PGP - DSBA

Advanced Statistics

Project Report – August 2022

Shruti Jha
8-14-2022

Contents

Problem 1

List of Tables

List of Figures

Problem 2

List of Tables

List of Figures

Problem 1 – Individuals Salary Analysis

Executive Summary

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

Introduction

The purpose of this case study is to explore the dataset and perform Analysis of Variance (ANOVA) to study the implications of 2 features (Education and Occupation levels) can have on Salary. We are going to observe the effects on Salary caused by each individual feature and both features combined.

Data Description

1. Education: Educational qualification (High school graduate, Bachelor, and Doctorate)
2. Occupation: Job or source of income (Administrative or clerical, Sales, Professional or specialty, and Executive or managerial)
3. Salary: Income

Sample of dataset

Head function shows top 5 rows of the dataset.

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

Table 1: Dataset Sample

- Dataset has 3 variables. There are 3 types of educational levels and 4 types of occupations, as explained in the Executive Summary.
- Salary column denotes the salary earned by 40 individuals with a specific educational level, at a certain occupation level.

Types of variables in the dataset

```
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)
```

- Education and Occupation are in object format and Salary is in numerical format.
- There are a total of 40 observations (rows) under 3 features (columns).

Missing values in the dataset

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

- There are no missing values in the dataset.

Descriptive Statistics

Describe function provides a table indicating the count of variables, mean, standard deviation and other values for the 5-point summary that includes (min, 25%, 50%, 75% and max). 50% in the table is also known as median.

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

Table 2: Dataset Description

- The highest number of candidates, i.e. 16, have educational qualification of Doctorate level.
- The highest number of candidates, i.e. 13, work at the Professional or specialty level.
- Salary ranges between INR 50,103 and INR 260,151.

Problem 1A – Individuals Salary Analysis

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

Null and Alternate Hypothesis for Education:

H_0 = The mean salary is same for all 3 education levels

H_1 = The mean salary is different for at least 1 of the 3 education levels

Null and Alternate Hypothesis for Occupation:

H_0 = The mean salary is same for all 4 occupations

H_1 = The mean salary is different for at least 1 of the 4 occupations

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

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

Table 3: Education ANOVA

- Since the p-value (1.257709e-08) is less than the significance level (0.05), we can reject the null hypothesis (H_0).
- Hence, we can conclude the there is a difference in the mean salary for at least 1 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.

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	0.884144	0.458508
Residual	36.0	1.528092e+11	4.244701e+09	NaN	NaN

Table 4: Occupation ANOVA

- Since the p-value (0.458508) is more than the significance level (0.05), we can accept the null hypothesis (H_0).
- Hence, we can conclude that the mean salary is same for all 4 levels of occupation.

Problem 1B – Individuals Salary Analysis

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

We have observed how the treatments affect Salary individually. Now we will analyse how both the treatments, combined, affect the Salary by establishing interaction between them.

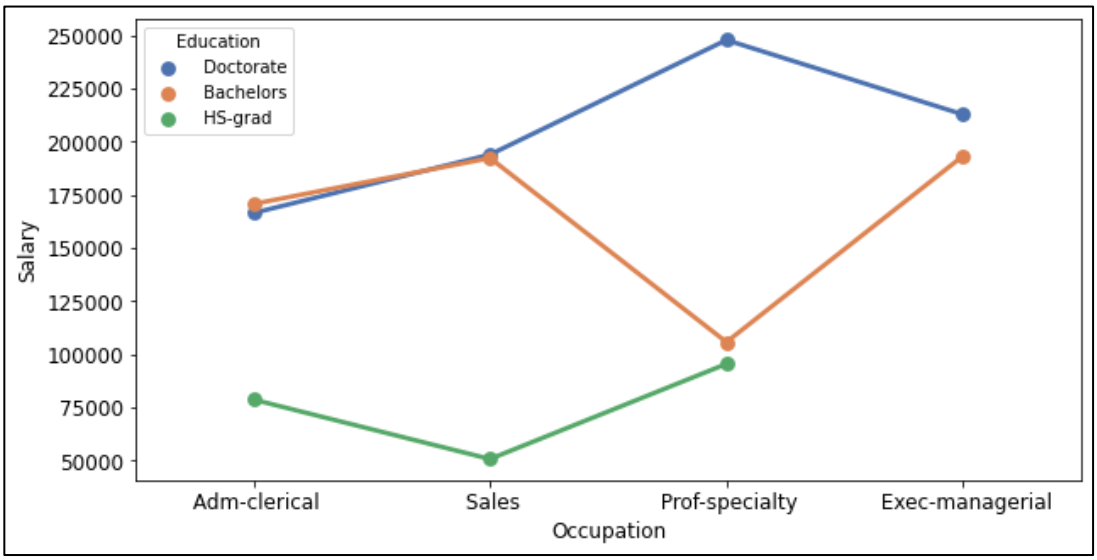


Figure 1: Interaction Plot

- By looking at this point plot, we can infer that there is interaction present between the two variables, as the line segments are not parallel.
- For the individuals with Doctorate and Bachelors degree:
  - They are earning almost equal salaries for Admin-clerical, Sales and Exec-managerial jobs positions.
  - There is a huge difference between the salaries they are earning at Prof-speciality job positions.
- For the individuals with Bachelors and High school graduation degree:
  - They are earning almost equal salaries for Prof-speciality job position.
  - There is a huge difference between the salaries they are earning at Admin-clerical and Sales job positions.
- High school graduates reach the Exec-managerial job level. And their salaries are very low as compared to the individuals with doctorate degree for rest of the 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?

Null and Alternate Hypothesis for interaction effect between Education and Occupation on Salary:

H_0 = There is no interaction effect between the 2 independent variables, Education and Occupation, on the mean salary

H_1 = There is some interaction effect between the 2 independent variables, Education and Occupation, on the mean salary

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

Table 5: Education & Occupation ANOVA

As we can see the p-value (2.232500e-05) for the interaction between Education and Occupation is less than the significance level of 0.05. So, we can reject the null hypothesis (H_0).

Hence, we can conclude that there is a difference between the salaries individuals are earning dependent on the combination of Education and Occupation levels combined.

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

As we observed using one-way and two-way analysis of variance (ANOVA):

- There is significantly less impact of Occupation level alone on the Salary.
    - However, if we plot a graph to see the salary trend in terms of occupation of the individuals, we observe that salary of individuals increases as the occupation level increases admin-clarical jobs to executive-managerial positions.

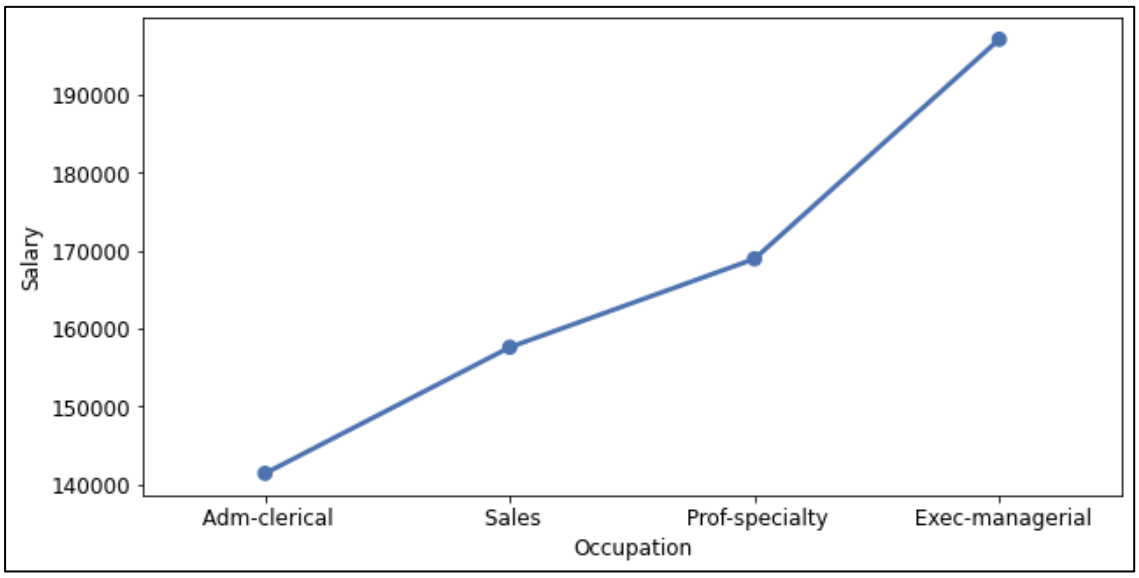


Figure 2: Occupation to Salary Plot

- The Salary is significantly impacted by the level of Education of an individual.
    - From the graph we observe that as the level of education increases from high school graduate to Doctorate degree, the salary also increases.

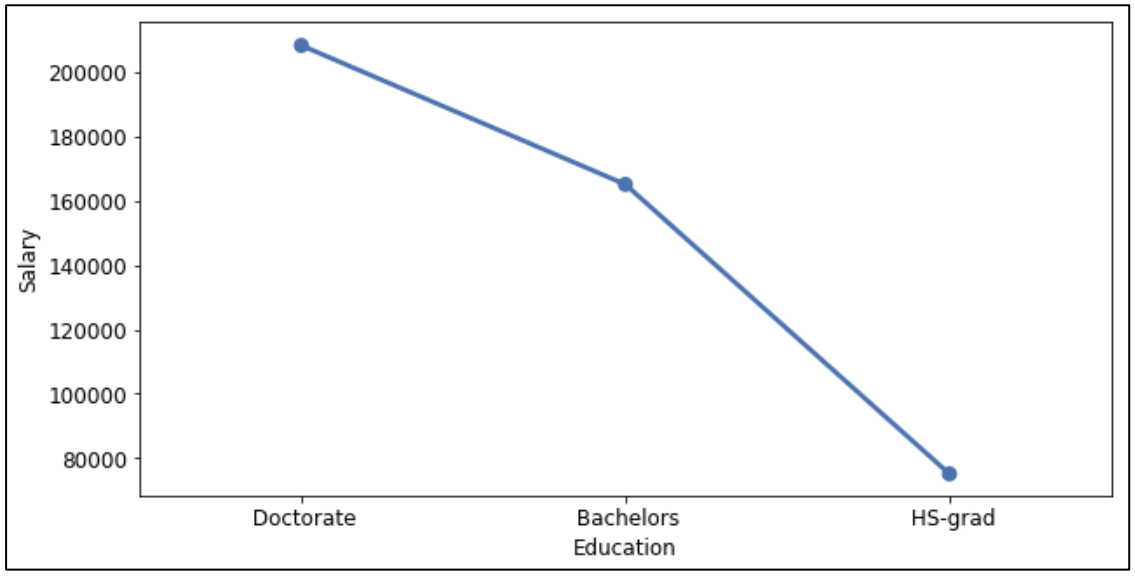


Figure 3: Education to Salary Plot

- The combination of Education and Occupation levels combined, also impact the Salary of an individual.

When we observe the impact of education and occupation levels together, we can see that:

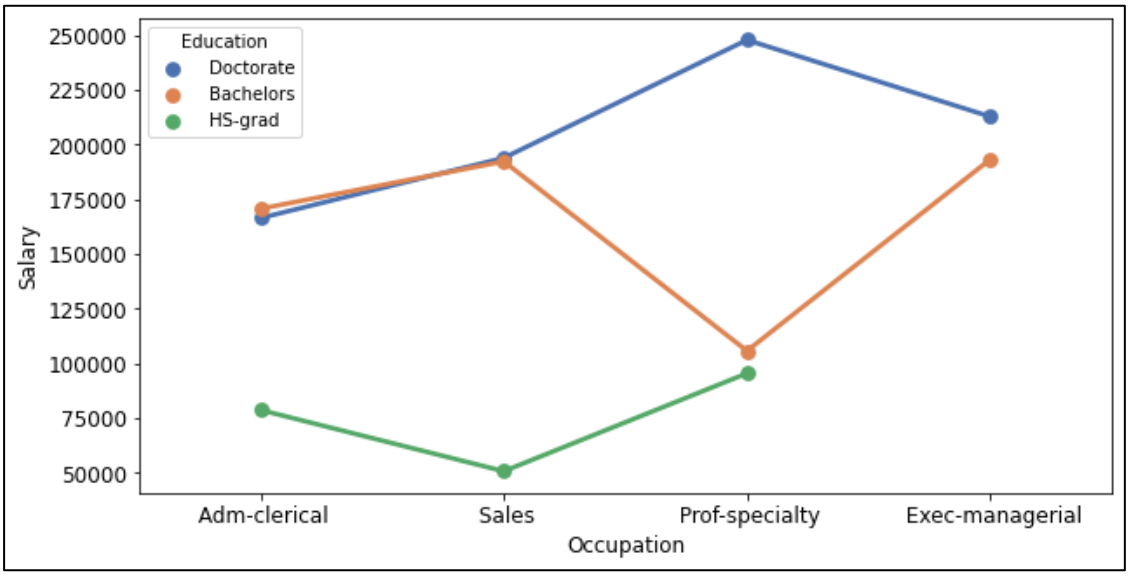


Figure 4: Occupation & Education Interaction Plot

- Individuals with Doctorate and Bachelors degree are suitable for all levels of occupation and Doctorate degree being the higher level of education does not define that those individuals are paid a lot high than the individuals with bachelors degree.
- High school graduates are paid much less than individuals with higher education and are not considered for the executive or managerial positions.

The HR department of a firm, for whom this salary trends may come in handy while deciding the salary of an individual based on the education and occupation levels individually or combined.

Problem 2 – College Data Analysis

Executive Summary

In the dataset, information on various colleges is provided. It consists of data about students, faculties and alumni. In this problem statement we will explore the different attributes associated with various colleges.

Introduction

The purpose of this case study is to explore the dataset, perform Exploratory Data Analysis and Principal Component Analysis and explain the business implication. The data consists of information from 777 colleges with 18 attributes.

Data Description

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

Sample of dataset

Head function shows top 5 rows of the dataset.

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

Table 1: Dataset Sample 1

Featuring Engineering - From the existing columns, we may be able to derive additional feature(s) which may be more relevant considering the objective to be achieved and may give us better insights.

For the purpose of better understanding the expenses made by students, an additional column "TotalExpns" has been included in the dataset to show combined cost of room, books and personal expenses that students make.

This is how the new sample of dataset looks like:

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	TotalExpns
Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60	5950
Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56	8700
Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54	5315
Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59	6775
Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15	6420

Table 2: Dataset Sample 2

There are 19 variables, 'Names' has unique college names listed. Each college has different sets of information provided against them.

Types of variables in the dataset

```
RangeIndex: 777 entries, 0 to 776
Data columns (total 19 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  
 18  TotalExpns   777 non-null    int64  
dtypes: float64(1), int64(17), object(1)
```

All the variables are in numeric (integer and float) format except 'Names' which is in object format.

There are a total of 777 rows and 19 columns in the dataset.

