ASU Ira A. Fulton Schools of Engineering
Arizona State University

Report on Student Performance Dataset

Team Number: 43

Authors:

- Mitali Kamal Bagadia | mbagadia@asu.edu | 1234311613
- Sharan Kumar Varma Chekuri | schekur3@asu.edu | 1234208003
- Aditya Rallapalli | arallapa@asu.edu | 1236173642
- Kruthika Suresh | ksures21@asu.edu | 1233969895

Date: 08/05/2025

Abstract

This study examines how parents' educational attainment affects their children's reading, writing, and math academic achievement. We investigate how family background and socioeconomic factors influence academic outcomes using a dataset of 1,000 high school students, descriptive analysis, hypothesis testing, and traditional and regularized linear regression modeling. According to our research, children whose parents have college degrees routinely perform better than their peers, with the most potent effects observed in reading and writing. Other factors like the type of lunch and the completion of test-prep courses also play a significant role. When trained solely on sociodemographic variables, regularized models such as Lasso and Elastic Net confirmed the predictive value of these variables and outperformed simple models. The full regression model explained up to 94.8% of the variance in writing scores, highlighting the combined impact of contextual and educational factors. These revelations emphasize the necessity of equity-driven academic interventions that assist students from underprivileged backgrounds.

1) Problem Definition & Context

A) Background and Motivation:

Understanding student performance factors is a key concern in educational data science. A key recurring factor is the academic level of parents—it is often regarded as a factor that significantly influences student outcomes. Since more and more education decisions are now based on data, examining this connection can provide actionable insights for educators and policymakers to inform evidence-based decisions.

B) Dataset Description and Collection Method:

For this study, we worked with a dataset called StudentsPerformance.csv, which has 1,000 anonymized student records. It's a popular open dataset often used in education research and machine learning projects. The data comes from standardized test results and some demographic surveys, though we don't know which institution collected it. It includes a mix of categorical features like gender, race/ethnicity, lunch type, parental education, test prep course, and numerical math, reading, and writing scores—making it well-suited for exploratory and inferential analysis.

C) Features Overview and Variable Types:

- **a) Categorical Features:**
  - i) Gender
  - ii) Race/Ethnicity
  - iii) Parental Level of Education
  - iv) Lunch
  - v) Test Preparation Course
- **b) Numerical Features:**
  - i) Math Score
  - ii) Reading Score
  - iii) Writing Score

The score columns are all integer values between 0 and 100. Notably, there are no missing values in the dataset. As for the parental education feature—it's an ordinal category, so we'll handle it that way when we get to the modeling part.

D) Research Objective and Hypotheses:

In this study, we wanted to see if a student's academic performance in reading, writing, and math is linked to their parents' education. The idea is simple—students with more educated parents might do better in school.

a) Main Hypothesis (H_0 vs H_1):

H_0: There's no real difference in scores based on parental education level.

H_1: Students with more educated parents tend to score higher in math, reading, and writing.

b) Sub-Hypotheses:

H_1: Students whose parents have a master's degree will score higher than those whose parents only finished high school.

H_2: Parent's education might affect reading and writing scores more than math.

H_3: Students whose parents went to college (associate's, bachelor's, or master's) are likely to perform better than those whose parents didn't.

E) Importance of the Analysis:

Figuring out how much parental education affects student performance can help schools use their resources better and design support systems that yield measurable improvements in academic outcomes. If we know students from less-educated families need more help, schools can step in with extra tutoring or mentorship programs. This kind of analysis doesn't just help students—it also makes education fairer and assists with planning future policies and support systems.

F) Literature Review and Related Work:

Recent studies have examined many demographic and socio-economic factors to understand what drives academic success, especially in high school. One thing that keeps showing up across different studies is the parents' education level (PED)—it's often one of the strongest indicators of how well students do.

Brew et al. [1] conducted an exhaustive literature review and found six general categories of academic performance predictors. Among them, demographic variables such as gender, parents' education, family size, and family income were found to be significant. Parents' education and family income were most significantly and strongly related to student achievement. Specifically, mothers' higher education was linked with more support at home, academic help, and higher educational aspirations, positively impacting student learning.

