ADVANCED STATISTICS PROJECT

Submitted by,

JIYA JACOB

PGP-DSBA ONLINE

JULY-B 2021

DATE:17/10/2021

SALARY DATA

CONTENTS

TOPIC	PAGE NO
Executive summary	5
Introduction	5
Data summary: Sample of data set	5
Exploratory data analysis	
Checking the types of variables in the data frame	6
Checking for the missing values in the dataset	6
1.1 State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.	6
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.	7
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.	7
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**.**	8
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.	8
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?	9
1.7 Explain the business implications of performing ANOVA for this particular case study.	10

LIST OF FIGURES

Fig 1: Interaction plot between the treatments with confidence intervals	9
Fig 2: Interaction plot between the treatments without confidence intervals	9

LIST OF TABLES

Table 1: Dataset sample	5
Table 2: Summary of Data	6
Table 3: The table showing the different class combinations of the education levels and their corresponding p-values	8
Table 4: The table showing the result of two-way ANOVA conducted (along with interaction) on the dataset.	10

EXECUTIVE SUMMARY

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals 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.

INTRODUCTION

The purpose of this whole exercise is to perform one-way ANOVA and two-way ANOVA under the assumption that the given dataset follows a normal distribution. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. Formulate hypothesis and analyze the influence of each of the independent variable on the dependent variable with the help of one-way ANOVA. Finding out the interaction between the two independent variable and analyze the interaction between two treatments using the two-way ANOVA. This assignment should help the student in exploring the summary statistics, hypothesis formulation, performing one-way and two-way ANOVA.

DATA DESCRIPTION

1. Education- 3 levels; doctorate, bachelors and HS-grad
2. Occupation- Prof-Speciality, Sales, Adm- Clerical and Exec- managerial
3. Salary-continuous from 50103 to 260151

Sample of the dataset:

	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: Dataset Sample

The data consists of salary of 40 individuals along with their educational qualifications, mainly of three levels and occupation which are of four kinds.

EXPLORATORY DATA ANALYSIS

Let us check the types of variables in the data frame.

```
Education     object
Occupation    object
Salary         int64
dtype: object
```

There are total 40 rows and 3 columns in the dataset. Out of 3, 2 columns are of object type and the remaining one is of integer data type.

Check for missing values in the dataset:

```
<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)
```

From the above results we can see that there is no missing value present in the dataset.

Summary of the dataset:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Education	40	3	Doctorate	16	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Occupation	40	4	Prof-specialty	13	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	40.0	NaN	NaN	NaN	162186.875	64860.407506	50103.0	99897.5	169100.0	214440.75	260151.0

TABLE 2: SUMMARY OF THE DATA

From the descriptive statistics, we can see that there are 3 different education levels of the employees namely, Doctorate, Bachelors and HS-grad and the occupation levels of these people are of four types, i.e. Prof-Specialty, Sales, Adm- Clerical and Exec-managerial. From the above table, we can see that the maximum salary among the 40 people was 260151.0 whereas the minimum salary sums up to just 50103.0

Q1.1: **State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.**

1.Stating the null and alternate hypothesis for conducting one way ANOVA for education

Null Hypothesis $H0$: The mean salary is same with different categories of education

Alternate Hypothesis HA : The mean salary is different in at-least one category of education

2.Stating the null and alternate hypothesis for conducting one way ANOVA for occupation

Null Hypothesis $H0$: The mean salary is same with different categories of occupation

Alternate Hypothesis HA : The mean salary is different in at-least one category of occupation

Q1.2: 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

Conclusion:

Since the p value is lesser than the significance level (0.05), we can reject the null hypothesis which states that the mean salary due to different levels of education are the same. It means that different levels of education have a significant effect on the mean salary.

Q1.3: 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

Conclusion:

Since the p value is greater than the significance level (0.05), we cannot reject the null hypothesis which states that the mean salary due to different levels of education are the same. It means that different levels of occupation do not have a significant effect on the mean salary.

Q1.4: 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.

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 3: The given above table shows the different class combinations of the education levels and their corresponding p-values

Interpretation:

Since the mean difference between Bachelor-Doctorate is more it means that their mean salaries are different while the mean salaries for the bachelors-HS grad and Doctorate-HS grad combination are not much. Here we cannot rely on p-value because all the three values are less than 0.05 which means that we reject the null hypothesis in all the three cases.

Q1.5: 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.

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation):C(Education)	11.0	1.434497e+11	1.304088e+10	18.339811	3.441555e-10
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

Given below shows the Interaction Effect: Occupation v/s Education

Fig 1: Interaction plot between the treatments with confidence intervals

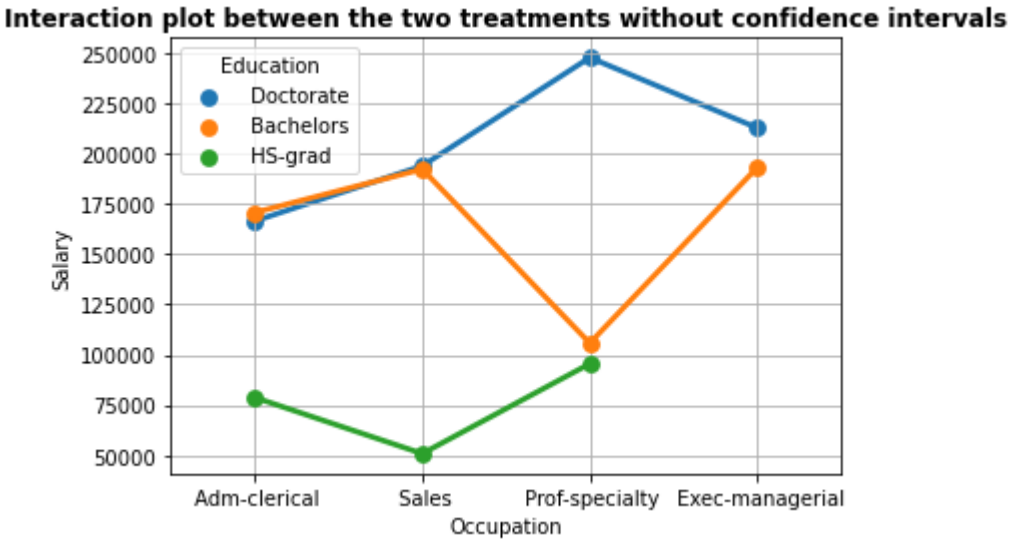

Fig 2: Interaction plot between the treatments without confidence intervals

From the above analysis, we can conclude than as occupation and education interaction is less than 0.05, there seems to be a significant statistical interaction between the both factors.

Q1.6: 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?

Null Hypothesis($H0$): The means of 'Salary' variable with respect to each occupation category and educational level is equal.

Alternative Hypothesis($H1$): At least one of the means of 'Salary' variable with respect to each occupation category and educational level is unequal.

	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 4: The table showing the result of two-way ANOVA conducted (along with interaction) on the dataset.

Interpretation of the result:

1. Considering both the factors (education and occupation), education is a significant factor as the p value is <0.05 whereas occupation is not a significant variable as p value of occupation is >0.05
2. As occupation and education interaction is less than 0.05, there seems to be a significant statistical interaction between the both factors.

Q1.7: Explain the business implications of performing ANOVA for this particular case study.

ANOVA means Analysis of variance. It is a technic that helps us to compare two population means and test the equality between the two. ANOVA performs the test of equality for more than two population means by actually analysing the variance. The primary purpose of two-way ANOVA to understand if there is an interaction between the two independent variables on the dependent variable.

In the given dataset, it is about the various details of the salary and the different education and occupation categories. We formulated null and alternative hypothesis and conducted one-way ANOVA taking into consideration both occupation and education levels separately. By conducting the one-way ANOVA, we were able to find out which independent variable has direct influence on the mean salary, that is the dependent variable. The two-way ANOVA was done to understand whether there is an interaction between the two independent variables on the dependent variable. After doing all these tests, we concluded that Education had a direct impact on mean salary while there is found to be a significant statistical interaction between education and occupation. So we can conclude that people with better education is expected to be paid more salary than in comparison to the occupation categories.

EDUCATION

POST 12TH

STANDARD

DATA

CONTENTS

TOPIC	PAGE NO
Executive summary	15
Introduction	15
Data summary: Sample of data set	15
Exploratory data analysis	
Checking the types of variables in the data frame	16
Checking for the missing values in the dataset	17
2.1. Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?	17
2.2. Is scaling necessary for PCA in this case? Give justification and perform scaling.	27
2.3. Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data]	28
2.4. Check the dataset for outliers before and after scaling. What insight do you derive here?	30
2.5. Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both]	31
2.6. Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features	33
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]	34
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?	34
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]	35

LIST OF FIGURES

Fig 1.1: The box plot and the histogram showing the distribution of data of 'Apps' variable	17
Fig 1.2: The box plot and the histogram showing the distribution of data of 'Accept' variable	18
Fig 1.3: The box plot and the histogram showing the distribution of data of 'Enroll' variable	18
Fig 1.4: The box plot and the histogram showing the distribution of data of 'Top10perc' variable	19
Fig 1.5: The box plot and the histogram showing the distribution of data of 'Top25perc' variable	19
Fig 1.6: The box plot and the histogram showing the distribution of data of 'F.Undergrad' variable	20
Fig 1.7: The box plot and the histogram showing the distribution of data of 'P.Undergrad' variable	20
Fig 1.8: The box plot and the histogram showing the distribution of data of 'Outstate' variable	21
Fig 1.9: The box plot and the histogram showing the distribution of data of 'Room.Board' variable	21
Fig 1.10: The box plot and the histogram showing the distribution of data of 'Books' variable	22
Fig 1.11: The box plot and the histogram showing the distribution of data of 'Personal' variable	22
Fig 1.12: The box plot and the histogram showing the distribution of data of 'PhD' variable	23
Fig 1.13: The box plot and the histogram showing the distribution of data of 'Terminal' variable	23
Fig 1.14: The box plot and the histogram showing the distribution of data of 'S.F.Ratio' variable	24
Fig 1.15: The box plot and the histogram showing the distribution of data of 'perc.alumni' variable	24
Fig 1.16: The box plot and the histogram showing the distribution of data of 'Expend' variable	25
Fig 1.17: The box plot and the histogram showing the distribution of data of 'Grad.Rate' variable	25
Fig 1.18: Heat map is portrayed for the multi variate analysis of the data.	26
Fig 1.19: The pair plot showing the multivariate analysis	27
Fig 2: The box plots shown the data before scaling.	30
Fig 3: The boxplot shows the data after scaling	31

Fig 4: The heat map to portray the multicollinearity after the PCA is applied to the data set.	33
Fig 5: Scree plot to determine the number of principal components to be selected based on their contribution to the variation in the dataset.	35

LIST OF TABLES

Table 1: Dataset sample	15
Table 2. Exploratory Data analysis	
Table 2.1. Checking the variables types in the data set	16
Table 2.2. Checking for the missing values in the dataset	17
Table 3: The head of the dataset that has been scaled	28
Table 4: The above table shows the correlation matrix	28
Table 5: The above table shows the covariance matrix	29
Table 6: The above given table shows the list of eigen vectors.	32
Table 7: PCA is performed and the data is exported into the data frame with original features.	33

EXECUTIVE SUMMARY

The dataset provided contains information on various colleges and Principal Component Analysis is done for this case study according to the instructions given.

