ADVANCE STATISTICS
FINAL PROJECT

Balaji M P
PGP DSBA Online -March' 22
Date: 26.06.2022

greatlearning
Power Ahead

TABLE OF CONTENTS

Salary Analysis.....3
Education PCA Analysis.....14

LIST OF FIGURES

Fig.1	Education Qualifications	03
Fig.2	Details of the dataset columns	05
Fig.3	Box-plot of Salary w.r.t Education	09
Fig.4	Interaction plot of Education & Occupation w.r.t Salary	10
Fig.5	Education PCA Analysis	14
Fig.6	Details of the dataset columns	19
Fig.7	Descriptive Summary, Histogram & Boxplot of Apps column	18
Fig.8	Descriptive Summary, Histogram & Boxplot of Accept column	19
Fig.9	Descriptive Summary, Histogram & Boxplot of Enroll column	20
Fig.10	Descriptive Summary, Histogram & Boxplot of Top10perc column	21
Fig.11	Descriptive Summary, Histogram & Boxplot of Top25perc column	22
Fig.12	Descriptive Summary, Histogram & Boxplot of F.Undergrad column	23
Fig.13	Descriptive Summary, Histogram & Boxplot of P.Undergrad column	24
Fig.14	Descriptive Summary, Histogram & Boxplot of Outstate column	25
Fig.15	Descriptive Summary, Histogram & Boxplot of Room.Board column	26
Fig.16	Descriptive Summary, Histogram & Boxplot of Books column	27
Fig.17	Descriptive Summary, Histogram & Boxplot of Personal column	28
Fig.18	Descriptive Summary, Histogram & Boxplot of PhD column	29
Fig.19	Descriptive Summary, Histogram & Boxplot of Terminal column	30
Fig.20	Descriptive Summary, Histogram & Boxplot of S.F.Ratio column	31
Fig.21	Descriptive Summary, Histogram & Boxplot of perc.alumni column	32
Fig.22	Descriptive Summary, Histogram & Boxplot of Expend column	33
Fig.23	Descriptive Summary, Histogram & Boxplot of Grad.Rate column	34
Fig.24	Distribution of Colleges & University	35
Fig.25	Pair plot of all the features	36
Fig.26	Correlation Heatmap	37
Fig.27	Boxplot of the features without scaling	39
Fig.28	Boxplot of the features after scaling	40
Fig.29	Covariance Heatmap	41
Fig.30	Correlation Heatmap	42

Fig.31	Boxplot of the features without scaling	43
Fig.32	Boxplot of the features after scaling	44
Fig.33	Eigen Values	45
Fig.34.1	Eigen Vectors	45
Fig.34.2	Eigen Vectors	46
Fig.35	Scree Plot	47
Fig.36	Individual & Cumulative variance	48
Fig.37	Information of new Data frame with original features	49
Fig.38	Cumulative Variance Explained for 17 PC's	51
Fig.39	Variance & Cumulative Variance	51
Fig.40	Correlation Heatmap of PC's	53
Fig.41	Feature classifications using Heatmap	54

LIST OF TABLES

Table 1	Sample of first 10 rows of the dataset	04
Table 2	ANOVA Table (Education w.r.t Salary)	06
Table 3	ANOVA Table (Education w.r.t Salary)	07
Table 4	One way ANOVA Table (Education w.r.t Salary)	09
Table 5	Sample of first 5 rows of the dataset	16
Table 6	New data frame with original features	49
Table 7	Transposed data-frame rounded off to 2 decimal places	50

Salary Analysis

Executive Summary

The educational background and employment of each individual are recorded in a dataset that includes the wages of 40 people. There are three stages of educational qualification: 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.

Fig.1 Education Qualifications

Introduction

The purpose of this **report** is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. The data consists of **Salaries, Educational Background and Occupation of 40 individuals.**

Data Description

1. **Education: Doctorate, Bachelors, HS-grad**
2. **Occupation: Prof-specialty, Sales, Adm-clerical, Exec-managerial Salary**
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
5	Doctorate	Sales	219420
6	Doctorate	Sales	237920
7	Doctorate	Sales	160540
8	Doctorate	Sales	180934
9	Doctorate	Prof-specialty	248156

Table 1. Sample of first 10 rows of the dataset

Dataset has 3 columns with 3 different types of Education, 4 different types Occupation and Individual's salary.

Exploratory Data Analysis:

Let us check the types of variables in the data frame and 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)
memory usage: 1.1+ KB
```

Fig.2 Details of the dataset columns

There are total of 40 rows and 3 columns in the dataset. Out of 3 columns, 2 columns are string type and the remaining column is integer. Also, from the above results we can see that there are no missing values present in dataset.

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

Framing the hypothesis:

One way ANOVA for Education assuming a confidence level of 95% (α = 0.05)

Education has three levels Doctorate, Bachelors & HS-grad. Let us assume the population means to be μ1, μ2, μ3 respectively.

H_0: $\mu 1 = \mu 2 = \mu 3$, the mean salary is same across all levels of education

H_1: For at least one pair of education level, the mean salary is different

One way ANOVA for Occupation assuming a confidence level of 95% (α = 0.05)

Occupation has four levels Prof-specialty, Sales, Adm-clerical, Exec-managerial. Let us assume the population means to be μ1, μ2, μ3, μ4 respectively.

H0: $\mu 1 = \mu 2 = \mu 3 = \mu 4$, the mean salary is same across all levels of occupation.

H1: For at least one pair of occupation level, the mean salary is different.

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.

The below table shows the results of ANOVA for Education with respect to Salary.

Table Info:

df: Degrees of freedom for model terms

sum_sq: Sum of squares for model terms

mean_sq: Mean of Sum of squares for model terms

F: F statistic value for significance of adding model terms

PR(>F): P-value for significance of adding model terms

	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 2. ANOVA Table (Education w.r.t Salary)

Inference:

At the level of 5% significance, **p-value is lesser than significance level** (α=0.05). Since **p-value < α**, we reject the null hypothesis. Hence, the **mean salary is different for at least one pair of Education level**.

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.

The below table shows the results of ANOVA for Occupation with respect to Salary.

Table Info:

df: Degrees of freedom for model terms

sum_sq: Sum of squares for model terms

mean_sq: Mean of Sum of squares for model terms

F: F statistic value for significance of adding model terms

PR(>F): P-value for significance of adding model terms

	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 3. ANOVA Table (Education w.r.t Salary)

Inference:

At the level of 5% significance, **p-value is greater than significance level** (α=0.05). Since **p-value > α**, we fail to reject the null hypothesis. Hence, the **mean salary is same across all levels of Occupation**.

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.

We failed to reject the null hypothesis in 1.3. However, the null hypothesis is rejected in 1.2, **Tukey's Multi comparison test** is performed to find out which education level has significantly different mean salary.

Framing the hypothesis:

H_0: All pairs of group means are equal

H_1: At least one group mean is different from the rest.

In this case, as there are only 3 pairs to be considered, we may write the null and alternative hypothesis as:

H_0: $\mu 1 = \mu 2$ and $\mu 1 = \mu 3$ and $\mu 2 = \mu 3$

H_1: $\mu 1 \neq \mu 2$ or $\mu 1 \neq \mu 3$ or $\mu 2 \neq \mu 3$

respectively, where $\mu 1$ represents mean salary when education type is Doctorate, $\mu 2$ represents mean salary when education type is Bachelors and $\mu 3$ is the same for HS-Grad.

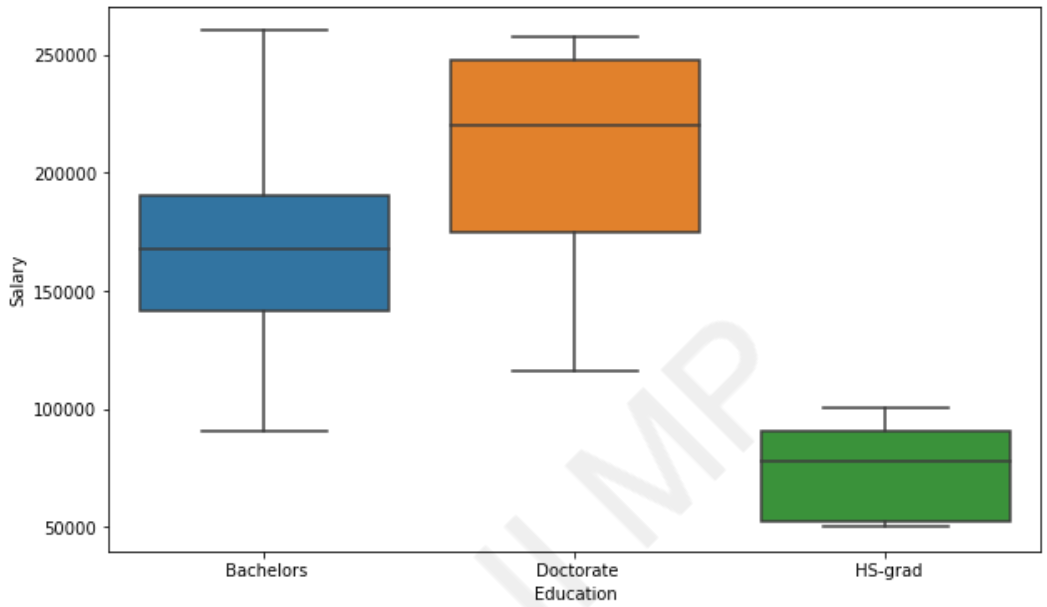


Fig.3 Box-plot of Salary w.r.t Education

From the above boxplot we can observe median salary of each Education level seems to vary from each other by a large margin.

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 4. One way ANOVA Table (Education w.r.t Salary)

Inference:

Based on the results from above table, we see that, the null hypothesis is rejected for all groups of Education level revealing that each group mean is significantly different from each other, thus also supporting the box-plot observations.

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.

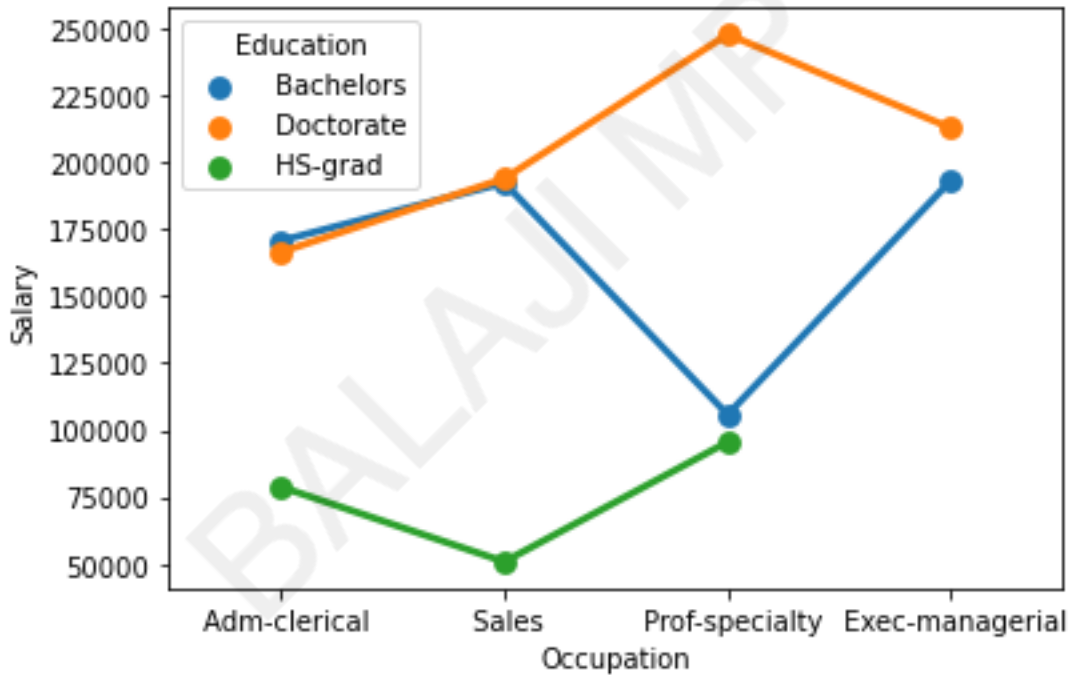


Fig.4 Interaction plot of Education & Occupation w.r.t Salary

From the interaction plot we can observe the following things:

- **Administrative** and **sales professionals** with **bachelor's** and **doctoral degrees** get similar earnings, whereas individuals with only **high school diplomas** receive low pay.
- Salary levels for **prof-speciality** professionals with **bachelor's degrees** and **high school diplomas** are similar, while those **with doctorates** make more money.
- **Executive-managerial** professionals with **PhD degrees** and **bachelor's degrees** earn comparable compensation, but people with only a **high school diploma** are not given preference for these roles.

Conclusion:

We can conclude from the above observations that an **interaction exists** between Education Occupation levels.

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?

Framing the hypothesis:

Let us assume a confidence level of 95% (α = 0.05) in all our hypotheses tests.

For Education Factor:

Education has three levels Doctorate, Bachelors & HS-grad. Let us assume the population means to be µ1, µ2, µ3 respectively.

H₀: $\mu 1 = \mu 2 = \mu 3$, the mean salary is same across all levels of education

H₁: $Not\ all\ \mu i$ are equal

For Occupation Factor:

Occupation has four levels Prof-specialty, Sales, Adm-clerical, Exec-managerial. Let us assume the population means to be µ1, µ2, µ3, µ4 respectively.

H₀: $\mu 1 = \mu 2 = \mu 3 = \mu 4$, the mean salary is same across all levels of occupation.

H₁: $Not\ all\ \mu i$ are equal

For Interaction:

Occupation has four levels Prof-specialty, Sales, Adm-clerical, Exec-managerial. Let us assume the population means to be μ1, μ2, μ3, μ4 respectively.

H_0: *The interaction effect does not exist*

H_1: *An interaction effect exist*

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	31.257677	1.981539e-08
C(Occupation)	3.0	5.519946e+09	1.839982e+09	1.120080	3.545825e-01
Residual	34.0	5.585261e+10	1.642724e+09	NaN	NaN

Table 4. Two Way ANOVA table without Interaction (Education, Occupation w.r.t Salary)

According to the following data, the mean earnings for all occupations are equal without interaction, although they differ for at least one pair of educational levels.

	df	sum_sq	mean_sq	F	\
C(Education)	2.0	1.026955e+11	5.134773e+10	72.211958	
C(Occupation)	3.0	5.519946e+09	1.839982e+09	2.587626	
C(Education):C(Occupation)	6.0	3.634909e+10	6.058182e+09	8.519815	
Residual	29.0	2.062102e+10	7.110697e+08	NaN	

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

Table 5. Two Way ANOVA table with Interaction (Education, Occupation w.r.t Salary)

Based on the above results, since the p-value is less than 0.05 for Education and Interaction, we reject the null hypothesis and since it is greater than 0.05 for Occupation, we fail to reject the null hypothesis. Hence, the **mean salaries across all Occupation levels are same and different for at least one pair of Education level**. Also, **an interaction effect exists between Education & Occupation**.

Inference:

Due to the effect of interaction, we see the p-value for the occupation component has been greatly reduced, as can be seen from the findings above, and is now closer to 0.05. However, as it is still higher than 0.05, the mean incomes across all occupation levels are still seen as being similar.

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

We can determine that Salary is dependent on Education and is driven by Occupation to a certain extent by doing an ANOVA on the provided data set.

Along with the connection between education and occupation, taking into account both factors while occupation is a non-significant variable with a P value of > 0.05, education is a significant factor with a P value of 0.05.

HS graduates had poor income packages in every occupation compared to those with Bachelor's and Doctorate degrees in administration and administration-clerical and sales, respectively.

When it comes to the business implications, education level has a significant impact on salary, while occupation paired with education level has a mild impact though not statistically significant.

Education PCA Analysis

Executive Summary

The collection includes statistics about a variety of colleges and universities. For this case study, a Principal Component Analysis must be performed in accordance with the provided guidelines.

Fig.5 Education PCA Analysis

Introduction

The purpose of this **report** is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. The data consists of **information on 777 colleges and universities.**

Data Description

1. **Names:** Names of various university and colleges (777 Unique values)
2. **Apps:** Continuous from 81 to 48094
3. **Accept:** Continuous from 72 to 26330
4. **Enroll:** Continuous from 35 to 6392
5. **Top10perc:** Continuous from 1 to 96
6. **Top25perc:** Continuous from 9 to 100
7. **F.Undergrad:** Continuous from 139 to 31643
8. **P.Undergrad:** Continuous from 1 to 21836
9. **Outstate:** Continuous from 2340 to 21700
10. **Room.Board:** Continuous from 1780 to 8124
11. **Books:** Continuous from 96 to 2340
12. **Personal:** Continuous from 250 to 6800
13. **PhD:** Continuous from 8 to 103
14. **Terminal:** Continuous from 24 to 100
15. **S.F.Ratio:** 2.5 to 39.8
16. **perc.alumni:** Continuous from 0 to 64
17. **Expend:** Continuous from 3186 to 56233
18. **Grad.Rate:** Continuous from 10 to 118

Sample of the dataset

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

Table.5 Sample of first 5 rows of the dataset

Exploratory Data Analysis

Let us check the types of variables in the data frame and check for missing values in the dataset

