KIRAN.N

Great Learning

Advanced STatistics Project Report

Table of Contents

[**Problem 1** 5](#_Toc97411006)

[Executive Summary 5](#_Toc97411007)

[Data Dictionary 5](#_Toc97411008)

[Exploratory Data Analysis 5](#_Toc97411009)

[Q 1.1 State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually. 6](#_Toc97411010)

[Q 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](#_Toc97411011)

[Q 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](#_Toc97411012)

[Q 1.4 If the null hypothesis is rejected in either (1.2) or in (1.3), find out which class means are significantly different. Interpret the result. 7](#_Toc97411013)

[Q 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. 9](#_Toc97411014)

[Q 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? 10](#_Toc97411015)

[Q 1.7 Explain the business implications of performing ANOVA for this particular case study. 10](#_Toc97411016)

[**Problem 2** 11](#_Toc97411017)

[Executive Summary 11](#_Toc97411018)

[Data Dictionary 11](#_Toc97411019)

[Exploratory Data Analysis 12](#_Toc97411020)

[Q 2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA? 13](#_Toc97411021)

[Q 2.2 Is scaling necessary for PCA in this case? Give justification and perform scaling. 21](#_Toc97411022)

[Q 2.3 Comment on the comparison between the covariance and the correlation matrices from this data[on scaled data] 22](#_Toc97411023)

[Q 2.4 Check the dataset for outliers before and after scaling. What insight do you derive here? 23](#_Toc97411024)

[Q 2.5 Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both] 24](#_Toc97411025)

[Q 2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features 25](#_Toc97411026)

[Q 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] 25](#_Toc97411027)

[Q 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? 27](#_Toc97411028)

[Q 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] 27](#_Toc97411029)

List Of Figures

[Figure 1 : Educational Categories verses Mean Salary Bar Plot 8](#_Toc97411030)

[Figure 2 : Occupational Categories verses Mean Salary Bar Plot 8](#_Toc97411031)

[Figure 3 : Education and Occupation Interaction Plot 9](#_Toc97411032)

[Figure 4 : Histogram and Box Plot of Apps 14](#_Toc97411033)

[Figure 5 : Histogram and Box Plot of Accept 14](#_Toc97411034)

[Figure 6 : Histogram and Box Plot of Enroll 14](#_Toc97411035)

[Figure 7 : Histogram and Box Plot of Top10perc 15](#_Toc97411036)

[Figure 8 : Histogram and Box Plot of Top25perc 15](#_Toc97411037)

[Figure 9 : Histogram and Box Plot of F.Undergrad 15](#_Toc97411038)

[Figure 10 : Histogram and Box Plot of P.Undergrad 16](#_Toc97411039)

[Figure 11 : Histogram and Box Plot of Outstate 16](#_Toc97411040)

[Figure 12 : Histogram and Box Plot of Room.Board 16](#_Toc97411041)

[Figure 13 : Histogram and Box Plot of Books 17](#_Toc97411042)

[Figure 14 : Histogram and Box Plot of Personal 17](#_Toc97411043)

[Figure 15: Histogram and Box Plot of PhD 17](#_Toc97411044)

[Figure 16 : Histogram and Box Plot of Terminal 18](#_Toc97411045)

[Figure 17 : Histogram and Box Plot of S.F.Ratio 18](#_Toc97411046)

[Figure 18 : Histogram and Box Plot of perc.alumni 18](#_Toc97411047)

[Figure 19 : Histogram and Box Plot of Expend 19](#_Toc97411048)

[Figure 20 : Histogram and Box Plot of Grad.Rate 19](#_Toc97411049)

[Figure 21 : Pair Plot 20](#_Toc97411050)

[Figure 22 : Correlation Plot 20](#_Toc97411051)

[Figure 23 : Box Plots after Treating Outliers 21](#_Toc97411052)

[Figure 24 : Box Plots Before Scaling 23](#_Toc97411053)

[Figure 25 : Box Plots After Scaling 23](#_Toc97411054)

[Figure 26 :Cumulative Explained Ratio verses PC Index 27](#_Toc97411055)

