Student Performance Analysis

Names: Elliott Hood, Jonathan Lopez

Introduction:

The datasets we chose for this project provide information about student demographics, economic factors, and academic performance. It includes 1,000 students, and each row represents one student. For each student we know their gender, ethnic group, and whether or not they completed a test preparation course. We also have details about their parents' education level, and what type of lunch plan they use.

The dataset records students' scores separately as reading, writing, and math scores. However, these were then used to create total scores and average scores, meaning there were variable dependencies we had to account for when building our models.

Question 1: Which racial/ethnic group has the highest total score on average? Are there statistically significant differences in performance between racial/ethnic groups, and what might these differences suggest about potential educational disparities?

Methods

We performed a linear regression analysis to identify the racial/ethnic group with the highest average total scores and investigate any discrepancies. Group A was used as the baseline when we first encoded the race/ethnicity data into dummy variables. This enabled us to evaluate the degree to which the average total score of each group deviated from that of Group A. To make sure our model is dependable, we divided the data into training and testing sets (80% training, 20% testing). The R2 score, which gauges the impact of race and ethnicity in overall scores, was used to assess the model's effectiveness. We used a boxplot to illustrate score distributions and a bar plot to illustrate the degree to which each group's scores deviated from the baseline according to the regression coefficients.

Results

Group E outperformed all other groups by a large margin, according to the regression model, and received the highest overall scores.

- Group E scored **25.97 points higher on average** than Group A.
- Group D scored **14.00 points higher**, Group C scored **8.33 points higher**, and Group B scored **2.08 points higher**.

This pattern was supported by the boxplot, which revealed that Group A had the lowest median total scores and Group E the highest. Also, the R2 value of 0.015 showed that race and ethnicity alone only partially explains the variance in the overall score.

Discussion

These findings demonstrate that, on average, Group E kids perform the best academically, while Group A children routinely receive the lowest scores. The obvious discrepancies across groups raise significant concerns about potential educational injustices, even though the R2 score indicates that race/ethnicity alone is not a strong predictor of total scores. Systemic factors such as unequal access to resources, variations in the quality of education, or differing degrees of assistance outside of school could be the cause of these inequities.

These results give stakeholders proof that focused measures could be required to guarantee fair opportunities for every kid. Examining variables such as socioeconomic position, parental education, or access to test-prep classes, for instance, may assist pinpoint areas in need of development. By creating a more inclusive and fair learning environment, closing these disparities can help not only the children who are immediately impacted but also the larger educational system.

