PGP - Data Science and Business Analytics

ADVANCED STATISTICS PROJECT REPORT

Akshaya Nallathambi

13th June, 2021

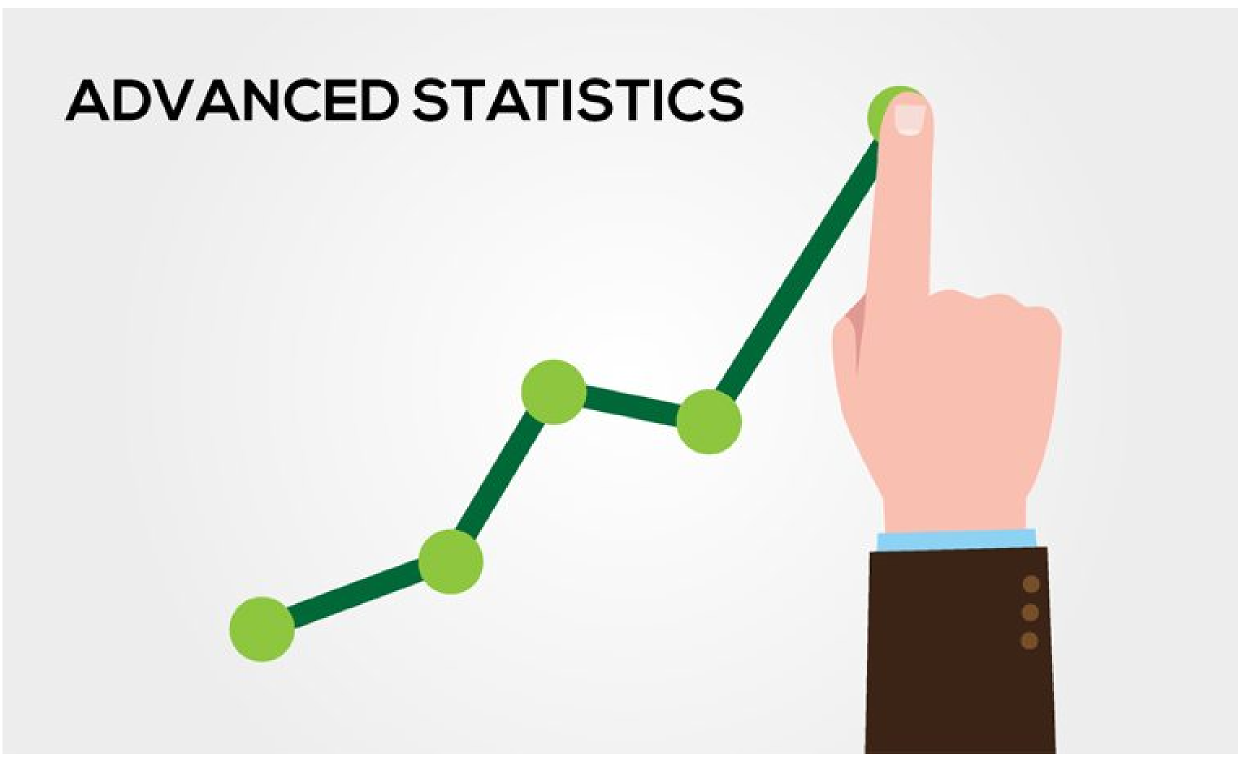


Table Of Contents

Problem 1

Problem statement **6**

Data Description **6**

Sample of the dataset **7**

Types of variables in the data frame **7**

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

Problem statement **15**

Data Description **15**

Sample of the dataset **17**

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**

2.2 Is scaling necessary for PCA in this case? Give justification and perform scaling **29**

2.3 Comment on the comparison between the covariance and the correlation matrices from this data.[on scaled data] **31**

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

2.5 Extract the eigenvalues and eigenvectors.[print both] **36**

2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features **39**

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] **40**

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? **42**

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] **44**

List of Figures

FIGURE 1	7
FIGURE 2	10
FIGURE 3	12
FIGURE 4	12
FIGURE 5	17
FIGURE 6	17
FIGURE 7	20
FIGURE 8	29
FIGURE 9	30
FIGURE 10	30
FIGURE 11	30
FIGURE 12	31
FIGURE 13	36
FIGURE 14	37
FIGURE 15	40
FIGURE 16	45

List of Tables

TABLE 1	7
TABLE 2	8
TABLE 3	9

TABLE 4 9

TABLE 5 12

TABLE 6 13

TABLE 7 18

TABLE 8 19

TABLE 9 32

TABLE 10 42

List of Graphs

GRAPH 1 11

GRAPH 2 14

GRAPH 3 21

GRAPH 4 26

GRAPH 5 28

GRAPH 6 33

GRAPH 7 35

GRAPH 8 39

GRAPH 9 43

GRAPH 10 44

Problem 1

Problem statement-

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.

Data Description-

Education: Doctorate, Bachelor, HS-grad

Occupation: Prof-specialty, Sales, Adm-clerical, Exec-managerial

Salary: Various salary level (ranging from 50,103 to 2,60,151)

Sample of the dataset-

FIGURE 1

	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
5	Doctorate	Sales	219420
6	Doctorate	Sales	237920
7	Doctorate	Sales	160540
8	Doctorate	Sales	180934

There are 3 variables out of which 2 are categorical and 1 is continuous. The data given is for 40 individuals. There are no null values.

Types of variables in the data frame-

TABLE 1

Education	object	Categorical
Occupation	object	Categorical
Salary	int64	Continuous

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

TABLE 2

H_0	Null Hypothesis
H_a	Alternative Hypothesis
μ	Hypothesized mean
α	Significance level

Salary and Education

H_0: The mean salary of individual is same with different categories of educational qualification.

H_a: The mean salary of individual is different in at-least one category of educational qualification.

Salary and Occupation

H_0: The mean salary of individual is same with different categories of occupation.

H_a: The mean salary of individual is different in at-least one category of occupation.

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

The below table gives the ANOVA output for Education with respect to the variable Salary -

TABLE 3

	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

Since the p value (1.257709e-08) is less than the significance level (α = 0.05), we can reject the null hypothesis and conclude that there is a difference in the mean salary of individual in at-least one category of educational qualification.

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.

The below table gives the ANOVA output for Occupation with respect to the variable Salary -

TABLE 4

	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

Since the p (0.458508) value is greater than the significance level (α = 0.05), we fail to reject the null hypothesis and conclude that there is a no difference in the mean salary of individual with respect to different categories of occupation.

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

The Tukey Test (or Tukey procedure), also called Tukey's Honest Significant Difference test, is a post-hoc test based on the studentized range distribution. An ANOVA test can tell you if your results are significant overall, but it won't tell you exactly where those differences lie. After you have run an ANOVA and found significant results, then you can run Tukey's HSD to find out which specific groups' means (compared with each other) are different. The test compares all possible pairs of means. [1]

The null hypothesis was rejected for Education with respect to the variable Salary. The below image shows the output of Tukey's HSD and it can be understood that all the combinations of Education are significantly different.

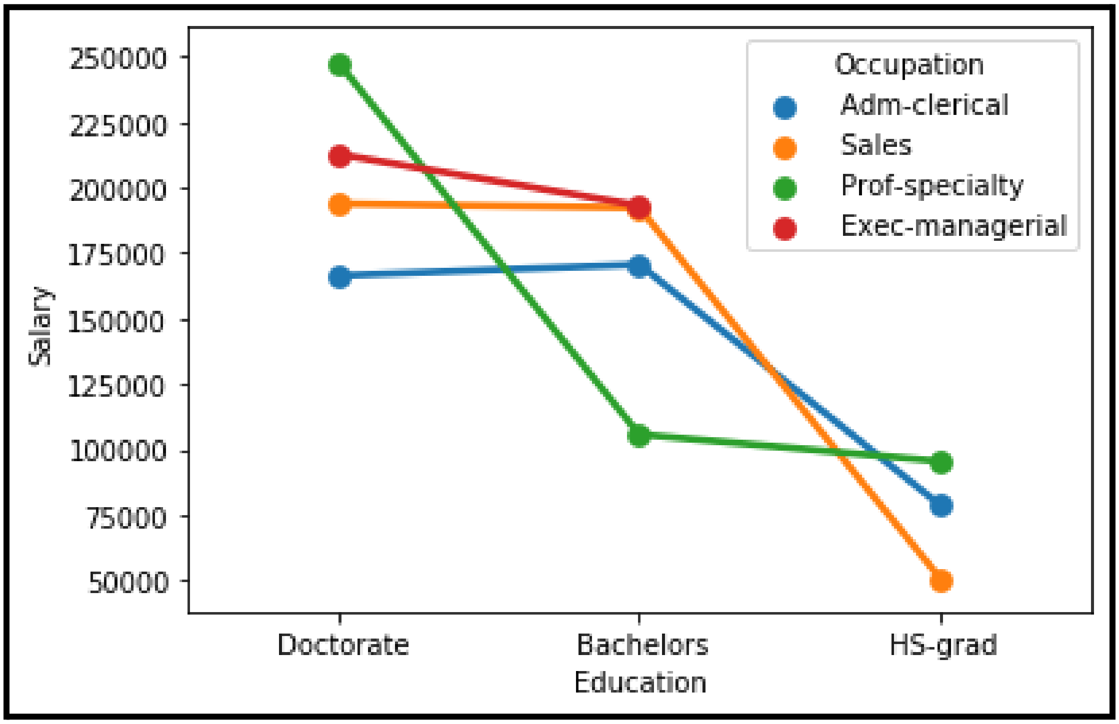
FIGURE 2

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

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.

GRAPH 1



The above chart shows that the salary is significantly high for the individuals with Doctorate in all the occupations . The individuals with Exec-managerial and Sales as occupation with an education qualification as Bachelors have the same salary. It is clear that the individuals with education qualification as HS-grad have the least salary. Also, the occupation Exec managerial has the education qualification as only Bachelors or Doctorate.

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?

TABLE 5

H_0	Null Hypothesis
H_a	Alternative Hypothesis
μ	Hypothesized mean
α	Significance level

Salary with Educational qualification and Occupation

H_0: The mean salary of an individual is the same with different categories of educational qualification and occupation.

H_a: The mean salary of an individual is different in at-least one category of educational qualification and occupation.

FIGURE 3

	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

FIGURE 4

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	2.284576	9.648715e-02
C(Education)	2.0	9.695663e+10	4.847831e+10	29.510933	3.708479e-08
Residual	34.0	5.585261e+10	1.642724e+09	NaN	NaN

Figure 3 & Figure 4 is the ANOVA output for the combination of Education-Occupation and Occupation-Education.

The below table gives the two way ANOVA output for *Salary* with Educational qualification, Occupation and the interaction of both Educational qualification and Occupation -

TABLE 6

	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

The p-value in both the treatments is less than the significance level (α = 0.05).

Due to the inclusion of the interaction effect term, we can see a change in the p-value of the first two treatments as compared to the Two-Way ANOVA without the interaction effect terms. And we see that the p-value of the interaction effect term of 'Education' and 'Occupation' suggests that the Null Hypothesis is rejected in this case.

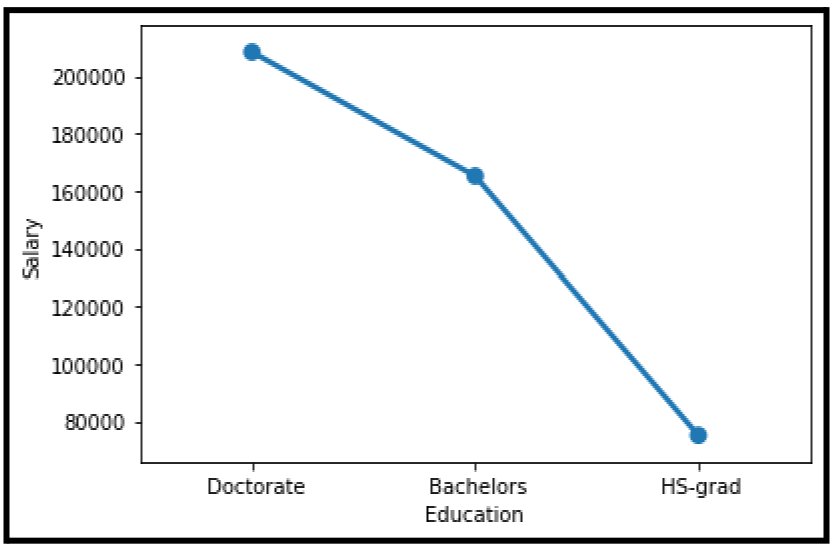
Therefore, the mean salary of an individual is different in at-least one category of educational qualification and occupation.

