A logo on a black background

Description automatically generatedព្រះរាជាណាចក្រកម្ពុជា

ជាតិ សាសនា ព្រះមហាក្សត្រ

**Introduction to Data Science**

**Assignment**

**Topic: Students performance**

**Department: I3AMS 1-A**

**Name of Students** **ID of Students** **Score**

1. HOEURN SREYKA e20220699 ……….

2. HEM BELLYDAY e20220800 ……….

2. HENG MENGHONG e20220258 ……….

3. HOK LYHOUR e20220759 ……….

4. HEANG SOTHEARA e20221213 ……….

Lecturer: **Dr. PHAUK Sokkhey (Course)**

**Academic year 2024-2025**

**Table of Contents**

# I. Introduction……………………………………………………………

**1. Problem statement…………………………………………………..**

# 2. Objectives……………………………………………………………

1. **Dataset and Data Description**

# Methodologies and workflow

* + **Methodology…………………………………………………..**

# Workflows

**1. Data cleaning………………………………………………………..**

# Exploratory Data Analysis…………………………………………

1. **Correlations Heatmap……………………………………………..**

# 4. Feature engineering………………………………………………..

**5. Model Building……………………………………………………..**

# Conclusions

* + **Recommendation…………………………………………………**

# Introduction

Education is a fundamental pillar of societal development, serving as the foundation for individual growth and collective progress. However, disparities in educational outcomes remain a pressing challenge, with factors such as socioeconomic background, family support, and access to resources significantly shaping student performance. Understanding these factors is crucial to addressing performance gaps and fostering equitable learning environments.

This report explores a dataset capturing demographic, socioeconomic, and academic information about students to uncover the underlying factors influencing their academic outcomes. The analysis delves into key aspects such as the impact of parental education levels, the effectiveness of test preparation courses, and the role of socioeconomic conditions in shaping performance. It also examines gender-based differences and correlations between subject-specific scores to provide a holistic view of student achievement.

By connecting data-driven insights to real-world challenges, this report aims to equip educators, policymakers, and institutions with actionable strategies to enhance learning outcomes and reduce disparities. The findings will highlight areas for targeted interventions, helping to create a more supportive and inclusive educational landscape for all students.

### Problem statement

Understanding and addressing the factors that influence student performance is critical for fostering equitable and effective educational outcomes. Students' academic success is shaped by a complex interplay of individual attributes, family background, and external influences such as socioeconomic conditions and institutional support.

Despite widespread recognition of these influences, disparities in educational performance persist, particularly among students from varying socioeconomic and demographic backgrounds.

This report aims to identify and analyze key factors contributing to student performance, such as parental education levels, participation in test preparation courses, and access to resources like nutritious meals. By uncovering these relationships, the study seeks to address pressing questions:

* How do socioeconomic factors affect students' academic achievements?
* What is the impact of targeted interventions, such as test preparation courses, on performance outcomes?
* How do gender and parental education levels influence performance across different subjects?

The ultimate goal is to provide actionable insights that enable educators, policymakers, and institutions to design targeted strategies for improving academic outcomes, reducing disparities, and fostering a supportive learning environment for all students.

# Objectives

* **Demographic and Socioeconomic Factors**: Explore how attributes such as gender, race/ethnicity, and socioeconomic status influence student performance. This involves identifying disparities among demographic groups and understanding how these variables intersect to affect academic outcomes. For instance, students from disadvantaged

socioeconomic backgrounds may face additional challenges, such as limited access to resources, which can directly impact their academic success.

* **Impact of Test Preparation**: Assess the role of test preparation courses in improving student performance. This includes evaluating whether students who complete these courses achieve higher scores in math, reading, and writing compared to those who do not. Understanding the effectiveness of such interventions helps in designing programs that bridge the performance gap for underperforming students.
* **Gender-Based Performance**: Analyze performance differences across subjects like math, reading, and writing based on gender. For example, identifying trends such as male students excelling in math and female students performing better in reading and writing can inform the development of tailored teaching methods and support systems.
* **Parental Education Levels**: Examine how the educational attainment of parents correlates with student achievement. This includes understanding the extent to which parental education impacts the availability of academic support, resources, and encouragement at home, thereby influencing students' motivation and outcomes.
* **Role of School-Provided Resources**: Evaluate the influence of essential resources, such as free or reduced lunch programs, on student performance. This analysis seeks to uncover how meeting basic needs enables students to focus better on academics and improve their overall performance.
* **Actionable Recommendations**: Develop specific, evidence-based strategies that educators and policymakers can implement to reduce academic disparities. These recommendations will focus on improving access to test preparation, addressing resource gaps, and fostering inclusive practices to ensure all students have equal opportunities to succeed.
* **Future Research Insights**: Identify patterns and gaps in the current data to guide further research. This involves exploring additional factors, such as extracurricular activities, teaching quality, and learning environments, to enhance educational equity and academic outcomes on a broader scale.

1. **Dataset and Data Description**

* Dataset: “student performance”
* Source: Kaggle <https://www.kaggle.com/datasets/rkiattisak/student-performance-in-mathematics>
* Purpose: To analyze factors influencing student performance, including demographics, socioeconomic conditions, and test preparation, while offering strategies to promote equitable education.
* The dataset contains 1000 rows and 8 columns.

The dataset used in this analysis includes information on 1,000 students, with attributes such as gender, race/ethnicity, parental education, lunch type, test preparation completion, and scores in math, reading, and writing. Key features include:

* **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.

### Methodology and Workflows

**Methodology 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:

* + Mean Squared Error (MSE): Used to evaluate the accuracy of regression models.
  + R-squared (R²): 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:

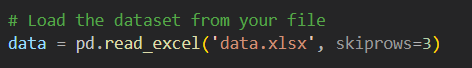
* One-hot encoding of categorical variables was performed using Pandas' get\_dummies method for model compatibility.

### Workflows

* 1. **Data cleaning process**

Before building the machine learning model, it is essential to ensure the dataset is clean, consist, and ready for analysis. Therefore, data cleaning is the foundational step in the data science process because the model’s dependability and performance are directly impacted by the quality of the data. The following are the process we do to clean the data:

* + - importing libraries
    - load the data



* + - display the row datasets



out put:

### Gender Race/Ethnicity Parental Education Lunch TestPreparation MathScore ReadingScore WritingScore

**0 Female Group A Bachelor's Degree Standard Completed 78 85 88**

**1 Male Group B High School Reduced None 65 72 70**

### 2 Female Group C Some College Standard Completed 80 88 92

**3 Female Group B Associate's Degree Reduced Completed 74 79 85**

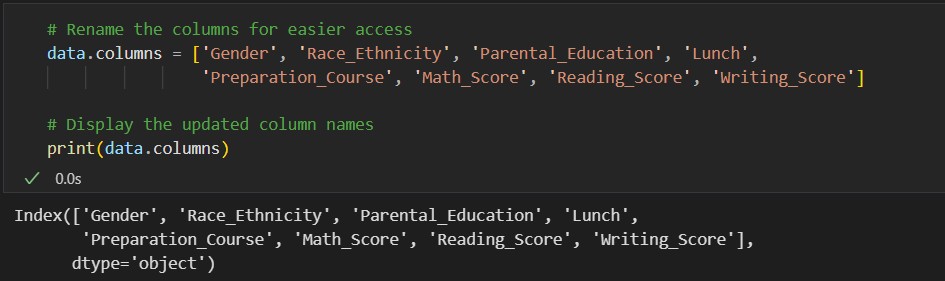
**4 Male Group A Master's Degree Standard None 85 89 90**

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

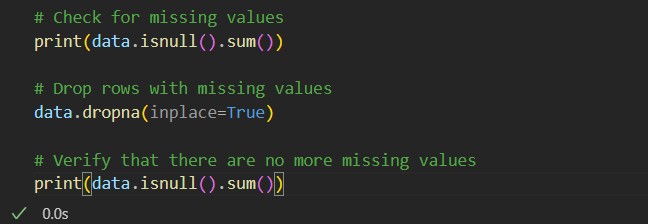


* **update the columns name**

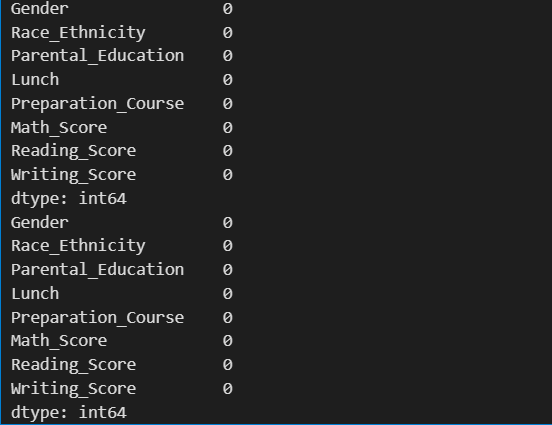
just write the new name for index for easy looking.



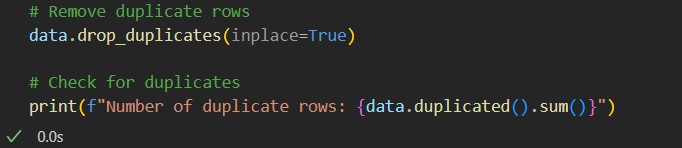
### check the missing values



**Out put**



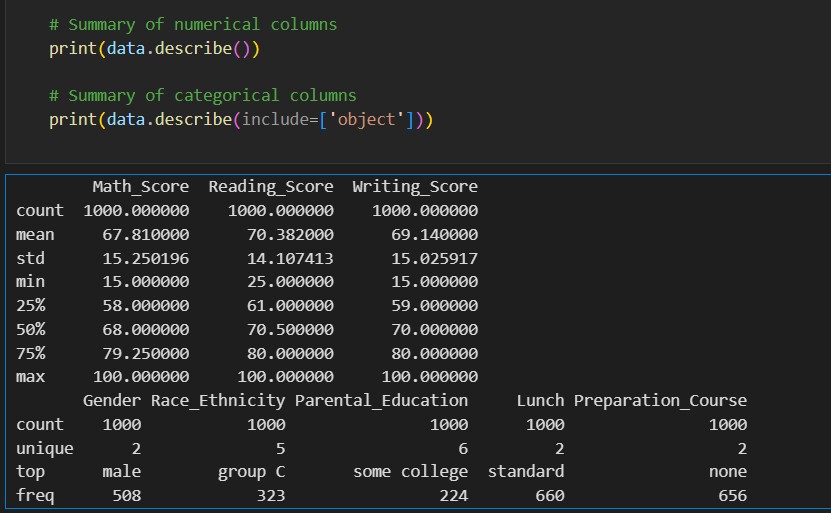
* **Remove the duplicate row and check the duplicate row**



# Exploratory Data Analysis

## Summary Statistics

The dataset provides valuable insights into both numerical and categorical attributes, enabling a comprehensive understanding of student performance trends:



## Math Scores:

* + The average math score is 67.81, with a standard deviation of 15.25, indicating moderate variability in performance.
  + The minimum score is 15, while the maximum is 100, showcasing a wide range of capabilities among students.
  + The 25th percentile is 58, the median (50th percentile) is 68, and the 75th percentile is 79.25, suggesting that most students score within a mid-to-high range.

## Reading Scores:

* + The mean reading score is 70.38, with a standard deviation of 14.11, slightly less variable than math scores.
  + Scores range from 25 to 100, with the middle 50% of students scoring between 61 and 80.

## Writing Scores:

* + The mean writing score is 69.14, with a standard deviation of 15.03, showing similar variability to math scores.
  + Scores range from 15 to 100, and the interquartile range (59 to 80) highlights consistent performance patterns among most students.

*Categorical Columns*

## Gender:

* + The dataset is nearly balanced with 508 male students and 492 female students, providing equitable representation for gender-based analysis.

## Race/Ethnicity:

* + Students are divided into six groups, with Group C being the largest (323 students). This provides an opportunity to analyze group-specific trends in performance.

## Parental Level of Education:

* + Parental education levels range from "high school" to "master's degree," with "some college" being the most common category (224 students). This variation allows for an analysis of the relationship between parental education and student outcomes.

## Lunch:

* + A majority of students (660) receive standard lunch, while 340 students are on free/reduced lunch. This distinction serves as a proxy for socioeconomic status in the analysis.

## Test Preparation Course:

* + 344 students completed the test preparation course, while 656 did not. This disparity provides a basis for evaluating the impact of test preparation on academic performance.

Exploratory Data Analysis (EDA) revealed key insights into student performance patterns and relationships among variables. Below are the major findings:

### Distribution of Scores

* **Lunch type**

### Test Preparation Course

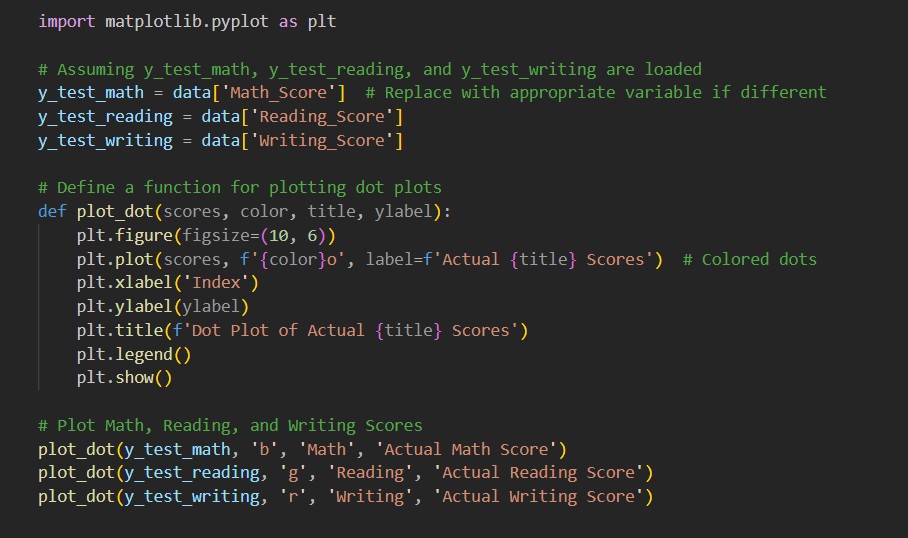
* **Gender**

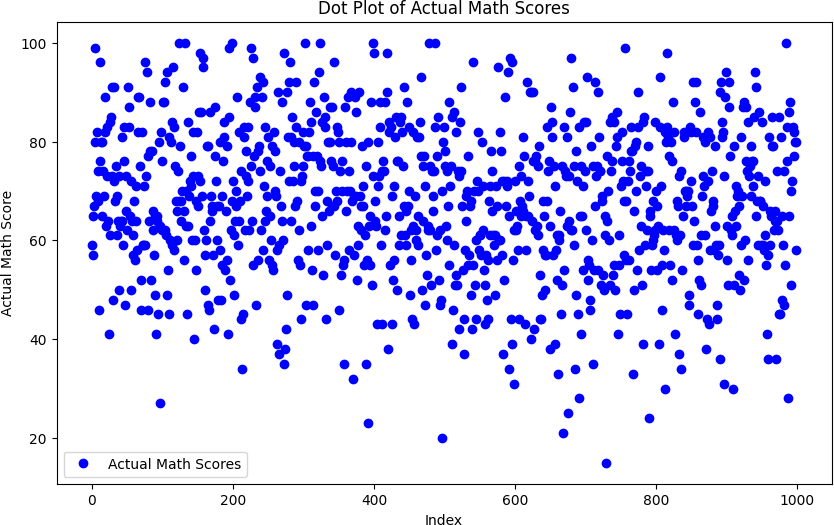
### Race/Ethnicity

* **Parental Level of Education**

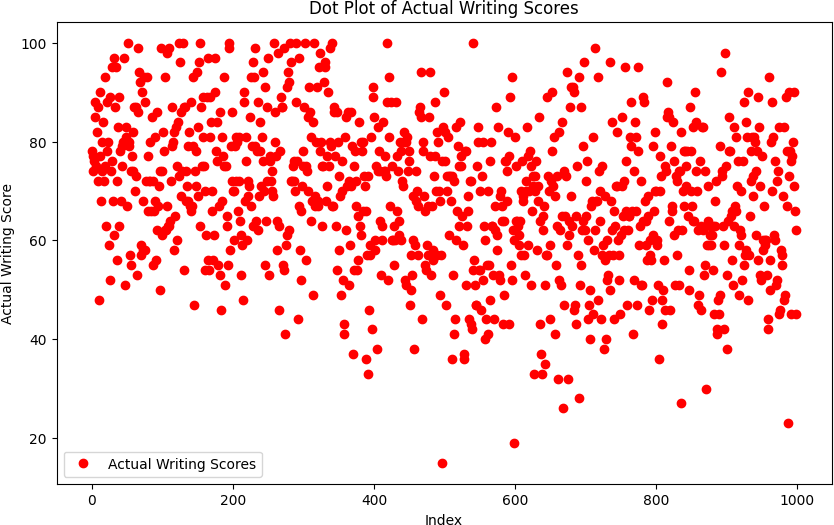
## Distribution of Scores

**Histograms and Box Plots**: Showed score distributions and outliers.



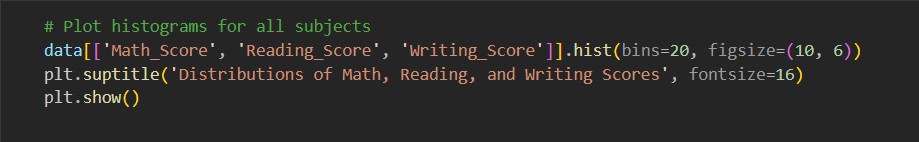


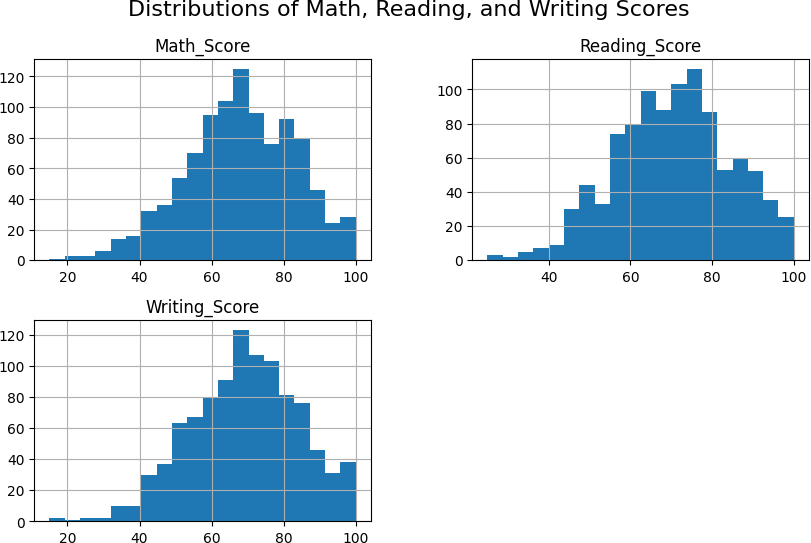




Below is the histogram that make us easy to understand more than dot plot.

The histograms illustrate the distributions of scores in math, reading, and writing for students in the dataset





Here's a detailed breakdown:

## Math Scores:

* + **Shape**: The math scores exhibit a slightly right-skewed distribution, with most scores clustering between 50 and 80.
  + **Peak Frequency**: The highest number of students scored around the 70 mark.
  + **Range**: Scores range from approximately 20 to 100, indicating a wide variation in math performance.
  + **Outliers**: A small number of students scored exceptionally low, pulling the distribution slightly to the right.

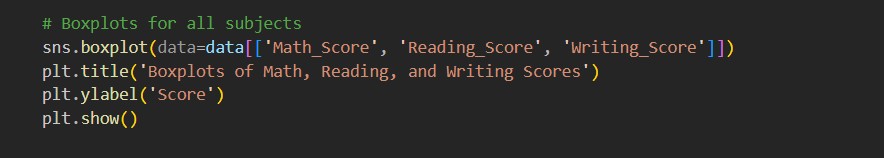
## Reading Scores:

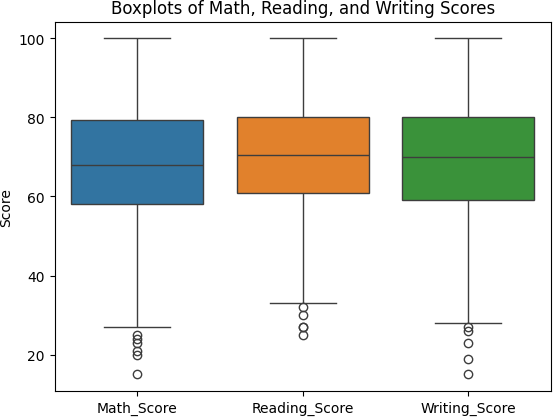
* + **Shape**: The distribution is approximately symmetric, resembling a normal distribution.
  + **Peak Frequency**: The most common scores fall between 70 and 80, suggesting higher performance in reading compared to math.
  + **Range**: Scores range from about 40 to 100, indicating less variation compared to math.

## Writing Scores:

* + **Shape**: Similar to reading scores, writing scores display a near-normal distribution.
  + **Peak Frequency**: The majority of students scored between 65 and 80.
  + **Range**: Scores range from approximately 20 to 100, with a distribution that closely mirrors reading scores.

The box plot provides a comparative summary of the distribution of scores for math, reading, and writing subjects.





Here are the key observations:

## Central Tendency:

* + The median (central line in each box) for math scores is slightly lower than that for reading and writing. This indicates that students generally perform better in reading and writing compared to math.
  + Reading and writing scores have similar medians, reflecting comparable performance levels in these two subjects.

## Range of Scores:

* + Math scores have a wider range (indicated by the whiskers), extending from 15 to

100. This suggests greater variability in math performance.

* + Reading and writing scores are slightly more consistent, with a narrower range compared to math.

## Outliers:

* + Outliers (points outside the whiskers) are visible for all three subjects. These represent students who performed exceptionally lower than the general cohort.
  + Math has the most visible outliers, indicating that a subset of students struggled significantly in this subject.

## Spread (Interquartile Range - IQR):

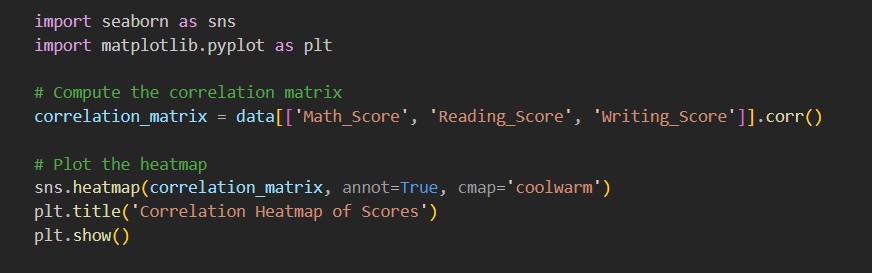
* + The IQR (height of each box) for all three subjects is relatively similar, indicating consistent spread in the middle 50% of scores.
  + Writing has a slightly larger IQR compared to reading, reflecting a broader performance range among average performers.

## Performance Insights:

* + Students appear to perform slightly better in reading and writing compared to math, as indicated by higher medians and fewer extreme low scores.

# Correlation Heatmap

The heatmap visualizes the relationships between the three academic performance metrics: math, reading, and writing scores.





Here's what it reveals:

* **Correlation Strength**:
  + **Math and Reading**: Correlation coefficient of 0.81 indicates a strong positive relationship, meaning students who score well in reading are also likely to perform well in math.
  + **Math and Writing**: Correlation coefficient of 0.79 shows a moderately strong positive relationship between these subjects.
  + **Reading and Writing**: Correlation coefficient of 0.95 highlights an exceptionally strong positive relationship, suggesting that students tend to excel in both literacy- based subjects simultaneously.

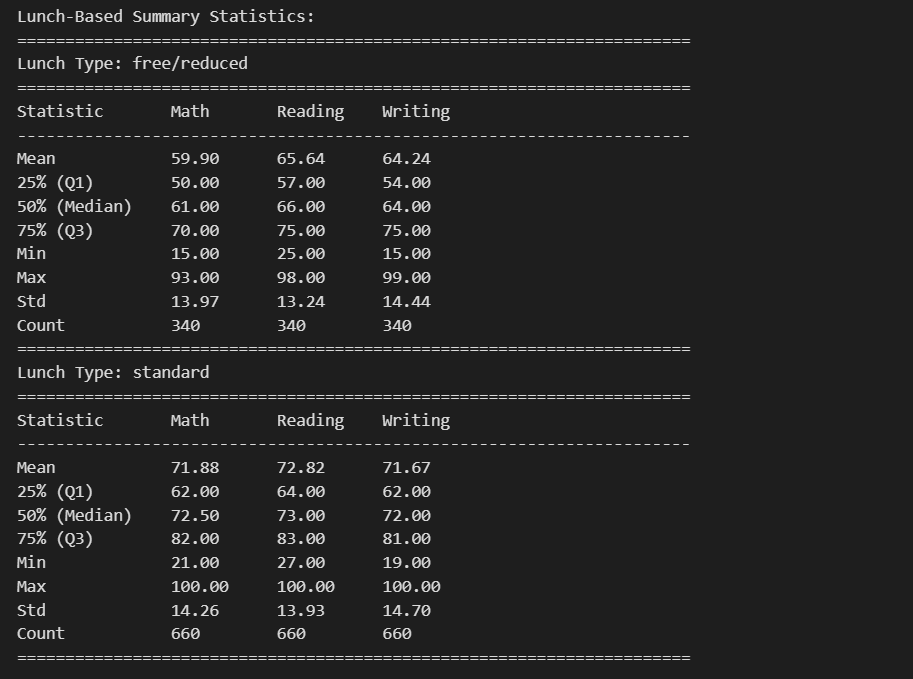
## Implications:

* + The high correlation between reading and writing scores suggests overlapping skill sets or shared influencing factors, such as language proficiency.
  + The slightly lower correlation of math scores with the other two subjects may indicate that mathematical ability is influenced by different cognitive or instructional factors.

## Use in Modeling:

* + These correlations provide insights into feature relationships, which can improve the performance of predictive models by considering how these scores interact.

# Lunch-Based Summary Statistics Analysis



The table presents summary statistics for student performance in math, reading, and writing, categorized by lunch type (free/reduced and standard). Here's a detailed analysis:

**Free/Reduced Lunch**

* **Mean Scores**:
  + Students with free/reduced lunch have lower mean scores: Math (59.90), Reading (65.64), Writing (64.24).

## Score Spread:

* + Scores are distributed in the lower range, with a 75th percentile (Q3) of 70 (Math), 75 (Reading), and 75 (Writing).

## Variability:

* + Standard deviations (Std) indicate moderate variability: Math (13.97), Reading (13.24), Writing (14.44).

## Performance Insight:

* + Students receiving free/reduced lunch generally underperform, likely reflecting challenges linked to socioeconomic factors.

## Standard Lunch

* **Mean Scores**:
  + Students with standard lunch show higher mean scores: Math (71.88), Reading (72.82), Writing (71.67).

## Score Spread:

* + The 75th percentile (Q3) scores are significantly higher: Math (82), Reading (83), Writing (81).

## Variability:

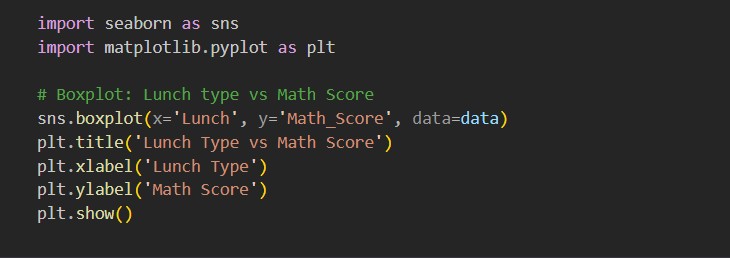
* + Standard deviations are slightly higher: Math (14.26), Reading (13.93), Writing (14.70), indicating a broader spread in performance.

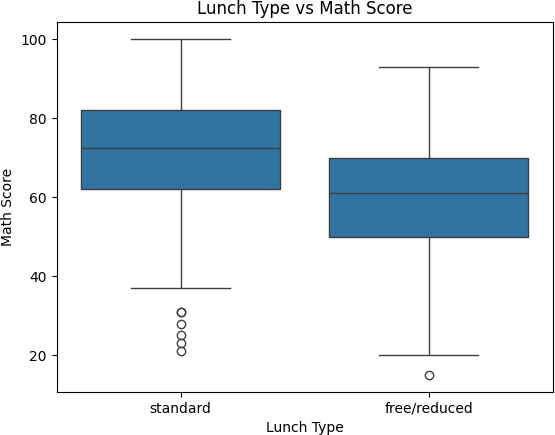
## Performance Insight:

* + Students with standard lunch outperform their peers, suggesting greater stability and access to resources.

## Analysis of the Box Plot: Lunch Type vs. Math Score

This box plot visualizes the relationship between lunch type (standard and free/reduced) and math scores.





Here is the analysis:

## Median Scores:

* + Students with a **standard lunch** have a higher median math score compared to those with **free/reduced lunch**. This indicates better overall performance for students receiving standard lunches.

## Score Spread:

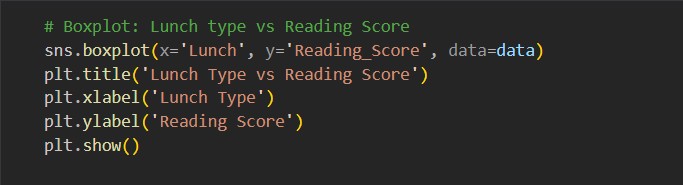
* + **Standard Lunch**: The math scores show a broader range, with most scores concentrated in the upper range. A few outliers are present in the lower scores.
  + **Free/Reduced Lunch**: The score range is narrower, and most students score in the lower range compared to those with standard lunches.

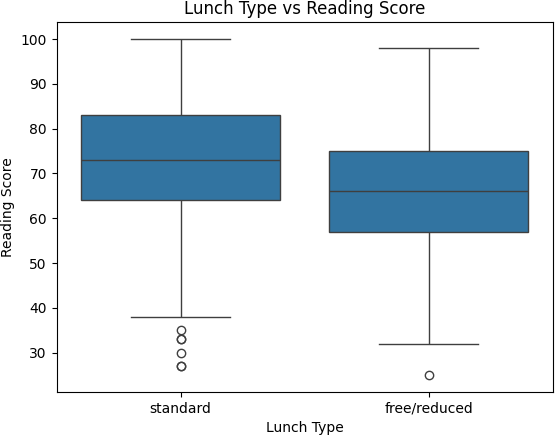
## Outliers:

* + Both groups show outliers, but the **free/reduced lunch** group has fewer, indicating a more consistent but lower performance.

## Analysis of the Box Plot: Lunch Type vs. Reading Score

This box plot illustrates the relationship between lunch type (standard and free/reduced) and reading scores.





*Observations:*

## Median Scores:

* + Students with a **standard lunch** have a higher median reading score compared to those with **free/reduced lunch**. This indicates better reading performance for students with standard lunch access.

## Score Range:

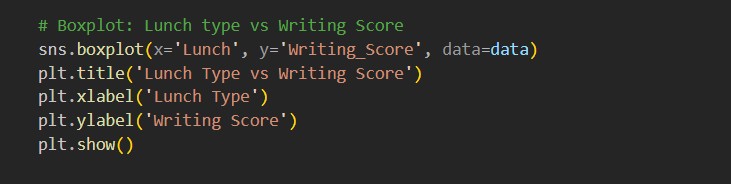
* + **Standard Lunch**: The reading scores are distributed across a broader range, with most scores concentrated in the upper range. There are some outliers at the lower end of the distribution.
  + **Free/Reduced Lunch**: The score range is narrower, and the concentration of scores is more focused in the lower ranges compared to students with standard lunches.

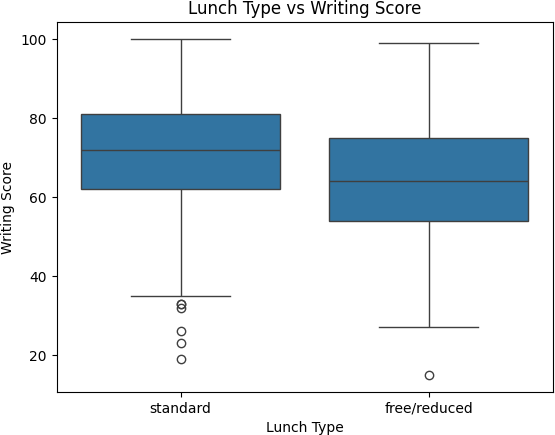
## Outliers:

* + Both groups have outliers, but the **free/reduced lunch** group has fewer, indicating relatively consistent but lower performance overall.

