BUSINESS REPORT
ON
ADVANCED STATISTICS

Name: Yashveer Kothari. A

Project: Advanced Statistics

Course: PGP-DSBA-2022

Date:07th June 2022

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PROBLEM 1

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 whole aim of this exercise is to explore the data using ANOVA, BOTH One way and two ANOVA. We will analyse the data by applying different Hypothesis and proving if the hypothesis can be accepted or rejected and thus understand the Significant Inter -dependency of the variables on each other or not i.e., if there is any dependency of any variable on the other available variables.

DATA DESCRIPTION:

1. Education: The Education qualification of the individual consisting of High School Graduate, Bachelor and Doctorate. It is of Object Data Type consisting of 40 entries
2. Occupation: The Occupation of the individual consisting four levels Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. It also Object Data Type and consists of 40 entries.
3. Salary: The salary of the individuals. It consists of 40 entries and is of Integer Data Type.

SAMPLE 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

	Education	Occupation	Salary
35	Bachelors	Exec-managerial	173935
36	Bachelors	Exec-managerial	212448
37	Bachelors	Exec-managerial	173664
38	Bachelors	Exec-managerial	212760
39	Doctorate	Exec-managerial	212781

Table 1: Data Set Sample

- The Dataset has 40 entries, 3 Variables.

DESCRIBING 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: Data Description

- There are 40 entries in all the variables.
- There are a greater number of people with qualifications as a doctor.
- There are a greater number of people with occupation as Prof-Speciality
- The Highest salary is 2,60,151 and lowest being 50,103.

DATASET INFORMATION:

```
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)
```

- There are a total of 40 entries, 3 variables namely Education, Occupation and salary, where Salary is Integer Data Type and other two are Object Data Type.
- There are 40 rows and 3 Columns.

IDENTIFYING THE MISSING VALUES:

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

- There are no missing values in any of the variables.

COUNTING THE DIFFERENT LEVELS IN EDUCATION:

```
Doctorate    16
Bachelors    15
HS-grad       9
```

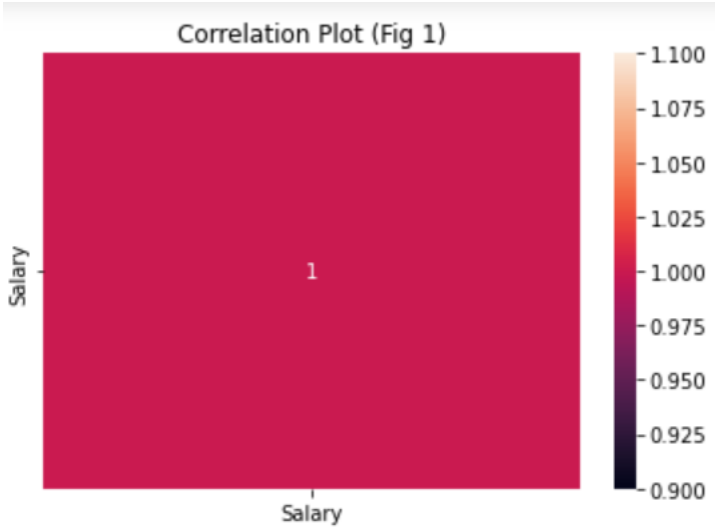
- There is a total of 16 Doctors, 15 Bachelors and 9 High school Graduates in variable Education.

COUNTING THE DIFFERENT LEVELS IN OCCUPATION:

```
Prof-specialty     13
Sales              12
Adm-clerical       10
Exec-managerial     5
```

- There is a total of 13 Prof-Speciality, 12 Sales, 10 Adm-clerical, 5 Exec- Managerial.
- In salary, every person has different salary.

CORRELATION PLOT:



Correlation Plot

There is no correlation between the variables.

PAIRPLOT:

Fig 2: Salary Pair plot

- Since there is no correlation there is no pair plot except for salary itself

PROBLEM 1A:

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

A. Hypothesis for Education using One way ANOVA:

Null Hypothesis HO: The mean salary value is same for all the categories of Education

Alternate Hypothesis HA: The mean salary value is different in at least one category of Education.

B. Hypothesis for Occupation using One way ANOVA:

Null Hypothesis HO: The mean salary value is same for all the categories of Occupation

Alternate Hypothesis HA: The mean salary value is different in at least one category of Occupation.

Q2. Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	30.95628	1.257709e-08
Residual	37.0	6.137256e+10	1.658718e+09	NaN	NaN

Table 3: One-Way ANOVA on Salary w.r.t Education

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 rej ect the null hypothesis and conclude that there is a significant difference in the mean salaries f or at least one category of education.

Q3. Perform a one-way ANOVA on Salary with respect to Occupation. State whether th e 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 4: One-Way ANOVA on Salary w.r.t Occupation

The above is the ANOVA table for Occupation variable

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

Q4. If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result. (Non-Graded)

PROBLEM 1B

Q1. What is the interaction between two treatments? Analyse the effects of one variable on the other (Education and Occupation) with the help of an interaction plot. [hint: use the 'point plot' function from the 'seaborn' function].

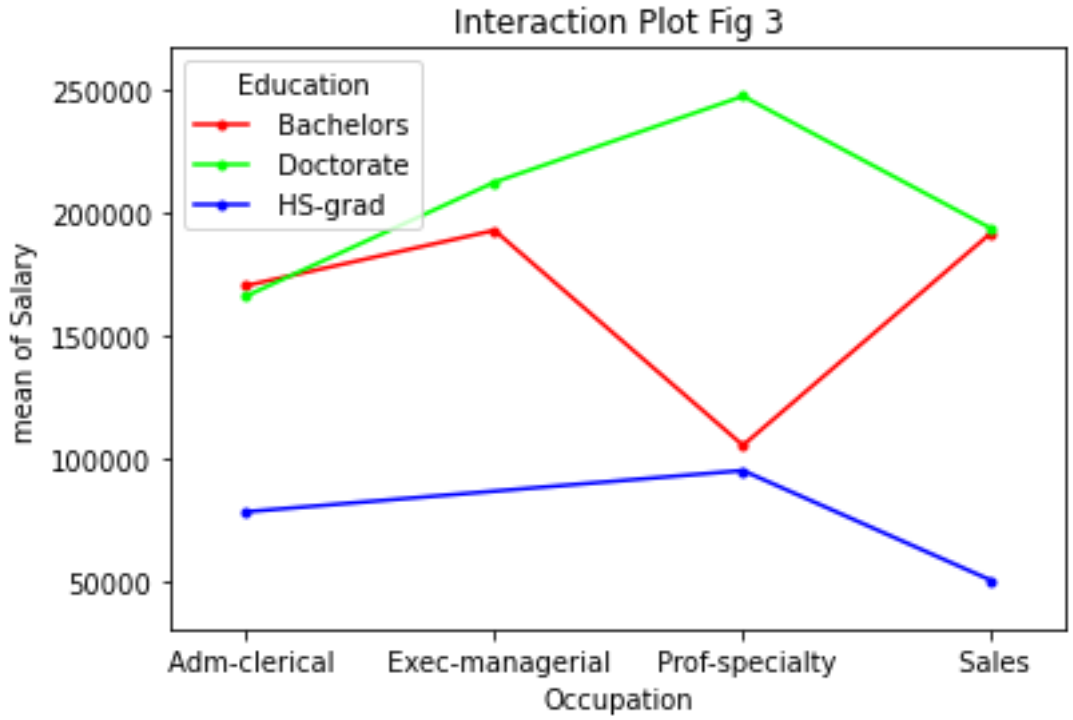


Fig 3: Interaction Plot for all Variables.

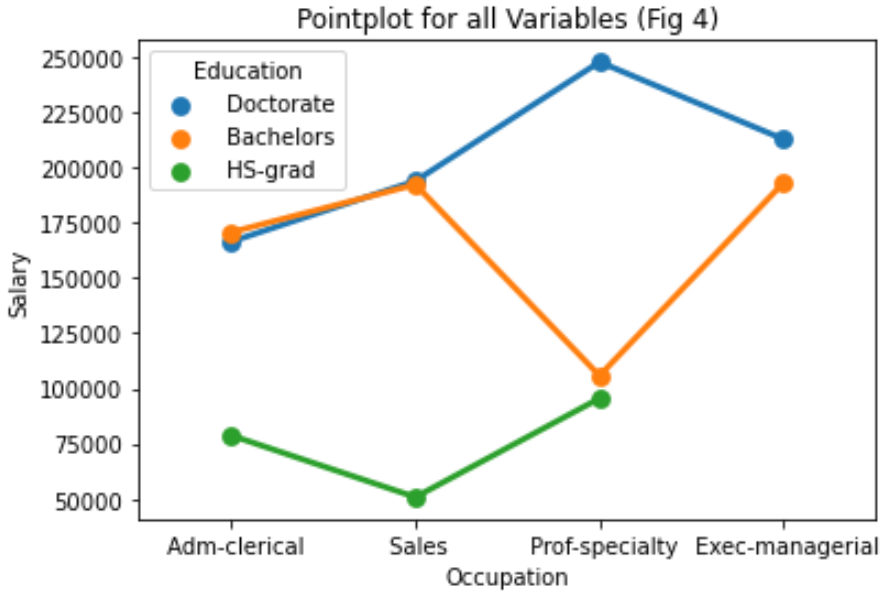


Fig 4: Point Plot for all variables.

Observation from the Interaction Plot Fig 3 and Point Plot Fig 4

1. People having a qualification as Bachelors and occupation as Sales and Executive Managerial have the same salaries

2. People having qualification of HS Grad have minimum salary

3. People with qualification of HS grad do not have occupation as Exec_Maagerial

4. People having Prof Speciality have the highest salaries

5. Sales Person with Bachelors or doctorate education earn similar salaries, earning higher than people with HS grad education.

6. Adm clerical with HS grad education earn the lowest comparatively.

7. 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.

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

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

10. 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.

11. Similarly, people with education as Bachelors and occupation as Prof-Specialty earn lesser than people with education as Bachelors and occupation Exec-Managerial whereas people with education as Doctorate and occupation as Prof-Specialty earn higher than people with education as Doctorate and occupation Exec-Managerial.

Q2. Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?

Null Hypothesis H0: There is no interaction effect between the independent variables, education and occupation on the mean salary.

Alternate Hypothesis H1: There is an interaction effect between the independent variable education and occupation on the mean salary.

	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 5: Two-way ANOVA on salary w.r.t both Education and Occupation.

INTERPRETATION:

Since the P-value of 2.232500e-05 is less than Alpha 0.05, we reject the null hypothesis and accept the alternate Hypothesis which states that the There is an interaction between Education and Occupation on the mean salary.

Q3. Explain the business implications of performing ANOVA for this particular case study.

From the ANOVA method and the interaction plot, we see that:

1. Education combined with occupation results in higher and better salaries among the people.

2. It is clearly seen that people with education as Doctorate earn maximum salaries and people with education HS-grad earn the least.

3. Thus, it can be concluded that Salary is dependent on educational qualifications and occupation.

PROBLEM 2:

EXECUTIVE SUMMARY:

The dataset contains information on various colleges. The application received by the colleges, the applications accepted, students enrolled in the colleges or universities, The Part time Graduates, Full- time Graduates, The cost of the course, personal cost and expenditure, the details about the faculties.

INTRODUCTION:

The main aim of this exercise is to explore the dataset using Univariate and Multi variate analysis, find whether the data set is significant to perform Principal component analysis using various methods and reduce the multi-collinearity in the dataset to build models and do further analysis efficiently.

