Table of Contents

[Problem 1 : Research laboratory new compound experiment 6](#_Toc63609926)

[Problem Statement: 6](#_Toc63609927)

[Data Description: 6](#_Toc63609928)

[Domain: 6](#_Toc63609929)

[Context: 6](#_Toc63609930)

[Attribute Information: 6](#_Toc63609931)

[Basic EDA summary:- 7](#_Toc63609932)

[1.1) State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables ‘A’ and ‘B’ individually. [both statement and statistical form like Ho=mu, Ha>mu] 7](#_Toc63609933)

[Formulate the Null and Alternate Hypothesis for Variable 'A' 7](#_Toc63609934)

[Formulate the Null and Alternate Hypothesis for Variable 'B' 7](#_Toc63609935)

[1.2) Perform one-way ANOVA for variable ‘A’ with respect to the variable ‘Relief’. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results. 8](#_Toc63609936)

[Plotting histogram and boxplot for different level of ingredient A 8](#_Toc63609937)

[Anova Assumptions 10](#_Toc63609938)

[Test for Normality - Shapiro- Wilk Test 10](#_Toc63609939)

[Levene's Test :- Computes homogeneity of variance across groups. 10](#_Toc63609940)

[One Way Anova for Ingredient ‘A’ 11](#_Toc63609941)

[One Way Anova Test Result: 11](#_Toc63609942)

[Conclusion:- 11](#_Toc63609943)

[Tukey Honest Significant Difference for Ingredient A 11](#_Toc63609944)

[Test Result:- 11](#_Toc63609945)

[Conclusion:- 11](#_Toc63609946)

[Point Plot for ingredient A 11](#_Toc63609947)

[1.3) Perform one-way ANOVA for variable ‘B’ with respect to the variable ‘Relief’. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results. 12](#_Toc63609948)

[Plotting histogram and boxplot for different level of ingredient B 12](#_Toc63609949)

[Anova Assumptions 14](#_Toc63609950)

[Test for Normality - Shapiro- Wilk Test 14](#_Toc63609951)

[Levene's Test :- Computes homogeneity of variance across groups. 14](#_Toc63609952)

[One Way Anova for Ingredient B 15](#_Toc63609953)

[One Way Anova Test Result: 15](#_Toc63609954)

[Conclusion:- 15](#_Toc63609955)

[Tukey Honest Significant Difference for Ingredient B 15](#_Toc63609956)

[Test Result:- 15](#_Toc63609957)

[Conclusion:- 15](#_Toc63609958)

[Point Plot for ingredient B 16](#_Toc63609959)

[1.4) Analyse the effects of one variable on another with the help of an interaction plot. What is the interaction between the two treatments? 16](#_Toc63609960)

[Point plot for checking interaction effect 17](#_Toc63609961)

[1.5) Perform a two-way ANOVA based on the different ingredients (variable ‘A’ & ‘B’ along with their interaction 'A\*B') with the variable 'Relief' and state your results. 17](#_Toc63609962)

[Test for Normality - Shapiro- Wilk Test 17](#_Toc63609963)

[Two Way Anova for Ingredient A, B and their interaction 18](#_Toc63609964)

[Conclusion:- 18](#_Toc63609965)

[1.6) Mention the business implications of performing ANOVA for this particular case study. 18](#_Toc63609966)

[Observations 18](#_Toc63609967)

[Assumptions 19](#_Toc63609968)

[Recommendation to the Business :- 20](#_Toc63609969)

[Problem 2 : Research laboratory new compound experiment 21](#_Toc63609970)

[Problem Statement: 21](#_Toc63609971)

[Data Description: 21](#_Toc63609972)

[Domain: 21](#_Toc63609973)

[Context: 21](#_Toc63609974)

[Attribute Information: 21](#_Toc63609975)

[2.1) Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented. 22](#_Toc63609976)

[Basic EDA summary:- 22](#_Toc63609977)

[Univariate Analysis 22](#_Toc63609978)

[Summary 22](#_Toc63609979)

[Checking IQR, Coeffiecient of Variation, IQR, lower range and upper range of numerical cols with summary 23](#_Toc63609980)

[Histogram 23](#_Toc63609981)

[Box Plots 25](#_Toc63609982)

[Multivariate analysis 27](#_Toc63609983)

[Correlation matrix 27](#_Toc63609984)

[Heat Map 27](#_Toc63609985)

[Pairplots 28](#_Toc63609986)

[Pairplot for columns 'Apps','Accept','Enroll','FUndergrad' 29](#_Toc63609987)

[Pairplot for columns 'Top10perc','Top25perc','Expend','Outstate','GradRate 30](#_Toc63609988)

[Pairplot for columns 'PhD','Terminal' 31](#_Toc63609989)

[Pairplot for 'FUndergrad','PUndergrad' 31](#_Toc63609990)

[Pairplot for columns 'Outstate','Expend','RoomBoard','PercAlumni','GradRate', 'SFRatio' 32](#_Toc63609991)

[Outliers Treatment 33](#_Toc63609992)

[Box plot before outlier treatment 34](#_Toc63609993)

[Box plot after outlier treatment 35](#_Toc63609994)

[Pandas profile report 36](#_Toc63609995)

[2.2) Scale the variables and write the inference for using the type of scaling function for this case study. 36](#_Toc63609996)

[Variance/Standard Deviation 36](#_Toc63609997)

[Standardizing 37](#_Toc63609998)

[2.3) Comment on the comparison between covariance and the correlation matrix. 38](#_Toc63609999)

[Covariance matrix of original dataset 38](#_Toc63610000)

[Correlation matrix of original data set 39](#_Toc63610001)

[Summary of scaled dataset 39](#_Toc63610002)

[Covariance matrix of scaled dataset 40](#_Toc63610003)

[Correlation matrix of scaled data set 40](#_Toc63610004)

[Conclusion 40](#_Toc63610005)

[Q2.4 Check the dataset for outliers before and after scaling. Draw your inferences from this exercise. 41](#_Toc63610006)

[Plotting box plot on initial data set without outlier treatment 41](#_Toc63610007)

[Plotting box plot after scaling and without any outlier treatment 42](#_Toc63610008)

[Conclusion 42](#_Toc63610009)

[2.5) Build the covariance matrix and calculate the eigenvalues and the eigenvector. 42](#_Toc63610010)

[Create covariance matrix 42](#_Toc63610011)

[Get eigen values and eigen vector 44](#_Toc63610012)

[2.6) Write the explicit form of the first PC (in terms of Eigen Vectors). 47](#_Toc63610013)

[Eigen vector value with 0 index 47](#_Toc63610014)

[PC1 explicit form 47](#_Toc63610015)

[2.7) Discuss 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? Perform PCA and export the data of the Principal Component scores into a data frame. 48](#_Toc63610016)

[Cumulative Distribution of Eigenvalues 48](#_Toc63610017)

[Scree Plot 48](#_Toc63610018)

[Plot Cumulative explained variance and individual explained variance vs Principal Components 49](#_Toc63610019)

[Plot Cumulative explained variance vs No of features 49](#_Toc63610020)

[Decision on the optimum number of principal components:- 50](#_Toc63610021)

[Perform PCA and export the data of the Principal Component scores into a data frame 51](#_Toc63610022)

[Data frame showing PC components 51](#_Toc63610023)

[Correlation between components and features 52](#_Toc63610024)

[2.8) Mention the business implication of using the Principal Component Analysis for this case study. 52](#_Toc63610025)

[PCA 52](#_Toc63610026)

[Business Implication: 52](#_Toc63610027)

# Problem 1 : Research laboratory new compound experiment

## ****Problem Statement:****

A research laboratory was developing a new compound for the relief of severe cases of hay fever. In an experiment with 36 volunteers, the amounts of the two active ingredients (A & B) in the compound were varied at three levels each. Randomization was used in assigning four volunteers to each of the nine treatments. The data on hours of relief can be found in the following .csv file: Fever.csv

## 

## Data Description:

The data consists of 36 volunteers, the amounts of the two active ingredients (A & B) in the compound were varied at three levels each. Randomization was used in assigning four volunteers to each of the nine treatments. The data on hours of relief can be found in the following .csv file: Fever.csv

## 

## Domain:

Healthcare

## 

## Context:

A research laboratory was developing a new compound for the relief of severe cases of hay fever. In an experiment with 36 volunteers, the amounts of the two active ingredients (A & B) in the compound were varied at three levels each. Randomization was used in assigning four volunteers to each of the nine treatments.

## Attribute Information:

* A : Ingredient A in compound
* B : Ingredient B in compound
* Volunteer : Volunteer no (1 to 4)
* Relief : Hours of reilef

## Basic EDA summary:-

* Data contains 36 rows and 4 columns (A,B, Volunteer, Relief)
* Column A, B and Volunteer seems to be a discrete numeric variable while Relief is a continuous numeric variable.
* There is no null data and duplicate data in any of these columns
* There are 12 rows each for 3 different ingredient values of categorical variable A (1,2,3)
* There are 12 rows each for 3 different ingredient values of categorical variable B (1,2,3)
* There are 9 rows each for 4 different Volunteers (1,2,3,4)
* From box plot and data summary (shown below), it can be seen that data doesn’t contains an outlier.

## 1.1) State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables ‘A’ and ‘B’ individually. [both statement and statistical form like Ho=mu, Ha>mu]

### Formulate the Null and Alternate Hypothesis for Variable 'A'

**Null Hypothesis (H0)** : The mean relief in hours is same for all three levels (1,2,3) of active ingredient A.

**𝐻0:μ1 =μ2=μ3** where μ is the sample mean and 1,2,3 are the levels for ingredient A

**Alternate Hypothesis (HA)** : The mean relief in hours is different for at-least one level (1,2 and 3) of active ingredient A

**HA : Not All μj are equal** where μj is the sample mean for jth level for ingredient A.

j can be 1,2 or 3

### Formulate the Null and Alternate Hypothesis for Variable 'B'

**Null Hypothesis (H0) : The mean relief in hours is same for all three levels (1,2,3) of active ingredient B.**

**𝐻0:μ1 =μ2=μ3** where μ is the sample mean and 1,2,3 are the levels for ingredient B

**Alternate Hypothesis**HA**: The mean relief in hours is different for at-least one level (1,2 and 3) of active ingredient B**

HA**: Not All**μj**are equal** where μj is the sample mean for jth level for ingredient B.

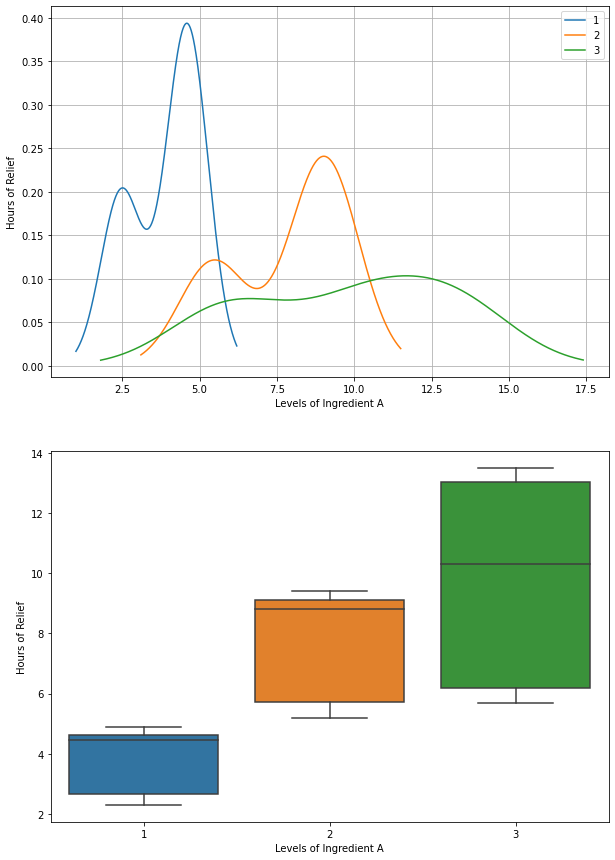
j can be 1,2 or 3

## 1.2) Perform one-way ANOVA for variable ‘A’ with respect to the variable ‘Relief’. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.

### Plotting histogram and boxplot for different level of ingredient A

From the graphs below we can say that :-

* Data for level 1 & 2 looks highly skewed towards left side and doesn’t looks Normal.
* Hours of Relief appears to be more in level 3 as compared to 2 and 1



### Anova Assumptions

1. Independent Sample - Sample should be selected randomly( Equally likely events).

2. Normal Distribution - Distribution of each group should be normal

3. Homogenous Group - Variance between the group should be the same.

### Test for Normality - Shapiro- Wilk Test

Null Hypothesis : The given observation is normally distributed

Alternative Hypothesis : The given observation is not normally distributed

From python, we can find the p value of Relief in hours of active ingredient A level 1, 2 and 3 :-

0.004211828112602234 0.001616060733795166 0.03468279168009758

#### p-value is lower than alpha =0.05, that means, we go ahead with alternative hypothesis as the observations are not normally distributed

### Levene's Test :- Computes homogeneity of variance across groups.

Null Hypothesis: Ho -The variances in the given dataset is the same

Alternative Hypothesis: Ha - The variances in the given dataset is different

From python, we can find the p value of Levene test as

LeveneResult(statistic=4.511350350740447, pvalue=0.018535088623493387)

#### Since p-value is less than 0.05, we shall proceed with Alternative hypothesis that the variances are different

#### Point 2 & 3 of ANOVA's assumption are not true in our case but as per requirement it is told to us that we have to assume the assumptions of ANOVA as true, we are proceeding with the one way Anova test statistics for comparison

### One Way Anova for Ingredient ‘A’

### One Way Anova Test Result:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | df | sum\_sq | mean\_sq | F | PR(>F) |
| C(A) | 2 | 220.02 | 110.01 | 23.46539 | 4.58E-07 |
| Residual | 33 | 154.71 | 4.688182 | NaN | NaN |

### Conclusion:-

Since the p value is less than the significance level (0.05), **we reject the null hypothesis** and conclude that with 95% confidence that the **mean relief in hours is different for at-least one level (1,2 and 3) of active ingredient A**

### Tukey Honest Significant Difference for Ingredient A

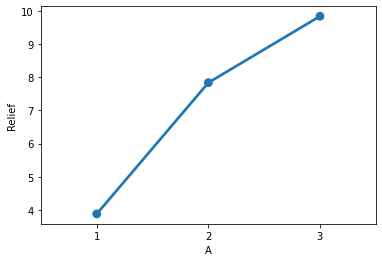
### Test Result:-

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Multiple Comparison of Means - Tukey HSD, FWER=0.05 | | | | | | |
| group1 | group2 | meandiff | p-adj | lower | upper | reject |
|  |  |  |  |  |  |  |
| 1 | 2 | 3.95 | 0.001 | 1.7814 | 6.1186 | TRUE |
| 1 | 3 | 5.95 | 0.001 | 3.7814 | 8.1186 | TRUE |
| 2 | 3 | 2 | 0.0755 | -0.1686 | 4.1686 | FALSE |

### Conclusion:-

From Tukey HSD test, we can say statistically with 95% confidence that mean relief is different for level 1 & 2 and level 1 & 3 of ingredient A. However mean relief is same for ingredient Level 2 and 3

### Point Plot for ingredient A



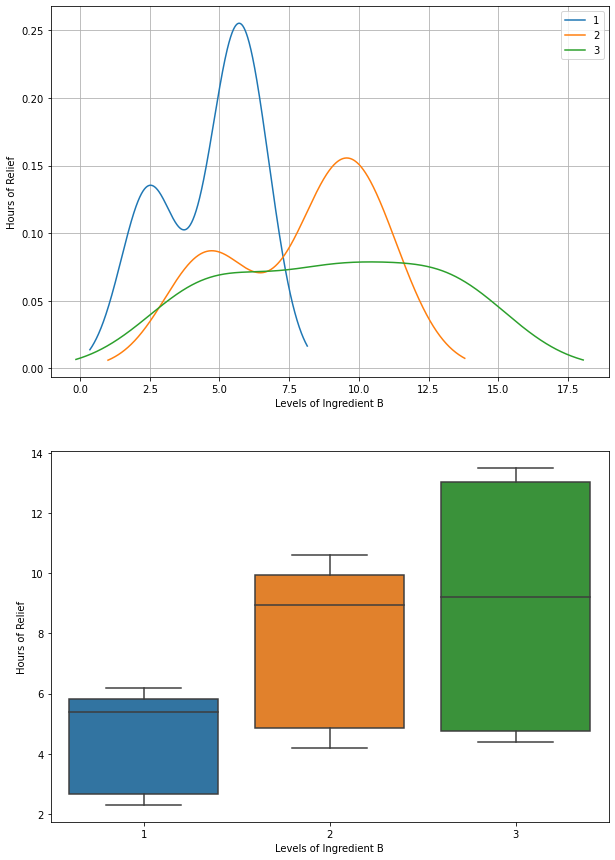
#### As seen in the point plot when level of Ingredient A goes from 2 to 3, slope of the line decreases i.e. mean Relief in hours is not increased much. It is already proven statistically via Tukey HSD test (above).

