Great Learning

Advanced Statistics Project - Business Report

Submitted by:

N. Aishwarya
PGP-DSBA

March 2022

Table of Contents

List of Figures

List of Tables

List of Formulas

Executive summary - Conceptual understanding of ANOVA:

This business report solves the above problem using Analysis of Variance (ANOVA) technique, which belongs to the domain called “Experimental Designs”.

ANOVA helps in establishing in an exact way, the Cause- Effect relation between variables. From the statistical inference point of view, ANOVA is an extension of independent t test for testing the equality of two population means. When more than two population means have to be compared, ANOVA technique is used. In this case, the null hypothesis (H_0) is defined as

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \ldots\ldots = \mu_k$ for testing the equality of population means for k populations where μ denotes the mean of the population.

The Alternative hypothesis(H_1) is defined as,

H_1: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \ldots\ldots \neq \mu_k$ for testing the inequality of population means for k populations where μ denotes the mean of the population.

Business problem 1A

Problem Statement:

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals are collected and each person’s educational qualification and occupation are noted.

Educational qualification is at three levels, High school graduate, Bachelor, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education - occupation combination.

Solution Approach:

The purpose of this whole exercise is to explore the dataset using one-way classification and two-way classification. The data consists of 40 different person’s salary based on their education and occupation. There are 3 educational levels and 4 Occupation levels.

This assignment will help in exploring the summary of dataset and analyzing whether salary is dependent on Education and Occupation using One-way Anova and Two-way Anova concepts. In this work, an analysis of salary data has been performed and the results and business insights drawn are inferred.

Exploratory Data Analysis:

Salary dataset data is loaded using pandas and the dataset has 40 rows and 3 columns. Description of variables are as below, to understand the data better:

- **Education**: Denotes each person's education qualifications and has the below categories:
  - Doctorate 16 (40% of the data)
  - Bachelors 15 (37.5% of the data)
  - HS-grad 9 (22.5% of the data)

- **Occupation:** Denotes each person's Occupations and has the below categories:
  - Prof-specialty 13 (32.5% of the data)
  - Sales 12 (30% of the data)
  - Adm-clerical 10 (25% of the data)
  - Exec-managerial 5 (12.5% of the data)

- **Salary:** Denotes each person's salary details and is a continuous variable

Sample of the dataset:

The top 10 records of the given dataset are given below:

	Education	Occupation	Salary
0	Doctorate	Adm-clerical	153197
1	Doctorate	Adm-clerical	115945
2	Doctorate	Adm-clerical	175935
3	Doctorate	Adm-clerical	220754
4	Doctorate	Sales	170769
5	Doctorate	Sales	219420
6	Doctorate	Sales	237920
7	Doctorate	Sales	160540
8	Doctorate	Sales	180934
9	Doctorate	Prof-specialty	248156

Dataset has 3 variables (Education, Occupation and Salary) with 3 Educational levels – "Doctorate", "Bachelors" and "HS-grad" and 4 occupation levels – "Prof-specialty ", "Sales", "Adm-clerical", "Exec-managerial" and the salary based on their education and occupation is defined.

Let us check the types of variables and missing values 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)
memory usage: 1.1+ KB
```

There is a total of 40 rows with 3 columns in the dataset. Out of these 3 variables, two variables “Education” and “Occupation” are object datatypes whereas only one variable “Salary” is of integer type. From the above results, we can observe that there is no missing value present in the dataset.

Below is the summary of the data, providing descriptive statistical variables:

Descriptive statistics help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are measures of center: the mean, median, and mode, which are used at almost all levels of math and statistics.

Table 1 - Summary of Salary data

	Education	Occupation	Salary
count	40	40	40.000
unique	3	4	NaN
top	Doctorate	Prof-specialty	NaN
freq	16	13	NaN
mean	NaN	NaN	162186.875
std	NaN	NaN	64860.408

min	NaN	NaN	50103.000
25%	NaN	NaN	99897.500
50%	NaN	NaN	169100.000
75%	NaN	NaN	214440.750
max	NaN	NaN	260151.000

Checking distinct values of Education:

```
 Doctorate     16
 Bachelors     15
 HS-grad        9
Name: Education, dtype: int64
```

Checking distinct values of Occupation:

```
 Prof-specialty      13
 Sales               12
 Adm-clerical        10
 Exec-managerial      5
Name: Occupation, dtype: int64
```

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

Education:

Null Hypothesis H_0: The mean salary is the same across all the 3 categories of education (Doctorate, Bachelors, HS-Grad).

Alternate Hypothesis H_1: The mean salary is different in at least one category of education.

Occupation:

Null Hypothesis H_0: The mean salary is the same across all the 4 categories of occupation (Prof-Specialty, Sales, Adm-clerical, Exec-Managerial).

Alternate Hypothesis H_1: The mean salary is different in at least one category of occupation.

Analysis 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.

Anova analysis is done using the OLS method, and below is the output:

Table 2 - one-way ANOVA for Education with salary

	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

Observation:

For the given problem sum of squares due to the factor Education (SSB) is 1.026955e+11 and the sum of squares due to error (SSW) is 6.137256e+10. The total sum of squares (SST) for the data is (1026955+6137256=7164211). Since the factor has 3 levels, DF corresponding to Education is 3 – 1 = 2. Total DF is 40 –1 = 39. Hence DF due to error is 39 –2 = 37.

Mean sum of squares is obtained by dividing the sums of squares by corresponding DF. The value of the F-statistic is approximately 30.95 or 31 and the p-value is highly significant.

Conclusion:

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 a significant difference in the mean salaries for at least one category of education.

Analysis 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.

Anova analysis is done using the OLS method, and below is the output:

Table 3 - one-way ANOVA for Occupation with salary

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

Observation:

For the given problem sum of squares due to the factor Education (SSB) is 1.125878e+10 and the sum of squares due to error (SSW) is 1.528092e+11. The total sum of squares (SST) for the data is (1125878+1528092=2653970) e+11 approximately. Since the factor has 4 levels, DF

corresponding to Occupation is 4 – 1 = 3. Total degrees of freedom are 40 –1 = 39. Hence degrees of freedom due to error is 39 –3 = 36. Mean sum of squares is obtained by dividing the sums of squares by corresponding DF. The value of the F-statistic is approximately 30.95 or 31 and the p-value is highly significant.

Conclusion:

Since the p value = 0.458508 is greater than the significance level (alpha = 0.05), we fail to reject the null hypothesis (i.e., we accept H0) and conclude that there is no significant difference in the mean salaries across the 4 categories of occupation.

Drawing a Point Plot (For both Education and Occupation with Salary)

Fig.1 – Point Plot for Education with Salary

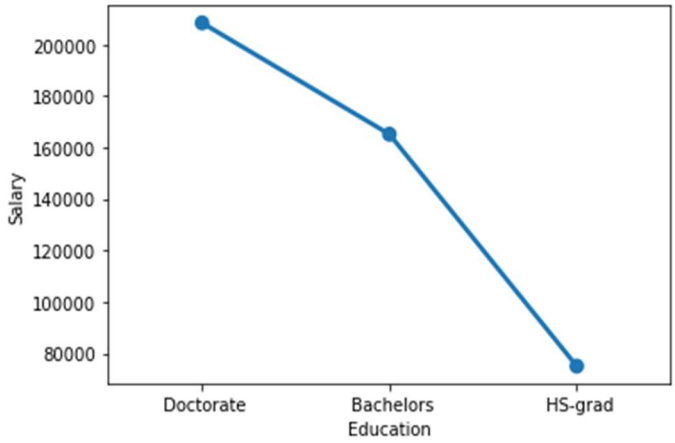
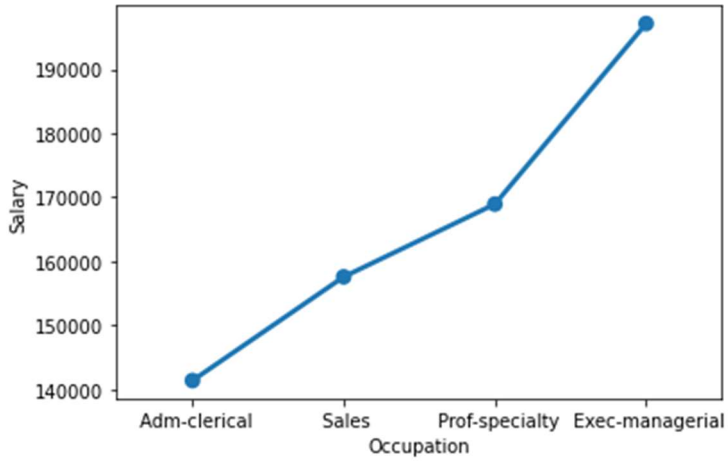


Fig.2 – Point Plot for Occupation with Salary



Conclusion:

We have observed that Education and Occupation individually have a significant impact on Salary as null hypotheses that group means are equal have been rejected in one case and accepted in the other case. However, we have not been able to determine which mean is different from the rest or whether all pairs of means are different. There are special tests (called post hoc tests) of the differences between all pairs of means. These tests are also called multiple comparison tests.

Analysis 1.4: If the null hypothesis is rejected in either (1.2) or in (1.3), find out which class means are significantly different. Interpret the result.

- A one-way ANOVA is used to determine whether or not there is a statistically significant difference between the means of three or more independent groups. If the overall p-value from the ANOVA table is less than some significance level, then we have sufficient evidence to say that at least one of the means of the groups is different from the others.

- However, this doesn't tell us which groups are different from each other. It simply tells us that not all of the group means are equal. In order to find out exactly which groups are different from each other; we must conduct a post hoc test.

One of the most commonly used post hoc tests is Tukey's Test, which allows us to make pairwise comparisons between the means of each group

Multiple comparison tests for Education:

In order to identify for which Education, the salary group means are different from other groups, the hypotheses may be stated as:

H_0: All pairs of group means are equal.

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

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

Conclusion:

The table shows that since the p- values (p-adj in the table) are lesser than the significance level for all the three categories of education, this implies that the mean salaries across all categories of education are different.

Multiple comparison tests for Occupation:

In order to identify for which occupation, the salary group means are different from other groups, the hypotheses may be stated as:

H_0: All pairs of group means are equal.

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

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

Conclusion:

For the category occupation, the Tukey Honest Significant Difference test has further confirmed that the mean salaries across all occupation classes are significantly same. We can observe that all p-values are greater than 0.05.

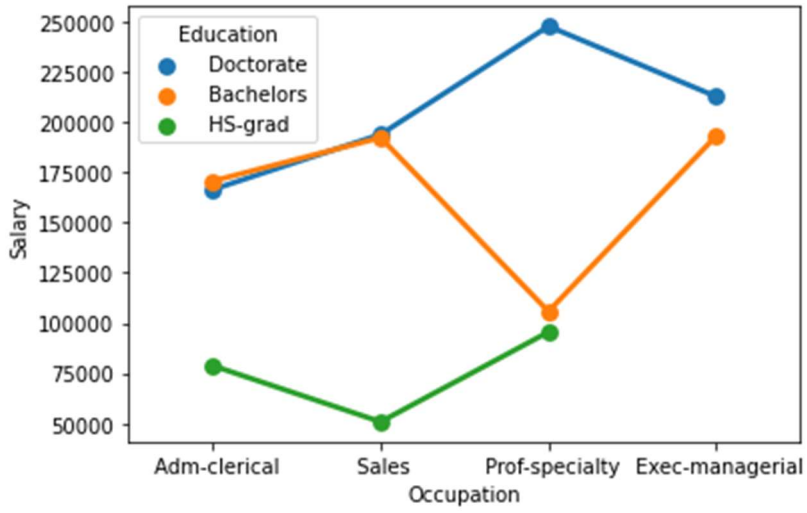
Business problem 1B:

TWO-WAY ANOVA

Analysis 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.

We analyze the effects of one variable on the other (Education and Occupation) with the help of a point plot.

Fig.3 – Interaction plot to analyze the effects of Education on Occupation



The interaction plot shows that there is significant amount of interaction between the categorical variables, Education and Occupation. Below are some of the key 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 170,000 –190,000).
- 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.

- People with **education as Bachelors and occupation Sales earn higher than people with education as Bachelors and occupation Prof-Specialty** whereas people with education as Doctorate and occupation Sales earn lesser than people with Doctorate and occupation Prof-Specialty. We see a reversal in this part of the plot.
- Similarly, people with **education as Bachelors and occupation as Prof-Specialty earn lesser than people with education as Bachelors and occupation Exec-Managerial** whereas people with education as Doctorate and occupation as Prof-Specialty earn higher than people with education as Doctorate and occupation Exec-Managerial. There is a reversal in this part of the plot too.
- **Salespeople with Bachelors or Doctorate education earn the same salaries** and earn higher than people with education as HS-grad.

- **Adm clerical people with education as HS-grad earn the lowest salaries** when compared to people with education as Bachelors or Doctorate.
- **Prof-Specialty people with education as Doctorate earn maximum salaries** and people with education as HS-Grad earn the minimum.
- **People with education as HS -Grad** earn the **minimum salaries**.
- There are **no people with education as HS -grad who hold Exec-managerial occupation**.
- People with **education as Bachelors and occupation, Sales and Exec-Managerial earn the same salaries**.

Analysis 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?

H_0**:** 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.

H_1: There is an interaction effect between the independent variable 'education' and the independent variable 'occupation' on the mean salary.

Table 4 - Two-way ANOVA for Education and Occupation with 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

Conclusion:

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.

Analysis 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.

Hence, we can conclude that Salary is dependent on educational qualifications and occupation.

Business 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.

Data Description:

The given dataset consists of data points of names of various university and college which has number of application received, accepted, and enrolled, percentage of new students from top 10% of higher secondary class, percentage of new students from top 25% of higher secondary class, Number of fulltime undergraduates, Number of parttime undergraduate students, Number of students for whom the particular college is out of state tuition, cost of room and board, estimated book costs for a student, estimated personal spending for a student, percentage of faculties with PHD, percentage of faculties with terminal degree, student/faculty ratio, percentage of alumni who donate, The instructional expenditure per student, Graduation Rate.

Sample of the dataset:

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

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

