Final Project - Step 2 (15 Points)

PSTAT100: Data Science Concepts and Analysis

STUDENT NAME

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

Below are the summary statistics for all numeric variables.

```
por_df %>%
  select(where(is.numeric)) %>%
  summary()
```

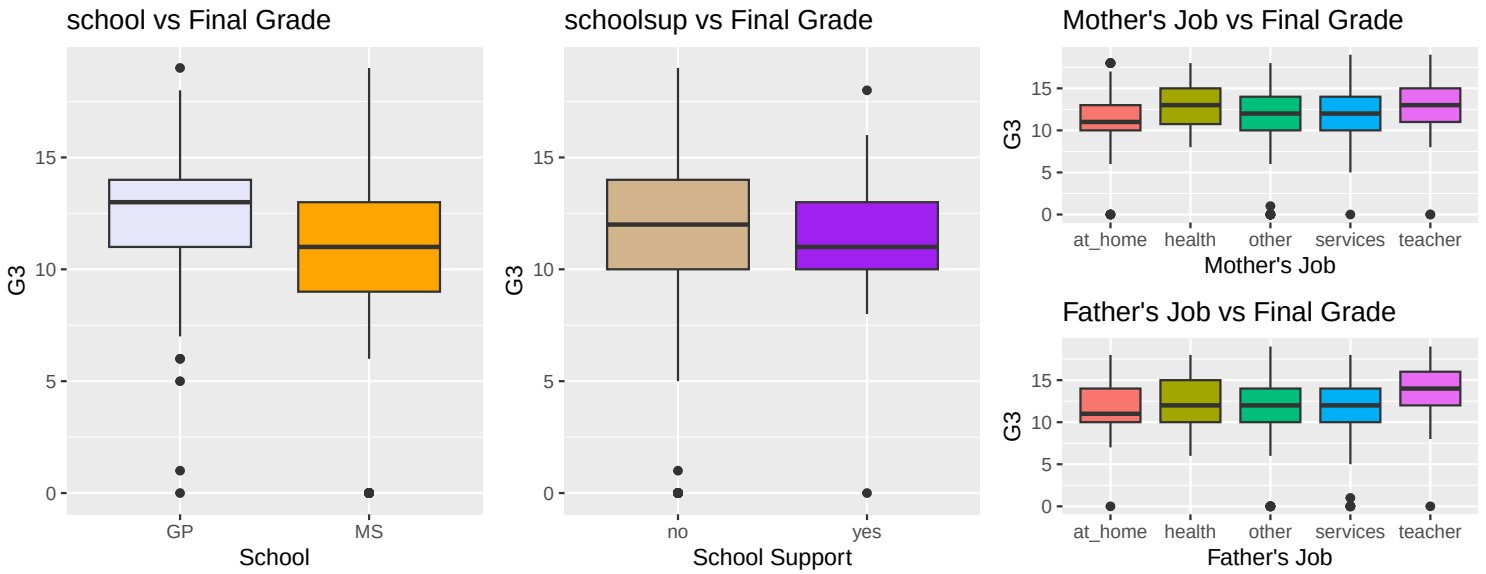
```
      age             Medu            Fedu          traveltime   
 Min.   :15.00   Min.   :0.000   Min.   :0.000   Min.   :1.000  
 1st Qu.:16.00   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000  
 Median :17.00   Median :2.000   Median :2.000   Median :1.000  
 Mean   :16.74   Mean   :2.515   Mean   :2.307   Mean   :1.569  
 3rd Qu.:18.00   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:2.000  
 Max.   :22.00   Max.   :4.000   Max.   :4.000   Max.   :4.000  
   studytime        failures          famrel        freetime    
 Min.   :1.000   Min.   :0.0000   Min.   :1.000   Min.   :1.00  
 1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:4.000   1st Qu.:3.00  
 Median :2.000   Median :0.0000   Median :4.000   Median :3.00  
 Mean   :1.931   Mean   :0.2219   Mean   :3.931   Mean   :3.18  
 3rd Qu.:2.000   3rd Qu.:0.0000   3rd Qu.:5.000   3rd Qu.:4.00  
 Max.   :4.000   Max.   :3.0000   Max.   :5.000   Max.   :5.00  
     goout           health         absences            G1       
 Min.   :1.000   Min.   :1.000   Min.   : 0.000   Min.   : 0.0  
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.: 0.000   1st Qu.:10.0  
 Median :3.000   Median :4.000   Median : 2.000   Median :11.0  
 Mean   :3.185   Mean   :3.536   Mean   : 3.659   Mean   :11.4  
 3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.: 6.000   3rd Qu.:13.0  
 Max.   :5.000   Max.   :5.000   Max.   :32.000   Max.   :19.0  
       G2              G3       
 Min.   : 0.00   Min.   : 0.00  
 1st Qu.:10.00   1st Qu.:10.00  
 Median :11.00   Median :12.00  
 Mean   :11.57   Mean   :11.91  
 3rd Qu.:13.00   3rd Qu.:14.00  
 Max.   :19.00   Max.   :19.00  
```