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

The ANOVA tests clearly show that the occupation has no effect on the salary. That is the mean salary of an individual is the same with different categories of occupation. Whereas the education has definitely an effect on the salary offered. That is the mean salary of an individual is different in at-least one category of educational qualification.

Thus I can conclude that higher the educational qualification, higher the salary. This can also be visually represented with the below graph,

GRAPH 2



Problem 2

Problem statement-

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 file: *Data Dictionary.xlsx.*

Data Description-

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

Sample of the dataset-

FIGURE 5

	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

FIGURE 6

perc.alumni	Expend	Grad.Rate
12	7041	60
16	10527	56
30	8735	54
37	19016	59
2	10922	15

There are 18 variables out of which 1 is categorical and 17 are numerical. The data given is for 777 colleges. There are no null values.

Types of variables in the data frame-

TABLE 7

Names	object	Categorical
Apps	int64	Continuous
Accept	int64	Continuous
Enroll	int64	Continuous
Top10perc	int64	Continuous
Top25perc	int64	Continuous
F.Undergrad	int64	Continuous
P.Undergrad	int64	Continuous
Outstate	int64	Continuous
Room.Board	int64	Continuous
Books	int64	Continuous
Personal	int64	Continuous
PhD	int64	Continuous
Terminal	int64	Continuous
S.F.Ratio	float64	Continuous
perc.alumni	int64	Continuous
Expend	int64	Continuous
Grad.Rate	int64	Continuous

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

There is only one categorical variable “Name” and each value is unique (777 values) in it. So univariate and bivariate analysis is done only in the numerical variables of the given data set.

Univariate analysis:

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data. [2]

The histograms are used for numerical variables to perform univariate analysis.

It is clear from the graph (Graph 3) that all the numerical variables are skewed.

Skewness:

Skewness is a number that indicates to what extent a variable is asymmetrically distributed. The below table represents the level of skewness and the respective skewness value. [3]

TABLE 8

Skewness level	Value
Symmetrical or Not Skewed	0
Less Skewed Data	± 0.5 to 1
Highly Skewed Data	Greater than ±1

When the skewness value is positive it is considered as right skewed data and when the skewness value is negative it is considered as left skewed data.

The table below shows the skewness value corresponding to each variable in the given data set.

FIGURE 7

	Skewness
Apps	3.716557
Accept	3.411126
Enroll	2.685268
Top10perc	1.410487
Top25perc	0.258839
F.Undergrad	2.605416
P.Undergrad	5.681358
Outstate	0.508294
Room.Board	0.476434
Books	3.478293
Personal	1.739131
PhD	-0.766686
Terminal	-0.814965
S.F.Ratio	0.666146
perc.alumni	0.605719
Expend	3.452640
Grad.Rate	-0.113558

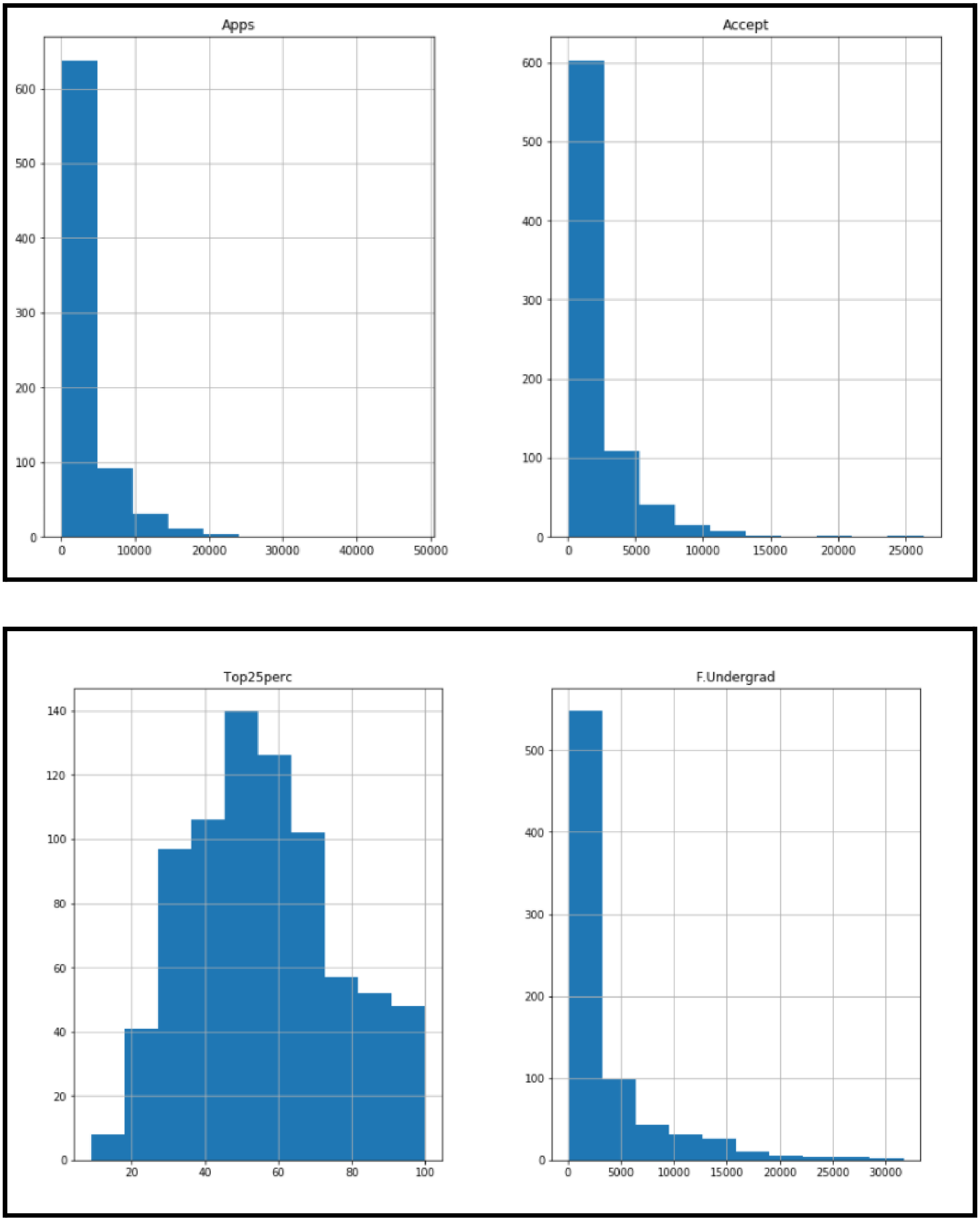
Right skewed variables:

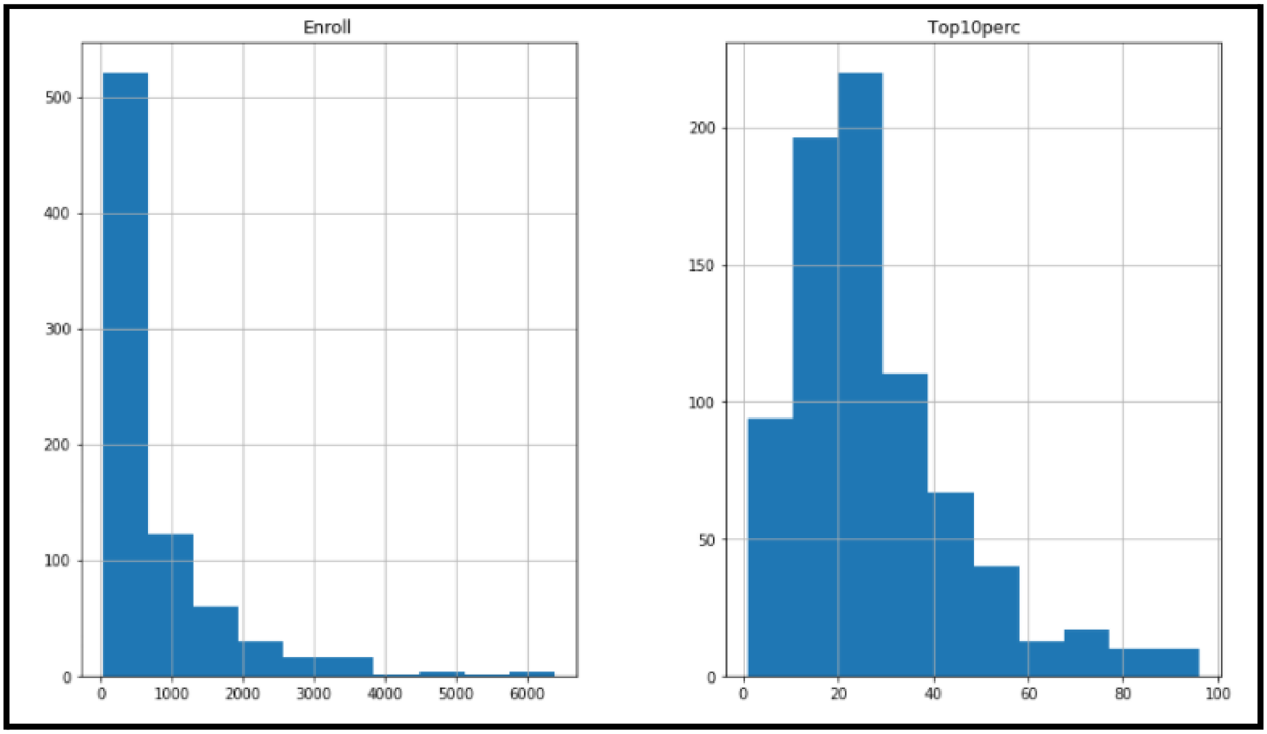
Apps, Accept, Enroll, Top10perc, Top25perc, F.Undergrad, P.Undergrad, Outstate, Room.Board, Books, Personal, S.F.Ratio, perc.alumni, Expend

