STAT 139 Project Proposal

Group Members: Chia Chi (Michelle) Ho, Yuyue Wang, Xiangru Shu, Chengye Liu
Project Title: How do social relationships affect high school student math grade?
Data: Student Performance Data Set
(https://archive.ics.uci.edu/ml/datasets/Student+Performance#)

Project Overview: The ability to predict a student's school performance is lucrative, especially from the educators' perspective. It allows the school to identity students who may need more attention and/or assistance. A student's school performance is affected by many factors. In this study, we would like to explore this topic from a social standpoint and ask how social relationships affect high school student math grade. We aim to build a predictive model to predict math grade based on information about a student's social relationships.

Hypotheses of Interest:

1. Determine whether high school student math grade is associated with the status of his/her interpersonal/social relationships (i.e. family size, parent's cohabitation status, guardian, family educational support, romantic relationship, quality of family relationships, goout).
2. If there is a significant association, determine whether the association is still significant controlling for other factors.
3. Build a best predictive model to predict student math grade.

Variables of Interest:

Response: G1 - first period grade (numeric: from 0 to 20); G2 - second period grade (numeric: from 0 to 20); and G3 - final grade (numeric: from 0 to 20)

Predictors:

- sex - student's sex
- famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
- Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
- Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
- guardian - student's guardian (nominal: 'mother', 'father' or 'other')
- 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)
- studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- failures - number of past class failures (numeric: n if 1<=n<3, else 4)
- schoolsup - extra educational support (binary: yes or no)
- famsup - family educational support (binary: yes or no)
- paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

- activities - extra-curricular activities (binary: yes or no)
- higher - wants to take higher education (binary: yes or no)
- internet - Internet access at home (binary: yes or no)
- romantic - with a romantic relationship (binary: yes or no)
- famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- health - current health status (numeric: from 1 - very bad to 5 - very good)
- absences - number of school absences (numeric: from 0 to 93)

Analysis Plan: Exploratory analysis will be performed to determine if any data transformation is needed to meet linear regression assumptions. Data will be cleaned and transformed as necessary. For example, quantitative variables will be standardized/normalized, and right-skewness will be corrected with logarithmic (or other proper) transformation. Several predictors as listed above, with an emphasis on variables related to interpersonal/social relationships status (social variables), will be used to determine statistical association with the response variables of interest (student's G1, G2 and G3 grades). For instance, we will use a two-sample t-test and an ANOVA test to determine whether being in a romantic relationship or the status of a student's guardianship affect a student's math grade, respectively. Several regression models will be considered and explored. Specifically, we will build univariate linear regression models from each of the social variables, then we will use model selection (step forward or backward), regularization and cross-validation to build a best predictive model from the set of social variables and selected interaction terms. Finally, we will build the final, best predictive model, again using model selection, cross-validation and regularization, including all of the listed variables and selected interaction terms to determine what factors are most associated with high school student's math grade.