DSBDAL Assignment 3

- Name :- Jash Shah
- Roll no :- 31457
- Completion date :- 1/02/22
- Submission date :- 1/02/22
- Title :- Descriptive statistics - Measure of central tendency
- Problem statement :-
  Perform the following operations on any open-source dataset.
  1. Provide summary statistics for a dataset with numeric variables grouped by one of the qualitative variable.
  2. Write a python program to display some basic statistical details like percentile, mean, std deviation etc. of the species of 'Iris-setosa', 'Iris-versicolor', and 'Iris-Verginica' of iris dataset
- Learning objectives :-
  1. We will be able to load dataset and perform all basic preprocessing on the dataset.
  2. We will be able to group data and find the grouped statistics for a variable.
- Learning Outcomes :-
  1. Learnt about grouping data according to a catagory and find basic ~~statistics~~ statistics of the grouped data.
  2. Learnt about mean, median, std deviation percentile etc.

- Theory :-

- ~~Stastics~~

1) Statistics :-

It is the science of collecting data and analyzing them to infer proportions (sample) that are representative of the population. In other words, ~~stastc~~ statistics is interpreting data in order to make predictions for the population.

- Branches of statistics

There are two branches of statistics :-

- Descriptive statistics.
- Inferential statistics.

Descriptive statistics :-

It is summarizing 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 generalization or inference beyond what is available. This means that the descriptive statistics are just the representation of the data available and not based on any theory of probability.

- Commonly used Measures

1. Measures of Central Tendency
2. Measure of Dispersion (or Variability)

Page No.
Date.

- Measures of central Tendency:

A measure of central tendency is a one number summary of the data that typically describes the center of the data. These one number summary is of three types :-

1. Mean :- It is defined as the ratio of the sum of all the observations. This is also known as Average.

2. Median :- It is the point which divides the entire data into two equal halves. one half of the data is less than the median, and the other half is greater then the same.

3. Mode :- It is the number which has the maximum frequency in the entire data set.

- Measures of Dispersion (or Variability)

Measures of Dispersion describes the spread of the data around the central value.

1. Absolute Deviation from Mean :- It is also called as Mean absolute deviation, describes the variation in the data set, in sense that it tells the average absolute distance of each data point in the set.

$$MAD = \frac{1}{N} \sum_{i=1}^{N} |X_i - \bar{X}|$$

2. Variance :- It measures how far are data points spread out from the mean.

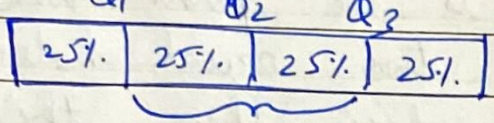
$$\text{Variance} : \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

Page No.
Date.

3. Standard Deviation :- The square root of variance is called the Standard Deviation.

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

4. Quartiles :- Quartiles are the points in the data set that divides the data set into 4 equal parts. Q_1, Q_2, Q_3 are the 1st, 2nd and 3rd quartile of data set.



Interquartile range
$= Q_3 - Q_1$

2) Pandas groupby :-
It is used for grouping the data according to the categories. It also helps to aggregate data efficiently.

- Steps and Functions :-

1. Data Preprocessing :- Process of transforming raw data into an understandable format.
Functions :- 1) read_csv() 2) head() 3) isnull()
4) describe() 5) fillna 6) astype 7) value_counts.

2. Calculating grouped statistics
Part 1 :-
1. Age is converted to age groups by binning method.

2. grouped the data according to the age group ~~and~~ & using groupby() function. And then calculated the statistics like mean, median related to salary using describe() function

Part 2:-

1. The data is grouped according to the 3 species using groupby and then found out all basic ~~stastics~~ statistics.

- Conclusion:- Successfully calculated the ~~its~~ measures of central tendency and measure of dispersion for grouped data for both the datasets