1 ANALYSIS ON SALARY DATA USING ANOVA TECHNIQUE

INFERENCE OF THE DATASET:-

1. The shape of the dataset seems to be with 40 rows and 3 columns.
2. The columns seems to be integer or object values.
3. We also can see they are no duplicates in the dataset.
4. The entire dataset does not have missing values or null values.

```
Education     0
Occupation    0
Salary        0
dtype: int64
```

Q.1- STATE THE NULL AND THE ALTERNATE HYPOTHESIS FOR CONDUCTING ONE-WAY ANOVA FOR BOTH EDUCATION AND OCCUPATION INDIVIDUALLY.

One way ANOVA (Education)

Null Hypothesis $H0$: The mean salary is the same across all the 3 categories of education (Doctorate, Bachelors, and HS-Grad).

Alternate Hypothesis $H1$: The mean salary is different in at least one category of education.

One way ANOVA (Occupation)

Null Hypothesis $H0$: The mean salary is the same across all the 4 categories of occupation (Prof-Specialty, Sales, Adm-clerical, and Exec-Managerial).

Alternate Hypothesis $H1$: The mean salary is different in at least one category of occupation.

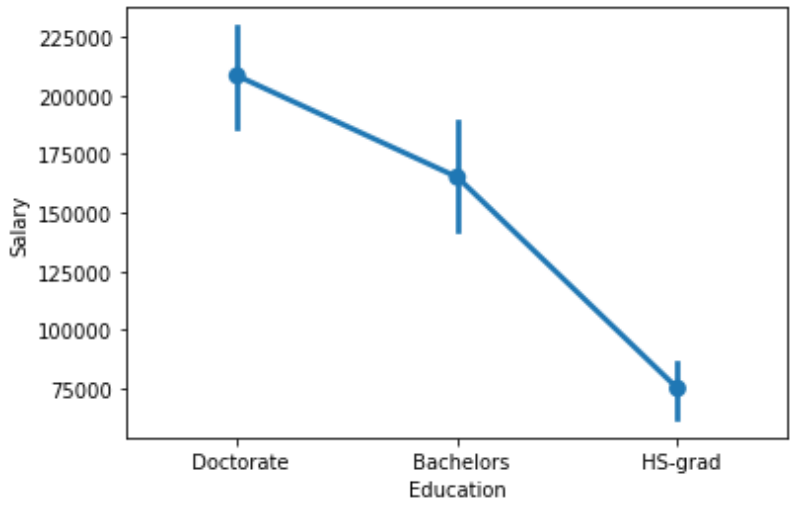
Q.2- PERFORM A ONE-WAY ANOVA ON SALARY WITH RESPECT TO EDUCATION. STATE WHETHER THE NULL HYPOTHESIS IS ACCEPTED OR REJECTED BASED ON THE ANOVA RESULTS.

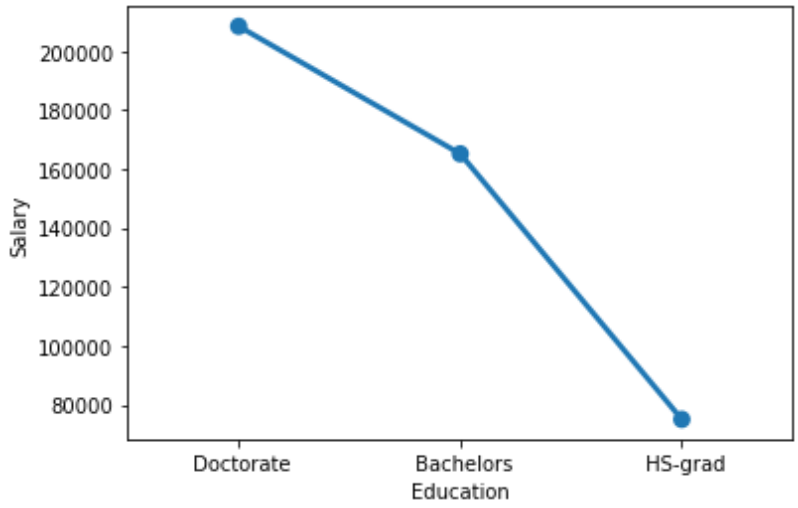
	DF	sum sq.	Mean sq.	F	PR(>F)
C (Education)	2.0	1.026955e+11	5.134773e+10	30.95628	1.257709e-08
Residual	37.0	6.137256e+10	1.658718e+09	Nan	Nan

Above is the ANOVA table for Education variable.

Since the p value = 1.257709e-08 is less than the significance level (alpha = 0.05), we can reject the null hypothesis and conclude that there is a significant difference in the mean salaries for at least one category of education.

Here's a Graph for better and easy understanding:-





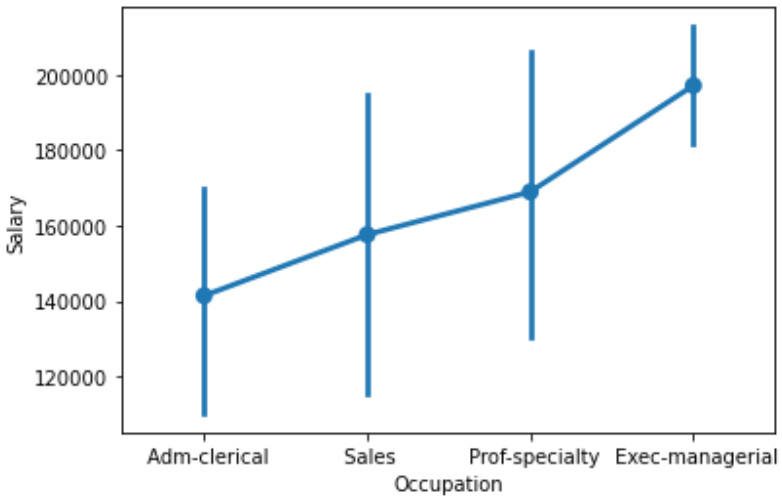
Q.3- Perform a one-way ANOVA on Salary with respect to Occupation. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

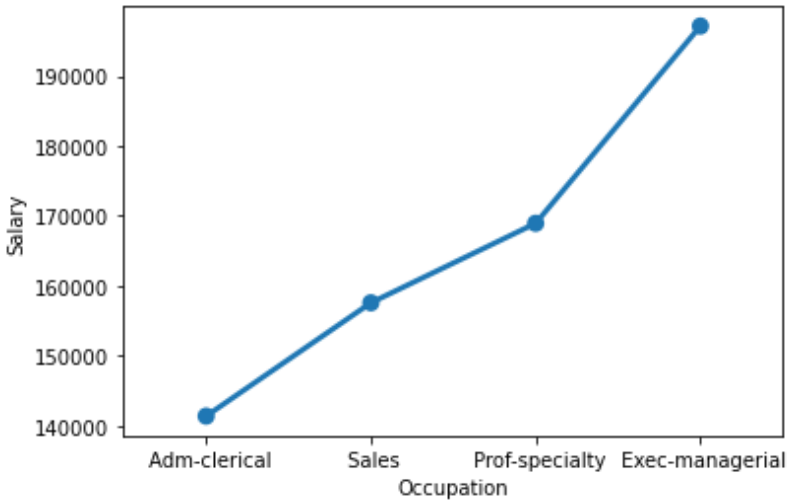
	df	sum sq.	Mean sq.	F	PR(>F)
C (Occupation)	3.0	1.125878e+10	3.752928e+09	0.884144	0.458508
Residual	36.0	1.528092e+11	4.244701e+09	Nan	Nan

The above is the ANOVA table for Occupation variable.

