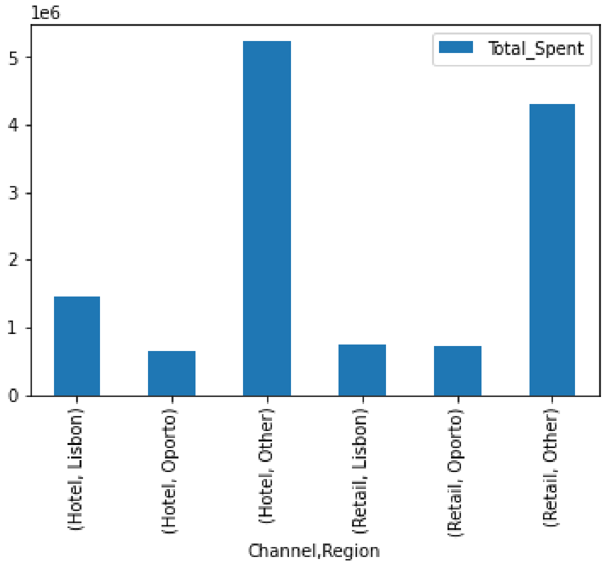
- Plot Spends against “Regions”.

- pd.pivot_table(data, index=['Region'], values=['Total_Spent'],aggfunc=np.sum).plot.bar()
- oporto region has less spent than and highest in other region



c.)

- Plot spends against with both "Channels" and "Regions" combined: pd.pivot_table(data, index=['Channel','Region'], values=['Total_Spent'],aggfunc=np.sum).plot.bar()

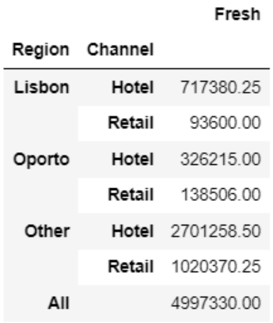


d.) The region and channel that spend more is Hotel in Other areas.
e.) The region and channel that spend the least is Hotel in oporto.

# BIVARIATE ANALYSIS:

a.) Pivot Table and find aggregate spends on fresh items:

```
pd.pivot_table(data,'Fresh',index=['Region','Channel'],margins=True,aggfunc=np.sum)
```

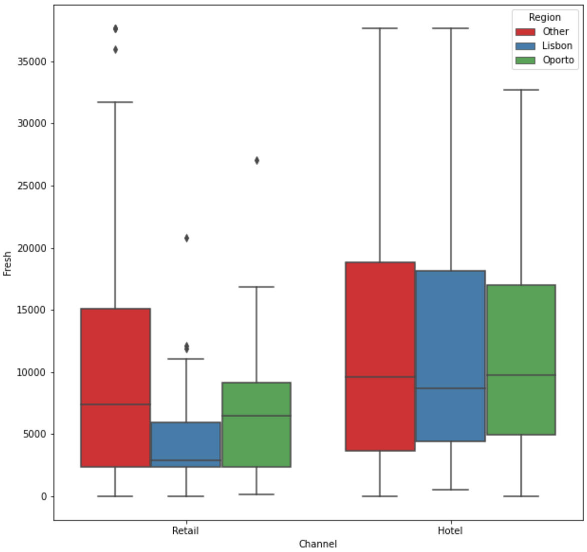


| | | Fresh |
|---|---|---|
| **Region** | **Channel** | |
| **Lisbon** | **Hotel** | 717380.25 |
| | **Retail** | 93600.00 |
| **Oporto** | **Hotel** | 326215.00 |
| | **Retail** | 138506.00 |
| **Other** | **Hotel** | 2701258.50 |
| | **Retail** | 1020370.25 |
| **All** | | 4997330.00 |

Gives us all the values for Hotel, retail for every region and gives us an insight of how much was spent on each region.

b.) Plot the spends on fresh products against channels and regions by colour:

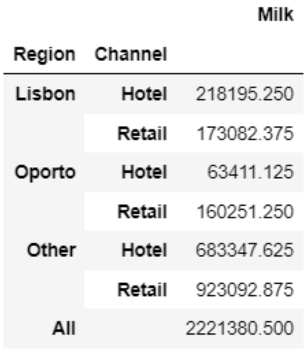
```
plt.figure(figsize=(10,10))
sns.boxplot(x='Channel',y='Fresh',hue='Region',data=data,palette='Set1')
plt.show()
```



- The Retail and hotel are right skewed and
- From the above IQR we infer that Hotel has spent the least than retail.
- In hotel its other region has more spent amount than the rest and the least spent here is Oporto.
- In Retail it’s again the Other region that spend the most as compared to the rest. The least is made by Lisbon.
- There Outliers for Retail which means values can be inconsistent for other and Lisbon.
- So overall we see that for fresh product is much more sold in Retail stores.

