ADVANCED sTATISTICS pROJECT

VISHAL.S

G3 PGPDSBA JULY

|  |  |  |
| --- | --- | --- |
| SNO | TITLE | PAGE NUMBER |
| A | INTRODUCTION | 2 |
| 1 | EXPOLATORY DATA ANALYSIS | 3 |
| 4 | SOLUTION | 3 |
| 4.1 | PROBLEM 1 | 3 |
| 4.2 | PROBLEM 2 | 3 |
|  |  | 4 |
|  |  | 4 |
|  |  | 4 |
|  |  | 4 |
|  |  | 4 |
|  |  | 6 |
| 4 | SOLUTION | 9 |
| 4.1 | PROBLEM 1 | 9 |
| 4.2 | PROBLEM 2 | 10 |
| 5 | CONCLUSION | 12 | |
| 5.1 | PROBLEM 1 | 12 | |
| 5.2 | PROBLEM 2 | 12 | |

TABLE OF CONTENTS

INTRODUCTION

Problem 1

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [[SalaryData.csv](https://olympus.greatlearning.in/courses/53915/files/3876960/download?verifier=wNllJ2eK76wljWYHTgU5FhDGkDkAHkE9R8YQuPx6&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. ANOVA is expected to be done on this data set

Problem 2

The dataset [Education - Post 12th Standard.csv](https://olympus.greatlearning.in/courses/53915/files/3110239/download?verifier=CCw6ItHHum7cHik3MXzchykkv9ODrRRLe7utcfpX&wrap=1) contains information on various colleges. Principal Component Analysis is expected to be done on this dataset

SAMPLE OF THE DATA SET

Table

Description automatically generated

Table 1. Dataset Sample

Table

Description automatically generated

Table 2. Dataset Sample for Problem 2

Exploratory Data Analysis

Types of Variables in the Datasets

Text

Description automatically generated

Table 3. Types of Data in Salary data set

Table

Description automatically generated

Table 4. Types of Data in Education data set

Checking for missing values in the dataset:

Graphical user interface, text

Description automatically generated with medium confidence

Table 5. Missing Values in Salary dataset

A picture containing table

Description automatically generated Table 6. Missing Values in Education dataset

There are no missing values present in both datasets.

SOLUTION

Problem 1:

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

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

Solution: Formulation of hypothesis for conducting one-way ANOVA for education qualification w.r.t salary

* H0: Salary depend on education qualification
* Ha: Salary does not depend on education
* Confidence level = 0.05

Formulation of hypothesis for conducting one-way ANOVA for occupation w.r.t salary

* H0: Salary depend on occupation
* Ha: Salary does not depend on occupation
* Confidence level = 0.05
  1. 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.

Solution: 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. We get following output:

Text

Description automatically generated

Table 7. AOV Table for Education with respect to the variable Salary

From the above table, we find that the P value is less than 0.05, hence we reject the null hypothesis.

* 1. 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.

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

Text

Description automatically generated Table 8. AOV Table for Occupation with respect to the variable Salary

From the above table, we find that the P value is greater than 0.05, hence we do not reject the null hypothesis

* 1. 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

Chart, line chart

Description automatically generated

Figure 1. Point Plot

* 1. 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?

Solution: Formulation of hypothesis for conducting two-way ANOVA based on education and occupation w.r.t salary.

• H0: Salary depends on both categories - education and occupation

• Ha: Salary does not depend on at least one of the categories - education and occupation • Confidence level = 0.05

Text

Description automatically generated

Table 9. Two Way AOV Table for variables Occupation and Education with respect to the variable Salary

Considering both education and occupation, education is a significant factor as the P value is 0.05

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

Solution: By performing ANOVA on the given data set, we can conclude that salary is dependent on occupation.

Problem 2:

The dataset Education - Post 12th Standard.csv contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions given.

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

Solution: Firstly, after importing all the relevant libraries on Jupyter notebook, we load the data set. Then, we perform EDA to extract and see patterns in the given data set. The given data set has a shape of (777, 18). Also, we check the top 5 rows of the data set then went on to see if there are any missing values in it – as per the output there are no missing values.

The analysis of all these variables includes:

• Statistical description of the numeric variable

• Distribution of the column with histogram or distplot

Chart, bar chart

Description automatically generated

Figure 2. Box Plot for Apps

The output displays, total 17\*3 = 51 distinct charts/columns. Hence I have put the screenshot of only one variable i.e. apps (Please refer the python notebook for your perusal). Further, we perform multivariate analysis, using correlation function in which we get below output.

