INST737 – Spring 2016

Digging into Data

Project Report

Student Alcohol Consumption

Submitted by:

Ekta Ahuja, emahuja@umd.edu
Kristen Bertch, kristennbertch@gmail.com
Amar Kurane, akurane@umd.edu

Why this is A Good Idea

Drinking is a common and even expected occurrence for college students in the United States. Unfortunately, drinking can lead to many negative consequences. As explained by the National Institute on Alcohol Abuse and Alcoholism, "1,825 college students between the ages of 18 and 24 die each year from alcohol-related unintentional injuries. More than 690,000 students between the ages of 18 and 24 are assaulted by another student who has been drinking. More than 150,000 students develop an alcohol-related health problem and between 1.2 and 1.5 percent of students indicate that they tried to commit suicide within the past year due to drinking or drug use."

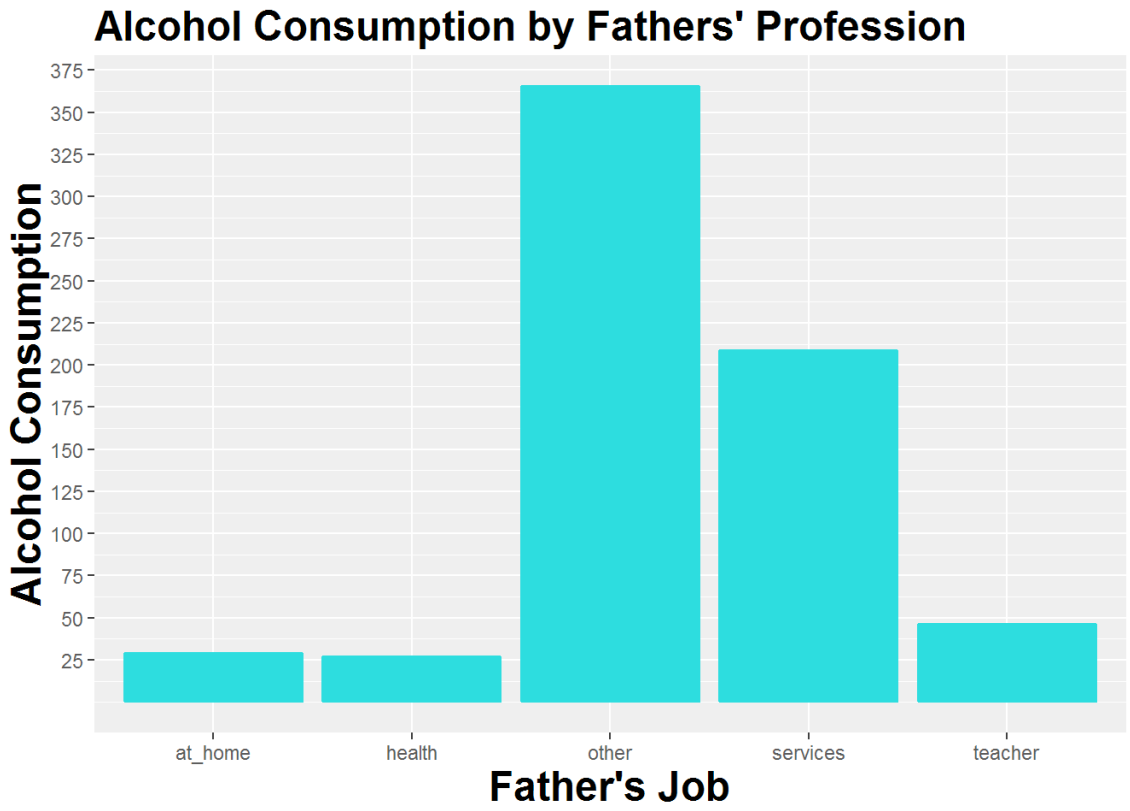
Drinking is a serious problem that many students are faced with. The purpose of this research project is to identify the signs that a student is likely to consume high amounts of alcohol. If schools and counselors know which groups of students are a high risk of drinking heavily, they will be able to focus their efforts on the students who need the most help. The more we know about high risk students, the more we will be able to help them make the right decisions.