Since the p value = 0.458508 is greater than the significance level (alpha = 0.05), we fail to reject the null hypothesis (i.e. we accept H0) and conclude that there is no significant difference in the mean salaries across the 4 categories of occupation.

Here's a Graph for better and easy understanding:-



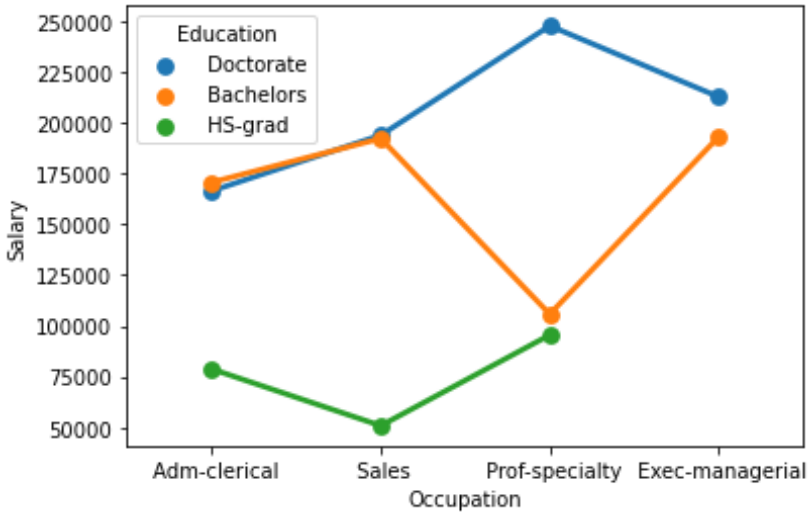


Q.4- IF THE NULL HYPOTHESIS IS REJECTED IN EITHER (2) OR IN (3), FIND OUT WHICH CLASS MEANS ARE SIGNIFICANTLY DIFFERENT. INTERPRET THE RESULT.

If we compare both the table regarding education and occupation. We see that education have a lesser P-value from the significance level i.e. (alpha = 0.05) and for occupation we can see a higher P-value which is greater than significance value. So from the P-value of both the results we can conclude that P-value being lower in education implies that the mean salaries across all categories of education are different and the P-value being greater implies that the mean salaries across all occupation classes are significantly same.

PROBLEM 1B:

Q.1- WHAT IS THE INTERACTION BETWEEN TWO TREATMENTS? ANALYZE THE EFFECTS OF ONE VARIABLE ON THE OTHER (EDUCATION AND OCCUPATION) WITH THE HELP OF AN INTERACTION PLOT.



The interaction plot shows that there is significant amount of interaction between the categorical variables, Education and Occupation.

The following are some of the observations from the interaction plot:

· People with HS-grad education do not reach the position of Exec-managerial and they hold only Adm-clerk, Sales and Prof-Specialty occupations.

· People with education as Bachelors or Doctorate and occupation as Adm-clerical and Sales almost earn the same salaries (salaries ranging from 170000–190000).

· People with education as Bachelors and occupation as Prof-Specialty earn lesser than people with education as Bachelors and occupations as Adm-clerical and Sales.

· People with education as Bachelors and occupation Sales earn higher than people with education as Bachelors and occupation Prof-Specialty whereas people with education as Doctorate and occupation Sales earn lesser than people with Doctorate and occupation Prof-Specialty. We see a reversal in this part of the plot.

· Similarly, people with education as Bachelors and occupation as Prof-Specialty earn lesser than people with education as Bachelors and occupation Exec-Managerial whereas people with education as Doctorate and occupation as Prof-Specialty earn higher than people with education as Doctorate and occupation Exec-Managerial. There is a reversal in this part of the plot too.

· Salespeople with Bachelors or Doctorate education earn the same salaries and earn higher than people with education as HS-grad.

· Adm clerical people with education as HS-grad earn the lowest salaries when compared to people with education as Bachelors or Doctorate.

· Prof-Specialty people with education as Doctorate earn maximum salaries and people with education as HS-Grad earn the minimum.

· People with education as HS -Grad earn the minimum salaries.

· There are no people with education as HS -grad who hold Exec-managerial occupation.

· People with education as Bachelors and occupation, Sales and Exec-Managerial earn the same salaries.

Q.2- Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?

Two way ANOVA

$H0$: The effect of the independent variable 'education' on the mean 'salary' does not depend on the effect of the other independent variable 'occupation' (i. e. there is no interaction effect between the 2 independent variables, education and occupation).

$H1$: There is an interaction effect between the independent variable 'education' and the independent variable 'occupation' on the mean Salary.

	df	sum sq.	Mean sq.	F\
C (Education)	**2.0**	**1.026955e+11**	**5.134773e+10**	**72.211958**
C (Occupation)	**3.0**	**5.519946e+09**	**1.839982e+09**	**2.587626**
C (Education): C (Occupation)	**6.0**	**3.634909e+10**	**6.058182e+09**	**8.519815**
Residual	**29.0**	**2.062102e+10**	**7.110697e+08**	**Nan**

	PR(>F)
C (Education)	**5.466264e-12**
C (Occupation)	**7.211580e-02**
C (Education):C(Occupation)	**2.232500e-05**
Residual	**Nan**

From the table, we see that there is a significant amount of interaction between the variables, Education and Occupation.

As p value = 2.232500e-05 is lesser than the significance level (alpha = 0.05), we reject the null hypothesis.

Thus, we see that there is an interaction effect between education and occupation on the mean salary.

Q.3- Explain the business implications of performing ANOVA for this particular case study.

From the ANOVA method and the interaction plot, we see that education combined with occupation results in higher and better salaries among the people. It is clearly seen that people with education as Doctorate draw the maximum salaries and people with education HS-grad earn the least. Thus, we can conclude that Salary is dependent on educational qualifications and occupation.