```
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
```

Inference of the dataset:

- The shape of the dataset is 777 rows and 18 columns.
- All the columns seem to be integer or float values.
- The Names column alone is a categorical value.
- We also can see they are no duplicates in the dataset.
- The entire dataset does not have missing values or null values.

Below is the summary of the data, providing descriptive statistical variables:

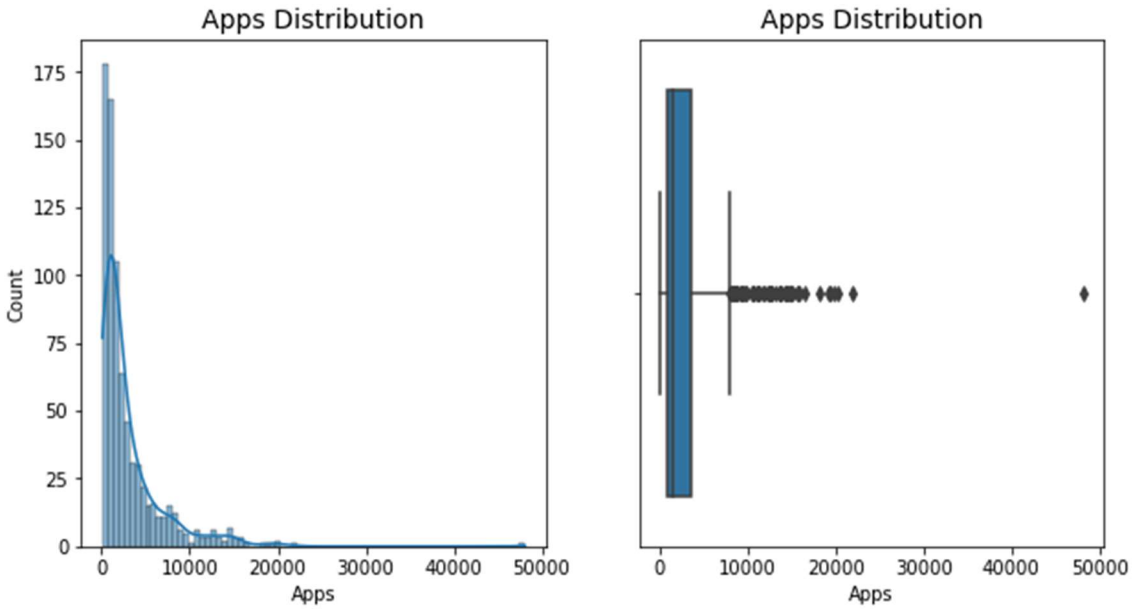
	count	mean	std	min	25%	50%	75%	max
Apps	777.0	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	779.972973	929.176190	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	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

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

Univariate analysis:

Helps us to understand the distribution of data in the dataset. With univariate analysis we can find patterns and we can summarize the data for

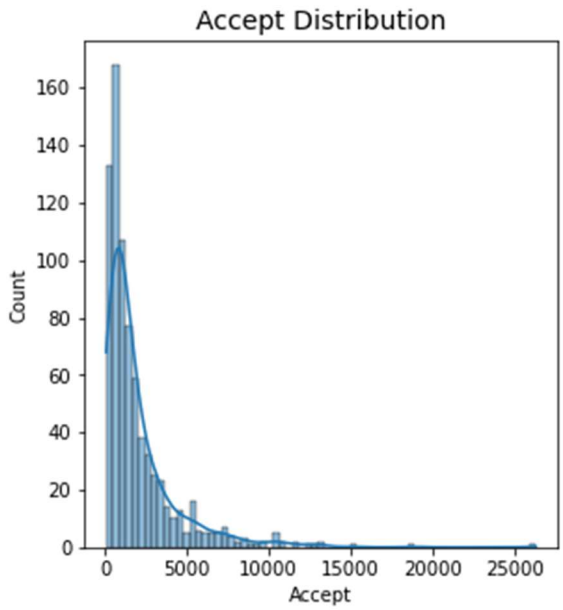
1. Apps Distribution:

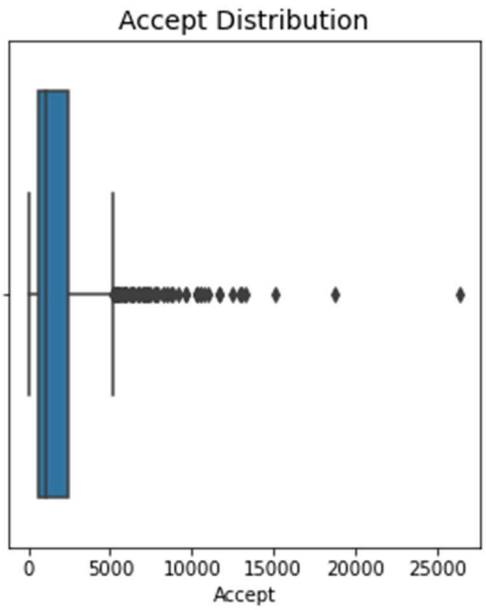


Conclusion:

- For Univariate Analysis of Apps, we are using histplot and boxplot to find information or patterns in the data.
- The Boxplot of Apps variable seems to have outliers. The distribution of the data is also skewed. We also understand that each college or university offers application in the range 3000 to 5000. The maximum application seems to be around 50,000.

2. Accept Distribution:

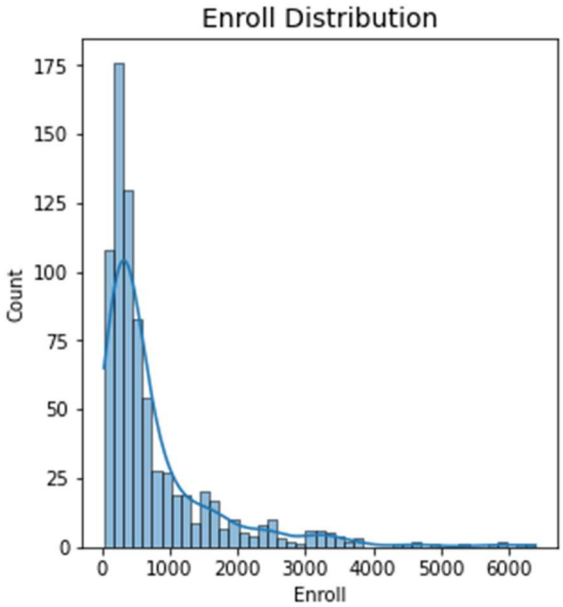


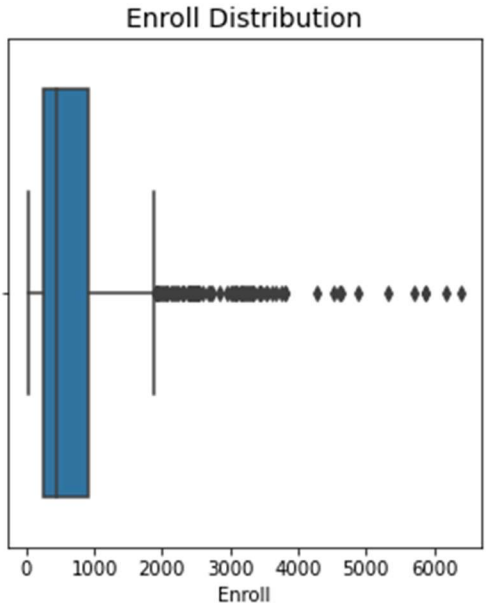


Conclusion:

- The accept variable seems to have outliers.
- The above plot shows that the majority of applications accepted from each university are in the range from 70 to 1500.
- The Accept variable seems to be positively skewed.

3. Enroll Distribution:

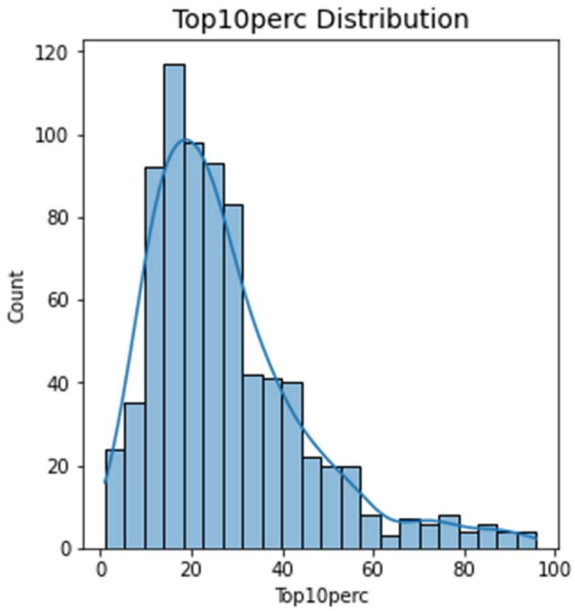


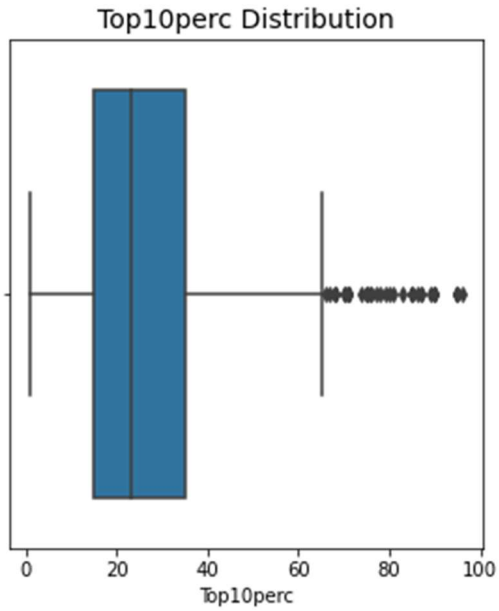


Conclusion:

- The boxplot of the Enroll variable also have outliers.
- The distribution of the data is positively skewed.
- From the histplot we can understand majority of the colleges have enrolled students in the range from 200 to 500 students.

4. Top10perc Distribution:

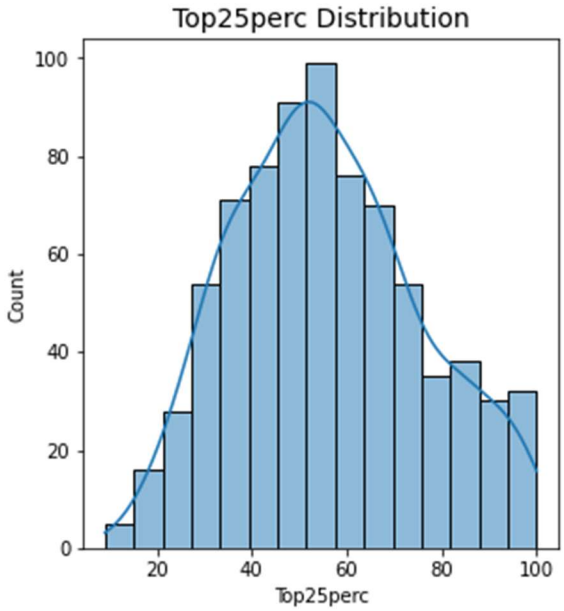


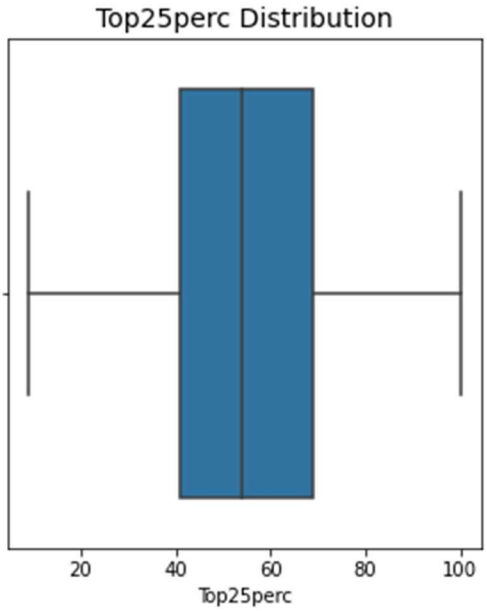


Conclusion:

- The Boxplot of the students from Top 10 percentage of higher secondary class seems to have outliers.
- The distribution seems to be positively skewed.
- There is a good amount of intake about 30 to 50 students from top 10 percentage of higher secondary class.

5. Top25perc Distribution:

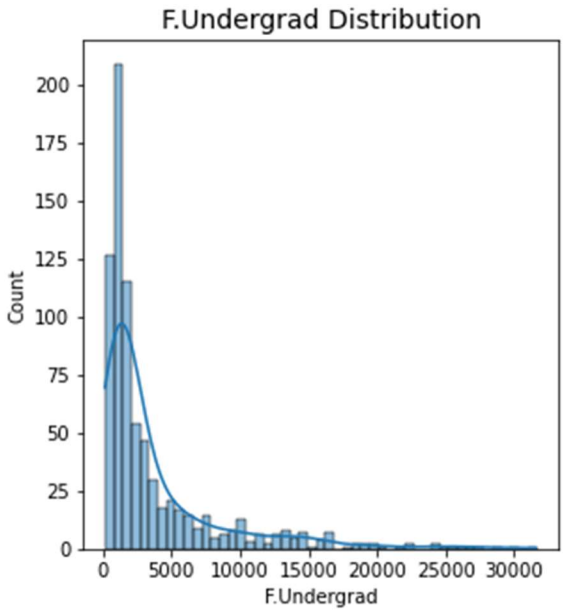


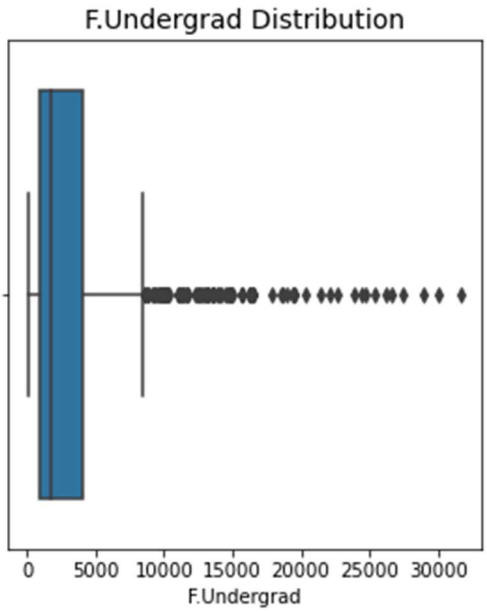


Conclusion:

- The Boxplot for the top 25% has no outliers.
- The distribution is almost normally distributed.
- Majority of the students are from top 25% of higher secondary school.

6. Full Time Undergraduate:

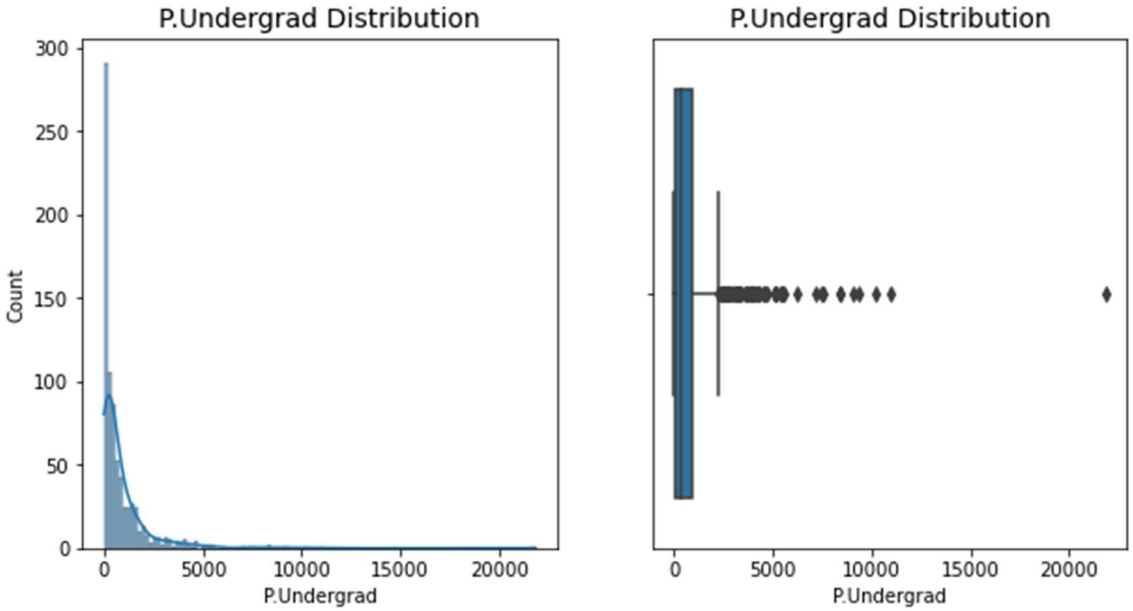




