PROJECT REPORT

ADVANCED STASTICS

PGP-DSBA

MODULE - 3

Contents

[Problem 1A: 3](#_Toc90227041)

[Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [SalaryData.csv] are collected and each person’s educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination. 3](#_Toc90227042)

[[Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.] 3](#_Toc90227043)

[1.1 State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually. 3](#_Toc90227044)

[1.2 Perform one-way ANOVA for Education with respect to the variable ‘Salary’. State whether the null hypothesis is accepted or rejected based on the ANOVA results. 4](#_Toc90227045)

[1.3 Perform one-way ANOVA for variable Occupation with respect to the variable ‘Salary’. State whether the null hypothesis is accepted or rejected based on the ANOVA results. 5](#_Toc90227046)

[1.4 If the null hypothesis is rejected in either (1.2) or in (1.3), find out which class means are significantly different. Interpret the result 6](#_Toc90227047)

[1.5 What is the interaction between the two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot. 7](#_Toc90227048)

[1.6 Perform a two-way ANOVA based on the Education and Occupation (along with their interaction Education\*Occupation) with the variable ‘Salary’. State the null and alternative hypotheses and state your results. How will you interpret this result? 9](#_Toc90227049)

[1.7 Explain the business implications of performing ANOVA for this particular case study. 10](#_Toc90227050)

[Problem 2: 11](#_Toc90227051)

[The dataset Education - Post 12th Standard.csv 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. 11](#_Toc90227052)

[2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA? 11](#_Toc90227053)

[2.2 Is scaling necessary for PCA in this case? Give justification and perform scaling. 22](#_Toc90227054)

[2.3 Comment on the comparison between the covariance and the correlation matrices from this data. [on scaled data] 23](#_Toc90227055)

[2.4 Check the dataset for outliers before and after scaling. What insight do you derive here? 28](#_Toc90227056)

[2.5 Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both] 30](#_Toc90227057)

[2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features. 33](#_Toc90227058)

[2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features] 34](#_Toc90227059)

[2.8 Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate? 34](#_Toc90227060)

[2.9 Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [Hint: Write Interpretations of the Principal Components Obtained] 38](#_Toc90227061)

[**Figure 1 CAT plot of Education VS Salary** 4](#_Toc90227062)

[**Figure 2 One way ANOVA for ‘Education’** 5](#_Toc90227063)

[**Figure 3 CAT plot of Occupation Vs Salary** 5](#_Toc90227064)

[**Figure 4 One way ANOVA for ‘Occupation’** 6](#_Toc90227065)

[Figure 5 Tukey HSD for variable ‘Education’ 6](#_Toc90227066)

[**Figure 6 Tukey HSD for variable ‘Occupation’** 7](#_Toc90227067)

[**Figure 7 Interaction Plot** 7](#_Toc90227068)

[**Figure 8 Two-way ANOVA table** 10](#_Toc90227069)

[Figure 9 Z-score formula 22](file:///E:\GREAT%20LEARNING\MOD%203%20ADVANCED%20STATISTICS\Project\Project.docx#_Toc90227070)

[Figure 10 Snippet of scaled data 23](#_Toc90227071)

[Figure 11 Covariance Formula 24](#_Toc90227072)

[Figure 12 Correlation Formula 27](file:///E:\GREAT%20LEARNING\MOD%203%20ADVANCED%20STATISTICS\Project\Project.docx#_Toc90227073)

[Figure 13 Correlation Matrix 28](#_Toc90227074)

[Figure 14 Boxplot Before Scaling 29](#_Toc90227075)

[Figure 15 Boxplot of Data after Scaling 29](#_Toc90227076)

[Figure 16 Eigen Values 33](#_Toc90227077)

[Figure 17 DataFrame after performing PCA 33](#_Toc90227078)

[Figure 18 Highly Reduced Collinearity after Performing PCA 34](#_Toc90227079)

[Figure 19 Scree Plot 35](#_Toc90227080)

[Figure 20 Cumulative Variance Explained 36](#_Toc90227081)

[Figure 21 DataFrame of Selected PCs 36](#_Toc90227082)

[Figure 22 Original Features and selected PCs 37](#_Toc90227083)

[Figure 23 Heatmap of Selected PCs 37](#_Toc90227084)

# **Problem 1A:**

# Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [[SalaryData.csv](https://olympus.greatlearning.in/courses/59237/files/4139567/download?verifier=s0Kzy5GUh2GUbwwy2IBZsc4TFeNikCBS8fcEf4xU&wrap=1)] are collected and each person’s educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination.

# [Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.]

**INFERENCE OF THE DATASET**

The shape of the dataset seems to be with 40 rows and 3 columns.

Education and Occupation are categorical and Salary is integer.

The entire dataset does not have missing values or null values.

## 1.1 State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.

**One way ANOVA(Education)**

Null Hypothesis 𝐻0: The mean salary is the same across all the 3 categories of education (Doctorate, Bachelors, HS-Grad).

Alternate Hypothesis 𝐻1: The mean salary is different in at least one category of education.

**One way ANOVA(Occupation)**

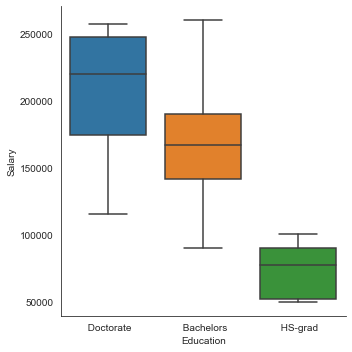
Null Hypothesis 𝐻0: The mean salary is the same across all the 4 categories of occupation (Prof-Specialty, Sales, Adm-clerical, Exec-Managerial).

Alternate Hypothesis 𝐻1: The mean salary is different in at least one category of occupation.

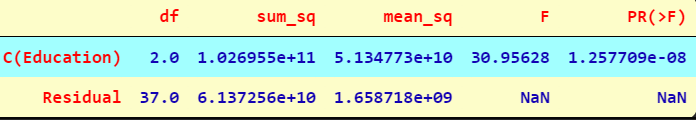
## 1.2 Perform one-way ANOVA for Education with respect to the variable ‘Salary’. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

First, we check the mean difference between salaries by using boxplot for the different categories under ‘Education’ (figure 1). We observe that significant difference is there. We move ahead and perform ANOVA test.

**Figure 1 CAT plot of Education VS Salary**



**Figure 2 One way ANOVA for ‘Education’**



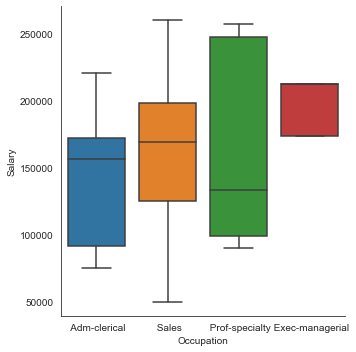
The above is the ANOVA table for Education variable.

Since the p value = 1.257709e-08 is less than the significance level (alpha = 0.05), we can reject the null hypothesis and conclude that there is a significant difference in the mean salaries for at least one category of education.

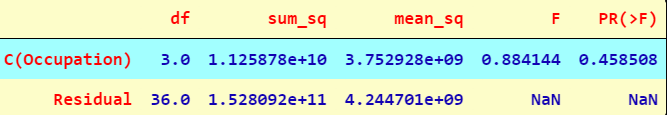
## 1.3 Perform one-way ANOVA for variable Occupation with respect to the variable ‘Salary’. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

First, we check the mean difference between salaries by using boxplot for the different categories under ‘Occupation’ (figure 3). We observe, there is not much significant difference between them. We move ahead and perform ANOVA test.

**Figure 3 CAT plot of Occupation Vs Salary**



**Figure 4 One way ANOVA for ‘Occupation’**



The above is the ANOVA table for Occupation variable.

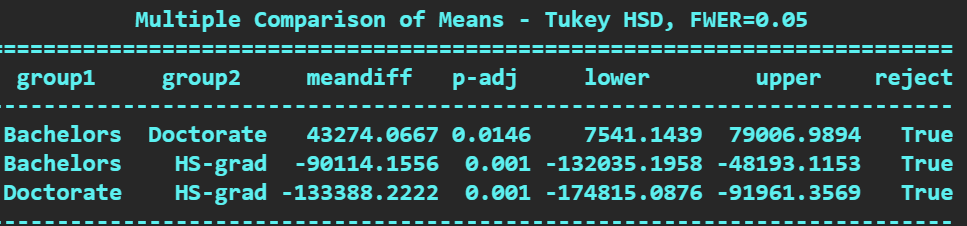
Since the p value = 0.458508 is greater than the significance level (alpha = 0.05), we fail to reject the null hypothesis (i.e., we accept H0) and conclude that there is no significant difference in the mean salaries across the 4 categories of occupation

1.4 If the null hypothesis is rejected in either (1.2) or in (1.3), find out which class means are significantly different. Interpret the result**.**

To find out which class means are significantly different, the Tukey Honest Significant Difference test is performed.

Using, the Tukey Honest Significant Difference test, we get the following table for the category education:

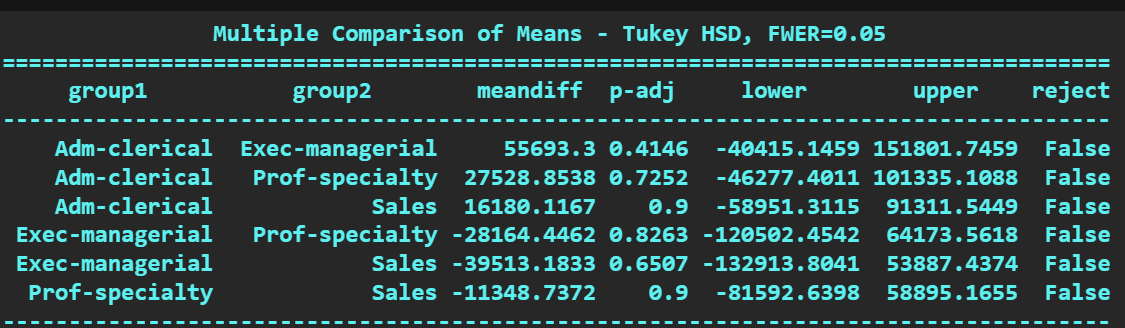
Figure Tukey HSD for variable ‘Education’



The table shows that since the p- values (p-adj in the table) are lesser than the significance level for all the three categories of education, this implies that the mean salaries across all categories of education are different.

For the category occupation, the Tukey Honest Significant Difference test has further confirmed that the mean salaries across all occupation classes are significantly same. The table below confirms the same, wherein we see that all p-values are greater than 0.05.

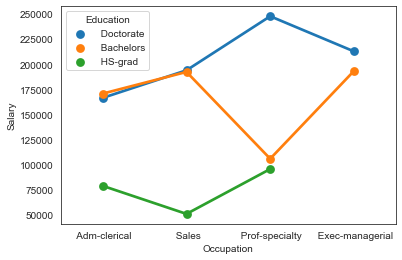
**Figure 6 Tukey HSD for variable ‘Occupation’**



## 1.5 What is the interaction between the two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.

We analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.

