BUSINESS REPORT OF STATISTICAL METHODS FOR DECISION MAKING

****

|  |
| --- |
| 1.1.1 Use methods of descriptive statistics to summarize data. |
| 1.1.2 Which Region and which Channel spent the most? |
| 1.1.3 Which Region and which Channel spent the least? |
| 1.2. 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. |
| 1.3 On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour? |
| 1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments. |
| 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. |
| Perform Exploratory Data Analysis [Univariate, Bivariate, and Multivariate analysis to be performed]. What insight do you draw from the EDA? |
| Quality of Business Report (Please refer to the Evaluation Guidelines for Business report checklist. Marks in this criteria are at the moderator's discretion) |

**Wholesale Customer 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.

## Data Description:

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).

## Domain:

Retail

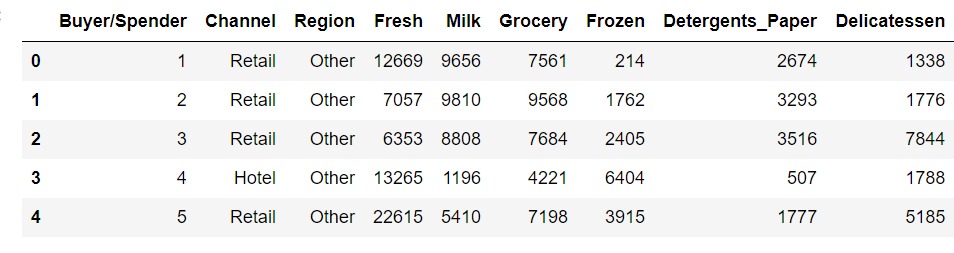
## Context:

Leveraging sales information is paramount for customers. In the case of a retail chains, analysis of products of retail chains sold across different region under different sales channel are very crucial in making business decisions. Hence, knowing to explore and generate value out of such data can be an invaluable skill to have.

## Attribute Information:

* Buyer/Spender
* Channel
* Region
* Fresh
* Milk
* Grocery
* Frozen
* Detergents Paper
* Delicatessen

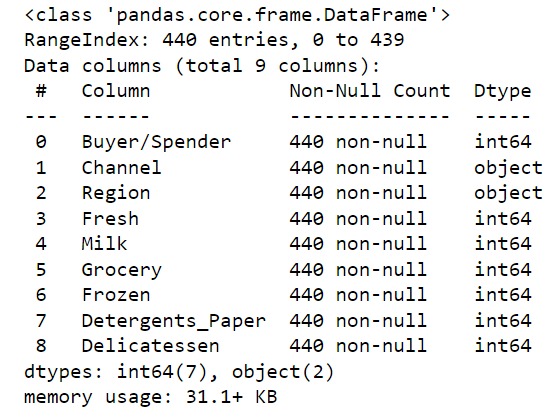
**Sample of the dataset**



Dataset has 9 variables with 2 different types of channel, 3 different region and 6 different types of items of 440 buyer/spender

## Exploratory Data Analysis

## Checking the data types /information’s



The data set contains 440 observations of data and 9 variables. Only Channel and Region are categorical while rest is numeric data. Since non null count is same in every column variable and is equal to total rows, hence there is no null data in any of these columns.

## Checking for missing values

## 

We can confirm that there is no NULL value.



Total no of rows are 440 and columns are 9

## Checking for summary

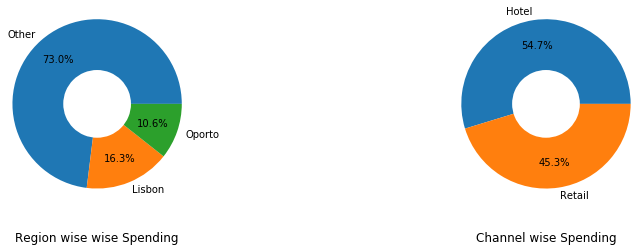
## 

Ignoring first column which is just a reference of Buyer/Spender no, we can see that:-

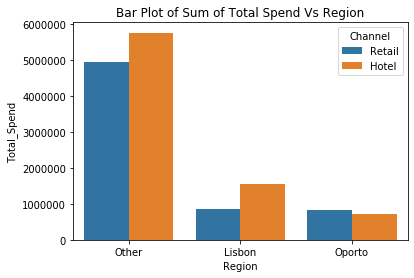
* max value of all the column attributes are high as compared to the median value. Hence there seems to be many outliers in this data.
* On checking the median values (50%), it appears that retailers spend more on Fresh products and grocery as compared to others.
* 75% of 440 retailers spend only 1820 or less on Delicatessen. So annual spend of Delicatessen appears to be least among all.

# 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?

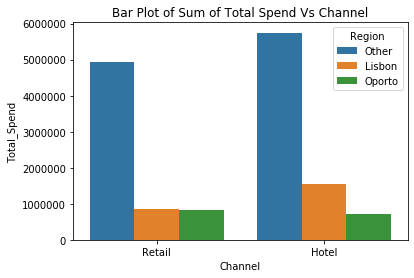
Summary statistics is already present in the EDA done above.



From the pie chart table, it can be seen that, a significant proportion of customers around 73% belongs to Other region and 54.7% of them span across hotel channel. The number of customers of hotel channel are double the number of customers who preferred retail.

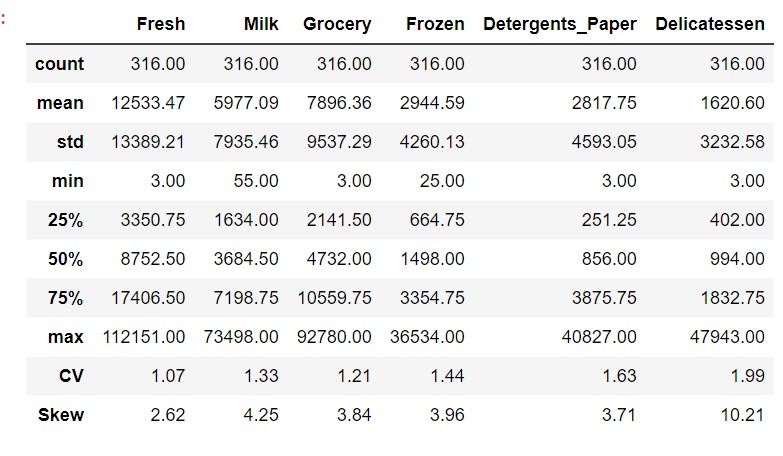


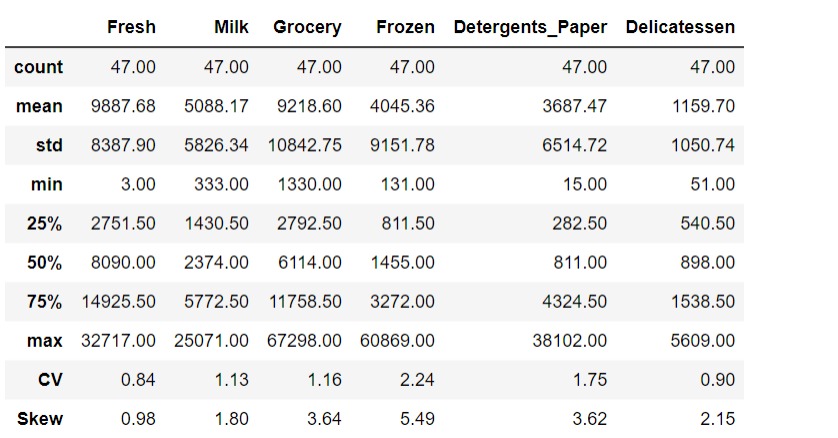
From the bar graph it is evident that, Other region recorded the highest spending and Oporto region being the least spent.

