AS Project Report

Deepti Yadav
April’22
Date : 29/05/2022

Table of Contents

List of Figures

List of Tables

Problem 1A : Salary

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

[Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.]

Table 1- Dataset Description

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

Table 2 - Dataset Information

```
<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
```

There are total 40 rows and 3 columns in the dataset. Out of 3, 2 columns are of object type and 1 is integer.

Table 3 - Missing values Check

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

From the above results we can see that there is no missing value present in the dataset.The data also does not have a duplicate value.

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

One way ANOVA For the Education:
H0_Edu : The mean salary is the same across all the three levels (High school graduate, Bachelor, and Doctorate).
Ha_Edu: The mean salary is different in at least one level.
$\alpha = 0.05$

One way ANOVA For the Occupation:
H0_Occ : The mean salary is the same across all the four levels (Administrative and clerical, Sales, Professional or specialty, and Executive or managerial)
Ha_Occ : The mean salary is different in at least level.
$\alpha = 0.05$

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

Table 4 – one-way ANOVA (Salary-Education)

	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

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

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

Table 5 – one-way ANOVA (Salary-Occupation)

	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

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

1.4 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)

To find out which class means are significantly different, the Tukey Honest Significant Difference test is performed.

Table 6 - Tukey HSD for variable 'Education'

Multiple Comparison of Means - Tukey HSD, FWER=0.05

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

For Category education the table above shows that since the p- values(p-adj in the table) are lesser than the significance level for all the three levels of education, this implies that the all classes mean are significantly different.

Table 7 - Tukey HSD for variable 'Occupation'

Multiple Comparison of Means - Tukey HSD, FWER=0.05

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

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

Problem 1b : Salary

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

When doing linear modeling or ANOVA it's useful to examine whether or not the effect of one variable depends on the level of one or more variables. If it does then we have what is called an "interaction".

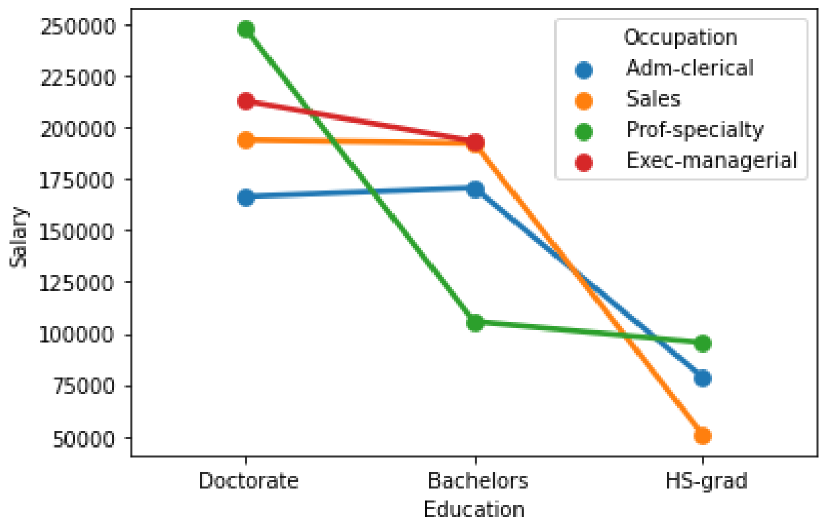


Figure 1 - Interaction Plot-1

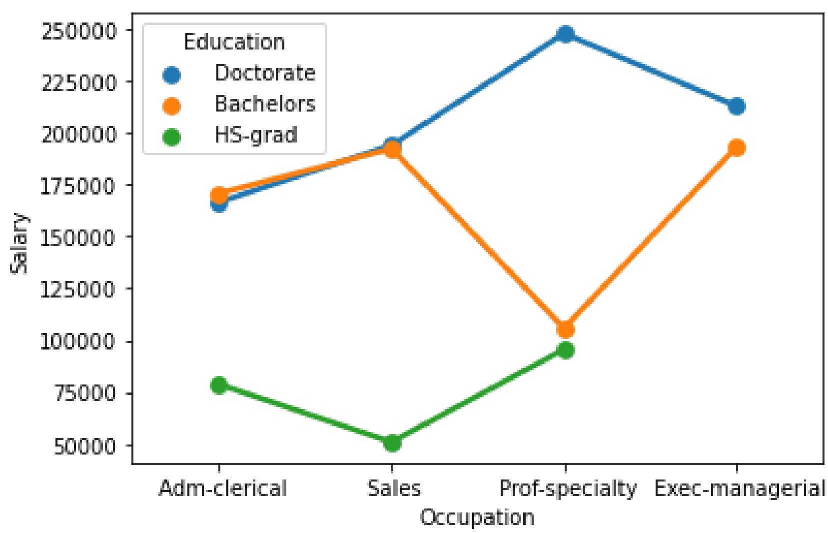


Figure 2 - Interaction Plot-2

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

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

- People with HS-grad education do not reach the position of Exec-managerial and they hold only Adm-clerk, Sales and Prof-Specialty occupations.
- People with education as Bachelors or Doctorate and occupation as Adm-clerical and Sales almost earn the same salaries (salaries ranging from 170000–190000).
- People with education as Bachelors and occupation as Prof-Specialty earn lesser than people with education as Bachelors and occupations as Adm-clerical and Sales.
- Prof-Specialty people with education as Doctorate earn maximum salaries and people with education as HS-Grad earn the minimum.
- Of all profession with HS-grad, Adm clerical people earn the lowest salaries
- People with education as Bachelors and occupation, Sales and Exec-Managerial earn almost the same salaries.
- Salespeople with Bachelors or Doctorate education earn the same salaries and earn higher than people with education as HS-grad.

1.6 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?

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

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

Table 8 – Two-way ANOVA

	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

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

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

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

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

Problem 2

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

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

EDA

Table 9 - Dataset Description

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

Table 10 - Dataset Information

```
<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
```

Table 11 - Missing values Check

```
Names           0
Apps            0
Accept          0
Enroll          0
Top10perc       0
Top25perc       0
F.Undergrad     0
P.Undergrad     0
Outstate        0
Room.Board      0
Books           0
Personal        0
PhD             0
Terminal        0
S.F.Ratio       0
perc.alumni     0
Expend          0
Grad.Rate       0
dtype: int64
```

Table 12 – Data set 5 point summary

	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

Observations:

1) The data set has 777 observations and 18 variables in the data set

2) Out of 18, only 1 column is object while rest are either integer or float.

3) There are no missing values in data set.

4) There are no duplicate rows present.

6) Total number of application received is 48094 for all universitities.

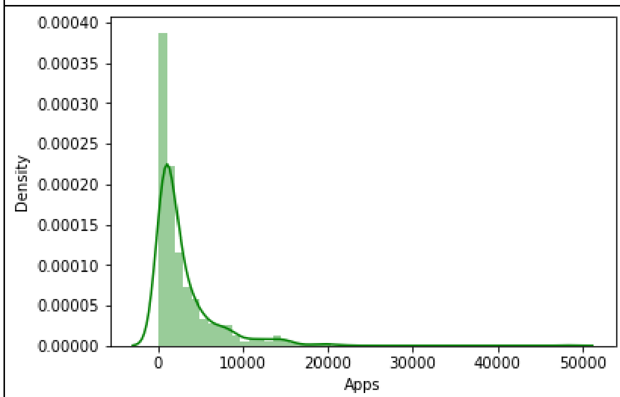
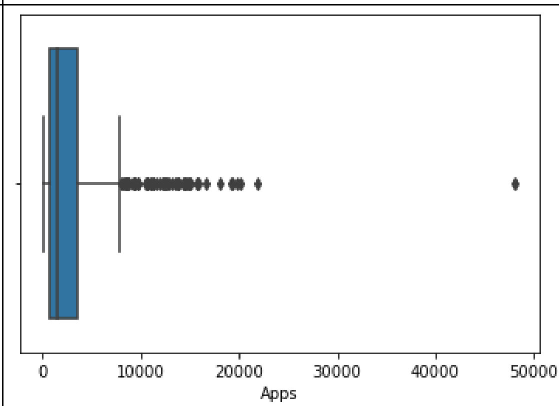
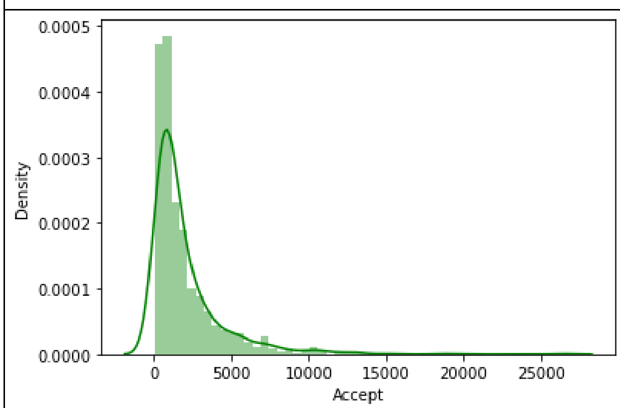
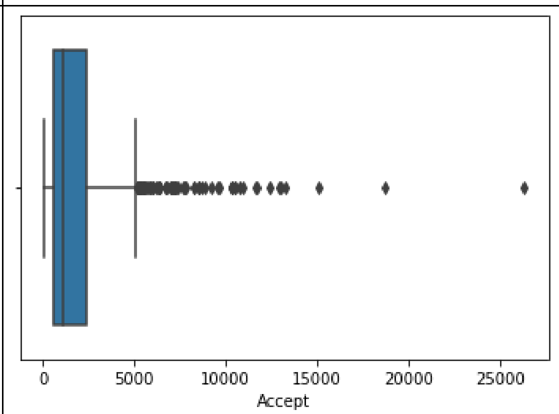
7) Number of applications accepted is 26330 for all universitities

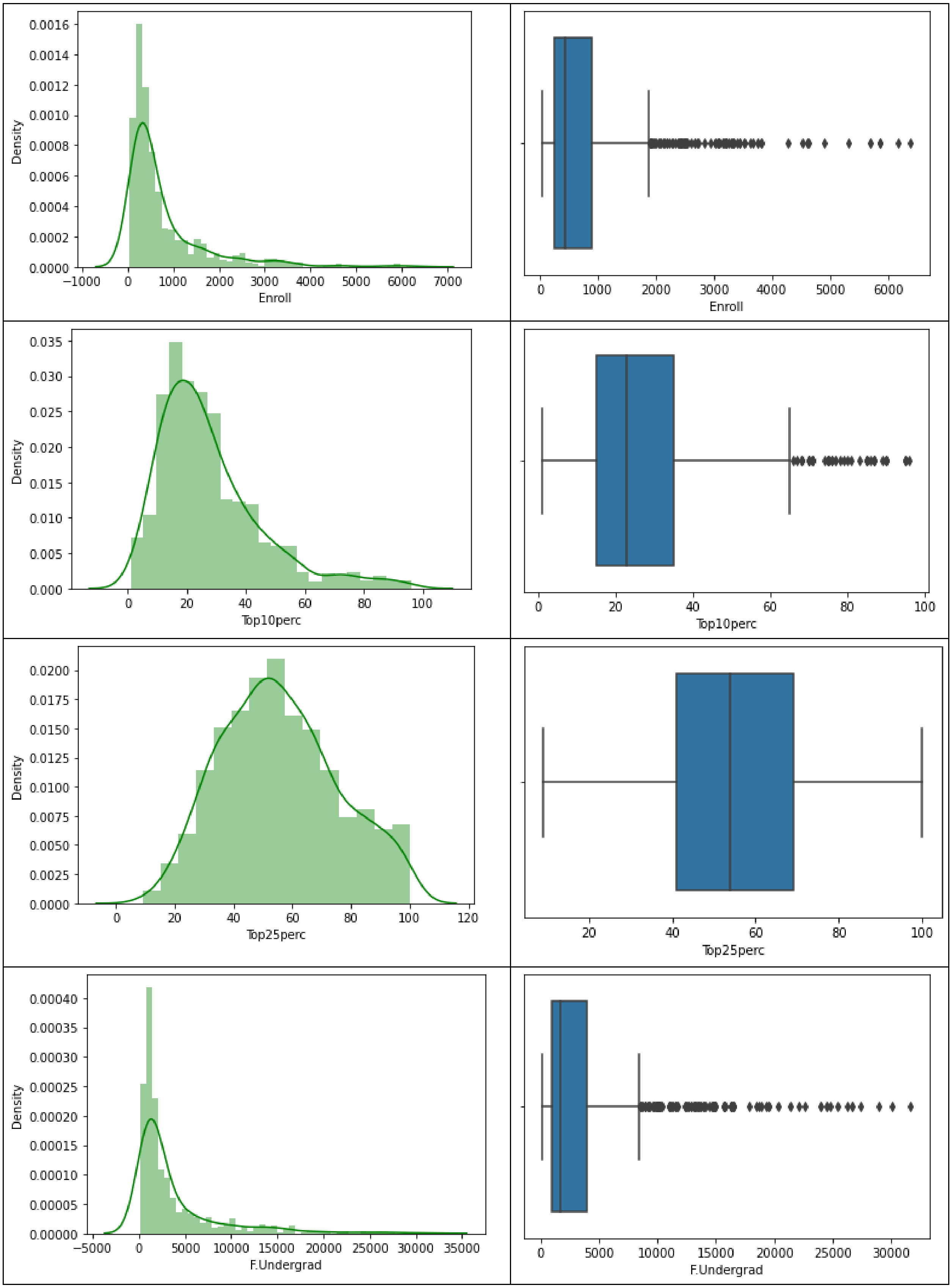
9) Maximum Graduation rate is 118% which looks like incorrect data.

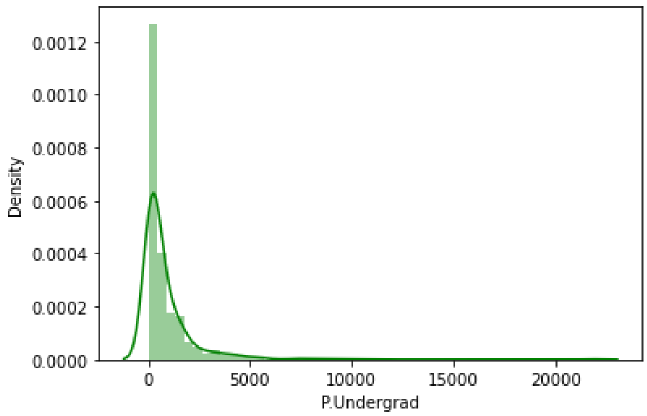
10) Maximum Percentage of faculties with Ph.D.’s is 103 which looks like incorrect data.

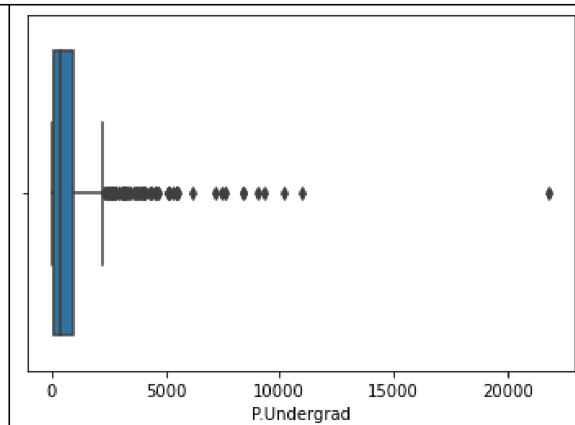
UNIVARIATE ANALYSIS

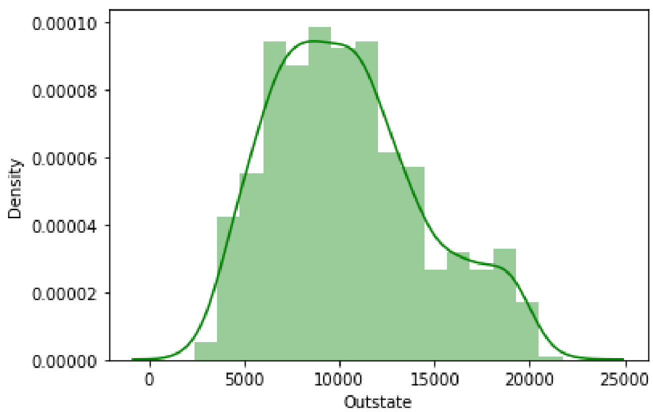
Table 13 – Univariate Analysis

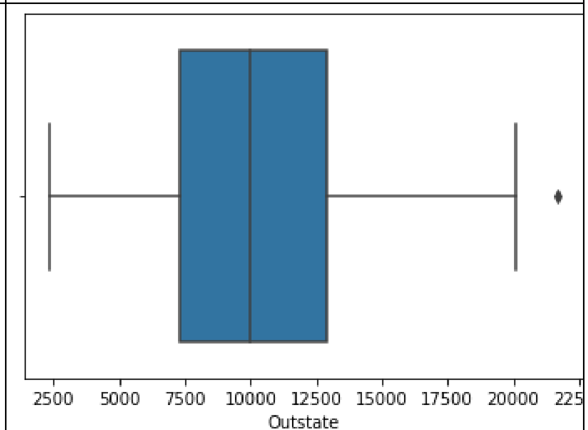
Histogram	Boxplot
	
	

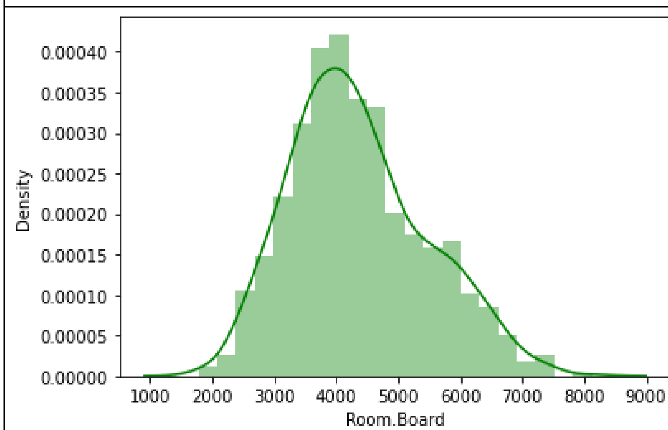

Density
Enroll
Enroll
Density
Top10perc
Top10perc
Density
Top25perc
Top25perc
Density
F.Undergrad
F.Undergrad

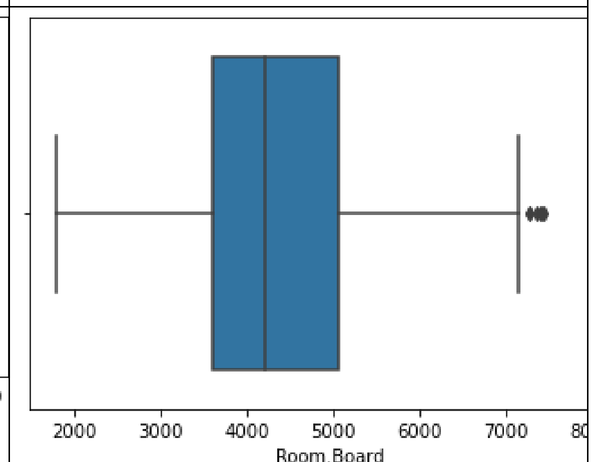

Density
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
5000
10000
15000
20000
P.Undergrad

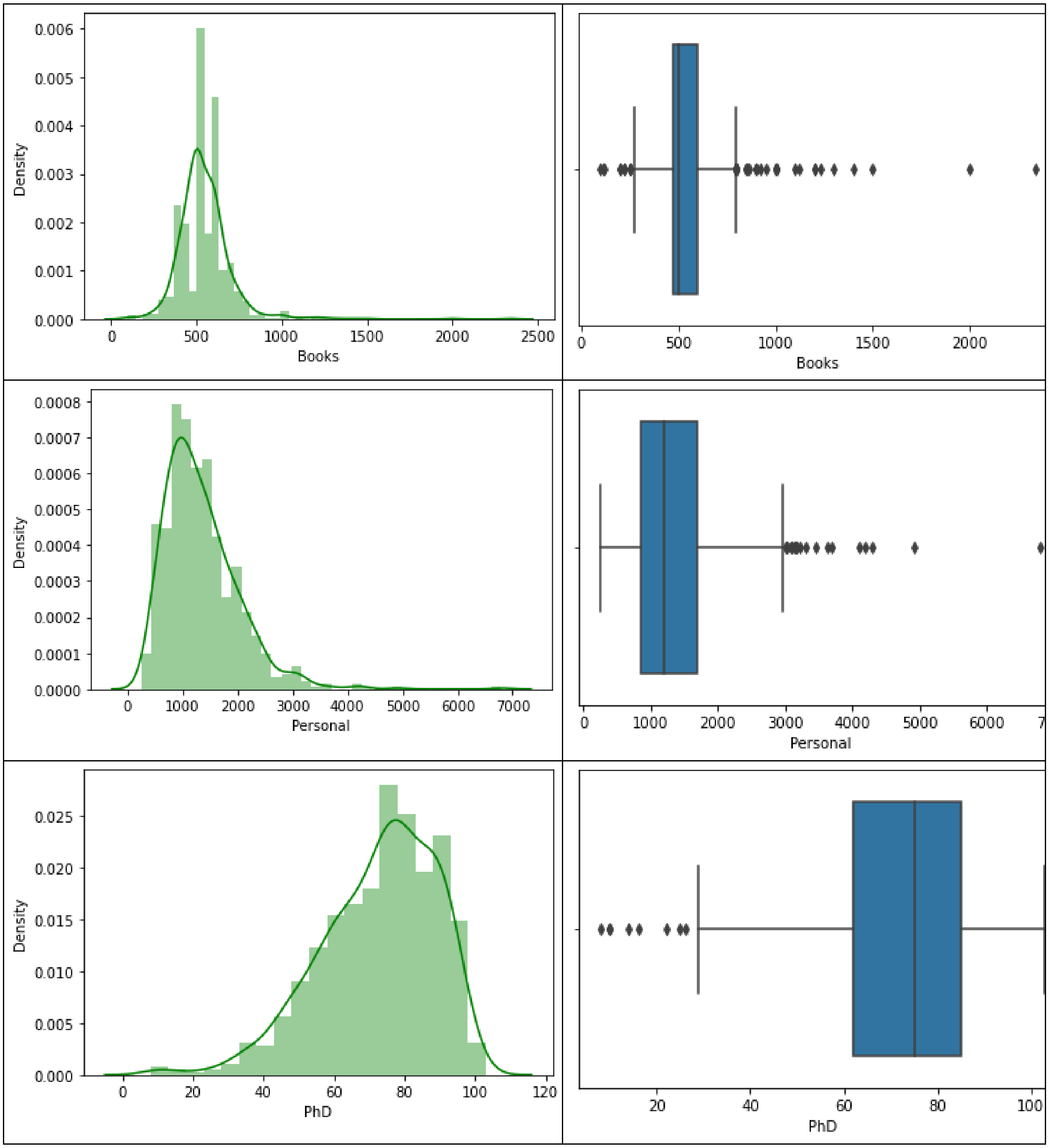

0
5000
10000
15000
20000
P.Undergrad

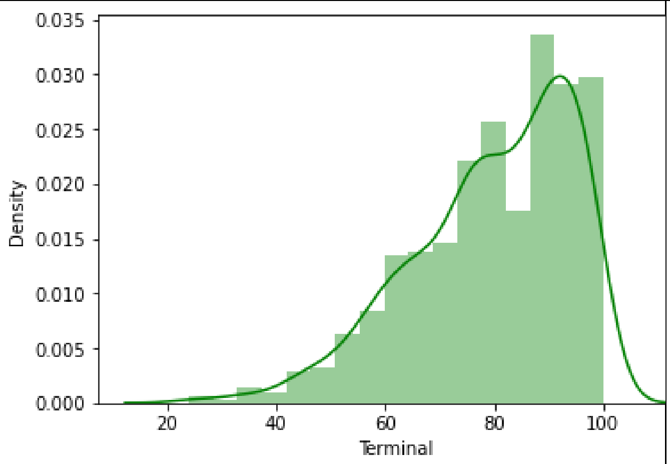

Density
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
5000
10000
15000
20000
25000
Outstate

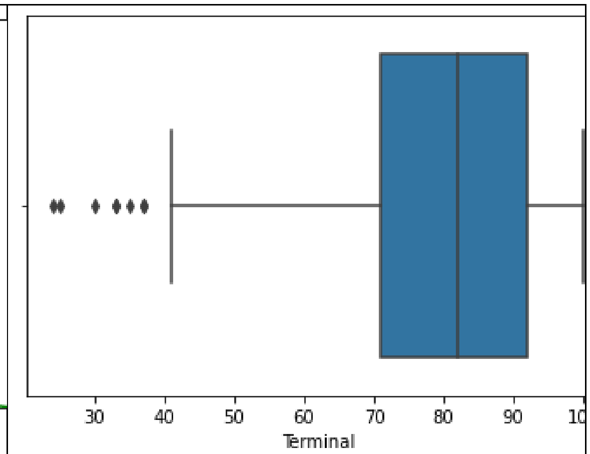

2500
5000
7500
10000
12500
15000
17500
20000
Outstate

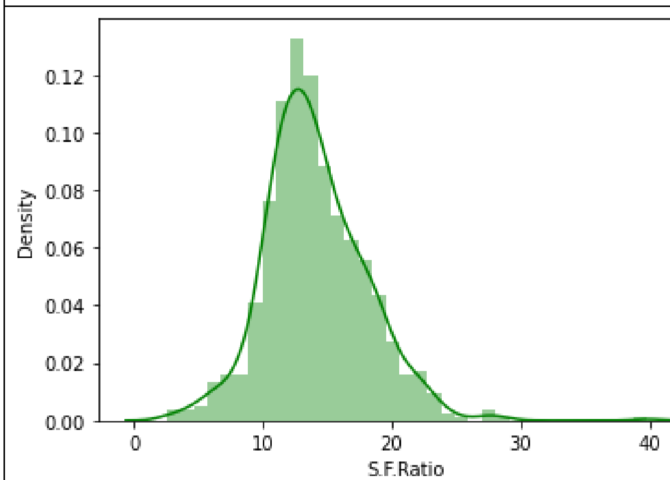

Density
0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
1000
2000
3000
4000
5000
6000
7000
8000
9000
Room.Board

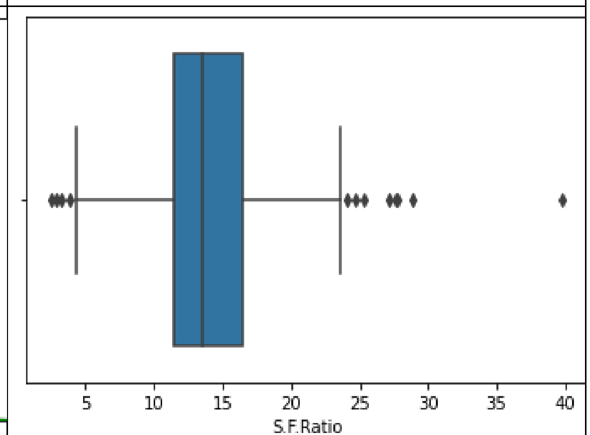

2000
3000
4000
5000
6000
7000
Room.Board

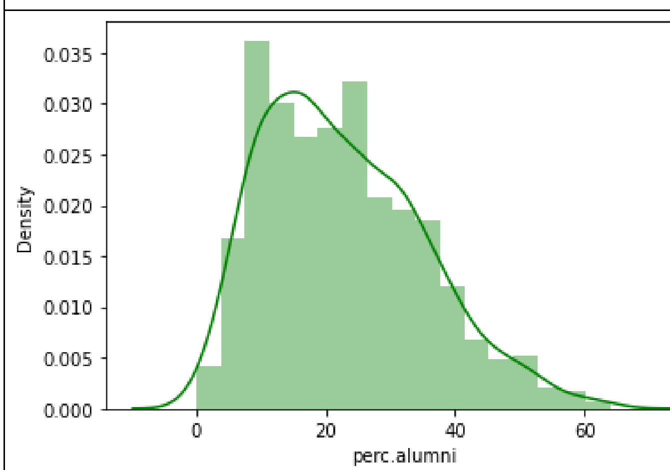

Density
Books
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
500
1000
1500
2000
2500
Books
0
500
1000
1500
2000
Density
Personal
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
1000
2000
3000
4000
5000
6000
7000
Personal
0
1000
2000
3000
4000
5000
6000
Density
PhD
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
120
PhD
20
40
60
80
100

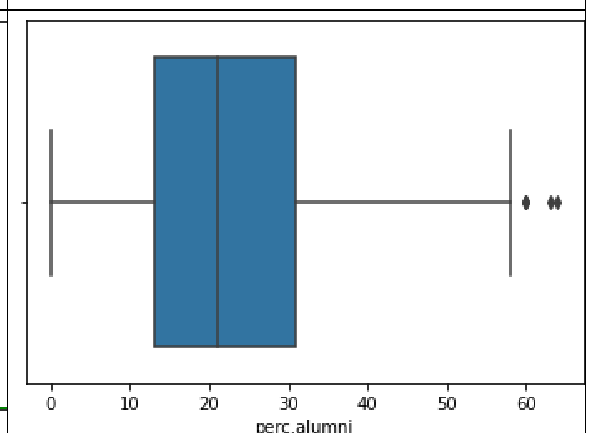

Density
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
20
40
60
80
100
Terminal

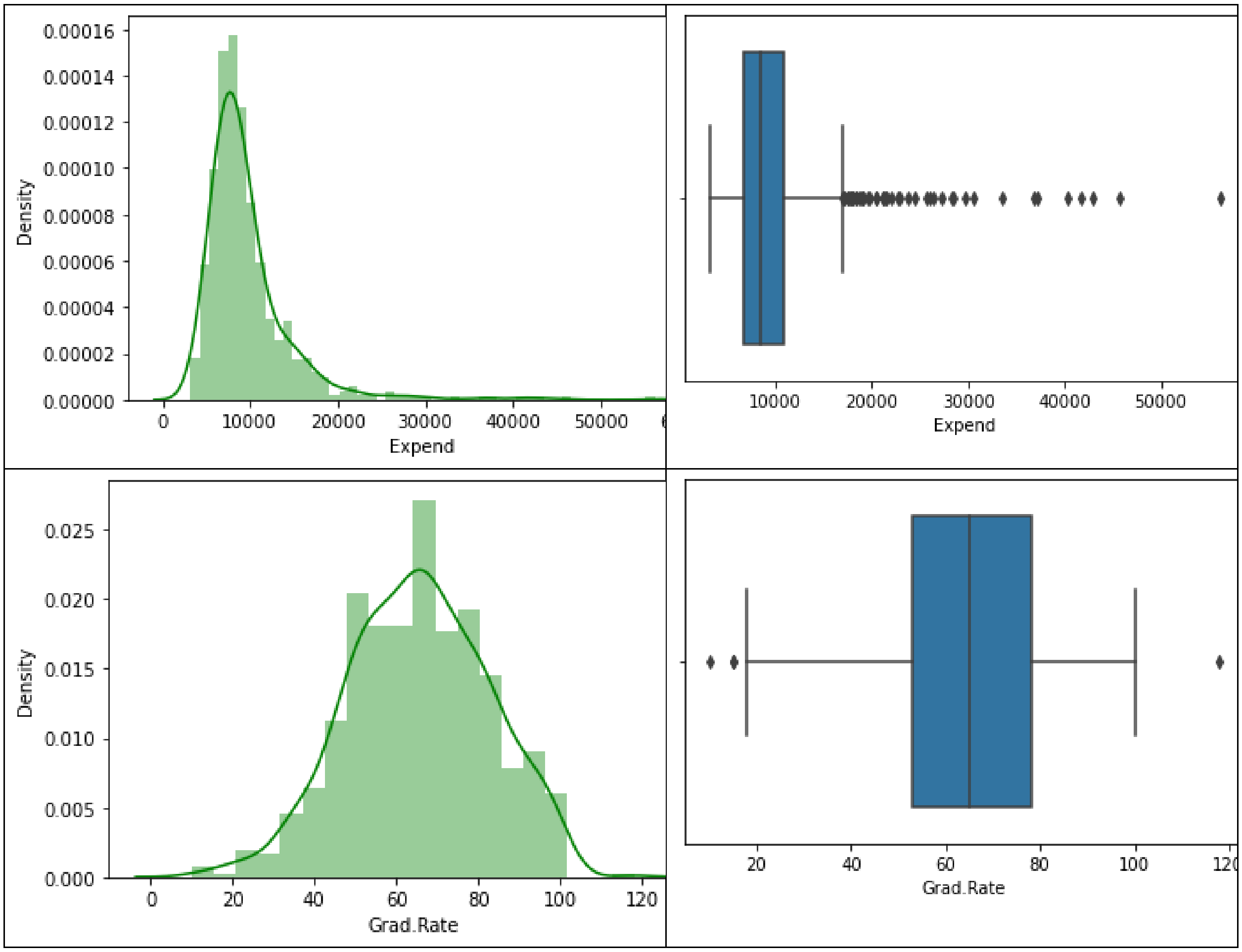

30
40
50
60
70
80
90
Terminal


Density
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
S.F.Ratio


5
10
15
20
25
30
35
40
S.F.Ratio


Density
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
perc.alumni


0
10
20
30
40
50
60
perc.alumni



Observations:

1) There are outliers present in all numeric variables except "Top25perc".

2) PhD and Terminal are highly left skewed.

3) Outliers should be treated.

MULTIIVARIATE ANALYSIS

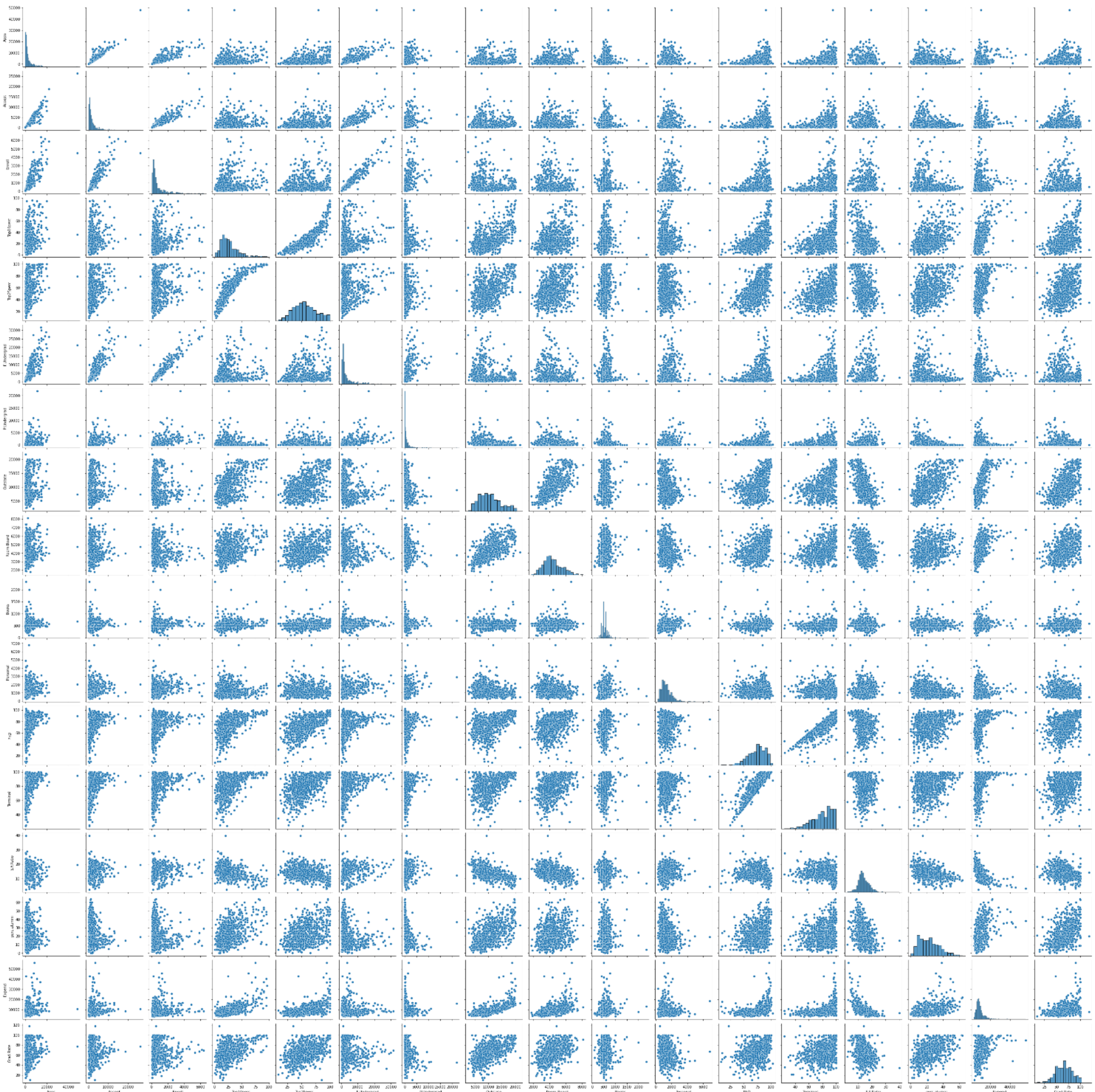


Figure 3 – Pair Plot

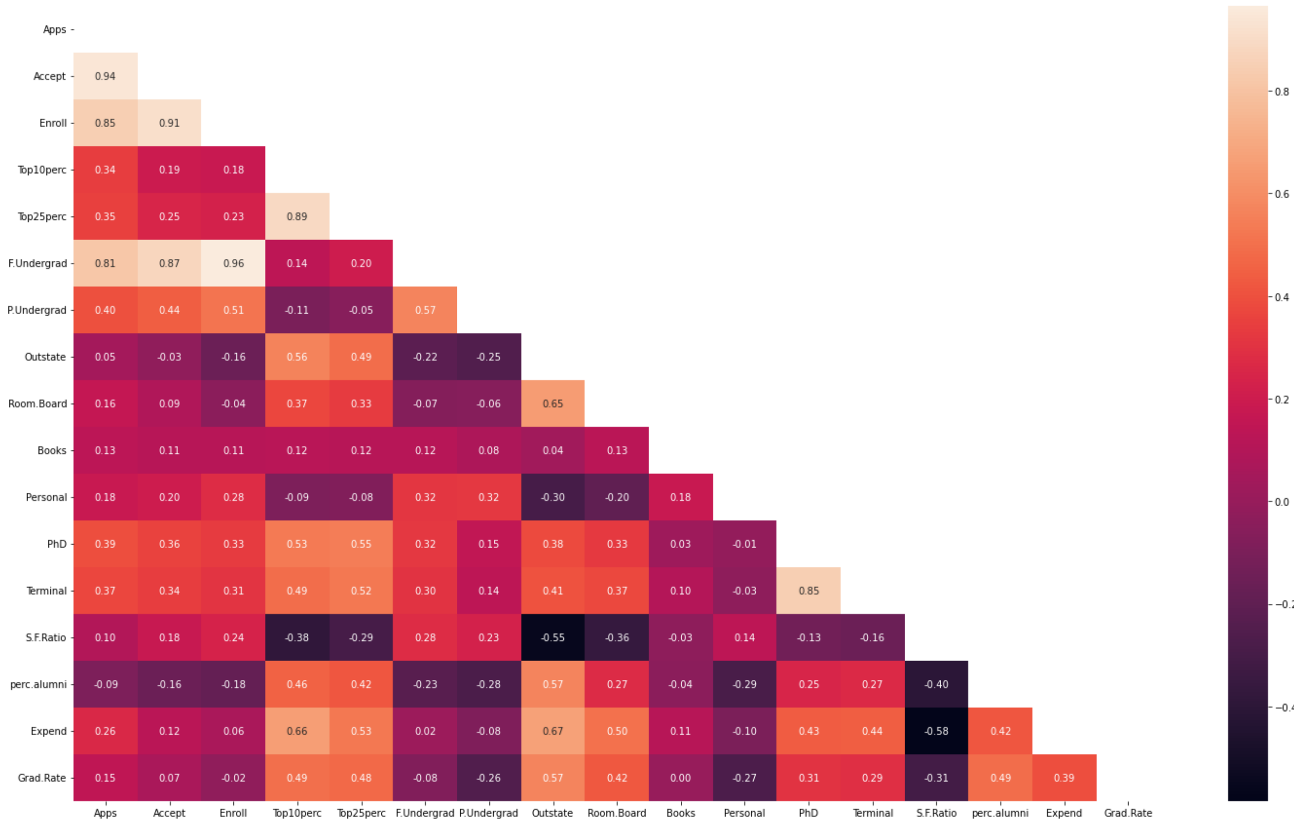


Figure 4 – Heat Map

Observations:

1) There are considerable number of features that are highly correlated. Correlation ranges from -1 to 1. Values closer to 0 shows no correaltion between varaiables.Closer to 1 is positively correlated

2) Enroll shows high correlation with 'Apps'&'Accept'.

3) F.Undergrad shows high correlation with 'Apps','Accept'& 'Enroll'.

4) Top25perc shows high correlation with Top10perc.

5) Terminal shows high correlation with PhD.

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

Often the variables of the data set are of different scales i.e. one variable is in millions and other in only 100. For e.g. in our data set Apps is having values in thousands and Grad.Rate in just two digits. Since the data in these variables are of different scales, it is tough to compare these variables.

Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

In this method, we convert variables with different scales of measurements into a single scale.

Scaling normalizes the data using the formula (x-mean)/standard deviation. Standard deviation becomes 1 and mean becomes zero.

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

Covariance Matrix

```
np.round(df_num_scaled.cov(),2)
```

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.00	0.94	0.85	0.34	0.35	0.82	0.40	0.05	0.17	0.13	0.18	0.39	0.37	0.10	-0.09	0.26	0.15
Accept	0.94	1.00	0.91	0.19	0.25	0.88	0.44	-0.03	0.09	0.11	0.20	0.36	0.34	0.18	-0.16	0.12	0.07
Enroll	0.85	0.91	1.00	0.18	0.23	0.97	0.51	-0.16	-0.04	0.11	0.28	0.33	0.31	0.24	-0.18	0.06	-0.02
Top10perc	0.34	0.19	0.18	1.00	0.89	0.14	-0.11	0.56	0.37	0.12	-0.09	0.53	0.49	-0.39	0.46	0.66	0.50
Top25perc	0.35	0.25	0.23	0.89	1.00	0.20	-0.05	0.49	0.33	0.12	-0.08	0.55	0.53	-0.30	0.42	0.53	0.48
F.Undergrad	0.82	0.88	0.97	0.14	0.20	1.00	0.57	-0.22	-0.07	0.12	0.32	0.32	0.30	0.28	-0.23	0.02	-0.08
P.Undergrad	0.40	0.44	0.51	-0.11	-0.05	0.57	1.00	-0.25	-0.06	0.08	0.32	0.15	0.14	0.23	-0.28	-0.08	-0.26
Outstate	0.05	-0.03	-0.16	0.56	0.49	-0.22	-0.25	1.00	0.66	0.04	-0.30	0.38	0.41	-0.56	0.57	0.67	0.57
Room.Board	0.17	0.09	-0.04	0.37	0.33	-0.07	-0.06	0.66	1.00	0.13	-0.20	0.33	0.38	-0.36	0.27	0.50	0.43
Books	0.13	0.11	0.11	0.12	0.12	0.12	0.08	0.04	0.13	1.00	0.18	0.03	0.10	-0.03	-0.04	0.11	0.00
Personal	0.18	0.20	0.28	-0.09	-0.08	0.32	0.32	-0.30	-0.20	0.18	1.00	-0.01	-0.03	0.14	-0.29	-0.10	-0.27
PhD	0.39	0.36	0.33	0.53	0.55	0.32	0.15	0.38	0.33	0.03	-0.01	1.00	0.85	-0.13	0.25	0.43	0.31
Terminal	0.37	0.34	0.31	0.49	0.53	0.30	0.14	0.41	0.38	0.10	-0.03	0.85	1.00	-0.16	0.27	0.44	0.29
S.F.Ratio	0.10	0.18	0.24	-0.39	-0.30	0.28	0.23	-0.56	-0.36	-0.03	0.14	-0.13	-0.16	1.00	-0.40	-0.58	-0.31
perc.alumni	-0.09	-0.16	-0.18	0.46	0.42	-0.23	-0.28	0.57	0.27	-0.04	-0.29	0.25	0.27	-0.40	1.00	0.42	0.49
Expend	0.26	0.12	0.06	0.66	0.53	0.02	-0.08	0.67	0.50	0.11	-0.10	0.43	0.44	-0.58	0.42	1.00	0.39
Grad.Rate	0.15	0.07	-0.02	0.50	0.48	-0.08	-0.26	0.57	0.43	0.00	-0.27	0.31	0.29	-0.31	0.49	0.39	1.00

```
np.round(df_num_scaled.corr(),2)
```