INTRODUCTION

The purpose of this whole exercise is to do the exploratory data analysis and then implement Principal component analysis on it. This assignment should help the student in knowing the art of dimensionality reduction and scaling down the data by taking into consideration only the influential factors without losing any of the useful data..

DATA DESCRIPTION

1. Names: Names of various university and colleges.
2. Apps: Number of applications received continuous from 81 to 48094.
3. Accept: Number of applications accepted continuous from 72 to 26330.
4. Enroll: Number of students enrolled continuous from 35 to 6392.
5. Top10perc: Percentage of new students from top 10% of Higher Secondary class from 1 to 96.
6. Top25perc: Percentage of new students from top 25% of Higher Secondary class from 9 to 100.
7. F.Undergrad: Number of full-time undergraduate students continuous from 139 to 31643.
8. P.Undergrad: Number of part-time undergraduate students continuous from 1 to 21836.
9. Outstate: Number of students for whom the particular college or university is Out-of-state tuition continuous from 2340 to 21700.
10. Room.Board: Cost of Room and board continuous from 1780 to 8124.
11. Books: Estimated book costs for a student continuous from 96 to 2340.
12. Personal: Estimated personal spending for a student continuous from 250 to 6800.
13. PhD: Percentage of faculties with Ph.D.'s from 8 to 103
14. Terminal: Percentage of faculties with terminal degree from 24 to 100.
15. S.F.Ratio: Student/faculty ratio 2.5 from 81 to 39.8
16. perc.alumni: Percentage of alumni who donate from 0 to 64.
17. Expend: The Instructional expenditure per student continuous from 3186 to 56233.
18. Grad.Rate: Graduation rate continuous from 10 to 118.

Sample of the dataset:

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
2	Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
3	Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
4	Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15

TABLE 1: Dataset Sample

Dataset has 777 entries with 18 variables of 3 datatypes. A study is done on understanding the influential factors in the given dataset.