**Figure 7 Interaction Plot**



The interaction plot shows that there is significant amount of interaction between the categorical variables, Education and Occupation.

The following are some of the observations from the interaction plot:

· People with HS-grad education do not reach the position of Exec-managerial and they hold only Adm-clerk, Sales and Prof-Specialty occupations.

· People with education as Bachelors or Doctorate and occupation as Adm-clerical and Sales almost earn the same salaries (salaries ranging from 170000–190000).

· People with education as Bachelors and occupation as Prof-Specialty earn lesser than people with education as Bachelors and occupations as Adm-clerical and Sales.

· People with education as Bachelors and occupation Sales earn higher than people with education as Bachelors and occupation Prof-Specialty whereas people with education as Doctorate and occupation Sales earn lesser than people with Doctorate and occupation Prof-Specialty. We see a reversal in this part of the plot.

· Similarly, people with education as Bachelors and occupation as Prof-Specialty earn lesser than people with education as Bachelors and occupation Exec-Managerial whereas people with education as Doctorate and occupation as Prof-Specialty earn higher than people with education as Doctorate and occupation Exec-Managerial. There is a reversal in this part of the plot too.

· Salespeople with Bachelors or Doctorate education earn the same salaries and earn higher than people with education as HS-grad.

· Adm clerical people with education as HS-grad earn the lowest salaries when compared to people with education as Bachelors or Doctorate.

· Prof-Specialty people with education as Doctorate earn maximum salaries and people with education as HS-Grad earn the minimum.

· People with education as HS -Grad earn the minimum salaries.

· There are no people with education as HS -grad who hold Exec-managerial occupation.

· People with education as Bachelors and occupation, Sales and Exec-Managerial earn the same salaries.

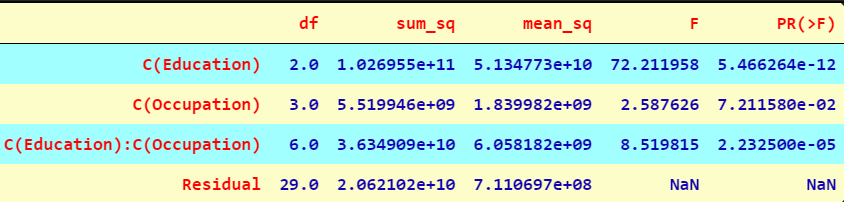
## 1.6 Perform a two-way ANOVA based on the Education and Occupation (along with their interaction Education\*Occupation) with the variable ‘Salary’. State the null and alternative hypotheses and state your results. How will you interpret this result?

𝐻0: The effect of the independent variable ‘Education’ on the mean ‘Salary’ does not depend on the effect of the other independent variable ‘Occupation’ (i. e. there is no interaction effect between the 2 independent variables, education and occupation).

𝐻1: There is an interaction effect between the independent variable ‘Education’ and the independent variable ‘Occupation’ on the mean ‘Salary’.

By performing two-way ANOVA, we get the following table:

**Figure 8 Two-way ANOVA table**



From the table, we see that there is a significant amount of interaction between the variables, Education and Occupation.

As p value = 2.232500e-05 is lesser than the significance level (alpha = 0.05), we reject the null hypothesis.

Thus, we see that there is an interaction effect between education and occupation on the mean salary.

## 1.7 Explain the business implications of performing ANOVA for this particular case study.

From the ANOVA method and the interaction plot, we see that education combined with occupation results in higher and better salaries among the people. It is clearly seen that people with education as Doctorate draw the maximum salaries and people with education HS-grad earn the least. Thus, we can conclude that Salary is dependent on educational qualifications and occupation.

# **Problem 2:**

# The dataset [Education - Post 12th Standard.csv](https://olympus.greatlearning.in/courses/59237/files/3628026/download?verifier=27lW89sX4nTc0E7xfH3NXb0qcQGg0LR770NWvmd4&wrap=1) 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](https://olympus.greatlearning.in/courses/59237/files/3628025/download?verifier=T48KvfUdRBF2vSBRxuKOA4AhwjEiBH8dnPtnY5Jy&wrap=1).

The given dataset consists of data points of names of various university and college which has number of application received, accepted, and enrolled, percentage of new students from top 10% of higher secondary class, percentage of new students from top 25% of higher secondary class, Number of fulltime undergraduates, Number of part time undergraduate students, Number of students for whom the particular college is out of state tuition, cost of room and board, estimated book costs for a student, estimated personal spending for a student, percentage of faculties with PHD, percentage of faculties with terminal degree, student/faculty ratio, percentage of alumni who donate, The instructional expenditure per student, Graduation Rate.

**INFERENCE OF THE DATASET**

The shape of the dataset seems to be with 777 rows and 18 columns.

All the columns seem to be integer or float values.

The Names column alone is a categorical value.

We also can see they are no duplicates in the dataset.

The entire dataset does not have missing values or null values.

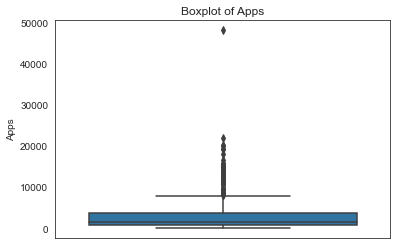
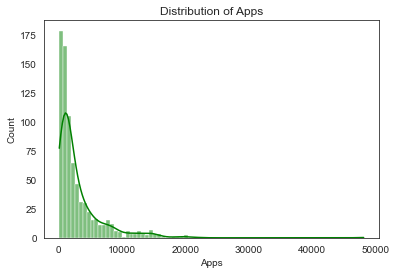
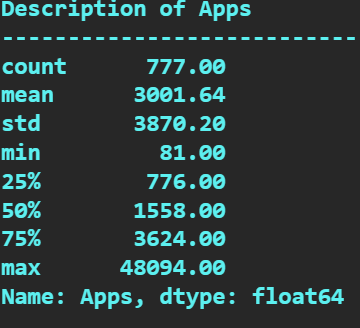
## 2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?

**UNIVARIATE ANALYSIS**

Helps us to understand the distribution of data in the dataset. With univariate analysis we can find patterns and we can summarize the data.

For univariate analysis we are using box plot and hist plot to find information or patterns in the data.

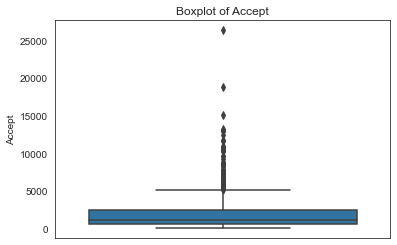
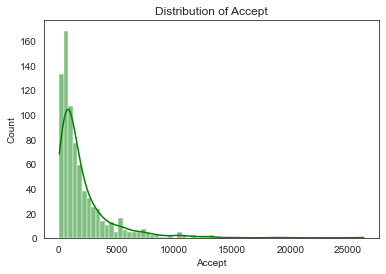
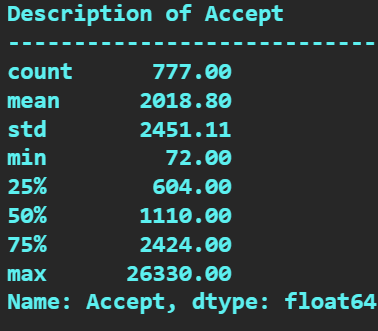
**APPS**

****

The Box plot of ‘Apps’ variable seems to have outliers; the distribution of the data is skewed we could also understand that mean applications offered by each college or university is 3000 approximately.

The max applications seem to be around 48094 and minimum is 81.

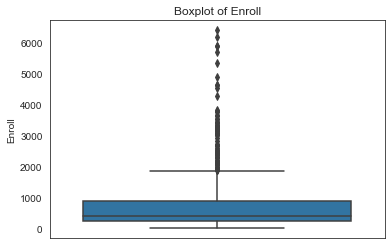
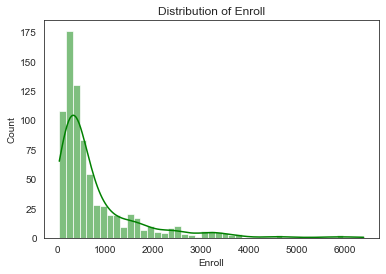
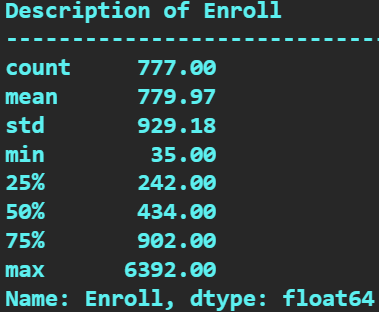
**ACCEPT**

****

The Box plot of ‘Accept’ variable seems to have outliers, the distribution of the data is skewed we could also understand that mean applications accepted by each college or university is 2018 approximately.

The max applications accepted seem to be around 26330 and minimum is 72.

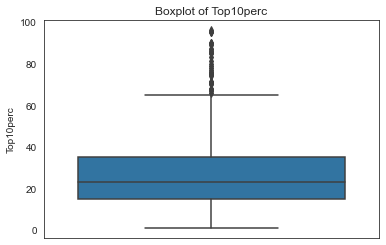
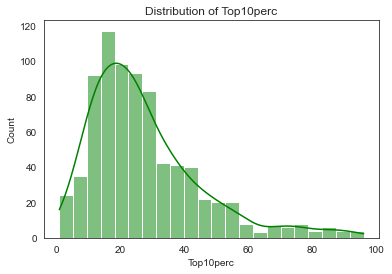
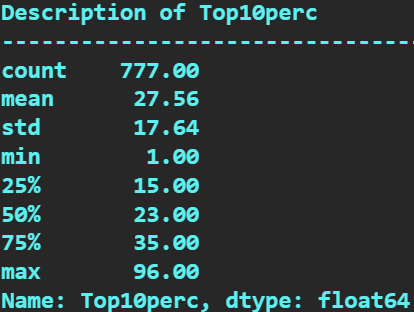
**ENROLL**

****

The Box plot of ‘Enroll’ variable also seems to have outliers, the distribution of the data is skewed we could also understand that mean enrolment by each college or university is 779 approximately.

The max enrolment seems to be around 6392 and minimum is 35.

