PROJECT REPORT ON ADVANCED STATISTICS

Akshaya Parthasarathy

Batch: PGPDSBA_online_July E 2020

Problem 1:

ANOVA – Analysis of Variance

Problem statement:

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

[Assume all the ANOVA assumptions are satisfied]

Exploratory Data Analysis:

Head of the dataset: Verify whether the dataset is loaded correctly

	A	B	Volunteer	Relief
0	1	1	1	2.4
1	1	1	2	2.7
2	1	1	3	2.3
3	1	1	4	2.5
4	1	2	1	4.6

Information of the dataset: There are four variables (A, B, Volunteer and Relief). All of which are either int or float. While performing ANOVA analysis the variables A and B will be considered as category.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36 entries, 0 to 35
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   A          36 non-null     int64
 1   B          36 non-null     int64
 2   Volunteer  36 non-null     int64
 3   Relief     36 non-null     float64
dtypes: float64(1), int64(3)
memory usage: 1.2 KB
```

Shape of the dataset:

There are 36 rows and 4 columns in the Fever dataset.

Summary statistics of the dataset:

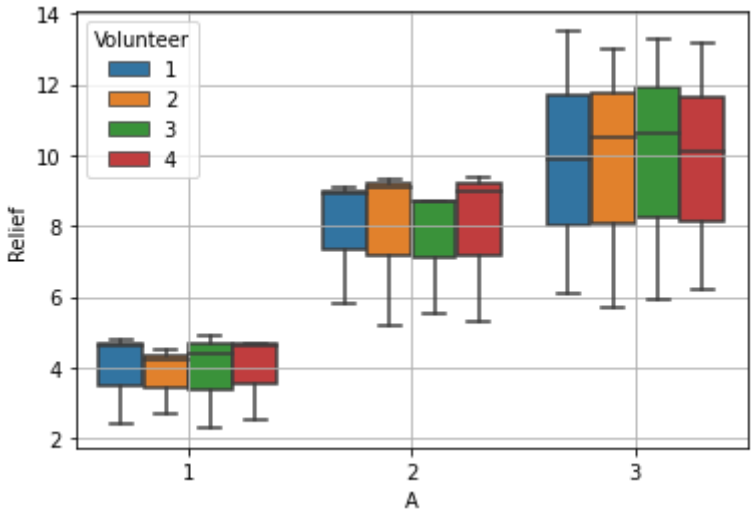
	A	B	Volunteer	Relief
count	36.000000	36.000000	36.000000	36.000000
mean	2.000000	2.000000	2.500000	7.183333
std	0.828079	0.828079	1.133893	3.272090
min	1.000000	1.000000	1.000000	2.300000
25%	1.000000	1.000000	1.750000	4.675000
50%	2.000000	2.000000	2.500000	6.000000
75%	3.000000	3.000000	3.250000	9.325000
max	3.000000	3.000000	4.000000	13.500000

Overall mean of the Relief column(dependent variable) is 7.183 and standard deviation is 3.272.

Missing values/Outliers in the dataset:

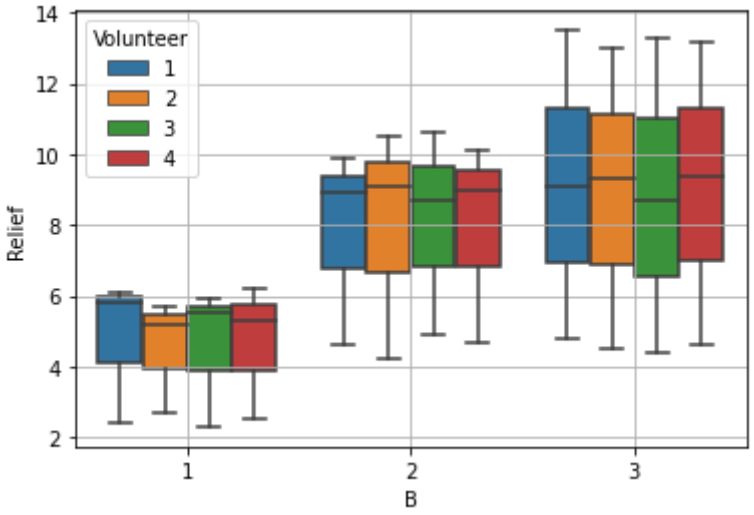
There are no missing values or outliers in the dataset.

Boxplot of ingredient A with Relief for different volunteers:



We can infer from the above plot that the third level of ingredient A has comparatively higher mean hours on Relief and the spread of data is quite large compared to other levels.

Boxplot of ingredient B with Relief for different volunteers:



From the above plot, the mean hours of Relief for second and third level of ingredient B is approximately same.

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

Independent variables: A, B

Dependent variable: Relief

For variable A:

Null Hypothesis:
$H0$: The mean hours of 'Relief' with respect to three levels of ingredient 'A' is equal. $\mu1=\mu2=\mu3$.

Alternate Hypothesis:
$H1$: At least one of the mean hours of 'Relief' with respect to three levels of ingredient 'A' is unequal. $\mu1=\mu2\neq\mu3$ (or) $\mu1\neq\mu2=\mu3$ (or) $\mu1\neq\mu3=\mu2$

For variable B:

Null Hypothesis:
$H0$: The mean hours of 'Relief' with respect to three levels of ingredient 'B' is equal. $\mu1=\mu2=\mu3$

Alternate Hypothesis:
$H1$: At least one of the mean hours of 'Relief' with respect to three levels of ingredient 'B' is unequal. $\mu1=\mu2\neq\mu3$ (or) $\mu1\neq\mu2=\mu3$ (or) $\mu1\neq\mu3=\mu2$

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.

ANOVA table:

	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

Inference:

Variation between the three levels of ingredient 'A' is 23 times the variation within the levels of ingredient 'A'. Probability that (this F-stat is large than 23) is small.

Here the p-value < alpha (0.05), the three levels of ingredient 'A' is a significant factor on the hours of 'Relief'. Reject Null Hypothesis that the mean hours of Relief variable with respect to the three levels of ingredient 'A' are equal.

Hence different levels of ingredient 'A' has significant effect on 'Hours of Relief' variable.

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.

ANOVA table:

	df	sum_sq	mean_sq	F	PR(>F)
C(B)	2.0	123.66	61.830000	8.126777	0.00135
Residual	33.0	251.07	7.608182	NaN	NaN

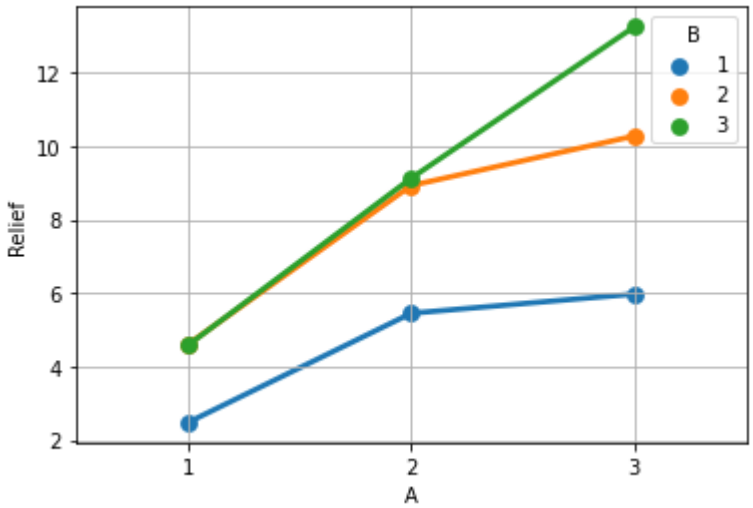
Inference:

Variation between the three levels of ingredient 'B' is 8 times the variation within the levels of ingredient 'A'. Probability that (this F-stat is large than 8) is small.

Here the p-value < alpha (0.05), the three levels of ingredient 'B' is a significant factor on the hours of 'Relief'. Reject Null Hypothesis that the mean hours of Relief variable with respect to the three levels of ingredient 'B' are equal.

Hence different levels of ingredient 'B' has significant effect on 'Hours of Relief' variable.

1.4) Analyse the effects of one variable on another with the help of an interaction plot.
What is the interaction between the two treatments?
[hint: use the 'pointplot' function from the 'seaborn' function]



Two treatments here is the effect of different levels of active ingredient A and B on the hours of relief.

Interaction is the simultaneous effect of two or more treatments on the response output(Relief). It occurs when the effect of one independent change depending on the level of another independent variable.

As seem from the above plot, there seems to be a little interaction effect between the two active ingredients(A & B) of the compound. In other way it can be stated that, the effect of one independent variable(active ingredient A) is not the same for all the levels of other independent variable(active ingredient B). For example, if we take the third level of active ingredient A, when its combined with the first level of ingredient B the mean hours of relief is 6(approximately), if its combined with the second level of ingredient B the hours of relief is around 10 and if the combination is with the third level of ingredient B then we get maximum hours of relief, 12 in this case.

