|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| Project: Advanced Statistics  Business Report |
| |  |  |  | | --- | --- | --- | |  |  |  | |

Table of Contents:

Problem 1: SALARY DATA ANALYSIS

**1)Executive Summary………………………………………………………………..……………4**

**2)Introduction………………………………………………………………………………….……5**

**3)Data Description………………………………………………………………………………....6**

**4)Description Analysis……………………………………………………………………..……7,8**

**Problem 1A:**

1. a) State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually……………………………………………………………………………....……….**8,9**
2. b) Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the Anova result…………………………………………………….………**9**
3. c) Perform a one-way ANOVA on Salary with respect to Occupation. State whether the null hypothesis is accepted or rejected based on the Anova result…………………………………………………………**.10**
4. d) If the null hypothesis is rejected in either (2) or in (3) find out which class means are significantly different. Interpret the result……………………………………………………………………….…….…**10,11**

**Problem 1B:**

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

[hint: use the ‘point plot’ function from the ‘seaborn’ function] ……………………….……………..**12**

1. Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education\*Occupation). State the null and alternative hypothesis and state your results. How will you interpret this result?........................................................................................................**13**
2. Explain the business implications of performing ANOVA for this particular case study………….….**14**

**FIGURES/TABLES:**

* Data Ranges/ Data Samples………………………………………………………………………5
* Data Types and entries…………………………………………………………………………….6
* Count of null values…………………………………………………………………………………6
* Salary column data description……………………………………………………………………6
* Graph for different data comparisons……………………………………………………………6-7
* One-way Anova…………………………………………………………………………………….8,9
* Tukey HSD for Variables………………………………………………………………………….9,10
* Interaction plot………………………………………………………………………………………10
* Point Plot……………………………………………………………………………………………...11
* One way-Two-way Anova result…………………………………………………………………...12

[Education - Post 12th Standard](https://olympus.greatlearning.in/courses/54964/files/3234944/download?verifier=PFivZbcUn6RVWwAQmYhwrgEQZvyBaaqyZxFBC1jn&wrap=1)

1)Executive Summary……………………………………………………………………………………………….……**16**

2)Introduction………………………………………………………………………………………………………………**16**

3)Data Description…………………………………………………………………………………………………………**17**

4)Data Visualization……………………………………………………………………………………………………**18-19**

2.1) Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA……………………………………………………………………………………………**19-29**

2.2) Is scaling necessary for PCA in this case? Give justification and perform scaling……………………………**30**

2.3) Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data] ……………………………………………………………………………………………………………………..**31-33**

2.4) Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]……………………………………………………………………………..**34**

2.5) Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both]……………………………….**35-36**

2.6) Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features……………………………………………………………………………………………………………**36**

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]…**37-39**

2.8) 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?..............................................................**39**

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] …………..**37**

**FIGURES/TABLES:**

* Data summary…………………………………………………………………………………………………2
* Data dictionary/Data headings…………………………………………………………………………2.2,2.3,2.4
* Data Information…………………………………………………………………………………………….2.5
* Null Values……………………………………………………………………………………………………2.6
* Data Visualization……………………………………………………………………………………….2.7,2.8,2.9
* Univariate………………………………………………………………………………………………….2.10-2.25
* Multivariate…………………………………………………………………………………………………2.26
* Heat Map…………………………………………………………………………………………………….2.27
* ZSCORE…………………………………………………………………………………………………….2.28
* Covariance……………………………………………………………………………………………………2.29
* Corelation…………………………………………………………………………………………………….2.30
* Scaling…………………………………………………………………………………………………………2.31
* Eigen Vectors……………………………………………………………………………………………….2.32
* Eigen Values………………………………………………………………………………………………….2.32
* Eigen Vectors…………………………………………………………………………………………………2.33
* PCA Components…………………………………………………………………………………………….2.34
* First Components……………………………………………………………………………………………..2.35
* Cumulative Values……………………………………………………………………………………………2.36
* New Data Frame……………………………………………………………………………………………….2.37

**Problem 1:**

SALARY DATA ANALYSIS



Picture URL- https://www.federalbank.co.in/fed-salary

Executive Summary

Salary is hypothesized to depend on educational qualification and occupation. Hence the data collected for salary and the educational qualification and occupation are noted. Educational qualification is based on 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.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. The data consists of 40 rows. The number of columns being 3 namely: Education, Occupation and Salary.



Table 1.1- Data Ranges

The data sample looks as below divided into 3 columns and 40 rows:

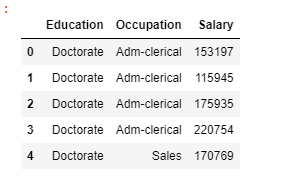


Table 1.2- data sample

Data Description

The exploratory data analysis of the dataset is:

a) Education as an object

b) Occupation as an object

c) Salary as an integer value

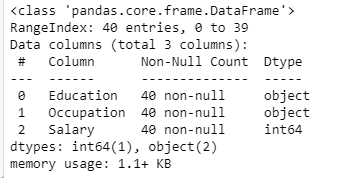


Table 1.3- Data types and number of entries

We will also check for some missing values in the data set:

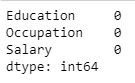


Table 1.4 Count of Null Values

To know what is the mean median and maximum value for the data set we are working on:

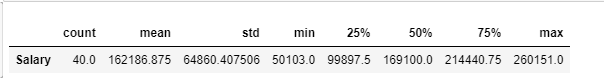


Table 1.5 Salary Column data described

Descriptive Analysis

To make basic checks as to which Occupation and Educational Qualification is getting the maximum or the minimum salary.

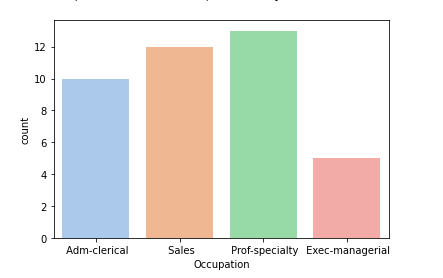


Table 1.6 Graphical representation of Occupation and Salary

Inference:

* The maximum salary is given to occupation of Prof-specialty
* Then Sales is the second followed by Adm-clerical.
* The least salary is given to occupation of Exec-managerial.

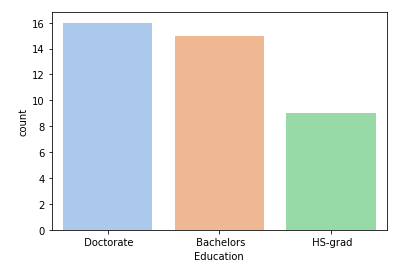


Table 1.7 Graphical representation of Education and Salary

Inference:

* The maximum salary is given to occupation of Doctorate.
* Then comes Bachelors Education.
* The least salary is given to occupation of HS-grad.

**Problem 1A:**

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

Before performing one-way Anova we checked data is normally distributed:

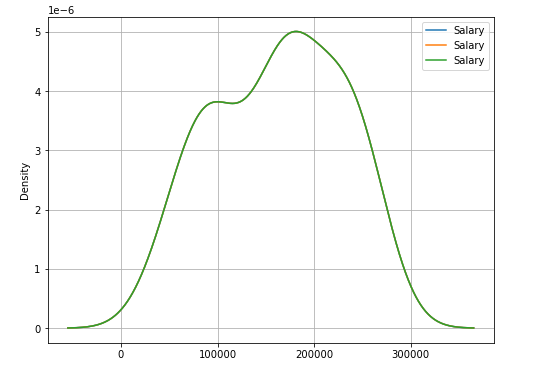


Table 1.8 Graphical representation of Education and Salary

**Performing ONE WAY ANOVA for:**

**We will consider hypothesis in Education as:**

* Null Hypothesis 𝐻0: The mean salary is the same across all the 3 categories of education namely: Doctorate, Bachelors, HS-Grad).
* Alternate Hypothesis 𝐻1: The mean salary is different in at least one category of education.
* Significance level= 0.05

**We will consider hypothesis in Occupation as:**

* Null Hypothesis 𝐻0: The mean salary is the same across all the 4 categories of occupation namely: Adm-clerical, Exec-managerial, Prof-specialty, Sales).
* Alternate Hypothesis 𝐻1: The mean salary is different in at least one category of occupation.
* Significance level= 0.05

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

To perform one-way ANOVA for education w.r.t the variable ‘Salary’, we apply the ANOVA formula in the Jupyter notebook and run the AOV table:

# ***Null Hypothesis 𝐻0: The mean salary is the same across all the 3 categories of education namely: Doctorate, Bachelors, HS-Grad).***

***Alternate Hypothesis 𝐻1: The mean salary is different in at least one category of education***

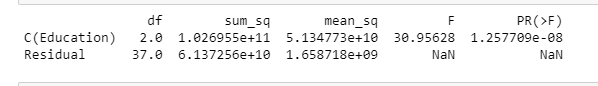


Table 1.9.1 One-way Anova for Education

Since the p value = 1.257709e-08 is less than the significance level (= 0.05), we reject the null hypothesis of equality in the mean. We are 95% confident about our rejection. The variance between the treatments is 53 times within each other which is significantly a large difference.