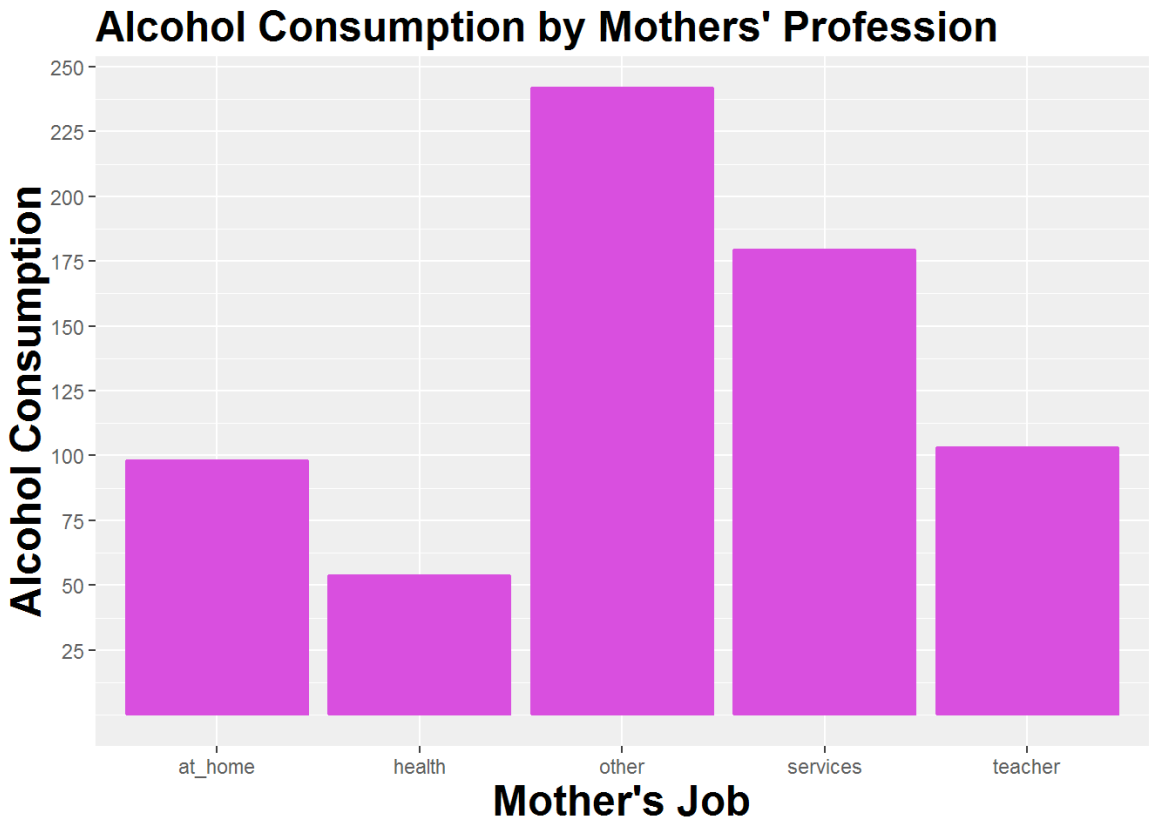
What We Did

Our first step was to combine the weekday drinking (Dalc column) and the weekend drinking (Walc) in order to get the average the student drinks throughout the week. We did this by multiplying the weekday-drinking column by five, multiplying the weekend drinking column by two, and then adding these numbers together. This gave us the average weekly consumption for each student.

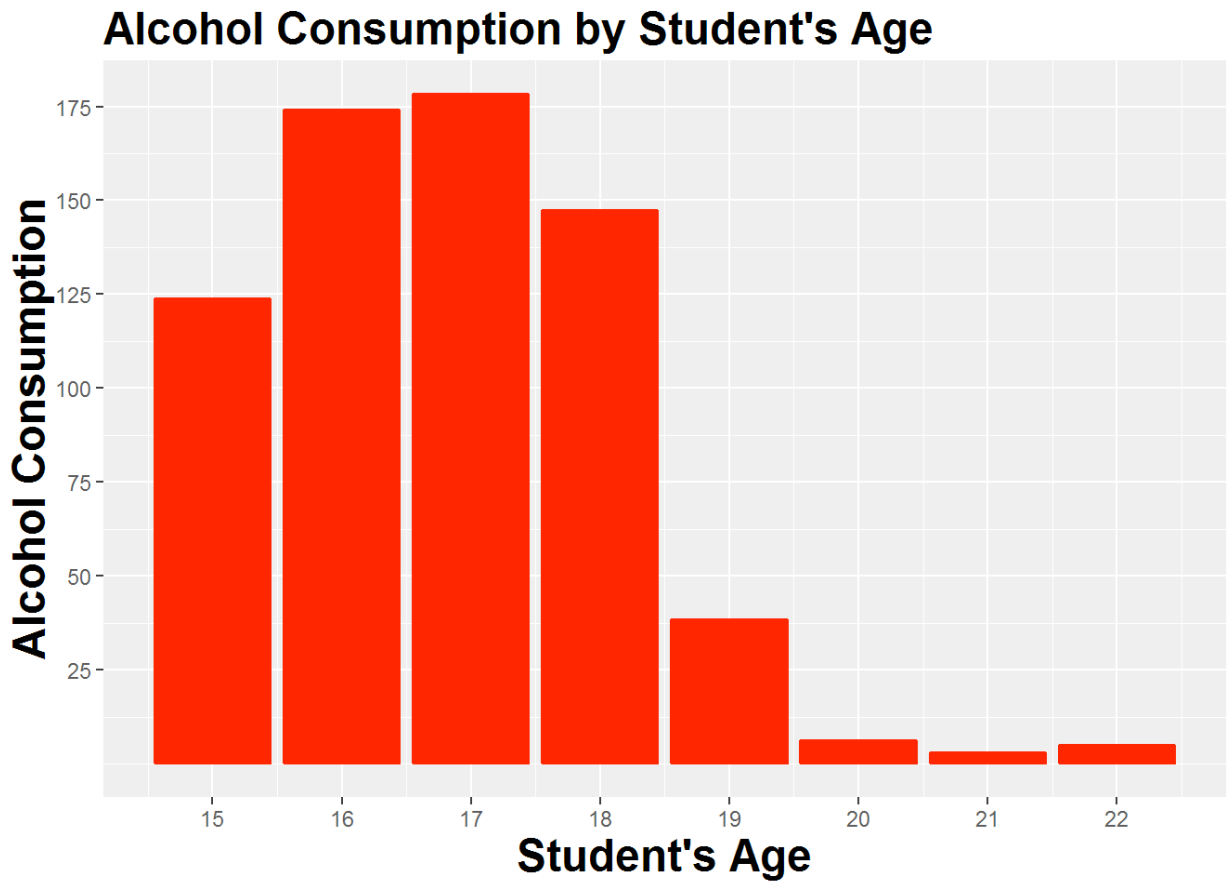
Next we made two categories: students who were likely to be heavy drinkers and students who were not. The drinking amounts are on a one to five scale. If students are from zero to three then they are in the not considered heavy drinkers (or for purposes of the write-up a "drinker"). If they are four to five then they are considered likely to drink heavily. We created the avg_consumption_class column where we defined these parameters.