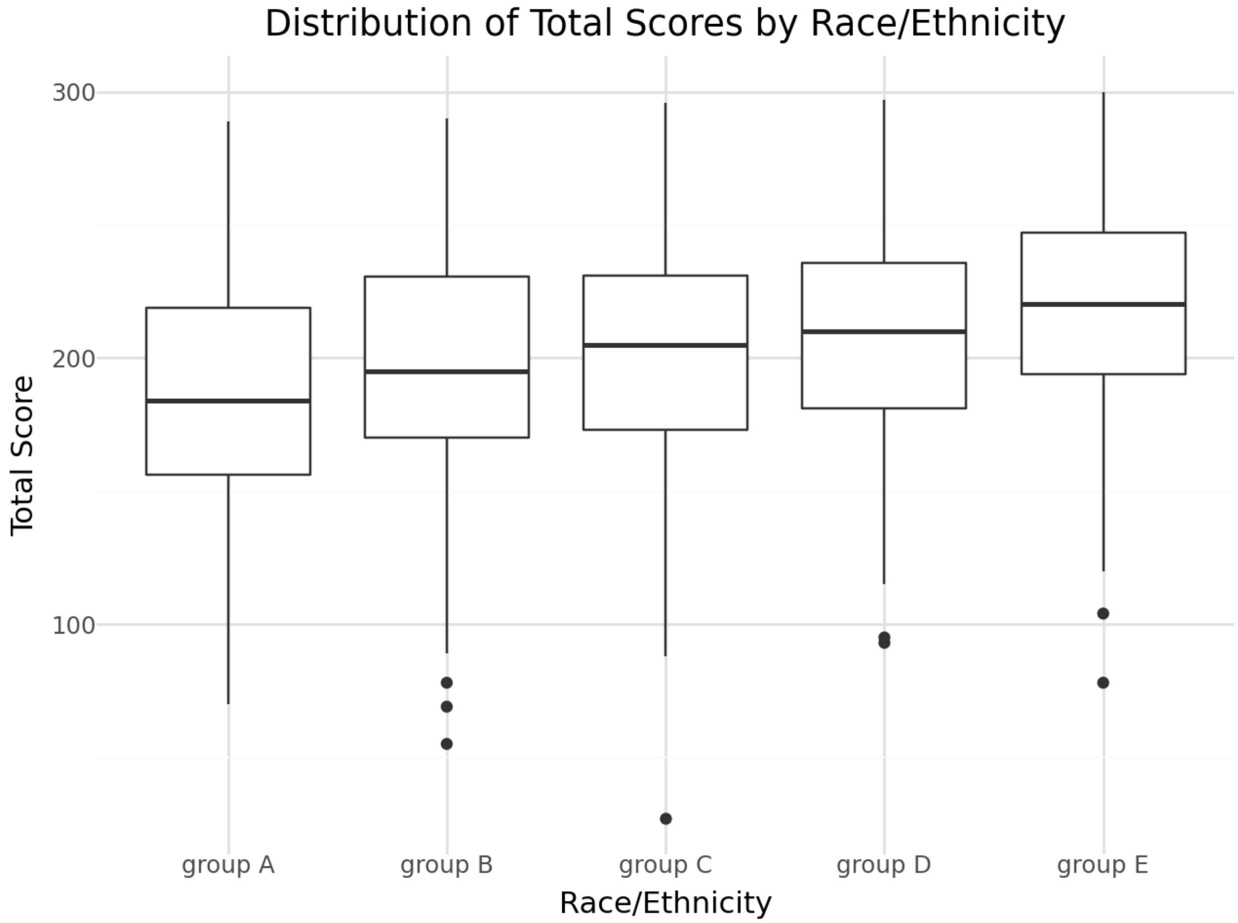

Figure 1: Box plot representing relations between race/ethnicity and total score

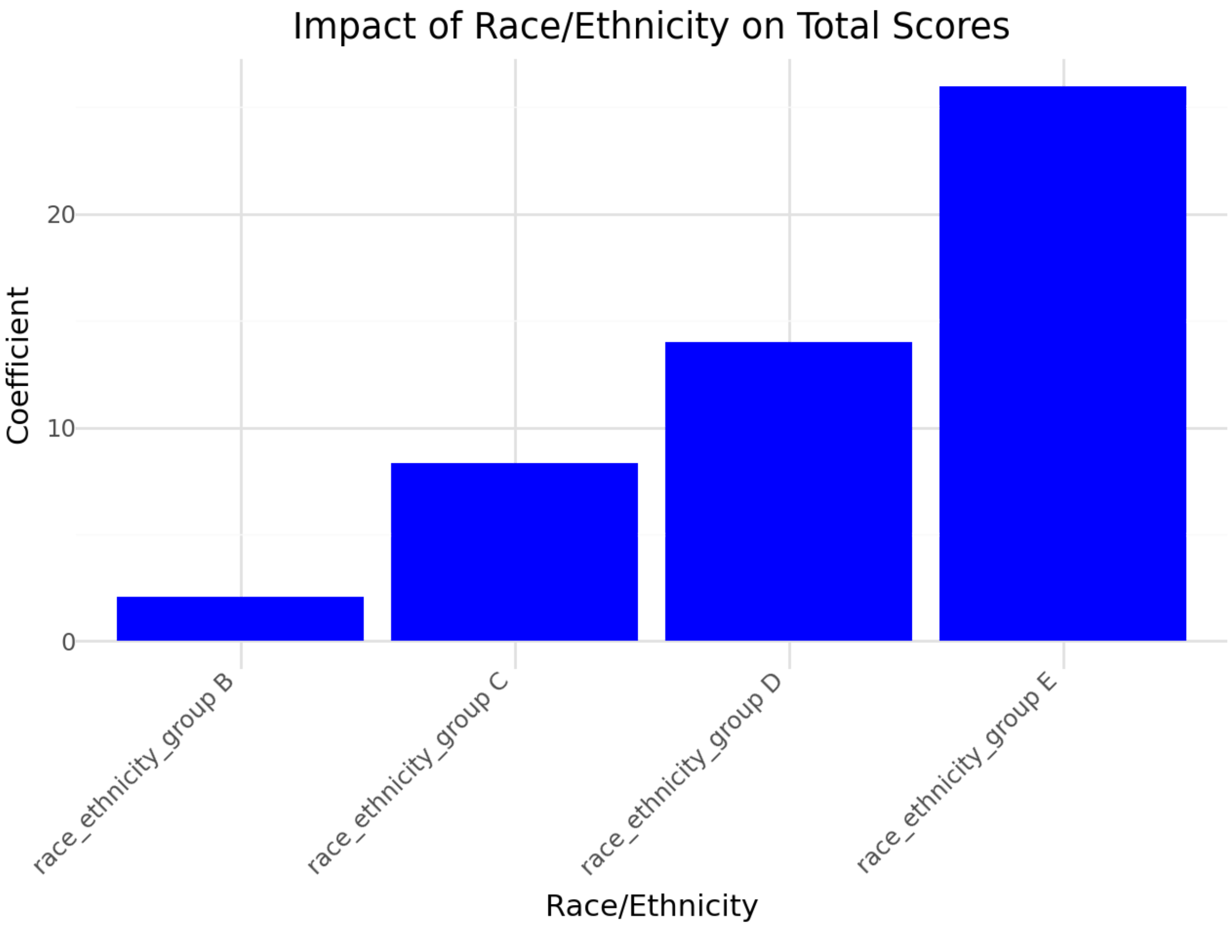

Figure 2: A bar graph of the impact of belonging to a specific race on a student's test results

Question 2: When clustering students based on math, reading, and writing scores, what distinct groups emerge? How do these clusters differ by characteristics like parental education, test preparation, and what insights can we draw from these differences?