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

Considering One-way Anova on Salary and Occupation:

***Null Hypothesis 𝐻0: The mean salary is the same across all the 4 categories of occupation namely: Adm-clerical, Exec-managerial, Prof-specialty, Sales).***

***Alternate Hypothesis 𝐻1: The mean salary is different in at least one category of occupation.***

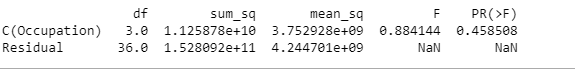


Table 1.9.2 One-way Anova for Occupation

Since the p value = 0.458508 is more than the significance level (= 0.05), we fail to reject the null hypothesis. That is, we accept H0. Concluding that there is no significant difference in the mean salaries across the 4 categories of occupation. We are 95% confident about failing to reject. The variance between the treatments is 0.884144 times within each other which is a small difference.

1.4 If the null hypothesis is rejected in either (2) or in (3) find out which class means are significantly different. Interpret the result.

To prove the difference in class mean we will **Perform Tukey HSD Test**. Since we rejected the null hypothesis is case 2 -Education. Here is the output from python:

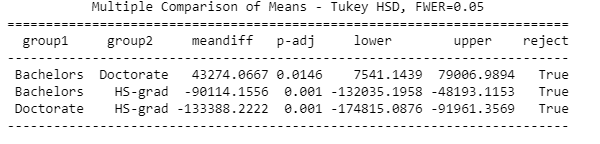


Table 1.10 Tukey HSD for variable ‘Education’

Inference:

* The table 1.10 shows p-adj value is less than 0.05(our level of significance) which indicates that mean salaries across all categories of education are different.

To prove the difference in class mean we will **Perform Tukey HSD Test**. Since we failed to reject the null hypothesis is case 2 -Occupation. Here is the output from python:

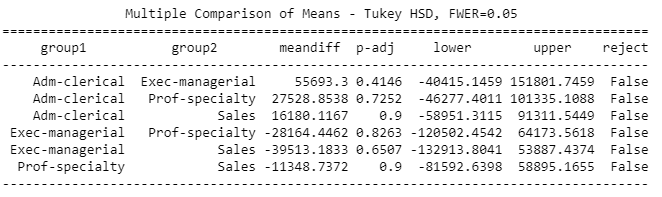


Table 1.11 Tukey HSD for variable ‘Occupation’

Inference:

* The table 1.11 shows p-adj value is more than 0.05(our level of significance) which indicates that the mean salaries across all occupation classes are significantly same.

To analyse the cause and effect of variables on each other we have plotted below the interaction plot for the categorical variable: Education and Occupation.

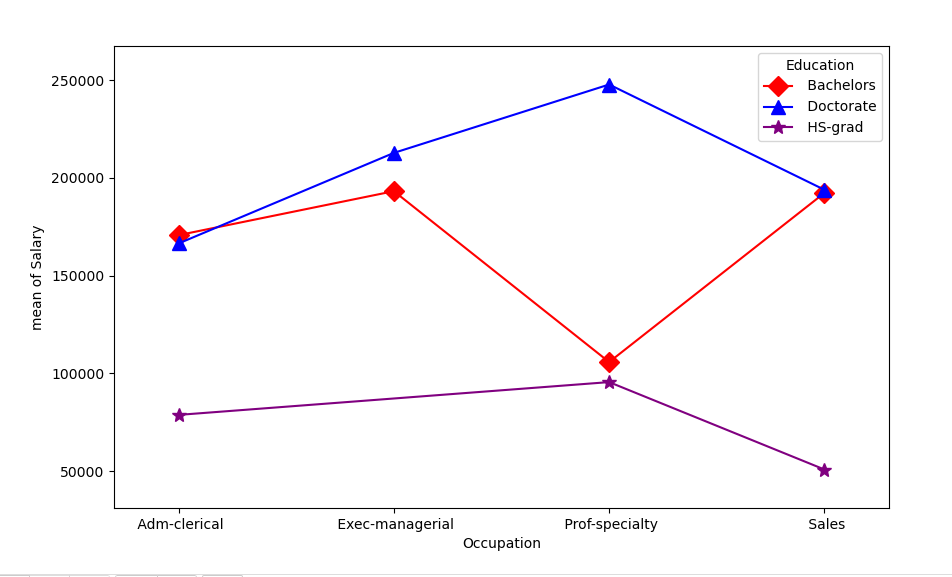


Table 1.12 Interaction Plot of Education and Occupation with respect to Salary

Inferences of Table 1.12:

* People who are HS-Grad do not hold Exec-managerial position but only have positions in: Adm-clerical, Prof-Specialty and Sales.
* People with Bachelors hold position in all three occupations namely: Adm-clerical, Exec-managerial, Prof-specialty and Sales.
* People who are Doctorate hold position in all three occupations namely: Adm-clerical, Exec-managerial, Prof-specialty and Sales. Highest paid Doctorate in the occupation Prof-specialty.
* People who are Bachelors and Doctorate hold the same salary in Adm-Clerical and in Sales.

**Problem 1B:**

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

We take up the point plot in jupyter notebook and get the below output:

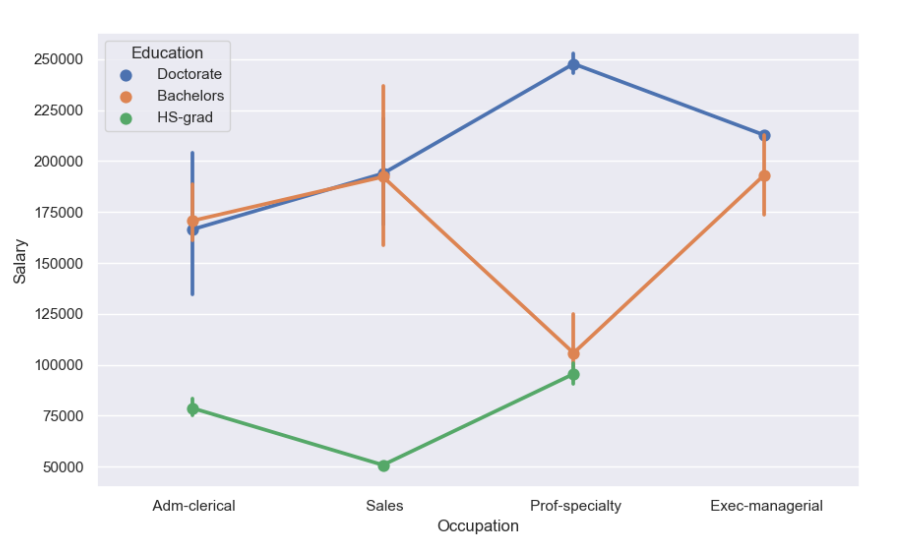


Table 1.13 Point Plot of Education and Occupation with respect to Salary

Inferences of Table 1.13:

* The HS-grad occupation is hired only in Adm-Clerical, Sales and Prof-specialty.
* The Bachelors degree is in the occupation of Adm-Clerical, Sales, Exec-managerial and Prof-specialty.
* The Doctorate is in the occupation of Adm-Clerical, Sales, Exec-managerial and Prof-specialty.
* The salary for Adm-clerical in the educational qualification of Doctorate and Bachelors is almost the same.
* The salary for Sales in the educational qualification of Doctorate and Bachelors is the same which is in between 1,75,000-2,00,000.
* The highest paid amongst all is occupation for Prof-specialty with the educational qualification of Doctorate.
* The lowest paid occupation is Sales with the qualification of HS-grad which is 50,000.
* The lowest paid in all educational qualification is Sales Occupation.

1.6 Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education\*Occupation). State the null and alternative hypothesis and state your results. How will you interpret this result?

By performing two-way ANOVA, we get the following table:

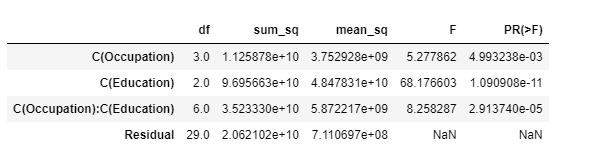


Table 1.14 Two-way Anova Test

* 𝐻0: The effect of the independent variable ‘education’ on the mean ‘salary’ does not depend on the effect of the other independent variable ‘occupation’ (i.e., there is no interaction effect between the 2 independent variables, education and occupation).
* 𝐻1: There is an interaction effect between the independent variable ‘education’ and the independent variable ‘occupation’ on the mean salary.
* Considering significance level 0.05.

Inference from table 1.14:

* As p value = 2.913740e-05 is lesser than the significance level ( = 0.05), we reject the null hypothesis.
* We see that there is an interaction effect between education and occupation on the mean salary.

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

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

**Problem 2:**

[Education - Post 12th Standard](https://olympus.greatlearning.in/courses/54964/files/3234944/download?verifier=PFivZbcUn6RVWwAQmYhwrgEQZvyBaaqyZxFBC1jn&wrap=1)



Picture URL- https://www.shiksha.com/courses-after-12th

We will be analysing by using principal component analysis on the data of the various Institutions in detail. We will also be using univariate and multivariate analysis. We will be dealing with a total of 777 rows and 18 columns.