EXPLORATORY DATA ANALYSIS

Let us check the types of variables in the data frame.

```
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. Checking the variables types in the data set

There are total 777 rows and 18 columns in the dataset. Out of 18, there is one column each for object data type and float data type and the rest are of int data type.

Check for missing values in the dataset:

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

Table 2.2. Checking for the missing values in the dataset

From the above results we can see that there is no missing value present in the dataset.

Q2.1. Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?

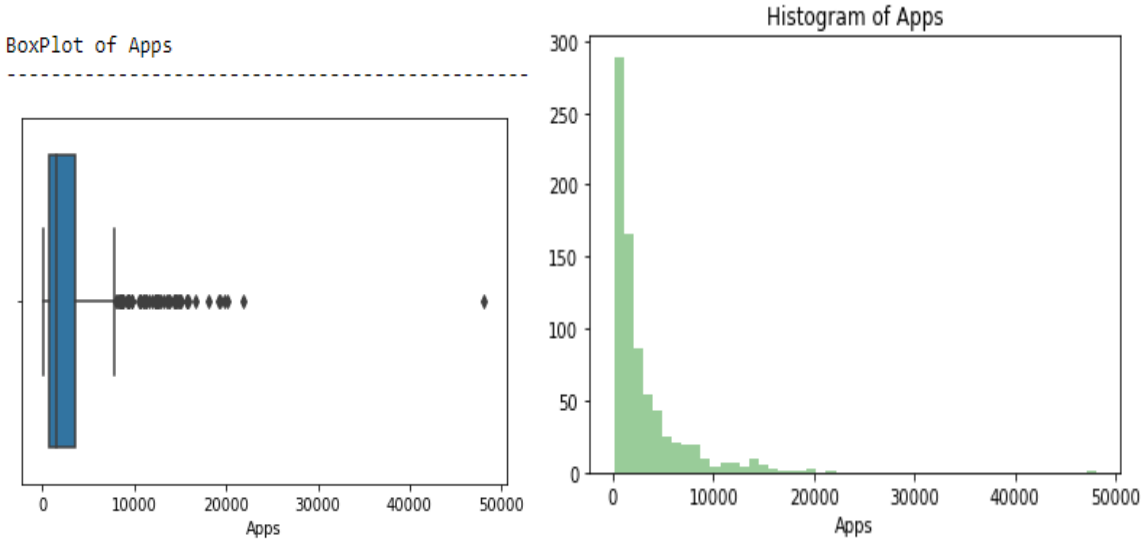

Fig 1.1: The box plot and the histogram showing the distribution of data of 'Apps' variable

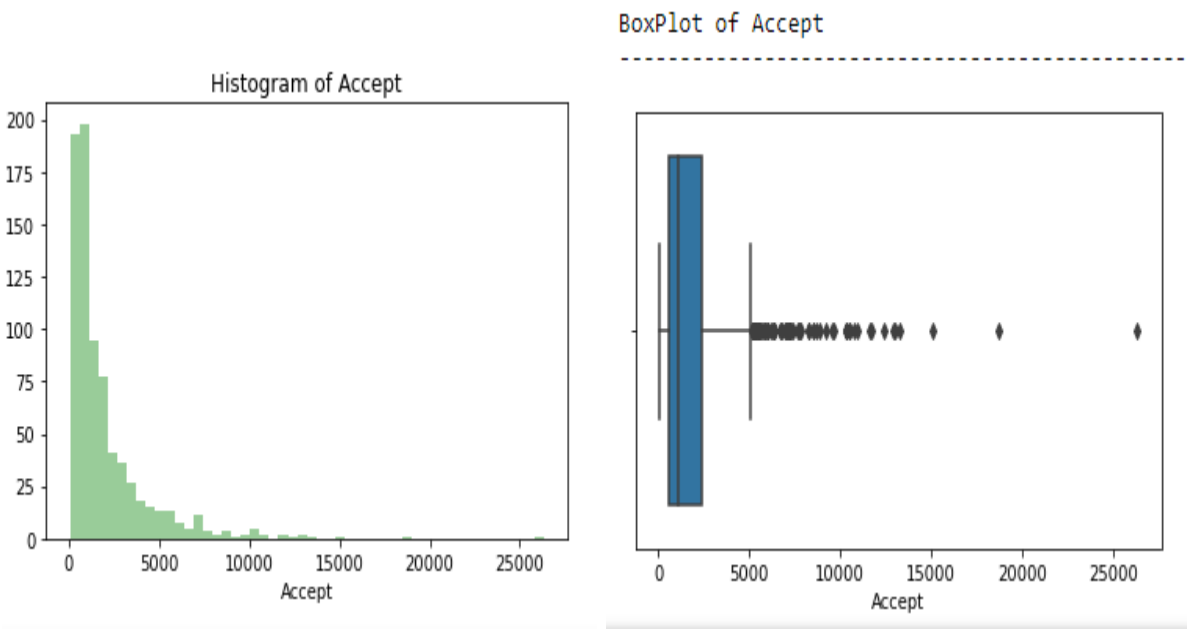

Fig 1.2: The box plot and the histogram showing the distribution of data of 'Accept' variable

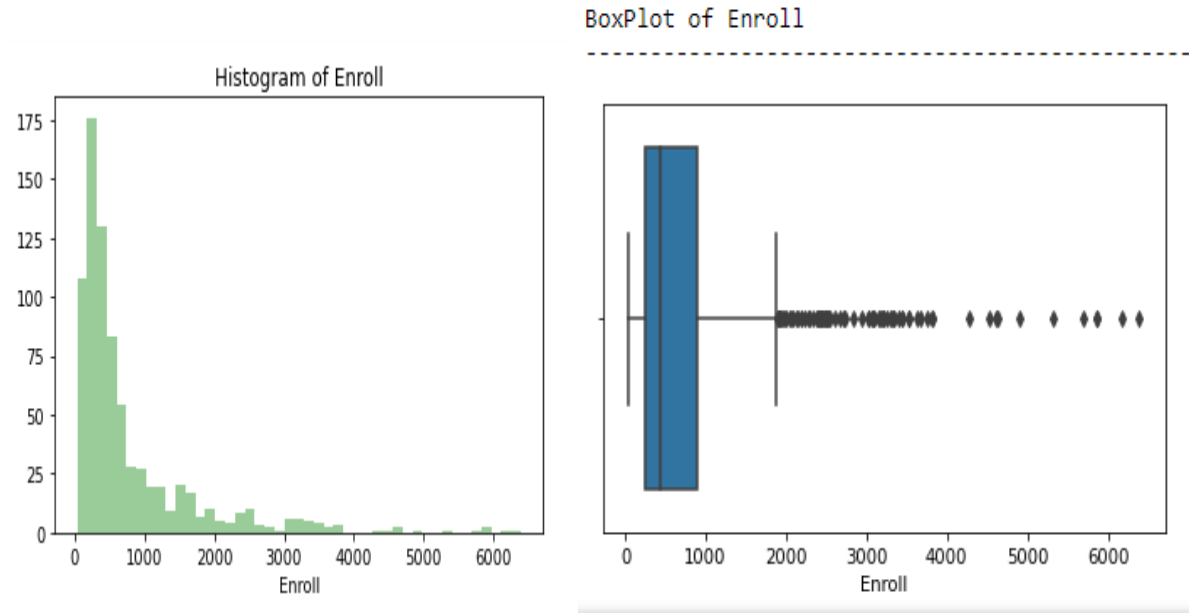

Fig 1.3: The box plot and the histogram showing the distribution of data of 'Enroll' variable

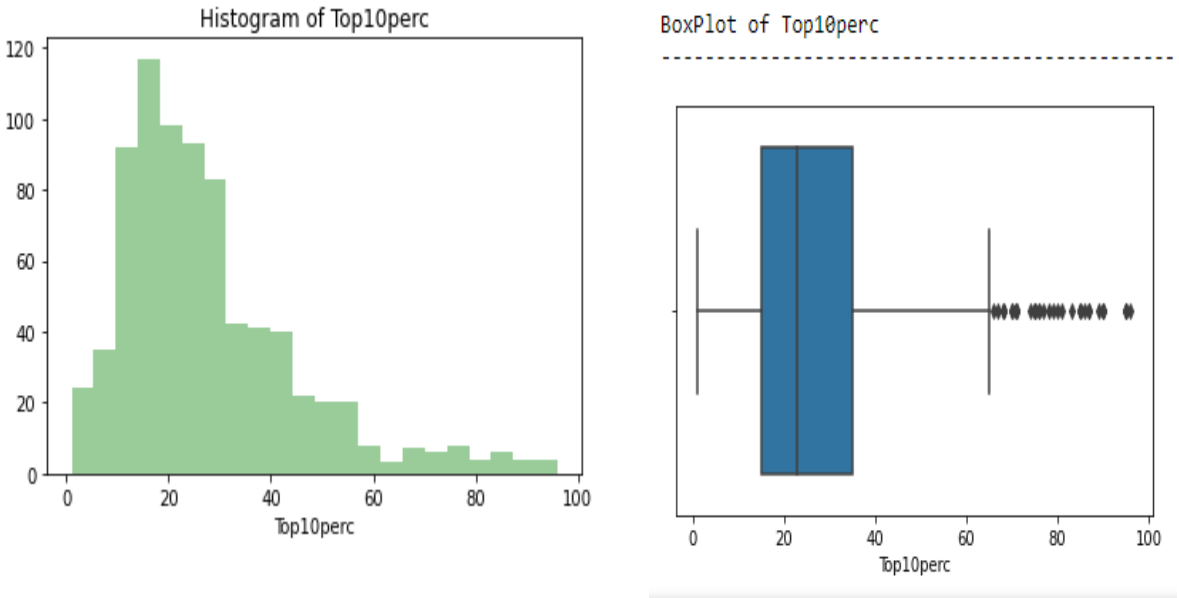

Fig 1.4: The box plot and the histogram showing the distribution of data of 'Top10perc' variable

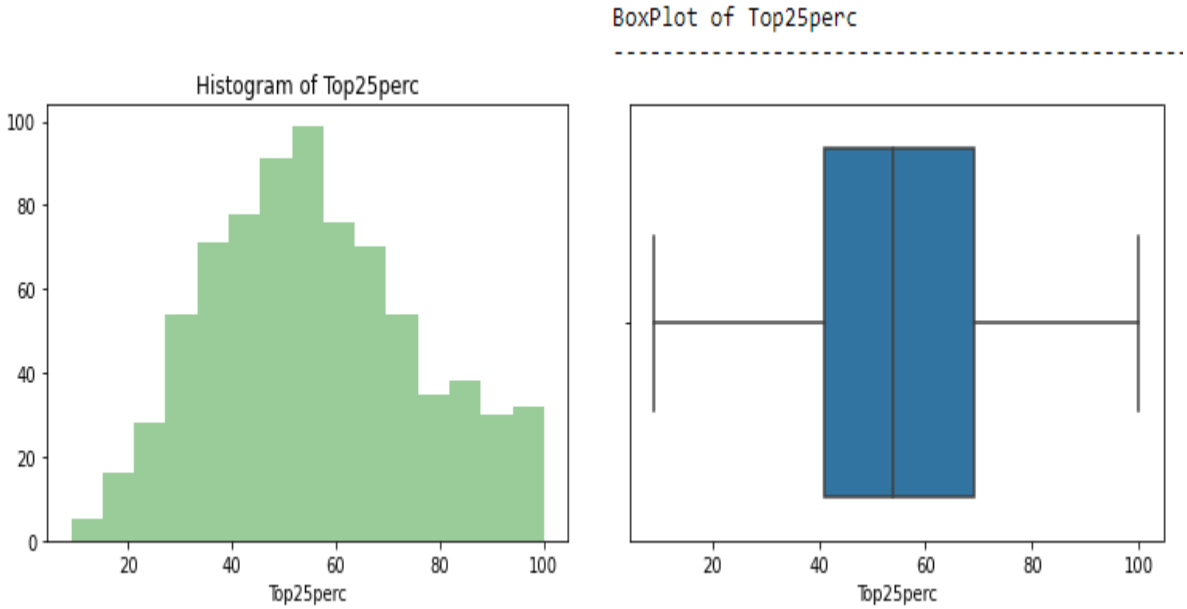

Fig 1.5: The box plot and the histogram showing the distribution of data of 'Top25perc' variable

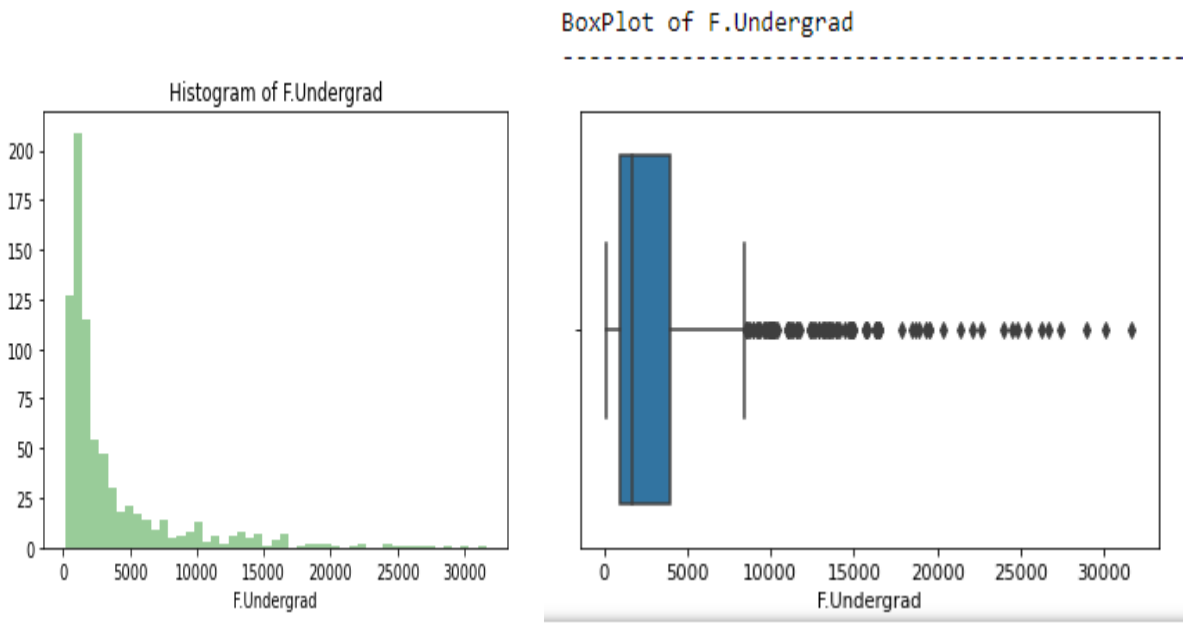

Fig 1.6: The box plot and the histogram showing the distribution of data of 'F.Undergrad' variable

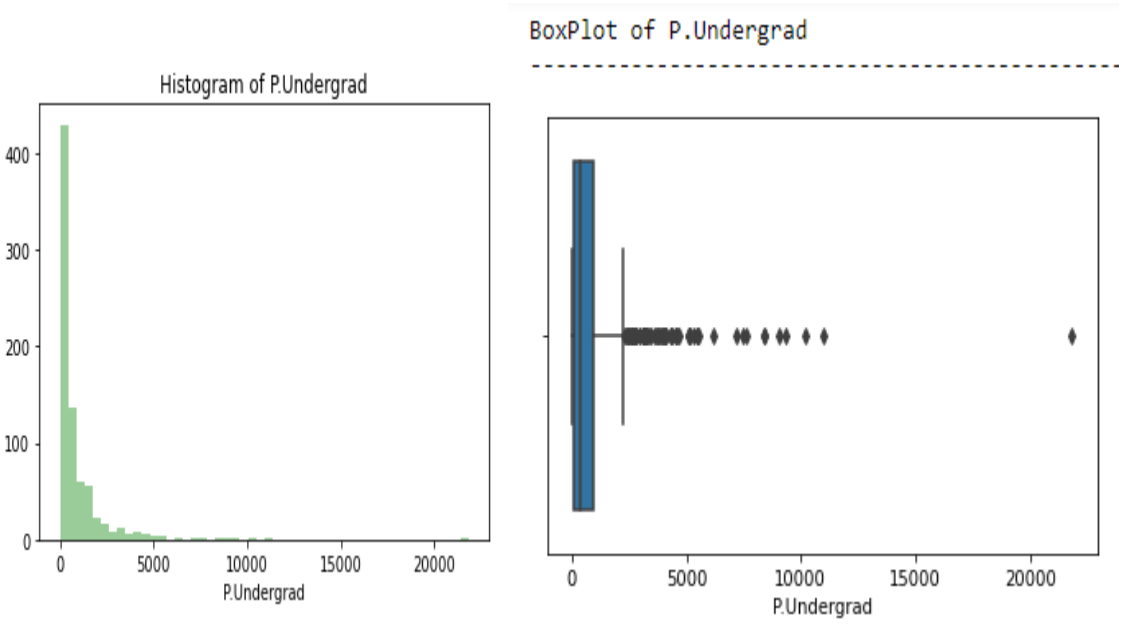

Fig 1.7: The box plot and the histogram showing the distribution of data of 'P.Undergrad' variable

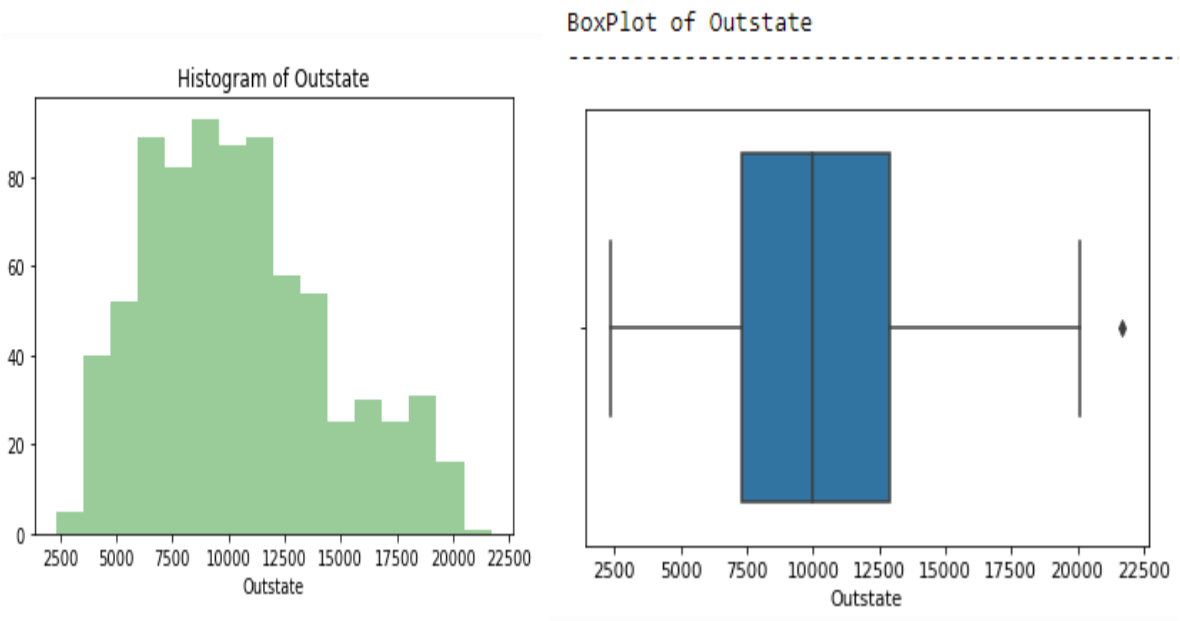

Fig 1.8: The box plot and the histogram showing the distribution of data of 'Outstate' variable

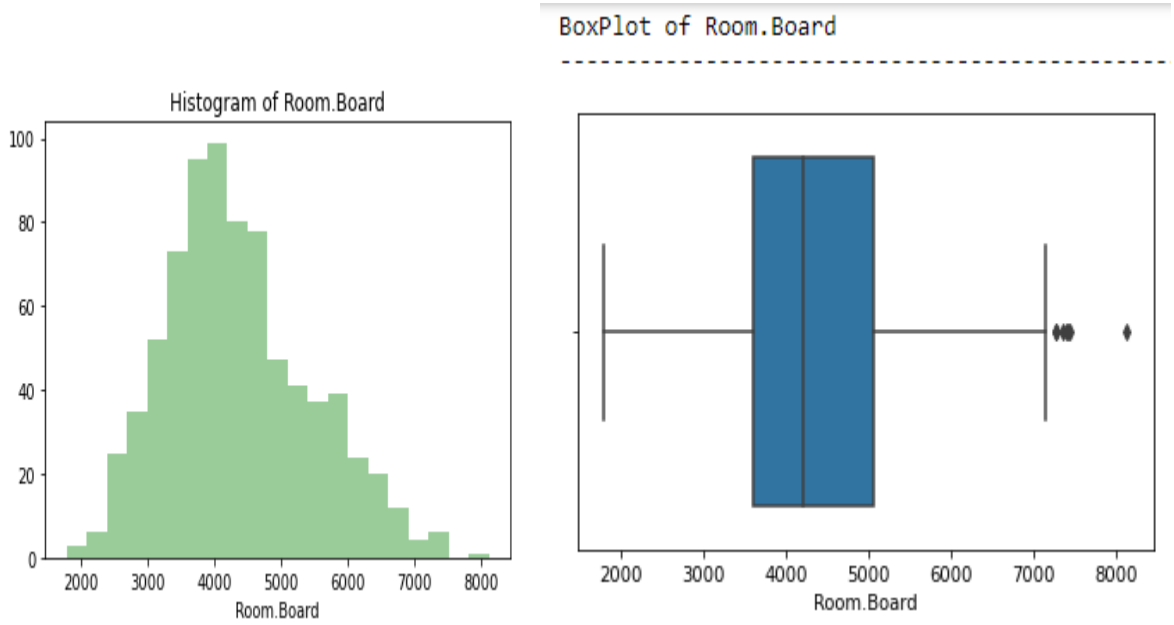