Methods

To group students based on their academic performance, we used the KMeans clustering algorithm on their math, reading, and writing scores. To guarantee fairness in clustering, these scores were standardized. Additionally, we included categorical features like gender and test preparation course participation. We determined the optimal number of clusters using the silhouette score, which suggested dividing the students into two distinct clusters. Once clustered, we analyzed how the groups differed based on characteristics like parental education, test preparation, and gender.

Results

Two groupings emerged from the analysis, we'll call them clusters 1 and 2. In cluster 1, math, reading, and writing scores were higher for students in this group. A higher percentage of male kids, pupils with parents with advanced degrees, and students who had taken test-prep classes were included in this group. Almost half of Cluster 0 finished a course on test preparation. However, this changes dramatically when compared to cluster 2, in which students performed significantly worse overall. Less than 25% of these children took test-prep classes, there were more female students in this group, and a larger proportion of pupils had parents with only a high school diploma or less.

Discussion

The results of this clustering study show distinct trends in student performance. Cluster 1, which is linked to higher scores, highlights how important parental education and test-taking chances are to academic success. On the other hand, Cluster 2 identifies possible problems that students may have, such as not preparing for the test properly and receiving little assistance from parents with less education. The performance gap may be closed by addressing these differences with focused interventions, such as providing more easily accessible test-prep courses and assisting families with lower educational attainment.

