**Problem 1:**

Context and given data:

The problem is about finding the best compound for the treatment of Hay Fever.

Compound A: Level 1,2 & 3.

Compound B: Level 1,2 & 3.

Volunteers: 1,2,3,4

Relief: Days of Relief after the specific treatment is given to the specific volunteer.

**1.1.)** State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables ‘A’ and ‘B’ individually.

Solution:

Compound A:

The Null hypothesis is that the mean relief time at different levels of the compound A is the same across the volunteers.

H0: MuA1 = MuA2 = MuA3

where, MuA1, MuA2 and MuA3 is the mean relief time of the compound A at levels 1, 2 & 3 respectively.

Alternate hypothesis: The mean relief time for at least one of the levels of the compound A is not the same across the volunteers.

MuA1 = MuA2 ≠ MuA3 (or) MuA1 ≠ MuA2 = MuA3 (or) MuA1 ≠ MuA2 ≠ MuA3

Compound B:

The Null hypothesis is that the mean relief time at different levels of the compound B is the same across the volunteers.

H0: MuB1 = MuB2 = MuB3

Where, MuB1, MuB2 and MuB3 is the mean relief time of the compound B at levels 1,2 & 3 respectively.

Alternate hypothesis: The mean relief time for at least one of the levels of the compound B is not the same across the volunteers.

MuB1 = MuB2 ≠ MuB3 (or) MuB1 ≠ MuB2 = MuB3 (or) MuB1 ≠ MuB2 ≠ MuB3

Interaction effect only to be mentioned for Two Way ANOVA:

Null Hypothesis: The Null hypothesis states that there is no interaction between the various sublevels of compound A (1,2 & 3) with the various sub levels of compound B (1,2 & 3).

The mean relief time of compound A and compound B between its various sublevels is the same.

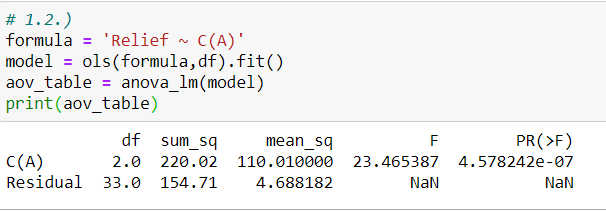
(A1B1 – A1B2 – A1B3) = (A2B1 – A2B2 – A3B3) = (A3B1 – A3B2 – A3B3) = 0

Alternate hypothesis: The Alternate hypothesis states that there is interaction between the various sublevels of compound A (1,2 & 3) with the various sub levels of compound B (1,2 & 3).

(A1B1 – A1B2 – A1B3) = (A2B1 – A2B2 – A3B3) = (A3B1 – A3B2 – A3B3) ≠ 0

[Reference Link: https://stats.stackexchange.com/questions/5617/what-is-the-null-hypothesis-for-interaction-in-a-two-way-anova#comment8475\_5618](Reference%20Link:%20https:/stats.stackexchange.com/questions/5617/what-is-the-null-hypothesis-for-interaction-in-a-two-way-anova%23comment8475_5618)

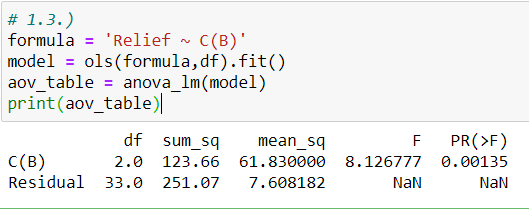
**1.2.)** One Way ANOVA for variable A with respect to the variable Relief:



The value of alpha is not mentioned. Hence, we take it as 5% or 0.05.

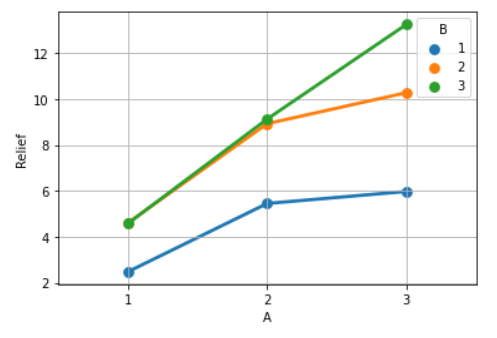
Since the p value is less than alpha, we reject the null hypothesis for compound A and conclude by saying: The mean relief time for at least one of the levels of the compound A is not the same across the volunteers.

1.3.) One Way ANOVA for variable B with respect to the variable Relief:



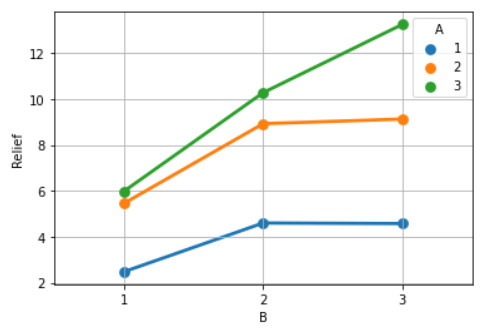
Since the p value is less than alpha, we reject the null hypothesis for compound B and conclude by saying: The mean relief time for at least one of the levels of the compound B is not the same across the volunteers.

**1.4.)** Effects of one variable on another with an Interaction plot:



From the above plot it is seen that at level 1 of the variable A, there is no interaction between level 1 of variable B and the levels 2 and 3 of variable B.

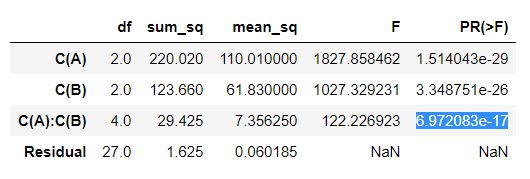
There is good interaction between the levels 2 and 3 of level B. The mean of level 2 and 3 of the variable B is the same at level 1 of the variable A. At level 2 of the variable A, the means of level 2 and 3 of the variable B is almost the same. At level 3 of the variable A, there is no interaction between any of the levels of the variable B.



There is almost no interaction between the different levels of the variable A at levels 2 and 3 of the variable B.

At level 1 of the variable B, the means of the levels 2 and 3 of the variable A are almost the same.

**1.5.)** Two-way ANOVA based on variable ‘A’ & ‘B’:



Here it is seen that the interaction between variables: A and B have a significant effect as the p value of the interaction is less than 0.05. This means that the different levels of A interact with the different levels of B significantly.

**1.6.)** Business implications for ANOVA study:

* It is seen from the interaction plot that the level 3 of the variable B has the greatest positive impact on the relief variable. The relief time is highest in level 3 of the variable B when clubbed with the level 3 of the variable A.
* It is seen from the interaction plot that the level 3 of the variable A has the greatest positive impact on the relief variable. The relief time is highest in level 3 of the variable A when clubbed with the level 3 of the variable B.
* The level 1 of both the variables A and B can be removed from testing because they do not have any significant impact on the relief time. Investment capital can be saved on this and instead invest more into the levels 2 and 3 of both the variables A and B.
* The medicines at the levels 2 of both the variables A and B can be together researched upon to further increase the interaction with an intention of increasing our target variable: Relief Time.
* The level 3 of the variables A and B have the largest relief time and should be considered as a success compound with respect to this study data.

**Problem 2:**

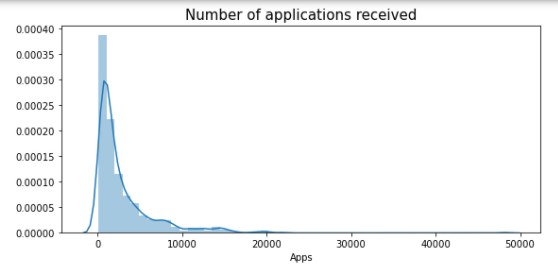
**2.1.)** Exploratory Data Analysis:

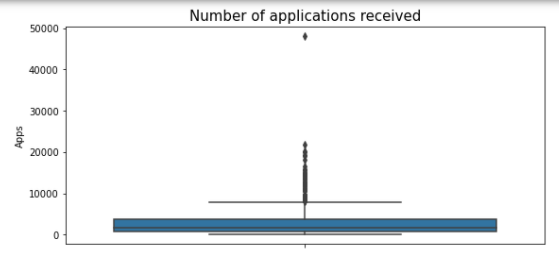
Context and given data: Data contains the information of 777 colleges with the information of various characteristics with respect to education post 12th standard in the United States of America. Dataset contains 18 columns out of which only the Names of universities is the only column containing textual information as opposed to the other 17 columns with numeric data. It is observed that the dataset does not contain any missing values.

Univariate Analysis:

For a univariate analysis, we plot a histogram and a boxplot for each of the numeric column variables:

Apps: Number of applications received:

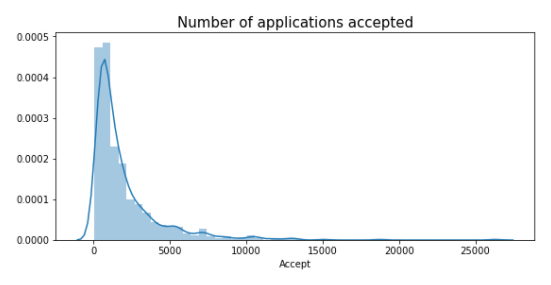




The histogram shows us that the data for the number of applications received is highly right skewed.

The boxplot of the same shows that there are many outliers present in this column beyond the upper whisker.

Accept: Number of applications accepted:

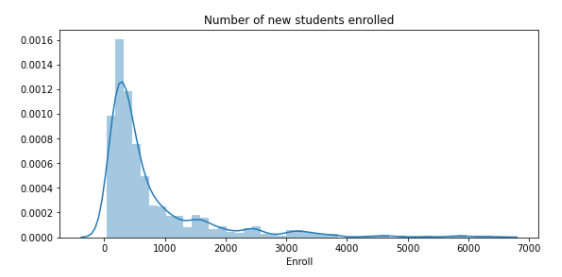


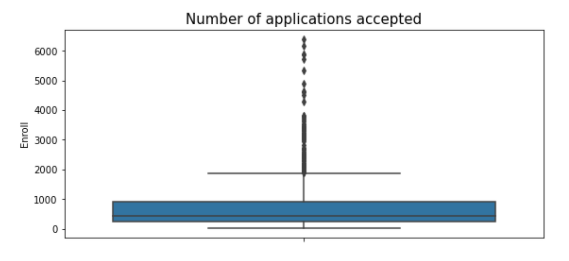


Based on the histogram, we can say that the data for the Accept variable is highly right skewed.

According to the boxplot, it is seen that there are many outliers beyond the upper whisker.

Enroll: Number of new students enrolled:

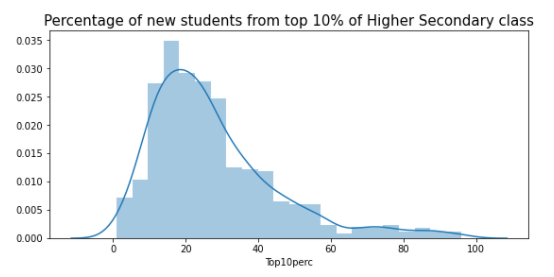


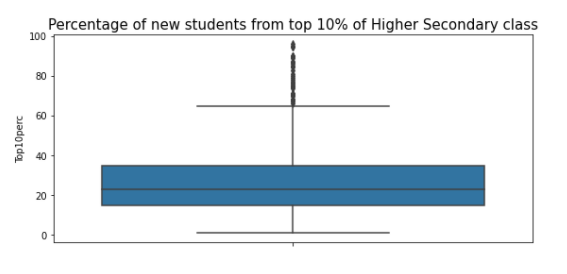


Again, for this variable the histogram shows us that the data is highly right skewed.

