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 apply and deepen our understanding of descriptive statistics concepts through the analysis of the Iris dataset. We will compute key statistical measures such Central tendency and dispersion or variability. This hands-on experience will enhance our ability to summarize and interpret data effectively.

Prerequisite :

1. Basic understanding of statistics, particularly descriptive statistics.
2. Knowledge of the Pandas, numpy library for data manipulation and analysis.
3. Experience with data visualization techniques using libraries like Matplotlib or Seaborn (optional).
4. Text editor and basic knowledge of file handling in Python.

Theory :

1. Descriptive Statistics :-

It is a branch of statistics that deals with the summarization and description of the main features of a dataset. Descriptive statistics are numbers that are used to describe and summarize the data. It provides a simple summary about the sample and the measures. The summary measures include measures of central tendency (mean, median and mode) and measures of variability (variance, standard deviation, IQR (Interquartile Range)).

2. Measures of central tendency :-

Central tendency refers to a central value that describes a probability distribution, representing the "center" or location of the data. The primary measures of central tendency are mean, median, and mode. While the mean is the most common measure, the median is often preferred for skewed distributions or when outliers are present. So, median is more robust measure than the mean.

I. Mean –

- Mean is also known as the simple average.
- It is denoted by greek letter μ for population and by $\bar{x}$ for sample.
- We can find mean of a number of elements by adding all the elements in a dataset and then dividing by the number of elements in the dataset.
- It is the most common measure of central tendency but it has a following drawback.
- The mean is affected by the presence of outliers.
- So, mean alone is not enough for making business decisions.

II. Median –

- Median is the number which divides the dataset into two equal halves.
- To calculate the median, we have to arrange our dataset of n numbers in ascending order.
- The median of the dataset is the number at (n+1)/2 th position, if n is odd.
- If n is even, then the median is the average of the (n/2)th number and (n+2)/2 th number.
- Median is robust to outliers.
- So, for skewed distribution or when there is concern about outliers, the median may be preferred.

III. Mode –

- Mode of a dataset is the value that occurs most often in the dataset.
- Mode is the value that has the highest frequency of occurrence in the dataset

3. Measures of dispersion or variability :-

Dispersion is an indicator of how far away from the center, we can find the data values. The most common measures of dispersion are **variance**, **standard deviation** and **interquartile range (IQR). Variance** is the standard measure of spread. The **standard deviation** is the square root of the variance. The **variance** and **standard deviation** are two useful measures of spread.

I. Variance –

- Variance measures the dispersion of a set of data points around their mean value.
- It is the mean of the squares of the individual deviations.
- Variance gives results in the original units squared.

II. Standard deviation –

- It is the square-root of the variance.
- For Normally distributed data, approximately 95% of the values lie within 2 s.d. of the mean.
- Standard deviation gives results in the original units.

III. Coefficient of Variation (CV)

- Coefficient of Variation (CV) is equal to the standard deviation divided by the mean.
- It is also known as relative standard deviation.

IV. IQR (Interquartile range)

- The IQR is calculated using the boundaries of data situated between the 1st and the 3rd quartiles.
- The interquartile range (IQR) can be calculated as follows:- IQR = Q3 – Q1
- IQR is a more robust measure of spread than variance and standard deviation and should therefore be preferred for small or asymmetrical distributions.
- It is a robust measure of spread.

Algorithm (if any to achieve the objective)

1. Import necessary libraries (Pandas, NumPy, Seaborn).
2. Load the Iris dataset from a CSV file.
3. Calculate descriptive statistics:
   - Mean, Median, Mode
   - Variance, Standard Deviation, Coefficient of Variation
   - Quartiles and IQR
4. Display the calculated statistics.

Code & Output –

```
[1]: # Importing the pandas library
     import pandas as pd

     # pd.read_csv() is used to load data from a CSV file into a DataFrame
     df = pd.read_csv('C:/Users/dnyan/FODS Assignments/Datasets/iris.csv')

     # df.head() shows the first 5 rows
     df.head()
```

