2022

# Advanced Statistics Business Report

Akashatra Sharma

GREAT LEARING
5/15/2022

# Table of Contents



# Table Of Figures

# Executive Summary

There are basically two types of Dataset provided which gives us a lot of information. The first data set was named as "Salary Data" and it consists of the Salary for different Employees/Individuals based upon their Education as well as Occupation. Next data we were provided was named as "Education Post 12th Standard" which contains information on various Colleges & Universities. In both of the datasets, we performed different analytical techniques and statistical operations in order to get better understanding and give business implications regarding each case study of data set.

# Introduction

The purpose of this assignment is to explore the data sets. For that, we'll do different analytical & statistical operations in order to get the most of out the data.

Starting with the data sets, we had gone through the both the data sets and the briefing of the data sets are as follows :

- First Data set that is 'Salary Data' consists of means Salary of 40 Employees/Individuals having different Occupation as well as Education.
- Second data set that is 'Education post 12TH Standard' consists variety of information of 777 different University and Colleges.

# Data Dictionary

The data dictionary is mainly for the understanding of meaning of columns provided in the second data set that is 'Education Post 12TH Standard'. It is as follows :

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

**Figure 1 : Data Dictionary**

# Data Description

Description of both data sets are as follows :

**1) Salary Data**

- Education : 3 Types of Education levels ( HS-Grad, Bachelors, Doctorate)
- Occupation : 4 Types of Occupation (Adm-Clerical, Sales, Prof-Specialty, Exec-Managerial)
- Salary : Continuous Data from 50103.00 to 260151.

**2) Education Post 12$^{TH}$ Standard**

- Names : Categorical Data (University/College Name) from Abilene Christian University to York College of Pennsylvania
- Apps : Continuous Data from 81 to 48094
- Accept : Continuous Data from 72 to 26330
- Enroll : Continuous Data from 35 to 6392
- Top10perc : Continuous Data from 1% to 96%
- Top25perc : Continuous Data from 9% to 100%
- F.Undergrad : Continuous Data from 139 to 31643
- P.Undergrad : Continuous Data from 1 to 21836
- Outstate : Continuous Data from 2340 to 21700
- Room.Board : Continuous Data from 1780 to 8124
- Books : Continuous Data from 96 to 2340
- Personal : Continuous Data from 250 to 6800
- PhD : Continuous Data from 8% to 103%
- Terminal : Continuous Data from 24% to 100%
- S.F.Ratio : Continuous Data from 2.5 to 39.8
- perc.alumni : Continuous Data from 0% to 64%
- Expend : Continuous Data from 3186 to 56233
- Grad.Rate : Continuous Data from 10 to 118

# Datasets

- **Salary Data**
  Here are the first five observation of this data set :

| | 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 |

**Figure 2 : Salary Data First 5 Observations**

- **Education Post 12$^{TH}$ Standard**
  Here are the first five observation of this data set :

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

**Figure 3 : Education Post 12$^{th}$ Standard First 5 Observations**

# Data Analysis

- **Data Types**
  The data types of variables in the data set are :
  - **Salary Data**

```
Education     object
Occupation    object
Salary         int64
```

**Figure 4 : Salary Data Types**

- **Education Post 12th Standard**

```
Names           object
Apps             int64
Accept           int64
Enroll           int64
Top10perc        int64
Top25perc        int64
F.Undergrad      int64
P.Undergrad      int64
Outstate         int64
Room.Board       int64
Books            int64
Personal         int64
PhD              int64
Terminal         int64
S.F.Ratio      float64
perc.alumni      int64
Expend           int64
Grad.Rate        int64
```

**Figure 5 : Education Post 12th Standard Data Types**

# Checking For Null Values

- **Salary Data**

```
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
```

**Figure 6 : Null Values Checking (Salary Data)**

Hence, we inferred that there are zero null values in the provided data set.
Also, we note that the shape of data set is (40,3) which means that there are 40 entries and 3 columns in the data set.

- **Education Post 12th Standard**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 777 entries, 0 to 776
Data columns (total 18 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Names        777 non-null    object
 1   Apps         777 non-null    int64
 2   Accept       777 non-null    int64
 3   Enroll       777 non-null    int64
 4   Top10perc    777 non-null    int64
 5   Top25perc    777 non-null    int64
 6   F.Undergrad  777 non-null    int64
 7   P.Undergrad  777 non-null    int64
 8   Outstate     777 non-null    int64
 9   Room.Board   777 non-null    int64
 10  Books        777 non-null    int64
 11  Personal     777 non-null    int64
 12  PhD          777 non-null    int64
 13  Terminal     777 non-null    int64
 14  S.F.Ratio    777 non-null    float64
 15  perc.alumni  777 non-null    int64
 16  Expend       777 non-null    int64
 17  Grad.Rate    777 non-null    int64
dtypes: float64(1), int64(16), object(1)
memory usage: 109.4+ KB
```

**Figure 7 : Null Values Checking (Education Post 12th Standard)**

Hence, we inferred that there are zero null values in the provided data set.
Also, we note that the shape of data set is (777,18) which means that there are 777 entries and 18 columns in the data set.

# Problem 1 – Salary Data

## Problem 1A :

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

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

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

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. ANOVA test is use to determine the influence that independent variables have on the dependent variable in a regression study.

Now, as per the question, the null hypothesis and alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually are :

**a) Variable : 'Education' -**

Null hypothesis states that the mean salary of individuals in ('Salary') for every type of Education are equal.

Alternative hypothesis states that there will be an effect of "Salary" on at least one of the levels in Education. The mean salary of individual from ('Salary') for at least one category of Education are unequal.

$H_0$: $\mu_{M1} = \mu_{M2} = \mu_{M3}$

$H_0$: Mean 'Salary' for any Education for all the Employees/Individuals are equal.

$H_1$: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$ or $H_1$: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$ or $H_1$: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$ or $H_1$: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

$H_1$: At least one of the mean 'Salary' for the various categories of education are unequal.

**b) Variable :'Occupation' -**

Null hypothesis states that the mean salary of Employees/Individuals in ('Salary') for every type of Occupation are equal.

Alternative hypothesis states that there will be an effect of "Salary" on at least one of the levels in Occupation. The mean salary of individual from ('Salary') for at least one category of Occupation are unequal.

$H_0$: $\mu_{T1} = \mu_{T2} = \mu_{T3}$

$H_0$: Mean 'Salary' of Employees/Individuals of the any Occupation are equal.

$H_1$: $\mu_{T1} \neq \mu_{T2} = \mu_{T3}$ or $H_1$: $\mu_{T1} = \mu_{T2} \neq \mu_{T3}$ or $H_1$: $\mu_{T1} = \mu_{T3} \neq \mu_{T2}$ or $H_1$: $\mu_{T1} \neq \mu_{T2} \neq \mu_{T3}$

$H_1$: At least one of the means between the 'Salary' with respect to the various occupations are unequal.



**Figure 8 : Null & Alternate Hypothesis (Education & Occupation Individually)**

## Question 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.

In this data set, we found out that "Education" was an 'Independent Variable' while on the other hand "Salary" was a 'Dependent Variable'.

Before performing ANOVA test , the independent variable data type must be converted into categorical data type for getting accurate results.

The conversion of object to categorical data type is shown below :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Education   40 non-null     object
 1   Occupation  40 non-null     object
 2   Salary      40 non-null     int64
dtypes: int64(1), object(2)
memory usage: 1.1+ KB
```

