STATS 15 Final Project - Portuguese Secondary School Academic Performance Report

Cindy Li, Eleanore Zhu, Yinbo Zhang, Joseph Liu, Haoxiang Gao

2025-05-31

Executive Summary

This project investigates how demographic, social, and behavioral factors correlate to student academic performance using data from two public secondary schools in Portugal. Our analysis reveals that academic outcomes are shaped by multiple variables, sometimes even in unexpected ways.

Demographic factors show that older students and those from rural areas tend to underperform in Portuguese, while male students perform better in Math and female students in Portuguese. Behavioral factors like study time, travel time, free time, and alcohol consumption have a stronger influence on Portuguese performance. Confirming normal intuition, students in better health condition and with fewer past failures in class tend to perform better in both subjects. Social support also plays a key role: students with good family relationships and desire for higher education are more likely to achieve top grades (A or higher).

Further analysis investigated some cross section variable interactions. While romantic relationships were initially hypothesized to negatively affect study time and thus grades, the data did not support this. In contrast, parental education—especially maternal education—shows a very strong positive correlation with academic performance, especially when the mother is the student's guardian. Finally, we discovered that Math and Portuguese respond to different types of influences. Overall, students show better performance as well as greater over-time grade improvement in Portuguese than in Math. There are some differences in the strength of correlation between some variables and grades in each subject.

Section 1 - Introduction

1.1 - Motivating Questions and Scope of Analysis

All our group members grew up in education-emphasizing Asian families. Though we completed high school in different countries, we share the experience of working hard towards higher education. So, we are drawn to the topic of academic performance and the underlying determinants. In particular, we would like to investigate how factors other than intelligence can impact educational attainment. Rather than focusing our exploration on China or the U.S., whose education systems we are already familiar with, we decided to challenge ourselves with Portugal, which has a different educational structure and less societal emphasis on education. We are especially curious about what shapes student performance in a context that differs vastly from our own.

In this project, we focus on two public schools in Portugal and use exploratory data analysis (EDA) to answer the overarching research question: **How do demographic, social support, and behavioral factors relate to academic performance – measured by test scores over time (G1, G2, G3) – and are there subject-based differences in these relationships?**

1.2 - Portugal's Educational Landscape: Historical and Social Context

Education in Portugal was established by the Constitution of the Republic in 1976 according to the democratic principles of the freedom to teach and learn("Overview"). Ten years later, the Education Act that defines educational objectives, structures, and modes of organization was also introduced, aiming to expand access to compulsory education. However, the underinvestment, regional disparities, and value differences still impact student educational outcomes today.

Despite Portugal's substantial efforts to improve its educational system, in 2006, the country had an early school dropout rate of 40% for 18 to 24-year-olds, lagging behind the 15% average for various other European Union nations (Cortez & Silva, 2008). Among the challenges that Portugal faces are teacher shortages, demographic shifts, and unequal access to quality instruction. In particular, many students have reported studying without teachers in at least one subject for extended periods of time ("Overview"). The systemic difficulties of lacking educational resources not only affect immediate academic performance, but also limit long-term opportunities for higher education and career advancement. This is especially true for students from disadvantaged backgrounds or remote areas – individuals who lack opportunities to access quality education and succeed in academic pursuits.

Therefore, it is critical to understand what factors influence student achievement in this context, and by analyzing how behavioral, social, and demographic variables interact with academic outcomes in Portugal, we hope to gain insights into an education system that faces significant historical and modern challenges.

1.3 - Structure of the Portuguese Education System

According to the Education Act of 1986, Portugal's education system is divided into three levels. The system begins with pre-school education, an optional education opportunity for children between the ages of three and six. Compulsory schooling spans twelve years (ages 6 - 18) and is divided into basic education and upper secondary education.

Basic education lasts nine years and is subdivided into three cycles. The 1st cycle is from Grades 1-4 (ages 6-10). The 2nd cycle: Grades 5 - 6 (ages 10-12). The 3rd cycle: Grades 7-9 (ages 12-15). This is followed by upper secondary education from grades 10-12 (aged 15-18), during which students may choose among general academic tracks (such as science and humanities), vocational training, or specialized artistic programs ("Overview"). The different pathways give students opportunities to personalize future plans based on their own circumstances; whether it be continuing towards higher education and university, polytechnics that emphasize technical training, or directly entering the labor market upon graduation. However, despite the structural flexibility, all upper secondary students are evaluated under a standard 0-20 grading scale, 20 being the highest possible score ("Overview"). This scale is used across general and vocational tracks, and is also applied in the dataset we have chosen. For better understanding, Portugal's 20-point grading scale can be mapped onto a conventional US letter grading scale as follows: A (16 - 20), B (14-15), C (12-13), D (10-1), and F(0-9); the scores represent the following achievements: A is excellent/very good, B is goo,. C is satisfactory, D is sufficient, F is failed (Cortez & Silva (2008)).

1.4 - Data Set Overview

The dataset we are analyzing was obtained from the UCI Machine Learning Repository, contributed by Paulo Cortez, a professor at the Department of Information Systems, University of Minho, Portugal (Cortez). The data contains the academic outcomes of students from two public secondary schools in Portugal. It includes the academic score of two courses, Math and Portuguese, over a period of time with more detailed attributes of each of the students. The students' scores were collected from school reports, on paper sheets, and including a few attributes (including the three-period grades (beginning, middle, and final course grades) and number of school absences) during the 2005-2006 school year. A questionnaire is also given to the students to complement other information. The questionnaires were designed with closed questions related to demographic, social/emotional, and behavioral variables. The questionnaires were first reviewed by professionals and pre-tested by a small group of students. The final version, with 37 questions, was answered by 788 students, among whom 111 were discarded due to lack of identification (Cortez & Silva, 2008). The resulting data was sorted into two datasets, one related to math scores and one to Portuguese language scores.

Additionally, the data was sampled from two public secondary schools – Gabriel Pereira (GP) and Mousinho da Silveira (MS) – located in the Alentejo region of southern Portugal. Alentejo is known for its agricultural economy, with extensive regions that are essentially rural and sparsely populated(Alentejo, Turismo do.). Its lagging economic development compared to urban centers like Lisbon would very likely influence the educational experiences and outcomes of students in this dataset. For instance, how students are living in rural or urban areas, the travel times to school, and limited access to education support may contribute to the differences we observe in student performance.

As the region of Alentejo makes up about one-third of Portugal(Alentejo, Turismo do.), we hope to gain a more nuanced understanding of how local Portuguese students' academic achievement is affected by various factors by situating our analysis within this specific region. This adds depth to our research question and allows for a more grounded interpretation of the dataset's patterns.

1.5 - Observational Units and Variable Structure

Each row in the dataset represents one student enrolled in a single subject course, either Mathematics or Portuguese. Since the data was split into two subject-specific files without identifying factors for us to match the student, the same student may appear in both datasets but as distinct entries. For the purposes of our project, we analyze each subject separately, treating each record as an independent observational unit. In the end, we will try to compare and analyze how the trends observed in the two courses differ.

During our initial data exploration, we did not find any missing values. This observation is supported by the original paper published that analyzes this dataset, which notes that entries lacking identification details were filtered out before publication (Cortez & Silva, 2008). We also checked for potential duplicate entries. However, since identifying information has been removed, it is not possible to confirm the presence of duplicates, although the structure and origin of the data suggest that such instances are unlikely.

1.6 - Variable Explanation

#	Variable	Type	Description
1	G1	Numeric, Discrete	First-period grade (0–20)
2	G2	Numeric, Discrete	Second-period grade (0–20)
3	G3	Numeric, Discrete	Final course grade (0–20)
4	Age	Numeric, Discrete	Student's age (15–22)
5	Sex	Character, Binary	Student's sex ('F' = female, 'M' = male)
6	School	Character, Binary	Student's school ('GP' = Gabriel Pereira, 'MS' = Mousinho da Silveira)
7	Address	Character, Binary	Home address type ('U' = urban, 'R' = rural)
8	Guardian	Character, Nominal	Student's guardian ('mother', 'father', 'other')
9	Medu	Numeric, Ordinal	Mother's education (0 = none to 4 = higher ed)
10	Fedu	Numeric, Ordinal	Father's education (0 = none to 4 = higher ed)
11	Mjob	Character, Nominal	Mother's job ('teacher', 'health', 'services', 'at_home', 'other')
12	Fjob	Character, Nominal	Father's job ('teacher', 'health', 'services', 'at_home', 'other')

#	Variable	Type	Description
13	Studytime	Numeric, Ordinal	Weekly study time (1 = <2h to 4 = >10h)
14	Traveltime	Numeric, Ordinal	Travel time to school (1 = <15min to 4 = >1h)
15	Failures	Numeric, Discrete	Past class failures (0 to 4+)
16	Paid	Character, Binary	Took extra paid classes ('yes', 'no')
17	Absences	Numeric, Discrete	Number of school absences (0–93)
18	Health	Numeric, Ordinal	Current health status (1 = very bad to 5 = very good)
19	Pstatus	Character, Binary	Parent cohabitation ('T' = together, 'A' = apart)
20	Schoolsup	Character, Binary	School educational support ('yes', 'no')
21	Famsup	Character, Binary	Family educational support ('yes', 'no')
22	Activities	Character, Binary	Participates in extracurriculars ('yes', 'no')
23	Higher	Character, Binary	Plans for higher education ('yes', 'no')
24	Nursery	Character, Binary	Attended nursery school ('yes', 'no')
25	Internet	Character, Binary	Internet access at home ('yes', 'no')
26	Romantic	Character, Binary	In a romantic relationship ('yes', 'no')
27	Famrel	Numeric, Ordinal	Family relationship quality (1 = very bad to 5 = very good)
28	Goout	Numeric, Ordinal	Going out with friends (1 = very low to 5 = very high)
29	Dalc	Numeric, Ordinal	Weekday alcohol consumption (1 = very low to 5 = very high)
30	Walc	Numeric, Ordinal	Weekend alcohol consumption (1 = very low to 5 = very high)
31	Freetime	Numeric, Ordinal	Free time after school (1 = very low to 5 = very high)
32	Famsize	Character, Binary	Family size ('LE3' <= 3, 'GT3' > 3 members)
33	Reason	Character, Nominal	Reason for choosing school ('home', 'reputation', 'course', 'other')

1.7 - Related Research

The original collector of this dataset explored the data in-depth and intended to predict student achievement in secondary education using Business Intelligence/Data Mining techniques. This paper provides us with some insights into possible relationships we could further investigate.

The following is the method used by the paper. It focuses on binary classification and five-level classification to establish data and it uses Naive Predictor, Decision Tree, Random Forest, and Neural Network to analyze the data (Cortez & Silva, 2008). These methods can identify the key elements of achieving success and process a large amount of data. Its primary purpose is to help predict students' final grades (G3) and to analyze how various factors contribute to academic success or failure. By using supervised machine learning models like Decision Trees, Random Forests, Support Vector Machines, and Neural Networks, we aim to understand these relationships better and improve the accuracy of predicting student outcomes (Cortez & Silva, 2008). Like Failure and G2, G1 is more influential than absence and school support. Mothers' educational experiences influence the grades of students, but not such influentially. On top of that, students' drinking behavior is negatively correlated with grades (Cortez & Silva, 2008). Overall, this analysis offers valuable insights into the complex factors influencing student performance in Portugal. Through examining the dataset and model results, it becomes clear that while academic history plays the most significant role, social behaviors, and family background also contribute in meaningful ways.

If the original research paper from Cortez & Silva (2008) focused on the evaluation of predictive accuracy across three tasks: binary classification (pass/fail), five-level classification (grade bands), and regression (numeric final grade G3) using the data of Portuguese and Mathematics separately, the other paper from Ali Khan (2020) took a different approach. He uses a two-variable decision tree to analyze student performance factors and rank variable importance using MIC (Maximal Information Coefficient). The focus of the paper is exploratory variable ranking and classification through a customized tree structure, prioritizing interpretability and policy implications with the use of both Portuguese and Math grades together(Ali Khan). Both papers provided a meaningful exploration and insight into the dataset. We aim to conduct further research based on this paper's dataset to investigate new and insightful relationships between variables.

References

Alentejo, Turismo do. "Visitalentejo." Turismo Do Alentejo, www.visitalentejo.pt/en/. Accessed 21 May 2025.

Ali Khan, Yousaf. "Factors Influencing Secondary School Student's Performance through Variable Decision Tree Data Mining Technique." International Journal of Data Science and Analysis, vol. 6, no. 5, 2020, p. 120, https://doi.org/10.11648/j.ijdsa.20200605.11.

Cortez, P. and A. M. Gonçalves Silva. "Using data mining to predict secondary school student performance." (2008).

"Overview." Europa.eu, 2024, eurydice.eacea.ec.europa.eu/eurypedia/portugal/overview. Accessed 18 May 2025.

Section 2 - Data Cleaning

```
library(knitr)
library(tidyverse)
library(ggplot2)
library(stringr)
library(broom)
library(tidyr)
library(dplyr)
library(paletteer)
library(patchwork)
library(lmtest)
library(readr)
```

2.1 - Loading Data

First, we load in our data set. The original data set contains two .csv files: one of Math grade and one of Portuguese grade. Each with the same 33 varibles as introduced above in the introduction section.

```
math_data <- read.csv2("student-mat.csv")
port_data <- read.csv2("student-por.csv")
```

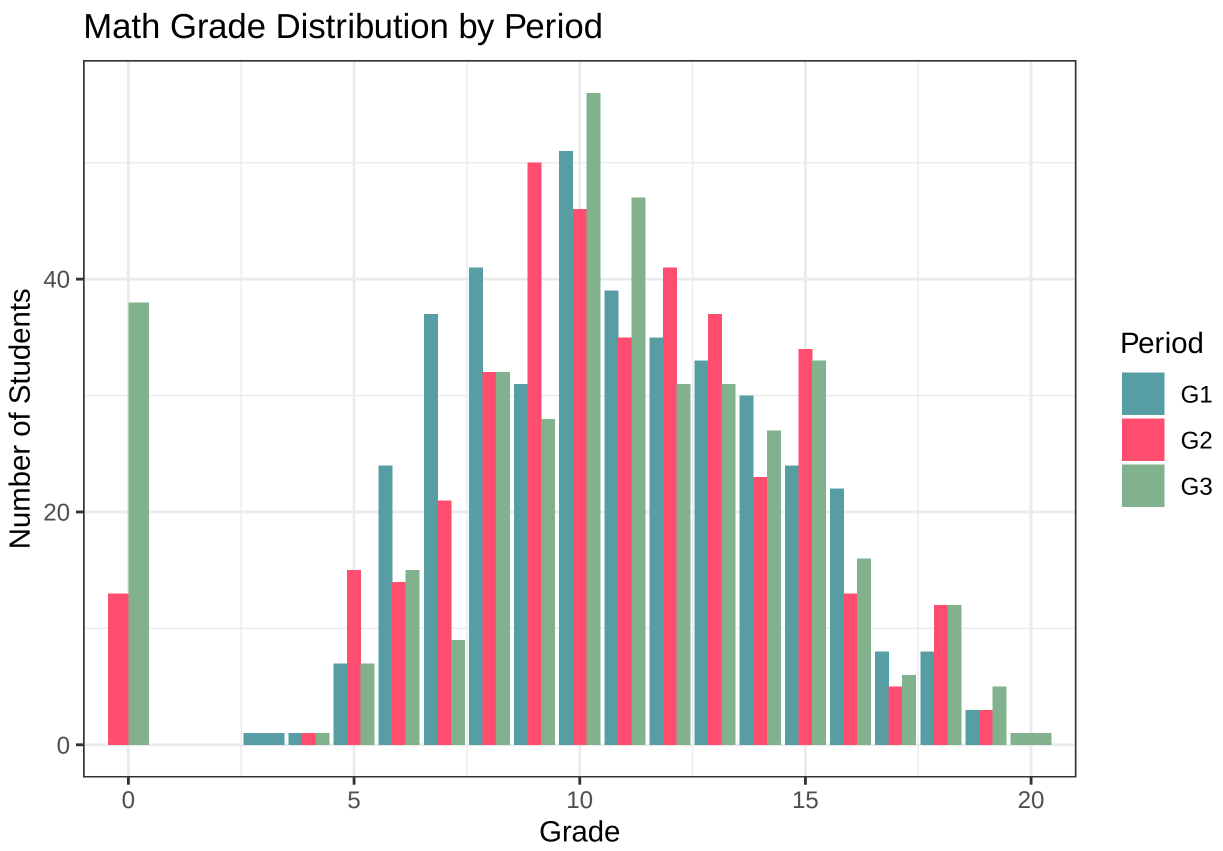
2.2 - Exploring the Data Set

While the data set was already cleaned by its provider, excluding missing values and data lacking identifications, We will still explore it first to see if there is any other errors or outliers.

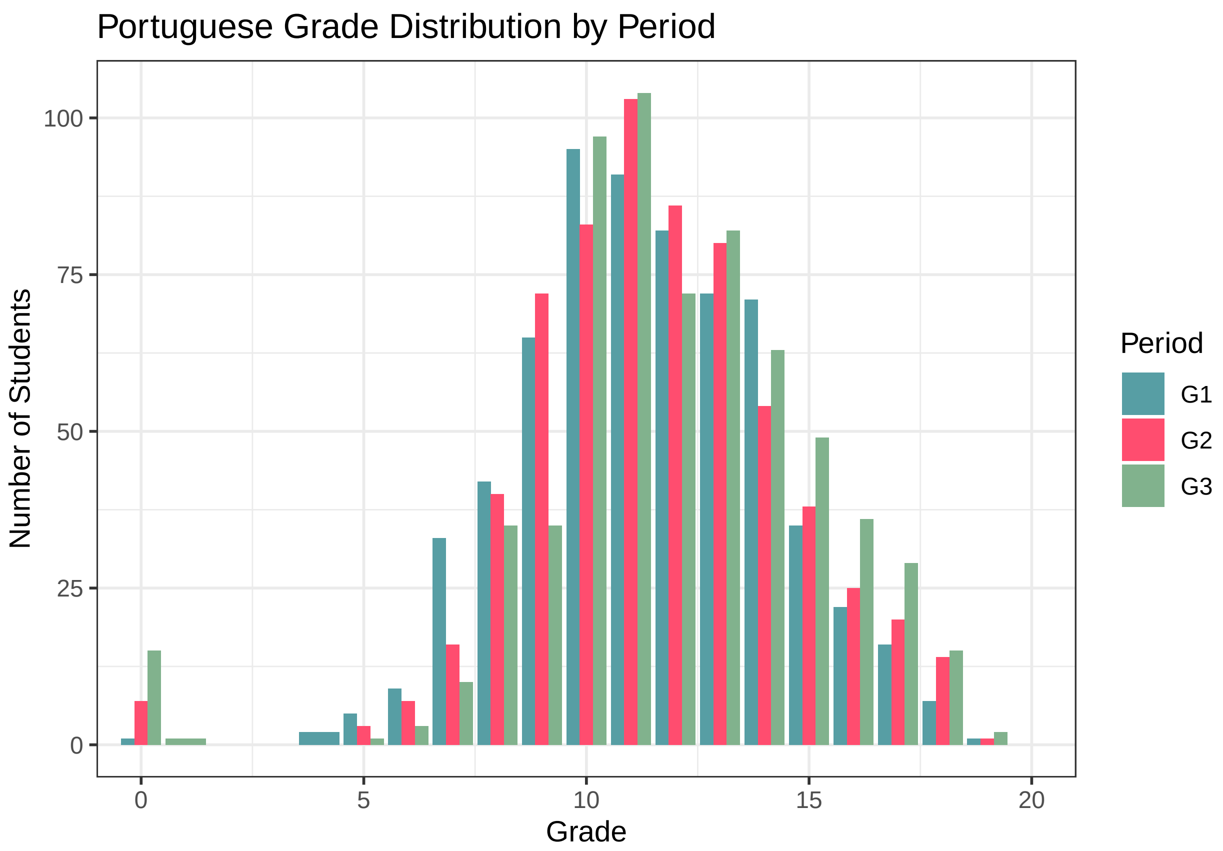
We plot the grade distributions of each subject and period first.

```
math_data_longer <- math_data %>%
  pivot_longer(cols = G1:G3, names_to = "Period", values_to = "Grade")
port_data_longer <- port_data %>%
  pivot_longer(cols = G1:G3, names_to = "Period", values_to = "Grade")
```

```
math_data_longer %>%
  ggplot(aes(x=Grade, fill=Period))+
  geom_bar(position="dodge")+
  coord_cartesian(xlim = c(0, 20)) +
  labs(title="Math Grade Distribution by Period", x="Grade", y="Number of Students")+
  scale_fill_paletteer_d("ggthemes::Classic_Blue_Red_12") +
  scale_fill_manual(values = as.character(paletteer::paletteer_d("MoMAColors::Klein")[c(2,1,7)]))+
  theme_bw()
```



```
port_data_longer %>%
  ggplot(aes(x=Grade, fill=Period))+
  geom_bar(position="dodge")+
  coord_cartesian(xlim = c(0, 20)) +
  labs(title="Portuguese Grade Distribution by Period",x="Grade", y="Number of Students")+
  scale_fill_manual(values = as.character(paletteer::paletteer_d("MoMAColors::Klein")[c(2,1,7)]))+
  theme_bw()
```



We notice that an increasing number of students got a 0 in G2 and G3 in both math and Portuguese class, which is not plausible if that was their actual grade. We will further analyze why these data appears.

We first filter out the students who score 0 in either G1, G2, or G3.

```
math_data %>%
  filter(G1 == 0 | G2==0 | G3 == 0) %>%
  select(G1,G2,G3) %>%
  arrange(G3) %>%
  arrange(G2) %>%
  arrange(G1) %>%
  head(10)
```

```
##    G1 G2 G3
## 1   4  0  0
## 2   5  0  0
## 3   5  0  0
## 4   6  0  0
## 5   6  0  0
## 6   6  5  0
## 7   6  5  0
## 8   6  5  0
## 9   6  5  0
## 10  6  5  0
```

```
port_data %>%
  filter(G1 == 0 | G2==0 | G3 == 0) %>%
  select(G1,G2,G3)%>%
  arrange(G3) %>%
  arrange(G2) %>%
  arrange(G1) %>%
  head(10)
```

```
##    G1 G2 G3
## 1   0 11 11
## 2   4  0  0
## 3   5  0  0
## 4   5  0  0
## 5   5  8  0
## 6   7  0  0
## 7   7  0  0
## 8   7  5  0
## 9   7  7  0
## 10  7  7  0
```

We notice that every student who got a 0 in G2 also got a 0 in G3, which suggests that 0 likely represents the student dropping the class rather than their actual score. Therefore, in our further analysis, we should distinguish these students with those who took the actual test.

Specifically, there is one student in Portuguese class who receive a 0 in G1 but 11 in G2 and G3. It is possible that this student scored 0 in G1 but improved their grade in G2 and G3. It is also possible that this student did not take the test in G1 but attended G2 and G3 exams. For this reason, we will not exclude this data from our future analysis.

2.3 - Relationship between Dropping Class and Other Variables

We would like to see if there is a relationship between other variables and whether the student would drop the class. Some factors are related to whether students are more likely to dropout.

For the purpose of easier analysis, we will create a combined table of Math and Portuguese grades, adding three new variables: bool variable G2_zero and G3_zero indicating whether the student drop the class in G2 and G3, and subject indicating which class (Math/Portuguese) the grade is.

```
library(tidyverse)
math_data <- math_data %>% mutate(G2_zero = G2 == 0)
port_data <- port_data %>% mutate(G2_zero = G2 == 0)
math_data <- math_data %>% mutate(G3_zero = G3 == 0)
port_data <- port_data %>% mutate(G3_zero = G3 == 0)
math_data$subject <- "Math"
port_data$subject <- "Portuguese"
combined_uncleaned <- bind_rows(math_data, port_data)
combined_uncleaned %>%
  select(G1, G2, G3, G2_zero, G3_zero) %>%
  head()
```

```
##   G1 G2 G3 G2_zero G3_zero
## 1  5  6  6   FALSE   FALSE
## 2  5  5  6   FALSE   FALSE
## 3  7  8 10   FALSE   FALSE
## 4 15 14 15   FALSE   FALSE
## 5  6 10 10   FALSE   FALSE
## 6 15 15 15   FALSE   FALSE
```

After some preliminary trials, we found that there is a relationship between the following two variables and whether the students are more likely to drop out or not.

- School Support
- Higher Education

School Support

- Null hypothesis: Students with or without school support score the same.

```
library(tidyverse)
table_drop_schoolsup <- table(combined_uncleaned$G3_zero, combined_uncleaned$schoolsup)
print(table_drop_schoolsup)
```

```
##
##          no yes
##   FALSE 874 117
##   TRUE   51   2
```

```
fisher.test(table_drop_schoolsup)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table_drop_schoolsup
## p-value = 0.07692
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.03411772 1.14096165
## sample estimates:
## odds ratio
##  0.2931678
```

- p-value - 0.0768. We fail to reject the null hypothesis at alpha = 0.05, but at a more lenient alpha = 0.10, we have a marginally significant relationship. The odds ratio = 0.293 suggests that school support may reduce dropout odds.

Higher Education

```
library(tidyverse)
table_drop_higher <- table(combined_uncleaned$G3_zero, combined_uncleaned$higher)
print(table_drop_higher)
```

```
##
##          no yes
##   FALSE  78 913
##   TRUE   11  42
```

```
fisher.test(table_drop_higher)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table_drop_higher
## p-value = 0.003525
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1573751 0.7324109
## sample estimates:
## odds ratio
##  0.3267197
```

- p-value < 0.05. We reject the null hypothesis.
- There is a strong relationship between whether the student aim for higher education and whether they would drop classes, with an odds ratio of 0.32.

2.4 Data Clearning Conclusion

We find some students with 0 in G2 and G3 grades. We deduce that these are due to the students dropping the class. Therefore we will exclude these data from our analysis of variables and grades in future sections.

We also create four cleaned data sets excluding students who drop out: two cleaned data sets of math grades and Portuguese grades respectively,

```
math_data_cleaned <- math_data %>%
  filter(G2 != 0 & G3 != 0)
port_data_cleaned <- port_data %>%
  filter(G2 != 0 & G3 != 0)
```

We create another data frame that combines these two data set, with a new column named "subject".

```
math_data_cleaned$subject <- "Math"
port_data_cleaned$subject <- "Portuguese"
combined_data <- bind_rows(math_data_cleaned, port_data_cleaned)
combined_data %>%
  select(age, sex, subject, G1, G2, G3) %>%
  head()
```

```
##   age sex subject G1 G2 G3
## 1  18   F    Math  5  6  6
## 2  17   F    Math  5  5  6
## 3  15   F    Math  7  8 10
## 4  15   F    Math 15 14 15
## 5  16   F    Math  6 10 10
## 6  16   M    Math 15 15 15
```

Furthermore, for coding convenience, we have another data frame that is pivoted longer, putting G1, G2 and G3 grade into the same column, adding a new "period" column.

```
longer_data <- combined_data %>%
  pivot_longer(cols = G1:G3, names_to = "Period", values_to = "Grade") %>%
  mutate(GradeGroup = case_when(
    Grade >= 16 ~ "A",           # Excellent/Very Good
    Grade >= 14 ~ "B",           # Good
    Grade >= 12 ~ "C",           # Satisfactory
    Grade >= 10 ~ "D",           # Sufficient
    TRUE        ~ "F"            # Fail
    ))
longer_data %>%
  select(age, sex, subject, Period, Grade, GradeGroup) %>%
  head()
```

```
## # A tibble: 6 x 6
##     age sex   subject Period Grade GradeGroup
##   <int> <chr> <chr>   <chr>  <int> <chr>
## 1    18 F     Math    G1         5 F
## 2    18 F     Math    G2         6 F
## 3    18 F     Math    G3         6 F
## 4    17 F     Math    G1         5 F
## 5    17 F     Math    G2         5 F
## 6    17 F     Math    G3         6 F
```

In further sections, we will conduct an exploratory data analysis on various variables and academic performances.

Section 3 - Demographic

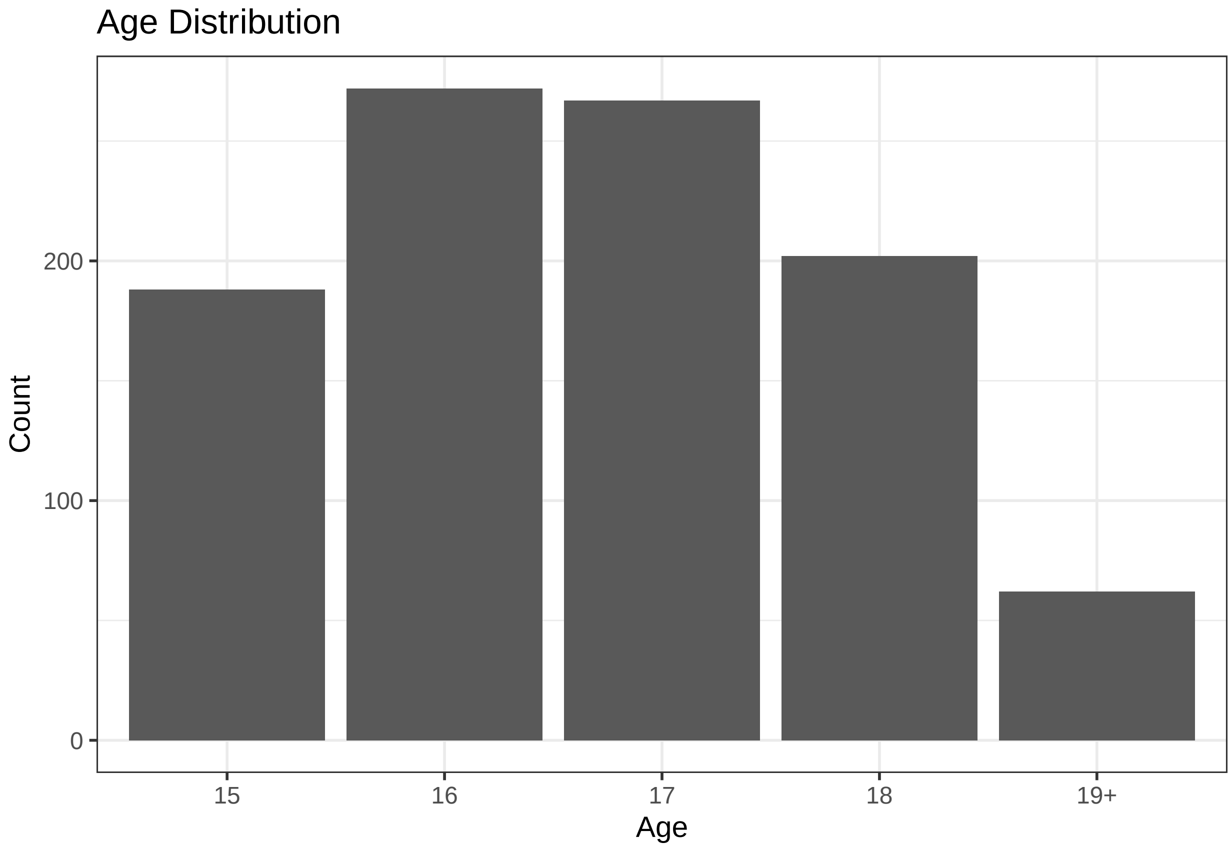
3.1 - Age

Age Distribution

```
data_19plus <- combined_data%>%
  mutate(age_group = ifelse(age >= 19, "19+", as.character(age)))

summary1 <- data_19plus %>%
  group_by(age_group) %>%
  count()

summary1
```

```
## # A tibble: 5 x 2
## # Groups:   age_group [5]
##   age_group     n
##   <chr>     <int>
## 1 15          188
## 2 16          272
## 3 17          267
## 4 18          202
## 5 19+          62
```

```
ggplot(data_19plus, aes(x = age_group)) +
  geom_bar() +
  labs(title = "Age Distribution",
       x = "Age",
       y = "Count")+
  theme_bw()
```

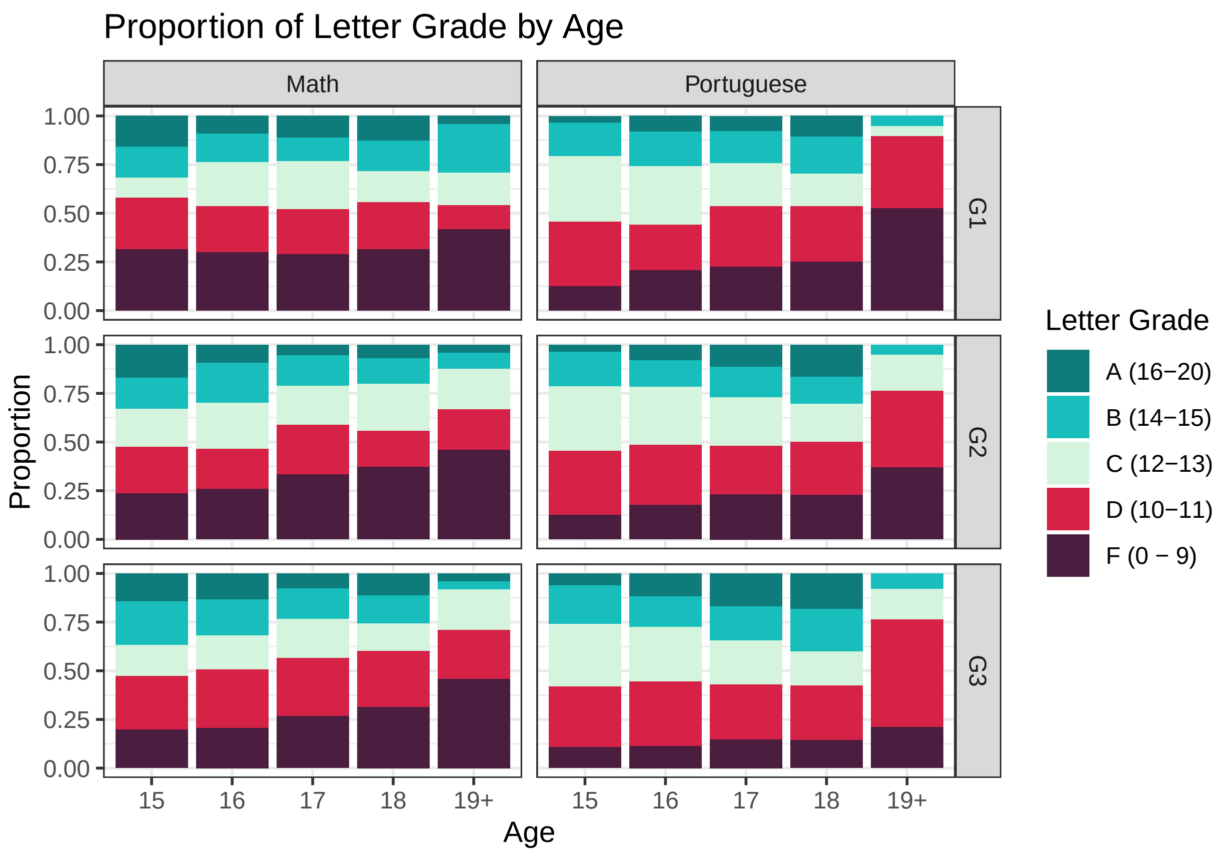


1. Ages 16 and 17 are the most common, each with the highest number of students.
2. Age 15 and 18 have moderate representation.
3. The 19+ group has the fewest students.

Q1: How does the proportion of letter grades vary by age group across periods?

- Null Hypothesis: The proportion of letter grades does not differ by age group across periods.

```
longer_data_19plus <- longer_data %>%
  mutate(age_group = ifelse(age >= 19, "19+", as.character(age)))

ggplot(longer_data_19plus, aes(x = age_group, fill = GradeGroup)) +
  geom_bar(position = "fill") +
  facet_grid(Period ~ subject) +
  labs(title = "Proportion of Letter Grade by Age",
       x = "Age", y = "Proportion",
       fill = "Letter Grade") +
  scale_fill_paletteer_d("PrettyCols::Beach", labels = c(
    "A (16-20)",
    "B (14-15)",
    "C (12-13)",
    "D (10-11)",
    "F (0 - 9)")) +
  theme_bw()
```



```r
longer_data_19plus %>%
  filter(subject == "Math", Period == "G1") %>%
  with(table(GradeGroup, age_group)) %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 13.059, df = 16, p-value = 0.6684
```

```r
longer_data_19plus %>%
  filter(subject == "Portuguese", Period == "G1") %>%
  with(table(GradeGroup, age_group)) %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 51.15, df = 16, p-value = 1.501e-05
```

```r
longer_data_19plus %>%
  filter(subject == "Math", Period == "G2") %>%
  with(table(GradeGroup, age_group)) %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 16.672, df = 16, p-value = 0.4071
```

```r
longer_data_19plus %>%
  filter(subject == "Portuguese", Period == "G2") %>%
  with(table(GradeGroup, age_group)) %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 38.815, df = 16, p-value = 0.001157
```

```r
longer_data_19plus %>%
  filter(subject == "Math", Period == "G3") %>%
  with(table(GradeGroup, age_group)) %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 15.336, df = 16, p-value = 0.5002
```

```r
longer_data_19plus %>%
  filter(subject == "Portuguese", Period == "G3") %>%
  with(table(GradeGroup, age_group)) %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 36.927, df = 16, p-value = 0.002147
```

Visual Observation

As age increases, students in both classes across all periods show an increases in lower letter grade proportion.

Statistical Test (Chi-squared)

- Portuguese p-values (1.501e-05, 0.001157, 0.002147) are all < 0.05, reject H0, proportion of letter grades does differ by age group across periods.

- Math p-values (0.6684, 0.4071, 0.5002) are all > 0.05, fail to reject H0, proportion of letter grades does not differ by age group across periods.

Answer

This indicates that age is associated with letter grade outcomes in Portuguese at each time point — older student groups consistently receive higher proportions of high grades. The only exception is age group 19+ which consistently performs the worst across all periods for both classes. However, for math, students performance isn't impacted by their age.

3.2 - Address

Address Distribution

```r
summary2 <- combined_data%>%
  group_by(address) %>%
  count()

summary2
```

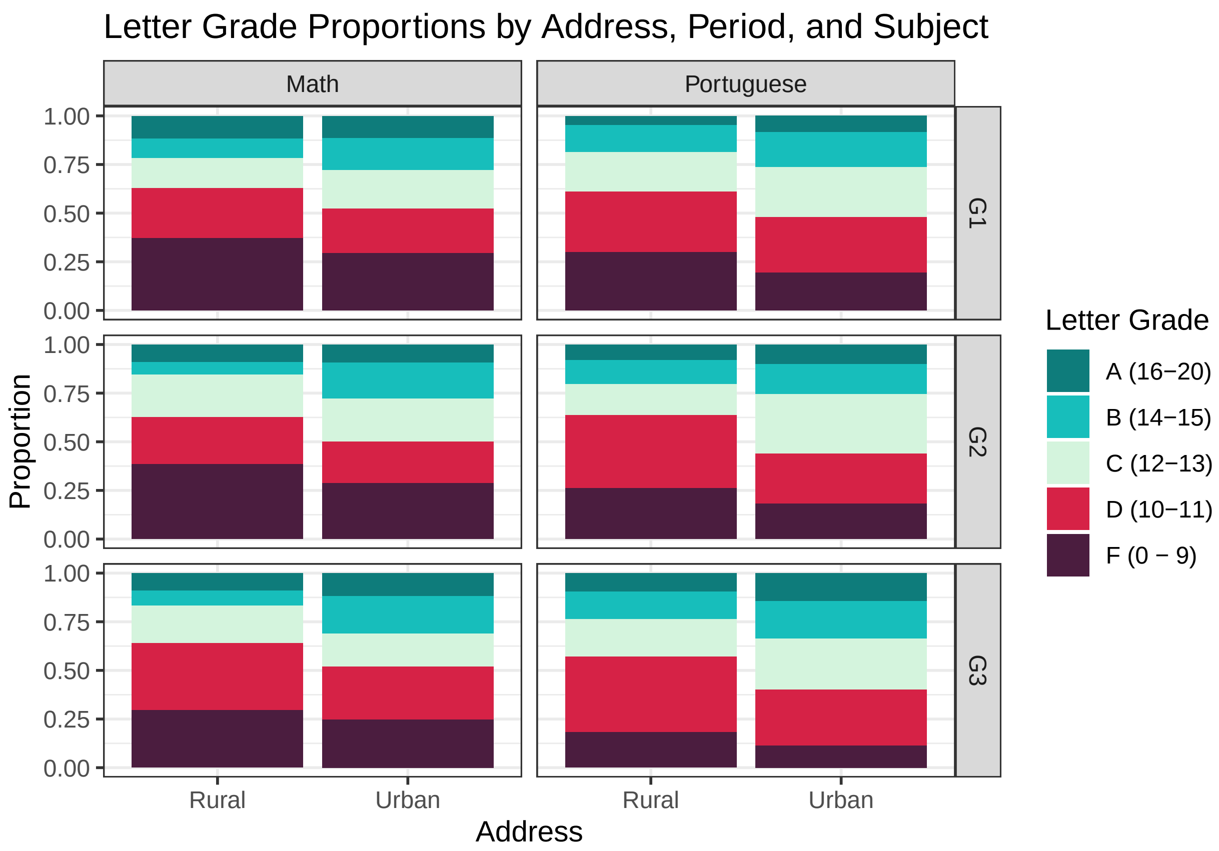
```
## # A tibble: 2 x 2
## # Groups:   address [2]
##   address     n
##   <chr>   <int>
## 1 R         265
## 2 U         726
```

The number of urban students is almost three times the number of rural students.

Q1: How does letter grade distribution vary between urban and rural students across different periods and subjects?

- Null Hypothesis: There is no difference in the distribution of letter grades between urban and rural students, and this distribution is the same across all periods and subjects.

```r
ggplot(longer_data, aes(x = address, fill = GradeGroup)) +
  geom_bar(position = "fill") +
  facet_grid(Period ~ subject) +
  labs(title = "Letter Grade Proportions by Address, Period, and Subject",
       fill = "Letter Grade",
       x = "Address",
       y = "Proportion") +
  scale_fill_paletteer_d("PrettyCols::Beach", labels = c(
    "A (16-20)",
    "B (14-15)",
    "C (12-13)",
    "D (10-11)",
    "F (0 - 9)")) +
  theme_bw()+
  scale_x_discrete(labels = c("R" = "Rural", "U" = "Urban"))
```



```
longer_data %>%
  filter(subject == "Math") %>%
  with(table(GradeGroup, address)) %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 16.612, df = 4, p-value = 0.002299
```

```
longer_data %>%
  filter(subject == "Portuguese") %>%
  with(table(GradeGroup, address)) %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 45.58, df = 4, p-value = 3.012e-09
```

Visual Observation

Overall, rural students seem to obtain lower letter grade than urban students across both classes and all periods.

Statistical Test (Chi-squared)

1. Portuguese: p-value = 3.012e-09 < 0.05, reject H0, address significantly impacts grade in Portuguese class
2. Math: p-value = 0.002299 < 0.05, reject H0, address significantly impacts grade in math class

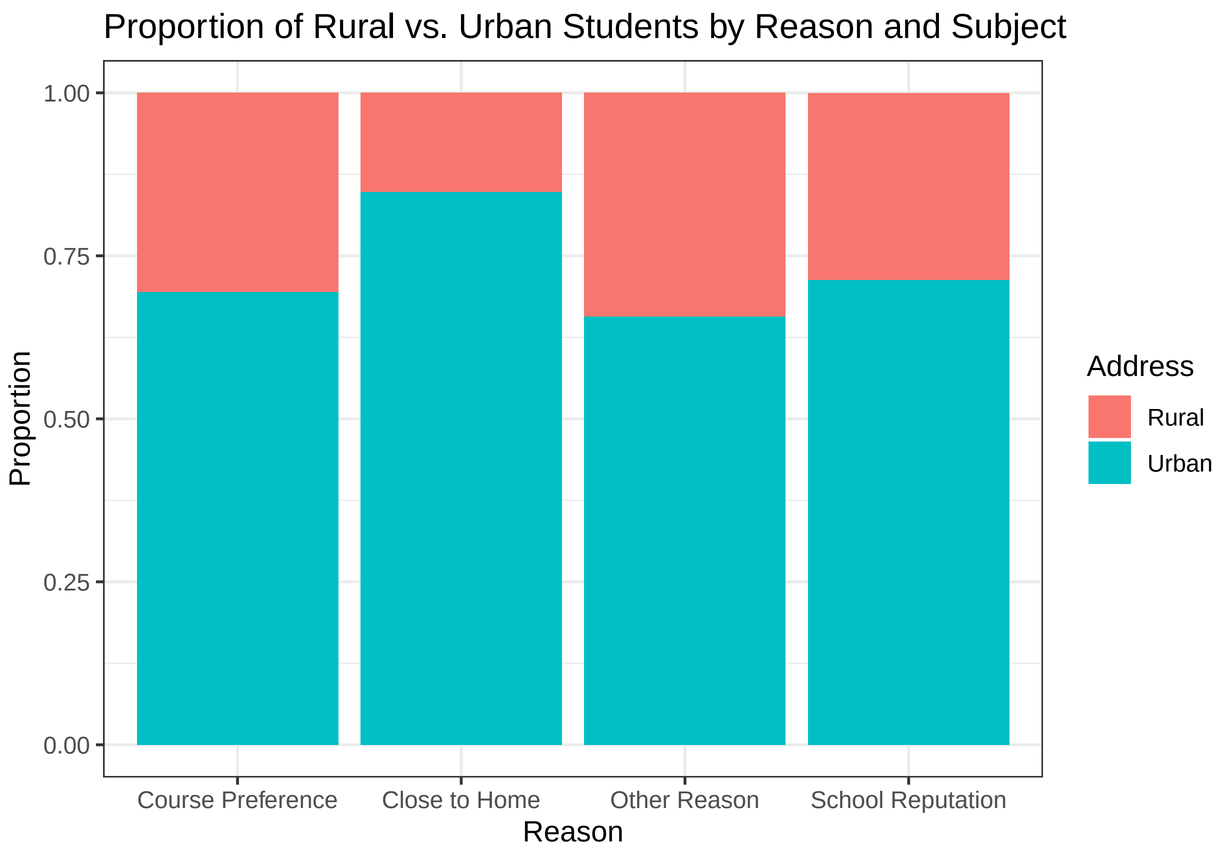
Answer

Address variable impacts letter grade over time for both subject. In both classes, there is a significant difference in letter grade between urban and rural students for every period. This means rural students consistently perform worse than urban students in both subjects throughout the year. This made us curious about if the reason to attend school corresponds to this difference.

Q2: How does the proportion of rural and urban students vary across reason to attend school?

- Null Hypothesis: There is no association between student address (urban vs. rural) and reason for attending school; the proportion of rural and urban students is the same across all reasons.

```r
ggplot(combined_data, aes(x = reason, fill = address)) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of Rural vs. Urban Students by Reason and Subject",
       x = "Reason",
       y = "Proportion",
       fill = "Address") +
  scale_fill_discrete(labels = c(
    "R" = "Rural",
    "U" = "Urban")) +
  scale_x_discrete(labels = c(
    "course" = "Course Preference",
    "home" = "Close to Home",
    "other" = "Other Reason",
    "reputation" = "School Reputation"))+
  theme_bw()
```



```r
table2_data <- table(combined_data$reason, combined_data$address)
chisq.test(table2_data)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table2_data
## X-squared = 22.921, df = 3, p-value = 4.195e-05
```

Visual Observation

1. The reason with largest proportion of urban students is a school that is close to home, second largest being school reputation.
2. The reason with largest proportion of rural students is other, second largest being course preference.

Statistical Test (Chi-squared)

Chi-square p-value of 4.195e-05 $<$ 0.05, reject H0, shows that reason and address variables are dependent

Answer

Address variable impacts grade over time by shaping students' motivation for attending school, which may indirectly influence their academic performance even when attending the same school.

The significant difference in reasons for attending school ($p < 0.001$) between urban and rural students suggests that rural students are more likely to attend school for practical reasons, while urban students more often choose school based on proximity or reputation.

Although they attend the same schools, these differences in motivation may reflect underlying disparities in academic preparation, expectations, or external support, which could help explain why rural students tend to perform worse. For instance, urban students can choose school close to home, indicating they had been closer to better studying resources in the city. This reinforces the idea that address impacts student performance over time — not because of school differences, but due to differences in student context and reason for enrollment.

3.3 - Sex

Sex Distribution

```
summary3 <- combined_data%>%
  group_by(sex) %>%
  count()

summary3
```

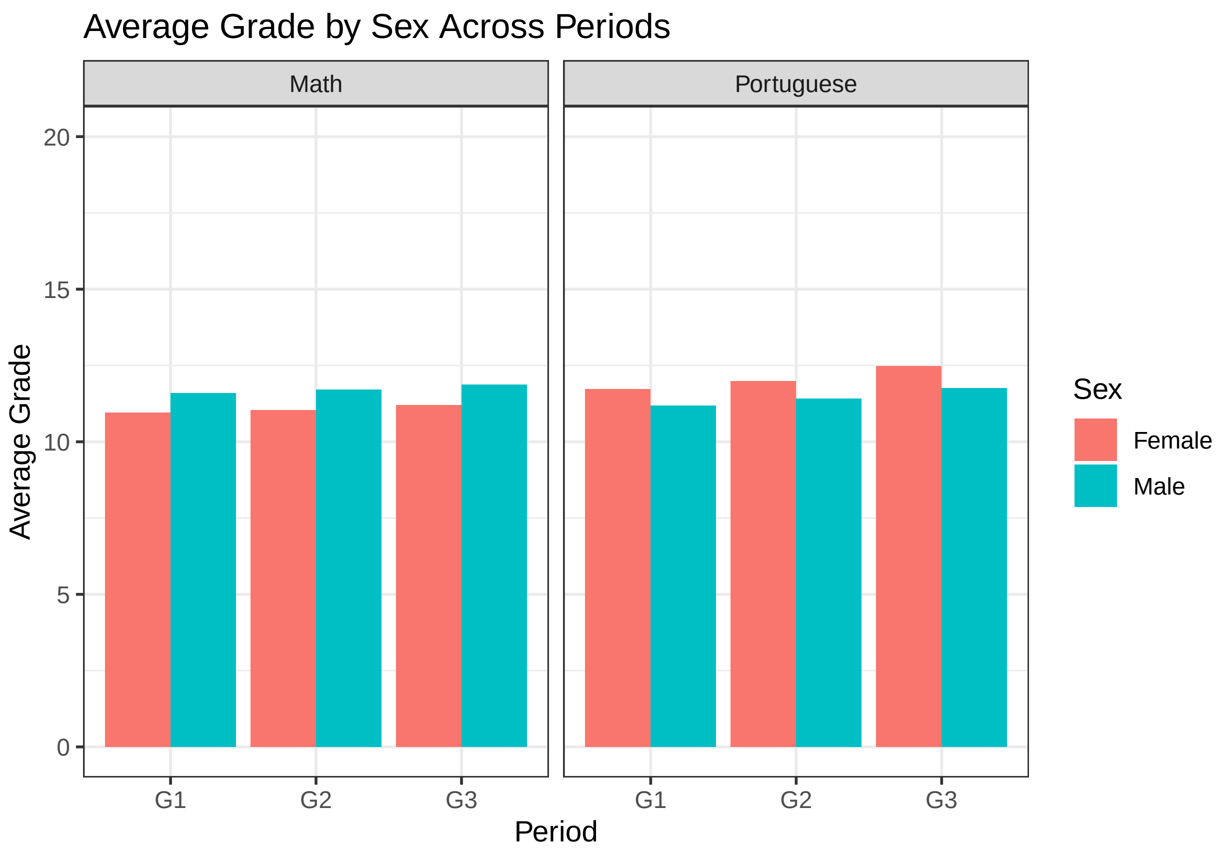
```
## # A tibble: 2 x 2
## # Groups:   sex [2]
##   sex       n
##   <chr> <int>
## 1 F       561
## 2 M       430
```

There are more female students than male students.

Q1: How does student average grade over time differ by sex, and does this pattern vary between Math and Portuguese?

- Null Hypothesis: There is no difference in average grade in both subjects by sex.

```
avg_grade_sex <- longer_data %>%
  group_by(sex, Period, subject) %>%
  summarize(average_grade = mean(Grade), .groups = "drop")

ggplot(avg_grade_sex, aes(x = Period, y = average_grade, fill = sex)) +
  geom_col(position = "dodge") +
  facet_wrap(~subject) +
  labs(title = "Average Grade by Sex Across Periods",
       x = "Period",
       y = "Average Grade",
       fill = "Sex") +
  scale_fill_discrete(labels = c(
    "F" = "Female",
    "M" = "Male"))+
  theme_bw()+
  ylim(0,20)
```



```
sex_test1 <- aov(Grade ~ sex * Period * subject, data = longer_data)
summary(sex_test1)
```

```
##                      Df Sum Sq Mean Sq F value   Pr(>F)
## sex                   1     23   22.60   2.766 0.096417 .
## Period                2    143   71.49   8.750 0.000163 ***
## subject               1    120  120.05  14.694 0.000129 ***
## sex:Period            2      3    1.28   0.157 0.855012
## sex:subject           1    274  273.92  33.527 7.76e-09 ***
## Period:subject        2     21   10.36   1.268 0.281445
## sex:Period:subject    2      1    0.61   0.075 0.927989
## Residuals          2961  24192    8.17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Visual Observation

On average, male students perform better in math class than female students, and vice versa in Portuguese class.

Statistical Test (ANOVA)

1. Sex: p-value = 0.09642 > 0.05, fail to reject H0, sex alone doesn't impact grade significantly
2. sex:subject p-value = 7.8e-09 < 0.05, reject H0, sex significantly impacts grade differently by subject

Answer

Although sex itself isn't an indicator of grade (p-value = 0.09642), the interaction p-value above, 7.8e-09, means that the impact of sex on grade depending on subject is highly significant. This means that gender-based performance differences are subject-specific, which is confirmed by the plot where male student on average perform better in math class and worse in Portuguese class.

Demographic Section Conclusion

There are three main variables in demographics: age (15-19+), address (urban & rural), and sex (male & female).

Age: older student groups consistently receive higher proportions of high grades. The only exception is age group 19+ which consistently performs the worst across all periods for both classes. However, for math, students performance isn't impacted by their age.

Address: In both classes, there is a significant difference in letter grade between urban and rural students for every period. This means rural students consistently perform worse than urban students in both subjects throughout the year. Also, the different emphasis of reason to attend school in may reflect underlying disparities between students from different addresses.

Sex: Gender-based performance differences are subject-specific, which is confirmed by the plot where male student on average perform better in math class and worse in Portuguese class.

Section 4 - Social Support

EDA Overview: Social Support Variables

In this section, we explore select variables associated with "social support." These capture family, school, and external support systems that may assist or influence a student's academic success. Additionally, we explore how the social support variables may interact with demographic variables, such as age and sex.

Data Wrangling & Summary Statistics

```r
df <- read.csv("combined_data_cleaned.csv")

df_ss <- df |> select(age,sex,higher,famrel,G1,G2,G3,subject)

df_ss <- df_ss |> mutate( higher = ifelse(higher=="yes",1,0),
                          male = ifelse(sex=="M",1,0),
                          age_group = ifelse(age >= 19, "19+", as.character(age)) )

longer_data <- df %>%
pivot_longer(cols = G1:G3, names_to = "Period", values_to = "Grade") %>%
mutate(GradeGroup = case_when(
Grade >= 16 ~ "A", # Excellent/Very Good
Grade < 16 ~ "Not A" # Good
))
```

```r
mean_sum <- df_ss |> select(-subject) |>
  summarize(across(everything(), ~ signif(mean(.x, na.rm = TRUE), 4)))

kable(mean_sum|>select(famrel,higher),caption="Mean Values for Famrel and Higher")
```

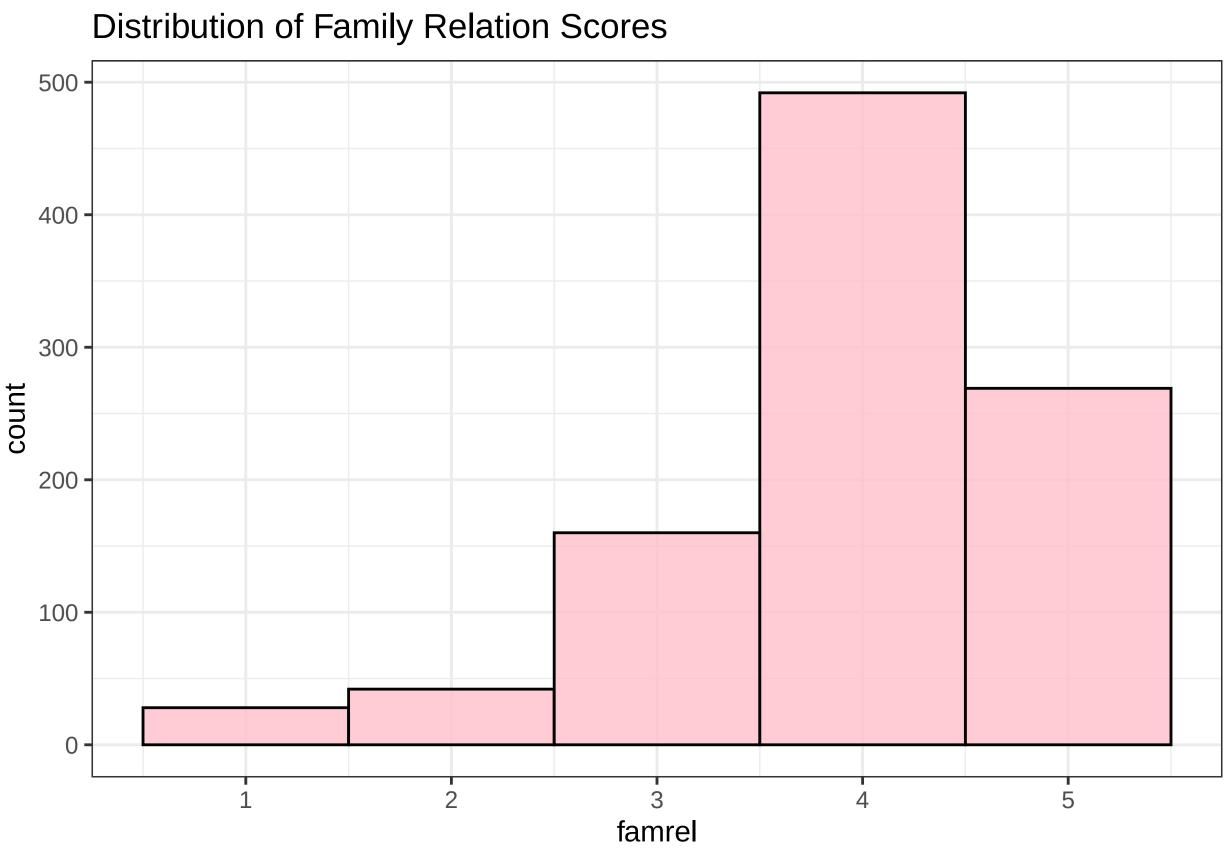
Table 2: Mean Values for Famrel and Higher

famrel	higher
3.94	0.9213

Section 1: Family Relationship

The *famrel* variable measures the quality of a student's family relationships (numeric: from 1 – very bad to 5 – excellent).

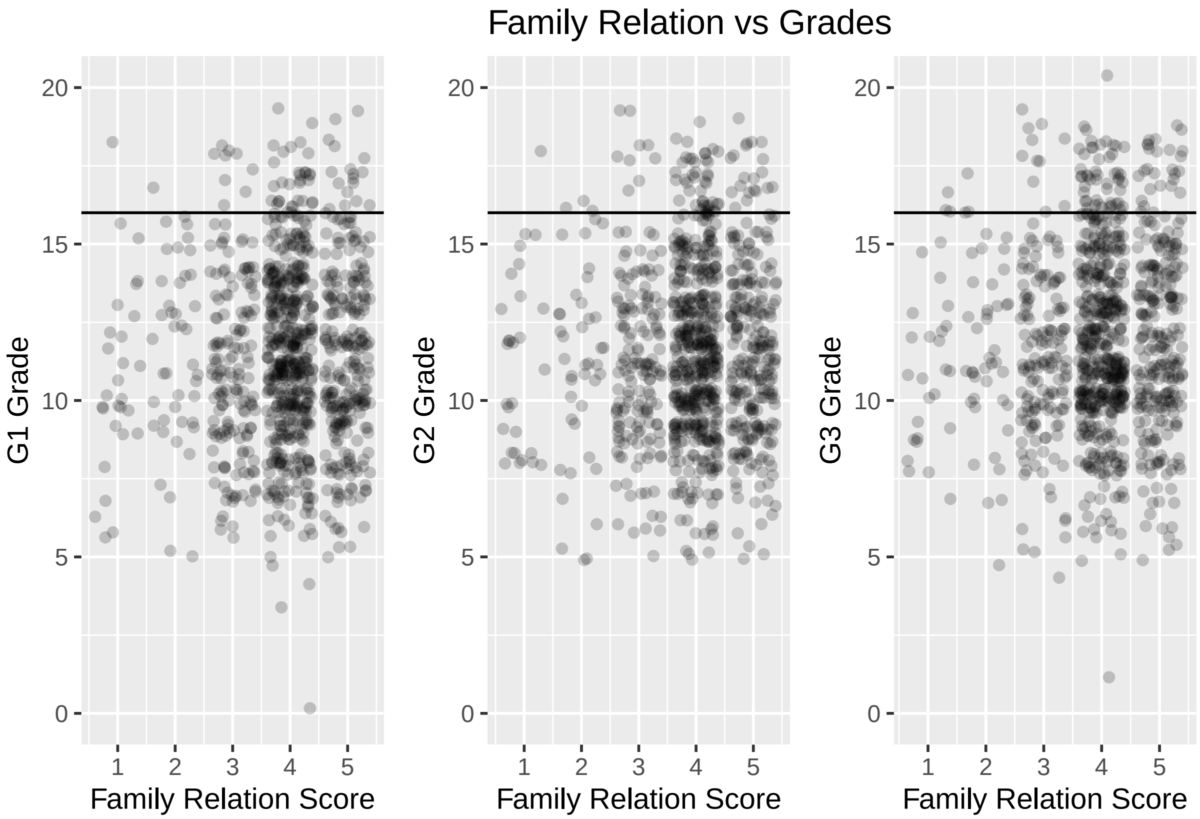
```r
ggplot(df_ss, aes(x=famrel))+geom_histogram(binwidth = 1,color="black",fill="pink",alpha=.8)+
  labs(title="Distribution of Family Relation Scores")+
  theme_bw()
```



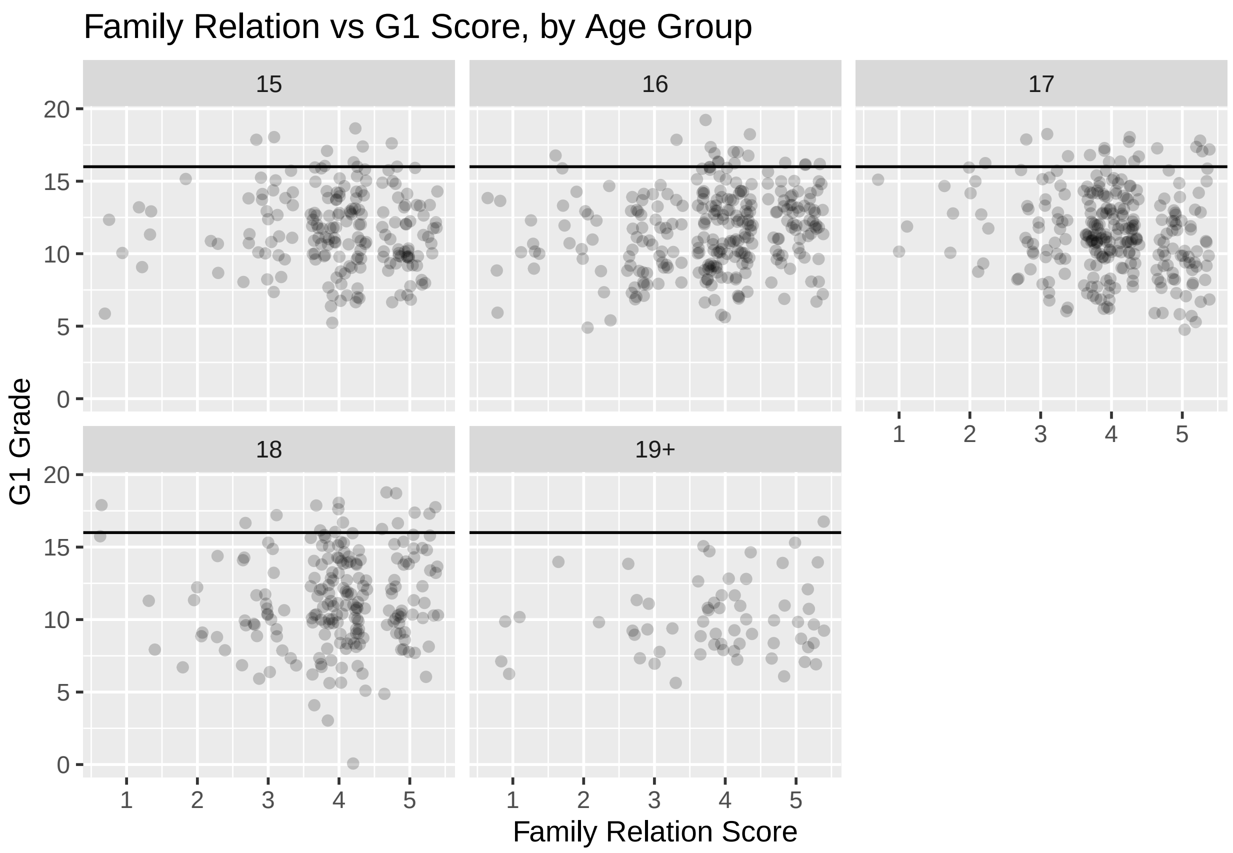
From the histogram, we see that $famrel$ is right-skewed, and from our summary statistic table, the mean $famrel$ score is 3.93.

Q1: Does family relationship quality affect grades?

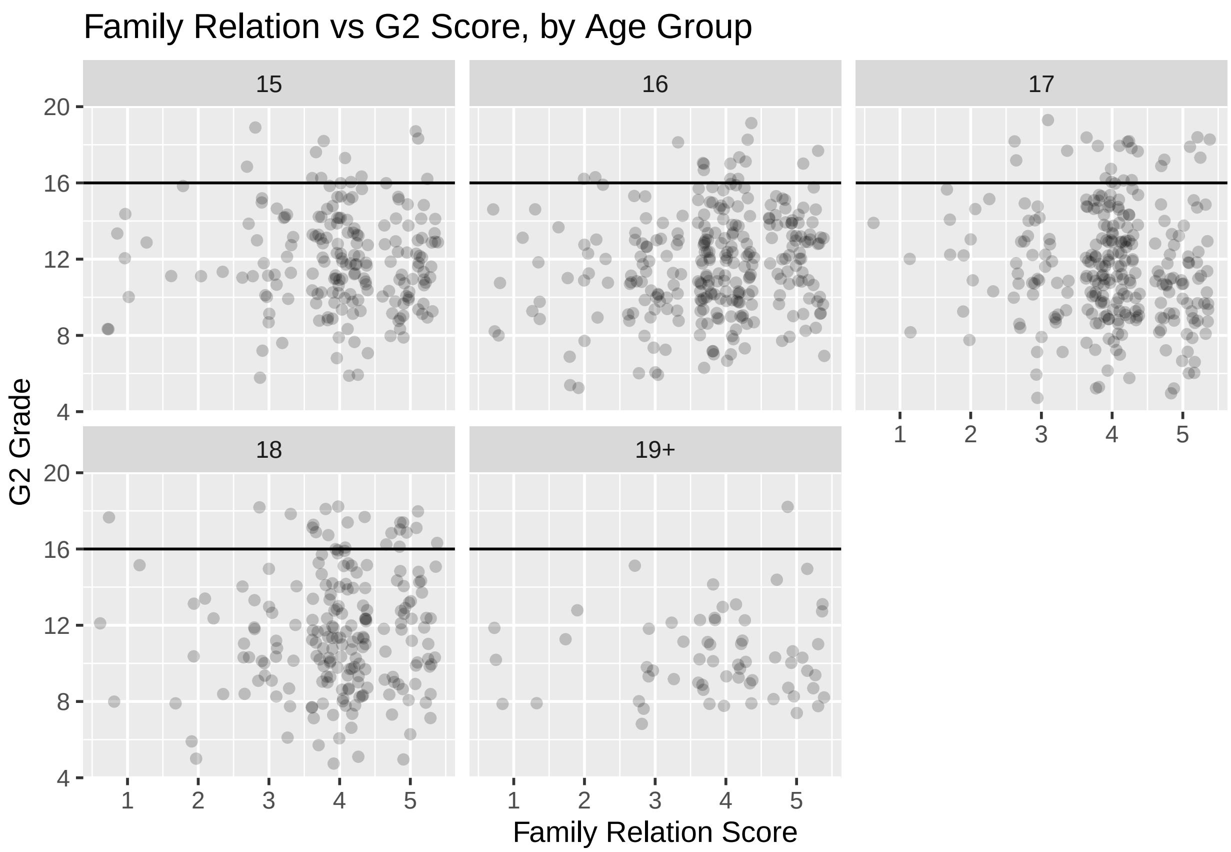
```
g1 <- ggplot(df_ss, aes(x = famrel, y = G1)) +
  geom_jitter(alpha = 0.2, height = NULL) +
  geom_hline(yintercept = 16) +
  coord_cartesian(ylim = c(0, 20)) +
  labs(x = "Family Relation Score", y = "G1 Grade")

g2 <- ggplot(df_ss, aes(x = famrel, y = G2)) +
  geom_jitter(alpha = 0.2, height = NULL) +
  geom_hline(yintercept = 16) +
  coord_cartesian(ylim = c(0, 20)) +
  labs(
    x = "Family Relation Score",
    y = "G2 Grade",
    title = "Family Relation vs Grades"
  )

g3 <- ggplot(df_ss, aes(x = famrel, y = G3)) +
  geom_jitter(alpha = 0.2, height = NULL) +
  geom_hline(yintercept = 16) +
  coord_cartesian(ylim = c(0, 20)) +
  labs(x = "Family Relation Score", y = "G3 Grade")

g1 + g2 + g3 + plot_layout(nrow = 1)
```



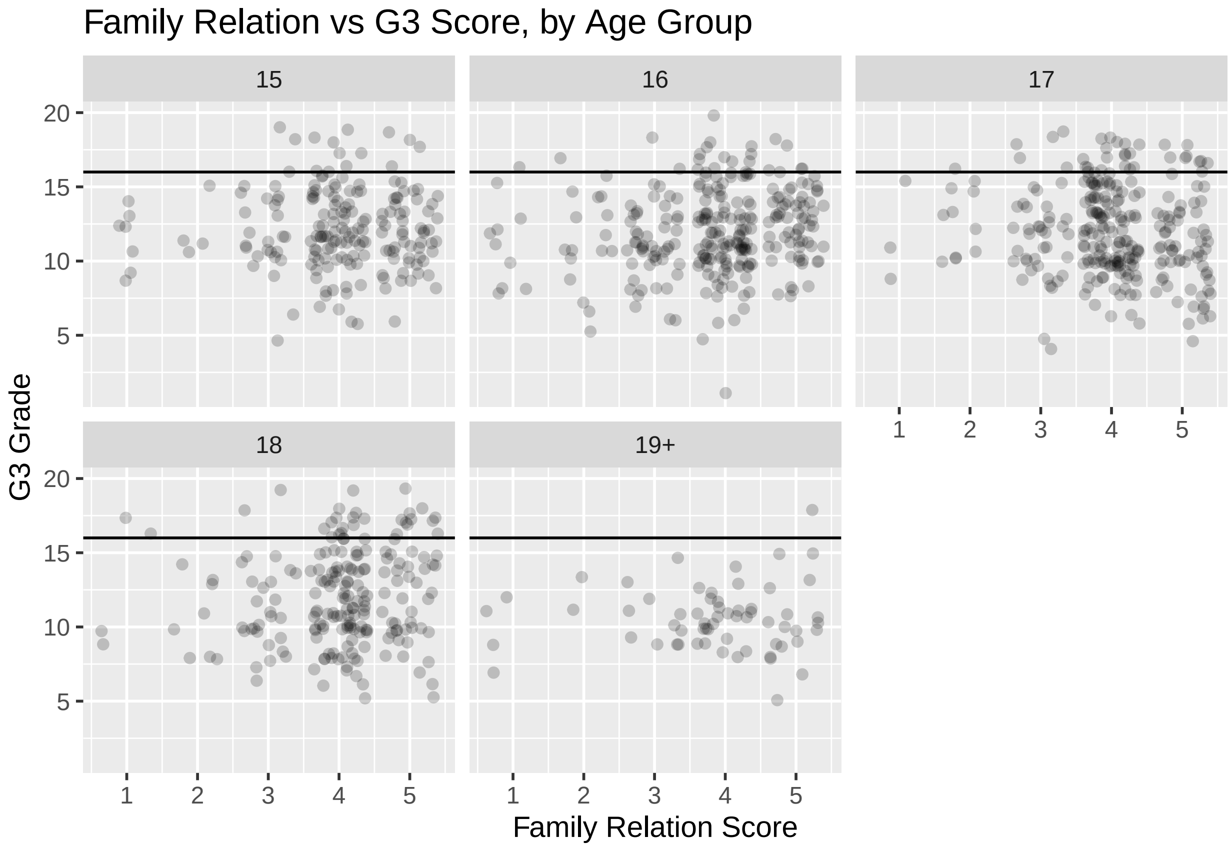
```
ggplot(df_ss,aes(x=famrel,y=G1))+
  geom_jitter(alpha=0.2,height=NULL) + geom_hline(yintercept=16)+
  facet_wrap(~age_group)+
  labs(x= "Family Relation Score", y = "G1 Grade",
       title="Family Relation vs G1 Score, by Age Group")
```



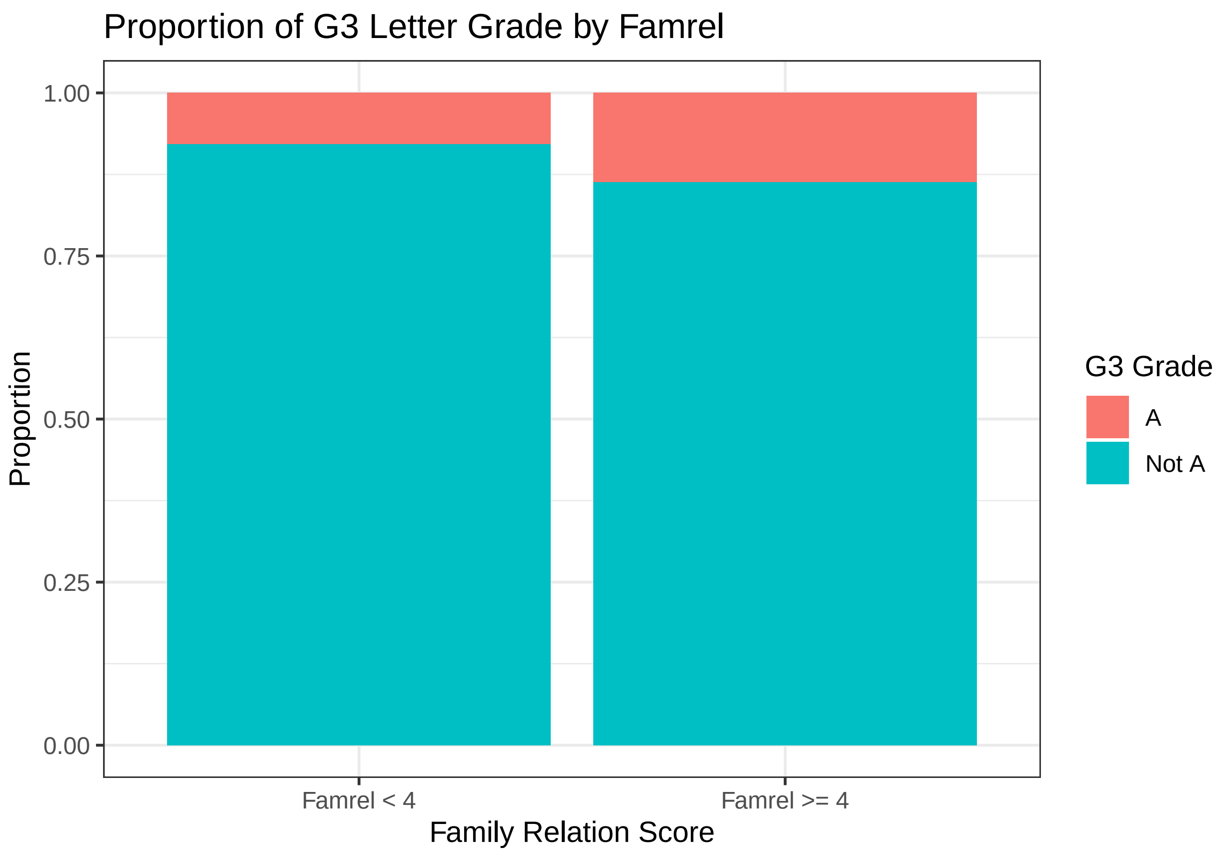
```
ggplot(df_ss,aes(x=famrel,y=G2))+
  geom_jitter(alpha=0.2,height=NULL) + geom_hline(yintercept=16)+
  facet_wrap(~age_group)+
  labs(x= "Family Relation Score", y = "G2 Grade",
       title="Family Relation vs G2 Score, by Age Group")
```



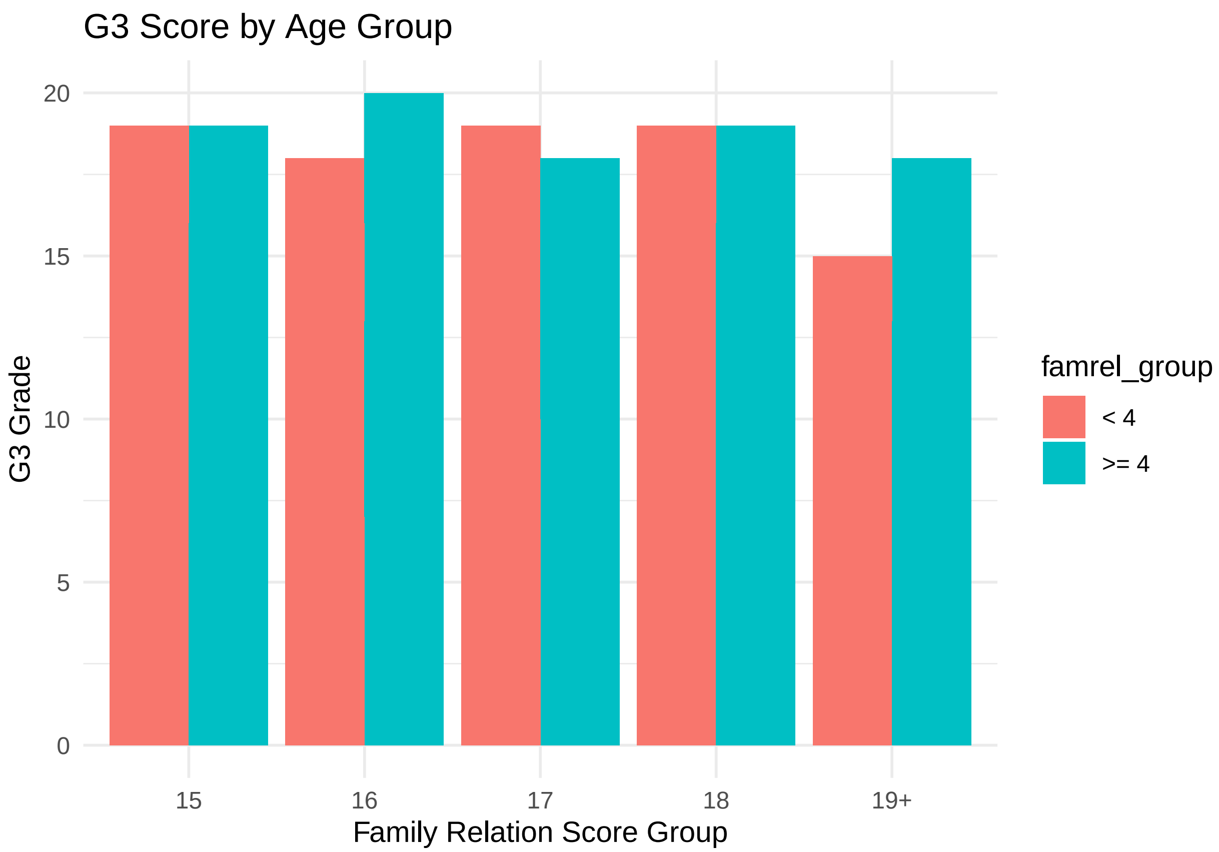
```r
ggplot(df_ss,aes(x=famrel,y=G3))+
  geom_jitter(alpha=0.2,height=NULL) + geom_hline(yintercept=16)+
  facet_wrap(~age_group)+
  labs(x= "Family Relation Score", y = "G3 Grade",
       title="Family Relation vs G3 Score, by Age Group")
```



```
longer_data_19plus <- longer_data %>%
mutate(famrel_4 = factor(ifelse(famrel>=4, 1,0)),age_group = ifelse(age >= 19, "19+", as.character(age)
  filter(Period == "G3")
ggplot(longer_data_19plus, aes(x = famrel_4, fill = GradeGroup)) +
geom_bar(position = "fill") +
scale_x_discrete(labels=c("Famrel < 4","Famrel >= 4"))+
labs(title = "Proportion of G3 Letter Grade by Famrel",
x = "Family Relation Score", y = "Proportion",
fill = "G3 Grade") +
theme_bw()
```



```
# Step 1: Create famrel group
df_ss <- df_ss %>%
  mutate(famrel_group = ifelse(famrel >= 4, ">= 4", "< 4"))

# Step 2: Plot
ggplot(df_ss, aes(x = age_group, y = G3, fill = famrel_group)) +
  geom_col(position="dodge") +
  labs(
    x = "Family Relation Score Group",
    y = "G3 Grade",
    title = "G3 Score by Age Group"
  ) +
  theme_minimal()
```



From the scatter plots, we see that across all test periods (G1-G3),those with famrel scores of 3 or more have more scores greater than 16 (an A equivalent). Faceting based on Age Group also shows a similar trend, especially for students 18 or under. It is possible that having good family relationship increases the likelihood of receiving an A. The next question is how to define a "good" family relationship.

```
df_ss_g1 <- df_ss |> mutate(gradeA = G1 >= 16, goodFamrel = famrel >= 4)
chisq_test <- chisq.test(table(df_ss_g1$gradeA, df_ss_g1$goodFamrel))
chisq_test
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(df_ss_g1$gradeA, df_ss_g1$goodFamrel)
## X-squared = 0.51225, df = 1, p-value = 0.4742
```

```
df_ss_g2 <- df_ss |> mutate(gradeA = G2 >= 16, goodFamrel = famrel >= 4)
chisq_test <- chisq.test(table(df_ss_g2$gradeA, df_ss_g2$goodFamrel))
chisq_test
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(df_ss_g2$gradeA, df_ss_g2$goodFamrel)
## X-squared = 2.4647, df = 1, p-value = 0.1164
```

```
df_ss_g3 <- df_ss |> mutate(gradeA = G3 >= 16, goodFamrel = famrel >= 4)
chisq_test <- chisq.test(table(df_ss_g3$gradeA, df_ss_g3$goodFamrel))
chisq_test
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(df_ss_g3$gradeA, df_ss_g3$goodFamrel)
## X-squared = 5.0524, df = 1, p-value = 0.02459
```

We find that when defining a "good" family relationship to be $famrel \geq 4$, at the 5% level we reject H_0 of the Chi-squared test for periods G3, so there is a relationship between (1) getting an A and (2) having a good family relationship in those periods. Notably, we fail to reject H_0 for G1 and G2. We continue the exploration using this definition: goodFamrel = $(famrel \geq 4)$.

```
famrel_prop <- df_ss_g3 |> group_by(goodFamrel) |>
  count(gradeA) |>
  mutate(prop_4_A = n / sum(n))
kable(famrel_prop)
```

goodFamrel	gradeA	n	prop_4_A
FALSE	FALSE	212	0.9217391
FALSE	TRUE	18	0.0782609
TRUE	FALSE	657	0.8633377
TRUE	TRUE	104	0.1366623

```
x <- c(18, 104)
n_total <- c(230, 761)
prop.test(x = x, n = n_total, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  x out of n_total
## X-squared = 5.5802, df = 1, p-value = 0.01816
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.10083251 -0.01597032
## sample estimates:
##     prop 1     prop 2
## 0.07826087 0.13666229
```

In G3, the percentage of students with As is 13.67% for those with good family relationship, whereas it is 7.83% for those without a good family relationship. Running a two-proportion z-test, we reject H_0; the 5.84% difference in proportions of students getting As is statistically significant.

Section 2: Students' Desire to Pursue Higher Education

Higher is a binary variable indicating whether or not the student intends to pursue higher education after high school. We hypothesize that wanting to attend higher education will increase G3 grades in school.

From the earlier table of means, we know that 92.13% of the students want to attend higher education (i.e. higher = 1). Does this differ with sex or age?

Q1: Does Wanting to Pursue Higher Education Affect Grades?

```
higher_prop <- df_ss_g1 |> group_by(higher) |> count(gradeA) |> mutate(prop_higher_A = n / sum(n))
kable(higher_prop,caption="Contingency Table: Higher vs A-Proportion (G1)")
```

Table 4: Contingency Table: Higher vs A-Proportion (G1)

higher	gradeA	n	prop_higher_A
0	FALSE	78	1.0000000
1	FALSE	826	0.9047097
1	TRUE	87	0.0952903

```
tbl <- matrix(c(78, 0, 826, 87), nrow = 2, byrow = TRUE)
rownames(tbl) <- c("higher = 0", "higher = 1")
colnames(tbl) <- c("not_A", "A")
fisher.test(tbl)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tbl
## p-value = 0.001188
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  2.123597      Inf
## sample estimates:
## odds ratio
##        Inf
```

```
higher_prop <- df_ss_g2 |> group_by(higher) |> count(gradeA) |> mutate(prop_higher_A = n / sum(n))
kable(higher_prop,caption="Contingency Table: Higher vs A-Proportion (G2)")
```

Table 5: Contingency Table: Higher vs A-Proportion (G2)

higher	gradeA	n	prop_higher_A
0	FALSE	78	1.000000
1	FALSE	820	0.898138
1	TRUE	93	0.101862

```
tbl <- matrix(c(78, 0, 820, 93), nrow = 2, byrow = TRUE)
rownames(tbl) <- c("higher = 0", "higher = 1")
colnames(tbl) <- c("not_A", "A")
fisher.test(tbl)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tbl
## p-value = 0.0007667
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  2.289602      Inf
## sample estimates:
## odds ratio
##        Inf
```

```
higher_prop <- df_ss_g3 |> group_by(higher) |> count(gradeA) |> mutate(prop_higher_A = n / sum(n))
kable(higher_prop,caption="Contingency Table: Higher vs A-Proportion (G3)")
```

Table 6: Contingency Table: Higher vs A-Proportion (G3)

higher	gradeA	n	prop_higher_A
0	FALSE	78	1.0000000
1	FALSE	791	0.8663746
1	TRUE	122	0.1336254

```
tbl <- matrix(c(78, 0, 791, 122), nrow = 2, byrow = TRUE)
rownames(tbl) <- c("higher = 0", "higher = 1")
colnames(tbl) <- c("not_A", "A")
fisher.test(tbl)
```

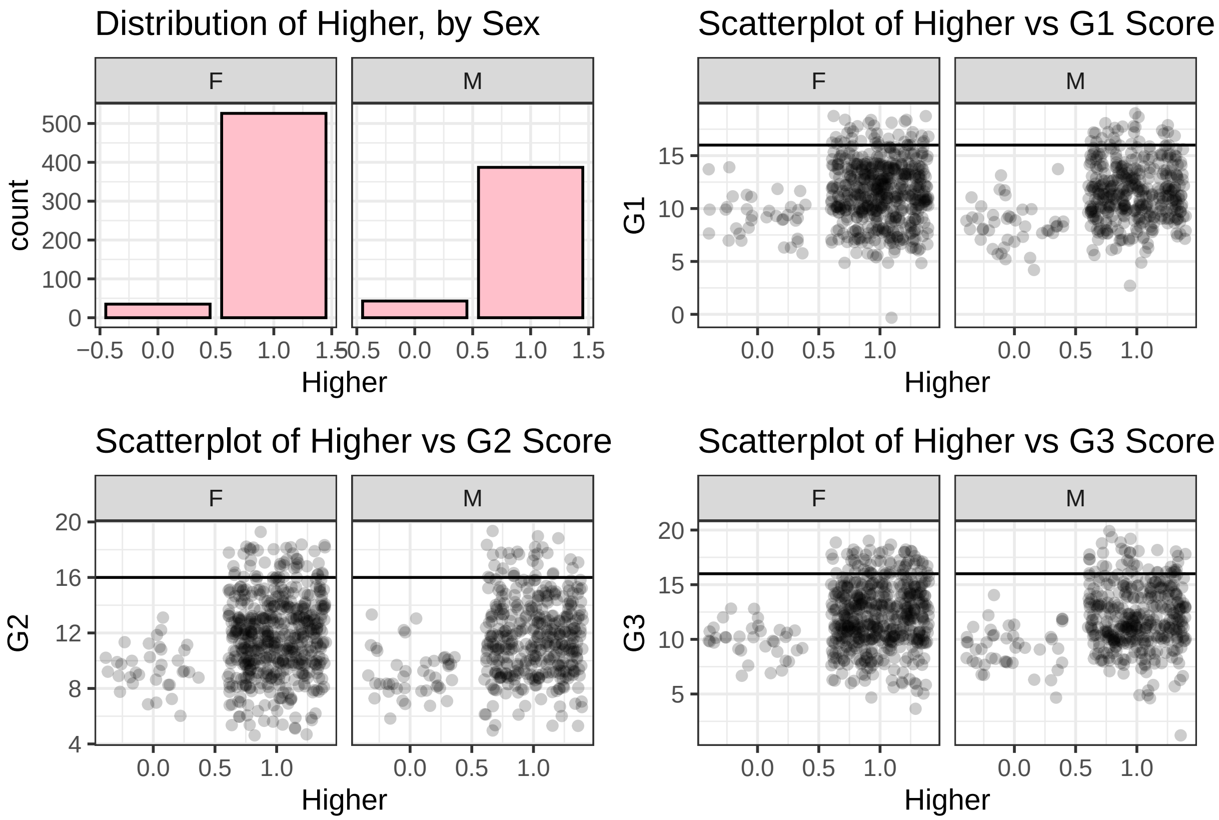
```
##
##  Fisher's Exact Test for Count Data
##
## data:  tbl
## p-value = 4.021e-05
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  3.127664      Inf
## sample estimates:
## odds ratio
##        Inf
```

First, we want to determine if higher is correlated with getting an A. Running respective Fisher Tests, we find similar conclusions for scores in all periods - we reject H_0. There is a very strong (technically infinite) association between higher and getting an A, so a relationship between the two variables is likely.

Q2: Does the effect of the variable "higher" grades depend on student sex?

```
higher_hist <- ggplot(df_ss,aes(x=higher))+
  geom_bar(fill="pink",color="black") + facet_wrap(~sex) +
  labs(title="Distribution of Higher, by Sex",
       x="Higher","Count") + theme_bw()
```

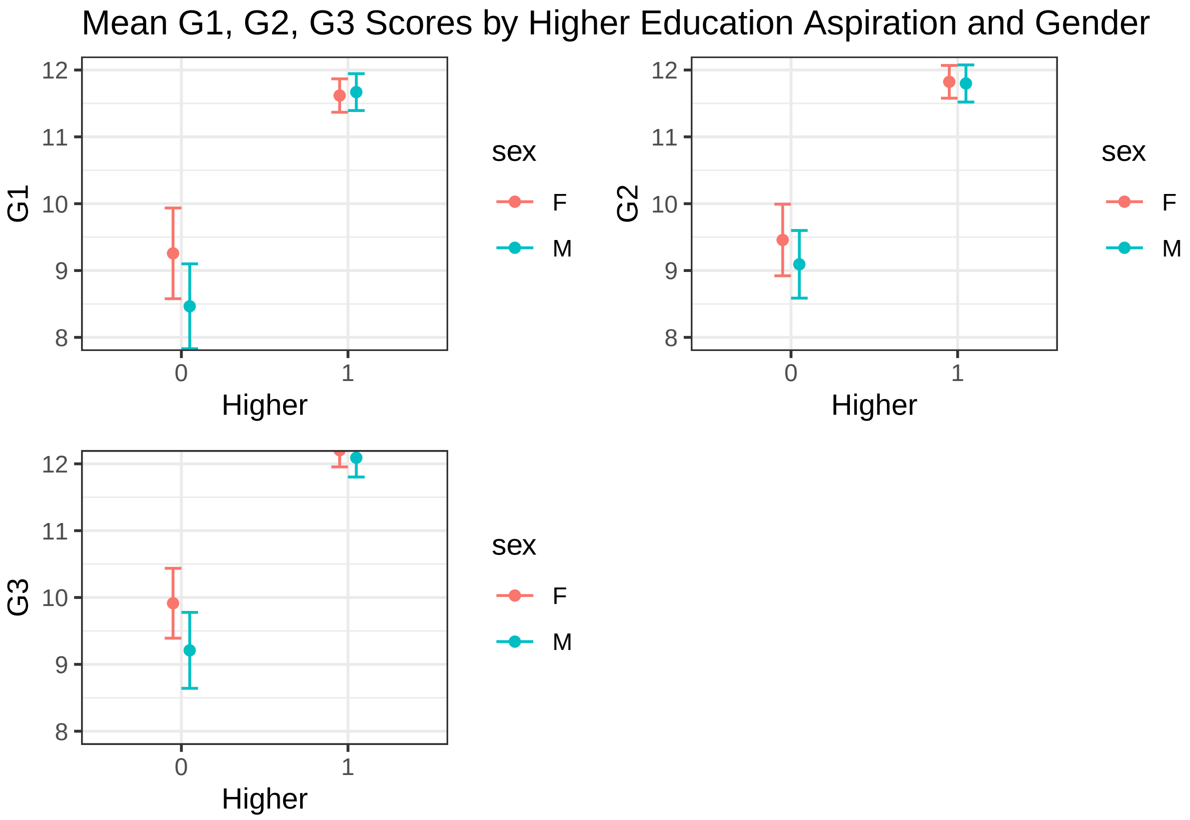
```r
jitter_g1 <- ggplot(df_ss,aes(x=higher,y=G1))+geom_jitter(alpha=0.2,height=NULL) +
  facet_wrap(~sex) + geom_hline(yintercept=16) +
  labs(title="Scatterplot of Higher vs G1 Score",x="Higher") +theme_bw()

jitter_g2 <- ggplot(df_ss,aes(x=higher,y=G2))+geom_jitter(alpha=0.2,height=NULL) +
  facet_wrap(~sex) + geom_hline(yintercept=16) +
  labs(title="Scatterplot of Higher vs G2 Score",x="Higher") +theme_bw()

jitter_g3 <- ggplot(df_ss,aes(x=higher,y=G3))+geom_jitter(alpha=0.2,height=NULL) +
  facet_wrap(~sex) + geom_hline(yintercept=16) +
  labs(title="Scatterplot of Higher vs G3 Score",x="Higher") +theme_bw()

higher_hist + jitter_g1 + jitter_g2 + jitter_g3 + plot_layout(ncol=2)
```



From the dotplot, for both ages, no one with *higher* $= 0$ got an A (in any period), so it is unlikely that the effect of *higher* on $G3$ differs based on gender.

```r
# G1
error_g1 <- ggplot(df_ss, aes(x = factor(higher), y = G1, color = sex)) +
  stat_summary(fun = mean, geom = "point", position = position_dodge(0.2)) +
  stat_summary(fun.data = mean_cl_normal, geom = "errorbar",
               width = 0.2, position = position_dodge(0.2)) +
  coord_cartesian(ylim = c(8, 12)) +   # <-- set y-axis scale
  labs(x = "Higher", y = "G1") +
  theme_bw()
```

```
# G2
error_g2 <- ggplot(df_ss, aes(x = factor(higher), y = G2, color = sex)) +
  stat_summary(fun = mean, geom = "point", position = position_dodge(0.2)) +
  stat_summary(fun.data = mean_cl_normal, geom = "errorbar",
               width = 0.2, position = position_dodge(0.2)) +
  coord_cartesian(ylim = c(8, 12)) +   # <-- set y-axis scale
  labs(x = "Higher", y = "G2") +
  theme_bw()

# G3
error_g3 <- ggplot(df_ss, aes(x = factor(higher), y = G3, color = sex)) +
  stat_summary(fun = mean, geom = "point", position = position_dodge(0.2)) +
  stat_summary(fun.data = mean_cl_normal, geom = "errorbar",
               width = 0.2, position = position_dodge(0.2)) +
  coord_cartesian(ylim = c(8, 12)) +   # <-- set y-axis scale
  labs(x = "Higher", y = "G3") +
  theme_bw()

# Combine into one layout
error_g1 + labs(title = "Mean G1, G2, G3 Scores by Higher Education Aspiration and Gender") +
  error_g2 + error_g3 +
  plot_layout(ncol = 2)
```



Since the error bars for mean score by gender overlap (for both values of *higher*), we do not have enough evidence to say that the effect of *higher* on G1, G2, and G3 varies based on sex. This confirms our initial suspicion.

Q3: Does proportion of students wanting to pursue higher education change with age?

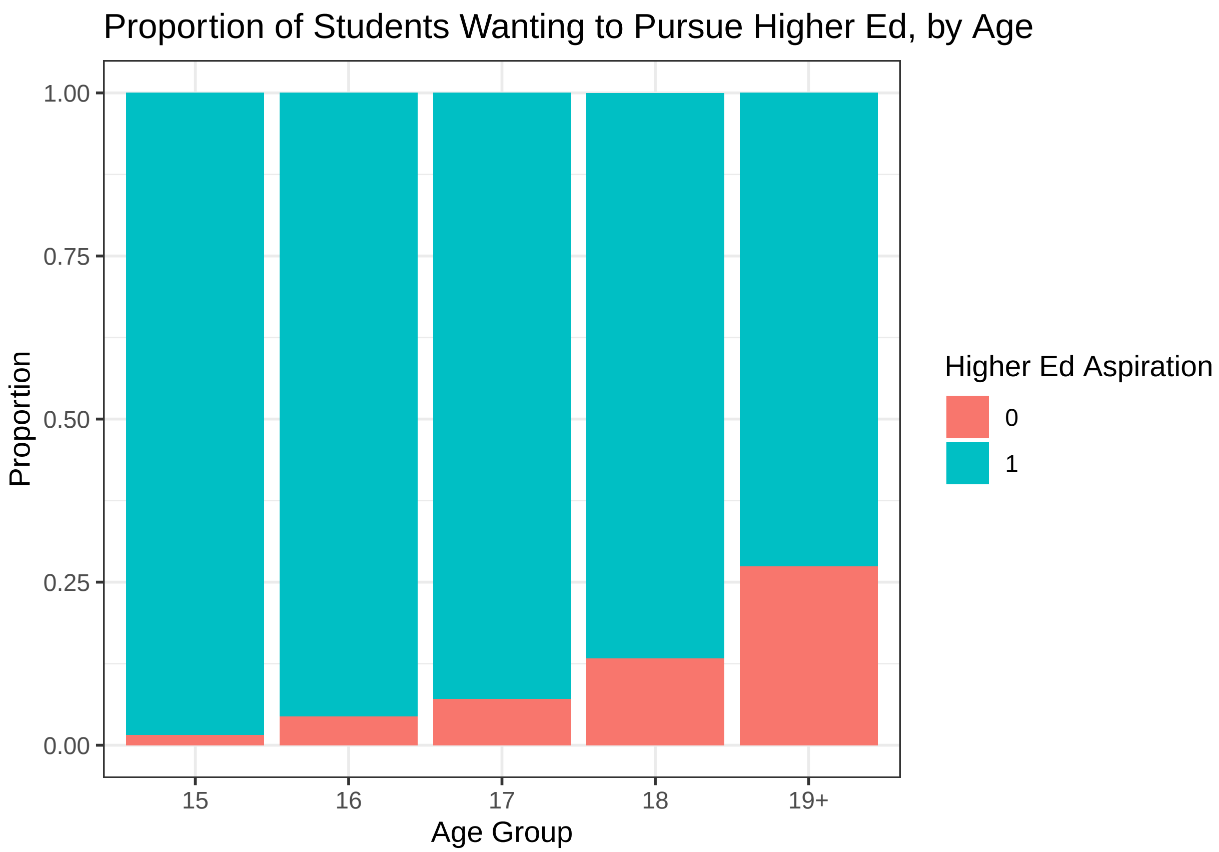
We would also like to explore if proportion of students with *higher* = 1 decreases with age. This is a different exploration that focuses on the interaction between a social and demographic factor, rather than just focusing on grades.

```r
df_higher_prop <- df_ss |> group_by(age_group) |>
  count(higher) |>  mutate(prop = n / sum(n)) #|> filter(higher==1)
kable(df_higher_prop|>select(-higher))
```

age_group	n	prop
15	3	0.0159574
15	185	0.9840426
16	12	0.0441176
16	260	0.9558824
17	19	0.0711610
17	248	0.9288390
18	27	0.1336634
18	175	0.8663366
19+	17	0.2741935
19+	45	0.7258065

The table shows that the proportion of students who want to attend higher education decreases with age, and the graph below provides a way to visualize this negative correlation.

```r
df_higher_prop <- df_higher_prop |> mutate(higher = factor(higher,levels=c(0,1)))

ggplot(df_higher_prop, aes(x = age_group, y = prop, fill = factor(higher), group = 1)) +
  geom_col(position=position_stack(reverse=F)) +
  labs(title = "Proportion of Students Wanting to Pursue Higher Ed, by Age",
       x = "Age Group", y = "Proportion",
       fill = "Higher Ed Aspiration") +
  theme_bw()
```



From the graph above, we can visualize the negative correlation between *age* and *higher*. As age increases, higher education aspiration decreases.

Social Support Section Conclusion

Exploring *famrel* and *higher* from the social support variable category, we find that the proportion of students getting As in G3 is almost 6% higher when family relationship is good. We also find that having a desire to achieve higher education is correlated with getting an A, but there is no gender-based difference in this "effect." This is true for all grade periods. Finally, we find that the proportion of students wanting to pursue higher education decreases with age, with a negative correlation of around -0.233. This aligns with a finding in the Demographic section - that older students get lower grades, particularly in Portuguese.

Section 5 - Behavioural Vairables

5.1 - Traveltime, Studytime, Freetime, Goout

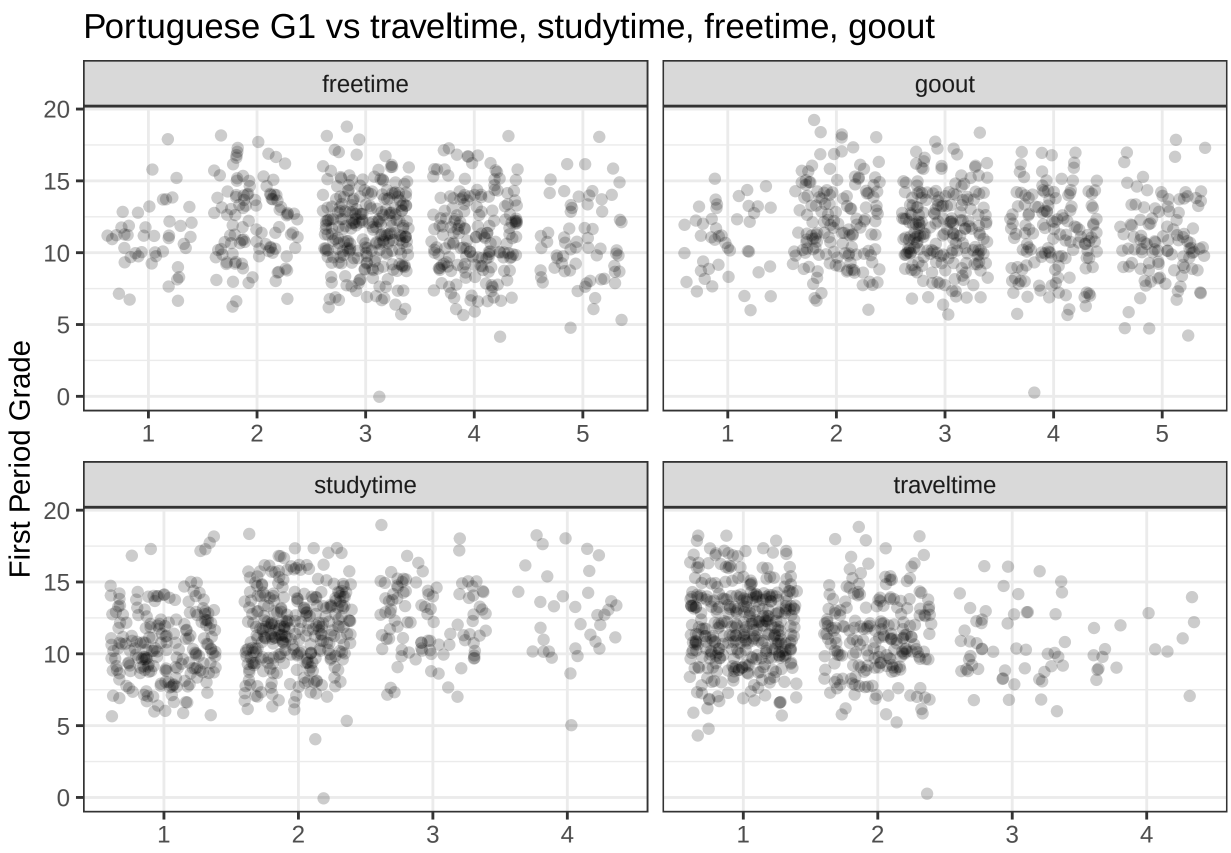
There are 10 variables in this section, each with a different y-axis, so they are divided into four groups. The first group is time-related variables, including traveltime, studytime, freetime, and goout. The second group focuses on alcohol consumption, with Dalc and Walc. The third group includes activities, which is a binary variables (yes/no). The remaining three variables — health, absences, and failures — form the fourth group. We will analyze each group separately.

Q1: How do studytime, traveltime, freetime, and goout affect students' grades in Portuguese and Mathematics across G1–G3?

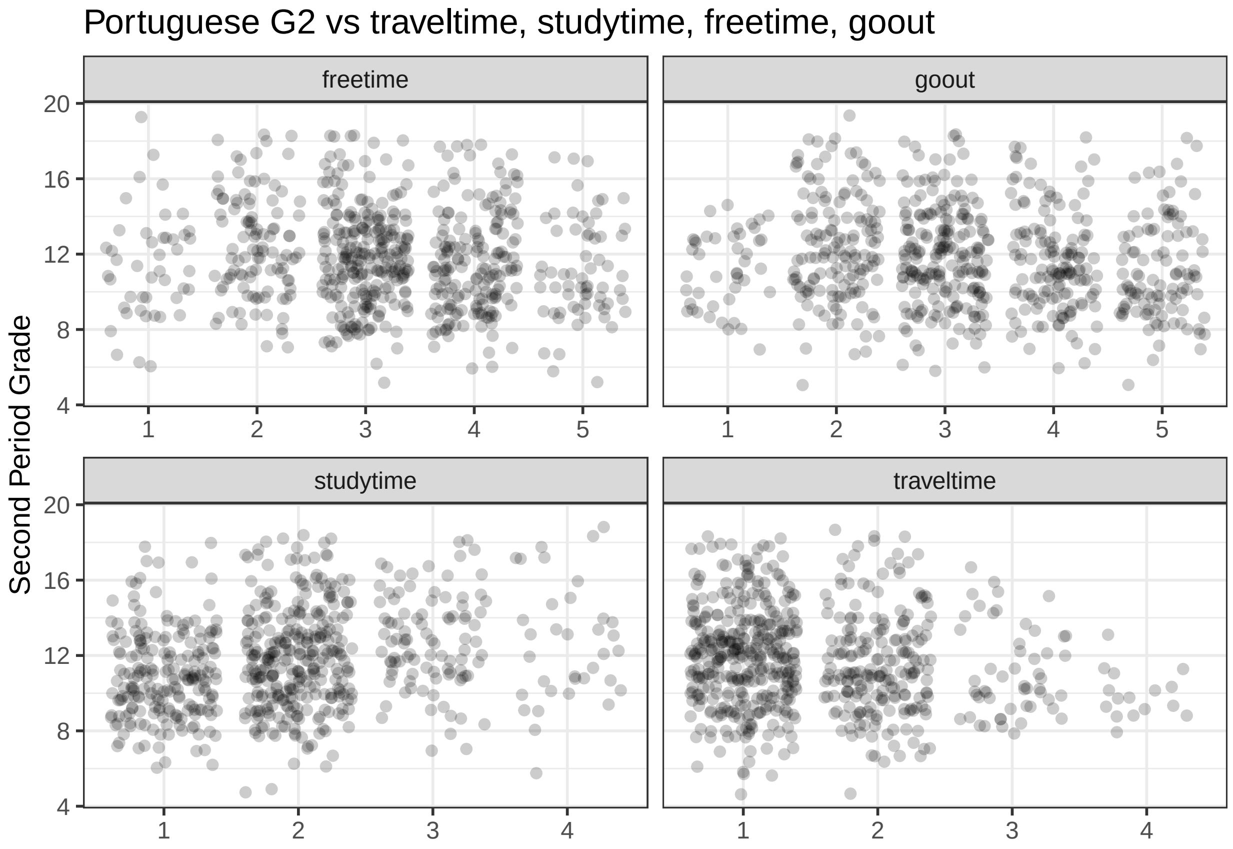
- Null Hypothesis: There is no significant difference in students' letter grades across different levels of traveltime, studytime, freetime, and goout.

Traveltime, Studytime, Freetime, Goout Distribution

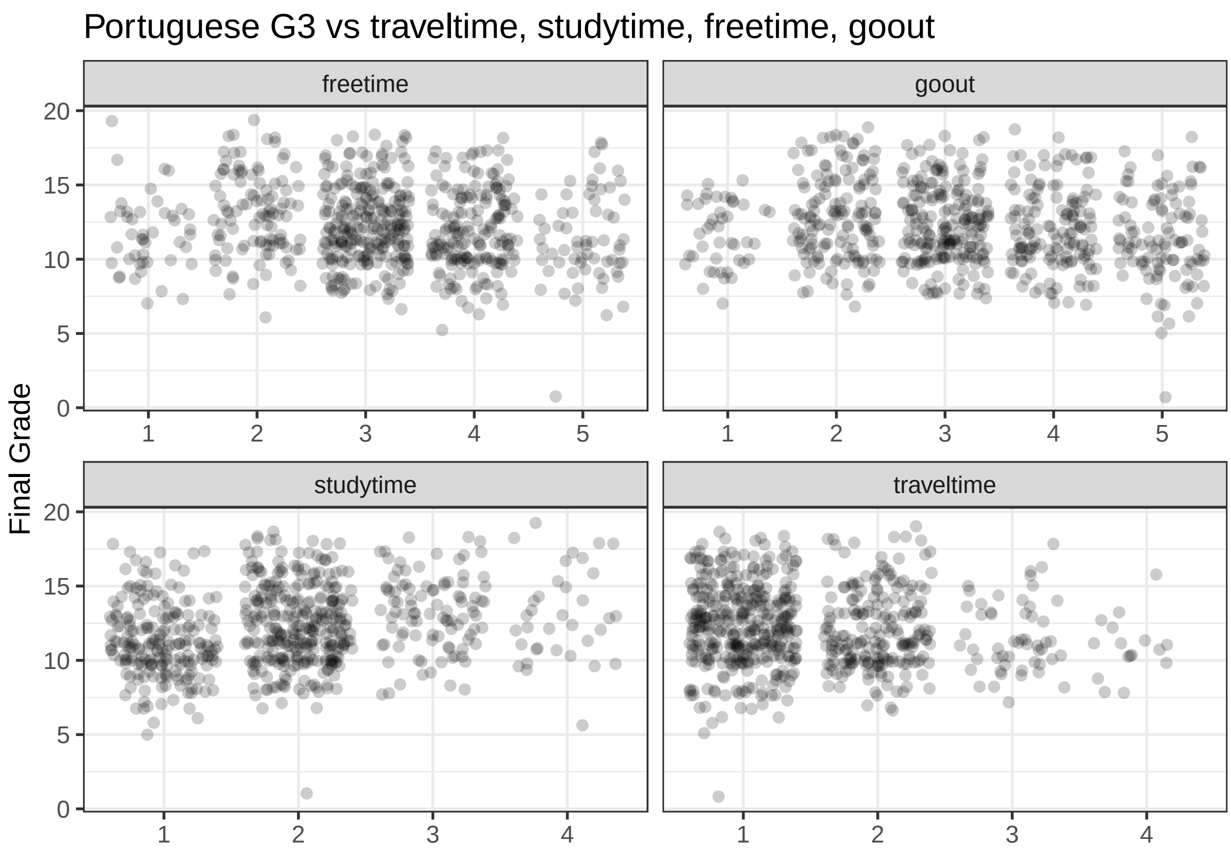
```
g1 <- read_csv("port_data_cleaned.csv") |>
  select(traveltime, studytime, freetime, goout, G1) |>
  mutate(across(-G1, as.character)) |>
  pivot_longer(-G1, names_to = "Variable", values_to = "Value")

ggplot(g1, aes(x = as.factor(Value), y = G1)) +
  geom_jitter(alpha = 0.2, height = NULL) +
  facet_wrap(~ Variable, scales = "free_x") +
  labs(title = "Portuguese G1 vs traveltime, studytime, freetime, goout", x = NULL, y = "First Period G
  theme_bw()
```



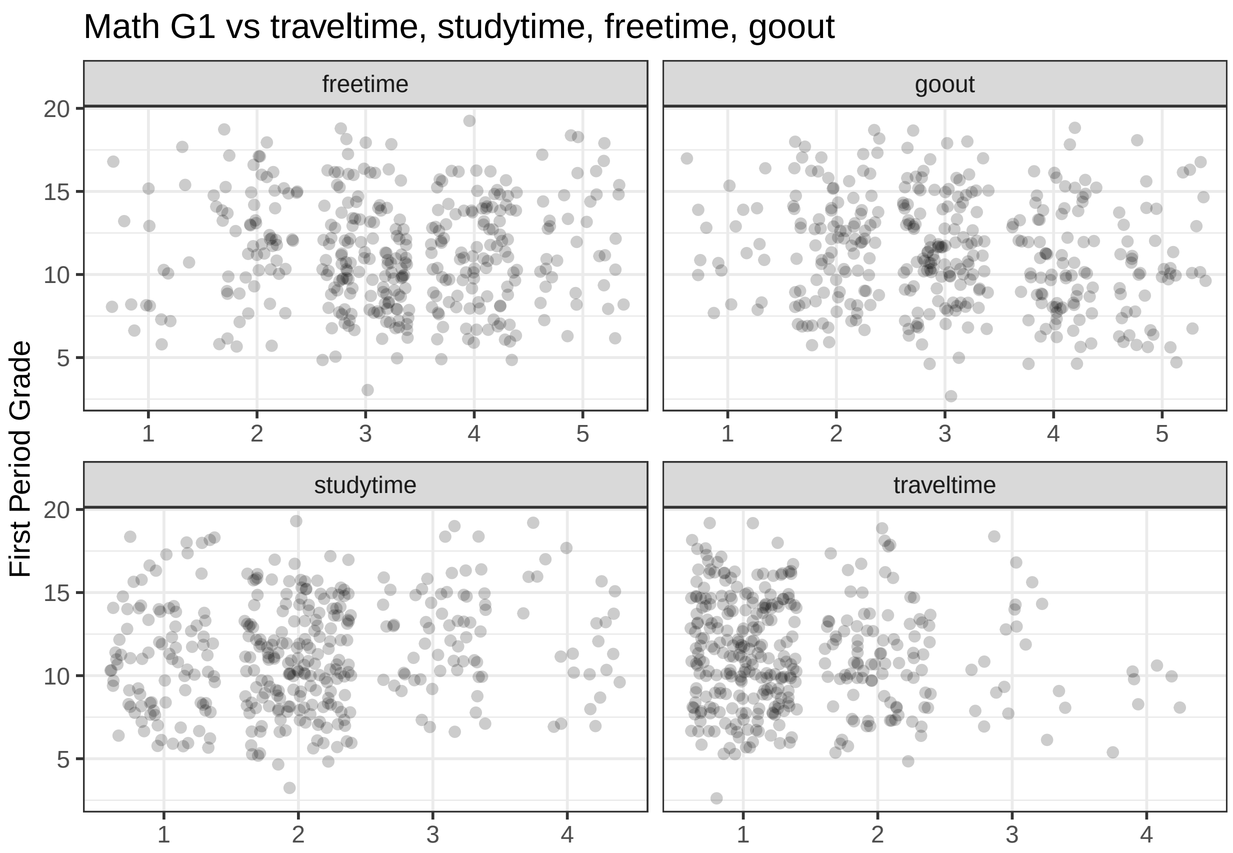
```
g2 <- read_csv("port_data_cleaned.csv") |>
  select(traveltime, studytime, freetime, goout, G2) |>
  mutate(across(-G2, as.character)) |>
  pivot_longer(-G2, names_to = "Variable", values_to = "Value")

ggplot(g2, aes(x = as.factor(Value), y = G2)) +
  geom_jitter(alpha = 0.2, height = NULL) +
  facet_wrap(~ Variable, scales = "free_x") +
  labs(title = "Portuguese G2 vs traveltime, studytime, freetime, goout", x = NULL, y = "Second Period
  theme_bw()
```



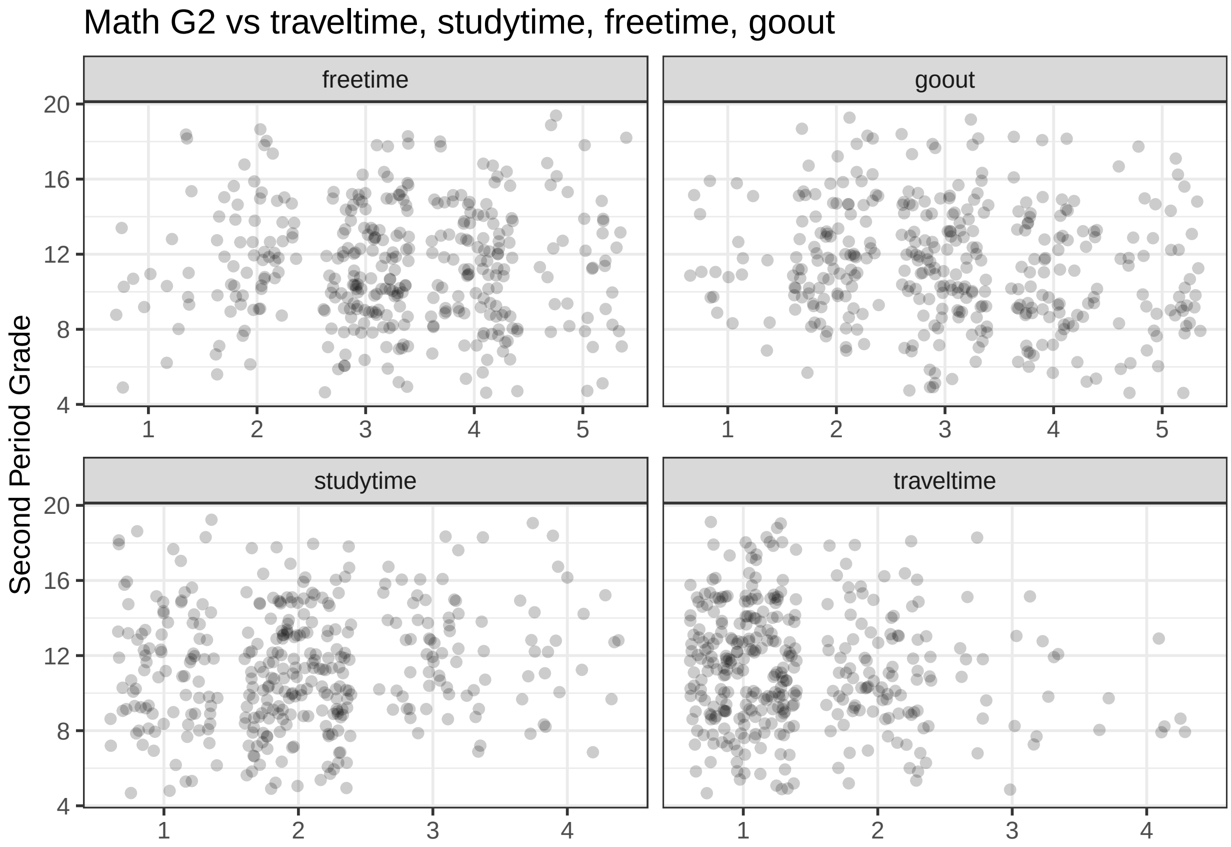
```
g3 <- read_csv("port_data_cleaned.csv") |>
  select(traveltime, studytime, freetime, goout, G3) |>
  mutate(across(-G3, as.character)) |>
  pivot_longer(-G3, names_to = "Variable", values_to = "Value")

ggplot(g3, aes(x = as.factor(Value), y = G3)) +
  geom_jitter(alpha = 0.2, height = NULL) +
  facet_wrap(~ Variable, scales = "free_x") +
  labs( title = "Portuguese G3 vs traveltime, studytime, freetime, goout", x = NULL,  y = "Final Grade"
  theme_bw()
```



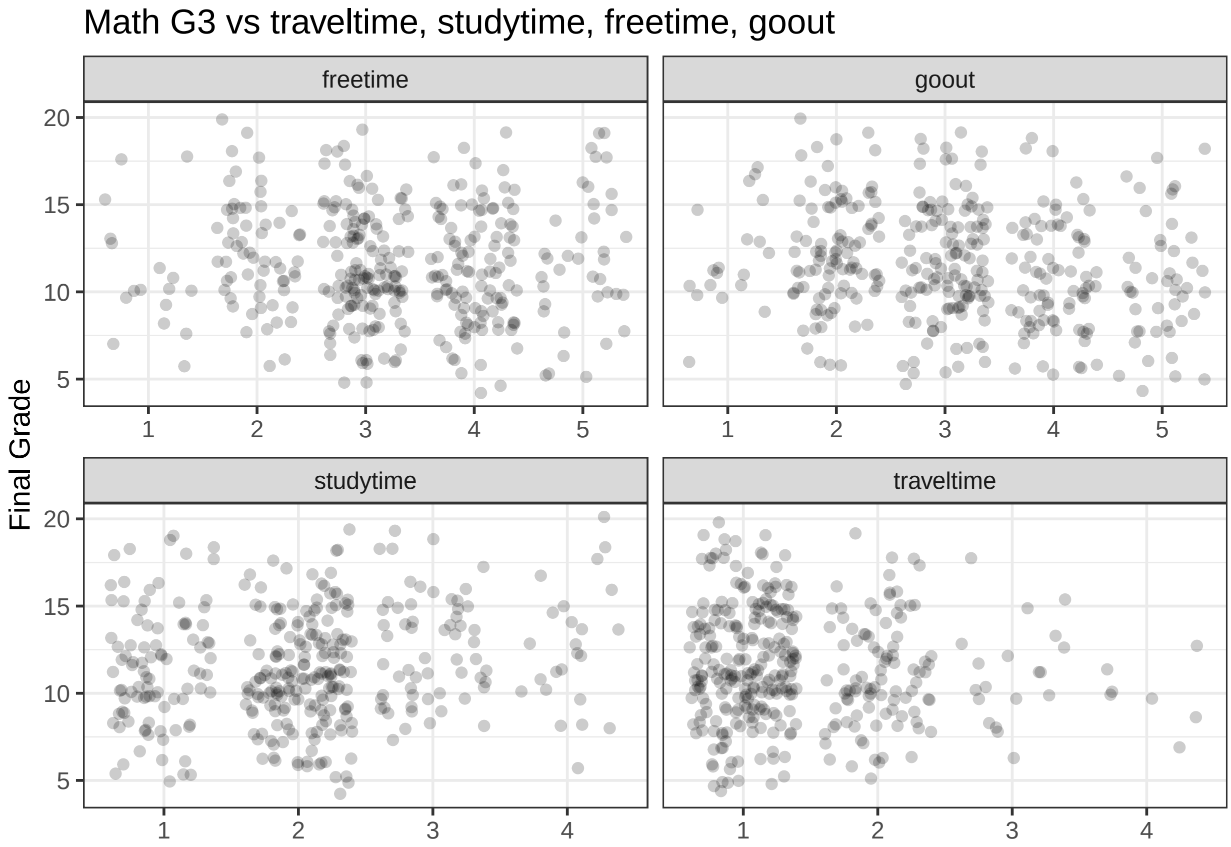
```
g1 <- read_csv("math_data_cleaned.csv") |>
  select(traveltime, studytime, freetime, goout, G1) |>
  mutate(across(-G1, as.character)) |>
  pivot_longer(-G1, names_to = "Variable", values_to = "Value")

ggplot(g1, aes(x = as.factor(Value), y = G1)) +
  geom_jitter(alpha = 0.2, height = NULL) +
  facet_wrap(~ Variable, scales = "free_x") +
  labs(title = "Math G1 vs traveltime, studytime, freetime, goout", x = NULL, y = "First Period Grade")
  theme_bw()
```



```
g2 <- read_csv("math_data_cleaned.csv") |>
  select(traveltime, studytime, freetime, goout, G2) |>
  mutate(across(-G2, as.character)) |>
  pivot_longer(-G2, names_to = "Variable", values_to = "Value")

ggplot(g2, aes(x = as.factor(Value), y = G2)) +
  geom_jitter(alpha = 0.2, height = NULL) +
  facet_wrap(~ Variable, scales = "free_x") +
  labs(title = "Math G2 vs traveltime, studytime, freetime, goout", x = NULL, y = "Second Period Grade"
  theme_bw()
```



```r
g3 <- read_csv("math_data_cleaned.csv") |>
  select(traveltime, studytime, freetime, goout, G3) |>
  mutate(across(-G3, as.character)) |>
  pivot_longer(-G3, names_to = "Variable", values_to = "Value")

ggplot(g3, aes(x = as.factor(Value), y = G3)) +
  geom_jitter(alpha = 0.2, height = NULL) +
  facet_wrap(~ Variable, scales = "free_x") +
  labs( title = "Math G3 vs traveltime, studytime, freetime, goout", x = NULL,  y = "Final Grade") +
  theme_bw()
```



Visual Observation

studytime: Students who study more tend to have higher scores, especially those at level 3 or level 4. This trend is more clearly observed in G1 and G3.

traveltime: Travel time appears to be slightly related to grades.

freetime: Free time also seems to have a slight relationship with grades.

goout: Students who go out more tend to have lower grades. In particular, those who go out very frequently often score below 10.

```r
grade_group <- function(score) {
  case_when(
    score >= 16 ~ "A",
    score >= 14 ~ "B",
    score >= 12 ~ "C",
    score >= 10 ~ "D",
    TRUE        ~ "F"
  )
}

prepare_data <- function(file) {
  read_csv(file) %>%
    pivot_longer(cols = G1:G3, names_to = "Period", values_to = "Grade") %>%
    mutate(
      GradeGroup = grade_group(Grade),
      traveltime = as.character(traveltime),
      studytime = as.character(studytime),
      freetime = as.character(freetime),
      goout = as.character(goout)
    )
}

run_tests <- function(df, subject) {
  cat(paste0("======== ", subject, " ========\n"))
  for (v in c("traveltime", "studytime", "freetime", "goout")) {
    cat("\n==========", v, "==========\n")
    for (p in c("G1", "G2", "G3")) {
      temp <- df %>% filter(Period == p)
      tbl <- table(temp[[v]], temp$GradeGroup)
      test <- chisq.test(tbl)
      cat(v, "vs", p, ": p-value =", signif(test$p.value, 4), "\n")
    }
  }
}

port <- prepare_data("port_data_cleaned.csv")
math <- prepare_data("math_data_cleaned.csv")

run_tests(port, "Portuguese")
```

```
## ======== Portuguese ========
##
## ========== traveltime ==========

## traveltime vs G1 : p-value = 0.163

## traveltime vs G2 : p-value = 0.003895

## traveltime vs G3 : p-value = 0.01226
##
## ========== studytime ==========

## studytime vs G1 : p-value = 4.083e-08

## studytime vs G2 : p-value = 6.418e-06

## studytime vs G3 : p-value = 1.416e-06
##
## ========== freetime ==========

## freetime vs G1 : p-value = 0.04864

## freetime vs G2 : p-value = 0.09138
## freetime vs G3 : p-value = 0.007875
##
## ========== goout ==========
## goout vs G1 : p-value = 0.1777

## goout vs G2 : p-value = 0.04095
## goout vs G3 : p-value = 0.005948
```

```r
run_tests(math, "Math")
```

```
## ======== Math ========
##
## ========== traveltime ==========

## traveltime vs G1 : p-value = 0.1132

## traveltime vs G2 : p-value = 0.1916

## traveltime vs G3 : p-value = 0.6713
##
## ========== studytime ==========

## studytime vs G1 : p-value = 0.1214

## studytime vs G2 : p-value = 0.164

## studytime vs G3 : p-value = 0.1413
##
## ========== freetime ==========

## freetime vs G1 : p-value = 0.1173

## freetime vs G2 : p-value = 0.4534

## freetime vs G3 : p-value = 0.4283
##
## ========== goout ==========

## goout vs G1 : p-value = 0.2411

## goout vs G2 : p-value = 0.08992

## goout vs G3 : p-value = 0.02307
```

Statistical Test (Chi-Square Test) — Mathematics

studytime

- G1: p = 0.1214 > 0.05 -> fail to reject H0
- G2: p = 0.1640 > 0.05 -> fail to reject H0
- G3: p = 0.1413 > 0.05 -> fail to reject H0

traveltime

- G1: p = 0.1132 > 0.05 -> fail to reject H0
- G2: p = 0.1916 > 0.05 -> fail to reject H0
- G3: p = 0.6713 > 0.05 -> fail to reject H0

freetime - G1: p = 0.1173 > 0.05 -> fail to reject H0 - G2: p = 0.4534 > 0.05 -> fail to reject H0 - G3: p = 0.4283 > 0.05 -> fail to reject H0

goout - G1: p = 0.2411 > 0.05 -> fail to reject H0 - G2: p = 0.0899 > 0.05 -> fail to reject H0 - G3: p = 0.0231 < 0.05 -> reject H0

Statistical Test (Chi-Square Test) — Portuguese

studytime

- G1: p = 4.08e-08 < 0.05 -> reject H0
- G2: p = 6.42e-06 < 0.05 -> reject H0
- G3: p = 1.42e-06 < 0.05 -> reject H0

traveltime

- G1: p = 0.1630 > 0.05 -> fail to reject H0
- G2: p = 0.0039 < 0.05 -> reject H0
- G3: p = 0.0123 < 0.05 -> reject H0

freetime

- G1: p = 0.0486 < 0.05 -> reject H0
- G2: p = 0.0914 > 0.05 -> fail to reject H0
- G3: p = 0.0079 < 0.05 -> reject H0

goout

- G1: p = 0.1777 > 0.05 -> fail to reject H0
- G2: p = 0.0410 < 0.05 -> reject H0
- G3: p = 0.0059 < 0.05 -> reject H0

Answer

We can see that students with more study time tend to achieve higher scores, especially in Portuguese. Students who have longer travel times tend to receive lower scores over time. The grade distributions for G1, G2, and G3 are more scattered when it comes to free time and going out. According to the chi-square test, study time, travel time, free time, and going out are strongly related to grades in Portuguese across G1 to G3. In contrast, for math, only going out shows a strong influence on G3, while the other variables have only a slight effect on the course.

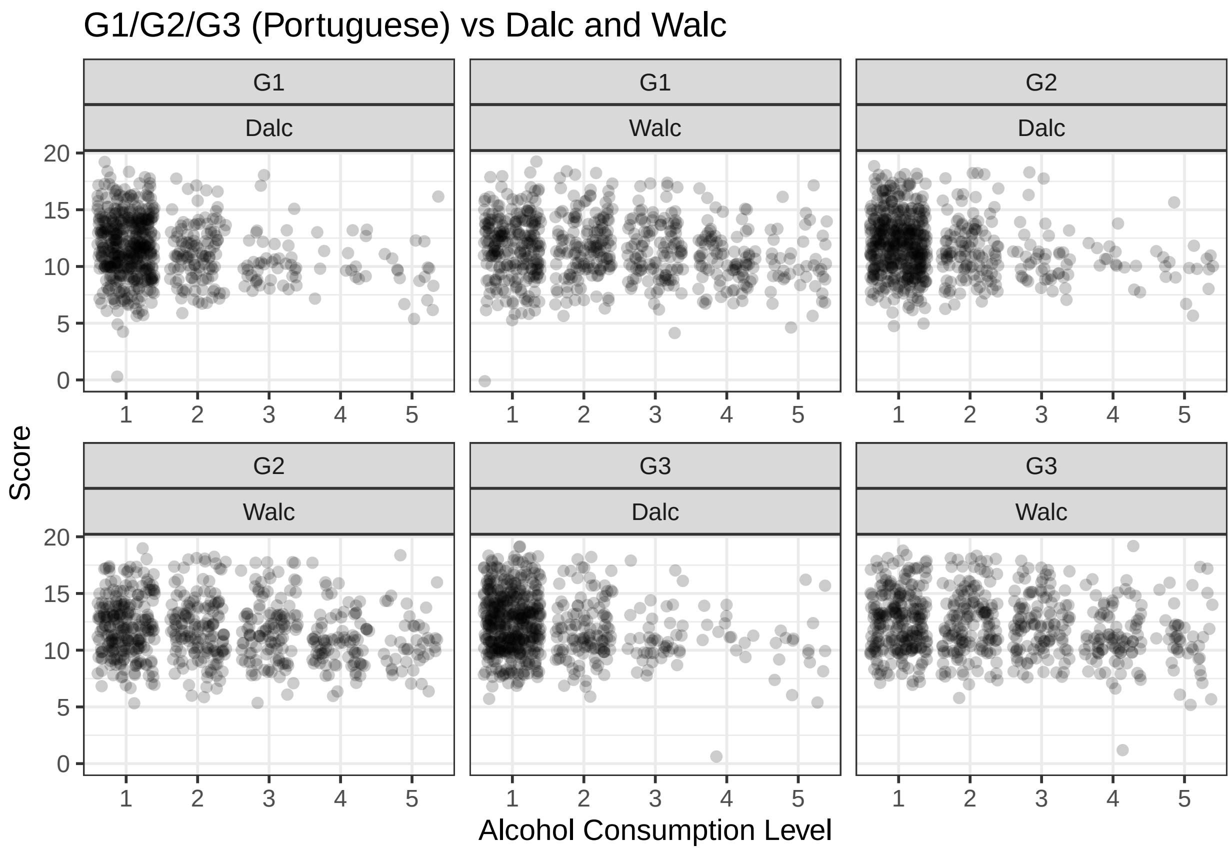
5.2 - Alcohol Consumption

Q1: How does alcohol affect student performance in Portuguese and Math courses?

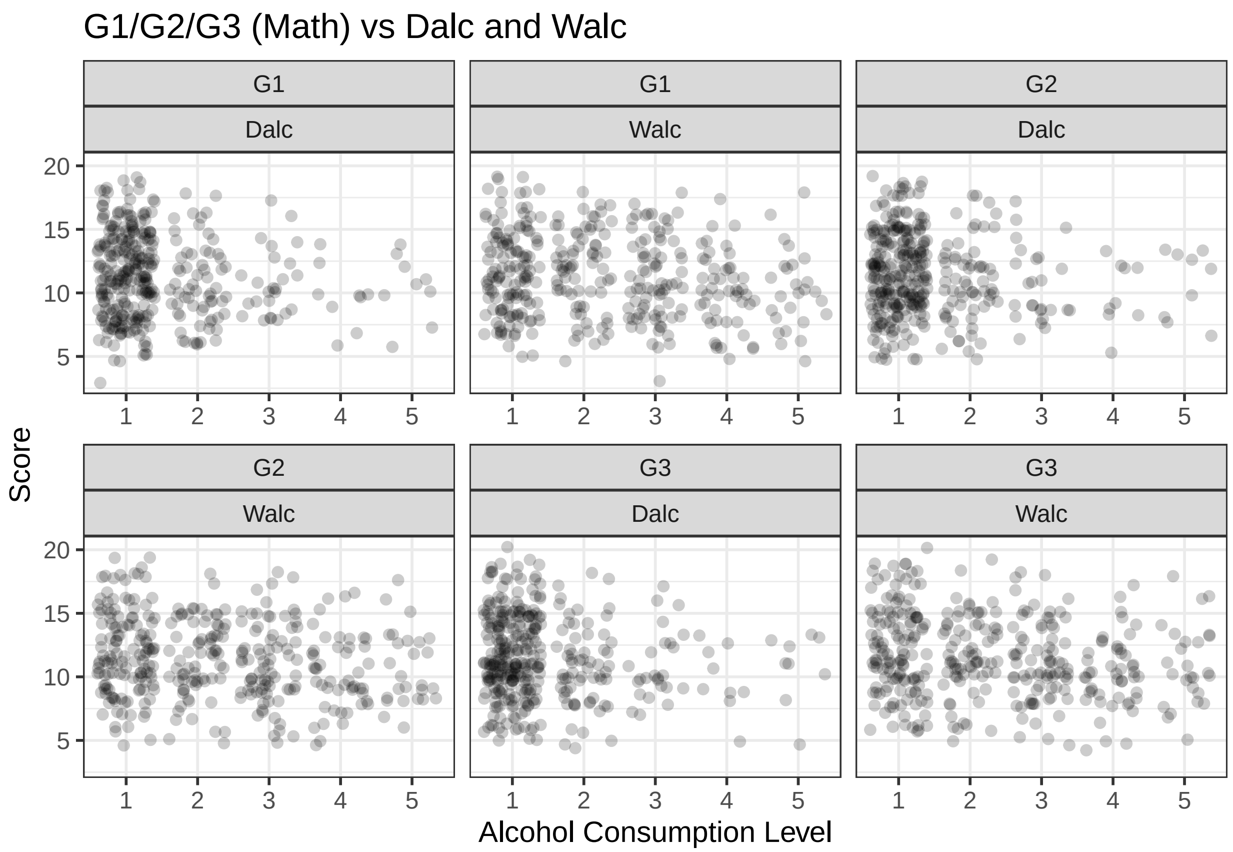
- Null Hypothesis: The proportion of letter grades does not differ by alcohol consumption level.

Alcohol Consumption Distribution

```
port <- read_csv("port_data_cleaned.csv") |>
  select(Dalc, Walc, G1, G2, G3) |>
  pivot_longer(cols = starts_with("G"), names_to = "Exam", values_to = "Score") |>
  pivot_longer(cols = c(Dalc, Walc), names_to = "AlcoholType", values_to = "Level") |>
  mutate(Level = as.factor(Level))

ggplot(port, aes(x = Level, y = Score)) +
  geom_jitter(alpha = 0.2, height = NULL) +
  facet_wrap(~ Exam + AlcoholType, scales = "free_x") +
  labs(title = "G1/G2/G3 (Portuguese) vs Dalc and Walc",
       x = "Alcohol Consumption Level", y = "Score") +
  theme_bw()
```



```
math <- read_csv("math_data_cleaned.csv") |>
  select(Dalc, Walc, G1, G2, G3) |>
  pivot_longer(cols = starts_with("G"), names_to = "Exam", values_to = "Score") |>
  pivot_longer(cols = c(Dalc, Walc), names_to = "AlcoholType", values_to = "Level") |>
  mutate(Level = as.factor(Level))

ggplot(math, aes(x = Level, y = Score)) +
  geom_jitter(alpha = 0.2, height = NULL) +
  facet_wrap(~ Exam + AlcoholType, scales = "free_x") +
  labs(title = "G1/G2/G3 (Math) vs Dalc and Walc",
       x = "Alcohol Consumption Level", y = "Score") +
  theme_bw()
```



Visual Observation

Higher alcohol levels (Dalc & Walc) are associated with more students scoring below 10 across all periods. Most high scores cluster at low alcohol levels (1 and 2), especially in G1 and G2. Both subjects show similar patterns — increased drinking links to lower scores.

```
grade_group <- function(score) {
  case_when(
    score >= 16 ~ "A",
    score >= 14 ~ "B",
    score >= 12 ~ "C",
    score >= 10 ~ "D",
    TRUE        ~ "F"
  )
}

prepare_data <- function(file) {
  read_csv(file) %>%
    pivot_longer(cols = G1:G3, names_to = "Period", values_to = "Grade") %>%
    mutate(
      GradeGroup = grade_group(Grade),
      Dalc = as.character(Dalc),
      Walc = as.character(Walc)
    )
}

run_tests <- function(df, subject, variables) {
  cat(paste0("======== ", subject, " ========\n"))
  for (v in variables) {
    cat("\n======", v, "======\n")
    for (p in c("G1", "G2", "G3")) {
      temp <- df %>% filter(Period == p)
      tbl <- table(temp[[v]], temp$GradeGroup)
      test <- chisq.test(tbl)
      cat(v, "->", p, ": p-value =", signif(test$p.value, 4), "\n")
    }
  }
}

math <- prepare_data("math_data_cleaned.csv")
port <- prepare_data("port_data_cleaned.csv")

run_tests(math, "Mathematics", c("Dalc", "Walc"))
```

```
## ======== Mathematics ========
##
## ====== Dalc ======

## Dalc -> G1 : p-value = 0.4231

## Dalc -> G2 : p-value = 0.02916

## Dalc -> G3 : p-value = 0.1031
##
## ====== Walc ======

## Walc -> G1 : p-value = 0.3674

## Walc -> G2 : p-value = 0.001082

## Walc -> G3 : p-value = 0.09464
```

```
run_tests(port, "Portuguese", c("Dalc", "Walc"))
```

```
## ======== Portuguese ========
##
## ====== Dalc ======

## Dalc -> G1 : p-value = 0.005585

## Dalc -> G2 : p-value = 0.0001362

## Dalc -> G3 : p-value = 0.001375
##
## ====== Walc ======

## Walc -> G1 : p-value = 0.01275
## Walc -> G2 : p-value = 0.006368
## Walc -> G3 : p-value = 0.006629
```

Statistical Test (Chi-Square Test) — Mathematics

Dalc

- G1: p = 0.4231 > 0.05 -> fail to reject H0
- G2: p = 0.0292 < 0.05 -> reject H0
- G3: p = 0.1031 > 0.05 -> fail to reject H0

Walc

- G1: p = 0.3674 > 0.05 -> fail to reject H0
- G2: p = 0.0011 < 0.05 -> reject H0
- G3: p = 0.0946 > 0.05 -> fail to reject H0

Statistical Test (Chi-Square Test) — Portuguese

Dalc

- G1: p = 0.0056 < 0.05 -> reject H0
- G2: p = 0.0001 < 0.05 -> reject H0
- G3: p = 0.0014 < 0.05 -> reject H0

Walc

- G1: p = 0.0128 < 0.05 -> reject H0
- G2: p = 0.0064 < 0.05 -> reject H0
- G3: p = 0.0066 < 0.05 -> reject H0

Answer

From the image, it can be seen that the frequency of drinking (especially on weekdays) has a significantly stronger impact on Portuguese language performance than on mathematics performance, and this impact remains significant throughout all stages of learning. In contrast, the impact of drinking on mathematics performance is weaker and less stable.

Drinking is related to both courses. In math, drinking is only related to G2, and it is slightly related to G1 and G3. In Portuguese, drinking is related to G1, G2, and G3 grades, especially on weekdays. Overall, these results show that drinking has a negative relationship with grades.

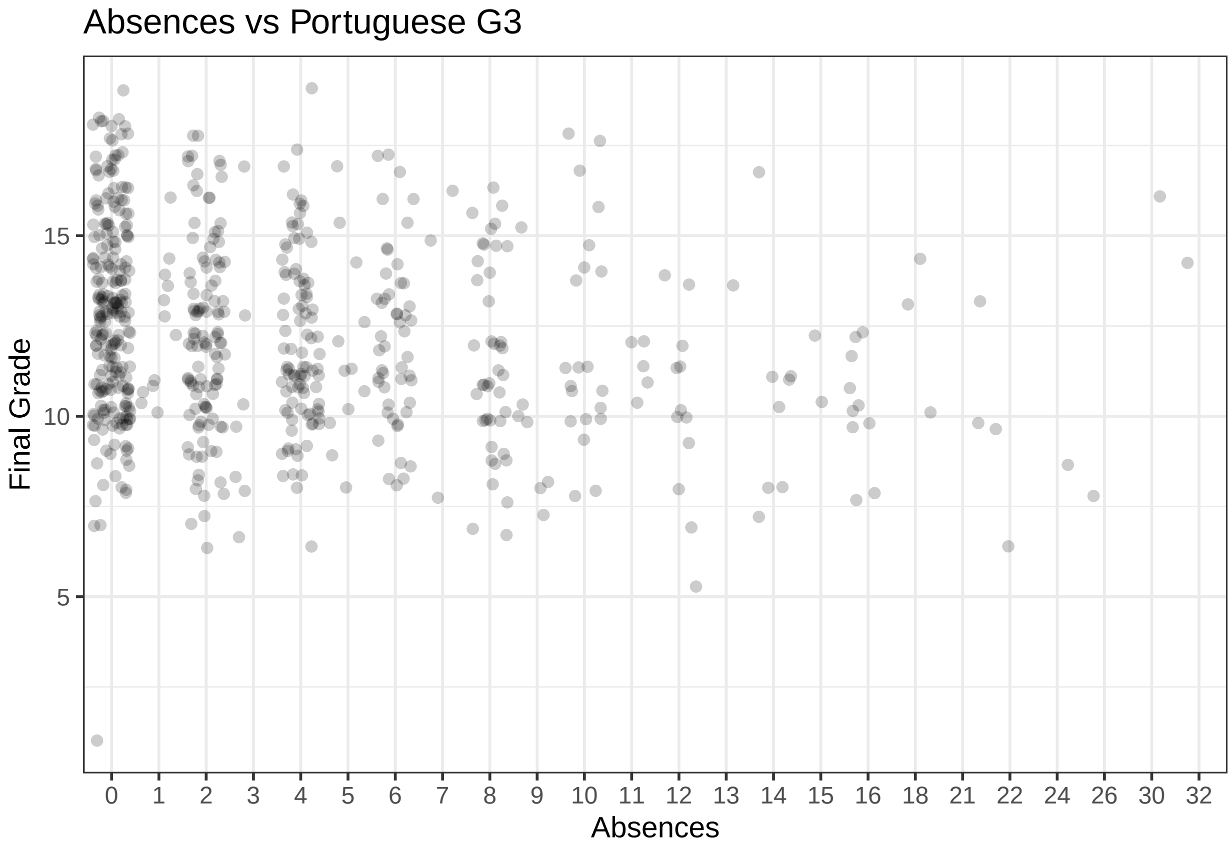
5.3 - Absences

Q1: How do absences affect student performance in Portuguese and Math courses?

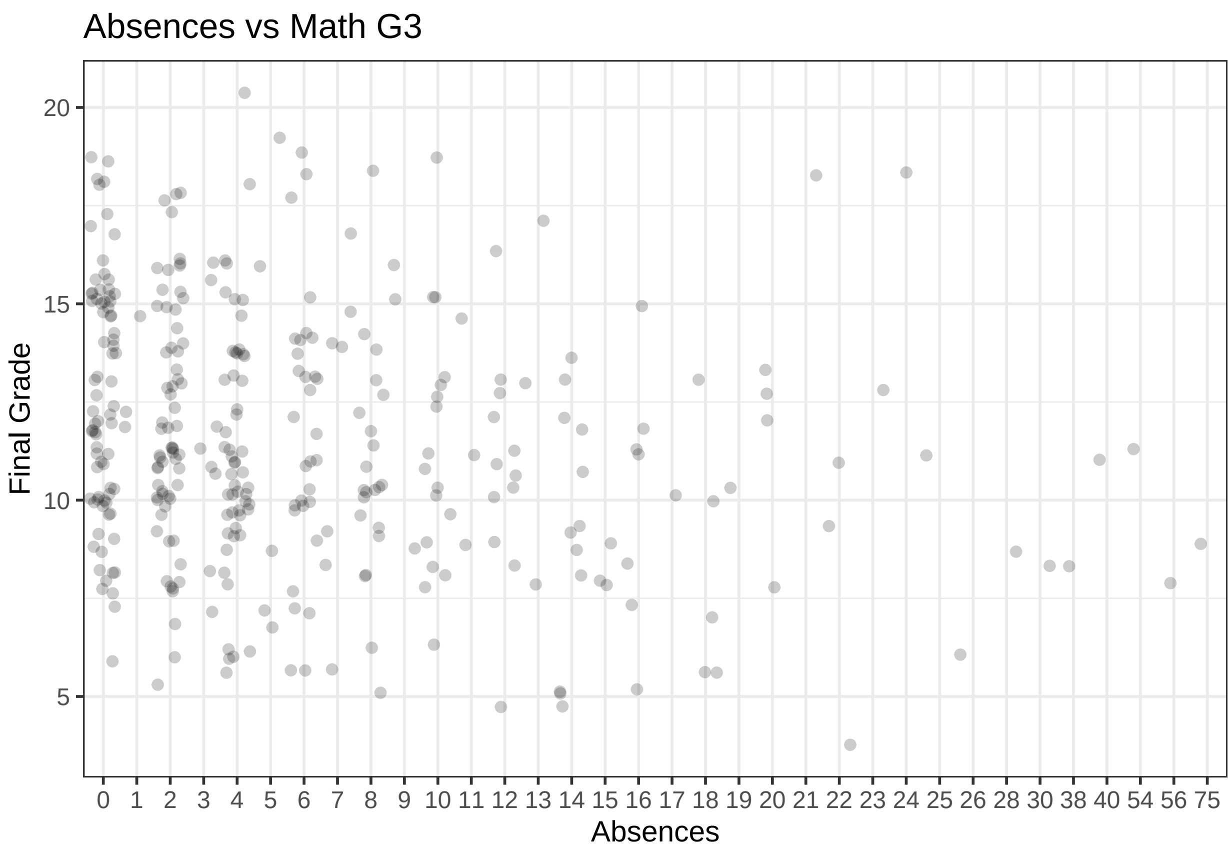
- Null Hypothesis: The proportion of letter grades does not differ by absences.

Absences Distribution

```
port <- read_csv("port_data_cleaned.csv")

ggplot(port, aes(x = as.factor(absences), y = G3)) +
  geom_jitter(alpha = 0.2, height = NULL) +
  labs(title = "Absences vs Portuguese G3",
       x = "Absences", y = "Final Grade") +
  theme_bw()
```



```
math <- read_csv("math_data_cleaned.csv")

ggplot(math, aes(x = as.factor(absences), y = G3)) +
  geom_jitter(alpha = 0.2, height = NULL) +
  labs(title = "Absences vs Math G3",
       x = "Absences", y = "Final Grade") +
  theme_bw()
```



Visual Observation

In Portuguese classes, students who miss more lessons generally receive lower final grades, indicating a clear negative impact of absences.

In Math classes, absences also appear to influence grades, but the relationship is weaker and less consistent among students.

```
predata <- function(file, subject) {
  read_csv(file) %>%
    pivot_longer(cols = G1:G3, names_to = "Period", values_to = "Grade") %>%
    mutate(
      Subject = subject,
      GradeGroup = case_when(
        Grade >= 16 ~ "A",
        Grade >= 14 ~ "B",
        Grade >= 12 ~ "C",
        Grade >= 10 ~ "D",
        TRUE        ~ "F"
      ),
      absences = case_when(
        absences == 0 ~ "None",
        absences <= 3 ~ "Low",
        absences <= 10 ~ "Medium",
        absences > 10 ~ "High"
      )
    )
```

```r
}

run_chisq <- function(df, subject) {
  cat("\n========", subject, "========\n")
  cat("\n====== Absences ======\n")
  for (p in c("G1", "G2", "G3")) {
    temp <- df %>% filter(Period == p)
    cat("Absences ->", p, ":\n")
    print(chisq.test(table(temp$absences, temp$GradeGroup)))
  }
}

math <- predata("math_data_cleaned.csv", "Math")
port <- predata("port_data_cleaned.csv", "Portuguese")

run_chisq(math, "Mathematics")
```

```
##
## ======== Mathematics ========
##
## ====== Absences ======
## Absences -> G1 :
##
##  Pearson's Chi-squared test
##
## data:  table(temp$absences, temp$GradeGroup)
## X-squared = 20.507, df = 12, p-value = 0.05808
##
## Absences -> G2 :
##
##  Pearson's Chi-squared test
##
## data:  table(temp$absences, temp$GradeGroup)
## X-squared = 24.673, df = 12, p-value = 0.01645
##
## Absences -> G3 :
##
##  Pearson's Chi-squared test
##
## data:  table(temp$absences, temp$GradeGroup)
## X-squared = 32.421, df = 12, p-value = 0.001191
```

```r
run_chisq(port, "Portuguese")
```

```
##
## ======== Portuguese ========
##
## ====== Absences ======
## Absences -> G1 :
```

```
##
##  Pearson's Chi-squared test
```

```
##
## data:  table(temp$absences, temp$GradeGroup)
## X-squared = 37.71, df = 12, p-value = 0.0001713
##
## Absences -> G2 :
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(temp$absences, temp$GradeGroup)
## X-squared = 43.216, df = 12, p-value = 2.075e-05
##
## Absences -> G3 :
##
##  Pearson's Chi-squared test
##
## data:  table(temp$absences, temp$GradeGroup)
## X-squared = 36.056, df = 12, p-value = 0.0003174
```

Statistical Test (Chi-Square Test) — Mathematics

Absences

- G1: p = 0.0581 > 0.05 -> fail to reject H0
- G2: p = 0.0165 < 0.05 -> reject H0
- G3: p = 0.0012 < 0.05 -> reject H0

Statistical Test (Chi-Square Test) — Portuguese

Absences

- G1: p = 0.0002 < 0.05 -> reject H0
- G2: p = 0.000021 < 0.05 -> reject H0
- G3: p = 0.0003 < 0.05 -> reject H0

Answer

From the plot, we can see that in Portuguese, students with more absences tend to have lower scores. In Math, although a similar trend exists, the distribution is not as strongly related as in Portuguese.

In the chi-square test, G1, G2, and G3 grades in Portuguese are associated with absences. However, in Math, absences are only significantly related to G2 and G3.

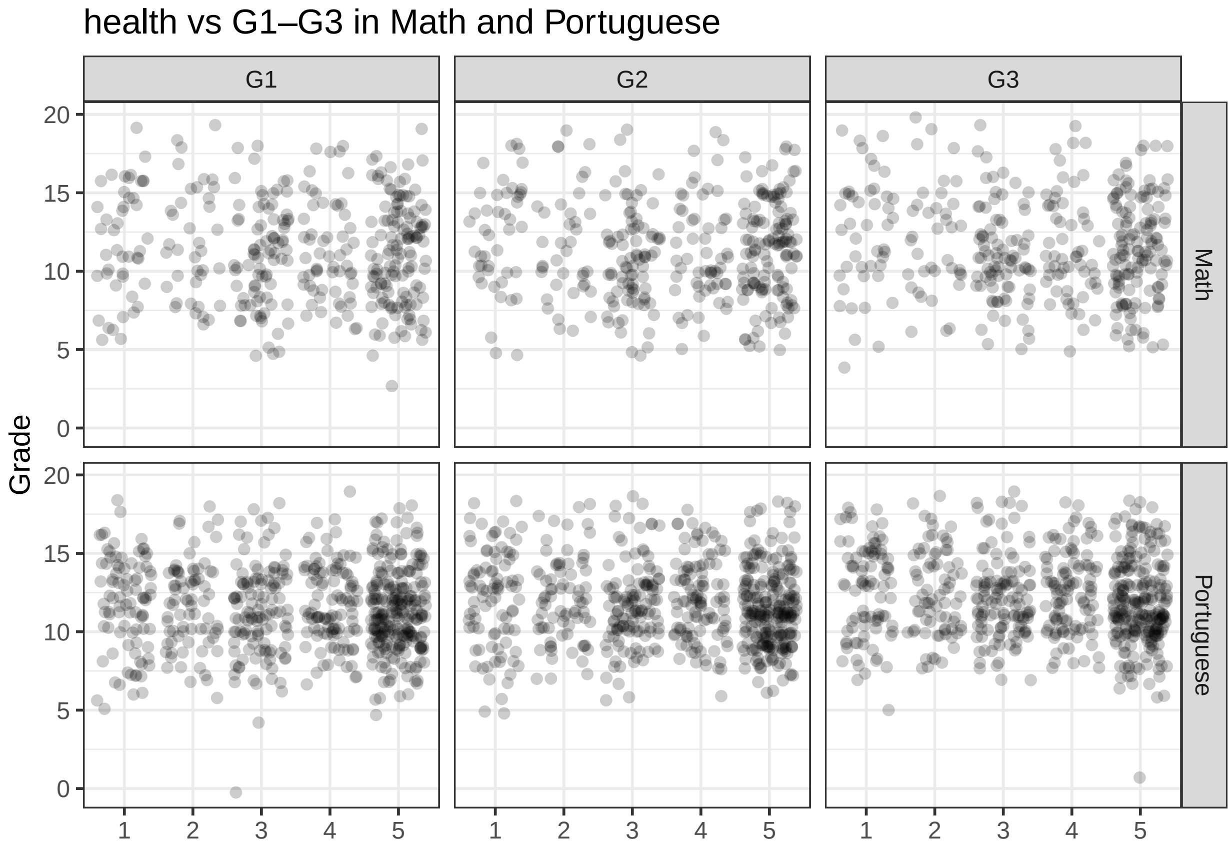
5.4 - Health

Q1: How do health affect student performance in Portuguese and Math courses?

- Null Hypothesis: The proportion of letter grades does not differ by health.

Health Distribution

```r
g1 <- read_csv("port_data_cleaned.csv") |>
  select(health, G1) |>
  mutate(health = as.character(health)) |>
  pivot_longer(-G1, names_to = "Variable", values_to = "Value") |>
  mutate(course = "Portuguese", Period = "G1", Grade = G1)

g2 <- read_csv("port_data_cleaned.csv") |>
  select(health, G2) |>
  mutate(health = as.character(health)) |>
  pivot_longer(-G2, names_to = "Variable", values_to = "Value") |>
  mutate(course = "Portuguese", Period = "G2", Grade = G2)

g3 <- read_csv("port_data_cleaned.csv") |>
  select(health, G3) |>
  mutate(health = as.character(health)) |>
  pivot_longer(-G3, names_to = "Variable", values_to = "Value") |>
  mutate(course = "Portuguese", Period = "G3", Grade = G3)

g1math <- read_csv("math_data_cleaned.csv") |>
  select(health, G1) |>
  mutate(health = as.character(health)) |>
  pivot_longer(-G1, names_to = "Variable", values_to = "Value") |>
  mutate(course = "Math", Period = "G1", Grade = G1)

g2math <- read_csv("math_data_cleaned.csv") |>
  select(health, G2) |>
  mutate(health = as.character(health)) |>
  pivot_longer(-G2, names_to = "Variable", values_to = "Value") |>
  mutate(course = "Math", Period = "G2", Grade = G2)

g3math <- read_csv("math_data_cleaned.csv") |>
  select(health, G3) |>
  mutate(health = as.character(health)) |>
  pivot_longer(-G3, names_to = "Variable", values_to = "Value") |>
  mutate(course = "Math", Period = "G3", Grade = G3)

all_data <- bind_rows(g1, g2, g3, g1math, g2math, g3math)

ggplot(all_data, aes(x = as.factor(Value), y = Grade)) +
  geom_jitter(alpha = 0.2, height = NULL) +
  facet_grid(course ~ Period) +
  labs(
    title = "health vs G1-G3 in Math and Portuguese",
    x = NULL,
    y = "Grade"
  ) +
  theme_bw()
```



Visual Observation

In Portuguese, students with lower health scores tend to have lower grade distributions, while those with higher health ratings generally achieve better scores. This pattern is particularly noticeable in G3. In contrast, math grades appear to be more evenly distributed across all health levels, suggesting that students' performance in math is less affected by their health condition.

```
grade_group <- function(score) {
  case_when(
    score >= 16 ~ "A",
    score >= 14 ~ "B",
    score >= 12 ~ "C",
    score >= 10 ~ "D",
    TRUE        ~ "F"
  )
}

prepare_health_data <- function(file) {
  read_csv(file) %>%
    mutate(
      GradeGroup_G1 = grade_group(G1),
      GradeGroup_G2 = grade_group(G2),
      GradeGroup_G3 = grade_group(G3),
      health = as.character(health)
    )
}
```

```r
run_health_tests <- function(df, subject) {
  cat(paste0("======== ", subject, " ========\n"))
  for (g in c("G1", "G2", "G3")) {
    grp <- paste0("GradeGroup_", g)
    cat("Health vs", g, ":\n")
    print(chisq.test(table(df$health, df[[grp]])))
    cat("\n")
  }
}

port <- prepare_health_data("port_data_cleaned.csv")
math <- prepare_health_data("math_data_cleaned.csv")

run_health_tests(port, "Portuguese")
```

```
## ======== Portuguese ========
## Health vs G1 :
##
##  Pearson's Chi-squared test
##
## data:  table(df$health, df[[grp]])
## X-squared = 17.627, df = 16, p-value = 0.3462
##
##
## Health vs G2 :
##
##  Pearson's Chi-squared test
##
## data:  table(df$health, df[[grp]])
## X-squared = 24.642, df = 16, p-value = 0.07639
##
##
## Health vs G3 :
##
##  Pearson's Chi-squared test
##
## data:  table(df$health, df[[grp]])
## X-squared = 29.765, df = 16, p-value = 0.01926
```

```r
run_health_tests(math, "Mathematics")
```

```
## ======== Mathematics ========
## Health vs G1 :

##
##  Pearson's Chi-squared test
##
## data:  table(df$health, df[[grp]])
## X-squared = 15.459, df = 16, p-value = 0.4913
##
##
## Health vs G2 :
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$health, df[[grp]])
## X-squared = 21.059, df = 16, p-value = 0.1762
##
##
## Health vs G3 :
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$health, df[[grp]])
## X-squared = 13.314, df = 16, p-value = 0.6497
```

Statistical Test (Chi-Square Test) – Mathematics

Health:

- G1: p = 0.4913 > 0.05 -> fail to reject H0
- G2: p = 0.1762 > 0.05 -> fail to reject H0
- G3: p = 0.6497 > 0.05 -> fail to reject H0

Statistical Test (Chi-Square Test) – Portuguese

Health:

- G1: p = 0.3462 > 0.05 -> fail to reject H0
- G2: p = 0.0764 > 0.05 -> fail to reject H0
- G3: p = 0.0193 < 0.05 -> reject H0

Answer

Students in better health tend to perform better, especially in Portuguese G3, where those in poorer health are more concentrated in the lower score range. This trend is not as clear in math, where grade distribution is more balanced across health levels. Overall, better health is associated with stronger performance in Portuguese but not clearly in math.

According to the Chi-Square test, G1 to G3 in math show weak or no significant relationship with health. However, in Portuguese, G3 shows a strong and statistically significant connection. This indicates that health has more influence on Portuguese performance than on math.

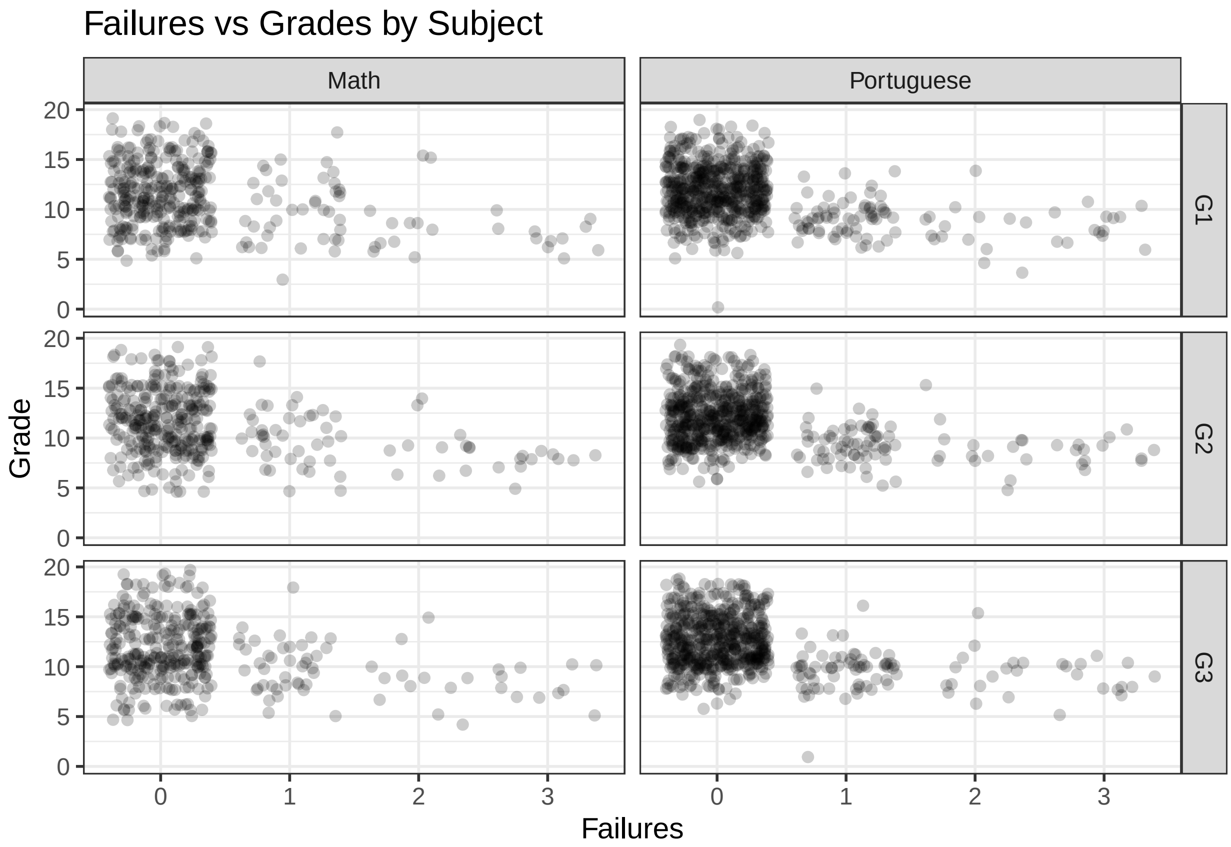
5.5 - failure

Q1: How do failure affect student performance in Portuguese and Math courses?

- Null Hypothesis: The proportion of letter grades does not differ by failures.

Failure Distribution

```r
port<- read_csv("port_data_cleaned.csv") |>
  select(failures, G1, G2, G3) |>
  pivot_longer(cols = G1:G3, names_to = "Period", values_to = "Grade") |>
  mutate(failures = as.character(failures), Subject = "Portuguese")

math<- read_csv("math_data_cleaned.csv") |>
  select(failures, G1, G2, G3) |>
  pivot_longer(cols = G1:G3, names_to = "Period", values_to = "Grade") |>
  mutate(failures = as.character(failures), Subject = "Math")

alldata <- bind_rows(port, math)

ggplot(alldata, aes(x = failures, y = Grade)) +
  geom_jitter(alpha = 0.2, height = NULL) +
  facet_grid(Period ~ Subject, scales = "free_x") +
  labs(title = "Failures vs Grades by Subject", x = "Failures", y = "Grade") +
  theme_bw()
```



Visual Observation

In both Math and Portuguese, as the number of failures increases from 0 to 3, grades from G1 to G3 are increasingly concentrated in the lower range.

```r
grade_group <- function(score) {
  case_when(
    score >= 16 ~ "A",
    score >= 14 ~ "B",
    score >= 12 ~ "C",
    score >= 10 ~ "D",
    TRUE        ~ "F"
  )
}

prepare_data <- function(file) {
  read_csv(file) %>%
    mutate(
      GradeGroup_G1 = grade_group(G1),
      GradeGroup_G2 = grade_group(G2),
      GradeGroup_G3 = grade_group(G3),
      failures = as.character(failures)
    )
}

run_failures_test <- function(df, subject) {
  cat(paste0("======== ", subject, " ========\n"))
  for (g in c("G1", "G2", "G3")) {
    grp <- paste0("GradeGroup_", g)
    cat("failures vs", g, ":\n")
    print(chisq.test(table(df$failures, df[[grp]])))
    cat("\n")
  }
}

df_port <- prepare_data("port_data_cleaned.csv")
df_math <- prepare_data("math_data_cleaned.csv")

run_failures_test(df_port, "Portuguese")
```

```
## ======== Portuguese ========
## failures vs G1 :

##
##  Pearson’s Chi-squared test
##
## data:  table(df$failures, df[[grp]])
## X-squared = 141.56, df = 12, p-value < 2.2e-16
##
##
## failures vs G2 :

##
##  Pearson’s Chi-squared test
##
## data:  table(df$failures, df[[grp]])
## X-squared = 132.4, df = 12, p-value < 2.2e-16
##
##
## failures vs G3 :
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$failures, df[[grp]])
## X-squared = 132.25, df = 12, p-value < 2.2e-16
```

```
run_failures_test(df_math, "Mathematics")
```

```
## ======== Mathematics ========
## failures vs G1 :

##
##  Pearson's Chi-squared test
##
## data:  table(df$failures, df[[grp]])
## X-squared = 40.329, df = 12, p-value = 6.341e-05
##
##
## failures vs G2 :

##
##  Pearson's Chi-squared test
##
## data:  table(df$failures, df[[grp]])
## X-squared = 49.983, df = 12, p-value = 1.406e-06
##
##
## failures vs G3 :

##
##  Pearson's Chi-squared test
##
## data:  table(df$failures, df[[grp]])
## X-squared = 45.859, df = 12, p-value = 7.337e-06
```

Statistical Test (Chi-Square Test) – Mathematics

Failures

- G1: p = 6.34e-05 < 0.05 -> reject H0
- G2: p = 1.41e-06 < 0.05 -> reject H0
- G3: p = 7.34e-06 < 0.05 -> reject H0

Statistical Test (Chi-Square Test) – Portuguese

Failures

- G1: p < 2.2e-16 < 0.05 -> reject H0
- G2: p < 2.2e-16 < 0.05 -> reject H0
- G3: p < 2.2e-16 < 0.05 -> reject H0

Answer

More failures students have, the lower their scores tend to be from G1 to G3 in both Math and Portuguese.

According to the chi-square test, the number of failures has a strong correlation with G1, G2, and G3 in both subjects. For Portuguese, the p-values are extremely small, indicating a very strong relationship with failures. Math also shows a strong correlation, but not as strong as Portuguese

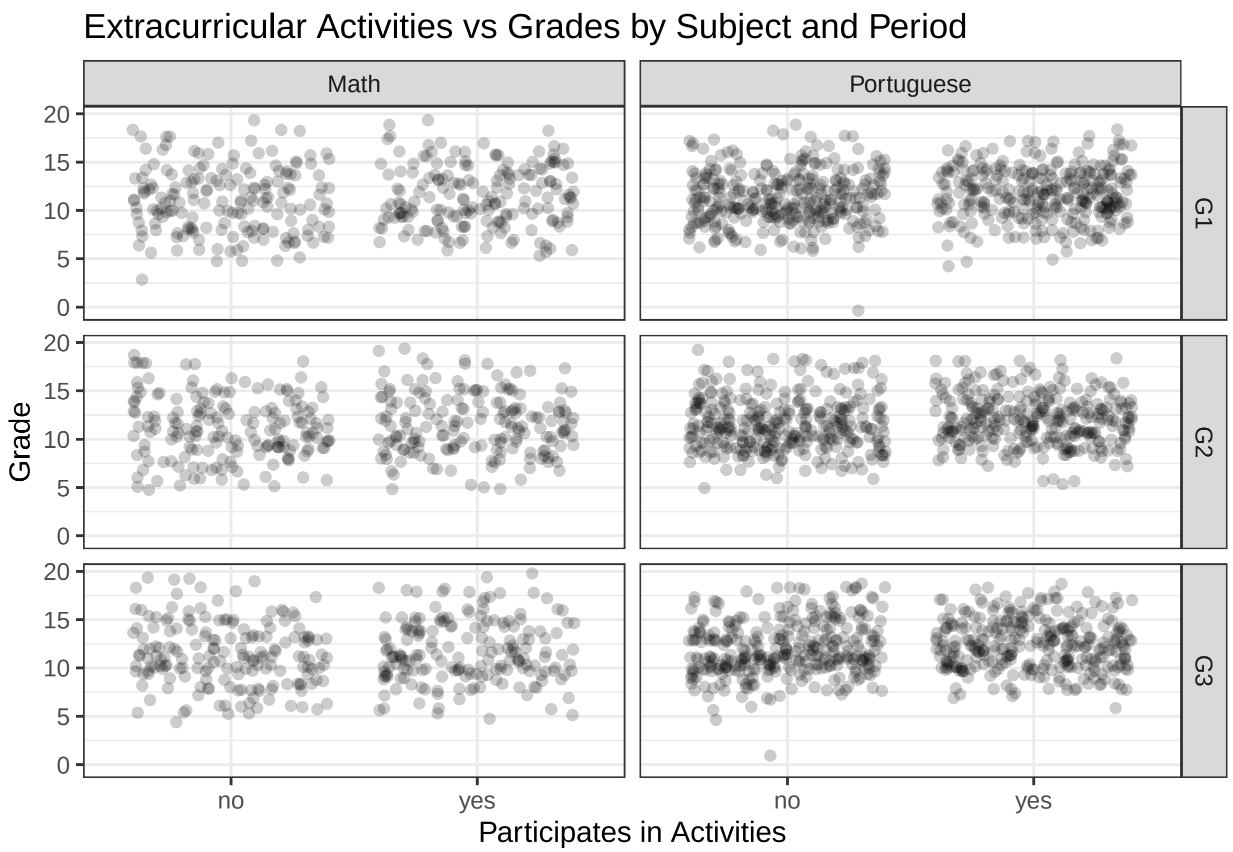
5.6 - activities

Q1: How do activities affect student performance in Portuguese and Math courses?

- Null Hypothesis: The proportion of letter grades does not differ by activities.

Activities Distribution

```r
port_g1 <- read_csv("port_data_cleaned.csv") |>
  select(activities, G1) |>
  rename(Grade = G1) |>
  mutate(Subject = "Portuguese", Period = "G1")

math_g1 <- read_csv("math_data_cleaned.csv") |>
  select(activities, G1) |>
  rename(Grade = G1) |>
  mutate(Subject = "Math", Period = "G1")

port_g2 <- read_csv("port_data_cleaned.csv") |>
  select(activities, G2) |>
  rename(Grade = G2) |>
  mutate(Subject = "Portuguese", Period = "G2")

math_g2 <- read_csv("math_data_cleaned.csv") |>
  select(activities, G2) |>
  rename(Grade = G2) |>
  mutate(Subject = "Math", Period = "G2")

port_g3 <- read_csv("port_data_cleaned.csv") |>
  select(activities, G3) |>
  rename(Grade = G3) |>
  mutate(Subject = "Portuguese", Period = "G3")

math_g3 <- read_csv("math_data_cleaned.csv") |>
  select(activities, G3) |>
  rename(Grade = G3) |>
  mutate(Subject = "Math", Period = "G3")

df <- bind_rows(port_g1, math_g1, port_g2, math_g2, port_g3, math_g3)

ggplot(df, aes(x = activities, y = Grade)) +
  geom_jitter(alpha = 0.2, height = NULL) +
  facet_grid(Period ~ Subject) +
  labs(
    title = "Extracurricular Activities vs Grades by Subject and Period",
    x = "Participates in Activities",
    y = "Grade"
  ) +
  theme_bw()
```



Visual Observation

For Portuguese, students who participate in activities tend to have higher G3 scores than those who do not. For Math, participation in activities does not appear to influence grades from G1 to G3.

```
grade_group <- function(score) {
  case_when(
    score >= 16 ~ "A",
    score >= 14 ~ "B",
    score >= 12 ~ "C",
    score >= 10 ~ "D",
    TRUE        ~ "F"
  )
}

prepare_activities_data <- function(file) {
  read_csv(file) %>%
    mutate(
      GradeGroup_G1 = grade_group(G1),
      GradeGroup_G2 = grade_group(G2),
      GradeGroup_G3 = grade_group(G3),
      activities = as.character(activities)
    )
}

run_activities_tests <- function(df, subject) {
  cat(paste0("======== ", subject, " ========\n"))
  for (g in c("G1", "G2", "G3")) {
    grp <- paste0("GradeGroup_", g)
    cat("activities vs", g, ":\n")
    print(chisq.test(table(df$activities, df[[grp]])))
    cat("\n")
  }
}

port <- prepare_activities_data("port_data_cleaned.csv")
math <- prepare_activities_data("math_data_cleaned.csv")

run_activities_tests(port, "Portuguese")
```

```
## ======== Portuguese ========
## activities vs G1 :
##
##  Pearson's Chi-squared test
##
## data:  table(df$activities, df[[grp]])
## X-squared = 8.1604, df = 4, p-value = 0.08588
##
##
## activities vs G2 :
##
##  Pearson's Chi-squared test
##
## data:  table(df$activities, df[[grp]])
## X-squared = 9.0401, df = 4, p-value = 0.0601
##
##
## activities vs G3 :
##
##  Pearson's Chi-squared test
##
## data:  table(df$activities, df[[grp]])
## X-squared = 10.875, df = 4, p-value = 0.028
```

```
run_activities_tests(math, "Mathematics")
```

```
## ======== Mathematics ========
## activities vs G1 :
##
##  Pearson's Chi-squared test
##
## data:  table(df$activities, df[[grp]])
## X-squared = 1.6863, df = 4, p-value = 0.7932
##
##
## activities vs G2 :
##
##  Pearson's Chi-squared test
```

```
##
## data:  table(df$activities, df[[grp]])
## X-squared = 1.382, df = 4, p-value = 0.8473
##
##
## activities vs G3 :
##
##  Pearson's Chi-squared test
##
## data:  table(df$activities, df[[grp]])
## X-squared = 2.5059, df = 4, p-value = 0.6436
```

Statistical Test (Chi-Square Test) Portuguese

- G1: p = 0.0859 > 0.05 -> fail to reject H0
- G2: p = 0.0601 > 0.05 -> fail to reject H0
- G3: p = 0.0280 < 0.05 -> reject H0

Statistical Test (Chi-Square Test) Mathematics

- G1: p = 0.7932 > 0.05 -> fail to reject H0
- G2: p = 0.8473 > 0.05 -> fail to reject H0
- G3: p = 0.6436 > 0.05 -> fail to reject H0

Answer

From the graph, we can see that students who participate in activities tend to have higher scores in Portuguese compared to those who don't. For Math, participation in activities doesn't appear to significantly affect grades.

According to the Chi-square test, the p-value for Portuguese G3 shows a strong relationship with activity participation. However, for Math, as well as G1 and G2 in Portuguese, there is no significant relationship with activities.

5.7 - Does behavioral factors affect grades different?

We are interested in how factors related to school performance and behavior impact grades in Math and Portuguese differently. We will explore the following factors:

- absences
- study time
- number of failures in the past
- going out with friends
- extracurricular activities
- romantic relationships
- alcohol consumption

Class Attendence and Absences

- Null hypothesis: There is no linear correlation between the number of absences and grade.

```
longer_data %>%
  group_by(subject, Period) %>%
  summarise(
    cor_test = list(cor.test(absences, Grade, method = "pearson", use = "complete.obs")),
    .groups = "drop"
  ) %>%
  mutate(
    correlation = map_dbl(cor_test, ~ .x$estimate),
    p_value = map_dbl(cor_test, ~ .x$p.value)
  ) %>%
  select(subject, Period, correlation, p_value)
```

```
## # A tibble: 6 x 4
##   subject    Period correlation     p_value
##   <chr>      <chr>        <dbl>       <dbl>
## 1 Math       G1          -0.120 0.0230
## 2 Math       G2          -0.200 0.000147
## 3 Math       G3          -0.213 0.0000492
## 4 Portuguese G1          -0.184 0.00000297
## 5 Portuguese G2          -0.197 0.000000556
## 6 Portuguese G3          -0.196 0.000000654
```

For math:

- p-value < 0.05 for all periods. We reject the null hypothesis.
- Correlation increases from little or no correlation for G1 to weak negative correlation for G2 and G3.

For Portuguese:

- p-value < 0.05 for all periods. We reject the null hypothesis.
- There is a weak negative correlation for all periods.

There is a growing negative correlation between absences and math grade, while the correlation between absences and Portuguese grade does not strengthen over time. This suggests that cumulative content and skill building in Math is more sensitive to missed instruction. In contrast, for Portuguese, language-based skills may be less disrupted by occasional absences or are more recoverable.

Study Time

- Null hypothesis: There is no correlation between study time and grade in the two subjects.

```
longer_data %>%
  group_by(subject, Period) %>%
  summarise(
    cor_test = list(cor.test(studytime, Grade, method = "pearson", use = "complete.obs")),
    .groups = "drop"
  ) %>%
  mutate(
    correlation = map_dbl(cor_test, ~ .x$estimate),
    p_value = map_dbl(cor_test, ~ .x$p.value)
  ) %>%
  select(subject, Period, correlation, p_value)
```

```
## # A tibble: 6 x 4
##   subject    Period correlation  p_value
##   <chr>      <chr>        <dbl>    <dbl>
## 1 Math       G1           0.141 7.79e- 3
## 2 Math       G2           0.120 2.36e- 2
## 3 Math       G3           0.127 1.66e- 2
## 4 Portuguese G1           0.249 2.13e-10
## 5 Portuguese G2           0.237 1.60e- 9
## 6 Portuguese G3           0.246 3.44e-10
```

For math:

- p-value < 0.05 for all periods. We reject the null hypothesis.
- There is a very small positive correlation between study time and Math grade for all periods, with cor < 0.15.

For Portuguese:

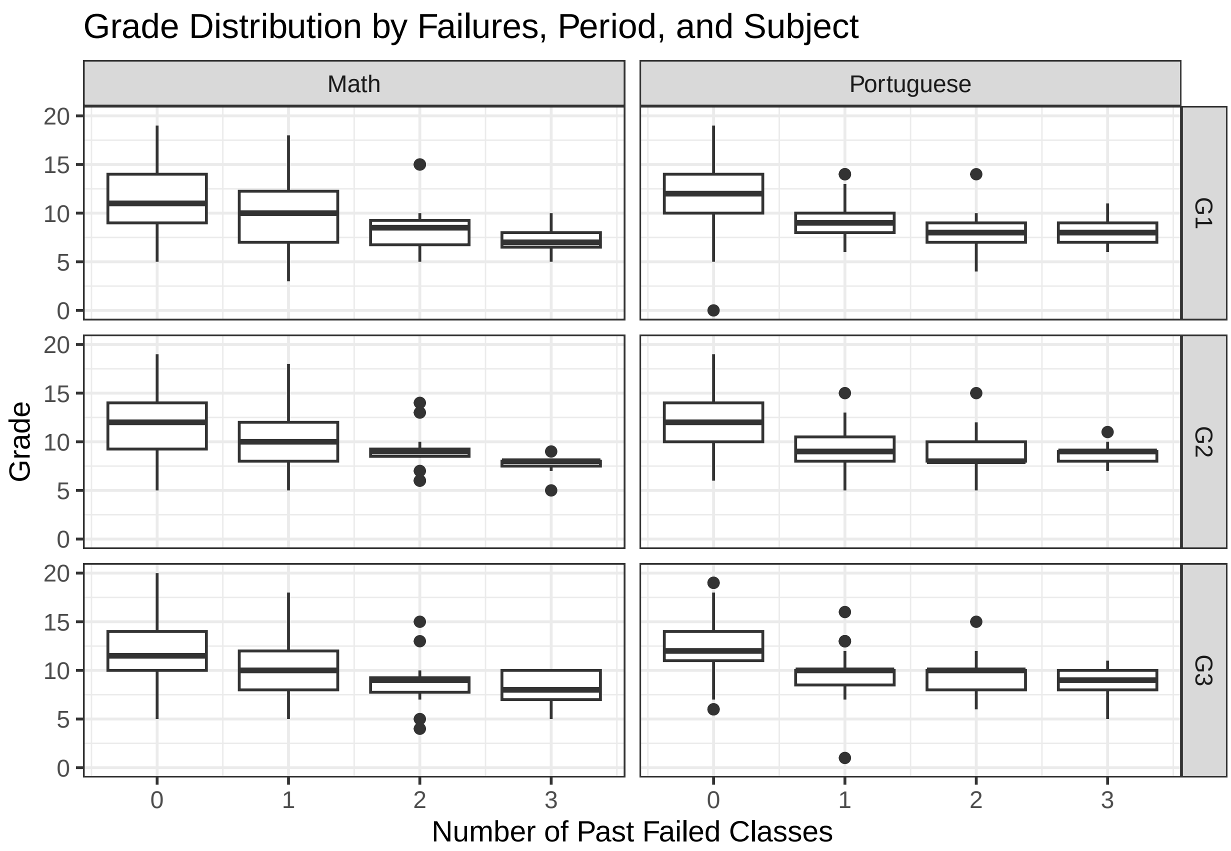
- p-value < 0.05 for all periods. We reject the null hypothesis.
- There is a weak to moderate positive correlation between study time and Portuguese grade for all periods, with cor around 0.24.

Study time positively correlates to grade stronger in Portuguese than in Math. This can possibly be explained by the nature and learning pattern of the subjects. Language classes relies more heavily on reading and writing, which benefit from sufficient time commitment, while math requires more intuition and conceptual understanding and thus does not depend as much on pure time commitment.

Past Failures

- Null hypothesis: There is no significant correlation between past failures and grade.

```r
ggplot(longer_data, aes(x = failures, y=Grade, group=failures)) +
  geom_boxplot() +
  facet_grid(Period ~ subject) +
  labs(title = "Grade Distribution by Failures, Period, and Subject",
       x = "Number of Past Failed Classes",
       y = "Grade")+
  theme_bw()
```



```
longer_data %>%
  group_by(subject, Period) %>%
  summarise(
    cor_test = list(cor.test(failures, Grade, method = "pearson", use = "complete.obs")),
    .groups = "drop"
  ) %>%
  mutate(
    correlation = map_dbl(cor_test, ~ .x$estimate),
    p_value = map_dbl(cor_test, ~ .x$p.value)
  ) %>%
  select(subject, Period, correlation, p_value)
```

```
## # A tibble: 6 x 4
##   subject    Period correlation  p_value
##   <chr>      <chr>        <dbl>    <dbl>
## 1 Math       G1          -0.302 5.74e- 9
## 2 Math       G2          -0.301 6.29e- 9
## 3 Math       G3          -0.294 1.53e- 8
## 4 Portuguese G1          -0.380 3.76e-23
## 5 Portuguese G2          -0.370 5.60e-22
## 6 Portuguese G3          -0.388 3.59e-24
```

For math:

- p-value < 0.05 for all periods. We reject the null hypothesis.

- There is a moderate negative correlation between failures and grade for all periods, with cor around 0.3.

For Portuguese:

- p-value < 0.05 for all periods. We reject the null hypothesis.
- There is a moderate negative correlation between failures and grade for all periods, with cor around 0.37.

There is a statistically significant moderate negative correlation between number of past failures and grade. The effect is slightly stronger in Portuguese than in Math.

Going out

- Null hypothesis: There is no significant correlation between socialization and grade.

```
longer_data %>%
  group_by(subject, Period) %>%
  summarise(
    cor_test = list(cor.test(goout, Grade, method = "pearson", use = "complete.obs")),
    .groups = "drop"
  ) %>%
  mutate(
    correlation = map_dbl(cor_test, ~ .x$estimate),
    p_value = map_dbl(cor_test, ~ .x$p.value)
  ) %>%
  select(subject, Period, correlation, p_value)
```

```
## # A tibble: 6 x 4
##   subject    Period correlation  p_value
##   <chr>      <chr>        <dbl>    <dbl>
## 1 Math       G1          -0.151  0.00437
## 2 Math       G2          -0.155  0.00330
## 3 Math       G3          -0.177  0.000761
## 4 Portuguese G1          -0.0920 0.0205
## 5 Portuguese G2          -0.106  0.00740
## 6 Portuguese G3          -0.120  0.00253
```

For Math:

- p-value < 0.05 for all periods. We reject the null hypothesis.
- There is a weak negative correlation between socialization and grade. The effect is roughly the same for all three periods.

For Portuguese:

- p-value < 0.05 for all periods. We reject the null hypothesis.
- There is also a statistically significant negative correlation between socialization and grade, but the effect is weaker than in Math.

Going out is negatively associated with academic performance in both Math and Portuguese, although the correlation is weak; the effect is also slightly stronger in Math, indicating that frequent socializing may interfere more with the concentration needed for mathematical problem solving.

Extracurricular Activities

- Null hypothesis: there is no correlation between students' participation in extracurricular activities and grades.

```
longer_data %>%
  group_by(subject, Period) %>%
  summarise(
    test = list(t.test(Grade ~ activities)),
    .groups = "drop"
  ) %>%
  mutate(
    tidied = map(test, tidy)
  ) %>%
  unnest(tidied) %>%
  select(subject, Period, estimate1, estimate2, p.value, conf.low, conf.high) %>%
  rename(
    mean_activities_no = estimate1,
    mean_activities_yes = estimate2,
    p_value = p.value,
    CI_lower = conf.low,
    CI_upper = conf.high
  )
```

```
## # A tibble: 6 x 7
##   subject    Period mean_activities_no mean_activities_yes p_value CI_lower
##   <chr>      <chr>               <dbl>               <dbl>   <dbl>    <dbl>
## 1 Math       G1                   11.1                11.5  0.220    -1.10
## 2 Math       G2                   11.1                11.6  0.208    -1.07
## 3 Math       G3                   11.3                11.7  0.270    -1.05
## 4 Portuguese G1                   11.3                11.7  0.0325   -0.872
## 5 Portuguese G2                   11.5                12.0  0.0187   -0.898
## 6 Portuguese G3                   12.0                12.4  0.0346   -0.869
## # i 1 more variable: CI_upper <dbl>
```

For math:

- p-value $>$ 0.05 for all periods. We fail to reject the null hypothesis.
- There is no significant difference between the grades of students with and without extracurricular activities.

For Portuguese:

- p-value $<$ 0.05 for all periods. We reject the null hypothesis.
- There is a slight difference in means of students participating or not participating in extracurricular activities. Students who do score roughly 0.5 points higher than those who don't.

There is a significant correlation between extracurricular activities and Portuguese grades but not for Math grades. This might suggest that students who participate in extracurricular activities tend to develop skills better related to language learning, such as reading, communication, and collaboration. In contrast, Math performance does not appear to be significantly influenced by extracurricular participation, indicating that achievement may rely more on structured academic inputs rather than soft-skill developments.

Romantic Relationship

- Null hypothesis: there is no correlation between having romantic relationships and grades.

```
longer_data %>%
  group_by(subject, Period) %>%
  summarise(
    test = list(t.test(Grade ~ romantic)),
    .groups = "drop"
  ) %>%
  mutate(
    tidied = map(test, tidy)
  ) %>%
  unnest(tidied) %>%
  select(subject, Period, estimate1, estimate2, p.value, conf.low, conf.high) %>%
  rename(
    mean_romantic_no = estimate1,
    mean_romantic_yes = estimate2,
    p_value = p.value,
    CI_lower = conf.low,
    CI_upper = conf.high
  )
```

```
## # A tibble: 6 x 7
##   subject    Period mean_romantic_no mean_romantic_yes p_value CI_lower CI_upper
##   <chr>      <chr>             <dbl>             <dbl>   <dbl>    <dbl>    <dbl>
## 1 Math       G1                 11.3              11.3  0.997   -0.692     0.695
## 2 Math       G2                 11.5              11.2  0.375   -0.367     0.969
## 3 Math       G3                 11.6              11.3  0.320   -0.338     1.03
## 4 Portuguese G1                 11.6              11.3  0.0915  -0.0601    0.802
## 5 Portuguese G2                 11.9              11.5  0.0846  -0.0509    0.793
## 6 Portuguese G3                 12.3              12.0  0.138   -0.108     0.779
```

For math:

- p-value $>$ 0.05 for all periods. We fail to reject the null hypothesis.
- There is no significant difference between the grades of students with and without romantic relationships.

For Portuguese:

- p-value $>$ 0.05 for all periods. We fail to reject the null hypothesis. However the p-value is significantly smaller than that for Math.
- There is no significant difference between the grades of students with and without romantic relationships either.

Romantic relationship is not significantly correlated with grades in either subject.

Alcohol Consumption

- Null hypothesis: students who consume alcohol on weekdays/weekends do not have a difference in grade from those who do not.

```
longer_data %>%
  group_by(subject, Period) %>%
  summarise(
    cor_test = list(cor.test(Walc, Grade, method = "pearson", use = "complete.obs")),
    .groups = "drop"
  ) %>%
  mutate(
    correlation = map_dbl(cor_test, ~ .x$estimate),
    p_value = map_dbl(cor_test, ~ .x$p.value)
  ) %>%
  select(subject, Period, correlation, p_value)
```

```
## # A tibble: 6 x 4
##   subject    Period correlation    p_value
##   <chr>      <chr>        <dbl>      <dbl>
## 1 Math       G1          -0.177 0.000807
## 2 Math       G2          -0.177 0.000801
## 3 Math       G3          -0.190 0.000305
## 4 Portuguese G1          -0.155 0.0000853
## 5 Portuguese G2          -0.169 0.0000199
## 6 Portuguese G3          -0.185 0.00000274
```

For Math:

- p-value < 0.05 for all periods. We reject the null hypothesis.
- There is a weak negative correlation between weekend alcohol consumption and grade. The effect does not show significant change over time.

For Portuguese

- p-value < 0.05 for all periods. We reject the null hypothesis.
- There is also a weak negative correlation between weekend alcohol consumption and grade. There is a slight increase in the strength from G1 to G3, but the change is moderate.

```
longer_data %>%
  group_by(subject, Period) %>%
  summarise(
    cor_test = list(cor.test(Dalc, Grade, method = "pearson", use = "complete.obs")),
    .groups = "drop"
  ) %>%
  mutate(
    correlation = map_dbl(cor_test, ~ .x$estimate),
    p_value = map_dbl(cor_test, ~ .x$p.value)
  ) %>%
  select(subject, Period, correlation, p_value)
```

```
## # A tibble: 6 x 4
##   subject    Period correlation      p_value
##   <chr>      <chr>         <dbl>        <dbl>
## 1 Math       G1           -0.129 0.0149
## 2 Math       G2           -0.128 0.0159
## 3 Math       G3           -0.141 0.00776
## 4 Portuguese G1           -0.190 0.00000154
## 5 Portuguese G2           -0.185 0.00000259
## 6 Portuguese G3           -0.211 0.0000000819
```

For Math:

- p-value < 0.05 for all periods. We reject the null hypothesis.
- There is a little to no correlation between weekday alcohol consumption and grade.

For Portuguese

- p-value < 0.05 for all periods. We reject the null hypothesis.
- There is also a weak negative correlation between weekend alcohol consumption and grade.

In general, there is a stronger negative correlation between alcohol consumption and grades in Portuguese than in Math. Students who drink more on weekdays tend to do worse in Portuguese.

- Weekend drinking negatively impacts performance in both subjects about equally, with weak correlations.
- Weekday drinking appears to have a stronger effect in Portuguese, suggesting students who drink during school days may struggle more with tasks like reading and writing.

Interestingly, for Math, weekend drinking has a slightly greater negative effect, but for Portuguese, weekday drinking has a slightly greater negative effect.

Conclusion

There is a growing negative correlation between absences and math grade over time. The negative correlation between absences and Portuguese grade stays roughly the same over time.

Study time positively correlates to grade stronger in Portuguese than in Math.

There is a significant negative correlation between number of past failures and grade in both subject, though the effect is stronger in Portuguese.

Going out is negatively associated with academic performance in both Math and Portuguese, although the correlation is weak; the effect is also slightly stronger in Math.

There is a significant correlation between extracurricular activities and Portuguese grades but not for Math grades.

Romantic relationship is not significantly correlated with grades in either subject.

Weekend drinking negatively impacts performance in both subjects about equally, with weak correlations. Weekday drinking appears to have a stronger effect in Portuguese. Math grade is affected more by weekend drinking than weekday drinking.

Behavioural Section Conclusion

Behavioral factors such as study time, travel time, free time, and social activities like going out significantly impact Portuguese grades, while only going out have an impact in Math.

In math, drinking is only related to G2, and it is slightly related to G1 and G3. In Portuguese, drinking is related to G1, G2, and G3 grades, especially on weekdays.

Portuguese are associated with absences. However, in Math, absences are only significantly related to G2 and G3.

Better health is associated with stronger performance in Portuguese G3 but not clearly in math.

The number of failures has a strong correlation with G1, G2, and G3 in both subjects.

Participating in activities positively correlates with higher Portuguese grade in G3, but shows little to no effect on Math grades and Portuguese G1 and G2.

Section 6 - Cross Sectional Variable Analysis

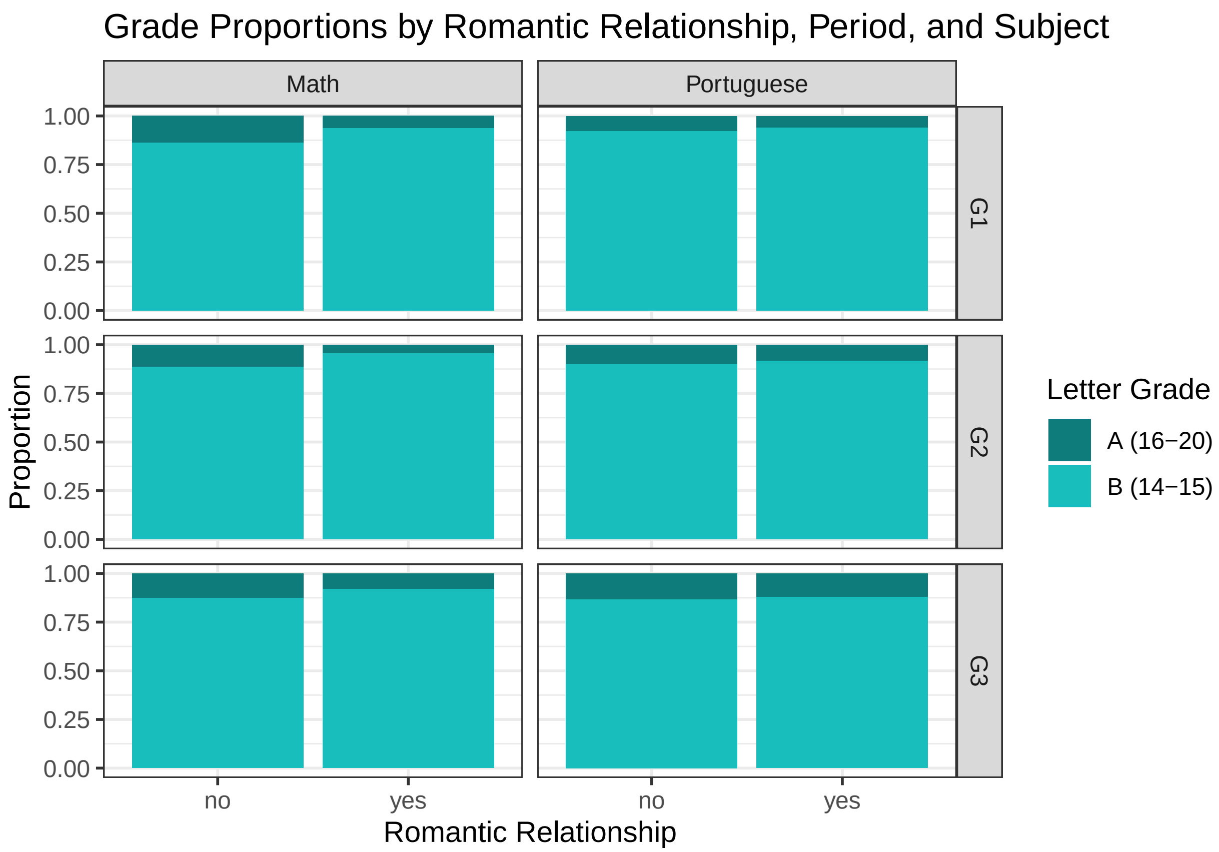
6.1 - Does being in a romantic relationship correlate with lower academic performance, and is this relationship potentially mediated by study time?

Q1: How do grades differ based on romantic relationship status?

- Null Hypothesis: There is no difference in grade distribution between students who are and are not in a romantic relationship.

```r
ggplot(longer_data, aes(x = romantic, fill = GradeGroup)) +
  geom_bar(position = "fill") +
  facet_grid(Period ~ subject) +
  labs(title = "Grade Proportions by Romantic Relationship, Period, and Subject",
       x = "Romantic Relationship",
       y = "Proportion",
       fill = "Letter Grade") +
  scale_fill_discrete(labels = c(
    "A" = "A (16-20)", "B" = "B (14-15)",
    "C" = "C (12-13)", "D" = "D (10-11)",
    "F" = "F (0 - 9)"
  )) +
    scale_fill_paletteer_d("PrettyCols::Beach", labels = c(
    "A (16-20)",
    "B (14-15)",
    "C (12-13)",
    "D (10-11)",
    "F (0 - 9)")) +
  theme_bw()
```

```
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```



Visual interpretation

Across all time periods and both subjects, students not in a romantic relationship ("no") consistently show a slightly higher proportion of top grade (A) compared to those in relationships ("yes"). Conversely, students in a romantic relationship tend to have slightly higher proportions of lower grades (C–F), particularly evident in the final period (G3). The trend is more noticeable in Math than in Portuguese, suggesting that romantic relationships may have a greater impact on performance in math coursework.

```
test_groups <- longer_data %>%
  group_by(Period, subject) %>%
  group_split()

for (g in test_groups) {
  cat("\n--- Period:", unique(g$Period), "| Subject:", unique(g$subject), "---\n")
  tbl <- table(g$GradeGroup, g$romantic)
  print(chisq.test(tbl))
}
```

```
##
## --- Period: G1 | Subject: Math ---
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl
## X-squared = 3.6806, df = 1, p-value = 0.05505
##
```

```
##
## --- Period: G1 | Subject: Portuguese ---
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl
## X-squared = 0.48529, df = 1, p-value = 0.486
##
##
## --- Period: G2 | Subject: Math ---
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl
## X-squared = 3.6523, df = 1, p-value = 0.05599
##
##
## --- Period: G2 | Subject: Portuguese ---
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl
## X-squared = 0.4091, df = 1, p-value = 0.5224
##
##
## --- Period: G3 | Subject: Math ---
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl
## X-squared = 1.2157, df = 1, p-value = 0.2702
##
##
## --- Period: G3 | Subject: Portuguese ---
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl
## X-squared = 0.094315, df = 1, p-value = 0.7588
```

Statistical Test

Summary of p-values:

- G1 Math: p = 0.03836; statistically significant, reject H0
- G1 Portuguese: p = 0.3487; not significant
- G2 Math: p = 0.06236; marginal, not significant at 0.05
- G2 Portuguese: p = 0.3872; not significant
- G3 Math: p = 0.1799; not significant
- G3 Portuguese: p = 0.67; not significant

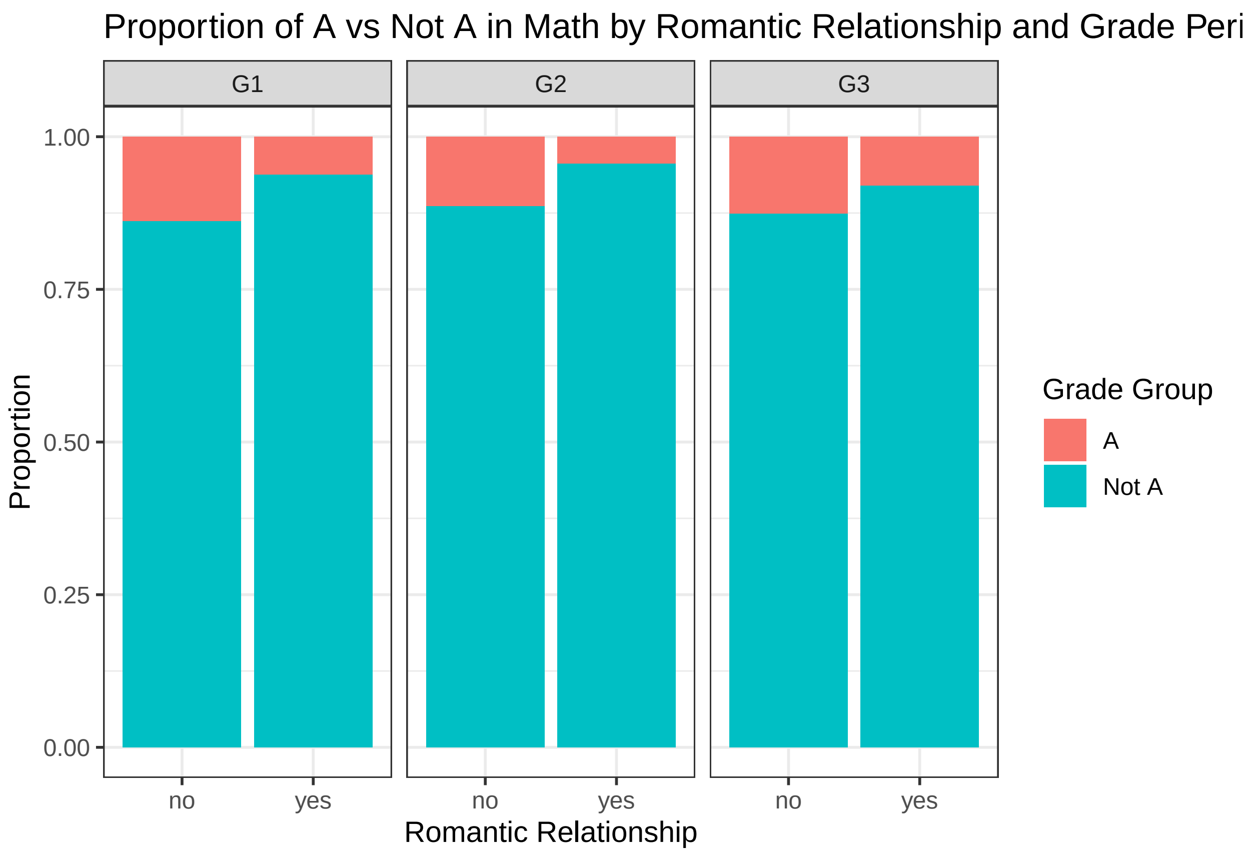
Answer

We reject the null hypothesis for G1 Math, indicating a statistically significant difference in grade distribution between students in and not in a romantic relationship during the first grade period of Math.

For all other combinations of subject and period, we fail to reject the null hypothesis. This suggests that, in general, there is no strong statistical evidence that romantic relationship status affects grade across most periods and subjects, except for G1 Math. A possible explanation is that students in romantic relationships are equally capable of managing their academic responsibilities, not distracted as what many people believe, and any early differences in G1 Math may be due to random variation or other unmeasured factors rather than the relationship status itself.

However, personally I am still curious if there is a possible correlation between top students who are getting As and who are not, particularly for math subject which shown a more noticeable trend in the previous graph. The following investigation focused on exploring the potential relationship of romantic relationship status with the proportion of getting an A in math subject.

```r
longer_math_data_cleaned_A <- math_data_cleaned %>%
  pivot_longer(cols = G1:G3, names_to = "Period", values_to = "Grade") %>%
  mutate(
    GradeGroup = ifelse(Grade >= 16, "A", "Not A")
  )

ggplot(longer_math_data_cleaned_A, aes(x = romantic, fill = GradeGroup)) +
  geom_bar(position = "fill") +
  facet_wrap(~ Period) +
  labs(
    title = "Proportion of A vs Not A in Math by Romantic Relationship and Grade Period",
    x = "Romantic Relationship",
    y = "Proportion",
    fill = "Grade Group"
  ) +
  theme_bw()
```



Visual interpretation

The chart shows that in all three periods, students not in a romantic relationship consistently have a higher proportion of A grades in Math compared to those who are. This suggests a possible link between relationship status and academic performance for top students who achieve the highest grade. A possible reason is that students not in relationships may have more time or focus for studying, which will help them in understanding the material. This is what hints the subsequent question, aiming to explore if there is a correlation between romantic relationship with study time. However, other factors could also be at play, so this pattern should be interpreted with caution.

```r
a_test_data <- longer_data %>%
  filter(subject == "Math") %>%
  mutate(A_Group = ifelse(GradeGroup == "A", "A", "Not A"))

a_test_groups <- a_test_data %>%
  group_by(Period) %>%
  group_split()

for (g in a_test_groups) {
  cat("\n--- Period:", unique(g$Period), "---\n")
  tbl <- table(g$A_Group, g$romantic)
  print(chisq.test(tbl))
}
```

```
##
## --- Period: G1 ---
```

```
##
##  Pearson’s Chi-squared test with Yates’ continuity correction
##
## data:  tbl
## X-squared = 3.6806, df = 1, p-value = 0.05505
##
##
## --- Period: G2 ---
##
##  Pearson’s Chi-squared test with Yates’ continuity correction
##
## data:  tbl
## X-squared = 3.6523, df = 1, p-value = 0.05599
##
##
## --- Period: G3 ---
##
##  Pearson’s Chi-squared test with Yates’ continuity correction
##
## data:  tbl
## X-squared = 1.2157, df = 1, p-value = 0.2702
```

Statistical test

- G1: p = 0.05505
- G2: p = 0.05599
- G3: p = 0.2702

Interpretation

In both G1 and G2, the p-values (~0.055) are just above the conventional significance threshold of 0.05, suggesting a marginal, but not statistically significant association between romantic relationship status and the likelihood of receiving an A in Math. In G3, the p-value is much higher (0.2702), indicating no significant relationship at all. These results imply that any differences in A-grade proportions between students with or without a romantic relationship are not strong enough to be considered statistically significant, though the G1 and G2 results hint at a possible weak trend that might worth further investigation with larger sample sizes.

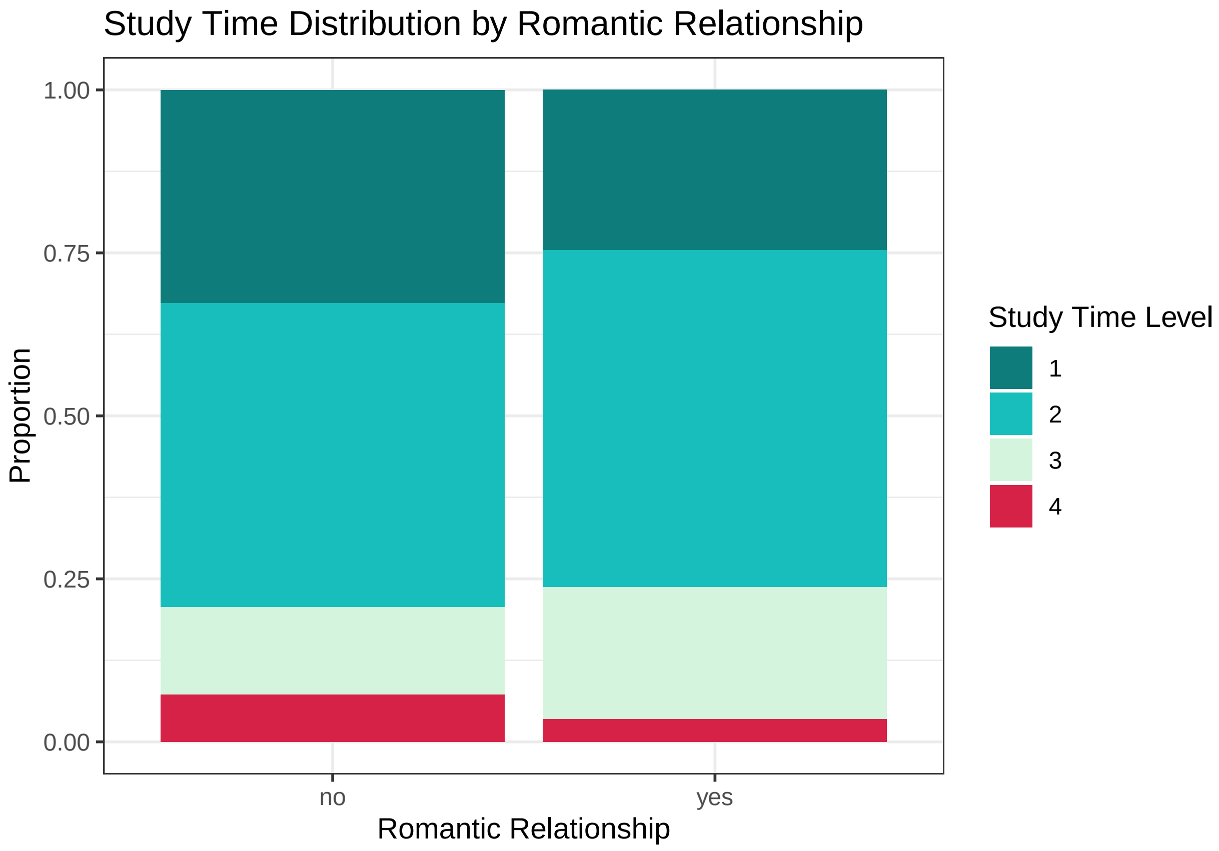
Correlation Between Romantic Relationship and A Grades in Math

In general, the graph shows a visible trend: students not in a romantic relationship tend to receive more A grades in Math across all periods. However, the chi-squared tests did not find this difference to be statistically significant, with p-values slightly above 0.05 in G1 and G2, and clearly non-significant in G3. This suggests that while there may be a pattern, we don’t have strong enough evidence to confidently claim a relationship between romantic status and top performance in Math.

Q2: Does being in a romantic relationship correlate with less study time?

- Null Hypothesis: Study time does not differ between students in romantic relationships and those who are not.

```r
ggplot(longer_data, aes(x = romantic, fill = as.factor(studytime))) +
  geom_bar(position = "fill") +
  labs(title = "Study Time Distribution by Romantic Relationship",
       x = "Romantic Relationship",
       y = "Proportion",
       fill = "Study Time Level")+
    scale_fill_paletteer_d("PrettyCols::Beach") +
  theme_bw()
```



Visual interpretation

Students not in a romantic relationship have a higher proportion of individuals in the lowest study time category (1 = <2 hours) compared to those in a relationship. Conversely, students in a romantic relationship show a slightly higher proportion in moderate study time levels (2 and 3 = 2–10 hours). The proportion of students studying more than 10 hours (level 4) is low overall, but appears slightly lower among those in a romantic relationship.

```
chisq.test(table(longer_data$studytime, longer_data$romantic))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(longer_data$studytime, longer_data$romantic)
## X-squared = 53.63, df = 3, p-value = 1.346e-11
```

Statistical Test (Chi-squared)

p-value = 1.346×10^{-11}

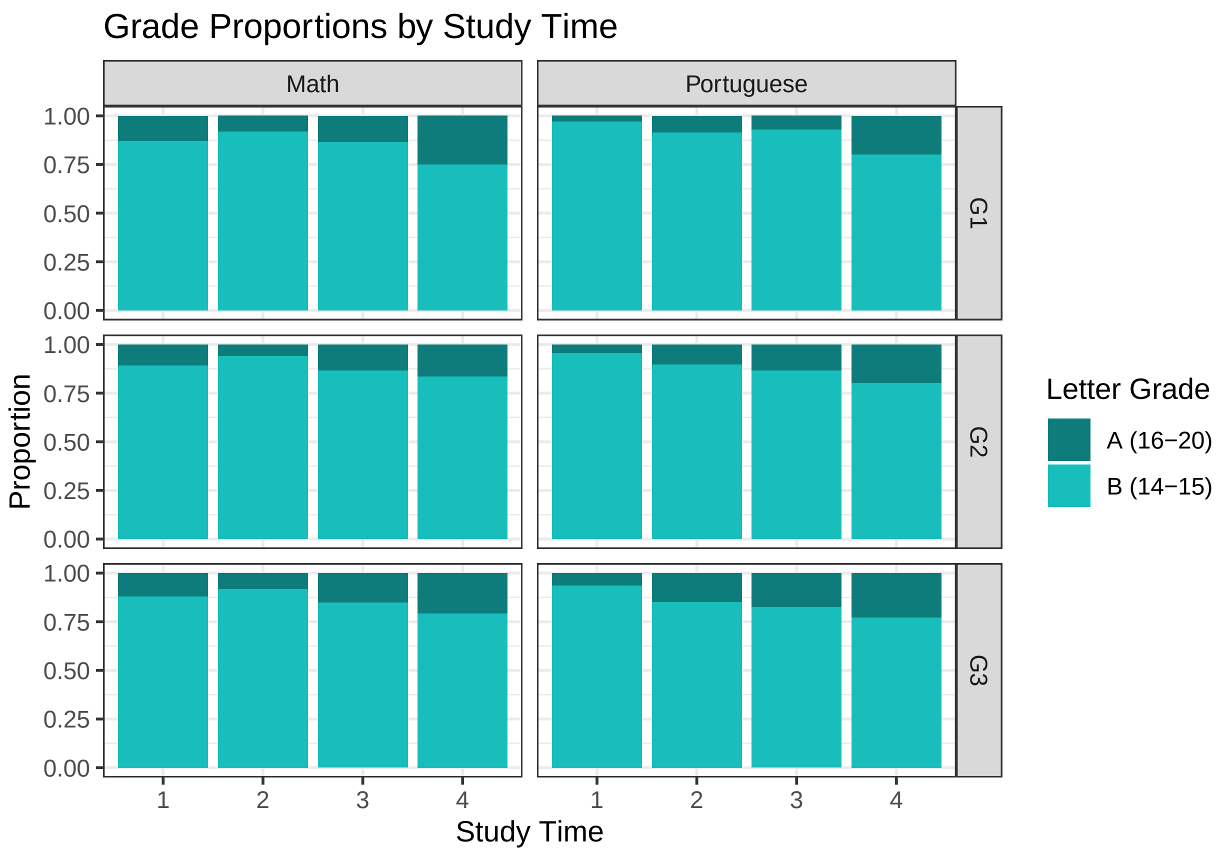
Answer

Because the p-value is far below 0.05, we reject the null hypothesis. This provides strong statistical evidence that study time distribution significantly differs between students with and without a romantic relationship. This result supports what was observed visually: romantic relationship status is associated with how much students study per week.

Q3: Does study time impact student grades?

- Null Hypothesis: The relationship between average grade and consistency is the same for Math and Portuguese.

```r
ggplot(longer_data, aes(x = as.factor(studytime), fill = GradeGroup)) +
  geom_bar(position = "fill") +
  facet_grid(Period ~ subject) +
  labs(
    title = "Grade Proportions by Study Time",
    x = "Study Time",
    y = "Proportion",
    fill = "Letter Grade"
  ) +
    scale_fill_paletteer_d("PrettyCols::Beach", labels = c(
    "A (16-20)",
    "B (14-15)",
    "C (12-13)",
    "D (10-11)",
    "F (0 - 9)")) +
  theme_bw()
```



Visual interpretation

In both subjects, across all periods, higher study time levels (3 and 4) are generally associated with higher proportions of A and B grades, and lower proportions of F grades. Conversely, lower study times (1 and 2) show larger proportions of D and F grades, especially for Math. The trend becomes more pronounced in G3, suggesting that consistent or increasing study time might accumulate to better final grades. So, there is a clear positive correlation between study time and academic performance. Students who invest more hours weekly in studying tend to earn better grades, supporting the idea that increased study time is beneficial for the study of both Math and Portuguese course.

```
test_groups <- longer_data %>%
  group_by(Period, subject) %>%
  group_split()

for (g in test_groups) {
  cat("\n--- Period:", unique(g$Period), "| Subject:", unique(g$subject), "---\n")
  tbl <- table(g$GradeGroup, g$studytime)
  print(chisq.test(tbl))
}
```

```
##
## --- Period: G1 | Subject: Math ---


##
##  Pearson's Chi-squared test
##
```

```
## data:  tbl
## X-squared = 6.665, df = 3, p-value = 0.08338
##
##
## --- Period: G1 | Subject: Portuguese ---

##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 15.047, df = 3, p-value = 0.001777
##
##
## --- Period: G2 | Subject: Math ---

##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 5.3972, df = 3, p-value = 0.1449
##
##
## --- Period: G2 | Subject: Portuguese ---

##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 12.674, df = 3, p-value = 0.005397
##
##
## --- Period: G3 | Subject: Math ---

##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 4.8671, df = 3, p-value = 0.1818
##
##
## --- Period: G3 | Subject: Portuguese ---

##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 13.563, df = 3, p-value = 0.003564
```

Statistical Test

Period, subject, p value, significance

- G1, Math,0.1214, Not Significant

- G1, Portuguese,4.08×10^{-8},Significant
- G2, Math, 0.1640, Not Significant
- G2, Portuguese, 6.42×10^{-6}, Significant
- G3, Math, 0.1413, not Significant
- G3, Portuguese, 1.42×10^{-6}, Significant

Answer

Portuguese grades show a statistically significant relationship with study time across all three grading periods (G1, G2, G3). This supports the idea that study time is associated with improved performance in Portuguese. Math grades do not show significant differences by study time, suggesting a weaker or more complex relationship that may be influenced by other variables. These findings partially support the thought that study time affects academic performance—stronger in Portuguese than Math.

Answer to 6.1

In general, there is no consistent or strong evidence that being in a romantic relationship correlates with lower academic performance, and the hypothesis that this is mediated by reduced study time is not supported by the data.

In fact: Students in relationships are not studying less; they are slightly more represented in moderate study time categories, even though the students who get As tend to not be within a romantic relationship. Also, only G1 Math shows a statistically significant difference in grades between romantic groups, which is not enough to suggest a strong pattern. While more study time does correlate with better performance, the link between romantic relationships and study time does not follow the expected direction.

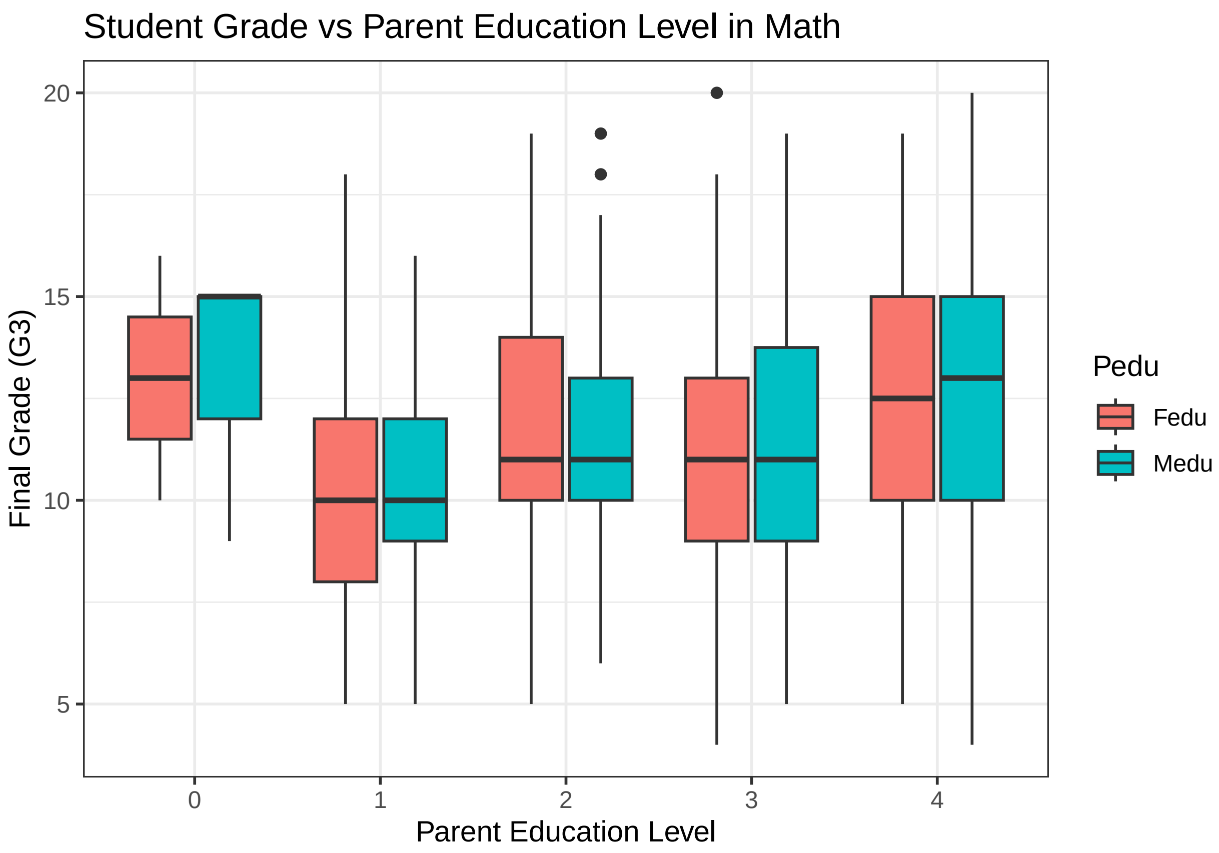
Overall, the data does not support the idea that romantic relationships are detrimental to academic performance via reduced study time. Any observed differences are minor, inconsistent, or even contradict initial assumptions that romantic relationship would correlate with a low study time and a low score on test.

6.2 - Does the educational attainment of a student's parent correlate with academic performance, and is this relationship stronger when the educated parent is the student's guardian?

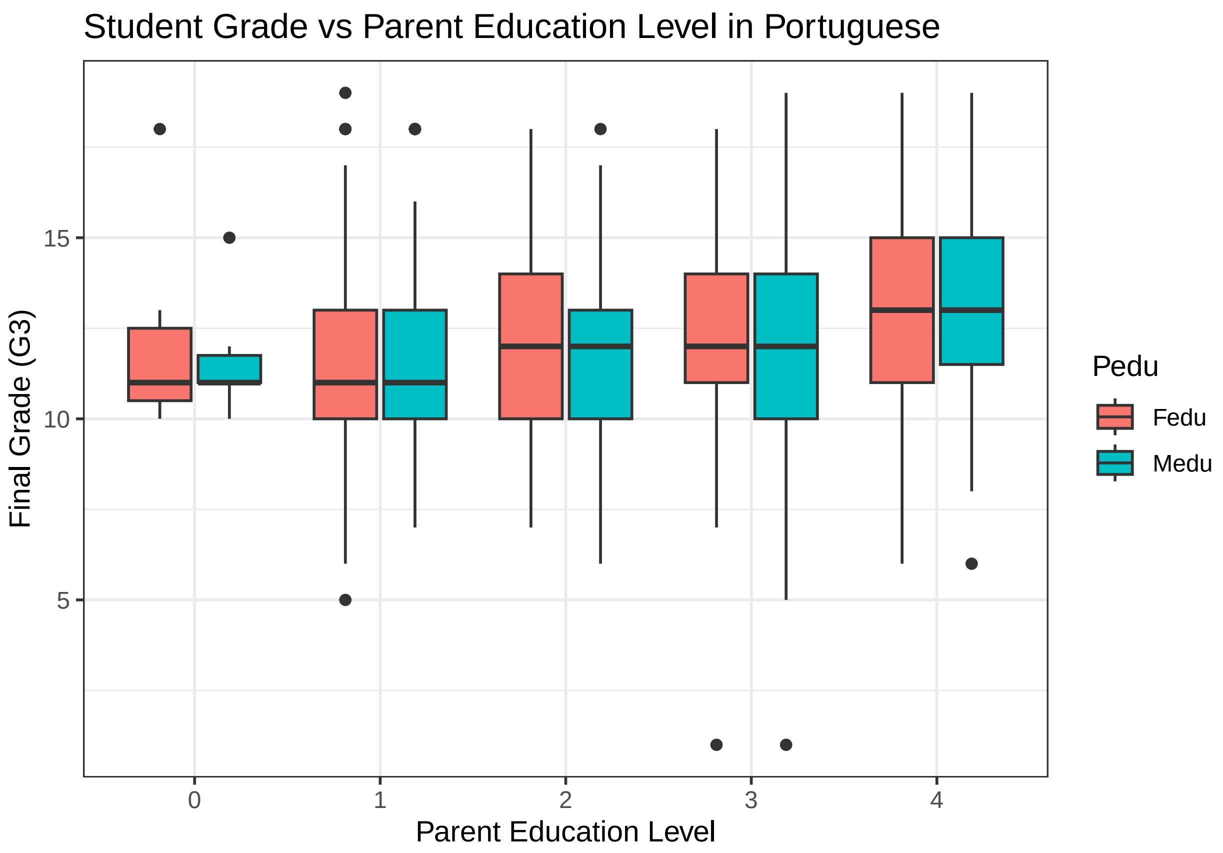
Q1: Is there a correlation between parental education level and final grades?

- Null Hypothesis: There is no association between either parent's education level and the student's final grade (G3).

```r
byparent_math <- math_data_cleaned %>%
  pivot_longer(cols=Medu:Fedu,names_to="Pedu",values_to="eduLevel")
byparent_port <- port_data_cleaned %>%
  pivot_longer(cols=Medu:Fedu,names_to="Pedu",values_to="eduLevel")

ggplot(byparent_math, aes(x=as.factor(eduLevel), y = G3, fill=Pedu)) +
  geom_boxplot() +
  labs(x = "Parent Education Level", y = "Final Grade (G3)",
       title = "Student Grade vs Parent Education Level in Math")+
  theme_bw()
```



```
ggplot(byparent_port, aes(x=as.factor(eduLevel), y = G3, fill=Pedu)) +
  geom_boxplot() +
  labs(x = "Parent Education Level", y = "Final Grade (G3)",
       title = "Student Grade vs Parent Education Level in Portuguese")+
  theme_bw()
```



Visual interpretation

1. Math: Both Mother's and Father's Education Level vs Student Final Grade show a general positive trend: as the parent's education level increases, the median student grade tends to rise. Higher parental education levels—both maternal and paternal—are generally associated with higher student academic performance. The relationship is more consistently positive for mothers than for fathers in this dataset. However there is one exception where parent's education level with 0, which is none, student showed a high median of final grade. But overall, these patterns still suggest a correlational link between parents' educational attainment and student outcomes, likely due to factors like educational support at home, value on education, and socio-economic advantages.
2. Portuguese: Both Mother's and Father's Education Level vs Student Final Grade also show a general positive trend: as the parent's education level increases, the median student grade tends to rise. Notice that students' with parental education attainment being none shows a lowest median grade, which matches with the normal intuition that lower parental education attainment might link to smaller positive influence and support in student's education. This also suggest that the values for parent education being 0 in Math class might be inaccurate or influenced by other external factors.

```
summary(aov(G3 ~ factor(Medu), data = math_data_cleaned))
```

```
##                  Df Sum Sq Mean Sq F value  Pr(>F)
## factor(Medu)   4    179   44.83   4.471 0.00155 **
## Residuals    352   3530   10.03
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(aov(G3 ~ factor(Fedu), data = math_data_cleaned))
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)   
## factor(Fedu)   4    139   34.63   3.414 0.00933 **
## Residuals    352   3571   10.14                  
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
summary(aov(G3 ~ factor(Medu), data = port_data_cleaned))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## factor(Medu)   4    392   98.06   14.71 1.72e-11 ***
## Residuals    629   4194    6.67                    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
summary(aov(G3 ~ factor(Fedu), data = port_data_cleaned))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## factor(Fedu)   4    199   49.64   7.116 1.32e-05 ***
## Residuals    629   4388    6.98                    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Statistical Test (ANOVA)

Math subject

- Mother’s Education (Medu) $p = 0.00155$, statistically significant
- Father’s Education (Fedu) $p = 0.00933$, statistically significant

Portuguese Subject

- Mother’s Education (Medu) $p = 1.72 \times 10^{-11}$, statistically significant
- Father’s Education (Fedu) $p = 1.32 \times 10^{-5}$, statistically significant

Both maternal and paternal education levels are significantly associated with student academic performance in both Math and Portuguese. The effect is stronger for Portuguese than for Math, based on lower p-values. In both subjects, mother’s education has a slightly stronger statistical impact than father’s education.

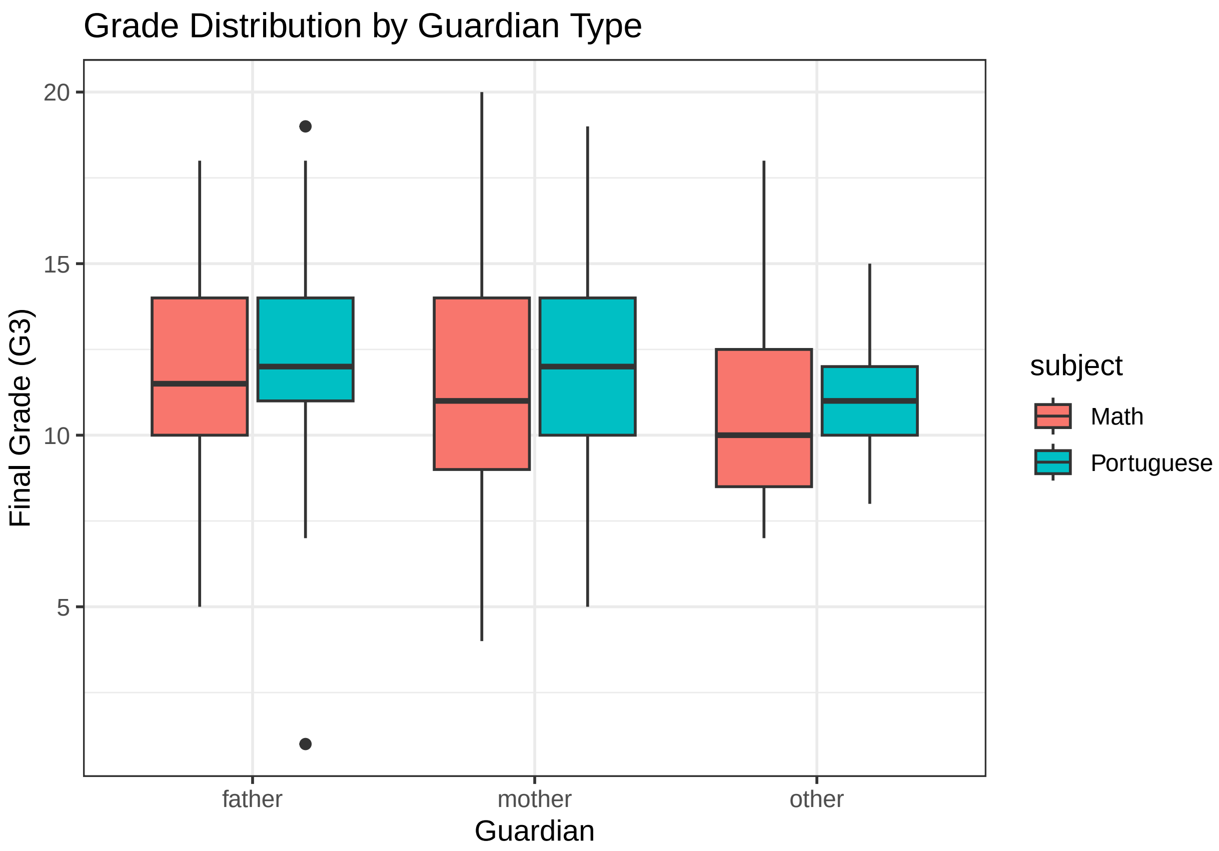
Answer

These results support the conclusion that parental education—particularly the mother’s is correlated with student achievement, and that higher levels of parental education are associated with higher final grades. A possible reason is that more highly educated parents, especially mothers, may be better equipped to support their children academically through help with homework, setting goals, or fostering a home environment that values learning.

Q2: Is there a difference in grades based on the guardian identity?

- Null Hypothesis: Students’ final grades do not vary based on whether their guardian is their mother, father, or someone else.

```r
ggplot(combined_data, aes(x = guardian, y = G3,fill=subject)) +
  geom_boxplot() +
  labs(x = "Guardian", y = "Final Grade (G3)",
       title = "Grade Distribution by Guardian Type")+
  theme_bw()
```



Visual interpretation

Math: Father and Mother guardianship groups show similar median grades, around 10–12. Students with guardians listed as “other” have a lower median grade, indicating generally lower performance and less variability.

Portuguese: Again, students whose guardian is listed as “other” have slightly lower median grades and a more compressed score distribution. Those with mother or father guardianship show higher median grades and more variation in performance. The difference between “father” and “mother” guardianship is small, with both medians around 12–13.

```r
summary(aov(G3 ~ guardian, data = math_data_cleaned))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## guardian      2     20   10.08   0.968  0.381
## Residuals   354   3689   10.42
```

```r
summary(aov(G3 ~ guardian, data = port_data_cleaned))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## guardian      2     51  25.544   3.554 0.0292 *
## Residuals   631   4536   7.188
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Statistical Test

- Math: p = 0.381; Not significant.
- Portuguese: p = 0.0292 ; Statistically significant.

Guardian type has an effect on Portuguese grades, but not on Math grades. Students with parental guardians (father or mother) tend to do better.

Answer

The impact of guardian identity on academic performance depends on the subject: For Math, guardian identity does not matter significantly. For Portuguese, guardian identity does influence outcomes, with parental guardianship linked to higher grades. Thus, we reject the null hypothesis for Portuguese but fail to reject it for Math. A possible reason is that Portuguese performance may be more influenced by home language environment and communication, where the presence of a parent as guardian could provide more consistent language environment. In contrast, Math performance might rely more on school-based instruction and practice, making it less sensitive to guardian identity.

Q3: Does parental education correlate more strongly with grades if that parent is the student's guardian?

- Null Hypothesis: There is no difference in the correlation between parental education and student grades, regardless of whether the parent is the guardian.

```
mother_guardian_math <- math_data_cleaned %>% filter(guardian == "mother")
father_guardian_math <- math_data_cleaned %>% filter(guardian == "father")

summary(aov(G3 ~ factor(Medu), data = mother_guardian_math))
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## factor(Medu)    4  137.6   34.39   3.231 0.0131 *
## Residuals     243 2586.0   10.64
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(G3 ~ factor(Fedu), data = father_guardian_math))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## factor(Fedu)  3   41.9  13.972   1.497  0.222
## Residuals    78  728.2   9.336
```

```
mother_guardian_port <- port_data_cleaned %>% filter(guardian == "mother")
father_guardian_port <- port_data_cleaned %>% filter(guardian == "father")

summary(aov(G3 ~ factor(Medu), data = mother_guardian_port))
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)    
## factor(Medu)   4  346.5   86.64   12.93 5.86e-10 ***
## Residuals    439 2942.2    6.70                     
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
summary(aov(G3 ~ factor(Fedu), data = father_guardian_port))
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## factor(Fedu)   3     28   9.337    1.28  0.284
## Residuals    146   1065   7.295               
```

Answer

Mother’s Education (Medu): When mother is the guardian:

- Math: p = 0.0131; Significant
- Portuguese: p = 5.86×10^{-10}; significant

When the mother is the guardian, her education level significantly predicts the student’s academic performance, especially in Portuguese. This suggests a stronger influence of the mother’s education when she is actively involved in caregiving. A possible reason for this strong influence being especially strong from mothers is that mothers are who often take on caregiving and educational roles in many households, which plays a more significant role in the child’s life.

Father’s Education (Fedu) When father is the guardian:

- Math: p = 0.222; not significant
- Portuguese: p = 0.284; not significant

When the father is the guardian, his education does not significantly impact the student’s final grades in either subject. This suggests that father’s role as a guardian does not enhance the influence of his education on the child’s performance.

Answer to 6.2

Yes, parental education is positively correlated with student academic performance. This correlation is stronger when the mother is the guardian, especially in Portuguese. The relationship is less pronounced or not significant when the father is the guardian or for Math grades. This suggests that maternal educational influence may have a more direct academic benefit when the mother also serves as the primary guardian.

Section 7 - Systematic Difference in Performance Between Subjects

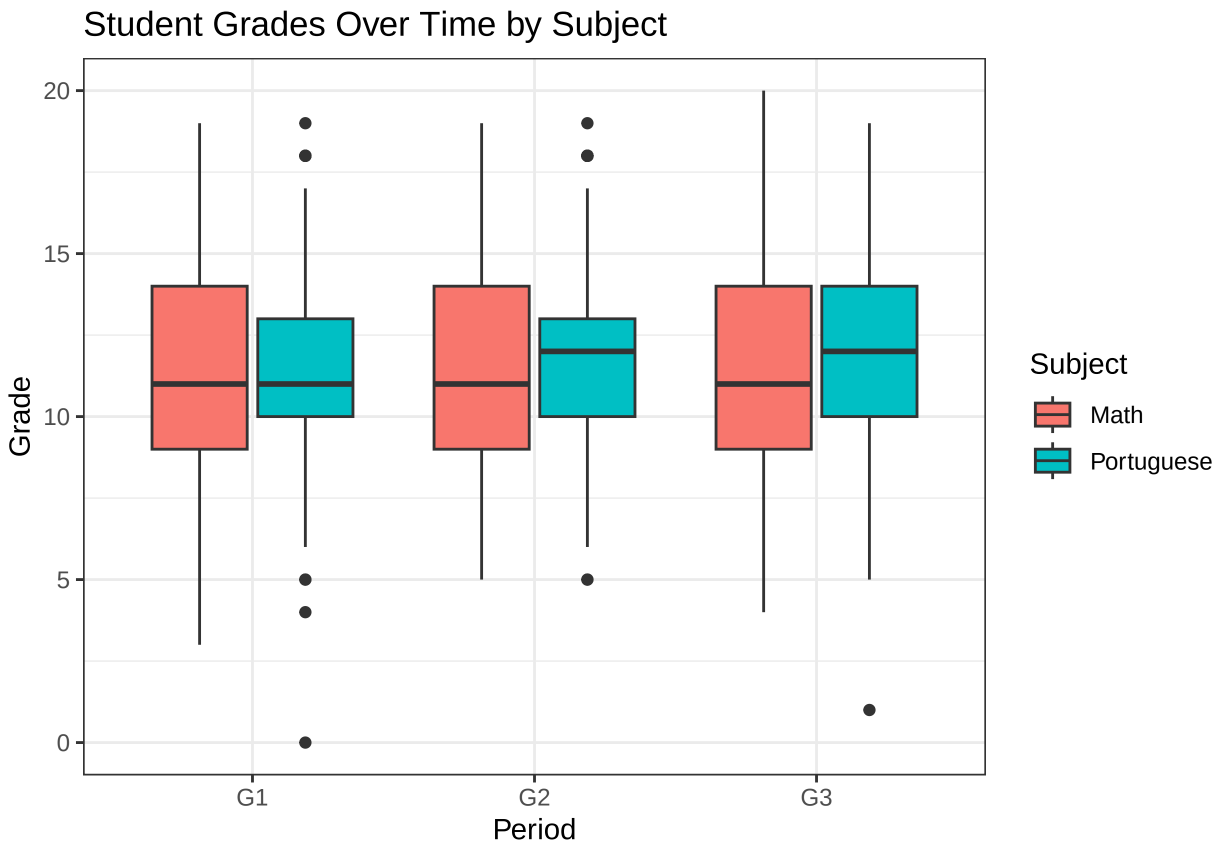
We are interested in whether there is a difference in the strength of correlation between variables and grades based on subject, addressing the second part of the overarching research question: are there subject-based systematic differences?

For the variables we selected, we will conduct statistical tests to find their correlation with grades based on subjects and periods. We will analyze whether there exists a statistically significant correlation, and if there is any, how strong these correlations are for each subject in each period. We will compare these results between subjects in order to answer our sectional research question.

7.1 - How do grades change over periods, and do these trends differ between Math and Portuguese courses?

- Null Hypothesis: The mean grades are equal across all periods and subjects, the effect of period on grade is the same for both subjects.

```
ggplot(longer_data, aes(x = Period, y = Grade, fill = subject)) +
  geom_boxplot() +
  labs(title = "Student Grades Over Time by Subject",
       fill = "Subject")+
  theme_bw()
```



```r
subject_test1 <- aov(Grade ~ subject  * Period, data = longer_data)
summary(subject_test1)
```

```
##                   Df Sum Sq Mean Sq F value   Pr(>F)    
## subject            1    127  127.07  15.399  8.9e-05 ***
## Period             2    143   71.49   8.664 0.000177 ***
## subject:Period     2     22   10.83   1.312 0.269325    
## Residuals       2967  24484    8.25                     
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Visual Observation

1. Math: grades appear to have no obvious change over periods. 2. Portuguese: grades shows an upward trend with increasing median and third quartile from the boxplot.

Statistical Test (ANOVA)

1. Subject: p-value = 8.9e-05 < 0.05. We reject the null hypothesis.
2. subject:Period: p-value = 0.269325 > 0.05. We fail to reject the null hypothesis. There is not a difference in the pattern of grade change by subject.

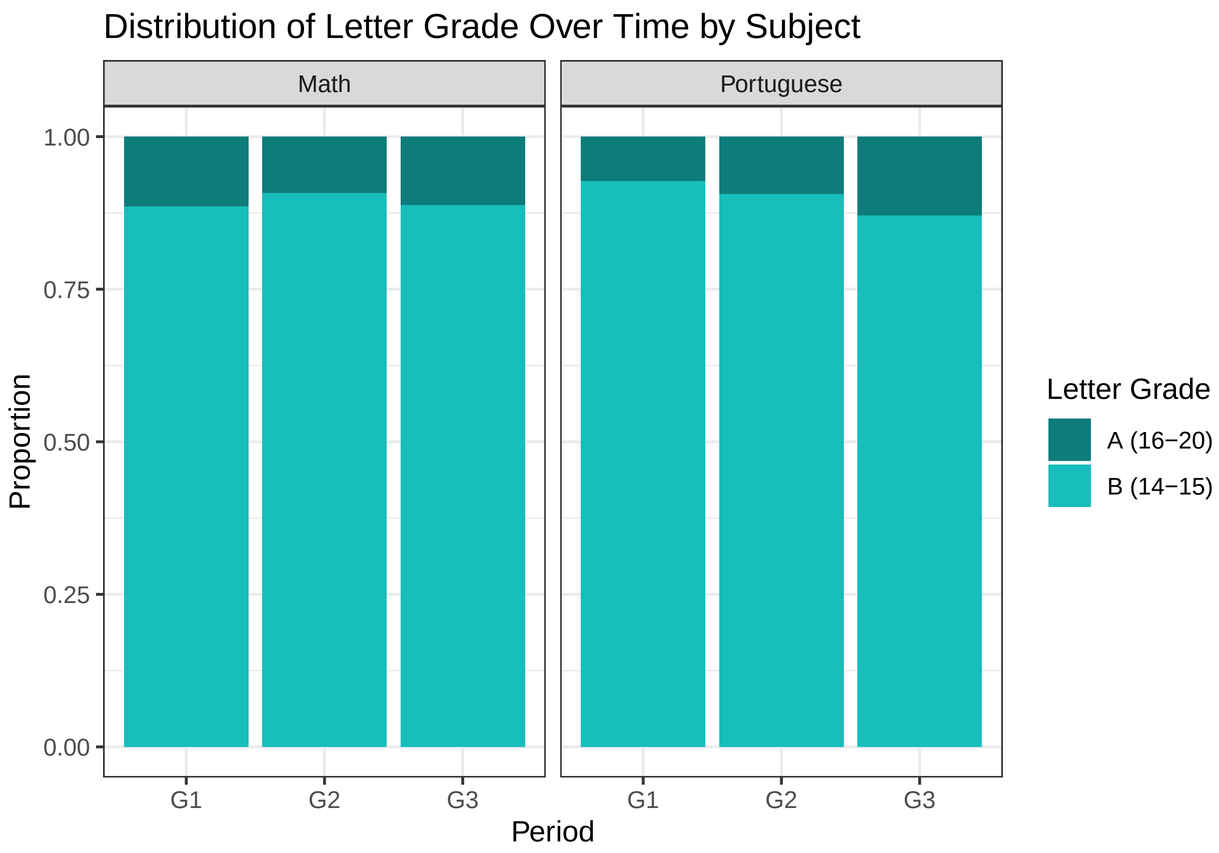
Conclusion

There is significantly differed grade based on subject over time. Specifically, students perform better in Portuguese than in math overall. Grades also improve over time periods, especially in Portuguese. This trend is further proven by the non-significant interaction between period and subject with grade, meaning the pattern of grade change doesn’t change by subject based on periods, both subjects show the same pattern of increasing over time.

7.2 - How does the distribution of letter grades shift over time by subject?

- Null Hypothesis: Letter grade isn’t impacted by period.

```r
ggplot(longer_data, aes(x = Period, fill = GradeGroup)) +
  geom_bar(position = "fill") +
  labs(title = "Distribution of Letter Grade Over Time by Subject",
       fill = "Letter Grade",
       y = "Proportion") +
  facet_wrap(~subject)+
    scale_fill_paletteer_d("PrettyCols::Beach", labels = c(
    "A (16-20)",
    "B (14-15)",
    "C (12-13)",
    "D (10-11)",
    "F (0 - 9)")) +
  theme_bw()
```



```r
longer_data %>%
  filter(subject == "Math") %>%
  with(table(GradeGroup, Period)) %>%
  chisq.test()
```

```
##
##  Pearson’s Chi-squared test
##
## data:  .
## X-squared = 1.1191, df = 2, p-value = 0.5715
```

```r
longer_data %>%
  filter(subject == "Portuguese") %>%
  with(table(GradeGroup, Period)) %>%
  chisq.test()
```

```
##
##  Pearson’s Chi-squared test
##
## data:  .
## X-squared = 11.663, df = 2, p-value = 0.002933
```

Visual Observation

1. Math: letter grade proportion doesn’t show a trend of change.

2. Portuguese: letter grade proportion shows an upward trend of increasing grade by period.

Statistical Test (Chi-squared)

1. Math: p-value = $0.3368 > 0.05$, fail to reject H0, math letter grade isn't impacted by period.
2. Portuguese: p-value = $0.0002564 < 0.05$, reject H0, Portuguese letter grade is impacted by period.

By-Subject Section Conclusion

Subject variable impacts grade over time by showing how letter grade change across periods based on subjects. Specifically, math shows no significant shift in letter grade distribution over time, indicating more consistent student performance across periods. In contrast, Portuguese shows a significant upward trend in letter grades, with more higher grades appearing over time by period.

The difference in performance could be due to differences in subject difficulty, and learning curves.

Section 8 Conclusion, Insight, Improvement

8.1 Improvement

If we were to approach this project again, we would like to conduct more second-layer analyses, such as exploring correlations, interaction effects, and potential mediating variables, building on the first-layer direct relationships we analyzed this time. This additional depth could help us better understand the underlying mechanisms and contribute to clearer cause-and-effect interpretations, although we recognize that establishing true causality would require careful modeling and possibly experimental designs.

Another improvement would be ensuring that all students involved in the data collection process have a thorough understanding of the meaning of each scale value (1 to 5) for the survey questions they receive. This would help ensure that the data collected more accurately reflects students' true perspectives and conditions, reducing potential measurement bias.

Furthermore, in Section 7 we found that there may exists a structural differences between Math and Portuguese study, likely due to the foundation differences in Math (a science-based class) and Portuguese (a language-based class). However, the exact reason for these differences is hypothesized but not explored fully due to the limitations of the data set. Further collection of data and possibly follow-up questionnaires is needed to answer these questions and build a more effective model to predict grades based on variables involved in this study. There are also other variables, such as demographic variables, that are not explored in this section. In future research, we can conduct similar analysis on these variables.

Lastly, our entire analysis was conducted based on our own manipulations and interpretations of the data, without applying methods from the original paper (Cortez & Silva, 2008). In future iterations, we would like to explore and implement techniques used in the paper, such as data mining approaches, and also other techniques like linear regression and models to enhance our interpretations and refine our analytical framework.

8.2 Overall Conclusion

This data exploration aimed to investigate if there is a potential correlation between various variables and the student's academic performance. Our analysis revealed that academic performance is shaped by a combination of demographic, social, and behavioral factors—but not always in the ways we expected.

Demographic factors like age, address, and sex show different associations with academic performance. Older students tend to perform worse in Portuguese, though this trend is not evident in Math. Also, rural students consistently underperform compared to urban students in both subjects, suggesting potential disparities in educational support. Lastly, gender impacts are subject-specific. Male students perform better in Math, while female students tend to outperform in Portuguese.

Behavioral factors such as study time, travel time, free time, and social activities like going out significantly impact Portuguese grades, while only going out have an impact in Math. Alcohol consumption, especially on weekdays, has a greater influence on Portuguese across all periods, while its impact on Math is weaker and less consistent. Students in better health tend to perform better. Prior course failures are strongly associated with lower grades in both subjects. Lastly, participating in activities positively correlates with higher Portuguese grade, particularly in G3, but shows little to no effect on Math grades.

Regarding social support, we find that the proportion of students getting As is almost 6% higher when family relationship is good. We also find that having a desire to achieve higher education is correlated with getting an A, but there is no gender-based difference in this "effect." Finally, we find that the proportion of students wanting to pursue higher education decreases with age. This aligns with a finding in the Demographic section - that older students get lower grades, particularly in Portuguese.

During the more complex layered analysis exploring how two variables might interact and relate to academic performance, we uncovered several key insights. Regarding the correlation between romantic relationships, study time and grade, we initially expected to find that students in relationships might study less and perform worse. However, the data didn't consistently support this assumption. There was no strong evidence showing a negative impact. In contrast, parental education, especially for mother's, did show a clear positive correlation with student grades, and this effect was particularly pronounced when the mother was also the student's guardian.

Lastly, during our analysis investigating the systematic difference in what types of variables might affect grades in Math and Portuguese each, it was found that Math appears more affected by acute disruptions and structural challenges, such as absences and lack of focus in class. Portuguese is more sensitive to cumulative behaviors and lifestyle consistency, such as study time and weekday alcohol consumption. These differences suggest that interventions should be subject-specific: success in Math requires maintaining attentiveness in problem-solving, while Portuguese success relies more on building supportive habits and long-term habits.