[1]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

```
[2]: # View dimensions of dataset
     df.shape
```

```
[2]: (150, 6)
```

```
[3]: # View summary of dataset
     df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Id             150 non-null    int64
 1   SepalLengthCm  150 non-null    float64
 2   SepalWidthCm   150 non-null    float64
 3   PetalLengthCm  150 non-null    float64
 4   PetalWidthCm   150 non-null    float64
 5   Species        150 non-null    object
dtypes: float64(4), int64(1), object(1)
```

```
[4]: # Check for missing values
     df.isnull().sum()
```

```
[4]: Id              0
     SepalLengthCm    0
     SepalWidthCm     0
     PetalLengthCm    0
     PetalWidthCm     0
     Species          0
     dtype: int64
```

```
[5]: # gives summary statistics of numeric columns only.
     df.describe()
```

[5]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	75.500000	5.843333	3.054000	3.758667	1.198667
std	43.445368	0.828066	0.433594	1.764420	0.763161
min	1.000000	4.300000	2.000000	1.000000	0.100000
25%	38.250000	5.100000	2.800000	1.600000	0.300000
50%	75.500000	5.800000	3.000000	4.350000	1.300000
75%	112.750000	6.400000	3.300000	5.100000	1.800000
max	150.000000	7.900000	4.400000	6.900000	2.500000

```
[7]: # gives the summary statistics of the SepalWidthCm columns.
     df['SepalWidthCm'].describe()
```

```
[7]: count    150.000000
     mean       3.054000
     std        0.433594
     min        2.000000
     25%        2.800000
     50%        3.000000
     75%        3.300000
     max        4.400000
     Name: SepalWidthCm, dtype: float64
```

```
[8]: # # Get descriptive statistics for all columns
     df.describe(include='all')
```

[8]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
count	150.000000	150.000000	150.000000	150.000000	150.000000	150
unique	NaN	NaN	NaN	NaN	NaN	3
top	NaN	NaN	NaN	NaN	NaN	Iris-setosa
freq	NaN	NaN	NaN	NaN	NaN	50
mean	75.500000	5.843333	3.054000	3.758667	1.198667	NaN
std	43.445368	0.828066	0.433594	1.764420	0.763161	NaN
min	1.000000	4.300000	2.000000	1.000000	0.100000	NaN
25%	38.250000	5.100000	2.800000	1.600000	0.300000	NaN
50%	75.500000	5.800000	3.000000	4.350000	1.300000	NaN
75%	112.750000	6.400000	3.300000	5.100000	1.800000	NaN
max	150.000000	7.900000	4.400000	6.900000	2.500000	NaN

[9]:
```
# The different categories of Species
df.Species.unique()
```

[9]:
```
array(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'], dtype=object)
```

[10]:
```
groups = df.groupby('Species',as_index= False)["Id"].count()
groups
```

[10]:

	Species	Id
0	Iris-setosa	50
1	Iris-versicolor	50
2	Iris-virginica	50

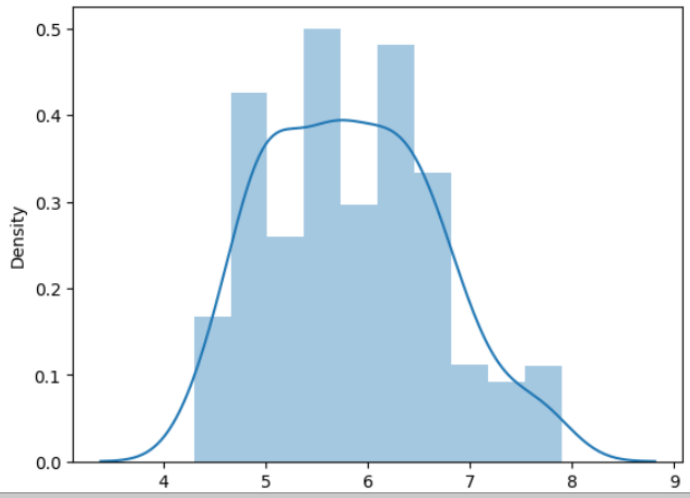
[11]:
```
# calculation of central tendency for SepalLengthCm column
mean_value = df['SepalLengthCm'].mean()
median_value = df['SepalLengthCm'].median()
mode_value = df['SepalLengthCm'].mode()

print("Mean Sepal Length: ", mean_value)
print("Median Sepal Length: ", median_value)
print("Mode Sepal Length: ", mode_value)
```

```
Mean Sepal Length:  5.843333333333334
Median Sepal Length:  5.8
Mode Sepal Length:  0    5.0
Name: SepalLengthCm, dtype: float64
```

