ADVANCED STATISTICS PROJECT

Question 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]

Variable A:

- Null Hypothesis: Ho_A: $mu_1 = mu_2 = mu_3$

The means of 'relief' variable with respect to each Variable 'A' category are equal.

- Alternate Hypothesis: Ha: Not all means are equal.

At least one of the means of 'relief' variable with respect to each A category is unequal.

Variable A: Categorical Values: 1,2,3

Variable B:

- Null Hypothesis: Ho_B : $mu_1 = mu_2 = mu_3$

The means of 'relief' variable with respect to each Variable 'B' category is equal

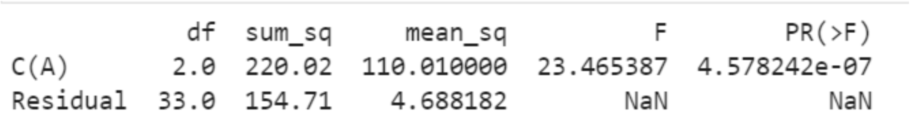
Alternate Hypothesis: Ha: Not all means are equal.

At least one of the means of 'relief' variable with respect to each B category is unequal.

Variable B: Categorical Values: 1,2,3

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 for Variable A with respect to Relief.

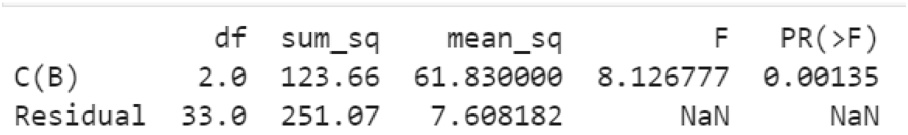


```
            df  sum_sq    mean_sq          F        PR(>F)
C(A)       2.0  220.02  110.010000  23.465387  4.578242e-07
Residual  33.0  154.71    4.688182        NaN           NaN
```

Since the p value is less than the significance level (0.05), we can reject the null hypothesis and conclude that there is a difference in the mean relief rate is different in at-least one category of Variable 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.

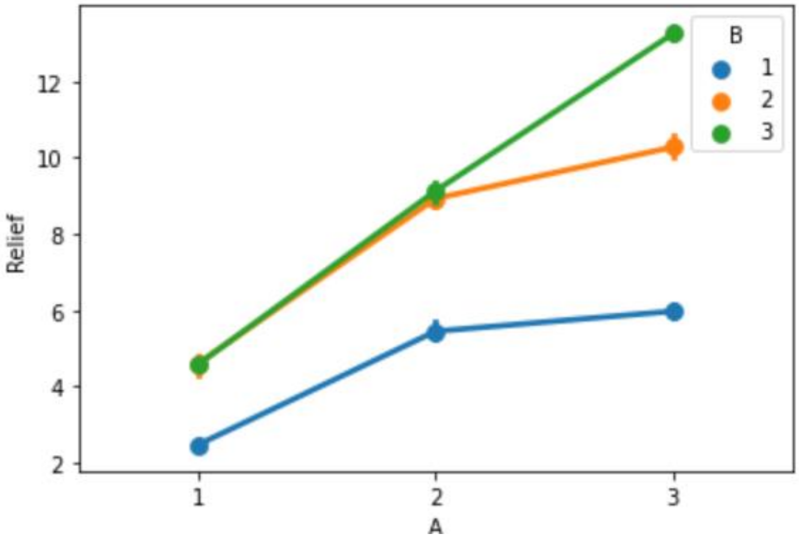
One-way ANOVA for Variable B with respect to Relief:



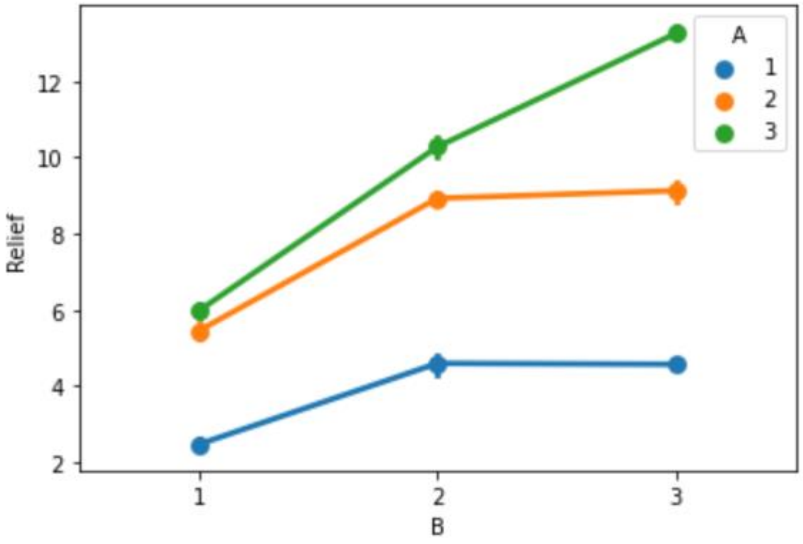
```
            df  sum_sq    mean_sq         F   PR(>F)
C(B)       2.0  123.66  61.830000  8.126777  0.00135
Residual  33.0  251.07   7.608182       NaN      NaN
```

Since the p value is less than the significance level (0.05), we can reject the null hypothesis and conclude that there is a difference in the mean relief rate is different in at-least one category of Variable 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]



```
<matplotlib.axes._subplots.AxesSubplot at 0xb6f3748>
```



Difference in lines suggests that there is a main effect of the treatment with respect to Relief i.e the relief of severe cases depends on the ingredients A and B.

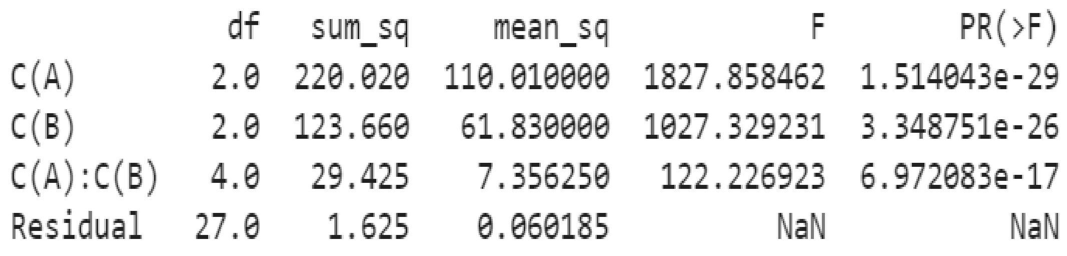
In other words, there is significant interaction between ingredients A and B because the distance between the means across the three levels are not the same.

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.

Formulating Null Hypothesis and Alternate Hypothesis:

Null Hypothesis:

H_0: The means of 'Relief' variable with respect to each 'A' category and 'B' is equal. (There between the levels of ingredient A and ingredient B)
H_1 : At least one of the means of 'Relief' variable with respect to each 'A' category and 'B' is unequal.(There is interaction)



```
               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
```

We can reject Null Hypothesis, and accept Alternate Hypothesis

Both the variables 'A' and 'B' are significant factors since p-values are less than 0.05.

As A and B interaction value is less than 0.05, There seems to be a statistical interaction.

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

To see if there is any significant interaction between the two ingredients with respect to relief variable, analysis of variance is used. So if there is any interaction found, then a test can be conducted to find out the effectiveness of the combination of both the ingredients as a unit and see how it impacts the treatment of hay fever.

For manufacturing companies, this type of test is extremely useful to see which treatment is effective and to check if the two treatments as a combination is more effective.

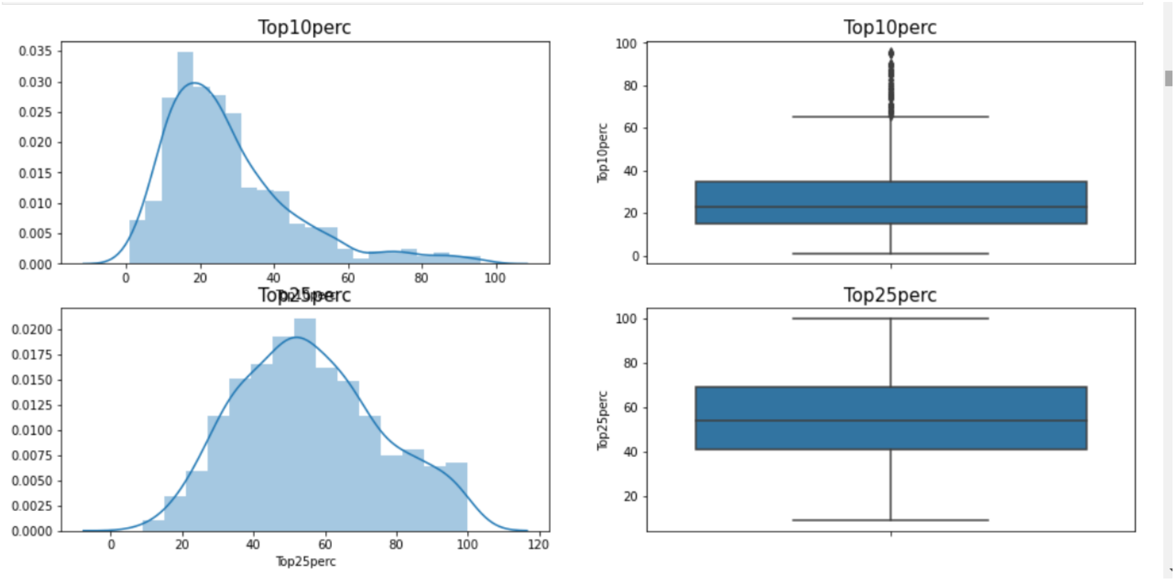
Problem 2:

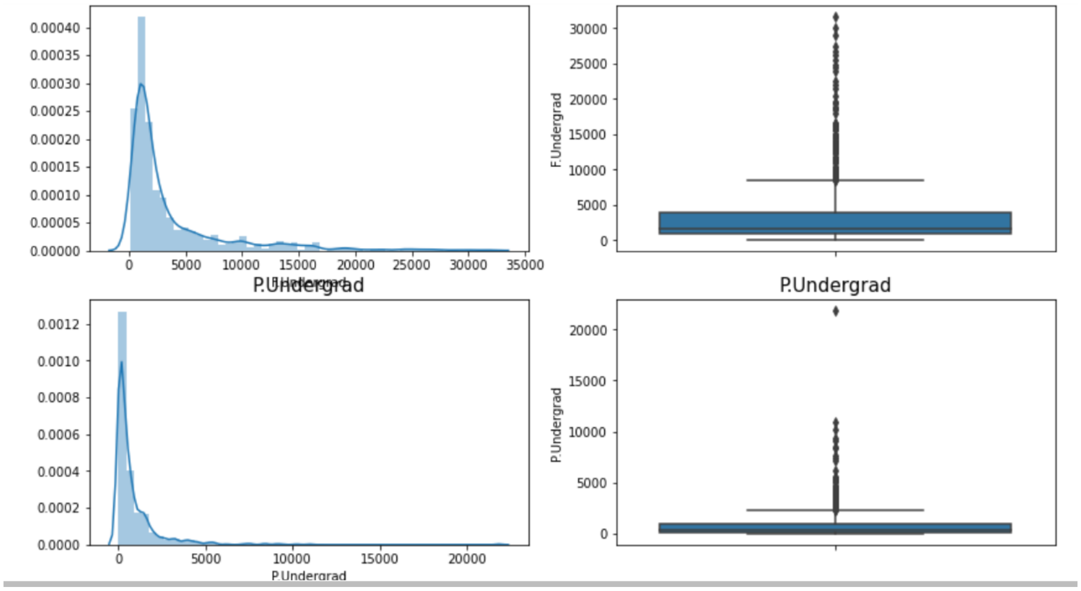
The dataset Education - Post 12th Standard.csv is a dataset that contains the names of various colleges. This particular case study is based on various parameters of various institutions. You are expected to do Principal Component Analysis for this case study according to the instructions given in the 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.

