PROJECT ADVANCE STATISTICS - NIKHIL

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. The data on hours of relief can be found in the following .csv file: Fever.csv

[Assume all of the ANOVA assumptions are satisfied]

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]

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.

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.

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]

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.

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

Sample data

	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

Shape:

(36,4)

Information of the dataframe

```
 #   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)
```

Summary of the dataframe

	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**

Checking for missing value

A	0
B	0
Volunteer	0
Relief	0

There are no missing values present

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]

For variable A

Ho :The mean hours of relief for different levels of active ingredient A are equal H0: u1A=u2A=U3A

Ha :At least at one of the levels, the mean hours of relief is different from the other. Ha:u1A!=u2A=U3A

For variable B

Ho :The mean hours of relief for different levels of active ingredient B are equal H0: u1B=u2B=U3B

Ha :At least at one of the levels, the mean hours of relief is different from the other Ha:u1B!=u2B=U3B

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.

One Way Anova:

Ho:The mean hours of relief for different levels of active ingredients A are equal .

Ha:At least at one of the levels, the mean hours of relief is different from the other

	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

Since the p value is less than the significance level, we can reject the null hupothesis and conclude that the mean number of hours of relief is different for the levels in compound A

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.

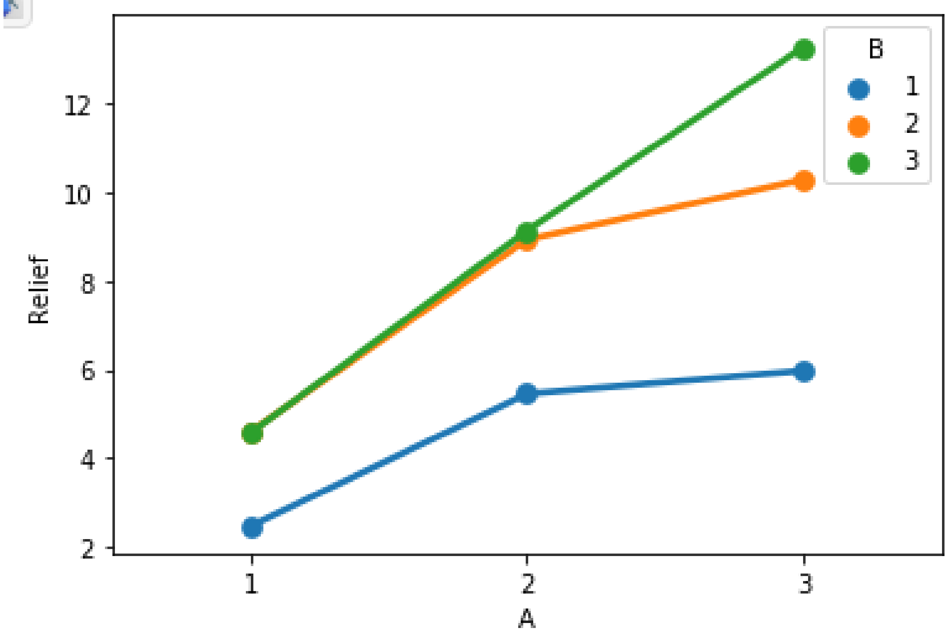
Ho:The mean hours of relief for different levels of active ingredients B are equal Ha:At least at one of the levels the mean hours of relief is different from the other

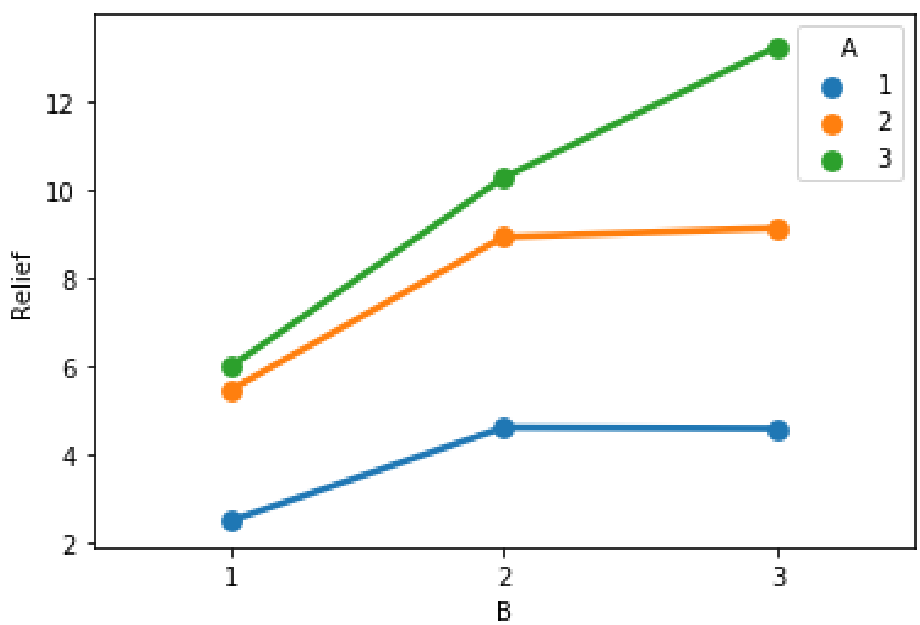
	df	sum_sq	mean_sq	F	PR(>F)
C(B)	2.0	123.66	61.830000	8.126777	0.00135
Residual	33.0	251.07	7.608182	NaN	NaN

Since the p value is less than the significance level, we can reject the null hupothesis and conclude that the mean number of hours of relief is different for the levels in compound B

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]

Drawing a Point Plot





By looking at the interaction plot between Variable A and Variable B

<u>We can say that there is an interaction between Variable A and Variable B based on response variable Relief.</u>

The responses significantly differs across the levels of both active ingredients the A and B. Relief hours are affected by interaction of A and B and relief number of hours increases significantly by level 3 of A and B .

Number of relief hours in not much impacted by level 2 and 3 of A combining with level 1 and 2 of B .

But it is not the case with level 3 of B. Relief hours change when level 3 of B interacts with A.

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.

	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

Two-way ANOVA –

At least at one of the levels, the mean hours of relief is different from the other

Ho: The mean hours of relief for different levels of active ingredients A and B are equal

H1: At least at one of the levels, the mean hours of relief of active ingredients A and B is different from the other .

By performing Two way anova p value is 6.972083e-17 which is way less than 0.05 .

Hence null hypothesis Ho is rejected that means At least at one of the levels the mean hours of relief of active ingredients A and B is different from the other. There is interaction between the levels of ingredient A and ingredient B. Considering both the factors(A and B) both are significant factor as the p value is <0.05.

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

The one-way ANOVA compares the means between the groups you are interested in and determines whether any of those means are statistically significantly different from each other.

In this case study after performing ANOVA , we can make the inference and conclude that both the active ingredients A and B have a significant interaction between the levels and both active ingredients A and B plays a significant and vital role in the cure of severe cases of hay fever.

When we performed Two way anova and checked the interaction between both variable A and B we found that interaction between the two ingredients is not significant while if we can see that A individually is the contributing factor in relief and B individually is a contributing factor in relief.

Variance between each level of ingredient A is 1827 times the variance within each level of ingredient A

Variance between each level of ingredient B is 1027 times the variance within each level of ingredient B

But when variance between of both A and B is taken together is 122 times the variation within which is very less as compared to when A and B are contributing individually and thus we can conclude that for preparing the cure for the relief of severe cases of hay fever we can consider that both the ingredients have significant impact if used individually but using both together is not that significant.

②

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.

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

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

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

2.5) Build the covariance matrix, eigenvalues, and eigenvector

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

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.

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

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

Shape:

(777,18)

Datatype of the dataframe

```
 #   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)
```

Summary of the dataframe

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

Checking for duplicates

```
0 duplicates are present
```

Checking for Missing values

```
Names          0
Apps           0
Accept         0
Enroll         0
Top10perc      0
Top25perc      0
F.Undergrad    0
P.Undergrad    0
Outstate       0
Room.Board     0
```

