**ADVANCED STATISTICS**

# Project Report

1. **PROJECT OBJECTIVE**

The objective of the report is to explore the data set for the 2 problem sets in Python and generate insights and recommendations for the same. This report will consist of the following:

* + Importing the dataset in python;
  + Understanding the structure of the data set;
  + Graphical exploration;
  + Descriptive statistics; and
  + Insights from the dataset and recommendations.

1. **ASSUMPTIONS**

We will make the following assumptions about the data set:

* Dependent variable to be measured at the continuous level
* There should be no significant outliers
* Dependent variable should be approximately normally distributed
* Randomization and replication of observations

1. **EXPLORATORY DATA ANALYSIS**

A Typical Data exploration activity consists of the following steps:

1. Environment Set up and Data Import;
2. Variable Identification
3. Univariate Analysis
4. Bi-Variate Analysis
5. Missing Value Treatment
6. Outlier Treatment
7. Variable Transformation / Feature Creation; and
8. Feature Exploration.

We shall follow these steps in exploring the provided datasets.

# (a) Environment Setup and Data Import

1. *Importing the libraries*

Before analysing the problem, the first step is to import all the required libraries such as, numpy, pandas, seaborn, matplotlib etc, based on our requirements. Doing this in the beginning helps the readability of the code

1. *Setting up Working Directory*

Setting up a working directory at the start of the Python session makes importing and exporting data files and code files easier. Basically, working directory is the location/ folder on the PC where you have the data, codes etc. related to the project. You can also upload the required files directly into jupyter notebook and import the required files from there

Since we have 2 different problems, we will look at each of them individually for all the steps, from exploratory data analysis, to addressing the questions, and drawing insights to conclude and give recommendations

1. **ANALYSIS OF PROBLEM 1**

# A research laboratory was developing a new compound for the relief of severe cases of hay fever. In an experiment with 36 volunteers, the amounts of the two active ingredients (A & B) in the compound were varied at three levels each. Randomization was used in assigning four volunteers to each of the nine treatments. The data on hours of relief can be found in the following .csv file: [Fever.csv](https://olympus.greatlearning.in/courses/7688/files/590457/download?verifier=OU735DhIvxi110mui1ZkWBICvrJpxrrORhugYiij&wrap=1)

# Import the Data Set

We use the command ‘read.csv’ to import the file in python– Fever.csv.

# Variable Identification

1. *Data view*

After importing the data, we use the df.head() function to view the data to see if the data has been imported properly.

A screenshot of a cell phone

Description automatically generated

1. *Checking the summary of the data*

We use the df.describe() and df.info function to view the summary of the data. This includes the 5 point summary and other specific details regarding the data. With the info function we can also view the data type and if there are any null values

A screenshot of a cell phone

Description automatically generated

1. *Checking for null values*

We use the df.isnull() function to check for any blank or NA cells in the data. By using the isnull function by itself we only get a boolean output. We use the df.isnull().sum() function to get an output of the count of null values in each column. This is helpful in cleaning the data. The rows with NA values can be removed and then one can move further with the analysis. The treat of blank cells will be looked at in different steps

# Univariate Analysis

We use the df.hist() function to plot a histogram for all the variables. This helps us to understand the distribution of each variable, whether it is normally distributed, skewed, or uniformly distributed.

A picture containing text

Description automatically generated

From the histograms we can note that variables A B and volunteers are categorical variables and hence the output is staggered histograms.

For Relief variable we can see that it is a continuous variable and its distribution is skewed

* 1. State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables ‘A’ and ‘B’ individually

Variable A:

Ho: µ1=µ2=µ3 (All population means are equal- The mean relief from each level of variation of ingredient A is the same)

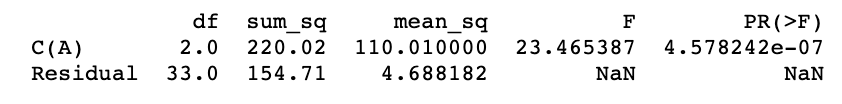
Ha: For at least one pair the population means are unequal (The mean relief for atleast one pair of level of variation of ingredient A is the same)

Variable B:

Ho: µ1=µ2=µ3 (All population means are equal - The mean relief from each level of variation of ingredient B is the same)

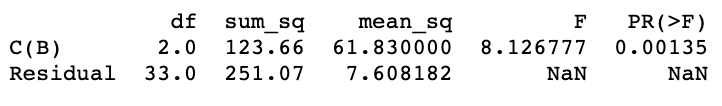
Ha: For at least one pair the population means are unequal ((The mean relief for atleast one pair of level of variation of ingredient B is the same)

* 1. Perform one-way ANOVA for variable ‘A’ with respect to the variable ‘Relief’. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.