DATA DESCRIPTION:

1) Names: Names of various university and colleges
2) Apps: Number of applications received
3) Accept: Number of applications accepted
4) Enrol: Number of new students enrolled
5) Top10perc: Percentage of new students from top 10% of Higher Secondary class
6) Top25perc: Percentage of new students from top 25% of Higher Secondary class
7) F. Undergrad: Number of full-time undergraduate students
8) P. Undergrad: Number of part-time undergraduate students
9) Outstate: Number of students for whom the particular college or university is Out-of-state tuition
10) Room. Board: Cost of Room and board
11) Books: Estimated book costs for a student
12) Personal: Estimated personal spending for a student
13) PhD: Percentage of faculties with Ph.D.'s
14) Terminal: Percentage of faculties with terminal degree
15) S.F. Ratio: Student/faculty ratio
16) perc. alumni: Percentage of alumni who donate

17) Expend: The Instructional expenditure per student
18) Grad. Rate: Graduation rate

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

SAMPLE DATASET:

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio
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

perc.alumni	Expend	Grad.Rate
12	7041	60
16	10527	56
30	8735	54
37	19016	59
2	10922	15

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio
772	Worcester State College	2197	1515	543	4	26	3089	2029	6797	3900	500	1200	60	60	21.0
773	Xavier University	1959	1805	695	24	47	2849	1107	11520	4960	600	1250	73	75	13.3
774	Xavier University of Louisiana	2097	1915	695	34	61	2793	166	6900	4200	617	781	67	75	14.4
775	Yale University	10705	2453	1317	95	99	5217	83	19840	6510	630	2115	96	96	5.8
776	York College of Pennsylvania	2989	1855	691	28	63	2988	1726	4990	3560	500	1250	75	75	18.1

perc.alumni	Expend	Grad.Rate
14	4469	40
31	9189	83
20	8323	49
49	40386	99
28	4509	99

Table 6: Data Set Sample

DATASET INFORMATION:

```
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
```

There are a total of 777 entries and 17 columns. The names column is the object type and S.F Ratio is Float Data Type whereas all other variables are Integer data type. The shape of the data set is 777 rows and 18 columns.

IDENTIFYING THE MISSING VALUES:

```
Names          0
Apps           0
Accept         0
Enroll         0
Top10perc      0
Top25perc      0
F.Undergrad    0
P.Undergrad    0
Outstate       0
Room.Board     0
Books          0
Personal       0
PhD            0
Terminal       0
S.F.Ratio      0
perc.alumni    0
Expend         0
Grad.Rate      0
```

DESCRIBING THE DATASET:

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

Table 7: Data Description

INTERPRETATION OF DESCRIPTIVE ANALYSIS:

1. There were a total of 48094 applications received in Rutgers at New Brunswick, with total average applications recd were 3001

2. The least application received were 81 received in Christendom College

3. A total of 26330 students were accepted which is the highest applications accepted among all universities (Rutgers at New Brunswick

4. The least application accepted were 72 applications in Christendom College.

5. The highest students enrolled were 6392 (Texas A&M Univ. at College Station) and lowest were 35 (Capitol College)

6. In Massachusetts Institute of Technology, there were 96 students who were students from the Top10% of Higher secondary class, whereas Centre of Creative Studies had only 1 student from the top 10% of Higher secondary class.

7. In Suny at Buffalo and University of California at Berkely (Highest number of students from Top 10 % Higher secondary class) there were 100 students who were from the Top 25% of Higher secondary class, whereas Huron University (Least number of students from top 10% Higher Secondary class) had only 9 students from the top 25% of the higher secondary class

8. In Texas A&M Univ. at College Station, there were 31643 full time students which is the highest number of Students opting for full time Undergraduate and Concordia College at St. Paul had the lowest students opting for Full time Under graduation which is 139 students.

9. 21836 students opted for Part Time under graduation in University of Minnesota Twin Cities (Highest among all the part time opting students), and 1 student from Claremont McKenna College opting for part time undergraduate (Lowest among all the part time opting students).

10. There were a total of 21700 students from outstate visiting Bennington College (It is the highest number of outstate students) and 2340 students visiting Brigham Young University at Provo which having the least number of Outstate Students.

11. The highest cost for Room Board is 8124 in Barnard College and least amount of 1780 in North Carolina A. & T. State University.

12. The highest estimated cost for a student is 2340 in Centre of Creative studies, and the lowest of 96 at Appalachian State University.

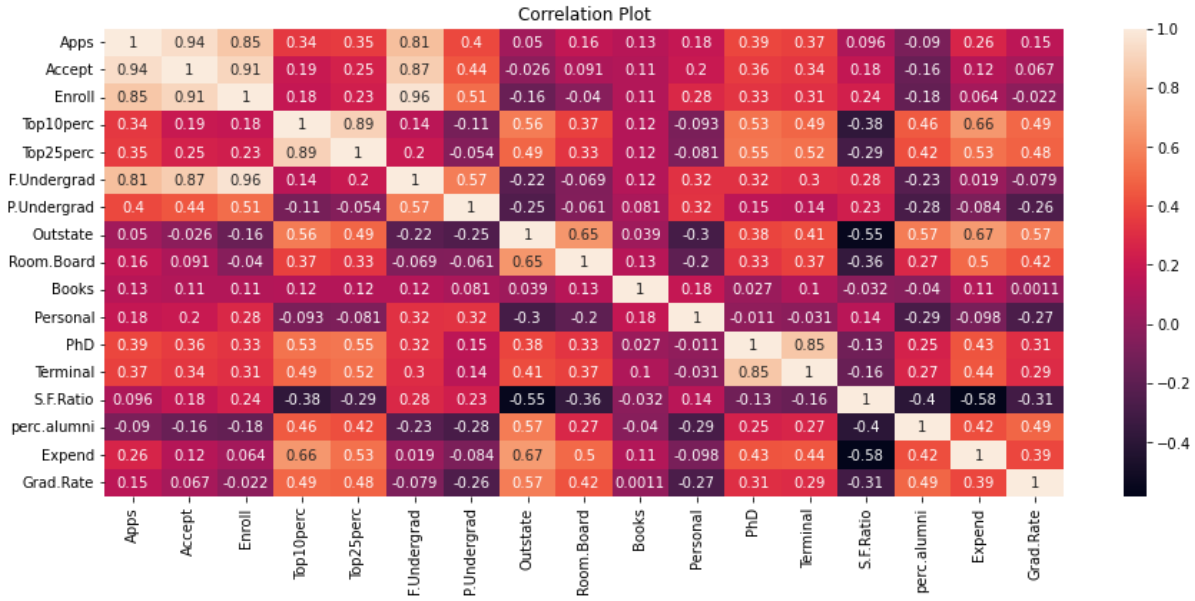
13. The Highest personal expenses is of 6800 (Saint Louis University) and the lowest is 250 (Benedictine College).

14. The percentage of Faculties with PHD's are highest in Texas A&M University at Galveston 103 faculties and lowest in Centre for Creative Studies 8 faculties

15. The number of faculties with terminal degrees are high in 14 universities with 100 faculties each and lowest in Salem-Teikyo University 24 Faculties.

16. The Highest Graduation rate is in Cazenovia College with 118% and Texas Southern University with 10 percentage having the lowest graduation rate.

PERFORMING MULTI-VARIATE ANALYSIS:



Correlation Plot -Multi Variate Analysis

INTERPRETATION OF THE CORRELATION PLOT:

The Correlation Plot or Heat Map gives us the correlation between two numerical values.

1. There is a higher correlation between the Applications received and the Applications accepted 0.91

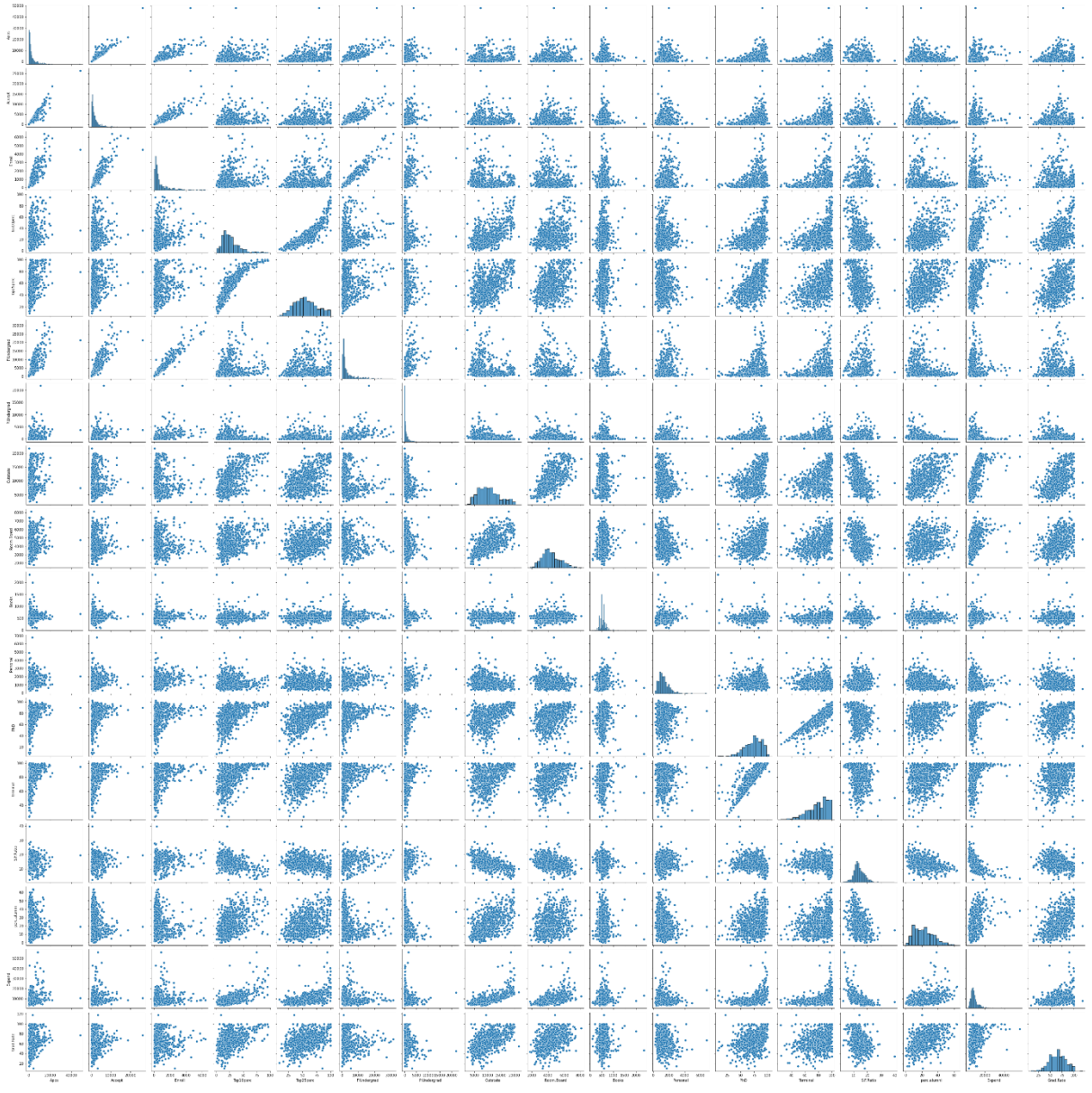
2. Similarly the correlation between the Applications accepted and Students enrolled is also high 0.94

3. There is no correlation between SF ratio and outstate -0.55, SF Ratio and Expend -0.58

4. There is a negative correlation between Application received and the Perc. Alumni. which indicates that not all students are a part of alumni of the college or University.

5. The applications with Top 10, 25 of Higher secondary class, outstate, room board, books, personal, PhD, terminal, expenditure and graduation ratio are positively skewed.

