AUTOMOBILE DATA ANALYSIS

Mohamed Rifaz Ali K S

PGP-DSBA Online

September’ 21

Contents

Problem 1 6

Executive Summary 6

Introduction 6

Data Description 6

Sample of the Salary Data dataset: 7

Let us check the types of variables in the data frame. 7

Check for missing values in the dataset: 7

Problem 1A: 8

1.1. State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually. 8

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 9

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. 9

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. 10

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. 11

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? 12

1.7. Explain the business implications of performing ANOVA for this particular case study. 14

Problem 2: 15

Executive Summary 16

Introduction 16

Data Description 16

Sample of the Education+-+Post+12th+Standard dataset: 17

Exploratory Data Analysis and PCA 18

Let us check the types of variables in the data frame. 18

2.1. Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA? 19

Univariate:........ 20

Multivariate........ 23

2.2. Is scaling necessary for PCA in this case? Give justification and perform scaling. 25

Histogram before scaling 26

Histogram after applying z-score(scaling)........ 27

2.3. Comment on the comparison between the covariance and the correlation matrices from this data.[on scaled data] 30

Strong correlation 31

Weak Correlation 32

2.4. Check the dataset for outliers before and after scaling. What insight do you derive here?........ 33

Outliers before scaling 33

Outliers after scaling 34

2.5. Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both] 37

Eigen vectors 38

Eigen values........ 40

2.6. Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features........ 41

Scree plot 42

Cumulative Explained Variance Vs. number of features 43

Absolute loadings of PCs........ 44

Correlation between components and features........ 45

Heat map to represent original features influence various PCs 45

Heat map to show correlation between features & PCs itself 46

Final PCA table with new labels 47

Heat map to show correlations among PCs........ 48

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] 49

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? 50

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] 52

List of Tables

Table 1. Dataset Sample 7
Table 2. One-Way ANOVA for Education 9
Table 3. F-Significant value for Education 9
Table 4. One-Way ANOVA for Occupation 9
Table 5. F-Significant value for Occupation 10
Table 6. Tukey's HSD test for Education 10
Table 7. Two-Way ANOVA 12
Table 8. Two-Way ANOVA with interaction 13
Table 9. Sample Education+-+Post+12th+Standard Dataset 17
Table 9. Duplicates check in the Dataset 17
Table 10. Information about the Dataset 18
Table 11. Describe the Dataset 19
Table 12. Sample Dataset after scaling 25
Table 13. Summarize the Dataset before scaling 28
Table 14. Summarize the Dataset after scaling 29
Table 15. Summarize the Dataset before scaling 35
Table 16. Summarize the Dataset after scaling 36
Table 17. Principal Component Dataset with original features 41
Table 18. Correlation between PC's and features 45
Table 19. Final PCA dataset 47

List of Figures

Figure 1. Interaction between Education and Occupation using interaction plot 11
Figure 2. Interaction between Education and Occupation using pointplot........ 13
Figure 3. Histogram of all numerical variables to find the distribution........ 21
Figure 4. Boxplot of all numerical variables to find the outliers........ 22
Figure 5. Heat map to check the correlation of the data........ 23
Figure 6. Pair plot to visualize the relationship among variables 24
Figure 7. z-score formula 25
Figure 8. Histograms before scaling........ 26
Figure 9. Histograms after scaling........ 27
Figure 10. Heat map to check the correlation of the data post scaling........ 30
Figure 11. Strong correlation representation between variables using regplot 31
Figure 12. Weak correlation representation between variables using regplot........ 32
Figure 13. Boxplot to identify the outliers for the original dataset(before scaling) 33
Figure 14. Boxplot to identify the outliers after applying z-score(after scaling) 34
Figure 15. Bartlett Sphericity test 37
Figure 16. Kaiser-Meyer-Olkin test 38
Figure 17. Screeplot to identify the principal components 42
Figure 18. Plot to identify the number of components needed to explain variance 43
Figure 19.Representation of each principal components........ 44
Figure 20.Heatmap to represent original features influence various PCs........ 45
Figure 21.Heatmap to show correlation between features & PCs itself 46
Figure 21.Heatmap to show correlations among PCs........ 48
Figure 22.Plot to identify number of PCs........ 51
Figure 23.Area plot to show the variances at each component level(PC2)........ 53
Figure 24. Area plot to show the variances at each component level(PC7)........ 53

Problem 1

Executive Summary

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals 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. In this problem, we will explore the mean salary of education – occupation at their levels are same or at least in any one of the levels it is getting varied.

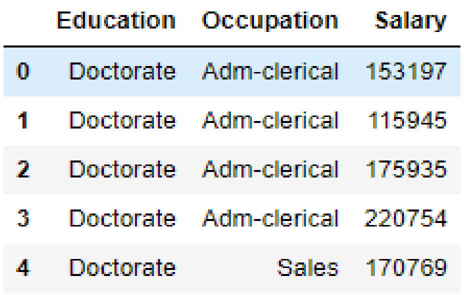
Introduction

The purpose of this problem is to figure out the aggregate variability of the 40 salaried individuals is present inside the observed dataset(SalaryData.csv) at education – occupation at any one of the levels using the Analysis of Variance(ANOVA) statistical model technique. In Problem 1A, we will deal with One-Way ANOVA technique to figure out the variation among groups at each independent variables(i.e., education – occupation). In Problem 1B, we will deal with Two-Way ANOVA technique to figure out the variation between the groups that have been split on two independent variables(called factors)(i.e., education and occupation).

Data Description

- Education: at three levels, High school graduate, Bachelor, and Doctorate
- Occupation: at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial
- Salary: Salary

Sample of the Salary Data dataset:



	Education	Occupation	Salary
0	Doctorate	Adm-clerical	153197
1	Doctorate	Adm-clerical	115945
2	Doctorate	Adm-clerical	175935
3	Doctorate	Adm-clerical	220754
4	Doctorate	Sales	170769

Table 1. Dataset Sample

Let us check the types of variables in the data frame.

- Education – Object
- Occupation – Object
- Salary – int64

There are 40 rows and 3 columns present in the dataset. Out of 3 columns, 2 columns are 'objects' and 1 is 'integer'.

Check for missing values in the dataset:

```
RangeIndex: 40 entries, 0 to 39
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Education   40 non-null     object
 1   Occupation  40 non-null     object
 2   Salary      40 non-null     int64 
dtypes: int64(1), object(2)
memory usage: 1.1+ KB
```

From the above results we can see that there is no missing value present in the dataset.

Problem 1A:

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.

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

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

Assumptions for Education

H_0: Mean salary of the Education at three levels are same.

H_A: At least one of the mean salary of the Education is NOT same.

Assumptions for Occupation

H_0: Mean salary of the Occupation at four levels are same.

H_A: At least one of the mean salary of the Occupation is NOT same.

If the p-value is < 0.05, then we reject the H_0.

If the p-value is >= 0.05, then we fail to reject the H_0, H_A is considered.

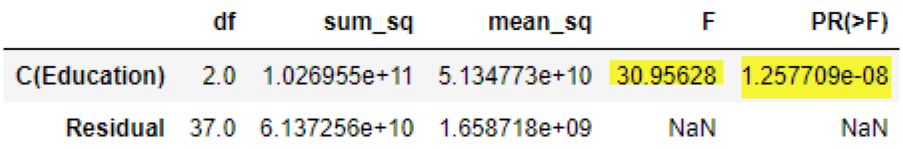
Where 0.05 represents the Alpha(significance level)

H_0 – Null Hypothesis

H_A – Alternate Hypothesis

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.

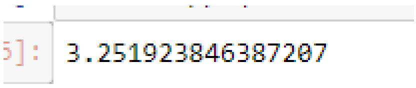
ANOVA table for Education



	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	30.95628	1.257709e-08
Residual	37.0	6.137256e+10	1.658718e+09	NaN	NaN

Table 2. One-Way ANOVA for Education

F-Significant value



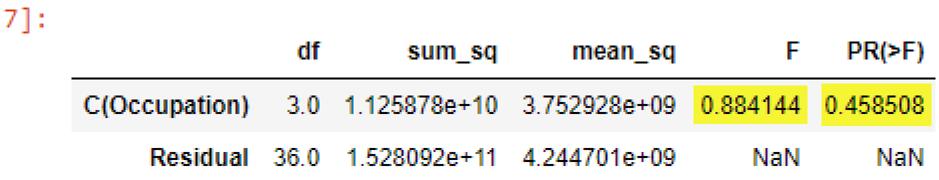
```
5]: 3.251923846387207
```

Table 3. F-Significant value for Education

Since the p-value is less than the α (significance level), we **can reject the null hypothesis and states that there is a difference in the mean salary at least in one of the different education levels.** Also, F-statistic value is greater than F-significant value, hence reject $H0$.

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.

ANOVA table for Occupation

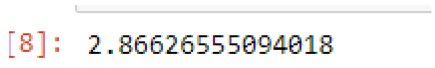


7]:

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	0.884144	0.458508
Residual	36.0	1.528092e+11	4.244701e+09	NaN	NaN

Table 4. One-Way ANOVA for Occupation

F-Significant value



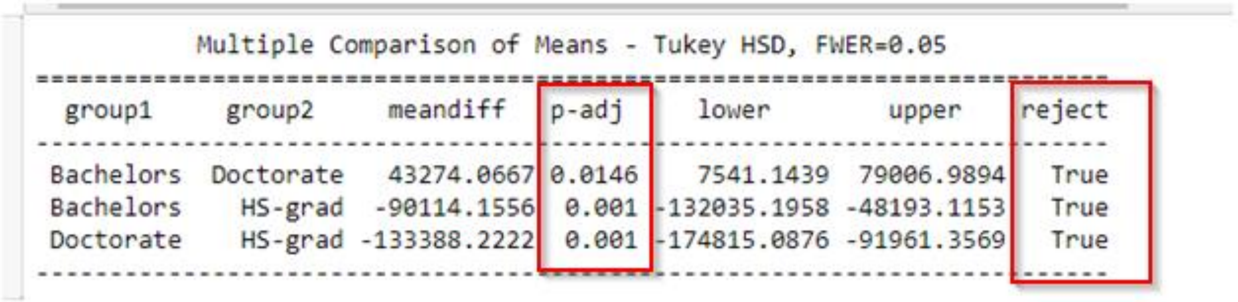
```
[8]:  2.86626555094018
```

Table 5. F-Significant value for Occupation

Since the p-value is greater than the α (significance level), there is **no evidence to reject the null hypothesis and hence the mean salaries of the individuals at different Occupation levels are equal.** Also, F-statistic value is lesser than F-significant value, hence we cannot reject $H0$.

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.

ANOVA test helps us to find out the whether the differences between groups of data are statistically significant or not, but it does not tell you 'where' the difference lies. In this case, Tukey HSD test helps to figure out which specific group the difference lies. It will compare at each combination of groups.



```
          Multiple Comparison of Means - Tukey HSD, FWER=0.05
=========================================================================
  group1     group2     meandiff   p-adj     lower         upper     reject
-------------------------------------------------------------------------
 Bachelors Doctorate   43274.0667 0.0146    7541.1439   79006.9894    True
 Bachelors   HS-grad  -90114.1556  0.001 -132035.1958  -48193.1153    True
 Doctorate   HS-grad -133388.2222  0.001 -174815.0876  -91961.3569    True
-------------------------------------------------------------------------
```

Table 6. Tukey's HSD test for Education

Above results from Tukey's HSD suggests that all the pairwise comparisons for treatments rejects null hypothesis (p-value < 0.05, reject=True) and indicates statistical **significant differences among the salaries at all levels of the Education.** Considering, Doctorate salary is greater than Bachelors & High-School graduates.

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.

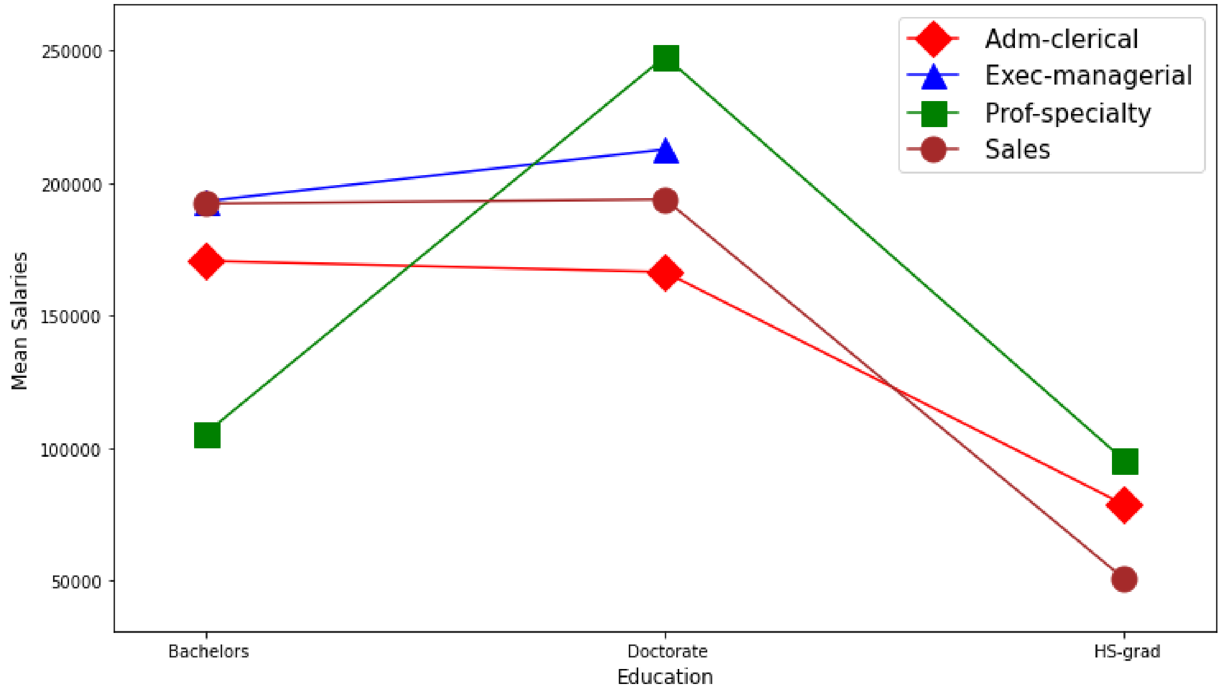


Figure 1. Interaction between Education and Occupation using interaction plot

Observations

- Doctorates: Certainly, the individuals who pursued the Doctorate Education offered the higher salaries in the Professional Specialty jobs. And, again in Exec-Managerial and Sales jobs, they fall into higher salaried brackets. With Adm-clerical, average salaries for Doctorates slightly less compared to Bachelors & few Doctorates are working.
- Bachelors: They fall into middle-income brackets. Average salaries with Sales & Exec-Managerial job roles are exactly same. For Adm-Clerical job roles, their mean income is slightly greater than the even the Doctorates. Few are preferred to work as a Professional Specialty jobs.

- HS-Grad: They fall into lower-income groups compared to the other two. None of them are working in Exec-Managerial job role.
- With 'Professional Specialty' job role, the average salaries for Bachelors and High School Graduates are mostly same.
- With 'Sales' & 'Adm-clerical' job role, the average salaries for Doctorates and Bachelors are mostly same. High School graduates salaries are nowhere nearer to them.
- 'Exec-Managerial' job role has no interaction with the High School graduates.

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?

H_0: The means of 'salary' variable with respect to each Education & Occupation level is equal.

H_A: At least one of the means of 'salary' variable with respect to each Education & Occupation level is unequal.

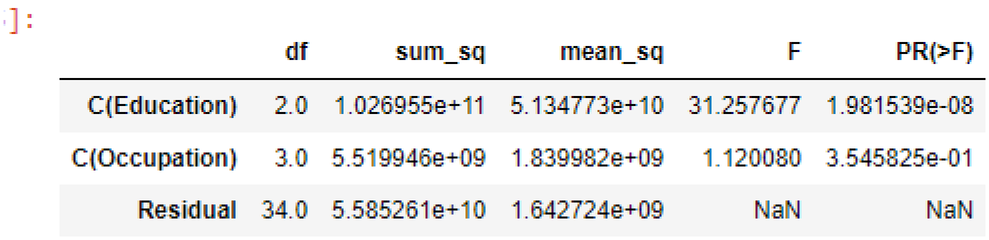
If the p-value is < 0.05, then we reject the H_0.

If the p-value is >= 0.05, then we fail to reject the H_0, H_A is considered.

Where 0.05 represents the Alpha(α - significance level)

H_0 – Null Hypothesis

H_A – Alternate Hypothesis



]:

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	31.257677	1.981539e-08
C(Occupation)	3.0	5.519946e+09	1.839982e+09	1.120080	3.545825e-01
Residual	34.0	5.585261e+10	1.642724e+09	NaN	NaN

Table 7. Two-Way ANOVA

We see the changes in the p-value in Two Way ANOVA, so there is some sort of interaction between the two treatments.

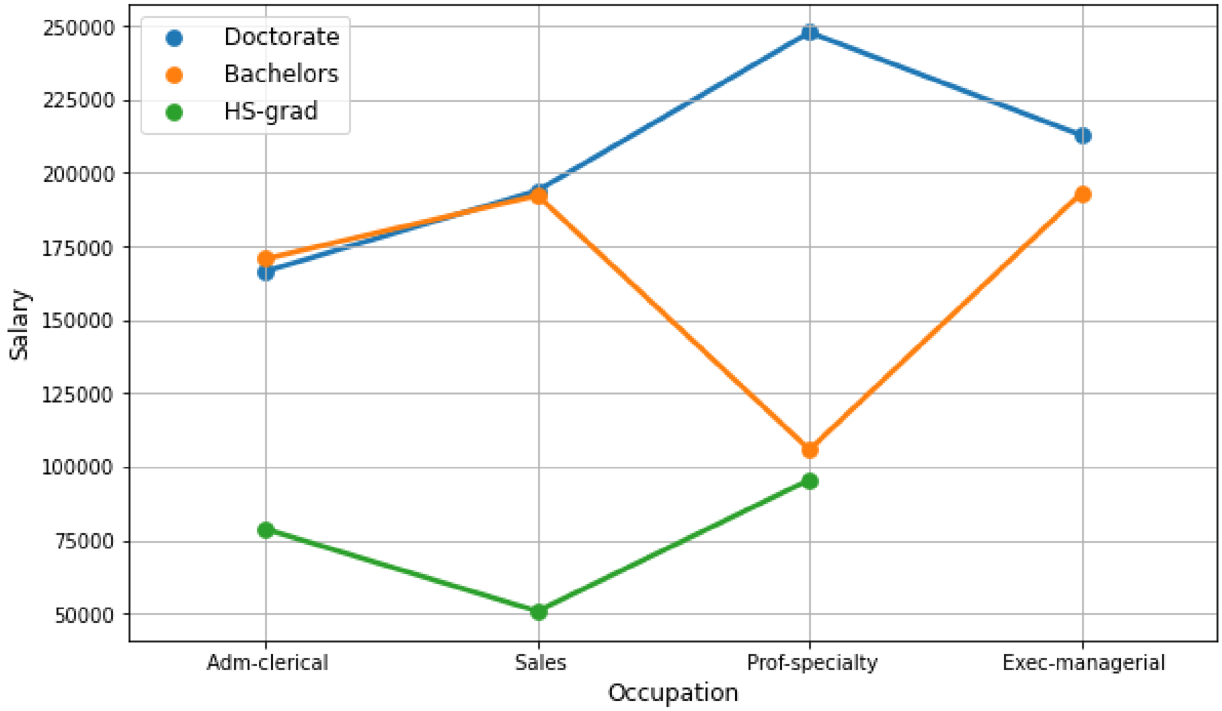
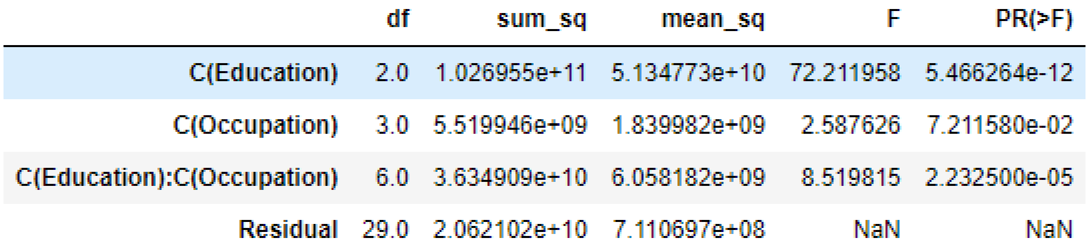


Figure 2. Interaction between Education and Occupation using pointplot

Also, from the interaction plot, we could see that there is some sort of interaction between the two treatments. So, we will introduce a new term while performing the Two Way ANOVA.