**Figure 9 : Before Conversion**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Education   40 non-null     category
 1   Occupation  40 non-null     category
 2   Salary      40 non-null     int64
dtypes: category(2), int64(1)
memory usage: 864.0 bytes
```

**Figure 10 : After Conversion**

As you can see, the output before and after conversion has been done successfully.

Now, after successful conversion, we proceeded towards performing one-way ANOVA test for 'Education' with respect to 'Salary'.

We did take the Level Of Significance(alpha) = 0.05 by default.

Here is the output of the ANOVA table :

| | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| 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 |

**Figure 11 : ANOVA Table (Salary & Education)**

**Interpretation :** The P-value obtained from ANOVA analysis for 'Education' is less than **α** (0.05). Thus, we Reject the Null Hypothesis(H0), since P-value < Level of significance (P-value < 0.05).

Thus, from one-way ANOVA test, we interpret that at least one of the Mean 'Salary' of Employees/Individuals for different type of Education are unequal.

## Question 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.

As we had already converted the independent variables that are 'Education' and 'Occupation' into categorical data type, so now we need to directly perform one-way ANOVA test for 'Occupation' with respect to 'Salary'.

Heading on to ANOVA Analysis, the output is as follows :

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

**Figure 12 : ANOVA Table ( Salary & Occupation )**

**Interpretation :**

The p-value obtained from ANOVA analysis for 'Occupation' is greater than **α** (0.05). Thus, we Fail to Reject the Null Hypothesis (H0) since p-value > Level of significance (p-value > 0.05). In a simple way we can say that , we do not have enough evidence to reject the null hypothesis.

Thus, we inferred that the Mean 'Salary' of Employees/Individuals any category of Occupation are equal.

## Question 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. (Non-Graded Question)

From the above questions (1.2) and (1.3), we found out that null hypothesis is rejected in Question 1.2 which means that at least one of the Mean **'Salary'** of Employees/Individuals for different type of Education are unequal.

Now, afterwards we found out which class means are significantly different. For that, we conducted **"Tukey's HSD Test"**.

Before proceeding further we must know what is Tukey's HSD test.

The **Tukey Test** , also called **Tukey's Honest Significant Difference Test**, is a post-hoc test based on the studentized range distribution. An ANOVA test can tell you if your results are significant overall, but it won't tell you exactly where those differences lie. After you have run an ANOVA and found significant results, then you can run Tukey's HSD to find out which specific group's means (compared with each other) are different. The test compares all possible pairs of means.

Mathematically, the relevant statistic is as follows :

$$q = \frac{\bar{x}_{max} - \bar{x}_{min}}{s.e.} \text{ where } s.e. = \sqrt{MS_W/n}$$

**Figure 13 : Relevant Statistic Formula**

and $n$ = the size of each of the group samples. The statistic ($q$) has a distribution called the studentized range ($q$). The critical values for this distribution are presented in the Studentized Range q Table based on the values of $\alpha$, $k$ (the number of groups) and $df_W$. If $q > q_{crit}$ then the two means are significantly different.
In mathematical terms, this test is equivalent to :

$$\bar{x}_{max} - \bar{x}_{min} > q_{crit}\sqrt{MS_W/n}$$

**Figure 14 : Tukey Test Mean Difference Formula**

Picking the largest pairwise difference in means allows us to control the experiment-wise error rate for all possible pairwise contrasts; in fact, Tukey's HSD keeps experiment-wise $\alpha = .05$ for the largest pairwise contrast, and is conservative for all other comparisons.

Note that the statistic $q$ is related to the usual $t$ statistic by $q = \sqrt{2}\,t$. Thus we can use the following $t$ statistic :

$$t = \frac{\bar{x}_{max} - \bar{x}_{min}}{\sqrt{2 \cdot MS_W/n}}$$

**Figure 15 : Main Formula Tukey Test**

The critical value for $t$ is now given by $t_{crit} = q_{crit} / \sqrt{2}$. If $t > t_{crit}$ then we reject the null hypothesis that $H_0$: $\mu_{max} = \mu_{min}$, and similarly for other pairs.

After getting the concept of Tukey's HSD Test, we applied the same in Jupyter notebook using Multi-comparison function between 'Salary' & 'Education' and got the output as follows :

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 |

**Figure 16 : Tukey's HSD Test Result**

**Interpretation :**

- Above results from **Tukey's HSD Test** suggests that the mean difference between Doctorate and HS-grad is maximum among all the other groups which indicates that employees who have completed Doctorate has a higher Salary as compared to employees who have are only HS-grad and hence it proves that there is statistical significant difference.

- Moreover, we can further interpret that employees/individuals who have completed their Bachelors have higher Salary then employees who have only completed HS-grad as there is clear big difference in means of Salary of both Bachelors and HS-grad. Adding on, from the last pair which is of Bachelors and Doctorate we can conclude that there is minimum mean difference as compared to other groups and also employees who completed there Doctorate have higher Salary than the ones who have done only Bachelors.

- In a nutshell we can state/conclude that Education plays a significant role in deciding the salary of the employee and Occupations plays less role in determining the Salary as compared to Education variable.

The **Tukey's HSD Test** for Occupation is not done as we fail to reject the null hypothesis and so it indicates that the confidence level of 95% we take by default is very much correct and thus it clearly means that the Occupation does not effect that much as compared to Education for the determination of the Employees/Individual Salary.

# Problem 1B :

## Question 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 used interaction plot function as well as point plot function for plotting between mean 'Salary' and 'Education' while keeping 'Occupation' as hue in order to differentiate all categories.

The Interaction Plot as per the question is as follows :

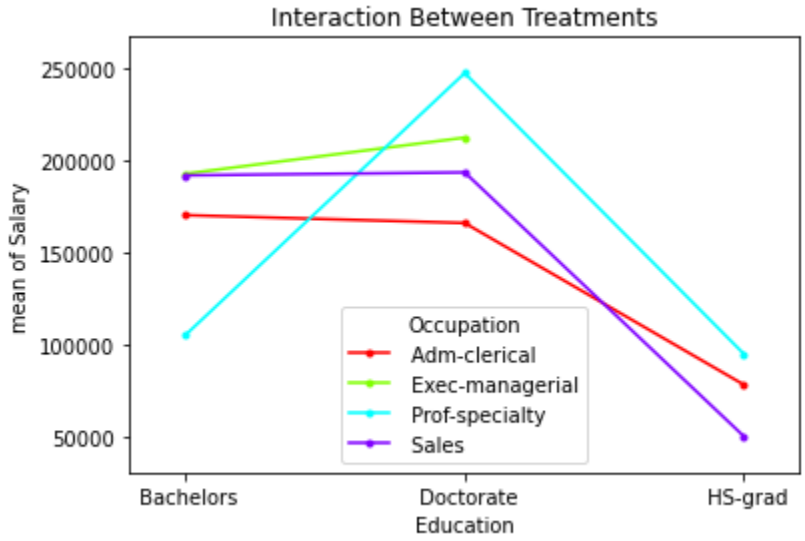



**Figure 17: Interaction Plot**

And the Point Plot for the same is as follows :

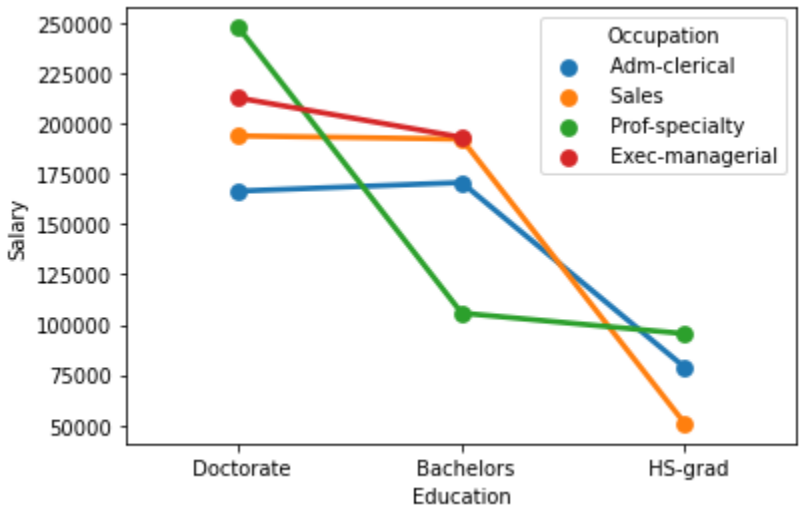



**Figure 18 : Point Plot**

**Interpretation :**

From the interaction plot as well as point plot we inferred that there was an **"2-Way Interaction Effect"** taking place between all levels and types of Education & Occupation which the significant p-value for Education*Occupation term confirms us while doing two-way ANOVA test.

- The graph shows that a Individual/Employee who has completed Doctorate have highest Salary when his occupation is Prof-Specialty while on the other hand, a person with a Doctorate with Occupation Adm-Clerical has a lower Salary as compared to above one.

- Moving on to the next one, Exec-Managerial with a Doctorate level of Education has 2nd best Salary among others.

- It is interesting to know that Sales Employees have almost equal amount of Salary despite of the fact that his education level is Doctorate or Bachelors. Also, Exec-Managerial Employee also have the same amount of Salary as that of Sales employee in Bachelors. We also noticed that Exec-Managerial employee should at least have an education level of Bachelors or higher(Doctorate) as there is no employee of Exec-managerial whose education is only HS-grad.

- If we look closely, then we could say that there is a slight increase in the graph of Adm-Clerical when the level of Education changes from Doctorate to Bachelors which means that Adm-Clerical Employees with a Bachelors have higher Salary as compared to the Adm-Clerical employees with a Doctorate level of education.

- Adding on, there is a significant decline in Salary of employees of Prof-Specialty when their education level changes from Doctorate to Bachelors. But, when the education level of Prof-Specialty Employee again changes from Bachelors to HS-grad then the decline of Salary is very less.

- Furthermore, we can say that there is a remarkable decrease in Salary of Sales Employee when his/her education levels changes from Bachelors to HS-grad and also the lowest Salary among rest of the other pairs of education & occupation. Similarly, the Salary for Adm-Clerical individual also decreases very much when his/her education levels changes from Bachelors to HS-grad.

**Conclusion :**

- It indicates that a person must have knowledge and education regarding the field of expertise in order to get a decent Occupation as well as good amount of Salary.

- From the dataset we can inferred that, among HS-grad the Employees with a Prof-Specialty earns maximum Salary as compared to the rest while Sales Employees earns the lowest as compared to other HS-grad Employees.

- On the other hand, when we look at the Bachelors Employees, then Exec-Managerial & Sales Employees/Individuals get maximum Salary and Prof-Specialty with a Bachelors get the lowest Salary among other Bachelors Employees.

- And the final one, we can clearly tell that Prof-Specialty Employees gets the maximum Salary among all types of Education & Occupation. Next to it is the Exec-Managerial Employees with the Doctorate which stands just after Prof-Specialty with Doctorate. And the Occupation which gets minimum Salary even with a Doctorate level of Education is Adm-Clerical.

## Question 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?

Before, conducting Two-way ANOVA test, we stated the null and alternate hypothesis for 'Education' & 'Occupation' that is 'Education*Occupation'.

Also, we take the Level Of Significance = 0.05 by default.

Here are the statements of null and alternate hypothesis :

$H_0$: $\mu_{M1} = \mu_{M2} = \mu_{M3}$

$H_0$: There is no Interaction Effect between the treatments(Education & Occupation).

$H_1$: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$ or $H_1$: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$ or $H_1$: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$ or $H_1$: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

$H_1$: There is an Interaction Effect between the treatments(Education & Occupation).

**Level of significance:**

$\alpha = 0.05$

**Figure 19 : Null & Alternate Hypothesis**

In two-way, we used the interaction effect for 'Education' & 'Occupation' that is 'Education*Occupation'. We performed ANOVA test and got the output table as follows :

| | 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 |

**Figure 20 : Two-way ANOVA Table (With Interaction Effect)**

**Interpretation:**

The P-Value obtained from ANOVA for Education is statistically significant (P-Value < 0.05) & for Occupation is not statistically significant (P-Value > 0.05) but the Interaction effect for both Education and Occupation (Education*Occupation) together is statistically significant (P-Value < 0.05).

We conclude that the type of levels of Education separately significantly affect the Salary variable while the types of levels of Occupation do not significantly affect the Salary variable but the Interaction Effect of both (Education & Occupation) significantly affects the Salary outcome, that is Interaction Effect of both (Education*Occupation) plays an important role in determining the Salary of an Employee/Individual .

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

In this case study, after conducting the ANOVA test i.e. one-way ANOVA & two-way ANOVA on both variables that is 'Education' & 'Occupation' we conclude that the type of levels of Education separately significantly affect the Salary variable because (P-Value < 0.05) while the types of levels of Occupation do not significantly affect the Salary variable because (P-Value > 0.05) but the Interaction of both (Education*Occupation) significantly affects the Salary outcome .

**Conclusion:**

- It indicates that a Person/Individual must have good education as well a good field of expertise in order to get a high profile Occupation as well as good amount of Salary.

- From the dataset we can inferred that, among HS-grad the Employee with a Prof-Specialty earns maximum Salary as compared to the rest Hs-Grad Employees while Sales Employees earns the lowest as compared to other HS-grad employees.

- On the other hand, when we look at the Bachelors Employees, then Exec-Managerial & Sales Employees/Individuals get maximum Salary and Prof-Specialty with a Bachelors get the lowest Salary among other Bachelors Employees.

- And the final one, we can clearly tell that Prof-Specialty Employees gets the maximum Salary among all types of Education & Occupation. Next to it is the Exec-Managerial Employees with the Doctorate which stands just after Prof-Specialty Employees with Doctorate. And the Occupation which gets minimum Salary even with a Doctorate level of Education is Adm-Clerical.

At last, we could say that Higher Education is directly proportional to the Occupation we desire and according to that the Salary is to be determined.

# Problem 2 – Education Post 12th Standard

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

At first we loaded the Data Dictionary for understanding of column name for data set "Education Post 12th Standard".

The Data Dictionary is as follows :

**Data Dictionary**

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

**Figure 21 : Data Dictionary**

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

## EDA

**Exploratory Data Analysis (EDA)** is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

After getting a brief understanding of what is EDA, we did analyze the data and here is what we found

## Descriptive Statistics :

| | 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 |

**Figure 22 : Descriptive Statistics**

## Check for Duplicate Data

We checked for that and the output is as follows :