## 1.3) Perform one-way ANOVA for variable ‘B’ with respect to the variable ‘Relief’. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.

### Plotting histogram and boxplot for different level of ingredient B

From the graphs shown below we can see that :-

* Data for level 1 & 2 looks highly skewed towards left side and doesn’t looks Normal.
* Hours of Relief appears to be more in level 3 as compared to 2 and 1



### Anova Assumptions

1. Independent Sample - Sample should be selected randomly( Equally likely events).

2. Normal Distribution - Distribution of each group should be normal

3. Homogenous Group - Variance between the group should be the same.

### Test for Normality - Shapiro- Wilk Test

Null Hypothesis : The given observation is normally distributed

Alternative Hypothesis : The given observation is not normally distributed

From python, we can find the p value of Relief in hours of active ingredient A level 1, 2 and 3 :-

0.003910908009856939 0.01112875435501337 0.031581711024045944

#### p-value is lower than alpha =0.05, that means, we go ahead with alternative hypothesis as the observations are not normally distributed

### Levene's Test :- Computes homogeneity of variance across groups.

Null Hypothesis: Ho -The variances in the given dataset is the same

Alternative Hypothesis: Ha - The variances in the given dataset is different

From python, we can find the p value of Levene test as

LeveneResult(statistic=2.941356517208818, pvalue=0.06675699295483081)

#### Since p-value is less than 0.05, we shall proceed with Alternative hypothesis that the variances are different

#### Point 2 & 3 of ANOVA's assumption are not true in our case but as per requirement it is told to us that we have to assume the assumptions of ANOVA as true, we are proceeding with the one way Anova test statistics for comparison

### One Way Anova for Ingredient B

### One Way Anova Test Result:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | df | sum\_sq | mean\_sq | F | PR(>F) |
| C(B) | 2 | 123.66 | 61.83 | 8.12678 | 0.00135 |
| Residual | 33 | 251.07 | 7.60818 | NaN | NaN |
|  |  |  |  |  |  |

### Conclusion:-

**Since the p value is less than the significance level (0.05), we reject the null hupothesis and conclude that with 95% confidence that the mean relief in hours is different for at-least one level (1,2 and 3) of active ingredient B**

### Tukey Honest Significant Difference for Ingredient B

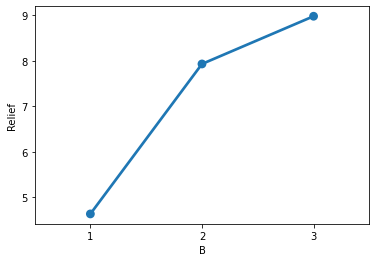
### Test Result:-

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Multiple Comparison of Means - Tukey HSD, FWER=0.05 | | | | | | |
| group1 | group2 | meandiff | p-adj | lower | upper | reject |
|  |  |  |  |  |  |  |
| 1 | 2 | 3.3 | 0.0164 | 0.5374 | 6.0626 | TRUE |
| 1 | 3 | 4.35 | 0.0014 | 1.5874 | 7.1126 | TRUE |
| 2 | 3 | 1.05 | 0.6164 | -1.7126 | 3.8126 | FALSE |

### Conclusion:-

From Tukey HSD test, we can say statistically with 95% confidence that mean relief is different for level 1 & 2 and level 1 & 3 of ingredient B. **However mean relief is same for ingredient B Level 2 and 3**

### Point Plot for ingredient B

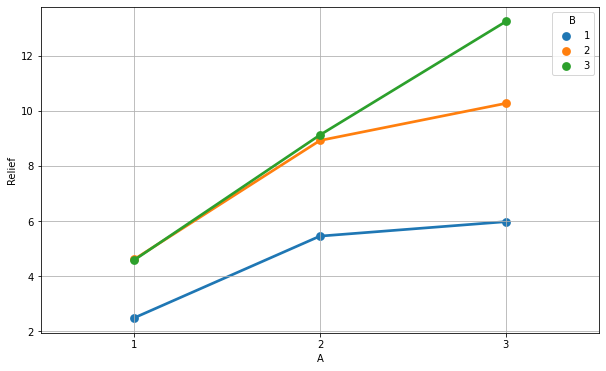


#### As seen in the point plot when level of Ingredient B goes from 2 to 3, slope of the line decreases i.e. mean Relief in hours is not increased much. It is already proven statistically via Tukey HSD test (above).

## 1.4) Analyse the effects of one variable on another with the help of an interaction plot. What is the interaction between the two treatments?

[hint: use the ‘pointplot’ function from the ‘seaborn’ function]

### Point plot for checking interaction effect



From point plot we can see that :-

* Mean relief in hours increases as the level of active ingredient A changes from 1 to 2 to 3. So it appears that percentage level of Ingredient A increases with its Level
* Mean relief in hours increases as the level of active ingredient B changes from 1 to 2 to 3. So it appears that percentage level of Ingredient B increases with its Level
* There is interaction between ingredient A & B
* Mean Relief in hours is approximately the same for level 2 & 3 of ingredient B and level 1 & 2 of ingredient A
* Rise in relief is significantly high if level 3 is used for both ingredient B and ingredient A
* It appears that for ingredient B level 3 concentration level should be used for level 3 concentration of ingredient A, otherwise we would not get substantial benefit in relief

## 1.5) Perform a two-way ANOVA based on the different ingredients (variable ‘A’ & ‘B’ along with their interaction 'A\*B') with the variable 'Relief' and state your results.

### Test for Normality - Shapiro- Wilk Test

Null Hypothesis : The given observation is normally distributed

Alternative Hypothesis : The given observation is not normally distributed

From python, we can find the p value of Relief in hours of active ingredient A level 1, 2 and 3 :-

0.02178293839097023

#### p-value is lower than alpha =0.05, that means, we go ahead with alternative hypothesis as the observations are not normally distributed

### Two Way Anova for Ingredient A, B and their interaction

For 2 way ANOVA's assumption is there that dependent variable has to be approximately normal which is not true in our case. So more samples should be taken but as per requirement it is told to us that we have to assume the assumptions of ANOVA as true, we are proceeding with the two way Anova test statistics for comparison

### Conclusion:-

#### From the 2 way Anova test results we can see that p value is less than the significane level 0.05 for all the three factors (ingredient A, ingredient B, interaction of ingredient A with B).

#### So statistically with 95% confidence we can say that Mean Relief is affected by ingredient A, Ingredient B and interaction of ingredient A & B.

#### p value is least for A, so the most significant factor affecting Relief is ingredient A followed by ingredient B and then the interaction of both the ingredients

## 1.6) Mention the business implications of performing ANOVA for this particular case study.

From the Anova test results (one way and two way) and the visual representation graphs like histograms, box plots, point plot we can see following observations:-

### Observations

#### Independent Variable 'A' affect on Dependent variable Hours of 'Relief'

* Ingredient A has 3 levels (1, 2 and 3). From summary of the sample data it can be seen that A ingredient level 3 has the highest mean relief of 9.8 as compared to level 2 (7.8) and level 1 (3.88)
* On comparing the mean relief of sample data across different levels of ingredient A via 1 way Anova test, we found with 95% confidence level that mean relief for all levels is not equal i.e. for atleast one pair of level 1, 2 and 3 mean relief is different.
* Then on doing Tukey HSD test we have found that mean relief for 2 pair of ingredient A levels are not equal. i.e. Mean relief is same for level 2 & level 3 of ingredient A but it’s different for (level 1 & level 2) and (level 1 & level 3) of ingredient A.
* Further we have added the other independent variables like ingredient B and interaction of ingredient A & B in 2 way Anova Test. Then we have found that ingredient A is the most important attribute in deciding the value for dependent variable Relief.

#### Independent Variable 'B' affect on Dependent variable Hours of 'Relief'

* Ingredient B has 3 levels (1, 2 and 3). From summary of the sample data it can be seen that A ingredient level 3 has the highest mean relief of 9.98 as compared to level 2 (7.9) and level 1 (4.6)
* On comparing the mean relief of sample data across different levels of ingredient B via 1 way Anova test, we found with 95% confidence level that mean relief for all levels is not equal i.e. for atleast one pair of level 1, 2 and 3 mean relief is different.
* Then on doing Tukey HSD test we have found that mean relief for 2 pair of ingredient A levels are not equal. i.e. Mean relief is same for level 2 & level 3 of ingredient B but its different for (level 1 & level 2) and (level 1 & level 3) of ingredient B.
* Further we have added the other independent variables like ingredient A and interaction of ingredient A & B in 2 way Anova Test. Then we have found that ingredient B is the second most important in deciding the value for dependent variable Relief.

#### Independent Variable 'A' and 'B' interaction affect on Dependent variable Hours of 'Relief'

* In 2 way Anova test and point plot , it can be seen that interaction of ingredient A and ingredient B is also important in affecting the value of dependent variable Relief
* Mean Relief in hours is approximately the same for level 2 & 3 of ingredient B and level 1 & 2 of ingredient A
* Rise in relief is significantly high if level 3 is used for both ingredient B and ingredient A
* It appears that for ingredient B level 3 concentration level should be used for level 3 concentration of ingredient A, otherwise we would not get substantial benefit in relief

### Assumptions

* As increasing the level from 1 to 2 and finally to 3, relief in hours is increased so we are assuming that Level 1 has the lowest concentration of ingredient A/B and Level 3 has the highest concentration.
* Level 1 concentration being in low concentration of ingredient A & B respectively, cost of compound will be lowest if level 1 is used for the treatment and similarly for level 3 it would be the highest.
* Level 1 concentration being in low concentration of ingredient A & B respectively, side effects(if any) of ingredients will be at lowest levels if level 1 is used for the treatment and similarly for level 3 it would be the highest.

### **Recommendation** to the Business :-

* Ingredient A , Ingredient B and their interaction are important factors in deciding the average relief in hours.
* Treatment in which ingredient A and ingredient B are at the lowest level (Level 1) gives the least relief to the patients. So business should avoid this treatment to the patients
* Treatment in which concentration is changed from level 2 to level 3 for either ingredient A or B, then correspondingly it should be changed to level 3 for other ingredient too. Otherwise we would not get sufficient increase in relief. For example in treatment with A 2, B 1 and in A 3, B 1 mean relief is not increased significantly. Similarly relief is same for treatment B2, A1 and B3, A1. So in these treatments even if we are increasing the concentration of only one ingredient to highest level 3 and ignoring other ingredient then we are not getting substantial rise in relief. So this should be avoided by Business.
* Treatment in which ingredient A and ingredient B are at the highest level (Level 3) gives the most relief to the patients. So if costing works out for the company and there are no side effects of highest concentration levels of ingredients A & B, this treatment can be chosen by the Business for treatment

# Problem 2 : Research laboratory new compound experiment

## ****Problem Statement:****

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

## Data Description:

The dataset Education - Post 12th Standard.csv is a dataset that contains the names of various colleges.

## Domain:

Education

## Context:

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

## Attribute Information:

* Names: Names of various university and colleges
* Apps: Number of applications received
* Accept: Number of applications accepted
* Enroll: Number of new students enrolled
* Top10perc: Percentage of new students from top 10% of Higher Secondary class
* Top25perc: Percentage of new students from top 25% of Higher Secondary class
* F.Undergrad: Number of full-time undergraduate students
* P.Undergrad: Number of part-time undergraduate students
* Outstate: Number of students for whom the particular college or university is Out-of-state tuition
* Room.Board: Cost of Room and board
* Books: Estimated book costs for a student
* Personal: Estimated personal spending for a student
* PhD: Percentage of faculties with Ph.D.’s
* Terminal: Percentage of faculties with terminal degree
* S.F.Ratio: Student/faculty ratio
* Perc.alumni: Percentage of alumni who donate
* Expend: The Instructional expenditure per student
* Grad.Rate: Graduation rate

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

### Basic EDA summary:-

* Data contains 777 rows and 18 columns.
* Except Names column, every other column is numeric in nature.
* In the numeric columns only one column ‘S.F.Ratio’ contains decimal data
* Columns heading contains dot and spaces which we have fixed by removing spaces and dots from them.
* There is no null/missing data or duplicate data in any of the columns.
* From box plot and data summary (shown below in Univariate analysis), it can be seen that data does contains many outliers.

### Univariate Analysis

### Summary

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **Apps** | 777 | 2571.35 | 2422.2 | 81 | 776 | 1558 | 3624 | 7896 |
| **Accept** | 777 | 1746.28 | 1523.29 | 72 | 604 | 1110 | 2424 | 5154 |
| **Enroll** | 777 | 660.39 | 570.13 | 35 | 242 | 434 | 902 | 1892 |
| **Top10perc** | 777 | 26.84 | 15.58 | 1 | 15 | 23 | 35 | 65 |
| **Top25perc** | 777 | 55.8 | 19.8 | 9 | 41 | 54 | 69 | 100 |
| **FUndergrad** | 777 | 2935.65 | 2700.23 | 139 | 992 | 1707 | 4005 | 8524.5 |
| **PUndergrad** | 777 | 655.88 | 716.27 | 1 | 95 | 353 | 967 | 2275 |
| **Outstate** | 777 | 10440.2 | 4021.71 | 2340 | 7320 | 9990 | 12925 | 21332.5 |
| **RoomBoard** | 777 | 4355.44 | 1090.67 | 1780 | 3597 | 4200 | 5050 | 7229.5 |
| **Books** | 777 | 539.43 | 115.23 | 275 | 470 | 500 | 600 | 795 |
| **Personal** | 777 | 1323.79 | 609.51 | 250 | 850 | 1200 | 1700 | 2975 |
| **PhD** | 777 | 72.77 | 15.95 | 27.5 | 62 | 75 | 85 | 103 |
| **Terminal** | 777 | 79.78 | 14.47 | 39.5 | 71 | 82 | 92 | 100 |
| **SFRatio** | 777 | 14.05 | 3.78 | 4 | 11.5 | 13.6 | 16.5 | 24 |
| **PercAlumni** | 777 | 22.72 | 12.33 | 0 | 13 | 21 | 31 | 58 |
| **Expend** | 777 | 9182.52 | 3396.5 | 3186 | 6751 | 8377 | 10830 | 16948.5 |
| **GradRate** | 777 | 65.47 | 17.14 | 15.5 | 53 | 65 | 78 | 115.5 |

From the summary, we can see that :-

* Huge difference in mean and median value of Apps, Accept, Enroll, FUndergrad, PUndergrad. Data appears to be right/positively skewed
* Except PhD and Terminal columns, maximum value of all other columns seems to be very high and appears to contain outliers.
* Phd and Terminal columns seems to have very low min value. Hence they also appear to contain outliers.
* Max value of GradRate is 118 which is more than 100%. It appears wrong