The boxplot shows us that there are many outliers present in this data beyond the upper whisker.

Top10perc: Percentage of new students from top 10% of Higher Secondary class:

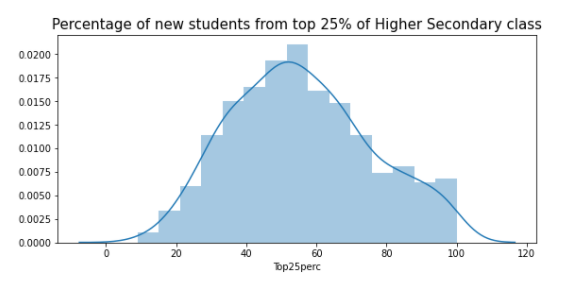


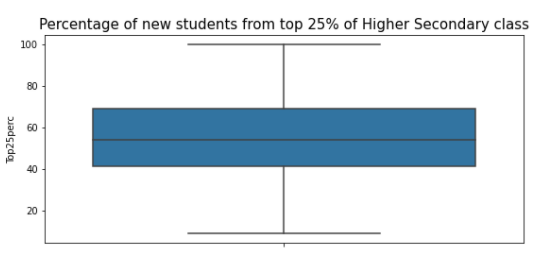


The data is almost normal. The histogram shows us that the data is slightly right skewed.

The boxplot shows us that there are many outliers present in this data beyond the upper whisker. And the fourth quartile is relatively closer to the outliers as compared to the earlier variables.

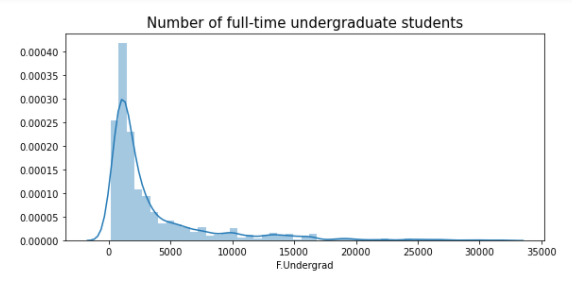
Top25perc: Percentage of new students from top 25% of Higher Secondary class:

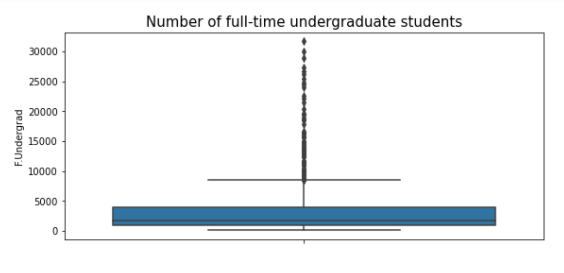




The data is normally distributed. Outliers are not present in this variable.

F.Undergrad: Number of full-time undergraduate students:

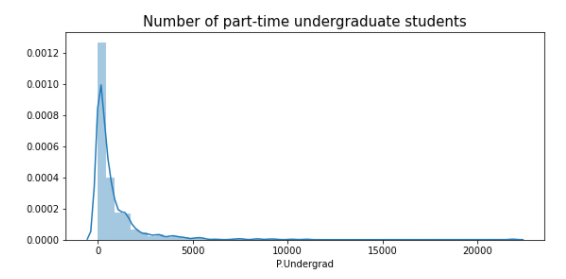


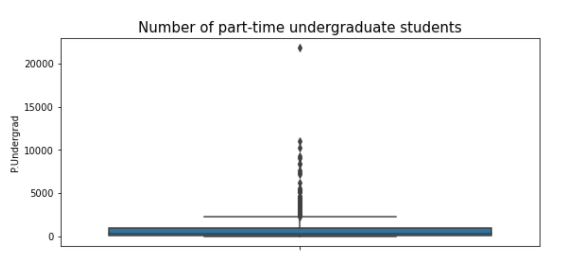


The histogram shows that the data for this variable is highly right skewed.

There are many outliers present in this variable beyond the upper whisker.

P.Undergrad: Number of part-time undergraduate students:

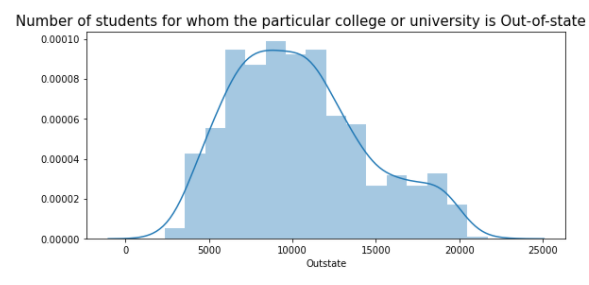


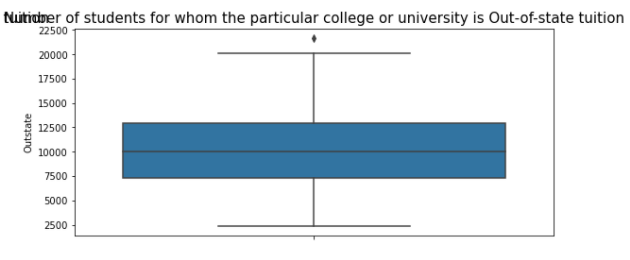


The data is extremely right skewed.

There are outliers present beyond the upper whisker in the boxplot.

Outstate: Number of students for whom the particular college or university is Out-of-state tuition:

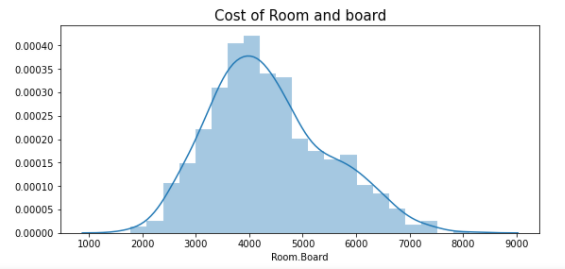


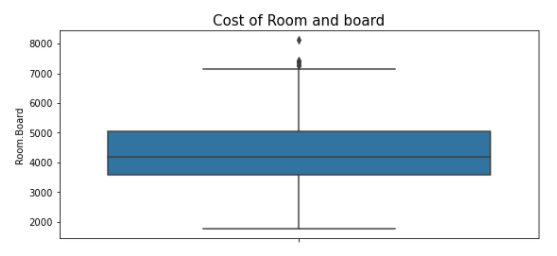


The data distantly resembles a Normal distribution.

Very few outliers are present beyond the upper whisker in the boxplot.

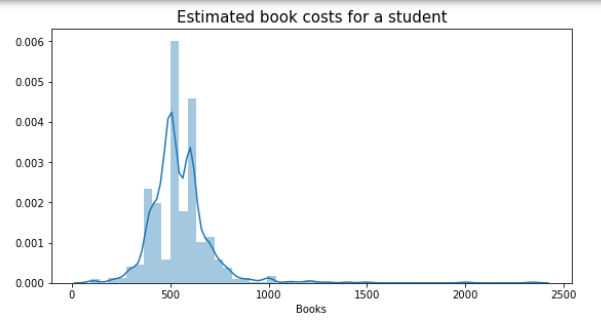
Room.Board: Cost of Room and board:

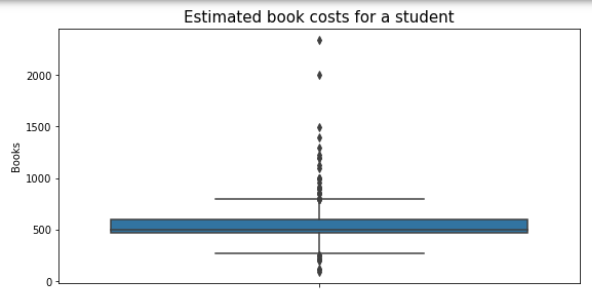




The cost of room and board variable is seen to be a fairly normal distribution with some outliers present beyond the upper whisker of the boxplot.

Books: Estimated book costs for a student:



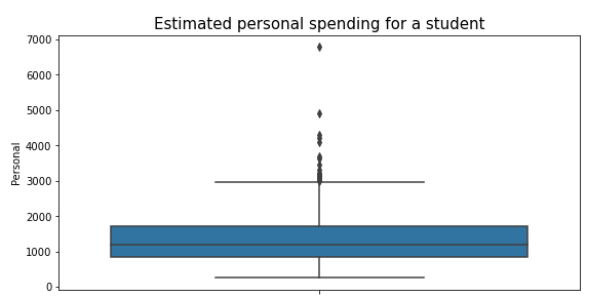


The distribution of the Estimated book costs for a student is right skewed.

It is interesting to note that it is the only variable so far to have outliers present beyond both the whiskers of the boxplot. However, it is seen that the maximum outliers are present beyond the upper whisker.

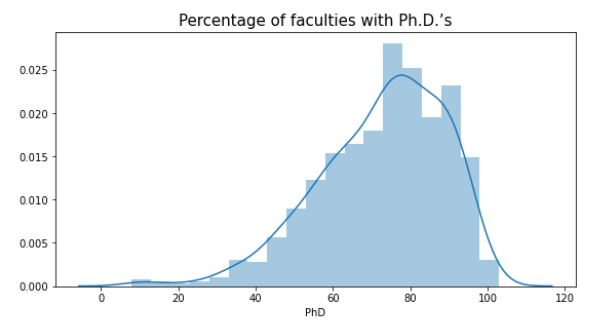
Personal: Estimated personal spending for a student:

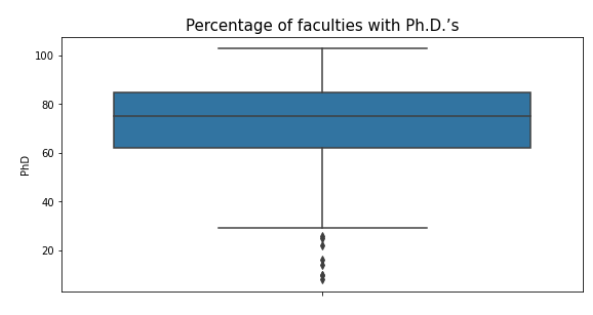




The data for this variable is right skewed in the distribution plot with outliers present beyond the upper whisker of the boxplot.

PhD: Percentage of faculties with Ph.D.’s:

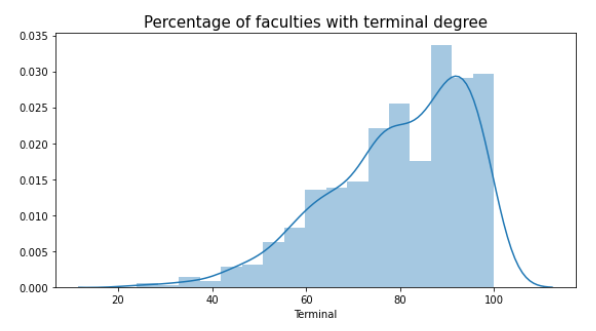


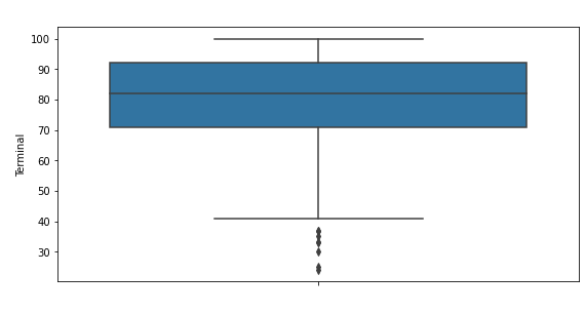


Interestingly, the data for this variable is left skewed with many faculties in the 80 percent section.