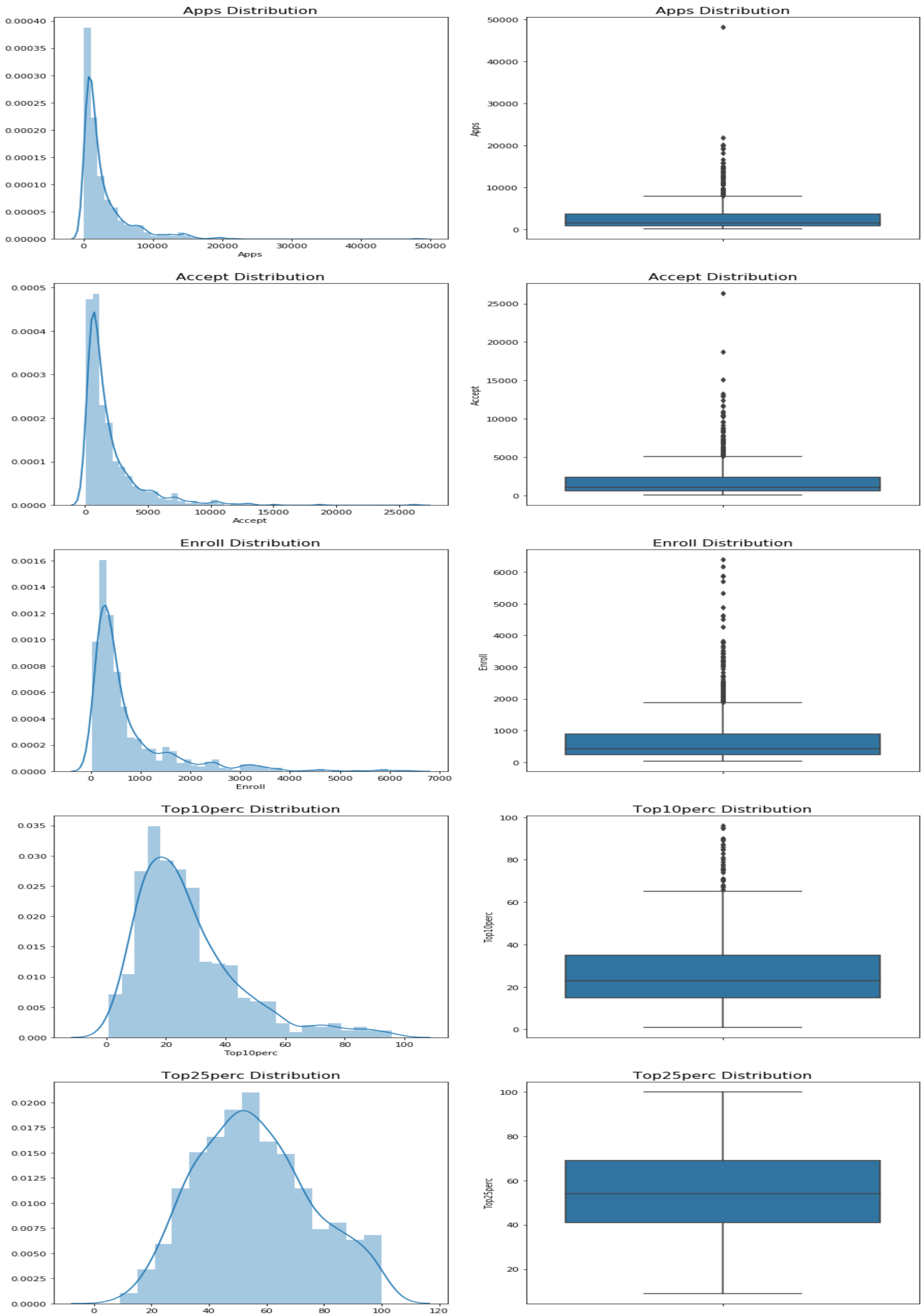
```
Books          0
Personal       0
PhD            0
Terminal       0
S.F.Ratio      0
perc.alumni    0
Expend         0
Grad.Rate      0
dtype: int64
```

No missing values are present.

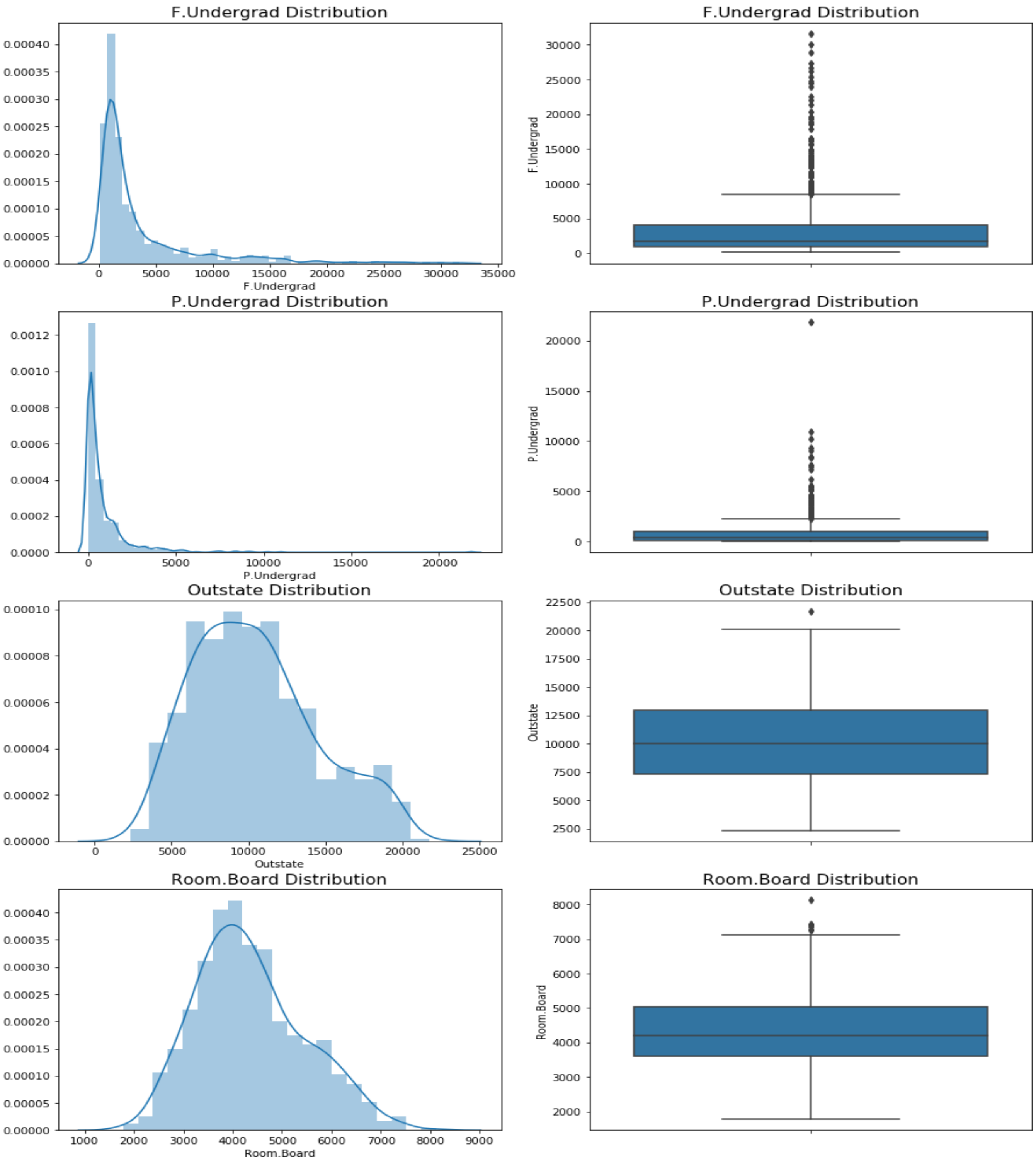
BOX PLOT AND DIST PLOT

Boxplot gives us a good indication of how the values in the data are spread out and also tells us if any outlier is present.

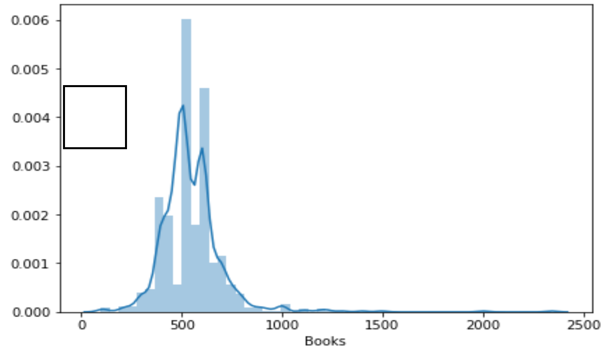
Displot shows us univariant set of observations .



