11 September 2021

Advanced Statistics Project

PGP-DSBA June-Batch

SOOREJ KUNNOOL BALAN

2021

Table of Contents

[1. Problem 1 3](#_Toc82196970)

[Data Description 3](#_Toc82196971)

[Sample of the dataset: 3](#_Toc82196972)

[Exploratory Data Analysis: 4](#_Toc82196973)

[Let us check the type of variables in the data frame 4](#_Toc82196974)

[Check for missing values in the dataset 4](#_Toc82196975)

[Check for duplicate observations in the dataset 4](#_Toc82196976)

[Data summary 4](#_Toc82196977)

[Outliers in the data 5](#_Toc82196978)

[Data type conversion 5](#_Toc82196979)

[Problem 1A 6](#_Toc82196980)

[1.1 State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually. 6](#_Toc82196981)

[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. 6](#_Toc82196982)

[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. 7](#_Toc82196983)

[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 results. 8](#_Toc82196984)

[Problem 1B 8](#_Toc82196985)

[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. 8](#_Toc82196986)

[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? 9](#_Toc82196987)

[1.7 Explain the business implications of performing ANOVA for this particular case study. 10](#_Toc82196988)

[2. Problem 2 11](#_Toc82196989)

[2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA? 11](#_Toc82196990)

[Data Description 11](#_Toc82196991)

[Sample of the dataset: 12](#_Toc82196992)

[Exploratory Data Analysis: 12](#_Toc82196993)

[2.2 Is scaling necessary for PCA in this case? Give justification and perform scaling. 40](#_Toc82196994)

[2.3 Comment on the comparison between the covariance and the correlation matrices from this data.[on scaled data] 41](#_Toc82196995)

[2.4 Check the dataset for outliers before and after scaling. What insight do you derive here? 42](#_Toc82196996)

[2.5 Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both] 43](#_Toc82196997)

[2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features. 45](#_Toc82196998)

[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] 46](#_Toc82196999)

[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? 46](#_Toc82197000)

[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] 47](#_Toc82197001)

[THE END 49](#_Toc82197002)

List of Tables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table-1.1 Dataset Sample…………………………………………………………. | 5 |  | Table-2.35 Top 5 colleges w.r.t perc.alumni………………………… | | 39 |
| Table-1.2 Concise data summary………………………………………………….. | 5 |  | Table-2.36 Bottom 5 colleges w.r.t perc.alumni……………………. | | 40 |
| Table-1.3 Data summary………………………………………………………….. | 6 |  | Table-2.37 Top 5 colleges w.r.t perc.alumni………………………… | | 40 |
| Table-1.4 Concise data summary…………………………………………………. | 7 |  | Table-2.38 Grad.Rate Data Error……………………………………. | | 40 |
| Table-1.5 Shapiro-Wilk Test: Salary/Education………………………………….. | 8 |  | Table-2.39 Data summary…………………………………………… | | 41 |
| Table-1.6 Levene’s Test: Salary/Education………………………………………. | 8 |  | Table-2.40 Bottom 5 colleges w.r.t Grad.Rate………………………. | | 42 |
| Table-1.7 One Way ANOVA: Salary/Education…………………………………. | 8 |  | Table-2.41 Colleges with 100% Grad.Rate…………………………. | | 42 |
| Table-1.8 Shapiro-Wilk Test: Salary/Occupation………………………………… | 9 |  | Table-2.42 Correlation matrix………………………………………. | | 43 |
| Table-1.9 Levene’s Test: Salary/Occupation……………………………………… | 9 |  | Table-2.43 Scaled – Data Summary…………………………………. | | 45 |
| Table-1.10 One Way ANOVA: Salary/Occupation………………………………. | 9 |  | Table-2.44 Scaled – Data Correlation matrix………………………... | | 46 |
| Table-1.11 Tukey Test: Salary/Education………………………………………… | 10 |  | Table-2.45 Scaled – Data Covariance matrix………………………... | | 47 |
| Table-1.12 Two Way ANOVA results……………………………………………. | 11 |  | Table-2.46 Eigenvectors against original features…………………… | | 51 |
| Table-2.1 Dataset Sample…………………………………………………………. | 14 |  |  | |  |
| Table-2.2 Concise data summary…………………………………………………. | 14 |  |  | |  |
| Table-2.3 Data Summary………………………………………………………….. | 15 |  |  | |  |
| Table-2.4 Bottom 5 colleges w.r.t Apps………………………………………….. | 17 |  |  | |  |
| Table-2.5 Top 5 colleges w.r.t Apps……………………………………………… | 17 |  |  | |  |
| Table-2.6 Bottom 5 colleges w.r.t Accept………………………………………… | 18 |  |  | |  |
| Table-2.7 Top 5 colleges w.r.t Accept…………………………………………….. | 18 |  |  | |  |
| Table-2.8 Bottom 5 colleges w.r.t Enroll…………………………………………. | 20 |  |  | |  |
| Table-2.9 Top 5 colleges w.r.t Enroll……………………………………………… | 20 |  |  | |  |
| Table-2.10 Bottom 5 colleges w.r.t Top10perc…………………………………… | 21 |  |  | |  |
| Table-2.11 Top 5 colleges w.r.t Top10perc………………………………………. | 21 |  |  | |  |
| Table-2.12 Bottom 5 colleges w.r.t Top25perc…………………………………… | 23 |  |  | |  |
| Table-2.13 Colleges with Top25perc at 100%......................................................... | 23 |  |  | |  |
| Table-2.14 Bottom 5 colleges w.r.t F.Undergrad…………………………………. | 24 |  |  |  | |
| Table-2.15 Top 5 colleges w.r.t F.Undergrad……………………………………… | 24 |  |  |  | |
| Table-2.16 Bottom 5 colleges w.r.t P.Undergrad…………………………………. | 26 |  |  |  | |
| Table-2.17 Top 5 colleges w.r.t P.Undergrad…………………………………….. | 26 |  |  |  | |
| Table-2.18 Bottom 5 colleges w.r.t Outstate……………………………………… | 27 |  |  |  | |
| Table-2.19 Top 5 colleges w.r.t Outstate…………………………………………. | 27 |  |  |  | |
| Table-2.20 Bottom 5 colleges w.r.t Room.Board…………………………………. | 29 |  |  |  | |
| Table-2.21 Top 5 colleges w.r.t Room.Board…………………………………….. | 29 |  |  |  | |
| Table-2.22 Bottom 5 colleges w.r.t Books………………………………………… | 30 |  |  |  | |
| Table-2.23 Top 5 colleges w.r.t Books……………………………………………. | 30 |  |  |  | |
| Table-2.24 Bottom 5 colleges w.r.t Personal……………………………………… | 32 |  |  |  | |
| Table-2.25 Top 5 colleges w.r.t Personal…………………………………………. | 32 |  |  |  | |
| Table-2.26 PhD Data Error……………………………………………………….. | 32 |  |  |  | |
| Table-2.27 Data summary…………………………………………………………. | 32 |  |  |  | |
| Table-2.28 Bottom 5 colleges w.r.t PhD………………………………………….. | 34 |  |  |  | |
| Table-2.29 Top 5 colleges w.r.t PhD………………………………………………. | 34 |  |  |  | |
| Table-2.30 Bottom 5 colleges w.r.t Terminal……………………………………… | 35 |  |  |  | |
| Table-2.31 Colleges with Terminal at 100%............................................................ | 35 |  |  |  | |
| Table-2.32 Top 5 colleges w.r.t S.F.Ratio………………………………………… | 37 |  |  |  | |
| Table-2.33 Bottom 5 colleges w.r.t S.F.Ratio…………………………………….. | 37 |  |  |  | |
| Table-2.34 Bottom 5 colleges w.r.t perc.alumni………………………………….. | 39 |  |  |  | |

List of Figures

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Figure-1.1 Histogram of Salary…………………………………………………… | 6 |  |  | |  |
| Figure-1.2 Boxplot of Salary……………………………………………………… | 7 |  |  | |  |
| Figure-1.3 Interaction plot of Salary vs Occupation/Education………………….. | 10 |  |  | |  |
| Figure-2.1 Histogram: Apps………………………………………………………. | 16 |  |  | |  |
| Figure-2.2 Boxplot: Apps…………………………………………………………. | 16 |  |  | |  |
| Figure-2.3 Histogram: Accept……………………………………………………... | 17 |  |  | |  |
| Figure-2.4 Boxplot: Accept……………………………………………………….. | 18 |  |  | |  |
| Figure-2.5 Histogram: Enroll……………………………………………………… | 19 |  |  | |  |
| Figure-2.6 Boxplot: Enroll………………………………………………………… | 19 |  |  | |  |
| Figure-2.7 Histogram: Top10perc………………………………………………… | 20 |  |  | |  |
| Figure-2.8 Boxplot: Top10perc……………………………………………………. | 21 |  |  | |  |
| Figure-2.9 Histogram: Top25perc………………………………………………… | 22 |  |  | |  |
| Figure-2.10 Boxplot: Top25perc………………………………………………….. | 22 |  |  | |  |
| Figure-2.11 Histogram: F.Undergrad……………………………………………… | 23 |  |  | |  |
| Figure-2.12 Boxplot: F.Undergrad……………………………………………….. | 24 |  |  | |  |
| Figure-2.13 Histogram: P.Undergrad…………………………………………….. | 25 |  |  | |  |
| Figure-2.14 Boxplot: P.Undergrad………………………………………………… | 25 |  |  | |  |
| Figure-2.15 Histogram: Outstate………………………………………………….. | 26 |  |  | |  |
| Figure-2.16 Boxplot: Outstate…………………………………………………….. | 27 |  |  | |  |
| Figure-2.17 Histogram: Room.Board……………………………………………… | 28 |  |  | |  |
| Figure-2.18 Boxplot: Room.Board………………………………………………… | 28 |  |  | |  |
| Figure-2.19 Histogram: Books……………………………………………………. | 29 |  |  | |  |
| Figure-2.20 Boxplot: Books………………………………………………………. | 30 |  |  | |  |
| Figure-2.21 Histogram: Personal………………………………………………….. | 31 |  |  | |  |
| Figure-2.22 Boxplot: Personal…………………………………………………….. | 31 |  |  | |  |
| Figure-2.23 Histogram: PhD……………………………………………………… | 33 |  |  |  | |
| Figure-2.24 Boxplot: PhD………………………………………………………… | 33 |  |  |  | |
| Figure-2.25 Histogram: Terminal………………………………………………… | 34 |  |  |  | |
| Figure-2.26 Boxplot: Terminal……………………………………………………. | 35 |  |  |  | |
| Figure-2.27 Histogram: S.F.Ratio…………………………………………………. | 36 |  |  |  | |
| Figure-2.28 Boxplot: S.F.Ratio…………………………………………………… | 37 |  |  |  | |
| Figure-2.29 Histogram: perc.alumni………………………………………………. | 38 |  |  |  | |
| Figure-2.30 Boxplot: perc.alumni…………………………………………………. | 38 |  |  |  | |
| Figure-2.31 Histogram: Expend…………………………………………………… | 39 |  |  |  | |
| Figure-2.32 Boxplot: Expend……………………………………………………… | 40 |  |  |  | |
| Figure-2.33 Histogram: Grad.Rate………………………………………………… | 42 |  |  |  | |
| Figure-2.34 Boxplot of Grad.Rate………………………………………………… | 42 |  |  |  | |
| Figure-2.35 Boxplot of Scaled Data………………………………………………. | 47 |  |  |  | |
| Figure-2.36 Boxplot of Original Data…………………………………………….. | 48 |  |  |  | |
| Figure-2.37 Scree Plot…………………………………………………………….. | 52 |  |  |  | |
| Figure-2.38 Cumulative Explained Variance Ratio Plot………………………….. | 52 |  |  |  | |

# 

# Problem 1

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.]



## Data Description

1. Education : type of Education (Doctorate, Bachelors, HS-grad).
2. Occupation : type of Occupation (Adm-clerical, Sales, Prof-specialty, Exec-managerial).
3. Salary : continuous from 50103 to 260151.

## Sample of the dataset:

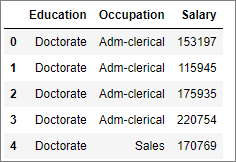


Table-1.1 Dataset Sample

## Exploratory Data Analysis:

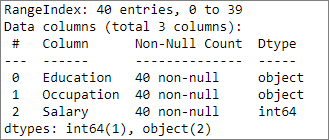


Table-1.2 Concise data summary

### Let us check the type of variables in the data frame

There are a total of 40 observations and 3 columns in the dataset. Of the 3 columns, 2 are of object type and 1 is of integer data type.

### Check for missing values in the dataset

From Table-1.2 we can see that all the columns have 40 non-null values and hence we have no missing values in the dataset.

### Check for duplicate observations in the dataset

There are no duplicate rows in the dataset.

### Data summary

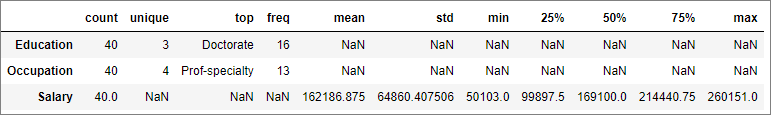


Table-1.3 Data summary

Salary is the only numerical field in the data, which is a numerical continuous variable with no missing data as per the count.

Mean – 162186.875

Standard Deviation – 64860.408

Q1(1st Quartile) – 99897.5

Q2(2nd Quartile)/Median – 169100.00

Q3(3rd Quartile) – 214440.75

IQR(Inter-Quartile Range) = Q3- Q1 = 114543.25

Minimum value = 50103.00

Maximum value = 260151.00

Range = 260151.00 – 50103.00 = 210048.00

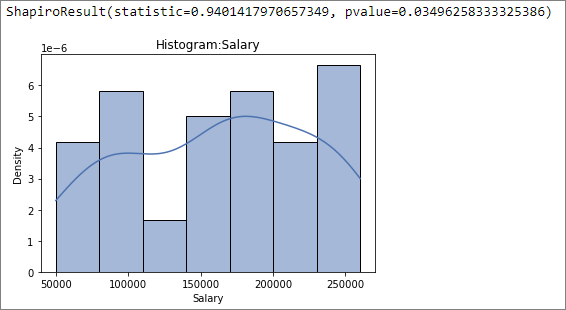


Figure-1.1 Histogram of Salary

Figure-1.1 depicts the histogram of “Salary” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 0.035 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Salary” column data does not have a normal distribution, as per the provided observations.