Executive Summary



Table 2. Rows, Columns

Introduction

The data consists of various columns as under which will be analysed in detail:

|  |
| --- |
| 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 |

Table 2.2 Column Names as per Data Dictionary

The data looks like as below and has the following columns:

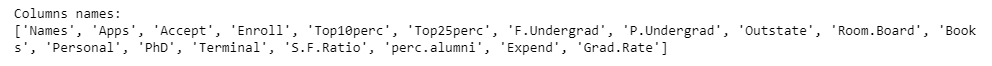


Table 2.3 Column Names as per Jupyter Notebook

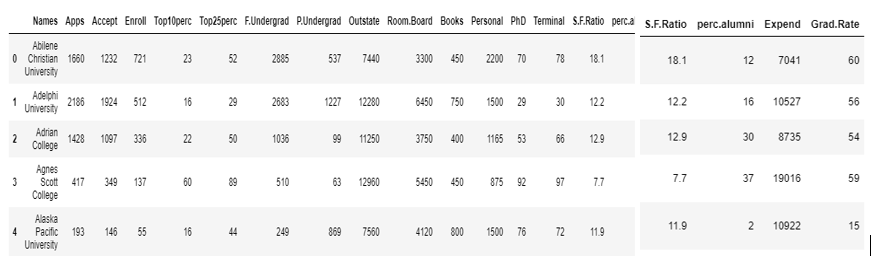


Table 2.4 Data headings

Data Description

The dataset has the below integer and decimal values. In the table 2.5 below:

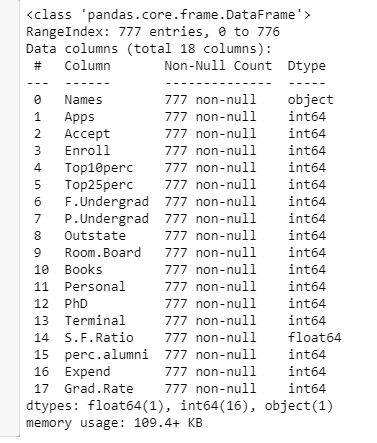
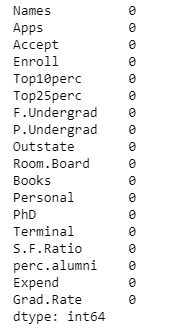
 

Table 2.6 Null Values

Table 2.5 Data Information

* Int64 means the data is in numerical format.
* Float64 means the data is in decimal number format
* Object is for alphabetical data.
* Total rows are 777 and the total columns is 18.
* The data has no null or empty cells for any headers. (Reverified in table 2.6)

To check if we have any outliers in the data:

Data Visualization

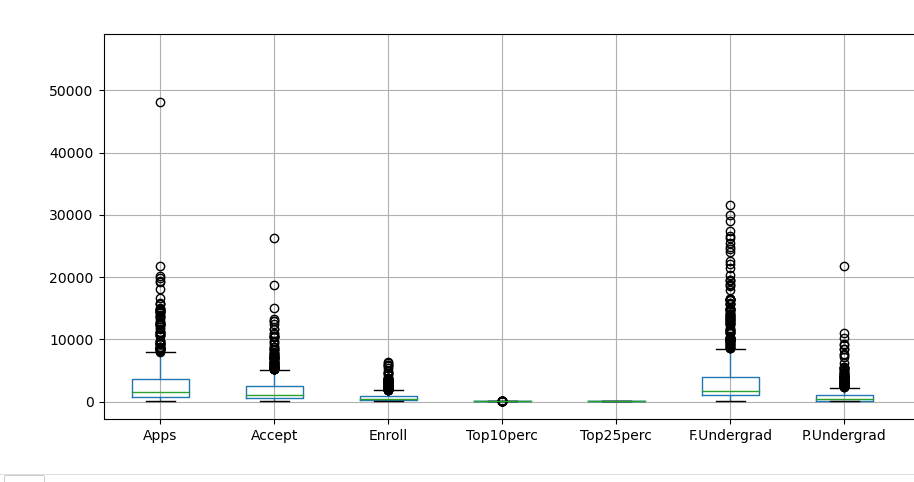


Table 2.7 Outliers(1)

* The data has outliers as seen in the figure 2.7 wherein highest is Number of full-time undergraduate students.
* The least being in the data is Percentage of new students from top 25% higher secondary class. Zero outliers as seen above.

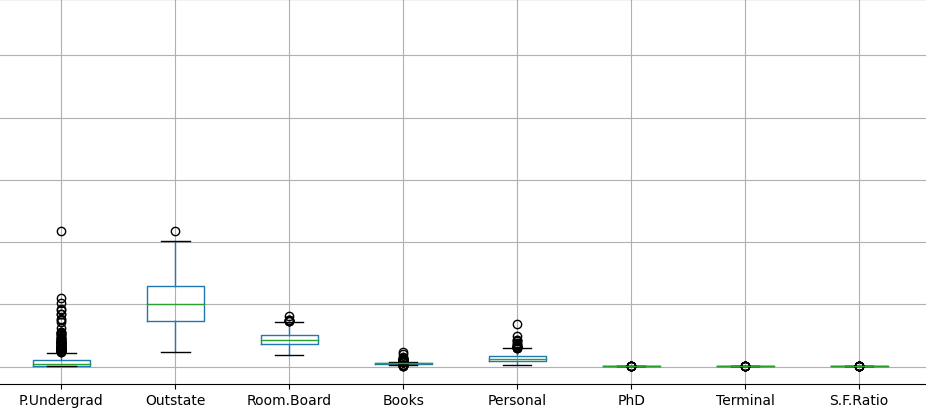


Table 2.8 Outliers(2)

* Here we see the number of outliers is maximum at Number of part time undergraduate student.
* PHD, Terminal and Student faculty ratio (S.F. Ratio) being at the same.
* Books also have a lot of outliers whereas outstate seems to have only one.
* Room board having lesser outliers than Number of part time undergraduate student, Books and personal.

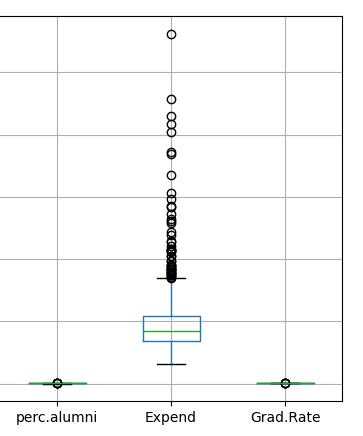


Table 2.9 Outliers(3)

* Expend has the maximum outliers and Perc alumni and Grad rate the least and the same.

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

**Univariate analysis** means an analysis is of the fact that only accounts for one variable's effect on a dependent variable. Primarily, Univariate Analysis simply takes data and provides a summary and associated pattern.  We have done Univariate analysis as:

**Apps: Number of applications received:**

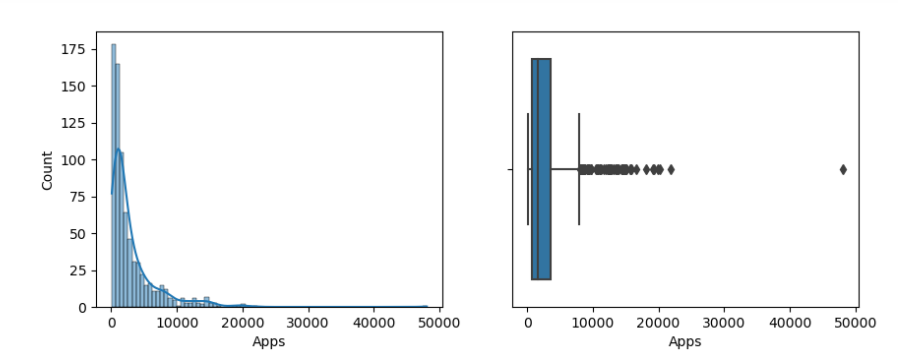


Table 2.10 Univariate for Applications

* The data has been represented with dist. plot and outliers. The box plot references to a lot of outliers in the data too.
* The count of application as seen in x axis is maximum at 50,000.
* The data is a skewed data and as seen in the y axis colleges or universities offer applications in the count of 3000-5000 as seen in the box plot.

**Accept: Number of applications accepted:**

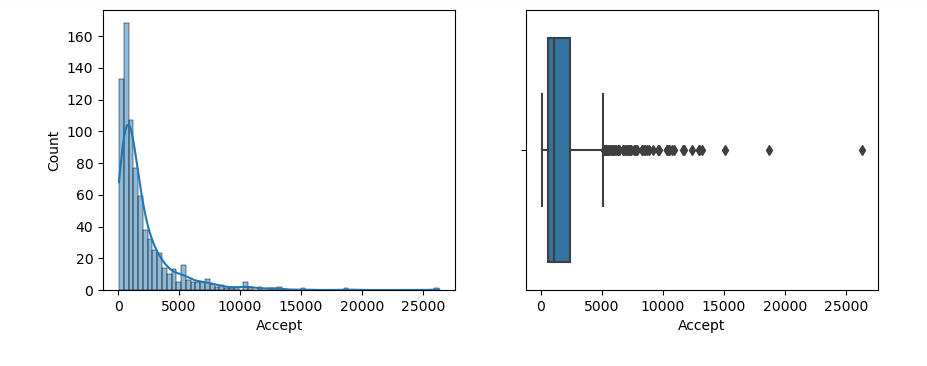


Table 2.11 Number of applications accepted

* The data has outliers as seen in the right side of the diagram 2.11. Where in the maximum outlier is 25000.
* The data has skewness as seen in the left image of dist. Plot. A positively skewed data.
* The number of applications accepted in each University are between 72-1500.