Chart, treemap chart

Description automatically generated

Figure 3. HeatMap

Insights:

• The average (mean) number of applications received by the listed universities is around 3,001

• The number of applications accepted ranges from 72 to 26,330

• Average student enrolment is around ~880

• Median of new students from top 10% of higher secondary class is 23%

• Average book cost is around 550

• The minimum S.F. ratio is around 2.5

• Average percentage of faculties with Ph.D.’s is 72.6

• There are considerable number of variables that are highly correlated

• “Apps” has high correlation with “Accept”, and ”Enroll”

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

Solution: Yes, it is necessary to perform scaling for PCA. For instance, in given data set, applications and other variables are having values in thousands and few variables such as percentile is in just two digits. So, the data in these variables are of different scales, it is tough to compare these variables. The PCA calculates a new projection of the given data set and the new axis are based on the standard deviation of the variables. So a variable with a high standard deviation in the data set will have a higher weight for the calculation of axis than a variable with a low standard deviation. By performing scaling, we can easily compare these variables.

Graphical user interface, application

Description automatically generated

Table 10. Z Score Scaling table

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

Table

Description automatically generated

Figure 4. Covariance matrix

Table, Excel

Description automatically generated

Figure 5. Correlation matrix

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

Solution: Before scaling, let’s plot a boxplot to check the outliers in all the variables. We get the following output:

Chart, box and whisker chart

Description automatically generated

Figure 6. Box Plot Before scaling

Post scaling, let’s plot a boxplot to check the outliers in all the variables. We get the following output:

Chart, box and whisker chart

Description automatically generated

Figure 7. Box Plot after scaling

Insights:

• By scaling, all variables have the same standard deviation, thus all variables have the same weight and thus resulting in PCA calculating relevant axis.

• Before scaling, we only had one variable with no outliers (top25 perc); Post scaling, we have multiple variables with negligible outliers – this is achieved by normalizing the scale of the variables

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

Text, table

Description automatically generated

Figure 8. Eigen Vectors

Text

Description automatically generated

Figure 9. Eigen Values

2.6 Perform PCA and export the data of the Principal Component scores into a data frame.

Solution: For performing PCA, we need to follow below steps:

# Step 1: Generate the covariance matrix

# Step 2: Get eigenvalues and eigenvector

# Step 3: View Scree Plot to identify the number of components to be built

# Step 4: We can perform PCA on the scaled data set by importing PCA from sklearn. decomposition.

We get following component output:

Graphical user interface, text, application

Description automatically generated

Table 11. Principal Components Score

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

Solution: Eigen Vector of First PC

[2.42 3.24 9.77 -1.02 2.28 -4.76 1.23 -3.41 -1.84 -1.34 -6.79 -1.51 5.73 2.54 -3.50 4.76 -2.73]

If we sort the eigenvectors in descending order with respect to their eigenvalues, we will have that the first eigenvector accounts for the largest spread among data, the second one for the second largest spread and so on.

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?

From the below screenshot of cumulative values of the eigenvalues, we can see that around 8 principal components explained over 90% of the variance. Thus, the optimum number of principal components can be 8.

Text

Description automatically generated

Figure 10. Cumulative Variance

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]

Solution: We know that the principal components describe the amount of the total variance that can be explained by a single dimension of the data. As mentioned above, we have generated only 8 PCA dimensions. These 8 PCA can be used for further analysis, representing more than 90% of the variance.

In this case study, we had 17 numeric variables to be assessed, with PCA we did dimensionality reduction from 17 to 8 (representing more than 90% of the variance). However, we can see from the above-mentioned cumulative variance that even 5 PCA dimensions represent around 80% of the variance.

But to be on a safer side, we have considered to go with 90% variance. Thus, as far as business implication of using PCA is concerned, in this case, we are reducing a high dimensional space (with 17 variables) and converting it to a lower dimensional space without (theoretically) losing much of the explanatory power.

Chart

Description automatically generated

Figure 10. Heat map 2

Following are the interpretations from the obtained PCs

• PC1: Explains No. of students for whom the college or university is Out-of-state tuition and instructional expenditure per student

• PC2: Represents the highly correlated variables such as Apps, Enroll and Accept

• PC3: Highlights the estimated cost of books for a student

• PC4: Represents % of faculties with Ph.D.’s and terminal degree

• PC5: Explains percentage of new students from top 10% and 25% of higher secondary class including cost of room and board

• PC6: Details about student/faculty ratio

• PC7: Highlights estimated personal spending for a student and graduation rate

• PC8: Explains number of part-time undergraduate students and alumni who donate