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).

Sample of dataset:

	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

Dataset has 9 variables, Buyer/Spender has unique row number for every transaction detail. There are 2 types of Channel (Hotel & Retail). There are 3 Regions (Other, Lisbon & Oporto) and rest are the 6 varieties for which the spending has been provided.

Let us check the types of variables in the data frame.

```
Buyer/Spender       int64
Channel            object
Region             object
Fresh               int64
Milk                int64
Grocery             int64
Frozen              int64
Detergents_Paper    int64
Delicatessen        int64
dtype: object
```

All the variables are in numerical format except Region and Channel which are in object format

There are total 440 rows and 8 columns in the dataset

Check for missing values in the dataset:

```
RangeIndex: 440 entries, 0 to 439
Data columns (total 8 columns):
Channel             440 non-null object
Region              440 non-null object
Fresh               440 non-null int64
Milk                440 non-null int64
Grocery             440 non-null int64
Frozen              440 non-null int64
Detergents_Paper    440 non-null int64
Delicatessen        440 non-null int64
dtypes: int64(6), object(2)
memory usage: 27.6+ KB
```

From the above results we can see that there is no missing value present in the dataset.

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?**

   Descriptive statistics help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are measures of centre: the mean, median, and mode, which are used at almost all levels of math and statistics.

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

Describe function will provide a table indicating the count of the variables, mean, standard deviation and other values for the 5 point summary that includes (min, 25%, 50%, 75% and max). 50% in the table is also known as Median.

The above descriptive statistics shows that average spending on Fresh is 1200, Milk is 5796, Grocery is 7951, Frozen is 3071, Detergents_Paper is 2881 and Delicatessen is 1525. From this result we can say that highest spending amount is on Grocery.

Now calculate median for all the variables

The "median" is the "middle" value in the sorted list of numbers.

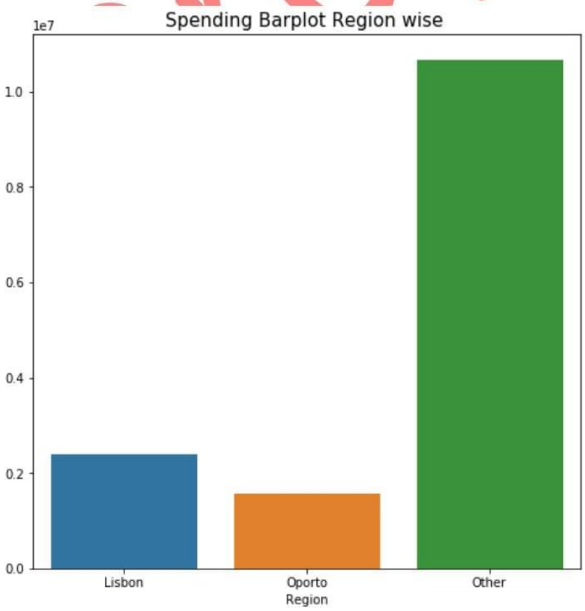
The Median of Fresh is 8504.0
The Median of Milk is 3627.0
The Median of Grocery is 4755.5
The Median of Frozen is 1526.0
The Median of Detergents_Paper is 816.5
The Median of Delicatessen is 965.5

The above results shows that the median of the data is same as the 50th percentile of the da taset. Since the mean and median of six variables are not same and there is very huge differe nce, when can say that the variables are highly skewed.

Mode Calculation

The "mode" is the value that occurs most often.

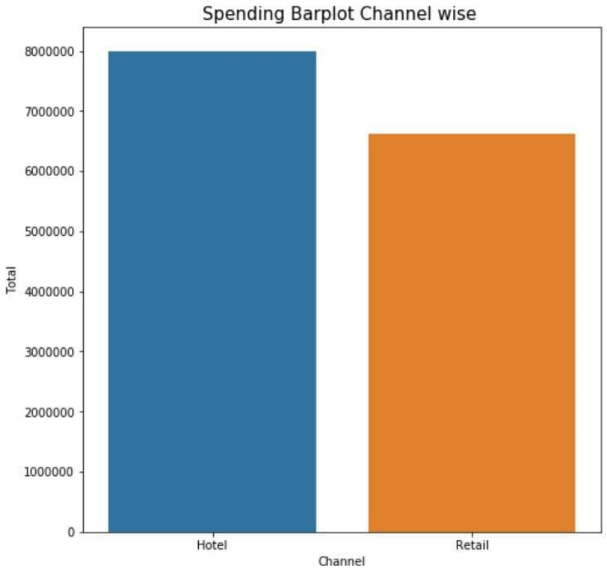
Since all the variables except Region and Channel is unique numerical values, there will be no mode. We can find the mode of Region and Channel