Since the p value is less than the significance level of 0.05, we can reject the null hypothesis and state that there is a difference in the mean hours of relief provided by different levels of the amount of active ingredient A

* 1. Perform one-way ANOVA for variable ‘B’ with respect to the variable ‘Relief’. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.



Since the p value is less than the significance level of 0.05, we can reject the null hypothesis and states that there is a difference in the mean hours of relief provided by different levels of the amount of active ingredient B

* 1. Analyse the effects of one variable on another with the help of an interaction plot.  
     What is an interaction between two treatments? [hint: use the ‘pointplot’ function from the ‘seaborn’ function]

A close up of a map

Description automatically generated

With the help of the point plot we can see that there may be an interaction between both the active ingredients A and B leading to an effect on the Relief variable

By interaction we essentially mean that there is a combined effect of factors or treatments on the dependent variable. The influence of one treatment depends on the level of the other treatment

* 1. Perform a two-way ANOVA based on the different ingredients (variable ‘A’ & ‘B’) with the variable 'Relief' and state your results.

First we formulate our null and alternate hypothesis for each treatment –

Variable A:

Ho: µ1=µ2=µ3 (All population means are equal)

Ha: For at least one pair the population means are unequal

Variable B:

Ho: µ1=µ2=µ3 (All population means are equal)

Ha: For at least one pair the population means are unequal)

Interaction:

Ho: There is no interaction between ingredients A&B

Ha: There is interaction between ingredients A&B

A screenshot of a cell phone

Description automatically generated

On performing two way Anova with interaction we can observe that with 0.05 significant level, we will reject the null hypothesis in case of each variable and conclude the following –

1. Ingredient A does has a significant effect on the Relief variable
2. Ingredient B does has a significant effect on the Relief variable
3. There is interaction between A&B ingredients which has a significant effect on relief variable
   1. Mention the business implications of performing ANOVA for this particular case study.

Anova or analysis of variance is a statistical tool which helps to aggregate the variability found in a data set into two parts – systematic and random factors. Using anova analysts can determine the impact the independent variables have on the dependent variable is significant or not.

For our study, we are studying the impact of different levels of two types of ingredients on hours of relief. With the help of Anova analysts can determine the influence of these different levels of ingredients and as well as their interaction on the no. of hours of relied in order to produce a more effective drugs

1. **ANALYSIS OF PROBLEM 2**

# The dataset [Education - Post 12th Standard.csv](https://olympus.greatlearning.in/courses/7688/files/590456/download?verifier=G4gBVMTe1s9PgatcSHIbjqR9ZD44lQtZmuotxji5&wrap=1) is a dataset which contains the names of various colleges. This particular case study is based on various parameters of various institutions. You are expected to do Principal Component Analysis for this case study according to the instructions given in the following rubric. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: [Data Dictionary.xlsx](https://olympus.greatlearning.in/courses/7688/files/590455/download?verifier=XE7sljqOxA2aNPrxfVw75Ku7EE5r1RB13s80Kcgx&wrap=1)

# Import the Data Set

We use the command ‘read.csv’ to import the file Education-Post 12th Standard.csv into python

# Variable Identification

1. *Data view*

After importing the data, we use the df.head() function to view the data to see if the data has been imported properly.

A screenshot of a cell phone

Description automatically generated

1. *Checking the summary of the data*

We use the df.describe() and df.info function to view the summary of the data. This includes the 5-point summary and other specific details regarding the data. With the info function we can also view the data type and if there are any null values.

1. *Checking for null values and duplicate values*

We use the df.isnull() function to check for any blank or NA cells in the data. By using the isnull() function by itself we only get a Boolean output. We use the df.isnull().sum() function to get an output of the count of null values in each column. This is helpful in cleaning the data. The rows with NA values can be removed and then one can move further with the analysis. However we need to be mindful of cases when there are a large number of NA values and the treatment of the same needs to be carried out.

There are no null values in the data set.

We can also use the df.duplicated.sum() function to check if there are any duplicate values in the dataset and remove the same so that it does not hamper the outcome

2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented.

# Univariate Analysis

We use the df.hist() function to plot a histogram for all the variables. This helps us to understand the distribution of each variable, whether it is normally distributed, skewed, or uniformly distributed.