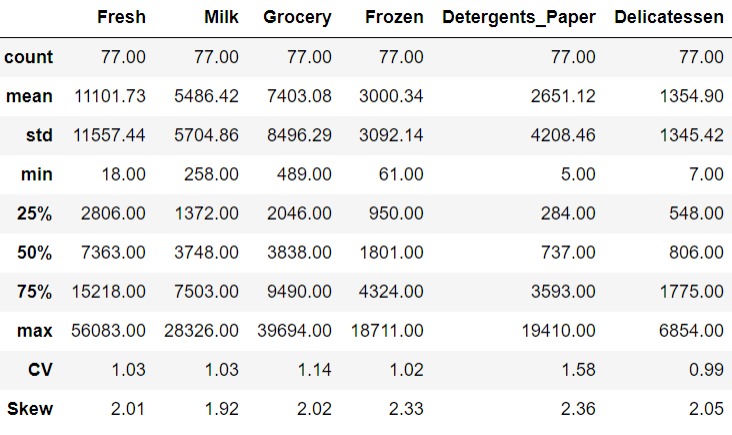


The bar graph shows that the hotel channel recorded highest spending and the retail channel being the lowest.

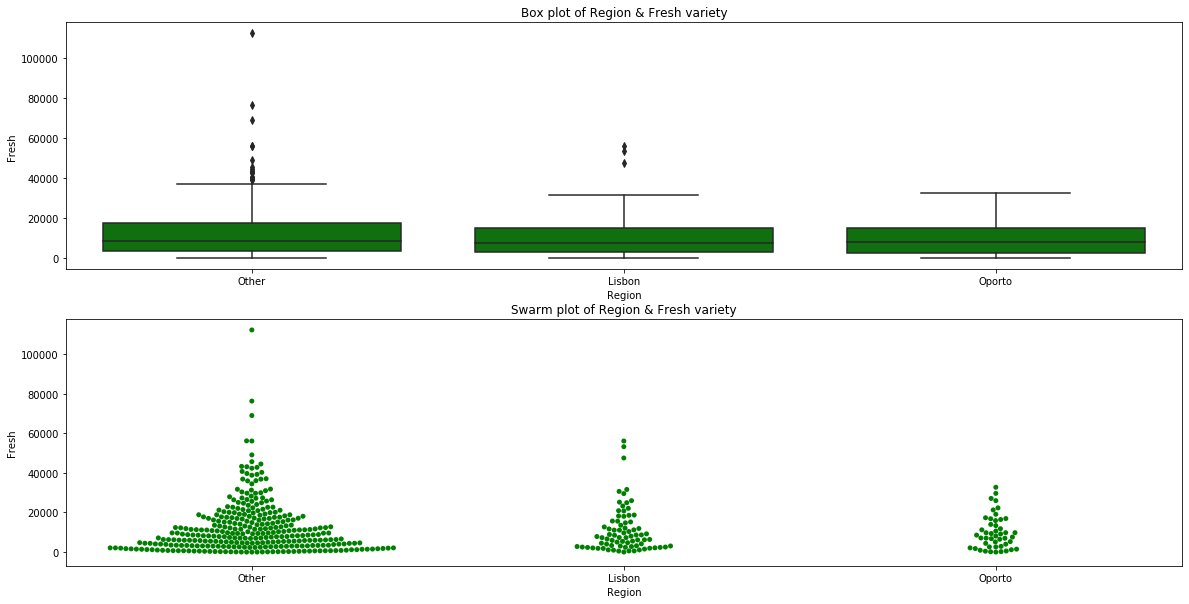
# 1.2) There are 6 different varieties of items are considered. Do all varieties show similar behaviour across Region and Channel? Provide justification for your answer.







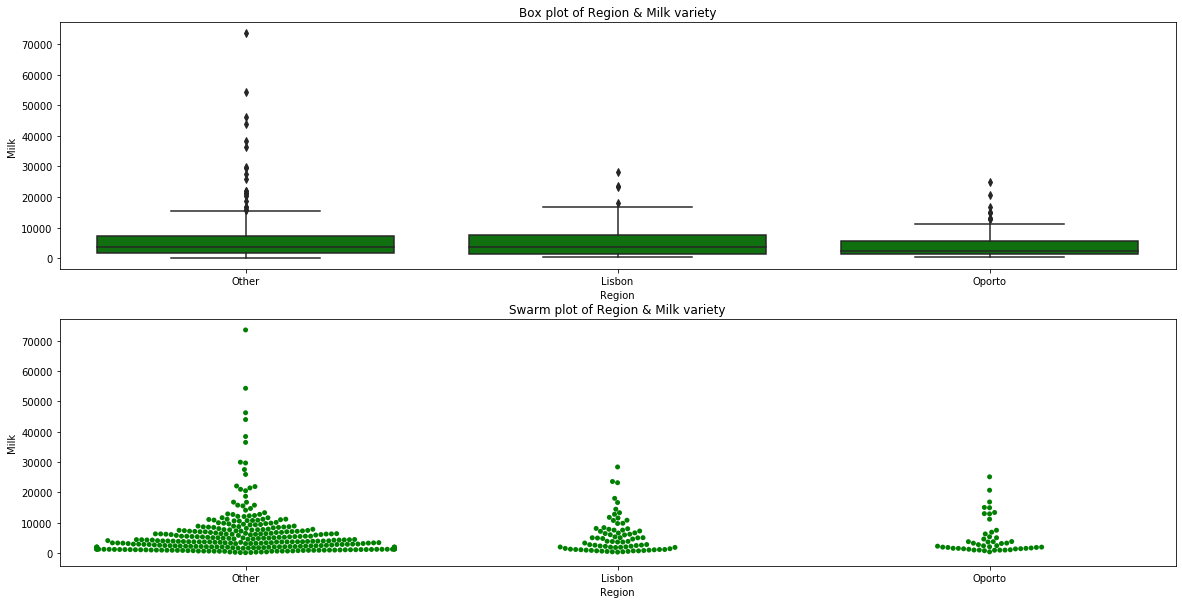




As seen in the summary statistics and swarm plot/box plot,

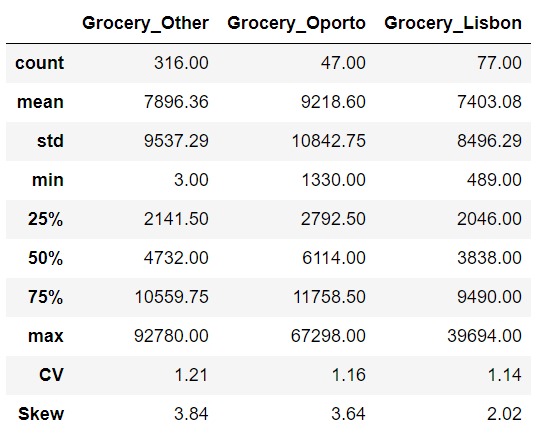
* Data contains more retailers in Other region as compared to Lisbon and Oporto.
* Except Oporto region, 'Other' and 'Lisbon' region Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
* Maximum annual spending in 'Other' region is very high as compared to other regions.
* Annual median spend of Other region is slightly higher(8752) than of Lisbon(7363) and Oporto region(8090)
* Although mean and median value of 'Other' region is highest but its volatility is also high i.e. it is the most inconsistent region for Fresh Variety
* Spread of data looks similar across all regions with distribution being right/positive skewed and 75% of retailers spending less than 17.5K annually across all three regions.
* Footfall is more for buyers under 'Other' Region and majority of buyers(75%) are spending less than around 8K across all three regions.

