Question 1 :

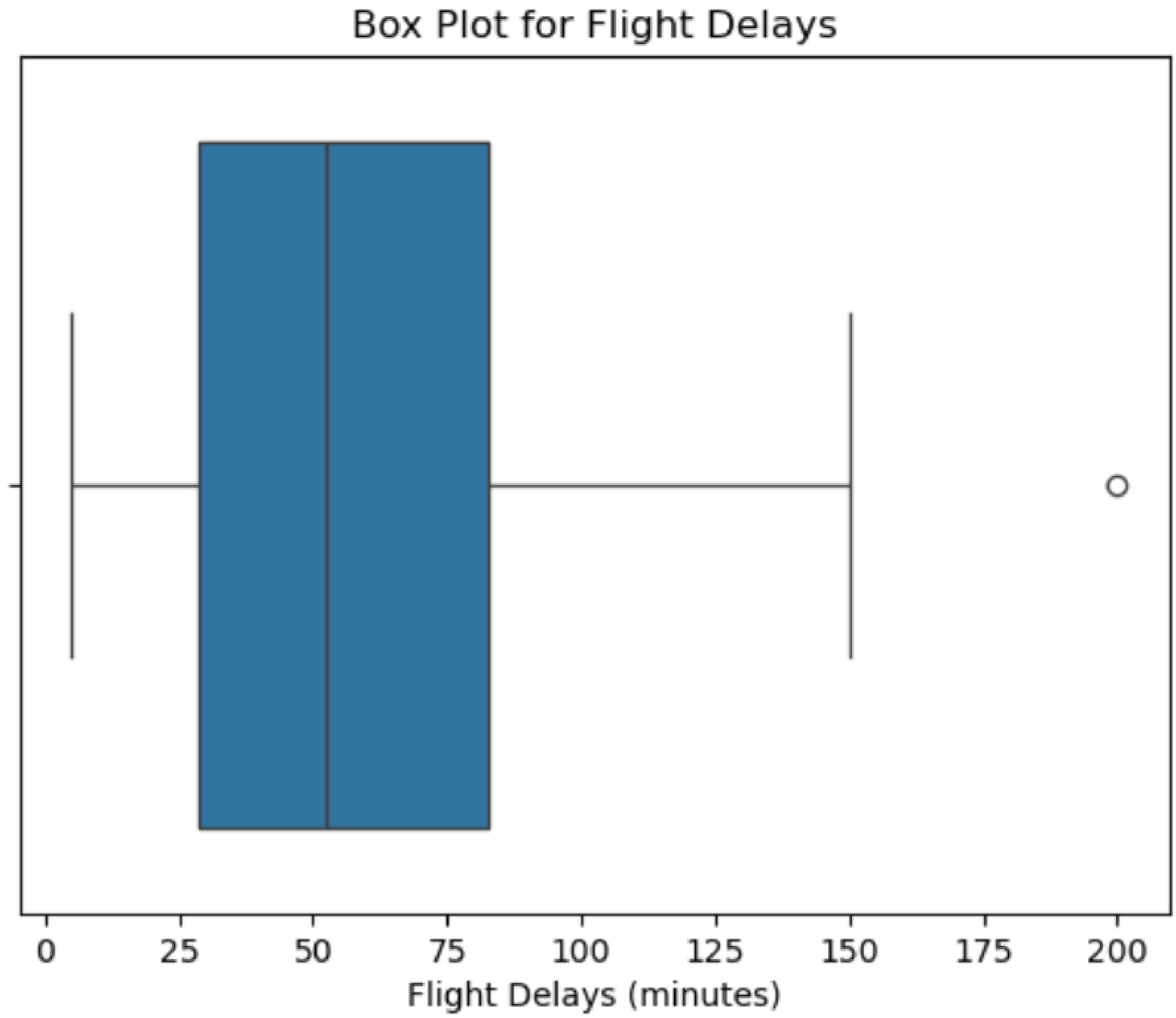
An airline tracks flight delays (in minutes) for 20 flights. Analyze the flight delays to calculate percentiles, detect outliers, and evaluate the overall distribution.

