Sourav Kotkar
TE-3 (N3)
31378

Performance Date - 25/01/2022
Submission Date - 25/01/2022

Page No.:
Date:

DSBDAL

Group - A : Assignment - 3

* **Title :-**
Descriptive Statistics - Measures of Central Tendency and Variability

* **Problem Statement :-**
Perform the following operations on any open-source dataset.
1. Perform 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 variable 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', 'Iris-versicolor' and 'Iris-virginica' of iris.csv dataset.

* **Learning Objectives :-**
To understand the measures of central tendency and variability like mean, median, mode, minimum, maximum, standard deviation, etc.

* **Learning Outcomes :-**
After completion of this assignment, student will be able to implement the measures of central tendency and variability on a dataset using Python programming language.

* **S/W and H/W requirements :-**
Python 3.9, Jupyter notebook, Windows 10, 8GB RAM laptop

* Theory :-

⇒ Central Tendency :-

In statistics, the central tendency is the descriptive summary of a data set. Through the single value from the dataset, it reflects the centre of the data distribution. Moreover, it does not provide information regarding individual data from the dataset, where it gives a summary of the dataset. Generally, the central tendency of a dataset can be defined using some of the measures in statistics.

Measures of Central Tendency :-

1) Mean
2) Median
3) Mode

⇒ Mean :-

The mean represents the average value of the dataset. It can be calculated as the sum of all the values in the dataset divided by the number of values. The formula to calculate the mean value is given as : $\frac{x_1 + x_2 + \ldots + x_n}{n}$

E.g. : Data = 5, 4, 2, 2, 3, 1, 5, 4, 5

$$\text{Mean} = \frac{5+4+2+2+3+1+5+4+5}{9}$$

$$= \frac{31}{9}$$

$$= 3.44$$

⇒ **Median :-**

Median is the middle value of the dataset in which the dataset is arranged in the ascending order or in descending order. When the dataset contains an even number of values, then the median value of the dataset can be found by taking the mean of the middle two values.

E.g. : Data = 23, 21, 18, 16, 15, 13, 12, 10, 9, 7, 6, 5, 2
Median = 12

⇒ **Mode :-**

The mode represents the frequently occuring value in the dataset. Sometimes the dataset may contain multiple modes and in some cases, it does not contain any mode at all.

E.g. : Data = 5, 4, 2, 3, 2, 1, 5, 4, 5
Mode = 5

⇒ **Variability :-**

Variability describes how far apart data points lie from each other and from the center of a distribution. Variability is also referred to as spread, scatter or dispersion.

Measures of Variability :-

1) Range
2) Interquartile range
3) Standard deviation
4) Variance

⇒ Range :-

The range tells you the spread of your data from the lowest to the highest value in the distribution.
To find the range, simply subtract the lowest value from the highest value in the dataset.

⇒ Interquartile Range :-

It gives you the spread of the middle of your distribution.
For any distribution, that's ordered from low to high, the IQR contains half of the values. While the first quartile (Q1) contains the first 25% of values, the fourth quartile (Q4) contains the last 25% of values.
IQR = Q3 - Q1

⇒ Standard Deviation :-

It is the average amount of variability in the dataset.
It tells, on average, how far each score lies from the mean. The larger the standard deviation, the more variable the dataset is.

Standard deviation, $\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$

where σ = std. deviation
X = each value
μ = mean
N = no. of values

⇒ Variance :-

It is the average of squared deviations from the mean. Variance is the square of the standard deviation.

$$\text{Variance} = \sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$$

→ Methods & Functions used :-

1) read-csv() - It is used to read a csv file into a DataFrame.

2) isnull() - It returns a boolean same-sized object indicating if the values are NA.

3) dtypes - It returns a series with the datatype of each column.

4) describe() - It gives descriptive statistics of the dataset.

5) groupby () - It involves some combination of splitting the object, applying a function, and combining the results.

6) min () - Returns the minimum of the values over the requested axis.

7) max () - Returns the maximum of the values over the requested axis.

8) mean () - Returns the mean of the values over the requested axis.

9) median () - Returns the median of the values over the requested axis.

10) std() - Returns sample standard deviation over requested axis.

11) value-counts () - Return a series containing counts of unique values.

12) agg() - Aggregate using one or more operations over the specified axis.

⇒ Packages / Libraries Used :-

1> Pandas :-
Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.

2> Matplotlib :-
Matplotlib is a low level graph plotting library in python that serves as a visualization utility.

3> Seaborn :-
Seaborn is a library that uses Matplotlib underneath to plot graphs. It is used to visualize random distributions.

⇒ Plots Used :-

1> Histplot :-
A histogram is a classic visualization tool that represents the distribution of one or more variables by counting the number of observations that fall with discrete bins. This function can normalize the statistic computed within each bin to estimate frequency, density or probability mass.

2> Facet Grid with Scatterplot :-
It is used to initialize the matplotlib figure and FacetGrid object.
A scatterplot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data.

* Observations :-

1) Mall Customers Dataset :-

- The dataset consists of attributes such as Customer ID, Gender, Age, Annual Income (k $), Spending Score (1-100).
- The dataset can be grouped by Gender which is the categorical variable in the dataset.
- We get the statistical summary of age, annual income & spending score of people grouped by gender (male / female).

2) Iris dataset :-

- The dataset includes three species of flowers with attributes of sepal length, sepal width, petal length, petal width.
- The dataset can be grouped by species (Iris-setosa, Iris-versicolor, Iris-virginica) which is the categorical variable in the dataset.
- From the dataset, we can predict that Iris-setosa has petal length of 1-2 cm & petal width of 0-0.75 cm, Iris-versicolor has petal length of 3-5 cm & petal width of 1-1.75 cm, Iris-virginica has petal length of 4.5-7 cm & petal width of 1.5-2.5 cm.

* Conclusion :-

Through this assignment, we learnt and implemented measures of central tendency and variability for mall customers and iris dataset using Python programming language.