```
Number of Duplicated Observations in the data set are 0
Hence, no observations are being duplicated in the data set
```

Hence, zero observations are duplicated in the provided dataset.

## Univariate Analysis

For Univariate analysis, we plotted a Distribution plot and a Boxplot for each column provided in the data set except for ‘Names’ column as it was a categorical data.

The Distribution plot was used for univariant set of observations and visualizes it through a histogram i.e. only one observation and hence we choose one particular column of the dataset.

And, Boxplot was used as a measure of how well the data is distributed in a data set. It divides the data set into three quartiles. This graph represents the minimum, maximum, median, first quartile and third quartile in the data and also shows us whether there are outliers or not.

Here the Distribution Plot and Boxplot for each column :

**Apps :**

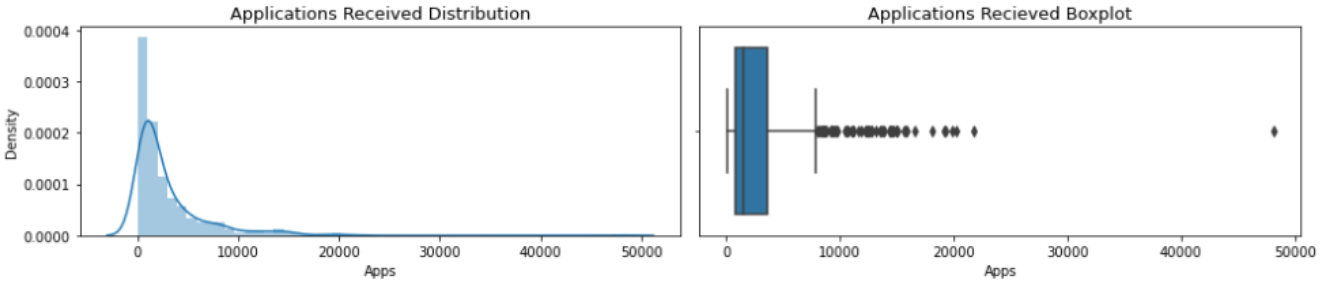



**Figure 23 : Distplot & Boxplot (Apps)**

The 'Apps' variables seems to have outliers as shown in the Boxplot. The Distplot shows us that the 'Apps' variable are positively right skewed. It also indicates that the majority of Application Received to the Universities/Colleges are ranging from 776 to 3624.

**Accept :**

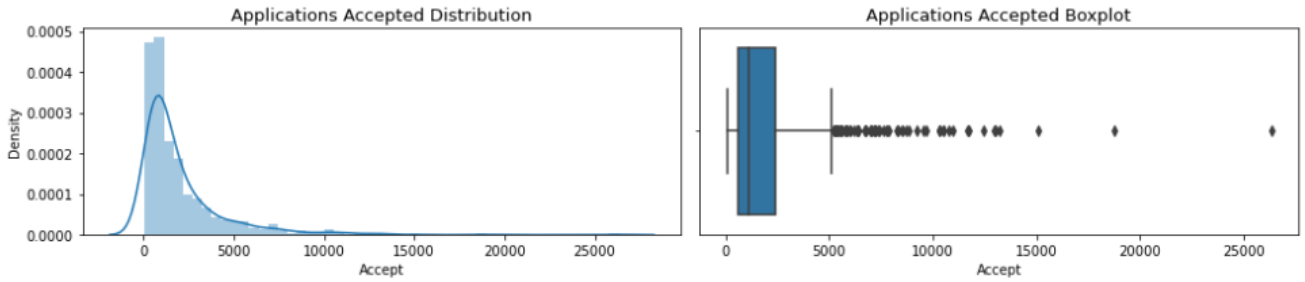



**Figure 24 : Distplot & Boxplot(Accept)**

The 'Accept' variables showed outliers as per the Boxplot. Furthermore, we clearly noticed that Inter-Quartile Range of applications received by Universities/Colleges are ranging from 604 to 2424. Also from Distplot , we noticed that 'Accept' variable a positively right skewed.

**Enroll :**

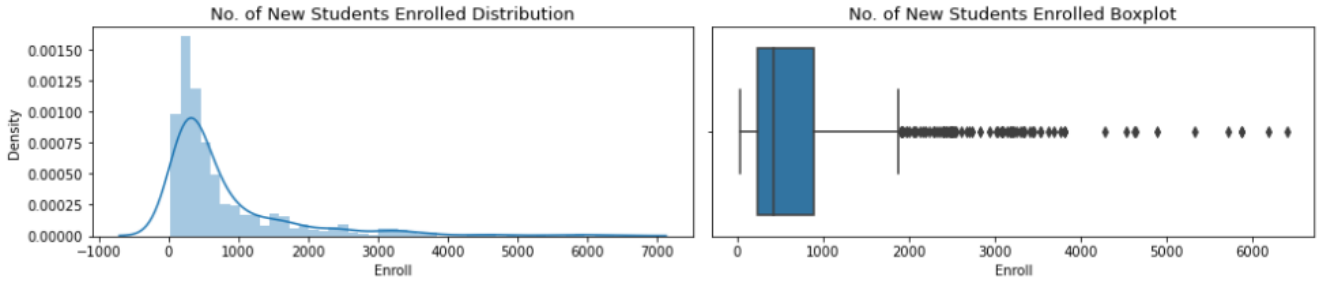



**Figure 25 : Distplot & Boxplot (Enroll)**

The 'Enroll' variables showed many outliers as per the Boxplot which indicated that there are many no. of new students who enrolled in respective Universities/Colleges more than the maximum(As per the 5 Point Summary) Adding on, with the help of Distplot, we noticed that 'Enroll' is positively right skewed. Also, the Inter-Quartile Range of new students enrolled in respective Universities/Colleges are ranging from 242 to 942.

**Top10perc :**

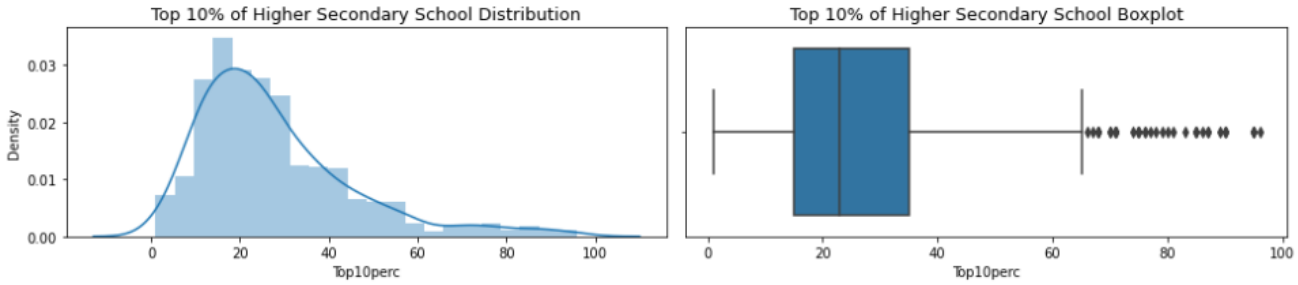



**Figure 26 : Distplot & Boxplot (Top25perc)**

The 'Top10perc' variable showed many outliers as shown in the Boxplot which means that many percentage of new students with Top 10% of Higher Secondary School are outside the maximum value(as per the 5 Point Summary). The Distplot showed us that 'Top10perc' variable is slightly positively right skewed.

**Top25perc :**

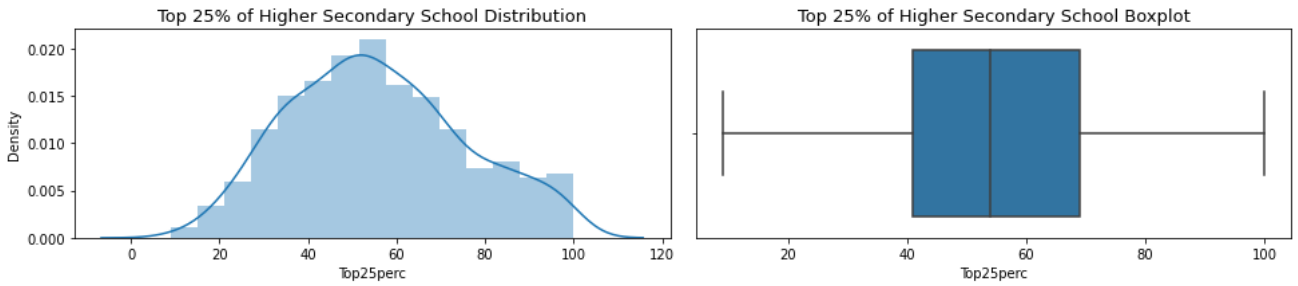
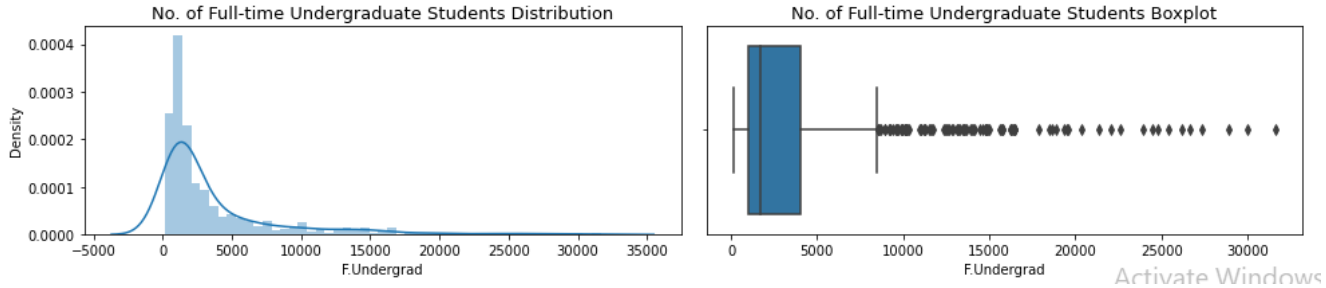



**Figure 27 : Distplot & Boxplot (Top25perc)**

The 'Top25perc' variable showed no outliers as per the Boxplot which means all values were laying inside the range from minimum to maximum. As per the Distplot, we got to know that 'Top25perc' was almost a normally distribution (neither left nor right skewed) but as we saw to the right side of the Distplot, we saw there were values present in high amount which breaks the normal distribution curve but it was close to normal distribution.

**F.Undergrad :**




The 'F.Undergrad' variable showed many outliers in the Boxplot which indicated that there are many Full-time Undergraduate Students which exceed the maximum value (from 5 point summary). As per we go to the Distplot, we got to know that 'F.Undergrad' was a positively right skewed distribution.

**P.Undergrad :**

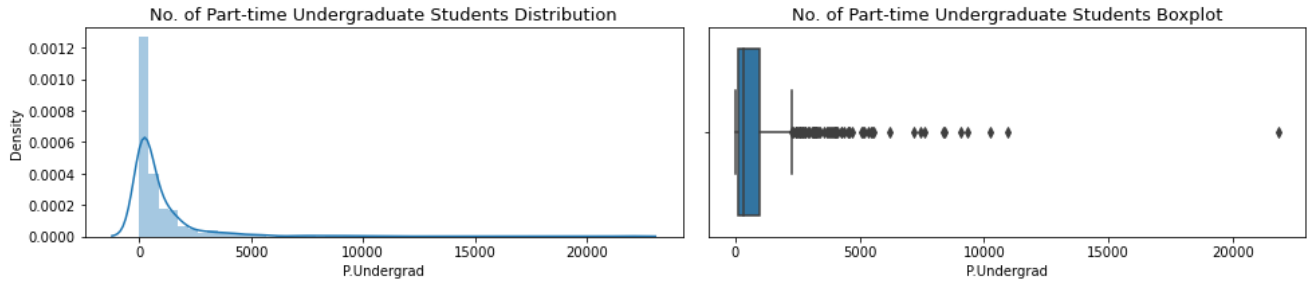



**Figure 28 : Distplot & Boxplot (P.Undergrad)**

The 'P.Undergrad' variable showed many outliers in the Boxplot which indicated that there were many Part-time Undergraduate students which exceed the maximum value (from 5 point summary). As we go to the Distplot, we got to know that 'P.Undergrad' was positively right skewed distribution.

**Outstate :**

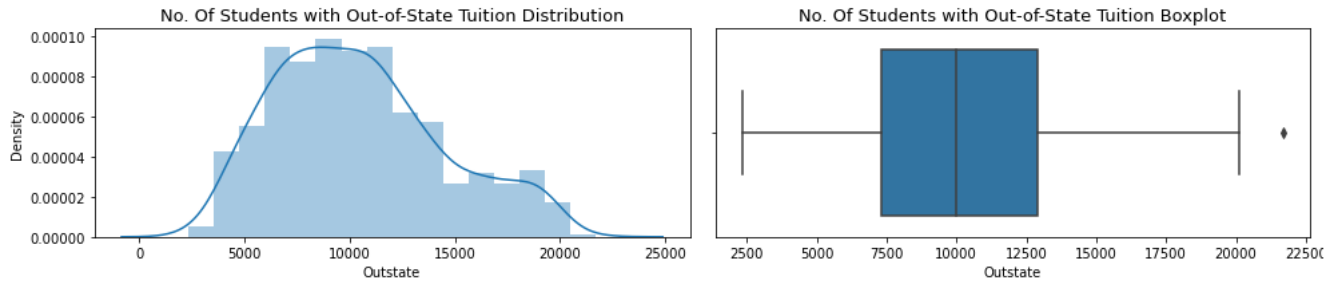



**Figure 29 : Distplot & Boxplot(Outstate)**

The 'Outstate' variable showed only 1 outlier as per the Boxplot which means only 1 value was laying outside the maximum range (as per the 5 Point summary). As per the Distplot, we got to know that 'Outstate' was almost a normally distribution (neither left nor right skewed) but as we saw to the right side of the Distplot, we saw there were values present in high amount which breaks the normal distribution curve but it was close to normal distribution.

**Room.Board :**

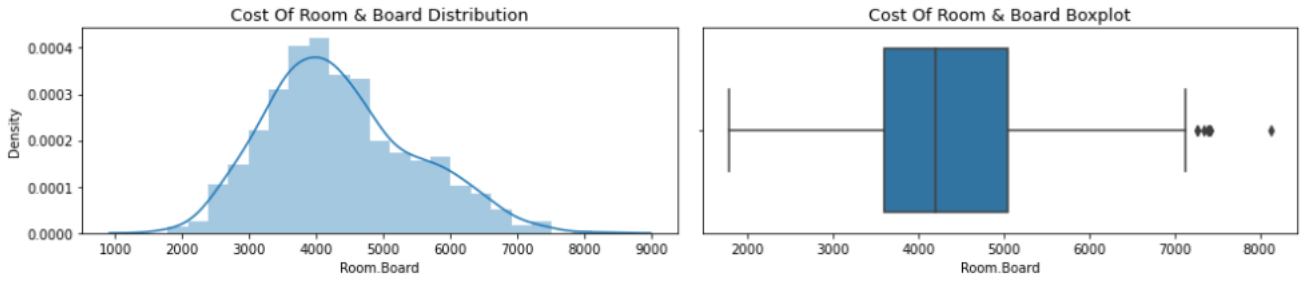



**Figure 30 : Distplot & Boxplot(Room.Board)**

The 'Room.Board' variable showed some outliers as per the Boxplot which means the cost of room and board were higher than the maximum range (as per the 5 Point summary). As per the Distplot, we got to know that 'Room.Board' was almost a normally distribution (neither left nor right skewed) but as we

saw to the right side of the Distplot, we saw there were values present in high amount which breaks the normal distribution curve but it was very close to normal distribution.

**Books :**

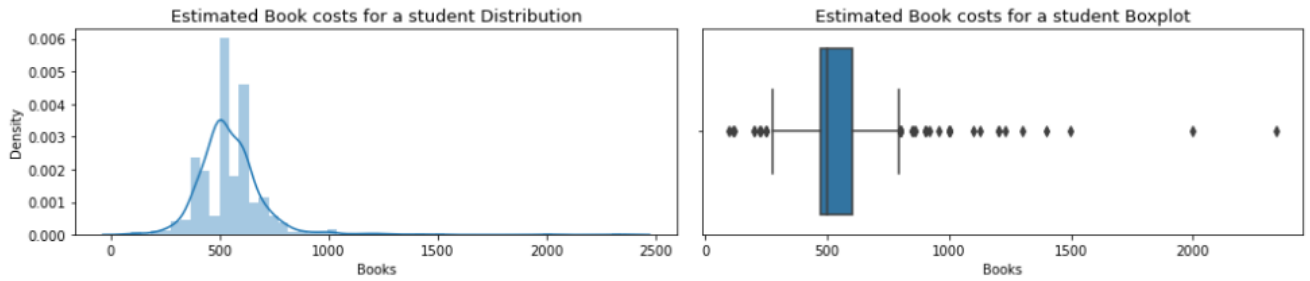



**Figure 31 : Distplot & Boxplot(Books)**

The 'Books' variable showed many outliers as per the Boxplot before the minimum and maximum range(As per the 5 Point Summary). As per the Distplot, we inferred that 'Books' was a positively right skewed distribution .

**Personal :**

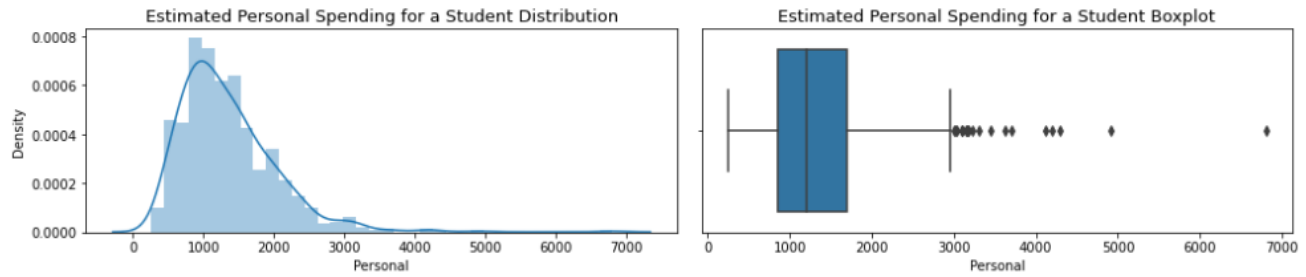



**Figure 32 : Distplot & Boxplot(Personal)**

The 'Personal' variable showed many outliers in the Boxplot which indicated that there were many Personal Spending per student that exceeded the maximum value (from 5 point summary). As we go to the Distplot, we got to know that 'Personal' was positively right skewed distribution.

**PhD :**

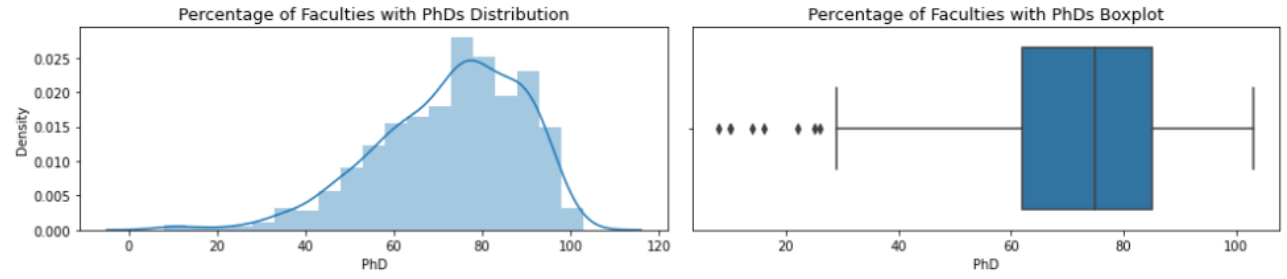



**Figure 33 : Distplot & Boxplot (PhD)**

The 'PhD' variable showed many outliers were present in the Boxplot which indicated that there were very less Percentage of Faculties with PhD's that were below minimum value (from 5 point summary). As we go to the Distplot, we got to know that 'PhD' was negatively left skewed distribution.