The most often occur value (mode) of Region is **Other**
The most often occur value (mode) of Channel is **Hotel**

That means that the most frequent value present in the Region column is **Other** and most frequent value in the Channel is **Hotel**.

From the barplot, we can see that Other region is spending the highest and Oporto Region is spending the least



From the barplot, we can see that Hotel Channel is spending the highest and Retail Channel is spending the least.

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.**
   To check the behavior of 6 different varieties, we will subset the dataset with respect to region and channel and analyze the descriptive statistics.

   Analysis of varieties in different Channel

- `Retail.describe()`

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	142.000000	142.000000	142.000000	142.000000	142.000000	142.000000
mean	8904.323944	10716.500000	16322.852113	1652.612676	7269.507042	1753.436620
std	8987.714750	9679.631351	12267.318094	1812.803662	6291.089697	1953.797047
min	18.000000	928.000000	2743.000000	33.000000	332.000000	3.000000
25%	2347.750000	5938.000000	9245.250000	534.250000	3683.500000	566.750000
50%	5993.500000	7812.000000	12390.000000	1081.000000	5614.500000	1350.000000

75%	12229.750000	12162.750000	20183.500000	2146.750000	8662.500000	2156.000000
max	44466.000000	73498.000000	92780.000000	11559.000000	40827.000000	16523.000000

Total count of spending done by Retail is 142. From the varying standard deviation ranging from (1953 to 12267) with high range, we found that all the variables don't show similar behaviour.

The minimum amount spend on Grocery is the highest and Delicatessen is the lowest.

- `Hotel.describe()`

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Count	298.000000	298.000000	298.000000	298.000000	298.000000	298.000000
Mean	13475.560403	3451.724832	3962.137584	3748.251678	790.560403	1415.956376
Std	13831.687502	4352.165571	3545.513391	5643.912500	1104.093673	3147.426922
Min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	4070.250000	1164.500000	1703.750000	830.000000	183.250000	379.000000
50%	9581.500000	2157.000000	2684.000000	2057.500000	385.500000	821.000000
75%	18274.750000	4029.500000	5076.750000	4558.750000	899.500000	1548.000000
Max	112151.000000	43950.000000	21042.000000	60869.000000	6907.000000	47943.000000

Total count of spendings done by Hotel is 298. From the varying standard deviation ranging from (1104 to 13832) with high range, we found that all the variables don't show similar behaviour.

The minimum amount spend on Milk is the highest. There are other 4 variables on which hotel has spent the same amount of minimum amount of 3.

Analysis of varieties in different Region

```
Lisbon.describe()
```

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000
mean	11101.727273	5486.415584	7403.077922	3000.337662	2651.116883	1354.896104

std	11557.438575	5704.856079	8496.287728	3092.143894	4208.462708	1345.423340
min	18.000000	258.000000	489.000000	61.000000	5.000000	7.000000
25%	2806.000000	1372.000000	2046.000000	950.000000	284.000000	548.000000
50%	7363.000000	3748.000000	3838.000000	1801.000000	737.000000	806.000000
75%	15218.000000	7503.000000	9490.000000	4324.000000	3593.000000	1775.000000
max	56083.000000	28326.000000	39694.000000	18711.000000	19410.000000	6854.000000

Total count of spending done by Lisbon is 77. From the varying standard deviation ranging from (1345 to 11557) with high range, we found that all the variables don't show similar behaviour for Lisbon Region.

