Project Proposal

November 17

Ilseop Lee, Ramil Mammadov, Tursunai Turumbekova, Yirang Liu

Load Packages

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(cowplot)
```

Dataset 1

- **Dataset Name**: Obesity Levels
- **Source**: Kaggle (Obesity Levels Dataset); Original research source: Estimation of Obesity Levels with a Trained Neural Network Approach Optimized by the Bayesian Technique, DOI: 10.3390/app13063875
- **Brief description**: This dataset originates from a study aimed at estimating obesity levels based on demographic, dietary, and lifestyle factors, containing 2,111 records from individuals in Mexico, Peru, and Colombia. Of the data, 23% was directly collected from users via a web platform, while 77% was synthetically generated using the Weka tool and the SMOTE filter to ensure a balanced representation of obesity levels. Each row in the dataset represents a unique individual, capturing their demographic details, lifestyle habits, dietary patterns, and obesity classification. The target variable categorizes individuals into seven obesity levels.

Research Question 1: How does the interaction between family history of overweight (`family_history_with_overweight`) and the frequency of physical activity (`FAF`) influence an individual's BMI?

- **Outcome Variable**: `BMI` (Continuous) - Body Mass Index calculated using the formula $BMI = \frac{Weight(kg)}{Height(m)^2}$, representing an individual's body composition.

- **Explanatory Variables**: `family_history_with_overweight` (Binary: Yes/No), `FAF` (Continuous: Frequency of physical activity per week in days)
- **Interaction Term**: `family_history_with_overweight * FAF`. This interaction will help evaluate whether physical activity affects BMI differently based on whether an individual has a family history of overweight.

Research Question 2: How does the frequency of alcohol consumption (`CALC`) and meal frequency (`NCP`) jointly influence obesity levels?

- **Outcome Variable**: `NObeyesdad` (Ordinal) - Obesity level classified into seven categories: Insufficient Weight, Normal Weight, Overweight Level I/II, and Obesity Type I/II/III.
- **Explanatory Variables**: `CALC` (Nominal: No, Sometimes, Frequently, Always), `NCP` (Continuous: Number of main meals per day)
- **Interaction Term**: `CALC * NCP` This interaction will help explore whether the frequency of alcohol consumption modifies the relationship between meal frequency and obesity levels.

Load the data and provide a `glimpse()`:

```
# Load the dataset
obesity_data <- read_csv(
  paste0(
    "https://github.com/cathylyirang/IDS702_Project_Group_5/",
    "raw/refs/heads/main/Data_Obesity/",
    "ObesityDataSet_raw_and_data_sinthetic.csv"
  ))
```

```
Rows: 2111 Columns: 17
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (9): Gender, CALC, FAVC, SCC, SMOKE, family_history_with_overweight, CAE...
dbl (8): Age, Height, Weight, FCVC, NCP, CH2O, FAF, TUE

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

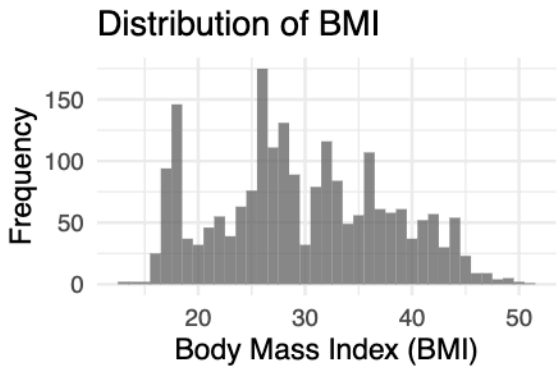
```
obesity_data <- obesity_data |>
  mutate(BMI = Weight / (Height^2))
glimpse(obesity_data)
```

```
Rows: 2,111
Columns: 18
$ Age                             <dbl> 21, 21, 23, 27, 22, 29, 23, 22, 24, 22,~
$ Gender                          <chr> "Female", "Female", "Male", "Male", "Ma~
$ Height                          <dbl> 1.62, 1.52, 1.80, 1.80, 1.78, 1.62, 1.5~
$ Weight                          <dbl> 64.0, 56.0, 77.0, 87.0, 89.8, 53.0, 55.~
$ CALC                            <chr> "no", "Sometimes", "Frequently", "Frequ~
$ FAVC                            <chr> "no", "no", "no", "no", "no", "yes", "y~
$ FCVC                            <dbl> 2, 3, 2, 3, 2, 2, 3, 2, 3, 2, 3, 2, 3, ~
$ NCP                             <dbl> 3, 3, 3, 3, 1, 3, 3, 3, 3, 3, 3, 3, 3, ~
$ SCC                             <chr> "no", "yes", "no", "no", "no", "no", "n~
$ SMOKE                           <chr> "no", "yes", "no", "no", "no", "no", "n~
$ CH2O                            <dbl> 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 3, ~
$ family_history_with_overweight <chr> "yes", "yes", "yes", "no", "no", "no", ~
$ FAF                             <dbl> 0, 3, 2, 2, 0, 0, 1, 3, 1, 1, 2, 2, 2, ~
$ TUE                             <dbl> 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 2, 1, 0, ~
$ CAEC                            <chr> "Sometimes", "Sometimes", "Sometimes", ~
$ MTRANS                          <chr> "Public_Transportation", "Public_Transp~
$ NObeyesdad                      <chr> "Normal_Weight", "Normal_Weight", "Norm~
$ BMI                             <dbl> 24.38653, 24.23823, 23.76543, 26.85185,~
```