[Figure 27 :Absolute Values of Each Original Feature in Different Principal Components 28](#_Toc97411056)

[Figure 28 : Correlation Plot of Principal Components 29](#_Toc97411057)

List of Tables

[Table 1 : Sample Dataset 5](#_Toc97411058)

[Table 2 : One-way ANOVA Table for Education 6](#_Toc97411059)

[Table 3 : One-way ANOVA Table for Occupation 7](#_Toc97411060)

[Table 4 : Educational Categories verses Mean Salary 7](#_Toc97411061)

[Table 5: Occupational Categories verses Mean Salary 8](#_Toc97411062)

[Table 6 : Two-Way ANOVA Table 9](#_Toc97411063)

[Table 7 : Two-Way ANOVA Table with Interaction Variable 10](#_Toc97411064)

[Table 8 : Sample Dataset 11](#_Toc97411065)

[Table 9 : Descriptive summary of the data 13](#_Toc97411066)

[Table 10 : Sample Scaled Data 22](#_Toc97411067)

[Table 11 : Sample Correlation Matrix After Scaling 22](#_Toc97411068)

[Table 12 : Sample Covariance Matrix After Scaling 22](#_Toc97411069)

[Table 13 : Dataframe of PC and Original Features 25](#_Toc97411070)

[Table 14 : Sample Dataset with Principal Components. 29](#_Toc97411071)

# **Problem 1**

## Executive Summary

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [[SalaryData.csv](https://olympus.mygreatlearning.com/courses/64476/files/4277857/download?verifier=S20eCCO7edkahXbj5Z9sOgJ2pJzSpveqfyO12mNd&wrap=1)] 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.

### Data Dictionary

1. Education: Education category High School Graduate, Bachelor or Doctorate.
2. Occupation: Occupation category Administrative and clerical, Sales, Professional or specialty, and Executive or managerial.
3. Salary: Salary details.

Sample of the dataset

Table

Description automatically generated

Table : Sample Dataset

Dataset has 3 columns with 40 rows. Each row in the dataset corresponds to one individual with their education, occupation and their salary details.

### Exploratory Data Analysis

Let us check the types of variables in the data frame.

Education object

Occupation object

Salary int64

There are total 40 rows and 3 columns in the dataset. Out of 3, 2 columns are of object type and rest 1 is of integer data type.

#### Check for missing values in the dataset

RangeIndex: 40 entries, 0 to 39

Data columns (total 3 columns):

# Column Non-Null Count Dtype

0 Education 40 non-null object

1 Occupation 40 non-null object

2 Salary 40 non-null int64

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

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

1. Let us formulate the hypothesis for conducting one-way ANOVA by considering only education as categorical field on which salary is dependent:
   * H0 (null hypothesis): Mean salaries remain same across various educational categories doctorate, bachelors and HS-grade.

μD = μB = μH

* + H1 (alternate hypothesis): Mean salaries aren’t same across various educational categories

μD != μB != μH

1. Let us formulate the hypothesis for conducting one-way ANOVA by considering only occupation as categorical field on which salary is dependent:
   * H0 (null hypothesis): Mean salaries remain same across various occupational categories Prof-specialty, Sales, Adm-clerical and Exec-managerial.

μP = μS = μA = μE

* + H1 (alternate hypothesis): Mean salaries aren’t same across various occupational categories

μP != μS != μA != μE

## Q 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.

Let us define a formula stating Salary is a o/p variable dependent on Categorical Variable Education.

*formula1 = 'Salary ~ C(Education)'*

Using the above formula, we will construct one-way ANOVA table as follows:

Graphical user interface, text, application

Description automatically generated

Table : One-way ANOVA Table for Education

From the above table p-value for Education is 1.25e-08. Since the p-values for Education is less than 0.05 we reject the Null Hypothesis, this means that Education has a statistically significant effect on Salary.

## Q 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.

Let us define a formula stating Salary is a o/p variable dependent on Categorical Variable Occupation.