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.00	0.94	0.85	0.34	0.35	0.81	0.40	0.05	0.16	0.13	0.18	0.39	0.37	0.10	-0.09	0.26	0.15
Accept	0.94	1.00	0.91	0.19	0.25	0.87	0.44	-0.03	0.09	0.11	0.20	0.36	0.34	0.18	-0.16	0.12	0.07
Enroll	0.85	0.91	1.00	0.18	0.23	0.96	0.51	-0.16	-0.04	0.11	0.28	0.33	0.31	0.24	-0.18	0.06	-0.02
Top10perc	0.34	0.19	0.18	1.00	0.89	0.14	-0.11	0.56	0.37	0.12	-0.09	0.53	0.49	-0.38	0.46	0.66	0.49
Top25perc	0.35	0.25	0.23	0.89	1.00	0.20	-0.05	0.49	0.33	0.12	-0.08	0.55	0.52	-0.29	0.42	0.53	0.48
F.Undergrad	0.81	0.87	0.96	0.14	0.20	1.00	0.57	-0.22	-0.07	0.12	0.32	0.32	0.30	0.28	-0.23	0.02	-0.08
P.Undergrad	0.40	0.44	0.51	-0.11	-0.05	0.57	1.00	-0.25	-0.06	0.08	0.32	0.15	0.14	0.23	-0.28	-0.08	-0.26
Outstate	0.05	-0.03	-0.16	0.56	0.49	-0.22	-0.25	1.00	0.65	0.04	-0.30	0.38	0.41	-0.55	0.57	0.67	0.57
Room.Board	0.16	0.09	-0.04	0.37	0.33	-0.07	-0.06	0.65	1.00	0.13	-0.20	0.33	0.37	-0.36	0.27	0.50	0.42
Books	0.13	0.11	0.11	0.12	0.12	0.12	0.08	0.04	0.13	1.00	0.18	0.03	0.10	-0.03	-0.04	0.11	0.00
Personal	0.18	0.20	0.28	-0.09	-0.08	0.32	0.32	-0.30	-0.20	0.18	1.00	-0.01	-0.03	0.14	-0.29	-0.10	-0.27
PhD	0.39	0.36	0.33	0.53	0.55	0.32	0.15	0.38	0.33	0.03	-0.01	1.00	0.85	-0.13	0.25	0.43	0.31
Terminal	0.37	0.34	0.31	0.49	0.52	0.30	0.14	0.41	0.37	0.10	-0.03	0.85	1.00	-0.16	0.27	0.44	0.29
S.F.Ratio	0.10	0.18	0.24	-0.38	-0.29	0.28	0.23	-0.55	-0.36	-0.03	0.14	-0.13	-0.16	1.00	-0.40	-0.58	-0.31
perc.alumni	-0.09	-0.16	-0.18	0.46	0.42	-0.23	-0.28	0.57	0.27	-0.04	-0.29	0.25	0.27	-0.40	1.00	0.42	0.49
Expend	0.26	0.12	0.06	0.66	0.53	0.02	-0.08	0.67	0.50	0.11	-0.10	0.43	0.44	-0.58	0.42	1.00	0.39
Grad.Rate	0.15	0.07	-0.02	0.49	0.48	-0.08	-0.26	0.57	0.42	0.00	-0.27	0.31	0.29	-0.31	0.49	0.39	1.00

Covariance matrix signifies the direction of the linear relationship between the two variables. By direction we mean if the variables are directly proportional or inversely proportional to each other. The values of covariance can be any number between the two opposite infinities.

Correlation matrix not only shows the kind of relation (in terms of direction) but also how strong the relationship is. Thus, we can say the correlation values have standardized notions, whereas the covariance values are not standardized and cannot be used to compare how strong or weak the relationship is because the magnitude has no direct significance. It can assume values from -1 to +1.

On scaled data, Covariance matrix is same as the correlation matrix of variables.

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

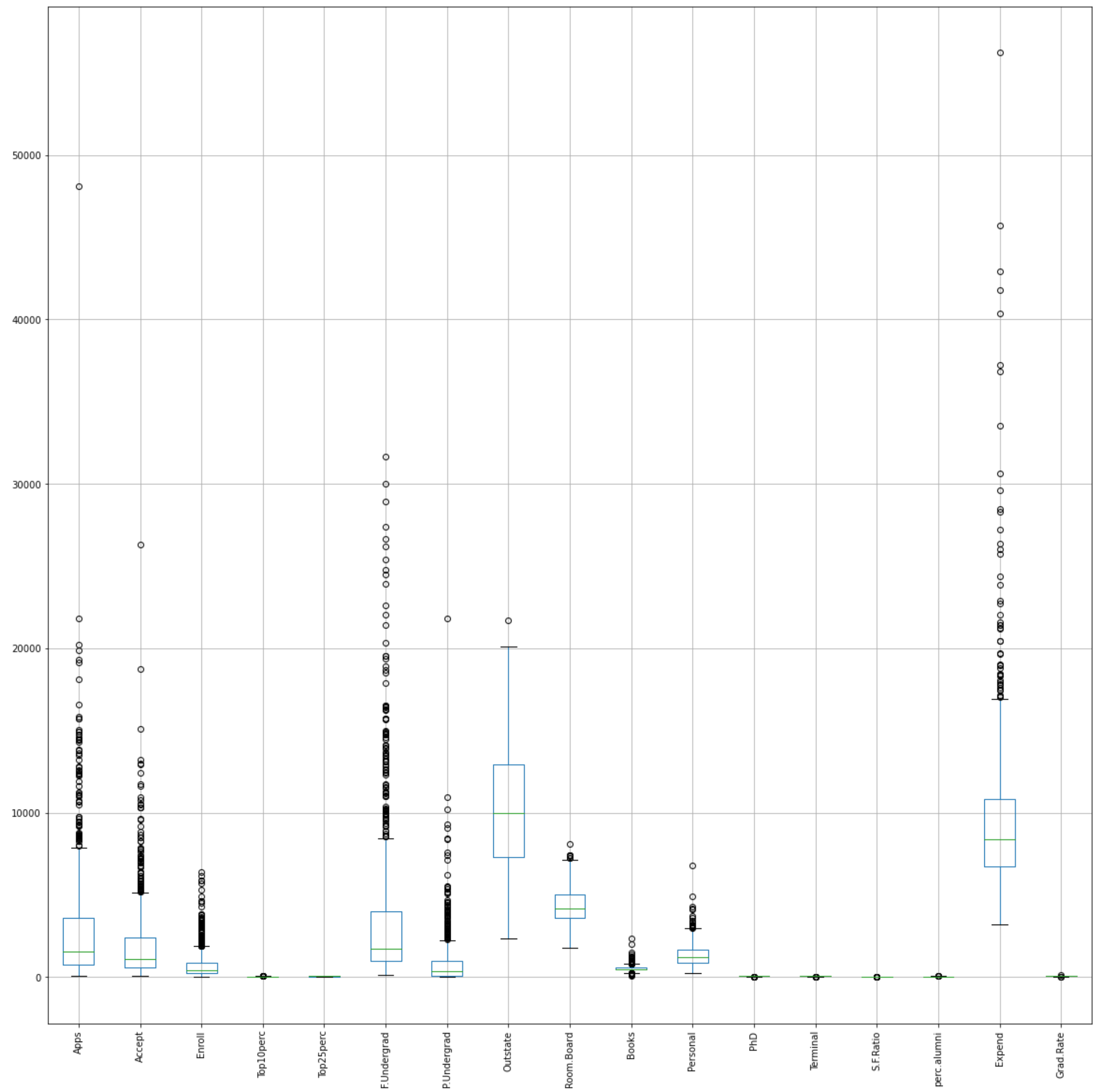


Figure 5 – Boxplot before scaling

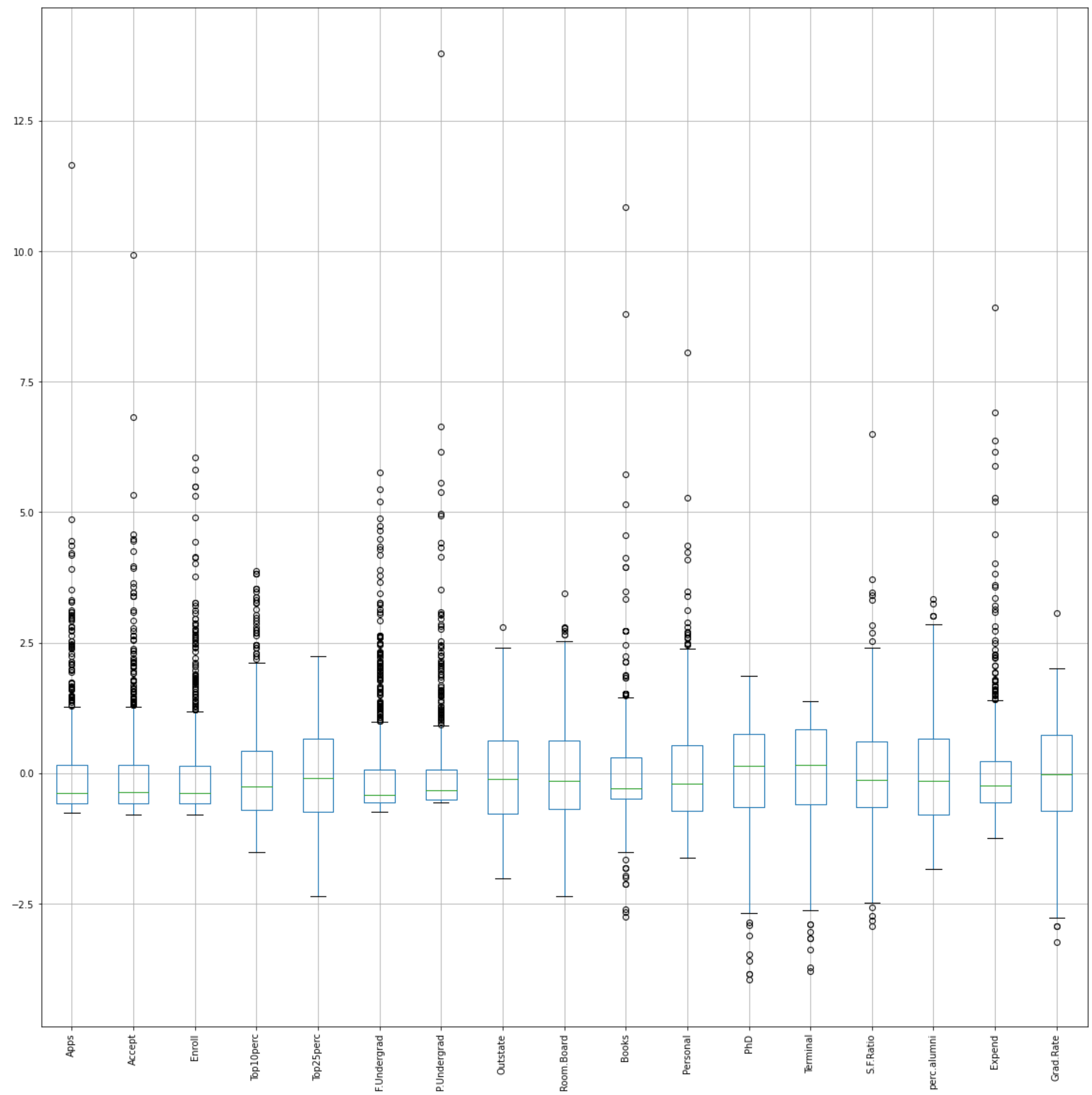


Figure 6 – Boxplot after scaling

Scaling ensures that attribute means are all 0 and variances 1. Medians are also close to each other.