Exploratory Plots:

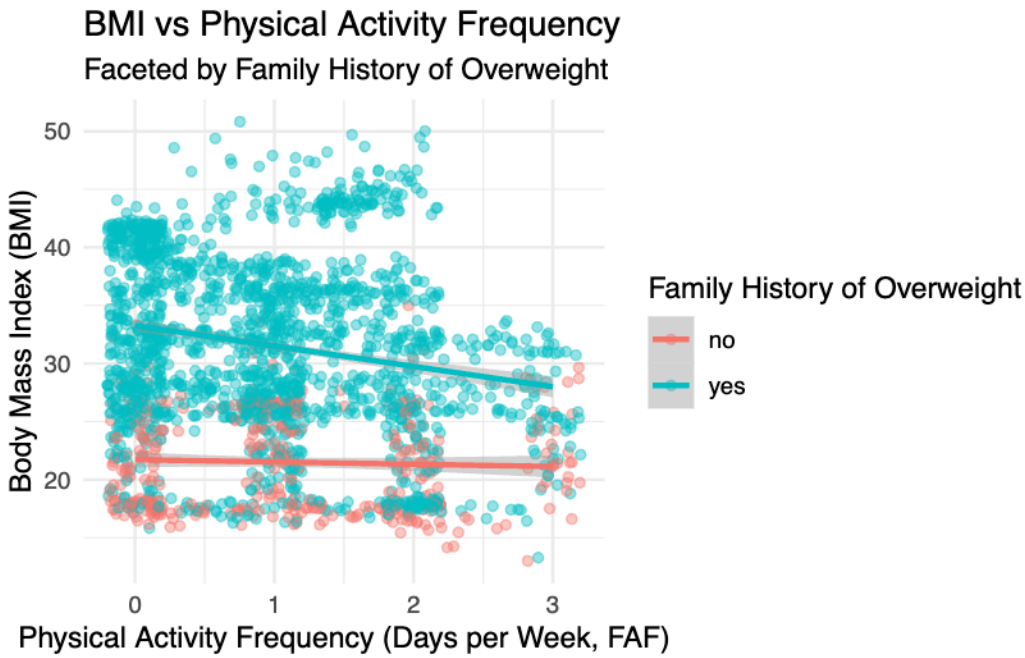
How does the interaction between family history of overweight (`family_history_with_overweight`) and physical activity frequency (`FAF`) influence an individual's BMI?

```
obesity_data |>ggplot(aes(x = BMI)) +
  geom_histogram(binwidth = 1, alpha = 0.7) +labs(title = "Distribution of BMI",
  x = "Body Mass Index (BMI)", y = "Frequency") +theme_minimal()
```