END

ANALYSIS OF "EDUCATION - POST 12TH STANDARD" DATASET USING EDA (EXPLORATORY DATA ANALYSIS) & PCA (PRINCIPAL COMPONENT ANALYSIS) TECHNIQUE'S :-

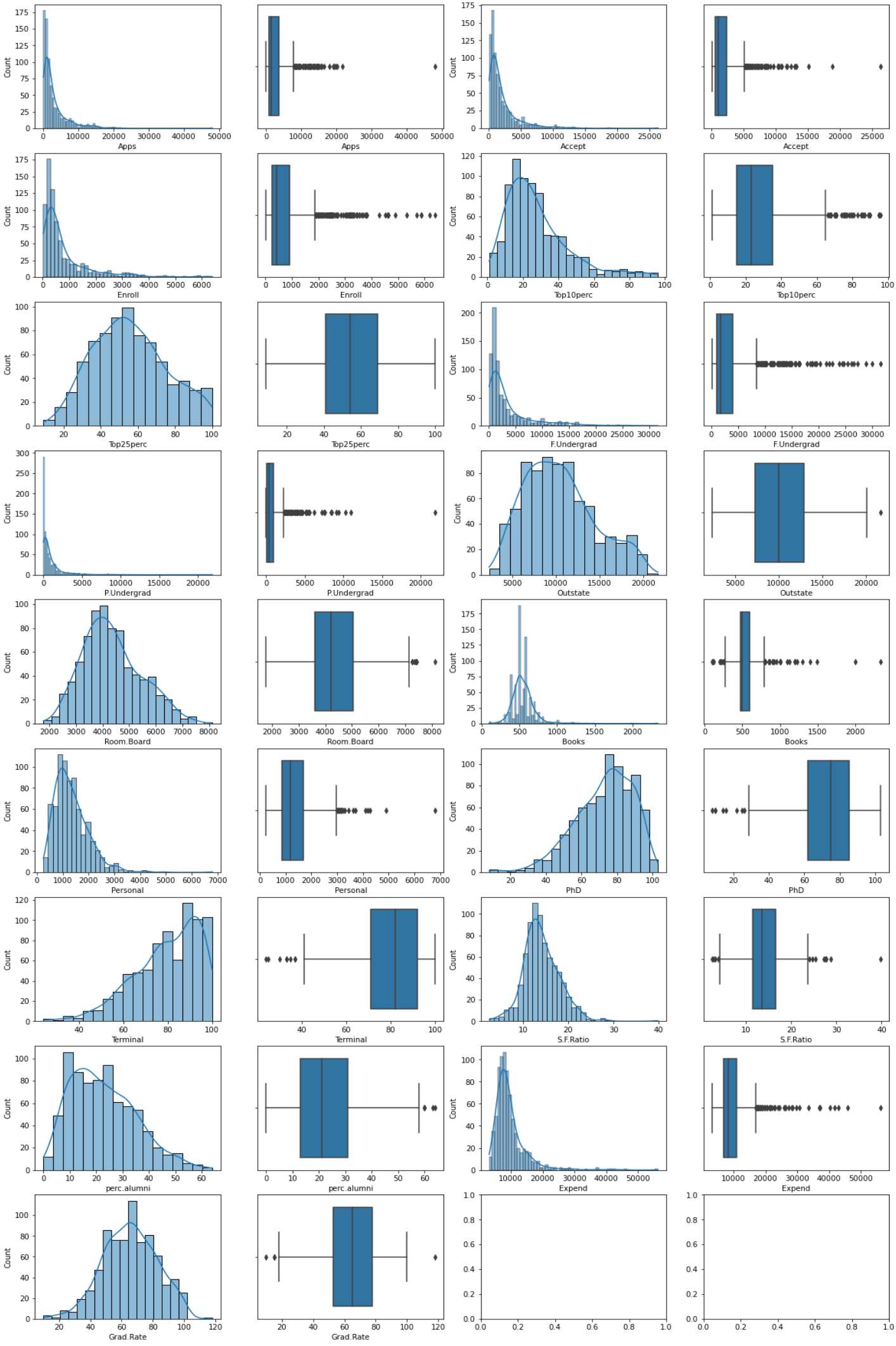
The given dataset consists of data points of names of various university and college which has number of application received, accepted, and enrolled, percentage of new students from top 10% of higher secondary class, percentage of new students from top 25% of higher secondary class, Number of fulltime undergraduates, Number of part-time undergraduate students, Number of students for whom the particular college is out of state tuition, cost of room and board, estimated book costs for a student, estimated personal spending for a student, percentage of faculties with PHD, percentage of faculties with terminal degree, student/faculty ratio, percentage of alumni who donate, The instructional expenditure per student, Graduation Rate.