The minimum amount spend on Grocery is the highest and Detergents_Paper is the lowest.

```
Oporto.describe()
```

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	47.000000	47.000000	47.000000	47.000000	47.000000	47.000000
mean	9887.680851	5088.170213	9218.595745	4045.361702	3687.468085	1159.702128
std	8387.899211	5826.343145	10842.745314	9151.784954	6514.717668	1050.739841
min	3.000000	333.000000	1330.000000	131.000000	15.000000	51.000000
25%	2751.500000	1430.500000	2792.500000	811.500000	282.500000	540.500000
50%	8090.000000	2374.000000	6114.000000	1455.000000	811.000000	898.000000
75%	14925.500000	5772.500000	11758.500000	3272.000000	4324.500000	1538.500000
max	32717.000000	25071.000000	67298.000000	60869.000000	38102.000000	5609.000000

Total count of spending done by Oporto is 47. From the varying standard deviation ranging from (1050 to 10843) with high range, we found that all the variables don't show similar behaviour for Oporto Region.

The minimum amount spend on Grocery is the highest and Fresh is the lowest.

```
Other.describe()
```

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	316.000000	316.000000	316.000000	316.000000	316.000000	316.000000

mean	12533.471519	5977.085443	7896.363924	2944.594937	2817.753165	1620.601266
std	13389.213115	7935.463443	9537.287778	4260.126243	4593.051613	3232.581660
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	3350.750000	1634.000000	2141.500000	664.750000	251.250000	402.000000
50%	8752.500000	3684.500000	4732.000000	1498.000000	856.000000	994.000000
75%	17406.500000	7198.750000	10559.750000	3354.750000	3875.750000	1832.750000
max	112151.000000	73498.000000	92780.000000	36534.000000	40827.000000	47943.000000

Total count of spending done by Other Region is 316. From the varying standard deviation ranging from (3232 to 13389) with high range, we found that all the variables don't show similar behaviour for Lisbon Region.

The minimum amount spend on Milk is the highest. There are other 4 variables on which Other has spent the same amount of minimum amount of 3.

III. On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour?

Descriptive measures of variability are used to describe the amount of variability or spread in a set of data. The most common measures of variability are the range, the interquartile range (IQR), variance, standard deviation, and coefficient of variation. We will use coefficient of variation here.

The coefficient of variation (CV) which is a statistical measure of the dispersion of data points in a data series around the mean. It is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another.

$$CV = \sigma / \mu$$

Where σ=standard deviation and μ=mean

- The Coefficient of Variation for Fresh is **1.053**
- The Coefficient of Variation for Milk is 1.272
- The Coefficient of Variation for Grocery is 1.194
- The Coefficient of Variation for Frozen is 1.579
- The Coefficient of Variation for Detergents Paper is 1.653
- The Coefficient of Variation for Delicatessen is **1.847**

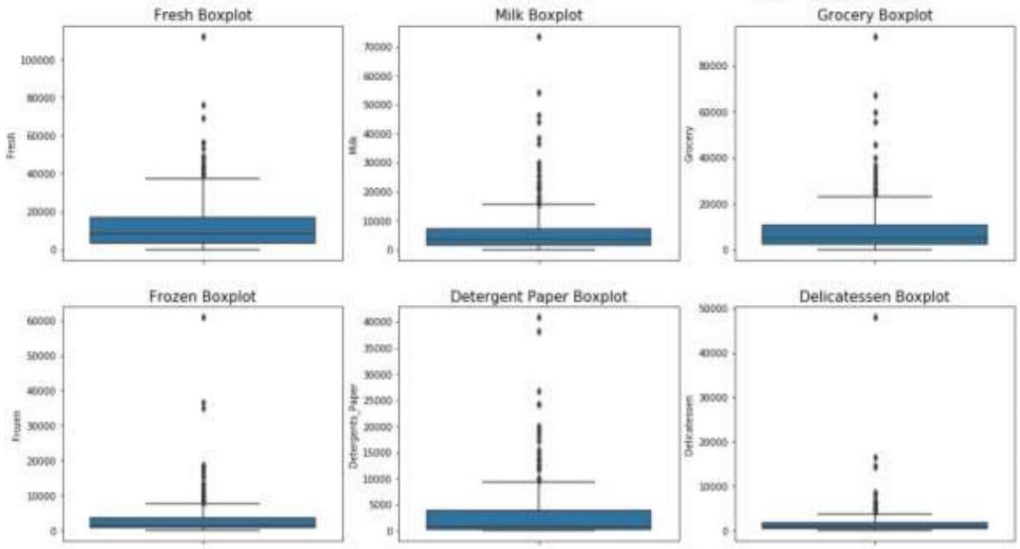
From the above results, we found that **Delicatessen** shows the most inconsistent behaviour and **Fresh** shows the least inconsistent behaviour