**Enrol: Number of new students enrolled:**

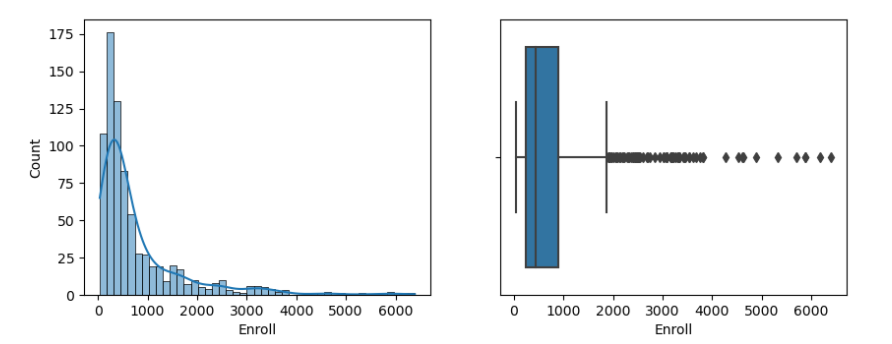


Table 2.12 Number of new students enrolled

* The Number of new students’ data has outliers as shown in the box plot, with a positively skewed data as shown in dist. Plot.
* The number of students enrolled minimum is 242 and maximum is 902.

**Top10perc: Percentage of new students from top 10% of Higher Secondary class:**

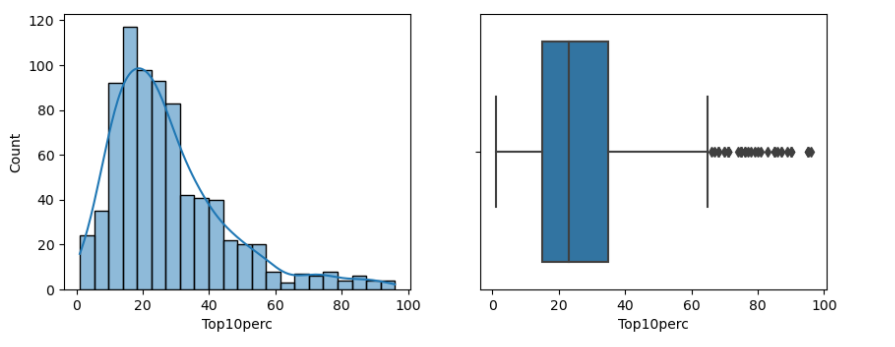


Table 2.13 Percentage of new students from top 10% of Higher Secondary class

* The data is posivitely skewed as per the distribution plot.
* There are many outliers in the data as seen in the box plot.
* The data has the minimum intake of students from 10% higher secondary class as 15 and maximum to 35.

**Top25perc: Percentage of new students from top 25% of Higher Secondary class:**

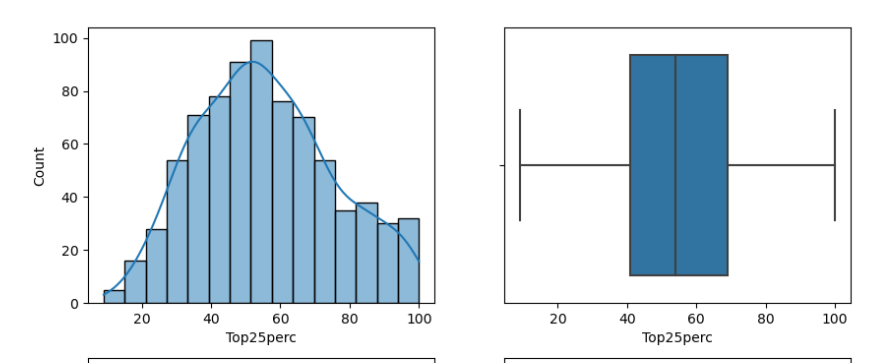


Table 2.14 Percentage of new students from top 25% of Higher Secondary class

* The box plot has zero outliers.
* The distribution plot is a normally distributed data.
* The Majority of new students are from 25% higher secondary class.

**F. Undergrad: Number of full-time undergraduate students:**

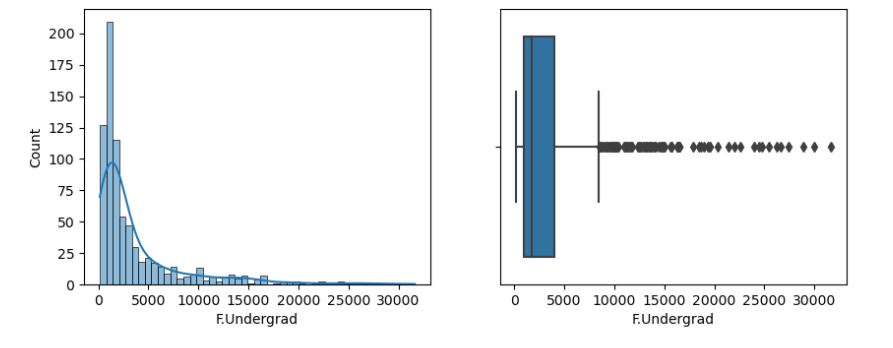


Table 2.15 Number of full-time undergraduate students

* The number of full-time undergraduate students have a lot of outliers which is beyond 30,000.
* The data is a positively skewed data as seen in the distribution plot.
* The range is 3000-5000 full time undergraduate students studying in all universities.

**P.Undergrad: Number of part-time undergraduate students:**

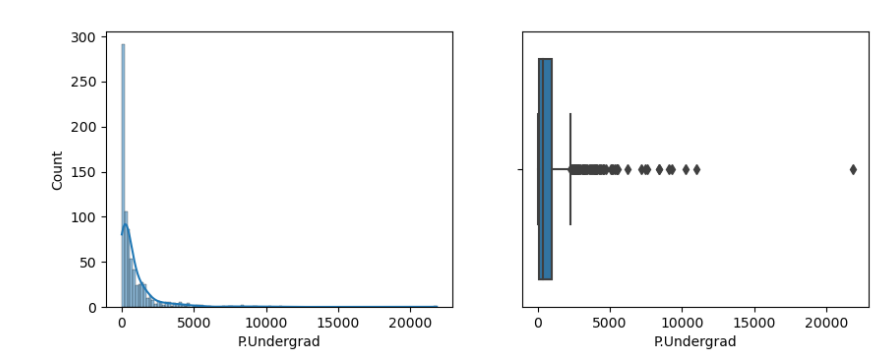


Table 2.16 Number of part-time undergraduate students

* The data has a lot of outliers as seen in the box plot.
* The distribution is a positively skewed data.
* 1000-3000 is the ranger for the students who are part time undergraduates studying in all universities.

**Outstate: Number of students for whom the particular college or university is Out-of-state tuition:**

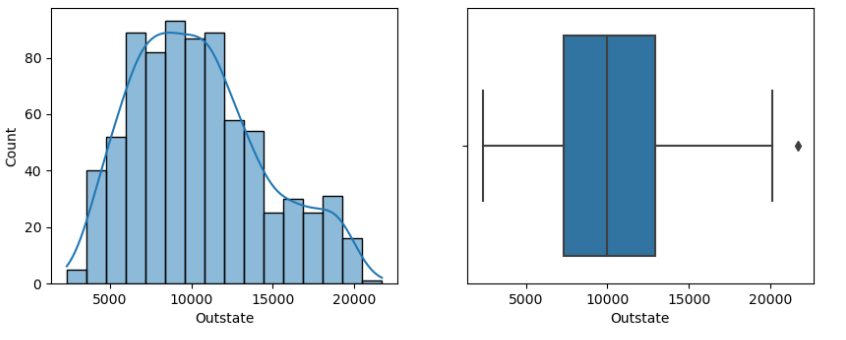


Table 2.16 Number of students for whom the particular college or university is Out-of-state tuition

* Has only one outlier as seen in the above figure of box plot.
* The pair plot is a normally distributed data.

**Room. Board: Cost of Room and board:**

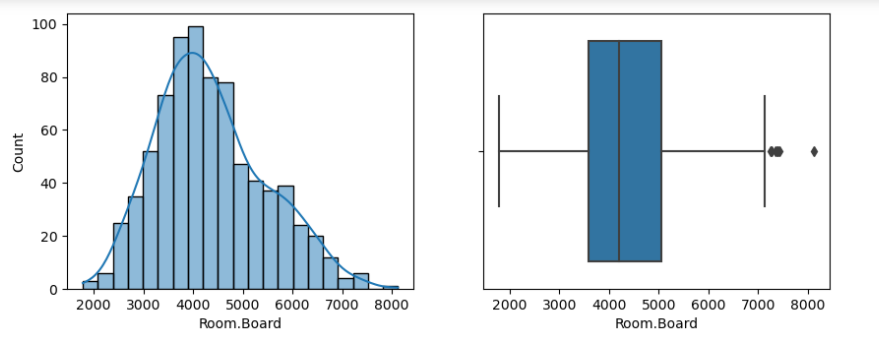


Table 2.17 Cost of Room and board

* Has only one outlier as seen in the above figure of box plot.
* The distribution plot is a normally distributed data.