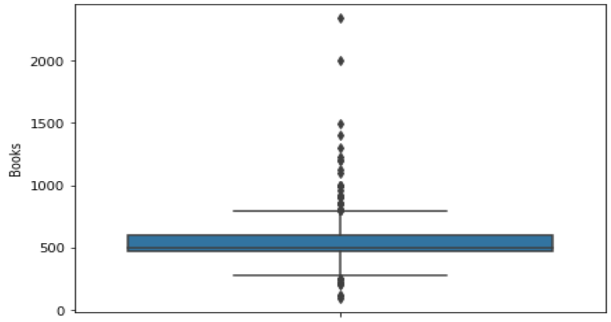
Apps Distribution
Apps
Apps Distribution
Apps
Accept Distribution
Accept
Accept Distribution
Accept
Enroll Distribution
Enroll
Enroll Distribution
Enroll
Top10perc Distribution
Top10perc
Top10perc Distribution
Top10perc
Top25perc Distribution
Top25perc
Top25perc Distribution
Top25perc


F.Undergrad Distribution
F.Undergrad
F.Undergrad Distribution
F.Undergrad
P.Undergrad Distribution
P.Undergrad
P.Undergrad Distribution
P.Undergrad
Outstate Distribution
Outstate
Outstate Distribution
Outstate
Room.Board Distribution
Room.Board
Room.Board Distribution
Room.Board