:



	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
C(Occupation)	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
C(Education):C(Occupation)	6.0	3.634909e+10	6.058182e+09	8.519815	2.232500e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

Table 8. Two-Way ANOVA with interaction

We can see the changes in the p-value of the two treatments when we introduce the new interaction variable. Here, p-value is < 0.05, we can **reject the null hypothesis. Interaction between the Education and the Occupation significantly(P<0.05) affects the Salary** from the above result.

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

- ANOVA stands for Analysis of Variance. One-Way Analysis of Variance tells you if there are any statistical differences between the means **of three or more independent groups**. A two-way ANOVA is used to estimate how the mean of a quantitative variable changes according to the levels of two categorical variables.
- In our case, we would use ANOVA to understand how the independent variables 'education' and 'occupation' groups respond to the dependent variable 'salary', with a null hypothesis for the test that the means of the different groups are equal. If there is a statistically significant result, then it means that the two populations are unequal (or different).
- ‘Education’ levels are Doctorate, Bachelors and High School graduates. Whereas, ‘Occupation’ levels of those are Adm-clerical, Professional specialty, Sales and Exec-managerial job roles.
- In general, whoever pursued the highest level of education earns more salary. In our case, Doctorates earns more salary than Bachelors and HS-grad. Bachelors earns more salary than HS-grad.
- Doctorates fall into high range income brackets, Bachelors into middle income range brackets and HS-grad fall into low-range buckets.
- There are 40 samples of mean salaries are collected from these categories.
- Observed from our dataset, the above assumption is not always true. There are samples where Bachelors earns more than Doctorates. For eg., Adm-clerical role with Bachelors education group(row 19) earns 188729/- which is greater than few of the Doctorates with Adm-clerical role. Similar observations goes with HS-grad with other levels of education groups.

- For some of the job role, especially professional-specialty, Doctorates clearly stands out from the other education groups.
- So, from the ANOVA test, we can say pursing higher levels education or working in some job roles alone does not fall the individuals into higher salary buckets. Rather, the **combination of education and the levels of job role plays significantly impacts the salary range of the employees.**

Problem 2:

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.

- Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?
- Is scaling necessary for PCA in this case? Give justification and perform scaling.
- Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].
- Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]
- Extract the eigenvalues and eigenvectors.[Using Sklearn PCA Print Both]
- Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features
- 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]
- 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?
- 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]

Executive Summary

The dataset contains the information on various colleges/universities in USA. It has the information about the college/university names, part/full time UG students, costs of books, board, room, personal estimated spending of the student, toppers from High Schools, percentage of faculties with PhD & terminal degree, Student/Faculty ratio and so on. In this problem statement, we will explore the different attributes of colleges/universities in USA.

Introduction

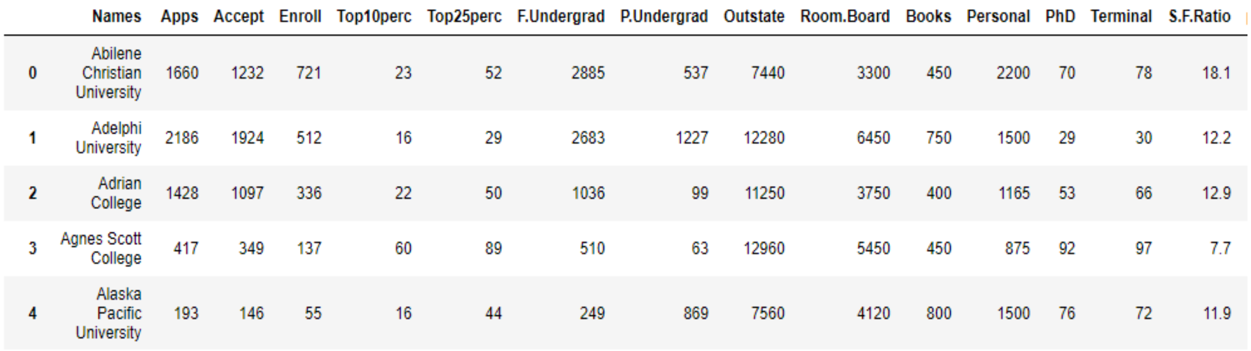
The purpose of this whole exercise is to explore the dataset(Education+-+Post+12th+Standard.csv). Do the Exploratory Data Analysis and Principal Component Analysis. The data consists of 777 different colleges/universities with 18 attributes of colleges. Analyze the different attributes of the college features which can help in analyzing the characteristics of the college.

Data Description

1. Names: Names of various university and colleges
2. Apps: Number of applications received
3. Accept: Number of applications accepted
4. Enroll: Number of new students enrolled
5. Top10perc: Percentage of new students from top 10% of Higher Secondary class
6. Top25perc: Percentage of new students from top 25% of Higher Secondary class
7. F.Undergrad: Number of full-time undergraduate students
8. P.Undergrad: Number of part-time undergraduate students
9. Outstate: Number of students for whom the particular college or university is Out-of-state tuition
10. Room.Board: Cost of Room and board
11. Books: Estimated book costs for a student
12. Personal: Estimated personal spending for a student
13. PhD: Percentage of faculties with Ph.D.'s
14. Terminal: Percentage of faculties with terminal degree

15. S.F.Ratio: Student/faculty ratio
16. perc.alumni: Percentage of alumni who donate
17. Expend: The Instructional expenditure per student
18. Grad.Rate: Graduation rate

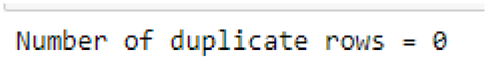
Sample of the Education+-+Post+12th+Standard dataset:



	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio
0	Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2
2	Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9
3	Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7
4	Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9

Table 9. Sample Education+-+Post+12th+Standard Dataset

Dataset has 777 rows and 18 columns. Each college has different sets of attributes.



```
Number of duplicate rows = 0
```

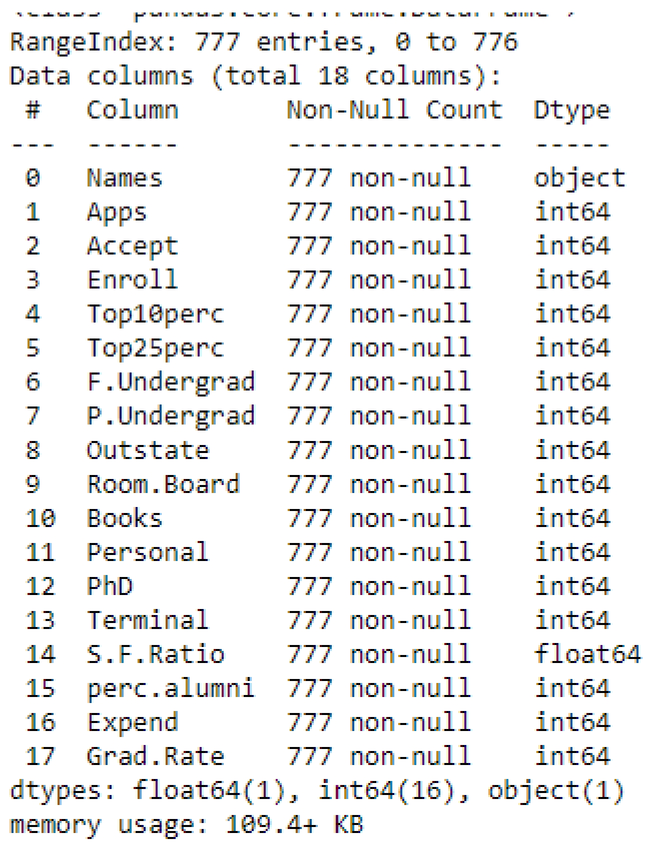
Table 10. Duplicates check in the Dataset

There are no duplicate samples present in the entire dataset.

Exploratory Data Analysis and PCA

Let us check the types of variables in the data frame.

There are 777 rows and 18 columns present in the dataset. Out of 18 columns, 16 of them are in 'integer' and one of them(S.F.Ratio) is in 'float' and one of them(Names) in 'object' data types. From the above results we can see that there is no missing value present in the dataset.