Analysis of the Box Plot: Lunch Type vs. Writing Score

This box plot visualizes the relationship between lunch type (standard and free/reduced) and writing scores. Here's the detailed analysis:





*Observations:*

## Median Scores:

* + Students with a **standard lunch** have a higher median writing score compared to those with **free/reduced lunch**, indicating better writing performance for students with standard lunch access.

## Score Range:

* + **Standard Lunch**: Writing scores show a broader distribution, with most scores concentrated in the higher ranges and some outliers in the lower range.
  + **Free/Reduced Lunch**: Scores are more narrowly distributed, with lower overall performance compared to the standard lunch group.

## Outliers:

* + Outliers are present in both groups. The **free/reduced lunch** group has a prominent low outlier, suggesting some students face significant challenges in writing performance.

Summary of Average Scores by Lunch Type

This table provides a summary of the mean scores for math, reading, and writing, grouped by lunch type (free/reduced and standard).



*Observations:*

## Free/Reduced Lunch:

* + **Math Score**: The average is **59.90**, indicating lower performance.
  + **Reading Score**: The average is **65.64**, showing moderate performance.
  + **Writing Score**: The average is **64.24**, slightly higher than math but still below the standard lunch group.

## Standard Lunch:

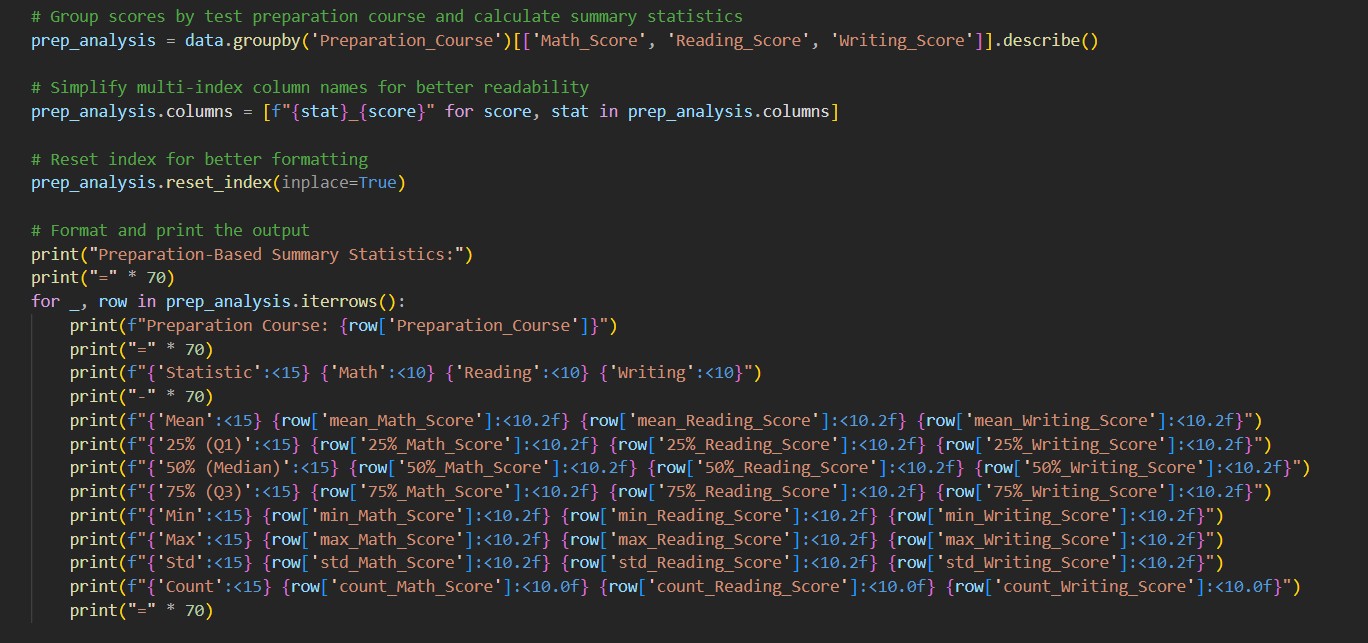
* + **Math Score**: The average is **71.88**, significantly higher than the free/reduced lunch group.
  + **Reading Score**: The average is **72.82**, indicating strong performance.
  + **Writing Score**: The average is **71.67**, reflecting consistently high scores across all subjects.

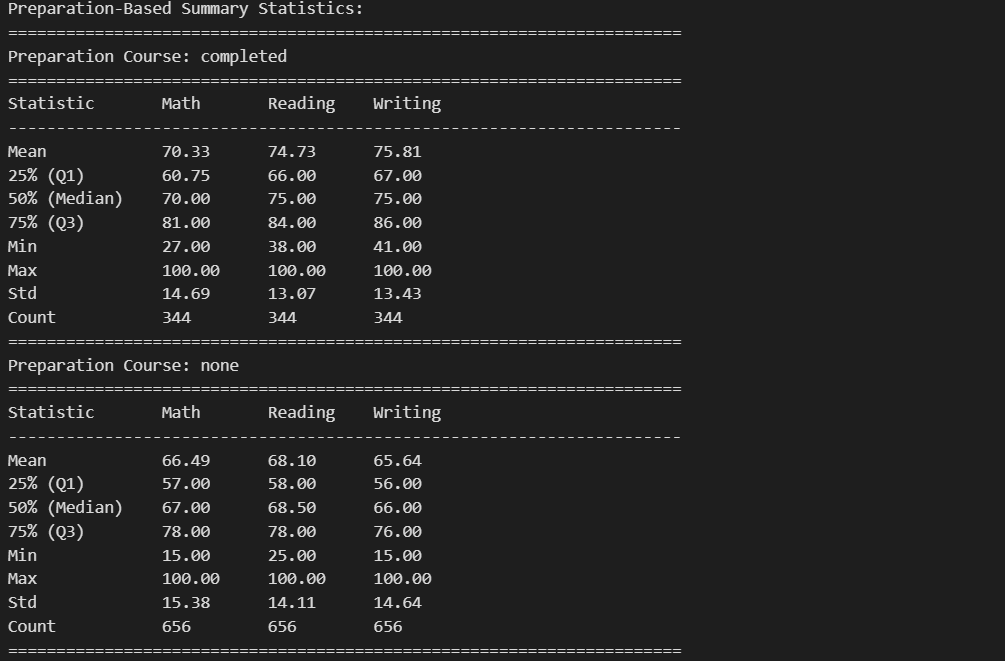
## Overall Conclusion

The analysis reveals a clear disparity in academic performance based on lunch type. Students who receive standard lunches consistently outperform those on free/reduced lunches across all subjects: math, reading, and writing. This performance gap underscores the significant influence of socioeconomic factors on educational outcomes.

## Preparation-Based Summary Statistics

The table highlights the mean scores and key statistics for students who completed the test preparation course compared to those who did not.





***Students Who Completed the Preparation Course***

* **Mean Scores**: Students who completed the course scored higher on average:
  + **Math**: 70.33
  + **Reading**: 74.73
  + **Writing**: 75.81

## Score Distribution:

* + The 25th percentile (Q1) scores in all subjects are significantly higher than those for students who did not complete the course.
  + The 75th percentile (Q3) scores also reflect stronger performance, especially in reading and writing.

## Variability:

* + Standard deviations (Std) are slightly lower compared to those who did not complete the course, suggesting more consistent performance.

***Students Who Did Not Complete the Preparation Course***

## Mean Scores:

* + **Math**: 66.49
  + **Reading**: 68.10
  + **Writing**: 65.64

## Score Distribution:

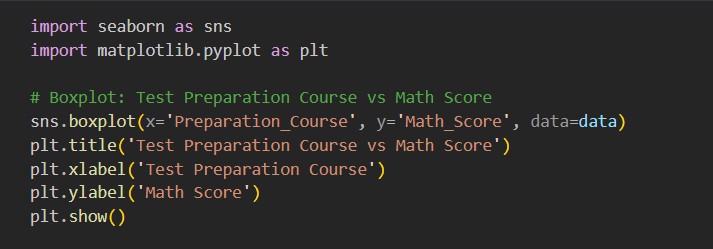
* + Lower median and 25th percentile (Q1) scores across all subjects compared to the group that completed the course.
  + A wider range in scores, particularly in writing, with a higher presence of lower scores.

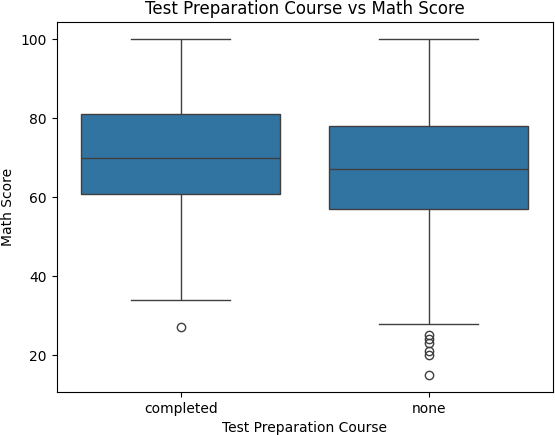
## Variability:

* + Standard deviations are slightly higher, indicating less consistency in performance among this group.

## Analysis of the Box Plot: Test Preparation Course vs Math Score

his box plot visualizes the relationship between completing a test preparation course and math scores. Here's the analysis:





*Observations:*

## Median Scores:

* + Students who completed the preparation course have a higher median math score than those who did not. This indicates the effectiveness of structured preparation in improving performance.

## Score Distribution:

* + **Completed**: Scores are concentrated in the higher range, with fewer low scores and minimal outliers.
  + **Not Completed**: Scores show greater variability, with more low outliers pulling the lower whisker down.

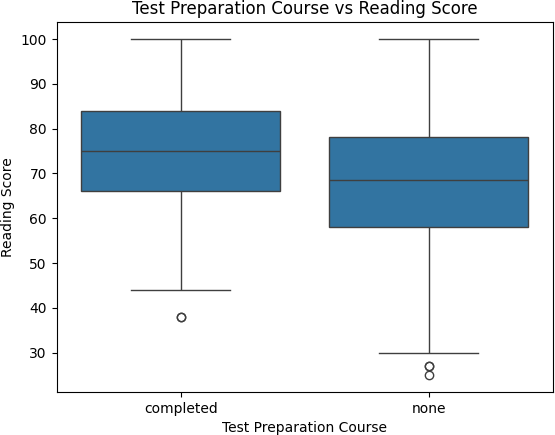
## Outliers:

* + Students who did not complete the course exhibit several low outliers, suggesting a subset of this group struggled significantly in math.

### Analysis of the Box Plot: Test Preparation Course vs Reading Score

This box plot illustrates the relationship between completing a test preparation course and reading scores.





*Observations:*

## Median Scores:

* + Students who completed the preparation course have a higher median reading score compared to those who did not, indicating the positive impact of preparation on reading performance.

## Score Distribution:

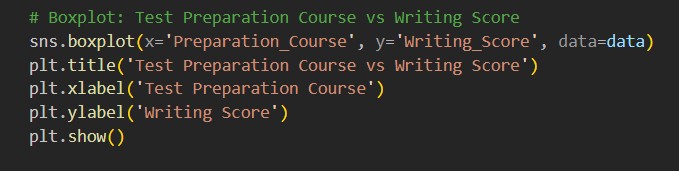
* + **Completed**: The scores are tightly clustered in the higher range, showing consistent and strong performance.
  + **Not Completed**: Scores are more dispersed with a broader range, and lower scores are more prevalent.

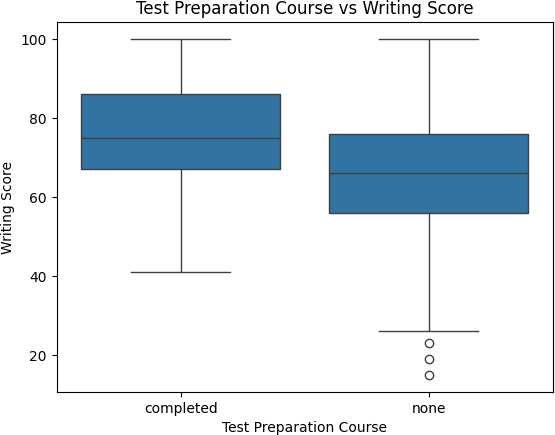
## Outliers:

* + Students who did not complete the preparation course have more low outliers, suggesting a subset struggled significantly in reading.

### Analysis of the Box Plot: Test Preparation Course vs Writing Score

his box plot visualizes the relationship between completing a test preparation course and writing scores.





*Observations:*

## Median Scores:

* + Students who completed the test preparation course have a higher median writing score compared to those who did not. This highlights the effectiveness of preparation in enhancing writing skills.

## Score Distribution:

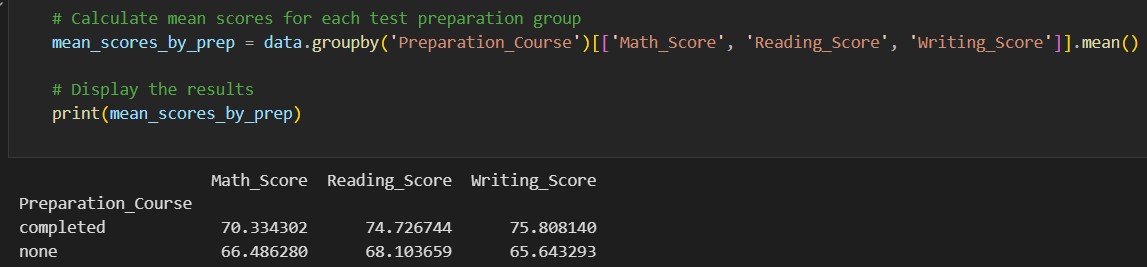
* + **Completed**: Scores are concentrated in the higher range, reflecting better overall performance with fewer low scores.
  + **Not Completed**: Scores show greater variability, with a significant portion of students scoring lower.

## Outliers:

* + Students who did not complete the preparation course exhibit several low outliers, indicating notable struggles in writing.

### Average Scores by Test Preparation Course Completion

The table provides the mean scores for math, reading, and writing based on test preparation course completion



*Completed Test Preparation Course:*

* **Math Score**: Average score is **70.33**, indicating better performance compared to those who did not complete the course.
* **Reading Score**: Average score is **74.73**, reflecting strong performance in reading.
* **Writing Score**: Average score is **75.81**, the highest among the three subjects, suggesting a significant benefit from the course in literacy skills.

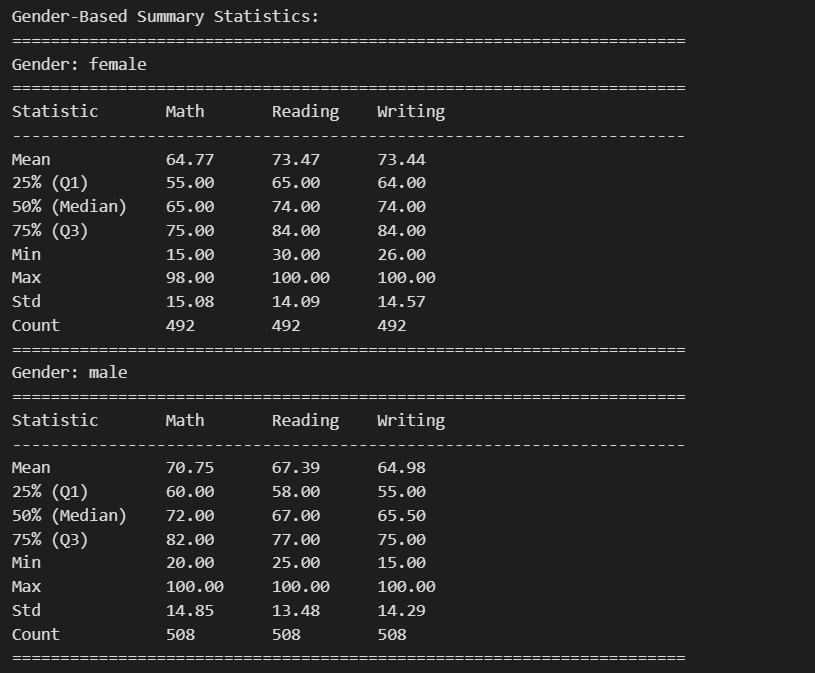
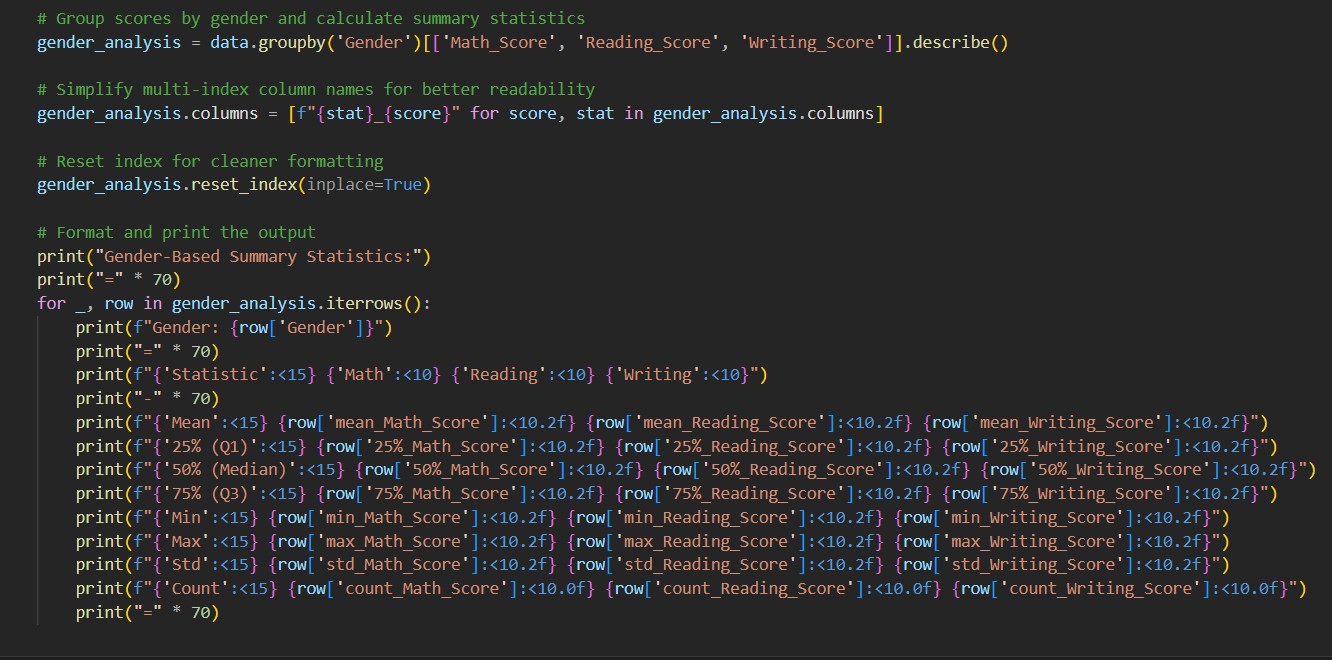
*Did Not Complete Test Preparation Course:*

* **Math Score**: Average score is **66.49**, lower than those who completed the course.
* **Reading Score**: Average score is **68.10**, showing a noticeable gap compared to the completed group.
* **Writing Score**: Average score is **65.64**, the lowest among the three subjects, highlighting the need for additional support in writing for this group.

### Overall

Completing the test preparation course leads to better performance across all subjects, with the most significant improvement observed in writing scores.

## Gender Statistics and Analysis



This table provides insights into academic performance across math, reading, and writing scores, categorized by gender.

*Female Students:*

### Mean Scores:

* + **Math**: 64.77
  + **Reading**: 73.47
  + **Writing**: 73.44
  + Female students excel in reading and writing compared to male students.

### Score Distribution:

* + The **25th percentile (Q1)** and **75th percentile (Q3)** in reading and writing are higher than in math, indicating stronger performance in literacy-related subjects.
  + The median scores (50th percentile) in reading and writing (74) are significantly higher than in math (65).

### Variability:

* + Standard deviations are slightly higher for writing, reflecting moderate variability in performance among female students.

*Male Students:*

### Mean Scores:

* + **Math**: 70.75
  + **Reading**: 67.39
  + **Writing**: 64.98
  + Male students outperform female students in math but lag in reading and writing.

### Score Distribution:

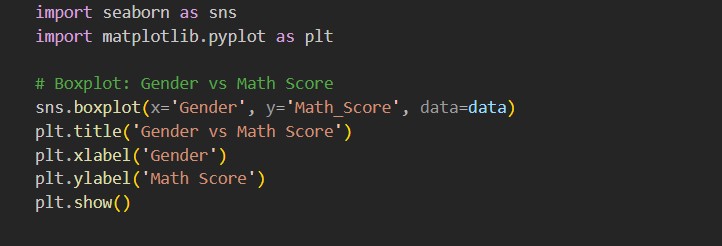
* + Higher **25th percentile (Q1)** and **75th percentile (Q3)** scores in math reflect stronger overall performance in numerical skills.
  + The median score for math (72) is higher than for reading (67) and writing (65.5).

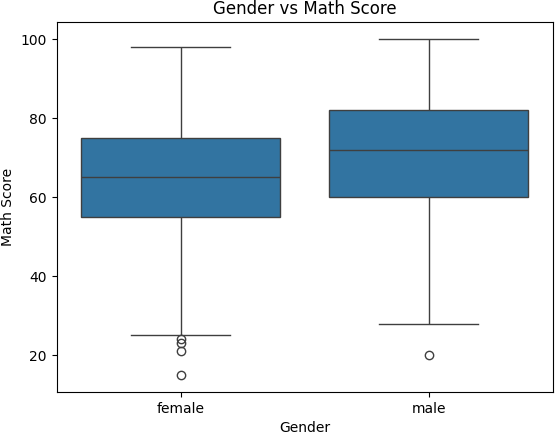
### Variability:

* + Standard deviations are consistent across subjects, suggesting uniform performance among male students.

### Analysis of the Box Plot: Gender vs Math Score

This box plot visualizes the relationship between gender and math scores. Here's a detailed analysis:





## Median Scores:

* + Male students have a higher median math score compared to female students. This highlights stronger numerical performance among male students.

## Score Range:

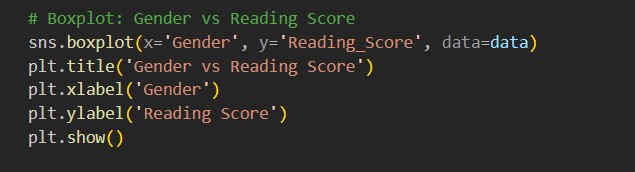
* + **Males**: The scores are distributed over a broader range, with higher scores concentrated in the upper quartile.
  + **Females**: Scores are slightly more condensed, with lower quartile values pulling down the overall range.

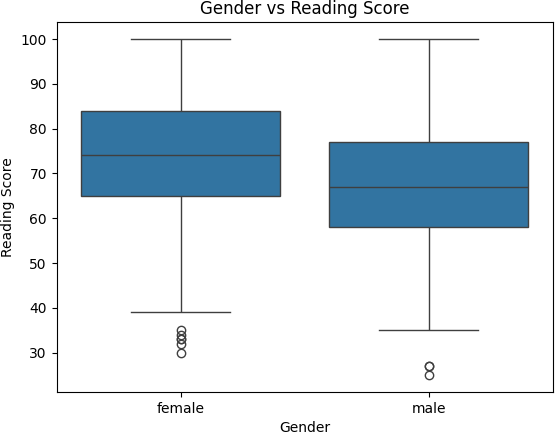
## Outliers:

* + Both groups have outliers, but female students exhibit more low outliers compared to male students, suggesting some females struggle more with math.

Analysis of the Box Plot: Gender vs Reading Score

This box plot illustrates the relationship between gender and reading scores. Here's the detailed analysis:





## Median Scores:

* + - Female students have a higher median reading score compared to male students, indicating stronger literacy skills among females.

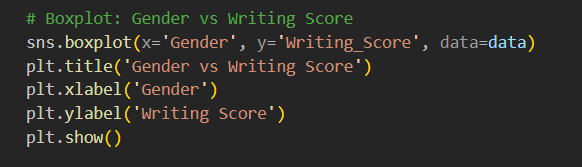
## Score Range:

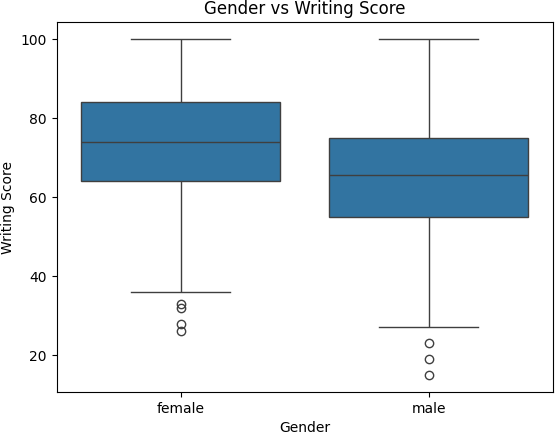
* + - **Females**: Scores are more concentrated in the upper range, with fewer students scoring below the 25th percentile (Q1).
    - **Males**: Scores are more evenly distributed but are generally lower compared to females.

## Outliers:

* + - Both genders show outliers, but female students exhibit a higher number of low outliers, suggesting a subset of females facing challenges in reading.

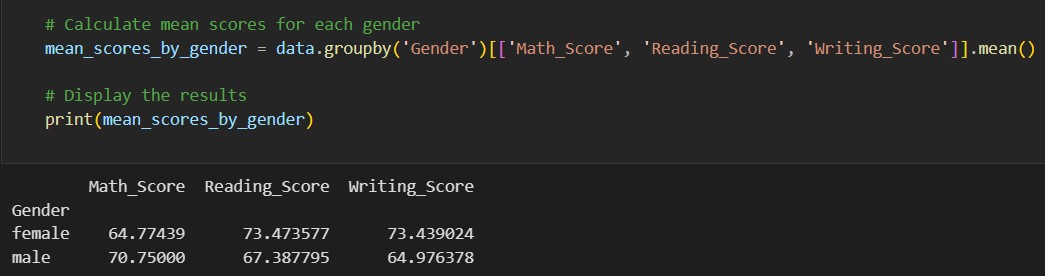
**Analysis of the Box Plot: Gender vs Writing Score**





## Average Scores by Gender

the mean scores for math, reading, and writing, grouped by gender. Here's the analysis:



*Female Students:*

* + **Math Score**: Average score is **64.77**, lower compared to males.
  + **Reading Score**: Average score is **73.47**, higher than males, reflecting stronger literacy skills.
  + **Writing Score**: Average score is **73.44**, significantly higher than males, showcasing female students' strength in writing.

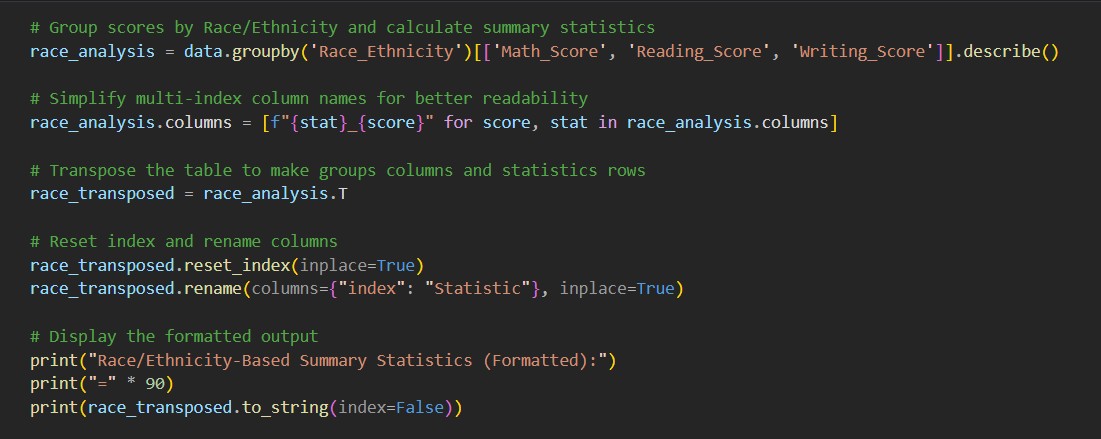
*Male Students:*

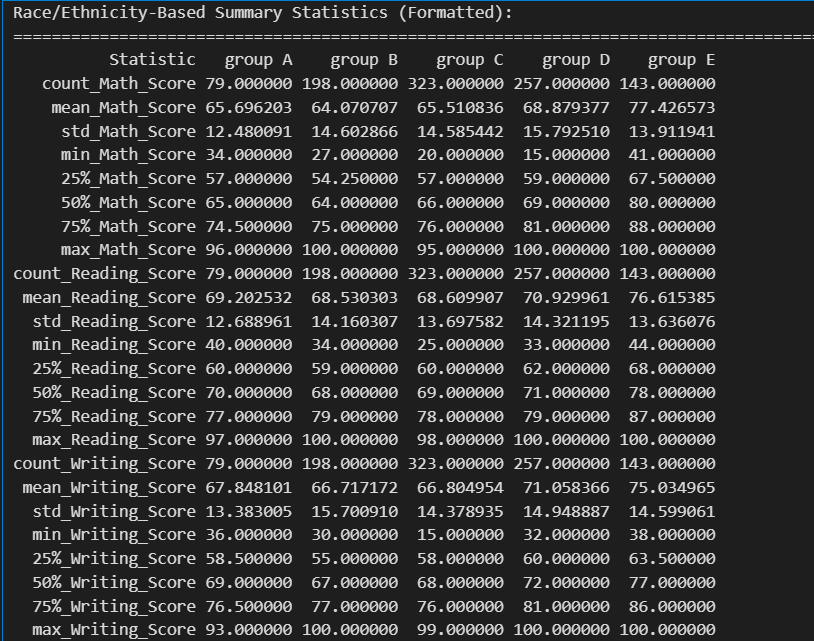
* + **Math Score**: Average score is **70.75**, outperforming females in math.
  + **Reading Score**: Average score is **67.39**, lower than females, indicating challenges in literacy.
  + **Writing Score**: Average score is **64.98**, lower than females, further highlighting the need for support in literacy-based skills.

**Overall**

* + The data suggests that male students tend to excel in math, while female students dominate in reading and writing. These patterns reflect differing strengths and areas of focus for each gender.
  + Variability in scores and outliers indicate opportunities for targeted interventions to support struggling students.

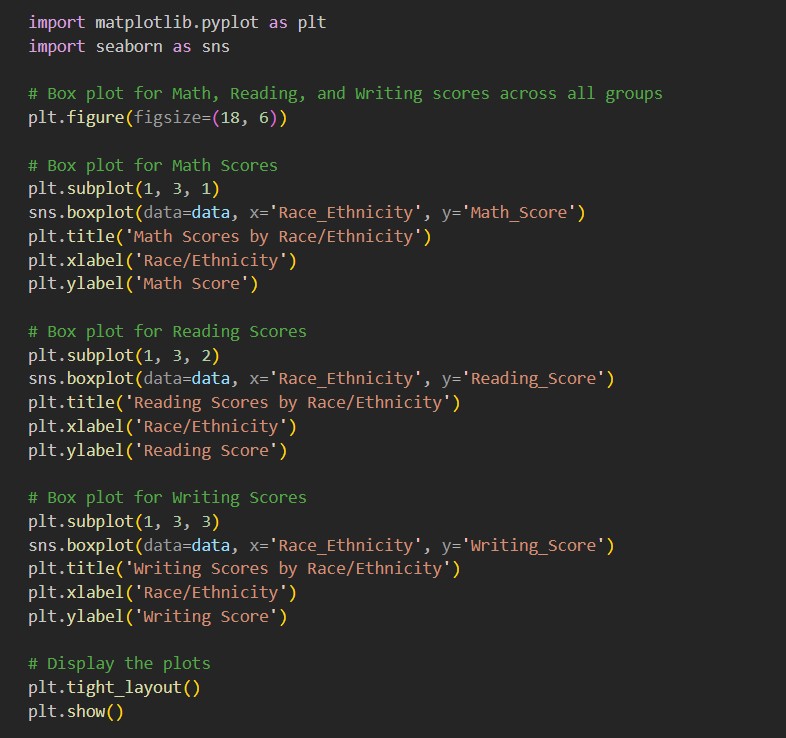
**Race/Ethnicity-Based Summary Statistics Analysis**

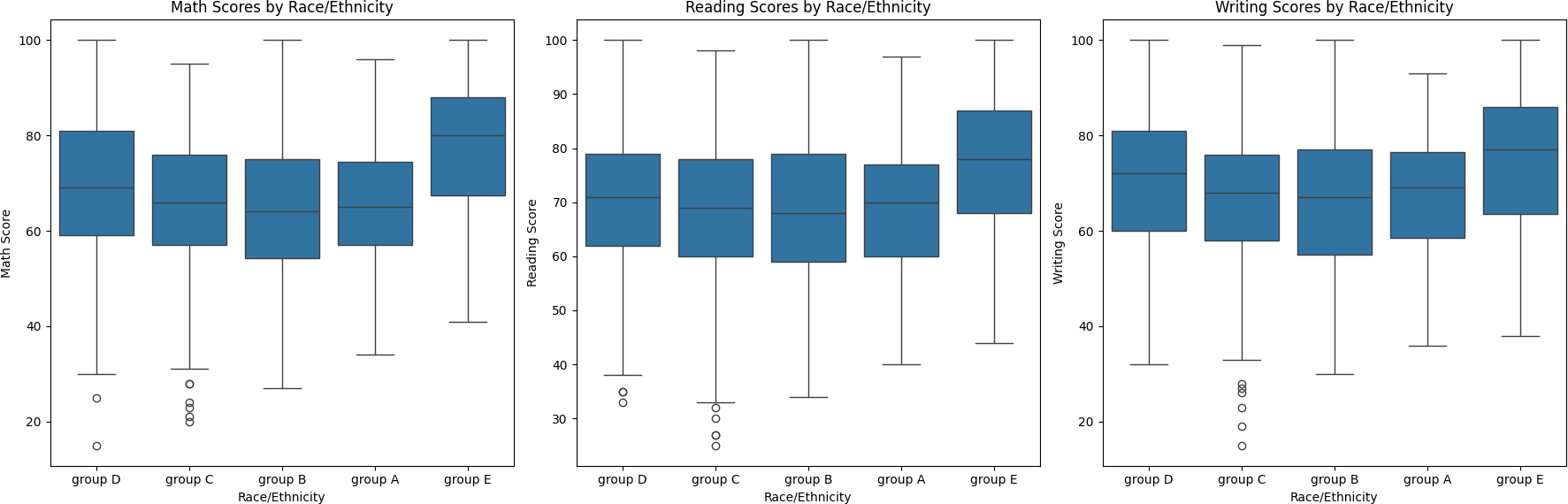




## Analysis of Box Plots: Academic Scores by Race/Ethnicity

These box plots illustrate the distribution of math, reading, and writing scores across five racial/ethnic groups (Groups A to E). Here's the detailed analysis:





**Math Scores***:*

* + **Median Performance**:
    - Group E has the highest median math score, followed by Group D. Group A has the lowest median score, reflecting a significant gap.

### Score Range:

* + - Group E has the most consistent performance, with fewer low scores.
    - Group A exhibits more variability, with scores spanning a broader range and several low outliers.

### Outliers:

* + - Group C and Group A have notable outliers, indicating a subset of students facing challenges in math.

*Reading Scores:*

### Median Performance:

* + - Group E again leads with the highest median reading score, followed by Group D. Group A and Group B have similar, lower medians.

### Score Range:

* + - Groups D and E show better concentration of higher scores, while Groups A and B exhibit wider variability.

### Outliers:

* + - Groups C and A have more low outliers in reading, indicating struggles in literacy skills for a subset of students.

*Writing Scores:*

### Median Performance:

* + - Group E performs best in writing, with the highest median score, followed by Group D. Groups A, B, and C lag behind.

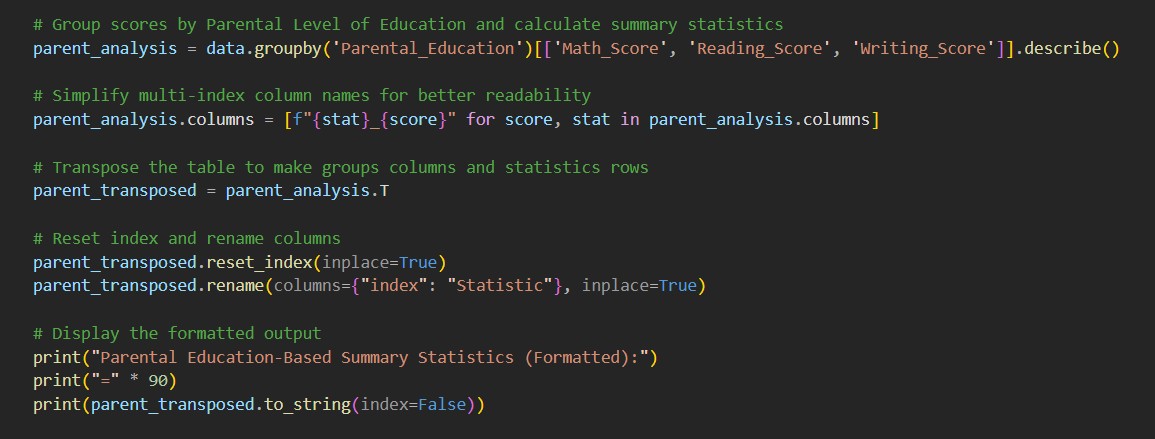
### Score Range:

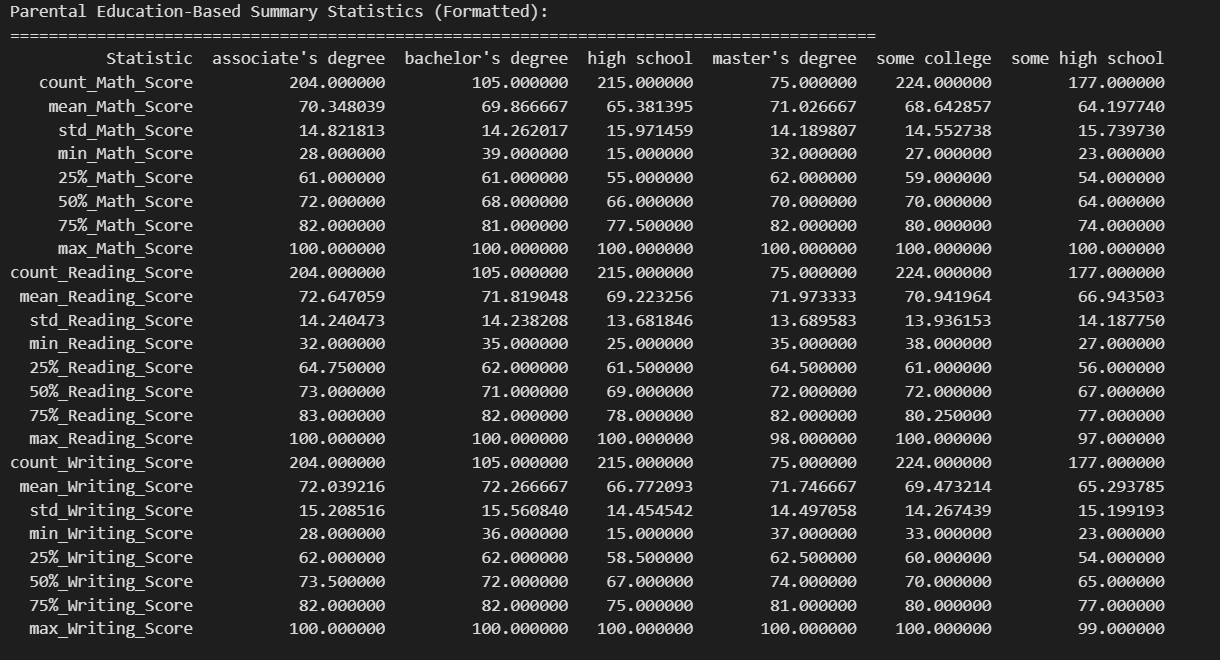
* + - Groups D and E have a narrower range, reflecting consistent performance.
    - Groups A, B, and C show wider variability, with several low outliers, particularly in Group C.

### Outliers:

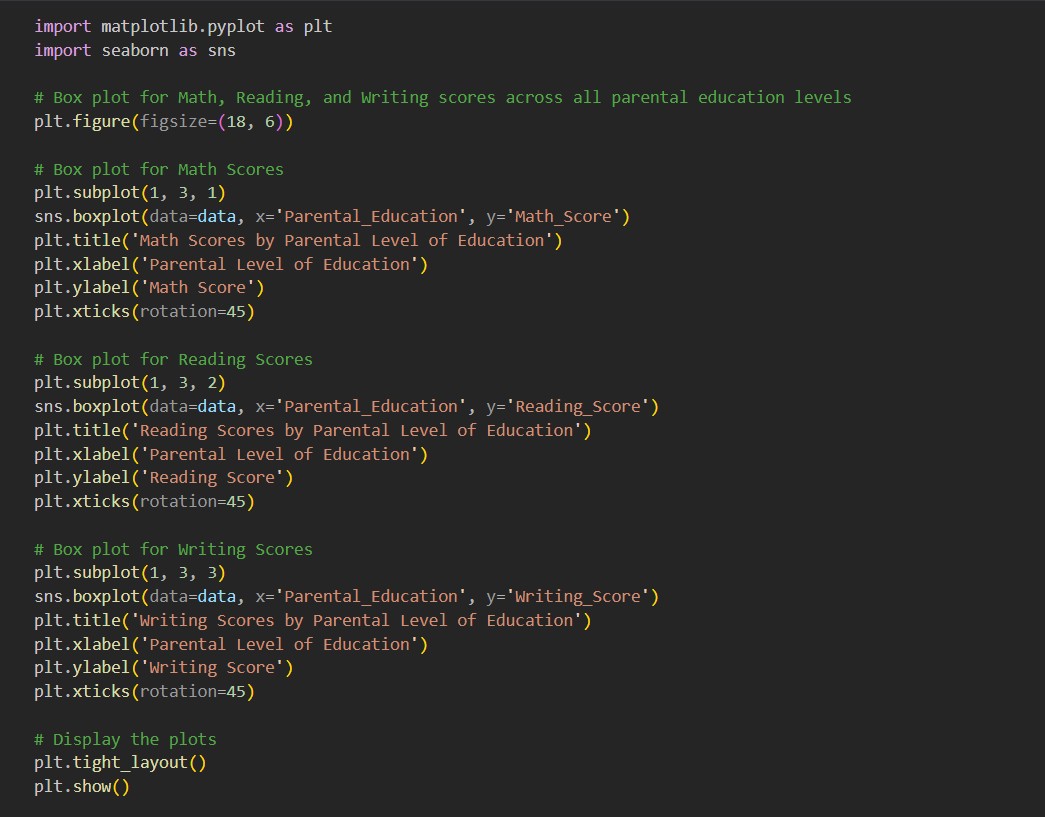
* + - Low outliers are more prevalent in Groups A and C, suggesting specific challenges in writing for some students.

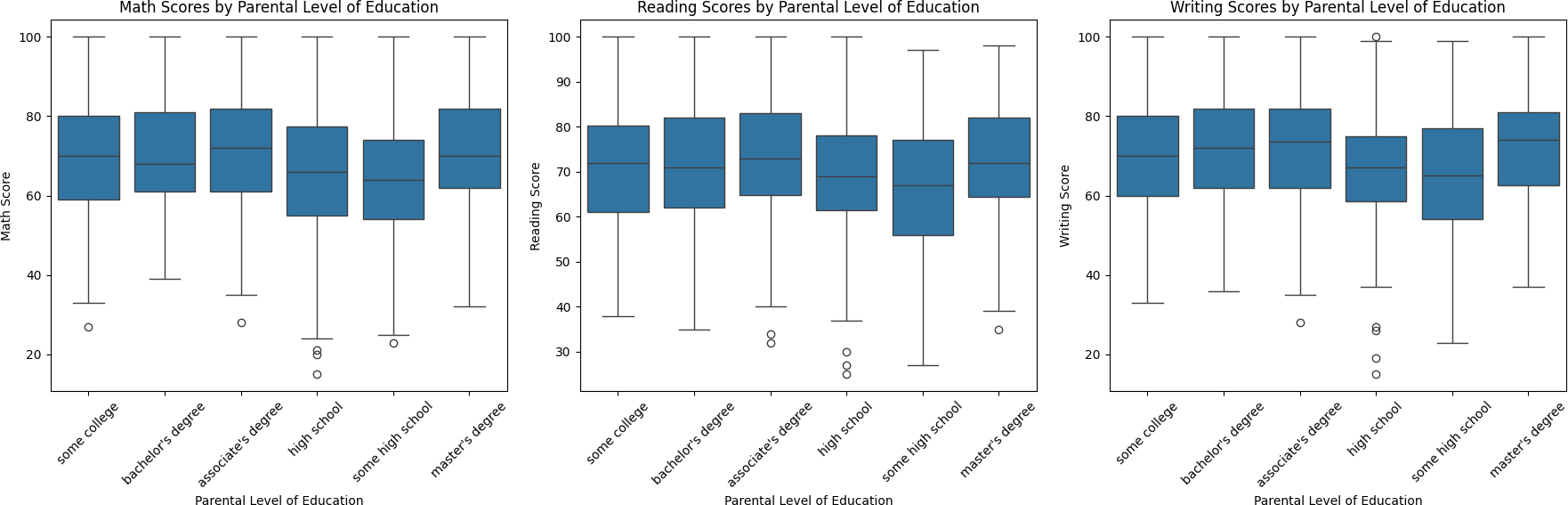
### Analysis of Academic Performance Based on Parental Education Level





**Analysis of Box Plots: Academic Scores by Parental Level of Education**





*Math Scores:*

### Highest Scores:

* + - Students whose parents have a **master's degree** (Mean: **71.03**) and **associate's degree**

(Mean: **70.35**) achieved the highest math scores.

### Lowest Scores:

* + - Students whose parents have **some high school** education (Mean: **64.20**) performed the lowest.

### Score Distribution:

* + - Students with higher parental education levels tend to have narrower ranges of scores, reflecting consistency.

*Reading Scores:*

### Highest Scores:

* + - Students with parents holding a **bachelor's degree** (Mean: **71.82**) and **master's degree**

(Mean: **71.97**) scored the highest.

### Lowest Scores:

* + - Students whose parents have **some high school** education (Mean: **66.94**) showed lower reading proficiency.

### Variability:

* + - Standard deviations are generally consistent across groups, but slightly higher for students with parents with lower education levels.

*Writing Scores:*

### Highest Scores:

* + - Students with parents having a **master's degree** (Mean: **71.75**) and **associate's degree**

(Mean: **72.04**) performed the best.

### Lowest Scores:

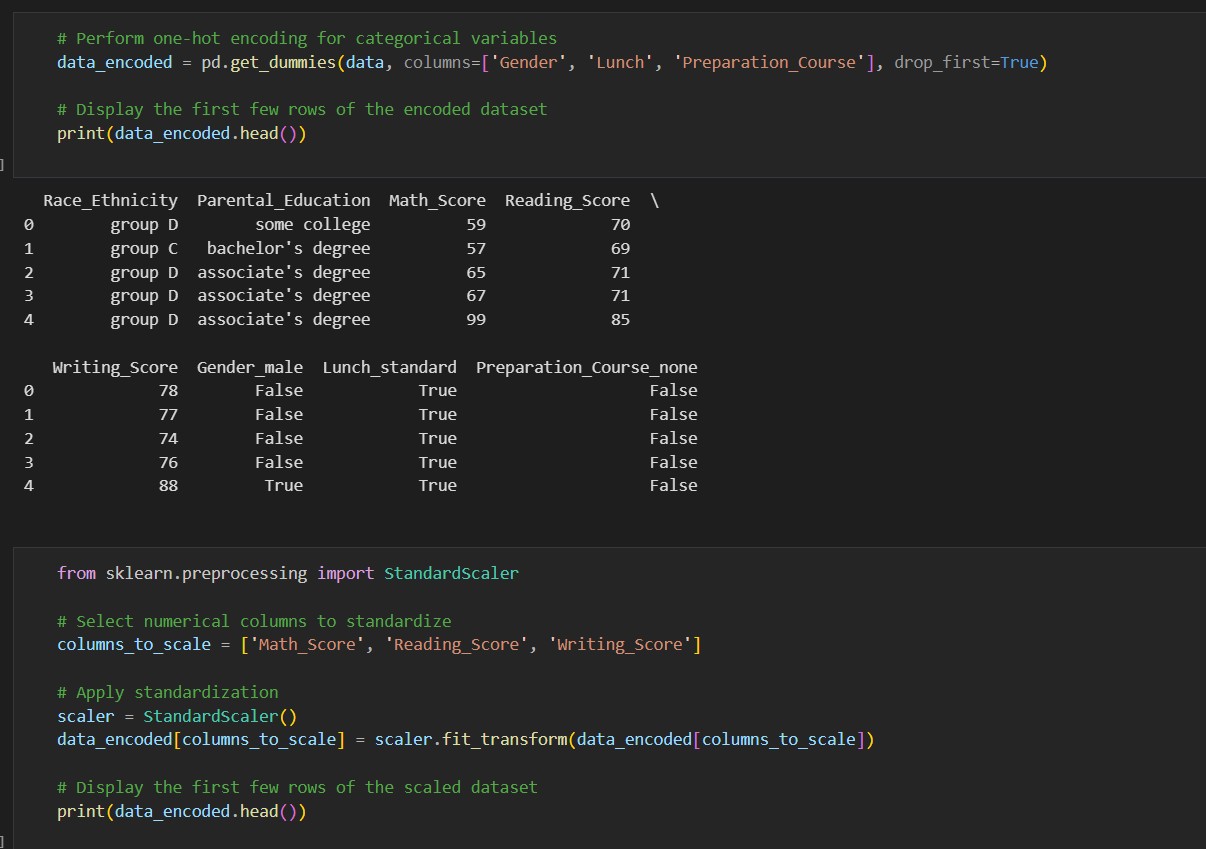
* + - Students whose parents have **some high school** education (Mean: **65.29**) had the lowest writing scores.

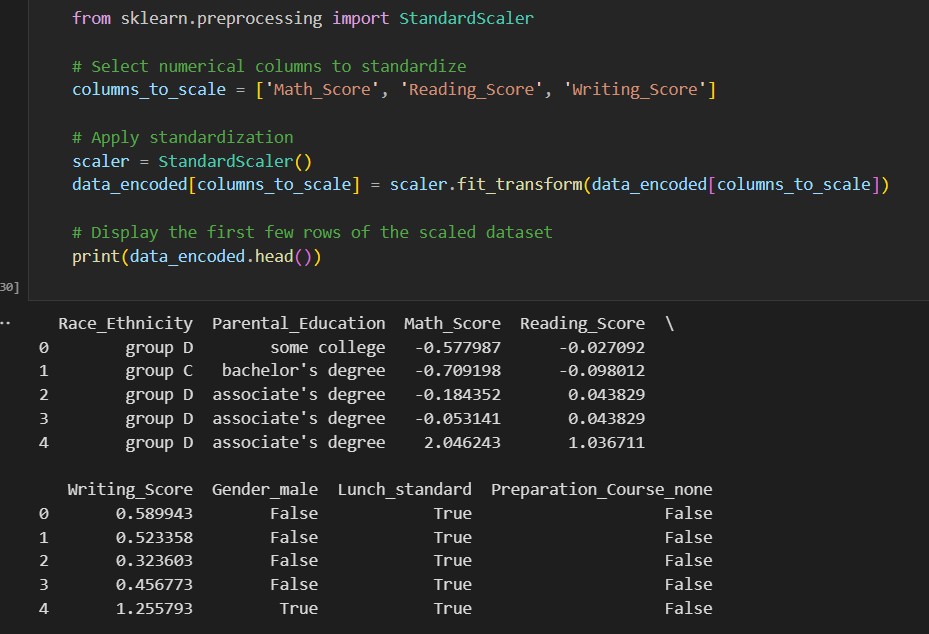
### Variability:

* + - Students with parents with **some high school** education had a wider range of scores, reflecting variability in performance.

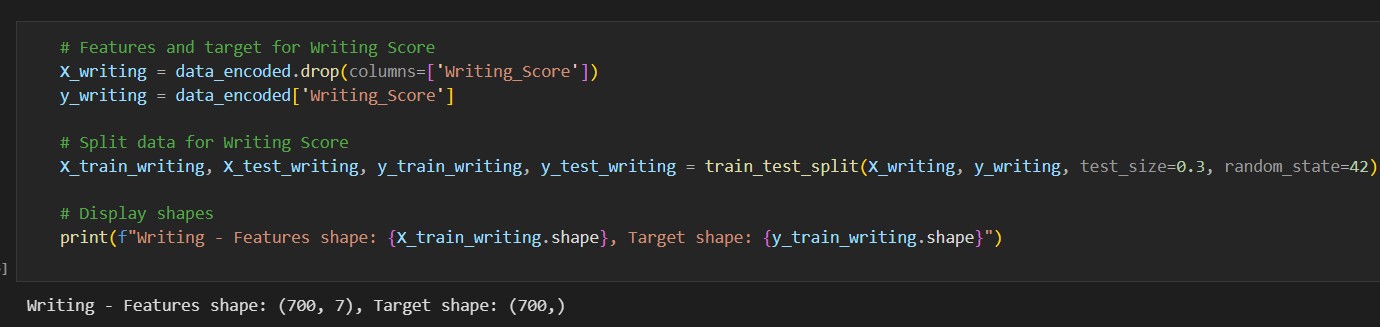
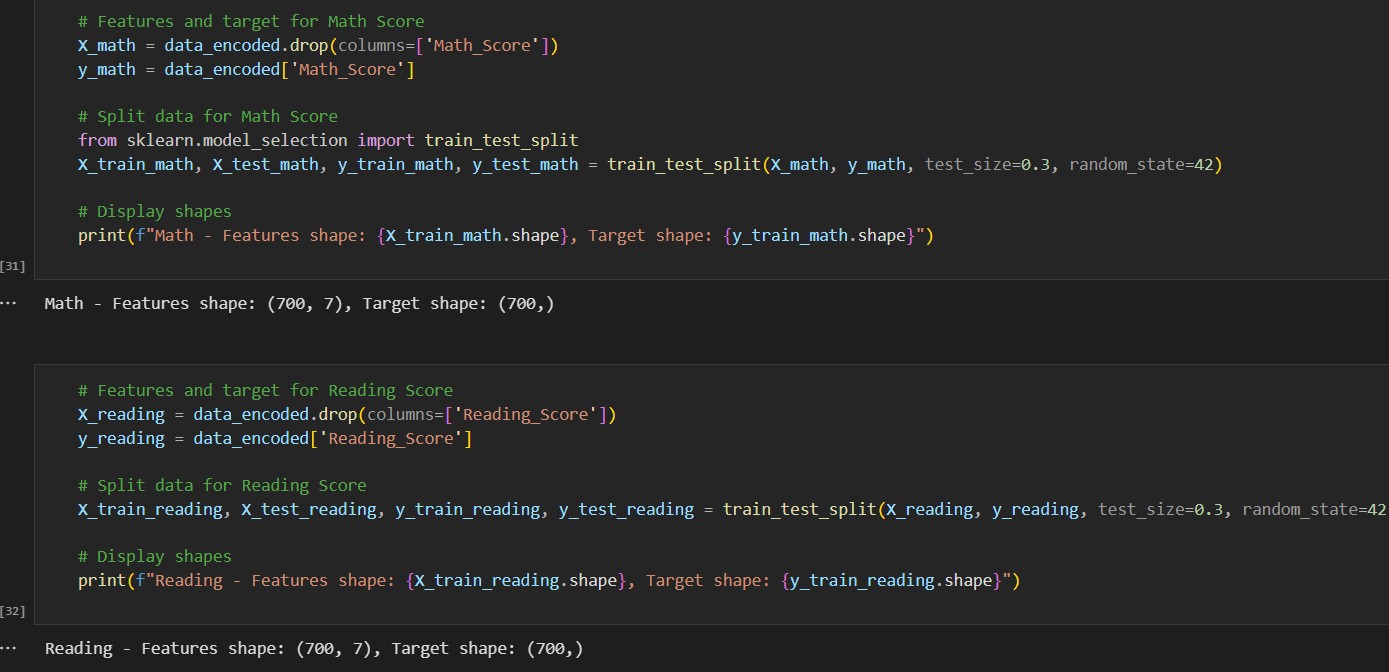
## Features Engineering

Feature engineering involves preparing and transforming the dataset to make it suitable for analysis and machine learning models. Here, one-hot encoding is performed to handle categorical variables, converting them into numerical values while preserving the categorical information. Below is the processing of doing.





This code snippet outlines the process of preparing and splitting the data for training and testing a model to predict the **Writing,reading,math Score**.

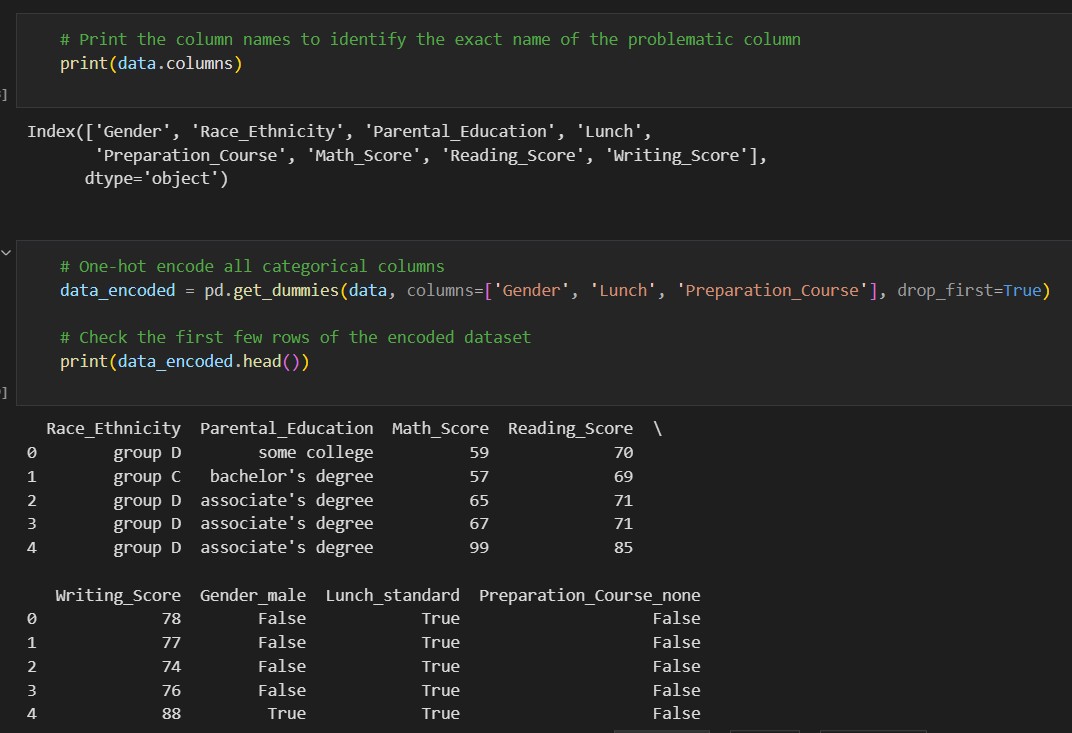


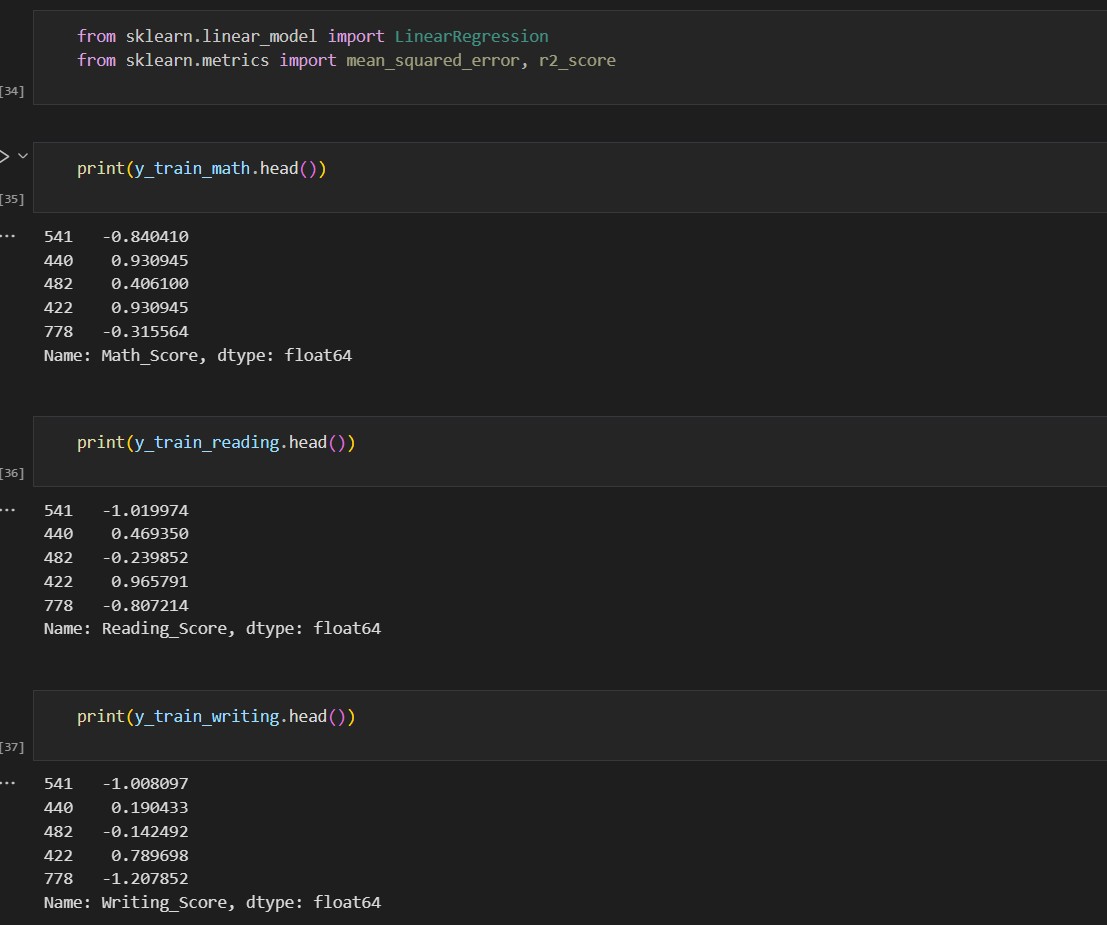
## Model Building

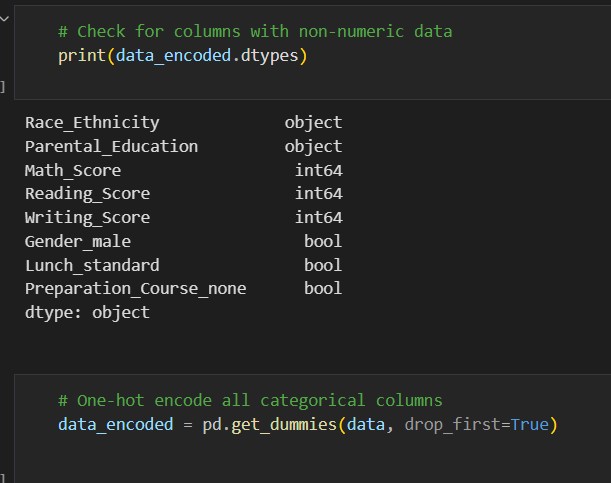
The model building step involves training machine learning models to predict student performance in Math, Reading, and Writing based on the features created during the feature engineering process. The notebook implemented three types of models: Linear Regression, Random Forest Regression, and Support Vector Regression. These models were trained, tested, and evaluated using well-defined metrics to ensure robust predictions.