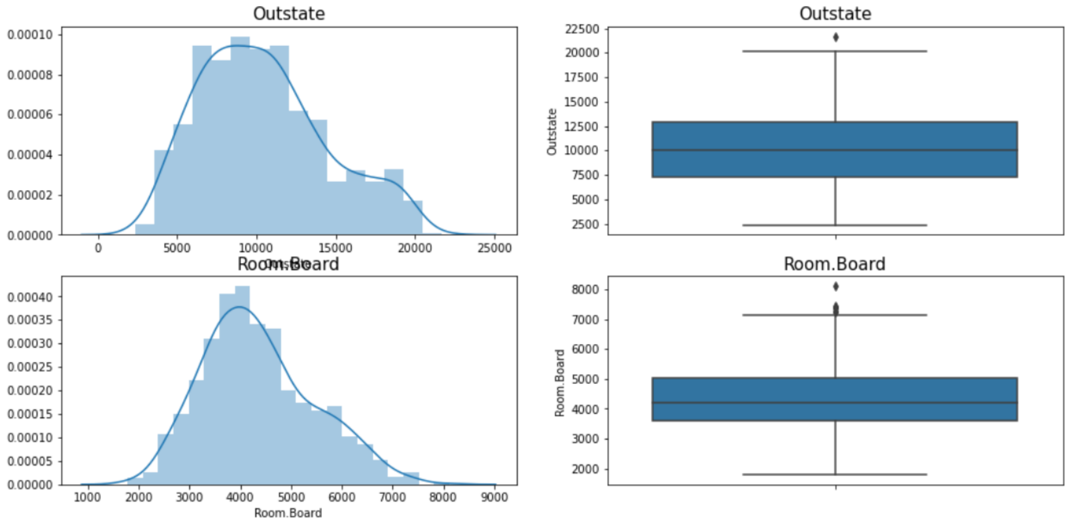
Univariate Analysis:

a. ***Dist Plot and Box Plot for all the variables respectively***

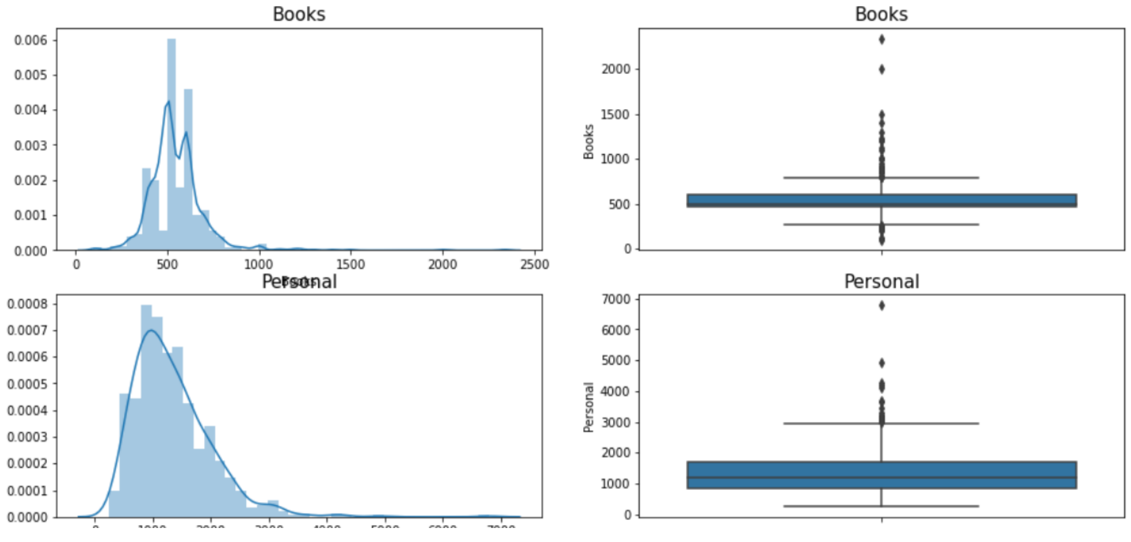




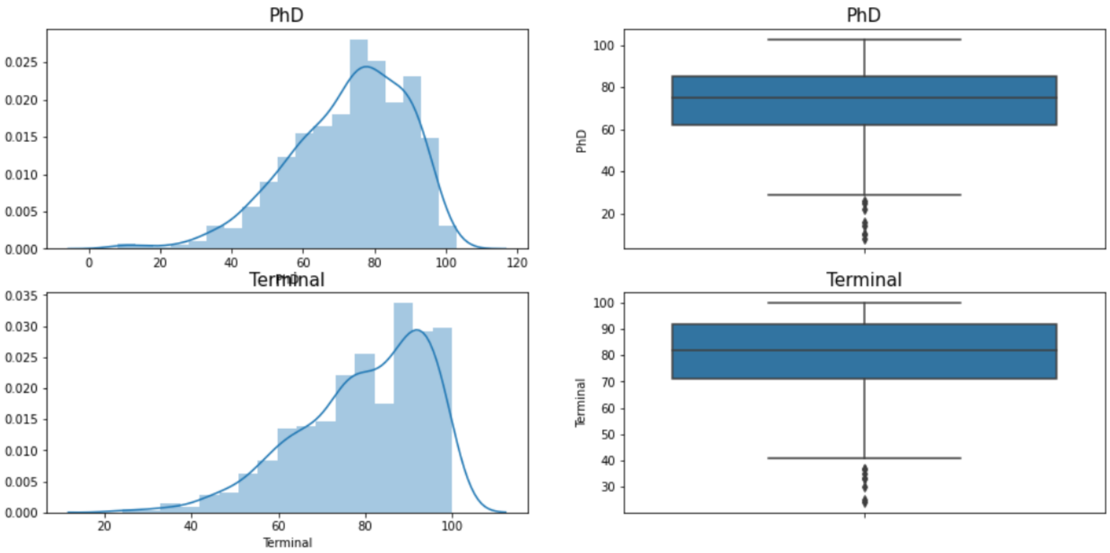
0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
5000
10000
15000
20000
25000
30000
35000
F.Undergrad
F.Undergrad
30000
25000
20000
15000
10000
5000
0
P.Undergrad
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
5000
10000
15000
20000
P.Undergrad
P.Undergrad
P.Undergrad
20000
15000
10000
5000
0

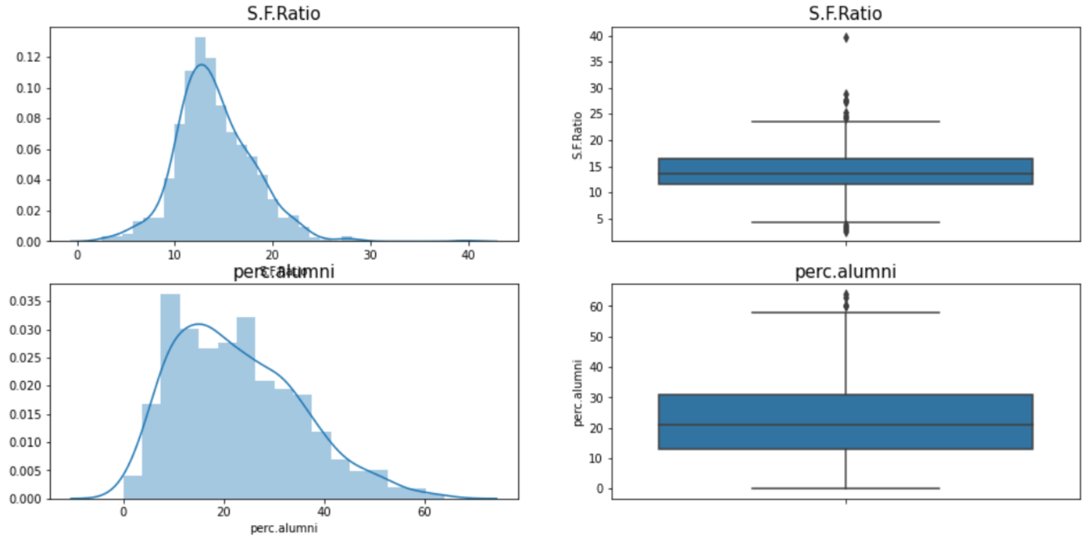


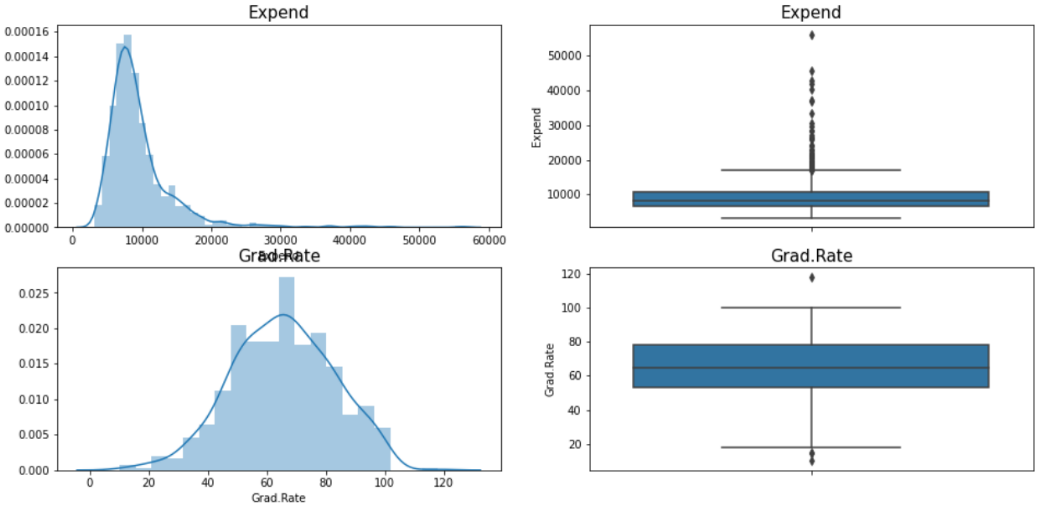
Outstate
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
5000
10000
15000
20000
25000
Outstate
Outstate
Outstate
22500
20000
17500
15000
12500
10000
7500
5000
2500
Room.Board
0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
1000
2000
3000
4000
5000
6000
7000
8000
9000
Room.Board
Room.Board
Room.Board
8000
7000
6000
5000
4000
3000
2000



Books
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
500
1000
1500
2000
2500
Books
Books
Books
2000
1500
1000
500
0
Personal
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
1000
2000
3000
4000
5000
6000
7000
Personal
Personal
7000
6000
5000
4000
3000
2000
1000
0

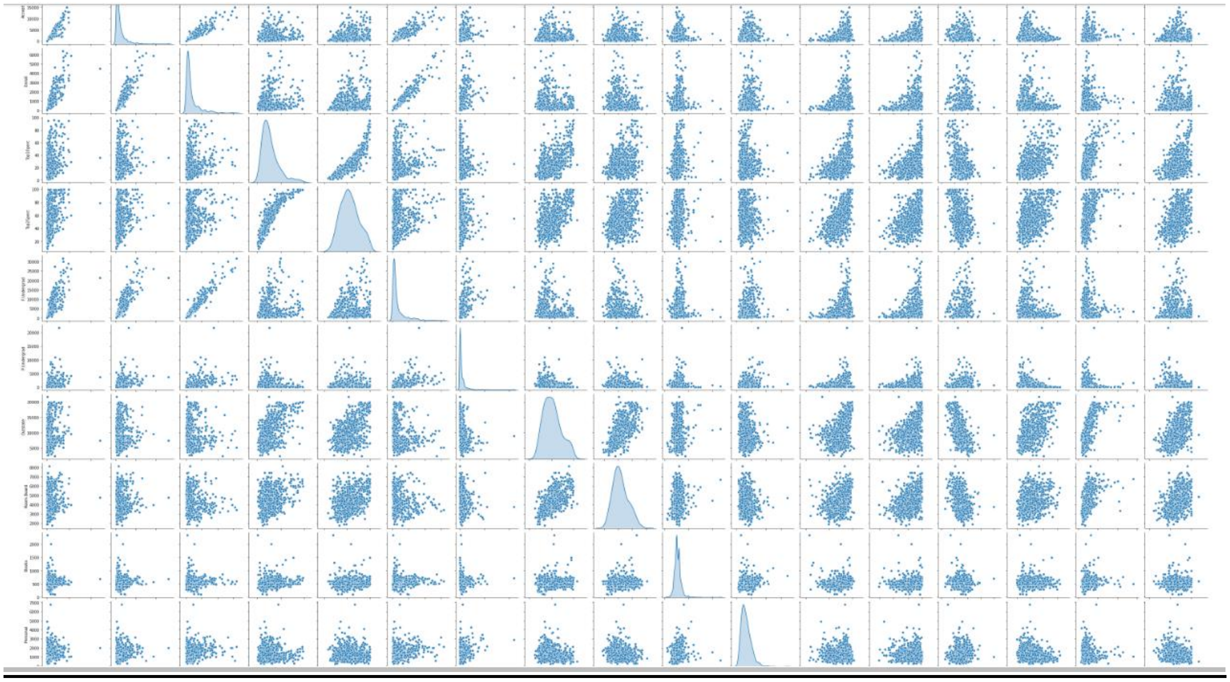

PhD
0.000
0.005
0.010
0.015
0.020
0.025
0 20 40 60 80 100 120
PhD
PhD
100
80
60
40
20
PhD
Terminal
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
20 40 60 80 100
Terminal
Terminal
100
90
80
70
60
50
40
30
Terminal


