**Experiment: 3.3**

**Student Name:** SANJIV GUPTA **UID:** 21BCS-*3478*

**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.*

1. **Tools/Resource Used:**

*1. Python programming language.*

*2. Jupyter Notebook.*

1. **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.*

1. **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()*

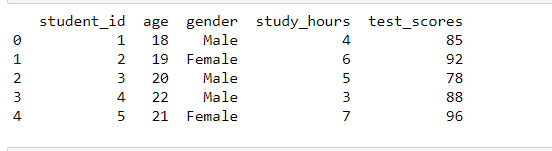
*plt.figure(figsize=(8, 6))*

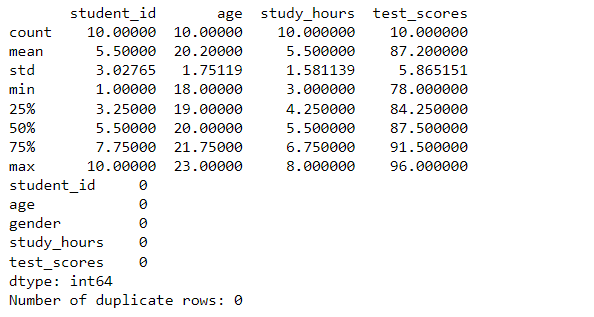
*sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm')*