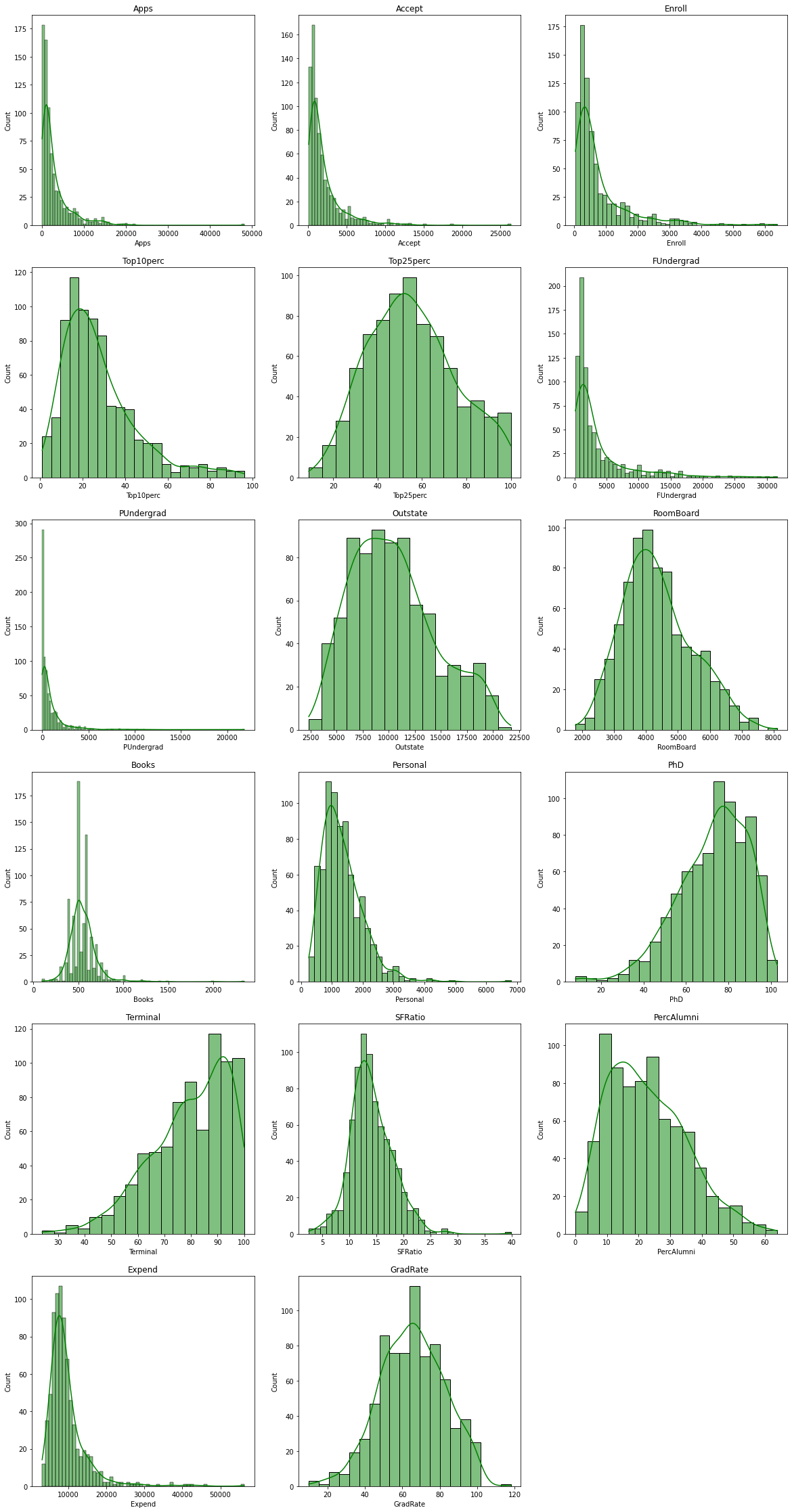
### Checking IQR, Coeffiecient of Variation, IQR, lower range and upper range of numerical cols with summary

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | mean | std | min | 25% | 50% | 75% | max | CV | Skew | IQR | UR | LR |
| Apps | 777 | 3001.64 | 3870.2 | 81 | 776 | 1558 | 3624 | 48094 | 1.29 | 3.72 | 2848 | 7896 | -3496 |
| Accept | 777 | 2018.8 | 2451.11 | 72 | 604 | 1110 | 2424 | 26330 | 1.21 | 3.42 | 1820 | 5154 | -2126 |
| Enroll | 777 | 779.97 | 929.18 | 35 | 242 | 434 | 902 | 6392 | 1.19 | 2.69 | 660 | 1892 | -748 |
| Top10perc | 777 | 27.56 | 17.64 | 1 | 15 | 23 | 35 | 96 | 0.64 | 1.41 | 20 | 65 | -15 |
| Top25perc | 777 | 55.8 | 19.8 | 9 | 41 | 54 | 69 | 100 | 0.35 | 0.26 | 28 | 111 | -1 |
| FUndergrad | 777 | 3699.91 | 4850.42 | 139 | 992 | 1707 | 4005 | 31643 | 1.31 | 2.61 | 3013 | 8524.5 | -3527.5 |
| PUndergrad | 777 | 855.3 | 1522.43 | 1 | 95 | 353 | 967 | 21836 | 1.78 | 5.69 | 872 | 2275 | -1213 |
| Outstate | 777 | 10440.67 | 4023.02 | 2340 | 7320 | 9990 | 12925 | 21700 | 0.39 | 0.51 | 5605 | 21332.5 | -1087.5 |
| RoomBoard | 777 | 4357.53 | 1096.7 | 1780 | 3597 | 4200 | 5050 | 8124 | 0.25 | 0.48 | 1453 | 7229.5 | 1417.5 |
| Books | 777 | 549.38 | 165.11 | 96 | 470 | 500 | 600 | 2340 | 0.3 | 3.49 | 130 | 795 | 275 |
| Personal | 777 | 1340.64 | 677.07 | 250 | 850 | 1200 | 1700 | 6800 | 0.5 | 1.74 | 850 | 2975 | -425 |
| PhD | 777 | 72.66 | 16.33 | 8 | 62 | 75 | 85 | 103 | 0.22 | -0.77 | 23 | 119.5 | 27.5 |
| Terminal | 777 | 79.7 | 14.72 | 24 | 71 | 82 | 92 | 100 | 0.18 | -0.82 | 21 | 123.5 | 39.5 |
| SFRatio | 777 | 14.09 | 3.96 | 2.5 | 11.5 | 13.6 | 16.5 | 39.8 | 0.28 | 0.67 | 5 | 24 | 4 |
| PercAlumni | 777 | 22.74 | 12.39 | 0 | 13 | 21 | 31 | 64 | 0.54 | 0.61 | 18 | 58 | -14 |
| Expend | 777 | 9660.17 | 5221.77 | 3186 | 6751 | 8377 | 10830 | 56233 | 0.54 | 3.46 | 4079 | 16948.5 | 632.5 |
| GradRate | 777 | 65.46 | 17.18 | 10 | 53 | 65 | 78 | 118 | 0.26 | -0.11 | 25 | 115.5 | 15.5 |

Few things which can be noticed here is :-

* From skewness column also we can see that most of the columns are positively skewed/ right skewed
* From LR/UR we can see that most of the numeric variables have min/max values beyond these limits and hence they have outliers

### Histogram

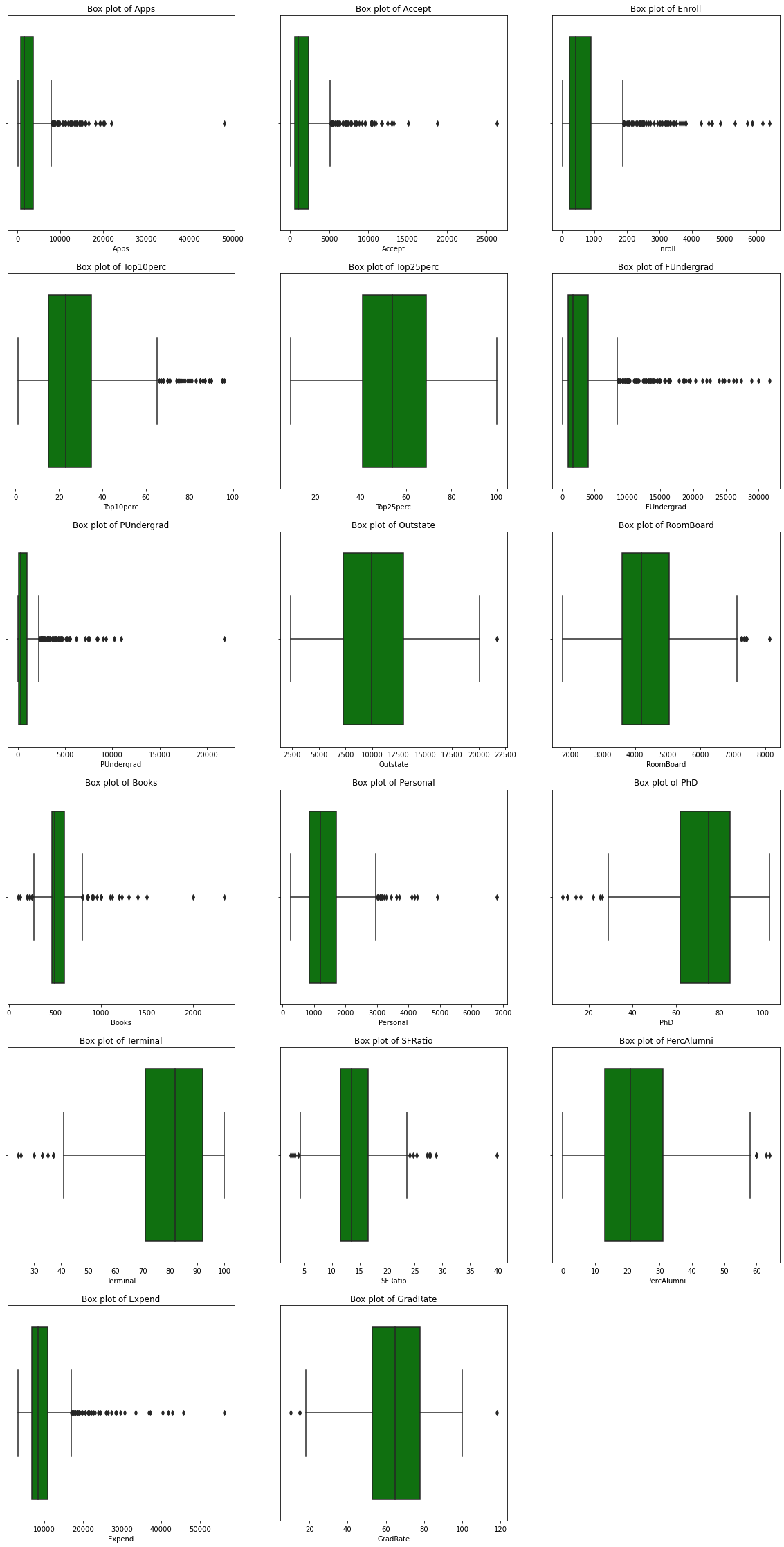


As seen in the histograms,

* Application count goes till 50K, accepted application goes till 25k and enrolled application count goes till 6K.
* Due to outliers present in almost all the columns, data is skewed
* Majority of full time undergrad student in colleges are less than 10k
* Majority of part time undergrad student in colleges are less than 5k
* Books cost contain few outliers whose value seems to be very high and its causing long tail in right
* Majority of faculty in colleges have high PHD degrees and high Terminal degress with them.
* Personal expenditure of most of the students is in 5K to 25K range
* student to faculty ratio is in range of 5 to 25 with 12-16 have the highest colleges
* In majority of the colleges 10-40 perecent of faculty donates
* Instutional expenditure per student is in range of 3k to 20k. Some colleges have very high institutional expenditure
* GradRate of most of the colleges goes from 40 to 100%.

### Box Plots

As evident from box plots shown below, there are outliers present in almost every numeric column except Top25perc.



### Multivariate analysis

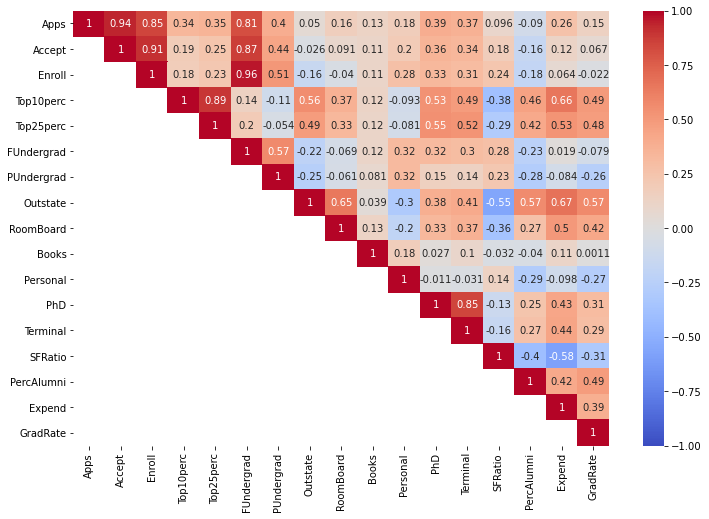
### Correlation matrix

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Apps** | **Accept** | **Enroll** | **Top10perc** | **Top25perc** | **FUndergrad** | **PUndergrad** | **Outstate** | **RoomBoard** | **Books** | **Personal** | **PhD** | **Terminal** | **SFRatio** | **PercAlumni** | **Expend** | **GradRate** |
| Apps | 1 | 0.94 | 0.85 | 0.34 | 0.35 | 0.81 | 0.4 | 0.05 | 0.16 | 0.13 | 0.18 | 0.39 | 0.37 | 0.1 | -0.09 | 0.26 | 0.15 |
| Accept | 0.94 | 1 | 0.91 | 0.19 | 0.25 | 0.87 | 0.44 | -0.03 | 0.09 | 0.11 | 0.2 | 0.36 | 0.34 | 0.18 | -0.16 | 0.12 | 0.07 |
| Enroll | 0.85 | 0.91 | 1 | 0.18 | 0.23 | 0.96 | 0.51 | -0.16 | -0.04 | 0.11 | 0.28 | 0.33 | 0.31 | 0.24 | -0.18 | 0.06 | -0.02 |
| Top10perc | 0.34 | 0.19 | 0.18 | 1 | 0.89 | 0.14 | -0.11 | 0.56 | 0.37 | 0.12 | -0.09 | 0.53 | 0.49 | -0.38 | 0.46 | 0.66 | 0.49 |
| Top25perc | 0.35 | 0.25 | 0.23 | 0.89 | 1 | 0.2 | -0.05 | 0.49 | 0.33 | 0.12 | -0.08 | 0.55 | 0.52 | -0.29 | 0.42 | 0.53 | 0.48 |
| FUndergrad | 0.81 | 0.87 | 0.96 | 0.14 | 0.2 | 1 | 0.57 | -0.22 | -0.07 | 0.12 | 0.32 | 0.32 | 0.3 | 0.28 | -0.23 | 0.02 | -0.08 |
| PUndergrad | 0.4 | 0.44 | 0.51 | -0.11 | -0.05 | 0.57 | 1 | -0.25 | -0.06 | 0.08 | 0.32 | 0.15 | 0.14 | 0.23 | -0.28 | -0.08 | -0.26 |
| Outstate | 0.05 | -0.03 | -0.16 | 0.56 | 0.49 | -0.22 | -0.25 | 1 | 0.65 | 0.04 | -0.3 | 0.38 | 0.41 | -0.55 | 0.57 | 0.67 | 0.57 |
| RoomBoard | 0.16 | 0.09 | -0.04 | 0.37 | 0.33 | -0.07 | -0.06 | 0.65 | 1 | 0.13 | -0.2 | 0.33 | 0.37 | -0.36 | 0.27 | 0.5 | 0.42 |
| Books | 0.13 | 0.11 | 0.11 | 0.12 | 0.12 | 0.12 | 0.08 | 0.04 | 0.13 | 1 | 0.18 | 0.03 | 0.1 | -0.03 | -0.04 | 0.11 | 0 |
| Personal | 0.18 | 0.2 | 0.28 | -0.09 | -0.08 | 0.32 | 0.32 | -0.3 | -0.2 | 0.18 | 1 | -0.01 | -0.03 | 0.14 | -0.29 | -0.1 | -0.27 |
| PhD | 0.39 | 0.36 | 0.33 | 0.53 | 0.55 | 0.32 | 0.15 | 0.38 | 0.33 | 0.03 | -0.01 | 1 | 0.85 | -0.13 | 0.25 | 0.43 | 0.31 |
| Terminal | 0.37 | 0.34 | 0.31 | 0.49 | 0.52 | 0.3 | 0.14 | 0.41 | 0.37 | 0.1 | -0.03 | 0.85 | 1 | -0.16 | 0.27 | 0.44 | 0.29 |
| SFRatio | 0.1 | 0.18 | 0.24 | -0.38 | -0.29 | 0.28 | 0.23 | -0.55 | -0.36 | -0.03 | 0.14 | -0.13 | -0.16 | 1 | -0.4 | -0.58 | -0.31 |
| PercAlumni | -0.09 | -0.16 | -0.18 | 0.46 | 0.42 | -0.23 | -0.28 | 0.57 | 0.27 | -0.04 | -0.29 | 0.25 | 0.27 | -0.4 | 1 | 0.42 | 0.49 |
| Expend | 0.26 | 0.12 | 0.06 | 0.66 | 0.53 | 0.02 | -0.08 | 0.67 | 0.5 | 0.11 | -0.1 | 0.43 | 0.44 | -0.58 | 0.42 | 1 | 0.39 |
| GradRate | 0.15 | 0.07 | -0.02 | 0.49 | 0.48 | -0.08 | -0.26 | 0.57 | 0.42 | 0 | -0.27 | 0.31 | 0.29 | -0.31 | 0.49 | 0.39 | 1 |

### Heat Map

We can see in heatmap that

* Apps , Accept, Enroll & Fundergrad are highly correlated with each other
* Top 10perc is positively correlated with top25perc , Expend , Outstate, GradRate
* PhD is highly correlated with Terminal
* FUndergrad is correlated with Pundergrad
* Outstate is positively correleted with Room Board, Expend, PercAlumni, GradRate and negatively correlated with SF ratio
* SF ratio is negatively correlated with Expend