S.F.Ratio
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0 10 20 30 40
S.F.Ratio
S.F.Ratio
40
35
30
25
20
15
10
5
S.F.Ratio
perc.alumni
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0 20 40 60
perc.alumni
perc.alumni
60
50
40
30
20
10
0
perc.alumni


Expend
0.00000
0.00002
0.00004
0.00006
0.00008
0.00010
0.00012
0.00014
0.00016
0 10000 20000 30000 40000 50000 60000
Expend
Expend
50000
40000
30000
20000
10000
Expend
Grad.Rate
0.000
0.005
0.010
0.015
0.020
0.025
0 20 40 60 80 100 120
Grad.Rate
Grad.Rate
120
100
80
60
40
20
Grad.Rate

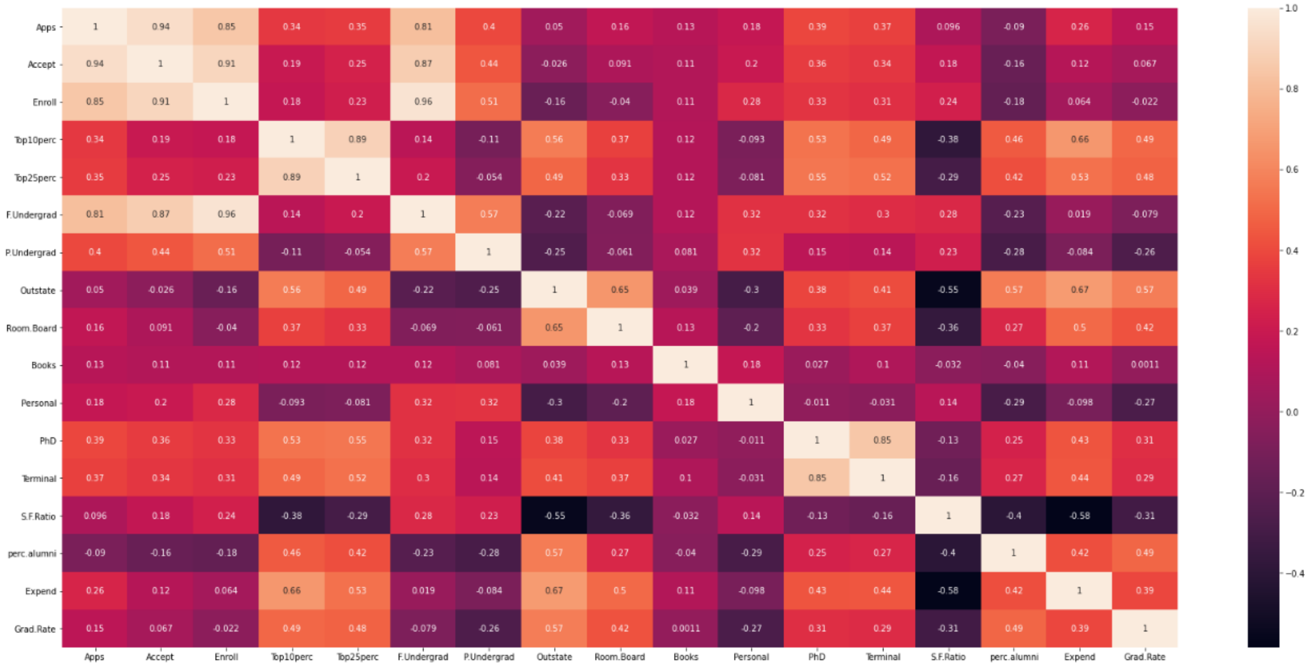
As seen from the above graphs, there are outliers present in all the variables except for Top25perc variable.

Multivariate Analysis:



Dist plots and pair plots suggest that There is large variance in the dataset and hence it is necessary to scale the data and make it a normally distributed data.

Heat map



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1	0.94	0.85	0.34	0.35	0.81	0.4	0.05	0.16	0.13	0.18	0.39	0.37	0.096	-0.09	0.26	0.15
Accept	0.94	1	0.91	0.19	0.25	0.87	0.44	-0.026	0.091	0.11	0.2	0.36	0.34	0.18	-0.16	0.12	0.067
Enroll	0.85	0.91	1	0.18	0.23	0.96	0.51	-0.16	-0.04	0.11	0.28	0.33	0.31	0.24	-0.18	0.064	-0.022
Top10perc	0.34	0.19	0.18	1	0.89	0.14	-0.11	0.56	0.37	0.12	-0.093	0.53	0.49	-0.38	0.46	0.66	0.49
Top25perc	0.35	0.25	0.23	0.89	1	0.2	-0.054	0.49	0.33	0.12	-0.081	0.55	0.52	-0.29	0.42	0.53	0.48
F.Undergrad	0.81	0.87	0.96	0.14	0.2	1	0.57	-0.22	-0.069	0.12	0.32	0.32	0.3	0.28	-0.23	0.019	-0.079
P.Undergrad	0.4	0.44	0.51	-0.11	-0.054	0.57	1	-0.25	-0.061	0.081	0.32	0.15	0.14	0.23	-0.28	-0.084	-0.26
Outstate	0.05	-0.026	-0.16	0.56	0.49	-0.22	-0.25	1	0.65	0.039	-0.3	0.38	0.41	-0.55	0.57	0.67	0.57
Room.Board	0.16	0.091	-0.04	0.37	0.33	-0.069	-0.061	0.65	1	0.13	-0.2	0.33	0.37	-0.36	0.27	0.5	0.42
Books	0.13	0.11	0.11	0.12	0.12	0.12	0.081	0.039	0.13	1	0.18	0.027	0.1	-0.032	-0.04	0.11	0.0011
Personal	0.18	0.2	0.28	-0.093	-0.081	0.32	0.32	-0.3	-0.2	0.18	1	-0.011	-0.031	0.14	-0.29	-0.098	-0.27
PhD	0.39	0.36	0.33	0.53	0.55	0.32	0.15	0.38	0.33	0.027	-0.011	1	0.85	-0.13	0.25	0.43	0.31
Terminal	0.37	0.34	0.31	0.49	0.52	0.3	0.14	0.41	0.37	0.1	-0.031	0.85	1	-0.16	0.27	0.44	0.29
S.F.Ratio	0.096	0.18	0.24	-0.38	-0.29	0.28	0.23	-0.55	-0.36	-0.032	0.14	-0.13	-0.16	1	-0.4	-0.58	-0.31
perc.alumni	-0.09	-0.16	-0.18	0.46	0.42	-0.23	-0.28	0.57	0.27	-0.04	-0.29	0.25	0.27	-0.4	1	0.42	0.49
Expend	0.26	0.12	0.064	0.66	0.53	0.019	-0.084	0.67	0.5	0.11	-0.098	0.43	0.44	-0.58	0.42	1	0.39
Grad.Rate	0.15	0.067	-0.022	0.49	0.48	-0.079	-0.26	0.57	0.42	0.0011	-0.27	0.31	0.29	-0.31	0.49	0.39	1

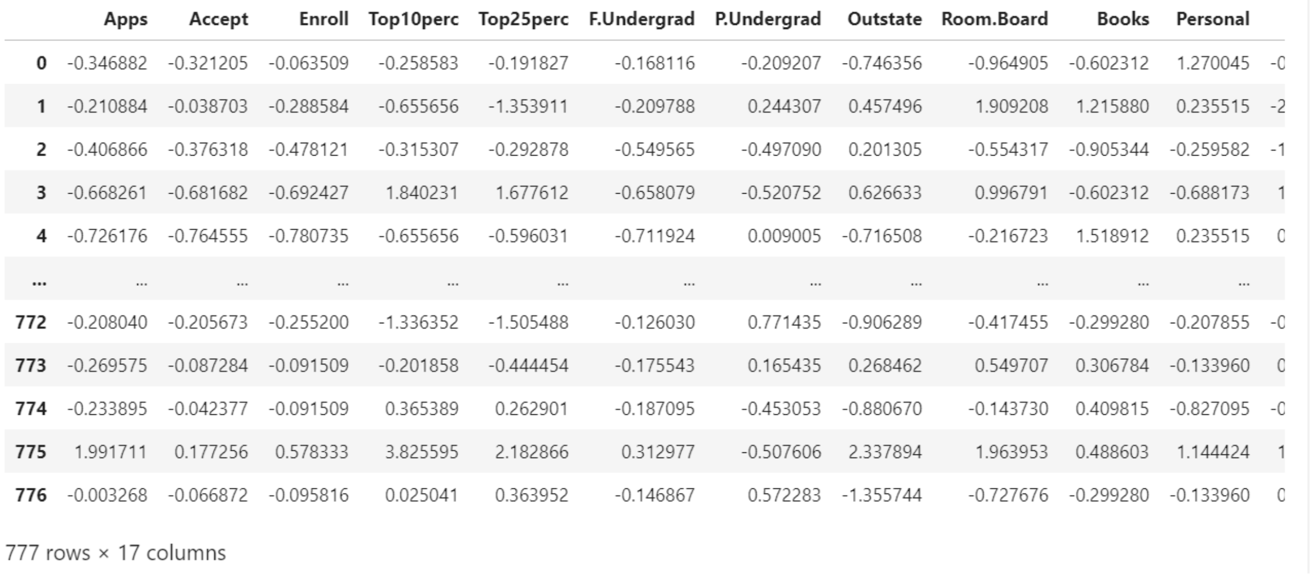
For this particular dataset, the highest correlation is between F.Undergrad and Enroll.

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

It is recommended to standardize or normalize the variables in the dataset to avoid large variability between the columns, as it can result in abnormal results.

Standardization is defined as shifting the distribution of each value to have a mean of zero and standard deviation of unit variance. This approach is considered extremely useful before applying PCA to the dataset.

I used sklearn's standard scalar function to scale the dataset. Below is a screenshot of the scaled variables that we are going to apply PCA on.

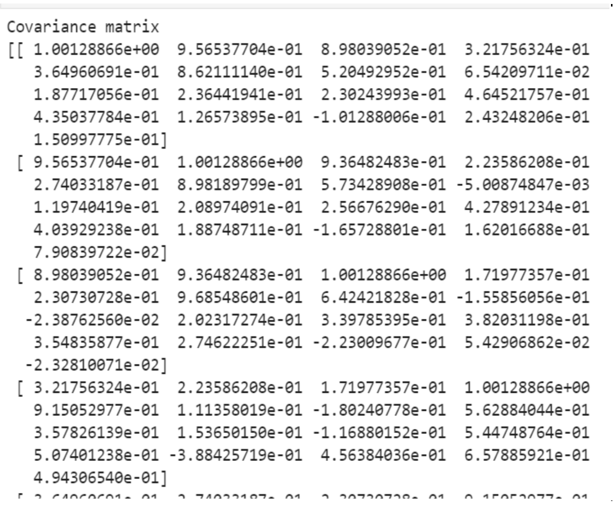


	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	
0	-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.964905	-0.602312	1.270045	-0
1	-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496	1.909208	1.215880	0.235515	-2
2	-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.497090	0.201305	-0.554317	-0.905344	-0.259582	-1
3	-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633	0.996791	-0.602312	-0.688173	1
4	-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005	-0.716508	-0.216723	1.518912	0.235515	0
...	...	...	...	...	...	...	...	...	...	...	...	
772	-0.208040	-0.205673	-0.255200	-1.336352	-1.505488	-0.126030	0.771435	-0.906289	-0.417455	-0.299280	-0.207855	-0
773	-0.269575	-0.087284	-0.091509	-0.201858	-0.444454	-0.175543	0.165435	0.268462	0.549707	0.306784	-0.133960	0
774	-0.233895	-0.042377	-0.091509	0.365389	0.262901	-0.187095	-0.453053	-0.880670	-0.143730	0.409815	-0.827095	-0
775	1.991711	0.177256	0.578333	3.825595	2.182866	0.312977	-0.507606	2.337894	1.963953	0.488603	1.144424	1
776	-0.003268	-0.066872	-0.095816	0.025041	0.363952	-0.146867	0.572283	-1.355744	-0.727676	-0.299280	-0.133960	0

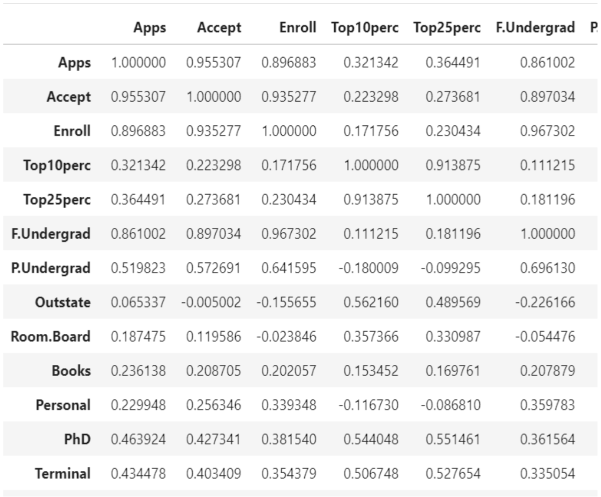
777 rows × 17 columns

In this approach, data on all dimensions are subtracted from their means to shift the data points to the origin and hence the data is centered on the origin.

