greatlearning
Power Ahead

Advanced Statistics Project

Made by:
Anmol Tripathi
Data Science and Business Analytics (PGP-DSBA)
Online September_A 2021-22

greatlearning
Power Ahead

Table of Contents:

Problem 1:

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Problem 2:

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

List of Graphs:

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

List of Formulas:

List of Tables:

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Problem 1:

I. Executive Summary:

The given data set has number columns and rows in them. The given data provides us the information with regards to various people's education, their current occupation and their salary details. ANOVA is a technique which belongs to the domain called "Experimental Designs". It helps in establishing in an exact way, the Cause- Effect relation between variables. From the statistical inference point of view, ANOVA is an extension of independent t test for testing the equality of two population means.

II. Background of the problem:

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [SalaryData.csv] are collected and each person's educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination.

III. Data Description:

The shape of the data set seems to be with 40 rows and 3 columns.

Two columns are object (string / categorical) in nature and Salary column is integer.

There are no duplicate values present in the data set.

Also, the entire data set does not have any null or missing values.

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

Table 1.1 No null Values

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Education   40 non-null     object
 1   Occupation  40 non-null     object
 2   Salary      40 non-null     int64
dtypes: int64(1), object(2)
memory usage: 1.1+ KB
```

Table 1.2 Data type

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

IV. Sample Data Set:

	Education	Occupation	Salary
0	Doctorate	Adm-clerical	153197
1	Doctorate	Adm-clerical	115945
2	Doctorate	Adm-clerical	175935
3	Doctorate	Adm-clerical	220754
4	Doctorate	Sales	170769

Table 1.3 Data type

The above table represents the actual representation of the data set.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

1.1 State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.

One way ANOVA(Education):

Null Hypothesis *H*0: The mean salary is the same across all the 3 categories of education (Doctorate, Bachelors, HS-Grad).

Alternate Hypothesis *H*1: The mean salary is different in at least one category of education.

One way ANOVA(Occupation):

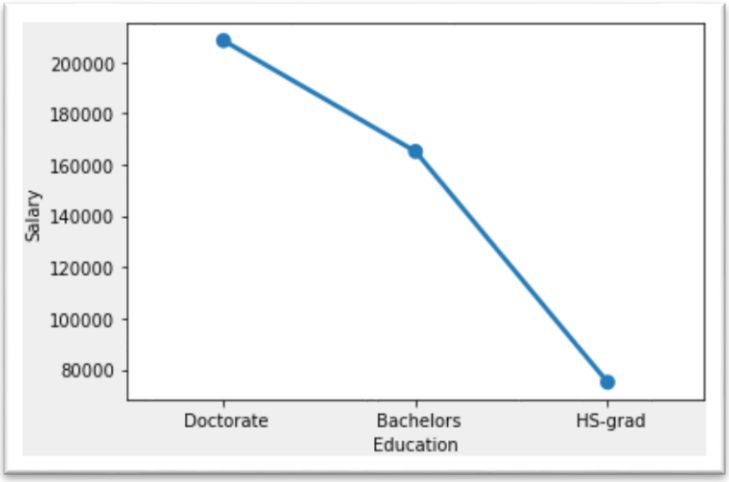
Null Hypothesis *H*0: The mean salary is the same across all the 4 categories of occupation (Prof-Specialty, Sales, Adm-clerical, Exec-Managerial).

Alternate Hypothesis *H*1: The mean salary is different in at least one category of occupation.

1.2 Perform one-way ANOVA for Education with respect to the variable 'Salary'. 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

Table 1.4 One Way ANOVA Table



Graph 1.1 Interaction of Education with salary

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

The above is the ANOVA table for Education variable.

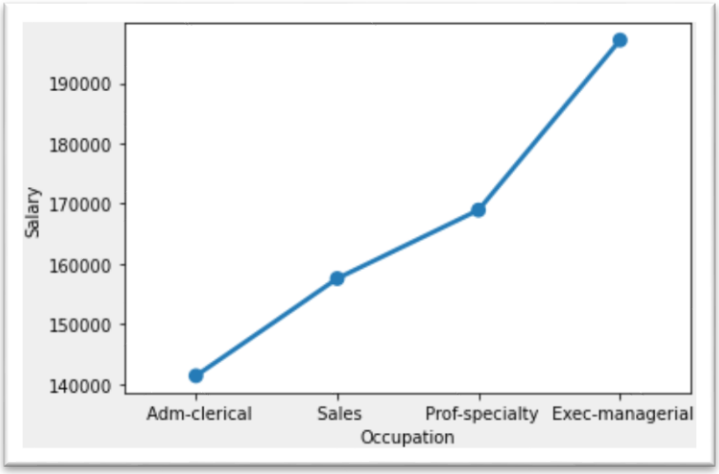
Since the p value = 1.257709e-08 is less than the significance level (alpha = 0.05), we can reject the null hypothesis and conclude that there is a significant difference in the mean salaries for at least one category of education.

1.3 Perform one-way ANOVA for variable Occupation with respect to the variable 'Salary'. 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

Table 1.5 One Way ANOVA Table

The above is the ANOVA table for Occupation variable.



Graph 1.2 Interaction of Occupation with salary

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.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

1.4 If the null hypothesis is rejected in either (1.2) or in (1.3), find out which class means are significantly different. Interpret the result.

To find out which class means are significantly different, the Tukey Honest Significant Difference test is performed.
Using, the Tukey Honest Significant Difference test, we get the following table for the category education:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Bachelors	Doctorate	43274.0667	0.0146	7541.1439	79006.9894	True
Bachelors	HS-grad	-90114.1556	0.001	-132035.1958	-48193.1153	True
Doctorate	HS-grad	-133388.2222	0.001	-174815.0876	-91961.3569	True

Table 1.6 Turkey Honest Significant Difference test table

The table shows that since the p- values (p-adj in the table) are lesser than the significance level for all the three categories of education, this implies that the mean salaries across all categories of education are different.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Adm-clerical	Exec-managerial	55693.3	0.4146	-40415.1459	151801.7459	False
Adm-clerical	Prof-specialty	27528.8538	0.7252	-46277.4011	101335.1088	False
Adm-clerical	Sales	16180.1167	0.9	-58951.3115	91311.5449	False
Exec-managerial	Prof-specialty	-28164.4462	0.8263	-120502.4542	64173.5618	False
Exec-managerial	Sales	-39513.1833	0.6507	-132913.8041	53887.4374	False
Prof-specialty	Sales	-11348.7372	0.9	-81592.6398	58895.1655	False

Table 1.7 Turkey Honest Significant Difference test table

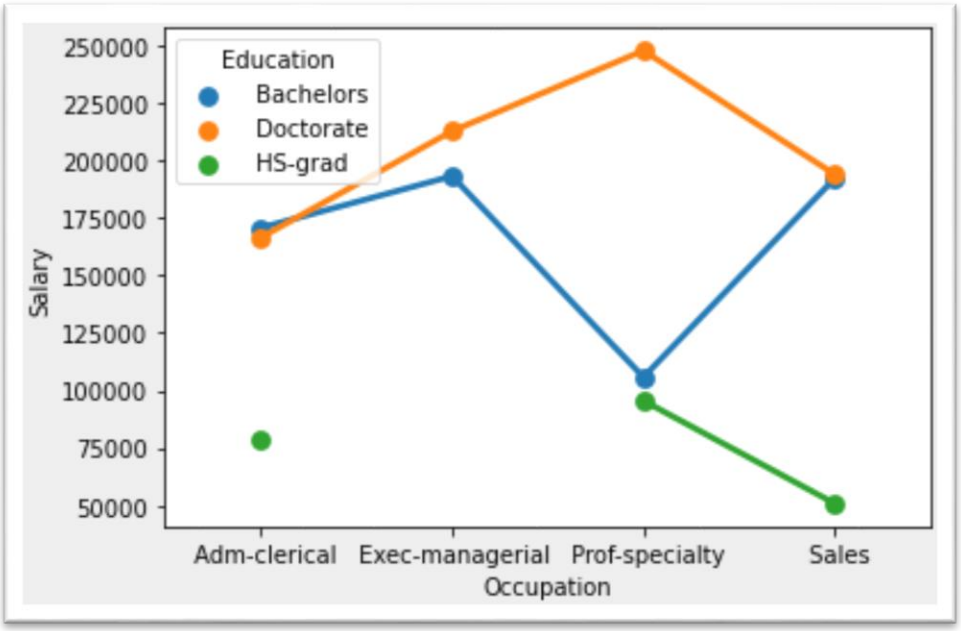
For the category occupation, the Tukey Honest Significant Difference test has further confirmed that the mean salaries across all occupation classes are significantly same. The table below confirms the same, wherein we see that all p-values are greater than 0.05.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

1.5 What is the interaction between the two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.

We 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.



Graph 1.3 Interaction of Occupation and education with salary

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

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

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.

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

*H*0: 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).
*H*1: There is an interaction effect between the independent variable 'education' and the independent variable 'occupation' on the mean Salary.

By performing two way ANOVA, we get the following table:

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
C(Occupation)	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
C(Education):C(Occupation)	6.0	3.634909e+10	6.058182e+09	8.519815	2.232500e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

Table 1.8 Two Way ANOVA Table

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

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.

1.7 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.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Problem 2:

I. Executive Summary:

The given data set consist of data points of names of various university and college which has number of applications received, accepted and enrolled. It also has percentage of new students from top 10% of higher secondary class, percentage of new students from top 25% of higher secondary class. The data set also has the list of undergrad students (full time / part time), Number of students for which particular college is out of state tuition, cost of room, and board. It also has estimate expenditure details for books and personal spending. It also has percentage of PHD faculties, percentage of faculties with terminal degree, student/faculty ratio, percentage of alumni who donate, graduation rate etc.

II. Background of the Problem:

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

III. Data Description:

The shape of the data set seems to be with 777 rows and 18 columns.

All the columns seem to be integer or float values.

The name of the column is a categorical value whereas Student/faculty ratio is float datatype.

There are no duplicate values present in the data set.

Also, the entire data set does not have any null or missing values.

Names	object
Apps	int64
Accept	int64
Enroll	int64
Top10perc	int64
Top25perc	int64
F.Undergrad	int64
P.Undergrad	int64
Outstate	int64
Room.Board	int64
Books	int64
Personal	int64
PhD	int64
Terminal	int64
S.F.Ratio	float64
perc.alumni	int64
Expend	int64
Grad.Rate	int64
dtype: object	

Table 2.1. Data Type

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

From the above data set, we can see what all inferences I have made above.