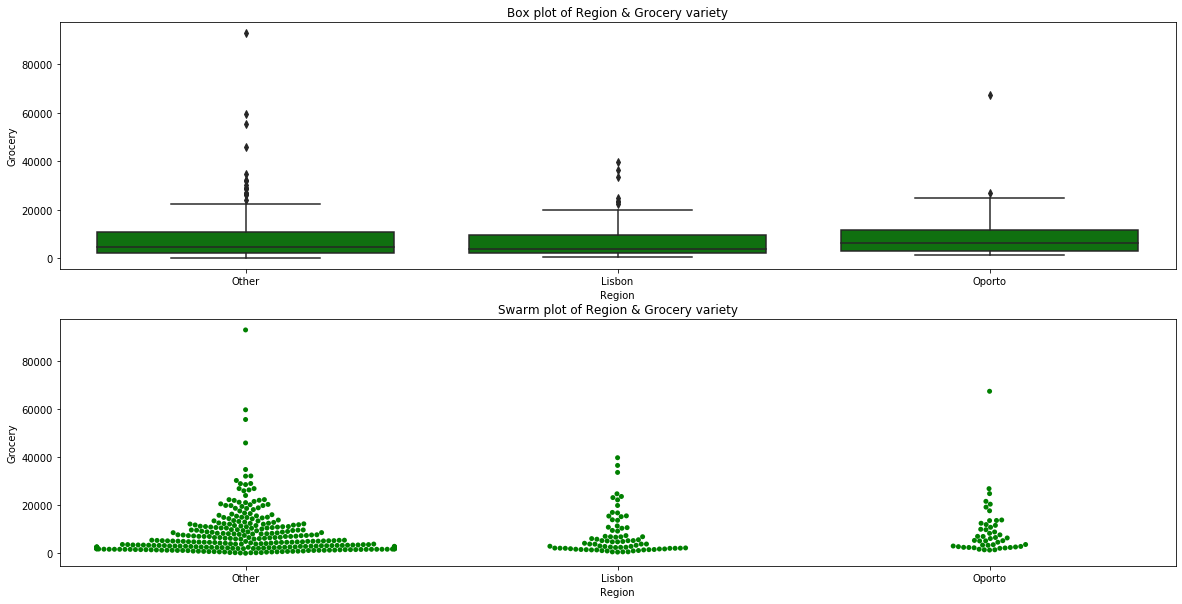




As seen in the summary statistics and swarm plot/box plot,

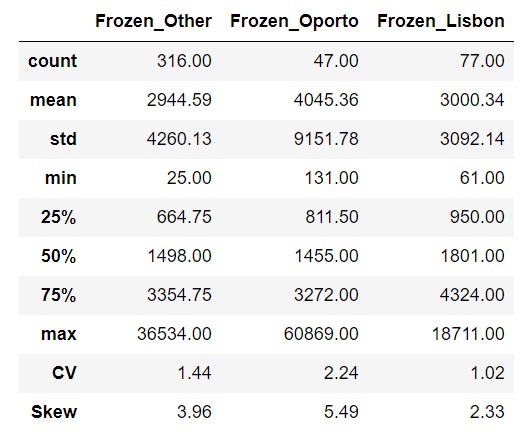
* Data contains more retailers in Other region as compared to Lisbon and Oporto.
* All three regions Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
* Maximum annual spending in 'Other' region is very high as compared to Lisbon/Oporto regions.
* Minimum annual spending in 'Other' region is low as compared to other Lisbon/Oporto regions.
* Annual median spend of Oporto region is slightly lower(2374) than of Lisbon(3748) and Other region(3684)
* Volatility of other region is highest among all i.e. it is the most inconsistent region for Milk Variety
* Spread of data looks similar across all regions with distribution being right/positive skewed and 75% of retailers spending less than 7.5K annually across all three regions.
* Footfall is more for buyers under 'Other' Region and majority of buyers(75%) are spending less than around 3K across all three regions.