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

Two-way ANOVA table(without interaction)

	df	sum_sq	mean_sq	F	PR(>F)
C(A)	2.0	220.02	110.010000	109.832850	8.514029e-15
C(B)	2.0	123.66	61.830000	61.730435	1.546749e-11
Residual	31.0	31.05	1.001613	NaN	NaN

Considering both the ingredients A & B, they are a significant factor on hours of Relief since the p-value for both is less than alpha(0.05).

Two-way ANOVA(with interaction)

Null Hypothesis:

There is no difference in the mean hours of Relief for different combinations of levels of the active ingredients A and B(meaning there is no interaction effect)

Alternative Hypothesis:

There is a difference in the mean hours of Relief for different combinations of levels of the active ingredients A and B(meaning there is an interaction effect)

If the alternate hypothesis is accepted, further analysis should be performed to explore where the individual differences are.

Table

	df	sum_sq	mean_sq	F	PR(>F)
C(A)	2.0	220.020	110.010000	1827.858462	1.514043e-29
C(B)	2.0	123.660	61.830000	1027.329231	3.348751e-26
C(A):C(B)	4.0	29.425	7.356250	122.226923	6.972083e-17
Residual	27.0	1.625	0.060185	NaN	NaN

The p-value of the interaction is less than alpha (0.05) which means statistically the interaction is meaningful and it has a significant effect on the hours of Relief. Upon introducing the interaction between the ingredients, A & B, the p-value for both the ingredients distinctly has also amplified which denotes the effect on Relief.

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

From the interaction plot, it can be seen that, overall, the hours of Relief increase with the levels of ingredient A and B with the third level of both the ingredients having maximum hours of relief. Even though the lines are not exactly intersecting, we cannot say that they are parallel as well , it can be said that there seems to be some kind of interaction among the active ingredients of the compound because the magnitude of difference between each level of active ingredient B is different at each levels of ingredient A.

Looking at the ANOVA table, p-value is $1.514e^{-29}$, $3.348e^{-26}$ and $6.972e^{-17}$ for ingredient A, ingredient B and interaction of ingredients A and B, respectively. Since all of these values are less than alpha(0.05), both active ingredients are needed, as well as their interaction, to explain the hours of relief.

From the results, it can be said that there is a strong evidence that the mean hours of relief vary with different levels of active ingredients A and B. The presence of interaction between the active ingredients A and B means that the hours of relief changes for different levels of A depends on varied levels of B ingredient and vice versa. If the other factors are well randomized for, then we can say that different levels of active ingredients are the main cause for relief(which is the effect).

Overall, third level of both active ingredient A and B has the maximum hours of relief.

Problem 2:

PCA – Principal Component Analysis

Problem statement:

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

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

Exploratory data analysis:

Head of the dataset: Verify that the dataset is loaded correctly.

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.
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	

Information of the dataset: There are a total of 18 variables out of which Names are object type and others are either float or int. We can also see that there are no missing values in the dataset.

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

Shape of the dataset:

There are 777 rows and 18 columns in the dataset.

Also, there are no duplicate values in the dataset.

Summary statistics of the dataset:

	Apps	Accept	Enroll	Top10perc	Top25perc
count	777.000000	777.000000	777.000000	777.000000	777.000000
mean	3001.638353	2018.804376	779.972973	27.558559	55.796654
std	3870.201484	2451.113971	929.176190	17.640364	19.804778
min	81.000000	72.000000	35.000000	1.000000	9.000000
25%	776.000000	604.000000	242.000000	15.000000	41.000000
50%	1558.000000	1110.000000	434.000000	23.000000	54.000000
75%	3624.000000	2424.000000	902.000000	35.000000	69.000000
max	48094.000000	26330.000000	6392.000000	96.000000	100.000000

	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books
count	777.000000	777.000000	777.000000	777.000000	777.000000
mean	3699.907336	855.298584	10440.669241	4357.526384	549.380952
std	4850.420531	1522.431887	4023.016484	1096.696416	165.105360
min	139.000000	1.000000	2340.000000	1780.000000	96.000000
25%	992.000000	95.000000	7320.000000	3597.000000	470.000000
50%	1707.000000	353.000000	9990.000000	4200.000000	500.000000
75%	4005.000000	967.000000	12925.000000	5050.000000	600.000000
max	31643.000000	21836.000000	21700.000000	8124.000000	2340.000000

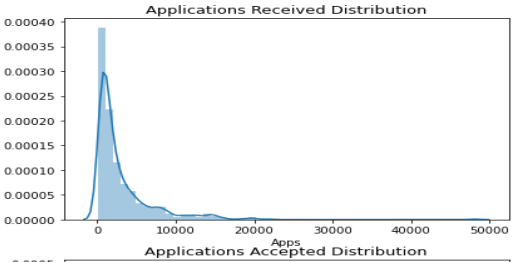
	Personal	PhD	Terminal	S.F.Ratio	perc.alumni
count	777.000000	777.000000	777.000000	777.000000	777.000000
mean	1340.642214	72.660232	79.702703	14.089704	22.743887