*formula2 = 'Salary ~ C(*Occupation*)'*

Using the above formula, we will construct one-way ANOVA table as follows:

Text

Description automatically generated

Table : One-way ANOVA Table for Occupation

From the above table p-value for Occupation is 0.46. Since the p-values for Occupation is more than 0.05 we cannot reject the Null Hypothesis, this means that Occupation doesn’t have a statistically significant effect on Salary.

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

We have rejected the null hypothesis in 1.2.

Let us compute the mean salaries across educational categories.

Table

Description automatically generated

Table : Educational Categories verses Mean Salary

Chart, bar chart

Description automatically generated

Figure : Educational Categories verses Mean Salary Bar Plot

Let us compute the mean salaries across occupational categories.

Table

Description automatically generated with medium confidence

Table : Occupational Categories verses Mean Salary

Chart, bar chart

Description automatically generated

Figure : Occupational Categories verses Mean Salary Bar Plot

From the above tables and plots we infer Mean Salaries across various educational categories are significantly different.

## Q 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.

Let us now perform the Two-Way ANOVA. We will now analyse the effect of both the treatments Education and Occupation on the 'Salary' variable using following formula:

formula = 'Salary ~ C(Education) + C(Occupation)'

Using the above formula, we will construct two-way ANOVA table as follows:

Text

Description automatically generated

Table : Two-Way ANOVA Table

**From the above table p-value for Education 1.98e-08 is less than 0.05, this means that Education has a statistically significant effect on Salary.**

**And p-value for Occupation 3.55e-01 is greater than 0.05, this means that Occupation doesn’t have a statistically significant effect on Salary.**

**Let us check whether there is any interaction effect between the treatments.**

Chart, line chart

Description automatically generated

Figure : Education and Occupation Interaction Plot

**From the above intersection plot we can see that there is some sort of interaction between the two treatments.**

## Q 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?

**Since we saw that there was some sort of interaction between the two treatments, we will introduce a new interaction term Education\*Occupation while performing the Two-Way ANOVA.**

**formula = 'Salary ~ C(Education) + C(Occupation) + C(Education):C(Occupation)'**

Using the above formula, we will construct two-way ANOVA table as follows:

Text

Description automatically generated

Table : Two-Way ANOVA Table with Interaction Variable

Let us formulate the Hypothesis:

* + H0 (null hypothesis):
    - Mean salaries remain same across various educational categories.
    - Mean salaries remain same across various occupation categories.
    - There is no interaction between Education and Occupation.
  + H1 (alternate hypothesis):
    - Mean salaries aren’t same across various educational categories.
    - Mean salaries aren’t same across various occupational categories.
    - There is interaction between Education and Occupation.

**From the above table p-values for both Education 5.46e-12 and the interaction term 2.23e-05 are less than 0.05 so we can reject the Null Hypothesis, this means that Education has a statistically significant effect on Salary and there** is interaction between Education and Occupation.

**And p-value for Occupation 7.21e-02 is greater than 0.05 so we cannot reject the Null Hypothesis, this means that Occupation doesn’t have a statistically significant effect on Salary.**

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

From the above ANOVA analysis on give case study we observe Salary of an individual is not dependent on his Occupation Level. Salary is dependent on the educational level of an individual. Also, there is some sort of dependency of Salary on the interaction between Educational and Occupational Level.

People with higher educational level can expect better salary to be paid irrespective of the Occupational level they work.

Higher the Educational level better you get paid.

# **Problem 2**

## Executive Summary

The dataset contains information on various colleges. Here we will be performing Principal Component Analysis for this case study according to the instructions given.

### Data Dictionary

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

A screenshot of a computer

Description automatically generated with medium confidence

Table : Sample Dataset

Dataset consists of 777 rows and 18 columns.

### Exploratory Data Analysis

Let us check the types of variables in the data frame.

Names object

Apps int64

Accept int64

Enroll int64

Top10perc int64

Top25perc int64

F.Undergrad int64

P.Undergrad int64

Outstate int64

Room.Board int64

Books int64

Personal int64

PhD int64

Terminal int64

S.F.Ratio float64

perc.alumni int64

Expend int64

Grad.Rate int64

There are total 18 columns in the dataset. Out of which 16 columns are of integer type, 1 column is of object type and 1 column is of float type.

