Assignment No. 2

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

Objective: The Objective of this assignment is to analyze the Iris dataset, which includes measurements of different iris flower species, by computing basic descriptive statistics such as summary statistics and quartiles. This will help us understand the dataset's distribution, central tendency (mean and median), and variability (standard deviation). By calculating these statistics, we aim to gain insights into the data's characteristics and develop skills in data analysis using Python, preparing us for more advanced analytical tasks in the future.

Prerequisite :

1. **Python and Libraries**: Basic knowledge of Python programming and familiarity with libraries like Pandas for data manipulation and Seaborn for visualization.
2. **Statistical Concepts**: Understanding of key statistical measures, including mean, median, standard deviation, and quartiles, to analyze data distribution.
3. **Environment Setup**: Ensure a Python environment is set up with necessary libraries installed (e.g., using Anaconda or pip).

Theory :

1) Iris Dataset:

The Iris dataset is a well-known dataset in both machine learning and statistical analysis. It contains 150 samples, each representing an iris flower from one of three species: *Iris setosa*, *Iris versicolor*, and *Iris virginica*. For each sample, four features are recorded: **sepal length**, **sepal width**, **petal length**, and **petal width**. These features are measured in centimeters and serve as the input variables to classify the species of the flower. The Iris dataset is widely used for teaching data analysis techniques due to its simplicity and well-defined structure.

2) Descriptive Statistics:

Descriptive statistics provide an overview of a dataset by summarizing its central tendency, dispersion, and overall distribution. These statistics help in understanding the structure and characteristics of the data.

- **a) Mean:**

The mean is the arithmetic average of a dataset, calculated by summing all the values and dividing by the number of observations. It is a common measure of central tendency but can be sensitive to outliers.

- **b) Median:**
  The median is the middle value when the data points are ordered from smallest to largest. It provides a better measure of central tendency for skewed datasets as it is not affected by extreme values.
- **c) Standard Deviation:**
  The standard deviation measures how much the data points deviate from the mean. A low standard deviation indicates that the data points are close to the mean, while a high standard deviation suggests a wider spread of values.
- **d) Quartiles:**
  Quartiles divide the dataset into four equal parts. These include:
  - Q1 (25th percentile) – the median of the lower half of the data.
  - Q2 (50th percentile) – the overall median.
  - Q3 (75th percentile) – the median of the upper half. Quartiles help in identifying the spread of the dataset and detecting outliers.

3) Data Visualization:

Visualizing data is a crucial step in exploratory data analysis as it helps to uncover patterns, trends, and potential outliers in the dataset.

- **a) Box Plots:**
  A box plot is a graphical representation that shows the distribution of a dataset based on its quartiles. It displays the minimum, first quartile, median, third quartile, and maximum. Box plots are particularly useful for detecting outliers.
- **b) Histograms:**
  A histogram displays the frequency distribution of a numerical dataset. It divides the data into bins and counts the number of observations in each bin, providing a clear view of the data's distribution shape (e.g., normal, skewed, bimodal).

4) Data Manipulation with Pandas:

The **Pandas** library in Python is a powerful tool for data manipulation and analysis. It provides easy-to-use data structures, such as **DataFrames**, for handling tabular data efficiently.

- **.describe():**
  This function in Pandas generates a summary of descriptive statistics for the dataset, including the count, mean, standard deviation, and quartiles for each numerical feature.

- **.quantile():**
  This function calculates specific percentiles (e.g., Q1, Q2, Q3) of the dataset, which helps in identifying the spread and distribution of the data.

Algorithm (if any to achieve the objective):

Step 1: Import necessary libraries (Pandas, NumPy, Matplotlib, Seaborn).

Step 2: Load the Iris dataset (using `pandas.read_csv()` or `seaborn.load_dataset()`).

Step 3: Display the first few rows of the dataset and check for missing values.

Step 4: Compute summary statistics using the `describe()` function.

Step 5: Calculate the quartiles (Q1, Q2, Q3) using the `quantile()` function.

Step 6: Calculate mean, median, and standard deviation for each feature.

Step 7: Visualize data using box plots to show quartiles and detect outliers.

Step 8: Visualize the frequency distribution of features using histograms.

Step 9: Analyze and interpret the results based on descriptive statistics and visualizations.

References :

1. **UCI Machine Learning Repository - Iris Dataset:**
   - The original source of the Iris dataset used in this practical analysis.
   - **Reference:** Fisher, R. A. (1936). "The use of multiple measurements in taxonomic problems." *Annals of Eugenics*, 7(2), 179-188.
2. **Pandas Documentation:**
   - For the usage of `describe()`, `quantile()`, and other functions to calculate descriptive statistics.
   - **Reference:** The Pandas Development Team (2023). "Pandas Documentation."
3. **Seaborn Documentation:**
   - For loading the Iris dataset and using Seaborn for visualization (box plots, histograms).
   - **Reference:** Michael Waskom (2023). "Seaborn: Statistical Data Visualization." Available at: https://seaborn.pydata.org/
4. **Python for Data Analysis - Book by Wes McKinney:**
   - A comprehensive guide for data manipulation and analysis using Pandas, including descriptive statistics and visualization.
   - **Reference:** McKinney, W. (2017). *Python for Data Analysis: Data Wrangling*

with Pandas, NumPy, and IPython. O'Reilly Media.

Conclusion :

In this assignment, we analyzed the Iris dataset using descriptive statistics and data visualization. Key metrics like mean, median, standard deviation, and quartiles helped us understand the central tendency, variability, and distribution of features such as sepal and petal dimensions.

Visualizations, including box plots and histograms, revealed patterns and outliers among the species, with notable differences in petal characteristics between *Iris setosa* and the other species. This analysis provided valuable insights into the dataset and built a foundation for more advanced data analysis tasks.