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



```
Covariance matrix
[[ 1.00128866e+00  9.56537704e-01  8.98039052e-01  3.21756324e-01
   3.64960691e-01  8.62111140e-01  5.20492952e-01  6.54209711e-02
   1.87717056e-01  2.36441941e-01  2.30243993e-01  4.64521757e-01
   4.35037784e-01  1.26573895e-01 -1.01288006e-01  2.43248206e-01
   1.50997775e-01]
 [ 9.56537704e-01  1.00128866e+00  9.36482483e-01  2.23586208e-01
   2.74033187e-01  8.98189799e-01  5.73428908e-01 -5.00874847e-03
   1.19740419e-01  2.08974091e-01  2.56676290e-01  4.27891234e-01
   4.03929238e-01  1.88748711e-01 -1.65728801e-01  1.62016688e-01
   7.90839722e-02]
 [ 8.98039052e-01  9.36482483e-01  1.00128866e+00  1.71977357e-01
   2.30730728e-01  9.68548601e-01  6.42421828e-01 -1.55856056e-01
  -2.38762560e-02  2.02317274e-01  3.39785395e-01  3.82031198e-01
   3.54835877e-01  2.74622251e-01 -2.23009677e-01  5.42906862e-02
  -2.32810071e-02]
 [ 3.21756324e-01  2.23586208e-01  1.71977357e-01  1.00128866e+00
   9.15052977e-01  1.11358019e-01 -1.80240778e-01  5.62884044e-01
   3.57826139e-01  1.53650150e-01 -1.16880152e-01  5.44748764e-01
   5.07401238e-01 -3.88425719e-01  4.56384036e-01  6.57885921e-01
   4.94306540e-01]
 [ 3.64960691e-01  2.74033187e-01  2.30730728e-01  9.15052977e-01
```

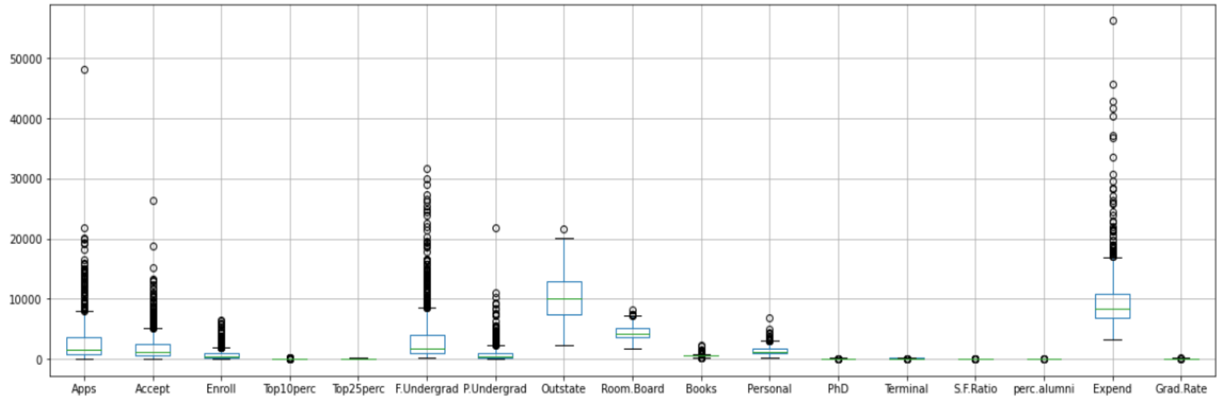


	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P
Apps	1.000000	0.955307	0.896883	0.321342	0.364491	0.861002	
Accept	0.955307	1.000000	0.935277	0.223298	0.273681	0.897034	
Enroll	0.896883	0.935277	1.000000	0.171756	0.230434	0.967302	
Top10perc	0.321342	0.223298	0.171756	1.000000	0.913875	0.111215	
Top25perc	0.364491	0.273681	0.230434	0.913875	1.000000	0.181196	
F.Undergrad	0.861002	0.897034	0.967302	0.111215	0.181196	1.000000	
P.Undergrad	0.519823	0.572691	0.641595	-0.180009	-0.099295	0.696130	
Outstate	0.065337	-0.005002	-0.155655	0.562160	0.489569	-0.226166	
Room.Board	0.187475	0.119586	-0.023846	0.357366	0.330987	-0.054476	
Books	0.236138	0.208705	0.202057	0.153452	0.169761	0.207879	
Personal	0.229948	0.256346	0.339348	-0.116730	-0.086810	0.359783	
PhD	0.463924	0.427341	0.381540	0.544048	0.551461	0.361564	
Terminal	0.434478	0.403409	0.354379	0.506748	0.527654	0.335054	

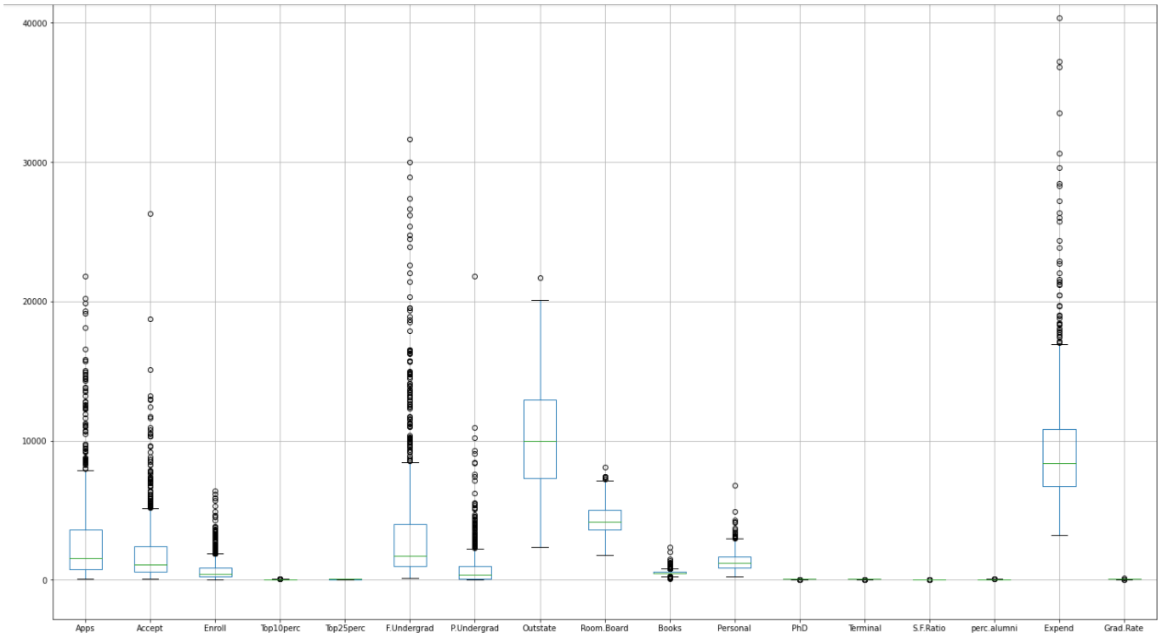
After interpreting the results obtained from covariance matrix and correlation of the scaled variables/standardized values**, co-relation and covariation have the same values after scaling**.

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

As observed from the below boxplot, there are outliers present in many columns **before scaling**

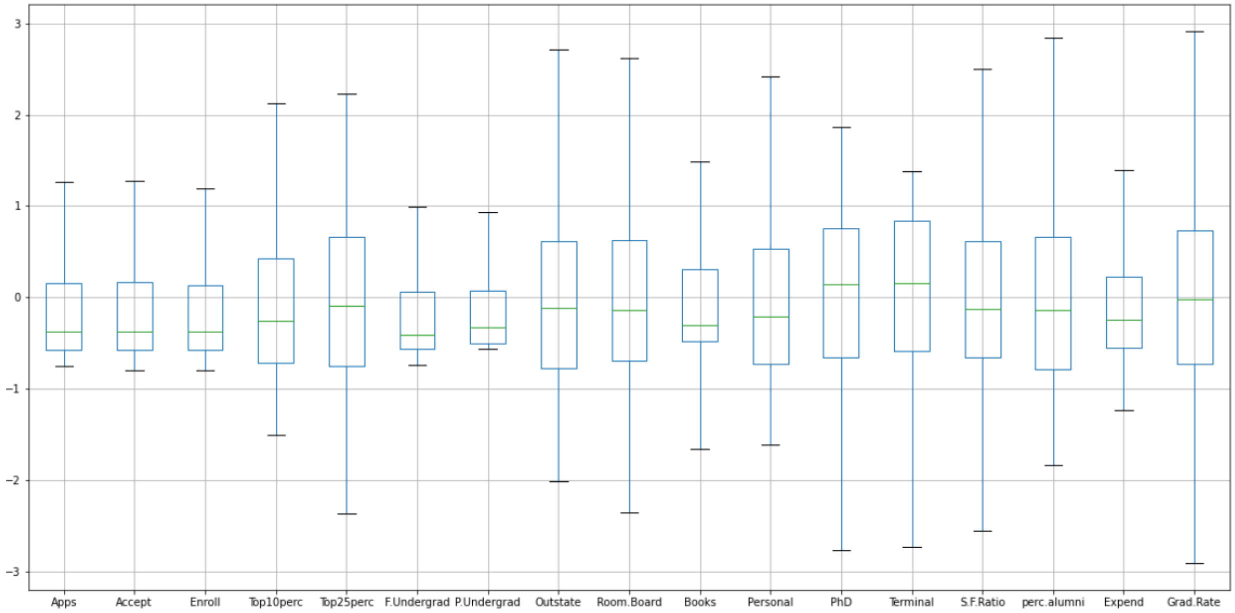


Boxplot obtained after scaling:



There is no difference in the outliers present in the dataset before and after scaling.

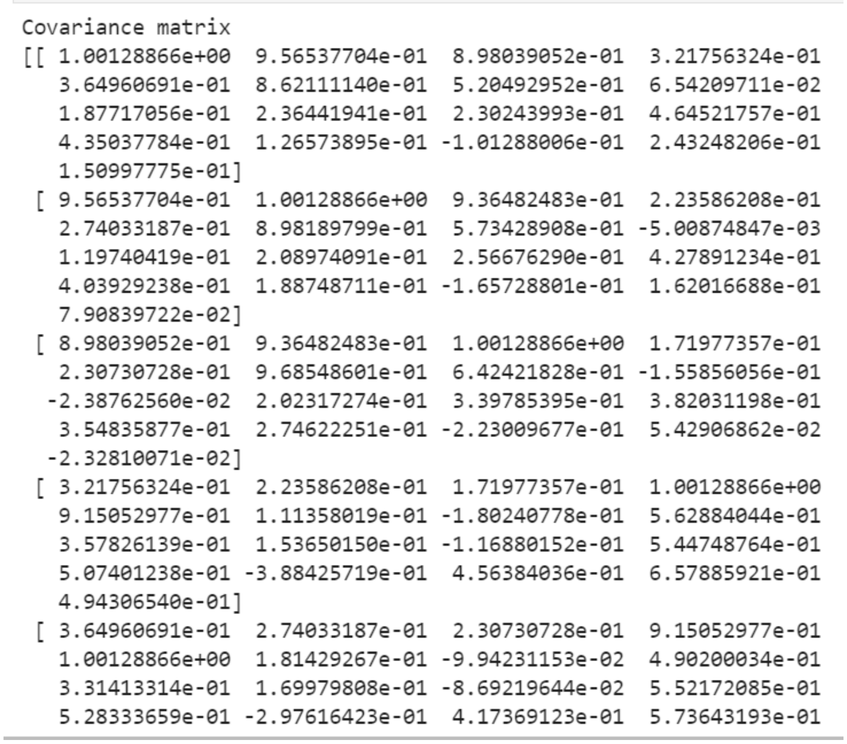
I have used a custom function to deal with the outliers present in the dataset.



