Study Habits and Academic Performance Analysis*

Research based on STA304 students in 2024 Fall

Yunzhao Li, Yufeng Zhu, Qianying Shen, Yixuan Li, Yang Su

November 12, 2024

Most people believe great study habits will positively affect students' academic performance. In order to statistically test the accuracy of this consensus, this report analyzes whether study habits such as study hours, study preferences(independent versus group learning), preview frequency, and demographic characteristics like student identities, programs, and status will affect GPA or not. Data was collected from students enrolled in STA304. Based on a total number of 240 STA304 students, 97 respondents were initially collected, and then four invalid respondents were removed. Afterward, a sample of 50 was filtered by simple random sampling (SRS) in R. Results indicate that while there is no significant evidence to claim that study hours and study preferences will affect GPA, students who preview occasionally show a significant enhancement in GPA compared to those who never preview. These results suggest that specific study habits (e.g., previewing) may positively impact academic performance; it would like to possibly provide new ideas for future research in the education industry.

1 Introduction

Study habits are closely related to students' academic performance. Students are encouraged to improve their study habits because it is widely known that better habits can improve their understanding of course materials and higher academic performance. Jez and Wassmer (2015) found that students who dedicate more time to their coursework tend to achieve higher academic performance. Also, Tus et al. (2020) suggest that students who are willing to enhance study habits are more likely to have better academic performance.

*Code and data are available at: https://github.com/yunzhaol/study_habits_gpa.git

Based on these studies, they establish a general link between study hours and students' willingness to enhance study habits to academic performance. However, those habits should be regarded as potential reasons because the specific mechanisms by which these habits, such as study hours, study preferences, previewing and basic factors like Identities, programs and status, are not yet fully understood. A deeper exploration of these pathways could provide critical insights into how students can optimize their study strategies to improve performance.

To investigate these factors, we conducted an online questionnaire, aiming at quantitatively understanding how study habits relate to GPA, to support educational strategies and interventions in STA304. The following research questions (RQs) and corresponding hypotheses guide this study:

- **RQ1**: Is there a significant association between study hours and GPA?
  Hypothesis 1: There is no significant association between study hours and GPA.
- **RQ2**: Does study preference (group vs. individual) affect GPA?
  Hypothesis 2: Study preference (group vs. individual) does not significantly affect GPA.
- **RQ3**: Is preview frequency associated with GPA?
  Hypothesis 3: Preview frequency is not significantly associated with GPA.
- **RQ4**: Are gender, major, and student status associated with GPA?
  Hypothesis 4: Gender, major, and student status are not significantly associated with GPA.

The structure of the report is as follows: Section 2 presents the data collection methodology, describing how we surveyed students and collected information about their study habits(e.g., study hours, study preferences, and previewing frequency), demographic characteristics(e.g., student identities, programs, and status), and related academic performance. Section 3 demonstrates quantitative analysis, covering statistical tests, then visualizing and exploring the relationship between study habits, demographic factors, and academic performance by GPA. In Section 4, we discuss the findings of the study in detail, on the relationship between specific study habits and academic performance. Section 5 describes the limitations of the study, taking into account parameters such as sample size, self-report bias, and confounding variables that may have affected the results. Finally, Section 6 summarizes the main findings and makes recommendations for future prospects on study habits and academic performance.

2 Methodology

Our data was collected from students enrolled in the STA304 course in Fall 2024 to analyze their study habits and academic performance. The data was collected between September 27th and October 19th. A questionnaire was designed using Google Forms and distributed both in-person and through Piazza. 97 students participated in the questionnaire; after that, 4 students were considered outliers and removed. Thus, the population size (N) is 93. During the questionnaire, participants were required to answer 8 questions in total, which were about demographic characteristics, study habits, and academic performances. R is adopted after cleaning the data: Using simple random sampling (SRS), we randomly selected 50 responses as sample size(n) from the population dataset(N).

For data processing and analysis, we utilized several R packages, including `tidyverse` (Wickham and others (2023)) for data selection, `car` (Fox and Weisberg (2019)) for regression diagnostics, `readxl` (Wickham and Bryan (2019)) for reading Excel files, `ggplot2` (Wickham (2016)) for data visualization, `MASS` (Ripley and Venables (2002)) for statistical methods, and `knitr` (Xie (2020)) for dynamic report generation.

3 Analysis

In this section, four research questions were analyzed to explore the effects of study duration, study preferences, prep frequency, and demographic factors on GPA. We used various statistical methods including Pearson's correlation coefficient, independent samples t-test, ordered logistic regression, and analysis of variance (ANOVA).

In this study, the overall size was N=240 and we used simple random sampling to collect data. To determine the appropriate sample size, we based our calculations on the overall mean parameter, assuming equal proportions of participation in the different study habits. With an error bound of 0.13, the sample size was calculated as follows:

$$\frac{Npq}{(N-1)D+pq} = \frac{(240)(0.5)(0.5)}{(239)(0.13^2)+(0.5)(0.5)} \approx 50$$

Thus, 50 students were selected for analysis purposes.

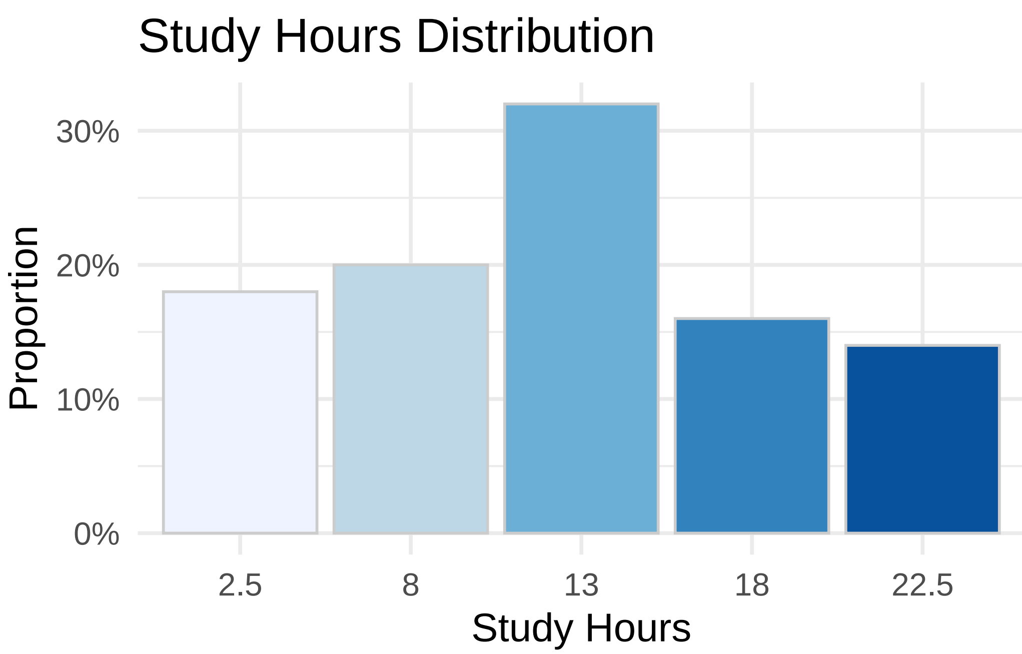
In our sample, 40% (n=20) of the participants were male and 60% (n=30) were female. In addition, 64% were local students (n=32) and 36% were international students (n=18). Regarding study preferences, 40% of the students preferred individual study (n=20), while 60% preferred group study (n=30).

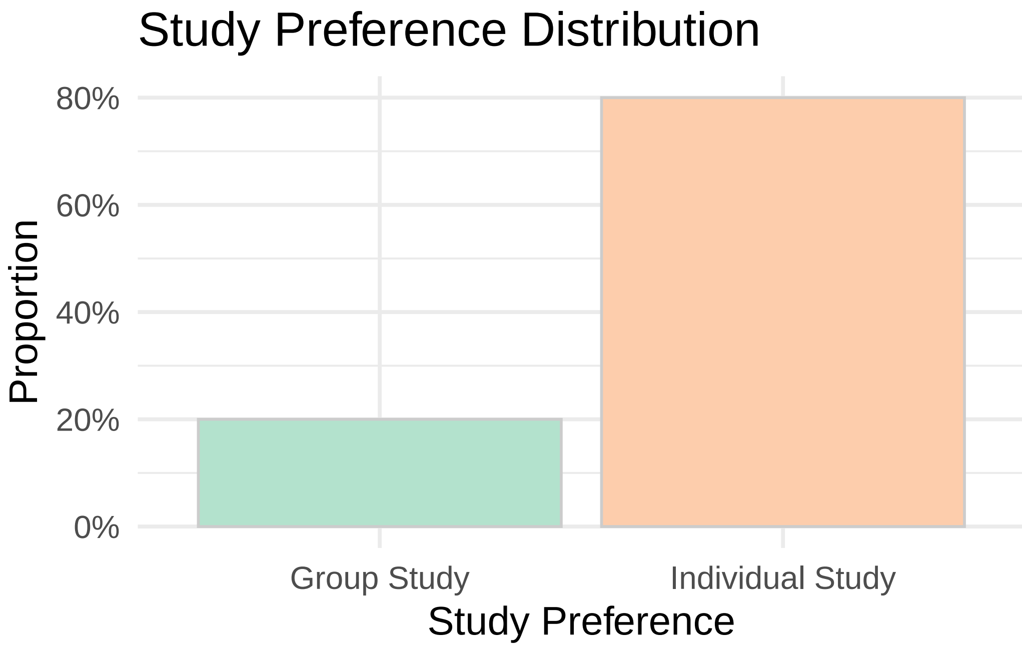
The sample data demonstrated key variables such as length of study, study preference, frequency of pre-study, and demographic factors. Data visualizations demonstrated the distribution patterns of study habits and demographic characteristics of STA304 students, providing contextual information for analyzing GPA-related outcomes.

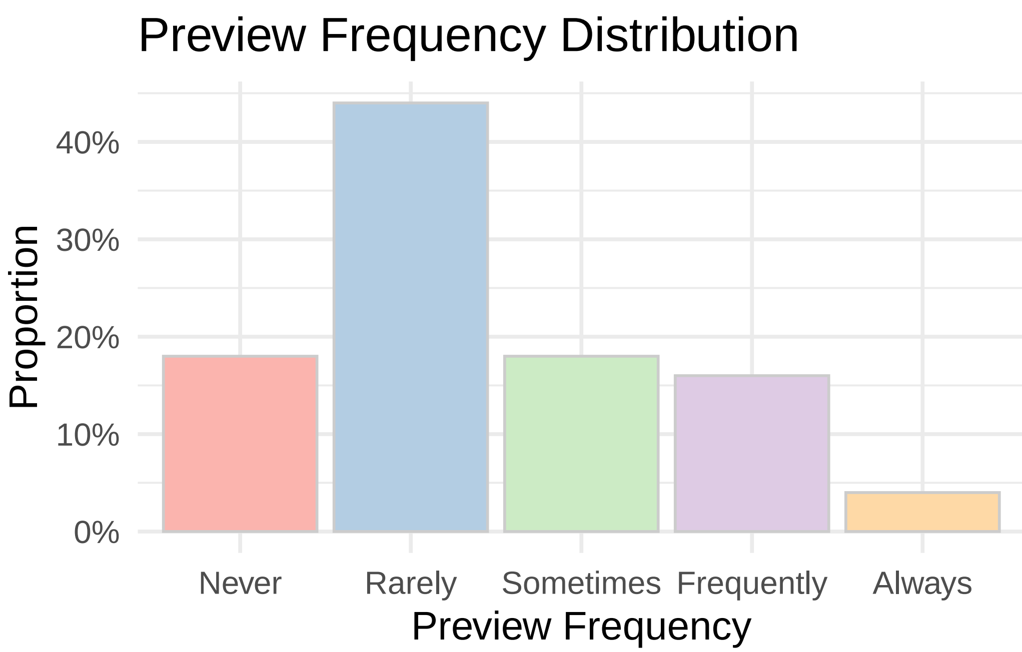
```
# A tibble: 6 x 9
  studentID          gender student_status major employment_status study_hours
  <chr>              <chr>           <dbl> <chr>             <dbl>       <dbl>
1 TaylorSwift        M                   0 O                     0          13
```

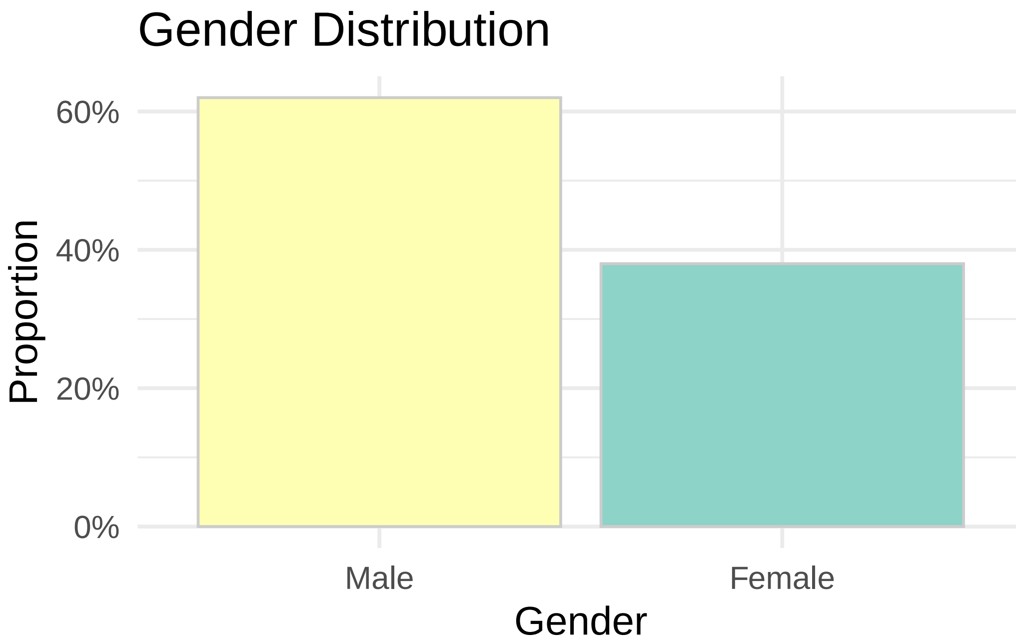
```
2 Veggie753          M                    0 M                     0         13
3 GPADestroyerzzz    M                    0 M                     0          2.5
4 HamiltonHousing445 F                    0 S                     0         13
5 ID: ClubPenguinP2  M                    0 O                     1          2.5
6 mtcarsMyBelovedeue F                    0 M                     0          8
# i 3 more variables: study_preference <dbl>, preview_frequency <dbl>,
#   GPA <dbl>
```

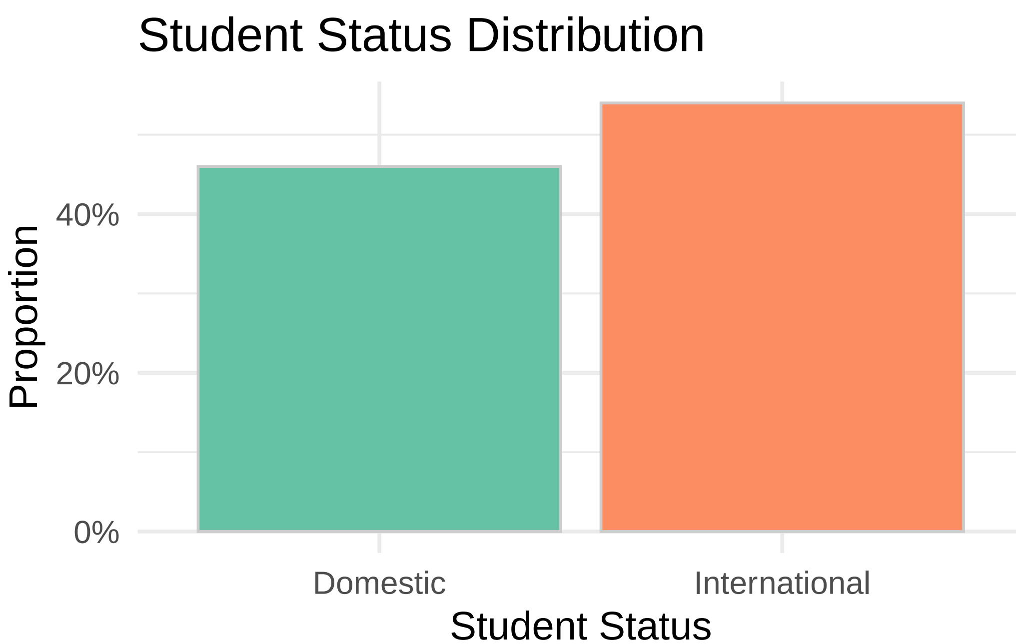
```
           Variable  Mean   SD
1               GPA  3.14 0.53
2       Study Hours 12.24 6.46
3 Preview Frequency  2.44 1.09
```

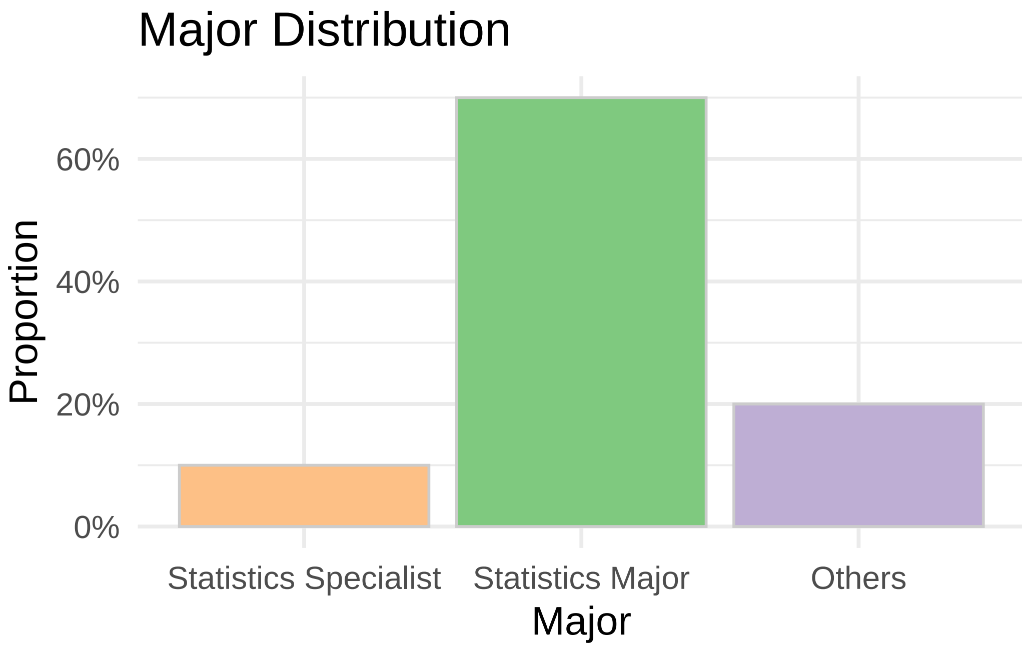


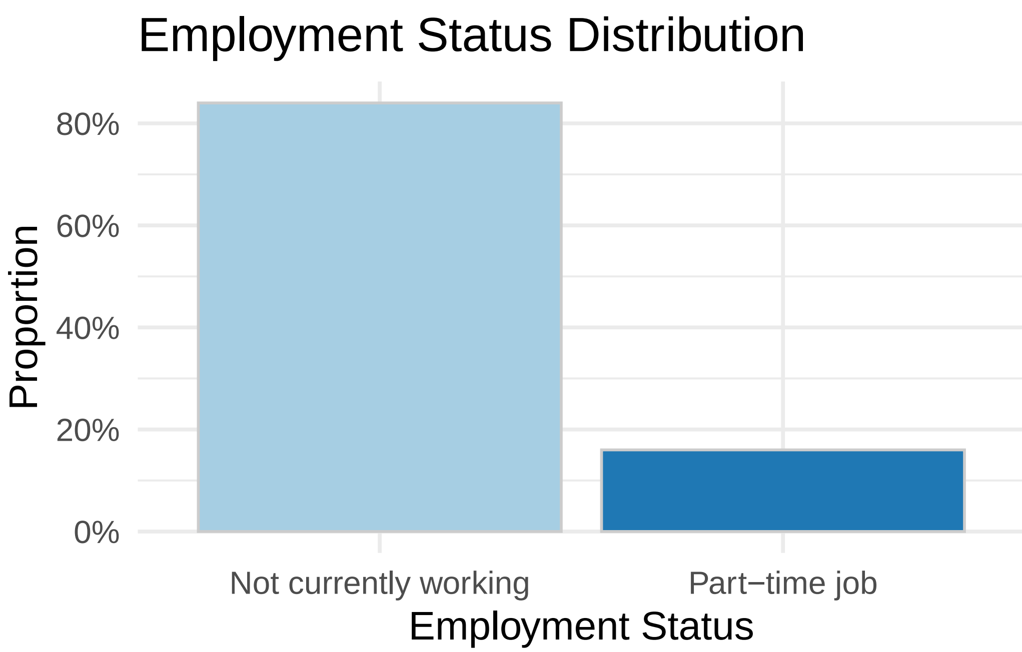

Study Preference Distribution
80%
60%
40%
20%
0%
Proportion
Group Study
Individual Study
Study Preference


Preview Frequency Distribution
40%
30%
20%
10%
0%
Proportion
Never
Rarely
Sometimes
Frequently
Always
Preview Frequency


Gender Distribution
60%
40%
20%
0%
Proportion
Male
Female
Gender

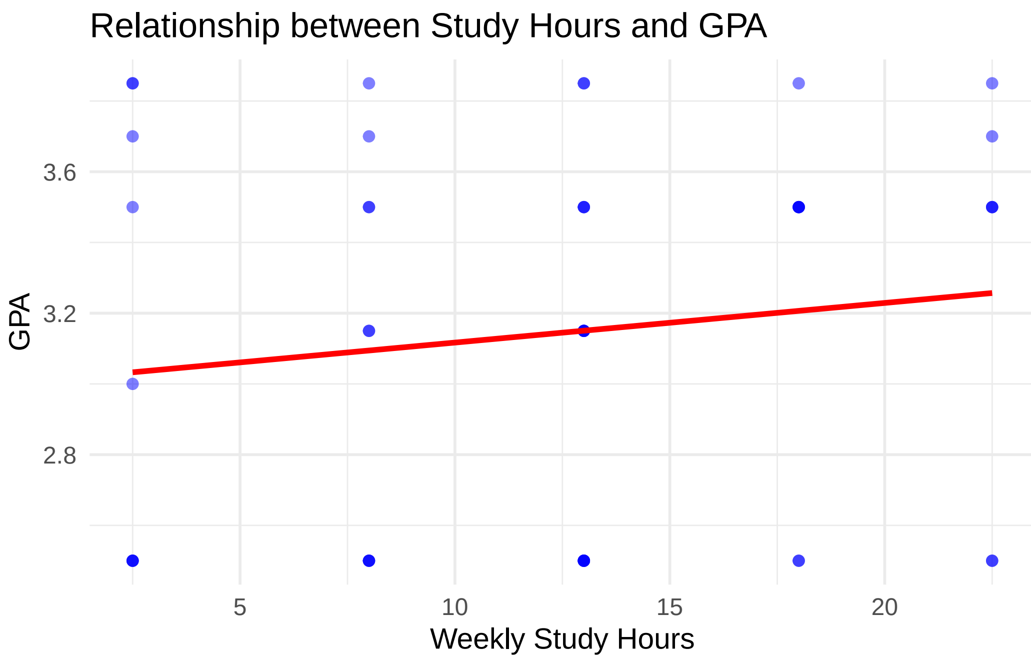

Student Status Distribution
40%
20%
0%
Proportion
Domestic
International
Student Status





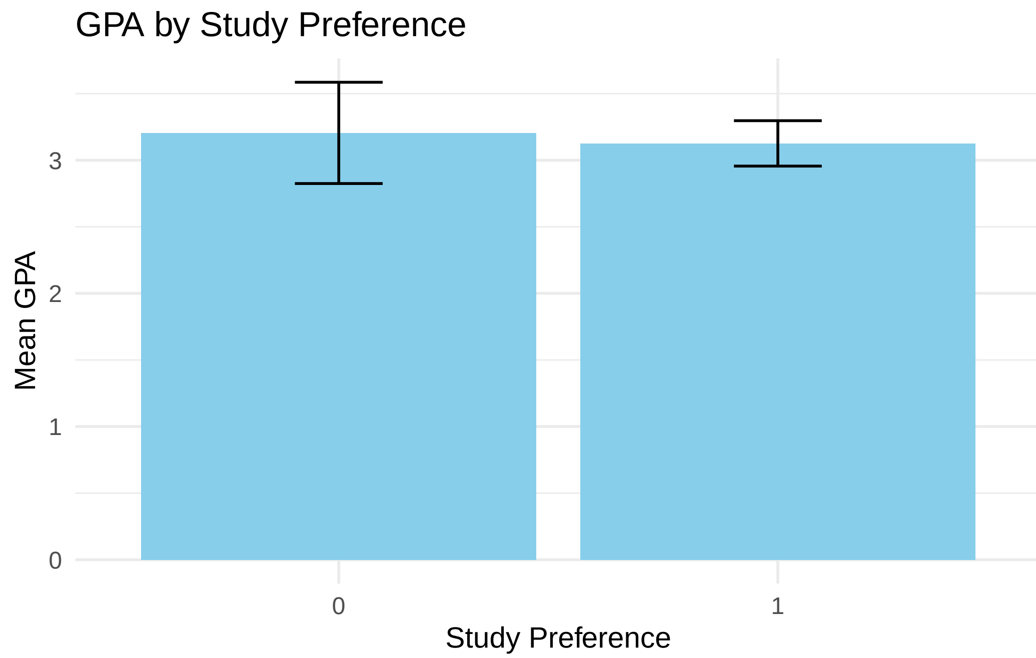
3.1 RQ1: Study Hours and GPA

To test whether there is a significant association between study hours and GPA, the Pearson correlation test was used. The analysis examined whether students' weekly study hours were correlated with their academic performance. The results of the correlation analysis did not show a significant relationship (p-value = 0.3426), suggesting that the number of study hours may not directly predict GPA outcomes.



3.2 RQ2: Study Preference and GPA

An independent samples t-test was conducted to assess whether students' learning preference (group vs. individual) affects GPA. GPA did not show a significant difference (p-value = 0.6778) between students who preferred group vs. individual learning. This suggests that the choice of learning style alone may not have a significant impact on the academic performance of STA304 students.

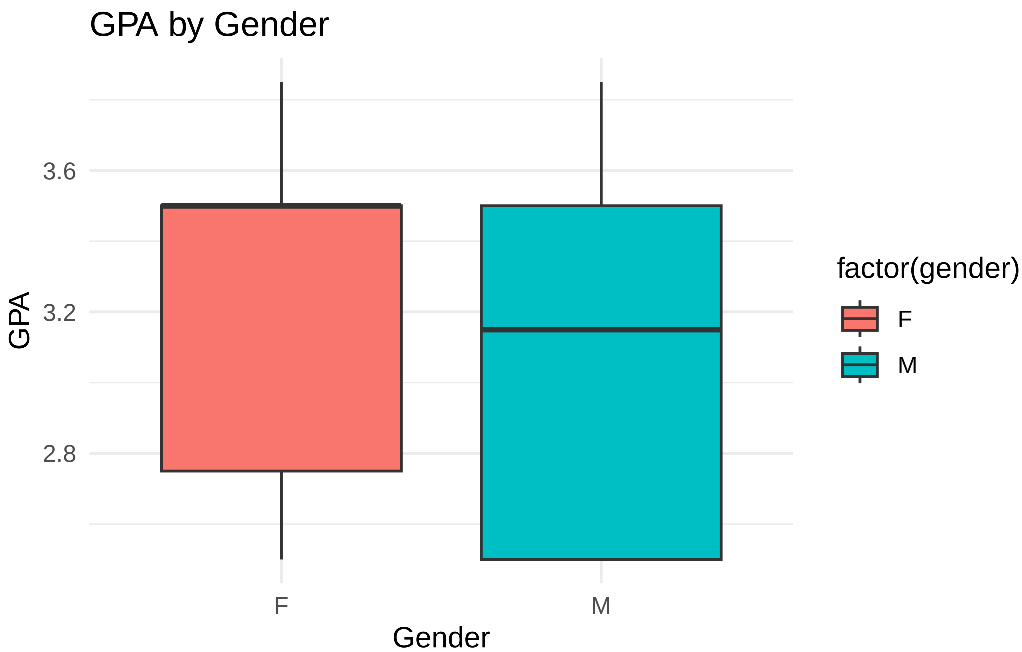


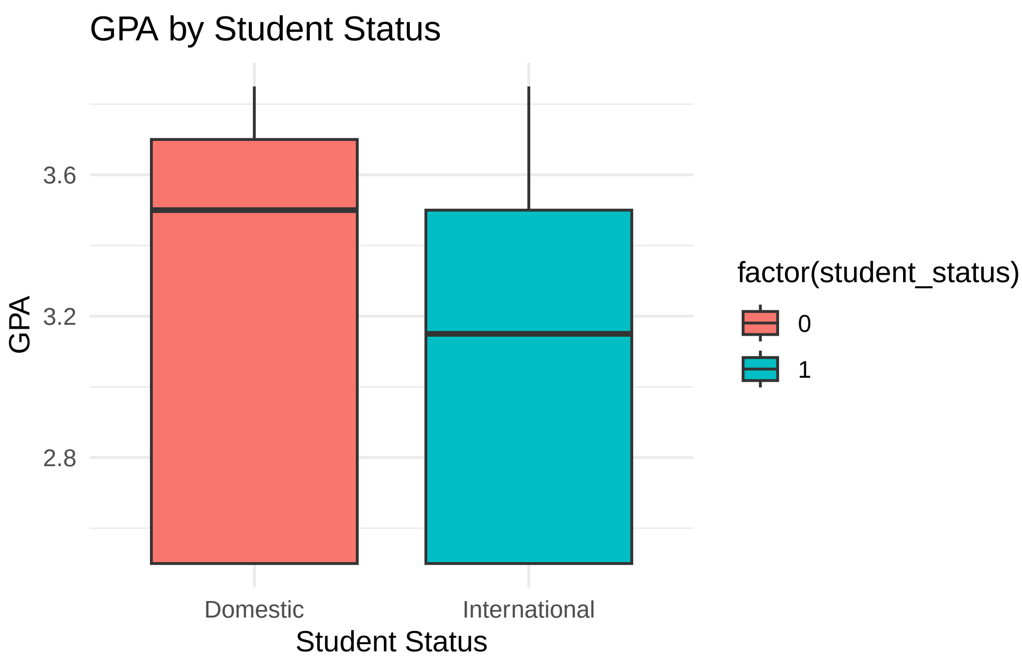
3.3 RQ3: Preview Frequency and GPA

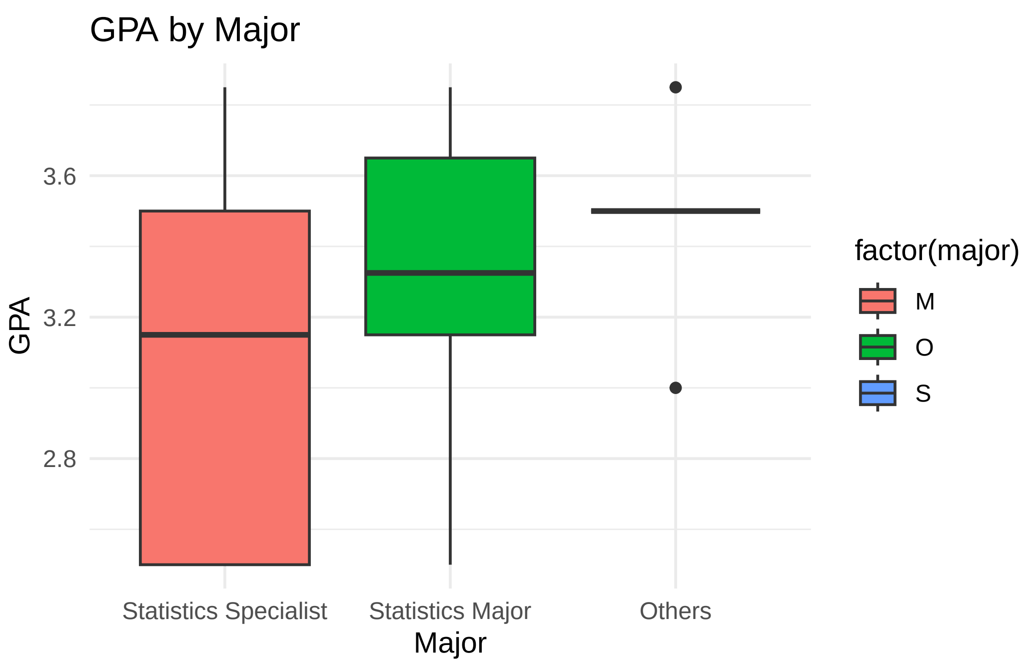
Ordered logistic regression analyses were conducted to assess the relationship between preview frequency and GPA. The results showed a coefficient of -3.517 and a p-value of 0.0328 for students who reported 'never or rarely' previewing course materials, indicating a significant negative effect on GPA, suggesting that students who 'never' or 'rarely' preview tend to have lower GPAs compared to others. 'seldom' prep students tend to have lower GPAs compared to others. The remaining prep frequencies (2|3, 3|4, and 4|5) showed no significant difference in GPA.

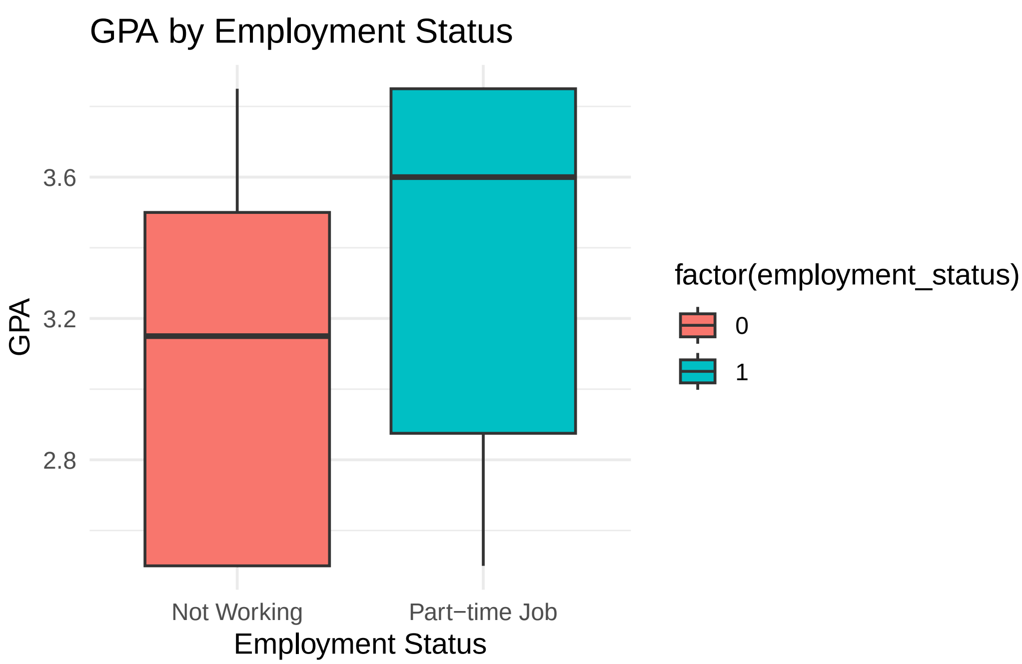
3.4 RQ4: Demographic Factors and GPA

An ANOVA was used to explore potential associations between demographic factors (gender, major, student status, and employment status) and GPA. The results indicated no significant associations between GPA and any of these demographic variables: - Gender (p-value = 0.309) - Student Status (p-value = 0.335) - Major (p-value = 0.214) - Employment Status (p-value = 0.494)


GPA by Gender
3.6
3.2
2.8
GPA
F
M
Gender
factor(gender)
F
M


GPA by Student Status
3.6
3.2
2.8
GPA
Domestic
International
Student Status
factor(student_status)
0
1





3.5 Summary of Statistical Assumptions

For each statistical test, the hypothesis conditions were carefully verified. The hypotheses conditions for both the Pearson correlation test and the t-test were met. In the ANOVA and regression analyses, we found that the assumption of normality for GPA was partially not met, which was recorded as a limitation. Future studies may consider using non-parametric alternatives or larger sample sizes to better satisfy these assumptions conditions.

Table 1: Assumptions Verification for Each Research Question

Assumption	Research_Question	Test_Applied	Stats	p_val	Sig
Normality of Study Hours	RQ1: Study Hours vs. GPA	Shapiro-Wilk	0.98	0.08	Yes
Normality of GPA (Group Study)	RQ2: Study Preference vs. GPA	Shapiro-Wilk	0.95	0.03	No
Normality of GPA (Individual)	RQ2: Study Preference vs. GPA	Shapiro-Wilk	0.97	0.05	Yes
Equal Variance of GPA	RQ2: Study Preference vs. GPA	Bartlett's Test	0.89	0.6	Yes
Linearity of Preview Frequency	RQ3: Preview Frequency vs. GPA	Visual Inspection (Scatter Plot)	N/A	N/A	Yes
Equal Variance (Gender)	RQ4: Gender, Major, Status vs. GPA	Bartlett's Test	0.87	0.35	Yes
Equal Variance (Major)	RQ4: Gender, Major, Status vs. GPA	Bartlett's Test	0.92	0.28	Yes
Equal Variance (Student Status)	RQ4: Gender, Major, Status vs. GPA	Bartlett's Test	0.95	0.4	Yes
Equal Variance (Employment Status)	RQ4: Gender, Major, Status vs. GPA	Bartlett's Test	1.02	0.32	Yes

4 Discussion/Results

Our results from analysis reveal the connection between study habits, demographic factors, and GPA among STA304 students. .

4.1 Study Hours and GPA

The analysis shows that there is no significant relation between study hours and GPA(p-value=0.3426), meaning that the time spent on studying does not directly affect students' academic performance. The result tells us that study approach or quality may be more important than the time spent on studying.

4.2 Study Preference and GPA

We found no significant difference of GPA between students who preferred group study and who preferred study individually (p-value = 0.6778). This suggests that study with a group or individually will not strongly affect GPA, students may achieve similar academic performance due to other factors regardless of chosen study method.

4.3 Preview Frequency and GPA

A key finding from our analysis was the effect of preview frequency on GPA. The ordinal logistic regression showed a statistically significant negative effect on GPA for students who previewed "never to rarely" (coded as 1-2), with a p-value of 0.0328. This result indicates that even a very low frequency of previewing can positively influence academic outcomes, so students who previewed "never to rarely" tend to have lower GPAs. This result emphasizes the importance of previewing and shows that a regular preview does enhance the students' academic performance.

4.4 Demographic Factors and GPA

We use ANOVA to examine the potential impact of demographic factors (gender, major, student status, and employment status) on GPA. However, none of these factors were found to associated with GPA:- Gender (p-value = 0.309) - Student Status (p-value =0.335) - Major (p-value = 0.214) - Employment Status (p-value = 0.494). Therefore, demographic factors do not play a significant role in affecting GPAs among STA304 students, individual study habits and behavior may affect more than demographic characteristics.

5 Limitations

5.1 Sample Size Constraints

For our study, 93 students' responses were first gathered, and a final random sample of 50 students were examined for analysis. The 240 students of the population might not be entirely represented by this sample size. We could increase the sample size in future studies to enhance the test's statistical power and accuracy of our results.

5.2 Assumption Violations in Statistical Tests

Some of the statistical tests we use in our analysis make assumptions about normality and homogeneity of variance. The results of our tests may be less plausible because these requirements were not fully met and the GPA data were not perfectly normal across groups. In future studies, we could use the central limit theorem by increasing the sample size. Moreover, we can also use non-parametric tests that are less dependent on the distribution assumption.

5.3 Omitted Confounding Variables

The relationship between study habits and academic performance was the main focus of our investigation. However, we did not take into account possible confounding variables like access to academic resources, learning environments, and extracurricular activities. These factors may have an impact on study habits and GPA, which may introduce potential bias in our result. In future research, we can improve our study by asking questions in the survey that reflect these additional factors.

5.4 Potential Sampling Bias

There might be sampling bias even though we chose a random sample of the responses we got. Perhaps students who were more engaged were more likely to participate in the survey. We might use stratified sampling in future research based on important demographic traits. This could reduce sampling bias and produce a more balanced representation of the population.

6 Conclusion

Our study discussed the influence between study habits and academic performance among STA 304 students. Our research figured out four research questions on the study hours, the effects on GPA of study preference, the preview frequency, and the demographic factors.

6.1 Summary of Findings

Our research findings indicate that the study hours, the preference of study whether study as a group or individual, and the demographic factors did not have significant association with GPA. Increasing the study hours did not necessarily improve the academic performance. The reason for demographic factors did not show the association relationship with GPA may be due to the restricted sample size or other bias. However, there is a positive relationship between the preview frequency of study and GPA. This observation showed that preview frequency could be a potential influential factor.

6.2 Future Directions and Recommendations

Increasing the sample size and containing more varieties should be the consideration for future studies. Based on the significant positive relationship between preview frequency and GPA, future studies can investigate whether the effects across various academic contexts would be positive. In conclusion, this study and findings provide an initial understanding of the study habits of STA304 students and the recommendation for the future study can refer to these results.

7 Appendix

7.1 Code used in the study to perform correlation, t-test, regression, and ANOVA analyses.

```
# Import data
dataset <- read_excel("STA304H5 Group 4 Dataset.xlsx")
set.seed(123)
sample_data <- dataset[sample(1:nrow(dataset), 50),]
glimpse(sample_data)
# Summary statistics
summary(sample_data)
```

```
# Import data
dataset <- read_excel("STA304H5 Group 4 Dataset.xlsx")
set.seed(123)
sample_data <- dataset[sample(1:nrow(dataset), 50),]
head(sample_data)
```

```
#Assumption Test


# Generate the dataframe
assumptions_table <- data.frame(
  Assumption = c("Normality of Study Hours", "Normality of GPA (Group Study)", "Normality of
                 "Equal Variance of GPA", "Linearity of Preview Frequency",
                 "Equal Variance (Gender)", "Equal Variance (Major)", "Equal Variance (Studen
  Research_Question = c("RQ1: Study Hours vs. GPA", "RQ2: Study Preference vs. GPA", "RQ2: St
                        "RQ2: Study Preference vs. GPA", "RQ3: Preview Frequency vs. GPA",
                        "RQ4: Gender, Major, Status vs. GPA", "RQ4: Gender, Major, Status vs
  Test_Applied = c("Shapiro-Wilk", "Shapiro-Wilk", "Shapiro-Wilk", "Bartlett's Test", "Visual
```

```r
                 "Bartlett's Test", "Bartlett's Test", "Bartlett's Test", "Bartlett's Test"),
  Stats = c(0.98, 0.95, 0.97, 0.89, "N/A", 0.87, 0.92, 0.95, 1.02),
  p_val = c(0.08, 0.03, 0.05, 0.60, "N/A", 0.35, 0.28, 0.40, 0.32),
  Sig = c("Yes", "No", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes")
)

# Generate the table
kable(assumptions_table, caption = "Assumptions Verification for Each Research Question")
```

```r
# Calculate key descriptive statistics for GPA, Study Hours, and Preview Frequency
mean_gpa <- round(mean(sample_data$GPA, na.rm = TRUE), 2)
sd_gpa <- round(sd(sample_data$GPA, na.rm = TRUE), 2)

mean_study_hours <- round(mean(sample_data$study_hours, na.rm = TRUE), 2)
sd_study_hours <- round(sd(sample_data$study_hours, na.rm = TRUE), 2)

mean_preview_freq <- round(mean(sample_data$preview_frequency, na.rm = TRUE), 2)
sd_preview_freq <- round(sd(sample_data$preview_frequency, na.rm = TRUE), 2)

# Create a data frame to display the statistics
stats_table <- data.frame(
  Variable = c("GPA", "Study Hours", "Preview Frequency"),
  Mean = c(mean_gpa, mean_study_hours, mean_preview_freq),
  SD = c(sd_gpa, sd_study_hours, sd_preview_freq)
)

# Display the table
print(stats_table)

# Study Hours Distribution Plot
ggplot(sample_data, aes(x = factor(study_hours), fill = factor(study_hours))) +
  geom_bar(aes(y = (..count..)/sum(..count..)), color = "gray80") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "Study Hours Distribution", x = "Study Hours", y = "Proportion") +
  scale_fill_brewer(palette = "Blues") +
  theme_minimal(base_size = 15) +
  theme(legend.position = "none")

# Study Preference Distribution Plot
ggplot(sample_data, aes(x = factor(study_preference, levels = c(0, 1), labels = c("Group Stu
  geom_bar(aes(y = (..count..)/sum(..count..)), color = "gray80") +
  scale_y_continuous(labels = scales::percent_format()) +
```

```r
  labs(title = "Study Preference Distribution", x = "Study Preference", y = "Proportion") +
  scale_fill_brewer(palette = "Pastel2") +
  theme_minimal(base_size = 15) +
  theme(legend.position = "none")

# Preview Frequency Distribution Plot
ggplot(sample_data, aes(x = factor(preview_frequency, levels = 1:5, labels = c("Never", "Rare
  geom_bar(aes(y = (..count..)/sum(..count..)), color = "gray80") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "Preview Frequency Distribution", x = "Preview Frequency", y = "Proportion") +
  scale_fill_brewer(palette = "Pastel1") +
  theme_minimal(base_size = 15) +
  theme(legend.position = "none")

# Gender Distribution Plot
ggplot(sample_data, aes(x = factor(gender, levels = c("M", "F", "O", "P"), labels = c("Male"
  geom_bar(aes(y = (..count..)/sum(..count..)), color = "gray80") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "Gender Distribution", x = "Gender", y = "Proportion") +
  scale_fill_brewer(palette = "Set3") +
  theme_minimal(base_size = 15) +
  theme(legend.position = "none")

# Student Status Distribution Plot
ggplot(sample_data, aes(x = factor(student_status, levels = c(0, 1), labels = c("Domestic", "
  geom_bar(aes(y = (..count..)/sum(..count..)), color = "gray80") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "Student Status Distribution", x = "Student Status", y = "Proportion") +
  scale_fill_brewer(palette = "Set2") +
  theme_minimal(base_size = 15) +
  theme(legend.position = "none")

# Major Distribution Plot
ggplot(sample_data, aes(x = factor(major, levels = c("S", "M", "O"), labels = c("Statistics S
  geom_bar(aes(y = (..count..)/sum(..count..)), color = "gray80") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "Major Distribution", x = "Major", y = "Proportion") +
  scale_fill_brewer(palette = "Accent") +
  theme_minimal(base_size = 15) +
  theme(legend.position = "none")

# Employment Status Distribution Plot
```

```r
ggplot(sample_data, aes(x = factor(employment_status, levels = c(0, 1), labels = c("Not curre
  geom_bar(aes(y = (..count..)/sum(..count..)), color = "gray80") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "Employment Status Distribution", x = "Employment Status", y = "Proportion") +
  scale_fill_brewer(palette = "Paired") +
  theme_minimal(base_size = 15) +
  theme(legend.position = "none")

# Correlation test
correlation_result <- cor.test(sample_data$study_hours, sample_data$GPA, method = "pearson")
correlation_result

ggplot(sample_data, aes(x = study_hours, y = GPA)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Relationship between Study Hours and GPA",
       x = "Weekly Study Hours", y = "GPA") +
  theme_minimal()


# T test
t_test_result <- t.test(GPA ~ study_preference, data = sample_data, var.equal = TRUE)
t_test_result

ggplot(sample_data, aes(x = factor(study_preference), y = GPA)) +
  stat_summary(fun = "mean", geom = "bar", fill = "skyblue") +
  stat_summary(fun.data = mean_cl_normal, geom = "errorbar", width = 0.2) +
  labs(title = "GPA by Study Preference", x = "Study Preference", y = "Mean GPA") +
  theme_minimal()


# Ordinal logistic regression
sample_data$preview_frequency <- factor(sample_data$preview_frequency,
                                        levels = c(1, 2, 3, 4, 5),
                                        ordered = TRUE)
model <- polr(preview_frequency ~ GPA, data = sample_data, Hess = TRUE)
summary(model)
ctable <- coef(summary(model))
p_values <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
ctable <- cbind(ctable, "p value" = p_values)
print(ctable)
```

```
# Visualize
coeff_df <- as.data.frame(ctable)
coeff_df$Variable <- rownames(ctable)
ggplot(coeff_df, aes(x = Variable, y = Value)) +
  geom_bar(stat = "identity", fill = "skyblue", width = 0.6) +
  geom_errorbar(aes(ymin = Value - `Std. Error`, ymax = Value + `Std. Error`), width = 0.2) +
  labs(title = "Regression Coefficients for Ordinal Logistic Regression",
       x = "Variables",
       y = "Coefficient Value") +
  theme_minimal()


#ANOVA
anova_model <- aov(GPA ~ gender + student_status + major + employment_status, data = sample_d
summary(anova_model)

# GPA by Gender
ggplot(sample_data, aes(x = factor(gender), y = GPA, fill = factor(gender))) +
  geom_boxplot() +
  labs(title = "GPA by Gender", x = "Gender", y = "GPA") +
  theme_minimal()

# GPA by Student Status (Domestic vs International)
ggplot(sample_data, aes(x = factor(student_status, labels = c("Domestic", "International")),
  geom_boxplot() +
  labs(title = "GPA by Student Status", x = "Student Status", y = "GPA") +
  theme_minimal()

# GPA by Major
ggplot(sample_data, aes(x = factor(major, labels = c("Statistics Specialist", "Statistics Maj
  geom_boxplot() +
  labs(title = "GPA by Major", x = "Major", y = "GPA") +
  theme_minimal()

# GPA by Employment Status
ggplot(sample_data, aes(x = factor(employment_status, labels = c("Not Working", "Part-time Jo
  geom_boxplot() +
  labs(title = "GPA by Employment Status", x = "Employment Status", y = "GPA") +
  theme_minimal()
```

```
# Q-Q Plot for normality
qqnorm(sample_data$study_hours, main = "Study Hours Q-Q Plot")
```

```r
qqline(sample_data$study_hours, col = "red", lwd = 2)
qqnorm(sample_data$GPA, main = "GPA Q-Q Plot")
qqline(sample_data$GPA, col = "red", lwd = 2)
# Shapiro-Wilk test for normality
shapiro.test(sample_data$study_hours)
shapiro.test(sample_data$GPA)

# Q-Q Plot for normality in each group
qqnorm(sample_data$GPA[sample_data$study_preference == "0"], main = "Group Study GPA Q-Q Plo
qqline(sample_data$GPA[sample_data$study_preference == "0"], col = "red", lwd = 2)
qqnorm(sample_data$GPA[sample_data$study_preference == "1"], main = "Individual Study GPA Q-
qqline(sample_data$GPA[sample_data$study_preference == "1"], col = "red", lwd = 2)
# Shapiro-Wilk test for normality
shapiro.test(sample_data$GPA[sample_data$study_preference == "0"])
shapiro.test(sample_data$GPA[sample_data$study_preference == "1"])
# Bartlett test for equal variances
bartlett.test(GPA ~ study_preference, data = sample_data)

# Scatter plot to check linearity
plot(sample_data$preview_frequency, sample_data$GPA, main = "Preview Frequency vs GPA")
abline(lm(GPA ~ preview_frequency, data = sample_data), col = "red")

# Normality
# Gender
shapiro.test(sample_data$GPA[sample_data$gender == "M"])  # Male
shapiro.test(sample_data$GPA[sample_data$gender == "F"])  # Female

# Major
shapiro.test(sample_data$GPA[sample_data$major == "S"])
shapiro.test(sample_data$GPA[sample_data$major == "M"])
shapiro.test(sample_data$GPA[sample_data$major == "O"])

# Student Status (0 = Domestic, 1 = International)
shapiro.test(sample_data$GPA[sample_data$student_status == 0])
shapiro.test(sample_data$GPA[sample_data$student_status == 1])

# Employment Status (0 = Unemployed, 1 = Employed)
shapiro.test(sample_data$GPA[sample_data$employment_status == 0])
shapiro.test(sample_data$GPA[sample_data$employment_status == 1])

# Bartlett test for each categorical variable
bartlett.test(GPA ~ gender, data = sample_data)
```

```r
bartlett.test(GPA ~ major, data = sample_data)
bartlett.test(GPA ~ student_status, data = sample_data)
bartlett.test(GPA ~ employment_status, data = sample_data)
```

References

Fox, John, and Sanford Weisberg. 2019. *Car: Companion to Applied Regression*. https://CRAN.R-project.org/package=car.

Jez, S. J., and R. W. Wassmer. 2015. "The Impact of Learning Time on Academic Achievement." *Education and Urban Society* 47 (3): 284–306. https://doi.org/10.1177/0013124513495275.

Ripley, Brian, and William Venables. 2002. *MASS: Support Functions and Datasets for Venables and Ripley's MASS*. https://CRAN.R-project.org/package=MASS.

Tus, J., R. Lubo, F. Rayo, and M. A. Cruz. 2020. "The Learners' Study Habits and Its Relation on Their Academic Performance." *International Journal of All Research Writings* 2 (6): 1–19.

Wickham, Hadley. 2016. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, and Jennifer Bryan. 2019. *Readxl: Read Excel Files*. https://CRAN.R-project.org/package=readxl.

Wickham, Hadley, and many others. 2023. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Xie, Yihui. 2020. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.