EXP NO: 5	EDA – DATA VISUALIZATION WITH MATPLOTLIB

AIM

The Python code aims to perform exploratory data analysis (EDA) by applying preprocessing steps and creating visualizations with Matplotlib. This helps to identify trends, compare group statistics, and observe data distributions using line charts, bar charts, and histograms.

PROBLEM STATEMENT

Raw datasets often contain large amounts of information that are not immediately meaningful. Without proper preprocessing and exploratory data analysis (EDA. Visualization techniques such as line charts, bar charts, and histograms help in summarizing the data and gaining insights.

ALGORITHM

Step 1: Import pandas for data handling, matplotlib for visualization, and sklearn scalers for preprocessing.

Step 2: Read the StudentsPerformance.csv dataset into a Pandas DataFrame.

Step 3: Display the first few rows of the dataset using df.head() to understand its structure.

Step 4: Group data by reading score and plot average math scores.

Step 5: Plot separate lines for categories (e.g., gender) for comparison.

Step 6: Group data by gender and calculate the average writing score.

Step 7: Plot a bar chart to compare gender-based averages.

Step 8: Plot the distribution of math scores to observe frequency patterns.

Step 9: Apply StandardScaler or MinMaxScaler for feature normalization if needed for further analysis.

Step 10: Analyze visualizations to identify trends, relationships, and score distributions.

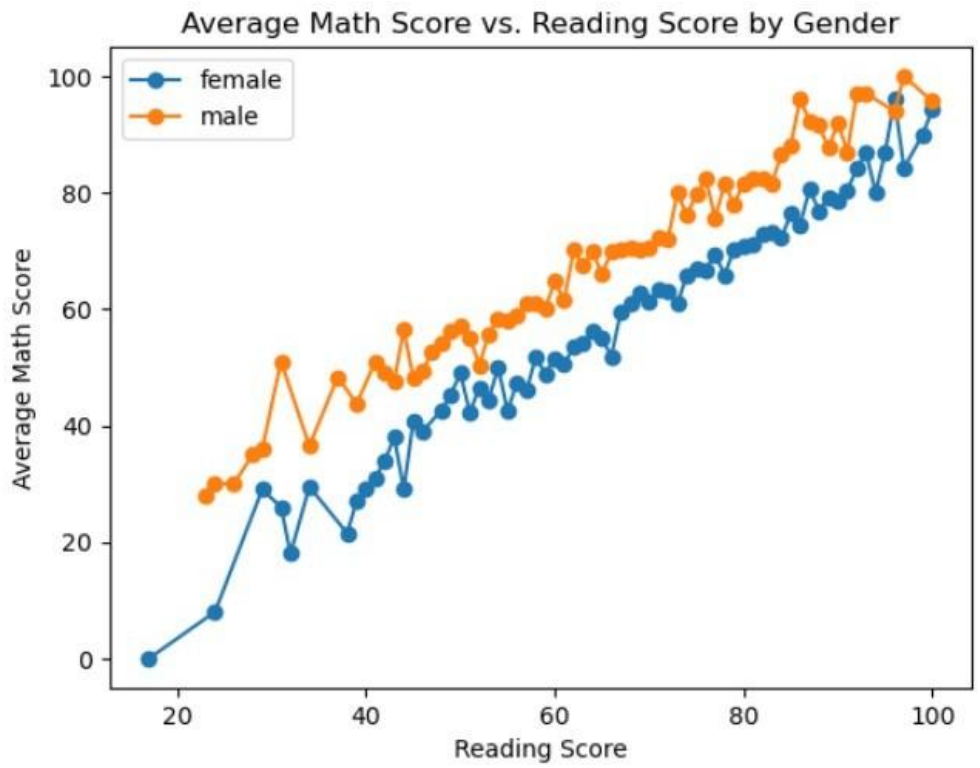
SAMPLE CODE

```
import pandas as pd
from sklearn.preprocessing import StandardScaler, MinMaxScaler
```

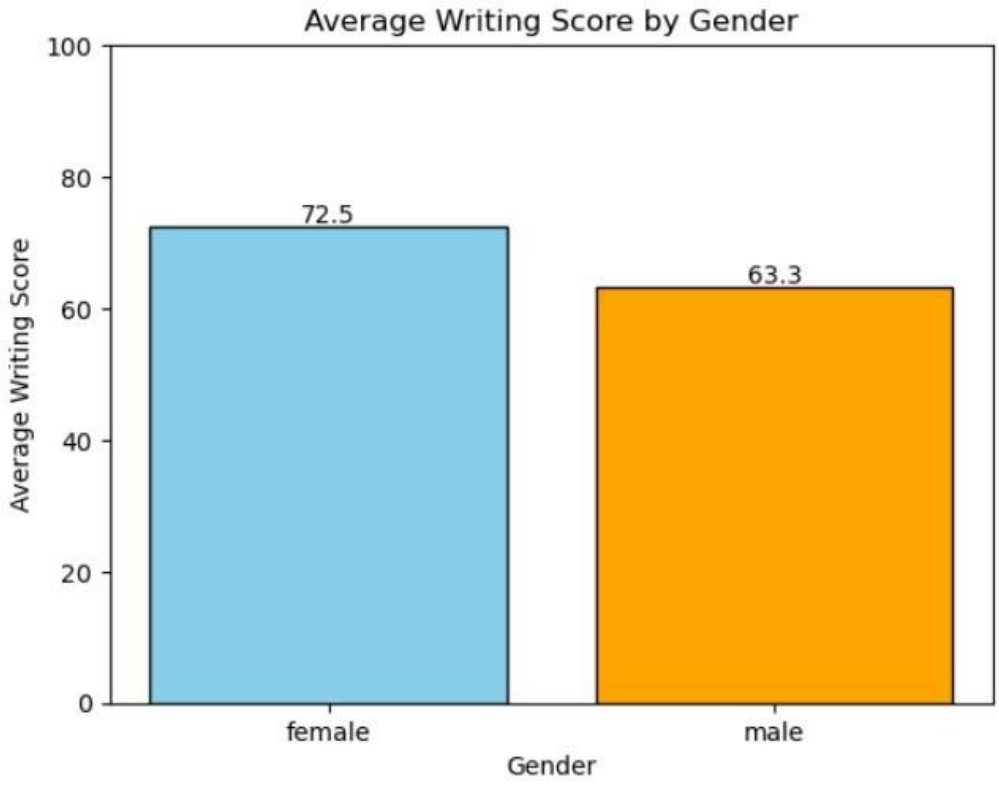
```
import matplotlib.pyplot as plt


# Step 1: Load dataset
df = pd.read_csv('StudentsPerformance.csv')
df.head()
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

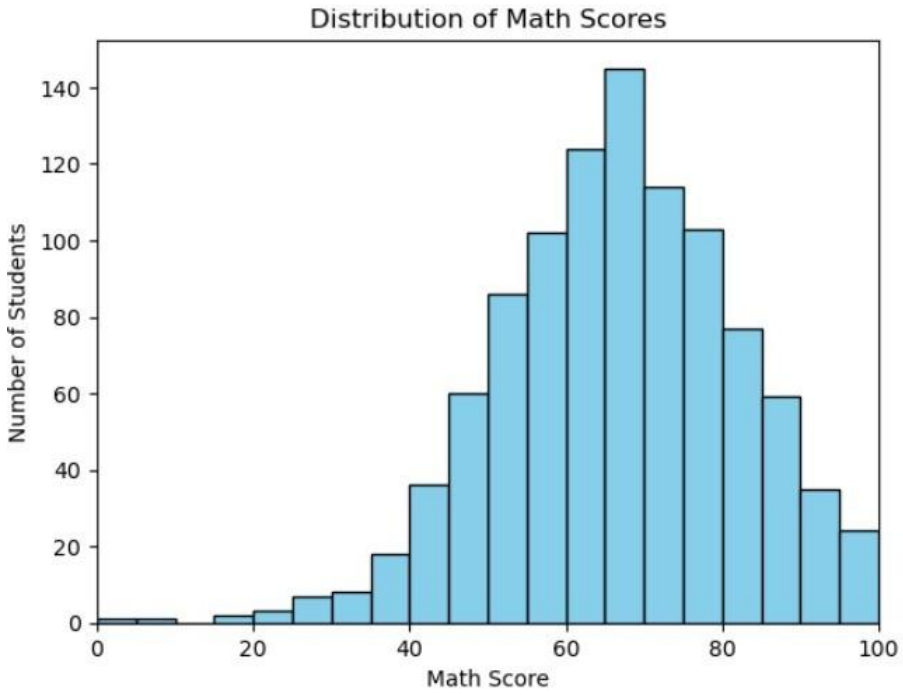
```
# Step 2: Line Chart - Average math score across reading score levels by gender
for gender in df["gender"].unique():
    avg_scores = df[df["gender"] == gender].groupby("reading score")["math score"].mean()
    plt.plot(avg_scores.index, avg_scores.values, marker='o', label=gender)

plt.title("Average Math Score vs. Reading Score by Gender")
plt.xlabel("Reading Score")
plt.ylabel("Average Math Score")
plt.legend()
plt.show()
```




```
# Step 3: Bar Chart - Average writing score by gender
avg_writing = df.groupby("gender")["writing score"].mean()

plt.bar(avg_writing.index, avg_writing.values,
        color=['skyblue', 'orange'], edgecolor='black')

# Add values on top of bars
for i, val in enumerate(avg_writing.values):
    plt.text(i, val + 0.5, round(val, 1), ha='center', fontsize=10)

plt.title("Average Writing Score by Gender")
plt.xlabel("Gender")
plt.ylabel("Average Writing Score")
plt.ylim(0, 100)  # keep y-axis within score range
plt.show()
```




```
# Step 4: Histogram - Distribution of math scores
plt.hist(df["math score"], bins=20, edgecolor='black', color='skyblue')
plt.title("Distribution of Math Scores")
plt.xlabel("Math Score")
plt.ylabel("Number of Students")
plt.xlim(0, 100)  # since scores are between 0–100
plt.show()
```



RESULT:

Thus, the EDA with data visualization with matplotlib was done using line, bar, and histogram charts.