Assignment No. 03

Aim: Descriptive Statistics - Measures of Central Tendency and variability Perform the following operations on any open source dataset (e.g., 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 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-versicolor' of iris.csv dataset.

Prerequisites:
1. Basics of Python programming.
2. Google Colab

Objectives: To learn the concept of how to display summary statistics for each feature available in the dataset. Implement a dataset into a dataframe.
Implement the following operations:
1. Display data set details.
2. Calculate min, max ,mean, range, standard deviation, variance.
3. Create histograph using hist function.
4. Create boxplot using boxplot function.

Theory:
How to Find the Mean, Median, Mode, Range, and Standard Deviation ?
Simplify comparisons of sets of number, especially large sets of number, by calculating the center values using mean, mode and median. Use the ranges and standard deviations of the sets to examine the variability of data.
Calculating Mean
The mean identifies the average value of the set of numbers. For example, consider the data set containing the values 20, 24, 25, 36, 25, 22, 23.
Formula
To find the mean, use the formula: Mean equals the sum of the numbers in the data set divided by the number of values in the data set. In mathematical terms: Mean=(sum of all terms)÷(how many terms or values in the set).
Adding Data Set
Add the numbers in the example data set: 20+24+25+36+25+22+23=175.
Finding Divisor
Divide by the number of data points in the set. This set has seven values so divide by 7.
Finding Mean
Insert the values into the formula to calculate the mean. The mean equals the sum of the values (175) divided by the number of data points (7). Since 175÷7=25, the mean of this data set equals 25. Not all mean values will equal a whole number.
Calculating Range
Range shows the mathematical distance between the lowest and highest values in the data set. Range measures the variability of the data set. A wide range indicates greater variability in the data, or perhaps a single outlier far from the rest of the data. Outliers may skew, or shift, the mean value enough to impact data analysis.
Identifying Low and High Values
In the sample group, the lowest value is 20 and the highest value is 36.
Calculating Range

To calculate range, subtract the lowest value from the highest value. Since 36-20=16, the range equals 16.

Calculating Standard Deviation

Standard deviation measures the variability of the data set. Like range, a smaller standard deviation indicates less variability.

Formula

Finding standard deviation requires summing the squared difference between each data point and the mean [$\sum(x-\mu)^2$], adding all the squares, dividing that sum by one less than the number of values (N-1), and finally calculating the square root of the dividend.

Mathematically, start with calculating the mean.

Calculating the Mean

Calculate the mean by adding all the data point values, then dividing by the number of data points. In the sample data set, 20+24+25+36+25+22+23=175. Divide the sum, 175, by the number of data points, 7, or 175÷7=25. The mean equals 25.

Squaring the Difference

Next, subtract the mean from each data point, then square each difference. The formula looks like this: $\sum(x-\mu)^2$, where $\sum$ means sum, x represents each data set value and μ represents the mean value. Continuing with the example set, the values become: 20-25=-5 and $-5^2=25$; 24-25=-1 and $-1^2=1$; 25-25=0 and $0^2=0$; 36-25=11 and $11^2=121$; 25-25=0 and $0^2=0$; 22-25=-3 and $-3^2=9$; and 23-25=-2 and $-2^2=4$.

Adding the Squared Differences

Adding the squared differences yields: 25+1+0+121+0+9+4=160. The example data set has 7 values, so N-1 equals 7-1=6. The sum of the squared differences, 160, divided by 6 equals approximately 26.6667.

Standard Deviation

Calculate the standard deviation by finding the square root of the division by N-1. In the example, the square root of 26.6667 equals approximately 5.164. Therefore, the standard deviation equals approximately 5.164.

Evaluating Standard Deviation

Standard deviation helps evaluate data. Numbers in the data set that fall within one standard deviation of the mean are part of the data set. Numbers that fall outside of two standard deviations are extreme values or outliers. In the example set, the value 36 lies more than two standard deviations from the mean, so 36 is an outlier. Outliers may represent erroneous data or may suggest unforeseen circumstances and should be carefully considered when interpreting data.

Facilities: Windows/Linux Operating Systems, Python

Application:

1. The histogram is suitable for visualizing distribution of numerical data over a continuous interval, or a certain time period. The histogram organizes large amounts of data, and produces visualization quickly, using a single dimension.
2. The box plot allows quick graphical examination of one or more data sets. Box plots may seem more primitive than a histogram but they do have some advantages. They take up less space and are therefore particularly useful for comparing distributions between several groups or sets of data. Choice of number and width of bins techniques can heavily influence the appearance of a histogram, and choice of bandwidth can heavily influence the appearance of a kernel density estimate.
3. Data Visualization Application lets you quickly create insightful data visualizations, in minutes. Data visualization tools allow anyone to organize and present information intuitively. They enables users to share data visualizations with others.

Input:

Structured Dataset: Iris

Dataset File: iris.csv

Output:
1. Display Dataset Details.
2. Calculate Min, Max, Mean, Variance value and Percentiles of probabilities also Display Specific use quantile.
3. Display the Histogram using Hist Function.
4. Display the Boxplot using Boxplot Function.

Conclusion:
Hence, we have studied using dataset into a dataframe and compare distribution and identify outliers.

Questions:
1. What is Data visualization?
2. How to calculate min,max,range and standard deviation?
3. How to create boxplot for each feature in the daset?
4. How to create histogram?
5. What is dataset?