Experiment - 01

Aim: To explore the descriptive statistics on the given dataset

Theory:

1. Introduction to Descriptive Statistics

Descriptive statistics summarize and describe the key features of a dataset, providing an overview of its structure and distribution. They include measures of central tendency (mean, median, and mode) and measures of variability (variance, standard deviation, range, IQR).

- **Measures of Central Tendency**:
  - **Mean**: The average of all data points. It is sensitive to outliers.
  - **Median**: The middle value when data is ordered. It is robust to outliers.
  - **Mode**: The most frequently occurring value.
- **Measures of Variability**:
  - **Variance**: Measures the spread of data around the mean in squared units.
  - **Standard Deviation**: The square root of variance; indicates how much data deviates from the mean.
  - **IQR (Interquartile Range)**: Difference between the 3rd and 1st quartiles (Q3 - Q1), robust against outliers.
  - **Coefficient of Variation (CV)**: Relative measure of variability calculated as standard deviation divided by the mean.

2. Measures of Shape

These include skewness and kurtosis, which describe the distribution's symmetry and peakedness.

- **Skewness**: Indicates the symmetry of data:
  - **Negative Skew**: The distribution has a long tail on the left side. The data values are concentrated on the right, and extremely small values pull the mean downward.
    Relationship: Mean < Median < Mode.
  - **Zero Skew**: The distribution is symmetrical, appearing balanced on both sides of the central value. It often resembles a bell shape.
    Relationship: Mean = Median = Mode.
  - **Positive Skew**: The distribution has a long tail on the right side. The data values are concentrated on the left, and extremely large values pull the mean upward.
    Relationship: Mean > Median > Mode.
- **Kurtosis**: Reflects the sharpness of the peak in data:
  - High kurtosis: Distinct peak and heavy tails.
  - Low kurtosis: Flat peak and lighter tails.

Code:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris

# Load the Iris dataset
data = load_iris()
iris_df = pd.DataFrame(data.data, columns=data.feature_names)
iris_df['species'] = data.target

# Print basic information
print(iris_df.shape)
print(iris_df.head())
print(iris_df.info())
print(iris_df.isnull().sum())
print(iris_df.describe())

# Calculate mean, median, and mode
mean = iris_df['sepal length (cm)'].mean()
print("\nMean:", mean)

median = iris_df['sepal length (cm)'].median()
print("Median:", median)

mode = iris_df['sepal length (cm)'].mode()
print("Mode:", mode)

# Distribution plot
sns.histplot(iris_df['sepal length (cm)'], bins=10, kde=True,
color='blue')
plt.title("Distribution Plot of Sepal Length (cm)")
plt.xlabel("Sepal Length (cm)")
plt.ylabel("Frequency")
plt.legend(labels=['sepal length (cm)'])
plt.show()

print(" ")
# Boxplot
sns.boxplot(x=iris_df['sepal length (cm)'], color='green')
plt.title("Boxplot of Sepal Length (cm)")
plt.xlabel("Sepal Length (cm)")
plt.show()
```

```
# Calculate other statistics
print("Min:", iris_df['sepal length (cm)'].min())
print("Max:", iris_df['sepal length (cm)'].max())
print("Range:", iris_df['sepal length (cm)'].max() - iris_df['sepal
length (cm)'].min())
print("Variance:", iris_df['sepal length (cm)'].var())
print("Standard Deviation:", iris_df['sepal length (cm)'].std())

# Interquartile Range (IQR)
Q1 = iris_df['sepal length (cm)'].quantile(0.25)
Q2 = iris_df['sepal length (cm)'].quantile(0.5)
Q3 = iris_df['sepal length (cm)'].quantile(0.75)
IQR = Q3 - Q1

print("Q1:", Q1)
print("Q2 (Median):", Q2)
print("Q3:", Q3)
print("IQR:", IQR)

# Skewness and Kurtosis
print("Skewness:", iris_df['sepal length (cm)'].skew())
print("Kurtosis:", iris_df['sepal length (cm)'].kurt())
```

Output:

Basic Information: shape, head, info, isnull, describe