Fig 1.9: The box plot and the histogram showing the distribution of data of 'Room.Board' variable

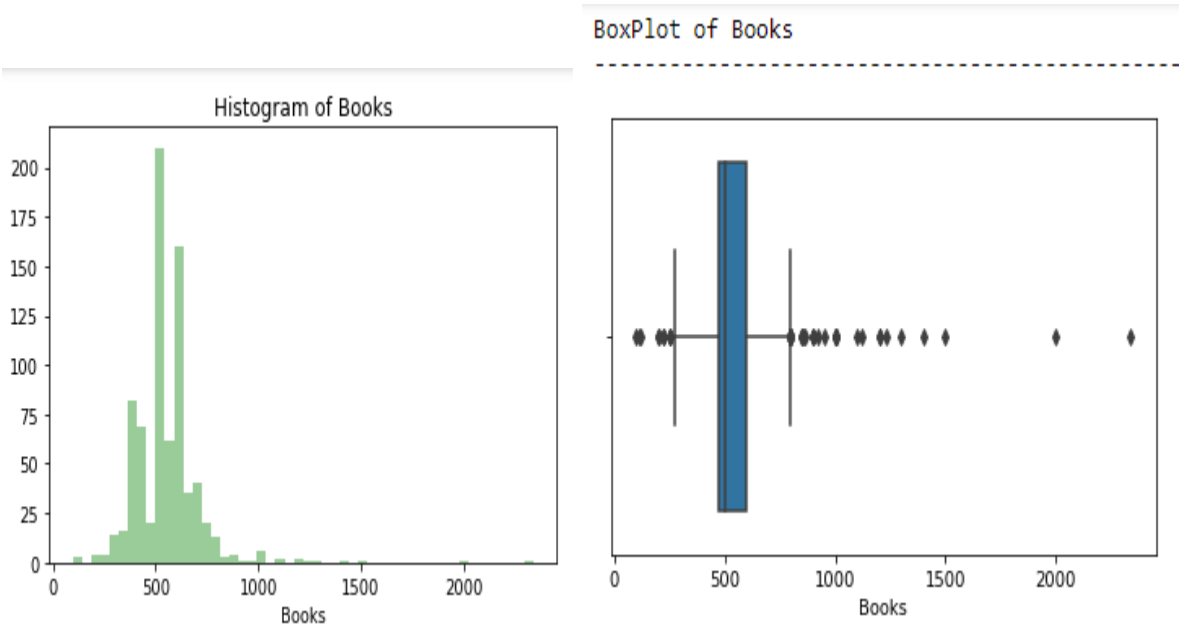

Fig 1.10: The box plot and the histogram showing the distribution of data of 'Books' variable

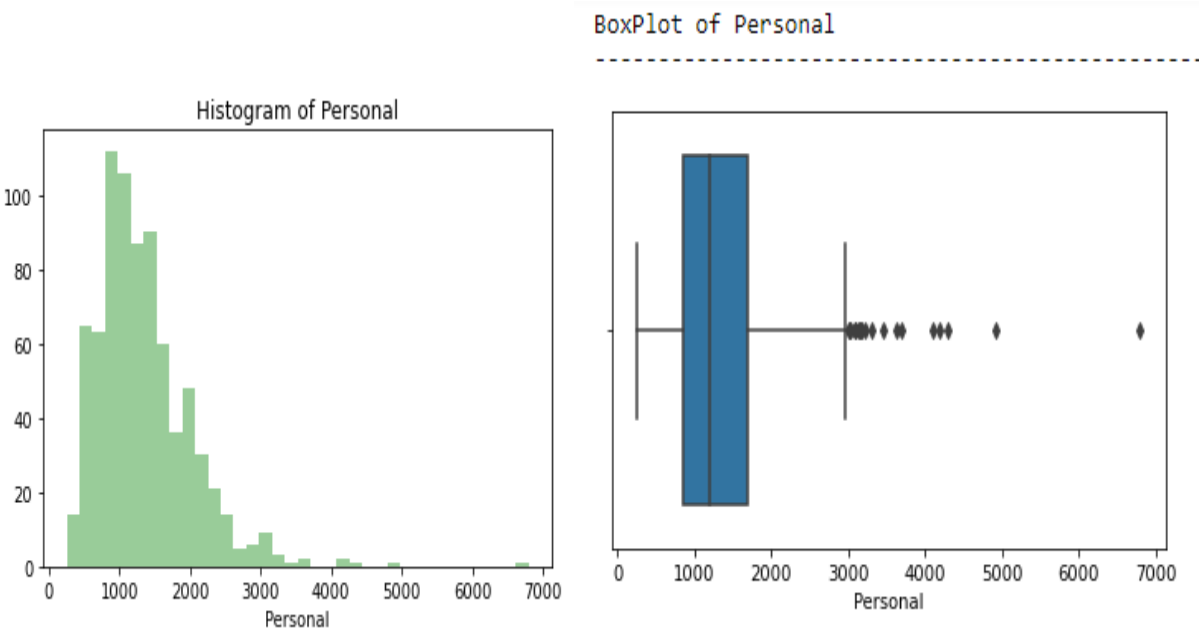

Fig 1.11: The box plot and the histogram showing the distribution of data of 'Personal' variable

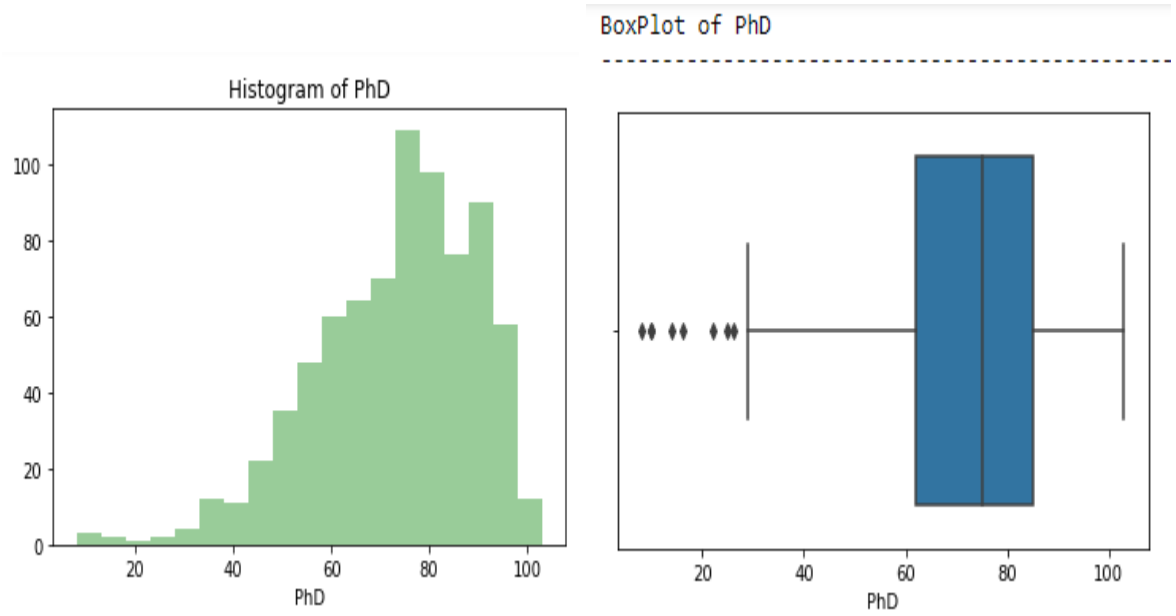

Fig 1.12: The box plot and the histogram showing the distribution of data of 'PhD' variable

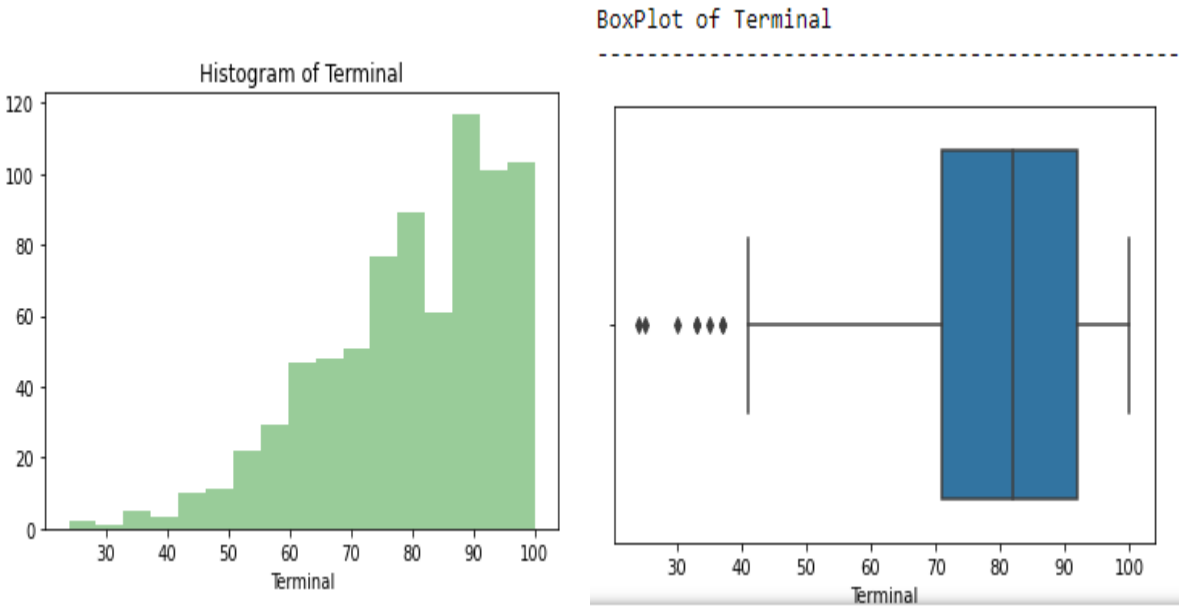

Fig 1.13: The box plot and the histogram showing the distribution of data of 'Terminal' variable

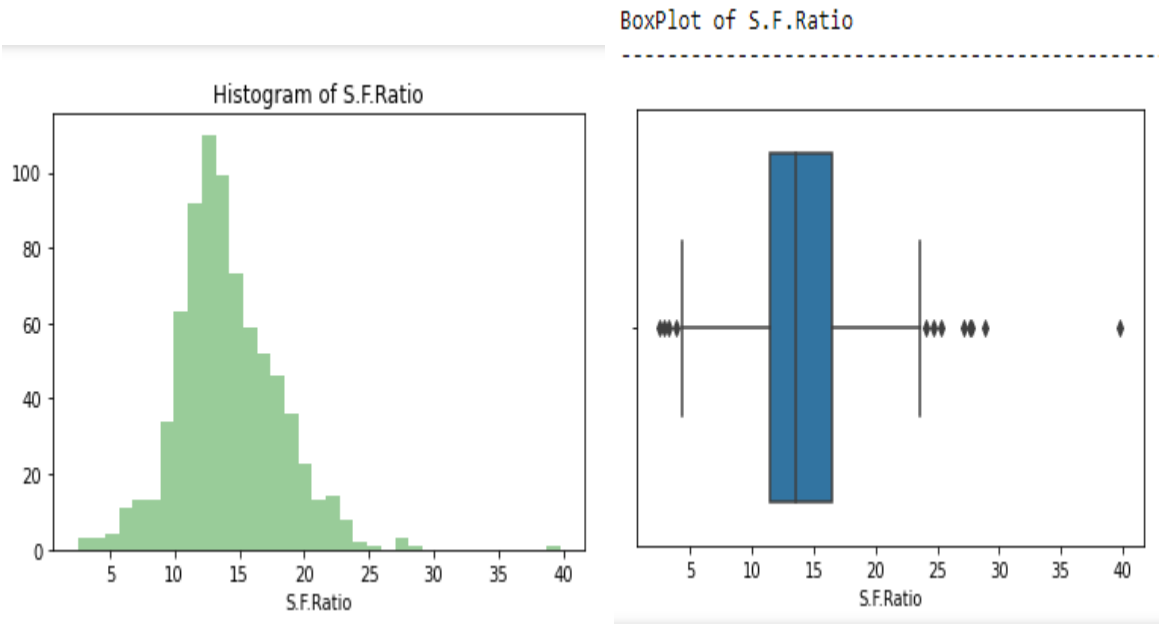

Fig 1.14: The box plot and the histogram showing the distribution of data of 'S.F.Ratio' variable

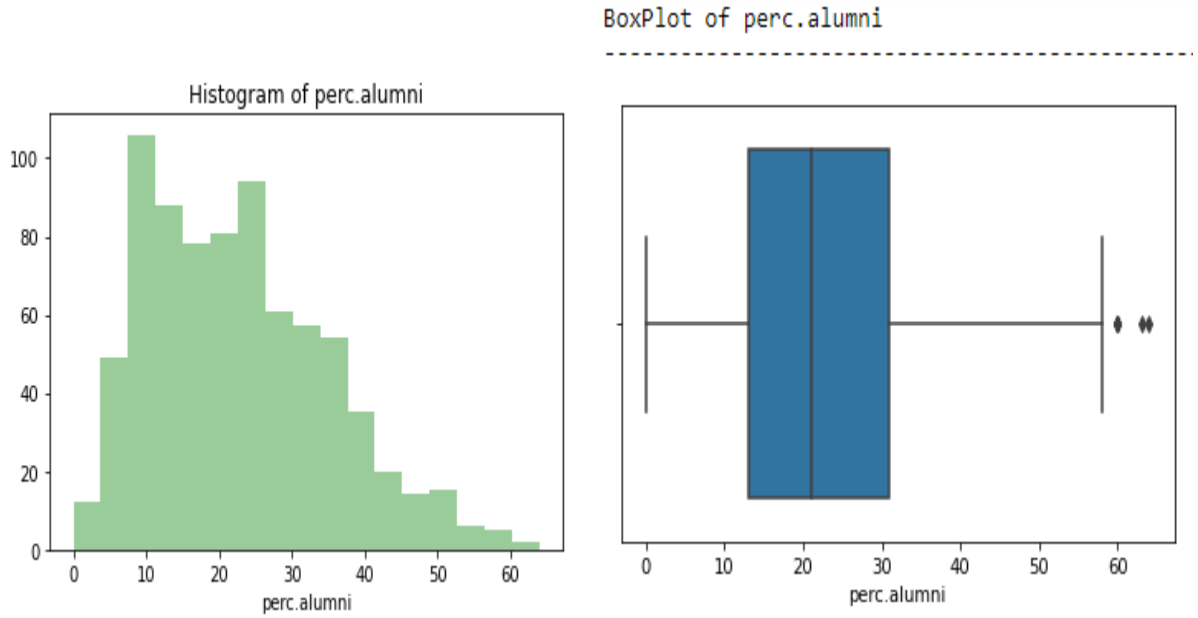

Fig 1.15: The box plot and the histogram showing the distribution of data of 'perc.alumni' variable

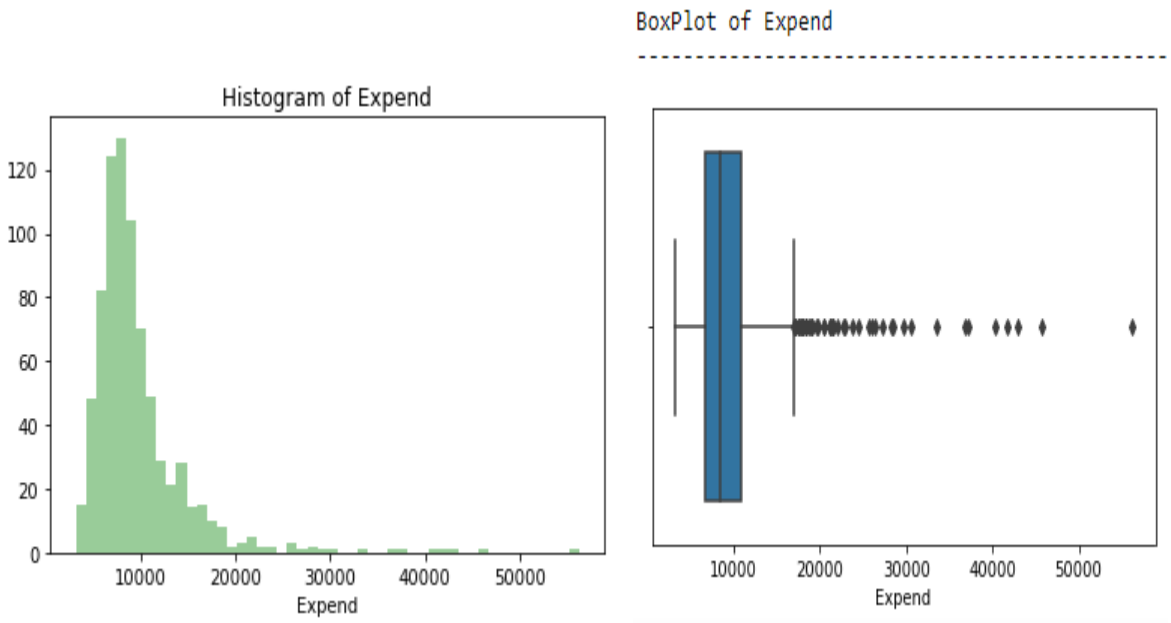

Fig 1.16: The box plot and the histogram showing the distribution of data of 'Expend' variable

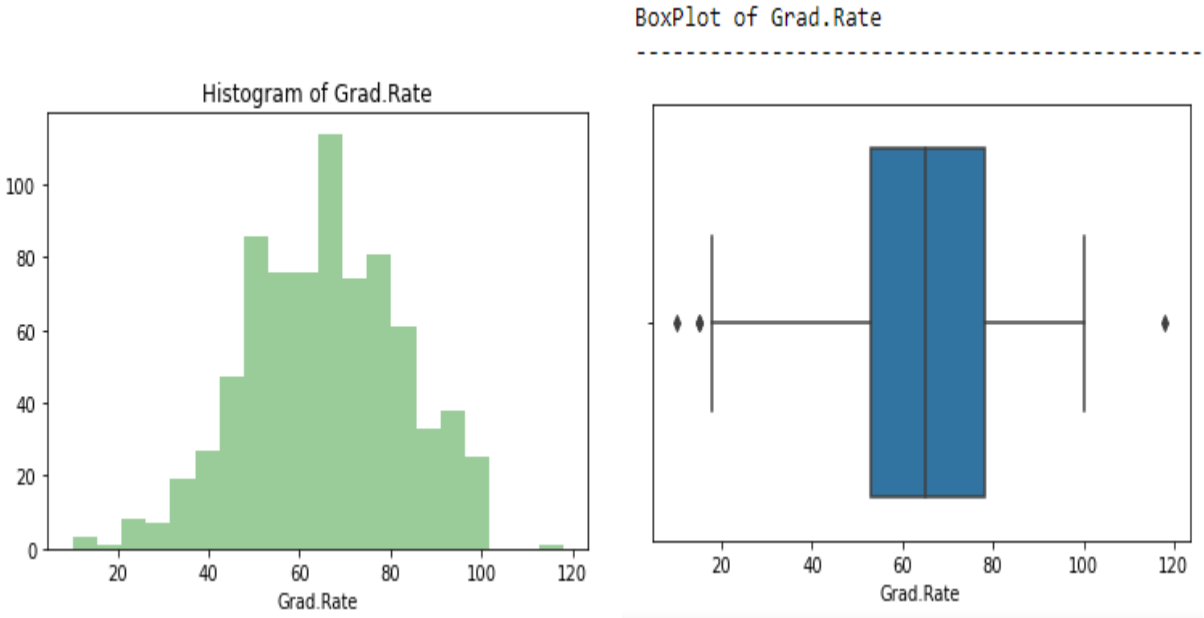

Fig 1.17: The box plot and the histogram showing the distribution of data of 'Grad.Rate' variable

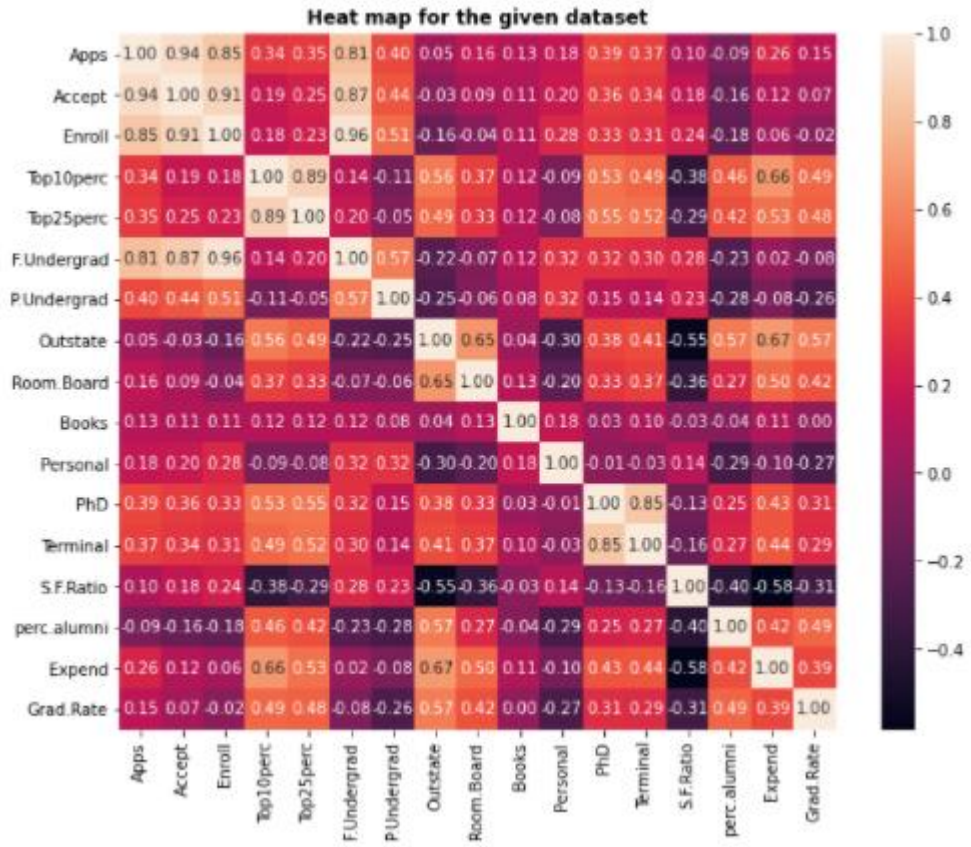

Fig 1.18: Heat map is portrayed for the multi variate analysis of the data.

Insights:

1. From the univariate analysis, we can understand that the category ‘Top25 perc’ is normally distributed and doesn’t have any outliers.
2. Also, the data in each of the columns namely, ‘Outstate’, ’Room Board’, ’Grad. Rate’,’ S.F.Ratio’, ’Perc. Alumni’ shows normal distribution in spite of having some outliers.
3. The distribution of the data in each of the columns namely ‘Ph.D’ and ‘Terminal’ are shown to be negatively skewed.