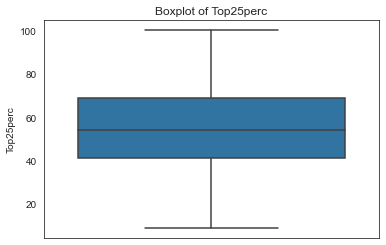
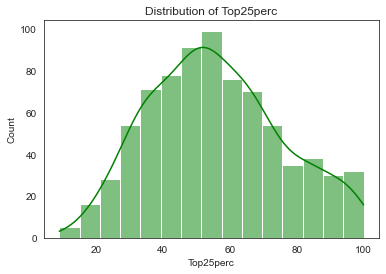
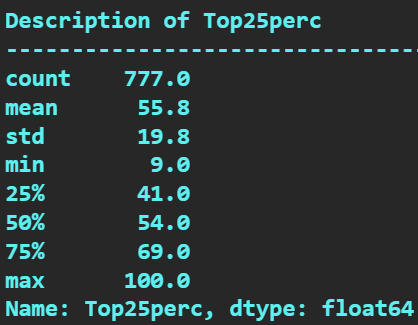
**TOP 10 PERC**

****

The Box plot of ‘Top 10 Perc’ variable also seems to have outliers, the distribution of the data is skewed. There is good amount of intake about 30 to 50 students from top 10 percentage of higher secondary class.

The max intake seems to be around 96 and minimum is 1.

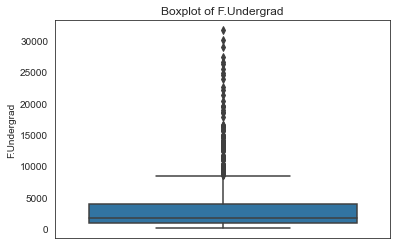
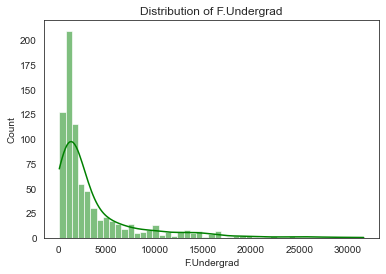
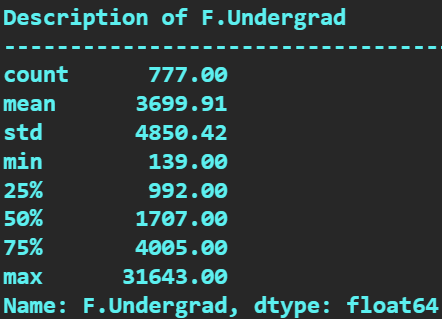
**TOP 25 PERC**

****

The Box plot of ‘Top 25 Perc’ variable have no outliers, the distribution of the data is almost normally distributed. Majority of the students are from top 25% of higher secondary class.

The max intake seems to be around 100 and minimum is 9.

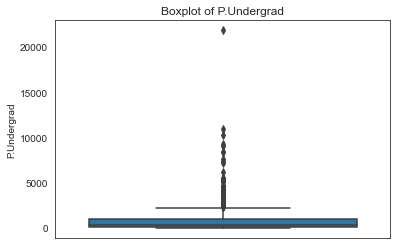
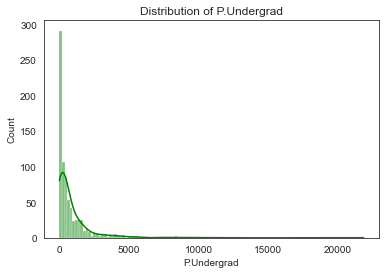
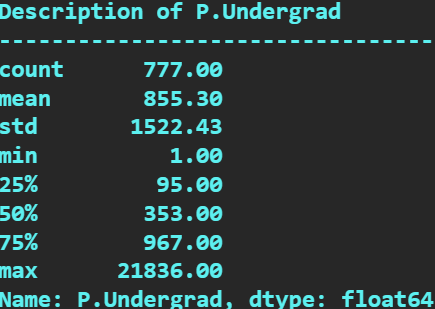
**FULL TIME UNDERGRADUATE**



The box plot of the full-time graduates has outliers. The distribution of the data is positively skewed. In the range about 3000 to 5000 they are full time graduates studying in all the university

The max intake under this category seems to be around 31643 and minimum is 139.

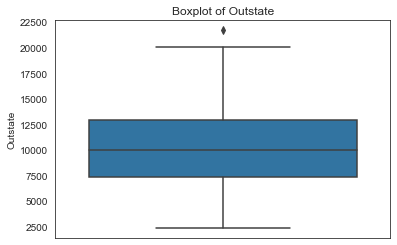
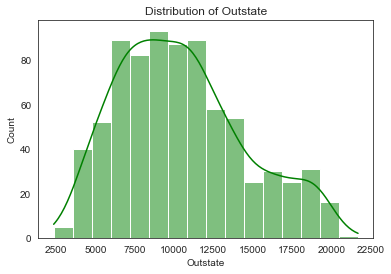
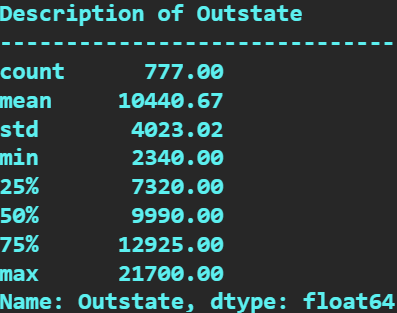
**PART TIME UNDERGRADUATE**

****

The box plot of the part time graduates has outliers. The distribution of the data is positively skewed. In the range about 1000 to 3000 they are part-time graduates studying in all the university.

The max intake under this category seems to be around 21836 and minimum is 1.

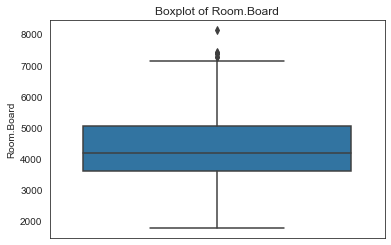
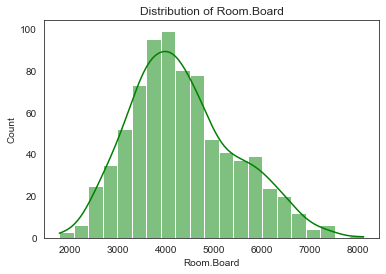
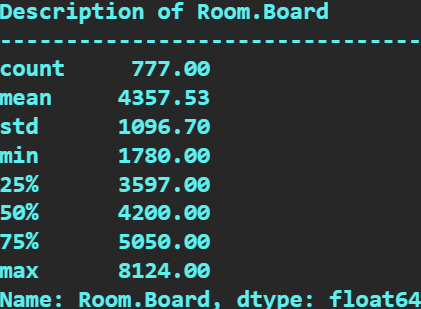
**OUTSTATE**

****

The box plot of outstate has only one outlier. The distribution is almost normally distributed.

The max intake under this category seems to be around 21700 and minimum is 2340.

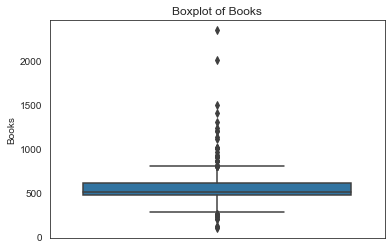
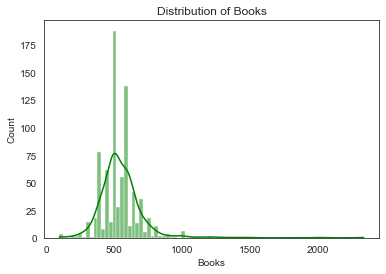
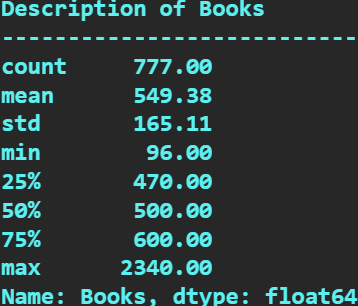
**ROOM BOARD**

****

The Room board has few outliers. The distribution is normally distributed.

The max intake under this category seems to be around 8124 and minimum is 1780.

**BOOKS**

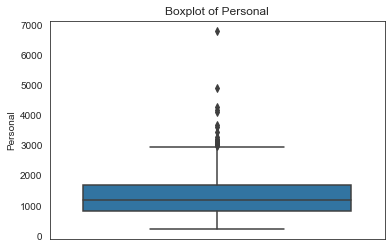
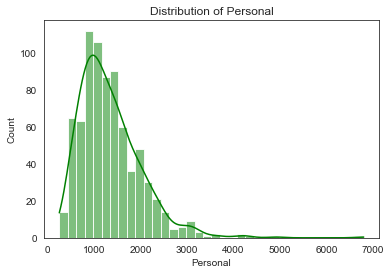
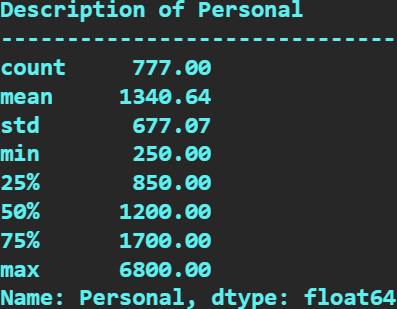
****

The box plot of books has outliers. The distribution seems to be bimodal. The cost of books per student seems to be in the range of 500 to 1000.

The maximum cost of books per student seems to be 2340.

The minimum cost of books per student seems to be 96.

**PERSONAL**

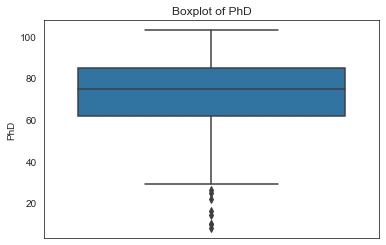
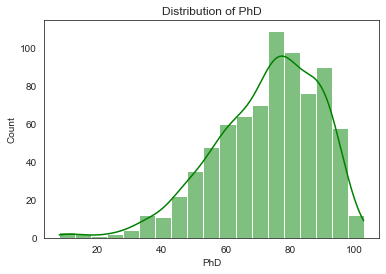
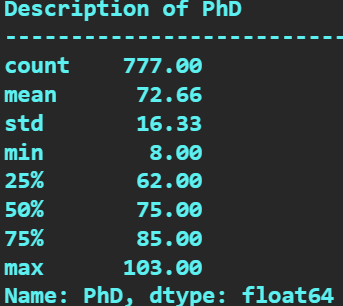
****

The box plot of personal expense has outliers. Some student’s personal expense is way bigger than the rest of the students. The distribution seems to be positively skewed.

Maximum personal expense is 6800.

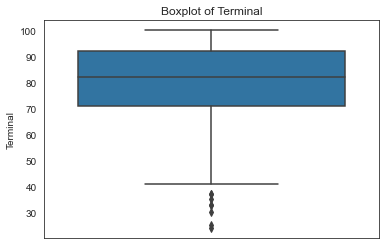
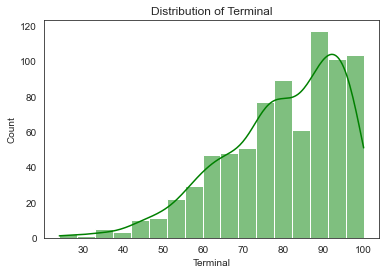
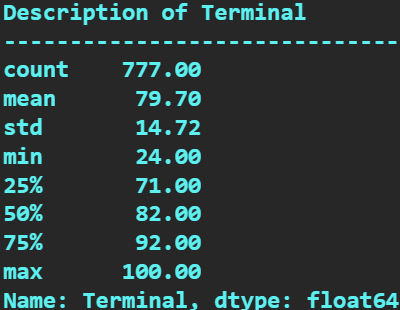
Minimum personal expense is 250.