**Terminal :**

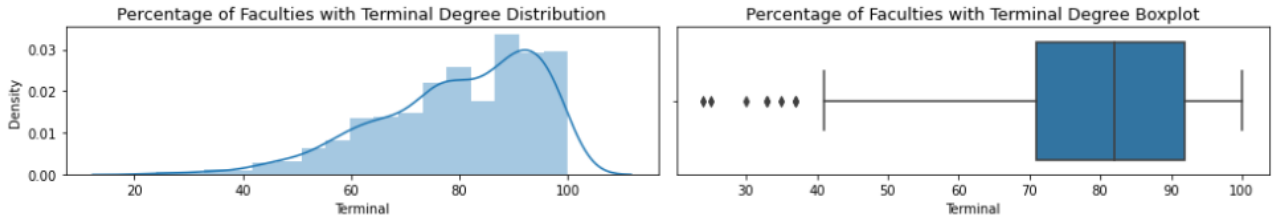



**Figure 34 : Distplot & Boxplot (Terminal)**

The 'Terminal' variable showed many outliers were present in the Boxplot which indicated that there were very less Percentage of Faculties with terminal degree that were below minimum value (from 5 point summary). As we go to the Distplot, we got to know that 'Terminal' was negatively left skewed distribution.

**S.F.Ratio :**

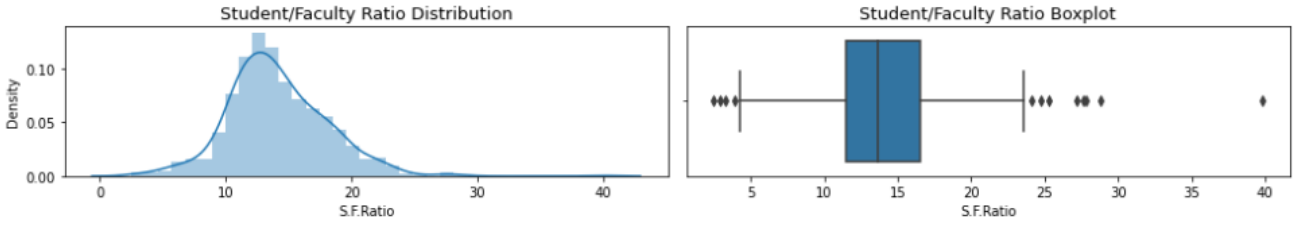



**Figure 35 : Distplot & Boxplot( S.F.Ratio )**

The 'S.F.Ratio' variable showed many outliers as per the Boxplot before the minimum and maximum range which means the Student/Faculty Ratio is varying very much to the left and at the right side of Boxplot (As per the 5 Point Summary). As per the Distplot, we inferred that 'S.F.Ratio' was a positively right skewed distribution .

**Perc.alumni**

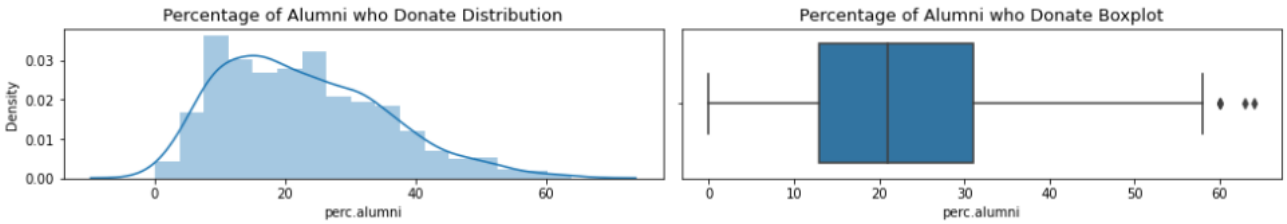



**Figure 36 : Distplot & Boxplot (perc.alumni)**

The 'perc.alumni' variable showed some outliers in the Boxplot which indicates that there were some percentage of alumni who donated that exceeded the maximum value (from 5 point summary). As we go to the Distplot, we got to know that 'perc.alumni' was positively right skewed distribution.

**Expend :**

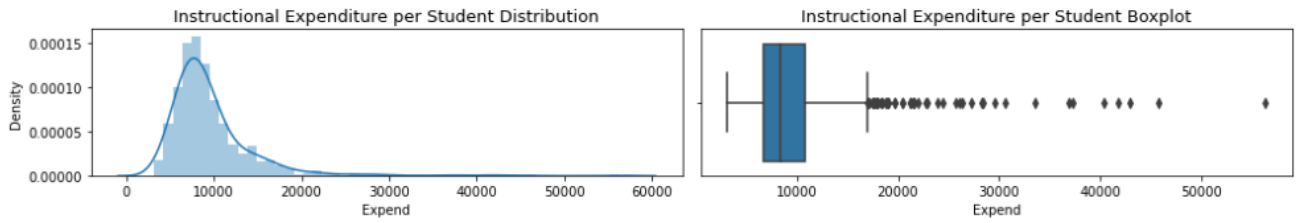



**Figure 37 : Distplot & Boxplot (Expend)**

The 'Expend' variable showed many outliers in the Boxplot which indicated that there were many Instructional Expenditure per student that exceeded the maximum value (from 5 point summary). As we go to the Distplot, we got to know that 'Expend' was positively right skewed distribution.

**Grad.Rate :**

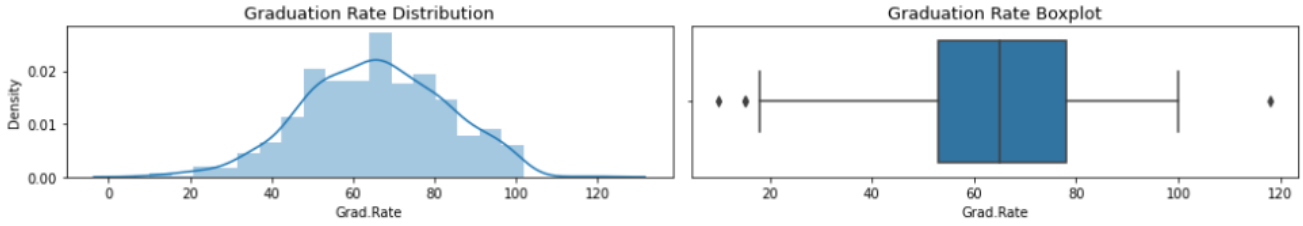



**Figure 38 : Distplot & Boxplot (Grad.Rate)**

The 'Grad.Rate' variable showed some outliers as per the Boxplot before the minimum and maximum range which indicated the Graduation rate varied very much to the left and at the right side of Boxplot (As per the 5 Point Summary). As per the Distplot, we inferred that 'Grad.Rate' was a slightly negatively left skewed distribution .

## Multivariate Analysis

**Multivariate analysis** (**MVA**) is a Statistical procedure for analysis of data involving more than one type of measurement or observation. It may also mean solving problems where more than one dependent variable is analyzed simultaneously with other variables.

Multivariate analysis is one of the most useful methods to determine relationships and analyse patterns among large sets of data. It is particularly effective in minimizing bias if a structured study design is employed. However, the complexity of the technique makes it a less sought-out model for novice research enthusiasts. Therefore, although the process of designing the study and interpretation of results is a tedious one, the techniques stand out in finding the relationships in complex

In this we found out the correlation matrix of the original dataset as follows :

| | Apps | Accept | Enroll | Top10perc | Top25perc | F.Undergrad | P.Undergrad | Outstate | Room.Board | Books | Personal |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Apps** | 1.000 | 0.943 | 0.847 | 0.339 | 0.352 | 0.814 | 0.398 | 0.050 | 0.165 | 0.133 | 0.179 |
| **Accept** | 0.943 | 1.000 | 0.912 | 0.192 | 0.247 | 0.874 | 0.441 | -0.026 | 0.091 | 0.114 | 0.201 |
| **Enroll** | 0.847 | 0.912 | 1.000 | 0.181 | 0.227 | 0.965 | 0.513 | -0.155 | -0.040 | 0.113 | 0.281 |
| **Top10perc** | 0.339 | 0.192 | 0.181 | 1.000 | 0.892 | 0.141 | -0.105 | 0.562 | 0.371 | 0.119 | -0.093 |
| **Top25perc** | 0.352 | 0.247 | 0.227 | 0.892 | 1.000 | 0.199 | -0.054 | 0.489 | 0.331 | 0.116 | -0.081 |
| **F.Undergrad** | 0.814 | 0.874 | 0.965 | 0.141 | 0.199 | 1.000 | 0.571 | -0.216 | -0.069 | 0.116 | 0.317 |
| **P.Undergrad** | 0.398 | 0.441 | 0.513 | -0.105 | -0.054 | 0.571 | 1.000 | -0.254 | -0.061 | 0.081 | 0.320 |
| **Outstate** | 0.050 | -0.026 | -0.155 | 0.562 | 0.489 | -0.216 | -0.254 | 1.000 | 0.654 | 0.039 | -0.299 |
| **Room.Board** | 0.165 | 0.091 | -0.040 | 0.371 | 0.331 | -0.069 | -0.061 | 0.654 | 1.000 | 0.128 | -0.199 |
| **Books** | 0.133 | 0.114 | 0.113 | 0.119 | 0.116 | 0.116 | 0.081 | 0.039 | 0.128 | 1.000 | 0.179 |
| **Personal** | 0.179 | 0.201 | 0.281 | -0.093 | -0.081 | 0.317 | 0.320 | -0.299 | -0.199 | 0.179 | 1.000 |
| **PhD** | 0.391 | 0.356 | 0.331 | 0.532 | 0.546 | 0.318 | 0.149 | 0.383 | 0.329 | 0.027 | -0.011 |
| **Terminal** | 0.369 | 0.338 | 0.308 | 0.491 | 0.525 | 0.300 | 0.142 | 0.408 | 0.375 | 0.100 | -0.031 |
| **S.F.Ratio** | 0.096 | 0.176 | 0.237 | -0.385 | -0.295 | 0.280 | 0.233 | -0.555 | -0.363 | -0.032 | 0.136 |
| **perc.alumni** | -0.090 | -0.160 | -0.181 | 0.455 | 0.418 | -0.229 | -0.281 | 0.566 | 0.272 | -0.040 | -0.286 |
| **Expend** | 0.260 | 0.125 | 0.064 | 0.661 | 0.527 | 0.019 | -0.084 | 0.673 | 0.502 | 0.112 | -0.098 |
| **Grad.Rate** | 0.147 | 0.067 | -0.022 | 0.495 | 0.477 | -0.079 | -0.257 | 0.571 | 0.425 | 0.001 | -0.269 |

**Figure 39 : Correlation Matrix (Original Data - Part 1)**

| PhD | Terminal | S.F.Ratio | perc.alumni | Expend | Grad.Rate |
|---|---|---|---|---|---|
| 0.391 | 0.369 | 0.096 | -0.090 | 0.260 | 0.147 |
| 0.356 | 0.338 | 0.176 | -0.160 | 0.125 | 0.067 |
| 0.331 | 0.308 | 0.237 | -0.181 | 0.064 | -0.022 |
| 0.532 | 0.491 | -0.385 | 0.455 | 0.661 | 0.495 |
| 0.546 | 0.525 | -0.295 | 0.418 | 0.527 | 0.477 |
| 0.318 | 0.300 | 0.280 | -0.229 | 0.019 | -0.079 |
| 0.149 | 0.142 | 0.233 | -0.281 | -0.084 | -0.257 |
| 0.383 | 0.408 | -0.555 | 0.566 | 0.673 | 0.571 |
| 0.329 | 0.375 | -0.363 | 0.272 | 0.502 | 0.425 |
| 0.027 | 0.100 | -0.032 | -0.040 | 0.112 | 0.001 |
| -0.011 | -0.031 | 0.136 | -0.286 | -0.098 | -0.269 |
| 1.000 | 0.850 | -0.131 | 0.249 | 0.433 | 0.305 |
| 0.850 | 1.000 | -0.160 | 0.267 | 0.439 | 0.290 |
| -0.131 | -0.160 | 1.000 | -0.403 | -0.584 | -0.307 |
| 0.249 | 0.267 | -0.403 | 1.000 | 0.418 | 0.491 |
| 0.433 | 0.439 | -0.584 | 0.418 | 1.000 | 0.390 |