After we created the columns specified above we then moved on to examining the dataset as a whole. First, we just looked at a summary of the data. We then decided to compare some of the individual variables to alcohol consumption in order to better understand our dataset. First we looked at alcohol consumption in relation to the type of jobs the students' fathers had and then the type of job the students' mothers had.


Alcohol Consumption by Fathers' Profession
Alcohol Consumption
375
350
325
300
275
250
225
200
175
150
125
100
75
50
25
at_home
health
other
services
teacher
Father's Job

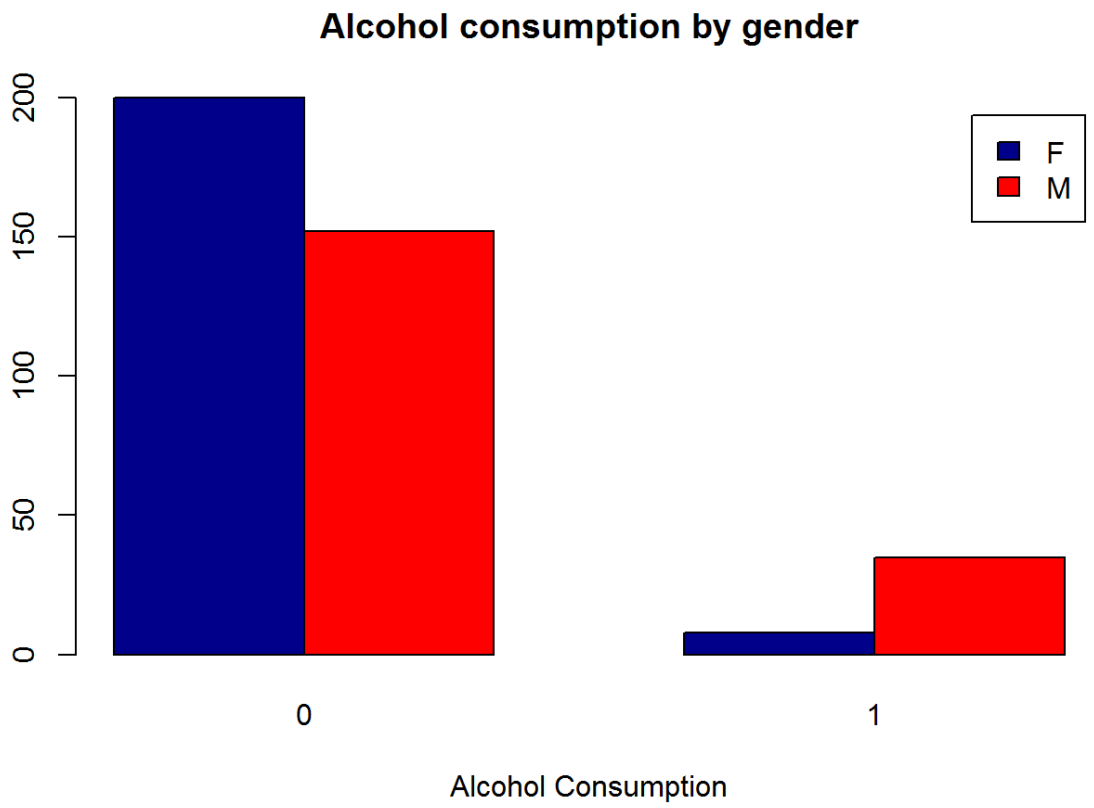

Alcohol Consumption by Mothers' Profession
Alcohol Consumption
250
225
200
175
150
125
100
75
50
25
at_home
health
other
services
teacher
Mother's Job

The information about the mother and fathers' jobs was not very enlightening because in both cases the highest level of alcohol consumption was in the "other" category. Next we looked at drinking based on age.



The graph above is very interesting. It shows that sixteen and seventeen-year-olds consume more alcohol than students who are closer to the age of twenty-one. Our report does not delve deeply into this topic, but a good research topic could be on exploring age specifically in relation to drinking. For example if young men and women really greatly decrease the amount they drink one it is legal, it may add some valuable facts to the debate concerning the drinking age. Of course, this would need to be explored further to make that type of argument.