There are many outliers present before the lower whisker of the boxplot.

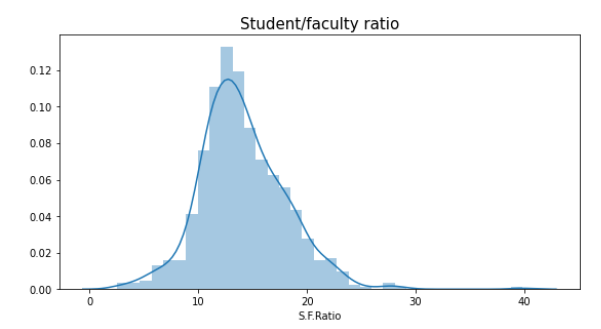
Terminal: Percentage of faculties with terminal degree:

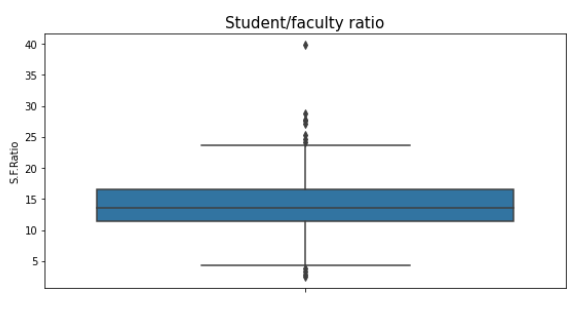




The data for the percentage of faculties with a terminal degree is highly left skewed in the distribution plot with many outliers before the lower whisker of the boxplot.

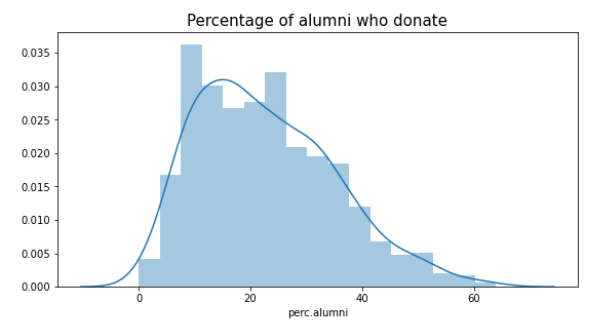
S.F.Ratio: Student/faculty ratio:

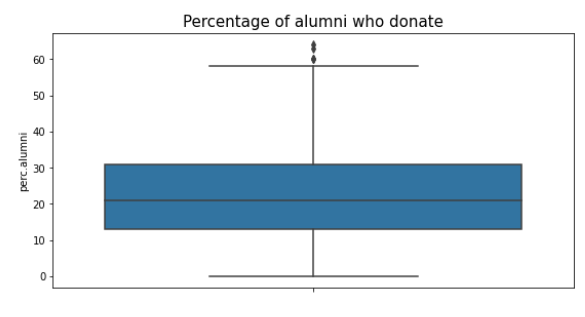




The student/faculty ratio follows a nearly perfect normal distribution with outliers present on either side of the whiskers. It is also interesting to note that the lower outliers are more closer to the lower whisker as compared to the distance between the upper outliers and the upper whisker.

perc.alumni: Percentage of alumni who donate:

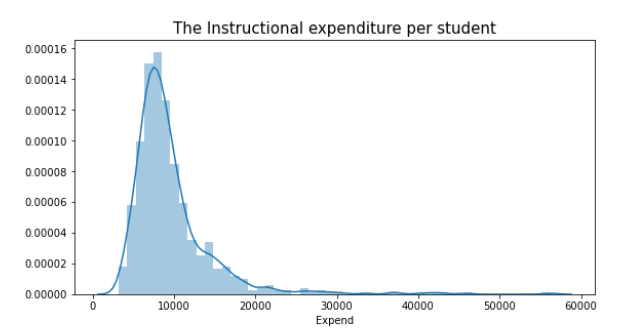


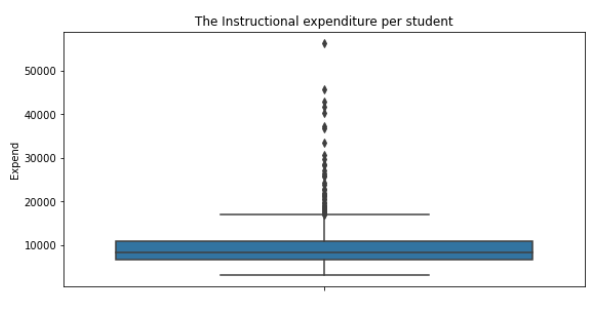


The data for this variable follows an approximate normal distribution with a minor right skewness.

Few outliers also are present beyond the upper whisker of the boxplot.

Expend: The Instructional expenditure per student:

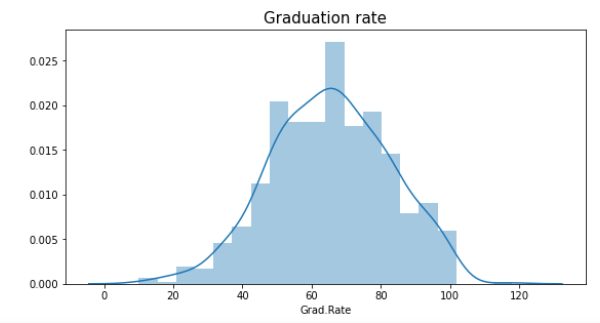


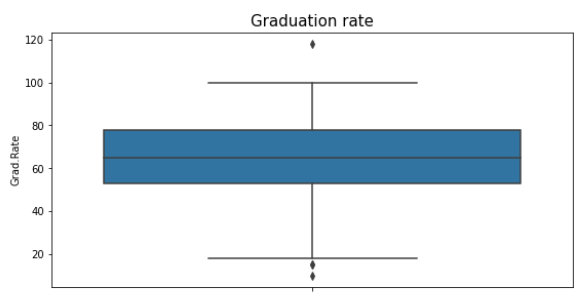


The data is highly right skewed with outliers extending up to the 6000 mark.

The outliers seem to be uniformly present up to the 4500 mark.

Grad.Rate: Graduation rate:





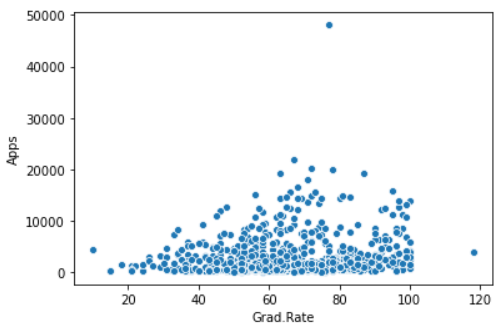
The graduation rate is almost normally distributed with a mild left skewness.

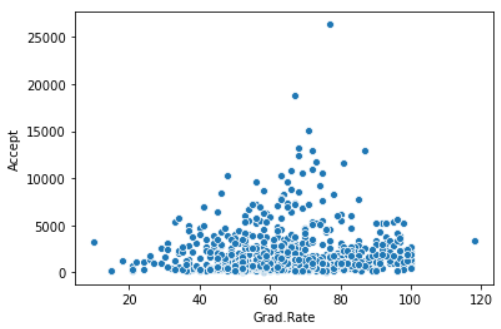
This variable is desired to have a high left skewness with many outliers present beyond the upper whisker.

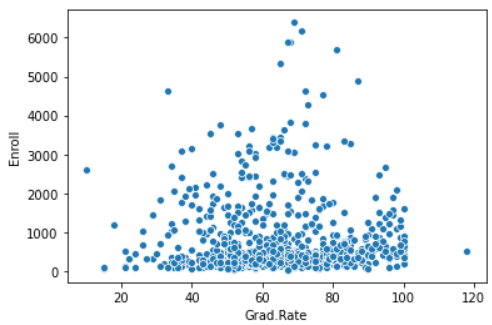
In the data given to us, there are few outliers present of which the ones below the lower whisker outnumber the ones beyond the upper whisker.

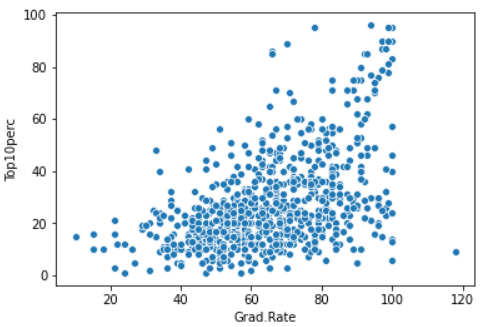
Bivariate Analysis:

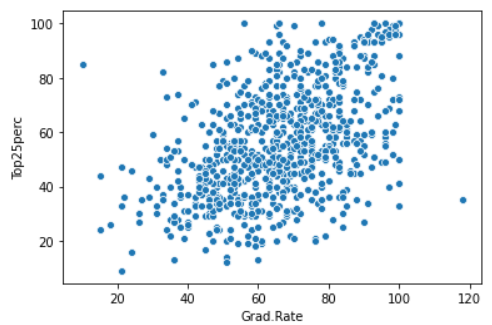
Each variable has been plotted against the variable: Grad.Rate using a scatterplot:

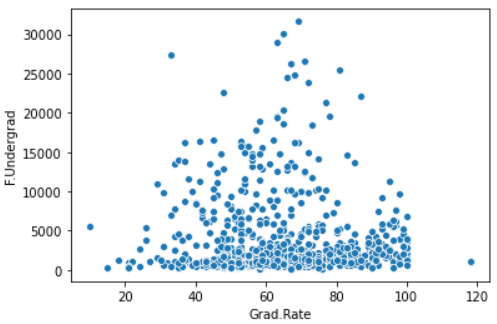


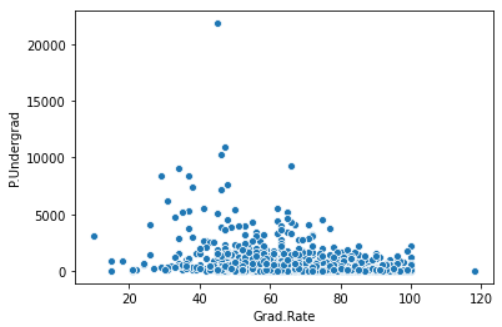


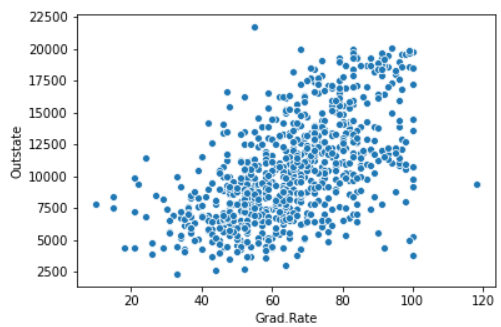


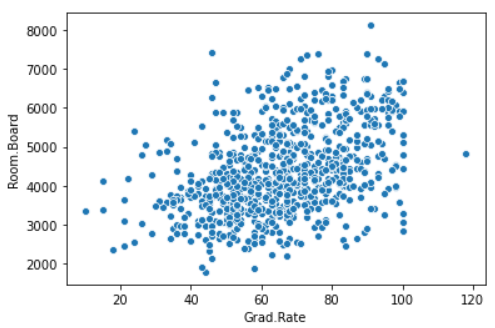


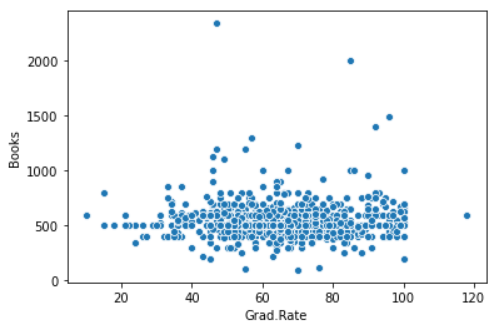


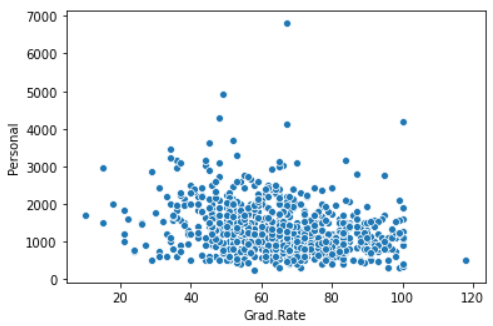


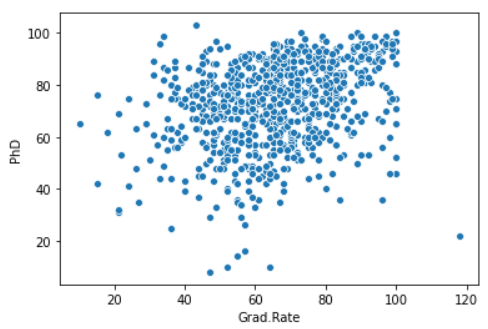


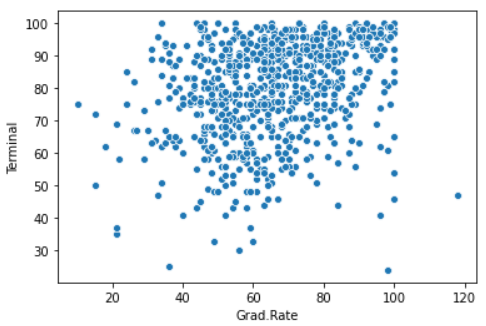


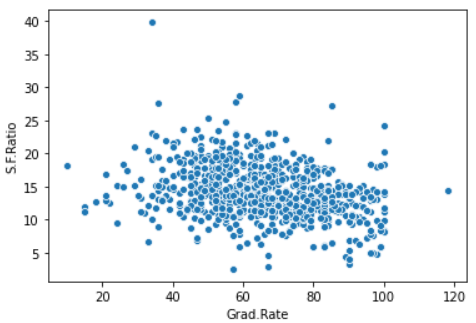


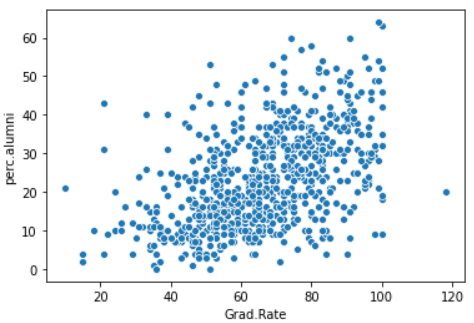


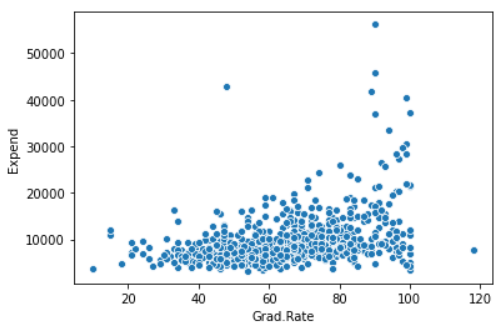




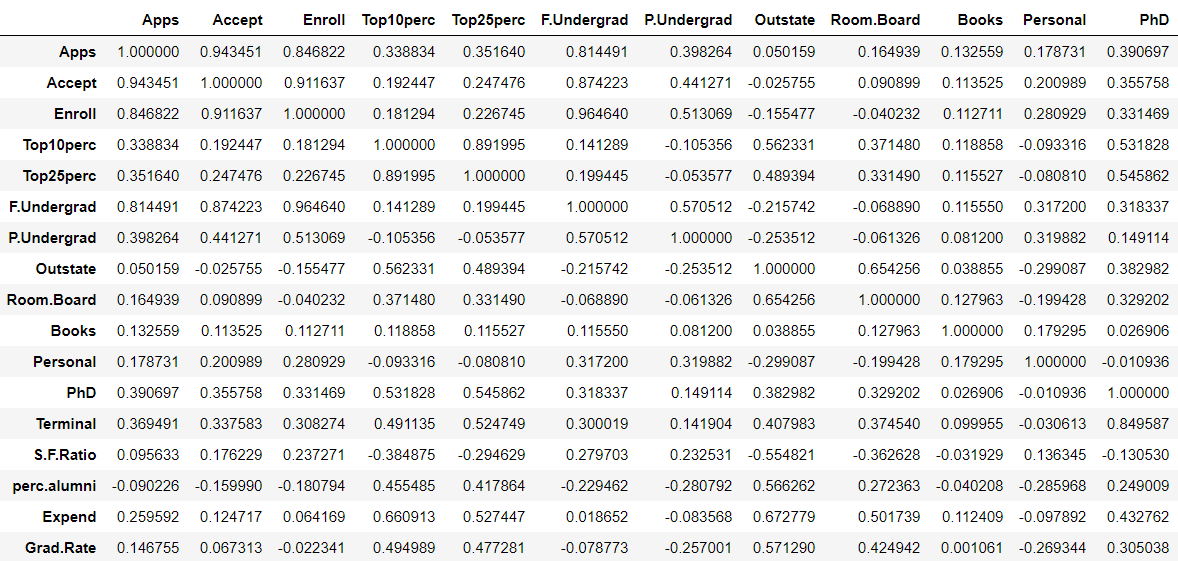


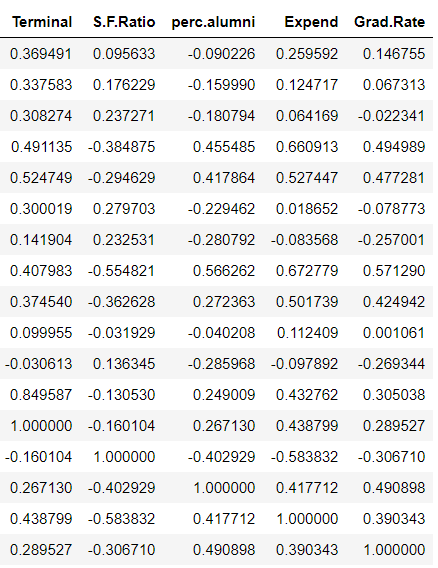
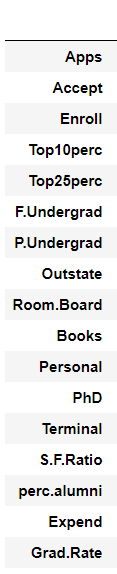




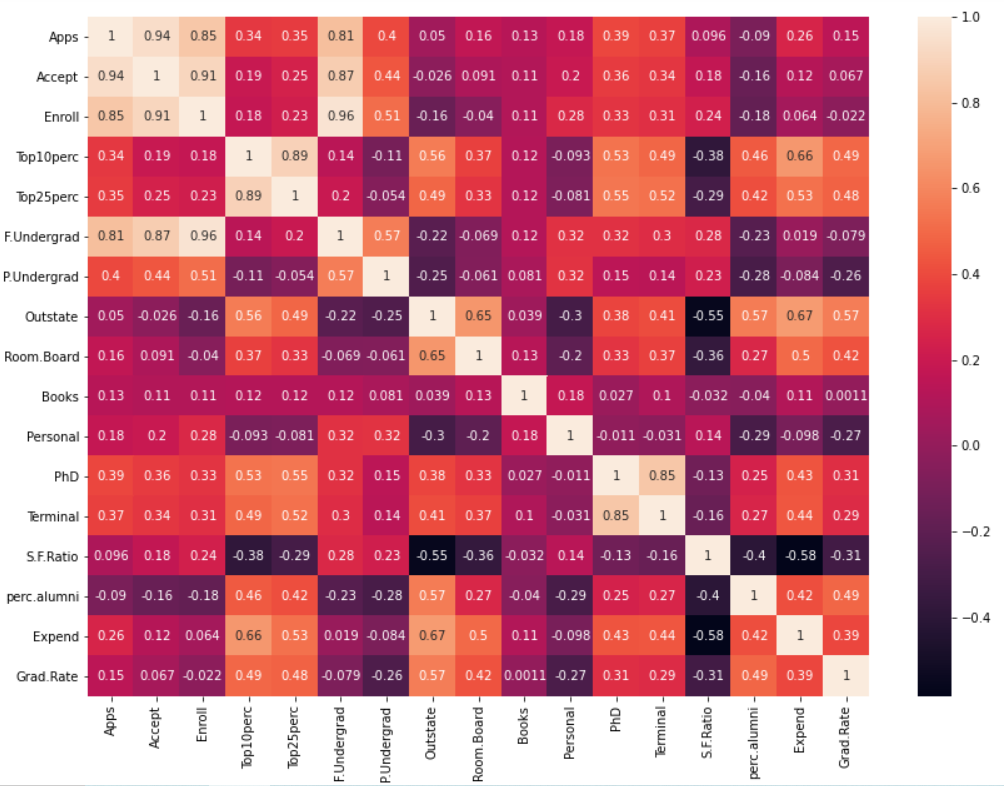


Correlation Plot using the Pearson’s method:

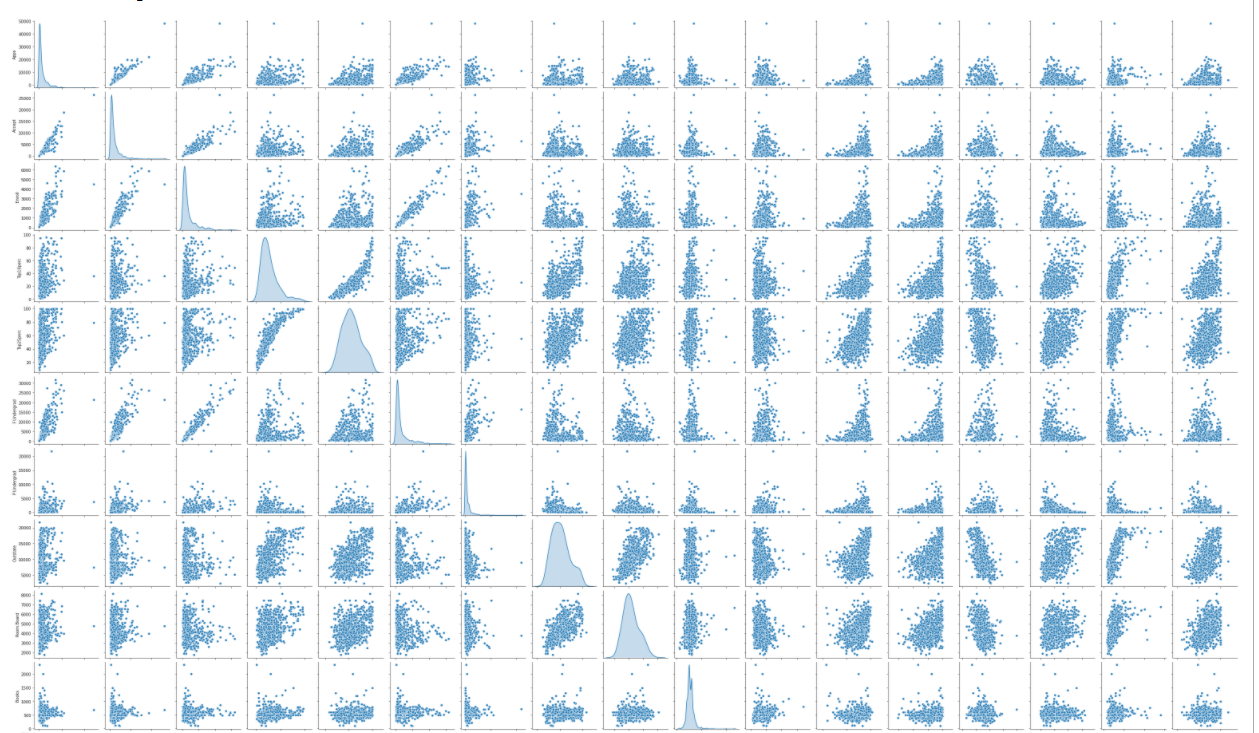


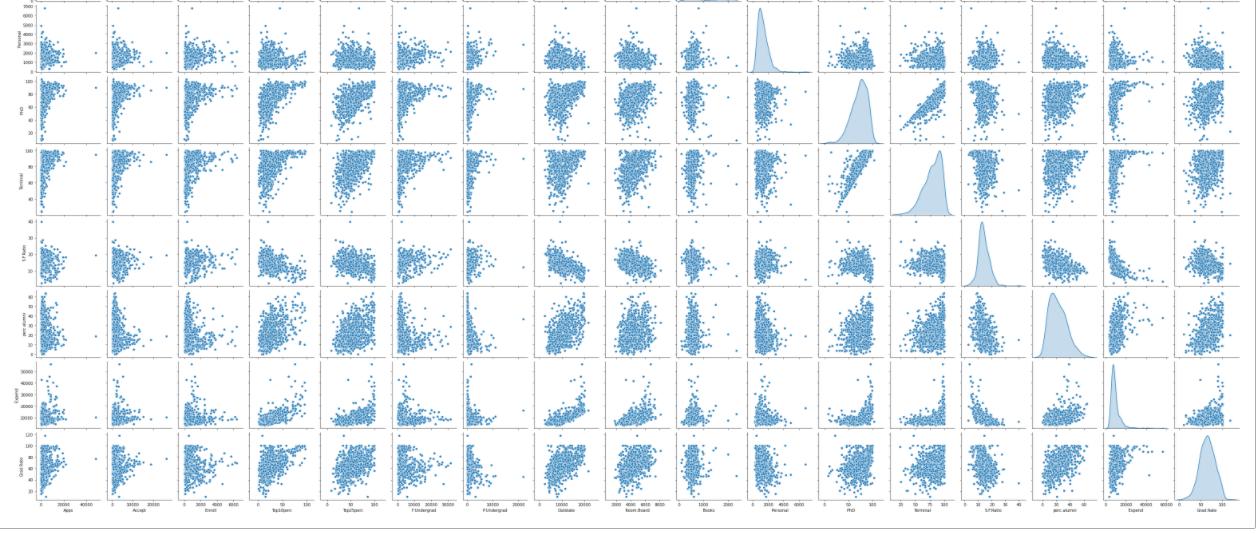


Heatmap for the same:



Pairplot for each variable:

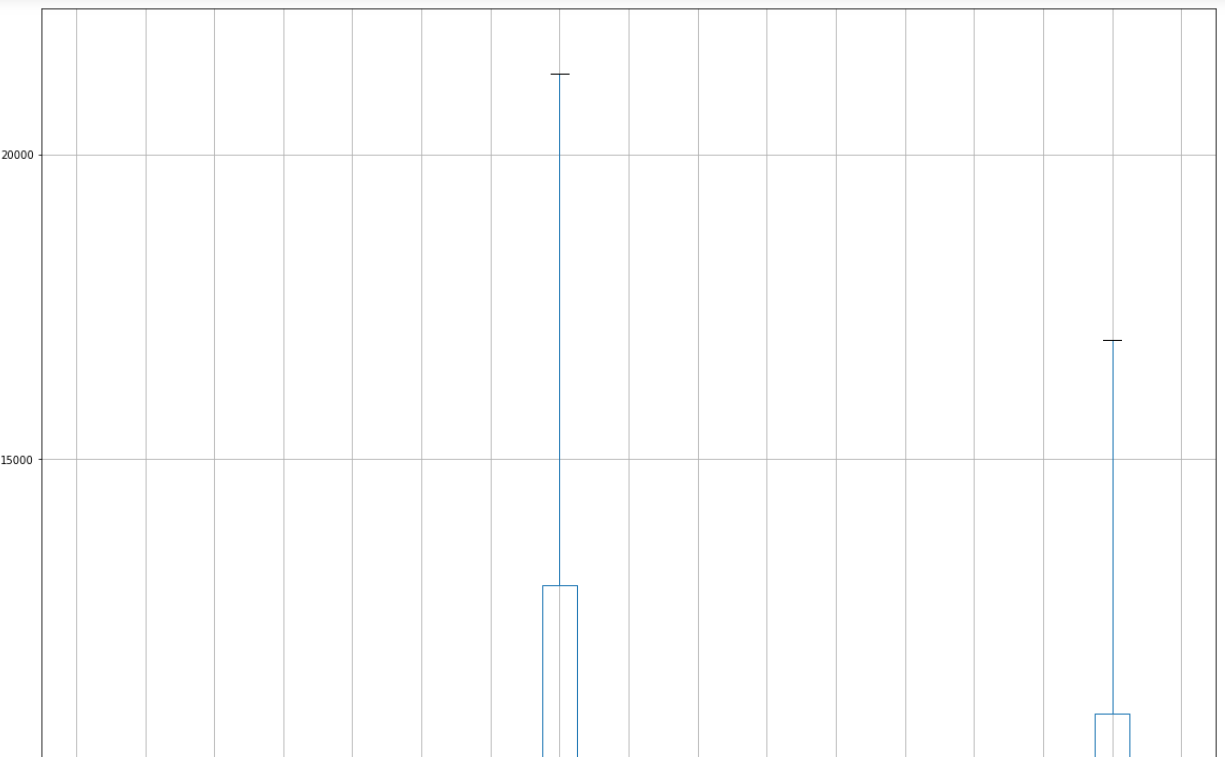


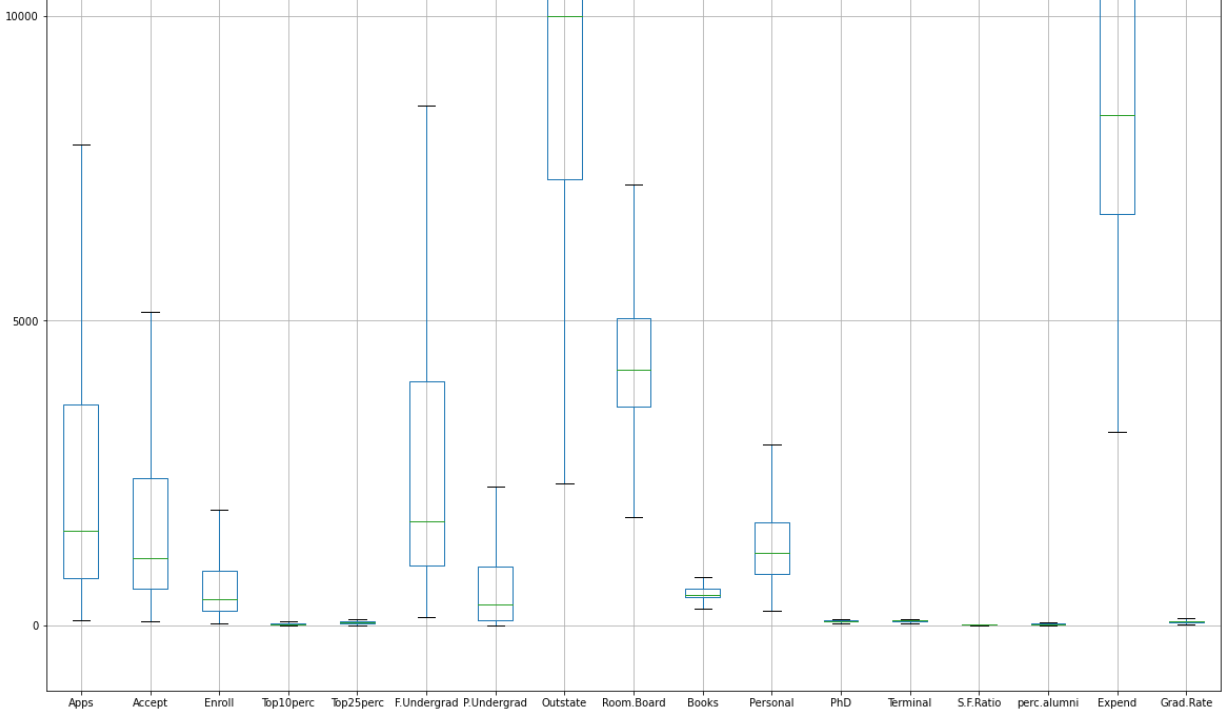


Boxplot before the Outlier Treatment:



Boxplot After treating the outliers:





* There is high degree of correlation observed between the following variables:

Number of applications received and the number of new students enrolled, the number of applications accepted and the number of full-time undergraduate students.

The number of new students enrolled and the number of part time undergraduate students.

The percentage of new students from top 10% & 25% of higher secondary class are highly correlated with the number of students for whom the particular college or university is out-of-state tuition.

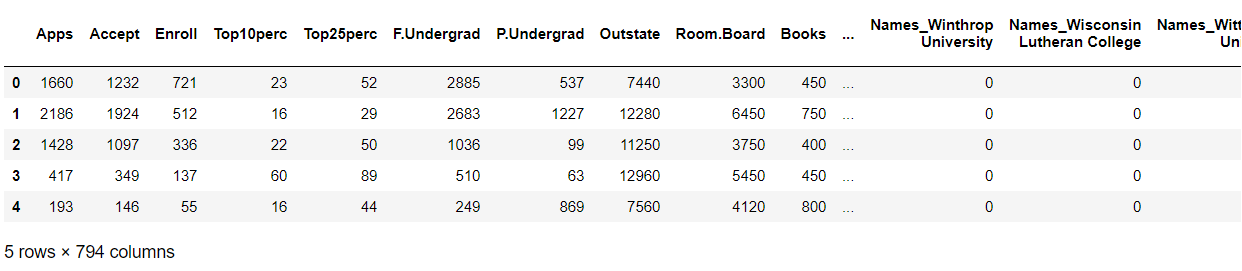
The percentage of faculties with a Ph.D and a terminal degree is high with the percentage of new students from top 10% and 25% of higher secondary class.

The number of students for whom the particular college or university is out-of-state tuition are highly correlated with the graduation rate.

The instructional expenditure per student is particularly low with the number of full-time undergraduate students. It is also showing a low negative correlation with the number of part time undergraduate students.