A close up of text on a white background

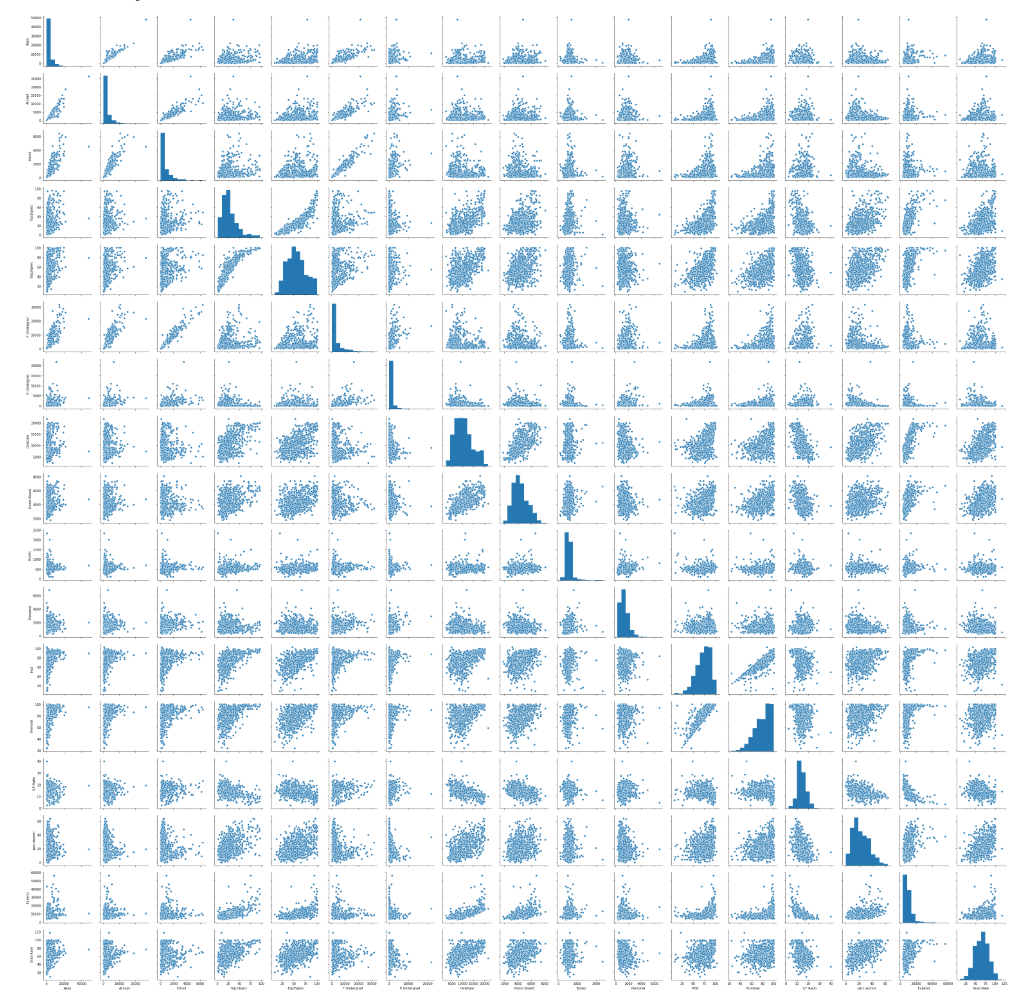
Description automatically generated

Interpretation of histogram

1. From the above histograms we can see that Accept, Apps, Books, Enroll, Expend,F.Undergrad, P.Undergrad, Personal are highly right skewed
2. Grad rate, Outstate and Top25perc seem to be more symmetrically or normally distributed
3. Terminal, PhD are highly left skewed
4. The remaining variables Room.Board, SF Ration, Top 10perc and perc.alumni are slightly skewed

# Bivariate Analysis

# We use the sns.pairplot() function to plot the relationship between all the numerical variables in the dataset. The histograms on the diagonal shows distribution of a each variable whereas the scatter plots on the upper and lower triangles show the relationship between two variables. We can see that all the plots show a linear relationship between the variables



We can also do the bivariate analysis to see the relationship between two variables using the heatmap to plot the correlation. We use the sns.heatmap() function to plot the same

A picture containing object

Description automatically generatedSimilar to the pair plot, the diagonal shows us the correlation between single variables whereas the triangles above and below the diagonals show us the correlation between each pair of 2 variables in the dataset

From the heatmap we can observe that Apps, Enroll, Accept and F.Undergrad are highly positively correlated

PhD and Terminal also have a high positive correlation of 0.8

2.2) Scale the variables and write the inference for using the type of scaling function for this case study.