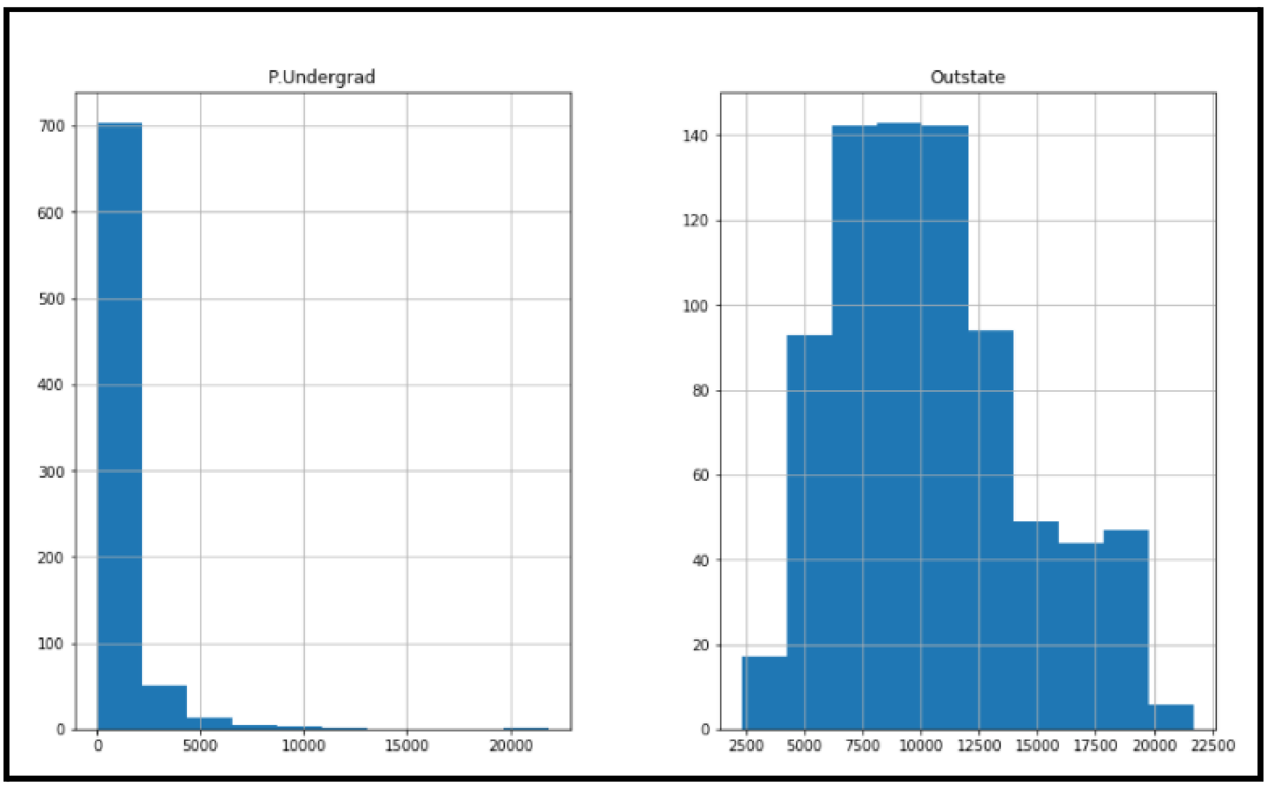
Left skewed variables:

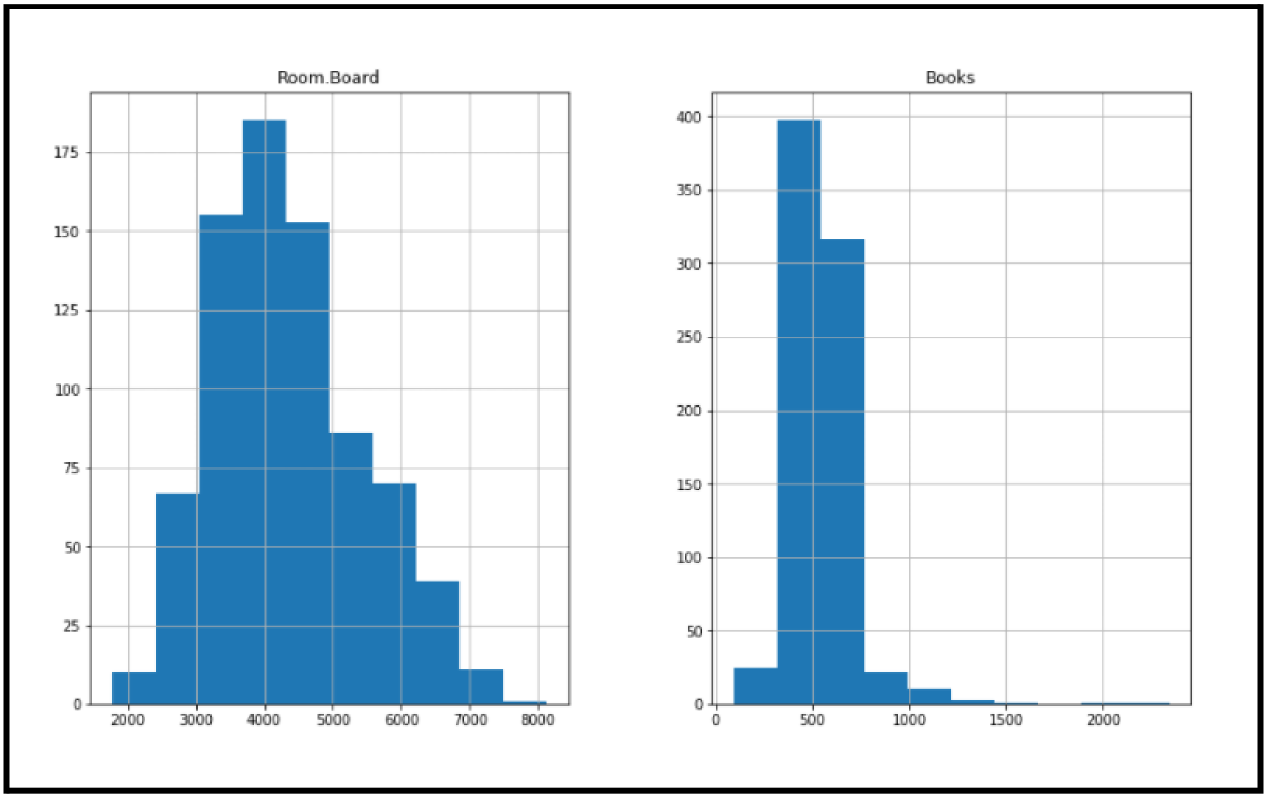
PhD, Terminal, Grad.Rate

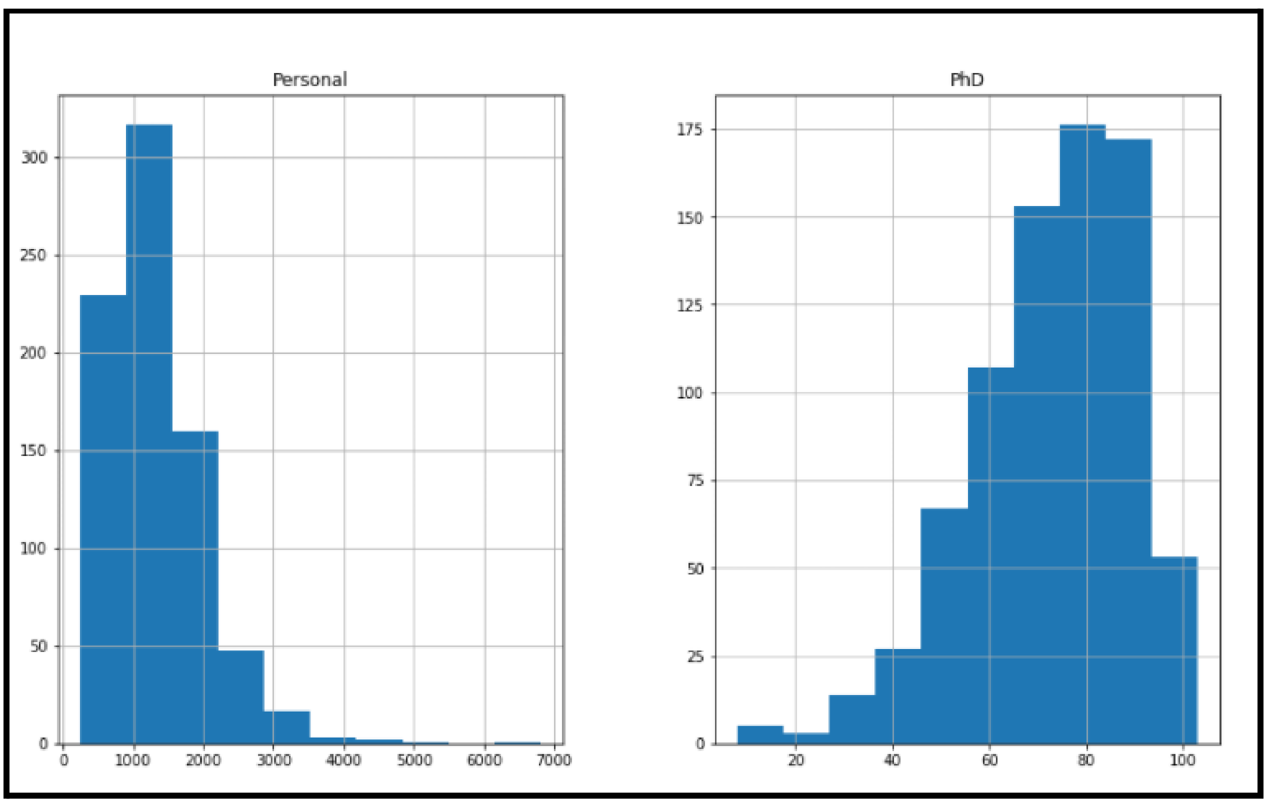
GRAPH 3

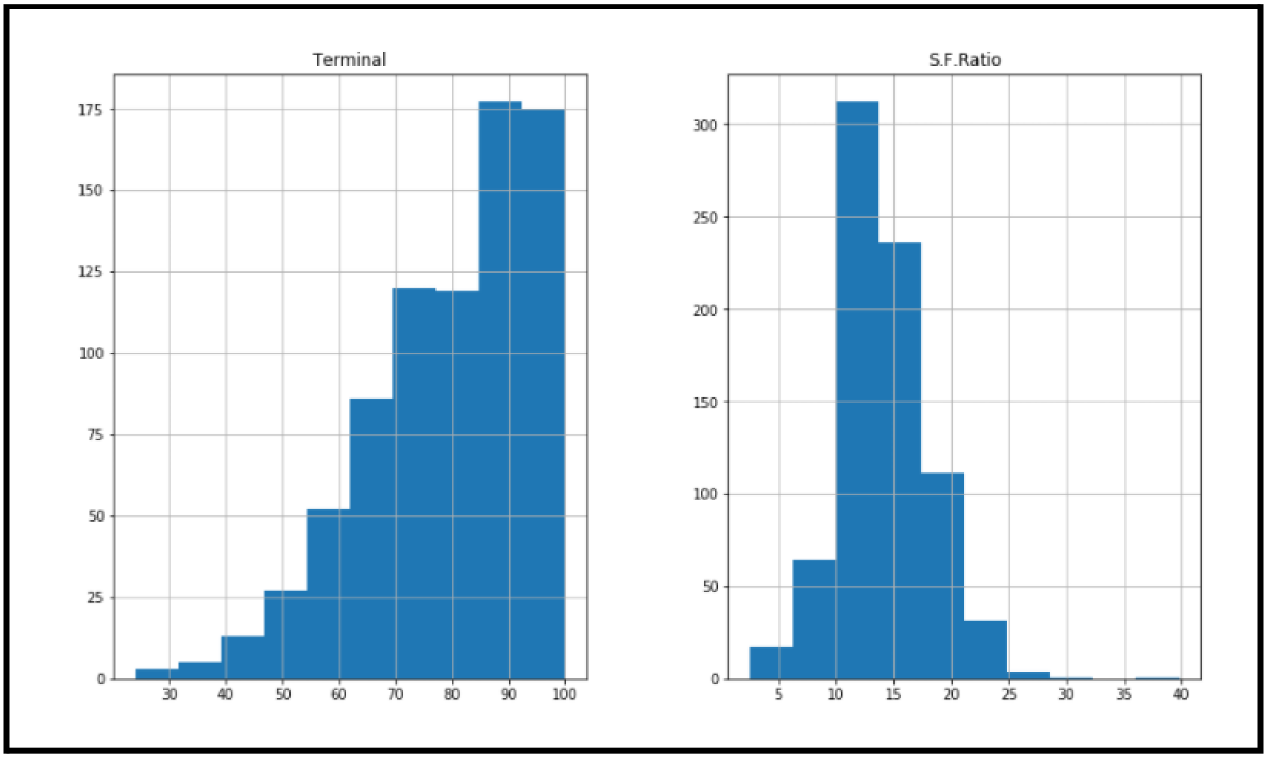

Apps
0
100
200
300
400
500
600
0
10000
20000
30000
40000
50000
Accept
0
100
200
300
400
500
600
0
5000
10000
15000
20000
25000
Top25perc
0
20
40
60
80
100
120
140
20
40
60
80
100
F.Undergrad
0
100
200
300
400
500
0
5000
10000
15000
20000
25000
30000