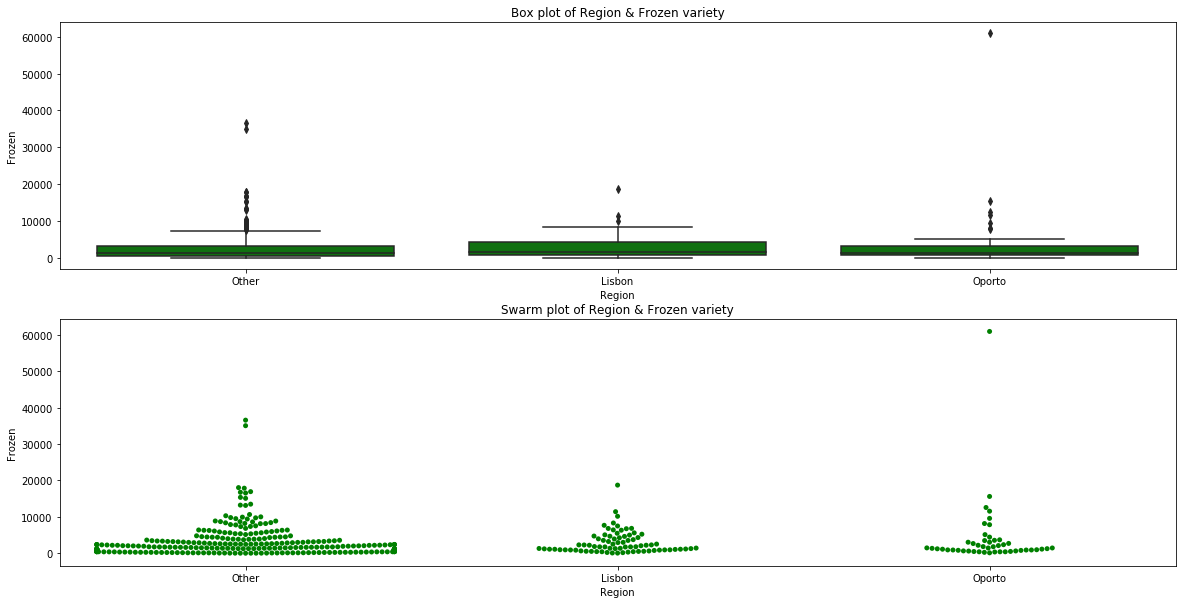




As seen in the summary statistics and swarm plot/box plot,

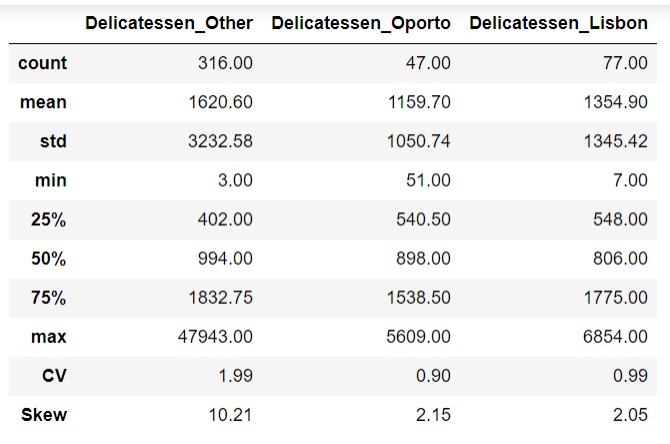
* Data contains more retailers in Other region as compared to Lisbon and Oporto.
* All three regions Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
* Maximum annual spending in 'Other' region is very high as compared to Lisbon/Oporto regions.
* Minimum annual spending in 'Other' region is low as compared to other Lisbon/Oporto regions.
* Annual median spend of Oporto region is highest(6114) as compared to Lisbon(3838) and Other region(4732)
* Volatility of other region is highest among all i.e. it is the most inconsistent region.
* Spread of data looks similar across all regions with distribution being right/positive skewed and 75% of retailers spending less than 11.7K annually across all three regions.
* Footfall is more for buyers under 'Other' Region and majority of buyers(75%) are spending less than around 10-11K across all three regions.

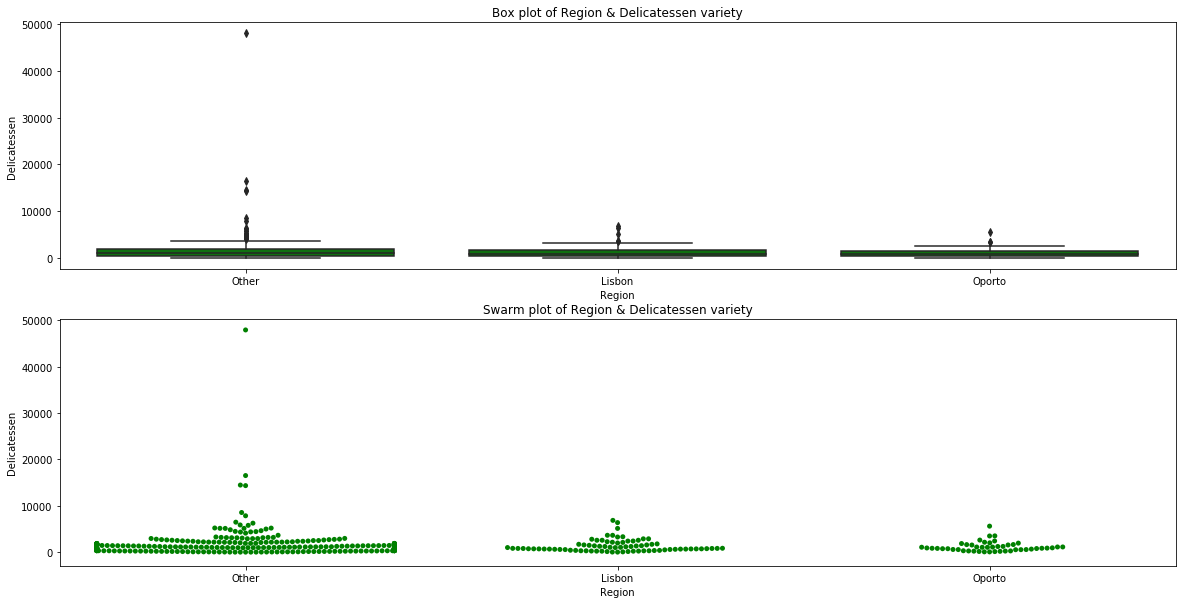




As seen in the summary statistics and swarm plot/box plot,

* Data contains more retailers in Other region as compared to Lisbon and Oporto.
* All three regions Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
* Maximum annual spending in 'Oporto' region is very high as compared to Lisbon/Other regions.
* Minimum annual spending in 'Other' region is low as compared to other Lisbon/Oporto regions.
* Annual median spend of Lisbon region is highest(1801) as compared to Oporto(1455) and Other region(1498)
* Volatility of 'Oporto' region is highest among all i.e. it is the most inconsistent region.
* Spread of data looks similar across all regions with distribution being right/positive skewed and 75% of retailers spending less than 4.3K annually across all three regions.
* Footfall is more for buyers under 'Other' Region and majority of buyers(75%) are spending less than around 3-4K across all three regions.





As seen in the summary statistics and swarm plot/box plot,

Data contains more retailers in Other region as compared to Lisbon and Oporto.

 All three regions Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.

 Maximum annual spending in 'Other' region is very high as compared to Lisbon/Oporto regions.

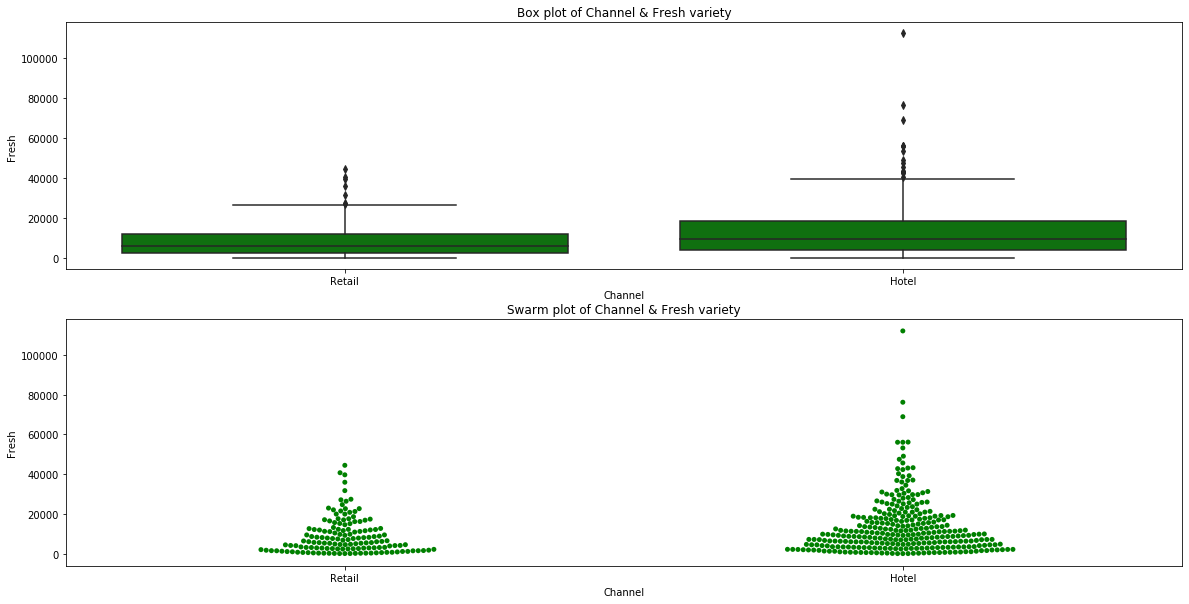
 Minimum annual spending in 'Other' region is low as compared to other Lisbon/Oporto regions.