c.) Pivot Tables and find aggregate spends on Milk items:

```
pd.pivot_table(data,'Milk',index=['Region','Channel'],margins=True,aggfunc=np.sum)
```

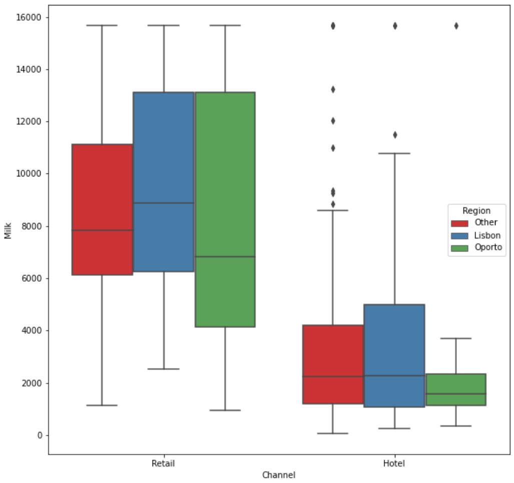


|  |  | Milk |
|---|---|---|
| Region | Channel |  |
| Lisbon | Hotel | 218195.250 |
|  | Retail | 173082.375 |
| Oporto | Hotel | 63411.125 |
|  | Retail | 160251.250 |
| Other | Hotel | 683347.625 |
|  | Retail | 923092.875 |
| All |  | 2221380.500 |

Gives us all the values for Hotel, retail for every region and gives us an insight of how much was spent on each region.

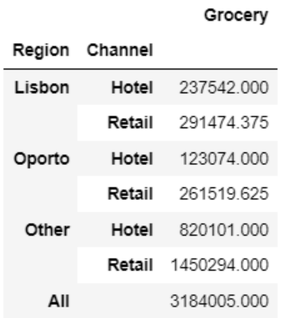
d.) Plot the spends on fresh products against channels and regions by colour:

```
plt.figure(figsize=(10,10))
sns.boxplot(x='Channel',y='Milk',hue='Region',data=data,palette='Set1')
plt.show()
```



- From the above plot we see that more is spent in retail than hotel.
- From the retail IQR as well we see more customers from Oporto have spent the most and least is from other areas.
- From the hotel IQR we see that customer have spent less and among this Lisbon has spent the most in the hotel and least is oporto.
- There are outliers for Hotel where spend is inconsistent for other areas and Lisbon.
- Again we see that Retail sells for milk than in Hotels.

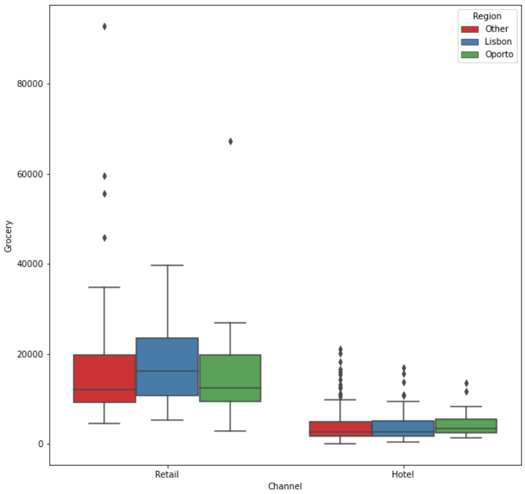
e.) Pivot Tables and find aggregate spends on Grocery items:
pd.pivot_table(data,'Grocery',index=['Region','Channel'],margins=True,aggfunc=np.sum)



| Region | Channel | Grocery |
|---|---|---|
| Lisbon | Hotel | 237542.000 |
| | Retail | 291474.375 |
| Oporto | Hotel | 123074.000 |
| | Retail | 261519.625 |
| Other | Hotel | 820101.000 |
| | Retail | 1450294.000 |
| All | | 3184005.000 |

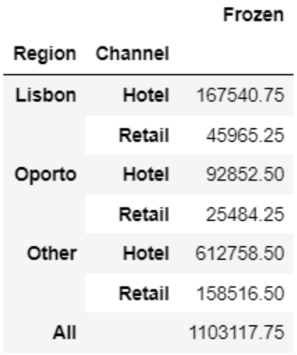
Gives us all the values for Hotel, retail for every region and gives us an insight of how much was spent on each region.

f.) Plot Spends on Groceries against Channels and Regions by colour:
plt.figure(figsize=(10,10))
sns.boxplot(x='Channel',y='Grocery',hue='Region',data=data,palette='Set1')
plt.show()