Missing values in the dataset

```
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
TotalExpns     0
```

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

Descriptive Statistics

Describe function provides a table indicating the count of variables, mean, standard deviation and other values for the 5-point summary that includes (min, 25%, 50%, 75% and max). 50% in the table is also known as median.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Names	777	777	Abilene Christian University	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Apps	777.0	NaN	NaN	NaN	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	NaN	NaN	NaN	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	NaN	NaN	NaN	779.972973	929.17619	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	NaN	NaN	NaN	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	NaN	NaN	NaN	55.796654	19.804778	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	NaN	NaN	NaN	3699.907336	4850.420531	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	NaN	NaN	NaN	855.298584	1522.431887	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	NaN	NaN	NaN	10440.669241	4023.016484	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	NaN	NaN	NaN	4357.526384	1096.696416	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	NaN	NaN	NaN	549.380952	165.10536	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	NaN	NaN	NaN	1340.642214	677.071454	250.0	850.0	1200.0	1700.0	6800.0
PhD	777.0	NaN	NaN	NaN	72.660232	16.328155	8.0	62.0	75.0	85.0	103.0
Terminal	777.0	NaN	NaN	NaN	79.702703	14.722359	24.0	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	NaN	NaN	NaN	14.089704	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777.0	NaN	NaN	NaN	22.743887	12.391801	0.0	13.0	21.0	31.0	64.0
Expend	777.0	NaN	NaN	NaN	9660.171171	5221.76844	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	NaN	NaN	NaN	65.46332	17.17771	10.0	53.0	65.0	78.0	118.0
TotalExpns	777.0	NaN	NaN	NaN	6247.54955	1216.013036	3452.0	5400.0	6100.0	6958.0	12330.0

Table 3: Dataset Description 1

From the above descriptive statistics, we can infer:

- We have 1 categorical field 'Names'.
- On an average, colleges received close to 3001 applications, out of which around 2018 applications got accepted and 780 students enrolled.
- The number of applications received ranges between 81 and 48094.
- The number of applications accepted ranges between 72 and 26330.
- The number of new students enrolled ranges between 35 and 6392.
- In Massachusetts Institute of Technology 96% new students fall under top 10% of higher secondary class.

`

	Names	Top10perc
354	Massachusetts Institute of Technology	96

Table 4: College with Top10perc

- In the below listed colleges all the new students fall under top 25% of higher secondary class:

	Names	Top25perc
60	Bowdoin College	100
250	Harvard University	100
251	Harvey Mudd College	100
562	SUNY at Buffalo	100
605	University of California at Berkeley	100
606	University of California at Irvine	100
663	University of Pennsylvania	100

Table 5: College with Top25perc

- On an average, there are 10440 students for whom the particular college or university is out-of-state tuition.
- Students from Missouri Southern State College spend least total amount and students from Saint Louis University spend highest total amount on room, books and personal expenses. Students spend an average amount of USD 6247.55.

	Names	TotalExpns
377	Missouri Southern State College	3452

Table 6: Lowes Total Expense

	Names	TotalExpns
497	Saint Louis University	12330

Table 7: College with Top25perc

- The maximum percentage of faculties with PhD is given 103%, and the maximum graduation rate is given 118%. Both need to be cleaned, as the maximum percentage cannot exceed 100%.

NaN shows that the values cannot be calculated for that particular variable. Like we cannot calculate mean for a categorical/object type variable, and in a same way unique value for a numerical variable.

Fixing the exceeding percentage value under PhD and graduation rate columns: Using np.where() function the percentage exceeding 100% have been replaced to be withing the limit.

Let's have a look at the final data description:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Names	777	777	Abilene Christian University	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Apps	777.0	NaN	NaN	NaN	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	NaN	NaN	NaN	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	NaN	NaN	NaN	779.972973	929.17619	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	NaN	NaN	NaN	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	NaN	NaN	NaN	55.796654	19.804778	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	NaN	NaN	NaN	3699.907336	4850.420531	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	NaN	NaN	NaN	855.298584	1522.431887	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	NaN	NaN	NaN	10440.669241	4023.016484	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	NaN	NaN	NaN	4357.526384	1096.696416	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	NaN	NaN	NaN	549.380952	165.10536	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	NaN	NaN	NaN	1340.642214	677.071454	250.0	850.0	1200.0	1700.0	6800.0
PhD	777.0	NaN	NaN	NaN	72.656371	16.321324	8.0	62.0	75.0	85.0	100.0
Terminal	777.0	NaN	NaN	NaN	79.702703	14.722359	24.0	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	NaN	NaN	NaN	14.089704	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777.0	NaN	NaN	NaN	22.743887	12.391801	0.0	13.0	21.0	31.0	64.0
Expend	777.0	NaN	NaN	NaN	9660.171171	5221.76844	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	NaN	NaN	NaN	65.440154	17.118804	10.0	53.0	65.0	78.0	100.0
TotalExpns	777.0	NaN	NaN	NaN	6247.54955	1216.013036	3452.0	5400.0	6100.0	6958.0	12330.0

Table 8: Dataset Description 2

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

Before Exploratory Data Analysis, we'll perform below mentioned checks:

1. **Check for duplicate records**

```
Number of duplicate entries:  0
```

Using "duplicated()" function, we can derive that there are no duplicate records in the data.

2. **Check for missing values**

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

Using "isnull().sum()" functions we can derive that there are no missing values in the dataset.

3. **Check for outliers**

To check for outliers, box plots have been plotted:

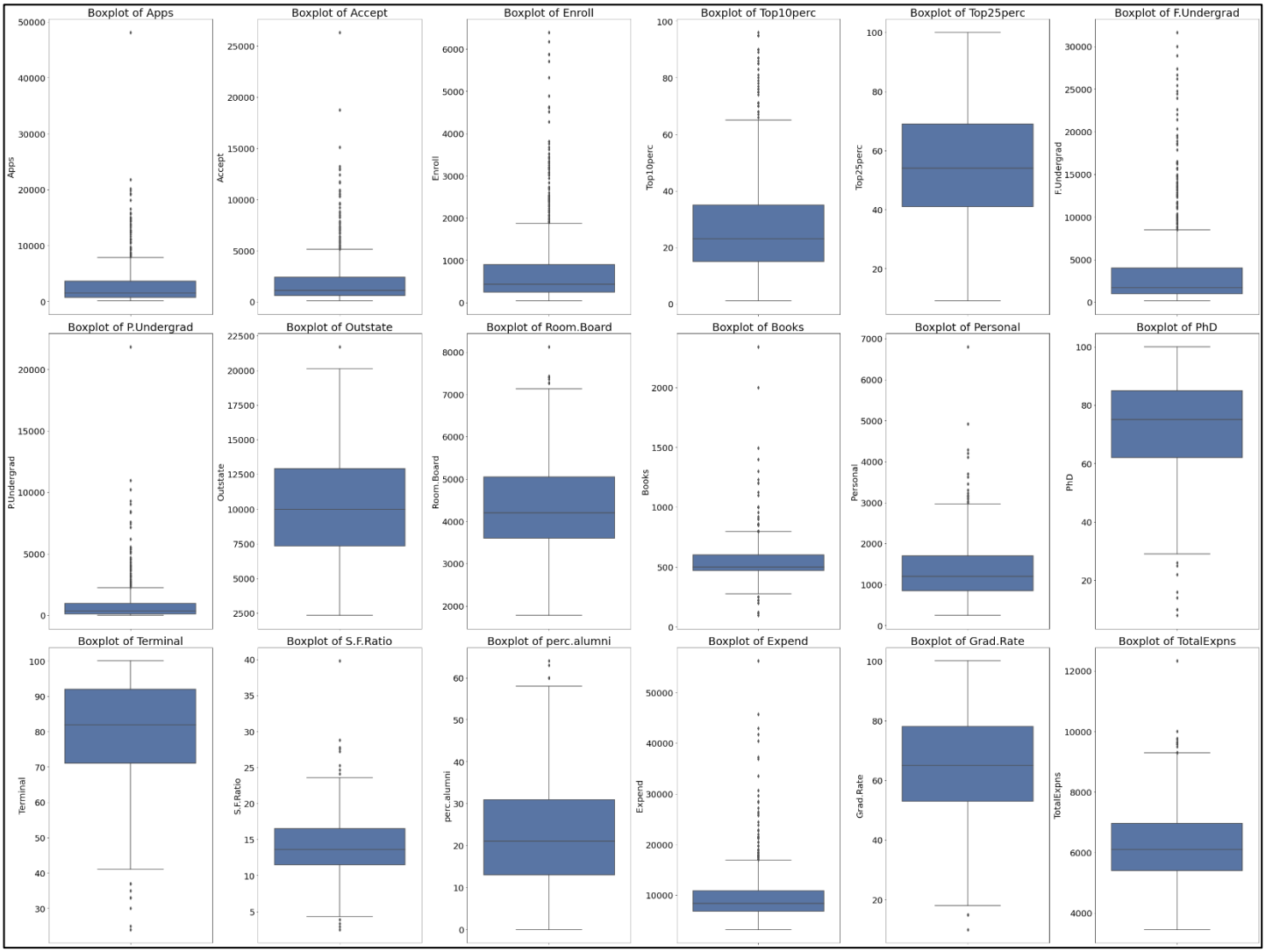


Figure 1: Outliers Check

- The small dots outside the whiskers of boxplots denote outliers. As we can infer from the above plots, all the variables have outliers / extreme values present in them, except Top25perc column.
- The outlier treatment is not required, as the extreme values in all the columns are relevant to the dataset.

2.1.1 Univariate analysis

This is the sample of data set with only numerical variables:

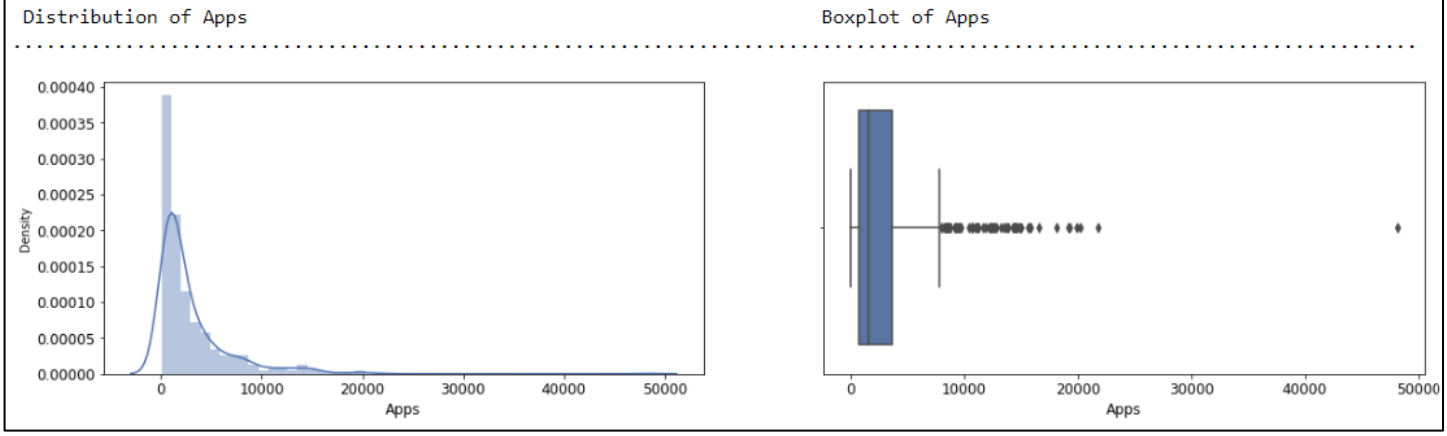
Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate	TotalExpns
1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60	5950
2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56	8700
1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54	5315
417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59	6775
193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15	6420

```
(777, 18)
```

Table 9: Numeric Dataset Sample

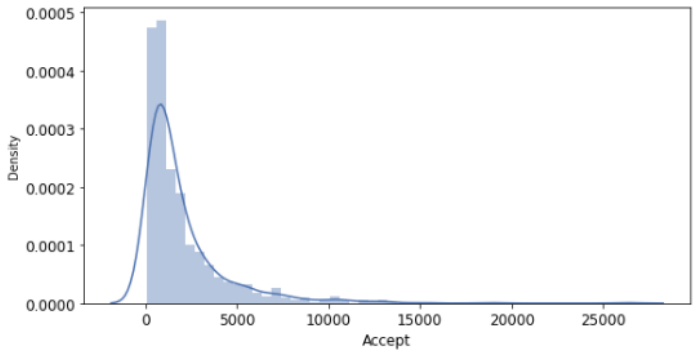
Univariate analysis is performed for all the numeric variables individually to display their statistical description. Visualized the variables using distplot to view the distribution and the box plot to view 5-point summary and outliers if any.

```
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
```

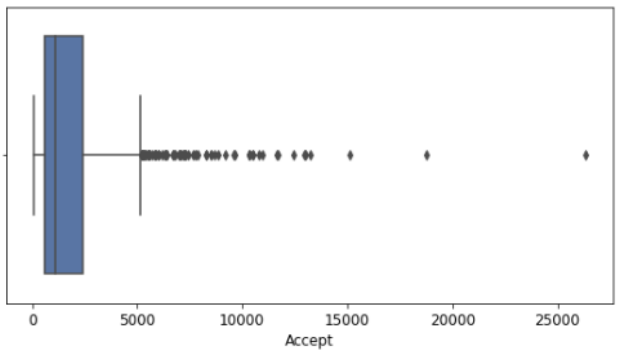


```
Description of Accept
.....................................................................
count      777.000000
mean      2018.804376
std       2451.113971
min         72.000000
25%        604.000000
50%       1110.000000
75%       2424.000000
max      26330.000000
Name: Accept, dtype: float64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

Distribution of Accept

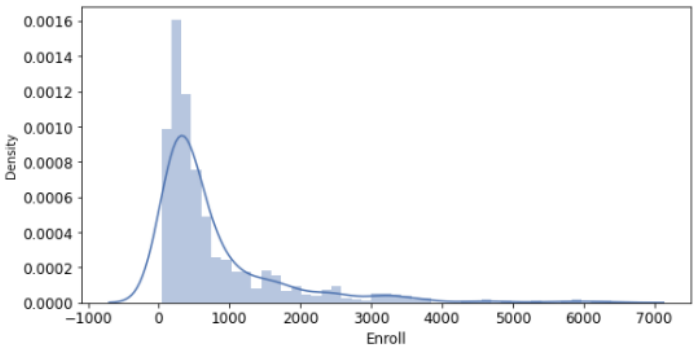


Boxplot of Accept

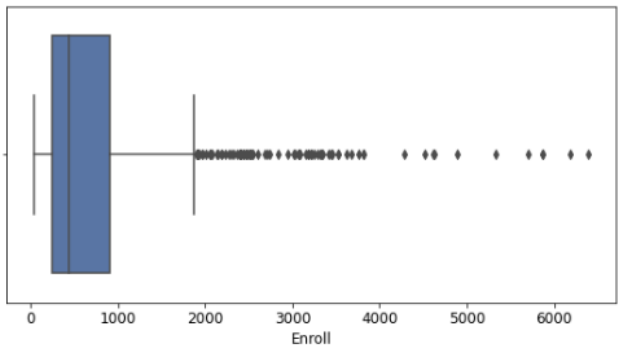