```
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
```

Table 2.2. No null values

From the above data set, we can see what all inferences I have made above.

IV. Sample Dataset:

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.a
0	Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	
2	Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	
3	Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	
4	Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	

Table 2.3. Data Type

The above table represents the actual representation of the data set. Few columns of the data set are not visible in the above diagram.

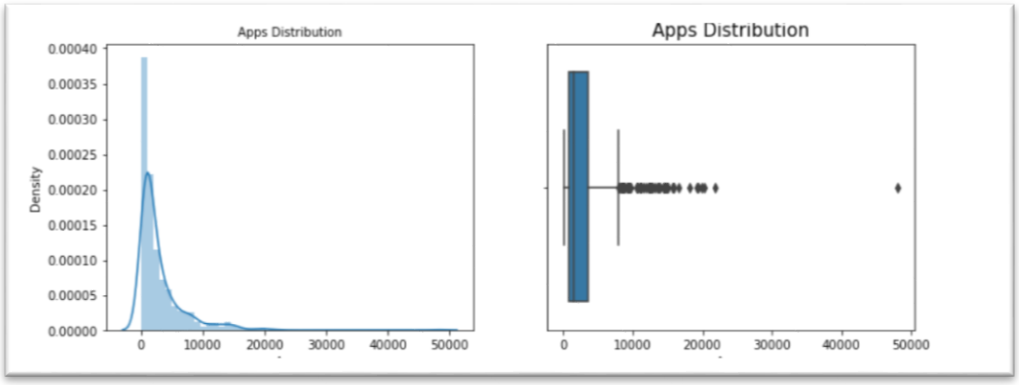
Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

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

Univariate Analysis:

It helps us to understand the data distribution of the data in the dataset. With the help of Univariate analysis, we can find patterns and we can easily summarize the data.

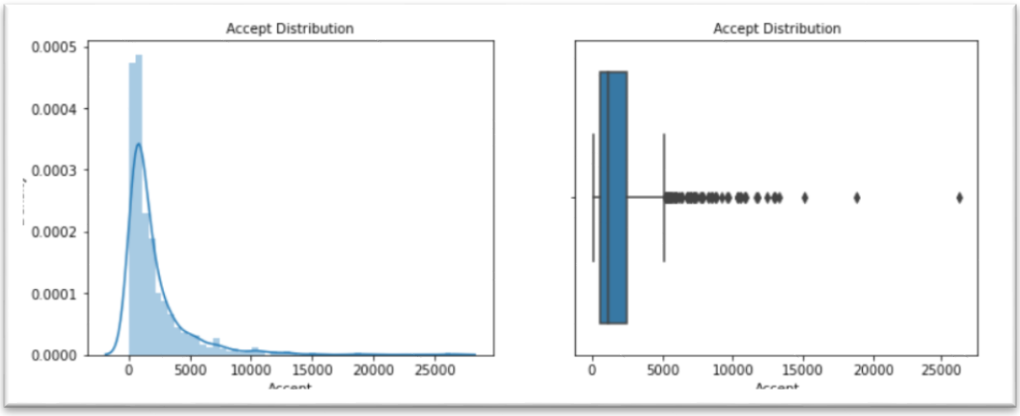
APPS:



Graph 2.1 EDA graph

The Box plot of the Apps variable seems to have the outliers, distribution of the data is skewed. We can also understand that each college or university offers application in the range 3000 to 5000. The max applications seem to be around 50,000. For univariate analysis of apps, we are using box plot and dist plot to find the information or patterns in the data. So we can clearly understand from the box plot we have outliers in the dataset.

ACCEPT:



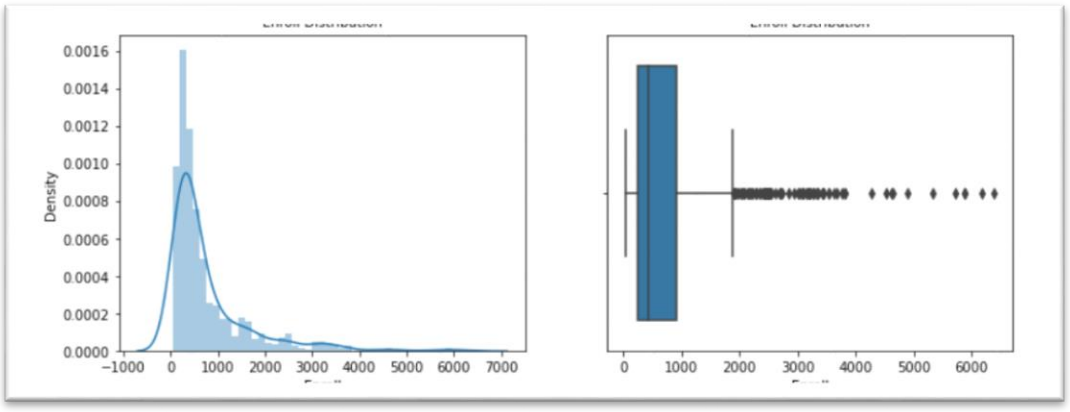
Graph 2.2 EDA graph

The accept variable seems to have outliers. The dis plot shows us the majority of applications

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

accepted from each university are in the range from 70 to 1500. The accept variable seems to be positively skewed.

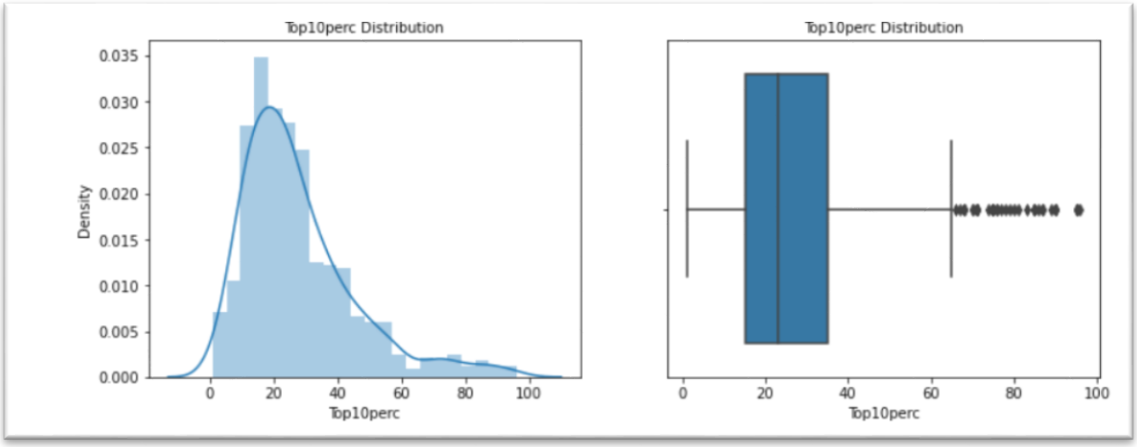
ENROLL:



Graph 2.3 EDA graph

The box plot of the Enroll variable also has outliers. The distribution of the data is positively skewed. From the dist plot, we can understand majority of the colleges have the enrolled students in the range of 200 to 500 students.

TOP 10 PERCENT:

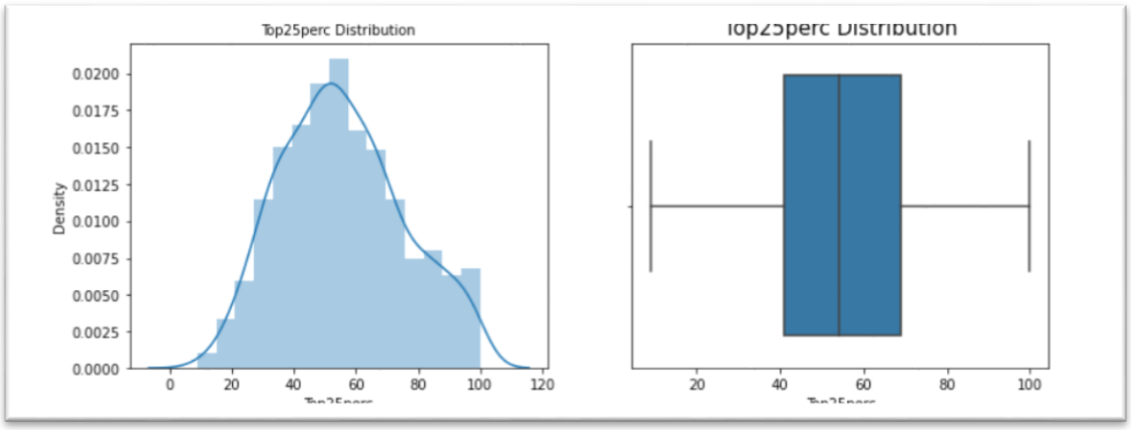


Graph 2.4 EDA graph

The Box plot of students from Top 10 percentage of higher secondary class seem to have outliers. The distribution seems to have positively skewed. There is good amount of intake about 30 to 50 students from top 10 percentage of the higher secondary class.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