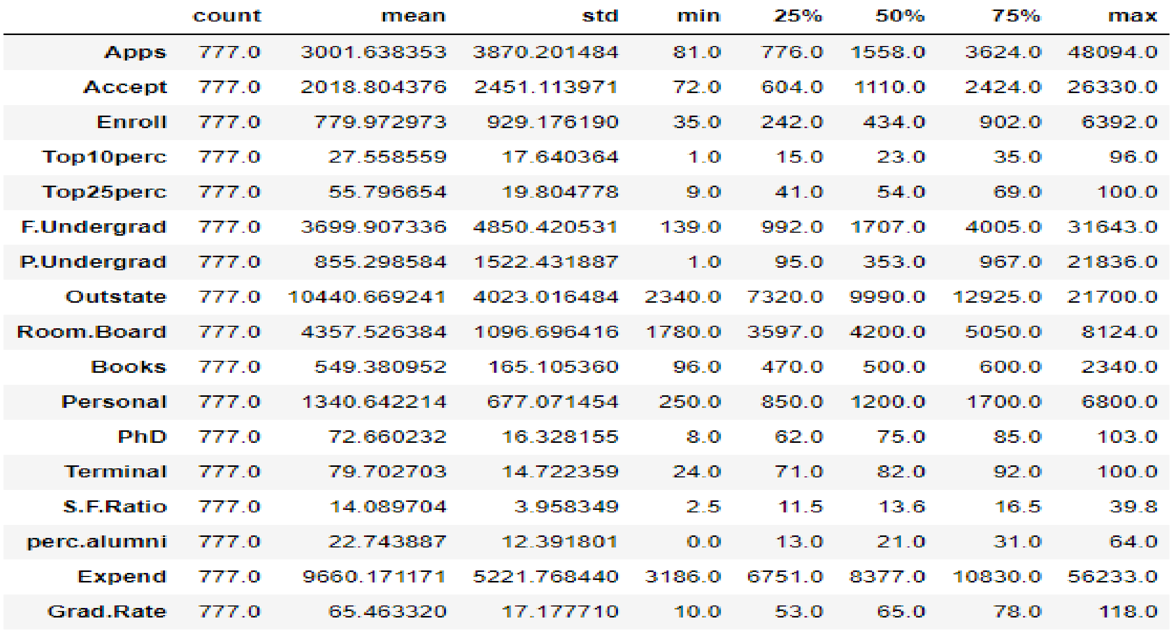


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 777 entries, 0 to 776
Data columns (total 18 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Names        777 non-null    object 
 1   Apps         777 non-null    int64  
 2   Accept       777 non-null    int64  
 3   Enroll       777 non-null    int64  
 4   Top10perc    777 non-null    int64  
 5   Top25perc    777 non-null    int64  
 6   F.Undergrad  777 non-null    int64  
 7   P.Undergrad  777 non-null    int64  
 8   Outstate     777 non-null    int64  
 9   Room.Board   777 non-null    int64  
 10  Books        777 non-null    int64  
 11  Personal     777 non-null    int64  
 12  PhD          777 non-null    int64  
 13  Terminal     777 non-null    int64  
 14  S.F.Ratio    777 non-null    float64
 15  perc.alumni  777 non-null    int64  
 16  Expend       777 non-null    int64  
 17  Grad.Rate    777 non-null    int64  
dtypes: float64(1), int64(16), object(1)
memory usage: 109.4+ KB
```

Table 11. Information about the Dataset

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

- For EDA, there are few steps to consider.
- Import the all the necessary libraries and the data.
- Understand the data using the describe() function and familiar with it. Also, to see if any **anomalies** present in the dataset.
- Check the data and its data types.
- Find whether there are any **duplicate samples** present in the dataset. Drop the duplicated rows to get the distinct records.
- Check the **outliers** using the boxplot. Based on the requirement, either drop the outlier values or treat the outlier values using Inter Quartile Range.
- Check if **missing values** are present in the dataset. Based on the requirement, either you drop it or replace it with mean, median and mode values.
- Standardize the data(Scaling by **z-score** technique)
- **Encoding** Technique.



	count	mean	std	min	25%	50%	75%	max
Apps	777.0	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	779.972973	929.176190	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	55.796654	19.804778	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	3699.907336	4850.420531	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	855.298584	1522.431887	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	10440.669241	4023.016484	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	4357.526384	1096.696416	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	549.380952	165.105360	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	1340.642214	677.071454	250.0	850.0	1200.0	1700.0	6800.0
PhD	777.0	72.660232	16.328155	8.0	62.0	75.0	85.0	103.0
Terminal	777.0	79.702703	14.722359	24.0	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	14.089704	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777.0	22.743887	12.391801	0.0	13.0	21.0	31.0	64.0
Expend	777.0	9660.171171	5221.768440	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	65.463320	17.177710	10.0	53.0	65.0	78.0	118.0

Table 12. Describe the Dataset

Observations:

- Out of 18 columns, 17 of them are numerical columns such as Apps , Accept , Enroll , Top10perc , Top25perc , F.Undergrad , P.Undergrad , Outstate , Room.Board , Books , Personal , PhD , Terminal , S.F.Ratio , perc.alumni , Expend , Grad.Rate and one Names is a categorical column.
- No duplicates and missing values are present.
- Upon looking at the average and minimum/maximum values, it clearly indicates most of the columns are having outliers and it needs to be cleaned.
- The maximum value of PhD and Grad.Rate is 103% and 118% respectively. Percentage of these columns should not go beyond 100% which needs to be cleaned. It can be corrected using the Median values.
- The columns such as Apps, Accept, Enroll, F.Undergrad, P.Undergrad, Books and Personal, their minimum values and maximum values are drastically differ, which needs a cleanup as well.

Univariate:

The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it. In a dataset, it explores each variable separately.

Using the histogram, we will see the distribution of the variables present in the dataset.

In our Education+-+Post+12th+Standard dataset, we will distribution of all the 17 numerical continuous variables such as Apps, Accept, Enroll, Top10perc, F.Undergrad, P.Undergrad, Books, Personal, S.F.Ratio, perc.alumini, Expend, PhD, Terminal, Top25perc, Outstate, Room.Board and Grad. Rate using the histogram.

Histogram to see the distribution

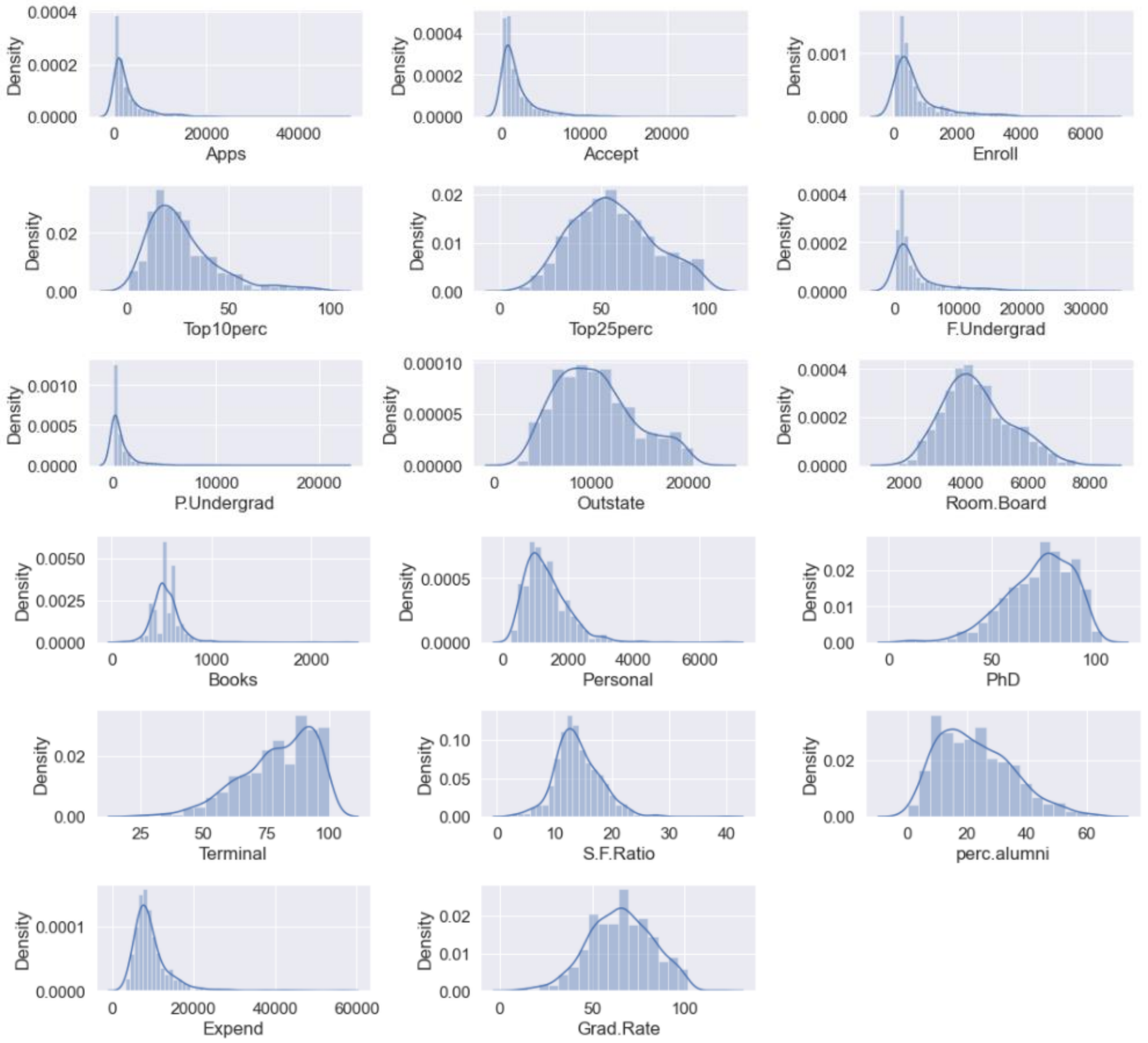


Figure 3. Histogram of all numerical variables to find the distribution

Left Skewed variables: Apps, Accept, Enroll, Top10perc, F.Undergrad, P.Undergrad, Books, Personal, S.F.Ratio, perc.alumini, Expend

Right Skewed variables: PhD, Terminal

Normal Distribution variables: Top25perc, Outstate, Room.Board, Grad. Rate

Boxplot to see the outliers

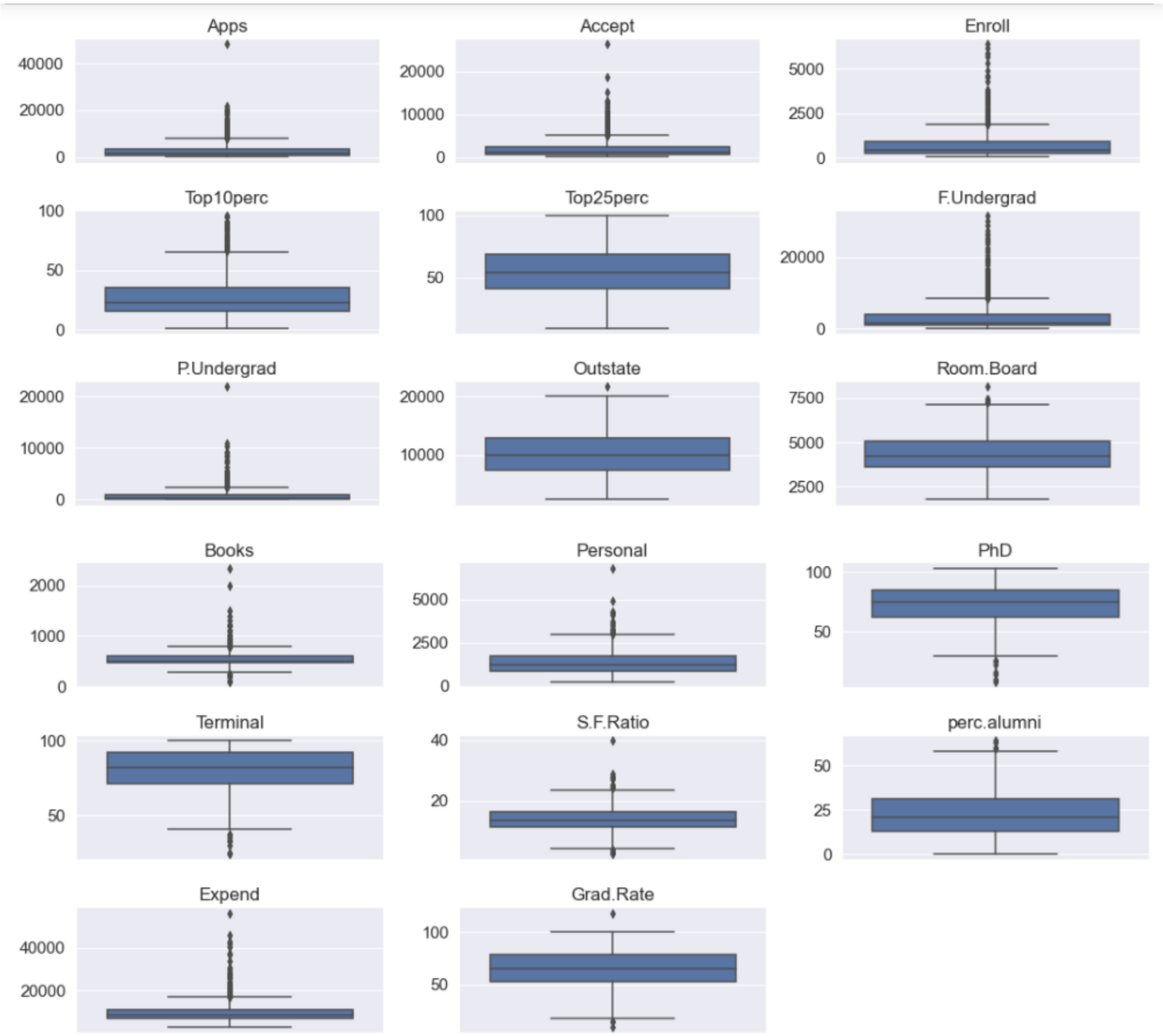


Figure 4. Boxplot of all numerical variables to find the outliers

- Normally distributed variables such as Top25perc has no outliers present and Outstate, Room.Board, and Grad. Rate has very few outliers.
- All the others (right/left)skewed variables has a large number of outliers present in it.

Multivariate

Heat map is to observe the correlation between the features.

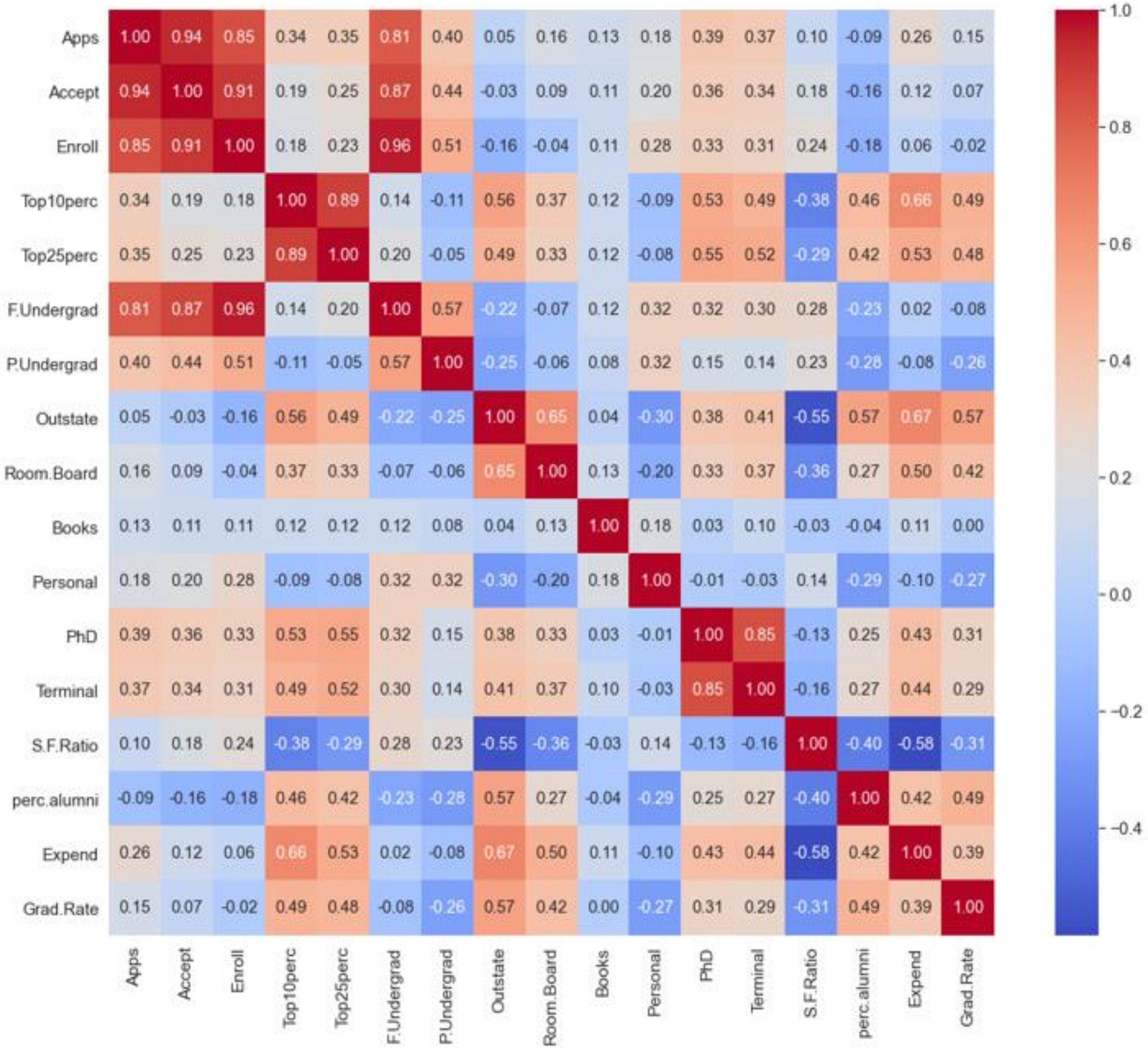


Figure 5. Heat map to check the correlation of the data

From the correlation heat map, we can see that attributes of the colleges are highly correlated in many places. You could see it in orange/brown shaded area.

- Application received and Applications Accepted are very highly correlated.
- Application received and students enrollment are highly correlated.

- Applications Accepted and students enrollment are very highly correlated.
- Percentage of PhD and Terminal faculties are highly correlated.
- Top25perc and Top10perc are very highly correlated.
- Full Time Undergraduate students and students enrollment are very highly correlated.
- Full Time Undergraduate students and Application received/Applications Accepted are highly correlated.

Pair plot

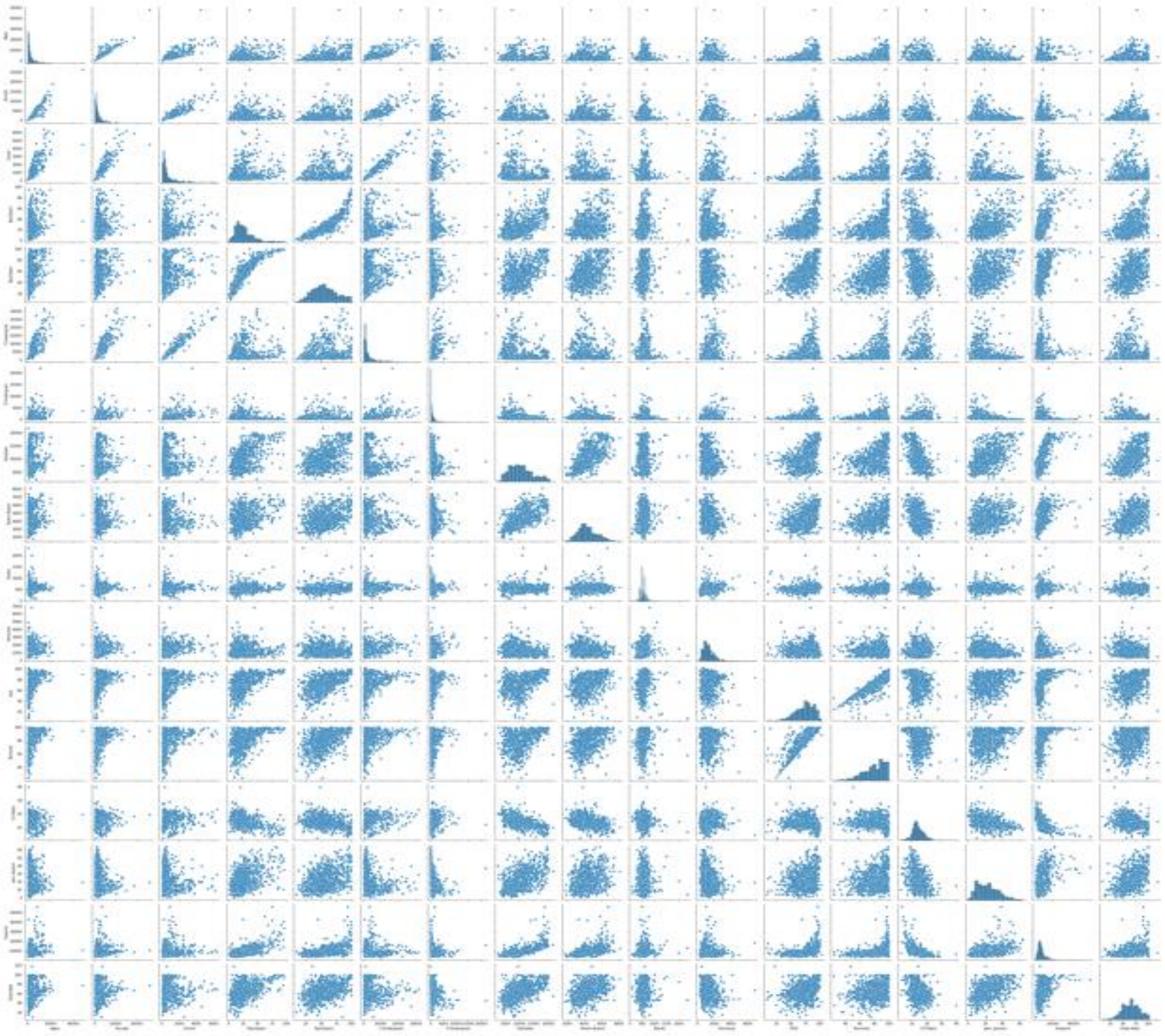


Figure 6. Pair plot to visualize the relationship among variables

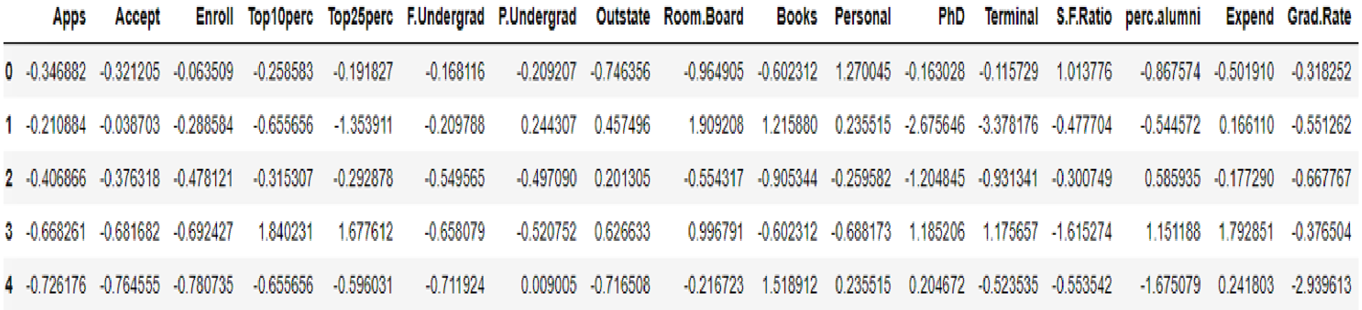
2.2. Is scaling necessary for PCA in this case? Give justification and perform scaling.

- There are 17 numerical columns present in the dataset thus we will get 17 principal components.
- The main idea of principal component analysis (PCA) is to reduce the dimensionality of a dataset consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent.