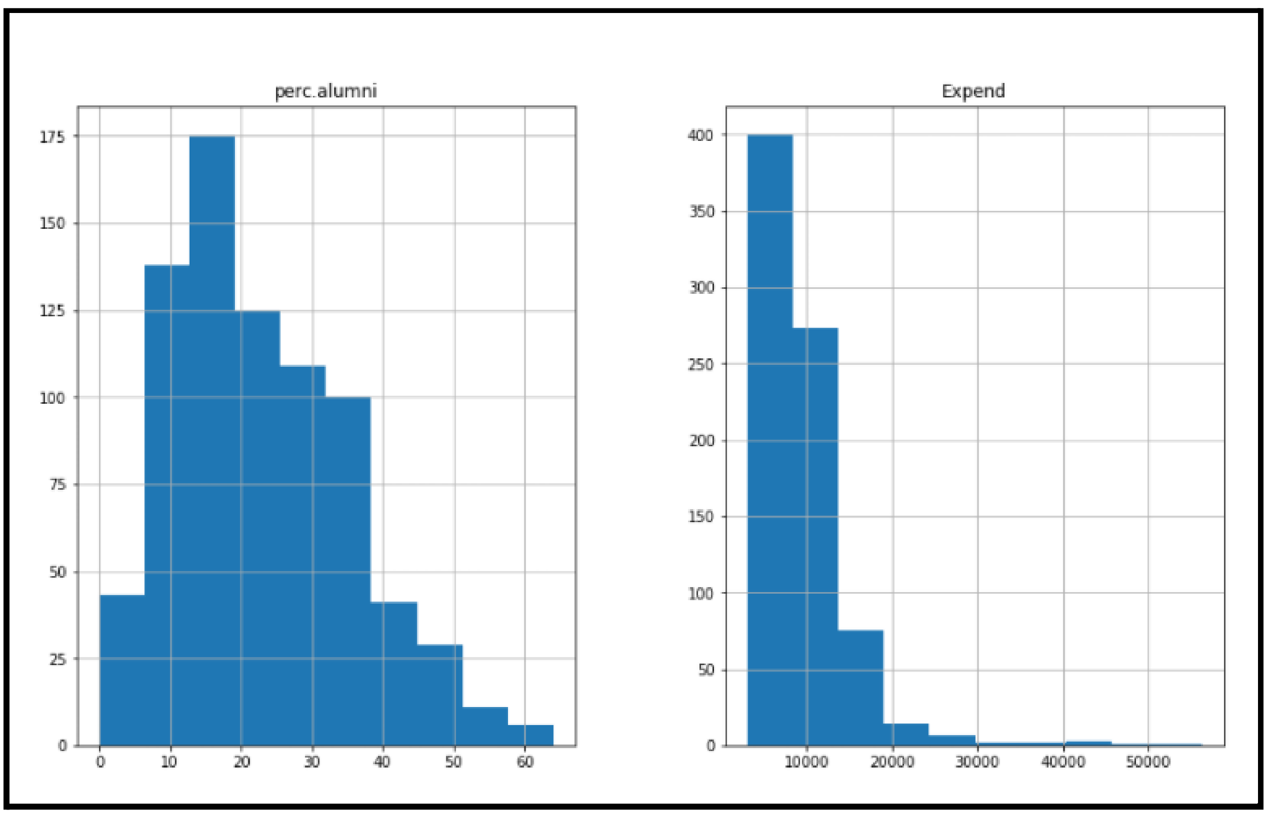

Enroll
0
100
200
300
400
500
0 1000 2000 3000 4000 5000 6000
Top10perc
0
50
100
150
200
0 20 40 60 80 100

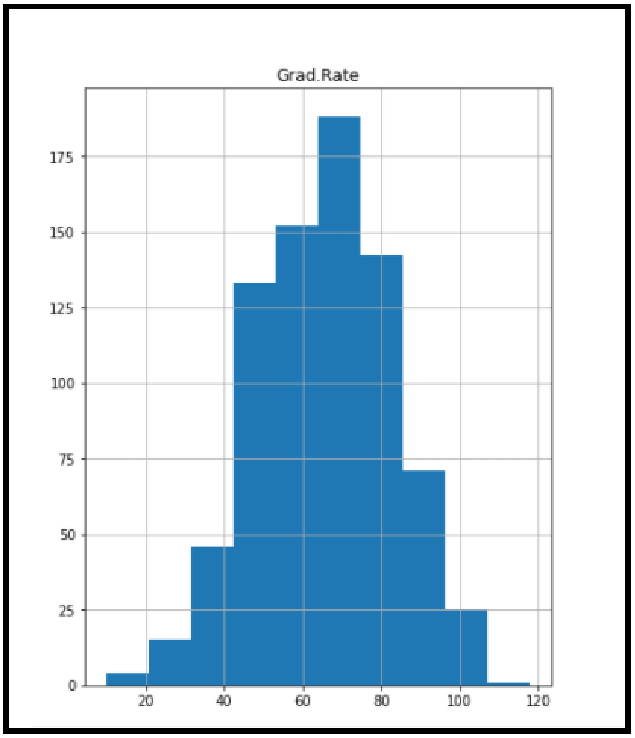

P.Undergrad
0
100
200
300
400
500
600
700
0 5000 10000 15000 20000
Outstate
0
20
40
60
80
100
120
140
2500 5000 7500 10000 12500 15000 17500 20000 22500


Room.Board
0
25
50
75
100
125
150
175
2000
3000
4000
5000
6000
7000
8000
Books
0
50
100
150
200
250
300
350
400
0
500
1000
1500
2000


Personal
0
50
100
150
200
250
300
0
1000
2000
3000
4000
5000
6000
7000
PhD
0
25
50
75
100
125
150
175
20
40
60
80
100


Terminal
175
150
125
100
75
50
25
0
30 40 50 60 70 80 90 100
S.F.Ratio
300
250
200
150
100
50
0
5 10 15 20 25 30 35 40


perc.alumni
175
150
125
100
75
50
25
0
0 10 20 30 40 50 60
Expend
400
350
300
250
200
150
100
50
0
10000 20000 30000 40000 50000



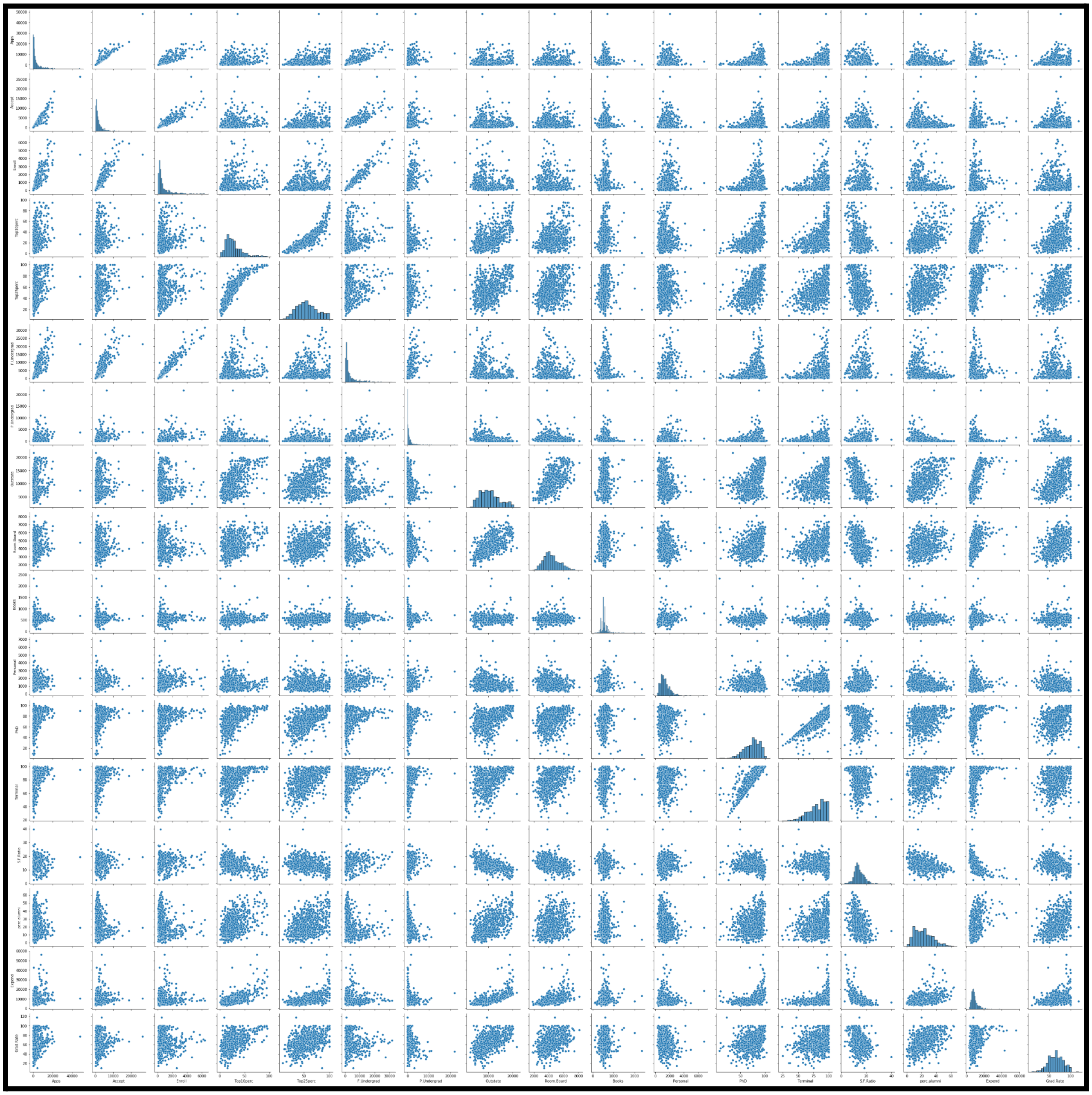
Bivariate analysis:

Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them. Bivariate analysis can be helpful in testing simple hypotheses of association. [4]

The pairplot is generally used for numerical variables to perform bivariate analysis.

A pairplot plots a pairwise relationship in a dataset. The pairplot function creates a grid of Axes such that each variable in data will be shared in the y-axis across a single row and in the x-axis across a single column. [5]

