EXPERIMENT NO: 3

AIM: Explore data visualization techniques.

THEORY:

Data Visualization is the graphical representation of data and information. It uses visual elements like charts, graphs, and maps to provide an accessible way to understand trends, outliers, and patterns in data. The primary goal of data visualization is to communicate information clearly and effectively through graphical means.

Techniques Used in the Tasks:

Feature Types: In any dataset, features can be of different types such as:

Numeric: Features with numerical values that can be measured and quantified.

Nominal: Features that represent categories or labels without any inherent order.

Histograms:

A histogram is a graphical representation of the distribution of numerical data.

It consists of a series of bars, where the height of each bar represents the frequency of data within a specified range (bin).

Histograms are useful for understanding the central tendency, variability, and shape of a dataset's distribution.

Boxplots:

A boxplot (box-and-whisker plot) is a method for graphically depicting groups of numerical data through their quartiles.

It displays the distribution of data based on a five-number summary: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.

Boxplots are helpful in identifying outliers, understanding the spread of data, and comparing distributions between different groups.

Pairplots:

A pairplot is a type of plot in seaborn library that creates a matrix of scatterplots and histograms for each pair of features in a dataset.

It allows us to visualize the relationships between multiple variables simultaneously. Pairplots are useful for identifying patterns, correlations, and outliers in multidimensional datasets.

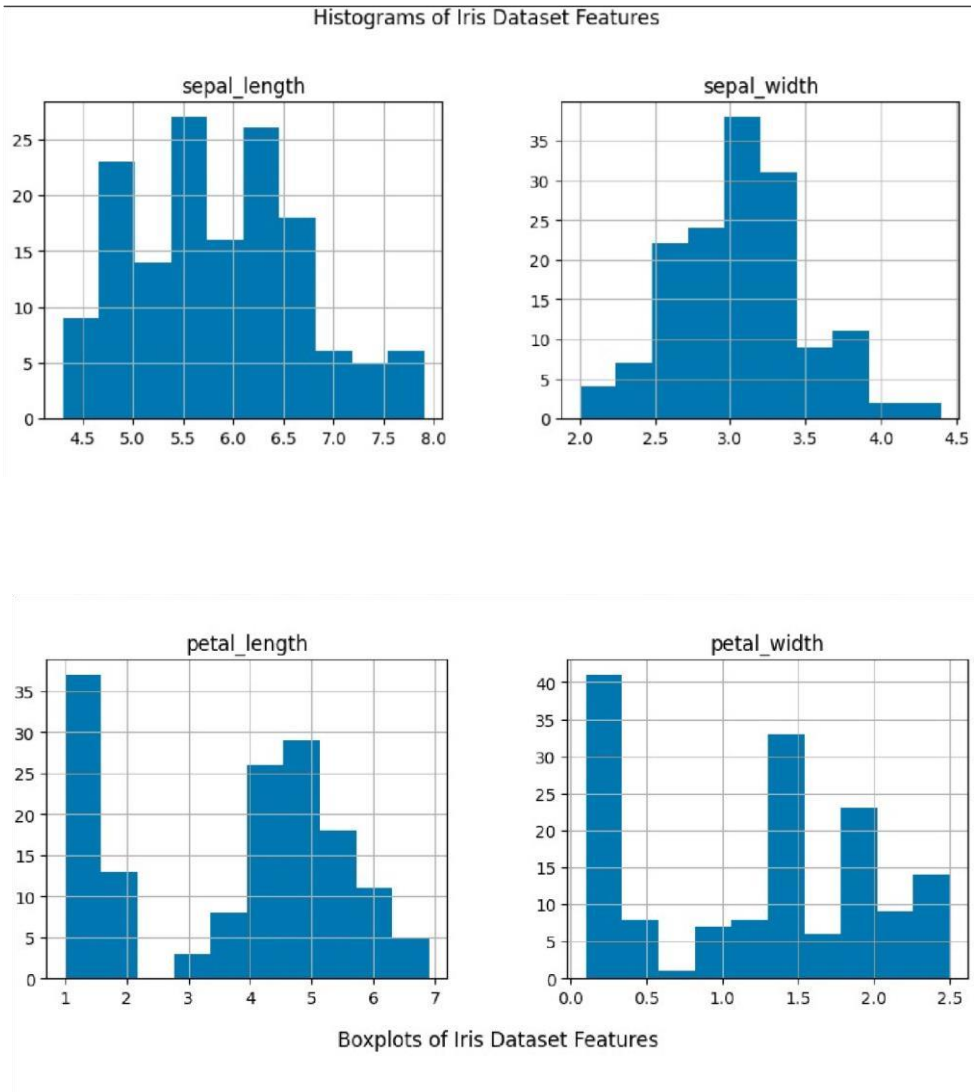
CODE:

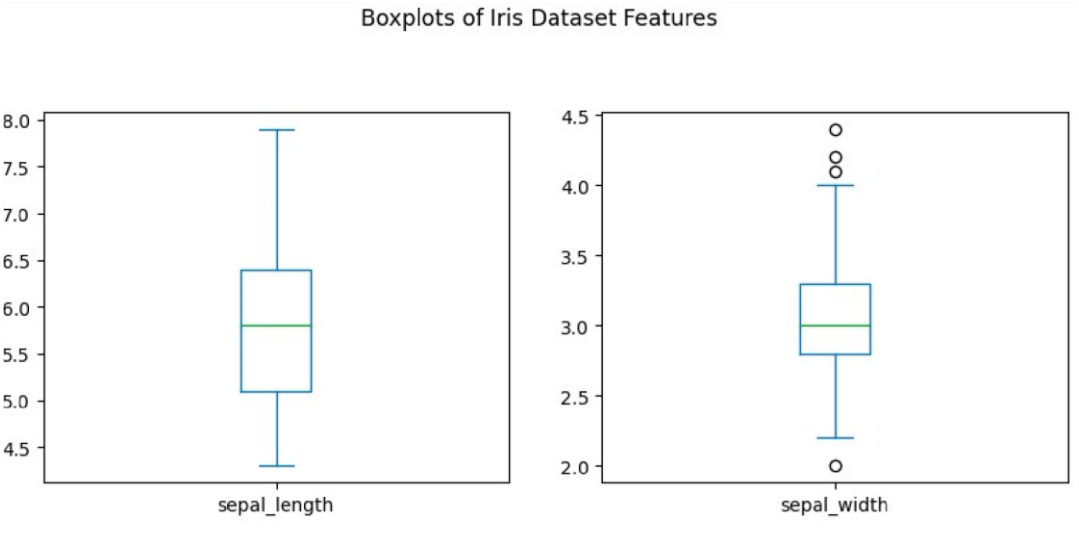
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load Iris dataset
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
iris_columns = ["sepal_length", "sepal_width", "petal_length", "petal_width", "class"]
iris_df = pd.read_csv(url, names=iris_columns)

# 1. List down the features and their types
features_and_types = iris_df.dtypes
print(features_and_types)

# 2. Create a histogram for each feature
iris_df.hist(figsize=(10, 8))
plt.suptitle("Histograms of Iris Dataset Features")
plt.show()

# 3. Create a boxplot for each feature
iris_df.plot(kind='box', subplots=True, layout=(2, 2), figsize=(10, 8))
plt.suptitle("Boxplots of Iris Dataset Features") plt.show()

# 4. Compare distributions and identify outliers
# We can identify outliers by observing points outside the whiskers of the boxplots.
# Now let's move to the Titanic dataset
# Load Titanic dataset titanic_df = sns.load_dataset('titanic')

# 5. Use Seaborn to find any patterns in the data
sns.pairplot(titanic_df, hue='survived',
diag_kind='hist') plt.title("Pairplot of Titanic Dataset")
plt.show()

# 6. Plot a histogram for ticket price distribution
plt.figure(figsize=(8, 6))
sns.histplot(titanic_df['fare'], bins=30,
kde=True) plt.title("Histogram of Ticket Price Distribution")
plt.xlabel("Fare")
plt.ylabel("Frequency") plt.show()

