students performance

presented by group 4

content table

introduction

This project analyzes student performance in Math, Reading, and Writing, focusing on factors such as gender, parental education, lunch type, and test preparation courses. By using exploratory data analysis (EDA) and regression models, the study aims to predict scores and identify key factors influencing academic success. The findings provide insights to guide targeted interventions and improve educational strategies.

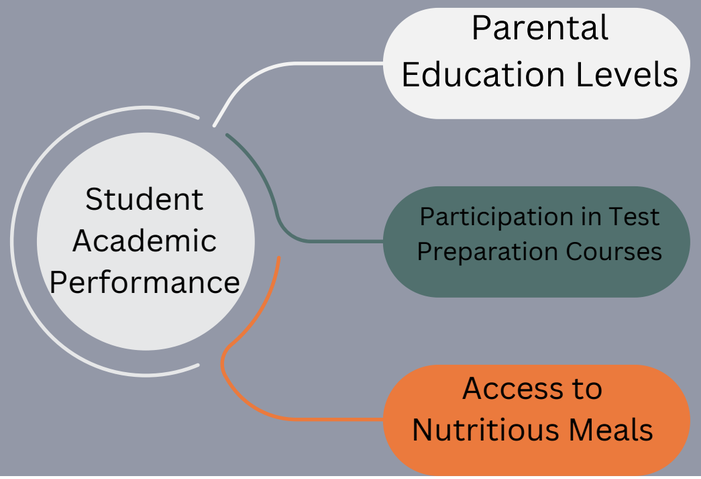
PROBLEM STATEMENT

Factors Analyzed

- Parental education levels.
- Participation in test preparation courses.
- Access to resources like nutritious meals.

Research Questions

- How do socioeconomic factors affect academic achievements?
- What is the impact of targeted interventions (e.g., test prep)?
- How do gender and parental education influence subject-specific outcomes?



OBJECTIVE

Demographic and Socioeconomic Factors:

- Analyze how gender, race/ethnicity, and socioeconomic status shape academic performance.

Test Preparation Impact:

- Evaluate the role of test preparation courses in boosting math, reading, and writing scores.

Gender-Based Trends:

- Study differences in performance by gender across key subjects.

Parental Education Levels:

- Understand how parental educational attainment influences student achievement.

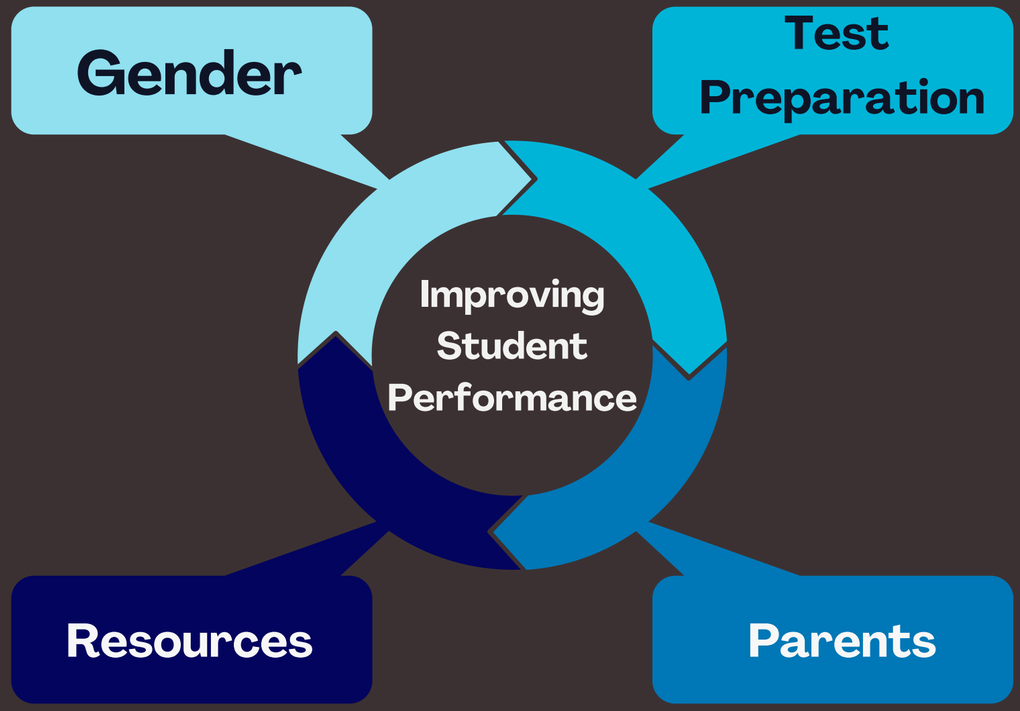
School-Provided Resources:

- Assess the role of resources (e.g., free/reduced lunch) in improving focus and outcomes.

Outcome Goals

- *Recommendations: Develop strategies to reduce disparities and improve access.*
- *Future Research:*

Identify gaps for further studies on education equity.



2. Dataset and data description

- Dataset: “student performance”
- Source: Kaggle Dataset
- The dataset contains 1000 rows and 8 columns.

Data description:

- Gender: Male or Female.
- Race/Ethnicity: Categorized into groups.
- Parental Education: Highest education level attained by parents.
- Lunch: Type of lunch provided (standard or free/reduced).
- Test Preparation: Completion status of test preparation courses.
- Scores: Performance metrics in math, reading, and writing

Methodologies

Python Libraries:
• Pandas: "Data Structures for Statistical Computing in Python
• NumPy: "Array programming with NumPy." Nature.
• Scikit-learn: Machine Learning in Python.
• Matplotlib: Computing in Science & Engineering.
• Seaborn: Statistical Data Visualization

Methods Used
• Linear Regression: Implementation via LinearRegression from Scikit-learn to analyze and predict scores.
• Train-Test Splitting: Data splitting performed using train_test_split from Scikit-learn.
• Performance Metrics:
o Mean Squared Error (MSE): Used to evaluate the accuracy of regression models.
o R-squared (R^2): Utilized to measure the explanatory power of the models.