IV. Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.
To check the outliers in the dataset, there are couple of graphs and methods are available to perform.

Plots which we can use to check for the outliers are Boxplot, scatterplot. Data point far from other data point present in the plots would be considered as an outlier.

We can calculate IQR Interquartile range for the variable. If the data point is outside the IQR range, then it would consider as outlier.

Below are the boxplots for the different varieties present in the dataset.



From the boxplots, we found that all the variables have outliers as there are not included in the box of observations i.e. nowhere near the quartiles.

V. 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.
From the all the analysis done, below are the Observations & Recommendations:

1. Out of all the regions, Other region is spending the highest and Oporto is spending the lowest.
2. Hotel is spending more than Retail.
3. Out of all the 6 varieties, the highest spending was done on Fresh followed by Grocery, Milk, Frozen, Detergents, Paper, and Delicatessen.
4. There are outliers present in the dataset.

Appendix Code

Wholesale Customers Analysis

Basic python packages load

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns


from scipy.stats import iqr #To calculate the IQR - Interquartile Range
import statistics as stat # To calculate the MODE
from statistics import stdev # To calculate the standard deviation
#stdev(data['Fresh'])
```

To set the working directory

[3]:
```
# import os
# os.chdir("D:\\Academic Operations\\DSBA - Python\\Online\\SMDM\\Project")
```

Load the dataset

[2]:
```
data = pd.read_csv("Wholesale Customer.csv")
```

Exploratory Data Analysis

Let us check if the data has been loaded.

[3]:
```
data.head()
```

t[3]:

	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

Let us check the types of variables in the data frame.

[4]:
```
data.dtypes
```

t[4]:
```
Buyer/Spender         int64
Channel              object
Region               object
Fresh                 int64
Milk                  int64
Grocery               int64
Frozen                int64
Detergents_Paper      int64
Delicatessen          int64
dtype: object
```

From the above results, we can see that all the variables except "Region" and "Channel" are in integer format and other 2 are object format.

The "Buyer/Spender" field which is unique number which is unnecessary, so let's remove it!

[5]:
```
data.drop('Buyer/Spender',axis=1, inplace=True)
```

Again check for the dataset after removal of "Buyer/Spender" column from the dataset

[6]:
```
data.head()
```

t[6]:

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

Now, let us perform some measures of |Descriptive Statistics

Let us check the names of the columns in our data frame.

```
data.columns
```

```
Index(['Channel', 'Region', 'Fresh', 'Milk', 'Grocery', 'Frozen',
       'Detergents_Paper', 'Delicatessen'],
      dtype='object')
```

Let us check the number of rows(observations) and columns(variables) in the data frame.

```
row, col = data.shape
print("There are total {}".format(row), "rows and {}".format(col), "columns in the dataset")
```

```
There are total 440 rows and 8 columns in the dataset
```

Let us check if any value in the data frame is null.

We can directly use 'data.info()' instead of using 'data.dtypes'. The code snippet 'data.info()' gives us the complete information about the data frame.

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440 entries, 0 to 439
Data columns (total 8 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Channel           440 non-null    object
 1   Region            440 non-null    object
 2   Fresh             440 non-null    int64
 3   Milk              440 non-null    int64
 4   Grocery           440 non-null    int64
 5   Frozen            440 non-null    int64
 6   Detergents_Paper  440 non-null    int64
 7   Delicatessen      440 non-null    int64
dtypes: int64(6), object(2)
memory usage: 27.6+ KB
```

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

```
Channel             0
Region              0
Fresh               0
Milk                0
Grocery             0
Frozen              0
Detergents_Paper    0
Delicatessen        0
dtype: int64
```

From the above result, it is evident that there is no missing/null values in the dataset.

