ASSIGNMENT NO. 1

AIM: Assignment of exploring data analysis.

PREREQUISITE: Statistics and Python programming

THEORY:

Exploratory data analysis or "EDA" is a critical first step in analyzing the data from an experiment. Here are the main reasons we use EDA: • detection of mistakes • checking of assumptions • preliminary selection of appropriate models • determining relationships among the explanatory variables, and • assessing the direction and rough size of relationships between explanatory and outcome variables. Loosely speaking, any method of looking at data that does not include formal statistical modeling and inference falls under the term exploratory data analysis.

- **Measure of Central tendency**

The central tendency or "location" of a distribution has to do with typical or middle values. The common, useful measures of central tendency are the statistics called (arithmetic) mean, median, and sometimes mode.

Means, such as geometric, harmonic, truncated, or Winsorized means, are used as measures of centrality. While most authors use the term "average" as a synonym for the arithmetic mean, some use average in a broader sense to also include geometric, harmonic, and other means. Assuming that we have n data values labeled x1 through x_n, the formula for calculating the sample (arithmetic) mean is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$

The median is another measure of central tendency. The sample median is the middle value after all of the values are put in an ordered list. If there are an even number of values, take the average of the two middle values.

- **Measure of variability**

Spread

Several statistics are commonly used as a measure of the spread of a distribution, including variance, standard deviation, and interquartile range. Spread is an indicator of how far away from the center we are still likely to find data values.

The **variance** is a standard measure of spread. It is calculated for a list of numbers, e.g., the n observations of a particular measurement labeled x1 through xn, based on the n sample deviations (or just "deviations"). Then for any data value, xi , the corresponding deviation is (xi − x¯), which is the signed (- for lower and + for higher) distance of the data value from the mean of all of the n data values. It is not hard to prove that the sum of all of the deviations of a sample is zero. The variance of a population is defined as the mean squared deviation (see section 3.5.2). The sample formula for the variance of observed data conventionally has n−1 in the denominator instead of n to achieve the property of "unbiasedness", which roughly means that when calculated for many different random samples from the same population, the average should match the corresponding population quantity (here, σ 2). The most commonly used symbol for sample variance is s 2 , and the formula is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}$$

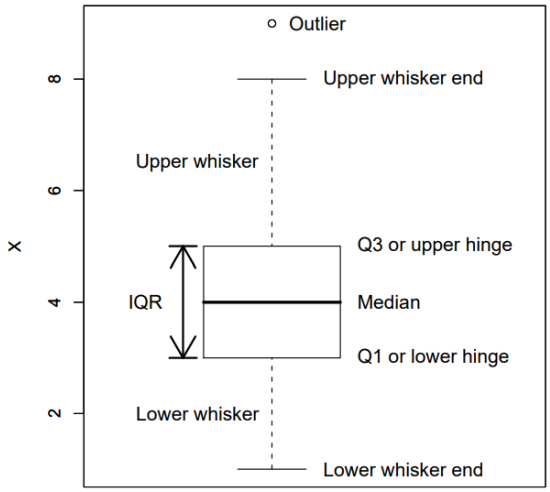
which is essentially the average of the squared deviations, except for dividing by n − 1 instead of n. This is a measure of spread, because the bigger the deviations from the mean, the bigger the variance gets.

The **standard deviation** is simply the square root of the variance. Therefore it has the same units as the original data, which helps make it more interpretable. The sample standard deviation is usually represented by the symbol s. For a theoretical Gaussian distribution, we learned in the previous chapter that mean plus or minus 1, 2 or 3 standard deviations holds 68.3, 95.4 and 99.7% of the probability respectively, and this should be approximately true for real data from a Normal distribution.

A third measure of spread is the **interquartile range**. To define IQR, we first need to define the concepts of quartiles. The quartiles of a population or a sample are the three values which divide the distribution or observed data into even fourths. So one quarter of the data fall below the first quartile, usually written Q1; one half fall below the second quartile (Q2); and three fourths fall below the third quartile (Q3). The astute reader will realize that half of the values fall above Q2, one quarter fall above Q3, and also that Q2 is a synonym for the median. Once the quartiles are defined, it is easy to define the IQR as IQR = Q3 − Q1. By definition, half of the values (and specifically the middle half) fall within an interval whose width equals the IQR. If the data are more spread out, then the IQR tends to increase, and vice versa.

Outlier Identification

Boxplots Another very useful univariate graphical technique is the boxplot. The boxplot will be described here in its vertical format, which is the most common, but a horizontal format also is possible. An example of a boxplot is shown in the following figure, which again represents the



data in EDA1.dat.

Here you can see that the boxplot consists of a rectangular box bounded above and below by "hinges" that represent the quartiles Q3 and Q1, respectively, and with a horizontal "median" line through it. You can also see the upper and lower "whiskers", and a point marking an "outlier". The vertical axis is in the units of the quantitative variable.

REFERENCES:

1. Coursera Course on "What is Data Science?" offered by IBM. Available at https://www.coursera.org/learn/what-is-datascience?specialization=ibm-data-science

2. Getting Started with Business Analytics: Insightful Decision-Making, David Roi Hardoon, Galit Shmueli, CRC Press

CONCLUSION:

Exploratory Data Analysis (EDA) is a crucial first step in analyzing data, helping identify patterns, relationships, and potential mistakes in the dataset. It involves various techniques such as calculating measures of central tendency (mean, median), measures of variability (variance, standard deviation, interquartile range), and detecting outliers using methods like boxplots. EDA guides further analysis by suggesting appropriate models and offering insights into the data's structure, making it an essential process in the data science workflow. Understanding and applying these concepts is key to extracting meaningful insights and making informed decisions based on the data.