4. The distribution of the data in each of the columns namely ‘Accept’,’Enroll’,’F. Undergrad’, ’P.Undergrad’, ’Personal’, ’Books’, ’Expend’ are shown to be positively skewed.
5. Heat map gives the correlation between the two numerical values. From the heat map, we can conclude that the Application variable ‘Apps’ is highly positively correlated with application accepted, students enrolled and full-time graduates. This correlation can be interpreted in the way that when a student submits the application and get accepted, it means

that there is a high chance that the student has enrolled as a ful-time graduate.

6. We can find a negative correlation between application and perc.alumni. This indicates that not all students are part of the alumni of their college or university.

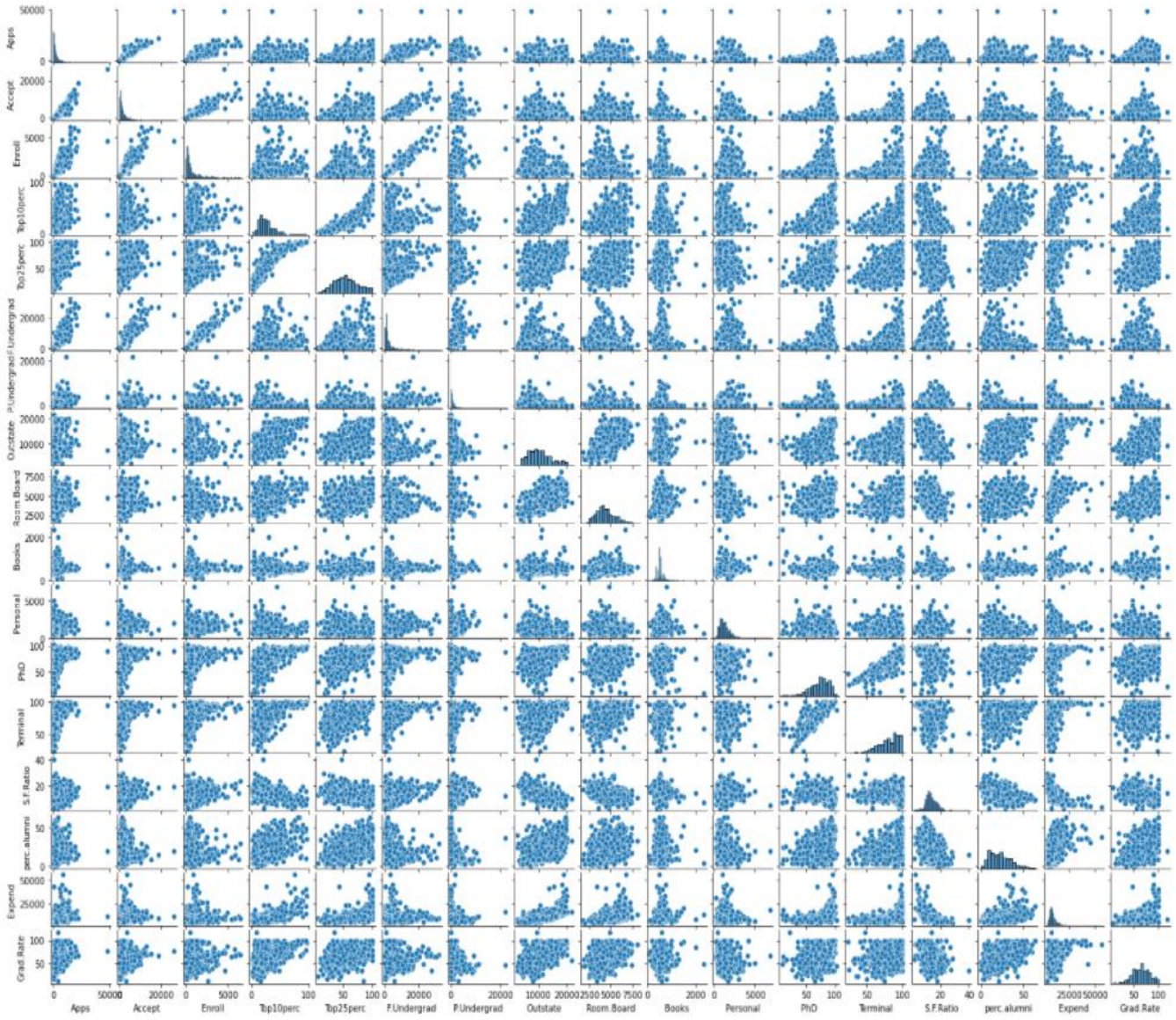

Fig 1.19: The pair plot showing the multivariate analysis

Q2.2. Is scaling necessary for PCA in this case? Give justification and perform scaling.

Justification:

Before scaling, the 'names' variable which is a categorical variable is dropped. Now the dataset consists of only numerical values. Now, if we have a look at our data set we can see that all the data is of different format. For example, 'top10perc','top25perc','PhD' and "Terminal','perc.alumni' are showing the percentage of the students while the 'Room. Board', 'Expend','books' and 'Personal' are values of money. Similarly,'S.F.Ratio' and 'Grad.Rate' are of ratio kind while the rest are of numerical kind. The algorithm will treat all these different numerical data as mere numbers and do the PCA which won't be correct. To ensure equal treatment of data, we center the data(scale) using Z-score technique.

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

Table 3: The head of the dataset that has been scaled

Q2.3. Comment on the comparison between the covariance and the correlation matrices from this data. [on scaled data]

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
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	0.369491	0.095633	-0.090226	0.259592	0.146755
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	0.337583	0.176229	-0.159990	0.124717	0.067313
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	0.308274	0.237271	-0.180794	0.064169	-0.022341
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	0.491135	-0.384875	0.455485	0.660913	0.494989
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	0.524749	-0.294629	0.417864	0.527447	0.477281
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	0.300019	0.279703	-0.229462	0.018652	-0.078773
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	0.141904	0.232531	-0.280792	-0.083568	-0.257001
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	0.407983	-0.554821	0.566262	0.672779	0.571290
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	0.374540	-0.362628	0.272363	0.501739	0.424942
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	0.099955	-0.031929	-0.040208	0.112409	0.001061
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	-0.030613	0.136345	-0.285968	-0.097892	-0.269344
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	0.849587	-0.130530	0.249009	0.432762	0.305038
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	1.000000	-0.160104	0.267130	0.438799	0.289527
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	-0.160104	1.000000	-0.402929	-0.583832	-0.306710
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	0.267130	-0.402929	1.000000	0.417712	0.490898
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	0.438799	-0.583832	0.417712	1.000000	0.390343
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	0.289527	-0.306710	0.490898	0.390343	1.000000

Table 4: The above table shows the correlation matrix