- Based on our principal component variance, we will come to know how many principal components are suffice to build the model. If we get the cumulative 80% from the principal component variance, it is good enough to reject the rest of the components.
- Importantly, the dataset on which PCA technique is to be used must be scaled.
- In our dataset, each numerical columns are in different units. So, when we do the scaling, data comes to the origin with the unit standard deviation.
- It is necessary to standardize the data before proceeding with PCA. Here, we can use the common scaling technique called z-score.

$$Z = \frac{x - \mu}{\sigma}$$

Figure 7. z-score formula

Z - standard score
x – observed value
µ - mean of the sample
σ – standard deviation



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.964905	-0.602312	1.270045	-0.163028	-0.115729	1.013776	-0.867574	-0.501910	-0.318252
1	-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496	1.909208	1.215880	0.235515	-2.675646	-3.378176	-0.477704	-0.544572	0.166110	-0.551262
2	-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.497090	0.201305	-0.554317	-0.905344	-0.259582	-1.204845	-0.931341	-0.300749	0.585935	-0.177290	-0.667767
3	-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633	0.996791	-0.602312	-0.688173	1.185206	1.175657	-1.615274	1.151188	1.792851	-0.376504
4	-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005	-0.716508	-0.216723	1.518912	0.235515	0.204672	-0.523535	-0.553542	-1.675079	0.241803	-2.939613

Table 13. Sample Dataset after scaling

Histogram before scaling

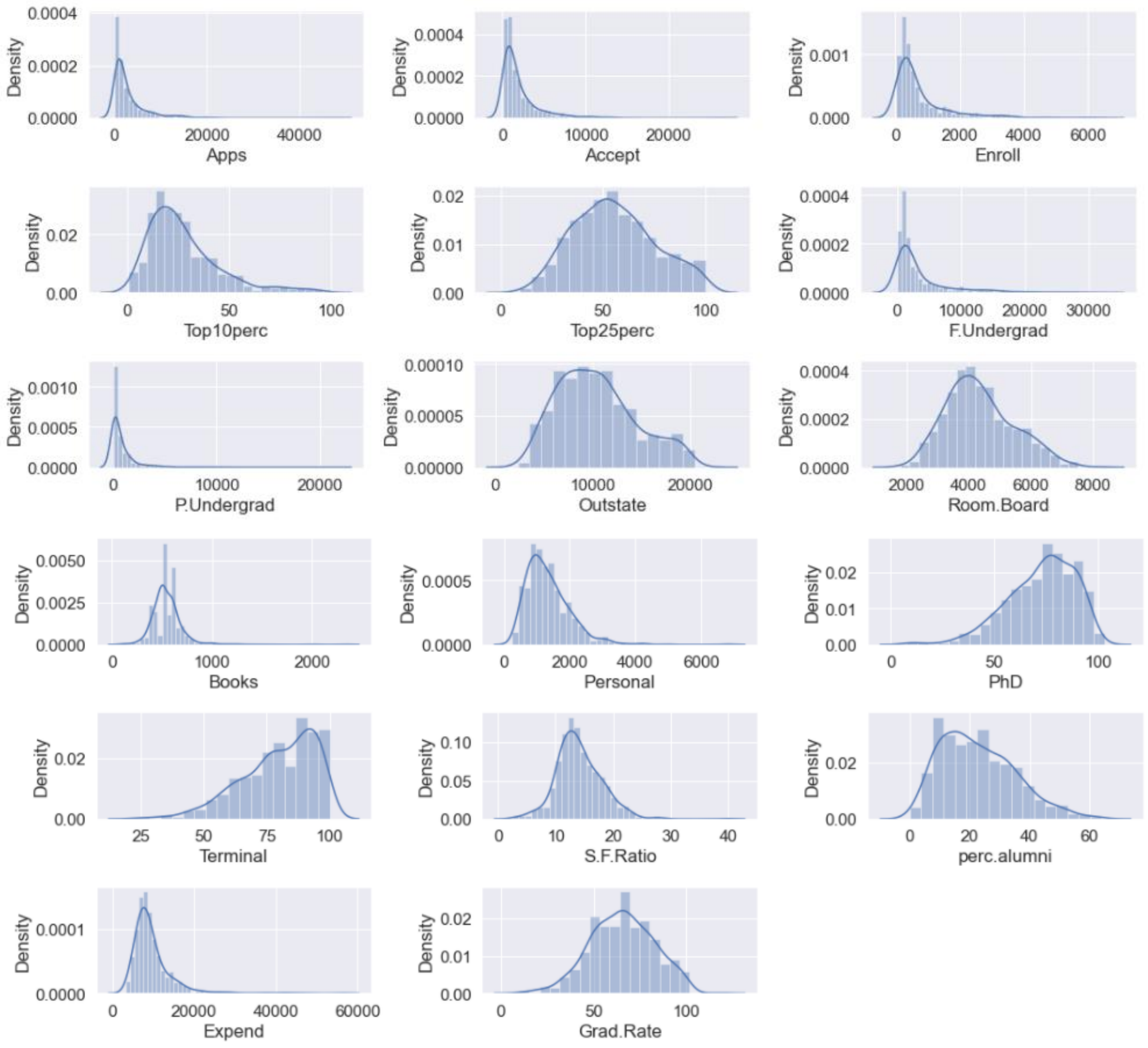


Figure 8. Histograms before scaling

Histogram after applying z-score(scaling)

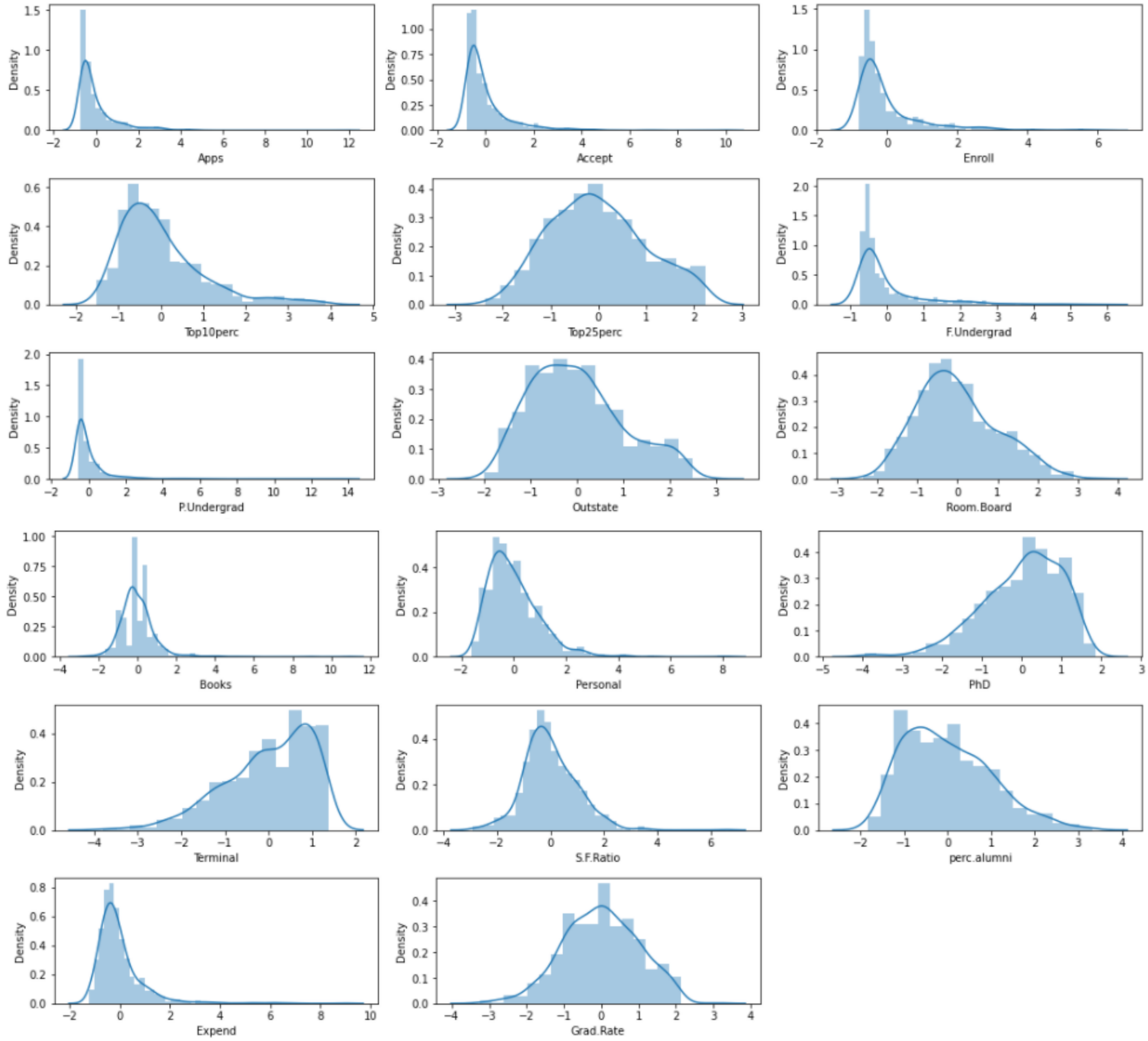
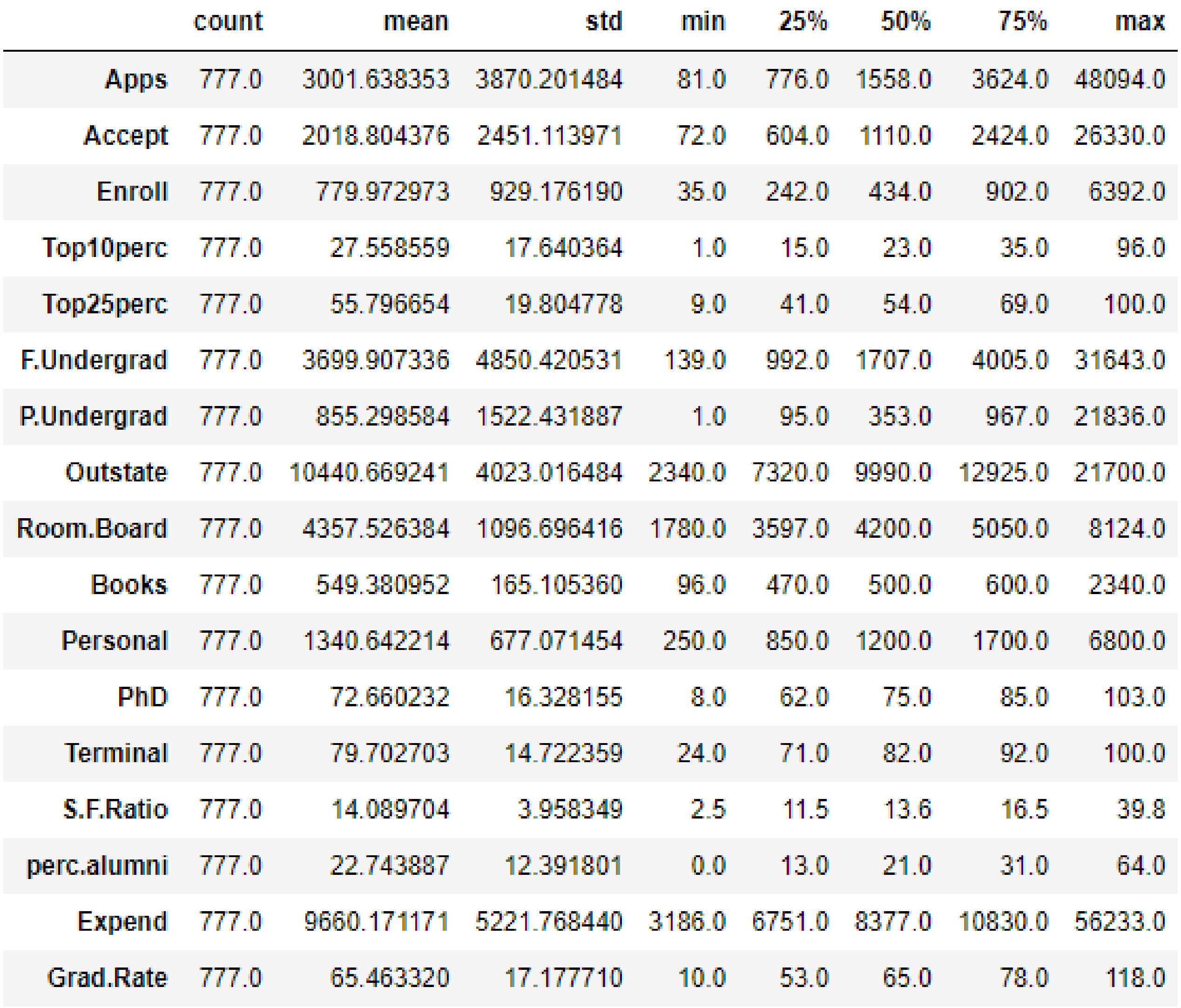


Figure 9. Histograms after scaling

Note the scaling ranges of the each variables before and after scaling. After applying the z-score on the dataset, all the variables are come nearer to the origin(it doesn't change anything in the data) & scales are all same. It is just another way of representing the dataset.

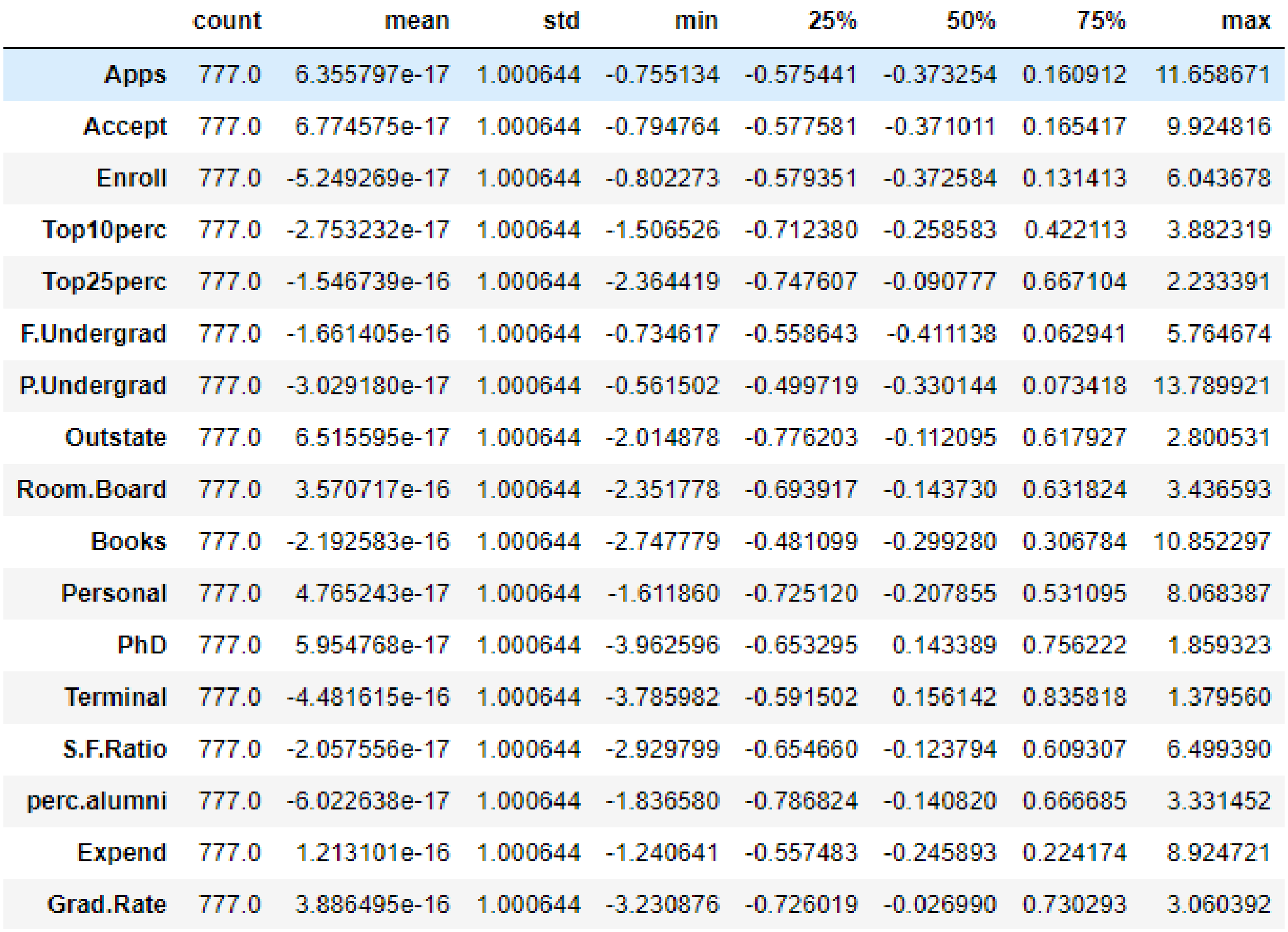
Summarize the dataset before scaling:



	count	mean	std	min	25%	50%	75%	max
Apps	777.0	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	779.972973	929.176190	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	55.796654	19.804778	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	3699.907336	4850.420531	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	855.298584	1522.431887	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	10440.669241	4023.016484	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	4357.526384	1096.696416	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	549.380952	165.105360	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	1340.642214	677.071454	250.0	850.0	1200.0	1700.0	6800.0
PhD	777.0	72.660232	16.328155	8.0	62.0	75.0	85.0	103.0
Terminal	777.0	79.702703	14.722359	24.0	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	14.089704	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777.0	22.743887	12.391801	0.0	13.0	21.0	31.0	64.0
Expend	777.0	9660.171171	5221.768440	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	65.463320	17.177710	10.0	53.0	65.0	78.0	118.0

Table 14. Summarize the Dataset before scaling

Summarize the data after scaling:

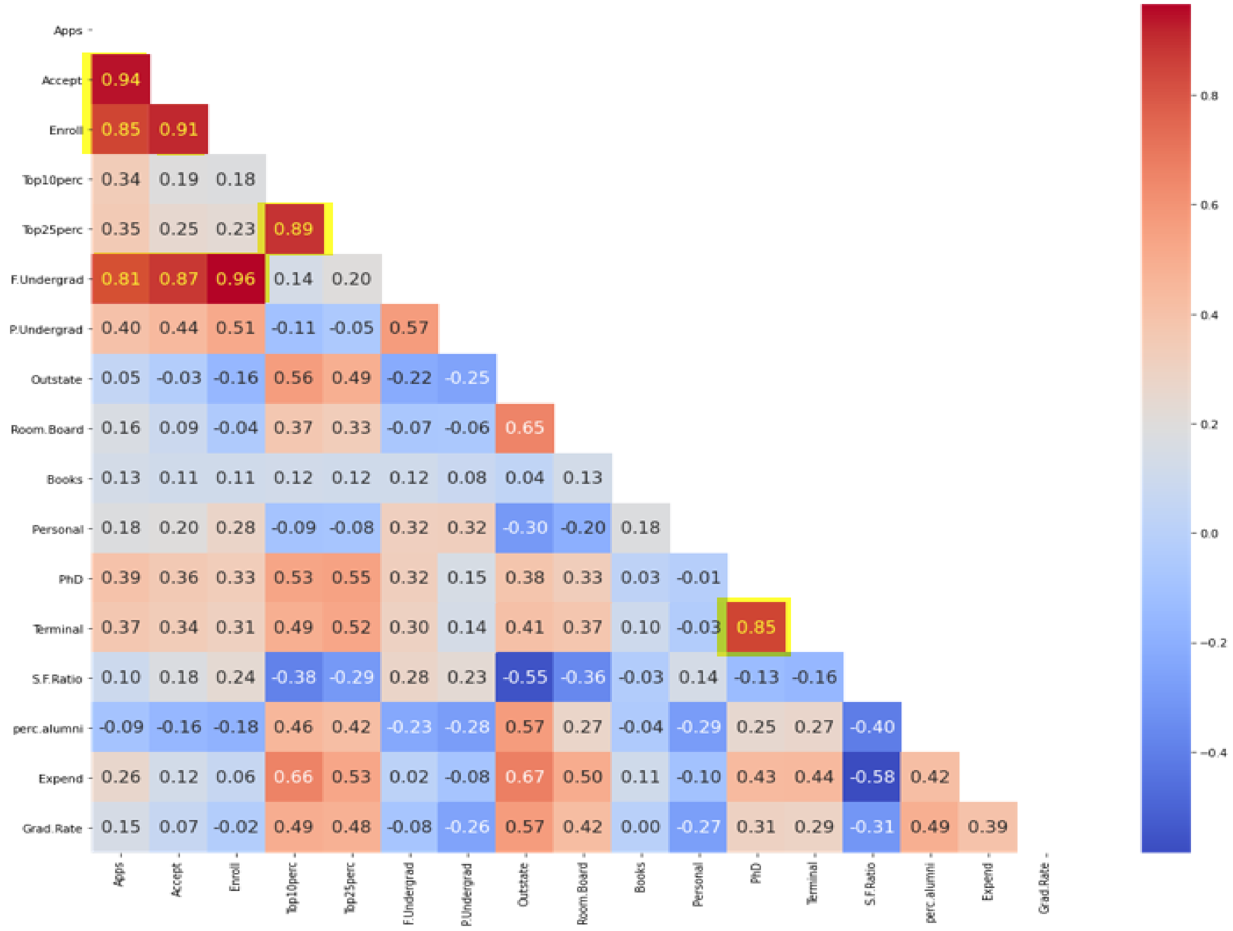


	count	mean	std	min	25%	50%	75%	max
Apps	777.0	6.355797e-17	1.000644	-0.755134	-0.575441	-0.373254	0.160912	11.658671
Accept	777.0	6.774575e-17	1.000644	-0.794764	-0.577581	-0.371011	0.165417	9.924816
Enroll	777.0	-5.249269e-17	1.000644	-0.802273	-0.579351	-0.372584	0.131413	6.043678
Top10perc	777.0	-2.753232e-17	1.000644	-1.506526	-0.712380	-0.258583	0.422113	3.882319
Top25perc	777.0	-1.546739e-16	1.000644	-2.364419	-0.747607	-0.090777	0.667104	2.233391
F.Undergrad	777.0	-1.661405e-16	1.000644	-0.734617	-0.558643	-0.411138	0.062941	5.764674
P.Undergrad	777.0	-3.029180e-17	1.000644	-0.561502	-0.499719	-0.330144	0.073418	13.789921
Outstate	777.0	6.515595e-17	1.000644	-2.014878	-0.776203	-0.112095	0.617927	2.800531
Room.Board	777.0	3.570717e-16	1.000644	-2.351778	-0.693917	-0.143730	0.631824	3.436593
Books	777.0	-2.192583e-16	1.000644	-2.747779	-0.481099	-0.299280	0.306784	10.852297
Personal	777.0	4.765243e-17	1.000644	-1.611860	-0.725120	-0.207855	0.531095	8.068387
PhD	777.0	5.954768e-17	1.000644	-3.962596	-0.653295	0.143389	0.756222	1.859323
Terminal	777.0	-4.481615e-16	1.000644	-3.785982	-0.591502	0.156142	0.835818	1.379560
S.F.Ratio	777.0	-2.057556e-17	1.000644	-2.929799	-0.654660	-0.123794	0.609307	6.499390
perc.alumni	777.0	-6.022638e-17	1.000644	-1.836580	-0.786824	-0.140820	0.666685	3.331452
Expend	777.0	1.213101e-16	1.000644	-1.240641	-0.557483	-0.245893	0.224174	8.924721
Grad.Rate	777.0	3.886495e-16	1.000644	-3.230876	-0.726019	-0.026990	0.730293	3.060392

Table 15. Summarize the Dataset after scaling

Note here the minimum, maximum value ranges and all are come nearer to the origin(especially 50%). Also, observe the standard deviation is common for all the features.

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



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps																	
Accept	0.94																
Enroll	0.85	0.91															
Top10perc	0.34	0.19	0.18														
Top25perc	0.35	0.25	0.23	0.89													
F.Undergrad	0.81	0.87	0.96	0.14	0.20												
P.Undergrad	0.40	0.44	0.51	-0.11	-0.05	0.57											
Outstate	0.05	-0.03	-0.16	0.56	0.49	-0.22	-0.25										
Room.Board	0.16	0.09	-0.04	0.37	0.33	-0.07	-0.06	0.65									
Books	0.13	0.11	0.11	0.12	0.12	0.12	0.08	0.04	0.13								
Personal	0.18	0.20	0.28	-0.09	-0.08	0.32	0.32	-0.30	-0.20	0.18							
PhD	0.39	0.36	0.33	0.53	0.55	0.32	0.15	0.38	0.33	0.03	-0.01						
Terminal	0.37	0.34	0.31	0.49	0.52	0.30	0.14	0.41	0.37	0.10	-0.03	0.85					
S.F.Ratio	0.10	0.18	0.24	-0.38	-0.29	0.28	0.23	-0.55	-0.36	-0.03	0.14	-0.13	-0.16				
perc.alumni	-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	-0.40			
Expend	0.26	0.12	0.06	0.66	0.53	0.02	-0.08	0.67	0.50	0.11	-0.10	0.43	0.44	-0.58	0.42		
Grad.Rate	0.15	0.07	-0.02	0.49	0.48	-0.08	-0.26	0.57	0.42	0.00	-0.27	0.31	0.29	-0.31	0.49	0.39	

Figure 10. Heat map to check the correlation of the data post scaling