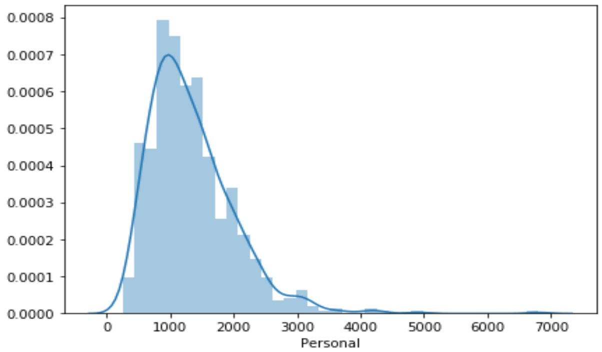
Books Distribution



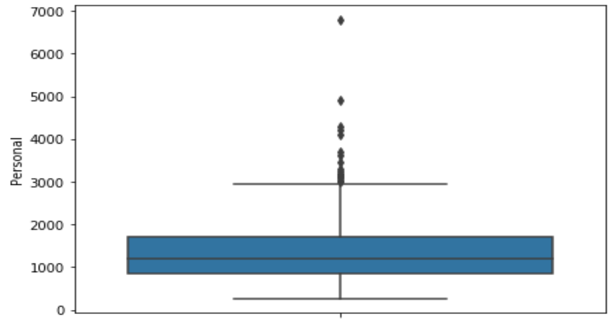
Books Distribution



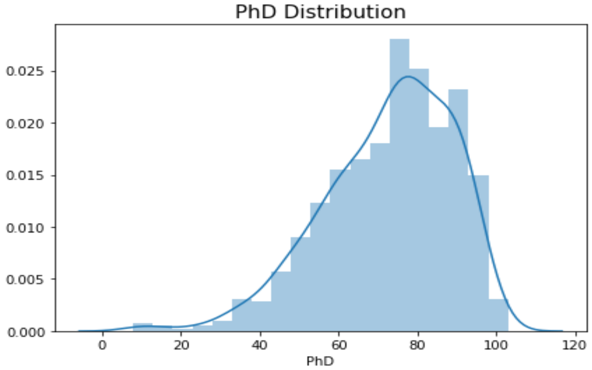
Personal Distribution



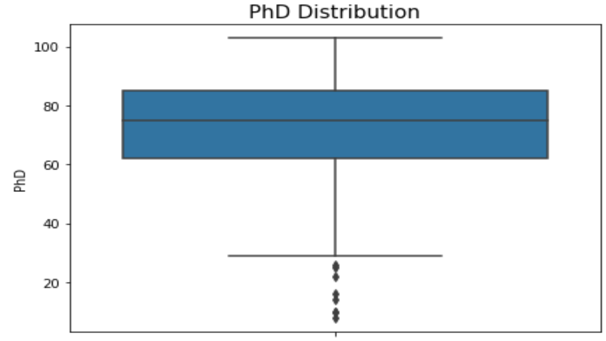
Personal Distribution



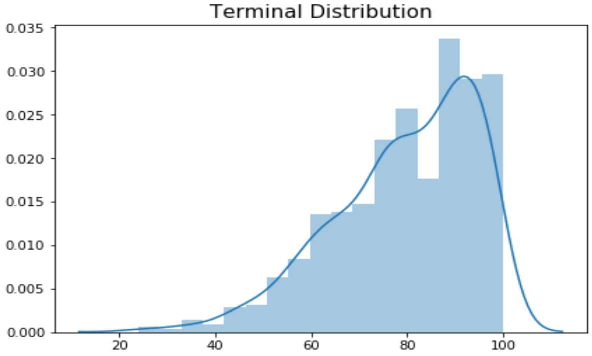
PhD Distribution



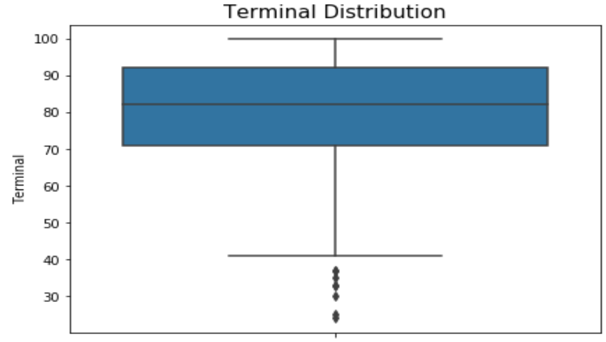
PhD Distribution



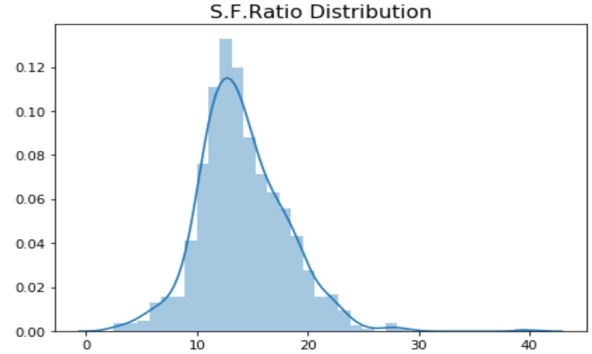
Terminal Distribution



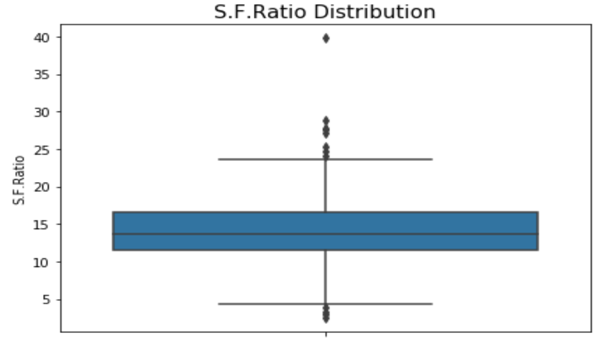
Terminal Distribution

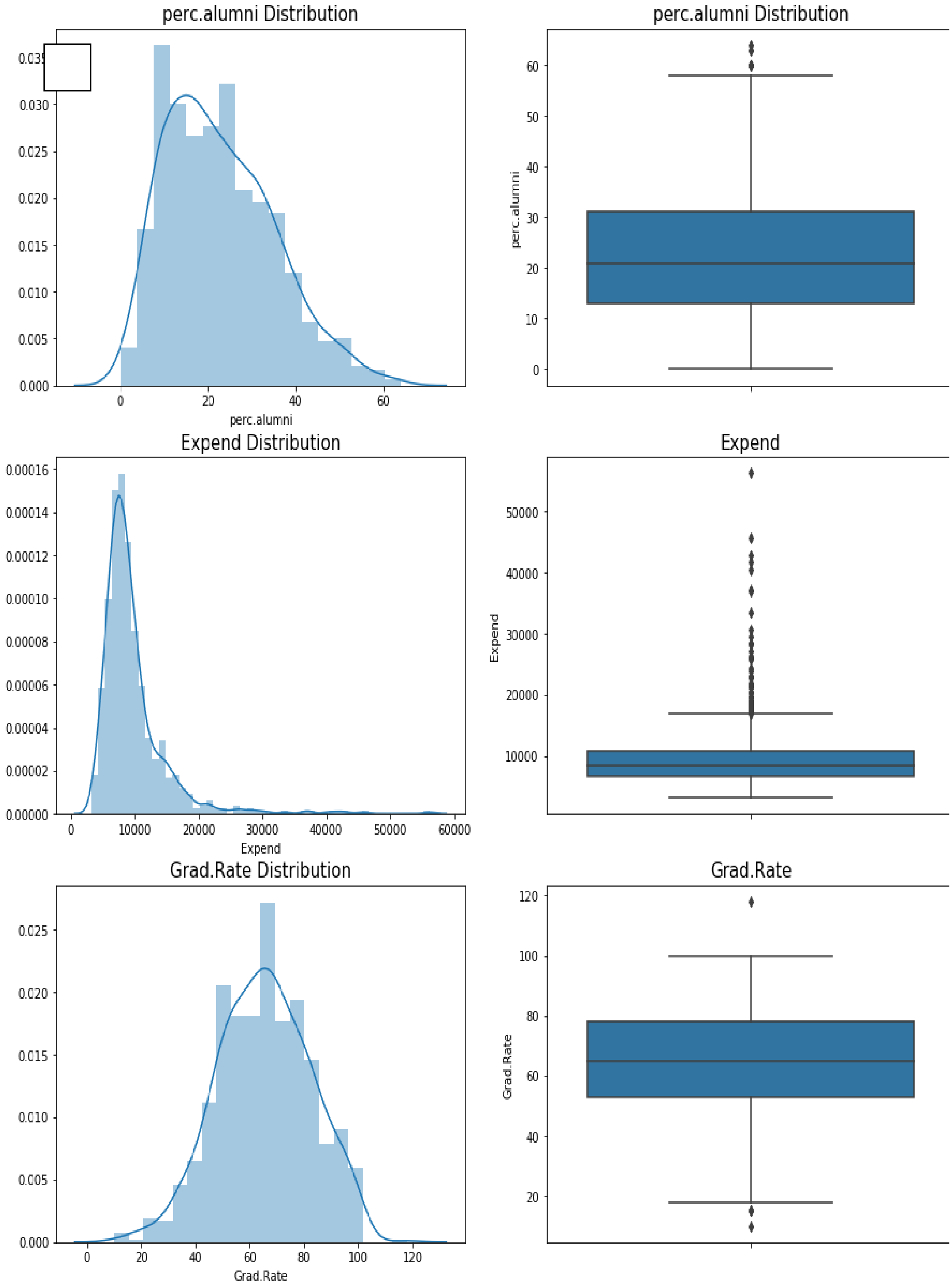


S.F.Ratio Distribution

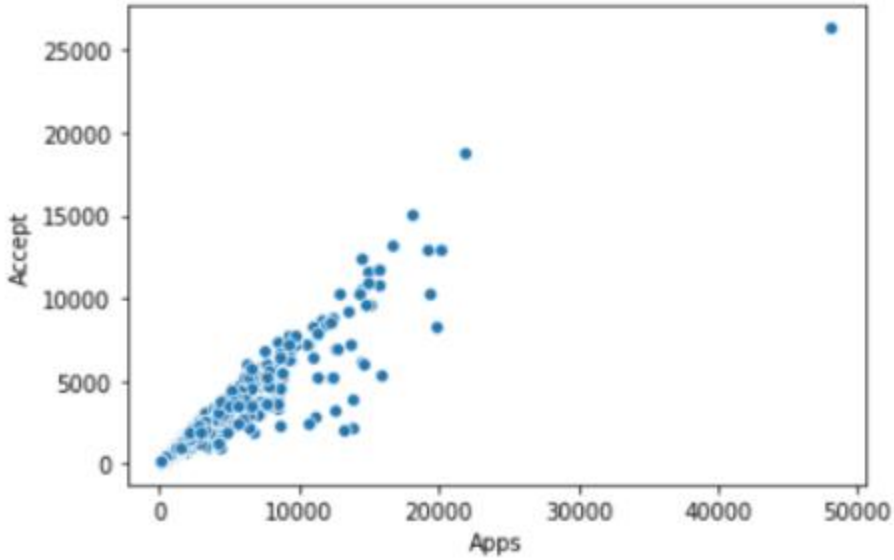


S.F.Ratio Distribution




perc.alumni Distribution
perc.alumni
perc.alumni Distribution
perc.alumni
Expend Distribution
Expend
Expend
Expend
Grad.Rate Distribution
Grad.Rate
Grad.Rate
Grad.Rate

SCATTER PLOT



SKEWNESS

```
Apps            3.723750
Accept          3.417727
Enroll          2.690465
Top10perc       1.413217
Top25perc       0.259340
F.Undergrad     2.610458
P.Undergrad     5.692353
Outstate        0.509278
Room.Board      0.477356
Books           3.485025
Personal        1.742497
PhD            -0.768170
Terminal       -0.816542
S.F.Ratio       0.667435
perc.alumni     0.606891
Expend          3.459322
Grad.Rate      -0.113777
```

Univariate analysis refer to the analysis of a single variable.

The purpose of univariate analysis is to summarize and find patterns in the data.

No null value found

No missing value found

Mean of Apps is 3001.6 ,Median is 1558 and S.D is 3870.2 whereas Max value is 48094.0

Mean of Accept is 2018.8, Median is 1110 and S.D is 2451.1 whereas Max value is 26330.0

By seeing at the distplot for Apps we can see that most of the colleges have received applications in the range 0-5000

By seeing at the distplot for Accept we can see that most of the colleges have received applications in the range 0-2500

Highest mean is of outstate 10440.66 and lowest mean is of S.F ratio 14.08

Highest median is of outstate 9990 and lowest is of S.F ratio 13.6

Max value is of expand 56233

If Mode< Median< Mean then the distribution is positively skewed.

If Mode> Median> Mean then the distribution is negatively skewed.

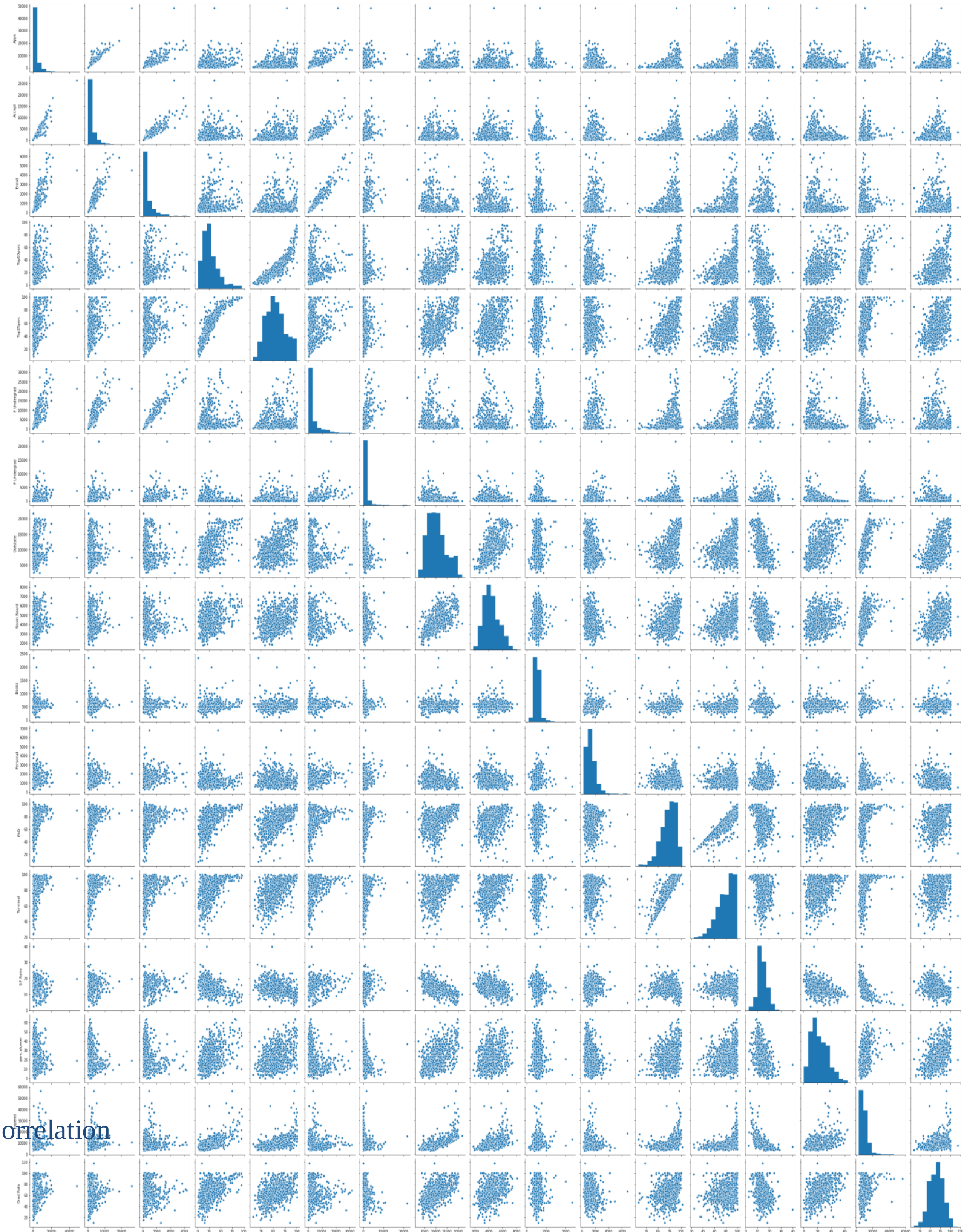
Positively skewed: Most frequent values are low and tail is towards high values. Negatively skewed: Most frequent values are high and tail is towards low values.

Presence of skew in data with long thick tail can impact the effective of PCA which hurts the notion of variance and therefore hurts the notion of variance and therefore hurts the notion of PCA.