****

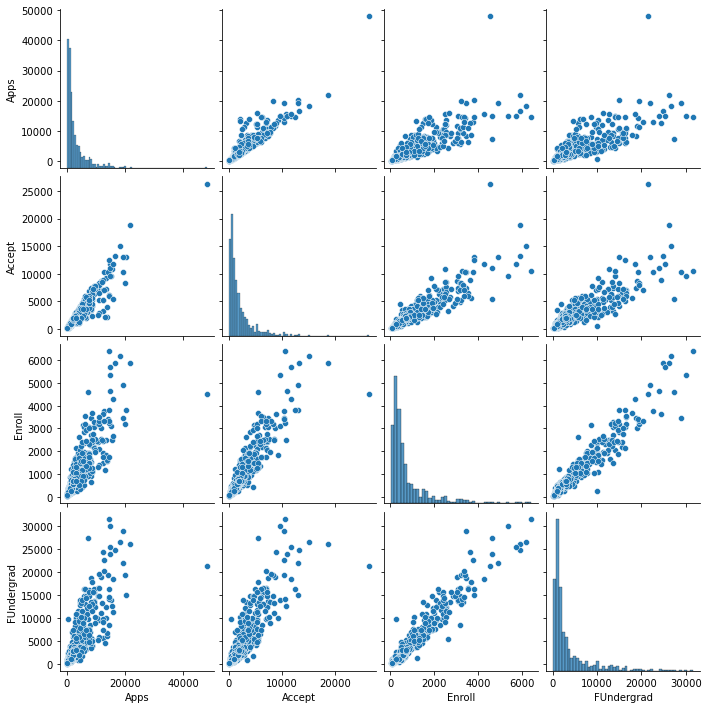
From heatmap and correlation matrix we can find that :-

* Apps , Accept, Enroll & Fundergrad are highly correlated with each other
* Top 10perc is positively correlated with top25perc , Expend , Outstate, GradRate
* PhD is highly correlated with Terminal
* FUndergrad is correlated with Pundergrad
* Outstate is positively correleted with Room Board, Expend, PercAlumni, GradRate and negatively correlated with SF ratio
* SF ratio is negatively correlated with Expend

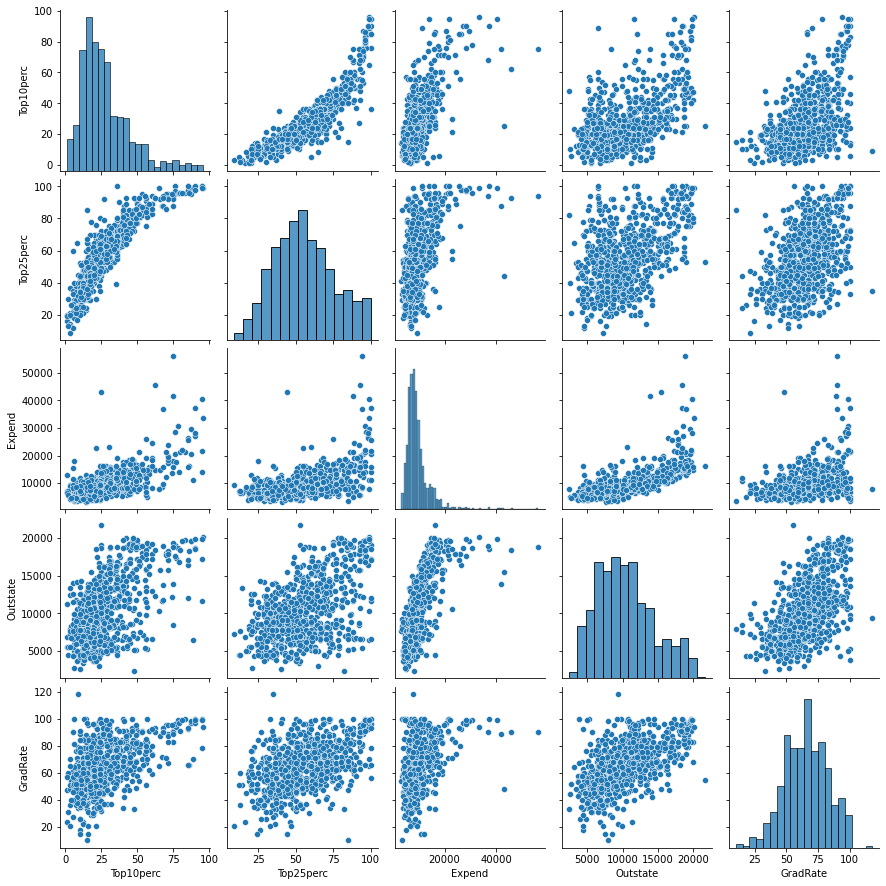
### Pairplots

To check the correlation of 2 columns in more detail we can draw **pairplots/Scatter Diagram**

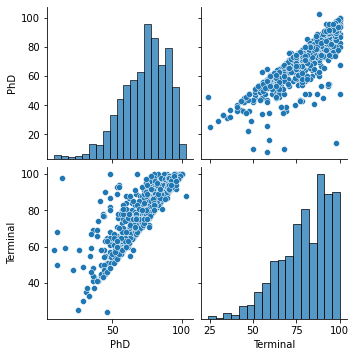
### Pairplot for columns 'Apps','Accept','Enroll','FUndergrad'



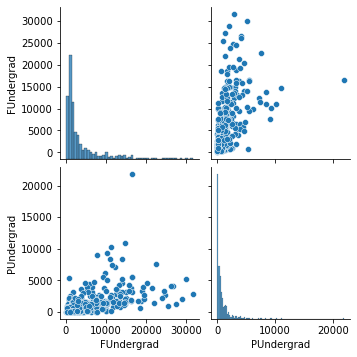
### Pairplot for columns 'Top10perc','Top25perc','Expend','Outstate','GradRate



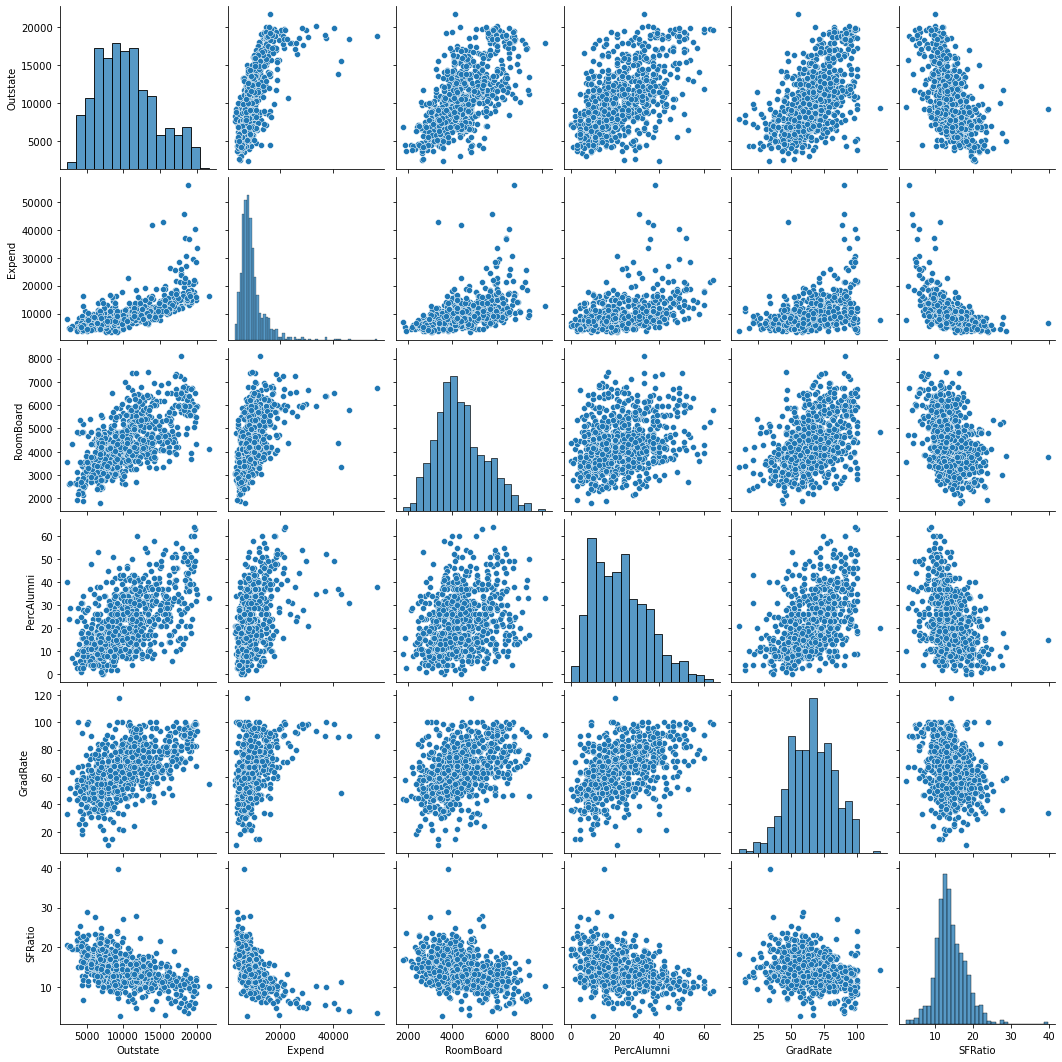
### Pairplot for columns 'PhD','Terminal'



### Pairplot for 'FUndergrad','PUndergrad'



### Pairplot for columns 'Outstate','Expend','RoomBoard','PercAlumni','GradRate', 'SFRatio'



* In these pairplots, we can see positive correlation between all these fields except SF ratio
* For SF ratio we can see the negative correlation with almost all these columns

### Outliers Treatment

As seen above, almost all the columns contains outliers and are causing skewness in the data so we need to treat them.

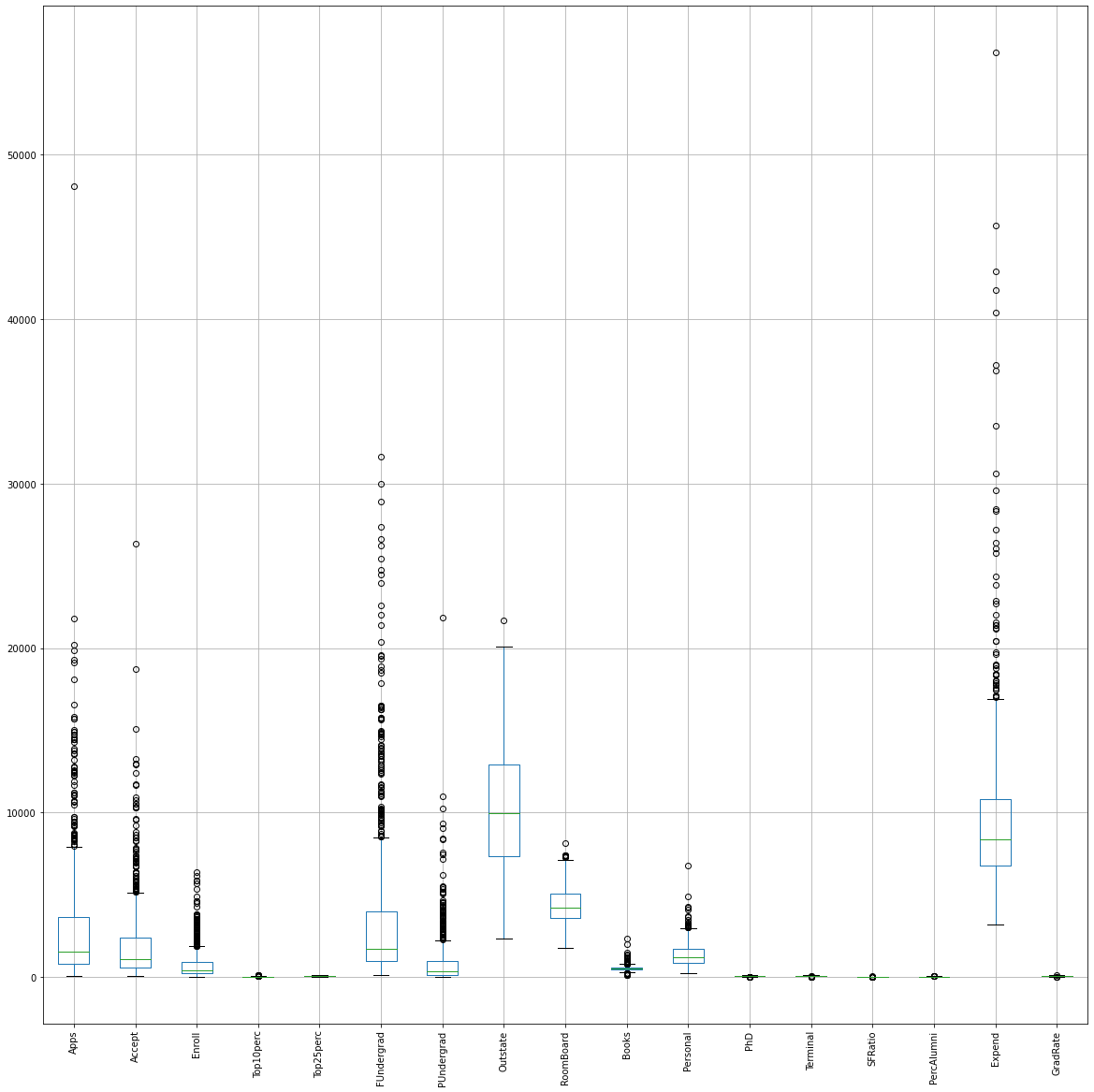
We have treated following columns for the outliers and we have capped their outer extreme value with the outer range and floor the lower extreme value with the lower range. Report of treatment done is as shown below :-

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column Name** | **Lower Range** | **Upper Range** | **No of outliers** | **Outlier %** |
| Apps | -3496 | 7896 | 70 | 9.01 |
| Accept | -2126 | 5154 | 73 | 9.4 |
| Enroll | -748 | 1892 | 79 | 10.17 |
| Top10perc | -15 | 65 | 39 | 5.02 |
| FUndergrad | -3527.5 | 8524.5 | 97 | 12.48 |
| PUndergrad | -1213 | 2275 | 67 | 8.62 |
| Outstate | -1087.5 | 21332.5 | 1 | 0.13 |
| RoomBoard | 1417.5 | 7229.5 | 7 | 0.9 |
| Books | 275 | 795 | 46 | 5.92 |
| Personal | -425 | 2975 | 20 | 2.57 |
| PhD | 27.5 | 119.5 | 8 | 1.03 |
| Terminal | 39.5 | 123.5 | 8 | 1.03 |
| SFRatio | 4 | 24 | 12 | 1.54 |
| PercAlumni | -14 | 58 | 5 | 0.64 |
| Expend | 632.5 | 16948.5 | 48 | 6.18 |
| GradRate | 15.5 | 115.5 | 4 | 0.51 |

As seen in the report, outlier % is high (around 10%) for Apps , Accept, Enroll , Fundergrad, Pundergrad

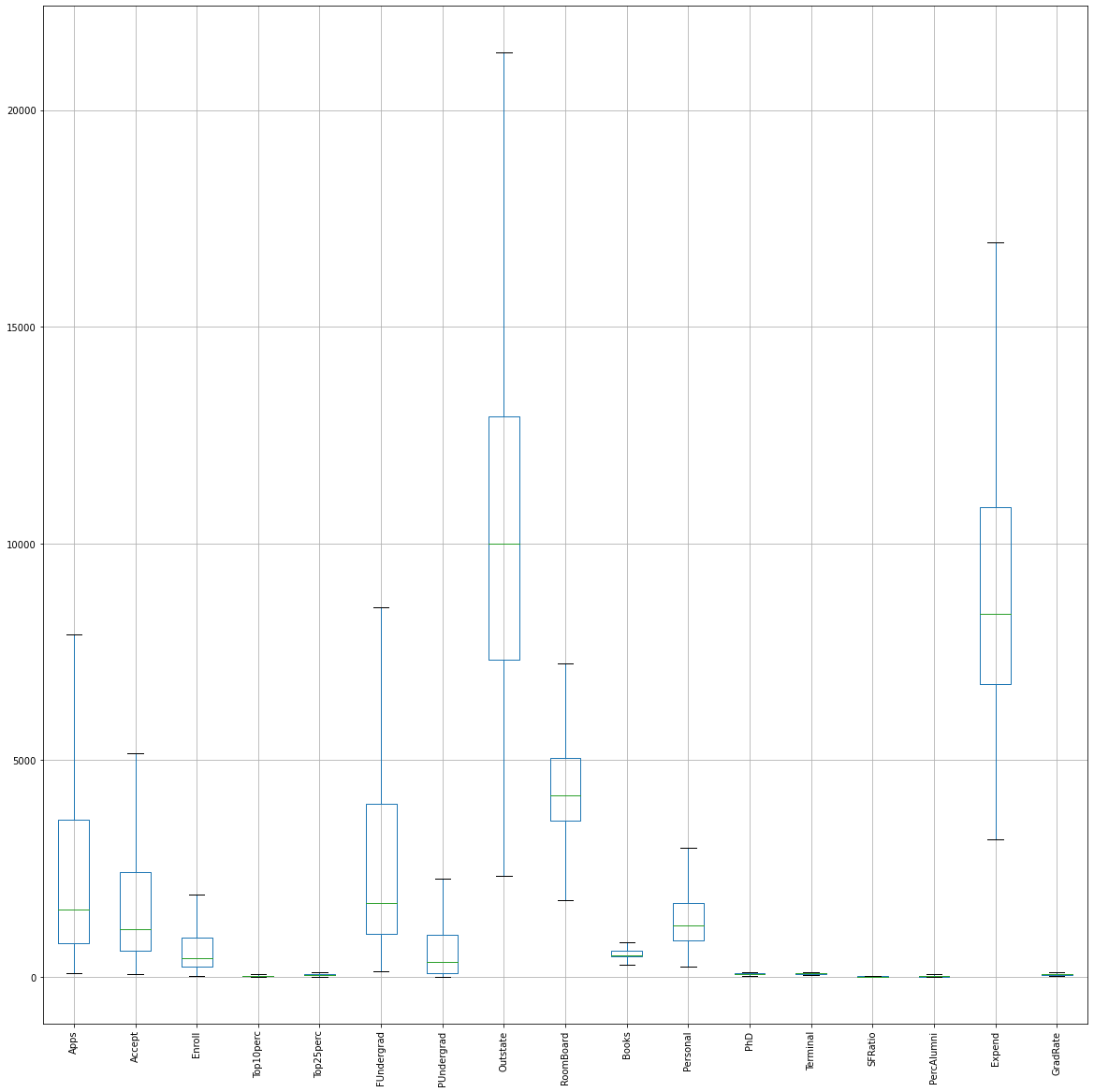
The treatment effect can be seen in the box plots before and after treatment

