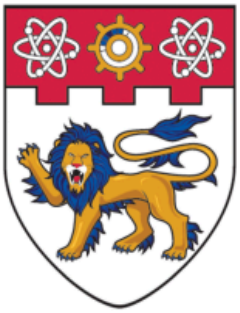


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

AY2024/2025 Semester 2
MH3511 Data Analysis in Computer
Group 13

Students' Performance in Examinations: Background, Preparation and Results

Matriculation Number	**Student Name**
U2340654A	Makayan Gwen Ashley Gravador
U2340894J	Tan Yong Kee
U2340227C	Tan Ranen
U2340547A	Chow Jia Hui
U2340638B	Valerie Ang Wei Ning

Content Page

1 Abstract

Examinations have long been an integral part of every student's academic journey. It has been highly debated whether examinations should be completely removed from the education systems to alleviate the overwhelming stress placed on students. In response to these concerns, Singapore's Ministry of Education has made numerous changes in the grading systems of national examinations in hopes of reducing the pressure on students (The Straits Times, 2020). In this project, our group explored a dataset of 500 students, intending to produce insights into any possible relationships between students' profiles and their final grades. Specifically, we want to assess factors such as the student's gender, weekly total number of study hours, attendance rate, past examination scores, family background, internet availability at home, and whether the student is involved in extracurricular activities, which will impact the final examination score received. Using various statistical techniques, our group uncovered meaningful relationships between these variables.

2 Introduction

Examinations can be considered as a milestone to test a student's understanding of the concepts taught in school. With many students taking final examinations at the end of the year to evaluate a year's worth of understanding, the availability of detailed data on students opens new doors for empirical analysis. As such, our group used the information in a dataset of 500 students to investigate the various factors that may affect a student's final examination score. Based on this dataset, we specifically aim to explore the following key and vital questions:

1. Is there a difference in the passing rates between male and female students?
2. Is there a trend in the passing rates among students with low, medium, and high study hours per week?
3. Does the number of study hours vary across different levels of parental education?
4. How do the students' final examination scores compare to their past examination scores?
5. How does parental education level influence students' performance in the final examination?
6. Does internet access at home relate to how much students improve their examination scores?
7. Is there a difference in final examination performance between students participating in extracurricular activities and those not?
8. Do students perform consistently across attendance rate, past and final examination scores?

3 Data Description

To analyse the relationships between different student profiles and final examination grades, we used a dataset titled "Student Performance Dataset" from an online machine learning and data science community, Kaggle. The dataset, "student_performance_dataset.csv", was uploaded on kaggle.com with the purpose of allowing users to perform predictive modeling, data visualisation, and educational analytics on it.

Before our group went on to analyse the dataset, we filtered and cleaned it to ensure data quality:

1) Any duplicated entries on the "***Student_ID***" column were removed to avoid any redundancy.
2) Our group went on the check and ensured that there were no rows or columns that contained "NA" before any analysis.

After filtering and cleaning up the dataset, 500 observations (students) with 10 variables were kept for analysis:

1. Unique identification for each student, ***Student_ID***
2. Gender of the student (Male/Female), ***Gender***
3. Average number of study hours per week, ***Study_Hours_per_Week***
4. Attendance percentage (50% to 100%), ***Attendance_Rate***
5. Average score of previous examinations (50 to 100), ***Past_Exam_Scores***
6. Education level of their parents (High School, Bachelors, Masters, PhD), ***Parental_Education_Level***
7. Internet accessibility at home (Yes/No), ***Internet_Access_at_Home***
8. Whether the student is involved in extracurricular activities (Yes/No), ***Extracurricular_Activities***
9. Final examination score of the student (50-100, integer values), ***Final_Exam_Score***
10. Student's status (Pass/Fail), ***Pass_Fail***

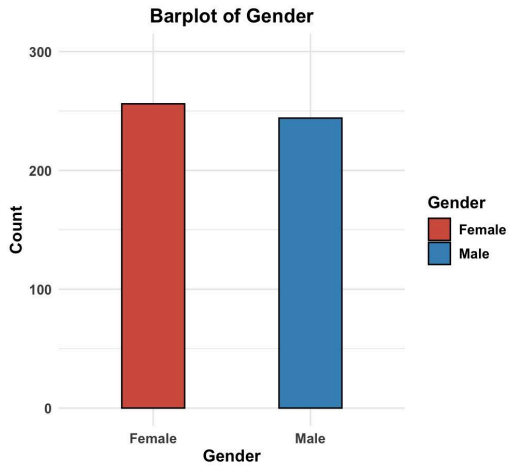
4 Description and Cleaning of Dataset

In this section, our group conducted a detailed assessment of every variable. Each variable was investigated individually for possible outliers or significantly skewed data. Although we applied log transformation to some of the variables in the data set, we found that the data still did not follow a normal distribution. Hence, our group has decided that we will not do any data transformation despite the skewed data.

4.1 Summary statistics for various variables

The histogram, the boxplot, and/or the transformation applied and the outliers removed from the variables are tabulated in the following subsections.

4.1.1 Gender of the student (Male/Female), *Gender*

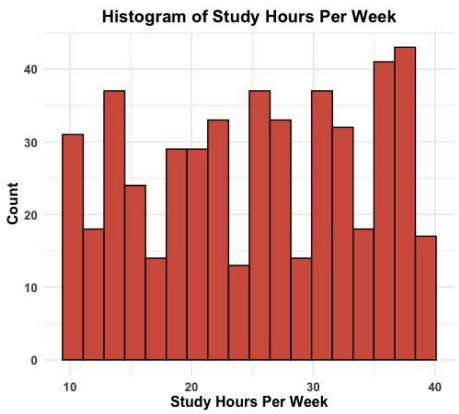
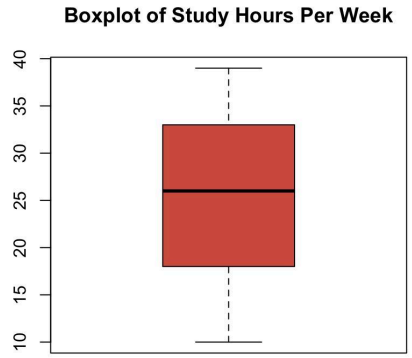


The variable is binary categorical, so no transformation is applied.

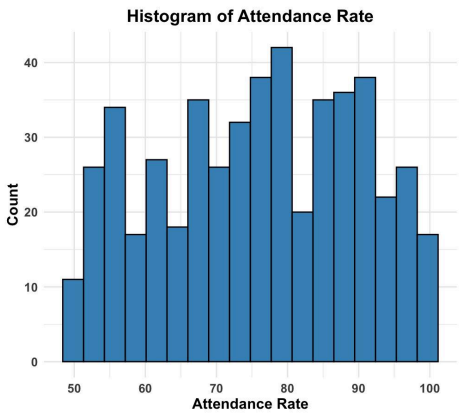
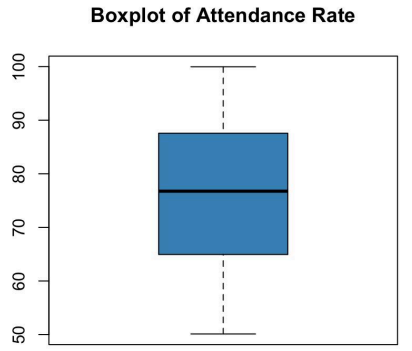
No outlying values of *Gender* are removed.

There was a similar number of females and males surveyed.

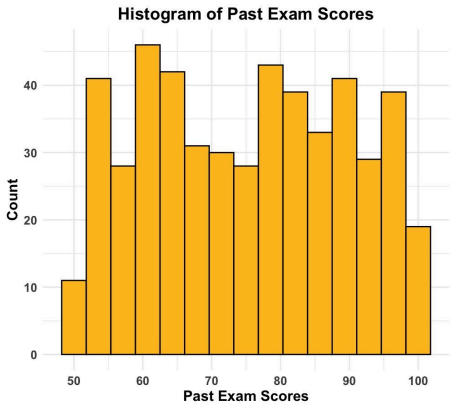
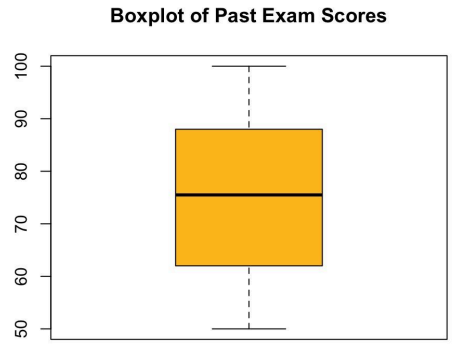
4.1.2 Average number of study hours per week, *Study_Hours_per_Week*

		The variable is continuous numerical. No outlying value of *Study_Hours_per_Week* is removed. The distribution of study hours fluctuates with no significant skewness.

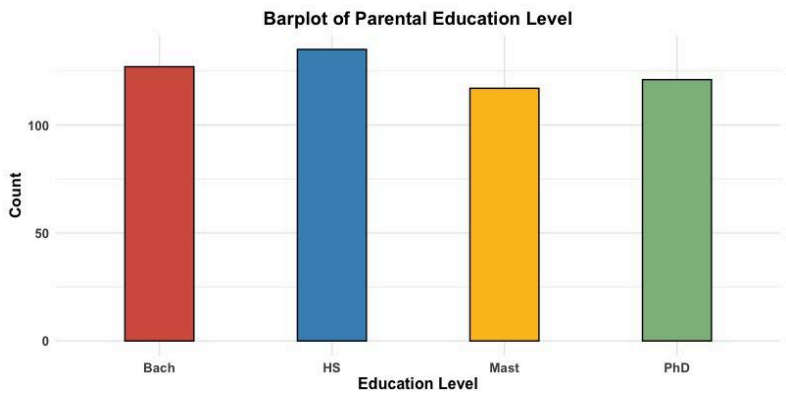
4.1.3 Attendance percentage (50% to 100%), *Attendance_Rate*

		The variable is continuous numerical. No outlying value of *Attendance_Rate* is removed. The distribution shows a concentration around higher attendance rates above 70%.

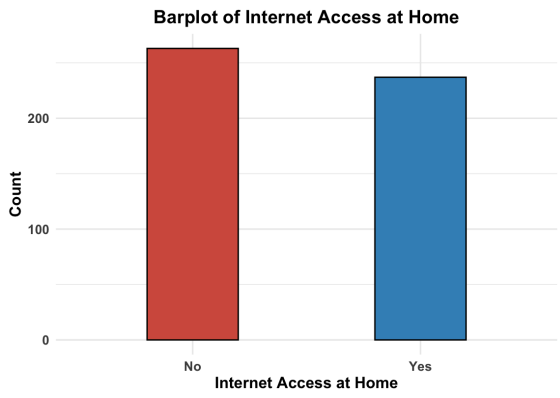
4.1.4 Average score of previous examinations (50 to 100), *Past_Exam_Scores*

		The variable is continuous numerical. No outlying value of *Past_Exam_Scores* is removed. The distribution of past exam scores fluctuates with no significant skewness.

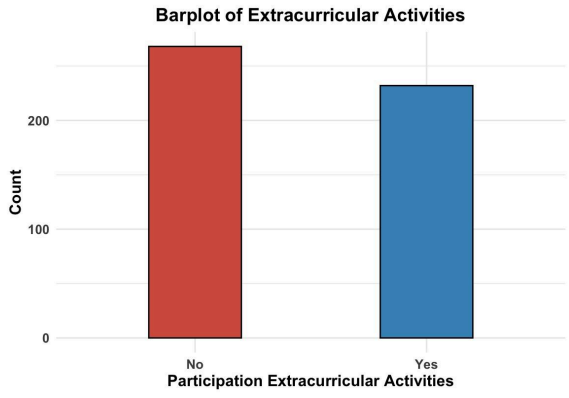
4.1.5 Education Level of Parents (High School, Bachelors, Masters, PhD), *Parental_Education_Level*

	The variable is categorical, so no transformation is applied. No outlying values of *Parental_Education_Level* are removed. The distribution is relatively balanced across the four education levels.

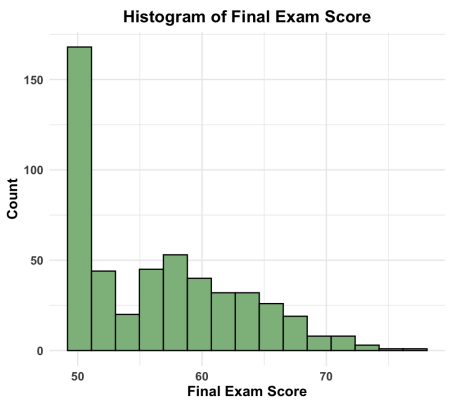
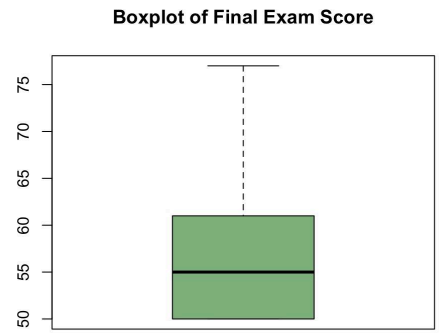
4.1.6 Internet Accessibility at Home (Yes/No), *Internet_Access_at_Home*

	The variable is binary categorical, so no transformation is applied. No outlying values of *Internet_Access_at_Home* are removed. A slightly larger number of students do not have internet access at home.

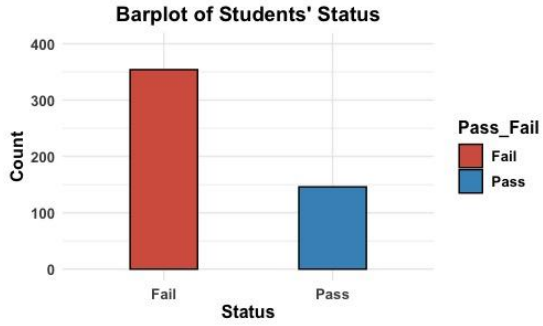
4.1.7 Whether the student is involved in extracurricular activities (Yes/No), *Extracurricular_Activities*

	The variable is binary categorical, so no transformation is applied. No outlying values of *Extracurricular_Activities* are removed. The number of students involved in extracurricular activities is nearly equal to those who are not.

4.1.8 Final Exam Score of the student (50-100, integer values), *Final_Exam_Score*

		The variable is continuous numerical. No outlying value of *Final_Exam_Score* is removed. The data is heavily right-skewed, with the majority of students scoring a 50 on the final examination.

4.1.9 Student Status (Pass/Fail), *Pass_Fail*

	The variable is binary categorical, so no transformation is applied. No outlying values of *Pass_Fail* are removed. Students have to score more than 60 to pass. The majority of students have failed.

4.2 Final dataset for analysis

Based on the above analysis, the final dataset has 500 observations (students) and 10 variables. *Study_Hours_per_Week*, *Attendance_Rate*, *Past_Exam_Scores* and *Final_Exam_Score* each do not follow the normal distribution. Even with log transformations applied to these variables, their p-values are still less than 0.05 when we conducted the Shapiro-Wilk Test. Hence, no data transformation was applied, and these variables remain non-normally distributed.

log(*Study_Hours_per_Week)*	log(*Attendance_Rate)*	log(*Past_Exam_Scores)*	log(*Final_Exam_Score)*
Shapiro-Wilk normality test data: ds$log_Study_Hours_per_Week W = 0.92005, p-value = 1.253e-15	Shapiro-Wilk normality test data: ds$log_Attendance_Rate W = 0.94874, p-value = 3.826e-12	Shapiro-Wilk normality test data: ds$log_Past_Exam_Scores W = 0.9512, p-value = 8.676e-12	Shapiro-Wilk normality test data: ds$log_Final_Exam_Score W = 0.89445, p-value < 2.2e-16

5 Statistical Analysis

5.1 Statistical Tests

5.1.1 Chi-Squared Test of *Pass_Fail* and *Gender*

Is there a difference in pass rates between male and female students?
In this analysis, we found both variables to be categorical. Hence, we opted for a Chi-Squared Test of Independence to analyse whether there is a significant relationship between a student's gender and the likelihood of passing their final exam.

H_0: There is no association between the gender of students and the likelihood of passing.
H_1: There is an association between the gender of students and the likelihood of passing.

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  gender_data
X-squared = 0.87286, df = 1, p-value = 0.3502
```

Figure 1: Result of chi-squared test of *Pass_Fail* and *Gender*

At a significance level of $\alpha = 0.05$, the obtained p-value is greater than 0.05. This means that we fail to reject the null hypothesis and conclude that there is no statistical association between the gender of students and their likelihood of passing. Therefore, the gender of a student does not influence a student's results.

5.1.2 Chi-Squared Test of *Pass_Fail* and *Study_Hours_per_Week*

Is there a trend in pass rates among students with low, medium, and high study hours per week?
To examine whether there is a trend in pass rates with students of different study hours per week, we categorise them by the number of hours they study a week using equal-width binning. We observe that there was a range from 10 to 39 hours per week. We divide this range into three intervals to form the categories below:

- Low: 10-19 hours
- Medium: 20-29 hours
- High: 30-39 hours

To determine whether there is an association between the number of hours studied per week and the passing rate, we use the Chi-Square Test of Independence. We selected this test as both variables are categorical. We developed the following hypotheses:

H_0: There is no association between different study hour categories and the likelihood of passing.
H_1: There is an association between different study hour categories and the likelihood of passing.

```
        Pearson's Chi-squared test

data:  table_hours
X-squared = 44.756, df = 2, p-value = 1.911e-10
```

Figure 2: Result of chi-squared test of *Pass_Fail* and *Study_Hours_per_Week*

From the Chi-Squared Test, the p-value is $1.911e^{-10}$. At the 0.05 level of significance, the p-value is less than 0.05. Hence, we reject the null hypothesis and conclude that there is a statistically significant association between the different study hour categories and the likelihood of passing. In short, this analysis suggests that the amount of time a student spent studying may influence their chances of passing the exam.

5.1.3 ANOVA Test of *Study_Hours_per_Week* and *Parental_Education_Level*

Do study hours vary across different levels of parental education?

To find out whether the amount of time a student spends studying each week varies based on the different levels of parental education, we examine the relationship between *Study_Hours_per_Week* and *Parental_Education_Level*. We find that the *Study_Hours_per_Week* is a continuous variable while *Parental_Education_Level* is a categorical variable, hence, we believe that a one-way Analysis of Variance (ANOVA) is an appropriate test for this analysis.

We plot a boxplot to illustrate the distributions of study hours per week among the different parental education levels.

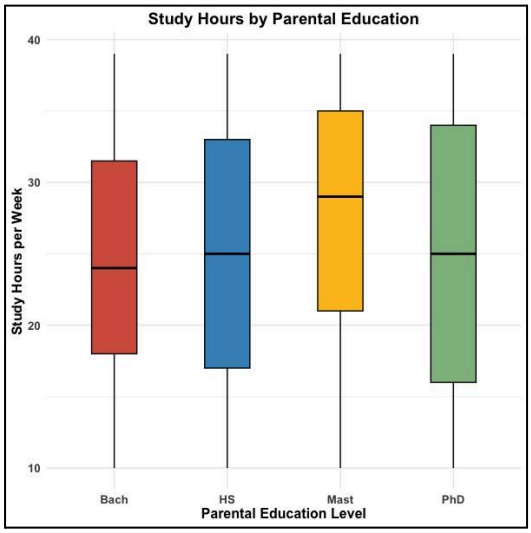


Figure 3: Boxplot of *Study_Hours_per_Week* by *Parental_Education_Level*

Before performing the ANOVA Test, the data must fulfill three conditions.

(1) Independence between the different Parental Education Level
It is satisfied as each student can only belong to one Parental Education Level Category.

(2) Normality
Since the sample size for each Parental Education Level exceeds 30. Hence, we can use Central Limit Theorem (CLT) to assume the normality of the samples and thus condition (2) is satisfied.

(3) Homogeneity of Variance

We use F-test to check whether 2 different Parental Education Levels have the same variance in study hours per week. In this F-test, we used the example of our Master Education Level versus the PhD Education Level.

H_0: $\sigma_{MAST} = \sigma_{PhD}$

H_1: $\sigma_{MAST} \neq \sigma_{PhD}$

```
        F test to compare two variances

data:  masters_hours and phd_hours
F = 0.85284, num df = 116, denom df = 120, p-value = 0.3894
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5937625 1.2263543
sample estimates:
ratio of variances
         0.8528435
```

Figure 4: Result of variance test of Master vs PhD Parental Education level

Using a significance level of α at 0.05, we found that all our F-tests yielded a p-value of more than 0.05. In our example, the comparison between students with parents holding a Master's and a PhD degree produced a p-value of 0.3894, indicating no significant difference in variances.

Since all the F-tests done failed to reject null hypothesis, we conclude that the assumption of homogeneity of variances holds. Hence, we deemed the ANOVA test appropriate for testing the equality of the means (μ_i). We test:

$H_0 : \mu_{Bach} = \mu_{HS} = \mu_{Mast} = \mu_{PhD}$ against $H_{1 :}$ not all μ_i are equal

```
                           Df Sum Sq Mean Sq F value Pr(>F)
Parental_Education_Level    3    526  175.26   2.278 0.0787 .
Residuals                 496  38155   76.93
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5: Result of ANOVA test of *Study_Hours_per_Week* and *Parental_Education_Level*

The ANOVA test returns a p-value of 0.0787, which means that we fail to reject the null hypothesis. This suggests that there is insufficient evidence to conclude that the mean study hours per week differ across the different parental education levels.

5.1.4 Wilcoxon Signed-Rank Test of *Past_Exam_Scores* and *Final_Exam_Score*

How do students' final examination scores compare to their past examination scores?
We will conduct a Wilcoxon Signed-Rank Test on *Past_Exam_Scores* and *Final_Exam_Score* to determine if there is a significant difference between the two distributions.

H_0: The distributions of *Past_Exam_Scores* and *Final_Exam_Score* are the same.
H_1: The distributions of *Past_Exam_Scores* and *Final_Exam_Score* are not the same.

```
        Wilcoxon signed rank test with continuity correction

data:  ds$Past_Exam_Scores and ds$Final_Exam_Score
V = 120225, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Figure 6: Result of Wilcoxon Signed-Rank Test of *Past_Exam_Scores* and *Final_Exam_Score*

From the Wilcoxon Signed-Rank Test, the p-value is less than $2.2e^{-16}$. At 0.05 level of significance, the p-value is less than 0.05. Hence, we reject H_0 and conclude that there is not enough evidence to support that the distributions of *Past_Exam_Scores* and *Final_Exam_Score* are the same.

Furthermore, the density distributions between *Past_Exam_Scores* and *Final_Exam_Scores* below clearly show the differences in distribution.

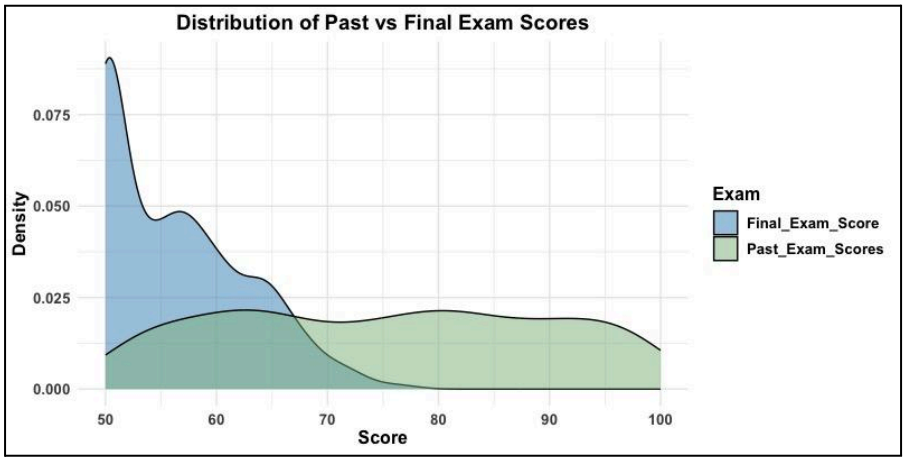


Figure 7: Density graph of *Past_Exam_Scores* vs *Final_Exam_Score*

Therefore, students generally did not perform better in their final examinations than in previous examinations, suggesting that they may be more difficult.

5.1.5 Kruskal-Wallis Test of *Final_Exam_Score* and *Parental_Education_Level*

How does parental education level influence students' performance in the final examination?
The Kruskal-Wallis Test will be used to test for differences in final exam score distributions across four levels of parental education: High School, Bachelor's, Master's, and PhD.

H_0: The distributions of *Final_Exam_Score* are the same across all levels of parental education.
H_1: At least one parental education level group has a different distribution of *Final_Exam_Score*.

```
        Kruskal-Wallis rank sum test

data:  Final_Exam_Score by Parental_Education_Level
Kruskal-Wallis chi-squared = 0.89001, df = 3, p-value = 0.8278
```

Figure 8: Result of Kruskal-Wallis Test of *Final_Exam_Score* and *Parental_Education_Level*

From the Kruskal-Wallis Test, the p-value is 0.8275. At the 0.05 level of significance, the p-value is greater than 0.05. Hence, we fail to reject H_0, and conclude that there is no statistically significant difference in the *Final_Exam_Score* distributions across all parental education levels.

This result is further proven by the violin plot shown below. The shapes of the distributions (e.g. widths of violin plot) are similar across all groups, with most scores concentrated between 50 and 70.

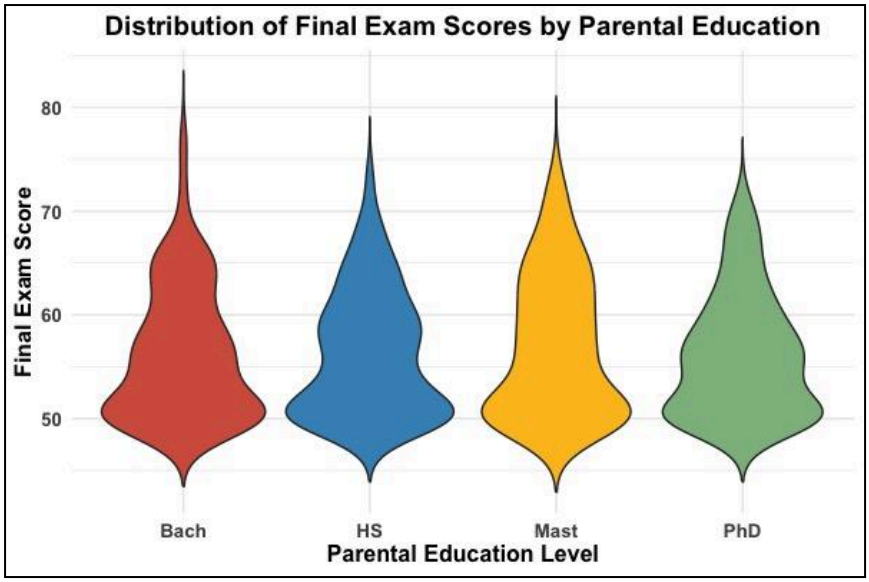


Figure 9: Violin plot of *Final_Exam_Score* by *Parental_Education_Level*

Hence, parental education level does not influence students' performance in the final examination.

5.1.6 Wilcoxon Rank Sum Test of *Past_Exam_Score*, *Final_Exam_Score* and *Internet_Access_at_Home*

Does internet access at home relate to how much students improve their examination scores?
We will use the Wilcoxon rank sum test to compare two independent samples of students with or without internet access at home and investigate whether internet access influences students' examination performance. We create a column, *Score_Change,* to store the changes in scores from *Past_Exam_Score* to *Final_Exam_Score*.

H_0: The distributions of *Score_Change* with or without access to the internet at home are identical.
H_1: The distributions of *Score_Change* with or without access to the internet at home are not identical.

```
        Wilcoxon rank sum test with continuity correction

data:  ds$Score_Change by ds$Internet_Access_at_Home
W = 30386, p-value = 0.6293
alternative hypothesis: true location shift is not equal to 0
```

Figure 10: Result of Wilcoxon Rank Sum Test of *Past_Exam_Score*, *Final_Exam_Score* and *Internet_Access_at_Home*

From the Wilcoxon Rank Sum Test, the p-value is 0.6293. At the 0.05 level of significance, the p-value is greater than 0.05. Hence, we fail to reject H_0 and conclude that there is no statistically significant difference in *Score_Change* between students with internet access at home and those without. The distributions of *Score_Change* with or without internet access at home are identical.

Hence, *Internet_Access_at_Home* does not affect the changes in students' examination scores.

5.1.7 Wilcoxon Rank Sum Test of *Final_Exam_Score* and *Extracurricular_Activities*

Is there a difference in final examination performance between students participating in extracurricular activities and those not?
The Wilcoxon Rank Sum Test will be used to compare the distribution of the final exam scores of students who participated in extracurricular activities vs those who did not, i.e. determine if their distributions are the same.

H_0: The distributions of *Final_Exam_Score* for students who participated in Extracurricular Activities and those who did not are the same.
H_1: The distributions of *Final_Exam_Score* for the two groups are different.

```
        Wilcoxon rank sum test with continuity correction

data:  Final_Exam_Score by Extracurricular_Activities
W = 29522, p-value = 0.3239
alternative hypothesis: true location shift is not equal to 0
```

Figure 11: Result of Wilcoxon Rank Sum Test of *Final_Exam_Score* and *Extracurricular_Activities*

From the Wilcoxon Rank Sum Test, the p-value is 0.3239. At the 0.05 level of significance, the p-value is greater than 0.05. Hence, we fail to reject H_0 and conclude that there is no statistically significant difference in the *Final_Exam_Score* distributions between students who participate in extracurricular activities and those who do not.

The boxplot shown below further proves this result. The median and interquartile range of both groups of students are similar, indicating similar distributions.

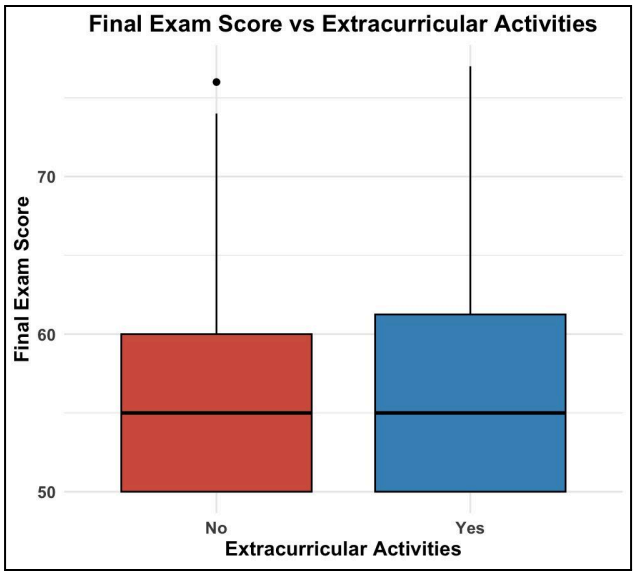


Figure 12: Boxplot of *Final_Exam_Score* vs *Extracurricular_Activities*

Thus, there is no difference in final examination performance between students participating in extracurricular activities and those not.

5.1.8 Friedman Rank Sum Test of *Attendance_Rate*, *Past_Exam_Scores* and *Final_Exam_Score*

Do students perform consistently across attendance rate, past and final examination scores?
We will conduct a Friedman Rank Sum Test to determine if there are statistically significant differences between three performance-related variables: *Attendance_Rate*, *Past_Exam_Score*, and *Final_Exam_Score*. The Friedman Test is suitable here because these variables were measured on the same group of students and expressed as a value out of 100.

H_0: All three measurements have identical distributions.
H_1: Not all three measurements have identical distributions.

```
        Friedman rank sum test

data:  Value and Measurement and Student_ID
Friedman chi-squared = 595.06, df = 2, p-value < 2.2e-16
```

Figure 13: Result of Friedman Rank Sum Test of *Attendance_Rate*, *Past_Exam_Scores* and *Final_Exam_Score*

From the Friedman Rank Sum Test, the p-value is less than $2.2e^{-16}$. At 0.05 level of significance, p-value is less than 0.05. Hence, we reject H_0 and conclude that there is not enough evidence to support the null hypothesis that all three measurements have the same distributions. This suggests that students do not perform consistently across the three variables.

To identify which pairs of measurements differ, a Pairwise Wilcoxon Rank Sum Test is conducted.

```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  ds1$Value and ds1$Measurement

                Attendance_Rate Final_Exam_Score
Final_Exam_Score <2e-16         -
Past_Exam_Scores 0.52           <2e-16

P value adjustment method: none
```

Figure 14: Result of Pairwise Wilcoxon Rank Sum Test of *Attendance_Rate*, *Past_Exam_Scores* and *Final_Exam_Score*

At a significance level of 0.05, there is sufficient evidence to conclude that *Attendance_Rate* and *Final_Exam_Score* are significantly different. However, there is insufficient evidence to conclude that *Attendance_Rate* and *Past_Exam_Scores* differ significantly. Additionally, our result supports the conclusion from Section 5.1.4 that *Past_Exam_Score* and *Final_Exam_Score* are not identical.

A possible interpretation is that past examination scores influence students' decisions to attend classes. Moreover, in the long term, whether a student attends class frequently may not be a crucial factor in determining final examination performance.

5.2 Correlation and Regression

5.2.1 Correlations between *Final_Exam_Score* and other continuous variables, *Study_Hours_per_Week*, *Attendance_Rate* and *Past_Exam_Scores*

In this section, we investigate the correlations between each continuous variable. Since they do not follow normal distributions, Spearman's correlation method was applied.

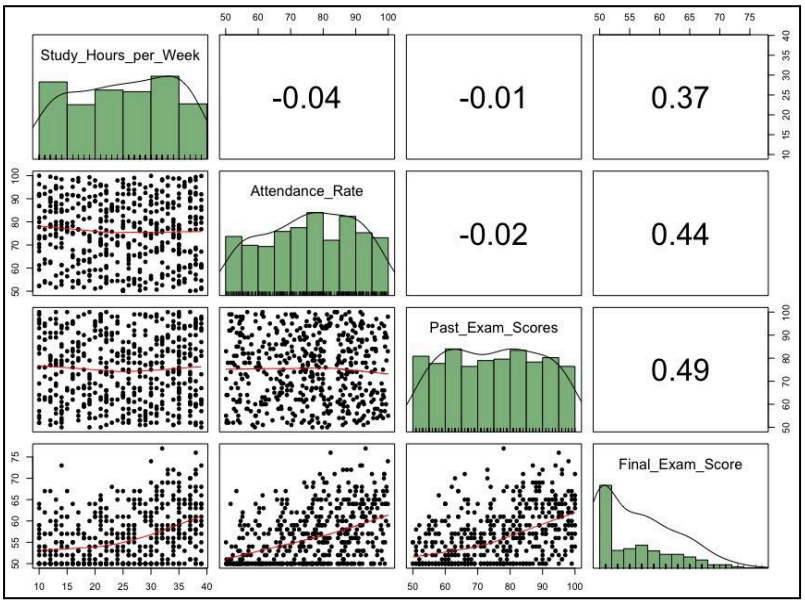


Figure 15: Relations between *Final_Exam_Score*, and *Study_Hours_per_Week*, *Attendance_Rate* and *Past_Exam_Scores*

Comparisons	Correlation Coefficient	Observations
Past_Exam_Scores vs *Final_Exam_Score*	0.49	The correlation coefficient is the highest at 0.49, meaning a moderate positive correlation. Students with higher scores in their past examinations are more likely to score higher in final examinations.
Attendance_Rate vs *Final_Exam_Score*	0.44	The correlation coefficient is 0.44, so there is a moderate positive correlation. Students who attend classes more frequently tend to score higher in final examinations.
Study_Hours_per_Week vs *Final_Exam_Score*	0.37	The correlation coefficient is lower at 0.37. There is a moderate positive correlation, but it is weaker than *Past_Exam_Scores* and *Attendance_Rate*. Students who study longer may improve their final examination scores, but it is not a strong predictor of their final scores.
Study_Hours_per_Week vs *Past_Exam_Scores*	-0.01	There is no significant correlation, and the scatter plot is spread out.
Attendance_Rate vs *Past_Exam_Scores*	-0.02	There is no significant correlation, and the scatter plot is spread out.
Study_Hours_per_Week vs *Past_Exam_Scores*	-0.04	There is no significant correlation, and the scatter plot is spread out.

5.2.2 Regression

```
Call:
lm(formula = Final_Exam_Score ~ Study_Hours_per_Week + Attendance_Rate +
    Past_Exam_Scores, data = ds2)

Residuals:
    Min      1Q  Median      3Q     Max
-10.485  -2.796  -0.451   2.556  14.521

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        17.47801    1.50630   11.60   <2e-16 ***
Study_Hours_per_Week  0.28444    0.02064   13.78   <2e-16 ***
Attendance_Rate       0.20758    0.01301   15.96   <2e-16 ***
Past_Exam_Scores      0.21266    0.01257   16.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.055 on 496 degrees of freedom
Multiple R-squared:  0.5838,    Adjusted R-squared:  0.5813
F-statistic: 231.9 on 3 and 496 DF,  p-value: < 2.2e-16
```

Figure 16: Linear Regression Model of *Final_Exam_Score*, and *Study_Hours_per_Week*, *Attendance_Rate* and *Past_Exam_Scores*

As an hour is added to the number of hours a student studied per week, the final examination score increases by 0.284 marks. As the attendance rate of a student increases by one percent, the final examination score increases by 0.208 marks. Additionally, a one-mark addition to past examination scores contributes 0.213 marks to the final examination score.

At a 0.05 level of significance, all three predictors — the number of study hours per week, the attendance rate, and a student's past examination scores — are statistically significant. This indicates that all three predictors have a higher likelihood of influencing a student's final examination score. The multiple R-squared value of 0.5838 suggests that these three variables can explain 58.38% of the variability in *Final_Exam_Score*, although other factors may also impact students' final examination scores. The y-intercept is estimated to be 17.478 and is likewise statistically significant. The residual standard error is 4.055, which refers to the deviation of the actual and predicted *Final_Exam_Score* being 4.055 marks.

Overall, our model captures some crucial predictors, but some variability is unexplained.

Thus, our fitted linear regression model is:

$$Final\_Exam\_Score = 17.478 + 0.284(Study\_Hours\_per\_Week) + 0.208(Attendance\_Rate) + 0.213(Past\_Exam\_Scores)$$

6 Conclusion and Discussion

In this report, our group attempted to answer some of the questions that many parents or students may have about what factors could affect a student's final examination score. In this report, we conclude that:

1. There is no significant association between a student's gender and their likelihood of passing their final examination.
2. There is a significant relationship between the number of hours spent studying a week and the passing rate. This means that students who study more per week have a higher chance of passing.
3. Study hours in a week do not significantly differ across the different parental education levels. This means that a student's parents' education level does not affect how much a student studies.
4. Students generally perform worse in their final examination as compared to their past examinations.
5. Parents' education level does not influence a student's performance in their final examination.
6. Having internet access at home does not affect the change in examination scores from past to final examinations.
7. Participating in extracurricular activities does not affect a student's final examination scores.
8. Past examination scores influence a student's decision to attend classes.

While these conclusions and findings may offer valuable insights to parents, teachers and even students themselves, we need to note that this dataset is limited in scope. There are many factors that may affect a student's final academic performance that were not captured in this dataset. Further research and analysis are required to provide more accurate and specific insights to many decisions.

7 Appendix

Appendix 1: ***Code for installing packages, loading library and cleaning of dataset***

```
# install packages
install.packages("dplyr")
install.packages("ggplot2")
install.packages("psych")

# loading libraries
library(dplyr)
library(ggplot2)
library(psych)

# cleaning of dataset
ds <- read.csv("student_performance_dataset.csv")
ds <- na.omit(ds)
ds <- ds %>% distinct(Student_ID, .keep_all = T)
```

Appendix 2: ***Code for plotting bar plot of Gender***

```
ggplot(ds, aes(x = Gender, fill = Gender)) +
  geom_bar(col = "black", width = 0.4) +
  scale_fill_manual(values = c("Male" = "#4393C3", "Female" = "#D6604D")) +
  labs(title = "Barplot of Gender", x = "Gender", y = "Count") +
  ylim(0, 300) +
  theme_minimal() +
  theme(plot.margin = margin(5, 5, 5, 5),
        plot.title = element_text(face = "bold", hjust = 0.5, size = 14),
        axis.title = element_text(face = "bold", hjust = 0.5, size = 12),
        axis.text = element_text(face = "bold", size = 10),
        legend.text = element_text(face = "bold", size = 10),
        legend.title = element_text(face = "bold", size = 12))
```

Appendix 3: ***Code for histogram and boxplot of Study_Hours_per_Week***

```
ggplot(ds, aes(x = Study_Hours_per_Week)) +
  geom_histogram(fill = "#D6604D", col = "black", bins = 18) +
  scale_fill_manual(values = c("Male" = "#4393C3", "Female" = "#D6604D")) +
  labs(title = "Histogram of Study Hours Per Week",
       x = "Study Hours Per Week",
       y = "Count") +
  theme_minimal() +
  theme(plot.margin = margin(5, 5, 5, 5),
        plot.title = element_text(face = "bold", hjust = 0.5, size = 14),
        axis.title = element_text(face = "bold", hjust = 0.5, size = 12),
        axis.text = element_text(face = "bold", size = 10),
        legend.text = element_text(face = "bold", size = 10),
        legend.title = element_text(face = "bold", size = 12))

boxplot(ds$Study_Hours_per_Week,
        main = "Boxplot of Study Hours Per Week",
```

```
      col = "#D6604D")
```

Appendix 4: ***Code for histogram and boxplot of Attendance_Rate***

```
ggplot(ds, aes(x = Attendance_Rate)) +
  geom_histogram(fill = "#4393C3", col = "black", bins = 18) +
  labs(title = "Histogram of Attendance Rate",
       x = "Attendance Rate",
       y = "Count") +
  theme_minimal() +
  theme(plot.margin = margin(5, 5, 5, 5),
        plot.title = element_text(face = "bold", hjust = 0.5, size = 14),
        axis.title = element_text(face = "bold", hjust = 0.5, size = 12),
        axis.text = element_text(face = "bold", size = 10),
        legend.text = element_text(face = "bold", size = 10),
        legend.title = element_text(face = "bold", size = 12))

boxplot(ds$Attendance_Rate,
        main = "Boxplot of Attendance Rate",
        col = "#4393C3")
```

Appendix 5: ***Code for histogram and boxplot of Past_Exam_Scores***

```
ggplot(ds, aes(x = Past_Exam_Scores)) +
  geom_histogram(fill = "#FFC125", col = "black", bins = 15) +
  labs(title = "Histogram of Past Exam Scores",
       x = "Past Exam Scores",
       y = "Count") +
  theme_minimal() +
  theme(plot.margin = margin(5, 5, 5, 5),
        plot.title = element_text(face = "bold", hjust = 0.5, size = 14),
        axis.title = element_text(face = "bold", hjust = 0.5, size = 12),
        axis.text = element_text(face = "bold", size = 10),
        legend.text = element_text(face = "bold", size = 10),
        legend.title = element_text(face = "bold", size = 12))

boxplot(ds$Past_Exam_Scores,
        main = "Boxplot of Past Exam Scores",
        col = "#FFC125",
        ylab = "Past Exam Scores")
```

Appendix 6: ***Code for barplot of Parental_Education_Level***

```
ds$Parental_Education_Short <- recode(ds$Parental_Education_Level,
                                      "High School" = "HS",
                                      "Bachelors" = "Bach",
                                      "Masters" = "Mast",
                                      "PhD" = "PhD")

ggplot(ds, aes(x = Parental_Education_Short, fill = Parental_Education_Short)) +
  geom_bar(col = "black", width = 0.4) +
  scale_fill_manual(values = c("HS" = "#4393C3",
```

```
                    "Bach" = "#D6604D",
                    "Mast" = "#FFC125",
                    "PhD" = "#8FBC8F")) +
  labs(title = "Barplot of Parental Education Level",
       x = "Education Level",
       y = "Count") +
  theme_minimal() +
  theme(legend.position = "none",
        plot.margin = margin(5, 5, 5, 5),
        plot.title = element_text(face = "bold", hjust = 0.5, size = 14),
        axis.title = element_text(face = "bold", hjust = 0.5, size = 12),
        axis.text = element_text(face = "bold", size = 10),
        legend.text = element_text(face = "bold", size = 10),
        legend.title = element_text(face = "bold", size = 12))
```

Appendix 7: *Code for barplot of Internet_Access_at_Home*

```
ggplot(ds, aes(x = Internet_Access_at_Home, fill = Internet_Access_at_Home)) +
  geom_bar(col = "black", width = 0.4) +
  scale_fill_manual(values = c("Yes" = "#4393C3",
                    "No" = "#D6604D")) +
  labs(title = "Barplot of Internet Access at Home",
       x = "Internet Access at Home",
       y = "Count") +
  theme_minimal() +
  theme(legend.position = "none",
        plot.margin = margin(5, 5, 5, 5),
        plot.title = element_text(face = "bold", hjust = 0.5, size = 14),
        axis.title = element_text(face = "bold", hjust = 0.5, size = 12),
        axis.text = element_text(face = "bold", size = 10),
        legend.text = element_text(face = "bold", size = 10),
        legend.title = element_text(face = "bold", size = 12))
```

Appendix 8: *Code for barplot of Extracurricular_Activities*

```
ggplot(ds, aes(x = Extracurricular_Activities, fill = Extracurricular_Activities)) +
  geom_bar(col = "black", width = 0.4) +
  scale_fill_manual(values = c("Yes" = "#4393C3", "No" = "#D6604D")) +
  labs(title = "Barplot of Extracurricular Activities",
       x = "Participation Extracurricular Activities",
       y = "Count") +
  theme_minimal() +
  theme(legend.position = "none",
        plot.margin = margin(5, 5, 5, 5),
        plot.title = element_text(face = "bold", hjust = 0.5, size = 14),
        axis.title = element_text(face = "bold", hjust = 0.5, size = 12),
        axis.text = element_text(face = "bold", size = 10),
        legend.text = element_text(face = "bold", size = 10),
        legend.title = element_text(face = "bold", size = 12))
```

Appendix 9: *Code for histogram and boxplot of Final_Exam_Score*

```
ggplot(ds, aes(x = Final_Exam_Score)) +
  geom_histogram(fill = "#8FBC8F", col = "black", bins = 15) +
  labs(title = "Histogram of Final Exam Score",
       x = "Final Exam Score",
       y = "Count") +
  theme_minimal() +
  theme(plot.margin = margin(5, 5, 5, 5),
        plot.title = element_text(face = "bold", hjust = 0.5, size = 14),
        axis.title = element_text(face = "bold", hjust = 0.5, size = 12),
        axis.text = element_text(face = "bold", size = 10),
        legend.text = element_text(face = "bold", size = 10),
        legend.title = element_text(face = "bold", size = 12))

boxplot(ds$Final_Exam_Score,
        main = "Boxplot of Final Exam Score",
        col = "#8FBC8F")
```

Appendix 10: *Code for barplot of Pass_Fail*

```
ggplot(ds, aes(x = Pass_Fail, fill = Pass_Fail)) +
  geom_bar(col = "black", width = 0.4) +
  scale_fill_manual(values = c("Pass" = "#4393C3", "Fail" = "#D6604D")) +
  labs(title = "Barplot of Students' Status",
       x = "Status",
       y = "Count") +
  theme_minimal() +
  ylim(0, 400) +
  theme(plot.margin = margin(5, 5, 5, 5),
        plot.title = element_text(face = "bold", hjust = 0.5, size = 14),
        axis.title = element_text(face = "bold", hjust = 0.5, size = 12),
        axis.text = element_text(face = "bold", size = 10),
        legend.text = element_text(face = "bold", size = 10),
        legend.title = element_text(face = "bold", size = 12))
```

Appendix 11*: Code for Shapiro-Wilk Normality Test*

```
ds$log_Study_Hours_per_Week <- log(ds$Study_Hours_per_Week)
ds$log_Attendance_Rate <- log(ds$Attendance_Rate)
ds$log_Past_Exam_Scores <- log(ds$Past_Exam_Scores)
ds$log_Final_Exam_Score <- log(ds$Final_Exam_Score)

shapiro.test(ds$log_Attendance_Rate)
shapiro.test(ds$log_Study_Hours_per_Week)
shapiro.test(ds$log_Past_Exam_Scores)
shapiro.test(ds$log_Final_Exam_Score)
```

Appendix 12*: Code for Chi-Squared Test for Gender and Pass_Fail*

```
gender_data <- table(ds$Gender, ds$Pass_Fail)
chisq.test(gender_data)
```

Appendix 13: ***Code for Chi-Squared Test for Pass_Fail and Study_Hours_per_week***

```
ds$Study_Hour_Category <- cut(ds$Study_Hours_per_Week,
                        breaks = c(9, 19, 29, 39),
                        labels = c("Low", "Medium", "High"),
                        include.lowest = TRUE)

table_hours <- table(ds$Study_Hour_Category, ds$Pass_Fail)

chisq.test(table_hours)
```

Appendix 14: ***Code for ANOVA Test of Study_Hours_per_Week and Parental_Education_Level***

```
ds$Parental_Education_Short <- recode(ds$Parental_Education_Level,
                                "High School" = "HS",
                                "Bachelors" = "Bach",
                                "Masters" = "Mast",
                                "PhD" = "PhD")

ggplot(ds, aes(x = Parental_Education_Short,
           y = Study_Hours_per_Week,
           fill = Parental_Education_Short)) +
  geom_boxplot(col = "black", width = 0.4) +
  scale_fill_manual(values = c("HS" = "#4393C3",
                         "Bach" = "#D6604D",
                         "Mast" = "#FFC125",
                         "PhD" = "#8FBC8F")) +
  labs(title = "Study Hours by Parental Education",
      x = "Parental Education Level",
      y = "Study Hours per Week") +
  theme_minimal() +
  theme(legend.position = "none",
       plot.margin = margin(5, 5, 5, 5),
       plot.title = element_text(face = "bold", hjust = 0.5, size = 14),
       axis.title = element_text(face = "bold", hjust = 0.5, size = 12),
       axis.text = element_text(face = "bold", size = 10),
       legend.text = element_text(face = "bold", size = 10),
       legend.title = element_text(face = "bold", size = 12))

anova_model <- aov(Study_Hours_per_Week ~ Parental_Education_Level, data = ds)

summary(anova_model)
```

Appendix 15: ***Code for Wilcoxon Signed-Rank Test Past_Exam_Scores and Final_Exam_Score***

```
wilcox.test(ds$Past_Exam_Scores, ds$Final_Exam_Score, paired = T)

# create new dataset scores with different structure to get density plot
scores <- data.frame(Exam = c(rep("Past_Exam_Scores", nrow(ds)),
                        rep("Final_Exam_Score", nrow(ds))),
                  Score = c(ds$Past_Exam_Scores, ds$Final_Exam_Score))
```

```
ggplot(scores, aes(x = Score, fill = Exam)) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribution of Past vs Final Exam Scores",
       x = "Score",
       y = "Density") +
  scale_fill_manual(values = c("#4393C3", "#8FBC8F")) +
  theme_minimal() +
  theme(plot.margin = margin(5, 5, 5, 5),
        plot.title = element_text(face = "bold", hjust = 0.5, size = 14),
        axis.title = element_text(face = "bold", hjust = 0.5, size = 12),
        axis.text = element_text(face = "bold", size = 10),
        legend.text = element_text(face = "bold", size = 10),
        legend.title = element_text(face = "bold", size = 12))
```

Appendix 16: ***Code for Kruskal Wallis Test and Violin plot of Final_Exam_Score by Parental_Education_Level***

```
kruskal.test (Final_Exam_Score ~ Parental_Education_Level, data = ds)

ds$Parental_Education_Short <- recode(ds$Parental_Education_Level,
                                      "High School" = "HS",
                                      "Bachelors" = "Bach",
                                      "Masters" = "Mast",
                                      "PhD" = "PhD")

ggplot(ds, aes(x = Parental_Education_Short,
               y = Final_Exam_Score,
               fill = Parental_Education_Short)) +
  geom_violin(trim = FALSE) +
  scale_fill_manual(values = c("HS" = "#4393C3",
                               "Bach" = "#D6604D",
                               "Mast" = "#FFC125",
                               "PhD" = "#8FBC8F")) +
  labs(title = "Distribution of Final Exam Scores by Parental Education",
       x = "Parental Education Level",
       y = "Final Exam Score") +
  theme_minimal() +
  theme(legend.position = "none",
        plot.margin = margin(5, 5, 5, 5),
        plot.title = element_text(face = "bold", hjust = 0.5, size = 14),
        axis.title = element_text(face = "bold", hjust = 0.5, size = 12),
        axis.text = element_text(face = "bold", size = 10),
        legend.text = element_text(face = "bold", size = 10),
        legend.title = element_text(face = "bold", size = 12))
```

Appendix 17: ***Code for Wilcoxon Rank Sum Test of Past_Exam_Score, Final_Exam_Score and Internet_Access_at_Home***

```
# create new column Score_Change (final - past)
ds$Score_Change <- ds$Final_Exam_Score - ds$Past_Exam_Scores
ds$Score_Change
wilcox.test(ds$Score_Change ~ ds$Internet_Access_at_Home)
```

Appendix 18: ***Code for Wilcoxon Rank Sum Test of Final_Exam_Score and Extracurricular_Activities***

```
wilcox.test(Final_Exam_Score ~ Extracurricular_Activities, data = ds)

# boxplot of Final Exam Score vs Extracurricular Activities
ggplot(ds, aes(x = Extracurricular_Activities, y = Final_Exam_Score)) +
  geom_boxplot(color = "black", fill = c("#D6604D", "#4393C3")) +
  labs(title = "Final Exam Score vs Extracurricular Activities",
       x = "Extracurricular Activities",
       y = "Final Exam Score") +
  theme_minimal() +
  theme(plot.margin = margin(5, 5, 5, 5),
        plot.title = element_text(face = "bold", hjust = 0.5, size = 14),
        axis.title = element_text(face = "bold", hjust = 0.5, size = 12),
        axis.text = element_text(face = "bold", size = 10),
        legend.text = element_text(face = "bold", size = 10),
        legend.title = element_text(face = "bold", size = 12))
```

Appendix 19: ***Code for Friedman Rank Sum Test of Attendance_Rate, Past_Exam_Scores and Final_Exam_Score***

```
# select relevant variables and create new dataset ds1 with different structure
ds1 <- data.frame(Student_ID = rep(ds$Student_ID, times = 3),
                  Measurement = c(rep("Past_Exam_Scores", nrow(ds)),
                                  rep("Attendance_Rate", nrow(ds)),
                                  rep("Final_Exam_Score", nrow(ds))),
                  Value = c(ds$Past_Exam_Scores, ds$Attendance_Rate, ds$Final_Exam_Score))
friedman.test(Value ~ Measurement | Student_ID, data = ds1)
pairwise.wilcox.test(ds1$Value, ds1$Measurement, p.adjust.method = "none")
```

Appendix 20: ***Code for Correlations between Final_Exam_Score and other continuous variables, Study_Hours_per_Week, Attendance_Rate and Past_Exam_Scores***

```
# create new dataset ds2 with continuous variables only
ds2 <- ds[, c(3, 4, 5, 9)]
pairs.panels(ds2,
             method = "spearman",     # variables are not normal
             pch = 16,
             density = TRUE,
             ellipses = FALSE,
             hist.col = "darkseagreen",
             cex.cor = 1.5,
             cex.labels = 1.5)
```

Appendix 21: ***Code for*** **Regression of** ***Final_Exam_Score*****, and** ***Study_Hours_per_Week*****,** ***Attendance_Rate*** **and** ***Past_Exam_Scores***

```
ds2 <- ds[, c(3, 4, 5, 9)]

model1 <- lm(Final_Exam_Score ~ Study_Hours_per_Week + Attendance_Rate + Past_Exam_Scores,
             data = ds2)

summary(model1)
```

8 Reference

1. *How the education system changes can help more kids succeed*. The Straits Times. (2020, April 21). https://www.straitstimes.com/singapore/how-education-system-changes-can-help-more-kids-succeed-0 (The Straits Times, 2020)