PAIRPLOT:

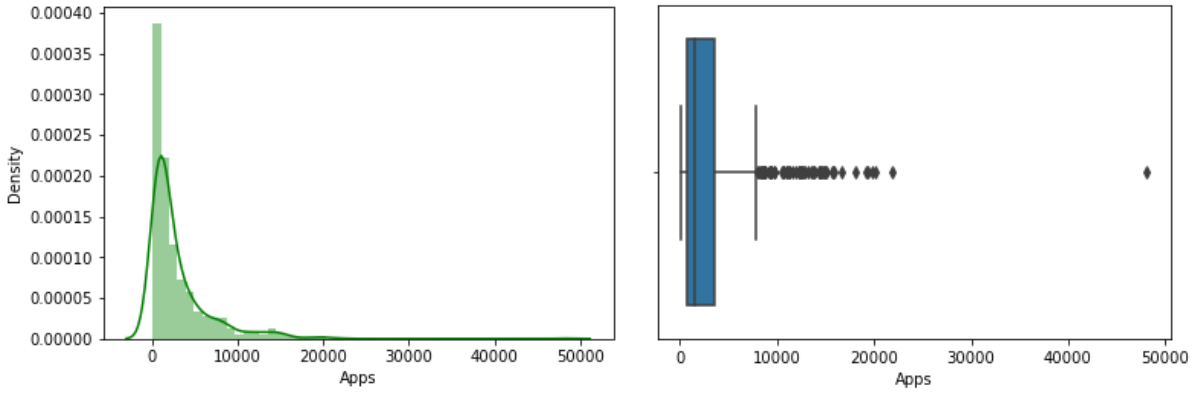


Pair Plot-Multi Variate Analysis.

PERFORMING UNI-VARIATE ANALYSIS FOR:

- Univariate Analysis: helps to understand the distribution of the data in the dataset. With Univariate analysis we can find different patterns, trends and summarize the data.
- Boxplot helps to identify outliers in data.
- Distribution Plot helps to identify the patterns of the data.
- APPS:

```
Description of Apps
-----------------------------------------------
count      777.000000
mean      3001.638353
std       3870.201484
min         81.000000
25%        776.000000
50%       1558.000000
75%       3624.000000
max      48094.000000
Name: Apps, dtype: float64 Distribution of Apps
-----------------------------------------------
```



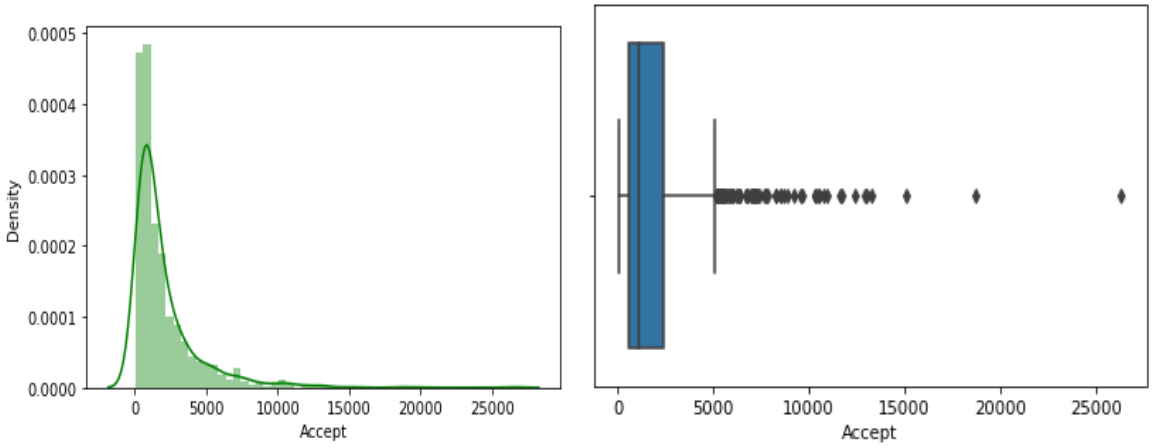
Distribution plot and Boxplot for Apps

INTERPRETATION FOR APPS:

Apps column have outliers, the Boxplot is Right skewed, The Inter quartile range ranges from 776 to 3624 and Median being 1558. The Outlier ranges from 7896 to 49000 approx. ((IQR*1.5) +Q3=Any point above the result is an outlier=Q3-Q1=3624-776= 2848 (2848*1.5) +3624=7896). The distribution of the data is skewed from which we can understand that each college or university offers application in the range of 3000 to 5000. The maximum application is around 50,000

- ACCEPT:

```
Description of Accept
-----------------------------------------------------
count      777.000000
mean      2018.804376
std       2451.113971
min         72.000000
25%        604.000000
50%       1110.000000
75%       2424.000000
max      26330.000000
Name: Accept, dtype: float64 Distribution of Accept
```



Distribution Plot and Boxplot for Accept.

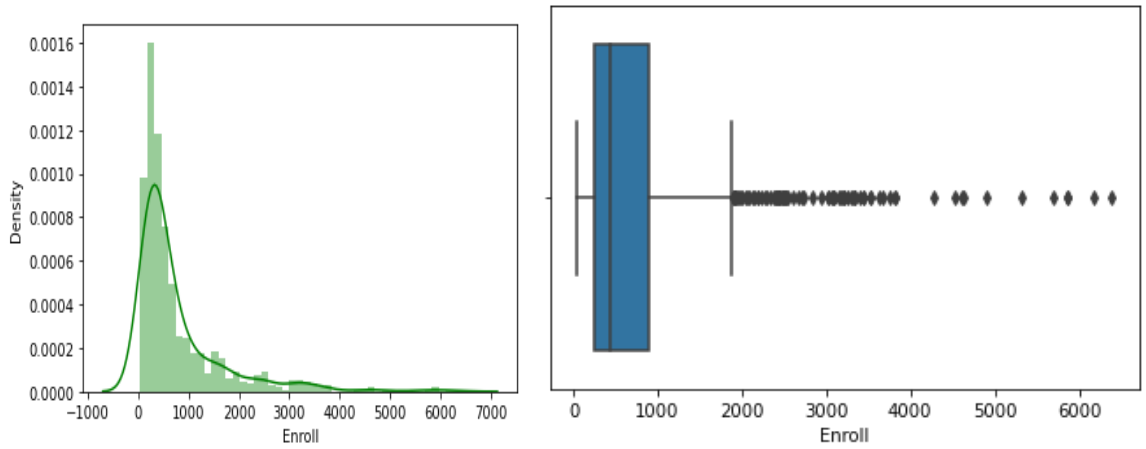
INTERPRETATION FOR ACCEPT:

Accept: Accept column has outliers, the boxplot is right skewed. The Inter quartile range ranges from 60 to 2424, median is 1110. The outlier ranges between 5970 to 25500 approx., (2424-60=2364, (2364*1.5) +2424 =5970).

The distribution Plot shows the majority of the applications accepted from each of the universities and colleges, ranging from 70 to 1500. The accept variable seems to be positively skewed.

- ENROLL:

```
Description of Enroll
-----------------------------------------------------
count     777.000000
mean      779.972973
std       929.176190
min        35.000000
25%       242.000000
50%       434.000000
75%       902.000000
max      6392.000000
Name: Enroll, dtype: float64 Distribution of Enroll
```



Distribution and Box Plot for Enrol

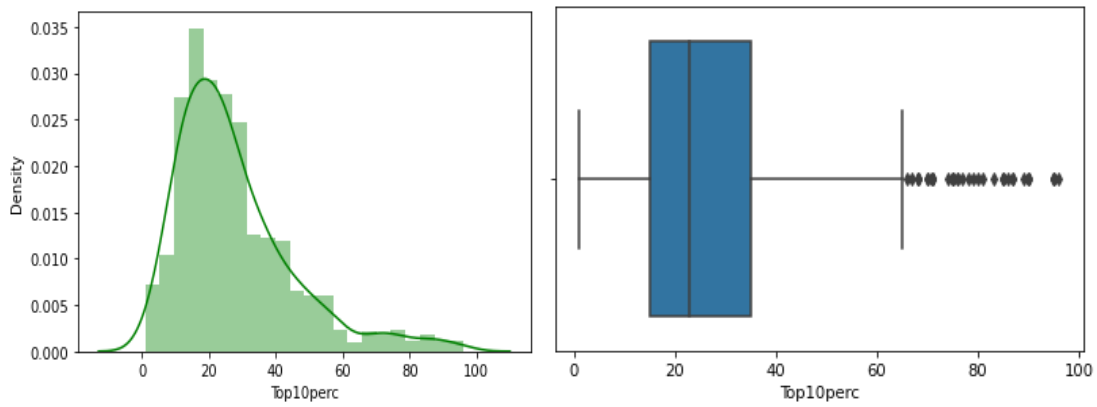
- **INTERPRETATION FOR ENROL:**

Enrol column has outliers, it is right skewed and ranges from 242 to 902 and Median is 434.The Outlier ranges between 1892 to 6600 approx. (IQR=902-242 =660 Outliers= (660*1.5) +902=1892).

The distribution plot is positively skewed and we can also understand that majority of the universities/Colleges have enrolled students in the range of 2000 to 5000.

- TOP 10 PERC:

```
Description of Top10perc
-------------------------
count    777.000000
mean      27.558559
std       17.640364
min        1.000000
25%       15.000000
50%       23.000000
75%       35.000000
max       96.000000
```



Distribution plot and Box plot for Top 10 perc

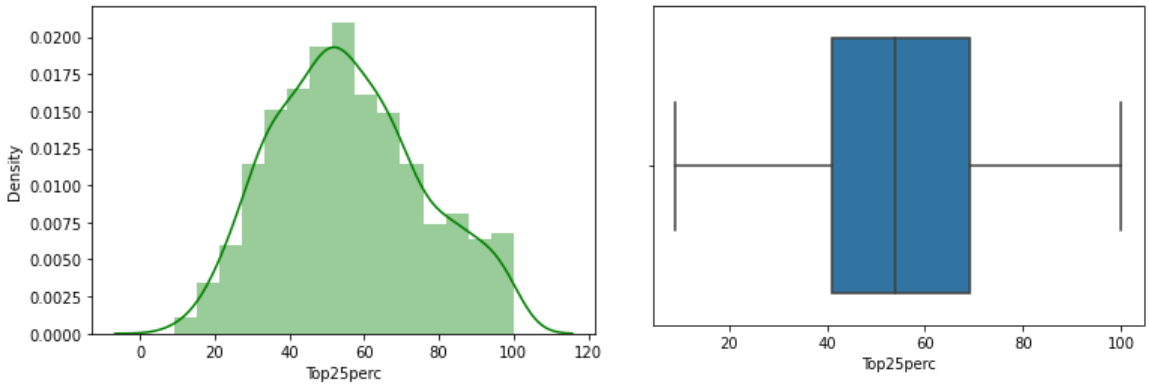
- **INTERPRETATION FOR TOP 10 PERC:**

Top 10 Perc: It has outliers, it is right skewed, The IQR ranges from 15 to 35 and Median is 23. The outliers range between 65 to 98 approx. (IQR=Q3-Q1= 35-15=20 and outliers= (20*1.5) +35=65)

The distribution seems to be positively skewed.. There are about 30 to 50 students that have taken up by the universities and colleges from these particular Top 10 high schools.