GRAPH 4

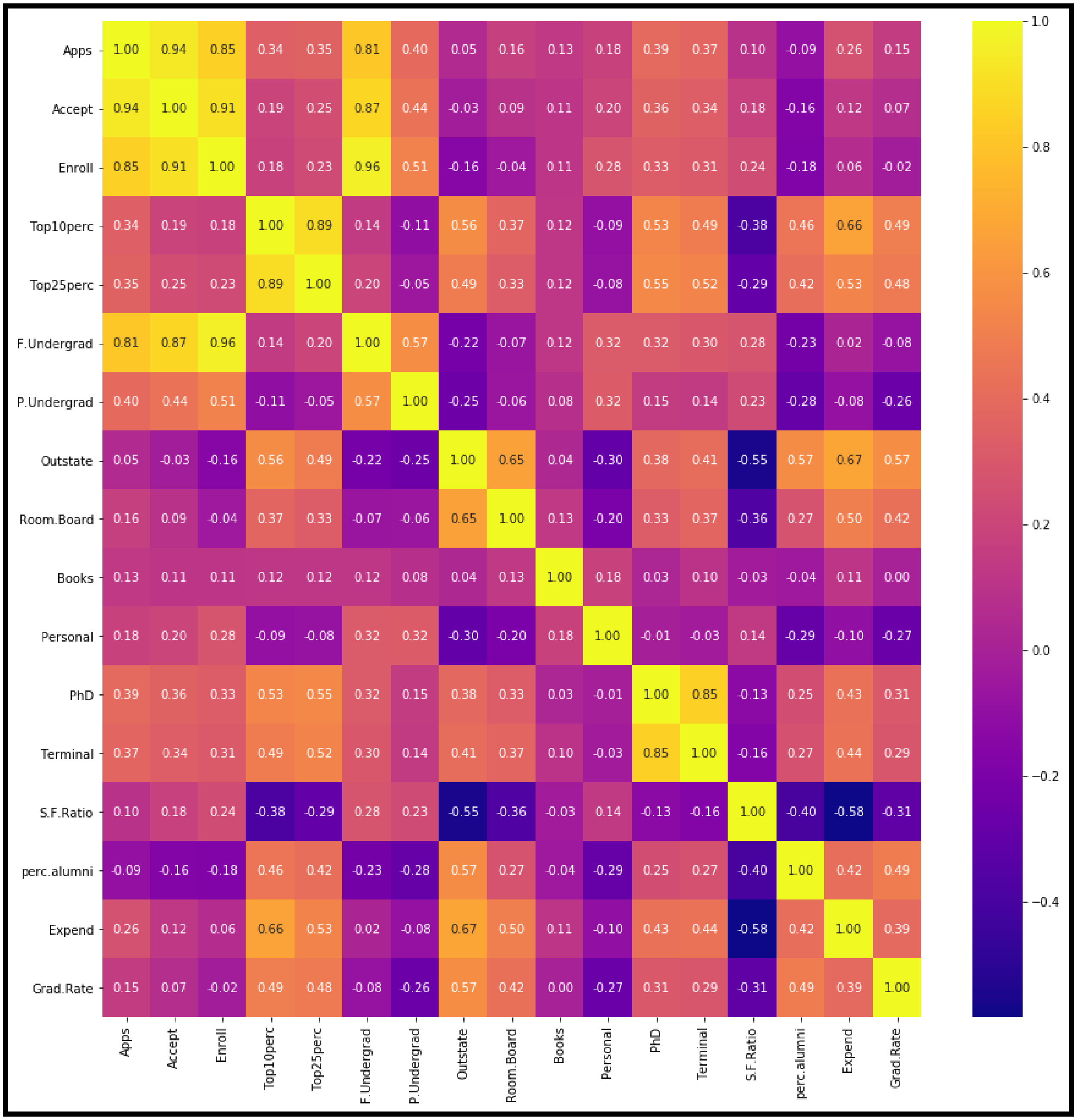


The below are the findings from the pairplot generated -

- The variable Apps is highly correlated with the variables Accept, Enroll and F.Undergrad.
- The variable Accept is highly correlated with the variables Enroll and F.Undergrad.
- The variable Enroll is highly correlated with the variable F.Undergrad.
- The variable Top10perc is highly correlated with the variable Top25perc.
- The variable PhD is highly correlated with the variable Terminal.

The heat map can also be used to check the association between two variables. All the boxes with a value higher than 0.8 are highly correlated. The heat map for all the numerical variable is below,

GRAPH 5



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

Scaling:

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. [6]

Scaling converts variables with different scales of measurements into a single scale. This is done only for the numerical variables.

The data is scaled using the formula $\frac{X-\mu}{\sigma}$.

μ: Mean

σ: Standard deviation

The process of scaling is necessary in the given data set as the variables of the data set are of different scales i.e. one variable has five digit numbers and other has only two digit numbers. For e.g. in our data set “Expend” is having values in thousands and “S.F.Ratio” in just two digits. Since the data in these variables are of different scales, it is tough to compare these variables. Therefore scaling is done in the given data set.

Before Scaling:

FIGURE 8

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

FIGURE 9

perc.alumni	Expend	Grad.Rate
12	7041	60
16	10527	56
30	8735	54
37	19016	59
2	10922	15

<u>After Scaling:</u>

FIGURE 10

Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio
-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
-0.210884	-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
-0.406866	-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
-0.668261	-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
-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

FIGURE 11

perc.alumni	Expend	Grad.Rate
-0.867574	-0.501910	-0.318252
-0.544572	0.166110	-0.551262
0.585935	-0.177290	-0.667767
1.151188	1.792851	-0.376504
-1.675079	0.241803	-2.939613

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

Covariance matrix:

A covariance matrix is a square matrix giving the covariance between each pair of elements of a given random vector. [7]

FIGURE 12

```
Covariance Matrix
%s [[ 1.00128866  0.94466636  0.84791332  0.33927032  0.35209304  0.81554018
   0.3987775   0.05022367  0.16515151  0.13272942  0.17896117  0.39120081
   0.36996762  0.09575627 -0.09034216  0.2599265   0.14694372]
 [ 0.94466636  1.00128866  0.91281145  0.19269493  0.24779465  0.87534985
   0.44183938 -0.02578774  0.09101577  0.11367165  0.20124767  0.35621633
   0.3380184   0.17645611 -0.16019604  0.12487773  0.06739929]
 [ 0.84791332  0.91281145  1.00128866  0.18152715  0.2270373   0.96588274
   0.51372977 -0.1556777  -0.04028353  0.11285614  0.28129148  0.33189629
   0.30867133  0.23757707 -0.18102711  0.06425192 -0.02236983]
 [ 0.33927032  0.19269493  0.18152715  1.00128866  0.89314445  0.1414708
  -0.10549205  0.5630552   0.37195909  0.1190116  -0.09343665  0.53251337
   0.49176793 -0.38537048  0.45607223  0.6617651   0.49562711]
 [ 0.35209304  0.24779465  0.2270373   0.89314445  1.00128866  0.19970167
  -0.05364569  0.49002449  0.33191707  0.115676   -0.08091441  0.54656564
   0.52542506 -0.29500852  0.41840277  0.52812713  0.47789622]
 [ 0.81554018  0.87534985  0.96588274  0.1414708   0.19970167  1.00128866
   0.57124738 -0.21602002 -0.06897917  0.11569867  0.31760831  0.3187472
   0.30040557  0.28006379 -0.22975792  0.01867565 -0.07887464]
 [ 0.3987775   0.44183938  0.51372977 -0.10549205 -0.05364569  0.57124738
   1.00128866 -0.25383901 -0.06140453  0.08130416  0.32029384  0.14930637
   0.14208644  0.23283016 -0.28115421 -0.08367612 -0.25733218]
 [ 0.05022367 -0.02578774 -0.1556777   0.5630552   0.49002449 -0.21602002
  -0.25383901  1.00128866  0.65509951  0.03890494 -0.29947232  0.38347594
   0.40850895 -0.55553625  0.56699214  0.6736456   0.57202613]
 [ 0.16515151  0.09101577 -0.04028353  0.37195909  0.33191707 -0.06897917
  -0.06140453  0.65509951  1.00128866  0.12812787 -0.19968518  0.32962651
   0.3750222  -0.36309504  0.27271444  0.50238599  0.42548915]
 [ 0.13272942  0.11367165  0.11285614  0.1190116   0.115676    0.11569867
   0.08130416  0.03890494  0.12812787  1.00128866  0.17952581  0.0269404
   0.10008351 -0.03197042 -0.04025955  0.11255393  0.00106226]
 [ 0.17896117  0.20124767  0.28129148 -0.09343665 -0.08091441  0.31760831
   0.32029384 -0.29947232 -0.19968518  0.17952581  1.00128866 -0.01094989
  -0.03065256  0.13652054 -0.2863366  -0.09801804 -0.26969106]
 [ 0.39120081  0.35621633  0.33189629  0.53251337  0.54656564  0.3187472
   0.14930637  0.38347594  0.32962651  0.0269404  -0.01094989  1.00128866
   0.85068186 -0.13069832  0.24932955  0.43331936  0.30543094]
 [ 0.36996762  0.3380184   0.30867133  0.49176793  0.52542506  0.30040557
   0.14208644  0.40850895  0.3750222   0.10008351 -0.03065256  0.85068186
   1.00128866 -0.16031027  0.26747453  0.43936469  0.28990033]
 [ 0.09575627  0.17645611  0.23757707 -0.38537048 -0.29500852  0.28006379
   0.23283016 -0.55553625 -0.36309504 -0.03197042  0.13652054 -0.13069832
  -0.16031027  1.00128866 -0.4034484  -0.5845844  -0.30710565]
 [-0.09034216 -0.16019604 -0.18102711  0.45607223  0.41840277 -0.22975792
  -0.28115421  0.56699214  0.27271444 -0.04025955 -0.2863366   0.24932955
   0.26747453 -0.4034484   1.00128866  0.41825001  0.49153016]
 [ 0.2599265   0.12487773  0.06425192  0.6617651   0.52812713  0.01867565
  -0.08367612  0.6736456   0.50238599  0.11255393 -0.09801804  0.43331936
   0.43936469 -0.5845844   0.41825001  1.00128866  0.39084571]
 [ 0.14694372  0.06739929 -0.02236983  0.49562711  0.47789622 -0.07887464
  -0.25733218  0.57202613  0.42548915  0.00106226 -0.26969106  0.30543094
   0.28990033 -0.30710565  0.49153016  0.39084571  1.00128866]]
```

Correlation matrix:

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses. [8]

TABLE 9

	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	1	0.943451	0.846822	0.338834	0.35164	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	0.369491	0.095633	-0.090226	0.259592	0.146755
Accept	0.943451	1	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899	0.113525	0.200989	0.355758	0.337583	0.176229	-0.15999	0.124717	0.067313
Enroll	0.846822	0.911637	1	0.181294	0.226745	0.96464	0.513069	-0.155477	-0.040232	0.112711	0.280929	0.331469	0.308274	0.237271	-0.180794	0.064169	-0.022341
Top10perc	0.338834	0.192447	0.181294	1	0.891995	0.141289	-0.105356	0.562331	0.37148	0.118858	-0.093316	0.531828	0.491135	-0.384875	0.455485	0.660913	0.494989
Top25perc	0.35164	0.247476	0.226745	0.891995	1	0.199445	-0.053577	0.489394	0.33149	0.115527	-0.08081	0.545862	0.524749	-0.294629	0.417864	0.527447	0.477281
F.Undergrad	0.814491	0.874223	0.96464	0.141289	0.199445	1	0.570512	-0.215742	-0.06889	0.11555	0.3172	0.318337	0.300019	0.279703	-0.229462	0.018652	-0.078773
P.Undergrad	0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	1	-0.253512	-0.061326	0.0812	0.319882	0.149114	0.141904	0.232531	-0.280792	-0.083568	-0.257001
Outstate	0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	-0.253512	1	0.654256	0.038855	-0.299087	0.382982	0.407983	-0.554821	0.566262	0.672779	0.57129
Room.Board	0.164939	0.090899	-0.040232	0.37148	0.33149	-0.06889	-0.061326	0.654256	1	0.127963	-0.199428	0.329202	0.37454	-0.362628	0.272363	0.501739	0.424942
Books	0.132559	0.113525	0.112711	0.118858	0.115527	0.11555	0.0812	0.038855	0.127963	1	0.179295	0.026906	0.099955	-0.031929	-0.040208	0.112409	0.001061
Personal	0.178731	0.200989	0.280929	-0.093316	-0.08081	0.3172	0.319882	-0.299087	-0.199428	0.179295	1	-0.010936	-0.030613	0.136345	-0.285968	-0.097892	-0.269344
PhD	0.390697	0.355758	0.331469	0.531828	0.545862	0.318337	0.149114	0.382982	0.329202	0.026906	-0.010936	1	0.849587	-0.13053	0.249009	0.432762	0.305038
Terminal	0.369491	0.337583	0.308274	0.491135	0.524749	0.300019	0.141904	0.407983	0.37454	0.099955	-0.030613	0.849587	1	-0.160104	0.26713	0.438799	0.289527
S.F.Ratio	0.095633	0.176229	0.237271	-0.384875	-0.294629	0.279703	0.232531	-0.554821	-0.362628	-0.031929	0.136345	-0.13053	-0.160104	1	-0.402929	-0.583832	-0.30671
perc.alumni	-0.090226	-0.15999	-0.180794	0.455485	0.417864	-0.229462	-0.280792	0.566262	0.272363	-0.040208	-0.285968	0.249009	0.26713	-0.402929	1	0.417712	0.490898
Expend	0.259592	0.124717	0.064169	0.660913	0.527447	0.018652	-0.083568	0.672779	0.501739	0.112409	-0.097892	0.432762	0.438799	-0.583832	0.417712	1	0.390343
Grad.Rate	0.146755	0.067313	-0.022341	0.494989	0.477281	-0.078773	-0.257001	0.57129	0.424942	0.001061	-0.269344	0.305038	0.289527	-0.30671	0.490898	0.390343	1

