**Assignment No. 2**

**Name: Aditya Wanave**

**PRN No: 122B1F135**

**Problem Statement:**

**Write a Python Script to Find Basic Descriptive Statistics Using Summary, Quartile Function, etc., on the Iris Dataset**

**Objective:**  
The objective of this assignment is to write a Python script that computes basic descriptive statistics, including measures like mean, median, quartiles, and a summary, for the **Iris dataset**. We aim to understand and analyze the key statistical metrics of the dataset using Python's libraries, primarily pandas and numpy.

**Prerequisites:**

1. **Python environment** with the following libraries installed:
   * **pandas** for data manipulation and analysis.
   * **numpy** for numerical calculations.
   * **seaborn or sklearn.datasets** (optional) for loading the Iris dataset if not available locally.
2. A **text editor** or **IDE** to write and run the Python script.
3. **Basic knowledge** of Python and descriptive statistics concepts like mean, median, standard deviation, and quartiles.

**Theory:**

The **Iris dataset** is a classic dataset used for machine learning and statistical analysis. It contains 150 records of iris flowers, with 50 samples from each of the three species: *Iris setosa*, *Iris versicolor*, and *Iris virginica*. Each sample includes four features:

* **Sepal Length**
* **Sepal Width**
* **Petal Length**
* **Petal Width**

For our task, we will focus on calculating basic descriptive statistics, such as:

1. **Mean:** The average of the values in a dataset.
2. **Median:** The middle value when the data is sorted.
3. **Standard Deviation:** A measure of how spread out the values are from the mean.
4. **Quartiles:** Values that divide the dataset into four equal parts.
5. **Summary Statistics:** A consolidated overview of basic statistics like count, mean, min, max, etc.

**Algorithm for Computing Descriptive Statistics:**

**Step 1: Import necessary libraries**

* Load the **Iris dataset** into a pandas DataFrame.

**Step 2: Generate basic descriptive statistics**

* Use pandas' describe() function to get a summary of the dataset.
* Calculate specific statistics like mean, median, standard deviation, and quartiles.

**Step 3: Output the results to the console or save them into a file**

**Python Script Example:**

*# Step 1: Import necessary libraries*

*import pandas as pd*

*import numpy as np*

*from sklearn.datasets import load\_iris*

*# Load the Iris dataset*

*iris = load\_iris()*

*df\_iris = pd.DataFrame(data=iris.data, columns=iris.feature\_names)*

*# Step 2: Generate descriptive statistics*

*# 2.1 Basic summary statistics*

*summary\_stats = df\_iris.describe()*

*print("Summary Statistics:")*

*print(summary\_stats)*

*# 2.2 Mean for each feature*

*mean\_values = df\_iris.mean()*

*print("\nMean Values:")*

*print(mean\_values)*

*# 2.3 Median for each feature*

*median\_values = df\_iris.median()*

*print("\nMedian Values:")*

*print(median\_values)*

*# 2.4 Standard deviation for each feature*

*std\_dev = df\_iris.std()*

*print("\nStandard Deviation:")*

*print(std\_dev)*

*# 2.5 Quartiles for each feature*

*q1 = df\_iris.quantile(0.25)*

*q3 = df\_iris.quantile(0.75)*

*print("\n1st Quartile (Q1):")*

*print(q1)*

*print("\n3rd Quartile (Q3):")*

*print(q3)*

*# Step 3: Optionally, save the results to a CSV file*

*summary\_stats.to\_csv('iris\_summary\_statistics.csv')*

**Step-by-step explanation:**

1. **Loading the Dataset:** We use sklearn.datasets.load\_iris() to load the Iris dataset, which is readily available in the sklearn library. The dataset is then converted into a pandas DataFrame for easier manipulation.
2. **Summary Statistics:** The describe() function provides basic statistics such as count, mean, standard deviation, minimum, maximum, and the quartiles.
3. **Specific Statistics:**
   * The **mean** is calculated using mean().
   * The **median** is calculated using median().
   * The **standard deviation** is obtained using std().
   * The **quartiles** (Q1 and Q3) are computed using quantile().
4. **Saving Results:** The summary statistics are saved to a CSV file (iris\_summary\_statistics.csv) for later reference or further analysis.

**References:**

* Pandas Documentation
* NumPy Documentation
* Scikit-learn Documentation

**Conclusion:**

By executing this script, we are able to compute and analyze the basic descriptive statistics for the Iris dataset, including mean, median, quartiles, and standard deviation. This process helps us understand the distribution and spread of the data, which is crucial for further data analysis and modeling tasks.