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

Let us check mean and the various measures of dispersion. From the below code snippet, we also get a proper idea about various necessary common measures of descriptive statistics.

```
data.describe(include = 'all')
```

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

From the above descriptive statistics, average spending on Fresh is 1200, Milk is 5796, Grocery is 7951, Frozen is 3071, Detergents_Paper is 2881 and Delicatessen is 1525. From this result we can say that Highgest spending amount is on Grocery.

Now calculate median for all the variables

The "median" is the "middle" value in the sorted list of numbers.

```
# Median
#columns = ['Fresh','Milk','Grocery','Frozen','Detergents_Paper','Delicatessen']
med1 = np.median(data['Fresh'])
print("The Median of Fresh is {}".format(med1))
med2 = np.median(data['Milk'])
print("The Median of Milk is {}".format(med2))
med3 = np.median(data['Grocery'])
print("The Median of Grocery is {}".format(med3))
med4 = np.median(data['Frozen'])
print("The Median of Frozen is {}".format(med4))
med5 = np.median(data['Detergents_Paper'])
print("The Median of Detergents_Paper is {}".format(med5))
med6 = np.median(data['Delicatessen'])
print("The Median of Delicatessen is {}".format(med6))
```

```
The Median of Fresh is 8504.0
The Median of Milk is 3627.0
The Median of Grocery is 4755.5
The Median of Frozen is 1526.0
The Median of Detergents_Paper is 816.5
The Median of Delicatessen is 965.5
```

The above results shows that the median of the data is same as the 50th percetile of the dataset. Since the mean and median of six variables are not same and there is very huge difference, when can say that the variables are highly skewed.

Mode Calculation

The "mode" is the value that occurs most often.

Since all the variables except Region and Channel is uniuqe numerical values, there will be no mode. We can find the mode of Region and Channel

```
# Mode
from statistics import mode
mod1 = mode(data['Region'])
print("The most often occuring value (mode) of Region is {}".format(mod1))
mod2 = mode(data['Channel'])
print("The most often occuring value (mode) of Channel is {}".format(mod2))
```

```
The most often occuring value (mode) of Region is Other
The most often occuring value (mode) of Channel is Hotel
```

Now, let us check the the different observations and the counts of each of these observations under the variables "Region" and "Channel".

```
data["Region"].value_counts()
```

```
Other     316
Lisbon     77
Oporto     47
Name: Region, dtype: int64
```

There are 47 observations of Oporto, 77 for Lisbon and 316 for Other Regions.

```
data["Channel"].value_counts()
```

```
Hotel     298
Retail    142
Name: Channel, dtype: int64
```

There are 142 observations of Retail and 298 for Hotel Channel.

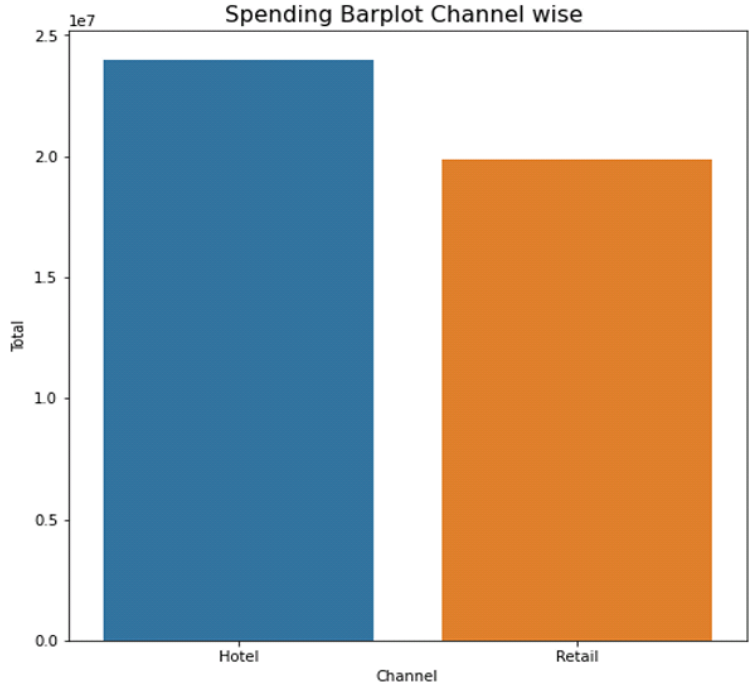
Now, we will create a new column of total of spendings by 6 different varieties

```
#### Comment by Mehak
##### Can't remove total column as the value in the plot below is the sum of amount spent regionwise a
d channelwise


## Adding row totals to the data frame
data['Total'] = data.sum(axis = 1)
data.head()
```