std	677.071454	16.328155	14.722359	3.958349	12.391801
min	250.000000	8.000000	24.000000	2.500000	0.000000
25%	850.000000	62.000000	71.000000	11.500000	13.000000
50%	1200.000000	75.000000	82.000000	13.600000	21.000000
75%	1700.000000	85.000000	92.000000	16.500000	31.000000
max	6800.000000	103.000000	100.000000	39.800000	64.000000

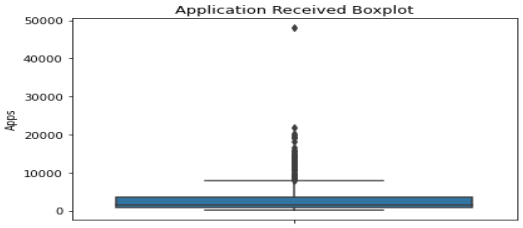
	Expend	Grad.Rate
count	777.000000	777.00000
mean	9660.171171	65.46332
std	5221.768440	17.17771
min	3186.000000	10.00000
25%	6751.000000	53.00000
50%	8377.000000	65.00000
75%	10830.000000	78.00000
max	56233.000000	118.00000

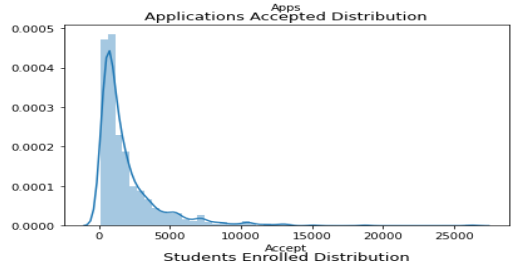
Almost all the variables have their mean greater than median, which means the data is right skewed. Except for the variable PhD and Terminal for which median is greater than mean and for Graduation Rate, mean and median are approximately equal. Therefore, we can conclude that there is skewness and outliers in the dataset which is further explained in the plots below.

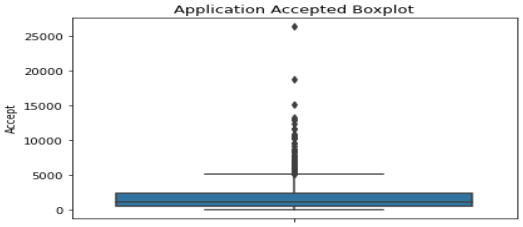
Univariate analysis: Taking into consideration single variable for analysis.

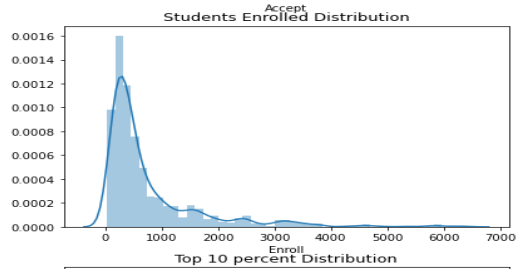
Boxplot and Distplot of all variables:

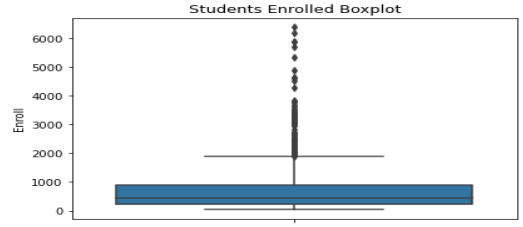

Applications Received Distribution
Apps

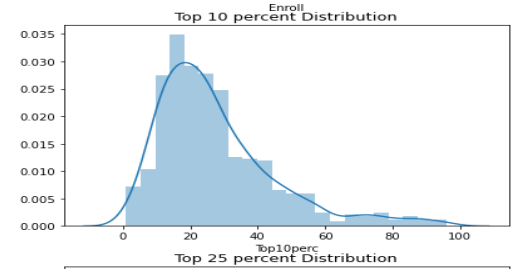

Application Received Boxplot
Apps

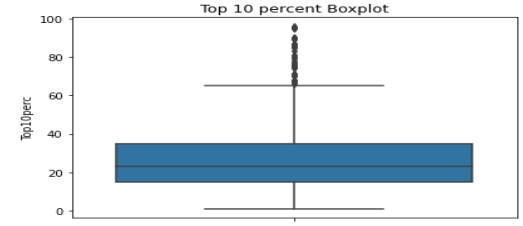

Applications Accepted Distribution
Accept

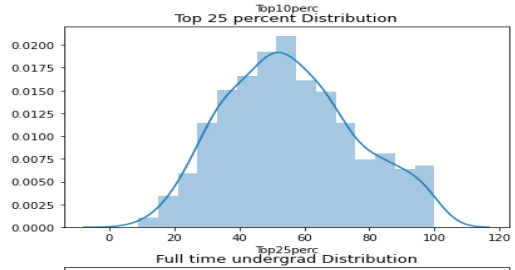

Application Accepted Boxplot
Accept

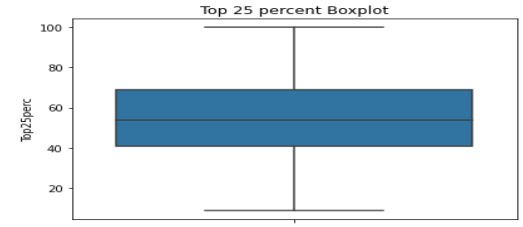

Students Enrolled Distribution
Enroll

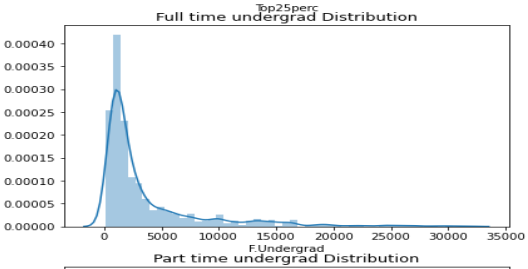

Students Enrolled Boxplot
Enroll

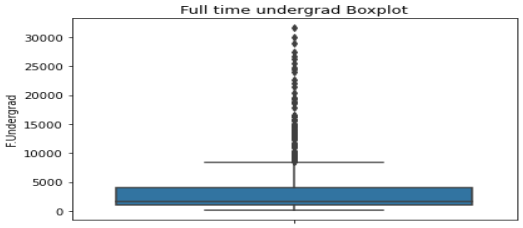

Top 10 percent Distribution
Top10perc

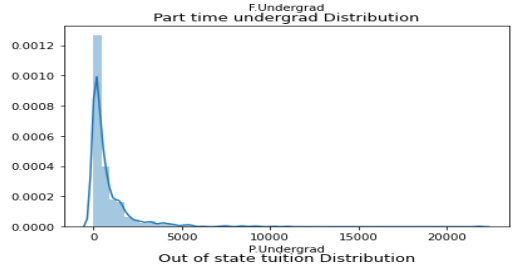

Top 10 percent Boxplot
Top10perc

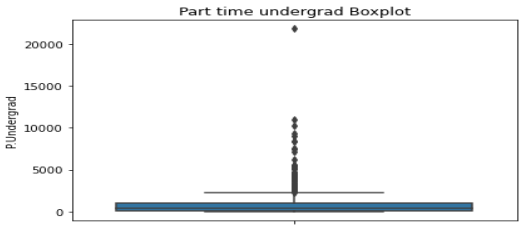

Top 25 percent Distribution
Top25perc

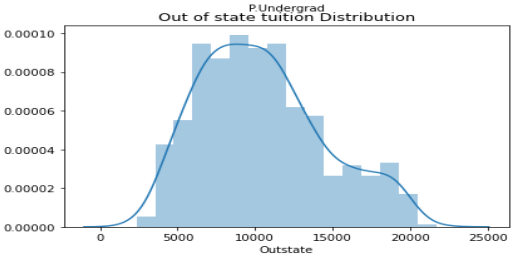

Top 25 percent Boxplot
Top25perc

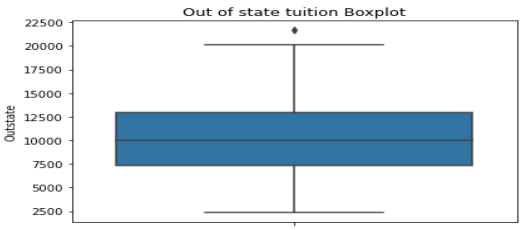

Full time undergrad Distribution
F.Undergrad

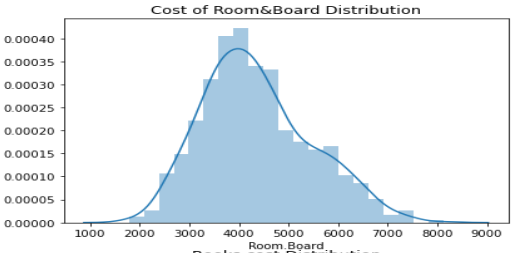

Full time undergrad Boxplot
F.Undergrad

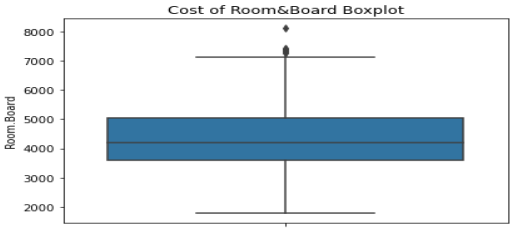

Part time undergrad Distribution
P.Undergrad

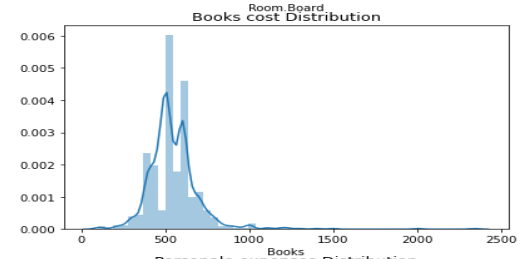

Part time undergrad Boxplot
P.Undergrad

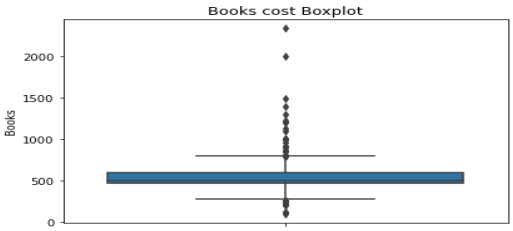

Out of state tuition Distribution
Outstate

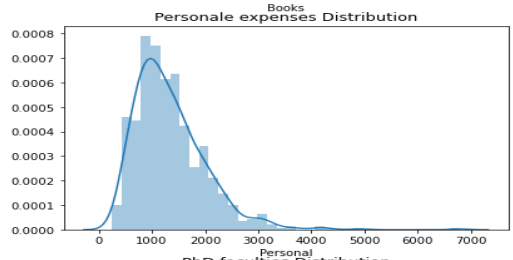

Out of state tuition Boxplot
Outstate

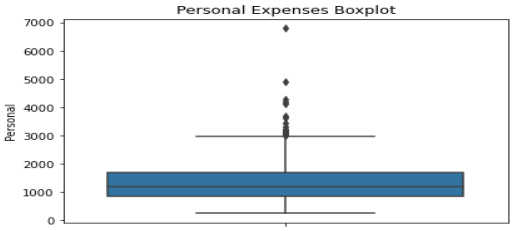

Cost of Room&Board Distribution
Room.Board

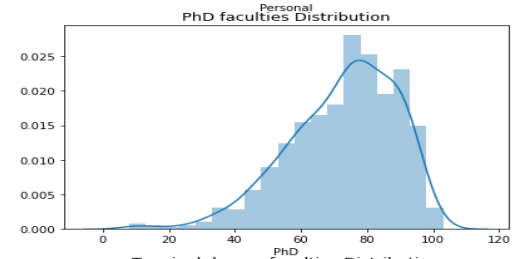

Cost of Room&Board Boxplot
Room.Board

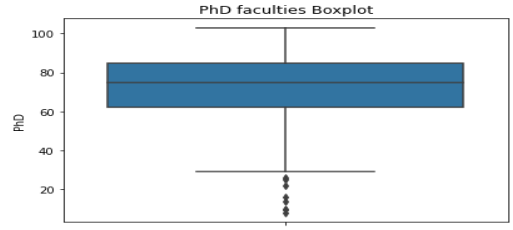

Books cost Distribution
Books

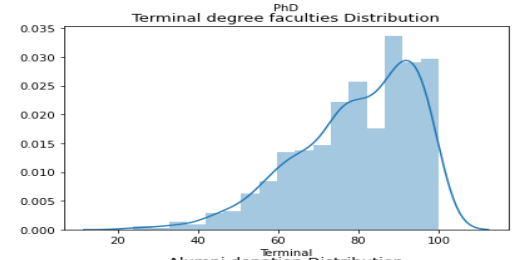

Books cost Boxplot
Books

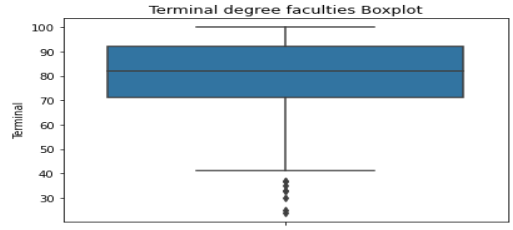

Personale expenses Distribution
Personal

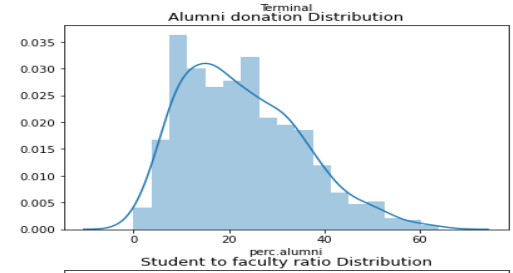

Personal Expenses Boxplot
Personal

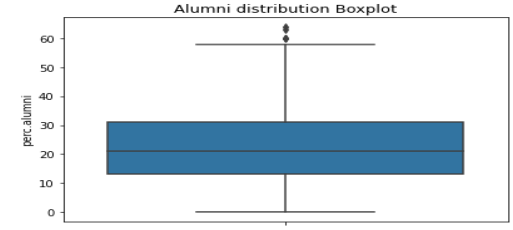

PhD faculties Distribution
PhD

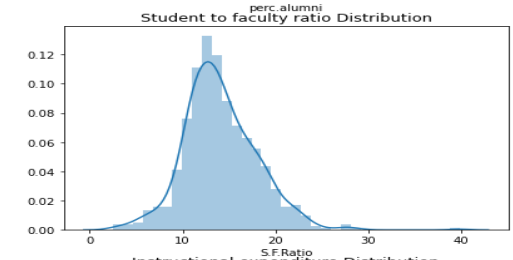

PhD faculties Boxplot
PhD

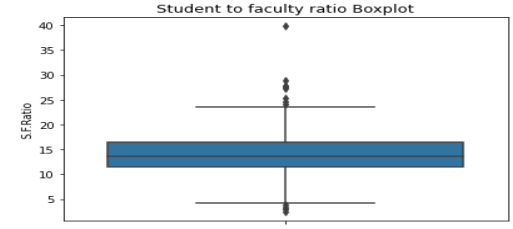

Terminal degree faculties Distribution
Terminal

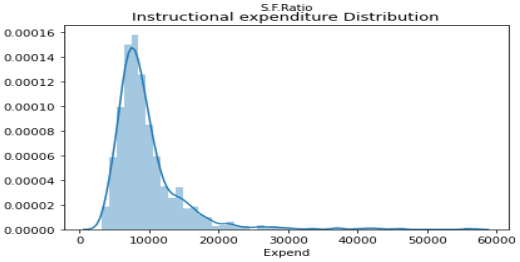

Terminal degree faculties Boxplot
Terminal

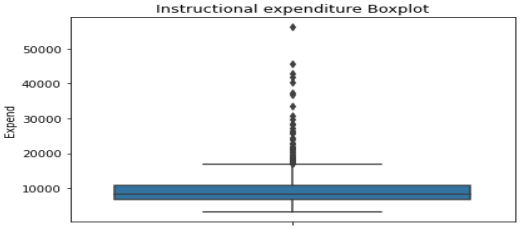

Alumni donation Distribution
perc.alumni

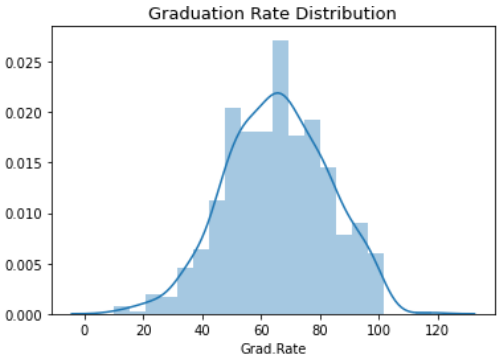

Alumni distribution Boxplot
perc.alumni

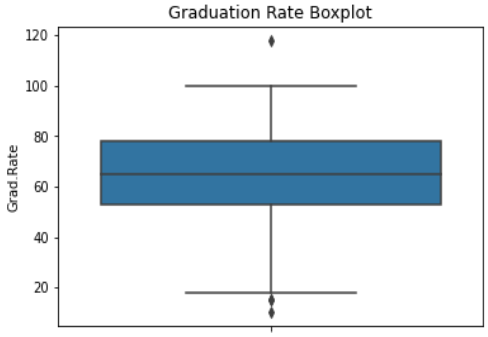

Student to faculty ratio Distribution
S.F.Ratio


Student to faculty ratio Boxplot
S.F.Ratio


Instructional expenditure Distribution
Expend


Instructional expenditure Boxplot
Expend





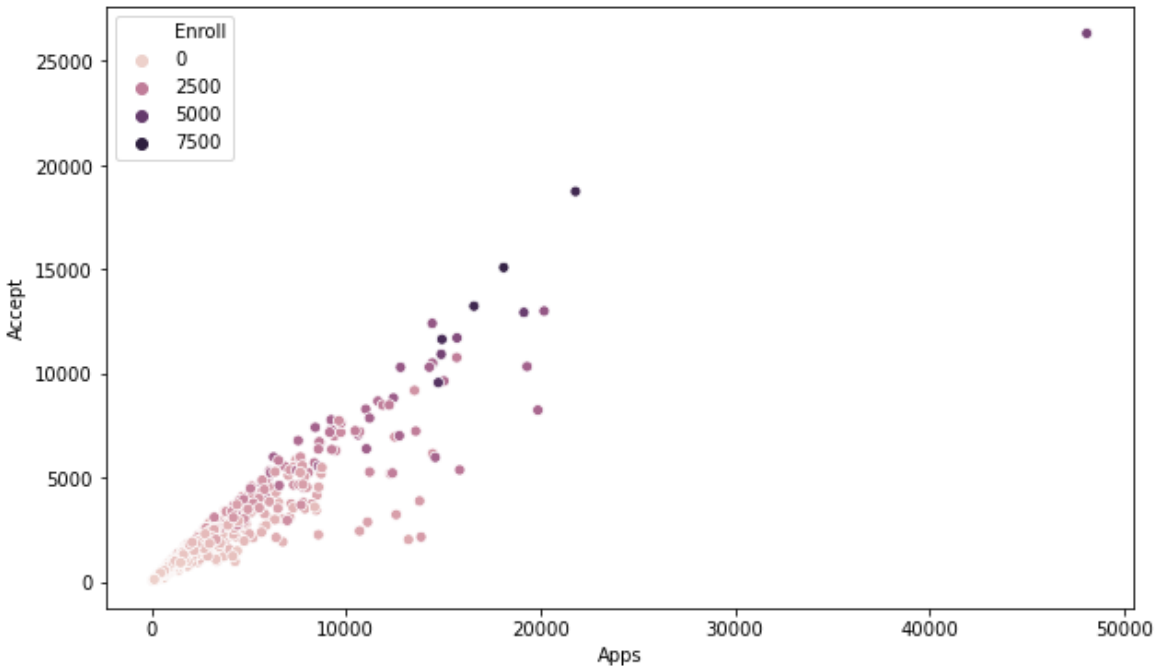
As observed from the above plots, out of 17 continuous variables, most of the variables have outliers except 'Top25percent'. Outstate, Alumni donations and Graduation rate variables has less outliers compared to the rest.

Skewness of the variables are visible in the dataset and as inferred before, all the variables are right skewed, except PhD and Terminal which are left skewed and Graduation rate approximately represents a normal distribution bell curve.

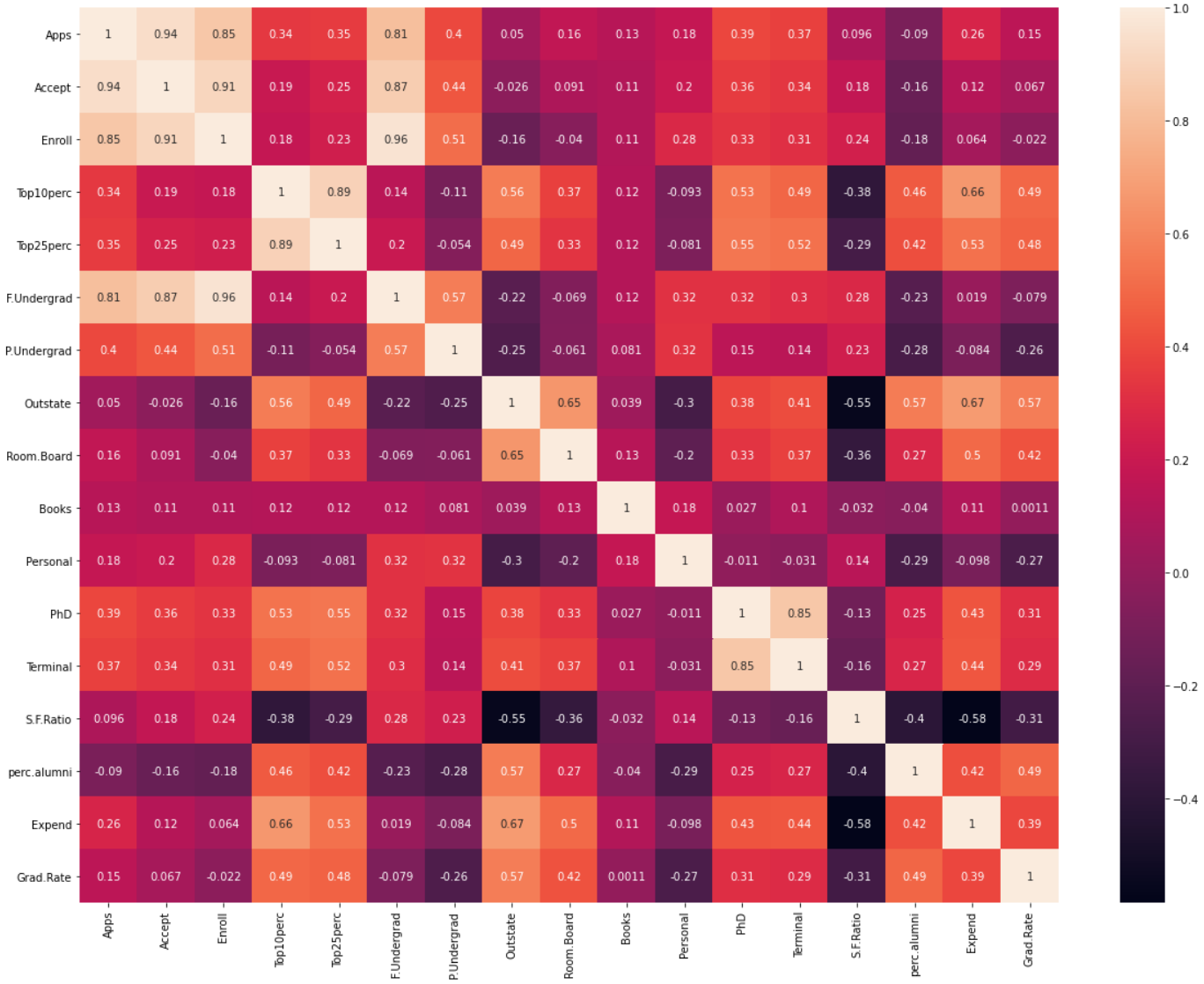
Bivariate and Multivariate Analysis of the dataset: Taking two or more variables into consideration for analysis.

Since, all the variables are continuous, we will be going with scatterplots and heatmap.

Scatter plot for Apps and Accept with hue as Enroll:



Heat map for the correlation:



From the above heatmap, we can see that the positive linear relationship is exhibited by the variable pairs Accept & Apps(0.94), Accept & Enroll(0.91), F.Undergrad & Enroll(0.96). And negative linear relationship exists between the variable pairs S.F.Ratio & Outstate(-0.55) and S.F.Ratio & Expend(-0.58). Other than that, all other variables correlation is not conclusive.

(The pair plot of above correlation is displayed the Notebook file which explains the above relation visually)

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

The dataset contains variables which does not belong to the same scale, like there are variables which indicates the number of students(Apps, Accept, Enroll, F.Undergrad, P.Undergrad, etc,.) and there are variables which indicates the fees(Books, Personal, Expend, etc,.) and ratio depicting variables as well. Since these variables are difficult to compare as such, we need to transform the variables.