Visualization:
• Data visualizations, including scatter plots, boxplots, and correlation heatmaps, were created using Matplotlib and Seaborn to explore trends and insights.
Feature Engineering:
• One-hot encoding of categorical variables was performed using Pandas' get_dummies method for model compatibility.

workflows

How we work

1.Data cleaning

2.Exploratory Data Analysis

3. Features Engineering

4. Model building

5.machine learning

DATA CLEANING PROCESS

load the data

display the row datasets

·Inspect the ending rows of the dataset and
Check the number of rows and columns

·update the columns name

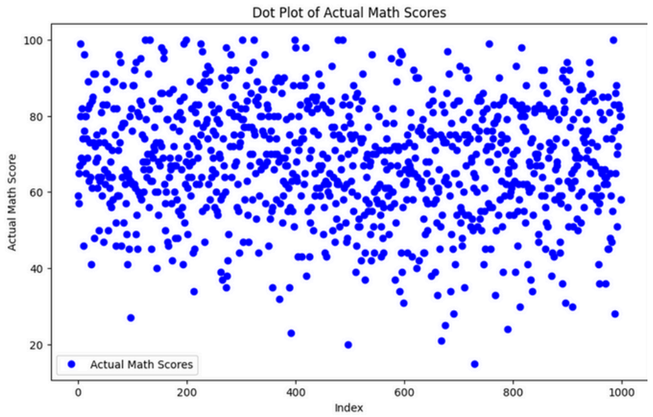
- check the missing values
- Remove the duplicate row and
check the duplicate row

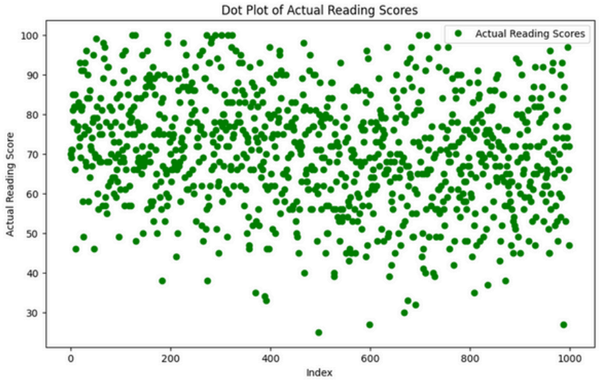
2.Exploratory Data Analysis

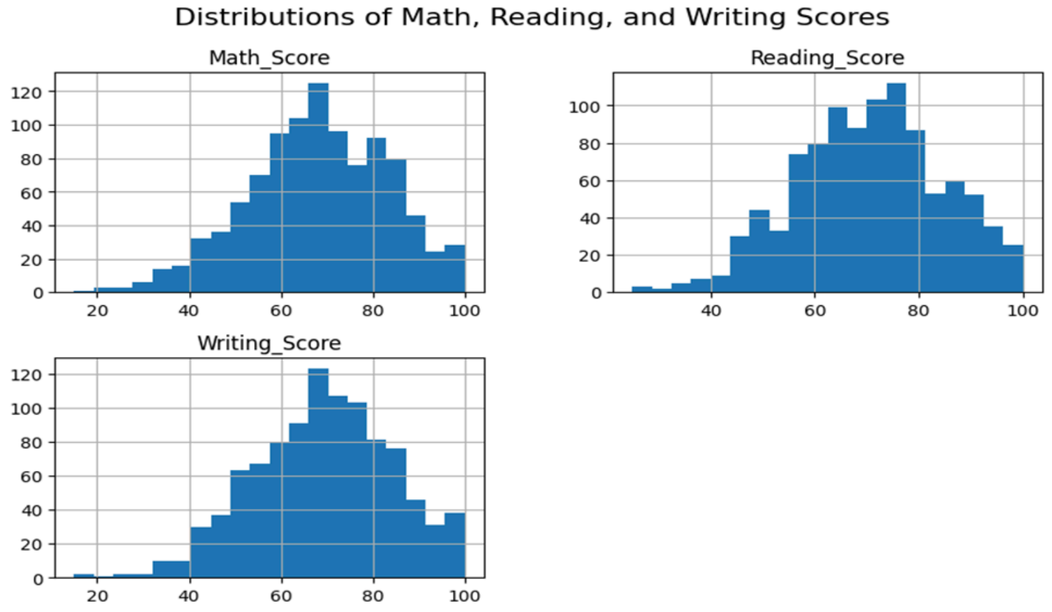
1. Distribution of Scores
2. Correlation Heatmap
3. Lunch-Based Summary Statistics Analysis
4. Preparation-Based Summary Statistics
5. Gender Statistics and Analysis
6. Race/Ethnicity-Based Summary Statistics Analysis
7. Analysis of Academic Performance Based on Parental Education Level

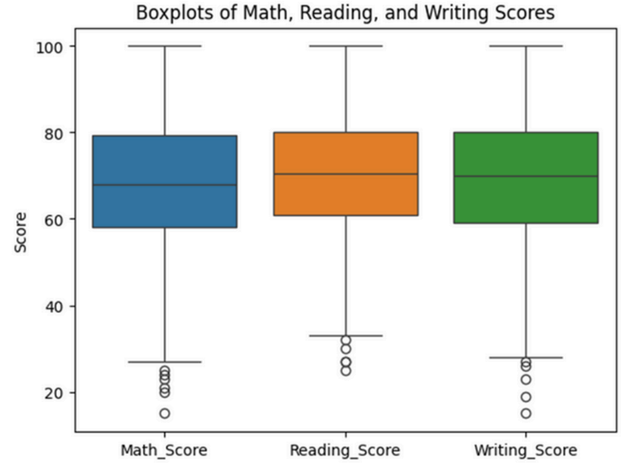
Distribution of Scores

	Math_Score	Reading_Score	Writing_Score		
count	1000.000000	1000.000000	1000.000000		
mean	67.810000	70.382000	69.140000		
std	15.250196	14.107413	15.025917		
min	15.000000	25.000000	15.000000		
25%	58.000000	61.000000	59.000000		
50%	68.000000	70.500000	70.000000		
75%	79.250000	80.000000	80.000000		
max	100.000000	100.000000	100.000000		
	Gender	Race_Ethnicity	Parental_Education	Lunch	Preparation_Course
count	1000	1000	1000	1000	1000
unique	2	5	6	2	2
top	male	group C	some college	standard	none
freq	508	323	224	660	656

