SMDB

PROJECT

CONTENT.

Problem statement

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.

2. Perform Exploratory Data Analysis [Univariate, Bivariate, and Multivariate analysis to be performed]. What insight do you draw from the EDA?

Problem 1

**Wholesale Customers Analysis**

**Problem Statement:**

A wholesale distributor operating in different regions of Portugal has information on the 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 channels (Hotel, Retail).

**1.1Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?**

1.1 .1 Use methods of descriptive statistics to summarize data.

Table 1: Top five rows of the dataset

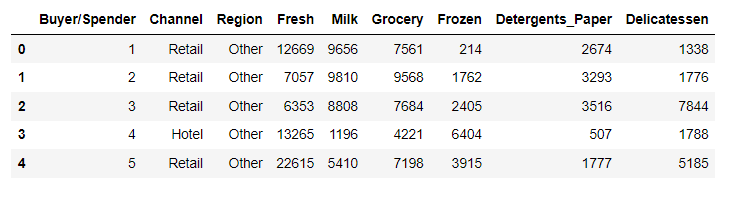


Table 2: Last five rows of the dataset

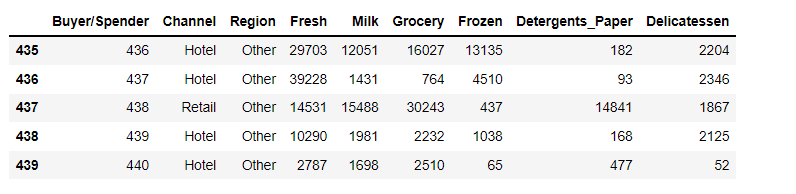
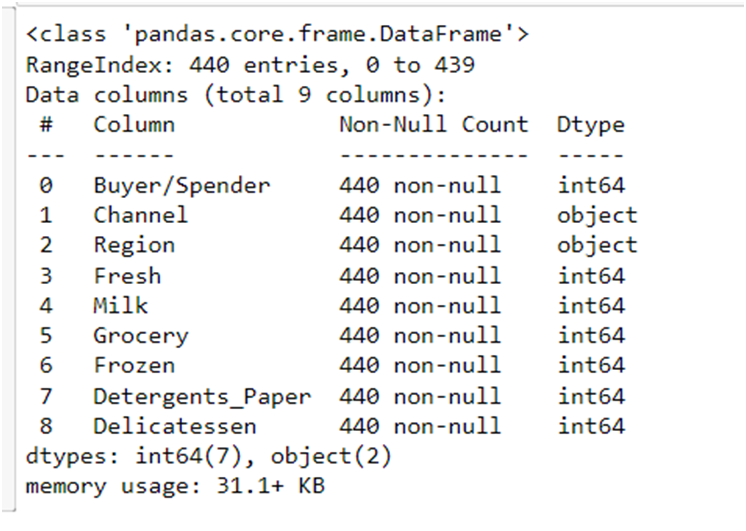


Table 3-Basic information of the dataset

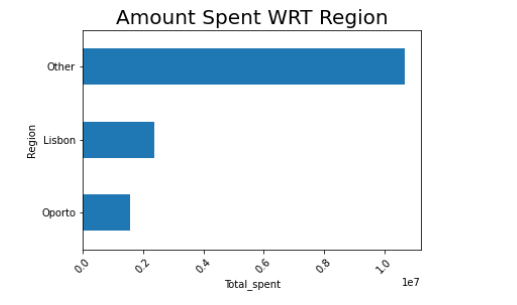


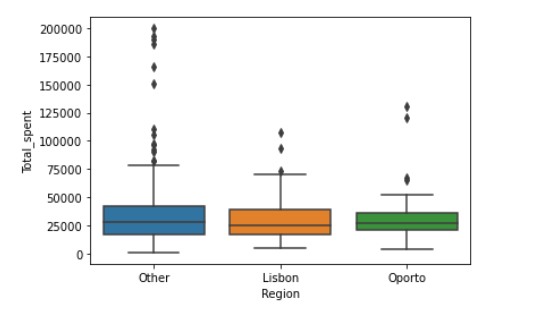
Looking at the dataset we can conclude the following-

* Dataset has 440 rows and 9 columns
* The variables present in the dataset are Buyer/sender, channel, region, fresh, milk, grocery, detergents paper, delicatessen.
  + Out of this only channel and region are categorical variables, others are numerical variables.
* There are no duplicate rows in the dataset.
* There are no null values in the dataset.