Nawang et al. [2] agreed with these points in their study on predictive models within educational settings. While demographic attributes like PED were commonly employed in model designs, the research highlighted earlier academic performance, school-level institutional infrastructure, and school-level inputs generally exert greater predictive effects. However, demographic and socio-economic status are underlying conditions that affect the ability of students to engage with and gain from schooling.

Khanna et al. [3] proposed the psychological variable, suggesting that students' attitudes and feelings also combine with control variables like PED to influence academic achievement. Similarly, Davaatseren et al. [4] were adamant that low-income students suffer cumulative disadvantages in the form of lack of learning resources, additional domestic work, and emotional stresses, which combined act against academic success.

Nutrition has also arisen as a likely but otherwise frequently overlooked driver of academic accomplishment. There has been evidence from various research showing that malnourishment or an irregular food pattern, for instance, skipping breakfast, are associated with poorer concentration and thinking, particularly in difficult areas such as maths [3].

One general pattern across several studies is the significance of school and teacher quality. Trained teacher availability, effective classroom setup, and adequately equipped libraries and laboratories have significantly improved student performance. In some cases, quality teaching has successfully compensated for disadvantages brought about by low PED or poor home support [2, 4].

Parents' educational level is among this larger context's best academic achievement and motivation predictors. Students with more educated parents receive higher academic encouragement and the advantages of learning-friendly home environments. As Hidayatullah and Csíkos [6] describe, PEDs influence students' motivation directly and indirectly, for instance, by promoting students' feelings of belonging in school, thereby raising engagement and achievement.

Also, Wang [5] found that higher PED levels are related to higher student grades, particularly in mathematics. Educated parents are better positioned to assist with homework and work through educational systems. Lower-PED students are twice as disadvantaged; they are less academically performing and are less motivated due to having fewer parental inputs and educational resources in the home [6].

These results demonstrate that multiple factors influence student performance. Numerous factors affect it, including mental health, school quality, family background, socioeconomic status, and nutrition. Parental education consistently emerges as one of the most effective of these. However, we still don't fully understand many things despite the overwhelming evidence connecting it to student motivation and performance. More research is required to determine how PEDs impact students in various contexts.

2) Exploratory Data Analysis and Initial Findings

A) Data Cleaning and Standardization:

Before proceeding with the analysis, we cleaned and standardized the dataset.

a) We made the column names lowercase and replaced spaces with underscores for consistency.
b) Then, we cleaned up string fields by removing extra spaces and special characters. This helped keep things clean and made it easier to work with the data during modeling and plotting.

B) Summary Statistics:

We began with a descriptive statistical summary of the three score columns: math, reading, and writing.

a) Math: 66.09
b) Reading: 69.17
c) Writing: 68.05

All three scores fall between 0 and 100, and the distributions were symmetric overall. Still, there were slight shifts when we separated things into different demographic groups.

Here's a quick snapshot of the stats:

Metric	Math Score	Reading Score	Writing Score
Mean	66.09	69.17	68.05
Median	66	70	69
Standard Deviation	15.2	14.6	15.0

Table 1: This table shows the basic statistics, mean, median, and standard deviation—for student scores in math, reading, and writing.

C) Group Analysis:

a) Score Distributions by Parental Education

i) Students whose parents had a master's degree achieved the highest average scores across all three subjects.
ii) In contrast, students whose parents had completed only high school or didn't finish school had the lowest average scores.
iii) Overall, there was a clear pattern: the more educated the parents, the better the student scores.
iv) Violin Plots for each subject, stratified by parental education:

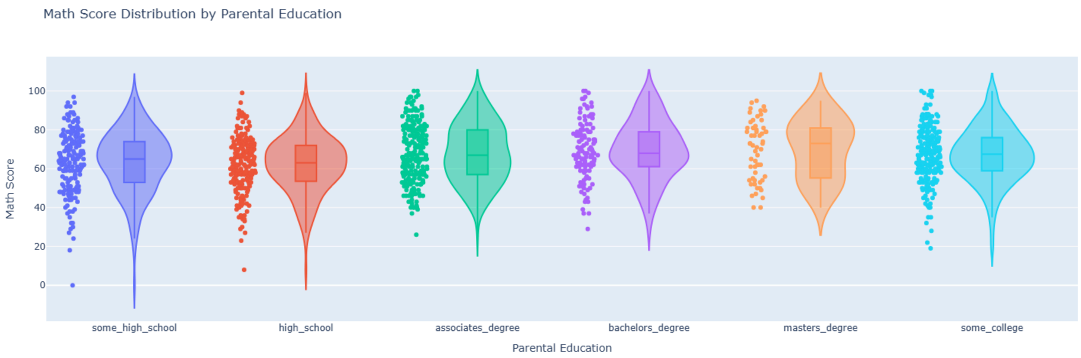


Figure 1: Violin plot displaying the distribution of math scores by parent's education level. Students whose parents have a master's degree showed higher median scores and narrower spread.

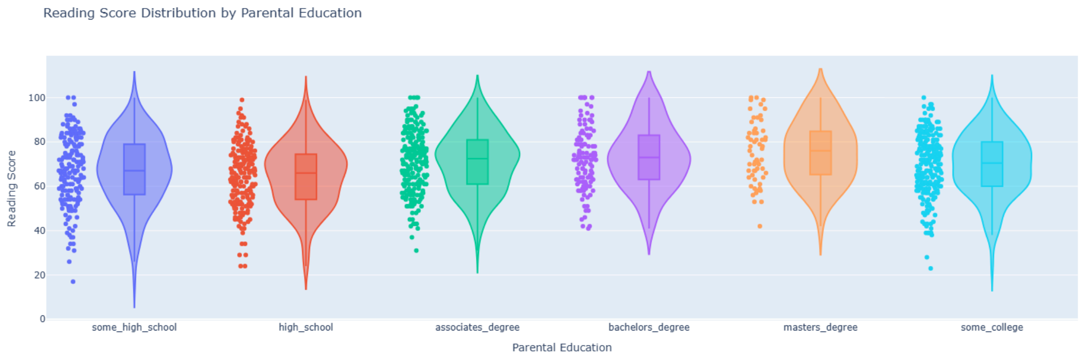


Figure 2: Violin plot comparing reading scores across parental education levels.

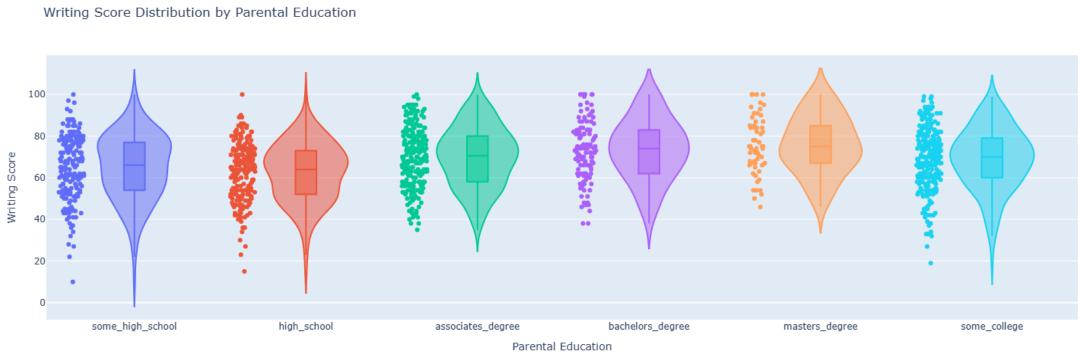


Figure 3: Violin plot for writing scores stratified by parental education, revealing wider performance gaps for language-based subjects.

1) The plots clearly showed that student scores vary based on parental education levels.

2) We also saw more outliers in the lower education groups, indicating higher variability in those categories.

b) **Demographic Breakdown:**

i) **Gender:** Females performed slightly better in reading and writing, while males showed slightly higher math scores.

ii) **Lunch Type:** Students with standard lunch significantly outperformed those on free/reduced lunch, indicating a socioeconomic gap.

iii) **Test Preparation Course:** Students who completed the course performed better across all subjects, most notably in writing.

iv) **Relevant Visuals:**

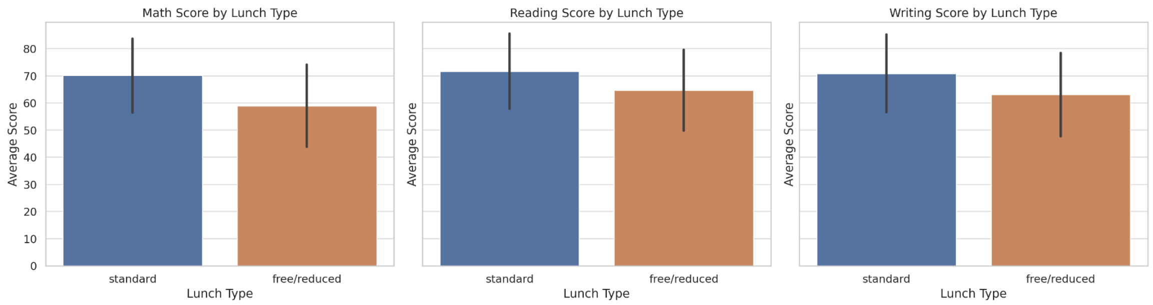
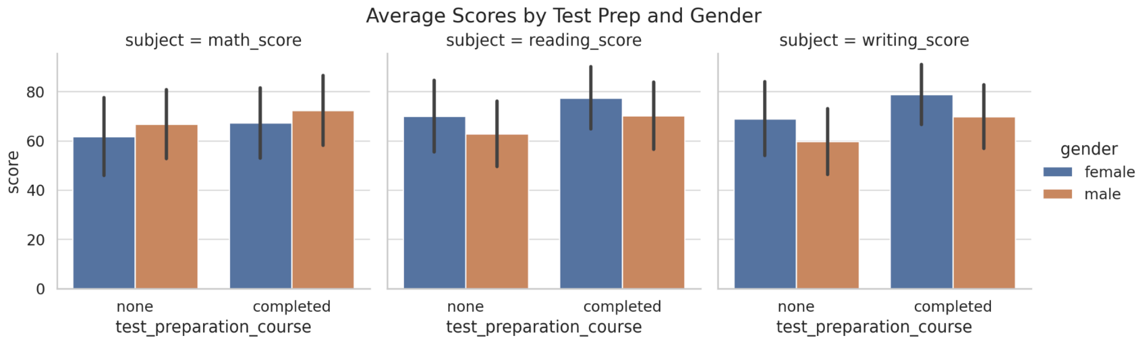


Figure 4: Grouped bar chart of math, reading, and writing scores based on lunch type. Students with standard lunches scored higher than those on free or reduced lunch.



lunchese 5: Facet grid bar plots of average scores by test preparation course and gender. Females scored higher in reading and writing; test preparation improved student outcomes.

D) **Findings:**

a) Students with parents holding higher education (especially master's or bachelor's degrees) consistently score better.

b) Those with "free/reduced lunch" or without test preparation courses perform lower on average.

c) Positive correlation between parental education and student scores—visible in all three subjects.

d) Disparity in distributions observed by demographic features.

E) **Issues Found During EDA:**

a) **Group Imbalance:** Yes, confirmed in the notebook using “groupby”. You can keep this as-is.

b) **Confounding Variables:** Yes, lunch and test prep were analyzed, affecting scores. Also valid.

c) **Skewness & Spread:** All three score distributions are approximately symmetric, with a slight left skew (i.e., more students scoring higher). There is no need for transformation or distributional correction before modeling.

3) Statistical Modeling and Inferential Analysis

A) Hypothesis Testing:

To evaluate our research hypotheses, we conducted a combination of One-Way ANOVA and independent samples t-tests, targeting specific hypotheses formulated in Section 1D.

Here’s how each hypothesis was tested and what we found:

1. **Main Hypothesis (H_0 vs H_1):**

H_0: There’s no real difference in student scores across different parental education levels.

H_1: Students with more educated parents tend to score better in math, reading, and writing.

We used a One-Way ANOVA test to examine whether average scores differed significantly across all parental education levels.

The results for all three subjects showed p-values < 0.001, indicating statistically significant differences in scores based on parental education level.

Subject	F-statistic	p-value
Math Score	6.522	0.00001
Reading Score	9.289	0.00000
Writing Score	14.442	0.00000

Table 2: ANOVA results testing mean score differences across parental education levels.

These results support our main Hypothesis.

2. **H1: Master's vs High School:**

We tested whether students whose parents had a master's degree scored significantly higher than those whose parents only completed high school using Welch's t-test (equal variance not assumed).

Subject	t-statistic	p-value	Mean
Math Score	3.412	0.00096	7.61
Reading Score	5.184	0.00000	10.67
Writing Score	6.449	0.00000	13.23

Table 3: Welch's t-test comparing scores between students with master's-educated parents and those with high school-educated parents.

3. **H2: Reading/Writing More Affected than Math:**

To examine this, we compared the effect sizes (Eta^2) from our ANOVA results:

Subject	Eta^2 (Effect Size)
Math Score	0.030
Reading Score	0.044
Writing Score	0.068

Table 4: Effect sizes (Eta^2) showing parental education's relative impact on math, reading, and writing scores.

Reading and writing scores showed **larger effect sizes**, suggesting that parental education had a **greater influence** on language-related subjects than math — which supports H2.

4. **H3: College vs Non-College:**

We grouped parents into:

College-educated: associate's, bachelor's, master's

Non-college: some high school, high school, some college

An independent samples t-test showed:

Subject	t-statistic	p-value	Cohen’s d
Math Score	4.309	0.00002	moderate
Reading Score	5.467	0.00000	moderate
Writing Score	6.507	0.00000	moderate

Table 5: Independent samples t-test results for college-educated vs non-college-educated parent groups.

These results support H3 — students with college-educated parents outperformed their peers.

B) Model Building:

To better understand the influence of parental education on academic performance, we built multiple linear regression models for each of the three target variables: math score, reading score, and writing score.

We considered the following predictor variables:

- parental_level_of_education (main variable of interest)
- gender
- lunch
- test_preparation_course
- race/ethnicity (included for completeness)

All categorical variables were encoded using dummy variables with appropriate reference categories, and the models were implemented using the “**statsmodels**” library in Python.

Model Variants Built:

1. **Simple Model:**

   Only included parental_level_of_education as a predictor.

   Example: writing_score ~ C(parental_level_of_education)

2. **Full Model:**

   Additional covariates (reading_score, math_score, test_preparation_course, lunch, gender, race_ethnicity) are added to control for potential confounders.

   Example: writing_score ~ C(parental_level_of_education) + C(test_preparation_course) + C(lunch) + C(gender) + C(race_ethnicity)

We calculated Variance Inflation Factors (VIFs) for all one-hot encoded socio-demographic variables to evaluate potential multicollinearity among predictors in the full regression model. All VIF values were well below the threshold of 5, with the highest being 3.17, indicating no severe multicollinearity. This confirms the stability of coefficient estimates in the full model.

Predictor	**VIF**
race/ethnicity_group C	3.17
race/ethnicity_group D	2.94
race/ethnicity_group B	2.56
race/ethnicity_group E	2.25
parental level of education_some college	1.57
parental level of education_high school	1.53
parental level of education_some high school	1.51
parental level of education_bachelor's degree	1.35
parental level of education_master's degree	1.20
test preparation course_none	1.02
gender_male	1.01
lunch_standard	1.01

Table 6: Multicollinearity Assessment Using VIF for Full Model Predictors

3. **Lasso and Elastic Net Models:**

We implemented regularized linear models — Lasso and Elastic Net regressions- to account for potential multicollinearity and irrelevant features. These models apply penalties to prevent overfitting and identify the most critical predictors of average student performance.

- Lasso helps in feature selection by shrinking some coefficients to zero.

- Elastic Net balances the strengths of Lasso (L1 penalty) and Ridge (L2 penalty), often producing more stable and robust models. Both models were trained using 5-fold cross-validation with one-hot encoded categorical predictors.

Justification for Using Linear Regression

- The outcome variables (scores) are continuous and roughly symmetric, as confirmed during EDA.

- No major violations of assumptions (normality of residuals, linearity) were observed in diagnostic tests.

- Regression allows us to estimate and interpret each factor's contribution while controlling for others.

C) Model Evaluation and Diagnostics:

We used the Akaike Information Criterion (AIC), residual diagnostics, R2 values, and coefficient significance as key metrics to evaluate the quality of our regression models.

Coefficient Significance:

Across all three models, we observed that the coefficients for parental_level_of_education were significant ($p < 0.05$) in both simple and full models. Other variables, such as test preparation course and lunch type, also showed statistically significant relationships with student performance.

Model Fit Comparison:

We compared two models predicting writing scores:

Model Type	**R^2 Value**	**AIC**
Simple Model	0.068	8220.8
Full Model	0.948	5353.7

Table 7: The full model, which includes demographic and academic predictors, significantly improves explanatory power ($R^2 = 0.948$) and model fit (AIC = 5353.7) compared to the simple model.

Model Type	R^2 Score	Mean Square Error
Lasso Regression	0.156	181.02
Elastic Net Regression	0.160	180.15

Table 8: Model Performance Comparison for Lasso and Elastic Net Regression Using Socio-Demographic Predictors

- The **simple model**, which included only parental education, explained ~6.8% of the variance.
- The **full model** added math and reading scores as predictors and demographic factors, explaining ~94.8% of the variance and drastically reducing the AIC.
- Including academic performance in related subjects drastically improves predictive accuracy.

Elastic Net offers the best performance among models using only socio-demographic predictors. **Lasso** performs similarly, confirming the robustness of key predictors such as parental education, lunch type, and test preparation course.

Residual Diagnostics:

To verify the assumptions of our linear regression, we conducted residual diagnostics.

- A **Shapiro-Wilk test** returned a p-value above 0.05, suggesting that residuals were approximately normally distributed.
- A **Breusch-Pagan test** confirmed no significant heteroskedasticity.
- The visual inspection of the **residuals vs fitted plot** revealed no obvious patterns, indicating linearity and homoscedasticity.
- The **Q–Q plot** showed that residuals followed a near-normal distribution.
- **Cook's Distance** was also computed to identify potential influential outliers.

These diagnostics confirm that our linear regression models satisfy the necessary assumptions for inference.

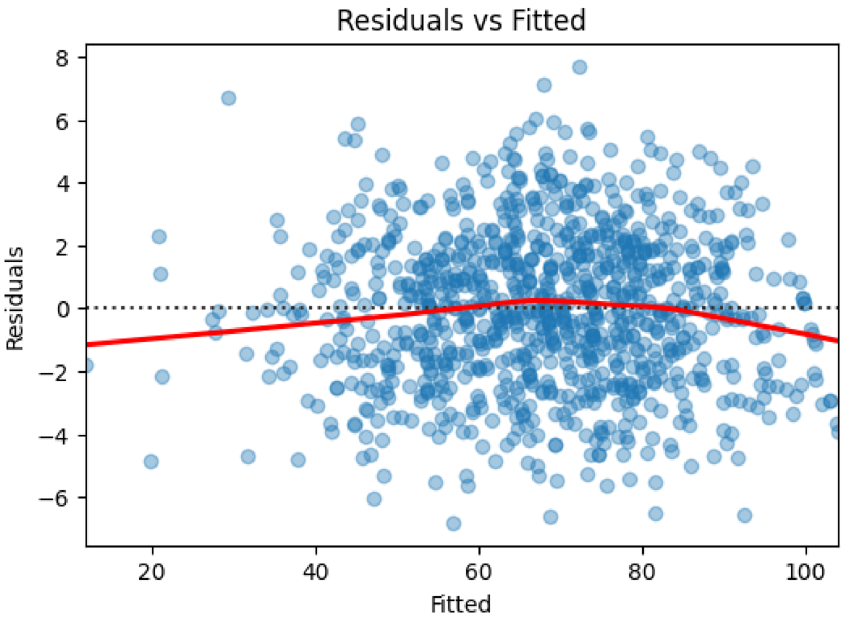


Figure 6: The scatterplot shows that the residuals are centered around zero with no clear pattern, which means the assumptions of linearity and constant variance (homoscedasticity) are holding up.