| 0.305 | 0.290 | -0.307 | 0.491 | 0.390 | 1.000 |



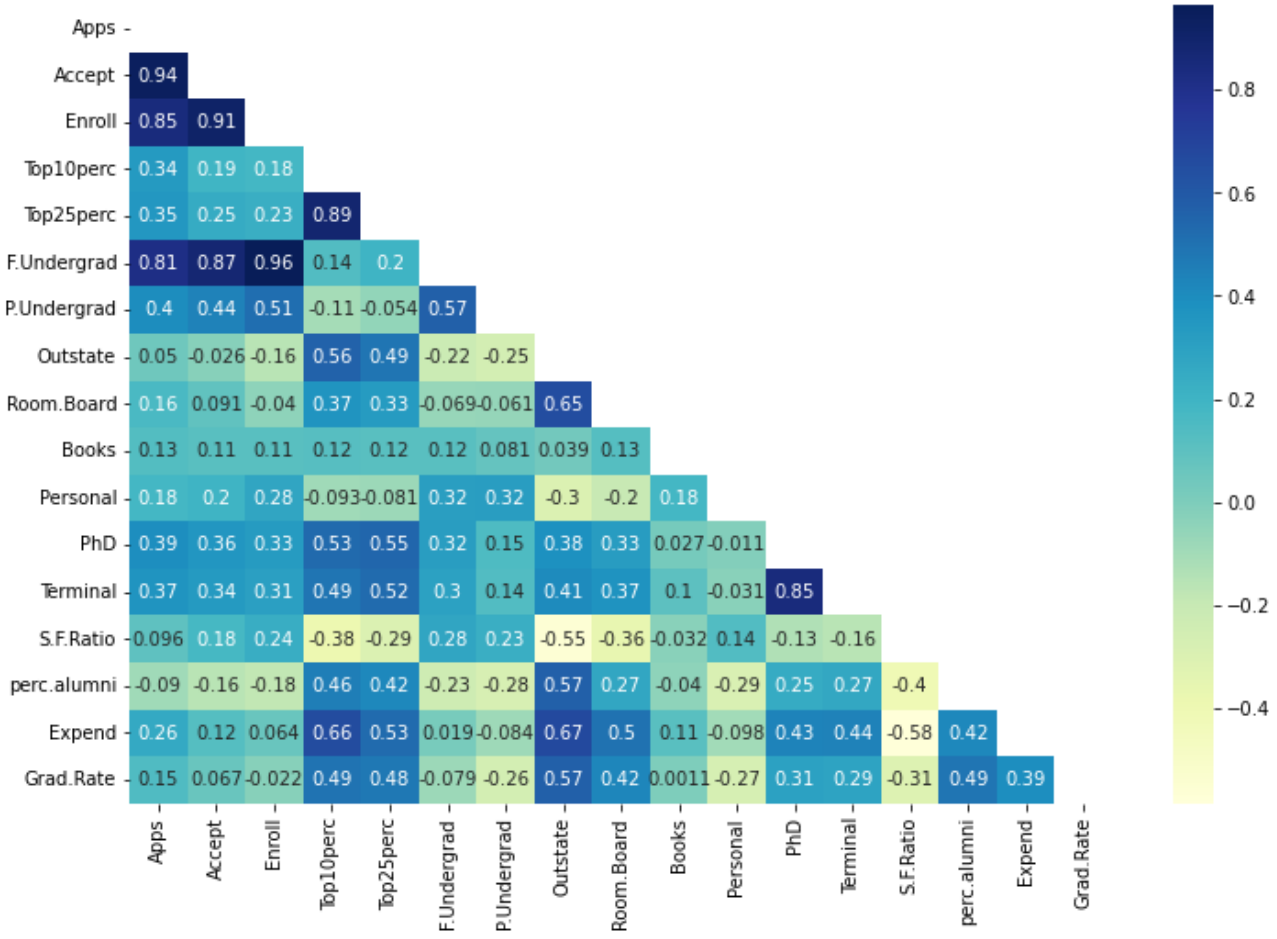
**Figure 40 : Correlation Matrix (Original Data - Part 2)**

After building the Correlation Matrix, we got the heatmap based on the correlation matrix of the original data set.

A **Correlation Heatmap** is a rectangular representation of data and it repeats the same data description twice because the categories are repeated on both axis for computing analysis. Hence, the same result is obtained twice. A correlation heatmap that presents data only once without repetition that is categories are correlated only once is known as a **Triangle Correlation Heatmap**.

Since data is symmetric across the diagonal from left-top to right bottom the idea of obtaining a triangle correlation heatmap is to remove data above it so that it is depicted only once. The elements on the diagonal are the parts where categories of the same type correlate.

**Triangle Correlation Heatmap** is as follows :




**Pairplot :**

**Pairplot** function allows the users to create an axis grid via which each numerical variable stored in data is shared across the X- and Y-axis in the structure of columns and rows. We can create the Scatter plots in order to display the pairwise relationships in addition to the distribution plot displaying the data distribution in the column diagonally.

The **pairplot()** function can also be used to showcase the subset of variables, or we can plot different types of variables on rows and columns.

The Pairplot for the original data set is as follows :

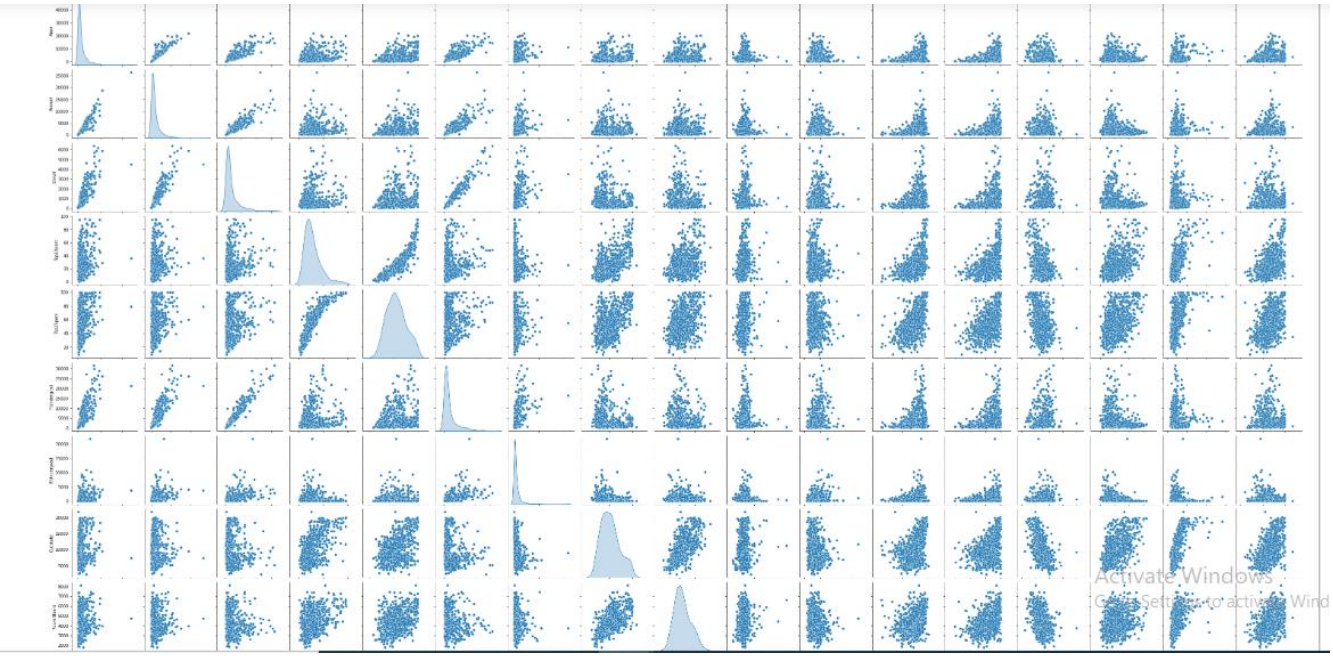


**Figure 41 : Pairplot (Original Data Part 1)**

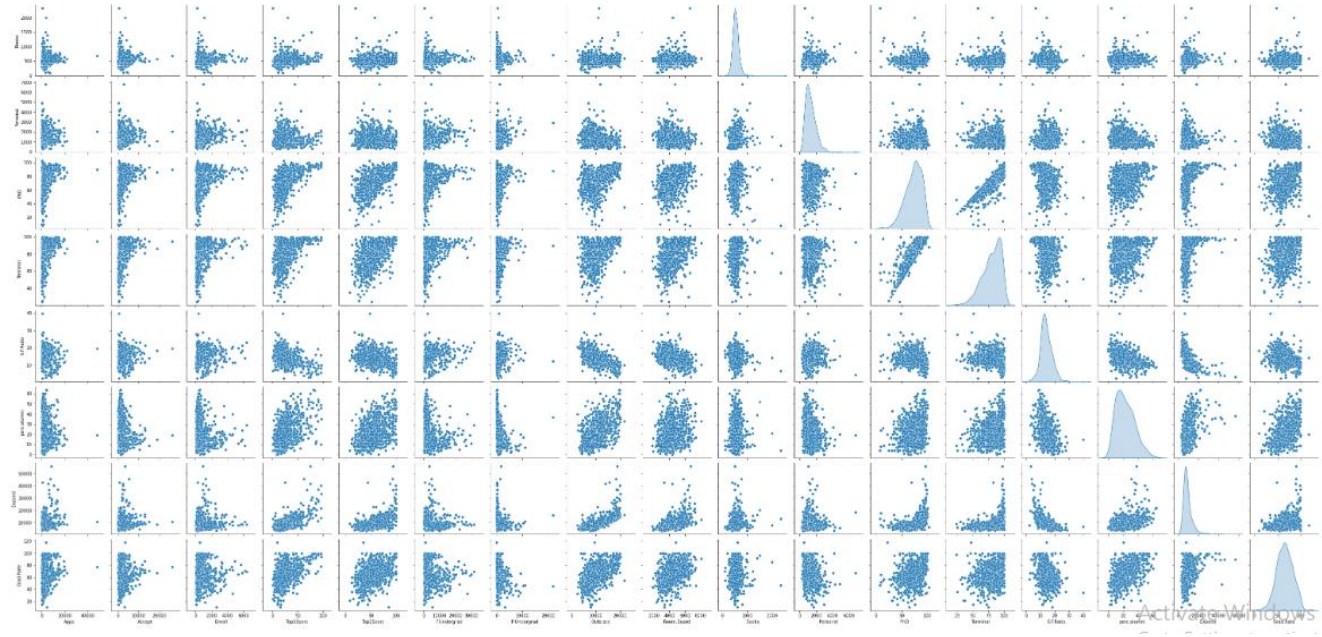


**Figure 42 : Pairplot (Original Data Part 2)**

## Interpretation :

- Firstly, we inferred that 94% of students whose application were received were accepted in their respective University/College.
- Secondly, we noted that 91% of students whose applications were accepted, got enrolled into that University/College.
- On the other hand, only 18% students were there who were in Top 10% of Higher Secondary School while only 23% students were there who were in Top 25% of Higher Secondary School.

- It was interesting to know that 96% of students who got enrolled are Full-time Undergraduate Students.
- We also noticed that 85% of Faculties who have Terminal degree also have PhD in their education.
- Moreover, we inferred that, 22% of Full-time Undergraduate students were giving out of state tuition fees for their respective University/College.

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

We dropped the 'Names' column for performing scaling & PCA as PCA should be performed only on continuous variables. We dropped 'Names' column and export the data into a new DataFrame .

Here it is :

| | Apps | Accept | Enroll | Top10perc | Top25perc | F.Undergrad | P.Undergrad | Outstate | Room.Board | Books | Personal | PhD | Terminal | S.F.Ratio | perc.alumni | Exp |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1660 | 1232 | 721 | 23 | 52 | 2885 | 537 | 7440 | 3300 | 450 | 2200 | 70 | 78 | 18.1 | 12 | 7 |
| 1 | 2186 | 1924 | 512 | 16 | 29 | 2683 | 1227 | 12280 | 6450 | 750 | 1500 | 29 | 30 | 12.2 | 16 | 10 |
| 2 | 1428 | 1097 | 336 | 22 | 50 | 1036 | 99 | 11250 | 3750 | 400 | 1165 | 53 | 66 | 12.9 | 30 | 8 |
| 3 | 417 | 349 | 137 | 60 | 89 | 510 | 63 | 12960 | 5450 | 450 | 875 | 92 | 97 | 7.7 | 37 | 19 |
| 4 | 193 | 146 | 55 | 16 | 44 | 249 | 869 | 7560 | 4120 | 800 | 1500 | 76 | 72 | 11.9 | 2 | 10 |

**Figure 43 : New Dataframe (Without 'Names' Column)**

Now the no. of columns have been reduced to 17.

Now, heading towards the question, 'Yes', **Scaling** is very important for PCA in this case. It's because PCA is affected by scale, so we need to scale the features in the data before applying PCA.

We use Standard Scaler from Scikit Learn to standardize the dataset features onto unit scale (mean = 0 and standard deviation = 1) which is a requirement for the optimal performance of many Machine Learning etc.

**Scaling the Data**

| | Apps | Accept | Enroll | Top10perc | Top25perc | F.Undergrad | P.Undergrad | Outstate | Room.Board | Books | Personal | PhD | Terminal | S.F.Ratio | perc.alumni |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | -0.35 | -0.32 | -0.06 | -0.26 | -0.19 | -0.17 | -0.21 | -0.75 | -0.96 | -0.60 | 1.27 | -0.16 | -0.12 | 1.01 | -0.87 |
| 1 | -0.21 | -0.04 | -0.29 | -0.66 | -1.35 | -0.21 | 0.24 | 0.46 | 1.91 | 1.22 | 0.24 | -2.68 | -3.38 | -0.48 | -0.54 |
| 2 | -0.41 | -0.38 | -0.48 | -0.32 | -0.29 | -0.55 | -0.50 | 0.20 | -0.55 | -0.91 | -0.26 | -1.20 | -0.93 | -0.30 | 0.59 |
| 3 | -0.67 | -0.68 | -0.69 | 1.84 | 1.68 | -0.66 | -0.52 | 0.63 | 1.00 | -0.60 | -0.69 | 1.19 | 1.18 | -1.62 | 1.15 |
| 4 | -0.73 | -0.76 | -0.78 | -0.66 | -0.60 | -0.71 | 0.01 | -0.72 | -0.22 | 1.52 | 0.24 | 0.20 | -0.52 | -0.55 | -1.68 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 772 | -0.21 | -0.21 | -0.26 | -1.34 | -1.51 | -0.13 | 0.77 | -0.91 | -0.42 | -0.30 | -0.21 | -0.78 | -1.34 | 1.75 | -0.71 |
| 773 | -0.27 | -0.09 | -0.09 | -0.20 | -0.44 | -0.18 | 0.17 | 0.27 | 0.55 | 0.31 | -0.13 | 0.02 | -0.32 | -0.20 | 0.67 |
| 774 | -0.23 | -0.04 | -0.09 | 0.37 | 0.26 | -0.19 | -0.45 | -0.88 | -0.14 | 0.41 | -0.83 | -0.35 | -0.32 | 0.08 | -0.22 |
| 775 | 1.99 | 0.18 | 0.58 | 3.83 | 2.18 | 0.31 | -0.51 | 2.34 | 1.96 | 0.49 | 1.14 | 1.43 | 1.11 | -2.10 | 2.12 |
| 776 | -0.00 | -0.07 | -0.10 | 0.03 | 0.36 | -0.15 | 0.57 | -1.36 | -0.73 | -0.30 | -0.13 | 0.14 | -0.32 | 1.01 | 0.42 |

777 rows × 17 columns



**Figure 44 : Scaled Data (Part 1)**

| Expend | Grad.Rate |
|---|---|
| -0.50 | -0.32 |
| 0.17 | -0.55 |
| -0.18 | -0.67 |
| 1.79 | -0.38 |
| 0.24 | -2.94 |
| ... | ... |
| -0.99 | -1.48 |
| -0.09 | 1.02 |
| -0.26 | -0.96 |
| 5.89 | 1.95 |
| -0.99 | 1.95 |

**Figure 45 : Scaled Data (Part 2)**

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

We build the covariance matrix and correlation matrix based on the scaled data. They are as follows :

