Problem 1:

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

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

Ans)

For "A" Compound:

H_0: $\mu_1 = \mu_2 = \mu_3$. The mean hours of relief provided by all the 3 levels of compound A are the same.

H_A: Not all population means are equal, for at least one pair of compound A.

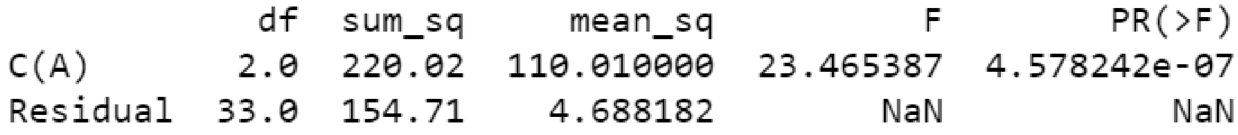
For "B" Compound:

H_0: $\mu_1 = \mu_2 = \mu_3$. The mean hours of relief provided by all the 3 levels of compound B are the same.

H_A: Not all population means are equal, for at least one pair of compound B.

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

Ans)



```
            df   sum_sq    mean_sq          F        PR(>F)
C(A)       2.0   220.02  110.010000  23.465387  4.578242e-07
Residual  33.0   154.71    4.688182        NaN           NaN
```

F-critical value for degrees of freedom 2(numerator), 33(denominator) and significance level 0.05: 3.285

Decision rule:

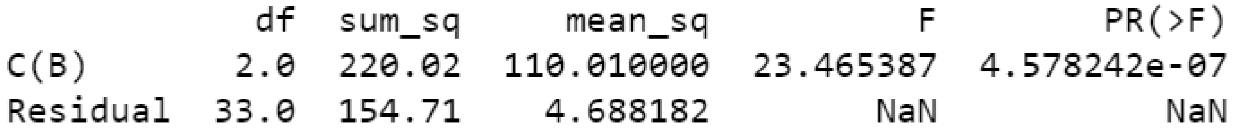
If p-value < significance level (0.05) or ($F \geq 3.285$); Reject the null hypothesis

If p-value > significance level (0.05) or ($F < 3.285$); Fail to reject the null hypothesis

Since the p value is less than the significance level (0.05), we can reject the null hypothesis, we can say with 95% confidence that the mean relief hours are not equal for all the three levels of compound A and different levels provide different levels of relief time to patients.

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

Ans)



```
             df   sum_sq    mean_sq          F        PR(>F)
C(B)        2.0   220.02  110.010000  23.465387  4.578242e-07
Residual   33.0   154.71    4.688182        NaN           NaN
```

F-critical value for degrees of freedom 2(numerator), 33(denominator) and significance level 0.05: 3.285

Decision rule:

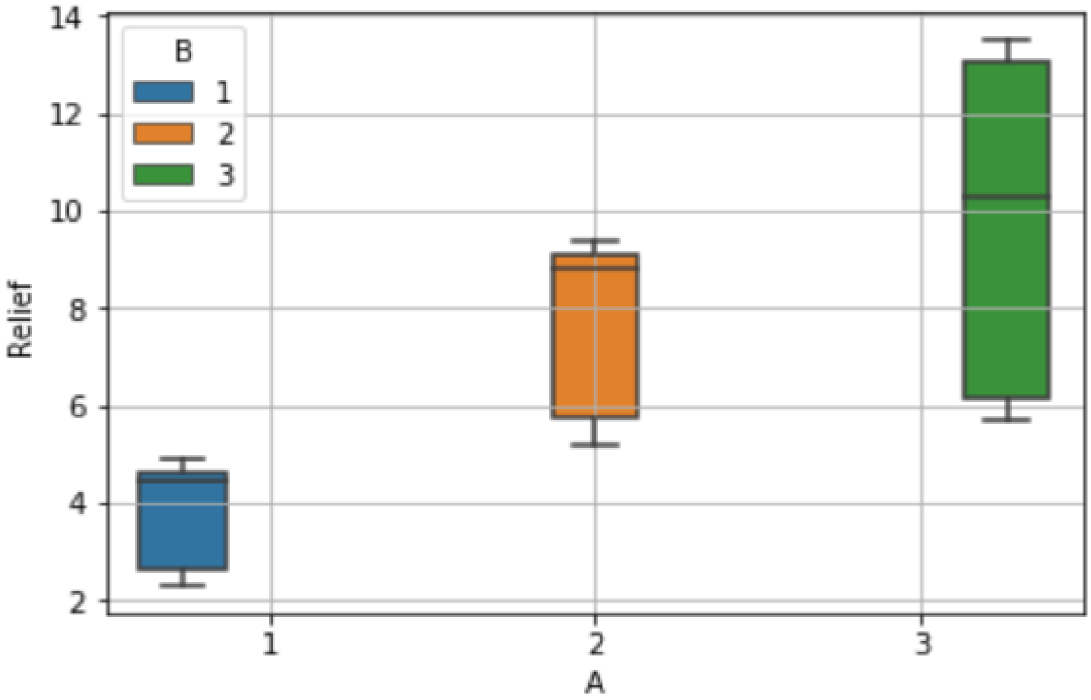
If p-value < significance level (0.05) or (F $\geq$ 3.285); Reject the null hypothesis

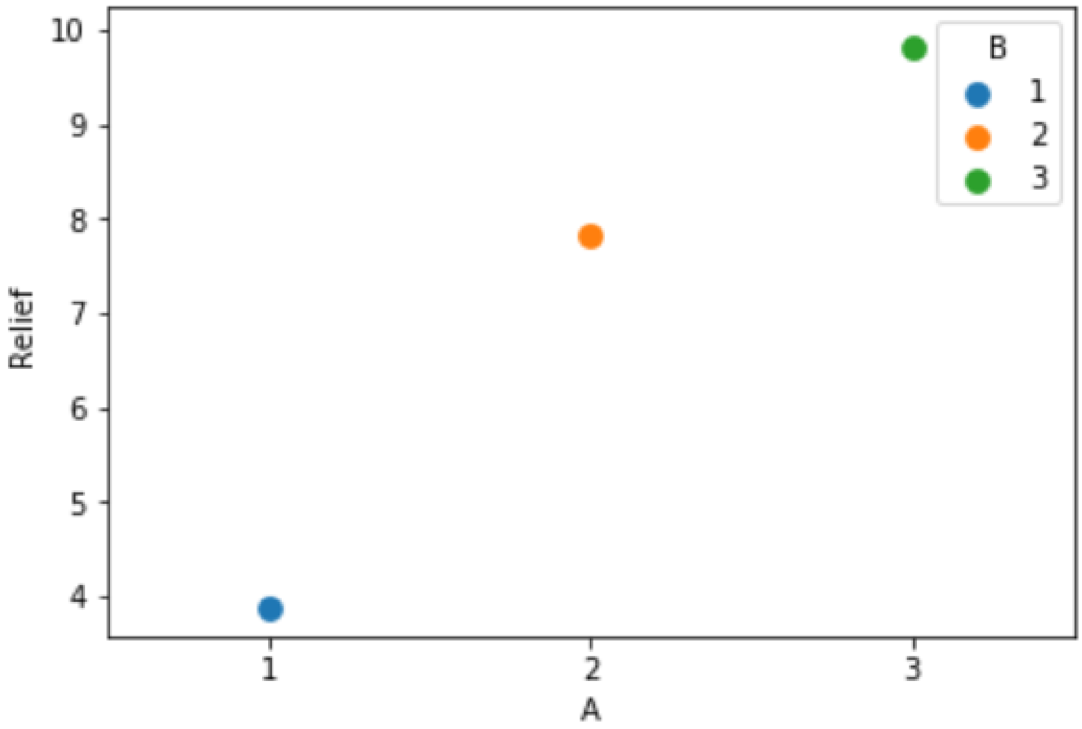
If p-value > significance level (0.05) or (F < 3.285); Fail to reject the null hypothesis

Since the p value is less than the significance level (0.05), we can reject the null hypothesis, we can say with 95% confidence that the mean relief hours is not equal for all the three levels of compound B and different levels of compound B provide different relief time to patients.

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

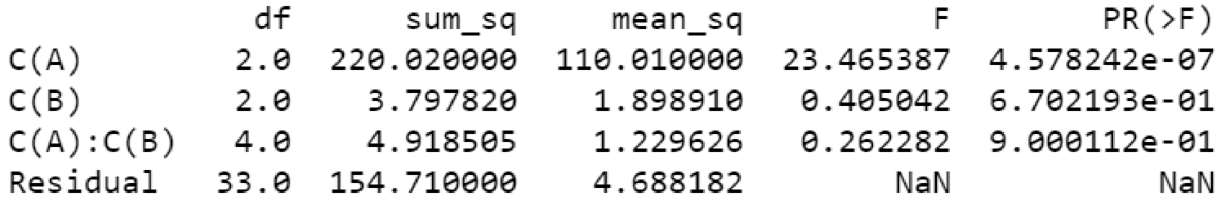
Ans)





The points are overlapping. As seen from the above two interaction plots, there seems to be no interaction amongst the two categorical variables.

By performing 2 way ANOVA and checking the interaction between A and B compounds statistically,



```
              df     sum_sq    mean_sq          F        PR(>F)
C(A)         2.0 220.020000 110.010000  23.465387  4.578242e-07
C(B)         2.0   3.797820   1.898910   0.405042  6.702193e-01
C(A):C(B)    4.0   4.918505   1.229626   0.262282  9.000112e-01
Residual    33.0 154.710000   4.688182        NaN           NaN
```

We see that p-value of interaction between A and B, i.e., 0.9 > 0.05, we failed to reject the null hypothesis (stated in question 1.5), thus there seems to be almost no statistical interaction between them. Which means there is no effect of levels of compound A on compound B and visa versa.

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

Ans)

For interaction:

H_0: The means 'Relief' time with respect to each level of compound A and B is equal. (There is an interaction effect)

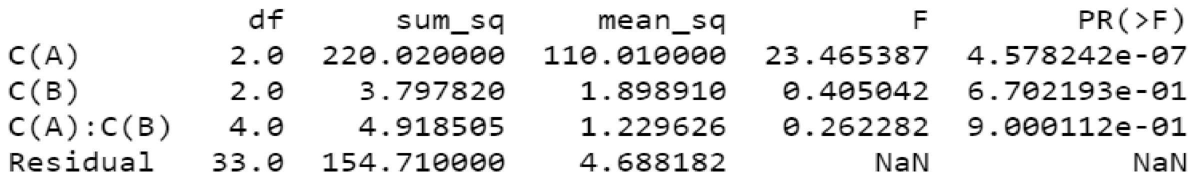
H_A: At least one of the means of 'Relief' time with respect to each level of compound A and B is unequal.

If the interaction p-value is statistically significant, then we conclude that the effect on the mean outcome of a change in one factor depends on the level of the other factor. More specifically, for at least one pair of levels of one factor the effect of a particular change in levels for the other factor depends on which level of the first pair we are focusing on.

Decision rule:

If p-value < significance level (0.05); Reject the null hypothesis (plays significant role because means are diff)

If p-value > significance level (0.05); Fail to reject the null hypothesis

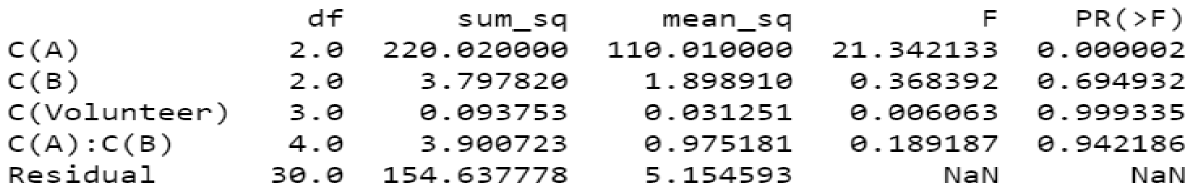


```
             df     sum_sq    mean_sq          F        PR(>F)
C(A)        2.0 220.020000 110.010000  23.465387  4.578242e-07
C(B)        2.0   3.797820   1.898910   0.405042  6.702193e-01
C(A):C(B)   4.0   4.918505   1.229626   0.262282  9.000112e-01
Residual   33.0 154.710000   4.688182        NaN           NaN
```

Here the p value for A < 0.05, which means we reject the null hypothesis for it. It plays a significant role in determining the mean relief time for the drug.

The p value for B > 0.05, which means we fail to reject the null hypothesis for it. It does not play a significant role in determining the mean relief time for the drug

The p value for interaction of A and B > 0.05, which means we fail to reject the null hypothesis for it. There is no conclusive evidence of there being an interaction between compounds A and B.



```
                df     sum_sq    mean_sq          F    PR(>F)
C(A)           2.0 220.020000 110.010000  21.342133  0.000002
C(B)           2.0   3.797820   1.898910   0.368392  0.694932
C(Volunteer)   3.0   0.093753   0.031251   0.006063  0.999335
C(A):C(B)      4.0   3.900723   0.975181   0.189187  0.942186
Residual      30.0 154.637778   5.154593        NaN       NaN
```

If we include the volunteer parameter too we can see that it doesn’t make much of a difference.

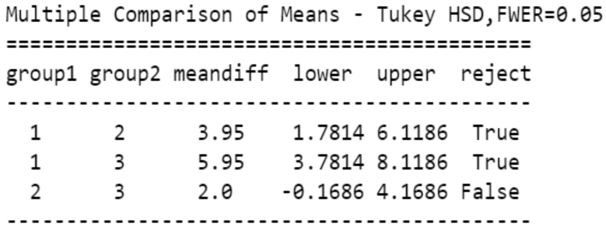
1.6) Mention the business implications of performing ANOVA for this particular case study.

Ans)

From the previous questions, we have deduced that there is not significant interaction between compounds A and B, and that the relief times of both compounds are same. We have also come to know from two way ANOVA that compound A plays a significant role in determining the mean relief time for the relief of severe cases of hay fever than compound B. Hence it should be preferred more.

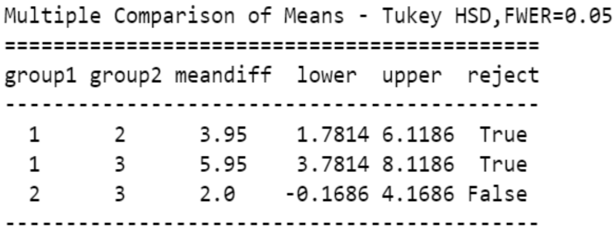
Now coming to which level provides better relief time, we run Tukey HSD test for A and B (both come to be similar). We can derive by looking at the graph that the amount of compound A at level 3 provides the most amount of relief from hay fever as compared to level 2 and 3. But statistically, by looking at the table it is said that the difference between the mean relief times of level 2 and 3 is not significant. So both the levels of concentrations can be selected.

For A

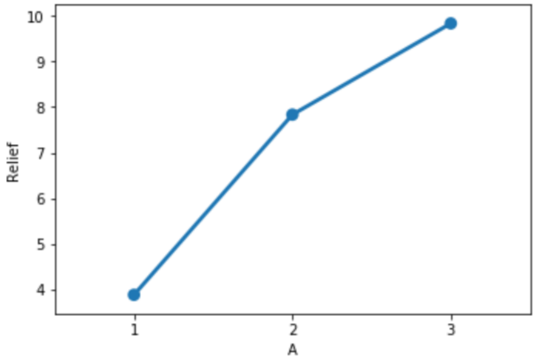


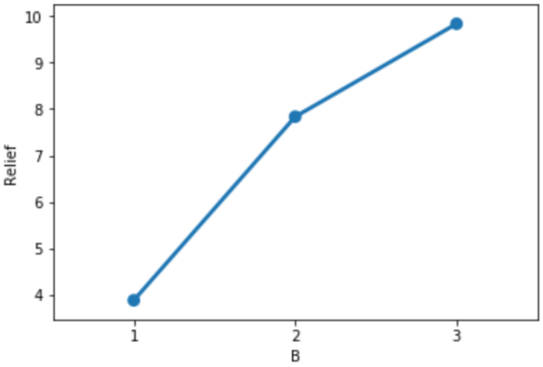
```
Multiple Comparison of Means - Tukey HSD,FWER=0.05
============================================
group1 group2 meandiff  lower  upper  reject
--------------------------------------------
  1      2      3.95    1.7814 6.1186  True
  1      3      5.95    3.7814 8.1186  True
  2      3      2.0    -0.1686 4.1686 False
--------------------------------------------
```

For B



```
Multiple Comparison of Means - Tukey HSD,FWER=0.05
============================================
group1 group2 meandiff  lower  upper  reject
--------------------------------------------
  1      2      3.95    1.7814 6.1186  True
  1      3      5.95    3.7814 8.1186  True
  2      3      2.0    -0.1686 4.1686 False
--------------------------------------------
```





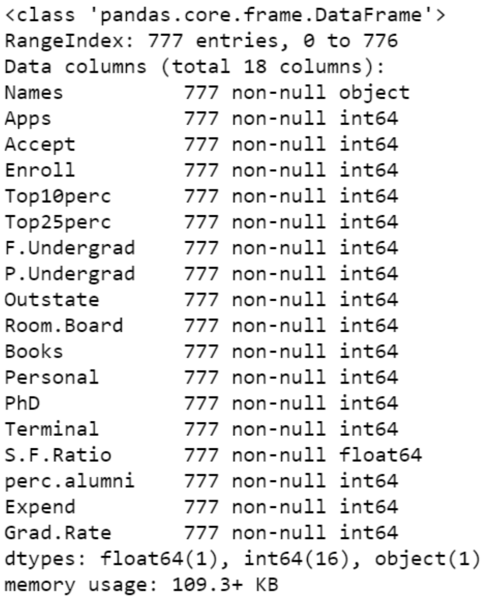
Problem 2:

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

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

Ans)

The dataset consists of data about 777 unique colleges with 18 variables providing information on it. The variables used and their data types are as follows:

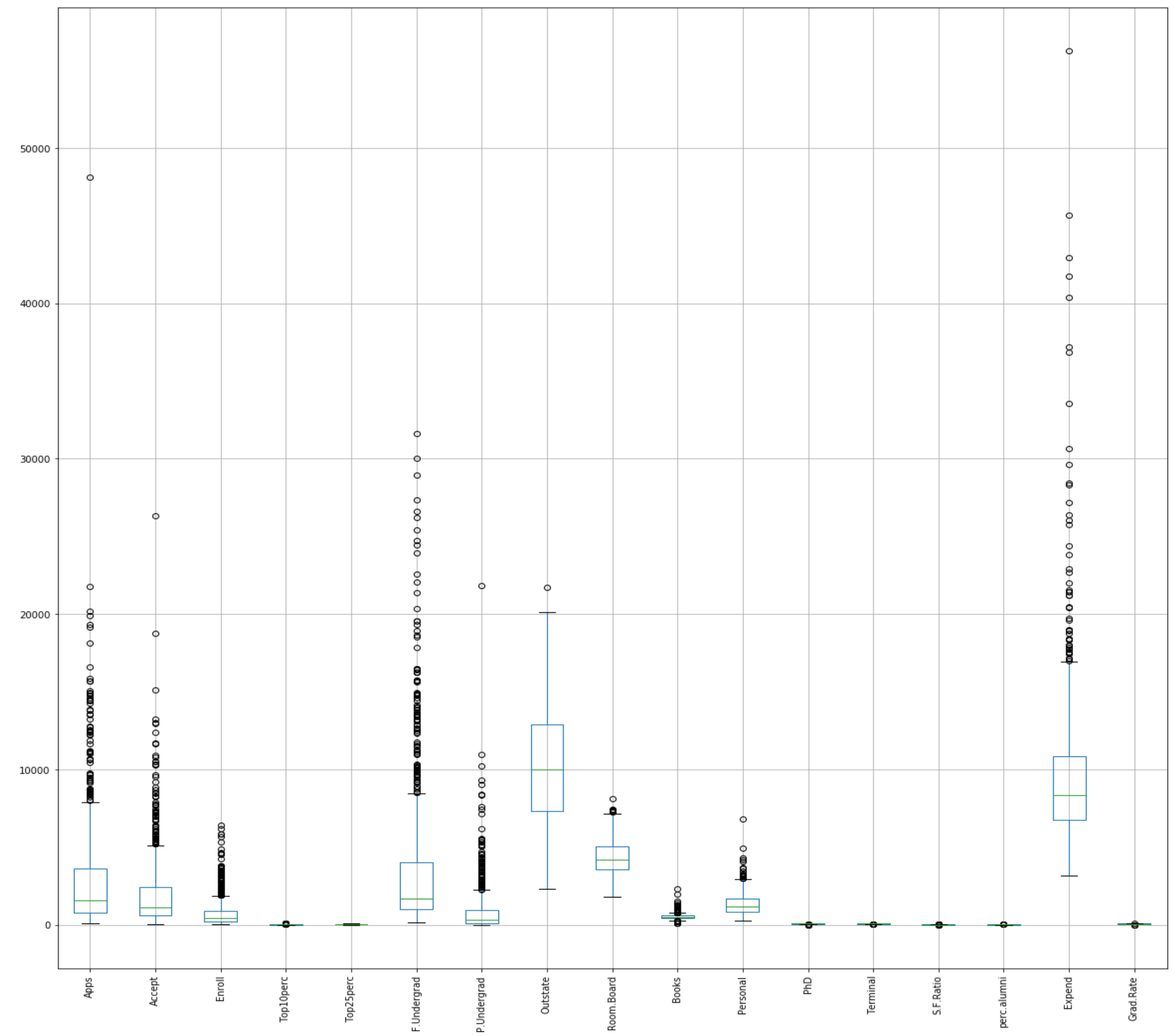


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 777 entries, 0 to 776
Data columns (total 18 columns):
Names          777 non-null object
Apps           777 non-null int64
Accept         777 non-null int64
Enroll         777 non-null int64
Top10perc      777 non-null int64
Top25perc      777 non-null int64
F.Undergrad    777 non-null int64
P.Undergrad    777 non-null int64
Outstate       777 non-null int64
Room.Board     777 non-null int64
Books          777 non-null int64
Personal       777 non-null int64
PhD            777 non-null int64
Terminal       777 non-null int64
S.F.Ratio      777 non-null float64
perc.alumni    777 non-null int64
Expend         777 non-null int64
Grad.Rate      777 non-null int64
dtypes: float64(1), int64(16), object(1)
memory usage: 109.3+ KB
```

Except names, all fields are numeric in nature, so in order to perform PCA analysis we will drop the names field.

There are **no missing values** or **duplicate values** in the dataset.

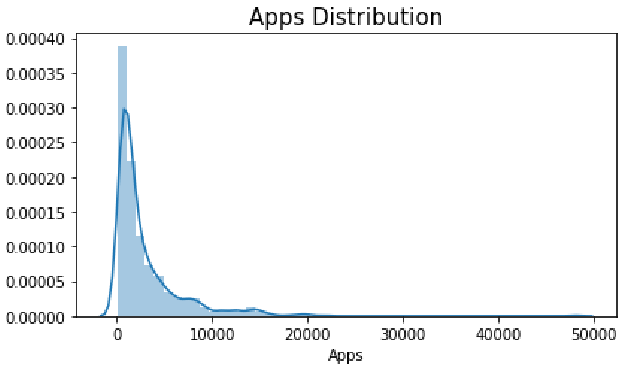
For checking the outliers, we do it using plotting a box plot.



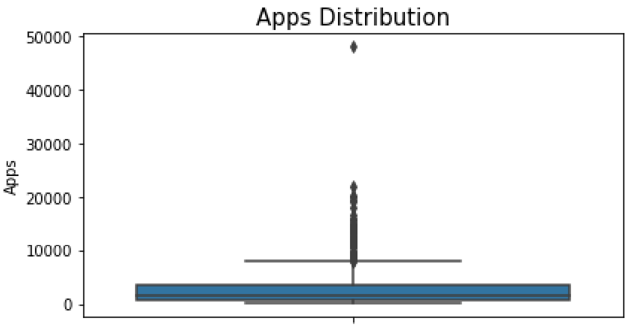
We can see that there are quite a few outliers in the dataset.

Uni-variate Analysis:

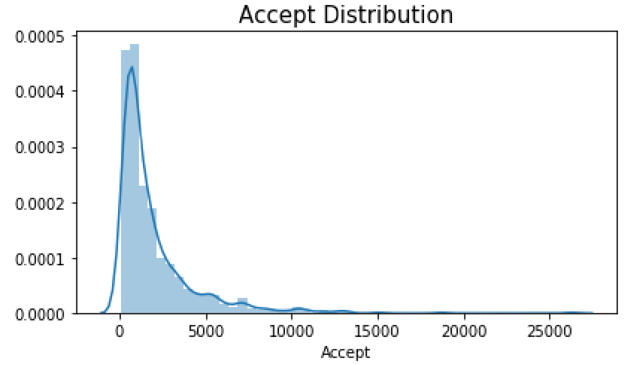
We will check the distribution of each variable by plotting their histogram and box plot.



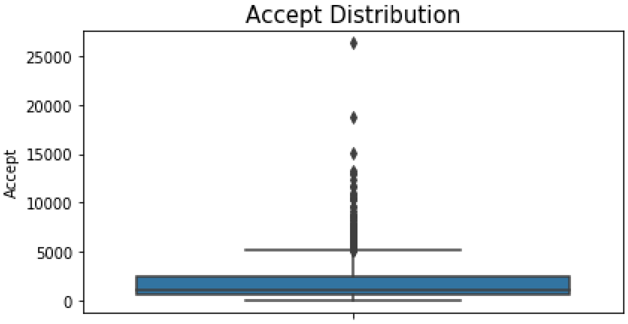
Apps Distribution



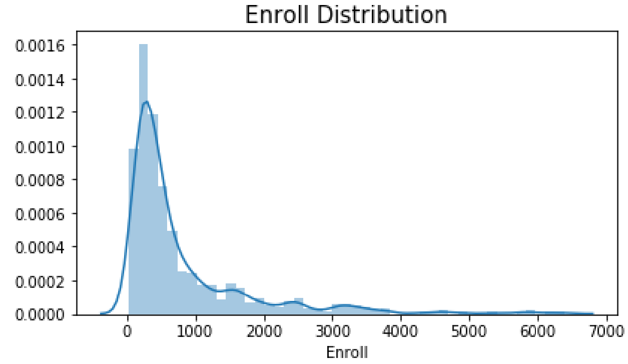
Apps Distribution



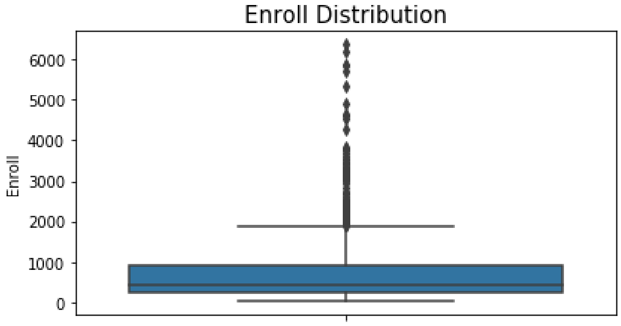
Accept Distribution



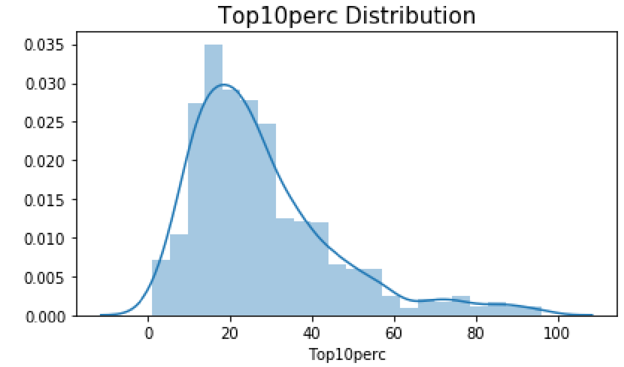
Accept Distribution



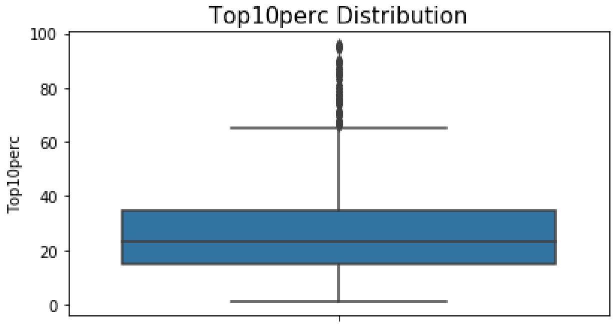
Enroll Distribution



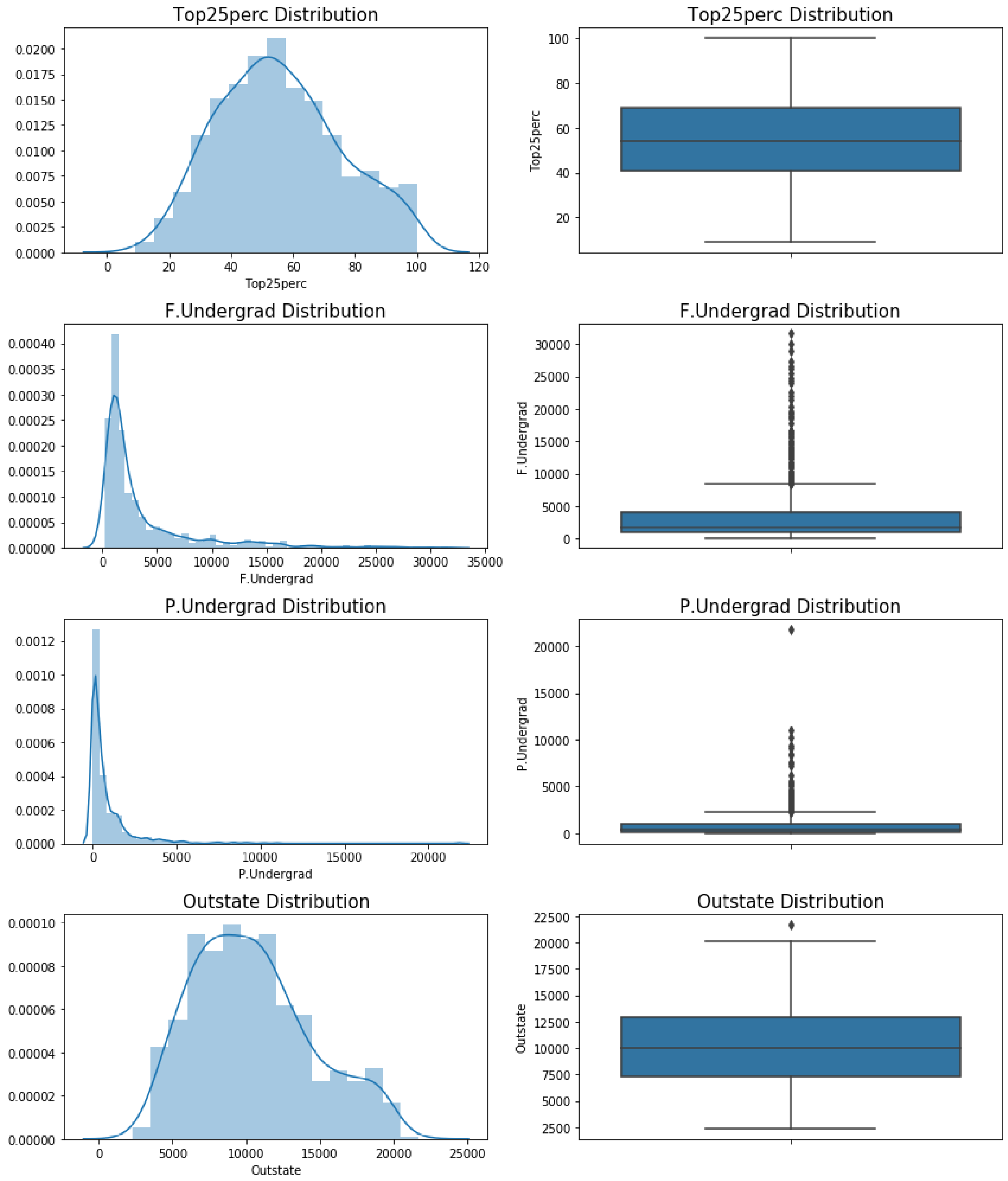
Enroll Distribution

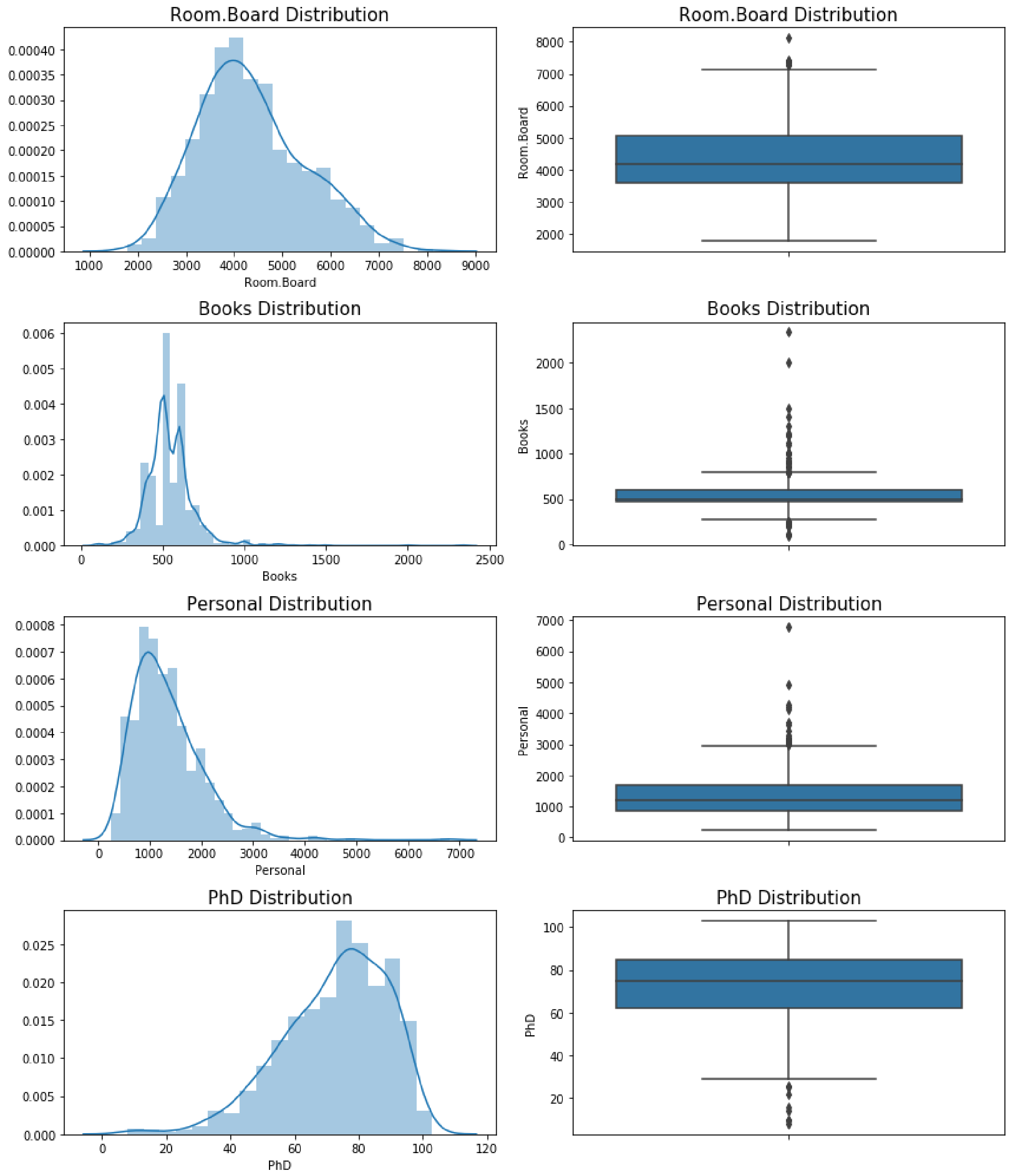


Top10perc Distribution

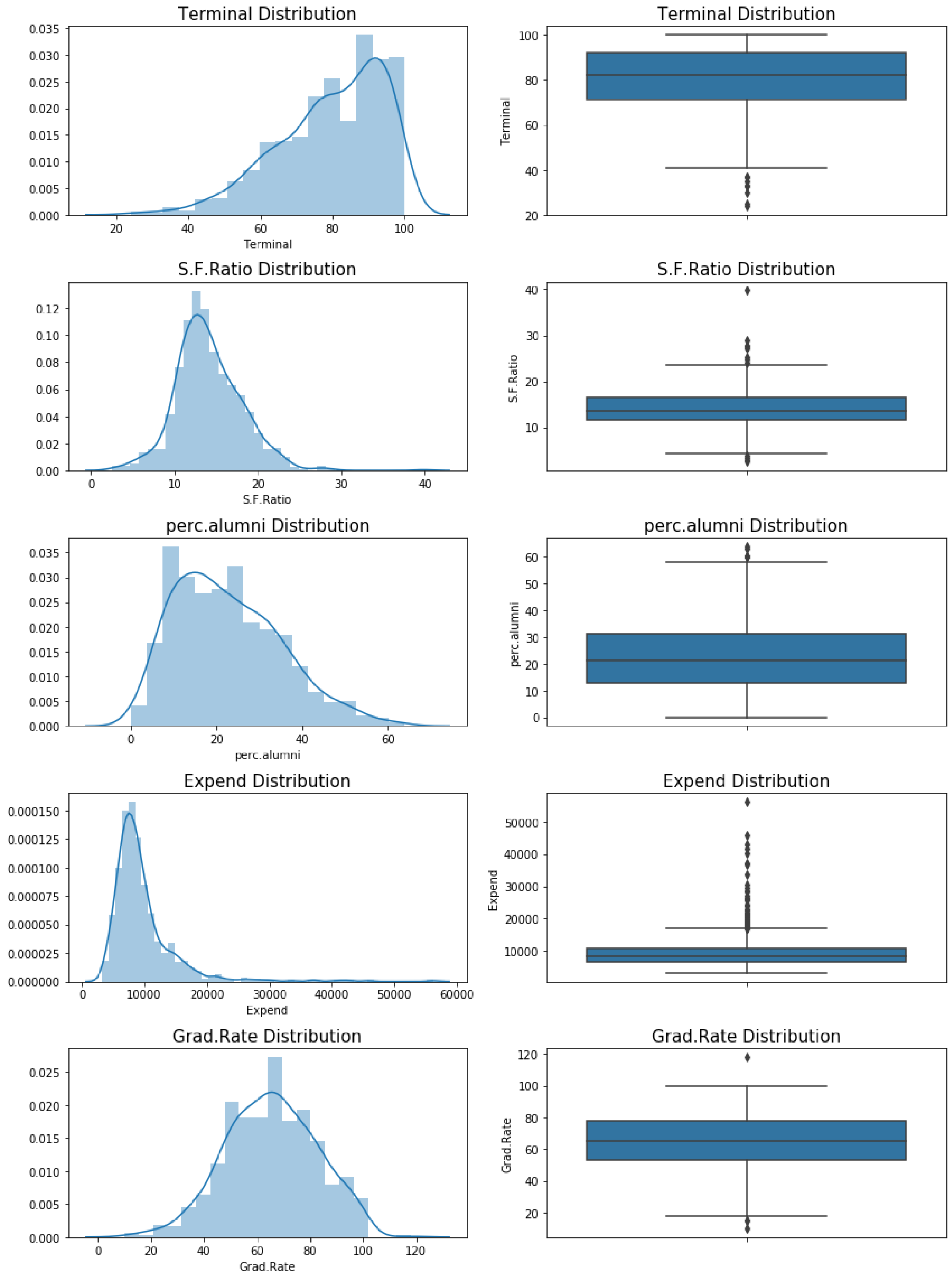


Top10perc Distribution


Top25perc Distribution
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
20
40
60
80
100
120
Top25perc
Top25perc Distribution
100
80
60
40
20
Top25perc
F.Undergrad Distribution
0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
5000
10000
15000
20000
25000
30000
35000
F.Undergrad
F.Undergrad Distribution
30000
25000
20000
15000
10000
5000
0
F.Undergrad
P.Undergrad Distribution
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
5000
10000
15000
20000
P.Undergrad
P.Undergrad Distribution
20000
15000
10000
5000
0
P.Undergrad
Outstate Distribution
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
5000
10000
15000
20000
25000
Outstate
Outstate Distribution
22500
20000
17500
15000
12500
10000
7500
5000
2500
Outstate



Room.Board Distribution
0.00000
0.00005
0.00010
0.00015
0.00020
0.00025
0.00030
0.00035
0.00040
1000
2000
3000
4000
5000
6000
7000
8000
9000
Room.Board
Room.Board Distribution
Room.Board
2000
3000
4000
5000
6000
7000
8000
Books Distribution
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0
500
1000
1500
2000
2500
Books
Books Distribution
Books
0
500
1000
1500
2000
Personal Distribution
0.0000
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
0.0007
0.0008
0
1000
2000
3000
4000
5000
6000
7000
Personal
Personal Distribution
Personal
0
1000
2000
3000
4000
5000
6000
7000
PhD Distribution
0.000
0.005
0.010
0.015
0.020
0.025
0
20
40
60
80
100
120
PhD
PhD Distribution
PhD
20
40
60
80
100



Terminal Distribution
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
20
40
60
80
100
Terminal
Terminal Distribution
100
80
60
40
20
Terminal
S.F.Ratio Distribution
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
S.F.Ratio
S.F.Ratio Distribution
40
30
20
10
S.F.Ratio
perc.alumni Distribution
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
perc.alumni
perc.alumni Distribution
60
50
40
30
20
10
0
perc.alumni
Expend Distribution
0.000150
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
0
10000
20000
30000
40000
50000
60000
Expend
Expend Distribution
50000
40000
30000
20000
10000
Expend
Grad.Rate Distribution
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
120
Grad.Rate
Grad.Rate Distribution
120
100
80
60
40
20
Grad.Rate

We observe

- By looking at the box plots that many variables have outliers within them.
- Variables like Top25perc Distribution, Outstate Distribution, Room.Board Distribution, perc.alumni Distribution and Grad.Rate Distribution appear to be almost normally distributed which also result in fewer outliers than others.
- Most of the variables are right skewed, i.e., they show positive skewness, and their tail ends in right.

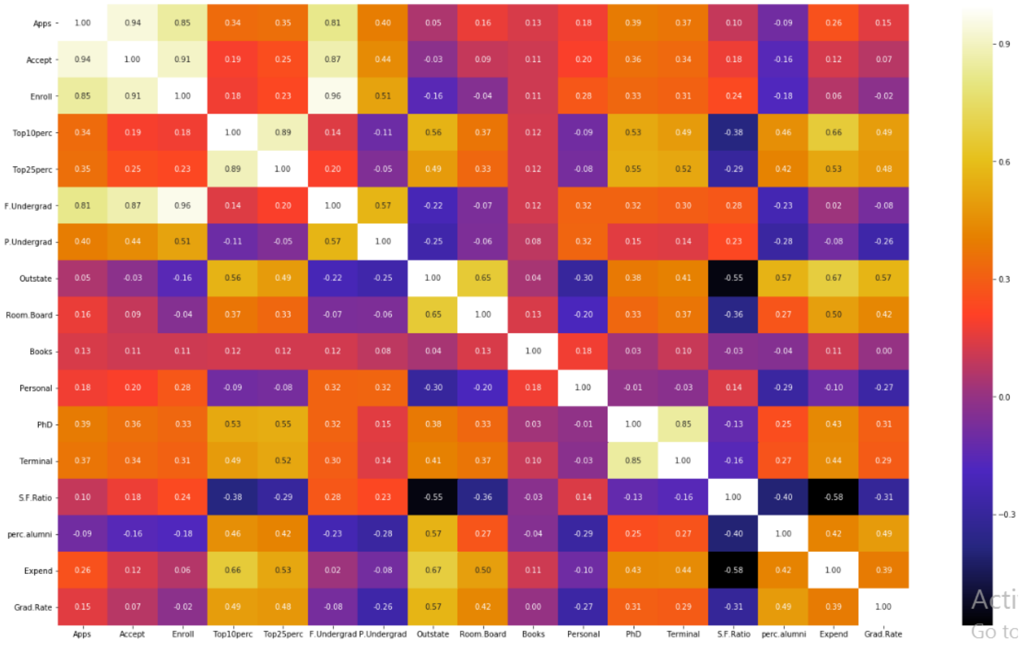
Bi-variate Analysis:

We start the bi-variate analysis by plotting a heat map which shows the correlation between variables. The darkest and the lightest colour signify high amount of correlation.

There is a high positive correlation between apps (Number of applications received), accept (Number of applications accepted), enroll (Number of new students enrolled), F.Undergrad (Number of full-time undergraduate students). We can infer that in most colleges the students enrolled majorly for full time graduation program.

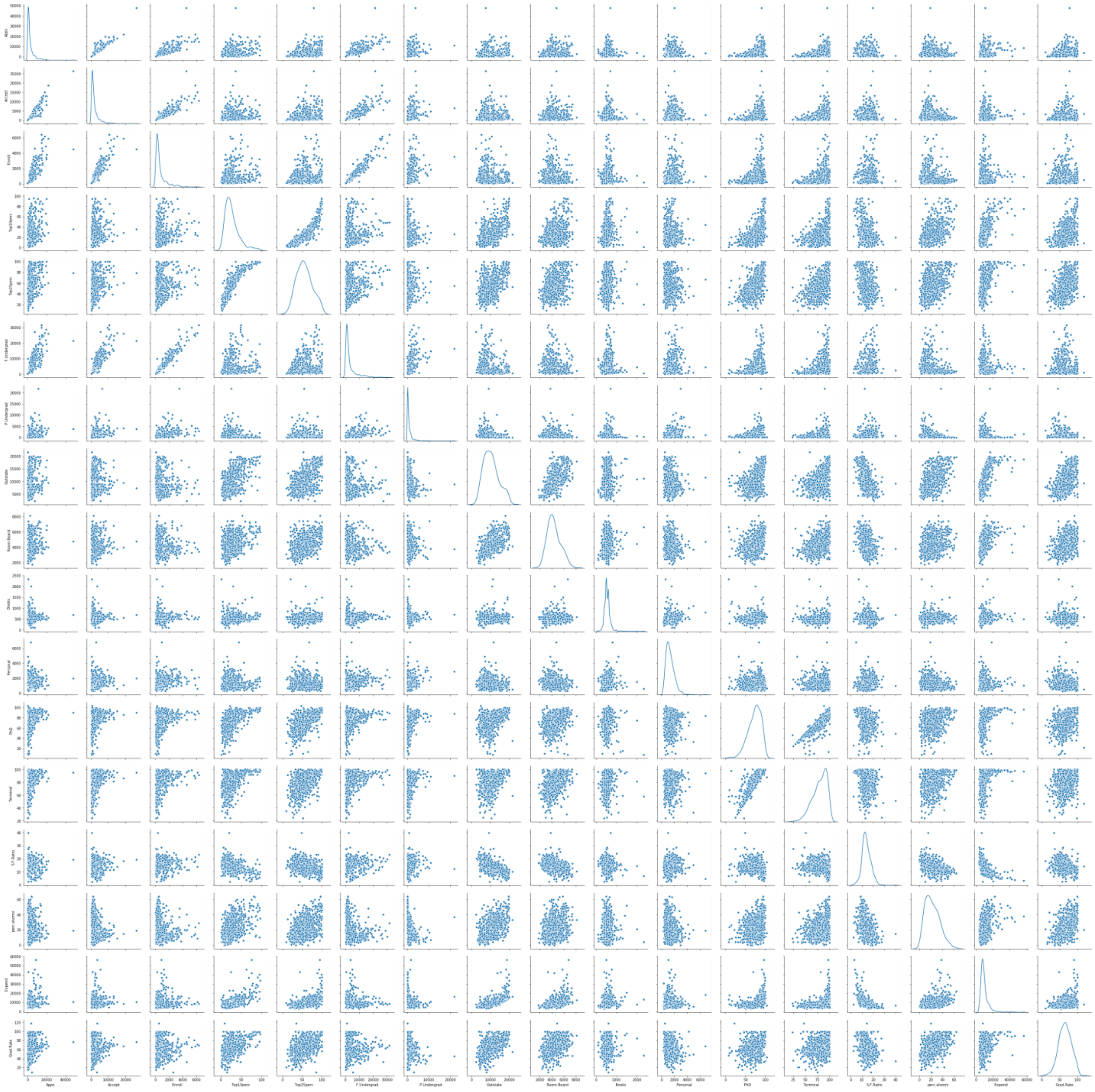
There is a high positive correlation between PhD (Percentage of faculties with Ph.D.'s) and Terminal (Percentage of faculties with terminal degree). We can infer that the faculties having PhD is their terminal degree indeed.

There is a high positive correlation between Top10perc (Percentage of new students from top 10% of Higher Secondary class) and Top25perc (Percentage of new students from top 25% of Higher Secondary class). We can infer that the new students from top 25% also belong to the top 10% of Higher Secondary class.

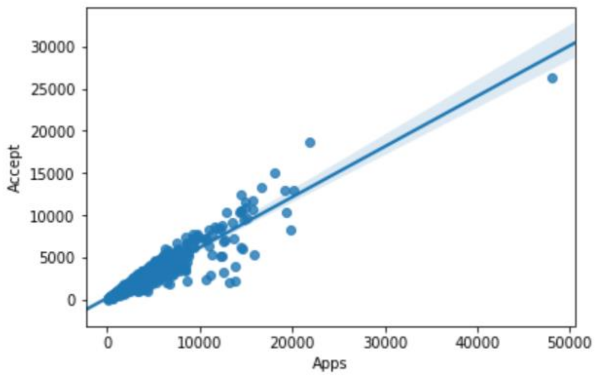


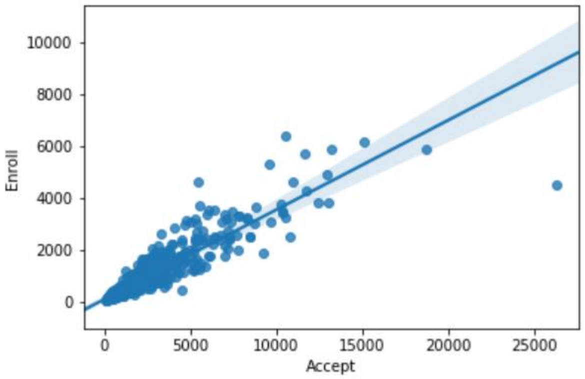
	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.00	0.94	0.85	0.34	0.35	0.81	0.40	0.05	0.16	0.13	0.18	0.39	0.37	0.10	-0.09	0.26	0.15
Accept	0.94	1.00	0.91	0.19	0.25	0.87	0.44	-0.03	0.09	0.11	0.20	0.36	0.34	0.18	-0.16	0.12	0.07
Enroll	0.85	0.91	1.00	0.18	0.23	0.96	0.51	-0.16	-0.04	0.11	0.28	0.33	0.31	0.24	-0.18	0.06	-0.02
Top10perc	0.34	0.19	0.18	1.00	0.89	0.14	-0.11	0.56	0.37	0.12	-0.09	0.53	0.49	-0.38	0.46	0.66	0.49
Top25perc	0.35	0.25	0.23	0.89	1.00	0.20	-0.05	0.49	0.33	0.12	-0.08	0.55	0.52	-0.29	0.42	0.53	0.48
F.Undergrad	0.81	0.87	0.96	0.14	0.20	1.00	0.57	-0.22	-0.07	0.12	0.32	0.32	0.30	0.28	-0.23	0.02	-0.08
P.Undergrad	0.40	0.44	0.51	-0.11	-0.05	0.57	1.00	-0.25	-0.06	0.08	0.32	0.15	0.14	0.23	-0.28	-0.08	-0.26
Outstate	0.05	-0.03	-0.16	0.56	0.49	-0.22	-0.25	1.00	0.65	0.04	-0.30	0.38	0.41	-0.55	0.57	0.67	0.57
Room.Board	0.16	0.09	-0.04	0.37	0.33	-0.07	-0.06	0.65	1.00	0.13	-0.20	0.33	0.37	-0.36	0.27	0.50	0.42
Books	0.13	0.11	0.11	0.12	0.12	0.12	0.08	0.04	0.13	1.00	0.18	0.03	0.10	-0.03	-0.04	0.11	0.00
Personal	0.18	0.20	0.28	-0.09	-0.08	0.32	0.32	-0.30	-0.20	0.18	1.00	-0.01	-0.03	0.14	-0.29	-0.10	-0.27
PhD	0.39	0.36	0.33	0.53	0.55	0.32	0.15	0.38	0.33	0.03	-0.01	1.00	0.85	-0.13	0.25	0.43	0.31
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	1.00	-0.16	0.27	0.44	0.29
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	1.00	-0.40	-0.58	-0.31
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	1.00	0.42	0.49
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	1.00	0.39
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	1.00

We also look at the pair plot for the dataset