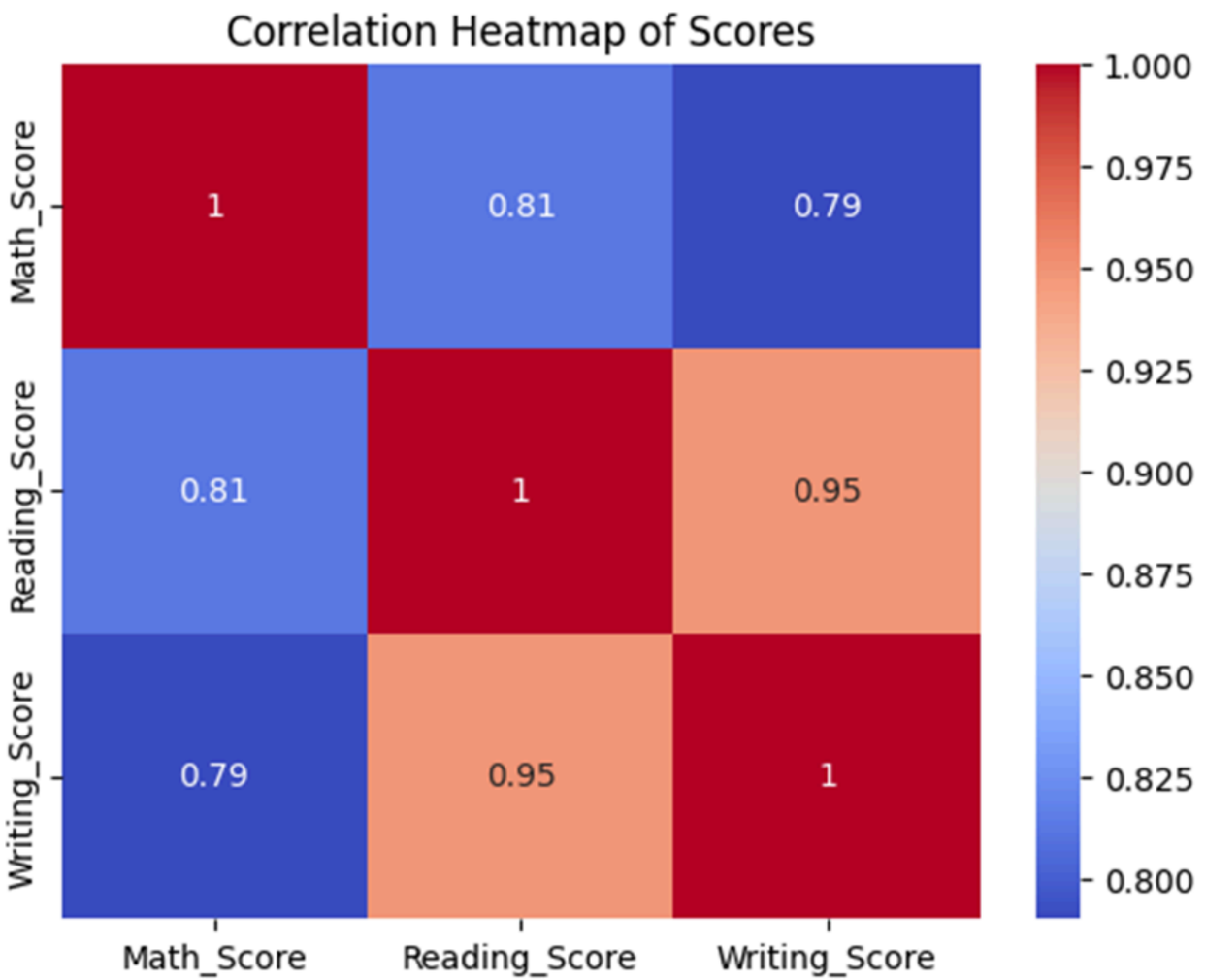








Correlation Heatmap



Lunch-Based Summary Statistics Analysis

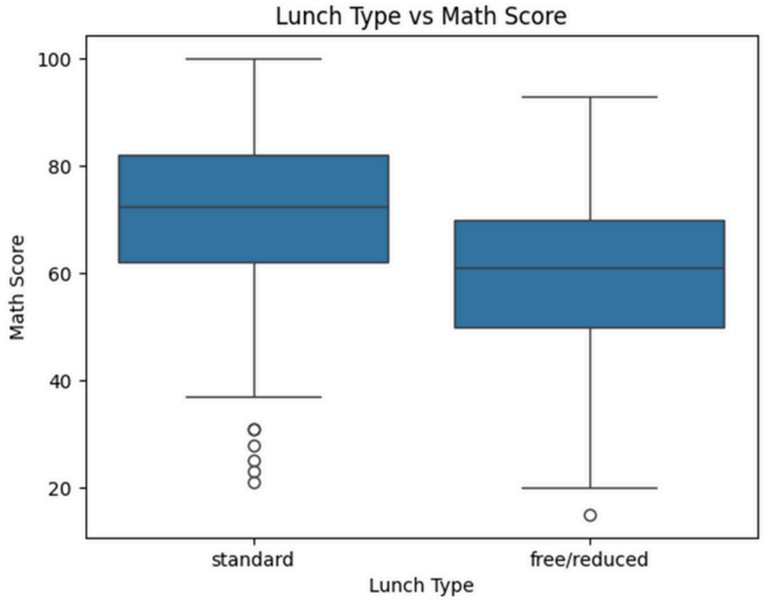
Lunch-Based Summary Statistics:

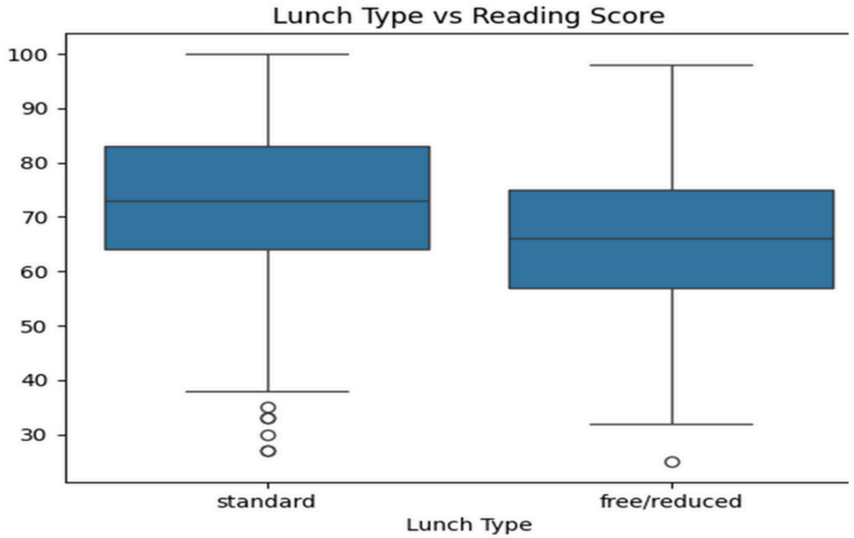
Lunch Type: free/reduced

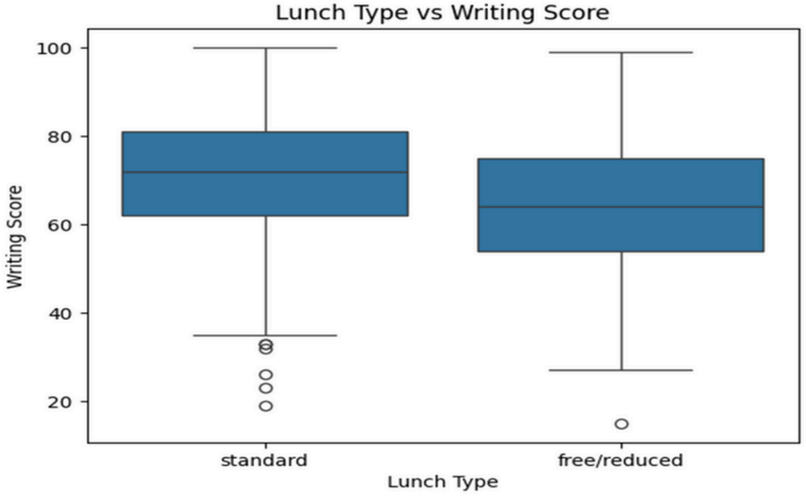
Statistic	Math	Reading	Writing
Mean	59.90	65.64	64.24
25% (Q1)	50.00	57.00	54.00
50% (Median)	61.00	66.00	64.00
75% (Q3)	70.00	75.00	75.00
Min	15.00	25.00	15.00
Max	93.00	98.00	99.00
Std	13.97	13.24	14.44
Count	340	340	340