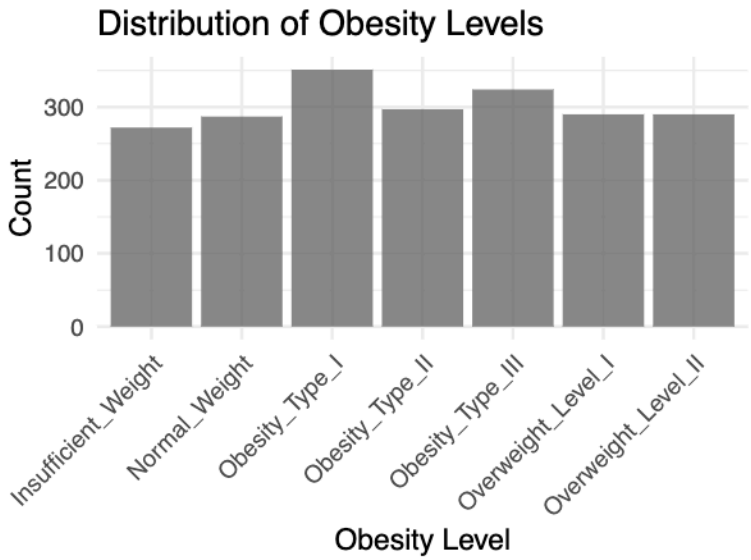

```
obesity_data |>
  ggplot(aes(x = FAF, y = BMI, color = family_history_with_overweight)) +
  geom_jitter(alpha = 0.4, width = 0.2) +
  geom_smooth(method = "lm") +
  labs(title = "BMI vs Physical Activity Frequency",
    subtitle = "Faceted by Family History of Overweight",
    x = "Physical Activity Frequency (Days per Week, FAF)",
    y = "Body Mass Index (BMI)", color = "Family History of Overweight") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

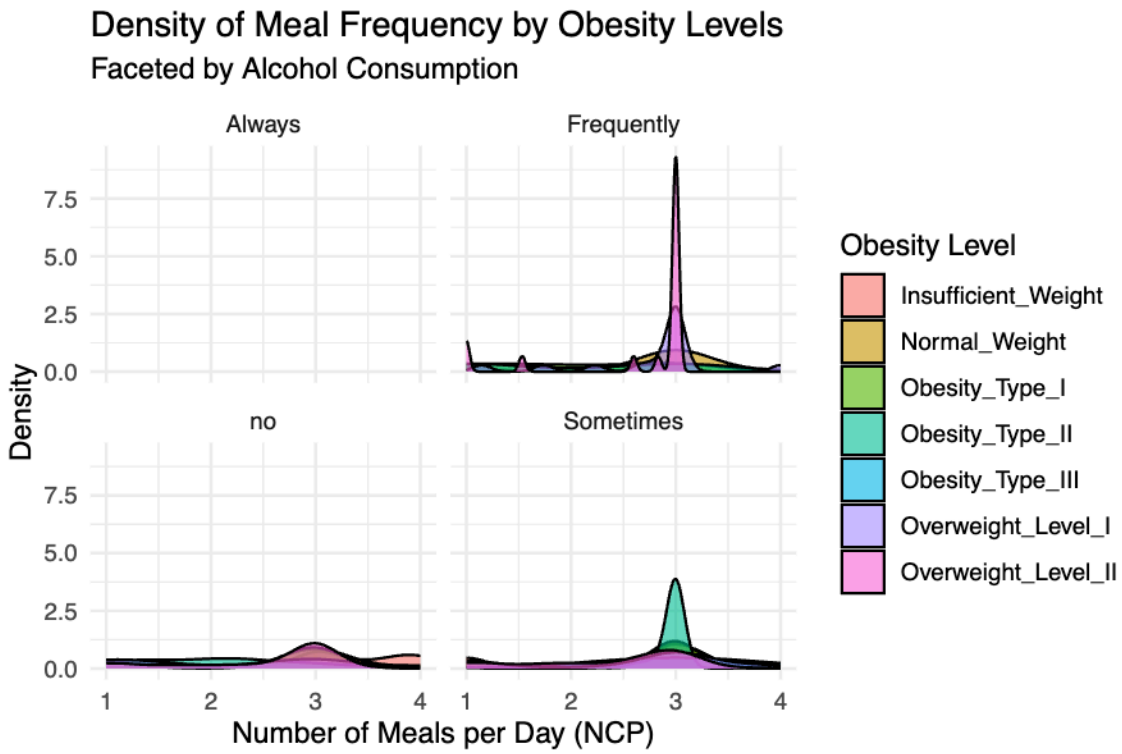


How does the frequency of alcohol consumption (CALC) and meal frequency (NCP) jointly influence obesity levels (NObeyesdad)?

```
obesity_data |>ggplot(aes(x = NObeyesdad)) + geom_bar(alpha = 0.7) +labs(
  title = "Distribution of Obesity Levels",x = "Obesity Level",
  y = "Count") +theme_minimal() + theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
obesity_data |>ggplot(aes(x = NCP, fill = NObeyesdad)) +geom_density(alpha = 0.6) +
               facet_wrap(~ CALC) +labs(
               title = "Density of Meal Frequency by Obesity Levels",
               subtitle = "Faceted by Alcohol Consumption",
               x = "Number of Meals per Day (NCP)", y = "Density",
               fill = "Obesity Level") + theme_minimal()
```



Dataset 2

- **Dataset Name**: Wine Quality(2009)
- **Source:** UCI Machine Learning Repository (http://archive.ics.uci.edu/dataset/186/wine+qualit)
- **Variables**: 14 variables and 6,497 observations of Portuguese "Vinho Verde" wine, 11 continuous variables (e.g., acidity, pH level), 2 categorical variables (red vs white, quality-low, medium, high), 1 interval variable (quality score which ranges from 1 to 10)

Brief description: Each row represents physicochecimal characteristic(e.g., acidity, pH level) of a wine sample, along with its quality score based on sensory evaluation.

Research question 1: What role do `alcohol` content and `citric acid` play in determining high-quality wine?

- **Outcome variable:** Quality Category(High, Medium, Low),
- **Interaction Term:** ***`alcohol`*** * ***`citric acid`***

Research question 2: How do `volatile acidity` and `total sulfur dioxide` levels interact to shape the wine quality score?

- **Outcome variable:** Wine quality categorized as low, medium, or high, based on the continuous quality score.
- **Interaction Terms**: ***`volatile acidity`*** * ***`total sulfur dioxide`**

Load the data and provide a `glimpse()`:

```
wine_data <- read_csv(
  paste0(
    "https://github.com/cathylyirang/IDS702_Project_Group_5/",
    "raw/refs/heads/main/Data_HR/",
    "winequality_integrated.csv"
  )
)
```

```
Rows: 6497 Columns: 14
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): type, quality_category
dbl (12): fixed_acidity, volatile_acidity, citric_acid, residual_sugar, chlo...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

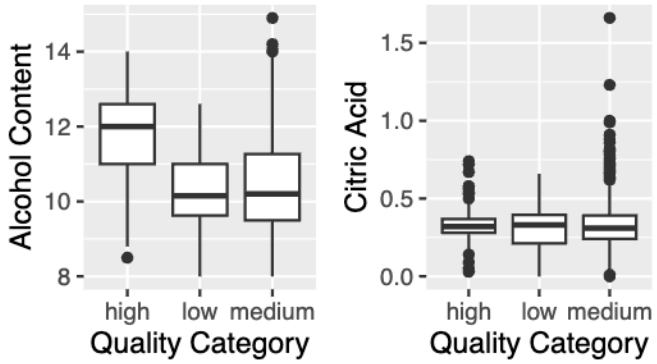
```r
glimpse(wine_data)
```

```
Rows: 6,497
Columns: 14
$ type                <chr> "white", "white", "white", "white", "white", "whi~
$ fixed_acidity       <dbl> 6.1, 7.1, 6.2, 7.5, 7.5, 9.4, 6.8, 6.7, 6.0, 6.8,~
$ volatile_acidity    <dbl> 0.260, 0.490, 0.255, 0.270, 0.230, 0.240, 0.290, ~
$ citric_acid         <dbl> 0.25, 0.22, 0.24, 0.31, 0.35, 0.29, 0.16, 0.26, 0~
$ residual_sugar      <dbl> 2.90, 2.00, 1.70, 5.80, 17.80, 8.50, 1.40, 1.55, ~
$ chlorides           <dbl> 0.047, 0.047, 0.039, 0.057, 0.058, 0.037, 0.038, ~
$ free_sulfur_dioxide <dbl> 289.0, 146.5, 138.5, 131.0, 128.0, 124.0, 122.5, ~
$ total_sulfur_dioxide <dbl> 440.0, 307.5, 272.0, 313.0, 212.0, 208.0, 234.5, ~
$ density             <dbl> 0.99314, 0.99240, 0.99452, 0.99460, 1.00241, 0.99~
$ pH                  <dbl> 3.44, 3.24, 3.53, 3.18, 3.44, 2.90, 3.15, 3.55, 3~
$ sulphates           <dbl> 0.64, 0.37, 0.53, 0.59, 0.43, 0.38, 0.47, 0.63, 0~
$ alcohol             <dbl> 10.5, 11.0, 9.6, 10.5, 8.9, 11.0, 10.0, 9.4, 9.4,~
$ quality             <dbl> 3, 3, 4, 5, 5, 3, 4, 3, 6, 5, 6, 6, 7, 8, 8, 5, 5~
$ quality_category    <chr> "low", "low", "medium", "medium", "medium", "low"~
```

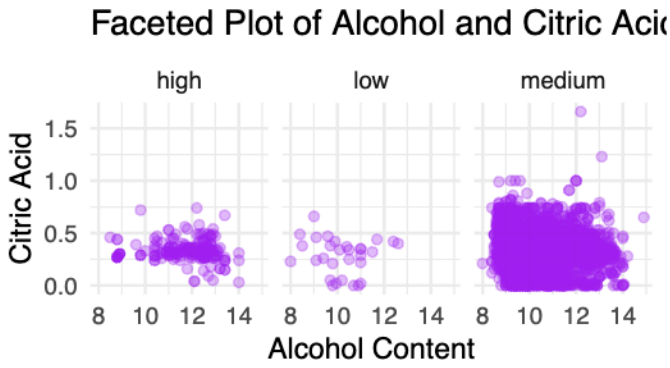
Exploratory Plots:

What role do `alcohol` content and `citric acid` play in determining high-quality wine?

```
plot1 <- ggplot(wine_data, aes(x = quality_category, y = alcohol)) +
  geom_boxplot() + labs(x = "Quality Category", y = "Alcohol Content")
plot2 <- ggplot(wine_data, aes(x = quality_category, y = citric_acid)) +
  geom_boxplot() + labs(x = "Quality Category", y = "Citric Acid")
plot_grid(plot1, plot2, ncol = 2, rel_widths = c(1, 1))
```