```
<class 'pandas.core.frame.DataFrame'>
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)
memory usage: 109.4+ KB
```

Fig.6 Details of the dataset columns

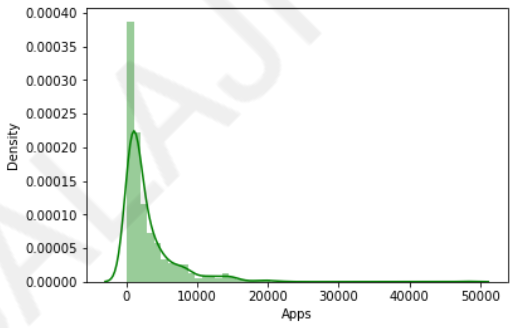
There are total of 777 rows and 18 columns in the dataset. Out of 18 columns, 16 columns are integer and the remaining 2 columns are float and string respectively. Also, from the above results we can see that there are no missing values present in dataset.

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

Univariate Analysis

```
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
```




```
----------------------------------------------------------------------
BoxPlot of Apps
```

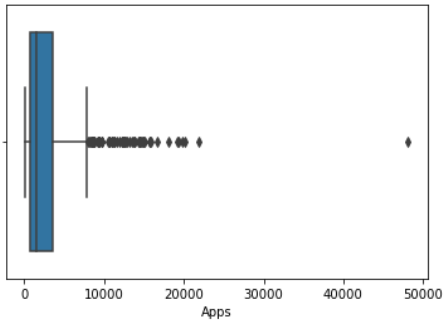
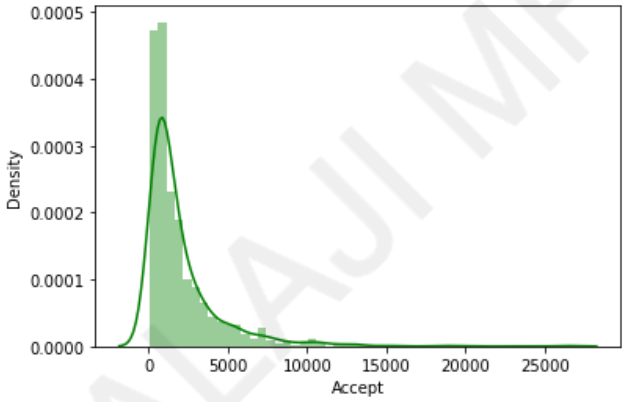


Fig.7 Descriptive Summary, Histogram & Boxplot of Apps column

With an average of 3001 applications, the number of applications received ranges from 81 to 48094. The histogram shows that the distribution of Apps is heavily right-skewed. A significant number of outliers are present, as seen from the box plot contributing to the skewness.

```
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
```




```
----------------------------------------------------------
BoxPlot of Accept
```

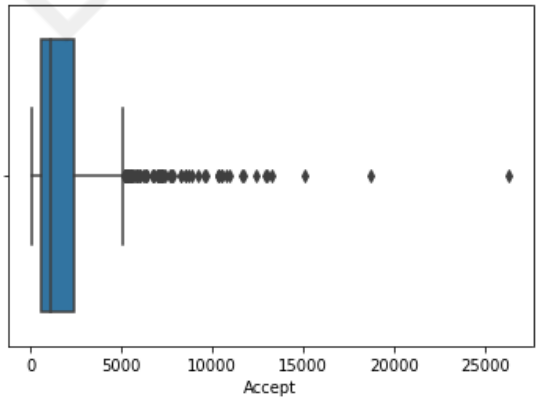
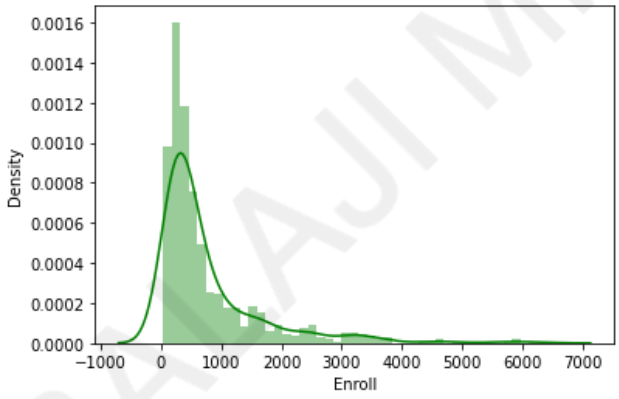


Fig.8 Descriptive Summary, Histogram & Boxplot of Accept column

With an average of 2018 applications being accepted by the college or university, the number of applications accepted ranges from 72 to 26330. The histogram shows that the distribution of Accept is heavily right-skewed. A significant number of outliers are present, as seen from the box plot contributing to the skewness.

```
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
```




```
------------------------------------------------------------
BoxPlot of Enroll
```

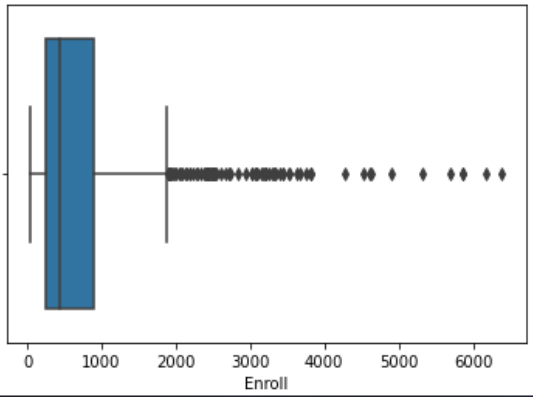
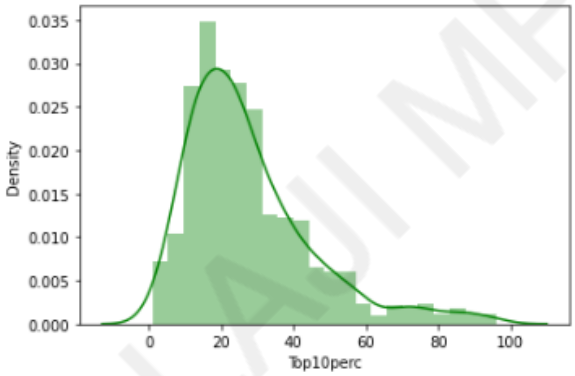


Fig.9 Descriptive Summary, Histogram & Boxplot of Enroll column

With an average of 779 students being enrolled, the number of enrollment ranges from 35 to 6392. The histogram shows that the distribution of Enroll is heavily right-skewed. A significant number of outliers are present, as seen from the box plot contributing to the skewness.

```
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
Name: Top10perc, dtype: float64
----------------------------------------------------------
Distribution of Top10perc
```




```
----------------------------------------------------------
BoxPlot of Top10perc
```

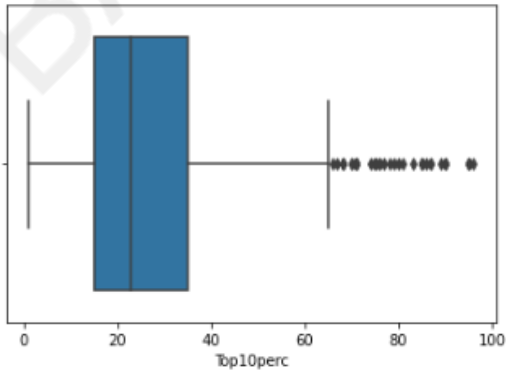
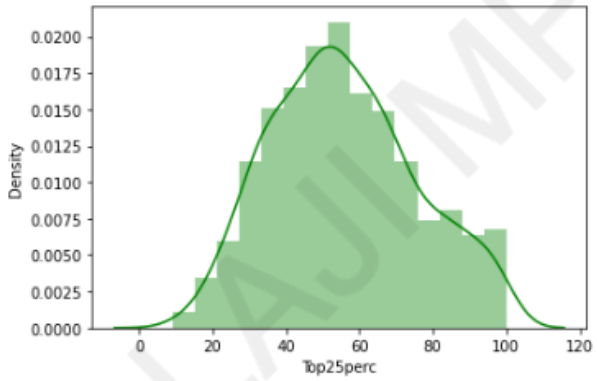


Fig.10 Descriptive Summary, Histogram & Boxplot of Top10perc column

An average of 28% of students from top 10% of Higher Secondary class are enrolled and the percentage ranges from 1 to 96. The histogram shows that the distribution of Top10perc is slightly right-skewed. A moderate number of outliers are present, as seen from the box plot contributing to the skewness.

```
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
```




```
----------------------------------------------------------
BoxPlot of Top25perc
```

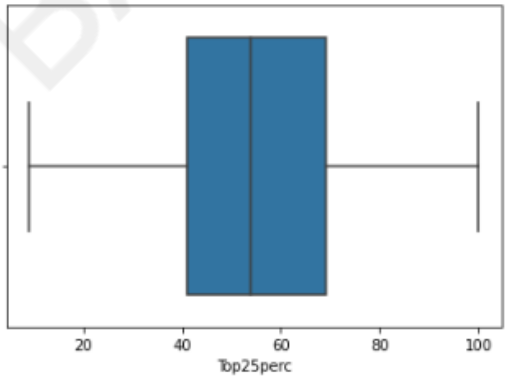
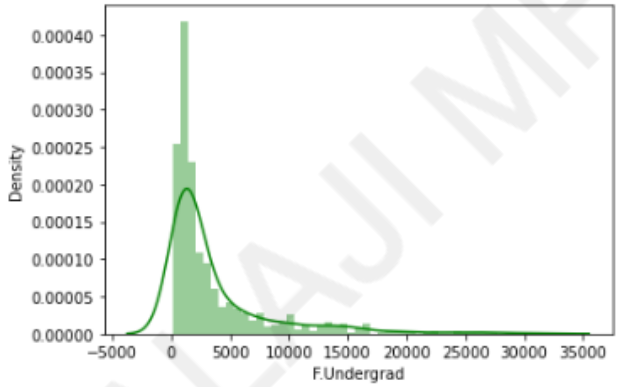


Fig.11 Descriptive Summary, Histogram & Boxplot of Top25perc column

An average of 55% of students from top 25% of Higher Secondary class are enrolled and the percentage ranges from 9 to 100. The histogram shows that the Top25perc seems to be normally distributed. No outliers are present, as seen from the box plot contributing to the normality.

