Alcohol and Student success

Wubba Lubba Dub Dub

Kate, Akshay, Purvi

Does alcohol affect student’s performance in school ?

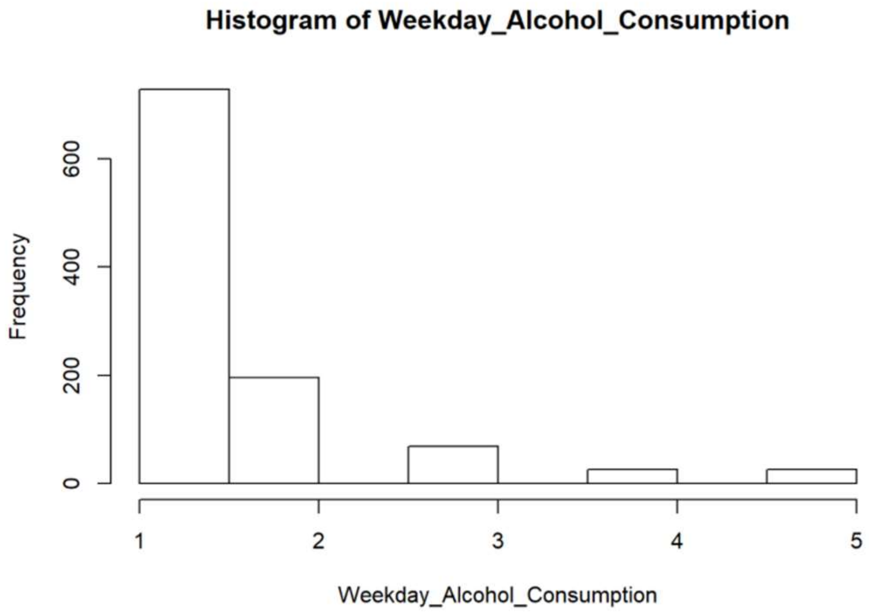
- **Background:** We were interested in determining how alcohol consumption related to academic success. GW emphasizes alcohol awareness and the impacts on class performance. We wanted to determine if these claims are supported by data.

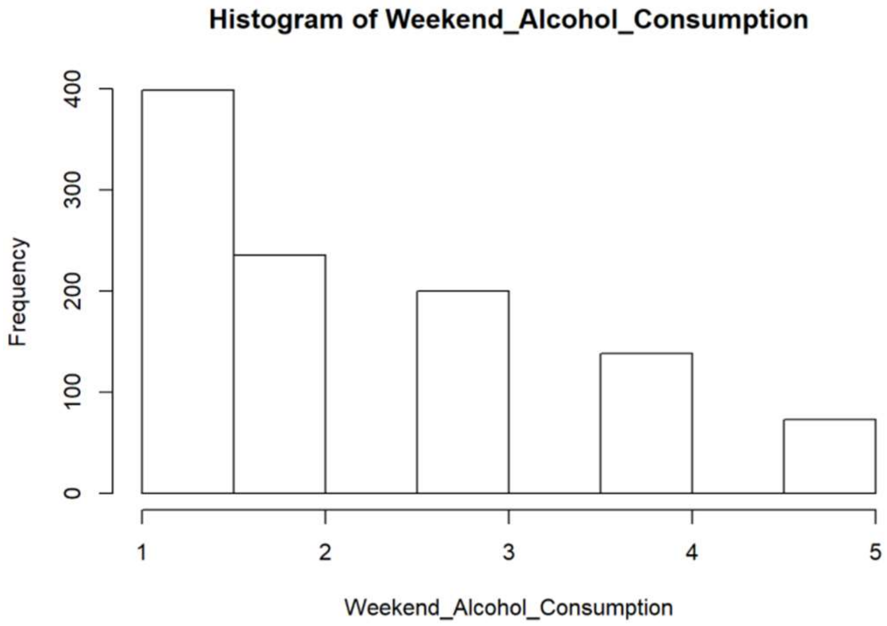
- **Data:** We found a dataset on Kaggle relating to alcohol consumption and grades for students in a secondary school. The data contained 35 variables on student life, and had observations for 1044 students.