Check for missing values in the dataset

RangeIndex: 777 entries, 0 to 776

Data columns (total 18 columns):

# Column Non-Null Count Dtype

0 Names 777 non-null object

1 Apps 777 non-null int64

2 Accept 777 non-null int64

3 Enroll 777 non-null int64

4 Top10perc 777 non-null int64

5 Top25perc 777 non-null int64

6 F.Undergrad 777 non-null int64

7 P.Undergrad 777 non-null int64

8 Outstate 777 non-null int64

9 Room.Board 777 non-null int64

10 Books 777 non-null int64

11 Personal 777 non-null int64

12 PhD 777 non-null int64

13 Terminal 777 non-null int64

14 S.F.Ratio 777 non-null float64

15 perc.alumni 777 non-null int64

16 Expend 777 non-null int64

17 Grad.Rate 777 non-null int64

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

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

EDA is an approach to analyse data using both nonvisual and visual techniques.

Let us have a look at descriptive summary of the data.

Table, Excel

Description automatically generated

Table : Descriptive summary of the data

As part of EDA Univariate analysis, we will plot histogram or distplot to view the distribution of data and the box plot to view 5-point summary and outliers if any for each numeric column present in the dataset.

1. Apps

Chart, histogram

Description automatically generated Chart

Description automatically generated

Figure : Histogram and Box Plot of Apps

From the above figure we observe data in Apps column is right skewed and outliers are present.

1. Accept

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

Figure : Histogram and Box Plot of Accept

From the above figure we observe data in Accept column is right skewed and outliers are present.

1. Enroll

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

Figure : Histogram and Box Plot of Enroll

From the above figure we observe data in Enroll column is right skewed and outliers are present.

1. Top10perc

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

Figure : Histogram and Box Plot of Top10perc

From the above figure we observe data in Top10perc column is right skewed and outliers are present.

1. Top25perc

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

Figure : Histogram and Box Plot of Top25perc

From the above figure we observe data in Top25perc column is normally distributed and outliers are not present.

1. F.Undergrad

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

Figure : Histogram and Box Plot of F.Undergrad

From the above figure we observe data in F.Undergrad column is right skewed and outliers are present.

1. P.Undergrad

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

Figure : Histogram and Box Plot of P.Undergrad

From the above figure we observe data in P.Undergrad column is right skewed and outliers are present.

1. Outstate

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Figure : Histogram and Box Plot of Outstate

From the above figure we observe data in Outstate column is normally distributed and one outlier is present.

1. Room.Board

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

Figure : Histogram and Box Plot of Room.Board

From the above figure we observe data in Room.Board column is almost normally distributed and few outliers are present.

1. Books

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

Figure : Histogram and Box Plot of Books

From the above figure we observe data in Books column is right skewed and outliers are present.

1. Personal

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

Figure : Histogram and Box Plot of Personal

From the above figure we observe data in Personal column is right skewed and outliers are present.

1. PhD

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

Figure : Histogram and Box Plot of PhD

From the above figure we observe data in PhD column is left skewed and outliers are present.

1. Terminal

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

Figure : Histogram and Box Plot of Terminal

From the above figure we observe data in Terminal column is left skewed and outliers are present.

1. S.F.Ratio

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

Figure : Histogram and Box Plot of S.F.Ratio

From the above figure we observe data in S.F.Ratio column is right skewed and outliers are present.

1. perc.alumni

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

Figure : Histogram and Box Plot of perc.alumni

From the above figure we observe data in perc.alumni column is right skewed and outliers are present.

1. Expend

Histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

Figure : Histogram and Box Plot of Expend

From the above figure we observe data in Expend column is right skewed and outliers are present.

1. Grad.Rate

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

Figure : Histogram and Box Plot of Grad.Rate

From the above figure we observe data in Grad.Rate column is almost normally distributed and few outliers are present.

Observations

* There are total 17 numeric columns.
* Outliers present needs to be treated.
* In the given dataset different columns have different weights, so scaling the data is necessary.

Multivariate Analysis

Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.

Correlation Plot help us to visualize the correlation between continuous variables.