**2.2)**

We first tried one hot encoding on the dataset for the Names column:



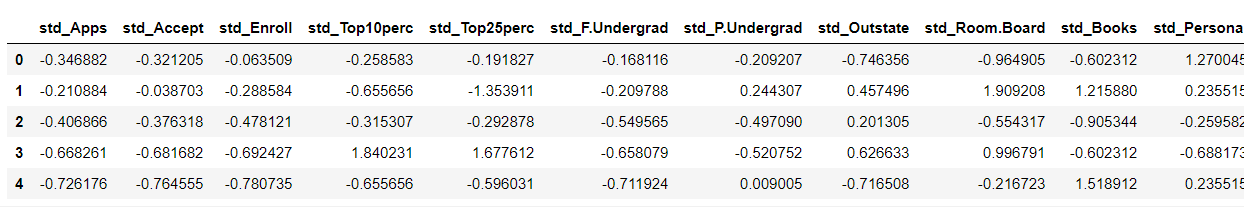
As we can see, this does not seem to be very appropriate for this dataset as the data in the categorical variable is not repetitive in nature.

So, we will choose to drop the Names column from this dataset for further processing:

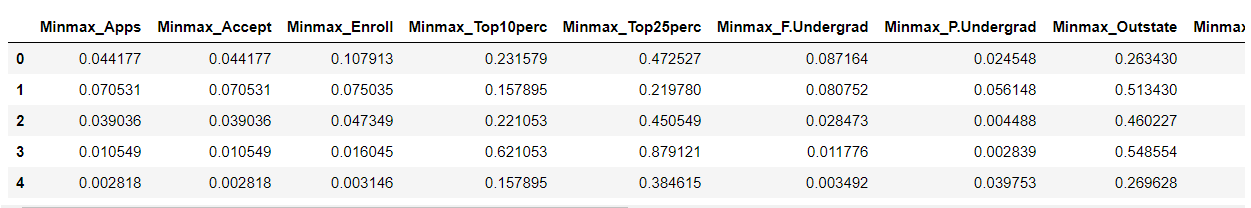
We first explore the different scaling options one by one and then choose which one fits the best:

In the process, we have prefixed every column with the name of the scaling option used to not tamper with the original dataset.

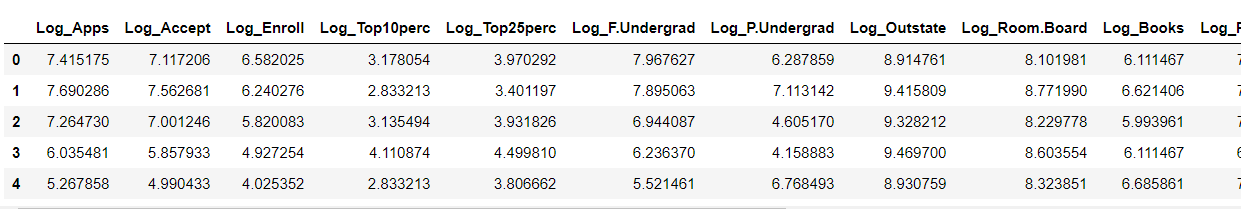
StandardScaler:



MinMaxScaler:



LogTransformer:



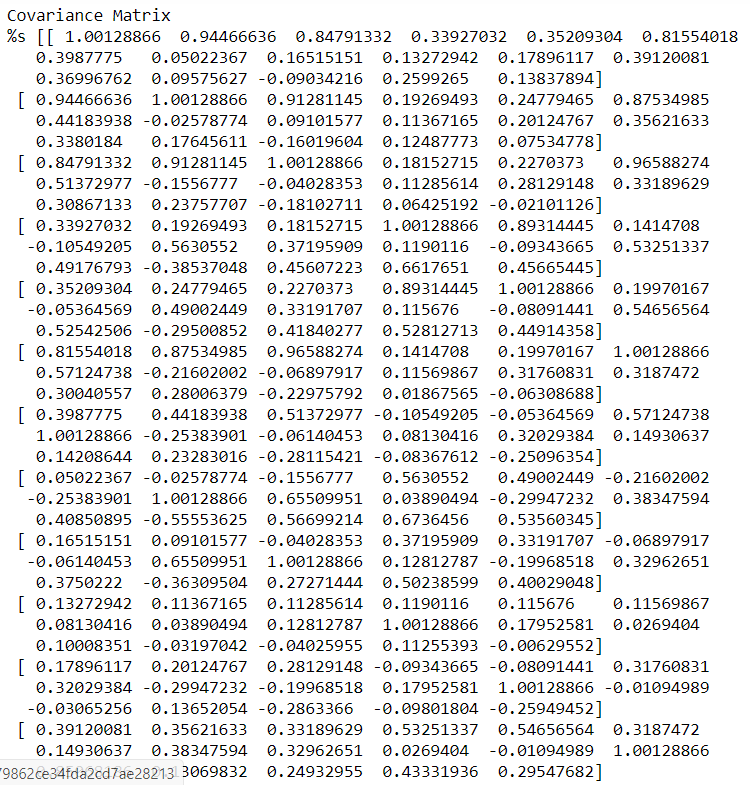
In this case, we have done the scaling without treating the outliers.

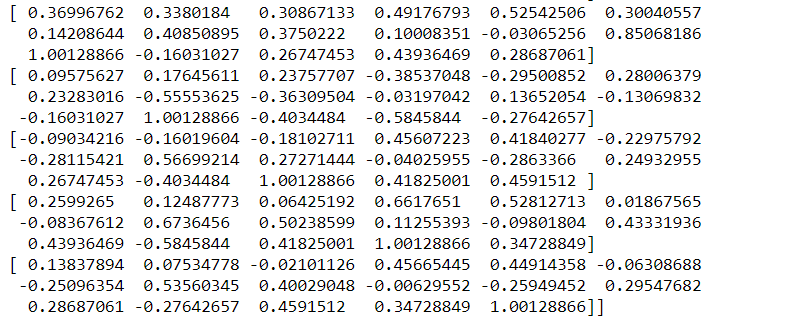
In the question 2.4 we will also see the effect of scaling after the outlier treatment.

We choose Standard Scaler as the chosen option of scaling as our objective is to bring the values of different variables to only one scale and it will also normalize the data.

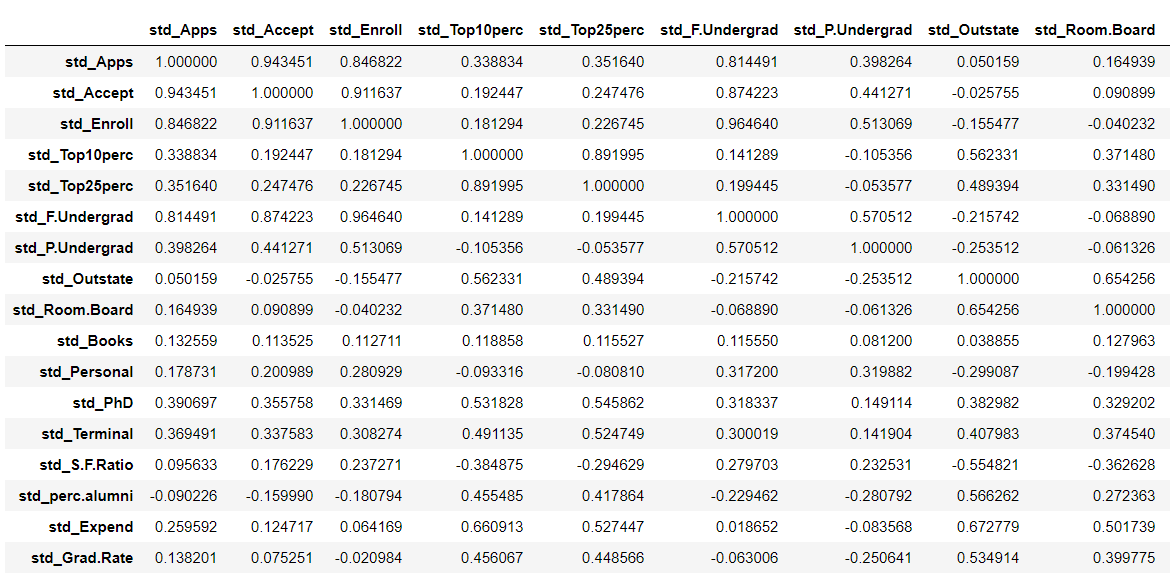
**Answers for Questions: 2.3.) & 2.4.) combined:**

i.) Covariance Matrix of the scaled data before outlier treatment:



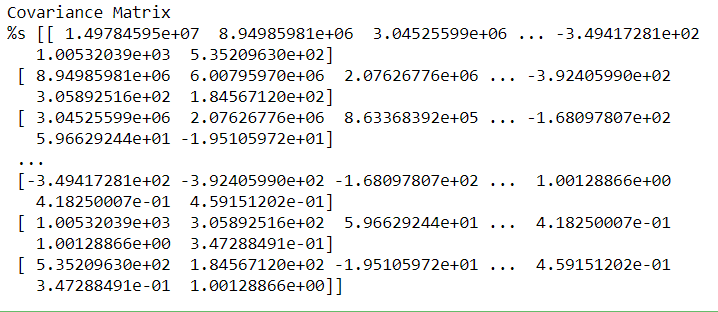


ii.) Correlation Matrix of the scaled data before outlier treatment:

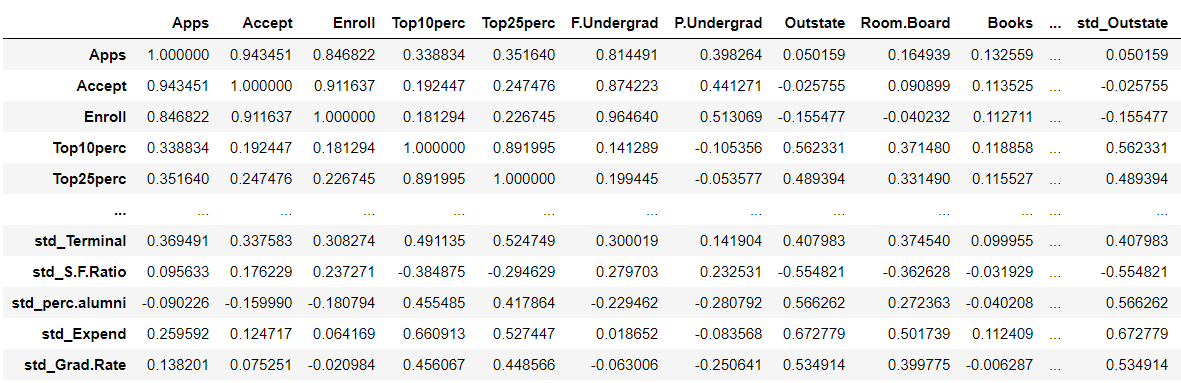




iii.) Covariance matrix of the unscaled data before outlier treatment:



iv.) Correlation matrix of the unscaled data before the outlier treatment:



It can be clearly observed from the above values that the correlation matrix is the same after scaling and before.

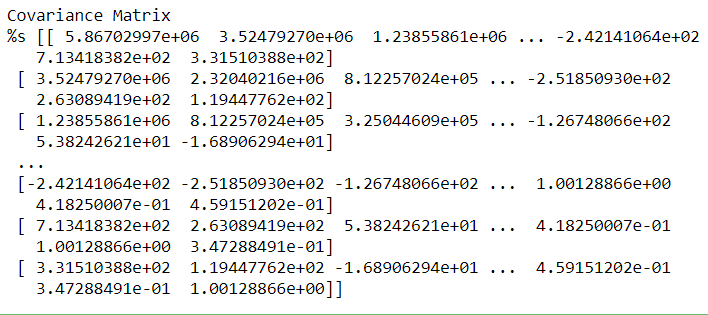
The covariance matrix however differs showing us how each variable is likely to change with respect to the other.