### Box plot before outlier treatment



Almost all the columns contains outliers and these are visible in boxplots

### Box plot after outlier treatment



### Pandas profile report

Pandas profile report has also been created and share in this report for reference. It contains all the column details



## 2.2) Scale the variables and write the inference for using the type of scaling function for this case study.

Now Name column has the unique name of the colleges and it has nothing to do with the further analysis. So it has been dropped before PCA analysis/ scaling

### Variance/Standard Deviation

In the dataset , standard deviation across difference columns is as shown below :-

Apps 2422.20

Accept 1523.29

Enroll 570.13

Top10perc 15.58

Top25perc 19.80

FUndergrad 2700.23

PUndergrad 716.27

Outstate 4021.71

RoomBoard 1090.67

Books 115.23

Personal 609.51

PhD 15.95

Terminal 14.47

SFRatio 3.78

PercAlumni 12.33

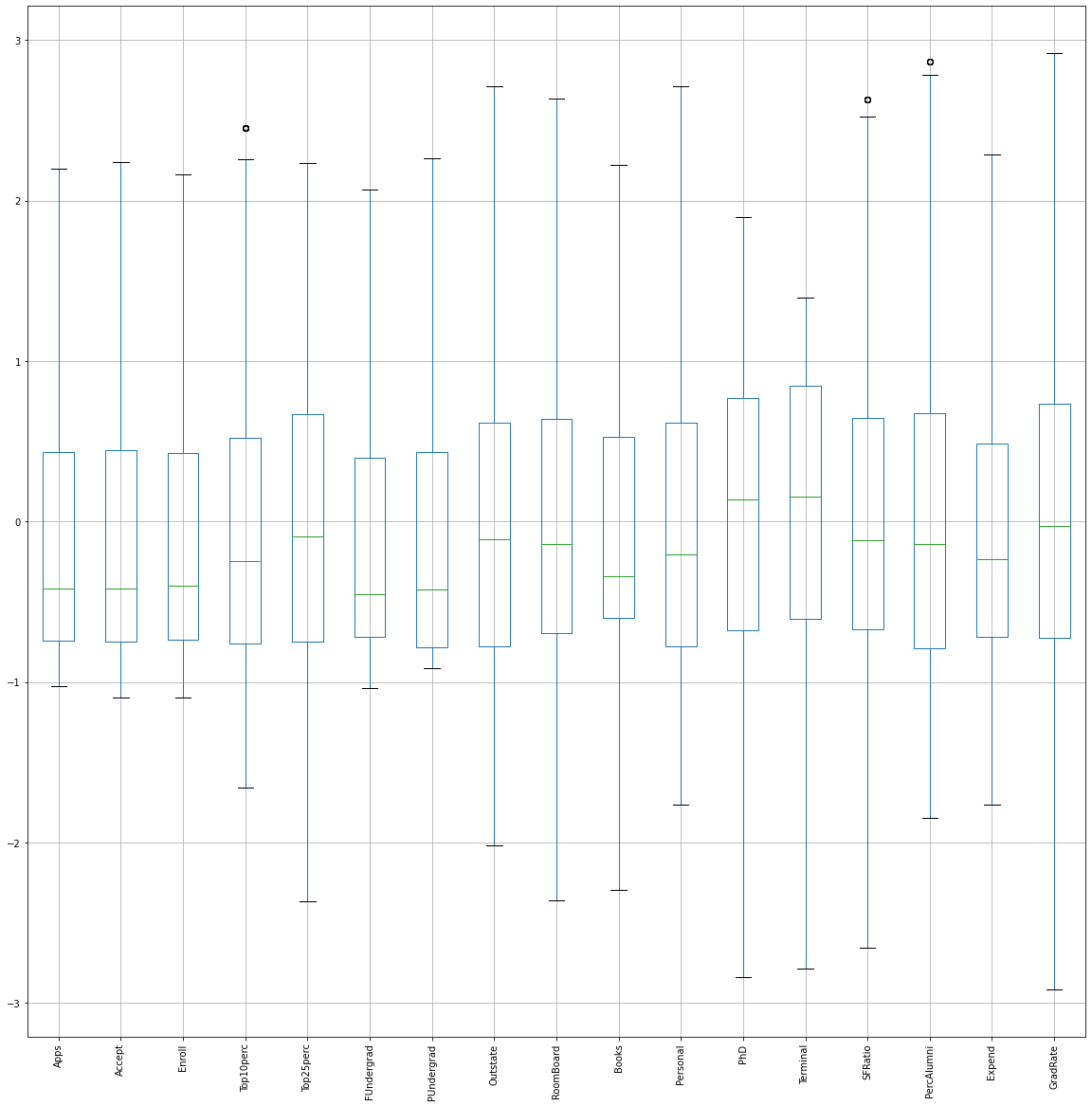
Expend 3396.50

GradRate 17.14

As we need to compare different columns and all the columns are on different scales. One is in percentage, one contains application count, application Enrolled count, one is the Student/Faculty ratio etc. Variance square (standard deviation) is very high for Expend, Outstate variable and for Phd and Terminal its very low. So in order to compare the data of these columns we need to bring them to one common scale as we can only compare Apples to Apples. Thus scaling is required here. We need to calculate (Z score) for each attribute to scale this data. It’s a standard scale and can be easily used here.

### Standardizing

After scaling, box plot looks like:-



**Now we can clearly see the difference in boxplots before scaling (outlier treatment section) and after scaling. As all columns are scaled up, all box plots are plotted and visible under same Y axis.**

## 2.3) Comment on the comparison between covariance and the correlation matrix.

#### Summary of original dataset after cleaning (outlier treatment)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| Apps | 777 | 2571.35 | 2422.20 | 81.0 | 776.0 | 1558.0 | 3624.0 | 7896 |
| Accept | 777 | 1746.28 | 1523.29 | 72.0 | 604.0 | 1110.0 | 2424.0 | 5154 |
| Enroll | 777 | 660.39 | 570.13 | 35.0 | 242.0 | 434.0 | 902.0 | 1892 |
| Top10perc | 777 | 26.84 | 15.58 | 1.0 | 15.0 | 23.0 | 35.0 | 65 |
| Top25perc | 777 | 55.80 | 19.80 | 9.0 | 41.0 | 54.0 | 69.0 | 100 |
| FUndergrad | 777 | 2935.65 | 2700.23 | 139.0 | 992.0 | 1707.0 | 4005.0 | 8524.5 |
| PUndergrad | 777 | 655.88 | 716.27 | 1.0 | 95.0 | 353.0 | 967.0 | 2275 |
| Outstate | 777 | 10440.2 | 4021.71 | 2340.0 | 7320.0 | 9990.0 | 12925.0 | 21332.5 |
| RoomBoard | 777 | 4355.44 | 1090.67 | 1780.0 | 3597.0 | 4200.0 | 5050.0 | 7229.5 |
| Books | 777 | 539.43 | 115.23 | 275.0 | 470.0 | 500.0 | 600.0 | 795 |
| Personal | 777 | 1323.79 | 609.51 | 250.0 | 850.0 | 1200.0 | 1700.0 | 2975 |
| PhD | 777 | 72.77 | 15.95 | 27.5 | 62.0 | 75.0 | 85.0 | 103 |
| Terminal | 777 | 79.78 | 14.47 | 39.5 | 71.0 | 82.0 | 92.0 | 100 |
| SFRatio | 777 | 14.05 | 3.78 | 4.0 | 11.5 | 13.6 | 16.5 | 24 |
| PercAlumni | 777 | 22.72 | 12.33 | 0.0 | 13.0 | 21.0 | 31.0 | 58 |
| Expend | 777 | 9182.52 | 3396.50 | 3186.0 | 6751.0 | 8377.0 | 10830.0 | 16948.5 |
| GradRate | 777 | 65.47 | 17.14 | 15.5 | 53.0 | 65.0 | 78.0 | 115.5 |

### Covariance matrix of original dataset

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Apps** | **Accept** | **Enroll** | **Top10perc** | **Top25perc** | **FUndergrad** | **PUndergrad** | **Outstate** | **RoomBoard** | **Books** | **Personal** | **PhD** | **Terminal** | **SFRatio** | **PercAlumni** | **Expend** | **GradRate** |
| Apps | 5867029.97 | 3524792.7 | 1238558.61 | 12128.73 | 17485.01 | 5631373.87 | 901869.92 | 636469.88 | 495273.82 | 65908.11 | 339481.46 | 17926.75 | 15231.3 | 1158.7 | -3020.03 | 1998621.84 | 6261.75 |
| Accept | 3524792.7 | 2320402.16 | 812257.02 | 5300.36 | 8256.49 | 3689705.47 | 624857.68 | -30645.21 | 198680.34 | 36633.57 | 238004.95 | 10384.88 | 8893.81 | 1086.63 | -3107.59 | 837169.14 | 2062.46 |
| Enroll | 1238558.61 | 812257.02 | 325044.61 | 1525.88 | 2601.88 | 1489137.68 | 262006.26 | -356900.27 | -14827.58 | 13274.24 | 117921.99 | 3470.22 | 2924.15 | 591.73 | -1565.09 | 104995.01 | -227.24 |
| Top10perc | 12128.73 | 5300.36 | 1525.88 | 242.82 | 282.03 | 4679.52 | -2009.14 | 35229.69 | 6073.55 | 275.53 | -1108.66 | 135.24 | 114.29 | -22.88 | 87.54 | 34774.48 | 131.87 |
| Top25perc | 17485.01 | 8256.49 | 2601.88 | 282.03 | 392.23 | 9689.9 | -1408.57 | 38993.75 | 7149.45 | 387.41 | -1047.9 | 174.23 | 151.24 | -22.28 | 101.75 | 38537.51 | 162.62 |
| FUndergrad | 5631373.87 | 3689705.47 | 1489137.68 | 4679.52 | 9689.9 | 7291258.52 | 1346390.41 | -2456058.98 | -160433.94 | 64681.06 | 592133.42 | 15575.16 | 13094.12 | 3315.86 | -9500.49 | 3399.28 | -3806.74 |
| PUndergrad | 901869.92 | 624857.68 | 262006.26 | -2009.14 | -1408.57 | 1346390.41 | 513048.46 | -1020372.83 | -52839.9 | 10113.02 | 150204.15 | 1458.78 | 1266.31 | 1004.54 | -3702.05 | -491257.73 | -3255.81 |
| Outstate | 636469.88 | -30645.21 | -356900.27 | 35229.69 | 38993.75 | -2456058.98 | -1020372.83 | 16174171.01 | 2875200.33 | 2368.07 | -798152.26 | 25106.68 | 24014.75 | -8730.88 | 28043.29 | 10590765.19 | 39466.61 |
| RoomBoard | 495273.82 | 198680.34 | -14827.58 | 6073.55 | 7149.45 | -160433.94 | -52839.9 | 2875200.33 | 1189552.34 | 13689.28 | -145952.4 | 5941.39 | 5986.88 | -1553.65 | 3661.77 | 2150881.26 | 7960.91 |
| Books | 65908.11 | 36633.57 | 13274.24 | 275.53 | 387.41 | 64681.06 | 10113.02 | 2368.07 | 13689.28 | 13277.89 | 16846.35 | 250.72 | 265.7 | -3.72 | -60.83 | 58700.04 | -15.9 |
| Personal | 339481.46 | 238004.95 | 117921.99 | -1108.66 | -1047.9 | 592133.42 | 150204.15 | -798152.26 | -145952.4 | 16846.35 | 371497.41 | -113.61 | -282.03 | 401.13 | -2296.95 | -338001.08 | -3039.4 |
| PhD | 17926.75 | 10384.88 | 3470.22 | 135.24 | 174.23 | 15575.16 | 1458.78 | 25106.68 | 5941.39 | 250.72 | -113.61 | 254.5 | 199.24 | -7.81 | 48.94 | 27662.86 | 84.78 |
| Terminal | 15231.3 | 8893.81 | 2924.15 | 114.29 | 151.24 | 13094.12 | 1266.31 | 24014.75 | 5986.88 | 265.7 | -282.03 | 199.24 | 209.47 | -8.27 | 47.46 | 25761.98 | 72.65 |
| SFRatio | 1158.7 | 1086.63 | 591.73 | -22.88 | -22.28 | 3315.86 | 1004.54 | -8730.88 | -1553.65 | -3.72 | 401.13 | -7.81 | -8.27 | 14.32 | -19.22 | -8410.74 | -20.01 |
| PercAlumni | -3020.03 | -3107.59 | -1565.09 | 87.54 | 101.75 | -9500.49 | -3702.05 | 28043.29 | 3661.77 | -60.83 | -2296.95 | 48.94 | 47.46 | -19.22 | 151.92 | 19379.52 | 103.83 |
| Expend | 1998621.84 | 837169.14 | 104995.01 | 34774.48 | 38537.51 | 3399.28 | -491257.73 | 10590765.19 | 2150881.26 | 58700.04 | -338001.08 | 27662.86 | 25761.98 | -8410.74 | 19379.52 | 11536186.08 | 24180.13 |
| GradRate | 6261.75 | 2062.46 | -227.24 | 131.87 | 162.62 | -3806.74 | -3255.81 | 39466.61 | 7960.91 | -15.9 | -3039.4 | 84.78 | 72.65 | -20.01 | 103.83 | 24180.13 | 293.87 |