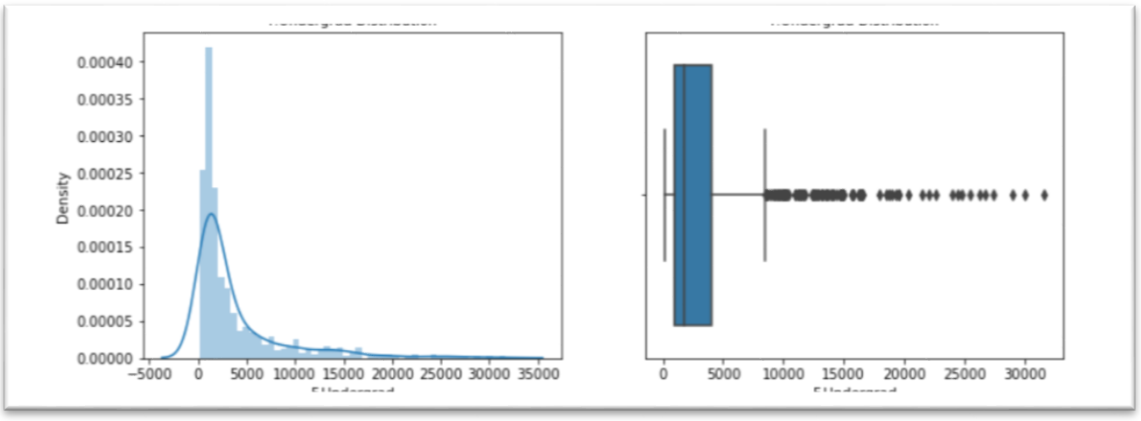
TOP 25 PERCENT:



Graph 2.5 EDA graph

The box plot of the full time graduates has outliers. The distribution of the data is positively skewed. In the range about 300 to 5000, they are full time graduates studying in all the university.

FULL TIME UNDERGRADUATE:

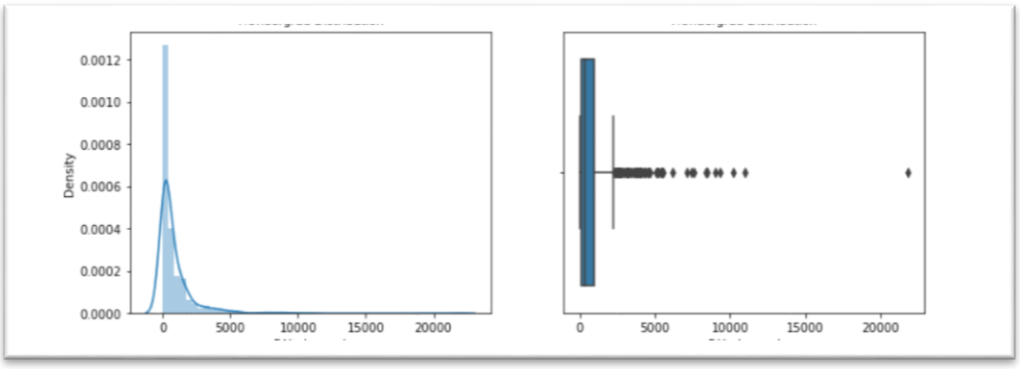


Graph 2.6 EDA graph

The box plot of the full-time graduates has outliers. The distribution of data is positively skewed. In range of about 3000 to 5000, they are full time graduates studying in all the university.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

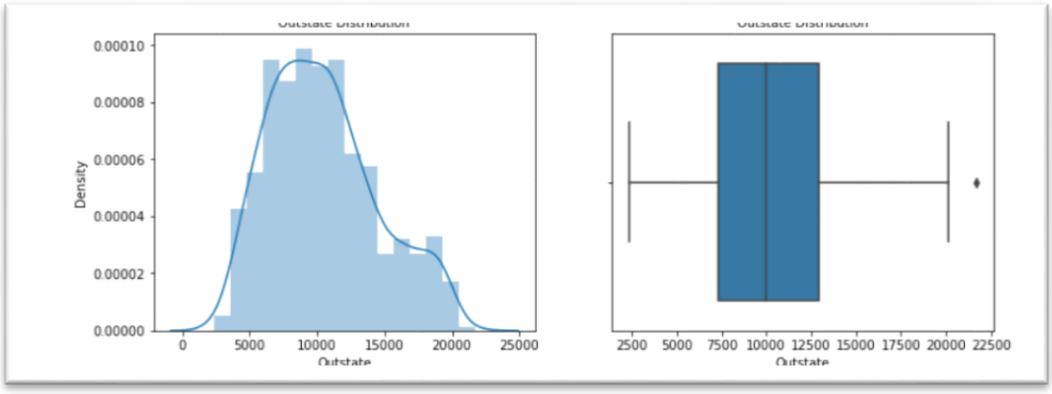
PART TIME UNDERGRADUATE:



Graph 2.7 EDA graph

The Box plot of the part time graduates has outliers. The distribution of the data is positively skewed. In the range about 1000 to 3000, they are part-time graduates studying in all the university.

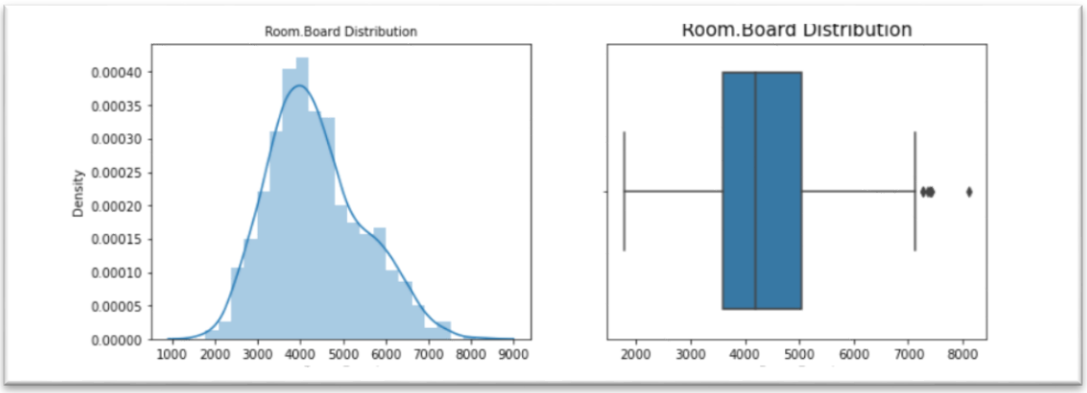
OUTSTATE:



Graph 2.8 EDA graph

The Box plot of outstate has only one outlier. The distribution is almost normally distributed.

ROOM BOARDS:

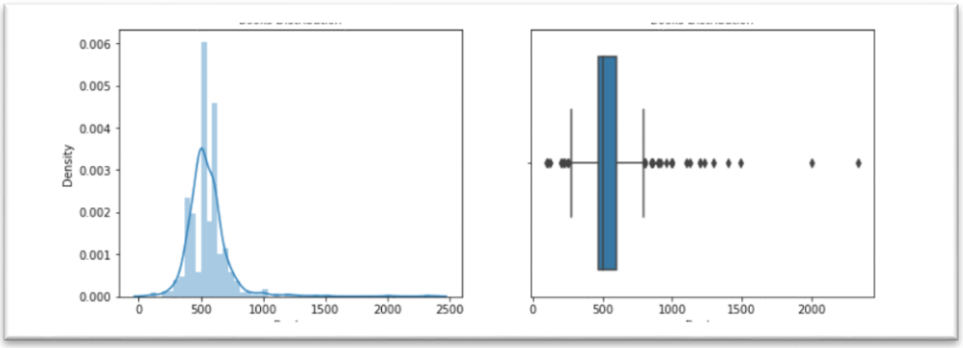


Graph 2.9 EDA graph

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

The room boards has few outliers. The distribution is normally distributed.

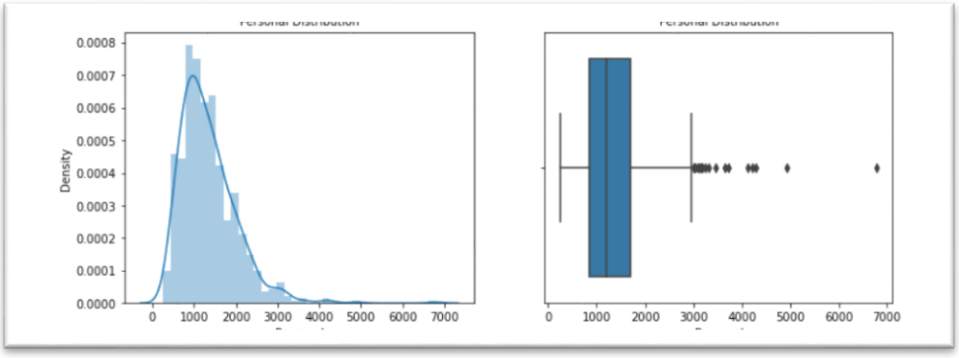
BOOKS:



Graph 2.10 EDA graph

The box plot of books has outliers. The distribution seems to be binomial. The cost of books per student seems to be in range of 500 to 100.

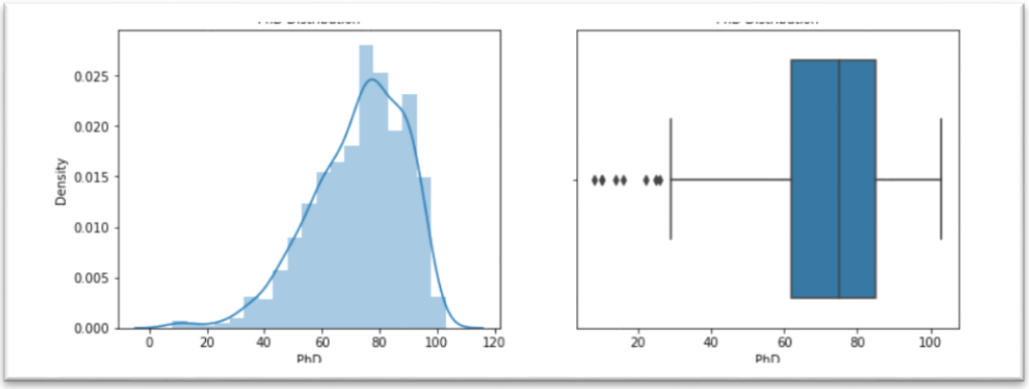
PERSONAL:



Graph 2.11 EDA graph

The Box Plot of personal expense has outliers. Some student's personal expense are way bigger that the rest of the rest of the students. The distribution seems to be positively skewed.

PHD:

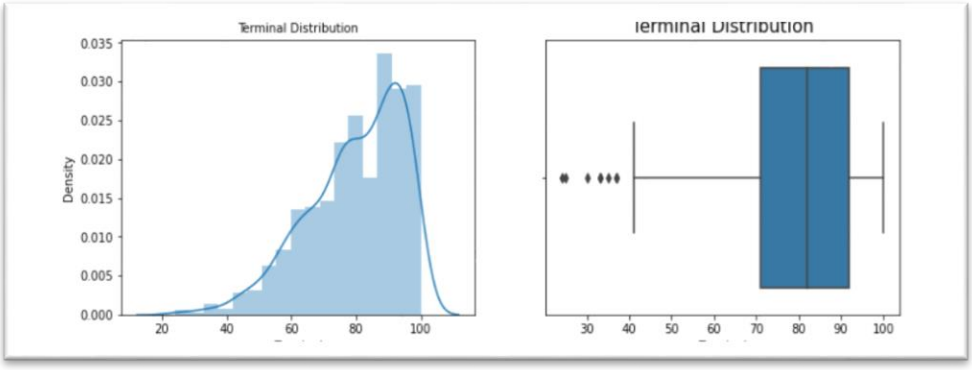


Graph 2.12 EDA graph

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

The box plot of PHD has outliers. The distribution seems to be skewed.

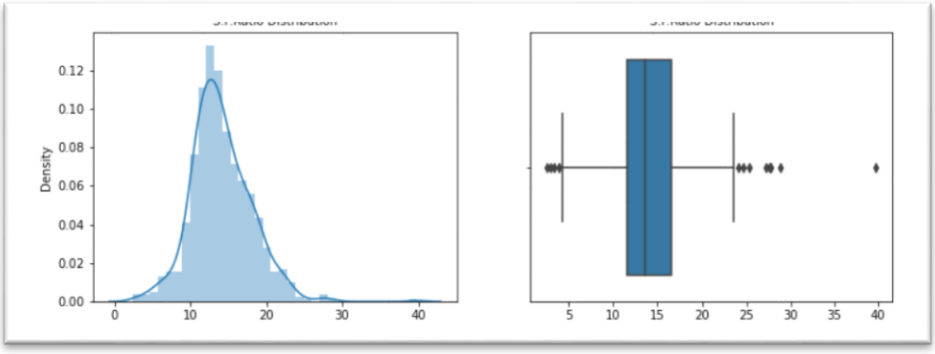
TERMINAL:



Graph 2.13 EDA graph

The Box plot of terminal seems to have outliers in the dataset. The distribution for the terminal also seems to be negatively skewed.

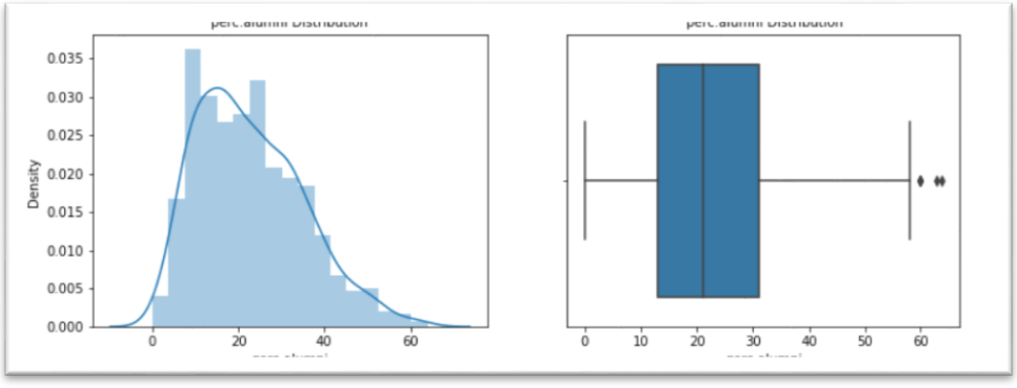
SF RATIO:



Graph 2.14 EDA graph

The SF Ratio variable also has outliers in the data set. The distribution is almost normally distributed. The student faculty ratio is almost same in all the university and colleges.

PERCENTAGE ALUMINI:

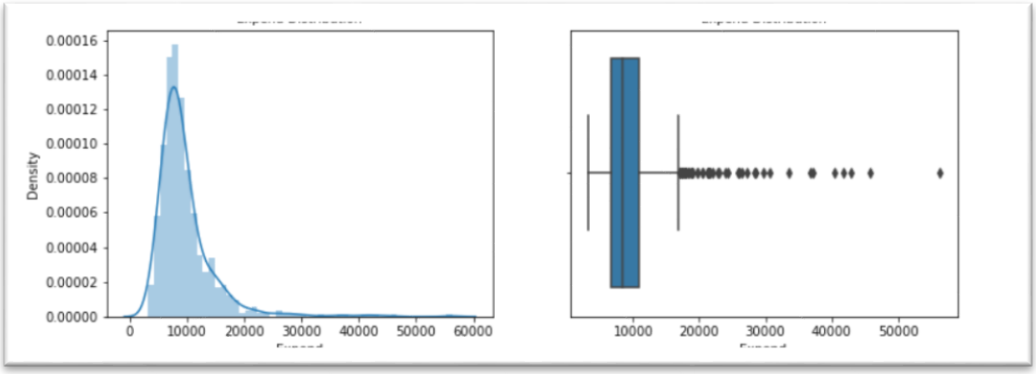


Graph 2.15 EDA graph