DataSet:

```
delays = [15, 30, 45, 20, 25, 100, 5, 60, 35, 50,

          120, 80, 10, 75, 90, 200, 55, 40, 70, 150]
```

Expected Output:

```
Percentiles (10th, 25th, 50th, 75th, 90th): [ 14.5   28.75  52.5   82.5  123.  ]
IQR: 53.75
Outliers: [200]
```

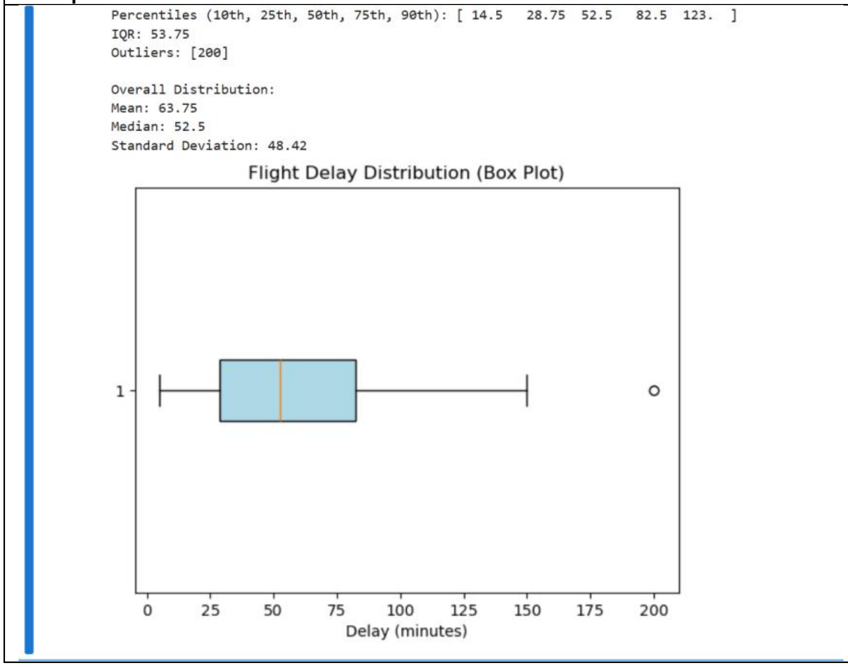


Answer:

```
import numpy as np
import matplotlib.pyplot as plt

# Dataset
delays = [15, 30, 45, 20, 25, 100, 5, 60, 35, 50,
          120, 80, 10, 75, 90, 200, 55, 40, 70, 150]

# Calculate Percentiles
percentiles = np.percentile(delays, [10, 25, 50, 75, 90])
print("Percentiles (10th, 25th, 50th, 75th, 90th):", percentiles)

# Calculate IQR
Q1 = percentiles[1]
Q3 = percentiles[3]
IQR = Q3 - Q1
print("IQR:", IQR)
```

```
#Find Outliers (1.5 * IQR rule)
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
outliers    = [x for x in delays if x < lower_bound or x >
upper_bound]
print("Outliers:", outliers)

#Evaluate overall distribution
mean   = np.mean(delays)
median = np.median(delays)
std    = np.std(delays)
print("\nOverall Distribution:")
print("Mean:", round(mean, 2))
print("Median:", median)
print("Standard Deviation:", round(std, 2))

#Box Plot Visualization
plt.boxplot(delays, vert=False, patch_artist=True,
boxprops=dict(facecolor='lightblue'))
plt.title("Flight Delay Distribution (Box Plot)")
plt.xlabel("Delay (minutes)")
plt.show()
```

Output



```
Percentiles (10th, 25th, 50th, 75th, 90th): [ 14.5   28.75  52.5   82.5  123.  ]
IQR: 53.75
Outliers: [200]

Overall Distribution:
Mean: 63.75
Median: 52.5
Standard Deviation: 48.42
```

Question 2 :