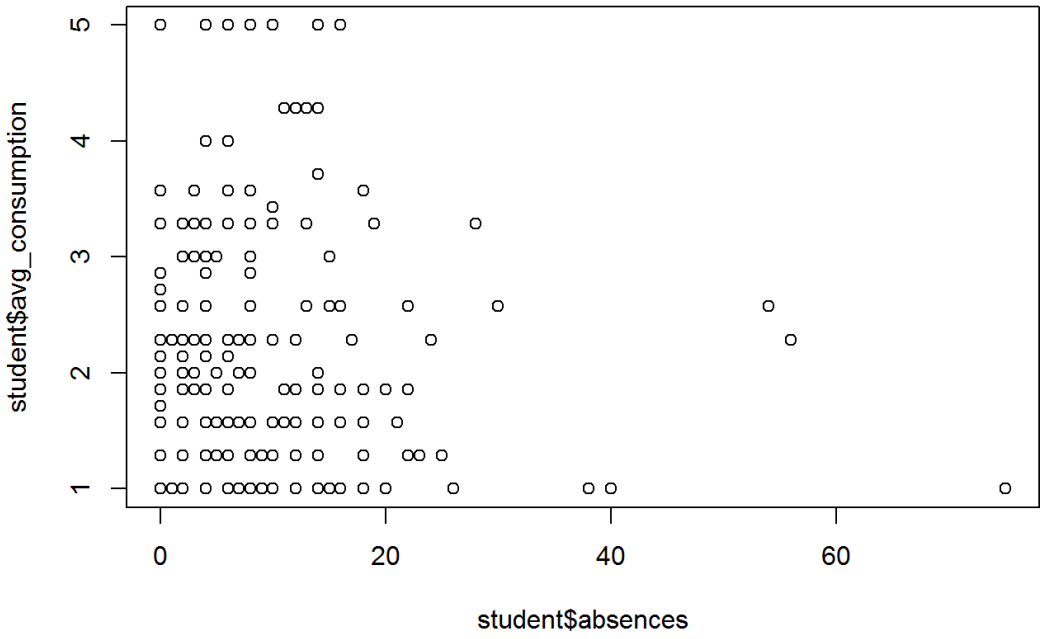
We proceeded to make multiple more graphs comparing different variables to alcohol consumption. As an example, I will display one of the graphs below and then I will simply explain the results of the others.



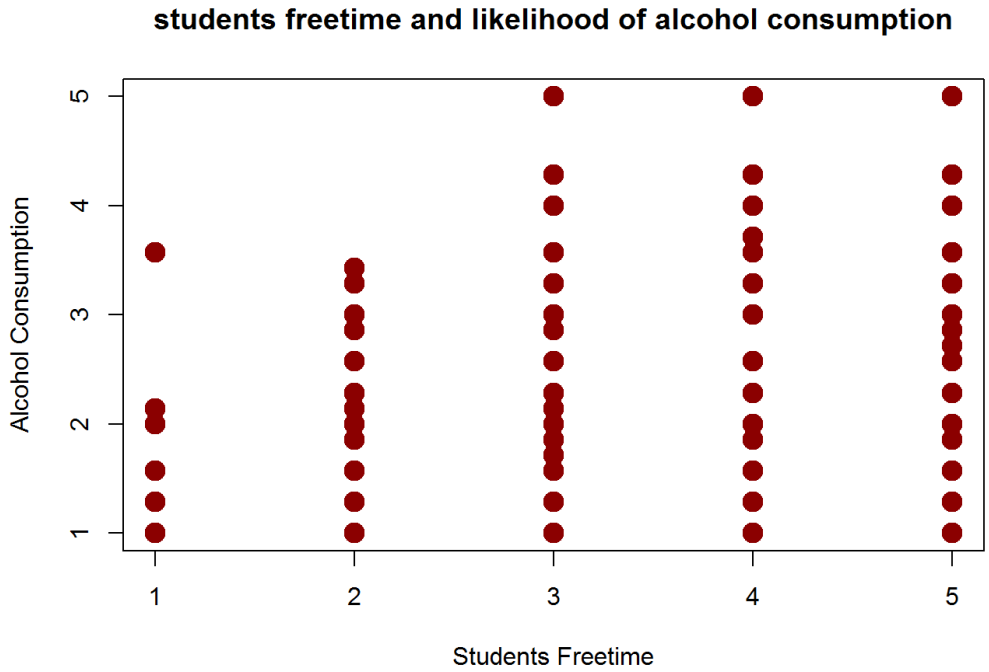
When looking at this chart the only side we care about is the side with the 1. The 1 column identifies the drinkers. So what this chart is showing is that of those who are drinkers, most of them are male. This means that gender is likely to be a good indication of whether a student will likely be a drinker. Of the drinkers, eight were female and thirty-five were male. This means that of the drinkers, about 81% were male.

Continuing to look at just the drinkers, we found indications that some variables were likely to be useful, whereas others were not. We found: 67% of drinkers were not in a romantic relationship, 93% of drinkers were not receiving school support, 54% of drinkers were receiving family educational support, 58% of drinkers were not involved in extra-curricular activities, 65% of drinkers came from an urban area, 84% of drinkers' parents were together, and 58% of drinkers had a family size that was greater than three.