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

The percentage of alumni box plot seems to have outliers in the data set. The distribution is almost normally distributed.

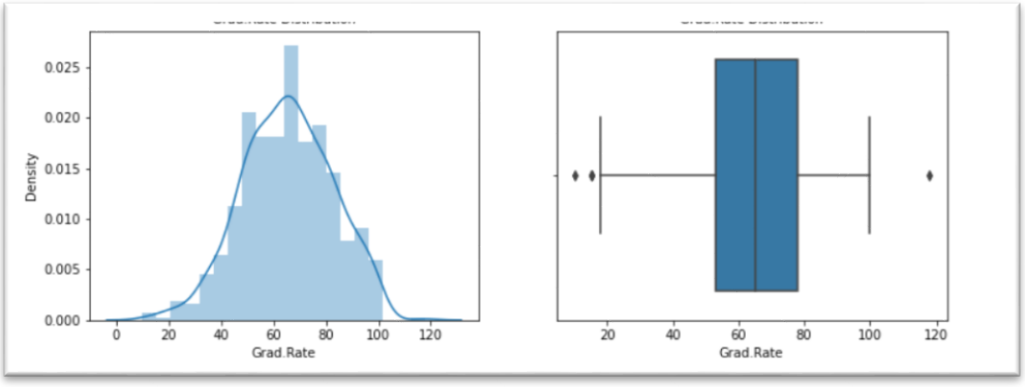
EXPENDITURE:



Graph 2.16 EDA graph

The expenditure variable also has outliers in the data set. The distribution of the expenditure is positively skewed.

GRADUATION RATIO:

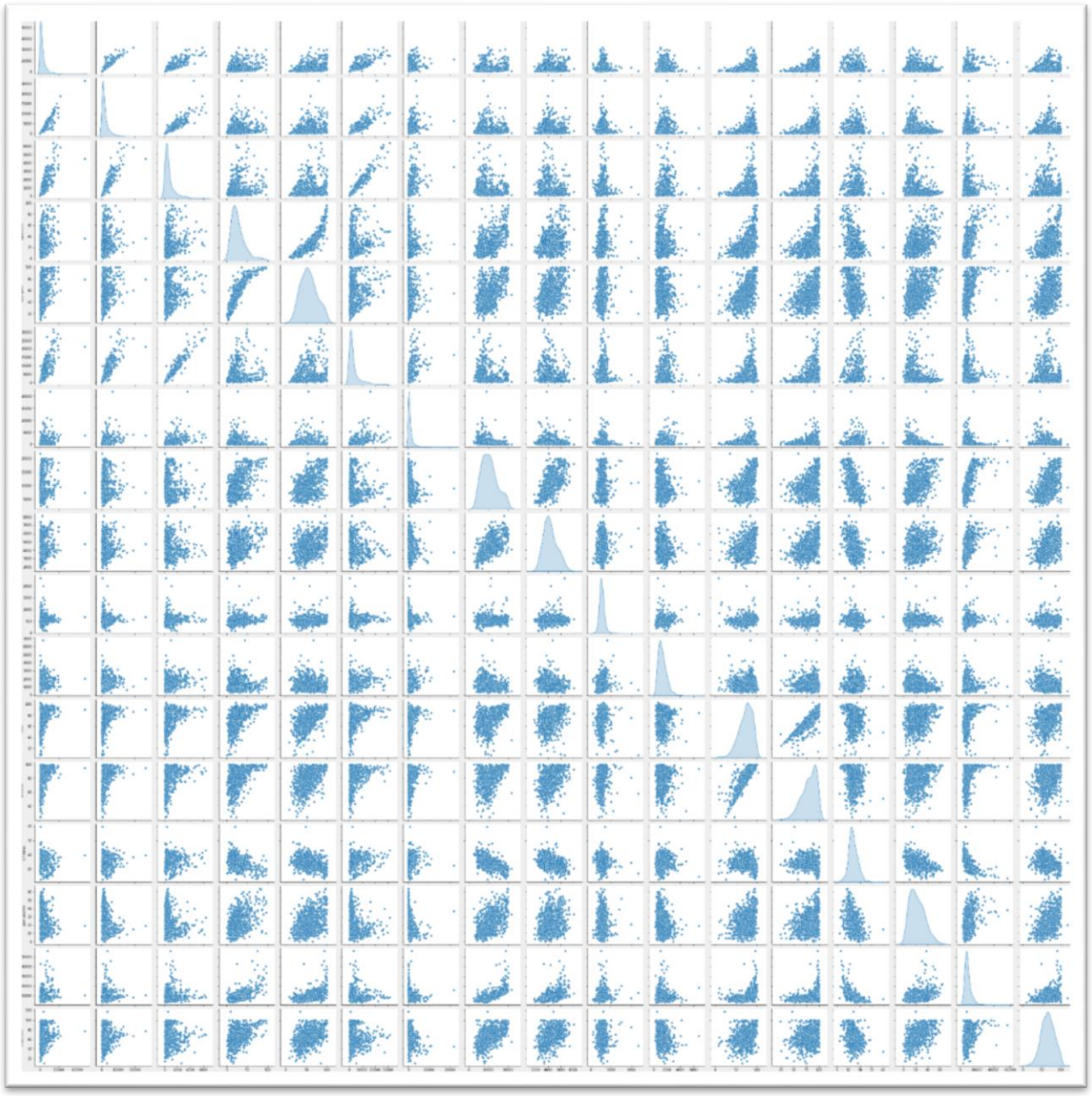


Graph 2.17 EDA graph

The graduation rate amongst the students in all the university above 60%. The box plot of the graduation rate has the outliers in the data set. The distribution is normally distributed.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Multivariate Analysis:

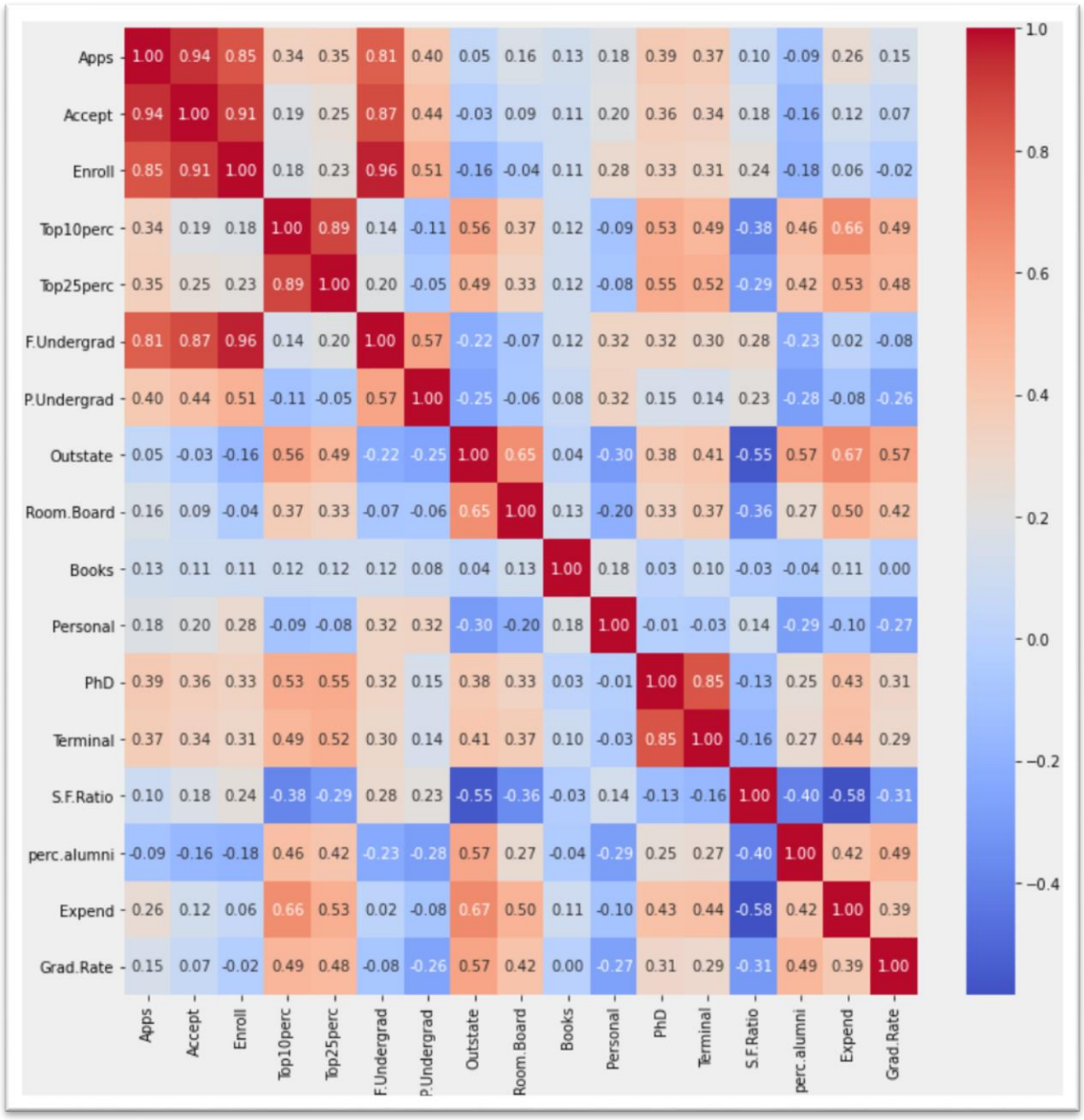


Graph 2.18 EDA graph

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 pattern or trends in the dataset

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Heat map:



Graph 2.19 Heat Map

This heat map gives us the correlation between two numerical values. We can easily detect from this that application variable is highly positively correlated with application accepted, students enrolled and full-time graduates. So, this relation ship gives the insights on when the student submits the application, it is accepted and the student is enrolled as full time 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, 25of higher secondary class, outstate, room board, books, personal, PhD, terminal, SF ratio, expenditure and graduation ratio are positively correlated.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

2.2 Is scaling necessary for PCA in this case? Give justification and perform scaling.

Before performing scaling, we have dropped the names variable as it is categorical in nature. Now, the data set consist of only numerical values. Then we have applied z-score method for this case study. We can also min max function to scale the variables.

Since the data set has 18 numerical columns with different scales. For example, the application, accepted application, enrolled fulltime graduates, part-time graduates, outstate are number of students. The top 10 percent and top 20 percent are students in which the values are given in percentage. Room board, books, and personal are values associated with money. The PhD, SF Ratio, percentage of alumni are percentage values of different combinations of student's alumni these are percentage values. The graduation rate is also percentage value of graduates who get graduated every year.