**PHD**

****

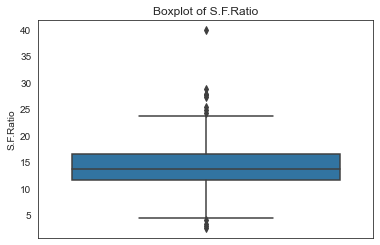
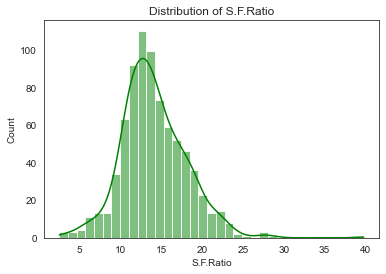
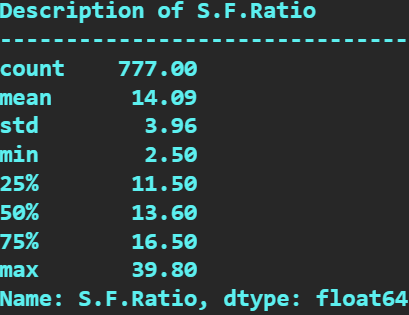
The box plot of PHD has outliers. The distribution seems to be negatively skewed.

**TERMINAL**

****

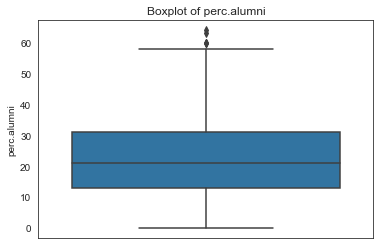
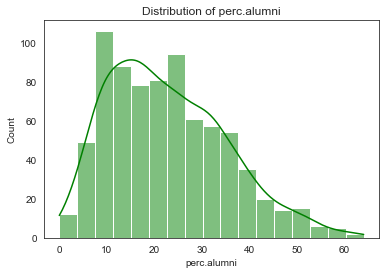
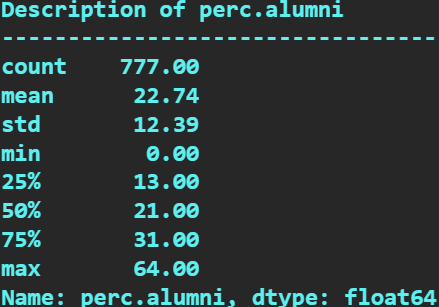
The box plot of terminal seems to have outliers in the dataset. The distribution for the terminal also seems to be negatively skewed.

**SF RATIO**

****

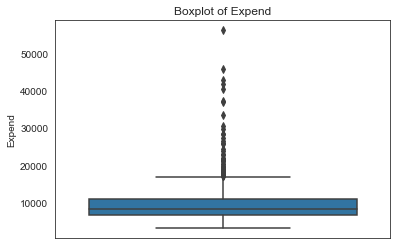
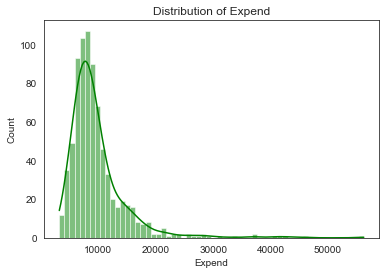
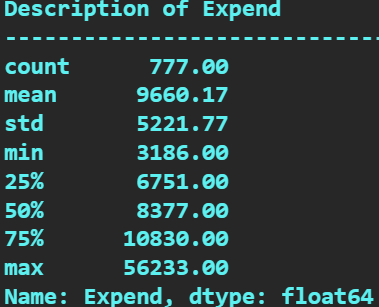
The SF ratio variable also has outliers in the dataset. The distribution is almost normally distributed. The student faculty ratio is almost same in all the university and colleges.

**PERCI ALUMINI**

****

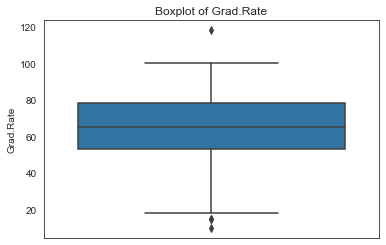
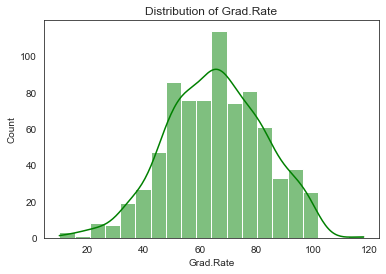
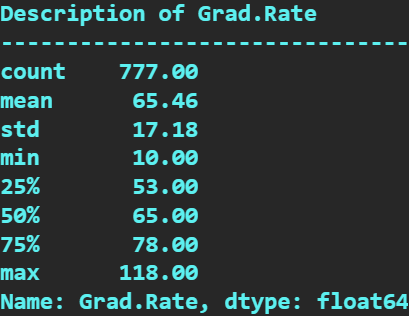
The percentage of alumni box plot seems to have outliers in the dataset. The distribution is almost normally distributed.

**EXPENDITURE**

****

The expenditure variable also has outliers in the dataset. The distribution of the expenditure is positively skewed.

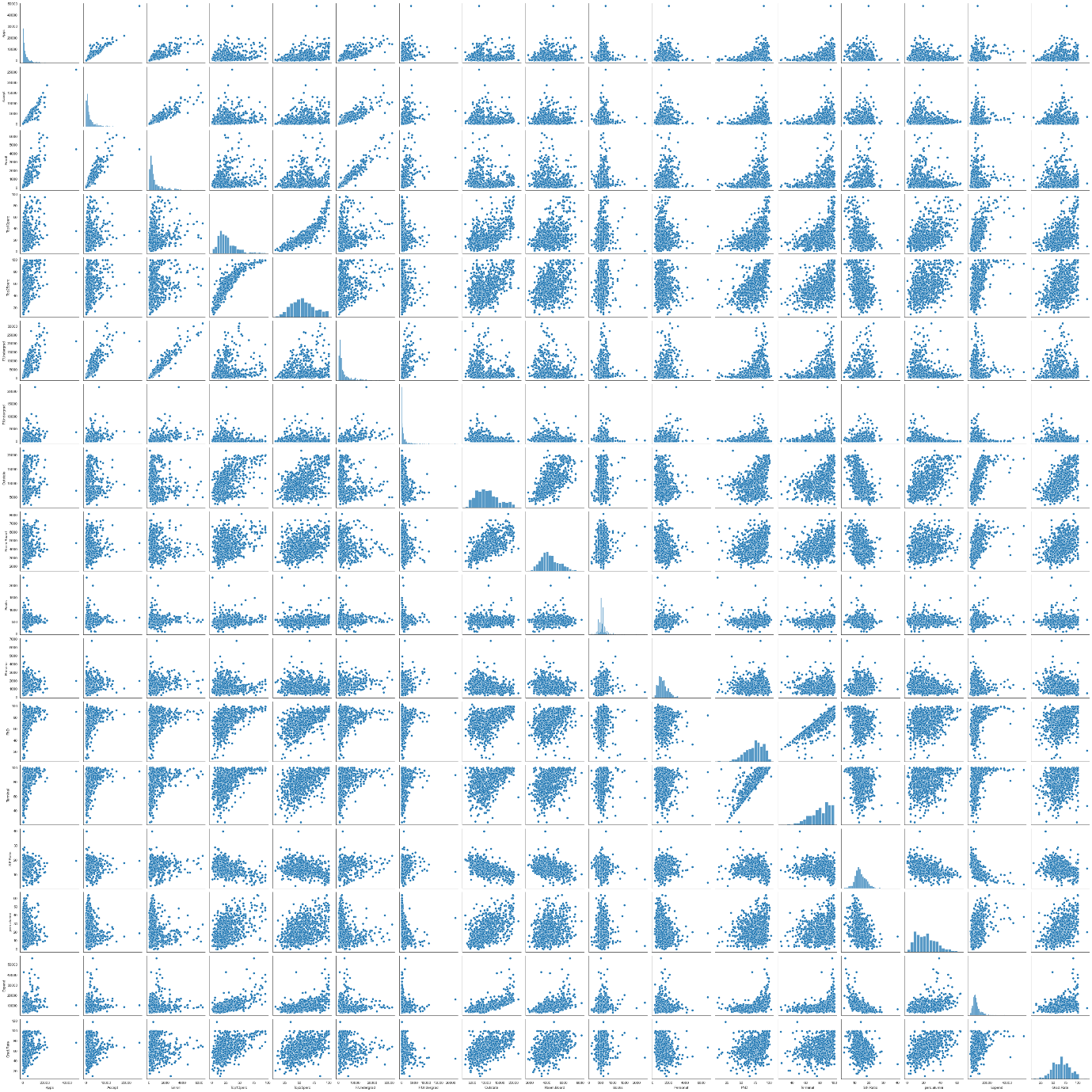
**GRAD RATE**

****

The graduation rate among the students in all the university above 60%. The box plot of the graduation rate has outliers in the dataset. The distribution is normally distributed.

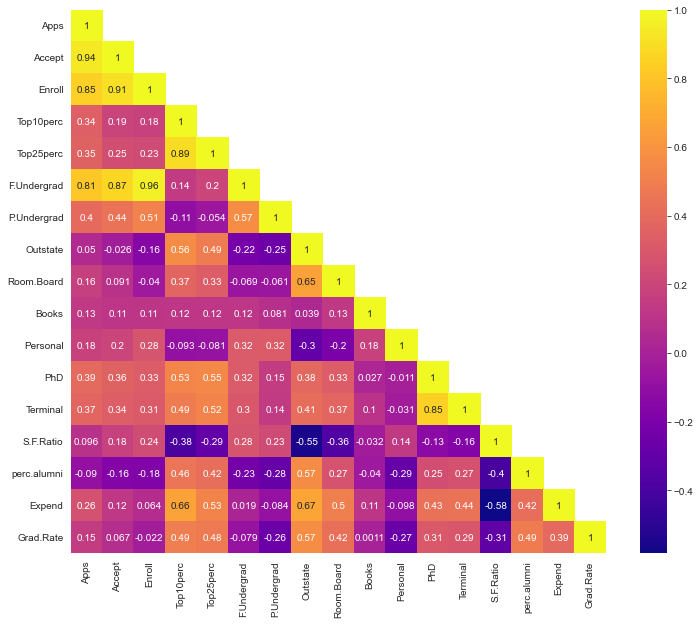
**MULTIVARIATE ANALYSIS**

**PAIRPLOT**

****

The pair plot helps us to understand the relationship between all the numerical values in the dataset. On comparing all the variables with each other we could understand the patterns or trends in the dataset.