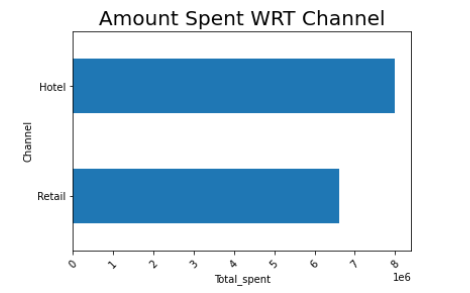
1.1.2 Which Region and which Channel spent the most-

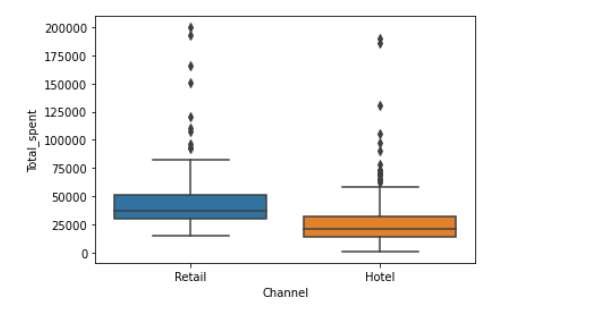
Annual spending of all the items in stores of different regions are as follows-   
Lisbon-2386813  
Oporto-1555088  
Other-10677599





Annual spending of all the items through different channels are as follows-   
Hotel-7999569  
Retail-6619931



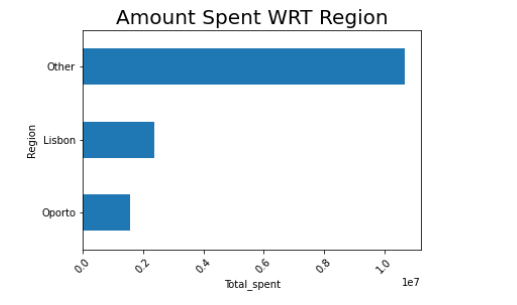


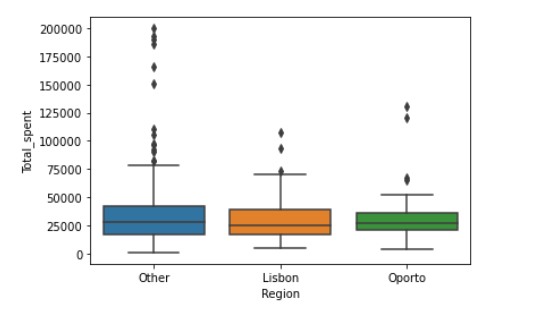
Annual spending of the region that spent the most is Other.

Annual spending of the Channel that spent the most is Hotel.

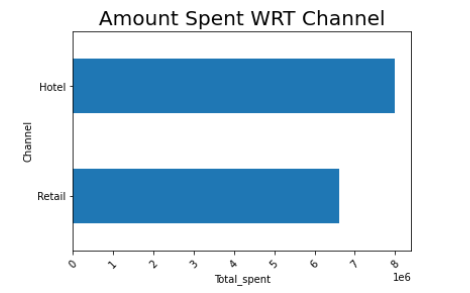
1.1.3 Which Region and which Channel spent the least?