Conclusion:

- The Boxplot of full time Undergraduates have outliers.
- The distribution of the data is positively skewed.
- In the range about 3000 to 5000 there are full time graduates studying in all the university.

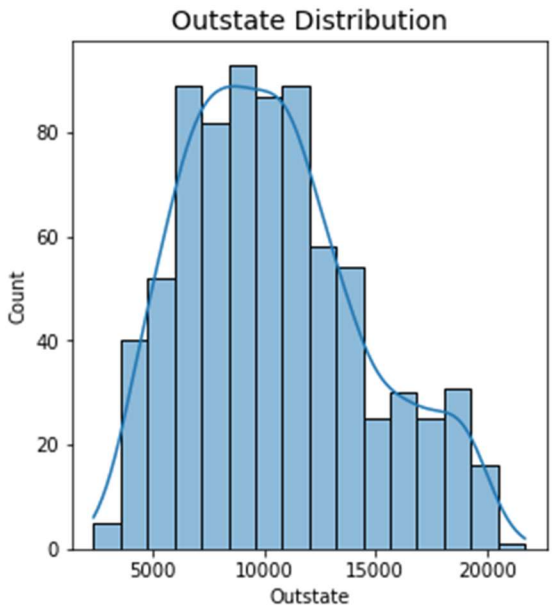
7. Part Time Undergraduate:

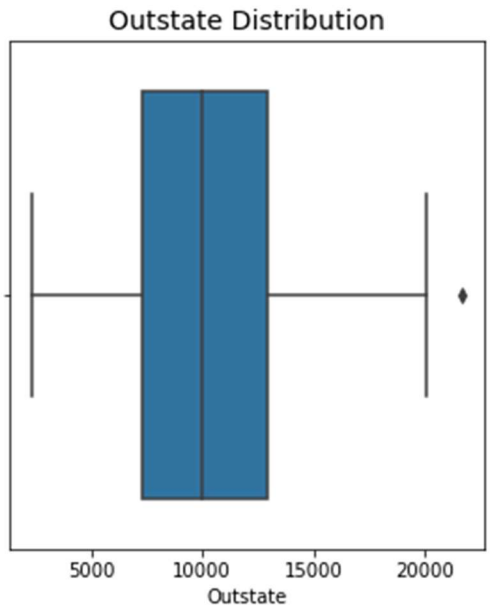


Conclusion:

- The Boxplot of Part time graduates have outliers.
- The distribution of the data is positively skewed.
- In the range about 1000 to 3000 they are part time graduates studying in all the university.

8. OUTSTATE:

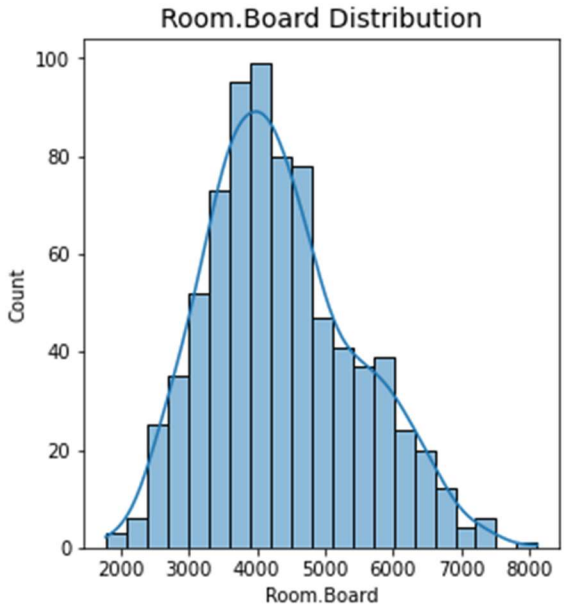


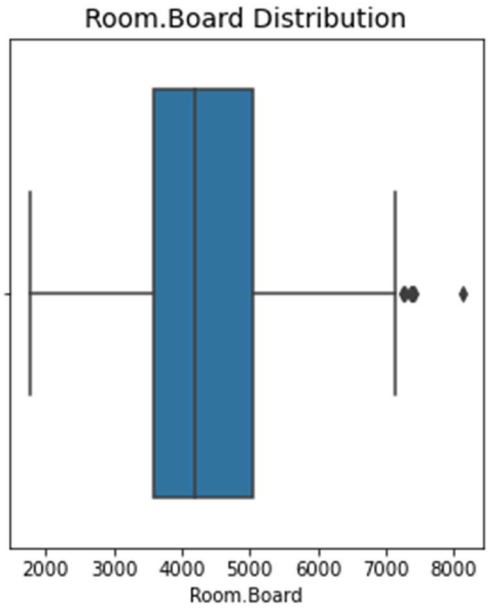


Conclusion:

- The Boxplot of Outstate also has outliers.
- The distribution is almost normally distributed.

9. ROOM BOARD:

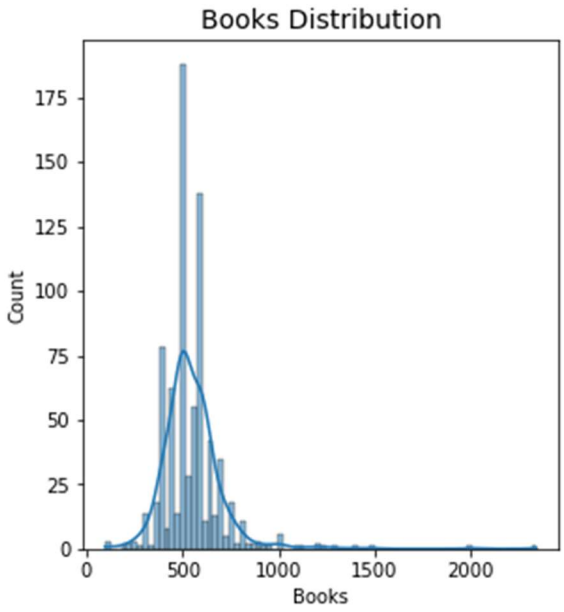


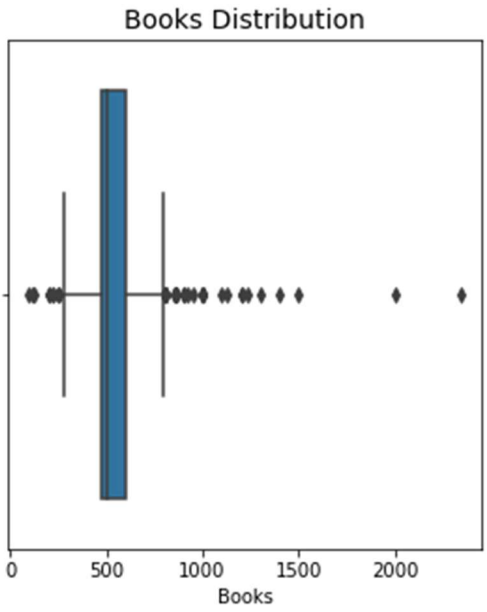


Conclusion:

- The Room Board has very few outliers.
- The distribution is normally distributed.

10. BOOKS:

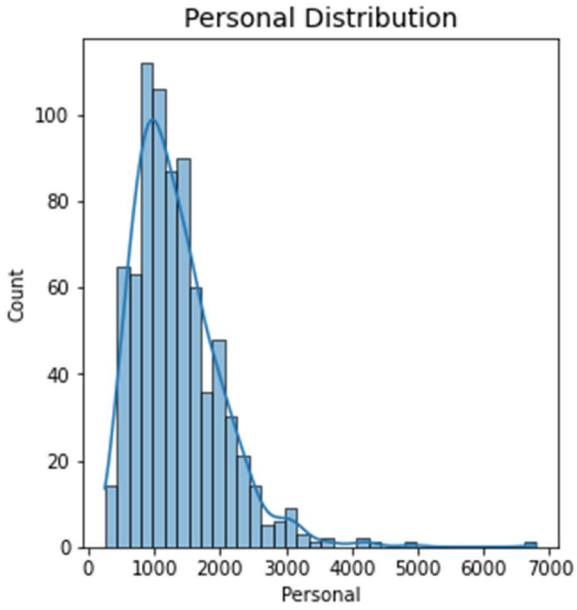


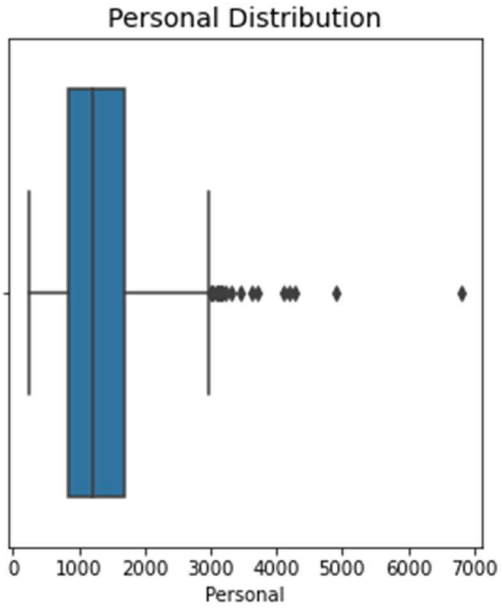


Conclusion:

- The Boxplot of Books has Outliers.
- The distribution seems to be bimodal.
- The Cost of books per student seems to be in the range of 100 to 500.

11. PERSONAL:

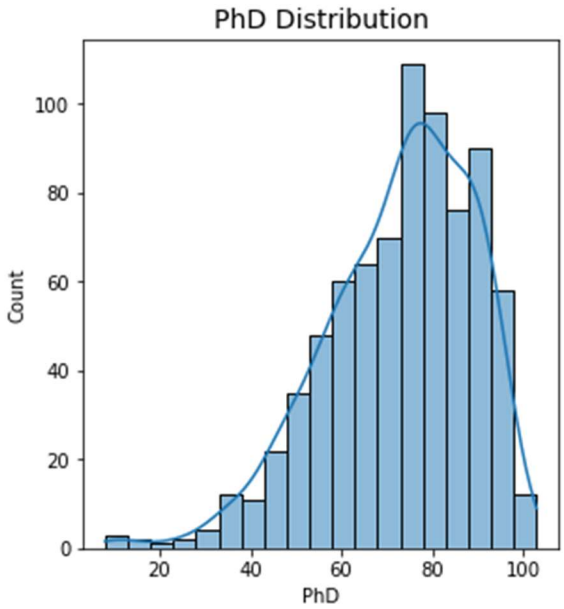


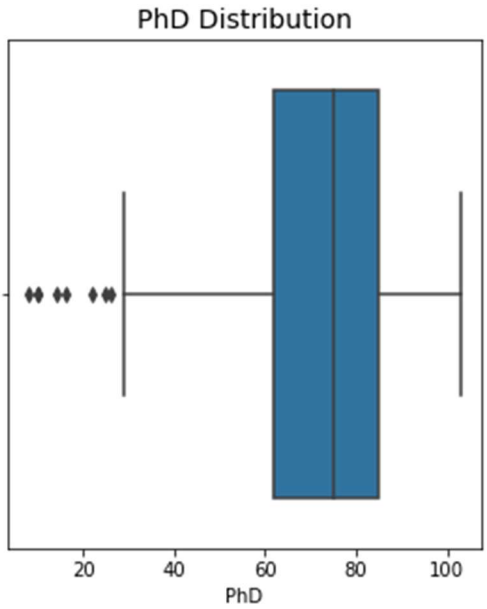


Conclusion:

- The Boxplot of Personal expenses have outliers.
- Some student's Personal expenses are more than other students.
- The distribution seems to be positively skewed.

12. PHD:

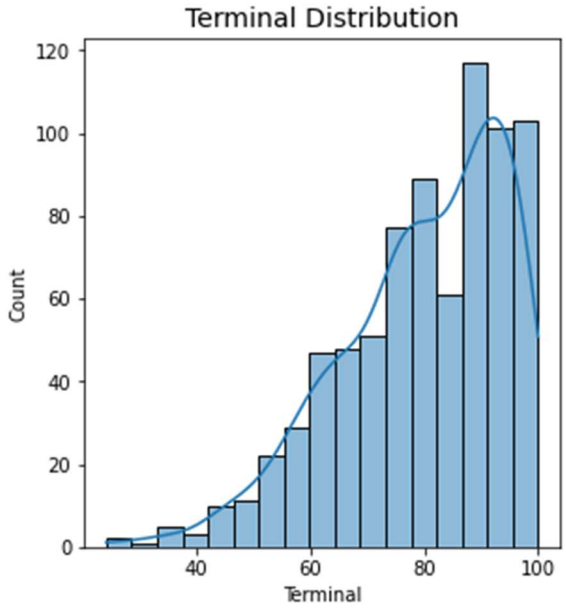


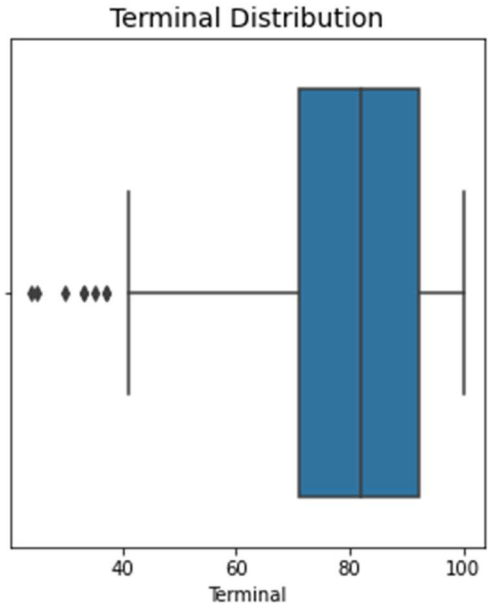


Conclusion:

- The Boxplot of PhD has Outliers.
- The distribution seems to be negatively skewed.

13. TERMINAL:

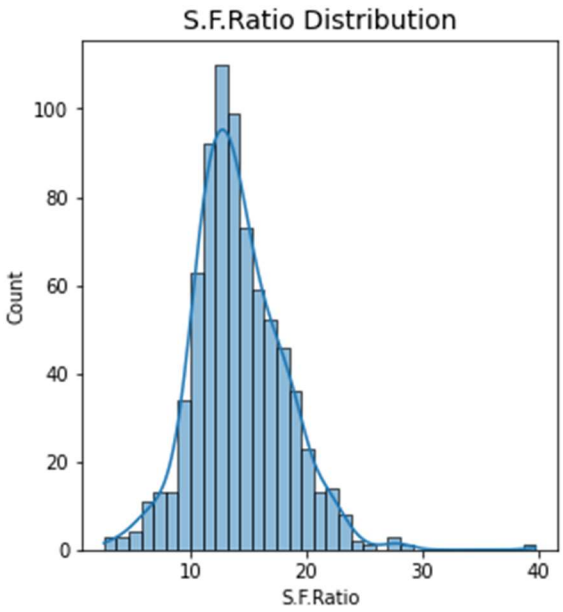


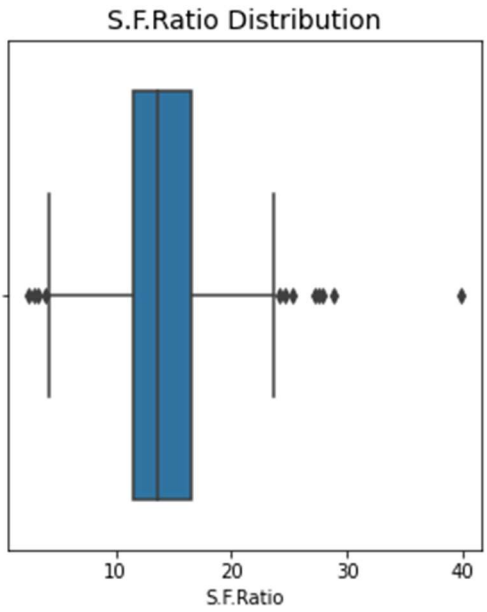


Conclusion:

- The Boxplot of Terminal seems to have outliers in the dataset.
- The distribution for Terminal seems to be negatively skewed.

14. SF RATIO:

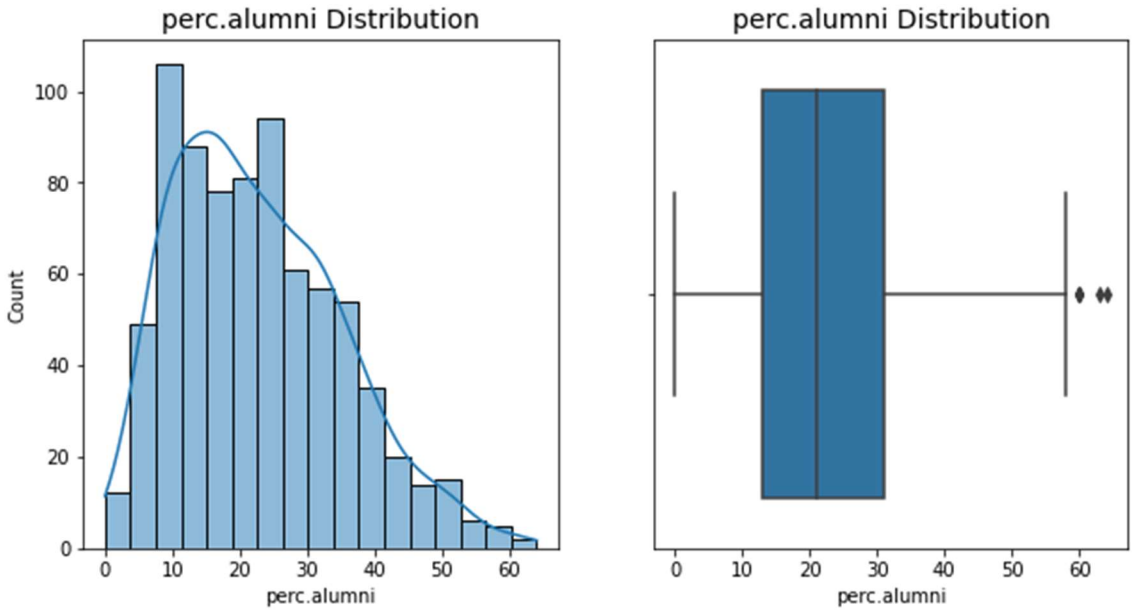




Conclusion:

- The S.F. Ratio also has outliers.
- S.F. Ratio is almost normally distributed.
- The student faculty ratio is almost same in all the university and college.

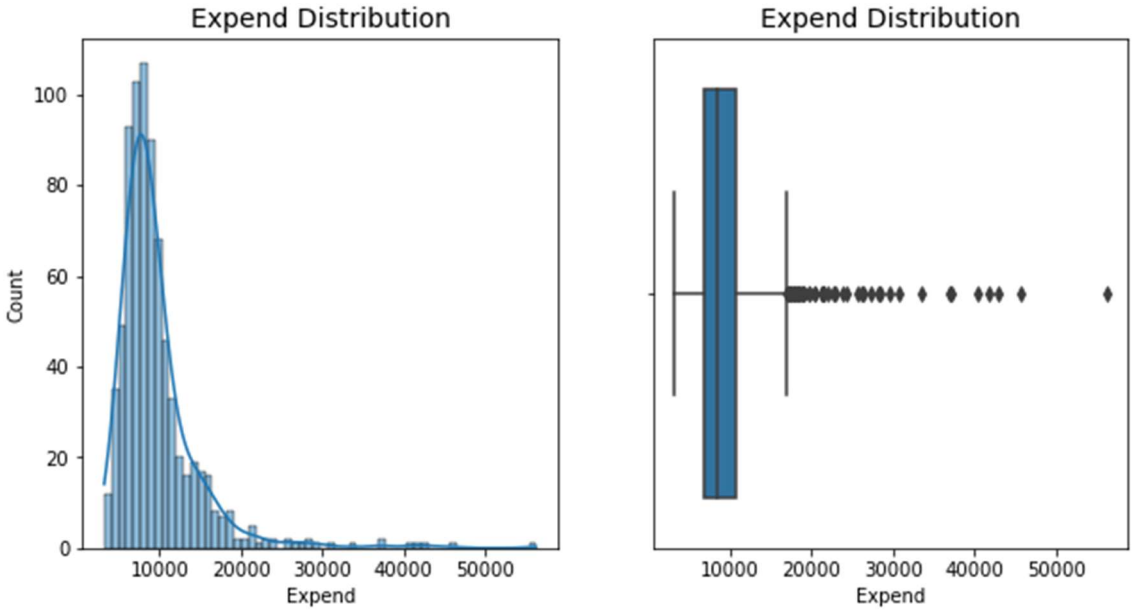
15. PERCI ALUMINI:



Conclusion:

- The percentage of Alumni boxplot seems to have Outliers.
- The distribution is almost normally distributed.

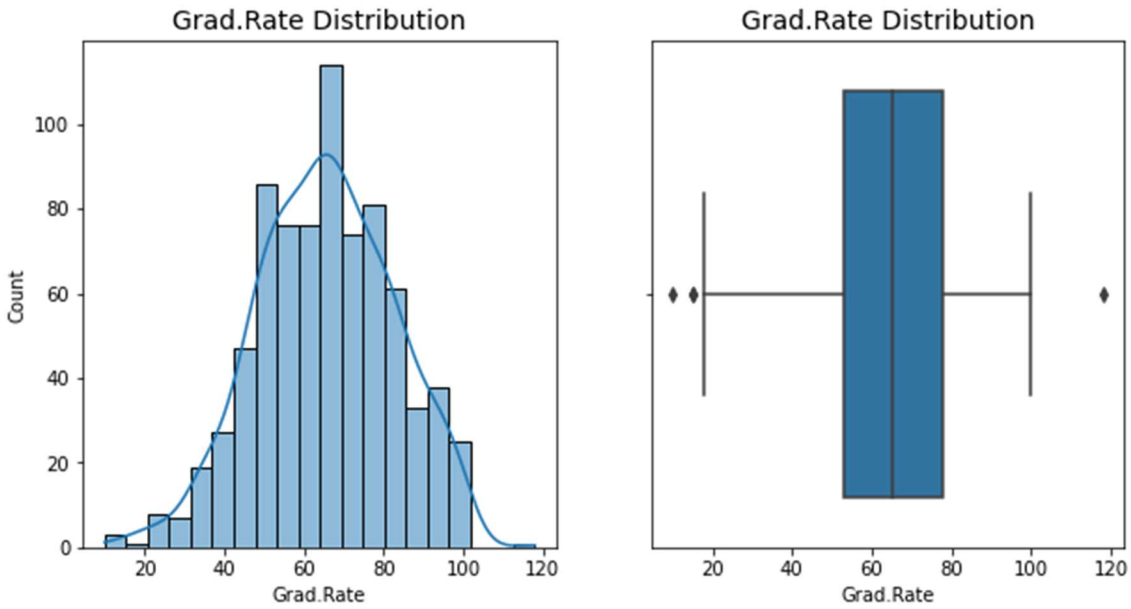
16. EXPENDITURE:



Conclusion:

- The Expenditure variable also has Outliers in the dataset.
- The distribution of the Expenditure variable is positively skewed.

17. GRAD RATE:



Conclusion:

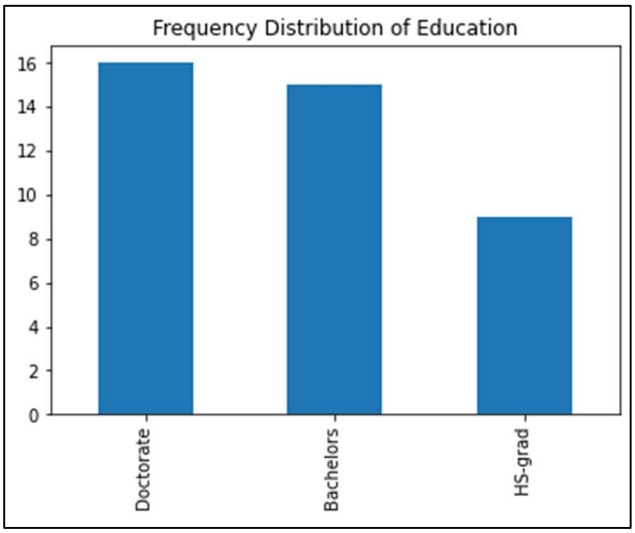
- The Boxplot of the Graduation has Outliers in the dataset.
- The Graduation rate among the students in all the university is above 60 %.
- The distribution is normally distributed.

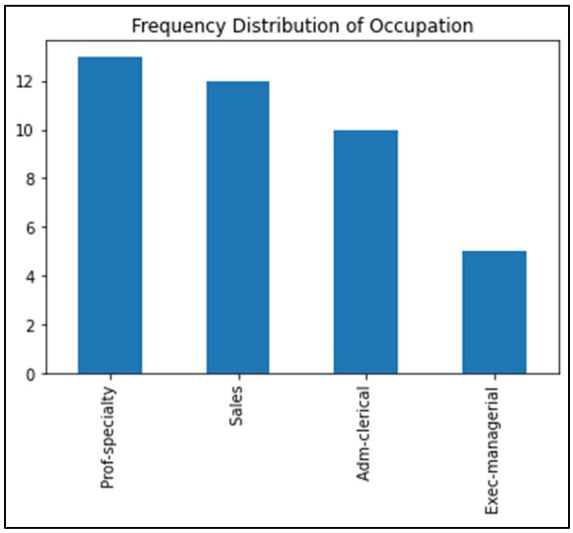
Univariate analysis of categorical variables.

