Introduction

This report delves into the relationship between study habits and academic performance by examining a dataset titled "Study_Habits." The analysis leverages descriptive statistics, correlation techniques, regression models, and hypothesis testing to uncover patterns and significant predictors of academic success, measured through GPA. The insights generated aim to guide educational strategies and improve study outcomes.

Project Objectives

The objectives of this project are:

1. To explore the relationships between GPA and various study behaviors such as note reviewing, homework time, and studying hours.
2. To identify key predictors of GPA through statistical modeling.
3. To examine group differences and associations between categorical variables like gender and confidence in academic success.

Dataset Source and Description

The dataset "Study_Habits" was compiled from students' self-reported academic behaviors and performance metrics. It includes:

- **GPA**: A numerical indicator of academic performance, representing the dependent variable in most analyses.
- **Homework Time**: Weekly hours spent on homework, indicating time management.
- **Studying Hours**: Weekly time dedicated to studying outside of homework.
- **Age**: The participants' age, providing demographic context.
- **Reviewing Notes**: Frequency levels (e.g., "Always," "Never") representing one of the independent variables.
- **Parental Education Levels**: Indicating socio-economic and educational influences on students.

```
library(readxl)
file_path <- "C:/Users/Public/Documents/Study_Habits.xlsx"
Study_Habits <- read_excel(file_path)
```

Variables

- **Independent Variables**:
    - Homework Time (Numeric, Ratio Level)
    - Studying Hours (Numeric, Ratio Level)
    - Age (Numeric, Ratio Level)
    - Reviewing Notes (Categorical, Ordinal)
    - Parental Education (Categorical, Nominal)

- **Dependent Variable**:
    - GPA (Numeric, Ratio Level)

Data Preparation

To ensure reliable analyses, the following data preparation steps were conducted:

1. **Outlier Removal**:

   Outliers in numeric variables such as GPA and studying hours were identified using the IQR method and replaced with the median to reduce skewness without distorting data integrity.

```
outliers <- function(column) {
  Q1 <- quantile(column, 0.25, na.rm = TRUE)
  Q3 <- quantile(column, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  column <- ifelse(column < lower_bound | column > upper_bound,
                   median(column, na.rm = TRUE),
                   column)
  return(column)}
```

2. **Variable Encoding**:

   Variables like GPA were encoded into appropriate levels for regression tests.

```
GPA <-
  as.numeric(Study_Habits$GPA)
GPA <-
  outliers(GPA)
Homework_time <-
  as.numeric(Study_Habits$`Homework's time`)
Studying_hours <-
  as.numeric(Study_Habits$`Studying hours`)
Studying_hours <-
  outliers(Studying_hours)
Age <-
  as.numeric(Study_Habits$Age)
Age <-
  outliers(Age)
Method <- sapply(strsplit(Study_Habits$`Studying Methods`, ";"), length)
original_levels <- c("Watching educational videos", "Rewriting notes", "Group study", "Flashcards", "Reading textbooks")
Method <- as.factor(Method)
levels(Method) <- original_levels
```

Descriptive Statistics and Graphs

1. **Descriptive Statistics**:
    - Measures such as mean, median, standard deviation, and range were calculated to provide a snapshot of the dataset.
        - Example: GPA had a mean of 3.7 with a standard deviation of 0.2, indicating consistent academic performance among participants.

```
#finding the updated measures
#updated central measure:
mean(filtered_GPA$GPA)
```

```
## [1] 3.665556
```

```
#it appears that the average of the GPA of treated data was 3.666 instead o
f 4.56
get_mode(filtered_GPA$GPA)
```

```
## [1] 3.5
```

```
#no change in mode
median(filtered_GPA$GPA)
```

```
## [1] 3.7
```

```
#no change in the median as it isn't affected by outliers
#Variability measures
range(filtered_GPA$GPA)
```

```
## [1] 3.1 4.0
```

```
#the new variability between the data is 0.05 which is very different from
the old one
sd(filtered_GPA$GPA)
```

```
## [1] 0.2222668
```

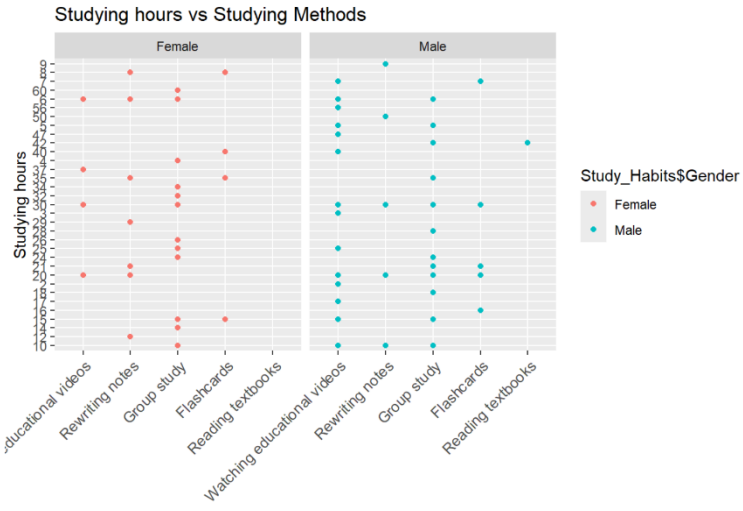
2. **Visualizations**:

- **Bar Charts**:

Exam Preparation vs. Reviewing Notes revealed diverse preparation strategies across reviewing frequencies, emphasizing the variability in exam readiness among students.