It is important to understand here that correlation remains same similar to what we observed in before scaling as well. Correlation is nothing but to study the strength of the relationship between the two features(variables). The value of correlation is bound on the upper by +1 and on the lower side by -1. Thus, it is a definite range. Whereas, Covariance signifies the direction(vectors) of the linear relationship between the two variables. The range of covariance is indefinite. It can take any positive value or any negative value (theoretically, the range is -∞ to +∞). If the value is positive, it has the positive correlation with other variable and if the value is negative, it has the negative correlation with the other variable.

Strong correlation

From the correlation heat map, we can see that attributes of the colleges are highly correlated in many places. You could see it in brown turn into dark brown area.

- Application received and Applications Accepted are very highly correlated.
- Applications Accepted and students enrollment are very highly correlated.
- Top25perc and Top10perc are very highly correlated.
- Full Time Undergraduate students and students enrollment are very highly correlated.
- Application received and students enrollment are highly correlated.
- Percentage of PhD and Terminal faculties are highly correlated.
- Full Time Undergraduate students and Application received/Applications Accepted are highly correlated.

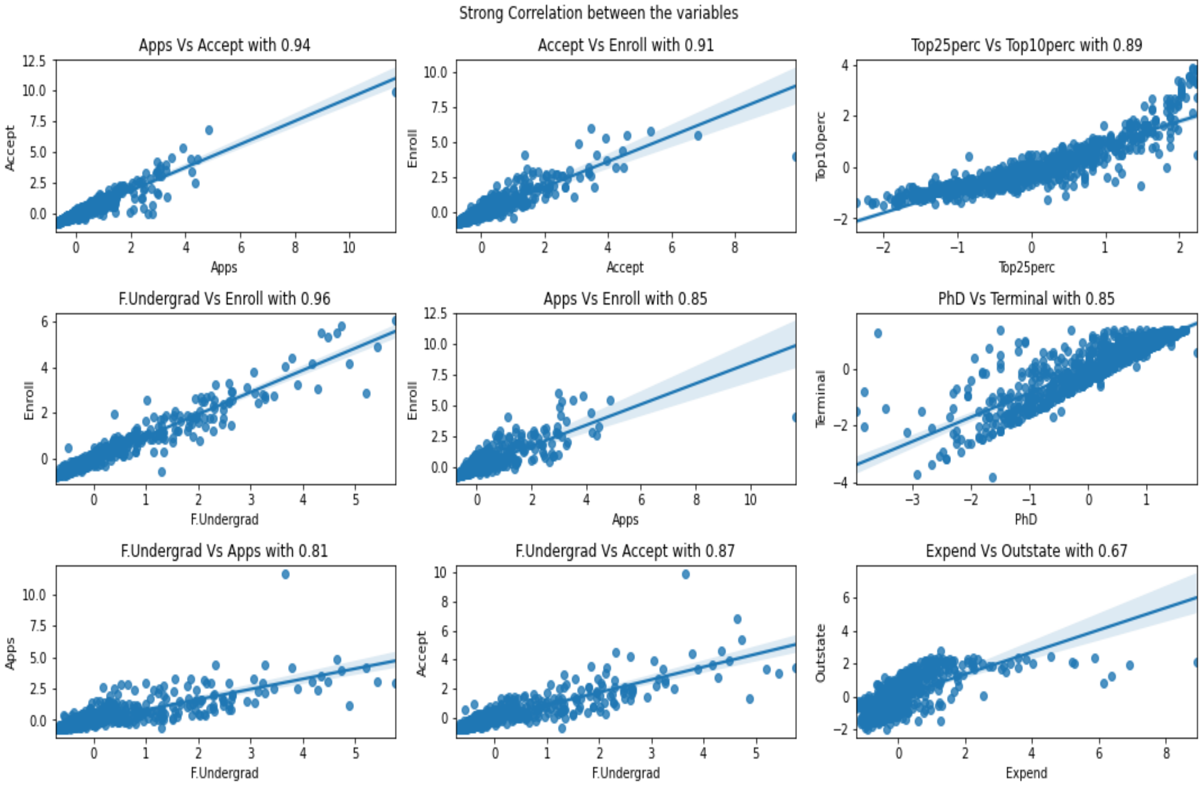


Figure 11. Strong correlation representation between variables using regplot

Weak Correlation

From the correlation heat map, we can see that some of the attributes of the colleges are weakly correlated in few places. You could see it in blue square area. Basically, **Student Faculty Ratio feature is weakly correlated with the other variables** such as Expend, Outstate, Percentage of Alumini, Percentage of new students from top 10% of Higher Secondary class, Cost of Room.Board and Graduation rate are weakly correlated.

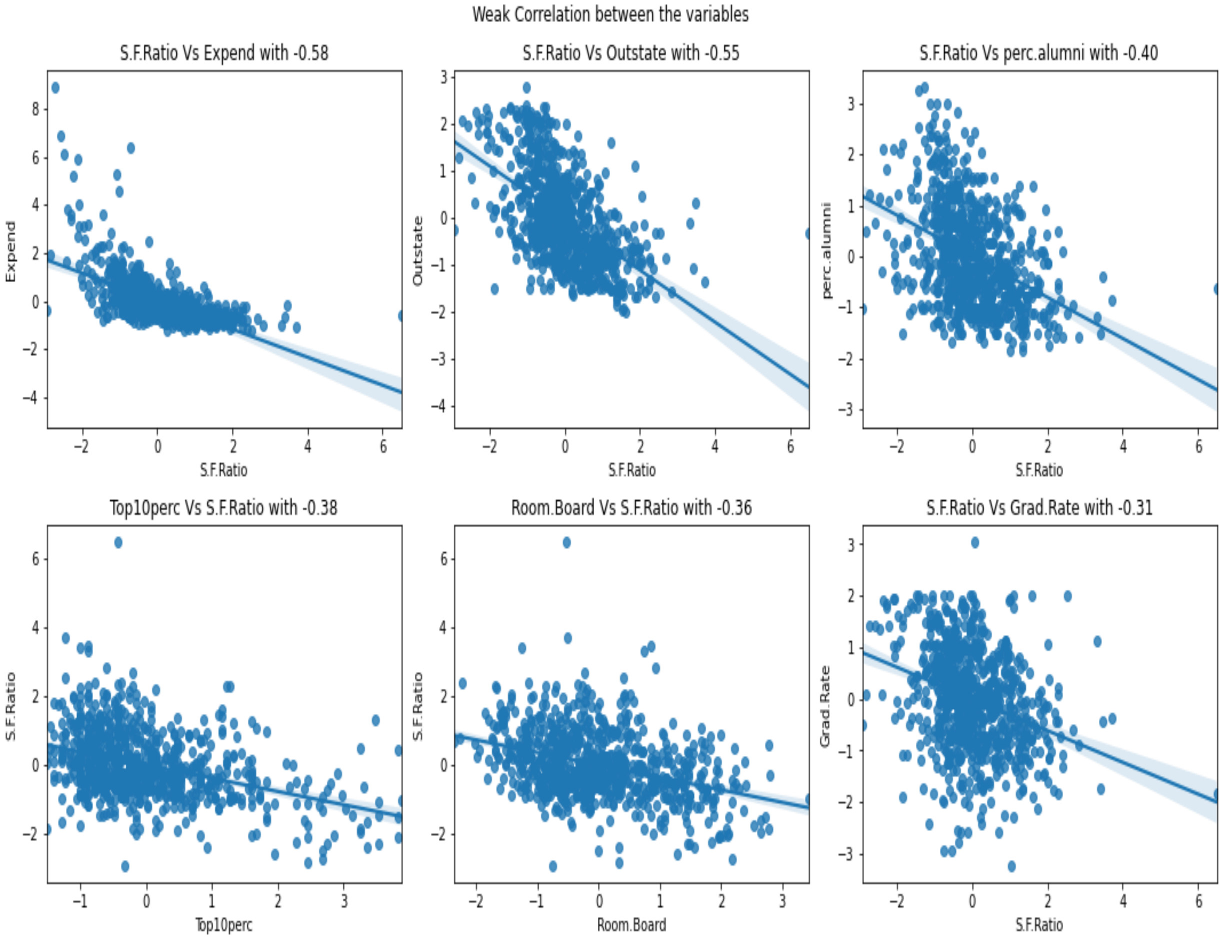


Figure 12. Weak correlation representation between variables using regplot

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

Using Boxplot is the simplest and easiest technique to identify the outliers present in the variable for the entire dataset.

Outliers before scaling

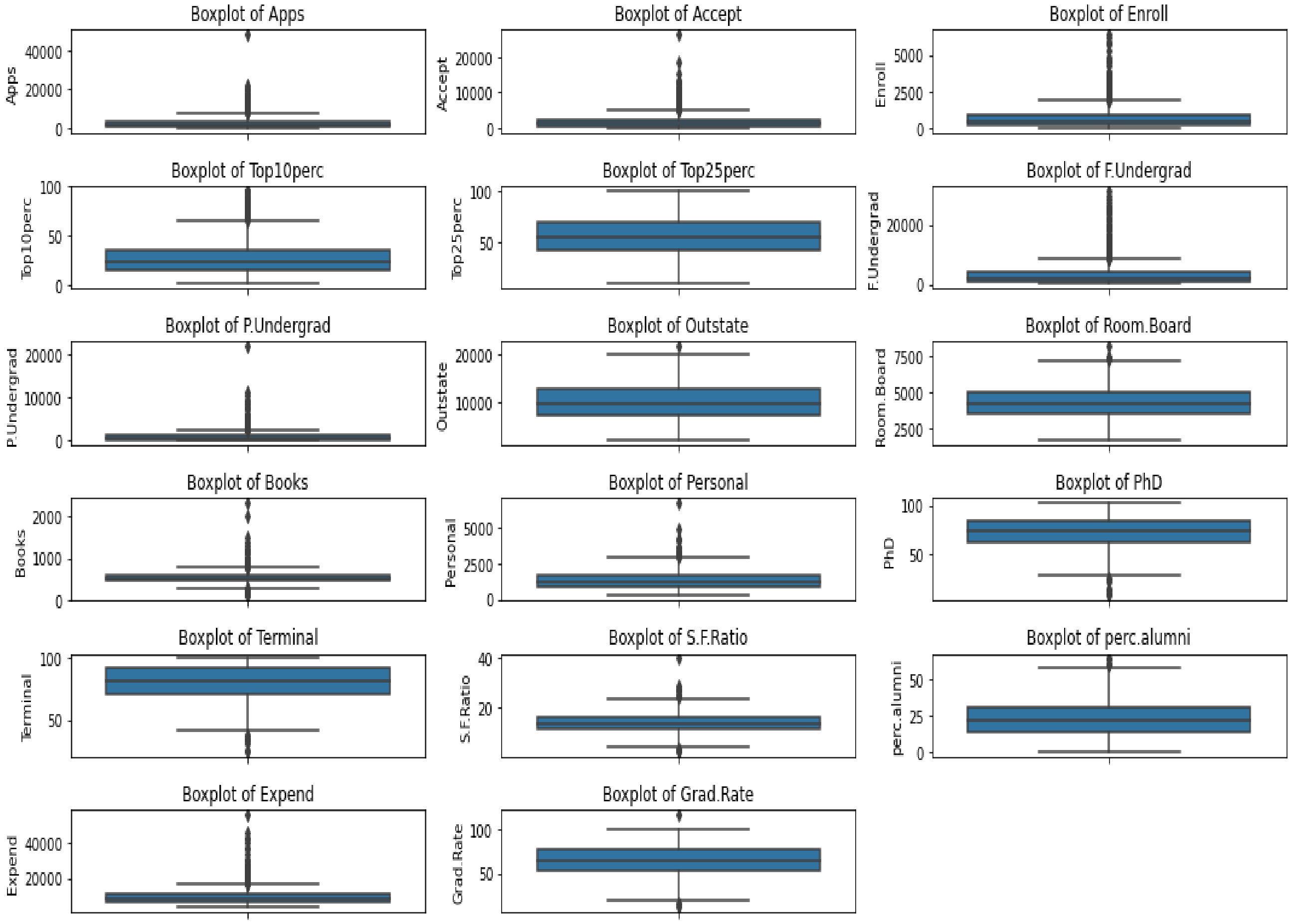


Figure 13. Boxplot to identify the outliers for the original dataset(before scaling)

Except Top25perc variable, all the other variables contains the outliers.

Outliers after scaling

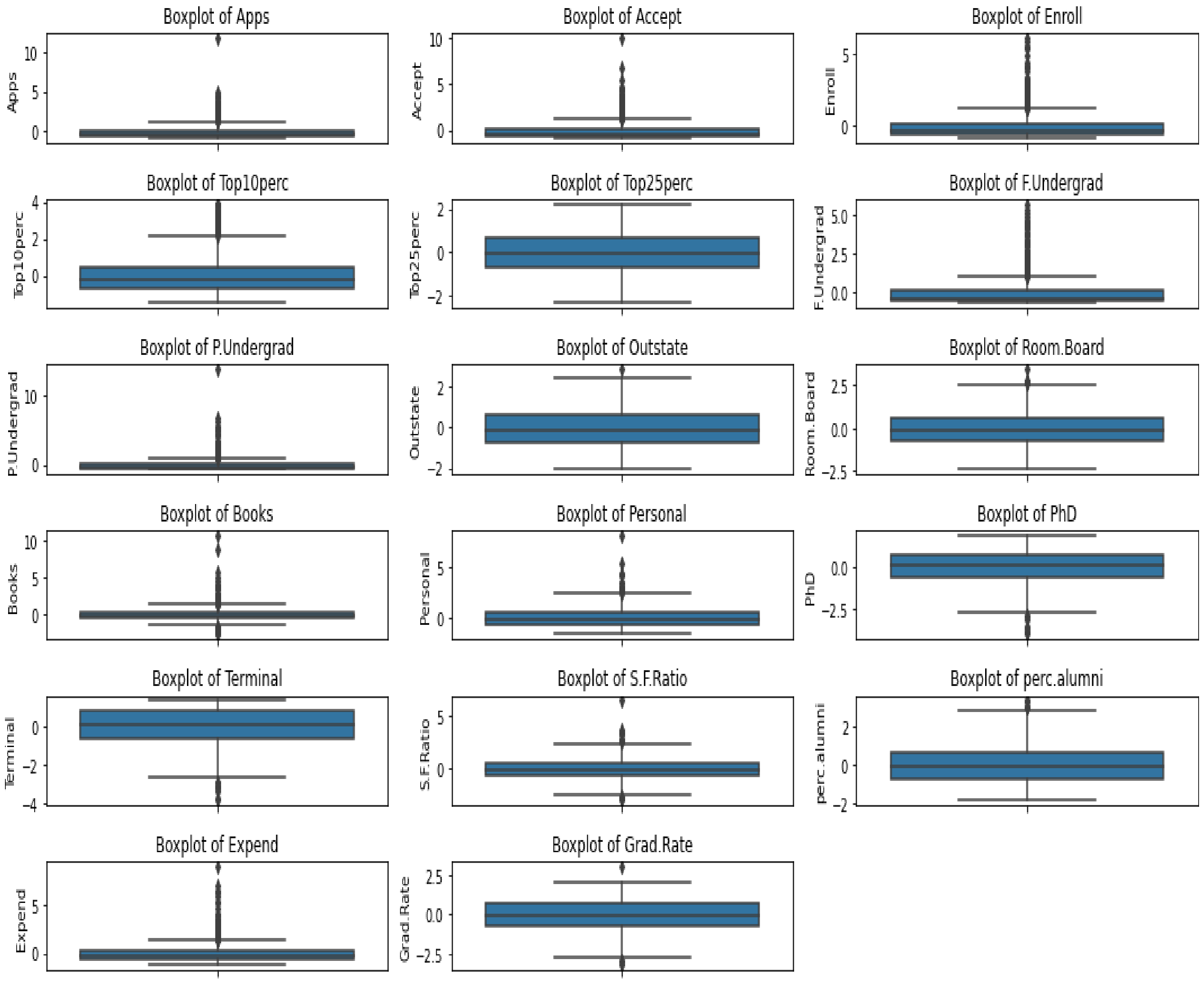
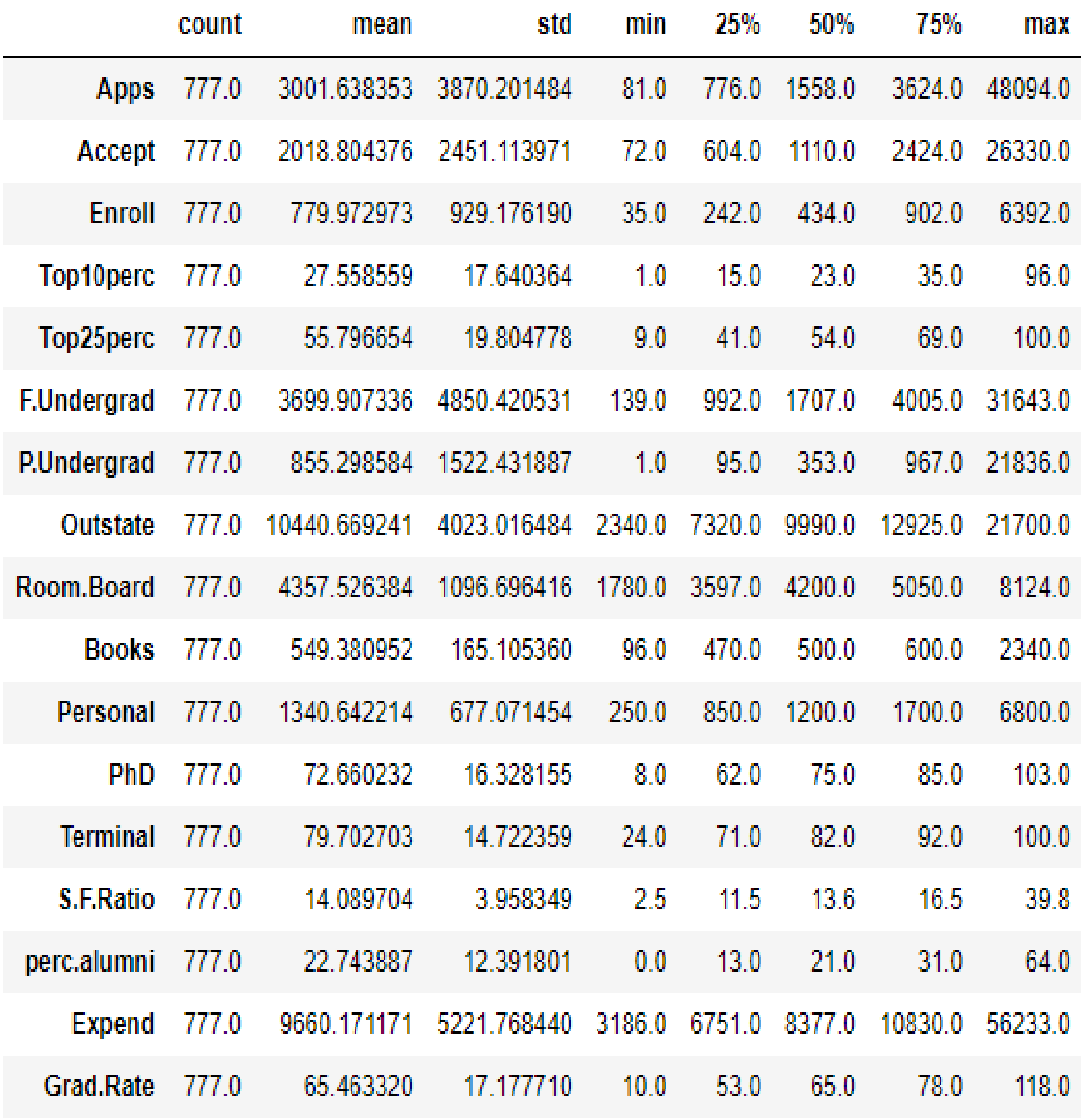


Figure 14. Boxplot to identify the outliers after applying z-score(after scaling)

Even after applying the z-score for the dataset, we don't see any difference in the outliers. With respect to outliers, before and after scaling, it looks alike. However, scaling has got reduced significantly in the post scaling representation, which tells the data comes around the origin.

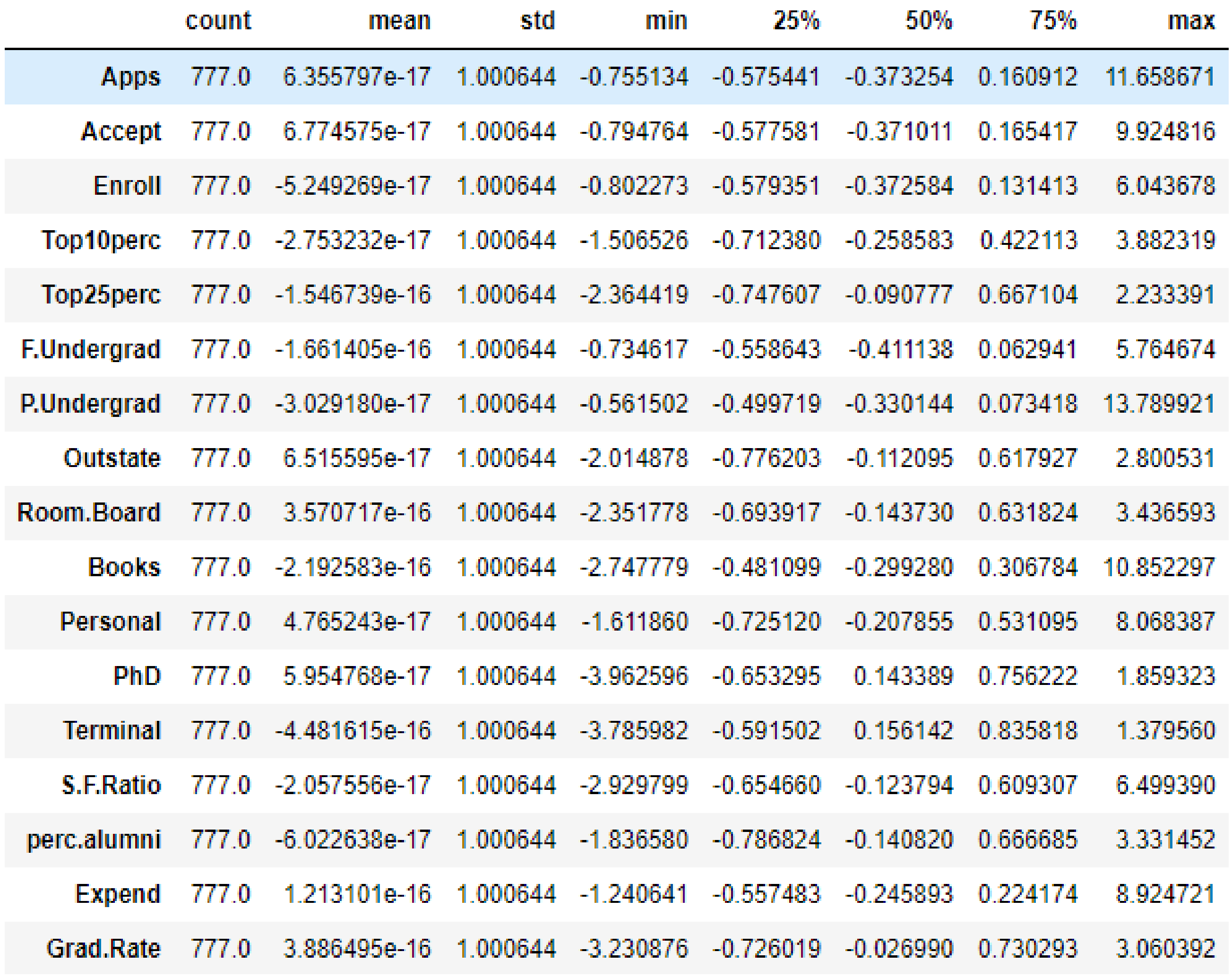
Summarize the data using the original dataset:



	count	mean	std	min	25%	50%	75%	max
Apps	777.0	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	779.972973	929.176190	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	55.796654	19.804778	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	3699.907336	4850.420531	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	855.298584	1522.431887	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	10440.669241	4023.016484	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	4357.526384	1096.696416	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	549.380952	165.105360	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	1340.642214	677.071454	250.0	850.0	1200.0	1700.0	6800.0
PhD	777.0	72.660232	16.328155	8.0	62.0	75.0	85.0	103.0
Terminal	777.0	79.702703	14.722359	24.0	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	14.089704	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777.0	22.743887	12.391801	0.0	13.0	21.0	31.0	64.0
Expend	777.0	9660.171171	5221.768440	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	65.463320	17.177710	10.0	53.0	65.0	78.0	118.0

Table 16. Summarize the Dataset before scaling

Summarize the data after scaling:



	count	mean	std	min	25%	50%	75%	max
Apps	777.0	6.355797e-17	1.000644	-0.755134	-0.575441	-0.373254	0.160912	11.658671
Accept	777.0	6.774575e-17	1.000644	-0.794764	-0.577581	-0.371011	0.165417	9.924816
Enroll	777.0	-5.249269e-17	1.000644	-0.802273	-0.579351	-0.372584	0.131413	6.043678
Top10perc	777.0	-2.753232e-17	1.000644	-1.506526	-0.712380	-0.258583	0.422113	3.882319
Top25perc	777.0	-1.546739e-16	1.000644	-2.364419	-0.747607	-0.090777	0.667104	2.233391
F.Undergrad	777.0	-1.661405e-16	1.000644	-0.734617	-0.558643	-0.411138	0.062941	5.764674
P.Undergrad	777.0	-3.029180e-17	1.000644	-0.561502	-0.499719	-0.330144	0.073418	13.789921
Outstate	777.0	6.515595e-17	1.000644	-2.014878	-0.776203	-0.112095	0.617927	2.800531
Room.Board	777.0	3.570717e-16	1.000644	-2.351778	-0.693917	-0.143730	0.631824	3.436593
Books	777.0	-2.192583e-16	1.000644	-2.747779	-0.481099	-0.299280	0.306784	10.852297
Personal	777.0	4.765243e-17	1.000644	-1.611860	-0.725120	-0.207855	0.531095	8.068387
PhD	777.0	5.954768e-17	1.000644	-3.962596	-0.653295	0.143389	0.756222	1.859323
Terminal	777.0	-4.481615e-16	1.000644	-3.785982	-0.591502	0.156142	0.835818	1.379560
S.F.Ratio	777.0	-2.057556e-17	1.000644	-2.929799	-0.654660	-0.123794	0.609307	6.499390
perc.alumni	777.0	-6.022638e-17	1.000644	-1.836580	-0.786824	-0.140820	0.666685	3.331452
Expend	777.0	1.213101e-16	1.000644	-1.240641	-0.557483	-0.245893	0.224174	8.924721
Grad.Rate	777.0	3.886495e-16	1.000644	-3.230876	-0.726019	-0.026990	0.730293	3.060392

Table 17. Summarize the Dataset after scaling

Standardizing the data doesn't have any impact on outliers rather all the data's comes around the origin (doesn't normalize the data). It is just an another way of representing the dataset.

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

Before decomposing the Eigen values and vectors, it is necessary to do run the Bartlett Sphericity and KMO test.

Bartlett Sphericity test is to validate the correlations significantly present in the dataset or not. Here in this case.

H_0: Correlations are not significant.

H_A: There are significant correlations

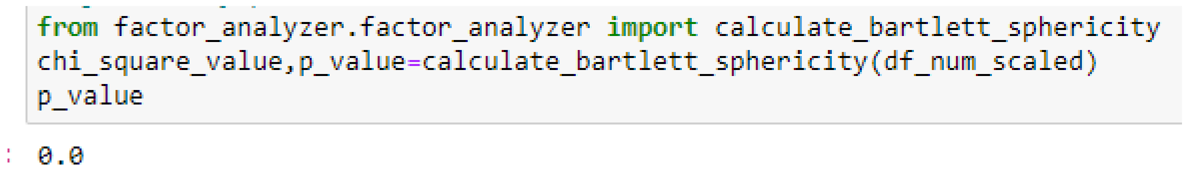
If the p-value is < 0.05, then we reject the H_0.

If the p-value is >= 0.05, then we fail to reject the H_0, H_A is considered.

Where 0.05 represents the Alpha(significance level)

H_0 – Null Hypothesis

H_A – Alternate Hypothesis



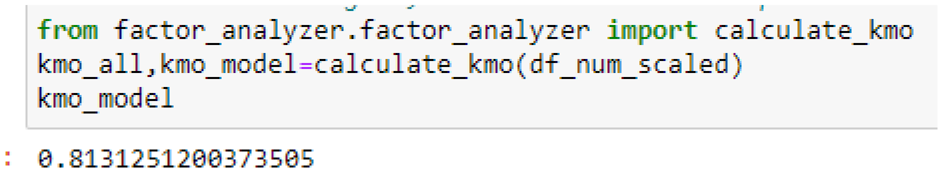
```
from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
chi_square_value,p_value=calculate_bartlett_sphericity(df_num_scaled)
p_value
```