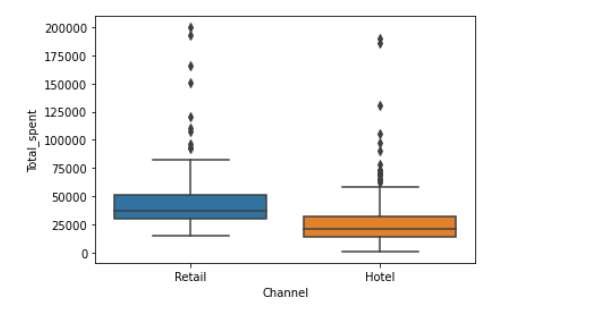
Annual spending of all the items in stores of different regions are as follows-   
Lisbon-2386813  
Oporto-1555088  
Other-10677599





Total spent through different channels.   
Hotel-7999569  
Retail-6619931



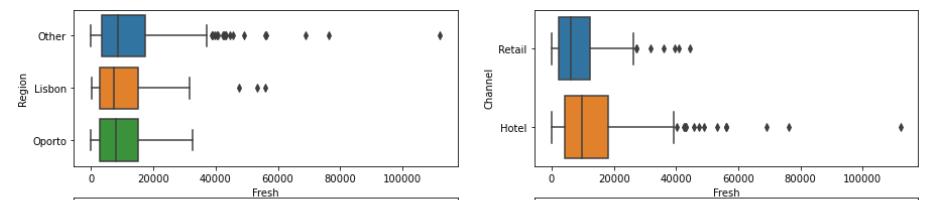


The least Annual spending was by the Oporto region.

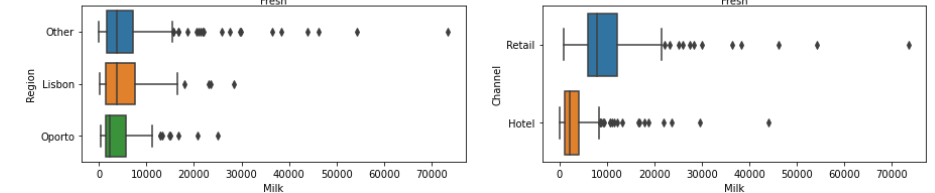
The least annual spending was by the Retail channel.

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.

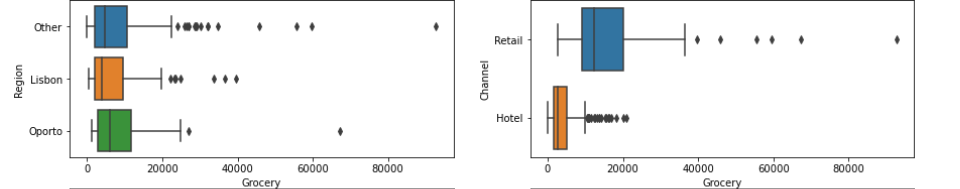
-If we compare the item fresh across Region and channel we can clearly see that annual spent have been more through the region other and through the channel Hotel. The wholesale distributor needs to increase it’s annual spending in the Lisbon region and retail channel.



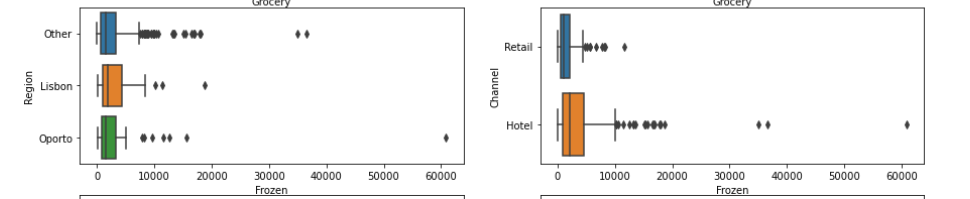
-If we compare annual amount spent on Milk by different channels and regions, we can clearly see that Other regions are the highest spenders on milk . In the channel category the milk sold through retail channel is the highest. We can see outliers in all the variables.



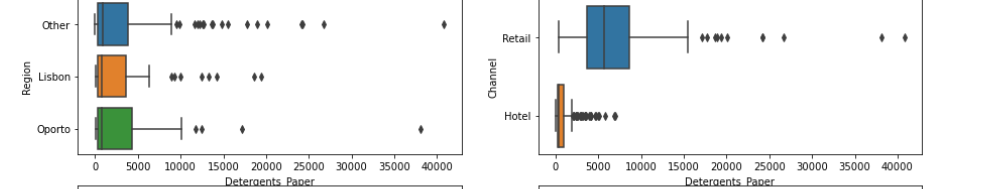
-For the variable Grocery we can see that this time Oporto has the highest number of spent in region category. The annual spent needs to be increased in Lisbon region. Also retail channel has the highest annual spent for groceries, also we can see that there are significant number of outliers in the hotel channel.