- Both are right skewed.
- From the above IQR we infer that neither of the places spend a lot on Groceries in general, but Retail revenues are more.
- We see that Lisbon spends more and has no outliers than rest.
- Hotel barely makes any income in regard to retail. But from here we must also take that all of the values are inconsistent due to outliers.

g.) Use Pivot Tables and find aggregate spends on Frozen items:
pd.pivot_table(data,'Frozen',index=['Region','Channel'],margins=True,aggfunc=np.sum)



| | | Frozen |
|---|---|---|
| Region | Channel | |
| Lisbon | Hotel | 167540.75 |
| | Retail | 45965.25 |
| Oporto | Hotel | 92852.50 |
| | Retail | 25484.25 |
| Other | Hotel | 612758.50 |
| | Retail | 158516.50 |
| All | | 1103117.75 |

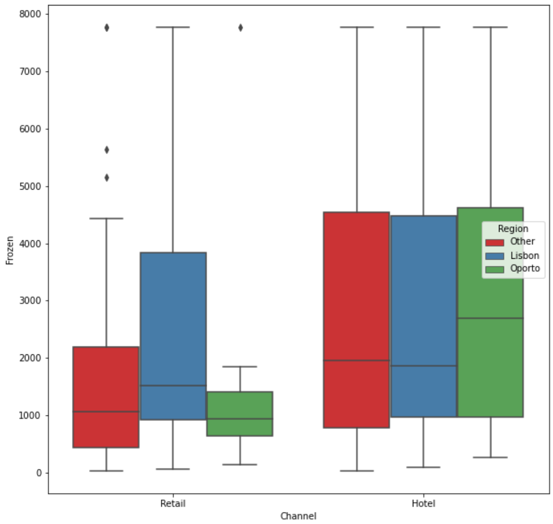
Gives us all the values for Hotel, retail for every region and gives us an insight of how much was spent on each region.

h.) Plot Spends on Frozen against Channels and Regions by colour

plt.figure(figsize=(10,10))

sns.boxplot(x='Channel',y='Frozen',hue='Region',data=data,palette='Set1')

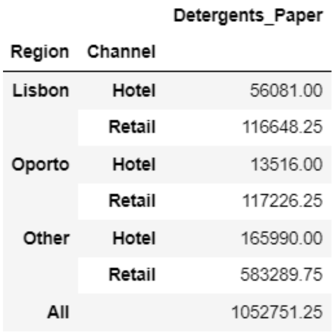
plt.show()



- From the above IQR we see Hotel is positively skewed.
- The more spread is Hotel than Retail and the maximum is oporto.
- There are Outliers for retail in region other and oporto hence values as inconsistent.
- Hotel customer spend the more than retail and skew is symmetric.

i.) Use Pivot Tables and find aggregate spends on Detergent Paper:

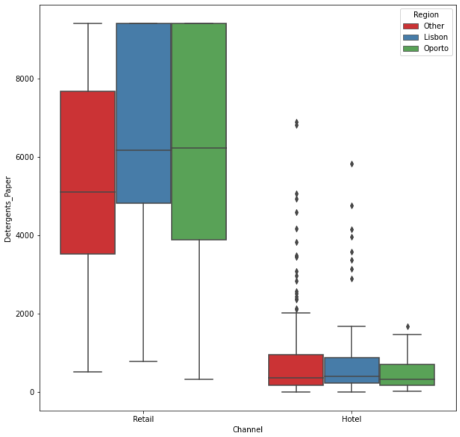
```
pd.pivot_table(data,'Detergents_Paper',index=['Region','Channel'],margins=True,aggfunc=np.sum)
```



| Region | Channel | Detergents_Paper |
|---|---|---|
| Lisbon | Hotel | 56081.00 |
| | Retail | 116648.25 |
| Oporto | Hotel | 13516.00 |
| | Retail | 117226.25 |
| Other | Hotel | 165990.00 |
| | Retail | 583289.75 |
| All | | 1052751.25 |

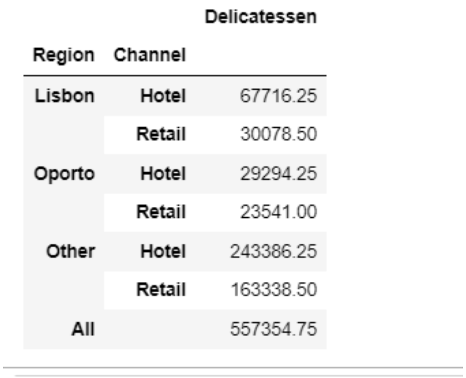
j.) Plot Spends on Detergent Paper against Channels and Regions by colour:

```
plt.figure(figsize=(10,10))
sns.boxplot(x='Channel',y='Detergents_Paper',hue='Region',data=data,palette='Set1')
plt.show()
```