**HEAT MAP**

****

This Heat map gives us the correlation between two numerical values.

We could understand the application variable is highly positively correlated with application accepted, students enrolled and full-time graduates. So, this relationship gives the insights on when student submits the application it is accepted and the student is enrolled as fulltime graduate.

We can find negative correlation between application and percentage of alumni. This indicates us not all students are part of alumni of their college or university.

The application with top 10, 25 perc of higher secondary class, outstate, room board, books, personal, PhD, terminal, S.F ratio, expenditure and Graduation ratio are positively correlated.

## 2.2 Is scaling necessary for PCA in this case? Give justification and perform scaling.

Before scaling I have dropped the names variable which is categorical.

Now, the dataset consists of only numerical values, I have applied z-score method for this case study. We can also use min max function to scale the variables.

Since, the dataset has 17 numerical columns with different scales.

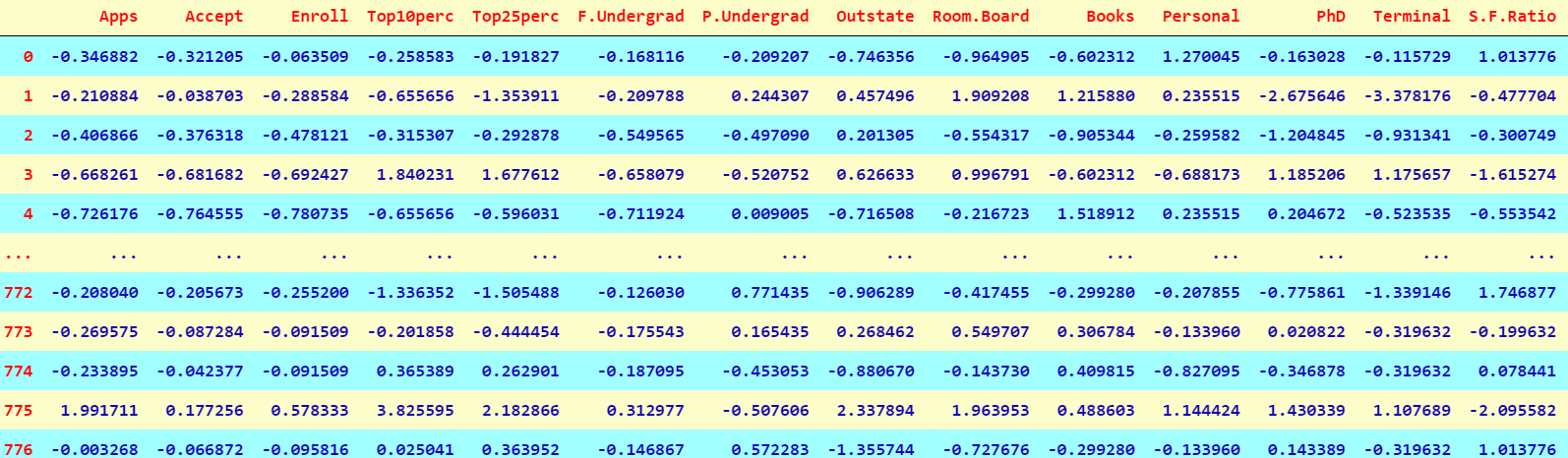
For example, the application, accepted application, enrolled fulltime graduates, part-time graduates, outstate are number of students. The top10 percent and top20 percent are students in which the values are given in percentage. Room board, books, and personal are values associated with money. The PhD, sf ratio, percentage of alumni are percentage values of different combinations of students, teachers, alumni. The graduation rate is also percentage value of graduates who get graduated every year.

Formula for Z-score is given below;

Below is a snippet of scaled data

Figure Z-score formula

Figure Snippet of scaled data



Z score tells us how many standard deviations is the point away from the mean and also the direction. Now, we can understand that all the variables are scaled by using z score function. Scaling is one of the most important methods to follow before implementing models.

## 2.3 Comment on the comparison between the covariance and the correlation matrices from this data. [on scaled data]

The comparison between the covariance and correlation matrix is that both of the terms measure the relationship and the dependency between two variables.

Scaling in general means representation of the dataset. The numbers will not change. We are bringing the dataset into one unit.

Covariance indicates the direction of the linear relationship between the variables whether it is positive or negative. By direction means it is directly proportional or inversely proportional.

Figure Covariance Formula



This below snippet is the covariance matrix on scaled dataset. We can clearly understand covariance matrix indicates direction of the linear relationship between the variables. By direction means it is directly proportional or inversely proportional.

**Covariance Matrix**

**[[ 1.00128866 0.94466636 0.84791332 0.33927032 0.35209304 0.81554018**

**0.3987775 0.05022367 0.16515151 0.13272942 0.17896117 0.39120081**

**0.36996762 0.09575627 -0.09034216 0.2599265 0.14694372]**

**[ 0.94466636 1.00128866 0.91281145 0.19269493 0.24779465 0.87534985**

**0.44183938 -0.02578774 0.09101577 0.11367165 0.20124767 0.35621633**

**0.3380184 0.17645611 -0.16019604 0.12487773 0.06739929]**

**[ 0.84791332 0.91281145 1.00128866 0.18152715 0.2270373 0.96588274**

**0.51372977 -0.1556777 -0.04028353 0.11285614 0.28129148 0.33189629**

**0.30867133 0.23757707 -0.18102711 0.06425192 -0.02236983]**

**[ 0.33927032 0.19269493 0.18152715 1.00128866 0.89314445 0.1414708**

**-0.10549205 0.5630552 0.37195909 0.1190116 -0.09343665 0.53251337**

**0.49176793 -0.38537048 0.45607223 0.6617651 0.49562711]**

**[ 0.35209304 0.24779465 0.2270373 0.89314445 1.00128866 0.19970167**

**-0.05364569 0.49002449 0.33191707 0.115676 -0.08091441 0.54656564**

**0.52542506 -0.29500852 0.41840277 0.52812713 0.47789622]**

**[ 0.81554018 0.87534985 0.96588274 0.1414708 0.19970167 1.00128866**

**0.57124738 -0.21602002 -0.06897917 0.11569867 0.31760831 0.3187472**

**0.30040557 0.28006379 -0.22975792 0.01867565 -0.07887464]**

**[ 0.3987775 0.44183938 0.51372977 -0.10549205 -0.05364569 0.57124738**

**1.00128866 -0.25383901 -0.06140453 0.08130416 0.32029384 0.14930637**

**0.14208644 0.23283016 -0.28115421 -0.08367612 -0.25733218]**

**[ 0.05022367 -0.02578774 -0.1556777 0.5630552 0.49002449 -0.21602002**

**-0.25383901 1.00128866 0.65509951 0.03890494 -0.29947232 0.38347594**

**0.40850895 -0.55553625 0.56699214 0.6736456 0.57202613]**

**[ 0.16515151 0.09101577 -0.04028353 0.37195909 0.33191707 -0.06897917**

**-0.06140453 0.65509951 1.00128866 0.12812787 -0.19968518 0.32962651**

**0.3750222 -0.36309504 0.27271444 0.50238599 0.42548915]**

**[ 0.13272942 0.11367165 0.11285614 0.1190116 0.115676 0.11569867**

**0.08130416 0.03890494 0.12812787 1.00128866 0.17952581 0.0269404**

**0.10008351 -0.03197042 -0.04025955 0.11255393 0.00106226]**

**[ 0.17896117 0.20124767 0.28129148 -0.09343665 -0.08091441 0.31760831**

**0.32029384 -0.29947232 -0.19968518 0.17952581 1.00128866 -0.01094989**

**-0.03065256 0.13652054 -0.2863366 -0.09801804 -0.26969106]**

**[ 0.39120081 0.35621633 0.33189629 0.53251337 0.54656564 0.3187472**

**0.14930637 0.38347594 0.32962651 0.0269404 -0.01094989 1.00128866**

**0.85068186 -0.13069832 0.24932955 0.43331936 0.30543094]**

**[ 0.36996762 0.3380184 0.30867133 0.49176793 0.52542506 0.30040557**

**0.14208644 0.40850895 0.3750222 0.10008351 -0.03065256 0.85068186**

**1.00128866 -0.16031027 0.26747453 0.43936469 0.28990033]**

**[ 0.09575627 0.17645611 0.23757707 -0.38537048 -0.29500852 0.28006379**

**0.23283016 -0.55553625 -0.36309504 -0.03197042 0.13652054 -0.13069832**

**-0.16031027 1.00128866 -0.4034484 -0.5845844 -0.30710565]**

**[-0.09034216 -0.16019604 -0.18102711 0.45607223 0.41840277 -0.22975792**

**-0.28115421 0.56699214 0.27271444 -0.04025955 -0.2863366 0.24932955**

**0.26747453 -0.4034484 1.00128866 0.41825001 0.49153016]**

**[ 0.2599265 0.12487773 0.06425192 0.6617651 0.52812713 0.01867565**

**-0.08367612 0.6736456 0.50238599 0.11255393 -0.09801804 0.43331936**

**0.43936469 -0.5845844 0.41825001 1.00128866 0.39084571]**

**[ 0.14694372 0.06739929 -0.02236983 0.49562711 0.47789622 -0.07887464**

**-0.25733218 0.57202613 0.42548915 0.00106226 -0.26969106 0.30543094**

**0.28990033 -0.30710565 0.49153016 0.39084571 1.00128866]]**

Correlation is the normalised version of covariance. Correlation measures the strength and the direction of the linear relationship between two variables. Strength is that is that positively correlated or negatively correlated.

