Assignment No. 2

Problem Statement: DESCRIPTIVE STATISTICS: Write a python script to find basic descriptive statistics using summary, quartile function, etc on iris datasets.

Objective: The objective is to compute basic descriptive statistics on the Iris dataset, including measures like mean, median, and quartiles. This helps summarize the dataset and identify patterns or outliers for further analysis.

Prerequisite :

1. A Python environment set up with libraries like pandas, xml.etree.ElementTree, and requests (for web access).
2. Internet connection (for reading datasets from the web).
3. Text editor and basic knowledge of file handling in Python.

Theory :

Descriptive Statistics:

Descriptive statistics are essential in summarizing and providing insight into a dataset's main characteristics. They are used to describe the central tendency, dispersion, and overall distribution of the data, making it easier to interpret large datasets. These statistics give us a clear overview of the dataset without delving into more complex analysis.

Key Descriptive Statistics:

1. **Mean (Average):** The mean is the sum of all data points divided by the number of data points. It represents the central value of the data but can be sensitive to outliers. The formula is:
   Mean=∑Xn\text{Mean} = \frac{\sum X}{n}Mean=n∑X
   where XXX represents the data points and nnn is the total number of data points.
2. **Median:** The median is the middle value when the data is arranged in ascending order. It is a robust measure of central tendency because it is not affected by extreme values. If the number of data points is odd, the median is the middle value; if even, it's the average of the two middle values.
3. **Mode:** The mode is the most frequently occurring value in the dataset. While some datasets may have no mode, others may have more than one, known as bimodal or multimodal distributions.

4. **Standard Deviation (SD):** Standard deviation measures the amount of variation or spread in a dataset. A low standard deviation indicates that the data points tend to be close to the mean, while a high standard deviation means the data points are spread out. The formula for standard deviation is:
   SD=1n∑(Xi−Mean)2\text{SD} = \sqrt{\frac{1}{n}\sum (X_i - \text{Mean})^2}SD=n1∑(Xi−Mean)2
   where XiX_iXi is each data point, and the mean is the average of the dataset.
5. **Variance:** Variance measures how much the data points differ from the mean. It is the square of the standard deviation, providing insights into the degree of spread within the data. A larger variance indicates more spread. The formula is:
   Variance=∑(Xi−Mean)2n\text{Variance} = \frac{\sum (X_i - \text{Mean})^2}{n}Variance=n∑(Xi−Mean)2
6. **Range:** The range gives the difference between the maximum and minimum values in the dataset:
   Range=Max Value−Min Value\text{Range} = \text{Max Value} - \text{Min Value}Range=Max Value−Min Value
   It is a simple measure of variability but can be affected by extreme values.
7. **Quartiles:** Quartiles divide the dataset into four equal parts, providing a deeper understanding of data distribution. The three key quartiles are:
   - Q1 (First Quartile/25th percentile): 25% of the data is below this value.
   - Q2 (Median/50th percentile): 50% of the data is below this value.
   - Q3 (Third Quartile/75th percentile): 75% of the data is below this value.
8. The difference between Q3 and Q1 is called the **Interquartile Range (IQR)**, which is a measure of statistical dispersion and is useful for identifying outliers:
   IQR=Q3−Q1\text{IQR} = Q3 - Q1IQR=Q3−Q1
9. **Skewness:** Skewness measures the asymmetry of the data distribution. A dataset is said to be:
   - Positively skewed (right-skewed) if the right tail is longer.
   - Negatively skewed (left-skewed) if the left tail is longer.
10. **Kurtosis:** Kurtosis measures the "tailedness" of the data distribution. It describes the shape of the data distribution in terms of the height and sharpness of the peak, and the presence of outliers.
    - High kurtosis indicates more outliers and sharper peaks.
    - Low kurtosis indicates fewer outliers and flatter distributions.

Importance of Descriptive Statistics:

- Data Summary: Descriptive statistics provide a quick and easy way to summarize large datasets, allowing researchers to understand the data's key characteristics without analyzing individual data points.

- Detecting Outliers: Measures such as IQR and standard deviation help identify outliers, which may distort the results of further analysis.
- Comparison: Descriptive statistics allow for comparison between different datasets, helping in identifying similarities or differences.
- Basis for Further Analysis: These statistics often serve as the foundation for more complex analyses like hypothesis testing, regression, and machine learning models.

References :

Database https://www.kaggle.com/datasets

Conclusion :

Descriptive statistics provide a quick summary of a dataset's main characteristics, including central tendency, variability, and distribution. They help identify patterns, trends, and outliers, offering a solid foundation for further analysis and data-driven decision-making.