INFERENCE OF THE DATASET:-

1. The shape of the dataset seems to be with 777 rows and 18 columns.
2. All the columns seems to be integer or float values.
3. The Names column alone is a categorical value.
4. We also can see they are no duplicates in the dataset.
5. The entire dataset does not have missing values or null 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
```

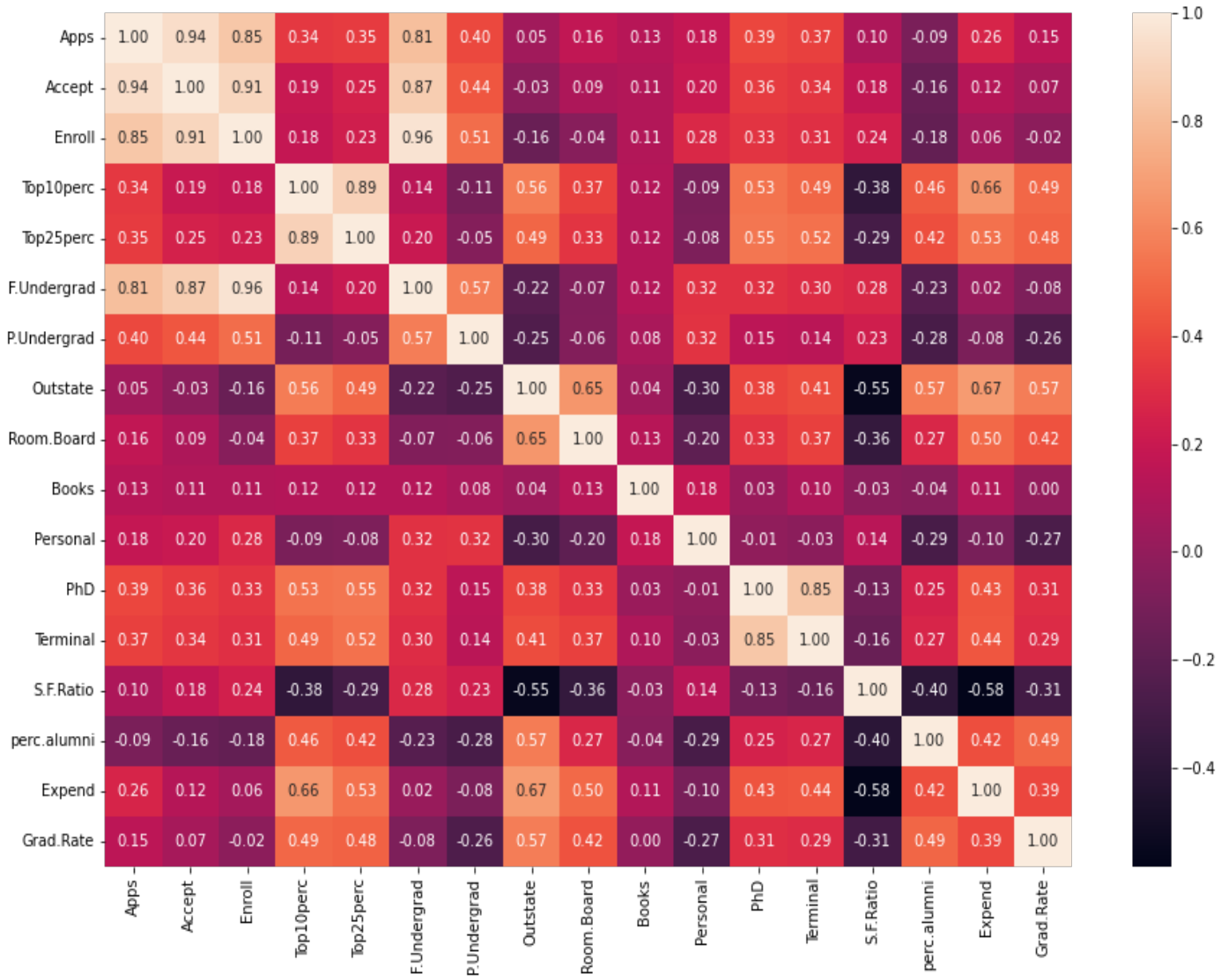
Q.1- Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?


Count
Apps
Apps
Accept
Accept
Enroll
Enroll
Top10perc
Top10perc
Top25perc
Top25perc
F.Undergrad
F.Undergrad
P.Undergrad
P.Undergrad
Outstate
Outstate
Room.Board
Room.Board
Books
Books
Personal
Personal
PhD
PhD
Terminal
Terminal
S.F.Ratio
S.F.Ratio
perc.alumni
perc.alumni
Expend
Expend
Grad.Rate
Grad.Rate

Insights:-

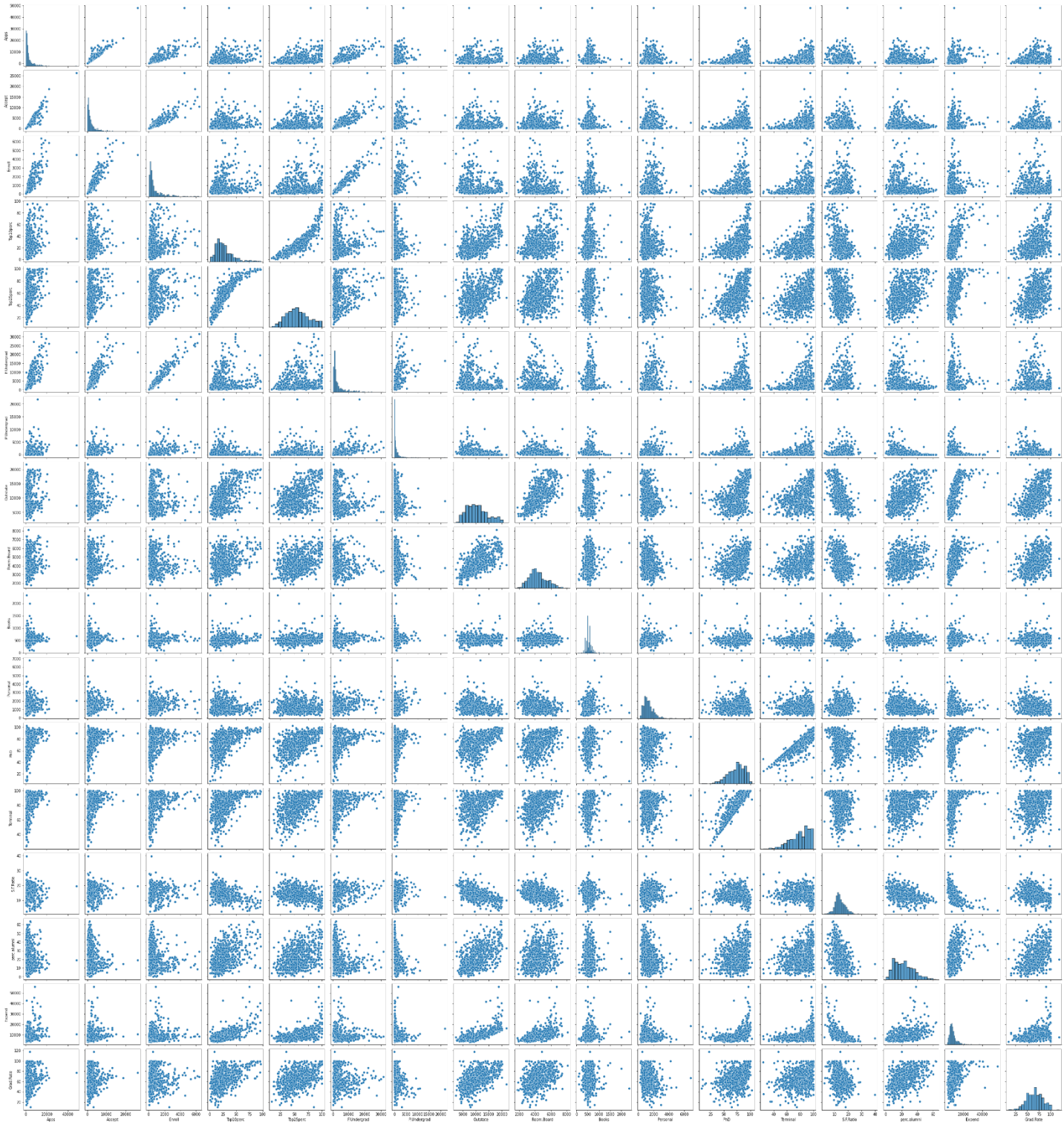
1. The Box plot of Apps variable seems to have outliers, the distribution of the data is skewed we could also understand that each college or university offers application. The max applications seems to be around 50,000.
2. The accept variable seems to have outliers. The distplot shows us the majority of applications accepted from each university.
3. The box plot of the Enroll variable also has outliers. The distribution of the data is positively skewed. From the distplot we can understand majority of the colleges have enrolled students.
4. The box plot of the students from top 10 percentage of higher secondary class seems to have outliers. There is good amount of intake.
5. The box plot for the top 25% has no outliers. The distribution is almost normally distributed. Majority of the students are from top 25% of higher secondary class.
6. The box plot of the full time graduates has outliers. The distribution of the data is positively skewed. There are full time graduates studying in all the university.
7. The box plot of the part time graduates has outliers. The distribution of the data is positively skewed. There are part-time graduates studying in all the university.
8. The box plot of outstate has only one outlier. The distribution is almost normally distributed.
9. The Room board has few outliers. The distribution is normally distributed.
10. The box plot of books has outliers. The distribution seems to be bimodal.
11. The box plot of personal expense has outliers. Some student's personal expense are way bigger than the rest of the students. The distribution seems to be positively skewed.
12. The box plot of PHD has outliers. The distribution seems to be negatively skewed.
13. The box plot of terminal seems to have outliers in the dataset. The distribution for the terminal also seems to be negatively skewed.
14. The SF ratio variable also has outliers in the dataset. The distribution is almost normally distributed. The student faculty ratio is almost same in all the university and colleges.
15. The percentage of alumni box plot seems to have outliers in the dataset. The distribution is almost normally distributed.
16. The expenditure variable also has outliers in the dataset. The distribution of the expenditure is positively skewed.
17. The graduation rate among the students in all the university above 60%. The box plot of the graduation rate has outliers in the dataset. The distribution is normally distributed.

HEATMAP:-



This Heat map gives us the correlation between two numerical values. We could understand the application variable is highly positively correlated with application accepted, students enrolled and full time graduates. So this relationship gives the insights on when student submits the application it is accepted and the student is enrolled as fulltime graduate. We can find negative correlation between application and percentage of alumni. This indicates us not all students are part of alumni of their college or university. The application with top 10, 25 of higher secondary class, outstate, room board, books, personal, PhD, terminal, S.F ratio, expenditure and Graduation ratio are positively correlated.

PAIRPLOT:-