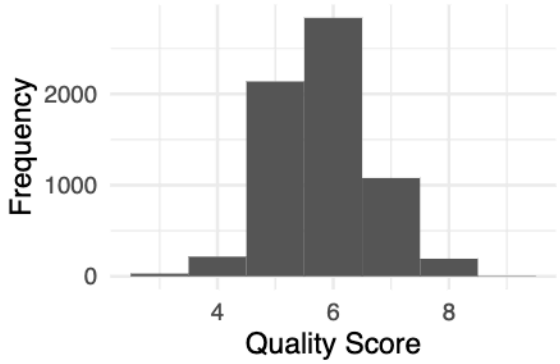



```
ggplot(wine_data, aes(x = alcohol, y = citric_acid)) +
  geom_point(alpha = 0.3, color = "purple") + facet_wrap(~ quality_category) +
  labs(title = "Faceted Plot of Alcohol and Citric Acid by Quality Category",
    x = "Alcohol Content", y = "Citric Acid") + theme_minimal()
```



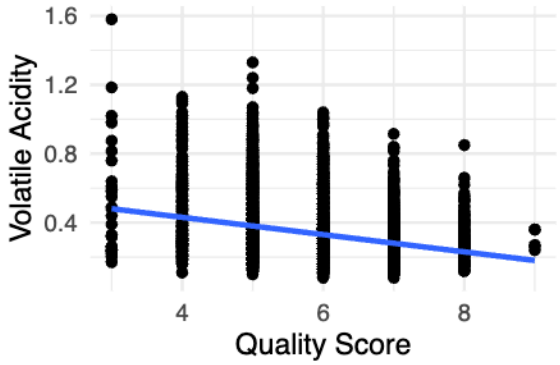
How do `volatile acidity` and `total sulfur dioxide` levels interact to shape the wine quality score?

```
# Histogram of Quality Scores
ggplot(wine_data, aes(x = quality)) + geom_histogram(binwidth = 1) +
  labs(x = "Quality Score", y = "Frequency") + theme_minimal()
```



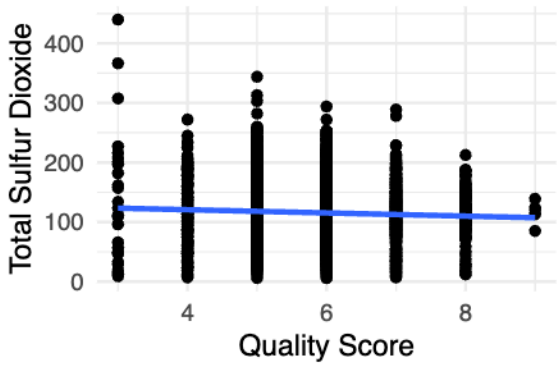

```
# Scatterplot of Quality vs. Volatile Acidity with Linear Trendline
ggplot(wine_data, aes(x = quality, y = volatile_acidity)) + geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Quality Score", y = "Volatile Acidity") +theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

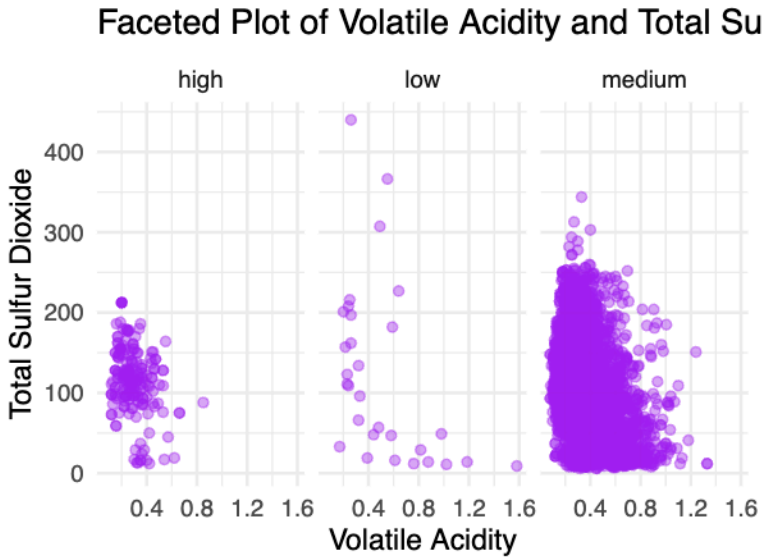



```
# Scatterplot of Quality vs. Total Sulfur Dioxide with Linear Trendline
ggplot(wine_data, aes(x = quality, y = total_sulfur_dioxide)) + geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Quality Score", y = "Total Sulfur Dioxide") + theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```
facet_plot <- ggplot(wine_data, aes(x = volatile_acidity, y = total_sulfur_dioxide)) +
  geom_point(alpha = 0.4, color = "purple") +facet_wrap(~ quality_category) +
labs(title = "Faceted Plot of Volatile Acidity and Total Sulfur Dioxide by Quality Category"
       x = "Volatile Acidity",y = "Total Sulfur Dioxide") +theme_minimal()
facet_plot
```



Dataset 3

Dataset Name: Student Performance

Data source: UC Irvine Machine Learning Repository (November 26, 2014), Cortez, P. (2008). Student Performance [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T.

Brief description: This dataset examines student achievement in secondary education at two Portuguese schools. Its attributes include student grades, demographic, social, and school-related features, and it was collected using school reports and questionnaires. The data specifically focuses on performance in Portuguese language classes. In [Cortez and Silva, 2008], the data set was modeled under binary/five-level classification and regression tasks.

Observations: The dataset contains 649 observations and 33 columns. Each row represents a unique student and captures a range of demographic, familial, academic, and personal attributes, including demographics, family background, academic details, and personal and social aspects. These attributes provide insights into each student's background and academic progress, making the dataset useful for analyzing factors that influence academic performance.

Research question 1: How does the amount of study time relate to final grade performance (Grade_3), and does it vary based on family support?

- **Outcome Variable**: Grade 3 (Continuous)
- **Primary Independent Variables**: Study time (continuous) and Family support (nominal, with interaction)

Research question 2: Is there an association between family relationship quality and school absences?

- **Outcome Variable**: school absences (ordinal)
- **Primary Independent Variable**: family relationship (ordinal)

Load the data and provide a `glimpse()`:

```
student_data <- read_delim(
  paste0(
    "https://github.com/Ramil-cyber/Student_data/",
    "raw/refs/heads/main/",
    "student_data.csv"
  ),
  delim = ";"
)
```

```
Rows: 649 Columns: 33
-- Column specification --------------------------------------------------------
Delimiter: ";"
chr (17): school, sex, address, family_size, Parent's_status, Mother's_job, ...
dbl (16): age, Mother's_education, Father's_education, travel_time, study_ti...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(student_data)
```

```
Rows: 649
Columns: 33
$ school                <chr> "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "~
$ sex                   <chr> "F", "F", "F", "F", "F", "M", "M", "F", "M", "M",~
$ age                   <dbl> 18, 17, 15, 15, 16, 16, 16, 17, 15, 15, 15, 15, 1~
$ address               <chr> "U", "U", "U", "U", "U", "U", "U", "U", "U", "U",~
$ family_size           <chr> "GT3", "GT3", "LE3", "GT3", "GT3", "LE3", "LE3", ~
$ `Parent's_status`     <chr> "A", "T", "T", "T", "T", "T", "T", "A", "A", "T",~
$ `Mother's_education`  <dbl> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4~
$ `Father's_education`  <dbl> 4, 1, 1, 2, 3, 3, 2, 4, 2, 4, 4, 1, 4, 3, 2, 4, 4~
$ `Mother's_job`        <chr> "at_home", "at_home", "at_home", "health", "other~
$ `Father's_job`        <chr> "teacher", "other", "other", "services", "other",~
$ reason                <chr> "course", "course", "other", "home", "home", "rep~
$ guardian              <chr> "mother", "father", "mother", "mother", "father",~
$ travel_time           <dbl> 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 2, 1, 1, 1~
$ study_time            <dbl> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3~
$ failures              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ school_support        <chr> "yes", "no", "yes", "no", "no", "no", "no", "yes"~