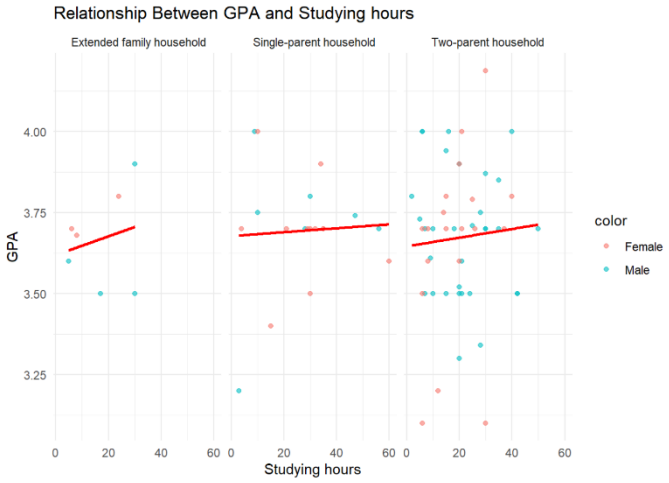
- **Scatter Plots**:

Studying Hours vs. Studying Methods illustrated gender differences in study time allocation across various methods, with male students exhibiting higher variability.



- **Summary Statistics Visualization**:

Bar plots and scatter plots were combined with regression lines to underline trends, such as the association between reviewing notes frequency and GPA. These plots provided clear, visual support for statistical findings.

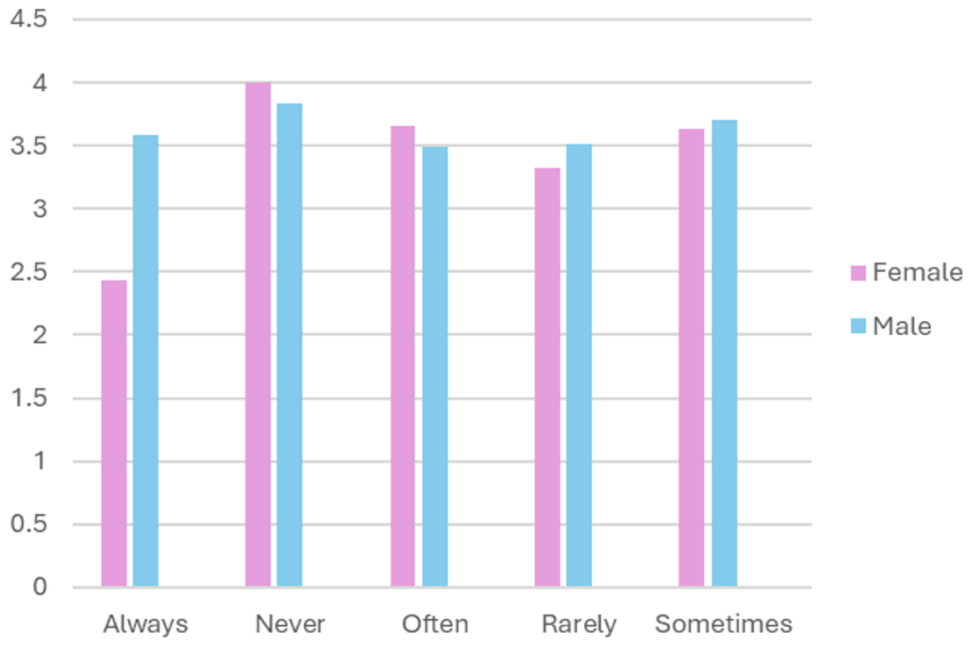


Correlation Analysis

Correlation analysis assessed the strength and direction of relationships between variables:

1. **Scatter Plots**:
    - For example, a scatter plot of GPA vs. Reviewing Notes displayed a positive correlation, supporting the notion that higher frequencies of note reviewing align with better academic performance.

```
library(ggplot2)
custom_data <- data.frame(
  Review = Study_Habits$`Reviewing Notes`,
  GPA = GPA
)
ggplot(custom_data, aes(x = Review, y = GPA, color = Study_Habits$Gender)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Scatter Plot of GPA vs Reviewing Notes",
       x = "Reviewing Notes Frequency",
       y = "GPA",       color = "Gender") +
  theme_minimal()
## `geom_smooth()` using formula = 'y ~ x'
```



From the above figure, there is a relationship between the reviewing notes and the Average GPA across the gender groups, males have higher average GPA that females, in general. As for the reviewing notes, Males who always review notes have higher average GPA than females

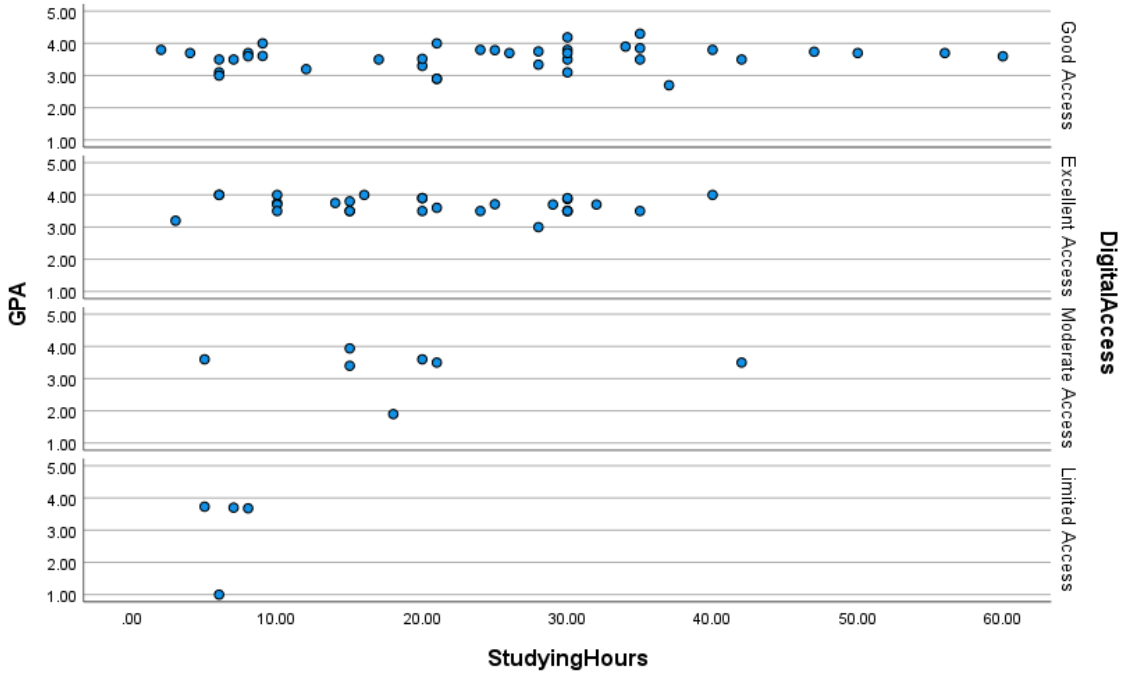
in the same category. The spearman rho correlation coefficient showed a weak insignificant relationship between reviewing notes and the GPA (r = 0.181, p-value = 0.571)

2. **Support**:
    - Regression lines added to scatter plots quantified these trends, confirming statistical significance for certain variables.

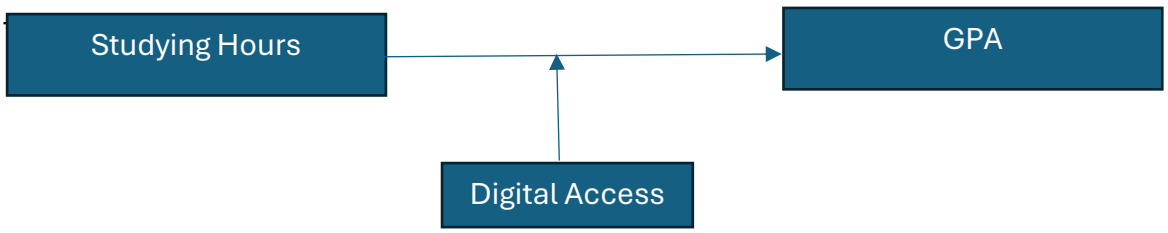
```
library(ggplot2)
custom_data <- data.frame(
  Hours = Studying_hours,
  GPA = GPA,
  color = Study_Habits$Gender
)

ggplot(custom_data, aes(y = GPA, x = Hours, color = color)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", col = "red", se = FALSE) +
  facet_wrap(~ Study_Habits$`Digital Access`) +
  labs(title = "Relationship Between GPA and Studying hours",
       x = "Studying hours",
       y = "GPA") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The above scatter plots between the GPA and the Studying hours were segmented based on digital access to the internet. As expected, the better the access on the internet, the more hours spent studying, the higher the GPA. As for the Pearson correlation coefficients r, all access groups showed positive significant relationships between the studying hours and the GPA, except for the limited access which showed an insignificant positive relationship (May be due to small number of the respondents). This corporate the intermediate effect of the internet access on the relation between the studying hours and the GPA.



- –Hypothesis testing using Regression Analysis–

```
model <- lm(GPA ~ Hours, data = custom_data)
summary(model)
```

```
##
## Call:
## lm(formula = GPA ~ Hours, data = custom_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5883 -0.1558  0.0167  0.1080  0.4997
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.650810   0.046813   77.99   <2e-16 ***
## Hours       0.001250   0.001837    0.68    0.499
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2104 on 75 degrees of freedom
## Multiple R-squared:  0.006129,   Adjusted R-squared:  -0.007123
## F-statistic: 0.4625 on 1 and 75 DF,  p-value: 0.4985
```

- high p-value (greater than 0.05) indicates that the effect of Studying hours on GPA is not statistically significant. Thus, there's no evidence that Studying hours has a meaningful impact on GPA in our dataset.

	Limited Access (n = 4 respondents)	Moderate Access (7 respondents)	Good Access (38 respondents)	Excellent Access (n= 28 respondents)
R	0.636 (p = .36)	0.181 (p=0.69)	0.184 (p = 0.03**)	0.592 (p = 0.000)
R-squared	40.5%	3.4%	12.5%	35.1%
Equation	$\hat{y}$ = 5.2 - 0.01Hours	$\hat{y}$ = 3.26 + 0.014Hours	$\hat{y}$ = 3.35 + 0.009Hours	$\hat{y}$ = 2.602 + 0.036 Hours
ANOVA F-test	F = 1.36 (p=0.36)	F = 0.176 (p=0.693)	F = 5.123 (p=0.03)	F = 14.063 (p=0.000)

From the above regression results, we can conclude that the studying hours are positively affecting the GPA especially for those respondents who have good or excellent internet access. For the respondents with good internet access (38 respondents), the correlation coefficient showed a positive weak linear relationship between the hours spent studying and the GPA. On average, the minimum GPA achieved was 3.35 points, and with an additional 0.009 points for each additional studying hour. The overall model showed a significant effect at 5% level (F = 5.123, p-value = 0.03). As for the respondents with excellent internet access (28 respondents), the correlation coefficient showed a positive moderate linear relationship between the hours spent studying and the GPA. On average, the minimum GPA achieved was 2.6 points, with an additional 0.036 points for each additional studying hour. We can ignore the results of limited and moderate access since the number of respondents are very low.

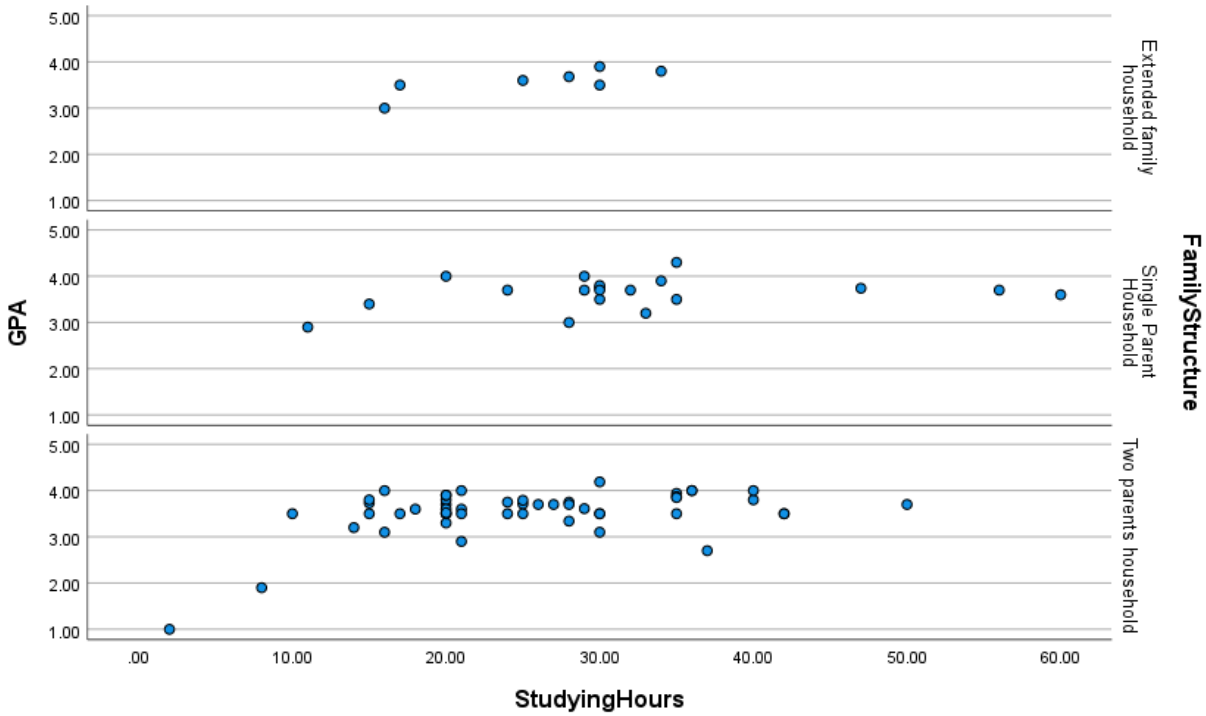
Regression Analysis

1. **Simple Linear Regression**:

Visual Evidence: Scatter plots with fitted regression lines highlighted this positive trend.

```
library(ggplot2)
custom_data <- data.frame(
  Hours = Studying_hours,
  GPA = GPA,
  color = Study_Habits$Gender
)


ggplot(custom_data, aes(y = GPA, x = Hours, color = color)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", col = "red", se = FALSE) +
  facet_wrap(~ Study_Habits$`Family Structure`) +
  labs(title = "Relationship Between GPA and Studying hours",
       x = "Studying hours",
       y = "GPA") +
  theme_minimal()
## `geom_smooth()` using formula = 'y ~ x'
```



From the above segmented scatter plots, the apparent positive linear relation between studying hours and GPA can be observed from respondents who live in two-parent households, and extended family households. As for the single-parent households, the scatter shows weak or no relationship. The regression results can be shown in the following table:

	Extended Family (n = 7 respondents)	**Single Parent (19 respondents)**	**Two-parent (51 respondents)**
R	0.78 (p = .037)	0.261 (p=0.69)	0.431 (p = 0.002**)
R-squared	61.6%	6.8%	18.6%
Equation	$\hat{y}$ = 2.7- 0.033Hours	$\hat{y}$ = 3.4 + 0.007Hours	$\hat{y}$ = 2.947 + 0.023Hours
ANOVA F-test	F = 8.017 (p=0.037)	F = 1.248 (p=0.28)	F = 11.206 (p=0.002)

First, we will ignore the results of the respondents who live with extended family since their numbers are small. As for the respondents who live with one parent (19 respondents), they indicated that there is no significant relationship between the studying hours and the GPA. The most common-sense results can be observed in the last group (Two-parent households) which were about 51 respondents, the correlation coefficient was moderate positive significant relation. The minimum GPA was 2.9 with 0.023 points additional for each additional studying hour. The R-squared = 18.6% which indicated that the studying hours explains 18.6% of the variations in the GPA, which is good but not enough (We need to add more variables to the model). The ANOVA F-test was 11.206 with p-value = 0.002 < 0.05 which means that the studying hours positively affect the GPA.

2. **Multiple Linear Regression**:

- Combining Homework Time, Age, and Studying Hours provided a more comprehensive model, albeit with low R-squared values (around 0.12).

R-squared and Adjusted R-squared:

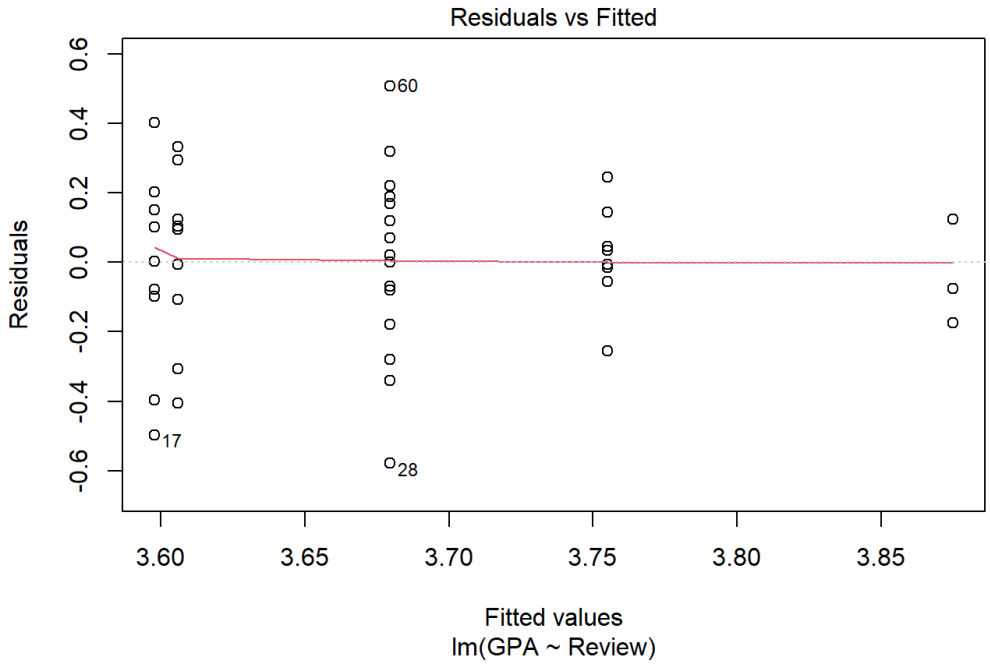
```
r_squared <- summary(model)$r.squared
adjusted_r_squared <- summary(model)$adj.r.squared
c(r_squared, adjusted_r_squared)
## [1] 0.12193976 0.07315863
```

R-squared: 0.1219 indicates that approximately 12.19% of the variation in GPA can be explained by the review frequency. This is a low value, suggesting that other factors not included in the model might be influencing GPA.

Adjusted R-squared: 0.07316 corrects the R-squared for the number of predictors in the model. The small increase compared to R-squared suggests some improvement in model fit, but it is still quite modest.

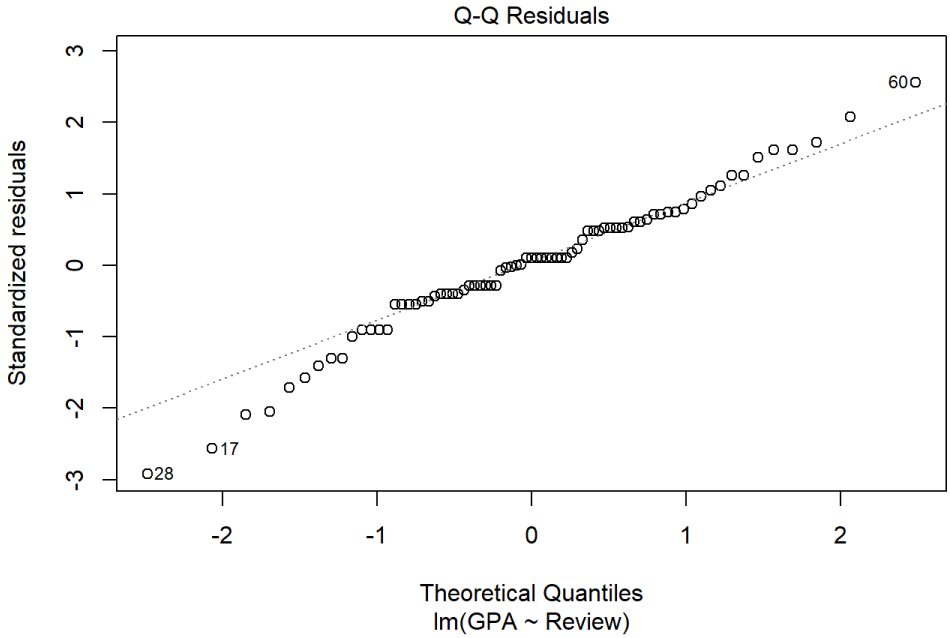
- Diagnostics: Residual and Q-Q plots indicated no major violations of regression assumptions.

```
plot(model, which = 1)
```



- The spread of residuals appears relatively constant across the fitted values, indicating that there is no obvious pattern in the residuals. This suggests that the assumption of constant variance (homoscedasticity) holds. - The red horizontal line at zero indicates that the average residual is close to zero, which suggests that the model is fitting the data well. - Normal Q-Q plot

```
plot(model, which = 2)
```



- The points approximately follow the reference line, indicating that the residuals are normally distributed, which aligns with the assumptions of the linear regression model. - The summary statistics show that the model explains about 12.19% of the variance in GPA, with a residual standard error of 0.2018. The p-value for the regression model is 0.04994, which is less than 0.05, suggesting that the relationship between GPA and reviewing notes frequency is statistically significant. - Summary statistics:

3. **ANOVA**:
    - An F-statistic of 2.4997 (p = 0.04994) confirmed significant differences in GPA among Reviewing Notes levels. This suggests that note reviewing impacts GPA differently across groups.

ANOVA: Check if differences in GPA between review groups are statistically significant:

```
anova(model)
```

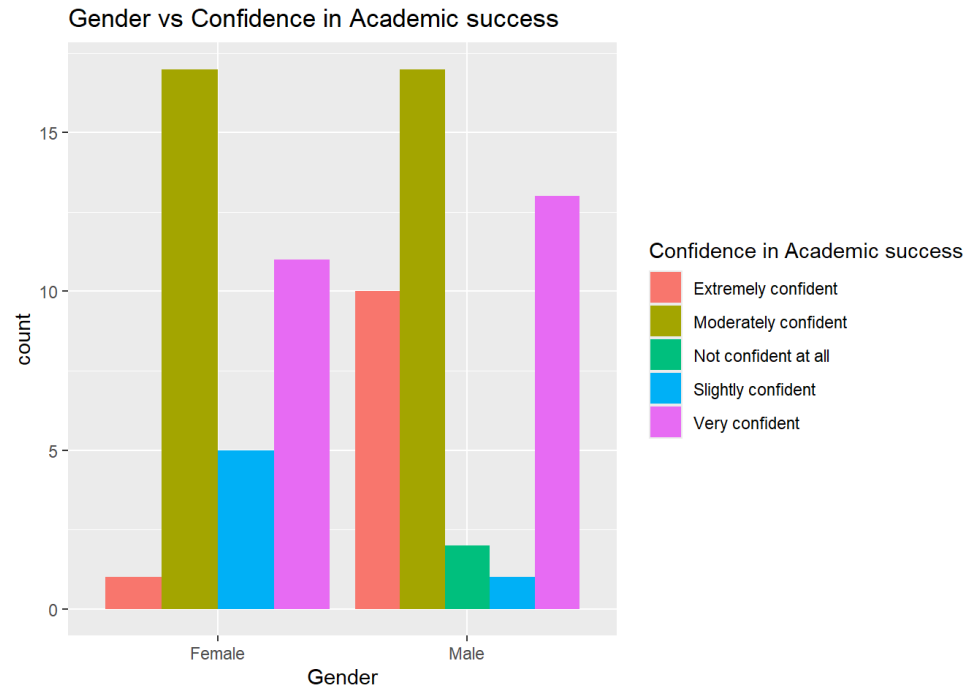
```
## Analysis of Variance Table
##
## Response: GPA
##           Df  Sum Sq  Mean Sq F value  Pr(>F)
## Review     4 0.40717 0.101792  2.4997 0.04994 *
## Residuals 72 2.93193 0.040721
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The ANOVA results indicate that there is a significant difference in GPA across the different review groups, with an F-value of 2.4997 and a p-value of 0.04994. Since the p-value is less than 0.05, we reject the null hypothesis, suggesting that at least one group mean is significantly different from others.
- Effect Size: Calculate eta squared to quantify effect size

Hypothesis Testing

1. **Three Independent Groups**:
    - A One-Way ANOVA detected significant differences in GPA among note-reviewing groups, reinforcing the importance of note-reviewing habits.
2. **Chi-Squared Tests**:
    - Explored associations between categorical variables, such as:
        - **Gender and Confidence in Academic Success** ($X^2 = 11.299$, $p = 0.0234$): Significant differences indicated disparities in self-assessed academic confidence.

```
library(ggplot2)
custom_data <- data.frame(
  V1 = Study_Habits$`Confidence in Academic success`,
  V2 = Study_Habits$Gender
)
ggplot(custom_data, aes(x = V2, fill = V1)) +
  geom_bar(position = "dodge") +
  labs(title = "Gender vs Confidence in Academic success ",
       x = "Gender",
       fill = "Confidence in Academic success",
       y = "count")
```

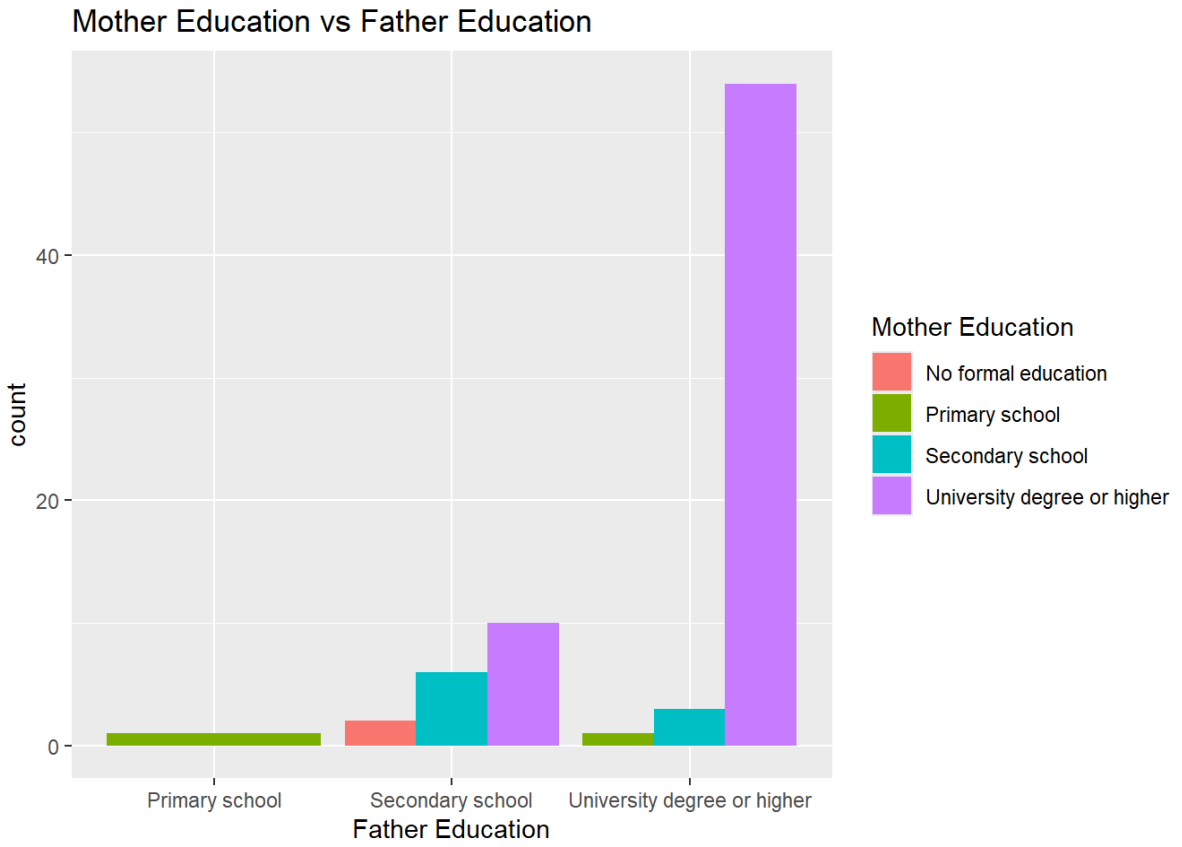


–Hypothesis testing using chi-squared test–

```
contingency_table <- table(Study_Habits$`Confidence in Academic success`, Study_Hab
its$Gender)
chi_squared_result <- chisq.test(contingency_table)
## Warning in stats::chisq.test(x, y, ...): Chi-squared approximation may be
## incorrect
print(chi_squared_result)
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 11.299, df = 4, p-value = 0.0234
```

- **Parental Education Levels** ($X^2 = 56.33$, $p < 0.001$): Highlighted strong correlations between fathers' and mothers' education, reflecting socio-economic patterns.

```
library(ggplot2)
custom_data <- data.frame(
  V1 = Study_Habits$`Parents Education Level`,
  V2 = Study_Habits$`Parents Education Level2`
)
ggplot(custom_data, aes(x = V1, fill = V2)) +
  geom_bar(position = "dodge") +
  labs(title = "Mother Education vs Father Education ",
       x = "Father Education",
       fill = "Mother Education",
       y = "count")
```



–Hypothesis testing using chi-squared test–

```
contingency_table <- table(Study_Habits$`Parents Education Level`, Study_Habits$`Pa
rents Education Level2`)

chi_squared_result <- chisq.test(contingency_table)

## Warning in stats::chisq.test(x, y, ...): Chi-squared approximation may be

## incorrect

print(chi_squared_result)

##

##  Pearson's Chi-squared test

##

## data:  contingency_table

## X-squared = 56.33, df = 6, p-value = 2.496e-10
```

Correlated Variables

1. **GPA vs Reviewing Notes**:
    - Evidence: Scatter plots, regression diagnostics, and ANOVA results all supported the relationship between note-reviewing frequency and GPA (R-squared = 0.1219).
    - Findings: Students who reviewed notes "Never" and "Often" achieved the highest GPAs, suggesting varying effects of reviewing frequency on performance.

2. **Homework's Time vs Studying Methods**:
    - Chi-squared analysis ($X^2 = 146.48$, df = 100, p = 0.0017) indicated a strong association.
    - Support: Bar plots visualized the differing study method preferences tied to time allocation.

3. **Confidence in Academic Success vs Gender**:
    - Evidence: Gender disparities were evident in bar charts, with significant differences confirmed by the chi-squared test ($X^2 = 11.299$, df = 4, p = 0.0234).

4. **Father's Education vs Mother's Education**:
    - Strong correlation revealed by the chi-squared test ($X^2 = 56.33$, df = 6, $p < 0.001$), supported by visuals showing interrelated educational patterns.

Summary of Further Analysis for GPA and frequency of reviewing notes

1. **Regression Diagnostics**:

Plots confirmed assumptions of normality and constant variance, validating regression results.

2. **Group Comparisons**:

Note-reviewing frequencies influenced GPA differently, with "Never" and "Often" groups outperforming others.

3. **Effect Size**:

Eta squared (0.1219) quantified the small to moderate effect of Reviewing Notes on GPA, emphasizing its relevance despite low R-squared values.

Conclusion

This analysis underscores the importance of study habits, such as note reviewing, on academic outcomes. While some relationships, like Reviewing Notes and GPA, were significant, the low explanatory power of regression models suggests additional factors influence academic performance. These findings highlight the need for targeted interventions to optimize study strategies and support diverse learning needs.

References

Dataset Source (Study_Habits Dataset. (2024). *Student self-reported academic behaviors and performance metrics*. Collected for statistical analysis. URL: https://forms.office.com/pages/responsepage.aspx?id=4FmN4ifxLUaWvZewJ_Cr4AaC5Ssij3BPpqE2Rlc7Jr1UMVM3R0tQRzcxQzBBT1AzUjBYS1pVM1JDMy4u&route=shorturl)

R Documentation (Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/)

Walck-Shannon, E. M., Rowell, S. F., & Frey, R. F. (2021). To What Extent Do Study Habits Relate to Performance?. CBE life sciences education, 20(1), ar6. https://doi.org/10.1187/cbe.20-05-0091