The pair plot helps us to understand the relationship between all the numerical values in the dataset. On comparing all the variables with each other we could understand the patterns or trends in the dataset.

Q.2 - Is scaling necessary for PCA in this case? Give justification and perform scaling.

First, we describe the data and see the values:-

	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

We can observe here is, that for all these variable's we can see, we find a lot of magnitude in these features:

1. We can see that minimum number of application received is 81 and max no of application received is 7896.
2. Minimum no of application that are accepted is 72 while the maximum no of application accepted is 5154 and so on and so forth.

So, from the descriptive data we can see that there is a difference in magnitude in it. So, this is a concern for us, we want to treat the data and bring the variable on the same scale because of this, scaling is necessary in this case and now we will perform scaling. So, now we will apply zscore transformation on the data it will give us a mean of 0 and standard deviation of 1 which will be applicable for all the column's .This would ensure that our data comes to standard normal form also known as standardization.

Now, we are applying Zscore Scaling:-

Apps	*Accept*	*Enroll*	*Top10perc*	*Top25perc*	*F.Undergrad*	*P.Undergrad*	*Outstate*
-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356
-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496
-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.49709	0.201305
-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633
-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005	-0.716508
Room Board	*Books*	*Personal*	*PhD*	*Terminal*	*S.F.Ratio*	*perc.alumni*	*Expend*
-0.964905	-0.602312	1.270045	-0.163028	-0.115729	1.013776	-0.867574	-0.50191
1.909208	1.21588	0.235515	-2.675646	-3.378176	-0.477704	-0.544572	0.16611
-0.554317	-0.905344	-0.259582	-1.204845	-0.931341	-0.300749	0.585935	-0.17729
0.996791	-0.602312	-0.688173	1.185206	1.175657	-1.615274	1.151188	1.792851
-0.216723	1.518912	0.235515	0.204672	-0.523535	-0.553542	-1.675079	0.241803

Grad.Rate
-0.318252
-0.551262
-0.667767
-0.376504
-2.939613

$$Z = \frac{x - \mu}{\sigma}$$

Z score tells us how many standard deviation is the point away from the mean and also the direction. Now, we can understand that all the variables are scaled by using zscore function. Scaling is one of the most important method to follow before implementing models.

Q.3 - COMMENT ON THE COMPARISON BETWEEN THE COVARIANCE AND THE CORRELATION MATRICES FROM THIS DATA. [ON SCALED DATA]

We took out the covariance matrix:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate
Apps	1.001289	0.944666	0.847913	0.33927	0.352093	0.81554	0.398777	0.050224
Accept	0.944666	1.001289	0.912811	0.192695	0.247795	0.87535	0.441839	-0.025788
Enroll	0.847913	0.912811	1.001289	0.181527	0.227037	0.965883	0.51373	-0.155678
Top10perc	0.33927	0.192695	0.181527	1.001289	0.893144	0.141471	-0.105492	0.563055
Top25perc	0.352093	0.247795	0.227037	0.893144	1.001289	0.199702	-0.053646	0.490024
F.Undergrad	0.81554	0.87535	0.965883	0.141471	0.199702	1.001289	0.571247	-0.21602
P.Undergrad	0.398777	0.441839	0.51373	-0.105492	-0.053646	0.571247	1.001289	-0.253839
Outstate	0.050224	-0.025788	-0.155678	0.563055	0.490024	-0.21602	-0.253839	1.001289
Room.Board	0.165152	0.091016	-0.040284	0.371959	0.331917	-0.068979	-0.061405	0.6551
Books	0.132729	0.113672	0.112856	0.119012	0.115676	0.115699	0.081304	0.038905
Personal	0.178961	0.201248	0.281291	-0.093437	-0.080914	0.317608	0.320294	-0.299472
PhD	0.391201	0.356216	0.331896	0.532513	0.546566	0.318747	0.149306	0.383476
Terminal	0.369968	0.338018	0.308671	0.491768	0.525425	0.300406	0.142086	0.408509
S.F.Ratio	0.095756	0.176456	0.237577	-0.38537	-0.295009	0.280064	0.23283	-0.555536
perc.alumni	-0.090342	-0.160196	-0.181027	0.456072	0.418403	-0.229758	-0.281154	0.566992
Expend	0.259927	0.124878	0.064252	0.661765	0.528127	0.018676	-0.083676	0.673646
Grad.Rate	0.146944	0.067399	-0.02237	0.495627	0.477896	-0.078875	-0.257332	0.572026

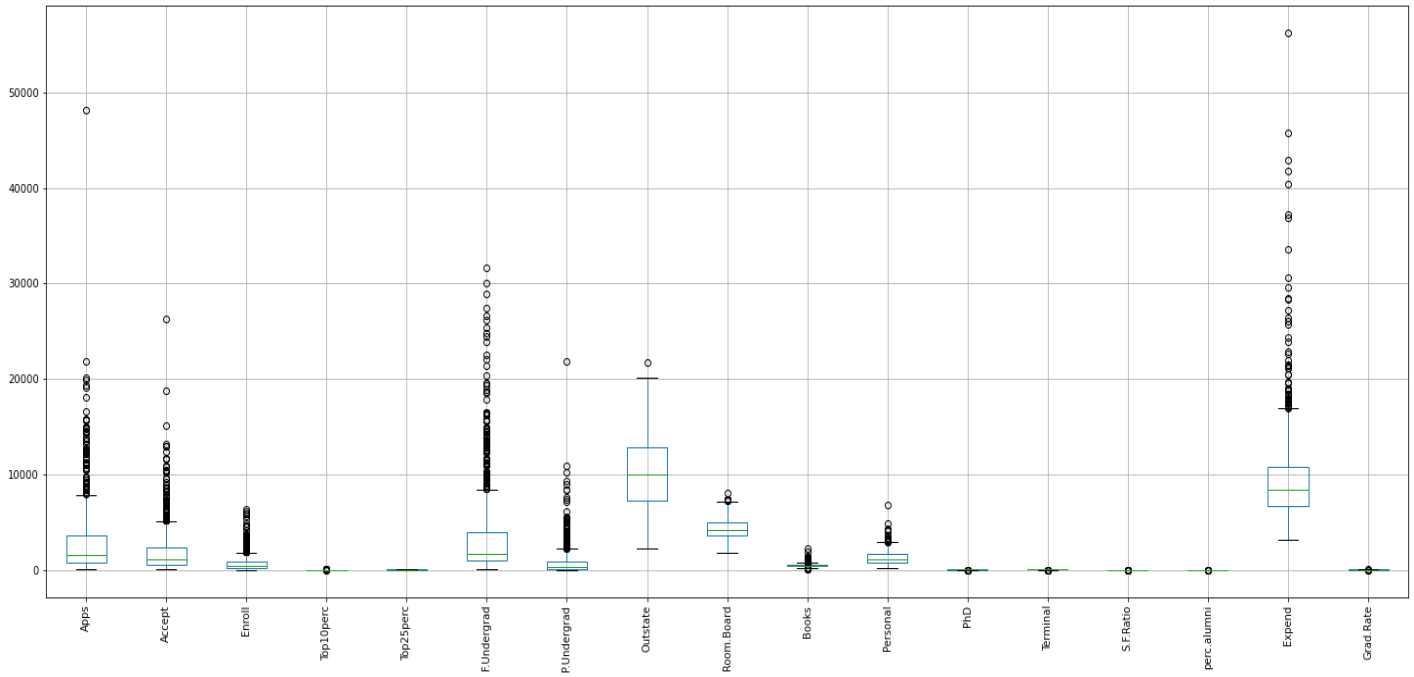
Now, we take out the correlation matrix:-

Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board
1	0.943451	0.846822	0.338834	0.35164	0.814491	0.398264	0.050159	0.164939
0.943451	1	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899
0.846822	0.911637	1	0.181294	0.226745	0.96464	0.513069	-0.155477	-0.040232
0.338834	0.192447	0.181294	1	0.891995	0.141289	-0.105356	0.562331	0.37148
0.35164	0.247476	0.226745	0.891995	1	0.199445	-0.053577	0.489394	0.33149
0.814491	0.874223	0.96464	0.141289	0.199445	1	0.570512	-0.215742	-0.06889
0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	1	-0.253512	-0.061326
0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	-0.253512	1	0.654256
0.164939	0.090899	-0.040232	0.37148	0.33149	-0.06889	-0.061326	0.654256	1
0.132559	0.113525	0.112711	0.118858	0.115527	0.11555	0.0812	0.038855	0.127963
0.178731	0.200989	0.280929	-0.093316	-0.08081	0.3172	0.319882	-0.299087	-0.199428
0.390697	0.355758	0.331469	0.531828	0.545862	0.318337	0.149114	0.382982	0.329202
0.369491	0.337583	0.308274	0.491135	0.524749	0.300019	0.141904	0.407983	0.37454
0.095633	0.176229	0.237271	-0.384875	-0.294629	0.279703	0.232531	-0.554821	-0.362628
-0.090226	-0.15999	-0.180794	0.455485	0.417864	-0.229462	-0.280792	0.566262	0.272363
0.259592	0.124717	0.064169	0.660913	0.527447	0.018652	-0.083568	0.672779	0.501739
0.146755	0.067313	-0.022341	0.494989	0.477281	-0.078773	-0.257001	0.57129	0.424942

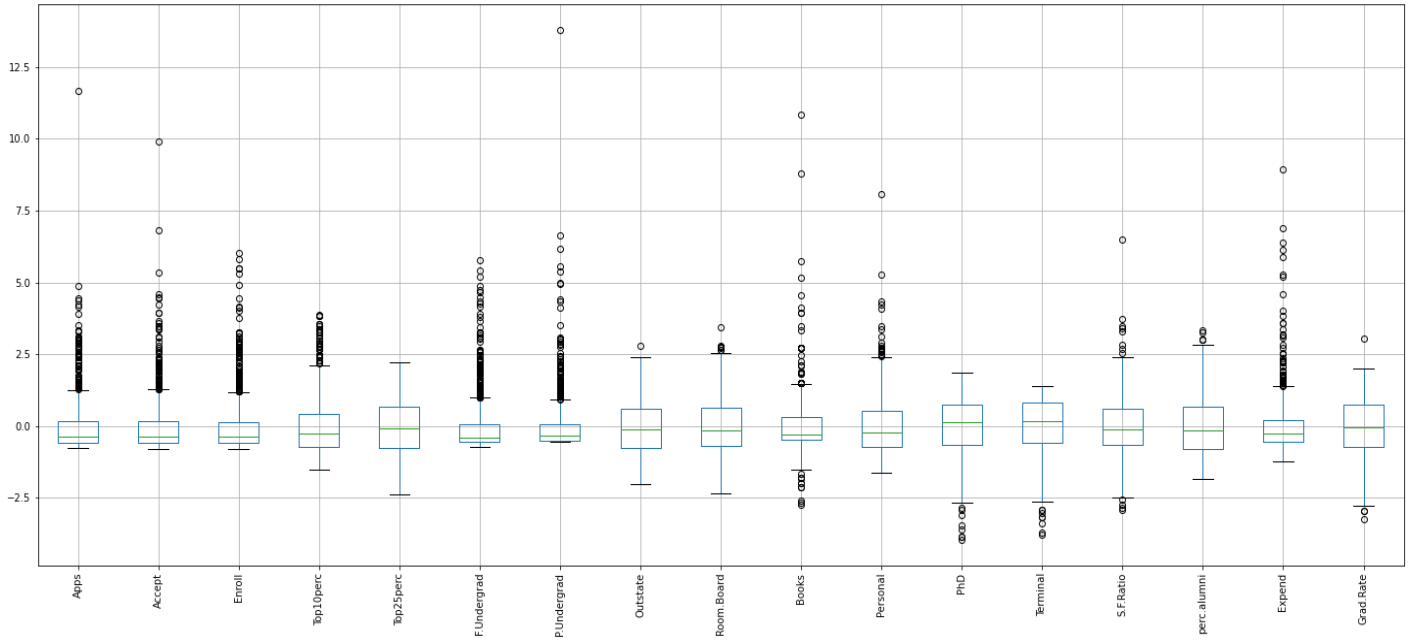
So, starting with covariance we know that it is used to know what is the relationship between 2 variables in a way it calculates directional relationship. When we calculate covariance the value which we get will either be positive value, negative value or can be zero also. So, from these values only we can make out if they are moving in the same direction or moving in opposite direction. So, covariance help's us to know the directions of the variable. But, we also want to know the strength of the relation between variables. So, for that we will use correlation i.e if there is strong relationship between variable's and Moderate relationship or Weak relationship and it also let us know the direction as well. We cannot know strength of a variable's until we know in which context it is used because the value of covariance that will come will be either positive or negative there will be no effect of that value until we measure it on a scale. So, when we calculate correlation all the values of covariance will come to us in the scale of -1 to 1 in covariance we can get value to any extend but in correlation we will definitely get a value between or as -1 to 1. So from these two things only we can get to know if the relationship is negative or positive. For Negative relationship one will increase and the other one will decrease but in positive relationship if one is increasing then other one will also increase with it. So, how to know if the relationship is weak, moderate or strong. We can judge that by observing the value in the data if it's -1 then it's a Perfect Negative correlation, if it's 1 then perfect positive correlation and if the value is 0.8 or above then it is strong positive correlation and if the value is lying between -0.8 or -1 then it is strong negative correlation and if the value is lying between 0.5 to less than 0.8 then it's a moderate relationship this same can be applied to negative moderate correlation as well i.e. -0.8 to -0.5 and for weak correlation its 0 to less than 0.5 same goes for negative weak correlation i.e. 0 to -0.5 and if the value is 0 then there is no correlation.

Q.4 Check the dataset for outliers before and after scaling. What insight do you derive here?

Before Scaling:-



After Scaling:-



Inference:
The outliers are still present in dataset.

Reason:
Scaling does not remove outliers scaling scales the values on a Z score distribution.
We can use any one method to remove outliers for further processes.
For example if we wish to remove outliers we can consider taking 3 standard deviations as outliers or either we can remove them or impute them with IQR values.

Q.5 - Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both]

Eigenvectors:-