$ family_support        <chr> "no", "yes", "no", "yes", "yes", "yes", "no", "ye~
$ extra_paid_classes    <chr> "no", "no", "no", "no", "no", "no", "no", "no", "~
$ activities            <chr> "no", "no", "no", "yes", "no", "yes", "no", "no",~
$ nursery_school        <chr> "yes", "no", "yes", "yes", "yes", "yes", "yes", "~
$ higher_school         <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", ~
$ internet_access       <chr> "no", "yes", "yes", "yes", "no", "yes", "yes", "n~
$ romantic              <chr> "no", "no", "no", "yes", "no", "no", "no", "no", ~
$ family_relationship   <dbl> 4, 5, 4, 3, 4, 5, 4, 4, 4, 5, 3, 5, 4, 5, 4, 4, 3~
$ free_time             <dbl> 3, 3, 3, 2, 3, 4, 4, 1, 2, 5, 3, 2, 3, 4, 5, 4, 2~
$ go_out                <dbl> 4, 3, 2, 2, 2, 2, 4, 4, 2, 1, 3, 2, 3, 3, 2, 4, 3~
$ workday_alcohol       <dbl> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ weekend_alcohol       <dbl> 1, 1, 3, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2~
$ healthy_status        <dbl> 3, 3, 3, 5, 5, 5, 3, 1, 1, 5, 2, 4, 5, 3, 3, 2, 2~
```

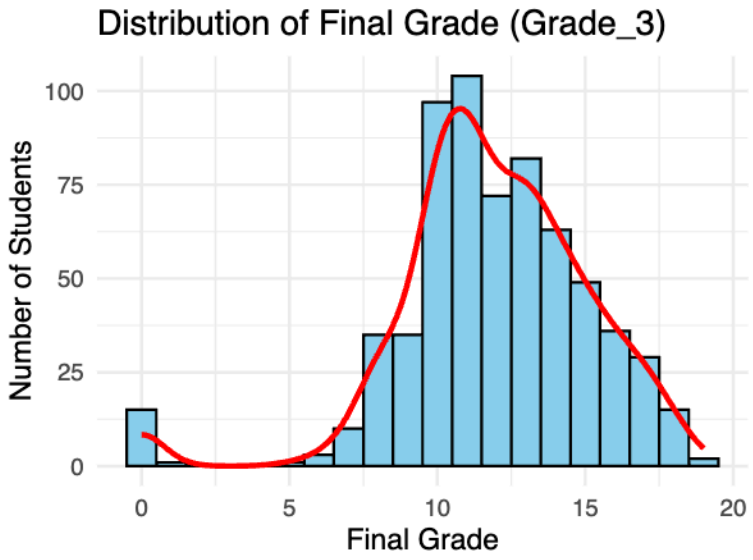
```
$ school_absences      <dbl> 4, 2, 6, 0, 0, 6, 0, 2, 0, 0, 2, 0, 0, 0, 0, 6, 1~
$ Grade_1              <dbl> 0, 9, 12, 14, 11, 12, 13, 10, 15, 12, 14, 10, 12,~
$ Grade_2              <dbl> 11, 11, 13, 14, 13, 12, 12, 13, 16, 12, 14, 12, 1~
$ Grade_3              <dbl> 11, 11, 12, 14, 13, 13, 13, 13, 17, 13, 14, 13, 1~
```

Exploratory Plots

How does the amount of study time relate to final grade performance (Grade_3), and does it vary based on family support?

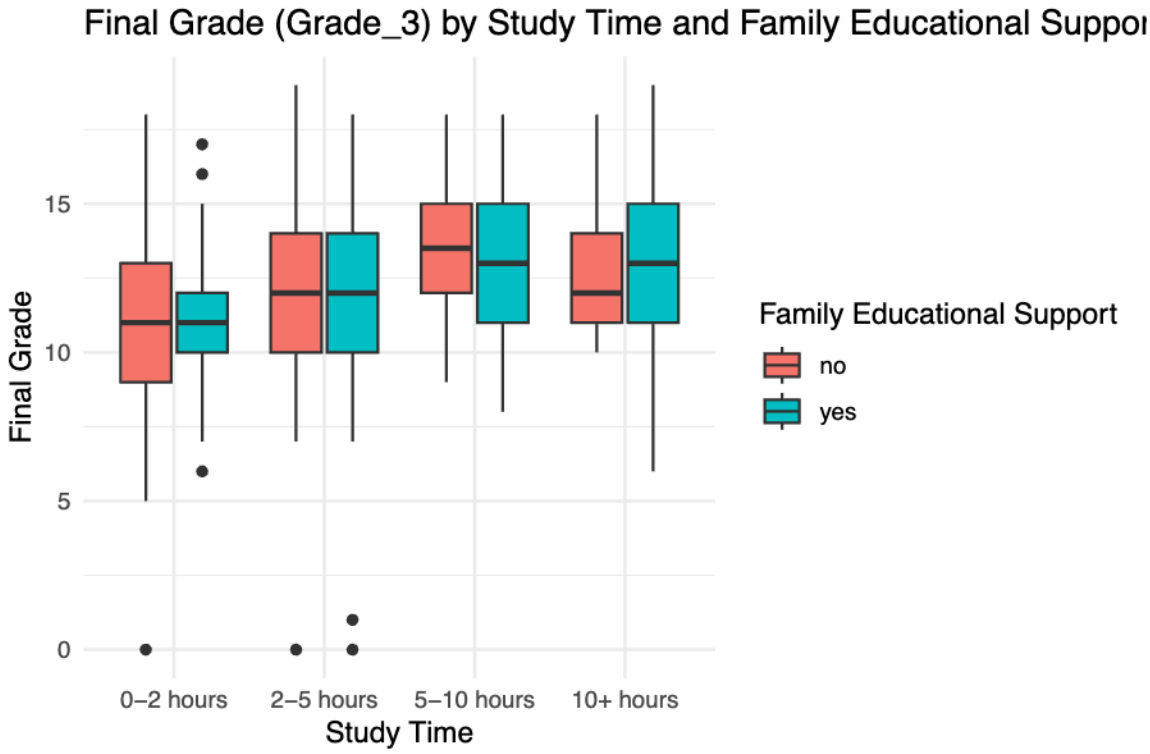
Grade Distribution: The distribution of `Grade_3` shows a range of final grades, suggesting variations in student performance.