### Correlation matrix of original data set

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Apps | Accept | Enroll | Top10perc | Top25perc | FUndergrad | PUndergrad | Outstate | RoomBoard | Books | Personal | PhD | Terminal | SFRatio | PercAlumni | Expend | GradRate |
| Apps | 1 | 0.96 | 0.9 | 0.32 | 0.36 | 0.86 | 0.52 | 0.07 | 0.19 | 0.24 | 0.23 | 0.46 | 0.43 | 0.13 | -0.1 | 0.24 | 0.15 |
| Accept | 0.96 | 1 | 0.94 | 0.22 | 0.27 | 0.9 | 0.57 | -0.01 | 0.12 | 0.21 | 0.26 | 0.43 | 0.4 | 0.19 | -0.17 | 0.16 | 0.08 |
| Enroll | 0.9 | 0.94 | 1 | 0.17 | 0.23 | 0.97 | 0.64 | -0.16 | -0.02 | 0.2 | 0.34 | 0.38 | 0.35 | 0.27 | -0.22 | 0.05 | -0.02 |
| Top10perc | 0.32 | 0.22 | 0.17 | 1 | 0.91 | 0.11 | -0.18 | 0.56 | 0.36 | 0.15 | -0.12 | 0.54 | 0.51 | -0.39 | 0.46 | 0.66 | 0.49 |
| Top25perc | 0.36 | 0.27 | 0.23 | 0.91 | 1 | 0.18 | -0.1 | 0.49 | 0.33 | 0.17 | -0.09 | 0.55 | 0.53 | -0.3 | 0.42 | 0.57 | 0.48 |
| FUndergrad | 0.86 | 0.9 | 0.97 | 0.11 | 0.18 | 1 | 0.7 | -0.23 | -0.05 | 0.21 | 0.36 | 0.36 | 0.34 | 0.32 | -0.29 | 0 | -0.08 |
| PUndergrad | 0.52 | 0.57 | 0.64 | -0.18 | -0.1 | 0.7 | 1 | -0.35 | -0.07 | 0.12 | 0.34 | 0.13 | 0.12 | 0.37 | -0.42 | -0.2 | -0.27 |
| Outstate | 0.07 | -0.01 | -0.16 | 0.56 | 0.49 | -0.23 | -0.35 | 1 | 0.66 | 0.01 | -0.33 | 0.39 | 0.41 | -0.57 | 0.57 | 0.78 | 0.57 |
| RoomBoard | 0.19 | 0.12 | -0.02 | 0.36 | 0.33 | -0.05 | -0.07 | 0.66 | 1 | 0.11 | -0.22 | 0.34 | 0.38 | -0.38 | 0.27 | 0.58 | 0.43 |
| Books | 0.24 | 0.21 | 0.2 | 0.15 | 0.17 | 0.21 | 0.12 | 0.01 | 0.11 | 1 | 0.24 | 0.14 | 0.16 | -0.01 | -0.04 | 0.15 | -0.01 |
| Personal | 0.23 | 0.26 | 0.34 | -0.12 | -0.09 | 0.36 | 0.34 | -0.33 | -0.22 | 0.24 | 1 | -0.01 | -0.03 | 0.17 | -0.31 | -0.16 | -0.29 |
| PhD | 0.46 | 0.43 | 0.38 | 0.54 | 0.55 | 0.36 | 0.13 | 0.39 | 0.34 | 0.14 | -0.01 | 1 | 0.86 | -0.13 | 0.25 | 0.51 | 0.31 |
| Terminal | 0.43 | 0.4 | 0.35 | 0.51 | 0.53 | 0.34 | 0.12 | 0.41 | 0.38 | 0.16 | -0.03 | 0.86 | 1 | -0.15 | 0.27 | 0.52 | 0.29 |
| SFRatio | 0.13 | 0.19 | 0.27 | -0.39 | -0.3 | 0.32 | 0.37 | -0.57 | -0.38 | -0.01 | 0.17 | -0.13 | -0.15 | 1 | -0.41 | -0.65 | -0.31 |
| PercAlumni | -0.1 | -0.17 | -0.22 | 0.46 | 0.42 | -0.29 | -0.42 | 0.57 | 0.27 | -0.04 | -0.31 | 0.25 | 0.27 | -0.41 | 1 | 0.46 | 0.49 |
| Expend | 0.24 | 0.16 | 0.05 | 0.66 | 0.57 | 0 | -0.2 | 0.78 | 0.58 | 0.15 | -0.16 | 0.51 | 0.52 | -0.65 | 0.46 | 1 | 0.42 |
| GradRate | 0.15 | 0.08 | -0.02 | 0.49 | 0.48 | -0.08 | -0.27 | 0.57 | 0.43 | -0.01 | -0.29 | 0.31 | 0.29 | -0.31 | 0.49 | 0.42 | 1 |

We can see that correlation matrix is not equal to covariance matrix of original dataset Covariance matrix values are high whereas correlation matrix values are in between -1 to 1

### Summary of scaled dataset

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | mean | std | min | 25% | 50% | 75% | max |
| Apps | 777 | 0 | 1 | -1.03 | -0.74 | -0.42 | 0.43 | 2.2 |
| Accept | 777 | 0 | 1 | -1.1 | -0.75 | -0.42 | 0.45 | 2.24 |
| Enroll | 777 | 0 | 1 | -1.1 | -0.73 | -0.4 | 0.42 | 2.16 |
| Top10perc | 777 | 0 | 1 | -1.66 | -0.76 | -0.25 | 0.52 | 2.45 |
| Top25perc | 777 | 0 | 1 | -2.36 | -0.75 | -0.09 | 0.67 | 2.23 |
| FUndergrad | 777 | 0 | 1 | -1.04 | -0.72 | -0.46 | 0.4 | 2.07 |
| PUndergrad | 777 | 0 | 1 | -0.91 | -0.78 | -0.42 | 0.43 | 2.26 |
| Outstate | 777 | 0 | 1 | -2.02 | -0.78 | -0.11 | 0.62 | 2.71 |
| RoomBoard | 777 | 0 | 1 | -2.36 | -0.7 | -0.14 | 0.64 | 2.64 |
| Books | 777 | 0 | 1 | -2.3 | -0.6 | -0.34 | 0.53 | 2.22 |
| Personal | 777 | 0 | 1 | -1.76 | -0.78 | -0.2 | 0.62 | 2.71 |
| PhD | 777 | 0 | 1 | -2.84 | -0.68 | 0.14 | 0.77 | 1.9 |
| Terminal | 777 | 0 | 1 | -2.79 | -0.61 | 0.15 | 0.84 | 1.4 |
| SFRatio | 777 | 0 | 1 | -2.66 | -0.67 | -0.12 | 0.65 | 2.63 |
| PercAlumni | 777 | 0 | 1 | -1.84 | -0.79 | -0.14 | 0.67 | 2.86 |
| Expend | 777 | 0 | 1 | -1.77 | -0.72 | -0.24 | 0.49 | 2.29 |
| GradRate | 777 | 0 | 1 | -2.92 | -0.73 | -0.03 | 0.73 | 2.92 |

We can see that after scaling standard deviation of each col is 1 and mean is 0

### Covariance matrix of scaled dataset

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Apps | Accept | Enroll | Top10perc | Top25perc | FUndergrad | PUndergrad | Outstate | RoomBoard | Books | Personal | PhD | Terminal | SFRatio | PercAlumni | Expend | GradRate |
| Apps | 1 | 0.96 | 0.9 | 0.32 | 0.36 | 0.86 | 0.52 | 0.07 | 0.19 | 0.24 | 0.23 | 0.46 | 0.44 | 0.13 | -0.1 | 0.24 | 0.15 |
| Accept | 0.96 | 1 | 0.94 | 0.22 | 0.27 | 0.9 | 0.57 | -0.01 | 0.12 | 0.21 | 0.26 | 0.43 | 0.4 | 0.19 | -0.17 | 0.16 | 0.08 |
| Enroll | 0.9 | 0.94 | 1 | 0.17 | 0.23 | 0.97 | 0.64 | -0.16 | -0.02 | 0.2 | 0.34 | 0.38 | 0.35 | 0.27 | -0.22 | 0.05 | -0.02 |
| Top10perc | 0.32 | 0.22 | 0.17 | 1 | 0.92 | 0.11 | -0.18 | 0.56 | 0.36 | 0.15 | -0.12 | 0.54 | 0.51 | -0.39 | 0.46 | 0.66 | 0.49 |
| Top25perc | 0.36 | 0.27 | 0.23 | 0.92 | 1 | 0.18 | -0.1 | 0.49 | 0.33 | 0.17 | -0.09 | 0.55 | 0.53 | -0.3 | 0.42 | 0.57 | 0.48 |
| FUndergrad | 0.86 | 0.9 | 0.97 | 0.11 | 0.18 | 1 | 0.7 | -0.23 | -0.05 | 0.21 | 0.36 | 0.36 | 0.34 | 0.32 | -0.29 | 0 | -0.08 |
| PUndergrad | 0.52 | 0.57 | 0.64 | -0.18 | -0.1 | 0.7 | 1 | -0.35 | -0.07 | 0.12 | 0.34 | 0.13 | 0.12 | 0.37 | -0.42 | -0.2 | -0.27 |
| Outstate | 0.07 | -0.01 | -0.16 | 0.56 | 0.49 | -0.23 | -0.35 | 1 | 0.66 | 0.01 | -0.33 | 0.39 | 0.41 | -0.57 | 0.57 | 0.78 | 0.57 |
| RoomBoard | 0.19 | 0.12 | -0.02 | 0.36 | 0.33 | -0.05 | -0.07 | 0.66 | 1 | 0.11 | -0.22 | 0.34 | 0.38 | -0.38 | 0.27 | 0.58 | 0.43 |
| Books | 0.24 | 0.21 | 0.2 | 0.15 | 0.17 | 0.21 | 0.12 | 0.01 | 0.11 | 1 | 0.24 | 0.14 | 0.16 | -0.01 | -0.04 | 0.15 | -0.01 |
| Personal | 0.23 | 0.26 | 0.34 | -0.12 | -0.09 | 0.36 | 0.34 | -0.33 | -0.22 | 0.24 | 1 | -0.01 | -0.03 | 0.17 | -0.31 | -0.16 | -0.29 |
| PhD | 0.46 | 0.43 | 0.38 | 0.54 | 0.55 | 0.36 | 0.13 | 0.39 | 0.34 | 0.14 | -0.01 | 1 | 0.86 | -0.13 | 0.25 | 0.51 | 0.31 |
| Terminal | 0.44 | 0.4 | 0.35 | 0.51 | 0.53 | 0.34 | 0.12 | 0.41 | 0.38 | 0.16 | -0.03 | 0.86 | 1 | -0.15 | 0.27 | 0.52 | 0.29 |
| SFRatio | 0.13 | 0.19 | 0.27 | -0.39 | -0.3 | 0.32 | 0.37 | -0.57 | -0.38 | -0.01 | 0.17 | -0.13 | -0.15 | 1 | -0.41 | -0.66 | -0.31 |
| PercAlumni | -0.1 | -0.17 | -0.22 | 0.46 | 0.42 | -0.29 | -0.42 | 0.57 | 0.27 | -0.04 | -0.31 | 0.25 | 0.27 | -0.41 | 1 | 0.46 | 0.49 |
| Expend | 0.24 | 0.16 | 0.05 | 0.66 | 0.57 | 0 | -0.2 | 0.78 | 0.58 | 0.15 | -0.16 | 0.51 | 0.52 | -0.66 | 0.46 | 1 | 0.42 |
| GradRate | 0.15 | 0.08 | -0.02 | 0.49 | 0.48 | -0.08 | -0.27 | 0.57 | 0.43 | -0.01 | -0.29 | 0.31 | 0.29 | -0.31 | 0.49 | 0.42 | 1 |

### Correlation matrix of scaled data set

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Apps | Accept | Enroll | Top10perc | Top25perc | FUndergrad | PUndergrad | Outstate | RoomBoard | Books | Personal | PhD | Terminal | SFRatio | PercAlumni | Expend | GradRate |
| Apps | 1 | 0.96 | 0.9 | 0.32 | 0.36 | 0.86 | 0.52 | 0.07 | 0.19 | 0.24 | 0.23 | 0.46 | 0.43 | 0.13 | -0.1 | 0.24 | 0.15 |
| Accept | 0.96 | 1 | 0.94 | 0.22 | 0.27 | 0.9 | 0.57 | -0.01 | 0.12 | 0.21 | 0.26 | 0.43 | 0.4 | 0.19 | -0.17 | 0.16 | 0.08 |
| Enroll | 0.9 | 0.94 | 1 | 0.17 | 0.23 | 0.97 | 0.64 | -0.16 | -0.02 | 0.2 | 0.34 | 0.38 | 0.35 | 0.27 | -0.22 | 0.05 | -0.02 |
| Top10perc | 0.32 | 0.22 | 0.17 | 1 | 0.91 | 0.11 | -0.18 | 0.56 | 0.36 | 0.15 | -0.12 | 0.54 | 0.51 | -0.39 | 0.46 | 0.66 | 0.49 |
| Top25perc | 0.36 | 0.27 | 0.23 | 0.91 | 1 | 0.18 | -0.1 | 0.49 | 0.33 | 0.17 | -0.09 | 0.55 | 0.53 | -0.3 | 0.42 | 0.57 | 0.48 |
| FUndergrad | 0.86 | 0.9 | 0.97 | 0.11 | 0.18 | 1 | 0.7 | -0.23 | -0.05 | 0.21 | 0.36 | 0.36 | 0.34 | 0.32 | -0.29 | 0 | -0.08 |
| PUndergrad | 0.52 | 0.57 | 0.64 | -0.18 | -0.1 | 0.7 | 1 | -0.35 | -0.07 | 0.12 | 0.34 | 0.13 | 0.12 | 0.37 | -0.42 | -0.2 | -0.27 |
| Outstate | 0.07 | -0.01 | -0.16 | 0.56 | 0.49 | -0.23 | -0.35 | 1 | 0.66 | 0.01 | -0.33 | 0.39 | 0.41 | -0.57 | 0.57 | 0.78 | 0.57 |
| RoomBoard | 0.19 | 0.12 | -0.02 | 0.36 | 0.33 | -0.05 | -0.07 | 0.66 | 1 | 0.11 | -0.22 | 0.34 | 0.38 | -0.38 | 0.27 | 0.58 | 0.43 |
| Books | 0.24 | 0.21 | 0.2 | 0.15 | 0.17 | 0.21 | 0.12 | 0.01 | 0.11 | 1 | 0.24 | 0.14 | 0.16 | -0.01 | -0.04 | 0.15 | -0.01 |
| Personal | 0.23 | 0.26 | 0.34 | -0.12 | -0.09 | 0.36 | 0.34 | -0.33 | -0.22 | 0.24 | 1 | -0.01 | -0.03 | 0.17 | -0.31 | -0.16 | -0.29 |
| PhD | 0.46 | 0.43 | 0.38 | 0.54 | 0.55 | 0.36 | 0.13 | 0.39 | 0.34 | 0.14 | -0.01 | 1 | 0.86 | -0.13 | 0.25 | 0.51 | 0.31 |
| Terminal | 0.43 | 0.4 | 0.35 | 0.51 | 0.53 | 0.34 | 0.12 | 0.41 | 0.38 | 0.16 | -0.03 | 0.86 | 1 | -0.15 | 0.27 | 0.52 | 0.29 |
| SFRatio | 0.13 | 0.19 | 0.27 | -0.39 | -0.3 | 0.32 | 0.37 | -0.57 | -0.38 | -0.01 | 0.17 | -0.13 | -0.15 | 1 | -0.41 | -0.65 | -0.31 |
| PercAlumni | -0.1 | -0.17 | -0.22 | 0.46 | 0.42 | -0.29 | -0.42 | 0.57 | 0.27 | -0.04 | -0.31 | 0.25 | 0.27 | -0.41 | 1 | 0.46 | 0.49 |
| Expend | 0.24 | 0.16 | 0.05 | 0.66 | 0.57 | 0 | -0.2 | 0.78 | 0.58 | 0.15 | -0.16 | 0.51 | 0.52 | -0.65 | 0.46 | 1 | 0.42 |
| GradRate | 0.15 | 0.08 | -0.02 | 0.49 | 0.48 | -0.08 | -0.27 | 0.57 | 0.43 | -0.01 | -0.29 | 0.31 | 0.29 | -0.31 | 0.49 | 0.42 | 1 |

### Conclusion

“Covariance” indicates the direction of the linear relationship between variables. “Correlation” on the other hand measures both the strength and direction of the linear relationship between the two variables. Correlation is a function of the covariance. Correlation is obtained using coefficient of two variables by dividing the covariance of these variables by the product of the standard deviations of the same values.