```
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 5: The above table shows the covariance matrix

Correlation matrix defines the quantity of strength(how much) and direction of the linear relationship between two variables. Strength means positively or negatively skewed. The correlation matrix before and after scaling remains the same. From the above correlation matrix, we can understand that the columns 'Apps','Accept','Enroll','F.Undergrad' ,'top25perc' and 'top10perc' are highly positively correlated.

Covariance indicates the proportionality between the variables; if the two variables are directly or inversely proportional to each other. Both covariance and correlation measure the relationship and dependency between two variables.

Q2.4. Check the dataset for outliers before and after scaling. What insight do you derive here?

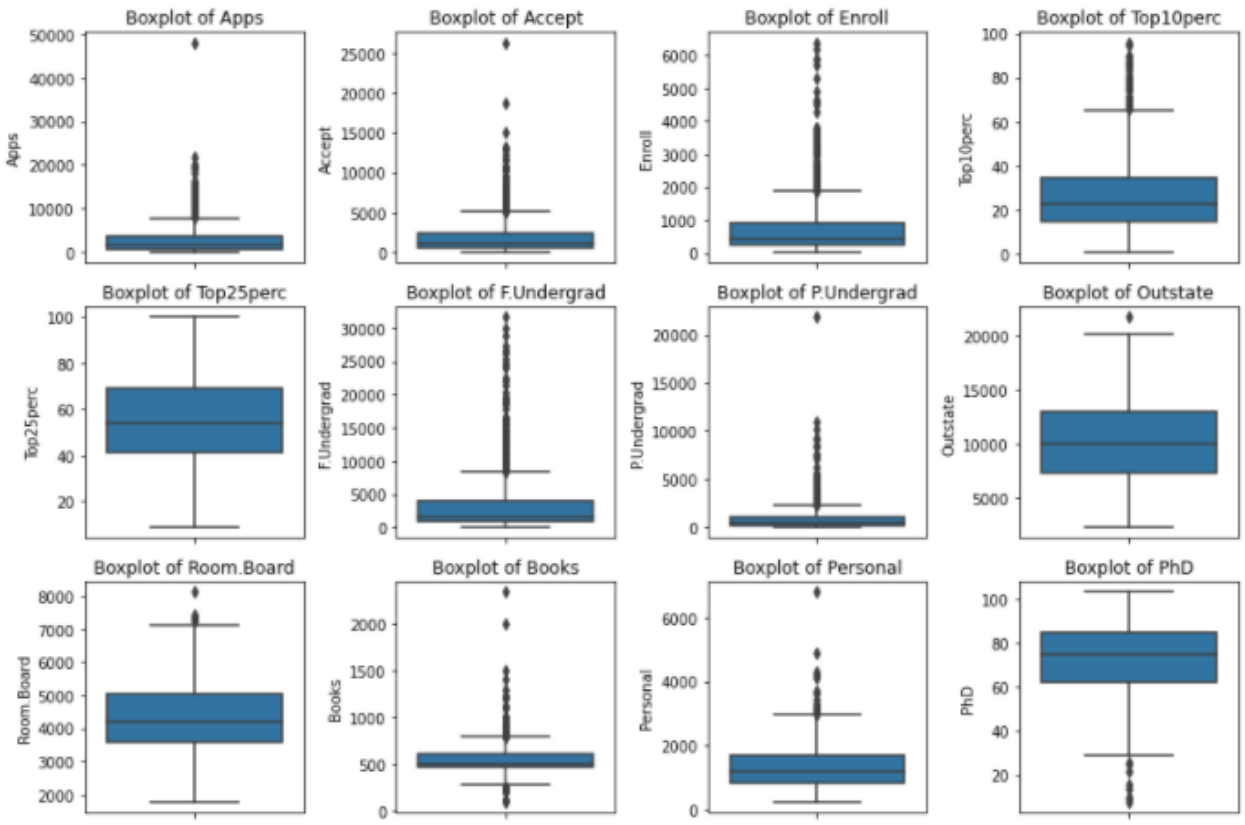

Fig 2: The box plots shown the data before scaling.

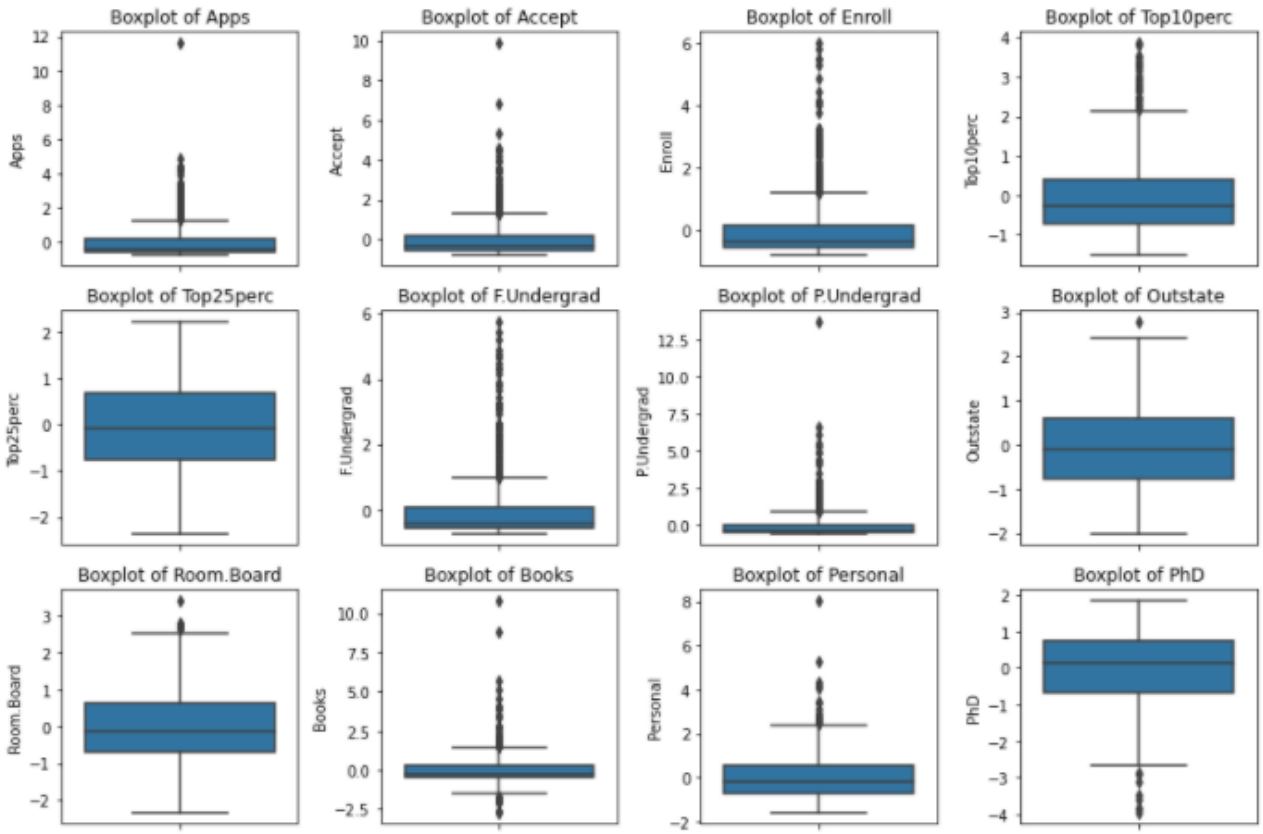

Fig 3: The boxplot shows the data after scaling

Insight:

Outliers are still present in the data. From this we can infer that the scaling of data doesn't remove the outliers in the data variables. Scaling just scales the value using Z-score technique and center the data.

Q2.5. Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both]

```
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]
```

The above given is the set of eigen values of the scaled data set.

```
Eigen Vectors
 %s [[-2.48765602e-01  3.31598227e-01 -6.30921033e-02  2.81310530e-01
  -5.74140964e-03 -1.62374420e-02 -4.24863486e-02 -1.03090398e-01
  -9.02270802e-02  5.25098025e-02 -3.58970400e-01  4.59139498e-01
  -4.30462074e-02  1.33405806e-01 -8.06328039e-02 -5.95830975e-01
   2.40709086e-02]
 [-2.07601502e-01  3.72116750e-01 -1.01249056e-01  2.67817346e-01
  -5.57860920e-02  7.53468452e-03 -1.29497196e-02 -5.62709623e-02
  -1.77864814e-01  4.11400844e-02  5.43427250e-01 -5.18568789e-01
   5.84055850e-02 -1.45497511e-01 -3.34674281e-02 -2.92642398e-01
  -1.45102446e-01]
 [-1.76303592e-01  4.03724252e-01 -8.29855709e-02  1.61826771e-01
   5.56936353e-02 -4.25579803e-02 -2.76928937e-02  5.86623552e-02
  -1.28560713e-01  3.44879147e-02 -6.09651110e-01 -4.04318439e-01
   6.93988831e-02  2.95896092e-02  8.56967180e-02  4.44638207e-01
   1.11431545e-02]
 [-3.54273947e-01 -8.24118211e-02  3.50555339e-02 -5.15472524e-02
   3.95434345e-01 -5.26927980e-02 -1.61332069e-01 -1.22678028e-01
   3.41099863e-01  6.40257785e-02  1.44986329e-01 -1.48738723e-01
   8.10481404e-03  6.97722522e-01  1.07828189e-01 -1.02303616e-03
   3.85543001e-02]
 [-3.44001279e-01 -4.47786551e-02 -2.41479376e-02 -1.09766541e-01
   4.26533594e-01  3.30915896e-02 -1.18485556e-01 -1.02491967e-01
   4.03711989e-01  1.45492289e-02 -8.03478445e-02  5.18683400e-02
   2.73128469e-01 -6.17274818e-01 -1.51742110e-01 -2.18838802e-02
  -8.93515563e-02]
 [-1.54640962e-01  4.17673774e-01 -6.13929764e-02  1.00412335e-01
   4.34543659e-02 -4.34542349e-02 -2.50763629e-02  7.88896442e-02
  -5.94419181e-02  2.08471834e-02  4.14705279e-01  5.60363054e-01
   8.11578181e-02  9.91640992e-03  5.63728817e-02  5.23622267e-01
   5.61767721e-02]
 [-2.64425045e-02  3.15087830e-01  1.39681716e-01 -1.58558487e-01
  -3.02385408e-01 -1.91198583e-01  6.10423460e-02  5.70783816e-01
   5.60672902e-01 -2.23105808e-01 -9.01788964e-03 -5.27313042e-02
  -1.00693324e-01  2.09515982e-02 -1.92857500e-02 -1.25997650e-01
  -6.35360730e-02]
 [-2.94736419e-01 -2.49643522e-01  4.65988731e-02  1.31291364e-01
  -2.22532003e-01 -3.00003910e-02  1.08528966e-01  9.84599754e-03
  -4.57332880e-03  1.86675363e-01 -5.08995918e-02  1.01594830e-01
  -1.43220673e-01  3.83544794e-02  3.40115407e-02  1.41856014e-01
  -8.23443779e-01]
 [-2.49030449e-01 -1.37808883e-01  1.48967389e-01  1.84995991e-01
  -5.60919470e-01  1.62755446e-01  2.09744235e-01 -2.21453442e-01
   2.75022548e-01  2.98324237e-01 -1.14639620e-03 -2.59293381e-02
   3.59321731e-01  3.40197083e-03  5.84289756e-02  6.97485854e-02
   3.54559731e-01]
 [-6.47575181e-02  5.63418434e-02  6.77411649e-01  8.70892205e-02
   1.27288825e-01  6.41054950e-01 -1.49692034e-01  2.13293009e-01
  -1.33663353e-01 -8.20292186e-02 -7.72631963e-04  2.88282896e-03
  -3.19400370e-02 -9.43887925e-03  6.68494643e-02 -1.14379958e-02
  -2.81593679e-02]
 [ 4.25285386e-02  2.19929218e-01  4.99721120e-01 -2.30710568e-01
   2.22311021e-01 -3.31398003e-01  6.33790064e-01 -2.32660840e-01
  -9.44688900e-02  1.36027616e-01  1.11433396e-03 -1.28904022e-02
   1.85784733e-02 -3.09001353e-03 -2.75286207e-02 -3.94547417e-02
  -3.92640266e-02]
 [-3.18312875e-01  5.83113174e-02 -1.27028371e-01 -5.34724832e-01
  -1.40166326e-01  9.12555212e-02 -1.09641298e-03 -7.70400002e-02
  -1.85181525e-01 -1.23452200e-01 -1.38133366e-02  2.98075465e-02
  -4.03723253e-02 -1.12055599e-01  6.91126145e-01 -1.27696382e-01
   2.32224316e-02]
 [-3.17056016e-01  4.64294477e-02 -6.60375454e-02 -5.19443019e-01
  -2.04719730e-01  1.54927646e-01 -2.84770105e-02 -1.21613297e-02
  -2.54938198e-01 -8.85784627e-02 -6.20932749e-03 -2.70759809e-02
   5.89734026e-02  1.58909651e-01 -6.71008607e-01  5.83134662e-02
   1.64850420e-02]
 [ 1.76957895e-01  2.46665277e-01 -2.89848401e-01 -1.61189487e-01
   7.93882496e-02  4.87045875e-01  2.19259358e-01 -8.36048735e-02
   2.74544380e-01  4.72045249e-01  2.22215182e-03 -2.12476294e-02
  -4.45000727e-01 -2.08991284e-02 -4.13740967e-02  1.77152700e-02
  -1.10262122e-02]
 [-2.05082369e-01 -2.46595274e-01 -1.46989274e-01  1.73142230e-02
   2.16297411e-01 -4.73400144e-02  2.43321156e-01  6.78523654e-01
  -2.55334907e-01  4.22999706e-01  1.91869743e-02  3.33406243e-03
   1.30727978e-01 -8.41789410e-03  2.71542091e-02 -1.04088088e-01
   1.82660654e-01]
 [-3.18908750e-01 -1.31689865e-01  2.26743985e-01  7.92734946e-02
  -7.59581203e-02 -2.98118619e-01 -2.26584481e-01 -5.41593771e-02
  -4.91388809e-02  1.32286331e-01  3.53098218e-02 -4.38803230e-02
  -6.92088870e-01 -2.27742017e-01 -7.31225166e-02  9.37464497e-02
   3.25982295e-01]
 [-2.52315654e-01 -1.69240532e-01 -2.08064649e-01  2.69129066e-01
   1.09267913e-01  2.16163313e-01  5.59943937e-01 -5.33553891e-03
   4.19043052e-02 -5.90271067e-01  1.30710024e-02 -5.00844705e-03
  -2.19839000e-01 -3.39433604e-03 -3.64767385e-02  6.91969778e-02
   1.22106697e-01]]
```

Table 6: The above given table shows the list of eigen vectors.

