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 seems to be spend more? Which Region and which Channel seems to spend less?**

	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

Descriptive statistics helps to understand the features of a specific data set through visual and numerical summarization. Among the most recognized types of descriptive statistics are measures of central tendency: mean, median, and mode.

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% (median), 75% and max, if the variable is continuous.

The above descriptive statistics shows that average spending on Fresh is 12000, Milk is 5796, Grocery is 7951, Frozen is 3071, Detergents_Paper is 2881 and Delicatessen is 1525. From this result we can say that the highest spending amount is on Fresh.

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

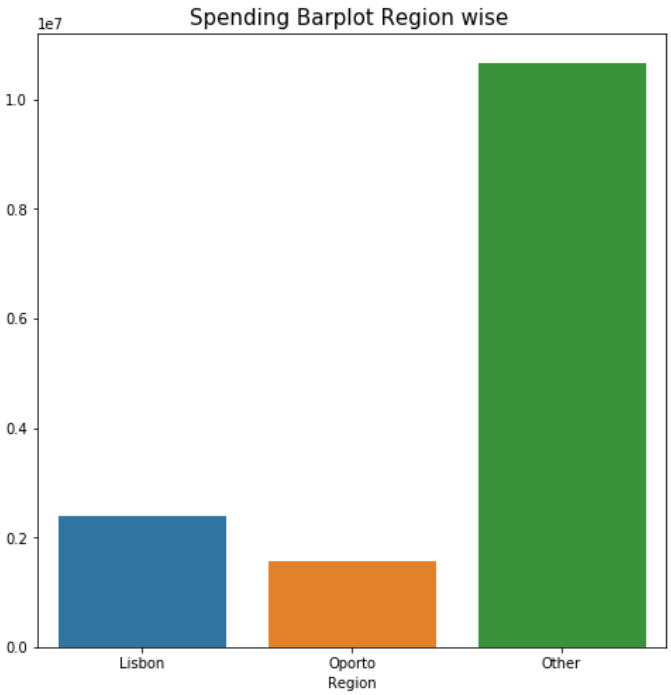
The above results show that The median is the 50th percentile of a variable. Since the mean and median of each of the six continuous variables show considerable difference, it may be noted that the variables are highly skewed.

Mode Calculation

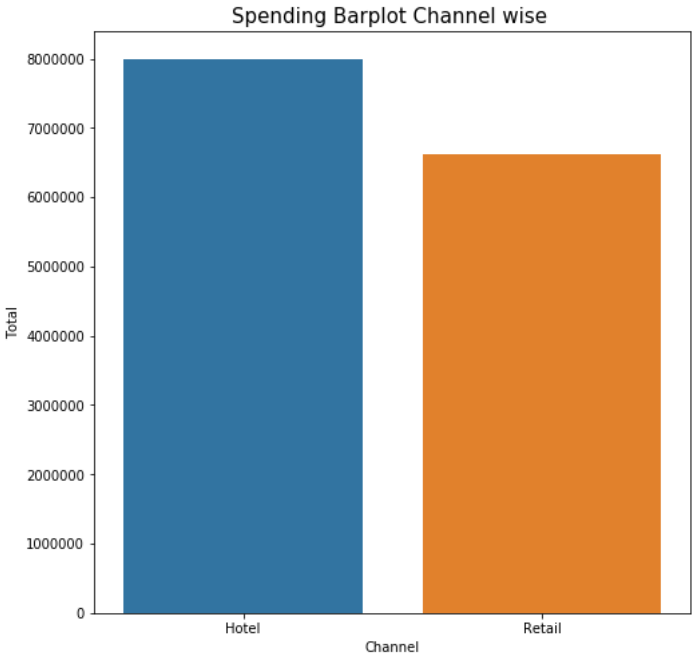
The "mode" is the value that occurs most often.

Since all the variables except Region and Channel is continuous, mode is not helpful for them. Typically, mode is used for discrete variables. We can find the mode of Region and Channel

The most frequently occurring value (mode) of Region is **Other**
The most frequently occurring value (mode) of 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 are considered. Do all varieties show similar behavior across Region and Channel?**

   To check the behavior of 6 different varieties, we will subset the dataset with respect to region and channel and consider 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

Retail has 142 observations. As the means, medians and standard deviations across the 6 variables show considerable differences, it is clear that spending is different in different category.

The minimum amount spend on Grocery is the highest and Delicatessen is the lowest. The maximum amount spent in the Frozen category is the lowest and Grocery maximum spent is the

highest. In fact, spending in Grocery is the highest in Retail, as its mean and the 3 quartiles have the highest numerical values compared to other variables. Note that its SD is also the highest.

```
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

Hotel has 298 observations.

The minimum amount spend on Milk is the highest. Fresh category is among those where minimum spent is the lowest, but the mean, the 3 quartiles and the maximum spent in this category is the highest. Note that this has the highest Standard Deviation too.

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

Lisbon has 77 observations. Here Fresh category has the highest value for all the statistics, except for minimum spent.