The correlation matrix shows us the strength of that variable.

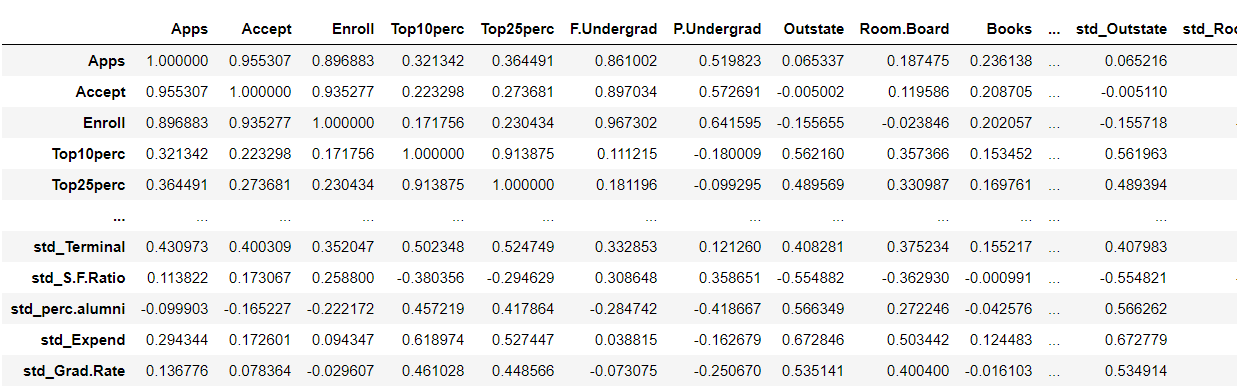
It is also interesting to note that the values of the covariance matrix and the correlation matrix are almost similar for the scaled data but not so for the unscaled data.

Now let us treat the outliers of the data:

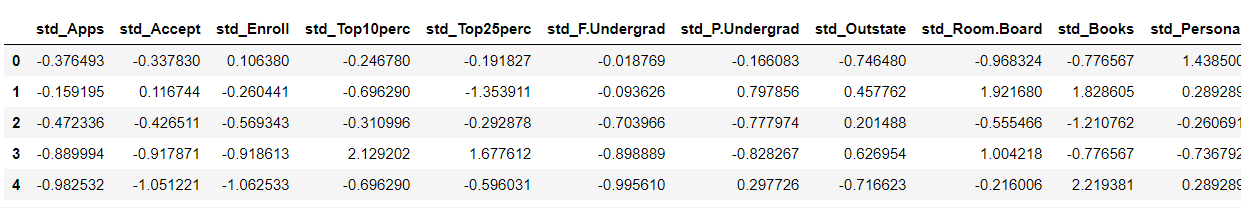
Covariance matrix of the unscaled data after the Outlier Treatment:



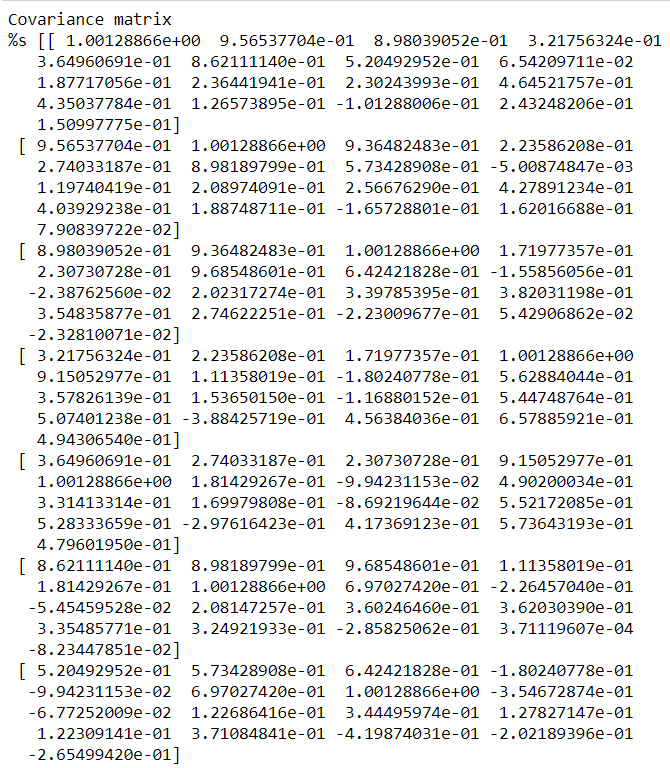
Correlation matrix of the unscaled data after the Outlier treatment:

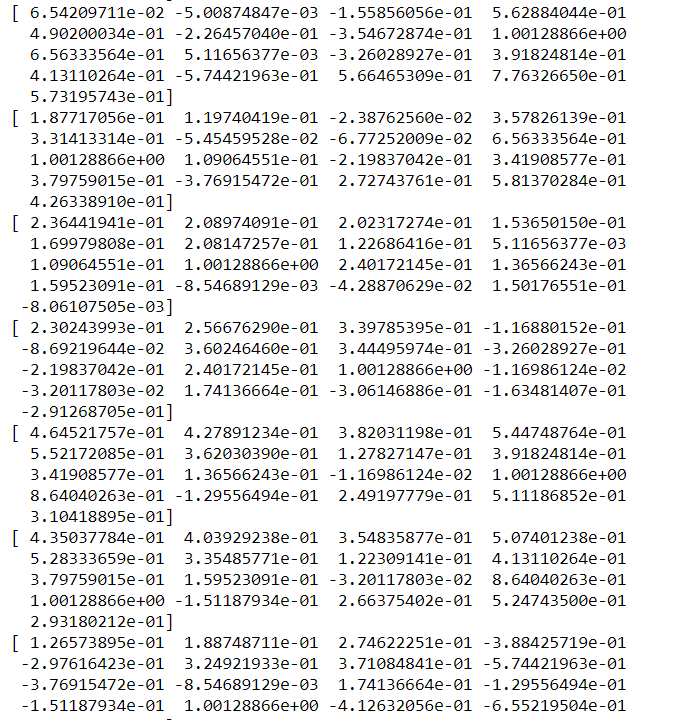


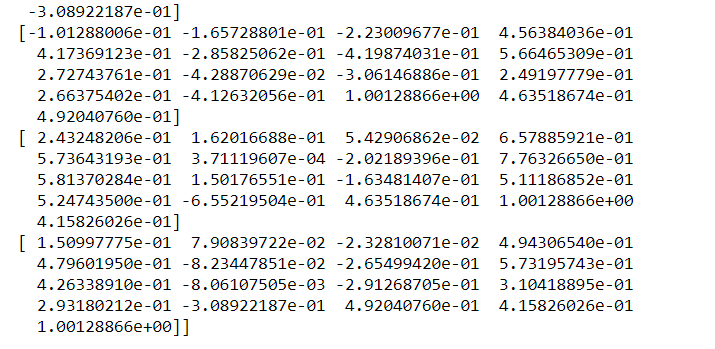
We then perform scaling on the data that has already been treated for outliers. For this we use the Standard Scaler function which returns the z scores. The head of the scaled data then looks like this:



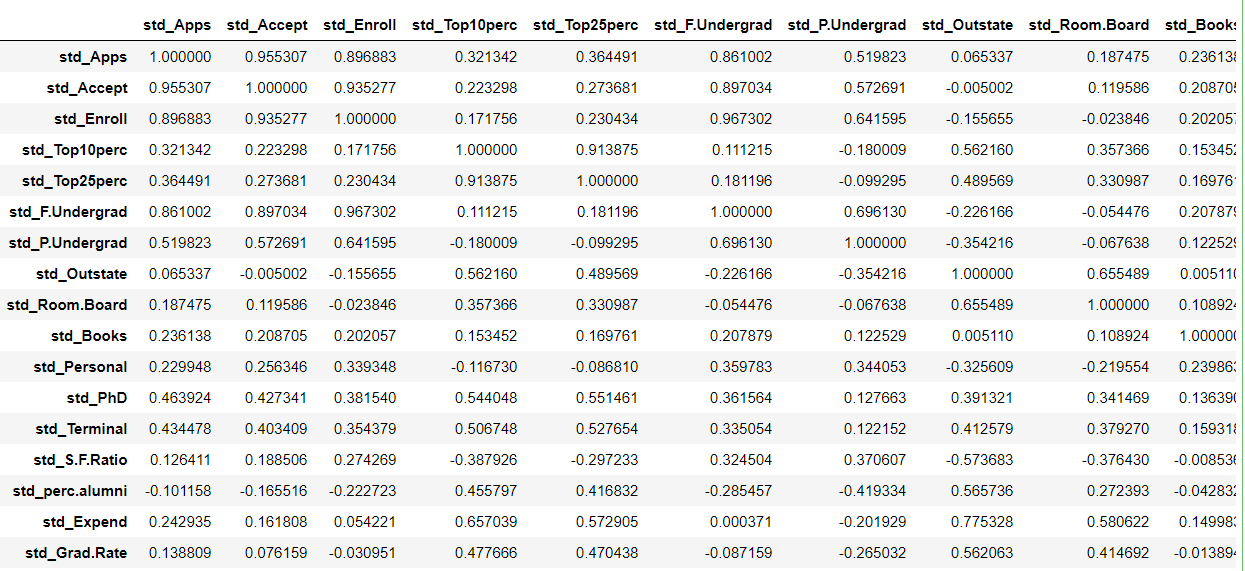
Covariance matrix of the scaled data after the Outlier Treatment:







Correlation matrix of the scaled data after the outlier treatment:



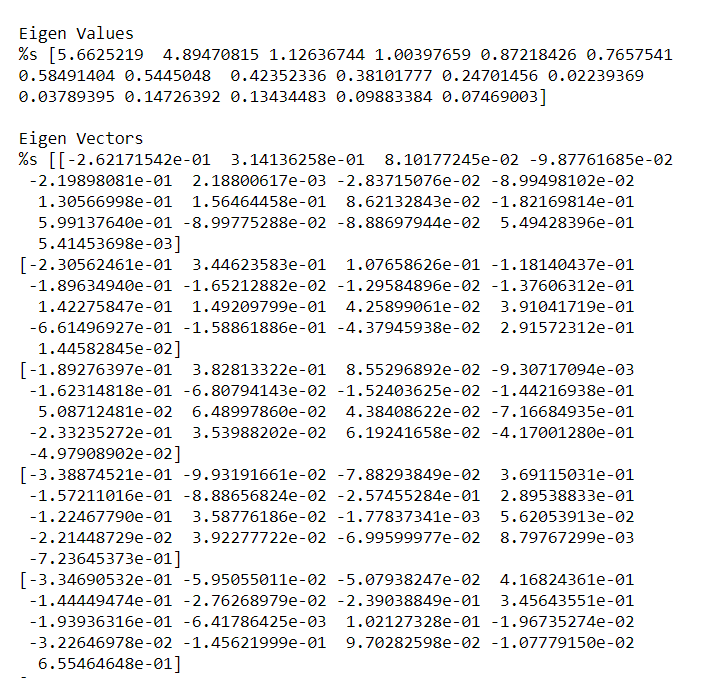
As we can see, the correlation matrix of the scaled and unscaled data after the treatment of outliers is the same.

