Project Proposal

Hypothesis Heroes - Drew Davison, Lisa Zhang, Ellie Culman, Austin Chang

```r
library(tidyverse)
library(tidymodels)
# add other packages as needed

# add code to load data
```

Introduction

All around the world, underage drinking among students is a significant public health issue and takes a huge toll on the quality of students' lives and education. There is plenty of existing documentation on the effects of socioeconomic status on risky drinking behavior amongst college students. For example, a paper by Susan E. Collins, PhD, professor and licensed clinical psychologist, found that individuals with lower socioeconomic status as well as people of racial and ethnic minorities and homelessness experience greater alcohol-related consequences. Additionally, studies from the National Institutes of Health have found that "aspects of college life—such as unstructured time, widespread availability of alcohol, inconsistent enforcement of underage drinking laws, and limited interactions with parents and other adults" drives up rates of underage drinking, and "college students have higher binge-drinking rates and a higher incidence of driving under the influence of alcohol than their noncollege peers." Since most existing literature on underage student drinking focuses on college students, we wanted to examine the factors contributing to drinking amongst secondary school students. Our research question is as follows:

How do social indicators affect student alcohol consumption in secondary schools?

We hypothesize that factors like familial status, family and school support, and other social and economic indicators will strongly influence the rates that secondary school students consume alcohol and that increased alcohol consumption is correlated with their school performance.

Data description

```
studentalc <- read_csv("~/project-Hypothesis-Heroes/data/student-por.csv")
```

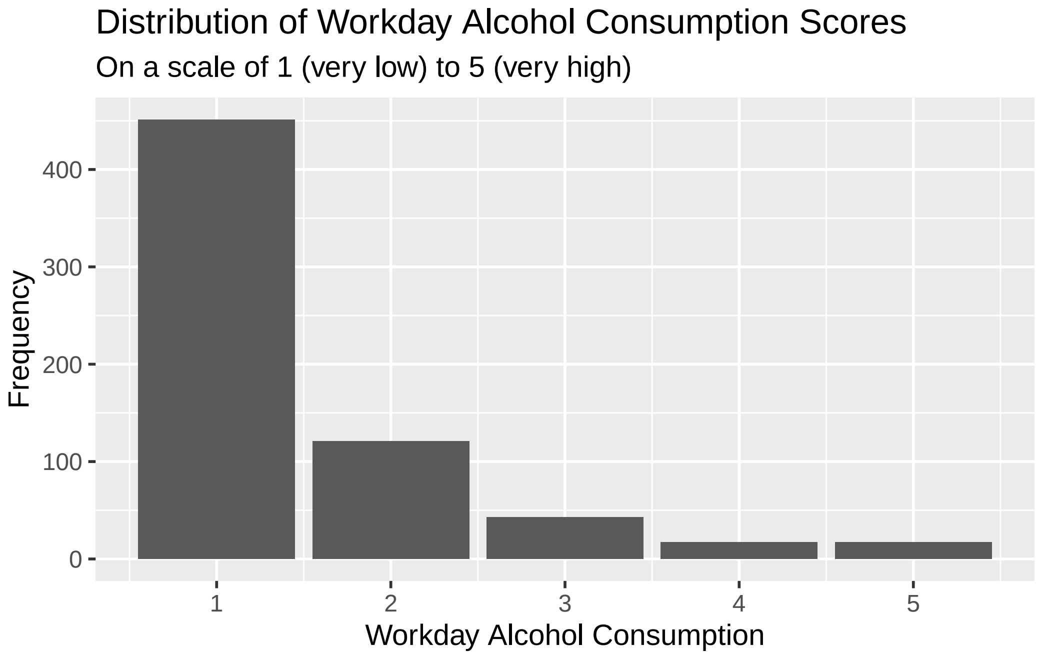
This data is from a Kaggle public dataset, originally sourced from the UC Irvine Machine Learning Repository. The data consists of information collected in 2008 on secondary school students from two schools in Portugal: Gabriel Pereira and Mousinho da Silveira. There are 649 observations, each one being a student, and 33 variables which cover a range of characteristics about each student's family, education, social situation, alcohol consumption, and grades in their Portuguese language class.