 Annual median spend of 'Other' region is highest(856) as compared to Oporto(811) and Lisbon region(737)

 Volatility of 'Oporto' region is highest among all i.e. it is the most inconsistent region.

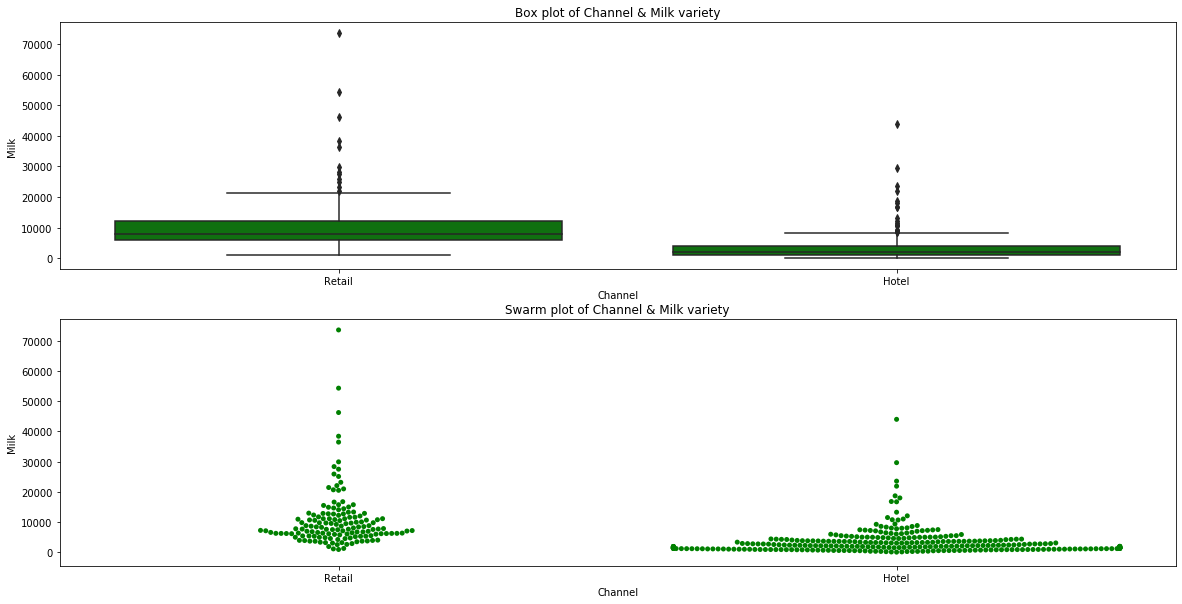
 Spread of data looks similar across all regions with distribution being right/positive skewed and 75% of retailers spending less than 4.3K annually across all three regions.

 Footfall is more for buyers under 'Other' Region and majority of buyers(75%) are spending less than 3-4K annually across all three regions.



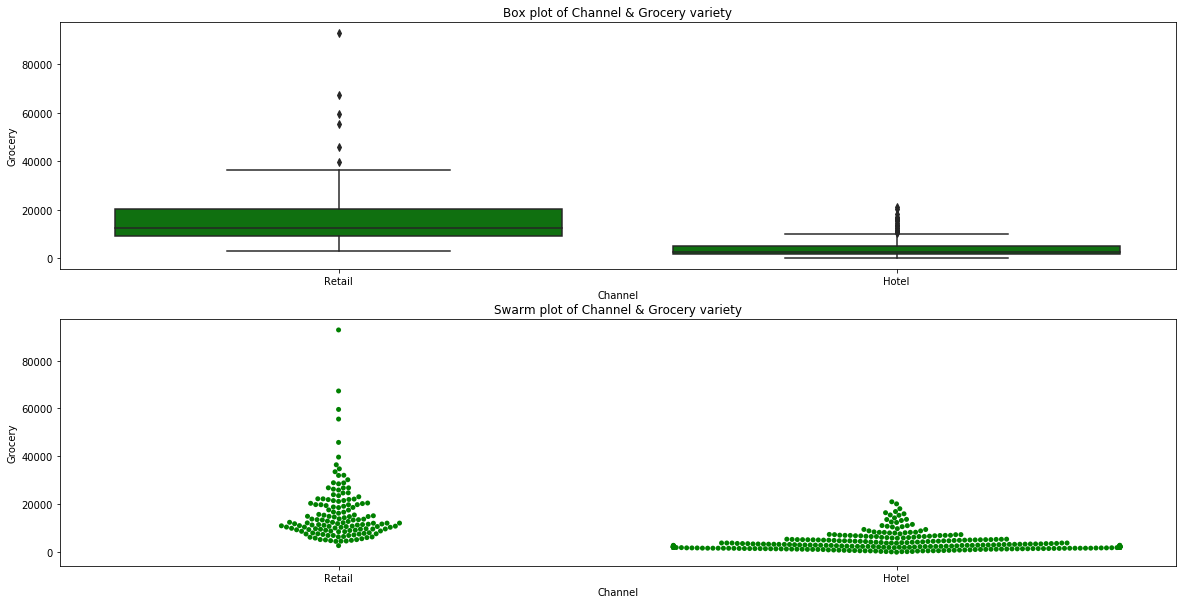
As seen in the summary statistics and swarm plot/box plot,

* Data contains more retailers in Hotel channel as compared to Retail channel
* Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
* Maximum annual spending in 'Hotel' Channel is high as compared to Retail Channel
* Min annual spending in 'Hotel' Channel is low as compared to Retail Channel
* Annual median spend of Hotel channel is higher(9581) than of Retail channel(5993)
* Volatility of both the channels is similar.
* Spread of data looks similar across both channels with distribution being right/positive skewed and 75% of retailers spending less than 18K annually on Hotel channel and 12.2K on Retail channel respectively.
* Footfall is more for buyers of Fresh variety under Hotel channel and buyers are spending much more money on Milk variety under Hotel channel.