This below snippet is the correlation matrix. We can clearly understand the correlation matrix which gives the strength and the relationship between the variables. The correlation matrix before scaling and after scaling will remain

the same. From this snippet we can understand which variables are highly positively correlated and the variables which are highly negatively correlated. We can also understand the variables which are moderately correlated with each other. We can see that application, acceptance, enrolment and fulltime graduates are highly positively correlated Also the top 10 percentage and top

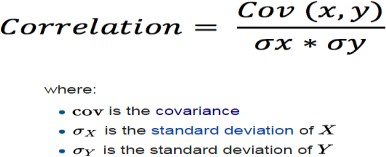
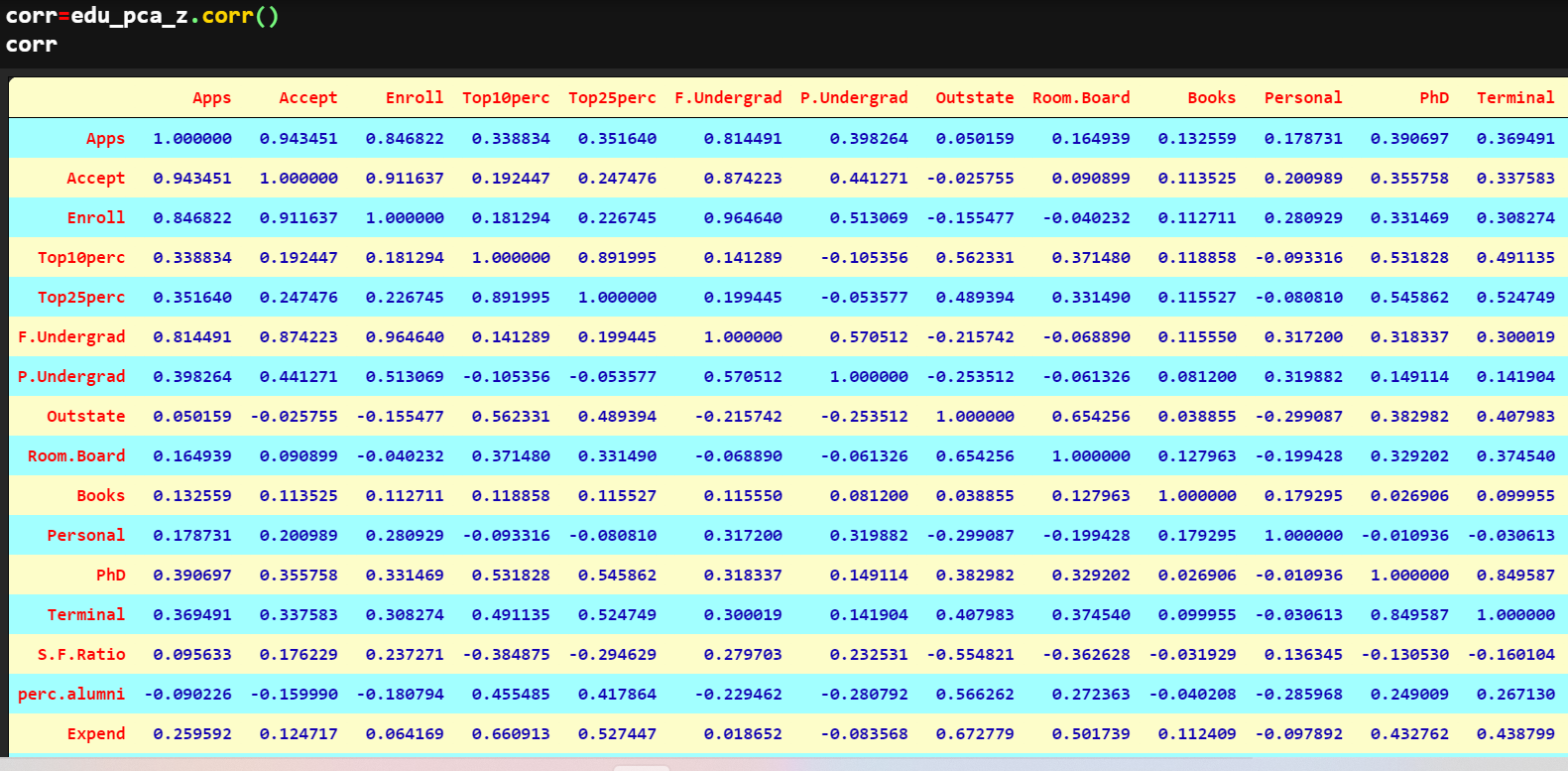
25 percentage are highly positively correlated.

Figure Correlation Formula

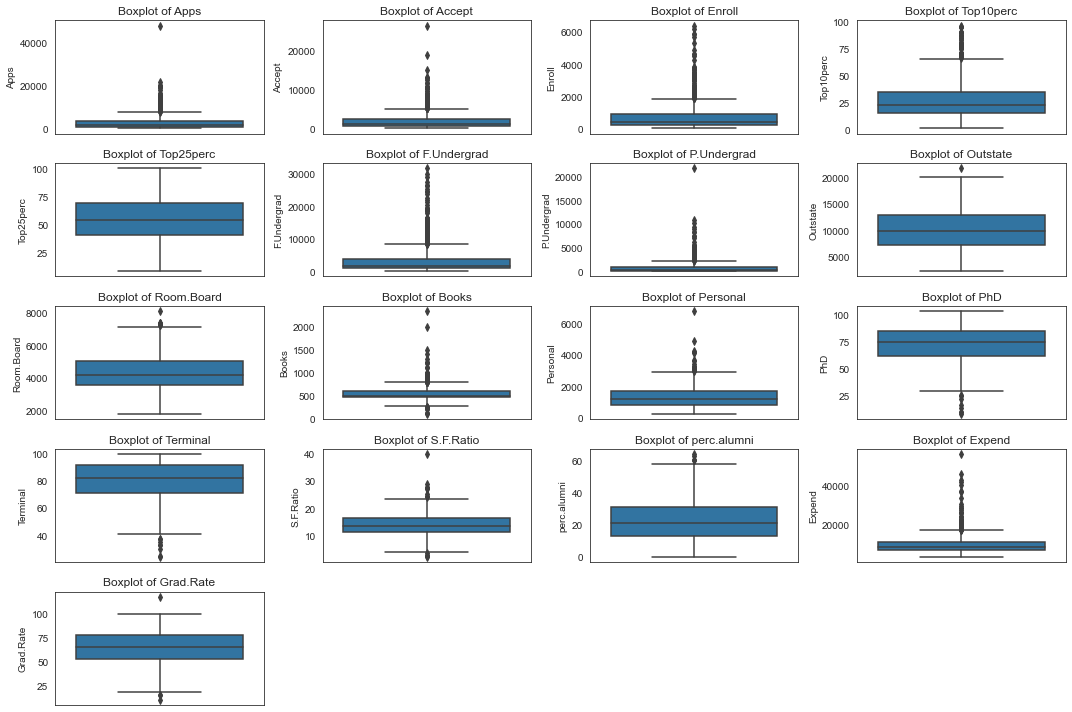
Figure Correlation Matrix



## 2.4 Check the dataset for outliers before and after scaling. What insight do you derive here?

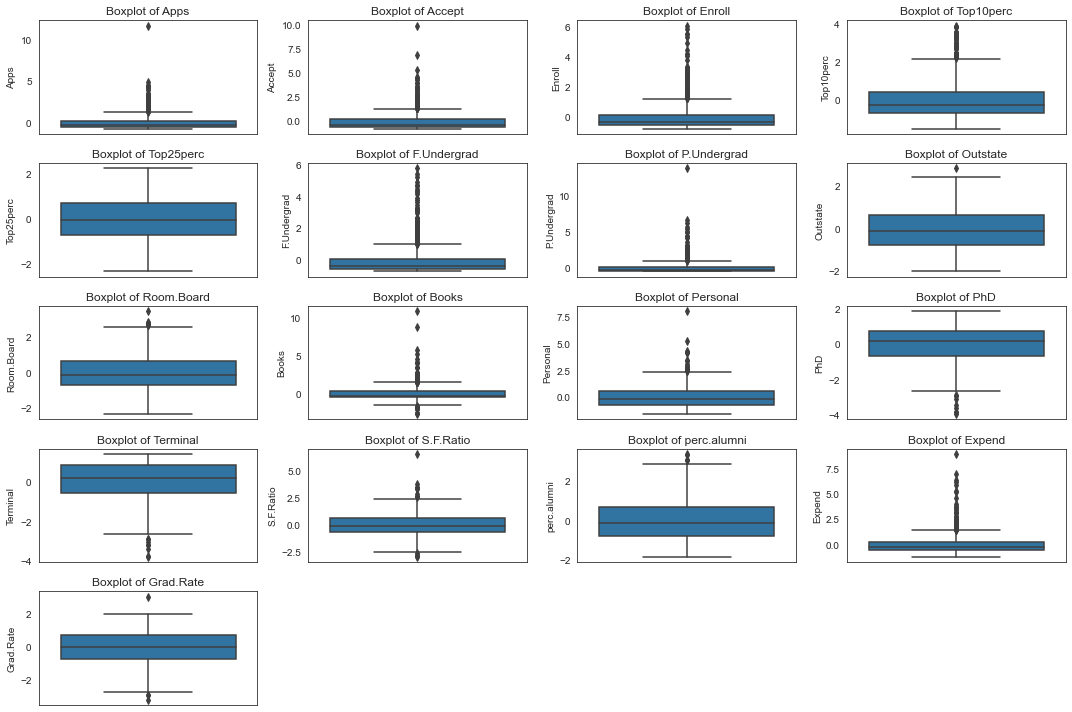
**Checking the data for outliers before scaling**

Figure Boxplot Before Scaling

****

**Checking the data for outliers after scaling**

Figure Boxplot of Data after Scaling

****

**Inference**

The outliers are still present in dataset.

Reason: Scaling does not remove outliers, scaling scales the values on a Z score distribution. We can further process the data to remove outliers.

For example, if we wish to remove outliers, we can consider taking 95 percentile value for higher outliers and 5 percentile value for lower-level outliers or either we can remove them or impute them with IQR values. Or we can impute them with IQRs.

## 2.5 Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both]

Before extracting eigen vectors and eigen values we will perform two tests.

* *Bartlett Sphericity test:*

Ho: Correlations are not significant

H1: There are significant correlations

Reject Ho if p-value is less than 0.05

After performing this test, we get the p-value, which is, 0.0, it is less the 0.05, so we reject null hypothesis and conclude that there is significant correlation between the variables. In case we fail to reject the null hypothesis then we cannot perform PCA.

* *KMO Test:* this test is used to check the adequacy of the data. If the KMO model value is above 0.7 is good, below 0.5 is not acceptable.

Here after performing the test, we get KMO model value is 0.8131. So, data is good to go forward.