- TOP 25 PERC:

```
Description of Top25perc
----------------------------------------------------------
count    777.000000
mean      55.796654
std       19.804778
min        9.000000
25%       41.000000
50%       54.000000
75%       69.000000
max      100.000000
Name: Top25perc, dtype: float64 Distribution of Top25perc
```



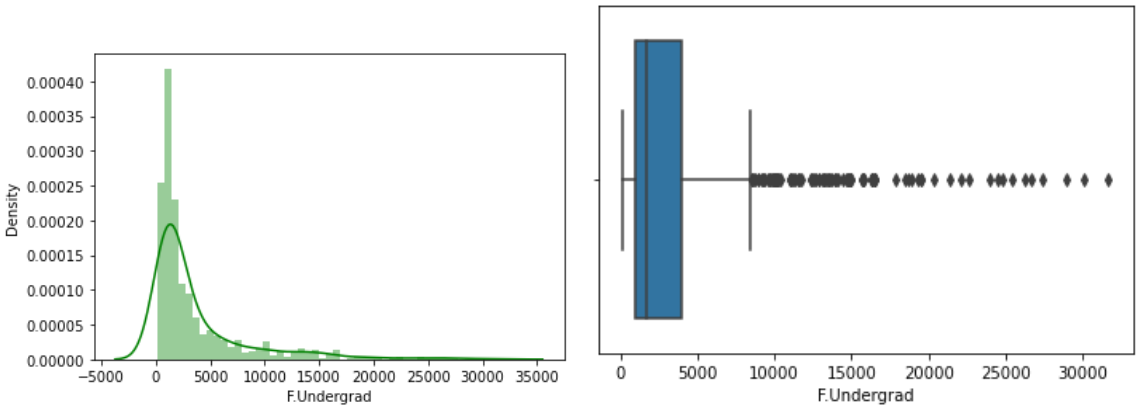
Distribution and Box Plot for Top 25 Perc

- **INTERPRETATION FOR TOP 25 PERC:**

Has no outliers and is normally distributed. IQR ranging between 41 to 69 and Median being 54. The distribution is almost normally distributed and majority of the students are from Top 25% of High schools.

- FULL TIME GRADUATE:

```
Description of F.Undergrad
-----------------------------------------------------------------
count      777.000000
mean      3699.907336
std       4850.420531
min        139.000000
25%        992.000000
50%       1707.000000
75%       4005.000000
max      31643.000000
Name: F.Undergrad, dtype: float64 Distribution of F.Undergrad
-----------------------------------------------------------------
```



Distribution plot and Box plot for Full Time Graduate

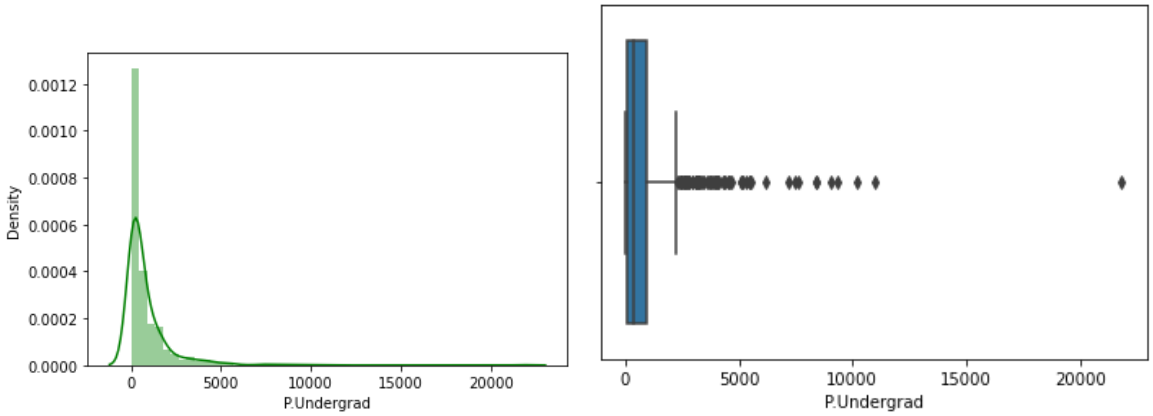
- **INTERPREATATION FOR FULL TIME GRADUATE:**

It has outliers and is right skewed. The IQR ranges between 992 to 4005 median is 1707. The Outliers ranges between 8524.5 to 35000 approx.. IQR= 4005-992= 3013, OUTLIER = (3013*1.5) +4005= 8524.5.

There are Full time Graduates fin all the universities ranging from 3000 to 5000.

- **PART TIME GRADUATE:**

```
Description of P.Undergrad
-----------------------------------------------------------
count      777.000000
mean       855.298584
std       1522.431887
min          1.000000
25%         95.000000
50%        353.000000
75%        967.000000
max      21836.000000
Name: P.Undergrad, dtype: float64 Distribution of P.Undergrad
```



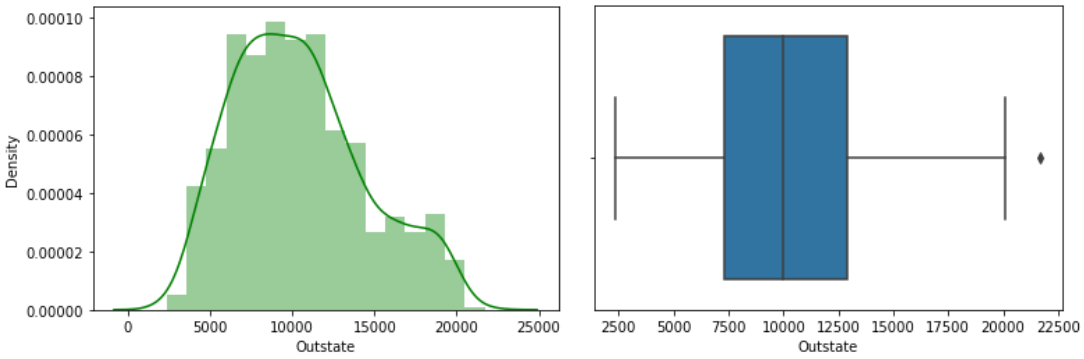
Distribution plot and Box plot for Part Time Graduate

It has outliers and is right skewed. The IQR ranges between 95 to 967, median is 353. The outliers range between 2275 and 25000 approx. IQR=967-95=872, outlier= (872*1.5) +967= 2275.

The distribution of the data is positively skewed. There are part time graduates in the range of 1000 to 3000 studying in all the universities and colleges.

- **OUTSTATE:**

```
Description of Outstate
-------------------------------------------------------
count      777.000000
mean     10440.669241
std       4023.016484
min       2340.000000
25%       7320.000000
50%       9990.000000
75%      12925.000000
max      21700.000000
Name: Outstate, dtype: float64 Distribution of Outstate
```



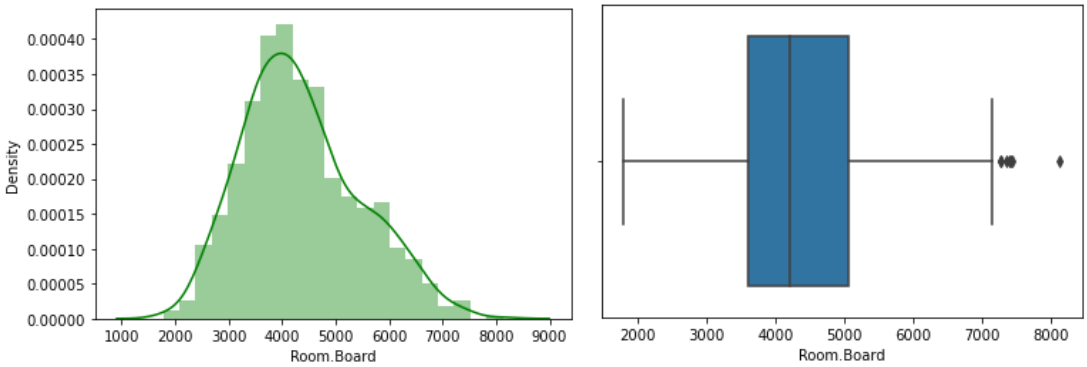
Distribution plot and Box plot of Outstate Students.

It has a single outlier and is positively skewed. The IQR ranges between 7320 to 12925 and median is 9990. The outlier is at. IQR= 12925-7320=5605, outlier= (5605*1.5) +12925= 21332.5.

The distribution is normally distributed.

- **ROOM BOARD:**

```
Description of Room.Board
----------------------------------------------------------.
count     777.000000
mean     4357.526384
std      1096.696416
min      1780.000000
25%      3597.000000
50%      4200.000000
75%      5050.000000
max      8124.000000
Name: Room.Board, dtype: float64 Distribution of Room.Board
```



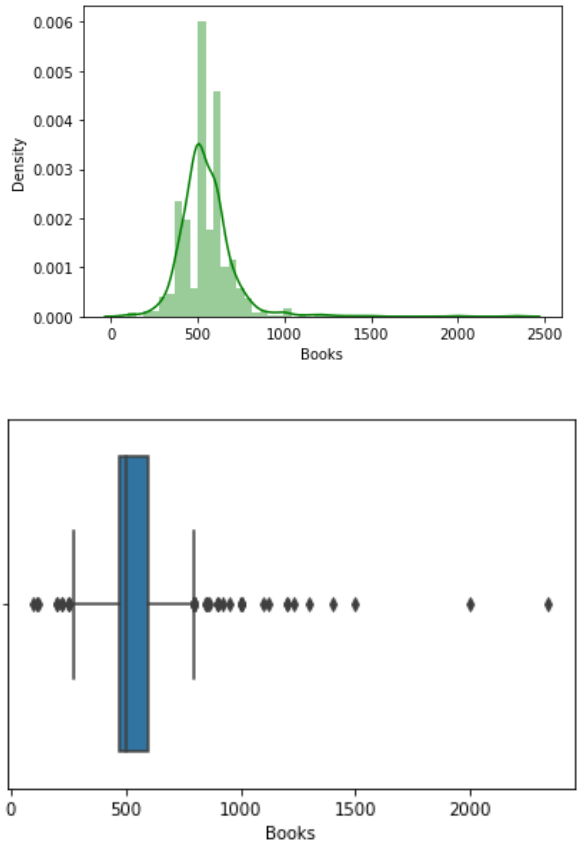
Distribution plot and Box plot of Room Board

It has outliers, it is normally distributed. The IQR ranges between 3597 to 5050, median is 4200. The outlier ranges between 7229.5 to 0-3597= 1453, outlier= (1453*1.5) +5050= 7229.5 to 8200 approx.

The Distribution is normally distributed.

- **BOOKS**:

```
Description of Books
-------------------------------------------------
count     777.000000
mean      549.380952
std       165.105360
min        96.000000
25%       470.000000
50%       500.000000
75%       600.000000
max      2340.000000
Name: Books, dtype: float64 Distribution of Books
```



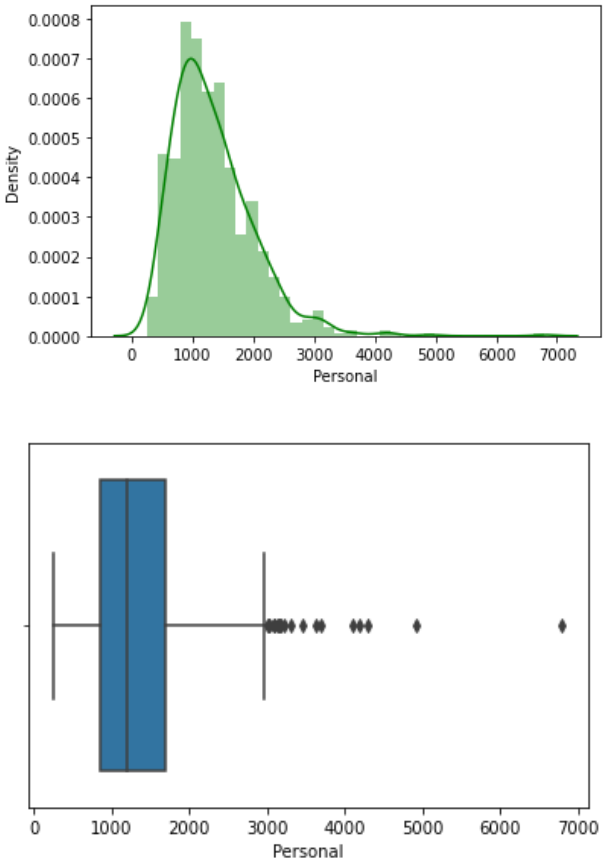
Distribution plot and Box plot of Books cost

- **INTERPRETATION OF BOOKS COST:**

It has outliers, it is right skewed. The IQR ranges between 470 to 600, median being 500. The outliers range below 275 and above 795. (IQR= 600-470=130, Outlier= (130*1.5) +600= 795 and (130*1.5)-470(Q1) =275). The Distribution is Bimodal and the cost of books per student ranges between 500 to 100.

