Student Performance Data Analysis

Tracey-Lee Swartz

June 2025

1 Introduction and Background

This analysis explores a dataset of 1,000 students, originally sourced from Kaggle, which includes demographic information and exam scores in math, reading, and writing. The aim of this project is to understand how certain background variables relate to academic performance and whether meaningful patterns emerge when examining student data.

The dataset includes features such as gender, parental level of education, lunch type (standard vs. free/reduced), and whether the student completed a standardized test preparation course. By combining statistical analysis and visual exploration, we seek to answer questions like: Does test prep actually improve scores? Are some academic skills more closely linked than others? And does a student's family background subtly shape their performance?

These questions are not only relevant for educators and policymakers, but also provide valuable practice in real-world data cleaning, transformation, and trend analysis using Python's data science stack.

2 Data Cleaning Summary

Before diving into analysis, the dataset was carefully examined for common quality issues. Fortunately, it was in good shape, but a few key checks were still performed:

- **Missing Values:** There were no missing or null values in any column. This meant no imputation or row removal was necessary.
- **Data Types:** All data types were already appropriate. Categorical data (e.g., gender, lunch type) was stored as object type, and exam scores were integers.
- **Column Name Standardization:** Original column names had spaces and mixed casing. These were converted to lowercase and underscores for ease of access (e.g., `parental level of education` became `parental level of education`).
- **Duplicate Records:** A check for duplicate rows showed that all entries were unique.

As a result, the dataset required minimal preprocessing and was analysis-ready after these verifications.

3 Trend 1: Does Completing the Test Preparation Course Help?

To investigate whether test preparation improves student performance, we grouped the dataset by the `test_preparation_course` variable, which has two values: `completed` and `none`. For each group, we calculated the average scores in math, reading, and writing.

A grouped bar chart visually highlighted the differences between the two groups across all subjects. In each case, students who completed the course scored higher, with the most significant improvements in reading and writing.

Insight: Completing the test preparation course is associated with consistently higher scores in all three subjects. The gains, while moderate, suggest that such preparation programs may provide students with strategies or confidence that translate into better exam performance.

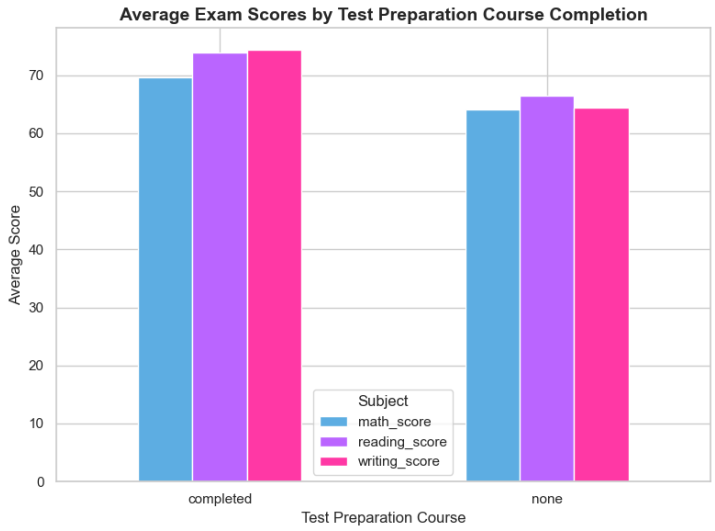


Figure 1: Average exam scores in math, reading, and writing for students who completed the test preparation course versus those who did not. Students who completed the course show consistently higher performance across all subjects.

4 Trend 2: Correlation Between Reading and Writing Scores

To explore the relationship between students' reading and writing abilities, we plotted each student's reading score against their writing score using a scatter plot. Each point represents a student, with their reading score on the x-axis and writing score on the y-axis.

The resulting plot showed a clear upward linear trend — as reading scores increased, so did writing scores. To quantify this relationship, we calculated the Pearson correlation coefficient, which came out to approximately 0.95.

This is a very strong positive correlation, indicating that students who perform well in reading also tend to perform well in writing. This is consistent with expectations, given that both skills involve comprehension, vocabulary, and language structure.

Insight: Reading and writing scores are closely linked. This suggests that interventions to improve one skill may positively impact the other, and that language proficiency tends to develop holistically rather than in isolation.

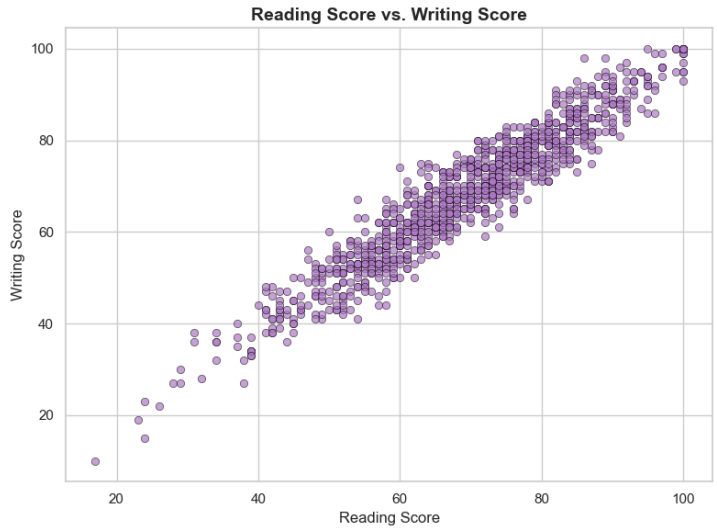


Figure 2: Scatter plot showing the relationship between reading and writing scores. The strong upward trend indicates a high positive correlation ($r \approx 0.95$), meaning students who score high in reading generally score high in writing as well.

5 Trend 3: Does Parental Education Affect Student Performance?

To examine the influence of parental education on student performance, we analyzed exam scores across six levels of `parental_level_of_education`, ranging from some high school to master's degree. We approached this analysis from two perspectives:

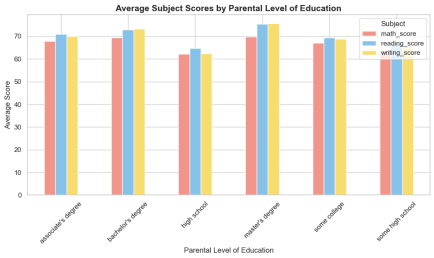
A. Subject-Specific Averages

We first calculated the average math, reading, and writing scores for each parental education level. The resulting bar chart revealed a general upward trend — students with more highly educated parents tended to score higher in all subjects. For example, students whose parents had a master's degree or a bachelor's degree generally outperformed those whose parents had only a high school diploma or less.

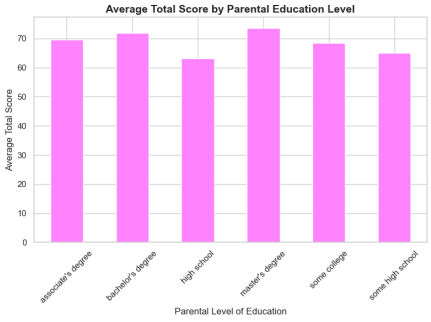
B. Overall Score Comparison

To simplify the view, we then created a new column, `average_score`, which is the mean of math, reading, and writing scores for each student. Grouping by parental education level, we plotted the average of these overall scores. This view helped to highlight the general trend without breaking it down by subject.

The differences in overall performance were modest but consistent. The highest average scores were observed in students whose parents had completed some college, a bachelor's degree, or a master's degree. The lowest averages occurred in students whose parents had only some high school education.



(a) Average scores in math, reading, and writing by parental level of education. Higher parental education levels are generally associated with better student performance across all subjects.



(b) Average total exam score (across math, reading, and writing) by parental education level. The chart suggests a modest but consistent increase in student performance as parental education level rises.

Bonus: Data Encryption Logic

In real-world scenarios, it's important to ensure that sensitive personal data is not exposed. To simulate privacy protection, we implemented a simple encryption approach using Python's `hashlib` to anonymize student identifiers.

For example, to encrypt a `student_id` column:

Listing 1: Encrypting a column using SHA-256

```
import hashlib

# Example: encrypting a 'student_id' column
def encrypt_value(val):
    return hashlib.sha256(str(val).encode()).hexdigest()

# Apply to a column
df['student_id_encrypted'] = df['student_id'].apply(encrypt_value)

# (Optional) drop original column if needed
df.drop('student_id', axis=1, inplace=True)
```

This ensures that any personally identifiable information is securely masked while still allowing the dataset to be analyzed anonymously.

6 Conclusion

This analysis explored how different background factors relate to student exam performance using a real-world dataset. Three main trends emerged:

- **Test Preparation Matters** – Students who completed a test preparation course consistently scored higher, especially in reading and writing, suggesting that targeted interventions can yield measurable academic gains.
- **Reading and Writing are Strongly Linked** – The extremely high correlation ($r \approx 0.95$) between reading and writing scores indicates that these skills develop in tandem. Supporting one may improve the other.
- **Parental Education Has a Modest Influence** – While not the strongest predictor, students with more highly educated parents tended to perform better across subjects. This supports the role of household academic support and socioeconomic factors in shaping outcomes.

Overall, the dataset highlights how academic performance is shaped not only by student effort but also by contextual factors like preparation and family background. These findings could inform targeted educational policies and underscore the value of comprehensive skill development.