```
(([[ 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]])
```

Eigenvalues:-

```
([5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123,
       0.84849117, 0.6057878 , 0.58787222, 0.53061262, 0.4043029 ,
       0.31344588, 0.22061096, 0.16779415, 0.1439785, 0.08802464,
       0.03672545, 0.02302787])
```

Explained Variance After sorting the eigenpairs, the next question is "how many principal compone nts are we going to choose for our new feature subspace?" A useful measure is the so-called "expl ained variance," which can be calculated from the eigenvalues. The explained variance tells us ho w much information (variance) can be attributed to each of the principal components. Below, we c an see that in scree plot.

Explained variance ratio:-

```
([0.32020628, 0.26340214, 0.06900917, 0.05922989, 0.05488405,
       0.04984701, 0.03558871, 0.03453621, 0.03117234, 0.02375192,
       0.01841426, 0.01296041, 0.00985754, 0.00845842, 0.00517126,
       0.00215754, 0.00135284])
```

The explained variance ratio is the percentage of variance that is attributed by each of the selected components. Ideally, you would choose the number of components t o include in your model by adding the explained variance ratio of each component u ntil you reach a total of around 0.8 or 80% to avoid over fitting.

Q.6 - Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features

Eigenvectors DataFrame with original feature's:-

Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board
0.248766	0.207602	0.176304	0.354274	0.344001	0.154641	0.026443	0.294736	0.24903
0.331598	0.372117	0.403724	-0.082412	-0.044779	0.417674	0.315088	-0.249644	-0.137809
-0.063092	-0.101249	-0.082986	0.035056	-0.024148	-0.061393	0.139682	0.046599	0.148967
0.281311	0.267817	0.161827	-0.051547	-0.109767	0.100412	-0.158558	0.131291	0.184996
0.005741	0.055786	-0.055694	-0.395434	-0.426534	-0.043454	0.302385	0.222532	0.560919
-0.016237	0.007535	-0.042558	-0.052693	0.033092	-0.043454	-0.191199	-0.03	0.162755
-0.042486	-0.01295	-0.027693	-0.161332	-0.118486	-0.025076	0.061042	0.108529	0.209744
-0.10309	-0.056271	0.058662	-0.122678	-0.102492	0.07889	0.570784	0.009846	-0.221453
-0.090227	-0.177865	-0.128561	0.3411	0.403712	-0.059442	0.560673	-0.004573	0.275023
0.05251	0.04114	0.034488	0.064026	0.014549	0.020847	-0.223106	0.186675	0.298324
0.043046	-0.058406	-0.069399	-0.008105	-0.273128	-0.081158	0.100693	0.143221	-0.359322
0.024071	-0.145102	0.011143	0.038554	-0.089352	0.056177	-0.063536	-0.823444	0.35456
0.595831	0.292642	-0.444638	0.001023	0.021884	-0.523622	0.125998	-0.141856	-0.069749
0.080633	0.033467	-0.085697	-0.107828	0.151742	-0.056373	0.019286	-0.034012	-0.058429
0.133406	-0.145498	0.02959	0.697723	-0.617275	0.009916	0.020952	0.038354	0.003402
0.459139	-0.518569	-0.404318	-0.148739	0.051868	0.560363	-0.052731	0.101595	-0.025929
0.35897	-0.543427	0.609651	-0.144986	0.080348	-0.414705	0.009018	0.0509	0.001146

Q.7 - Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only).

In generic first PC can be represented using linear combination of features and its coefficients/weights:

PC1=A1 * X1+A2 * X2+A3 * X3+A4 * X4..... where X1,X2,X3,X4... are original variables/features before transformation.

In this scenario PC1 can be represented as linear combination of below components [A1 A2 A3 A4....A17]=

[0.248766,0.207602, 0.176304, 0.354274, 0.344001, 0.154641, 0.026443, 0.294736, 0.249030, 0.064758, -0.042529, 0.318313, 0.317056, -0.176958, 0.205082, 0.318909, 0.252316]

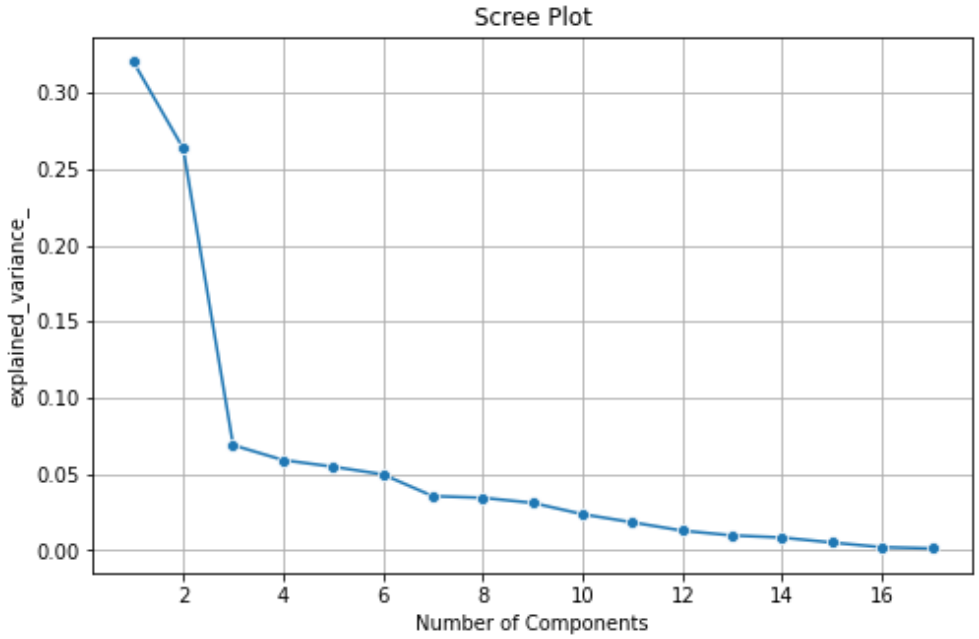
[X1 X2 X3.....X17]=[-0.35,-0.32,-0.06,-0.26,-0.19,-0.17,-0.21,-0.75,-0.96,-0.60,-0.27,-0.16,-0.12,1.01,-0.87,-0.50,-0.32] of scaled data.

Q.8 - Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?

Created a dataframe containing the loadings or coefficients of all PCs:-

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	-0.042486	-0.10309
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	-0.01295	-0.056271
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	-0.027693	0.058662
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	-0.161332	-0.122678
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	-0.118486	-0.102492
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454	-0.025076	0.07889
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199	0.061042	0.570784
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532	-0.03	0.108529	0.009846
Room.Board	0.24903	-0.137809	0.148967	0.184996	0.560919	0.162755	0.209744	-0.221453
Books	0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055	-0.149692	0.213293
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398	0.63379	-0.232661
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256	-0.001096	-0.07704
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.20472	0.154928	-0.028477	-0.012161
S.F.Ratio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046	0.219259	-0.083605
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.04734	0.243321	0.678524
Expend	0.318909	-0.13169	0.226744	0.079273	0.075958	-0.298119	-0.226584	-0.054159
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163	0.559944	-0.005336

Creating a Scree Plot:-



Cumulative Values of the eigenvalues:-