### Outliers in the data

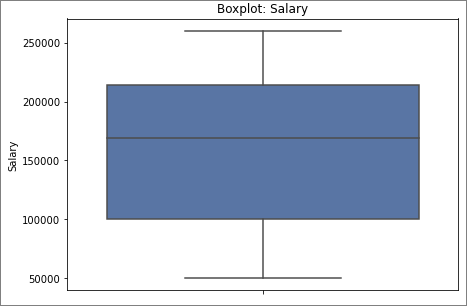


Figure-1.2 Boxplot of Salary

Figure-1.2 depicts the boxplot of “Salary” and we can see that there are no outliers present in the data.

### Data type conversion

Since columns, ‘Education’ and ‘Occupation’ are categorical, converted the data type of variables into categorical, as shown in Table 1.4

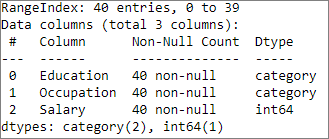


Table-1.4 Concise data summary

## Problem 1A

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

Null and alternate hypothesis for conducting one-way ANOVA for Education

Ho – Null Hypothesis – The means of Salary variable with respect to each Education type is same.

Ha – Alternate Hypothesis – At least one of the means of Salary variable with respect to each Education type is different.

Null and alternate hypothesis for conducting one-way ANOVA for Occupation

Ho – Null Hypothesis – The means of Salary variable with respect to each Occupation type is same.

Ha – Alternate Hypothesis – At least one of the means of Salary variable with respect to each Occupation type is different.

## 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.

Assumptions to perform one-way ANOVA are as follows:

1. All samples are drawn independent of each other.

Let’s assume the data provided are randomly drawn and independent of each other.

1. Each group sample are drawn from a normally distributed population.

Let’s perform Shapiro-Wilk test on the Salary sample for each of the education

groups- Doctorate, Bachelors, HS-grad. The results are given in Table 1.5.



Table-1.5 Shapiro-Wilk Test: Salary/Education

From the above table we can see that the p-value for each of the sample is greater than 0.05, and hence we don’t have enough evidence to reject the null hypothesis that the samples are normally distributed. Hence the salary samples for each education type are normally distributed.

1. All population groups have a common variance.

Let’s perform Levene’s test on the Salary sample for the education groups- Doctorate, Bachelors, HS-grad. The results are given in Table 1.6.



Table-1.6 Levene’s Test: Salary/Education

From the above table we can see that the p-value for Levene’s test is greater than 0.05, and hence we don’t have enough evidence to reject the null hypothesis that the samples have the same variance. Hence the salary samples for each education types have the same variance.

Let’s perform one-way ANOVA on Salary w.r.t Education. The hypothesis are as follows:

Ho – Null Hypothesis – The means of Salary variable with respect to each Education type is same.

Ha – Alternate Hypothesis – At least one of the means of Salary variable with respect to each Education type is different.

One-way ANOVA provides the below result (Table 1.7)

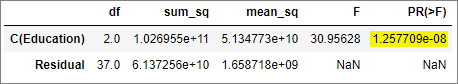


Table-1.7 One Way ANOVA: Salary/Education

Since the pvalue(highlighted in Table 1.7) is lesser than 0.05, we have enough evidence to reject the null hypothesis, hence at least one of the means of Salary variable with respect to Education type is different from others.

## 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.

Assumptions to perform one-way ANOVA are as follows:

1. All samples are drawn independent of each other.

Let’s assume the data provided are randomly drawn and independent of each other.

1. Each group sample are drawn from a normally distributed population.

Let’s perform Shapiro-Wilk test on the Salary sample for each of the occupation

groups- Adm-clerical, Sales, Prof-specialty, Exec-managerial. The results are given in Table 1.8.

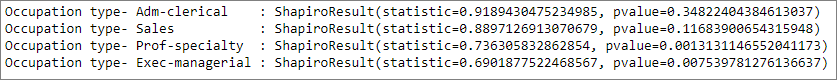


Table-1.8 Shapiro-Wilk Test: Salary/Occupation

From the above table we can see that the p-value for Salary samples against Occupation type Prof-specialty and Exec-managerial is lesser than 0.05, and hence we have enough evidence to reject the null hypothesis that the samples are normally distributed. But the question has asked to assume the samples as normal, hence we will continue with ANOVA.

1. All population groups have a common variance.

Let’s perform Levene’s test on the Salary sample for the occupation

groups- Adm-clerical, Sales, Prof-specialty, Exec-managerial. The results are given in Table 1.9.



Table-1.9 Levene’s Test: Salary/Occupation

From the above table we can see that the p-value for Levene’s test is greater than 0.05, and hence we don’t have enough evidence to reject the null hypothesis that the samples have the same variance. Hence the Salary samples against occupation types have the same variance.

Let’s perform one-way ANOVA on Salary w.r.t Occupation. The hypothesis are as follows:

Ho – Null Hypothesis – The means of Salary variable with respect to each Occupation type is same.

Ha – Alternate Hypothesis – At least one of the means of Salary variable with respect to each Occupation type is different.

One-way ANOVA provides the below result (Table 1.10)

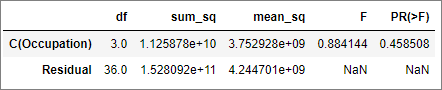


Table-1.10 One Way ANOVA: Salary/Occupation

Since the pvalue(highlighted in Table 1.10) is greater than 0.05, we do not have enough evidence to reject the null hypothesis, hence the means of Salary variable with respect to each Occupation type is the same.

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

We have determined that the one-way ANOVA null hypothesis is rejected in 1.2, where we have determined that at least one of the means of Salary variable with respect to each Education type is different.

Let’s see which Education means are significantly different, by executing Tukey test for the Salary/Education data. The results are as given below:

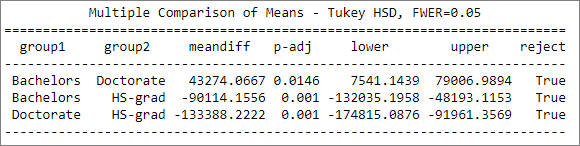


Table-1.11 Tukey Test: Salary/Education

The p-value (given as p-adj in the above table) is below 0.05 for all the 3 education pairs, hence the Salary means across education is significantly different from each other. We can see that Education level ‘Doctorate’ has a higher mean than ‘Bachelors’ and ‘Bachelors’ has a higher mean than ‘HS-grad’.

## Problem 1B

## 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.

Let us draw an interaction plot to see the interaction between the treatments (Education/Occupation).

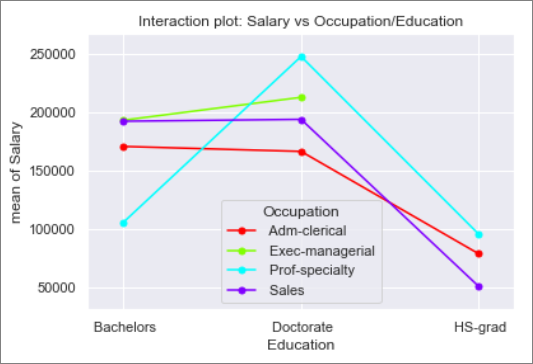


Figure-1.3 Interaction plot of Salary vs Occupation/Education

1. There seems to be a significant level of interaction between education levels – Bachelors & Doctorate over occupation groups – Prof-Specialty across the other 3 occupation types.
2. There seems to be a significant level of interaction between education levels – Bachelors & Doctorate over occupation – Sales and Adm-clerical.
3. It is expected that typically Doctorates will earn higher salary than Bachelors. But for Sales and Adm-Clerical jobs the salaries are comparable. Either work experience is playing a role to explain this anomaly or Doctorate education is not required for Sales/Adm-Clerical jobs, and hence they are overqualified for the job and are paid at the same scale as that of Bachelors.
4. In the Prof-specialty occupation, we can see Bachelors and HS-Grads have comparable salaries, but Doctorates have more than 2 times the salaries of Bachelors/HS-Grad.
5. In Exec-managerial occupation, HS-Grad’s does not seem to be eligible.
6. There seems to be a significant level of interaction between education levels –HS-Grad & Doctorate over occupation – Sales & Adm-Clerical.

## 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?

Let’s perform two-way ANOVA on Salary w.r.t treatments Education, Occupation and with their interaction. The hypothesis are as follows:

Education group hypothesis:

Ho – Null Hypothesis – The means of Salary variable with respect to education groups is the same.

Ha – Alternate Hypothesis – At least one of the means of Salary variable with respect to education group differs from others.

Occupation type hypothesis:

Ho – Null Hypothesis – The means of Salary variable with respect to occupation types is the same.

Ha – Alternate Hypothesis – At least one of the means of Salary variable with respect to occupation type differs from others.

Interaction between Education/Occupation type hypothesis:

Ho – Null Hypothesis – There is no interaction between education groups and occupation types.

Ha – Alternate Hypothesis – There is interaction between education groups and occupation types.

Two-way ANOVA provides the below results as shown in Table 1.12:

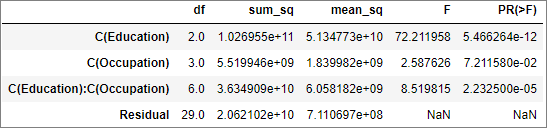


Table-1.12 Two Way ANOVA results

1. We can see that ANOVA result for Education groups have p-value as 5.47e-12, which is lesser than 0.05. Hence we have enough evidence to reject the Education group null hypothesis. So, at least one of the means of Salary variable with respect to education group differs from others.
2. The ANOVA result for Occupation types have p-value as 7.21e-2, which is greater than 0.05. Hence we do not have enough evidence to reject the Occupation type null hypothesis. So, the means of Salary variable with respect to occupation type is the same.
3. The ANOVA result for interaction effect between factors, Education and Occupation, has a p-value 2.23e-5 which is lesser than 0.05. Hence we have enough evidence to reject the Interaction between Education/Occupation type, null hypothesis. So, there is an interaction between education groups and occupation types on Salary. We have surmised this in question 1.5.

From the above results we can see that Education plays a significant role in an employee’s mean Salary. Although occupation type alone does not, statistically, seem to have an impact on Salary, the interaction between the factors Education and Occupation, have a significant impact on the mean Salary of an individual.

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

* An employee’s salary is significantly dependent on his/her education level.
* Statistically occupation type does not have an impact on Salary, yet the interaction between Education and Occupation factor, significantly impacts mean Salary.
* From the given samples we have seen Bachelor’s earning a higher average Salary than Doctorates in ‘Adm-clerical’ occupation type. This could be due to either the fact that Doctorates are

over-qualified for this job and the higher education level does not contribute to the job or we are comparing between bachelor groups having significantly higher work experience than Doctorates. Hence more data such as work experience, domain experience, whether working in a domain in which education is completed etc. are needed to increase the effectiveness/validity of the test.

* HS-Grads are the lowest paid among all occupation types and from the data, we can infer that they are not qualified for Exec-managerial job posts. They earn the lowest in ‘Sales’ occupation and highest in ‘Prof-specialty’.
* Doctorates command the highest salary across occupation types, except Sales, where Bachelor’s earn slightly higher average salary. Doctorates earn the maximum in Prof-specialty and least in Adm-clerical.
* Bachelors earn the maximum in Exec-managerial and the least in Prof-specialty.
* The data indicates that it is possible to have individuals working in the same job role, having varying salaries according to education level. Years of experience would also affect these dynamics.

# Problem 2

The dataset ‘[Education - Post 12th Standard.csv](https://olympus.greatlearning.in/courses/51551/files/2844602/download?verifier=mb31Q2fiwvYG4wpC6eXjitgZ9dt78sC4VOkQKUdI&wrap=1)’ 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](https://olympus.greatlearning.in/courses/51551/files/2844601/download?verifier=jE2L69JsTTAy7YawmYT6Ze1MWJsPTkGg9PiGO5Dp&wrap=1)

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

### Data Description

1. Names: Names of various university and colleges - nominal in nature.
2. Apps: Number of applications received – continuous from 81 to 48094.
3. Accept: Number of applications accepted – continuous from 72 to 26330.
4. Enroll: Number of new students enrolled – continuous from 35 to 6392.
5. Top10perc: Percentage of new students from top 10% of Higher Secondary class – continuous from 1 to 96.
6. Top25perc: Percentage of new students from top 25% of Higher Secondary class – continuous from 9 to 100.
7. F.Undergrad: Number of full-time undergraduate students – continuous from 139 to 31643.
8. P.Undergrad: Number of part-time undergraduate students – continuous from 1 to 21836.
9. Outstate: Number of students for whom the particular college or university is Out-of-state tuition –

continuous from 2340 to 21700.

1. Room.Board: Cost of Room and board – continuous from 1780 to 8124.
2. Books: Estimated book costs for a student – continuous from 96 to 2340.
3. Personal: Estimated personal spending for a student – continuous from 250 to 6800.
4. PhD: Percentage of faculties with Ph.D.’s – continuous from 8 to 103.
5. Terminal: Percentage of faculties with terminal degree – continuous from 24 to 100.
6. S.F.Ratio: Student/faculty ratio – continuous from 2.5 to 39.8.
7. perc.alumni: Percentage of alumni who donate – continuous from 0 to 64.
8. Expend: The Instructional expenditure per student – continuous from 3186 to 56233.
9. Grad.Rate: Graduation rate – continuous from 10 to 118.

