PICT, PUNE

BSBDA Lab

Assignment 03

31/18

Title: Descriptive Statistics - measures of central tendency & variability

Problem Statement:

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

i) Provide summary statistics (mean, median, min, max, standard deviation) for a dataset (age, income etc) with numeric variables grouped by one of the qualitative (categorical) variable.

For eg. if your categorical variable is age groups & 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.

ii) Write a python program to display some basic statistical details like percentile, mean, standard deviation etc of the species of "Iris-setosa", "Iris-versicolor" & "Iris-versicolor" of iris.csv dataset.

Provide the codes with outputs & explain everything that you do in this step.

Learning Objectives:

After performing this assignment one should able to implement the measures of central tendency & variability & find some conclusion on the basis of statistical variable.

PICT, PUNE

Software requirements:
- Jupyter notebook,
- python interpreter.

Hardware requirements:
- Windows 10, 8GB RAM
- Intel i5 - 8 Gen (4-core) processor.

Theory:

Measure of central tendency & variability:
1) It is also refered as mesure of centre @ central location.
2) It is a summary measure that attempts to describe a whole set of data with a single value that represents the middle @ centre of its distribution.
3) There are three main measures of central tendency
   i) mode - most common occuring value
   ii) median - midde value in the 'ordered' distribution
   iii) mean - it is sum of values divided by number of values in dataset.

Methods & functions Used:
1) Pandas. read_csv().
   read the csv file into data frame. It also supports optionally iterating / breaking of the file into chunks

PICT, PUNE

2) dataframe.head(limit): returns first limit rows of dataset (by default it returns 5).
eg df.head() // returns first 5
df.head(7) // return first 7.

3) dataframe.tail(limit): same as head but from end of the dataframe.

4) df.shape() : returns a tuple representing the dimensions i.e. no. of rows & columns of dataset.
eg df.shape()
(30, 25)

5) dataframe.dtypes: returns the datatypes of every column present in dataset.

6) dataframe.info(): It prints concise summary of dataframe, including index dtype & non-null values & memory usage.

7) dataframe.describe(): provides description statistics that summerizes the central tendency, dispersion & shape.

8) dataframe.isnull(): return dataframe with value "true" where it find null values & "false" when it encounters any type of data.

9) dataframe.groupby(): It groups dataframe using a mapped/ by a series of columns.
A grouping operation involves some combination of splitting the object, applying a function, & combining the results.
This can be used to group large amounts of data & compute operations on these groups.

PICT, PUNE

syntax: If we have to find a mean by grouping a gender

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

5) Scatter Plot:

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

The co-ordinates of each point are defined by two dataframe & 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 & latitude in a map or in general, any pair of metrices that can be plotted against each other.

syntax:

import matplotlib.pyplot as plt
plt.scatter(x,y, marker, label)

Packages / Modules / libraries Used:

1) Pandas:

i) Used for data manipulation & analysis
ii) free software released under the three clause BSD License.

syntax import pandas.

2) Numpy

i) Used for working with arrays.
ii) Also has a functions for working in domain of linear algebra, fourier transform & matrices.

PICT, PUNE

iii) It is open source project.
Syntax: import numpy as np

3) Scipy:
i) It is free & open source python library
ii) used for scientific & technical computing.
iii) contains modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal & image processing
Syntax:
import scipy.stats as stas

4) Matplotlib:
i) It is a plotting library for python
ii) It is a collection of functions that make matplotlib works like MATLAB.

5) Pylab:
i) It is a module that provides a matlab like namespace by importing functions from modules numpy & matplotlib
Syntax: import pylab

About datasets Used:

1) covid_19_india.csv: this consists of data, time, State/union territory, confirmed IndianNational, confirmed foreignNational, Cured, Deaths, Confirmed columns.

2) StudentPerformance.csv. dataset consists of gender.

PICT, PUNE

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

3) Iris.csv - dataset consists of sepal.length, petallength sepalwidth, petalwidth. & species (setosa, virginia, versicolor).

Analysis / Observations:

1) From the measure of central tendencies like ~~mean~~ mean we get the exact average no. of deaths per month in specific state/union territory from covid_19_dataset.
2) From student performance we get the mean of math scores of female & male and according to that we predict which gender students have to practice more.
3) From the iris dataset we can predict the type of species from mean sepallength/width.

Conclusion:

From this assigment we learned the measures of central tendency, how it is beneficial to predict some meaningful results & some visualization techniques.

✓.