```
: 0.0
```

Figure 15. Bartlett Sphericity test

After running the test, p-value is 0.0 indicates that there are significant correlations present in the dataset.

Next step is to run the KMO test to validate the sample size in the dataset. It is good to proceed fur ther if the resultant is above 0.5.



```
from factor_analyzer.factor_analyzer import calculate_kmo
kmo_all,kmo_model=calculate_kmo(df_num_scaled)
kmo_model
```

```
: 0.8131251200373505
```

Figure 16. Kaiser-Meyer-Olkin test

In our case, kmo_model result is `0.8131251200373505`

Both of our tests are results out passed as expected to proceed further with PCA analysis.

Eigen vectors

There are 17 numerical features present in the dataset, which tells us there will be 17 principal components.

```
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]])
```

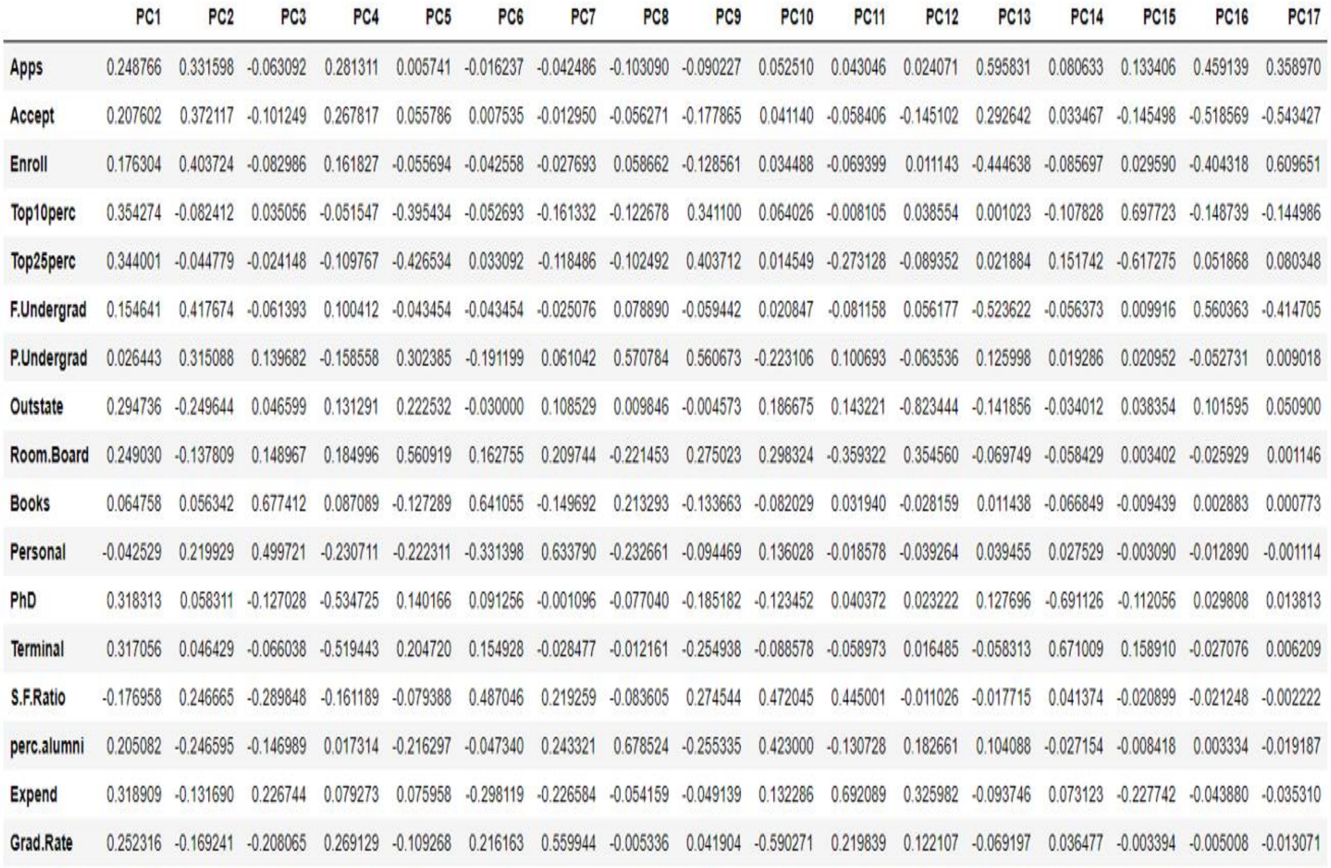
Eigen values

This is always returned in the descending order.

```
array([5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123,
       0.84849117, 0.6057878 , 0.58787222, 0.53061262, 0.4043029 ,
       0.31344588, 0.22061096, 0.16779415, 0.1439785 , 0.08802464,
       0.03672545, 0.02302787])
```

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

Below table shows all the 17 principal components with original features.



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	-0.042486	-0.103090	-0.090227	0.052510	0.043046	0.024071	0.595831	0.080633	0.133406	0.459139	0.358970
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	-0.012950	-0.056271	-0.177865	0.041140	-0.058406	-0.145102	0.292642	0.033467	-0.145498	-0.518569	-0.543427
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	-0.027693	0.058662	-0.128561	0.034488	-0.069399	0.011143	-0.444638	-0.085697	0.029590	-0.404318	0.609651
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	-0.161332	-0.122678	0.341100	0.064026	-0.008105	0.038554	0.001023	-0.107828	0.697723	-0.148739	-0.144986
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	-0.118486	-0.102492	0.403712	0.014549	-0.273128	-0.089352	0.021884	0.151742	-0.617275	0.051868	0.080348
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454	-0.025076	0.078890	-0.059442	0.020847	-0.081158	0.056177	-0.523622	-0.056373	0.009916	0.560363	-0.414705
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199	0.061042	0.570784	0.560673	-0.223106	0.100693	-0.063536	0.125998	0.019286	0.020952	-0.052731	0.009018
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000	0.108529	0.009846	-0.004573	0.186675	0.143221	-0.823444	-0.141856	-0.034012	0.038354	0.101595	0.050900
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755	0.209744	-0.221453	0.275023	0.298324	-0.359322	0.354560	-0.069749	-0.058429	0.003402	-0.025929	0.001146
Books	0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055	-0.149692	0.213293	-0.133663	-0.082029	0.031940	-0.028159	0.011438	-0.066849	-0.009439	0.002883	0.000773
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398	0.633790	-0.232661	-0.094469	0.136028	-0.018578	-0.039264	0.039455	0.027529	-0.003090	-0.012890	-0.001114
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256	-0.001096	-0.077040	-0.185182	-0.123452	0.040372	0.023222	0.127696	-0.691126	-0.112056	0.029808	0.013813
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928	-0.028477	-0.012161	-0.254938	-0.088578	-0.058973	0.016485	-0.058313	0.671009	0.158910	-0.027076	0.006209
S.F.Ratio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046	0.219259	-0.083605	0.274544	0.472045	0.445001	-0.011026	-0.017715	0.041374	-0.020899	-0.021248	-0.002222
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340	0.243321	0.678524	-0.255335	0.423000	-0.130728	0.182661	0.104088	-0.027154	-0.008418	0.003334	-0.019187
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119	-0.226584	-0.054159	-0.049139	0.132286	0.692089	0.325982	-0.093746	0.073123	-0.227742	-0.043880	-0.035310
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163	0.559944	-0.005336	0.041904	-0.590271	0.219839	0.122107	-0.069197	0.036477	-0.003394	-0.005008	-0.013071

Table 18. Principal Component Dataset with original features

Check the cumulative explained variance ratio to find a cut off for selecting the number of PCs