```
Details of Education
---------------------------------
 Doctorate     16
 Bachelors     15
 HS-grad        9
Name: Education, dtype: int64
```

```
Details of Occupation
---------------------------------
 Prof-specialty     13
 Sales              12
 Adm-clerical       10
 Exec-managerial     5
Name: Occupation, dtype: int64
```

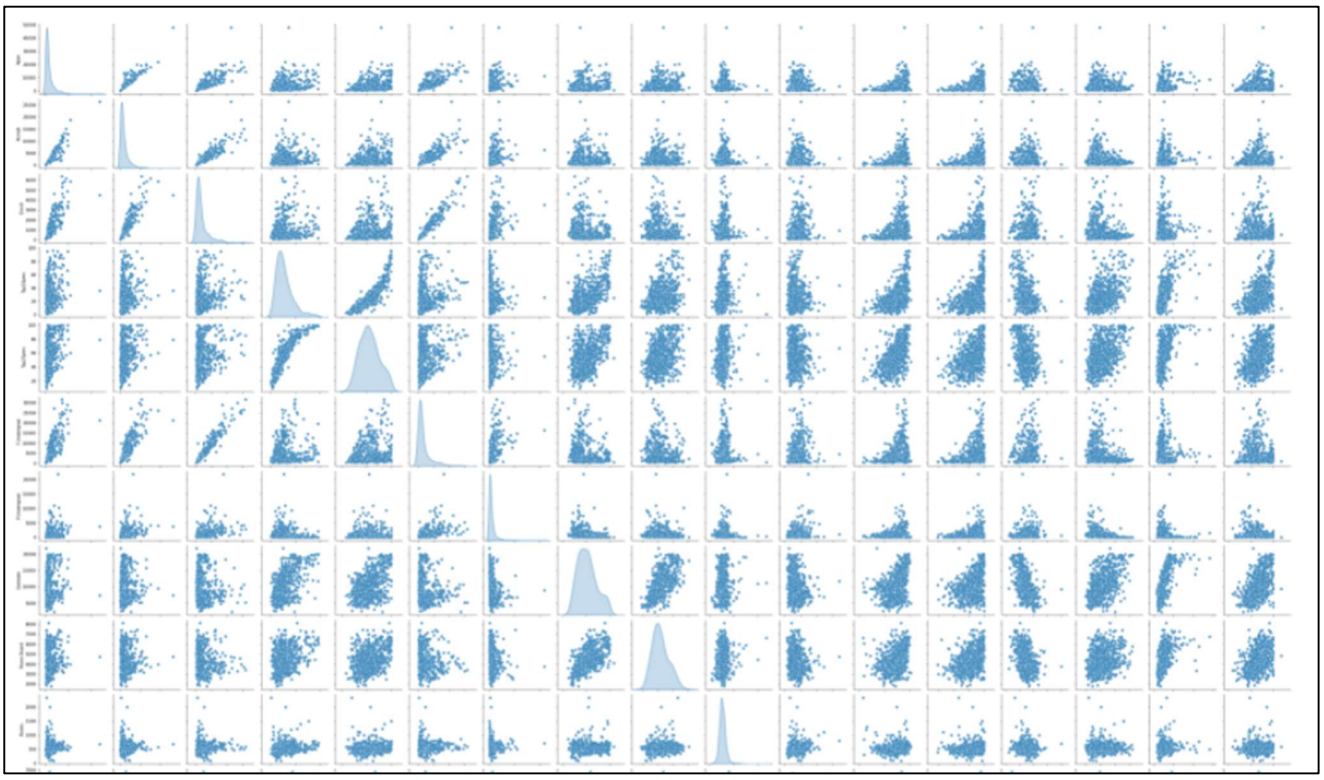
Fig.5 – Frequency distribution for categorical variables





MULTIVARIATE ANALYSIS:

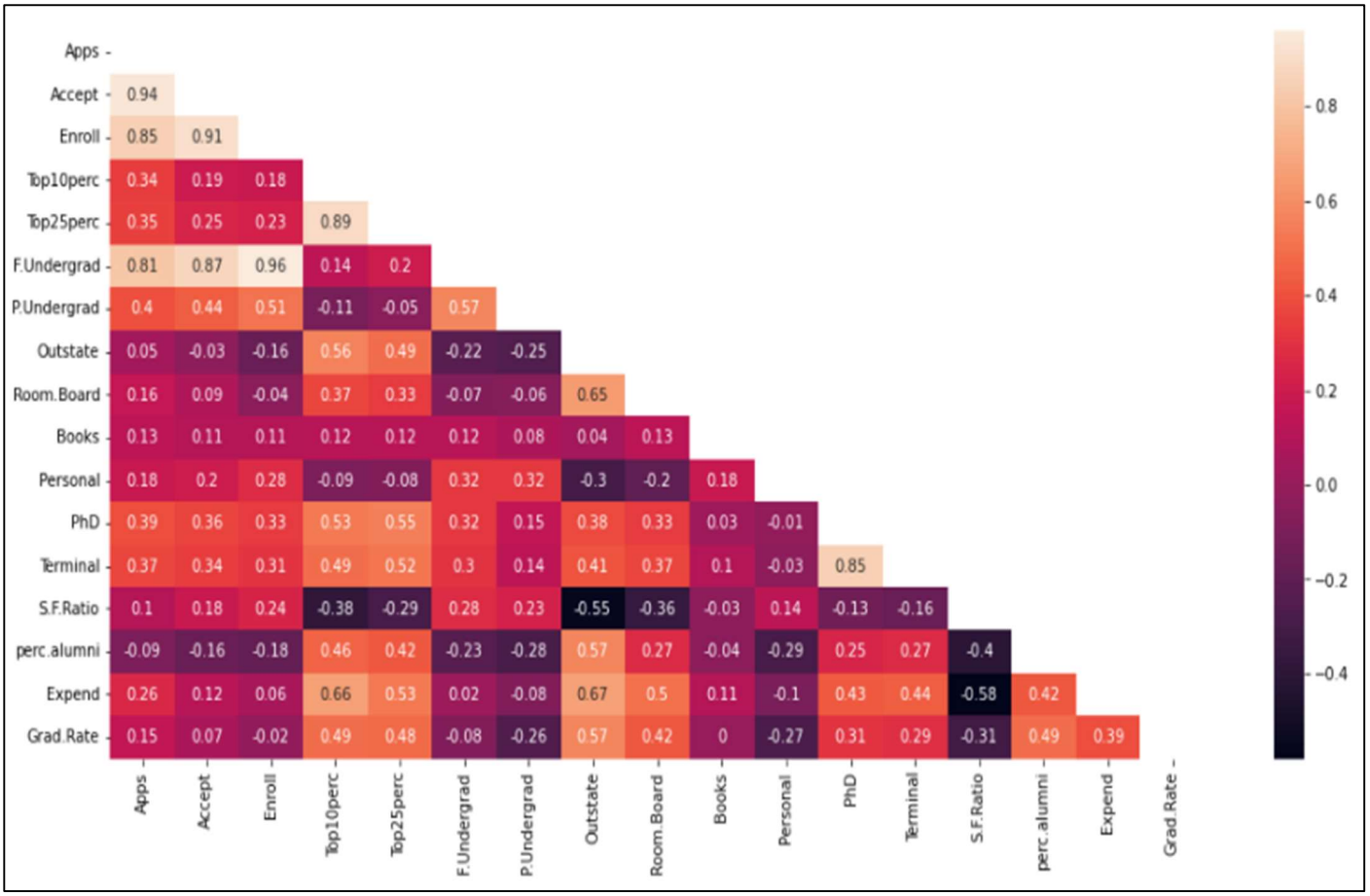
Figure 6 - Pair Plot:



Conclusion:

The Pair Plot helps us to understand the relationship between all the numerical values in the dataset. On comparing all the variables with each other we could understand the patterns. The pair plot function in seaborn makes it very easy to generate joint scatter plots for all the columns in the data.

Figure 7 - Heat Map - Correlation between two numerical values



Conclusion:

This Heat map gives us the correlation between two numerical values. There are considerable number of features that are highly correlated.

- We could understand the application variable is highly positively correlated with application accepted, students enrolled and full-time graduates. So, this relationship gives the insights on when student submits the application it is accepted and the student is enrolled as full time Graduates.
- The application with top 10 percent and 25 percent of higher secondary class, Outstate, Room Board, Books, Personal, PhD, Terminal, S.F. Ratio, expenditure and Graduation ratio are positively correlated.
- 'Enroll' shows high correlation with 'Full Time Undergraduate'.
- We can find negative correlation between application and percentage of alumni. This shows that not all students are part of alumni of their college or university.
- 'Top10perc' shows high correlation with 'Top25perc'.
- 'F. Undergraduate' shows less correlation with ‘Expend'.

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

Importance of 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', 'Enroll' is having positive integer values whereas S.F. Ratio is having decimal values. 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. It helps to normalize the data within a particular range. 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.
- Z score normalizes the data using the formula:

Formula 1. Z Score

$$z = \frac{x-\mu}{\sigma}$$

Where, $z =$ = standard score
$x =$ Observed value
$\mu =$ mean of the sample
$\sigma =$ standard deviation of the sample.

- Z score tells many standard deviations is the point away from the mean. It also allows us to determine how usual or unusual a data point is in a distribution.
- We will be doing this only for the numerical variables.

Scaling the Variables:

Before scaling I have dropped the names variable which is categorical. The dataset has 18 numerical columns with different scales.

For example, the application, accepted application, enrolled fulltime graduates, part-time graduates, outstate are number of students. The top10 percent and top20 percent are students in which the values are given in percentage. Room board, books, and personal are values associated with money. The PhD, sf ratio, percentage of alumni are percentage values of different combinations of students' teachers' alumni these are percentage values. The graduation rate is also percentage value of graduates who get graduated every year.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
0	-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.964905	-0.602312	1.270045	-0.163028	-0.115729
1	-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496	1.909208	1.215880	0.235515	-2.675646	-3.378176
2	-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.497090	0.201305	-0.554317	-0.905344	-0.259582	-1.204845	-0.931341
3	-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633	0.996791	-0.602312	-0.688173	1.185206	1.175657
4	-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005	-0.716508	-0.216723	1.518912	0.235515	0.204672	-0.523535

We can observe that all the variables are converted into same data types:

```
RangeIndex: 777 entries, 0 to 776
Data columns (total 17 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Apps         777 non-null    float64
 1   Accept       777 non-null    float64
 2   Enroll       777 non-null    float64
 3   Top10perc    777 non-null    float64
 4   Top25perc    777 non-null    float64
 5   F.Undergrad  777 non-null    float64
 6   P.Undergrad  777 non-null    float64
 7   Outstate     777 non-null    float64
 8   Room.Board   777 non-null    float64
 9   Books        777 non-null    float64
 10  Personal     777 non-null    float64
 11  PhD          777 non-null    float64
 12  Terminal     777 non-null    float64
 13  S.F.Ratio    777 non-null    float64
 14  perc.alumni  777 non-null    float64
 15  Expend       777 non-null    float64
 16  Grad.Rate    777 non-null    float64
dtypes: float64(17)
memory usage: 103.3 KB
```

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

- The comparison between the covariance and correlation matrix is that both of the terms measure the relationship and the dependency between two variables.
- Scaling in general means representation of the dataset. The numbers will not change. We are bringing the dataset into one unit.

Covariance:

Covariance indicates the direction of the linear relationship between the variables whether it is positive or negative. By direction means it is directly proportional or inversely proportional.

Population Covariance Formula:

Formula 2. Covariance

$$Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Sample Covariance Formula:

$$Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(N-1)}$$

The Covariance matrix for the scaled data is given below:

```
Covariance Matrix
%s [[ 1.00128866  0.94466636  0.84791332  0.33927032  0.35209304  0.81554018
   0.3987775   0.05022367  0.16515151  0.13272942  0.17896117  0.39120081
   0.36996762  0.09575627 -0.09034216  0.2599265   0.14694372]
 [ 0.94466636  1.00128866  0.91281145  0.19269493  0.24779465  0.87534985
   0.44183938 -0.02578774  0.09101577  0.11367165  0.20124767  0.35621633
   0.3380184   0.17645611 -0.16019604  0.12487773  0.06739929]
 [ 0.84791332  0.91281145  1.00128866  0.18152715  0.2270373   0.96588274
   0.51372977 -0.1556777  -0.04028353  0.11285614  0.28129148  0.33189629
   0.30867133  0.23757707 -0.18102711  0.06425192 -0.02236983]
 [ 0.33927032  0.19269493  0.18152715  1.00128866  0.89314445  0.1414708
  -0.10549205  0.5630552   0.37195909  0.1190116  -0.09343665  0.53251337
   0.49176793 -0.38537048  0.45607223  0.6617651   0.49562711]
 [ 0.35209304  0.24779465  0.2270373   0.89314445  1.00128866  0.19970167
  -0.05364569  0.49002449  0.33191707  0.115676   -0.08091441  0.54656564
   0.52542506 -0.29500852  0.41840277  0.52812713  0.47789622]
 [ 0.81554018  0.87534985  0.96588274  0.1414708   0.19970167  1.00128866
   0.57124738 -0.21602002 -0.06897917  0.11569867  0.31760831  0.3187472
   0.30040557  0.28006379 -0.22975792  0.01867565 -0.07887464]
 [ 0.3987775   0.44183938  0.51372977 -0.10549205 -0.05364569  0.57124738
   1.00128866 -0.25383901 -0.06140453  0.08130416  0.32029384  0.14930637
   0.14208644  0.23283016 -0.28115421 -0.08367612 -0.25733218]
 [ 0.05022367 -0.02578774 -0.1556777   0.5630552   0.49002449 -0.21602002
  -0.25383901  1.00128866  0.65509951  0.03890494 -0.29947232  0.38347594
   0.40850895 -0.55553625  0.56699214  0.6736456   0.57202613]
 [ 0.16515151  0.09101577 -0.04028353  0.37195909  0.33191707 -0.06897917
  -0.06140453  0.65509951  1.00128866  0.12812787 -0.19968518  0.32962651
   0.3750222  -0.36309504  0.27271444  0.50238599  0.42548915]
 [ 0.13272942  0.11367165  0.11285614  0.1190116   0.115676    0.11569867
   0.08130416  0.03890494  0.12812787  1.00128866  0.17952581  0.0269404
   0.10008351 -0.03197042 -0.04025955  0.11255393  0.00106226]
 [ 0.17896117  0.20124767  0.28129148 -0.09343665 -0.08091441  0.31760831
   0.32029384 -0.29947232 -0.19968518  0.17952581  1.00128866 -0.01094989
  -0.03065256  0.13652054 -0.2863366  -0.09801804 -0.26969106]
```

```
[ 0.39120081  0.35621633  0.33189629  0.53251337  0.54656564  0.3187472
  0.14930637  0.38347594  0.32962651  0.0269404  -0.01094989  1.00128866
  0.85068186 -0.13069832  0.24932955  0.43331936  0.30543094]
[ 0.36996762  0.3380184   0.30867133  0.49176793  0.52542506  0.30040557
  0.14208644  0.40850895  0.3750222   0.10008351 -0.03065256  0.85068186
  1.00128866 -0.16031027  0.26747453  0.43936469  0.28990033]
[ 0.09575627  0.17645611  0.23757707 -0.38537048 -0.29500852  0.28006379
  0.23283016 -0.55553625 -0.36309504 -0.03197042  0.13652054 -0.13069832
 -0.16031027  1.00128866 -0.4034484  -0.5845844  -0.30710565]
[-0.09034216 -0.16019604 -0.18102711  0.45607223  0.41840277 -0.22975792
 -0.28115421  0.56699214  0.27271444 -0.04025955 -0.2863366   0.24932955
  0.26747453 -0.4034484   1.00128866  0.41825001  0.49153016]
[ 0.2599265   0.12487773  0.06425192  0.6617651   0.52812713  0.01867565
 -0.08367612  0.6736456   0.50238599  0.11255393 -0.09801804  0.43331936
  0.43936469 -0.5845844   0.41825001  1.00128866  0.39084571]
[ 0.14694372  0.06739929 -0.02236983  0.49562711  0.47789622 -0.07887464
 -0.25733218  0.57202613  0.42548915  0.00106226 -0.26969106  0.30543094
  0.28990033 -0.30710565  0.49153016  0.39084571  1.00128866]]
```

Correlation:

Correlation measures the strength (how much?) and the direction of the linear relationship between two variables. Strength is that is that positively correlated or negatively correlated.

Formula 3. Correlation

$$Correlation = \frac{Cov(x, y)}{\sigma_x * \sigma_y}$$

Where,

$Cov(x, y)$ = Covariance of x and y
$\sigma_x =$ Standard deviation of x
$\sigma_y =$ Standard deviation of y

The Correlation matrix for the scaled data is given below:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899	0.113525	0.200989	0.355758
Enroll	0.846822	0.911637	1.000000	0.181294	0.226745	0.964640	0.513069	-0.155477	-0.040232	0.112711	0.280929	0.331469
Top10perc	0.338834	0.192447	0.181294	1.000000	0.891995	0.141289	-0.105356	0.562331	0.371480	0.118858	-0.093316	0.531828
Top25perc	0.351640	0.247476	0.226745	0.891995	1.000000	0.199445	-0.053577	0.489394	0.331490	0.115527	-0.080810	0.545862
F.Undergrad	0.814491	0.874223	0.964640	0.141289	0.199445	1.000000	0.570512	-0.215742	-0.068890	0.115550	0.317200	0.318337
P.Undergrad	0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	1.000000	-0.253512	-0.061326	0.081200	0.319882	0.149114
Outstate	0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	-0.253512	1.000000	0.654256	0.038855	-0.299087	0.382982
Room.Board	0.164939	0.090899	-0.040232	0.371480	0.331490	-0.068890	-0.061326	0.654256	1.000000	0.127963	-0.199428	0.329202
Books	0.132559	0.113525	0.112711	0.118858	0.115527	0.115550	0.081200	0.038855	0.127963	1.000000	0.179295	0.026906
Personal	0.178731	0.200989	0.280929	-0.093316	-0.080810	0.317200	0.319882	-0.299087	-0.199428	0.179295	1.000000	-0.010936
PhD	0.390697	0.355758	0.331469	0.531828	0.545862	0.318337	0.149114	0.382982	0.329202	0.026906	-0.010936	1.000000
Terminal	0.369491	0.337583	0.308274	0.491135	0.524749	0.300019	0.141904	0.407983	0.374540	0.099955	-0.030613	0.849587
S.F.Ratio	0.095633	0.176229	0.237271	-0.384875	-0.294629	0.279703	0.232531	-0.554821	-0.362628	-0.031929	0.136345	-0.130530
perc.alumni	-0.090226	-0.159990	-0.180794	0.455485	0.417864	-0.229462	-0.280792	0.566262	0.272363	-0.040208	-0.285968	0.249009
Expend	0.259592	0.124717	0.064169	0.660913	0.527447	0.018652	-0.083568	0.672779	0.501739	0.112409	-0.097892	0.432762
Grad.Rate	0.146755	0.067313	-0.022341	0.494989	0.477281	-0.078773	-0.257001	0.571290	0.424942	0.001061	-0.269344	0.305038

Conclusion:

- Covariance indicates the direction of the linear relationship between variables. Correlation measures both the strength and direction of the linear relationship between two variables.
- Both Covariance and Correlation Matrix measures the relationship and the dependency between two variables.
- The Correlation matrix before scaling and after scaling remains the same.
- From the above table we can observe the variable which are highly positively correlated and variable which are highly negatively correlated.
- We can see that application, acceptance, enrolment and fulltime graduates are highly positively correlated.
- Top 10 percentage and Top 25 percentage are highly positively correlated.

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

Figure 8 - Box Plot for Checking the outliers before scaling

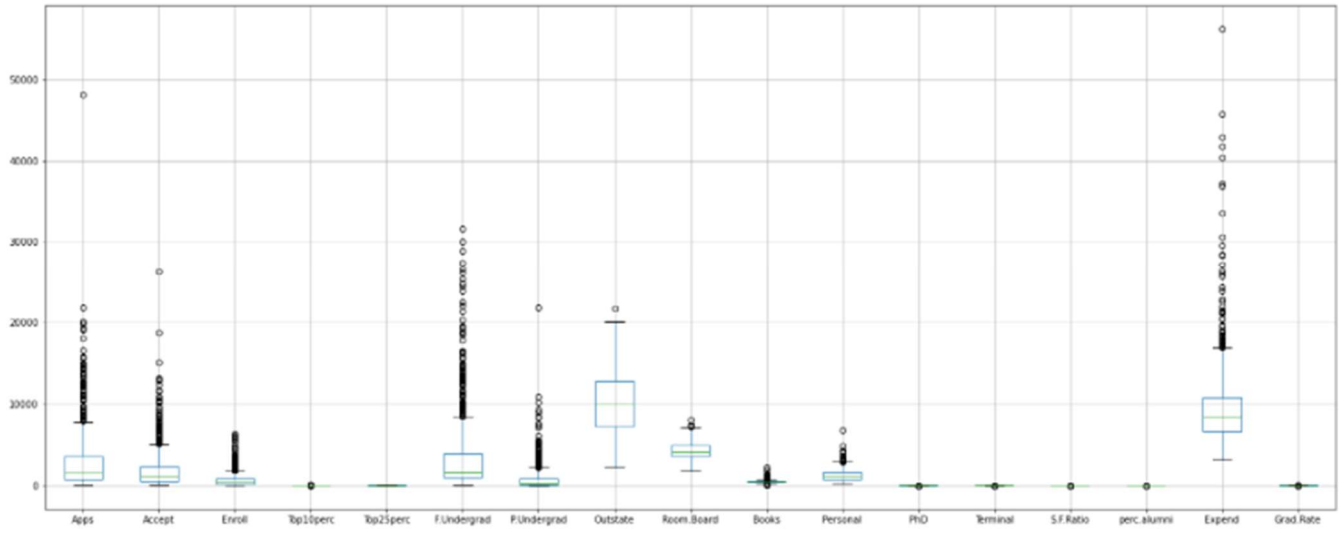
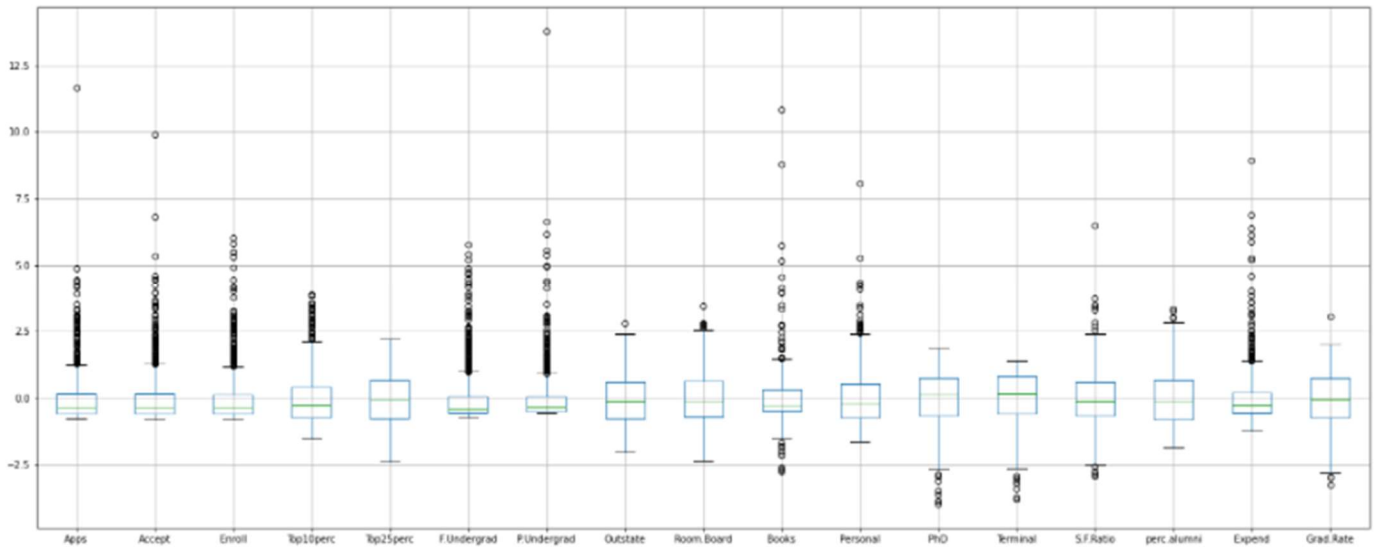


Figure 9 - Box Plot for Checking the outliers after scaling



Inference:

- The Outliers are still present in the dataset.

Reason:

- Scaling does not remove Outliers. Scaling scales the values on a Z score distribution. There are many methods to remove Outliers. For example, we can consider taking 3 standard deviations as outliers or either we can remove them or impute them with IQR values.

Analysis 2.5. Extract the eigenvalues and eigenvectors.

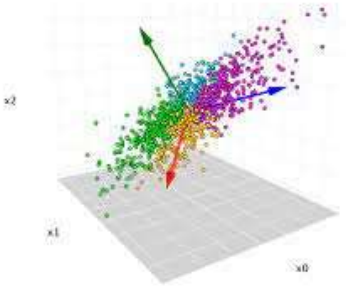
Eigenvector and Eigenvalues are below:

```
Eigen Vectors
%s [[-2.48765602e-01  3.31598227e-01 -6.30921033e-02  2.81310530e-01
  -5.74140964e-03 -1.62374420e-02 -4.24863486e-02 -1.03090398e-01
  -9.02270802e-02  5.25098025e-02 -3.58970400e-01  4.59139498e-01
  -4.30462074e-02  1.33405806e-01 -8.06328039e-02 -5.95830975e-01
   2.40709086e-02]
 [-2.07601502e-01  3.72116750e-01 -1.01249056e-01  2.67817346e-01
  -5.57860920e-02  7.53468452e-03 -1.29497196e-02 -5.62709623e-02
  -1.77864814e-01  4.11400844e-02  5.43427250e-01 -5.18568789e-01
   5.84055850e-02 -1.45497511e-01 -3.34674281e-02 -2.92642398e-01
  -1.45102446e-01]
 [-1.76303592e-01  4.03724252e-01 -8.29855709e-02  1.61826771e-01
   5.56936353e-02 -4.25579803e-02 -2.76928937e-02  5.86623552e-02
  -1.28560713e-01  3.44879147e-02 -6.09651110e-01 -4.04318439e-01
   6.93988831e-02  2.95896092e-02  8.56967180e-02  4.44638207e-01
   1.11431545e-02]
 [-3.54273947e-01 -8.24118211e-02  3.50555339e-02 -5.15472524e-02
   3.95434345e-01 -5.26927980e-02 -1.61332069e-01 -1.22678028e-01
   3.41099863e-01  6.40257785e-02  1.44986329e-01 -1.48738723e-01
   8.10481404e-03  6.97722522e-01  1.07828189e-01 -1.02303616e-03
   3.85543001e-02]
 [-3.44001279e-01 -4.47786551e-02 -2.41479376e-02 -1.09766541e-01
   4.26533594e-01  3.30915896e-02 -1.18485556e-01 -1.02491967e-01
   4.03711989e-01  1.45492289e-02 -8.03478445e-02  5.18683400e-02
   2.73128469e-01 -6.17274818e-01 -1.51742110e-01 -2.18838802e-02
  -8.93515563e-02]
 [-1.54640962e-01  4.17673774e-01 -6.13929764e-02  1.00412335e-01
   4.34543659e-02 -4.34542349e-02 -2.50763629e-02  7.88896442e-02
  -5.94419181e-02  2.08471834e-02  4.14705279e-01  5.60363054e-01
   8.11578181e-02  9.91640992e-03  5.63728817e-02  5.23622267e-01
   5.61767721e-02]
```

```
  [-2.64425045e-02  3.15087830e-01  1.39681716e-01 -1.58558487e-01
   -3.02385408e-01 -1.91198583e-01  6.10423460e-02  5.70783816e-01
    5.60672902e-01 -2.23105808e-01 -9.01788964e-03 -5.27313042e-02
   -1.00693324e-01  2.09515982e-02 -1.92857500e-02 -1.25997650e-01
   -6.35360730e-02]
  [-2.94736419e-01 -2.49643522e-01  4.65988731e-02  1.31291364e-01
   -2.22532003e-01 -3.00003910e-02  1.08528966e-01  9.84599754e-03
   -4.57332880e-03  1.86675363e-01 -5.08995918e-02  1.01594830e-01
   -1.43220673e-01  3.83544794e-02  3.40115407e-02  1.41856014e-01
   -8.23443779e-01]
  [-2.49030449e-01 -1.37808883e-01  1.48967389e-01  1.84995991e-01
   -5.60919470e-01  1.62755446e-01  2.09744235e-01 -2.21453442e-01
    2.75022548e-01  2.98324237e-01 -1.14639620e-03 -2.59293381e-02
    3.59321731e-01  3.40197083e-03  5.84289756e-02  6.97485854e-02
    3.54559731e-01]
  [-6.47575181e-02  5.63418434e-02  6.77411649e-01  8.70892205e-02
    1.27288825e-01  6.41054950e-01 -1.49692034e-01  2.13293009e-01
   -1.33663353e-01 -8.20292186e-02 -7.72631963e-04  2.88282896e-03
   -3.19400370e-02 -9.43887925e-03  6.68494643e-02 -1.14379958e-02
   -2.81593679e-02]
  [ 4.25285386e-02  2.19929218e-01  4.99721120e-01 -2.30710568e-01
    2.22311021e-01 -3.31398003e-01  6.33790064e-01 -2.32660840e-01
   -9.44688900e-02  1.36027616e-01  1.11433396e-03 -1.28904022e-02
    1.85784733e-02 -3.09001353e-03 -2.75286207e-02 -3.94547417e-02
   -3.92640266e-02]
  [-3.18312875e-01  5.83113174e-02 -1.27028371e-01 -5.34724832e-01
   -1.40166326e-01  9.12555212e-02 -1.09641298e-03 -7.70400002e-02
   -1.85181525e-01 -1.23452200e-01 -1.38133366e-02  2.98075465e-02
   -4.03723253e-02 -1.12055599e-01  6.91126145e-01 -1.27696382e-01
    2.32224316e-02]
[-3.17056016e-01  4.64294477e-02 -6.60375454e-02 -5.19443019e-01
 -2.04719730e-01  1.54927646e-01 -2.84770105e-02 -1.21613297e-02
 -2.54938198e-01 -8.85784627e-02 -6.20932749e-03 -2.70759809e-02
  5.89734026e-02  1.58909651e-01 -6.71008607e-01  5.83134662e-02
  1.64850420e-02]
[ 1.76957895e-01  2.46665277e-01 -2.89848401e-01 -1.61189487e-01
  7.93882496e-02  4.87045875e-01  2.19259358e-01 -8.36048735e-02
  2.74544380e-01  4.72045249e-01  2.22215182e-03 -2.12476294e-02
 -4.45000727e-01 -2.08991284e-02 -4.13740967e-02  1.77152700e-02
 -1.10262122e-02]
[-2.05082369e-01 -2.46595274e-01 -1.46989274e-01  1.73142230e-02
  2.16297411e-01 -4.73400144e-02  2.43321156e-01  6.78523654e-01
 -2.55334907e-01  4.22999706e-01  1.91869743e-02  3.33406243e-03
  1.30727978e-01 -8.41789410e-03  2.71542091e-02 -1.04088088e-01
  1.82660654e-01]
[-3.18908750e-01 -1.31689865e-01  2.26743985e-01  7.92734946e-02
 -7.59581203e-02 -2.98118619e-01 -2.26584481e-01 -5.41593771e-02
 -4.91388809e-02  1.32286331e-01  3.53098218e-02 -4.38803230e-02
 -6.92088870e-01 -2.27742017e-01 -7.31225166e-02  9.37464497e-02
  3.25982295e-01]
[-2.52315654e-01 -1.69240532e-01 -2.08064649e-01  2.69129066e-01
  1.09267913e-01  2.16163313e-01  5.59943937e-01 -5.33553891e-03
  4.19043052e-02 -5.90271067e-01  1.30710024e-02 -5.00844705e-03
 -2.19839000e-01 -3.39433604e-03 -3.64767385e-02  6.91969778e-02
  1.22106697e-01]]


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

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



- **First, we will check the eigenvalues by applying PCA taking all features.**

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

- **Check the explained variance for each PC.**

  Formula 4. Explained variance

  Explained variance = (eigen value of each PC)/ (sum of eigen values of all PCs)

  The output will be:

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

- **Extract Eigenvectors**

```
array([[ 2.48765602e-01,  2.07601502e-01,  1.76303592e-01,
         3.54273947e-01,  3.44001279e-01,  1.54640962e-01,
         2.64425045e-02,  2.94736419e-01,  2.49030449e-01,
         6.47575181e-02, -4.25285386e-02,  3.18312875e-01,
         3.17056016e-01, -1.76957895e-01,  2.05082369e-01,
         3.18908750e-01,  2.52315654e-01],
       [ 3.31598227e-01,  3.72116750e-01,  4.03724252e-01,
        -8.24118211e-02, -4.47786551e-02,  4.17673774e-01,
         3.15087830e-01, -2.49643522e-01, -1.37808883e-01,
         5.63418434e-02,  2.19929218e-01,  5.83113174e-02,
         4.64294477e-02,  2.46665277e-01, -2.46595274e-01,
        -1.31689865e-01, -1.69240532e-01],
       [-6.30921033e-02, -1.01249056e-01, -8.29855709e-02,
         3.50555339e-02, -2.41479376e-02, -6.13929764e-02,
         1.39681716e-01,  4.65988731e-02,  1.48967389e-01,
         6.77411649e-01,  4.99721120e-01, -1.27028371e-01,
        -6.60375454e-02, -2.89848401e-01, -1.46989274e-01,
         2.26743985e-01, -2.08064649e-01],
       [ 2.81310530e-01,  2.67817346e-01,  1.61826771e-01,
        -5.15472524e-02, -1.09766541e-01,  1.00412335e-01,
        -1.58558487e-01,  1.31291364e-01,  1.84995991e-01,
         8.70892205e-02, -2.30710568e-01, -5.34724832e-01,
        -5.19443019e-01, -1.61189487e-01,  1.73142230e-02,
         7.92734946e-02,  2.69129066e-01],
       [ 5.74140964e-03,  5.57860920e-02, -5.56936353e-02,
        -3.95434345e-01, -4.26533594e-01, -4.34543659e-02,
         3.02385408e-01,  2.22532003e-01,  5.60919470e-01,
        -1.27288825e-01, -2.22311021e-01,  1.40166326e-01,
         2.04719730e-01, -7.93882496e-02, -2.16297411e-01,
         7.59581203e-02, -1.09267913e-01],
```

```
[-1.62374420e-02,  7.53468452e-03, -4.25579803e-02,
 -5.26927980e-02,  3.30915896e-02, -4.34542349e-02,
 -1.91198583e-01, -3.00003910e-02,  1.62755446e-01,
  6.41054950e-01, -3.31398003e-01,  9.12555212e-02,
  1.54927646e-01,  4.87045875e-01, -4.73400144e-02,
 -2.98118619e-01,  2.16163313e-01],
[-4.24863486e-02, -1.29497196e-02, -2.76928937e-02,
 -1.61332069e-01, -1.18485556e-01, -2.50763629e-02,
  6.10423460e-02,  1.08528966e-01,  2.09744235e-01,
 -1.49692034e-01,  6.33790064e-01, -1.09641298e-03,
 -2.84770105e-02,  2.19259358e-01,  2.43321156e-01,
 -2.26584481e-01,  5.59943937e-01],
[-1.03090398e-01, -5.62709623e-02,  5.86623552e-02,
 -1.22678028e-01, -1.02491967e-01,  7.88896442e-02,
  5.70783816e-01,  9.84599754e-03, -2.21453442e-01,
  2.13293009e-01, -2.32660840e-01, -7.70400002e-02,
 -1.21613297e-02, -8.36048735e-02,  6.78523654e-01,
 -5.41593771e-02, -5.33553891e-03],
[-9.02270802e-02, -1.77864814e-01, -1.28560713e-01,
  3.41099863e-01,  4.03711989e-01, -5.94419181e-02,
  5.60672902e-01, -4.57332880e-03,  2.75022548e-01,
 -1.33663353e-01, -9.44688900e-02, -1.85181525e-01,
 -2.54938198e-01,  2.74544380e-01, -2.55334907e-01,
 -4.91388809e-02,  4.19043052e-02],
[ 5.25098025e-02,  4.11400844e-02,  3.44879147e-02,
  6.40257785e-02,  1.45492289e-02,  2.08471834e-02,
 -2.23105808e-01,  1.86675363e-01,  2.98324237e-01,
 -8.20292186e-02,  1.36027616e-01, -1.23452200e-01,
 -8.85784627e-02,  4.72045249e-01,  4.22999706e-01,
  1.32286331e-01, -5.90271067e-01],
[ 4.30462074e-02, -5.84055850e-02, -6.93988831e-02,
 -8.10481404e-03, -2.73128469e-01, -8.11578181e-02,
  1.00693324e-01,  1.43220673e-01, -3.59321731e-01,
  3.19400370e-02, -1.85784733e-02,  4.03723253e-02,
 -5.89734026e-02,  4.45000727e-01, -1.30727978e-01,
  6.92088870e-01,  2.19839000e-01],
[ 2.40709086e-02, -1.45102446e-01,  1.11431545e-02,
  3.85543001e-02, -8.93515563e-02,  5.61767721e-02,
 -6.35360730e-02, -8.23443779e-01,  3.54559731e-01,
 -2.81593679e-02, -3.92640266e-02,  2.32224316e-02,
  1.64850420e-02, -1.10262122e-02,  1.82660654e-01,
  3.25982295e-01,  1.22106697e-01],
[ 5.95830975e-01,  2.92642398e-01, -4.44638207e-01,
  1.02303616e-03,  2.18838802e-02, -5.23622267e-01,
  1.25997650e-01, -1.41856014e-01, -6.97485854e-02,
  1.14379958e-02,  3.94547417e-02,  1.27696382e-01,
 -5.83134662e-02, -1.77152700e-02,  1.04088088e-01,
 -9.37464497e-02, -6.91969778e-02],
[ 8.06328039e-02,  3.34674281e-02, -8.56967180e-02,
 -1.07828189e-01,  1.51742110e-01, -5.63728817e-02,
  1.92857500e-02, -3.40115407e-02, -5.84289756e-02,
 -6.68494643e-02,  2.75286207e-02, -6.91126145e-01,
  6.71008607e-01,  4.13740967e-02, -2.71542091e-02,
  7.31225166e-02,  3.64767385e-02],
[ 1.33405806e-01, -1.45497511e-01,  2.95896092e-02,
  6.97722522e-01, -6.17274818e-01,  9.91640992e-03,
  2.09515982e-02,  3.83544794e-02,  3.40197083e-03,
 -9.43887925e-03, -3.09001353e-03, -1.12055599e-01,
  1.58909651e-01, -2.08991284e-02, -8.41789410e-03,
 -2.27742017e-01, -3.39433604e-03],
```

```
[ 4.59139498e-01, -5.18568789e-01, -4.04318439e-01,
 -1.48738723e-01,  5.18683400e-02,  5.60363054e-01,
 -5.27313042e-02,  1.01594830e-01, -2.59293381e-02,
  2.88282896e-03, -1.28904022e-02,  2.98075465e-02,
 -2.70759809e-02, -2.12476294e-02,  3.33406243e-03,
 -4.38803230e-02, -5.00844705e-03],
[ 3.58970400e-01, -5.43427250e-01,  6.09651110e-01,
 -1.44986329e-01,  8.03478445e-02, -4.14705279e-01,
  9.01788964e-03,  5.08995918e-02,  1.14639620e-03,
  7.72631963e-04, -1.11433396e-03,  1.38133366e-02,
  6.20932749e-03, -2.22215182e-03, -1.91869743e-02,
 -3.53098218e-02, -1.30710024e-02]])
```

