Assignment No.2

Problem Statement: Analyzing Descriptive Statistics of the Iris Dataset

Objective:

The purpose of this assignment is to analyze the Iris dataset by calculating key descriptive statistics, such as summary measures and quartiles, to gain insights into the dataset's distribution.

Prerequisites:

- Python Installation:
  - Ensure that Python 3 is installed on your system.
- Required Libraries:
  - Use the following command to install the necessary libraries:
  - pip install pandas numpy seaborn
- Basic Python Knowledge:
  Familiarity with data analysis concepts and basic statistics.

Theory:

Descriptive statistics provide a concise summary of the key features of a dataset, including central tendency, dispersion, and the data's spread.

1. **Measures of Central Tendency:**
   - **Mean:**
     The average value of a dataset, calculated by summing all observations and dividing by the total number of observations. It offers a general sense of the dataset's level but can be affected by outliers.
   - **Median:**
     The middle value of a sorted dataset, ensuring that half of the data points fall above and half below it. The median is less affected by outliers, making it more representative of a dataset with extreme values.
   - **Mode:**
     The most frequently occurring value in the dataset, often used to understand the most common observation in a distribution.
2. **Measures of Dispersion:**
   - **Variance:**
     Variance indicates the spread of data points by showing how much they deviate from the mean. It's the average of the squared differences from the mean.
   - **Standard Deviation:**
     A more interpretable measure than variance, as it is in the same units as the data. It quantifies the spread or variability in the dataset. A lower standard deviation suggests data points are closer to the mean, while a higher value indicates greater variability.

- **Range:**
  The difference between the largest and smallest values. While it gives an immediate sense of spread, it doesn't convey details about the data's distribution.

3. **Quartiles and Interquartile Range (IQR):**
   - **Quartiles:**
     These divide the dataset into four equal parts. Q1 (the 25th percentile), Q2 (the median or 50th percentile), and Q3 (the 75th percentile) help summarize the distribution's spread.
   - **IQR (Interquartile Range):**
     The difference between Q3 and Q1, focusing on the middle 50% of the dataset. The IQR is useful for identifying outliers—values falling far outside this range can be considered atypical.

Dataset Overview:

The Iris dataset contains features such as sepal length, sepal width, petal length, and petal width across various species of iris flowers. Performing descriptive statistics on this dataset provides an essential overview of these features, offering insights into their central tendencies and variability. This foundational analysis can lead to deeper statistical analysis and machine learning applications.

Algorithm

1. Data Loading: Import the Iris dataset from the `seaborn` library.
2. Summary Statistics: Compute statistics such as mean, median, and standard deviation using the `describe()` method.
3. Quartile Calculation: Use the `quantile()` method to find the first, second, and third quartiles.
4. Output the Statistics: Print the results for interpretation.

Code

```
import pandas as pd
import seaborn as sns

iris = sns.load_dataset('iris')

summary = iris.describe()
print("Descriptive Statistics Summary:")
print(summary)

quartiles = iris.quantile([0.25, 0.5, 0.75])
print("\nQuartiles of the Dataset:")
print(quartiles)
```

References

- Fisher, R.A. (1936). "The use of multiple measurements in taxonomic problems." *Annals of Eugenics*, 7(2), 179–188.
- UCI Machine Learning Repository: [Iris Dataset](https://archive.ics.uci.edu/ml/datasets/iris)
- McKinney, W. "Python for Data Analysis," *O'Reilly Media*, 2017.

Conclusion

This analysis of descriptive statistics on the Iris dataset provides a foundational understanding of the dataset's characteristics, enabling deeper insights for further analysis and modeling.