We observe that all the variables have an outliers except Top25perc

Multivariate Analysis

Pair plot

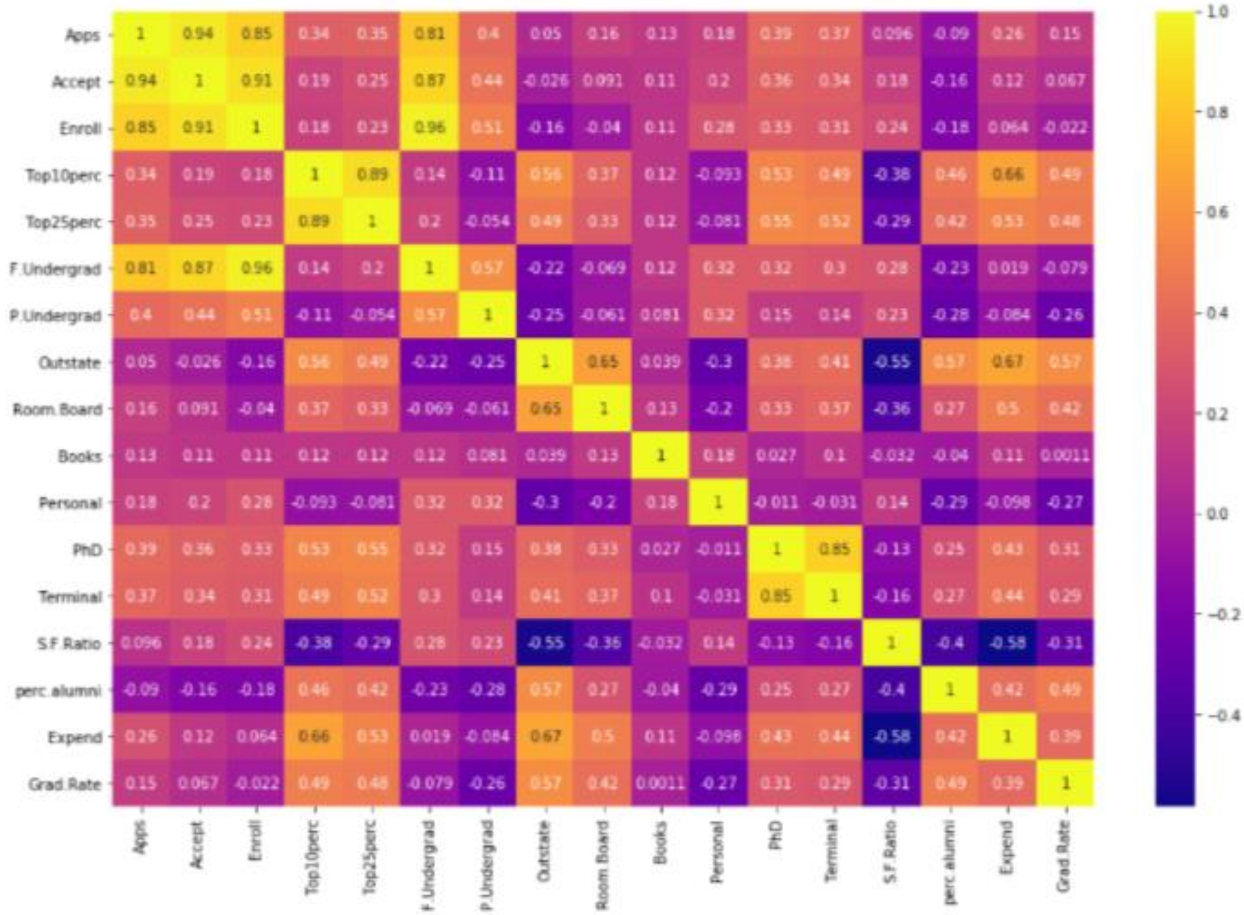


Correlation

	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

PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0.390697	0.369491	0.095633	-0.090226	0.259592	0.146755
0.355758	0.337583	0.176229	-0.159990	0.124717	0.067313
0.331469	0.308274	0.237271	-0.180794	0.064169	-0.022341
0.531828	0.491135	-0.384875	0.455485	0.660913	0.494989
0.545862	0.524749	-0.294629	0.417864	0.527447	0.477281
0.318337	0.300019	0.279703	-0.229462	0.018652	-0.078773
0.149114	0.141904	0.232531	-0.280792	-0.083568	-0.257001
0.382982	0.407983	-0.554821	0.566262	0.672779	0.571290
0.329202	0.374540	-0.362628	0.272363	0.501739	0.424942
0.026906	0.099955	-0.031929	-0.040208	0.112409	0.001061
-0.010936	-0.030613	0.136345	-0.285968	-0.097892	-0.269344
1.000000	0.849587	-0.130530	0.249009	0.432762	0.305038
0.849587	1.000000	-0.160104	0.267130	0.438799	0.289527
-0.130530	-0.160104	1.000000	-0.402929	-0.583832	-0.306710
0.249009	0.267130	-0.402929	1.000000	0.417712	0.490898
0.432762	0.438799	-0.583832	0.417712	1.000000	0.390343
0.305038	0.289527	-0.306710	0.490898	0.390343	1.000000

Heatmap



We can see that many columns are co-related to each other

There is strong correlation exists between the applications received and application accepted.

There is strong correlation exists between the applications received and applications enrolled.

Scatter plot gave us an idea of the association between different variables. It gives visual representation of the degree of correlation between any two columns.

<u>To obtain more precise information we have used the .corr() and heatmap to see that many columns are co-related to each other</u>

So we can perform PCA.

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

First we can drop the variable ‘Names’ than we can use standard scaling or z score for scaling the data

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

StandardScaler removes the mean and scales the data to unit variance.

Standard Scaler is the number of standard deviations by which the value of a raw score is above or below the mean value of what is being observed or measured.

Raw scores above the mean have positive standard scores, while those below the mean have negative standard scores.

StandardScaler removes the mean and scales the data to unit variance.

Often the variables of the data set are of different scales In this data set all variables are integers except SF Ratio is a float It gets difficult to compare the data when it is in different scales.

StandardScaler score the method used to standardize the range of features of data. Since, the range of values of data may vary widely,so we have choose Standard Scaler In this method, we convert variables with different scales of measurements into a single scale. StandardScaler normalizes the data using the

Formula (x-mean)/standard deviation.

We do this only for the numerical variables.

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

“Covariance” indicates the direction of the linear relationship between variables.

“Correlation” on the other hand measures both the strength and direction of the linear relationship between two variables.

Correlation is a function of the covariance.
You can obtain the correlation coefficient of two variables by dividing the covariance of these variables by the product of the standard deviations of the same values.

We can say that after scaling - the covariance and the correlation have the same values correlation value itself is scaled covariance value.

The covariance or correlation matrix captures relationships between different variables in their original dimensions. When the sign is positive,the variables are said to be positively correlated When the sign is negative, the variables are said to be negatively correlated When the sign is 0, the variables are said to be uncorrelated.

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

1.Eigen decomposition of the covariance matrix after standardizing the data.

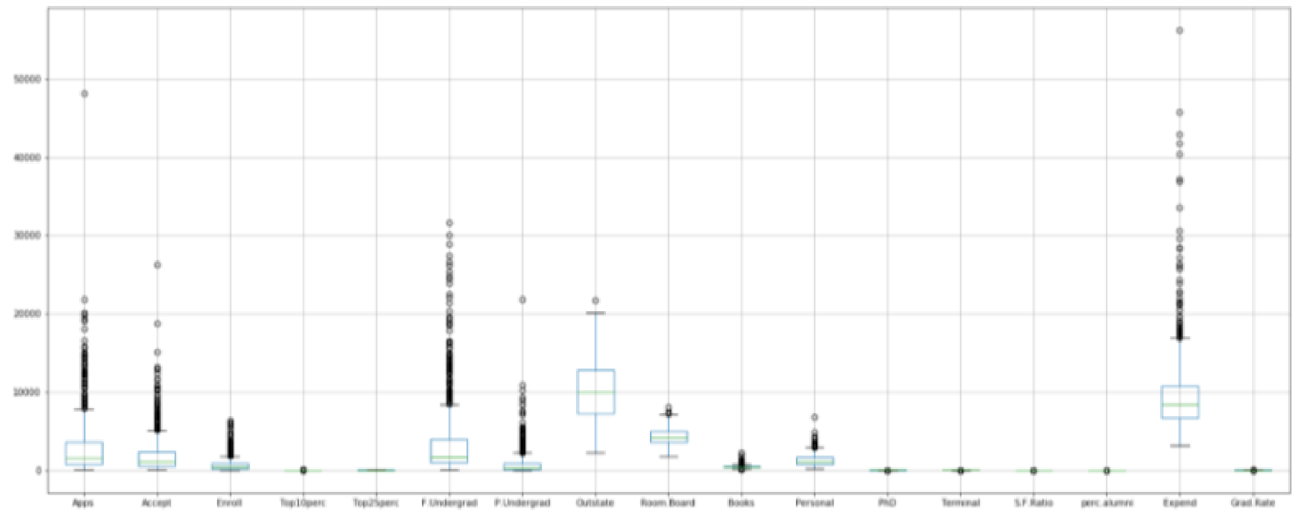
2.Eigen decomposition of the correlation matrix.

3.Eigen decomposition of the correlation matrix after standardizing the data.