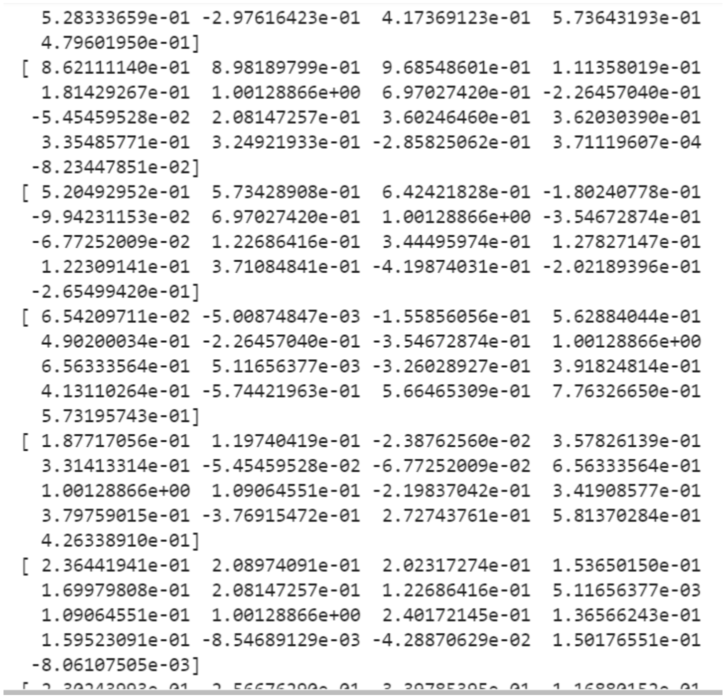
After applying the functions on the affected columns, the outliers are removed.

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

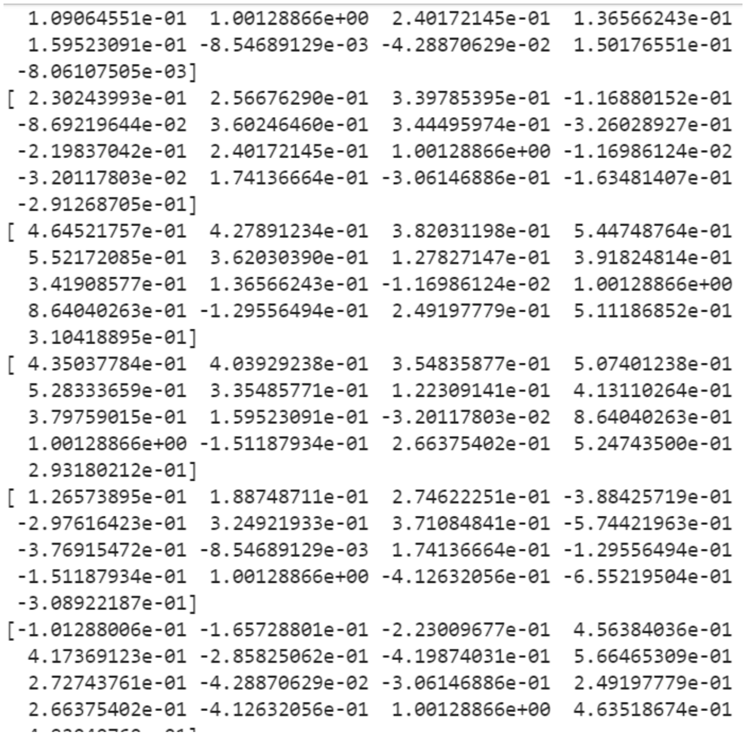
Covariance Matrix



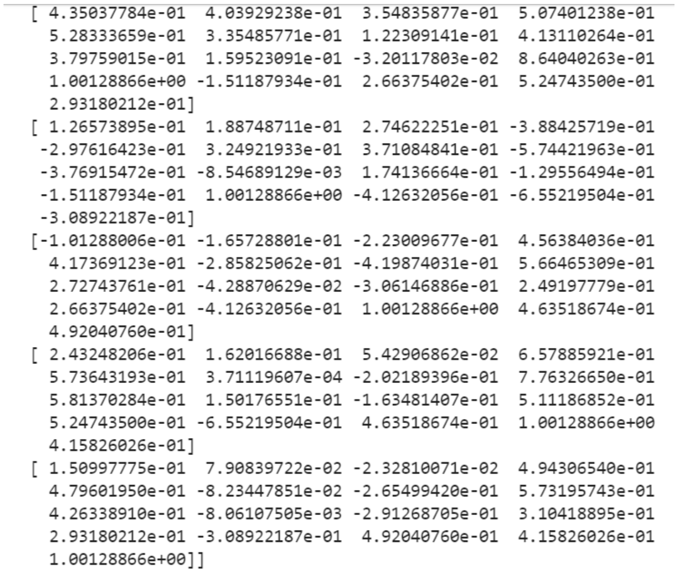
```
Covariance matrix
[[ 1.00128866e+00  9.56537704e-01  8.98039052e-01  3.21756324e-01
   3.64960691e-01  8.62111140e-01  5.20492952e-01  6.54209711e-02
   1.87717056e-01  2.36441941e-01  2.30243993e-01  4.64521757e-01
   4.35037784e-01  1.26573895e-01 -1.01288006e-01  2.43248206e-01
   1.50997775e-01]
 [ 9.56537704e-01  1.00128866e+00  9.36482483e-01  2.23586208e-01
   2.74033187e-01  8.98189799e-01  5.73428908e-01 -5.00874847e-03
   1.19740419e-01  2.08974091e-01  2.56676290e-01  4.27891234e-01
   4.03929238e-01  1.88748711e-01 -1.65728801e-01  1.62016688e-01
   7.90839722e-02]
 [ 8.98039052e-01  9.36482483e-01  1.00128866e+00  1.71977357e-01
   2.30730728e-01  9.68548601e-01  6.42421828e-01 -1.55856056e-01
  -2.38762560e-02  2.02317274e-01  3.39785395e-01  3.82031198e-01
   3.54835877e-01  2.74622251e-01 -2.23009677e-01  5.42906862e-02
  -2.32810071e-02]
 [ 3.21756324e-01  2.23586208e-01  1.71977357e-01  1.00128866e+00
   9.15052977e-01  1.11358019e-01 -1.80240778e-01  5.62884044e-01
   3.57826139e-01  1.53650150e-01 -1.16880152e-01  5.44748764e-01
   5.07401238e-01 -3.88425719e-01  4.56384036e-01  6.57885921e-01
   4.94306540e-01]
 [ 3.64960691e-01  2.74033187e-01  2.30730728e-01  9.15052977e-01
   1.00128866e+00  1.81429267e-01 -9.94231153e-02  4.90200034e-01
   3.31413314e-01  1.69979808e-01 -8.69219644e-02  5.52172085e-01
   5.28333659e-01 -2.97616423e-01  4.17369123e-01  5.73643193e-01
```



```
   5.28333659e-01 -2.97616423e-01  4.17369123e-01  5.73643193e-01
   4.79601950e-01]
 [ 8.62111140e-01  8.98189799e-01  9.68548601e-01  1.11358019e-01
   1.81429267e-01  1.00128866e+00  6.97027420e-01 -2.26457040e-01
  -5.45459528e-02  2.08147257e-01  3.60246460e-01  3.62030390e-01
   3.35485771e-01  3.24921933e-01 -2.85825062e-01  3.71119607e-04
  -8.23447851e-02]
 [ 5.20492952e-01  5.73428908e-01  6.42421828e-01 -1.80240778e-01
  -9.94231153e-02  6.97027420e-01  1.00128866e+00 -3.54672874e-01
  -6.77252009e-02  1.22686416e-01  3.44495974e-01  1.27827147e-01
   1.22309141e-01  3.71084841e-01 -4.19874031e-01 -2.02189396e-01
  -2.65499420e-01]
 [ 6.54209711e-02 -5.00874847e-03 -1.55856056e-01  5.62884044e-01
   4.90200034e-01 -2.26457040e-01 -3.54672874e-01  1.00128866e+00
   6.56333564e-01  5.11656377e-03 -3.26028927e-01  3.91824814e-01
   4.13110264e-01 -5.74421963e-01  5.66465309e-01  7.76326650e-01
   5.73195743e-01]
 [ 1.87717056e-01  1.19740419e-01 -2.38762560e-02  3.57826139e-01
   3.31413314e-01 -5.45459528e-02 -6.77252009e-02  6.56333564e-01
   1.00128866e+00  1.09064551e-01 -2.19837042e-01  3.41908577e-01
   3.79759015e-01 -3.76915472e-01  2.72743761e-01  5.81370284e-01
   4.26338910e-01]
 [ 2.36441941e-01  2.08974091e-01  2.02317274e-01  1.53650150e-01
   1.69979808e-01  2.08147257e-01  1.22686416e-01  5.11656377e-03
   1.09064551e-01  1.00128866e+00  2.40172145e-01  1.36566243e-01
   1.59523091e-01 -8.54689129e-03 -4.28870629e-02  1.50176551e-01
  -8.06107505e-03]
 [ 2.30243993e-01  2.56676290e-01  3.39785395e-01  1.16880152e-01
```



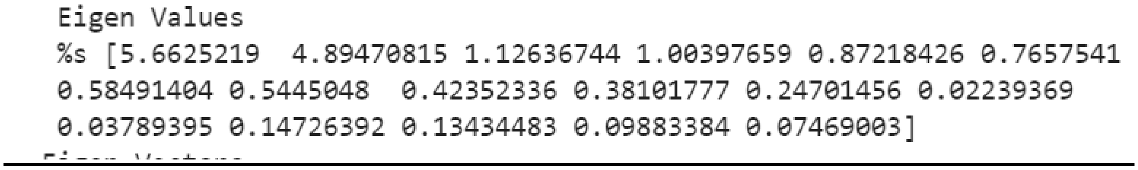
```
  1.09064551e-01  1.00128866e+00  2.40172145e-01  1.36566243e-01
  1.59523091e-01 -8.54689129e-03 -4.28870629e-02  1.50176551e-01
 -8.06107505e-03]
[ 2.30243993e-01  2.56676290e-01  3.39785395e-01 -1.16880152e-01
 -8.69219644e-02  3.60246460e-01  3.44495974e-01 -3.26028927e-01
 -2.19837042e-01  2.40172145e-01  1.00128866e+00 -1.16986124e-02
 -3.20117803e-02  1.74136664e-01 -3.06146886e-01 -1.63481407e-01
 -2.91268705e-01]
[ 4.64521757e-01  4.27891234e-01  3.82031198e-01  5.44748764e-01
  5.52172085e-01  3.62030390e-01  1.27827147e-01  3.91824814e-01
  3.41908577e-01  1.36566243e-01 -1.16986124e-02  1.00128866e+00
  8.64040263e-01 -1.29556494e-01  2.49197779e-01  5.11186852e-01
  3.10418895e-01]
[ 4.35037784e-01  4.03929238e-01  3.54835877e-01  5.07401238e-01
  5.28333659e-01  3.35485771e-01  1.22309141e-01  4.13110264e-01
  3.79759015e-01  1.59523091e-01 -3.20117803e-02  8.64040263e-01
  1.00128866e+00 -1.51187934e-01  2.66375402e-01  5.24743500e-01
  2.93180212e-01]
[ 1.26573895e-01  1.88748711e-01  2.74622251e-01 -3.88425719e-01
 -2.97616423e-01  3.24921933e-01  3.71084841e-01 -5.74421963e-01
 -3.76915472e-01 -8.54689129e-03  1.74136664e-01 -1.29556494e-01
 -1.51187934e-01  1.00128866e+00 -4.12632056e-01 -6.55219504e-01
 -3.08922187e-01]
[-1.01288006e-01 -1.65728801e-01 -2.23009677e-01  4.56384036e-01
  4.17369123e-01 -2.85825062e-01 -4.19874031e-01  5.66465309e-01
  2.72743761e-01 -4.28870629e-02 -3.06146886e-01  2.49197779e-01
  2.66375402e-01 -4.12632056e-01  1.00128866e+00  4.63518674e-01
```



```
[ 4.35037784e-01  4.03929238e-01  3.54835877e-01  5.07401238e-01
  5.28333659e-01  3.35485771e-01  1.22309141e-01  4.13110264e-01
  3.79759015e-01  1.59523091e-01 -3.20117803e-02  8.64040263e-01
  1.00128866e+00 -1.51187934e-01  2.66375402e-01  5.24743500e-01
  2.93180212e-01]
[ 1.26573895e-01  1.88748711e-01  2.74622251e-01 -3.88425719e-01
 -2.97616423e-01  3.24921933e-01  3.71084841e-01 -5.74421963e-01
 -3.76915472e-01 -8.54689129e-03  1.74136664e-01 -1.29556494e-01
 -1.51187934e-01  1.00128866e+00 -4.12632056e-01 -6.55219504e-01
 -3.08922187e-01]
[-1.01288006e-01 -1.65728801e-01 -2.23009677e-01  4.56384036e-01
  4.17369123e-01 -2.85825062e-01 -4.19874031e-01  5.66465309e-01
  2.72743761e-01 -4.28870629e-02 -3.06146886e-01  2.49197779e-01
  2.66375402e-01 -4.12632056e-01  1.00128866e+00  4.63518674e-01
```

