# Final Project - Step 2 (15 Points)

**PSTAT100: Data Science Concepts and Analysis**


**STUDENT NAME**

- Xinyao Song (xinyaosong)
- Matthew Chan (hchan837)
- STUDENT 3 (NetID 3)
- STUDENT 4 (NetID 4)
- STUDENT 5 (NetID 5)


Due Date

The deadline for this step is **Friday, May 9, 2025**.

Instructions

In this step, you will develop clear research questions and hypotheses based on your selected dataset, and conduct a thorough Exploratory Data Analysis (EDA). This foundational work is crucial for guiding your analysis in the following steps.

## 1 Step 2: Research Questions, Hypotheses, and Exploratory Data Analysis (EDA)

### 1.1 Research Questions

**Question 1**

- How does weekly study time, parents' education level, and internet access influence a student's average Portuguese grade (average of G1, G2, and G3)?

**Question 2**

- How is alcohol consumption among Portuguese high school students associated with their academic performance and social activities?

**Question 3**

- Does the average free time differ between students who are in a romantic relationship and those who are not?

## 1.2 Hypotheses

### Hypothesis 1

- At least one of the predictors (weekly study time, parents' education level, or internet access) has a significant effect on students' average Portuguese grade.

### Hypothesis 2

- Alcohol consumption is significantly associated with either academic performance or social activity levels among Portuguese high school students.

### Hypothesis 3

- There is a significant difference in means between students in a relationship and those who are not.

## 1.3 Exploratory Data Analysis (EDA)

## 1.4 Data Cleaning

```r
library(dplyr)
library(tidyr)
library(readr)
library(ggplot2)

# Preprocess student data frame
por_df <- readr::read_delim("../data/student-por.csv", delim = ";")

por_df <- por_df %>%
  mutate(
    school = as.factor(school),
    sex = as.factor(sex),
    address = as.factor(address),
    famsize = as.factor(famsize),
    Pstatus = as.factor(Pstatus),
    Mjob = as.factor(Mjob),
    Fjob = as.factor(Fjob),
    reason = as.factor(reason),
    guardian = as.factor(guardian),
    schoolsup = as.factor(schoolsup),
    famsup = as.factor(famsup),
    paid = as.factor(paid),
    activities = as.factor(activities),
    nursery = as.factor(nursery),
    higher = as.factor(higher),
    internet = as.factor(internet),
    romantic = as.factor(romantic),
    # ordered factor from 1-5 for very high, low etc.
    Dalc = factor(Dalc, levels = 1:5, ordered = TRUE),
    Walc = factor(Walc, levels = 1:5, ordered = TRUE)
  )

# Check for missing values
print(anyNA(por_df))
```

```
[1] FALSE
```

From the missing value check above, we can see that there are no missing values in this dataset. Therefore, no data cleaning is necessary.