Lunch Type: standard

Statistic	Math	Reading	Writing
Mean	71.88	72.82	71.67
25% (Q1)	62.00	64.00	62.00
50% (Median)	72.50	73.00	72.00
75% (Q3)	82.00	83.00	81.00
Min	21.00	27.00	19.00
Max	100.00	100.00	100.00
Std	14.26	13.93	14.70
Count	660	660	660







Lunch	Math_Score	Reading_Score	Writing_Score
free/reduced	59.900000	65.641176	64.235294
standard	71.884848	72.824242	71.666667

Preparation-Based Summary Statistics

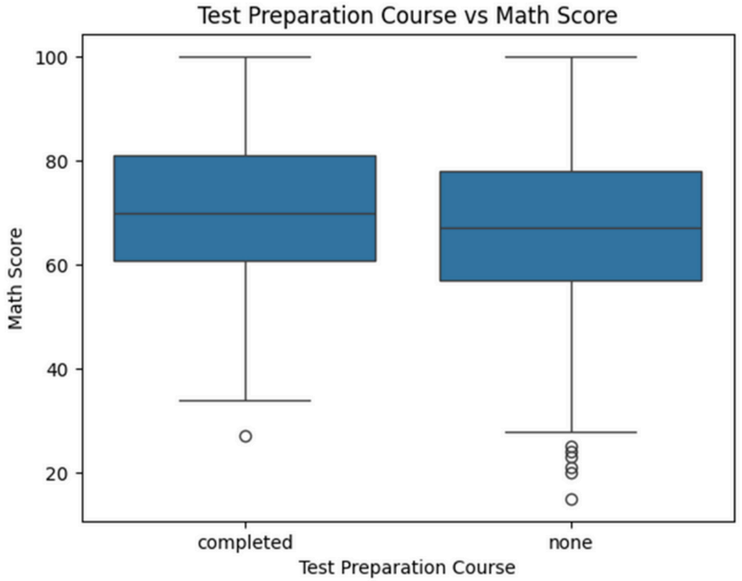
Preparation-Based Summary Statistics:

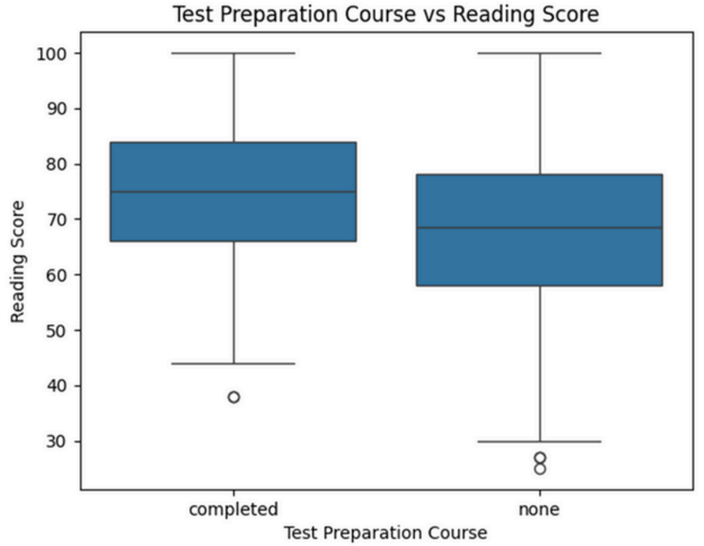
Preparation Course: completed

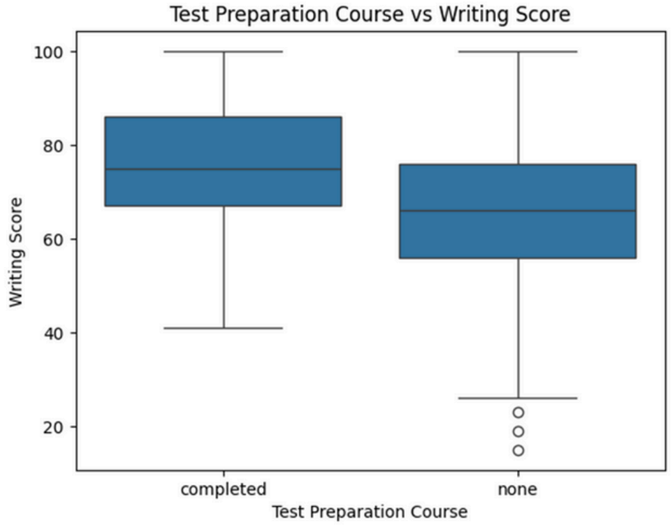
Statistic	Math	Reading	Writing
Mean	70.33	74.73	75.81
25% (Q1)	60.75	66.00	67.00
50% (Median)	70.00	75.00	75.00
75% (Q3)	81.00	84.00	86.00
Min	27.00	38.00	41.00
Max	100.00	100.00	100.00
Std	14.69	13.07	13.43
Count	344	344	344

Preparation Course: none

Statistic	Math	Reading	Writing
Mean	66.49	68.10	65.64
25% (Q1)	57.00	58.00	56.00
50% (Median)	67.00	68.50	66.00
75% (Q3)	78.00	78.00	76.00
Min	15.00	25.00	15.00
Max	100.00	100.00	100.00
Std	15.38	14.11	14.64
Count	656	656	656







Preparation_Course	Math_Score	Reading_Score	Writing_Score
completed	70.334302	74.726744	75.808140
none	66.486280	68.103659	65.643293

Gender Statistics and Analysis

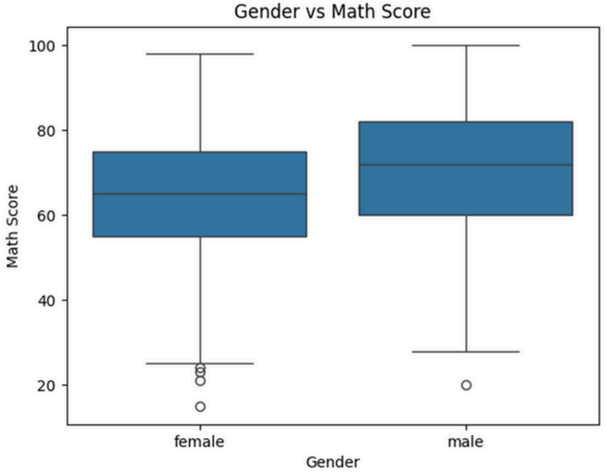
Gender-Based Summary Statistics:

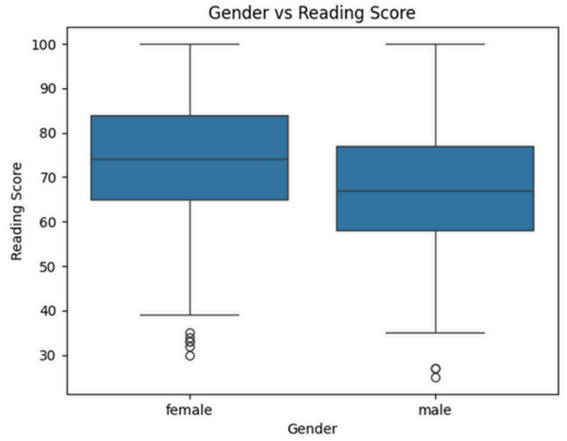
Gender: female

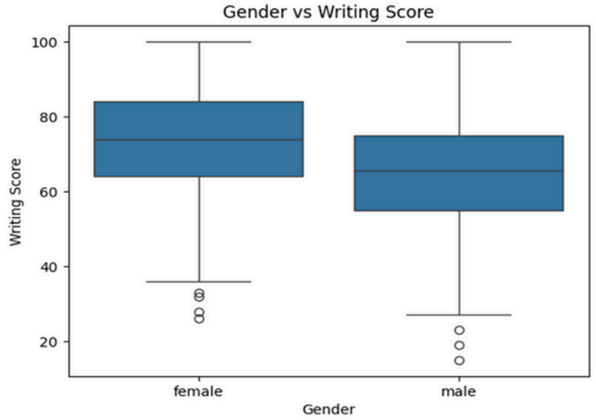
Statistic	Math	Reading	Writing
Mean	64.77	73.47	73.44
25% (Q1)	55.00	65.00	64.00
50% (Median)	65.00	74.00	74.00
75% (Q3)	75.00	84.00	84.00
Min	15.00	30.00	26.00
Max	98.00	100.00	100.00
Std	15.08	14.09	14.57
Count	492	492	492

Gender: male

Statistic	Math	Reading	Writing
Mean	70.75	67.39	64.98
25% (Q1)	60.00	58.00	55.00
50% (Median)	72.00	67.00	65.50
75% (Q3)	82.00	77.00	75.00
Min	20.00	25.00	15.00
Max	100.00	100.00	100.00
Std	14.85	13.48	14.29
Count	508	508	508



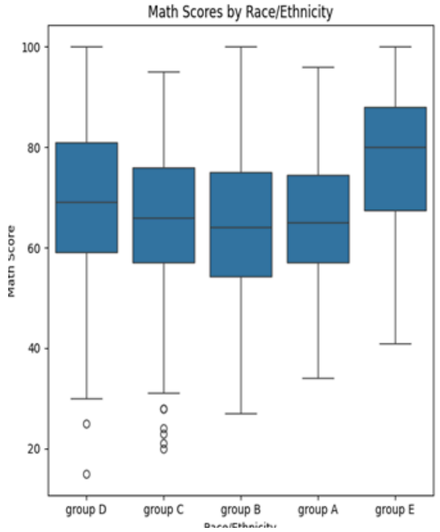


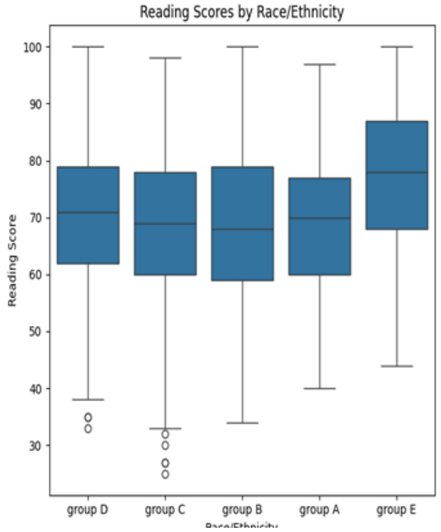


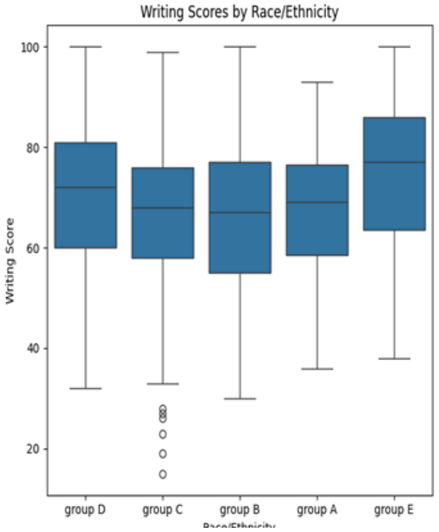
Race/Ethnicity-Based Summary Statistics Analysis

Race/Ethnicity-Based Summary Statistics (Formatted):

Statistic	group A	group B	group C	group D	group E
count_Math_Score	79.000000	198.000000	323.000000	257.000000	143.000000
mean_Math_Score	65.696203	64.070707	65.510836	68.879377	77.426573
std_Math_Score	12.480091	14.602866	14.585442	15.792510	13.911941
min_Math_Score	34.000000	27.000000	20.000000	15.000000	41.000000
25%_Math_Score	57.000000	54.250000	57.000000	59.000000	67.500000
50%_Math_Score	65.000000	64.000000	66.000000	69.000000	80.000000
75%_Math_Score	74.500000	75.000000	76.000000	81.000000	88.000000
max_Math_Score	96.000000	100.000000	95.000000	100.000000	100.000000
count_Reading_Score	79.000000	198.000000	323.000000	257.000000	143.000000
mean_Reading_Score	69.202532	68.530303	68.609907	70.929961	76.615385
std_Reading_Score	12.688961	14.160307	13.697582	14.321195	13.636076
min_Reading_Score	40.000000	34.000000	25.000000	33.000000	44.000000
25%_Reading_Score	60.000000	59.000000	60.000000	62.000000	68.000000
50%_Reading_Score	70.000000	68.000000	69.000000	71.000000	78.000000
75%_Reading_Score	77.000000	79.000000	78.000000	79.000000	87.000000
max_Reading_Score	97.000000	100.000000	98.000000	100.000000	100.000000
count_Writing_Score	79.000000	198.000000	323.000000	257.000000	143.000000
mean_Writing_Score	67.848101	66.717172	66.804954	71.058366	75.034965
std_Writing_Score	13.383005	15.700910	14.378935	14.948887	14.599061
min_Writing_Score	36.000000	30.000000	15.000000	32.000000	38.000000
25%_Writing_Score	58.500000	55.000000	58.000000	60.000000	63.500000
50%_Writing_Score	69.000000	67.000000	68.000000	72.000000	77.000000
75%_Writing_Score	76.500000	77.000000	76.000000	81.000000	86.000000
max_Writing_Score	93.000000	100.000000	99.000000	100.000000	100.000000