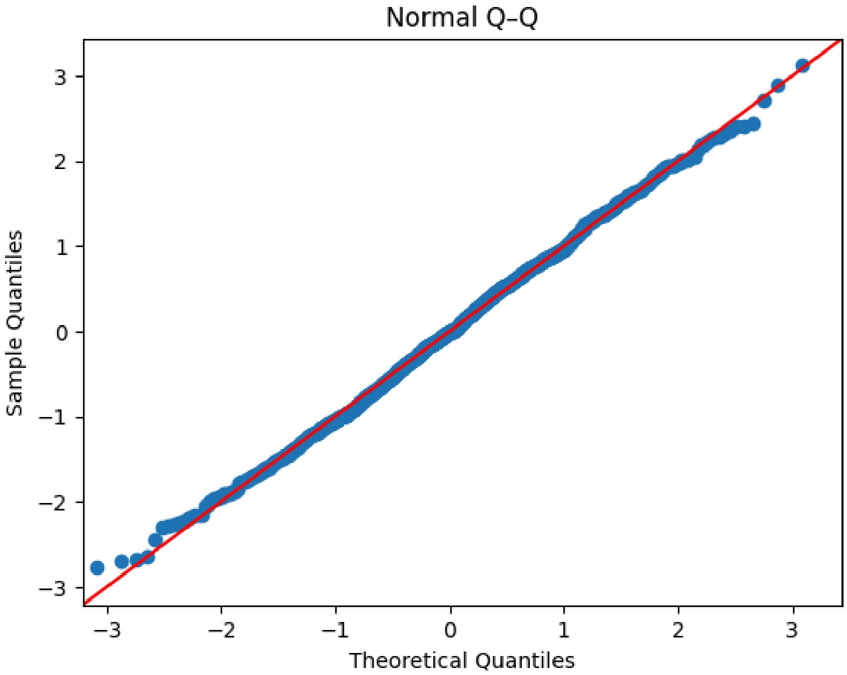


Figure 7: Q–Q plot of regression residuals. Points align closely with the 45° reference line, supporting the normality assumption.

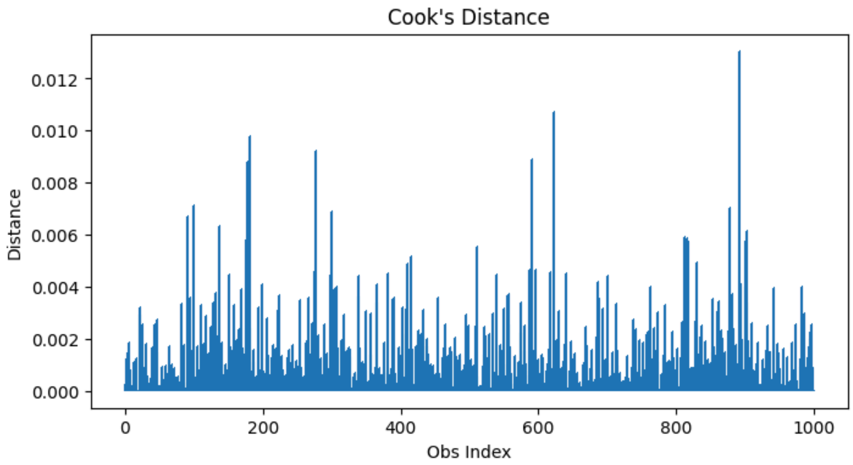


Figure 8: Cook's Distance values across observations. No extreme outliers were observed.

Assumptions and Diagnostics:

- **Normality of residuals:** Checked using the Shapiro-Wilk test; no substantial deviation from normality was observed.

- **Linearity and homoscedasticity:** Residual plots did not show major violations.

- **Multicollinearity:** We checked the VIF scores for all the one-hot encoded predictors, and since all of them were well below 5, there's no multicollinearity issue in the full regression.

  That means our linear regression setup is solid, and the results we're getting can be trusted.

D) Contributions to Solving the Problem:

This study explored whether a student's math, reading, and writing performance is connected to how educated their parents are. And the results were pretty straightforward—parental education plays a meaningful role, both statistically and in real life.

From the exploratory phase, we observed that students whose parents have higher academic qualifications consistently scored better in all three subjects. These patterns were visible in average scores, score distributions, and subgroup comparisons (e.g., by lunch type and test preparation status).

The hypothesis testing results provided strong statistical backing for these patterns. Using ANOVA and t-tests, we confirmed that performance differences across education levels are significant. Additionally, effect sizes showed that parental education more substantially influences reading and writing than math, aligning with our sub-hypotheses.