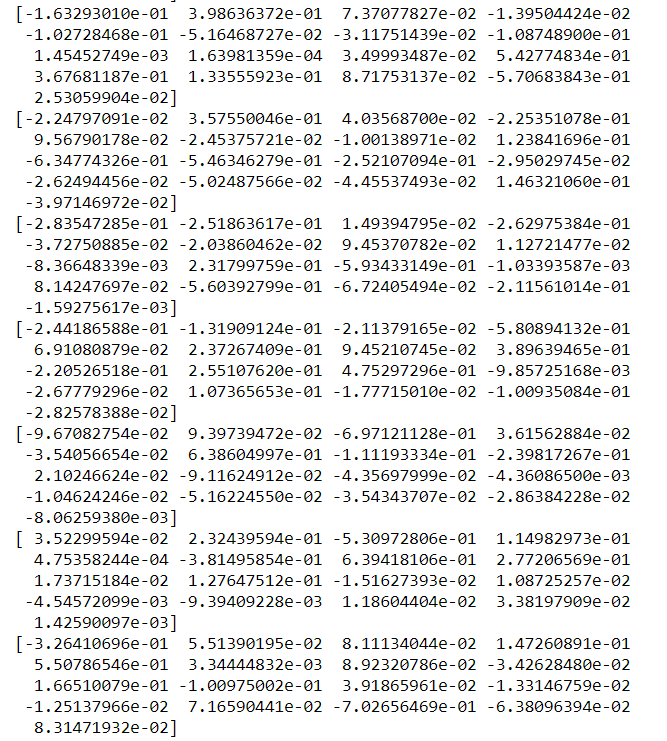
It can also be observed that the correlation between the variables has increased after the outlier treatment.

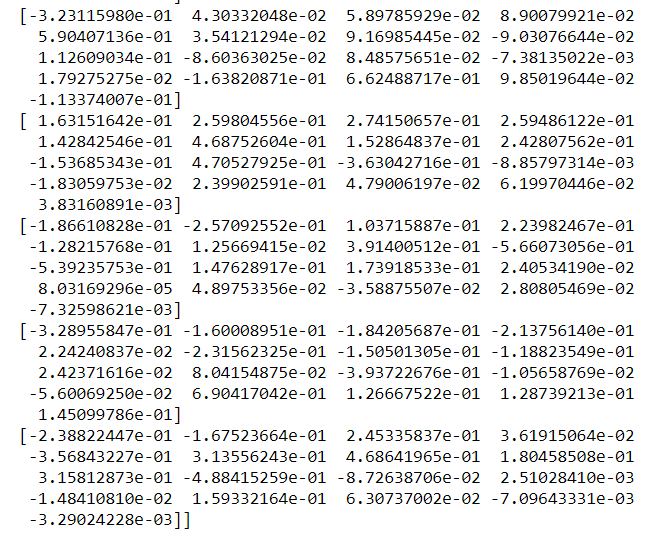
**2.5)**

For calculating the Eigen Values and the Eigen Vectors, let us use the covariance matrix of the scaled data after the outlier treatment since we see that the correlation values have improved post the outlier treatment:

In our calculation, it is the ‘cov\_matrix4’.

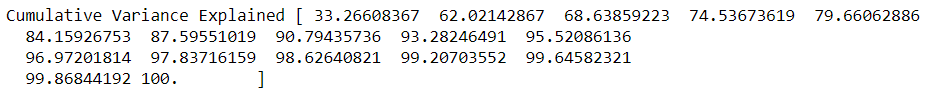






**2.6)** Explicit form of first PC (in terms of Eigen Vectors):

To find the number of principal components required for our data, let us first find out the cumulative distribution of Eigen Values:

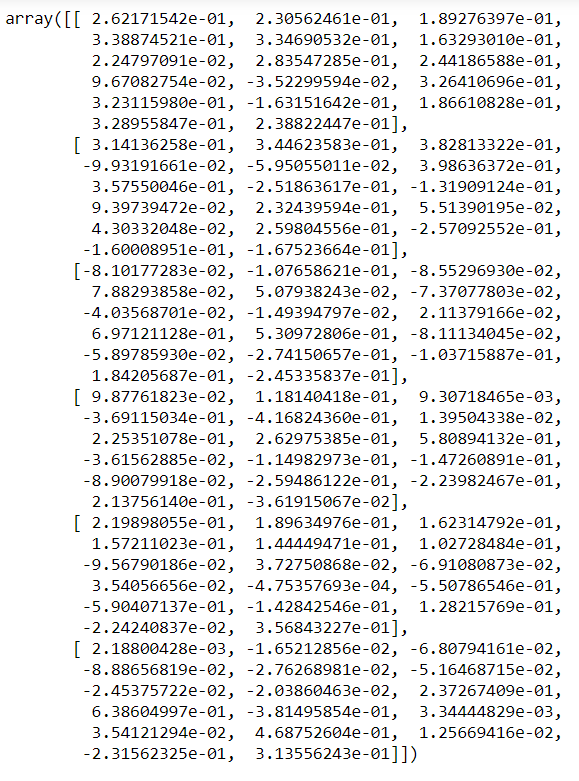


It is therefore evident that we need to go up to 6 principal components to find 84.15% of the variance explained.

Now we use the PCA command from Sklearn to find the principal components:



From the above we see that the complete values for the principal components are not mentioned. So we need to expand the same:



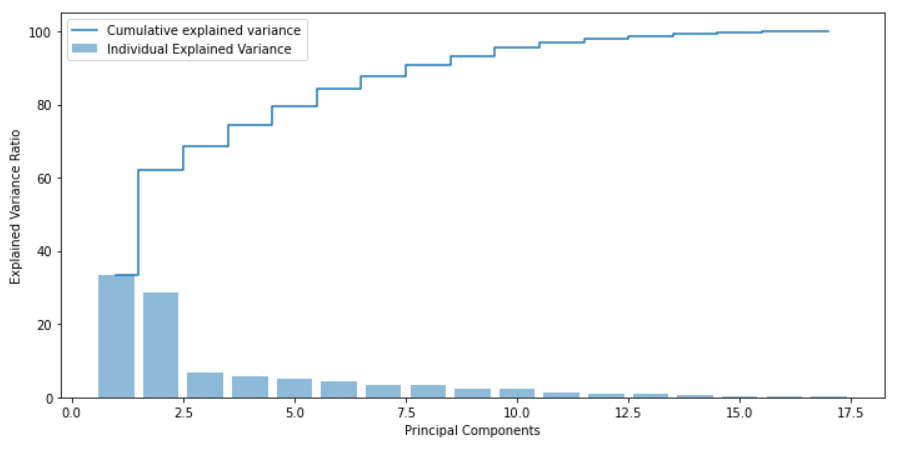
Now we can write the explicit form of the first principal component in terms of Eigen Vectors as:

PC1 = (0.964 x Number of Applications received) + (0.848 x Number of Applications Accepted) + (0.695 x Number of New Students Enrolled) + (1.246 x Percentage of new students from top 10% of Higher Secondary Class) + (1.23 x Percentage of new students from top 25% of Higher Secondary class) + (0.60 x Number of full-time undergraduate students) + (0.304 x Number of part-time undergraduate students) + (1.042 x Number of students for whom the particular college or university is Out-of-state tuition) + (0.897 x Cost of Room and board) + (1.308 x  Estimated book costs for a student) + ((-0.47) x Estimated personal spending for a student)) + (1.200 x Percentage of faculties with Ph.D.’s) + (1.188 x Percentage of faculties with terminal degree) + ((-0.599) x Student/faculty ratio) + (0.686 x Percentage of alumni who donate) + (1.209 x The Instructional expenditure per student) + (0.878 x Graduation Rate)

**2.7)** & **2.8)** The Cumulative values of the eigenvalues are as follows:

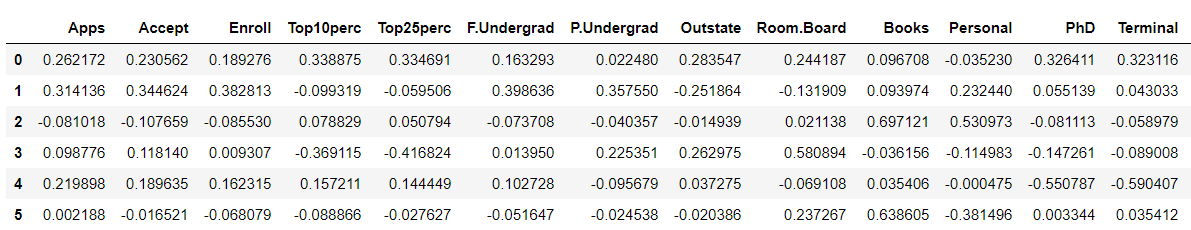


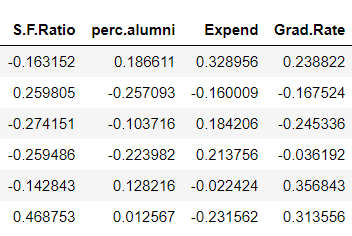
This can be explained better with a visual representation as follows:



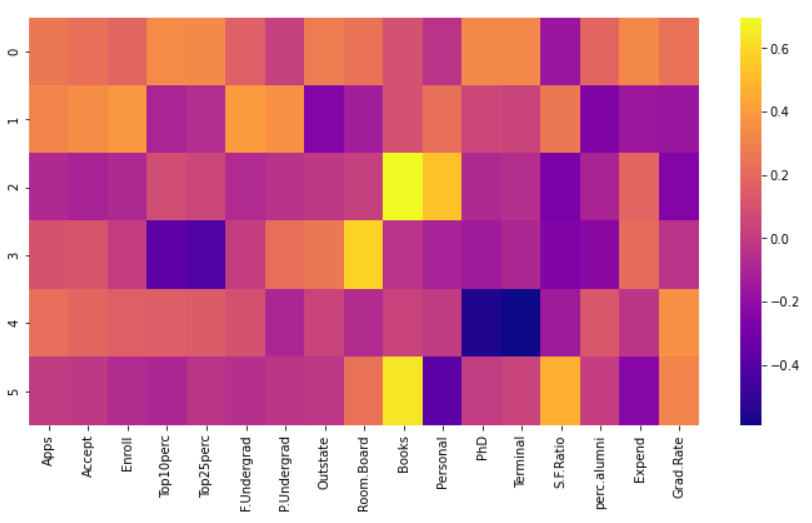
* The need for 6 + principal components to show more than 85% of variance is visually shown above.
* A a thumb rule, we stick to a maximum of 6 principal components and so we get to a variance percentage of 84.15, which is good enough.
* It is also imperative to note here, that the second principal component (28.76%) captures almost the same amount of variance as the first principal component (33.26%).

Now let us put the principal components in a data frame:





Now let us plot a heatmap to show the principal components with the column names:



* The PC1 is highly correlated with the entities: Percentage of new students from top 10% of Higher Secondary class, Percentage of new students from top 25% of Higher Secondary class, Percentage of faculties with Ph.D.’s and Student/faculty ratio.
* The PC2 is highly correlated with Percentage of new students from top 10% of Higher Secondary class, Number of students for whom the particular college or university is Out-of-state tuition, Cost of Room and board, Number of new students enrolled, Number of full-time undergraduate students, Percentage of alumni who donate, ‘The Instructional expenditure per student’ and the Graduation rate.
* The PC4 shows a high relationship between the variables Graduation Rate, Percentage of faculties with Ph.D.’s and Percentage of faculties with terminal degree. The relationship is understandable as Ph.D students and the faculties with Percentage of faculties with Ph.D.’s have a high probability of Graduation Rate owing to the nature of the section of the two.