SEMESTER PROJECT

STUDENT PERFORMANCE ANALYSIS

UNDERSTANDING THE INFLUENCIAL FACTORS ON STUDENT'S PERFORMANCE

STA 6704 – Data Mining II

Instructor: Dr. Aaron Smith

Student: Trung Lai Nguyen

1. Project Statement

The objective of this project is to investigate predictors of students' academic performance by considering several factors such as gender, parents' education levels, test preparation, and students' lunch meal using a multiple linear regression analysis. To specify, this project aims to address the following questions:

1. Whether the performance of female students significantly differs from that of male ones on average.

2. Whether the students' learning performance varies across the ethnic groups.

3. Whether the parents' education levels predict students' academic achievement.

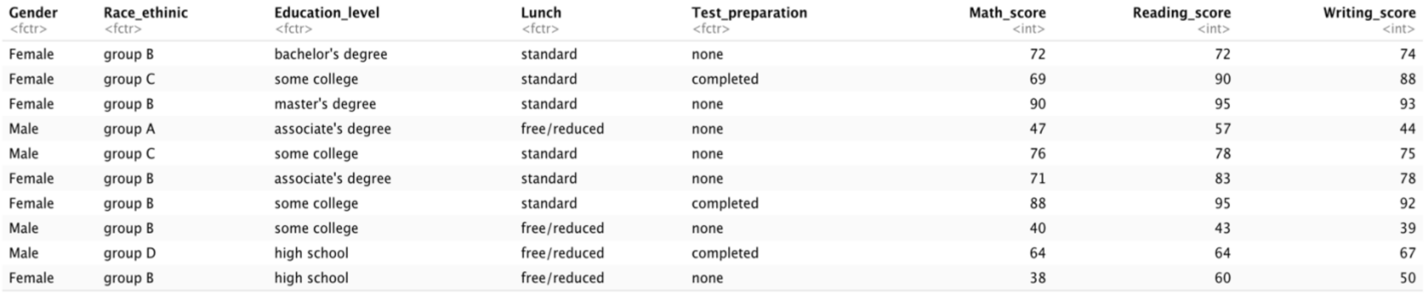
4. Whether test preparation plays a critical role in improving students' performance.

5. Whether there is any difference in academic performance of students who have free lunch versus those who have the standard meal.

2. Exploratory Data Analysis

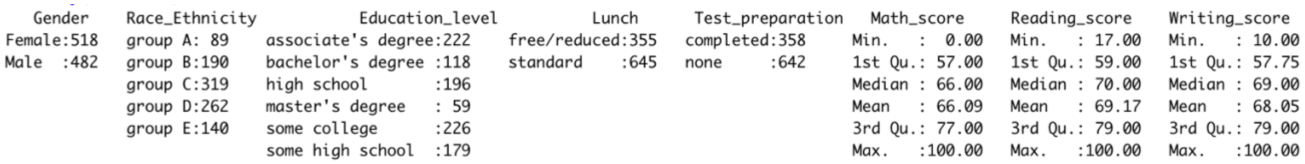
2.1. About the data set

The data set utilized in this project is gathered from Kaggle. It records the performances of 1,000 students in three different exams including math, reading and writing exams. The data set also provides descriptive information about the students including gender, race/ethnicity, parents' education levels, lunch meal and whether the student completed a test preparation course before the exams.



Gender	Race_ethinic	Education_level	Lunch	Test_preparation	Math_score	Reading_score	Writing_score
<fctr>	<fctr>	<fctr>	<fctr>	<fctr>	<int>	<int>	<int>
Female	group B	bachelor's degree	standard	none	72	72	74
Female	group C	some college	standard	completed	69	90	88
Female	group B	master's degree	standard	none	90	95	93
Male	group A	associate's degree	free/reduced	none	47	57	44
Male	group C	some college	standard	none	76	78	75
Female	group B	associate's degree	standard	none	71	83	78
Female	group B	some college	standard	completed	88	95	92
Male	group B	some college	free/reduced	none	40	43	39
Male	group D	high school	free/reduced	completed	64	64	67
Female	group B	high school	free/reduced	none	38	60	50

Figure 1: Summary of the Student Performance in Exams data set (Kaggle, 2022)



```
   Gender     Race_Ethnicity            Education_level          Lunch        Test_preparation   Math_score       Reading_score    Writing_score
 Female:518   group A: 89    associate's degree:222   free/reduced:355   completed:358     Min.   :  0.00   Min.   : 17.00   Min.   : 10.00
 Male  :482   group B:190    bachelor's degree :118   standard    :645   none     :642     1st Qu.: 57.00   1st Qu.: 59.00   1st Qu.: 57.75
              group C:319    high school       :196                                        Median : 66.00   Median : 70.00   Median : 69.00
              group D:262    master's degree   : 59                                        Mean   : 66.09   Mean   : 69.17   Mean   : 68.05
              group E:140    some college      :226                                        3rd Qu.: 77.00   3rd Qu.: 79.00   3rd Qu.: 79.00
                             some high school  :179                                        Max.   :100.00   Max.   :100.00   Max.   :100.00
```

Figure 2: Summary of the attributes

According to Figure 2, there are more female students (512) than male students (482) in the data set. In terms of the race and ethnicity, there are five groups which have been coded as group A, B, C, D and E, among which group C is the largest one. The education levels of students' parents are divided into six categories, including high school, some high school, some college, associate's degree, bachelor's degree and master's degree. The student's lunch meal consists of two categories as free/reduced and standard meals. Lastly, the test preparation attribute is classified as "completed" and "none", in which "completed" means that the student has completed a preparation course before the exams and "none" refers to no preparation course taken.

2.2. Explore the student performance

2.2.1. Which gender performs better?

As a common stereotype, males are believed to have better logical skills than female, especially in subjects related to natural science such as math, physics and chemistry, while female are expected to be better at social science subjects including law, economics and history. This infers that females generally have better reading and writing skills than males. At this stage, boxplots are utilized to confirm whether that common belief remains valid with this data set.

In the Math exam:

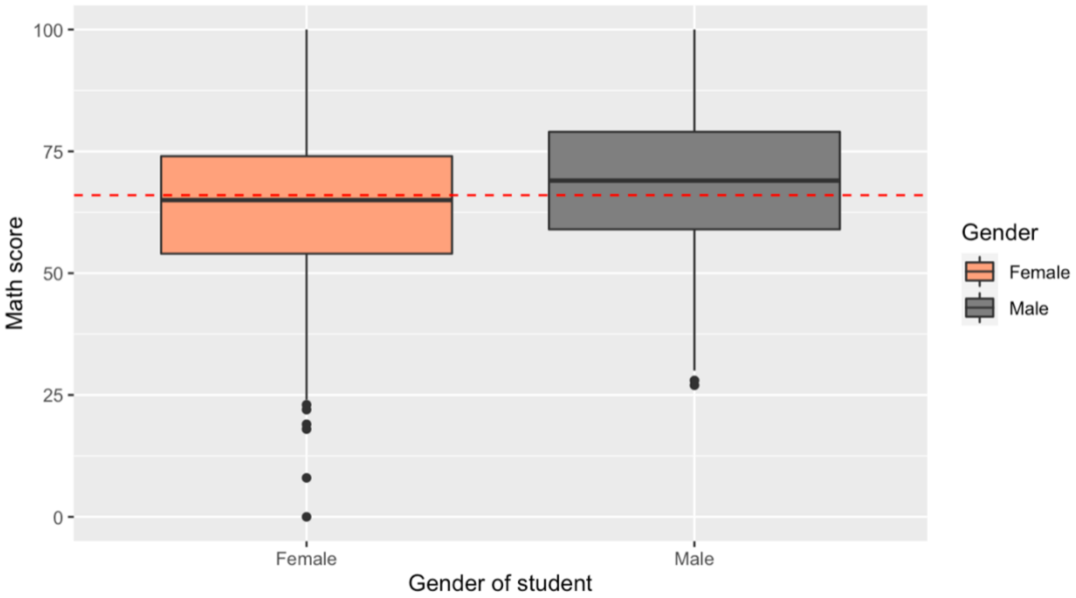


Figure 4: Math scores by gender of students

According to Figure 4, the average math score for 1,000 students is 66 which is presented by the red dashed line. It could be seen that in overall, male students performed better than female ones in the math exam as the median math score of the former was higher than that of the later. Figure 4 also reflects that the majority of male population has above average performance in the math exam while more than half of the female students performed below the average.

In the Writing exam:

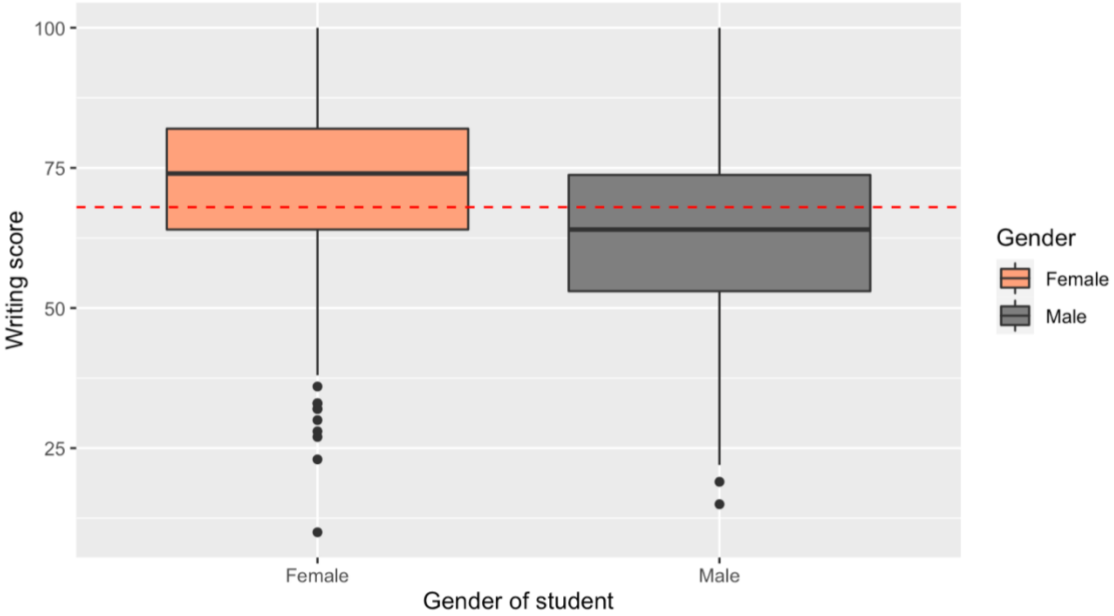


Figure 5: Writing scores by gender of students

In Figure 5, the average writing score for the entire pool is 68, represented by the red dashed line. In contrast with the math exam results, female students had better performance than male students in the writing exam because more than fifty percent of female students achieved a higher score than the average while more than half of the male students' scores is below the average.

In the Reading exam:

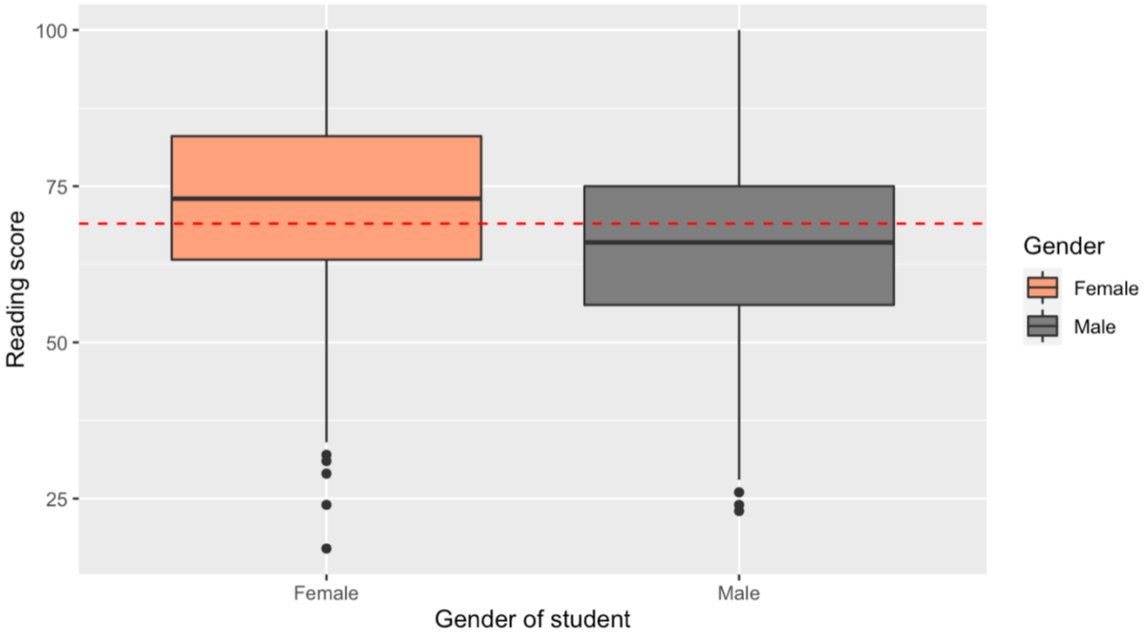


Figure 6: Reading scores by gender of students

From the boxplot in Figure 6, the average of reading score is 69, reflected by the red dashed line. While more than half of the female students had the scores higher than the average, the majority of male students performed lower than the average. Therefore, female students had better performance than male students in the reading exam.

From the above interpretation, the stereotype remains true with this data set. Specifically, male students are better in the math exam while female students surpass their counterparts in the reading and writing exams.

In Overall:

The overall performance of students is evaluated according to the total score of three exams. Thus, I will compute the GPA based on the overall score of each student in three exams including math, reading and writing scores. Specifically, a student's GPA is A if their overall score is between 90 and 100; from 80 to 89, GPA is B; between 70 and 79, GPA is C; from 60 to 69 is D and below 59 is F.

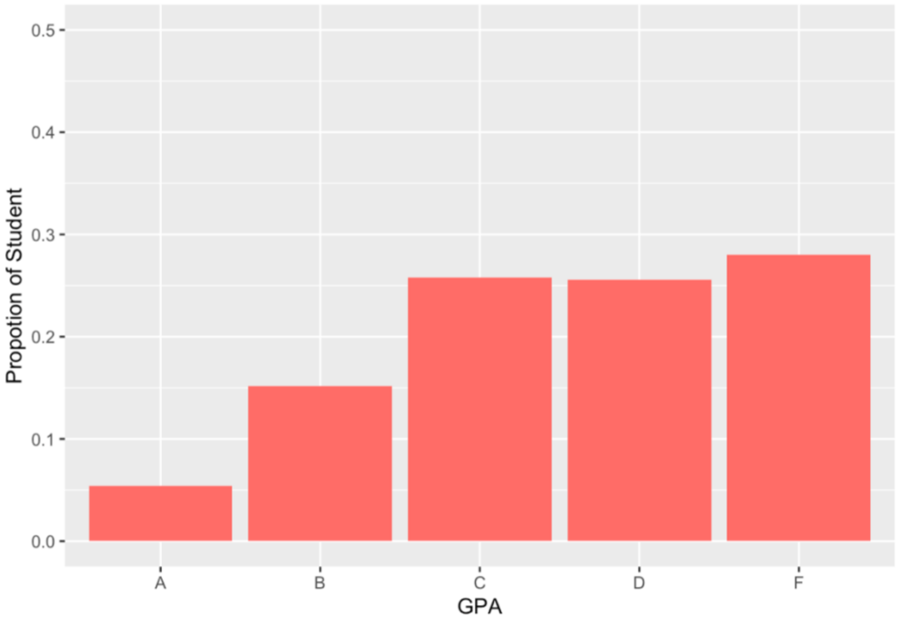


Figure 6: Proportion of students by GPA

In Figure 6, the number of students who achieved A comprise the smallest proportion while the larger proportions are belong to the ones with lower GPA. This is not a favorable result as the proportion of F is the highest.

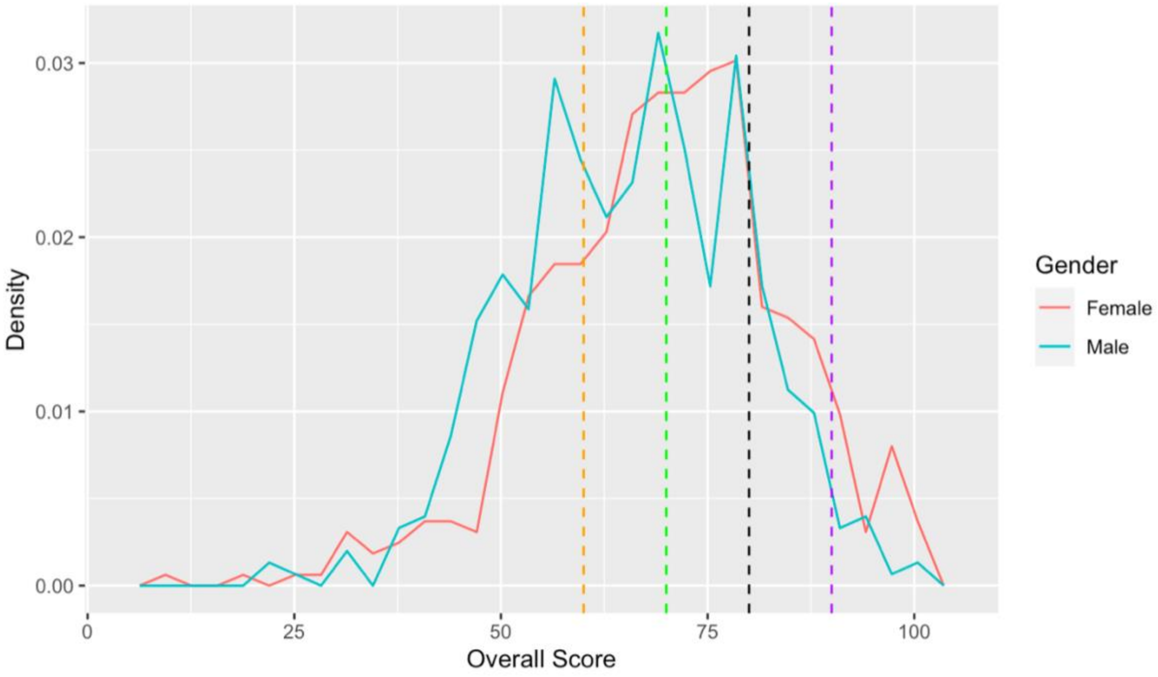


Figure 7: Density curves of overall score

Figure 7 demonstrates the density of overall scores categorized by gender. In particular, the proportions of male and female students are presented by the areas below the blue and red density curves respectively over their overall scores. The orange, green, black and purple dashed lines divide the overall scores into five ranges of GPA: F, D, C, B and A. In this case, the passing score (60) is reflected by the orange dashed line. Scores on the left of the orange dashed line are considered as failed (F). Scores between orange (60) and green (70) dashed line are GPA of D. The range from green (70) to black (80) dashed line is GPA of C. The interval among black and purple dashed lines is corresponding to GPA of B. The purple dashed line mark at 90, scores on the right of which are considered as GPA of A. The corresponding areas of the density curves for female students in the GPA range from B to A are greater than that of male students. It infers that the number of female students whose having GPA of A and B was higher than that of male students.

2.2.2. Race/Ethnicity and Student performance:

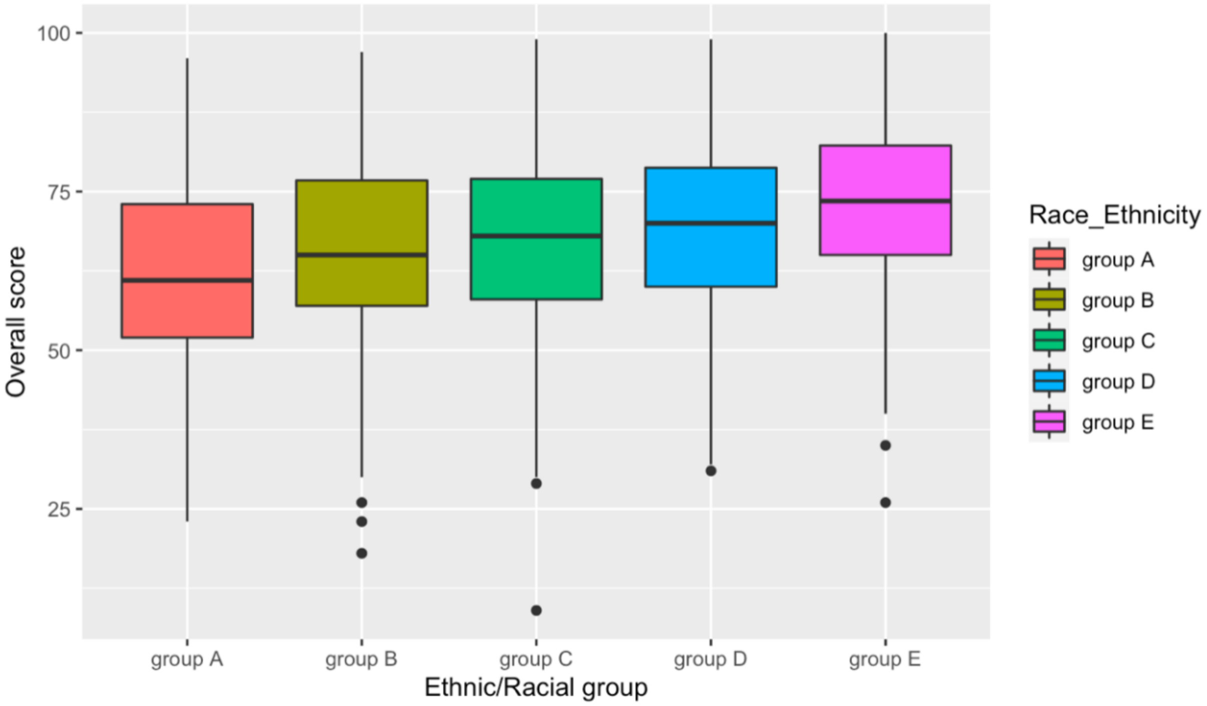


Figure 8: Overall scores of each race/ethnicity group

The boxplot in Figure 8 demonstrates the distribution of overall scores among racial/ethnic groups. Specifically, while the maximum score that students in Group E achieved is 100, the highest score of Group C and D is approximately 98 and that of Group A and B is about 95 and 96 correspondingly. Besides, the median scores split each group's overall scores into two halves. It can be seen that the upper half of the students in Group E obtained the highest scores, followed by that of Group B, C, D and A. This infers that in overall, students in Group E had the best performance while those in Group A had the lowest performance.

2.2.3. Parent's education levels and Student performance:

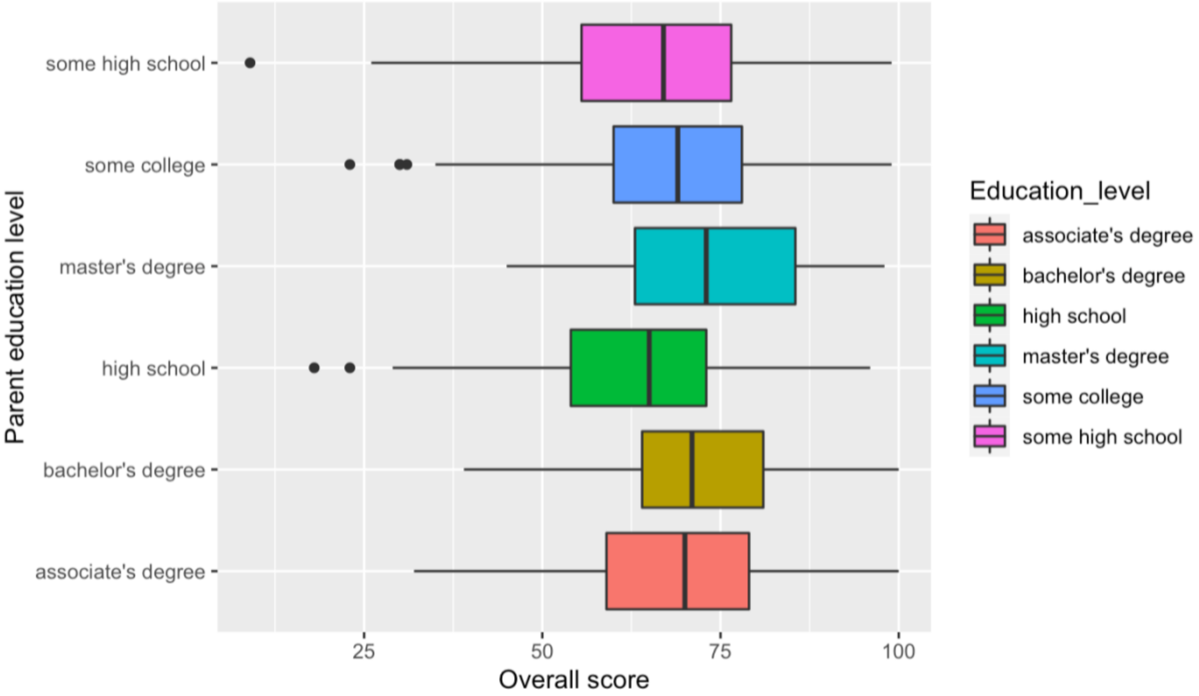


Figure 9: Distribution of overall scores respective to *parents' education level*

The boxplot in Figure 8 presents the distribution of overall scores across parents' education levels. In particular, while the students whose parents own bachelor's degree or associate's degree could earn up to 100 points, those whose parents are high school graduates could only achieve up to 90. In addition, the median scores show that the upper-half of students whose parents possess master's, bachelor's or associate's degree had better performance than the upper-half of those who have non-degree-parents.

2.2.4. Lunch meal and Student performance

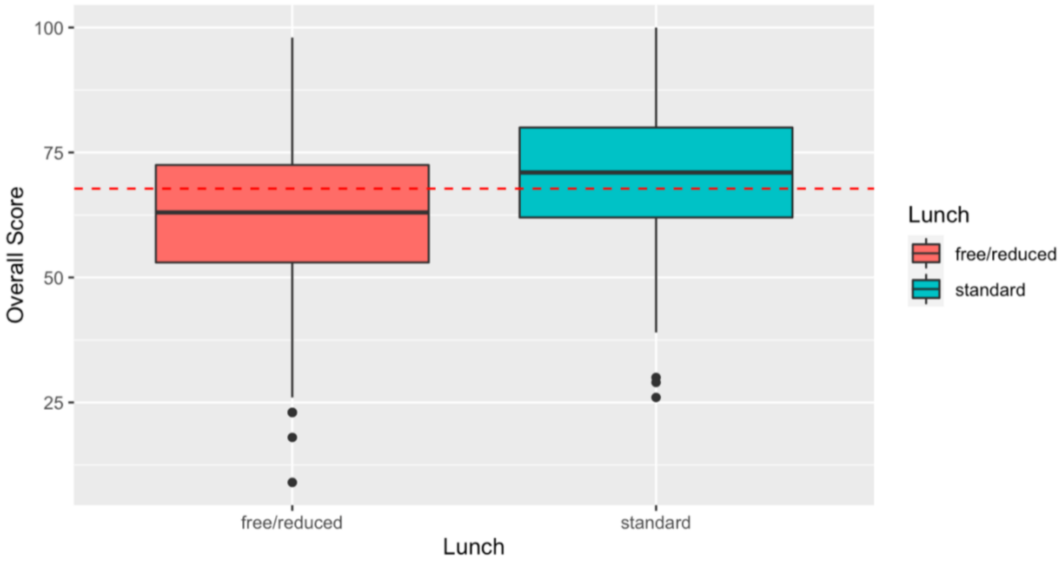


Figure 10: Distribution of overall scores respective to lunch meal

In Figure 10, the average of overall score for the entire pool is 67.8, demonstrated by the red dashed line. The median score of the students having standard lunch meal is higher than the average score, inferring that more than half of the students having standard lunch performed above the average. On the other hand, the majority of the students who received free/reduced meal had below average performance.

2.2.5. Test preparation course and Student performance

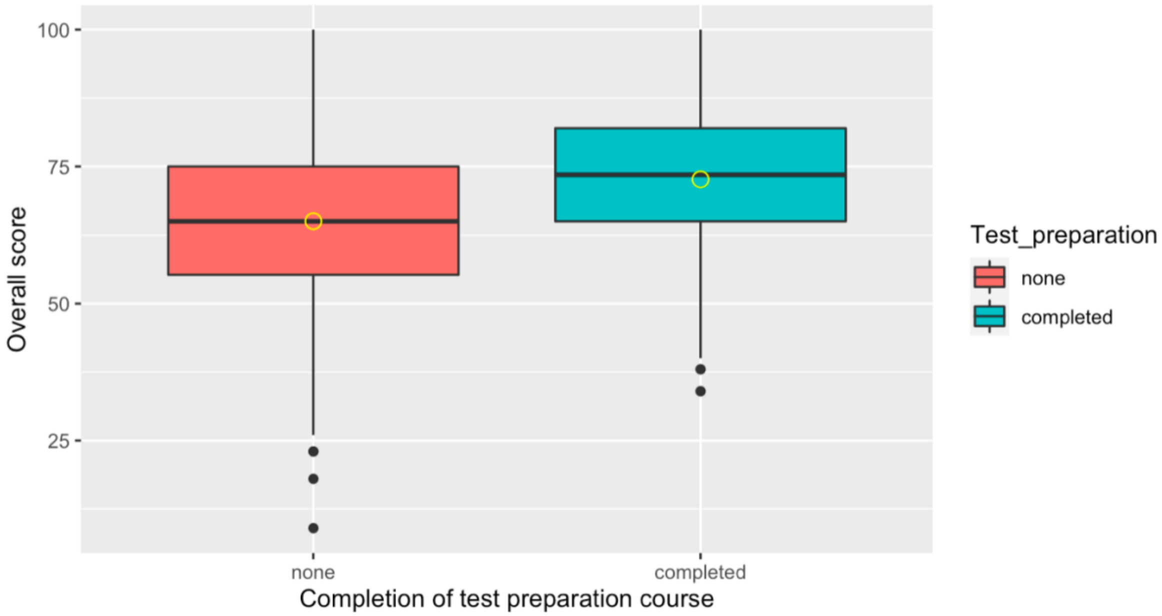


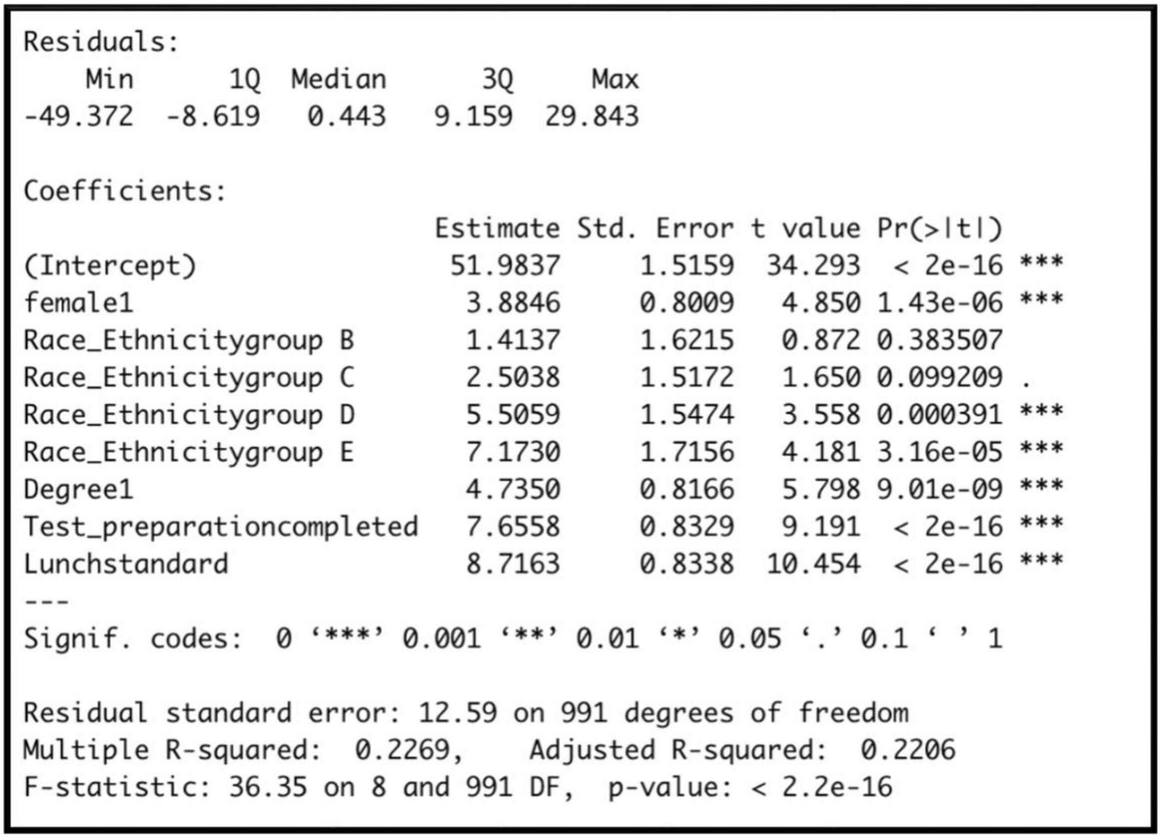
Figure 11: Distribution of overall scores respective to test preparation course

From the boxplot in Figure 11, the students who have done the test preparation course before the exams have higher median scores than those have not completed the course. This means that completing the preparation course could help students achieve a greater overall score.

3. Influential factors and Student performance

Based on the previous exploration, it is expected that the performance of students could be influenced by factors such as gender, racial/ethnicity, parents' education levels, lunch meal and whether students completed a test preparation course. In order to identify the impact of these factors on students' performance, a multiple linear regression is utilized. In this case, the dependent variable is the overall score and the predictors are five factors including gender, racial/ethnicity, parents' education levels, lunch meal and whether the student complete a test preparation course.

Next, a 5-fold cross validation is conducted to measure the model's accuracy. RMSE value achieved is approximately 12.65 which is considered as high. It could be explained by that all of the predictor is categorical.



```
Residuals:
    Min      1Q  Median      3Q     Max 
-49.372  -8.619   0.443   9.159  29.843 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                51.9837     1.5159  34.293  < 2e-16 ***
female1                     3.8846     0.8009   4.850 1.43e-06 ***
Race_Ethnicitygroup B       1.4137     1.6215   0.872 0.383507    
Race_Ethnicitygroup C       2.5038     1.5172   1.650 0.099209 .  
Race_Ethnicitygroup D       5.5059     1.5474   3.558 0.000391 ***
Race_Ethnicitygroup E       7.1730     1.7156   4.181 3.16e-05 ***
Degree1                     4.7350     0.8166   5.798 9.01e-09 ***
Test_preparationcompleted   7.6558     0.8329   9.191  < 2e-16 ***
Lunchstandard               8.7163     0.8338  10.454  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.59 on 991 degrees of freedom
Multiple R-squared:  0.2269,	Adjusted R-squared:  0.2206 
F-statistic: 36.35 on 8 and 991 DF,  p-value: < 2.2e-16
```

Figure 12: The model results

Figure 12 summarizes the model’s results. The adjusted R-squared is 0.2269, inferring that 22.7 percent of the overall score is explained by the considered predictors. To specify, the results show that gender (Estimate = 3.8846, *t* value = 4.850, $p < 0.05$) significantly influenced the overall score. For example, the overall score is predicted to increase 3.8846 points for female students. It means that female students achieved 3.8846 points higher than their male counterparts.

The model results present the significant effect of race/ethnicity on the overall score. For instance, students from Group D (Estimate = 5.5059, *t* value = 3.558, $p < 0.05$) and Group E (Estimate = 7.173, *t* value = 4.181, $p < 0.05$) have 5.5059 and 7.173 points higher than those from Group A, while there is no significant difference among students from Group A, B and C.

The education levels of parents (Estimate = 4.735, *t* value = 5.798, $p < 0.05$) are also found to significantly predict students' overall score. In particular, students whose parents possess degree have 4.735 points higher than those whose parents did not achieve any degree.

The test preparation course (Estimate = 7.6558, *t* value = 9.191, $p < 0.05$) is found to have significant impact on the overall score of students. It infers that the preparation course can help to improve the overall score. For instance, the student who completed the test preparation course can achieve 7.6558 points higher than those did not.

Lunch meal (Estimate = 8.7143, *t* value = 10.454, $p < 0.05$) also significantly influence the overall score. The students who have standard lunch have 8.7143 points higher than who have free/reduced meal.

4. Conclusion

In conclusion, most of students did not perform well in the exams. The proportion of students who had GPA of F is much greater than that of students having GPA of A and B. Furthermore, the analysis results also show that students' performance varies across the groups of gender, racial/ethnicity, parents' education levels, lunch meal types and test preparation course completion. This finding could be helpful for educators to figure out how to improve students' performance in the future in several ways:

- In terms of gender, to increase the overall score in three exams, female students should be encouraged to improve their math skills while male students should enhance their writing and reading skills.
- Completing the test preparation course will be beneficial for students to have better overall score in three exams. Teachers should encourage students to attend the course to enhance their academic performance.
- Students whose parents did not have any degree could have less motivation to study than those whose parents hold a degree. Therefore, achievement-oriented behaviors (i.e., studying and obtaining advanced degree) should be encouraged with students having non-degree parents to motivate their study performance.

Additionally, this project findings suggest future analysis to consider other variables as predictors of the students' performance, especially the numeric data such as students' IQ, study time, free time after school and family size. Those additional data could not only help to increase the accuracy of the predicted model but also provide more insights about students' performance.

Reference

Student Performance in Exams. Retrieved from: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams?datasetId=74977&sortBy=voteCount&language=R