```
(150, 5)
   sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm)  species
0                5.1               3.5                1.4               0.2        0
1                4.9               3.0                1.4               0.2        0
2                4.7               3.2                1.3               0.2        0
3                4.6               3.1                1.5               0.2        0
4                5.0               3.6                1.4               0.2        0
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   sepal length (cm)  150 non-null    float64
 1   sepal width (cm)   150 non-null    float64
 2   petal length (cm)  150 non-null    float64
 3   petal width (cm)   150 non-null    float64
 4   species            150 non-null    int64
dtypes: float64(4), int64(1)
```

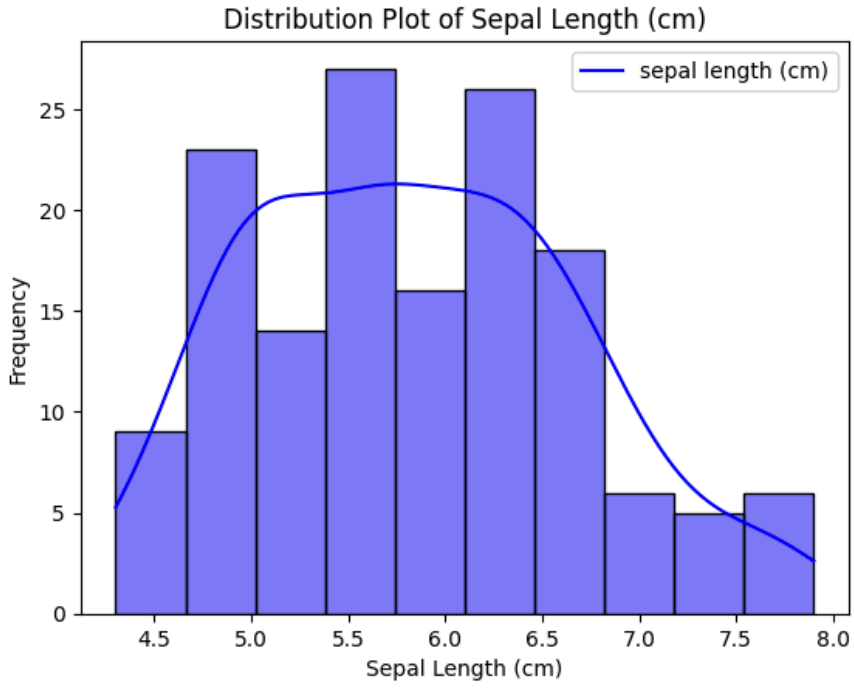
```
None
sepal length (cm)    0
sepal width (cm)     0
petal length (cm)    0
petal width (cm)     0
species              0
dtype: int64
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333	1.000000
std	0.828066	0.435866	1.765298	0.762238	0.819232
min	4.300000	2.000000	1.000000	0.100000	0.000000
25%	5.100000	2.800000	1.600000	0.300000	0.000000
50%	5.800000	3.000000	4.350000	1.300000	1.000000
75%	6.400000	3.300000	5.100000	1.800000	2.000000
max	7.900000	4.400000	6.900000	2.500000	2.000000

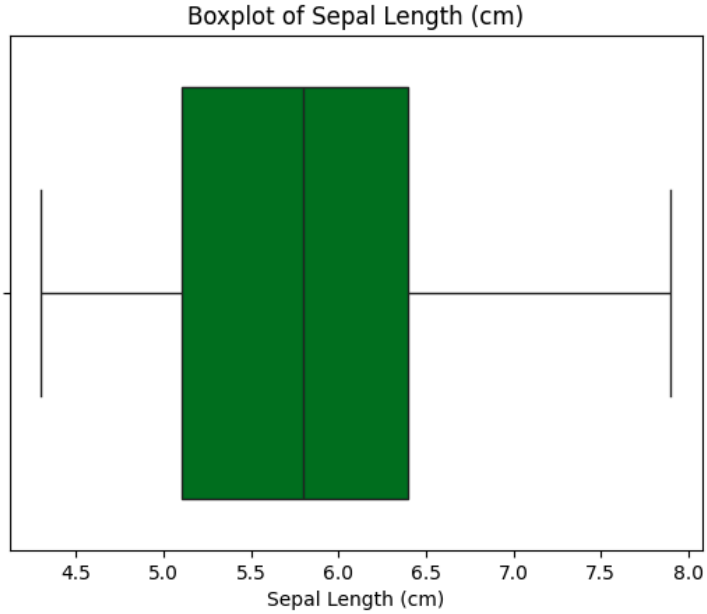
Mean, Median and Mode for the column “Sepal Length”

```
Mean: 5.843333333333334
Median: 5.8
Mode: 0    5.0
Name: sepal length (cm), dtype: float64
```

Distribution plot for the column “Sepal Length”



Boxplot for the column “Sepal Length”



Measures of dispersion or variability

```
Min: 4.3
Max: 7.9
Range: 3.6000000000000005
Variance: 0.6856935123042505
Standard Deviation: 0.8280661279778629
Q1: 5.1
Q2 (Median): 5.8
Q3: 6.4
IQR: 1.3000000000000007
Skewness: 0.3149109566369728
Kurtosis: -0.5520640413156395
```

Conclusion: Hence, we performed descriptive analysis on the iris dataset