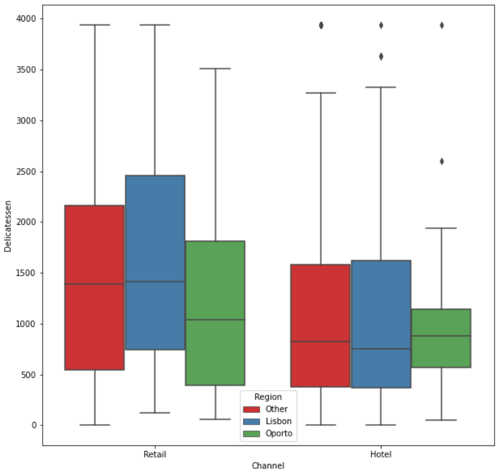
- From the above plot we that retail is left skewed and hotel is right skewed.
- IQR represents than retail spends more out of which customer spends more in Lisbon.
- There are outliers in case of hotel making the spend inconsistent.
- Retail makes a lot of income as compared to the hotel and main revenue comes from Lisbon region.

k.) Use Pivot Tables and find aggregate spends on Delicatessen:
pd.pivot_table(data,'Delicatessen',index=['Region','Channel'],margins=True,aggfunc=np.sum)



| | | Delicatessen |
|---|---|---|
| Region | Channel | |
| Lisbon | Hotel | 67716.25 |
| | Retail | 30078.50 |
| Oporto | Hotel | 29294.25 |
| | Retail | 23541.00 |
| Other | Hotel | 243386.25 |
| | Retail | 163338.50 |
| All | | 557354.75 |

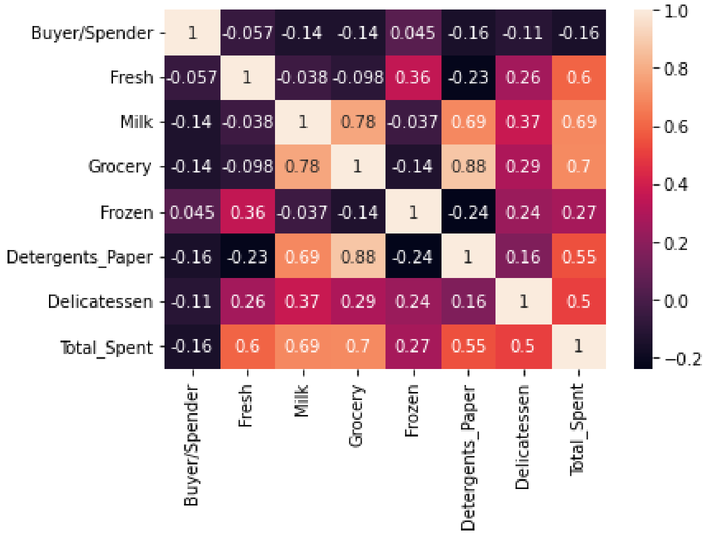
l.) Plot Spends on Delicatessen against Channels and Regions by colour:
plt.figure(figsize=(10,10))
sns.boxplot(x='Channel',y='Delicatessen',hue='Region',data=data,palette='Set1')
plt.show()



- From the IQR we see that the values for retail and hotel is right skewed.
- More spends on retail sector than hotel where customer in Lisbon spend more.
- Hotel is symmetrical but has outliers so its inconsistent.
- Least spent in oporto in both channels.

QUESTION 11:

a.) Plot correlation using heat map:
sns.heatmap(data.corr(), annot=True)



b.) Highly correlated are Grocery and detergents and 2nd highest is Grocery and milk.

c.) Grocery milk and detergent paper show multi-collinearity.

**Recommendation and Insights for the business:**

- From the analysis we see that hotel spends for on all these variables than retail. We must work towards increasing the spends in retail channel.
- Now from the inferences we see that delicateness of the detergent and the amount spent on this is more by the hotel in oporto than in retail similarly for the rest of the of the regions, hence we must increase the spends in the oporto regions for hotels.
- For detergent and paper Lisbon and oporto regions for channel hotel have the least spent as compared to the retail.SO we can increase the spend in retail sectors of Lisbon and oporto.
- For fresh product in hotels in oporto has least spends of all the other product as compared to retail and hotel. Increase the sales for Oporto.
- For frozen product maximum is spend in Hotel, region oporto. So we can increase the spends in retails of oporto.
- For grocery we see that least is spent in oporto, so increase the spends in hotels of Oporto since retails sales are pretty high.
- For milk again we that least spent on milk by Oporto, channel hotel, so increase the spends on this.
- But overall just increase the spends of retail than hotel for mostly the Oporto region. It has the least spend.