**Books: Estimated book costs for a student:**

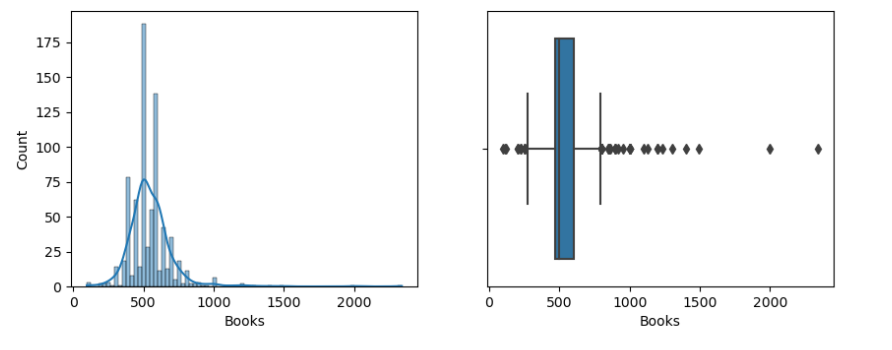


Table 2.18 Estimated book costs for a student

* Has outlier as seen in the above figure of box plot.
* The distribution plot is a normally distributed data.

**Personal: Estimated personal spending for a student:**

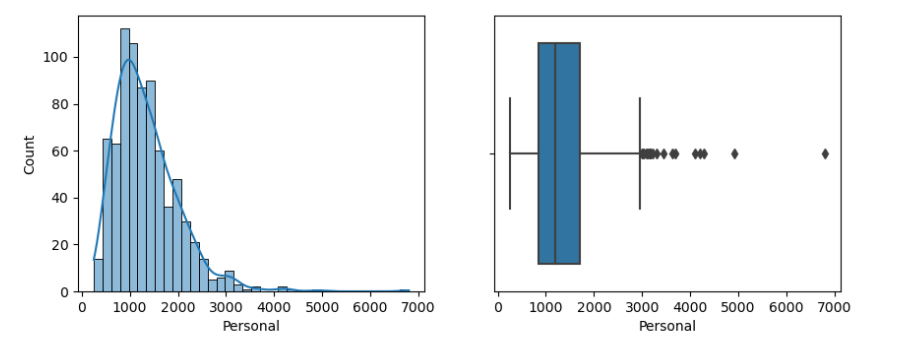


Table 2.19 Estimated personal spending for a student

* Has many outliers as seen in the boxplot.
* The distribution plot is a normally distributed data.
* The spending’s of some students is way higher than the others.

**PhD: Percentage of faculties with Ph.D.’s:**

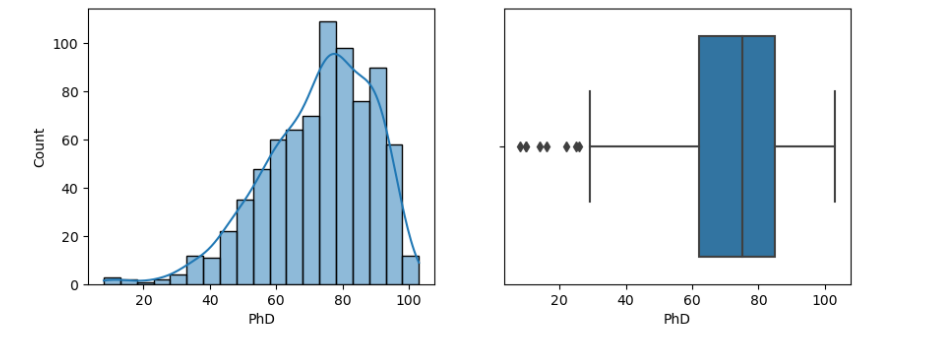


Table 2.20 Percentage of faculties with Ph.D.’s

* The data is negatively skewed as seen in the data of distribution plot.
* The data has outliers as seen in the box plot.

**Terminal: Percentage of faculties with terminal degree:**

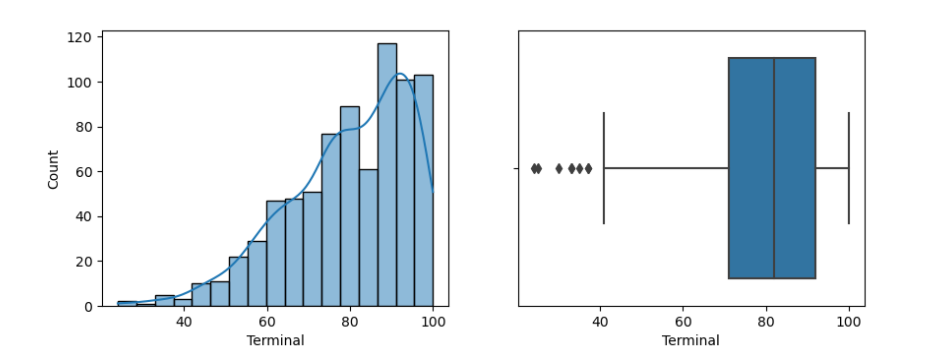


Table 2.21  Percentage of faculties with terminal degree

* The data is negatively skewed as seen in the data of distribution plot.
* The data has outliers as seen in the box plot.

**S.F.Ratio: Student/faculty ratio:**

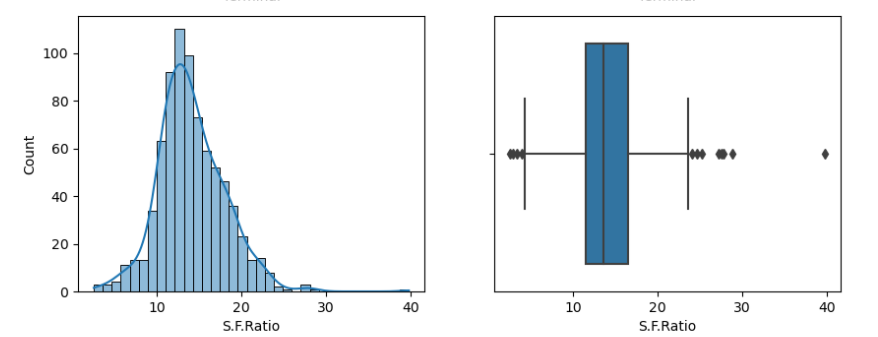


Table 2.22 Student/faculty ratio

* Has outlier as seen in the above figure of box plot.
* The distribution plot is a normally distributed data.
* The student faculty ratio is the same in all the university and colleges.

**Perc.alumni: Percentage of alumni who donate:**

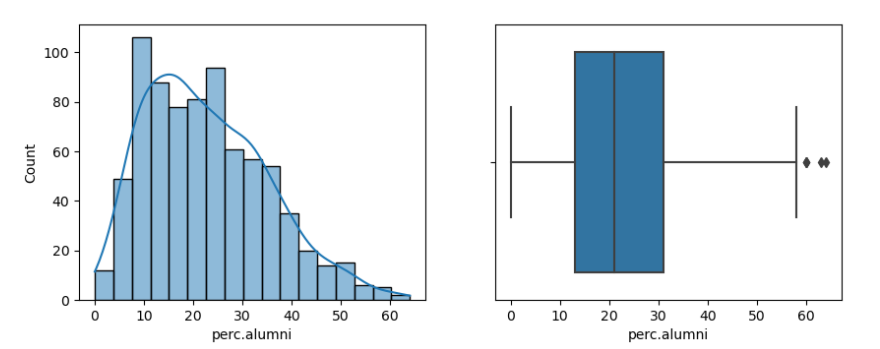


Table 2.23 Percentage of alumni who donate

* Has many outliers as seen in the boxplot.
* The distribution plot is a normally distributed data.

**Expend: The Instructional expenditure per student:**

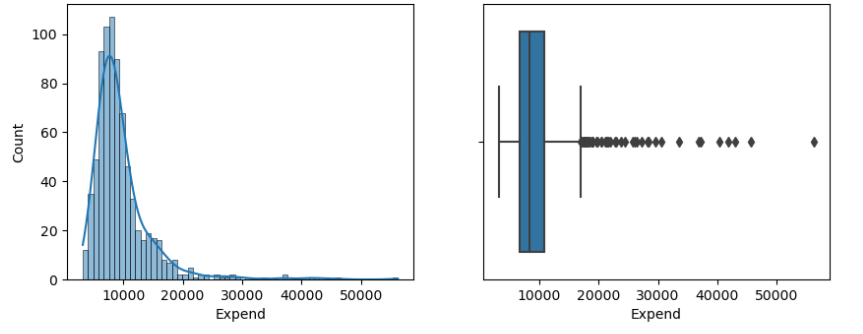


Table 2.24 The Instructional expenditure per student

* The data is positively skewed as seen in the data of distribution plot.
* The data has outliers as seen in the box plot.

**Grad. Rate: Graduation rate:**

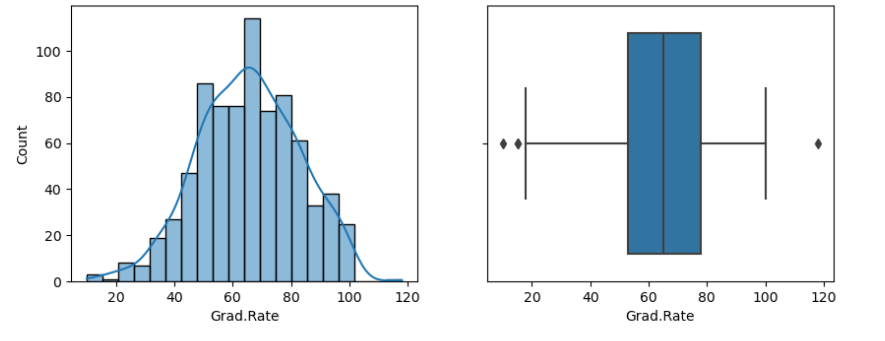


Table 2.25 Graduation rate

* The data is normally distributed data as per the distribution plot.
* The data has outliers as seen in the box plot.
* The graduation rate among the students in all University is above 60%.