```
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
```




```
------------------------------------------------------------
BoxPlot of F.Undergrad
```

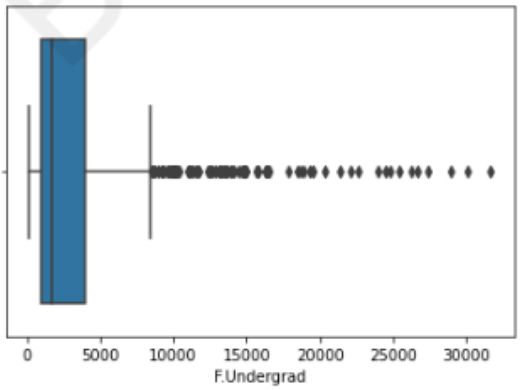
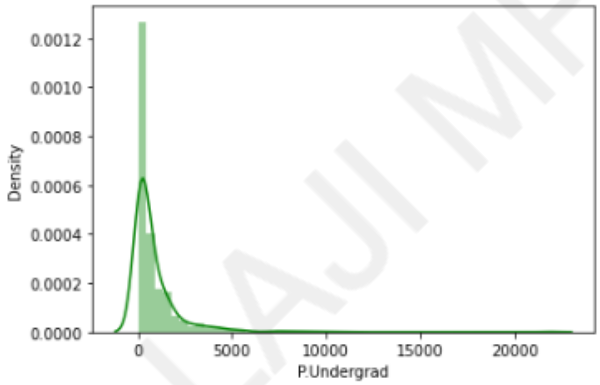


Fig.12 Descriptive Summary, Histogram & Boxplot of F.Undergrad column

On average, 3699 students are full-time undergraduate students and the number ranges from 139 to 992. The histogram shows that the distribution of F.Undergrad is heavily right-skewed. A significant number of outliers are present, as seen from the box plot contributing to the skewness.

```
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
```




```
----------------------------------------------------------
BoxPlot of P.Undergrad
```

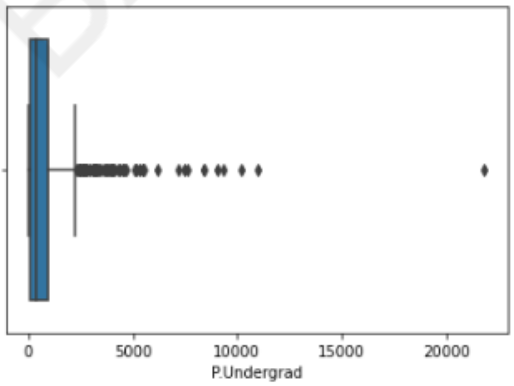


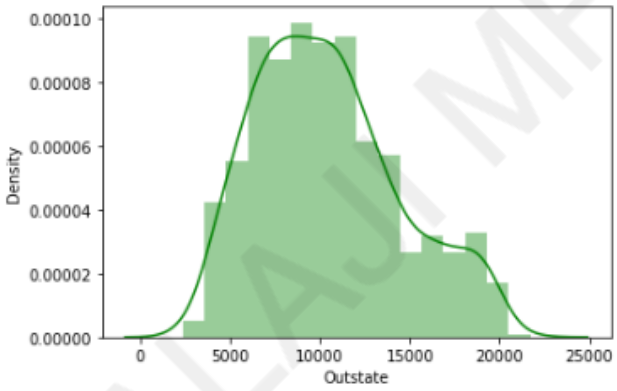
Fig.13 Descriptive Summary, Histogram & Boxplot of P.Undergrad column

On average, 856 students are part-time under-graduate students and the number ranges from 1 to 21836. The histogram shows that the distribution of P.Undergrad is heavily right-skewed. A significant number of outliers are present, as seen from the box plot contributing to the skewness.

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



BoxPlot of Outstate

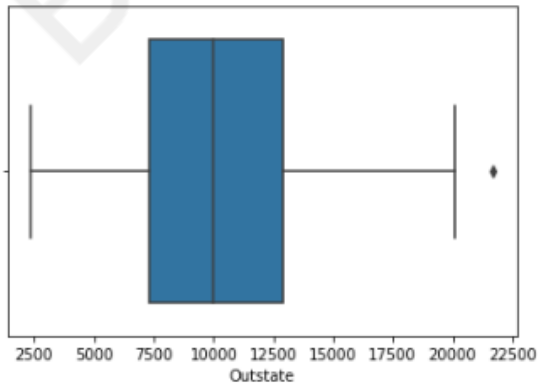
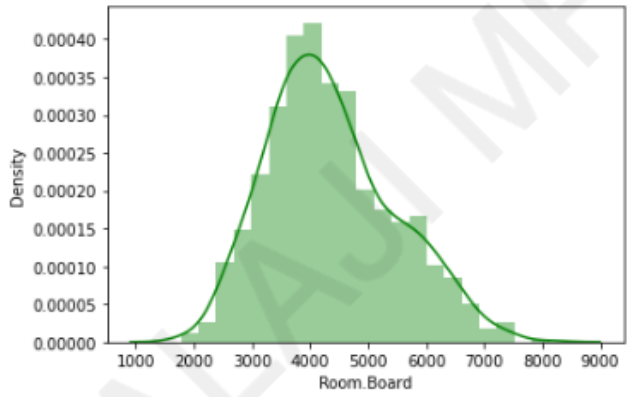


Fig.14 Descriptive Summary, Histogram & Boxplot of Outstate column

On average, for 10440 students a particular college or university is Out-of-state tuition and the number ranges from 2340 to 21700. The histogram shows that the Outstate seems to be normally distributed. Only one outlier is present, as seen from the box plot contributing to the normality.

```
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
```




```
----------------------------------------------------------
BoxPlot of Room.Board
```

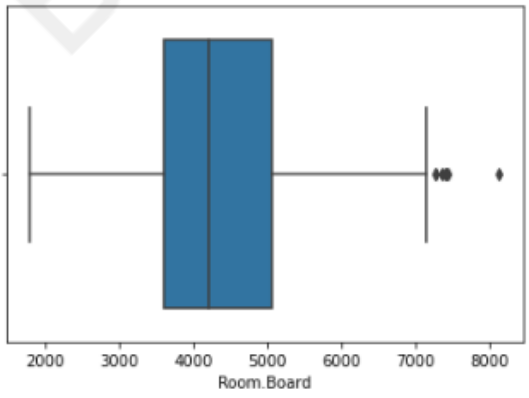
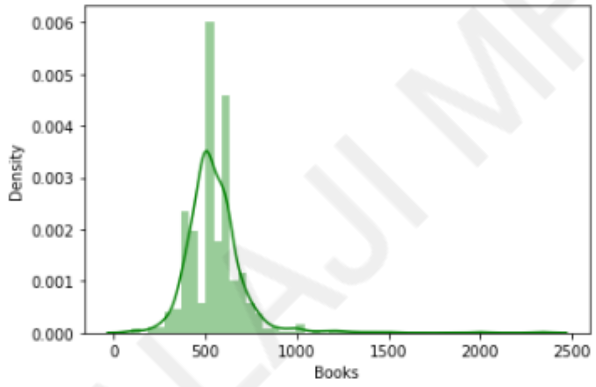


Fig.15 Descriptive Summary, Histogram & Boxplot of Room.Board column

On average, Cost of Room and board is 4357 and the cost ranges from 1780 to 8124. The histogram shows that the Room.Board seems to be normally distributed. Few outliers are present, as seen from the box plot.

```
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
```




```
----------------------------------------------------------
BoxPlot of Books
```

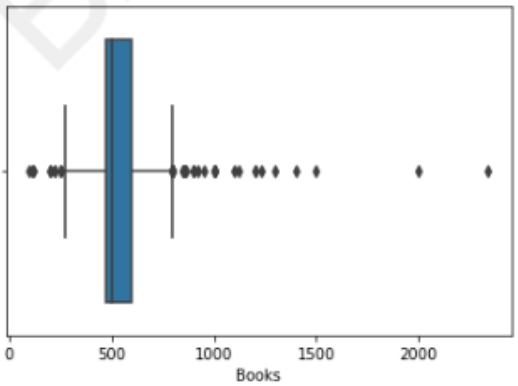
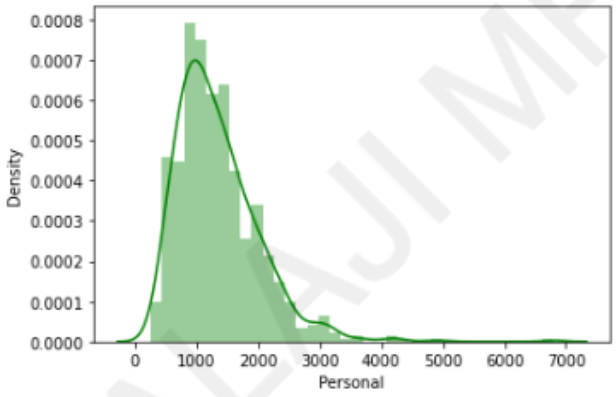


Fig.16 Descriptive Summary, Histogram & Boxplot of Books column

On average, estimated book costs for a student is 550 and the cost ranges from 96 to 2340. The histogram shows that the Books seems to be heavily right skewed. A significant number of outliers are present, as seen from the box plot contributing to the skewness.

```
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
```




```
----------------------------------------------------------
BoxPlot of Personal
```

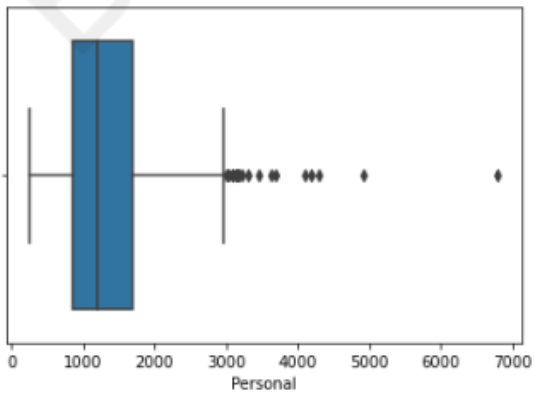
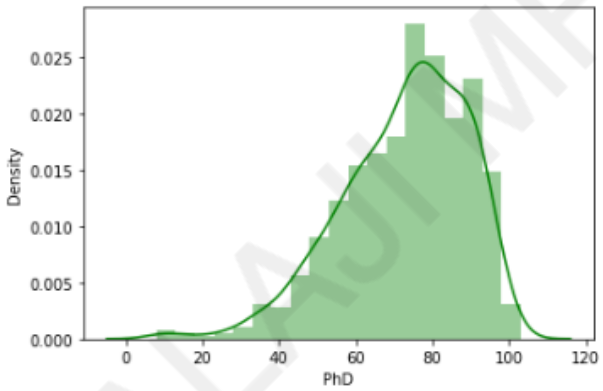


Fig.17 Descriptive Summary, Histogram & Boxplot of Personal column

On average, estimated personal spending for a student is 1340 and the spending ranges from 250 to 850. The histogram shows that the Personal seems to be slightly right skewed. A moderate number of outliers are present, as seen from the box plot contributing to the skewness.

```
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
```




```
------------------------------------------------------------
BoxPlot of PhD
```

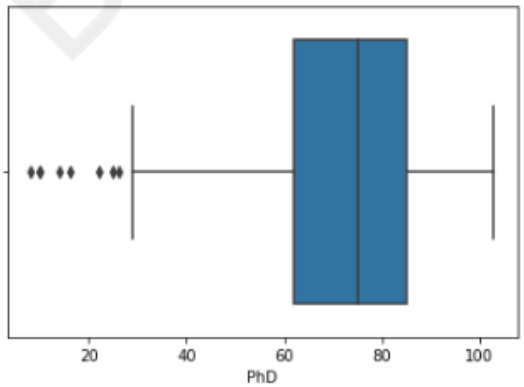
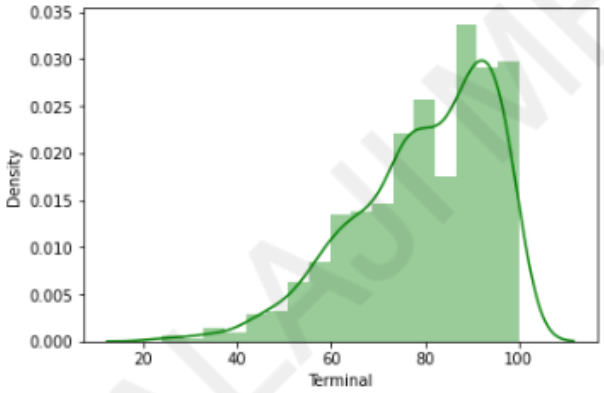


Fig.18 Descriptive Summary, Histogram & Boxplot of PhD column

With an average of 72% of faculties with Ph.D.'s, the percentage ranges from 8 to 103. The histogram shows that the PhD seems to be slightly left skewed. Very few outliers are present, as seen from the box plot contributing to the skewness.