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

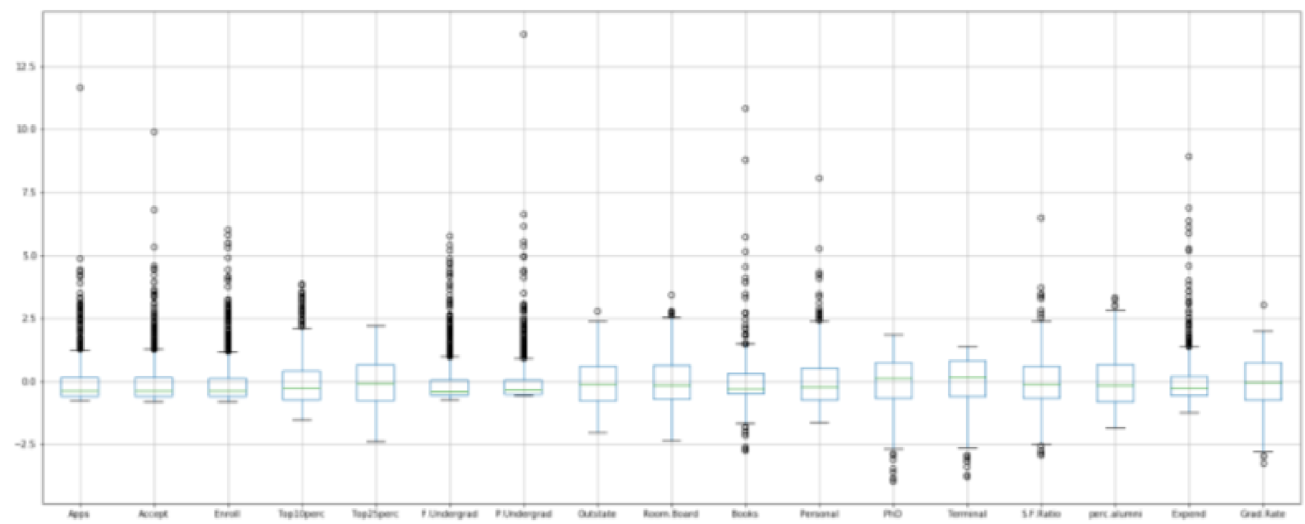
2.4) Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.

BEFORE SCALING



All have outliers except Top25perc

AFTER SCALING



In both cases before scaling and after scaling

We observed that almost all variables have outliers except Top25perc

So scaling doesn't remove outliers it just helps in standardising the values.

Before scaling the boxplot shows the amount of total variance explained in the different principal components where we have not normalized the data.

After scaling we have normalized the data. Here it is clear that PCA seeks to maximize the variance of each component.

<u>There is no significant change in the outlier pattern.</u>

Covariance matrix must be built on the scaled data and the same should be given as an input to calculate the Eigen values and vectors.

Outliers are valid and need to be retained inmodel in this case and should not be treated

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

COVARIANCE MATRIX

```
array([[ 1.001e+00,  9.450e-01,  8.480e-01,  3.390e-01,  3.520e-01,
         8.160e-01,  3.990e-01,  5.000e-02,  1.650e-01,  1.330e-01,
         1.790e-01,  3.910e-01,  3.700e-01,  9.600e-02, -9.000e-02,
         2.600e-01,  1.470e-01],
       [ 9.450e-01,  1.001e+00,  9.130e-01,  1.930e-01,  2.480e-01,
         8.750e-01,  4.420e-01, -2.600e-02,  9.100e-02,  1.140e-01,
         2.010e-01,  3.560e-01,  3.380e-01,  1.760e-01, -1.600e-01,
         1.250e-01,  6.700e-02],
       [ 8.480e-01,  9.130e-01,  1.001e+00,  1.820e-01,  2.270e-01,
         9.660e-01,  5.140e-01, -1.560e-01, -4.000e-02,  1.130e-01,
         2.810e-01,  3.320e-01,  3.090e-01,  2.380e-01, -1.810e-01,
         6.400e-02, -2.200e-02],
       [ 3.390e-01,  1.930e-01,  1.820e-01,  1.001e+00,  8.930e-01,
         1.410e-01, -1.050e-01,  5.630e-01,  3.720e-01,  1.190e-01,
        -9.300e-02,  5.330e-01,  4.920e-01, -3.850e-01,  4.560e-01,
         6.620e-01,  4.960e-01],
       [ 3.520e-01,  2.480e-01,  2.270e-01,  8.930e-01,  1.001e+00,
         2.000e-01, -5.400e-02,  4.900e-01,  3.320e-01,  1.160e-01,
        -8.100e-02,  5.470e-01,  5.250e-01, -2.950e-01,  4.180e-01,
         5.280e-01,  4.780e-01],
       [ 8.160e-01,  8.750e-01,  9.660e-01,  1.410e-01,  2.000e-01,
         1.001e+00,  5.710e-01, -2.160e-01, -6.900e-02,  1.160e-01,
         3.180e-01,  3.190e-01,  3.000e-01,  2.800e-01, -2.300e-01,
         1.900e-02, -7.900e-02],
       [ 3.990e-01,  4.420e-01,  5.140e-01, -1.050e-01, -5.400e-02,
         5.710e-01,  1.001e+00, -2.540e-01, -6.100e-02,  8.100e-02,
         3.200e-01,  1.490e-01,  1.420e-01,  2.330e-01, -2.810e-01,
        -8.400e-02, -2.570e-01],
       [ 5.000e-02, -2.600e-02, -1.560e-01,  5.630e-01,  4.900e-01,
        -2.160e-01, -2.540e-01,  1.001e+00,  6.550e-01,  3.900e-02,
        -2.990e-01,  3.830e-01,  4.090e-01, -5.560e-01,  5.670e-01,
         6.740e-01,  5.720e-01],
       [ 1.650e-01,  9.100e-02, -4.000e-02,  3.720e-01,  3.320e-01,
        -6.900e-02, -6.100e-02,  6.550e-01,  1.001e+00,  1.280e-01,
        -2.000e-01,  3.300e-01,  3.750e-01, -3.630e-01,  2.730e-01,
         5.020e-01,  4.250e-01],
       [ 1.330e-01,  1.140e-01,  1.130e-01,  1.190e-01,  1.160e-01,
         1.160e-01,  8.100e-02,  3.900e-02,  1.280e-01,  1.001e+00,
         1.800e-01,  2.700e-02,  1.000e-01, -3.200e-02, -4.000e-02,
         1.130e-01,  1.000e-03],
       [ 1.790e-01,  2.010e-01,  2.810e-01, -9.300e-02, -8.100e-02,
         3.180e-01,  3.200e-01, -2.990e-01, -2.000e-01,  1.800e-01,
         1.001e+00, -1.100e-02, -3.100e-02,  1.370e-01, -2.860e-01,
        -9.800e-02, -2.700e-01],
       [ 3.910e-01,  3.560e-01,  3.320e-01,  5.330e-01,  5.470e-01,
         3.190e-01,  1.490e-01,  3.830e-01,  3.300e-01,  2.700e-02,
        -1.100e-02,  1.001e+00,  8.510e-01, -1.310e-01,  2.490e-01,
         4.330e-01,  3.050e-01],
       [ 3.700e-01,  3.380e-01,  3.090e-01,  4.920e-01,  5.250e-01,
         3.000e-01,  1.420e-01,  4.090e-01,  3.750e-01,  1.000e-01,
```

```
 -3.100e-02,  8.510e-01,  1.001e+00, -1.600e-01,  2.670e-01,
  4.390e-01,  2.900e-01],
[ 9.600e-02,  1.760e-01,  2.380e-01, -3.850e-01, -2.950e-01,
  2.800e-01,  2.330e-01, -5.560e-01, -3.630e-01, -3.200e-02,
  1.370e-01, -1.310e-01, -1.600e-01,  1.001e+00, -4.030e-01,
 -5.850e-01, -3.070e-01],
[-9.000e-02, -1.600e-01, -1.810e-01,  4.560e-01,  4.180e-01,
 -2.300e-01, -2.810e-01,  5.670e-01,  2.730e-01, -4.000e-02,
 -2.860e-01,  2.490e-01,  2.670e-01, -4.030e-01,  1.001e+00,
  4.180e-01,  4.920e-01],
[ 2.600e-01,  1.250e-01,  6.400e-02,  6.620e-01,  5.280e-01,
  1.900e-02, -8.400e-02,  6.740e-01,  5.020e-01,  1.130e-01,
 -9.800e-02,  4.330e-01,  4.390e-01, -5.850e-01,  4.180e-01,
  1.001e+00,  3.910e-01],
[ 1.470e-01,  6.700e-02, -2.200e-02,  4.960e-01,  4.780e-01,
 -7.900e-02, -2.570e-01,  5.720e-01,  4.250e-01,  1.000e-03,
 -2.700e-01,  3.050e-01,  2.900e-01, -3.070e-01,  4.920e-01,
  3.910e-01,  1.001e+00]])
```

Eigen Vectors

```
% [[-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]]
```

Eigen Values

```
% [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]
```

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

Explicit form of the FIRST PRINCIPAL COMPONENT

PC1 = (2.48765602e-01)*Apps + (2.07601502e-01)*Accept + (1.76303592e-01)*Enroll + (3.54273947e-01)*Top10perc + (3.44001279e-01)*Top25perc + (1.54640962e-01)*F.Undergrad + (2.64425045e-02)*P.Undergrad + (2.94736419e-01)*Outstate + (2.49030449e-01)*Room.Board + (6.47575181e-02)*Books + (-4.25285386e-02)*Personal + (3.18312875e-01)*PhD +(3.17056016e-01)*Terminal + (-1.76957895e-01)*S.F.Ratio + (2.05082369e-01)*perc.alumni + (3.18908750e-01)*Expend +(2.52315654e-01)*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 eigen

vectors indicate? Perform PCA and export the data of the Principal Component scores into a data frame.