```
array([0.32020628, 0.58360843, 0.65261759, 0.71184748, 0.76673154,
       0.81657854, 0.85216726, 0.88670347, 0.91787581, 0.94162773,
       0.96004199, 0.9730024 , 0.98285994, 0.99131837, 0.99648962,
       0.99864716, 1.        ])
```

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.

In the above array we see that the first feature explains 32% of the variance within our data set while the first two components explains 58.4% and so on. If we employ 7 features we capture ~ 85% 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). With just 7 features, we are able to achieve 85% variation. So, from 17 components, it is reduced to 7 components.

Scree plot

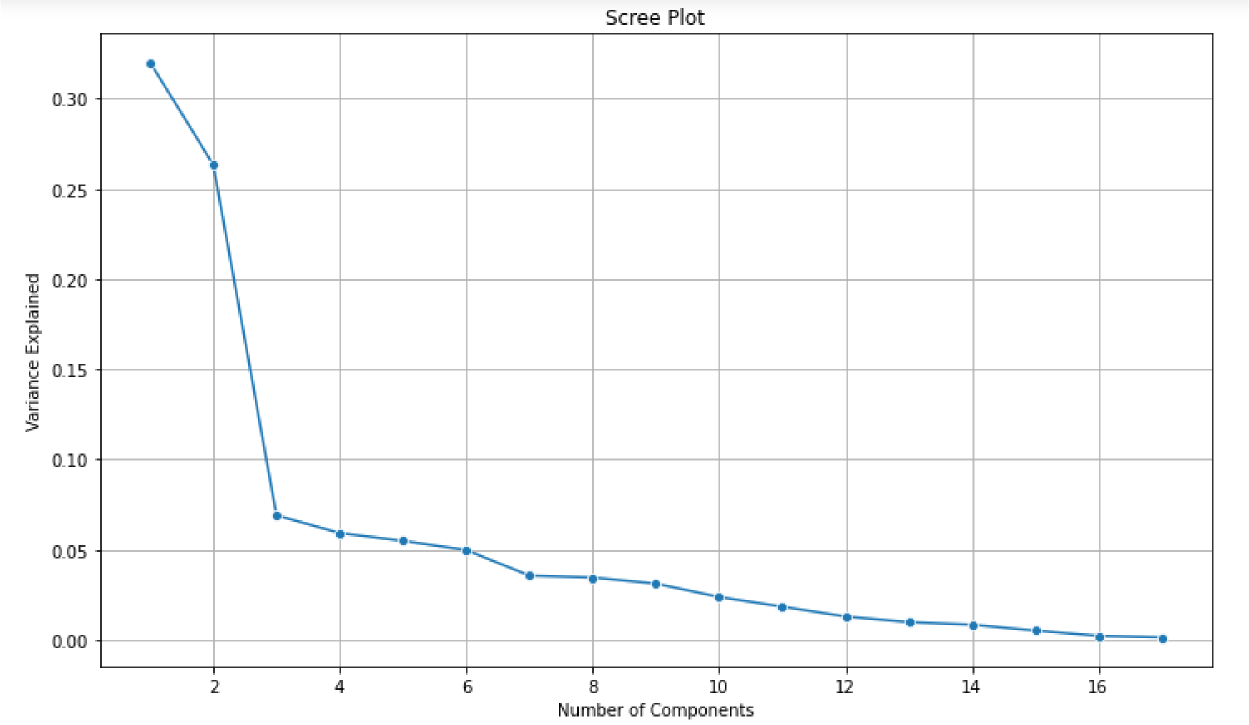


Figure 17. Screeplot to identify the principal components

- Visually we can observe that there is steep drop in variance explained with increase in number of PC's.
- Depending on the requirement 80% variation with 6 components or 7 components will also do good. But, here we will proceed with 7 principal components here with 85% variance.

Cumulative Explained Variance Vs. number of features

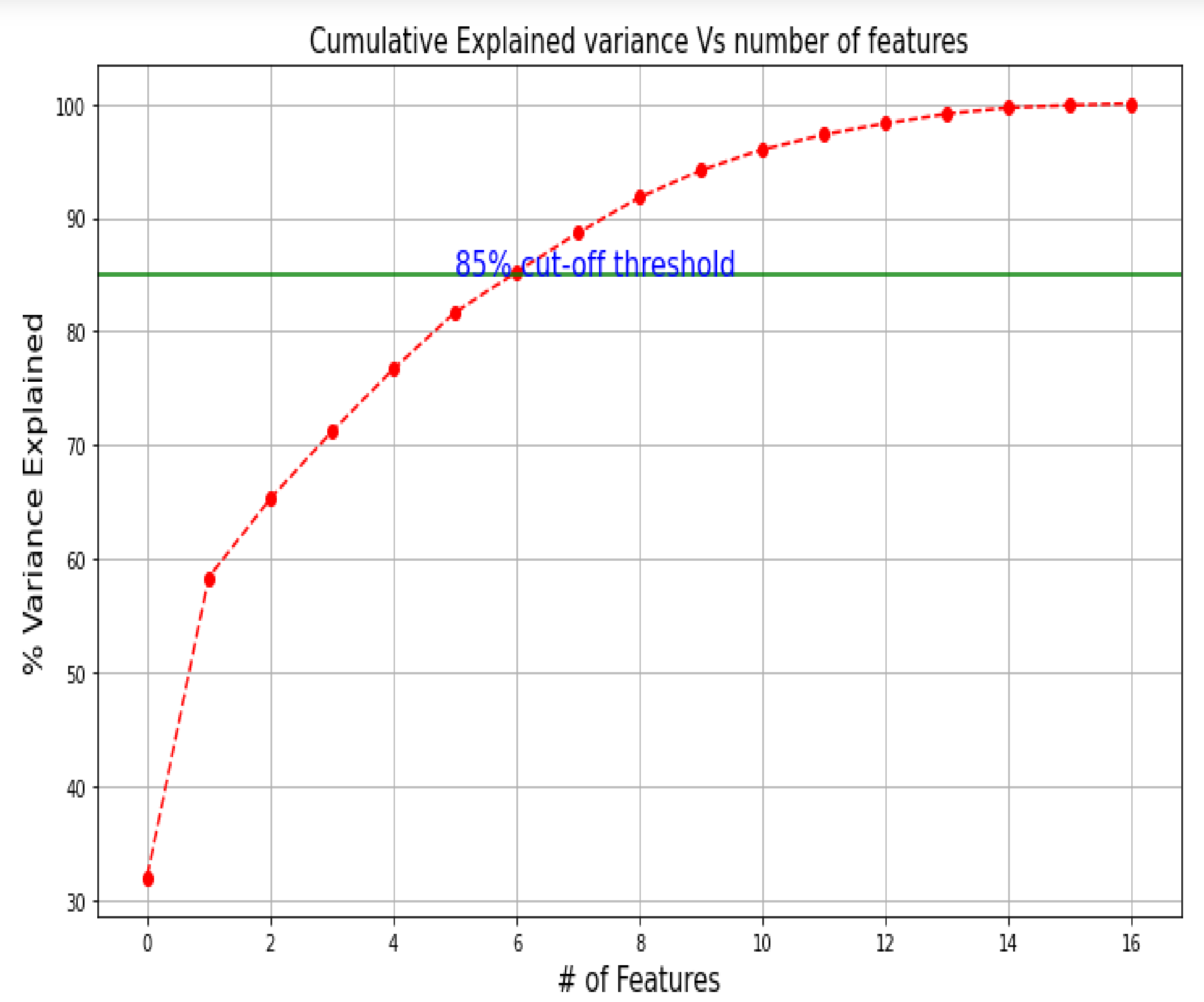


Figure 18. Plot to identify the number of components needed to explain variance

We can see from the above plot that indicates clearly with threshold cut-off at 85% requires 7 principal components to perform PCA.

Absolute loadings of PCs

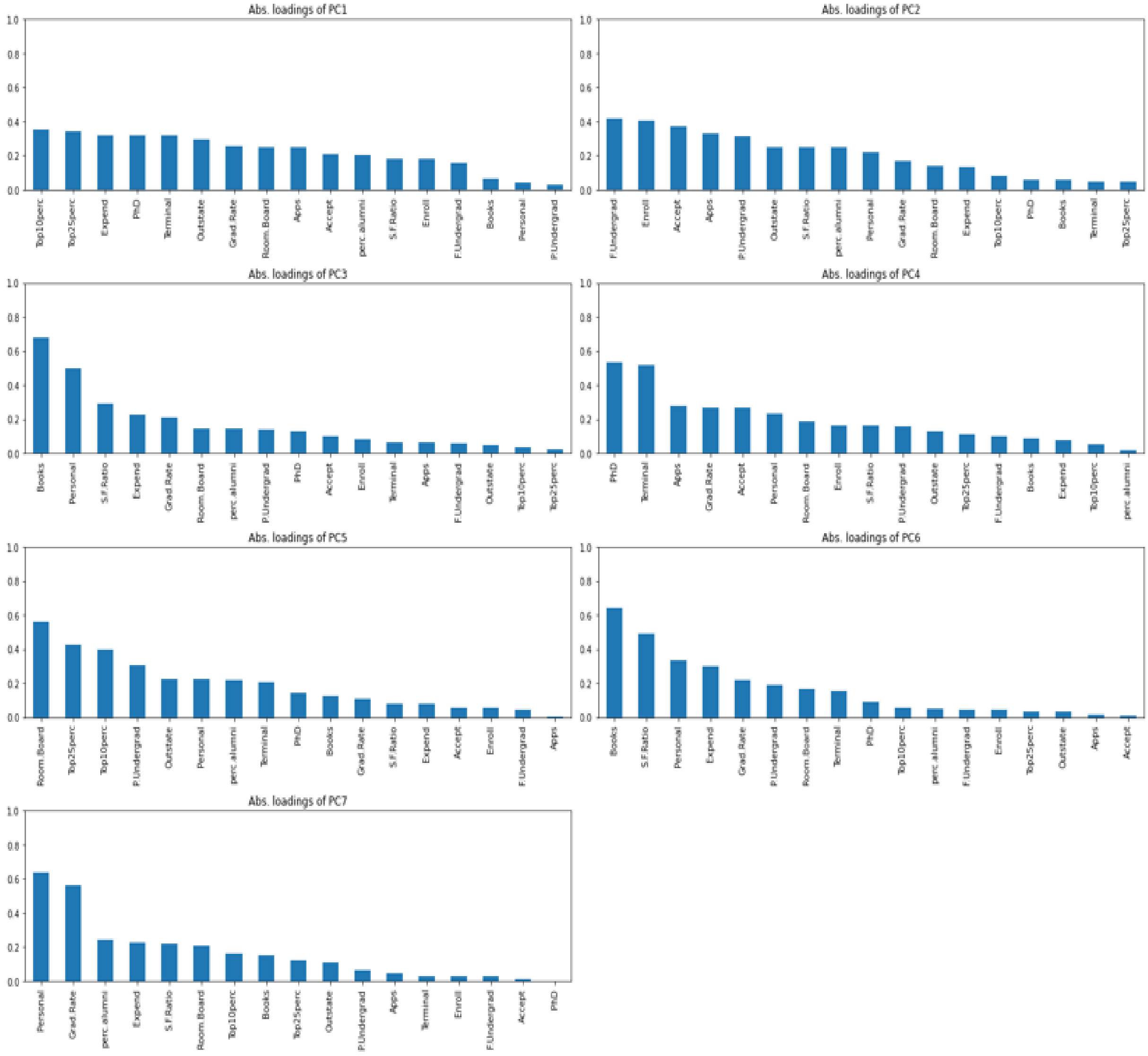
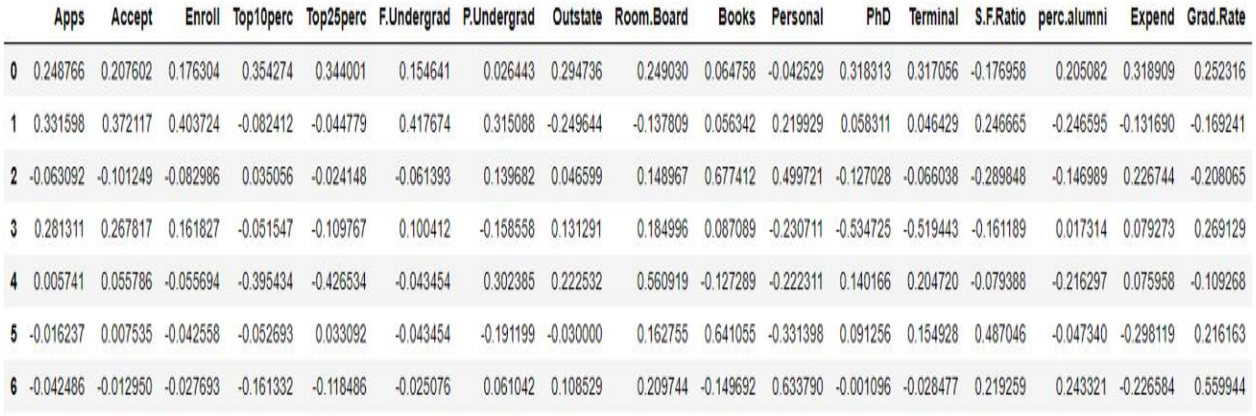


Figure 19.Representation of each principal components

The above plots shows that how the original features matter to each Principal Component. Also, note that we are considering only the absolute values. The variances of the cumulative sum of selected number of components(i.e., 7 components) is 85% as shown below.

```
array([32.02, 58.36, 65.26, 71.18, 76.67, 81.65, 85.21])
```

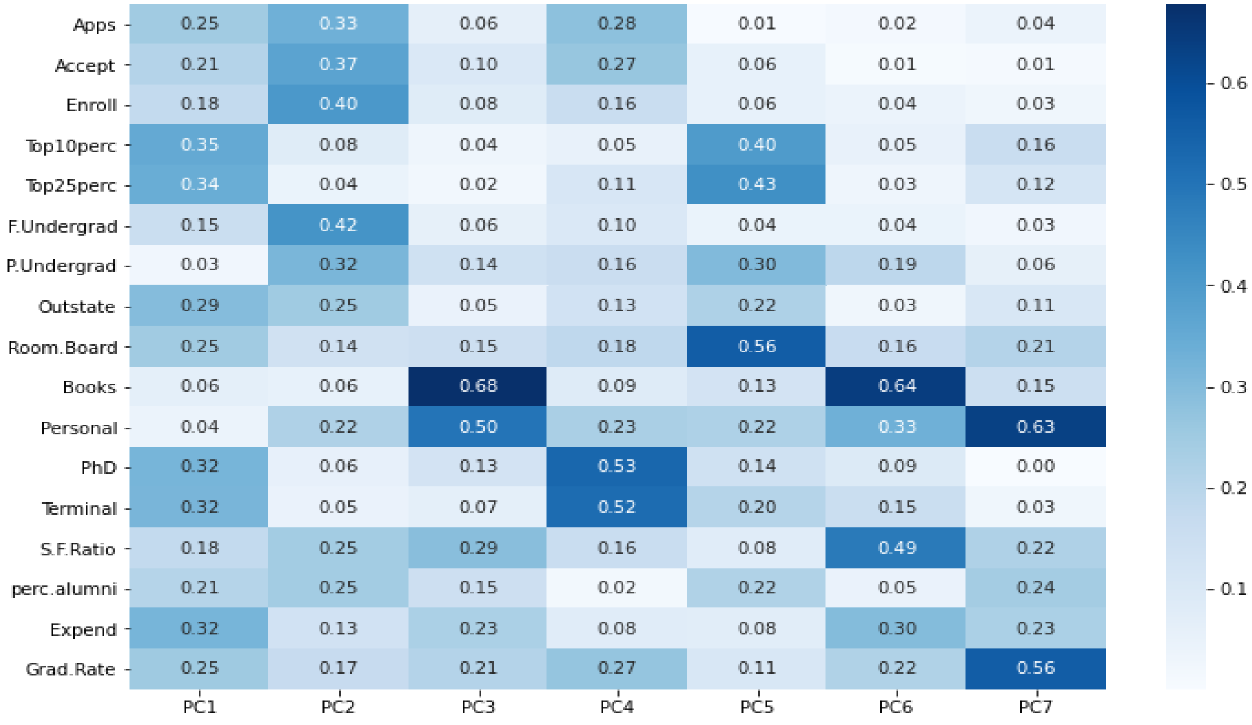
Correlation between components and features



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	0.248766	0.207602	0.176304	0.354274	0.344001	0.154641	0.026443	0.294736	0.249030	0.064758	-0.042529	0.318313	0.317056	-0.176958	0.205082	0.318909	0.252316
1	0.331598	0.372117	0.403724	-0.082412	-0.044779	0.417674	0.315088	-0.249644	-0.137809	0.056342	0.219929	0.058311	0.046429	0.246665	-0.246595	-0.131690	-0.169241
2	-0.063092	-0.101249	-0.082986	0.035056	-0.024148	-0.061393	0.139682	0.046599	0.148967	0.677412	0.499721	-0.127028	-0.066038	-0.289848	-0.146989	0.226744	-0.208065
3	0.281311	0.267817	0.161827	-0.051547	-0.109767	0.100412	-0.158558	0.131291	0.184996	0.087089	-0.230711	-0.534725	-0.519443	-0.161189	0.017314	0.079273	0.269129
4	0.005741	0.055786	-0.055694	-0.395434	-0.426534	-0.043454	0.302385	0.222532	0.560919	-0.127289	-0.222311	0.140166	0.204720	-0.079388	-0.216297	0.075958	-0.109268
5	-0.016237	0.007535	-0.042558	-0.052693	0.033092	-0.043454	-0.191199	-0.030000	0.162755	0.641055	-0.331398	0.091256	0.154928	0.487046	-0.047340	-0.298119	0.216163
6	-0.042486	-0.012950	-0.027693	-0.161332	-0.118486	-0.025076	0.061042	0.108529	0.209744	-0.149692	0.633790	-0.001096	-0.028477	0.219259	0.243321	-0.226584	0.559944

Table 19. Correlation between PC's and features

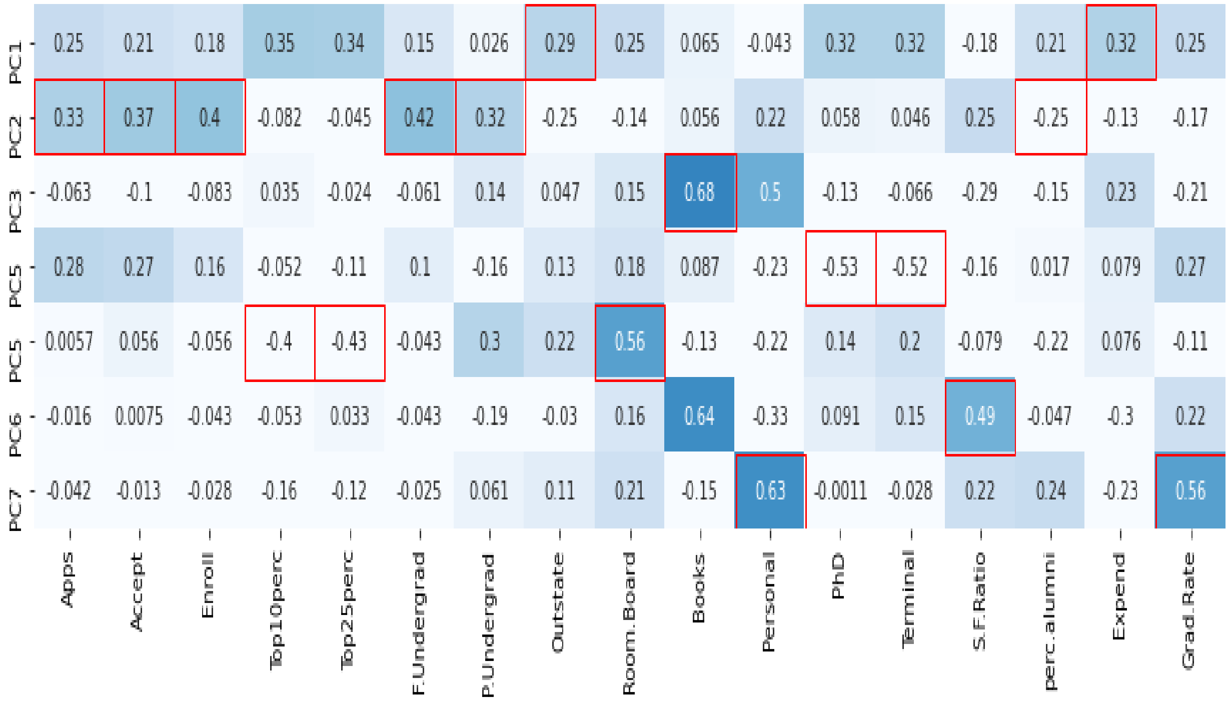
Heat map to represent original features influence various PCs



	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Apps	0.25	0.33	0.06	0.28	0.01	0.02	0.04
Accept	0.21	0.37	0.10	0.27	0.06	0.01	0.01
Enroll	0.18	0.40	0.08	0.16	0.06	0.04	0.03
Top10perc	0.35	0.08	0.04	0.05	0.40	0.05	0.16
Top25perc	0.34	0.04	0.02	0.11	0.43	0.03	0.12
F.Undergrad	0.15	0.42	0.06	0.10	0.04	0.04	0.03
P.Undergrad	0.03	0.32	0.14	0.16	0.30	0.19	0.06
Outstate	0.29	0.25	0.05	0.13	0.22	0.03	0.11
Room.Board	0.25	0.14	0.15	0.18	0.56	0.16	0.21
Books	0.06	0.06	0.68	0.09	0.13	0.64	0.15
Personal	0.04	0.22	0.50	0.23	0.22	0.33	0.63
PhD	0.32	0.06	0.13	0.53	0.14	0.09	0.00
Terminal	0.32	0.05	0.07	0.52	0.20	0.15	0.03
S.F.Ratio	0.18	0.25	0.29	0.16	0.08	0.49	0.22
perc.alumni	0.21	0.25	0.15	0.02	0.22	0.05	0.24
Expend	0.32	0.13	0.23	0.08	0.08	0.30	0.23
Grad.Rate	0.25	0.17	0.21	0.27	0.11	0.22	0.56

Figure 20. Heatmap to represent original features influence various PCs

Heat map to show correlation between features & PCs itself



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
PC1	0.25	0.21	0.18	0.35	0.34	0.15	0.026	0.29	0.25	0.065	-0.043	0.32	0.32	-0.18	0.21	0.32	0.25
PC2	0.33	0.37	0.4	-0.082	-0.045	0.42	0.32	-0.25	-0.14	0.056	0.22	0.058	0.046	0.25	-0.25	-0.13	-0.17
PC3	-0.063	-0.1	-0.083	0.035	-0.024	-0.061	0.14	0.047	0.15	0.68	0.5	-0.13	-0.066	-0.29	-0.15	0.23	-0.21
PC5	0.28	0.27	0.16	-0.052	-0.11	0.1	-0.16	0.13	0.18	0.087	-0.23	-0.53	-0.52	-0.16	0.017	0.079	0.27
PC5	0.0057	0.056	-0.056	-0.4	-0.43	-0.043	0.3	0.22	0.56	-0.13	-0.22	0.14	0.2	-0.079	-0.22	0.076	-0.11
PC6	-0.016	0.0075	-0.043	-0.053	0.033	-0.043	-0.19	-0.03	0.16	0.64	-0.33	0.091	0.15	0.49	-0.047	-0.3	0.22
PC7	-0.042	-0.013	-0.028	-0.16	-0.12	-0.025	0.061	0.11	0.21	-0.15	0.63	-0.0011	-0.028	0.22	0.24	-0.23	0.56

Figure 21.Heatmap to show correlation between features & PCs itself

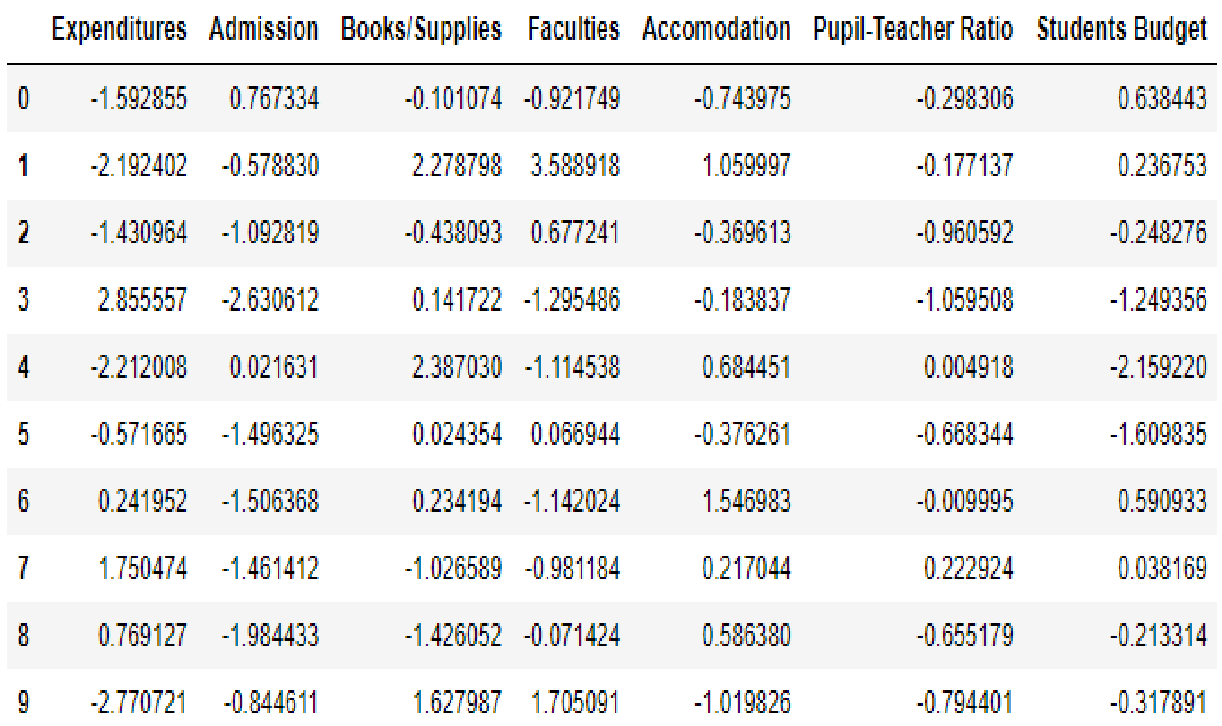
Inferences from the above Heat map

- This heat map and the color bar basically represents the correlation between the various features and the principal components itself.
- Depending on the relationship, we could go ahead and label relationship with features.
- Note, the red square blocks are based on the absolute values among the PC's.

Final PCA table with new labels

Based on the analysis, renamed the reduced principal components as follows.

PC1	Outstate, Expend	Expenditures
PC2	Apps, Accept, Enroll, F.Undergrad, P.Undergrad, perc.alumini	Admission
PC3	Books	Books/Supplies
PC4	PhD, Terminal	Faculties
PC5	Top10perc, Top25perc, Room.Board	Accommodation
PC6	S.F.Ratio	Pupil-Teacher Ratio
PC7	Personal, Grad. Rate	Students Budget



	Expenditures	Admission	Books/Supplies	Faculties	Accomodation	Pupil-Teacher Ratio	Students Budget
0	-1.592855	0.767334	-0.101074	-0.921749	-0.743975	-0.298306	0.638443
1	-2.192402	-0.578830	2.278798	3.588918	1.059997	-0.177137	0.236753
2	-1.430964	-1.092819	-0.438093	0.677241	-0.369613	-0.960592	-0.248276
3	2.855557	-2.630612	0.141722	-1.295486	-0.183837	-1.059508	-1.249356
4	-2.212008	0.021631	2.387030	-1.114538	0.684451	0.004918	-2.159220
5	-0.571665	-1.496325	0.024354	0.066944	-0.376261	-0.668344	-1.609835
6	0.241952	-1.506368	0.234194	-1.142024	1.546983	-0.009995	0.590933
7	1.750474	-1.461412	-1.026589	-0.981184	0.217044	0.222924	0.038169
8	0.769127	-1.984433	-1.426052	-0.071424	0.586380	-0.655179	-0.213314
9	-2.770721	-0.844611	1.627987	1.705091	-1.019826	-0.794401	-0.317891

Table 20. Final PCA dataset

Heat map to show correlations among PCs

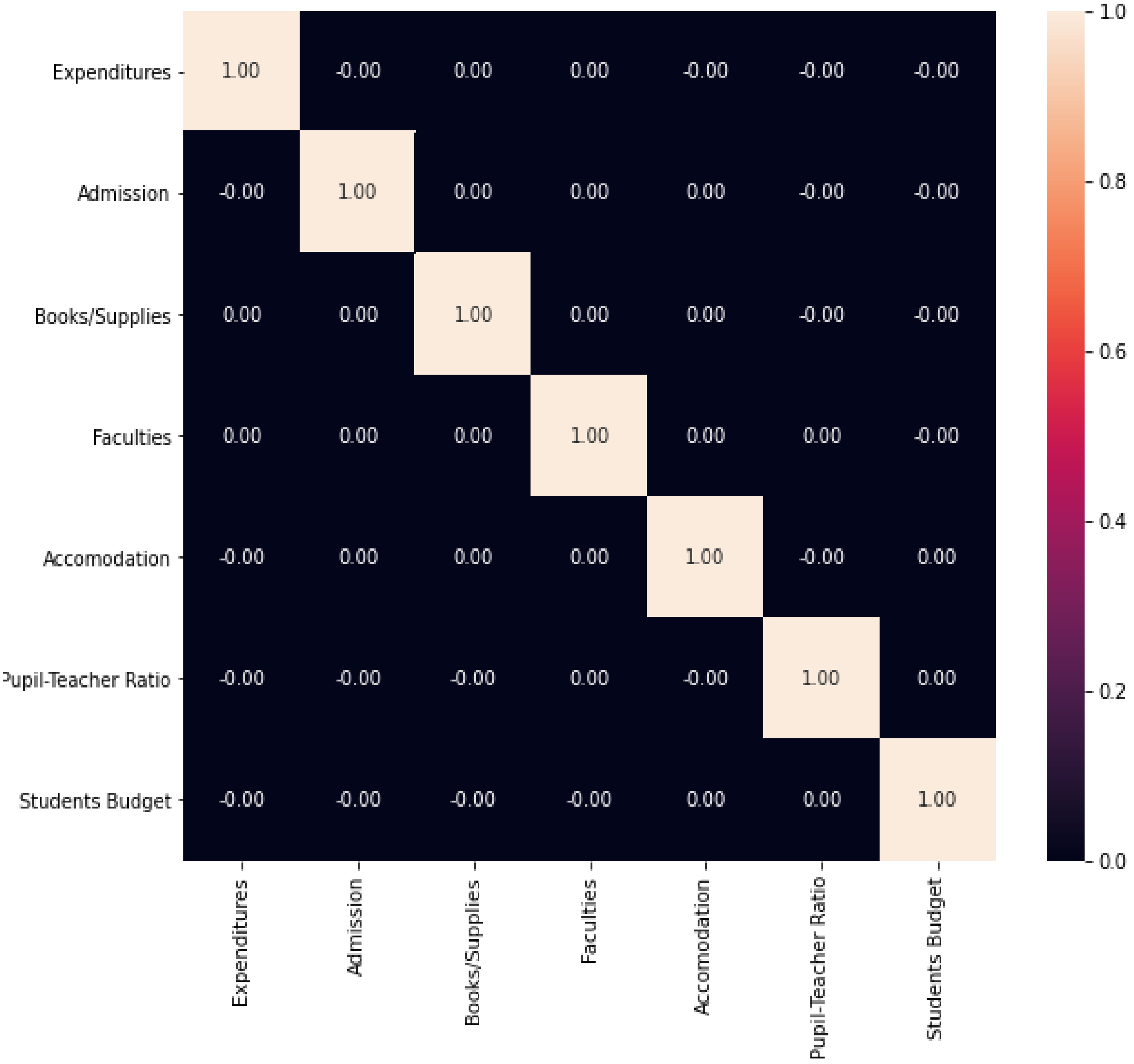


Figure 21.Heatmap to show correlations among PCs

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]

We can summarize the calculation of the covariance matrix via the following matrix equation:

$$\Sigma = \frac{1}{n-1}\left((\mathbf{X}-\bar{\mathbf{x}})^T\,(\mathbf{X}-\bar{\mathbf{x}})\right)$$

where $\bar{\mathbf{X}}$ is mean vector $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i.$

First of all, we need to calculate the covariance matrix. From the transposed covariance matrix (*orthogonal to each other*), we can decompose Eigen values and Eigen vectors using linear algebraic Eigen calculation.

To calculate PC1

PC1 = a1X1+a2X2+a3X3+a4X4+….anXn

X is the sample variables

a is the weightage(Eigen vectors)

PC1 is the 1st principal component.

The explicit form of the first principal component in terms of Eigen vectors with just 2 decimals