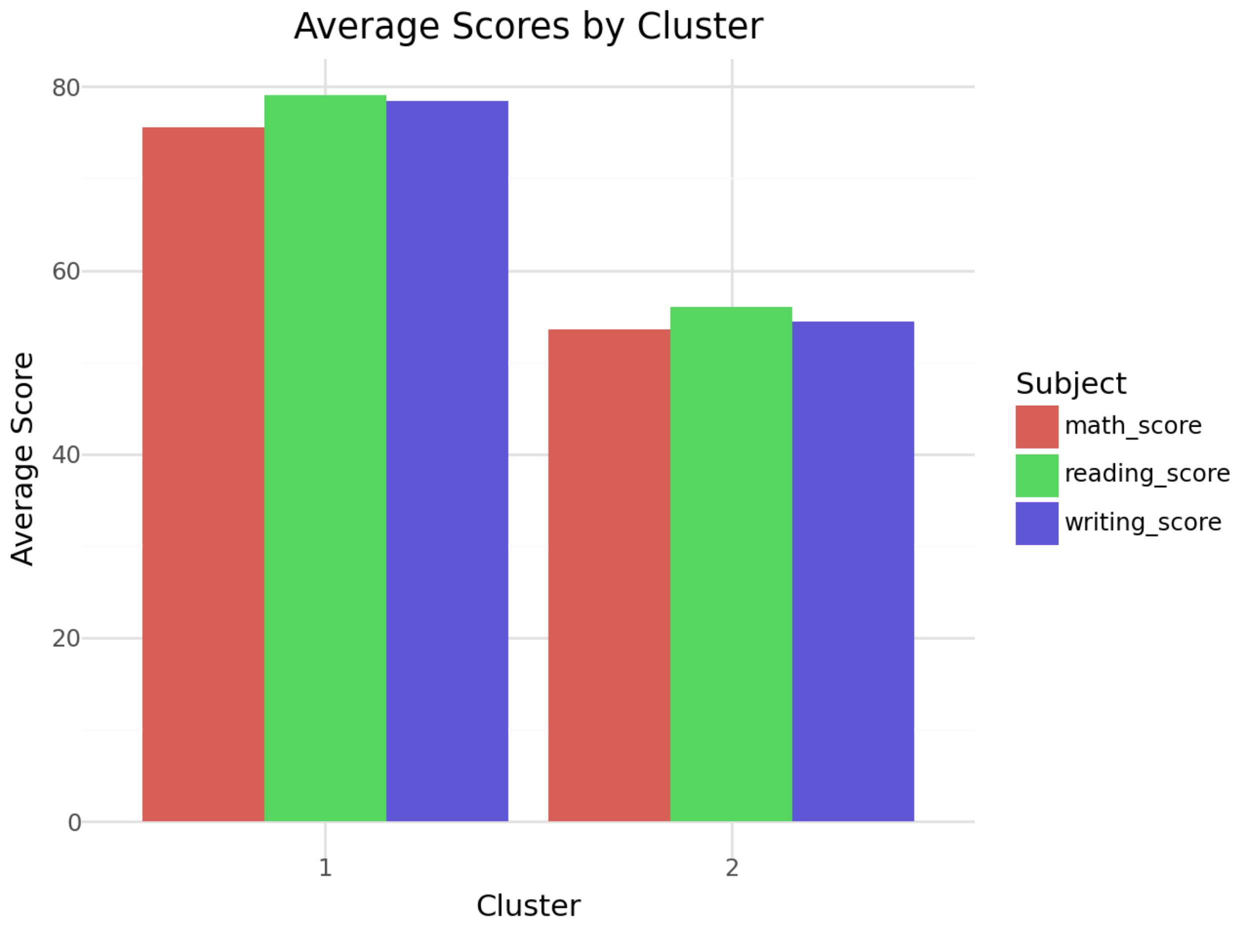

Figure 1: The average scores for each student based on their cluster

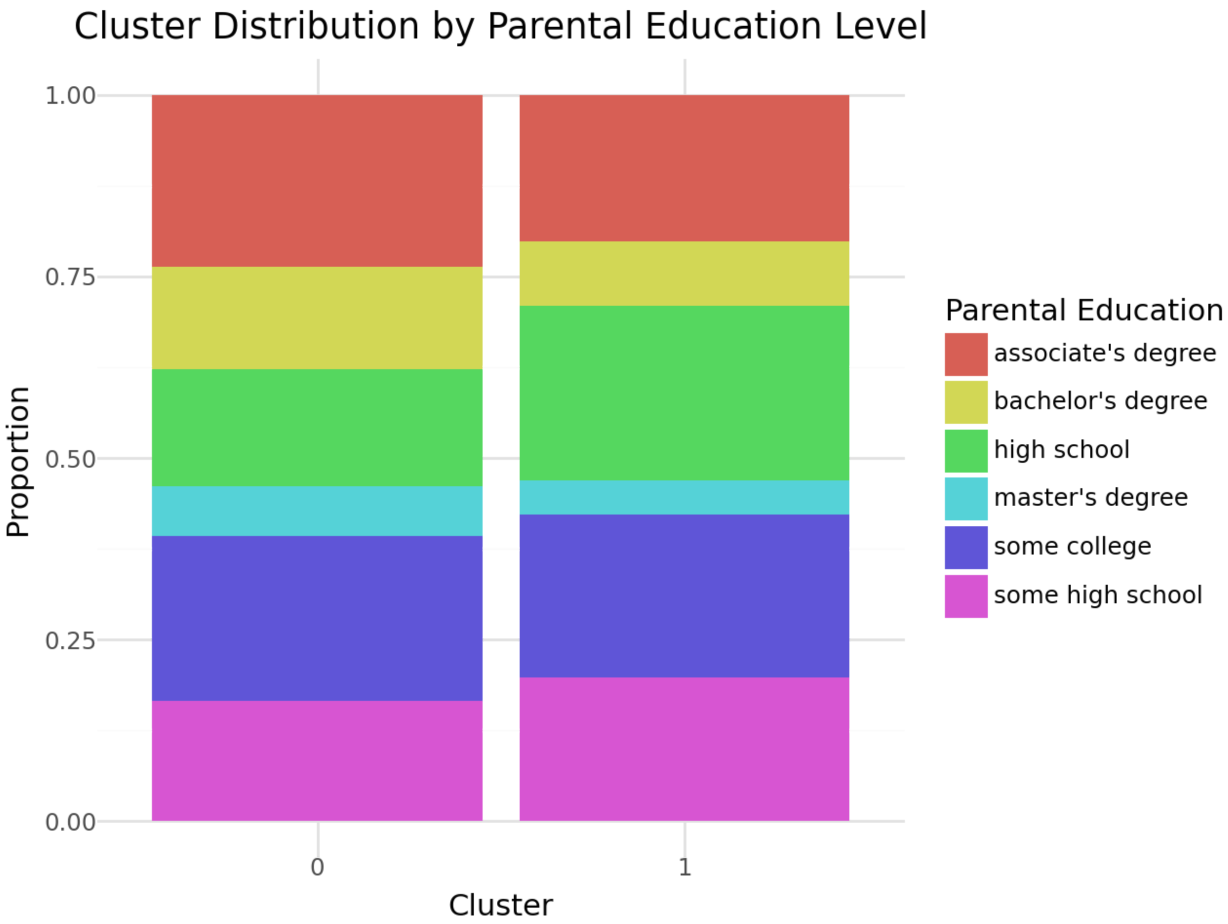

Figure 2: A graph comparing the proportions of parental education between the two clusters.