## 1.5 Descriptive Statistics

```
summary(por_df)
```

```
 school   sex          age        address famsize   Pstatus      Medu
 GP:423   F:383   Min.   :15.00   R:197   GT3:457   A: 80   Min.   :0.000
 MS:226   M:266   1st Qu.:16.00   U:452   LE3:192   T:569   1st Qu.:2.000
                  Median :17.00                             Median :2.000
                  Mean   :16.74                             Mean   :2.515
                  3rd Qu.:18.00                             3rd Qu.:4.000
                  Max.   :22.00                             Max.   :4.000
     Fedu             Mjob           Fjob            reason      guardian
 Min.   :0.000   at_home :135   at_home : 42   course    :285   father:153
 1st Qu.:1.000   health  : 48   health  : 23   home      :149   mother:455
 Median :2.000   other   :258   other   :367   other     : 72   other : 41
 Mean   :2.307   services:136   services:181   reputation:143
 3rd Qu.:3.000   teacher : 72   teacher : 36
 Max.   :4.000
  traveltime      studytime        failures      schoolsup famsup     paid
 Min.   :1.000   Min.   :1.000   Min.   :0.0000   no :581   no :251   no :610
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   yes: 68   yes:398   yes: 39
 Median :1.000   Median :2.000   Median :0.0000
 Mean   :1.569   Mean   :1.931   Mean   :0.2219
 3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
 Max.   :4.000   Max.   :4.000   Max.   :3.0000
 activities nursery   higher    internet  romantic      famrel
 no :334    no :128   no : 69   no :151   no :410   Min.   :1.000
 yes:315    yes:521   yes:580   yes:498   yes:239   1st Qu.:4.000
                                                    Median :4.000
                                                    Mean   :3.931
                                                    3rd Qu.:5.000
                                                    Max.   :5.000
   freetime        goout       Dalc    Walc        health
 Min.   :1.00   Min.   :1.000   1:451   1:247   Min.   :1.000
 1st Qu.:3.00   1st Qu.:2.000   2:121   2:150   1st Qu.:2.000
 Median :3.00   Median :3.000   3: 43   3:120   Median :4.000
 Mean   :3.18   Mean   :3.185   4: 17   4: 87   Mean   :3.536
 3rd Qu.:4.00   3rd Qu.:4.000   5: 17   5: 45   3rd Qu.:5.000
 Max.   :5.00   Max.   :5.000                   Max.   :5.000
   absences            G1              G2              G3
 Min.   : 0.000   Min.   : 0.0   Min.   : 0.00   Min.   : 0.00
 1st Qu.: 0.000   1st Qu.:10.0   1st Qu.:10.00   1st Qu.:10.00
 Median : 2.000   Median :11.0   Median :11.00   Median :12.00
 Mean   : 3.659   Mean   :11.4   Mean   :11.57   Mean   :11.91
 3rd Qu.: 6.000   3rd Qu.:13.0   3rd Qu.:13.00   3rd Qu.:14.00
 Max.   :32.000   Max.   :19.0   Max.   :19.00   Max.   :19.00
```

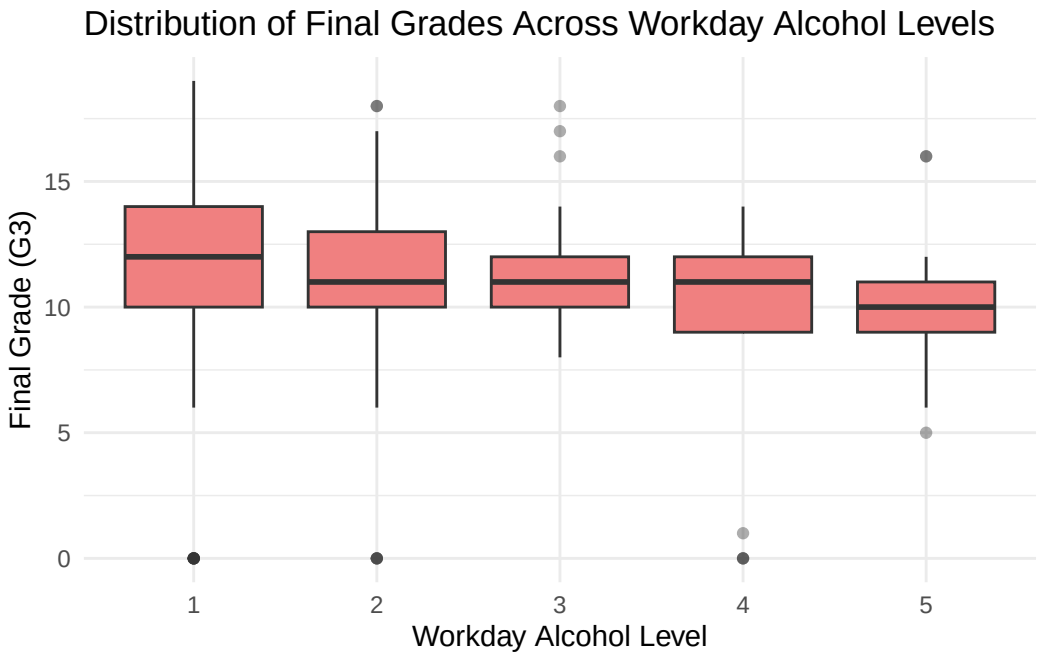
## 1.6 Data Visualization

### 1.6.1 Research Question 2: How is alcohol consumption among Portuguese high school students associated with their academic performance and social activities?