```
ggplot(student_data, aes(x = Grade_3)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  geom_density(aes(y = after_stat(count)), color = "red", linewidth = 1) +
  labs(title = "Distribution of Final Grade (Grade_3)", x = "Final Grade",
       y = "Number of Students") +theme_minimal()
```



Grade and Study Time with Family Support: The box plot highlights the interaction between study time and family support, suggesting that the positive impact of family support is most evident when students already study for at least 5 hours. This implies that family support is effective when it supplements a student's efforts. In the lowest study time group (0-2 hours), grades are generally low and variable, regardless of family support, indicating that without a minimum level of study time, family support alone is insufficient to drive high academic achievement.

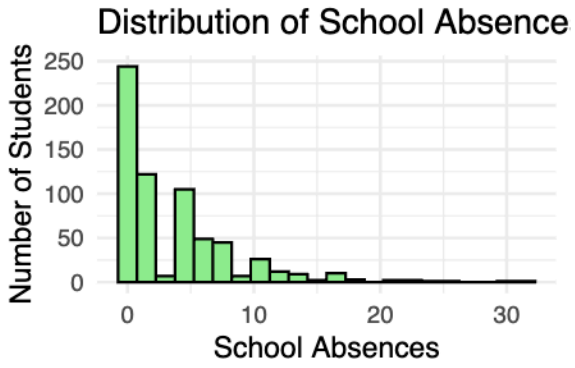
```r
student_data$study_time_binned <- cut(student_data$study_time,
                                      breaks = 4,
                                      labels = c("0-2 hours", "2-5 hours",
                                                 "5-10 hours", "10+ hours"))
ggplot(student_data, aes(x = study_time_binned, y = Grade_3, fill = family_support)) +
  geom_boxplot() +
  labs(title = "Final Grade (Grade_3) by Study Time and Family Educational Support",
       x = "Study Time", y = "Final Grade", fill = "Family Educational Support") +
  theme_minimal()
```



Is there an association between family relationship quality and school absences?

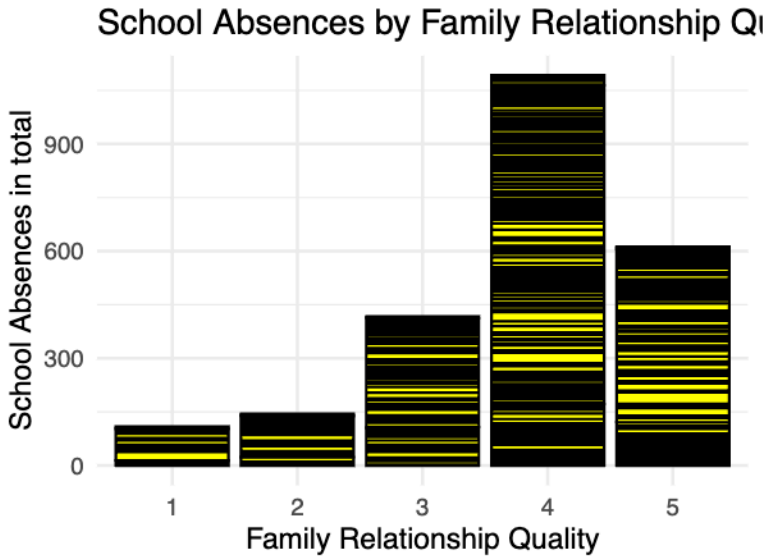
School Absences: The distribution of `school_absences` is right-skewed, indicating that most students have relatively few absences.

```r
ggplot(student_data, aes(x = school_absences)) +
  geom_histogram(binwidth = 1.5, fill = "lightgreen", color = "black") +
  labs(title = "Distribution of School Absences", x = "School Absences",
       y = "Number of Students") +theme_minimal()
```



School Absences and Family Relationship: The bar chart shows how family relationship quality influences school absences. Strong family relationships (category 5) may support better attendance, while moderate relationships (category 4) are linked to higher absences, possibly due to emotional or logistical challenges. Interestingly, students with very poor family relationships (categories 1 and 2) have the fewest absences, perhaps due to discomfort at home or a lack of support to address personal issues by missing school.

```
ggplot(student_data, aes(x = as.factor(family_relationship), y = school_absences)) +
  geom_bar(stat = "identity", fill = "yellow", color = "black") +
  labs(title = "School Absences by Family Relationship Quality",
       x = "Family Relationship Quality", y = "School Absences in total") +
  theme_minimal()
```



Team Charter

When will you meet as a team to work on the project components? Will these meetings be held in person or virtually?

Regular Meetings:

- **Timing:** Every Tuesday.
- **Format:** In-person until Nov 27. After Nov 27, meetings will be held virtually to accommodate end-of-semester schedules.

Additional Meetings:

- **Virtual Kickback Calls:** These will be scheduled as needed to discuss project details outside of regular Tuesday meetings.

What is your group policy on missing team meetings (e.g., how much advance notice should be provided)?

Policy on Missing Team Meetings:

- Members are required to give **at least 24 hours' notice** if they anticipate missing a meeting, except in emergencies. This allows the team to reassign tasks if needed.

How will your team communicate (email, Slack, text messages)? What is your policy on appropriate response time (within a certain number of hours? Nights/weekends?)?

Communication Channels and Response Times:

- **Primary Communication:** Slack will be used for regular discussions and document sharing.
- **Emergency Communication:** WhatsApp/Text Messages will be used for urgent situations.
- **Response Time:** Team members are expected to respond within **12 hours on weekdays**. Responses on nights and weekends are encouraged but not mandatory, except in urgent situations or near project deadlines.