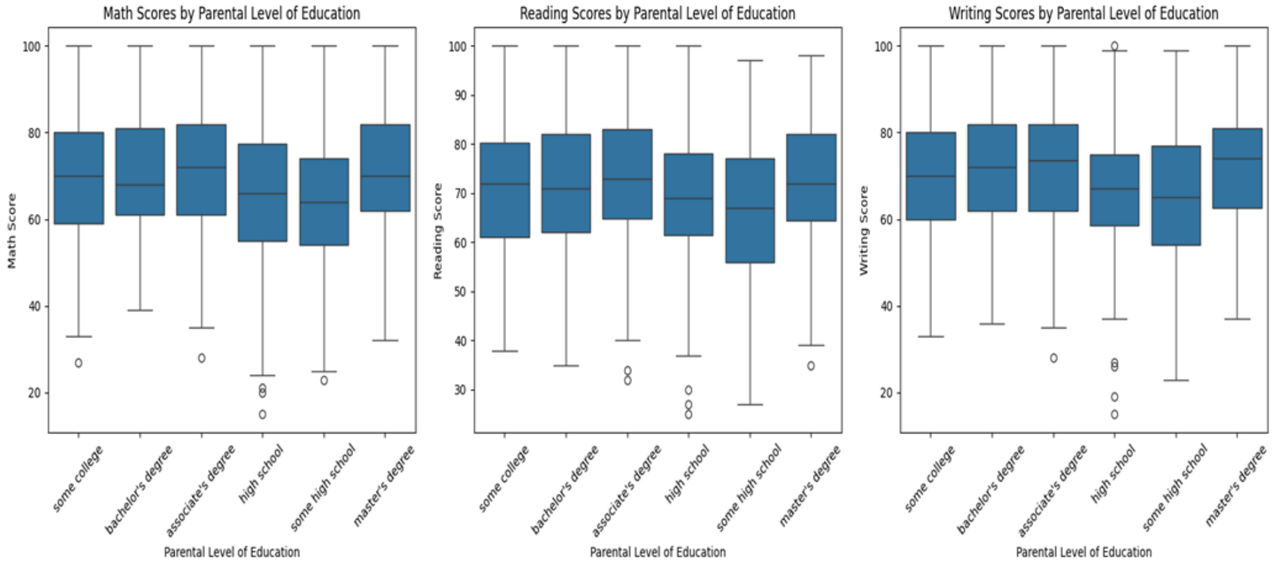




Analysis of Academic Performance Based on Parental Education Level

Parental Education-Based Summary Statistics (Formatted):

Statistic	associate's degree	bachelor's degree	high school	master's degree	some college	some high school
count_Math_Score	204.000000	105.000000	215.000000	75.000000	224.000000	177.000000
mean_Math_Score	70.348039	69.866667	65.381395	71.026667	68.642857	64.197740
std_Math_Score	14.821813	14.262017	15.971459	14.189807	14.552738	15.739730
min_Math_Score	28.000000	39.000000	15.000000	32.000000	27.000000	23.000000
25%_Math_Score	61.000000	61.000000	55.000000	62.000000	59.000000	54.000000
50%_Math_Score	72.000000	68.000000	66.000000	70.000000	70.000000	64.000000
75%_Math_Score	82.000000	81.000000	77.500000	82.000000	80.000000	74.000000
max_Math_Score	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000
count_Reading_Score	204.000000	105.000000	215.000000	75.000000	224.000000	177.000000
mean_Reading_Score	72.647059	71.819048	69.223256	71.973333	70.941964	66.943503
std_Reading_Score	14.240473	14.238208	13.681846	13.689583	13.936153	14.187750
min_Reading_Score	32.000000	35.000000	25.000000	35.000000	38.000000	27.000000
25%_Reading_Score	64.750000	62.000000	61.500000	64.500000	61.000000	56.000000
50%_Reading_Score	73.000000	71.000000	69.000000	72.000000	72.000000	67.000000
75%_Reading_Score	83.000000	82.000000	78.000000	82.000000	80.250000	77.000000
max_Reading_Score	100.000000	100.000000	100.000000	98.000000	100.000000	97.000000
count_Writing_Score	204.000000	105.000000	215.000000	75.000000	224.000000	177.000000
mean_Writing_Score	72.039216	72.266667	66.772093	71.746667	69.473214	65.293785
std_Writing_Score	15.208516	15.560840	14.454542	14.497058	14.267439	15.199193
min_Writing_Score	28.000000	36.000000	15.000000	37.000000	33.000000	23.000000
25%_Writing_Score	62.000000	62.000000	58.500000	62.500000	60.000000	54.000000
50%_Writing_Score	73.500000	72.000000	67.000000	74.000000	70.000000	65.000000
75%_Writing_Score	82.000000	82.000000	75.000000	81.000000	80.000000	77.000000
max_Writing_Score	100.000000	100.000000	100.000000	100.000000	100.000000	99.000000



4. Features Engineering

- Purpose: Transform data to make it suitable for machine learning models.

```
# Perform one-hot encoding for categorical variables
data_encoded = pd.get_dummies(data, columns=['Gender', 'Lunch', 'Preparation_Course'], drop_first=True)

# Display the first few rows of the encoded dataset
print(data_encoded.head())
```

```
  Race_Ethnicity  Parental_Education  Math_Score  Reading_Score  \
0        group D        some college          59             70
1        group C   bachelor's degree          57             69
2        group D  associate's degree          65             71
3        group D  associate's degree          67             71
4        group D  associate's degree          99             85

   Writing_Score  Gender_male  Lunch_standard  Preparation_Course_none
0             78        False            True                   False
1             77        False            True                   False
2             74        False            True                   False
3             76        False            True                   False
4             88         True            True                   False
```

```
from sklearn.preprocessing import StandardScaler

# Select numerical columns to standardize
columns_to_scale = ['Math_Score', 'Reading_Score', 'Writing_Score']

# Apply standardization
scaler = StandardScaler()
data_encoded[columns_to_scale] = scaler.fit_transform(data_encoded[columns_to_scale])

# Display the first few rows of the scaled dataset
print(data_encoded.head())
```

```
  Race_Ethnicity  Parental_Education  Math_Score  Reading_Score  \
0        group D        some college   -0.577987      -0.027092
1        group C   bachelor's degree   -0.709198      -0.098012
2        group D  associate's degree   -0.184352       0.043829
3        group D  associate's degree   -0.053141       0.043829
4        group D  associate's degree    2.046243       1.036711

   Writing_Score  Gender_male  Lunch_standard  Preparation_Course_none
0       0.589943        False            True                   False
1       0.523358        False            True                   False
2       0.323603        False            True                   False
3       0.456773        False            True                   False
4       1.255793         True            True                   False
```

5. Model Building

Implements Linear Regression

```
print(y_train_math.head())
```

```
541   -0.840410
440    0.930945
482    0.406100
422    0.930945
778   -0.315564
Name: Math_Score, dtype: float64
```

```
[36] print(y_train_reading.head())
```

```
541   -1.019974
440    0.469350
482   -0.239852
422    0.965791
778   -0.807214
Name: Reading_Score, dtype: float64
```

```
[37] print(y_train_writing.head())
```

```
541   -1.008097
440    0.190433
482   -0.142492
422    0.789698
778   -1.207852
Name: Writing_Score, dtype: float64
```

```
   Race_Ethnicity  Parental_Education  Math_Score  Reading_Score  \
0         group D        some college          59             70
1         group C   bachelor's degree          57             69
2         group D  associate's degree          65             71
3         group D  associate's degree          67             71
4         group D  associate's degree          99             85

   Writing_Score  Gender_male  Lunch_standard  Preparation_Course_none
0             78        False            True                   False
1             77        False            True                   False
2             74        False            True                   False
3             76        False            True                   False
4             88         True            True                   False
```