```
Covariance Matrix
%s [[ 1.    0.94  0.85  0.34  0.35  0.82  0.4   0.05  0.17  0.13  0.18  0.39
   0.37  0.1  -0.09  0.26  0.15]
 [ 0.94  1.    0.91  0.19  0.25  0.88  0.44 -0.03  0.09  0.11  0.2   0.36
   0.34  0.18 -0.16  0.12  0.07]
 [ 0.85  0.91  1.    0.18  0.23  0.97  0.51 -0.16 -0.04  0.11  0.28  0.33
   0.31  0.24 -0.18  0.06 -0.02]
 [ 0.34  0.19  0.18  1.    0.89  0.14 -0.11  0.56  0.37  0.12 -0.09  0.53
   0.49 -0.39  0.46  0.66  0.5 ]
 [ 0.35  0.25  0.23  0.89  1.    0.2  -0.05  0.49  0.33  0.12 -0.08  0.55
   0.53 -0.3   0.42  0.53  0.48]
 [ 0.82  0.88  0.97  0.14  0.2   1.    0.57 -0.22 -0.07  0.12  0.32  0.32
   0.3   0.28 -0.23  0.02 -0.08]
 [ 0.4   0.44  0.51 -0.11 -0.05  0.57  1.   -0.25 -0.06  0.08  0.32  0.15
   0.14  0.23 -0.28 -0.08 -0.26]
 [ 0.05 -0.03 -0.16  0.56  0.49 -0.22 -0.25  1.    0.66  0.04 -0.3   0.38
   0.41 -0.56  0.57  0.67  0.57]
 [ 0.17  0.09 -0.04  0.37  0.33 -0.07 -0.06  0.66  1.    0.13 -0.2   0.33
   0.38 -0.36  0.27  0.5   0.43]
 [ 0.13  0.11  0.11  0.12  0.12  0.12  0.08  0.04  0.13  1.    0.18  0.03
   0.1  -0.03 -0.04  0.11  0.  ]
 [ 0.18  0.2   0.28 -0.09 -0.08  0.32  0.32 -0.3  -0.2   0.18  1.   -0.01
  -0.03  0.14 -0.29 -0.1  -0.27]
 [ 0.39  0.36  0.33  0.53  0.55  0.32  0.15  0.38  0.33  0.03 -0.01  1.
   0.85 -0.13  0.25  0.43  0.31]
 [ 0.37  0.34  0.31  0.49  0.53  0.3   0.14  0.41  0.38  0.1  -0.03  0.85
   1.   -0.16  0.27  0.44  0.29]
 [ 0.1   0.18  0.24 -0.39 -0.3   0.28  0.23 -0.56 -0.36 -0.03  0.14 -0.13
  -0.16  1.   -0.4  -0.58 -0.31]
 [-0.09 -0.16 -0.18  0.46  0.42 -0.23 -0.28  0.57  0.27 -0.04 -0.29  0.25
   0.27 -0.4   1.    0.42  0.49]
 [ 0.26  0.12  0.06  0.66  0.53  0.02 -0.08  0.67  0.5   0.11 -0.1   0.43
   0.44 -0.58  0.42  1.    0.39]
 [ 0.15  0.07 -0.02  0.5   0.48 -0.08 -0.26  0.57  0.43  0.   -0.27  0.31
   0.29 -0.31  0.49  0.39  1.  ]]
```

**Figure 46 : Covariance Matrix (Scaled Data)**

Correlation Matrix :

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



**Figure 47 : Correlation Matrix (Part 1)**

| perc.alumni | Expend | Grad.Rate |
|---|---|---|
| -0.09 | 0.26 | 0.15 |
| -0.16 | 0.12 | 0.07 |
| -0.18 | 0.06 | -0.02 |
| 0.46 | 0.66 | 0.49 |
| 0.42 | 0.53 | 0.48 |
| -0.23 | 0.02 | -0.08 |
| -0.28 | -0.08 | -0.26 |
| 0.57 | 0.67 | 0.57 |
| 0.27 | 0.50 | 0.42 |
| -0.04 | 0.11 | 0.00 |
| -0.29 | -0.10 | -0.27 |
| 0.25 | 0.43 | 0.31 |
| 0.27 | 0.44 | 0.29 |
| -0.40 | -0.58 | -0.31 |
| 1.00 | 0.42 | 0.49 |
| 0.42 | 1.00 | 0.39 |
| 0.49 | 0.39 | 1.00 |

**Figure 48 : Correlation Matrix ( Part 2 )**

Correlation matrix of the original unscaled data is shown below :

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



**Figure 49 : Correlation matrix of the Original Unscaled Data (Part 1)**

| S.F.Ratio | perc.alumni | Expend | Grad.Rate |
|---|---|---|---|
| 0.10 | -0.09 | 0.26 | 0.15 |
| 0.18 | -0.16 | 0.12 | 0.07 |
| 0.24 | -0.18 | 0.06 | -0.02 |
| -0.38 | 0.46 | 0.66 | 0.49 |
| -0.29 | 0.42 | 0.53 | 0.48 |
| 0.28 | -0.23 | 0.02 | -0.08 |
| 0.23 | -0.28 | -0.08 | -0.26 |
| -0.55 | 0.57 | 0.67 | 0.57 |
| -0.36 | 0.27 | 0.50 | 0.42 |
| -0.03 | -0.04 | 0.11 | 0.00 |
| 0.14 | -0.29 | -0.10 | -0.27 |
| -0.13 | 0.25 | 0.43 | 0.31 |
| -0.16 | 0.27 | 0.44 | 0.29 |
| 1.00 | -0.40 | -0.58 | -0.31 |
| -0.40 | 1.00 | 0.42 | 0.49 |
| -0.58 | 0.42 | 1.00 | 0.39 |
| -0.31 | 0.49 | 0.39 | 1.00 |



**Figure 50 : Correlation matrix of the Original Unscaled Data (Part 2)**

**Interpretation :**

By comparing both Covariance and Correlation Matrices we inferred that the values of both the matrices are same. Note that all variances in Covariance Matrix(Scaled Data) are now 1 (main diagonal). In fact, this matrix is same as the correlation matrix of the original (unscaled) variables.

Hence, we found out that the output of correlation matrix of unscaled data and correlation matrix of scaled data and covariance matrix of scaled data are same and identical.

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

We plotted the Boxplot for both data sets that is the original unscaled data and the new scaled data.
The Original Data Boxplot is as follows :

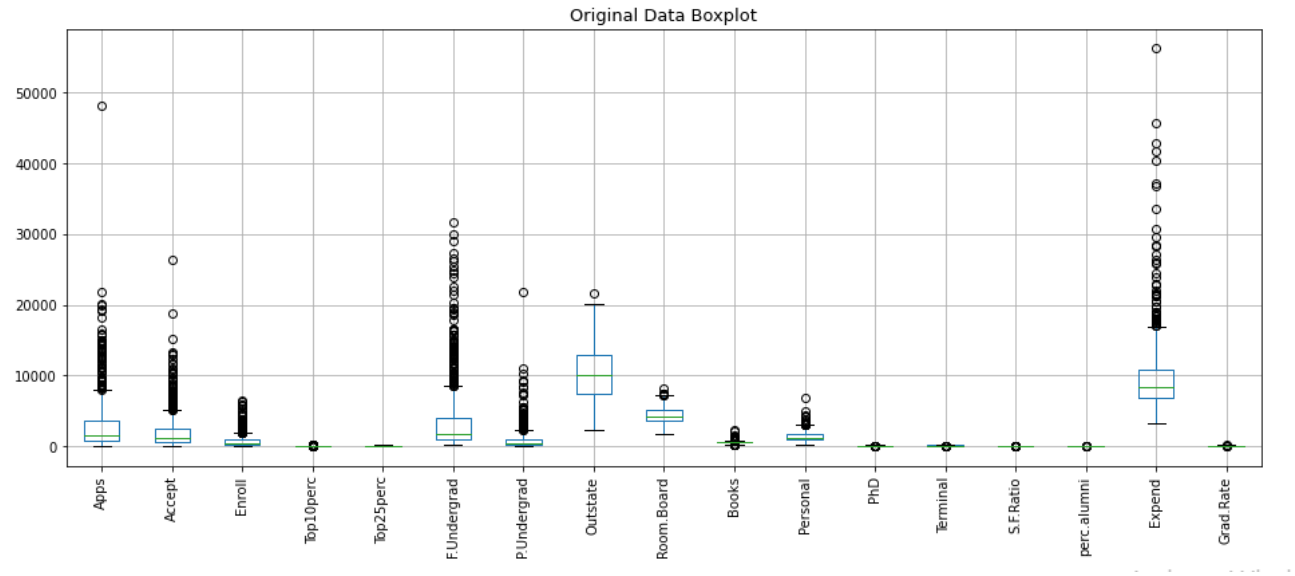



**Figure 51 : Original Data Boxplot**

The New Scaled Data Boxplot is as follows :

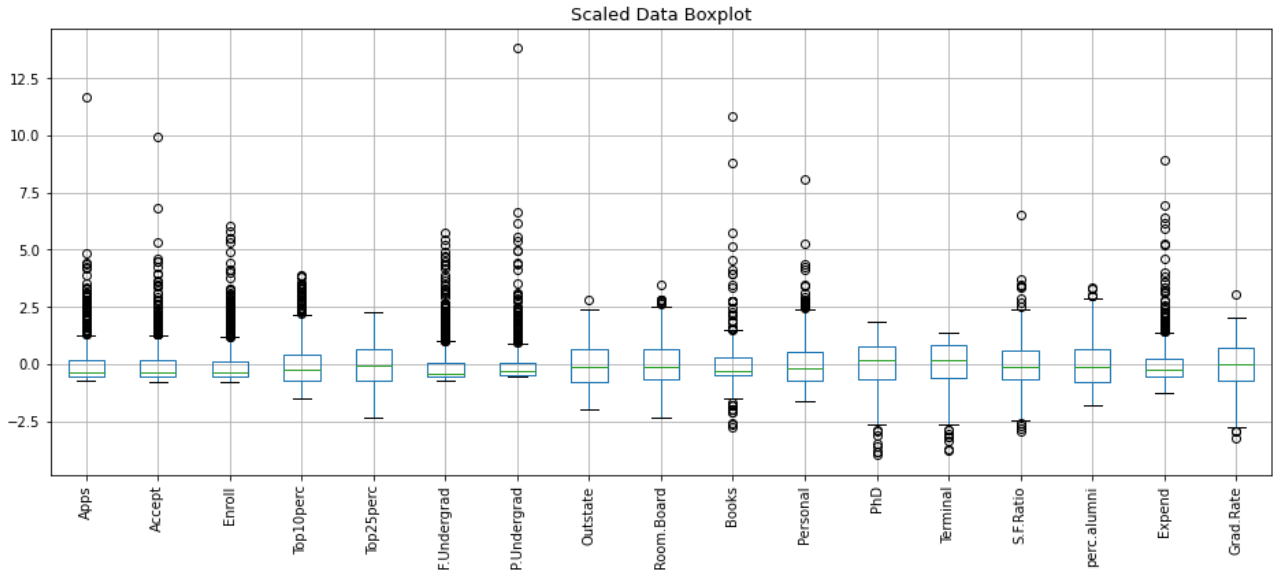



**Figure 52 : Scaled Data Boxplot**

**Interpretation :**

Yes, Outliers are being detected in both Boxplots that is before scaling and after scaling.

We can also note that in the Scaled Boxplot all the boxes of boxplot are on the same line indicating that scaling had standardized all the variables. While in the Original Data Boxplot we can clearly see that many boxes are on different levels according to their respective variables which indicates that all variables are not scaled.

Thus, this led to know that scaling plays an important and crucial role while performing PCA.

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

We used Sklearn and covariance matrix of scaled data and extracted the eigen values and eigen vectors. The output of eigen vectors and eigen values are rounded off and then the output is generated for ease.

Here are the eigen vectors :

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | -0.248 | 0.332 | -0.070 | 0.281 | -0.004 | -0.016 | -0.110 | 0.005 | -0.079 | 0.064 | -0.173 | -0.550 | 0.055 | -0.150 | 0.141 | -0.586 | 0.010 |
| 1 | -0.207 | 0.372 | -0.112 | 0.260 | -0.050 | 0.010 | -0.061 | -0.004 | -0.177 | 0.030 | 0.304 | 0.676 | -0.067 | 0.187 | 0.057 | -0.292 | -0.152 |
| 2 | -0.176 | 0.404 | -0.090 | 0.159 | 0.063 | -0.039 | 0.042 | 0.044 | -0.132 | 0.026 | -0.708 | 0.151 | -0.082 | -0.041 | -0.116 | 0.446 | 0.027 |
| 3 | -0.354 | -0.083 | 0.041 | -0.052 | 0.396 | -0.050 | -0.156 | 0.106 | 0.346 | 0.078 | 0.091 | 0.223 | -0.019 | -0.684 | -0.117 | -0.026 | 0.046 |
| 4 | -0.345 | -0.044 | -0.014 | -0.118 | 0.417 | 0.038 | -0.116 | 0.083 | 0.413 | 0.010 | -0.060 | -0.109 | -0.278 | 0.611 | 0.161 | 0.006 | -0.082 |
| 5 | -0.155 | 0.419 | -0.059 | 0.100 | 0.048 | -0.042 | 0.063 | 0.053 | -0.071 | 0.014 | 0.602 | -0.368 | -0.079 | -0.023 | -0.120 | 0.497 | 0.070 |
| 6 | -0.026 | 0.314 | 0.146 | -0.149 | -0.307 | -0.212 | 0.577 | 0.119 | 0.539 | -0.220 | -0.028 | 0.051 | 0.102 | -0.028 | 0.034 | -0.113 | -0.064 |
| 7 | -0.294 | -0.250 | 0.046 | 0.136 | -0.225 | -0.037 | 0.054 | -0.097 | -0.014 | 0.191 | -0.009 | -0.109 | 0.071 | -0.056 | -0.057 | 0.154 | -0.826 |
| 8 | -0.249 | -0.137 | 0.143 | 0.200 | -0.564 | 0.156 | -0.132 | -0.264 | 0.265 | 0.284 | -0.005 | 0.026 | -0.350 | 0.008 | -0.069 | 0.056 | 0.383 |
| 9 | -0.065 | 0.056 | 0.678 | 0.102 | 0.115 | 0.638 | 0.154 | 0.215 | -0.135 | -0.091 | -0.005 | -0.001 | 0.039 | 0.007 | -0.061 | -0.021 | -0.029 |
| 10 | 0.043 | 0.220 | 0.504 | -0.221 | 0.219 | -0.320 | -0.025 | -0.680 | -0.102 | 0.137 | -0.005 | 0.013 | -0.026 | 0.011 | 0.030 | -0.034 | -0.038 |
| 11 | -0.318 | 0.059 | -0.115 | -0.537 | -0.146 | 0.094 | -0.082 | -0.023 | -0.174 | -0.137 | -0.023 | -0.041 | 0.054 | 0.116 | -0.672 | -0.194 | 0.017 |
| 12 | -0.318 | 0.047 | -0.055 | -0.518 | -0.215 | 0.151 | -0.036 | 0.027 | -0.251 | -0.085 | 0.002 | 0.027 | -0.050 | -0.170 | 0.662 | 0.121 | 0.022 |
| 13 | 0.177 | 0.247 | -0.280 | -0.169 | 0.062 | 0.489 | 0.014 | -0.226 | 0.280 | 0.499 | -0.003 | 0.024 | 0.417 | 0.020 | 0.032 | 0.034 | -0.046 |
| 14 | -0.206 | -0.246 | -0.150 | 0.006 | 0.219 | -0.039 | 0.718 | -0.022 | -0.288 | 0.395 | 0.017 | 0.004 | -0.137 | 0.017 | -0.020 | -0.113 | 0.195 |
| 15 | -0.318 | -0.131 | 0.221 | 0.086 | -0.067 | -0.304 | -0.123 | 0.200 | -0.049 | 0.173 | 0.012 | 0.053 | 0.704 | 0.228 | 0.059 | 0.107 | 0.275 |
| 16 | -0.254 | -0.168 | -0.214 | 0.263 | 0.109 | 0.220 | 0.159 | -0.531 | 0.053 | -0.579 | 0.013 | 0.007 | 0.261 | 0.002 | 0.042 | 0.078 | 0.105 |

**Figure 53 : Eigen Vectors**

And here are the Eigen Values :