TOTAL(sum of eigen values)=17.021907216494846

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.        ]
```

First Principal component explains 32.02 percent variance

Second Principal component explains 58.36 percent variance

Third Principal component explains 65.26 percent variance

Forth Principal component explains 71.18 percent variance

Fifth Principal component explains 76.67 percent variance

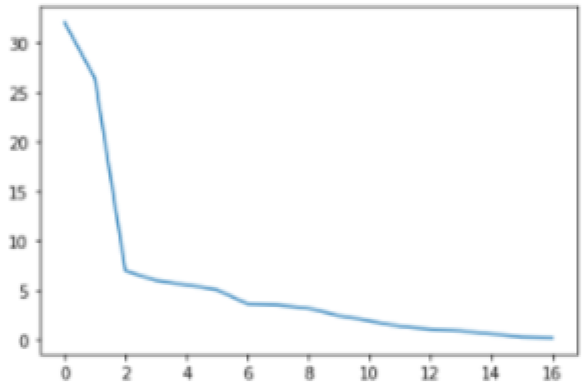
Sixth Principal component explains 81.65 percent variance

Seventh Principal component explains 85.21 percent variance

Eight Principal component explains 88.67 percent variance

Nineth Principal component explains 91.78 percent variance

Scree plot



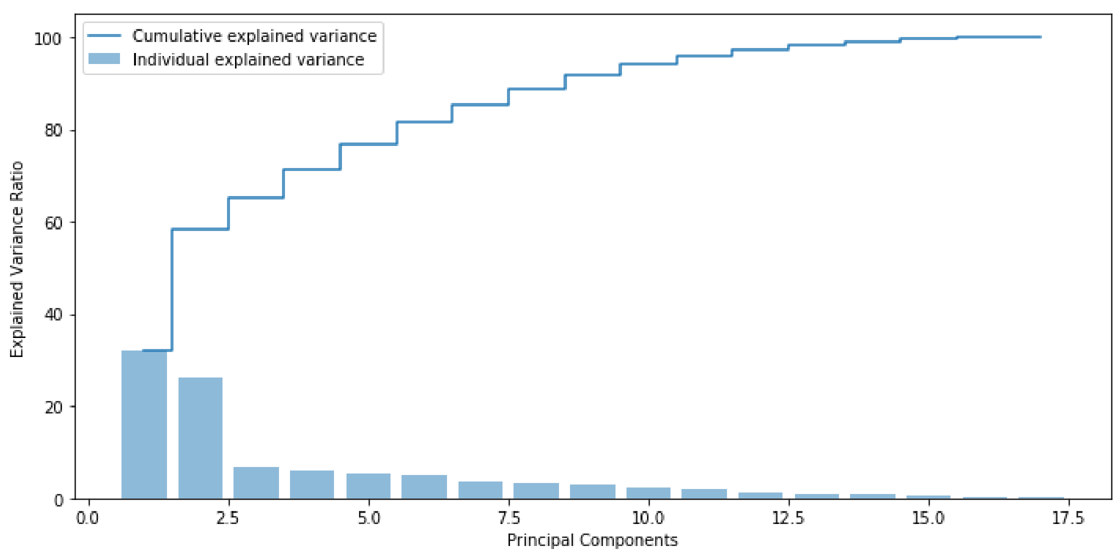
Screeplot is also called Elbow curve

Based on this graph we will take a decision on how many PC can be taken

where ever there is a sudden change in the graph that is taken as the Principal

Component.

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



The Cumulative % gives the percentage of variance accounted for by the n components.

For example, the cumulative percentage for the second component is the sum of the percentage of variance for the first and second components.

It helps in deciding the number of components by selecting the components which explained the high variance .

For about 60percent variance we need 2PC

For about 80percent variance we need 5PC

For about 85percent variance we need 7PC

For about 90percent variance we need 9PC

We will proceed with 7 components here which explains ~ 85% of the variance within the dataset.

Eigen vector indicate

Eigen vector points in the direction where the maximum variance is explained.

Eigen vector determine the directions of the new feature space, and the eigenvalues determine their magnitude.

Eigenvectors are the coefficients of new feature components which is obtained by multiplying the eigen vector values by the features.

Eigenvectors can help us in calculating an approximation of a large matrix as a smaller vector.

Eigenvectors are used to make linear transformation understandable.

Eigenvectors and eigenvalues are used to reduce noise in data. They can help us improve efficiency in intensive tasks.

pca.components_

```
array([[ 2.48765602e-01,  2.07601502e-01,  1.76303592e-01,
         3.54273947e-01,  3.44001279e-01,  1.54640962e-01,
         2.64425045e-02,  2.94736419e-01,  2.49030449e-01,
         6.47575181e-02, -4.25285386e-02,  3.18312875e-01,
         3.17056016e-01, -1.76957895e-01,  2.05082369e-01,
         3.18908750e-01,  2.52315654e-01],
       [ 3.31598227e-01,  3.72116750e-01,  4.03724252e-01,
        -8.24118211e-02, -4.47786551e-02,  4.17673774e-01,
         3.15087830e-01, -2.49643522e-01, -1.37808883e-01,
         5.63418434e-02,  2.19929218e-01,  5.83113174e-02,
         4.64294477e-02,  2.46665277e-01, -2.46595274e-01,
        -1.31689865e-01, -1.69240532e-01],
```