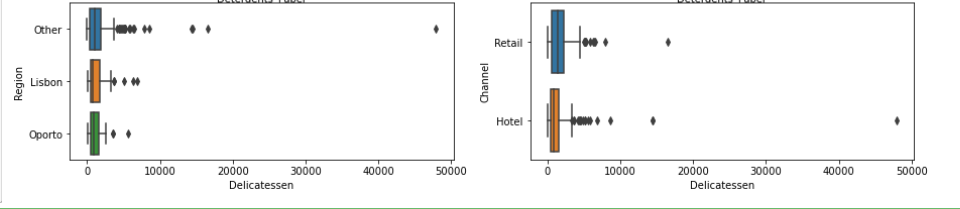
-For the variable Frozen we can clearly see that for Lisbon we have the highest total spent and lowest total spent for Oporto. In the channel medium we can see Hotel channel has the highest annual spent.



-For the variable Detergents Paper the annual spent is highest for Oporto and least for Lisbon. The distributors need to spend more in Lisbon. For channel we can clearly see that Retail has the highest annual spent and we need to increase more annual spending in Hotel channel.



-For the variable Delicatessen Other region has the highest annual spent and Lisbon and Oporto has the almost same amount of total annual spent in the region. For the retail and hotel chain they both have almost same annual spent with retail having a bit higher than hotel.



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?

Standard deviation for all the categories are-

Buyer/Spender 127.16  
Fresh 12647.33  
Milk 7380.38  
Grocery 9503.16  
Frozen 4854.67  
Detergents\_Paper 4767.85  
Delicatessen 2820.11  
Total\_spent 26356.30

Fresh item have highest Standard deviation So that is Inconsistent.

Delicatessen item have smallest Standard deviation, So that is consistent

Now after finding coefficient of variation for all the categories we get following values.

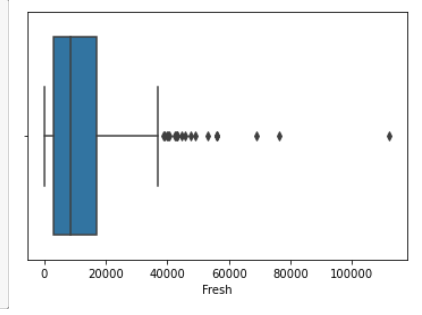
|  |  |
| --- | --- |
| Coefficient of variation |  |
|  |  |
| Fresh | 1.0527196084948245 |
| Milk | 1.2718508307424503 |
| Grocery | 1.193815447749267 |
| Frozen | 1.5785355298607762 |
| Detergents\_Paper | 1.6527657881041729 |
| Delicatessen | 1.8473041039189306 |
|  |  |

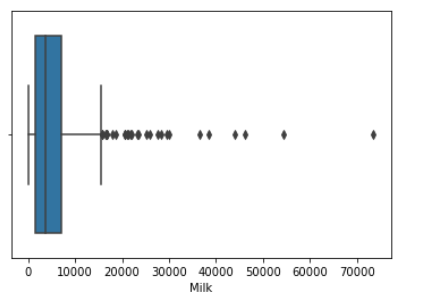
Fresh item have lowest coefficient of Variation So that is consistent.

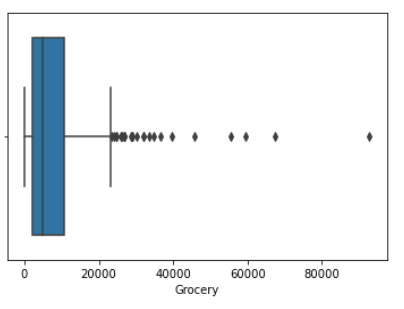
Delicatessen item have highest coefficient of Variation, So that is Inconsistent.