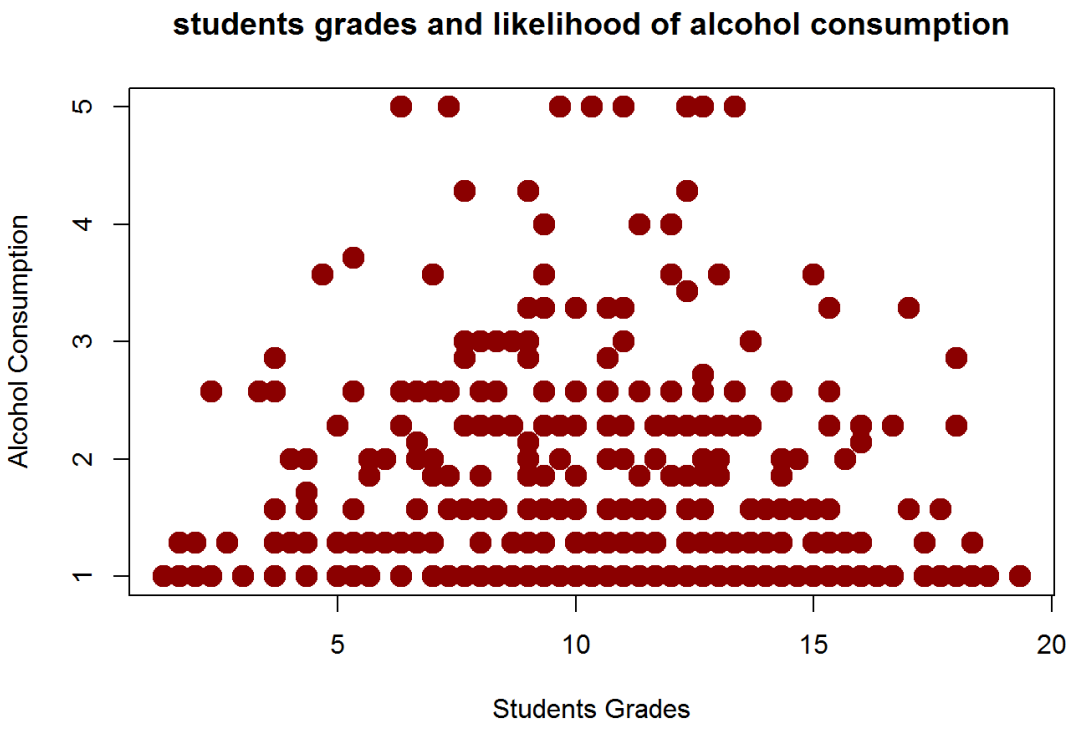
After finding the relationship between alcohol consumption and the variables above, we charted a few more variables so that we would have a good visualization of the data.



The number of absences goes from zero to ninety-three. Surprisingly, the data shows that the less a student is absent, the more likely he/she is a drinker.



The chart above shows that the more free time students had, the more likely they were to be a drinker.



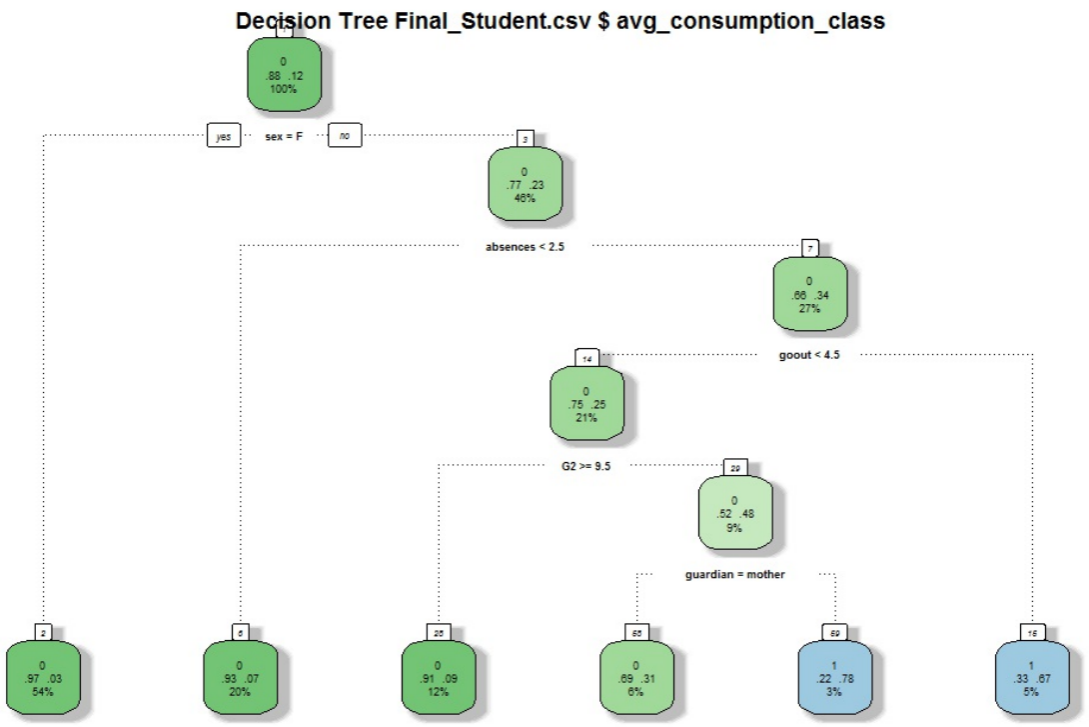
In this dataset, grades are on a scale from 1 to 20. This was another surprising finding, but it does not appear that there is a strong correlation between drinking and grades.