For this case study, standard z-score scaling is used which converts the group of data in our distribution such that the mean is 0 and standard deviation is 1. Z-score is expressed in terms of standard deviations from their means. It converts the dataset within a range of (-3,3) provided the dataset is free from outliers or skewness.

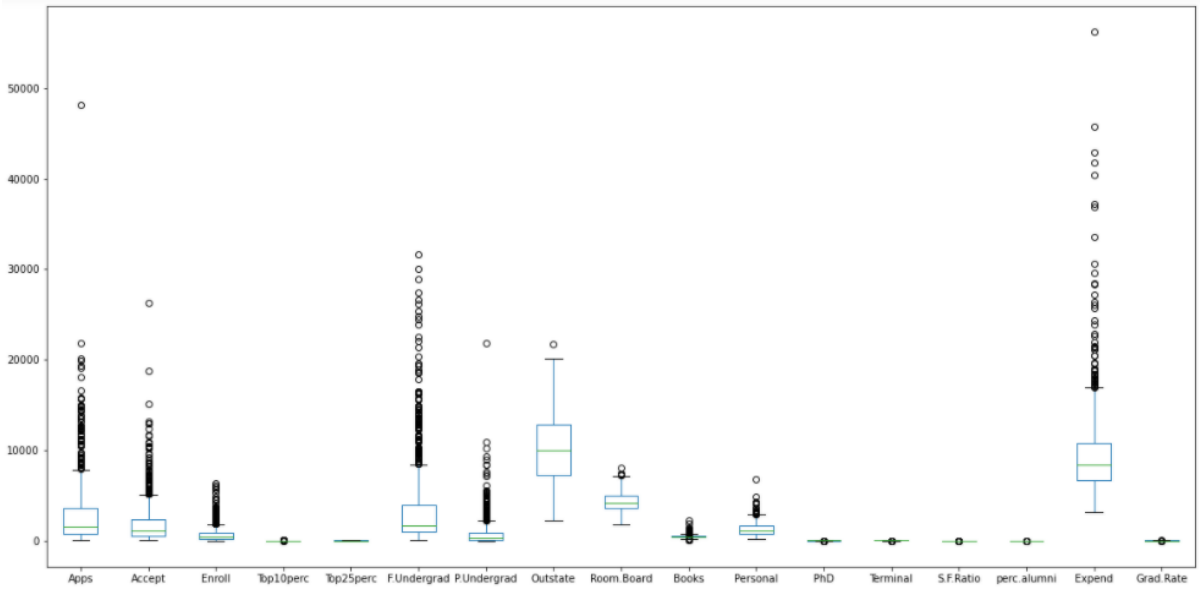
Calculating the Standard Score (Z-Score)

$$\text{Standard Score, } z = \frac{X - \mu}{\sigma}$$

TERMS:
μ = mean (pronounced 'mu')
X = score
σ = standard deviation (pronounced 'sigma')

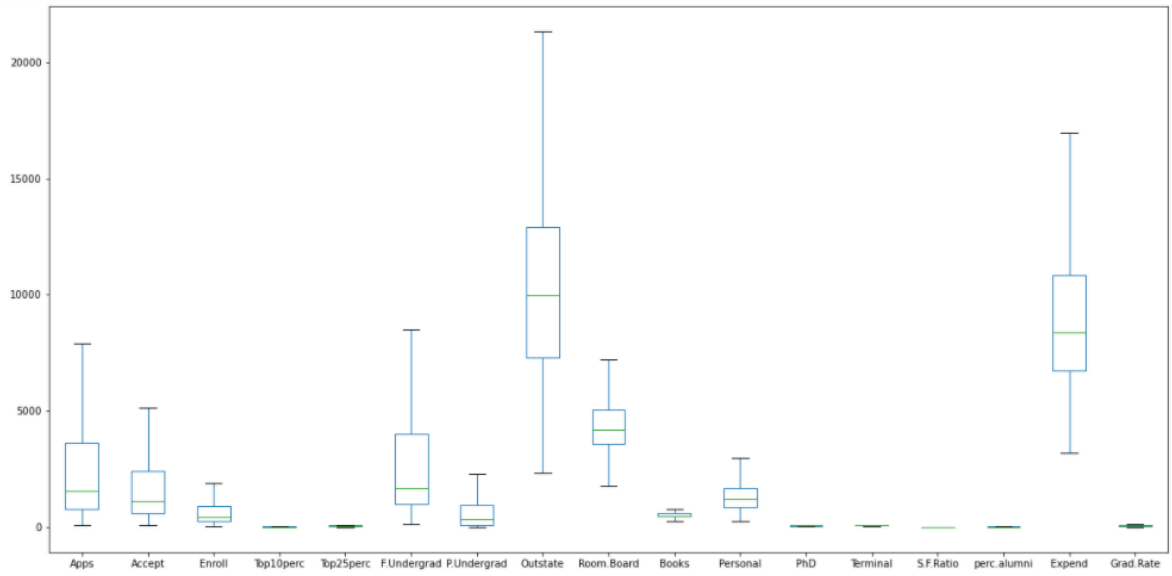
Outlier treatment: Before scaling the variables, since the dataset contains outliers, we have to treat the outliers in order for the output to be valid because if the scaling is done on the dataset with outliers then it would result in meaningless mean and standard deviation.

Box plot of all variables:



IQR treatment for outliers: Custom function is defined which takes column as input and returns two output for a particular column if the value is greater than maximum limit or less than minimum limit. Loop the function for all the variables such that it replaces the values greater than maximum limit by that limit and vice versa.

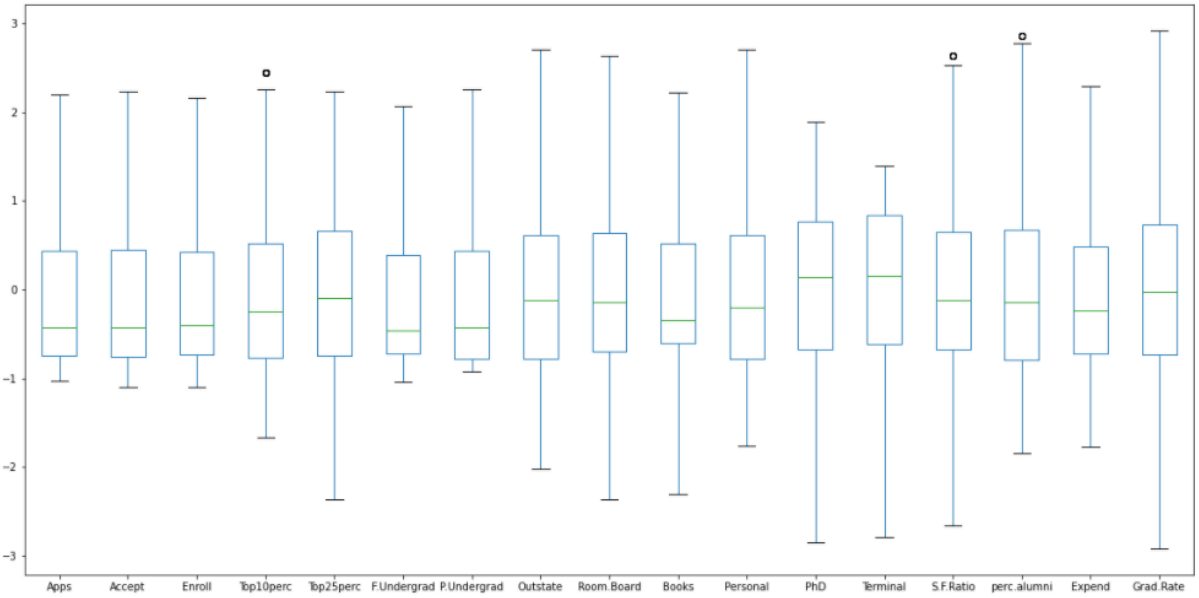
Boxplot of all variables after the outlier treatment:



Result after doing z-score scaling for the new dataset,

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.R
0	-0.376493	-0.337830	0.106380	-0.246780	-0.191827	-0.018769	-0.166083	-0.746480	-0.968324	-0.776567	1.438500	-0.174045	-0.123239	1.070
1	-0.159195	0.116744	-0.260441	-0.696290	-1.353911	-0.093626	0.797856	0.457762	1.921680	1.828605	0.289289	-2.745731	-2.785068	-0.489
2	-0.472336	-0.426511	-0.569343	-0.310996	-0.292878	-0.703966	-0.777974	0.201488	-0.555466	-1.210762	-0.260691	-1.240354	-0.952900	-0.304
3	-0.889994	-0.917871	-0.918613	2.129202	1.677612	-0.898889	-0.828267	0.626954	1.004218	-0.776567	-0.736792	1.205884	1.190391	-1.679
4	-0.982532	-1.051221	-1.062533	-0.696290	-0.596031	-0.995610	0.297726	-0.716623	-0.216006	2.219381	0.289289	0.202299	-0.538069	-0.568

Boxplot of the above scaled dataset:



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

Covariance matrix explains total variances of individual dimensions through one matrix in which the diagonal represents the variances of each variable with itself and off-diagonal elements explains the covariance of variables with respect to other variables. It gives the direction of the linear relationship between variables.

If covariance is done for centred variables the diagonal becomes 1, and off-diagonal represents the correlation. Scaling converts the covariance matrix to **Correlation matrix** for standardized variables. Correlation represents both the strength and direction of the linear relationship.

Covariance matrix after scaling

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Termi
Apps	1.001289	0.956538	0.898039	0.321756	0.364961	0.862111	0.520493	0.065421	0.187717	0.236442	0.230244	0.464522	0.435
Accept	0.956538	1.001289	0.936482	0.223586	0.274033	0.898190	0.573429	-0.005009	0.119740	0.208974	0.256676	0.427891	0.403
Enroll	0.898039	0.936482	1.001289	0.171977	0.230731	0.968549	0.642422	-0.155856	-0.023876	0.202317	0.339785	0.382031	0.354
Top10perc	0.321756	0.223586	0.171977	1.001289	0.915053	0.111358	-0.180241	0.562884	0.357826	0.153650	-0.116880	0.544749	0.507
Top25perc	0.364961	0.274033	0.230731	0.915053	1.001289	0.181429	-0.099423	0.490200	0.331413	0.169980	-0.086922	0.552172	0.528
F.Undergrad	0.862111	0.898190	0.968549	0.111358	0.181429	1.001289	0.697027	-0.226457	-0.054546	0.208147	0.360246	0.362030	0.335
P.Undergrad	0.520493	0.573429	0.642422	-0.180241	-0.099423	0.697027	1.001289	-0.354673	-0.067725	0.122686	0.344496	0.127827	0.122
Outstate	0.065421	-0.005009	-0.155856	0.562884	0.490200	-0.226457	-0.354673	1.001289	0.656334	0.005117	-0.326029	0.391825	0.413
Room.Board	0.187717	0.119740	-0.023876	0.357826	0.331413	-0.054546	-0.067725	0.656334	1.001289	0.109065	-0.219837	0.341909	0.379
Books	0.236442	0.208974	0.202317	0.153650	0.169980	0.208147	0.122686	0.005117	0.109065	1.001289	0.240172	0.136566	0.159
Personal	0.230244	0.256676	0.339785	-0.116880	-0.086922	0.360246	0.344496	-0.326029	-0.219837	0.240172	1.001289	-0.011699	-0.032
PhD	0.464522	0.427891	0.382031	0.544749	0.552172	0.362030	0.127827	0.391825	0.341909	0.136566	-0.011699	1.001289	0.864
Terminal	0.435038	0.403929	0.354836	0.507401	0.528334	0.335486	0.122309	0.413110	0.379759	0.159523	-0.032012	0.864040	1.001
S.F.Ratio	0.126574	0.188749	0.274622	-0.388426	-0.297616	0.324922	0.371085	-0.574422	-0.376915	-0.008547	0.174137	-0.129556	-0.151
perc.alumni	-0.101288	-0.165729	-0.223010	0.456384	0.417369	-0.285825	-0.419874	0.566465	0.272744	-0.042887	-0.306147	0.249198	0.266
Expend	0.243248	0.162017	0.054291	0.657886	0.573643	0.000371	-0.202189	0.776327	0.581370	0.150177	-0.163481	0.511187	0.524
Grad.Rate	0.150998	0.079084	-0.023281	0.494307	0.479602	-0.082345	-0.265499	0.573196	0.426339	-0.008061	-0.291269	0.310419	0.293

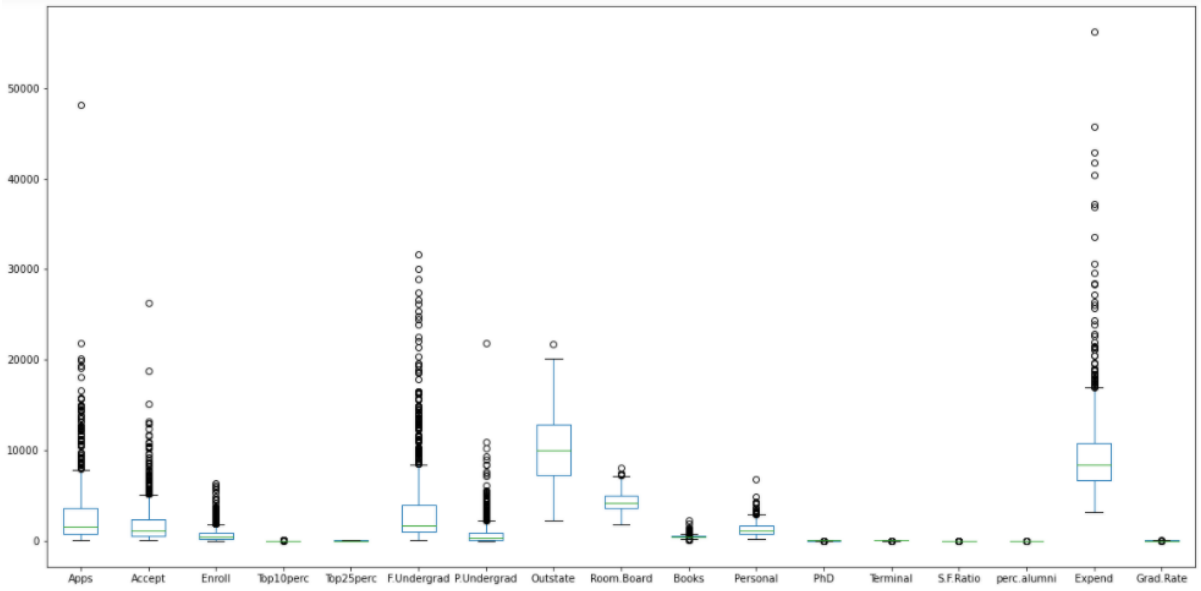
Correlation matrix after scaling