```
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
```




```
------------------------------------------------------------
BoxPlot of Terminal
```

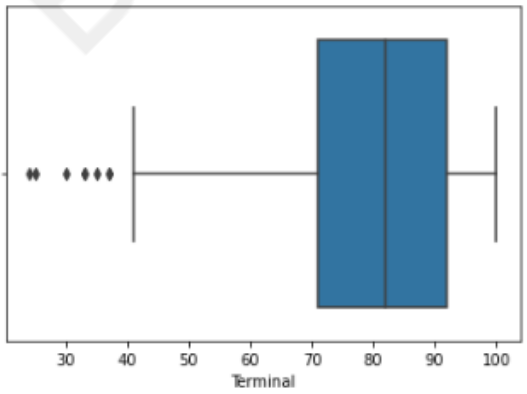
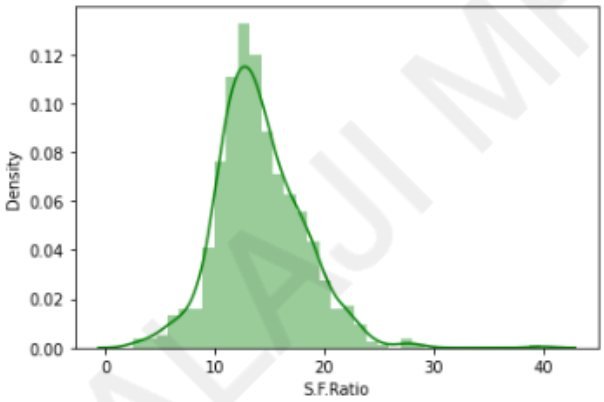


Fig.19 Descriptive Summary, Histogram & Boxplot of Terminal column

With an average of 80% of faculties with terminal degree, the percentage ranges from 24 to 100. The histogram shows that the Terminal seems to be slightly left skewed. Very few outliers are present, as seen from the box plot contributing to the skewness.

```
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
```




```
----------------------------------------------------
BoxPlot of S.F.Ratio
```

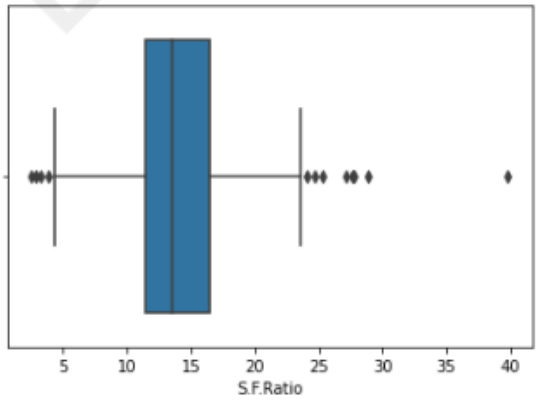
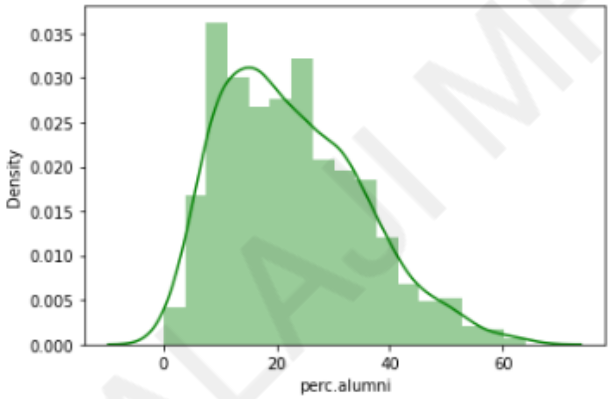


Fig.20 Descriptive Summary, Histogram & Boxplot of S.F.Ratio column

On average, the ratio of students to faculty is 14 and it ranges from 2.5 to 40. The histogram shows that the S.F.Ratio seems to be slightly right skewed. Very few outliers are present, as seen from the box plot contributing to the skewness.

```
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
```




```
----------------------------------------------------------
BoxPlot of perc.alumni
```

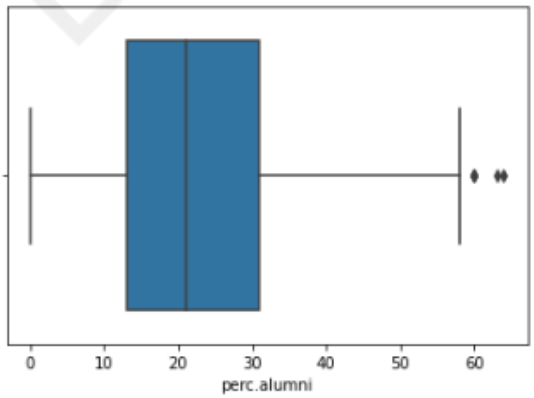
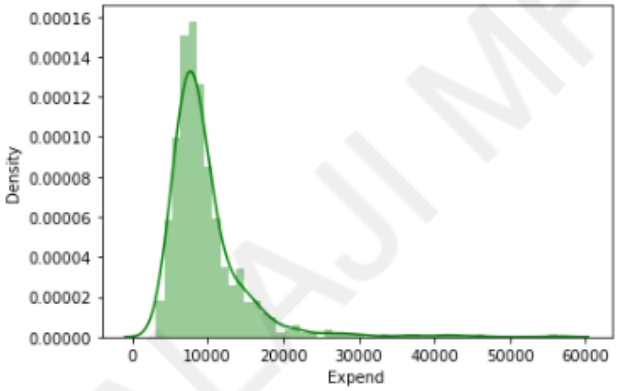


Fig.21 Descriptive Summary, Histogram & Boxplot of perc.alumni column

On average, 23% of alumni donates to their college or university and the percentage ranges from 0 to 64. The histogram shows that the perc.alumni seems to be normally distributed. Few outliers are present, as seen from the box plot.

```
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
```




```
------------------------------------------------------------
BoxPlot of Expend
```

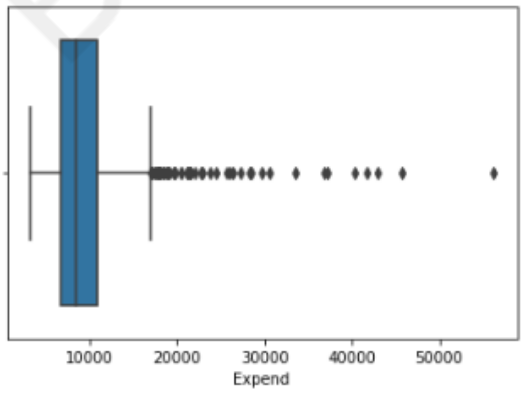
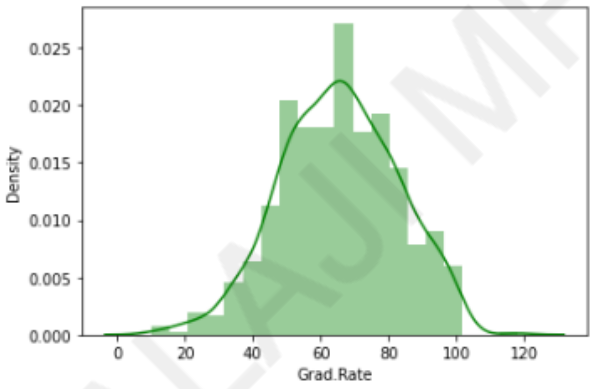


Fig.22 Descriptive Summary, Histogram & Boxplot of Expend column

The average Instructional expenditure per student is 9660 and it ranges from 3186 to 56233. The histogram shows that the Expend is heavily right skewed. A significant number of outliers are present, as seen from the box plot contributing to the skewness.

