Group contract

Group number: L21G05
Name: Lakshya Sakhuja
GitHub link: https://github.sydney.edu.au/mwan0680/L21G05-GROUP.git

I agree to:

- Abide by the terms of this contract in relation to the group assessment for DATA2002/2902.
- Store all my written and code contributions to the assessment in the GitHub repository.
- Keep a record of my other contributions to the assessment (e.g. discussions, emails, meetings attended). A copy of this may be requested by the coordinator.
- Work cooperatively, treat each other with respect, act honestly and ethically and not engage in any activities that could be perceived as bullying or harassment, as detailed in the Student Charter
- Communicate in two main ways: informal discussions on Slack and using the "Issues" functionality on GitHub to provide updates on specific tasks, including tagging responsibility to specific group members.
- Treat group members with respect and work collaboratively and contribute equitably to group work.
- Check the Whatsapp group multiple times daily and check in with GitHub at least twice a week and more regularly as we get closer to the deadline. If something on GitHub is urgent, it will be highlighted on the Whatsapp Group.
- Attend labs in the weeks before the tasks are due and meet for lunch on the day of the lab to give us time to informally discuss any issues we're facing. Other meetings will be held via Zoom or Google Meet.
- Communicate proactively if you're unable to complete a task or will be unavailable for a period, so the group can adjust timelines and redistribute work if needed.
- Review each other's work before submission to ensure consistency in style, accuracy, and alignment with the assignment requirements.
- Respect deadlines and shared schedules, ensuring all individual contributions are completed on time to allow adequate review and integration before submission.

I understand that:

- My agreement to these terms is indicated through the act of submitting this in Canvas.
- If I fail to meet my obligations as detailed in this group contract, then I have failed to meet the assessment requirements for DATA2002/2902 and may be awarded a mark of zero for some or all of the project components.

Exploratory data analysis

Data set: Student + Performance
Dependent variable: G3 (Final Grade)

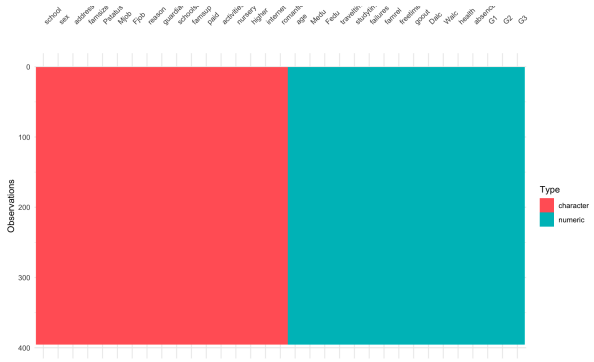
Data Loading Process

The analysis begins by loading the required libraries and importing the student performance datasets. The datasets are semicolon-separated files containing comprehensive information about students' academic performance and various demographic, family, and lifestyle factors.

```{r load-libraries, warning=FALSE, message=FALSE}
# Load required libraries
suppressWarnings({
  library(tidyverse)
  library(corrplot)
  library(gridExtra)
  library(VIM)
  library(ggplot2)
  library(dplyr)
  library(readr)
  library(scales)
  library(visdat)
})
```

```{r load-data}
# Load the datasets
math_data <- read_delim("student+performance/student/student-mat.csv", delim = ";")
portuguese_data <- read_delim("student+performance/student/student-por.csv", delim = ";")
```

Loading Libraries and Datasets



visdat Visualisation (No missing data)

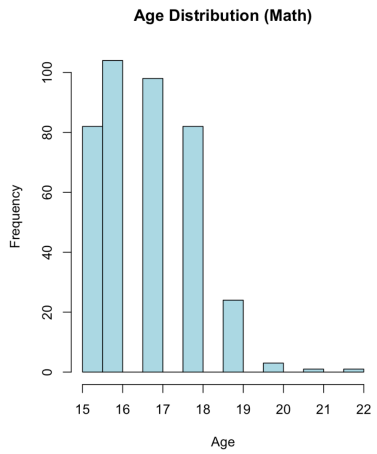
Student Characteristics and Demographics

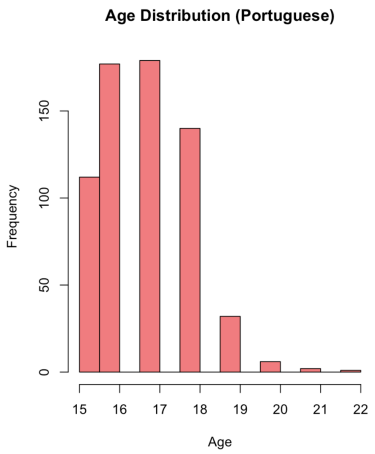
This comprehensive section examines the fundamental characteristics of students and their family backgrounds. The analysis explores age distributions, gender representation, school enrolment patterns, parental education levels, and family structure. These demographic factors provide crucial context for understanding student performance patterns and help identify potential relationships between background characteristics and academic outcomes. The visualisations reveal important insights about the student population composition and family socioeconomic factors that may influence educational achievement

The dataset contains **33 variables** including *demographics* (school, sex, age), *family background* (parent education, jobs), *academic factors* (study time, failures), and *lifestyle variables* (alcohol consumption, social activities). All variables are properly formatted with appropriate data types for analysis.

```
Math Dataset Summary:
    school              sex                 age          address
famsize            Pstatus              Medu
 Length:395         Length:395         Min.   :15.0   Length:395
Length:395         Length:395         Min.   :0.000
 Class :character   Class :character   1st Qu.:16.0   Class
:character   Class :character   Class :character   1st Qu.:2.000
 Mode  :character   Mode  :character   Median :17.0   Mode
:character   Mode  :character   Mode  :character   Median :3.000
                                      Mean   :16.7
Mean   :2.749
                                      3rd Qu.:18.0
3rd Qu.:4.000
                                      Max.   :22.0
Max.   :4.000
      Fedu            Mjob              Fjob               reason
guardian           traveltime      studytime
 Min.   :0.000   Length:395         Length:395
Length:395         Length:395         Min.   :1.000   Min.
:1.000
 1st Qu.:2.000   Class :character   Class :character   Class
:character   Class :character   1st Qu.:1.000   1st Qu.:1.000
 Median :2.000   Mode  :character   Mode  :character   Mode
:character   Mode  :character   Median :1.000   Median :2.000
 Mean   :2.522
```

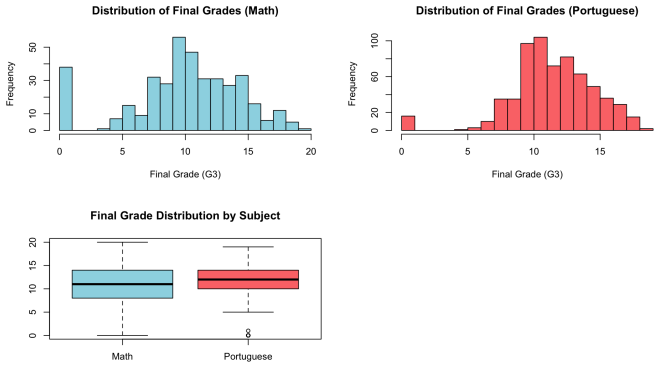
```
Portuguese Dataset Summary:
    school              sex                 age          address
famsize            Pstatus              Medu
 Length:649         Length:649         Min.   :15.00
Length:649         Length:649         Length:649         Min.
:0.000
 Class :character   Class :character   1st Qu.:16.00   Class
:character   Class :character   Class :character   1st Qu.:2.000
 Mode  :character   Mode  :character   Median :17.00   Mode
:character   Mode  :character   Mode  :character   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           traveltime      studytime
 Min.   :0.000   Length:649         Length:649
Length:649         Length:649         Min.   :1.000   Min.
:1.000
 1st Qu.:1.000   Class :character   Class :character   Class
:character   Class :character   1st Qu.:1.000   1st Qu.:1.000
 Median :2.000   Mode  :character   Mode  :character   Mode
:character   Mode  :character   Median :1.000   Median :2.000
 Mean   :2.307
```



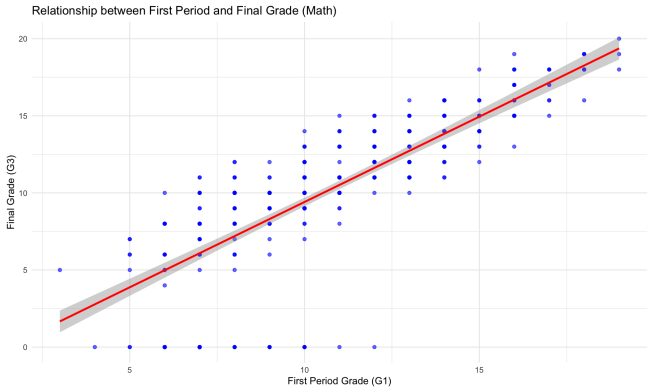


Target Variable Analysis

The analysis reveals significant differences between Math and Portuguese course performance. Portuguese students demonstrate superior academic outcomes with higher median grades and more consistent performance distribution. Math grades show a concerning bimodal pattern with a substantial peak at zero, indicating many students struggle significantly. The age distribution is consistent across subjects, with most students aged 16-17. The grade progression analysis shows strong correlation between periods, suggesting early performance is predictive of final outcomes. These findings highlight the need for targeted interventions in mathematics education and suggest Portuguese language instruction may be more effective or accessible to students.



Grade Progression (G1, G2, G3)



Grade Progression Scatter Plot (Math)

Conclusion

This exploratory data analysis examined 1,044 student records across mathematics and Portuguese courses from two Portuguese secondary schools. The analysis revealed significant performance differences between subjects, with Portuguese students consistently outperforming mathematics students. Key factors influencing academic achievement include gender, parental education, study habits, and lifestyle choices. The findings highlight the need for targeted interventions in mathematics education and early identification of at-risk students.

Key Findings:

- **Subject Performance Gap**: Portuguese students achieve higher grades with more consistent performance than mathematics students
- **Gender Differences**: Female students significantly outperform male students in both subjects
- **Family Impact**: Parental education levels strongly correlate with student academic success
- **Academic Factors**: Past failures and study time are key predictors of final performance
- **Early Prediction**: First-period grades strongly predict final outcomes, enabling early intervention

The analysis provides valuable insights for educational improvement, emphasising the need for mathematics curriculum reform, gender-sensitive teaching approaches, and family engagement programs to enhance student achievement.