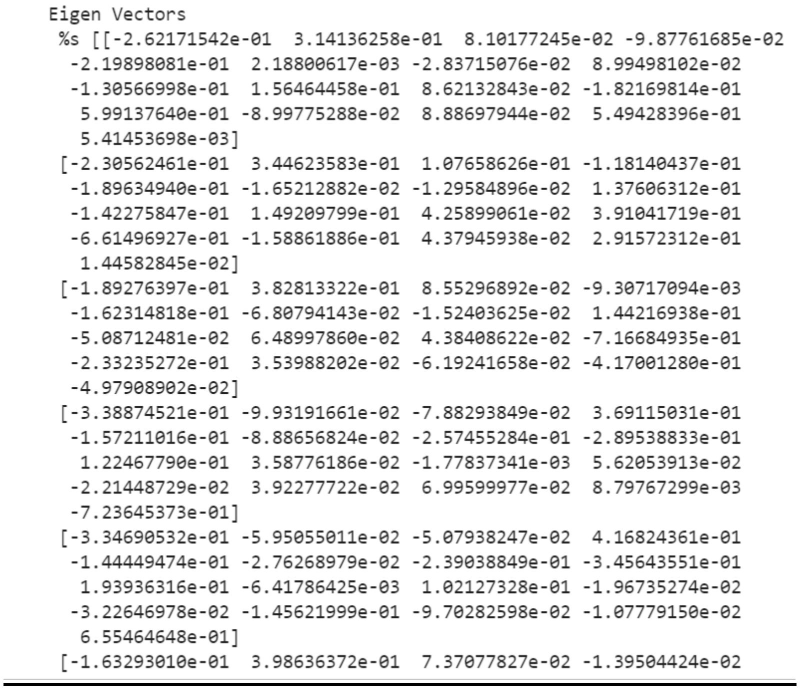
```
  4.92040760e-01]
[ 2.43248206e-01  1.62016688e-01  5.42906862e-02  6.57885921e-01
  5.73643193e-01  3.71119607e-04 -2.02189396e-01  7.76326650e-01
  5.81370284e-01  1.50176551e-01 -1.63481407e-01  5.11186852e-01
  5.24743500e-01 -6.55219504e-01  4.63518674e-01  1.00128866e+00
  4.15826026e-01]
[ 1.50997775e-01  7.90839722e-02 -2.32810071e-02  4.94306540e-01
  4.79601950e-01 -8.23447851e-02 -2.65499420e-01  5.73195743e-01
  4.26338910e-01 -8.06107505e-03 -2.91268705e-01  3.10418895e-01
  2.93180212e-01 -3.08922187e-01  4.92040760e-01  4.15826026e-01
  1.00128866e+00]]
```

Eigen Values:

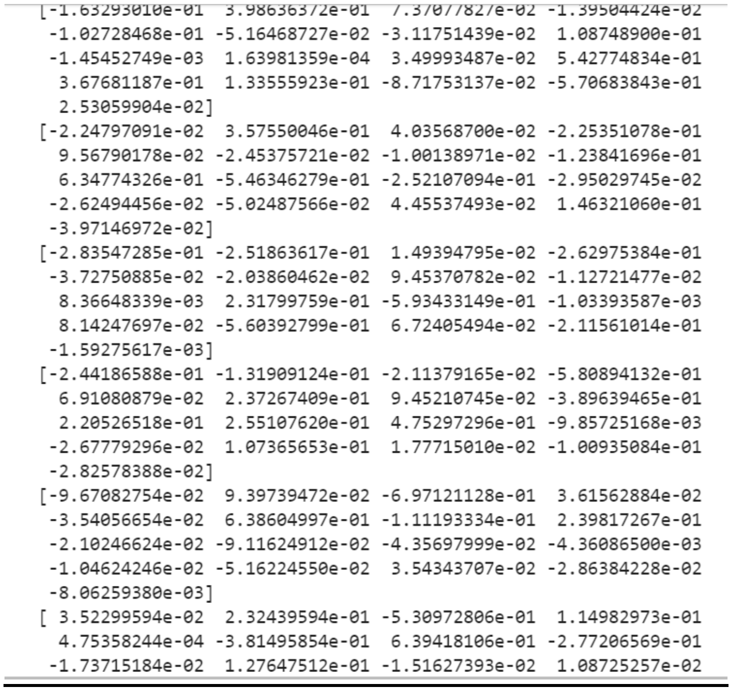


```
Eigen Values
%s [5.6625219  4.89470815 1.12636744 1.00397659 0.87218426 0.7657541
0.58491404 0.5445048  0.42352336 0.38101777 0.24701456 0.02239369
0.03789395 0.14726392 0.13434483 0.09883384 0.07469003]
```

Eigen Vectors:

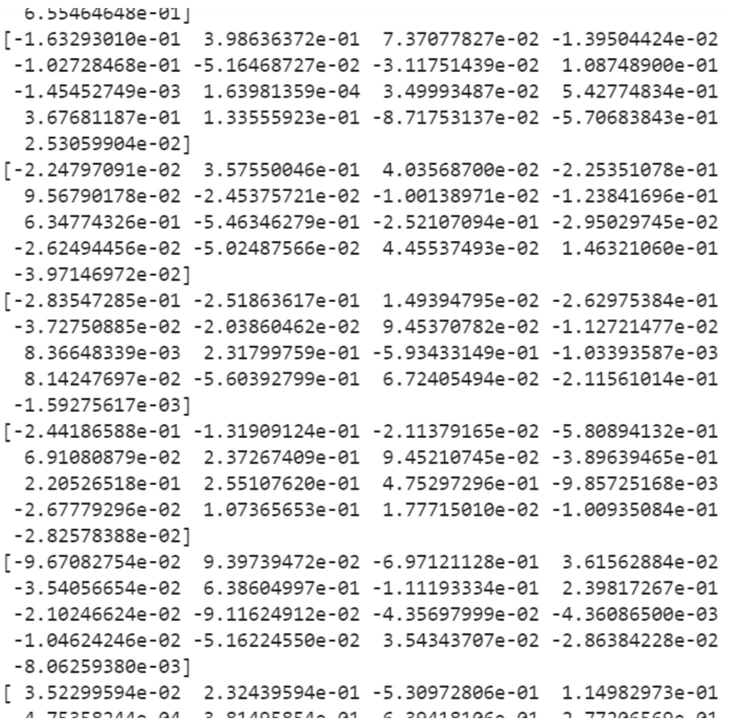


```
Eigen Vectors
 %s [[-2.62171542e-01  3.14136258e-01  8.10177245e-02 -9.87761685e-02
  -2.19898081e-01  2.18800617e-03 -2.83715076e-02  8.99498102e-02
  -1.30566998e-01  1.56464458e-01  8.62132843e-02 -1.82169814e-01
   5.99137640e-01 -8.99775288e-02  8.88697944e-02  5.49428396e-01
   5.41453698e-03]
 [-2.30562461e-01  3.44623583e-01  1.07658626e-01 -1.18140437e-01
  -1.89634940e-01 -1.65212882e-02 -1.29584896e-02  1.37606312e-01
  -1.42275847e-01  1.49209799e-01  4.25899061e-02  3.91041719e-01
  -6.61496927e-01 -1.58861886e-01  4.37945938e-02  2.91572312e-01
   1.44582845e-02]
 [-1.89276397e-01  3.82813322e-01  8.55296892e-02 -9.30717094e-03
  -1.62314818e-01 -6.80794143e-02 -1.52403625e-02  1.44216938e-01
  -5.08712481e-02  6.48997860e-02  4.38408622e-02 -7.16684935e-01
  -2.33235272e-01  3.53988202e-02 -6.19241658e-02 -4.17001280e-01
  -4.97908902e-02]
 [-3.38874521e-01 -9.93191661e-02 -7.88293849e-02  3.69115031e-01
  -1.57211016e-01 -8.88656824e-02 -2.57455284e-01 -2.89538833e-01
   1.22467790e-01  3.58776186e-02 -1.77837341e-03  5.62053913e-02
  -2.21448729e-02  3.92277722e-02  6.99599977e-02  8.79767299e-03
  -7.23645373e-01]
 [-3.34690532e-01 -5.95055011e-02 -5.07938247e-02  4.16824361e-01
  -1.44449474e-01 -2.76268979e-02 -2.39038849e-01 -3.45643551e-01
   1.93936316e-01 -6.41786425e-03  1.02127328e-01 -1.96735274e-02
  -3.22646978e-02 -1.45621999e-01 -9.70282598e-02 -1.07779150e-02
   6.55464648e-01]
 [-1.63293010e-01  3.98636372e-01  7.37077827e-02 -1.39504424e-02
```

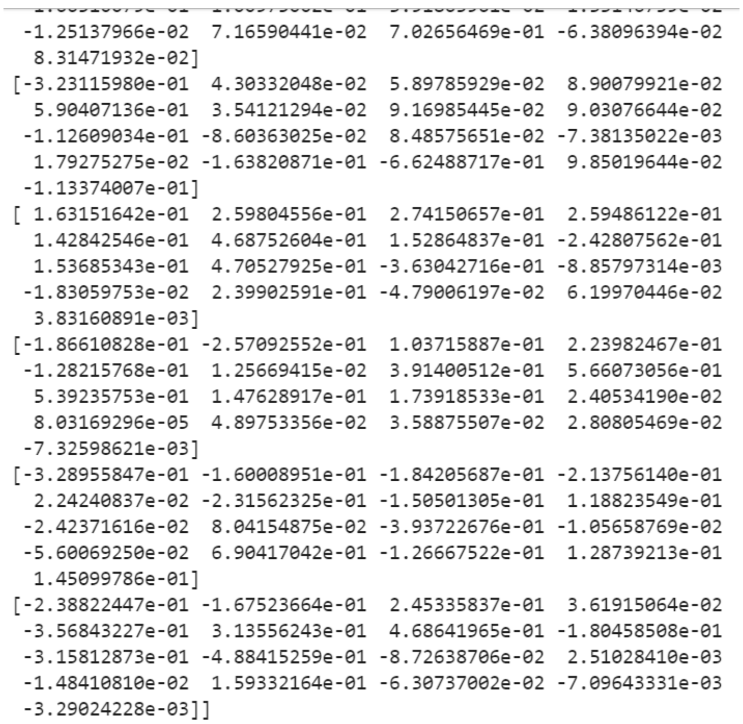


```
 [-1.63293010e-01  3.98636372e-01  7.37077827e-02 -1.39504424e-02
```

```
  -1.02728468e-01 -5.16468727e-02 -3.11751439e-02  1.08748900e-01
  -1.45452749e-03  1.63981359e-04  3.49993487e-02  5.42774834e-01
   3.67681187e-01  1.33555923e-01 -8.71753137e-02 -5.70683843e-01
   2.53059904e-02]
 [-2.24797091e-02  3.57550046e-01  4.03568700e-02 -2.25351078e-01
   9.56790178e-02 -2.45375721e-02 -1.00138971e-02 -1.23841696e-01
   6.34774326e-01 -5.46346279e-01 -2.52107094e-01 -2.95029745e-02
  -2.62494456e-02 -5.02487566e-02  4.45537493e-02  1.46321060e-01
  -3.97146972e-02]
 [-2.83547285e-01 -2.51863617e-01  1.49394795e-02 -2.62975384e-01
  -3.72750885e-02 -2.03860462e-02  9.45370782e-02 -1.12721477e-02
   8.36648339e-03  2.31799759e-01 -5.93433149e-01 -1.03393587e-03
   8.14247697e-02 -5.60392799e-01  6.72405494e-02 -2.11561014e-01
  -1.59275617e-03]
 [-2.44186588e-01 -1.31909124e-01 -2.11379165e-02 -5.80894132e-01
   6.91080879e-02  2.37267409e-01  9.45210745e-02 -3.89639465e-01
   2.20526518e-01  2.55107620e-01  4.75297296e-01 -9.85725168e-03
  -2.67779296e-02  1.07365653e-01  1.77715010e-02 -1.00935084e-01
  -2.82578388e-02]
 [-9.67082754e-02  9.39739472e-02 -6.97121128e-01  3.61562884e-02
  -3.54056654e-02  6.38604997e-01 -1.11193334e-01  2.39817267e-01
  -2.10246624e-02 -9.11624912e-02 -4.35697999e-02 -4.36086500e-03
  -1.04624246e-02 -5.16224550e-02  3.54343707e-02 -2.86384228e-02
  -8.06259380e-03]
 [ 3.52299594e-02  2.32439594e-01 -5.30972806e-01  1.14982973e-01
   4.75358244e-04 -3.81495854e-01  6.39418106e-01 -2.77206569e-01
  -1.73715184e-02  1.27647512e-01 -1.51627393e-02  1.08725257e-02
```