```
[-6.30921033e-02, -1.01249056e-01, -8.29855709e-02,
  3.50555339e-02, -2.41479376e-02, -6.13929764e-02,
  1.39681716e-01,  4.65988731e-02,  1.48967389e-01,
  6.77411649e-01,  4.99721120e-01, -1.27028371e-01,
 -6.60375454e-02, -2.89848401e-01, -1.46989274e-01,
  2.26743985e-01, -2.08064649e-01],
[ 2.81310530e-01,  2.67817346e-01,  1.61826771e-01,
 -5.15472524e-02, -1.09766541e-01,  1.00412335e-01,
 -1.58558487e-01,  1.31291364e-01,  1.84995991e-01,
  8.70892205e-02, -2.30710568e-01, -5.34724832e-01,
 -5.19443019e-01, -1.61189487e-01,  1.73142230e-02,
  7.92734946e-02,  2.69129066e-01],
[ 5.74140964e-03,  5.57860920e-02, -5.56936353e-02,
 -3.95434345e-01, -4.26533594e-01, -4.34543659e-02,
  3.02385408e-01,  2.22532003e-01,  5.60919470e-01,
 -1.27288825e-01, -2.22311021e-01,  1.40166326e-01,
  2.04719730e-01, -7.93882496e-02, -2.16297411e-01,
  7.59581203e-02, -1.09267913e-01],
[-1.62374420e-02,  7.53468452e-03, -4.25579803e-02,
 -5.26927980e-02,  3.30915896e-02, -4.34542349e-02,
 -1.91198583e-01, -3.00003910e-02,  1.62755446e-01,
  6.41054950e-01, -3.31398003e-01,  9.12555212e-02,
  1.54927646e-01,  4.87045875e-01, -4.73400144e-02,
 -2.98118619e-01,  2.16163313e-01],
[-4.24863486e-02, -1.29497196e-02, -2.76928937e-02,
 -1.61332069e-01, -1.18485556e-01, -2.50763629e-02,
  6.10423460e-02,  1.08528966e-01,  2.09744235e-01,
 -1.49692034e-01,  6.33790064e-01, -1.09641298e-03,
 -2.84770105e-02,  2.19259358e-01,  2.43321156e-01,
 -2.26584481e-01,  5.59943937e-01],
[-1.03090398e-01, -5.62709623e-02,  5.86623552e-02,
 -1.22678028e-01, -1.02491967e-01,  7.88896442e-02,
  5.70783816e-01,  9.84599754e-03, -2.21453442e-01,
  2.13293009e-01, -2.32660840e-01, -7.70400002e-02,
 -1.21613297e-02, -8.36048735e-02,  6.78523654e-01,
 -5.41593771e-02, -5.33553891e-03],
[-9.02270802e-02, -1.77864814e-01, -1.28560713e-01,
  3.41099863e-01,  4.03711989e-01, -5.94419181e-02,
  5.60672902e-01, -4.57332880e-03,  2.75022548e-01,
 -1.33663353e-01, -9.44688900e-02, -1.85181525e-01,
 -2.54938198e-01,  2.74544380e-01, -2.55334907e-01,
 -4.91388809e-02,  4.19043052e-02],
[ 5.25098025e-02,  4.11400844e-02,  3.44879147e-02,
  6.40257785e-02,  1.45492289e-02,  2.08471834e-02,
 -2.23105808e-01,  1.86675363e-01,  2.98324237e-01,
 -8.20292186e-02,  1.36027616e-01, -1.23452200e-01,
 -8.85784627e-02,  4.72045249e-01,  4.22999706e-01,
  1.32286331e-01, -5.90271067e-01],
[ 4.30462074e-02, -5.84055850e-02, -6.93988831e-02,
 -8.10481404e-03, -2.73128469e-01, -8.11578181e-02,
```

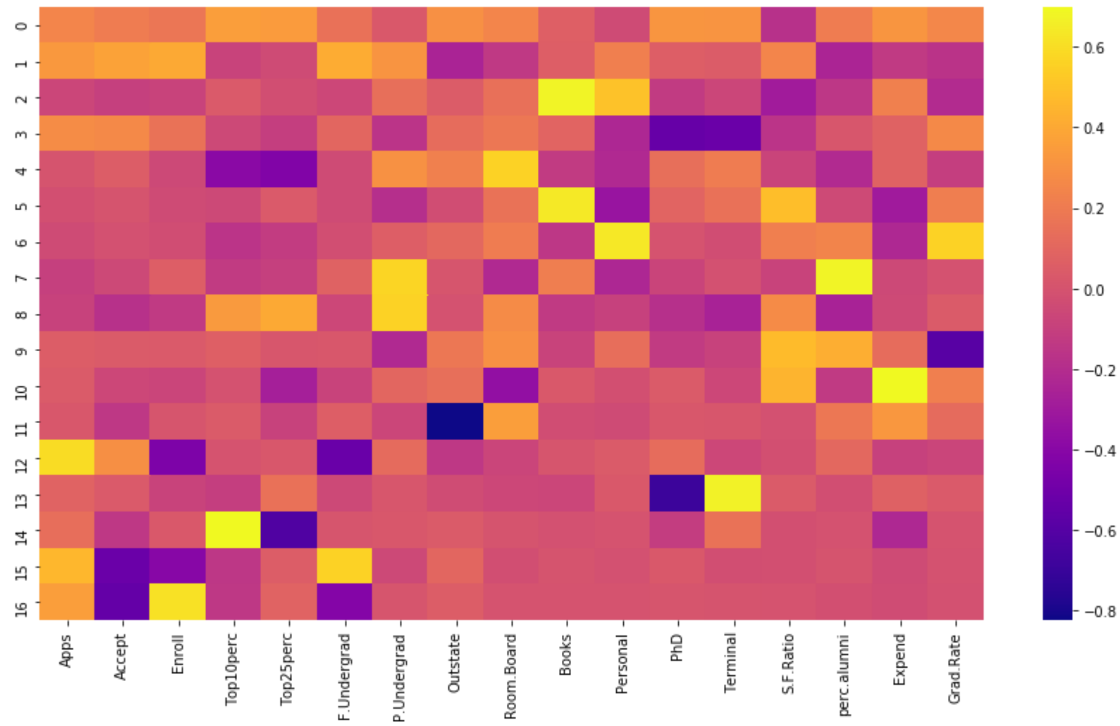
```
  1.00693324e-01,  1.43220673e-01, -3.59321731e-01,
  3.19400370e-02, -1.85784733e-02,  4.03723253e-02,
 -5.89734026e-02,  4.45000727e-01, -1.30727978e-01,
  6.92088870e-01,  2.19839000e-01],
[ 2.40709086e-02, -1.45102446e-01,  1.11431545e-02,
  3.85543001e-02, -8.93515563e-02,  5.61767721e-02,
 -6.35360730e-02, -8.23443779e-01,  3.54559731e-01,
 -2.81593679e-02, -3.92640266e-02,  2.32224316e-02,
  1.64850420e-02, -1.10262122e-02,  1.82660654e-01,
  3.25982295e-01,  1.22106697e-01],
[ 5.95830975e-01,  2.92642398e-01, -4.44638207e-01,
  1.02303616e-03,  2.18838802e-02, -5.23622267e-01,
  1.25997650e-01, -1.41856014e-01, -6.97485854e-02,
  1.14379958e-02,  3.94547417e-02,  1.27696382e-01,
 -5.83134662e-02, -1.77152700e-02,  1.04088088e-01,
 -9.37464497e-02, -6.91969778e-02],
[ 8.06328039e-02,  3.34674281e-02, -8.56967180e-02,
 -1.07828189e-01,  1.51742110e-01, -5.63728817e-02,
  1.92857500e-02, -3.40115407e-02, -5.84289756e-02,
 -6.68494643e-02,  2.75286207e-02, -6.91126145e-01,
  6.71008607e-01,  4.13740967e-02, -2.71542091e-02,
  7.31225166e-02,  3.64767385e-02],
[ 1.33405806e-01, -1.45497511e-01,  2.95896092e-02,
  6.97722522e-01, -6.17274818e-01,  9.91640992e-03,
  2.09515982e-02,  3.83544794e-02,  3.40197083e-03,
 -9.43887925e-03, -3.09001353e-03, -1.12055599e-01,
  1.58909651e-01, -2.08991284e-02, -8.41789410e-03,
 -2.27742017e-01, -3.39433604e-03],
[ 4.59139498e-01, -5.18568789e-01, -4.04318439e-01,
 -1.48738723e-01,  5.18683400e-02,  5.60363054e-01,
 -5.27313042e-02,  1.01594830e-01, -2.59293381e-02,
  2.88282896e-03, -1.28904022e-02,  2.98075465e-02,
 -2.70759809e-02, -2.12476294e-02,  3.33406243e-03,
 -4.38803230e-02, -5.00844705e-03],
[ 3.58970400e-01, -5.43427250e-01,  6.09651110e-01,
 -1.44986329e-01,  8.03478445e-02, -4.14705279e-01,
  9.01788964e-03,  5.08995918e-02,  1.14639620e-03,
  7.72631963e-04, -1.11433396e-03,  1.38133366e-02,
  6.20932749e-03, -2.22215182e-03, -1.91869743e-02,
 -3.53098218e-02, -1.30710024e-02]])
```

Principal Component scores into a data frame

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board
0	0.248766	0.207602	0.176304	0.354274	0.344001	0.154641	0.026443	0.294736	0.249030
1	0.331598	0.372117	0.403724	-0.082412	-0.044779	0.417674	0.315088	-0.249644	-0.137809
2	-0.063092	-0.101249	-0.082986	0.035056	-0.024148	-0.061393	0.139682	0.046599	0.148967
3	0.281311	0.267817	0.161827	-0.051547	-0.109767	0.100412	-0.158558	0.131291	0.184996
4	0.005741	0.055786	-0.055694	-0.395434	-0.426534	-0.043454	0.302385	0.222532	0.560919
5	-0.016237	0.007535	-0.042558	-0.052693	0.033092	-0.043454	-0.191199	-0.030000	0.162755
6	-0.042486	-0.012950	-0.027693	-0.161332	-0.118486	-0.025076	0.061042	0.108529	0.209744

Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0.064758	-0.042529	0.318313	0.317056	-0.176958	0.205082	0.318909	0.252316
0.056342	0.219929	0.058311	0.046429	0.246665	-0.246595	-0.131690	-0.169241
0.677412	0.499721	-0.127028	-0.066038	-0.289848	-0.146989	0.226744	-0.208065
0.087089	-0.230711	-0.534725	-0.519443	-0.161189	0.017314	0.079273	0.269129
-0.127289	-0.222311	0.140166	0.204720	-0.079388	-0.216297	0.075958	-0.109268
0.641055	-0.331398	0.091256	0.154928	0.487046	-0.047340	-0.298119	0.216163
-0.149692	0.633790	-0.001096	-0.028477	0.219259	0.243321	-0.226584	0.559944

HEATMAP



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

<u>This case study has 17 variables and to analyse we need reduction of variables and to do that we need Principal Component Analysis</u>

<u>After doing analysis we can consider 7 PCAs to get about 85% variability.</u>

PCA is a statistical technique and uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables.

PCA also is a tool to reduce multidimensional data to lower dimensions while retaining most of the information.

Principal Component Analysis (PCA) is a well-established mathematical technique for reducing the dimensionality of data, while keeping as much variation as possible.

PCA can only be done on continous variables. Large datasets are increasingly common and are often difficult to interpret.

Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance.

After scaling we have normalized the data. Here it is clear that PCA seeks to maximize the variance of each component.

There is no significant change in the outlier pattern.

Covariance matrix is built on the scaled data and the same has been given as an input to calculate the Eigen values and vectors.

Outliers are valid and need to be retained inmodel in this case and should not be treated