Student Performance Factors Analysis

An ETL & Data Visualization Project

By Kanyinsola Fakehinde

Project Goals & Motivation

The Challenge: Education is complex; performance is influenced by more than just time in the classroom. We need a clear view of the *non-academic* factors driving student success.

Dataset Focus: Analyzing the **Student Performance Factors Dataset** (6,607 students, 20 attributes) to find patterns impacting the final Exam_Score.

Primary Project Objectives

- **ETL Pipeline:** Design and implement a complete, robust ETL process in Python (Pandas/NumPy).
- **Key Insight:** Identify the strongest positive and negative correlations with Exam_Score.

The ETL Pipeline: From Raw to Ready

The most critical step was converting messy, qualitative data into a format suitable for numerical analysis.

- **Extract & Load (E/L)**
- **Source:** Data was downloaded from Kaggle using the kagglehub API.
- **Load:** Final, cleaned data was saved to a persistent CSV file, marking the completion of the ETL process.

The ETL Pipeline: From Raw to Ready

The most critical step was converting messy, qualitative data into a format suitable for numerical Transformation (T)

- **Data Imputation:** Handled missing values (e.g., in Teacher_Quality) using Mode Imputation to preserve the distribution of the categorical data.
- **Categorical Encoding:** Converted all ordinal features (e.g., Parental Involvement, Family Income) into numerical scales (1, 2, 3) for correlation analysis.
- **Feature Engineering:** Created three powerful composite features to drive the analysis:

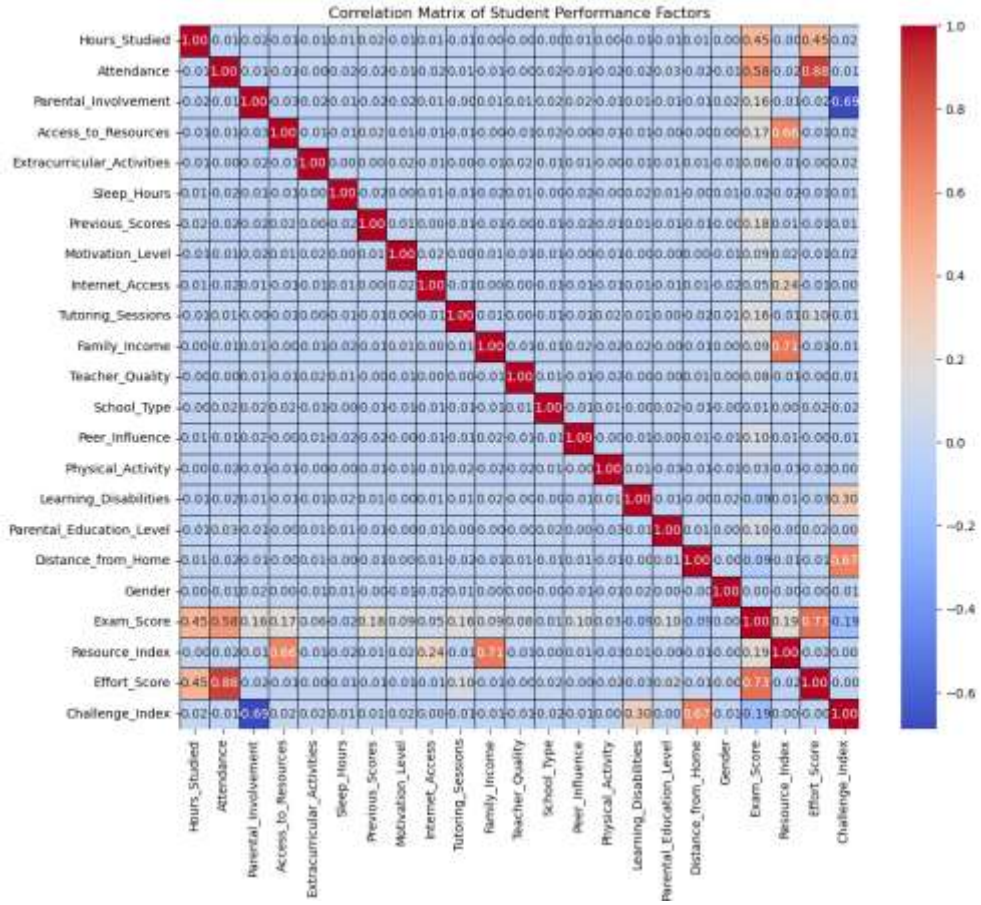
Engineered Score	Purpose
Effort Score	Quantifies total student effort.
Resource Index	Measures socioeconomic and access-based support.
Challenge Index	Measures cumulative student adversity (inverted to align with challenge).

Feature Engineering

The Three Engineered Factors

Feature	Purpose	Key Components
Effort Score	Quantifies total engagement.	Hours_Studied, Attendance, Tutoring_Sessions
Resource Index	Measures external support.	Family_Income, Internet_Access, Access_to_Resources
Challenge Index	Measures cumulative adversity.	Distance_from_Home, Learning_Disabilities, **Inverted** Parental_Involvement

Finding 1: The Overall Relationship (Heatmap)



The correlation heatmap provided the essential, high-level map of influence, quantifying how every feature relates to the final Exam_Score.

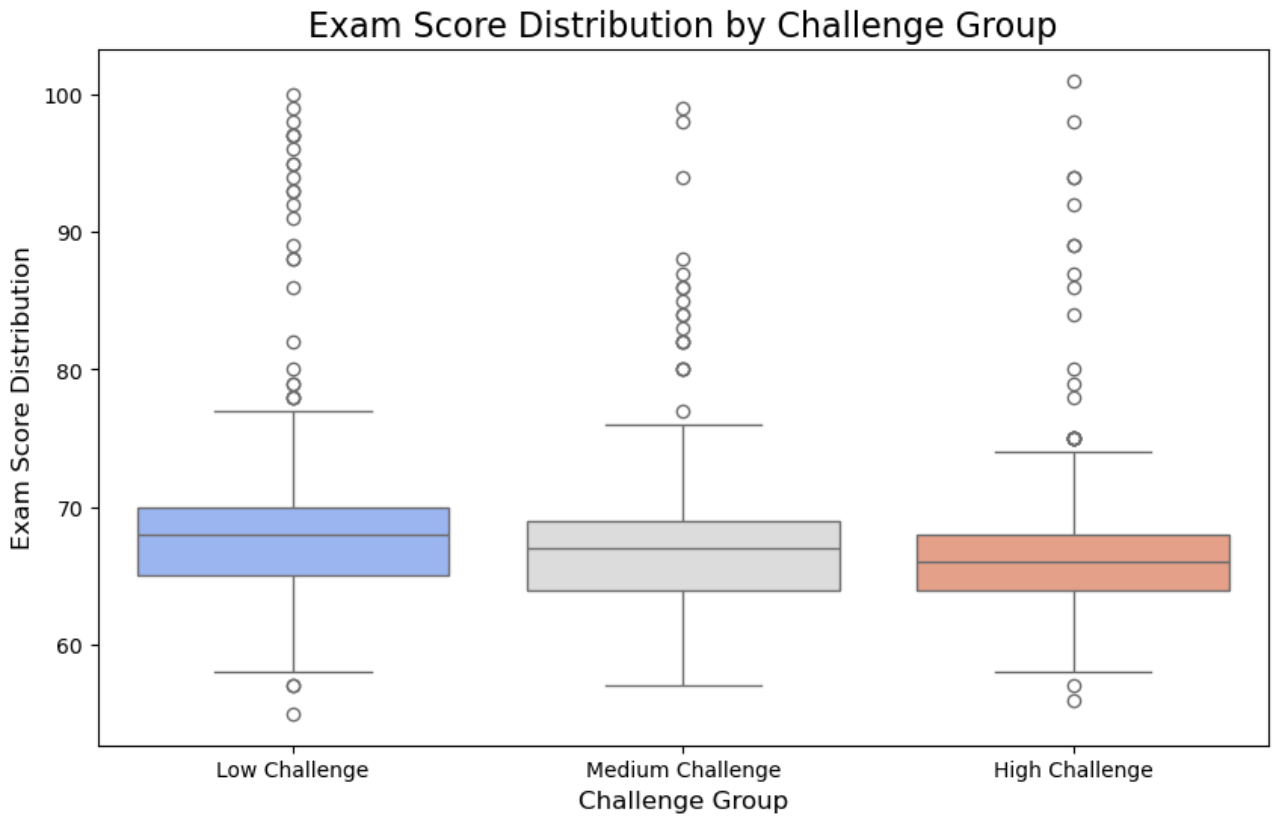
Strongest Positive Predictors:

- **Effort Score:** Showed the highest positive correlation, reinforcing that student effort (hours, attendance) directly correlates with achievement.
- **Previous Scores:** Also strongly positive, indicating historical performance is a powerful baseline.

Strongest Negative Predictor:

- **Challenge Index:** Showed the strongest negative correlation (approx. **-0.69**), confirming that adversity and environmental disadvantage are the biggest inhibitors of high scores.

Finding 2: The Impact of "Challenge" (Box Plot)



This box plot directly visualizes the effect of the **Challenge Index** by comparing the median score and score spread across three derived groups.

- **Median Score Drops:** The median (typical) score is highest for the **Low Challenge** group (approx. 70) and drops significantly for the Medium and High Challenge groups (approx. 66). This confirms that a lower level of challenge leads to a reliably higher *typical* score.

- **High Scores Are Possible:** All three groups show high outliers (scores up to 100). This suggests that while challenge inhibits the typical student, it **does not block** exceptional students who may have high motivation or previous scores.

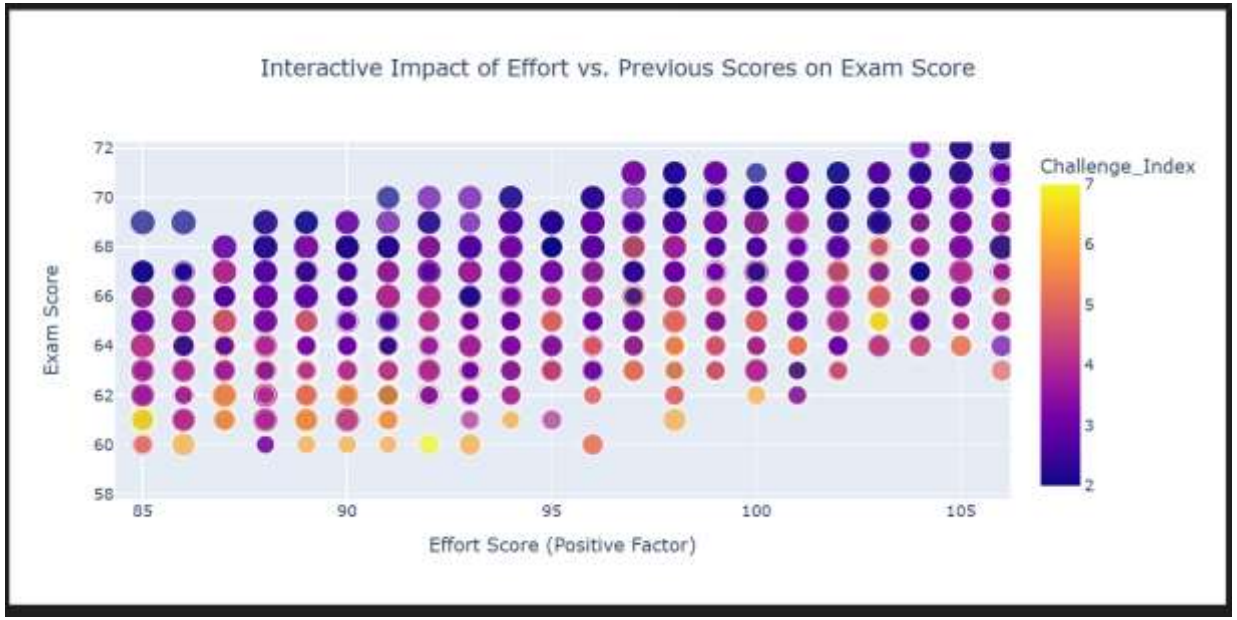
Finding 3: The Full Story (Interactive Plot)

This multi-dimensional interactive scatter plot reveals the complex relationship between the key factors engineered during the ETL phase.

- **The Two Bands:** The data clearly splits into two major bands: a lower band (scores 55-80) and an upper band (scores 80-100). This visually shows that **Previous Scores (size of the bubble)** act as a prerequisite for reaching the highest tier.

- **Challenge on the Margins:** The high Challenge Index (yellow/red bubbles) is concentrated at the lower end of both bands, visually confirming that high adversity makes high scores less likely, even with high effort.

- **Condition for Top Success:** The very highest scores (90+) are achieved almost exclusively by students with **high Effort Score** (far right) and **low Challenge Index** (dark purple color), demonstrating that low adversity is a necessary condition for peak performance.



The diagram has been zoomed in for clarity, so it does not display all records.

Summary of Key Insights

Key Insights and Data-Driven Conclusions

- **Effort is the Engine (Positive Drivers):**
  - The Effort_Score (Hours, Attendance, Tutoring) and Previous_Scores are the strongest positive predictors of academic achievement.

- **Challenge is the Inhibitor (Negative Drivers):**
  - The engineered Challenge_Index (Inverted Parental Involvement, Distance, Learning Disabilities) is the single strongest factor negatively correlated with the final score.

- **The Success Formula:**
  - While effort drives performance for the majority, breaking into the **top tier (scores 90+)** requires a low baseline of adversity. The highest scores are generally achieved by students with **High Previous Performance** and a **Low Challenge Index**.

- **Distribution:** The overall distribution of exam scores is slightly left-skewed, centered around the mean (approx. 66), suggesting a larger proportion of students fall slightly below the average.

Recommendations & Conclusion

Based on the data, here are three clear recommendations for improving student scores.

1. Focus on Students Who Need Help Now

- **Recommendation:** We can create support programs (like academic coaching or mentorship) for students who have **both** low Previous_Scores and a low Effort_Score.
- **Why?** This group has the biggest potential for improvement. Helping them increase their effort (attendance, study hours) is a clear first step to raising their scores.

2. Reduce Student Challenges

- **Recommendation:** We must find ways to help students in the "High Challenge" group by giving them extra support.
- **Why?** The Challenge_Index was the strongest *negative* factor. This means that providing resources (like tutoring) will likely have the biggest positive impact on lifting the average scores for these students.

3. Track Progress Over Time

- **Recommendation:** As a next step, we should track these same students over a longer period (e.g., 6 months).
- **Why?** This will let us prove that our recommendations actually *caused* their scores to improve, which helps us build an even better support model in the future.

Conclusion

Our project proved that a student's final score is not just about effort. The biggest factors are **Effort**, **Resources**, and the **Challenges** they face. The best way to improve overall performance is to help the students who face the most challenges.