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
0	Retail	Other	12669	9656	7561	214	2674	1338	102336
1	Retail	Other	7057	9810	9568	1762	3293	1776	99798
2	Retail	Other	6353	8808	7684	2405	3516	7844	109830
3	Hotel	Other	13265	1196	4221	6404	507	1788	82143
4	Retail	Other	22615	5410	7198	3915	1777	5185	138300

Create barplot to check the spending as per Channel

```
plt.figure(figsize=(8, 8))
ChannelAggregated = pd.DataFrame(data.groupby(["Channel"],sort=True)["Total"].sum()).reset_index()
b = sns.barplot(x=ChannelAggregated["Channel"], y=ChannelAggregated["Total"], data=ChannelAggregated)
b.set_title("Spending Barplot Channel wise ",fontsize=15)
plt.show()
```



From the above barplot, we can see that Hotel Channel is spending the highest and Retail Channel is spending the least

```
#Now drop the Total column
data.drop('Total',axis=1, inplace=True)
```

II. 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.

We will subset the dataset with respect to region and channel

```
# Channel wise data subset
Retail = data[data['Channel'] == "Retail"]
Hotel = data[data['Channel'] == "Hotel"]
```

To check the behaivour of the varities, we will do the descriptive analytics

```
Retail.describe()
```

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	142.000000	142.000000	142.000000	142.000000	142.000000	142.000000
mean	8904.323944	10716.500000	16322.852113	1652.612676	7269.507042	1753.436620
std	8987.714750	9679.631351	12267.318094	1812.803662	6291.089697	1953.797047
min	18.000000	928.000000	2743.000000	33.000000	332.000000	3.000000
25%	2347.750000	5938.000000	9245.250000	534.250000	3683.500000	566.750000
50%	5993.500000	7812.000000	12390.000000	1081.000000	5614.500000	1350.000000
75%	12229.750000	12162.750000	20183.500000	2146.750000	8662.500000	2156.000000
max	44466.000000	73498.000000	92780.000000	11559.000000	40827.000000	16523.000000

Total count of spendings done by Retail is 142. From the varying standard deviation ranging from (1953 to 12267) with high range, we found that all the variables don't show similar behavior.

The minimum amount spend on Grocery is the highest and Delicatessen is the lowest.

]:
```
Hotel.describe()
```

]:

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	298.000000	298.000000	298.000000	298.000000	298.000000	298.000000
mean	13475.560403	3451.724832	3962.137584	3748.251678	790.560403	1415.956376
std	13831.687502	4352.165571	3545.513391	5643.912500	1104.093673	3147.426922
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	4070.250000	1164.500000	1703.750000	830.000000	183.250000	379.000000
50%	9581.500000	2157.000000	2684.000000	2057.500000	385.500000	821.000000
75%	18274.750000	4029.500000	5076.750000	4558.750000	899.500000	1548.000000
max	112151.000000	43950.000000	21042.000000	60869.000000	6907.000000	47943.000000

Total count of spendings done by Hotel is 298. From the varying standard deviation ranging from (1104 to 13832) with high range, we found that all the variables don't show similar behavior.

The minimum amount spend on Milk is the highest. There are other 4 variables on which hotel has spend the same amount of minimum amount of 3.