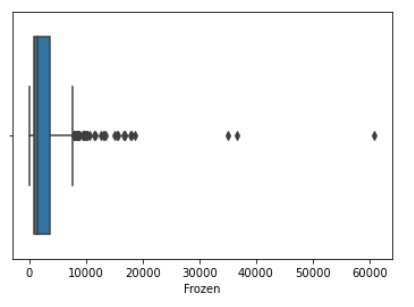
1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments

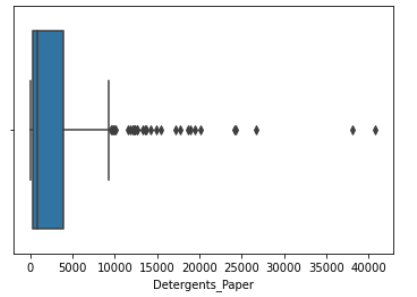
Using the boxplot we can clearly see that there are outliers in all the numerical values of Fresh, Milk, Grocery, Frozen, Detergents\_Paper and Delicatessen.

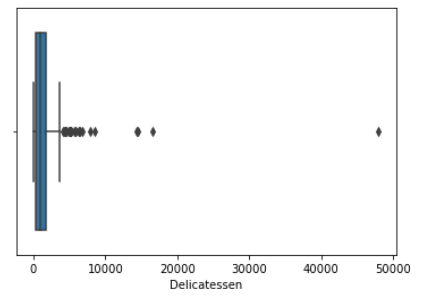












1.5 On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective.

The distributors need to increase their annual spent in in fresh for Lisbon region.

The distributors need to increase their annual spent in fresh for retail chanel.

The distributors need to increase their annual spent in Oporto region and Hotel channel for Milk.

The distributors need to increase their annual spent in Lisbon region and Hotel channel for grocerty

The distributors need to increase their annual spent in in Oporto region and retail channel for frozen.

The distributors need to increase their annual spent in Lisbon region and hotel channel for Detergents paper.

The distributors need to increase their annual spent in Delicatessen for Oporto region and Hotel channel.

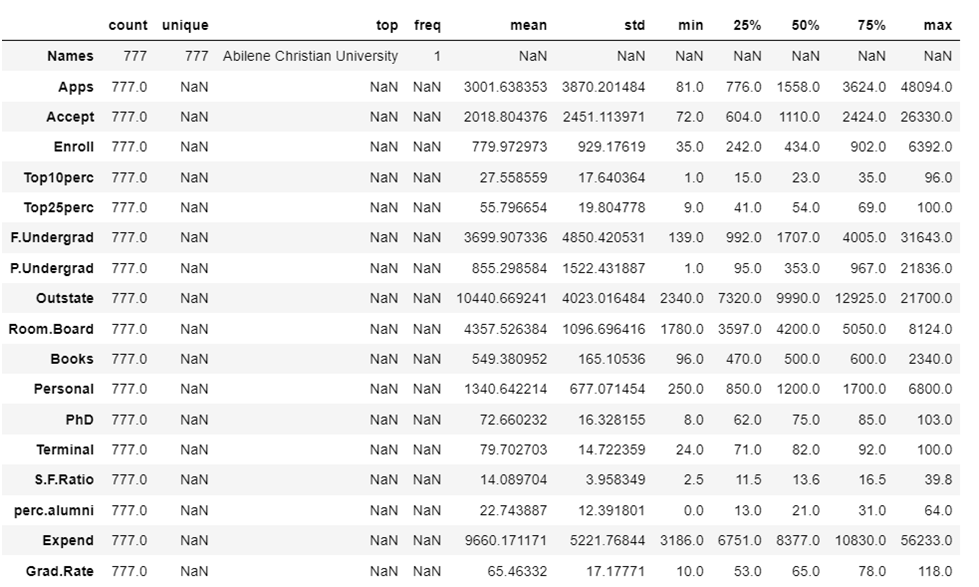
Also sicne the Delicatessen item have highest coefficient of variation it is highly inconsistent, so the distributors can work