```
from scipy.stats import zscore
df_z=df_1.apply(zscore)
df_z.head()
```

Formula 2.1 Applying Z score on the table

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Rat
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	1.01377
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	-0.47770
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	-0.30074
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	-1.61527
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	-0.55354

Table 2.4. New table with Z score

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

Formula 2.2 Z Score Formula

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

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

2.3 Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].

The comparison between the covariance and correlation matrix is that both of the terms measures the relationship and the dependency between two variables.

Scaling in general means representation of the dataset. The numbers will not change. We are bringing the dataset into one unit.

Covariance indicates the direction of the linear relationship between the variables weather it is positive or negative. By direction means, it is directly proportional or inversely proportional.

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

Formula 2.3 Covariance Formula

Correlation measures the strength and the direction of the linear relationship between two variables. Strength is that positively correlated or negatively correlated.

$$Correlation = \frac{Cov\,(x,y)}{\sigma x * \sigma y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

Formula 2.4 Correlation Formula

This below snippet is the covariance matrix on the scaled dataset. We can clearly understand covariance matrix indicates direction of the linear relationship between the variables. By direction means, it is directly proportional or inversely proportional.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

```
Covariance Matrix
%s [[ 1.00128866  0.94466636  0.84791332  0.33927032  0.35209304  0.81554018
   0.3987775   0.05022367  0.16515151  0.13272942  0.17896117  0.39120081
   0.36996762  0.09575627 -0.09034216  0.2599265   0.14694372]
 [ 0.94466636  1.00128866  0.91281145  0.19269493  0.24779465  0.87534985
   0.44183938 -0.02578774  0.09101577  0.11367165  0.20124767  0.35621633
   0.3380184   0.17645611 -0.16019604  0.12487773  0.06739929]
 [ 0.84791332  0.91281145  1.00128866  0.18152715  0.2270373   0.96588274
   0.51372977 -0.1556777  -0.04028353  0.11285614  0.28129148  0.33189629
   0.30867133  0.23757707 -0.18102711  0.06425192 -0.02236983]
 [ 0.33927032  0.19269493  0.18152715  1.00128866  0.89314445  0.1414708
  -0.10549205  0.5630552   0.37195909  0.1190116  -0.09343665  0.53251337
   0.49176793 -0.38537048  0.45607223  0.6617651   0.49562711]
 [ 0.35209304  0.24779465  0.2270373   0.89314445  1.00128866  0.19970167
  -0.05364569  0.49002449  0.33191707  0.115676   -0.08091441  0.54656564
   0.52542506 -0.29500852  0.41840277  0.52812713  0.47789622]
 [ 0.81554018  0.87534985  0.96588274  0.1414708   0.19970167  1.00128866
   0.57124738 -0.21602002 -0.06897917  0.11569867  0.31760831  0.3187472
   0.30040557  0.28006379 -0.22975792  0.01867565 -0.07887464]
 [ 0.3987775   0.44183938  0.51372977 -0.10549205 -0.05364569  0.57124738
   1.00128866 -0.25383901 -0.06140453  0.08130416  0.32029384  0.14930637
   0.14208644  0.23283016 -0.28115421 -0.08367612 -0.25733218]
 [ 0.05022367 -0.02578774 -0.1556777   0.5630552   0.49002449 -0.21602002
  -0.25383901  1.00128866  0.65509951  0.03890494 -0.29947232  0.38347594
   0.40850895 -0.55553625  0.56699214  0.6736456   0.57202613]
 [ 0.16515151  0.09101577 -0.04028353  0.37195909  0.33191707 -0.06897917
  -0.06140453  0.65509951  1.00128866  0.12812787 -0.19968518  0.32962651
   0.3750222  -0.36309504  0.27271444  0.50238599  0.42548915]
 [ 0.13272942  0.11367165  0.11285614  0.1190116   0.115676    0.11569867
   0.08130416  0.03890494  0.12812787  1.00128866  0.17952581  0.0269404
   0.10008351 -0.03197042 -0.04025955  0.11255393  0.00106226]
 [ 0.17896117  0.20124767  0.28129148 -0.09343665 -0.08091441  0.31760831
   0.32029384 -0.29947232 -0.19968518  0.17952581  1.00128866 -0.01094989
  -0.03065256  0.13652054 -0.2863366  -0.09801804 -0.26969106]
 [ 0.39120081  0.35621633  0.33189629  0.53251337  0.54656564  0.3187472
   0.14930637  0.38347594  0.32962651  0.0269404  -0.01094989  1.00128866
   0.85068186 -0.13069832  0.24932955  0.43331936  0.30543094]
[ 0.36996762  0.3380184   0.30867133  0.49176793  0.52542506  0.30040557
   0.14208644  0.40850895  0.3750222   0.10008351 -0.03065256  0.85068186
   1.00128866 -0.16031027  0.26747453  0.43936469  0.28990033]
[ 0.09575627  0.17645611  0.23757707 -0.38537048 -0.29500852  0.28006379
   0.23283016 -0.55553625 -0.36309504 -0.03197042  0.13652054 -0.13069832
  -0.16031027  1.00128866 -0.4034484  -0.5845844  -0.30710565]
[-0.09034216 -0.16019604 -0.18102711  0.45607223  0.41840277 -0.22975792
  -0.28115421  0.56699214  0.27271444 -0.04025955 -0.2863366   0.24932955
   0.26747453 -0.4034484   1.00128866  0.41825001  0.49153016]
[ 0.2599265   0.12487773  0.06425192  0.6617651   0.52812713  0.01867565
  -0.08367612  0.6736456   0.50238599  0.11255393 -0.09801804  0.43331936
   0.43936469 -0.5845844   0.41825001  1.00128866  0.39084571]
[ 0.14694372  0.06739929 -0.02236983  0.49562711  0.47789622 -0.07887464
  -0.25733218  0.57202613  0.42548915  0.00106226 -0.26969106  0.30543094
   0.28990033 -0.30710565  0.49153016  0.39084571  1.00128866]]
```

Table 2.5. Covariance Matrix

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

This below snippet is the correlation matrix. We can clearly understand the correlation matrix which gives the strength and the relationship between the variables.

The correlation matrix before the scaling will remain the same.

From the snippet, we can understand variables which are highly positively correlated and the variables which are highly negatively correlated. We can also understand the variables which are moderately correlated with each other.

We can see that application, acceptance, enrollment and fulltime graduates are highly positively correlated.

Also, the top 10 percentage and top 25 percentage are highly positively correlated.

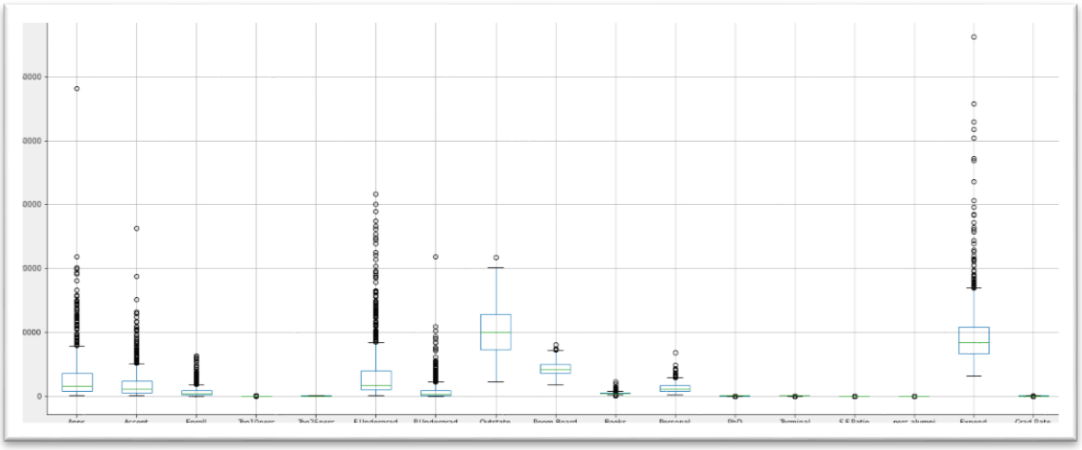
	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Termin
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	0.3694
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899	0.113525	0.200989	0.355758	0.3375
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	0.3082
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	0.4911
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	0.5247
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	0.3000
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	0.1419
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	0.4079
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	0.3745
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	0.0999
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	-0.0306
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	0.8495
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	1.0000
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	-0.1601
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	0.2671
Expend	0.259592	0.124717	0.064169	0.660913	0.527447	0.018652	-0.083568	0.672779	0.501739	0.112409	-0.097892	0.432762	0.4387
Grad.Rate	0.146755	0.067313	-0.022341	0.494989	0.477281	-0.078773	-0.257001	0.571290	0.424942	0.001061	-0.269344	0.305038	0.2895

Table 2.6. Correlation Matrix

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

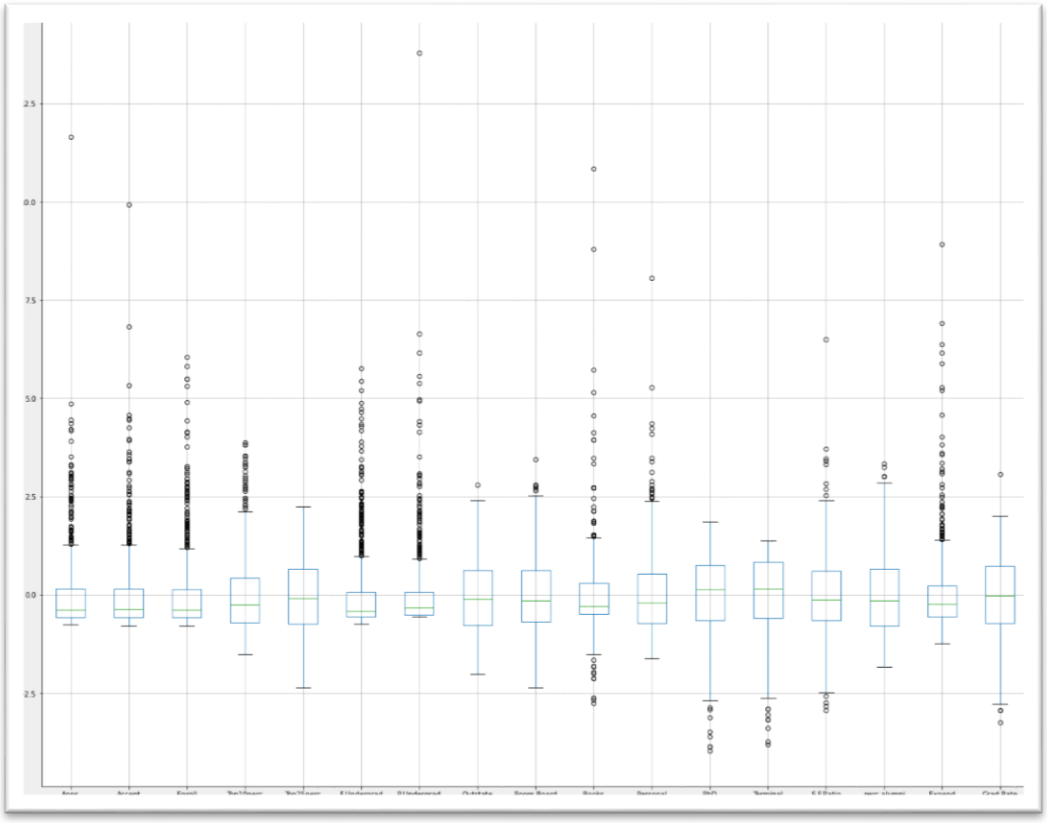
2.4 Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]

Checking for data before calling scaling:



Graph 2.20 Checking for outliers in the data

Checking the dataset after scaling:



Graph 2.21 Checking for outliers after performing scaling

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Inference:

The outliers are still present in the dataset.

Reason:

scaling does not remove outliers scaling values on a Z score distribution. We can use any method to remove outliers for further processes.

For example, if we wish to remove the outliers, we can consider taking 3 standard deviations as outliers or either we can remove them or impute them with the IQR values.

2.5 Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both]