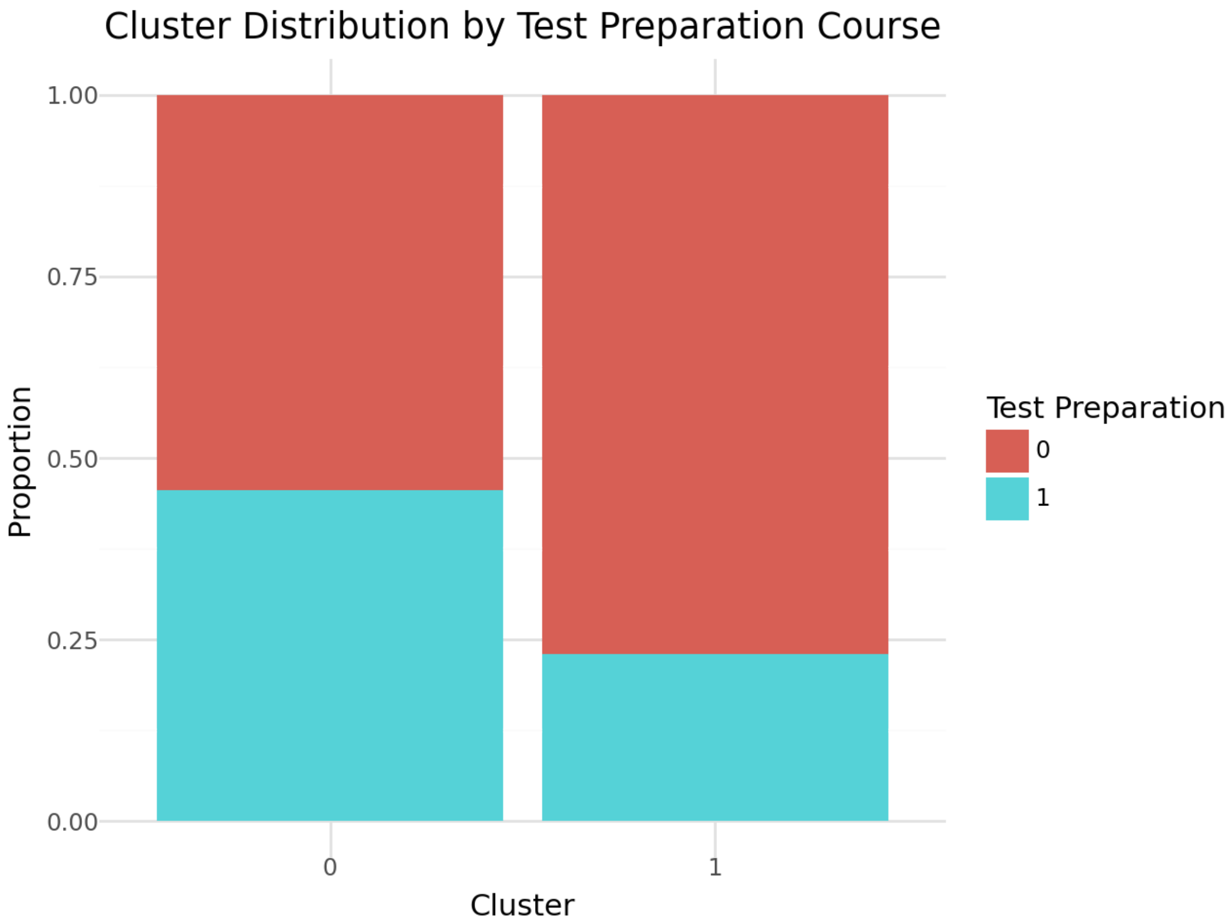

Figure 3: The proportion of students who underwent test preparation grouped by cluster.

Question 3: Which parental education level is associated with the highest scores across the three subjects (math, reading, and writing), and does the level of parental education influence consistency in student performance across all subjects?

Methods

We examined the test scores for each parental education group in order to look at the connection between parental education and student performance. We divided the three-dimensional subject score data into two principal components using Principal Component Analysis (PCA). This change made it easier to visualize the data, and check for any clusters formed from the data.

Results

The grouped bar chart revealed that students with parents holding a master's degree achieved the highest average scores across all three subjects, while those with parents whose education ended at high school or graduated high school scored the lowest. As for consistency, the PCA scatter plot showed no significant clustering, suggesting that parental education level does not reliably and consistently influence student performance across subjects.

Discussion

Higher levels of education are linked to improved academic performance, suggesting that parental education has a discernible effect on student achievement. The PCA scatterplot, however, shows that there is no discernible difference in performance consistency between subjects based on parental educational attainment. This implies that, similar to their counterparts with lower parental education levels, students with highly educated parents do better on average but maintain a balanced relative performance in math, reading, and writing. These results emphasize how crucial it is to assist children whose parents have less education in order to lessen performance gaps and advance fair academic results.