Correlation(x,y) = Covariance (x,y) / (std (x) \* std(y))

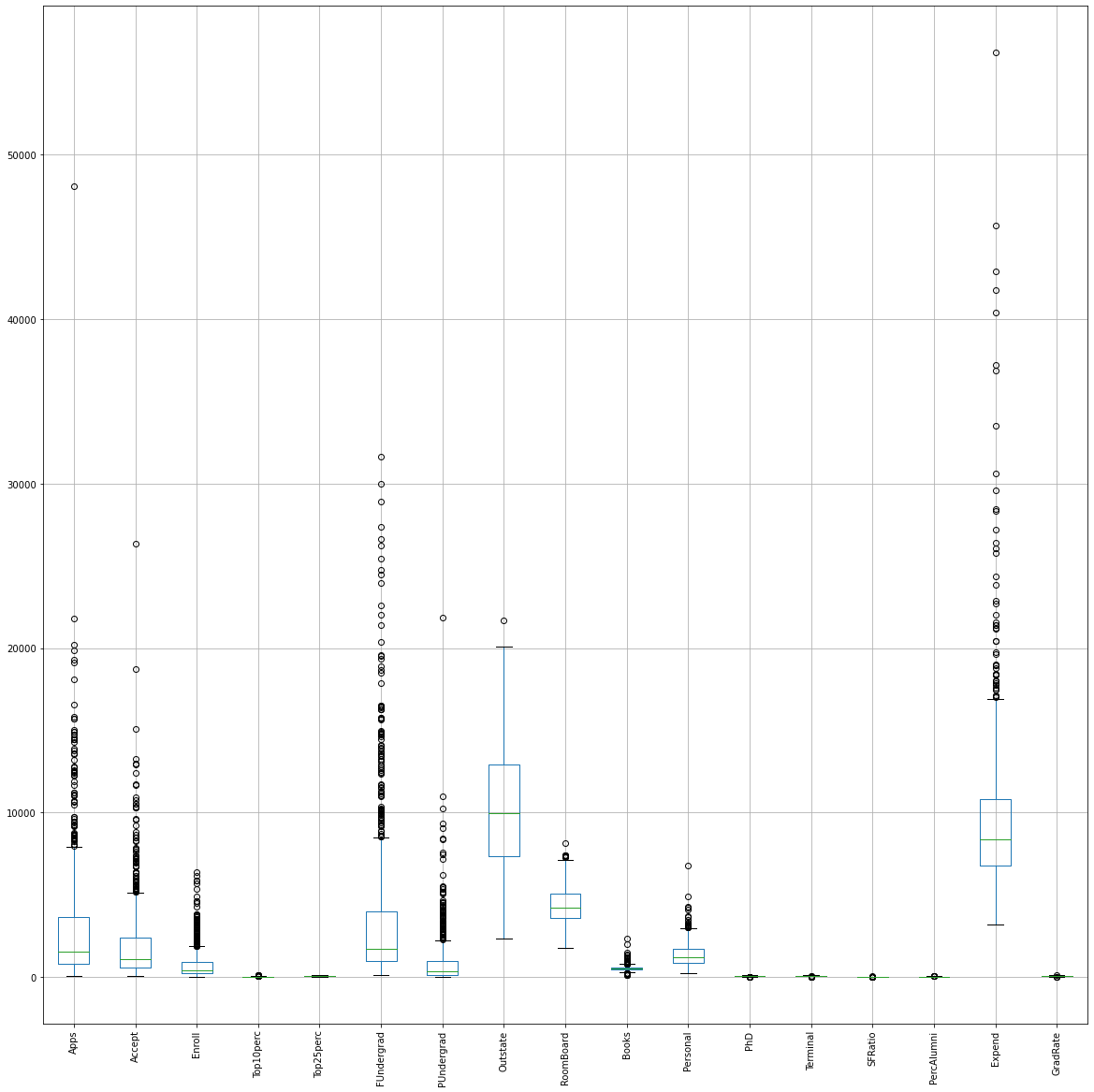
Now since standarad deviation of all columns is 1 for scaled dataset,

correlation(x,y) = Covariance (x,y)

Henec, we can say that Covariance matrix and correlation matrix are same for the scaled data set.

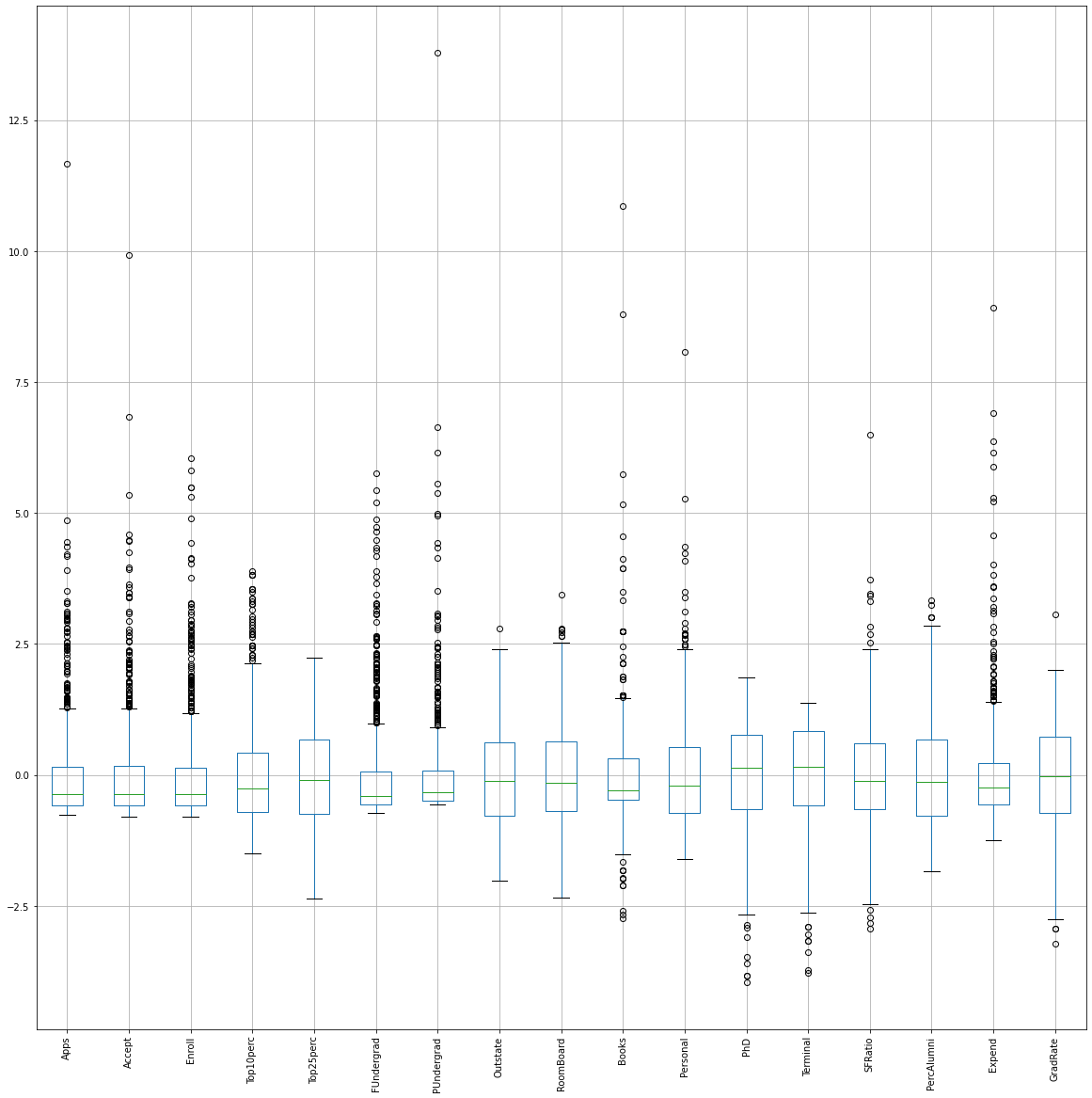
## Q2.4 Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.

### Plotting box plot on initial data set without outlier treatment



We can see many outliers in the dataset

### Plotting box plot after scaling and without any outlier treatment



We can see even after scaling, outliers are still there in the box plots.

### Conclusion

On comparing the 2 box plots of before and after scaling we have found that scaling has no effect on outliers i.e. after doing scaling outliers have not been removed from the dataset

## 2.5) Build the covariance matrix and calculate the eigenvalues and the eigenvector.

### Create covariance matrix

From Jupyter notebook, Covariance matrix is as shown below :-

Covariance Matrix

%s [[ 1.00128866e+00 9.56537704e-01 8.98039052e-01 3.21756324e-01

3.64960691e-01 8.62111140e-01 5.20492952e-01 6.54209711e-02

1.87717056e-01 2.36441941e-01 2.30243993e-01 4.64521757e-01

4.35037784e-01 1.26573895e-01 -1.01288006e-01 2.43248206e-01

1.50997775e-01]

[ 9.56537704e-01 1.00128866e+00 9.36482483e-01 2.23586208e-01

2.74033187e-01 8.98189799e-01 5.73428908e-01 -5.00874847e-03

1.19740419e-01 2.08974091e-01 2.56676290e-01 4.27891234e-01

4.03929238e-01 1.88748711e-01 -1.65728801e-01 1.62016688e-01

7.90839722e-02]

[ 8.98039052e-01 9.36482483e-01 1.00128866e+00 1.71977357e-01

2.30730728e-01 9.68548601e-01 6.42421828e-01 -1.55856056e-01

-2.38762560e-02 2.02317274e-01 3.39785395e-01 3.82031198e-01

3.54835877e-01 2.74622251e-01 -2.23009677e-01 5.42906862e-02

-2.32810071e-02]

[ 3.21756324e-01 2.23586208e-01 1.71977357e-01 1.00128866e+00

9.15052977e-01 1.11358019e-01 -1.80240778e-01 5.62884044e-01

3.57826139e-01 1.53650150e-01 -1.16880152e-01 5.44748764e-01

5.07401238e-01 -3.88425719e-01 4.56384036e-01 6.57885921e-01

4.94306540e-01]

[ 3.64960691e-01 2.74033187e-01 2.30730728e-01 9.15052977e-01

1.00128866e+00 1.81429267e-01 -9.94231153e-02 4.90200034e-01

3.31413314e-01 1.69979808e-01 -8.69219644e-02 5.52172085e-01

5.28333659e-01 -2.97616423e-01 4.17369123e-01 5.73643193e-01

4.79601950e-01]

[ 8.62111140e-01 8.98189799e-01 9.68548601e-01 1.11358019e-01

1.81429267e-01 1.00128866e+00 6.97027420e-01 -2.26457040e-01

-5.45459528e-02 2.08147257e-01 3.60246460e-01 3.62030390e-01

3.35485771e-01 3.24921933e-01 -2.85825062e-01 3.71119607e-04

-8.23447851e-02]

[ 5.20492952e-01 5.73428908e-01 6.42421828e-01 -1.80240778e-01

-9.94231153e-02 6.97027420e-01 1.00128866e+00 -3.54672874e-01

-6.77252009e-02 1.22686416e-01 3.44495974e-01 1.27827147e-01

1.22309141e-01 3.71084841e-01 -4.19874031e-01 -2.02189396e-01

-2.65499420e-01]

[ 6.54209711e-02 -5.00874847e-03 -1.55856056e-01 5.62884044e-01

4.90200034e-01 -2.26457040e-01 -3.54672874e-01 1.00128866e+00

6.56333564e-01 5.11656377e-03 -3.26028927e-01 3.91824814e-01

4.13110264e-01 -5.74421963e-01 5.66465309e-01 7.76326650e-01

5.73195743e-01]

[ 1.87717056e-01 1.19740419e-01 -2.38762560e-02 3.57826139e-01

3.31413314e-01 -5.45459528e-02 -6.77252009e-02 6.56333564e-01

1.00128866e+00 1.09064551e-01 -2.19837042e-01 3.41908577e-01

3.79759015e-01 -3.76915472e-01 2.72743761e-01 5.81370284e-01

4.26338910e-01]

[ 2.36441941e-01 2.08974091e-01 2.02317274e-01 1.53650150e-01

1.69979808e-01 2.08147257e-01 1.22686416e-01 5.11656377e-03

1.09064551e-01 1.00128866e+00 2.40172145e-01 1.36566243e-01

1.59523091e-01 -8.54689129e-03 -4.28870629e-02 1.50176551e-01

-8.06107505e-03]

[ 2.30243993e-01 2.56676290e-01 3.39785395e-01 -1.16880152e-01

-8.69219644e-02 3.60246460e-01 3.44495974e-01 -3.26028927e-01

-2.19837042e-01 2.40172145e-01 1.00128866e+00 -1.16986124e-02

-3.20117803e-02 1.74136664e-01 -3.06146886e-01 -1.63481407e-01

-2.91268705e-01]

[ 4.64521757e-01 4.27891234e-01 3.82031198e-01 5.44748764e-01

5.52172085e-01 3.62030390e-01 1.27827147e-01 3.91824814e-01

3.41908577e-01 1.36566243e-01 -1.16986124e-02 1.00128866e+00

8.64040263e-01 -1.29556494e-01 2.49197779e-01 5.11186852e-01

3.10418895e-01]

[ 4.35037784e-01 4.03929238e-01 3.54835877e-01 5.07401238e-01

5.28333659e-01 3.35485771e-01 1.22309141e-01 4.13110264e-01

3.79759015e-01 1.59523091e-01 -3.20117803e-02 8.64040263e-01

1.00128866e+00 -1.51187934e-01 2.66375402e-01 5.24743500e-01

2.93180212e-01]

[ 1.26573895e-01 1.88748711e-01 2.74622251e-01 -3.88425719e-01

-2.97616423e-01 3.24921933e-01 3.71084841e-01 -5.74421963e-01

-3.76915472e-01 -8.54689129e-03 1.74136664e-01 -1.29556494e-01

-1.51187934e-01 1.00128866e+00 -4.12632056e-01 -6.55219504e-01

-3.08922187e-01]

[-1.01288006e-01 -1.65728801e-01 -2.23009677e-01 4.56384036e-01

4.17369123e-01 -2.85825062e-01 -4.19874031e-01 5.66465309e-01

2.72743761e-01 -4.28870629e-02 -3.06146886e-01 2.49197779e-01

2.66375402e-01 -4.12632056e-01 1.00128866e+00 4.63518674e-01

4.92040760e-01]

[ 2.43248206e-01 1.62016688e-01 5.42906862e-02 6.57885921e-01

5.73643193e-01 3.71119607e-04 -2.02189396e-01 7.76326650e-01

5.81370284e-01 1.50176551e-01 -1.63481407e-01 5.11186852e-01

5.24743500e-01 -6.55219504e-01 4.63518674e-01 1.00128866e+00

4.15826026e-01]

[ 1.50997775e-01 7.90839722e-02 -2.32810071e-02 4.94306540e-01

4.79601950e-01 -8.23447851e-02 -2.65499420e-01 5.73195743e-01

4.26338910e-01 -8.06107505e-03 -2.91268705e-01 3.10418895e-01

2.93180212e-01 -3.08922187e-01 4.92040760e-01 4.15826026e-01

1.00128866e+00]]

### Get eigen values and eigen vector

Eigen values and Eigen vectors output is as shown below :-

Eigen Values

%s [5.6625219 4.89470815 1.12636744 1.00397659 0.87218426 0.7657541

0.58491404 0.5445048 0.42352336 0.38101777 0.24701456 0.02239369

0.03789395 0.14726392 0.13434483 0.09883384 0.07469003]

Eigen Vectors

%s [[-2.62171542e-01 3.14136258e-01 8.10177245e-02 -9.87761685e-02

-2.19898081e-01 2.18800617e-03 -2.83715076e-02 -8.99498102e-02

1.30566998e-01 -1.56464458e-01 -8.62132843e-02 1.82169814e-01

-5.99137640e-01 8.99775288e-02 8.88697944e-02 5.49428396e-01

5.41453698e-03]

[-2.30562461e-01 3.44623583e-01 1.07658626e-01 -1.18140437e-01

-1.89634940e-01 -1.65212882e-02 -1.29584896e-02 -1.37606312e-01

1.42275847e-01 -1.49209799e-01 -4.25899061e-02 -3.91041719e-01

6.61496927e-01 1.58861886e-01 4.37945938e-02 2.91572312e-01

1.44582845e-02]

[-1.89276397e-01 3.82813322e-01 8.55296892e-02 -9.30717094e-03

-1.62314818e-01 -6.80794143e-02 -1.52403625e-02 -1.44216938e-01

5.08712481e-02 -6.48997860e-02 -4.38408622e-02 7.16684935e-01

2.33235272e-01 -3.53988202e-02 -6.19241658e-02 -4.17001280e-01

-4.97908902e-02]

[-3.38874521e-01 -9.93191661e-02 -7.88293849e-02 3.69115031e-01

-1.57211016e-01 -8.88656824e-02 -2.57455284e-01 2.89538833e-01

-1.22467790e-01 -3.58776186e-02 1.77837341e-03 -5.62053913e-02

2.21448729e-02 -3.92277722e-02 6.99599977e-02 8.79767299e-03

-7.23645373e-01]

[-3.34690532e-01 -5.95055011e-02 -5.07938247e-02 4.16824361e-01

-1.44449474e-01 -2.76268979e-02 -2.39038849e-01 3.45643551e-01

-1.93936316e-01 6.41786425e-03 -1.02127328e-01 1.96735274e-02

3.22646978e-02 1.45621999e-01 -9.70282598e-02 -1.07779150e-02

6.55464648e-01]

[-1.63293010e-01 3.98636372e-01 7.37077827e-02 -1.39504424e-02

-1.02728468e-01 -5.16468727e-02 -3.11751439e-02 -1.08748900e-01

1.45452749e-03 -1.63981359e-04 -3.49993487e-02 -5.42774834e-01

-3.67681187e-01 -1.33555923e-01 -8.71753137e-02 -5.70683843e-01

2.53059904e-02]

[-2.24797091e-02 3.57550046e-01 4.03568700e-02 -2.25351078e-01

9.56790178e-02 -2.45375721e-02 -1.00138971e-02 1.23841696e-01

-6.34774326e-01 5.46346279e-01 2.52107094e-01 2.95029745e-02

2.62494456e-02 5.02487566e-02 4.45537493e-02 1.46321060e-01

-3.97146972e-02]