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Termi
Apps	1.000000	0.955307	0.896883	0.321342	0.364491	0.861002	0.519823	0.065337	0.187475	0.236138	0.229948	0.463924	0.434
Accept	0.955307	1.000000	0.935277	0.223298	0.273681	0.897034	0.572691	-0.005002	0.119586	0.208705	0.256346	0.427341	0.403
Enroll	0.896883	0.935277	1.000000	0.171756	0.230434	0.967302	0.641595	-0.155655	-0.023846	0.202057	0.339348	0.381540	0.354
Top10perc	0.321342	0.223298	0.171756	1.000000	0.913875	0.111215	-0.180009	0.562160	0.357366	0.153452	-0.116730	0.544048	0.506
Top25perc	0.364491	0.273681	0.230434	0.913875	1.000000	0.181196	-0.099295	0.489569	0.330987	0.169761	-0.086810	0.551461	0.527
F.Undergrad	0.861002	0.897034	0.967302	0.111215	0.181196	1.000000	0.696130	-0.226166	-0.054476	0.207879	0.359783	0.361564	0.335
P.Undergrad	0.519823	0.572691	0.641595	-0.180009	-0.099295	0.696130	1.000000	-0.354216	-0.067638	0.122529	0.344053	0.127663	0.122
Outstate	0.065337	-0.005002	-0.155655	0.562160	0.489569	-0.226166	-0.354216	1.000000	0.655489	0.005110	-0.325609	0.391321	0.412
Room.Board	0.187475	0.119586	-0.023846	0.357366	0.330987	-0.054476	-0.067638	0.655489	1.000000	0.108924	-0.219554	0.341469	0.379
Books	0.236138	0.208705	0.202057	0.153452	0.169761	0.207879	0.122529	0.005110	0.108924	1.000000	0.239863	0.136390	0.159
Personal	0.229948	0.256346	0.339348	-0.116730	-0.086810	0.359783	0.344053	-0.325609	-0.219554	0.239863	1.000000	-0.011684	-0.031
PhD	0.463924	0.427341	0.381540	0.544048	0.551461	0.361564	0.127663	0.391321	0.341469	0.136390	-0.011684	1.000000	0.862
Terminal	0.434478	0.403409	0.354379	0.506748	0.527654	0.335054	0.122152	0.412579	0.379270	0.159318	-0.031971	0.862928	1.000
S.F.Ratio	0.126411	0.188506	0.274269	-0.387926	-0.297233	0.324504	0.370607	-0.573683	-0.376430	-0.008536	0.173913	-0.129390	-0.150
perc.alumni	-0.101158	-0.165516	-0.222723	0.455797	0.416832	-0.285457	-0.419334	0.565736	0.272393	-0.042832	-0.305753	0.248877	0.266
Expend	0.242935	0.161808	0.054221	0.657039	0.572905	0.000371	-0.201929	0.775328	0.580622	0.149983	-0.163271	0.510529	0.524
Grad.Rate	0.150803	0.078982	-0.023251	0.493670	0.478985	-0.082239	-0.265158	0.572458	0.425790	-0.008051	-0.290894	0.310019	0.292

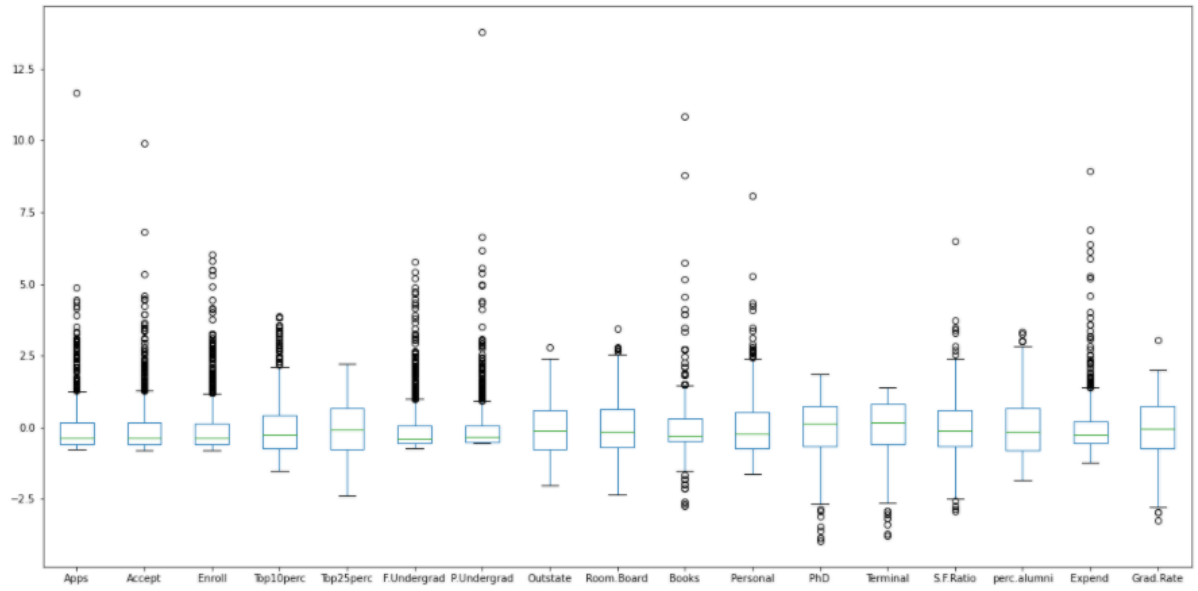
After scaling, both covariance and correlation matrix are one and the same.

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

Boxplot to show the outliers before scaling:



Boxplot for outliers after scaling(without outlier treatment):



We could see that most of the outliers are still present even after scaling. Scaling results are not significant if the original dataset have outliers.

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

PCA uses the concept of Eigen Decomposition, in which the covariance matrix is decomposed to eigen values and eigen vectors.

Eigen vectors are used to understand the directions of spread of our data and eigen values are those corresponding variances. In other words, Eigen vectors gives the principal components and eigen values gives the explained variances of the components

Covariance matrix:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Termi
Apps	1.001289	0.956538	0.898039	0.321756	0.364961	0.862111	0.520493	0.065421	0.187717	0.236442	0.230244	0.464522	0.435
Accept	0.956538	1.001289	0.936482	0.223586	0.274033	0.898190	0.573429	-0.005009	0.119740	0.208974	0.256676	0.427891	0.403
Enroll	0.898039	0.936482	1.001289	0.171977	0.230731	0.968549	0.642422	-0.155856	-0.023876	0.202317	0.339785	0.382031	0.354
Top10perc	0.321756	0.223586	0.171977	1.001289	0.915053	0.111358	-0.180241	0.562884	0.357826	0.153650	-0.116880	0.544749	0.507
Top25perc	0.364961	0.274033	0.230731	0.915053	1.001289	0.181429	-0.099423	0.490200	0.331413	0.169980	-0.086922	0.552172	0.528
F.Undergrad	0.862111	0.898190	0.968549	0.111358	0.181429	1.001289	0.697027	-0.226457	-0.054546	0.208147	0.360246	0.362030	0.335
P.Undergrad	0.520493	0.573429	0.642422	-0.180241	-0.099423	0.697027	1.001289	-0.354673	-0.067725	0.122686	0.344496	0.127827	0.122
Outstate	0.065421	-0.005009	-0.155856	0.562884	0.490200	-0.226457	-0.354673	1.001289	0.656334	0.005117	-0.326029	0.391825	0.413
Room.Board	0.187717	0.119740	-0.023876	0.357826	0.331413	-0.054546	-0.067725	0.656334	1.001289	0.109065	-0.219837	0.341909	0.379
Books	0.236442	0.208974	0.202317	0.153650	0.169980	0.208147	0.122686	0.005117	0.109065	1.001289	0.240172	0.136566	0.159
Personal	0.230244	0.256676	0.339785	-0.116880	-0.086922	0.360246	0.344496	-0.326029	-0.219837	0.240172	1.001289	-0.011699	-0.032
PhD	0.464522	0.427891	0.382031	0.544749	0.552172	0.362030	0.127827	0.391825	0.341909	0.136566	-0.011699	1.001289	0.864
Terminal	0.435038	0.403929	0.354836	0.507401	0.528334	0.335486	0.122309	0.413110	0.379759	0.159523	-0.032012	0.864040	1.001
S.F.Ratio	0.126574	0.188749	0.274622	-0.388426	-0.297616	0.324922	0.371085	-0.574422	-0.376915	-0.008547	0.174137	-0.129556	-0.151
perc.alumni	-0.101288	-0.165729	-0.223010	0.456384	0.417369	-0.285825	-0.419874	0.566465	0.272744	-0.042887	-0.306147	0.249198	0.266
Expend	0.243248	0.162017	0.054291	0.657886	0.573643	0.000371	-0.202189	0.776327	0.581370	0.150177	-0.163481	0.511187	0.524
Grad.Rate	0.150998	0.079084	-0.023281	0.494307	0.479602	-0.082345	-0.265499	0.573196	0.426339	-0.008061	-0.291269	0.310419	0.293

Eigen values: There are total of 17 eigen values

These are the