[18]:
```
# Plot the distribution
import seaborn as sns

sns.distplot(df['SepalLengthCm'], bins=10, hist=True, kde=True, label = 'SepalLengthCm')
```

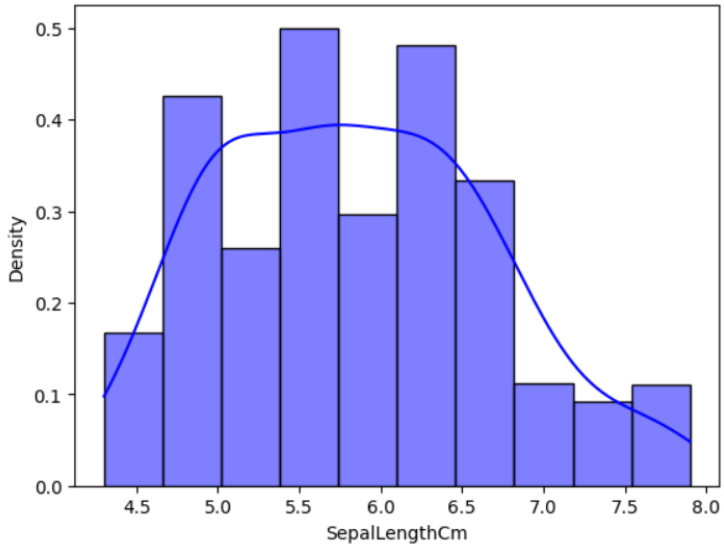
```
C:\Users\dnyan\AppData\Local\Temp\ipykernel_8756\3254176243.py:4: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df['SepalLengthCm'], bins=10, hist=True, kde=True, label = 'SepalLengthCm')
```

[18]:
```
<Axes: xlabel='SepalLengthCm', ylabel='Density'>
```



[19]:
```
sns.histplot(df['SepalLengthCm'], bins=10, kde=True, color='blue', label='Sepal Length', stat='density')
```

[19]:
```
<Axes: xlabel='SepalLengthCm', ylabel='Density'>
```



[21]:
```
# Computation of measures of dispersion or variability for PetalLengthCm column
import numpy as np
min_value = df['PetalLengthCm'].min()
max_value = df['PetalLengthCm'].max()
range_value = max_value - min_value
var = df['PetalLengthCm'].var()
std = df['PetalLengthCm'].std()
Q1 = df['PetalLengthCm'].quantile(0.25)
Q2 = df['PetalLengthCm'].quantile(0.5)
Q3 = df['PetalLengthCm'].quantile(0.75)
IQR = Q3  - Q1

print("Min value of Sepal Length: ", min_value)
print("Max value of Sepal Length: ", max_value)
print("Range of Sepal Length: ", range_value)
print("Variance: ", var)
print("Standrad Deviation: ", std)
print("Q1 or 25th percentile: ", Q1)
print("Median (Q2 or 50th percentile): ", Q2)
print("Q3 or 75th percentile: ", Q3)
print("Interquartile Range: ", IQR)
```

```
Min value of Sepal Length:  1.0
Max value of Sepal Length:  6.9
Range of Sepal Length:  5.9
Variance:  3.113179418344519
Standrad Deviation:  1.7644204199522626
Q1 or 25th percentile:  1.6
Median (Q2 or 50th percentile):  4.35
Q3 or 75th percentile:  5.1
Interquartile Range:  3.4999999999999996
```

[23]:
```
df['PetalWidthCm'].skew()
```

[23]:
```
np.float64(-0.10499656214412734)
```

[24]:
```
df['PetalWidthCm'].kurt()
```

[24]:
```
np.float64(-1.3397541711393433)
```

References :

1. **https://colab.research.google.com/drive/12F_1x3qy0xzfkvW561sFHFQJ9zaEtUlF#scrollTo=1I2UBAFIJiJ1**
2. **https://www.kaggle.com/code/saurav9786/descriptive-statistics**
3. **https://www.kaggle.com/code/bharath25/descriptive-statistics-and-machine-learning-iris**

Conclusion :

In this assignment, we analyzed the Iris dataset using descriptive statistics, gaining insights into central tendency and variability. By calculating the mean, median, mode, variance, standard deviation, and interquartile range, we summarized the dataset's characteristics, which is crucial for informed decision-making and further statistical modeling. Descriptive statistics serve as a powerful tool for exploring and interpreting data patterns effectively.