```
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

Oporto has 47 observations. In this region, mean spent on Fresh and Grocery are comparable, but neither the medians, nor the maximum spent amounts. SD for Fresh is also much smaller than Grocery.

```
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

Other Region has 316 observations.

The minimum amount spend on Milk is the highest. However, Fresh category has the highest numerical values for all other statistics.

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

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). may be considered as a normalized (standardized) value of SD. Typically, while comparing among variables whose means show wide differences, CV is used instead of Standard Deviation.

$$CV = \sigma / \mu$$

Where σ=standard deviation and μ=mean in the population. In a sample the population parameters are replaced by their sample estimates and

$$CV = s / \bar{x}$$

CV is often expressed as a %, i.e.

$$CV\% = s / \bar{x} \times 100$$

- The Coefficient of Variation for Fresh is **1.053 (or 105%)**
- The Coefficient of Variation for Milk is 1.272 (or 127%)
- The Coefficient of Variation for Grocery is 1.194 (or 119.5%)
- The Coefficient of Variation for Frozen is 1.579 (or 158%)
- The Coefficient of Variation for Detergents Paper is 1.653 (or 165%)
- The Coefficient of Variation for Delicatessen is **1.847 (or 185%)**

These values are obtained from the all 440 observations taken together.

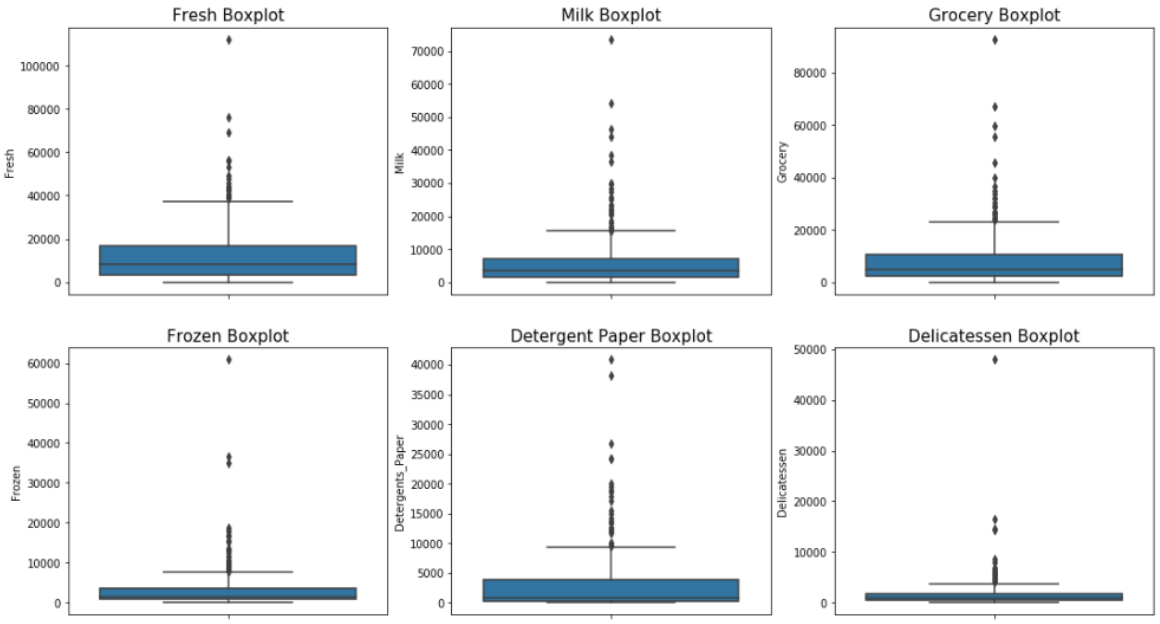
From the above results, we found that **Delicatessen** shows the most inconsistent behaviour (highest CV) and **Fresh** shows the least inconsistent behaviour (lowest CV)

IV. Are there any outliers in the data?

An outlier is an observation which is far from the main concentration of the observations. To check the outliers in the dataset, Boxplot is the easiest and the most useful technique. A boxplot is constructed using Q1, Q2, Q3 and IQR = Q3 – Q1. All the points outside of the range [Q1 – 1.5 IQR, Q3 + 1.5 IQR] may be considered outliers.

[*Alternatively, if the data set is large and the above criterion identifies a large proportion of points as outliers, [Q1 – 3 IQR, Q3 + 3 IQR] is also used. Only points outside of the ±3IQR range are termed as outliers*]

Below are the boxplots for the different varieties present in the dataset. These plots are constructed on all 440 observations.



It is clear from the boxplots above all 6 variables contain outliers. The whiskers below and above the boxes are the ±1.5 IQR values from Q1 and Q3 respectively. All the outliers are on the higher side of the distribution.

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

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

1. Out of the 3 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 is on Fresh followed by Grocery, Milk, Frozen, Detergents_Paper, and Delicatessen.
4. When analysing spend behaviour within Channels and within Region, Grocery spending is not consistently highest. For Chanel = Hotel and for Region = Lisbon and Other spending in Fresh category is the highest
5. There are outliers present in the dataset.