```
([0.32020628, 0.58360843, 0.65261759, 0.71184748, 0.76673154,
       0.81657854, 0.85216726, 0.88670347, 0.91787581, 0.94162773,
       0.96004199, 0.9730024 , 0.98285994, 0.99131837, 0.99648962,
       0.99864716, 1.        ])
```

In order to decide which eigenvector's can dropped without losing too much information for the construction of lower-dimensional subspace, we need to inspect the corresponding eigenvalues: The eigenvectors with the lowest eigenvalues bear the least information about the distribution of the data; those are the ones can be dropped. Below, with the help of scree plot we can observe which eigenvector contain's how much information and acoording to it we will drop the other eigenvector's which doesn't contain's much in formation.

To decide the optimum number of principal components:-

1. Check for cumulative variance up to 90%, check the corresponding associated with 90%
2. The incremental value between the components should not be less than five percent. 18 So basis on this we can decide the optimum number of principal components as 6, because after this the incremental value between the is less than 5%. So, we select 5 principal components for this case study.

The first components explain 32.02% variance in data

The second components explains 58.36% variance in data

The third components explains 65.26% variance in data

The fourth components explains 71.18% variance in data

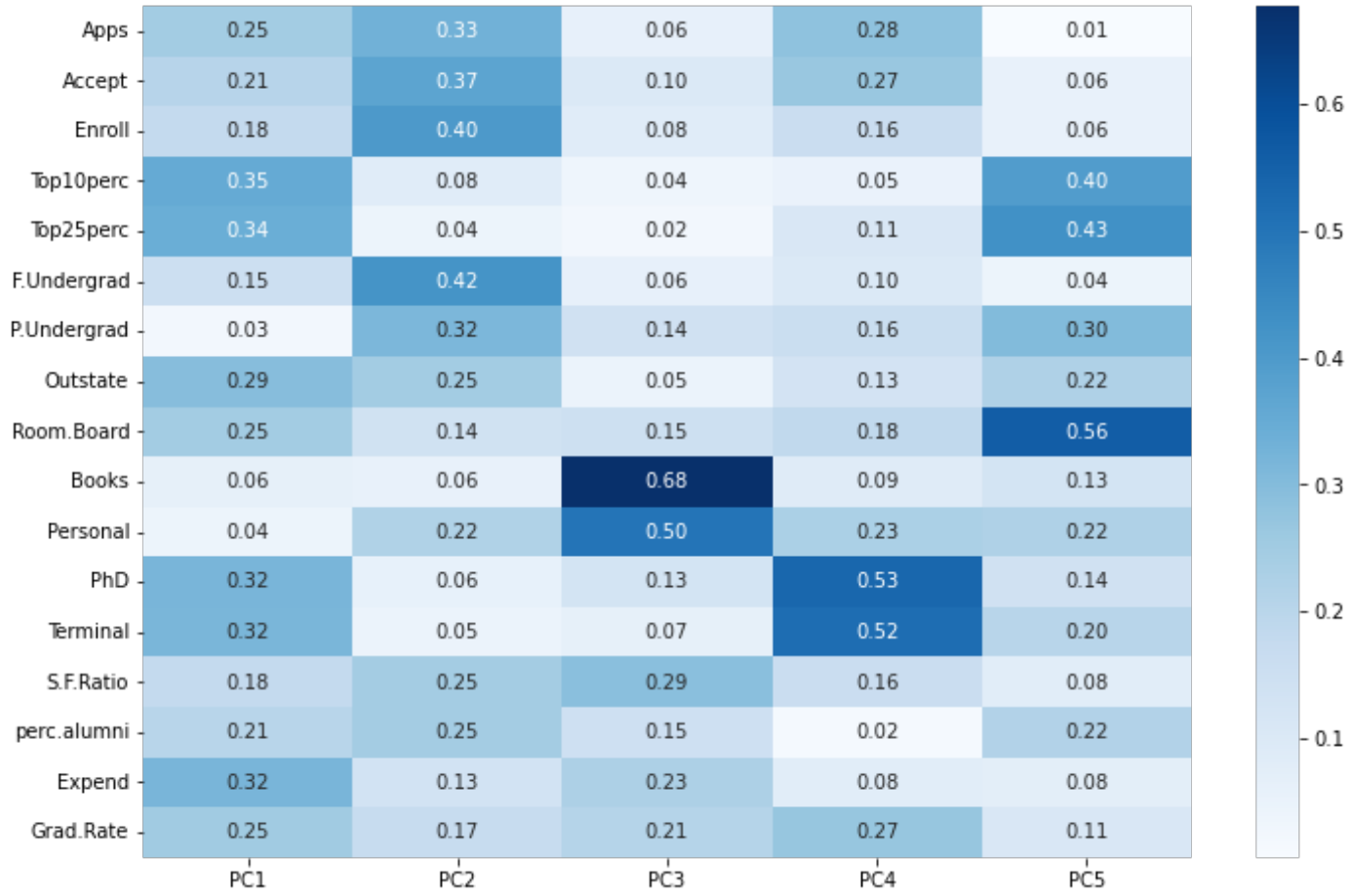
The fifth components explains 76.67% variance in data

The Eigen vectors or PC for this case study is five, we can understand how much each variable contributes to the principal components. In other words we can also say weights attached to each variable. With this Eigen vectors we can understand which variable has more weightage and influences the dataset in the principal components. The PCA reduces the multi collinearity and with this reduced collineraity we can runs models and improved efficiency scores. PCA

Chose PCs based on Cumulative explained variance:-

	PC1	PC2	PC3	PC4	PC5
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919
Books	0.064758	0.056342	0.677412	0.087089	-0.127289
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720
S.F.Ratio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268

PCA is performed and after PCA the multi collinearity is highly reduced.



Q.9 - Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis?

This business case study is about education dataset which contain the names of various colleges, which has various details of colleges and university. To understand more about the dataset we perform univariate analysis and multivariate analysis which gives us the understanding about the variables. From analysis we can understand the distribution of the dataset, skew, and patterns in the dataset. From multivariate analysis we can understand the correlation of variables. Inference of multivariate analysis shows we can understand multiple variables highly correlated with each other. The scaling helps the dataset to standardize the variable in one

scale.The principal component analysis is used reduce the multicollinearity between the variables. Depending on the variance of the dataset we can reduce the PCA components. The PCA components for this business case is 6 where we could understand the maximum variance of the dataset. Using the components we can now understand the reduced multicollinearity in the dataset