For analysis purpose we try to put a regression line through following variables and observe that there exists a positive linear relationship among them. As the number of apps increases, the number of accept increases, and as the number of accept increases, the number of enroll also increases.





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

Ans)

In the given dataset, although all the values are numeric, they differ in scales. Some provide count of students like number of applications received, accepted and enrolled, some denote the cost of amenities like room.board, books, expend which account for the higher numbers in the given dataset and some are just percentage values like S.F. Ratio and grad ratio whose values will always be below hundred.

PCA effectiveness depends upon the scales of the attributes. If data is having different scales, PCA has following drawbacks:

- PCA will pick variable with highest variance rather than picking up attributes based on correlation
- Changing scales of the variables can change the PCA
- Interpreting PCA can become challenging due to presence of discrete data
- Presence of skew in data with long thick tail can impact the effectiveness of the PCA (related to point 1)

Because it's trying to capture the total variance in the set of variables, PCA requires that the input variables have similar scales of measurement. Variables whose numbers are just larger will have much bigger variance just because the numbers are so big. So before starting with a covariance matrix, it's a good idea to standardize those variables before we begin so that the variables with the biggest scales don't overwhelm the PCA.

Standardization transforms the data such that the resulting distribution has a **mean of 0** and a **standard deviation of 1**. We should standardize the variables before applying PCA because it will give more emphasis to those variables having higher variances than to those variables with very low variances while identifying the right principle component.

Normalizing the data is sensitive to outliers, so if there are outliers in the data set it is a bad practice. Standardization creates a new data not bounded (unlike normalization).

Usually, the **standardization** (**Z-score normalization)** is preferred because **normalization** (**min-max scaling)** is prone to **over-fitting**.

- **Standardization** is mostly used on **unsupervised learning algorithms**. In our case, standardization is **more beneficial** than normalization.
- If you see a **bell-curve** in your data then **standardization** is more preferable. In our case we observed while doing uni-variate analysis that our data does contain bell curve formation hence **standardization** is preferred.
- If your dataset has **extremely high** or **low values** (**outliers**) then **standardization** is more preferred because usually, normalization will **compress** these values into a **small range**.It's also not influenced by maximum and minimum values in our data so if our data contains outliers it's good to go.

We thus standardize the data. Data on all the dimensions are subtracted from their means to shift the data points to the origin, i.e. the data is centered on the origins so that when the axis is rotated, the values don't change.

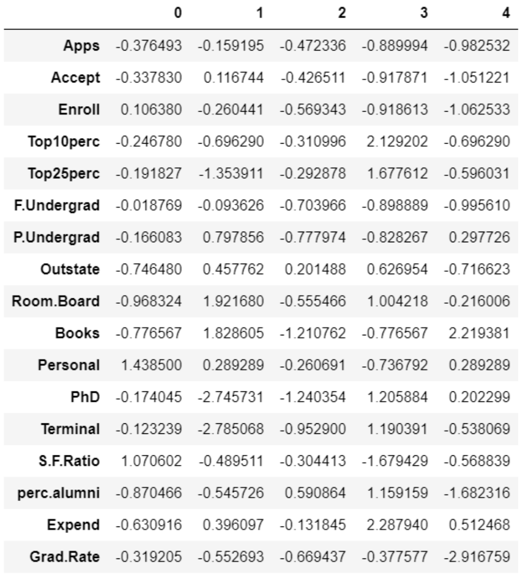
Here we use zscore method from scipy library that computes the relative Z-score of the input data, relative to the sample mean and standard deviation. It computes zscores for all columns at once.

Syntax:

scipy.stats.zscore(arr, axis=0, ddof=0) function

Parameters:
arr : *[array_like] Input array or object for which Z-score is to be calculated.*
axis : *Axis along which the mean is to be computed. By default axis = 0.*
ddof : *Degree of freedom correction for Standard Deviation.*

This is how the readings appear after scaling for first 5 observations



	0	1	2	3	4
Apps	-0.376493	-0.159195	-0.472336	-0.889994	-0.982532
Accept	-0.337830	0.116744	-0.426511	-0.917871	-1.051221
Enroll	0.106380	-0.260441	-0.569343	-0.918613	-1.062533
Top10perc	-0.246780	-0.696290	-0.310996	2.129202	-0.696290
Top25perc	-0.191827	-1.353911	-0.292878	1.677612	-0.596031
F.Undergrad	-0.018769	-0.093626	-0.703966	-0.898889	-0.995610
P.Undergrad	-0.166083	0.797856	-0.777974	-0.828267	0.297726
Outstate	-0.746480	0.457762	0.201488	0.626954	-0.716623
Room.Board	-0.968324	1.921680	-0.555466	1.004218	-0.216006
Books	-0.776567	1.828605	-1.210762	-0.776567	2.219381
Personal	1.438500	0.289289	-0.260691	-0.736792	0.289289
PhD	-0.174045	-2.745731	-1.240354	1.205884	0.202299
Terminal	-0.123239	-2.785068	-0.952900	1.190391	-0.538069
S.F.Ratio	1.070602	-0.489511	-0.304413	-1.679429	-0.568839
perc.alumni	-0.870466	-0.545726	0.590864	1.159159	-1.682316
Expend	-0.630916	0.396097	-0.131845	2.287940	0.512468
Grad.Rate	-0.319205	-0.552693	-0.669437	-0.377577	-2.916759

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

Ans)

Both covariance and correlation measure linear relationships between variables.

The covariance indicates the direction of the linear relationship between variables. A covariance of 0 indicates that two variables are totally unrelated. If the covariance is positive, the variables increase in the same direction, and if the covariance is negative, the variables change in opposite directions. The magnitude of the covariance depends on the scale of each variable. **Correlation on the other hand measures both the strength and direction of the linear relationship between two variables.** When the correlation coefficient is positive, an increase in one variable also results in an increase in the other. When the correlation coefficient is negative, the changes in the two variables are in opposite directions. When there is no relationship, there is no change in either.

The linear correlation between two features and is closely related to the covariance. In fact, correlation between two variables is a normalized version of the covariance. By dividing the covariance by the features' standard deviations, we ensure that the correlation between two features is in the range [-1, 1], which makes it more interpretable than the unbounded covariance.

In fact, it's a normalized version of the covariance However, note that **the covariance and correlation are exactly the same if the features are normalized to unit variance (e.g., via standardization or z-score normalization).** Two features are perfectly positively correlated if ρ=1ρ=1 and perfectly negatively correlated if ρ=−1ρ=−1. No correlation is observed if ρ=0ρ=0.

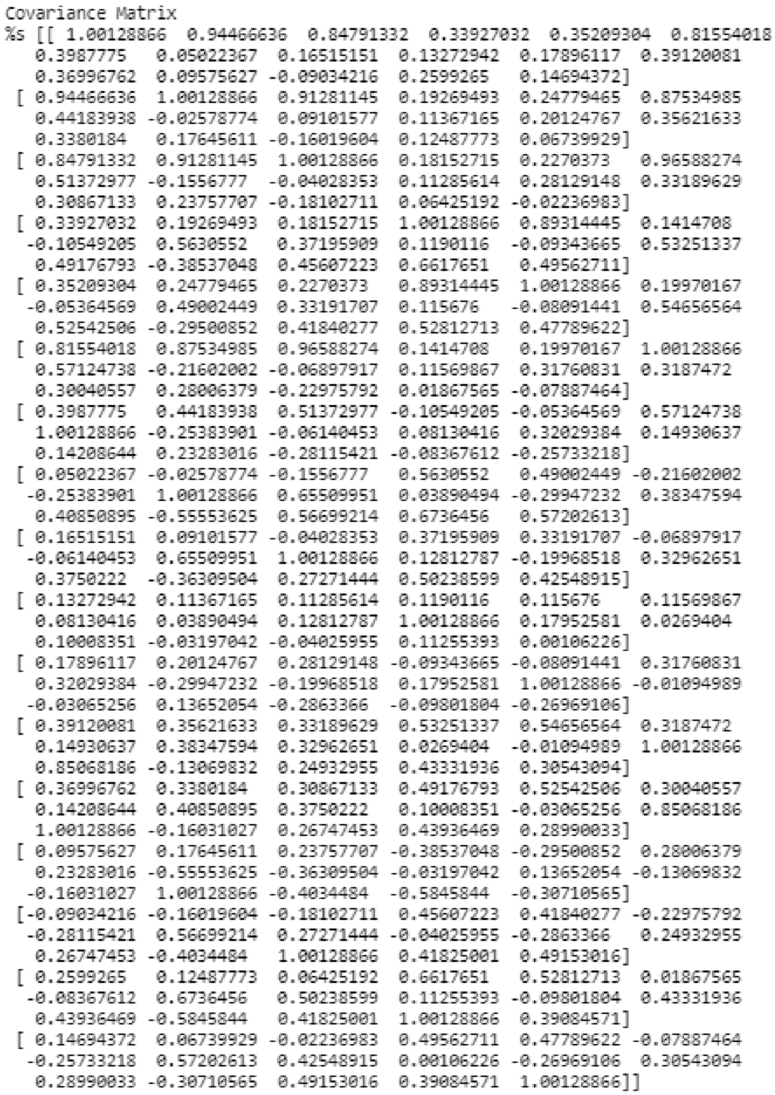
Thus we can state that above three approaches yield the same eigenvectors and eigenvalue pairs:

- Eigen decomposition of the covariance matrix after standardizing the data.
- Eigen decomposition of the correlation matrix.
- Eigen decomposition of the correlation matrix after standardizing the data.

Finally we can say that after scaling - the covariance and the correlation matrix have the same values.

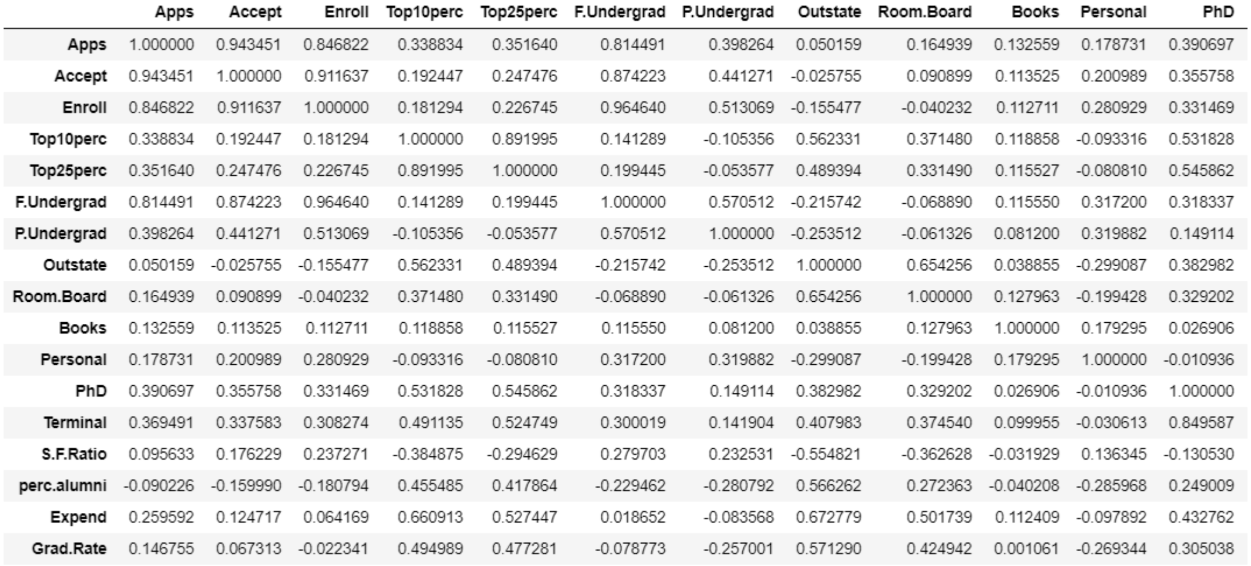
Below is comparison between covariance matrix and correlation matrix on scaled data. We can see that they have same values.

Covariance Matrix



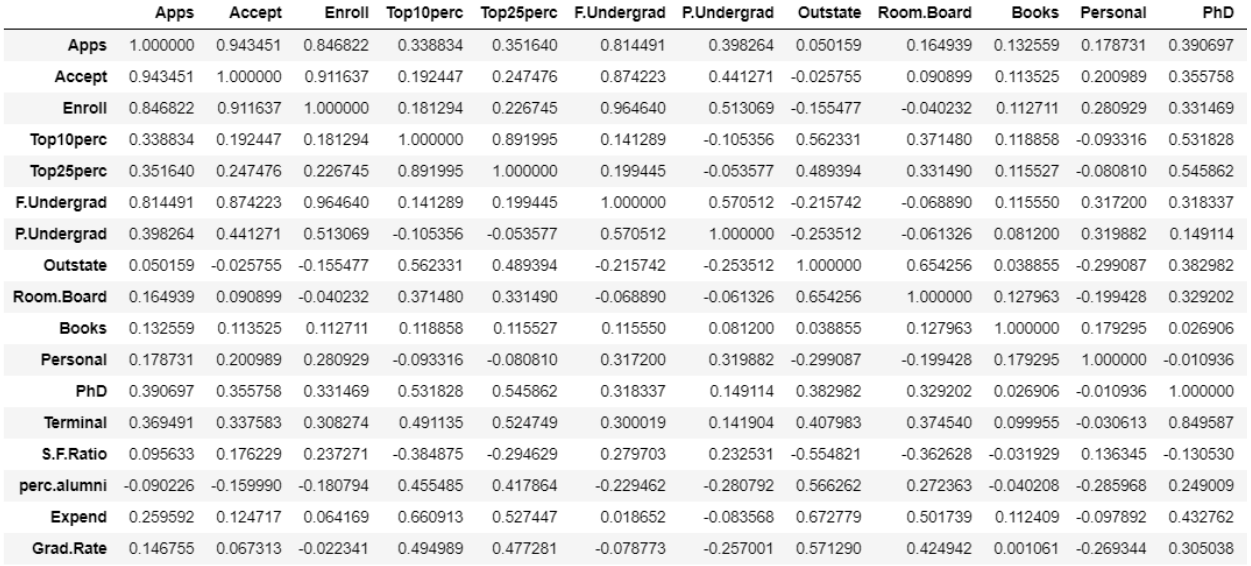
```
Covariance Matrix
%s [[ 1.00128866  0.94466636  0.84791332  0.33927032  0.35209304  0.81554018
    0.3987775   0.05022367  0.16515151  0.13272942  0.17896117  0.39120081
    0.36996762  0.09575627 -0.09034216  0.2599265   0.14694372]
  [ 0.94466636  1.00128866  0.91281145  0.19269493  0.24779465  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