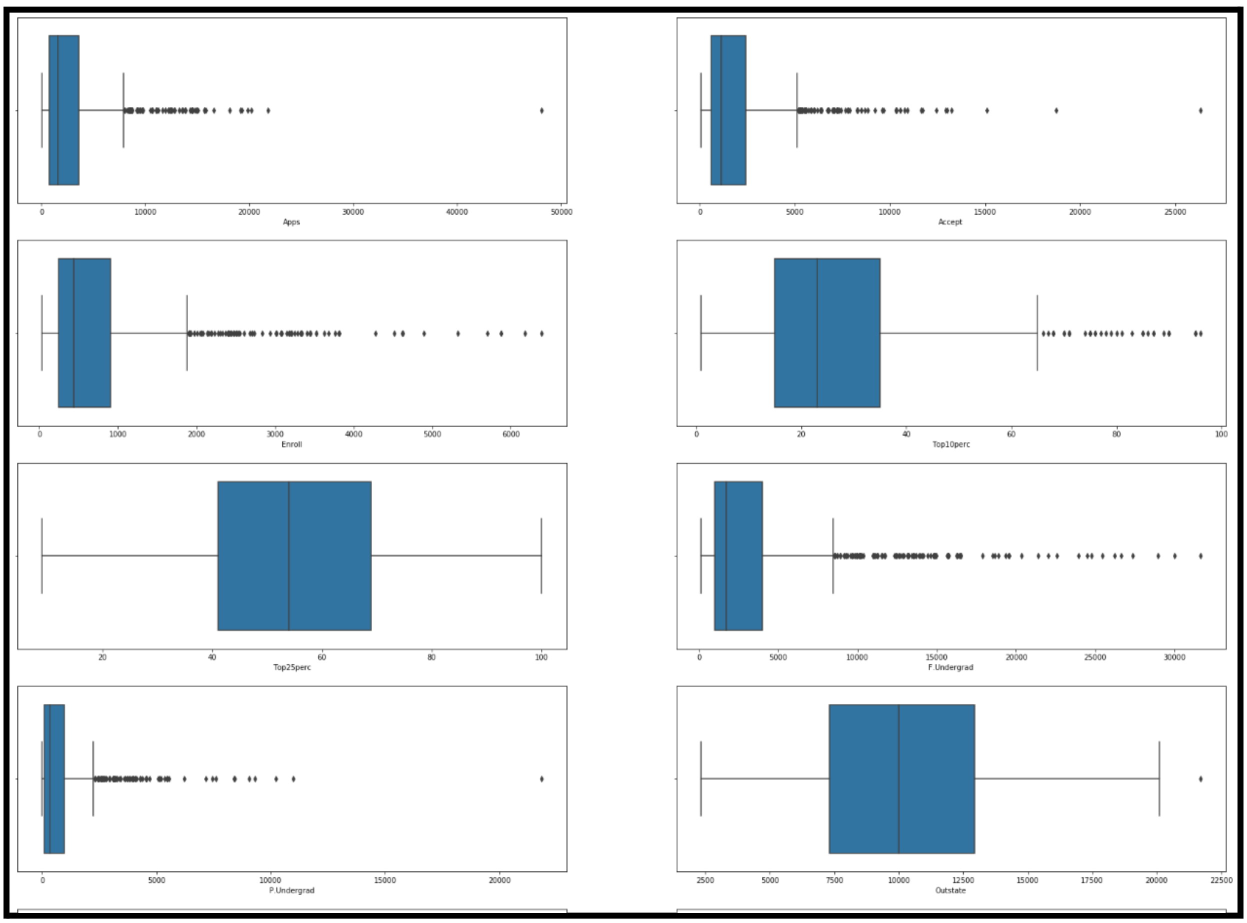
The correlation matrix of the standard scaled dataset is almost the same as the covariance matrix of the same standard scaled dataset.

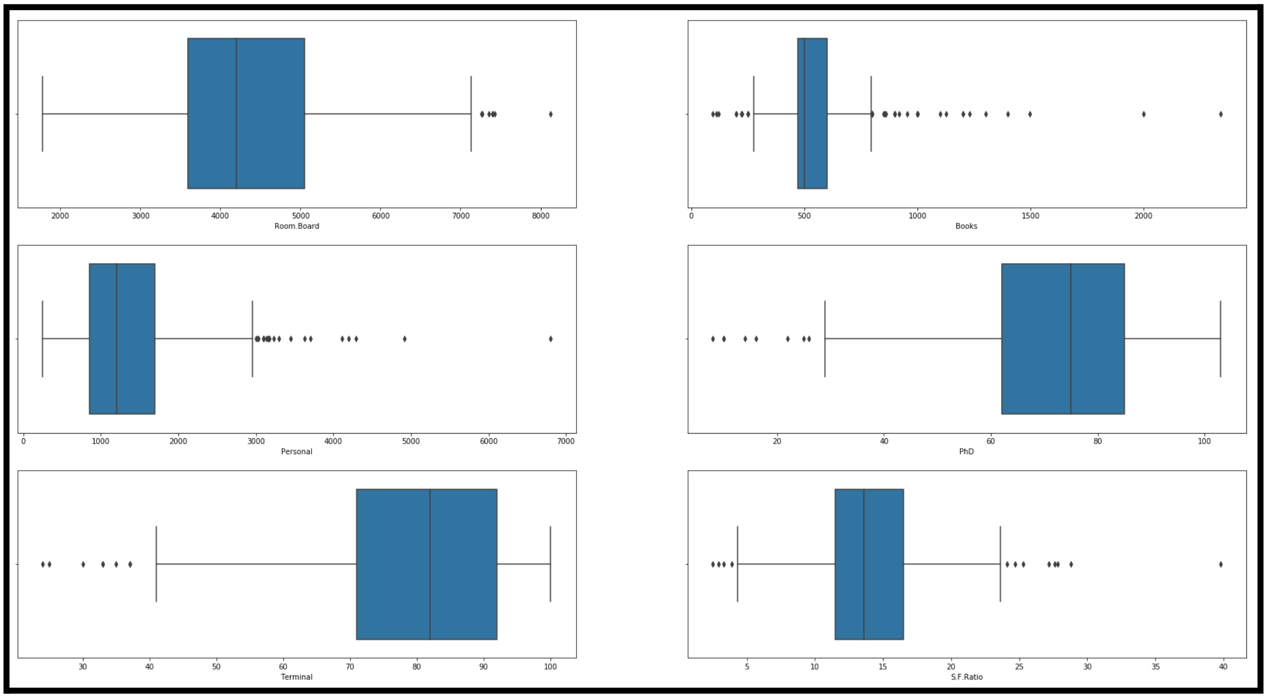
2.4 Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]

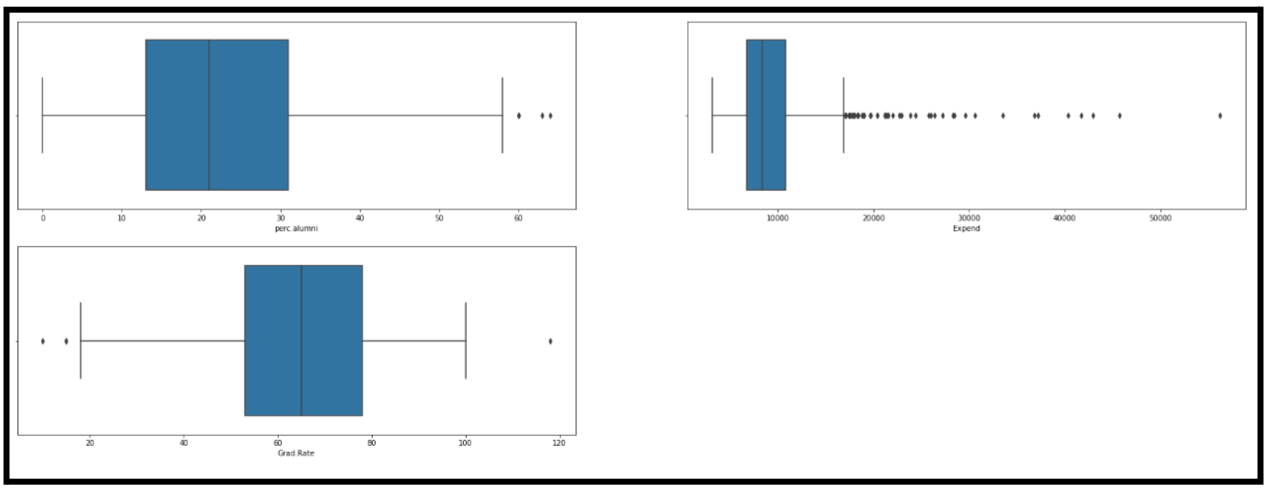
Outliers can be checked using boxplots. The given dataset has outliers in it.