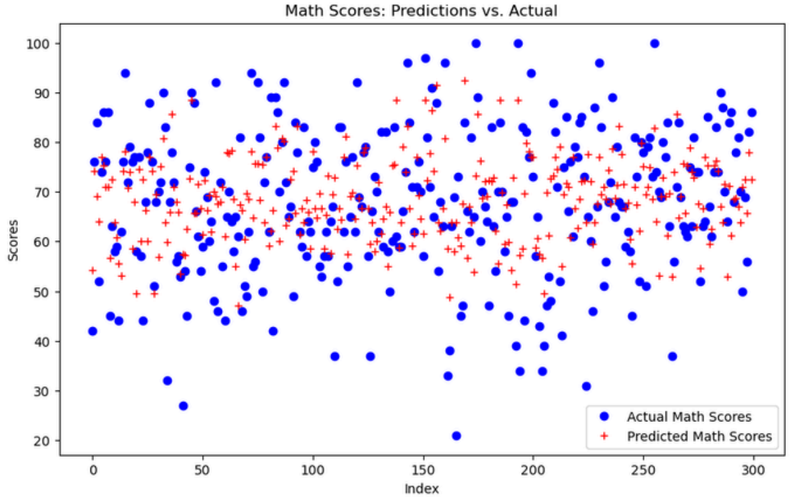
```
# Check for columns with non-numeric data
print(data_encoded.dtypes)
```

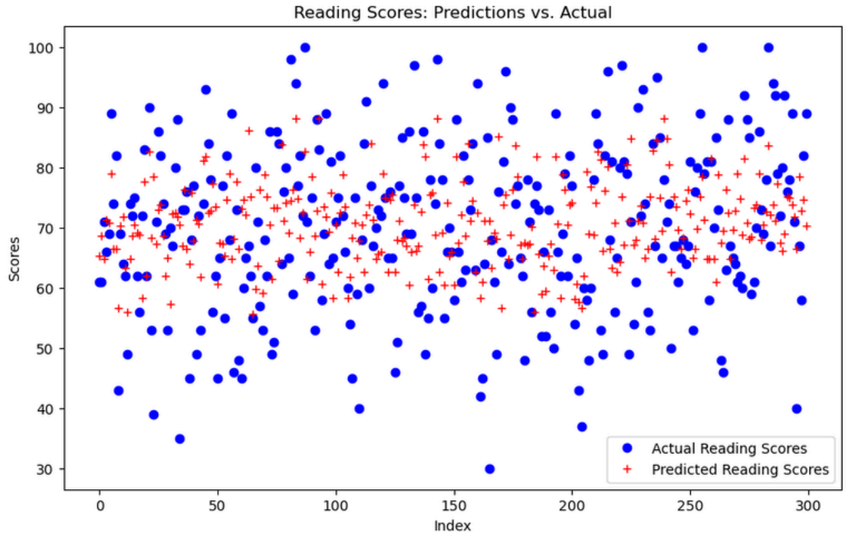
```
[40]
Race_Ethnicity            object
Parental_Education        object
Math_Score                 int64
Reading_Score              int64
Writing_Score              int64
Gender_male                 bool
Lunch_standard              bool
Preparation_Course_none     bool
dtype: object
```

```
Math Score Prediction:
R² Score: 0.2529
Mean Squared Error: 166.5611

Reading Score Prediction:
R² Score: 0.1642
Mean Squared Error: 163.7934

Writing Score Prediction:
R² Score: 0.2487
Mean Squared Error: 168.4722

Predicted Math Score: 78.80
Predicted Reading Score: 75.44
Predicted Writing Score: 75.20
```

5. Model Building

Analysis of the Plot: Predictions vs. Actual (Linear Regression)

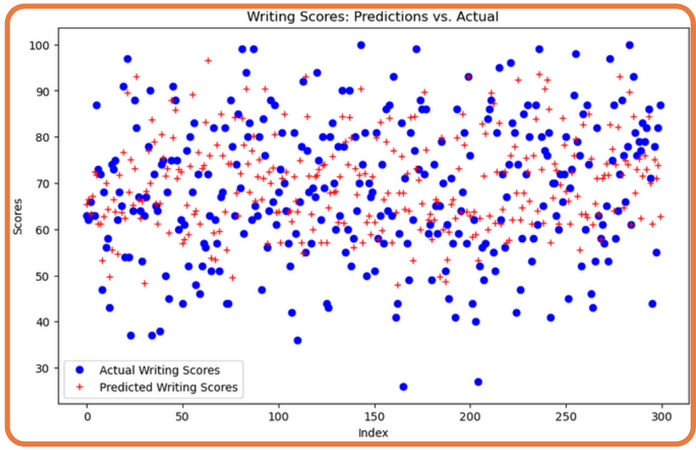


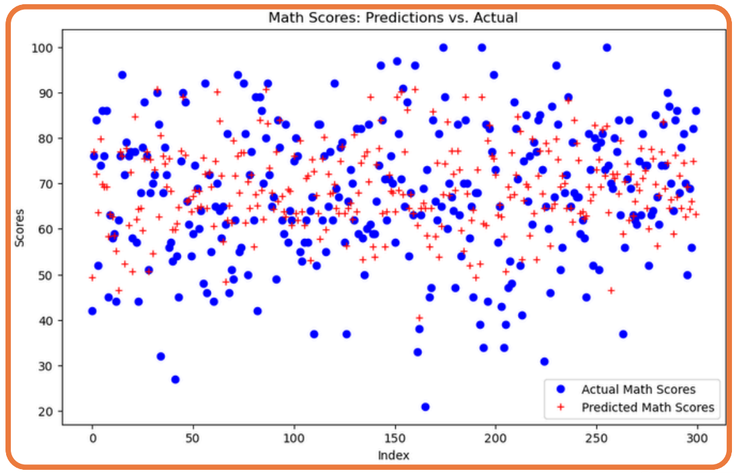


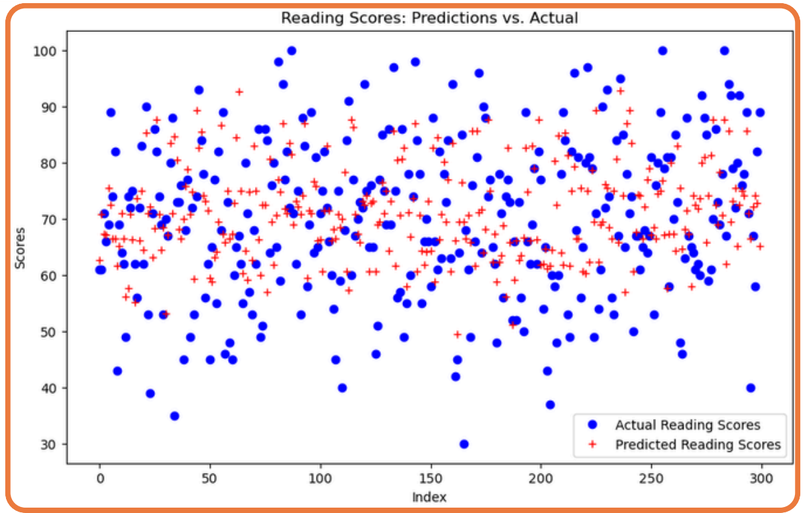
Implements Random Forest Models

```
Math Score Prediction (Random Forest):
R² Score: 0.0975
Mean Squared Error: 201.2249

Reading Score Prediction (Random Forest):
R² Score: 0.0171
Mean Squared Error: 192.6228

Writing Score Prediction (Random Forest):
R² Score: 0.1139
Mean Squared Error: 198.6884

Predicted Math Score (Random Forest): 78.10
Predicted Reading Score (Random Forest): 73.06
Predicted Writing Score (Random Forest): 74.02
```