Q2.6. Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
PC1	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
PC2	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	0.246665	-0.246595	-0.131690	-0.169241
PC3	-0.063092	-0.101249	-0.082986	0.035056	-0.024148	-0.061393	0.139682	0.046599	0.148967	0.677412	0.499721	-0.127028	-0.066038	-0.289848	-0.146989	0.226744	-0.208065
PC4	0.281311	0.267817	0.161827	-0.051547	-0.109767	0.100412	-0.158558	0.131291	0.184996	0.087089	-0.230711	-0.534725	-0.519443	-0.161189	0.017314	0.079273	0.269129
PC5	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	-0.079388	-0.216297	0.075958	-0.109268
PC6	-0.016237	0.007535	-0.042558	-0.052693	0.033092	-0.043454	-0.191199	-0.030000	0.162755	0.641055	-0.331398	0.091256	0.154928	0.487046	-0.047340	-0.298119	0.216163
PC7	-0.042486	-0.012950	-0.027693	-0.161332	-0.118486	-0.025076	0.061042	0.108529	0.209744	-0.149692	0.633790	-0.001096	-0.028477	0.219259	0.243321	-0.226584	0.559944
PC8	-0.103090	-0.056271	0.058662	-0.122678	-0.102492	0.078890	0.570784	0.009846	-0.221453	0.213293	-0.232661	-0.077040	-0.012161	-0.083605	0.678524	-0.054159	-0.005336

Table 7: PCA is performed and the data is exported into the data frame with original features.

It is noticed that after performing PCA, multicollinearity is highly reduced. It is evident through the heat map portrayed below.

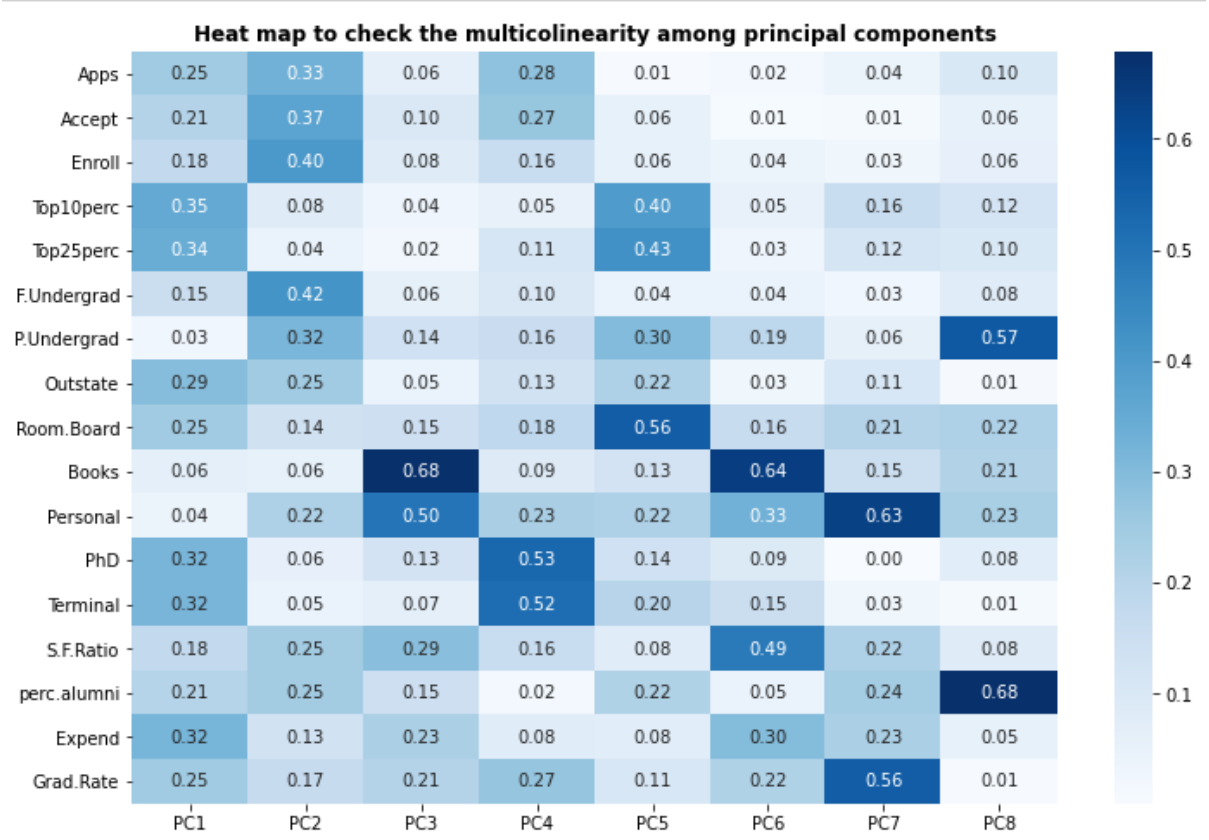

Fig 4: The heat map to portray the multicollinearity after the PCA is applied to the data set.

Q2.7 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]

```
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])
```

The above shown is the explicit form of the first PC.

```
the linear equation of first 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*R
oom.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+
```

The above shown is the linear equation of the first component.

Eigen vectors help us to understand which variable has more weightage and greater influences on the dataset of the principal components.

Q2.8. 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?

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

The above given array shows the cumulative values of the eigenvalues.

Adding all the eigen values, we will get a sum of 1. Eigen values and eigen vectors are the core of PCA and they are obtained through covariance or correlation matrix. The eigen vectors or the principal components determine the directions of the new feature space, and the eigenvalues determine their magnitude. Eigen vectors help us to understand which variable has more weightage and greater influences on the dataset of the principal components. We sort the principal components on the basis of the number of eigen values. Larger the number of eigen values, more important the principal component is. Principal component 1 will always be the greatest among all and has the maximum information captured.

Q2.9. 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]

Eight Principal components was found to be helpful in future analysis. While analysing the scree plot we can see that after the principal component 8, the variance is seen to be somewhat constant. So, we conclude that these 8 components have a larger contribution to the variance in the dataset while the rest of the principal components have negligible contribution.

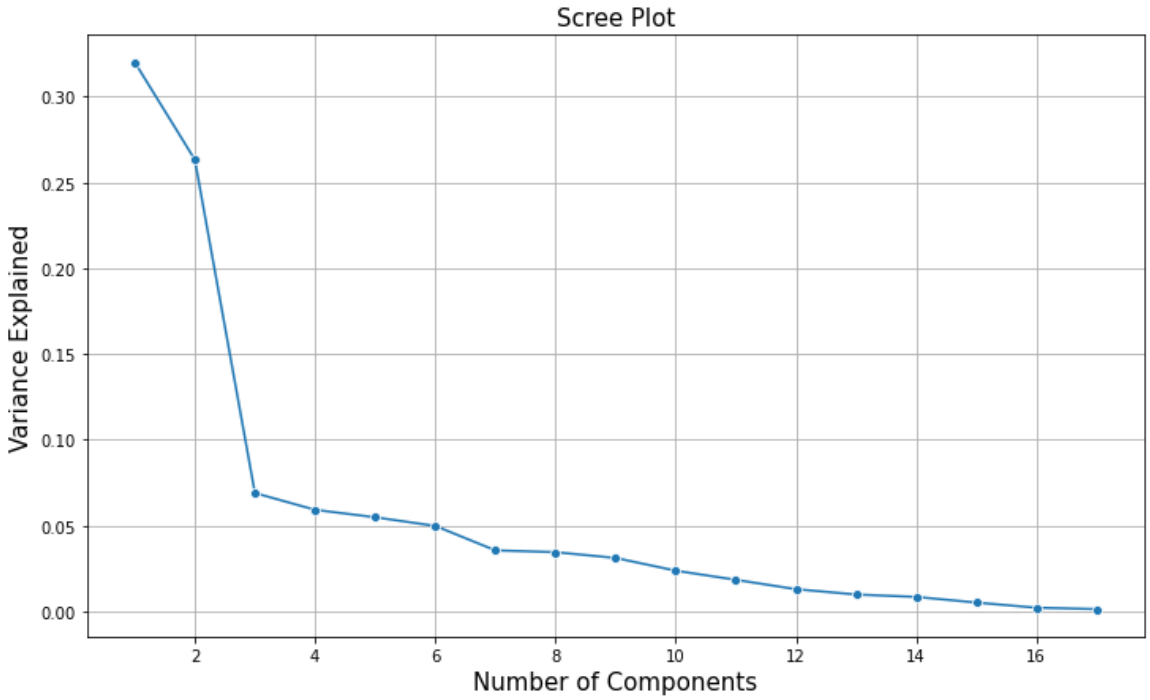

Fig 5: Scree plot to determine the number of principal components to be selected based on their contribution to the variation in the dataset.

PCA is used for three main purposes:

1.Increase SNR-signal to noise ratio; i.e. it maximises the signal or information content that we get and reduce the noise levels in the signal. In short, it means to increase the signal content and suppress the noise content. Greater the SNR, better the model will be.

2. It is used to eliminate the dependency between the independent dimensions. PCA is a great tool for dimensionality reduction.

In the given dataset, its about the various details of the colleges and universities. For better understanding of the dataset, we did univariate and multivariate analysis. From these analyses we were able to understand the

skewness, distribution, datatypes of the variables, correlation of variables with each other. We did scaling to standardize the data entries of various variables to a common data format for better modelling and studies. PCA is done to reduce the multicollinearity and dimensionality reduction of the dataset. Depending on the variance of the dataset we can reduce the PCA components. The PCA components for this study case is 8 where we could understand the maximum variance of the dataset. Thus, we attained a more efficient dataset with reduced dimensions which when given to machine learning models give you a better result.

THE END