Initial exploratory data analysis

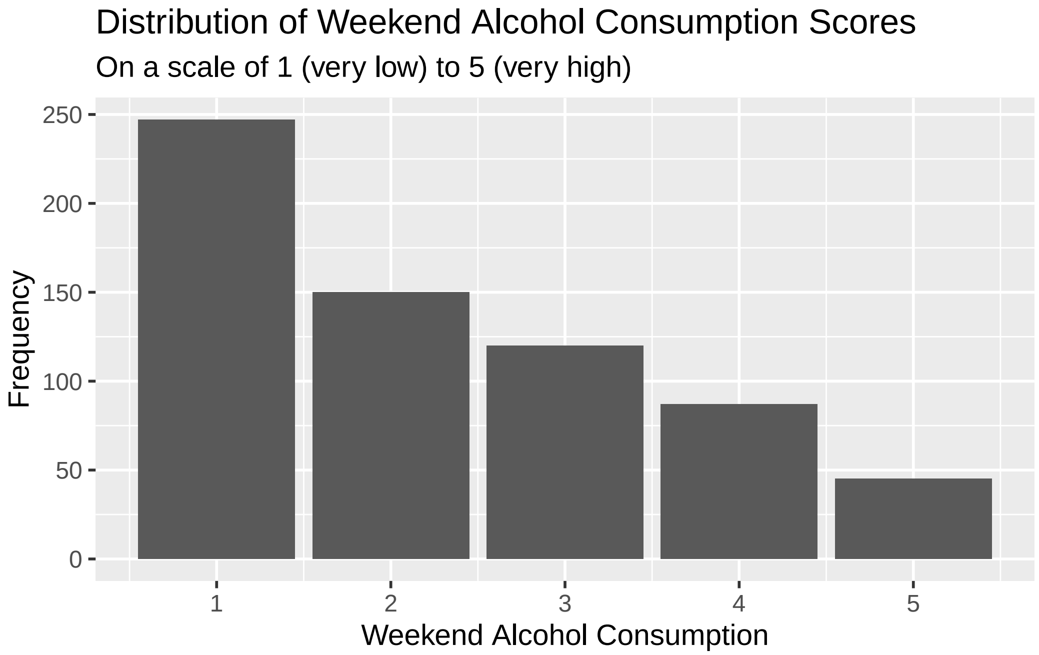
Response variable(s): Alcohol Consumption:

- Dalc (workday alcohol consumption)
- Walc (weekend alcohol consumption)