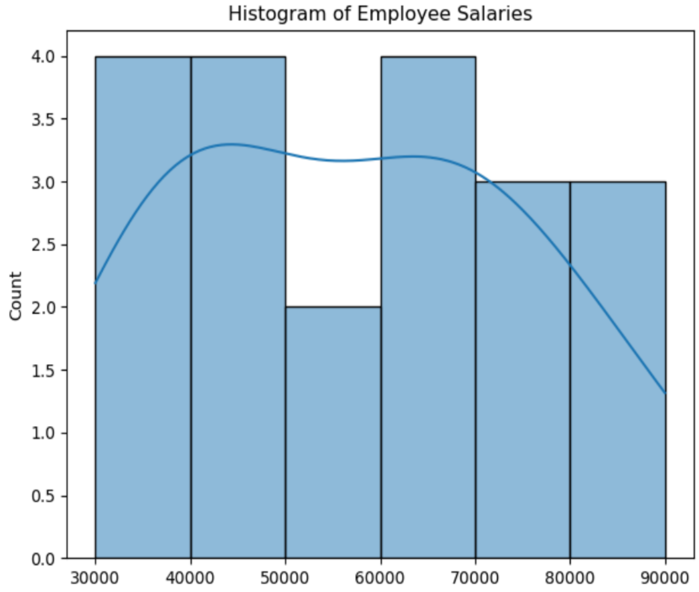
A company wants to analyze the salary distribution of its employees to understand the central tendency and determine whether the data is skewed.

DataSet:

```
salaries = [30000, 32000, 35000, 37000, 40000, 42000, 45000, 47000, 50000, 55000,

            60000, 62000, 65000, 67000, 70000, 72000, 75000, 80000, 85000, 90000]
```

Expected Output:

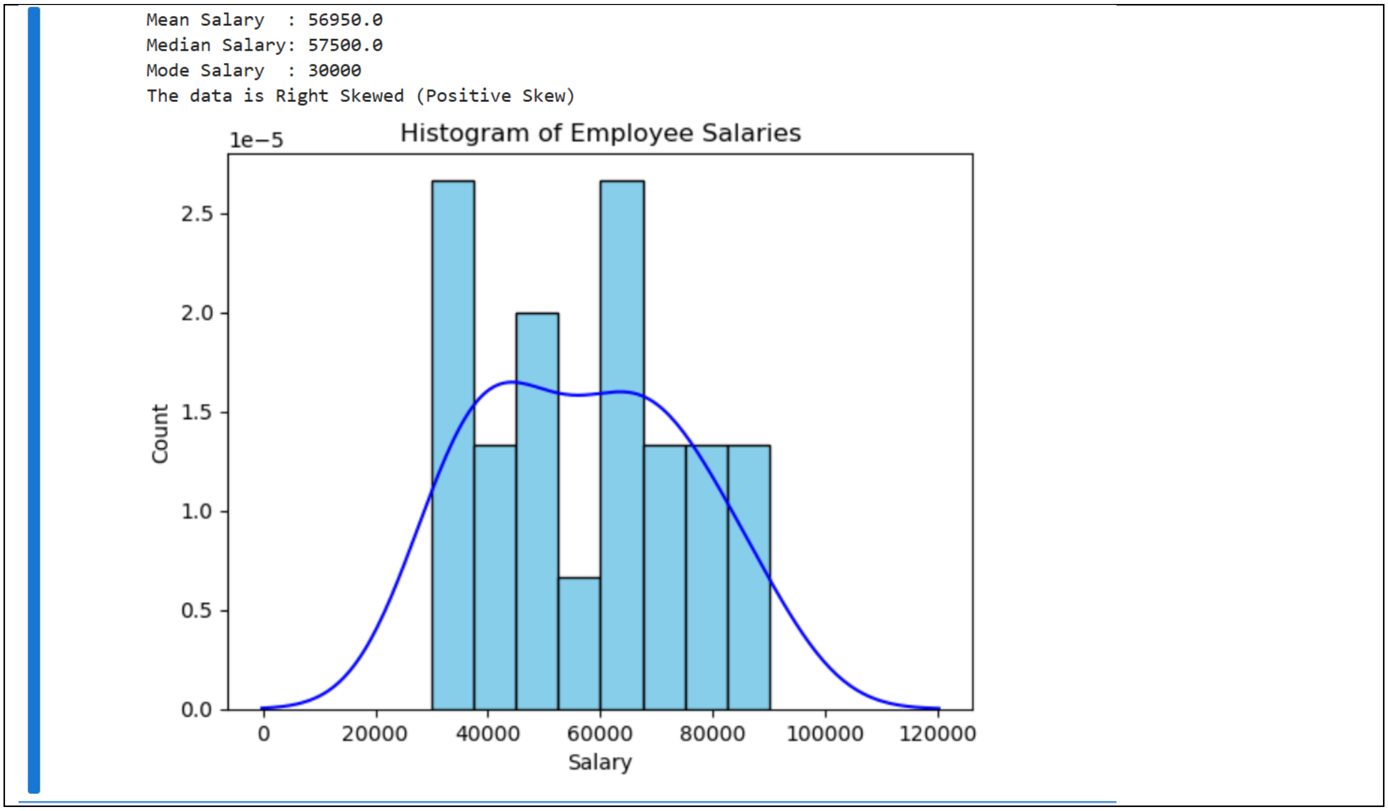
```
Mean Salary: 56950.0

Median Salary: 57500.0

Mode Salary: 30000

The data is Left Skewed (Negative Skew)
```