PCA is performed and it is exported into a data frame. After PCA the multi collinearity is highly reduced.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
0	0.248766	0.207602	0.176304	0.354274	0.344001	0.154641	0.026443	0.294736	0.249030	0.064758	-0.042529	0.318313	0.317056
1	0.331598	0.372117	0.403724	-0.082412	-0.044779	0.417674	0.315088	-0.249644	-0.137809	0.056342	0.219929	0.058311	0.046429
2	-0.063092	-0.101249	-0.082986	0.035056	-0.024148	-0.061393	0.139682	0.046599	0.148967	0.677412	0.499721	-0.127028	-0.066038
3	0.281310	0.267817	0.161827	-0.051547	-0.109767	0.100412	-0.158558	0.131291	0.184996	0.087089	-0.230711	-0.534725	-0.519443
4	0.005742	0.055786	-0.055694	-0.395434	-0.426534	-0.043454	0.302385	0.222532	0.560919	-0.127289	-0.222311	0.140166	0.204720

Analysis 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 explicit for the first PC is:

```
The Linear eq of 1st component:
0.249 * Apps + 0.208 * Accept + 0.176 * Enroll + 0.354 * Top10perc + 0.344 * Top25perc + 0.155 * F.Undergrad + 0.026 * P.Und
ergrad + 0.295 * Outstate + 0.249 * Room.Board + 0.065 * Books + -0.043 * Personal + 0.318 * PhD + 0.317 * Terminal + -0.177
* S.F.Ratio + 0.205 * perc.alumni + 0.319 * Expend + 0.252 * Grad.Rate +
```

Analysis 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?

The Cumulative values of the Eigenvalues are:

```
array([ 32.0206282 ,  58.36084263,  65.26175919,  71.18474841,
        76.67315352,  81.65785448,  85.21672597,  88.67034731,
        91.78758099,  94.16277251,  96.00419883,  97.30024023,
        98.28599436,  99.13183669,  99.64896227,  99.86471628,
       100.        ])
```

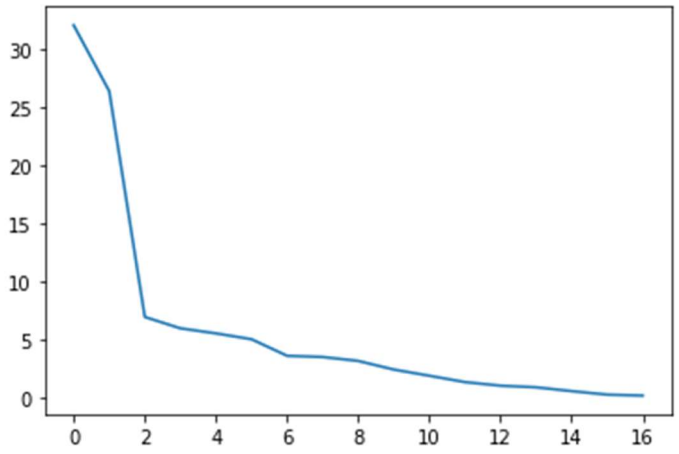
- **Adding the Eigenvalues, we will get sum of 100.**
- To decide the Optimum number of Principle Components,
  - Check for cumulative variance up to 90%
  - The incremental value between the components should not be less than five percent.

So based on this we can decide the optimum number of principal components as 6, because after this the incremental values are less than 5%.

So, we select 5 principal components for this case study.

```
array([[ 0.2487656 ,  0.2076015 ,  0.17630359,  0.35427395,  0.34400128,
         0.15464096,  0.0264425 ,  0.29473642,  0.24903045,  0.06475752,
        -0.04252854,  0.31831287,  0.31705602, -0.17695789,  0.20508237,
         0.31890875,  0.25231565],
       [ 0.33159823,  0.37211675,  0.40372425, -0.08241182, -0.04477866,
         0.41767377,  0.31508783, -0.24964352, -0.13780888,  0.05634184,
         0.21992922,  0.05831132,  0.04642945,  0.24666528, -0.24659527,
        -0.13168986, -0.16924053],
       [-0.0630921 , -0.10124906, -0.08298557,  0.03505553, -0.02414794,
        -0.06139297,  0.13968172,  0.04659887,  0.14896739,  0.67741165,
         0.49972112, -0.12702837, -0.06603755, -0.2898484 , -0.14698927,
         0.22674398, -0.20806465],
       [ 0.28131051,  0.26781737,  0.16182679, -0.05154724, -0.10976654,
         0.1004123 , -0.15855848,  0.13129136,  0.18499599,  0.08708922,
        -0.23071057, -0.53472483, -0.51944302, -0.16118949,  0.01731422,
         0.0792735 ,  0.26912907],
       [ 0.0057414 ,  0.0557861 , -0.05569362, -0.39543434, -0.4265336 ,
        -0.04345438,  0.30238541,  0.222532  ,  0.56091947, -0.12728883,
        -0.22231102,  0.14016632,  0.20471973, -0.07938825, -0.21629741,
         0.07595812, -0.10926791]])
```

Graphically:



Conclusion:

- The first Components explain 32.02% variance in data.
- The first two Components explain 58.36% variance in data.
- The first three Components explain 65.26% variance in data.
- The first four Components explain 71.18% variance in data.
- The first five Components explain 76.67% variance in data.

The Eigen vectors or PC for this case study is five, we can understand how much each variable contributes to the principal components. In other words, we can also say weights attached to each variable. With this Eigen vectors we can understand which variable has more weightage and influences the dataset in the principal components. The PCA reduces the multi collinearity and with this reduced collinearity we can runs models and improved efficiency scores

Fig.10 – Screen Plot with PCA and Eigen Value

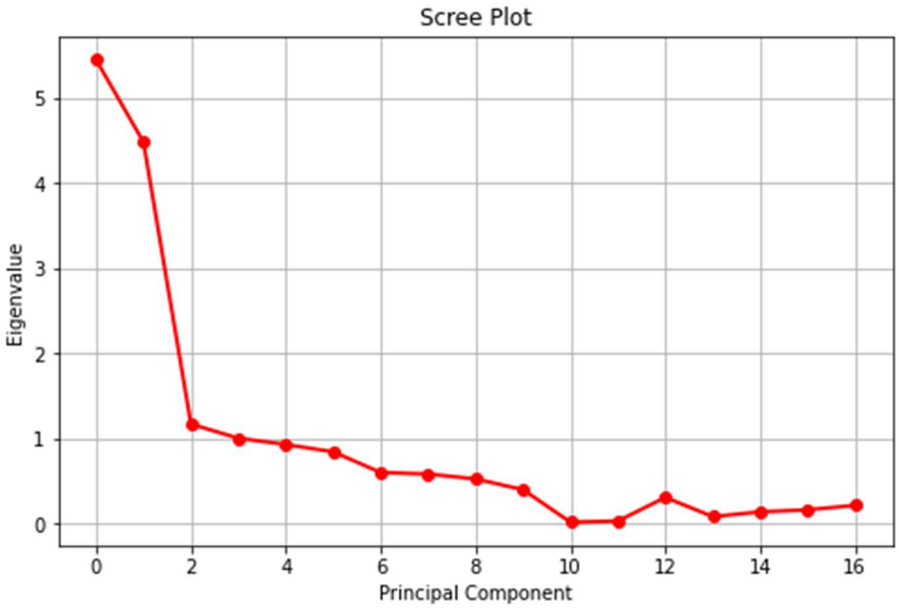
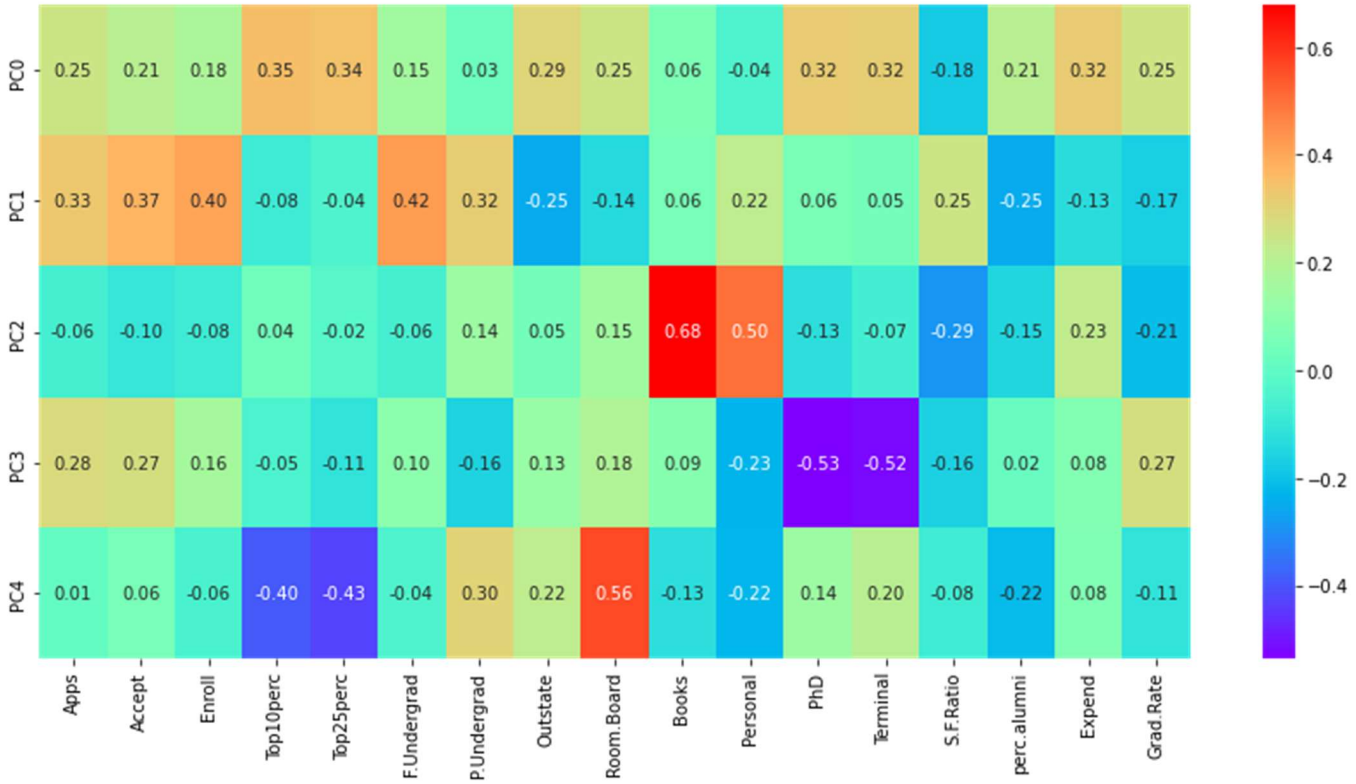


Fig.11 – Heat Map of PCAs:



Analysis 2.9. Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [Hint: Write Interpretations of the Principal Components Obtained]

This business case study is about education dataset which contain the names of various colleges, which has various details of colleges and university. To understand more about the dataset, we perform univariate analysis and multivariate analysis which gives us the understanding about the variables. From analysis we can understand the distribution of the dataset, skew, and patterns in the dataset. From multivariate analysis we can understand the correlation of variables. Inference of multivariate analysis shows we can understand multiple variables highly correlated with each other.

The scaling helps the dataset to standardize the variable in one scale. Outliers are imputed using IQR values once the values are imputed, we can perform PCA. The principal component analysis is used reduce the multicollinearity between the variables. Depending on the variance of the dataset we can reduce the PCA components. The PCA components for this business case is 5 where we could understand the maximum variance of the dataset. Using the components, we can now understand the reduced multicollinearity in the dataset.