1.6 Data Visualization

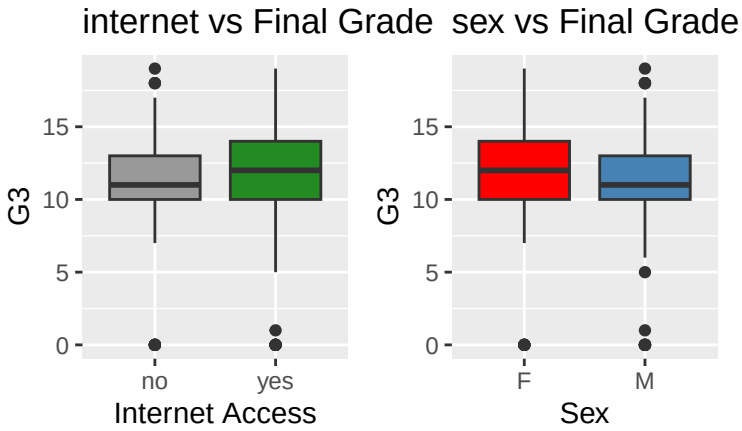
1.6.1 Research Question 1: Impact of Non-previous grade factors on grade

```
p1 <- make_boxplot(por_df, "school", c("GP" = "lavender", "MS" = "orange"), xlabel = "School")
p2 <- make_boxplot(por_df, "schoolsup", c("yes" = "purple", "no" = "tan"), xlabel = "School Support")
p3 <- make_boxplot(por_df, "Mjob", title = "Mother's Job vs Final Grade", xlabel = "Mother's Job")
p4 <- make_boxplot(por_df, "Fjob", title = "Father's Job vs Final Grade", xlabel = "Father's Job")
p5 <- make_boxplot(por_df, "internet", c("yes" = "forestgreen", "no" = "gray60"), xlabel = "Internet Ac
```

```r
p6 <- make_boxplot(por_df, "sex", c("F" = "red", "M" = "steelblue"), xlabel = "Sex")

(p1 | p2) | (p3 / p4)
```




```r
p5 | p6
```



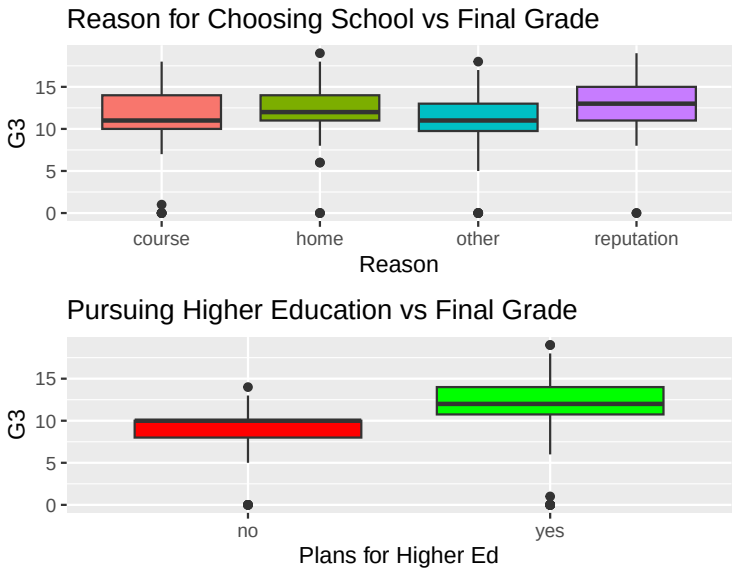
The GP school has a higher median G3 among the two. School support indicates struggling students–its median is slightly lower than those not receiving support.