## Grade Distributions by Alcohol Levels

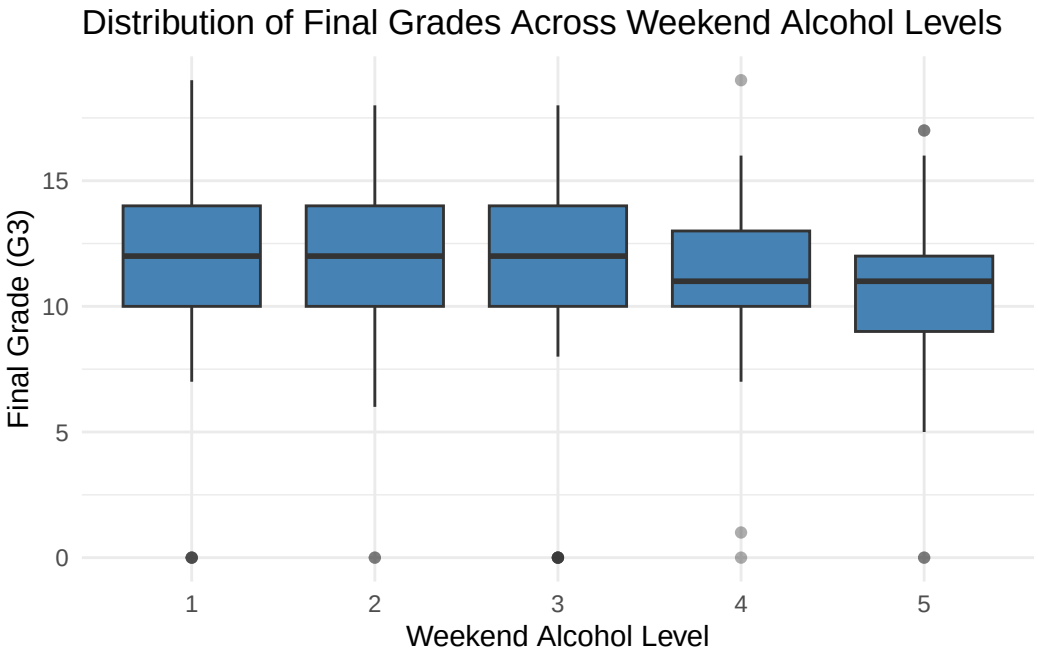
**Final Grade by Workday Alcohol Category:**

```
ggplot(por_df, aes(x = Dalc, y = G3)) +
  geom_boxplot(outlier.alpha = 0.4, fill = "lightcoral") +
  labs(
    title = "Distribution of Final Grades Across Workday Alcohol Levels",
    x = "Workday Alcohol Level",
    y = "Final Grade (G3)"
  ) +
  theme_minimal()
```




**Final Grade by Weekend Alcohol Category:**

```
ggplot(por_df, aes(x = Walc, y = G3)) +
  geom_boxplot(outlier.alpha = 0.4, fill = "steelblue") +
  labs(
    title = "Distribution of Final Grades Across Weekend Alcohol Levels",
    x = "Weekend Alcohol Level",
    y = "Final Grade (G3)"
  ) +
  theme_minimal()
```




---

## Social Activities vs Alcohol Consumption

**Going Out Frequency by Workday Alcohol:**

```
ggplot(por_df, aes(x = Dalc, y = goout)) +
  geom_boxplot(outlier.alpha = 0.4, fill = "lightcoral") +
  labs(
    title = "Going Out with Friends Across Workday Alcohol Levels",
    x = "Workday Alcohol Level",
    y = "Going Out Frequency (1=Very Low, 5=Very High)"
  ) +
  theme_minimal()
```

**Free Time After School by Weekend Alcohol:**

```r
ggplot(por_df, aes(x = Walc, y = freetime)) +
  geom_boxplot(fill = "steelblue") +
  labs(
    title = "Free Time After School Across Weekend Alcohol Levels",
    x = "Weekend Alcohol Level",
    y = "Free Time (1=Very Low, 5=Very High)"
  ) +
  theme_minimal()
```