**EIGEN VECTORS**

array([[ 2.48765602e-01, 2.07601502e-01, 1.76303592e-01,

3.54273947e-01, 3.44001279e-01, 1.54640962e-01,

2.64425045e-02, 2.94736419e-01, 2.49030449e-01,

6.47575181e-02, -4.25285386e-02, 3.18312875e-01,

3.17056016e-01, -1.76957895e-01, 2.05082369e-01,

3.18908750e-01, 2.52315654e-01],

[ 3.31598227e-01, 3.72116750e-01, 4.03724252e-01,

-8.24118211e-02, -4.47786551e-02, 4.17673774e-01,

3.15087830e-01, -2.49643522e-01, -1.37808883e-01,

5.63418434e-02, 2.19929218e-01, 5.83113174e-02,

4.64294477e-02, 2.46665277e-01, -2.46595274e-01,

-1.31689865e-01, -1.69240532e-01],

[-6.30921033e-02, -1.01249056e-01, -8.29855709e-02,

3.50555339e-02, -2.41479376e-02, -6.13929764e-02,

1.39681716e-01, 4.65988731e-02, 1.48967389e-01,

6.77411649e-01, 4.99721120e-01, -1.27028371e-01,

-6.60375454e-02, -2.89848401e-01, -1.46989274e-01,

2.26743985e-01, -2.08064649e-01],

[ 2.81310530e-01, 2.67817346e-01, 1.61826771e-01,

-5.15472524e-02, -1.09766541e-01, 1.00412335e-01,

-1.58558487e-01, 1.31291364e-01, 1.84995991e-01,

8.70892205e-02, -2.30710568e-01, -5.34724832e-01,

-5.19443019e-01, -1.61189487e-01, 1.73142230e-02,

7.92734946e-02, 2.69129066e-01],

[ 5.74140964e-03, 5.57860920e-02, -5.56936353e-02,

-3.95434345e-01, -4.26533594e-01, -4.34543659e-02,

3.02385408e-01, 2.22532003e-01, 5.60919470e-01,

-1.27288825e-01, -2.22311021e-01, 1.40166326e-01,

2.04719730e-01, -7.93882496e-02, -2.16297411e-01,

7.59581203e-02, -1.09267913e-01],

[-1.62374420e-02, 7.53468452e-03, -4.25579803e-02,

-5.26927980e-02, 3.30915896e-02, -4.34542349e-02,

-1.91198583e-01, -3.00003910e-02, 1.62755446e-01,

6.41054950e-01, -3.31398003e-01, 9.12555212e-02,

1.54927646e-01, 4.87045875e-01, -4.73400144e-02,

-2.98118619e-01, 2.16163313e-01],

[-4.24863486e-02, -1.29497196e-02, -2.76928937e-02,

-1.61332069e-01, -1.18485556e-01, -2.50763629e-02,

6.10423460e-02, 1.08528966e-01, 2.09744235e-01,

-1.49692034e-01, 6.33790064e-01, -1.09641298e-03,

-2.84770105e-02, 2.19259358e-01, 2.43321156e-01,

-2.26584481e-01, 5.59943937e-01],

[-1.03090398e-01, -5.62709623e-02, 5.86623552e-02,

-1.22678028e-01, -1.02491967e-01, 7.88896442e-02,

5.70783816e-01, 9.84599754e-03, -2.21453442e-01,

2.13293009e-01, -2.32660840e-01, -7.70400002e-02,

-1.21613297e-02, -8.36048735e-02, 6.78523654e-01,

-5.41593771e-02, -5.33553891e-03],

[-9.02270802e-02, -1.77864814e-01, -1.28560713e-01,

3.41099863e-01, 4.03711989e-01, -5.94419181e-02,

5.60672902e-01, -4.57332880e-03, 2.75022548e-01,

-1.33663353e-01, -9.44688900e-02, -1.85181525e-01,

-2.54938198e-01, 2.74544380e-01, -2.55334907e-01,

-4.91388809e-02, 4.19043052e-02],

[ 5.25098025e-02, 4.11400844e-02, 3.44879147e-02,

6.40257785e-02, 1.45492289e-02, 2.08471834e-02,

-2.23105808e-01, 1.86675363e-01, 2.98324237e-01,

-8.20292186e-02, 1.36027616e-01, -1.23452200e-01,

-8.85784627e-02, 4.72045249e-01, 4.22999706e-01,

1.32286331e-01, -5.90271067e-01],

[ 4.30462074e-02, -5.84055850e-02, -6.93988831e-02,

-8.10481404e-03, -2.73128469e-01, -8.11578181e-02,

1.00693324e-01, 1.43220673e-01, -3.59321731e-01,

3.19400370e-02, -1.85784733e-02, 4.03723253e-02,

-5.89734026e-02, 4.45000727e-01, -1.30727978e-01,

6.92088870e-01, 2.19839000e-01],

[ 2.40709086e-02, -1.45102446e-01, 1.11431545e-02,

3.85543001e-02, -8.93515563e-02, 5.61767721e-02,

-6.35360730e-02, -8.23443779e-01, 3.54559731e-01,

-2.81593679e-02, -3.92640266e-02, 2.32224316e-02,

1.64850420e-02, -1.10262122e-02, 1.82660654e-01,

3.25982295e-01, 1.22106697e-01],

[ 5.95830975e-01, 2.92642398e-01, -4.44638207e-01,

1.02303616e-03, 2.18838802e-02, -5.23622267e-01,

1.25997650e-01, -1.41856014e-01, -6.97485854e-02,

1.14379958e-02, 3.94547417e-02, 1.27696382e-01,

-5.83134662e-02, -1.77152700e-02, 1.04088088e-01,

-9.37464497e-02, -6.91969778e-02],

[ 8.06328039e-02, 3.34674281e-02, -8.56967180e-02,

-1.07828189e-01, 1.51742110e-01, -5.63728817e-02,

1.92857500e-02, -3.40115407e-02, -5.84289756e-02,

-6.68494643e-02, 2.75286207e-02, -6.91126145e-01,

6.71008607e-01, 4.13740967e-02, -2.71542091e-02,

7.31225166e-02, 3.64767385e-02],

[ 1.33405806e-01, -1.45497511e-01, 2.95896092e-02,

6.97722522e-01, -6.17274818e-01, 9.91640992e-03,

2.09515982e-02, 3.83544794e-02, 3.40197083e-03,

-9.43887925e-03, -3.09001353e-03, -1.12055599e-01,

1.58909651e-01, -2.08991284e-02, -8.41789410e-03,

-2.27742017e-01, -3.39433604e-03],

[ 4.59139498e-01, -5.18568789e-01, -4.04318439e-01,

-1.48738723e-01, 5.18683400e-02, 5.60363054e-01,

-5.27313042e-02, 1.01594830e-01, -2.59293381e-02,

2.88282896e-03, -1.28904022e-02, 2.98075465e-02,

-2.70759809e-02, -2.12476294e-02, 3.33406243e-03,

-4.38803230e-02, -5.00844705e-03],

[ 3.58970400e-01, -5.43427250e-01, 6.09651110e-01,

-1.44986329e-01, 8.03478445e-02, -4.14705279e-01,

9.01788964e-03, 5.08995918e-02, 1.14639620e-03,

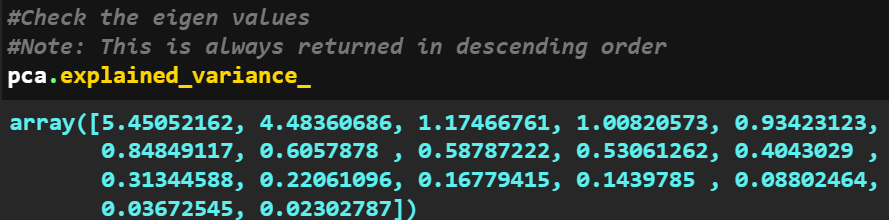
7.72631963e-04, -1.11433396e-03, 1.38133366e-02,

6.20932749e-03, -2.22215182e-03, -1.91869743e-02,

-3.53098218e-02, -1.30710024e-02]])

**EIGEN VALUES**

Figure Eigen Values

****

## 2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features.

PCA is performed and it is exported into a data frame. After PCA the multi collinearity is highly reduced.

Figure DataFrame after performing PCA

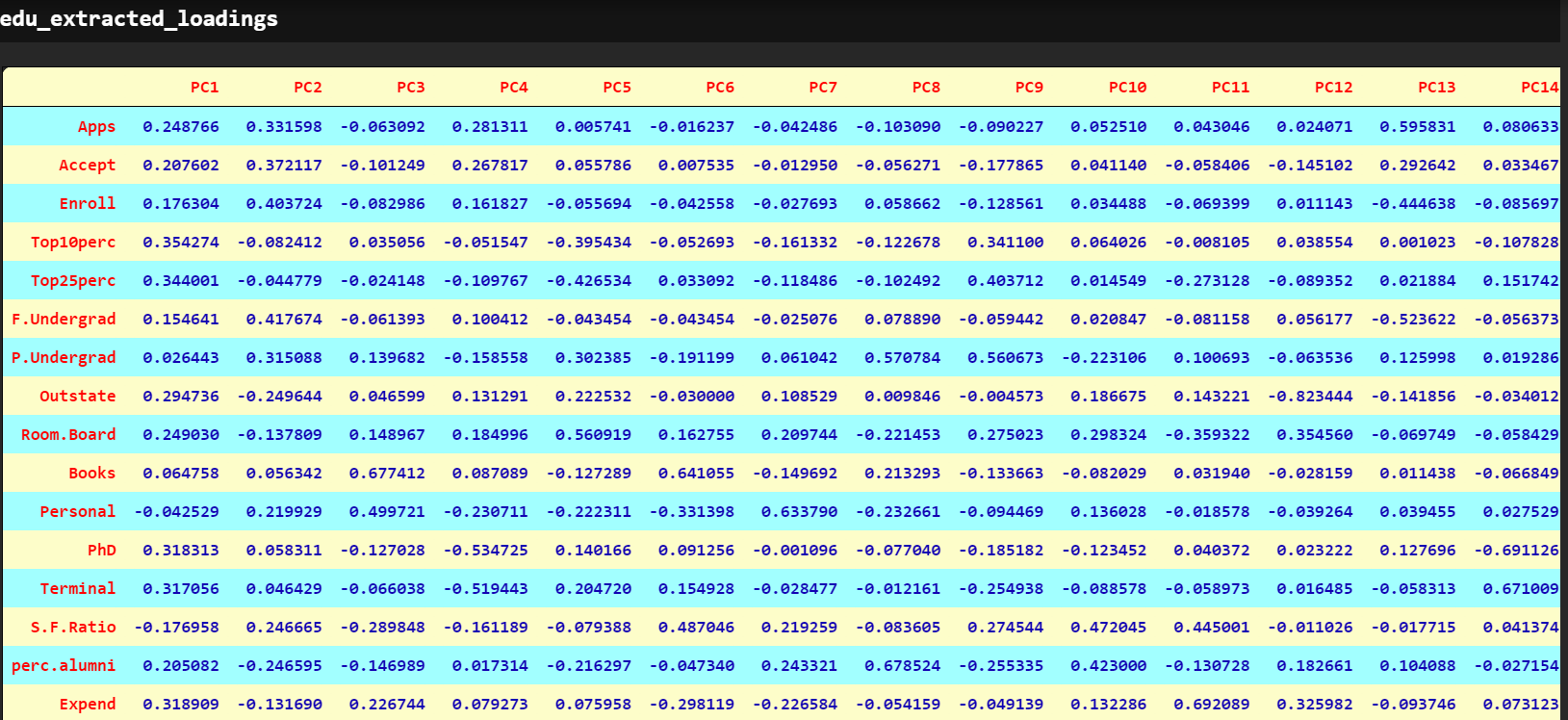
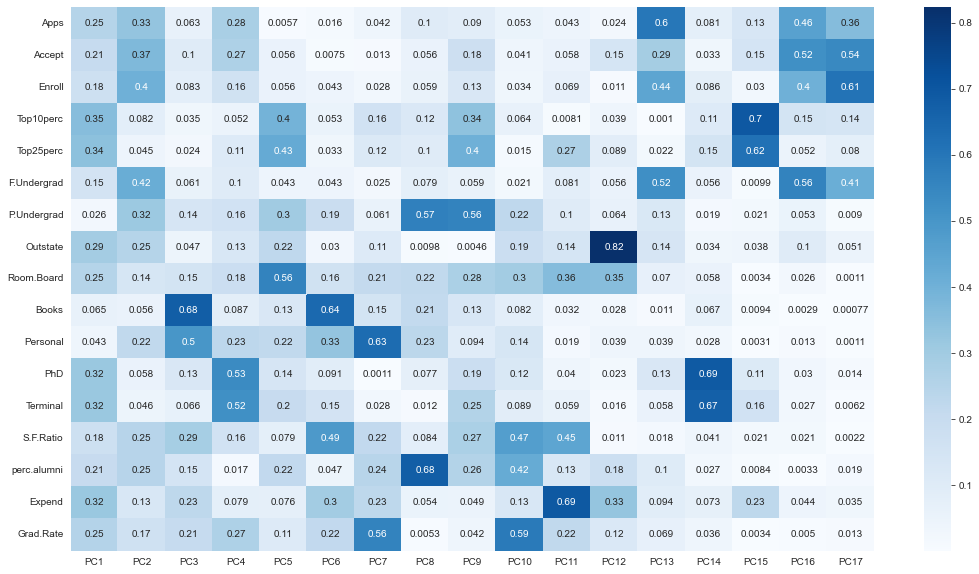