Before scaling -

GRAPH 6

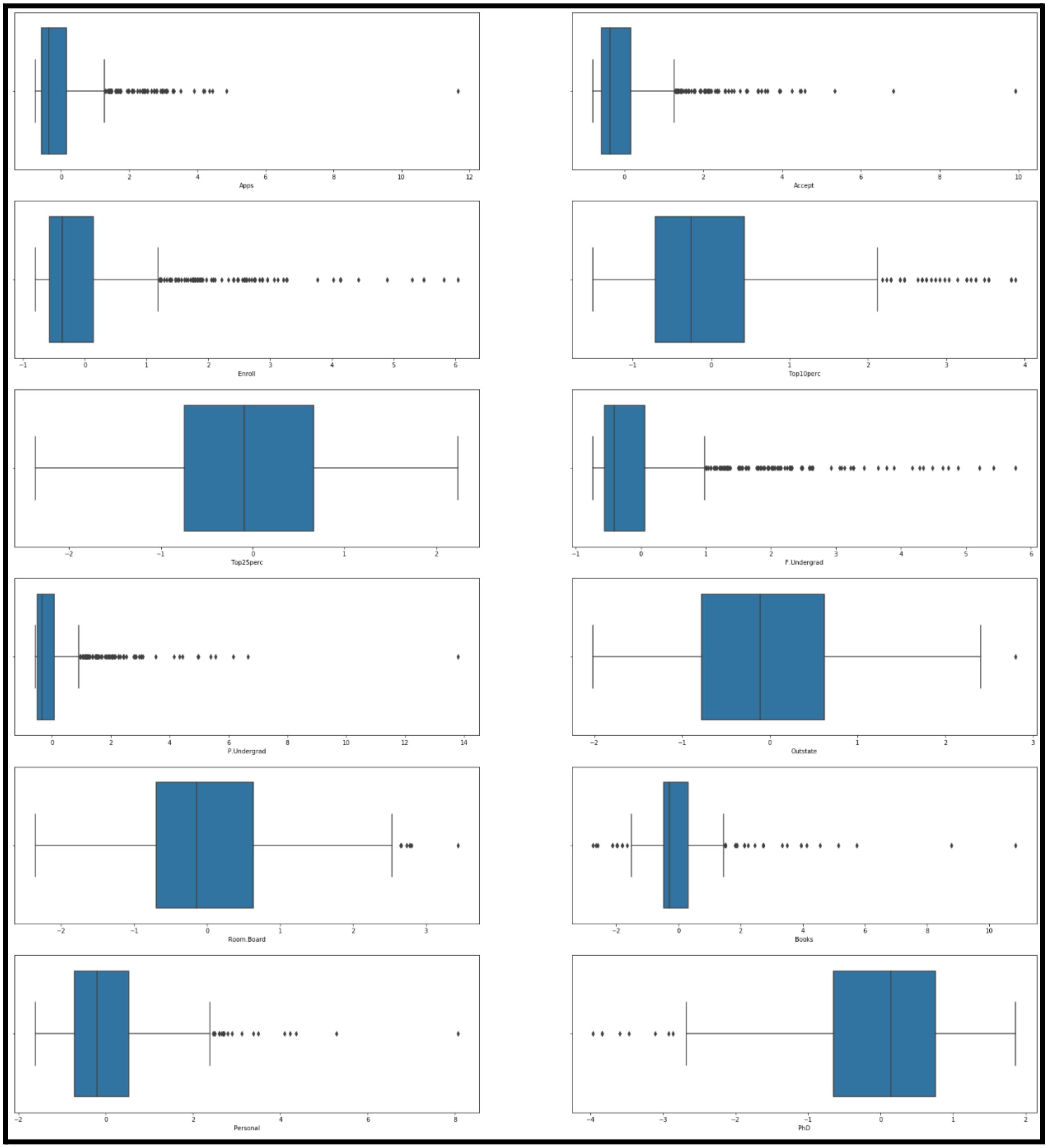


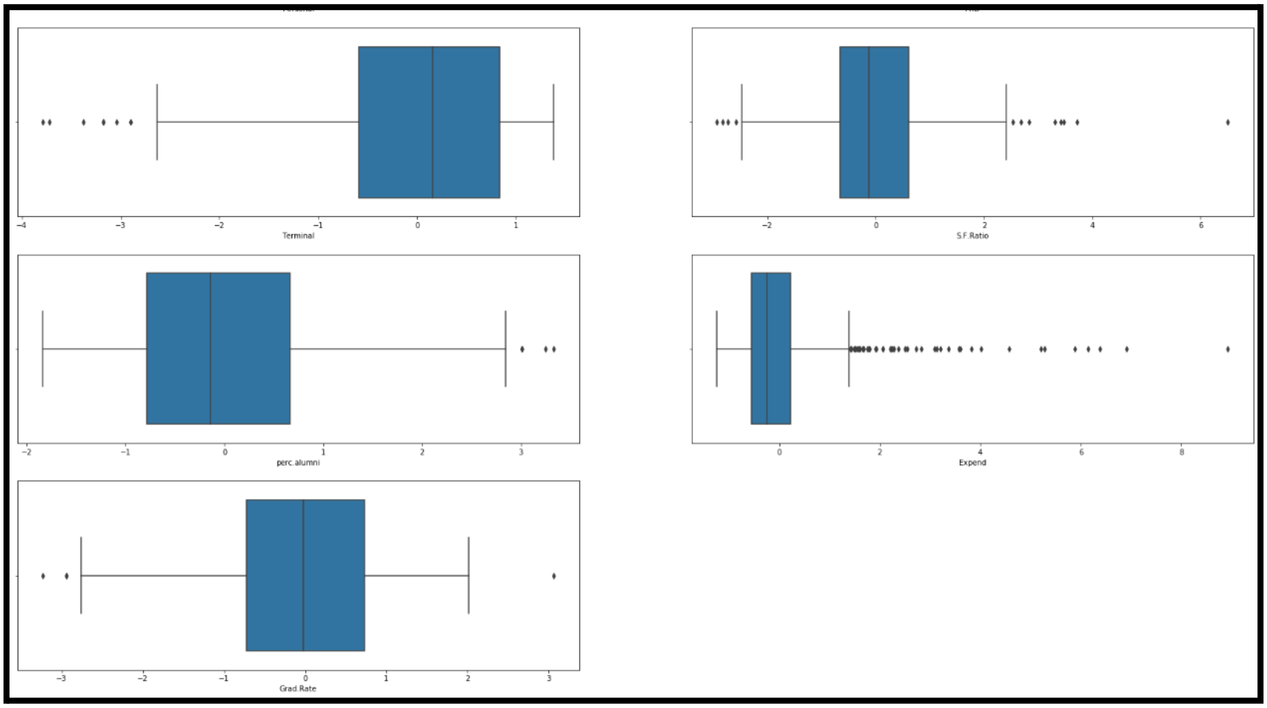

Room.Board
Books
Personal
PhD
Terminal
S.F.Ratio


perc.alumni
Expend
Grad.Rate

After scaling -

GRAPH 7





It is clear from the graphs that outliers are still present in the data even after scaling. Scaling only converts variables with different scales of measurements into a single scale. It will not necessarily remove the outliers present in the data.

2.5 Extract the eigenvalues and eigenvectors.[print both]

Eigenvalues:

FIGURE 13

```
Eigen Values
%s [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]
```

Eigenvectors:

FIGURE 14

```
Eigen Vectors
 %s [[-2.48765602e-01  3.31598227e-01  6.30921033e-02 -2.81310530e-01
   5.74140964e-03  1.62374420e-02  4.24863486e-02  1.03090398e-01
   9.02270802e-02 -5.25098025e-02  3.58970400e-01 -4.59139498e-01
   4.30462074e-02 -1.33405806e-01  8.06328039e-02 -5.95830975e-01
   2.40709086e-02]
 [-2.07601502e-01  3.72116750e-01  1.01249056e-01 -2.67817346e-01
   5.57860920e-02 -7.53468452e-03  1.29497196e-02  5.62709623e-02
   1.77864814e-01 -4.11400844e-02 -5.43427250e-01  5.18568789e-01
  -5.84055850e-02  1.45497511e-01  3.34674281e-02 -2.92642398e-01
  -1.45102446e-01]
 [-1.76303592e-01  4.03724252e-01  8.29855709e-02 -1.61826771e-01
  -5.56936353e-02  4.25579803e-02  2.76928937e-02 -5.86623552e-02
   1.28560713e-01 -3.44879147e-02  6.09651110e-01  4.04318439e-01
  -6.93988831e-02 -2.95896092e-02 -8.56967180e-02  4.44638207e-01
   1.11431545e-02]
 [-3.54273947e-01 -8.24118211e-02 -3.50555339e-02  5.15472524e-02
  -3.95434345e-01  5.26927980e-02  1.61332069e-01  1.22678028e-01
  -3.41099863e-01 -6.40257785e-02 -1.44986329e-01  1.48738723e-01
  -8.10481404e-03 -6.97722522e-01 -1.07828189e-01 -1.02303616e-03
   3.85543001e-02]
 [-3.44001279e-01 -4.47786551e-02  2.41479376e-02  1.09766541e-01
  -4.26533594e-01 -3.30915896e-02  1.18485556e-01  1.02491967e-01
  -4.03711989e-01 -1.45492289e-02  8.03478445e-02 -5.18683400e-02
  -2.73128469e-01  6.17274818e-01  1.51742110e-01 -2.18838802e-02
  -8.93515563e-02]
 [-1.54640962e-01  4.17673774e-01  6.13929764e-02 -1.00412335e-01
  -4.34543659e-02  4.34542349e-02  2.50763629e-02 -7.88896442e-02
   5.94419181e-02 -2.08471834e-02 -4.14705279e-01 -5.60363054e-01
  -8.11578181e-02 -9.91640992e-03 -5.63728817e-02  5.23622267e-01
   5.61767721e-02]
 [-2.64425045e-02  3.15087830e-01 -1.39681716e-01  1.58558487e-01
   3.02385408e-01  1.91198583e-01 -6.10423460e-02 -5.70783816e-01
  -5.60672902e-01  2.23105808e-01  9.01788964e-03  5.27313042e-02
   1.00693324e-01 -2.09515982e-02  1.92857500e-02 -1.25997650e-01
  -6.35360730e-02]
```

```
[-2.94736419e-01 -2.49643522e-01 -4.65988731e-02 -1.31291364e-01
  2.22532003e-01  3.00003910e-02 -1.08528966e-01 -9.84599754e-03
  4.57332880e-03 -1.86675363e-01  5.08995918e-02 -1.01594830e-01
  1.43220673e-01 -3.83544794e-02 -3.40115407e-02  1.41856014e-01
 -8.23443779e-01]
[-2.49030449e-01 -1.37808883e-01 -1.48967389e-01 -1.84995991e-01
  5.60919470e-01 -1.62755446e-01 -2.09744235e-01  2.21453442e-01
 -2.75022548e-01 -2.98324237e-01  1.14639620e-03  2.59293381e-02
 -3.59321731e-01 -3.40197083e-03 -5.84289756e-02  6.97485854e-02
  3.54559731e-01]
[-6.47575181e-02  5.63418434e-02 -6.77411649e-01 -8.70892205e-02
 -1.27288825e-01 -6.41054950e-01  1.49692034e-01 -2.13293009e-01
  1.33663353e-01  8.20292186e-02  7.72631963e-04 -2.88282896e-03
  3.19400370e-02  9.43887925e-03 -6.68494643e-02 -1.14379958e-02
 -2.81593679e-02]
[ 4.25285386e-02  2.19929218e-01 -4.99721120e-01  2.30710568e-01
 -2.22311021e-01  3.31398003e-01 -6.33790064e-01  2.32660840e-01
  9.44688900e-02 -1.36027616e-01 -1.11433396e-03  1.28904022e-02
 -1.85784733e-02  3.09001353e-03  2.75286207e-02 -3.94547417e-02
 -3.92640266e-02]
[-3.18312875e-01  5.83113174e-02  1.27028371e-01  5.34724832e-01
  1.40166326e-01 -9.12555212e-02  1.09641298e-03  7.70400002e-02
  1.85181525e-01  1.23452200e-01  1.38133366e-02 -2.98075465e-02
  4.03723253e-02  1.12055599e-01 -6.91126145e-01 -1.27696382e-01
  2.32224316e-02]
[-3.17056016e-01  4.64294477e-02  6.60375454e-02  5.19443019e-01
  2.04719730e-01 -1.54927646e-01  2.84770105e-02  1.21613297e-02
  2.54938198e-01  8.85784627e-02  6.20932749e-03  2.70759809e-02
 -5.89734026e-02 -1.58909651e-01  6.71008607e-01  5.83134662e-02
  1.64850420e-02]
[ 1.76957895e-01  2.46665277e-01  2.89848401e-01  1.61189487e-01
 -7.93882496e-02 -4.87045875e-01 -2.19259358e-01  8.36048735e-02
 -2.74544380e-01 -4.72045249e-01 -2.22215182e-03  2.12476294e-02
  4.45000727e-01  2.08991284e-02  4.13740967e-02  1.77152700e-02
 -1.10262122e-02]
[-2.05082369e-01 -2.46595274e-01  1.46989274e-01 -1.73142230e-02
 -2.16297411e-01  4.73400144e-02 -2.43321156e-01 -6.78523654e-01
  2.55334907e-01 -4.22999706e-01 -1.91869743e-02 -3.33406243e-03
 -1.30727978e-01  8.41789410e-03 -2.71542091e-02 -1.04088088e-01
  1.82660654e-01]
```