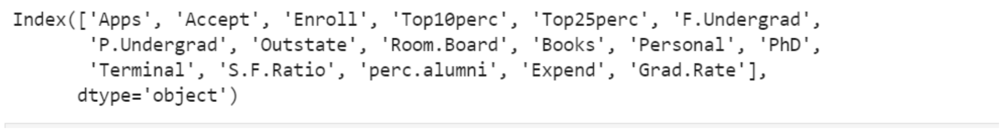
```
  6.55464648e-01]
 [-1.63293010e-01  3.98636372e-01  7.37077827e-02 -1.39504424e-02
  -1.02728468e-01 -5.16468727e-02 -3.11751439e-02  1.08748900e-01
  -1.45452749e-03  1.63981359e-04  3.49993487e-02  5.42774834e-01
   3.67681187e-01  1.33555923e-01 -8.71753137e-02 -5.70683843e-01
   2.53059904e-02]
 [-2.24797091e-02  3.57550046e-01  4.03568700e-02 -2.25351078e-01
   9.56790178e-02 -2.45375721e-02 -1.00138971e-02 -1.23841696e-01
   6.34774326e-01 -5.46346279e-01 -2.52107094e-01 -2.95029745e-02
  -2.62494456e-02 -5.02487566e-02  4.45537493e-02  1.46321060e-01
  -3.97146972e-02]
 [-2.83547285e-01 -2.51863617e-01  1.49394795e-02 -2.62975384e-01
  -3.72750885e-02 -2.03860462e-02  9.45370782e-02 -1.12721477e-02
   8.36648339e-03  2.31799759e-01 -5.93433149e-01 -1.03393587e-03
   8.14247697e-02 -5.60392799e-01  6.72405494e-02 -2.11561014e-01
  -1.59275617e-03]
 [-2.44186588e-01 -1.31909124e-01 -2.11379165e-02 -5.80894132e-01
   6.91080879e-02  2.37267409e-01  9.45210745e-02 -3.89639465e-01
   2.20526518e-01  2.55107620e-01  4.75297296e-01 -9.85725168e-03
  -2.67779296e-02  1.07365653e-01  1.77715010e-02 -1.00935084e-01
  -2.82578388e-02]
 [-9.67082754e-02  9.39739472e-02 -6.97121128e-01  3.61562884e-02
  -3.54056654e-02  6.38604997e-01 -1.11193334e-01  2.39817267e-01
  -2.10246624e-02 -9.11624912e-02 -4.35697999e-02 -4.36086500e-03
  -1.04624246e-02 -5.16224550e-02  3.54343707e-02 -2.86384228e-02
  -8.06259380e-03]
 [ 3.52299594e-02  2.32439594e-01 -5.30972806e-01  1.14982973e-01
   4.75358244e-04  3.81495854e-01  6.39418106e-01  2.77396569e-01
```



```
  -1.25137966e-02  7.16590441e-02  7.02656469e-01 -6.38096394e-02
   8.31471932e-02]
 [-3.23115980e-01  4.30332048e-02  5.89785929e-02  8.90079921e-02
   5.90407136e-01  3.54121294e-02  9.16985445e-02  9.03076644e-02
  -1.12609034e-01 -8.60363025e-02  8.48575651e-02 -7.38135022e-03
   1.79275275e-02 -1.63820871e-01 -6.62488717e-01  9.85019644e-02
  -1.13374007e-01]
 [ 1.63151642e-01  2.59804556e-01  2.74150657e-01  2.59486122e-01
   1.42842546e-01  4.68752604e-01  1.52864837e-01 -2.42807562e-01
   1.53685343e-01  4.70527925e-01 -3.63042716e-01 -8.85797314e-03
  -1.83059753e-02  2.39902591e-01 -4.79006197e-02  6.19970446e-02
   3.83160891e-03]
 [-1.86610828e-01 -2.57092552e-01  1.03715887e-01  2.23982467e-01
  -1.28215768e-01  1.25669415e-02  3.91400512e-01  5.66073056e-01
   5.39235753e-01  1.47628917e-01  1.73918533e-01  2.40534190e-02
   8.03169296e-05  4.89753356e-02  3.58875507e-02  2.80805469e-02
  -7.32598621e-03]
 [-3.28955847e-01 -1.60008951e-01 -1.84205687e-01 -2.13756140e-01
   2.24240837e-02 -2.31562325e-01 -1.50501305e-01  1.18823549e-01
  -2.42371616e-02  8.04154875e-02 -3.93722676e-01 -1.05658769e-02
  -5.60069250e-02  6.90417042e-01 -1.26667522e-01  1.28739213e-01
   1.45099786e-01]
 [-2.38822447e-01 -1.67523664e-01  2.45335837e-01  3.61915064e-02
  -3.56843227e-01  3.13556243e-01  4.68641965e-01 -1.80458508e-01
  -3.15812873e-01 -4.88415259e-01 -8.72638706e-02  2.51028410e-03
  -1.48410810e-02  1.59332164e-01 -6.30737002e-02 -7.09643331e-03
  -3.29024228e-03]]
```

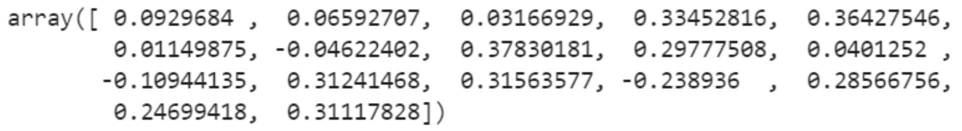
2.6) Write the explicit form of the first PC (in terms of Eigen Vectors).

Columns



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

Eigen Vector[0]

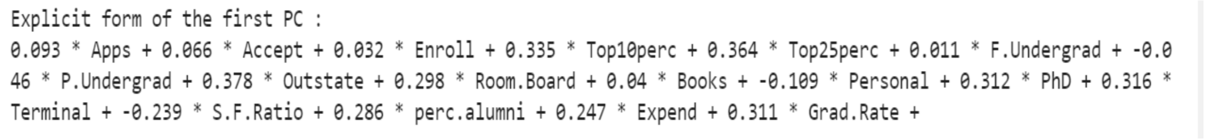


```
array([ 0.0929684 ,  0.06592707,  0.03166929,  0.33452816,  0.36427546,
        0.01149875, -0.04622402,  0.37830181,  0.29777508,  0.0401252 ,
       -0.10944135,  0.31241468,  0.31563577, -0.238936  ,  0.28566756,
        0.24699418,  0.31117828])
```

After performing matrix multiplication of columns and Eigen Vector[0]

Principal component 1 = EV[0][0] * Col1 + EV[0][1] * Col2 .... + EV[0][9] * col10

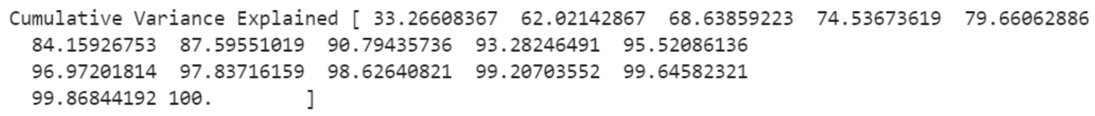
Therefore, explicit/ linear equation form of First Principal Component is:



```
Explicit form of the first PC :
0.093 * Apps + 0.066 * Accept + 0.032 * Enroll + 0.335 * Top10perc + 0.364 * Top25perc + 0.011 * F.Undergrad + -0.0
46 * P.Undergrad + 0.378 * Outstate + 0.298 * Room.Board + 0.04 * Books + -0.109 * Personal + 0.312 * PhD + 0.316 *
Terminal + -0.239 * S.F.Ratio + 0.286 * perc.alumni + 0.247 * Expend + 0.311 * 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.

Cumulative values of the eigenvalues are below:



```
Cumulative Variance Explained [ 33.26608367  62.02142867  68.63859223  74.53673619  79.66062886
  84.15926753  87.59551019  90.79435736  93.28246491  95.52086136
  96.97201814  97.83716159  98.62640821  99.20703552  99.64582321
  99.86844192 100.        ]
```

First component explains 33.2% of variance in data.

First two components explain 62% of variance in data.

First three components explain 68.6% of variance in data

First four components explain 74.5% of variance in data

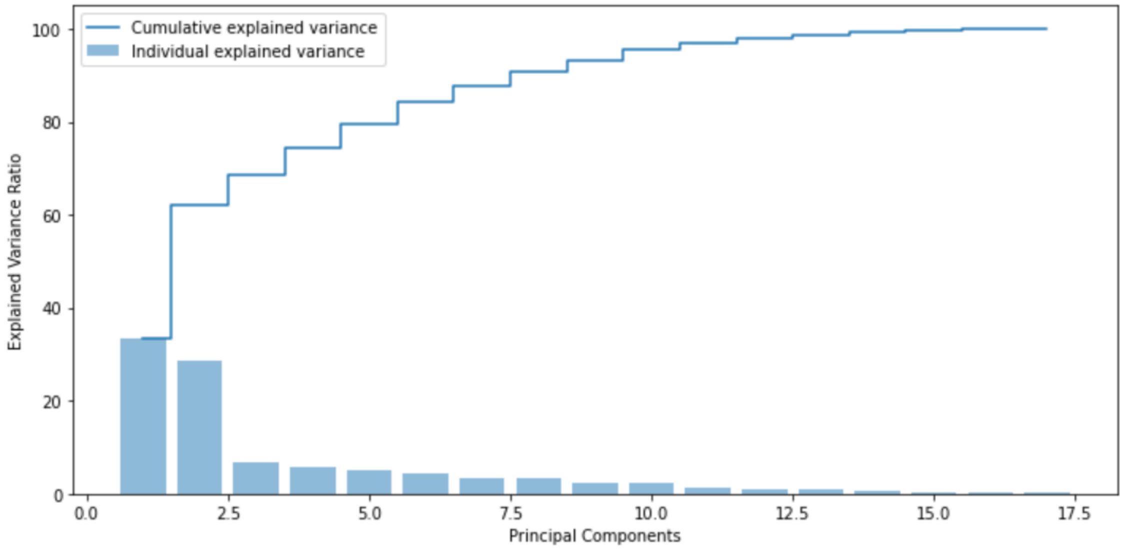
First five components explain 79.66% of variance in data.

First six components explain 84.1% of variance in data.

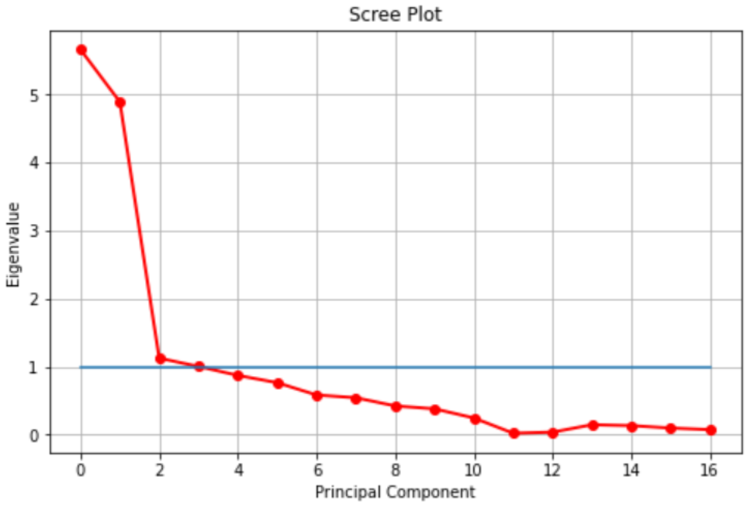
First seven components explain 87.5% of variance in data.

If we employ 4 features we capture ~ 74.5% of the variance within the dataset

Choosing optimal number of principal components



We can notice that Eigen vectors 10, 11 hold very little information i.e. Eigen vectors with insignificant contribution to total Eigen values can be removed from analysis.

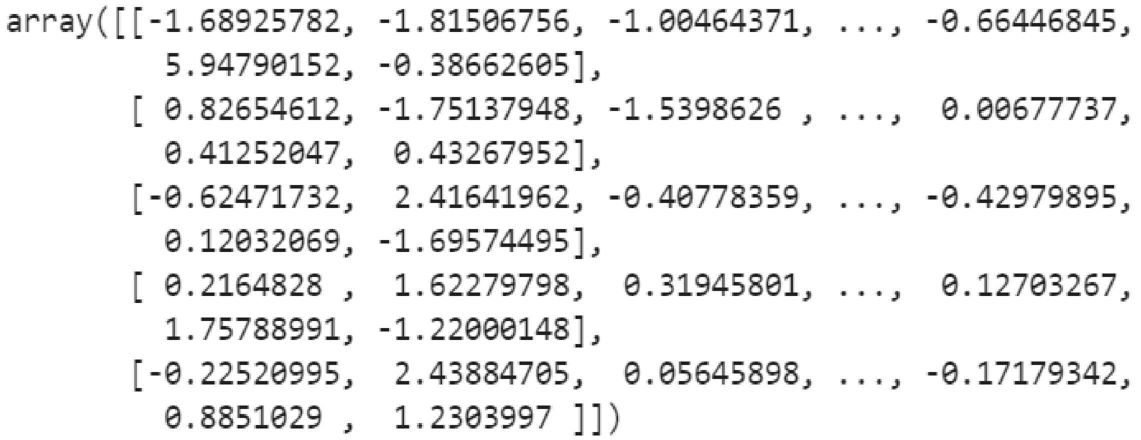


Principal component with eigenvalues greater than 1 are considered more significant. Initially, I decided to choose 4 number of principal components as principal component 4 is close to 1 and hence significant. But it captures only 74.5% of variance in the data, the information captured is comparatively not enough.

So Depending on the requirements of this problem, **5 seems to be an optimal number of principal components.**

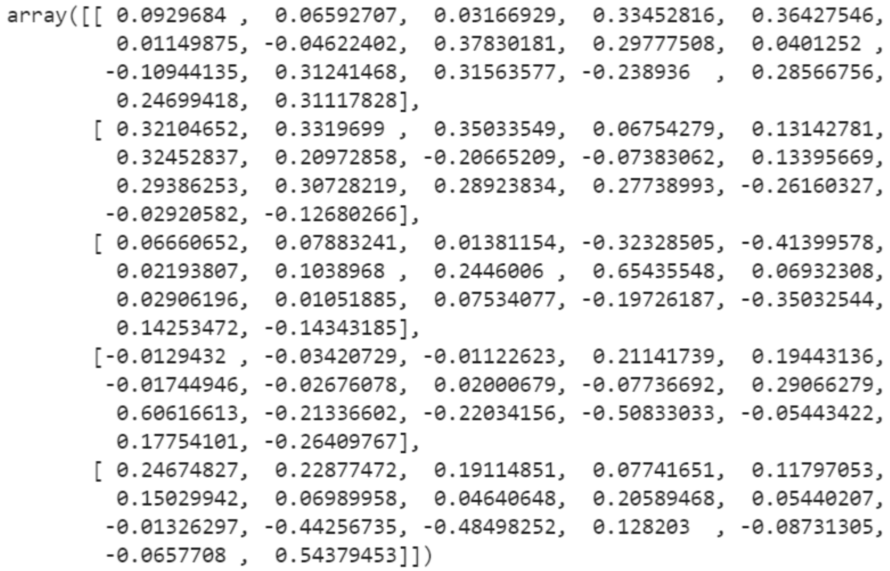
Eigen vectors are the principal components that determine the directions of the new feature space and the corresponding Eigen values which are the magnitudes of variance captured

PCA applied to dataset



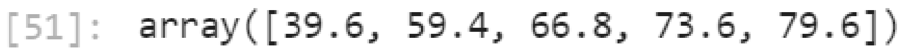
```
array([[-1.68925782, -1.81506756, -1.00464371, ..., -0.66446845,
         5.94790152, -0.38662605],
       [ 0.82654612, -1.75137948, -1.5398626 , ...,  0.00677737,
         0.41252047,  0.43267952],
       [-0.62471732,  2.41641962, -0.40778359, ..., -0.42979895,
         0.12032069, -1.69574495],
       [ 0.2164828 ,  1.62279798,  0.31945801, ...,  0.12703267,
         1.75788991, -1.22000148],
       [-0.22520995,  2.43884705,  0.05645898, ..., -0.17179342,
         0.8851029 ,  1.2303997 ]])
```

PCA Component scores



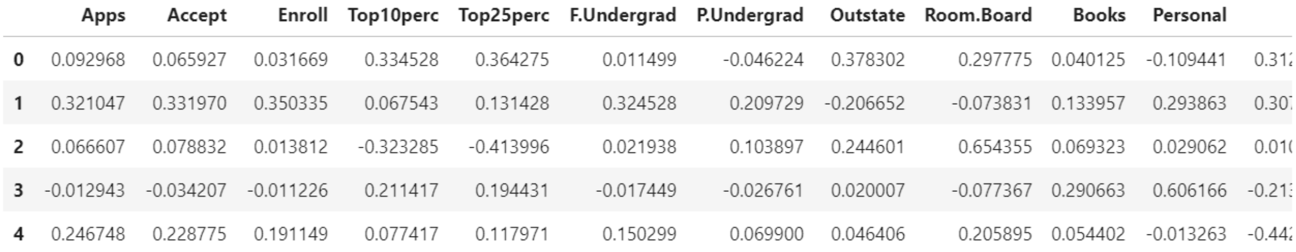
```
array([[ 0.0929684 ,  0.06592707,  0.03166929,  0.33452816,  0.36427546,
         0.01149875, -0.04622402,  0.37830181,  0.29777508,  0.0401252 ,
        -0.10944135,  0.31241468,  0.31563577, -0.238936  ,  0.28566756,
         0.24699418,  0.31117828],
       [ 0.32104652,  0.3319699 ,  0.35033549,  0.06754279,  0.13142781,
         0.32452837,  0.20972858, -0.20665209, -0.07383062,  0.13395669,
         0.29386253,  0.30728219,  0.28923834,  0.27738993, -0.26160327,
        -0.02920582, -0.12680266],
       [ 0.06660652,  0.07883241,  0.01381154, -0.32328505, -0.41399578,
         0.02193807,  0.1038968 ,  0.2446006 ,  0.65435548,  0.06932308,
         0.02906196,  0.01051885,  0.07534077, -0.19726187, -0.35032544,
         0.14253472, -0.14343185],
       [-0.0129432 , -0.03420729, -0.01122623,  0.21141739,  0.19443136,
        -0.01744946, -0.02676078,  0.02000679, -0.07736692,  0.29066279,
         0.60616613, -0.21336602, -0.22034156, -0.50833033, -0.05443422,
         0.17754101, -0.26409767],
       [ 0.24674827,  0.22877472,  0.19114851,  0.07741651,  0.11797053,
         0.15029942,  0.06989958,  0.04640648,  0.20589468,  0.05440207,
        -0.01326297, -0.44256735, -0.48498252,  0.128203  , -0.08731305,
        -0.0657708 ,  0.54379453]])
```

Cumulative sum of variance explained with 5 features. variance is explained by all the five components is approximately 80%. The information captured by five components altogether is 80% approximately.



```
[51]:  array([39.6, 59.4, 66.8, 73.6, 79.6])
```

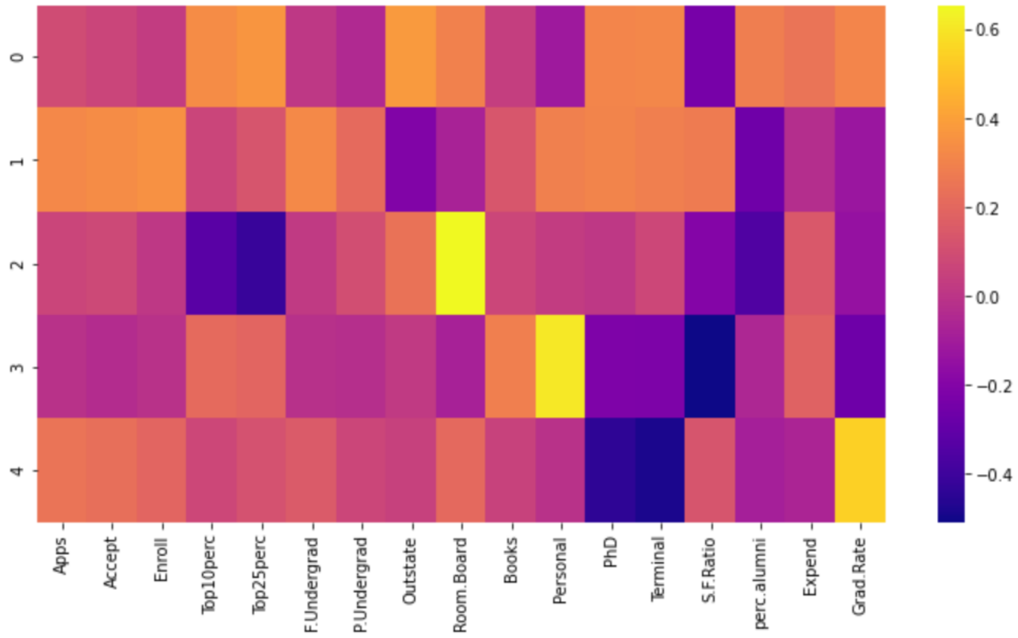
Reduced new Data frame



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	
0	0.092968	0.065927	0.031669	0.334528	0.364275	0.011499	-0.046224	0.378302	0.297775	0.040125	-0.109441	0.31
1	0.321047	0.331970	0.350335	0.067543	0.131428	0.324528	0.209729	-0.206652	-0.073831	0.133957	0.293863	0.30
2	0.066607	0.078832	0.013812	-0.323285	-0.413996	0.021938	0.103897	0.244601	0.654355	0.069323	0.029062	0.01
3	-0.012943	-0.034207	-0.011226	0.211417	0.194431	-0.017449	-0.026761	0.020007	-0.077367	0.290663	0.606166	-0.21
4	0.246748	0.228775	0.191149	0.077417	0.117971	0.150299	0.069900	0.046406	0.205895	0.054402	-0.013263	-0.44

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

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



Component 1 can be labelled as P.Undergrad as it looks more related.

Component 2 can be renamed as Apps.

This way, all co-related relationships can be labelled with features depending on the relationships and thus removing unnecessary features from the dataset.

Principal Component Analysis is used to remove the redundant features from the datasets without losing much information. If there is a research study that is to be conducted on student's motivation to enroll in an institution/university, this approach can be used to remove unnecessary features and reduce the dimensionality of the dataset.

Or another fine business Implication for this problem would be in identifying appropriate and significant factors/parameters for assessment of each university’s performance in the context of university rankings.

PCA analysis can be applied on 777 universities with a list of 17 factors to determine the most relevant components.

For example, Graduation rate seems to be one of the relevant and uncorrelated factor in assessing an institution’s performance in the context of university rankings.