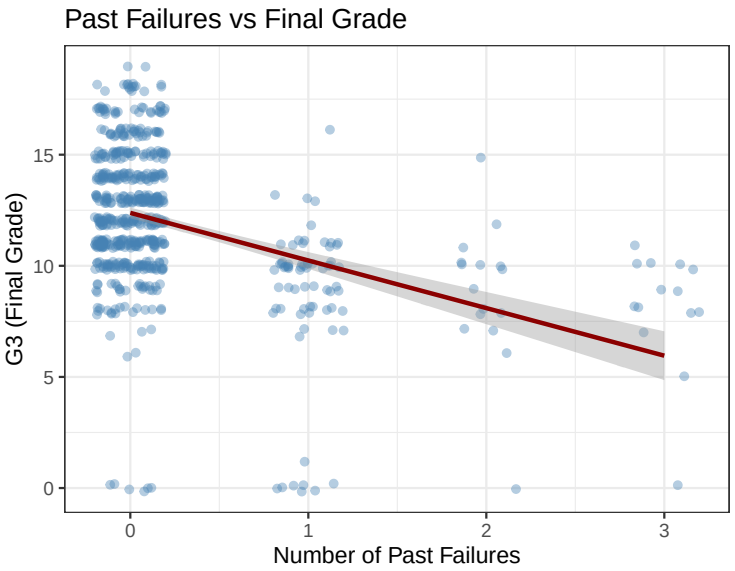
For parent professions, the 95% median line indicates teacher parent profession is highest, and at home is lowest. Health is highest with teacher for mother's, and in the middle range for father's.

Having internet access is significant as well, and so is the student's sex: having internet and being female having 95% confidence medians above the opposite groups.

```r
p_reason <- make_boxplot(por_df, "reason",
                         title = "Reason for Choosing School vs Final Grade",
                         xlabel = "Reason")

p_higher <- make_boxplot(por_df, "higher", fill_colors = c("yes" = "green", "no" = "red"),
  title = "Pursuing Higher Education vs Final Grade",
  xlabel = "Plans for Higher Ed")

```

```r

p_failures <- ggplot(por_df, aes(failures, G3)) +
  geom_jitter(width = 0.2, height = 0.2, alpha = 0.4, color = "steelblue") +
  geom_smooth(method = "lm", color = "darkred") +
  labs(title = "Past Failures vs Final Grade",
       x = "Number of Past Failures",
       y = "G3 (Final Grade)") +
  theme_bw()

(p_reason / p_higher) | p_failures
```





For the set of graph on the left, both of these factors indicate motivation. So, as is evident, they have significantly different medians between each of these two sets of groups.

For the graph on the right, it is the strongest predictor that is not based on G2 grade. Its R^2 is rather weak, but bt tracing a horizontal line from the left to the right, finding if the left's shaded bottom is above a point on the right's, one finds: - 0 failures has a significantly greater predicted final grade than the rest - 1 failure has a significantly greater predicted final grade than 3 failures.

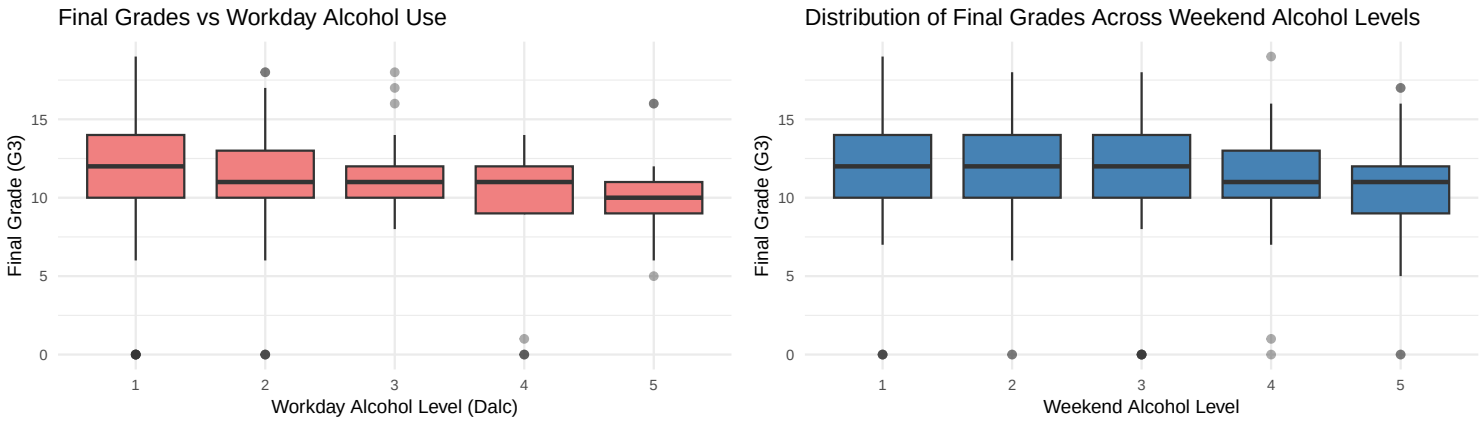
1.6.2 Research Question 2: How is alcohol consumption among Portuguese high school students associated with their academic performance and social activities?