- **Research:** We researched the grading scale for Portugal (where the study took place) and learned that a grade over 10 (out of 20) is considered passing. Other analysis which has been conducted on student success and alcohol is largely qualitative, while our analysis will be quantitative.

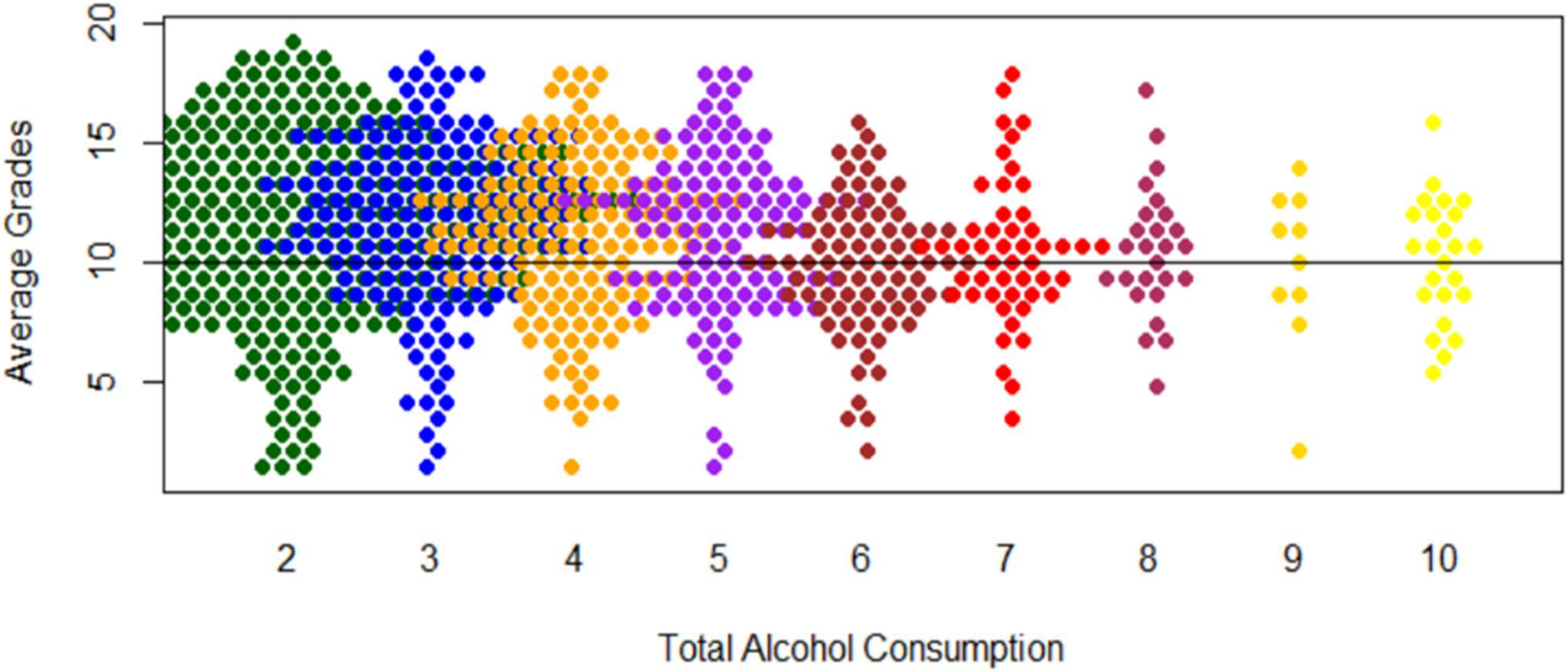
- **Question Selection:** We settled on the research question "**Does alcohol consumption affect student academic performance?**". We will measure alcohol consumption using the variables "Dalc" (Weekday alcohol consumption on a scale from 1 to 5) and "Walc" (Weekend alcohol consumption). We will measure academic success using average grade and whether or not a student passes the course.