```
Description of Enroll
.....................................................................
count      777.000000
mean       779.972973
std        929.176190
min         35.000000
25%        242.000000
50%        434.000000
75%        902.000000
max       6392.000000
Name: Enroll, dtype: float64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

Distribution of Enroll

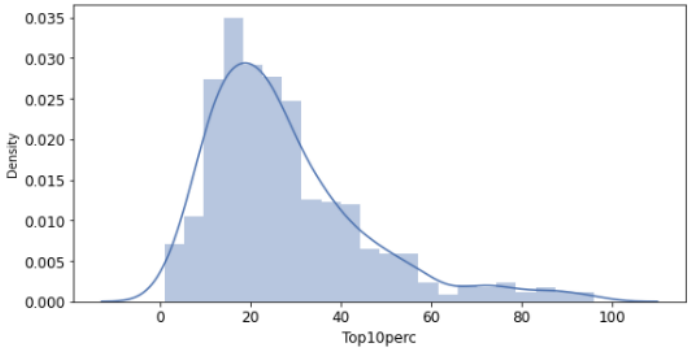


Boxplot of Enroll

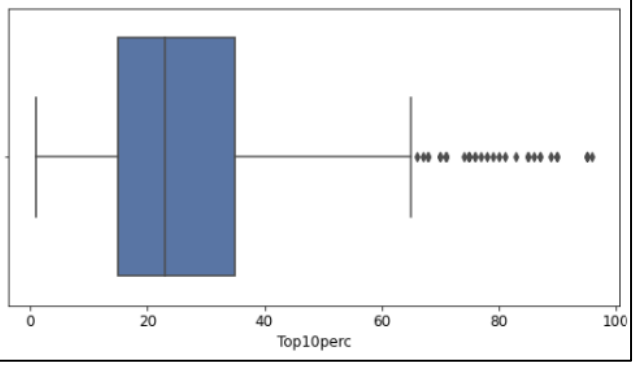


```
Description of Top10perc
.....................................................................
count    777.000000
mean      27.558559
std       17.640364
min        1.000000
25%       15.000000
50%       23.000000
75%       35.000000
max       96.000000
Name: Top10perc, dtype: float64
```

Distribution of Top10perc

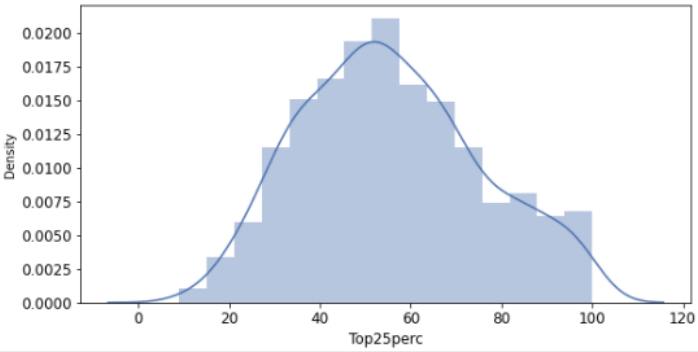


Boxplot of Top10perc

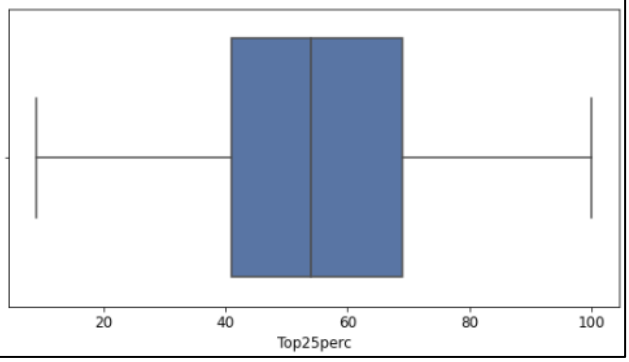



```
Description of Top25perc
.....................................................................
count    777.000000
mean      55.796654
std       19.804778
min        9.000000
25%       41.000000
50%       54.000000
75%       69.000000
max      100.000000
Name: Top25perc, dtype: float64
```

Distribution of Top25perc

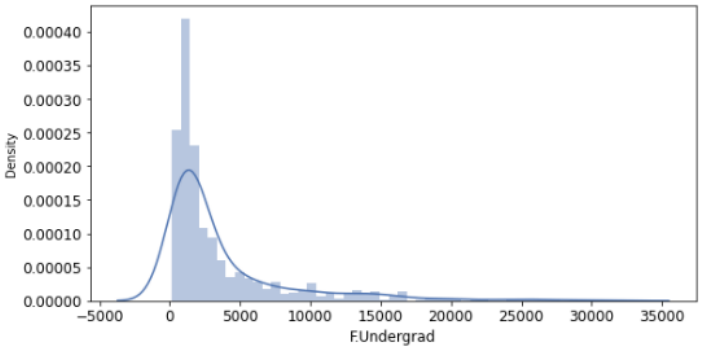


Boxplot of Top25perc

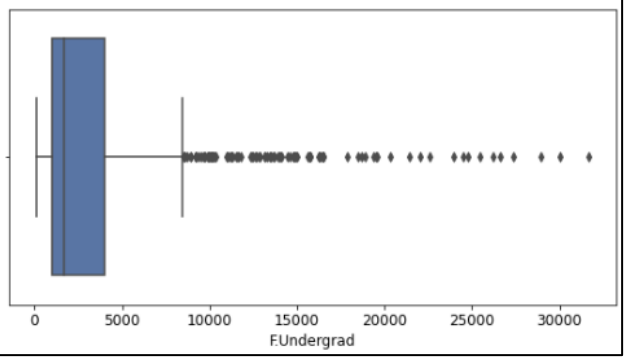


```
Description of F.Undergrad
.....................................................................
count      777.000000
mean      3699.907336
std       4850.420531
min        139.000000
25%        992.000000
50%       1707.000000
75%       4005.000000
max      31643.000000
Name: F.Undergrad, dtype: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

Distribution of F.Undergrad

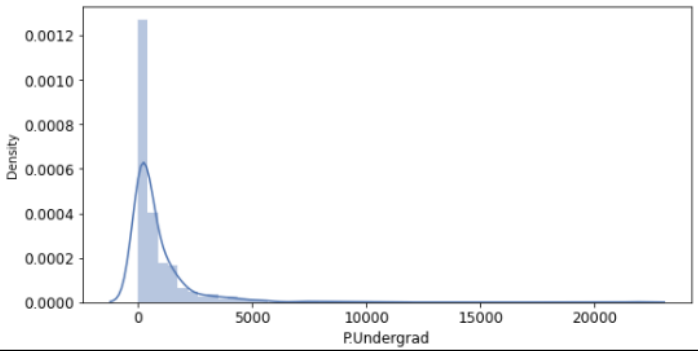


Boxplot of F.Undergrad

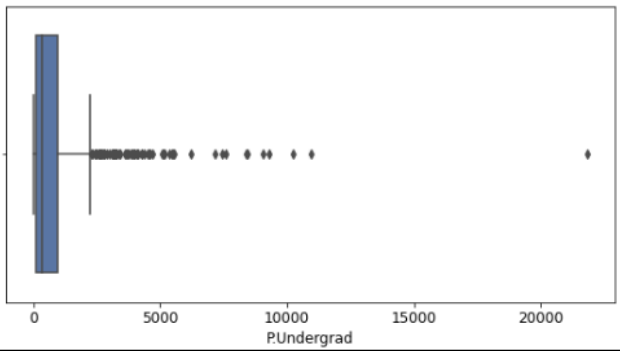



```
Description of P.Undergrad
.....................................................................
count      777.000000
mean       855.298584
std       1522.431887
min          1.000000
25%         95.000000
50%        353.000000
75%        967.000000
max      21836.000000
Name: P.Undergrad, dtype: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

Distribution of P.Undergrad

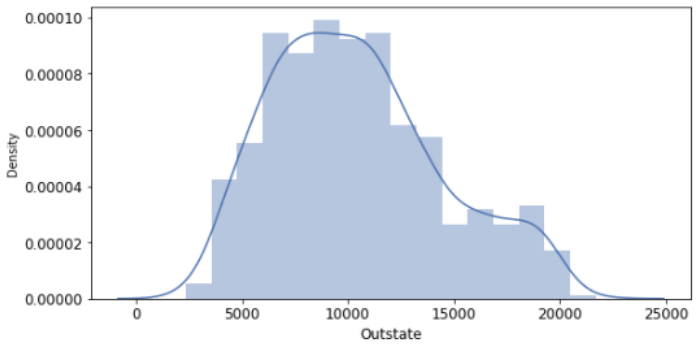


Boxplot of P.Undergrad

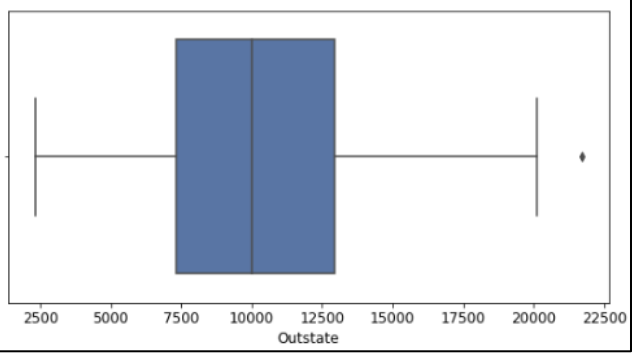


```
Description of Outstate
.....................................................................
count      777.000000
mean     10440.669241
std       4023.016484
min       2340.000000
25%       7320.000000
50%       9990.000000
75%      12925.000000
max      21700.000000
Name: Outstate, dtype: float64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

Distribution of Outstate

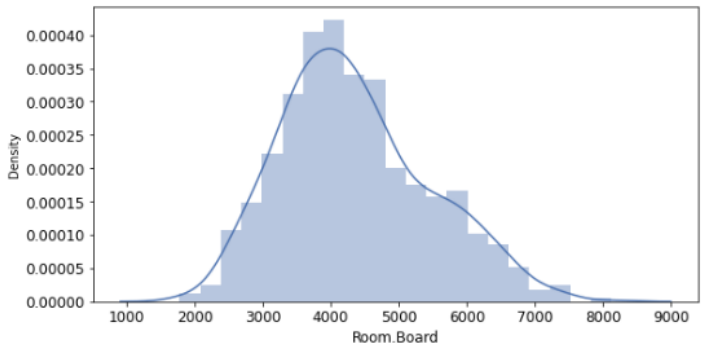


Boxplot of Outstate

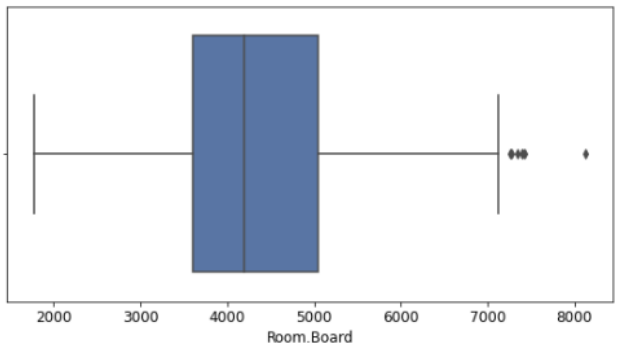



```
Description of Room.Board
.....................................................................
count     777.000000
mean     4357.526384
std      1096.696416
min      1780.000000
25%      3597.000000
50%      4200.000000
75%      5050.000000
max      8124.000000
Name: Room.Board, dtype: float64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

Distribution of Room.Board

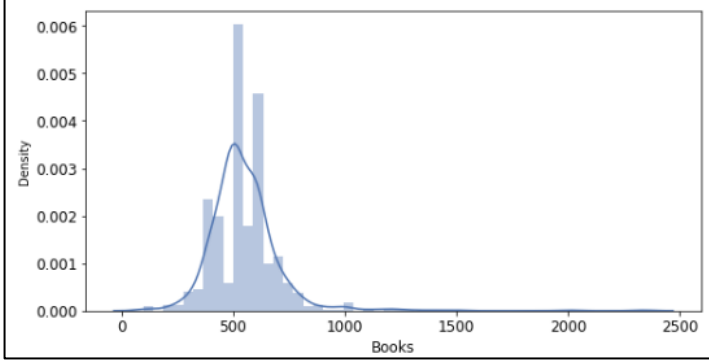


Boxplot of Room.Board

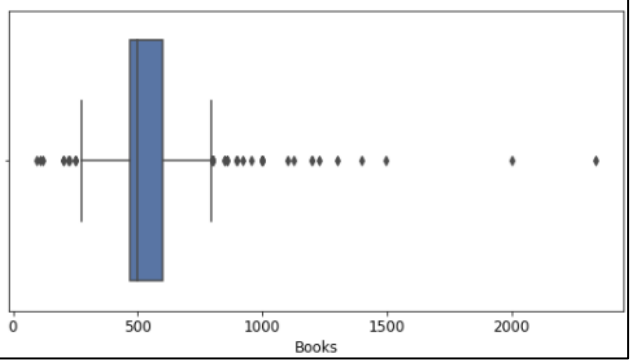


```
Description of Books
.....................................................................
count     777.000000
mean      549.380952
std       165.105360
min        96.000000
25%       470.000000
50%       500.000000
75%       600.000000
max      2340.000000
Name: Books, dtype: float64
```

Distribution of Books

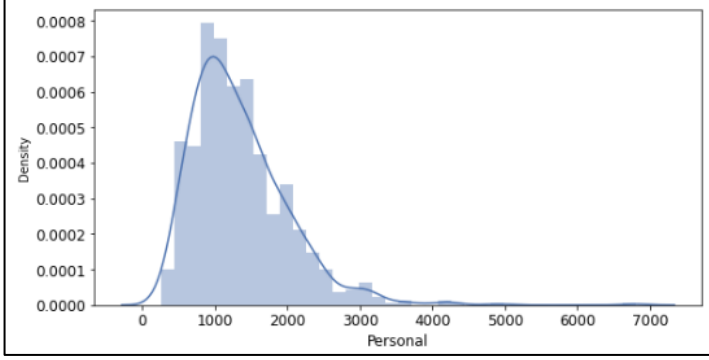


Boxplot of Books

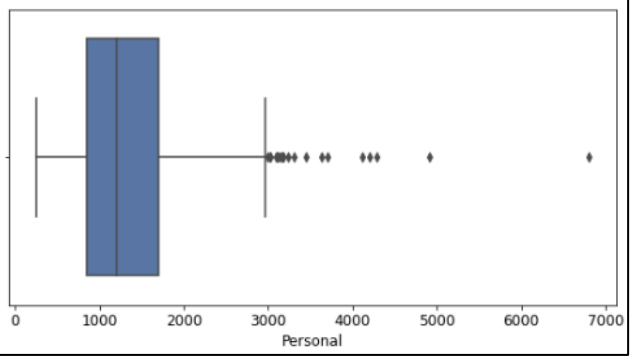



```
Description of Personal
.....................................................................
count     777.000000
mean     1340.642214
std       677.071454
min       250.000000
25%       850.000000
50%      1200.000000
75%      1700.000000
max      6800.000000
Name: Personal, dtype: float64
```