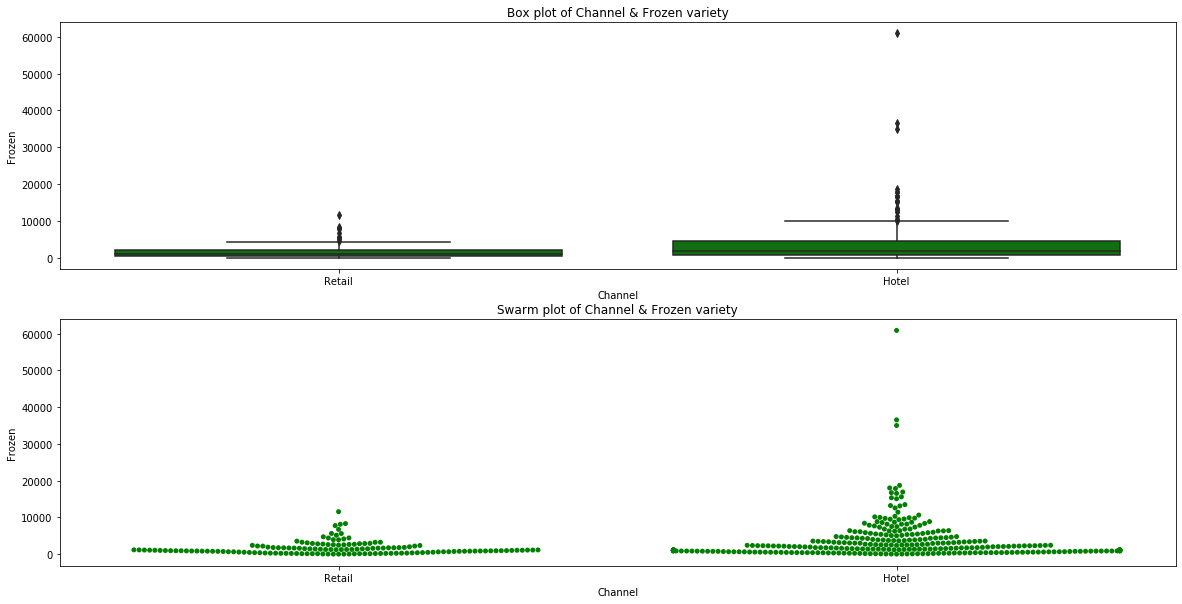
As seen in the summary statistics and swarm plot/box plot,

* Data contains more retailers in Hotel channel as compared to Retail channel
* Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
* Maximum annual spending in 'Retail' Channel is high as compared to Hotel Channel
* Min annual spending in 'Hotel' Channel is low as compared to Retail Channel
* Annual median spend of Hotel channel(2157) is very low as compared to Retail channel(7812)
* Volatility of Hotel channel is higher than of Retail channel.
* Spread of data looks similar across both channels with distribution being right/positive skewed and 75% of retailers spending less than 4K annually on Hotel channel and 12K on Retail channel respectively.
* Although footfall is more for buyers of Milk variety under Hotel channel but buyers are spending more money on Milk variety under Retail channel.



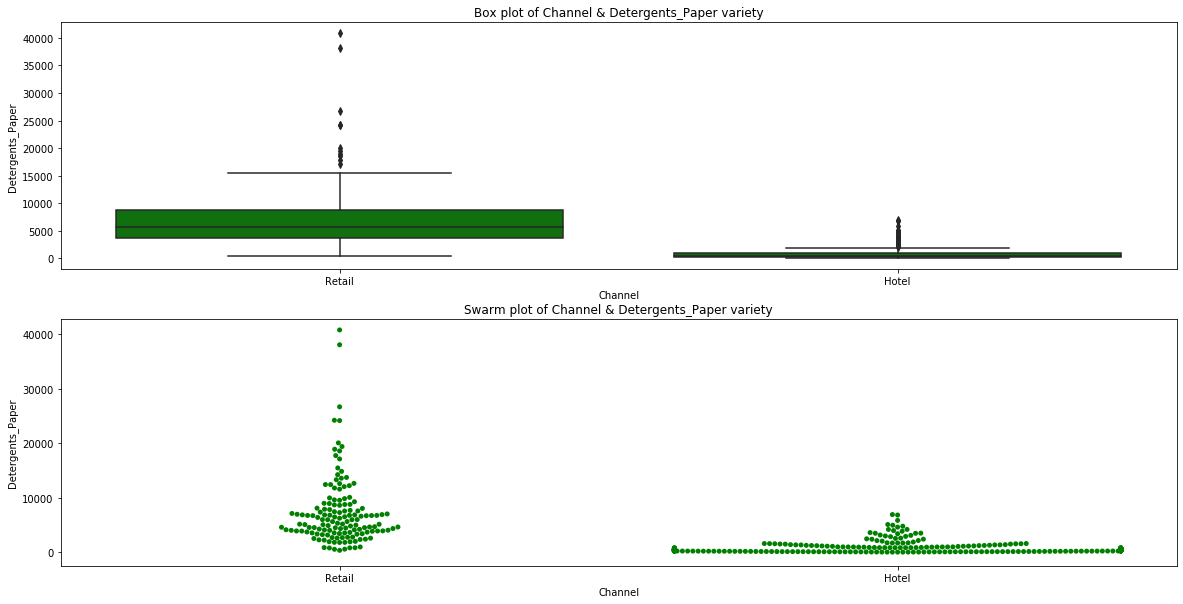
As seen in the summary statistics and swarm plot/box plot,

* Data contains more retailers in Hotel channel as compared to Retail channel
* Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
* Maximum annual spending in 'Retail' Channel is high as compared to Hotel Channel
* Min annual spending in 'Hotel' Channel is low as compared to Retail Channel
* Annual median spend of Hotel channel(2684) is very low as compared to Retail channel(12390)
* Volatility of Hotel channel is higher than of Retail channel.
* Spread of data looks similar across both channels with distribution being right/positive skewed and 75% of retailers spending less than 5K annually on Hotel channel and 20K on Retail channel respectively.
* Although Footfall is more for buyers of Grocery variety under Hotel channel but the buyers are spending more money on Grocery variety under Retail channel.



As seen in the summary statistics and swarm plot/box plot,

* Data contains more retailers in Hotel channel as compared to Retail channel
* Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
* Maximum annual spending in 'Hotel' Channel is high as compared to Retail Channel
* Min annual spending in 'Hotel' Channel is similar compared to Retail Channel
* Annual median spend of Hotel channel(2057) is high as compared to Retail channel(1081)
* Volatility of Hotel channel is higher than of Retail channel.
* Spread of data looks similar across both channels with distribution being right/positive skewed and 75% of retailers spending less than 4.5K annually on Hotel channel and 2.1K on Retail channel respectively.
* The Footfall is more for buyers under Hotel channel and the buyers are spending more money on Frozen variety under Hotel channel as compared to Retail channel.



As seen in the summary statistics and swarm plot/box plot,

 Data contains more retailers in Hotel channel as compared to Retail channel

 Data contains outliers as seen in box plot. Hence, we are using median values for comparison instead of mean.