**MULTIVARIATE ANALYSIS** is simultaneously studying various variables at the same time. This is more informative than univariate analysis.

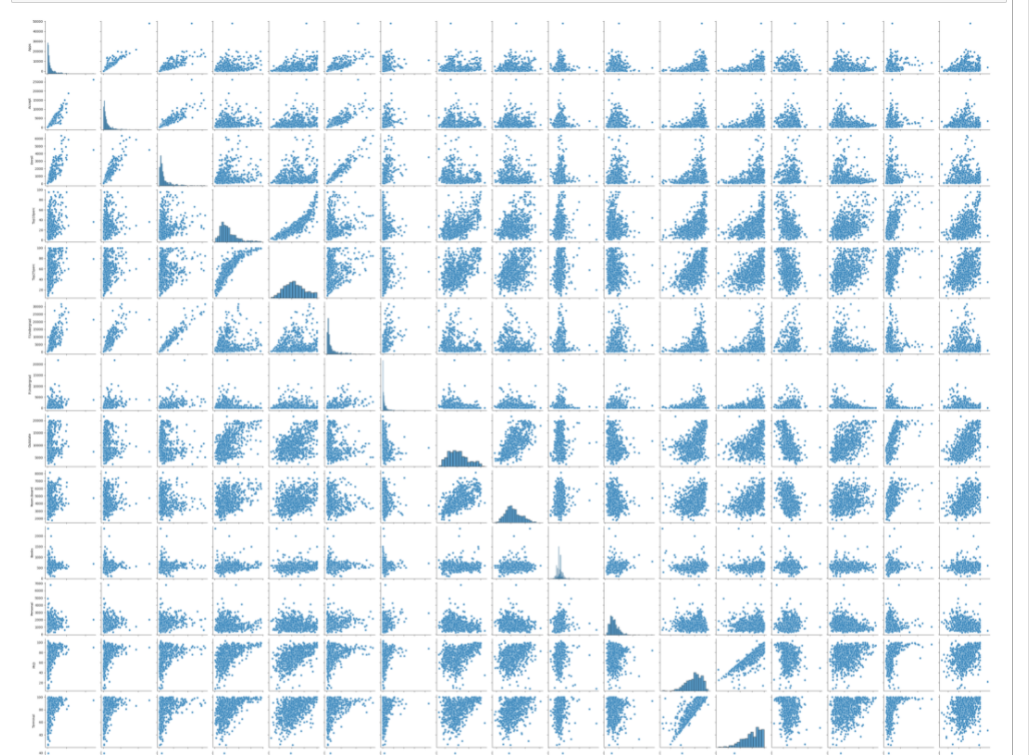
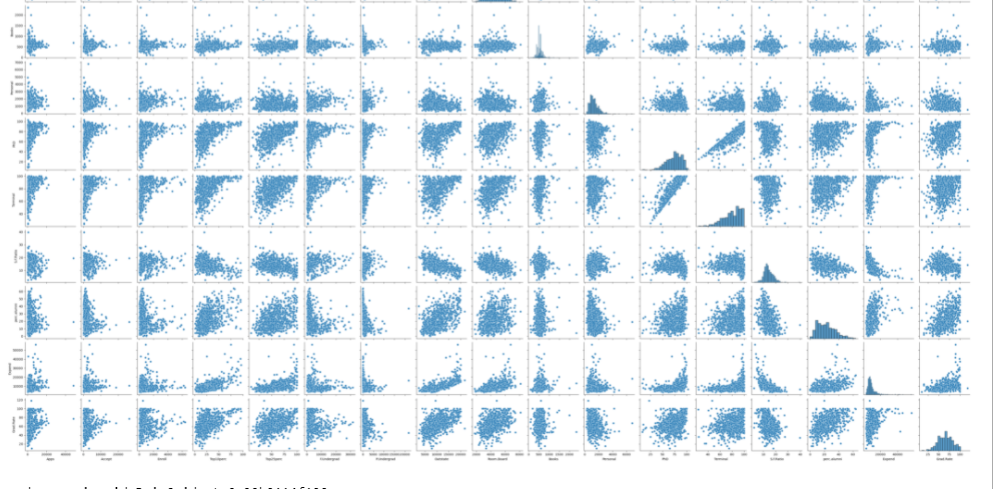
 

Table 2.26 Pair Plot

Inference:

* The multivariate analysis helps us above to know the different trends of the data as per the each other.
* The pair plot helps us understand the relationship between all numerical values in the data set.

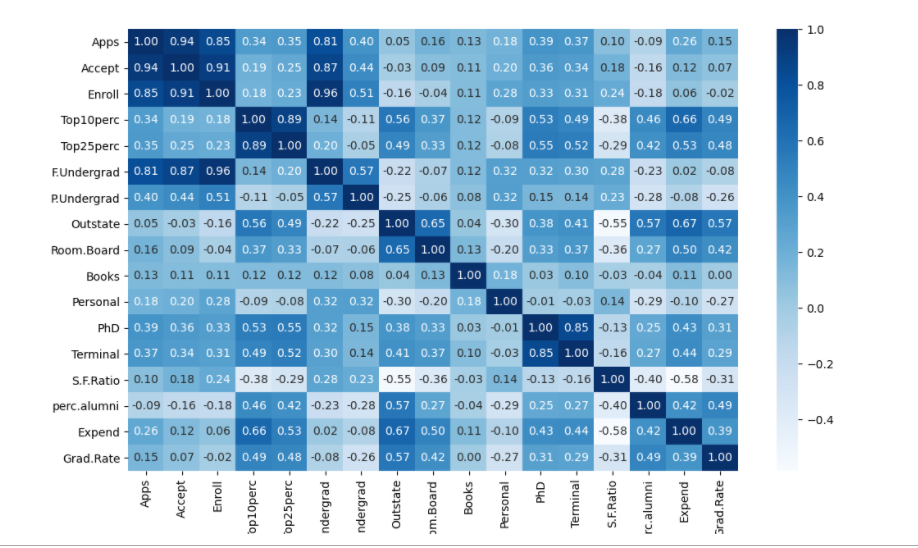


Table 2.27 Heat Map

Inference:

* Application variables are highly positively related to applications accepted, students enrolled and full- time graduates.
* When applications from the student is received the students are enrolled as full-time graduates.
* We can find negative corelation between application and percentage of alumni which indicates not all students are part of the alumni of their college or university.
* The below are highly co-related amongst each other:

|  |
| --- |
| * Top10perc: Percentage of new students from top 10% of Higher Secondary class |
| * Top25perc: Percentage of new students from top 25% of Higher Secondary class * Outstate: Number of students for whom the particular college or university is Out-of-state tuition * Room. Board: Cost of Room and board * Books: Estimated book costs for a student * Personal: Estimated personal spending for a student * PhD: Percentage of faculties with Ph.D.’s * Terminal: Percentage of faculties with terminal degree * S.F.Ratio: Student/faculty ratio * perc.alumni: Percentage of alumni who donate * Expend: The Instructional expenditure per student * Grad.Rate: Graduation rate |

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

Yes, it is necessary to normalize data before performing PCA. The PCA calculates a new projection of the data set. If we normalize our data, all variables have the same standard deviation, thus all variables have the same weight and our PCA calculates relevant axis.

Why scaling is important is justified below:

* The data has all in different units of measurement like application, accepted application, enrolled fulltime graduates, part-time graduates, outstate are all in numbers of students.
* The top 10% and 20% is all in percentage data.
* Room board, books, and personal are values associated with money.
* The PHD, sf ratio, percentage of alumni are percentage values of all combinations of teachers and students.
* The graduation rate is also a percentage.

**The data post dropping categorical column has now 777 rows and 17 columns.**

* We will go to perform zscore test (as below table 2.28) will help us scale and know how many standard deviations is the point away from the mean and also the direction.
* In scaling the below data numbers did not change only the difference of units were bought to same scale.

**We have scaled the data as below:**



Table 2.28 ZScore Test

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

**The terms covariance and correlation are very similar to each other in probability theory and statistics. Both the terms describe the extent to which a random variable or a set of random variables can deviate from the expected value.**

**Covariance** signifies the direction of the linear relationship between the two variables. **Variance** is the expectation of the squared deviation of a random variable from its mean. Informally, it measures how far a set of numbers are spread out from their average value.