Students’ Alcohol Consumption Distribution





Average Grades against Total Alcohol Consumption

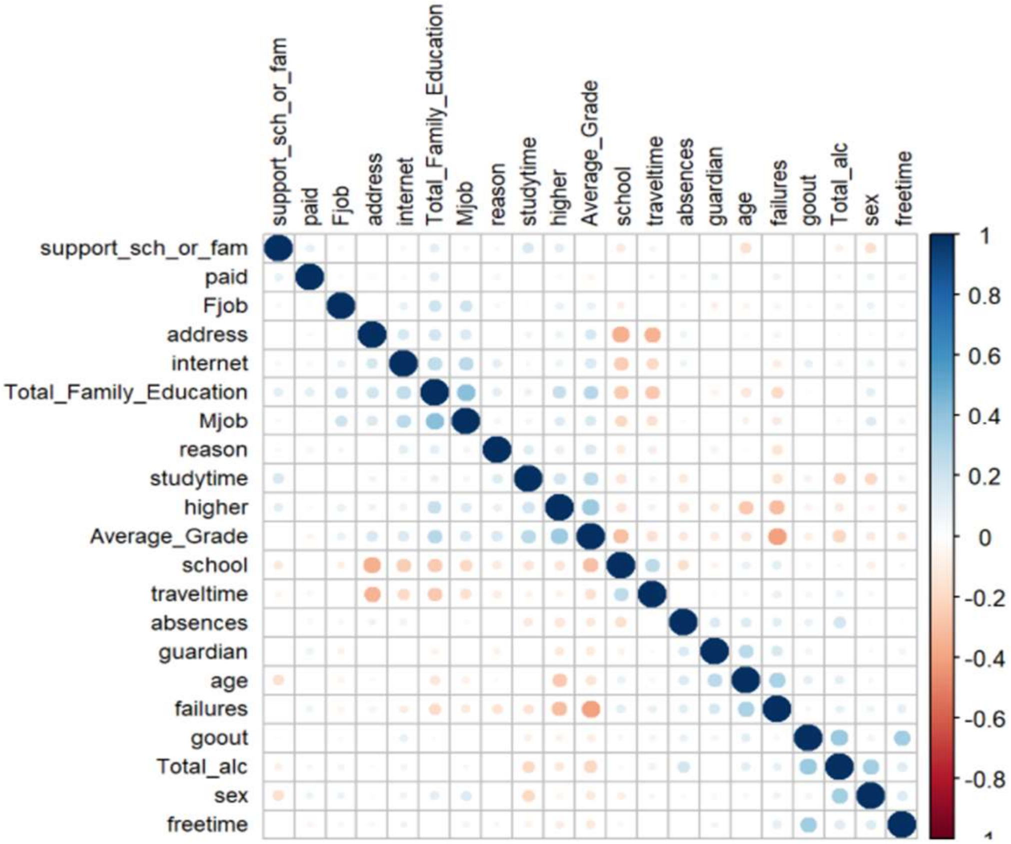


Correlation between the relevant variables

- Students taking Math did not show any interesting trends like the students taking Portuguese

- Students taking Portuguese showed many factors affecting their grades, highest correlation being past failures and average_grade