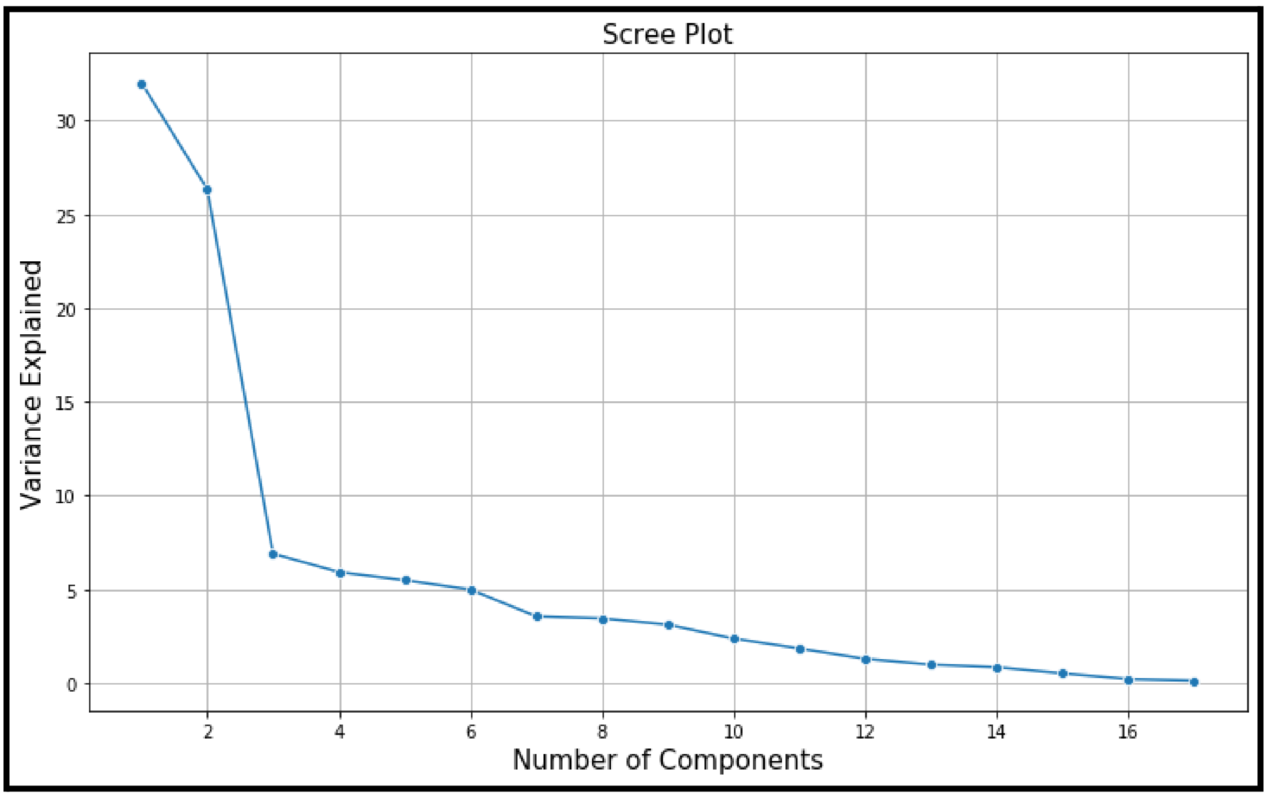
```
[-3.18908750e-01 -1.31689865e-01 -2.26743985e-01 -7.92734946e-02
  7.59581203e-02  2.98118619e-01  2.26584481e-01  5.41593771e-02
  4.91388809e-02 -1.32286331e-01 -3.53098218e-02  4.38803230e-02
  6.92088870e-01  2.27742017e-01  7.31225166e-02  9.37464497e-02
  3.25982295e-01]
[-2.52315654e-01 -1.69240532e-01  2.08064649e-01 -2.69129066e-01
 -1.09267913e-01 -2.16163313e-01 -5.59943937e-01  5.33553891e-03
 -4.19043052e-02  5.90271067e-01 -1.30710024e-02  5.00844705e-03
  2.19839000e-01  3.39433604e-03  3.64767385e-02  6.91969778e-02
  1.22106697e-01]]
```

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

In multivariate statistics, a scree plot is a line plot of the eigenvalues of factors or principal components in an analysis. The scree plot is used to determine the number of factors to retain in an exploratory factor analysis (FA) or principal components to keep in a principal component analysis (PCA). [9]

The below is the scree plot for the given data set-

GRAPH 8



There were originally 17 dimensions. After performing PCA, there are now 8 PCA components.

The below is the sample of the data after performing PCA-

FIGURE 15

pc_expenditure	pc_students	pc_books	pc_faculty	pc_top_schools	pc_s.f.ratio	pc_grad.rate	pc_alumni
-1.592855	0.767334	-0.101074	-0.921749	-0.743975	-0.298306	0.638443	-0.879386
-2.192402	-0.578830	2.278798	3.588918	1.059997	-0.177137	0.236753	0.046925
-1.430964	-1.092819	-0.438093	0.677241	-0.369613	-0.960592	-0.248276	0.308740
2.855557	-2.630612	0.141722	-1.295486	-0.183837	-1.059508	-1.249356	-0.147694
-2.212008	0.021631	2.387030	-1.114538	0.684451	0.004918	-2.159220	-0.624413
-0.571665	-1.496325	0.024354	0.066944	-0.376261	-0.668344	-1.609835	-0.529391
0.241952	-1.506368	0.234194	-1.142024	1.546983	-0.009995	0.590933	-0.329858
1.750474	-1.461412	-1.026589	-0.981184	0.217044	0.222924	0.038169	0.173929
0.769127	-1.984433	-1.426052	-0.071424	0.586380	-0.655179	-0.213314	-0.275114
-2.770721	-0.844611	1.627987	1.705091	-1.019826	-0.794401	-0.317891	-0.160687

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]

An equation that can be written in the form ax + by = c is called a linear equation. This is the standard form of a linear equation in two variables x and y. [10]

The below is general form of a first PC linear equation-

$$\boxed{PC_1 = W_1(Y_1) + W_2(Y_2) + ... + W_{17}(Y_{17})}$$

PC_1: First Principal Component

W_i: PCA component loadings *(i = 1,2, ...,17)*

Y_i: Features *(i = 1,2, ...,17)*

As there are 17 features in the given data set, the equation extends upto 17.

The below is first PC linear equation for the given data set (two places of decimals)-

$$PC_1 = 0.24\,(Apps) + 0.20\,(Accept) + 0.17\,(Enroll) + 0.35\,(Top10perc) + 0.34\,(To$$

$$+ 0.15\,(F.Undergrad) + 0.02\,(P.Undergrad) + 0.29\,(Outstate) + 0.24\,(Room.Bo$$

$$+ 0.06\,(Books) - 0.04\,(Personal) + 0.31\,(PhD) + 0.31\,(Terminal) - 0.17\,(S.F.Ra$$

$$+ 0.20\,(perc.alumni) + 0.31\,(Expend) + 0.25(Grad.Rate)$$

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

The cumulative variance gives the percentage of variance accounted for by the first 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. [11]

The below table gives the cumulative variance-

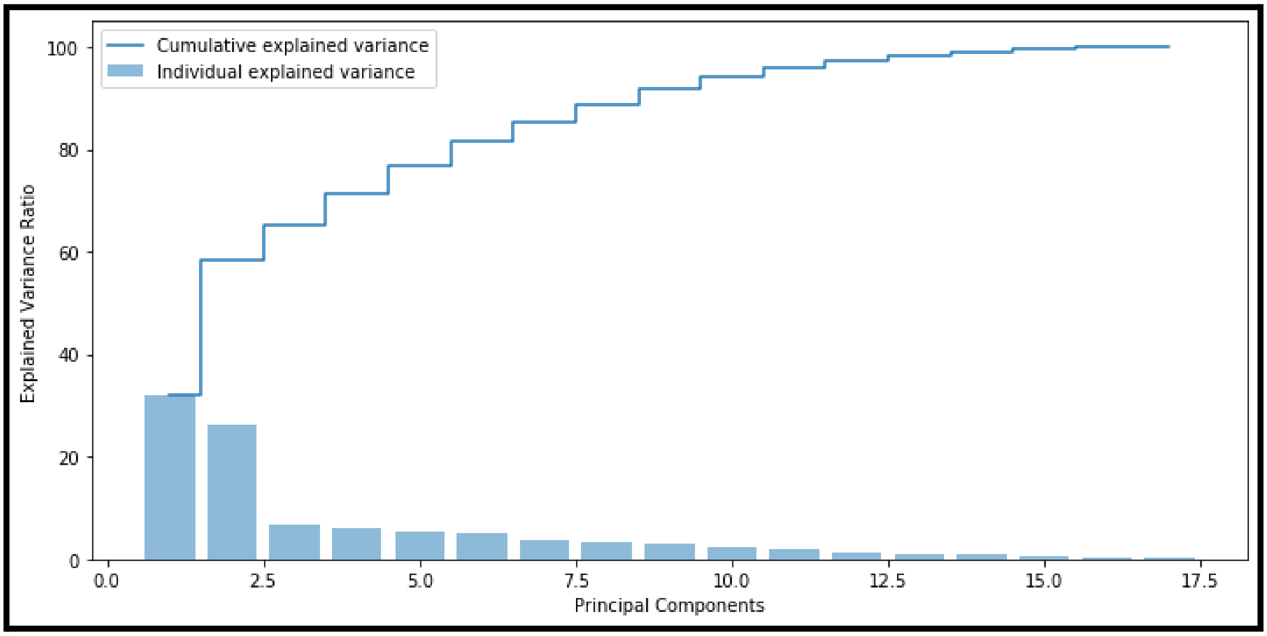
TABLE 10

Component	Cumulative Variance (%)
1	32.0206282
2	58.36084263
3	65.26175919
4	71.18474841
5	76.67315352
6	81.65785448
7	85.21672597
8	88.67034731
9	91.78758099
10	94.16277251
11	96.00419883
12	97.30024023
13	98.28599436
14	99.13183669
15	99.64896227

16	99.86471628
17	100

The graph below explains how much variance is covered within each component. This will help us reduce the number of dimensions while retaining the original features.

GRAPH 9



There are 17 components in the given data set. After the 8th component, the variance reduces to a very small number. Therefore we can consider the first 8 components for dimension reduction. Thus, cumulative variance helps to decide on the optimum number of principal components.

Eigenvectors are a special set of vectors associated with a linear system of equations (i.e., a matrix equation) that are sometimes also known as characteristic vectors, proper vectors, or latent vectors. [12]

Eigenvectors represent *direction* or *magnitude*. An individual Eigenvector is a particular “direction” in the scatterplot of data.

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]

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

The below correlation matrix shows the correlations between the PCs and the constituent variables,

GRAPH 10

The image below is the sample data of the reduced data set-

FIGURE 16

Names	pc_expenditure	pc_students	pc_books	pc_faculty	pc_top_schools	pc_s.f.ratio	pc_grad.rate	pc_alumni
Abilene Christian University	-1.592855	0.767334	-0.101074	-0.921749	-0.743975	-0.298306	0.638443	-0.879386
Adelphi University	-2.192402	-0.578830	2.278798	3.588918	1.059997	-0.177137	0.236753	0.046925
Adrian College	-1.430964	-1.092819	-0.438093	0.677241	-0.369613	-0.960592	-0.248276	0.308740
Agnes Scott College	2.855557	-2.630612	0.141722	-1.295486	-0.183837	-1.059508	-1.249356	-0.147694
Alaska Pacific University	-2.212008	0.021631	2.387030	-1.114538	0.684451	0.004918	-2.159220	-0.624413
Albertson College	-0.571665	-1.496325	0.024354	0.066944	-0.376261	-0.668344	-1.609835	-0.529391
Albertus Magnus College	0.241952	-1.506368	0.234194	-1.142024	1.546983	-0.009995	0.590933	-0.329858
Albion College	1.750474	-1.461412	-1.026589	-0.981184	0.217044	0.222924	0.038169	0.173929
Albright College	0.769127	-1.984433	-1.426052	-0.071424	0.586380	-0.655179	-0.213314	-0.275114
Alderson-Broaddus College	-2.770721	-0.844611	1.627987	1.705091	-1.019826	-0.794401	-0.317891	-0.160687

With the dimensions reduced, it will be easy for any algorithm to process the data. We can find the college with the most graduation rate and the one with the least. This can also be used to improvise the donations based on the expense.

There were 17 different variables in the original data set. The application of PCA has reduced the dimensions to 8 which is able to explain 88% of variance in the data.

Unsupervised learning like clustering can further be applied on the data to segment the colleges based on the components created and further analyzed.

References

Websites-

[1] https://www.statisticshowto.com/tukey-test-honest-significant-difference/

[2] https://www.statisticshowto.com/univariate/

[3] https://www.spss-tutorials.com/skewness/

[4] https://en.wikipedia.org/wiki/Bivariate_analysis

[5] https://pythonbasics.org/seaborn-pairplot/

[6] https://en.wikipedia.org/wiki/Feature_scaling

[7] https://en.wikipedia.org/wiki/Covariance_matrix

[8] https://www.displayr.com/what-is-a-correlation-matrix/

[9] https://en.wikipedia.org/wiki/Scree_plot

[10] https://www.cuemath.com/algebra/linear-equations/

[11] https://www.ibm.com/docs/en/spss-statistics/23.0.0?topic=reduction-total-variance-explained

[12] https://mathworld.wolfram.com/Eigenvector.html

End of Project