Answer:

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

# Dataset
salaries = [30000, 32000, 35000, 37000, 40000, 42000, 45000, 47000, 50000,
55000,
            60000, 62000, 65000, 67000, 70000, 72000, 75000, 80000, 85000,
90000]

#Calculate Central Tendency
mean_salary   = np.mean(salaries)
median_salary = np.median(salaries)
mode_salary   = stats.mode(salaries, keepdims=True)[0][0]

print("Mean Salary  :", mean_salary)
print("Median Salary:", median_salary)
print("Mode Salary  :", mode_salary)

#Determine Skewness
skew_value = stats.skew(salaries)
if skew_value > 0:
    skew_type = "Right Skewed (Positive Skew)"
elif skew_value < 0:
    skew_type = "Left Skewed (Negative Skew)"
else:
    skew_type = "Symmetrical"

print("The data is", skew_type)

#Plot Histogram
plt.hist(salaries, bins=8, color='skyblue', edgecolor='black', density=True)
plt.title("Histogram of Employee Salaries")
plt.xlabel("Salary")
plt.ylabel("Count")

# Overlay a smooth curve
sns.kdeplot(salaries, color='blue')

plt.show()
```

OUTPUT



```
Mean Salary  : 56950.0
Median Salary: 57500.0
Mode Salary  : 30000
The data is Right Skewed (Positive Skew)
```

Question 3:

A school wants to analyze the exam performance of students across three subjects: Mathematics, Science, and English. How can Data Science concepts be applied to understand their performance?

DataSet:

```
data = {
    'Student': ['A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J'],
    'Mathematics': [85, 78, 92, 88, 70, 95, 60, 80, 90, 76],
    'Science': [80, 85, 88, 70, 75, 92, 55, 82, 89, 78],
    'English': [78, 74, 85, 80, 68, 90, 50, 77, 83, 72]
}
```

Expected Output:

Descriptive Statistics
Histogram(graph)
Correlation Analysis(graph)
HeatMap(graph)

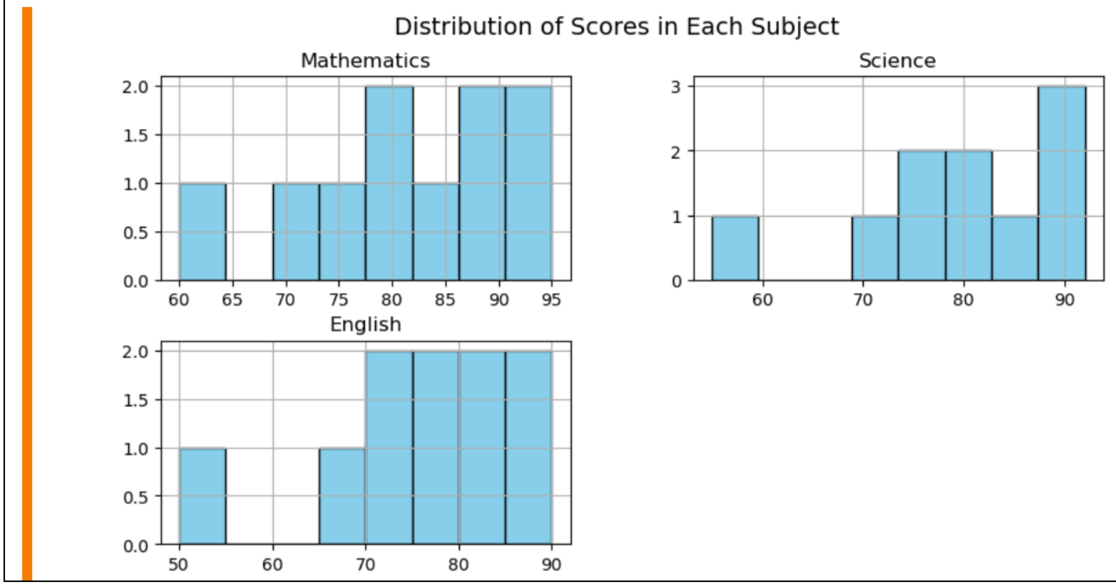
Answer

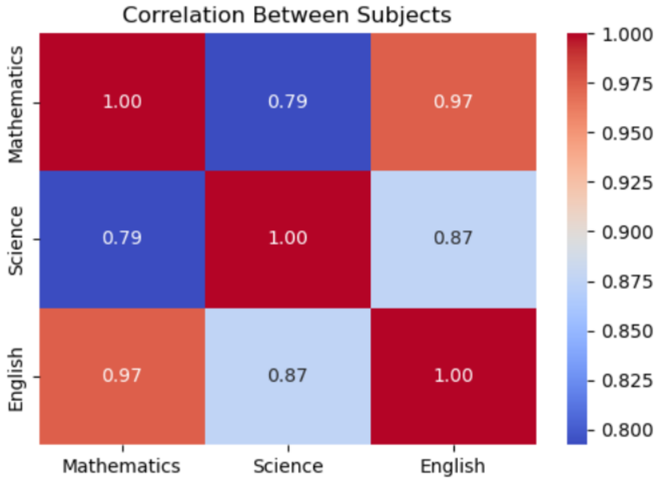
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Dataset
data = {
    'Student': ['A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J'],
    'Mathematics': [85, 78, 92, 88, 70, 95, 60, 80, 90, 76],
    'Science': [80, 85, 88, 70, 75, 92, 55, 82, 89, 78],
    'English': [78, 74, 85, 80, 68, 90, 50, 77, 83, 72]
}

# Convert to DataFrame
df = pd.DataFrame(data)

# Descriptive Statistics
print("Descriptive Statistics:\n")
print(df[['Mathematics', 'Science', 'English']].describe())

# Histogram for each subject
```

```
df[['Mathematics', 'Science', 'English']].hist(bins=8, figsize=(10,
5), color='skyblue', edgecolor='black')
plt.suptitle("Distribution of Scores in Each Subject", fontsize=14)
plt.show()

# Correlation Analysis
plt.figure(figsize=(6,4))
sns.heatmap(df[['Mathematics', 'Science', 'English']].corr(),
annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Correlation Between Subjects")
plt.show()
```

Output

```
Descriptive Statistics:

       Mathematics    Science    English
count    10.000000  10.000000  10.000000
mean     81.400000  79.400000  75.700000
std      10.844353  10.895463  11.086027
min      60.000000  55.000000  50.000000
25%      76.500000  75.750000  72.500000
50%      82.500000  81.000000  77.500000
75%      89.500000  87.250000  82.250000
max      95.000000  92.000000  90.000000
```




Correlation Between Subjects
Mathematics
Science
English
1.00
0.79
0.97
0.79
1.00
0.87
0.97
0.87
1.00
Mathematics
Science
English
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.800

WS

Question 4:

A pharmaceutical company conducted a clinical trial with two groups: one receiving medication and the other a placebo. How do you perform a hypothesis test to determine the effectiveness of the medication?