### Sample of the dataset:

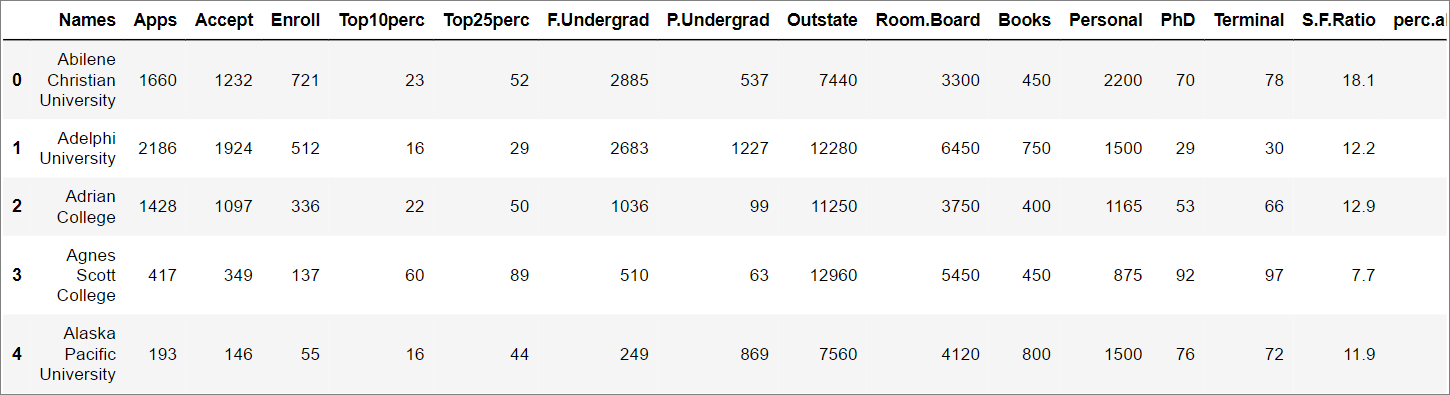


Table-2.1 Dataset Sample

### Exploratory Data Analysis:

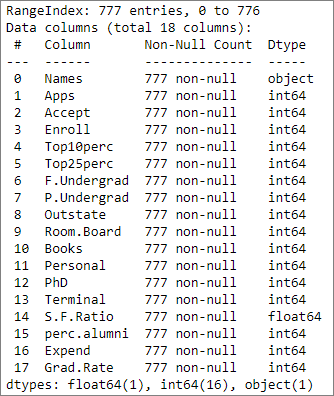


Table-2.2 Concise data summary

#### Let us check the type of variables in the data frame

There are a total of 777 observations and 18 columns in the dataset. Of the 18 columns, 1 is float, 16 are integer and 1 is of object data type.

#### Check for missing values in the dataset

From Table-2.2 we can see that all the columns have 777 non-null values and hence we have no missing values in the dataset.

#### Check for duplicate observations in the dataset

There are no duplicate rows in the dataset.

#### Data summary

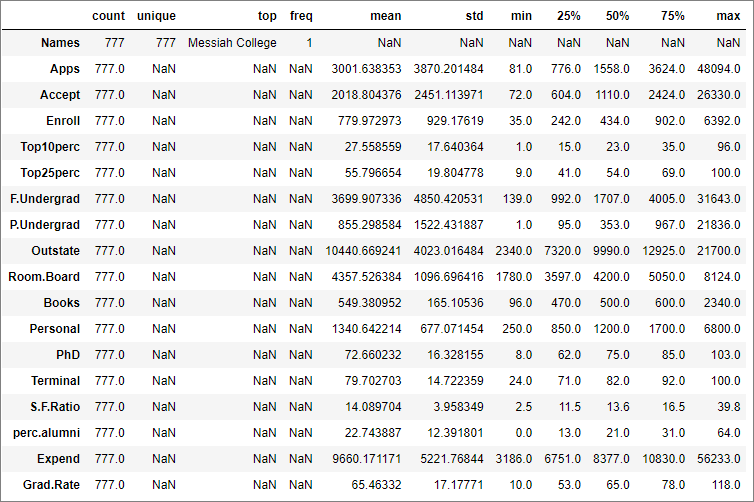


Table-2.3 Data Summary

The above data summary will be further explained in the univariate analysis section below.

#### Univariate Analysis:

The column ‘Names’ is ordinal in nature and will not be considered for the variable analysis. All the other variables are numeric and continuous in nature.

1. Apps

Number of applications received is a continuous variable with the below stats (refer Table 2.3):

Mean = 3001.64

Standard Deviation = 3870.20

Min value in dataset = 81

Max value in dataset = 48094

Range = Min – Max = 48013

Q1(1st Quartile) = 776

Q2(2nd Quartile)/Median = 1558

Q3(3rd Quartile) = 3624

IQR(Inter-Quartile Range) = Q3- Q1 = 2848

Quartile Min value = Q1 – 1.5 \* IQR = -3496

Quartile Max value = Q3 + 1.5 \* IQR = 7896

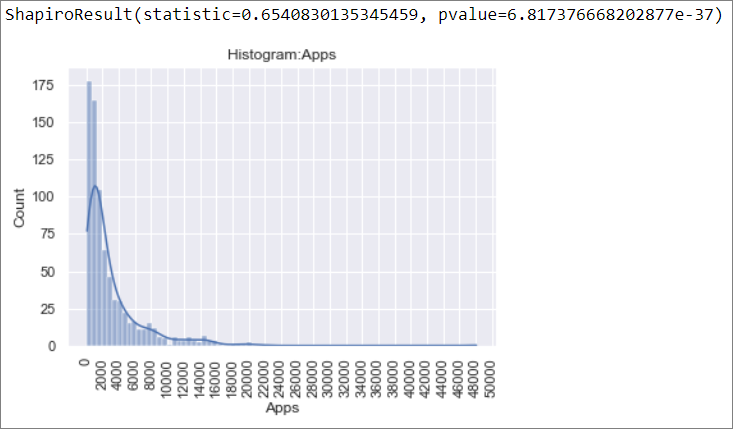


Figure-2.1 Histogram: Apps

Figure-2.1 depicts the histogram of “Apps” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 6.82e-37 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Apps” column data does not have a normal distribution, as per the provided observations.

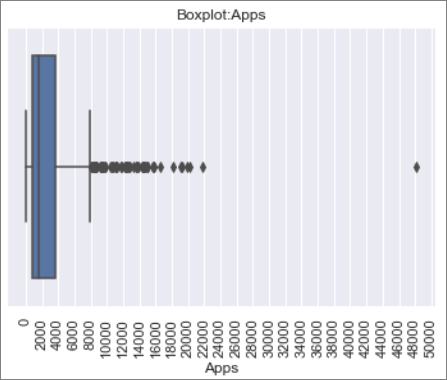


Figure-2.2 Boxplot: Apps

Figure-2.2 depicts the boxplot of “Apps” and we can see that there are outliers present in the data.

There are 70 observations which has outliers in Apps columns.

The top 5 and bottom 5 colleges w.r.t number of applications is given below:

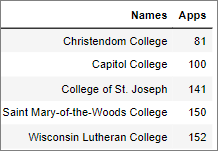
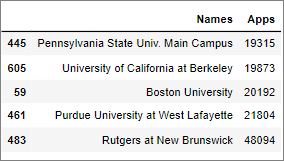
 

Table-2.4 Bottom 5 colleges w.r.t Apps Table-2.5 Top 5 colleges w.r.t Apps

1. Accept

Number of applications accepted is a continuous variable with the below stats (refer Table 2.3):

Mean = 2018.80

Standard Deviation = 2451.11

Min value in dataset = 72

Max value in dataset = 26330

Range = Min – Max = 26258

Q1(1st Quartile) = 604

Q2(2nd Quartile)/Median = 1110

Q3(3rd Quartile) = 2424

IQR(Inter-Quartile Range) = Q3- Q1 = 1820

Quartile Min value = Q1 – 1.5 \* IQR = -2126

Quartile Max value = Q3 + 1.5 \* IQR = 5154

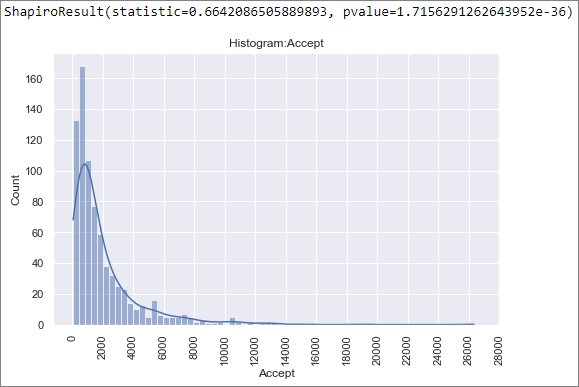


Figure-2.3 Histogram: Accept

Figure-2.3 depicts the histogram of “Accept” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 1.72e-36 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Accept” column data does not have a normal distribution, as per the provided observations.

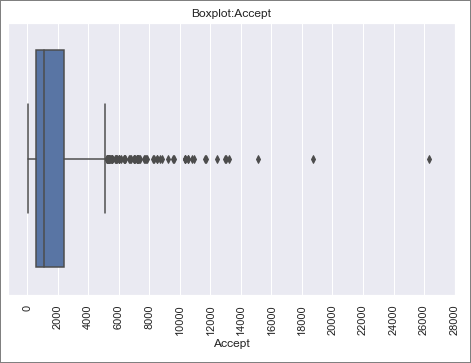


Figure-2.4 Boxplot: Accept

Figure-2.4 depicts the boxplot of “Accept” and we can see that there are outliers present in the data.

There are 73 observations which has outliers in Accept columns.

The top and bottom 5 colleges w.r.t number of Acceptance is given below:

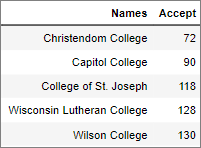
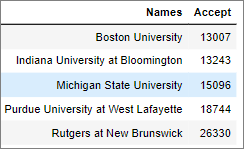
 

Table-2.6 Bottom 5 colleges w.r.t Accept Table-2.7 Top 5 colleges w.r.t Accept

1. Enroll

Number of new students enrolled is a continuous variable with the below stats (refer Table 2.3):

Mean = 779.97

Standard Deviation = 929.18

Min value in dataset = 35

Max value in dataset = 6392

Range = Min – Max = 6357

Q1(1st Quartile) = 242

Q2(2nd Quartile)/Median = 434

Q3(3rd Quartile) = 904

IQR(Inter-Quartile Range) = Q3- Q1 = 662

Quartile Min value = Q1 – 1.5 \* IQR = -420

Quartile Max value = Q3 + 1.5 \* IQR = 1566

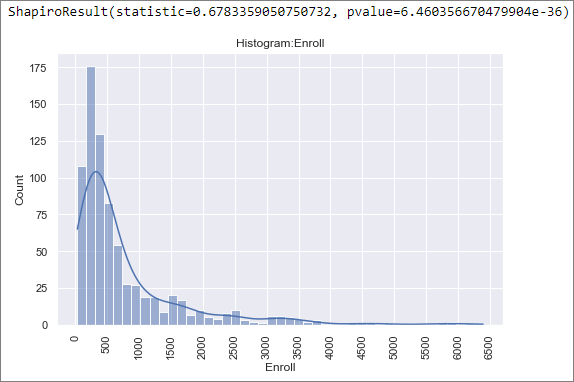


Figure-2.5 Histogram: Enroll

Figure-2.5 depicts the histogram of “Enroll” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 6.46e-36 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Enroll” column data does not have a normal distribution, as per the provided observations.

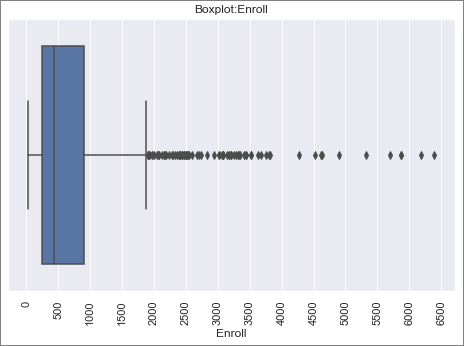


Figure-2.6 Boxplot: Enroll

Figure-2.6 depicts the boxplot of “Enroll” and we can see that there are outliers present in the data.

There are 6 observations which has outliers in Enroll columns.

The top and bottom 5 colleges w.r.t number of Enroll is given below:

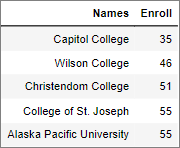
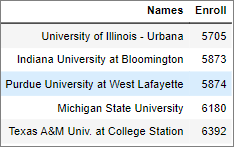
 

Table-2.8 Bottom 5 colleges w.r.t Enroll Table-2.9 Top 5 colleges w.r.t Enroll

1. Top10perc

Number of new students enrolled is a continuous variable with the below stats (refer Table 2.3):

Mean = 27.56

Standard Deviation = 17.64

Min value in dataset = 1

Max value in dataset = 96

Range = Min – Max = 95

Q1(1st Quartile) = 15

Q2(2nd Quartile)/Median = 23

Q3(3rd Quartile) = 35

IQR(Inter-Quartile Range) = Q3- Q1 = 20

Quartile Min value = Q1 – 1.5 \* IQR = -15

Quartile Max value = Q3 + 1.5 \* IQR = 65

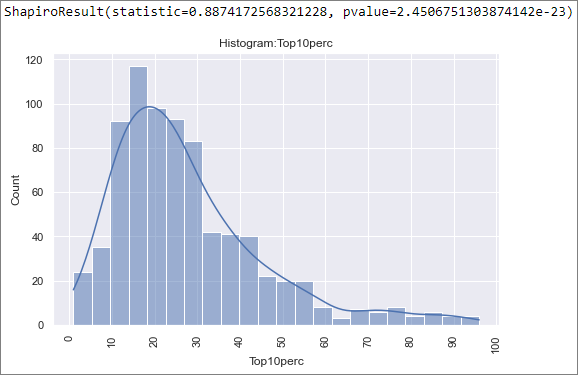


Figure-2.7 Histogram: Top10perc

Figure-2.7 depicts the histogram of “Top10perc” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 2.45e-23 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Top10perc” column data does not have a normal distribution, as per the provided observations.

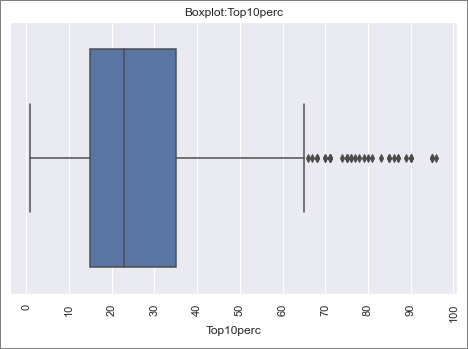


Figure-2.8 Boxplot: Top10perc

Figure-2.8 depicts the boxplot of “Top10perc” and we can see that there are outliers present in the data. There are 39 observations which has outliers in Top10perc columns.