```
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
```




```
----------------------------------------------------------
BoxPlot of Grad.Rate
```

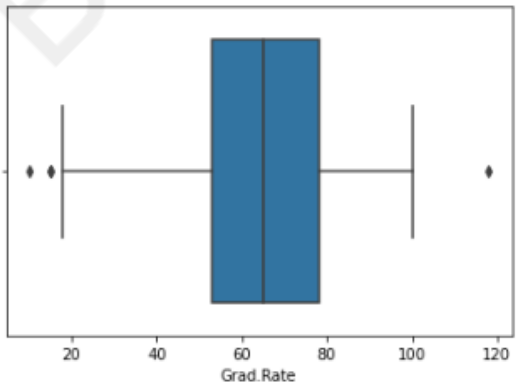


Fig.23 Descriptive Summary, Histogram & Boxplot of Grad.Rate column

On average, 66% of students graduate from college or university per year and the percentage ranges from 10 to 118. The histogram shows that the Grad Rate seems to be normally distributed. Few outliers are present, as seen from the box plot.

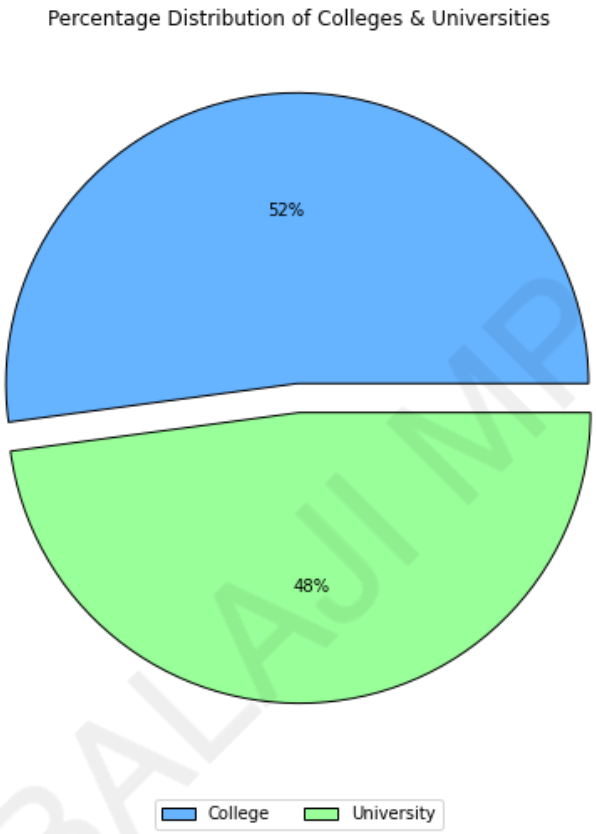


Fig.24 Distribution of Colleges & University

It can be seen from the above pie chart, 52% of dataset is from colleges and remaining 48% contains universities.

Bivariate Analysis

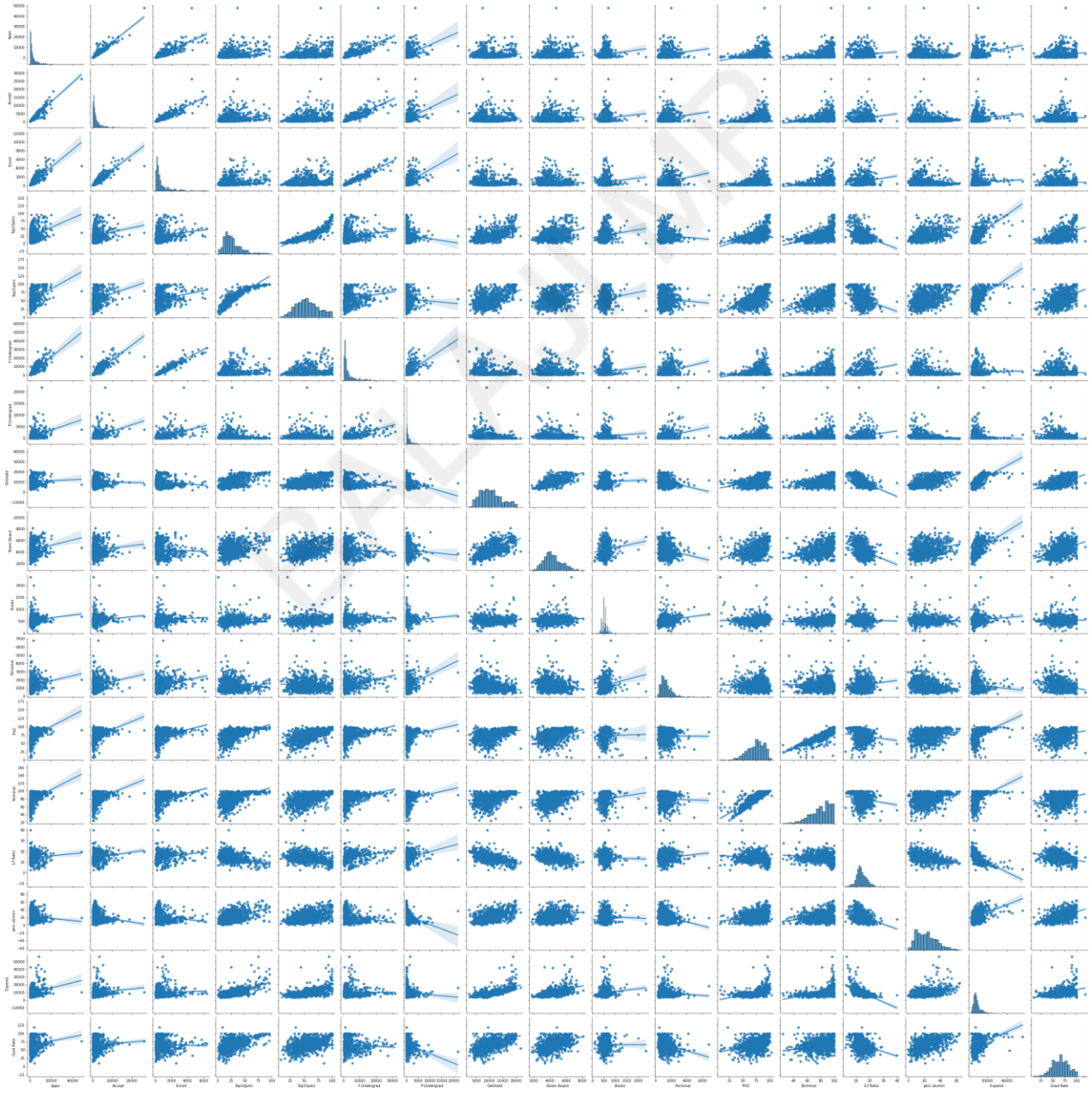


Fig.25 Pair plot of all the features

From the pair plot, we can see that some features are correlated to some extent. Let us confirm this with a heat map.

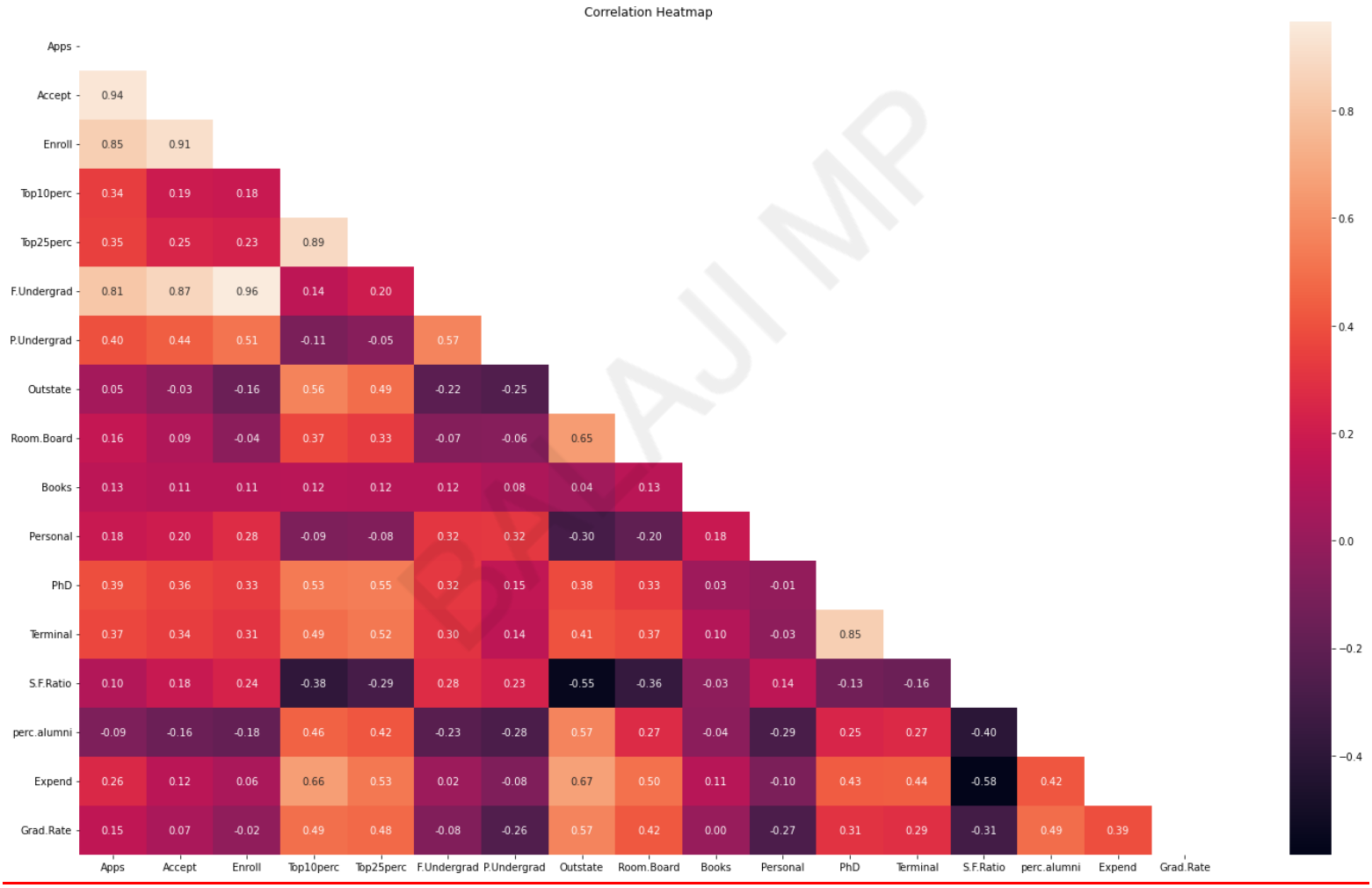


Fig.26 Correlation Heatmap

Inference:

- **There are a large number of highly connected variables.**
- **Apps is highly correlated with Accept and Enroll**
- **PhD and Terminal features show high correlation**

Insights from Univariate & Bivariate Analysis:

- **A college or university typically receives 3002 applications each year. Of those, 2019 applications (or almost 67 percent) are accepted, and nearly 40% of those accepted students enroll.**
- **There are typically four full-time undergraduate students for part-time undergraduate student (approximately)**
- **The cost of the book is typically approximately $550.**
- **About 75% of a student's personal expenses fall below $1700.**
- **Universities and colleges maintain a student to faculty ratio of no less than three and no more than forty.**
- **65 percent of people receive their degrees on average.**
- **There are a sizable number of highly correlated traits.**
- **“Apps”, "Accept" and "Enroll" have a strong link.**

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

Yes, scaling is required in order to move further with PCA analysis. When determining the principal component axis, PCA considers the highest variance that may be seen in a certain direction. A variable with a high variance or standard deviation would be given greater weight when determining the axis than a variable with a low variance if there was no scaling.

Scaling the features is therefore essential in order to properly determine the axis for the principal components.

