**Overview**

Student social habits and demographic information can often influence academic performance. More specifically, parental, extracurricular, academic, and social habit information could provide insight on student performance in school. For this analysis, I am aiming to determine if student alcohol consumption habits with demographic information can be used to predict final grade.

**Research questions**

The goal of this project is to determine if student grade can be predicted with information on demographics, academics, extracurriculars, social, and other factors. Some of the research questions include:

* What variable is the strongest predictor of student grade?
* How does alcohol consumption influence grade? Does weekend or weekday consumption play a larger role?
* What other interesting findings can be gathered from the data? Are there certain interesting variables that do not seem to influence grade at all?

**Data**

The data dictionary can be found at <https://www.kaggle.com/uciml/student-alcohol-consumption#student-mat.csv>. A different dataset combines the data from the two classes from the original link and adds a few more relevant variables for my analysis (<https://www.kaggle.com/joeycorea/mathportugeseyouthalcoholstudy>). First, a variable representing overall grade is added, aggregating the original three periodic grades into one grade where the most recent grade is weighted higher. Second, two binary variables are added for heavy drinkers (high alcohol consumption on weekday and weekend) and binge drinkers (high weekend but low weekday alcohol consumption scores).

This dataset also includes all of the original variables to total 50 features on 1,044 students from two schools. Demographic information given includes age, sex, and urban/rural address. Information on parents and family is also given, including parent status, education, and job category for both mothers and fathers in addition to family size and quality of family relationship. Academic-related features include travel time to school, study time, number of class failures, class absences, extra school support, extra paid classes, extracurricular activities, and higher education consideration. The last main category is social information, including how often the student goes out, weekday and weekend alcohol consumption, and the previously explained binary variables for binge and heavy drinkers. However, problems may be encountered as there are many variables in this dataset. Second, some variables are likely to be related. Thus, variable selection will likely be particularly important. The selected variables could be used to predict overall grade.

A previously published paper (Cortez, 2008) using this dataset focused on how different models (i.e. Decision Tree, Random Forest, Neural Networks, etc.) predicted grades but included all data in modeling, including previous period grades. In contrast, I will focus on identifying features that best predict overall grades rather than specific periodic grades using linear regression.

**Project plan**

Before beginning modeling, a significant amount of EDA needs to be done to have a better understanding of what variables may have interesting relationships with grade and with each other. This will inform variable selection for modeling. In order to predict overall grade, a linear model will first be explored to see how changes in each of the chosen variables influence grade. Because variable selection is important with this data, I plan on spending approximately two weeks exploring and choosing variables to better understand the possible variables. Then, I will spend the next week exploring models, focusing mostly on linear regression. This general plan ensures results ready before November 26th.