Next we did a linear regression to see which variable were most important to determining whether a student was likely to be a drinker.

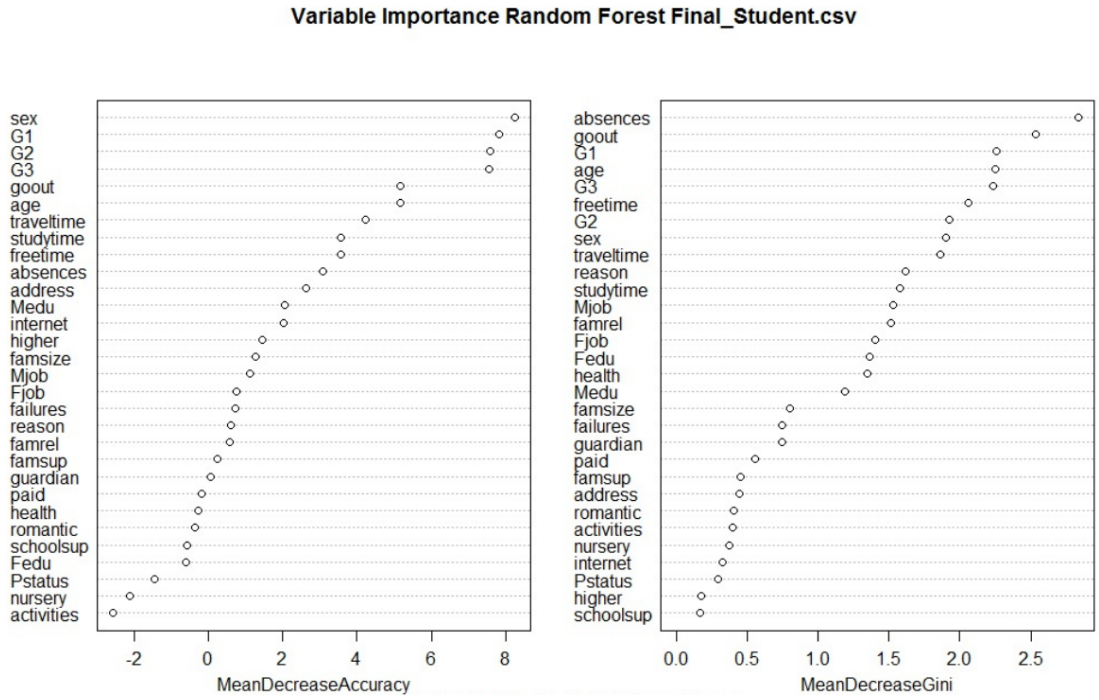
We deduced that the following variables has most effect on alcohol consumption:

1) Sex
2) reason (for going to that college)
3) paid (is getting paid)
4) activities (extra-curricular)
5) nursery (whether a nursery school was attended by the student)
6) famrel (quality of family relationship)
7) freetime
8) goout
9) absences

We also used a decision tree to show the interaction of these variables.



We then cross-referenced the above lm results using the results of our random forest below.