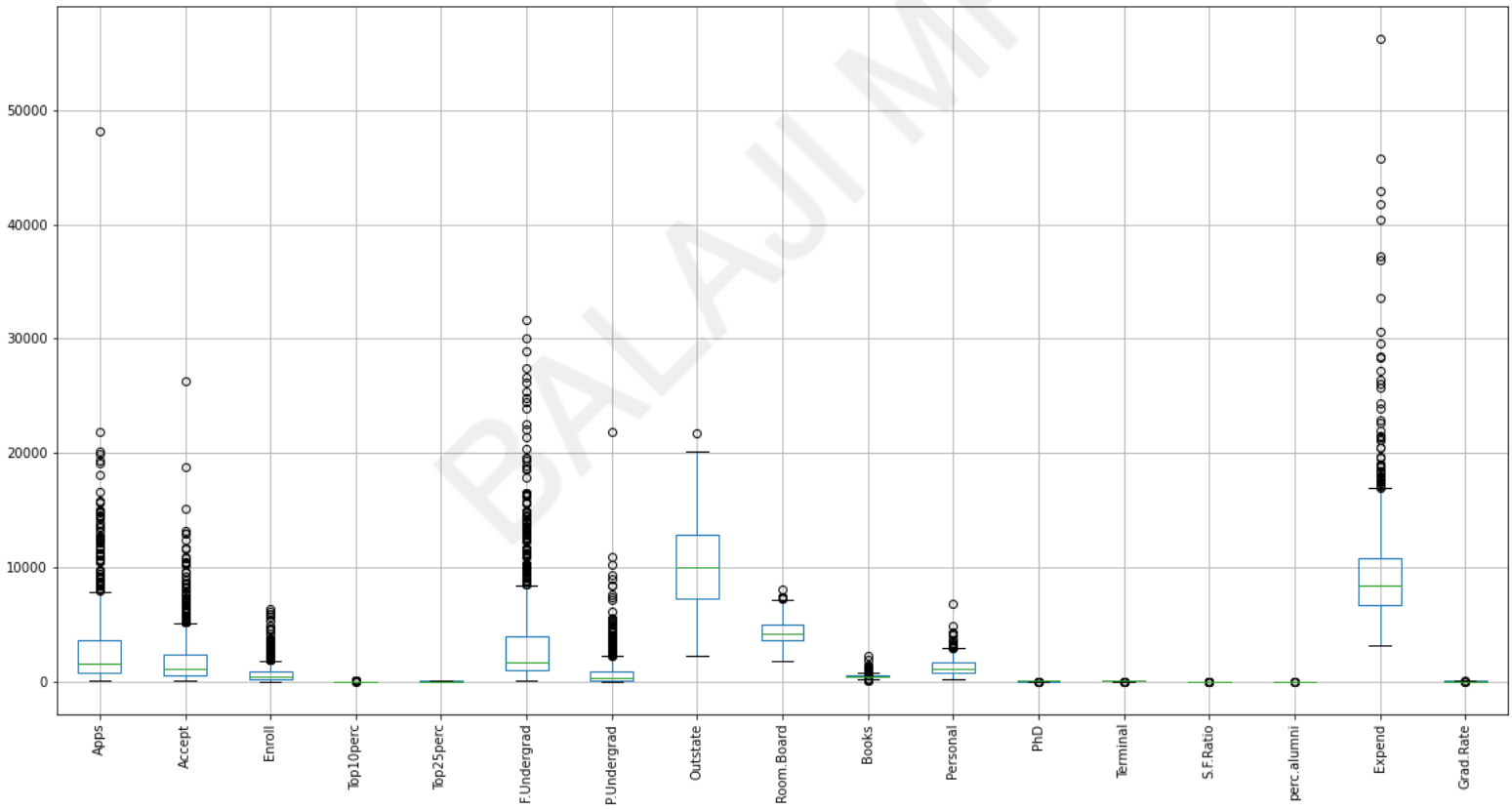


Fig.27 Boxplot of the features without scaling

As we can see, the units of measurement for applications, acceptance, and enrollment are all numerical, but those for room and board, books, and personal items are monetary and all of them have different means. We must thus combine them all into a single scale of measurement in order to conduct further analysis.

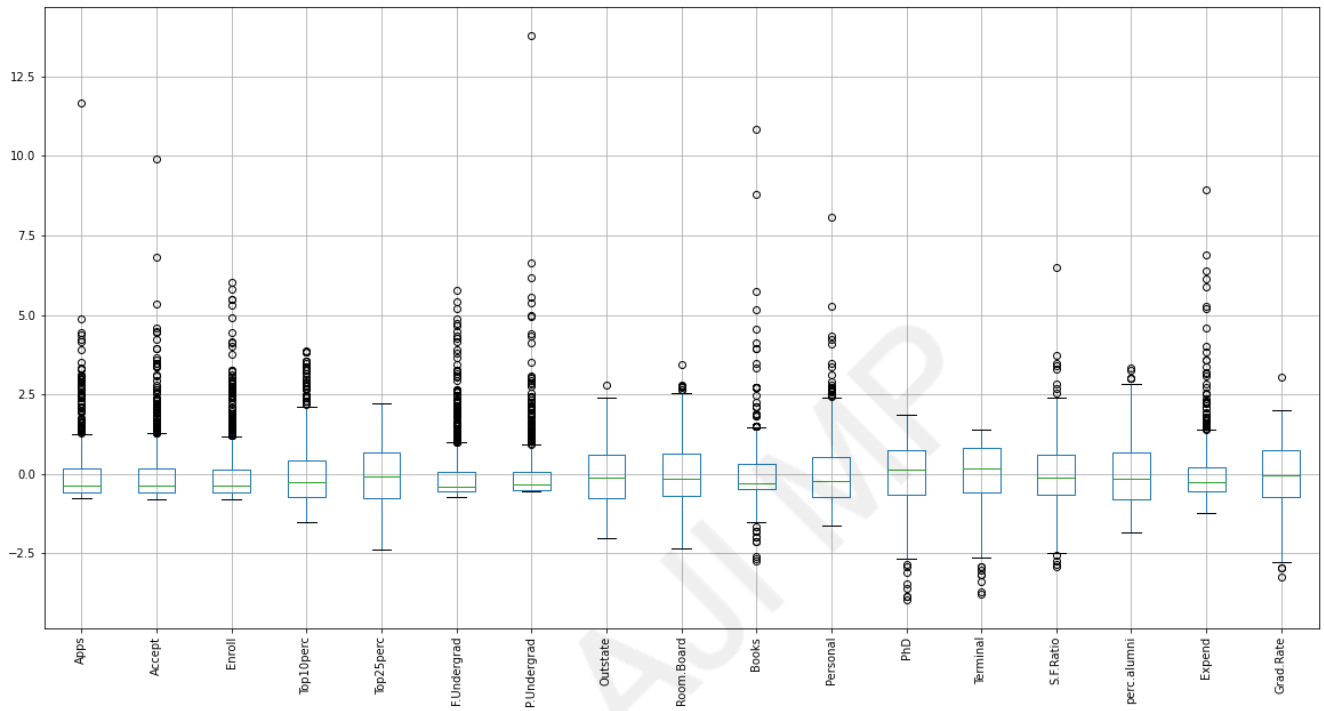


Fig.28 Boxplot of the features after scaling

All characteristics have now been scaled with a mean that is centered around 0 and a standard deviation of 1.

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

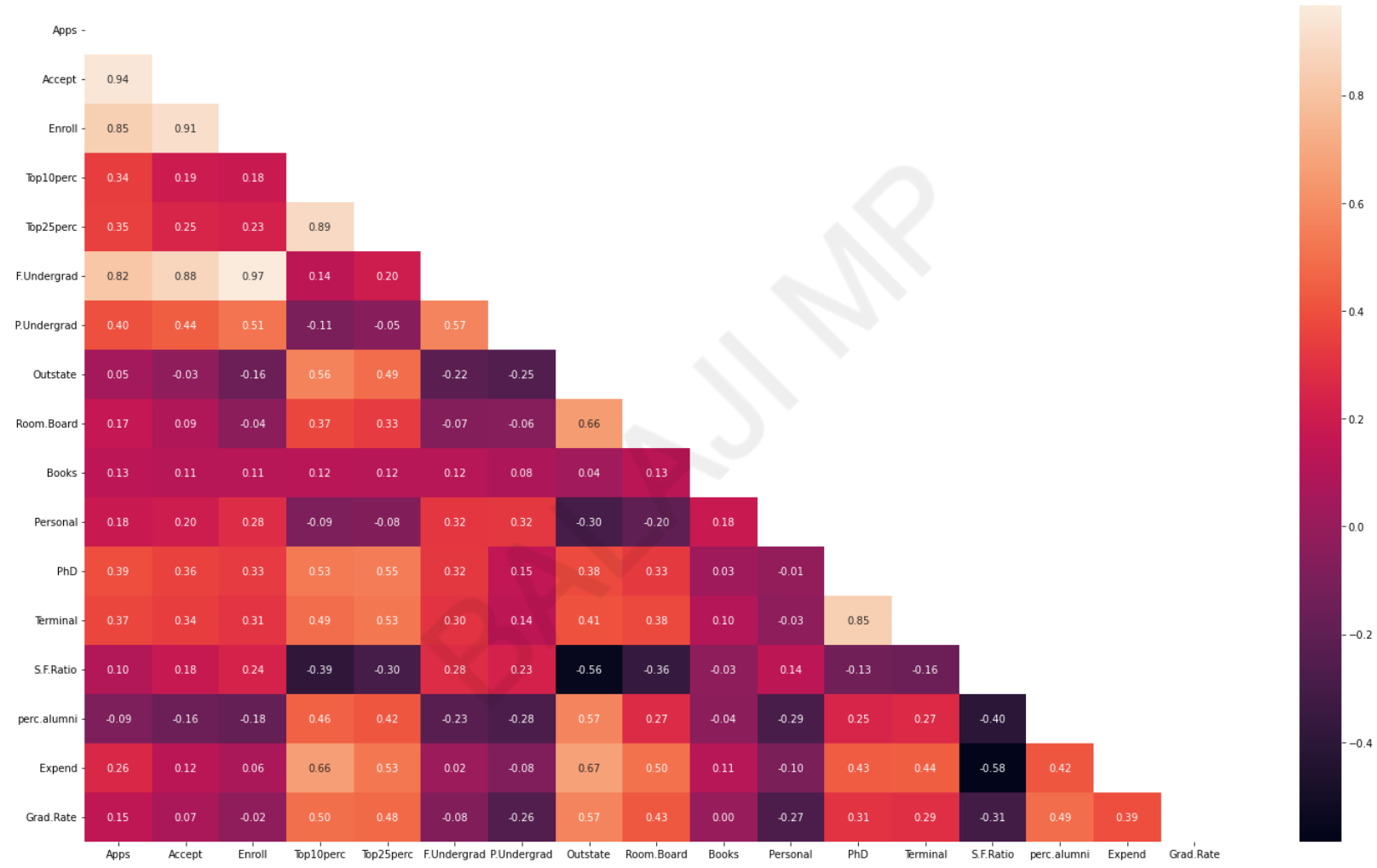


Fig.29 Covariance Heatmap

Covariance is a statistical term that refers to a systematic relationship between two random variables in which a change in the other reflects a change in one variable. The covariance value can range from -∞ to +∞, with a negative value indicating a negative relationship and a positive value indicating a positive relationship.

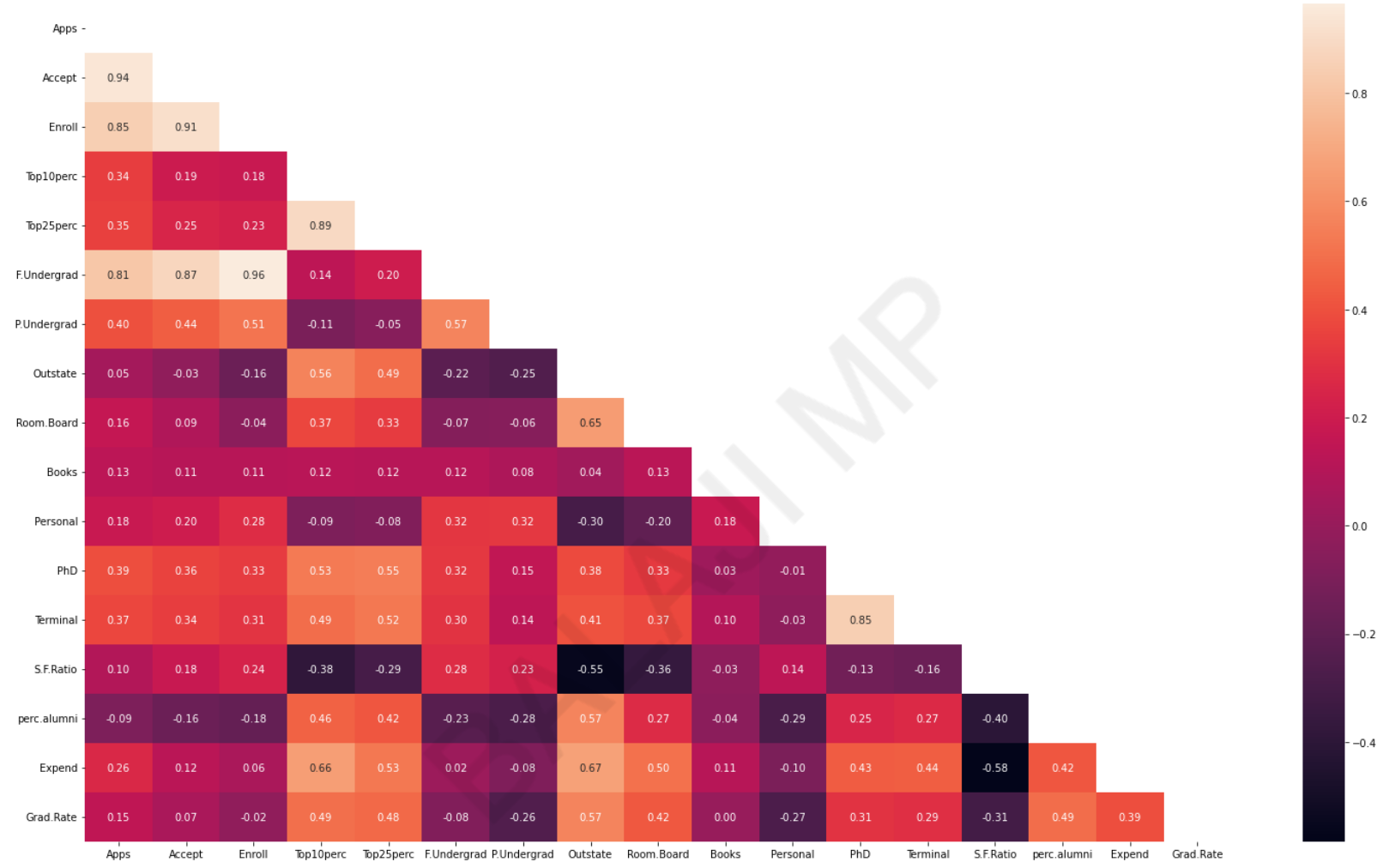


Fig.30 Correlation Heatmap

Correlation is a statistical measure that indicates how strongly two variables are related. The correlation value ranges from -1 to +1.

Inference:

- Correlation can also be considered as a scaled covariance, where usually covariance varies from -∞ to +∞ and correlation converts them into range -1 to +1.
- Since the correlation and covariance are calculated on the scaled dataset, the values are almost similar to each other as seen from comparing the two heatmaps with their respective covariance and correlation values.

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

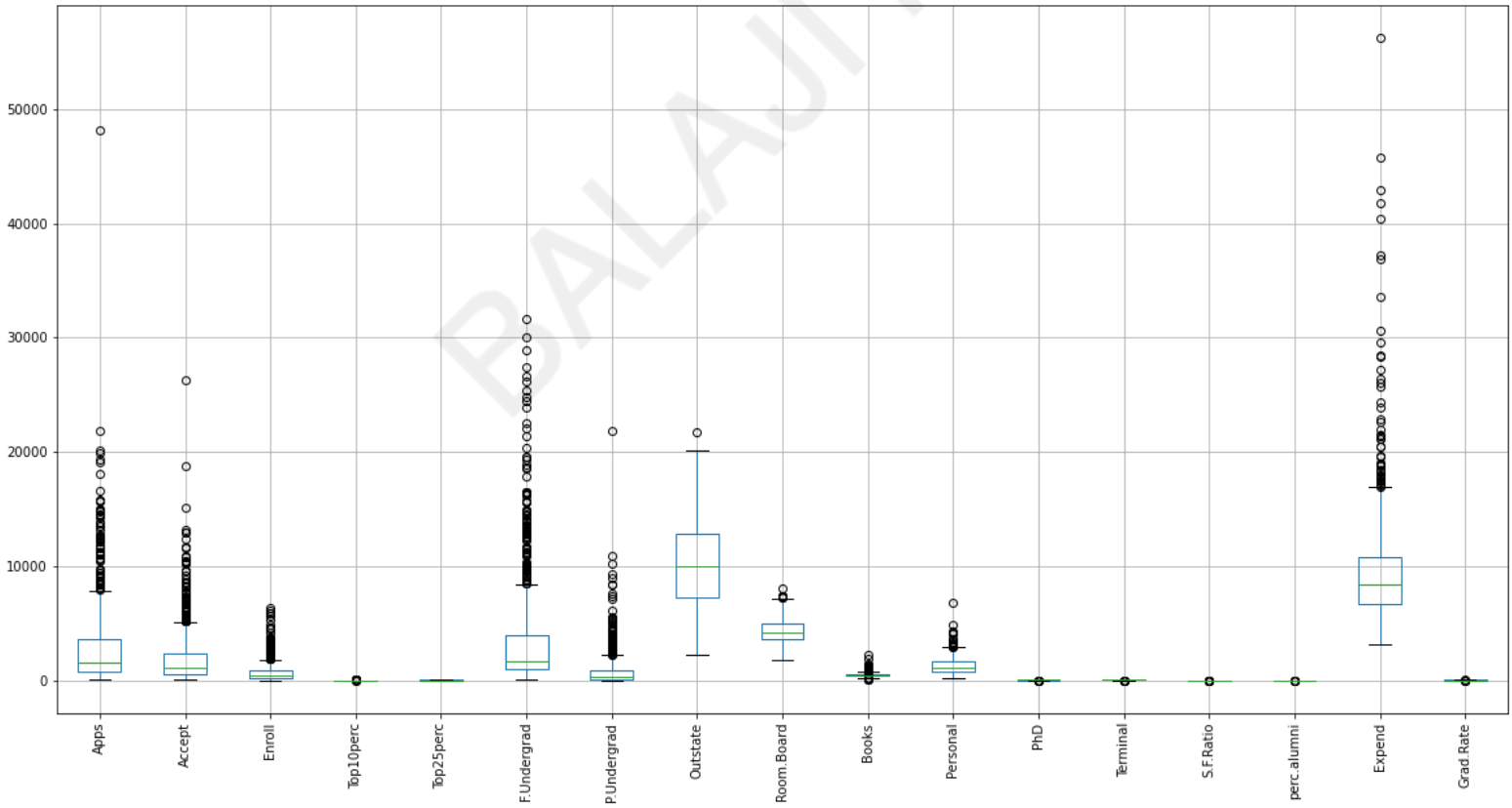


Fig.31 Boxplot of the features without scaling

After doing scaling using Z score, we obtain the following result

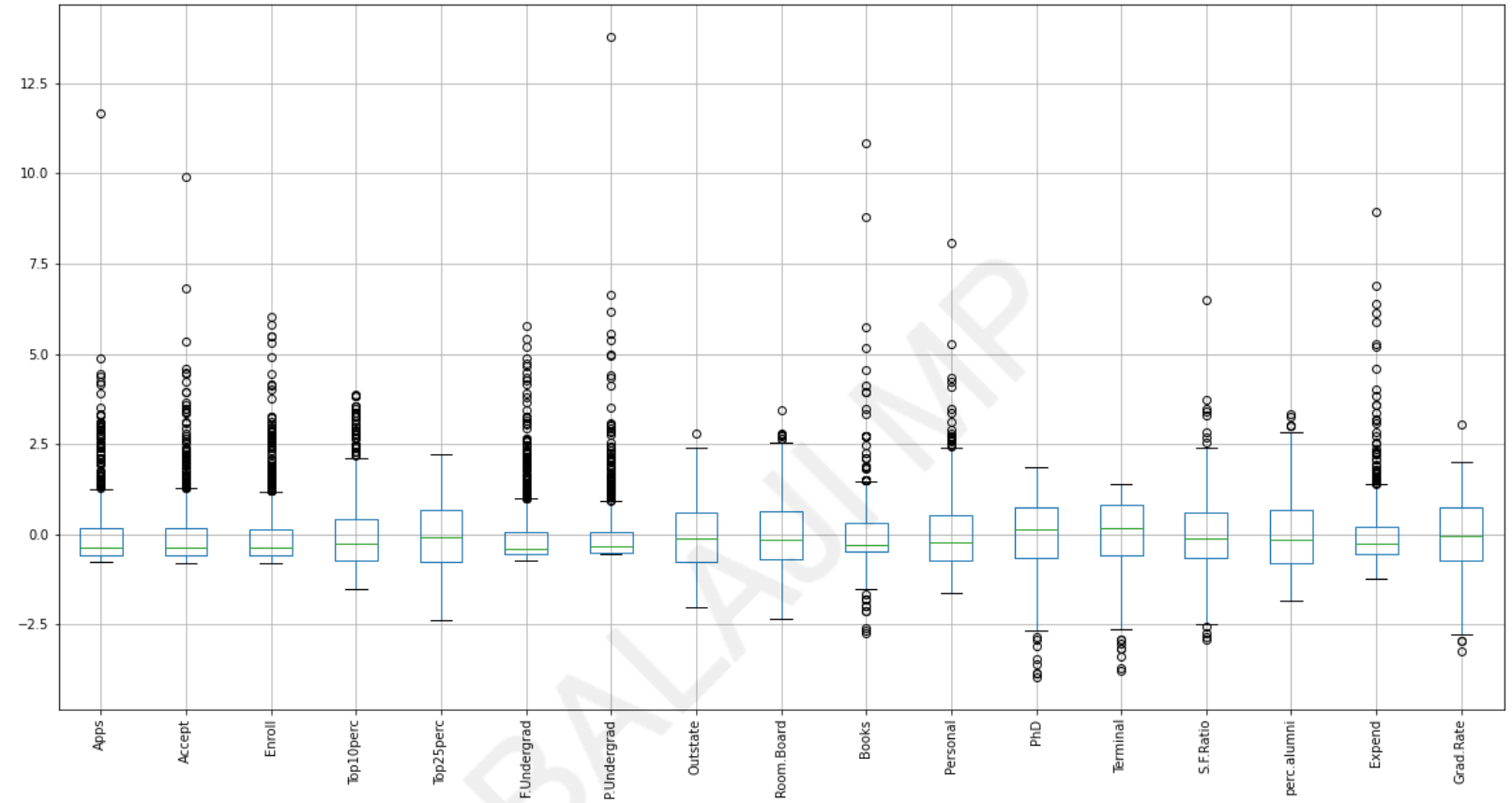


Fig.32 Boxplot of the features after scaling

Insights:

- **Both before and after scaling, there are outliers.**
- **All of the variables have unique means and corresponding standard deviations prior to scaling. All the outliers are positive (greater than 0).**
- **After scaling, each variable has a mean of 0 and a standard deviation of 1. This will guarantee that all the variables have the same weight, assisting PCA analysis. Additionally, we can see the existence of negative outliers brought on by scaling.**

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