```
[-0.25  0.33  0.06 -0.28  0.01  0.02  0.04  0.1   0.09 -0.05  0.36
 -0.46  0.04 -0.13  0.08 -0.6   0.02]
```

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?

Eigen values are

```
array([5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123,
       0.84849117, 0.6057878 , 0.58787222, 0.53061262, 0.4043029 ,
       0.02302787, 0.03672545, 0.31344588, 0.08802464, 0.1439785 ,
       0.16779415, 0.22061096])
```

Cumulative variance explained is calculated = Eigen value/total number of principal components.

In our dataset, original principal components is 17 because we have 17 features. We need to sort the values in descending order before calculating the cumulative variance.

```
[5.450521622150291, 4.483606861940844, 1.1746676129474873, 1.008205729
969502, 0.9342312255505789, 0.8484911715044989, 0.6057878032793995, 0.
5878722195930828, 0.5306126247005801, 0.40430289775168937, 0.313445879
8102977, 0.2206109646163884, 0.1677941521658084, 0.14397849747566185,
0.08802463699454269, 0.03672544741045179, 0.02302786863372936]
```

For example, let's take the first value 5.450521622150291/ 17 = 0.32058823529411764705882352941176. Let's multiply the result with 100 to get the variance in percentage and it will be 32.02%. Similarly, we can calculate for other values and results are shown below.

Cumulative Variance Explained

```
 [ 32.0206282   58.36084263  65.26175919  71.18474841  76.67315352
  81.65785448  85.21672597  88.67034731  91.78758099  94.16277251
  96.00419883  97.30024023  98.28599436  99.13183669  99.64896227
  99.86471628 100.        ]
```

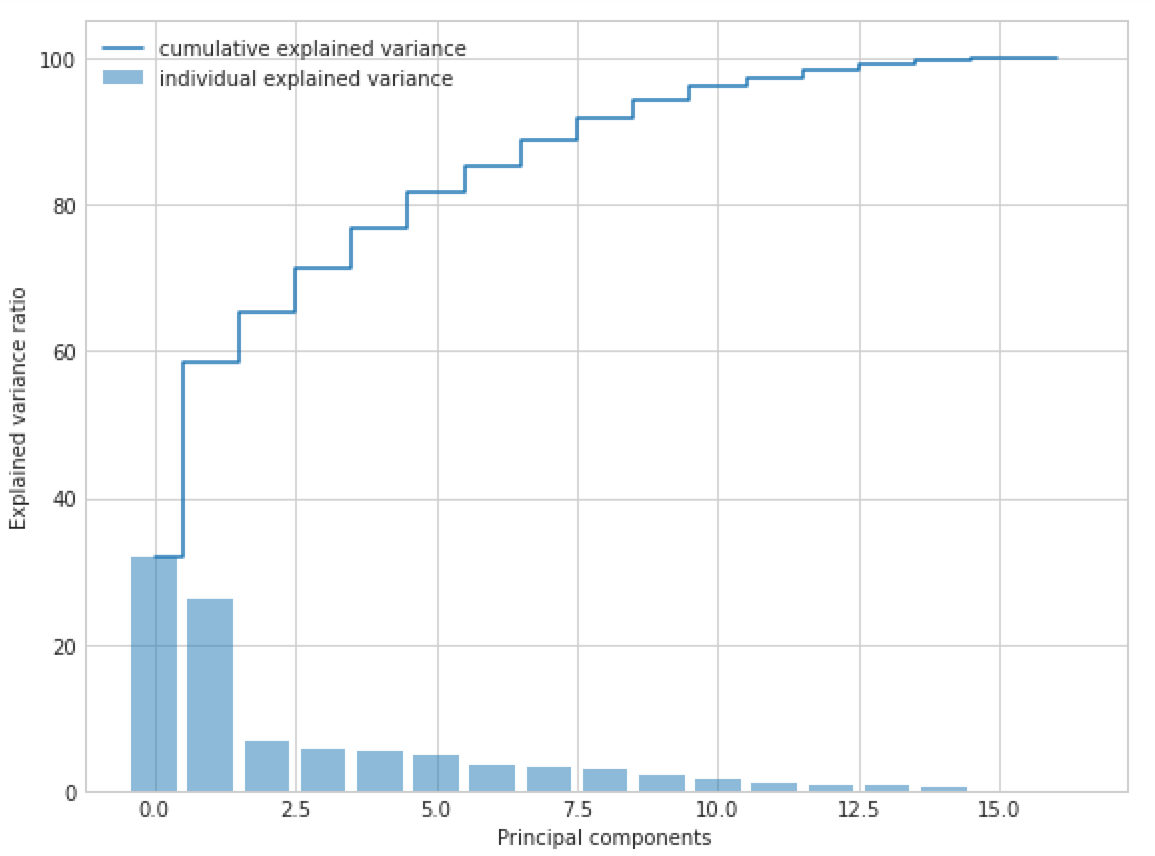


Figure 22.Plot to identify number of PCs

The plot above clearly shows that most of the variance (32.02% + 26.34%= 58.36% *of the variance to be precise*) can be explained by the 1st and 2nd principal components. If we employ 6 f eatures we capture ~81% of the variance within the dataset, thus it helps us to decide the number of principal components. But, in our case, we conclude with 7 principal components. i.e., with ~85% variance, while the other principal components can safely be dropped without losing to much information.

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]

PCA stands for Principal Component Analysis. It is one of the most popular multivariate statistical technique and used mainly for three reasons.

- Treats the multi correlation of the data.
- Dimensionality reduction technique.
- Transform the original dataset(orthogonal 90 degree to each other).

PCA relies on averages, so whatever tests depends on averages must not have outliers present in the dataset. Since the question was mentioned not to treat the outliers. We did not perform the outlier treatment steps.

Otherwise, following steps must be performed for PCA

- Remove outliers
- Standardize the dataset(apply z-score)
- Conduct p-test(Bartlett Sphericity)
- Conduct KMO test to see adequate sample size present in the dataset.
- See the correlations between the features using the heat map.
- Apply PCA.
- Cut-off dimensions.
- Label the components using suitable naming convention.
- Plot the Heat map using the final principal components, confirm the multi-collinearity of the data is removed.

In our Education - Post 12th Standard.csv dataset of various universities/colleges in USA. Initially, there are 17 numerical columns to determine the characteristics of the college. But, after applying the PCA, just with 7 features that explains over 85% of variance is sufficient to understand the universities.

With the PC1 and PC2, it explains the variance of 58.36%.

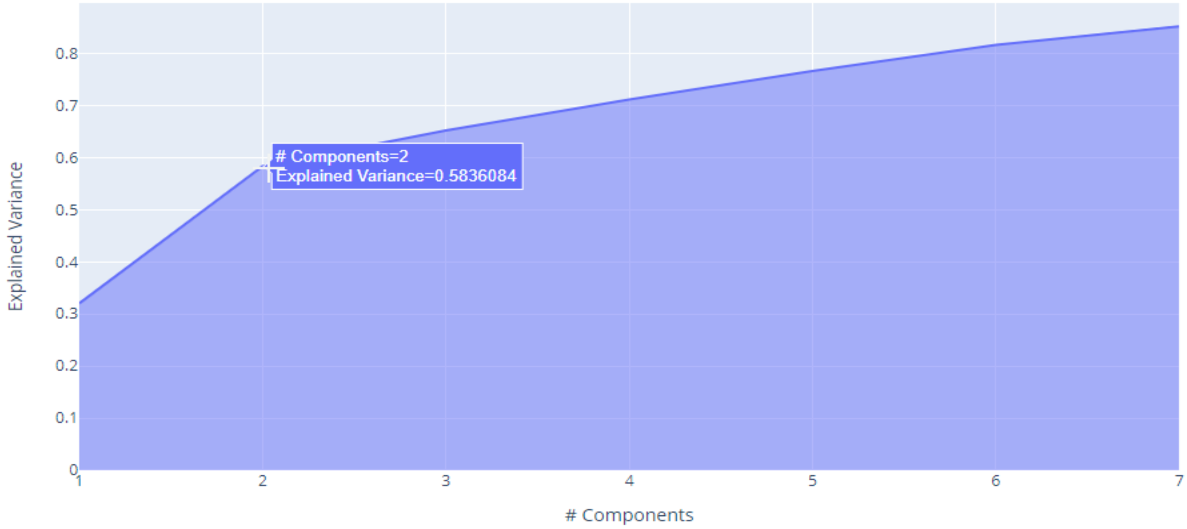


Figure 23.Area plot to show the variances at each component level(PC2)

With the components from PC1 to PC7, it explains the variance of 85.21%.

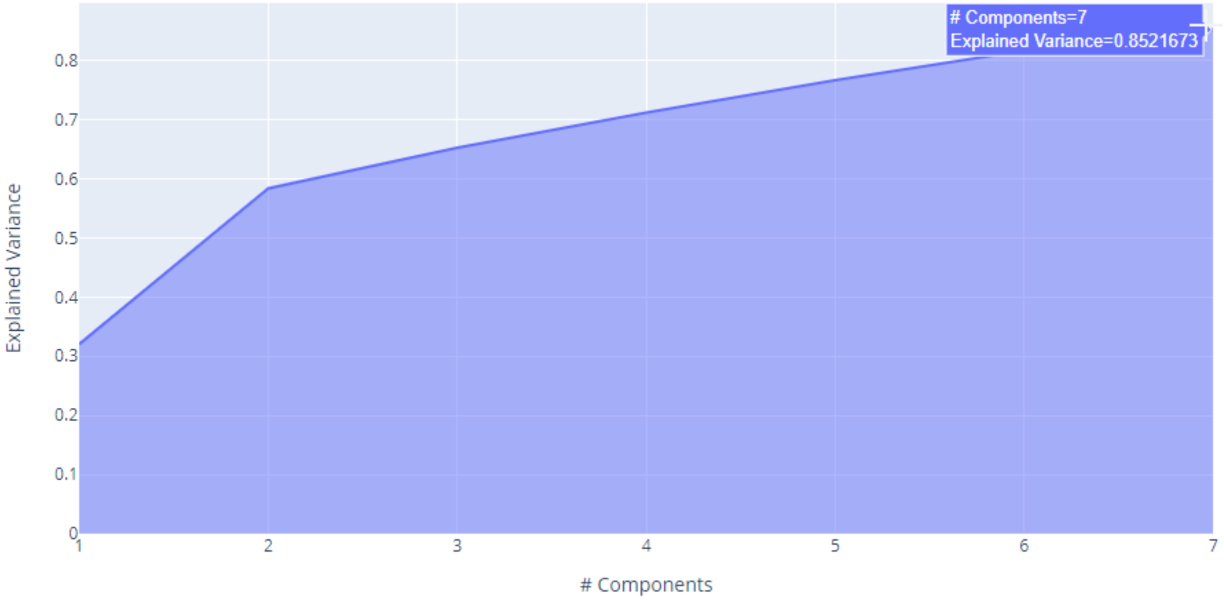


Figure 24. Area plot to show the variances at each component level(PC7)

THE END!