- So, we analyzed the student's taking Portuguese even further

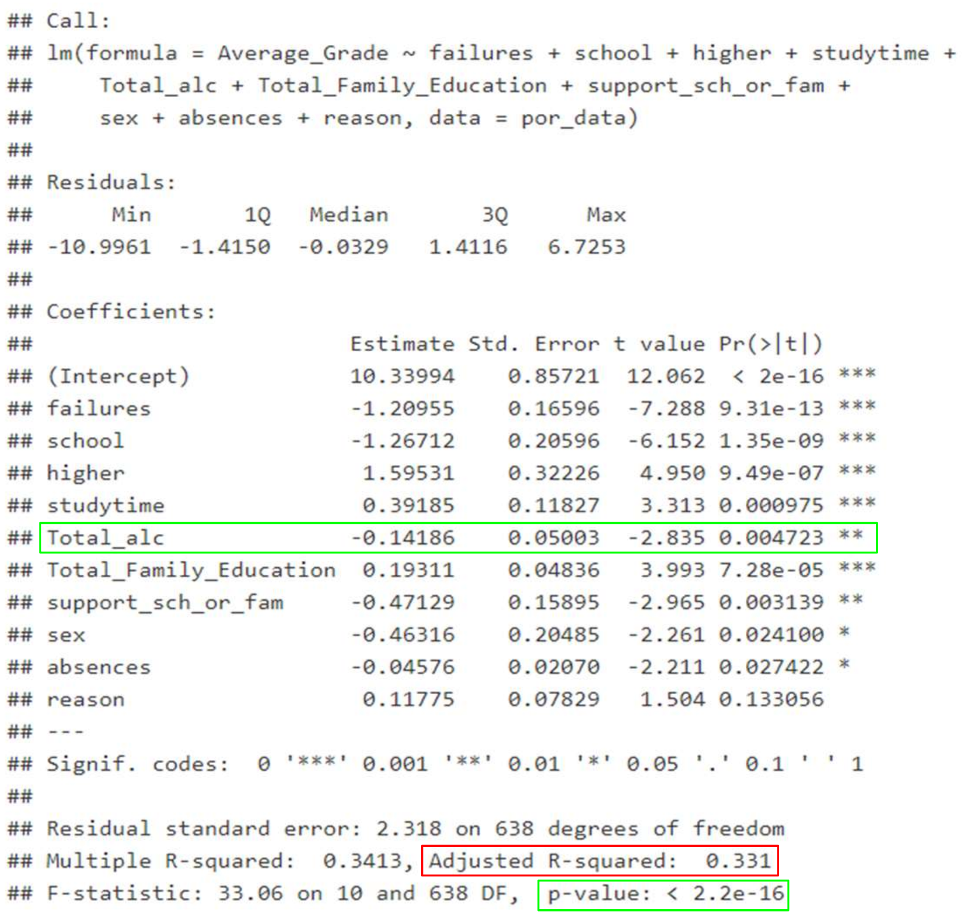


Model Selection

Multiple Linear Regression:

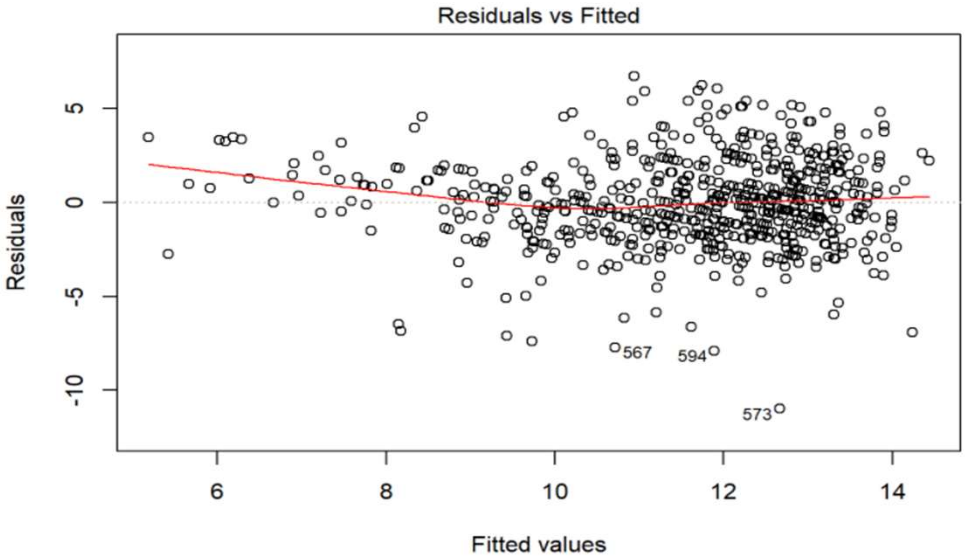
- Increased alcohol consumption has a negative effect on the student's average grade. The negative factor supports our question