```
Eigen Values
[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]
```

Fig.33 Eigen Values

```
Eigen Vectors
[[ 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]
```

Fig.34.1 Eigen Vectors

```
[ 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]]
```

Fig.34.2 Eigen Vectors

According to the total number of numerical variables, there are a total of 17 eigen values and their associated eigen vectors. Scikit Learn PCA is used to acquire them.

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

PCA steps to be followed:

- **Standardize the range of continuous numerical variables**
- **Compute the covariance matrix to identify correlations**
- **Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components**
- **Determine the number of principal components using scree plot and cumulative sum of variances**
- **Perform PCA on the scaled dataset for the decided number of components which explains 70% to 80% of the total variance.**

We have already performed the first 3 steps before. Let us continue from step 4.

Scree Plot:

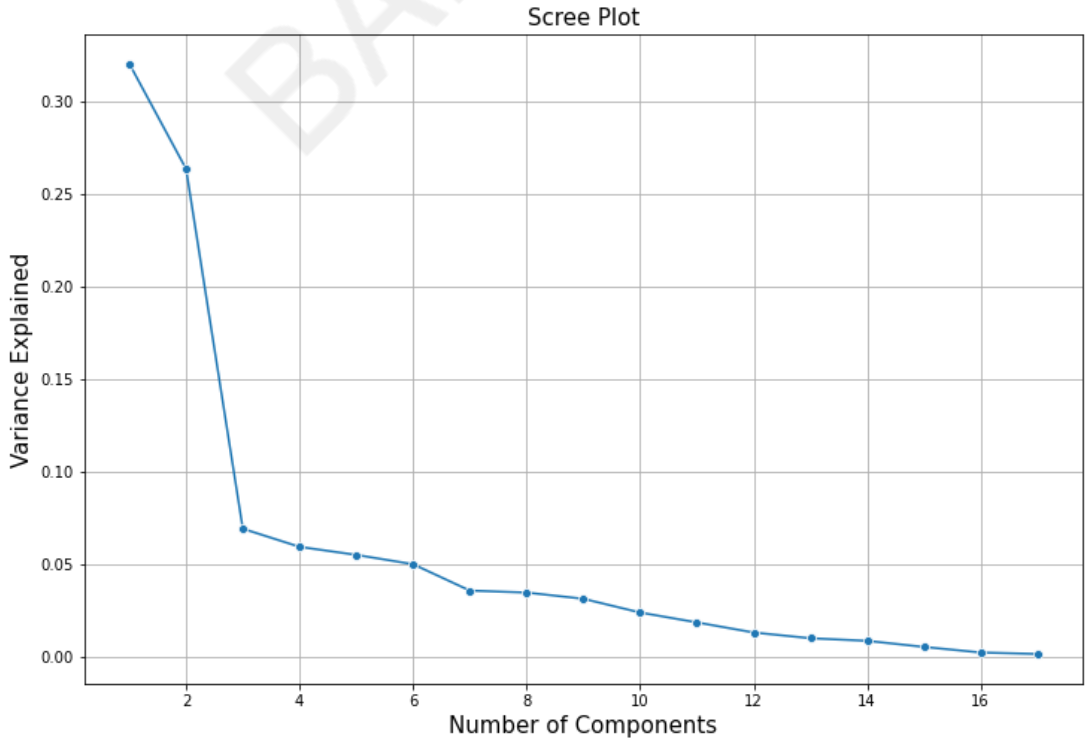


Fig.35 Scree Plot

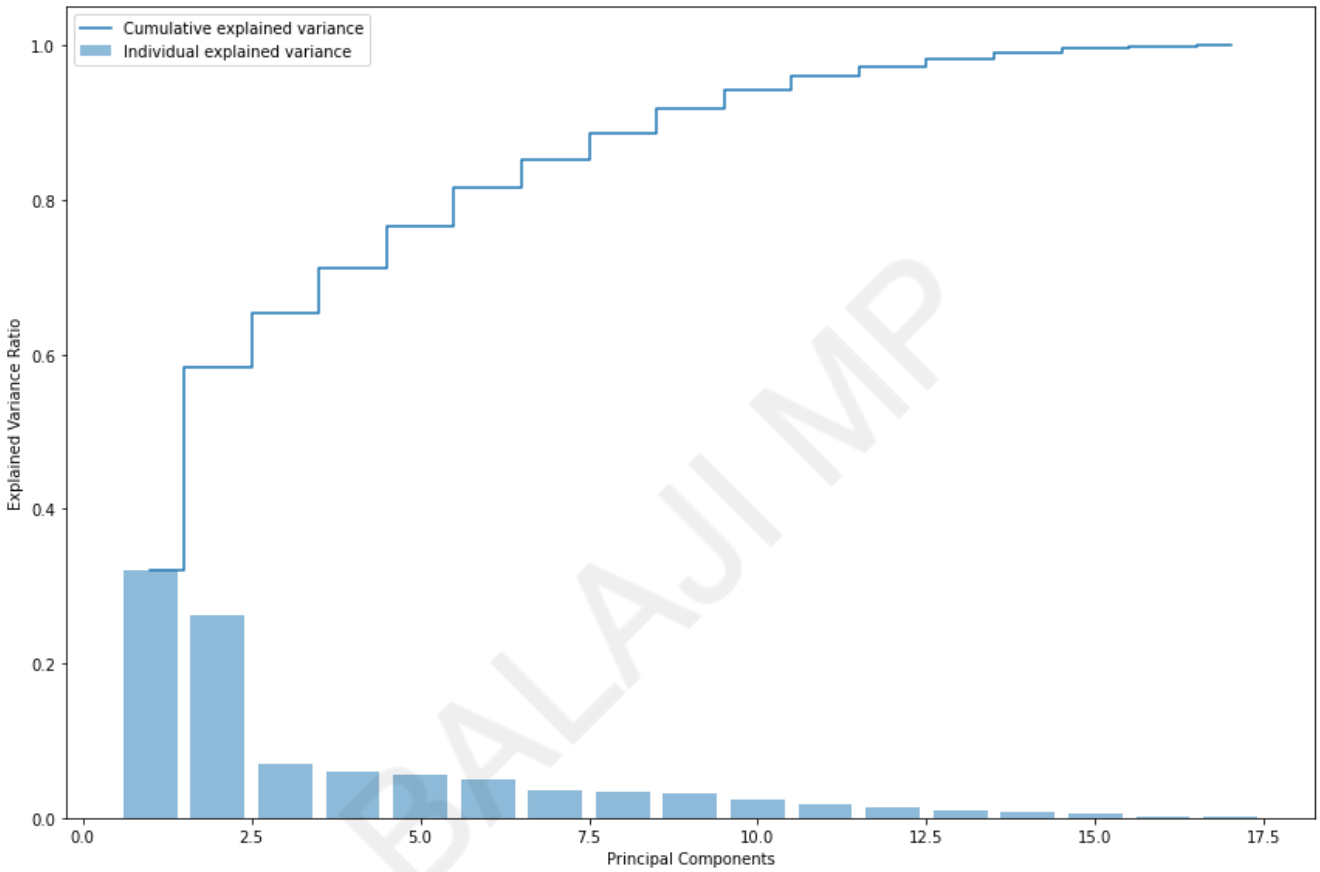


Fig.36 Individual & Cumulative variance

The first 6 principal components account for 81 percent of the total information (variance) in the dataset, according to the aforementioned 2 plots. So, 6 principal components may be used to conduct PCA.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 17 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Apps         6 non-null      float64
 1   Accept       6 non-null      float64
 2   Enroll       6 non-null      float64
 3   Top10perc    6 non-null      float64
 4   Top25perc    6 non-null      float64
 5   F.Undergrad  6 non-null      float64
 6   P.Undergrad  6 non-null      float64
 7   Outstate     6 non-null      float64
 8   Room.Board   6 non-null      float64
 9   Books        6 non-null      float64
 10  Personal     6 non-null      float64
 11  PhD          6 non-null      float64
 12  Terminal     6 non-null      float64
 13  S.F.Ratio    6 non-null      float64
 14  perc.alumni  6 non-null      float64
 15  Expend       6 non-null      float64
 16  Grad.Rate    6 non-null      float64
dtypes: float64(17)
memory usage: 944.0 bytes
```

Fig.37 Information of new Data frame with original features

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Rati
0	0.248766	0.207602	0.176304	0.354274	0.344001	0.154641	0.026443	0.294736	0.249030	0.064758	-0.042529	0.318313	0.317056	-0.1769
1	0.331598	0.372117	0.403724	-0.082412	-0.044779	0.417674	0.315088	-0.249644	-0.137809	0.056342	0.219929	0.058311	0.046429	0.2466
2	-0.063092	-0.101249	-0.082986	0.035056	-0.024148	-0.061393	0.139682	0.046599	0.148967	0.677412	0.499721	-0.127028	-0.066038	-0.2898
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	-0.1611
4	0.005741	0.055786	-0.055694	-0.395434	-0.426534	-0.043454	0.302385	0.222532	0.560919	-0.127289	-0.222311	0.140166	0.204720	-0.0793
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	0.4870

Table 6. New data frame with original features

First Principal Component

With all the features, a new data frame has been produced with the corresponding eigen vectors. Each row in this data frame corresponds to a principal component.