**Problem 2:**

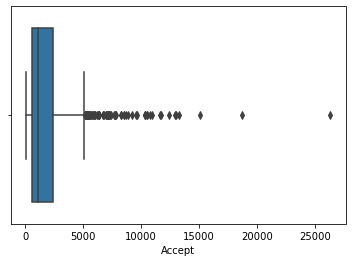
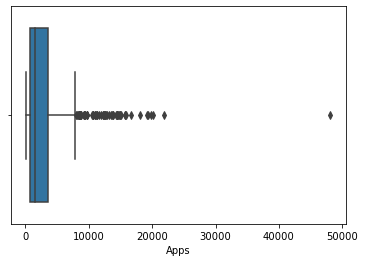
The dataset Education - Post 12th Standard.csv (attached) contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions given. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: Data Dictionary.xlsx (attached)

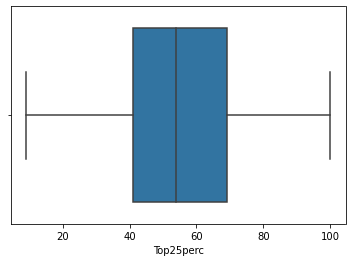
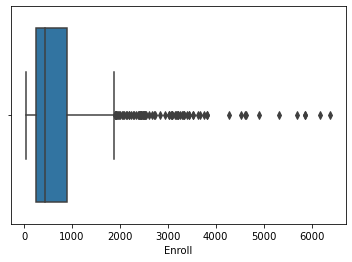
Perform Exploratory Data Analysis [Univariate, Bivariate, and Multivariate analysis to be performed]. What insight do you draw from the EDA?

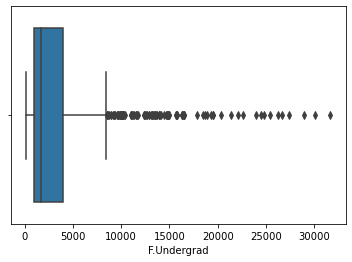
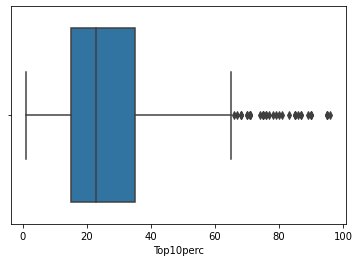
* The dataset contains of 777 Samples/rows and 18 Features/columns
* The data set contains single object, 16 int and single float data-type
* There are no null values in any of the features
* There are no duplicates values in the data set.
* The min, max, mean median standard deviation values are shown in the below figure for all the numerical features.

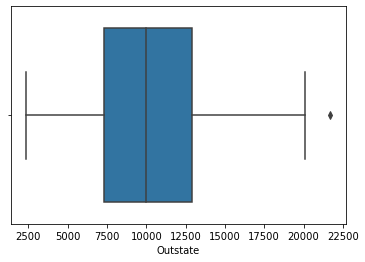
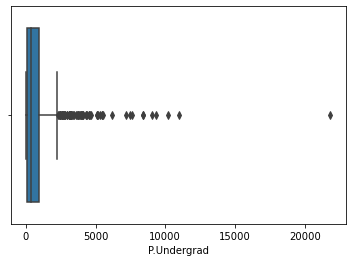