Distribution of Personal

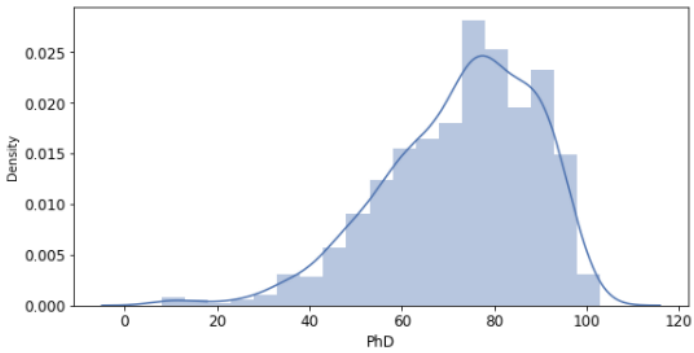


Boxplot of Personal

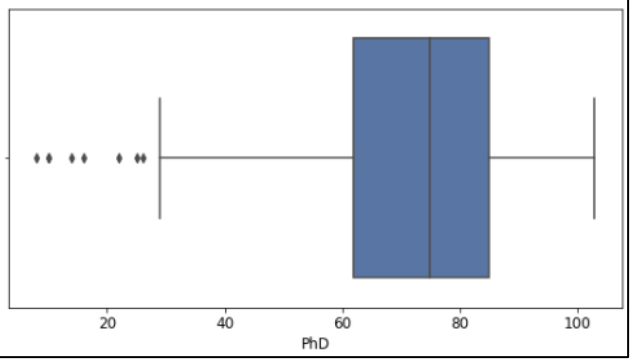


```
Description of PhD
.....................................................................
count    777.000000
mean      72.660232
std       16.328155
min        8.000000
25%       62.000000
50%       75.000000
75%       85.000000
max      103.000000
Name: PhD, dtype: float64
```

Distribution of PhD

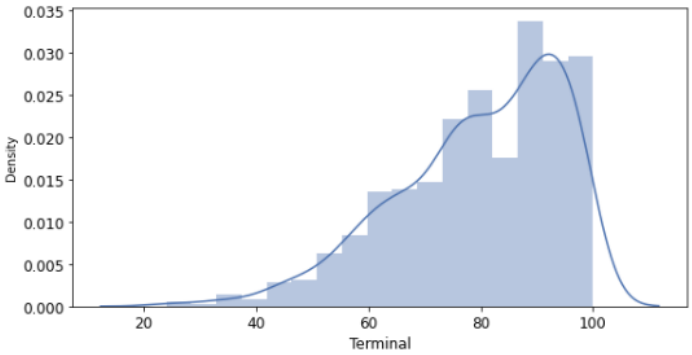


Boxplot of PhD

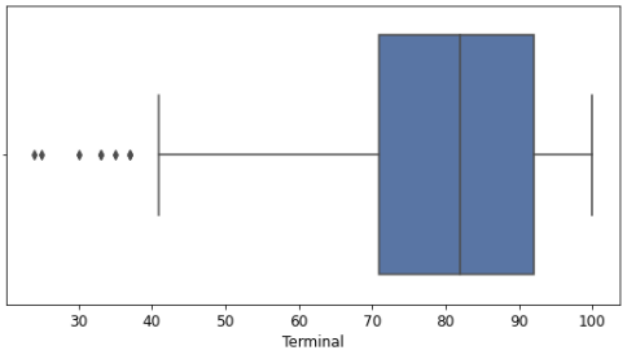



```
Description of Terminal
.....................................................................
count    777.000000
mean      79.702703
std       14.722359
min       24.000000
25%       71.000000
50%       82.000000
75%       92.000000
max      100.000000
Name: Terminal, dtype: float64
```

Distribution of Terminal

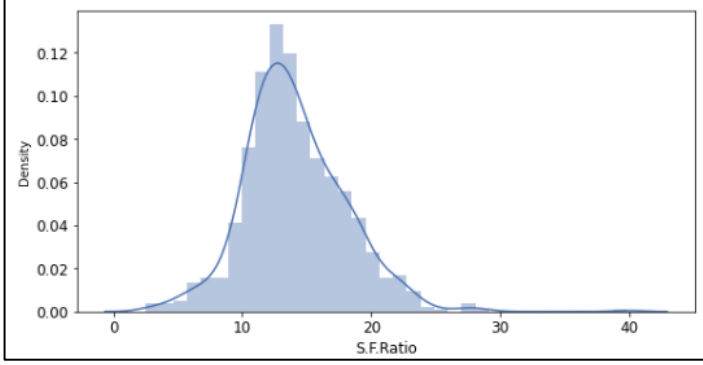


Boxplot of Terminal

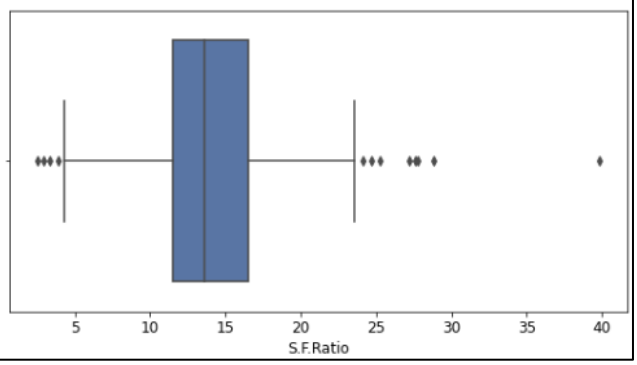


```
Description of S.F.Ratio
.....................................................................
count    777.000000
mean      14.089704
std        3.958349
min        2.500000
25%       11.500000
50%       13.600000
75%       16.500000
max       39.800000
Name: S.F.Ratio, dtype: float64
```

Distribution of S.F.Ratio

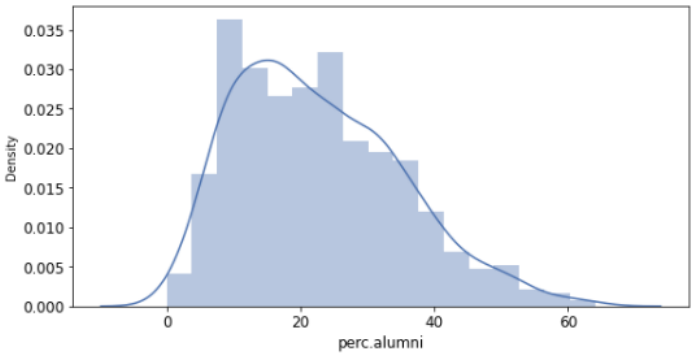


Boxplot of S.F.Ratio

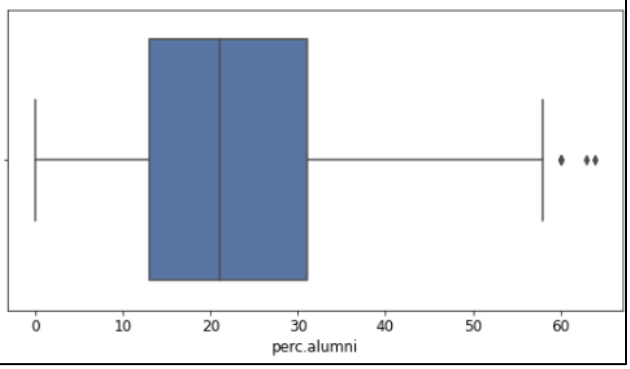



```
Description of perc.alumni
.....................................................................
count    777.000000
mean      22.743887
std       12.391801
min        0.000000
25%       13.000000
50%       21.000000
75%       31.000000
max       64.000000
Name: perc.alumni, dtype: float64
```

Distribution of perc.alumni

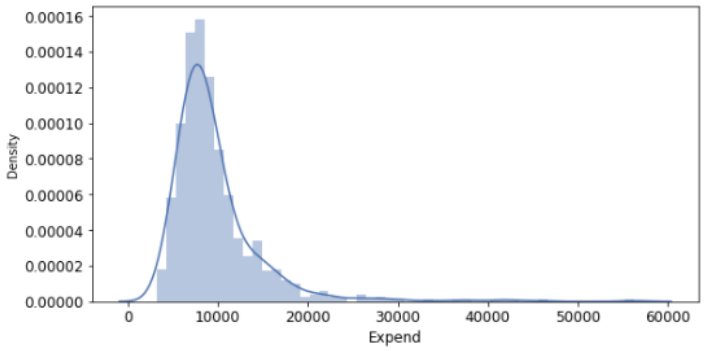


Boxplot of perc.alumni

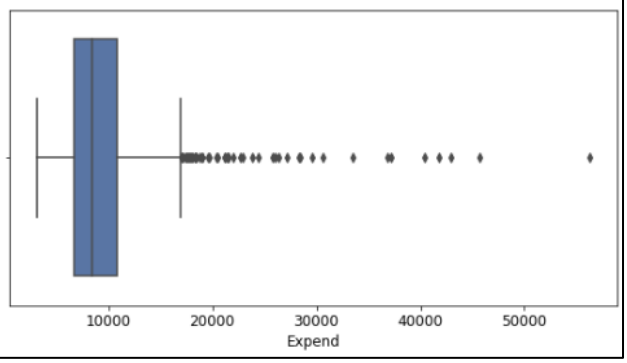


```
Description of Expend
.....................................................................
count      777.000000
mean      9660.171171
std       5221.768440
min       3186.000000
25%       6751.000000
50%       8377.000000
75%      10830.000000
max      56233.000000
Name: Expend, dtype: float64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

Distribution of Expend

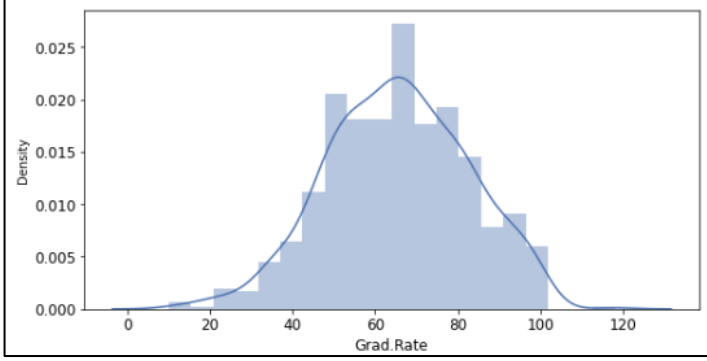


Boxplot of Expend

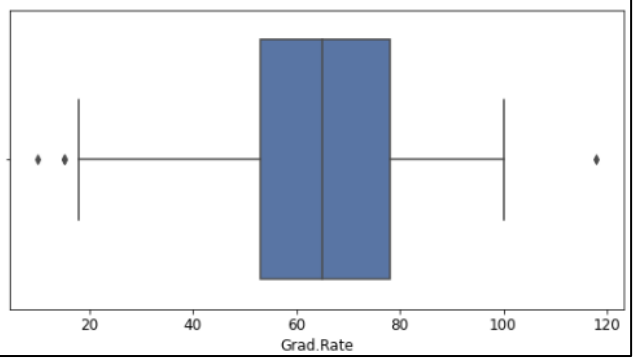



```
Description of Grad.Rate
.....................................................................
count    777.00000
mean      65.46332
std       17.17771
min       10.00000
25%       53.00000
50%       65.00000
75%       78.00000
max      118.00000
Name: Grad.Rate, dtype: float64
```

Distribution of Grad.Rate



Boxplot of Grad.Rate