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

Eigen values for all numeric variables

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

Eigen vectors for all numeric variables

```
[[-2.48765602e-01  3.31598227e-01  6.30921033e-02 -2.81310530e-01
   5.74140964e-03  1.62374420e-02  4.24863486e-02  1.03090398e-01
   9.02270802e-02 -5.25098025e-02  3.58970400e-01 -4.59139498e-01
   4.30462074e-02 -1.33405806e-01  8.06328039e-02 -5.95830975e-01
   2.40709086e-02]
 [-2.07601502e-01  3.72116750e-01  1.01249056e-01 -2.67817346e-01
   5.57860920e-02 -7.53468452e-03  1.29497196e-02  5.62709623e-02
   1.77864814e-01 -4.11400844e-02 -5.43427250e-01  5.18568789e-01
  -5.84055850e-02  1.45497511e-01  3.34674281e-02 -2.92642398e-01
  -1.45102446e-01]
 [-1.76303592e-01  4.03724252e-01  8.29855709e-02 -1.61826771e-01
  -5.56936353e-02  4.25579803e-02  2.76928937e-02 -5.86623552e-02
   1.28560713e-01 -3.44879147e-02  6.09651110e-01  4.04318439e-01
  -6.93988831e-02 -2.95896092e-02 -8.56967180e-02  4.44638207e-01
   1.11431545e-02]
 [-3.54273947e-01 -8.24118211e-02 -3.50555339e-02  5.15472524e-02
  -3.95434345e-01  5.26927980e-02  1.61332069e-01  1.22678028e-01
  -3.41099863e-01 -6.40257785e-02 -1.44986329e-01  1.48738723e-01
  -8.10481404e-03 -6.97722522e-01 -1.07828189e-01 -1.02303616e-03
   3.85543001e-02]
 [-3.44001279e-01 -4.47786551e-02  2.41479376e-02  1.09766541e-01
  -4.26533594e-01 -3.30915896e-02  1.18485556e-01  1.02491967e-01
  -4.03711989e-01 -1.45492289e-02  8.03478445e-02 -5.18683400e-02
  -2.73128469e-01  6.17274818e-01  1.51742110e-01 -2.18838802e-02
  -8.93515563e-02]
 [-1.54640962e-01  4.17673774e-01  6.13929764e-02 -1.00412335e-01
  -4.34543659e-02  4.34542349e-02  2.50763629e-02 -7.88896442e-02
   5.94419181e-02 -2.08471834e-02 -4.14705279e-01 -5.60363054e-01
  -8.11578181e-02 -9.91640992e-03 -5.63728817e-02  5.23622267e-01
   5.61767721e-02]
 [-2.64425045e-02  3.15087830e-01 -1.39681716e-01  1.58558487e-01
   3.02385408e-01  1.91198583e-01 -6.10423460e-02 -5.70783816e-01
  -5.60672902e-01  2.23105808e-01  9.01788964e-03  5.27313042e-02
   1.00693324e-01 -2.09515982e-02  1.92857500e-02 -1.25997650e-01
  -6.35360730e-02]
 [-2.94736419e-01 -2.49643522e-01 -4.65988731e-02 -1.31291364e-01
   2.22532003e-01  3.00003910e-02 -1.08528966e-01 -9.84599754e-03
   4.57332880e-03 -1.86675363e-01  5.08995918e-02 -1.01594830e-01
   1.43220673e-01 -3.83544794e-02 -3.40115407e-02  1.41856014e-01
  -8.23443779e-01]
```

```
[-2.49030449e-01 -1.37808883e-01 -1.48967389e-01 -1.84995991e-01
  5.60919470e-01 -1.62755446e-01 -2.09744235e-01  2.21453442e-01
 -2.75022548e-01 -2.98324237e-01  1.14639620e-03  2.59293381e-02
 -3.59321731e-01 -3.40197083e-03 -5.84289756e-02  6.97485854e-02
  3.54559731e-01]
[-6.47575181e-02  5.63418434e-02 -6.77411649e-01 -8.70892205e-02
 -1.27288825e-01 -6.41054950e-01  1.49692034e-01 -2.13293009e-01
  1.33663353e-01  8.20292186e-02  7.72631963e-04 -2.88282896e-03
  3.19400370e-02  9.43887925e-03 -6.68494643e-02 -1.14379958e-02
 -2.81593679e-02]
[ 4.25285386e-02  2.19929218e-01 -4.99721120e-01  2.30710568e-01
 -2.22311021e-01  3.31398003e-01 -6.33790064e-01  2.32660840e-01
  9.44688900e-02 -1.36027616e-01 -1.11433396e-03  1.28904022e-02
 -1.85784733e-02  3.09001353e-03  2.75286207e-02 -3.94547417e-02
 -3.92640266e-02]
[-3.18312875e-01  5.83113174e-02  1.27028371e-01  5.34724832e-01
  1.40166326e-01 -9.12555212e-02  1.09641298e-03  7.70400002e-02
  1.85181525e-01  1.23452200e-01  1.38133366e-02 -2.98075465e-02
  4.03723253e-02  1.12055599e-01 -6.91126145e-01 -1.27696382e-01
  2.32224316e-02]
[-3.17056016e-01  4.64294477e-02  6.60375454e-02  5.19443019e-01
  2.04719730e-01 -1.54927646e-01  2.84770105e-02  1.21613297e-02
  2.54938198e-01  8.85784627e-02  6.20932749e-03  2.70759809e-02
 -5.89734026e-02 -1.58909651e-01  6.71008607e-01  5.83134662e-02
  1.64850420e-02]
[ 1.76957895e-01  2.46665277e-01  2.89848401e-01  1.61189487e-01
 -7.93882496e-02 -4.87045875e-01 -2.19259358e-01  8.36048735e-02
 -2.74544380e-01 -4.72045249e-01 -2.22215182e-03  2.12476294e-02
  4.45000727e-01  2.08991284e-02  4.13740967e-02  1.77152700e-02
 -1.10262122e-02]
[-2.05082369e-01 -2.46595274e-01  1.46989274e-01 -1.73142230e-02
 -2.16297411e-01  4.73400144e-02 -2.43321156e-01 -6.78523654e-01
  2.55334907e-01 -4.22999706e-01 -1.91869743e-02 -3.33406243e-03
 -1.30727978e-01  8.41789410e-03 -2.71542091e-02 -1.04088088e-01
  1.82660654e-01]
[-3.18908750e-01 -1.31689865e-01 -2.26743985e-01 -7.92734946e-02
  7.59581203e-02  2.98118619e-01  2.26584481e-01  5.41593771e-02
  4.91388809e-02 -1.32286331e-01 -3.53098218e-02  4.38803230e-02
  6.92088870e-01  2.27742017e-01  7.31225166e-02  9.37464497e-02
  3.25982295e-01]
[-2.52315654e-01 -1.69240532e-01  2.08064649e-01 -2.69129066e-01
 -1.09267913e-01 -2.16163313e-01 -5.59943937e-01  5.33553891e-03
 -4.19043052e-02  5.90271067e-01 -1.30710024e-02  5.00844705e-03
  2.19839000e-01  3.39433604e-03  3.64767385e-02  6.91969778e-02
  1.22106697e-01]]
```

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

Bartletts Test of Sphericity

Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.

H0: All variables in the data are uncorrelated

Ha: At least one pair of variables in the data are correlated

If the null hypothesis cannot be rejected, then PCA is not advisable.

If the p-value is small, then we can reject the null hypothesis and agree that there is atleast one pair of vairbales in the data wihich are correlated hence PCA is recommended.

P_value = 0

Conclusion: Reject null which means At least one pair of variables in the data are correlated

KMO Test

The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.