```
[ 0.3987775   0.44183938  0.51372977 -0.10549205 -0.05364569  0.57124738
  1.00128866 -0.25383901 -0.06140453  0.08130416  0.32029384  0.14930637
  0.14208644  0.23283016 -0.28115421 -0.08367612 -0.25733218]
[ 0.05022367 -0.02578774 -0.1556777   0.5630552   0.49002449 -0.21602002
 -0.25383901  1.00128866  0.65509951  0.03890494 -0.29947232  0.38347594
  0.40850895 -0.55553625  0.56699214  0.6736456   0.57202613]
[ 0.16515151  0.09101577 -0.04028353  0.37195909  0.33191707 -0.06897917
 -0.06140453  0.65509951  1.00128866  0.12812787 -0.19968518  0.32962651
  0.3750222  -0.36309504  0.27271444  0.50238599  0.42548915]
[ 0.13272942  0.11367165  0.11285614  0.1190116   0.115676    0.11569867
  0.08130416  0.03890494  0.12812787  1.00128866  0.17952581  0.0269404
  0.10008351 -0.03197042 -0.04025955  0.11255393  0.00106226]
[ 0.17896117  0.20124767  0.28129148 -0.09343665 -0.08091441  0.31760831
  0.32029384 -0.29947232 -0.19968518  0.17952581  1.00128866 -0.01094989
 -0.03065256  0.13652054 -0.2863366  -0.09801804 -0.26969106]
[ 0.39120081  0.35621633  0.33189629  0.53251337  0.54656564  0.3187472
  0.14930637  0.38347594  0.32962651  0.0269404  -0.01094989  1.00128866
  0.85068186 -0.13069832  0.24932955  0.43331936  0.30543094]
[ 0.36996762  0.3380184   0.30867133  0.49176793  0.52542506  0.30040557
  0.14208644  0.40850895  0.3750222   0.10008351 -0.03065256  0.85068186
  1.00128866 -0.16031027  0.26747453  0.43936469  0.28990033]
[ 0.09575627  0.17645611  0.23757707 -0.38537048 -0.29500852  0.28006379
  0.23283016 -0.55553625 -0.36309504 -0.03197042  0.13652054 -0.13069832
 -0.16031027  1.00128866 -0.4034484  -0.5845844  -0.30710565]
[-0.09034216 -0.16019604 -0.18102711  0.45607223  0.41840277 -0.22975792
 -0.28115421  0.56699214  0.27271444 -0.04025955 -0.2863366   0.24932955
  0.26747453 -0.4034484   1.00128866  0.41825001  0.49153016]
[ 0.2599265   0.12487773  0.06425192  0.6617651   0.52812713  0.01867565
 -0.08367612  0.6736456   0.50238599  0.11255393 -0.09801804  0.43331936
  0.43936469 -0.5845844   0.41825001  1.00128866  0.39084571]
[ 0.14694372  0.06739929 -0.02236983  0.49562711  0.47789622 -0.07887464
 -0.25733218  0.57202613  0.42548915  0.00106226 -0.26969106  0.30543094
  0.28990033 -0.30710565  0.49153016  0.39084571  1.00128866]]
```


Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