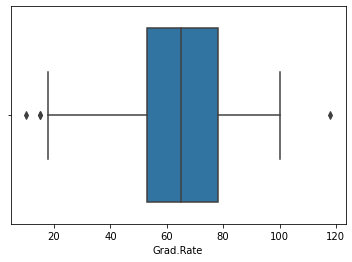
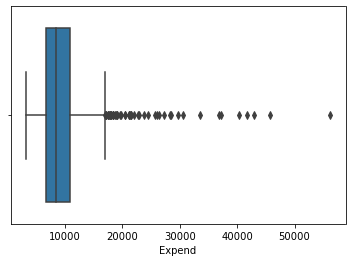
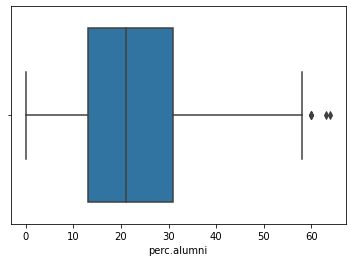
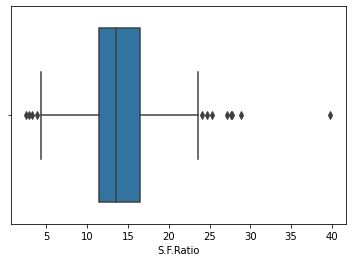
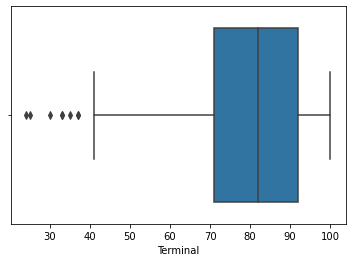
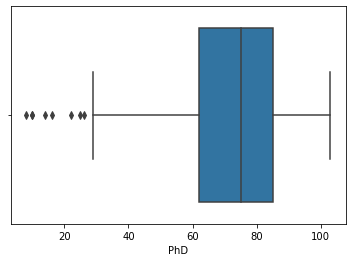
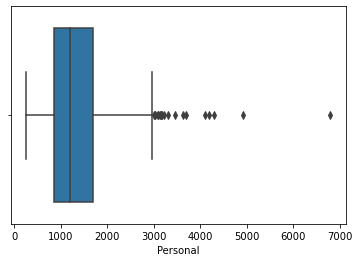
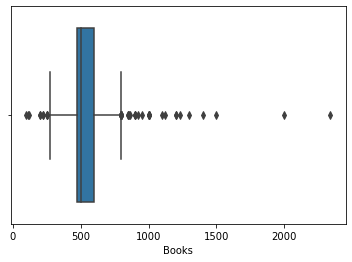
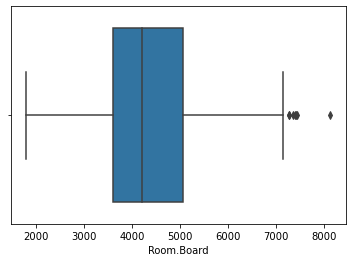
* There are outliers in multiple data set features mentioned below