]:
```
# Region wise data subset
Lisbon = data[data['Region'] == "Lisbon"]
Oporto = data[data['Region'] == "Oporto"]
Other = data[data['Region'] == "Other"]
```

]:
```
Lisbon.describe()
```

]:

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000
mean	11101.727273	5486.415584	7403.077922	3000.337662	2651.116883	1354.896104
std	11557.438575	5704.856079	8496.287728	3092.143894	4208.462708	1345.423340
min	18.000000	258.000000	489.000000	61.000000	5.000000	7.000000
25%	2806.000000	1372.000000	2046.000000	950.000000	284.000000	548.000000
50%	7363.000000	3748.000000	3838.000000	1801.000000	737.000000	806.000000
75%	15218.000000	7503.000000	9490.000000	4324.000000	3593.000000	1775.000000
max	56083.000000	28326.000000	39694.000000	18711.000000	19410.000000	6854.000000

Total count of spendings done by Lisbon is 77. From the varying standard deviation ranging from (1345 to 11557) with high range, we found that all the variables don't show similar behavior for Lisbon Region.

The minimum amount spend on Grocery is the highest and Detergents_Paper is the lowest.

]:
```
Oporto.describe()
```

]:

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	47.000000	47.000000	47.000000	47.000000	47.000000	47.000000
mean	9887.680851	5088.170213	9218.595745	4045.361702	3687.468085	1159.702128
std	8387.899211	5826.343145	10842.745314	9151.784954	6514.717668	1050.739841
min	3.000000	333.000000	1330.000000	131.000000	15.000000	51.000000
25%	2751.500000	1430.500000	2792.500000	811.500000	282.500000	540.500000
50%	8090.000000	2374.000000	6114.000000	1455.000000	811.000000	898.000000
75%	14925.500000	5772.500000	11758.500000	3272.000000	4324.500000	1538.500000
max	32717.000000	25071.000000	67298.000000	60869.000000	38102.000000	5609.000000

Total count of spendings done by Oporto is 47. From the varying standard deviation ranging from (1050 to 10843) with high range, we found that all the variables don't show similar behavior for Oporto Region.

The minimum amount spend on Grocery is the highest and Fresh is the lowest.

]:
```
Other.describe()
```

]:

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	316.000000	316.000000	316.000000	316.000000	316.000000	316.000000
mean	12533.471519	5977.085443	7896.363924	2944.594937	2817.753165	1620.601266
std	13389.213115	7935.463443	9537.287778	4260.126243	4593.051613	3232.581660
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	3350.750000	1634.000000	2141.500000	664.750000	251.250000	402.000000
50%	8752.500000	3684.500000	4732.000000	1498.000000	856.000000	994.000000
75%	17406.500000	7198.750000	10559.750000	3354.750000	3875.750000	1832.750000
max	112151.000000	73498.000000	92780.000000	36534.000000	40827.000000	47943.000000

Total count of spendings done by Other Region is 316. From the varying standard deviation ranging from (3232 to 13389) with high range, we found that all the variables don't show similar behavior for Lisbon Region.

The minimum amount spend on Milk is the highest. There are other 4 variables on which Other has spend the same amount of minimum amount of 3.

III. On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour?

Descriptive measures of variability is used to describe the amount of variability or spread in a set of data. The most common measures of variability are the range, the interquartile range (IQR), variance, standard deviation and coefficient of variation. We will use coefficient of variation.

The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a series around the mean. It is a useful statistic for comparing the degree of variation from one data series to another, even if the means are very different from one another.

This measure is the most appropriate measure for the current scenario

$$CV = \sigma/\mu$$

σ=standard deviation
μ=mean

```
from scipy.stats import variation
cols=['Fresh', 'Milk', 'Grocery', 'Frozen',
      'Detergents_Paper', 'Delicatessen']
for i in cols:
    print('The Coefficient of Variation for {} is {}'.format(i,round(variation(data[i]),3)))
```

```
The Coefficient of Variation for Fresh is 1.053
The Coefficient of Variation for Milk is 1.272
The Coefficient of Variation for Grocery is 1.194
The Coefficient of Variation for Frozen is 1.579
The Coefficient of Variation for Detergents_Paper is 1.653
The Coefficient of Variation for Delicatessen is 1.847
```

