Final Project

1. Student Data Set

The project data sets consist of data regarding students in secondary education of two Portuguese schools. There is one dataset for math and one for Portuguese. There are 33 variables on each data set. The data was collected by *Cortez, P. (2008). Student Performance [Dataset]. UCI Machine Learning Repository.* *https://doi.org/10.24432/C5TG7T*.

You should start your project by merging the two data sets.
The variables used to identify unique rows are:
by=c("school","sex","age","address","famsize","Pstatus","Medu","Fedu","Mjob","Fjob", "reason","nursery","internet")

After merging:
Please rename variables **G1**, **G2**, and **G3** to distinguish between math and Portuguese. The merging will create some other duplicate and redundant columns. Please remove these before beginning analysis.

The data set consists of the following variables:
1 **school** - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
2 **sex** - student's sex (binary: "F" - female or "M" - male)
3 **age** - student's age (numeric: from 15 to 22)
4 **address** - student's home address type (binary: "U" - urban or "R" - rural)
5 **famsize** - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
6 **Pstatus** - parent's cohabitation status (binary: "T" - living together or "A" - apart)
7 **Medu** - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8 **Fedu** - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9 **Mjob** - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
10 **Fjob** - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
11 **reason** - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
12 **guardian** - student's guardian (nominal: "mother", "father" or "other")
13 **traveltime** - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 **studytime** - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 **failures** - number of past class failures (numeric: n if 1<=n<3, else 4)
16 **schoolsup** - extra educational support (binary: yes or no)
17 **famsup** - family educational support (binary: yes or no)

18 **paid** - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 **activities** - extra-curricular activities (binary: yes or no)
20 **nursery** - attended nursery school (binary: yes or no)
21 **higher** - wants to take higher education (binary: yes or no)
22 **internet** - Internet access at home (binary: yes or no)
23 **romantic** - with a romantic relationship (binary: yes or no)
24 **famrel** - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 **freetime** - free time after school (numeric: from 1 - very low to 5 - very high)
26 **goout** - going out with friends (numeric: from 1 - very low to 5 - very high)
27 **Dalc** - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 **Walc** - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 **health** - current health status (numeric: from 1 - very bad to 5 - very good)
30 **absences** - number of school absences (numeric: from 0 to 93)
31 **G1** - first period grade (numeric: from 0 to 20)
31 **G2** - second period grade (numeric: from 0 to 20)
32 **G3** - final grade (numeric: from 0 to 20, output)

2. Final Project Evaluation

The final project includes two components: presentation and final project report. The percentage of each part to your final project score is given below:

1. Presentations on April 28, 30, and May 2 during our regular meeting time in the class lecture room. Each group will give a 15-20-minute presentation. This part contributes 30 percent to your final project score.
2. Final project report and the corresponding R code is due May 5, 2025. The final project report and the corresponding R code contribute 40 percent to your final project score. The final project report and R code will be evaluated based on the details given below.
3. The models and methods you explored and implemented in your final project contribute 30 percent to your final project. This will be evaluated based on the appropriateness of your models and methods, and the number of ways attempted.

3. Final Project Report and its Associated R code

The final project report is due May 5, 2025. The final project report contributes 40 percent to your final project score. Each team should only submit one report and one R code file. However, please specify the contribution from each team member clearly so that we can evaluate according to your contribution. Please submit your final project report and R code to Gradescope.com.

In your final project report, you should include the following components and list them clearly in each section of your report. You might not receive the corresponding credits if one of the following components is missing in your report and R code.

1. The practical or scientific questions your group would like to address and the corresponding solutions using statistical and machine learning methods.
2. Preliminary exploratory data analysis (such as graphs, plots, and summary statistics) that guide you to choose appropriate methods or models or provide some answers to your raised questions.
3. For each method you try and implement: (i) provide the corresponding model (if applicable). For example, if you apply a logistic regression model, please specify your outcome and predictors and specific models you fit in math equations; (ii) details about the variable selection and model selection steps; (iii) details of tuning parameter selection; (iv) details about your model fitting results including parameter estimations, statistical significance or fitted models; (v) interpretation of your model and how it addresses your raised questions; (vi) conclusion and any aspect for improvement.
4. To compare different methods, please evaluate the performance of each method and use the same training and test data set to evaluate the performance. To ensure reproducibility of your results, please set seeds so that your results can be reproduced. Recommend a few methods that you think are most appropriate and perform best to address your raised questions.

In the R code corresponding to your final project report, we will evaluate the following aspects. If one of the following is not satisfied, you might not receive the corresponding credits.

1. The R code is correctly written for your chosen methods.
2. The results presented in your final project report can be reproduced.
3. Your R code can run smoothly without errors.
4. Clear and detailed comments on each part of the code.
5. Code is efficient in time and memory.