Figure Highly Reduced Collinearity after Performing PCA



## 2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]

**The Linear Equation of first component:**

**0.25 \* Apps + 0.21 \* Accept + 0.18 \* Enroll + 0.35 \* Top10perc + 0.34 \* Top25perc + 0.15 \* F.Undergrad + 0.03 \* P.Undergrad + 0.29 \* Outstate + 0.25 \* Room.Board + 0.06 \* Books + -0.04 \* Personal + 0.32 \* PhD + 0.32 \* Terminal + -0.18 \* S.F.Ratio + 0.21 \* perc.alumni + 0.32 \* Expend + 0.25 \* Grad.Rate**

## 2.8 Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?

Adding the Eigen values, we will get sum of 100.

**array([ 32.02, 58.36, 65.26, 71.18, 76.67, 81.66, 85.22, 88.67,**

**91.79, 94.16, 96. , 97.3 , 98.29, 99.13, 99.65, 99.86,**

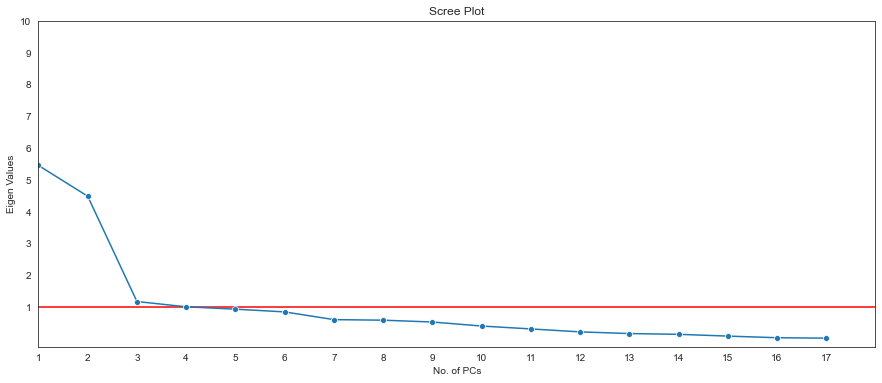
**100. ])**

To decide the optimum number of principal components we use following graphs and rules.

Rule I

As per the Kaiser's rule, we need to consider only those principal components that have an eigen value of more than 1, since if eigen value is less than 1, it means that the PC does not do the job of even a lone original variable. So, going by this rule it appears that 6 is the ideal number of principal components to take to reduce the dimensions of the original dataset. Because PC6 is just below 1, rest others are not doing the job of a lone original variable. See the graph below.

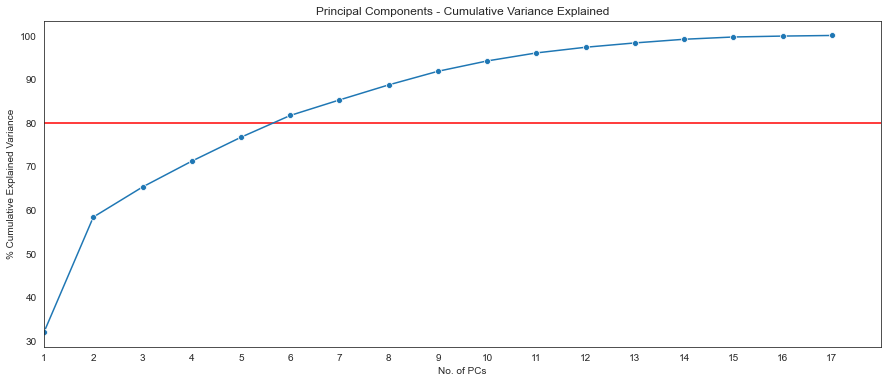
Figure Scree Plot



Rule II

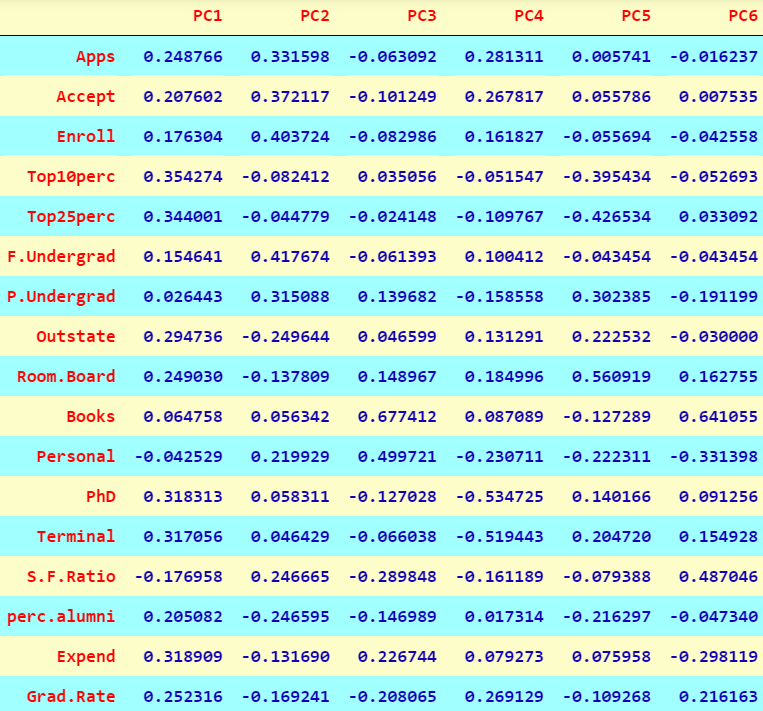
Looking at the cumulative variance explained plot, assuming that we are ok with 80% of the variance in data being explained while reducing the features, we see that, here too considering only the top 6 principal components does the job. So, we will consider only the top 6 principal components, thereby reducing the original dimensions of the data from 17 numeric features to just 6 components, while still being able to account for 80% of the original variations in data. Check the plot below.

Figure Cumulative Variance Explained



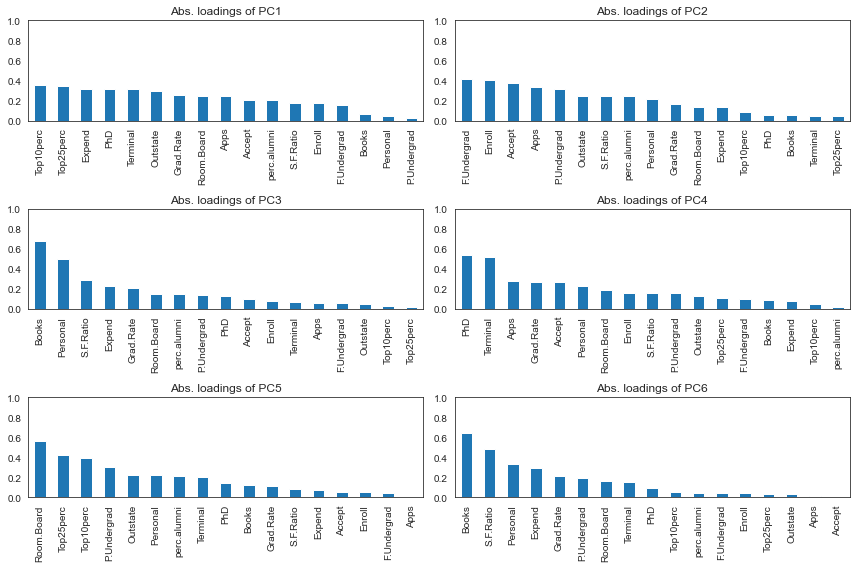
Using these six components we can reduce the dimensions and build a new dataframe.

Figure DataFrame of Selected PCs



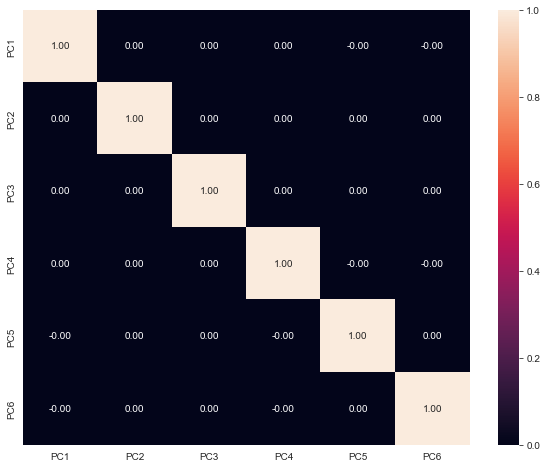
Now we can check graphically as to how the original features matter to each PC. We used absolute values in this graph.

Figure Original Features and selected PCs



Check the presence of correlations among the selected PCs. We can observe that there is no correlation between the selected PCs.

Figure Heatmap of Selected PCs



# 2.9 Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [Hint: Write Interpretations of the Principal Components Obtained]

This business case study is about education dataset which contain the names of various colleges, which has various details of colleges and university. To understand more about the dataset, we perform univariate analysis and multivariate analysis which gives us the understanding about the variables. From analysis we can understand the distribution of the dataset, skew, and patterns in the dataset. From multivariate analysis we can understand the correlation of variables. Inference of multivariate analysis shows we can understand multiple variables highly correlated with each other. The scaling helps the dataset to standardize the variable in one scale. Outliers can be imputed using IQR values once the values are imputed, we can perform PCA. The principal component analysis is used reduce the multicollinearity between the variables. Depending on the variance of the dataset we can reduce the PCA components. The PCA components for this business case is 6 where we could understand the maximum variance of the dataset. Using the components, we can now understand the reduced multicollinearity in the dataset. With this analysis we can perform further analysis and model building. PCA will improve the efficiency of machine learning models.