- **PERSONAL:**

```
Description of Personal
-------------------------------------------------------
count     777.000000
mean     1340.642214
std       677.071454
min       250.000000
25%       850.000000
50%      1200.000000
75%      1700.000000
max      6800.000000
Name: Personal, dtype: float64 Distribution of Personal
```



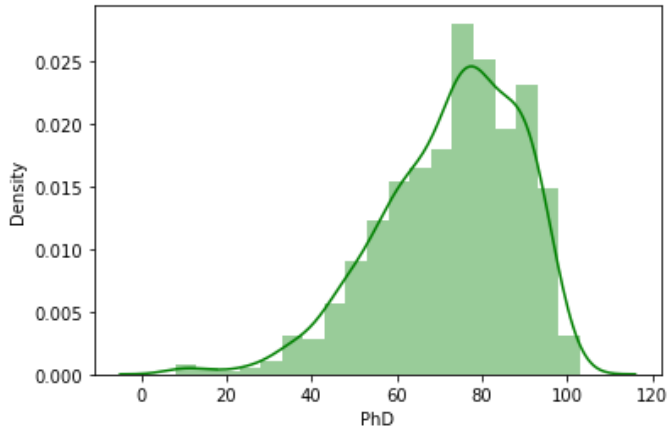
Distribution plot and Box plot of Personal cost

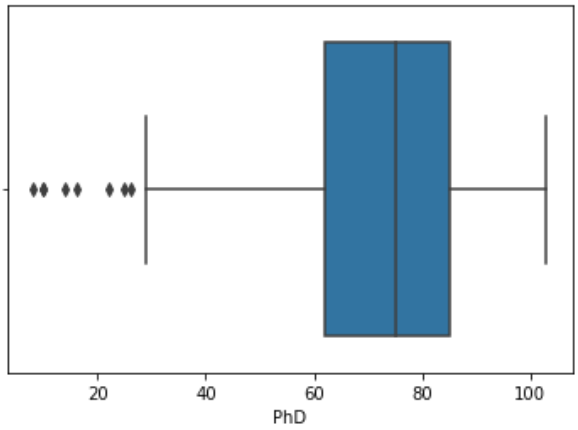
INTERPRETATION FOR PERSONAL COST:

It has Outliers and it is slightly Right skewed but almost normally distributed. The IQR ranges between 850 to 1700, median is 1200. The outlier ranges between 2975 to 6900 approx. (IQR= 1700-850=850, outlier= (850*1.5) +1700=2975). Few students' personal cost is higher than the others and distribution is positively skewed.

- **PhD:**

```
Description of PhD
-----------------------------------------------
count    777.000000
mean      72.660232
std       16.328155
min        8.000000
25%       62.000000
50%       75.000000
75%       85.000000
max      103.000000
Name: PhD, dtype: float64 Distribution of PhD
```





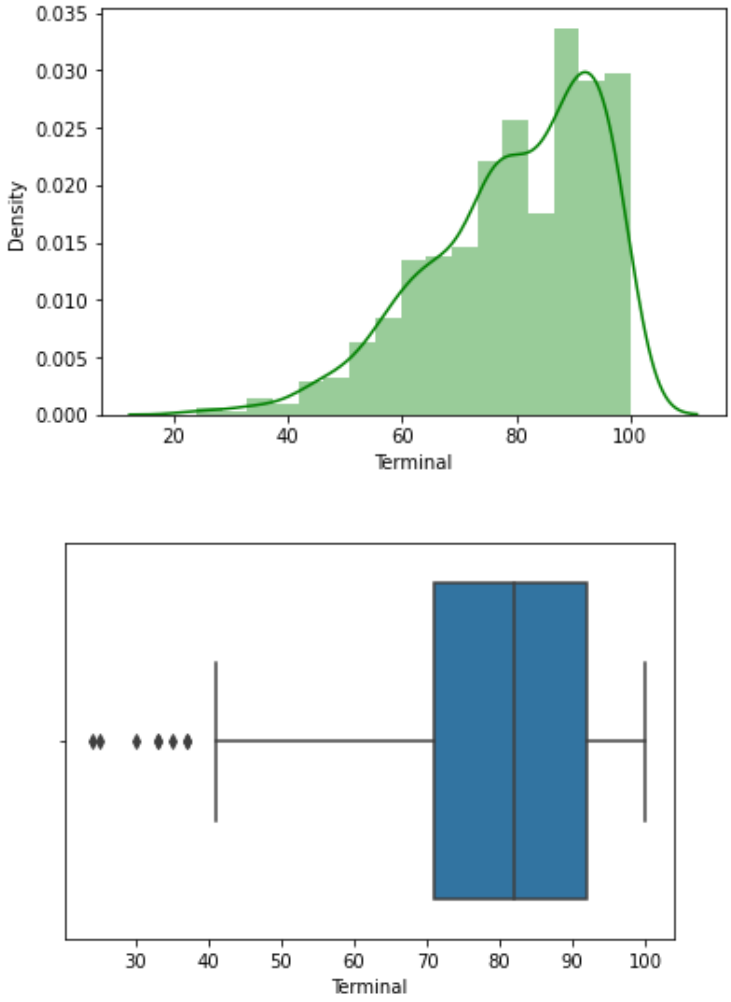
Distribution Plot and Box Plot of Faculties who have a PhD.

- **INTERPRETATION FOR PhD:**

It has Outliers and it is negatively skewed. The IQR ranges between 62 to85 and median is 75. The Outlier Ranges Between 37.5 to 10 approx. IQR= 85-60= 15, outlier= (15*1.5)-60=37.5. The Distribution is negatively skewed.

- **TERMINAL:**

```
Description of Terminal
-------------------------------------------------------
count    777.000000
mean      79.702703
std       14.722359
min       24.000000
25%       71.000000
50%       82.000000
75%       92.000000
max      100.000000
Name: Terminal, dtype: float64 Distribution of Terminal
```



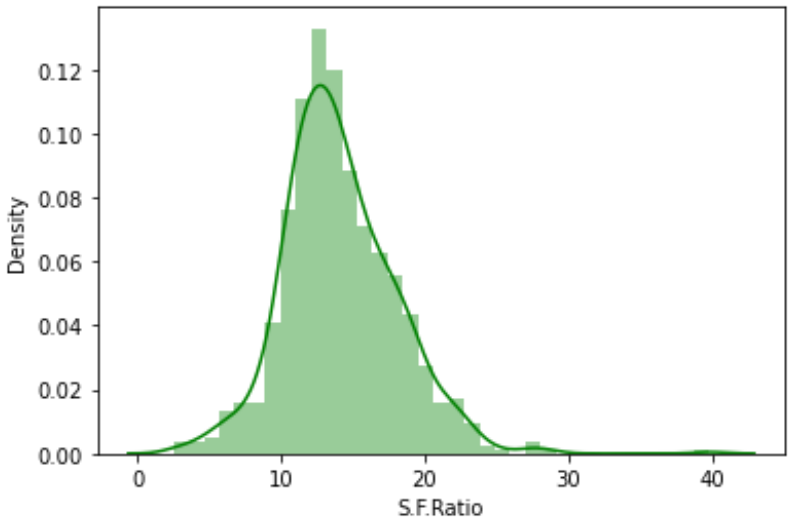
Distribution plot and Box plot of Terminal

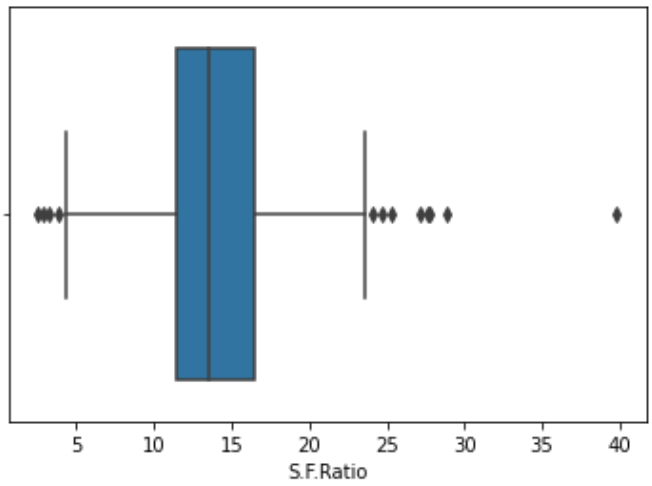
- **INTERPRETATION OF TERMINAL:**

It has Outliers and it is negatively skewed. The IQR ranges between 71 to 92, median is 82. Outlier ranges between 39.5 to 20 approx.= 92-71=21, outlier= (21*1.5)-71=39.5. The distribution is negatively skewed.

- **S.F. RATIO:**

```
Description of S.F.Ratio
------------------------------------------------------------
count    777.000000
mean      14.089704
std        3.958349
min        2.500000
25%       11.500000
50%       13.600000
75%       16.500000
max       39.800000
Name: S.F.Ratio, dtype: float64 Distribution of S.F.Ratio
```





Distribution plot and Box plot of S.F. Ratio

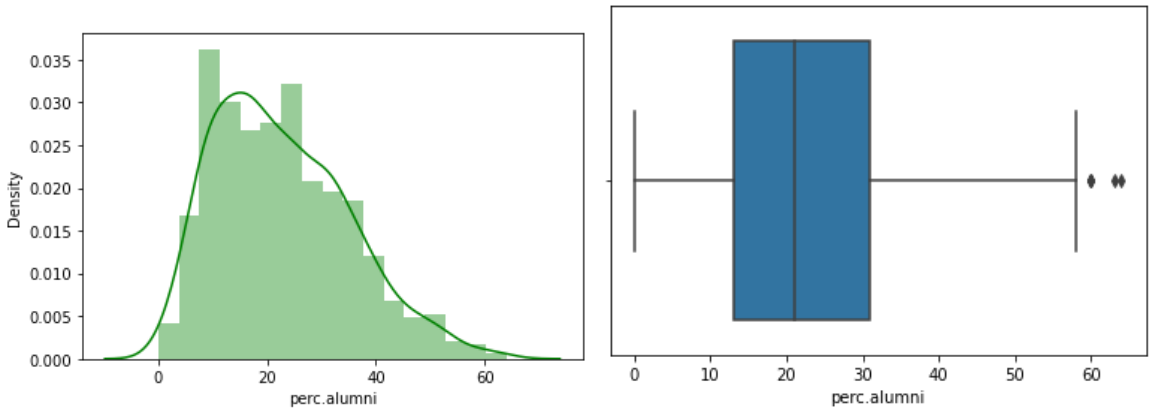
- **INTERPRETATION OF S.F. Ratio:**

It has outliers and is normally distributed. The IQR ranges between 11.5 to 16.5, median is 13.6. The Outlier ranges between 3.5 to 2.5 below the minimum whiskers IQR= 16.5-11.5=5,

outlier= (5*1.5)-11=3.5 and between 23.5 to 40 above the maximum whisker= 16.5-11.5=5, outlier= (5*1.5) +16.5=23.5. The distribution is normally distributed and the S.F. Ratio is the same in all colleges/Universities.