```
 Eigen Values are as follows :

%s [5.451 4.488 1.173 1.004 0.938 0.849 0.586 0.604 0.522 0.405 0.023 0.033
 0.312 0.091 0.143 0.161 0.217]
```

**Figure 54 : Eigen Values**

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

We used Sklearn decomposition and imported PCA. After that, we performed PCA. Then that data of the Principal Component(eigenvectors) was transformed into a new DataFrame called pca_score.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | -1.592855 | 0.767334 | -0.101074 | -0.921749 | -0.743975 | -0.298306 | 0.638443 | -0.879386 | 0.093084 | 0.048593 | 0.399747 | -0.089690 | -0.052098 | 0.180140 | 0.00 |
| **1** | -2.192402 | -0.578830 | 2.278798 | 3.588918 | 1.059997 | -0.177137 | 0.236753 | 0.046925 | 1.113780 | 0.965154 | -0.212509 | 0.097239 | -0.243518 | -0.744204 | 0.10 |
| **2** | -1.430964 | -1.092819 | -0.438093 | 0.677241 | -0.369613 | -0.960592 | -0.248276 | 0.308740 | -0.105452 | 0.640660 | -0.154993 | -0.344731 | 0.097551 | 0.227527 | -0.02 |
| **3** | 2.855557 | -2.630612 | 0.141722 | -1.295486 | -0.183837 | -1.059508 | -1.249356 | -0.147694 | 0.378997 | 0.461244 | -0.420651 | 0.687143 | -0.075461 | -0.003380 | -0.07 |
| **4** | -2.212008 | 0.021631 | 2.387030 | -1.114538 | 0.684451 | 0.004918 | -2.159220 | -0.624413 | -0.160383 | 0.363428 | -0.153339 | -0.050552 | 0.267207 | -0.614409 | -0.27 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **772** | -3.328458 | 1.220255 | -0.383388 | 0.108555 | 0.776996 | 0.309429 | -0.165021 | 0.347435 | 0.545218 | 0.876458 | 0.447952 | -0.029980 | 0.274732 | -0.392414 | 0.04 |
| **773** | 0.199389 | -0.686689 | 0.051564 | 0.562269 | 0.375191 | 0.373343 | 0.848453 | 0.626515 | -0.072041 | -0.311567 | 0.012782 | 0.201674 | -0.088843 | -0.339420 | 0.08 |
| **774** | -0.732561 | -0.077235 | -0.000406 | 0.054316 | -0.516021 | 0.468014 | -1.317492 | -0.128288 | 0.212375 | 0.300443 | -0.471931 | 0.448232 | 0.083219 | -0.041142 | 0.06 |
| **775** | 7.919327 | -2.068329 | 2.073564 | 0.852054 | -0.947755 | -2.069937 | 0.083328 | -0.552586 | 0.081969 | 0.924892 | 2.242207 | 1.363251 | 0.076113 | -0.074911 | 0.35 |
| **776** | -0.469508 | 0.366661 | -1.328915 | -0.108023 | -1.132176 | 0.839893 | 1.307313 | 0.627410 | 0.723562 | -1.208948 | 0.207169 | 0.785628 | 0.432977 | -0.101576 | -0.11 |

777 rows × 17 columns

**Figure 55 : PCA_Score(Eigen Vectors)**

## Question 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]

We first generated the loading/weights of each feature corresponding to eigen vector/component from PCA.

Afterwards , we extracted the first row of original scaled data _____(1)

Then, we extracted the pca_component[0] that is the first value of pca_component as per index used in python_____(2)

Finally with the help of (1) & (2), we generated the equation/explicit form of the first Principal Component.

The Equation/Explicit Form is as follows :

PC0 = 0.25*(-0.35) + 0.21*(-0.32) + 0.18*(-0.06) + 0.35*(-0.26) + 0.34*(-0.19) + 0.15*(-0.17) + 0.03*(-0.21) + 0.29*(-0.75)+ 0.25*(-0.96) + 0.06*(-0.6) + (-0.04*1.27) + 0.32*(-0.16) + 0.32*(-0.12) + (-0.18*1.01) + 0.21*(-0.87) + 0.32*(-0.5) + 0.25*(-0.32)

The output of this equation is -1.59 (rounded off to 2 decimals only).

At last, we used the iloc function and find the Principal Component [0] value from that.

From that also, we got the same output as from the above equation we formed. Thus, it confirms that we got the correct equation.

**Question 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?**

First we found out explained variance(eigen values) of PCA. It is as follows :