[-2.83547285e-01 -2.51863617e-01 1.49394795e-02 -2.62975384e-01

-3.72750885e-02 -2.03860462e-02 9.45370782e-02 1.12721477e-02

-8.36648339e-03 -2.31799759e-01 5.93433149e-01 1.03393587e-03

-8.14247697e-02 5.60392799e-01 6.72405494e-02 -2.11561014e-01

-1.59275617e-03]

[-2.44186588e-01 -1.31909124e-01 -2.11379165e-02 -5.80894132e-01

6.91080879e-02 2.37267409e-01 9.45210745e-02 3.89639465e-01

-2.20526518e-01 -2.55107620e-01 -4.75297296e-01 9.85725168e-03

2.67779296e-02 -1.07365653e-01 1.77715010e-02 -1.00935084e-01

-2.82578388e-02]

[-9.67082754e-02 9.39739472e-02 -6.97121128e-01 3.61562884e-02

-3.54056654e-02 6.38604997e-01 -1.11193334e-01 -2.39817267e-01

2.10246624e-02 9.11624912e-02 4.35697999e-02 4.36086500e-03

1.04624246e-02 5.16224550e-02 3.54343707e-02 -2.86384228e-02

-8.06259380e-03]

[ 3.52299594e-02 2.32439594e-01 -5.30972806e-01 1.14982973e-01

4.75358244e-04 -3.81495854e-01 6.39418106e-01 2.77206569e-01

1.73715184e-02 -1.27647512e-01 1.51627393e-02 -1.08725257e-02

4.54572099e-03 9.39409228e-03 -1.18604404e-02 3.38197909e-02

1.42590097e-03]

[-3.26410696e-01 5.51390195e-02 8.11134044e-02 1.47260891e-01

5.50786546e-01 3.34444832e-03 8.92320786e-02 -3.42628480e-02

1.66510079e-01 1.00975002e-01 -3.91865961e-02 1.33146759e-02

1.25137966e-02 -7.16590441e-02 7.02656469e-01 -6.38096394e-02

8.31471932e-02]

[-3.23115980e-01 4.30332048e-02 5.89785929e-02 8.90079921e-02

5.90407136e-01 3.54121294e-02 9.16985445e-02 -9.03076644e-02

1.12609034e-01 8.60363025e-02 -8.48575651e-02 7.38135022e-03

-1.79275275e-02 1.63820871e-01 -6.62488717e-01 9.85019644e-02

-1.13374007e-01]

[ 1.63151642e-01 2.59804556e-01 2.74150657e-01 2.59486122e-01

1.42842546e-01 4.68752604e-01 1.52864837e-01 2.42807562e-01

-1.53685343e-01 -4.70527925e-01 3.63042716e-01 8.85797314e-03

1.83059753e-02 -2.39902591e-01 -4.79006197e-02 6.19970446e-02

3.83160891e-03]

[-1.86610828e-01 -2.57092552e-01 1.03715887e-01 2.23982467e-01

-1.28215768e-01 1.25669415e-02 3.91400512e-01 -5.66073056e-01

-5.39235753e-01 -1.47628917e-01 -1.73918533e-01 -2.40534190e-02

-8.03169296e-05 -4.89753356e-02 3.58875507e-02 2.80805469e-02

-7.32598621e-03]

[-3.28955847e-01 -1.60008951e-01 -1.84205687e-01 -2.13756140e-01

2.24240837e-02 -2.31562325e-01 -1.50501305e-01 -1.18823549e-01

2.42371616e-02 -8.04154875e-02 3.93722676e-01 1.05658769e-02

5.60069250e-02 -6.90417042e-01 -1.26667522e-01 1.28739213e-01

1.45099786e-01]

[-2.38822447e-01 -1.67523664e-01 2.45335837e-01 3.61915064e-02

-3.56843227e-01 3.13556243e-01 4.68641965e-01 1.80458508e-01

3.15812873e-01 4.88415259e-01 8.72638706e-02 -2.51028410e-03

1.48410810e-02 -1.59332164e-01 -6.30737002e-02 -7.09643331e-03

-3.29024228e-03]]

## 2.6) Write the explicit form of the first PC (in terms of Eigen Vectors).

**Explicit form of the first PC would be the cross multiplication of variables and eigen vector of 0 index**

### Eigen vector value with 0 index

Output is as shown below:-

array([-0.262, -0.231, -0.189, -0.339, -0.335, -0.163, -0.022, -0.284,

-0.244, -0.097, 0.035, -0.326, -0.323, 0.163, -0.187, -0.329,

-0.239])

### PC1 explicit form

**PC1 = -.262 \* Apps - 0.231 \* Accept - 0.189 \* Enroll - 0.339 \* Top10perc -0.335 \* Top25perc - 0.163 \* FUndergrad -.022 \* PUndergrad -.284 \* Outstate - 0.244 \* RoomBoard -0.097 \* Books +.035 \* Personal - 0.326 \* PhD -0.323 \* Terminal +0.163 \* SFRatio - 0.187 \* PercAlumni - 0.329 \* Expend -0.239 \* GradRate**

**From SKlearn Eigen vector values are :-**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Apps | Accept | Enroll | Top10perc | Top25perc | FUndergrad | PUndergrad | Outstate | RoomBoard |
| PC1 | 0.262 | 0.231 | 0.189 | 0.339 | 0.335 | 0.163 | 0.022 | 0.284 | 0.244 |
|  | Books | Personal | PhD | Terminal | SFRatio | PercAlumni | Expend | GradRate |  |
| PC1 | 0.097 | -0.035 | 0.326 | 0.323 | -0.163 | 0.187 | 0.329 | 0.239 |  |

**PC1 = 0.262 \* Apps + 0.231 \* Accept + 0.189 \* Enroll + 0.339 \* Top10perc + 0.335 \* Top25perc + 0.163 \* FUndergrad +.022 \* PUndergrad +.284 \* Outstate + 0.244 \* RoomBoard + 0.097 \* Books -.035 \* Personal + 0.326 \* PhD +0.323 \* Terminal - 0.163 \* SFRatio + 0.187 \* PercAlumni + 0.329 \* Expend + 0.239 \* GradRate**

**We can see that in both the form coefficients are same but signs are reversed.**

## 2.7) Discuss 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? Perform PCA and export the data of the Principal Component scores into a data frame.

### Cumulative Distribution of Eigenvalues

From python, we can see the cumulative Variance explained (shown below):-

Cumulative Variance Explained [ 33.26608367 62.02142867 68.63859223 74.53673619 79.66062886

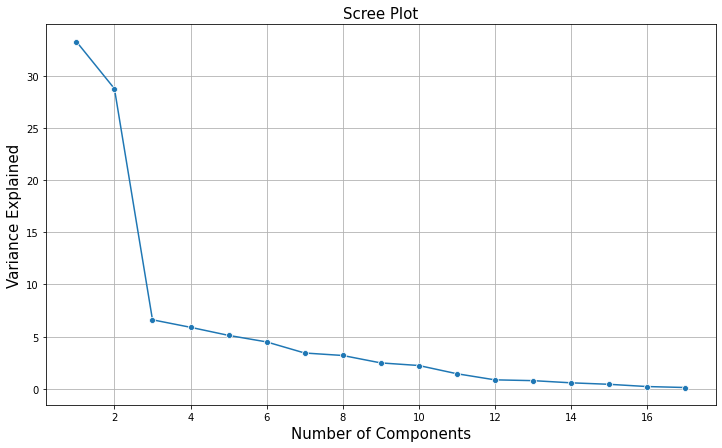
84.15926753 87.59551019 90.79435736 93.28246491 95.52086136

96.97201814 97.83716159 98.62640821 99.20703552 99.64582321

99.86844192 100. ]

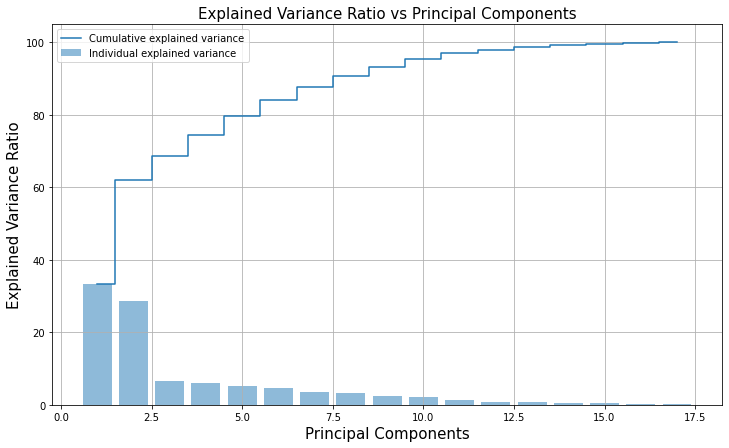
The Cumulative % gives the percentage of variance accounted for by the n components. For example, the cumulative percentage for the second component is the sum of the percentage of variance for the first and second components. It helps in deciding the number of components by selecting the components which explained the high variance.

### Scree Plot

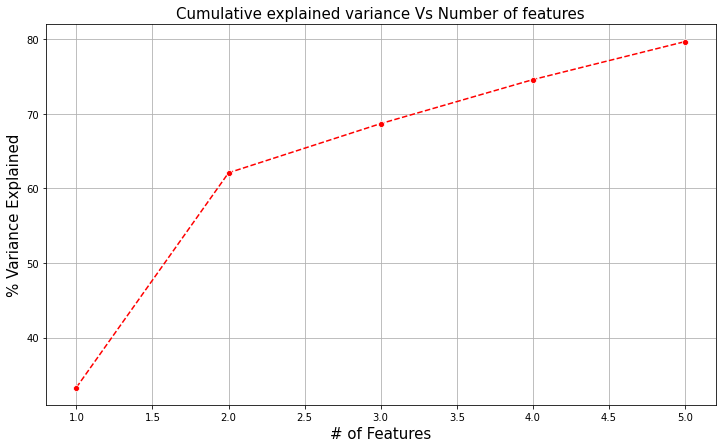


* **Visually we can observe that the steep drop in variance explained with increase in number of PC's.**
* **We will proceed with 5 components here. But depending on requirement for 90% variation, 8 components will also do good**

### Plot Cumulative explained variance and individual explained variance vs Principal Components



### Plot Cumulative explained variance vs No of features



### Decision on the optimum number of principal components:-

In the above array we see that the first feature explains 33.3% of the variance within our

data set while the first two explain 62.1 and so on. We can see that with PC5 around 80%

variance is explained. On-going from PC5 to PC6, there is a marginal (4%) increase in the

variation. If we employ 5 features we capture ~ 80% of the variance within the dataset, thus we gain very little by implementing an additional feature (think of this as diminishing marginal return on total variance explained). Hence we are stopping at PC5 only.

**What do the eigenvectors indicate?**

The eigenvectors and eigenvalues of a covariance (or correlation) matrix represent the “core” of a PCA: **The eigenvectors (principal components) determine the directions of the new feature space,** and the eigenvalues determine their magnitude.

Thus Eigenvectors provided the coefficients of all the PC components. For example in explicit form of the first PC (shown below), all these coefficient values are picked up from the Eigen Vectors.

**PC1 = 0.262 \* Apps + 0.231 \* Accept + 0.189 \* Enroll + 0.339 \* Top10perc + 0.335 \* Top25perc + 0.163 \* FUndergrad +.022 \* PUndergrad +.284 \* Outstate + 0.244 \* RoomBoard + 0.097 \* Books -.035 \* Personal + 0.326 \* PhD +0.323 \* Terminal - 0.163 \* SFRatio + 0.187 \* PercAlumni + 0.329 \* Expend + 0.239 \* GradRate**

This way for all PC, equation can be formed and PC score can be computed using Eigen Vectors. Thus these are extremely important in PCA.

### Perform PCA and export the data of the Principal Component scores into a data frame

As we have decided to go till PC5, table showing standard deviation, proportion of variance and cumulative proportion is as shown below :-

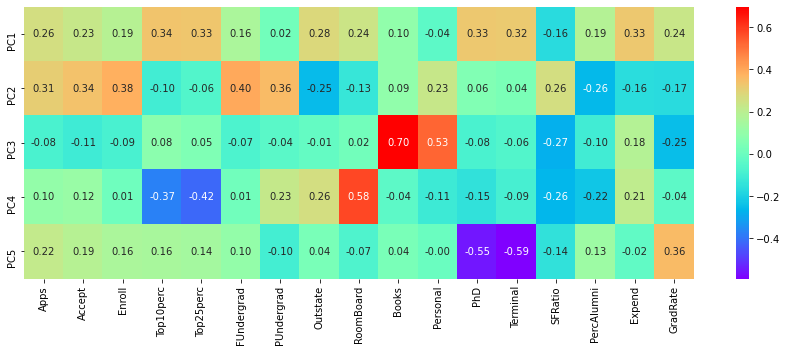
|  |  |  |  |
| --- | --- | --- | --- |
| PCs | Proportion Of Variance | Standard Deviation | Cumulative Proportion |
| PC1 | 0.33 | 2.38 | 0.33 |
| PC2 | 0.29 | 2.21 | 0.62 |
| PC3 | 0.07 | 1.06 | 0.69 |
| PC4 | 0.06 | 1 | 0.75 |
| PC5 | 0.05 | 0.93 | 0.8 |

### Data frame showing PC components

Dataframe showing PC components is as shown below :-

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Apps | Accept | Enroll | Top10perc | Top25perc | FUndergrad | PUndergrad | Outstate | RoomBoard |
| PC1 | 0.26 | 0.23 | 0.19 | 0.34 | 0.33 | 0.16 | 0.02 | 0.28 | 0.24 |
| PC2 | 0.31 | 0.34 | 0.38 | -0.1 | -0.06 | 0.4 | 0.36 | -0.25 | -0.13 |
| PC3 | -0.08 | -0.11 | -0.09 | 0.08 | 0.05 | -0.07 | -0.04 | -0.01 | 0.02 |
| PC4 | 0.1 | 0.12 | 0.01 | -0.37 | -0.42 | 0.01 | 0.23 | 0.26 | 0.58 |
| PC5 | 0.22 | 0.19 | 0.16 | 0.16 | 0.14 | 0.1 | -0.1 | 0.04 | -0.07 |
|  |  |  |  |  |  |  |  |  |  |
|  | Books | Personal | PhD | Terminal | SFRatio | PercAlumni | Expend | GradRate |  |
| PC1 | 0.1 | -0.04 | 0.33 | 0.32 | -0.16 | 0.19 | 0.33 | 0.24 |  |
| PC2 | 0.09 | 0.23 | 0.06 | 0.04 | 0.26 | -0.26 | -0.16 | -0.17 |  |
| PC3 | 0.7 | 0.53 | -0.08 | -0.06 | -0.27 | -0.1 | 0.18 | -0.25 |  |
| PC4 | -0.04 | -0.11 | -0.15 | -0.09 | -0.26 | -0.22 | 0.21 | -0.04 |  |
| PC5 | 0.04 | 0 | -0.55 | -0.59 | -0.14 | 0.13 | -0.02 | 0.36 |  |

### Correlation between components and features



**Inferences from the above Heatmap**

This heatmap and the color bar basically represent the correlation between the various feature and the principal component itself

Component 3 looks more related to student expense on Books and Personal Expense - We can label it as student expense

PC5 is more related to PhD and Terminal - we can label it as faculty qualification

PC4 is more related to cost of Room and Board and also with Top10perc and top25perc.

## 2.8) Mention the business implication of using the Principal Component Analysis for this case study.

### PCA

PCA is a statistical technique and uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. PCA also is a tool to reduce multidimensional data to lower dimensions while retaining most of the information. Principal Component Analysis (PCA) is a well-established mathematical technique for reducing the dimensionality of data, while keeping as much variation as possible.

### Business Implication:

We have seen that data set contains 18 columns of data in which one column contains name of the colleges and all other columns provides us some useful characteristics to enable us to rate the colleges. There is no target variable associated with the given mentioned data set.

As seen in the analysis, there are many columns which have relationship with each other. So there is multicollinearity problem in the given data and it can be problematic for many models. Also it is hard to compare 17 columns.

So using Principal component technique we have brought this dataset from higher dimension to a lower dimension i.e. with only 5 columns. **Using only 5 PC's we have attained around 80% variability in the data and we have reduced the dimensions from 17 to 5 which is a significant reduction.**

Further apart from dimension reduction, problem of multicollinearity is also solved as new 5 PC's are orthogonal in nature i.e. independent of each other.

Lastly on checking relationship between components and features, we have found that :-

Among these 5 PCs, PC3 is more related to student personal and books expense. We can label these columns to say student expense.

PC5 is more related to PhD and Terminal. Thus we can label it as faculty qualification.

PC4 is more related to cost of Room and Board and also with Top10perc and top25perc. We can discuss it further with the SMEs and label it accordingly.