Pair Plot

A picture containing table

Description automatically generated

Figure : Pair Plot

Correlation Plot

Chart, treemap chart

Description automatically generated

Figure : Correlation Plot

From the correlation plot, we can see the correlation among different variables. Correlation values near to 1 or -1 are highly positively correlated and highly negatively correlated respectively. Correlation values near to 0 are not correlated to each other.

From the above plot we observe following columns are highly positively correlated:

* Apps and Accept
* Accept and Enroll
* Enroll and F.Undergrad

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

Scaling or standardisation of variables is must before applying PCA because it will give more emphasis to those variables having higher variances than to those variables with very low variances while identifying the right principal component.

Here in our dataset few columns like Apps, Accept, Enroll have information in terms of count, few columns like Top10perc, Top25pers, Terminal have information in terms of percentages, few columns like Room.Board, Books, Personal have information in terms of cost and there are columns like S.F.Ratio have information in terms of ratios.

Since the data present in different columns are on different units scaling is necessary.

Consider Apps column and Enroll column, both columns have information in terms of count. Values in Apps column may lie in the region 81 - 84094 while values in Enroll column lie in the region 35 – 6392 since Apps column is having higher variance compared to Enroll column PCA will give more weight to Apps. Here standardization is required to tackle these issues.

First, we will treat the outliers and then scale the data.

Diagram

Description automatically generated

Figure : Box Plots after Treating Outliers

Z-score technique is used to perform scaling of data.

A picture containing diagram

Description automatically generated

Table : Sample Scaled Data

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

A screenshot of a computer

Description automatically generated with medium confidence

Table : Sample Correlation Matrix After Scaling

A screenshot of a computer

Description automatically generated with medium confidence

Table : Sample Covariance Matrix After Scaling

Complete covariance matrix and correlation matrix after scaling is present in ipynb file.

Covariance is when two variables vary with each other, whereas Correlation is when the change in one variable results in the change in another variable.

Covariance values will not be standardized. Correlation values will be standardized.

Since our data is already scaled [standardised] both covariance and correlation matrix will be almost same.

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

Let us analyse the outliers with the help of Box Plot.

Diagram

Description automatically generated

Figure : Box Plots Before Scaling

Diagram

Description automatically generated

Figure : Box Plots After Scaling

Scaling does not have any impact on outliers. Only values in the dataset will be scaled, as part of it even outliers also get scaled and they still continue to remain as outliers.

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

Eigen Vectors are as follows:

array([[ 0.262, 0.231, 0.189, 0.339, 0.335, 0.163, 0.022, 0.284,

0.244, 0.097, -0.035, 0.326, 0.323, -0.163, 0.187, 0.329,

0.239],

[ 0.314, 0.345, 0.383, -0.099, -0.06 , 0.399, 0.358, -0.252,

-0.132, 0.094, 0.232, 0.055, 0.043, 0.26 , -0.257, -0.16 ,

-0.168],

[-0.081, -0.108, -0.086, 0.079, 0.051, -0.074, -0.04 , -0.015,

0.021, 0.697, 0.531, -0.081, -0.059, -0.274, -0.104, 0.184,

-0.245],

[ 0.099, 0.118, 0.009, -0.369, -0.417, 0.014, 0.225, 0.263,

0.581, -0.036, -0.115, -0.147, -0.089, -0.259, -0.224, 0.214,

-0.036],

[ 0.22 , 0.19 , 0.162, 0.157, 0.144, 0.103, -0.096, 0.037,

-0.069, 0.035, -0. , -0.551, -0.59 , -0.143, 0.128, -0.022,

0.357],

[ 0.002, -0.017, -0.068, -0.089, -0.028, -0.052, -0.025, -0.02 ,

0.237, 0.639, -0.381, 0.003, 0.035, 0.469, 0.013, -0.232,

0.314],

[-0.028, -0.013, -0.015, -0.257, -0.239, -0.031, -0.01 , 0.095,

0.095, -0.111, 0.639, 0.089, 0.092, 0.153, 0.391, -0.151,

0.469],

[-0.09 , -0.138, -0.144, 0.29 , 0.346, -0.109, 0.124, 0.011,

0.39 , -0.24 , 0.277, -0.034, -0.09 , 0.243, -0.566, -0.119,

0.18 ],

[-0.131, -0.142, -0.051, 0.122, 0.194, -0.001, 0.635, 0.008,

0.221, -0.021, -0.017, -0.167, -0.113, 0.154, 0.539, -0.024,

-0.316],

[-0.156, -0.149, -0.065, -0.036, 0.006, -0. , 0.546, -0.232,

-0.255, 0.091, -0.128, 0.101, 0.086, -0.471, -0.148, -0.08 ,

0.488],

[-0.086, -0.043, -0.044, 0.002, -0.102, -0.035, 0.252, 0.593,

-0.475, 0.044, 0.015, -0.039, -0.085, 0.363, -0.174, 0.394,

0.087],

[-0.09 , -0.159, 0.035, 0.039, -0.146, 0.134, -0.05 , -0.56 ,

0.107, -0.052, -0.009, 0.072, -0.164, 0.24 , 0.049, 0.69 ,

0.159],

[-0.089, -0.044, 0.062, -0.07 , 0.097, 0.087, -0.045, -0.067,

-0.018, -0.035, 0.012, -0.703, 0.662, 0.048, -0.036, 0.127,

0.063],

[-0.549, -0.292, 0.417, -0.009, 0.011, 0.571, -0.146, 0.212,

0.101, 0.029, -0.034, 0.064, -0.099, -0.062, -0.028, -0.129,

0.007],

[ 0.005, 0.014, -0.05 , -0.724, 0.655, 0.025, -0.04 , -0.002,

-0.028, -0.008, 0.001, 0.083, -0.113, 0.004, -0.007, 0.145,

-0.003],

[ 0.599, -0.661, -0.233, -0.022, -0.032, 0.368, -0.026, 0.081,

-0.027, -0.01 , -0.005, -0.013, 0.018, -0.018, 0. , -0.056,

-0.015],

[-0.182, 0.391, -0.717, 0.056, -0.02 , 0.543, -0.03 , -0.001,

-0.01 , -0.004, 0.011, -0.013, -0.007, -0.009, 0.024, -0.011,

0.003]])

Eigen Values are as follows:

array([5.663, 4.895, 1.126, 1.004, 0.872, 0.766, 0.585, 0.545, 0.424,

0.381, 0.247, 0.147, 0.134, 0.099, 0.075, 0.038, 0.022])

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

Following is the dataframe of Principal Components with original features.

Table

Description automatically generated

Table : Dataframe of PC and Original Features

## Q 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]

From the above dataframe we will be able to write following linear equations:

PC1 = 0.26Apps + 0.23Accept + 0.19Enroll + 0.34Top10perc + 0.33Top25perc + 0.16F.Undergrad + 0.02P.Undergrad + 0.28Outstate + 0.24Room.Board + 0.1Books - 0.04Personal + 0.33PhD + 0.32Terminal - 0.16S.F.Ratio + 0.19perc.alumni + 0.33Expend + 0.24Grad.Rate

PC2 = 0.31Apps + 0.34Accept + 0.38Enroll - 0.1Top10perc - 0.06Top25perc + 0.4F.Undergrad + 0.36P.Undergrad - 0.25Outstate - 0.13Room.Board + 0.09Books + 0.23Personal + 0.06PhD + 0.04Terminal + 0.26S.F.Ratio - 0.26perc.alumni - 0.16Expend - 0.17Grad.Rate

PC3 = -0.08Apps - 0.11Accept - 0.09Enroll + 0.08Top10perc + 0.05Top25perc - 0.07F.Undergrad -0.04P.Undergrad - 0.01Outstate + 0.02Room.Board + 0.7Books + 0.53Personal - 0.08PhD -0.06Terminal - 0.27S.F.Ratio - 0.1perc.alumni + 0.18Expend - 0.25Grad.Rate

PC4 = 0.1Apps + 0.12Accept + 0.01Enroll - 0.37Top10perc - 0.42Top25perc + 0.01F.Undergrad + 0.23P.Undergrad + 0.26Outstate + 0.58Room.Board - 0.04Books - 0.11Personal - 0.15PhD -0.09Terminal - 0.26S.F.Ratio - 0.22perc.alumni + 0.21Expend - 0.04Grad.Rate

PC5 = 0.22Apps + 0.19Accept + 0.16Enroll + 0.16Top10perc + 0.14Top25perc + 0.1F.Undergrad -0.1P.Undergrad + 0.04Outstate - 0.07Room.Board + 0.04Books - 0.0Personal - 0.55PhD -0.59Terminal - 0.14S.F.Ratio + 0.13perc.alumni - 0.02Expend + 0.36Grad.Rate

PC6 = 0.0Apps - 0.02Accept - 0.07Enroll - 0.09Top10perc - 0.03Top25perc - 0.05F.Undergrad -0.02P.Undergrad - 0.02Outstate + 0.24Room.Board + 0.64Books - 0.38Personal + 0.0PhD + 0.04Terminal + 0.47S.F.Ratio + 0.01perc.alumni - 0.23Expend + 0.31Grad.Rate

PC7 = -0.03Apps - 0.01Accept - 0.02Enroll - 0.26Top10perc - 0.24Top25perc - 0.03F.Undergrad - 0.01P.Undergrad + 0.09Outstate + 0.09Room.Board - 0.11Books + 0.64Personal + 0.09PhD + 0.09Terminal + 0.15S.F.Ratio + 0.39perc.alumni - 0.15Expend + 0.47Grad.Rate

PC8 = -0.09Apps - 0.14Accept - 0.14Enroll + 0.29Top10perc + 0.35Top25perc - 0.11F.Undergrad + 0.12P.Undergrad + 0.01Outstate + 0.39Room.Board - 0.24Books + 0.28Personal - 0.03PhD - 0.09Terminal + 0.24S.F.Ratio - 0.57perc.alumni - 0.12Expend + 0.18Grad.Rate

PC9 = -0.13Apps - 0.14Accept - 0.05Enroll + 0.12Top10perc + 0.19Top25perc - 0.0F.Undergrad + 0.63P.Undergrad + 0.01Outstate + 0.22Room.Board - 0.02Books - 0.02Personal - 0.17PhD - 0.11Terminal + 0.15S.F.Ratio + 0.54perc.alumni - 0.02Expend - 0.32Grad.Rate

PC10 = -0.16Apps - 0.15Accept - 0.06Enroll - 0.04Top10perc + 0.01Top25perc - 0.0F.Undergrad + 0.55P.Undergrad - 0.23Outstate - 0.26Room.Board + 0.09Books - 0.13Personal + 0.1PhD + 0.09Terminal - 0.47S.F.Ratio - 0.15perc.alumni - 0.08Expend + 0.49Grad.Rate

PC11 = -0.09Apps - 0.04Accept - 0.04Enroll + 0.0Top10perc - 0.1Top25perc - 0.03F.Undergrad + 0.25P.Undergrad + 0.59Outstate - 0.48Room.Board + 0.04Books + 0.02Personal - 0.04PhD - 0.08Terminal + 0.36S.F.Ratio - 0.17perc.alumni + 0.39Expend + 0.09Grad.Rate

PC12 = -0.09Apps - 0.16Accept + 0.04Enroll + 0.04Top10perc - 0.15Top25perc + 0.13F.Undergrad - 0.05P.Undergrad - 0.56Outstate + 0.11Room.Board - 0.05Books - 0.01Personal + 0.07PhD - 0.16Terminal + 0.24S.F.Ratio + 0.05perc.alumni + 0.69Expend + 0.16Grad.Rate

PC13 = -0.09Apps - 0.04Accept + 0.06Enroll - 0.07Top10perc + 0.1Top25perc + 0.09F.Undergrad - 0.04P.Undergrad - 0.07Outstate - 0.02Room.Board - 0.04Books + 0.01Personal - 0.7PhD + 0.66Terminal + 0.05S.F.Ratio - 0.04perc.alumni + 0.13Expend + 0.06Grad.Rate

PC14 = -0.55Apps - 0.29Accept + 0.42Enroll - 0.01Top10perc + 0.01Top25perc + 0.57F.Undergrad - 0.15P.Undergrad + 0.21Outstate + 0.1Room.Board + 0.03Books - 0.03Personal + 0.06PhD - 0.1Terminal - 0.06S.F.Ratio - 0.03perc.alumni - 0.13Expend + 0.01Grad.Rate

PC15 = 0.01Apps + 0.01Accept - 0.05Enroll - 0.72Top10perc + 0.66Top25perc + 0.03F.Undergrad - 0.04P.Undergrad - 0.0Outstate - 0.03Room.Board - 0.01Books + 0.0Personal + 0.08PhD - 0.11Terminal + 0.0S.F.Ratio - 0.01perc.alumni + 0.15Expend - 0.0Grad.Rate

PC16 = 0.6Apps - 0.66Accept - 0.23Enroll - 0.02Top10perc - 0.03Top25perc + 0.37F.Undergrad - 0.03P.Undergrad + 0.08Outstate - 0.03Room.Board - 0.01Books - 0.0Personal - 0.01PhD + 0.02Terminal - 0.02S.F.Ratio + 0.0perc.alumni - 0.06Expend - 0.01Grad.Rate

PC17 = -0.18Apps + 0.39Accept - 0.72Enroll + 0.06Top10perc - 0.02Top25perc + 0.54F.Undergrad - 0.03P.Undergrad - 0.0Outstate - 0.01Room.Board - 0.0Books + 0.01Personal - 0.01PhD - 0.01Terminal - 0.01S.F.Ratio + 0.02perc.alumni - 0.01Expend + 0.0Grad.Rate

## Q 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?

Let us use cumulative explained variance ratio to find a cut off for selecting the number of PC’s.

Following is the cumulative explained variance ratio:

array([0.33266084, 0.62021429, 0.68638592, 0.74536736, 0.79660629,

0.84159268, 0.8759551 , 0.90794357, 0.93282465, 0.95520861,

0.96972018, 0.97837162, 0.98626408, 0.99207036, 0.99645823,

0.99868442, 1. ])

Chart, histogram

Description automatically generated

Figure :Cumulative Explained Ratio verses PC Index

From the above array and plot if 𝑘 = 6, cumulative proportion is 84.15%. Although there are 17 observed variables, the first 6 principal components can explain more than 80% of the total variation. Hence it is sufficient to use the first 6 PCs instead of the original 17 variables.

The eigenvectors and eigenvalues of a covariance matrix represent the “core” of a PCA: The eigenvectors determine the directions of the new feature space, and the eigenvalues determine their magnitude.

## Q 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]

For a model building activity, too much of data can be a bad thing. At a certain point, more features or dimensions can decrease a model’s accuracy since there is more data that needs to be generalized this is known as the curse of dimensionality.

Principal Component Analysis (PCA) is an unsupervised linear transformation technique. PCA aims to find the directions of maximum variance in high-dimensional data and projects it onto a new subspace with equal or fewer dimensions than the original one.

In the current case-study we have total 17 continuous columns which are 17 different dimensions. If we intend to build a model, we have to consider all 17 features for model building. As mentioned earlier model building with a greater number of features will be a complex task.

So, after applying PCA on current dataset we obtain 17 principal components. From the cumulative explained variance ratio we observe first 6 Principal components contain more than 80% of variance information which is sufficient. Hence it is sufficient to use the first 6 PCs in new subspace instead of the original 17 variables.

Each Principal component obtained is the linear combination of 17 original features.

Let us check as to how the original features matter to each PC by considering the absolute values.

Chart, bar chart, histogram, waterfall chart

Description automatically generated

Chart, histogram, waterfall chart

Description automatically generated

Figure :Absolute Values of Each Original Feature in Different Principal Components

Going further we can use these 6 Principal components instead of 17 features.

Let us obtain the scores as the dot product between the loadings and features. And using these scores we will create a dataframe out of fit transformed scaled data above.

Table

Description automatically generated

Table : Sample Dataset with Principal Components.

All the Principal Components obtained will be orthogonal to each other and there won’t be any correlations among them.

Chart

Description automatically generated

Figure : Correlation Plot of Principal Components