Final Project Report

PSTAT100: Data Science Concepts and Analysis

STUDENT NAME

- Xinyao Song (xinyaosong)
- Matthew Chan (hchan837)
- Darren Colianni (dcolianni)
- Nic Chan (nicholaschan)
- STUDENT 5 (NetID 5)

1 Abstract

This project investigates how various social, familial, and behavioral factors are correlated with academic performance and average free times among Portuguese high school students. Using the Student Performance dataset compiled by Cortez and Silva in 2008, we explore three specific research questions: the first two concerning the influence of parental education and alcohol consumption on students' final Portuguese grades (G3), and the last one focusing on how one's relationship status influence the amount of free time students report. From the result, we can see that [FINISH AT THE END]

2 Introduction

This project focuses on how different factors influence school performance in Portuguese secondary school students. The dataset our team used is the Student Performance dataset compiled by Cortez and Silva in 2008. The dataset contains 649 observations and 30 features collected through questionnaires from two Portuguese high schools, with features covering aspects of family background, social environment, academic records, and types of lifestyles. The primary objective of our analysis is to examine how different social and behavioral factors relate to the final Portuguese grade (G3), which is measured on a scale from 0 to 20. By answering the three focused research questions below, we aim to understand how various background and behavioral variables correlate with or predict academic performance. The results from this analysis could help inform educational strategies and potentially help students with their stress and mental health.

2.1 Research Question 1

After statistically controlling for weekly study time (on a 1–4 numeric scale where 1 indicates less than 2 hours and 4 indicates more than 10 hours) and internet access (coded as "yes" or "no"), how is the father's highest education level (on a 0–4 numeric scale where 0 indicates no formal education and 4 indicates higher education) associated with Portuguese high school students' final grades in Portuguese class (G3), which are measured on a 0–20 numeric scale?

Hypothesis:

H_0: After controlling for weekly study time and internet access, the father's highest education level (0 = no formal education to 4 = higher education) has no significant association with students' final Portuguese grade (G3, measured on a 0–20 scale).

H_1: After controlling for weekly study time and internet access, the father's highest education level (0 = no formal education to 4 = higher education) is significantly positively associated with students' final Portuguese grade (G3, 0–20 scale).

2.2 Research Question 2

Based on the number of alcoholic beverages consumed by Portuguese high school students per weekday and weekend (on a 1-5 numeric scale where 1 indicates very low and 5 indicating very high), how does it affect Portuguese high school students' final grades in Portuguese class (G3), which are measured on a 0–20 numeric scale?

Hypothesis:

H_0: There is no significant association between students' alcohol consumption levels during weekdays and weekends (1 = very low to 5 = very high) and their final Portuguese grade (G3, 0–20 scale).

H_1: Students' alcohol consumption levels during weekdays and weekends (1 = very low to 5 = very high) are significantly associated with their final Portuguese grade (G3, 0–20 scale).

2.3 Research Question 3

Is there a statistically significant difference in reported average free time (on a 1-5 categorical scale where 1 indicates very low and 5 indicates very high) between Portuguese high school students who are in romantic relationships (coded as 1), and those who are not in romantic relationships (coded as 0)?

Hypothesis:

H_0: There is no significant difference in average free time between students in a romantic relationship and those who are not.

H_1: There is a significant difference in average free time (1 = very low to 5 = very high) between students in a romantic relationship and students who are not in romantic relationships.

3 Data Processing

The Student Performance dataset does not contain any missing values or fields with unclear meanings. Therefore, no data cleaning or preprocessing was necessary for this project.

4 Research Question 1

4.1 Modeling Process

The aim is to observe if, after controlling for study time and internet access, father's education significantly influenced the student's final Portuguese grade. The question at hand calls for a partial significance test, which works by having one one the models be a subset to another. Therefore, if there is a significant difference, it is due to the predictors added in between.

The models were both linear and used the following predictors:

- Model 1: Study time and internet access.
- Model 2: Model 1's predictors as well as the education level of the student's father.

The two linear models were compared using an ANOVA test. Since the only predictor added in model 2 was the father's education, if there was a significant decrease in residual variance–as indicated by the resulting p-value–it would be due to this predictor.

