**CS 105 Final Project Proposal**

1. **Team Members:** Gabrielle French (gjfrench), Rocky Xie (rockyxie), Karam Yang (karamy)
2. **Dataset:** Our dataset, “Student Alcohol Consumption Data Set”, is from the UCI Machine Learning Repository and contains information of Secondary School Student

**URL (CSV)**: https://archive.ics.uci.edu/ml/machine-learning-databases/00356/

**URL (page)**: https://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION#

**Key Attributes:** school, sex, age, address, famsize, Pstatus, Medu, Fedu, Mjob, Fjob, reason, guardian, traveltime, studytime, failures, schoolsup, famup, paid, activities, nursery, higher, internet, romantic, famrel, freetime, goout, Dalc, Walc, health, absences, G1, G2, G3

1. **Data Mining:** This site features this dataset as both a study in factors influencing alcohol consumption in students (Dalc and Walc), as well as factors influencing the relative success of a student (failures, higher, G1, G2, G3). We may explore both of these channels. We will use *classification learning* to build a model that predicts either the alcohol consumption and/or the success of students (based on their social and economic advantages). *Association Learning* will be used to determine which attribute are most strongly correlated with success versus failure.
2. **Goal of our Model:** The goal of this data mining model is to analyze factors in a student’s abilities academically, and in their likelihood to drink; even further, we can find a relationship between the two results. Identifying challenges that students face in having successful academic careers and healthy lifestyles is important in extending opportunities to all students.
3. **Transformations:** We may keep all information in tact and in its current form, as all of the data has potential to be relevant to our goals, and all of the attributes are part of the same relationship; while we may look at alcohol consumption and success as two separate results, they are more than likely related factors. Potentially, Weekday (Dalc) and Weekend (Walc) alcohol consumption data could be combined; grades (G1, G2, G3) could potentially be averaged, so long as it does not affect the outcome too heavily. We may also be able to break down the tables into groups (Parents, Economics, Social, etc.), but it may be unnecessary.
4. **Other Components:**

Relational Tables: As of now, it seems to make sense for our data set to have one table featuring all of the data, which we would call StudentAttributes; other breakdowns may occur, as noted above.

Programs: SQLite to explore data, Python and SQL queries to find averages etc.

Visuals: Google Charts (pie chart, map, spin on a bar graph), possibly independent? We’d like to design figures for each department (ie, a cartoon teacher for the Education department) and size them proportionate to their relative incomes.

1. **Division of Work:**

Introduction: Gabrielle

Dataset description: Karam

Data preparation: All

Data analysis: All

Visuals: All

Results: Karam & Rocky  
 Conclusions & Appendix: Rocky & Gabby