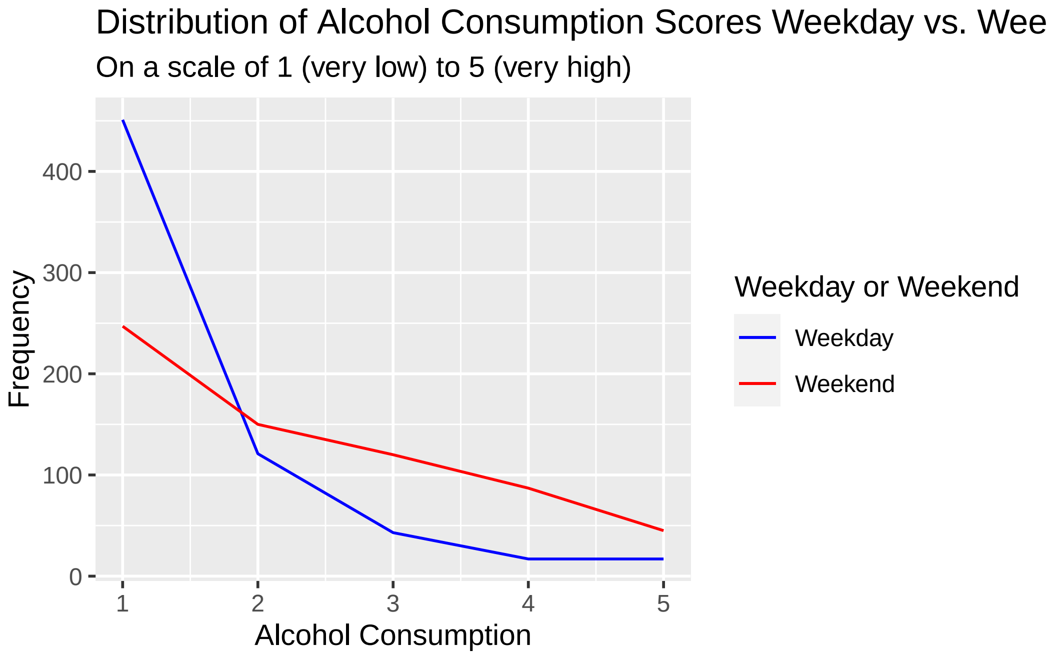
```
studentalc |>
  ggplot(aes(x = Dalc)) +
  geom_histogram(stat = "count") +
  labs(x = "Workday Alcohol Consumption",
       y = "Frequency",
       title = "Distribution of Workday Alcohol Consumption Scores",
       subtitle = "On a scale of 1 (very low) to 5 (very high)")
```



```
studentalc |>
  ggplot(aes(x = Walc)) +
  geom_histogram(stat = "count") +
  labs(x = "Weekend Alcohol Consumption",
       y = "Frequency",
       title = "Distribution of Weekend Alcohol Consumption Scores",
       subtitle = "On a scale of 1 (very low) to 5 (very high)")
```

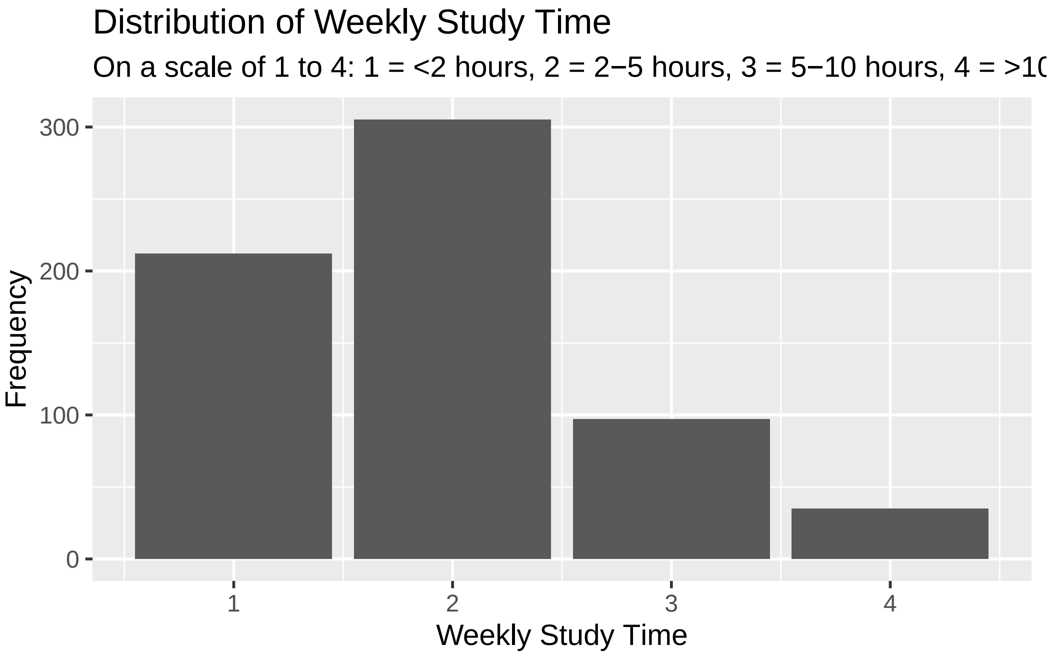


```
studentalc %>%
  ggplot(aes(x = Dalc)) +
  geom_line(aes(y = ..count.., color = "Dalc"), stat = 'count') +
  geom_line(aes(x = Walc, y = ..count.., color = "Walc"), stat = 'count') +
  labs(x = "Alcohol Consumption",
       y = "Frequency",
       color = "Weekday or Weekend",
       title = "Distribution of Alcohol Consumption Scores Weekday vs. Weekend",
       subtitle = "On a scale of 1 (very low) to 5 (very high)") +
  scale_color_manual(
    values = c("Dalc" = "blue", "Walc" = "red"),
    labels = c("Weekday", "Weekend")
  )
```