1.6.2.1 Grade Distributions by Alcohol Levels

Final Grade by Workday Alcohol Category & Final Grade by Weekend Alcohol Category

```r
p_dalc <- ggplot(por_df, aes(x = Dalc, y = G3)) +
  geom_boxplot(outlier.alpha = 0.4, fill = "lightcoral") +
  labs(
    title = "Final Grades vs Workday Alcohol Use",
    x = "Workday Alcohol Level (Dalc)",
    y = "Final Grade (G3)"
  ) +
  theme_minimal()+
  theme(text = element_text(size = 9))

p_walc <- ggplot(por_df, aes(x = Walc, y = G3)) +
```

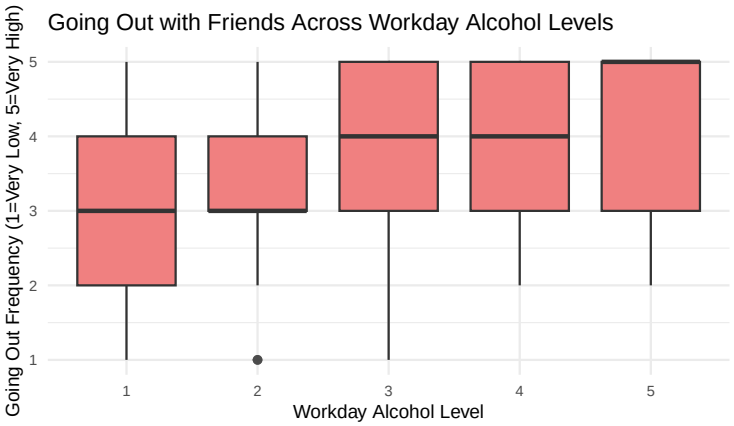
```
  geom_boxplot(outlier.alpha = 0.4, fill = "steelblue") +
  labs(
    title = "Distribution of Final Grades Across Weekend Alcohol Levels",
    x = "Weekend Alcohol Level",
    y = "Final Grade (G3)"
  ) +
  theme_minimal()+
  theme(text = element_text(size = 9))

p_dalc | p_walc
```

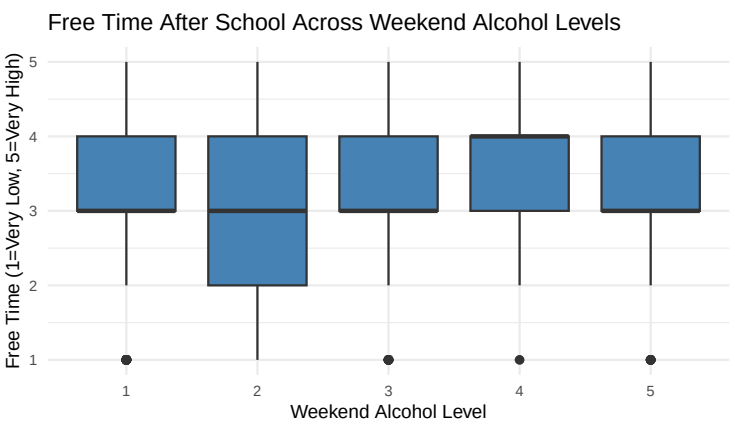


1.6.2.2 Social Activities vs Alcohol Consumption

Going Out Frequency by Workday Alcohol & Free Time After School by Weekend Alcohol

```
p_goout <- 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()+
  theme(text = element_text(size = 9))

p_freetime <- 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()+
  theme(text = element_text(size = 9))

p_goout | p_freetime
```





1.7 Question 3: Does the average free time differ between students who are in a romantic relationship and those who are not?

```r
#We create a variable to compare the means of people in relationships vs not in relationships
mean_df <- por_df %>% group_by(romantic) %>% summarize(mean_freetime = mean(freetime))

p_hist <- ggplot(por_df, aes(x = freetime)) +
  geom_histogram(aes(y = ..density.., fill = as.factor(romantic)),
                 alpha = 0.7, binwidth = 1, color = "black") +
  facet_wrap(~ romantic) +
  labs(title = "Density Histogram of Freetime by Relationship Status", x = "Freetime", y = "Density", f
  theme_minimal() +
  theme(text = element_text(size = 9))

p_density <- ggplot(por_df, aes(x = freetime, fill = as.factor(romantic))) +
  geom_density(alpha = 0.4, bw = 0.35) +
  labs(title = "Density Estimate of Freetime (Romantic vs Not)", x = "Freetime",y = "Density", fill = "
  theme_minimal() +
  theme(text = element_text(size = 9))

p_hist | p_density
```