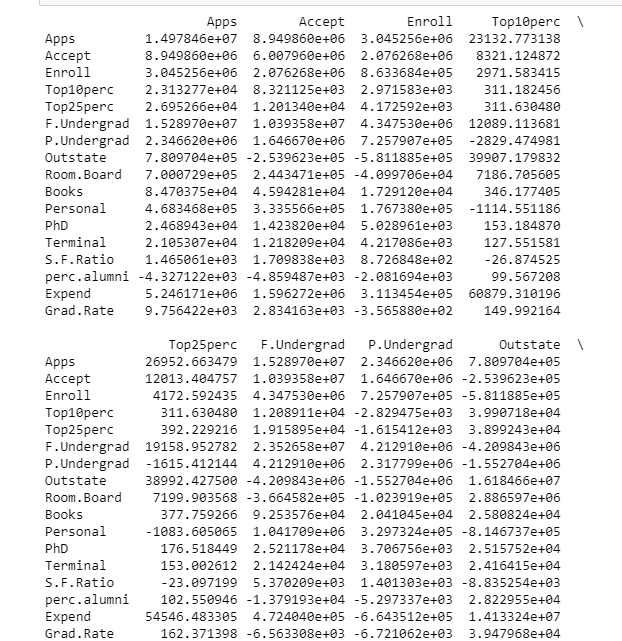


Table 2.29.1 covariance matrix

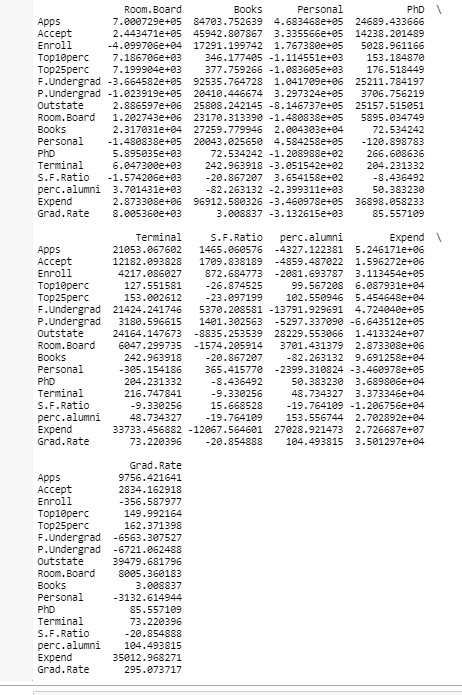
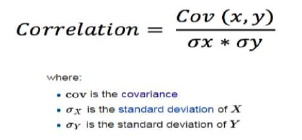


Table 2.29.2 covariance matrix

INFERENCE: The above figures 2.29.1 and 2.29.2 shows linear relationship between the variables. As the data has outliers and is only performed with scaling the data the covariances are high.

**Correlation** analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables.



**Correlation** post scaling is also the same as before scaling:

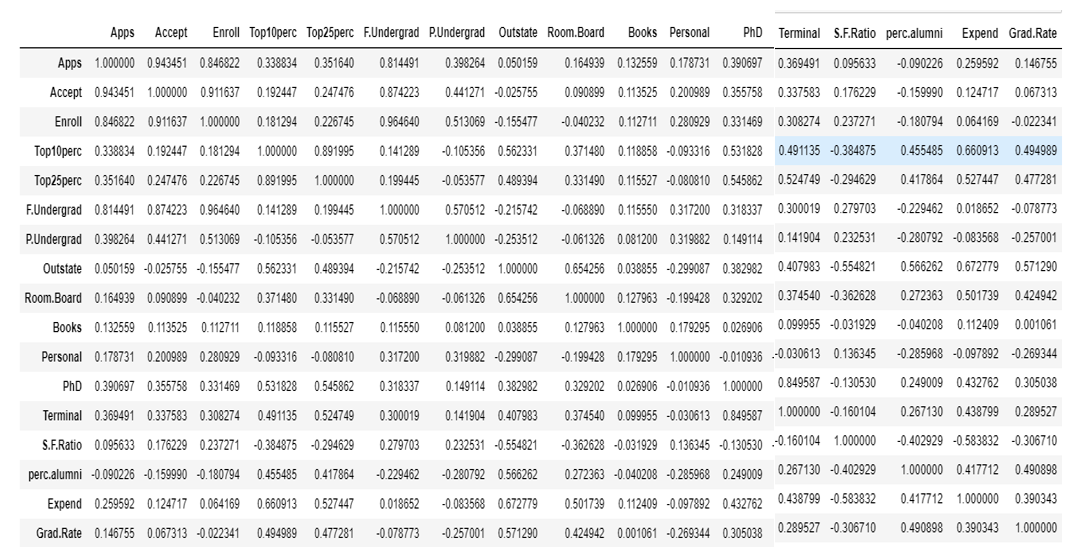


Table 2.30 Correlation Matrix

* We can see how some variables above are highly positively correlated and some negatively corelated. The variables moderately corelated are also seen.
* The application, acceptance, enrolment and full-time graduates are highly positively co-related with each other. The top 10 percentage and top 25 percentage are highly positively co-related.

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

Outliers post scaling:

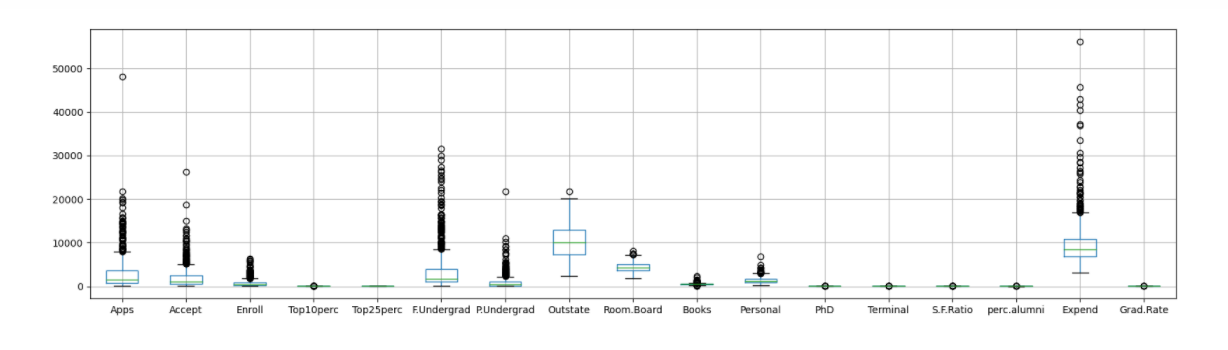


Table 2.31.1 Outlier post scaling

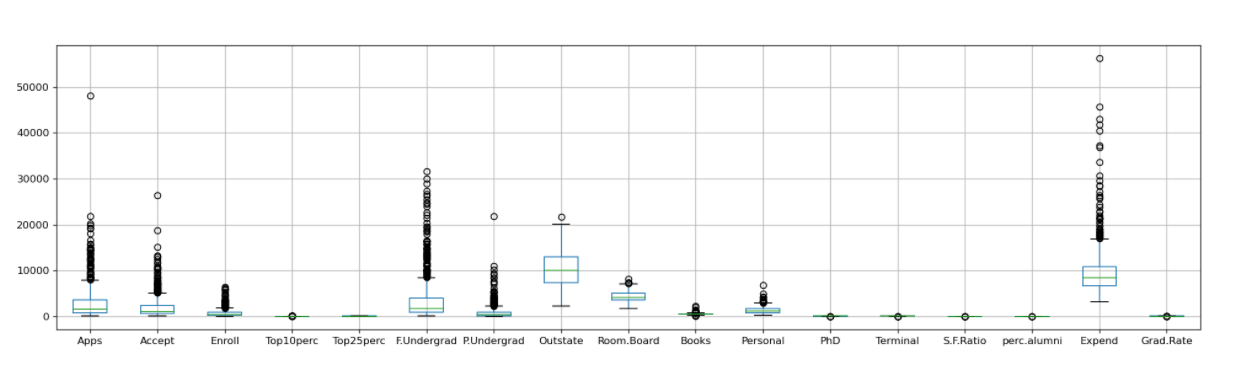
Outliers before scaling:

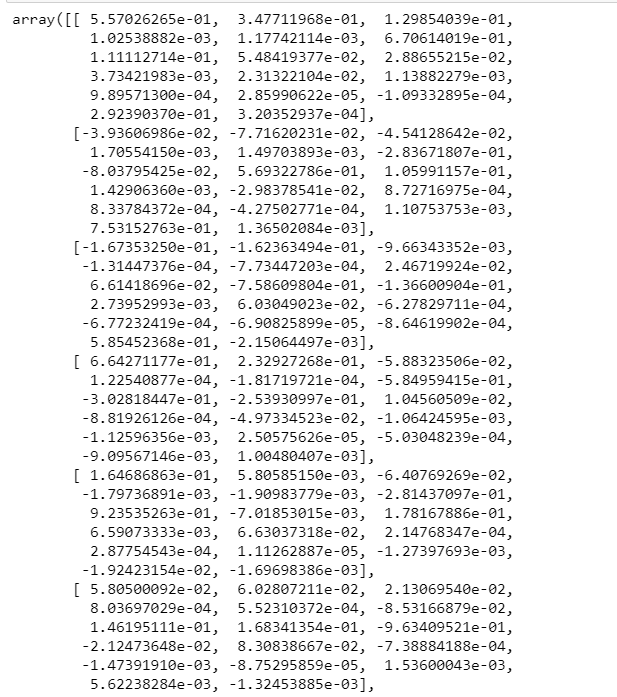
Table 2.31.2 Outliers pre-scaling

Inference:

* No change in outliers in above figures of Pre scaling and Post scaling.
* Scaling scaled the values in ZScore distribution.
* Both figures have the same inferences.
* The maximum outliers are for Expend and the minimum is for Top 25%.
* The PHD, top 10 %, Terminal, SF ration, Perc. Alumni and Grad rate all are at the same level.