PERC ALUMNI:

```
Description of perc.alumni
-----------------------------------------------------------
count    777.000000
mean      22.743887
std       12.391801
min        0.000000
25%       13.000000
50%       21.000000
75%       31.000000
max       64.000000
Name: perc.alumni, dtype: float64 Distribution of perc.alumni
```



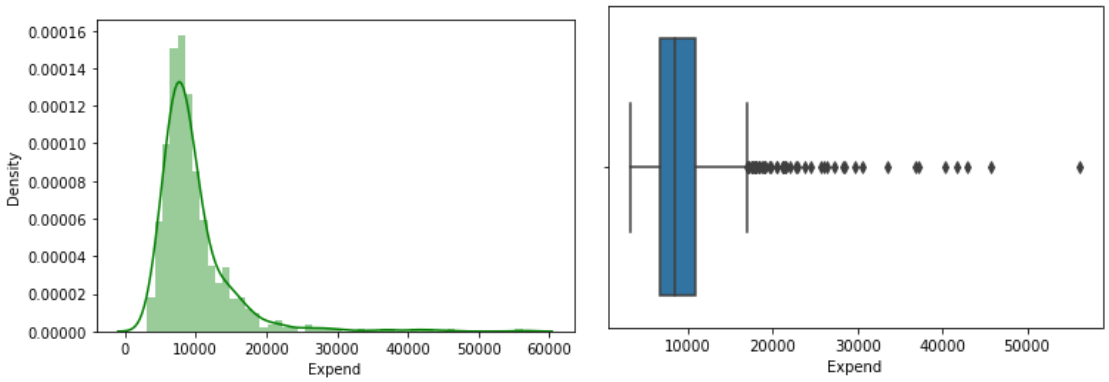
Distribution plot and Box plot of Perc Alumni

- **INTERPRETATION OF Perc Alumni:**

It has outliers and is slightly right skewed. The IQR ranges between 13 to 31, median is 21. The outliers range between 58 to 65.IQR= 31-13=18, outlier= (18*1.5) +31=58.The distribution is normally distributed.

- **EXPEND:**

```
Description of Expend
-----------------------------------------------------
count      777.000000
mean      9660.171171
std       5221.768440
min       3186.000000
25%       6751.000000
50%       8377.000000
75%      10830.000000
max      56233.000000
Name: Expend, dtype: float64 Distribution of Expend
```



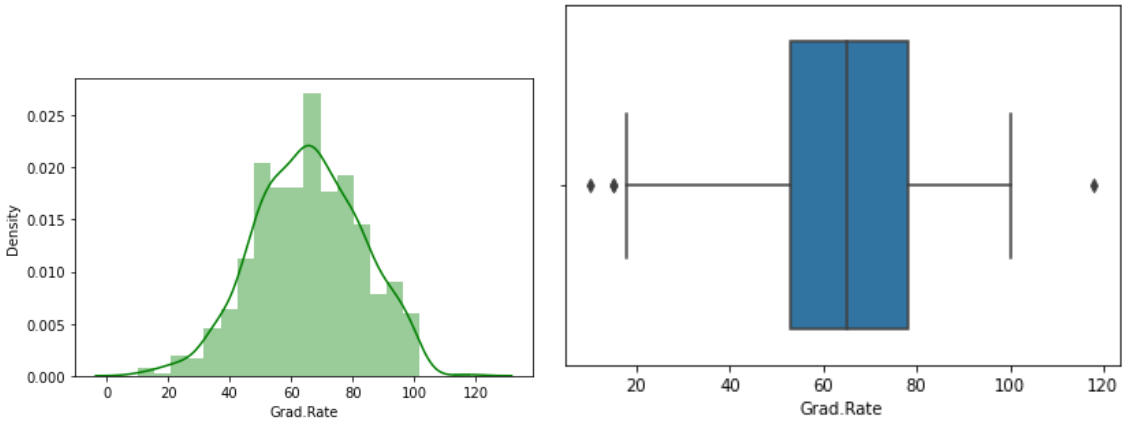
Distribution plot and Box plot of Expend

- **INTERPRETATION OF EXPEND:**

It has Outliers, it is normally distributed. The IQR ranges between 6751 to 10830, median is 8377. The outlier ranges between 16948.5 to 58000 approx. (IQR= 10830-6751=4079, outlier= (4079*1.5) +10830=16948.50). The Distribution is positively skewed.

- **GRAD RATE:**

```
Description of Grad.Rate
----------------------------------------------------------
count    777.00000
mean      65.46332
std       17.17771
min       10.00000
25%       53.00000
50%       65.00000
75%       78.00000
max      118.00000
Name: Grad.Rate, dtype: float64 Distribution of Grad.Rate
----------------------------------------------------------
```



Distribution plot and Box plot of Grad Rate

- **INTERPRETATION OF GRAD RATE:**

It has Outliers and it is slightly negatively skewed, but normally distributed. The IQR ranges between 53 to 78 and median is 65. The outlier ranges between 15.5 to 12 approx. below the minimum whisker point (IQR= 78-53=25, outlier= (25*1.5)-53=15.5) and at 115. 5 above the maximum whisker (IQR= 78-53=25, outlier= (25*1.5) +78=115.5). The distribution is normally distributed and there is a graduation rate of 60% in all the universities or colleges.

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

Yes, Scaling is necessary in this case, because as per the data dictionary and Info command There are 18 numerical columns with different scales.

Scaling in general means the representation of the data set into one unit.

Like,

1. The application, application accepted, Full time graduates, part time graduates, outstate are number of students,

2. The Top 10 percent and top 25 percent are values in percentage,

3. Room Board, books and personal are values in Units/Money,

4. The PhD, S.F. Ratio, percentage of alumni, Grad Rate are values in percentage of students, Faculties and Graduates.

- **SCALING USING Z SCORE TEST:**

After Dropping column "Names" from the Data Frame

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
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	-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
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
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
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

S.F.Ratio	perc.alumni	Expend	Grad.Rate
1.013776	-0.867574	-0.501910	-0.318252
-0.477704	-0.544572	0.166110	-0.551262
-0.300749	0.585935	-0.177290	-0.667767
-1.615274	1.151188	1.792851	-0.376504
-0.553542	-1.675079	0.241803	-2.939613

Table 8: Scaled Data frame

Z (Z score) = x-mu/standard deviation of the sample,

where,

x = the observed value,

Mu= The mean of the sample.

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

Covariance matrix

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad
Apps	1.001289	0.944666	0.847913	0.339270	0.352093	0.815540
Accept	0.944666	1.001289	0.912811	0.192695	0.247795	0.875350
Enroll	0.847913	0.912811	1.001289	0.181527	0.227037	0.965883
Top10perc	0.339270	0.192695	0.181527	1.001289	0.893144	0.141471
Top25perc	0.352093	0.247795	0.227037	0.893144	1.001289	0.199702
F.Undergrad	0.815540	0.875350	0.965883	0.141471	0.199702	1.001289
P.Undergrad	0.398777	0.441839	0.513730	-0.105492	-0.053646	0.571247
Outstate	0.050224	-0.025788	-0.155678	0.563055	0.490024	-0.216020
Room.Board	0.165152	0.091016	-0.040284	0.371959	0.331917	-0.068979
Books	0.132729	0.113672	0.112856	0.119012	0.115676	0.115699
Personal	0.178961	0.201248	0.281291	-0.093437	-0.080914	0.317608
PhD	0.391201	0.356216	0.331896	0.532513	0.546566	0.318747
Terminal	0.369968	0.338018	0.308671	0.491768	0.525425	0.300406
S.F.Ratio	0.095756	0.176456	0.237577	-0.385370	-0.295009	0.280064
perc.alumni	-0.090342	-0.160196	-0.181027	0.456072	0.418403	-0.229758
Expend	0.259927	0.124878	0.064252	0.661765	0.528127	0.018676
Grad.Rate	0.146944	0.067399	-0.022370	0.495627	0.477896	-0.078875

	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD
Apps	0.398777	0.050224	0.165152	0.132729	0.178961	0.391201
Accept	0.441839	-0.025788	0.091016	0.113672	0.201248	0.356216
Enroll	0.513730	-0.155678	-0.040284	0.112856	0.281291	0.331896
Top10perc	-0.105492	0.563055	0.371959	0.119012	-0.093437	0.532513
Top25perc	-0.053646	0.490024	0.331917	0.115676	-0.080914	0.546566
F.Undergrad	0.571247	-0.216020	-0.068979	0.115699	0.317608	0.318747
P.Undergrad	1.001289	-0.253839	-0.061405	0.081304	0.320294	0.149306
Outstate	-0.253839	1.001289	0.655100	0.038905	-0.299472	0.383476
Room.Board	-0.061405	0.655100	1.001289	0.128128	-0.199685	0.329627
Books	0.081304	0.038905	0.128128	1.001289	0.179526	0.026940
Personal	0.320294	-0.299472	-0.199685	0.179526	1.001289	-0.010950
PhD	0.149306	0.383476	0.329627	0.026940	-0.010950	1.001289
Terminal	0.142086	0.408509	0.375022	0.100084	-0.030653	0.850682
S.F.Ratio	0.232830	-0.555536	-0.363095	-0.031970	0.136521	-0.130698
perc.alumni	-0.281154	0.566992	0.272714	-0.040260	-0.286337	0.249330
Expend	-0.083676	0.673646	0.502386	0.112554	-0.098018	0.433319
Grad.Rate	-0.257332	0.572026	0.425489	0.001062	-0.269691	0.305431

	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	0.369968	0.095756	-0.090342	0.259927	0.146944
Accept	0.338018	0.176456	-0.160196	0.124878	0.067399
Enroll	0.308671	0.237577	-0.181027	0.064252	-0.022370
Top10perc	0.491768	-0.385370	0.456072	0.661765	0.495627
Top25perc	0.525425	-0.295009	0.418403	0.528127	0.477896
F.Undergrad	0.300406	0.280064	-0.229758	0.018676	-0.078875
P.Undergrad	0.142086	0.232830	-0.281154	-0.083676	-0.257332
Outstate	0.408509	-0.555536	0.566992	0.673646	0.572026
Room.Board	0.375022	-0.363095	0.272714	0.502386	0.425489
Books	0.100084	-0.031970	-0.040260	0.112554	0.001062
Personal	-0.030653	0.136521	-0.286337	-0.098018	-0.269691
PhD	0.850682	-0.130698	0.249330	0.433319	0.305431
Terminal	1.001289	-0.160310	0.267475	0.439365	0.289900
S.F.Ratio	-0.160310	1.001289	-0.403448	-0.584584	-0.307106
perc.alumni	0.267475	-0.403448	1.001289	0.418250	0.491530
Expend	0.439365	-0.584584	0.418250	1.001289	0.390846
Grad.Rate	0.289900	-0.307106	0.491530	0.390846	1.001289

