|  |
| --- |
| ESLSCA |
| Education Performance Analytics |
| Data Wrangling, Analysis, and Insights Report |

|  |
| --- |
| Abdelrahman Mansour (222300472), Hashem Zayed (222300400), Marwan Hany (222300457) & Paula Vargas (222300406) |

Contents

[Introduction 2](#_Toc185479267)

[Python Data Wrangling and Analysis 3](#_Toc185479268)

[Data Collection 3](#_Toc185479269)

[Data Cleaning 3](#_Toc185479270)

[Feature Engineering 4](#_Toc185479271)

[Exploratory Data Analysis (EDA) 4](#_Toc185479272)

[SQL Query Insights and EER Diagram Analysis 5](#_Toc185479273)

[Relational Database Structure 5](#_Toc185479274)

[SQL Query Insights 5](#_Toc185479275)

[Subject-Specific Performance 5](#_Toc185479276)

[Tableau Visualizations 6](#_Toc185479277)

[Top Performers and Underperformers 7](#_Toc185479278)

[Conclusion 7](#_Toc185479279)

[References 7](#_Toc185479280)

# Introduction

This report analyzes student performance to identify the factors that influence their grades and scores across four key subjects: math, reading, writing, and science. The primary focus is to examine the impact of ethnicity, test preparation, and parental education levels on student outcomes.

The analysis was conducted using a combination of tools and techniques to ensure thorough and accurate exploration of the data:

* Python: For data wrangling, cleaning, feature engineering, and exploratory analysis.
* SQL: For extracting insights and managing relational database structures.
* Tableau: For creating clear, interactive visualizations that highlight trends and patterns.
* Entity-Relationship Diagrams (ERDs): To document and understand the structure and relationships within the database.

This multi-step approach allows for:

* Identifying trends and patterns in student performance.
* Uncovering the effects of demographic and educational factors.
* Providing actionable insights to help improve academic outcomes.

By combining these tools and methods, the report offers a comprehensive view of the dataset and its key influences on student performance.

# Python Data Wrangling and Analysis

## Data Collection

The following datasets were utilized:

* **Original Dataset:** Student\_Performance.csv (10,000 rows, 12 columns).
* **SQL-Derived Datasets:**
  + **performance.csv:** Detailed subject scores (10,000 rows).
  + **student.csv:** Demographic and test preparation details (10,000 rows).
  + **subject\_grade.csv:** Subject-wise and overall grades (10,000 rows).
  + **SQL insights files:** Grade distributions, average scores, and demographic performance.
* **Updated Dataset:** clean\_student\_performance.csv after cleaning and enhancement.

## Data Cleaning

Python libraries such as pandas and scikit-learn were used for data cleaning and transformation:

* **Numerical Data:**
  + Missing values were addressed using median imputation to minimize skew caused by outliers.
  + Out-of-range values (e.g., scores above 100) were capped within acceptable ranges.
* **Categorical Data:**

Missing values in categorical fields (e.g., test\_preparation\_course) were predicted using a Random Forest Classifier.

* **Outlier Handling:**

Boxplots and z-scores were used to detect and cap extreme outliers, particularly in math and science scores.

* **Dropped Columns:**

The lunch column was removed after revealing no significant correlation with student performance.

## Feature Engineering

Additional metrics and features were created to enhance the dataset and support analysis:

* **Subject-Wise Grades:**  
  Grades were assigned for math, reading, writing, and science based on score ranges:
  + A: 90-100
  + B: 80-89
  + C: 70-79
  + D: 60-69
  + F: Below 60
* **New Columns:**
  + math\_grade
  + reading\_grade
  + writing\_grade
  + science\_grade
  + overall\_grade
* **Overall Grades:**  
  A weighted average of all subject scores was calculated, emphasizing reading and writing for balanced assessment.

## Exploratory Data Analysis (EDA)

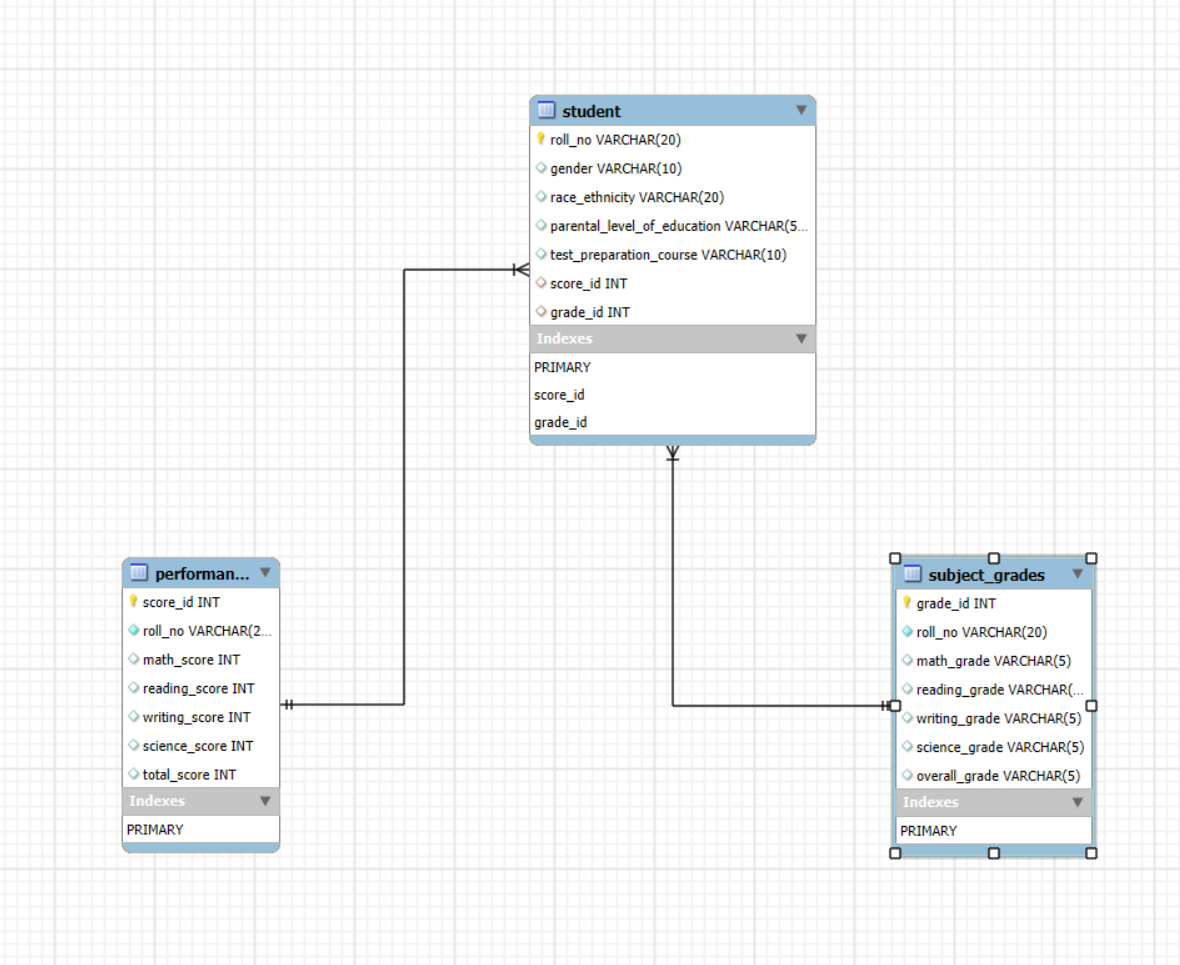
Using Python libraries (seaborn, matplotlib), the following patterns and trends were identified:

* **Data Distributions:**
  + Math scores exhibited a right-skewed distribution, normalized using PowerTransformer.
  + Reading and writing scores were more consistent and normally distributed.
* **Correlations:**
  + Strong positive correlation (r=0.78r = 0.78r=0.78) between math and science scores.
  + Minimal correlation (r<0.2r < 0.2r<0.2) between lunch and scores, supporting its exclusion.
* **Performance Trends:**

Female students excelled in reading and writing, while male students performed better in math.

# SQL Query Insights and EER Diagram Analysis

## Relational Database Structure



The SQL database was structured with normalized tables and clear relationships:

* **student ↔ performance:**
  + **Type:** One-to-Many (1: N).
  + **Connection:** score\_id (FK in student) links to score\_id in performance.
* **student ↔ subject\_grades:**
  + **Type:** One-to-Many (1: N).
  + **Connection:** grade\_id (FK in student) links to grade\_id in subject\_grades.

## SQL Query Insights

* **Ethnicity-Based Performance:**
  + Group C had the largest representation (2,938 students) and strongest performance.
  + Group A had the smallest representation (1,065 students) and struggled across subjects.
* **Test Preparation Impact:**
  + 38.78% of students completed the test preparation course.
  + Students who completed the course scored, on average, 12% higher.
* **Gender Distribution:**
  + Balanced representation of male (4,996) and female (5,004) students.
  + Female students excelled in reading and writing, while males performed slightly better in math.
* **Parental Education Influence:**

Students whose parents completed "some high school" had the highest average total score (266.98).

## Subject-Specific Performance

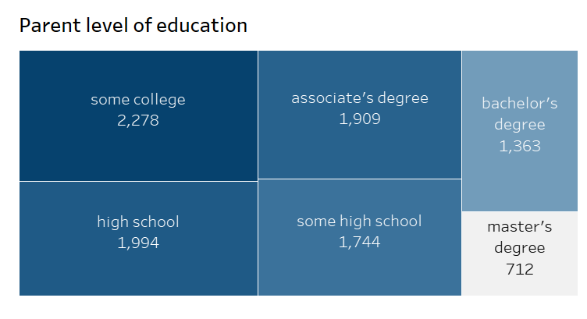
* **Math Scores:**
  + **Average score:** 57.18 (lowest among all subjects).
  + Variability was highest across demographics.
* **Reading and Writing Scores:**
  + **Reading average:** 70.13, Writing average: 71.42.
  + Consistent performance across most groups.
* **Science Scores:**

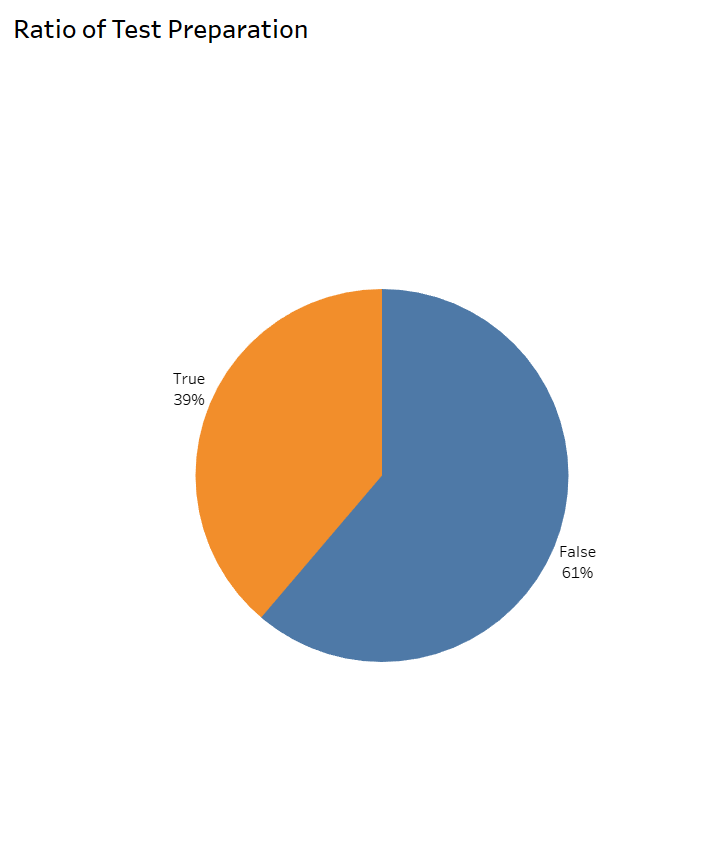
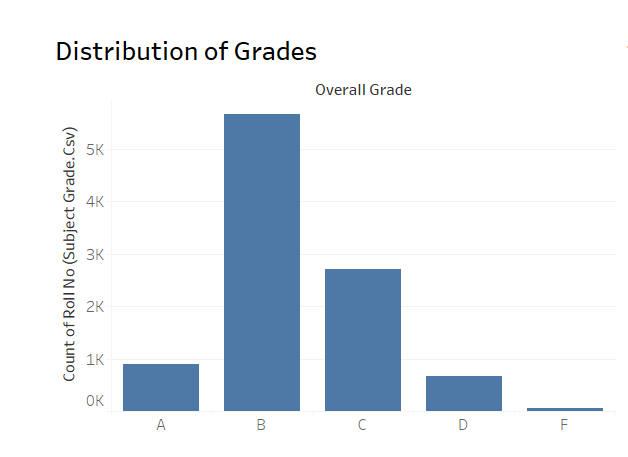
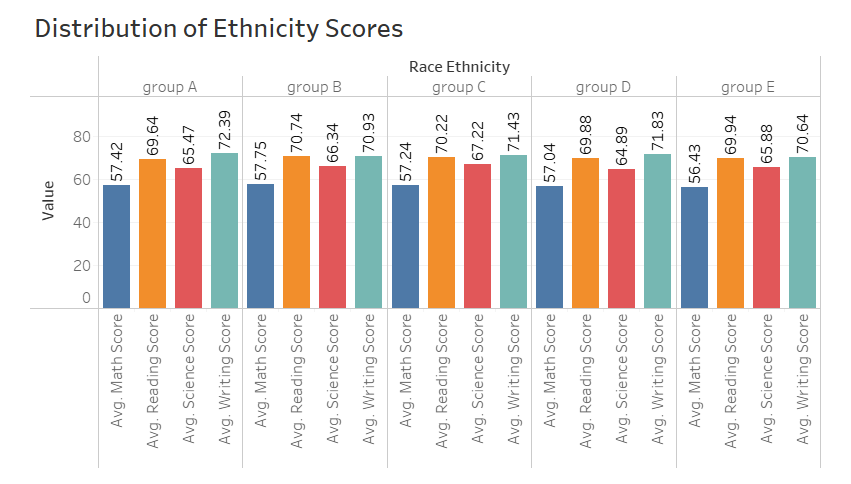
Average score: 66.07, showing strong correlations with math performance.

# Tableau Visualizations

Tableau visualizations were created to enhance understanding and validation of the analysis. The following visualizations were selected for their clarity and ability to highlight critical trends:

1. **A blue and orange circle with text

   Description automatically generatedGender Ratio**
   1. **Visualization:** Pie chart displaying a 50:50 ratio between male and female students.
   2. **Purpose:** Confirms balanced representation, ensuring unbiased gender-based performance comparisons.
2. **Parental Education Levels**
   1. **Visualization:** Treemap showing the distribution of parental education levels:
      1. **Most Common:** "Some college" (2,278 students).
      2. **Least Common:** "master’s degree" (712 students).
   2. **Purpose:** Highlights the influence of parental education on student outcomes, with students from "some high school" backgrounds achieving the highest average scores.
3. **A graph of race ethnicity

   Description automatically generatedEthnicity and Performance**
   1. **Visualization:** Bar charts showing ethnic group distribution and total scores:
      1. Group C leads in representation and performance (781,733 total scores).
      2. Group A struggles with the lowest representation and scores.
   2. **Purpose:** Identifies disparities between ethnic groups, supporting targeted interventions.
4. **Test Preparation Impact**
   1. **Visualization:** Bar chart comparing scores of students who completed test preparation versus those who did not, alongside a pie chart indicating that 39% of students completed the course.
   2. **Purpose:** Demonstrates the significant impact of test preparation on performance, providing evidence to expand these programs.
5. **Grade Distribution**
   1. **Visualization:** Bar charts displaying the distribution of grades:
      1. Most students achieved B grades, while A and F grades were less frequent.
   2. **Purpose:** Visualizes overall performance trends and highlights areas of underperformance.
6. **Subject-Specific Performance**
   1. **Visualization:** Stacked bar charts illustrating subject-specific scores across demographic groups:
      1. Math consistently underperforms, while reading and writing excel.
   2. **Purpose:** Reinforces the need for targeted math improvement programs and celebrates consistent strengths in reading and writing.

# Top Performers and Underperformers

* **Top Students:**
  + **Highest total score:** 383, achieved by multiple students (e.g., std-8035, std-8779).
  + Female students dominated top scores in math and science.
* **Low-Scoring Students:**
  + Students scoring below 60 in key subjects contributed to lower grades overall.
  + **Example:** Std-01 scored 26 in science and received an overall grade of C.

# Conclusion

This report integrates Python-driven analysis, SQL insights, and Tableau visualizations to provide a comprehensive understanding of student performance. Key takeaways include:

1. **Math Performance:** Requires focused improvement programs.
2. **Test Preparation Impact:** Significant improvement in outcomes for students completing preparation courses.
3. **Demographic Trends:** Highlights gaps in performance based on gender, ethnicity, and parental education.
4. **Tableau Visualizations:** Enabled intuitive, interactive exploration of the dataset, confirming and enhancing Python and SQL-based insights.

# links

* **Dataset**

<https://www.kaggle.com/datasets/nadeemajeedch/students-performance-10000-clean-data-eda>

* **Tableau public**

<https://public.tableau.com/views/Studentperformance_17345820163170/Main?:language=en-US&publish=yes&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link>

* **GitHub**