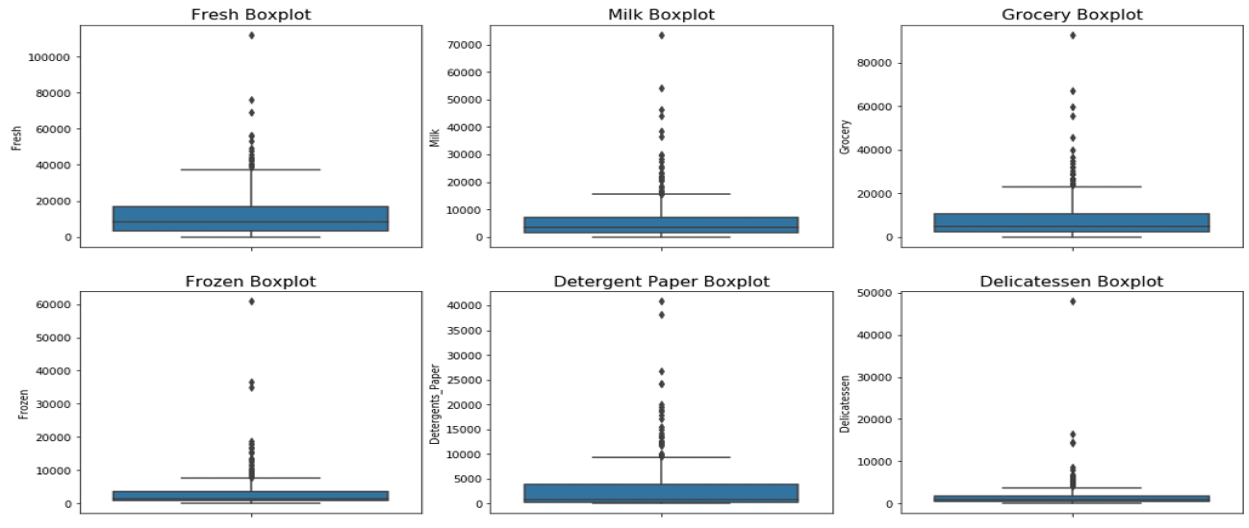
From the above result of CV , we found that Delicatessen shows the most inconsistent behaviour and Fresh shows the least inconsistent behaviour

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

In statistics, an outlier is an observation point that is distant from other observations.

In descriptive statistics, a box plot is a method for graphically depicting groups of numerical data through their quartiles.

```
fig, axes = plt.subplots(nrows=2,ncols=3)
fig.set_size_inches(18, 10)
a = sns.boxplot(data = data, x = "Fresh" , orient = "v" , ax=axes[0][0])
a.set_title("Fresh Boxplot",fontsize=15)

b = sns.boxplot(data = data, x = "Milk" , orient = "v" , ax=axes[0][1])
b.set_title("Milk Boxplot",fontsize=15)

c = sns.boxplot(data = data, x = "Grocery" , orient = "v" , ax=axes[0][2])
c.set_title("Grocery Boxplot",fontsize=15)

d = sns.boxplot(data = data, x = "Frozen" , orient = "v" , ax=axes[1][0])
d.set_title("Frozen Boxplot",fontsize=15)

e = sns.boxplot(data = data, x = "Detergents_Paper" , orient = "v" , ax=axes[1][1])
e.set_title("Detergent Paper Boxplot",fontsize=15)

f = sns.boxplot(data = data, x = "Delicatessen" , orient = "v" , ax=axes[1][2])
f.set_title("Delicatessen Boxplot",fontsize=15)

plt.show()
```



From the boxplots, we found that all the variables have outliers as there are not included in the box of observations i.e no where near the quartiles.

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

From the all the analysis done , below are the Observations & Recommendations:

1. Out of all the regions, Other region is spending the higest and Oporto is spending the lowest.
2. Hotel is spending more than Retail.
3. Out of all the 6 varities, the highest spending was done on Fresh followed by Grocery,Milk,Frozen,Detergents_Paper, Delicatessen.
4. There are outliers present in the dataset.

]: