University of Calgary

Faculty of Science

Department of Computer Science

DATA SCIENCE AND ANALYTICS 601

Group Project Report

Topic: Analysis of Student Performance Factors

Date Submitted:

15 October 2024

Submitted by:

Arthur Sumague

Agyei Osei Duodu

Ikeora Ekene

Stephen Opoku Bonsu

INTRODUCTION

Understanding the drivers underlying student achievement is crucial for educators, policymakers, and other stakeholders who desire to enhance the performance in educational progress. This study shall investigate other sets of determinants of student performance starting from the socio-economic backgrounds, parental attendance and involvement, school resources, and characteristics of individual students themselves. Investigation will look at a range of factors that influence the level of student attainment from socio-economic background and parental involvement to school resources and individual characteristics of students themselves.

This project gets into insights based on data exploration to find patterns and correlations between various variables and the achievements of students. This shall put forward a comprehensive overview of the most influential factors, with recommendations that can be taken for improving strategies and policies regarding education. This study hopefully will join the continuous effort to push for better educational standards and student success through rigorous analysis and intelligent interpretation.

ABOUT THE DATASET

The data used in the entire analysis was drawn from Kaggle.com. It is comprised of Seven numerical, 13 categorical variables and a total of 1000 rows.

No	Feature	Description
1	Hours_Studied	Number of hours spent studying per week.
2	Attendance	Percentage of classes attended.
3	Parental Involvement	Level of parental involvement in the student's education (Low, Medium, High).
4	Access_to_Resources	Availability of educational resources (Low, Medium, High).
5	Extracurricular_Activities	Participation in extracurricular activities (Yes, No).
6	Sleep_Hours	Average number of hours of sleep per night.
7	Previous_Scores	Scores from previous exams.
8	Motivation_Level	Student's level of motivation (Low, Medium, High).
9	Internet_Access	Availability of internet access (Yes, No).
10	Tutoring_Sessions	Number of tutoring sessions attended per month.
11	Family_Income	Family income level (Low, Medium, High).
12	Teacher_Quality	Quality of the teachers (Low, Medium, High).
13	School_Type	Type of school attended (Public, Private).
14	Peer_Influence	Influence of peers on academic performance (Positive, Neutral, Negative).
15	Physical_Activity	Average number of hours of physical activity per week.
16	Learning_Disabilities	Presence of learning disabilities (Yes, No).

17	Parental_Education_Level	Highest education level of parents (High School, College, Postgraduate).
18	Distance_from_Home	Distance from home to school (Near, Moderate, Far).
19	Gender	Gender of the student (Male, Female).
20	Exam_Score	Final exam score.

EDA AND DATA VISUALIZATION

As part of our EDA, we cleaned the dataset by removing rows with missing values and correcting invalid entries in categorical variables. The key aim of this phase was to understand the distribution of variables and their initial relationships with the exam scores.

Key EDA Visualizations

Pairwise plot: Displaying the relationship strength between numerical variables.

Scatter Plot: Attendance vs. Exam Score showing positive correlation despite a few outliers.

Scatter Plot: Hours Studied vs. Exam Score with a similar pattern of correlation and the presence of outliers.

Boxplots: Used to visualize the distribution of categorical factors like Parental Involvement and Access to Resources.

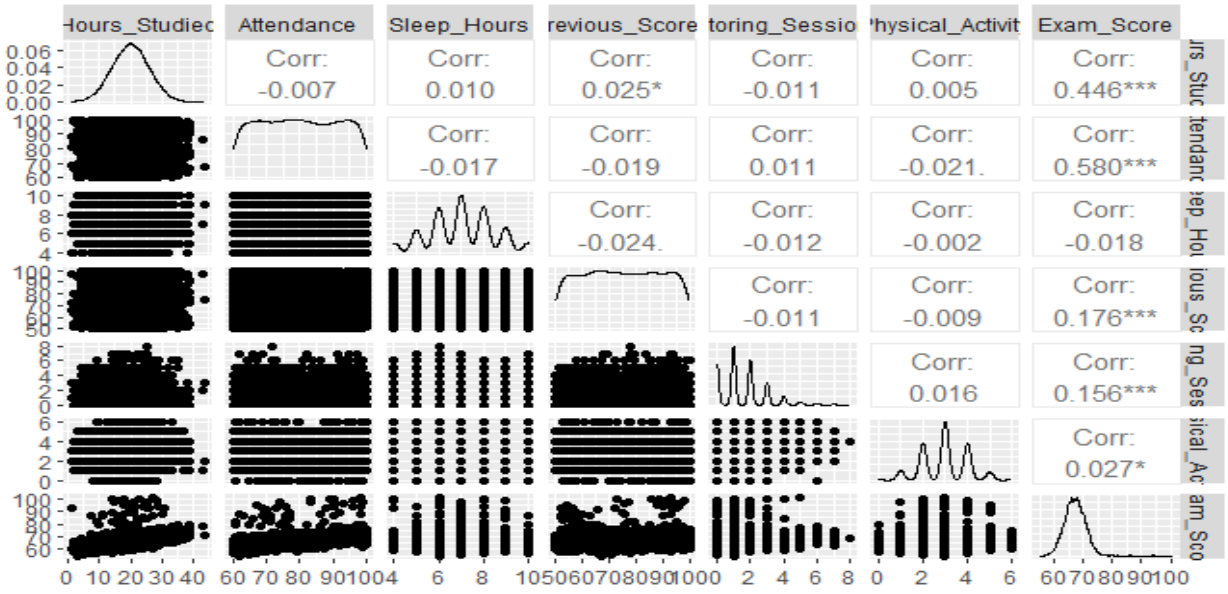


Figure 1: Pairwise plot on the numeric variables

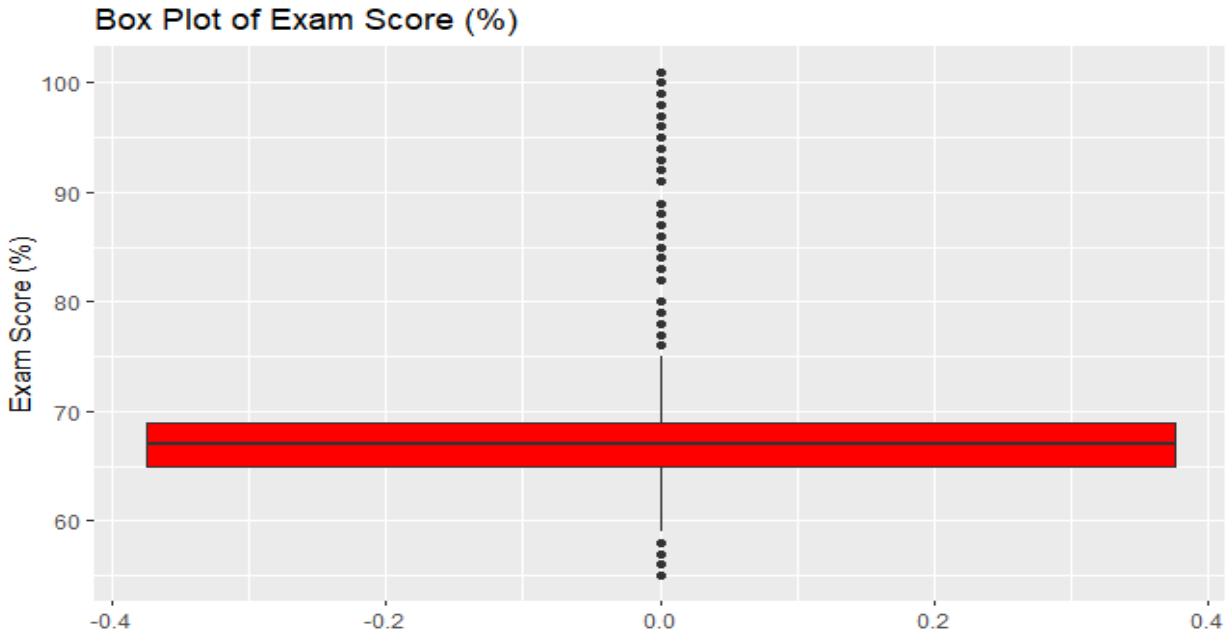


Figure 2: Box plot of Exam Score (%)

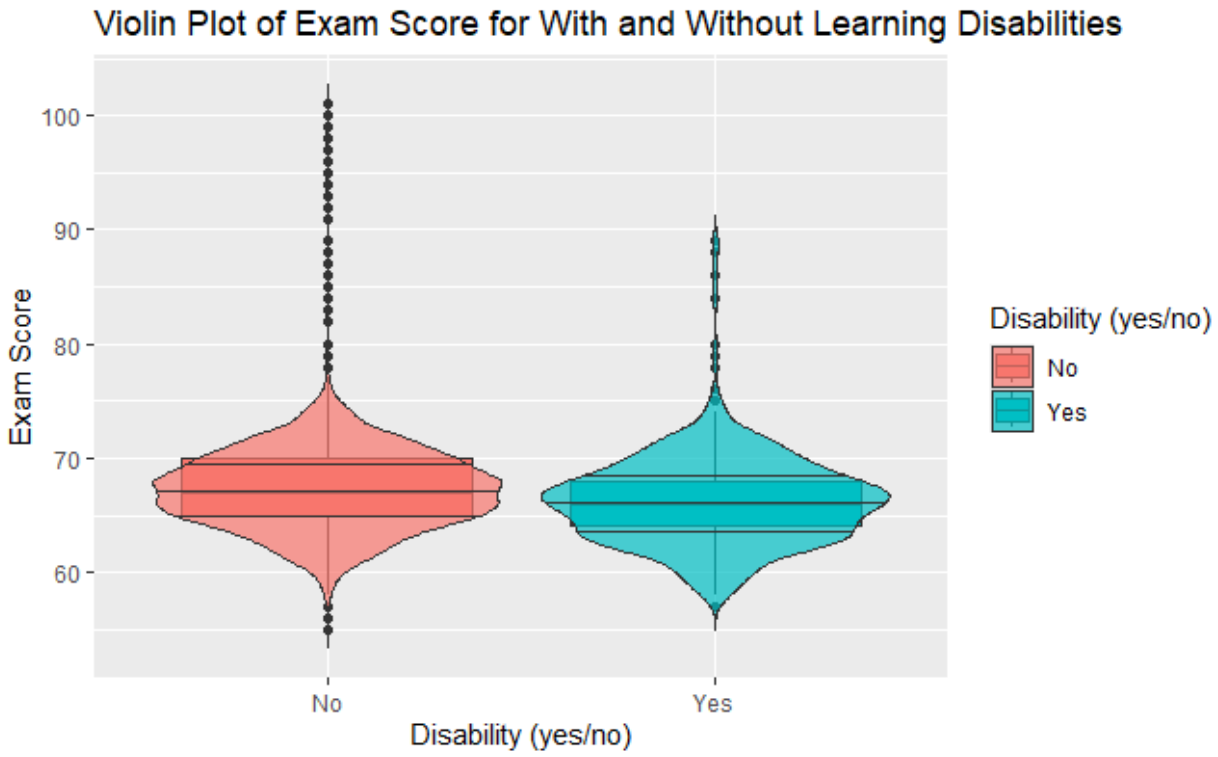


Figure 3: Violin plot- showing distribution of exam scores based on Disability

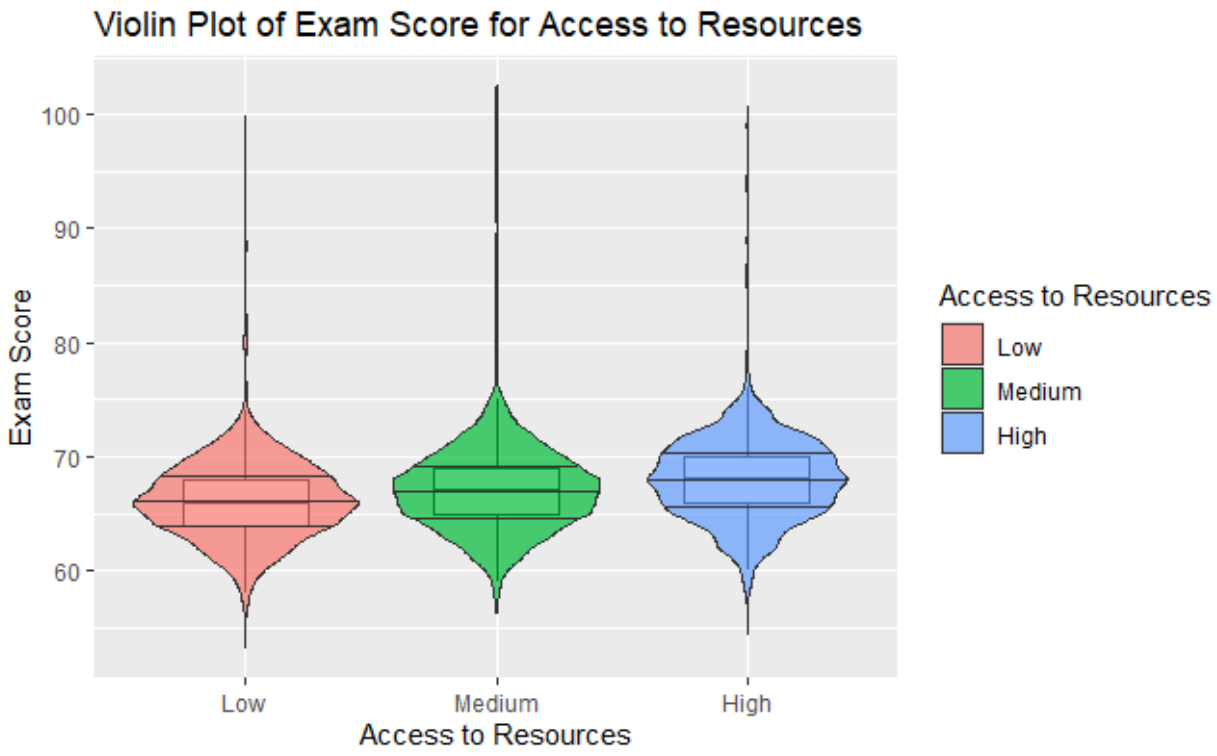


Figure 4: Violin plot- showing distribution of exam scores based on Access to Resources

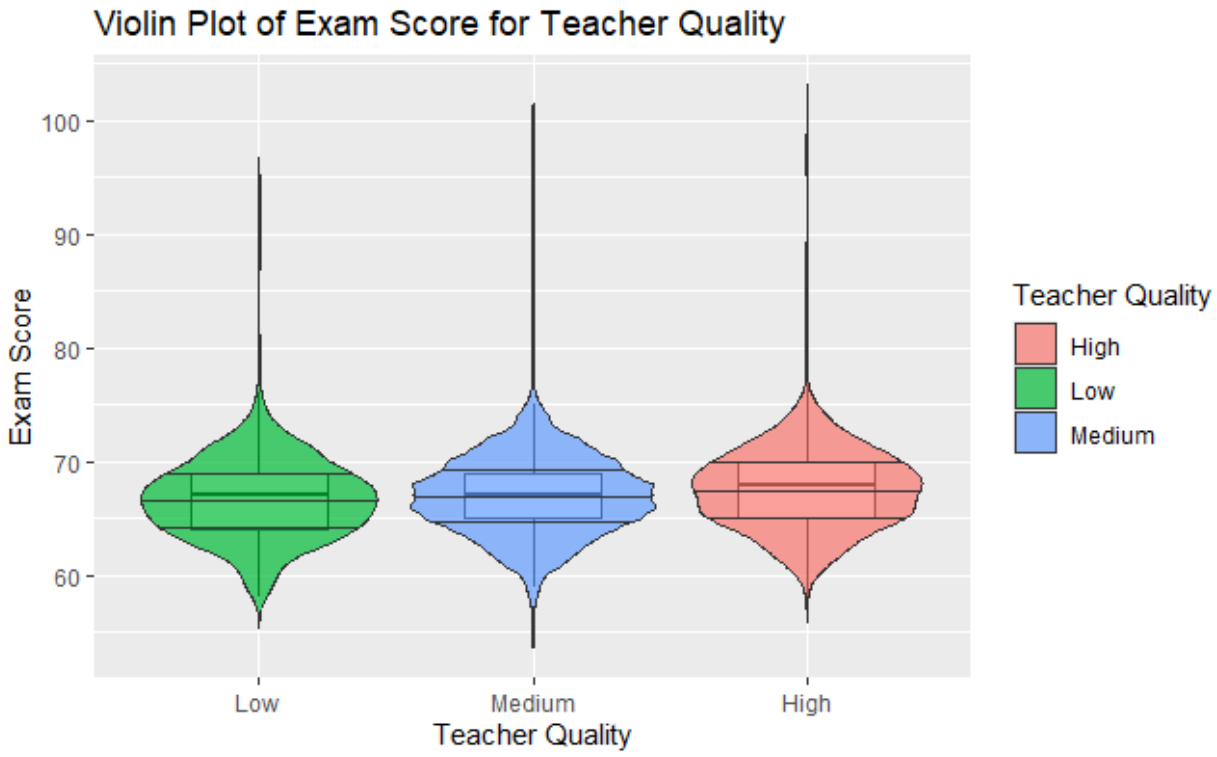


Figure 5: Violin plot- showing distribution of exam scores based on Teaching Quality

Key insights from EDA

- Correlation analysis revealed that Hours Studied and Attendance had the strongest positive correlations with the final exam score. This formed the basis for further analysis we carried out in the next section.
- The distribution of Exam Scores in figure 2 showed a slightly right-skewed distribution. We will gauge student performance by analyzing the resulting Exam Scores, taking into account all the factors in our dataset. The median Exam Score is 67%, while the mean is slightly higher at 67.25%. Additionally, the standard deviation of the scores is 3.91%, indicating the degree of variability around the mean.
- The Violin plots displays the relationship between the exam scores and the categorical variables considered to have an impact on exam scores. Figure 3 shows a slightly better score for those without learning disabilities. Both Figures 4 and 5 show a slight increase moving from low to high.

Note: the scatter plots for Exam Score vs. Attendance and Exam Score vs. Hours Studied are excluded from the EDA because they are included in" Hypothesis Test Results" section to avoid redundancy for the reader.

HYPOTHESIS TEST RESULTS

Hypothesis testing was conducted to statistically evaluate the significance of the relationship between key explanatory variables and the response variable (exam score).

Bootstrapping the Mean Exam Score for Without and With Disability

Hypothesis Statements:

H_0: $\mu_{\text{Exam Score – No Learning Disability}} \leq \mu_{\text{Exam Score – Learning Disability}}$

H_A: $\mu_{\text{Exam Score – No Learning Disability}} > \mu_{\text{Exam Score – Learning Disability}}$

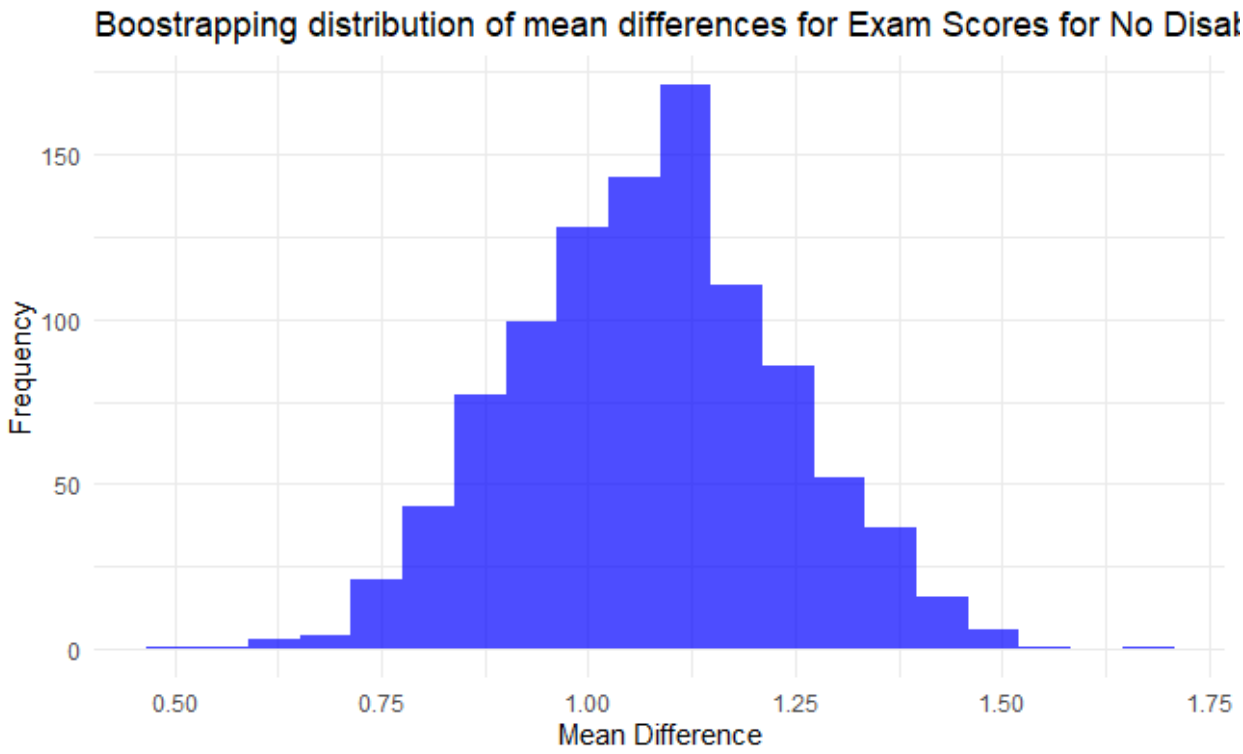


Figure 6: Bootstrap distribution of mean difference exam scores between no disability and disability

The graph of Bootstrapping distribution of mean differences for Exam Scores for No Disability and Disability shows us the outcome of our bootstrap test. The 95% Confidence Interval for the Bootstrapped Mean Difference for Exam Score for Individuals Without and With Learning Disabilities is 0.7638096 to 1.396271. Because of the 95% confidence interval (the range where the population mean difference occurs), which does not contain the value 0. We are enabled to state that there is a significant difference in the mean between the two populations, therefore we can reject the null hypothesis.

Bootstrapping the Mean Exam Score for Males and Females

H_0: $\mu_{\text{Exam Score\_Female}} = \mu_{\text{Exam Score\_Male}}$

H_A: $\mu_{\text{Exam Score \_Female}} \neq \mu_{\text{Exam Score\_Male}}$

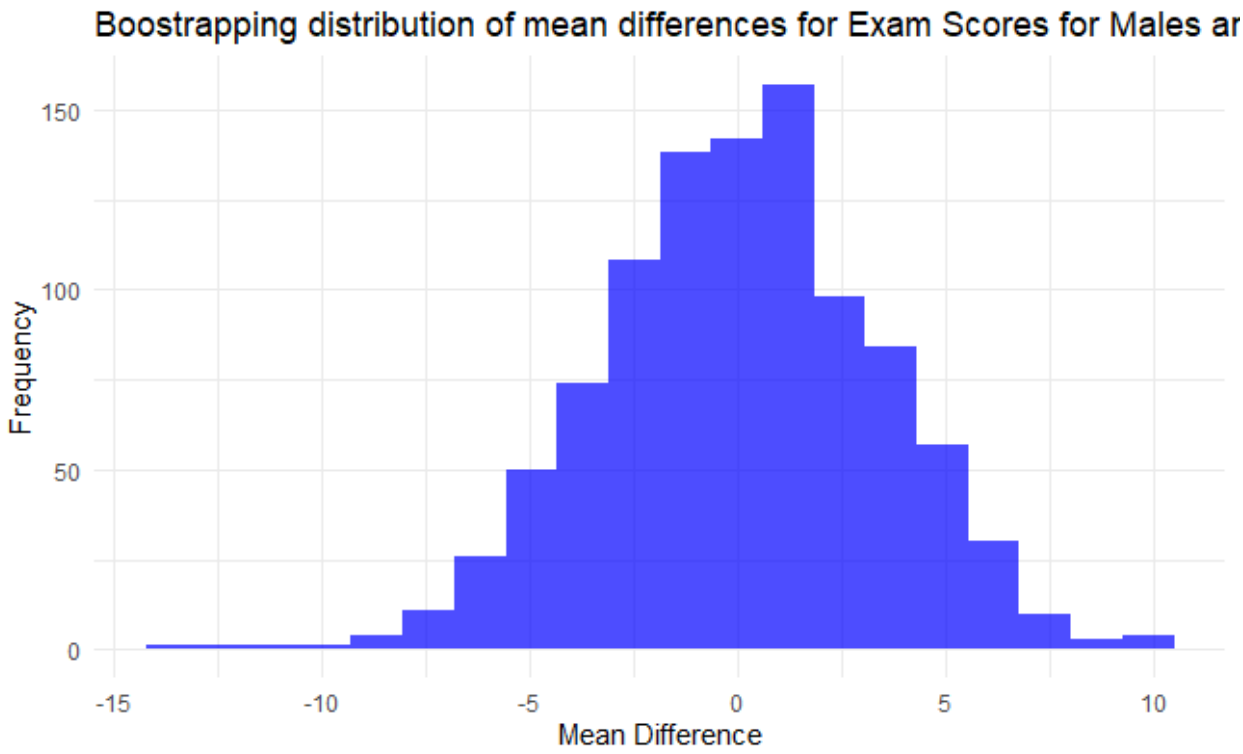


Figure 7: Bootstrap distribution of mean difference exam scores between male and female

The graph of Bootstrapping distribution of mean differences for Exam Scores for Females and Males shows us the outcome of our bootstrap test. The 95% Confidence Interval for the Bootstrapped Mean Difference for Exam Score for Males and Females is -6.418657 to 6.373881. Because of the 95% confidence interval (the range where the population mean difference occurs), which does the value 0. We are enabled to state that there is a no significant difference in the mean between the two populations, therefore we failt to reject the null hypothesis.

From our Exploratory and Data Analysis of numerical variables, we found that the variables, Attendance and Hours Studied to have the highest correlation with our main determinant of student performance, Exam Scores. Naturally, a linear regression test was imposed on both factors find a relationship. The following are shown.

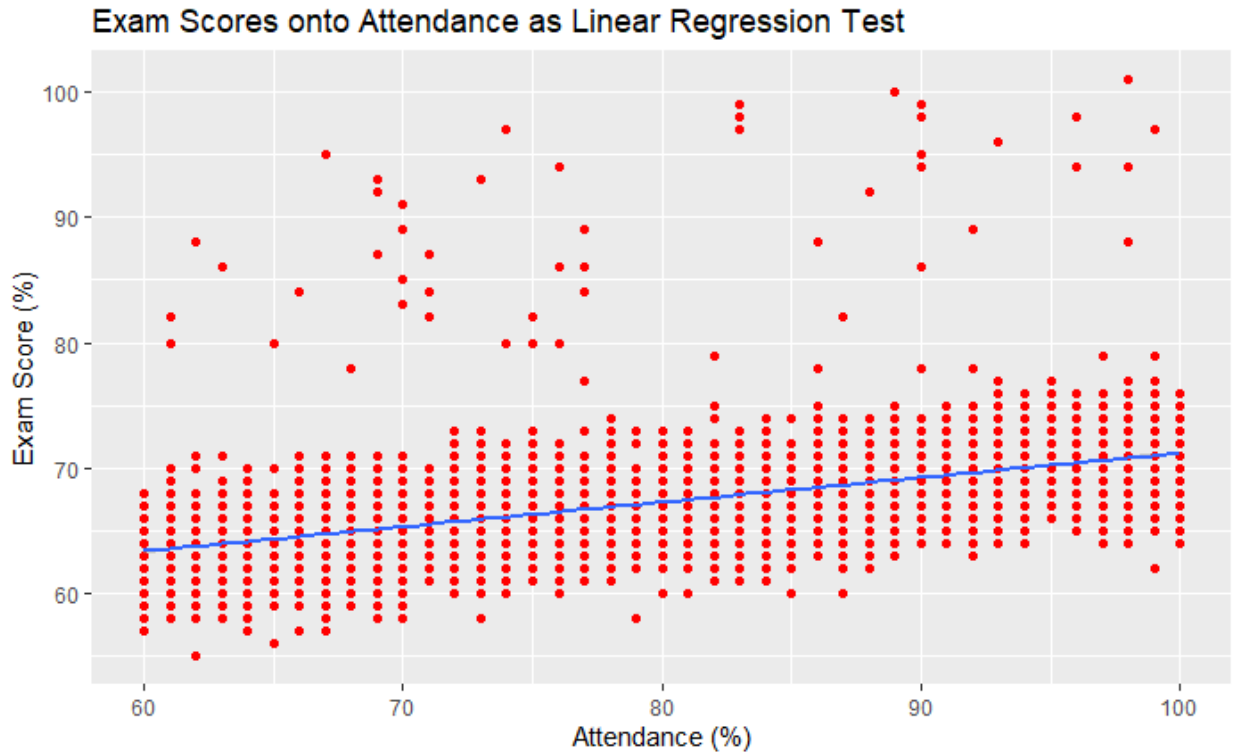
Linear Regression For Exam Scores Vs. Attendance

H_0: $\beta_1 = 0$

the slope of the linear model of Exam Scores onto Attendance is not significantly different to 0

H_0: $\beta_1 \neq 0$

the slope of the linear model of Exam Scores onto Attendance is significantly different to 0



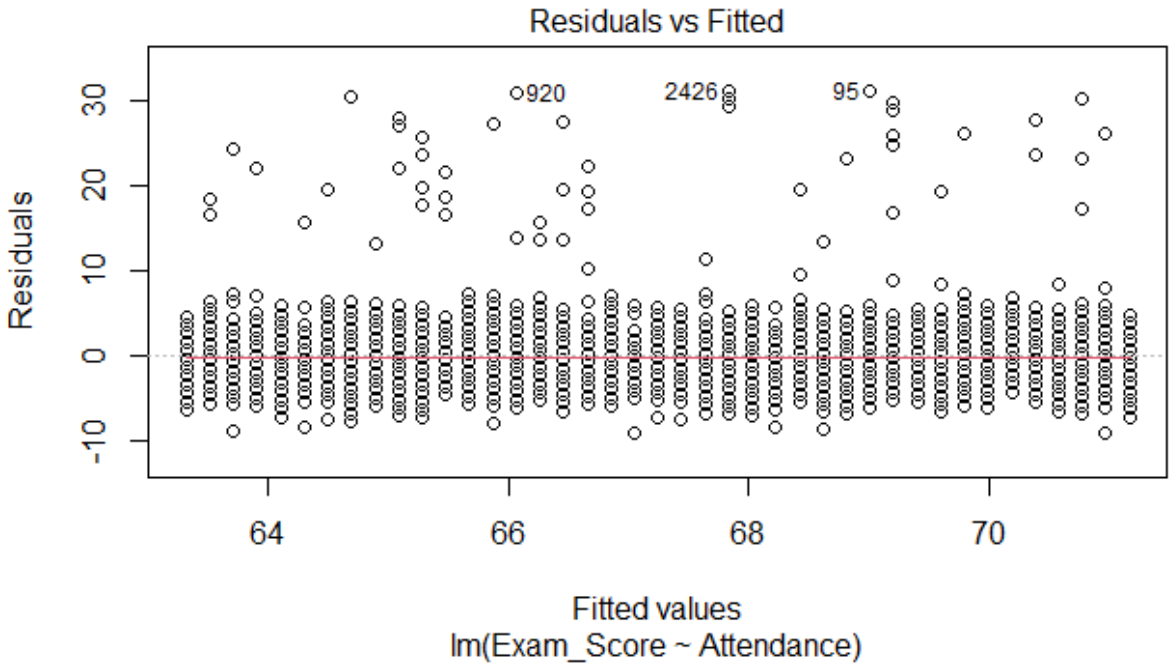
Output:

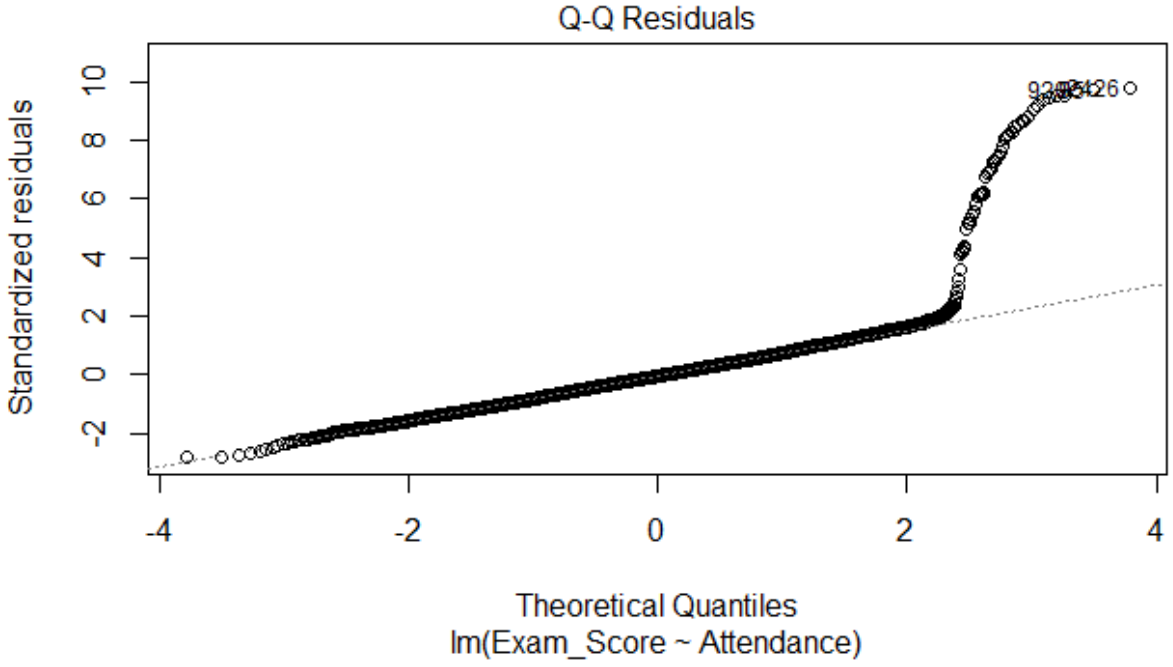
```
Call:
lm(formula = Exam_Score ~ Attendance, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0495 -1.8159 -0.1897  1.5020 31.1655

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 51.544540   0.277137  185.99   <2e-16 ***
Attendance   0.196265   0.003428   57.26   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.182 on 6462 degrees of freedom
Multiple R-squared:  0.3366,    Adjusted R-squared:  0.3365
F-statistic:  3278 on 1 and 6462 DF,  p-value: < 2.2e-16
```

Conclusion:

We can find that the p-value of this linear model is less than 2E-16, meaning that the slope is significantly different than positive linear relationship between exam scores and attendance, with $\beta_1 = 0.196265$ with a y-intercept of 51.544 and a R^2 value of 0.3366.

Assessing the Model:





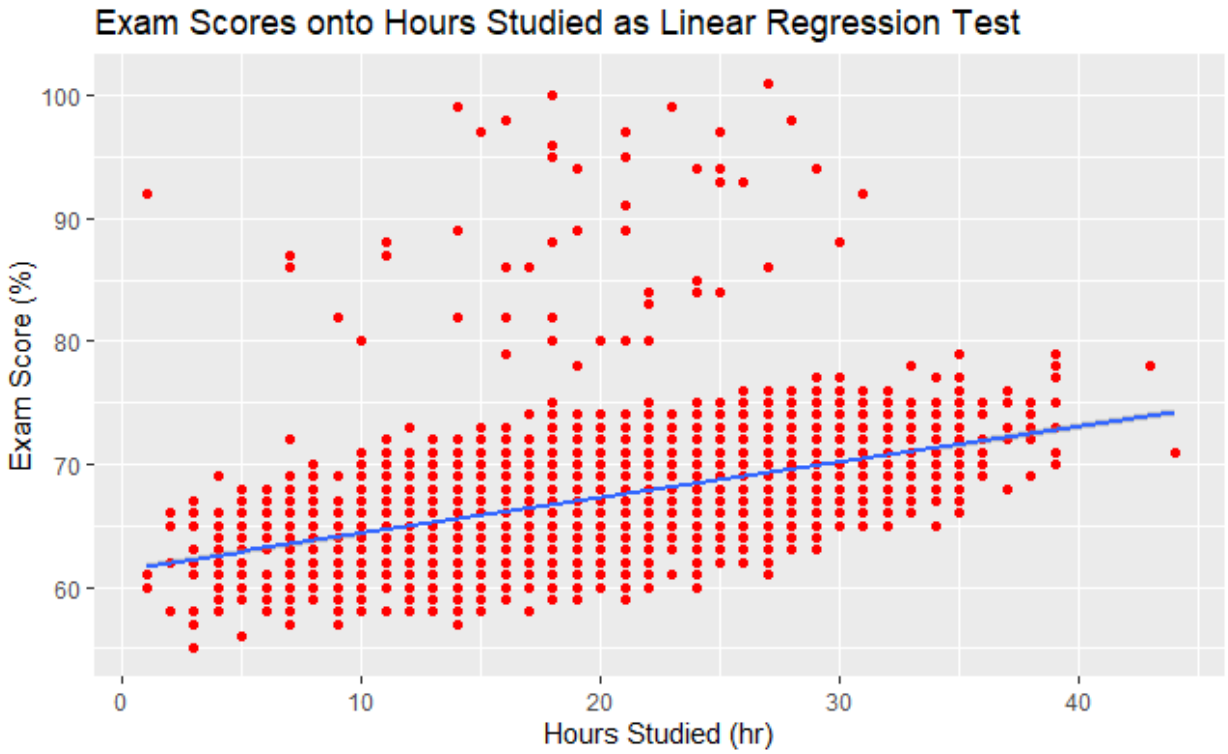
The coefficients of the linear regression model suggest that for every extra (%) attended to class, individuals can expect a 0.196265 % increase in their exam score. However, with no attendance (attendance (%) = 0), the data expects an individual to score 51.544% in their exam. However, the R^2 value of 0.3366 explains that the model does not linearly fit the data as well as it should (should be around 0.70 - 1.00).

However, we are seeing that the Residuals vs. Fitted plot does have points densed around 0, indicating that there are some points whose variation is not random, and stray away from the pattern. A similar issue arises with the QQ plot, where some data points do not follow normality.

Linear Regression Test for Exam Score Vs. Hours Studied

H_0: $\beta_1 = 0$
the slope of the linear model of Exam Scores onto Hours Studied is not significantly different to 0

H_0: $\beta_1 \neq 0$
the slope of the linear model of Exam Scores onto Hours Studied is significantly different to 0



Output:

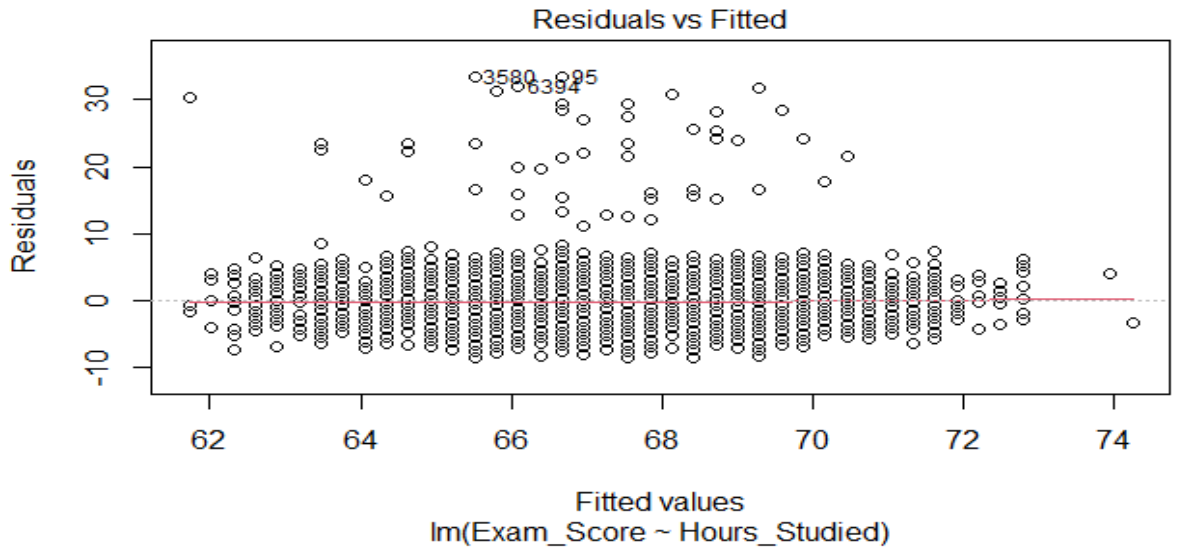
```
Call:
lm(formula = Exam_Score ~ Hours_Studied, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.545  -2.254  -0.127   2.037  33.492
```

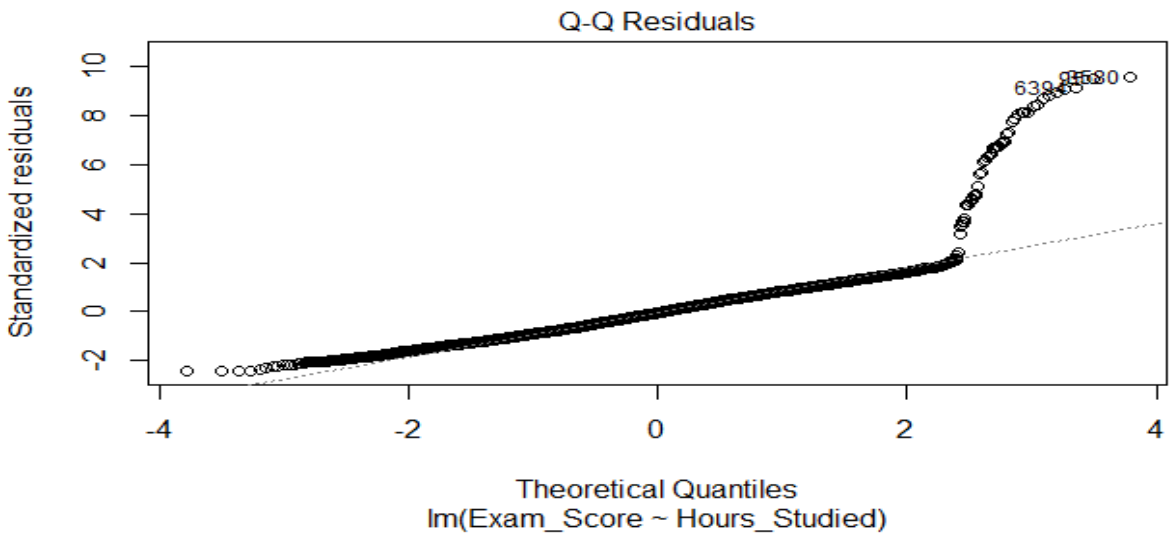
```
Coefficients:
              Estimate       Std.       Error       t       value       Pr(>|t|)
(Intercept)          61.434819          0.151519       405.46          <2e-16    ***
Hours_Studied       0.290971          0.007262          40.06          <2e-16    ***
---
Signif.   codes:     0   ‘***’   0.001   ‘**’   0.01   ‘*’   0.05   ‘.’   0.1   ‘   ’   1

Residual    standard    error:    3.497    on    6462    degrees    of    freedom
Multiple R-squared:  0.199,    Adjusted        R-squared:                0.1989
F-statistic:  1605 on 1 and 6462 DF,  p-value: < 2.2e-16
```

Conclusion:

We can find that the p-value of this linear model is less than 2E-16, meaning that the slope is significantly different than positive linear relationship between exam scores and hours studied, with $\beta_1 = 0.290971$ with a y-intercept of 61.434819 and a R^2 value of 0.199.

Assessing the Model:





The coefficients of the linear regression model suggest that for every extra hour studied, individuals can expect a 0.290971 % increase in their exam score. However, with no hours studied (Hours Studied = 0), the data expects an individual to score 61.435% in their exam. However, the R^2 value of 0.199 explains that the model does not linearly fit the data as well as it should (should be around 0.70 - 1.00).

However, we are seeing that the Residuals vs. Fitted plot does have points dense around 0, indicating that there are some points whose variation is not random, and stray away from the pattern. A similar issue arises with the QQ plot, where some data points do not follow normality.

Multilinear Regression

Initial Multilinear Regression Test with Uncleaned Data

The model was expanded to include multiple explanatory variables. This was conducted as an attempt to identify the highest contributors towards Exam Score via a linear regression test. Therefore, it included categorical variable, which were converted into numerical variables.

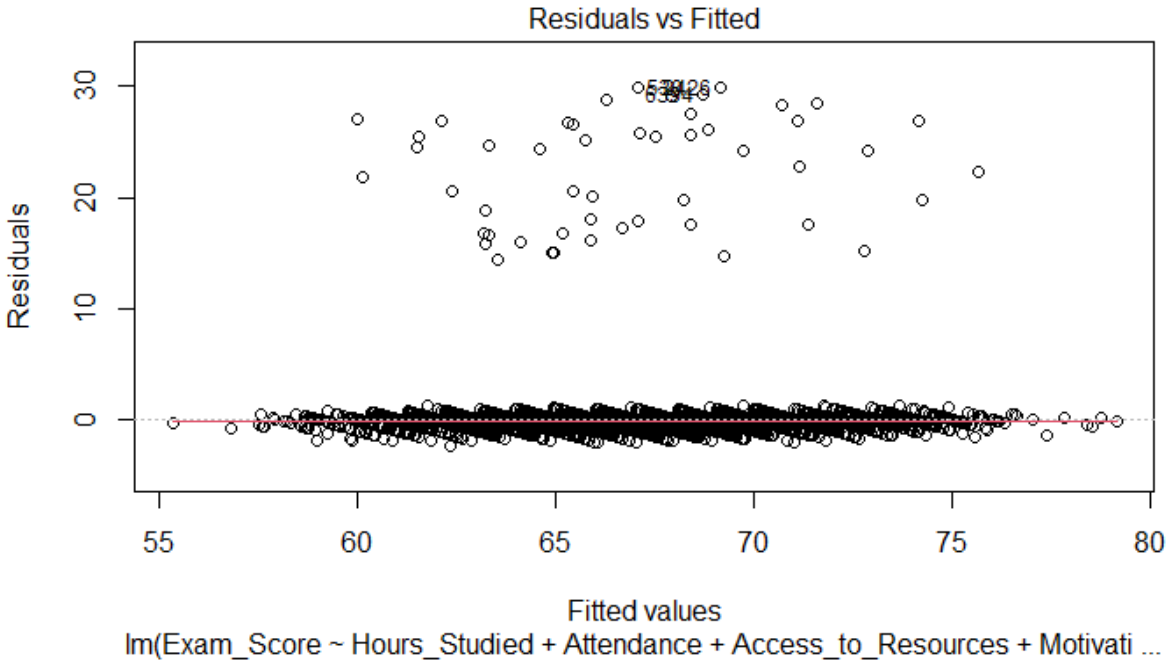
Output:

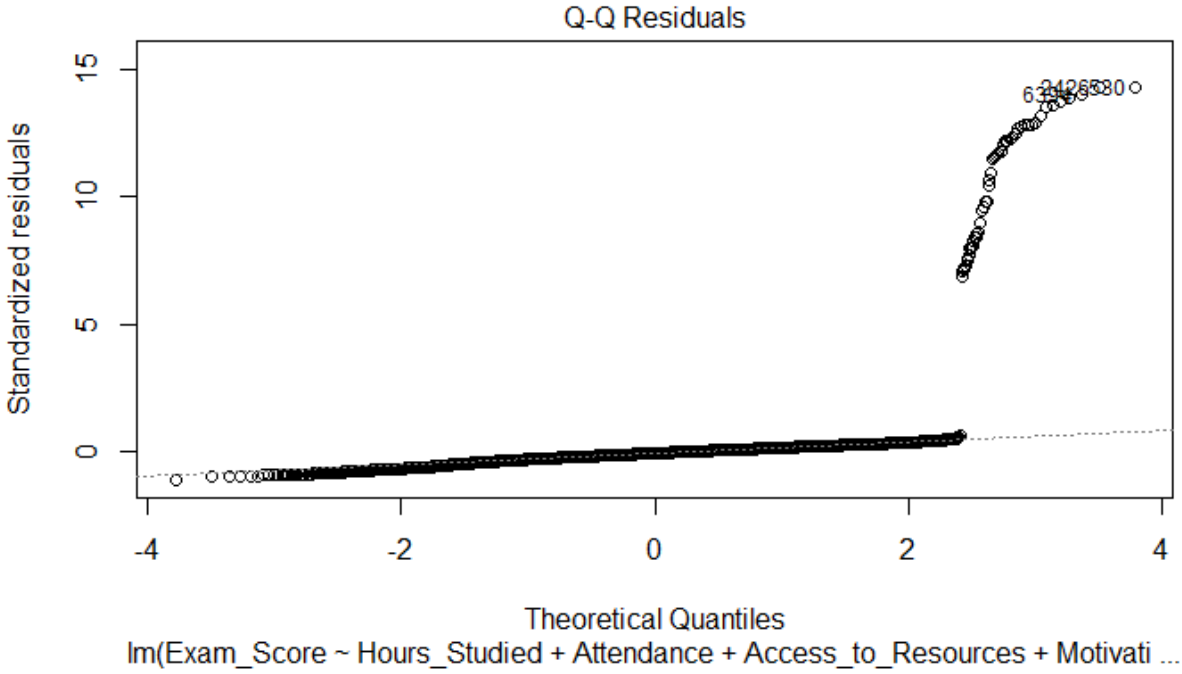
```
Call:
lm(formula = Exam_Score ~ Hours_Studied + Attendance + Access_to_Resources +
    Motivation_Level + Parental_Involvement + Teacher_Quality +
    Parental_Education_Level + Previous_Scores + Tutoring_Sessions +
    Internet_Access + Family_Income + Extracurricular_Activities +
    Peer_Influence + Distance_from_Home, data = data)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.3530 -0.4636 -0.1315  0.1796 29.9001
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 27.891638   0.397231  70.215   <2e-16 ***
Hours_Studied                0.295516   0.004382  67.434   <2e-16 ***
Attendance                   0.198872   0.002272  87.542   <2e-16 ***
Access_to_Resources          1.027570   0.037561  27.357   <2e-16 ***
Motivation_Level             0.529606   0.037703  14.047   <2e-16 ***
Parental_Involvement         0.992645   0.037778  26.275   <2e-16 ***
Teacher_Quality              0.529039   0.043626  12.127   <2e-16 ***
Parental_Education_Level     0.489244   0.033596  14.562   <2e-16 ***
Previous_Scores              0.048757   0.001823  26.750   <2e-16 ***
Tutoring_Sessions            0.499823   0.021251  23.520   <2e-16 ***
Internet_Access              0.905720   0.098955   9.153   <2e-16 ***
Family_Income                0.523750   0.035270  14.850   <2e-16 ***
Extracurricular_Activities   0.563491   0.053470  10.538   <2e-16 ***
Peer_Influence               0.515492   0.034679  14.865   <2e-16 ***
Distance_from_Home           0.478135   0.039127  12.220   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.093 on 6363 degrees of freedom
  (86 observations deleted due to missingness)
Multiple R-squared:  0.7147,    Adjusted R-squared:  0.7141
F-statistic:  1139 on 14 and 6363 DF,  p-value: < 2.2e-16
```

While it improved the R^2 value, the model risked overfitting. Although, this analysis let us to the most significant impact were Hours Studied, Attendance, Previous Exam Scores, Access to resources, and Parental Involvement. All values included in this regression model found to have β_1 to be significantly different than 0, indicating the coefficient provided does contribute to Exam Score in that magnitude within our dataset.

Assessing the model:

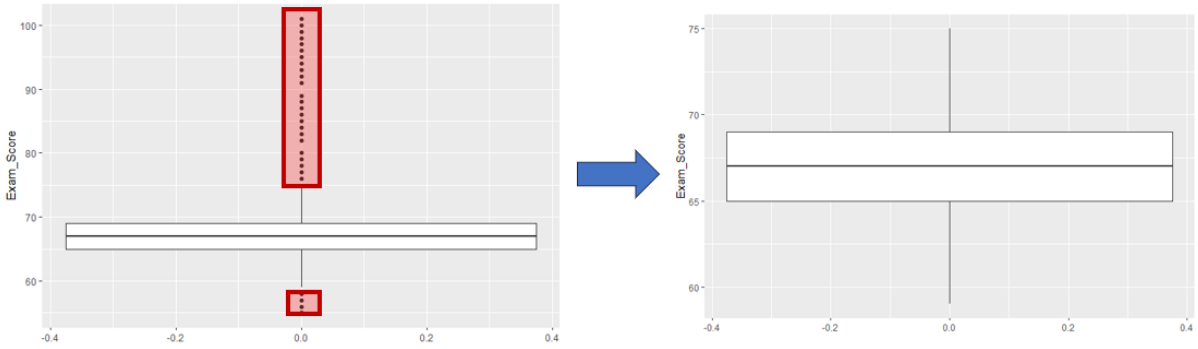




Seeing these Residual vs. Fitted Plots, we see that there are clear outliers present in our data, that makes these coefficients unusable because our linear model does not fit the data. Therefore, we must take steps

Cleaning the Data

The outlier We used the IQR function to remove outliers from the exam scores. Following this cleaning step; the resulting distribution



We can see that from the original data (fig. 2 also) that we removed the outliers (seen in red on the left plot) that the Exam Score distribution was cleaned and no longer contains outliers.

Final Multilinear Regression Test with Cleaned Data

Call:

lm(formula = Exam_Score ~ Hours_Studied + Attendance + Access_to_Resources +

Parental_Involvement + Previous_Scores, data = data_clean)

Residuals:

Min 1Q Median 3Q Max

-4.3450 -0.8161 -0.0150 0.8122 4.8344

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.675225	0.162734	231.51	<2e-16 ***
Hours_Studied	0.291862	0.002611	111.79	<2e-16 ***
Attendance	0.199186	0.001338	148.89	<2e-16 ***
Access_to_Resources	0.962203	0.022068	43.60	<2e-16 ***

Parental_Involvement 0.982198 0.022171 44.30 <2e-16 ***

Previous_Scores 0.047461 0.001070 44.36 <2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.226 on 6355 degrees of freedom

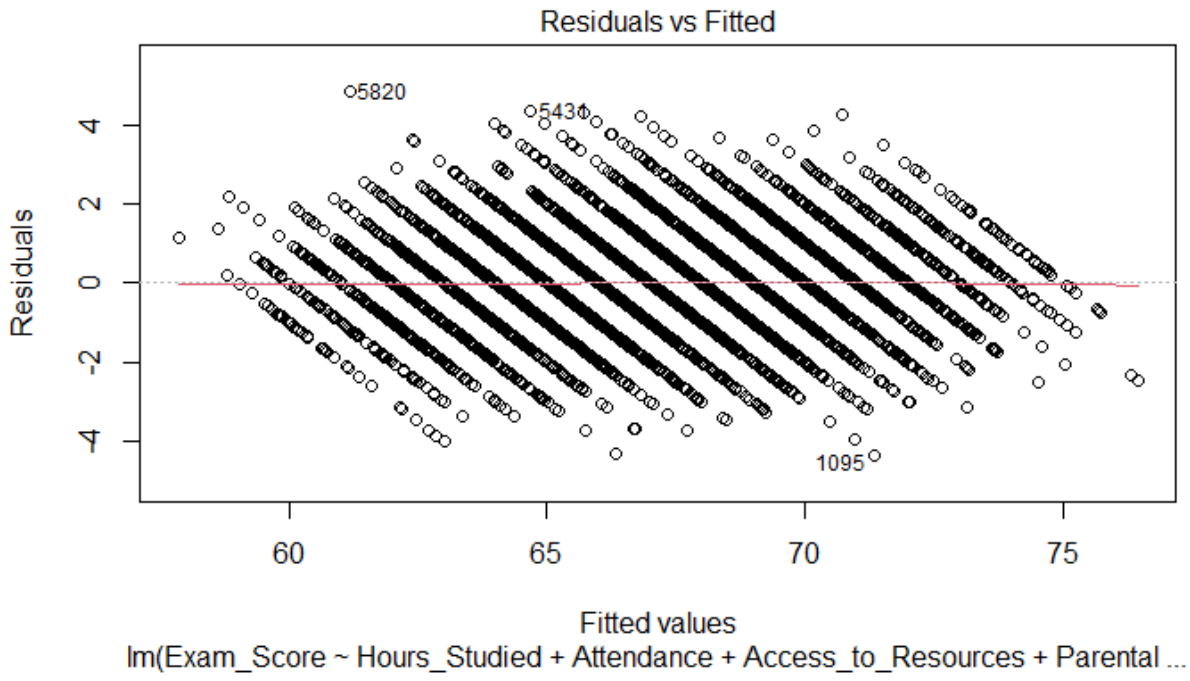
Multiple R-squared: 0.8581, Adjusted R-squared: 0.858

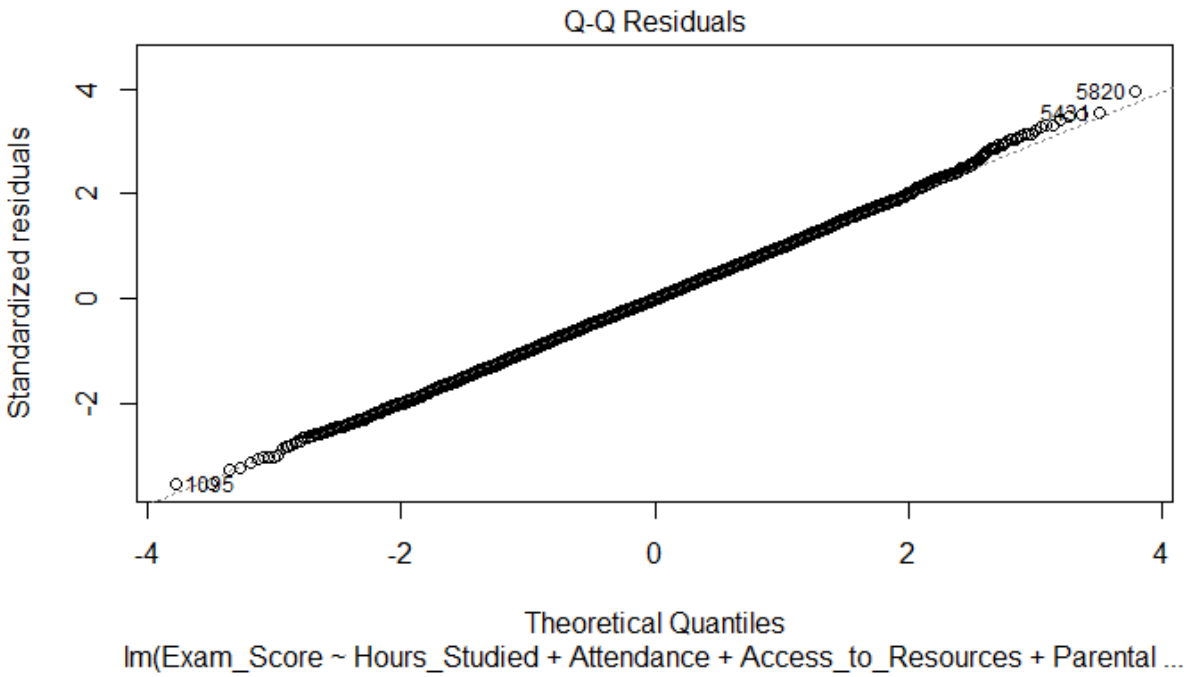
F-statistic: 7689 on 5 and 6355 DF, p-value: < 2.2e-16

Analysis:

Analysis of the output for the clean data indicate that the R^2 has shot up to 0.8581. This means that the model determines the proportion of variance in the Exam Score that can be explained by the independent variables in this multilinear regression. Additionally, we still find significance on the slope of the graphs.

Assessment:





We find that this cleaned data creates a much better model for us to create inferences from for the coefficients produced. We can find that normal distribution is followed by the entire data-set as well as the error is densely packed around 0 within the residuals vs. fitted plot.

CONCLUSION AND FUTURE STEPS

Key Findings

Following our extensive analysis of our dataset, we come to the following Top 5 contributors to student performance (gauged by Exam Score variable). We conclude that Hours Studied, Attendance (%), Access to Resources, Parental Involvement, and Previous Scores have statistically significant effects on the outcome of Exam Scores, with p-values less than 2×10-16. The overall model is significant and explains 85.8% of the variance in the Exam Score. To quantify their contribution towards the resulting exam score, for every; extra hour studied, there is a 0.291% increase in the exam score, extra (%) Attendance, there is a 0.199% increase in the exam score, increase in access to resources (from low to medium to high), exam scores improve by 0.962%, additional parental involvement (from low to medium to high), there is a 0.982% increase in the exam score, extra (%) in the previous scores, there is a 0.047% increase in the exam score, given that all the other variables remain constant for each of these cases.

Limitations

Going back to our issue of outliers, those can perhaps be best explained by the existing of confounding variables not included in our dataset. The analysis did not include factors like

intelligence quotient or emotional quotient, which could enhance model accuracy and explain some of the "weird" trends we were facing. Additionally, categorical data such as; Parental Involvement and Access to Resources were treated categorically, which limited precision. We can enhance our model by curating data that tabulated them as numerical data, so it would be easier to quantify the changes to these explanatory variables towards the resulting response variable (Exam Score).

Future Steps

Future steps include collecting more detailed datasets for a nuanced approach in analyzing categorical data. For example, instead of "Parental Involvement" we can use "Hours Spent with Parents" or instead of "Access to Resources" we can use "Money Spent on Education". Apart from our focus on linear regression on Exam Score, we can also analyze what contribute towards our Top 5 hits. For example, is there a relationship between "Attendance" and "Motivation Level". These inter-relations between expalantory variables were not explored unfortunately due the project's timeline, but can provide even more insight to proper steps to take to maximum one's own student performance.

Practical Applications

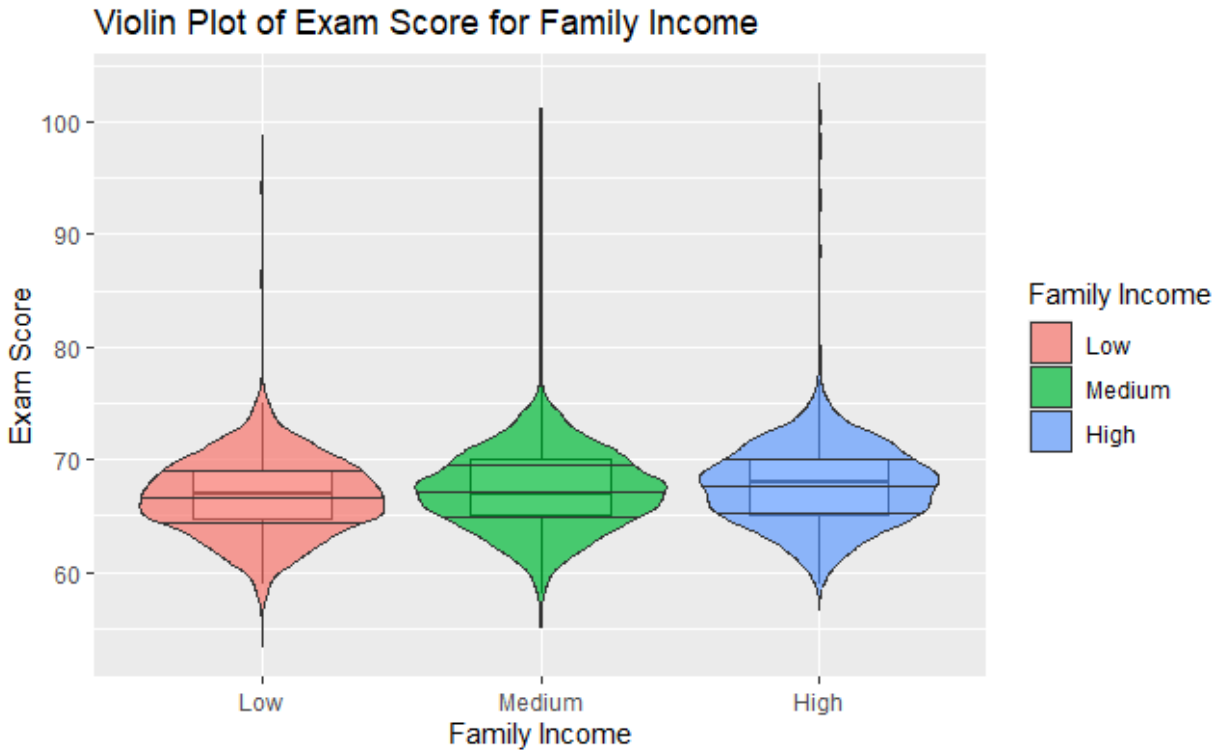
These insights can help educators and policymakers make data-driven decisions to improve student outcomes. As well, for some us being parents, we can take these steps to help our own children.

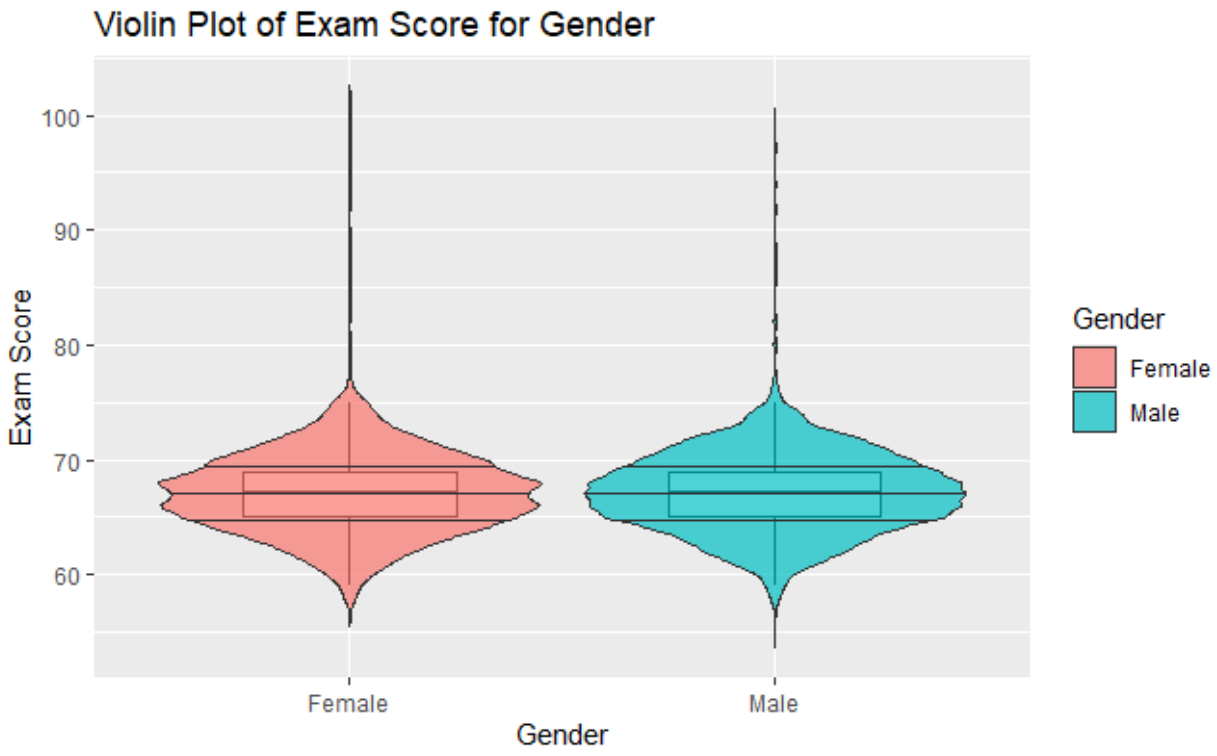
References

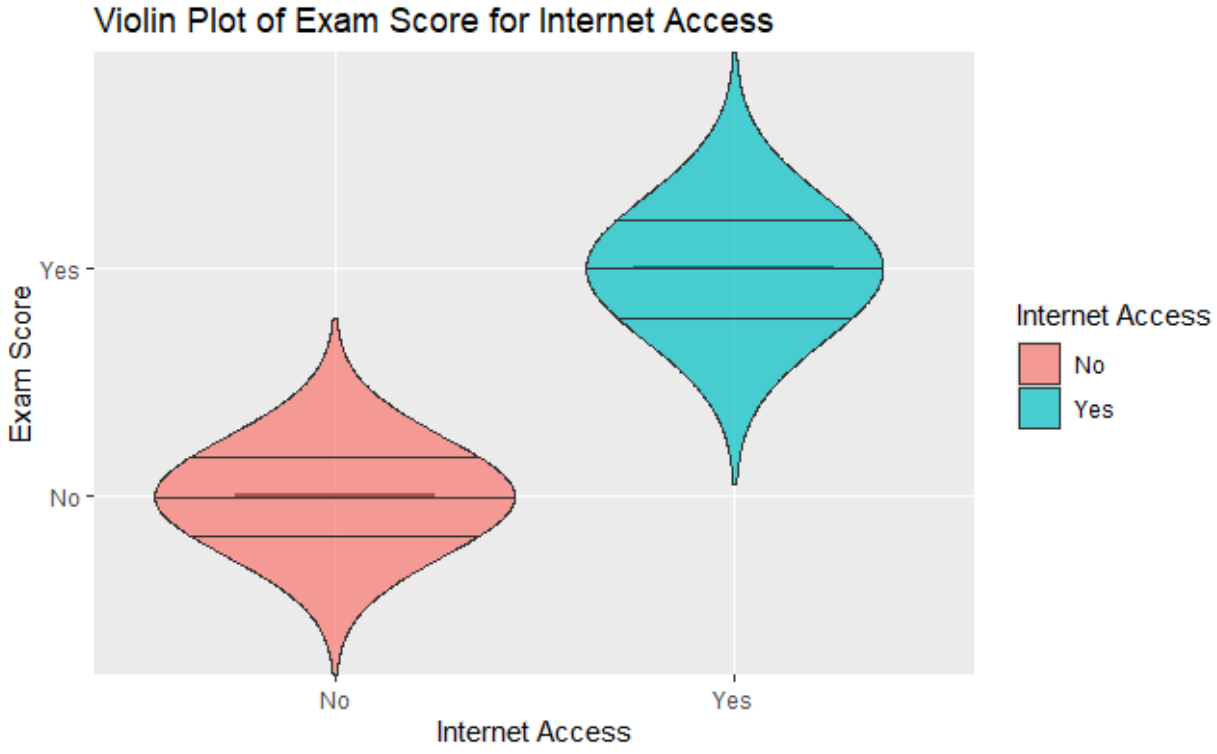
[1] C. Hanck and M. A. Hanck, *Introduction to Econometrics with R*. [Online]. Available: https://www.econometrics-with-r.org/6.2-tmrm.html. [Accessed: Oct. 14, 2024].

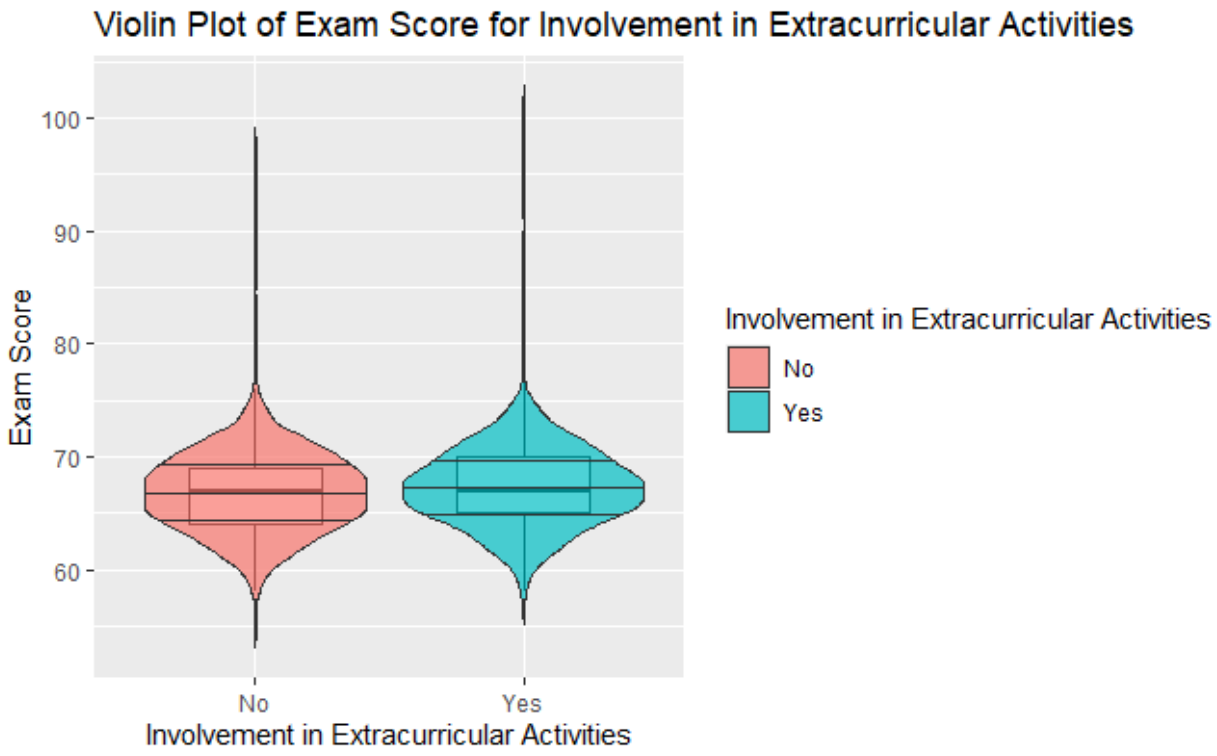
[2] L. Nguy, "Student Performance Factors," *Kaggle*, Aug. 2023. [Online]. Available: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors. [Accessed: Oct. 14, 2024].

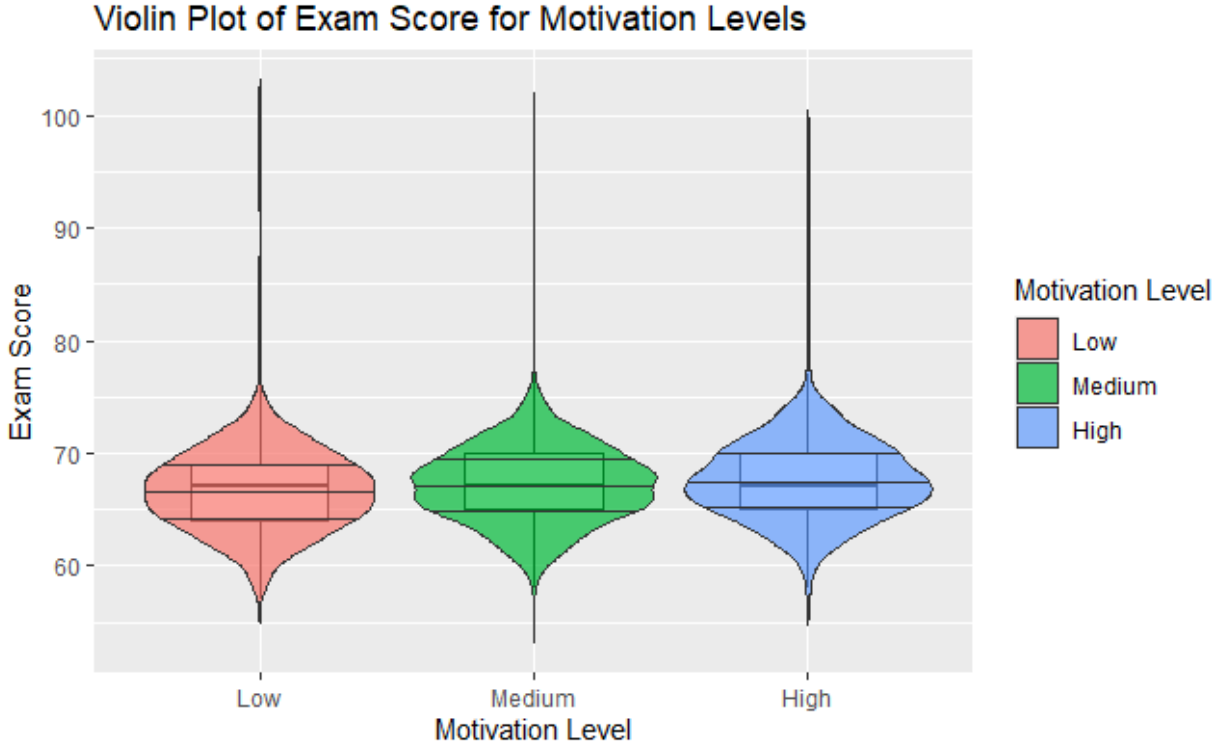
Supplementary

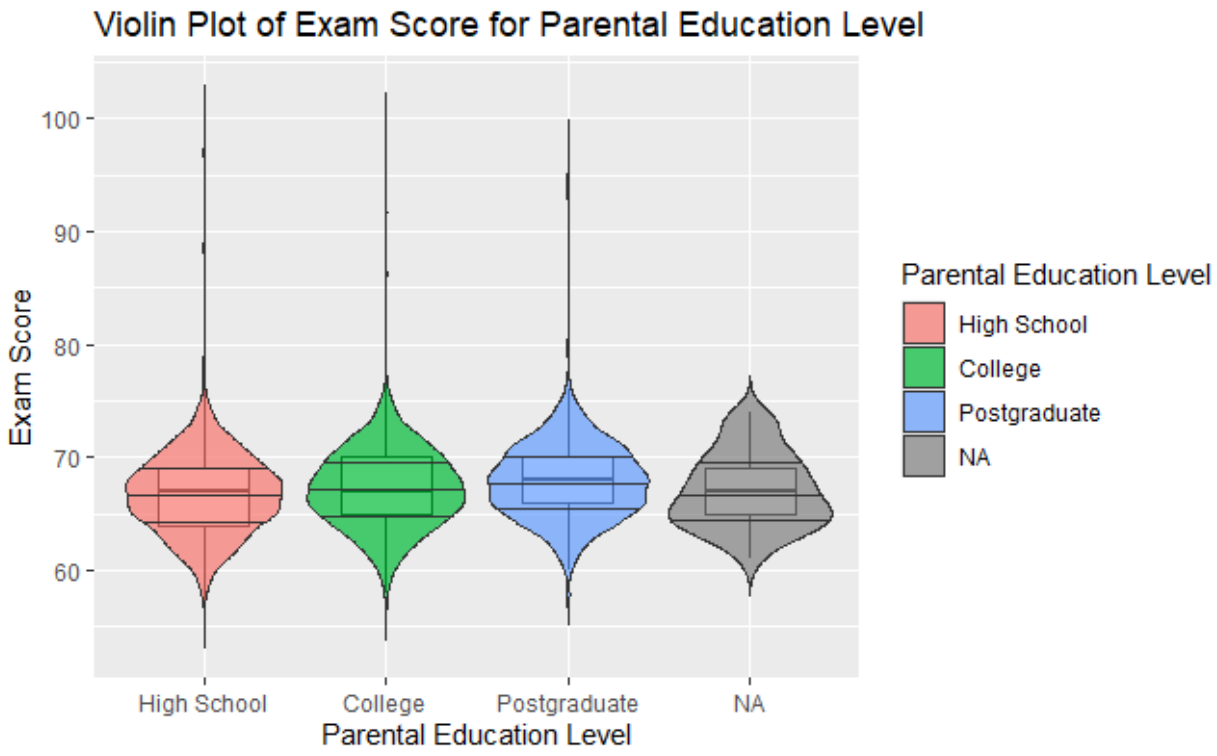


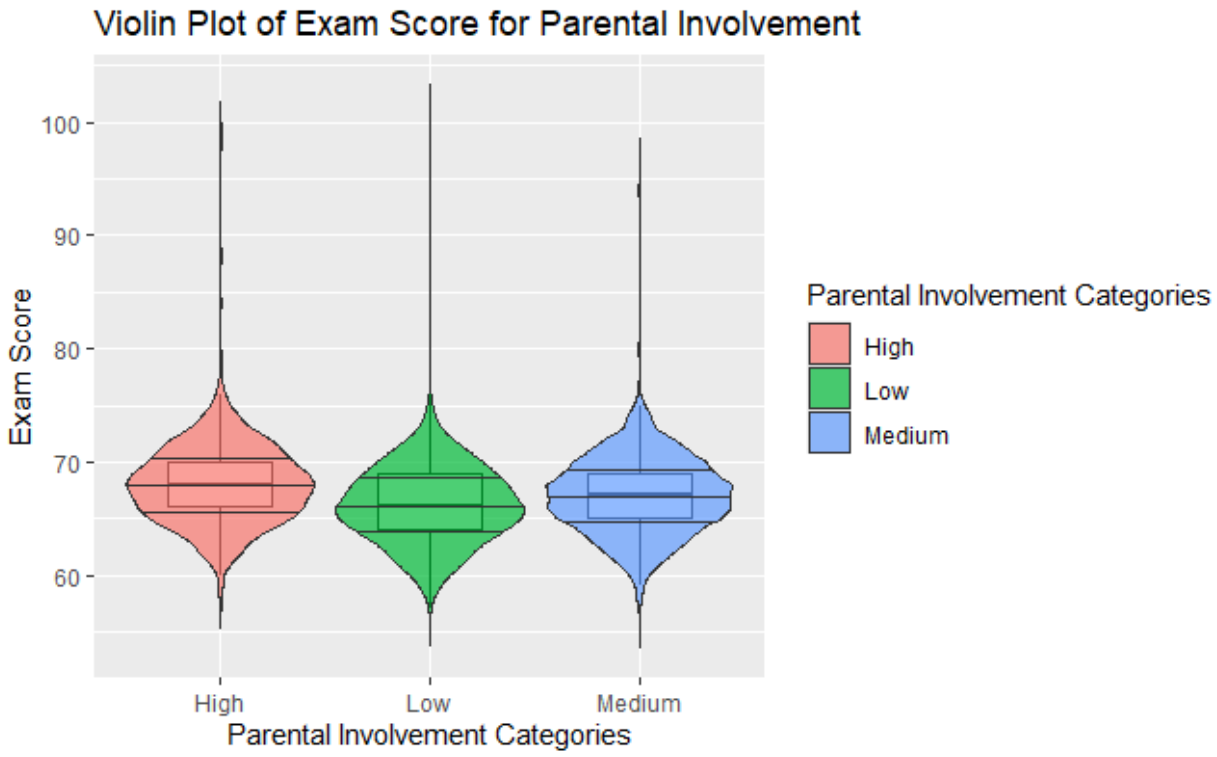

Violin Plot of Exam Score for Gender
Exam Score
100
90
80
70
60
Female
Male
Gender
Gender
Female
Male

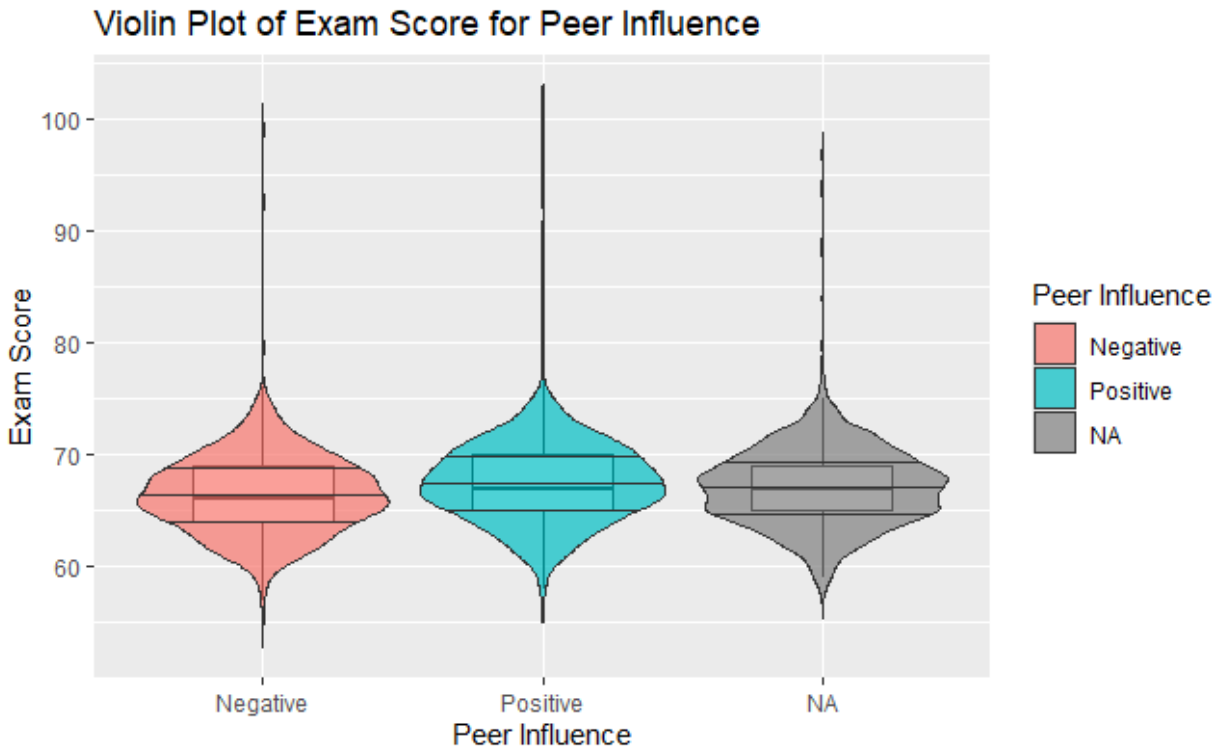

Violin Plot of Exam Score for Internet Access
Exam Score
Yes
No
No
Yes
Internet Access
Internet Access
No
Yes


Violin Plot of Exam Score for Involvement in Extracurricular Activities
Exam Score
100
90
80
70
60
No
Yes
Involvement in Extracurricular Activities
Involvement in Extracurricular Activities
No
Yes


Violin Plot of Exam Score for Motivation Levels
Exam Score
100
90
80
70
60
Low
Medium
High
Motivation Level
Motivation Level
Low
Medium
High


Violin Plot of Exam Score for Parental Education Level
Exam Score
100
90
80
70
60
High School
College
Postgraduate
NA
Parental Education Level
Parental Education Level
High School
College
Postgraduate
NA


Violin Plot of Exam Score for Parental Involvement
Exam Score
100
90
80
70
60
High
Low
Medium
Parental Involvement Categories
Parental Involvement Categories
High
Low
Medium


Violin Plot of Exam Score for Peer Influence
Exam Score
100
90
80
70
60
Negative
Positive
NA
Peer Influence
Peer Influence
Negative
Positive
NA