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

Eigenvectors:



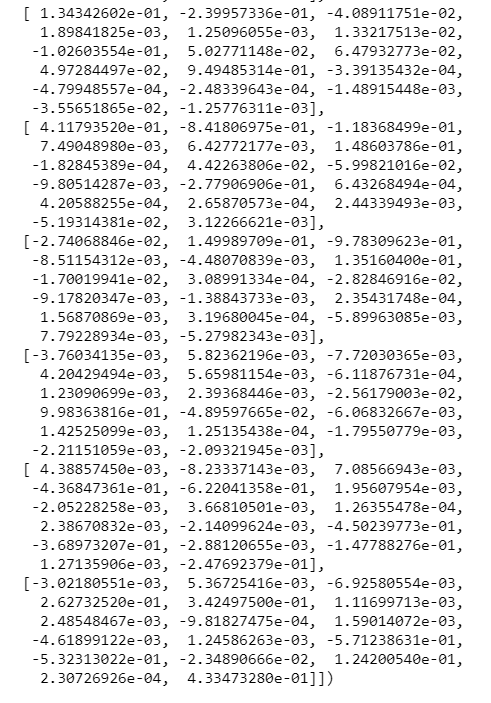


Table 2.32 Eigen Vectors

* The eigenvectors are also termed as characteristic roots.
* It is a non-zero vector that can be changed at most by its [scalar factor](https://byjus.com/maths/scale-factor/) after the application of linear transformations.
* Here we understand which variable contributed more to the PCA as per the weightage and numbers above.

Eigen Value:

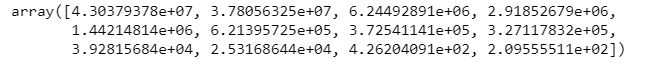


Table 2.32 Eigen Values

* The corresponding eigenvalue is the factor by which the eigenvector is scaled.
* We see all positive eigen values.
* Shows how much each variable is contributing to PCA.

2.**6** Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features

The below table shows how the data has been aligned newly to a dataset:

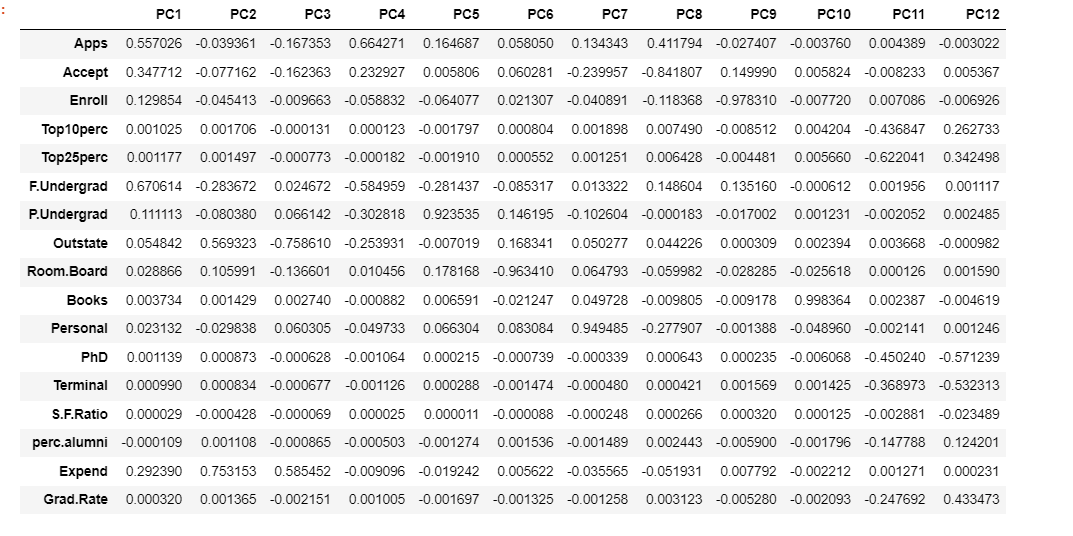
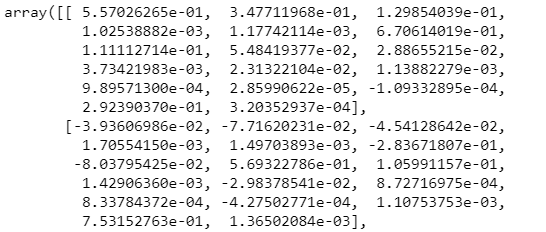


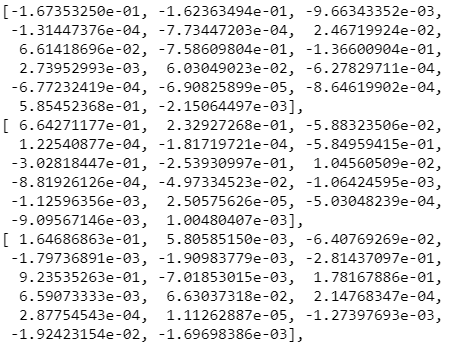
Table 2.33 Eigen Vectors into a data frame

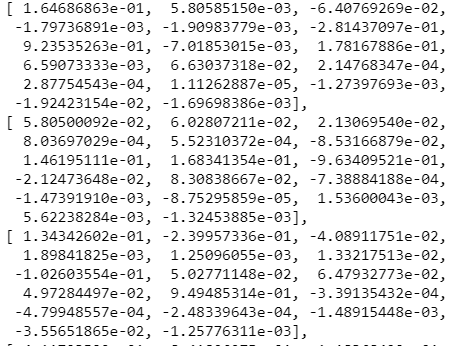
* The new shape of the data is 17 rows and 12 columns:
* The data has the following columns as PC1-PC12
* The rows are the data dictionary given to us.

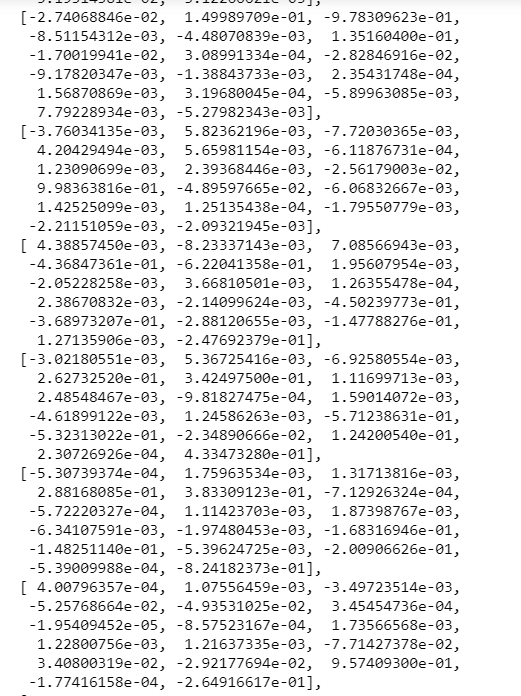
2.7Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]

Listing the PCA components first:









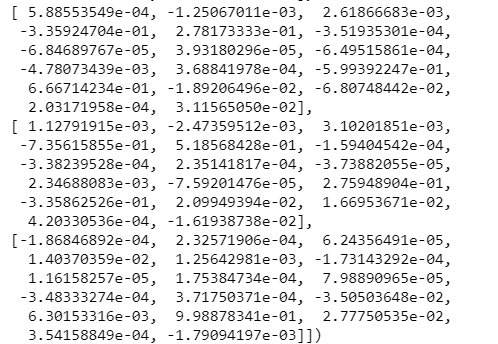


Table 2.34 PCA components

### The liner equation of the first component:

### 

Table 2.35 First components

2.8 Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?

The cumulative values are:

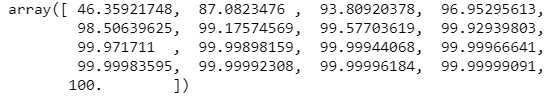


Table 2.36 Cumulative Values

* Once we add the eigen values we should get a summation of 100.

To check the optimum number of PCA:

* Check for cumulative variance up to 90%, check any corresponding variables up to 90%
* The increase amongst values should not be less than 5%.

5 is the optimum number of variable components. The incremental value in the 6th is less than 5%. Hence, we take up 5 components.

Post PCA on the selected data we get the below new data frame:



Table 2.37 New Data Frame

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

This business case study is about education dataset which contain the names of various colleges, which has various details of colleges and university. To understand more about the dataset, we perform univariate analysis and multivariate analysis which gives us the understanding about the variables. From analysis we can understand the distribution of the dataset, skew, and patterns in the dataset. From multivariate analysis we can understand the correlation of variables. Inference of multivariate analysis shows we can understand multiple variables highly correlated with each other. The scaling helps the dataset to standardize the variable in one scale. Outliers are imputed using IQR values once the values are imputed, we can perform PCA. The principal component analysis is used reduce the multicollinearity between the variables. Depending on the variance of the dataset we can reduce the PCA components. The PCA components for this business case are 5 where we could understand the maximum variance of the dataset. Using the components, we can now understand the reduced multicollinearity in the dataset. PCA reduces the multicollinearity of data as below:

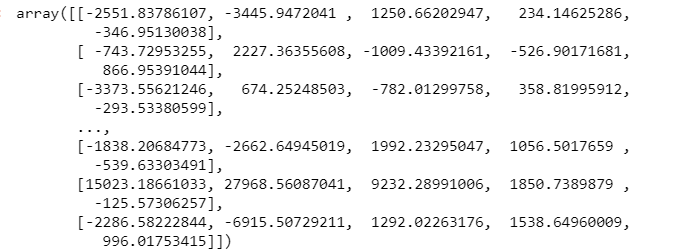


Table 2.39 PCA components