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]

	0	1	2	3	4	5
Apps	0.25	0.33	-0.06	0.28	0.01	-0.02
Accept	0.21	0.37	-0.10	0.27	0.06	0.01
Enroll	0.18	0.40	-0.08	0.16	-0.06	-0.04
Top10perc	0.35	-0.08	0.04	-0.05	-0.40	-0.05
Top25perc	0.34	-0.04	-0.02	-0.11	-0.43	0.03
F.Undergrad	0.15	0.42	-0.06	0.10	-0.04	-0.04
P.Undergrad	0.03	0.32	0.14	-0.16	0.30	-0.19
Outstate	0.29	-0.25	0.05	0.13	0.22	-0.03
Room.Board	0.25	-0.14	0.15	0.18	0.56	0.16
Books	0.06	0.06	0.68	0.09	-0.13	0.64
Personal	-0.04	0.22	0.50	-0.23	-0.22	-0.33
PhD	0.32	0.06	-0.13	-0.53	0.14	0.09
Terminal	0.32	0.05	-0.07	-0.52	0.20	0.15
S.F.Ratio	-0.18	0.25	-0.29	-0.16	-0.08	0.49
perc.alumni	0.21	-0.25	-0.15	0.02	-0.22	-0.05
Expend	0.32	-0.13	0.23	0.08	0.08	-0.30
Grad.Rate	0.25	-0.17	-0.21	0.27	-0.11	0.22

Table 7. Transposed data-frame rounded off to 2 decimal places

Linear Equation of First Principal Component:

$$PC1 = (0.25 * \text{Apps}) + (0.21 * \text{Accept}) + (0.18 * \text{Enroll}) + (0.35 * \text{Top10perc}) + (0.34 * \text{Top25perc}) + (0.15 * \text{F.Undergrad}) + (0.03 * \text{P.Undergrad}) + (0.29 * \text{Outstate}) + (0.25 * \text{Room.Board}) + (0.06 * \text{Books}) + (-0.04 * \text{Personal}) + (0.32 * \text{PhD}) + (0.32 * \text{Terminal}) + (-0.18 * \text{S.F.Ratio}) + (0.21 * \text{perc.alumni}) + (0.32 * \text{Expend}) + (0.25 * \text{Grad.Rate})$$

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

```
Cumulative Variance Explained
[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.        ]
```

Fig.38 Cumulative Variance Explained for 17 PC's

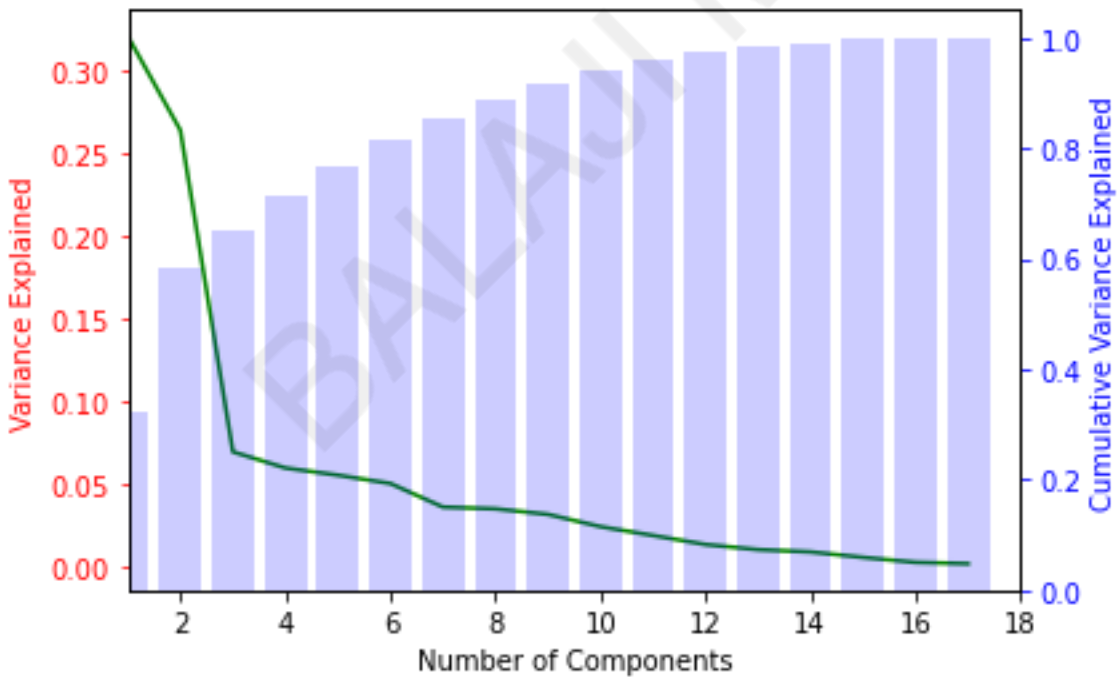


Fig.39 Variance & Cumulative Variance

Total variance of the PCs is equal to the total variance of the original attributes. Since the variances of the PCs are monotonically decreasing, it is possible to use first few PCs, so that a large proportion of total variance is explained by a significantly smaller number of PCs. The **cumulative values of the eigen values** help us to determine the number of principal components. The number of

PCs that are retained is subjective. Typically, the number is chosen so as to retain 70% - 90% of total variance.

The first six principal components account for roughly 81 percent of the overall variation, as can be seen from the cumulative variance explained.

Eigenvectors indicate the direction of the principal components, we can multiply the original data by the eigenvectors to re-orient our data onto the new axes. The eigenvector with the highest eigenvalue is therefore the principal component.

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]

Given a high dimensional dataset, it's a good idea to start with PCA in order to understand the main variance in the data and to understand its "real" dimensionality by plotting the explained variance vector and reducing the total number of dimensions without losing much information. PCA is particularly helpful and efficient only when there is high correlation between the features of the dataset.

18 characteristics from the dataset for our case study were provided to us for analysis. By using PCA, we were able to capture the majority of the information using only 6 principal components. These can also be used to represent more than 80% of the overall variance in the study. The principal component count is a subjective decision. As in our instance, we had the option of selecting the first 5 PCs, which would account for 76% of the overall variation, or the first 7 PCs, which would account for 85% of the whole variance.

However, since 6 primary components account for 81 percent of the overall variance, we have chosen to use them. In this scenario, we are reducing a high dimensional space to a lower dimensional space without losing much of the information and minimizing multicollinearity, which has implications for business.

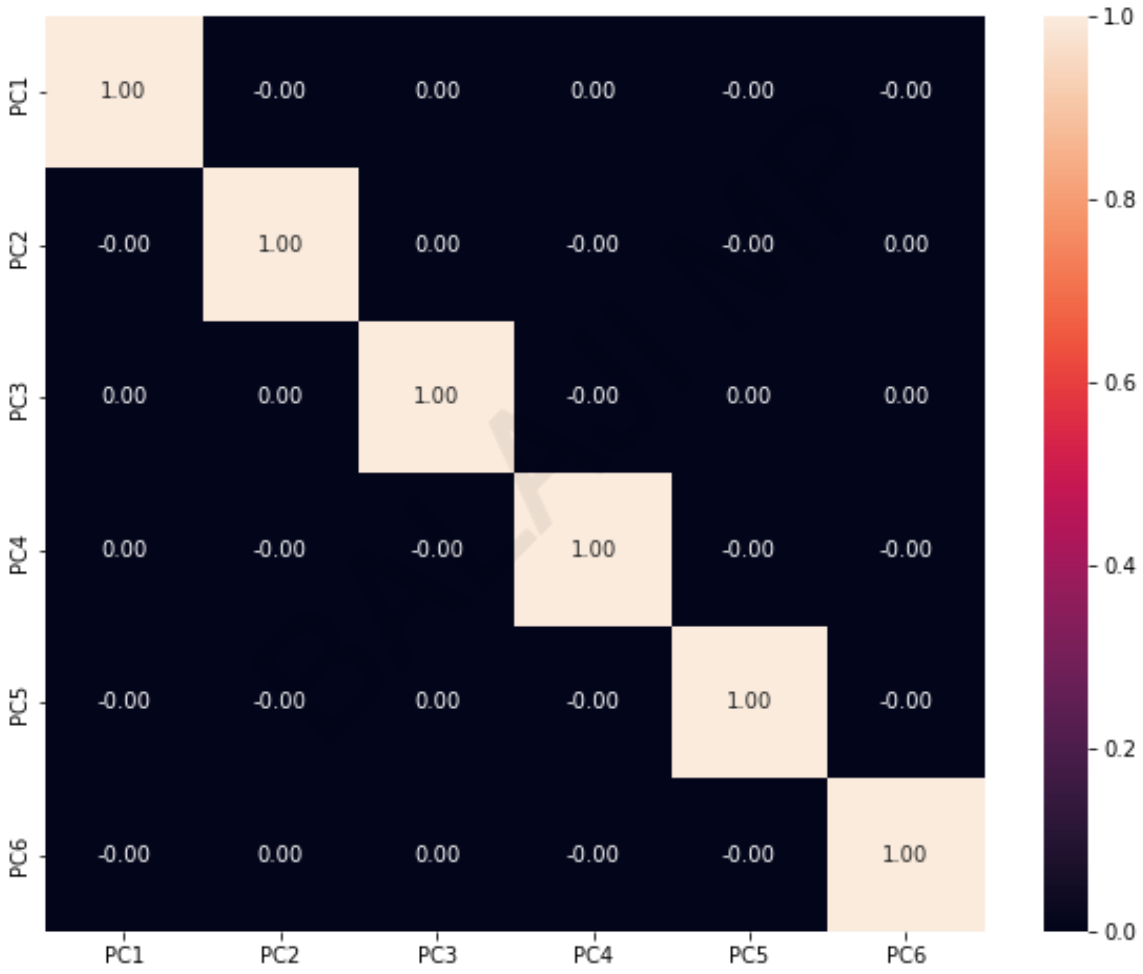


Fig.40 Correlation Heatmap of PC's

We can see that by using PCA, we were able to lessen the data's multicollinearity. Since they are orthogonal to one another, no two principal components are correlated.

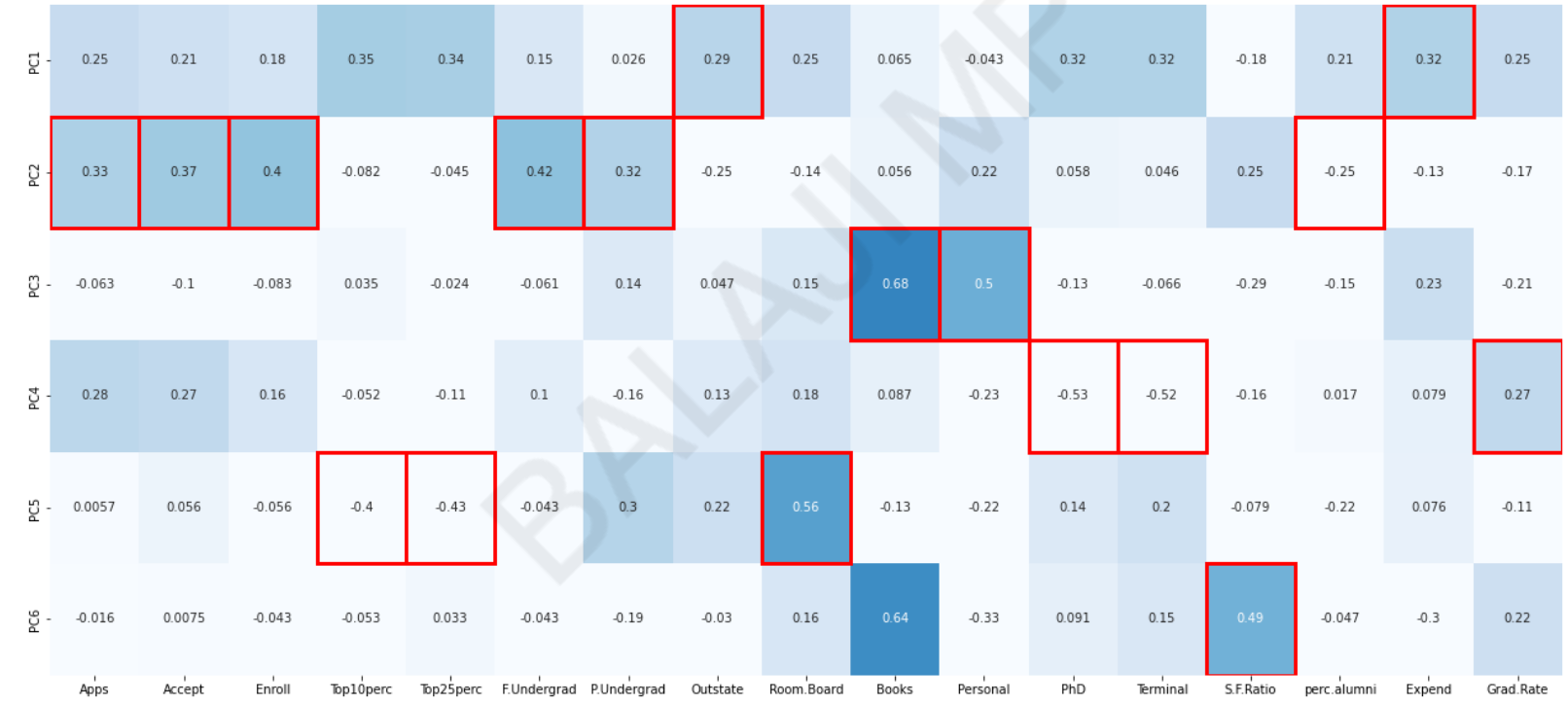


Fig.41 Heatmap

The highlighted cell in the heatmap above denotes the principal component in which the feature displays the greatest variance. This is helpful for classifying the features and conducting additional analysis on grouped features.

Insights from the above heatmap:

- **PC1**: Indicates the Number of students for whom the particular college or university is Out-of-state tuition and the Instructional expenditure per student.
- **PC2:** Apps, Accept, Enroll, F. Undergrad, P. Undergrad, and perc. alumni are examples of strongly correlated variables that are indicated.
- **PC3:** Highlights the projected cost of books for a student and their personal spending.
- **PC4:** Shows the proportion of professors who have earned a terminal degree and a Ph.D. The graduation rate is also shown.
- **PC5:** Covers the cost of rooms and board as well as the new students from the top 10 and 25 percent of the higher secondary class.
- **PC6:** Highlights the student to faculty ratio.

The information for a collection of characteristics is now displayed by all these primary components, as can be seen above. Therefore, with the proper identification of the principal components, additional research may be done to obtain the necessary data and provide valuable business insights.