To scale the variables we use the zscore function to help standardize the dataset’s feature onto a scale where the mean =0 and standard variance = 1. We know that the scale of a variable affects PCA hence we need to scale the variables before performing PCA. With the help of scaling, all variables are then on the same unit which helps to make the variables comparable with another and measure the magnitude during calculation of eigen vectors.

After scaling the data will be transformed wherein every value ranges from -1 to 1.

A screenshot of a cell phone

Description automatically generated

2.3 Comment on the comparison between covariance and the correlation matrix after scaling.

We use the df.cov() function to get the covariance matrix

**A close up of text on a white background

Description automatically generated**

Similarly we use the df. corr() function to get the correlation matrix

**A close up of a piece of paper

Description automatically generated**

The covariance matrix is a measure of how changes in one variable are associated with changes in another variable. It measures the degrees to which two variables are linearly associated. Similarly the correlation also tells us the extent to which two variables are associated in the data set. As we know, the formula of covariance assumes the units from the products of the units of the two variables considered. However, in case of correlation, it gives us a unit free relationship between the two variables.

In this case, both matrices are same as the data has already been scaled and standardized

2.4) Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.

We use the df.boxplot function to check for outliers

**A close up of a screen

Description automatically generated**

**We can observe that Top 25 perc does not have any outliers**

**A picture containing screenshot

Description automatically generated**

**Post scaling of the data we can observe that all the variables have outliers apart from Top25 perc**

2.5)  Build the covariance matrix, eigenvalues and eigenvector

We use the np.cov() function to building the covariance matrix. We then use the np.linalg.eig() function to calculate the eigen values and eigen vectors respectively.

An Eigen value is a number, telling us how much variance there is in the data in the direction of the Eigen vector. When Eigen value is 0, there is no variation at all.

Pls refer the appendix for the output

2.6) Write the explicit form of the first PC (in terms of Eigen Vectors).

The equation for first PC would be the cross multiplication of variables and the Eigen vectors of 0 index.

**Eigen vector for PC1:**

array([-0.2487656 ,0.33159823,0.0630921, -0.28131053,0.00574141,0.01623744,0.04248635,0.1030904 ,0.09022708, -0.0525098 ,0.3589704 , -0.4591395 , 0.04304621, -0.13340581, 0.0806328 ,-0.59583097, 0.02407091])

𝐏𝐂1:  -0.2487656Apps+0.33159823Accept+0.0630921Enroll-0.28131053Top10perc+0.00574141Top25perc+ 0.01623744F.Undergrad+0.04248635P.Undergrad+0.1030904Outstate+0.09022708Room.Board-0.0525098Books+0.3589704Personal-0.4591395PhD+0.04304621Terminal-0.13340581S.F.Ratio+0.0806328 perc.alumni-0.59583097Expend+ 0.02407091Grad.Rate

2.7) Discuss 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? Perform PCA and export the data of the Principal Component scores into a data frame.

After calculating the eigen values and vectors we will get 17 principal components as we have 17 variables. We will calculate the cumulative variance explained by these components

A close up of a logo

Description automatically generated

The first PC explains 32% of variance in the dataset. PC1 and PC2 explains 58% of the variance and so on.

We then plot the scree plot

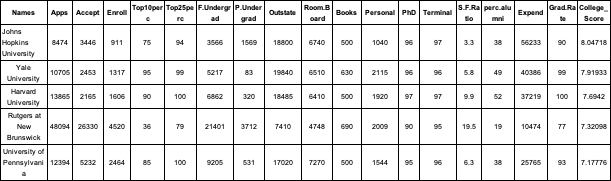
A screenshot of a cell phone

Description automatically generated

Visually we can observe that there is steep drop in variance explained with increase in number of PC's. 58.3% of the total variation is explained by first two PCs.

In the scree plot, the last big drop occurs between the second and third components and we choose the first 2 components

We then calculate the scores for PCA1 and result is following dataframe



2.8) Mention the business implication of using the Principal Component Analysis for this case study.

Principal component analysis is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. It is a tool that is used to reduce the dimensions of data while retaining most of the information. For the given dataset we have various parameters available for colleges. With the help of PCA, we can reduce the dimensions of the data, while retaining the information. With the help of PCA we ascertained that with the help of 2 principal components we can ascertain most of the variability in the data and with the help of PC1 we can calculate the college score which helps us to identify the top ranked college taking into consideration all the different parameters.