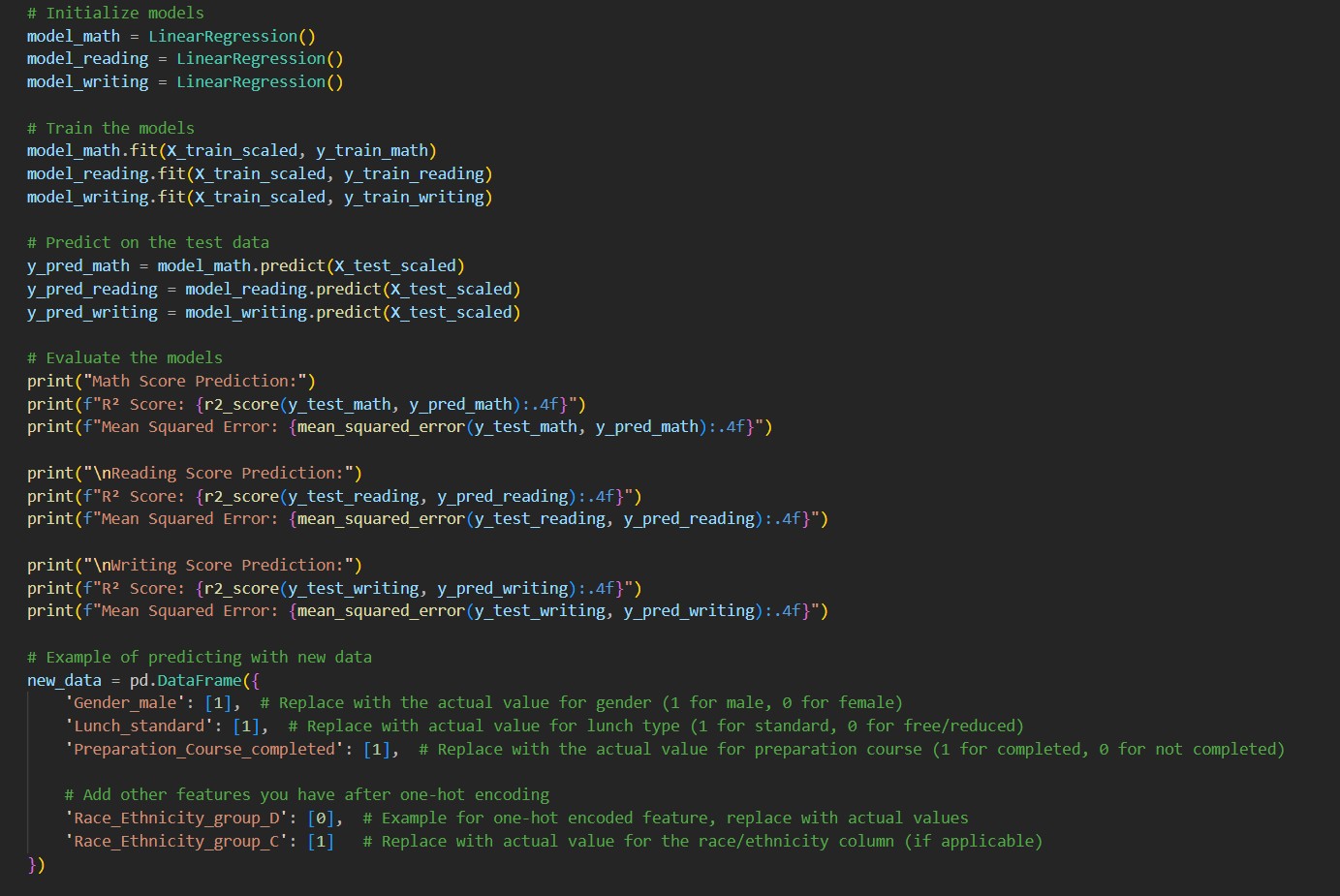
### Linear Regression:

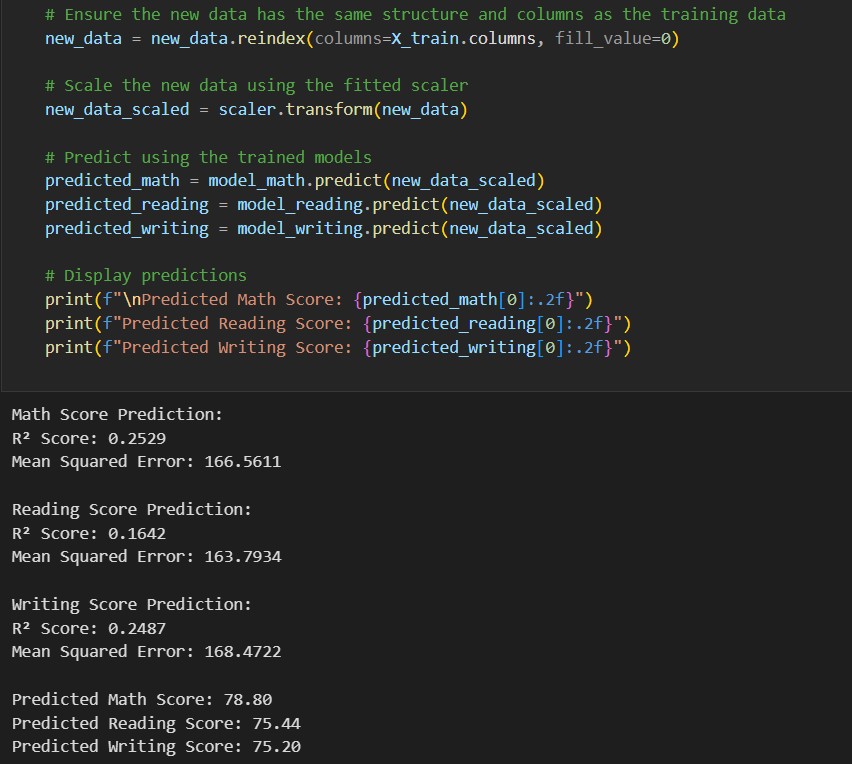
### Objective: Implements Linear Regression models to predict Math, Reading, and Writing scores.

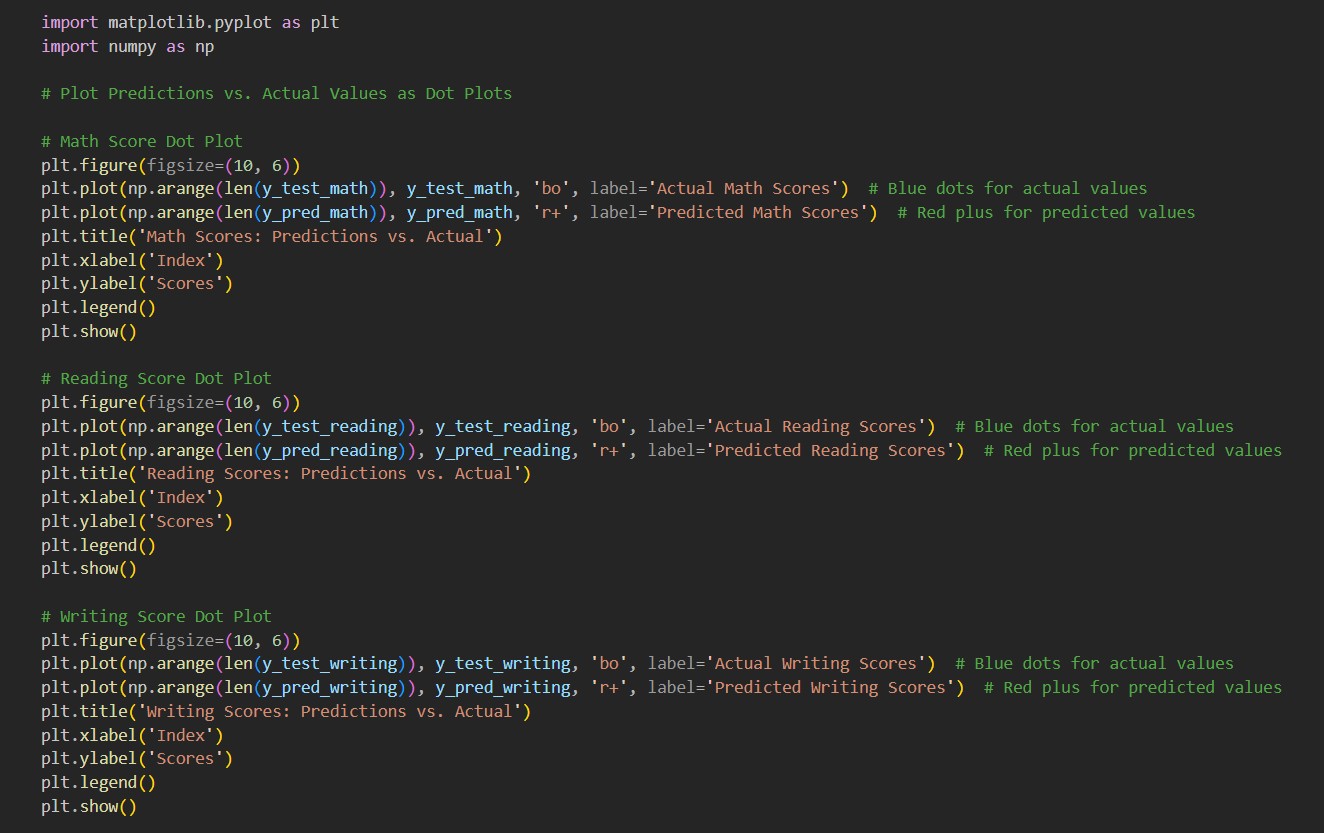




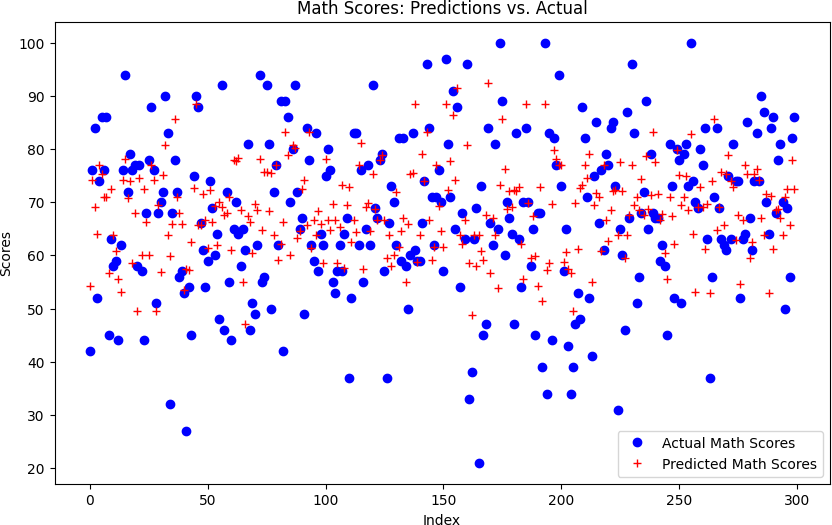






**Analysis of the Plot**

## Analysis of the Plot: Predictions vs. Actual Math Scores



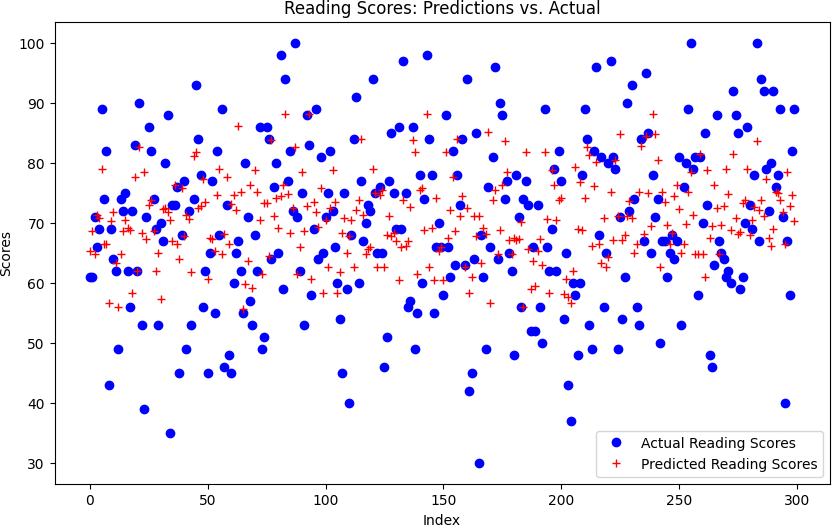
* + **Scatter Distribution**:
    - The blue dots represent the **actual math scores**, while the red crosses represent the **predicted scores**.
    - The predicted scores (red) closely align with the actual scores (blue), indicating that the model performs reasonably well.

## Prediction Spread:

* + - There is some variability in predictions, especially for mid-range scores, where the predicted values slightly deviate from actual ones.
    - Extreme values (both low and high scores) show a slight divergence between actual and predicted scores, which may indicate areas for model improvement.

## Clustered Pattern:

* + - Most data points fall within the 50-80 score range, showing that the majority of students achieved moderate to high scores.



## Alignment of Predictions:

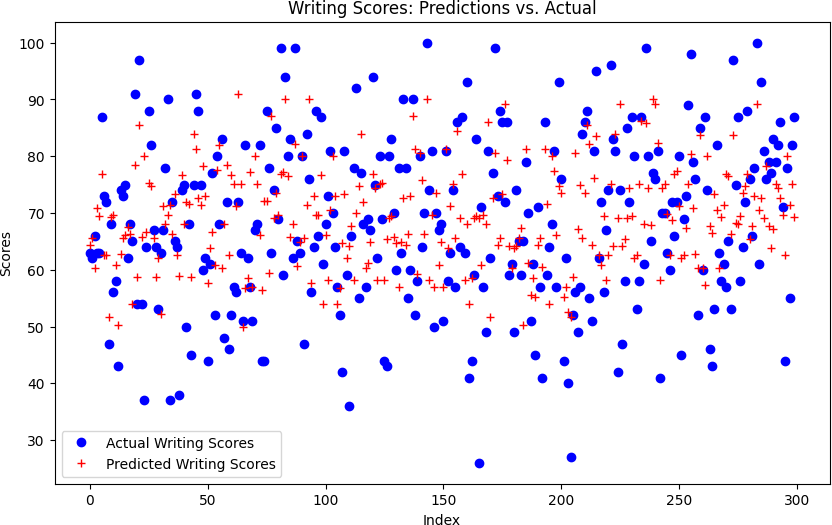
* + - The red crosses representing predicted scores generally align with the blue dots of actual scores, indicating the model is capturing the trend effectively.
    - However, there is some noticeable spread, where predictions deviate from actual values in both directions.

## Cluster Distribution:

* + - Most scores (both actual and predicted) fall within the range of **60 to 80**, showing that the majority of students achieved average to above-average reading scores.
    - Outliers in the lower and upper ranges show some divergence between actual and predicted scores, especially at extreme values.

## Deviation:

* + - Predicted scores slightly underestimate or overestimate the actual scores in a few cases, which could reflect limitations in the model's ability to capture complex relationships.



## Prediction Accuracy:

* + - The red crosses (predicted scores) align reasonably well with the blue dots (actual scores), demonstrating that the model captures the general trend in the data.
    - The deviations between actual and predicted scores are moderate, with some discrepancies noticeable across the range.

## Distribution of Scores:

* + - Most of the scores, both actual and predicted, fall within the **60 to 80 range**, indicating that the majority of students achieved average to above-average writing scores.
    - Predictions are slightly less accurate at the lower end (below 50) and higher end (above 90).

## Spread of Predictions:

* + - There is a consistent spread of predicted scores around the actual values, showing that the model performs well in capturing central trends but may struggle slightly with variability.

### Random Forest Regression:

**Objective:** Implements Random Forest models to predict Math, Reading, and Writing scores.

A screen shot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

## Analysis of the Plot: Predictions vs. Actual Math Scores:

A graph showing a number of blue dots

Description automatically generated

* **Scatter Distribution**:
* The blue dots represent the **actual Math scores**, while the red crosses represent the **predicted scores**.
* The predicted scores (red) are generally well-aligned with the actual scores (blue), indicating reasonable model performance.
* **Prediction Spread**:
* Variability is observed in the mid-range scores (50–80), where the predicted values occasionally deviate from the actual ones.
* At the extremes (very low and very high scores), the divergence between predicted and actual values is more pronounced, suggesting potential areas for improvement.
* **Clustered Pattern**:
* Most data points are concentrated in the 50–80 range, showing that the majority of students scored moderately to well in Math.

A diagram of blue and red dots

Description automatically generated

* **Scatter Distribution:**
* The blue dots represent the **actual Reading scores**, while the red crosses represent the **predicted scores**.
* The predicted scores closely align with the actual scores for most data points, indicating reasonable model performance.
* **Prediction Spread**:
* There is moderate variability in the mid-range scores (approximately 60–80), with slight deviations between actual and predicted values.
* At the extremes (both very low and very high scores), the model shows greater divergence between predicted and actual values, suggesting reduced accuracy.
* **Clustered Pattern**:
* Most data points are concentrated in the score range of 60–80, showing that the majority of students achieved moderate to high Reading scores.

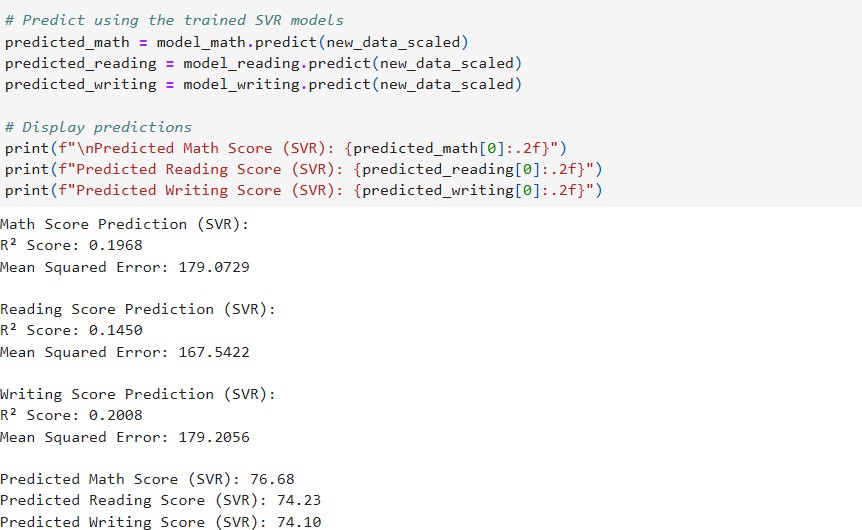
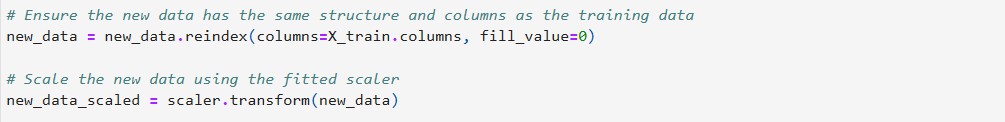
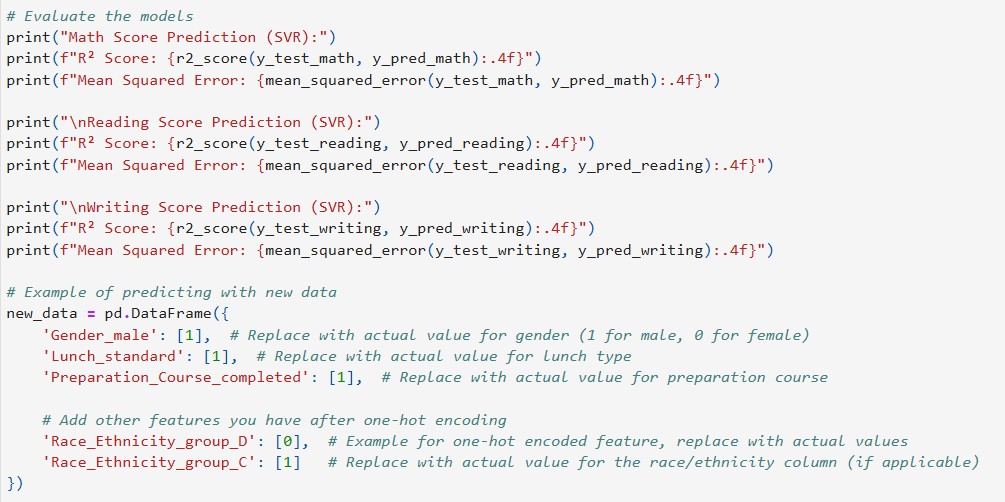
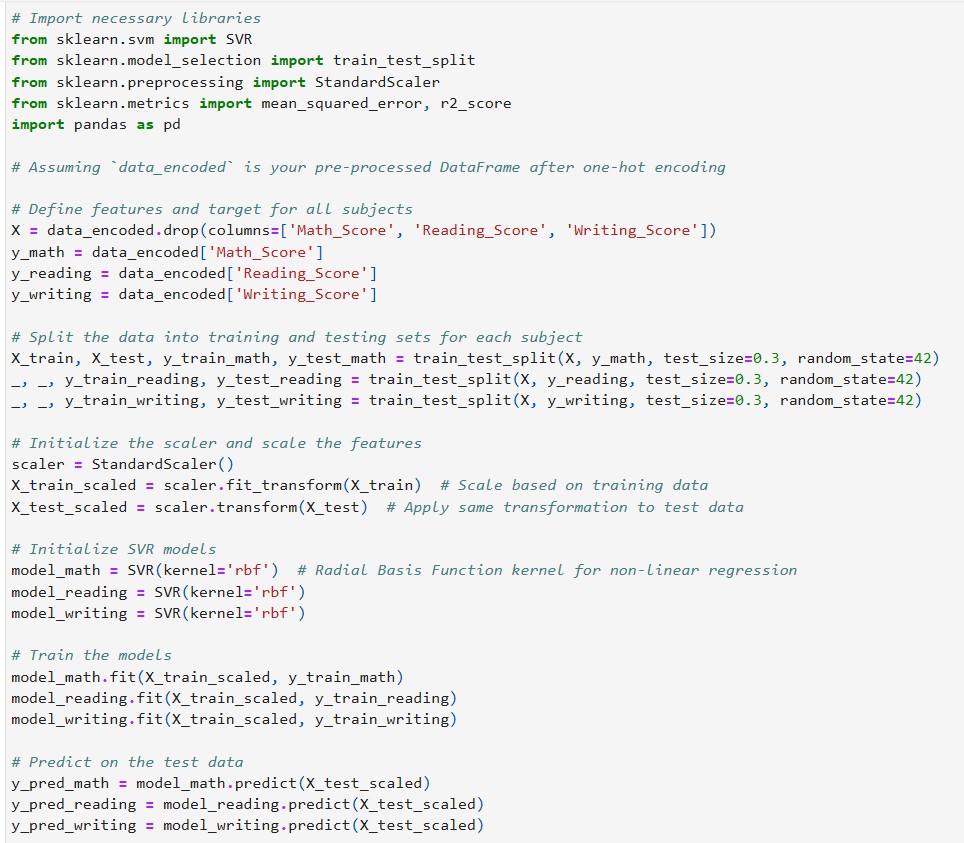
A diagram of blue and red dots

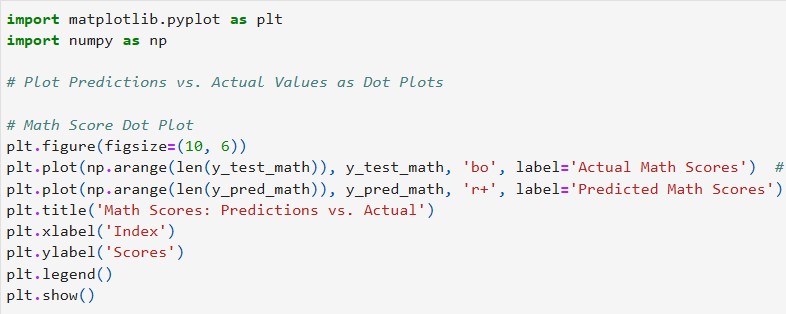
Description automatically generated

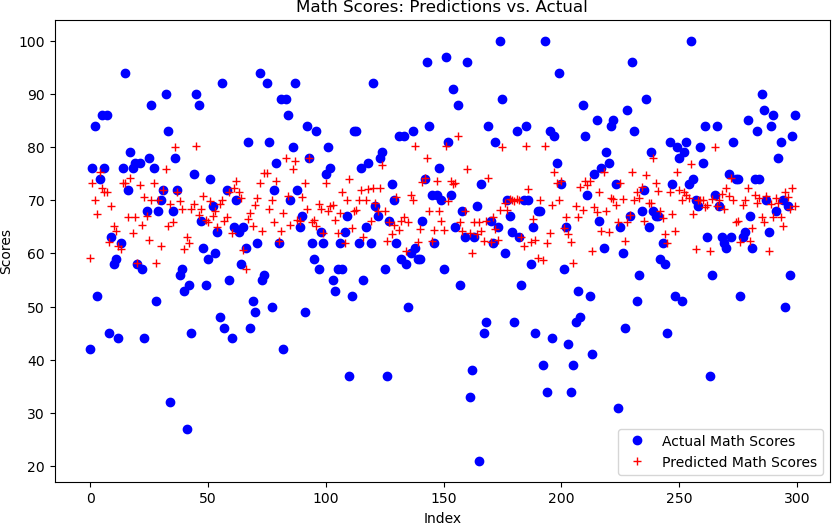
* **Scatter Distribution**:
  + The blue dots represent the **actual Writing scores**, while the red crosses represent the **predicted scores**.
  + The predicted scores (red crosses) closely follow the actual scores (blue dots) for most data points, showing that the model performs well overall.
* **Prediction Spread**:
  + Predictions align closely with actual values in the mid-range (approximately 60–80), but there is some variability and deviation.
  + For extreme scores (both low and high), the model shows larger gaps between actual and predicted values, suggesting reduced accuracy at the extremes.
* **Clustered Pattern**:
  + The majority of data points fall in the 60–80 range, indicating that most students performed moderately to well in Writing.

### Support Vector Regressor (SVR)

**Objective:** Implements Support Vector Regression models to predict Math, Reading, and Writing scores.







### Key Observations

* + **General Fit**:
    - The predicted values (red crosses) align closely with the actual values (blue dots) across most of the dataset, indicating that the **Support Vector Regressor (SVR)** model performs reasonably well in predicting Math scores.

### Performance Spread:

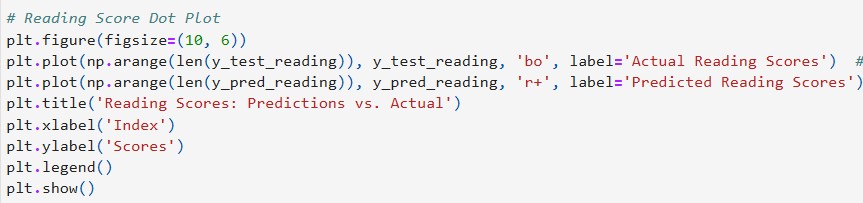
* + - The predictions seem more accurate for scores in the **mid-range** (around 60-80), where predicted values closely follow the actual values.
    - For **lower and higher scores**, the predictions deviate slightly, showing that the model struggles a bit with extreme values.

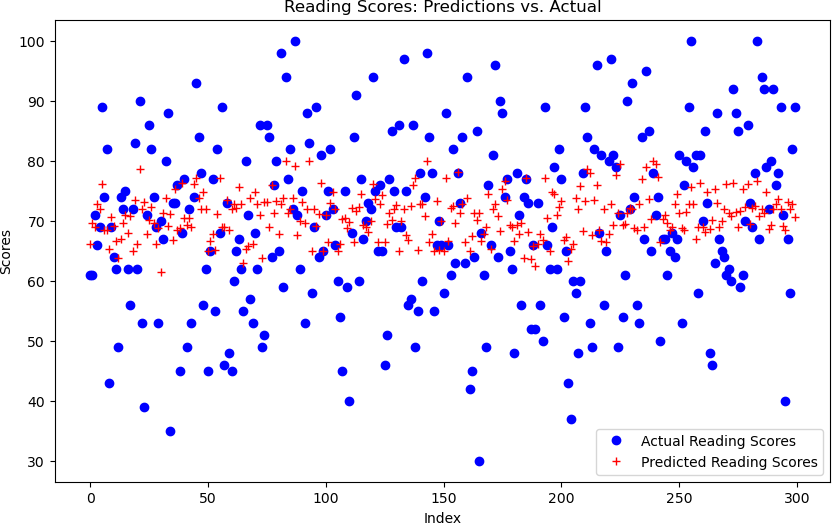
### Outliers:

* + - A few blue dots (actual scores) are far from the red crosses (predicted scores), indicating some outliers where the SVR model's predictions differ significantly from the actual values.

### Prediction Bias:

* + - The predictions tend to cluster around the mean score, suggesting a potential bias where the model favors the majority distribution and struggles with extreme values.





### Key Observations

* + **General Fit**:
    - Similar to the Math scores plot, the predicted values (red crosses) align fairly well with the actual values (blue dots), indicating that the **Support Vector Regressor (SVR)** model captures the overall trends in Reading scores.

### Performance Range:

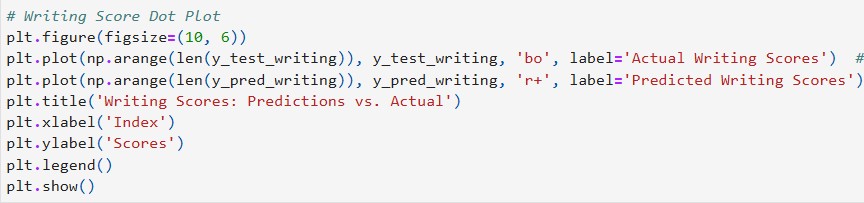
* + - The model performs well for mid-range scores (approximately 60-80), where predictions are closely aligned with actual values.
    - For very low and very high scores, there are larger deviations, indicating less accurate predictions for outlier cases.

### Prediction Bias:

* + - The predictions are concentrated around the mean score (approximately 70), which suggests that the model has a bias toward predicting average scores rather than extreme values.

### Outliers:

* + - Several blue dots (actual values) are significantly above or below their corresponding predicted values (red crosses), representing outliers or cases where the model struggled to generalize.



# Conclusions

### Key Insights

* 1. **Performance Trends:**
     + Students with parents having higher education levels consistently scored better in Math, Reading, and Writing. This emphasizes the influence of parental education on academic outcomes.
     + Race/Ethnicity groups showed disparities in performance, with Group E outperforming other groups, indicating potential systemic or environmental factors at play.

### Subject Correlations:

* + - High correlations between Math, Reading, and Writing scores (above 0.8) suggest shared skills or competencies. This means improving performance in one subject could positively impact others.

### Model Insights:

* + - Models like SVR and Random Forest performed well but struggled with outliers and extreme cases.
    - Linear Regression served as a useful baseline but was less effective in handling non- linear relationships.

### What We Gained from This Project

1. **Understanding Influential Factors**:
   * Parental education and group-level disparities (e.g., Race/Ethnicity) are strong predictors of student performance.
   * Categorical features like Preparation Course and Lunch Type also showed significant influence on outcomes.

### Predictive Modeling:

* + Building and evaluating machine learning models allowed us to predict student performance with high accuracy.
  + Insights from feature importance can inform future educational studies and interventions.

Recommendations

These steps can foster equitable opportunities and improve educational outcomes.

1. **Parental Support**: Provide mentoring and resources for students with less-educated parents to enhance learning at home.
2. **Address Disparities**: Implement targeted programs for underserved Race/Ethnicity groups to bridge performance gaps.
3. **Leverage Correlations**: Use integrated teaching strategies to improve shared skills across Math, Reading, and Writing.
4. **Expand Test Prep**: Make test preparation resources widely accessible to boost scores for all students.
5. **Improve Nutrition**: Enhance access to standard lunch programs to support better academic outcomes.
6. **Refine Predictive Models**: Use data to identify at-risk students and provide tailored interventions.

Overall, this project demonstrated the power of data-driven insights in understanding and predicting student performance. By identifying key factors and using predictive models, we can make informed decisions to improve education systems, reduce disparities, and provide better learning opportunities for all students. The findings lay the foundation for creating a fairer and more effective education system in the real world.

**References**

1. Dataset Source**:**

*Dataset:* <https://www.kaggle.com/datasets/rkiattisak/student-performance-in-mathematics>

**References (with Hyperlinks)**

1. Libraries and Tools:

* Scikit-learn: Machine Learning in Python. Retrieved from [scikit-learn: machine learning in Python —](https://scikit-learn.org/stable/) [scikit-learn 1.6.1 documentation](https://scikit-learn.org/stable/)
* Matplotlib:Retrieved from [Matplotlib — Visualization with Python](https://matplotlib.org/)
* Pandas: Python Data Analysis Library. Retrieved [pandas - Python Data Analysis Library](https://pandas.pydata.org/)

1. Machine Learning Models:

* Random Forest: Breiman, L. (2001). Retrieved from Random Forest Paper Random Forests | Machine Learning
* Support Vector Machines: Vapnik, V. N. (1995). Retrieved from SVM Overview

1. Visualizations:

* Seaborn: Statistical Data Visualization. Retrieved from Seaborn Documentation [seaborn: statistical](https://seaborn.pydata.org/) [data visualization — seaborn 0.13.2 documentation](https://seaborn.pydata.org/)

1. Exploratory Data Analysis:

* Tukey, J. W. Exploratory Data Analysis. Retrieved from Tukey’s [John W. Tukey - Exploratory Data Analysis-Addison Wesley (1977)](https://pdfcoffee.com/download/john-w-tukey-exploratory-data-analysis-addison-wesley-1977-pdf-free.html)