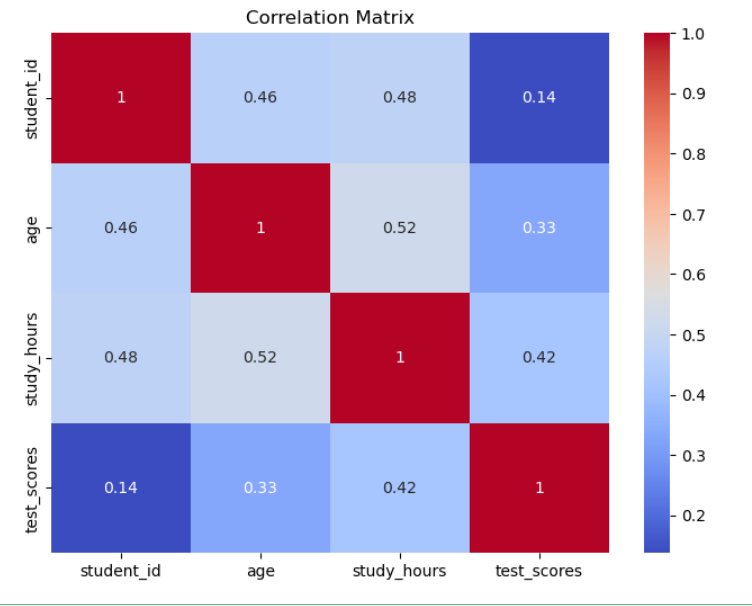
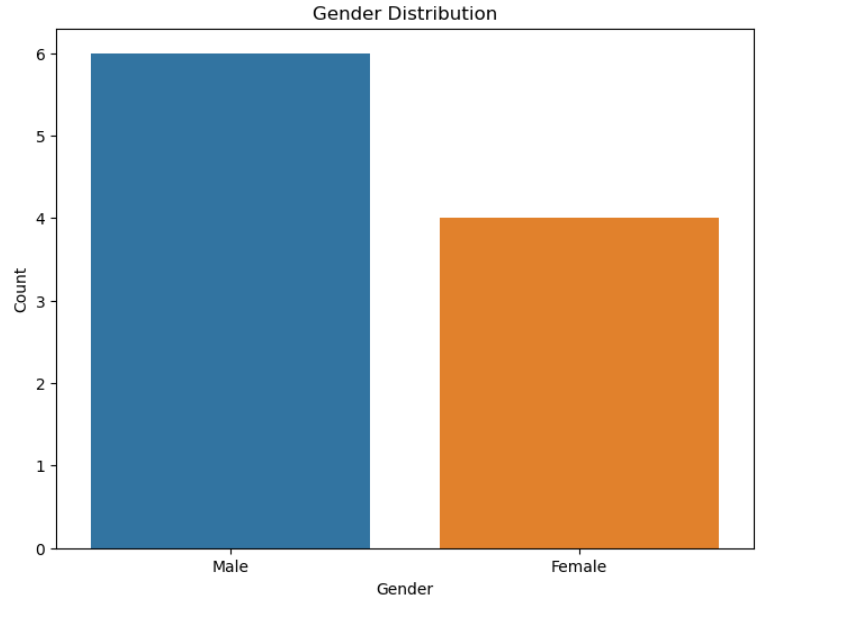
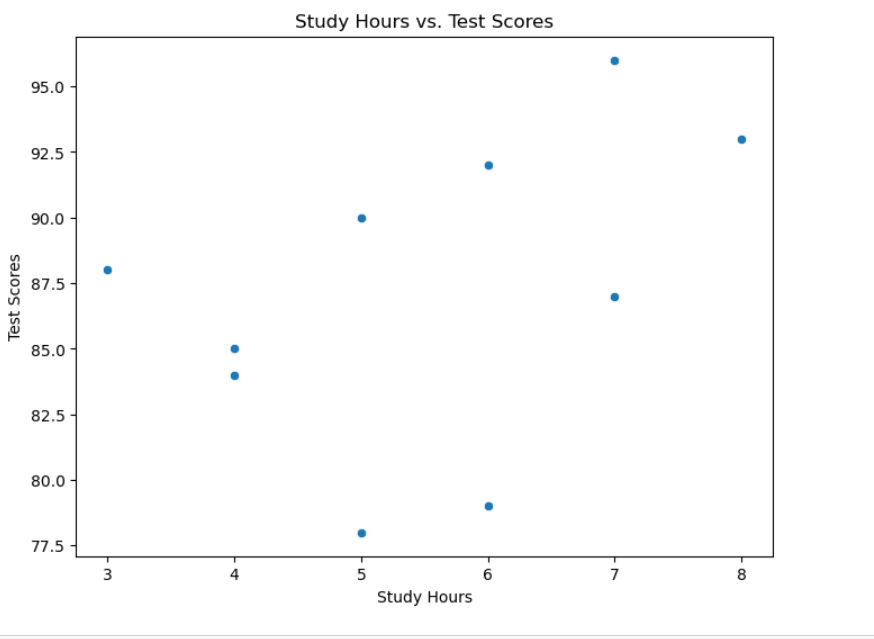
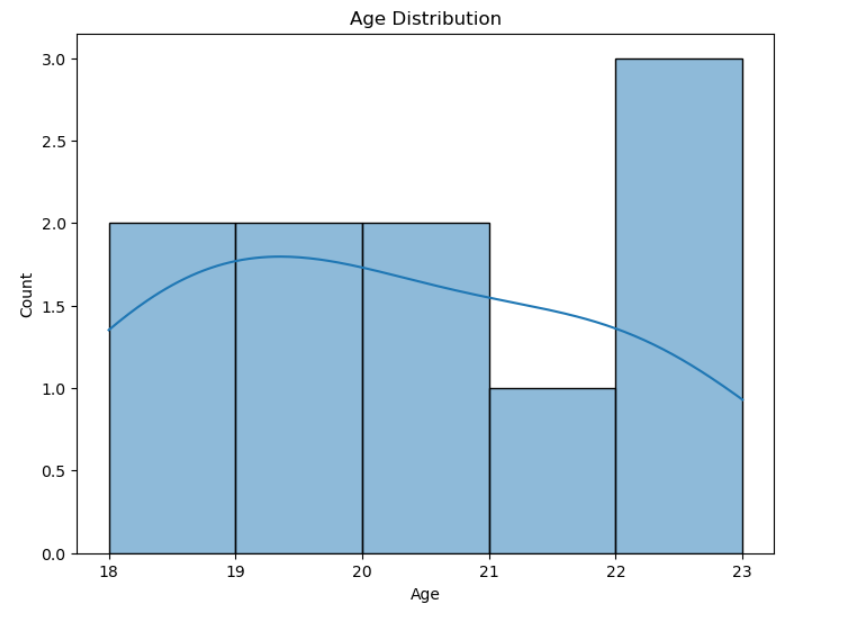
*plt.title('Correlation Matrix')*

*plt.show()*

1. **Output/Result:**

****

****

****

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