The top and bottom 5 colleges w.r.t percent value of Top10perc is given below:

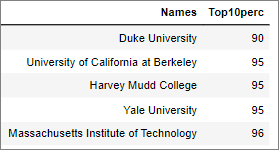
 

Table-2.10 Bottom 5 colleges w.r.t Top10perc Table-2.11 Top 5 colleges w.r.t Top10perc

1. Top25perc

Percentage of new students from top 25% of Higher Secondary class with the below stats (refer Table 2.3):

Mean = 55.80

Standard Deviation = 19.80

Min value in dataset = 9

Max value in dataset = 100

Range = Min – Max = 91

Q1(1st Quartile) = 41

Q2(2nd Quartile)/Median = 54

Q3(3rd Quartile) = 69

IQR(Inter-Quartile Range) = Q3- Q1 = 28

Quartile Min value = Q1 – 1.5 \* IQR = -1 which is less than min value of dataset, hence 9

Quartile Max value = Q3 + 1.5 \* IQR = 111 which is greater than max value of dataset hence 100

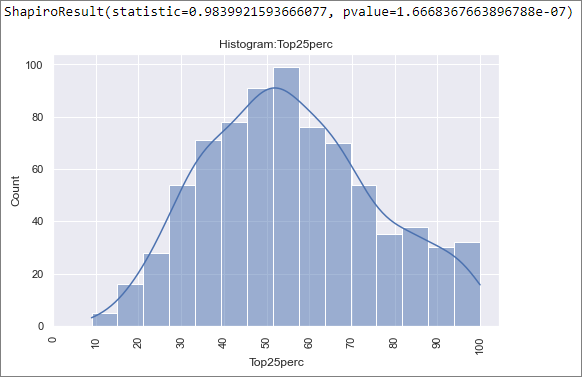


Figure-2.9 Histogram: Top25perc

Figure-2.9 depicts the histogram of “Top25perc” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 1.67e-7 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Top25perc” column data does not have a normal distribution, as per the provided observations.

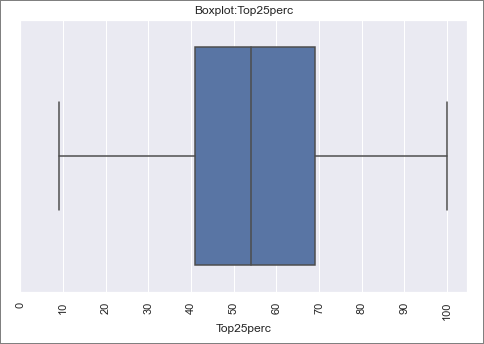


Figure-2.10 Boxplot: Top25perc

Figure-2.10 depicts the boxplot of “Top25perc” and we can see that there are no outliers present in the data.

The bottom 5 colleges w.r.t percent value and the colleges which has 100% of Top25perc is given below:

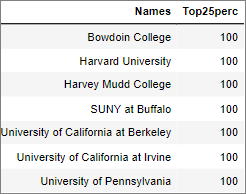
 

Table-2.12 Bottom 5 colleges w.r.t Top25perc Table-2.13 Colleges with Top25perc at 100%

1. F.Undergrad

Number of full-time undergraduate students with the below stats (refer Table 2.3):

Mean = 3699.91

Standard Deviation = 4850.42

Min value in dataset = 139

Max value in dataset = 31643

Range = Min – Max = 31504

Q1(1st Quartile) = 992

Q2(2nd Quartile)/Median = 1707

Q3(3rd Quartile) = 4005

IQR(Inter-Quartile Range) = Q3- Q1 = 3013

Quartile Min value = Q1 – 1.5 \* IQR = -3527.5

Quartile Max value = Q3 + 1.5 \* IQR = 8524.5

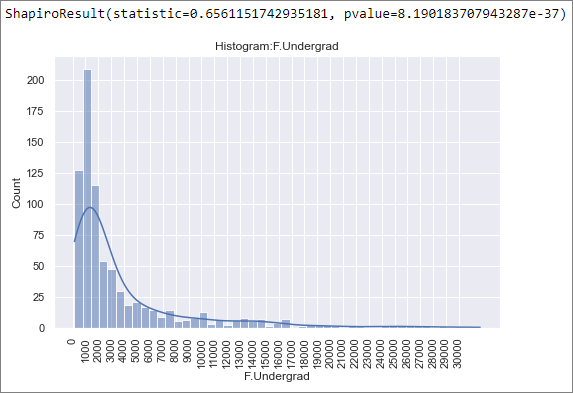


Figure-2.11 Histogram: F.Undergrad

Figure-2.11 depicts the histogram of “F.Undergrad” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 8.19e-37 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “F.Undergrad” column data does not have a normal distribution, as per the provided observations.

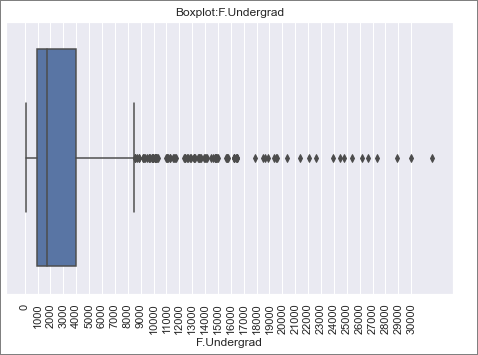


Figure-2.12 Boxplot: F.Undergrad

Figure-2.12 depicts the boxplot of “F.Undergrad” and we can see that there are outliers present in the data. There are 97 observations which has outliers in F.Undergrad column.

The top and bottom 5 colleges w.r.t value of F.Undergrad is given below:

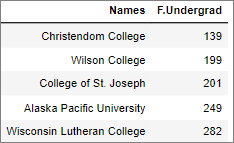
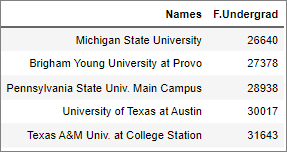
 

Table-2.14 Bottom 5 colleges w.r.t F.Undergrad Table-2.15 Top 5 colleges w.r.t F.Undergrad

1. P.Undergrad

Number of full-time undergraduate students with the below stats (refer Table 2.3):

Mean = 855.30

Standard Deviation = 1522.43

Min value in dataset = 1

Max value in dataset = 21836

Range = Min – Max = 21835

Q1(1st Quartile) = 95

Q2(2nd Quartile)/Median = 353

Q3(3rd Quartile) = 967

IQR(Inter-Quartile Range) = Q3- Q1 = 872

Quartile Min value = Q1 – 1.5 \* IQR = -1213

Quartile Max value = Q3 + 1.5 \* IQR = 2275

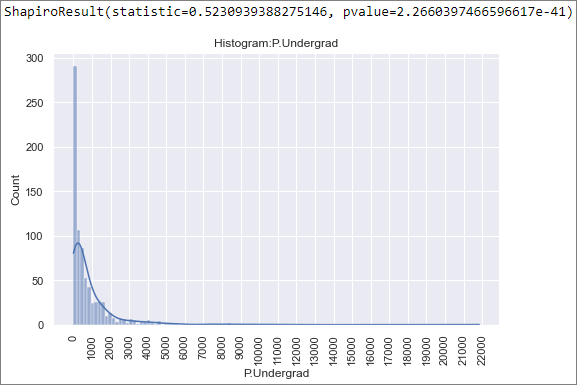


Figure-2.13 Histogram: P.Undergrad

Figure-2.13 depicts the histogram of “P.Undergrad” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 2.26e-41 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “P.Undergrad” column data does not have a normal distribution, as per the provided observations.

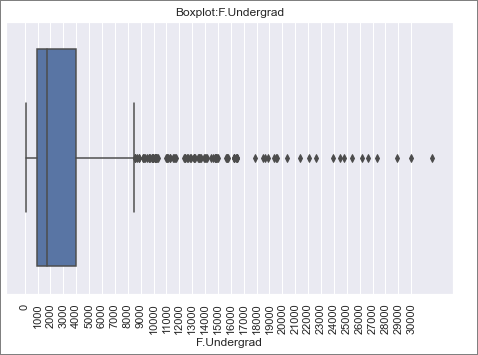


Figure-2.14 Boxplot: P.Undergrad

Figure-2.14 depicts the boxplot of “P.Undergrad” and we can see that there are outliers present in the data. There are 67 observations which has outliers in P.Undergrad column.

The top and bottom 5 colleges w.r.t value of P.Undergrad is given below:

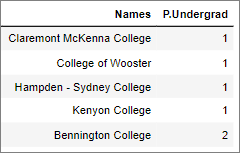
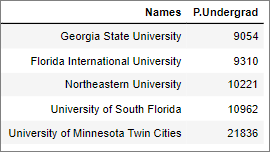
 

Table-2.16 Bottom 5 colleges w.r.t P.Undergrad Table-2.17 Top 5 colleges w.r.t P.Undergrad

1. Outstate

Number of students for whom the particular college or university is Out-of-state tuition, with the below stats (refer Table 2.3):

Mean = 10440.67

Standard Deviation = 4023.02

Min value in dataset = 2340

Max value in dataset = 21700

Range = Min – Max = 19360

Q1(1st Quartile) = 7320

Q2(2nd Quartile)/Median = 9990

Q3(3rd Quartile) = 12925

IQR(Inter-Quartile Range) = Q3- Q1 = 5605

Quartile Min value = Q1 – 1.5 \* IQR = -1087.5

Quartile Max value = Q3 + 1.5 \* IQR = 21332.5

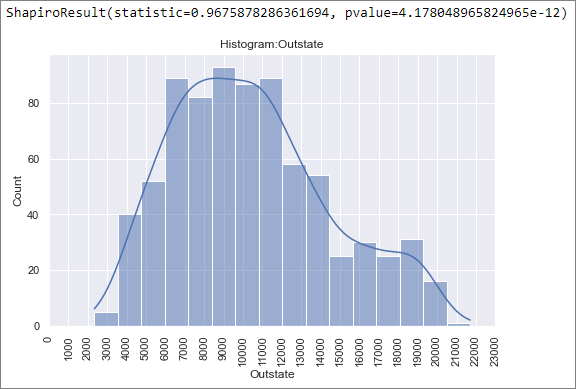


Figure-2.15 Histogram: Outstate

Figure-2.15 depicts the histogram of “Outstate” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 4.18e-12 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Outstate” column data does not have a normal distribution, as per the provided observations.

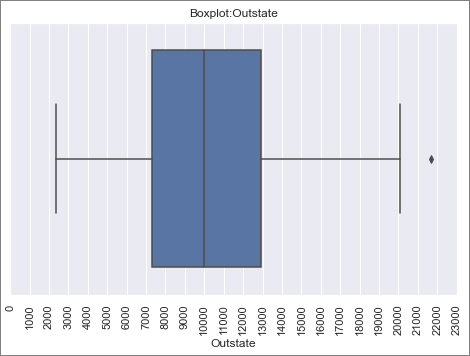


Figure-2.16 Boxplot: Outstate

Figure-2.16 depicts the boxplot of “Outstate” and we can see that there are outliers present in the data. There is 1 observation which has outliers in Outstate column.

The top and bottom 5 colleges w.r.t value of Outstate is given below:

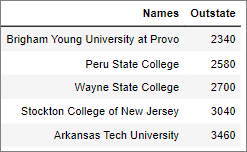
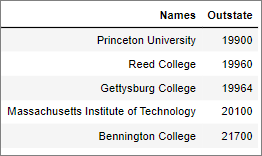
 

Table-2.18 Bottom 5 colleges w.r.t Outstate Table-2.19 Top 5 colleges w.r.t Outstate

1. Room.Board

Cost of Room and board is continuous with the below stats (refer Table 2.3):

Mean = 4357.53

Standard Deviation = 1096.7

Min value in dataset = 1780

Max value in dataset = 8124

Range = Min – Max = 6344

Q1(1st Quartile) = 3597

Q2(2nd Quartile)/Median = 4200

Q3(3rd Quartile) = 5050

IQR(Inter-Quartile Range) = Q3- Q1 = 1453

Quartile Min value = Q1 – 1.5 \* IQR = 1417.5

Quartile Max value = Q3 + 1.5 \* IQR = 7229.5

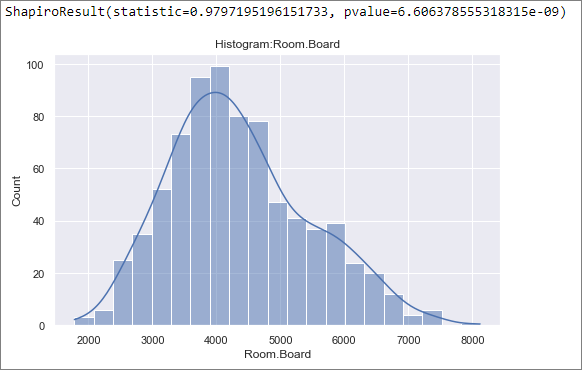


Figure-2.17 Histogram: Room.Board

Figure-2.17 depicts the histogram of “Room.Board” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 6.61e-9 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Room.Board” column data does not have a normal distribution, as per the provided observations.

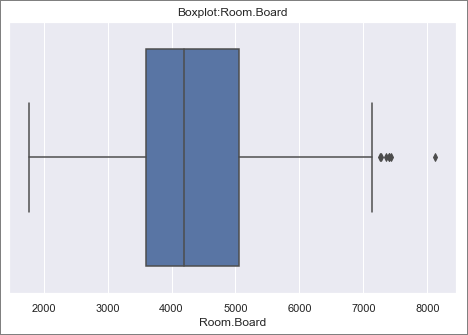


Figure-2.18 Boxplot: Room.Board

Figure-2.18 depicts the boxplot of “Room.Board” and we can see that there are outliers present in the data. There are 7 observation which has outliers in Room.Board column.

The top and bottom 5 colleges w.r.t value of Room.Board is given below:

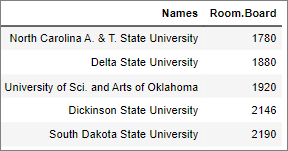
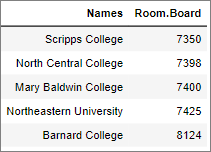
 

Table-2.20 Bottom 5 colleges w.r.t Room.Board Table-2.21 Top 5 colleges w.r.t Room.Board

1. Books

Estimated book costs for a student is continuous with the below stats (refer Table 2.3):

Mean = 549.38

Standard Deviation = 165.11

Min value in dataset = 96

Max value in dataset = 2340

Range = Min – Max = 2244

Q1(1st Quartile) = 470

Q2(2nd Quartile)/Median = 500

Q3(3rd Quartile) = 600

IQR(Inter-Quartile Range) = Q3- Q1 = 130

Quartile Min value = Q1 – 1.5 \* IQR = 275

Quartile Max value = Q3 + 1.5 \* IQR = 795

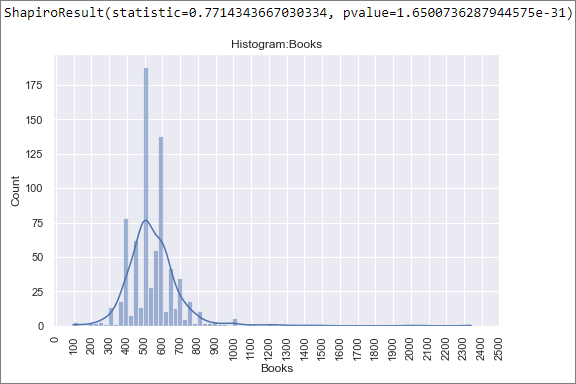


Figure-2.19 Histogram: Books

Figure-2.19 depicts the histogram of “Books” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 1.65e-31 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Books” column data does not have a normal distribution, as per the provided observations.