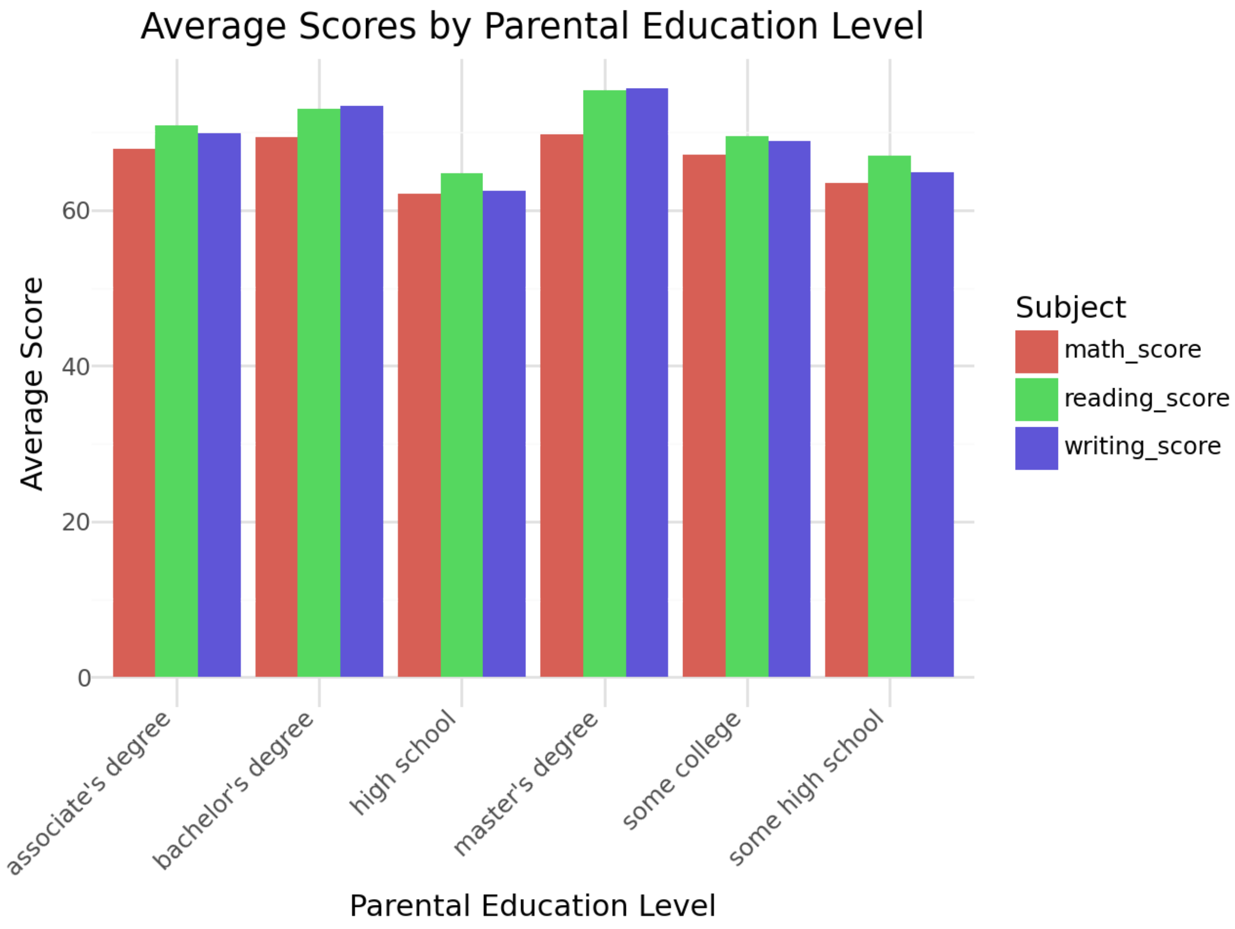

Figure 1: A bar graph displaying the average scores for each subject grouped by the parental level of education.