```
The eigen values are:
 [5.6625219  4.89470815 1.12636744 1.00397659 0.87218426 0.7657541
 0.58491404 0.5445048  0.42352336 0.38101777 0.24701456 0.02239369
 0.03789395 0.14726392 0.13434483 0.09883384 0.07469003]
```

Eigen vectors: Its dimension will be the total number of variables taken into consideration. Here it is 17 x 17.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	
0	-0.262172	0.314136	0.081018	-0.098776	-0.219898	0.002188	-0.028372	-0.089950	0.130567	-0.156464	-0.086213	0.182170	-0.599138	0.089978	0.0
1	-0.230562	0.344624	0.107659	-0.118140	-0.189635	-0.016521	-0.012958	-0.137606	0.142276	-0.149210	-0.042590	-0.391042	0.661497	0.158862	0.0
2	-0.189276	0.382813	0.085530	-0.009307	-0.162315	-0.068079	-0.015240	-0.144217	0.050871	-0.064900	-0.043841	0.716685	0.233235	-0.035399	-0.0
3	-0.338875	-0.099319	-0.078829	0.369115	-0.157211	-0.088866	-0.257455	0.289539	-0.122468	-0.035878	0.001778	-0.056205	0.022145	-0.039228	0.0
4	-0.334691	-0.059506	-0.050794	0.416824	-0.144449	-0.027627	-0.239039	0.345644	-0.193936	0.006418	-0.102127	0.019674	0.032265	0.145622	-0.0
5	-0.163293	0.398636	0.073708	-0.013950	-0.102728	-0.051647	-0.031175	-0.108749	0.001455	-0.000164	-0.034999	-0.542775	-0.367681	-0.133556	-0.0
6	-0.022480	0.357550	0.040357	-0.225351	0.095679	-0.024538	-0.010014	0.123842	-0.634774	0.546346	0.252107	0.029503	0.026249	0.050249	0.0
7	-0.283547	-0.251864	0.014939	-0.262975	-0.037275	-0.020386	0.094537	0.011272	-0.008366	-0.231800	0.593433	0.001034	-0.081425	0.560393	0.0
8	-0.244187	-0.131909	-0.021138	-0.580894	0.069108	0.237267	0.094521	0.389639	-0.220527	-0.255108	-0.475297	0.009857	0.026778	-0.107366	0.0
9	-0.096708	0.093974	-0.697121	0.036156	-0.035406	0.638605	-0.111193	-0.239817	0.021025	0.091162	0.043570	0.004361	0.010462	0.051622	0.0
10	0.035230	0.232440	-0.530973	0.114983	0.000475	-0.381496	0.639418	0.277207	0.017372	-0.127648	0.015163	-0.010873	0.004546	0.009394	-0.0
11	-0.326411	0.055139	0.081113	0.147261	0.550787	0.003344	0.089232	-0.034263	0.166510	0.100975	-0.039187	0.013315	0.012514	-0.071659	0.7
12	-0.323116	0.043033	0.058979	0.089008	0.590407	0.035412	0.091699	-0.090308	0.112609	0.086036	-0.084858	0.007381	-0.017928	0.163821	-0.6
13	0.163152	0.259805	0.274151	0.259486	0.142843	0.468753	0.152865	0.242808	-0.153685	-0.470528	0.363043	0.008858	0.018306	-0.239903	-0.0
14	-0.186611	-0.257093	0.103716	0.223982	-0.128216	0.012567	0.391401	-0.566073	-0.539236	-0.147629	-0.173919	-0.024053	-0.000080	-0.048975	0.0
15	-0.328956	-0.160009	-0.184206	-0.213756	0.022424	-0.231562	-0.150501	-0.118824	0.024237	-0.080415	0.393723	0.010566	0.056007	-0.690417	-0.1
16	-0.238822	-0.167524	0.245336	0.036192	-0.356843	0.313556	0.468642	0.180459	0.315813	0.488415	0.087264	-0.002510	0.014841	-0.159332	-0.0

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

Explicit form of the first Principal component:

0.262 * Apps + **0.231** * Accept + **0.189** * Enroll +
0.339 * Top10perc + **0.335** * Top25perc + **0.163** * F.Undergrad+
0.022 * P.Undergrad + **0.284** * Outstate + **0.244** * Room.Board +
0.097 * Books - **0.035** * Personal + **0.326** * PhD +
0.323 * Terminal - **0.163** * S.F.Ratio + **0.187** * perc.alumni
+ **0.329** * Expend + **0.239** * 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.

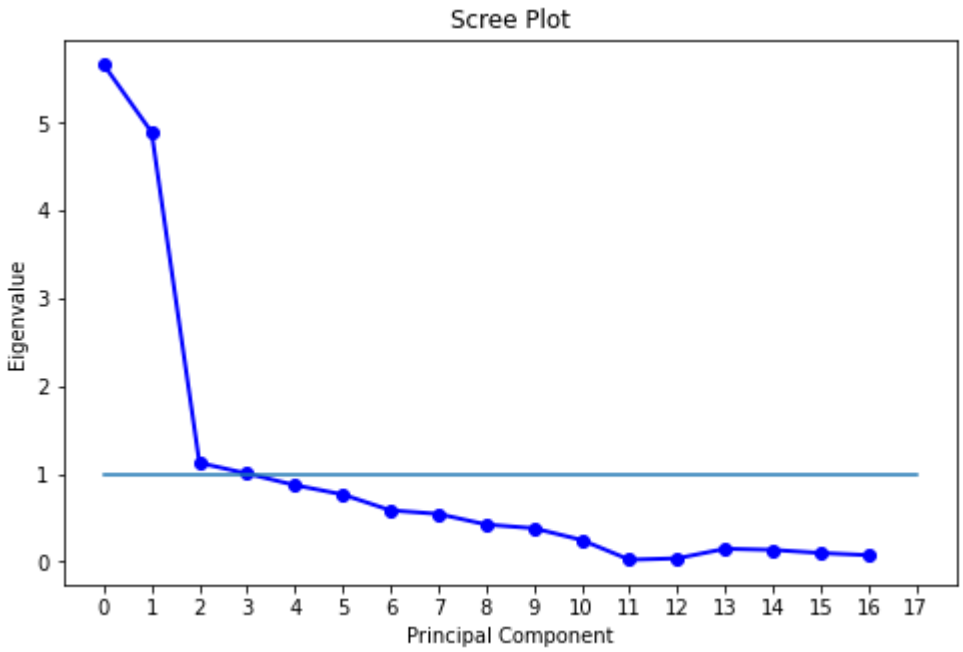
Eigen vectors of the covariance matrix indicates the direction of the axes where there is most variances(most information) which are also called as principal components. And eigen values are simply coefficients attached to the eigen vectors, which gives the amount of variance carried in each principal component.

Cumulative eigen values:

```
Cumulative Eigen Values [ 5.6625219  10.55723004 11.68359748 12.6875740
8 13.55975833 14.32551243 14.91042647 15.45493127 15.87845463 16.259472
39 16.50648695 16.52888065 16.5667746  16.71403851 16.84838334 16.94721
718 17.02190722]
```

For scaled data, the total eigen values will be approximately equal to the number of variables(or principal components).

Scree plot:



There is a total of four principal components which has eigen value as greater than(or equal to) 1.

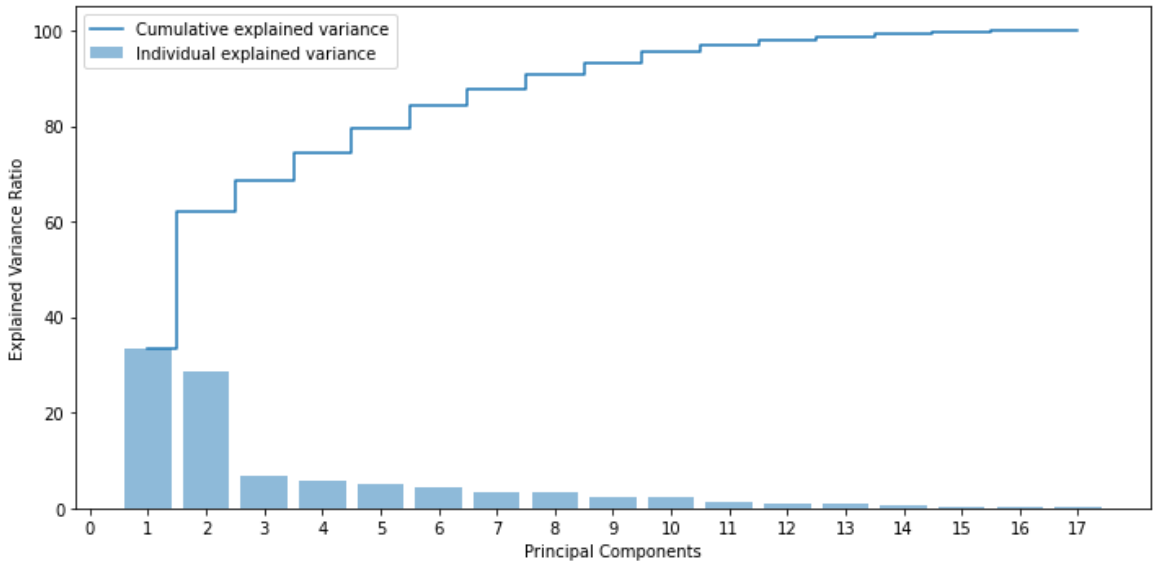
Individual explained variance by the eigen values:

```
array([33.3, 28.8,  6.6,  5.9,  5.1,  4.5,  3.4,  3.2,  2.5,  2.2,  1.5
,0.9,  0.8,  0.6,  0.4,  0.2,  0.1])
```

Cumulative variance explained by the eigen values:

```
Cumulative Variance Explained [ 33.26608367  62.02142867  68.63859223
74.53673619  79.66062886  84.15926753  87.59551019  90.79435736  93.282
46491  95.52086136  96.97201814  97.83716159  98.62640821  99.20703552
99.64582321  99.86844192 100.        ]
```

Their cumulative variance explained is the ratio of each eigen value to the sum of all eigen values.



Optimum number of principal components:

For determining the number of principal components to be considered, we can either go with **scree plot** which represents the Principal components along with their eigen values. Therefore, the PC's which have eigen value greater than 1 can be considered as the optimum number.

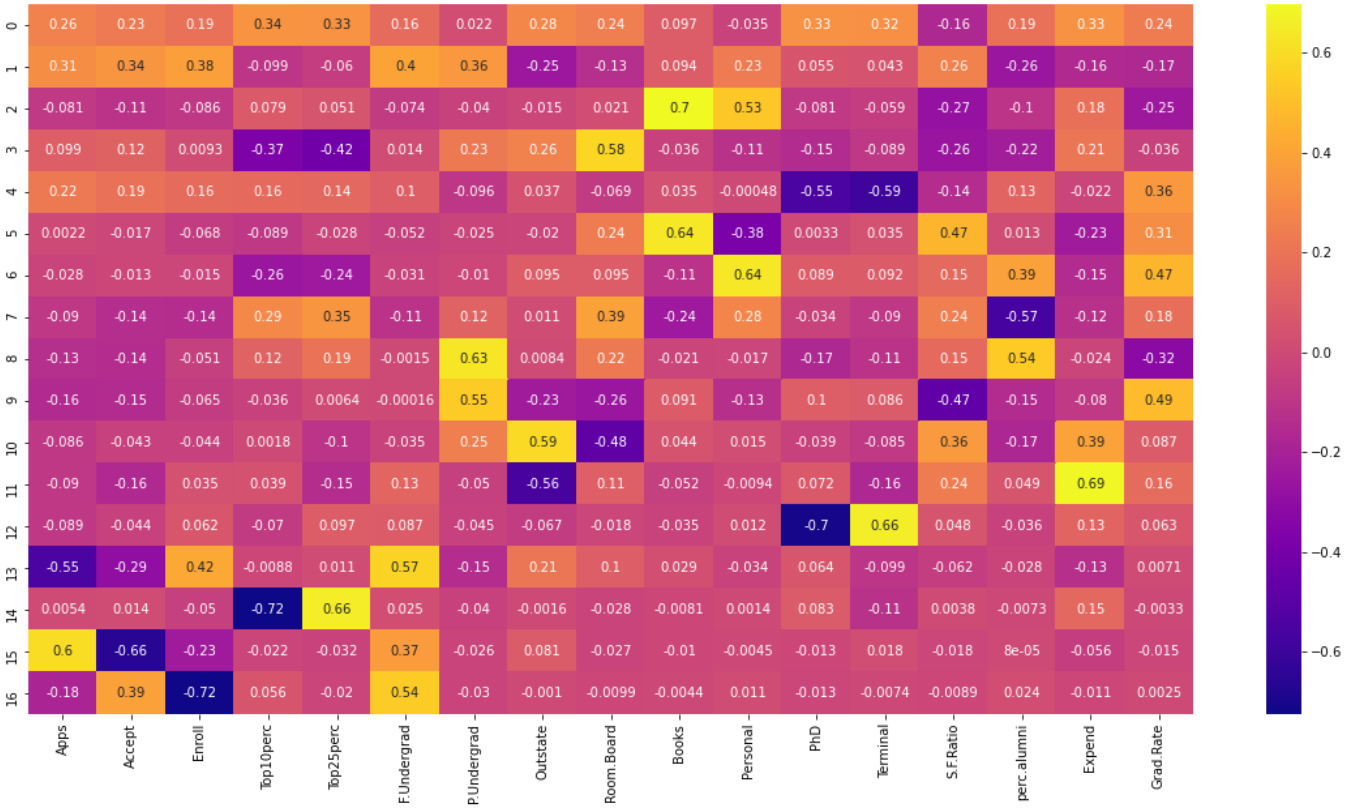
Else, we can consider the explained variance ratio vs Principal component plot, in which the principal components which contributes to 80% variation can be considered as the optimum number.

From this explained variance ratio plot, we can infer that 6 principal components have cumulative explained variance of 84.15%. Hence for this problem analysis, I will be going with 6 principal components.

After performing PCA, the principal component scores of the variables are:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	0.262172	0.230562	0.189276	0.338875	0.334691	0.163293	0.02248	0.283547	0.244187	0.096708	-0.03523	0.326411	0.323116	-0.163152	0.186611	0.328956	0.238822
1	0.314136	0.344624	0.382813	-0.099319	-0.059506	0.398636	0.35755	-0.251864	-0.131909	0.093974	0.23244	0.055139	0.043033	0.259805	-0.257093	-0.160009	-0.167524
2	-0.081018	-0.107659	-0.08553	0.078829	0.050794	-0.073708	-0.040357	-0.014939	0.021138	0.697121	0.530973	-0.081113	-0.058979	-0.274151	-0.103716	0.184206	-0.245336
3	0.098776	0.11814	0.009307	-0.369115	-0.416824	0.01395	0.225351	0.262975	0.580894	-0.036156	-0.114983	-0.147261	-0.089008	-0.259486	-0.223982	0.213756	-0.036192
4	0.219898	0.189635	0.162315	0.157211	0.144449	0.102728	-0.095679	0.037275	-0.069108	0.035406	-0.000475	-0.550787	-0.590407	-0.142843	0.128216	-0.022424	0.356843
5	0.002188	-0.016521	-0.068079	-0.088866	-0.027627	-0.051647	-0.024538	-0.020386	0.237267	0.638605	-0.381496	0.003344	0.035412	0.468753	0.012567	-0.231562	0.313556
6	-0.028372	-0.012958	-0.01524	-0.257455	-0.239039	-0.031175	-0.010014	0.094537	0.094521	-0.111193	0.639418	0.089232	0.091699	0.152865	0.391401	-0.150501	0.468642
7	-0.08995	-0.137606	-0.144217	0.289539	0.345644	-0.108749	0.123842	0.011272	0.389639	-0.239817	0.277207	-0.034263	-0.090308	0.242808	-0.566073	-0.118824	0.180459
8	-0.130567	-0.142276	-0.050871	0.122468	0.193936	-0.001455	0.634774	0.008366	0.220527	-0.021025	-0.017372	-0.16651	-0.112609	0.153685	0.539236	-0.024237	-0.315813
9	-0.156464	-0.14921	-0.0649	-0.035878	0.006418	-0.000164	0.546346	-0.2318	-0.255108	0.091162	-0.127648	0.100975	0.086036	-0.470528	-0.147629	-0.080415	0.488415
10	-0.086213	-0.04259	-0.043841	0.001778	-0.102127	-0.034999	0.252107	0.593433	-0.475297	0.04357	0.015163	-0.039187	-0.084858	0.363043	-0.173919	0.393723	0.087264
11	-0.089978	-0.158862	0.035399	0.039228	-0.145622	0.133556	-0.050249	-0.560393	0.107366	-0.051622	-0.009394	0.071659	-0.163821	0.239903	0.048975	0.690417	0.159332
12	-0.08887	-0.043795	0.061924	-0.06996	0.097028	0.087175	-0.044554	-0.067241	-0.017772	-0.035434	0.01186	-0.702656	0.662489	0.047901	-0.035888	0.126668	0.063074
13	-0.549428	-0.291572	0.417001	-0.008798	0.010778	0.570684	-0.146321	0.211561	0.100935	0.028638	-0.03382	0.06381	-0.098502	-0.061997	-0.028081	-0.128739	0.007096
14	0.005415	0.014458	-0.049791	-0.723645	0.655465	0.025306	-0.039715	-0.001593	-0.028258	-0.008063	0.001426	0.083147	-0.113374	0.003832	-0.007326	0.1451	-0.00329
15	0.599138	-0.661497	-0.233235	-0.022145	-0.032265	0.367681	-0.026249	0.081425	-0.026778	-0.010462	-0.004546	-0.012514	0.017928	-0.018306	0.00008	-0.056007	-0.014841
16	-0.18217	0.391042	-0.716685	0.056205	-0.019674	0.542775	-0.029503	-0.001034	-0.009857	-0.004361	0.010873	-0.013315	-0.007381	-0.008858	0.024053	-0.010566	0.00251

Heatmap for the above PC



2.8) Mention the business implication of using the Principal Component Analysis for this case study. [Hint: Write Interpretations of the Principal Components Obtained]

Principal component analysis is a statistical technique which uses orthogonal transformation to convert a set of possibly correlated variables into a set of values of linearly uncorrelated variables.

In this case study, there are 17 features(apart from Names of the colleges). By looking at the column names, we are not able to identify the un-important columns that can be ignored. And also, the columns are correlated to each other which makes it difficult to select features that are entirely independent. In order to reduce the dimensionality and represent the features without any loss of information PCA is being used.

PCA will result in another set of new dimensions, the one which captures maximum variance in the underlying data is the Principal Component 1. The next principal component will be orthogonal to it and so on. In this way we can reduce the total error in the representation as well.

Upon implementing PCA to this case study, out of a total of 17 features, we are able to come up with significant 6 Principal components, which is a linear combination of all the features in the original dataset, that are independent of each other. The cumulative explained variance of the 6 PC is 84.15%.

From the above heatmap, we can infer some of the characteristics that is being explained by each component

Principal component	**Individual Variance Explained(%)**	**Cumulative variance explained(%)**	**Dominant Features(having a PC score of greater than +3 and lesser than -3)**
PC1	33.3	33.3	Top10perc, Top25perc, PhD, Terminal and Expend.
PC2	28.8	62.1	Apps, Accept, Enroll, F.Undergrad and P.Undergrad. This can be related to the number of students.
PC3	6.6	68.7	Books and Personal. This feature represents the estimated cost spending of the students.
PC4	5.9	74.6	Top10perc & Top25perc(in negative direction) and Room board
PC5	5.1	79.7	This component also corresponds to PhD and Terminal(but in other direction) and also Grad.Rate.
PC6	4.5	84.1	This component corresponds to features like books, S.F.Ratio and Grad.Rate(in positive direction) and it captures the rest of the variance of personal feature.

Organizing principal component in this way reduces dimensionality without losing much information and thus by discarding the component with low information, we can come up with new combination of features to represent the original dataset.