```
Eigen Vectors
%s [[-2.48765602e-01  3.31598227e-01  6.30921033e-02 -2.81310530e-0
   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]
```


Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

```
[-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]]
```

Table 2.7. Eigen vector

```
 Eigen Values
%s [5.45052162 4.48360686 1.17466761 1.00820573 0.93423123 0.84849117
 0.6057878  0.58787222 0.53061262 0.4043029  0.02302787 0.03672545
 0.31344588 0.08802464 0.1439785  0.16779415 0.22061096]
```

Table 2.8. Eigen Values

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features

```
pca.explained_variance_
```

```
array([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])
```

Table 2.9. PCA Variance

```
df_z.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')
```

Table 2.10. Columns

```
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],
```


Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

```
[ 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],
```


Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

```
[ 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]])
```

Table 2.11. Array Table

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]

```
print('The Linear eq of 1st component: ')
for i in range(0,df_z.shape[1]):
    print('{} * {}'.format(np.round(pca.components_[0][i],3),df_z.columns[i]),end=' + ')
```

Formula 2.5. Formation for the linear equation

The Linear eq of 1st component:
0.249 * Apps + 0.208 * Accept + 0.176 * Enroll + 0.354 * Top10perc + 0.344 * Top25perc + 0.155 * F.Undergrad + 0.026 * P.Undergrad + 0.295 * Outstate + 0.249 * Room.Board + 0.065 * Books + -0.043 * Personal + 0.318 * PhD + 0.317 * Terminal + -0.177 * S.F.Ratio + 0.205 * perc.alumni + 0.319 * Expend + 0.252 * Grad.Rate +

2.8 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?

```
tot = sum(eig_vals)
var_exp = [( i /tot ) * 100 for i in sorted(eig_vals, reverse=True)]
cum_var_exp = np.cumsum(var_exp)
cum_var_exp
```

```
array([ 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.        ])
```

Formula 2.6. Formation of the sum of eigen values in an array form

Adding the Eigen values, we will get the sum of 100.
To decide the optimum number of principal components.
Check for cumulative variance up to 90%, check the corresponding associated with 90%.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

The incremental value between the components should not be less than five percent.

So, basis of this, we can decide the optimum number of principal components as 6, because after this, incremental value between them is less than 5%.

So, we select 5 principal components for this case study.