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899	0.113525	0.200989	0.355758
Enroll	0.846822	0.911637	1.000000	0.181294	0.226745	0.964640	0.513069	-0.155477	-0.040232	0.112711	0.280929	0.331469
Top10perc	0.338834	0.192447	0.181294	1.000000	0.891995	0.141289	-0.105356	0.562331	0.371480	0.118858	-0.093316	0.531828
Top25perc	0.351640	0.247476	0.226745	0.891995	1.000000	0.199445	-0.053577	0.489394	0.331490	0.115527	-0.080810	0.545862
F.Undergrad	0.814491	0.874223	0.964640	0.141289	0.199445	1.000000	0.570512	-0.215742	-0.068890	0.115550	0.317200	0.318337
P.Undergrad	0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	1.000000	-0.253512	-0.061326	0.081200	0.319882	0.149114
Outstate	0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	-0.253512	1.000000	0.654256	0.038855	-0.299087	0.382982
Room.Board	0.164939	0.090899	-0.040232	0.371480	0.331490	-0.068890	-0.061326	0.654256	1.000000	0.127963	-0.199428	0.329202
Books	0.132559	0.113525	0.112711	0.118858	0.115527	0.115550	0.081200	0.038855	0.127963	1.000000	0.179295	0.026906
Personal	0.178731	0.200989	0.280929	-0.093316	-0.080810	0.317200	0.319882	-0.299087	-0.199428	0.179295	1.000000	-0.010936
PhD	0.390697	0.355758	0.331469	0.531828	0.545862	0.318337	0.149114	0.382982	0.329202	0.026906	-0.010936	1.000000
Terminal	0.369491	0.337583	0.308274	0.491135	0.524749	0.300019	0.141904	0.407983	0.374540	0.099955	-0.030613	0.849587
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	-0.130530
perc.alumni	-0.090226	-0.159990	-0.180794	0.455485	0.417864	-0.229462	-0.280792	0.566262	0.272363	-0.040208	-0.285968	0.249009
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
Grad.Rate	0.146755	0.067313	-0.022341	0.494989	0.477281	-0.078773	-0.257001	0.571290	0.424942	0.001061	-0.269344	0.305038

Correlation Matrix of original data (will be same as for scaled data as normalization of it is already taken care in it)



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899	0.113525	0.200989	0.355758
Enroll	0.846822	0.911637	1.000000	0.181294	0.226745	0.964640	0.513069	-0.155477	-0.040232	0.112711	0.280929	0.331469
Top10perc	0.338834	0.192447	0.181294	1.000000	0.891995	0.141289	-0.105356	0.562331	0.371480	0.118858	-0.093316	0.531828
Top25perc	0.351640	0.247476	0.226745	0.891995	1.000000	0.199445	-0.053577	0.489394	0.331490	0.115527	-0.080810	0.545862
F.Undergrad	0.814491	0.874223	0.964640	0.141289	0.199445	1.000000	0.570512	-0.215742	-0.068890	0.115550	0.317200	0.318337
P.Undergrad	0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	1.000000	-0.253512	-0.061326	0.081200	0.319882	0.149114
Outstate	0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	-0.253512	1.000000	0.654256	0.038855	-0.299087	0.382982
Room.Board	0.164939	0.090899	-0.040232	0.371480	0.331490	-0.068890	-0.061326	0.654256	1.000000	0.127963	-0.199428	0.329202
Books	0.132559	0.113525	0.112711	0.118858	0.115527	0.115550	0.081200	0.038855	0.127963	1.000000	0.179295	0.026906
Personal	0.178731	0.200989	0.280929	-0.093316	-0.080810	0.317200	0.319882	-0.299087	-0.199428	0.179295	1.000000	-0.010936
PhD	0.390697	0.355758	0.331469	0.531828	0.545862	0.318337	0.149114	0.382982	0.329202	0.026906	-0.010936	1.000000
Terminal	0.369491	0.337583	0.308274	0.491135	0.524749	0.300019	0.141904	0.407983	0.374540	0.099955	-0.030613	0.849587
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	-0.130530
perc.alumni	-0.090226	-0.159990	-0.180794	0.455485	0.417864	-0.229462	-0.280792	0.566262	0.272363	-0.040208	-0.285968	0.249009
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
Grad.Rate	0.146755	0.067313	-0.022341	0.494989	0.477281	-0.078773	-0.257001	0.571290	0.424942	0.001061	-0.269344	0.305038

Covariance matrix of original data

```
Covariance Matrix
%s [[ 1.49784595e+07  8.94985981e+06  3.04525599e+06  2.31327731e+04
   2.69526635e+04  1.52897025e+07  2.34662015e+06  7.80970356e+05
   7.00072872e+05  8.47037526e+04  4.68346833e+05  2.46894337e+04
   2.10530676e+04  1.46506058e+03 -4.32712238e+03  5.24617110e+06
   9.75642164e+03]
```

```
[ 8.94985981e+06  6.00795970e+06  2.07626776e+06  8.32112487e+03
  1.20134048e+04  1.03935824e+07  1.64666972e+06 -2.53962285e+05
  2.44347147e+05  4.59428079e+04  3.33556631e+05  1.42382015e+04
  1.21820938e+04  1.70983819e+03 -4.85948702e+03  1.59627169e+06
  2.83416292e+03]
[ 3.04525599e+06  2.07626776e+06  8.63368392e+05  2.97158341e+03
  4.17259244e+03  4.34752988e+06  7.25790674e+05 -5.81188483e+05
 -4.09970592e+04  1.72911997e+04  1.76737970e+05  5.02896117e+03
  4.21708603e+03  8.72684773e+02 -2.08169379e+03  3.11345431e+05
 -3.56587977e+02]
[ 2.31327731e+04  8.32112487e+03  2.97158341e+03  3.11182456e+02
  3.11630480e+02  1.20891137e+04 -2.82947498e+03  3.99071798e+04
  7.18670561e+03  3.46177405e+02 -1.11455119e+03  1.53184870e+02
  1.27551581e+02 -2.68745252e+01  9.95672077e+01  6.08793102e+04
  1.49992164e+02]
[ 2.69526635e+04  1.20134048e+04  4.17259244e+03  3.11630480e+02
  3.92229216e+02  1.91589528e+04 -1.61541214e+03  3.89924275e+04
  7.19990357e+03  3.77759266e+02 -1.08360506e+03  1.76518449e+02
  1.53002612e+02 -2.30971994e+01  1.02550946e+02  5.45464833e+04
  1.62371398e+02]
[ 1.52897025e+07  1.03935824e+07  4.34752988e+06  1.20891137e+04
  1.91589528e+04  2.35265793e+07  4.21291009e+06 -4.20984304e+06
 -3.66458224e+05  9.25357647e+04  1.04170909e+06  2.52117842e+04
  2.14242417e+04  5.37020858e+03 -1.37919297e+04  4.72403958e+05
 -6.56330753e+03]
[ 2.34662015e+06  1.64666972e+06  7.25790674e+05 -2.82947498e+03
 -1.61541214e+03  4.21291009e+06  2.31779885e+06 -1.55270428e+06
 -1.02391862e+05  2.04104467e+04  3.29732427e+05  3.70675622e+03
  3.18059661e+03  1.40130256e+03 -5.29733709e+03 -6.64351154e+05
 -6.72106249e+03]
[ 7.80970356e+05 -2.53962285e+05 -5.81188483e+05  3.99071798e+04
  3.89924275e+04 -4.20984304e+06 -1.55270428e+06  1.61846616e+07
  2.88659739e+06  2.58082421e+04 -8.14673718e+05  2.51575151e+04
  2.41641477e+04 -8.83525354e+03  2.82295531e+04  1.41332357e+07
  3.94796818e+04]
[ 7.00072872e+05  2.44347147e+05 -4.09970592e+04  7.18670561e+03
  7.19990357e+03 -3.66458224e+05 -1.02391862e+05  2.88659739e+06
  1.20274303e+06  2.31703134e+04 -1.48083768e+05  5.89503475e+03
  6.04729974e+03 -1.57420591e+03  3.70143138e+03  2.87330848e+06
  8.00536018e+03]
[ 8.47037526e+04  4.59428079e+04  1.72911997e+04  3.46177405e+02
  3.77759266e+02  9.25357647e+04  2.04104467e+04  2.58082421e+04
  2.31703134e+04  2.72597799e+04  2.00430257e+04  7.25342415e+01
  2.42963918e+02 -2.08672067e+01 -8.22631321e+01  9.69125803e+04
```

