Ananya Agarwal, Jessica Laman, Angelina Liu, Yuqi Zhang

MA213 Paper

Group A9

May 03, 2023

**Influence of Different Variables on Student’s Success**

Introduction

Education is fundamental to young adults’ growth into adulthood and participation as active members in society. The earliest years, especially those spent in school, shape the minds of children for the rest of their lives. To better aid future generations, it is essential to provide effective education, and studying the performance of students provides an insight on how that can be done. Our research topic analyzed a student performance data set in which students of two Portuguese high schools were surveyed for their academic performance along with other demographics. This data collection contained a number of characteristics, such as the student's school, sex, age, type of neighborhood, family size, parental education, alcohol consumption, internet access, and more. The variety of variables made this data set interesting because some may have appeared unrelated but may have had the potential to influence a student's academic progress. We recognize the significance of early school years in regards to both higher education and future prospects as students who have gone through secondary education. It is important to study this data set to have a better understanding of factors that may improve a student’s academic success, especially if it were to be used to implement changes in the education system. Some factors we looked into were the effects of alcohol consumption on final grades, internet access, parental education and the impacts these had on student’s desire for pursuing higher education. We compared factors to find ones that lead to better or worse academic success. Although this survey was conducted in Portuguese schools, it is relevant for larger research on academic achievement since education is crucial for the development of young adults from all backgrounds.

Description of Data Set

Our research focused on data collected from two Portuguese public schools in the Alentejo region of Portugal between the 2005 and 2006 school year. The data was provided by school reports and questionnaires sent out by Paul Cortez and other researchers. They prepared a questionnaire of 37 questions and were able to collect data from 788 students, though they had to discard 111 results because of identification issues. The data we analyzed was merged from two data sets: Mathematics (395 values) and Portuguese (649 values) test scores. The merged data helped us to better identify common variables that would affect students’ scores in general. Our Y variables were the numeric student scores in Mathematics and Portuguese (score of 0 being the lowest and 20 being the highest) over three periods, and the X variables were demographical, social, emotional, and school related factors. Some specific examples include: student's school, sex, age, type of neighborhood, family size, parental education, alcohol consumption, internet access, and more. For binary qualitative variables (except school, sex, type of neighborhood, family size, and parental cohabitation status), we assigned them to either ‘True’ or ‘False’. Otherwise, they were nominal qualitative variables with responses that corresponded with the variable (parental occupation, reason to choose school, etc). For quantitative variables, we gave ranges for numeric values that varied depending on the type of variable. In total, the data comprised 382 rows (observations) and 53 columns (variables). A breakdown of all the variables we analyze in our paper is listed in the table below (copied from the source):

| Attribute | Description |
| --- | --- |
| Medu | mother’s education (numeric: from 0 to 4^a ) |
| Fedu | father’s education (numeric: from 0 to 4^a ) |
| Internet | internet access at home (binary: yes or no) |
| Walc | weekend alcohol consumption (numeric: from 1 – very low to 5 – very high) |
| Dalc | workday alcohol consumption (numeric: from 1 – very low to 5 – very high) |
| G3 | final grade (numeric: from 0 to 20) |

In respect to data cleaning, we found no missing data in any of our variables upon checking for missing variables in R studio. Additionally, there were no duplicates or errors in labeling as the data was previously used for regression as well. We did encounter some outliers in variables such as G1 and G2. We decided against removing them because they fall in our defined range and are important values to be analyzed for our research. These outliers provide insight on factors that could cause extreme academic performances – overachieving and underachieving.

Hypotheses

To formulate our research, we brainstormed hypotheses to understand how different factors affect each other. We looked at criterias such as alcohol consumption, internet use, time spent studying, parental education levels, and final grade. Concerning what effect alcohol consumption has on a students final grade, our null hypothesis is that alcohol consumption shows no effect on students final grade. Alcohol consumption was self-reported on a scale of 1 (very low) to 5 (very high) alcohol consumption. Weekday alcohol consumption was measured separately to weekend alcohol consumption, so in our analyses we will examine a correlation between final grades with weekend alcohol and weekday alcohol, respectively. Final grades are measured on a scale of 0 - 20. Our alternative hypothesis is that students who consumed more alcohol received lower final grades, shown by a negative correlation between alcohol consumed and test scores.

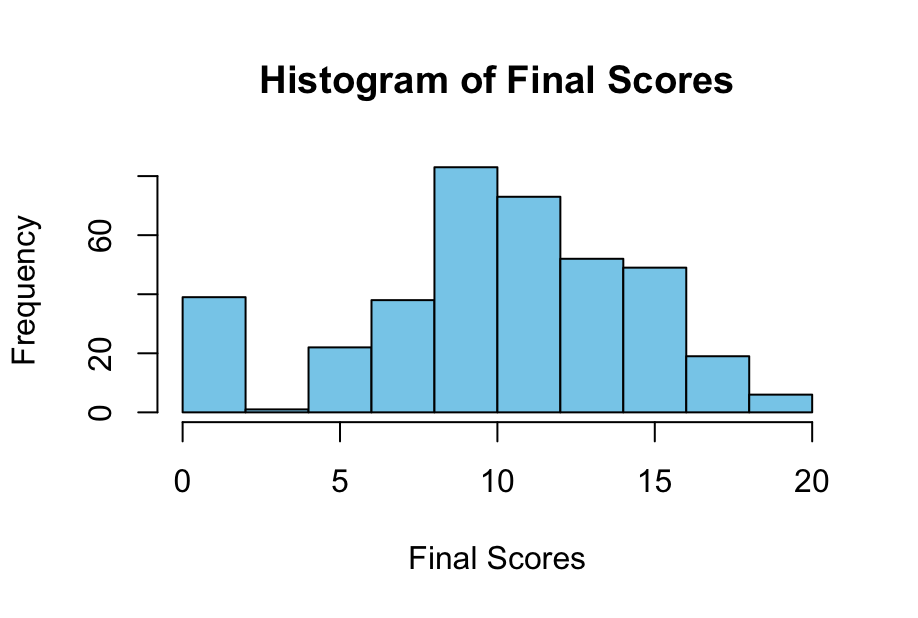
Another question we will explore is whether internet access affects the amount of time a student studies. This metric was measured binarily in the population, and students are recorded as with or without internet access. Therefore we will compare the mean time spent studying for these two groups. Our null hypothesis is that internet access showed no effect on the amount of time a student spends studying. Our alternative hypothesis is that the population with internet access will have a higher study time than students without internet access.

Thirdly, we will investigate whether there is correlation between parental education level, and students desire to pursue higher education. Our null hypothesis is that parental education level shows no effect on students' desire to take higher education. In this dataset parental higher education is measured on a scale of 1- 4, corresponding with primary, secondary, higher education, and professional degree. We hypothesize that higher parental education will be correlated with students wishing to pursue higher education, shown by a high positive correlation between parental education and the proportion of students who desire to pursue higher education.

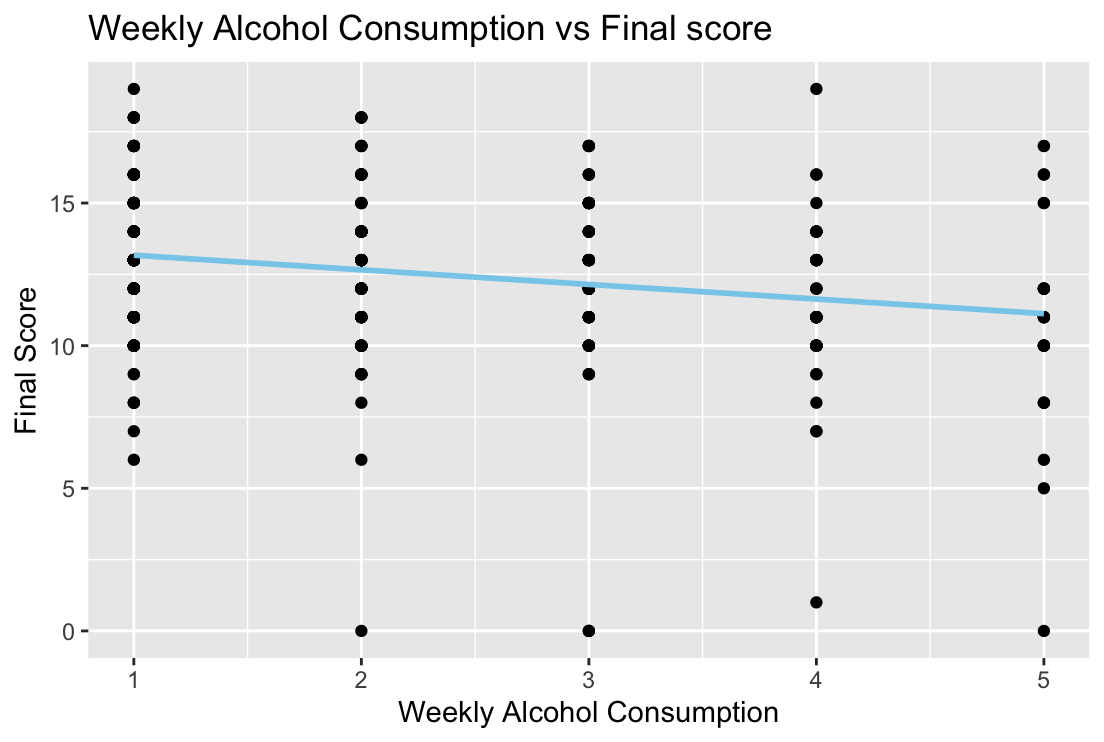
Lastly, we will explore whether students aspiring to pursue higher education typically perform better in school. The desire to pursue higher education was measured binarily, with students categorized into two populations: those who wish to pursue higher education and those who do not. Our null hypothesis is that these two populations show the same performance in school. Our alternative hypothesis is that students who aspire to higher education show higher performance in school.

Data Visualization

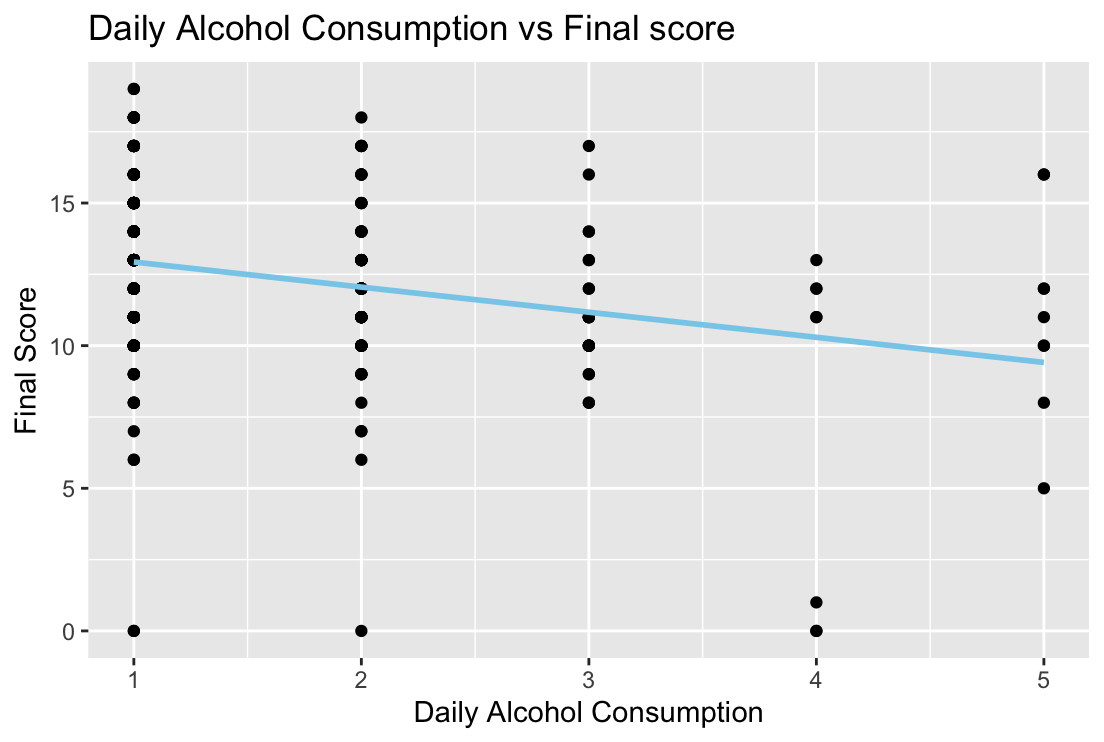
In order to better visualize the different factors we compared in our hypotheses, we utilized a variety of graphs using R Studio. Since these graphs are comparing two variables at a time, it could help us determine the correlations between factors that may impact academic success.



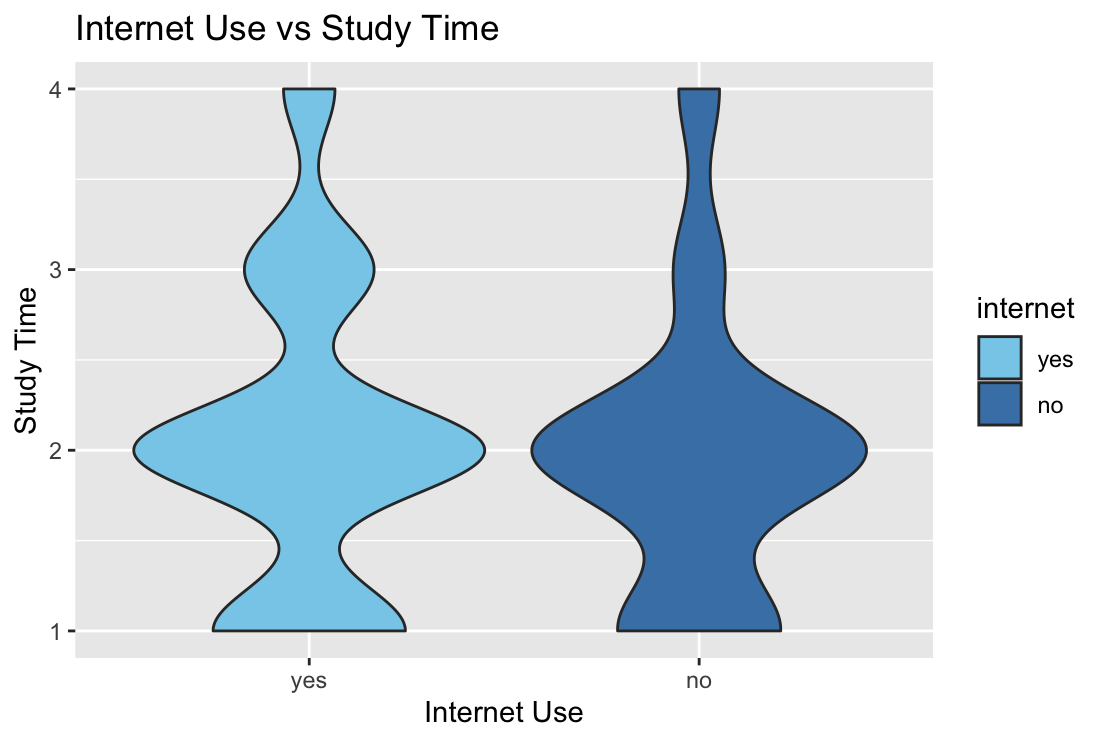
We used a histogram to visualize the frequency of students’ final scores, with the x axis being the scores and y axis being the frequency. This demonstrated the distribution of final grades across all of the students surveyed which acted as a reference to compare to graphs of other factors. Moreover, this shows that the final scores are normally distributed.



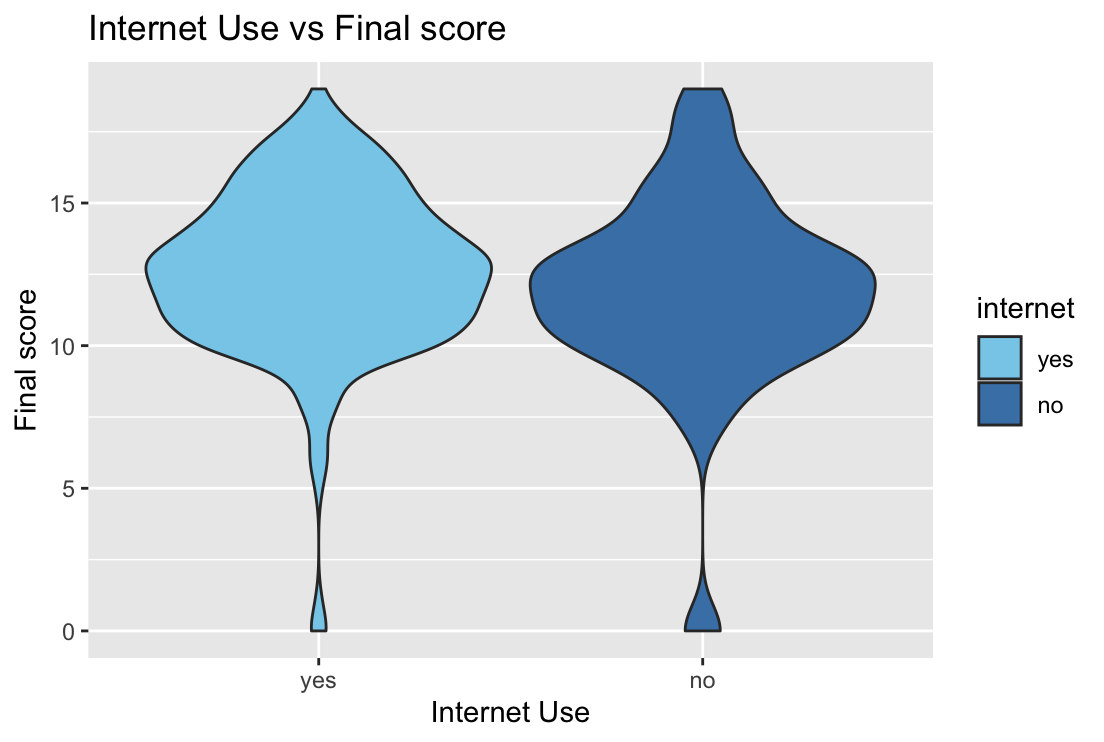
This scatter plot was made based on our alternative hypothesis that higher weekly alcohol consumption correlates with lower final grades, with the null hypothesis being there is no correlation between those two variables. Students’ weekly alcohol consumption was measured from 1 to 5, 1 being low consumption and 5 being high consumption, as shown on the x axis. The students’ corresponding final grades are on the y axis. Based on the linear regression line, it appeared that there is a negative correlation between weekly alcohol consumption and final scores.



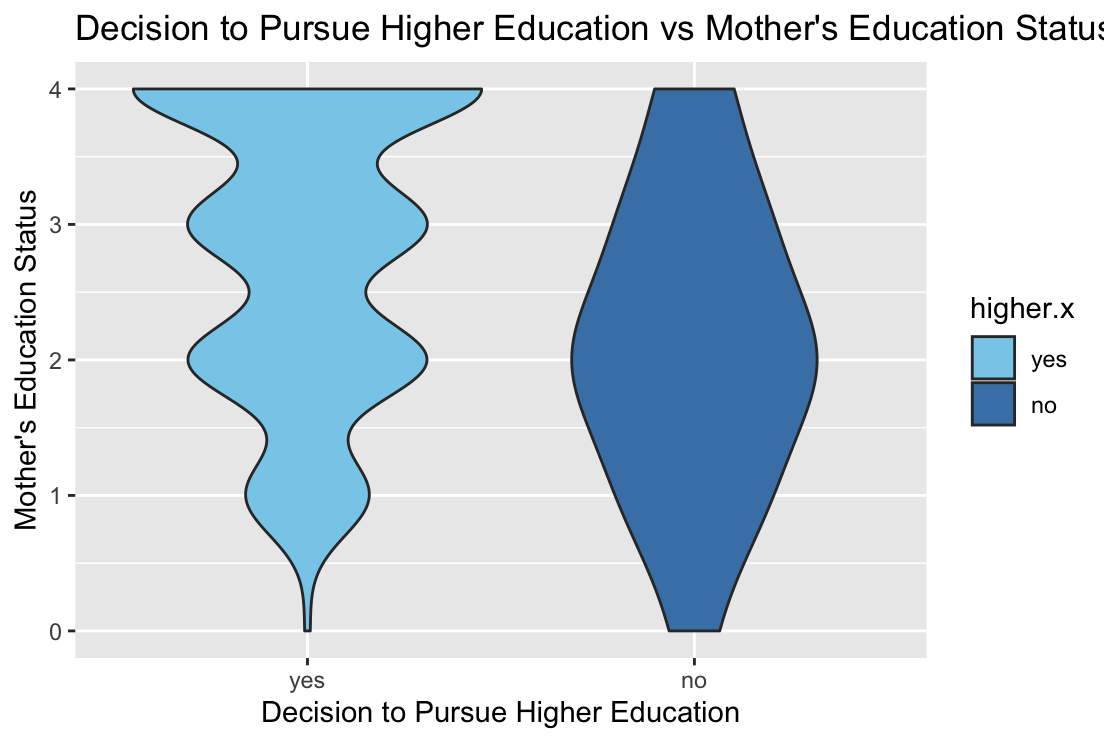
To better determine the correlation between alcohol consumption and final grades, we also created a scatter plot for daily alcohol consumption using the same null and alternative hypothesis. Compared to the last scatter plot, this one shows a stronger correlation with its downward slope, meaning there is evidence that higher daily alcohol consumption corresponds to lower final grade.

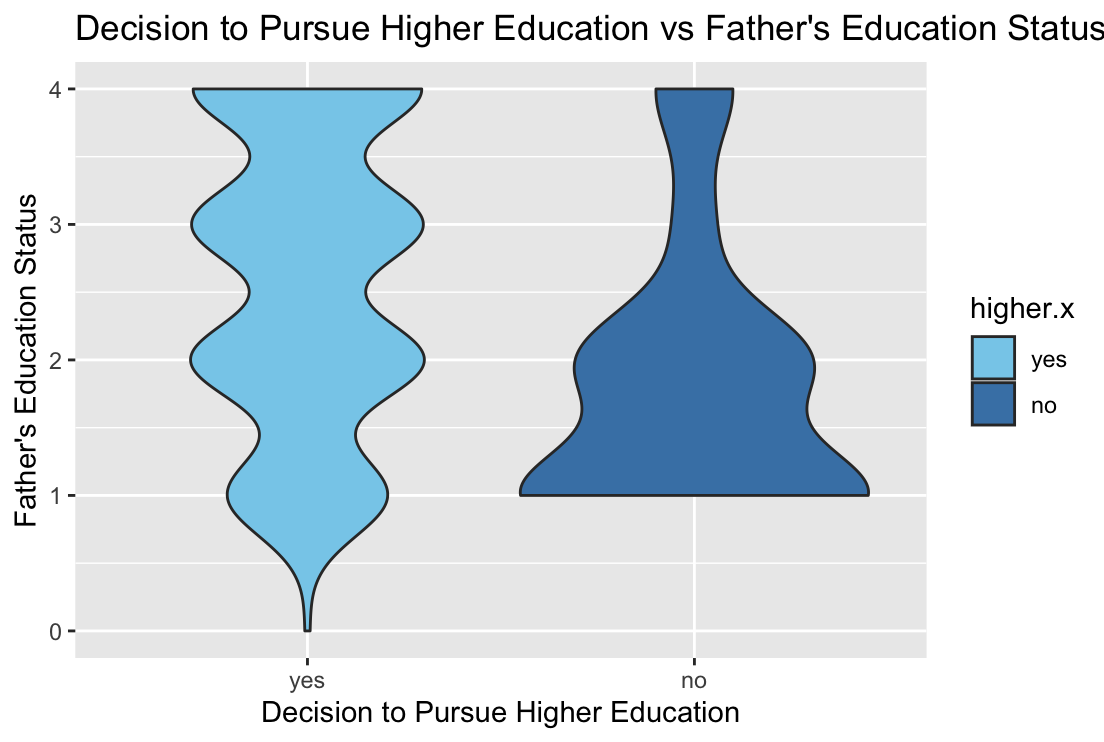


For visualizing the correlation between students’ internet access and their time spent on studying, we used a violin plot as its width is useful in showing density of a variable. Our null hypothesis was that there is no correlation between internet access and students’ study time which may be rejected because the violin plot demonstrated a skewed distribution favoring our alternative hypothesis. Since the width of the violin plot was greater at the higher study times for students’ with internet access, there was evidence that internet access positively influences students’ study time.

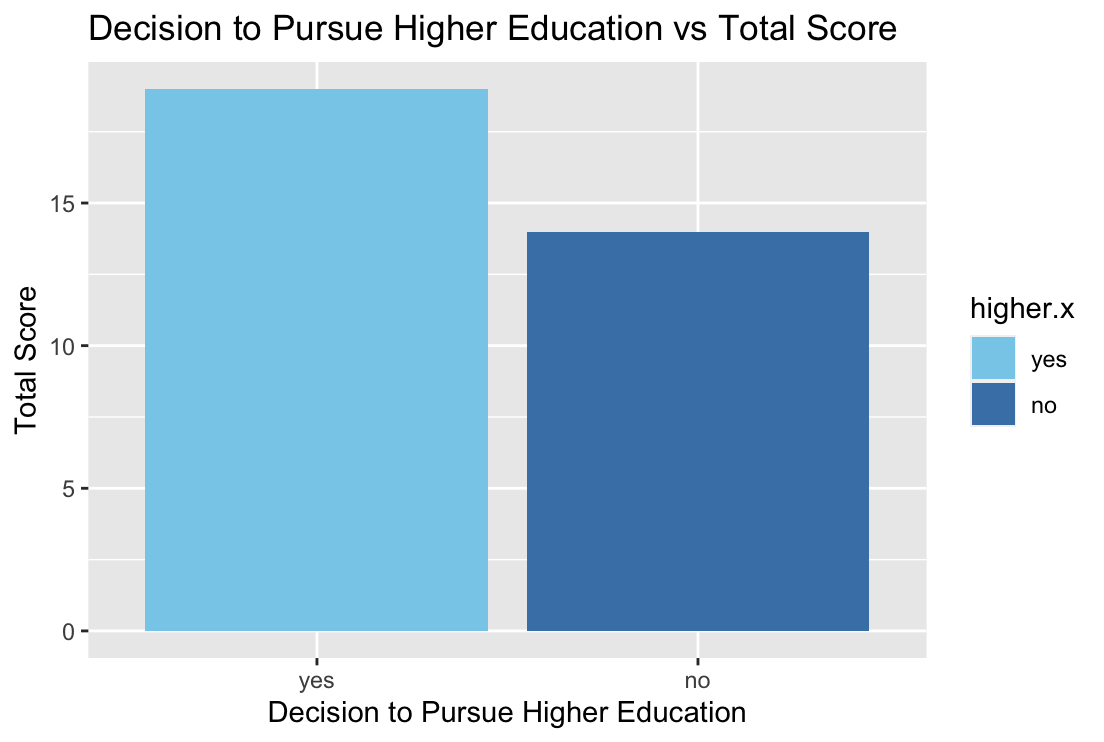


Because our research was conducted to evaluate factors that influence student grades, we also created a violin plot to visualize the correlation between internet access and students’ final grades. We found that the width of the violin plot was slightly greater for students who did have internet access. The width in this case showed the students’ that received higher final grades, which meant there was a positive correlation between students’ internet access and their final grades.





We utilized a violin plot because its width is good for displaying the density of a variable to visualize the relationship between students' desire to seek further education and their parents' educational level. Since the results for mothers’ and fathers’ education level had different correlations with students’ desire to pursue higher education, we created a graph for each. The violin plot for fathers’ education showed that there was a correlation between parental education and students’ desire for higher education due to the width of the violin graph, while there was no significant correlation for the mothers’ education level.



For our final hypothesis regarding desire for higher education and final grades, we used bar graphs. The graph on the left showed the final scores of students who wanted to pursue higher education while the graph on the right represented those who did not want to. Since the bar on the left was taller than the right one, there was evidence to support the alternative hypothesis that students who desired higher education also scored higher.

Hypothesis Testing

To find out whether or not we reject our null hypotheses for our four hypotheses, we conducted hypothesis testing for each. Since all of our hypotheses involved two different variables, we used two-sample hypothesis tests. Before we began, we had to establish our alpha values and determine whether our samples were normally distributed. To do this, we referred back to the graphs showing the correlation of each hypothesis. Once we know the variables are approximately normally distributed, we can set our alpha to the standard 0.05.

For our hypotheses regarding the effect of alcohol consumption on final grades, we conducted a two-sample t-test to decide whether or not to reject the null hypothesis that students who consume a high amount of alcohol obtain the same grades as students who consume a low amount of alcohol. For alcohol consumption and final grades, we calculated the means of each and conducted a t-test code in R assuming the variances were equal. The code we used can be seen below. After we computed the results, we had a t-test statistic of -2.374, and our p-value was 0.01905 which is less than our alpha value of 0.05. Because of this, we reject our null hypothesis, meaning the grades of students who consume a high amount of alcohol are significantly different from the grades of students who consume a low amount of alcohol. From this hypothesis test, it suggested that alcohol consumption has an impact on students’ final grades.

alc\_high <- d3[d3$Walc.x == 4, ]$G3.y

mean(alc\_high)

var(alc\_high)

hist(alc\_high)

alc\_low <- d3[d3$Walc.x == 2, ]$G3.y

mean(alc\_low)

var(alc\_low)

hist(alc\_low)

t.test(alc\_high, alc\_low,

var.equal = TRUE)

t = -2.374, df = 131, p-value = 0.01905

In order to determine whether or not to reject the null hypothesis that students who have internet connection get the same amount of studying time as students who don't, we did a two-sample t-test (Welch Two Sample t-test) for our hypotheses regarding the influence of Internet access on time spent studying. We used the same method as the first hypothesis test. We got a t-test statistic of 1.7429. Our p-value was 0.08215 which is more than our alpha value of 0.05. Therefore we failed to reject the null hypothesis, meaning there is no correlation between internet access and students’ study time. In other words, Internet access showed no effect on the amount of time a student spends studying. From this hypothesis test, it suggested that access to the internet has no impact on the amount of time a student spends studying.

Our third hypothesis test was concerning how parental education level influences students desire to pursue higher education. This was measured separately for mothers and fathers education levels. We selected an alpha value of 0.05 for both mother and father correlations. Computing the results for fathers correlation, the t-test statistic was 3.5859, and our p-value was 0.0004361, which is lower than our alpha value of 0.05. This led us to reject the null hypothesis and instead conclude that there is a correlation between father education and the child's desire to pursue higher education. After computing the same test for mothers, our results were that our p-value was 0.07534 which is greater than our alpha value of 0.05. Therefore we reject the null hypothesis, meaning that there is not a significant correlation between mothers education levels and a child's desire to pursue higher education.

Our final hypothesis tested whether final grades were correlated with desire to pursue higher education. To determine whether there was correlation, we conducted a two-sample t-test (Welch two sample t-test) comparing the mean final grades of those who said ‘yes’ to desiring to pursue higher education and those who said ‘no’ to pursuing higher education. This testing determined whether those desiring to pursue higher education had different final grades than those who did not intend to pursue higher education. We chose to use an alpha value of 0.05 because this data appeared to be normally distributed. After conducting this testing, we got a t-test statistic of 6.2109, and our p-value was 1.383e-09, which is lower than our alpha value of 0.05. Therefore, we were able to reject the null hypothesis as we have evidence to suggest there is a correlation between final grades and students desire to pursue higher education. Our findings suggest that students' final grades do have an impact on desire to pursue higher education.

Conclusion

Based on our hypothesis testing, we found that higher alcohol consumption was correlated with lower final grades, supporting our alternative hypothesis. So students who drank less alcohol had more success in school. We found that internet access had no significant impact on how much time students spend studying. This is important to understand whether or not internet access is a limiting factor for students' ability to study. Are students without the internet disadvantaged in school? Our analysis implies that internet access is not a limiting factor for students' academics, at least in terms of study time. Our next hypothesis tested whether parental education level was a predictor of students intentions to pursue higher education. Only fathers’ level of education had a significant impact on students’ desire to pursue higher education. We found that students who did not intend to continue in school, their father most likely did not continue past secondary education either. This could be the result of various factors, for example families without higher degrees may be in lower paying jobs and unable to pay for their child's higher education, in addition to family culture and expectations. An observation we made was there was not a significant correlation in the case of mothers' education levels and student’s decision to pursue higher education, perhaps due to cultural or socioeconomic factors which can be tied to patriarchal practices in Portugal. This is an area for further research comparing different countries with different cultures. Finally, supporting our alternative hypothesis, we found evidence that higher final grades were correlated with desire to pursue higher education. This could be explained in different ways, either students who succeed in high school are more likely to desire higher education or students planning on higher education are more motivated to succeed in highschool, perhaps to get into a good college. This finding could be a result of either or a combination of these factors.

Citations

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Total Contributions:

Angelina: drafted and edited introduction, brainstormed questions, edited description, brainstormed hypotheses, data visualization written descriptions (graphs 1 - 7), hypothesis testing paragraph (hypothesis 1), intro for data visualization and hypothesis testing, rewrote graph descriptions, final editing

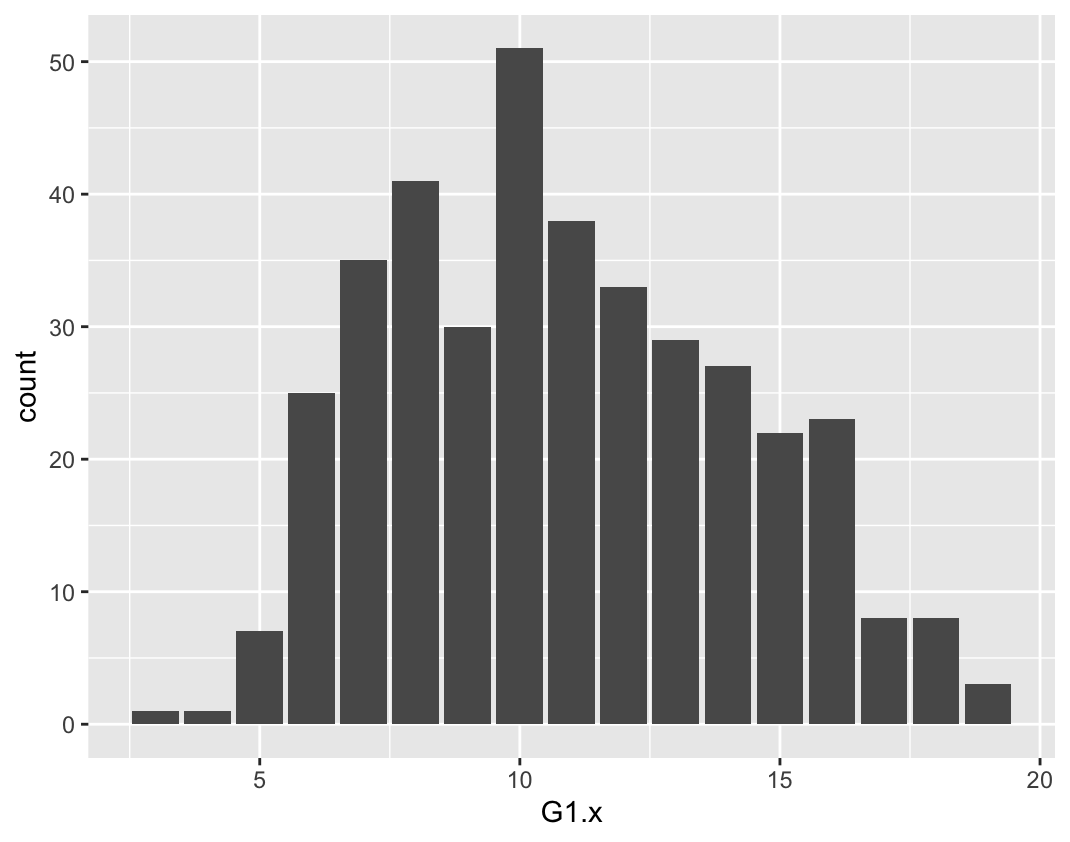
Yuqi: edited introduction, brainstormed questions, edited description, edited hypothesis, data visualization written descriptions (graphs 8 -15), hypothesis testing paragraph (hypothesis 2), final editing

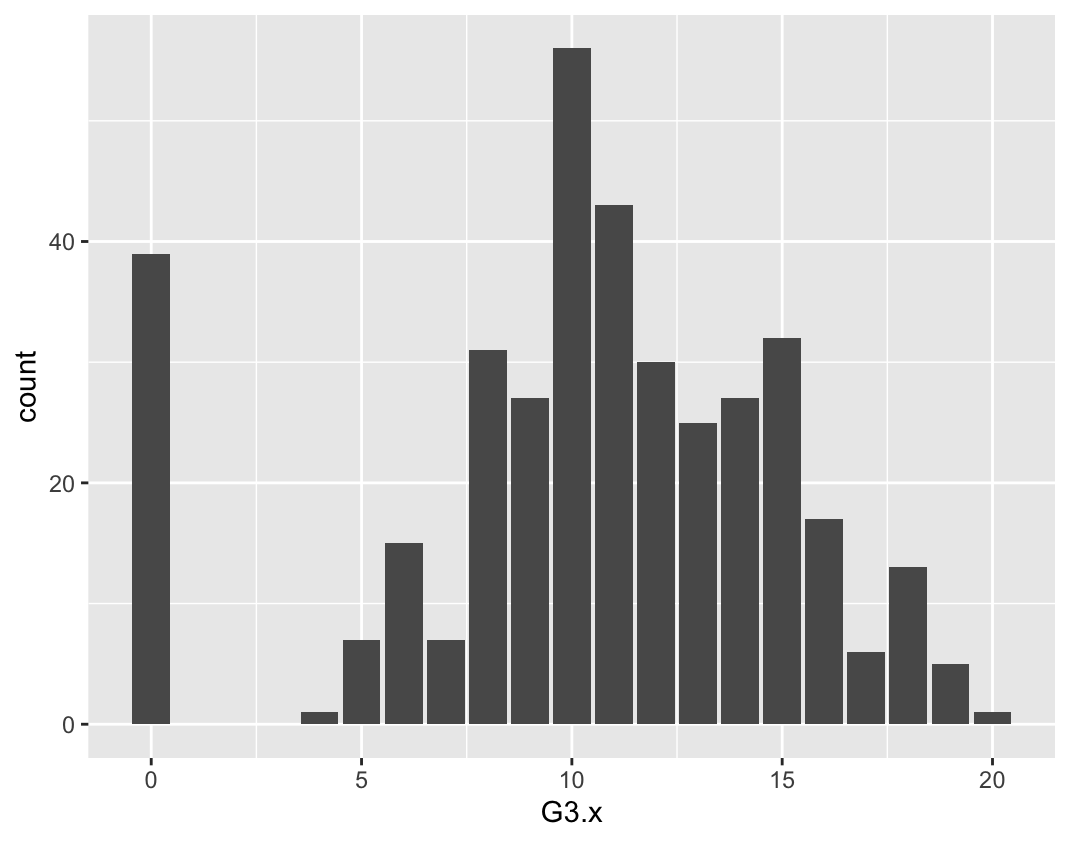
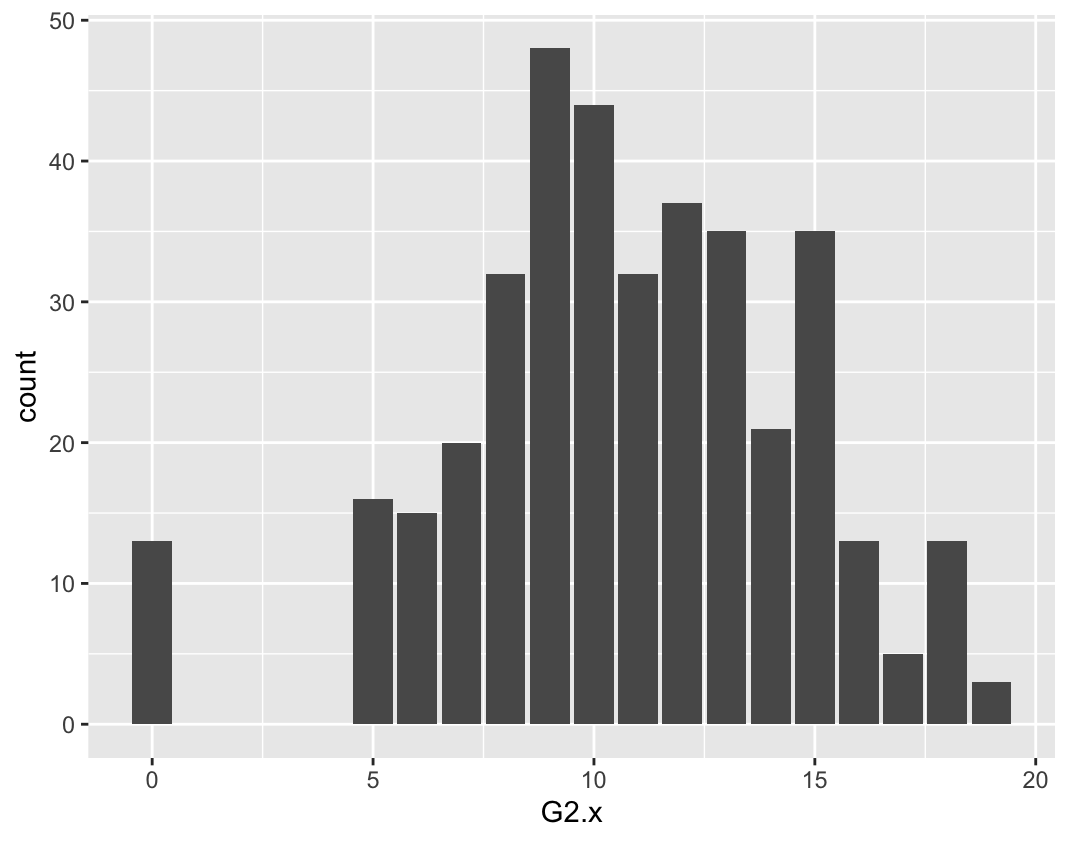
Ananya: brainstormed questions, drafted description, performed data cleaning, data visualization (coding and interpretation), R coding and editing for hypothesis testing, final editing

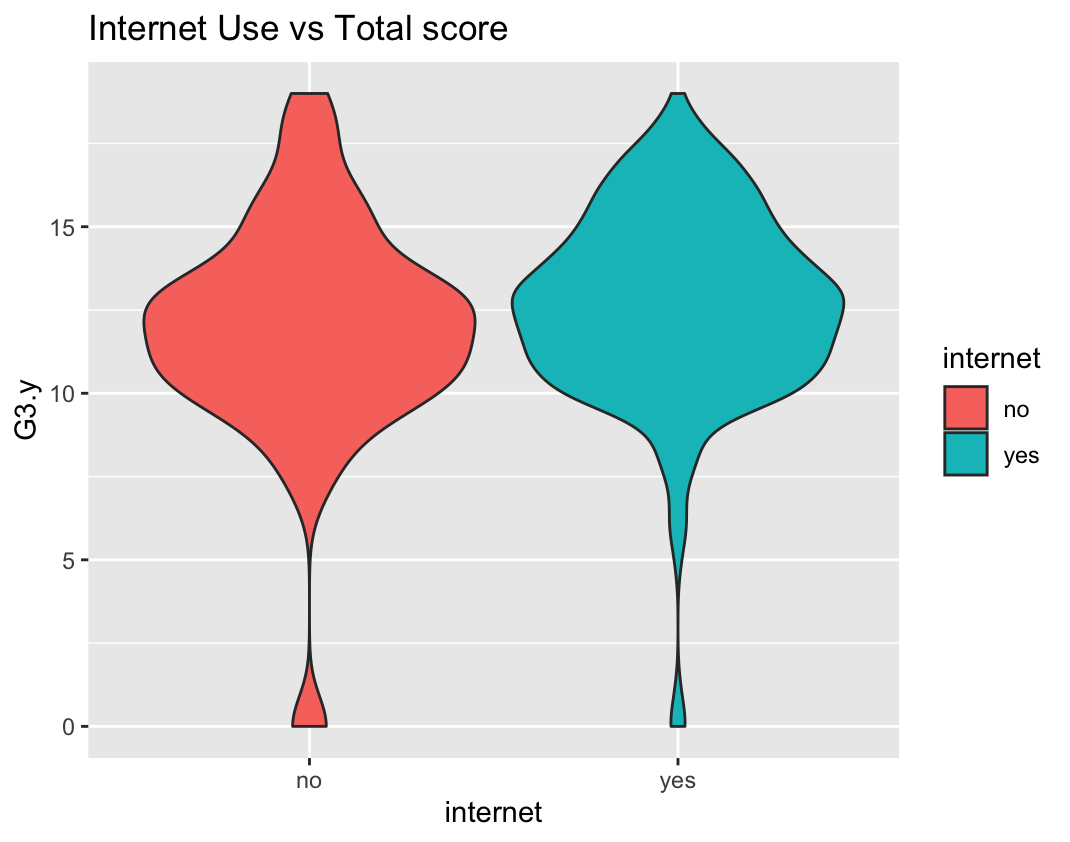
Jessica: Brainstormed hypotheses, drafted and edited hypothesis, hypothesis testing paragraphs (hypothesis 3 and 4) and hypothesis testing editing, wrote conclusion, final editing

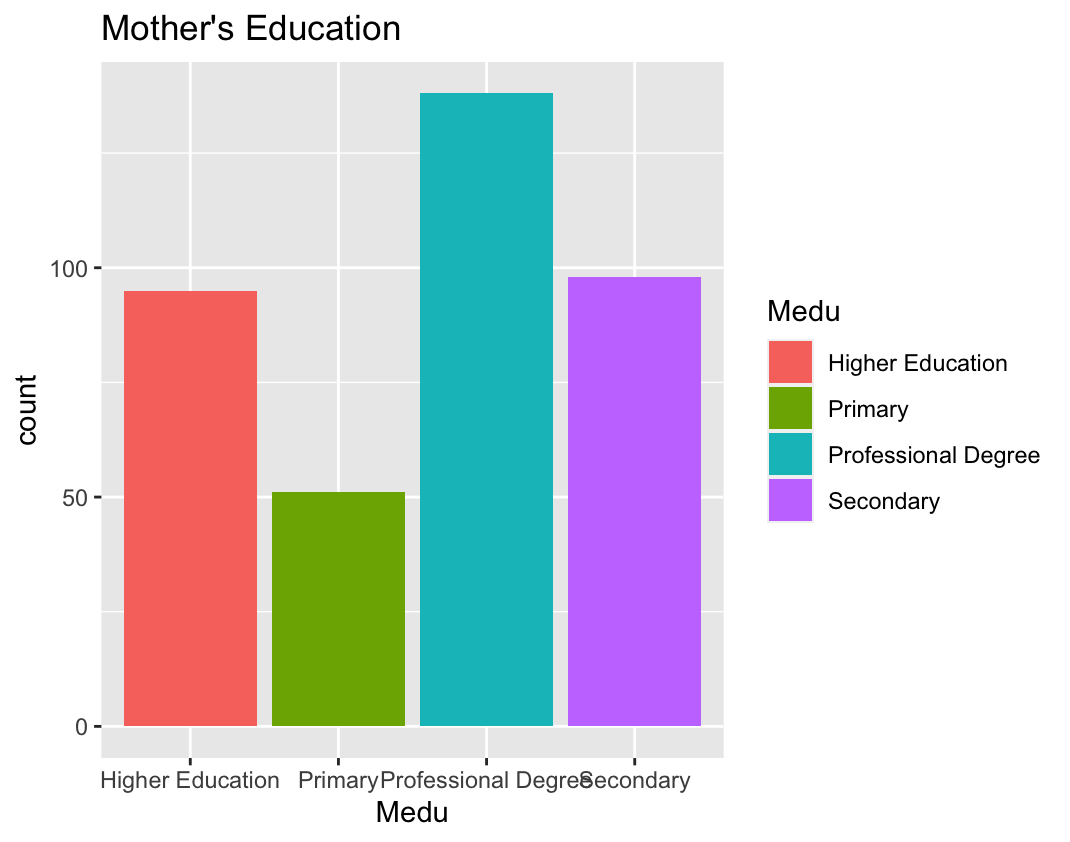
**Appendix**

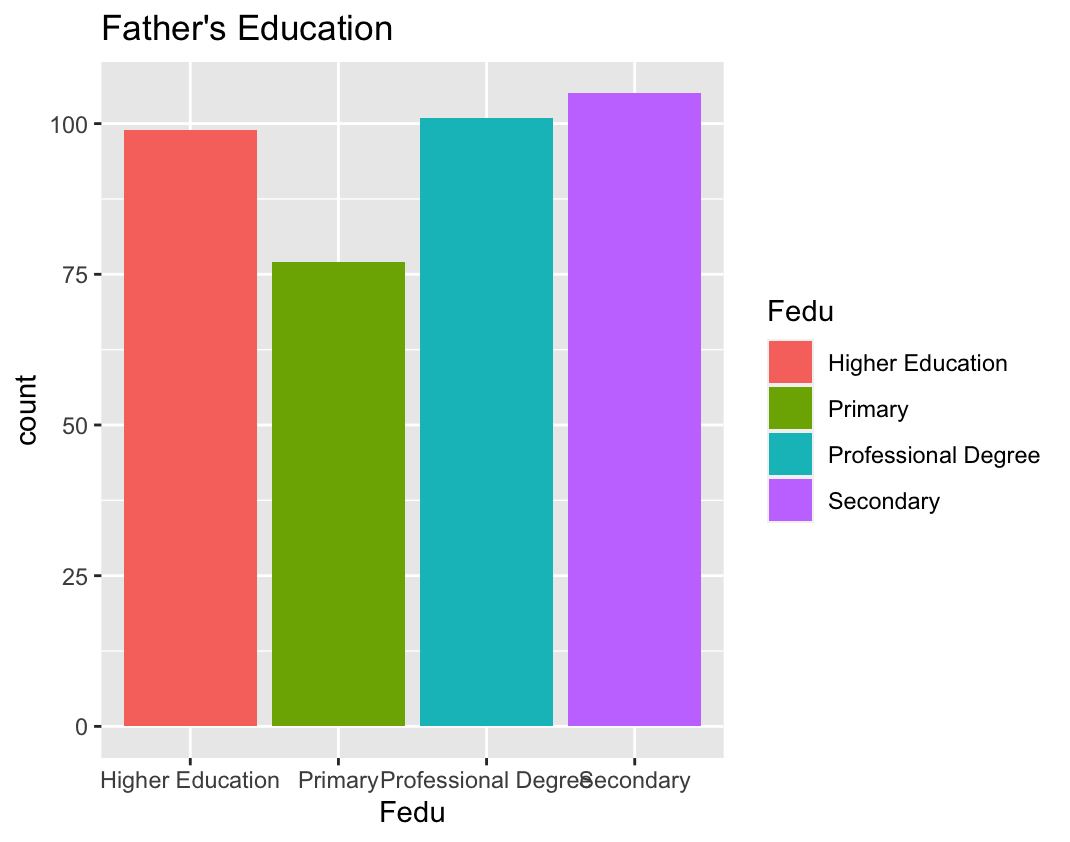
| **Attribute** | **Description (Domain)** |
| --- | --- |
| sex | student’s sex (binary: female or male) |
| age | student’s age (numeric: from 15 to 22) |
| school | student’s school (binary: Gabriel Pereira or Mousinho da Silveira) |
| address | student’s home address type (binary: urban or rural) |
| Pstatus | parent’s cohabitation status (binary: living together or apart) |
| Medu | mother’s education (numeric: from 0 to 4^a ) |
| Mjob | mother’s job (nominal^b ) |
| Fedu | father’s education (numeric: from 0 to 4^a ) |
| Fjob | father’s job (nominal^b ) |
| guardian | student’s guardian (nominal: mother, father or other) |
| famsize | family size (binary: ≤ 3 or > 3) |
| famrel | quality of family relationships (numeric: from 1 – very bad to 5 – excellent) |
| reason | reason to choose this school (nominal: close to home, school reputation, course preference or other) |
| traveltime | home to school travel time (numeric: 1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour). |
| studytime | weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours) |
| failures | number of past class failures (numeric: n if 1 ≤ n < 3, else 4) |
| schoolsup | extra educational school support (binary: yes or no) |
| famsup | family educational support (binary: yes or no) |
| activities | extra-curricular activities (binary: yes or no) |
| paidclass | extra paid classes (binary: yes or no) |
| internet | Internet access at home (binary: yes or no) |
| nursery | attended nursery school (binary: yes or no) |
| higher | wants to take higher education (binary: yes or no) |
| romantic | with a romantic relationship (binary: yes or no) |
| freetime | free time after school (numeric: from 1 – very low to 5 – very high) |
| goout | going out with friends (numeric: from 1 – very low to 5 – very high) |
| Walc | weekend alcohol consumption (numeric: from 1 – very low to 5 – very high) |
| Dalc | workday alcohol consumption (numeric: from 1 – very low to 5 – very high) |
| health | current health status (numeric: from 1 – very bad to 5 – very good) |
| absences | number of school absences (numeric: from 0 to 93) |
| G1 | first period grade (numeric: from 0 to 20) |
| G2 | second period grade (numeric: from 0 to 20) |
| G3 | final grade (numeric: from 0 to 20) |
| a 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.  b teacher, health care related, civil services (e.g. administrative or police), at home or other. | |

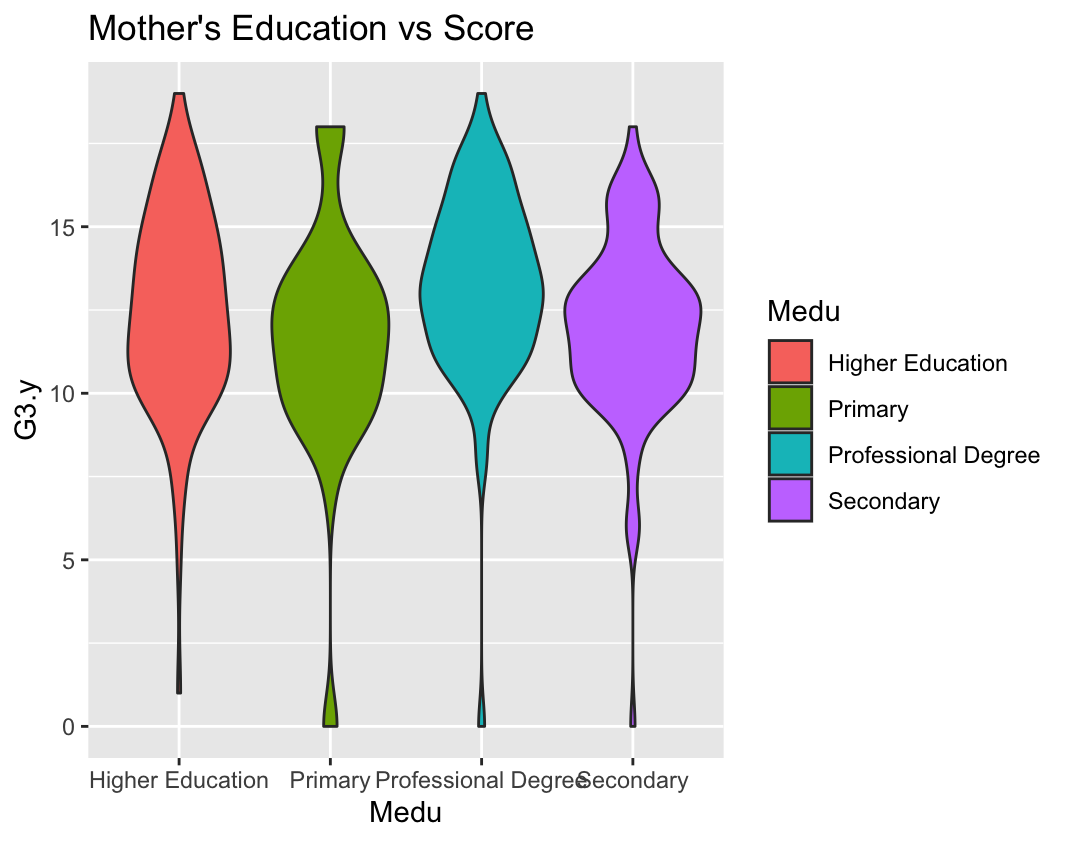


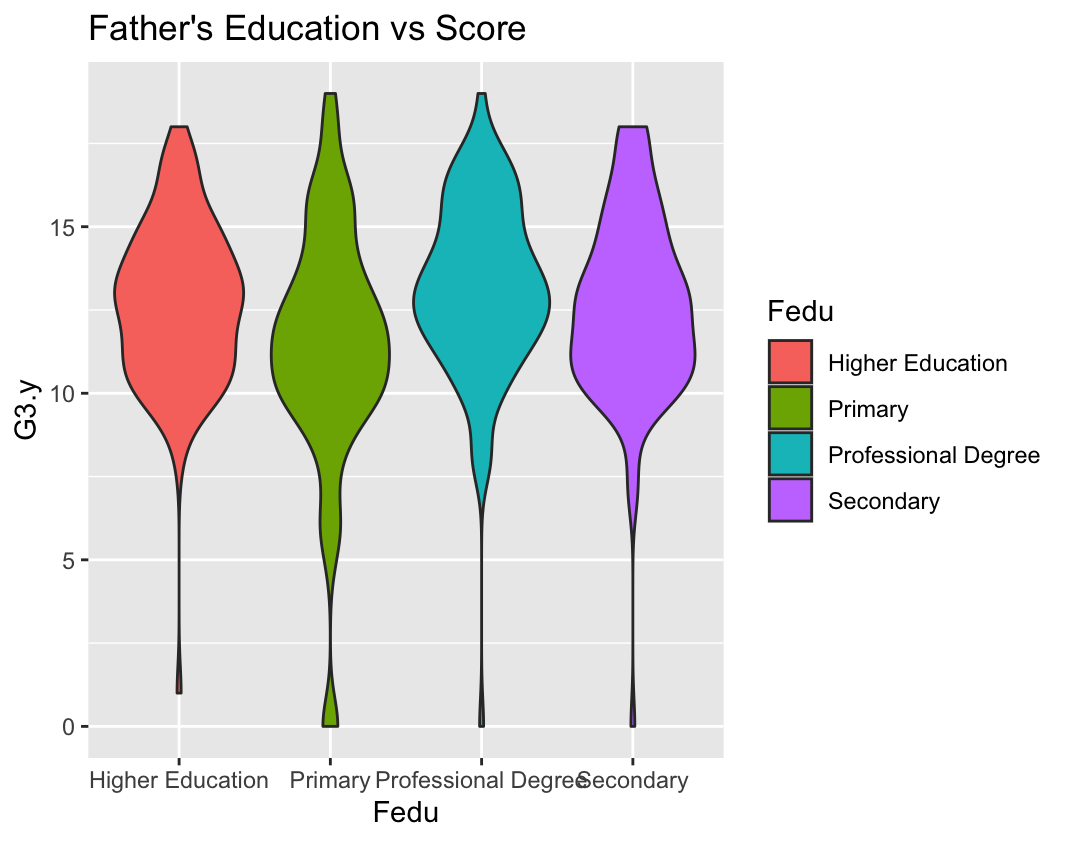


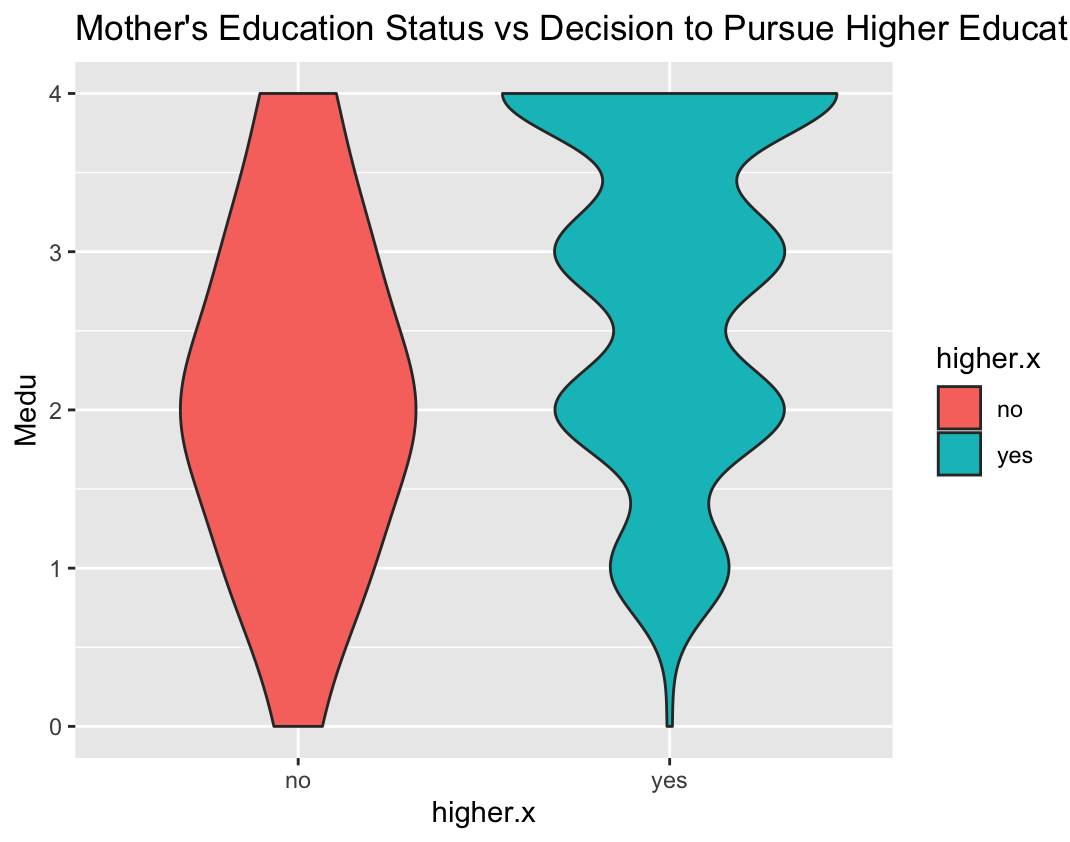












R code

d1=read.table("~/Downloads/student-mat.csv",sep=";",header=TRUE)

d2=read.table("~/Downloads/student-por.csv",sep=";",header=TRUE)

d3=merge(d1,d2,by=c("school","sex","age","address","famsize","Pstatus","Medu","Fedu","Mjob","Fjob","reason","nursery","internet"))

print(nrow(d3)) # 382 students

hist(d3$G3.x, col="skyblue", xlab = "Final Scores", main = "Histogram of Final Scores")

install.packages("ggplot2")

data(d3)

library(ggplot2)

ScatterPlot <- ggplot(data = d3, aes(Dalc.x,G3.y)) +

geom\_point() +

geom\_smooth(method = "lm", se = FALSE, color = "skyblue") +

ggtitle("Daily Alcohol Consumption vs Final score") +

labs(x = "Daily Alcohol Consumption", y = "Final Score")

print(ScatterPlot)

ScatterPlot1<- ggplot(data = d3, aes(Walc.x,G3.y)) +

geom\_point() +

geom\_smooth(method = "lm", se = FALSE, color = "skyblue") +

ggtitle("Weekly Alcohol Consumption vs Final score") +

labs(x = "Weekly Alcohol Consumption", y = "Final Score")

print(ScatterPlot2)

ViolinPlot <- ggplot(data = d3, aes(x = internet, y = G3.y, fill = internet)) +

geom\_violin() +

scale\_fill\_manual(values = c("skyblue", "steelblue")) +

ggtitle("Internet Use vs Final score") +

xlab("Internet Use") +

ylab("Final score")

print(ViolinPlot)

d3$internet <- factor(d3$internet, levels = c("yes", "no"))

ViolinPlot1 <- ggplot(data = d3, aes(x = internet, y = studytime.y , fill = internet)) +

geom\_violin() +

scale\_fill\_manual(values = c("skyblue", "steelblue")) +

ggtitle("Internet Use vs Study Time") +

xlab("Internet Use") +

ylab("Study Time")

print(ViolinPlot1)

mean\_ts <- mean(d3$studytime.x)

ViolinPlot <- ggplot(data = d3, aes(higher.x, y = Fedu, fill=higher.x)) +

geom\_violin() +

ggtitle("Father's Education Status vs Decision to Pursue Higher Education")

print(ViolinPlot)

ScatterPlot <- ggplot(data = d3, aes(Medu,higher.y)) + geom\_point() +

geom\_smooth(method = "lm", se = FALSE) + ggtitle("Internet Use vs Total score")

print(ScatterPlot)

ScatterPlot <- ggplot(data = d3, aes(x=log(Medu),y=log(G3.y))) + geom\_point() +

geom\_smooth(method = "lm", se = FALSE) + ggtitle("log(Weekly Alchol) vs log(Total score)")

print(ScatterPlot)

mean\_score1 <- subset(d3, higher.y == "yes")$G3.y

print(mean\_score1)

mean\_score2 <- subset(d3, higher.y == "no")$G3.y

ScatterPlot <- ggplot(data = d3, aes(mean\_score1, mean\_score2)) + geom\_point() +

geom\_smooth(method = "lm", se = FALSE) + ggtitle("Internet Use vs Total score")

print(ScatterPlot)

StackedBar <- ggplot(d3,aes(G3.x,fill=mean\_score1))+geom\_bar(position="fill") +

ggtitle("Percentage of colors based on cut") + ylab("Percentage")

print(StackedBar)

StackedBar <- ggplot(d3, aes(x =G3.x, fill = higher.y)) +

geom\_bar(position = "fill") +

ggtitle("Percentage of values based on G3.x and higher.y") +

ylab("Percentage")

print(StackedBar)

ScatterPlot <- ggplot(data = d3, aes(higher.x, G3.y)) + geom\_point() +

geom\_smooth(method = "lm", se = FALSE) + ggtitle("Decision to pursue higher education vs Total score")

print(ScatterPlot)

bb <- ggplot(data = d3, aes(x = higher.x, y = G3.y, fill = higher.x)) +

geom\_bar(stat = "identity", position = "dodge") +

scale\_fill\_manual(values = c("skyblue", "steelblue")) +

ggtitle("Decision to Pursue Higher Education vs Total Score") +

xlab("Decision to Pursue Higher Education") +

ylab("Total Score")

print(bb)

d3$

PresidentsBarPlot <- ggplot(data = d3, aes(x = G3.x, fill = higher.y)) +

geom\_bar() +

scale\_fill\_manual(values = c("skyblue", "steelblue")) +

ggtitle("Decision to Pursue Higher Education vs Final Score") +

xlab("Decision to Pursue Higher Education") +

ylab("Final Score")

print(PresidentsBarPlot)

d3$higher.x <- factor(d3$higher.x, levels = c("yes", "no"))

print(d3$Medu)

ViolinPlot1 <- ggplot(data = d3, aes(x = higher.x, y = Fedu, fill = higher.x)) +

geom\_violin() +

scale\_fill\_manual(values = c("skyblue", "steelblue")) +

ggtitle("Decision to Pursue Higher Education vs Father's Education Status") +

xlab("Decision to Pursue Higher Education") +

ylab("Father's Education Status")

print(ViolinPlot1)

d3$Medu <- ifelse(d3$Medu == 1, "Primary",

ifelse(d3$Medu == 2, "Secondary",

ifelse(d3$Medu == 3, "Higher Education", "Professional Degree")))

d3$Medu

d3$Fedu <- ifelse(d3$Fedu == 1, "Primary",

ifelse(d3$Fedu == 2, "Secondary",

ifelse(d3$Fedu == 3, "Higher Education", "Professional Degree")))

d3$Fedu

ViolinPlot <- ggplot(data = d3, aes(Medu, y = higher.y, fill=higher.x)) +

geom\_violin() +

ggtitle("Mother's Education Status vs Decision to Pursue Higher Education")

print(ViolinPlot)

BarPlot <- ggplot(data = d3, aes(x = Fedu,, fill = Fedu)) +

geom\_bar(stat = "count") +

ggtitle("Father's Education")

print(BarPlot)

ViolinPlot <- ggplot(data = d3, aes(x=Fedu, y = G3.y, fill=Fedu)) +

geom\_violin() +

ggtitle("Father's Education vs Score")

print(ViolinPlot)

**#1: Alcohol consumption and final grade**

#Hypothesis Testing

alc\_high <- d3[d3$Walc.x == 4, ]$G3.y

mean(alc\_high)

var(alc\_high)

hist(alc\_high)

alc\_low <- d3[d3$Walc.x == 2, ]$G3.y

mean(alc\_low)

var(alc\_low)

hist(alc\_low)

* t.test(alc\_high, alc\_low,

var.equal = TRUE)

**Result:**

**Two Sample t-test (use this)**

data: alc\_high and alc\_low

t = -2.374, df = 131, **p-value = 0.01905**

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-2.2323108 -0.2029833

sample estimates:

mean of x mean of y

11.50000 12.71765

* t.test(alc\_high,

alc\_low)

Welch Two Sample t-test

data: alc\_high and alc\_low

t = -2.4008, df = 100.97, p-value = 0.01819

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-2.2237465 -0.2115476

sample estimates:

mean of x mean of y

11.50000 12.71765

**#2: Internet access and time spent studying**

Code:

int\_yes <- d3[d3$internet == 'yes', ]$G3.y

mean(int\_yes)

var(int\_yes)

hist(int\_yes)

int\_no <- d3[d3$internet == 'no', ]$G3.y

mean(int\_no)

var(int\_no)

hist(int\_no)

t.test(int\_yes, int\_no,

var.equal = TRUE)

t.test(int\_yes, int\_no)

**Result:**

**Two Sample t-test**

data: int\_yes and int\_no

t = 1.7429, df = 380, p-value = **0.08215**

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.09352114 1.55350411

sample estimates:

mean of x mean of y

12.62654 11.89655

**Welch Two Sample t-test**

data: int\_yes and int\_no

t = 1.5347, df = 71.83, p-value = 0.1292

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.2182326 1.6782155

sample estimates:

mean of x mean of y

12.62654 11.89655

p-value = **0.08215** > alpha: 0.05

So it fail to reject the null hypothesis. null hypothesis was that there is no correlation between internet access and students’ study time. In other words, Internet access showed no effect on the amount of time a student spends studying.

**#3: Parental education and students’ desire to pursue higher education**

Code: #Hypothesis Testing for Father's education and students’ desire to pursue higher education

fedu\_high <- d3[d3$Fedu == 4, ]$G3.y

mean(fedu\_high)

var(fedu\_high)

hist(fedu\_high)

fedu\_low <- d3[d3$Fedu == 1, ]$G3.y

mean(fedu\_low)

var(fedu\_low)

hist(fedu\_low)

t.test(fedu\_high,fedu\_low,

var.equal = TRUE)

t.test(fedu\_high,

fedu\_low)

Result:

Two Sample t-test

Two Sample t-test

data: fedu\_high and fedu\_low

t = 3.5859, df = 174, **p-value = 0.0004361**

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.7726941 2.6645353

sample estimates:

mean of x mean of y

13.12121 11.40260

Welch Two Sample t-test (use this)

data: fedu\_high and fedu\_low

t = 3.4499, df = 133.9, p-value = 0.0007504

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.7333258 2.7039036

sample estimates:

mean of x mean of y

13.12121 11.40260

#Mother’s edu

medu\_high <- d3[d3$Medu == 3, ]$G3.y

mean(medu\_high)

var(medu\_high)

hist(medu\_high)

medu\_low <- d3[d3$Medu == 1, ]$G3.y

mean(medu\_low)

var(medu\_low)

hist(medu\_low)

t.test(medu\_high,medu\_low,

var.equal = TRUE)

t.test(medu\_high,

medu\_low)

Result

Two Sample t-test

data: medu\_high and medu\_low

t = 1.7914, df = 144, **p-value = 0.07534**

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.1010519 2.0556443

sample estimates:

mean of x mean of y

12.52632 11.54902

Welch Two Sample t-test

data: medu\_high and medu\_low

t = 1.7047, df = 89.013, p-value = 0.09174

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.1618396 2.1164320

sample estimates:

mean of x mean of y

12.52632 11.54902

**#4: Desire to pursue higher education and final grade**

high\_yes <- d3[d3$higher.x == 'yes', ]$G3.y

mean(high\_yes)

var(high\_yes)

hist(high\_yes)

high\_no <- d3[d3$higher.x == 'no', ]$G3.y

mean(high\_no)

var(high\_no)

hist(high\_no)

t.test(high\_yes,high\_no,

var.equal = TRUE)

t.test(high\_yes,high\_no)

Results:

Two Sample t-test

data: high\_yes and high\_no

t = 6.2109, df = 380, **p-value = 1.383e-09**

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

2.880138 5.548434

sample estimates:

mean of x mean of y

12.71429 8.50000

Welch Two Sample t-test

data: high\_yes and high\_no

t = 5.7377, df = 18.44, p-value = 1.765e-05

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

2.673800 5.754771

sample estimates:

mean of x mean of y

12.71429 8.50000