Table 9: Covariance Matrix

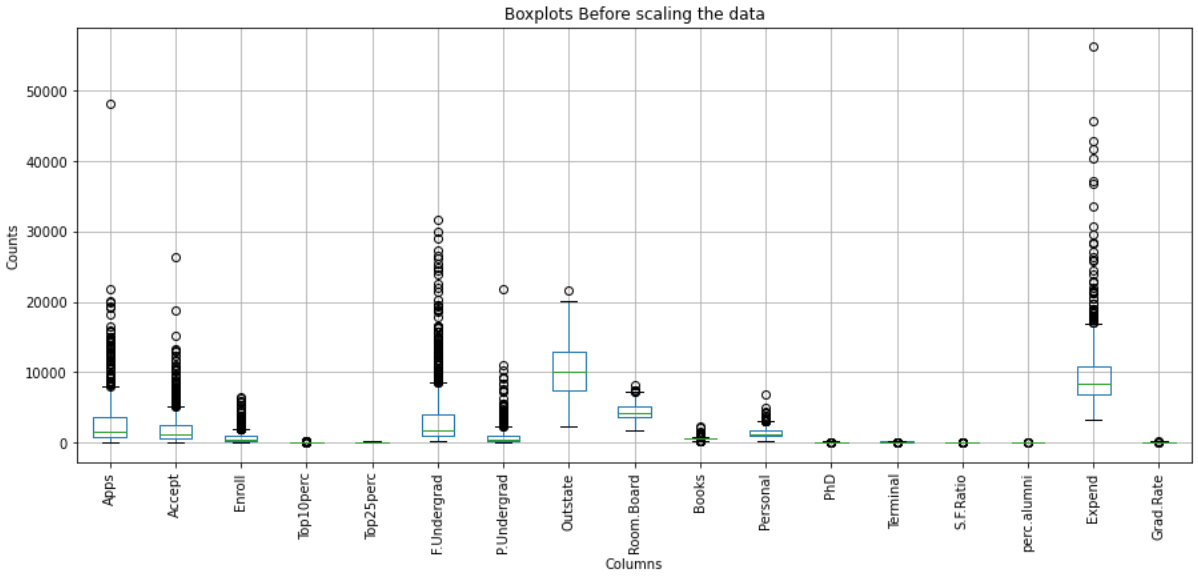
- **INFERENCE FOR Q.2.3**.

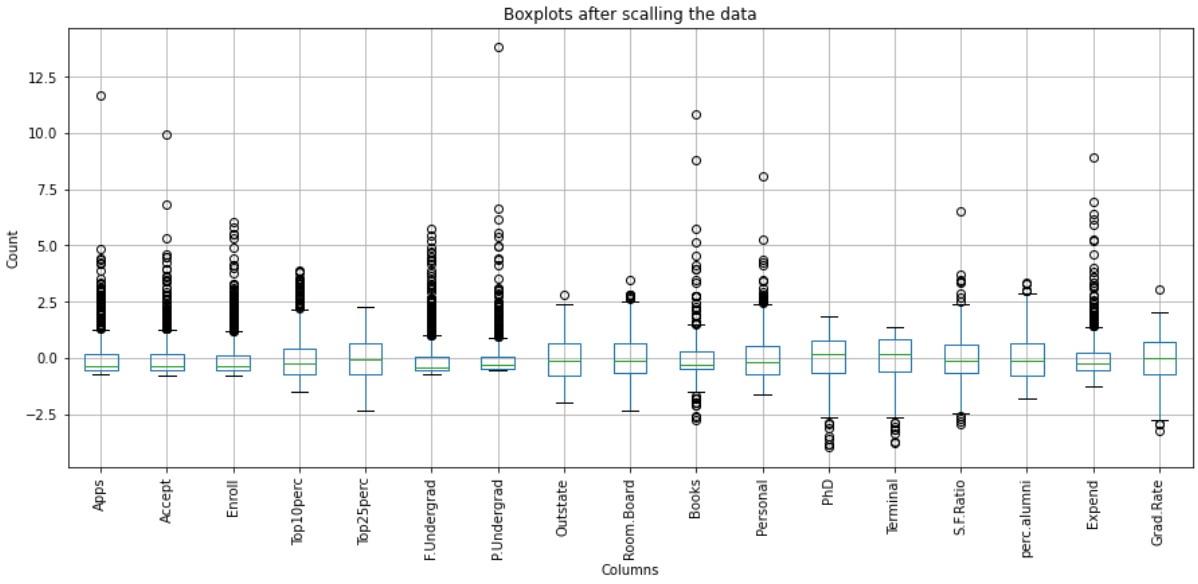
The comparison between the covariance and correlation matrix is the relationship measure and dependency between two variables.

Covariance indicates the proportionality (Direct or indirect) of the relationship between variables, if it is positive or negative.

Here in the above table, we can see that app, accept, enrol, F. Grad, top 10 percentage and top 25 percentage are highly positively correlated.

Q.2.4 Check the dataset for outliers before and after scaling. What insight do you derive here?





INTERPRETATION FOR Q.2.4:

There are still outliers in the scaled boxplot, it is because Z score test or scaling does not remove the Outliers from the data. In case outliers have to be removed, we can remove them from the data itself or impute median or mean values (IQR value).

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

- Applying SK learn using decomposition method:

```
array([[-1.59285540e+00,  7.67333510e-01, -1.01073537e-01, ...,
         1.75239502e-03, -9.31400698e-02,  9.35522023e-02],
       [-2.19240180e+00, -5.78829984e-01,  2.27879812e+00, ...,
         1.03709803e-01, -5.02556890e-02, -1.74057054e-01],
       [-1.43096371e+00, -1.09281889e+00, -4.38092811e-01, ...,
        -2.25582869e-02, -4.05268301e-03,  3.75875882e-03],
       ...,
       [-7.32560596e-01, -7.72352397e-02, -4.05641899e-04, ...,
         6.79013123e-02, -2.32023970e-01, -9.99380421e-02],
       [ 7.91932735e+00, -2.06832886e+00,  2.07356368e+00, ...,
         3.53597440e-01,  3.04416200e-01,  3.35104811e-01],
       [-4.69508066e-01,  3.66660943e-01, -1.32891515e+00, ...,
        -1.14873492e-01, -1.17076127e-01, -2.57218339e-03]])
```

- Extracting the Eigen Vectors:

```
array([[ 2.48765602e-01,  2.07601502e-01,  1.76303592e-01,
         3.54273947e-01,  3.44001279e-01,  1.54640962e-01,
         2.64425045e-02,  2.94736419e-01,  2.49030449e-01,
         6.47575181e-02, -4.25285386e-02,  3.18312875e-01,
         3.17056016e-01, -1.76957895e-01,  2.05082369e-01,
         3.18908750e-01,  2.52315654e-01],
       [ 3.31598227e-01,  3.72116750e-01,  4.03724252e-01,
        -8.24118211e-02, -4.47786551e-02,  4.17673774e-01,
         3.15087830e-01, -2.49643522e-01, -1.37808883e-01,
         5.63418434e-02,  2.19929218e-01,  5.83113174e-02,
         4.64294477e-02,  2.46665277e-01, -2.46595274e-01,
        -1.31689865e-01, -1.69240532e-01],
       [-6.30921033e-02, -1.01249056e-01, -8.29855709e-02,
         3.50555339e-02, -2.41479376e-02, -6.13929764e-02,
         1.39681716e-01,  4.65988731e-02,  1.48967389e-01,
         6.77411649e-01,  4.99721120e-01, -1.27028371e-01,
        -6.60375454e-02, -2.89848401e-01, -1.46989274e-01,
         2.26743985e-01, -2.08064649e-01],
       [ 2.81310530e-01,  2.67817346e-01,  1.61826771e-01,
        -5.15472524e-02, -1.09766541e-01,  1.00412335e-01,
        -1.58558487e-01,  1.31291364e-01,  1.84995991e-01,
         8.70892205e-02, -2.30710568e-01, -5.34724832e-01,
        -5.19443019e-01, -1.61189487e-01,  1.73142230e-02,
         7.92734946e-02,  2.69129066e-01],
```

```
[ 5.74140964e-03,  5.57860920e-02, -5.56936353e-02,
 -3.95434345e-01, -4.26533594e-01, -4.34543659e-02,
  3.02385408e-01,  2.22532003e-01,  5.60919470e-01,
 -1.27288825e-01, -2.22311021e-01,  1.40166326e-01,
  2.04719730e-01, -7.93882496e-02, -2.16297411e-01,
  7.59581203e-02, -1.09267913e-01],
[-1.62374420e-02,  7.53468452e-03, -4.25579803e-02,
 -5.26927980e-02,  3.30915896e-02, -4.34542349e-02,
 -1.91198583e-01, -3.00003910e-02,  1.62755446e-01,
  6.41054950e-01, -3.31398003e-01,  9.12555212e-02,
  1.54927646e-01,  4.87045875e-01, -4.73400144e-02,
 -2.98118619e-01,  2.16163313e-01],
[-4.24863486e-02, -1.29497196e-02, -2.76928937e-02,
 -1.61332069e-01, -1.18485556e-01, -2.50763629e-02,
  6.10423460e-02,  1.08528966e-01,  2.09744235e-01,
 -1.49692034e-01,  6.33790064e-01, -1.09641298e-03,
 -2.84770105e-02,  2.19259358e-01,  2.43321156e-01,
 -2.26584481e-01,  5.59943937e-01],
[-1.03090398e-01, -5.62709623e-02,  5.86623552e-02,
 -1.22678028e-01, -1.02491967e-01,  7.88896442e-02,
  5.70783816e-01,  9.84599754e-03, -2.21453442e-01,
  2.13293009e-01, -2.32660840e-01, -7.70400002e-02,
 -1.21613297e-02, -8.36048735e-02,  6.78523654e-01,
 -5.41593771e-02, -5.33553891e-03],
```

```
[-9.02270802e-02, -1.77864814e-01, -1.28560713e-01,
  3.41099863e-01,  4.03711989e-01, -5.94419181e-02,
  5.60672902e-01, -4.57332880e-03,  2.75022548e-01,
 -1.33663353e-01, -9.44688900e-02, -1.85181525e-01,
 -2.54938198e-01,  2.74544380e-01, -2.55334907e-01,
 -4.91388809e-02,  4.19043052e-02],
[ 5.25098025e-02,  4.11400844e-02,  3.44879147e-02,
  6.40257785e-02,  1.45492289e-02,  2.08471834e-02,
 -2.23105808e-01,  1.86675363e-01,  2.98324237e-01,
 -8.20292186e-02,  1.36027616e-01, -1.23452200e-01,
 -8.85784627e-02,  4.72045249e-01,  4.22999706e-01,
  1.32286331e-01, -5.90271067e-01],
[ 4.30462074e-02, -5.84055850e-02, -6.93988831e-02,
 -8.10481404e-03, -2.73128469e-01, -8.11578181e-02,
  1.00693324e-01,  1.43220673e-01, -3.59321731e-01,
  3.19400370e-02, -1.85784733e-02,  4.03723253e-02,
 -5.89734026e-02,  4.45000727e-01, -1.30727978e-01,
  6.92088870e-01,  2.19839000e-01],
[ 2.40709086e-02, -1.45102446e-01,  1.11431545e-02,
  3.85543001e-02, -8.93515563e-02,  5.61767721e-02,
 -6.35360730e-02, -8.23443779e-01,  3.54559731e-01,
 -2.81593679e-02, -3.92640266e-02,  2.32224316e-02,
  1.64850420e-02, -1.10262122e-02,  1.82660654e-01,
  3.25982295e-01,  1.22106697e-01],
```

```
[ 5.95830975e-01,  2.92642398e-01, -4.44638207e-01,
  1.02303616e-03,  2.18838802e-02, -5.23622267e-01,
  1.25997650e-01, -1.41856014e-01, -6.97485854e-02,
  1.14379958e-02,  3.94547417e-02,  1.27696382e-01,
 -5.83134662e-02, -1.77152700e-02,  1.04088088e-01,
 -9.37464497e-02, -6.91969778e-02],
[ 8.06328039e-02,  3.34674281e-02, -8.56967180e-02,
 -1.07828189e-01,  1.51742110e-01, -5.63728817e-02,
  1.92857500e-02, -3.40115407e-02, -5.84289756e-02,
 -6.68494643e-02,  2.75286207e-02, -6.91126145e-01,
  6.71008607e-01,  4.13740967e-02, -2.71542091e-02,
  7.31225166e-02,  3.64767385e-02],
[ 1.33405806e-01, -1.45497511e-01,  2.95896092e-02,
  6.97722522e-01, -6.17274818e-01,  9.91640992e-03,
  2.09515982e-02,  3.83544794e-02,  3.40197083e-03,
 -9.43887925e-03, -3.09001353e-03, -1.12055599e-01,
  1.58909651e-01, -2.08991284e-02, -8.41789410e-03,
 -2.27742017e-01, -3.39433604e-03],
[ 4.59139498e-01, -5.18568789e-01, -4.04318439e-01,
 -1.48738723e-01,  5.18683400e-02,  5.60363054e-01,
 -5.27313042e-02,  1.01594830e-01, -2.59293381e-02,
  2.88282896e-03, -1.28904022e-02,  2.98075465e-02,
 -2.70759809e-02, -2.12476294e-02,  3.33406243e-03,
 -4.38803230e-02, -5.00844705e-03],
[ 3.58970400e-01, -5.43427250e-01,  6.09651110e-01,
 -1.44986329e-01,  8.03478445e-02, -4.14705279e-01,
  9.01788964e-03,  5.08995918e-02,  1.14639620e-03,
  7.72631963e-04, -1.11433396e-03,  1.38133366e-02,
  6.20932749e-03, -2.22215182e-03, -1.91869743e-02,
 -3.53098218e-02, -1.30710024e-02]])
```