 Maximum annual spending in 'Retail' Channel is high as compared to Hotel Channel

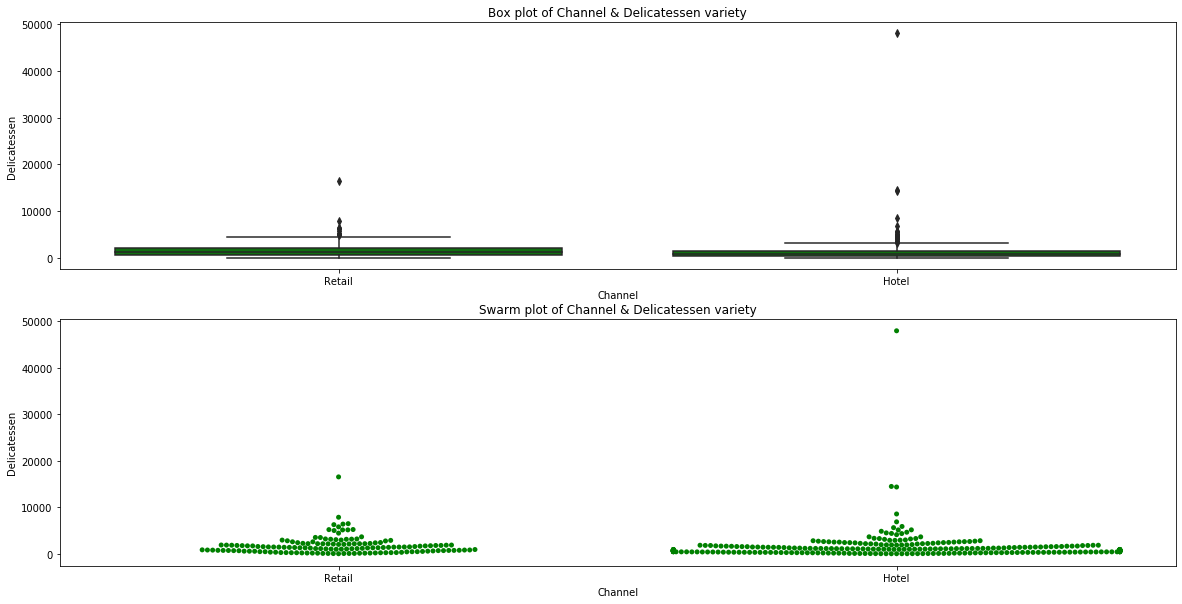
 Min annual spending in 'Hotel' Channel is low as compared to Retail Channel

 Annual median spend of Hotel channel(385) is very low as compared to Retail channel(5614)

 Volatility of Hotel channel is higher than of Retail channel.

 Spread of data looks similar across both channels with distribution being right/positive skewed and 75% of retailers spending less than 0.9K annually on Hotel channel and 8.6K on Retail channel respectively.

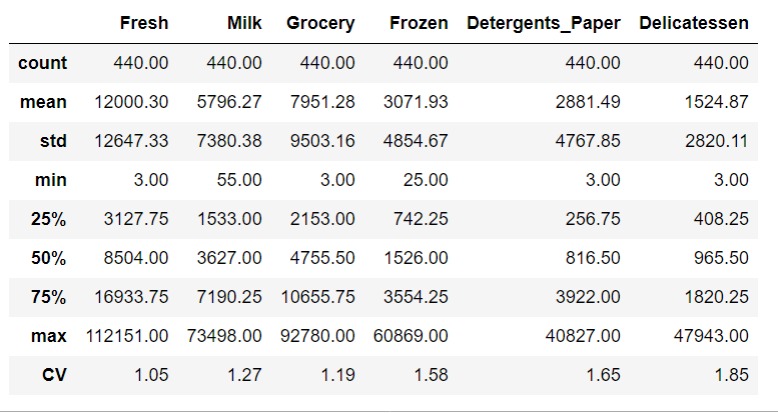
 Although Footfall is more for buyers under Hotel channel but the buyers are spending more money on Detergent Paper variety under Retail channel.



As seen in the summary statistics and swarm plot/box plot,

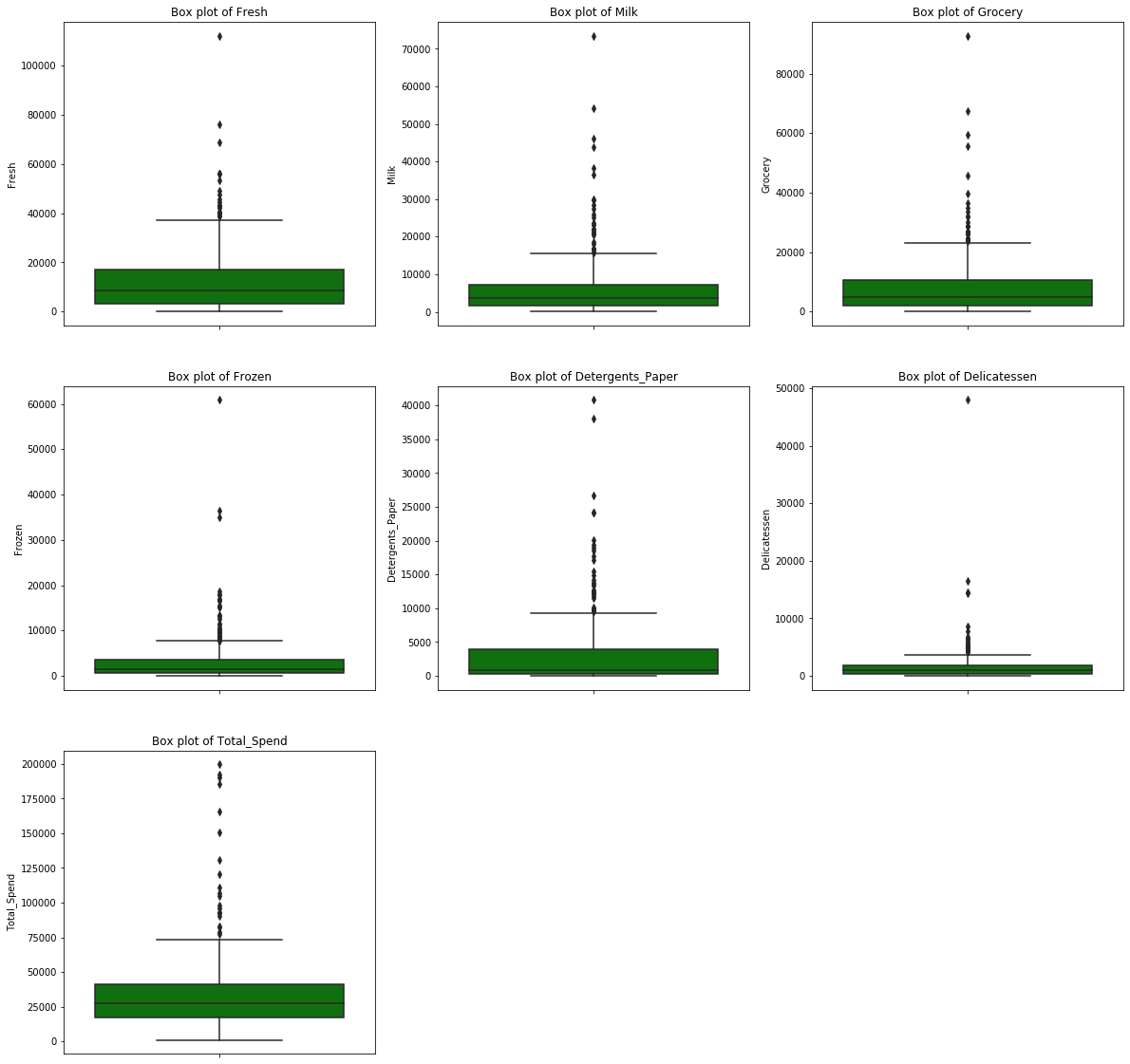
* Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
* Maximum annual spending in 'Hotel' Channel is high as compared to Retail Channel
* Min annual spending in 'Hotel' Channel is similar to Retail Channel
* Annual median spend of Hotel channel(821) is very low as compared to Retail channel(1350)
* Volatility of Hotel channel is higher than of Retail channel.
* Spread of data looks similar across both channels with distribution being right/positive skewed and 75% of retailers spending less than 1.5K annually on Hotel channel and 2.1K on Retail channel respectively.
* Although Footfall is more for buyers under Hotel channel but the buyers are spending more money on Delicateessen variety under Retail channel.

