Group: Alexander Tan, Andreea Ghenciu, Mridula Srivathsan, Nikhita (Nikki) Nair, Rishav Roy

Data: https://www.kaggle.com/datasets/uciml/student-alcohol-consumption

Code:
```
import pandas as pd
pd.read_csv("alcconsumption/student-por.csv")
```

Questions:

1. What factors have the greatest correlation with alcohol consumption in secondary students?
2. Which factors have a positive/negative correlation (risk factors vs. protective factors)?

Variables of Interest:

1. sex - student's sex (binary: 'F' - female or 'M' - male)
2. age - student's age (numeric: from 15 to 22)
3. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
4. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
5. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
6. Fedu - father's education (same as mother's)
7. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
8. Fjob - father's job (same as mother's)
9. studytime - weekly study time (numeric: 1 - <2 hrs, 2 - 2 to 5 hrs, 3 - 5 to 10 hrs, or 4 - >10 hrs)
10. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
11. famsup - family educational support (binary: yes or no)
12. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
13. internet - Internet access at home (binary: yes or no)
14. romantic - with a romantic relationship (binary: yes or no)
15. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
16. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
17. **Dalc (response)** - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
18. **Walc (response)** - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
19. health - current health status (numeric: from 1 - very bad to 5 - very good)
20. absences - number of school absences (numeric: from 0 to 93)
21. G3 - final grade (numeric: from 0 to 20)

All 33 variables will be considered during model selection, but these seem most relevant

Methods:

1. Lasso (or possibly ridge) regression model for feature selection
    a. We will decide if we want a sparse model (lasso) or a model with many covariates (ridge) after further data exploration
2. Using the features selected from the feature selections, we will run a multiple linear regression model to determine the positive/negative correlation (risk/protective factors) with weekend and weekday alcohol consumption