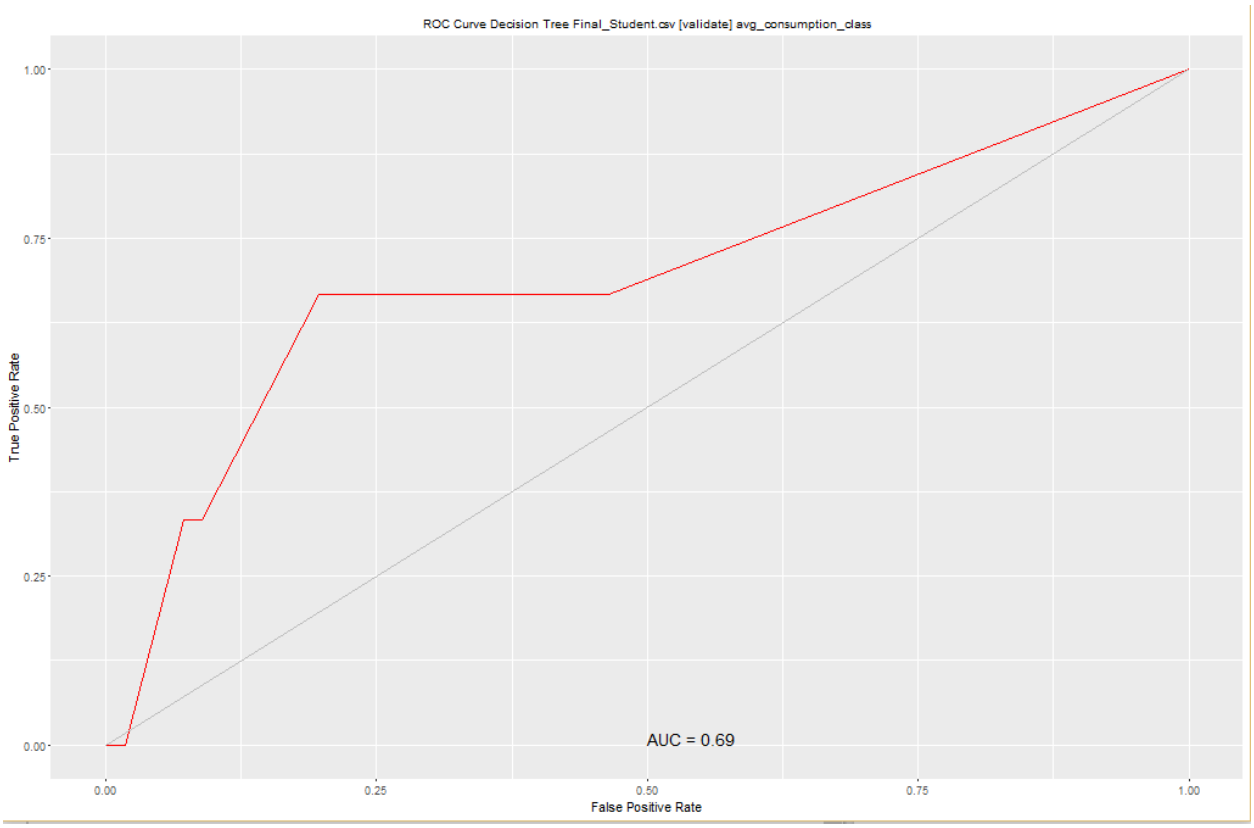


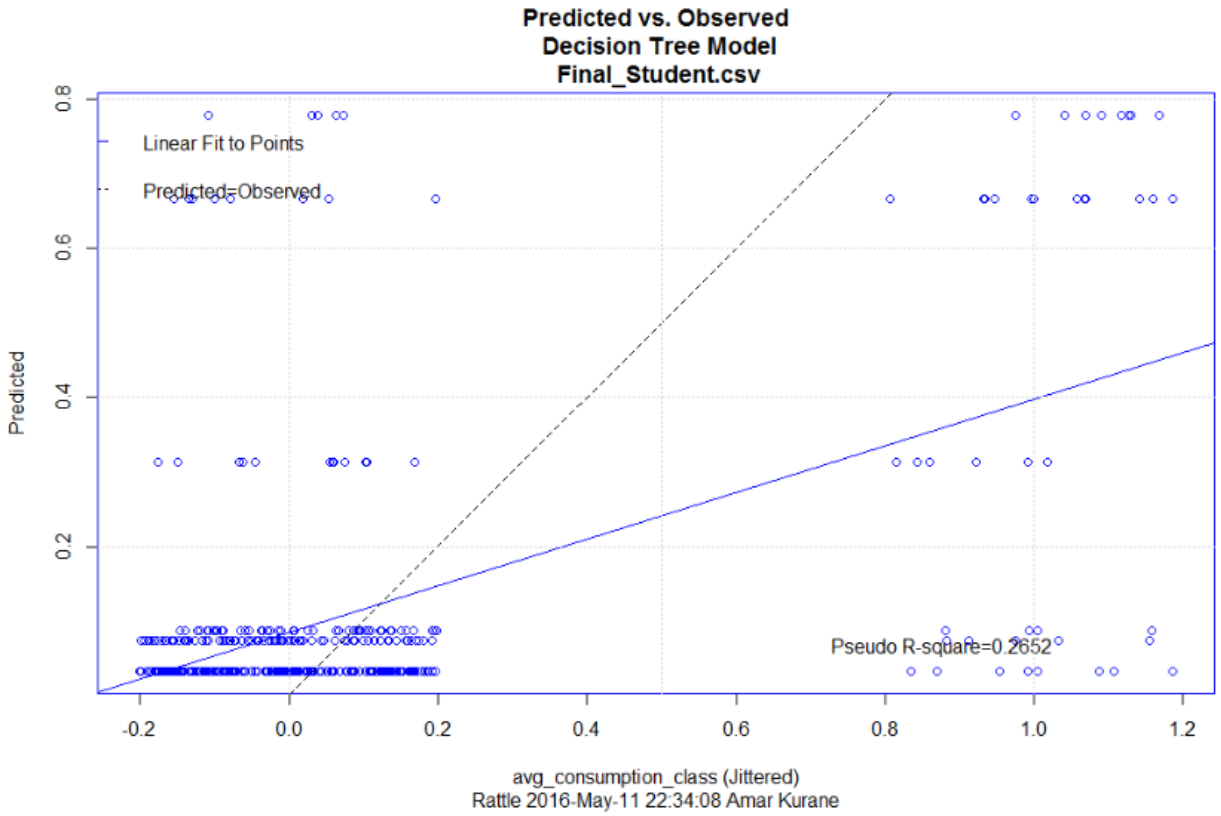
The random forest test highliged other variables, such as grades, that were not as prominent in our lm model. This indicated that we should consider these variables as well when doing our regression tests.