4.2 Results

Here were the results:

Table 1: ANOVA Comparison Table

Model	Res.Df	RSS	Df	Sum.of.Sq	F	Pr..F.	Signif
1	646	6207.3					
2	645	5995.9	1	211.42	22.74	2.30e-06	***

4.3 Interpretation

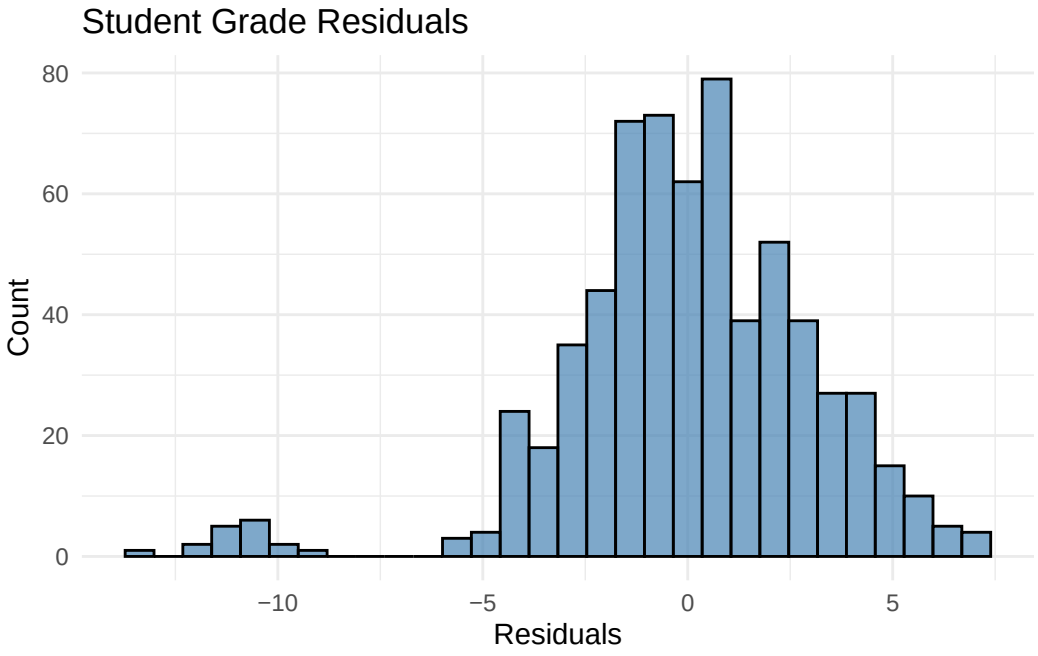
With a p-value of approximately $2.30 * 10^{-6}$, there is sufficient evidence to reject the null hypothesis,
H_0: that after controlling for weekly study time and internet access, the father's highest education level has no significant association with students' final Portuguese grade. Therefore, opting instead for the alternative hypothesis, the father's highest education level does play a role in the students' final Portuguese grade.

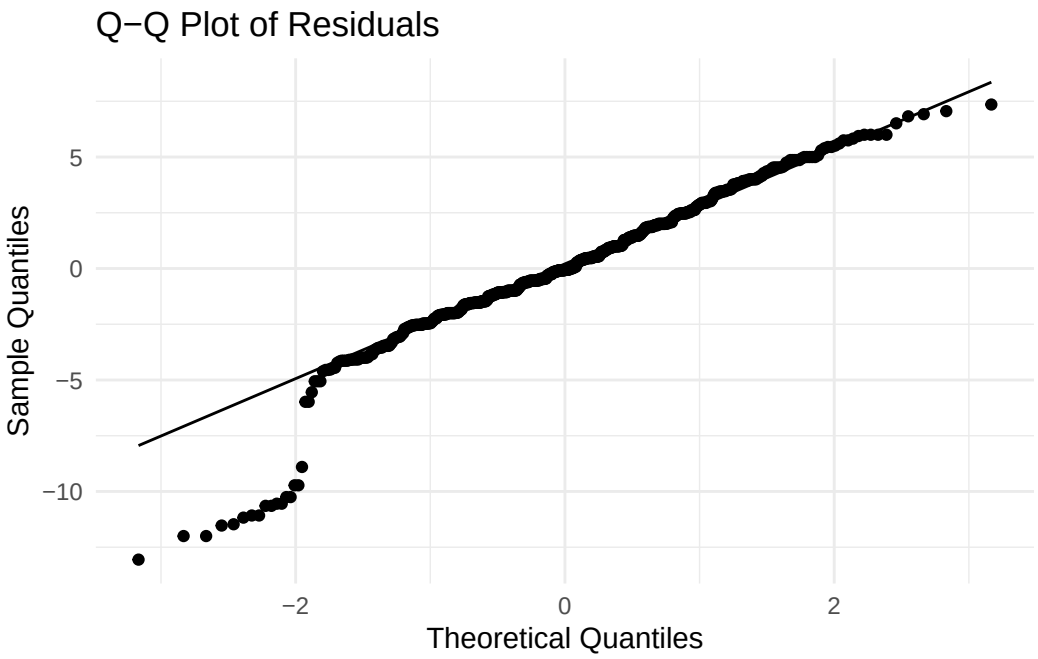
4.4 Visualization and Communication (Model Checking)

Model checking was performed to ensure this result from the previous section was accurate. To assess whether this model is appropriate, we examined Model 2, which contains the same predictors as Model 1 in addition to father's education. For the partial significance test to be valid, Model 2 must satisfy the standard linear regression assumptions: independence, no structure in residuals, normality of residuals, and homoscedasticity (equal variance across fitted values).

Independence is satisfied because the dataset includes the full population of students from each school during that year. Additionally, because this is an observational study with no temporal or treatment structure, residual independence is expected and visually confirmed. It only remains to be seen whether the remaining conditions—normality and equal variance—hold true.

4.4.1 Normality of Residuals

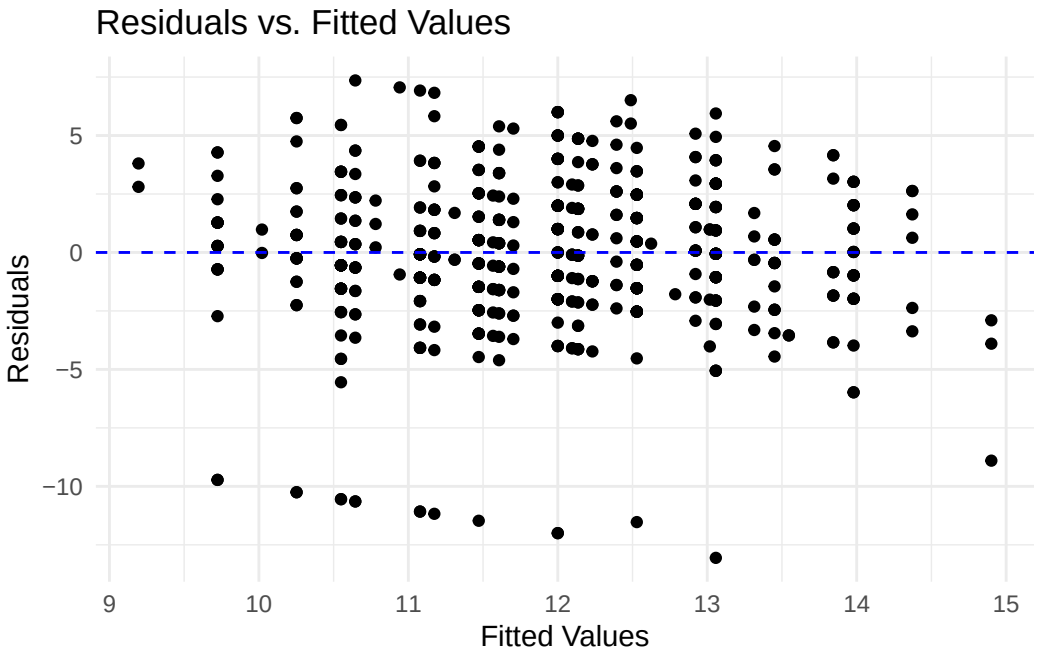




The Q-Q plot indicates a deviation from normality in the residuals, and likewise, a Shapiro-wilk test result of 2.83×10^{-15}(p-value < 0.05) corroborates this deviation. Nevertheless, as evident through the histogram, the deviation is primarily due to a relatively small group of students who scored significantly lower than predicted, causing a mild left skew. These values represent approximately 2.62% of the residuals. Aside from this small subset, the residuals are normally distributed and centered around zero.

Given the large sample size and that the deviation stems from a small portion of the data, the normality assumption is considered reasonably satisfied for inference purposes.

4.4.2 Equal Variances Across Fitted Values



The residuals appear to have approximately constant variance across the range of fitted values, with no clear funneling or systematic spread. Thus, there is no strong evidence to reject the homoscedasticity assumption.

Hence, all assumptions for the partial significance test performed were reasonably satisfied. Therefore, the previously found p-value of approximately $2.30 * 10^{-6}$ can be trusted.

4.5 Conclusion and Recommendation

After controlling for study time and internet access, the father's education level remains a significant positive predictor of students' final Portuguese grades. This likely reflects the academic support that more educated parents are able to provide at home; eg, helping with homework and explaining difficult topics. To support those lacking this advantage, schools should consider expanding tutoring and afterschool programs to foster a more equitable learning environment. Additionally, making greater use of school libraries—through activities like book clubs or guest author visits—could help motivate students to read more and strengthen their Portuguese language skills.

5 Research Question 2

5.1 Modeling Process

To investigate the association between alcohol consumption and students' final Portuguese grades (`G3`), we employed a two-way ANOVA. This method allows us to examine the main effects of weekday alcohol consumption (Dalc) and weekend alcohol consumption (`Walc`) on `G3`, as well as their interaction effect. The model is specified as `G3 ~ Dalc * Walc`, which includes both main effects and the interaction term. Both Dalc and Walc are treated as ordered factors on a 1-5 scale, representing consumption levels from very low to very high.

5.2 Results

The ANOVA test was performed to assess the significance of weekday and weekend alcohol consumption, and their interaction, on the final grade. The results are summarized in the table below.

Table 2: ANOVA for Alcohol Consumption's Effect on Final Grades

Term	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.	Signif
Dalc	4	327.51	81.880	8.282	0.0000	***
Walc	4	40.11	10.030	1.014	0.3992	
Dalc:Walc	14	206.78	14.770	1.494	0.1077	
Residuals	626	6188.90	9.886	NA	NA	

5.3 Interpretation

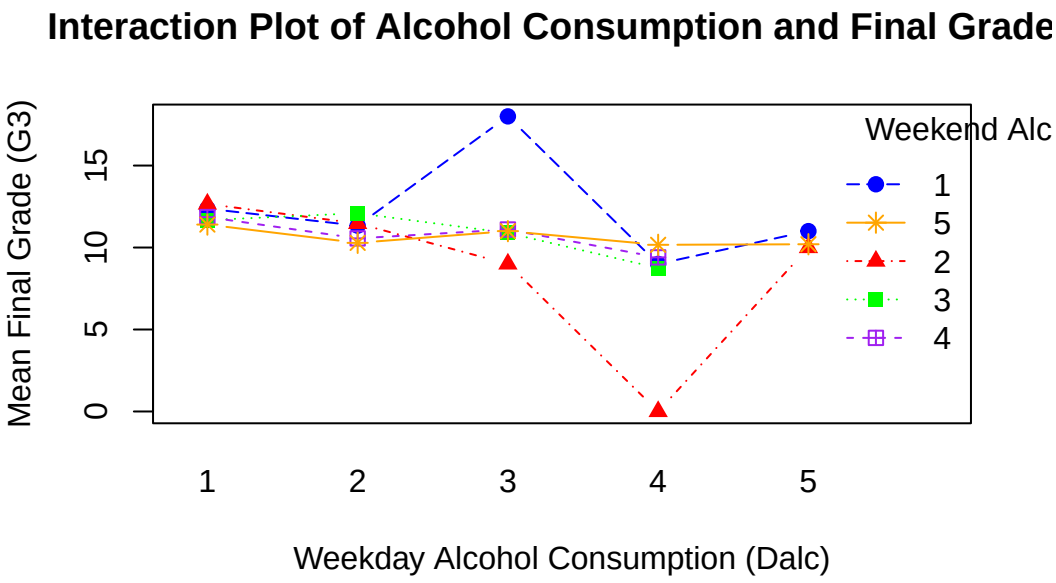
The results from the two-way ANOVA indicate that weekday alcohol consumption (`Dalc`) has a statistically significant main effect on students' final grades (`G3`), with a p-value of 0.0000. This suggests that higher levels of weekday drinking are significantly associated with differences in academic performance.

In contrast, weekend alcohol consumption (`Walc`) does not have a statistically significant effect on grades, with a p-value of 0.3992, which is well above the conventional 0.05 threshold. Therefore, weekend drinking alone does not show a significant association with final grades.

The interaction term `Dalc:Walc` has a p-value of 0.1077, which is also not statistically significant. This indicates that the effect of weekday alcohol consumption on grades does not significantly differ across levels of weekend alcohol consumption, and vice versa.

5.4 Visualization and Communication (Model Checking)

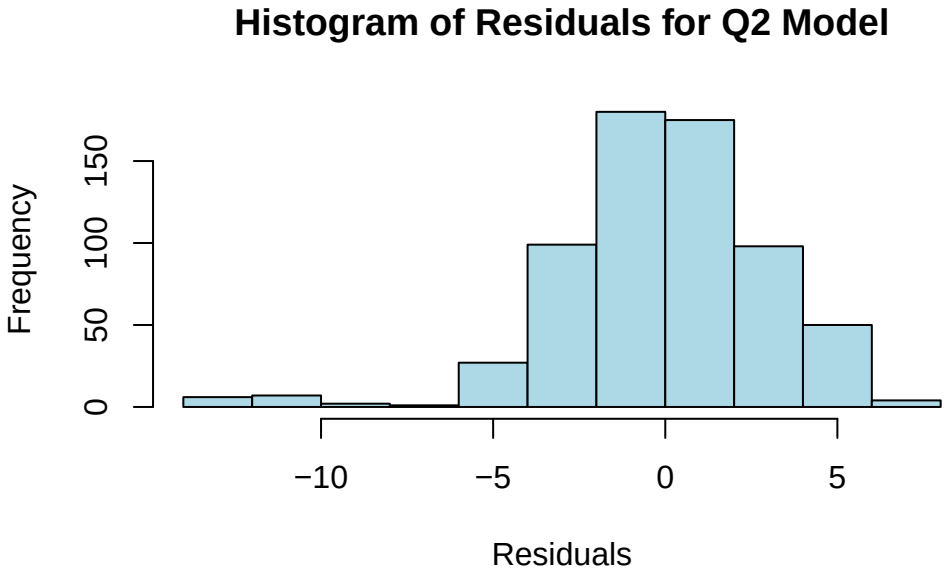
To better understand the relationship between alcohol consumption and academic performance, an interaction plot is generated. This plot helps to visualize the main effects of `Dalc` and `Walc` and the lack of a significant interaction between them:

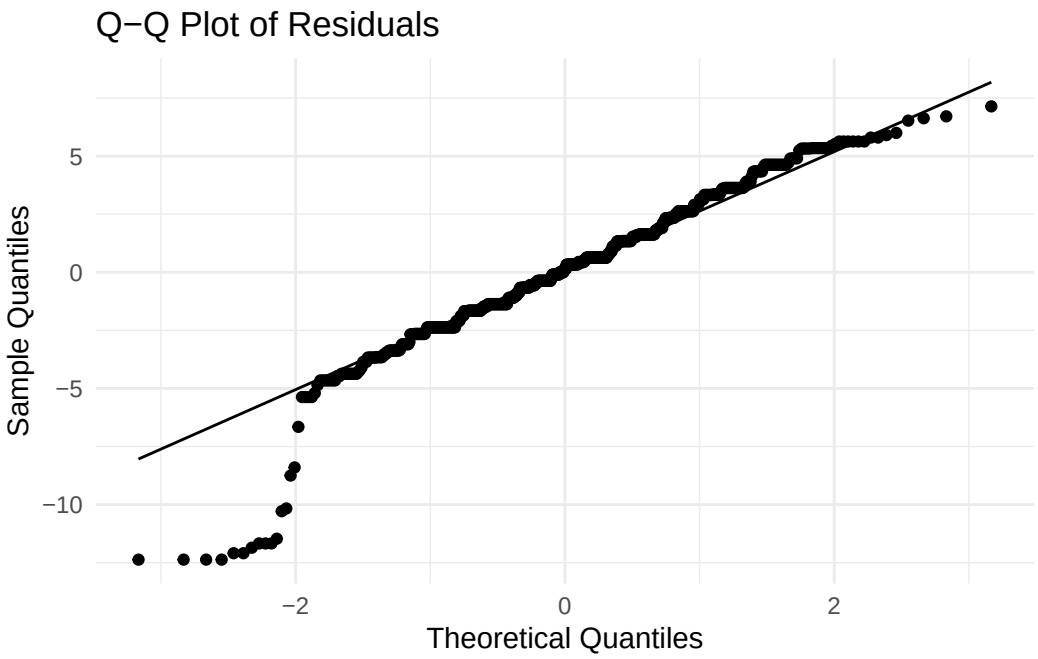


The interaction plot displays mean final grades by levels of weekday and weekend alcohol consumption. For all levels of weekend alcohol use, there is a general downward trend in academic performance as weekday alcohol use increases. The lines representing different levels of weekend alcohol use are mostly parallel, suggesting no strong interaction between weekday and weekend drinking. One exception is the sharp increase for `Walc = 1` at `Dalc = 3`, which may indicate variability or a small sample at that point. Overall, the plot suggests that weekday drinking is more consistently associated with lower grades than weekend drinking, and that the two do not strongly interact.

To check the assumptions of the ANOVA model, we analyze the residuals:

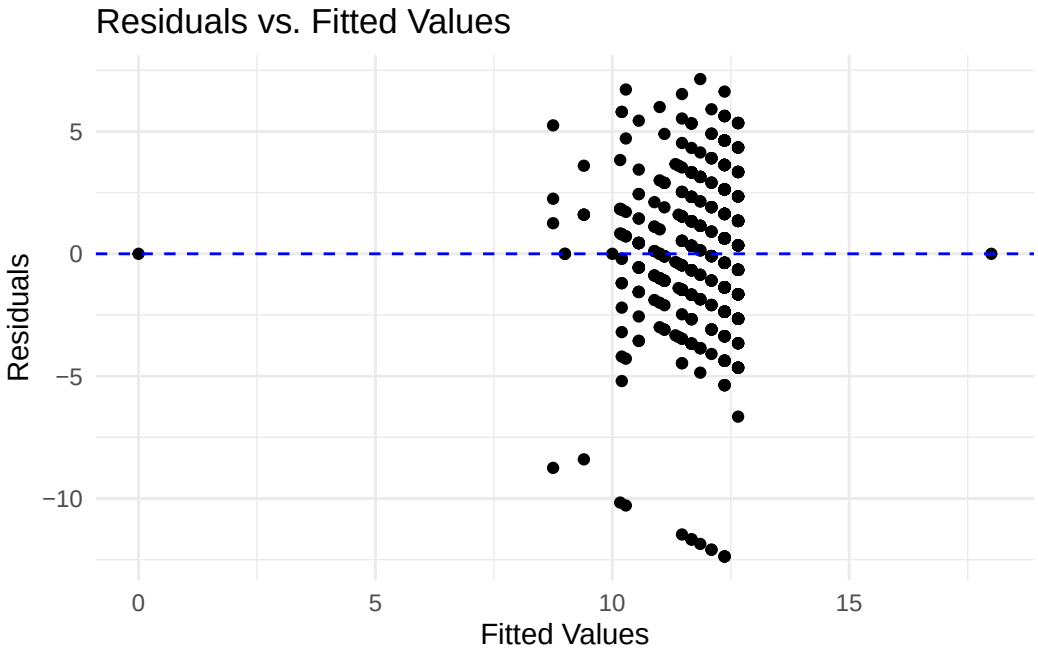
5.4.1 Normality of Residuals





The histogram of the residuals shows a distribution that is approximately normal as it is roughly symmetric and bell-shaped, centered around at 0. The Q–Q plot supports this conclusion for the most part, as the residuals closely follow the theoretical quantile line in the center of the distribution. However, there is some noticeable deviation in the lower tail, suggesting potential skewness or the presence of outliers. Despite these tail deviations, the residuals largely conform to the normality assumption required for ANOVA. Overall, the assumption of normality appears reasonably satisfied.

5.4.2 Equal Variances Across Fitted Values



While the Shapiro-Wilk test returned a highly significant p-value ($p \approx 4.94 \times 10^{-15}$), indicating a deviation from normality in the residuals, visual assessments such as the histogram and Q–Q plot suggest that the deviation is relatively minor and occurs primarily in the tails.

Additionally, the residuals appear to have approximately constant variance across the range of fitted values, with no clear funneling or systematic patterns—supporting the assumption of homoscedasticity. The presence of vertically aligned points is expected in an ANOVA model due to discrete factor levels, and the vertical spread within those columns is reasonably consistent.

Given the large sample size, ANOVA is generally robust to mild departures from normality. Therefore, despite the significant p-value from the Shapiro-Wilk test, there is no strong evidence of a meaningful violation of model assumptions. The assumptions of normality and homoscedasticity are considered reasonably satisfied, supporting the trustworthiness of the ANOVA results.

5.5 Conclusion

Our two-way ANOVA analysis examined the relationship between Portuguese high school students' alcohol consumption and their final grades in Portuguese class. The results revealed a significant negative association between weekday alcohol consumption and final grades, suggesting that higher levels of drinking during the school week are associated with lower academic performance.

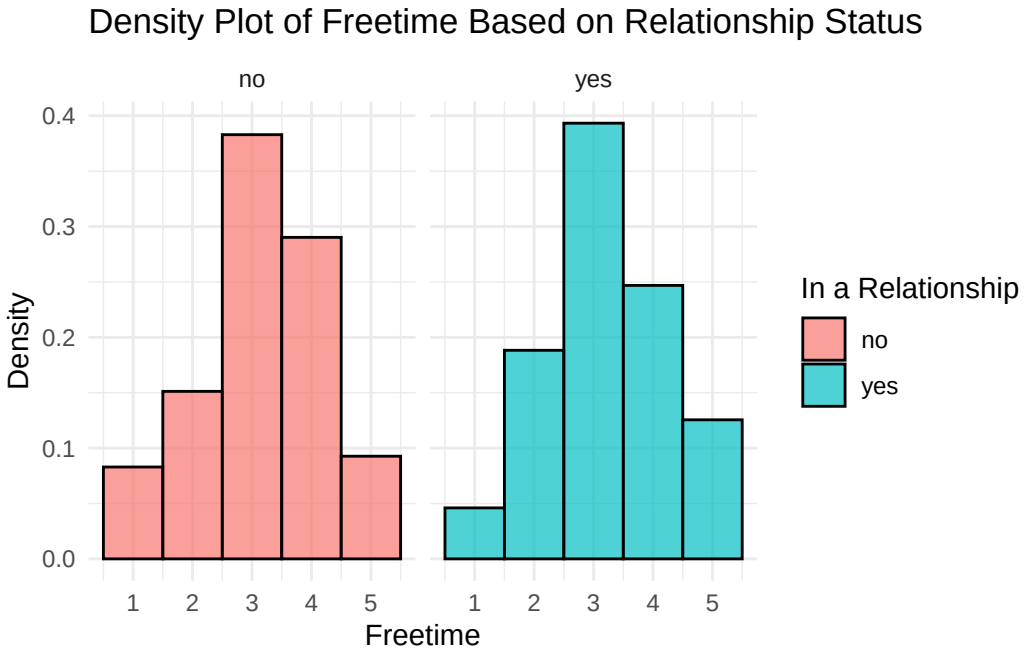
In contrast, weekend alcohol consumption did not show a statistically significant effect on grades. The interaction between weekday and weekend alcohol use was not significant, indicating that the effect of weekday drinking on academic outcomes does not vary meaningfully across levels of weekend drinking.

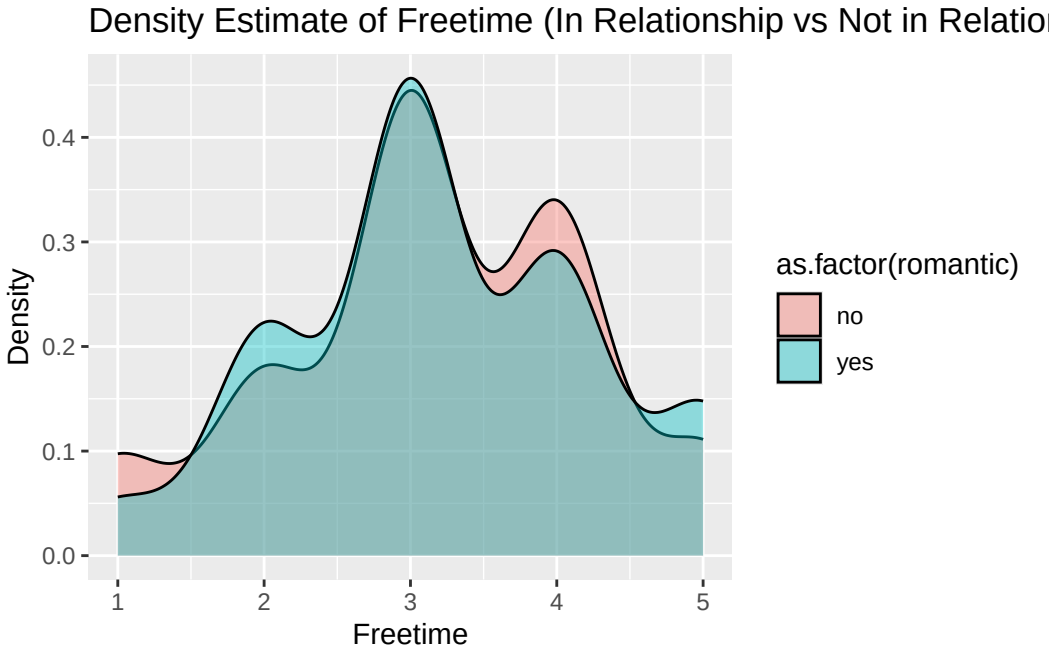
These findings suggest that weekday drinking may be disruptive to students' academic success, possibly due to its interference with school responsibilities, sleep quality, and cognitive performance. While the study does not establish causality, the strong association points to a behavioral pattern that may warrant intervention. Future research or policy efforts may consider addressing weekday alcohol consumption to support better academic outcomes.

6 Research Question 3

6.1 Modeling Process

When looking at Research Question 3, it was evident that I would be comparing the average amount of free time between two groups. Those groups being students in relationships versus students who are not in relationships. The first thing I did was create side by side density histograms showing the distributions of average free time between the two groups. I made sure it was a density histogram because there were more students not in relationships (412) than those in relationships (239). This ensured that the distributions could be compared 1 to 1 without being affected by unequal weighting. The next plot I created was an overlayed density plot of both distributions which allows the viewer to see the similarities and differences in distribution in both plots. I created these plots to get some preliminary insights on how the distributions looked before I used a t-test to see whether I could reject or fail to reject the null hypothesis that there is no significant difference in average free time between students in a romantic relationship and those who are not.





6.2 Results

Table 3: Summary of Welch's t-test for Freetime by Romantic Status

Statistic	Value
Mean (No)	3.159
Mean (Yes)	3.218
Variance (No)	1.122
Variance (Yes)	1.079
t-statistic	-0.693
df	505.830
p-value	0.488
95% CI Lower	-0.226
95% CI Upper	0.108

We use a Welch's two-sided t-test because the inputted data is independent, the sample sizes between the groups differ, and the variances are unequal. Based on the t-test, we get a t statistic of -0.67501, and a p value of 0.5. Because this is a two tailed test with a = 0.05, our alpha endpoints are approximately -1.96 and 1.96. The t statistic of -0.67501 falls between this interval, so we fail to reject the null hypothesis.

6.3 Interpretation

Based on the two-sided Welch's t-test, we fail to reject the null hypothesis. This indicates that there is no significant difference in average free time between students in relationships and students not in relationships. This finding agrees with the density histograms and density plot because when looking at the models, the two groups appear to follow similar distributions. Running the t-test confirms the observation that there is no statistically significant difference in average free time.

7 Conclusion