Generally, if MSA is less than 0.5, PCA is not recommended, since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

kmo_model = 0.81

Conclusion: PCA is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	-0.042486	-0.103090	-0.090227	0.052510	0.043046	0.024071	0.595831	0.080633	0.133406	0.459139	0.358970
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	-0.012950	-0.056271	-0.177865	0.041140	-0.058406	-0.145102	0.292642	0.033467	-0.145498	-0.518569	-0.543427
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	-0.027693	0.058662	-0.128561	0.034488	-0.069399	0.011143	-0.444638	-0.085697	0.029590	-0.404318	0.609651
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	-0.161332	-0.122678	0.341100	0.064026	-0.008105	0.038554	0.001023	-0.107828	0.697723	-0.148739	-0.144986
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	-0.118486	-0.102492	0.403712	0.014549	-0.273128	-0.089352	0.021884	0.151742	-0.617275	0.051868	0.080348
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454	-0.025076	0.078890	-0.059442	0.020847	-0.081158	0.056177	-0.523622	-0.056373	0.009916	0.560363	-0.414705
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199	0.061042	0.570784	0.560673	-0.223106	0.100693	-0.063536	0.125998	0.019286	0.020952	-0.052731	0.009018
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000	0.108529	0.009846	-0.004573	0.186675	0.143221	-0.823444	-0.141856	-0.034012	0.038354	0.101595	0.050900
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755	0.209744	-0.221453	0.275023	0.298324	-0.359322	0.354560	-0.069749	-0.058429	0.003402	-0.025929	0.001146
Books	0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055	-0.149692	0.213293	-0.133663	-0.082029	0.031940	-0.028159	0.011438	-0.066849	-0.009439	0.002883	0.000773
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398	0.633790	-0.232661	-0.094469	0.136028	-0.018578	-0.039264	0.039455	0.027529	-0.003090	-0.012890	-0.001114
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256	-0.001096	-0.077040	-0.185182	-0.123452	0.040372	0.023222	0.127696	-0.691126	-0.112056	0.029808	0.013813
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928	-0.028477	-0.012161	-0.254938	-0.088578	-0.058973	0.016485	-0.058313	0.671009	0.158910	-0.027076	0.006209
S.F.Ratio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046	0.219259	-0.083605	0.274544	0.472045	0.445001	-0.011026	-0.017715	0.041374	-0.020899	-0.021248	-0.002222
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340	0.243321	0.678524	-0.255335	0.423000	-0.130728	0.182661	0.104088	-0.027154	-0.008418	0.003334	-0.019187
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119	-0.226584	-0.054159	-0.049139	0.132286	0.692089	0.325982	-0.093746	0.073123	-0.227742	-0.043880	-0.035310
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163	0.559944	-0.005336	0.041904	-0.590271	0.219839	0.122107	-0.069197	0.036477	-0.003394	-0.005008	-0.013071

Figure 7 – All Principal components with original features

	PC1	PC2	PC3	PC4	PC5	PC6
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755
Books	0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928
S.F.Ratio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163

Figure 8 – Principal components with original features after concluding PCA

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 eq of 1st component:
0.25 * Apps + 0.21 * Accept + 0.18 * Enroll + 0.35 * Top10perc + 0.34 * Top25perc + 0.15 * F.Undergrad + 0.03 * P.Und
ergrad + 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 +
```

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?

```
np.cumsum(pca.explained_variance_ratio_)
```

```
array([0.32020628, 0.58360843, 0.65261759, 0.71184748, 0.76673154,
       0.81657854])
```

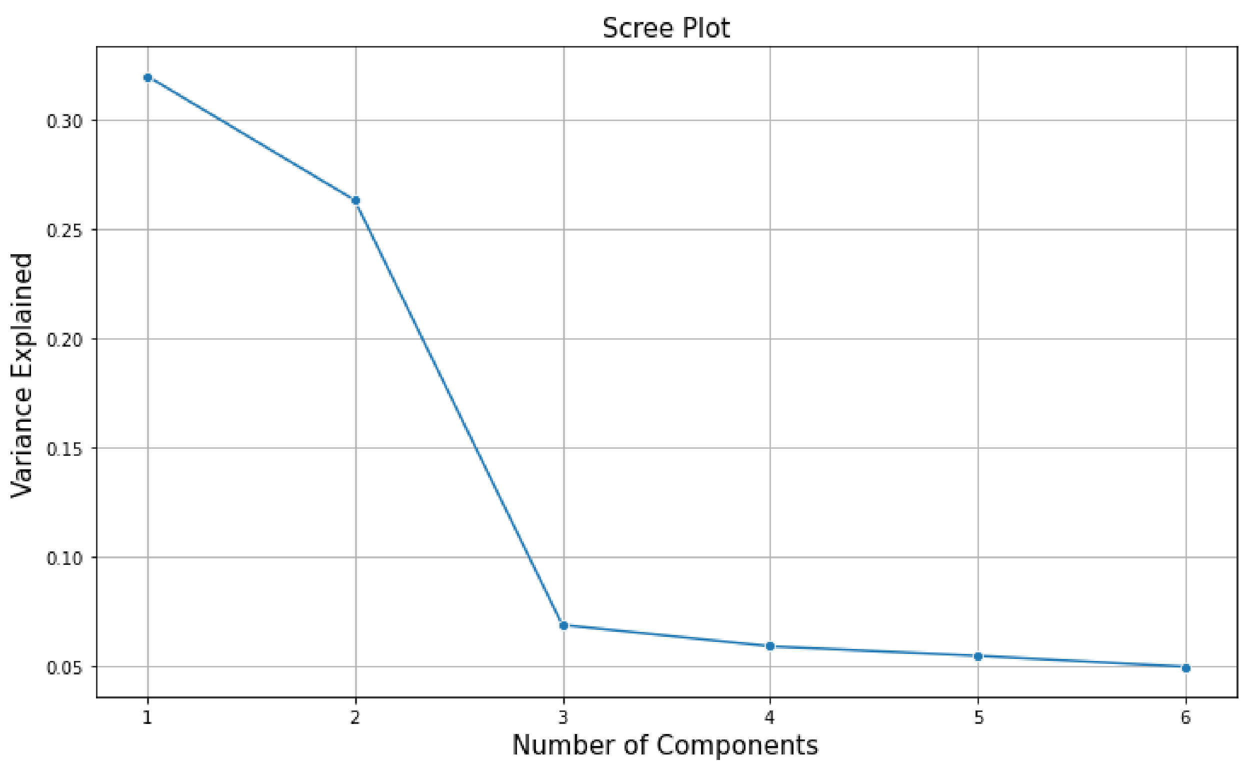


Figure 9 – Scree Plot

optimum number of principal components can be decided by principal components. Although there are 17 original attributes, more than 80% of the total variance can be explained with only the first 6 PC’s which helps to achieve goal of dimension reduction. In scree plot, there is a distinct break at 2. However, k cannot be taken to be 2 since the first two PCs explain only 58% of total variance. The PCs must be taken so as to explain between 70% - 90% of the total variance. If k = 6, then the first 6 PCs explain 81.6% of the total variance. One choice of k could have been 4. However, we have taken k = 6 so that the explained variance is above 81.6%

Eigen vectors indicate directions of the spread of our data.

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]

With the help of PCA, we are able to reduce 17 numeric variables or dimensions to 6 dimension which is able to explain 81.6% of variance in the data. This helps in saving the time and effort to analyze the data.

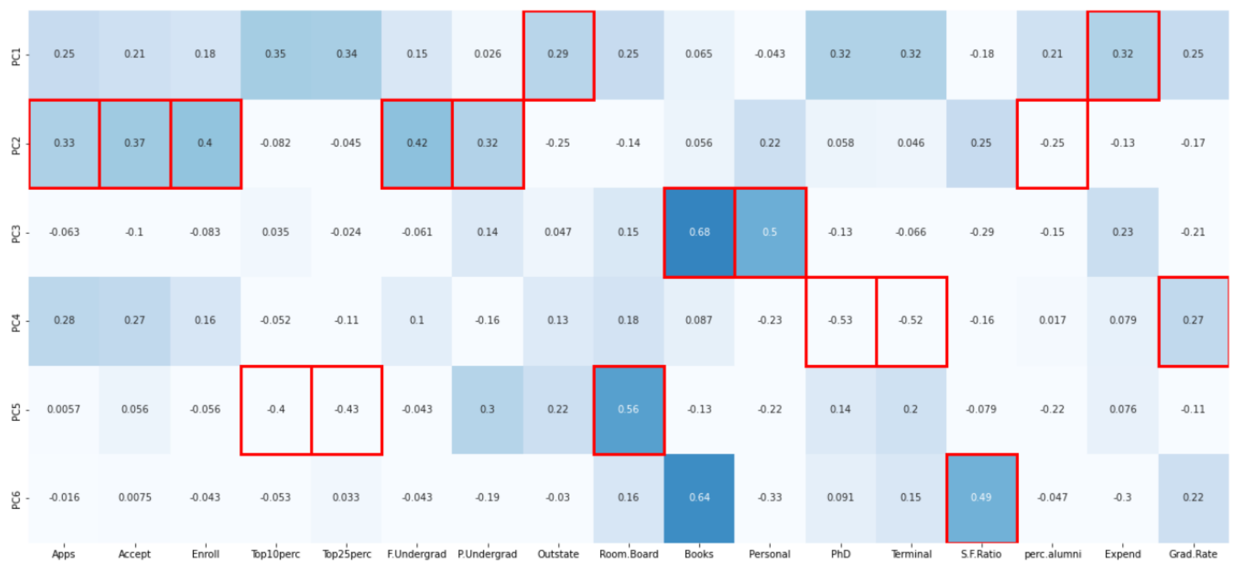


Figure 10 – PCA loading

With help of reduced components, we can observe some patterns. Using the components additional rules can be derived and analyzed.

With Figure 10 we can identify which features have maximum loading across the components. Features marked with rectangular red box are the one having maximum loading on the respective component. We consider these marked features to decide the context that the component represents with the help of subject matter expert. This way we can create new and more relevant features from the original features.

Unsupervised learning like clustering can further be applied on the data to segment the universities based on the components created and further analyzed.