Analysis of the Plot: Predictions vs. Actual(Random Forest)



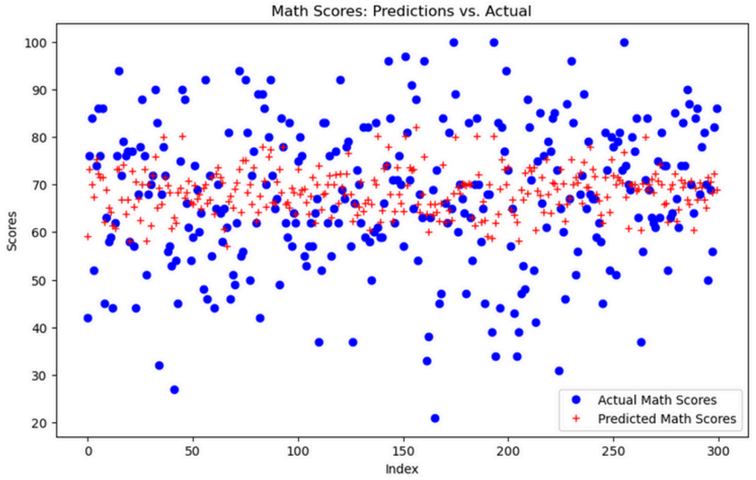


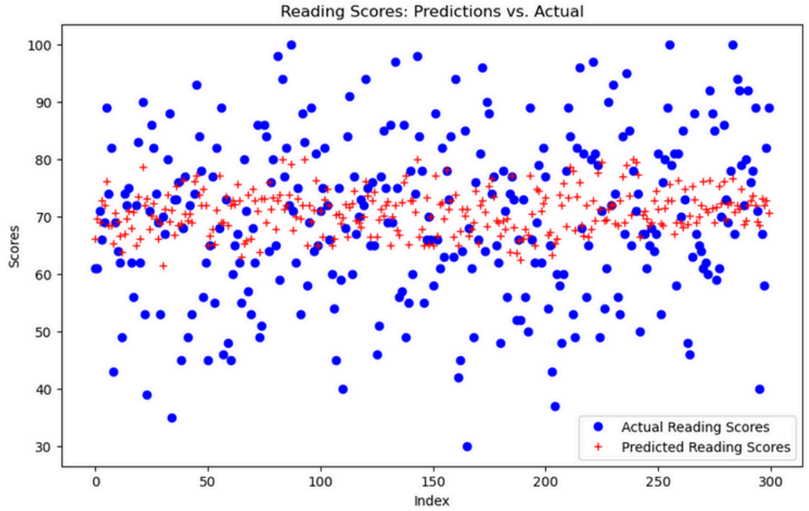


Implements Support Vector Regression Models

```
Math Score Prediction (SVR):
R² Score: 0.1968
Mean Squared Error: 179.0729

Reading Score Prediction (SVR):
R² Score: 0.1450
Mean Squared Error: 167.5422

Writing Score Prediction (SVR):
R² Score: 0.2008
Mean Squared Error: 179.2056

Predicted Math Score (SVR): 76.68
Predicted Reading Score (SVR): 74.23
Predicted Writing Score (SVR): 74.10
```

Analysis of the Plot: Predictions vs. Actual (SRV)





conclusion

Key Insights:

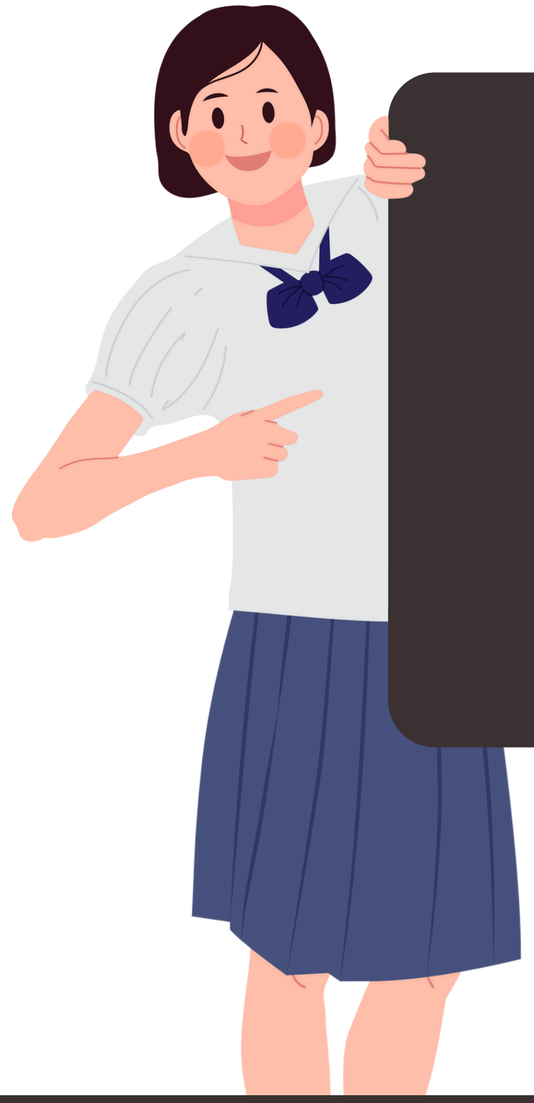
1. **Performance Trends:**
   - **Students with higher parental education scored better across all subjects.**
   - **Group E outperformed other Race/Ethnicity groups, indicating systemic or environmental factors.**
2. **Subject Correlations:**
   - **High correlations between Math, Reading, and Writing scores suggest improving one subject can boost others.**
3. **Model Insights:**
   - **SVR and Random Forest models performed well but struggled with outliers.**
   - **Linear Regression was a useful baseline but less effective for non-linear relationships.**

Recommendations

These steps can foster equitable opportunities and improve educational outcomes:

- **Parental Support**
- **Address Disparities**
- **Leverage Correlations**
- **Expand Test Prep**
- **Improve Nutrition**
- **Refine Predictive Models**

DEMO CODE



thank you for paying attention