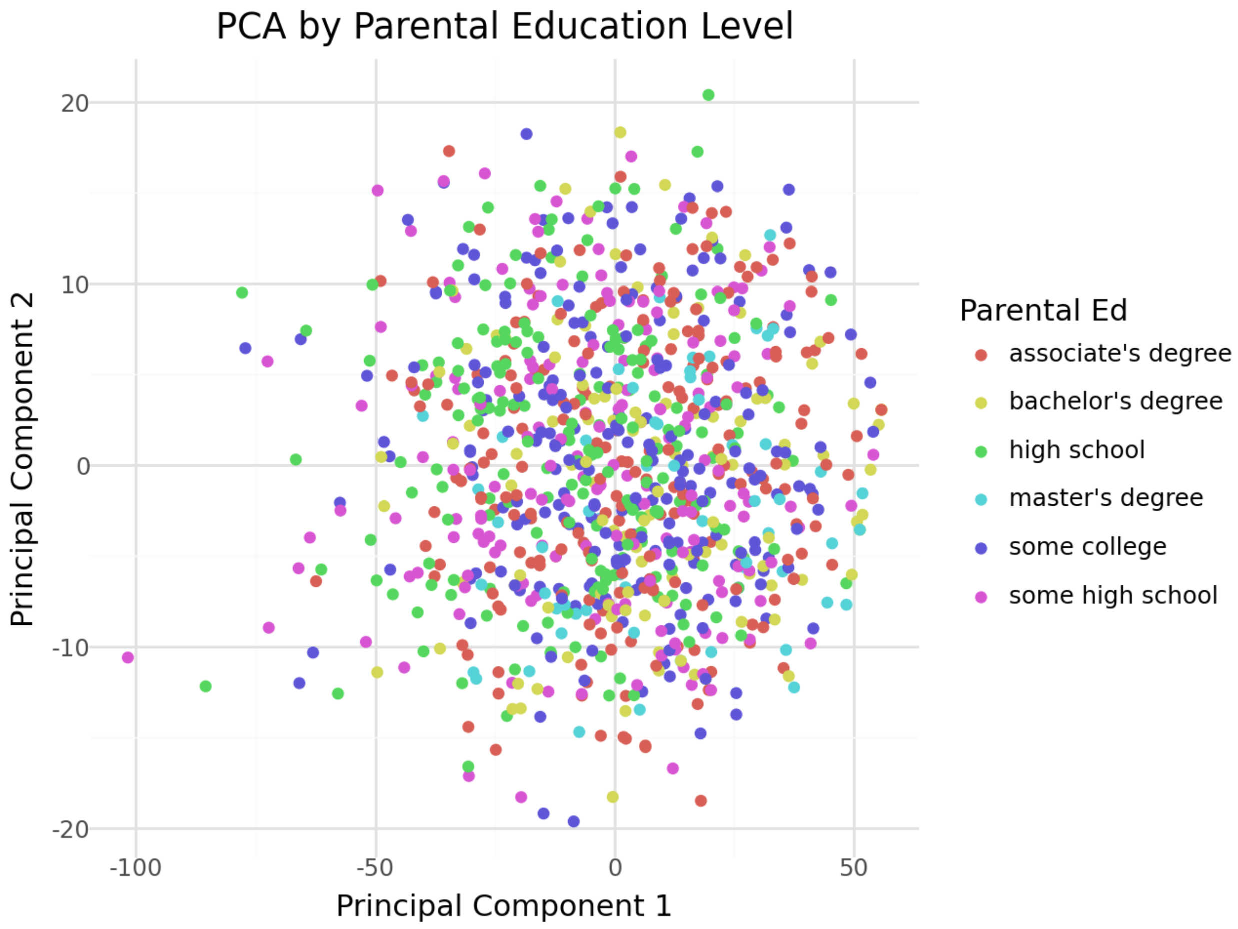

Figure 2: A cluster diagram displaying the relative distance between points in groups.

Question 4: When predicting the average score, which predictor (gender, race/ethnicity, parental level of education, lunch type, test preparation course) has the strongest effect on the model’sR^2 when removed and what does this reveal about the importance of each factor in academic performance?

Methods

We constructed a linear regression model using all predictors listed above to figure out which predictor has the most impact on the average score. Using all of the predictors included, we calculated the model's R2 score to evaluate its performance and compared it to R2 scores obtained by deleting each predictor separately. This enabled us to evaluate the effect of eliminating a predictor on the explanatory power of the model. To illustrate these variations, a bar chart was employed. Lastly, to visualize the prediction accuracy of the algorithm, we graphed the anticipated scores against the real scores.

Results

As expected, when all predictors were included, the entire model's performance score was at its greatest. The most notable drop in performance was observed when parental education levels (master's and college degrees) were eliminated, suggesting that this variable had a major impact on the model's capacity to forecast average scores. Removing group B resulted in the biggest decline in performance, followed by groups C, D, and E among the race groups. When eliminated, test preparation course and lunch type had the least impact on performance, indicating that they are poorer predictors. Predictions closely followed a diagonal line in the scatterplot of projected versus actual scores, indicating a strong alignment. At higher scores, however, such as when real scores approached 92, our model overpredicted, indicating that the relationship is nonlinear.

Discussion

According to our model, the most important factor affecting average scores is the education level of the parents, indicating that more educated parents give their children access to tools or circumstances that support improved academic achievement. Additionally important is race and ethnicity, especially for group B, which may reflect inequalities or systemic variables influencing academic achievement. The sort of lunch and the test-prep course, on the other hand, had little effect. This indicates that the test prep course is relatively ineffective, and indicates that it should either be restructured or removed by the school.

Although minor overpredictions at high scores would indicate that the model is not fully capturing the subtleties of high-achieving students, the projected vs. actual scatterplot validates the model's dependability. This information could be used by the school district to reallocate its resources to try to maximize student performance.