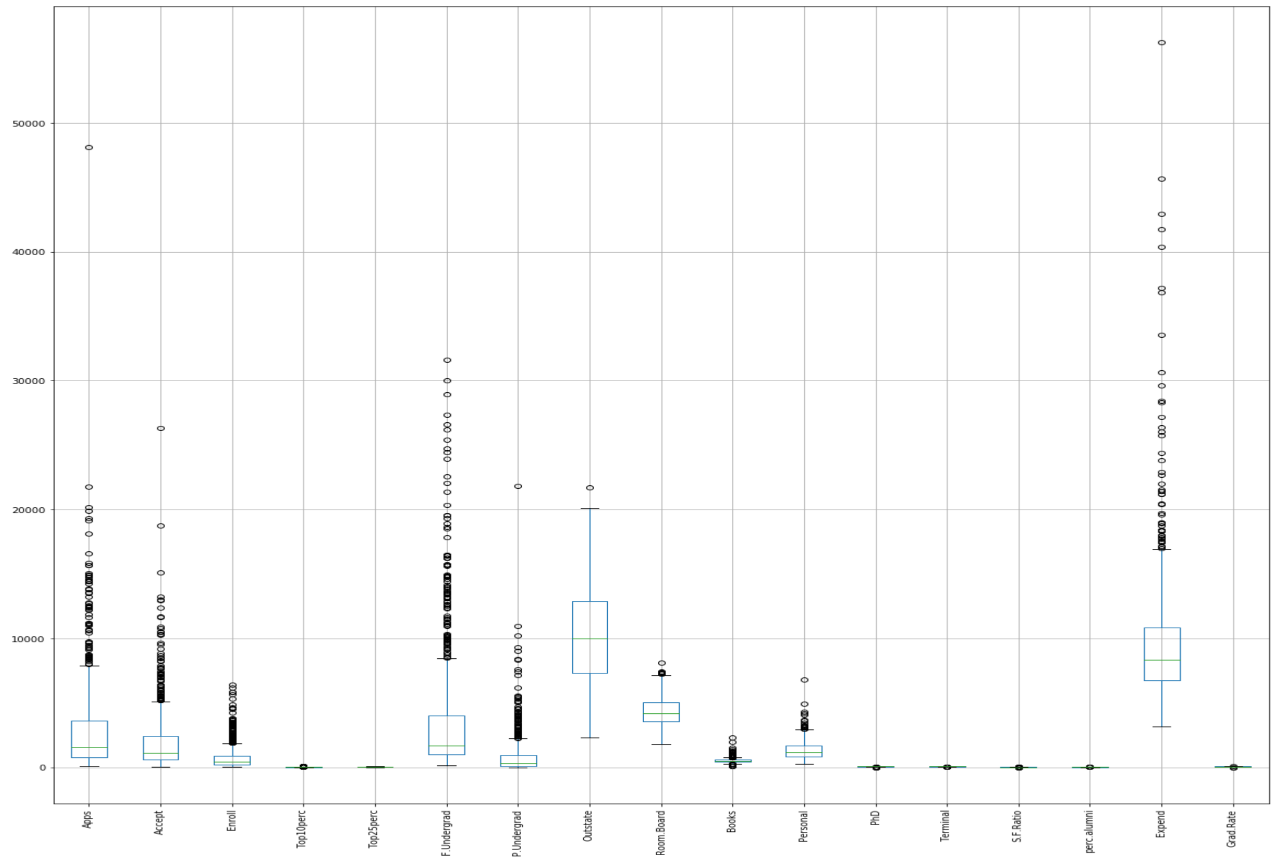
```
   3.00883652e+00]
 [ 4.68346833e+05  3.33556631e+05  1.76737970e+05 -1.11455119e+03
  -1.08360506e+03  1.04170909e+06  3.29732427e+05 -8.14673718e+05
  -1.48083768e+05  2.00430257e+04  4.58425753e+05 -1.20898783e+02
  -3.05154186e+02  3.65415770e+02 -2.39931082e+03 -3.46097802e+05
  -3.13261494e+03]
 [ 2.46894337e+04  1.42382015e+04  5.02896117e+03  1.53184870e+02
   1.76518449e+02  2.52117842e+04  3.70675622e+03  2.51575151e+04
   5.89503475e+03  7.25342415e+01 -1.20898783e+02  2.66608636e+02
   2.04231332e+02 -8.43649246e+00  5.03832295e+01  3.68980582e+04
   8.55571090e+01]
 [ 2.10530676e+04  1.21820938e+04  4.21708603e+03  1.27551581e+02
   1.53002612e+02  2.14242417e+04  3.18059661e+03  2.41641477e+04
   6.04729974e+03  2.42963918e+02 -3.05154186e+02  2.04231332e+02
   2.16747841e+02 -9.33025564e+00  4.87343271e+01  3.37334569e+04
   7.32203957e+01]
 [ 1.46506058e+03  1.70983819e+03  8.72684773e+02 -2.68745252e+01
  -2.30971994e+01  5.37020858e+03  1.40130256e+03 -8.83525354e+03
  -1.57420591e+03 -2.08672067e+01  3.65415770e+02 -8.43649246e+00
  -9.33025564e+00  1.56685279e+01 -1.97641094e+01 -1.20675646e+04
  -2.08548884e+01]
 [-4.32712238e+03 -4.85948702e+03 -2.08169379e+03  9.95672077e+01
   1.02550946e+02 -1.37919297e+04 -5.29733709e+03  2.82295531e+04
   3.70143138e+03 -8.22631321e+01 -2.39931082e+03  5.03832295e+01
   4.87343271e+01 -1.97641094e+01  1.53556744e+02  2.70289215e+04
   1.04493815e+02]
 [ 5.24617110e+06  1.59627169e+06  3.11345431e+05  6.08793102e+04
   5.45464833e+04  4.72403958e+05 -6.64351154e+05  1.41332357e+07
   2.87330848e+06  9.69125803e+04 -3.46097802e+05  3.68980582e+04
   3.37334569e+04 -1.20675646e+04  2.70289215e+04  2.72668656e+07
   3.50129683e+04]
 [ 9.75642164e+03  2.83416292e+03 -3.56587977e+02  1.49992164e+02
   1.62371398e+02 -6.56330753e+03 -6.72106249e+03  3.94796818e+04
   8.00536018e+03  3.00883652e+00 -3.13261494e+03  8.55571090e+01
   7.32203957e+01 -2.08548884e+01  1.04493815e+02  3.50129683e+04
   2.95073717e+02]]
```

Since normalization is not done here we can see there is a huge variation in the values of variances and co variances due to difference in scales of the data.

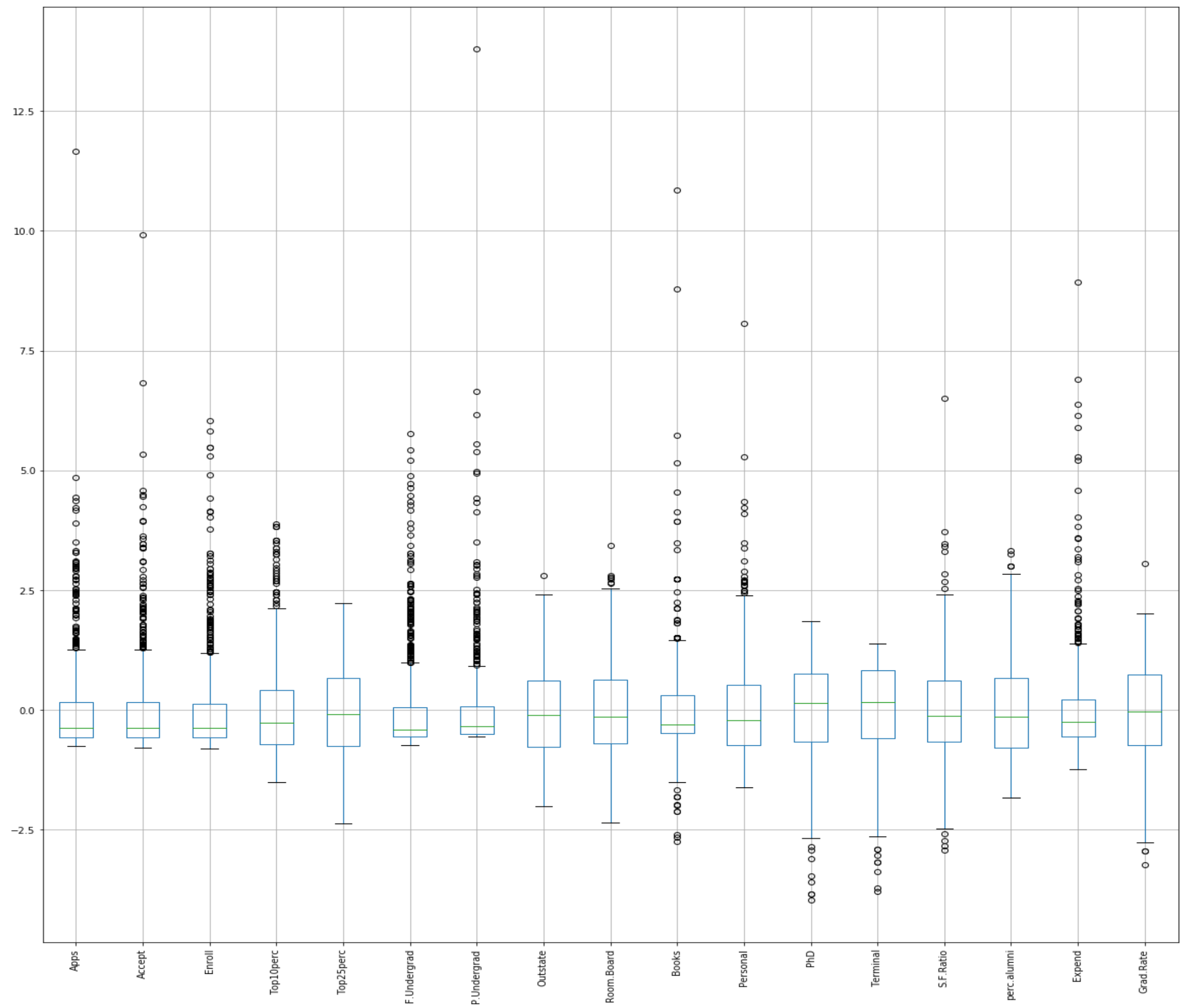
2.4) Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.

Ans)

<u>Before Scaling without outlier treatment:</u>



<u>After Scaling without outlier treatment:</u>



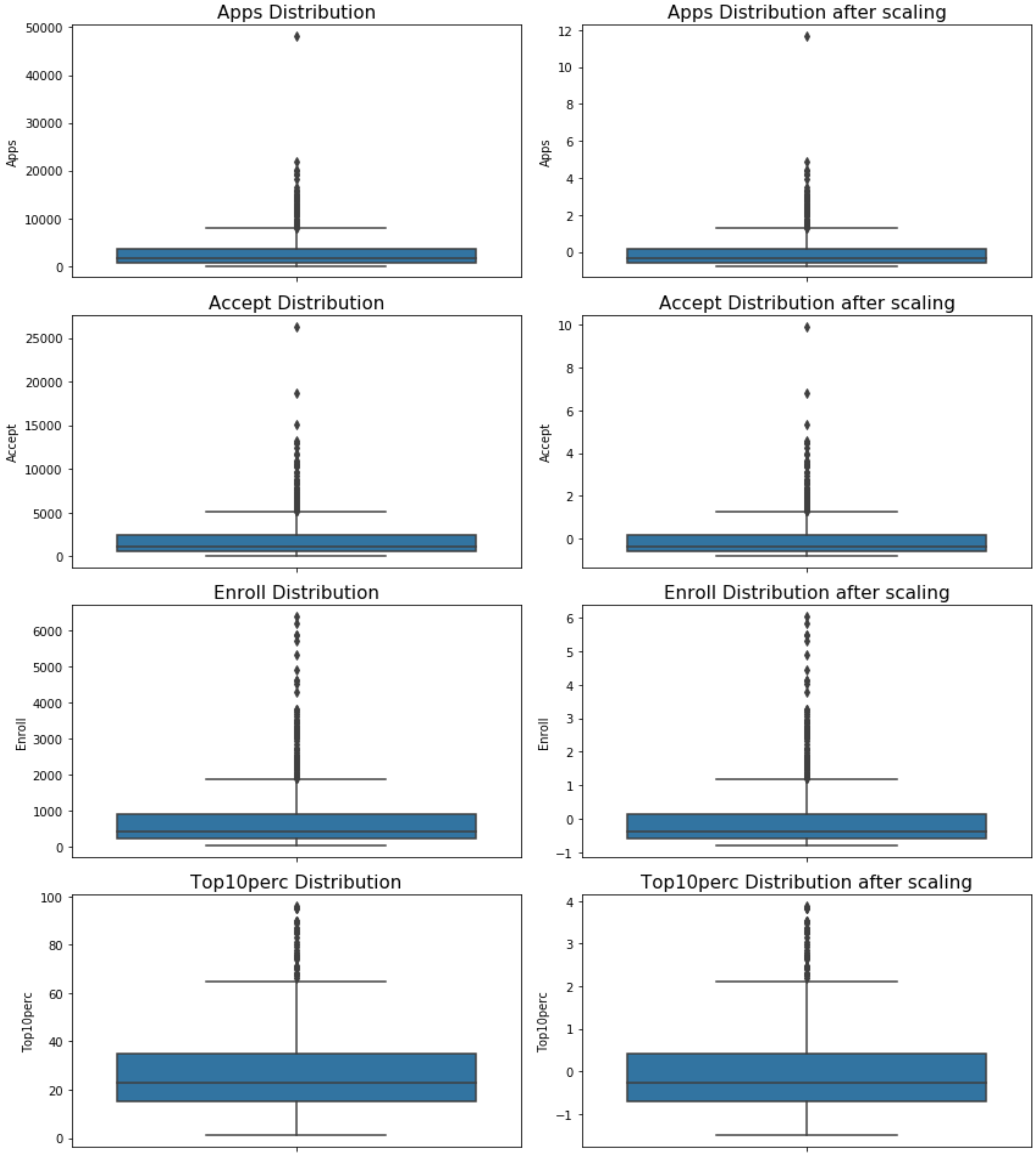
From the above two box plots, we can observe that the original numeric information is on different scales, like outstate and expend had values way higher than other variables.

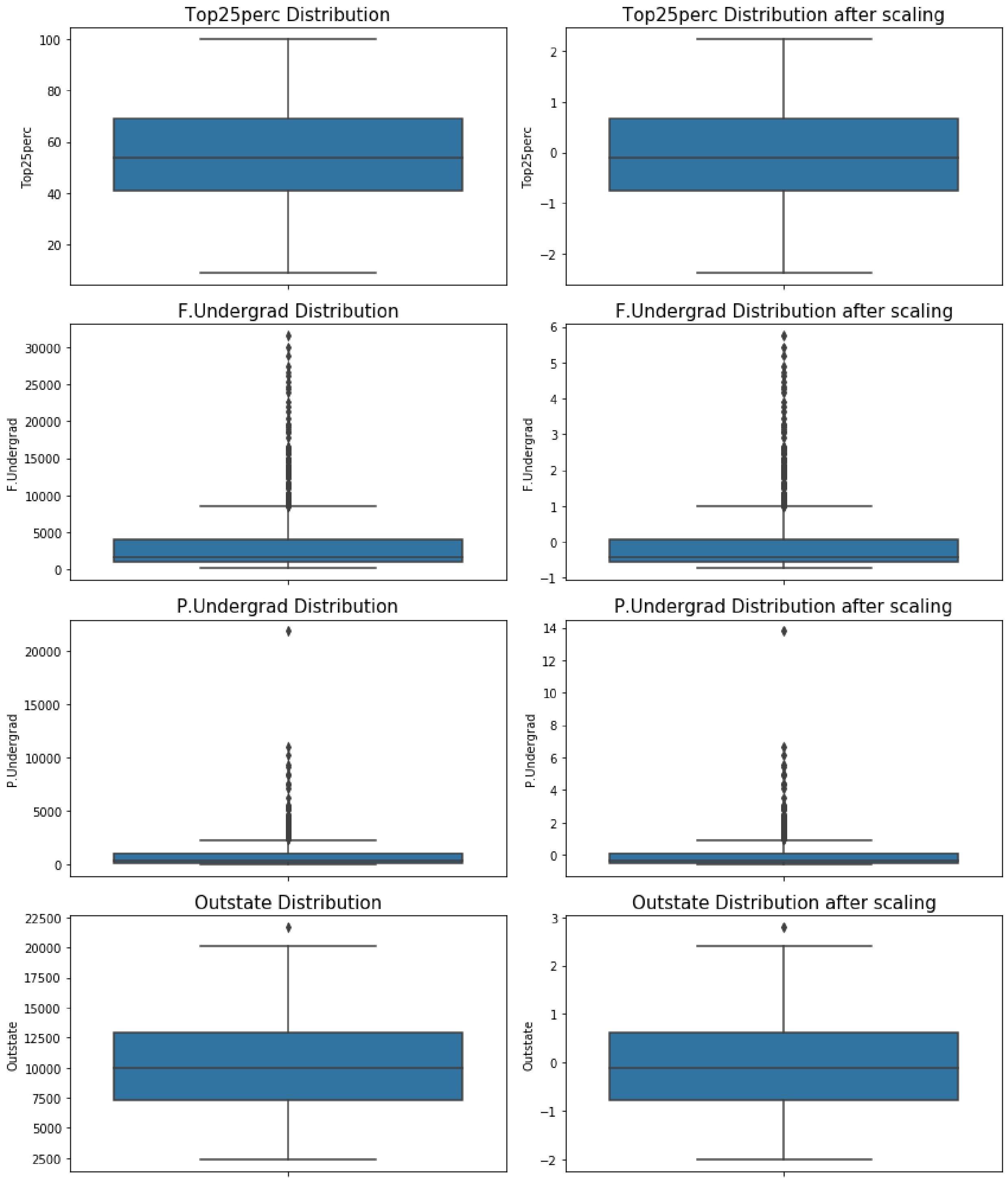
PCA effectiveness depends upon the scales of the attributes. If data is having different scales, PCA has following drawbacks:

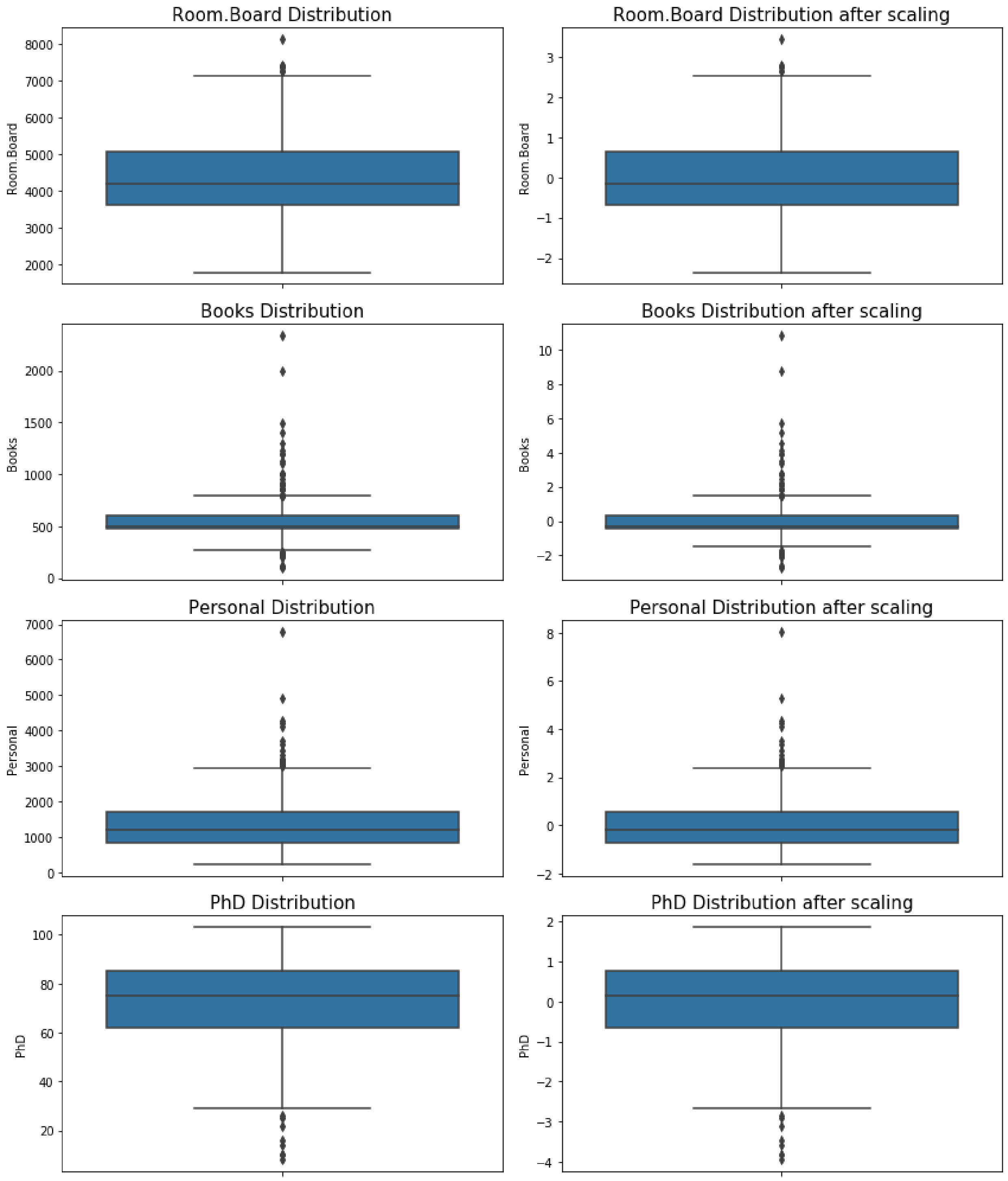
- PCA will pick variable with highest variance rather than picking up attributes based on correlation
- Changing scales of the variables can change the PCA
- Interpreting PCA can become challenging due to presence of discrete data
- Presence of skew in data with long thick tail can impact the effectiveness of the PCA (related to point 1)

Because it's trying to capture the total variance in the set of variables, PCA requires that the input variables have similar scales of measurement. Variables whose numbers are just larger will have much bigger variance just because the numbers are so big. So before starting with a covariance matrix, it’s a good idea to standardize those variables before we begin so that the variables with the biggest scales don’t overwhelm the PCA. So after scaling the variables, we can observe in box plot 2 that the variables have been standardized and are now on same scale, and now this data can be worked upon for PCA.

But we can see that we still have outliers in the scaled dataset.




Top25perc Distribution
Top25perc
100
80
60
40
20
Top25perc Distribution after scaling
Top25perc
2
1
0
−1
−2
F.Undergrad Distribution
F.Undergrad
30000
25000
20000
15000
10000
5000
0
F.Undergrad Distribution after scaling
F.Undergrad
6
5
4
3
2
1
0
−1
P.Undergrad Distribution
P.Undergrad
20000
15000
10000
5000
0
P.Undergrad Distribution after scaling
P.Undergrad
14
12
10
8
6
4
2
0
Outstate Distribution
Outstate
22500
20000
17500
15000
12500
10000
7500
5000
2500
Outstate Distribution after scaling
Outstate
3
2
1
0
−1
−2



Room.Board Distribution

Room.Board Distribution after scaling

Books Distribution

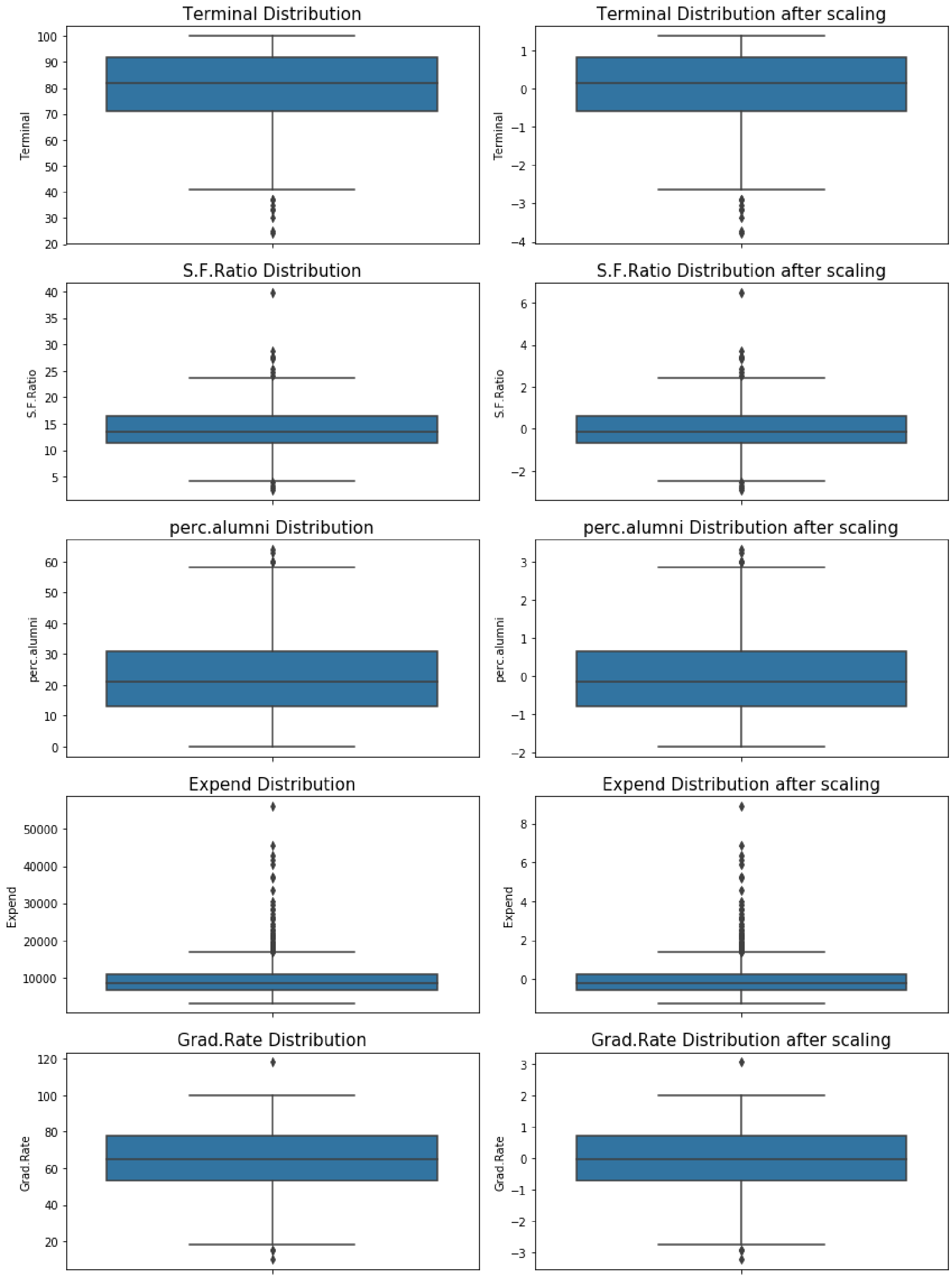
Books Distribution after scaling

Personal Distribution

Personal Distribution after scaling

PhD Distribution

PhD Distribution after scaling


Terminal Distribution
Terminal
100
90
80
70
60
50
40
30
20
Terminal Distribution after scaling
Terminal
1
0
−1
−2
−3
−4
S.F.Ratio Distribution
S.F.Ratio
40
35
30
25
20
15
10
5
S.F.Ratio Distribution after scaling
S.F.Ratio
6
4
2
0
−2
perc.alumni Distribution
perc.alumni
60
50
40
30
20
10
0
perc.alumni Distribution after scaling
perc.alumni
3
2
1
0
−1
−2
Expend Distribution
Expend
50000
40000
30000
20000
10000
Expend Distribution after scaling
Expend
8
6
4
2
0
Grad.Rate Distribution
Grad.Rate
120
100
80
60
40
20
Grad.Rate Distribution after scaling
Grad.Rate
3
2
1
0
−1
−2
−3

From the above side by side comparison we can see that scaling has not impacted the outliers present in the data. Treatment of outlier is sometimes essential and sometimes outliers are a kind of data that is important and should not be treated. Sometimes outliers indicate a mistake in data collection. Other times, though, they can influence a data set, so it's important to keep them to better understand the big picture.

Don't drop an outlier if:

- **Your results are critical**, so even small changes will matter a lot. For example, you can feel better about dropping outliers about people's favorite TV shows, not about the temperatures at which airplane seals fail.
- **There are a lot of outliers.** Outliers are rare by definition. If, for example, 30% of your data is outliers, then it actually means that there's something interesting going on with your data that you need to look further into.

All data distributions have a spread of values. Extreme values can occur, but they have lower probabilities. If the sample size is large enough as in this case, we're bound to obtain unusual values. Random chance might include extreme values in datasets. In other words, the process or population might produce weird values naturally. There's nothing wrong with these data points. They're unusual, but they are a normal part of the data distribution.

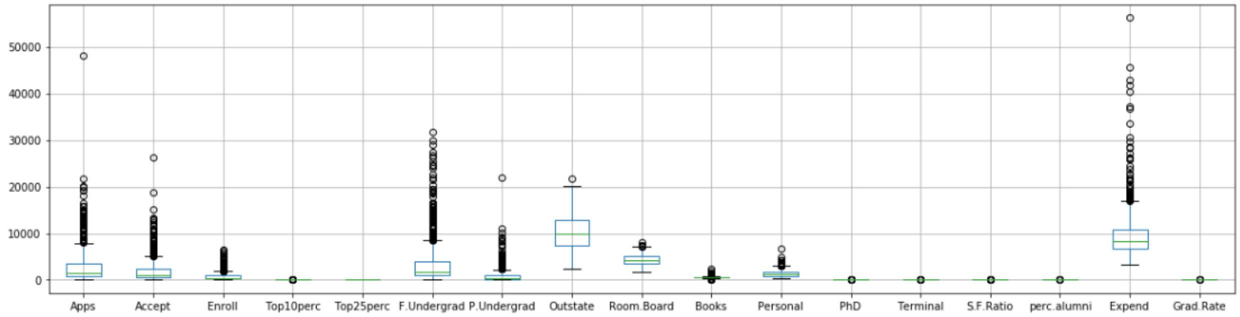
In this dataset, we can see that the data present in every column is not a typing error; they are all possible values which can actually belong to that column and are thus critical information for that particular column.

Sometimes it's best to keep outliers in your data, like the values in the particular dataset because it gives valuable information about critical data of a particular college, which helps us define the trend of that college. It captures valuable information that is part of a study area. Retaining these points can be hard, particularly when it reduces statistical significance! However, excluding extreme values solely due to their extremeness can distort the results by removing information about the variability inherent in the study area. You're forcing the subject area to appear less variable than it is in reality. So in this case we will be refraining from doing the outlier treatment for the sake of getting more accurate results while performing the principal component analysis.

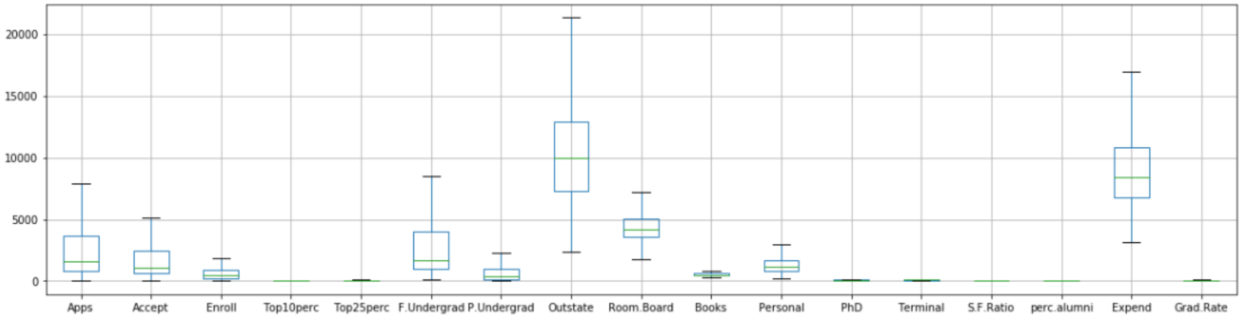
So to come to our final conclusions for our case study, we will not be treating dataset for outliers.

But for comparative analysis, we will be treating the outliers using the capping approach, i.e., replacing extreme values with the upper bound and lower bound in a box plot using IQR.

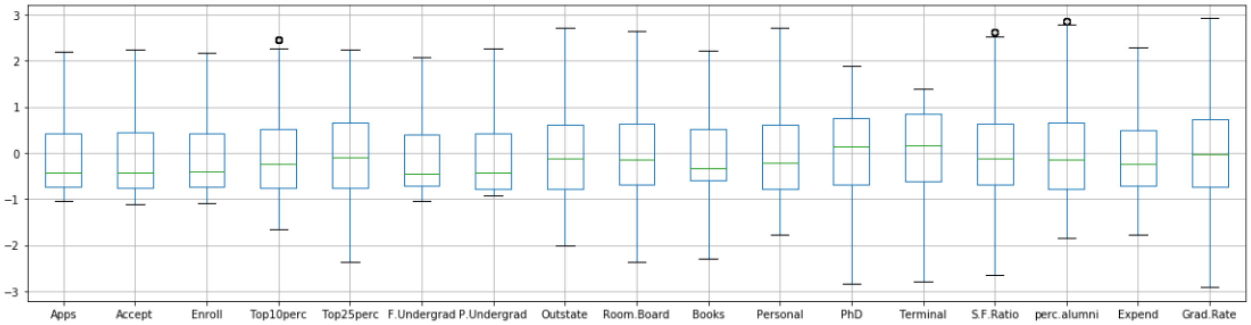
Boxplot of original data without treating outliers:



Boxplot of original data after treating outliers:

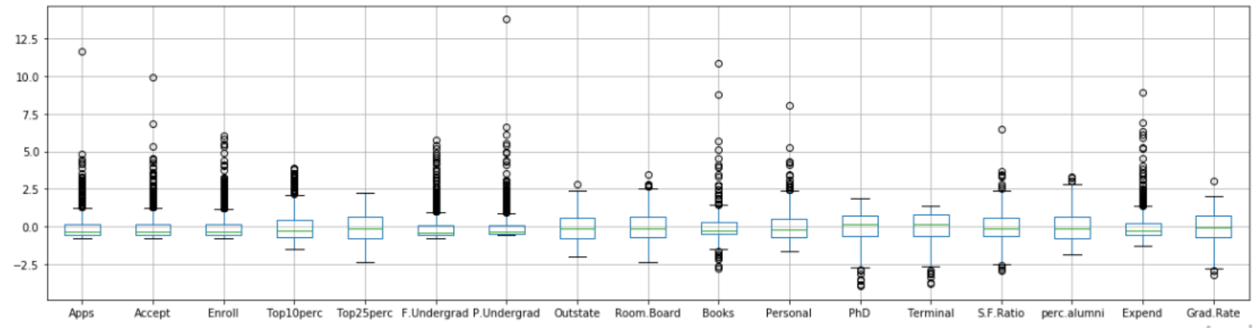


Boxplot after scaling above data (original data after treating outliers):

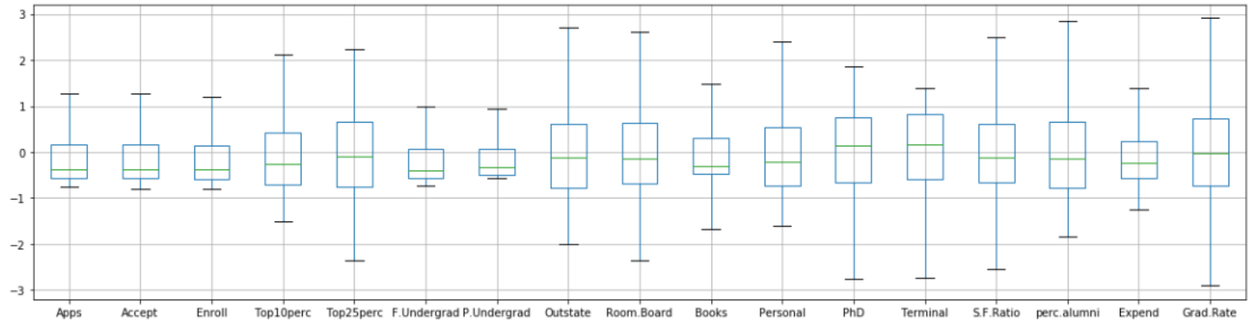


We can see from above plot that even though the dataset was treated for outliers before scaling, still we encountered some outliers after scaling.

