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 25, 2024

Study habits are key factors in students' academic performance. This paper examines the impact of specific study habits such as study hours, study preferences, and preview frequency—on GPA, along with demographic factors like gender, major, and student status. Out of a total population of 240 students in the STA304 course in 2024 Fall, data were collected from 97 respondents, with 4 invalid entries removed, and a final sample of 50 selected via simple random sampling in R. Results indicate that while study hours and study preferences for independent and group learning do not significantly affect GPA, students who preview occasionally show a significant positive impact on GPA compared to those who never preview. These findings suggest that even minimal previewing could enhance academic performance, highlighting potential areas for educational interventions and future research.

1 Introduction

Study habits are widely recognized as a key factor influencing students' academic performance. It is generally accepted that by improving their study habits, students can enhance their understanding of course materials, leading to better academic outcomes. Previous research has highlighted the importance of study habits; for example, Jez and Wassmer (2015) found that students who dedicate more time to their coursework tend to achieve higher academic performance. Similarly, Tus et al. (2020) suggested that students who actively improve their study habits are more likely to experience academic success.

While these studies establish a general link between study habits and academic performance, the specific mechanisms—such as study hours, study methods (group versus individual study),

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

and the frequency of previewing or reviewing materials are not yet fully understood. A deeper understanding of how these detailed parameters affect academic outcomes could provide valuable insights for educational strategies, allowing students to optimize their study approaches.

To investigate these factors, we conducted an online questionnaire survey in Fall 2024 with students in the STA304 course. This study aims to quantitatively examine the relationship between specific study habits and GPA to support the development of effective educational interventions. The following research questions (referred to as RQs) and corresponding hypotheses guide this study:

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

The structure of the paper is as follows: Section 2 outlines our data collection methods, describing the process by which we surveyed students and gathered information on their study habits, demographic characteristics, and academic performance. Section 3 presents our quantitative analysis, including statistical tests and visualizations to examine the relationships between study habits, demographic factors, and GPA. In Section 4, we discuss the findings in detail, focusing on whether and how specific study habits such as study hours, study preferences, and preview frequency relate to academic performance. Section 5 covers the limitations of our study, acknowledging factors such as sample size, self-reporting bias, and potential confounding variables that could impact our results. Finally, Section 6 concludes our analysis by summarizing key findings and suggesting directions for future research on the topic of 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 collection took place between September 27th and October 19th. A questionnaire was designed using Google Forms and distributed both in-person and through Piazza. We chose Simple Random Sampling (SRS) for this study, it ensures that every student in the STA304 population has an equal probability of being included in the sample. Compare to the other sampling methods, SRS makes the study easier to conduct

without further steps to divide different groups. We initially collected responses from 97 students; after removing 4 invalid responses without filling the id, we retained 93 valid entries. Each participant answered 8 questions covering basic information (e.g., identity, program, and academic performance), study hours, working hours, study preferences (individual or group), and preview frequency. After data cleaning, we used R to apply simple random sampling (SRS) to select a final sample of 50 responses from the dataset.

For data processing and analysis, we utilized several R packages, including `tidyverse` (Wickham and others (2023)) for data manipulation, `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

This section analyzes the four research questions to explore the impact of study hours, study preferences, preview frequency, and demographic factors on GPA. Various statistical methods were applied, including Pearson correlation, independent t-test, ordinal logistic regression, and ANOVA.

For our study, our population size is (N = 240), and we collected data using simple random sampling. To determine an appropriate sample size, we used a calculation focused on the mean population parameter, assuming an equal proportion among those who engage in various study habits. Given a bound of error of 0.13, the sample size calculation was as follows:

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

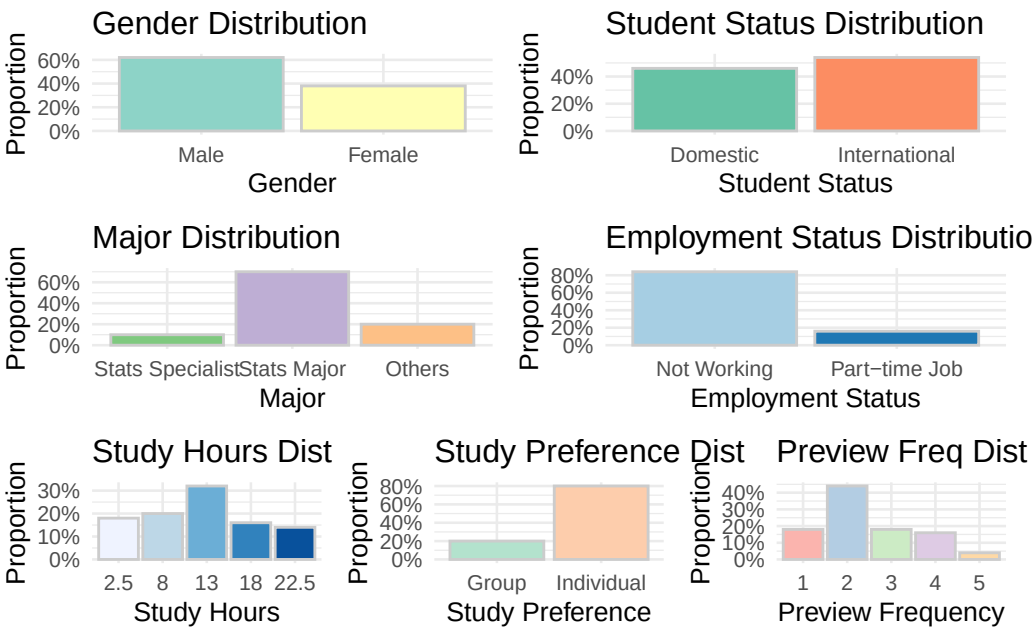
Thus, we sampled 50 students for our analysis.

Our sample included 40% male participants (n = 20) and 60% female participants (n = 30). Additionally, 64% were domestic students (n = 32), while 36% were international students (n = 18). For study preferences, 40% preferred individual study (n = 20), and 60% preferred group study (n = 30).

The sample data displays key variables like study hours, preferences, preview frequency, and demographic factors. Visualizations show distribution patterns for study habits and demographic characteristics among STA304 students. These insights provide context for analyzing GPA-related outcomes.

Variable	Mean	SD
GPA	3.14	0.53
Study Hours	12.24	6.46

Variable	Mean	SD
Preview Frequency	2.44	1.09



3.1 Summary of Statistical Assumptions

For each statistical test, assumptions were carefully verified. The Pearson correlation and t-test assumptions were satisfied. In the case of the ANOVA and regression analyses, we found that the assumption of normality was partially unmet for GPA, which was noted as a limitation. Future studies might consider using non-parametric alternatives or larger samples to better satisfy these assumptions.

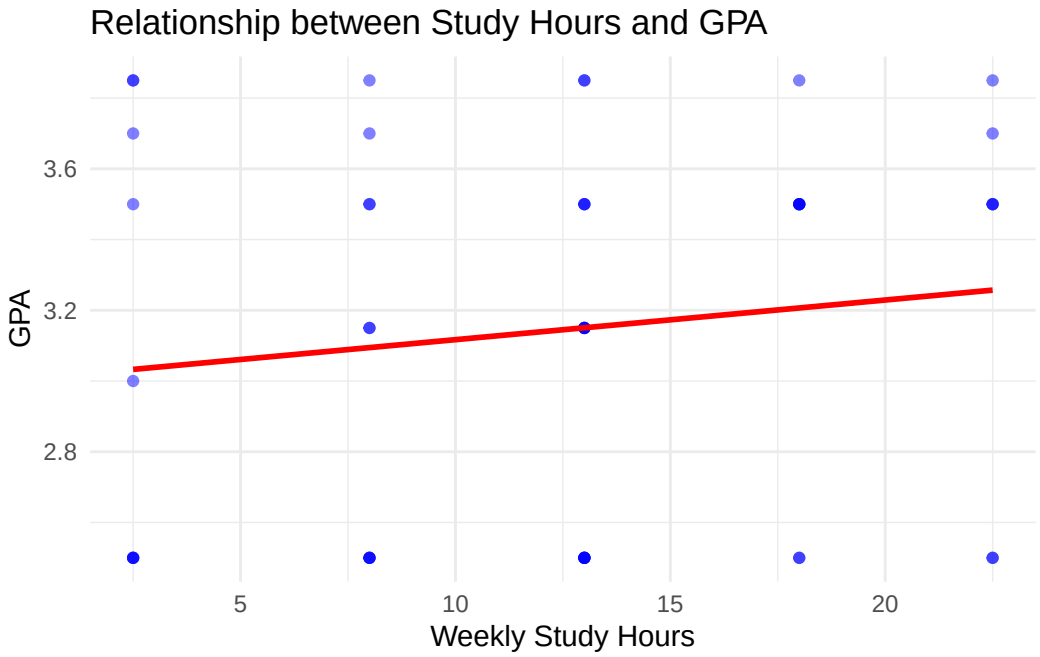
Table 2: Assumptions Verification for Each Research Question

Assumption	Research Question	Test Applied	Stats	PV	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

Assumption	Research Question	Test Applied	Stats	PV	Sig
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

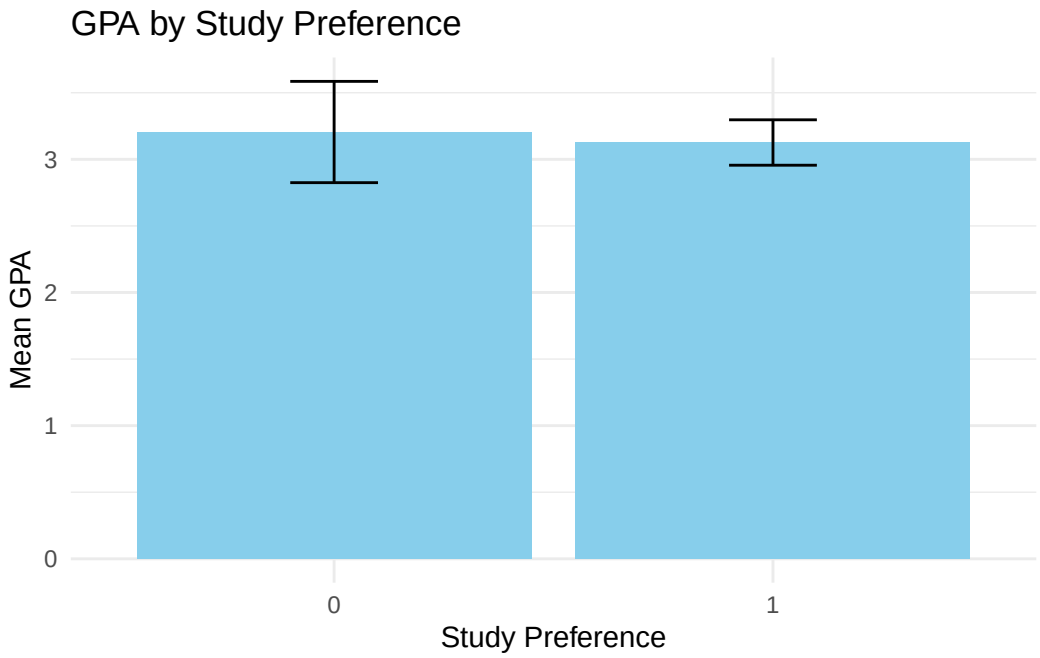
3.2 RQ1: Study Hours and GPA

To test whether there is a significant association between study hours and GPA, a Pearson correlation test was applied. The analysis examined if students' weekly study hours correlate with their academic performance. The results of the correlation analysis did not reveal a significant relationship between the two variables (p-value = 0.3426), suggesting that the amount of time spent studying may not directly predict GPA outcomes.



3.3 RQ2: Study Preference and GPA

We conducted an independent t-test to evaluate if students' study preference (group vs. individual) affects GPA. No significant difference was observed in GPA between those who preferred group study and those who favored individual study (p-value = 0.6778). This suggests that the choice of study method alone may not strongly influence academic performance among STA304 students.



3.4 RQ3: Preview Frequency and GPA

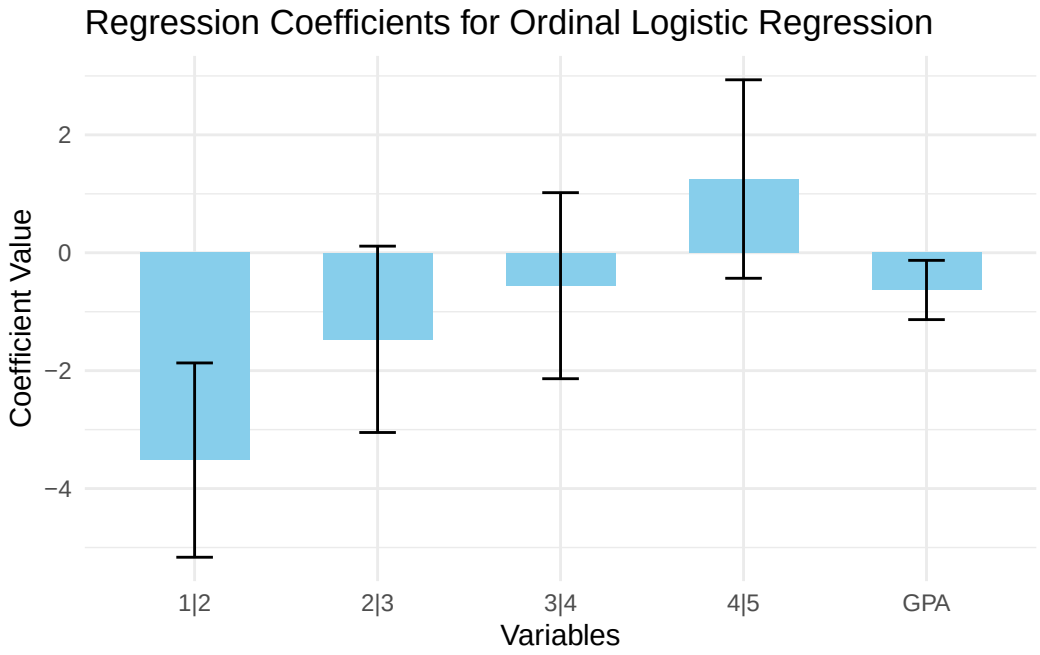
We use preview frequency to present how often a student preview upcoming lecture contents, from never to always, coded as 1 to 5. To assess the relationship between preview frequency and GPA, an ordinal logistic regression was conducted.

1|2, 2|3, 3|4, and 4|5 refer to the thresholds for transitioning between the ordered categories of the preview frequency levels. GPA represents the independent variable being analyzed for its relationship with the dependent variable.

The results showed a significant difference specifically for students who reported previewing course materials at a frequency of "never to rarely" (coded as 1-2), with a coefficient of -3.517 and a p-value of 0.0328, indicating a statistically significant negative impact on GPA. This suggests that students with very low preview frequency (i.e., those who preview "never" or "rarely") tend to have lower GPA outcomes compared to others. Other frequency levels (2|3, 3|4, and 4|5) did not show statistically significant differences in GPA.

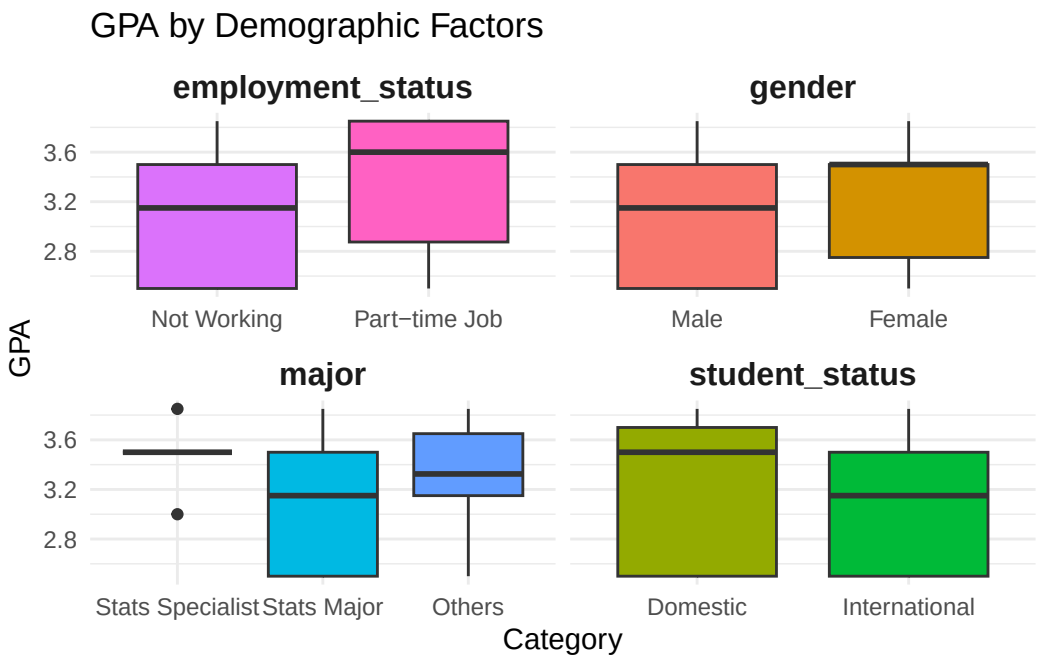
Table 3: Table: Ordinal Logistic Regression Results

Coefficient	Estimate	Std..Error	t.value	p.value
GPA	-0.631	0.503	-1.256	0.2091
1|2	-3.517	1.648	-2.134	0.0328
2|3	-1.468	1.580	-0.929	0.3528
3|4	-0.558	1.578	-0.354	0.7235
4|5	1.251	1.683	0.743	0.4576



3.5 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)



4 Discussion/Results

The results from our analysis reveal findings regarding the relationship between study habits, demographic factors, and GPA among STA304 students.

4.1 Study Hours and GPA

The correlation analysis between study hours and GPA did not show a significant association (p-value = 0.3426), suggesting that simply increasing the time spent studying does not necessarily translate to higher academic performance. This finding highlights that quality or method of study might be more crucial than the quantity of time spent.

4.2 Study Preference and GPA

Our analysis found no significant difference in GPA between students who preferred group study versus those who preferred individual study (p-value = 0.6778). This result suggests that study preference alone may not have a strong impact on GPA. It implies that students may achieve similar outcomes regardless of their study method, potentially due to other factors.

4.3 Preview Frequency and GPA

One of the more notable findings emerged from our analysis of preview frequency. The ordinal logistic regression showed a statistically significant negative impact on GPA for students with a preview frequency of "never to rarely" (coded as 1-2), with a p-value of 0.0328. This result indicates that even minimal previewing behavior can positively influence academic outcomes, as students who engaged in "never to rarely" previewing tended to have lower GPAs compared to their peers who previewed more frequently. This finding emphasizes the potential value of regular previewing in reinforcing understanding and supporting higher academic performance.

4.4 Demographic Factors and GPA

An ANOVA was conducted to examine the potential impact of demographic factors (gender, major, student status, and employment status) on GPA. None of these factors were found to be significantly 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)

These results suggest that demographic factors do not play a substantial role in determining GPA among this sample of STA304 students. This may imply that academic performance within this group is influenced more by individual study behaviors rather than by demographic characteristics.

5 Limitations

5.1 Sample Size Constraints

Our study included responses from 93 students, with a final random sample of 50 students used in analysis. This sample size may not fully capture the characteristics of the larger population of 240 students. An increased sample size could improve the precision of our findings and the statistical power of our tests.

5.2 Assumption Violations in Statistical Tests

Some statistical tests in our analysis assume normality and homogeneity of variances, conditions which were not fully met, particularly regarding the normality of GPA distributions across groups. This may affect the accuracy of the parametric test results. Future studies could address this limitation by either increasing the sample size to rely on the Central Limit Theorem for normality or by using non-parametric tests that are less sensitive to distributional assumptions.

5.3 Omitted Confounding Variables

Our survey focused on study habits and academic performance, but it did not account for potential confounding factors such as extracurricular activities, study environments, or access to academic resources. These factors could influence both study habits and GPA, introducing potential bias in our findings.

5.4 Potential Sampling Bias

Although our sampling method was random within the collected responses, it may still have introduced sampling bias. Students who are more diligent may have been more likely to participate. To reduce sampling bias in future studies, stratified sampling based on other demographic characteristics, for example, stratify based on Class Participation Frequency could provide a more balanced representation of the population.

6 Conclusion

Our study explored the relationship between study habits and academic performance among STA304 students, addressing four research questions on study hours, study preference, preview frequency, and demographic factors.

6.1 Summary of Findings

Our findings indicate no significant association between total study hours and GPA, suggesting that increasing study time alone may not improve academic performance. Study preference (group vs. individual) also showed no significant impact on GPA. However, preview frequency displayed a notable positive relationship with GPA, highlighting it as a potentially influential factor. Demographic factors like gender, major, and student status did not show significant associations with GPA, though this may be limited by sample size or other bias.

6.2 Future Directions and Recommendations

Future studies should consider increasing sample size and including more confounding variables, such as extracurricular activities and study environments, to gain a more comprehensive understanding. While non-parametric tests could be employed now to address violations of normality assumptions, this study prioritized parametric methods to maintain consistency and comparability with existing research. Given the significant result for preview frequency, further investigation into its impact across different academic contexts would be beneficial.

In conclusion, our findings offer initial insights into the study habits of STA304 students, with recommendations for future research to expand upon 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 GPA (Individual)",
                 "Equal Variance of GPA", "Linearity of Preview Frequency",
                 "Equal Variance (Gender)", "Equal Variance (Major)",
                 "Equal Variance (Student Status)",
                 "Equal Variance (Employment Status)"),
  Research_Question = c("RQ1: Study Hours vs. GPA",
                        "RQ2: Study Preference vs. GPA",
                        "RQ2: Study Preference vs. GPA",
                        "RQ3: Preview Frequency vs. GPA",
                        "RQ4: Gender, Major,
                        Status vs. GPA",
                        "RQ4: Gender, Major, Status vs. GPA",
```

```
                        "RQ4: Gender, Major, Status vs. GPA",
                        "RQ4: Gender, Major, Status vs. GPA"),
  Test_Applied = c("Shapiro-Wilk", "Shapiro-Wilk", "Shapiro-Wilk",
                   "Bartlett's Test", "Visual Inspection (Scatter Plot)",
                "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")
```

```
# 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 Study", "Individual Study")),
                        fill = factor(study_preference))) +
  geom_bar(aes(y = (..count..)/sum(..count..)), color = "gray80") +
  scale_y_continuous(labels = scales::percent_format()) +
  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", "Rarely",
                                              "Sometimes",
                                              "Frequently",
                                              "Always")),
                        fill = factor(preview_frequency))) +
  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", "Female",
                                              "Others",
                                              "Prefer not to disclose")),
                        fill = factor(gender))) +
  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") +
```

```r
  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", "International")),
                        fill = factor(student_status))) +
  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
                                              Specialist", "Statistics
                                              Major",
                                              "Others")),
                        fill = factor(major))) +
  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
ggplot(sample_data, aes(x = factor(employment_status,
                                   levels = c(0, 1),
                                   labels = c(
                                     "Not currently working",
                                     "Part-time job")),
                        fill = factor(employment_status))) +
  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") +
```

```r
  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)
```

```r
# 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_data)
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")),
                        y = GPA,
                        fill = factor(student_status))) +
  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 Major",
                                                     "Others")),
                        y = GPA,
                        fill = factor(major))) +
```

```
  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 Job")),
y = GPA,
fill = factor(employment_status))) +
  geom_boxplot() +
  labs(title = "GPA by Employment Status",
       x = "Employment Status", y = "GPA") +
  theme_minimal()
```

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

```r
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)
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.