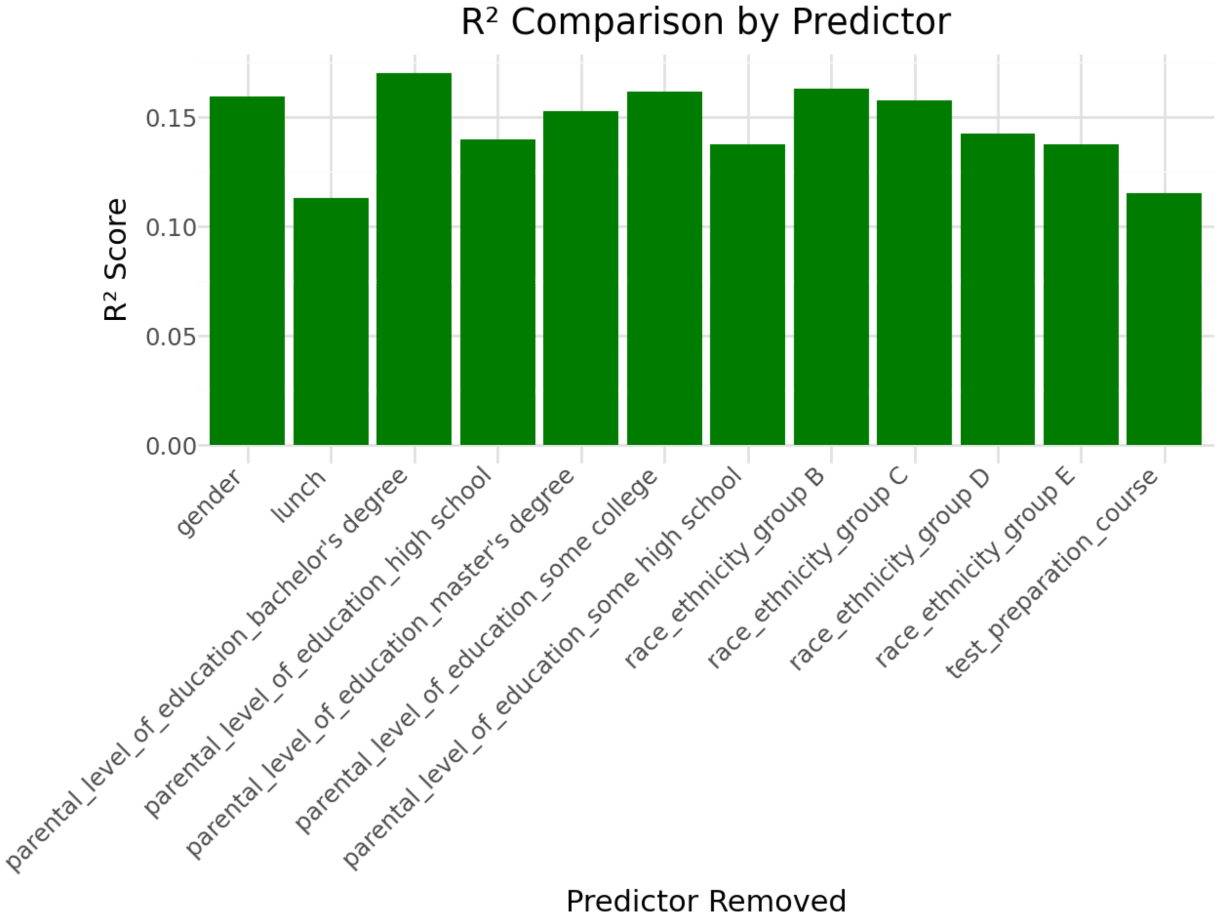

Figure 1: A bar graph displaying the R2 score of the model with each predictor removed independently.

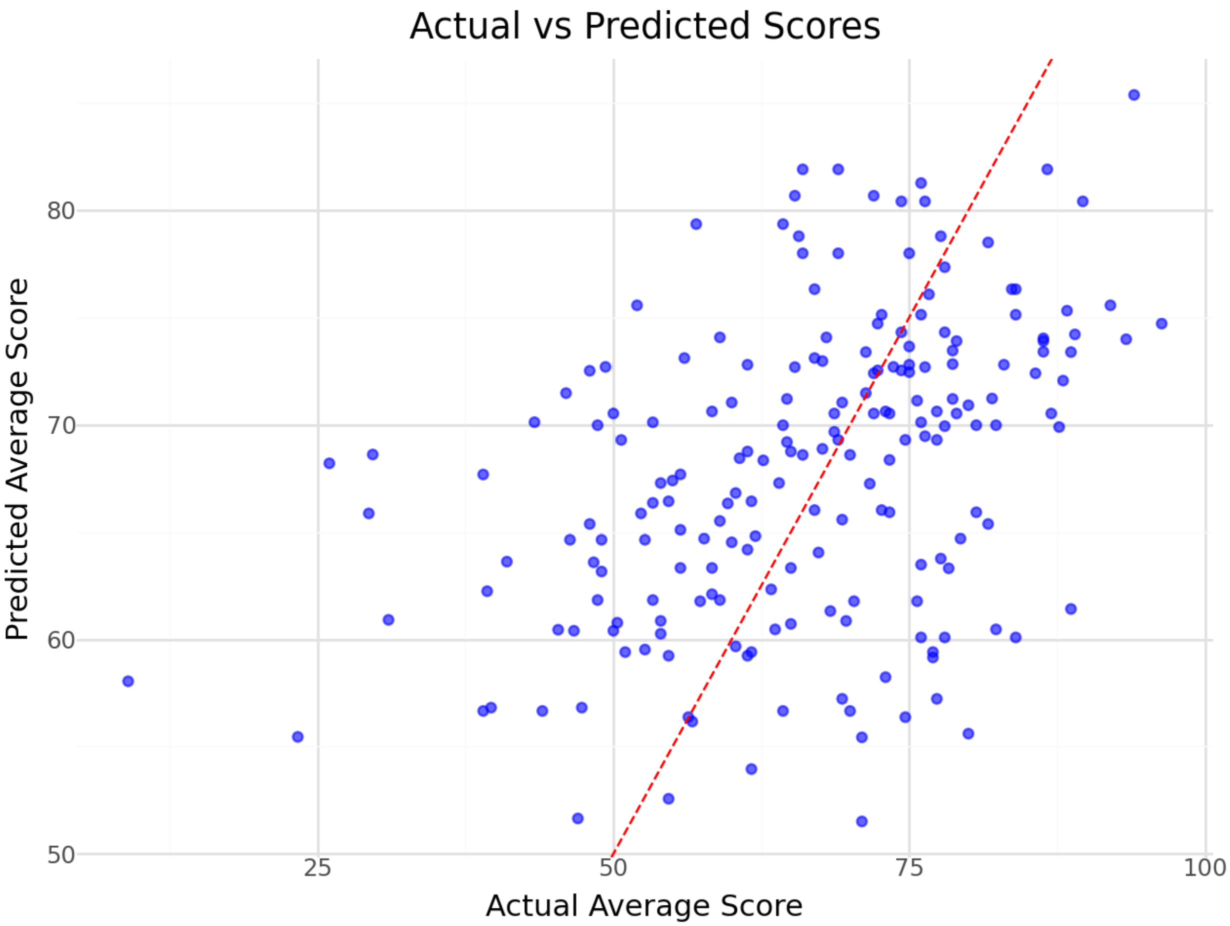

Figure 2: A linear approximation of a student's predicted score and their actual score made with our model.