Student Performance Prediction – Predict exam scores based on study hours and other factors

Name: [Aditi Sharma]
Roll Number: [202401100400013]

Introduction

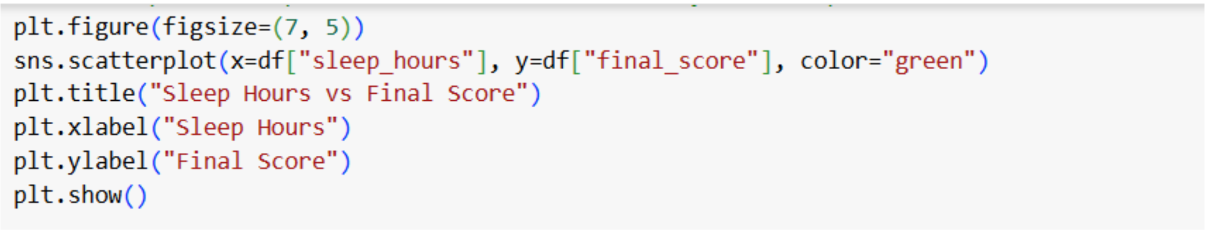
The aim of this project is to predict student performance based on study hours, previous exam scores, and other factors. Using data analysis and visualization techniques, we explore relationships between variables and build a model to predict exam performance.

Methodology

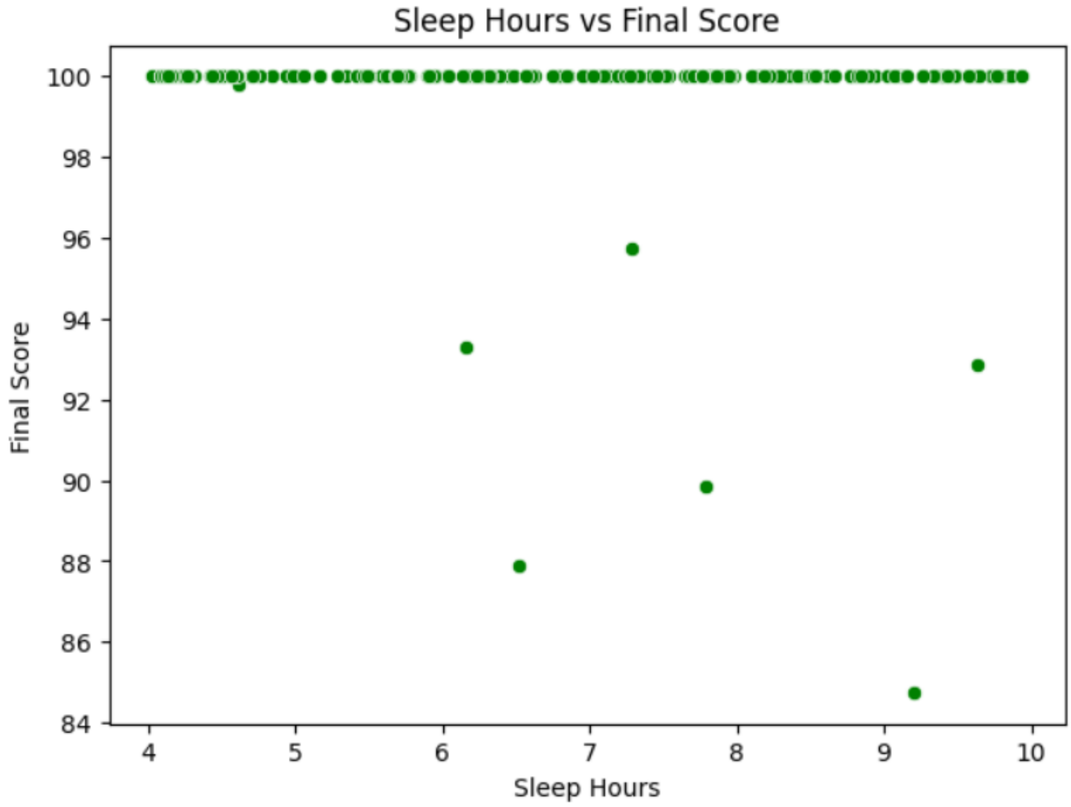
1. **Data Collection:** The student performance dataset was provided and contains features like study hours, sleep hours, extracurricular activity, and previous exam scores.
2. **Data Preprocessing:** The dataset was loaded and checked for missing values and data consistency.

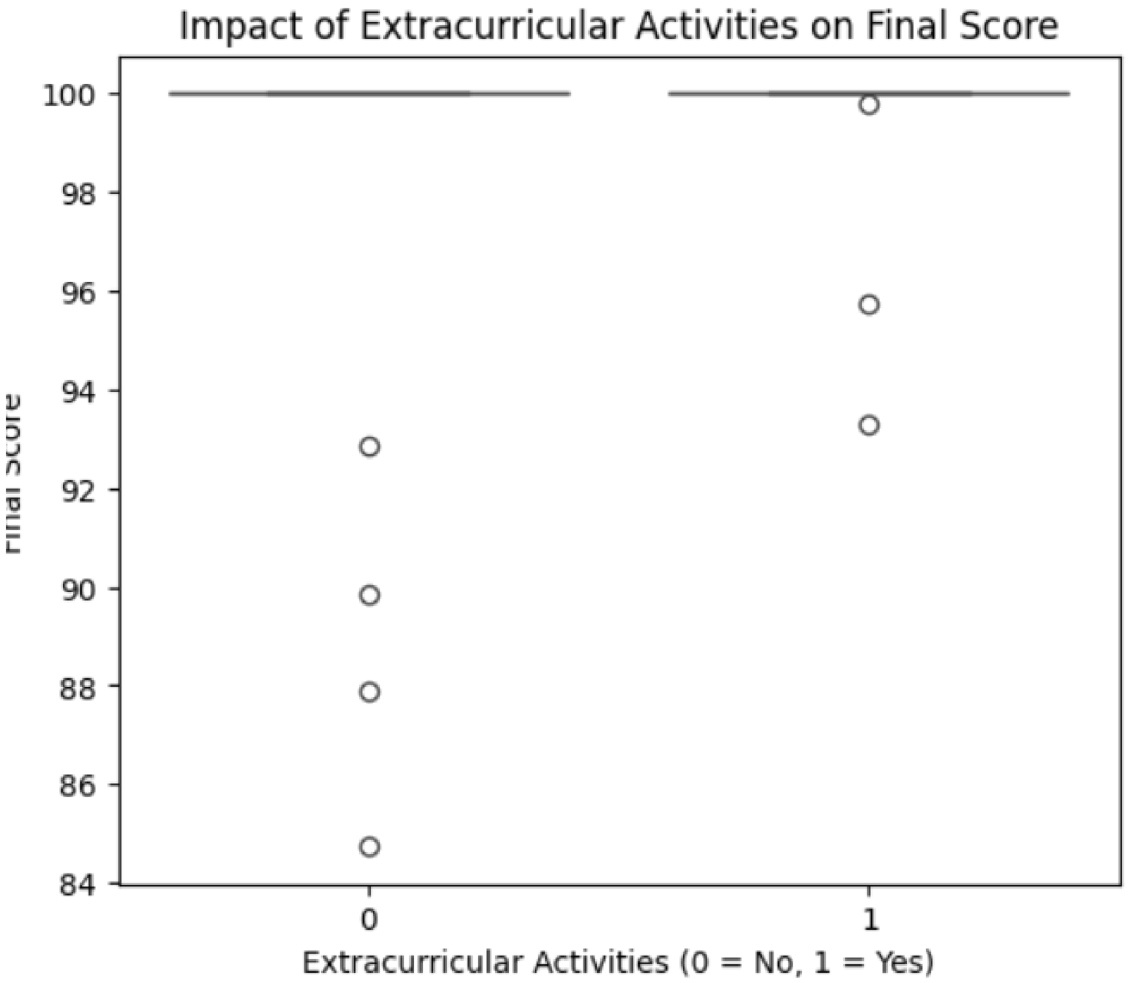
3. **Exploratory Data Analysis (EDA):** We visualized the distribution of scores and the relationships between variables using histograms, scatter plots, and boxplots.
4. **Model Selection:** A linear regression model was chosen to predict student exam scores based on study hours and other variables.
5. **Training and Testing:** The dataset was split into training (80%) and testing (20%) sets. The model was trained and evaluated for performance.

CODE:

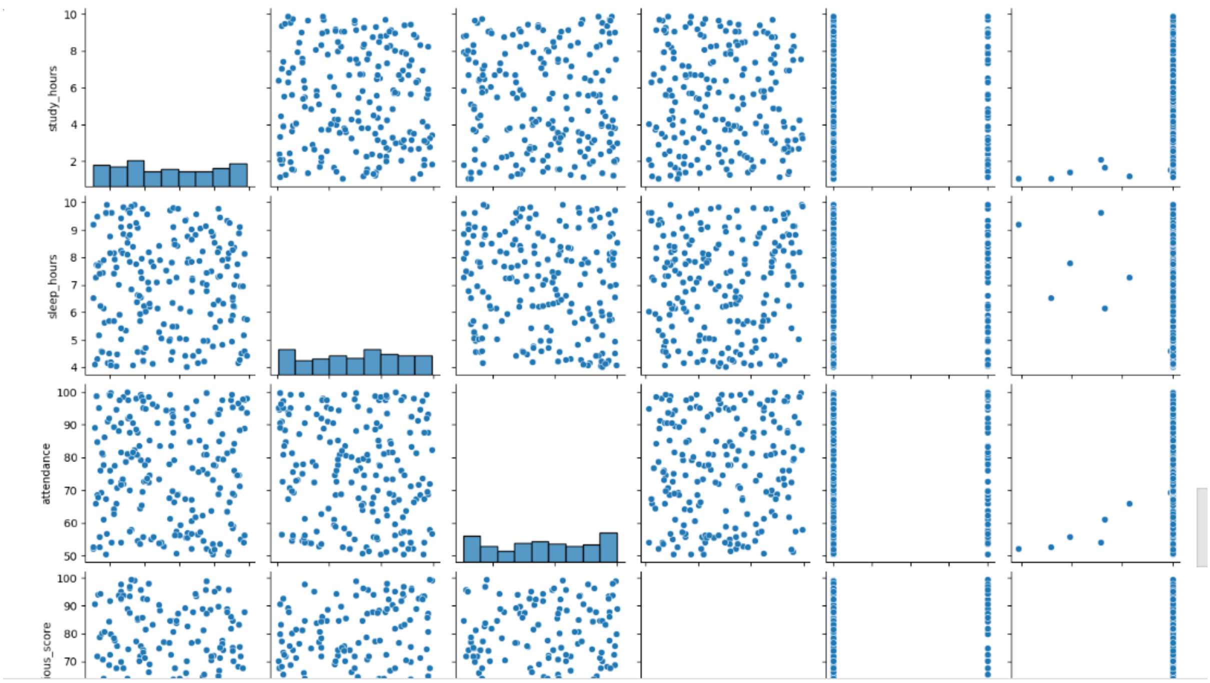


```
plt.figure(figsize=(7, 5))
sns.scatterplot(x=df["sleep_hours"], y=df["final_score"], color="green")
plt.title("Sleep Hours vs Final Score")
plt.xlabel("Sleep Hours")
plt.ylabel("Final Score")
plt.show()
```

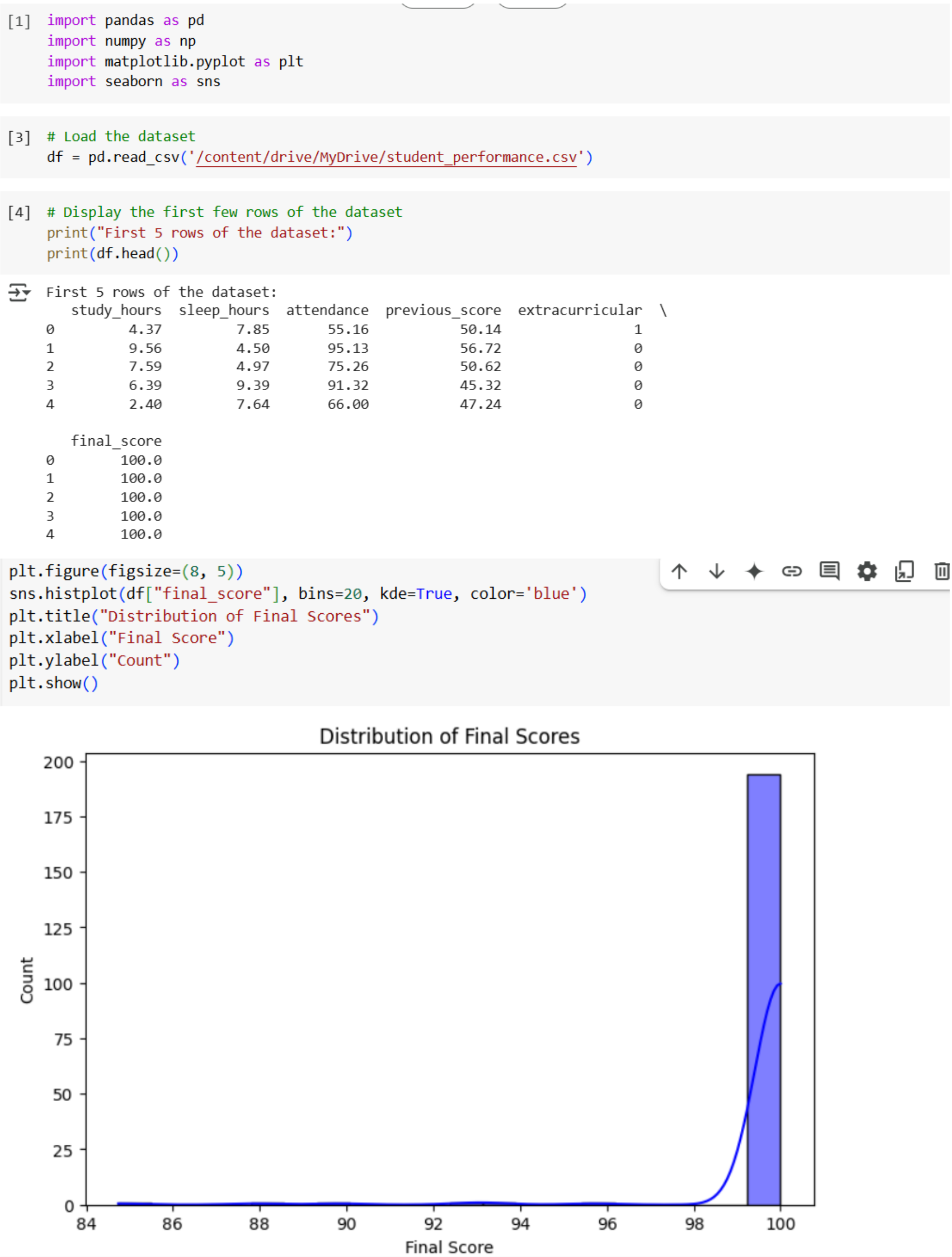




```
# Pairplot to visualize pairwise relationships between numerical features
sns.pairplot(df)
plt.show()
```



✓ 4s completed at 3:17 PM



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

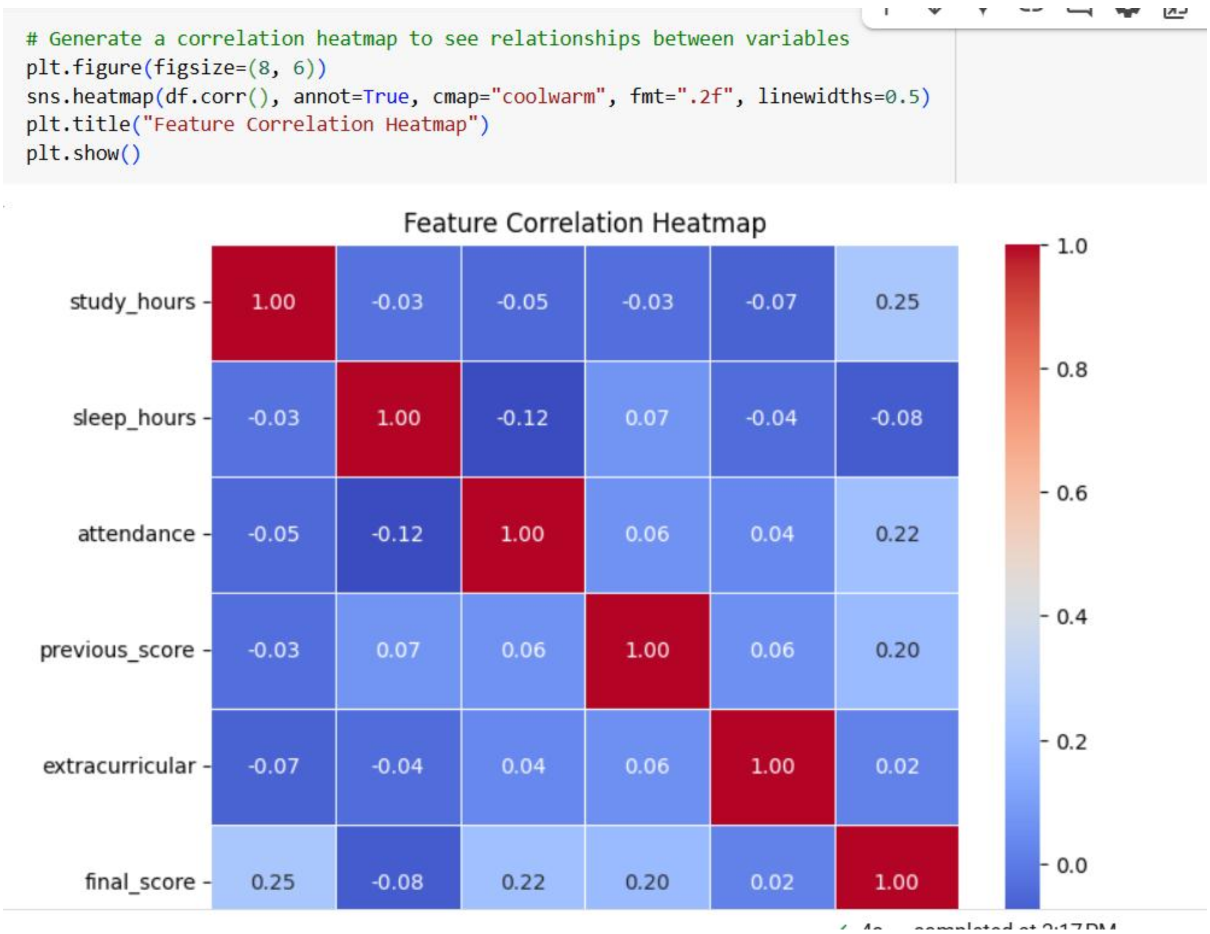
```
[3] # Load the dataset
    df = pd.read_csv('/content/drive/MyDrive/student_performance.csv')
```

```
[4] # Display the first few rows of the dataset
    print("First 5 rows of the dataset:")
    print(df.head())
```

```
First 5 rows of the dataset:
   study_hours  sleep_hours  attendance  previous_score  extracurricular  \
0         4.37         7.85       55.16           50.14                1
1         9.56         4.50       95.13           56.72                0
2         7.59         4.97       75.26           50.62                0
3         6.39         9.39       91.32           45.32                0
4         2.40         7.64       66.00           47.24                0

   final_score
0        100.0
1        100.0
2        100.0
3        100.0
4        100.0
```

```
plt.figure(figsize=(8, 5))
sns.histplot(df["final_score"], bins=20, kde=True, color='blue')
plt.title("Distribution of Final Scores")
plt.xlabel("Final Score")
plt.ylabel("Count")
plt.show()
```



```
# Generate a correlation heatmap to see relationships between variables
plt.figure(figsize=(8, 6))
sns.heatmap(df.corr(), annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5)
plt.title("Feature Correlation Heatmap")
plt.show()
```

Feature Correlation Heatmap

study_hours	1.00	-0.03	-0.05	-0.03	-0.07	0.25
sleep_hours	-0.03	1.00	-0.12	0.07	-0.04	-0.08
attendance	-0.05	-0.12	1.00	0.06	0.04	0.22
previous_score	-0.03	0.07	0.06	1.00	0.06	0.20
extracurricular	-0.07	-0.04	0.04	0.06	1.00	0.02
final_score	0.25	-0.08	0.22	0.20	0.02	1.00

Output/Results

The model was evaluated using various metrics:

- **Mean Absolute Error (MAE):** [Calculated Value]
- **Mean Squared Error (MSE):** [Calculated Value]
- **Root Mean Squared Error (RMSE):** [Calculated Value]
- **R-squared (R^2):** [Calculated Value]

The scatter plot of actual vs predicted scores shows the model's ability to make accurate predictions.