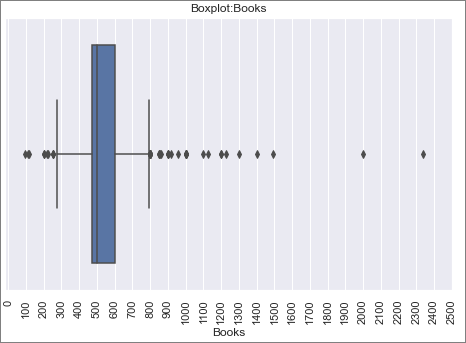


Figure-2.20 Boxplot: Books

Figure-2.20 depicts the boxplot of “Books” and we can see that there are outliers present in the data. There are 46 observations which has outliers in Book’s column.

The top and bottom 5 colleges w.r.t value of Books is given below:

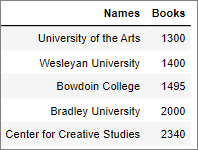
 

Table-2.22 Bottom 5 colleges w.r.t Books Table-2.23 Top 5 colleges w.r.t Books

1. Personal

Estimated personal spending for a student is continuous with the below stats (refer Table 2.3):

Mean = 1340.64

Standard Deviation = 677.07

Min value in dataset = 250

Max value in dataset = 6800

Range = Min – Max = 6550

Q1(1st Quartile) = 850

Q2(2nd Quartile)/Median = 1200

Q3(3rd Quartile) = 1700

IQR(Inter-Quartile Range) = Q3- Q1 = 850

Quartile Min value = Q1 – 1.5 \* IQR = -425

Quartile Max value = Q3 + 1.5 \* IQR = 2975

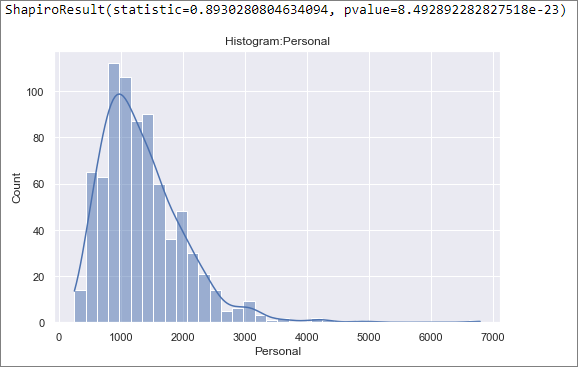


Figure-2.21 Histogram: Personal

Figure-2.21 depicts the histogram of “Personal” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 8.49e-23 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Personal” column data does not have a normal distribution, as per the provided observations.

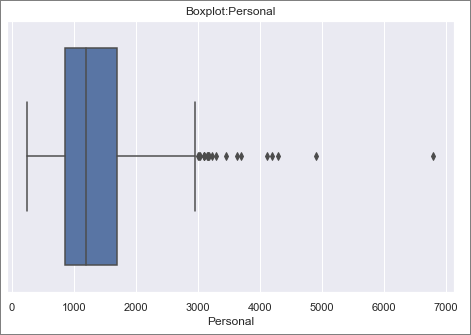


Figure-2.22 Boxplot: Personal

Figure-2.22 depicts the boxplot of “Personal” and we can see that there are outliers present in the data. There are 20 observations which has outliers in Personal column.

The top and bottom 5 colleges w.r.t value of Personal is given below:

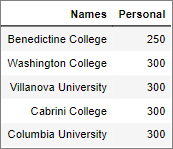
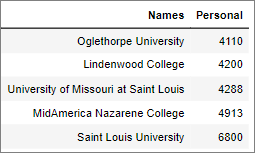
 

Table-2.24 Bottom 5 colleges w.r.t Personal Table-2.25 Top 5 colleges w.r.t Personal

1. PhD

Percentage of faculties with PhD’s is continuous.

There is one observation where percentage is greater than 100 as shown below:

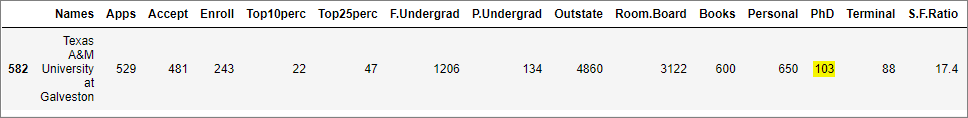


Table-2.26 PhD Data Error

We will impute the above erroneous PhD value with median value of PhD (75%).

Please find below the updated data summary table:

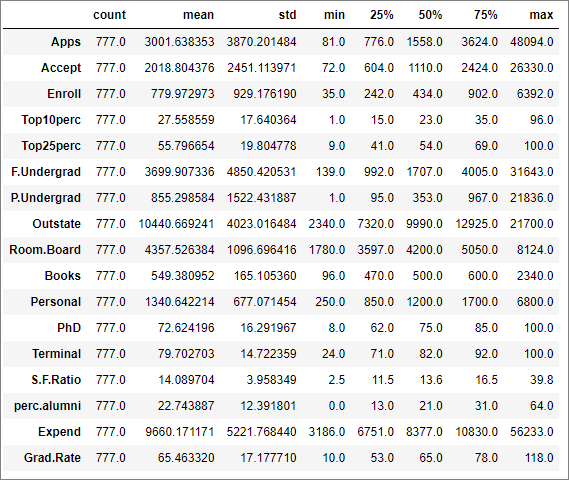


Table-2.27 Data summary

Let’s look at the important stats of PhD from Table 2.27

Mean = 72.62

Standard Deviation = 16.29

Min value in dataset = 8

Max value in dataset = 100

Range = Min – Max = 92

Q1(1st Quartile) = 62

Q2(2nd Quartile)/Median = 75

Q3(3rd Quartile) = 85

IQR(Inter-Quartile Range) = Q3- Q1 = 23

Quartile Min value = Q1 – 1.5 \* IQR = 27.5

Quartile Max value = Q3 + 1.5 \* IQR = 119.5, which is greater than dataset max value, hence 100.

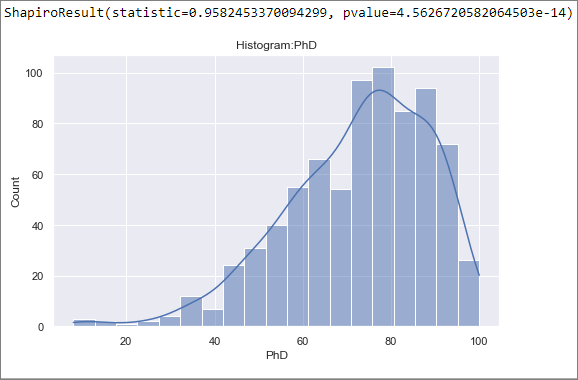


Figure-2.23 Histogram: PhD

Figure-2.23 depicts the histogram of “PhD” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 4.56e-14 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “PhD” column data does not have a normal distribution, as per the provided observations.

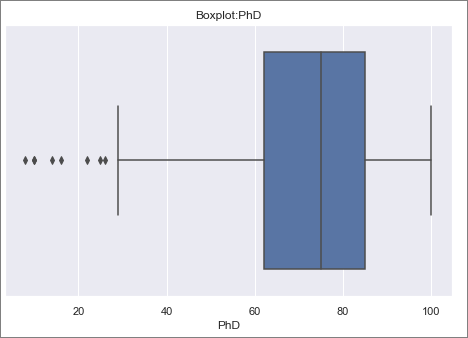


Figure-2.24 Boxplot: PhD

Figure-2.24 depicts the boxplot of “PhD” and we can see that there are outliers present in the data. There are 8 observations which has outliers in PhD column.

The top and bottom 5 colleges w.r.t value of PhD is given below:

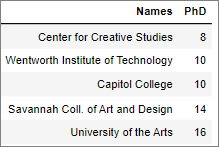
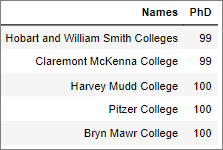
 

Table-2.28 Bottom 5 colleges w.r.t PhD Table-2.29 Top 5 colleges w.r.t PhD

1. Terminal

Percentage of faculties with terminal degree is continuous with the below stats (refer Table 2.27):

Mean = 79.70

Standard Deviation = 14.72

Min value in dataset = 24

Max value in dataset = 100

Range = Min – Max = 76

Q1(1st Quartile) = 71

Q2(2nd Quartile)/Median = 82

Q3(3rd Quartile) = 92

IQR(Inter-Quartile Range) = Q3- Q1 = 21

Quartile Min value = Q1 – 1.5 \* IQR = 39.5

Quartile Max value = Q3 + 1.5 \* IQR = 123.5, which is greater than max dataset value, hence 100.

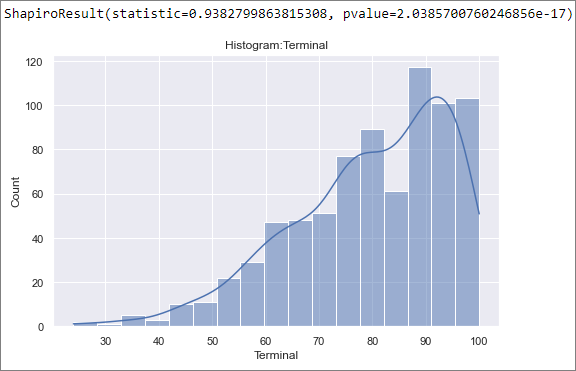


Figure-2.25 Histogram: Terminal

Figure-2.25 depicts the histogram of “Terminal” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 2.04e-17 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Terminal” column data does not have a normal distribution, as per the provided observations.

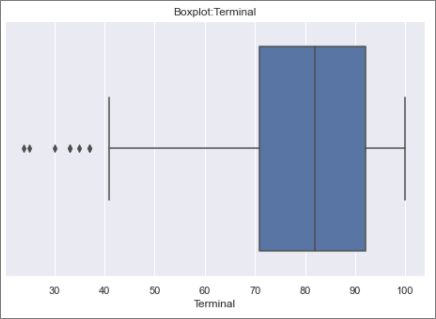


Figure-2.26 Boxplot: Terminal

Figure-2.26 depicts the boxplot of “Terminal” and we can see that there are outliers present in the data. There are 8 observations which has outliers in Terminal column.

The bottom 5 colleges w.r.t value of Terminal, and those with 100% Terminal value is given below:

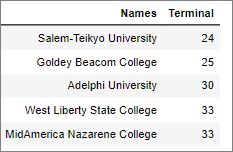
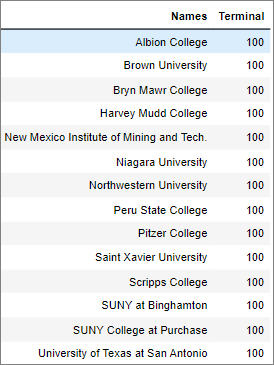
 

Table-2.30 Bottom 5 colleges w.r.t Terminal Table-2.31 Colleges with Terminal at 100%

1. S.F.Ratio

Student/faculty ratio is continuous with the below stats (refer Table 2.27):

Mean = 14.09

Standard Deviation = 3.96

Min value in dataset = 2.5

Max value in dataset = 39.8

Range = Min – Max = 37.3

Q1(1st Quartile) = 11.5

Q2(2nd Quartile)/Median = 13.6

Q3(3rd Quartile) = 16.5

IQR(Inter-Quartile Range) = Q3- Q1 = 5

Quartile Min value = Q1 – 1.5 \* IQR = 4

Quartile Max value = Q3 + 1.5 \* IQR = 24

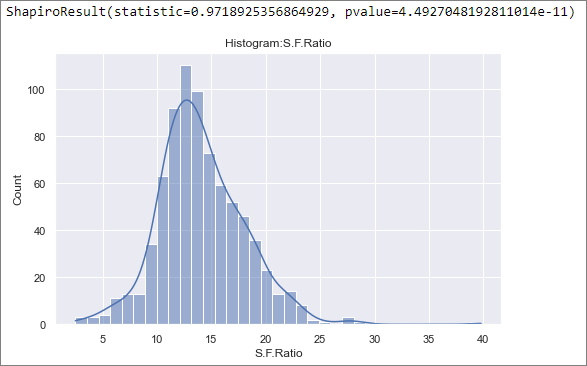


Figure-2.27 Histogram: S.F.Ratio

Figure-2.27 depicts the histogram of “S.F.Ratio” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 4.49e-11 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “S.F.Ratio” column data does not have a normal distribution, as per the provided observations.

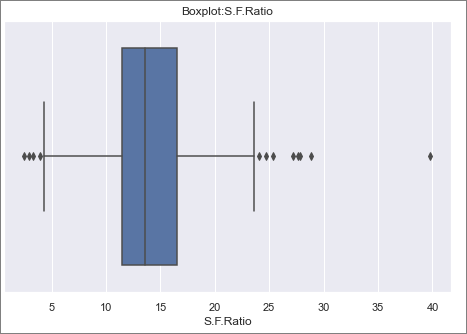


Figure-2.28 Boxplot: S.F.Ratio

Figure-2.28 depicts the boxplot of “S.F.Ratio” and we can see that there are outliers present in the data. There are 12 observations which has outliers in S.F.Ratio column.