```
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])
```

**Figure 56 : Explained Variance**

Then we found out explained variance ratio . It is as follows :

```
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])
```

**Figure 57 : Explained Variance Ratio**

Then, we found out the cumulative explained variance ratio . It is as follows :

```
Cumulative Variance Explained

 [ 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.        ]
```

Note: Cumulative variance of all 12 components is 1. i.e. all 12 components are able to explain 100% variance in data

**Figure 58 : Cumulative Variance Explained**

Next, we plotted the Scree Plot for visualization . It is as follows:

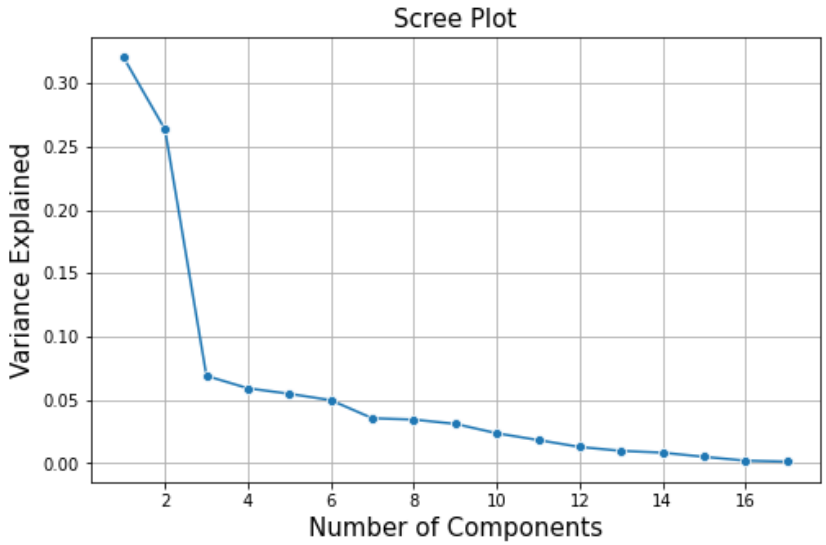



**Figure 59 : Scree Plot**

Next, we plotted the graph between Explained Variance Ratio vs Principal Components.

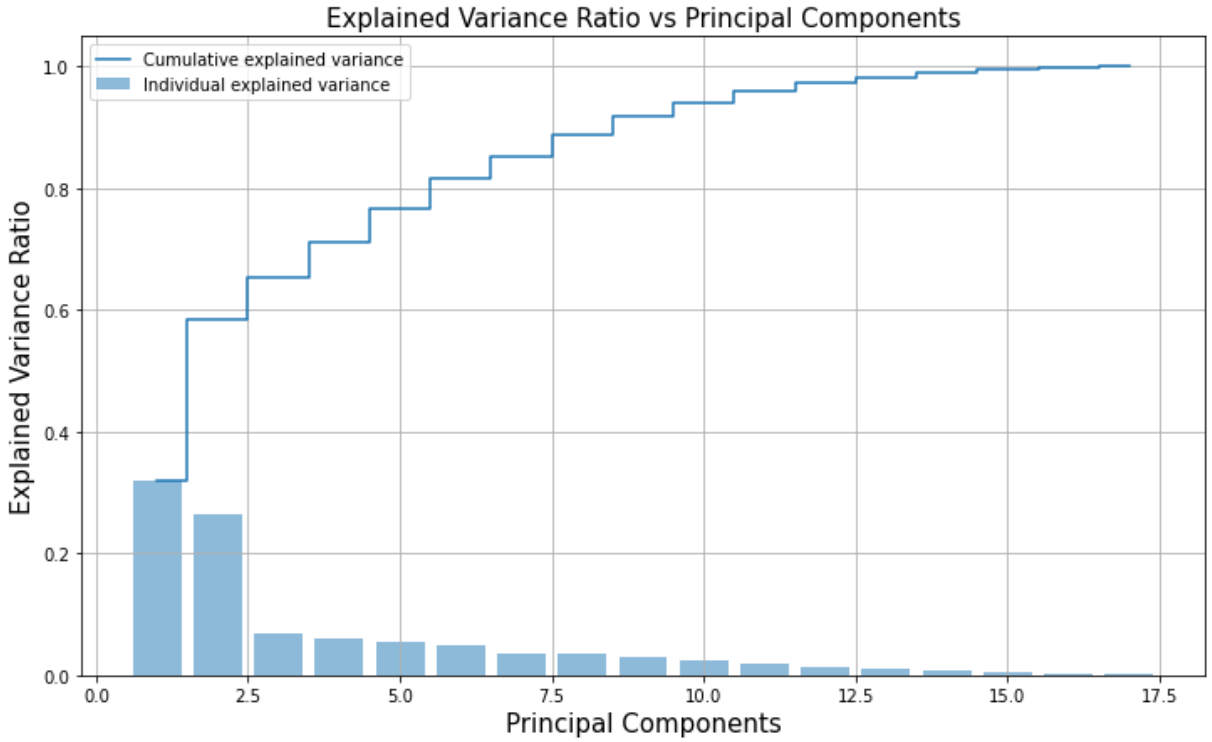



**Figure 60: Explained Variance vs Principal Components**

From scree plot and plot above, we see that we can retain only 4 components instead of 17 components as beyond 4 components the amount of variance explained in minimal and the first 4 components are able to explain 71.18% variance in data. Rest remaining components out of 17 were not considered as their contribution were very minimal and thus they can be neglected because the main 4 Principal components were able to explain majority 71.18% variance in the data.

## Question 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]

We reduced the linear dimension to 4 components with the help of PCA. It was named as PCA main. It is as follows:

```
array([[-1.59285540e+00,  7.67333506e-01, -1.01072665e-01,
        -9.21750291e-01],
       [-2.19240180e+00, -5.78829993e-01,  2.27879980e+00,
         3.58891645e+00],
       [-1.43096371e+00, -1.09281889e+00, -4.38092841e-01,
         6.77240573e-01],
       ...,
       [-7.32560597e-01, -7.72352585e-02, -4.01791814e-04,
         5.43122158e-02],
       [ 7.91932735e+00, -2.06832894e+00,  2.07358213e+00,
         8.52024253e-01],
       [-4.69508065e-01,  3.66660971e-01, -1.32892100e+00,
        -1.08013199e-01]])
```

**Figure 61 : PCA_Main (4 Components)**

Then we checked the shape of the new PCA_Main and found out that its dimensions were

```
(4, 17)
```

**Figure 62 : Dimensions of PCA_Main**

We can clearly see that now with the help of PCA , the number of rows for explaining majority of the data have been reduced to 4.

After that we plot a heatmap for 4 components of new PCA main.

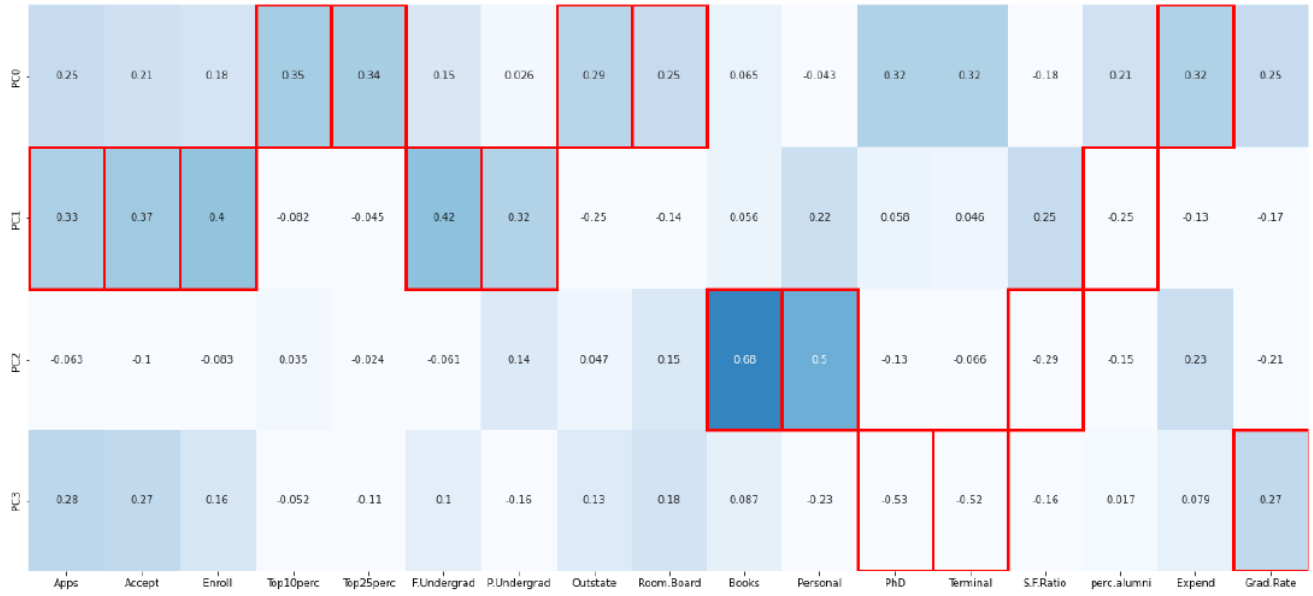



**Figure 63 : Heatmap of PCA main**

## Interpretations :

- The heatmap clearly shows that maximum percentage for Top10perc is35% and it was placed in PC0 while among Top25%perc the maximum percentage is 34% it was also placed in PC0.
- We inferred that only 33% of students application were received that were maximum and were placed in PC1. With that in consideration, only 37% students application accepted in PC1 and 40% of the students enrolled in their respective University/College.
- We inferred that 68% of estimated books cost for a student lay in PC2 while on the other hand, only 50% of estimated personal spending of a student were in PC2.
- At last, out of all PC components, the maximum graduation rate was considered in PC2 with 27% .

The first five observation of the new PCA main are :

| | PC0 | PC1 | PC2 | PC3 |
|---|---|---|---|---|
| 0 | -1.592855 | 0.767334 | -0.101073 | -0.921750 |
| 1 | -2.192402 | -0.578830 | 2.278800 | 3.588916 |
| 2 | -1.430964 | -1.092819 | -0.438093 | 0.677241 |
| 3 | 2.855557 | -2.630612 | 0.141719 | -1.295479 |
| 4 | -2.212008 | 0.021631 | 2.387012 | -1.114513 |

**Figure 64 : PCA (4 Components)**

Then, the 'Names' Column was added back to the PCA main and named as new Dataframe called df_new.

| | Names | PC0 | PC1 | PC2 | PC3 |
|---|---|---|---|---|---|
| 0 | Abilene Christian University | -1.592855 | 0.767334 | -0.101073 | -0.921750 |
| 1 | Adelphi University | -2.192402 | -0.578830 | 2.278800 | 3.588916 |
| 2 | Adrian College | -1.430964 | -1.092819 | -0.438093 | 0.677241 |
| 3 | Agnes Scott College | 2.855557 | -2.630612 | 0.141719 | -1.295479 |
| 4 | Alaska Pacific University | -2.212008 | 0.021631 | 2.387012 | -1.114513 |

**Figure 65 : PCA (df_new)**

A new correlation matrix was build for the new PCA Dataframe

| | PC0 | PC1 | PC2 | PC3 |
|---|---|---|---|---|
| PC0 | 1.000000e+00 | -2.314568e-17 | -2.159908e-17 | -6.163812e-17 |
| PC1 | -2.314568e-17 | 1.000000e+00 | 4.363905e-18 | -8.560740e-17 |
| PC2 | -2.159908e-17 | 4.363905e-18 | 1.000000e+00 | 6.021184e-17 |
| PC3 | -6.163812e-17 | -8.560740e-17 | 6.021184e-17 | 1.000000e+00 |

**Figure 66: Correlation Matrix PCA(df_new)**

At last we plot the heatmap for this new PCA .

It is as follows :

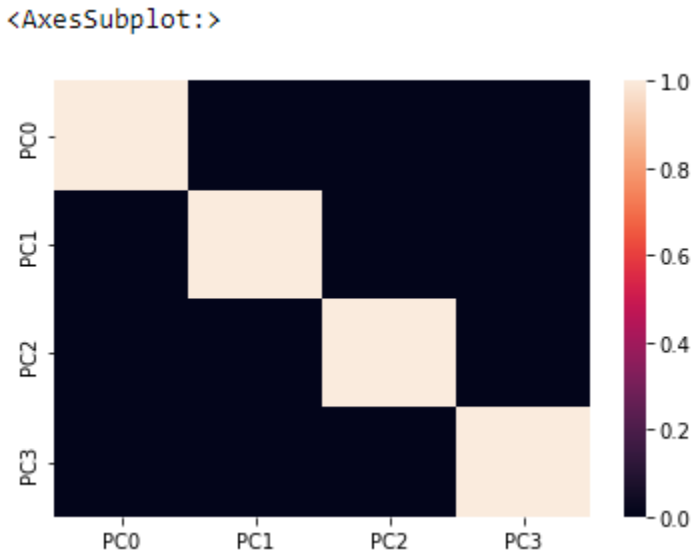



**Figure 67 : Heatmap of PCA (df_new)**

**Interpretations :**

We can see that all the components has explained most of the variance from the data and this can be used as PCA helped in getting the output. This heatmap justified that all first 4 components from the Original Data set were able to explain 71.18% variance which is sufficient.