Potential Quantitative Predictor Variable: studytime

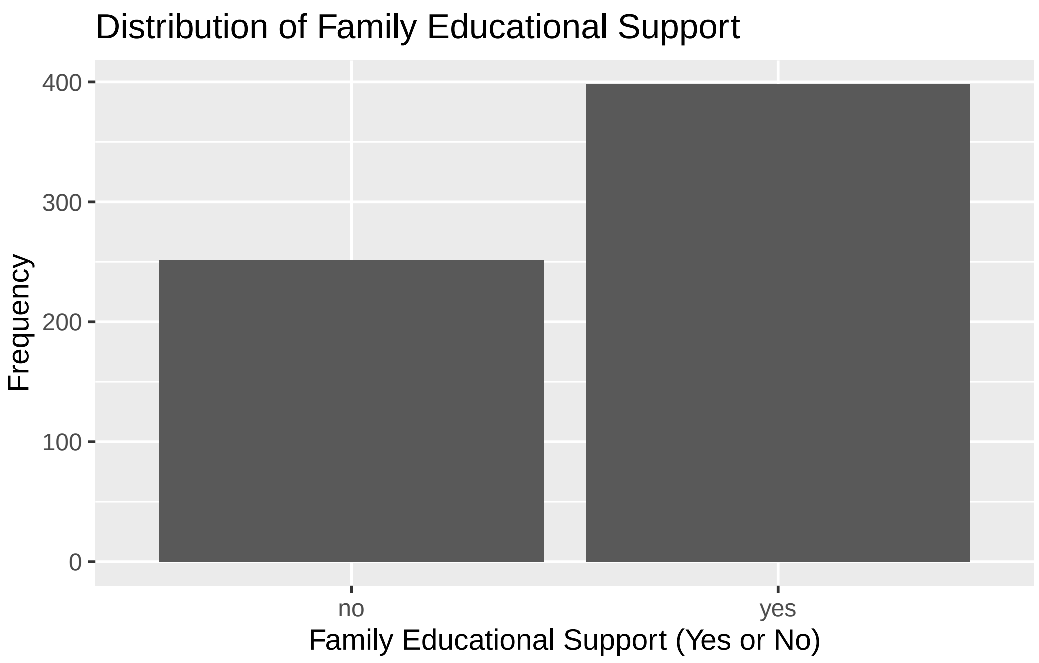
```
studentalc |>
  ggplot(aes(x = studytime)) +
  geom_histogram(stat = "count") +
  labs(x = "Weekly Study Time",
       y = "Frequency",
       title = "Distribution of Weekly Study Time",
       subtitle = "On a scale of 1 to 4: 1 = <2 hours, 2 = 2-5 hours, 3 = 5-10 hours, 4 =
```



The distribution of Weekly Study Time is unimodal at a self-reported score of 2, corresponding to 2-5 hours of studying a week. It appears that a significant majority of students study between 0 and 5 hours, while few students study more than 10 hours.

Potential Categorical Predictor Variable: famsup

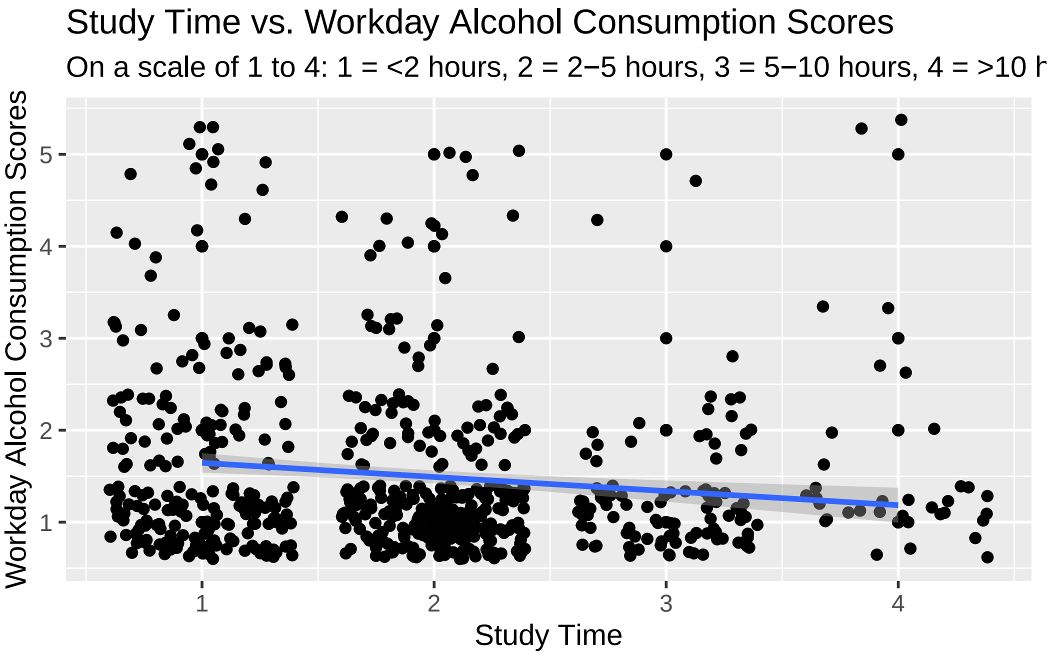
```
studentalc |>
  ggplot(aes(x = famsup)) +
  geom_histogram(stat = "count") +
  labs(x = "Family Educational Support (Yes or No)",
       y = "Frequency",
       title = "Distribution of Family Educational Support")
```



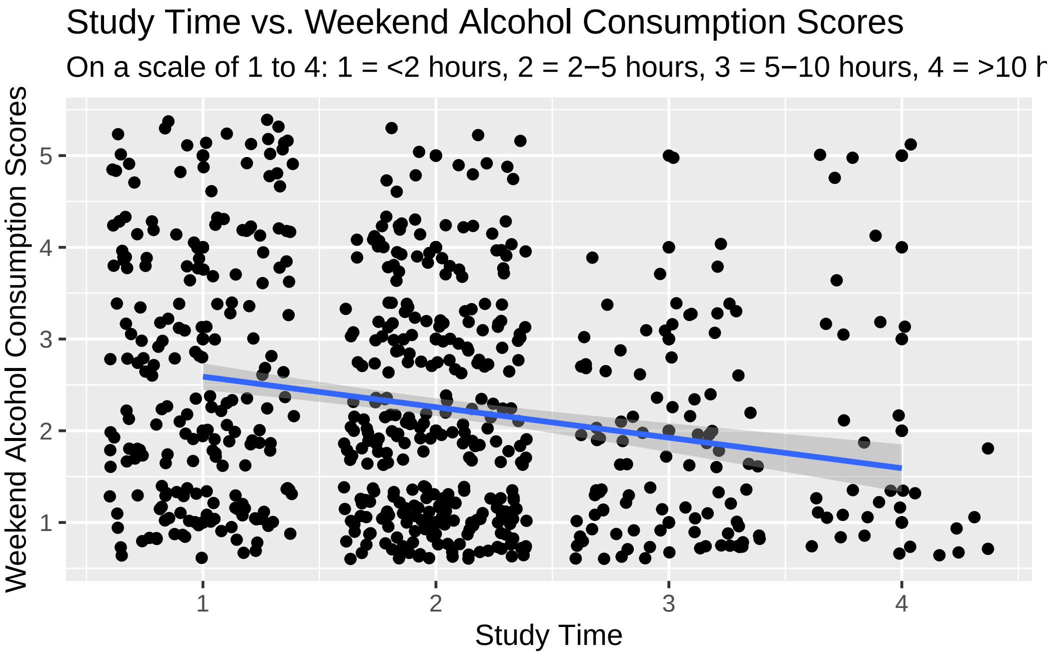
About 400 students in the sample receive family educational support, while around 250 do not.

Predictor-Response Relationships:

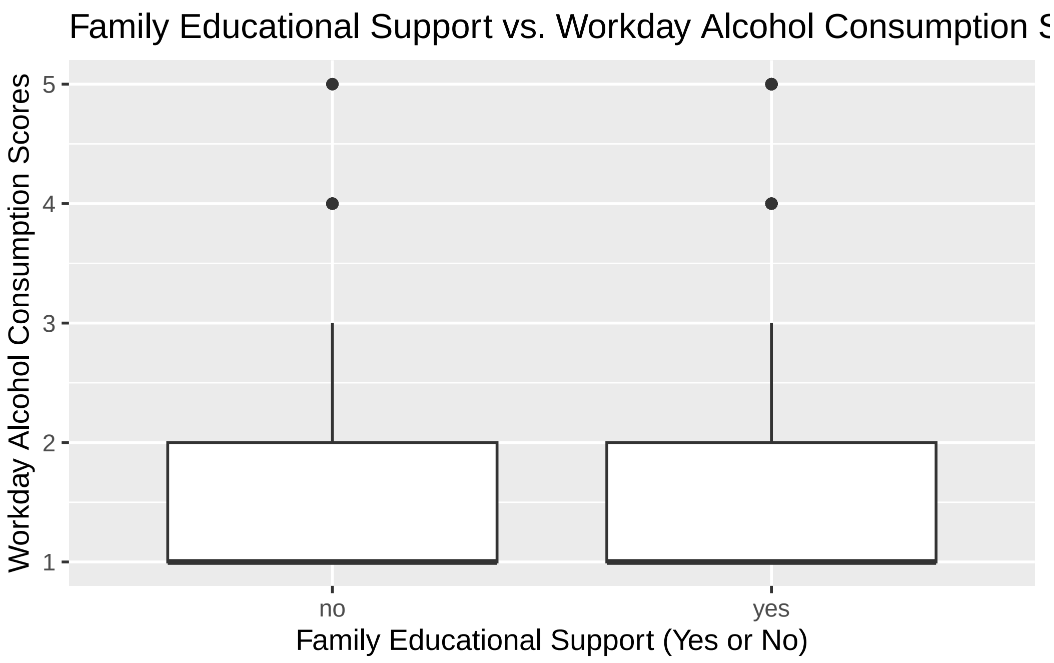
```
studentalc |>
  ggplot(aes(x = studytime, y = Dalc)) +
  geom_point() +
  geom_jitter() +
  geom_smooth(method = "lm") +
  labs(x = "Study Time",
       y = "Workday Alcohol Consumption Scores",
       title = "Study Time vs. Workday Alcohol Consumption Scores",
       subtitle = "On a scale of 1 to 4: 1 = <2 hours, 2 = 2-5 hours, 3 = 5-10 hours, 4 =
```



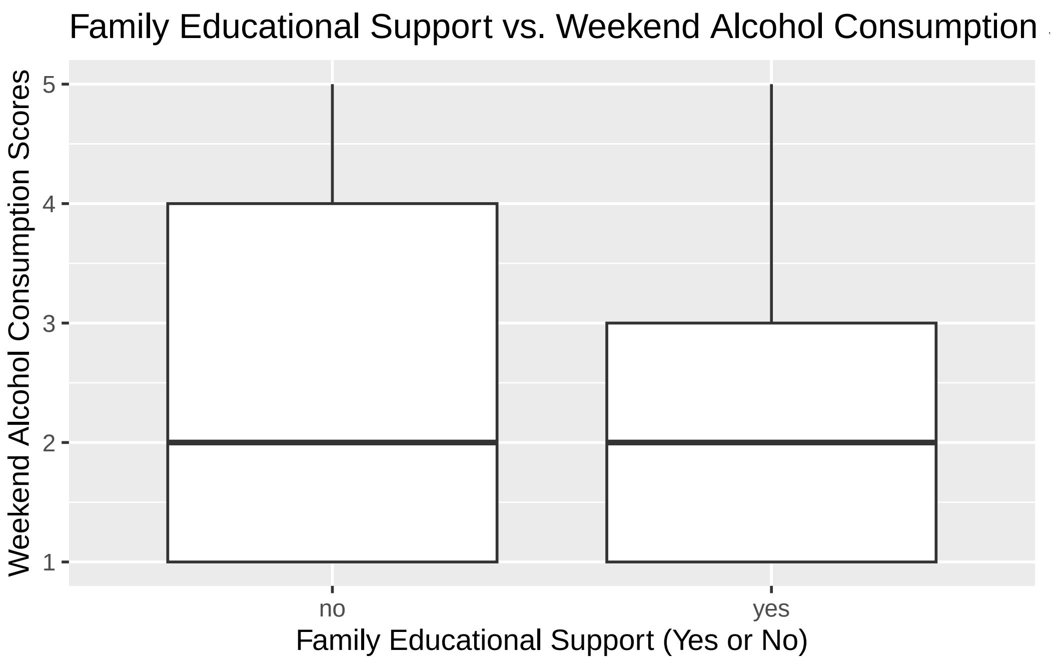
```
studentalc |>
  ggplot(aes(x = studytime, y = Walc)) +
  geom_point() +
  geom_jitter() +
  geom_smooth(method = "lm") +
  labs(x = "Study Time",
       y = "Weekend Alcohol Consumption Scores",
       title = "Study Time vs. Weekend Alcohol Consumption Scores",
       subtitle = "On a scale of 1 to 4: 1 = <2 hours, 2 = 2-5 hours, 3 = 5-10 hours, 4 =
```



```
studentalc |>
  ggplot(aes(x = famsup, y = Dalc)) +
  geom_boxplot() +
  labs(x = "Family Educational Support (Yes or No)",
       y = "Workday Alcohol Consumption Scores",
       title = "Family Educational Support vs. Workday Alcohol Consumption Scores")
```



```r
studentalc |>
  ggplot(aes(x = famsup, y = Walc)) +
  geom_boxplot() +
  labs(x = "Family Educational Support (Yes or No)",
       y = "Weekend Alcohol Consumption Scores",
       title = "Family Educational Support vs. Weekend Alcohol Consumption Scores")
```



Potential Interaction Effect: One interaction effect we're interested in exploring is the interaction between famsup and Pstatus, which is the interaction between Family Educational Support that a student receives and the parent's cohabitation status.

Analysis approach

We will be creating a model for alcohol consumption in secondary school students. Our response variables are weekday and weekend alcohol consumption, which may be mutated into a single variable. Our potential predictors include family educational support, study time, grades, cohabitation status, familial education, and economic status. Alcohol consumption is measured on a self-report scale from 1 to 5, 5 being the highest. Study time is stratified into low medium and high amounts, grades are scaled from 1 to 20, educational support is a yes or no boolean, cohabitation status, economic status are both multi-level categorical variables. We will use a multiple linear regression in order to fit the variables to our model. We do not predict interaction as of now, though we may find some.

Data dictionary

The data dictionary can be found here.