The top and bottom 5 colleges w.r.t value of S.F.Ratio (lower is better) is given below:

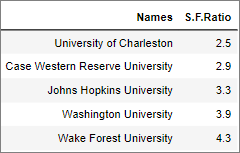
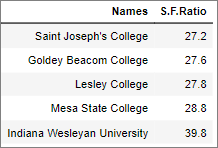
 

Table-2.32 Top 5 colleges w.r.t S.F.Ratio Table-2.33 Bottom 5 colleges w.r.t S.F.Ratio

1. perc.alumni

Percentage of alumni who donate is continuous with the below stats (refer Table 2.27):

Mean = 22.74

Standard Deviation = 12.39

Min value in dataset = 0

Max value in dataset = 64

Range = Min – Max = 64

Q1(1st Quartile) = 13

Q2(2nd Quartile)/Median = 21

Q3(3rd Quartile) = 31

IQR(Inter-Quartile Range) = Q3- Q1 = 18

Quartile Min value = Q1 – 1.5 \* IQR = -14, which is less than dataset min value, hence 0

Quartile Max value = Q3 + 1.5 \* IQR = 58

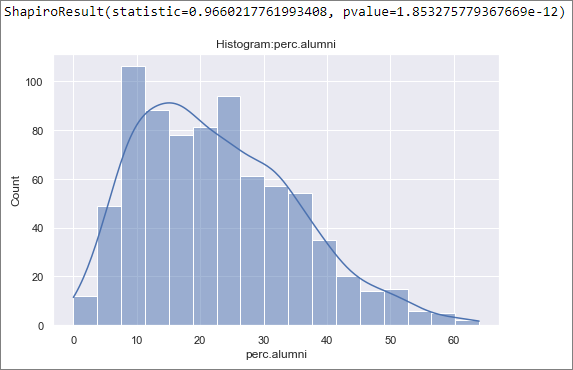


Figure-2.29 Histogram: perc.alumni

Figure-2.29 depicts the histogram of “perc.alumni” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 1.85e-12 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “perc.alumni” column data does not have a normal distribution, as per the provided observations.

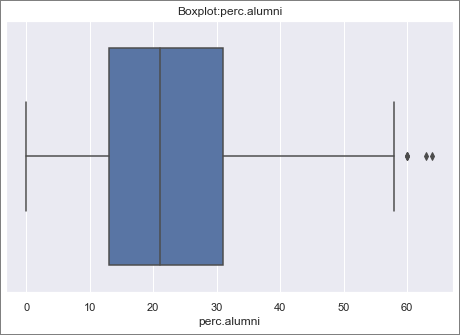


Figure-2.30 Boxplot: perc.alumni

Figure-2.30 depicts the boxplot of “perc.alumni” and we can see that there are outliers present in the data. There are 5 observations which has outliers in perc.alumni column.

The top and bottom 5 colleges w.r.t value of perc.alumni is given below:

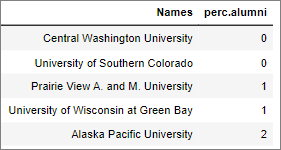
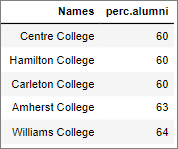
 

Table-2.34 Bottom 5 colleges w.r.t perc.alumni Table-2.35 Top 5 colleges w.r.t perc.alumni

1. Expend

The Instructional expenditure per student is continuous with the below stats (refer Table 2.27):

Mean = 9660.17

Standard Deviation = 5221.77

Min value in dataset = 3186

Max value in dataset = 56233

Range = Min – Max = 53047

Q1(1st Quartile) = 6751

Q2(2nd Quartile)/Median = 8377

Q3(3rd Quartile) = 10830

IQR(Inter-Quartile Range) = Q3- Q1 = 4079

Quartile Min value = Q1 – 1.5 \* IQR = 632.5, which is less than dataset min value, hence 3186

Quartile Max value = Q3 + 1.5 \* IQR = 16948.5

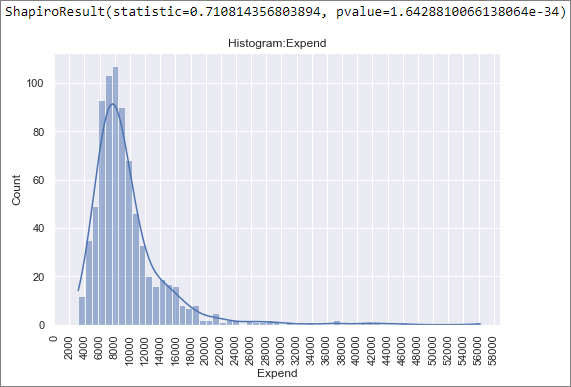


Figure-2.31 Histogram: Expend

Figure-2.31 depicts the histogram of “Expend” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 1.64e-34 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Expend” column data does not have a normal distribution, as per the provided observations.

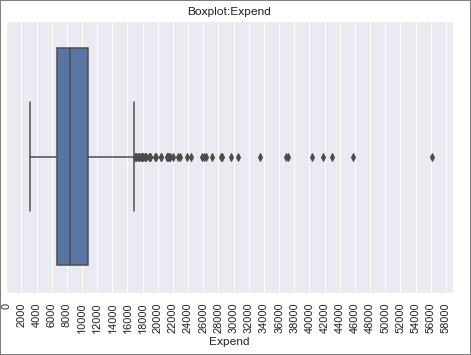


Figure-2.32 Boxplot: Expend

Figure-2.32 depicts the boxplot of “Expend” and we can see that there are outliers present in the data. There are 48 observations which has outliers in Expend column.

The top and bottom 5 colleges w.r.t value of Expend is given below:

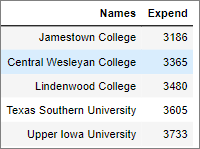
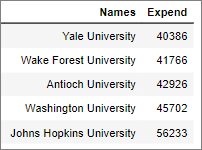
 

Table-2.36 Bottom 5 colleges w.r.t perc.alumni Table-2.37 Top 5 colleges w.r.t perc.alumni

1. Grad.Rate

Graduation rate is continuous.

There is one observation where percentage is greater than 100 as shown below:



Table-2.38 Grad.Rate Data Error

We will impute the above erroneous Grad.Rate value with median value of Grad.Rate(65).

Please find below the updated data summary table:

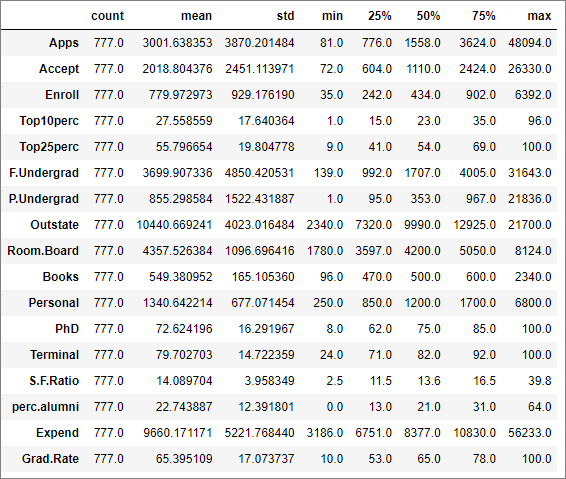


Table-2.39 Data summary

Let’s look at the important stats of Grad.Rate from Table 2.39

Mean = 65.40

Standard Deviation = 17.07

Min value in dataset = 10

Max value in dataset = 100

Range = Min – Max = 90

Q1(1st Quartile) = 53

Q2(2nd Quartile)/Median = 65

Q3(3rd Quartile) = 78

IQR(Inter-Quartile Range) = Q3- Q1 = 25

Quartile Min value = Q1 – 1.5 \* IQR = 15.5

Quartile Max value = Q3 + 1.5 \* IQR = 115.5, which is greater than dataset max value, hence 100.

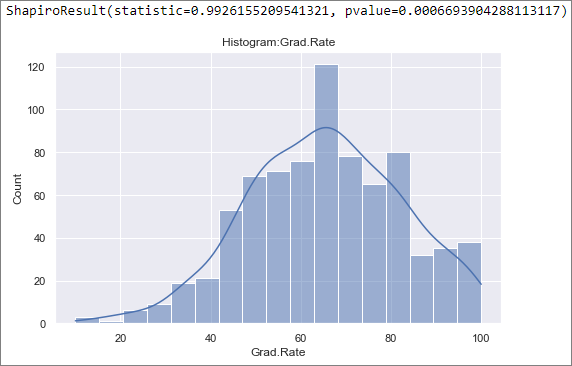


Figure-2.33 Histogram: Grad.Rate

Figure-2.33 depicts the histogram of “Grad.Rate” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 6.7e-4 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Grad.Rate” column data does not have a normal distribution.

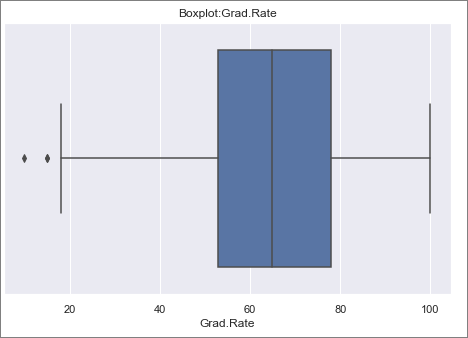


Figure-2.34 Boxplot of Grad.Rate

Figure-2.34 depicts the boxplot of “Grad.Rate” and we can see that there are outliers present in the data. There are 3 observations which has outliers in Grad.Rate column.

The bottom 5 colleges w.r.t Grad.Rate and colleges with 100% Grad.Rate is shown below:

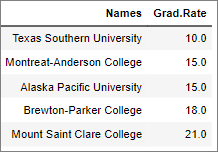
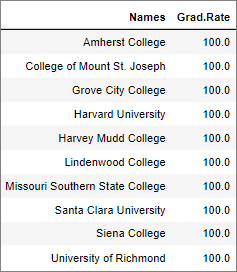
 

Table-2.40 Bottom 5 colleges w.r.t Grad.Rate Table-2.41 Colleges with 100% Grad.Rate

#### Bivariate Analysis:

Let us plot a heat map for the correlation matrix of given data frame.

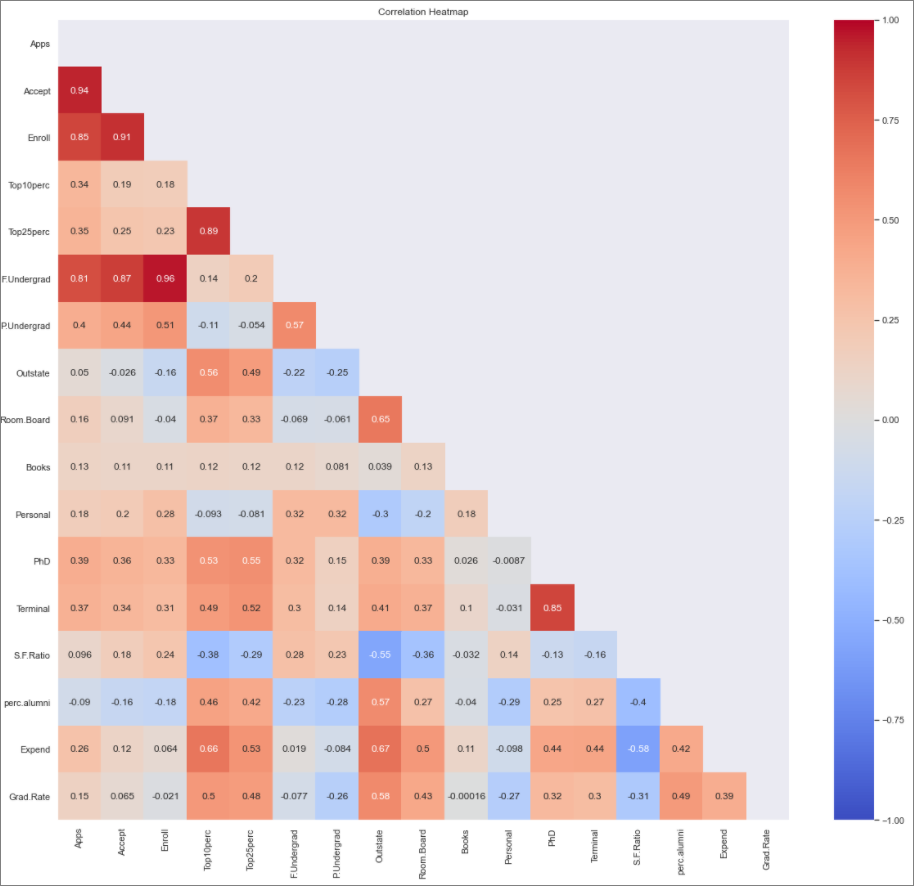


Table-2.42 Correlation matrix

We can see very strong positive correlation between the following variables:

* Apps and Accept

The number of applications vs number of acceptances by colleges follow a positive correlation, increase in one tends to increase in other. Either colleges are ready to accept in more students and increase the number of batches or students applying have a good idea on the possible number of seats and chances of acceptance.

* Enroll and Accept

The number of enrollments has a positive correlation with acceptance. Since students would apply to multiple colleges, the institutions must be accepting more applications than seats and then decide via response of enrolling students.

* Enroll and Apps

This tells us that all 3 variables Apps, Accept and Enroll follows a positive correlation and tends to increase or decrease together.

* F.Undergrad and Apps

Full time Undergraduate students has a positive correlation with applications and that is logical. This will also tell us that full time undergraduate will also have a positive correlation with Accept and Enroll.

* F.Undergrad and Accept
* F.Undergrad and Enroll
* PhD and Terminal

PhD is a terminal degree and hence has a direct positive correlation with percentage of faculties with terminal degrees.

We can see significant negative correlation between the following variables:

* S.F.Ratio and Outstate

This provides an important information to colleges. Out of state students will increase if the colleges are able to provide a greater number of faculty per student. Please note that for S.F.Ratio, the value when lower, is considered better as it means more faculty per student.

* S.F.Ratio and Expend

When a college provides better S.F.Ratio, it also follows that the college needs to spend more on the faculty. This is expected.

* S.F.Ratio and Top10perc

This is an important point of consideration for colleges. To attract students from the Top 10% of HS degree students, the college should be willing to offer more faculty per student.

#### Other Insights:

* Mean application acceptance ratio of all colleges is :74.69%

This is indication to prospect students to apply to multiple colleges to ensure application in at least one college.

* Of the number of students application accepted, average enrollment across colleges is :41.2%

This is indication to colleges as to, how many more applications they need to receive than seats to ensure that they have enough students for classes.

* We can see that students prefer colleges based on the academics of the faculty. All the colleges which have at least 52% of its students coming from Top10% of the Higher secondary class has, on an average, more than 90% of their faculty with PhD degrees. Colleges which have at least 90% of their students coming from Top25% of the Higher Secondary class, on an average more than 90% of their faculty with PhD Degrees.