- Extracting the Eigen Values:

```
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])
```

Q.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]

```
The Linear equation of the first components are:
0.25*Apps+0.21*Accept+0.18*Enroll+0.35*Top10perc+0.34*Top25perc+0.15*F.Undergrad+0.03*P.Undergrad+0.29*Outstate+0.25*Room.Board
+0.06*Books+-0.04*Personal+0.32*PhD+0.32*Terminal+-0.18*S.F.Ratio+0.21*perc.alumni+0.32*Expend+0.25*Grad.Rate+
```

Q.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?

- Explained Variance Ratio**:**

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

- Cumulative Explained Variance:

```
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.        ])
```

- **INFERENCE FOR Q.2.8**

To decide the optimum number of principal components

Adding the Eigen Values, we get total as 100

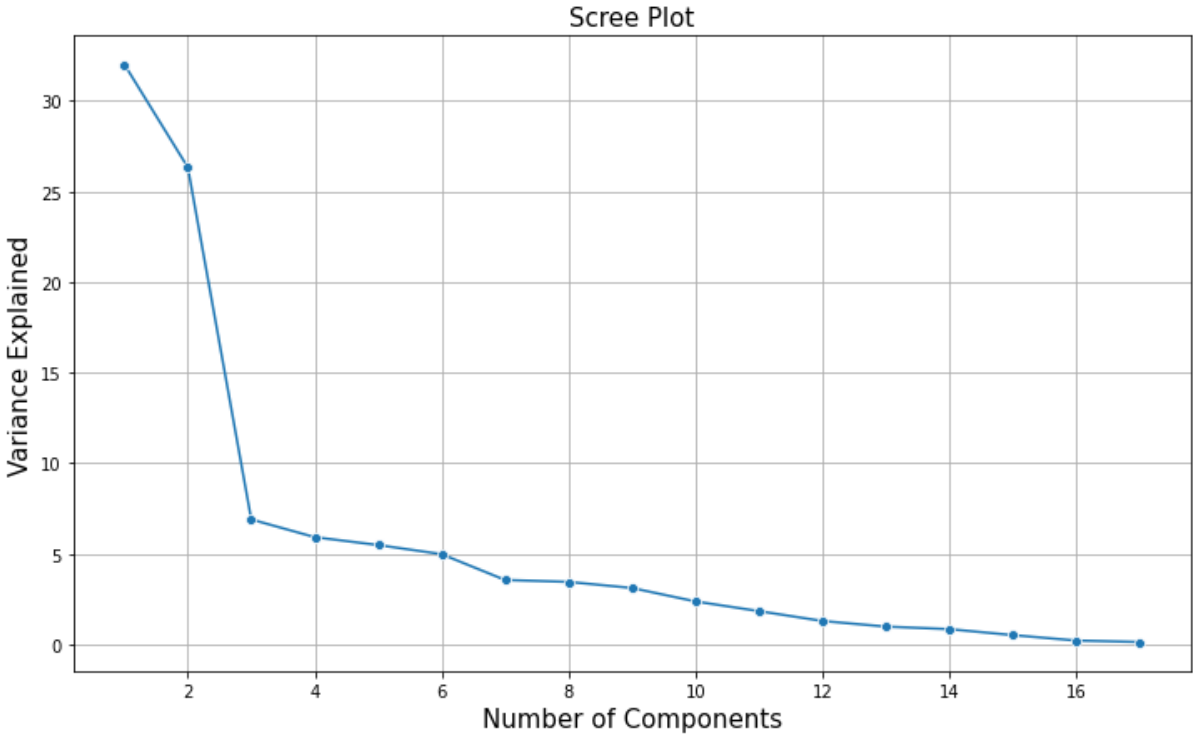
1. The incremental value between the components should not be less than 5%

2. Cumulative variance should be up to 90%-85%

Based on this we can decide that the optimal number of principal components is 6, as after 81.65 the incremental value between the components is less than 5%. Therefore there 5 principal components for this case.

Thus, it is understood that how much each variable contributes to the Principal components, with eigen vectors we can understand which variable has more weightage and influence the dataset in principal component analysis.

The PCA reduces the multi collinearity, thus with these reduced multi collinearity we run models efficiently.



From the above scree plot it is inferred that the multi-collinearity has been reduced using PCA.

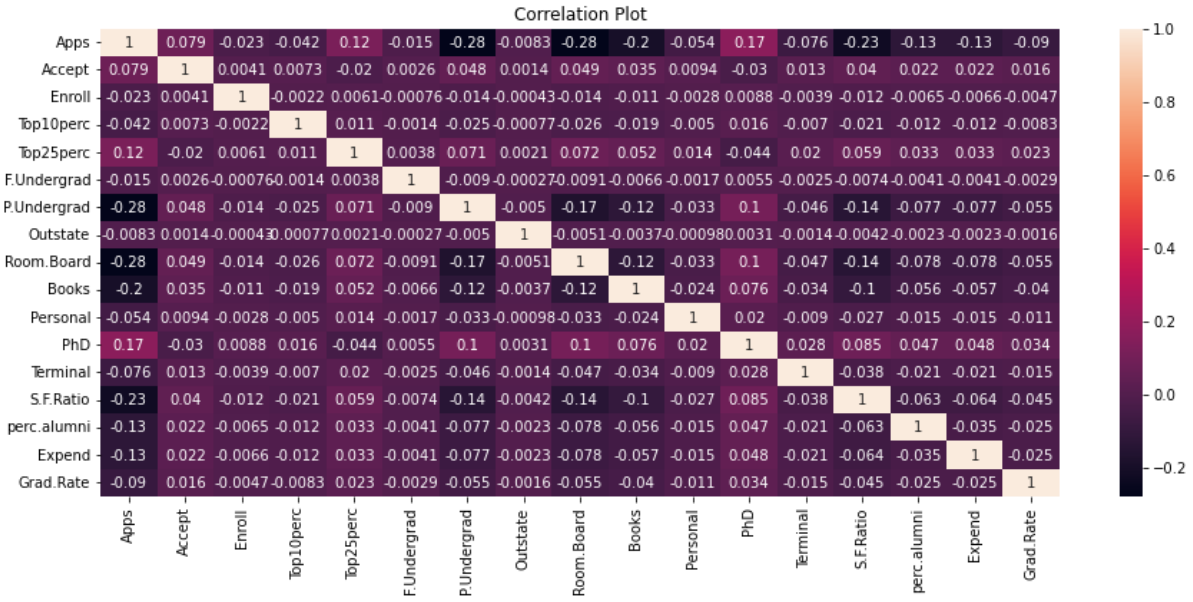
- PCA being exported into a new Data Frame:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
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
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
2	-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
3	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
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
5	-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
6	-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
7	-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
8	-0.090227	-0.177865	-0.128561	0.341100	0.403712	-0.059442	0.560673	-0.004573	0.275023	-0.133663	-0.094469	-0.185182	-0.254938
9	0.052510	0.041140	0.034488	0.064026	0.014549	0.020847	-0.223106	0.186675	0.298324	-0.082029	0.136028	-0.123452	-0.088578
10	0.043046	-0.058406	-0.069399	-0.008105	-0.273128	-0.081158	0.100693	0.143221	-0.359322	0.031940	-0.018578	0.040372	-0.058973
11	0.024071	-0.145102	0.011143	0.038554	-0.089352	0.056177	-0.063536	-0.823444	0.354560	-0.028159	-0.039264	0.023222	0.016485
12	0.595831	0.292642	-0.444638	0.001023	0.021884	-0.523622	0.125998	-0.141856	-0.069749	0.011438	0.039455	0.127696	-0.058313
13	0.080633	0.033467	-0.085697	-0.107828	0.151742	-0.056373	0.019286	-0.034012	-0.058429	-0.066849	0.027529	-0.691126	0.671009
14	0.133406	-0.145498	0.029590	0.697723	-0.617275	0.009916	0.020952	0.038354	0.003402	-0.009439	-0.003090	-0.112056	0.158910
15	0.459139	-0.518569	-0.404318	-0.148739	0.051868	0.560363	-0.052731	0.101595	-0.025929	0.002883	-0.012890	0.029808	-0.027076
16	0.358970	-0.543427	0.609651	-0.144986	0.080348	-0.414705	0.009018	0.050900	0.001146	0.000773	-0.001114	0.013813	0.006209

S.F.Ratio	perc.alumni	Expend	Grad.Rate
-0.176958	0.205082	0.318909	0.252316
0.246665	-0.246595	-0.131690	-0.169241
-0.289848	-0.146989	0.226744	-0.208065
-0.161189	0.017314	0.079273	0.269129
-0.079388	-0.216297	0.075958	-0.109268
0.487046	-0.047340	-0.298119	0.216163
0.219259	0.243321	-0.226584	0.559944
-0.083605	0.678524	-0.054159	-0.005336
0.274544	-0.255335	-0.049139	0.041904
0.472045	0.423000	0.132286	-0.590271
0.445001	-0.130728	0.692089	0.219839
-0.011026	0.182661	0.325982	0.122107
-0.017715	0.104088	-0.093746	-0.069197
0.041374	-0.027154	0.073123	0.036477
-0.020899	-0.008418	-0.227742	-0.003394
-0.021248	0.003334	-0.043880	-0.005008
-0.002222	-0.019187	-0.035310	-0.013071

Table 10: New Data Frame

- Heatmap after PCA:



Q.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 case study talks about the education, this dataset contains the names of colleges and universities with the number of applications received, accepted, enrolled and details about the faculties and other student cost.

To understand, we perform Univariate and Multi variate analysis, which provides us with detail understanding about the variables i.e., we get to know the following:

1. Distribution of the data in the dataset

2. The skewness of the data in the dataset

3. Different trends in the dataset.

From Multi variate analysis we get to know the Correlation of variables, whether they are highly correlated or no correlation is there.

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 PCA is used to reduce the multicollinearity between the variables. Depending on the variance of the dataset we can reduce the PCA components. The PCA components for this

case are 5 where er can understand that maximum variance in the dataset. With PCA we can perform further analysis and models with efficiency.

THE END