Group A

Assignment No: 3

Title of the Assignment: Descriptive stastics - Measures of central tendency and variability.

perform the following operations on any open source dataset (e.g data, csv)

1. Provide summary stastics (mean, median, minimum, maximum, standard deviation) for a dataset (age, income etc) with numeric variables grouped by of the qualitative (categorial) variables. for example. if your categorial variable is age groups quantitative variable is income. then provide summary static of income grouped by the age groups.
2. write a python program to display some basic stastical details like percentile, meen standard deviation etc

objective of the assignment: students should be able to perform the stastical operations using python on any open source dataset.

prerequisite.

1. Basic of python programming
2. concept of stastistics such as mean, median minimum. maximum, standard deviation etc

concepts for Theory:

1. summary stastistics
2. Types of variables.

3. summary ~~Stastis~~ Statistics of income grouped by age groups.
4. Display basic sta~~tis~~tistical details on the iris dataset.

1. Summary statistics:
what is statistics?
statistics is the science of collecting data and analysing them to infer proportions (sample) that are representative of the population.

Branches of ~~statistics~~: statistics:
There are two branches of statistics.
DESCRIPTIVE STATISTICS: Descriptive statistics is statistics or measure that describes the data.

Descriptive statistics:
Descriptive statistics is summarising the data at hand through certain numbers like mean, median etc. so as to make the understanding of the data easier. It does not involve any generalisation of inference beyond what is available.

Commonly used Measures
1. Measures of central Tendency
2. Measures of Dispersion (or variability)

- Measures of central Tendency
A measure of central tendency is one number summary of the data typically describes the centre of the Data

a] Mean: Mean is defined as the ratio of the sum of all the observations, in the data to the total number of observation.

b] Median: Median is the Point which devides the entire data into two equal halves. one-half of the data is less than the median is given band other half is greater than the Same.

c] Mode: Mode is the number - which has the maximum frequency in the entire dataset or in the other words, mode is the number that appears the maximum number of the times.

- If there is only one number that appears the maximum number of times. the data has one mode and is called unimodel.

- Measures of Dispersion (or variability)

measures of Dispersion describes the spread of the data around the central value for measures of central Tendency)

1. Absolute Deviation from Mean - The absolute Deviation from mean, also called mean absolute deviation (MAD), describes the variation in dataset.

$$\text{Mean Absolute Deviation} = \frac{1}{N} \sum_{i=1}^{N} | X_i - \bar{X} |$$

2) variance- Variance measures how far are data points spread out from the mean

$$\text{Variance} = \frac{1}{N}\sum_{i=1}^{x}(x_i - \bar{x})^2$$

3) standard Deviation- The square root of variance is called the standard Deviation it i's calculated as

$$\text{std Deviation} = \sqrt{\text{variance}} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

4] skewness- The measures of a symmetry in a probability distribution is defined by skewness. It can either be Positive negative or undefined

$$\text{Skewness} = \frac{3(\text{mean} - \text{median})}{\text{Std Deviation}}$$

2. Types of Variables:

A variable is a characteristics that can be measured and the can assume different values. Height, Age, income, Province of country of birth, grades obtained at school. and type of housing are all examples of variables.

Variables may be classified into two main categories

- categorical and
- Numeric

- Categorial Variable-

A categorical Variable (also called qualitative variables) refers to characteristics that can't be quantifiable.

- Norminal Variable-

  A norminal variable is onethat describes a name, label or category without natural order in the given table.

- Ordinal Variable -

  An ordinal Variable is a variable whose Values are defined by an order relation between the different categories.

- Numerical variables-

  A numeric variable(also called quantitative variable) is a quantifiable characteristics whose values are numbers.

- Continuous variables-

  A variable is said to be continuous it is can assume an infinite number of real values within a given interval.

  for instance, consider the height of student, The height can't take any values.

- Discreate variables:

  A Variable is said to be continuous if it can assume ony infinite number of real values within a given interval.

3 summary statistics of income grouped by the age groups:

problem statement: for example, if your categorical variable is age groups and quantitative variable is income. then provide Summary statistics income grouped by the age groups. Create a list that contains a numerical value for each response to the Categorial variable.

Categorial Variable : Genre
Quantitative variable: Age

Conclusion:

Descriptive statistics summarises or describes the characteristics of data set. Descriptive stastics consist of two basic categories of measures

- measures of central tendency and
- measures of variability (or spread)

Measures of central tendency describe the centre of data set. It includes the mean, median and mode.

Measures of variability or spread describe the dispersion of data within the set and it includes standard deviation, varience, minimum and maximum variables.