Experiment: 3.3

Student Name: Nishant Kumar Mehta
UID: 21BCS-*3402*
Branch: CSE
Section/Group: 21BCS-IOT-602B
Semester: 5th
Date: 02/11/23
Subject Name: AIML Lab
Subject Code: 21CSH-316

1. AIM: Implement Exploratory Data Analysis on any data set.

2. Objective:

To Learn about Meta-data.

3. Tools/Resource Used:

1. Python programming language.
2. Jupyter Notebook.

4. Algorithm:

- Import libraries: Use pandas, numpy, and data visualization tools. Load dataset.
- Display initial data overview.
- Check and handle missing values and duplicates.
- Explore data through univariate and bivariate analysis. Visualize correlations between numeric variables.
- Detect and address outliers if needed.
- Summarize findings and plan next steps.

5. Program Code:

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

data = {
    'student_id': range(1, 11),
```

```
    'age': [18, 19, 20, 22, 21, 20, 19, 18, 23, 22],
    'gender': ['Male', 'Female', 'Male', 'Male', 'Female', 'Male', 'Female', 'Male', 'Female', 'Male'],
    'study_hours': [4, 6, 5, 3, 7, 6, 5, 4, 8, 7],
    'test_scores': [85, 92, 78, 88, 96, 79, 90, 84, 93, 87]
}

df = pd.DataFrame(data)

print(df.head())

summary = df.describe()
print(summary)

missing_values = df.isnull().sum()
print(missing_values)

duplicates = df.duplicated().sum()
print("Number of duplicate rows:", duplicates)

# Remove duplicates if present
df = df.drop_duplicates()

# Example histogram for age
plt.figure(figsize=(8, 6))
sns.histplot(df['age'], kde=True)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.show()

# Example scatter plot for study hours vs. test
scores plt.figure(figsize=(8, 6))
sns.scatterplot(data=df, x='study_hours', y='test_scores')
plt.title('Study Hours vs. Test Scores') plt.xlabel('Study
Hours')
plt.ylabel('Test
Scores') plt.show()

# Example count plot for gender
plt.figure(figsize=(8, 6))
```

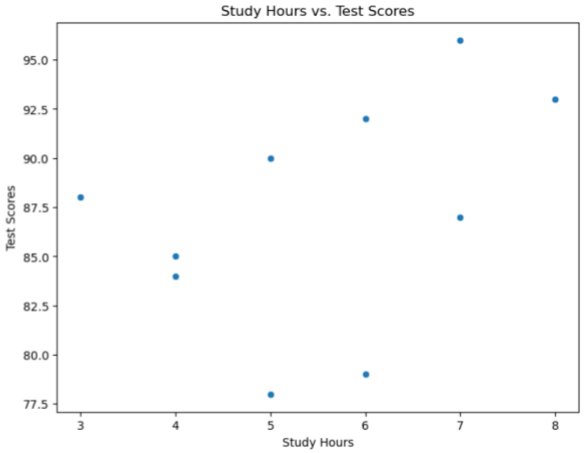
```
sns.countplot(data=df, x='gender')
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()

correlation_matrix = df.corr()
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

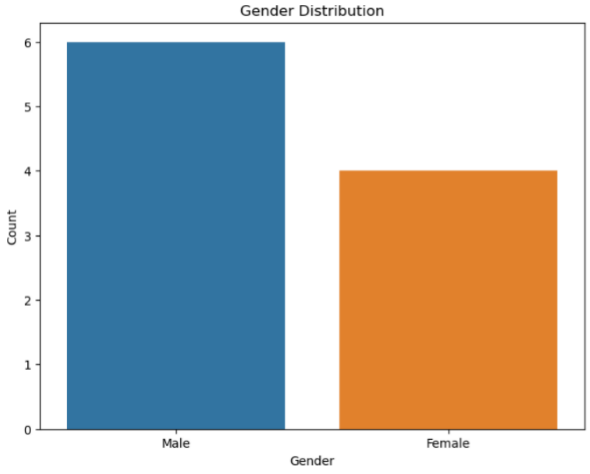
6. Output/Result:

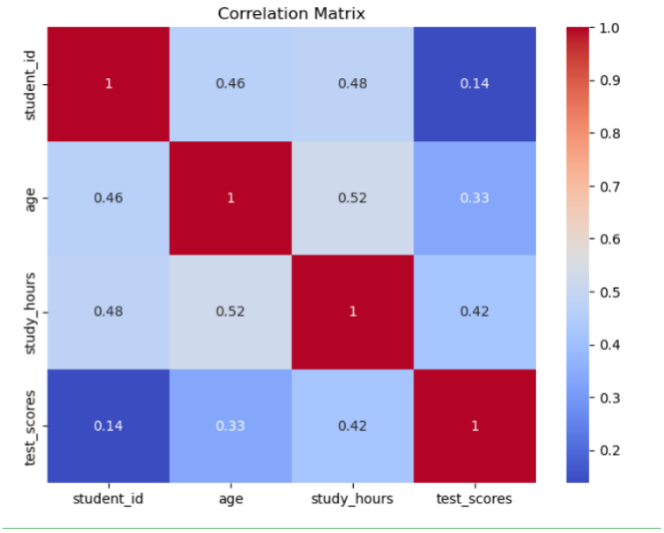
	student_id	age	gender	study_hours	test_scores
0	1	18	Male	4	85
1	2	19	Female	6	92
2	3	20	Male	5	78
3	4	22	Male	3	88
4	5	21	Female	7	96

	student_id	age	study_hours	test_scores
count	10.00000	10.00000	10.000000	10.000000
mean	5.50000	20.20000	5.500000	87.200000
std	3.02765	1.75119	1.581139	5.865151
min	1.00000	18.00000	3.000000	78.000000
25%	3.25000	19.00000	4.250000	84.250000
50%	5.50000	20.00000	5.500000	87.500000
75%	7.75000	21.75000	6.750000	91.500000
max	10.00000	23.00000	8.000000	96.000000

```
student_id     0
age            0
gender         0
study_hours    0
test_scores    0
dtype: int64
Number of duplicate rows: 0
```







7. Learning Outcomes:

1. *Implement to implement different python library.*
2. *Understand the concept of EDA process.*