The regression models added deeper insight. While the simple model confirmed that parental education alone can predict academic performance to some extent, the whole model highlighted the combined effects of other demographic and educational variables. It showed that test preparation, lunch type, and reading/math skills further explain variance in performance — but crucially, parental education remained a significant predictor even after accounting for these factors.

We also used Lasso and Elastic Net regularized regression models to ensure the results were solid and avoid problems like overfitting or multicollinearity. Even when we only used socio-demographic features, these models still picked out parental education, lunch type, and test prep status as key predictors of student performance.

Together, these results contribute toward:

- Validating the educational impact of family background.
- Highlighting the importance of equity-based interventions, especially for students from less-educated households.
- This kind of analysis can help shape better academic support—like ensuring struggling students get easier access to test prep or proper meals through lunch programs.

By combining descriptive stats, inference tests, and different regression models (including the regularized ones), this study clearly shows how parental education connects to student performance. It also shares some practical insights that can help educators and policymakers work toward closing the performance gap.

4) Key Insights and Conclusion

This study explores whether a student's academic performance is influenced by their parent's level of education — and to what extent other demographic factors might play a role. Through a combination of descriptive statistics, inferential testing, and regression modeling, we were able to arrive at several key insights:

Key Insights

- Parental education level matters — Students whose parents hold college degrees, particularly master's degrees, tend to score significantly higher in math, reading, and writing.
- Reading and writing are more sensitive to parental education than math scores regarding mean differences and effect sizes.
- Lunch type and test preparation emerged as significant secondary predictors, reinforcing the impact of socioeconomic factors on student achievement.
- Parental education explained only a small part of the variation ($R^2 = 0.068$). The prediction improved slightly when we used Lasso and Elastic Net with just socio-demographic features ($R^2 = 0.16$). But the big jump happened when we added academic scores into the full

regression model—it went up to $R^2 = 0.948$. That shows how much better the model gets when we look at everything together.

Conclusion

The results show that family background—especially how educated the parents are—plays a significant role in students' performance. But that's not the whole picture. Things like socioeconomic status and access to resources, like test prep, also make a big difference.

Even when we only used demographic features, regularized models picked out test prep, lunch type, and parental education as significant predictors. That just shows how these factors matter on their own, too.

Taken together, the results point toward a need for holistic academic support systems that account not only for what happens inside the classroom but also for students' home and social environments. Targeted interventions for students from less-educated or low-income households could help level the playing field — especially in reading and writing, where the performance gaps are more pronounced.

This study adds to a growing body of evidence advocating equity-focused education policy and data-informed academic support.

5) References

[1] Brew, E. A., Nketiah, B., & Koranteng, R. (2021). *A literature review of academic performance: An insight into factors and their influences on academic outcomes of students at senior high schools*. Open Access Library Journal, 8, 1–14. https://doi.org/10.4236/oalib.1107423

[2] Nawang, H., Makhtar, M., & Hamza, W. M. (2021). *A systematic literature review on student performance predictions*. International Journal of Advanced Technology and Engineering Exploration, 8(84), 1441–1453. https://doi.org/10.19101/IJATEE.2021.874521

[3] Khanna, L., Singh, S. N., & Alam, M. (2018). *Multidimensional analysis of psychological factors affecting students' academic performance*. arXiv. https://arxiv.org/abs/1806.03242

[4] Davaatseren, A., Myagmar, M., & Dulamsuren, N. (2024). *Factors affecting students' academic performance: In the case of the accounting study of National University of Mongolia*. In *Proceedings of the 4th International Conference on Education and Social Science Research* (pp. 39-44). Atlantis Press. https://doi.org/10.2991/978-94-6463-382-5_6

[5] Wang, A. (2023). *Parent's education and child's school performance* [Kaggle notebook]. https://www.kaggle.com/code/adastroabyssosque/parent-s-education-and-child-s-school-performance

[6] *Hidayatullah, A., & Csíkos, C. (2024). The role of students' beliefs, parents' educational level, and the mediating role of attitude and motivation in students' mathematics achievement. Asia-Pacific Education Researcher, 33, 253–262.* https://doi.org/10.1007/s40299-023-00724-2

[7] Mitali, K. B., Kruthika, S., Aditya, R., & Sharan, K. V. C. (2025). *Analysis of student performance* [Source code]. GitHub. https://github.com/student-performance-dse501/analysis-on-student-performance