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

In Principal Component Analysis(PCA) we are attempting to find the components which has maximum variance. Hence if the data variables have different scales, the change in one variable might bias the scales of variance in an unfair manner. Hence scaling is performed to normalize the data within a range. In this example we have number of applications, data in percentage form for Graduation rate, cost of living in currency rate etc. which follows their own scale. Hence its necessary to scale the data using methods like zscale or min-max scale. Scaling also helps in reducing the magnitude of processing required for PCA, thus making the process faster.

Before scaling we have seen the data summary in Table 2.39.

After performing scaling via zscore normalization the data summary is as follows:

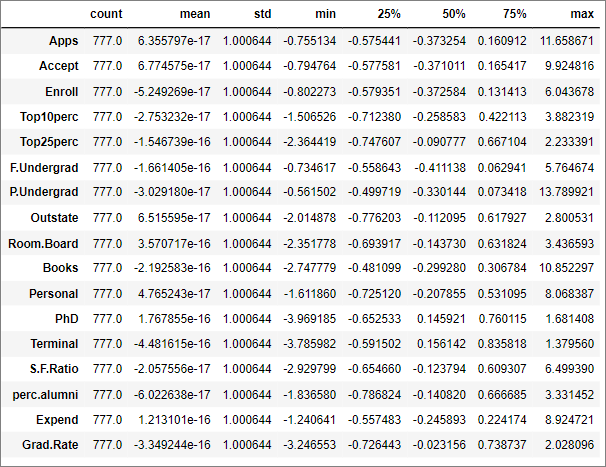


Table-2.43 Scaled – Data Summary\

From table 2.43, we can see that the mean of all the columns is near to 0 and standard deviation is near to 1. Hence the zscore normalization has rescaled the features to have the properties of a standard normal distribution.

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

Covariance between 2 variables indicates the direction of linear relationship between 2 variables. It is affected by the scale of the variables involved. Correlation between 2 variables indicates the direction as well as how strong the relationship is.

Covariance between 2 variables x and y is denoted as :

COV(x ,y) =

where n is the number of samples, is the mean of x and is the mean of y

Correlation between 2 variables x and y is denoted as:

COR(x ,y) =

where is the variance of x and is the variance of y.

For the zscore scaled data , standard deviation is 1, hence variance value will also be 1.

Hence for our scaled data, COR(x, y) = COV(x, y)

Let’s look at the correlation matrix as shown in Table 2.44:

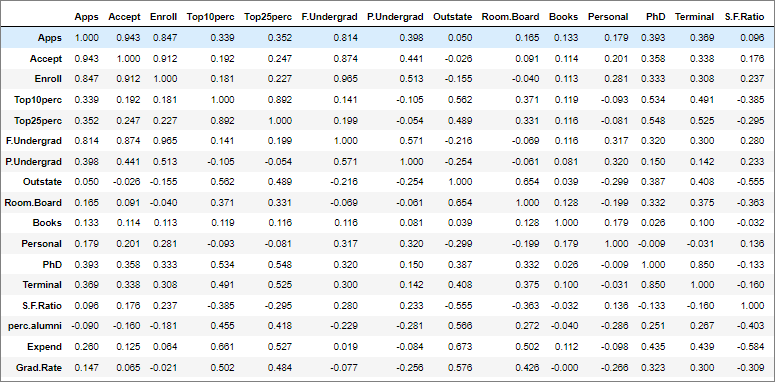


Table-2.44 Scaled – Data Correlation matrix

The covariance matrix for the scaled data is shown below in Table 2.45.

We can see that the values in both the tables are nearly the same.

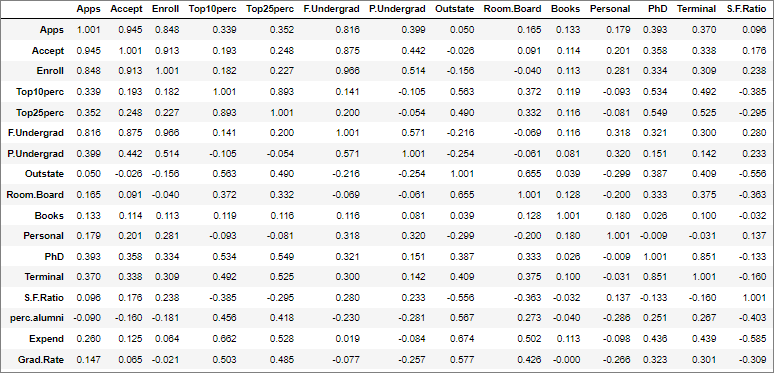


Table-2.45 Scaled – Data Covariance matrix

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

Let’s check the boxplot for scaled data as shown below in Figure 2.35

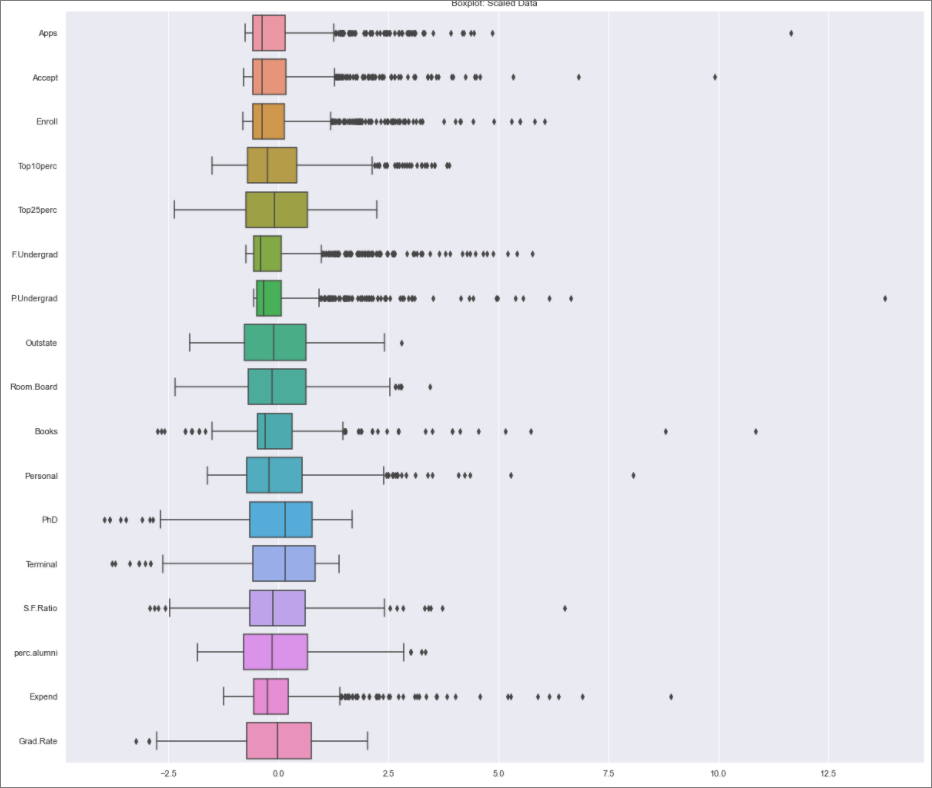


Figure-2.35 Boxplot of Scaled Data

Let’s check the boxplot for original data as shown below in Figure 2.36

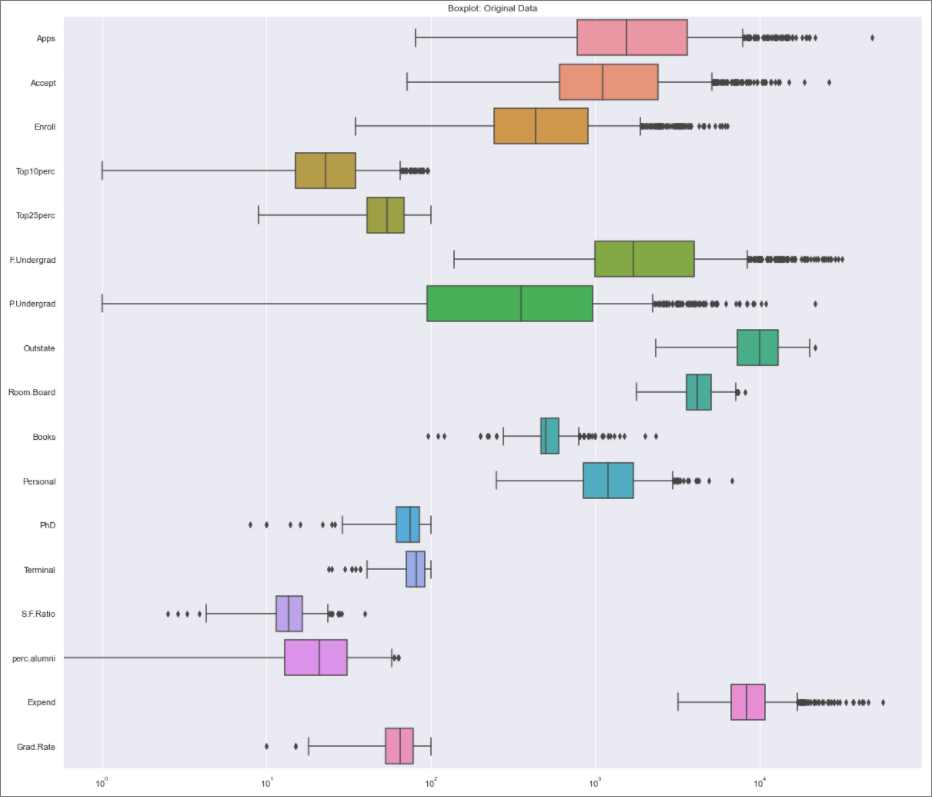


Figure-2.36 Boxplot of Original Data

For the original data, the scale of x axis has been changed to logarithmic, so that we can view all the data.

We can see that the same number of outliers are present in both the scaled and original data and they are relative to the position in scaled as it was in original. This tells us that the original data has been scaled to follow a standard normal distribution with mean 0 and standard deviation 1, but the displacement of data in relative to each other has been retained.

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

Please find below the eigenvalues derived from the scaled dataset.

array([5.46384062, 4.4841809 , 1.17453449, 0.99793705, 0.93465879,

0.84734458, 0.60586408, 0.58783511, 0.53014189, 0.40354672,

0.3131535 , 0.22030447, 0.16777512, 0.14299858, 0.08806661,

0.03671503, 0.02300969])

Please find below the eigenvectors for the scaled dataset.