# 7. Plot a box plot for distribution of age with respect to each gender and survival
plt.figure(figsize=(10, 8))
sns.boxplot(data=titanic_df, x='sex', y='age', hue='survived')
plt.title("Distribution of Age with respect to Gender and
```

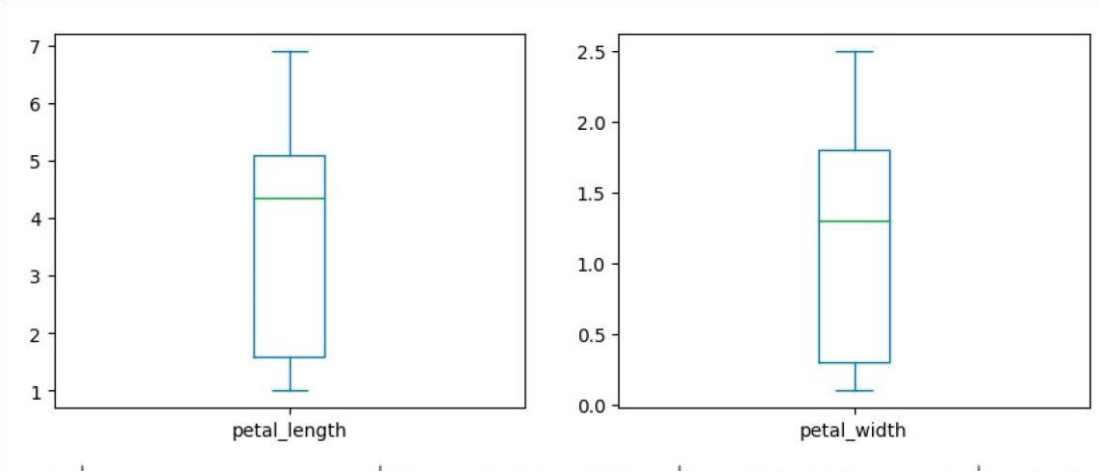
Survival") plt.xlabel("Gender") plt.ylabel("Age") plt.show()

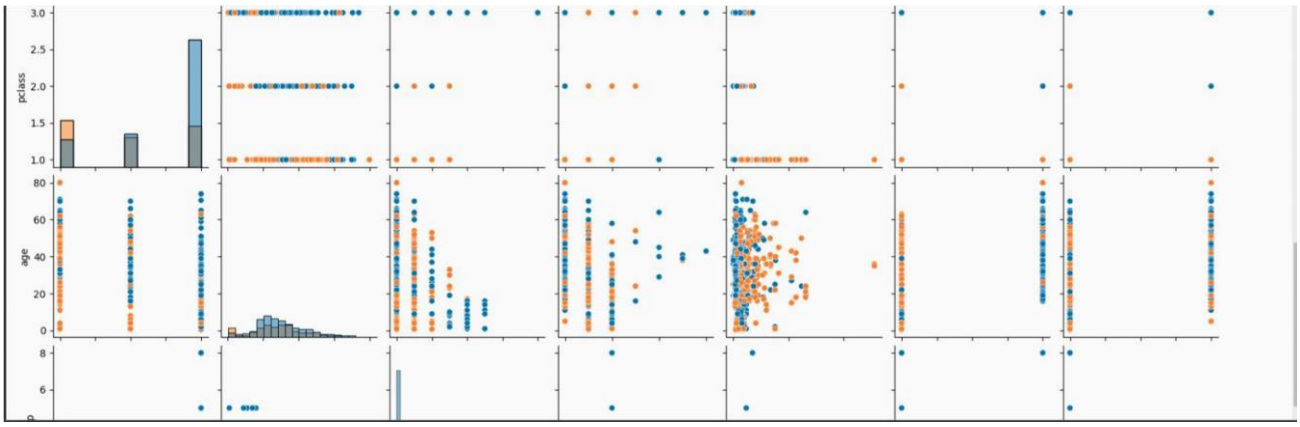
OUTPUT:

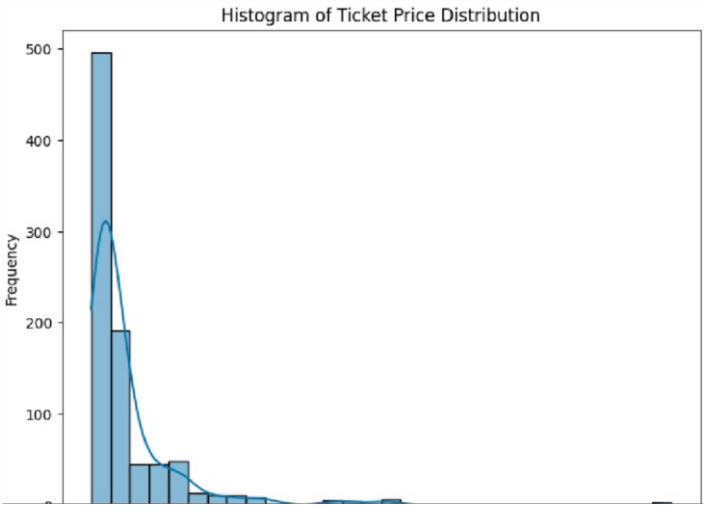
sepal_length float64
sepal_width float64
petal_length float64
petal_width float64
class object dtype:
object

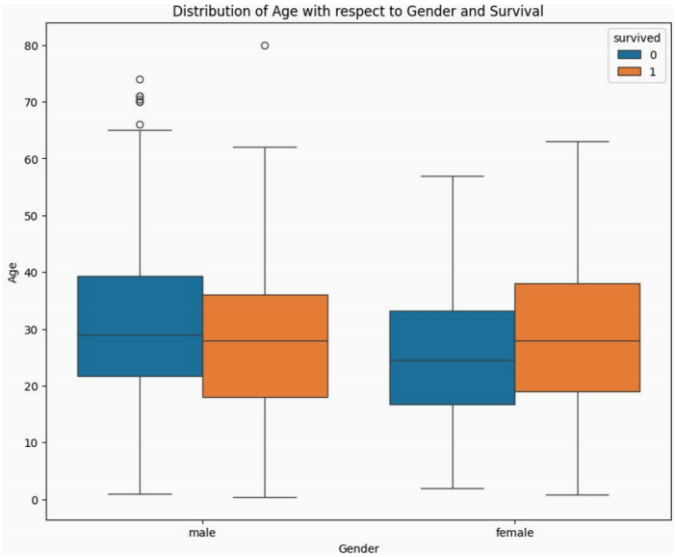



Boxplots of Iris Dataset Features
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
sepal_length
4.5
4.0
3.5
3.0
2.5
2.0
sepal_width


7
6
5
4
3
2
1
petal_length
2.5
2.0
1.5
1.0
0.5
0.0
petal_width


pclass
age





CONCLUSION:

In this experiment, data visualization techniques were employed to analyze both the Iris flower and Titanic datasets. Through histograms, boxplots, and pairplots, we delved into the distributions and relationships within the data. From understanding feature variability in the Iris dataset to exploring survival outcomes in the Titanic dataset, visualization provided crucial insights for informed analysis and decision-making.