Boxplot of the scaled original data not treated for outliers:



Boxplot of above data (scaled original data not treated for outliers) treated for outliers:



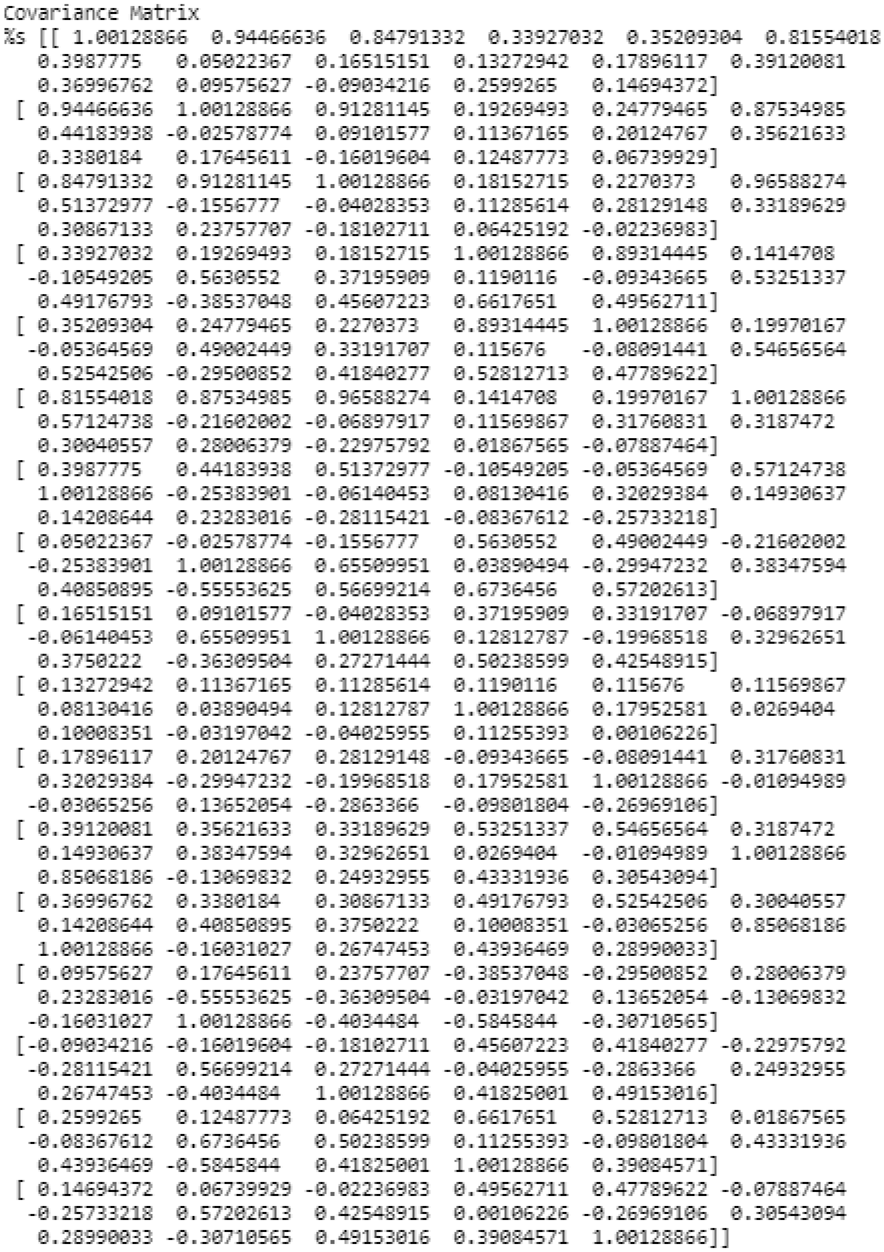
We can see from above graph that the outliers in the data have been successfully treated as compared to previous case.

If we were to treat the outliers, it is always a good practice to treat the outliers after scaling because, even if we treat the outliers before scaling, there are chances we might encounter additional outliers after scaling which might need further treatment. Thus for comparative analysis, we will be using data treated for outliers after scaling.

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

Ans)

Covariance matrix of data without treatment of outliers:



```
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]]
```

EigenValues and EigenVectors data without treatment of outliers:

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]
```

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]]
```

Covariance matrix of data after treatment of outliers:

```
Covariance Matrix
%s [[ 3.92202587e-01  3.72045278e-01  3.44861102e-01  1.77882547e-01
   2.28413447e-01  3.00372860e-01  1.53261035e-01  4.09309446e-02
   1.16838149e-01  1.03276949e-01  1.29720101e-01  2.84046975e-01
   2.67661215e-01  7.57322866e-02 -6.30526007e-02  9.90236070e-02
   9.43096150e-02]
 [ 3.72045278e-01  3.86719034e-01  3.57101106e-01  1.22742172e-01
   1.70302604e-01  3.10747826e-01  1.67663695e-01 -3.11176328e-03
   7.40055439e-02  9.06387432e-02  1.43597630e-01  2.59812502e-01
   2.46777922e-01  1.12140752e-01 -1.02443762e-01  6.54924734e-02
   4.90474515e-02]
 [ 3.44861102e-01  3.57101106e-01  3.76969419e-01  9.32127566e-02
   1.41572482e-01  3.30838988e-01  1.85453496e-01 -9.55996485e-02
  -1.45695116e-02  8.66382517e-02  1.87681520e-01  2.29023882e-01
   2.14034520e-01  1.61090704e-01 -1.36102632e-01  2.16676723e-02
  -1.42555842e-02]
 [ 1.77882547e-01  1.22742172e-01  9.32127566e-02  7.81305061e-01
   8.08308058e-01  5.47613345e-02 -7.49074558e-02  4.97059995e-01
   3.14346107e-01  9.47255125e-02 -9.29426027e-02  4.70148933e-01
   4.40620887e-01 -3.28019751e-01  4.00987210e-01  3.78002594e-01
   4.35749475e-01]
 [ 2.28413447e-01  1.70302604e-01  1.41572482e-01  8.08308058e-01
   1.00128866e+00  1.01001820e-01 -4.67766043e-02  4.90041139e-01
   3.29590971e-01  1.18631669e-01 -7.82479424e-02  5.39489477e-01
   5.19387111e-01 -2.84523563e-01  4.15135356e-01  3.73125871e-01
   4.78619938e-01]
 [ 3.00372860e-01  3.10747826e-01  3.30838988e-01  5.47613345e-02
   1.01001820e-01  3.10315170e-01  1.82563117e-01 -1.26027953e-01
  -3.01988039e-02  8.08715656e-02  1.80536480e-01  1.96913464e-01
   1.83602609e-01  1.72926770e-01 -1.58267436e-01  1.34384529e-04
  -4.57475459e-02]
 [ 1.53261035e-01  1.67663695e-01  1.85453496e-01 -7.49074558e-02
  -4.67766043e-02  1.82563117e-01  2.21636838e-01 -1.66812465e-01
  -3.16881569e-02  4.02847905e-02  1.45904561e-01  5.87588044e-02
   5.65696027e-02  1.66907512e-01 -1.96485156e-01 -6.18747157e-02
  -1.24656448e-01]
 [ 4.09309446e-02 -3.11176328e-03 -9.55996485e-02  4.97059995e-01
   4.90041139e-01 -1.26027953e-01 -1.66812465e-01  1.00063964e+00
   6.52513004e-01  3.56977546e-03 -2.93399069e-01  3.82701065e-01
   4.05983213e-01 -5.48973759e-01  5.63250944e-01  5.04797524e-01
   5.71836675e-01]
 [ 1.16838149e-01  7.40055439e-02 -1.45695116e-02  3.14346107e-01
   3.29590971e-01 -3.01988039e-02 -3.16881569e-02  6.52513004e-01
   9.90307358e-01  7.56993714e-02 -1.96811114e-01  3.32218551e-01
   3.71275537e-01 -3.58352687e-01  2.69792322e-01  3.76072607e-01
   4.23126447e-01]
 [ 1.03276949e-01  9.06387432e-02  8.66382517e-02  9.47255125e-02
   1.18631669e-01  8.08715656e-02  4.02847905e-02  3.56977546e-03
   7.56993714e-02  4.87714769e-01  1.50893084e-01  9.31226224e-02
   1.09448483e-01 -5.70259896e-03 -2.97713834e-02  6.81740268e-02
  -5.61443574e-03]
 [ 1.29720101e-01  1.43597630e-01  1.87681520e-01 -9.29426027e-02
  -7.82479424e-02  1.80536480e-01  1.45904561e-01 -2.93399069e-01
  -1.96811114e-01  1.50893084e-01  8.11420703e-01 -1.02893106e-02
  -2.83293174e-02  1.49863186e-01 -2.74121214e-01 -9.57249786e-02
  -2.61665861e-01]
 [ 2.84046975e-01  2.59812502e-01  2.29023882e-01  4.70148933e-01
   5.39489477e-01  1.96913464e-01  5.87588044e-02  3.82701065e-01
   3.32218551e-01  9.31226224e-02 -1.02893106e-02  9.55820539e-01
   8.29899302e-01 -1.21012174e-01  2.42170983e-01  3.24864101e-01
   3.02668013e-01]
 [ 2.67661215e-01  2.46777922e-01  2.14034520e-01  4.40620887e-01
   5.19387111e-01  1.83602609e-01  5.65696027e-02  4.05983213e-01
   3.71275537e-01  1.09448483e-01 -2.83293174e-02  8.29899302e-01
   9.67665090e-01 -1.42089297e-01  2.60463224e-01  3.35539361e-01
   2.87625501e-01]
 [ 7.57322866e-02  1.12140752e-01  1.61090704e-01 -3.28019751e-01
  -2.84523563e-01  1.72926770e-01  1.66907512e-01 -5.48973759e-01
  -3.58352687e-01 -5.70259896e-03  1.49863186e-01 -1.21012174e-01
  -1.42089297e-01  9.15128308e-01 -3.92368123e-01 -4.07438139e-01
  -2.94727250e-01]
 [-6.30526007e-02 -1.02443762e-01 -1.36102632e-01  4.00987210e-01
   4.15135356e-01 -1.58267436e-01 -1.96485156e-01  5.63250944e-01
   2.69792322e-01 -2.97713834e-02 -2.74121214e-01  2.42170983e-01
   2.60463224e-01 -3.92368123e-01  9.90599513e-01  2.99881835e-01
   4.88405261e-01]
 [ 9.90236070e-02  6.54924734e-02  2.16676723e-02  3.78002594e-01
   3.73125871e-01  1.34384529e-04 -6.18747157e-02  5.04797524e-01
   3.76072607e-01  6.81740268e-02 -9.57249786e-02  3.24864101e-01
   3.35539361e-01 -4.07438139e-01  2.99881835e-01  4.23629634e-01
   2.69919979e-01]
 [ 9.43096150e-02  4.90474515e-02 -1.42555842e-02  4.35749475e-01
   4.78619938e-01 -4.57475459e-02 -1.24656448e-01  5.71836675e-01
   4.23126447e-01 -5.61443574e-03 -2.61665861e-01  3.02668013e-01
   2.87625501e-01 -2.94727250e-01  4.88405261e-01  2.69919979e-01
   9.97192469e-01]]
```

EigenValues and EigenVectors data after treatment of outliers:

```
Eigen Values
 %s [4.75579369 2.3800885  0.88497491 0.81453646 0.72423975 0.52688069
 0.47958062 0.41127635 0.36620193 0.23942458 0.00793972 0.01435481
 0.03582106 0.12943793 0.09751277 0.08189987 0.06059116]


Eigen Vectors
 %s [[ 0.0929684   0.32104652  0.06660652 -0.0129432   0.24674827 -0.00650339
   0.2400326  -0.13180129 -0.01773119 -0.03400089  0.14346037 -0.59269472
   0.5569348   0.03784437  0.22445438  0.11116423 -0.00914552]
 [ 0.06592707  0.3319699   0.07883241 -0.03420729  0.22877472 -0.02487384
   0.27288698 -0.12917757 -0.0195307  -0.06133927 -0.32336365  0.70710813
   0.26755069  0.00907891  0.17576693  0.15058739  0.00281244]
 [ 0.03166929  0.35033549  0.01381154 -0.01122623  0.19114851 -0.03094669
   0.26523924 -0.12023338 -0.00760842 -0.00639234  0.69930928  0.13493584
  -0.49315933 -0.01188448  0.04255122  0.01589244  0.02985794]
 [ 0.33452816  0.06754279 -0.32328505  0.21141739  0.07741651  0.32096376
  -0.09974811  0.02170441  0.13379303 -0.00700079 -0.0310452   0.02159112
   0.00901027  0.08516089  0.14064341 -0.29511711  0.69375696]
 [ 0.36427546  0.13142781 -0.41399578  0.19443136  0.11797053  0.37686172
  -0.18713202  0.01531995  0.18486533  0.12749803  0.00890279  0.01682207
  -0.00267915 -0.13131824 -0.18030241  0.29567976 -0.511526  ]
 [ 0.01149875  0.32452837  0.02193807 -0.01744947  0.15029942 -0.01698553
   0.21021832 -0.10021068 -0.01175077  0.02574847 -0.61814444 -0.35140689
  -0.55093616 -0.01666731 -0.06116207 -0.05081468 -0.01136756]
 [-0.04622402  0.20972858  0.1038968  -0.02676078  0.06989958 -0.00474311
   0.08937877 -0.0537958   0.04181848  0.13366948  0.04589764  0.03885668
   0.23891433 -0.08576745 -0.8706629  -0.24738219  0.14019587]
 [ 0.37830181 -0.20665209  0.2446006   0.02000679  0.04640648 -0.05172404
   0.05471221 -0.02283038  0.17673816 -0.75697305 -0.00367566 -0.04554559
  -0.07289772 -0.0571668  -0.21583357  0.27224079  0.09997777]
 [ 0.29777508 -0.07383062  0.65435548 -0.07736692  0.20589468  0.0558972
  -0.31550426 -0.11989134  0.33073591  0.42976697  0.00556533  0.01161555
  -0.06810512  0.05833278  0.09487917  0.01923638  0.02476745]
 [ 0.0401252   0.13395669  0.06932308  0.29066279  0.05440207 -0.08091173
  -0.50411092 -0.48393685 -0.6116081  -0.10870293  0.00416292  0.00902118
  -0.01670698  0.0378997  -0.05038066  0.04789243  0.02826868]
 [-0.10944135  0.29386253  0.02906196  0.60616613 -0.01326297 -0.52880595
  -0.19661216  0.33142515  0.31893212 -0.03384734 -0.00630645  0.0047103
   0.02161283 -0.01497856  0.03722249  0.00703647 -0.0047994 ]
 [ 0.31241468  0.30728219  0.01051885 -0.21336602 -0.44256735 -0.07682269
  -0.04689621  0.18837226 -0.0983305   0.02010444  0.00804     0.00548016
  -0.00454371  0.70146927 -0.09315117  0.10037346 -0.0434578 ]
 [ 0.31563577  0.28923834  0.07534077 -0.22034156 -0.48498252 -0.10434108
  -0.06899073  0.09780965 -0.12229609  0.06046494  0.0048878  -0.00898686
   0.02707478 -0.67932403  0.11394885 -0.01716347  0.08863951]
 [-0.238936    0.27738993 -0.19726187 -0.50833033  0.128203   -0.06989583
  -0.48926256 -0.16012666  0.36247736 -0.31211523  0.00662521  0.01073745
   0.02278083  0.01040467  0.06491485 -0.20799672 -0.07241455]
 [ 0.28566756 -0.26160327 -0.35032544 -0.05443422 -0.08731305 -0.56468516
   0.1645982  -0.52854784  0.21127982  0.21023249 -0.01383698  0.00137953
   0.01480452  0.03469458 -0.02561499 -0.01145122  0.00883861]
 [ 0.24699418 -0.02920582  0.14253472  0.17754101 -0.0657708   0.05988252
   0.1378192  -0.05801938 -0.02433122 -0.20131984  0.01193343  0.04641723
   0.04412393  0.04429777  0.12534642 -0.76406889 -0.45737401]
 [ 0.31117828 -0.12680266 -0.14343185 -0.26409767  0.54379453 -0.34192987
  -0.11257333  0.47485834 -0.35725708  0.05772333 -0.00123347  0.00788293
  -0.01406045 -0.03267103 -0.03283629 -0.10963549 -0.02419997]]
```

Covariance matrix helps in seeing the variance in the data; it’s having variance stored in them. Eigen values help in explaining the variance and they are calculated based on covariance matrix. By comparing the eigenvalues for above two cases, i.e., with and without treatment of outliers, we can see that a significant amount of variance was not explained by the principal components after treatment of outliers which could be useful in our case study.

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

Ans)

First Principal component (PC) = Eigen vector with highest Eigen value

∴ PC = array([-0.2487656 , 0.33159823, 0.0630921 , -0.28131053, 0.00574141, 0.01623744, 0.04248635, 0.1030904 , 0.09022708, -0.0525098 , 0.3589704 , -0.4591395 , 0.04304621, -0.13340581, 0.0806328 , -0.59583097, 0.02407091])

Its Eigen value = 5.450521622150289

Columns of dataset = Index(['Apps', 'Accept', 'Enroll', 'Top10perc', 'Top25perc', 'F.Undergrad', 'P.Undergrad', 'Outstate', 'Room.Board', 'Books', 'Personal', 'PhD', 'Terminal', 'S.F.Ratio', 'perc.alumni', 'Expend', 'Grad.Rate']

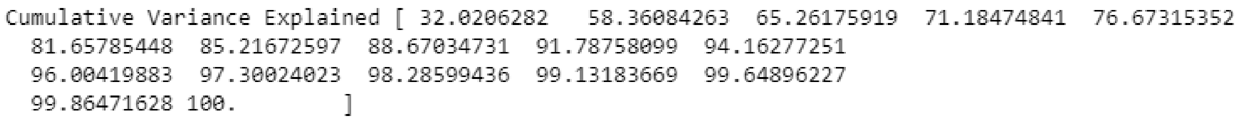
∴ Explicit form of first PC = PC * Columns of dataset

= (-0.2487656 * Apps) + (0.33159823 * Accept) + (0.0630921 * Enroll) + (-0.28131053 *Top10perc) + (0.00574141 * Top25perc) + (0.01623744 * F.Undergrad) + (0.04248635 * P.Undergrad) + (0.1030904 * Outstate) + (0.09022708 * Room.Board) + (-0.0525098 * Books) + (0.3589704 * Personal) + (-0.4591395 * PhD) + (0.04304621 * Terminal) + (-0.13340581 * S.F.Ratio) + (0.0806328 * perc.alumni) + (-0.59583097 * Expend) + (0.02407091 * Grad.Rate)

2.7) Discuss the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate? Perform PCA and export the data of the Principal Component scores into a data frame.

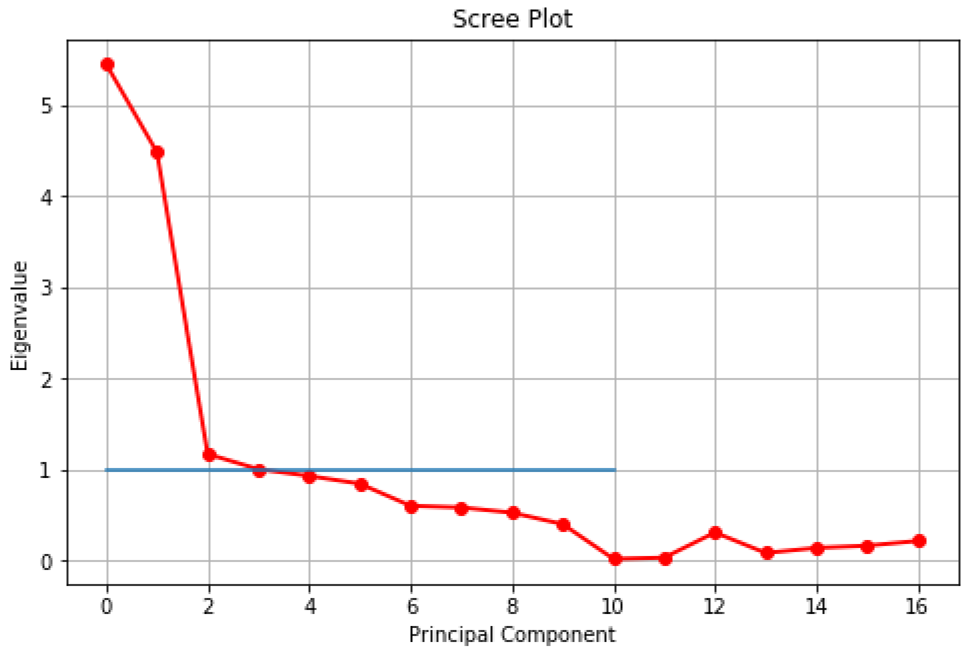
Ans)

The cumulative values of eigenvalues in ascending order of eigenvalues is as follows

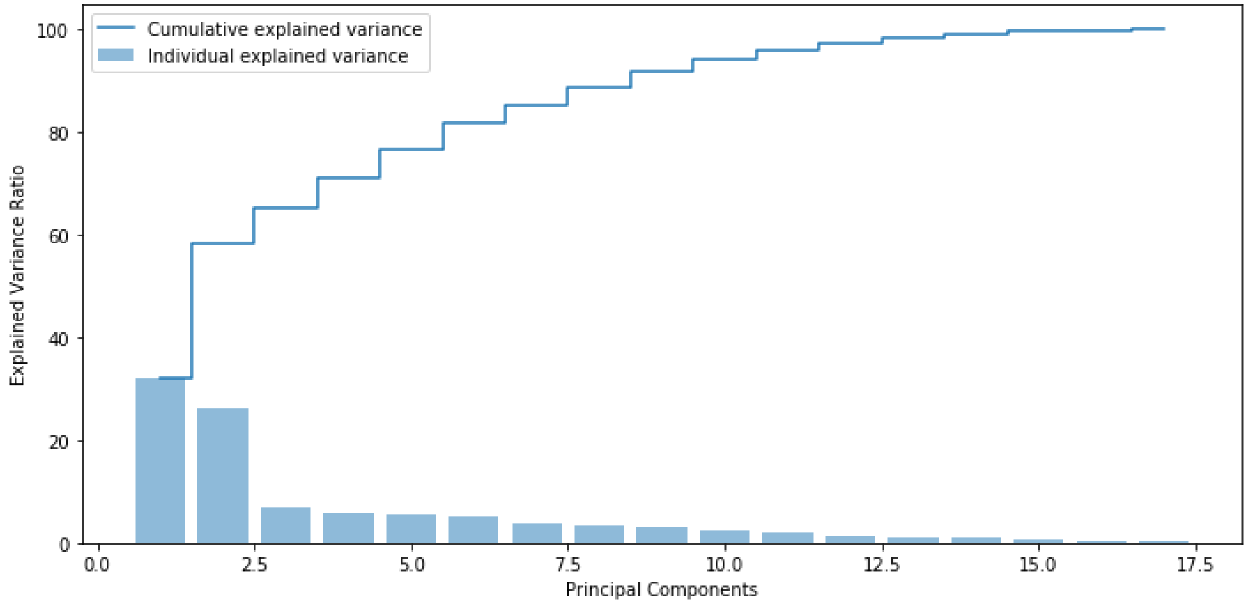


```
Cumulative Variance Explained [ 32.0206282   58.36084263  65.26175919  71.18474841  76.67315352
  81.65785448  85.21672597  88.67034731  91.78758099  94.16277251
  96.00419883  97.30024023  98.28599436  99.13183669  99.64896227
  99.86471628 100.        ]
```

Plotting the scree plot



Plot Cumulative explained variance and individual explained variance vs Principal Components

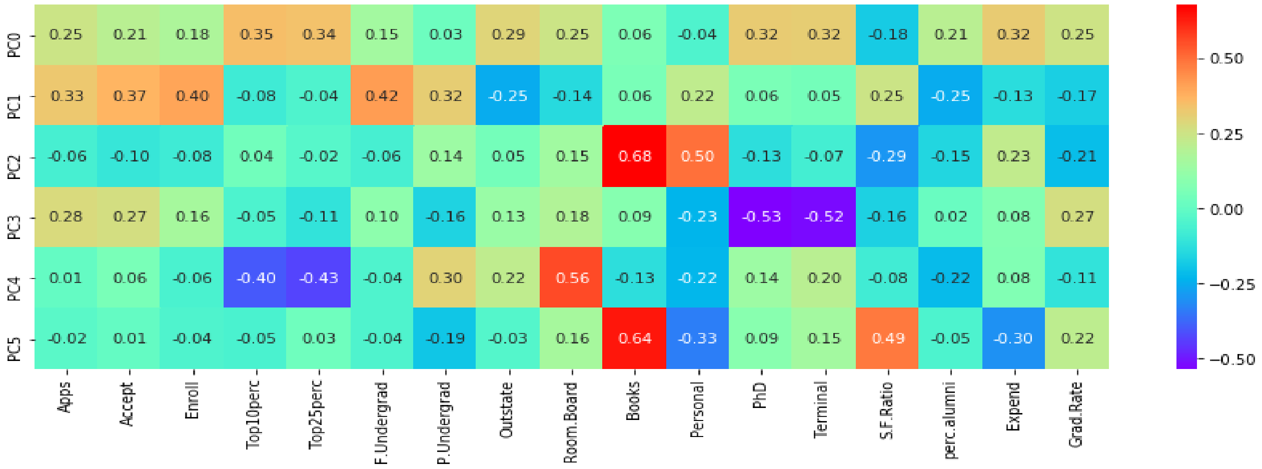


Visually we can observe that there is a steep drop in variance explained with increase in number of PC's.

There are a few criteria's which can be considered while selecting the number of principal components:

- 1 eigenvalue: There are 4 PCs with eigenvalues greater than 1 and generally these are considered more significant than others.
- For accurate representation the minimum number of PC's should be 1/4th of the total variables, in this case four components, should be taken into consideration. The first four components explain 71.19% of variance in data.
- One could select the top components such that the cumulative variance exceeds a threshold, such as 80%. In this case, 6 components.

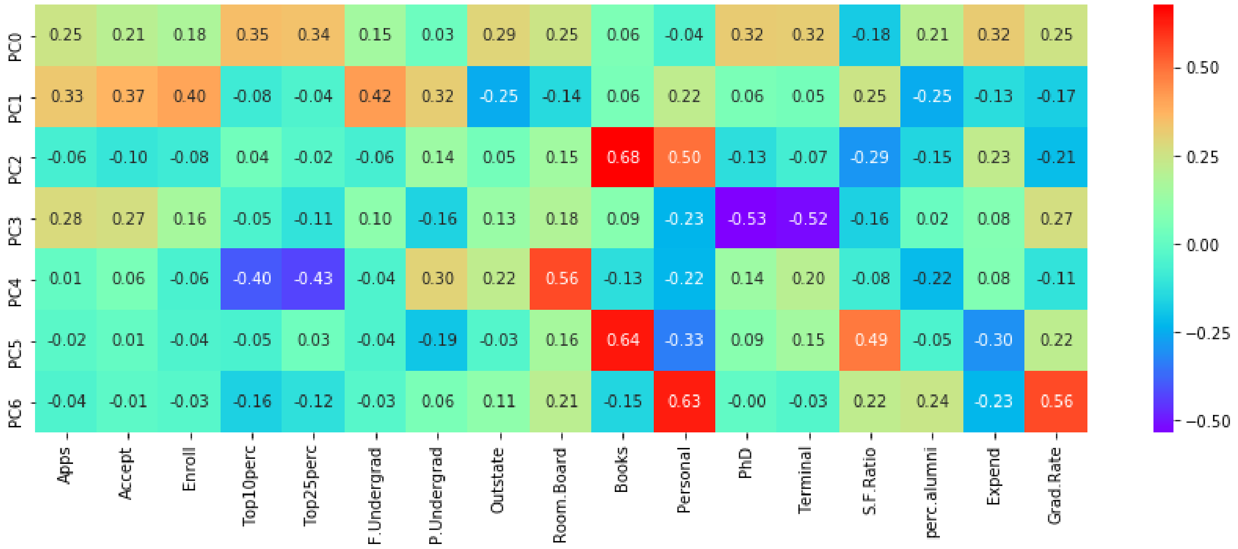
By taking a correlation plot of 6 PCs



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
PC0	0.25	0.21	0.18	0.35	0.34	0.15	0.03	0.29	0.25	0.06	-0.04	0.32	0.32	-0.18	0.21	0.32	0.25
PC1	0.33	0.37	0.40	-0.08	-0.04	0.42	0.32	-0.25	-0.14	0.06	0.22	0.06	0.05	0.25	-0.25	-0.13	-0.17
PC2	-0.06	-0.10	-0.08	0.04	-0.02	-0.06	0.14	0.05	0.15	0.68	0.50	-0.13	-0.07	-0.29	-0.15	0.23	-0.21
PC3	0.28	0.27	0.16	-0.05	-0.11	0.10	-0.16	0.13	0.18	0.09	-0.23	-0.53	-0.52	-0.16	0.02	0.08	0.27
PC4	0.01	0.06	-0.06	-0.40	-0.43	-0.04	0.30	0.22	0.56	-0.13	-0.22	0.14	0.20	-0.08	-0.22	0.08	-0.11
PC5	-0.02	0.01	-0.04	-0.05	0.03	-0.04	-0.19	-0.03	0.16	0.64	-0.33	0.09	0.15	0.49	-0.05	-0.30	0.22

We see that not all variables are explained properly with these many PCs, so we increase them gradually and check their correlation.

Plotting the correlation for 7 PCs



We can deduce that all the variables are getting covered by the 7 PCs in the following way:

PC0 = Outstate, Expend, Top10perc, Top25perc

PC1 = Apps, Accept, Enroll, F.Undergrad, P.Undergrad, perc.alumni

PC2 = Books

PC3 = PhD, Terminal

PC4 = Room.Board

PC5 = S.F.Ratio

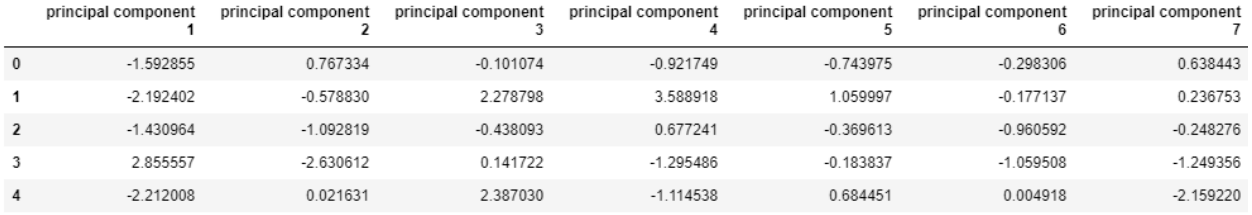
PC6 = Personal, Grad.Rate

In this manner, we are able to cover 85.22% variance in the data with minimum correlation between the components, reducing redundancy. The 7 principal components can then be used in place of the original 17 predictors, reducing dimensionality.

The eigenvectors are the principal components indeed which determine the directions of the new feature space, and the eigenvalues determine their magnitude. The first eigenvector represents a direction captured by multiplying weights (also known as factor loadings) with respective columns/variables/features in a linear manner which captures most of the variation in the data.

We created a new dataframe named ‘**principal_components_Df**’ that has the principal component values for all 777 colleges.

The first 5 rows of the dataset are as follows

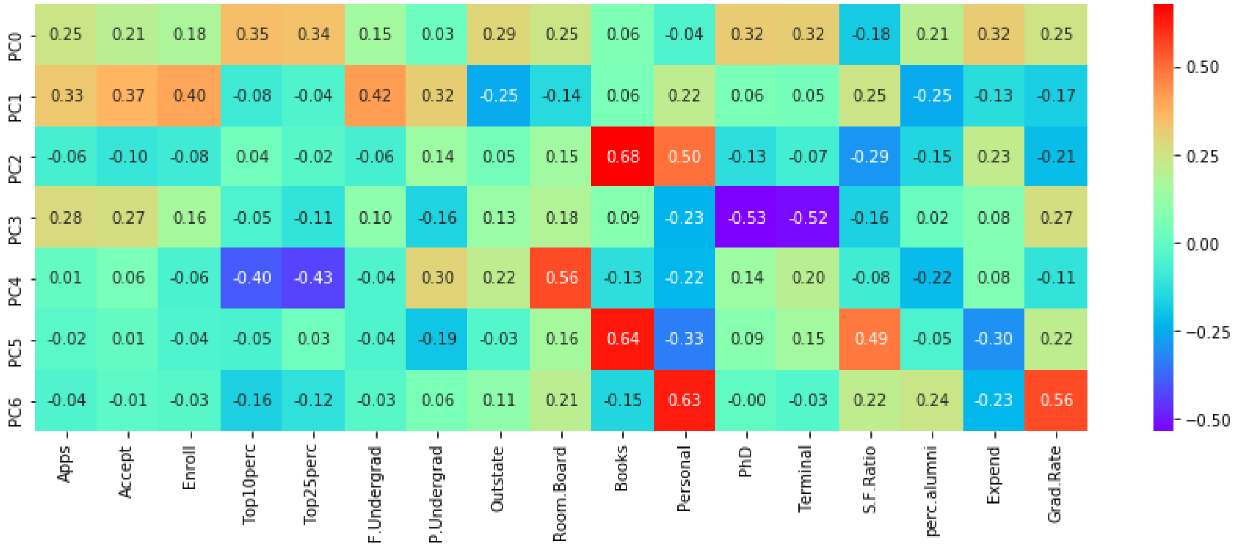


	principal component 1	principal component 2	principal component 3	principal component 4	principal component 5	principal component 6	principal component 7
0	-1.592855	0.767334	-0.101074	-0.921749	-0.743975	-0.298306	0.638443
1	-2.192402	-0.578830	2.278798	3.588918	1.059997	-0.177137	0.236753
2	-1.430964	-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

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

Ans)

As we have seen from the previous solution that we have decided to include 7 PCs in this case study, and after plotting the correlation for 7 PCs



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
PC0	0.25	0.21	0.18	0.35	0.34	0.15	0.03	0.29	0.25	0.06	-0.04	0.32	0.32	-0.18	0.21	0.32	0.25
PC1	0.33	0.37	0.40	-0.08	-0.04	0.42	0.32	-0.25	-0.14	0.06	0.22	0.06	0.05	0.25	-0.25	-0.13	-0.17
PC2	-0.06	-0.10	-0.08	0.04	-0.02	-0.06	0.14	0.05	0.15	0.68	0.50	-0.13	-0.07	-0.29	-0.15	0.23	-0.21
PC3	0.28	0.27	0.16	-0.05	-0.11	0.10	-0.16	0.13	0.18	0.09	-0.23	-0.53	-0.52	-0.16	0.02	0.08	0.27
PC4	0.01	0.06	-0.06	-0.40	-0.43	-0.04	0.30	0.22	0.56	-0.13	-0.22	0.14	0.20	-0.08	-0.22	0.08	-0.11
PC5	-0.02	0.01	-0.04	-0.05	0.03	-0.04	-0.19	-0.03	0.16	0.64	-0.33	0.09	0.15	0.49	-0.05	-0.30	0.22
PC6	-0.04	-0.01	-0.03	-0.16	-0.12	-0.03	0.06	0.11	0.21	-0.15	0.63	-0.00	-0.03	0.22	0.24	-0.23	0.56

Interpretation of the principal components is based on finding which variables are most strongly correlated with each component, i.e., which of these numbers are large in magnitude, the farthest from zero in either direction. And as per that we deduced all the variables are getting covered by the 7 PCs in the following way:

PC0 = Outstate, Expend, Top10perc, Top25perc
PC1 = Apps, Accept, Enroll, F.Undergrad, P.Undergrad, perc.alumni
PC2 = Books
PC3 = PhD, Terminal
PC4 = Room.Board
PC5 = S.F.Ratio
PC6 = Personal, Grad.Rate

First Principal Component Analysis – PC0

Variables: Outstate (Number of students for whom the particular college or university is Out-of-state tuition), Expend (The Instructional expenditure per student), Top10perc (Percentage of new students from top 10% of Higher Secondary class), Top25perc (Percentage of new students from top 25% of Higher Secondary class)

It is most positively correlated with these two variables as compared to others. We know that out-of-state tuition is typically more expensive than In-state tuition. Out-of-state students pay more simply because they do not pay taxes to the state in which the university is located. In-state residents, on the other hand, have been supporting the state, and thus indirectly funding the university, all their lives. Thus, lower tuition costs are the state's way of both rewarding its residents for their contributions and accounting for the tax dollars they have already paid to support their state's schools. And from this relation we can say the PC0 correlates with the fact that students for whom the particular college or university is Out-of-state have more expenditure when compared to other students.

Top10perc (Percentage of new students from top 10% of Higher Secondary class), Top25perc (Percentage of new students from top 25% of Higher Secondary class) are also positive significant contributors here, which can further tell us that more students are preferred if they belong to these two categories.

Second Principal Component Analysis – PC1

Variables: Apps (Number of applications received), Accept (Number of applications accepted), Enroll (Number of new students enrolled), F.Undergrad (Number of full-time undergraduate students), perc.alumni (Percentage of alumni who donate)

It is most positively correlated with these variables and slightly negatively but most related with perc.alumni as compared to others. Since there is positive correlation between Apps, Accept, Enroll, and F.Undergrad we can deduce that if the number of applications received is higher, then chances of number of applications accepted are also higher, and from them enrollments of new students are also higher. These new enrolled students are more likely to be full-time undergraduate students. It also shows that alumni are less likely to donate if the college has higher number of students, i.e. Apps, Accept and Enroll since the alumni might think the institution makes enough money from the higher number of students.

Third Principal Component Analysis – PC2

Variables: Books (Estimated book costs for a student)

This PC is closely positively correlated with books and also Personal (Estimated personal spending for a student) (**not as much as PC6 but still a significant amount but makes sense here**). It shows that books might constitute of a significant part of personal spending of students.

Fourth Principal Component Analysis – PC3

Variables: PhD (Percentage of faculties with Ph.D.'s), Terminal (Percentage of faculties with terminal degree)

This PC is highly negatively correlated with PhD and Terminal. But the values of them are same. It can imply that the faculties with PhD indeed have PhD as their terminal degree, i.e., the highest academic degree that can be awarded in a particular field. It can be a measure of faculties with highest level of qualification in their fields and there is a high chance that it is PhD.

Fifth Principal Component Analysis – PC4
Variables: Room.Board (Cost of Room and board)
It is most positively correlated with Room.Board. This component can be a measure of the Cost of Room and board for an institution.

Sixth Principal Component Analysis – PC5
Variables: S.F.Ratio (Student/faculty ratio)
It is most positively correlated with S.F.Ratio. This component can be a measure of the Student/faculty ratio for an institution.

Seventh Principal Component Analysis – PC6
Variables: Personal (Estimated personal spending for a student), Grad.Rate (Graduation rate)
It is most positively correlated with Personal and Grad.Rate. We can deduce that for the colleges having high graduation rate, the personal spending of students tends to be higher too.