# 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?



# As seen above, item which is showing most inconsistent behaviour is 'Delicatessen' and item which is showing lowest inconsistent behaviour is 'Fresh'

# 1.4 Are there any outliers in the data?



As evident from above box plots, all attributes have outliers. All outliers are on maximum side i.e. there are few retailers which are spending much more than the majority of the retailers. Data appears to be right skewed.

# 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.

###### After analysis , our recommendation for business is :-

Buyers are spending much more on Fresh Products, Grocery and Milk varieties. As seeen in summary 75% of buyers are spending less than 4k on Frozen and Detergents\_Paper. Delicateessen variety is consumed even less as 75% of buyers are

s spending only less than 2K on it. It means Fresh Products, Grocer and Milk varieties are high selling items and bussiness should ensure proper supply of these varieties.

- Delicateessen variety share is only 4.5% of total. May be price of the these items could be very high causing low sales.So Business either need to drop this variety or perhaps needs put in more sales promotion efforts to boost its sale.

- Spending pattern of all 6 items across region appears to be similar

- Spending pattern of 6 items across channels gives us following points

- Fresh and Frozen variety is purchased more via Hotel Channel (annual median around 1.5 to 2 times) as compared to Retail channel

- Milk, Grocery is purchased more via Retail Channel(annual median around 4 to 5 times) as compared to Hotel Channel

- Detergents\_Paper is purchased very very less via Hotel Channel as compared to Retail Channel (annualmedian around 14 times)

- Thus Business should focus on sale of Fresh and Frozen varieties under Hotel Channel and Milk, Grocery and Detergent\_Paper under Retail Channel.

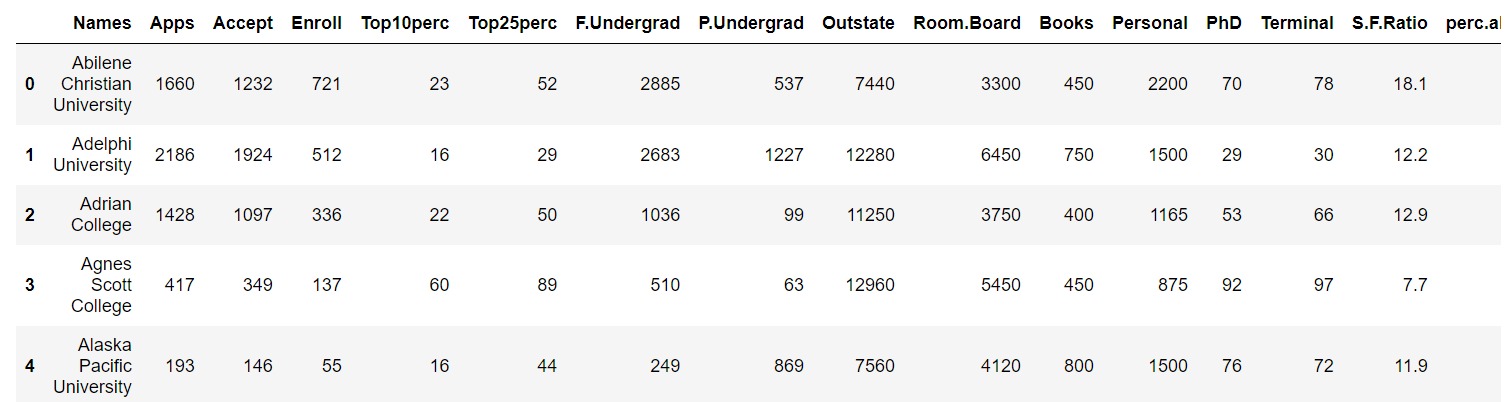
- Since there is a strong correlation between Grocery & Milk, Grocery & Detergent\_Paper, business could give combo offers for these varieties like BOGO scheme etc for sale promotion

- Also clubbing of negative correlated varieties like Detergent\_Paper & Frozen, Detergent\_Paper & Fresh should be avoided.

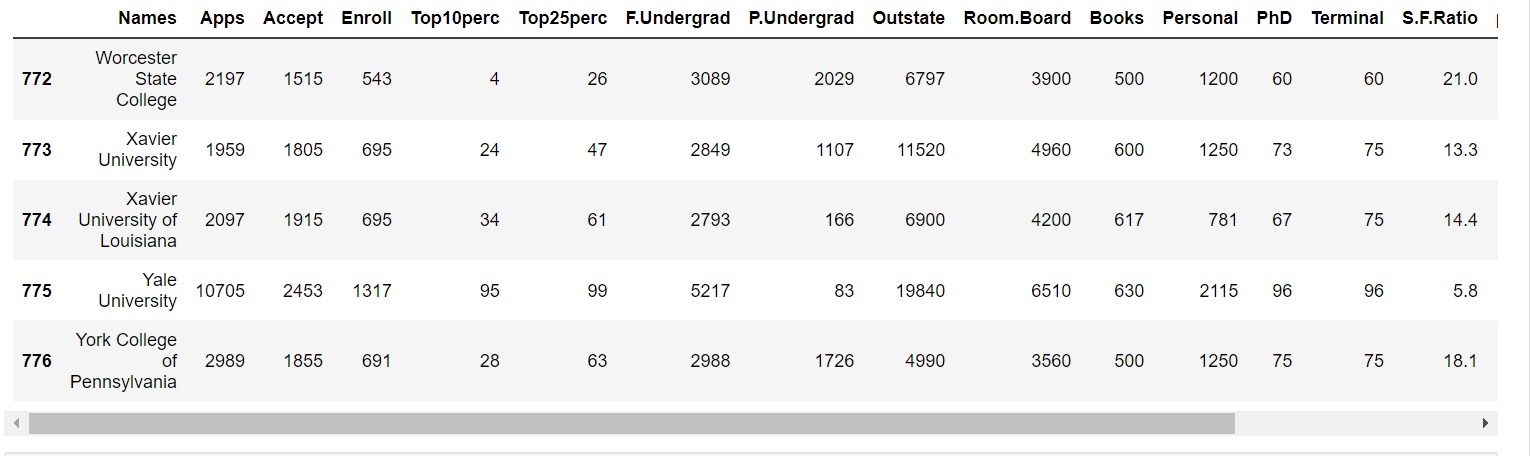
# **Problem 2**

The dataset Education - Post 12th Standard.csv is a dataset which contains the names of various colleges. This particular case study is based on various parameters of various institutions. You are expected to do Principal Component Analysis for this case study according to the instructions given in the following rubric. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: Data Dictionary.xlsx.

**2.1) Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented.**



Described all the heads.



Showed all the tails respectively.

Information’s of particular data