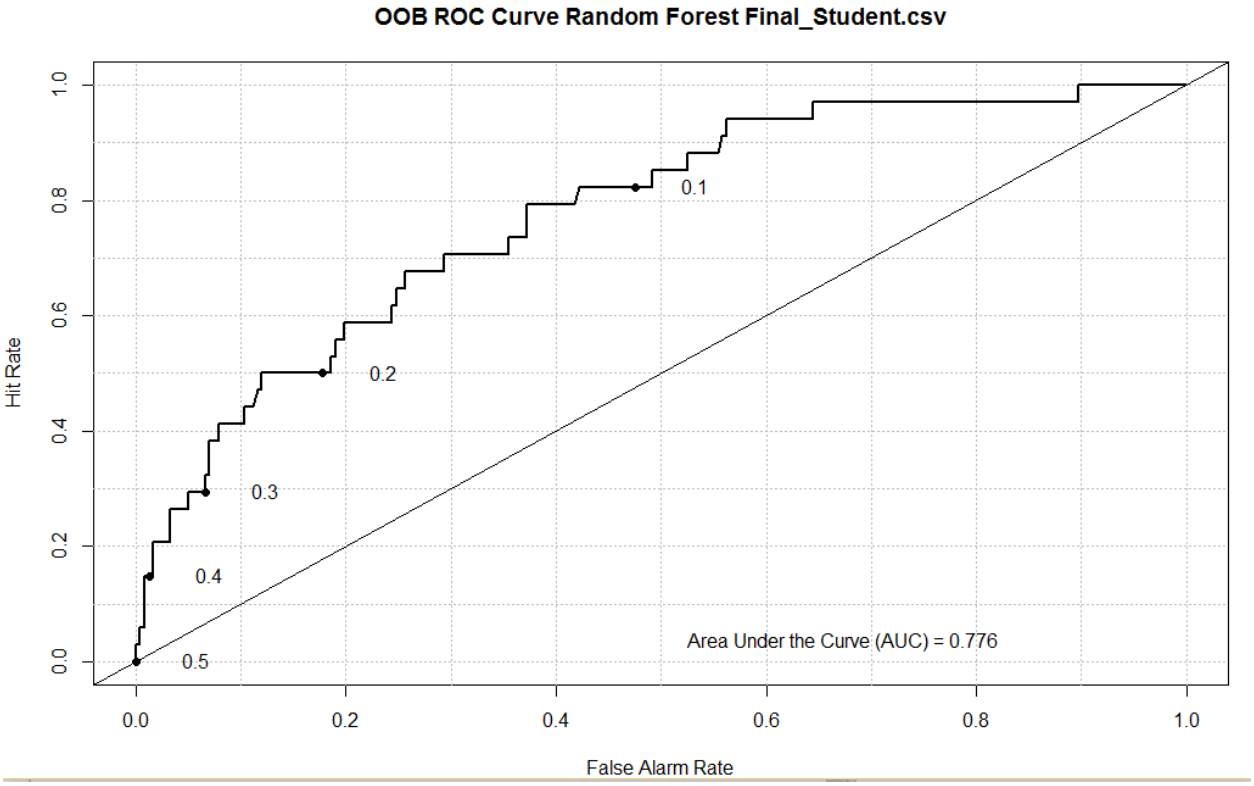
Finally, we split our data into test data and training data and then performed regresion tests. The goal was to find which set of variables most accuratly predicts whether a student will be a drinker.

We also implemented Naïve Bayes Classification algorithm to find out which set of variables were able to best predict whether a student is likely to be invloved in alcohol consumption. And we found out that ‘Age’ and ‘Sex’ of the student were the best predictors of alcohol consumption.

Then, to test the accuracy for our model, we plotted ROC graphs for Decision Tree and Random Forest Algorithms, which are as shown below:


ROC Curve Decision Tree Final_Student.csv [validate] avg_consumption_class
True Positive Rate
False Positive Rate
AUC = 0.69
0.00
0.25
0.50
0.75
1.00


Predicted vs. Observed
Decision Tree Model
Final_Student.csv
Linear Fit to Points
Predicted=Observed
Pseudo R-square=0.2652
Predicted
avg_consumption_class (Jittered)
Rattle 2016-May-11 22:34:08 Amar Kurane



Whether the Technique Worked

As the results show, we can predict with about 95% accuracy using Logistic Regression Model whether a student is likely to be a drinker. Our technique was effective in showing what schools should look for as early warning signs that students may be prone to drinking.

Limitations

- Limited number of records in the dataset
- Data is collected from the just two schools of same geograph

Who Did What

All of us searched for, discussed, and decided on a dataset to work with.

Kristen Bertch: I created some of the graphs and discussed our strategies, results, and conclusions with the group. I also wrote the Write-up and was point for forming the slideshow for our presentation.

Ekta Ahuja: I worked on finding correlations between alcohol consumption and other various parameters in the dataset. Also, I implemented Naïve Bayes algorithm. I also made the final edits in the report.

Amar Kurane: I have decided upon the dataset, analyzed it and created the classification. Also, I completed major part of data exploration and predictive algorithms like linear regression, logistic regression and decision tree.