Dataset:

```
medication_group = [110, 115, 108, 102, 107, 99, 111, 104, 109, 101]

placebo_group = [120, 125, 130, 122, 128, 119, 124, 127, 123, 126]
```

Expected Output:

```
T-Statistic: -9.201427649220966

P-Value: 3.163912817600812e-08

Reject the null hypothesis: The medication is effective.
```

Answer:

```
import scipy.stats as stats

# Dataset
medication_group = [110, 115, 108, 102, 107, 99, 111, 104, 109, 101]
placebo_group = [120, 125, 130, 122, 128, 119, 124, 127, 123, 126]

# Perform Independent Two-Sample t-Test
t_statistic, p_value = stats.ttest_ind(medication_group,
placebo_group)

# Print results
print("T-Statistic:", t_statistic)
print("P-Value:", p_value)

# Inference
alpha = 0.05  # significance level
if p_value < alpha:
    print("Reject the null hypothesis: The medication is effective.")
else:
    print("Fail to reject the null hypothesis: No significant
difference.")
```

Output

```
T-Statistic: -9.201427649220966
P-Value: 3.163912817600812e-08
Reject the null hypothesis: The medication is effective.
```

Question 5 : A company conducted a customer satisfaction survey where customers rated their experience on a scale of 1 to 10. Analyze the survey results to calculate descriptive statistics and visualize the distribution of customer satisfaction ratings.

Sample DataSet:

ratings = [8, 9, 7, 5, 6, 10, 9, 4, 7, 8,

6, 9, 10, 5, 8, 7, 6, 9, 10, 7]

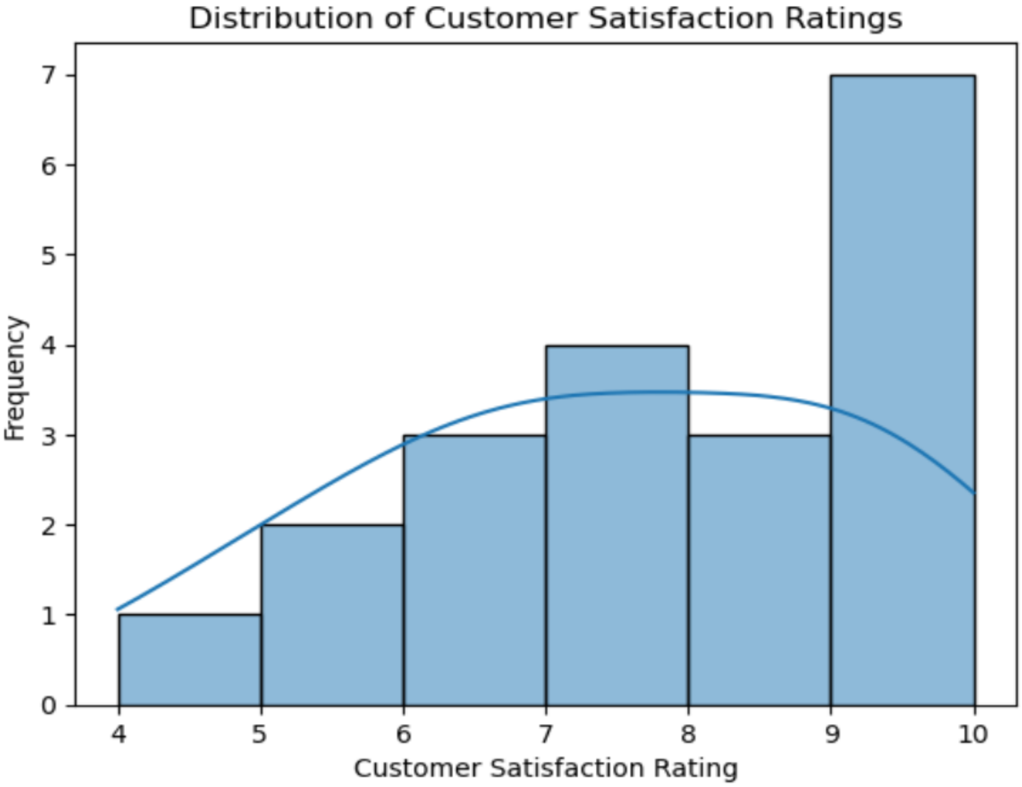
Expected Output:

Mean Rating: 7.5

Median Rating: 7.5

Mode Rating: 7

Standard Deviation: 1.746424919657298

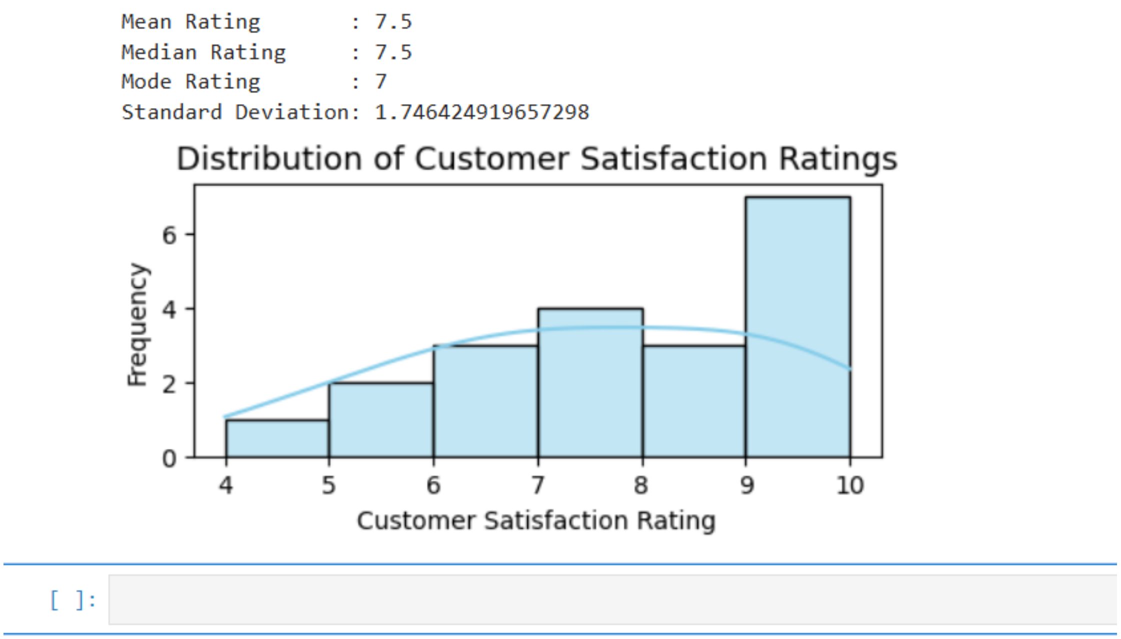


Answer

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
```

```
# DataSet
ratings = [8, 9, 7, 5, 6, 10, 9, 4, 7, 8,
           6, 9, 10, 5, 8, 7, 6, 9, 10, 7]

# Descriptive Statistics
mean_rating   = np.mean(ratings)
median_rating = np.median(ratings)
mode_rating   = stats.mode(ratings, keepdims=True)[0][0]
std_dev       = np.std(ratings)

print("Mean Rating        :", mean_rating)
print("Median Rating      :", median_rating)
print("Mode Rating        :", mode_rating)
print("Standard Deviation:", std_dev)

# Visualization
plt.figure(figsize=(5,2))
sns.histplot(ratings, bins=6, kde=True, color='skyblue',
edgecolor='black')
plt.title("Distribution of Customer Satisfaction Ratings",
fontsize=13)
plt.xlabel("Customer Satisfaction Rating")
plt.ylabel("Frequency")
plt.show()
```

OUTPUT



```
Mean Rating        : 7.5
Median Rating      : 7.5
Mode Rating        : 7
Standard Deviation: 1.746424919657298
```

Distribution of Customer Satisfaction Ratings
Frequency
0
2
4
6
4
5
6
7
8
9
10
Customer Satisfaction Rating
[ ]: