Experiment Number : 4 - Applying and interpreting different plots

---

Batch: FDS-2 **Roll Number: 16010422234** **Name: Chandana Ramesh Galgali**

Aim of the Experiment: Applying and interpreting different plots.

Program/ Steps:

1. Identify the attributes where it will be sensible to apply the below given plots.

a. Box Plot

b. Q Q Plot

c. Histogram

d. Scatter Plot

Apply the above mentioned plots on the identified attributes. Discuss the inferences from these plots in detail.

Code with Output/Result:

1. Importing Libraries and creating dataset:

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

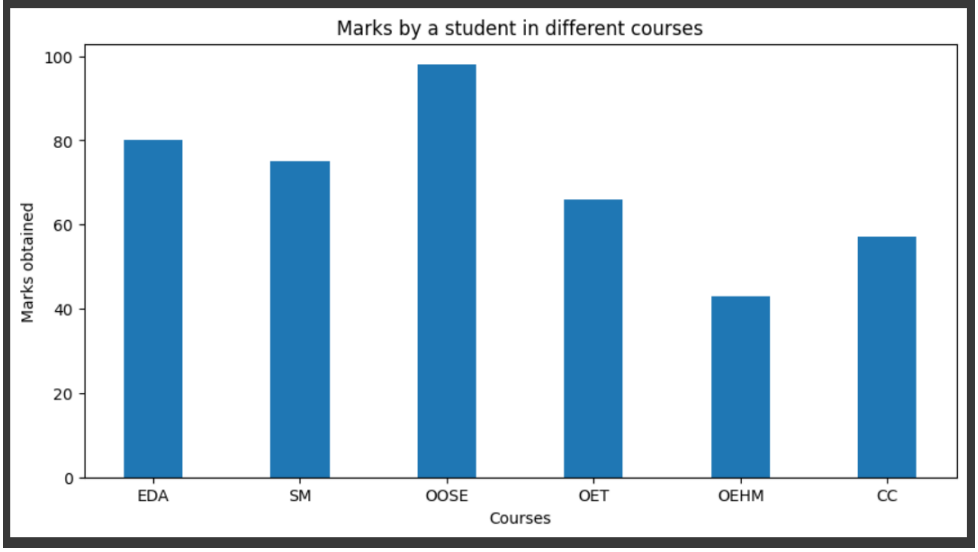
```
data={'EDA':80,'SM':75,'OOSE':98,'OET':66,'OEHM':43,'CC':57}
courses=list(data.keys())
marks=list(data.values())
```

```
print(data)
```

```
{'EDA': 80, 'SM': 75, 'OOSE': 98, 'OET': 66, 'OEHM': 43, 'CC': 57}
```

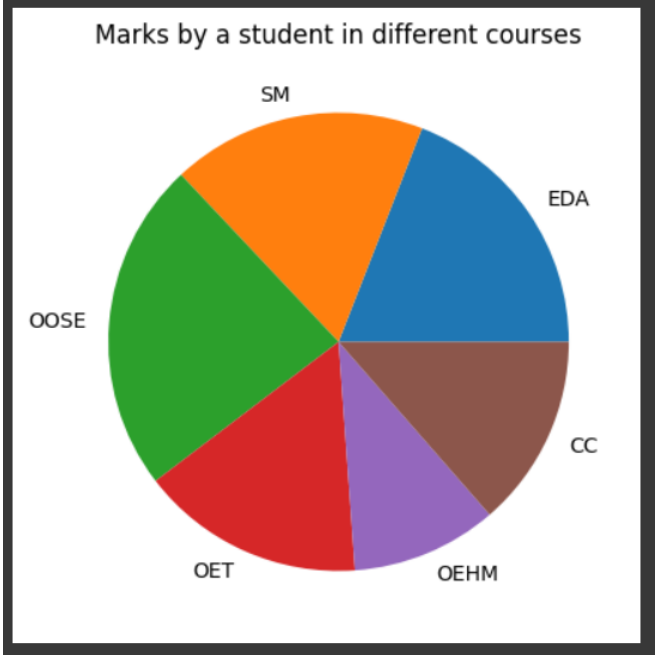
2. Bar Chart:

```
fig=plt.figure(figsize=(10,5))
plt.bar(courses, marks, width=0.4)
plt.xlabel("Courses")
plt.ylabel("Marks obtained")
plt.title("Marks by a student in different courses")
plt.show()
```



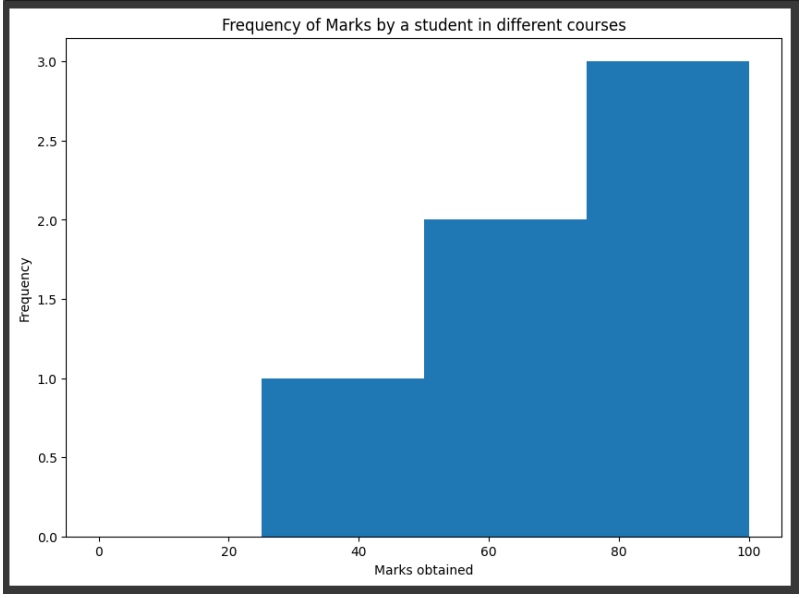
3. Pie Chart:

```
fig=plt.figure(figsize=(10,5))
plt.pie(marks, labels=courses)
plt.title("Marks by a student in different courses")
plt.show()
```



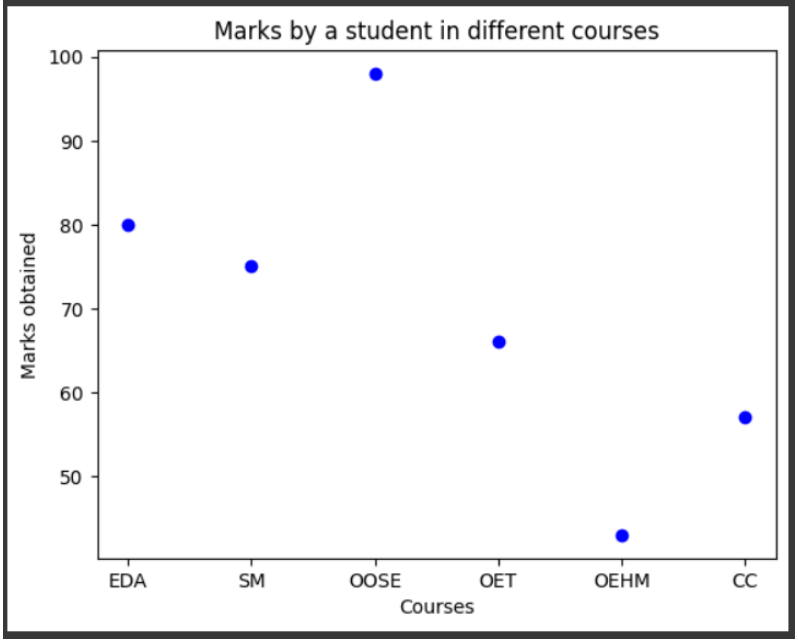
4. Histogram:

```
fig, ax = plt.subplots(figsize=(10,7))
ax.hist(marks, bins = [0, 25, 50, 75, 100])
plt.ylabel("Frequency")
plt.xlabel("Marks obtained")
plt.title("Frequency of Marks by a student in different courses")
plt.show()
```



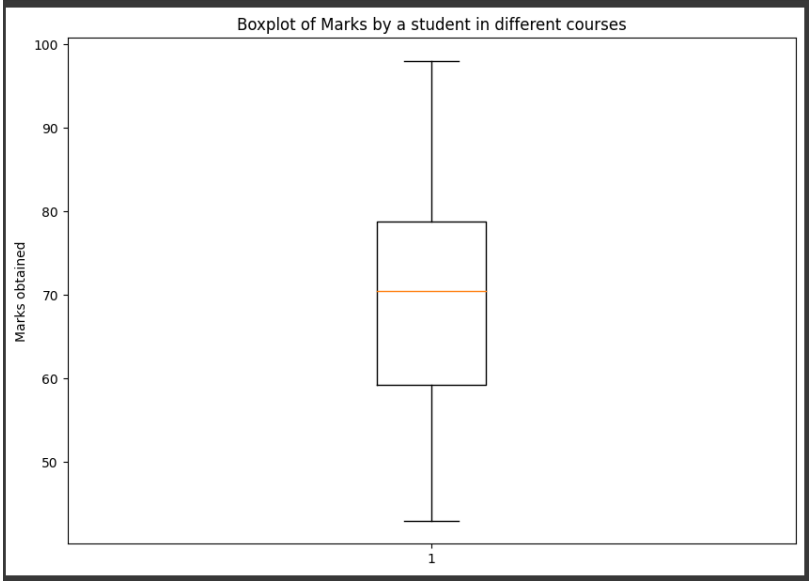
5. Scatter Plot:

```
plt.scatter(courses, marks, c="blue")
plt.xlabel("Courses")
plt.ylabel("Marks obtained")
plt.title("Marks by a student in different courses")
plt.show()
```



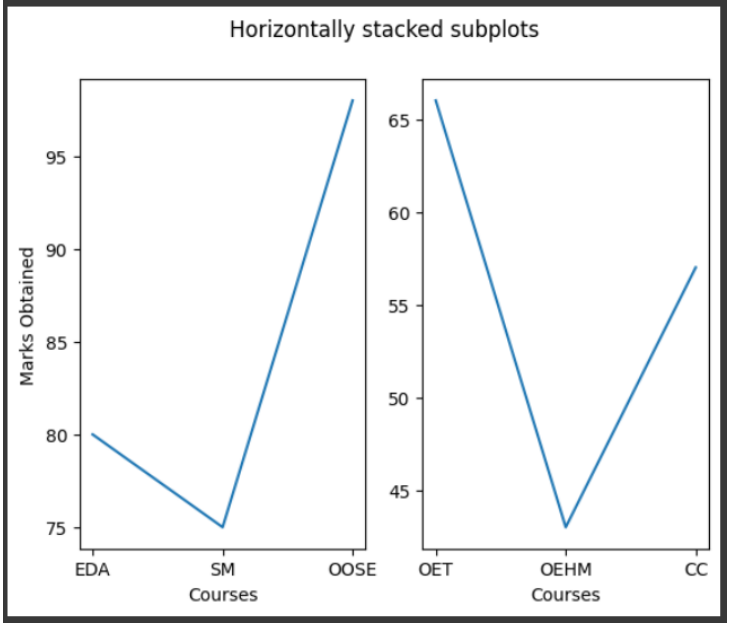
6. Quartile Plot (Boxplot):

```
fig=plt.figure(figsize=(10,7))
plt.boxplot(marks)
plt.ylabel("Marks obtained")
plt.title("Boxplot of Marks by a student in different courses")
plt.show()
```



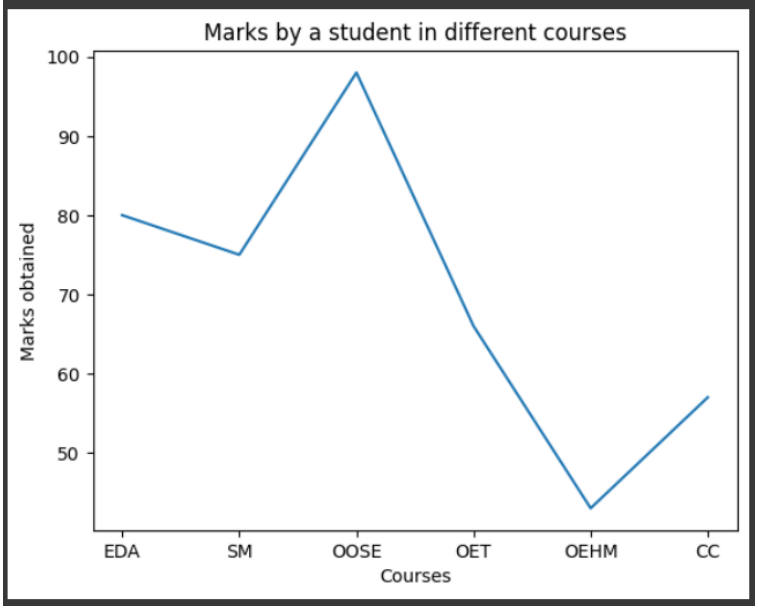
7. Subplots:

```
fig, (ax1, ax2)=plt.subplots(1, 2)
fig.suptitle('Horizontally stacked subplots')
ax1.plot(courses[:3], marks[:3])
ax1.set_xlabel("Courses")
ax1.set_ylabel("Marks Obtained")
ax2.plot(courses[3:], marks[3:])
ax2.set_xlabel("Courses")
plt.show()
```



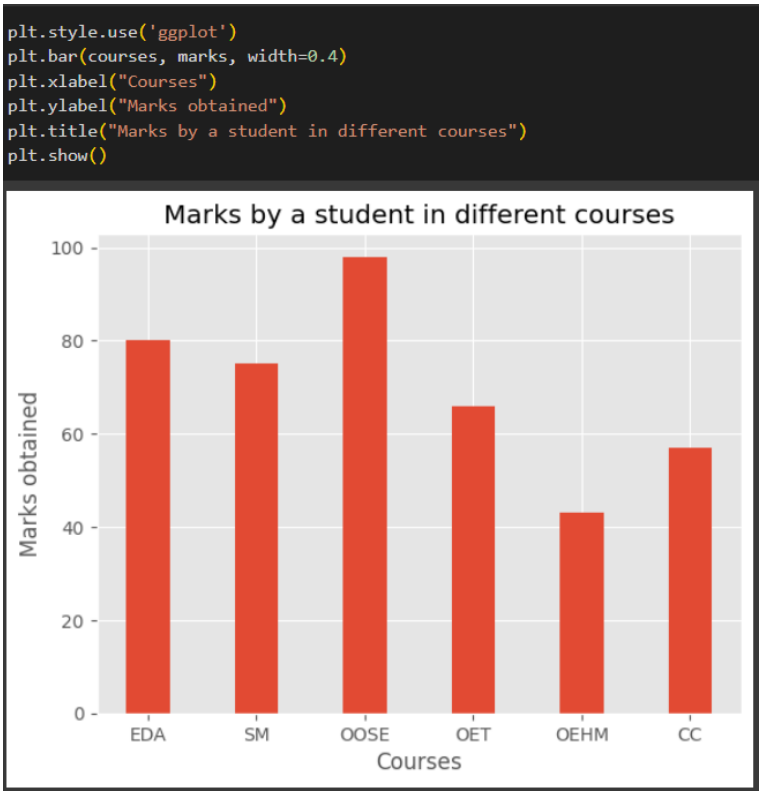
8. Line Chart:

```
plt.plot(courses, marks)
plt.xlabel("Courses")
plt.ylabel("Marks obtained")
plt.title("Marks by a student in different courses")
plt.show()
```



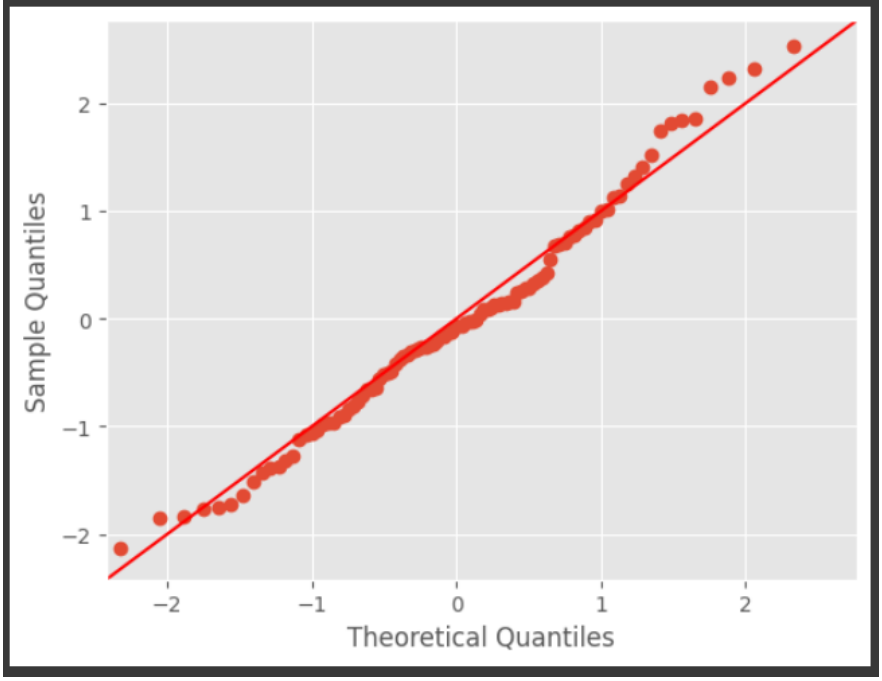
9. Ggplot:

```
plt.style.use('ggplot')
plt.bar(courses, marks, width=0.4)
plt.xlabel("Courses")
plt.ylabel("Marks obtained")
plt.title("Marks by a student in different courses")
plt.show()
```



10. QQ Plot:

```
import numpy as np
import statsmodels.api as sm
import pylab as py
data_points=np.random.normal(0, 1, 100)
sm.qqplot(data_points, line='45')
py.show()
```



Code:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm
import pylab as py
data = pd.read_csv(r'/Flight_delay.csv')
print("Dataset:\n",data)
column_data = data['AirTime']
column_data_1 = data['TaxiIn']
column_data_2 = data['ActualElapsedTime']
fig=plt.figure(figsize=(10,7))
plt.boxplot(column_data)
plt.ylabel("Air Time")
plt.title("Boxplot of Air Time")
plt.show()
data_points=np.random.normal(0, 1, 100)
sm.qqplot(data_points, line='45')
py.show()
fig, ax = plt.subplots(figsize=(10,7))
ax.hist(column_data_1, bins = [0, 5, 10, 15, 20, 25, 30, 35, 40])
plt.ylabel("Frequency")
plt.xlabel("TaxiIn")
plt.title("Frequency of Taxis In")
plt.show()
plt.scatter(column_data_2, column_data, c="blue")
plt.xlabel("Actual Elapsed Time")
plt.ylabel("Air Time")
plt.show()
```

Output:

Dataset:

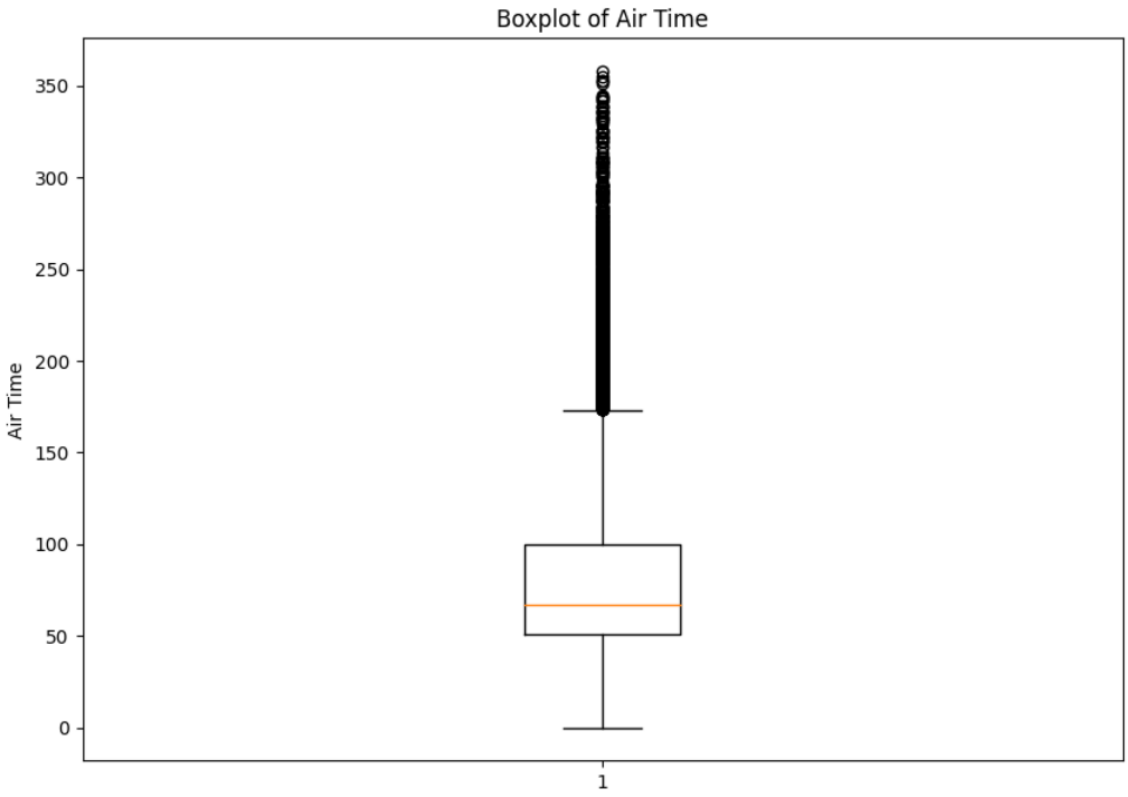
	DayOfWeek	Date	DepTime	ArrTime	CRSArrTime	UniqueCarrier	\
0	4	03-01-2019	1829	1959	1925	WN	
1	4	03-01-2019	1937	2037	1940	WN	
2	4	03-01-2019	1644	1845	1725	WN	
3	4	03-01-2019	1452	1640	1625	WN	
4	4	03-01-2019	1323	1526	1510	WN	
...	...	...	...	...	...	...	
22349	2	15-01-2019	1647	1831	1705	OO	
22350	2	15-01-2019	1738	1834	1710	OO	
22351	2	15-01-2019	1617	1723	1537	OO	
22352	2	15-01-2019	1308	1440	1313	OO	
22353	2	15-01-2019	1119	1229	1147	OO	

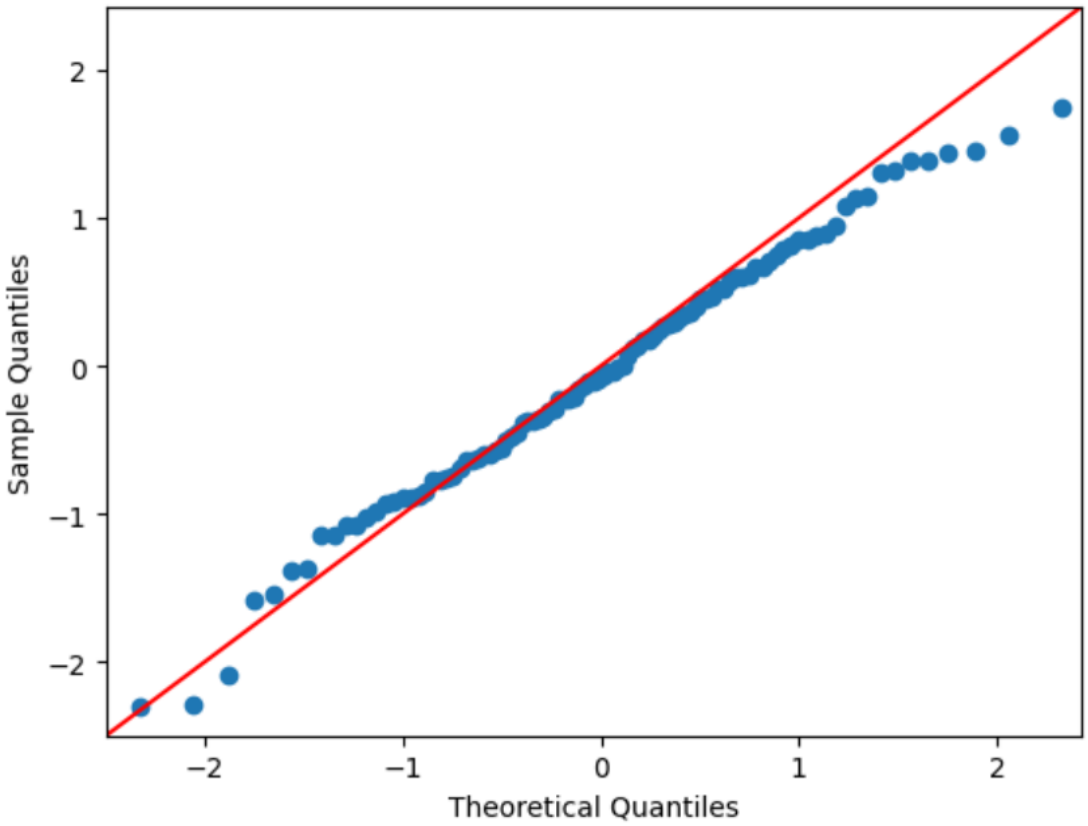
	Airline	FlightNum	TailNum	ActualElapsedTime	...	\
0	Southwest Airlines Co.	3920	N464WN	90	...	
1	Southwest Airlines Co.	509	N763SW	240	...	
2	Southwest Airlines Co.	1333	N334SW	121	...	
3	Southwest Airlines Co.	675	N286WN	228	...	
4	Southwest Airlines Co.	4	N674AA	123	...	
...	...	...	...	...	...	
22349	Skywest Airlines Inc.	5443	N220SW	104	...	
22350	Skywest Airlines Inc.	5445	N293SW	56	...	
22351	Skywest Airlines Inc.	5445	N293SW	66	...	
22352	Skywest Airlines Inc.	5446	N227SW	92	...	
22353	Skywest Airlines Inc.	5446	N227SW	70	...	

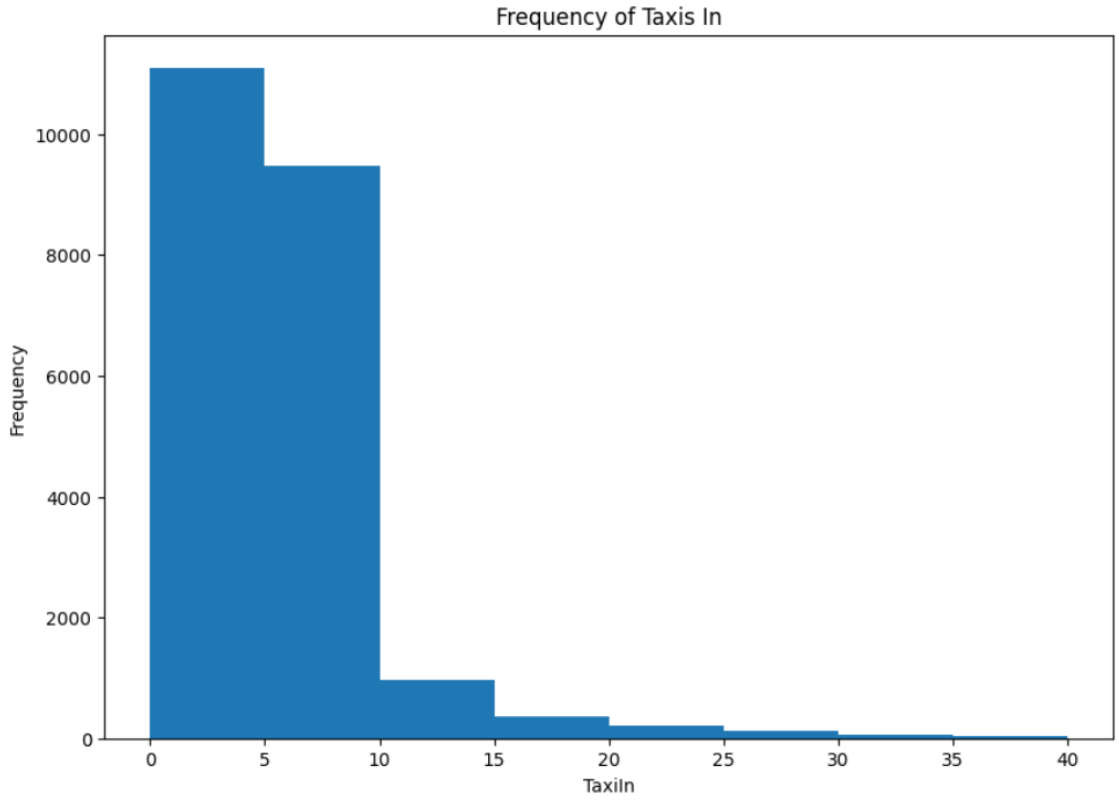
	TaxiIn	TaxiOut	Cancelled	CancellationCode	Diverted	CarrierDelay	\
0	3.0	10.0	0.0	N	0.0	2.0	
1	3.0	7.0	0.0	N	0.0	10.0	
2	6.0	8.0	0.0	N	0.0	8.0	
3	7.0	8.0	0.0	N	0.0	3.0	
4	4.0	9.0	0.0	N	0.0	0.0	
...	...	...	...	...	...	...	
22349	3.0	17.0	0.0	N	0.0	86.0	
22350	3.0	4.0	0.0	N	0.0	0.0	
22351	3.0	21.0	0.0	N	0.0	106.0	
22352	5.0	26.0	0.0	N	0.0	0.0	
22353	NaN	NaN	NaN	NaN	NaN	NaN	

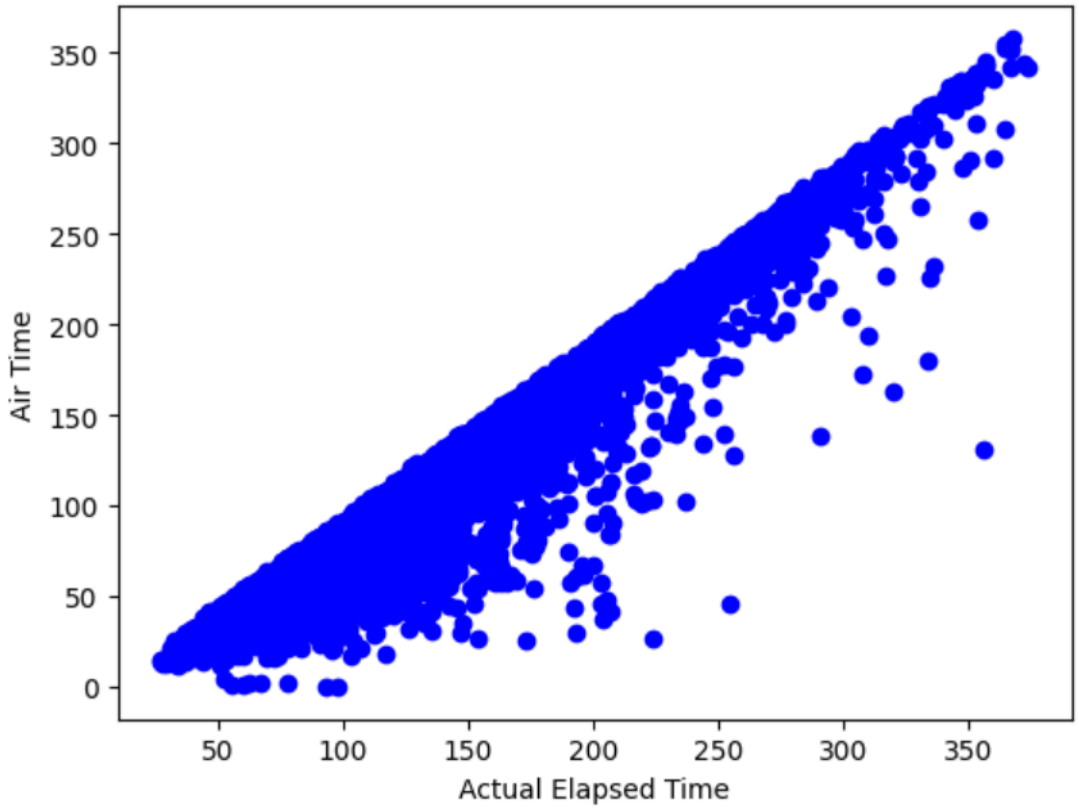
	WeatherDelay	NASDelay	SecurityDelay	LateAircraftDelay
0	0.0	0.0	0.0	32.0
1	0.0	0.0	0.0	47.0
2	0.0	0.0	0.0	72.0
3	0.0	0.0	0.0	12.0
4	0.0	0.0	0.0	16.0
...	...	...	...	...
22349	0.0	0.0	0.0	0.0
22350	0.0	84.0	0.0	0.0
22351	0.0	0.0	0.0	0.0
22352	0.0	87.0	0.0	0.0
22353	NaN	NaN	NaN	NaN

[22354 rows x 29 columns]


Boxplot of Air Time
Air Time
350
300
250
200
150
100
50
0
1


Sample Quantiles
Theoretical Quantiles
2
1
0
−1
−2
−2
−1
0
1
2


Frequency of Taxis In
Frequency
TaxiIn
0
2000
4000
6000
8000
10000
0
5
10
15
20
25
30
35
40


Air Time
Actual Elapsed Time
0
50
100
150
200
250
300
350
50
100
150
200
250
300
350

Post Lab Question-Answers:

1. Why is it important to measure the dispersion in the dataset?

Ans: Measuring dispersion in a dataset is important because it provides valuable information about the variability or spread of the data points. It helps us understand how the data points are distributed around the central tendency, such as the mean or median.

By measuring dispersion, we can assess the degree of variability within the dataset. This information is crucial for several reasons:

1. Understanding the spread: Dispersion measures, such as range, variance, or standard deviation, give us an idea of how far apart the data points are from each other. This helps us understand the range of values the data can take and how spread out they are.

2. Comparing datasets: Dispersion allows us to compare the variability between different datasets. For example, if we are comparing the performance of two groups, knowing the dispersion can help us determine if one group has more consistent results than the other.

3. Identifying outliers: Dispersion measures can help identify outliers, which are data points that significantly deviate from the rest of the dataset. Outliers can provide valuable insights or indicate errors in data collection.

4. Decision-making: Dispersion measures provide insights into the reliability and stability of the data. This information is crucial for making informed decisions. For example, if the data points are highly dispersed, it may indicate a higher level of uncertainty or risk.

Overall, measuring dispersion in a dataset helps us gain a deeper understanding of the data, make comparisons, identify patterns, and make more informed decisions based on the variability and spread of the data points.

2. Discuss the other purposes/advantages of the plots used in this experiment.

Ans: In addition to measuring dispersion, plots are widely used in data analysis for several purposes and offer various advantages. Here are some of the key purposes and advantages of using plots in experiments:

1. Visualizing patterns and trends: Plots provide a visual representation of the data, allowing us to identify patterns, trends, and relationships that may not be apparent from raw data alone. By plotting the data points, we can observe the overall shape of the distribution, detect clusters or gaps, and understand the general behavior of the data.

2. Assessing distribution: Plots such as histograms, box plots, or density plots help us understand the distribution of the data. They provide insights into the shape (e.g., symmetric, skewed), central tendency, and spread of the data. This information is crucial for selecting appropriate statistical methods and making accurate interpretations.

3. Identifying outliers: Plots can effectively highlight outliers, which are data points that deviate significantly from the rest of the dataset. Outliers may indicate errors, anomalies, or interesting

phenomena that require further investigation. Scatter plots, box plots, or modified z-scores are commonly used to identify outliers visually.

4. Comparing groups or conditions: Plots allow for easy visual comparison of different groups or conditions in an experiment. By plotting multiple datasets on the same graph, we can observe differences in central tendency, dispersion, or distribution between groups. This visual comparison aids in drawing meaningful conclusions and identifying significant differences.

5. Communicating results: Plots are powerful tools for effectively communicating research findings to others. They provide a concise and intuitive way to present complex data and statistical analyses. Well-designed plots can enhance the clarity and impact of research presentations, reports, and publications.

6. Exploratory data analysis: Plots play a crucial role in exploratory data analysis, where the goal is to gain insights, generate hypotheses, and identify interesting patterns or relationships. By visually exploring the data through plots, researchers can generate new ideas, formulate research questions, and guide further analysis.

7. Model diagnostics: Plots are used to assess the goodness-of-fit and assumptions of statistical models. Residual plots, Q-Q plots, and diagnostic plots help identify potential model violations, such as heteroscedasticity, nonlinearity, or outliers. These plots aid in refining models and improving the accuracy of predictions.

In summary, plots serve multiple purposes in data analysis, including visualizing patterns, assessing distribution, identifying outliers, comparing groups, communicating results, aiding exploratory analysis, and diagnosing statistical models. They provide a powerful visual representation of data, enabling researchers to gain insights, make informed decisions, and effectively communicate their findings.

Outcomes: Comprehend descriptive and proximity measures of data

Conclusion (based on the Results and outcomes achieved):

The experiment on applying and interpreting different plots in data analysis yielded positive outcomes. The use of plots proved to be highly advantageous, providing valuable insights, aiding in exploratory analysis, facilitating communication, and guiding decision-making. The experiment reinforced the importance of utilizing plots as a fundamental tool in data analysis, enabling researchers to gain a deeper understanding of the data and make informed conclusions.

References:

Books/ Journals/ Websites

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition

---