SMDM Project Report

PGP-DSBA Online

May ’22

Date: 28/08/2022

Contents

1 Problem Statement…………………………………………………………………………………………………………………………………………4

1.1 State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually……………………………………………………………………………………………………………………….4

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…………………………………………………………4

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………………………………………………………………………………………………………………………………………………….4

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……………………………………………………………………………………………………………….5

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………………………………………………..5

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?.......................................................................................6

1.7 Explain the business implications of performing ANOVA for this particular case study…………………..6

2 Problem Statement…………………………………………………………………………………………………………………………………………7

2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?................................................................................................7

2.2 Is scaling necessary for PCA in this case? Give justification and perform scaling…………………………..15

2.3 Comment on the comparison between the covariance and the correlation matrices from this data.[on scaled data]………………………………………………………………………………………………………………………….16

2.4 Check the dataset for outliers before and after scaling. What insight do you derive here?.............18

2.5 Extract the eigenvalues and eigenvectors……………………………………………………………………………………..21

2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features…………………………………………………………………………………………………………………………..22

2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only).……………………………………………………………………………………………………………………………….24

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?................................................25

2.9 Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis?.....................................................................................................28

**List of Figures**

Figure 1 One-Way ANOVA on Salary with respect to Education……………………………………………4

Figure 2 One-Way ANOVA on Salary with respect to Occupation…………………………………………4

Figure 3 Apps Distribution……………………………………………………………………………………………………5

Figure 4 Accept Distribution…………………………………………………………………………………………………5

Figure 5 Enroll Distribution…………………………………………………………………………………………………..6

Figure 6 Top10Perc Distribution…………………………………………………………………………………………..6

Figure 7 Top25Perc Distribution…………………………………………………………………………………………..6

Figure 8 F.Undergrad Distribution………………………………………………………………………………………..7

Figure 9 P.Undergrad Distribution………………………………………………………………………………………..7

Figure 10 Outstate Distribution…………………………………………………………………………………………….8

Figure 11 RoomBoard Distribution……………………………………………………………………………………….8

Figure 12 Books Distribution………………………………………………………………………………………………..8

Figure 13 Personal Distribution……………………………………………………………………………………………9

Figure 14 PhD Distribution…………………………………………………………………………………………………..9

Figure 15 Terminal Distribution………………………………………………………………………………………….10

Figure 16 S.F.Ratio Distribution………………………………………………………………………………………….10

Figure 17 PercAlumni Distribution………………………………………………………………………………………11

Figure 18 Expend Distribution…………………………………………………………………………………………….11

Figure 19 Grad Rate Distribution…………………………………………………………………………………………12

Figure 20 Multivariate analysis – Heat Map……………………………………………………………………….12

Figure 21 Covariance Matrix……………………………………………………………………………………………….14

Figure 22 Correlation Table…………………………………………………………………………………………………15

Figure 23 Correlation Heat Map………………………………………………………………………………………….16

Figure 24 Outlier check before Scaling…………………………………………………………………………………17

Figure 25 Outlier check after Scaling…………………………………………………………………………………..17

Figure 26 Outlier Treatment……………………………………………………………………………………………….19

Figure 27 Scree Plot……………………………………………………………………………………………………………..21

Figure 28 Eigen-Value Graph……………………………………………………………………………………………….24

**1 Problem Statement**

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/78180/files/6370335/download?verifier=BM88Y7a8Yw7gDjGGKEGEeuO2Px62yQY8qzZok30z&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.

**1.1 Hypothesis for conducting One-Way ANOVA for Education and Occupation**

|  |  |  |
| --- | --- | --- |
|  | **One-Way ANOVA for Education** | **One-Way ANOVA for Occupation** |
| **H0** | **M1=M2=M3=Mk - All Education level mean salaries are equal** | **M1=M2=M3=Mk - All Occupation level mean salaries are equal** |
| **H1** | **Any two salary means are not equal** | **Any two salary means are not equal** |

**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 results.**

Graphical user interface, text, application

Description automatically generated

Figure 1 One-Way ANOVA on Salary with respect to Education

P-value is less than the significance level - 0.05, hence we can reject the null-hypothesis and conclude that not all education level mean salaries are equal

**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 results.**

Graphical user interface, text

Description automatically generated with medium confidence

Figure 2 One-Way ANOVA on Salary with respect to Occupation

P-value is greater than the significance level, alpha = 0.05, hence we can accept the null-hypothesis and conclude that all occupation level mean salaries are equal

**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.**

Graphical user interface, text

Description automatically generated

The null-Hypothesis is rejected for education level salaries. From above, we can interpret that the HS-Grad class mean is significantly different from the entire population mean.

**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**

Chart, line chart

Description automatically generated

People with HS-Grad education do not acquire Exec-Managerial positions[¶](http://localhost:8888/notebooks/Desktop/GL/Adv%20Stats/AS_Project_SwaathiRamakrishnan_CodeFile.ipynb#People-with-HS-Grad-education-do-not-acquire-Exec-Managerial-positions)

People with HS-Grad education earn the least when compared to Docorate and Bachelors degree

People with Doctorate degree earn the highest among the 3 education levels

For Admin-Clerical and Sales, the salary pay is almost the same for both Doctorate and Bachelors education level

We can infer from the above that education level impacts the occupation which in turn will affect the salary level

**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?**

Text

Description automatically generated

**H0**: The effect of Education on Mean Salary is independent to the effect of Occupation on Mean Salary (there is no interaction between the 2 variables Occupation and Education)

**H1**: Interaction effect is present between the two variables Occupation and Education on the mean salary.

**Interpretation**: Since the p-value of the interaction between the variables is lesser than alpha = 0.05, the null hypothesis can be rejected. Hence it can be concluded that there is a significant interaction effect between occupation and education on the mean salary

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

From the above ANOVA analysis and Interaction Plots it is observed that salary is impacted by both education and occupation along with their interaction effect. People who have Doctorate seem to be drawing the maximum salary and people with HS-Grad draw the least. By performing the ANOVA tests it is clearly seen that only education or only occupation does not help in drawing good salaries. A combination of both produces better results.

**Problem 2:**

The dataset [Education - Post 12th Standard.csv](https://olympus.mygreatlearning.com/courses/78180/files/5510379/download?verifier=V7FXa5qDDiX6kKrYf6QnDsNFLacD0bddD8fB2orc&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.mygreatlearning.com/courses/78180/files/5510378/download?verifier=VV7P65zCUC8DHgEGTjlwSDaCH3PcDbJLHkBOqa4S&wrap=1).

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

Below a univariate analysis for the numerical data is performed to find patterns in the data and understand them.

Chart

Description automatically generated

*Figure 3 Apps Distribution*

The above plot indicates that we have outliers in Apps data. The maximum number of applications received seems to be at 50000.

Chart

Description automatically generated

*Figure 4 Accept Distribution*

The above plot indicates that we have many outliers in Accept data. The maximum number of applications accepted seems to be > 25000.

Chart

Description automatically generated

*Figure 5 Enroll Distribution*

From the above plots we can understand that most of the colleges have enrolled less than 1000 new students. The data is positively skewed and has many outliers.

Chart, histogram, box and whisker chart

Description automatically generated

*Figure 6 Top10Perc Distribution*

The above plot indicates that the colleges enroll 30 to 50 percentage of new students from top 10% of Higher Secondary class. The plot has outliers and is right skewed.

Chart, histogram

Description automatically generated

*Figure 7 Top25Perc Distribution*

The above plot indicates that the colleges enroll 40 to 70 percentage of new students from top 25% of Higher Secondary class. The plot is normally distributed and has no outliers.

A picture containing chart

Description automatically generated

*Figure 8 F.Undergrad Distribution*

The above plot indicates that the colleges have less than 5000 Fulltime students. The plot is right skewed and has outliers.

A picture containing chart

Description automatically generated

*Figure 9 P.Undergrad Distribution*

The above plot indicates that the colleges have less than 3000 Parttime students. The plot is right skewed and has outliers.

Chart, histogram

Description automatically generated

*Figure 10 Outstate Distribution*

The above plot indicates that the colleges have around 10000 outstation students. The plot is normally distributed and has 1 outlier.

Chart, histogram

Description automatically generated

*Figure 11 RoomBoard Distribution*

The above plot indicates that the cost of room and boarding in the colleges is below 5000. The plot is normally distributed and has outliers.

Chart

Description automatically generated

*Figure 12 Books Distribution*

The above plot indicates that the estimated cost of books for each student is around 500. The plot is right skewed and has outliers.

Chart

Description automatically generated

*Figure 13 Personal Distribution*

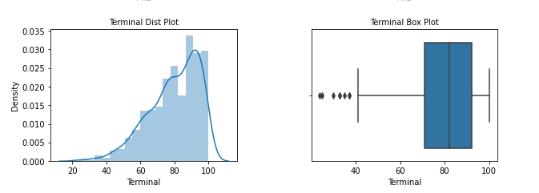
The above plot indicates that the estimated personal spending for each student is around 2000. The plot is right skewed and has outliers.

Chart, histogram

Description automatically generated

*Figure 14 PhD Distribution*

The above plot indicates that tthere are 60 to 80 faculties with PhD degrees. The plot is left skewed and has outliers.



*Figure 15 Terminal Distribution*

The above plot indicates that there are 70 to 100 faculties with terminal degrees. The plot is left skewed and has outliers.

Chart

Description automatically generated

*Figure 16 S.F.Ratio Distribution*

The above plot indicates that the colleges have 10-15 Student Faculty ratio. The plot is right skewed and has outliers.

Chart, histogram

Description automatically generated

*Figure 17 PercAlumni Distribution*

The above plot indicates that the colleges have 10-30 percentage of Alumni who donate. The plot is right skewed and has outliers.

Chart

Description automatically generated

*Figure 18 Expend Distribution*

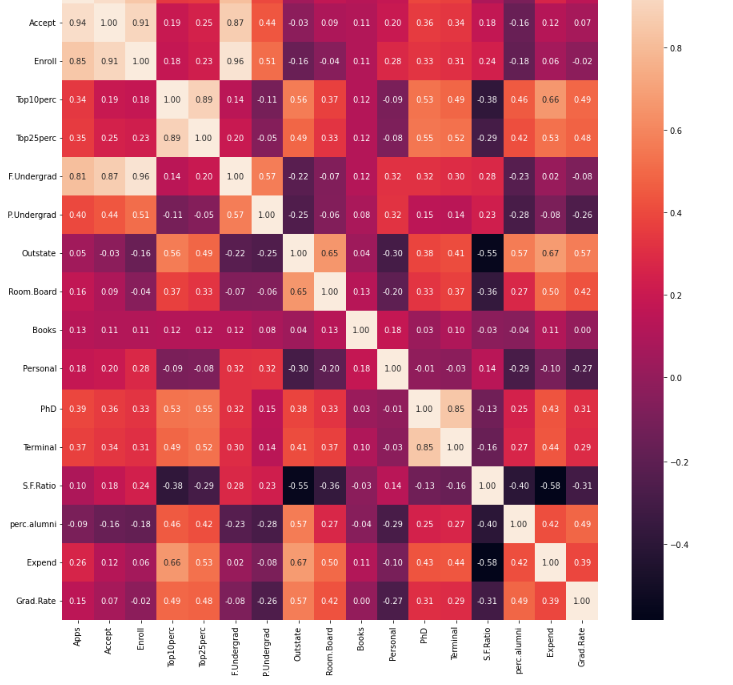
The above plot indicates that the Instructional expenditure per student is around 10000. The plot is right skewed and has outliers.

Chart, histogram

Description automatically generated

*Figure 19 Grad Rate Distribution*

The above plot indicates that the Graduation Rate is between 50 and 80 percent. The plot is left skewed and has outliers.



*Figure 20 Multivariate analysis – Heat Map*

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

Scaling is an important step in doing Exploratory Data Analysis before doing PCA because we need to transform all the features so that they are in the same scale. The categorical column “Names” is dropped and the scaling technique z-score is applied.

Graphical user interface

Description automatically generated with medium confidence

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

Both of these terms indicate the relationship between two variables. Covariance tells us if the relationship between the variables is positive or negative i.e directly proportional or inversely proportional to each other. Whereas correlation tells us how much the 2 variables are correlated i.e we can understand which variables are highly positively correlated or negatively correlated.

Table

Description automatically generated

*Figure 21 Covariance Matrix*

Graphical user interface, application, table, Excel

Description automatically generated

*Figure 22 Correlation Table*

Chart, treemap chart

Description automatically generated

*Figure 23 Correlation Heat Map*

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

Chart, box and whisker chart

Description automatically generated

*Figure 24 Outlier check before Scaling*

Chart, box and whisker chart

Description automatically generated

*Figure 25 Outlier check after Scaling*

We can infer from the above 2 plots that scaling does not help in treating the outliers in the data. We need to perform outlier treatment to remove outliers.

Outliers in the dataset are significantly different than the rest of the values in the feature. Hence treating them will help in further analysis of PCA.

Graphical user interface, text, application, email

Description automatically generated

Chart, box and whisker chart

Description automatically generated

*Figure 26 Outlier Treatment*

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

Eigenvectors are nothing but the components.

Table

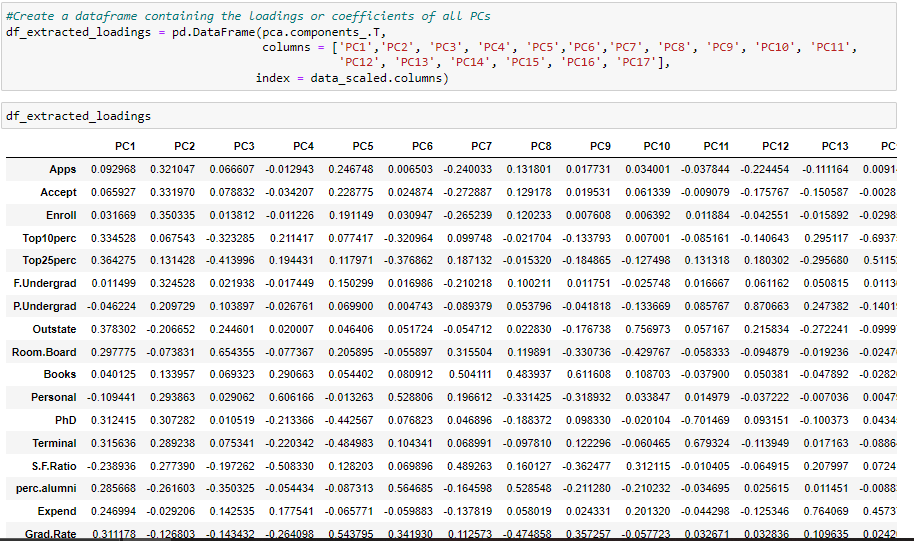
Description automatically generated

Eigen values explain the variance and are denoted in descending order to show us which component has the maximum variance

Text

Description automatically generated

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



Scree Plot

Chart, line chart

Description automatically generated

*Figure 27 Scree Plot*

Table

Description automatically generated

**2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only).**

Explicit form of 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 + 0.25 \* Grad.Rate

**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 sum of he eigen values help us decide the values that are upto 80%. Hence we select 8 principal components for further analysis.

Table

Description automatically generated

By plotting the cumulative eigen\_value graph below, we can understand that the eigen vectors with insignificant contribution to total eigen values can be removed from analysisChart, histogram

Description automatically generated

*Figure 28 Eigen-Value Graph*

A picture containing graphical user interface

Description automatically generated

**2.9 Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis?**

PCA is a dimension reduction technique to capture as much information as possible. Correlated variables are transformed to linear uncorrelated principal components. We extract important variables from a dataset that contains a large set of variables to bring out strong patterns in the dataset.

We begin by standardizing the dataset by shifting the data points to the origin. Then the eigen vectors (also known as components) and eigen values (the magnitudes of variance) are determined after this and sorted in descending order of eigen values, which gives us the first Principal Component that covers the maximum information from the Dataset.

We have 17 features in this dataset but selecting the top few components would explain 99% of the total variance. Hence, we select 8 components for our analysis.

In PCA, n\_components denote how many components are returned after fit and transformation. The amount of variance explained by each of the component is denoted by explained\_variance\_ and the explained\_variance\_ratio\_ is the percentage of that variance. From the scree plot and the cumulative explained variance ratio we choose only 8 components for our analysis. From this it is observed that 8 of the components help us understand the maximum variance in the dataset and the rest can be eliminated from further analysis.

The bar and line plot showing the explained variance and components helps us to understand the eigen vectors with insignificant contribution to eigen values, and thereby removed from the analysis.

Eigen-Value graph

Chart, histogram

Description automatically generated