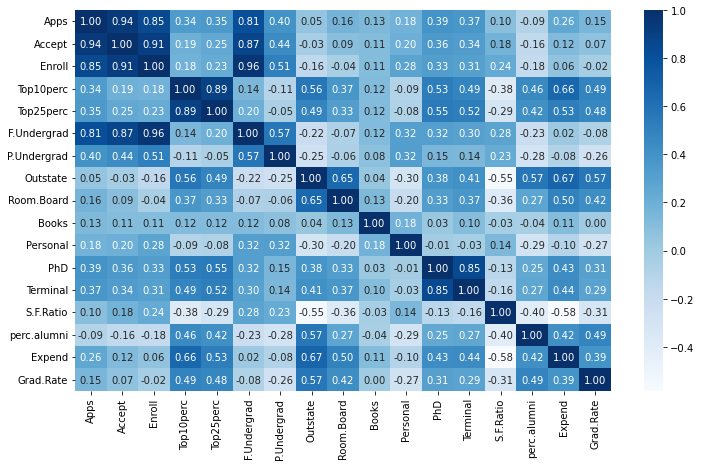




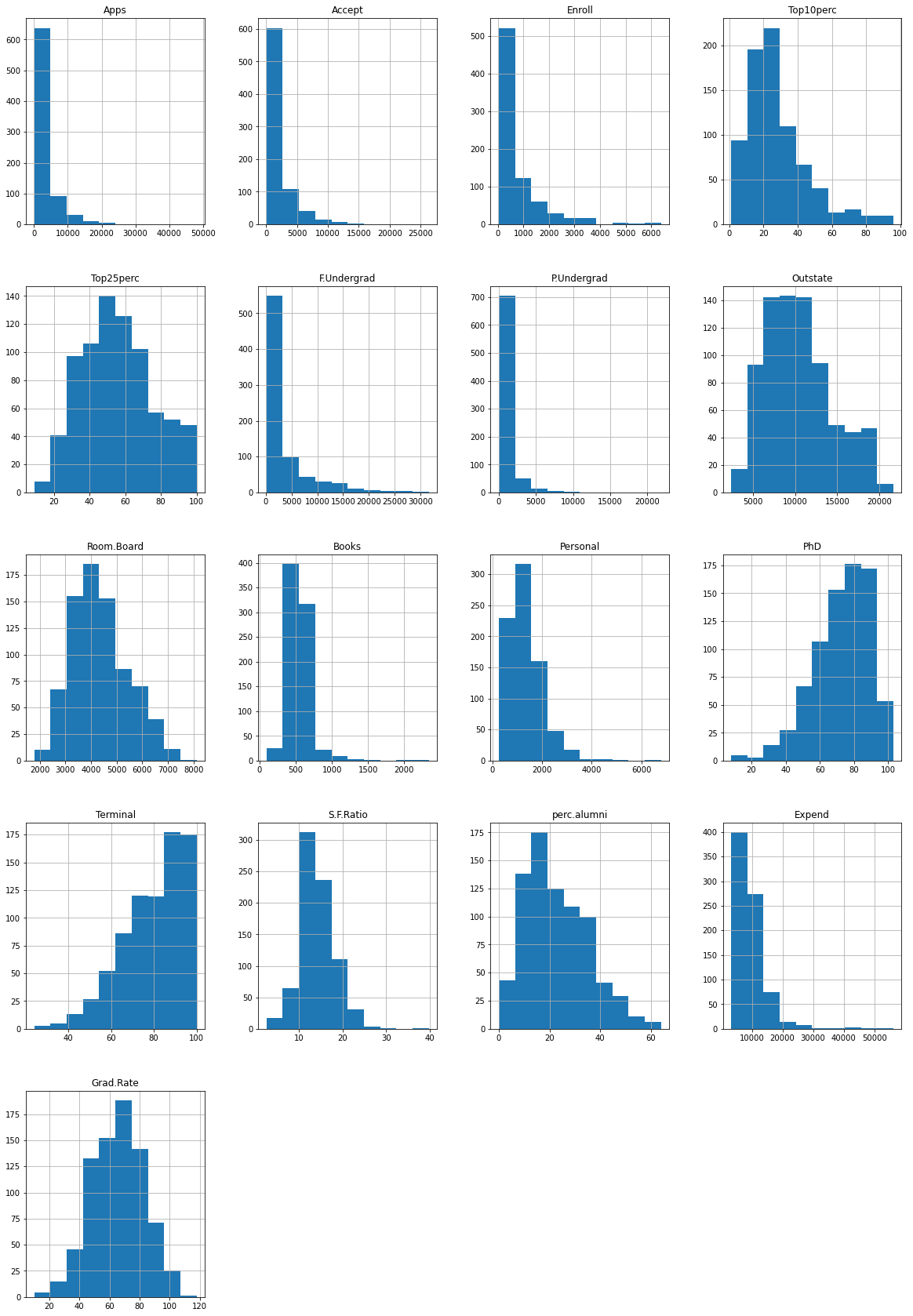




Correlation between the numerical data features-



Univariate

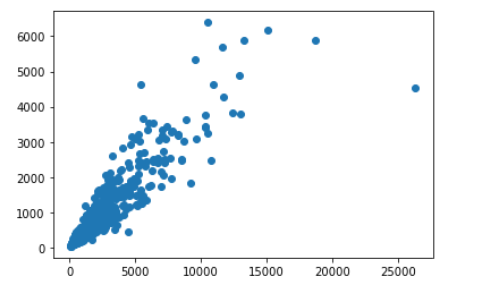


We can clearly see that graphs for Top25perc, Grad.Rate, perc.alumni, S.F ratio, Outstate are almost normal, the other features are highly skewed.

Right skewed graphs-Apps, accept, enroll, Top10perc, F.under, P.under , Books, personal, expend.

Bivariate Analysis.

* Below scatterplot between applications and enrollments show a positive relation, with increasing applications enrollments have increased.



* Below scatterplot shows a positive relationship between Applications received by universities and the number of applications accepted. We can clearly see that there is an outlier as well.

