DOP :- 25/01/2022 DOS :- 25/01/2022

Assignment no 3

* **Title :-** Descriptive Statistics - Measures of Central Tendency and variability.

* **Problem Statement :**

Perform the following operations on any open source dataset (eg data.csv)

① Provide summary statistics (mean, median, minimum, maximum, standard deviation) for a dataset (age, income etc) with numeric variables grouped by one of the qualitative (categorical) variable. For example if your categorical value is age groups and quantitative variable is income, then provide summary statistics of income grouped by the age groups. Create a list that contains a numeric value for each response to the categorical variable.

② Write a Python program to display some basic statistical details like percentile, mean, standard deviation etc. of the species of 'Iris-setosa' and 'Iris-versicolor' and 'Iris-versicolor' of iris.csv dataset.

* **Learning Objective**

① To understand the measures of central-Tendency like std deviation, mean, mode, median, etc.

* Learning Outcomes:

After performing this assignment one should able to

- ① To implement the measures of central tendency and variability and find some conclusion on the basis of statistics available.

* Software Requirements:

- Anaconda Navigator
- Jupyter Notebook
- Python 3.8

* Hardware Requirements:-

- Windows 10, 8GB RAM
- Intel i5 processor, 64 bit OS

* Theory:

* Measures of central Tendency & Variability:-

A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.

There are three main measures of central tendency: the mode, the median and the mean. Each of these measures describes a different indication of the typical or central value in the distribution.

① The Mode :-

The mode is the most commonly occuring value in a distribution

Eg: 54, 54, 54, 55, 56, 57, 59

Mode = 54 because its frequency is highest.

② The Median :-

The Median is the middle value in distribution when the values are arranged in ascending or descending order.

The Median divides the distribution in half (there are 50% of observations on either side of the median value)

Eg:- 52, 53, 54, 55, 56, 57, 58

Median = 55.

③ The Mean :-

The mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as "arithmetic average".

Eg:- 10, 11, 25, 30, 40

$$\text{Mean} = \frac{10+11+25+30+40}{5} = \underline{\underline{28.2}}$$

Mean = 28.2.

* Methods and Functions used :-

① Pandas.read_csv()

Read a comma seperated values (csv) file into a Dataframe.

Also supports optionally iterating or breaking of the file into chunks.

② dataframe.head(limit)

Returns the first 5 rows of dataset by default. To override a default we use the limit to get that number of rows.

Eg: df.head(6)

Returns first 6 rows of dataset.

③ dataframe.tail(limit)

Returns the last 5 rows of dataset by default. To override a default we use the limit to get that number of rows.

Eg :- df.tail(3)

Returns last 3 rows from dataset.

④ df.shape() or dataframe.shape

Returns a tuple representing the dimensions i.e no. of rows and no. of colums present in dataset.

Eg : df.shape

(13560, 21).

⑤ dataframe.dtypes

It returns the datatypes of every column present in dataset.

⑥ df.info() or dataframe.info()

Print a concise summary of a Dataframe. This method prints information about a dataframe including the index dtype and columns, non-null values and memory usage.

⑦ dataframe.describe()

Provides descriptive statistics that summarize the central tendency, dispersion and shape.

⑧ dataframe.isnull()

Returns dataframe with value 'True' where it finds Null values and 'False' where it encounters any type of value.

⑨ df.groupby ( ) :

Group Dataframe using a mapper or by a series of columns.

A groupby operation involves some combination of splitting the object, applying a function, and combining the results. This can be used to group large amounts of data and compute operations on these groups.

Syntax:- If we have to find the mean by grouping a gender

df.groupby ('gender') [column_name].mean ( ).

⑩ Scatter plot :-

Dataframe.plot.scatter (x, y, s = None, c = None)

The coordinates of each point are defined by two dataframe and filled circles are used to represent each point. This kind of plot is useful to see complex correlations between two variables. Points could be for instance natural 2D coordinates like longitude and latitude in a map or, in general, any pair of metrics that can be plotted against each other.

Syntax: import matplotlib.pyplot as plt

plt.scatter (x, y, marker, label).

* Packages / Module / Libraries used.

① Pandas :

Pandas is a software library written for the python programming language for data manipulation and analysis. In particular it offers data structures and operation for manipulating numerical tables and times series. It is free software release under the three clause BSD License. The name is derived from the term "Panel data" an econometrics terms for data sets that include observations over multiple time periods for the same individuals.

Syntax :- import pandas as pd.

② Numpy :-

Numpy is a python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform and matrices. Numpy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely. Numpy stands for Numerical Python.

Syntax : import numpy as np.

③ Scipy :-

Scipy is a free and open-source Python library used for scientific computing and technical computing. Scipy contains modules for optimization linear algebra, integration, interpolation, special functions, FFT, signal and image processing.
Syntax : Import scipy.stats as stats.

④ Matplotlib :-

Matplotlib is a plotting library for the python programming language and its Numerical mathematics extension Numpy. matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB.
Syntax : import matplotlib.pyplot as plt.

⑤ Pylab :

Pylab is a module that provides a Matlab like namespace by importing functions from the modules Numpy and Matplotlib.
Syntax : import pylab.

About Dataset Used :-

① covid_19_india.csv

Dataset consist of Date, Time, State/Union Territory, Confirmed Indian National, Confirmed Foreign National, Cured, Deaths, Confirmed Columns.

② StudentPerformance.csv

Dataset consist of gender, race/ethnicity, parental level of education, lunch, test preparation course, math score, reading score, writing score.

③ Iris.csv,

Dataset consist of Sepealength, Petallength, Sepal width, Petal width and species (setosa, virginica, versicolor).

Analysis / Observations :-

From the measure of central tendencies like mean we get the exact average no. of deaths per year month in specific state / Union territory from covid_19 dataset.

From the StudentPerformance we ~~~~ get the mean of math scores of ~~team~~ female and male and according to that we predict which gender students have to practice more.

From the iris dataset we can predict the type of species from mean sepallength or width Also it is observed that if petal length is bet^n 1-2 then it is setosa, 3-5 it is versicolor & 5-6 virginica.

Conclusion :-

From this assignment we learnt the measures of central tendency, how it is beneficial to predict some meaningful results and some visualization techniques.