array([[ 2.47532537e-01, 2.06299756e-01, 1.75138842e-01,

3.53990557e-01, 3.43702467e-01, 1.53527590e-01,

2.57859287e-02, 2.94965994e-01, 2.48896885e-01,

6.42827785e-02, -4.25981738e-02, 3.19579232e-01,

3.16776477e-01, -1.77164491e-01, 2.05418014e-01,

3.18605544e-01, 2.55626868e-01],

[ 3.32426877e-01, 3.72875058e-01, 4.04250640e-01,

-8.15077765e-02, -4.39140881e-02, 4.18089297e-01,

3.15119134e-01, -2.48748173e-01, -1.36927922e-01,

5.66066526e-02, 2.19634151e-01, 5.88540448e-02,

4.70008357e-02, 2.46076655e-01, -2.46023579e-01,

-1.30727369e-01, -1.68686285e-01],

[-5.96777039e-02, -9.77593828e-02, -8.13575684e-02,

3.42794420e-02, -2.55085723e-02, -6.05628562e-02,

1.37966832e-01, 4.81491197e-02, 1.52143504e-01,

6.79693620e-01, 4.95300078e-01, -1.31903121e-01,

-7.09715596e-02, -2.91240091e-01, -1.47090056e-01,

2.27918829e-01, -2.05241119e-01],

[ 2.85096690e-01, 2.71787742e-01, 1.63457087e-01,

-5.51796334e-02, -1.15855059e-01, 1.01292070e-01,

-1.58951251e-01, 1.36572002e-01, 1.91576166e-01,

7.13795514e-02, -2.49032745e-01, -5.29195380e-01,

-5.18144779e-01, -1.68468205e-01, 1.61146847e-02,

8.60666647e-02, 2.43113496e-01],

[ 1.70019622e-04, 5.06165527e-02, -5.87975580e-02,

-3.94019023e-01, -4.23913172e-01, -4.54078528e-02,

3.06058094e-01, 2.20139878e-01, 5.56677349e-01,

-1.31994951e-01, -2.17223922e-01, 1.50814264e-01,

2.14615639e-01, -7.70591919e-02, -2.15815334e-01,

7.49927681e-02, -1.15718370e-01],

[-1.22811932e-02, 1.12771266e-02, -4.01196268e-02,

-5.41942245e-02, 3.08687410e-02, -4.17274288e-02,

-1.93339422e-01, -2.67734158e-02, 1.67220434e-01,

6.40262169e-01, -3.37697849e-01, 8.30924584e-02,

1.49282189e-01, 4.85844391e-01, -4.71502314e-02,

-2.97457585e-01, 2.15656623e-01],

[-3.06596836e-02, -2.81197218e-03, -2.35400786e-02,

-1.64625571e-01, -1.25835762e-01, -2.40116146e-02,

2.51111263e-02, 1.12135600e-01, 2.18058092e-01,

-1.50371205e-01, 6.37299867e-01, -2.57764237e-03,

-4.17554494e-02, 2.13980087e-01, 2.21096515e-01,

-2.24278705e-01, 5.64575953e-01],

[-1.03577277e-01, -5.55282328e-02, 5.83157504e-02,

-1.29536151e-01, -1.08116507e-01, 7.82383577e-02,

5.70135018e-01, 1.46220489e-02, -2.12995569e-01,

2.07821399e-01, -2.08241997e-01, -7.72742010e-02,

-1.40630479e-02, -7.54935950e-02, 6.89202681e-01,

-6.26002869e-02, 1.48592841e-02],

[-8.93297613e-02, -1.76710689e-01, -1.28035155e-01,

3.38331896e-01, 4.01041645e-01, -5.92590638e-02,

5.61028863e-01, -1.20304645e-03, 2.80843531e-01,

-1.35859726e-01, -8.38764902e-02, -1.86611542e-01,

-2.59934804e-01, 2.79116404e-01, -2.46797747e-01,

-5.30219126e-02, 4.94830868e-02],

[ 5.07359465e-02, 4.05496164e-02, 3.14617978e-02,

6.40272957e-02, 1.43050699e-02, 1.74906259e-02,

-2.23974904e-01, 1.81786941e-01, 2.93448662e-01,

-8.50316393e-02, 1.41756567e-01, -1.18174458e-01,

-7.70099176e-02, 4.73970264e-01, 4.23595852e-01,

1.39085118e-01, -5.91786591e-01],

[ 4.31741045e-02, -5.86528967e-02, -6.83645378e-02,

-8.77395752e-03, -2.74016941e-01, -7.99532742e-02,

1.02020732e-01, 1.46199343e-01, -3.60912134e-01,

3.34809127e-02, -2.07268141e-02, 3.98157243e-02,

-6.25480083e-02, 4.42559093e-01, -1.33965467e-01,

6.89350125e-01, 2.24433233e-01],

[ 2.33859080e-02, -1.46702734e-01, 1.21051996e-02,

3.78844097e-02, -9.11239455e-02, 5.76487187e-02,

-6.28499474e-02, -8.22908398e-01, 3.52608505e-01,

-2.76901401e-02, -4.01144680e-02, 1.97930769e-02,

1.86203119e-02, -1.06883947e-02, 1.80051409e-01,

3.27702760e-01, 1.27378733e-01],

[ 5.97045093e-01, 2.92920428e-01, -4.46935450e-01,

-1.60339719e-03, 2.48611440e-02, -5.25319912e-01,

1.26509734e-01, -1.45891405e-01, -7.18354751e-02,

9.82435927e-03, 4.00538504e-02, 1.11116039e-01,

-3.94111128e-02, -1.66448115e-02, 1.02793833e-01,

-9.04328465e-02, -6.58020893e-02],

[ 6.60495278e-02, 2.62518280e-02, -7.39350561e-02,

-1.07536920e-01, 1.49296761e-01, -4.28876301e-02,

1.94333337e-02, -2.58416552e-02, -5.72153462e-02,

-6.82096145e-02, 3.01808078e-02, -6.96287228e-01,

6.69882232e-01, 4.10690876e-02, -3.13661076e-02,

7.37876749e-02, 4.72561238e-02],

[ 1.32822763e-01, -1.44994783e-01, 2.99576263e-02,

6.98036916e-01, -6.17396656e-01, 1.00315012e-02,

2.09983126e-02, 3.95392860e-02, 3.69908819e-03,

-9.57613442e-03, -2.65877604e-03, -1.10964608e-01,

1.57564764e-01, -2.10406761e-02, -8.21266099e-03,

-2.28296993e-01, -4.01950518e-03],

[ 4.59049537e-01, -5.17547918e-01, -4.05328352e-01,

-1.47935152e-01, 5.19055777e-02, 5.60675304e-01,

-5.31278335e-02, 1.02095362e-01, -2.56620894e-02,

2.85111489e-03, -1.31210253e-02, 2.98639738e-02,

-2.70252109e-02, -2.11914505e-02, 3.92402220e-03,

-4.42782385e-02, -7.66081425e-03],

[ 3.59723152e-01, -5.44054193e-01, 6.09048127e-01,

-1.44813035e-01, 8.04965700e-02, -4.14017608e-01,

8.72484281e-03, 5.14125438e-02, 1.23870180e-03,

7.40409608e-04, -1.20672459e-03, 1.40741118e-02,

6.18241305e-03, -2.22862123e-03, -1.88728215e-02,

-3.56405319e-02, -1.45932934e-02]])

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

Please find below the eigenvectors moved into a data frame with values against the original features.

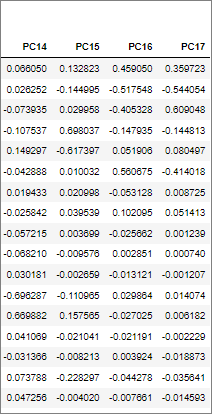
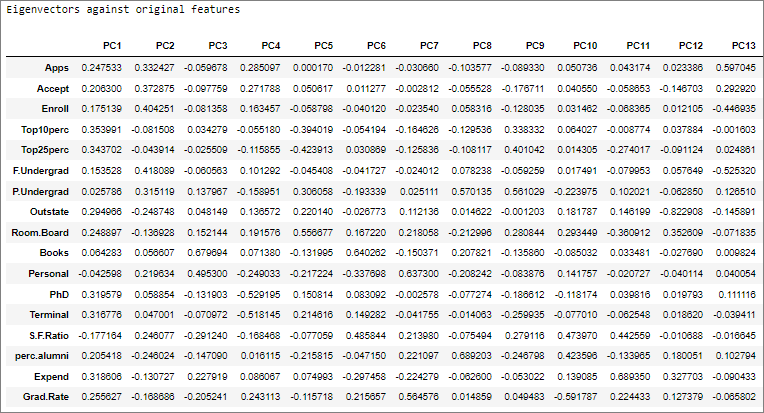


Table-2.46 Eigenvectors against original features

## 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 first principal component (PC1 in Table 2.46) can be expressed as shown below:

PC1 = 0.25 Apps + 0.21 Accept + 0.18 Enroll + 0.35 Top10perc + 0.34 Top25perc + 0.15 F.Undergrad +

0.03 P.Undergrad + 0.29 Outstate + 0.25 Room.Board + 0.06 Books - 0.04 Personal + 0.32 PhD +

0.32 Terminal – 0.18 S.F.Ratio + 0.21 perc.alumni + 0.32 Expend + 0.26 Grad.Rate

## 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 eigen values are as given below:

array([5.46384062, 4.4841809 , 1.17453449, 0.99793705, 0.93465879,

0.84734458, 0.60586408, 0.58783511, 0.53014189, 0.40354672,

0.3131535 , 0.22030447, 0.16777512, 0.14299858, 0.08806661,

0.03671503, 0.02300969])

The cumulative values of the eigen values are as follows:

array([ 5.46384062, 9.94802151, 11.12255601, 12.12049306, 13.05515185,

13.90249643, 14.50836051, 15.09619562, 15.6263375 , 16.02988422,

16.34303772, 16.56334219, 16.73111731, 16.87411589, 16.9621825 ,

16.99889753, 17.02190722])

The total variance adds up to 17.02190722, hence we can say that PC1 with eigen value 5.46384062

contribute to 5.46384062/17.02190722 = 0.32098874 % of the explained variance. Similarly, we can

calculate the contribution to variance by the other components, which is given below.

array([0.32098874, 0.26343587, 0.06900134, 0.05862663, 0.05490917,

0.04977965, 0.0355932 , 0.03453403, 0.03114468, 0.02370749,

0.01839709, 0.01294241, 0.00985642, 0.00840086, 0.00517372,

0.00215693, 0.00135177])

A scree plot displays how much variation each principal component captures from the data.

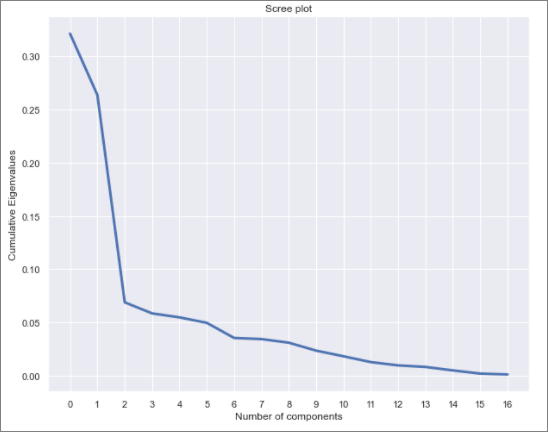


Figure-2.37 Scree Plot

If we consider cumulative values of the explained variance ratio and plot a graph as shown below:

array([0.32098874, 0.58442461, 0.65342596, 0.71205259, 0.76696176,

0.81674141, 0.8523346 , 0.88686863, 0.91801332, 0.94172081,

0.96011789, 0.9730603 , 0.98291673, 0.99131758, 0.9964913 ,

0.99864823, 1. ])

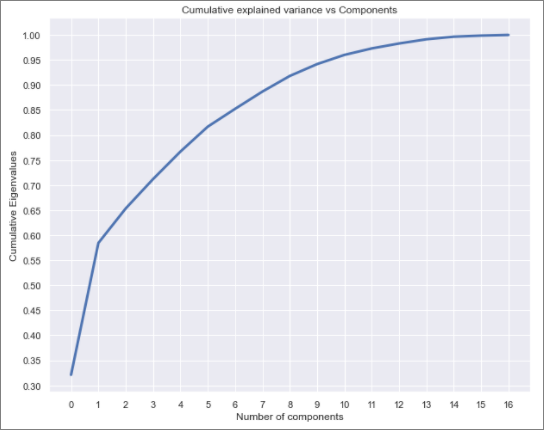


Figure-2.38 Cumulative Explained Variance Ratio Plot

The plot in Figure 2.38 tells us that the first 8 components capture 88.69% of the variation in data or

captures 88.69% of the information in the original 17 variables. This helps us in reducing dimensionality (from 17 to 8), while retaining significant information. An accepted practice is to consider enough

components that will explain at least 85% of the variance.

The eigenvectors indicate the direction of the principal components. When the original data is multiplied by the eigen vectors, they are re-oriented to form the orthogonal axes of the principal components.

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

Principal Component Analysis helps to simplify the complexity in data with high dimensions(17 in this case, transformed to 8 components), still retaining the trends and patterns. It helps to reduce the curse of dimensionality and addresses the multi-collinearity between data variables by converting the original data into orthogonal components which maximize variance. The components are chosen in such a way that information loss is minimized.

Let’s look at the 8 principal components chosen, which account for 88.69% of the information in the original dataset. For this we need to identify the features which have maximum loading across the components. We will first plot the component loading on a heatmap. For each feature, we find the maximum loading value across the components and highlight it. Features marked with rectangular red box are the one having maximum loading on the respective component. We consider these marked features to decide the context that the component represents. The heatmap is given below (Table 2.47).

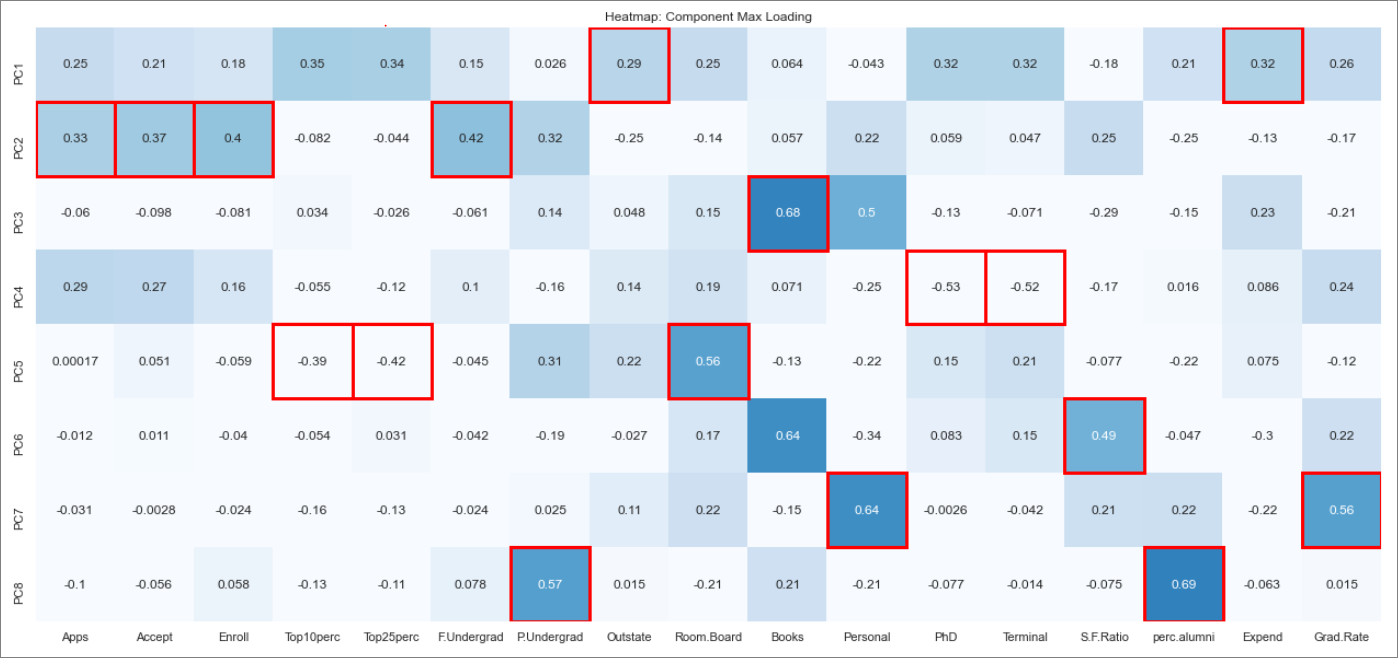


Table-2.47 Heatmap Component Max Loading

PC1: For first principal component we can see that loading is maximum for ‘Outstate’ and ‘Expend’ and

they have a positive correlation. We can surmise that, institutions which are opted by maximum out-of-state students, do so considering the prestige of the institution. Hence institutions will have to

spend more on faculty and facility to retain this edge. This component primarily is a measure of the institution’s expenditure per student and an indirect measure of how premium the institution is.

PC2: For second principal component we can see that loading is maximum for ‘Apps’, ’Accept’, ’Enroll’ and ‘F.Undergrad’ and they all are positively correlated. This component primarily is a measure of the institution’s admission related information, and number of full-time undergrad students is a result of the admission related information.

PC3: Third principal component has maximum loading ‘Books’, and the correlation is positive. This component primarily is a measure of the expenditure on books by students.

PC4: Fourth principal component has maximum loading on ‘PhD’ and ‘Terminal’. The relation is negative correlation. This component primarily is a measure of the qualification of the professors in the college. Hence higher the component value, lower is the qualification.

PC5: Fifth principal component has maximum loading on ‘Top10perc’, ’Top25perc’ and ‘Room.Board’. The relation is negative correlation for the first 2 variable and positive for the last. This component is a combination of the caliber of joining students and the cost of room and board, and the cost has a higher weightage.

PC6: Sixth principal component has maximum loading on ‘S.F.Ratio’ with a positive correlation. Hence higher the component value higher is the student to faculty ratio.

PC7: Seventh principal component has maximum loading on ‘Personal’ and ‘Grad.Rate’ with a positive correlation. This component is a combination of graduation rate as well as personal expenses for a student, with personal expenses having slightly more weightage.

PC8: Eighth principal component has maximum loading on ‘P.Undergrad’ and ‘perc.alumni’ with a positive correlation. This component is a combination of the number of part time undergraduate students as well as percentage of alumni who donates, with alumni donation having slightly more weightage.

# THE END