```
Description of TotalExpns
.....................................................................
count      777.000000
mean      6247.549550
std       1216.013036
min       3452.000000
25%       5400.000000
50%       6100.000000
75%       6958.000000
max      12330.000000
Name: TotalExpns, dtype: float64
```

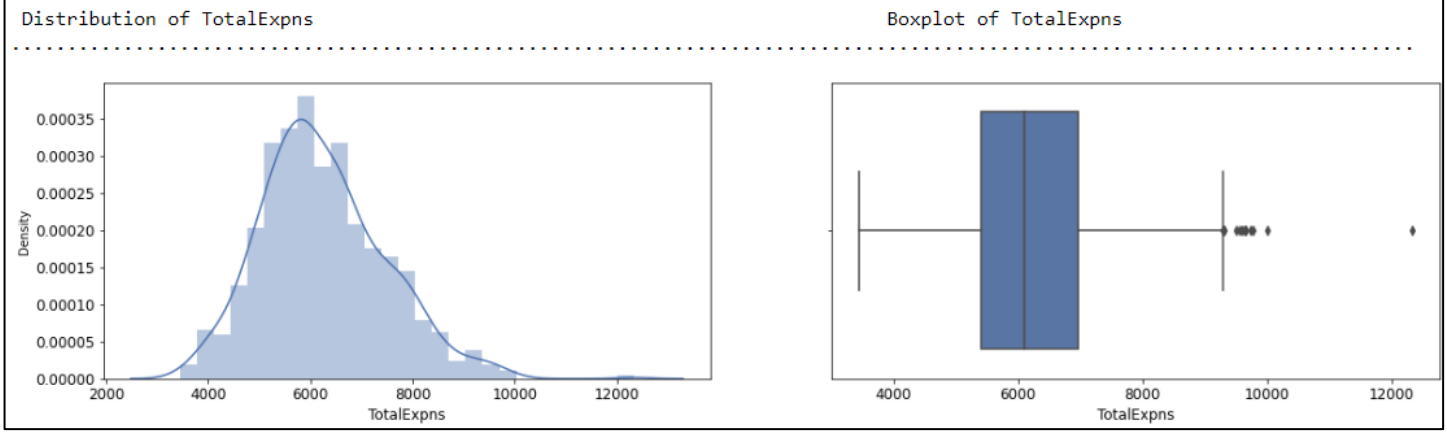


Figure 2: Univariate Analysis

Observations

- There are 18 numeric fields in the dataset.
- Data for most of the variables is right skewed. Except for PhD and Terminal, the data is left skewed for them.
- For variables - Top25perc, S.F.Ratio, Grad.Rate, Perc.alumni and TotalExpns, data is almost normally distributed.
- The outliers present in the data could be valid and not junk. Hence, we are not treating the outliers for this dataset.

2.1.2 Multivariate analysis

Pair plot:

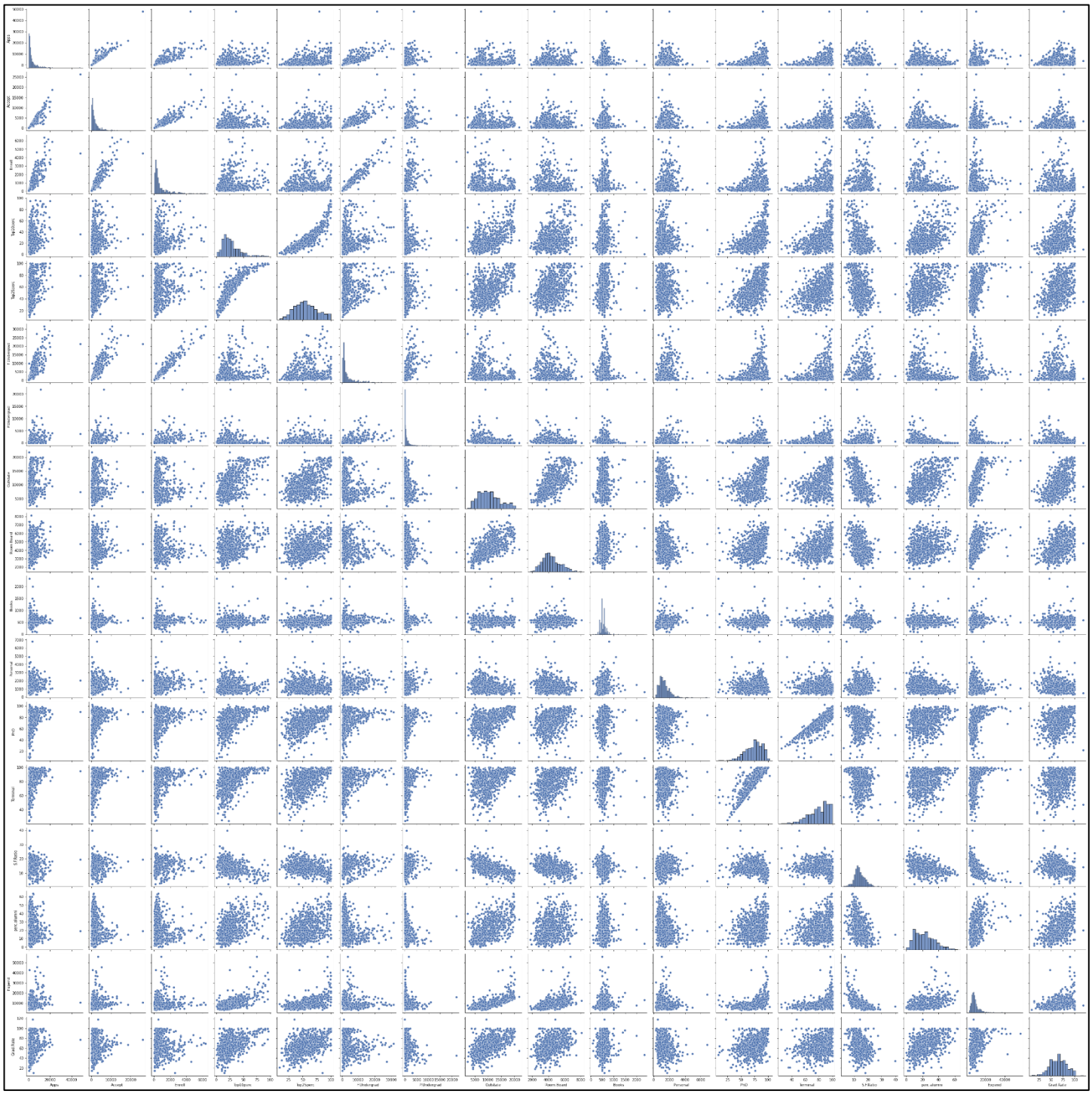


Figure 3: Pairplot

Since there are many numeric variables, the graphs in pair plot have turned out to be very small. They can be read properly when zoomed-in in this document or by doing a double click on the plot in Jupyter notebook.

For the sake of interpreting the pair plot result properly, some segments from the above pair plot have been provided below:

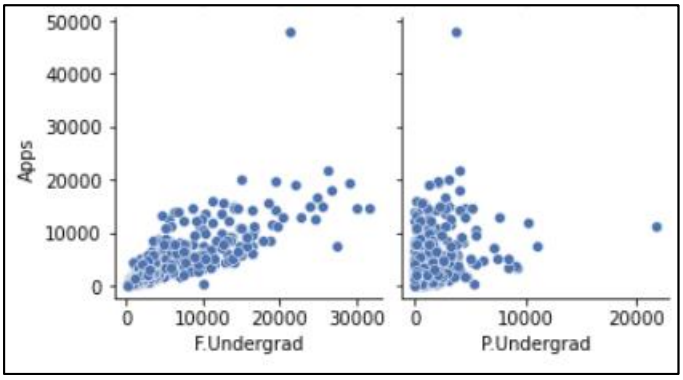


Figure 4: Pairplot 1.1

- The number of applications received are more from full-time undergraduate students and less from part-time undergraduate students.

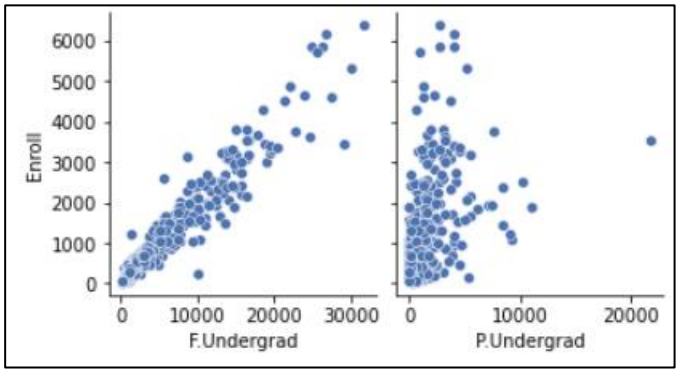


Figure 5: Pairplot 1.2

- More full-time undergraduate students enroll into colleges as compared to the part-time undergraduate students.

- From the above pattern we can say that more students enroll for colleges where the percentage of faculties with PhD and terminal (highest level) degree is higher.

- Students also consider 'Student to Faculty Ratio' which enrolling for a college. We can see that where the S.F Ratio is lesser, most of the students have enroll in those colleges. *(Note: Faculties would pay more precise attention to its students if the students to faculty ration is less)*

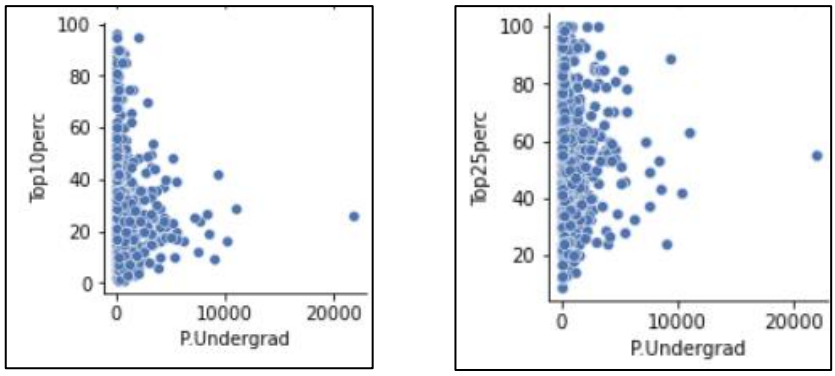


Figure 6: Pairplot 1.3

- Not many part-time students are from top 10% or 25% of higher secondary class

Correlation plot (Heatmap):

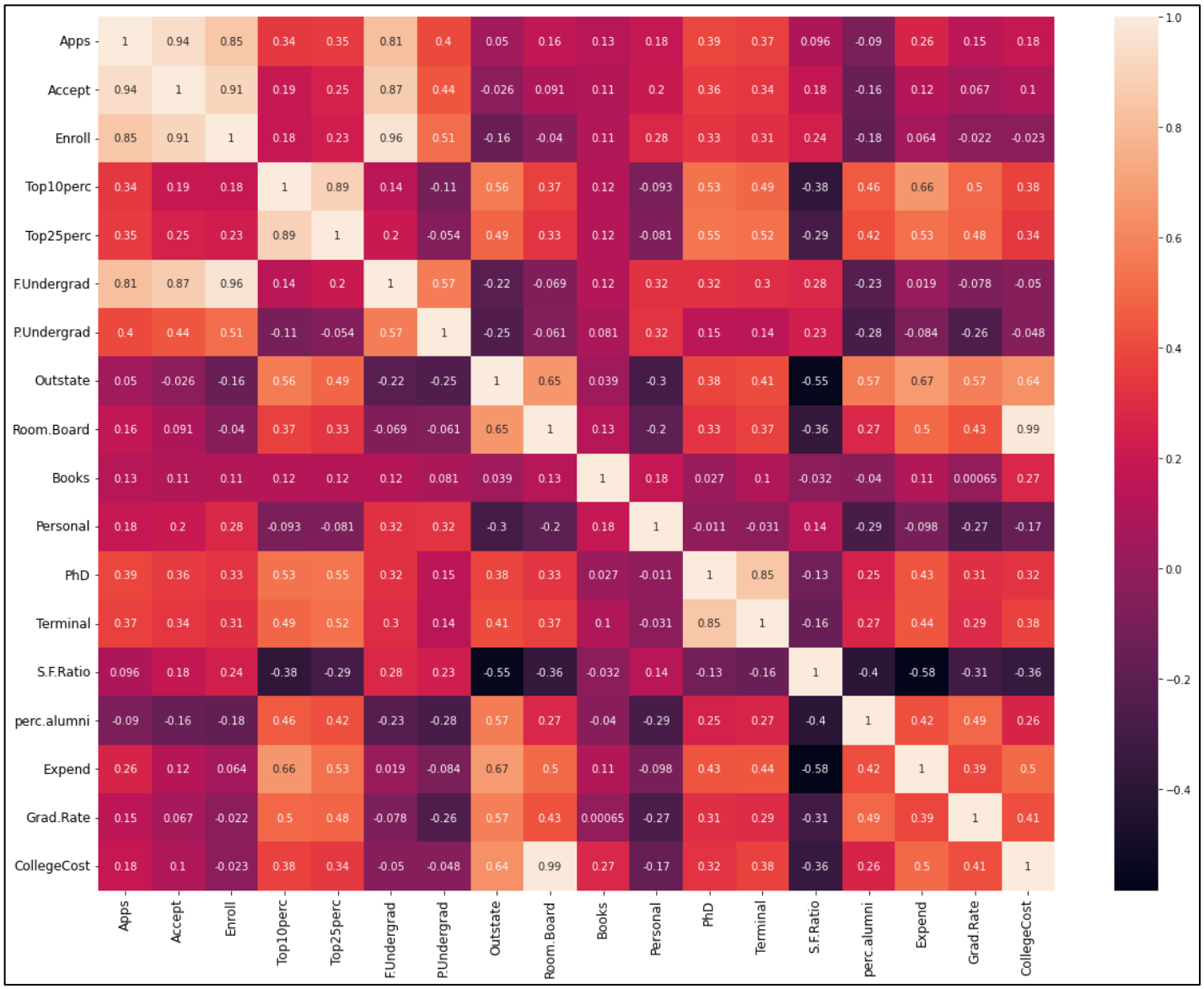


Figure 7: Heatmap

- The correlation is high for top 10% of higher secondary class students and high instructional expenditure by college. From this we can interpret that the enrolment of new students from top 10% of higher secondary class is high where high instructional expenditure is made on the students by the colleges.
- More number of out-of-state students enroll in colleges which make more instructional expenditure per student.
- Out-of-state students and cost of room & boarding shows considerably high corelation. We can say that out-of-state student tend to spend more on accommodation as they are away from their homes.
- Let's understand the negative correlation of student to faculty ratio with various variables. As explained earlier lesser the student to faculty ratio (S.F ratio) the better it is for the students.
    - More Student from the top 10% and 25% of higher secondary class enroll in the colleges where S.F ratio is less.
    - More outstation students enroll in colleges with low S.F ratio and hence they end up paying more on room and boarding.

- More PhD and terminal degree faculties are from the colleges with low S.F ratio.
- Colleges with low S.F ratio make a good amount on instructional expenditure per student as a result the graduation rate is also higher for such colleges
- With low S.F ratio, we can assume the quality of education provided is substantially good, hence we can say that more well-placed alumni tend to make donations to these colleges.

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

Scaling of the data is necessary when the variables of the dataset are of different scales, i.e. one variable is in thousands and other in only hundreds.

In the problem statement we have at hand, there are certain variables which have values of different scales, like Expend and Apps have values in thousands and all the percentage variables have a maximum of 100 as values. Since the data in these variables are of different scales, it is tough to compare them. Hence, the scaling of the variables for PCA is necessary in this case.

Expend	56233.0
Apps	48094.0
F.Undergrad	31643.0
Accept	26330.0
P.Undergrad	21836.0
Outstate	21700.0
TotalExpns	12330.0
Room.Board	8124.0
Personal	6800.0
Enroll	6392.0
Books	2340.0
Top25perc	100.0
PhD	100.0
Terminal	100.0
Grad.Rate	100.0
Top10perc	96.0
perc.alumni	64.0
S.F.Ratio	39.8

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 pre-processing while using machine learning algorithms.

In this method, we convert variables with different scales of measurements into a single scale. StandardScaler normalizes the data using the formula (x-mean)/standard deviation.

Note: *Now that we are done with EDA, separate variables showing expenses made by students - 'Room.Board', 'Personal' and 'Books' have been removed and only retained 'TotalExpns' variable which has all the expenses by students combined.*

We perform scaling only for the 15 numerical variables. Below is the sample of our dataset after scaling:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate	TotalExpns
0	-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.162859	-0.115729	1.013776	-0.867574	-0.501910	-0.317993	-0.244850
1	-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496	-2.676529	-3.378176	-0.477704	-0.544572	0.166110	-0.551805	2.018095
2	-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.497090	0.201305	-1.205112	-0.931341	-0.300749	0.585935	-0.177290	-0.668710	-0.767385
3	-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633	1.185939	1.175657	-1.615274	1.151188	1.792851	-0.376446	0.434033
4	-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005	-0.716508	0.204995	-0.523535	-0.553542	-1.675079	0.241803	-2.948375	0.141908

(777, 15)

Table 10: Scaled Dataset Sample

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

- Covariance indicates the direction of the linear relationship between variables. It gets affected when the scales of data are changed.
- Correlation on the other hand measures both the strength and direction of the linear relationship between two variables.

Covariance Matrix:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate	TotalExpns
Apps	1.00	0.94	0.85	0.34	0.35	0.82	0.40	0.05	0.39	0.37	0.10	-0.09	0.26	0.15	0.27
Accept	0.94	1.00	0.91	0.19	0.25	0.88	0.44	-0.03	0.36	0.34	0.18	-0.16	0.12	0.07	0.21
Enroll	0.85	0.91	1.00	0.18	0.23	0.97	0.51	-0.16	0.33	0.31	0.24	-0.18	0.06	-0.02	0.14
Top10perc	0.34	0.19	0.18	1.00	0.89	0.14	-0.11	0.56	0.53	0.49	-0.39	0.46	0.66	0.50	0.30
Top25perc	0.35	0.25	0.23	0.89	1.00	0.20	-0.05	0.49	0.55	0.53	-0.30	0.42	0.53	0.48	0.27
F.Undergrad	0.82	0.88	0.97	0.14	0.20	1.00	0.57	-0.22	0.32	0.30	0.28	-0.23	0.02	-0.08	0.13
P.Undergrad	0.40	0.44	0.51	-0.11	-0.05	0.57	1.00	-0.25	0.15	0.14	0.23	-0.28	-0.08	-0.26	0.13
Outstate	0.05	-0.03	-0.16	0.56	0.49	-0.22	-0.25	1.00	0.38	0.41	-0.56	0.57	0.67	0.57	0.43
PhD	0.39	0.36	0.33	0.53	0.55	0.32	0.15	0.38	1.00	0.85	-0.13	0.25	0.43	0.31	0.30
Terminal	0.37	0.34	0.31	0.49	0.53	0.30	0.14	0.41	0.85	1.00	-0.16	0.27	0.44	0.29	0.33
S.F.Ratio	0.10	0.18	0.24	-0.39	-0.30	0.28	0.23	-0.56	-0.13	-0.16	1.00	-0.40	-0.58	-0.31	-0.26
perc.alumni	-0.09	-0.16	-0.18	0.46	0.42	-0.23	-0.28	0.57	0.25	0.27	-0.40	1.00	0.42	0.49	0.08
Expend	0.26	0.12	0.06	0.66	0.53	0.02	-0.08	0.67	0.43	0.44	-0.58	0.42	1.00	0.39	0.41
Grad.Rate	0.15	0.07	-0.02	0.50	0.48	-0.08	-0.26	0.57	0.31	0.29	-0.31	0.49	0.39	1.00	0.23
TotalExpns	0.27	0.21	0.14	0.30	0.27	0.13	0.13	0.43	0.30	0.33	-0.26	0.08	0.41	0.23	1.00

Table 11: Covariance Matrix

Correlation Matrix:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate	TotalExpns
Apps	1.00	0.94	0.85	0.34	0.35	0.81	0.40	0.05	0.39	0.37	0.10	-0.09	0.26	0.15	0.27
Accept	0.94	1.00	0.91	0.19	0.25	0.87	0.44	-0.03	0.36	0.34	0.18	-0.16	0.12	0.07	0.21
Enroll	0.85	0.91	1.00	0.18	0.23	0.96	0.51	-0.16	0.33	0.31	0.24	-0.18	0.06	-0.02	0.14
Top10perc	0.34	0.19	0.18	1.00	0.89	0.14	-0.11	0.56	0.53	0.49	-0.38	0.46	0.66	0.50	0.30
Top25perc	0.35	0.25	0.23	0.89	1.00	0.20	-0.05	0.49	0.55	0.52	-0.29	0.42	0.53	0.48	0.27
F.Undergrad	0.81	0.87	0.96	0.14	0.20	1.00	0.57	-0.22	0.32	0.30	0.28	-0.23	0.02	-0.08	0.13
P.Undergrad	0.40	0.44	0.51	-0.11	-0.05	0.57	1.00	-0.25	0.15	0.14	0.23	-0.28	-0.08	-0.26	0.13
Outstate	0.05	-0.03	-0.16	0.56	0.49	-0.22	-0.25	1.00	0.38	0.41	-0.55	0.57	0.67	0.57	0.43
PhD	0.39	0.36	0.33	0.53	0.55	0.32	0.15	0.38	1.00	0.85	-0.13	0.25	0.43	0.31	0.29
Terminal	0.37	0.34	0.31	0.49	0.52	0.30	0.14	0.41	0.85	1.00	-0.16	0.27	0.44	0.29	0.33
S.F.Ratio	0.10	0.18	0.24	-0.38	-0.29	0.28	0.23	-0.55	-0.13	-0.16	1.00	-0.40	-0.58	-0.31	-0.26
perc.alumni	-0.09	-0.16	-0.18	0.46	0.42	-0.23	-0.28	0.57	0.25	0.27	-0.40	1.00	0.42	0.49	0.08
Expend	0.26	0.12	0.06	0.66	0.53	0.02	-0.08	0.67	0.43	0.44	-0.58	0.42	1.00	0.39	0.41
Grad.Rate	0.15	0.07	-0.02	0.50	0.48	-0.08	-0.26	0.57	0.31	0.29	-0.31	0.49	0.39	1.00	0.23
TotalExpns	0.27	0.21	0.14	0.30	0.27	0.13	0.13	0.43	0.29	0.33	-0.26	0.08	0.41	0.23	1.00

Table 12: Correlation Matrix

- As we can see that there is no difference in both, the covariance and the correlation matrices, since the covariance matrix is built on the scaled data.
- Scaling ensures that attribute means are all 0 and variances 1.

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

Outliers before scaling:

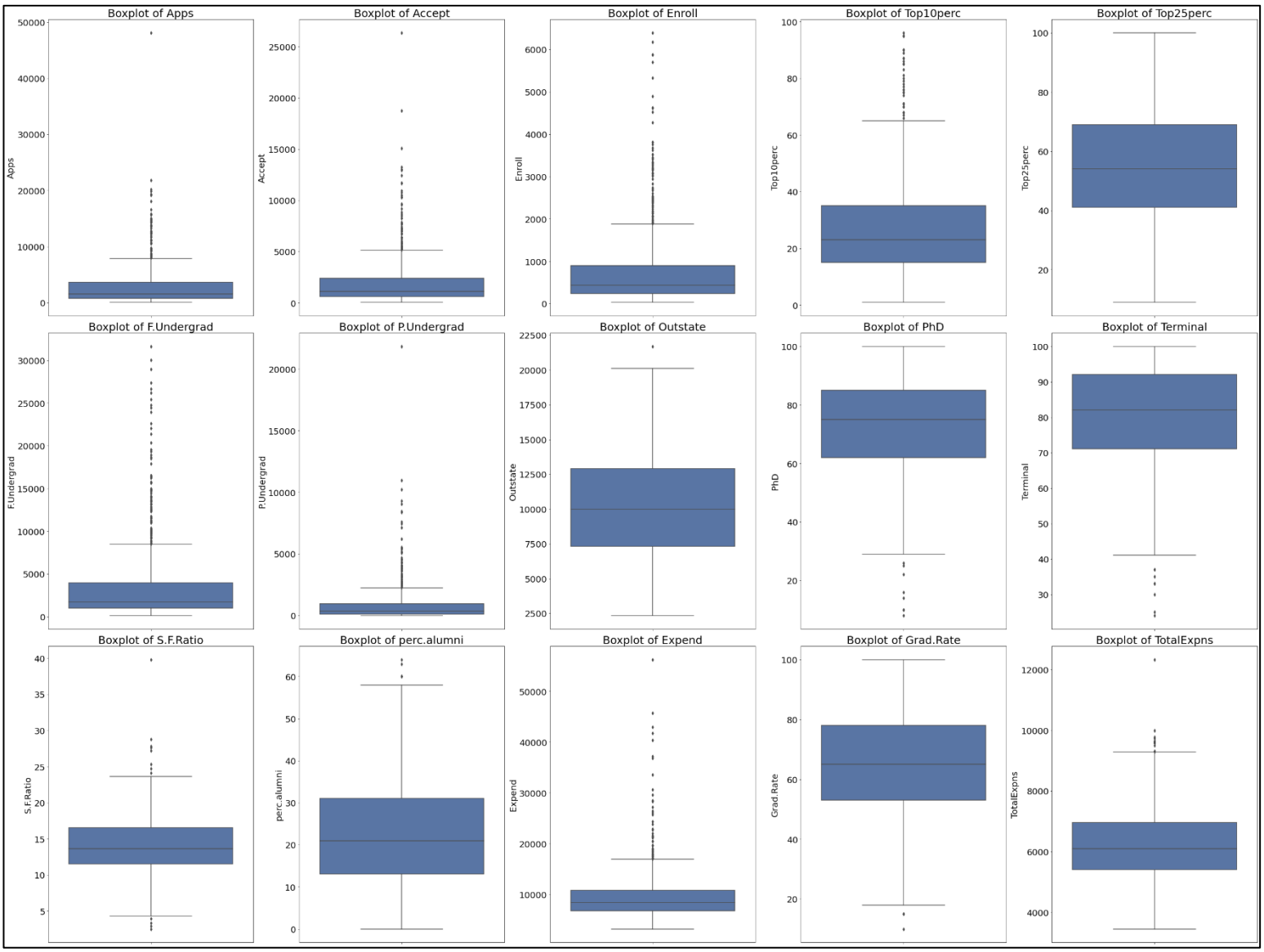


Figure 8: Before Scaling Boxplot

Outliers after scaling:

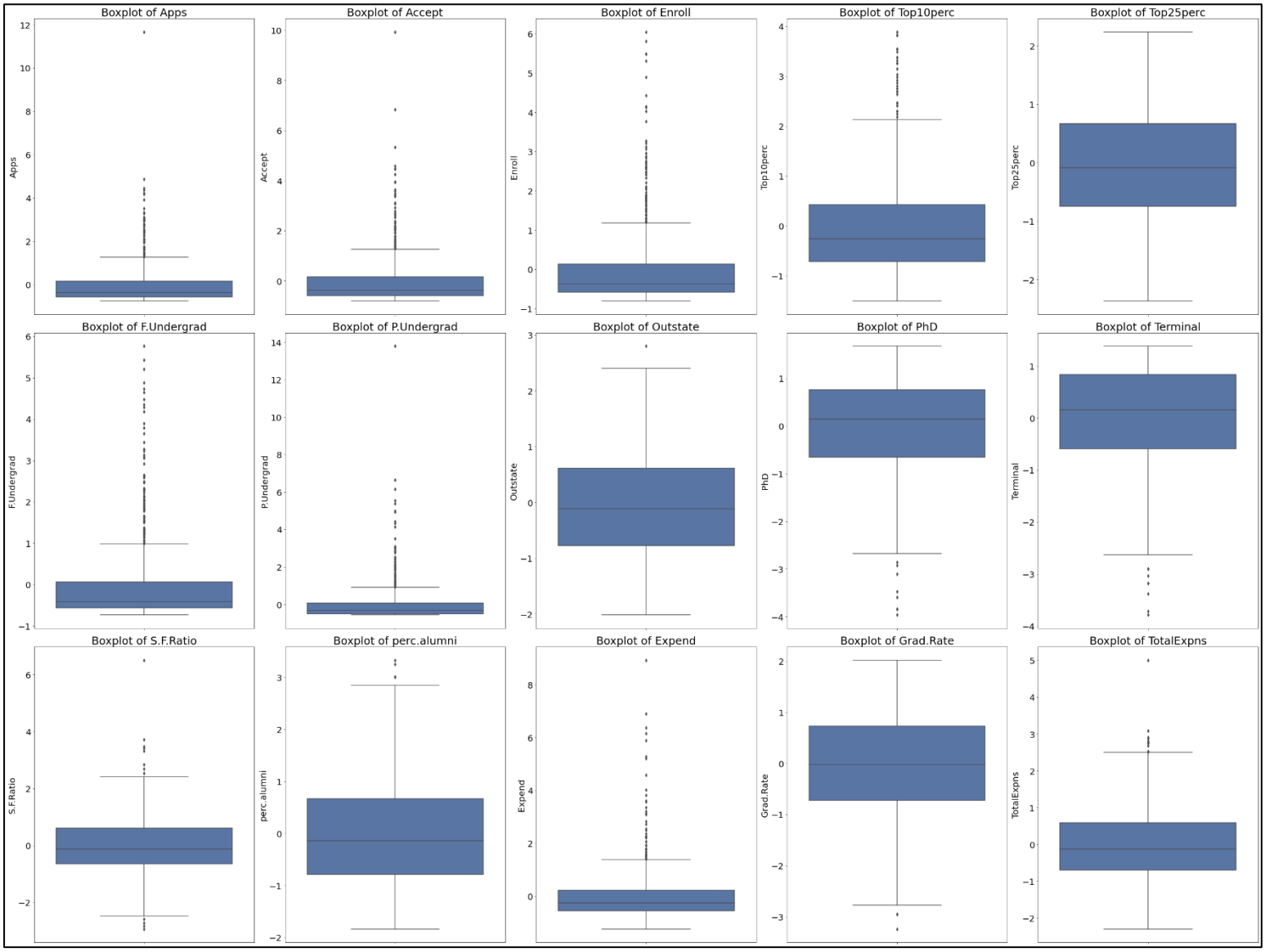


Figure 9: After Scaling Boxplot

We can tell from the above boxplots representing data before and after scaling:

1. There is no difference in the visualization of non-scaled and scaled data. The proportions of the data remain the same after scaling.
2. The outliers position also remains the same, as scaling doesn't impact or treat the outliers in the data.
3. The only difference that is observed in both the boxplots is that the scale of y-axis has changed. The boxplots of scaled data have lower scales as compared to the non-scaled data.
4. The mean of scaled data tends to 0 and standard deviation tends to 1.
5. Scaling is a practical approach when there is a requirement of comparing different variables in a dataset.

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

Eigenvectors of all 15 components:

```
[[ 2.89890298e-01  2.53768115e-01  2.28584726e-01  3.44148209e-01
   3.38430069e-01  2.09300605e-01  6.88978151e-02  2.58945625e-01
   3.26303936e-01  3.22068695e-01 -1.46510666e-01  1.72211895e-01
   3.01947999e-01  2.28494656e-01  2.14540846e-01]
 [ 3.03866701e-01  3.53858444e-01  3.85222710e-01 -1.51651829e-01
  -1.10150348e-01  4.01808457e-01  3.17019610e-01 -2.96703144e-01
   6.30714852e-03 -4.39309177e-03  2.89744132e-01 -2.87829759e-01
  -1.92514542e-01 -2.08497955e-01 -3.93336204e-02]
 [ 8.26058829e-03 -2.08299788e-03 -5.78766793e-02 -1.90747290e-01
  -2.76083633e-01 -4.09101317e-02  2.94555482e-01  1.75931731e-01
  -1.31951373e-01 -7.17945383e-02 -3.63587038e-01 -2.54836234e-01
   2.71141927e-01 -2.23905642e-01  6.49020898e-01]
 [ 2.62580503e-01  2.29310096e-01  1.72145461e-01  8.94603090e-02
   2.94642863e-02  1.09809984e-01 -2.24837524e-01  6.76539050e-02
  -5.54042692e-01 -5.72443262e-01 -2.09993390e-01  1.01569044e-01
   1.04095305e-01  2.62106142e-01 -6.45930604e-02]
 [ 5.63297770e-02  1.06713210e-01 -2.87222254e-02 -2.80431585e-01
  -2.26568226e-01 -5.44600036e-02 -2.14115281e-01  1.83500806e-01
   4.24656287e-02  6.91388929e-02  4.11399248e-01  2.52776566e-02
  -2.91815885e-01  5.65979453e-01  4.36112812e-01]
 [-2.49230922e-03  8.98599315e-02  9.50692641e-02 -3.88001230e-01
  -4.21593170e-01  7.16948519e-02  3.06656717e-01  1.50909684e-01
   6.40489126e-02  1.15544249e-01 -2.53983866e-01  5.77401354e-01
   2.68122922e-02  1.05129730e-01 -3.24836590e-01]
 [-1.60098836e-01 -1.99073327e-01 -6.36309728e-02  2.07718402e-01
   2.88936429e-01  4.99315957e-03  6.80935313e-01 -4.45982939e-02
  -2.00423039e-01 -2.23773822e-01  2.48403922e-01  3.10136192e-01
  -1.67726089e-01  1.48826634e-01  1.99891837e-01]
 [ 3.12741308e-02  4.76102140e-02  8.32336896e-02  1.55089096e-02
  -1.04189643e-02  6.45034381e-02 -3.23203967e-01  5.53453153e-02
  -9.76623433e-02 -1.74023801e-02  3.22399110e-01  5.62215243e-01
   4.18536570e-02 -5.97253785e-01  2.99101843e-01]
 [ 8.65260891e-02  4.73269407e-03 -1.03152136e-01  3.07086384e-02
  -1.67450578e-01 -1.19686918e-01  1.57642534e-01  3.95322150e-01
  -3.49486728e-02 -1.02877858e-01  5.55472573e-01 -1.68649137e-01
   5.75223882e-01 -1.66134340e-02 -2.78309062e-01]
 [-3.62411462e-02 -1.87154135e-01  4.14880361e-02  2.11614927e-02
  -2.04655323e-01  7.87664773e-02 -7.62436381e-02 -7.13966555e-01
   7.10296996e-02 -2.48871337e-03  8.71964037e-02  1.46539378e-01
   5.23321651e-01  2.60813783e-01  1.47728659e-01]
 [ 5.96922489e-01  2.75696221e-01 -4.41479395e-01  1.62509487e-03
   3.20079503e-02 -5.10076502e-01  1.22884371e-01 -2.51563595e-01
   8.76625637e-02 -2.14858470e-02 -2.46651476e-02  1.21594566e-01
  -6.88746404e-02 -6.99436568e-02  2.29687338e-03]
 [ 4.47008052e-02  1.15117293e-02 -5.65486653e-02 -1.05491217e-01
   1.42838109e-01 -2.18811939e-02  1.71747071e-02 -6.43469336e-02
  -6.97834527e-01  6.77338018e-01  3.51364337e-02 -2.80273671e-02
   8.12417995e-02  4.26117556e-02 -4.50582720e-02]
 [ 1.34223588e-01 -1.43260961e-01  2.85850086e-02  6.97370970e-01
  -6.18892297e-01  7.02673444e-03  2.15398222e-02  4.37867801e-02
  -1.07735814e-01  1.56007292e-01 -2.08930328e-02 -8.53153491e-03
  -2.28865874e-01 -2.69649551e-03 -4.14855555e-03]
 [ 4.58749011e-01 -5.19084568e-01 -4.03968380e-01 -1.47364801e-01
   5.21403914e-02  5.61584123e-01 -5.30769957e-02  9.91173198e-02
   2.79629704e-02 -2.60519137e-02 -2.12243327e-02  3.58561660e-03
  -4.39130700e-02 -6.65886465e-03 -2.36707463e-02]
 [ 3.59501725e-01 -5.43393560e-01  6.09145184e-01 -1.45042224e-01
   8.04852069e-02 -4.14858574e-01  8.85896130e-03  5.18927072e-02
   1.36109750e-02  6.55970510e-03 -2.10793291e-03 -1.91012940e-02
  -3.53927037e-02 -1.34034305e-02  3.46933225e-04]]
```

Eigenvalues of all 15 components:

```
[5.37282435 4.20425114 1.06423857 0.98995025 0.71689932 0.62766305
 0.56608727 0.41880236 0.34401279 0.24890415 0.16994101 0.14786614
 0.08810354 0.03676283 0.02302313]
```

For extracting optimal number of principal components, we have to first identify the number of components to be built. For that we create a Scree Plot of explained variance for 15 principal components:

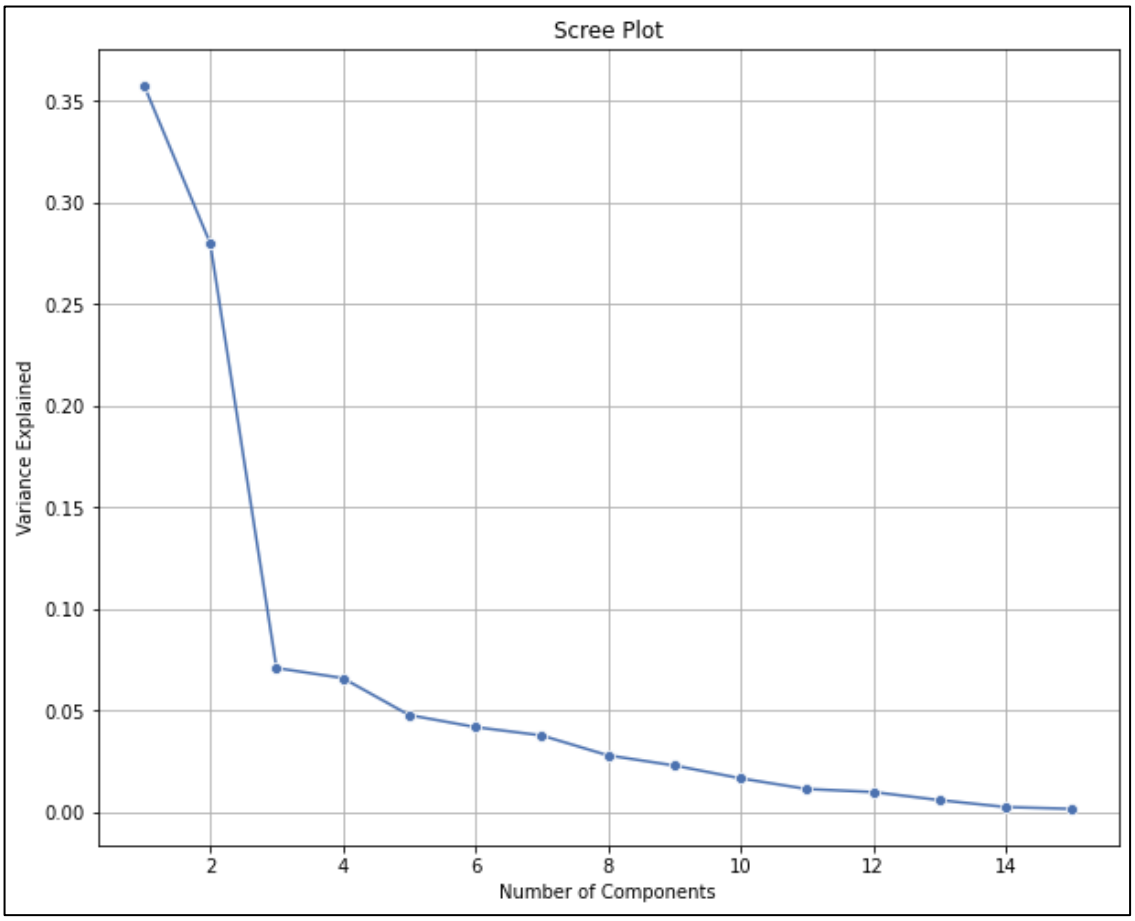


Figure 10: Scree Plot 15 Component

Below is the cumulative explained variance ratio to find a cut off for selecting the number of PCs. The values are in percentage, representing the percentage of explained variance each component contain.

```
[0.3577273  0.63764999 0.70850791 0.77441966 0.82215143 0.86394178
 0.90163237 0.92951659 0.95242126 0.96899351 0.98030833 0.99015339
 0.9960194  0.9984671  1.        ]
```

From the scree plot and cumulative explained variance ratio we decide to go with 7 principal components because these components explain 90% of the variance.

Now we extract the eigenvectors and eigenvalues of these 7 principal components.

Eigenvectors:

```
[[ 0.2898903   0.25376811  0.22858473  0.34414821  0.33843007  0.20930061
   0.06889782  0.25894563  0.32630394  0.3220687  -0.14651067  0.1722119
   0.301948    0.22849466  0.21454085]
 [ 0.3038667   0.35385844  0.38522271 -0.15165183 -0.11015035  0.40180846
   0.31701961 -0.29670314  0.00630715 -0.00439309  0.28974413 -0.28782976
  -0.19251454 -0.20849796 -0.03933362]
 [ 0.00826059 -0.002083   -0.05787668 -0.19074729 -0.27608363 -0.04091013
   0.29455548  0.17593173 -0.13195137 -0.07179454 -0.36358704 -0.25483623
   0.27114193 -0.22390564  0.6490209 ]
 [ 0.2625805   0.2293101   0.17214546  0.08946031  0.02946429  0.10980998
  -0.22483752  0.06765391 -0.55404269 -0.57244326 -0.20999339  0.10156904
   0.1040953   0.26210614 -0.06459306]
 [ 0.05632978  0.10671321 -0.02872223 -0.28043159 -0.22656823 -0.05446
  -0.21411528  0.18350081  0.04246563  0.06913889  0.41139925  0.02527766
  -0.29181589  0.56597945  0.43611281]
 [-0.00249231  0.08985993  0.09506926 -0.38800123 -0.42159317  0.07169485
   0.30665672  0.15090968  0.06404891  0.11554425 -0.25398387  0.57740135
   0.02681229  0.10512973 -0.32483659]
 [-0.16009884 -0.19907333 -0.06363097  0.2077184   0.28893643  0.00499316
   0.68093531 -0.04459829 -0.20042304 -0.22377382  0.24840392  0.31013619
  -0.16772609  0.14882663  0.19989184]]
```

Eigenvalues:

```
[5.37282435 4.20425114 1.06423857 0.98995025 0.71689932 0.62766305
 0.56608727]
```

Scree plot of 7 PC:

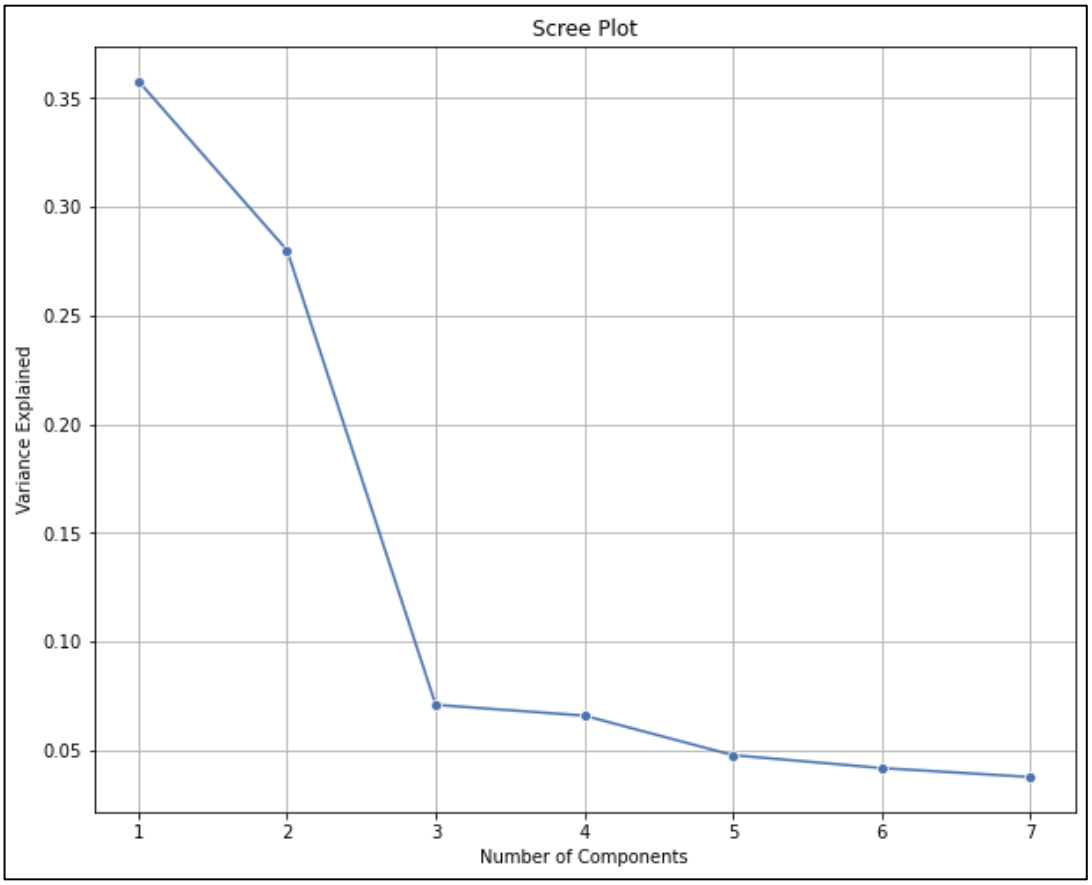


Figure 11: Scree Plot 7 Component

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

Dataframe with 15 Principal Components:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate	TotalExpns
PC1	0.289890	0.253768	0.228585	0.344148	0.338430	0.209301	0.068898	0.258946	0.326304	0.322069	-0.146511	0.172212	0.301948	0.228495	0.214541
PC2	0.303867	0.353858	0.385223	-0.151652	-0.110150	0.401808	0.317020	-0.296703	0.006307	-0.004393	0.289744	-0.287830	-0.192515	-0.208498	-0.039334
PC3	0.008261	-0.002083	-0.057877	-0.190747	-0.276084	-0.040910	0.294555	0.175932	-0.131951	-0.071795	-0.363587	-0.254836	0.271142	-0.223906	0.649021
PC4	0.262581	0.229310	0.172145	0.089460	0.029464	0.109810	-0.224838	0.067654	-0.554043	-0.572443	-0.209993	0.101569	0.104095	0.262106	-0.064593
PC5	0.056330	0.106713	-0.028722	-0.280432	-0.226568	-0.054460	-0.214115	0.183501	0.042466	0.069139	0.411399	0.025278	-0.291816	0.565979	0.436113
PC6	-0.002492	0.089860	0.095069	-0.388001	-0.421593	0.071695	0.306657	0.150910	0.064049	0.115544	-0.253984	0.577401	0.026812	0.105130	-0.324837
PC7	-0.160099	-0.199073	-0.063631	0.207718	0.288936	0.004993	0.680935	-0.044598	-0.200423	-0.223774	0.248404	0.310136	-0.167726	0.148827	0.199892
PC8	0.031274	0.047610	0.083234	0.015509	-0.010419	0.064503	-0.323204	0.055345	-0.097662	-0.017402	0.322399	0.562215	0.041854	-0.597254	0.299102
PC9	0.086526	0.004733	-0.103152	0.030709	-0.167451	-0.119687	0.157643	0.395322	-0.034949	-0.102878	0.555473	-0.168649	0.575224	-0.016613	-0.278309
PC10	-0.036241	-0.187154	0.041488	0.021161	-0.204655	0.078766	-0.076244	-0.713967	0.071030	-0.002489	0.087196	0.146539	0.523322	0.260814	0.147729
PC11	0.596922	0.275696	-0.441479	0.001625	0.032008	-0.510077	0.122884	-0.251564	0.087663	-0.021486	-0.024665	0.121595	-0.068875	-0.069944	0.002297
PC12	0.044701	0.011512	-0.056549	-0.105491	0.142838	-0.021881	0.017175	-0.064347	-0.697835	0.677338	0.035136	-0.028027	0.081242	0.042612	-0.045058
PC13	0.134224	-0.143261	0.028585	0.697371	-0.618892	0.007027	0.021540	0.043787	-0.107736	0.156007	-0.020893	-0.008532	-0.228866	-0.002696	-0.004149
PC14	0.458749	-0.519085	-0.403968	-0.147365	0.052140	0.561584	-0.053077	0.099117	0.027963	-0.026052	-0.021224	0.003586	-0.043913	-0.006659	-0.023671
PC15	0.359502	-0.543394	0.609145	-0.145042	0.080485	-0.414859	0.008859	0.051893	0.013611	0.006560	-0.002108	-0.019101	-0.035393	-0.013403	0.000347

Table 13: Dataframe with 15 Principal Components

Dataframe with 7 Principal Components:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate	TotalExpns
PC1	0.289890	0.253768	0.228585	0.344148	0.338430	0.209301	0.068898	0.258946	0.326304	0.322069	-0.146511	0.172212	0.301948	0.228495	0.214541
PC2	0.303867	0.353858	0.385223	-0.151652	-0.110150	0.401808	0.317020	-0.296703	0.006307	-0.004393	0.289744	-0.287830	-0.192515	-0.208498	-0.039334
PC3	0.008261	-0.002083	-0.057877	-0.190747	-0.276084	-0.040910	0.294555	0.175932	-0.131951	-0.071795	-0.363587	-0.254836	0.271142	-0.223906	0.649021
PC4	0.262581	0.229310	0.172145	0.089460	0.029464	0.109810	-0.224838	0.067654	-0.554043	-0.572443	-0.209993	0.101569	0.104095	0.262106	-0.064593
PC5	0.056330	0.106713	-0.028722	-0.280432	-0.226568	-0.054460	-0.214115	0.183501	0.042466	0.069139	0.411399	0.025278	-0.291816	0.565979	0.436113
PC6	-0.002492	0.089860	0.095069	-0.388001	-0.421593	0.071695	0.306657	0.150910	0.064049	0.115544	-0.253984	0.577401	0.026812	0.105130	-0.324837
PC7	-0.160099	-0.199073	-0.063631	0.207718	0.288936	0.004993	0.680935	-0.044598	-0.200423	-0.223774	0.248404	0.310136	-0.167726	0.148827	0.199892

Table 14: Dataframe with 7 Principal Components

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]

PCA components (Eigenvectors) of first PC:

```
Apps           0.29
Accept         0.25
Enroll         0.23
Top10perc      0.34
Top25perc      0.34
F.Undergrad    0.21
P.Undergrad    0.07
Outstate       0.26
PhD            0.33
Terminal       0.32
S.F.Ratio     -0.15
perc.alumni    0.17
Expend         0.30
Grad.Rate      0.23
TotalExpns     0.21
Name: PC1, dtype: float64
```

Linear equation of first PC:

(0.29 * Apps) + (0.25 * Accept) + (0.23 * Enroll) + (0.34 * Top10perc) + (0.34 * Top25perc) + (0.21 * F.Undergrad) + (0.07 * P.Undergrad) + (0.26 * Outstate) + (0.33 * PhD) + (0.32 * Terminal) + (-0.15 * S.F.Ratio) + (0.17 * perc.alumni) + (0.3 * Expend) + (0.23 * Grad.Rate) + (0.21 * TotalExpns)

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?

Below is the cumulative explained variance ratio of all 15 components:

```
[0.3577273  0.63764999 0.70850791 0.77441966 0.82215143 0.86394178
 0.90163237 0.92951659 0.95242126 0.96899351 0.98030833 0.99015339
 0.9960194  0.9984671  1.        ]
```

The values can be interpreted in percentage form, representing the percentage of explained variance (eigenvalues) each component contains. The optimal number of principal components should account for a significant percentage of explained variance. From the above cumulative values, we can say that top 7 components explain 90% of the total variance. Hence it is sufficient to use the first 7 PCs instead of the original 15 variables, thereby reducing the dimensions by almost half.

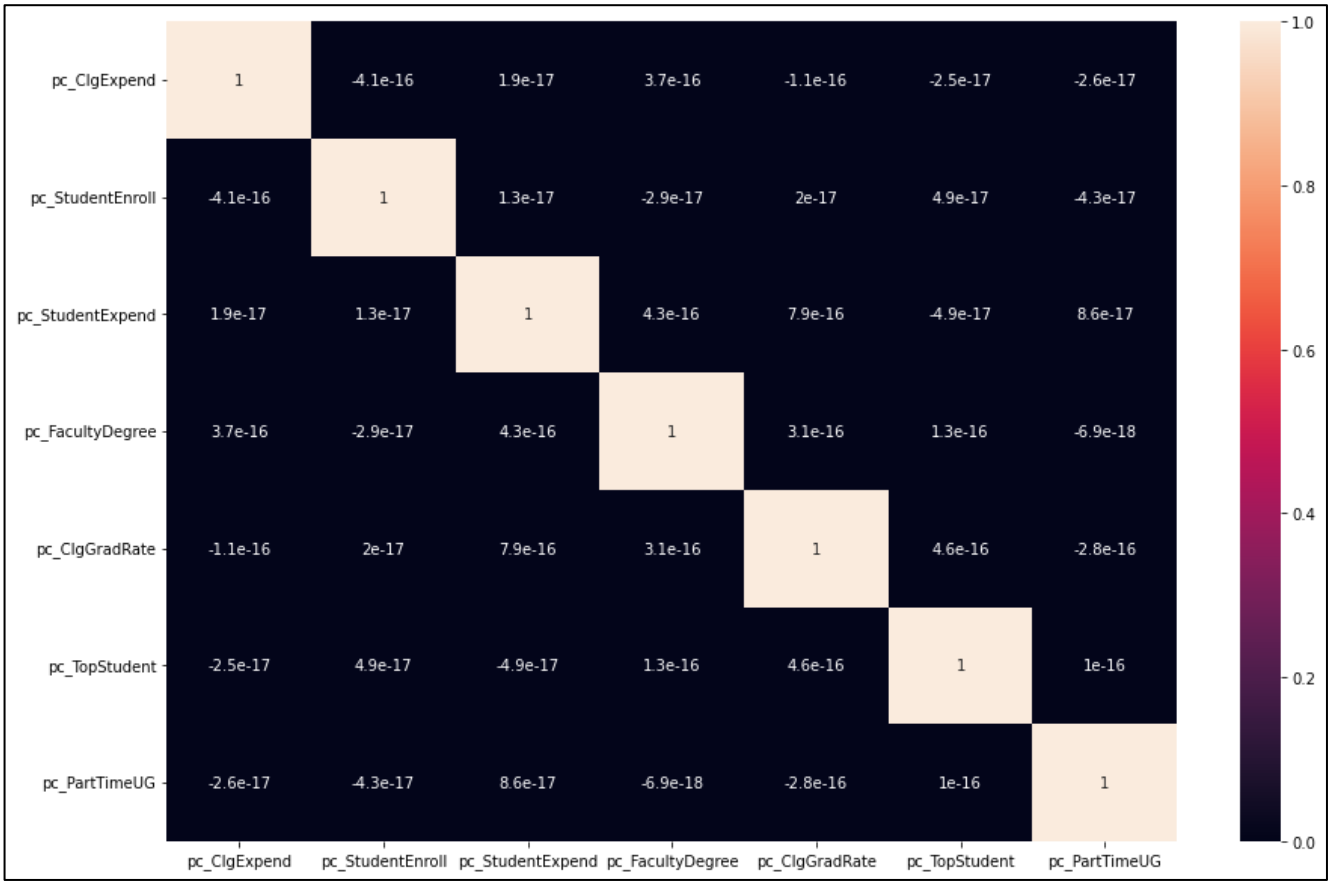
Eigenvectors are the new dimensions of the new feature space in PCA. These variables are basically multiplication of a vector with a metrics that changes the basis of the vector and also its direction in a linear transformation.

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]

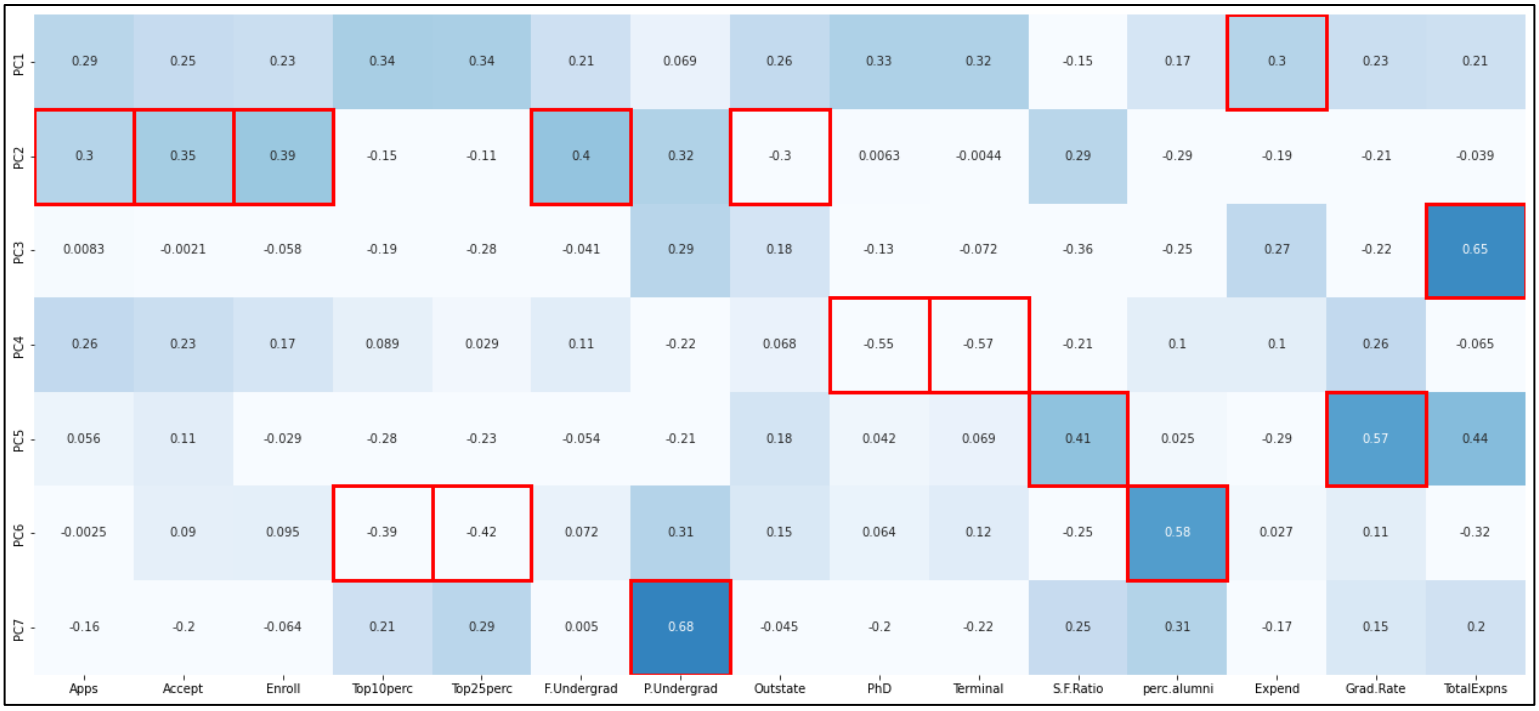
Principal Component Analysis (PCA) is useful when we have a large number of variables at hand, and it is difficult to interpret the data and it is hard to decide which variables to focus on. PCA is used to reduce the dimensions of the feature space, also called as "dimensionality reduction".

As we observed with the help of scree plot and cumulative explained variance ratio that the top 7 PCs can help in the further analysis of the data, since they explain 90% of the total variance. Using PCA we were able to reduce our dimensions by almost 50%, from 15 to 7 components.

Another use of PCA is to reduce multicollinearity of the variables. In the below heatmap, we can see that the correlation between components is very close to zero.



Here is a representation of what each PC explains:



- PC1 – Expenses made by colleges
- PC2 – Application received, accepted, number of students enrolled, full-time undergraduates and outstation students
- PC3 – Expenses made by students on boarding, books and personal usage
- PC4 – Qualification of faculties: PhD and Terminal degree
- PC5 – Student to faculty ratio and graduation rate of colleges
- PC6 – Top 10% and 25% of new students and percentage of alumni who donate
- PC7 – Part time undergraduate students

For further interpretation and business recommendations, we have renamed the PCs based on the variables they explain and combined them with the categorical variable – "Names" (college & university names). Here is the sample of the data:

	Names	pc_CIgExpend	pc_StudentEnroll	pc_StudentExpend	pc_FacultyDegree	pc_CIgGradRate	pc_TopStudent	pc_PartTimeUG
0	Abilene Christian University	-1.258454	0.619873	-0.423799	-0.490679	0.220902	-0.791249	-0.066417
1	Adelphi University	-2.357501	-0.062981	3.061478	2.944544	0.487783	-0.540248	0.975597
2	Adrian College	-1.546697	-0.854596	-0.132903	0.932095	-0.607153	0.410228	0.003925
3	Agnes Scott College	2.413697	-2.871099	0.110710	-0.963727	-1.775058	-0.504754	-0.047065
4	Alaska Pacific University	-2.223598	0.190453	1.692611	-1.329853	-1.850752	-1.015386	-0.987657

Table 15: Principal Components Dataset Sample

below is the list of top colleges, which spend high on instructional expenditure per student, have highly qualified faculties and high graduation rate:

	Names
274	Indiana University at Bloomington
279	James Madison University
365	Miami University at Oxford
432	Ohio University
483	Rutgers at New Brunswick
576	Syracuse University
614	University of Delaware
623	University of Illinois - Urbana
634	University of Massachusetts at Amherst
677	University of Southern California
713	Virginia Tech

Table 16: Top Colleges

Students can consider opting for these colleges.

Below is the list of colleges with low performance, which spend low on instructional expenditure per student, have low % of highly qualified faculties and low graduation rate:

	Names
4	Alaska Pacific University
142	Columbia College MO
146	Concordia Lutheran College
170	Dowling College
197	Fayetteville State University
247	Hardin-Simmons University
303	Lamar University
444	Pembroke State University
697	University of Wisconsin-Superior

Table 17: Bottom Colleges

These colleges should reconsider their strategy by improving on factors basis on which they have been in the bottom of the list.