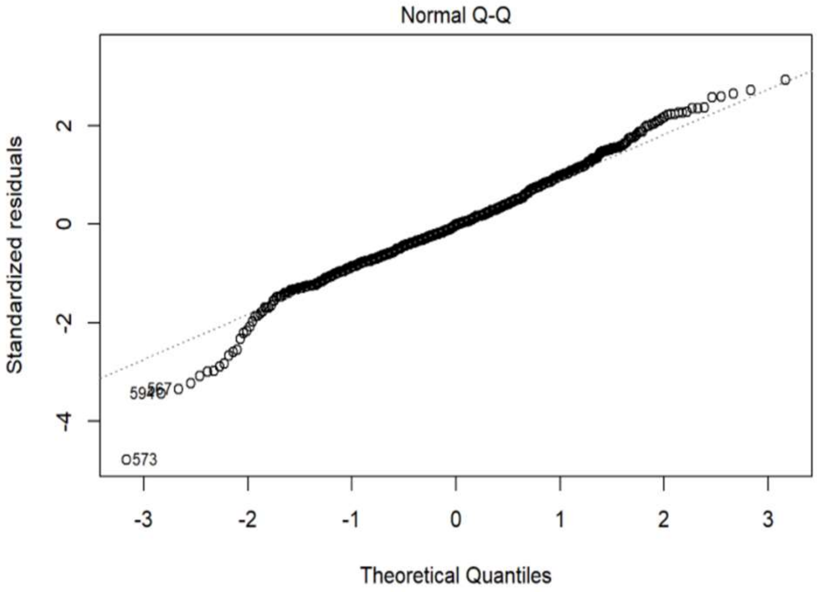
- The low adjusted R-squared value is compensated by a significant low p-value



```
## Call:
## lm(formula = Average_Grade ~ failures + school + higher + studytime +
##     Total_alc + Total_Family_Education + support_sch_or_fam +
##     sex + absences + reason, data = por_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9961  -1.4150  -0.0329   1.4116   6.7253
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.33994    0.85721  12.062  < 2e-16 ***
## failures               -1.20955    0.16596  -7.288 9.31e-13 ***
## school                 -1.26712    0.20596  -6.152 1.35e-09 ***
## higher                  1.59531    0.32226   4.950 9.49e-07 ***
## studytime               0.39185    0.11827   3.313 0.000975 ***
## Total_alc              -0.14186    0.05003  -2.835 0.004723 **
## Total_Family_Education  0.19311    0.04836   3.993 7.28e-05 ***
## support_sch_or_fam     -0.47129    0.15895  -2.965 0.003139 **
## sex                    -0.46316    0.20485  -2.261 0.024100 *
## absences               -0.04576    0.02070  -2.211 0.027422 *
## reason                  0.11775    0.07829   1.504 0.133056
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.318 on 638 degrees of freedom
## Multiple R-squared:  0.3413, Adjusted R-squared:  0.331
## F-statistic: 33.06 on 10 and 638 DF,  p-value: < 2.2e-16
```

Model Validation



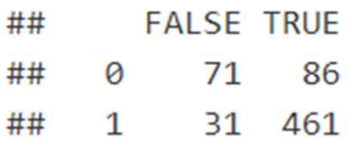


Model Selection

Logistic Regression:

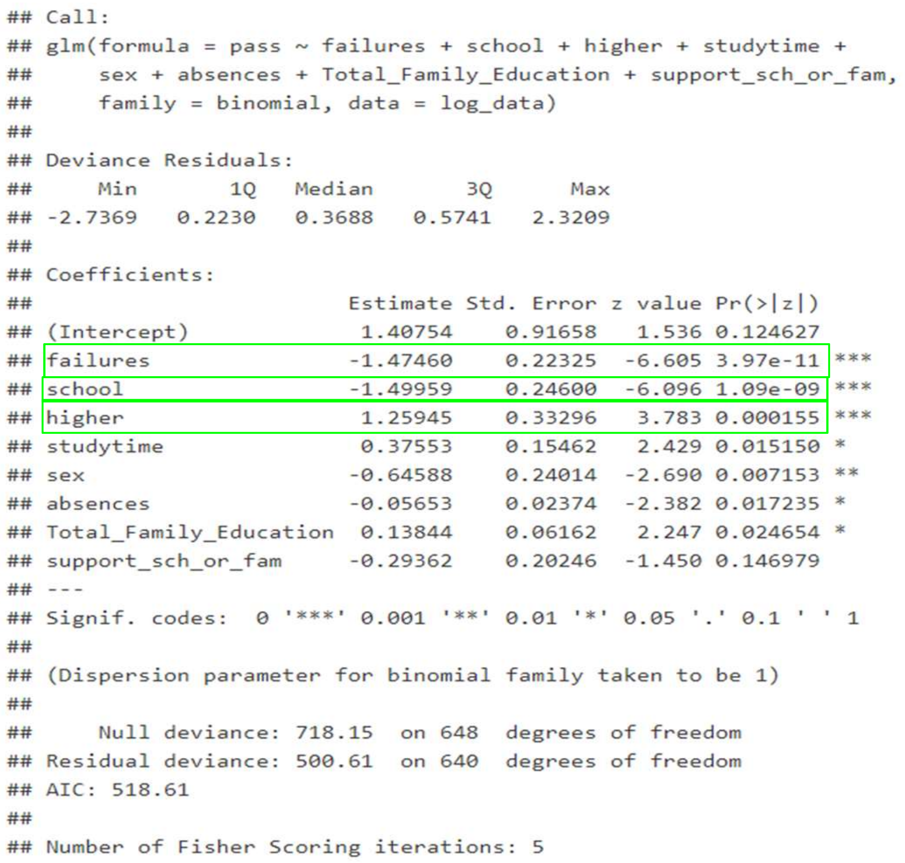
- There were factors other than alcohol consumption that affected a student passing or failing an exam

Confusion Matrix to validate our model:



```
##      FALSE TRUE
##   0     71   86
##   1     31  461
```

- This indicates a success rate of 82%, with a 13% false positive rate and a 5% false negative rate



```
## Call:
## glm(formula = pass ~ failures + school + higher + studytime +
##     sex + absences + Total_Family_Education + support_sch_or_fam,
##     family = binomial, data = log_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7369   0.2230   0.3688   0.5741   2.3209
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)             1.40754    0.91658   1.536 0.124627
## failures               -1.47460    0.22325  -6.605 3.97e-11 ***
## school                 -1.49959    0.24600  -6.096 1.09e-09 ***
## higher                  1.25945    0.33296   3.783 0.000155 ***
## studytime               0.37553    0.15462   2.429 0.015150 *
## sex                    -0.64588    0.24014  -2.690 0.007153 **
## absences               -0.05653    0.02374  -2.382 0.017235 *
## Total_Family_Education  0.13844    0.06162   2.247 0.024654 *
## support_sch_or_fam     -0.29362    0.20246  -1.450 0.146979
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 718.15  on 648  degrees of freedom
## Residual deviance: 500.61  on 640  degrees of freedom
## AIC: 518.61
##
## Number of Fisher Scoring iterations: 5
```

Conclusion

Overall, our models gave us a good idea about how alcohol and other factors affected student's performance in school. We understand that alcohol (surprisingly) has a relatively low effect on the students' grades. We had initially started out with the hypothesis that higher the alcohol consumption, higher the failure rate of the students. However, the effect isn't as dramatic as we had expected it to be.

Takeaways:

- Our multiple linear regression model was statistically significant. Diagnostic plots determined that no model assumptions were violated in the construction of this model.
- Alcohol consumption does predict a statistically significant decrease in student grade performance. Other predictors of student grades were past failures, school attended, study time, and desire for future education.
- We can predict the expected grade of a student based on past failures, school, desire for higher ed, time studying, alcohol consumption, family education, support, sex, and absences.
- Potential limitations include the size of the data, since we only had information on two schools. Additionally, if we had more data, we could have done training and testing for our logistic model.
- Another limitation was the structure of the data– a lot of our variables were categorical variables. Alcohol consumption was a categorical variable on a scale from "low" to "high", rather than a numeric variable. Similarly, grades were on a 20 point scale, rather than a continuous percentage scale.

Rdocs link:
http://rpubs.com/kjones409/339320

Datasource link: https://www.kaggle.com/uciml/student-alcohol-consumption