```
pca = PCA(n_components=5)
X_pca= pca.fit_transform(df_z)
```

```
pca.components_
```

```
array([[ 0.2487656 ,  0.2076015 ,  0.17630359,  0.35427395,  0.34400128,
         0.15464096,  0.0264425 ,  0.29473642,  0.24903045,  0.06475752,
        -0.04252854,  0.31831287,  0.31705602, -0.17695789,  0.20508237,
         0.31890875,  0.25231565],
       [ 0.33159823,  0.37211675,  0.40372425, -0.08241182, -0.04477865,
         0.41767377,  0.31508783, -0.24964352, -0.13780888,  0.05634184,
         0.21992922,  0.05831132,  0.04642945,  0.24666528, -0.24659527,
        -0.13168986, -0.16924053],
       [-0.06309202, -0.10124915, -0.08298568,  0.03505545, -0.02414788,
        -0.06139284,  0.1396817 ,  0.04659889,  0.14896738,  0.67741165,
         0.49972112, -0.12702836, -0.06603756, -0.2898484 , -0.14698927,
         0.226744  , -0.20806465],
       [ 0.28131045,  0.26781744,  0.16182688, -0.05154717, -0.1097666 ,
         0.10041219, -0.15855847,  0.13129134,  0.184996  ,  0.08708922,
        -0.23071057, -0.53472484, -0.519443  , -0.16118948,  0.01731422,
         0.07927348,  0.26912907],
       [ 0.00574144,  0.05578606, -0.05569368, -0.39543438, -0.42653357,
        -0.04345431,  0.3023854 ,  0.22253201,  0.56091947, -0.12728882,
        -0.22231102,  0.14016633,  0.20471972, -0.07938825, -0.21629741,
         0.07595812, -0.10926791]])
```

Table 2.12. PCA Components

The first components explain 32.02% variance in the data.
The first two components explain 58.36% variance in the data.
The first three components explain 65.26% variance in the data.
The first four components explain 71.18% variance in the data.
The first five components explain 76.67% variance in data.

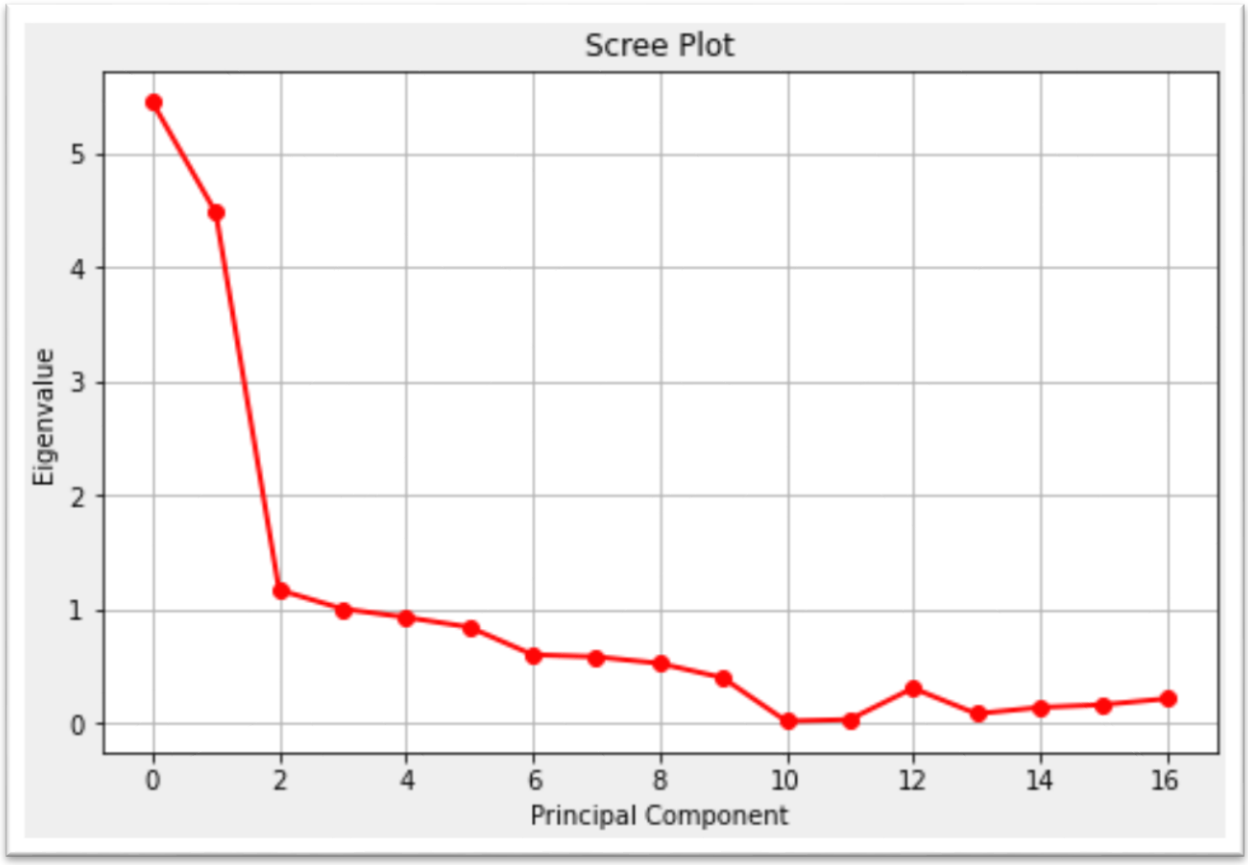
```
tot = sum(eig_vals)
var_exp = [( i /tot ) * 100 for i in sorted(eig_vals, reverse=True)]
cum_var_exp = np.cumsum(var_exp)
cum_var_exp
```

```
array([ 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.        ])
```

Formula 2.7. Formation of the sum of eigen values in an array form

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

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 collinearity we run models and improved efficiency scores.



Graph 2.22 Scree Plot

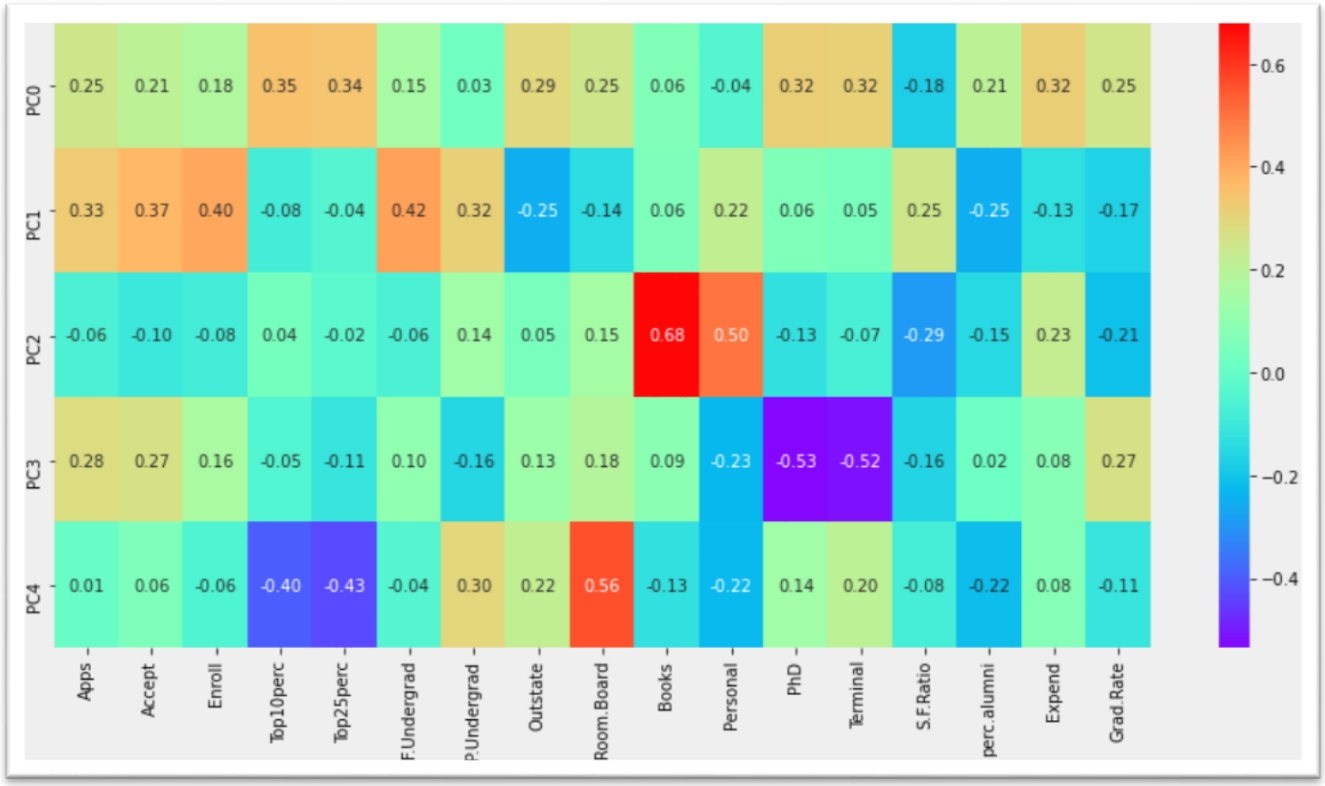
PCA is performed and it is exported into the data frame. After PCA the multi collinearity is highly reduced.

```
df_comp = pd.DataFrame(pca.components_,columns=list(df_z))
df_comp
```

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Rat
0	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	-0.1769
1	0.331598	0.372117	0.403724	-0.082412	-0.044779	0.417674	0.315088	-0.249644	-0.137809	0.056342	0.219929	0.058311	0.046429	0.2466
2	-0.063092	-0.101249	-0.082986	0.035055	-0.024148	-0.061393	0.139682	0.046599	0.148967	0.677412	0.499721	-0.127028	-0.066038	-0.2898
3	0.281310	0.267817	0.161827	-0.051547	-0.109767	0.100412	-0.158558	0.131291	0.184996	0.087089	-0.230711	-0.534725	-0.519443	-0.1611
4	0.005741	0.055786	-0.055694	-0.395434	-0.426534	-0.043454	0.302385	0.222532	0.560919	-0.127289	-0.222311	0.140166	0.204720	-0.0793

Table 2.13. Data frame

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited



Graph 2.23 Heat Map

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

2.9 Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [Hint: Write Interpretations of the Principal Components Obtained]

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 data set 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. Outliers are imputed using IQR values once the values are imputed, we can perform PCA. 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 5 where we could understand the maximum variance of the dataset. Using the components, we can now understand the reduced multicollinearity in the dataset.

With this analysis, we can perform further analysis and model building PCA will improve the efficiency of machine learning models.

END

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited