# Final Report

due November 16, 2021 by 11:59 PM

Danielle Mensah, Haby Sow, Colin Lee

11/12/2021

## Abstract

This report outlines the effects of alcohol consumption on students' efficacy of learning and whether family circumstances play a role in the amount of alcohol consumed by these students. We used data from a dataset we found on Kaggle that was posted by UC Irvine. The dataset collected data on student self-reported alcohol consumption in Portuguese math class in two secondary schools, Gabriel Pereira and Moushino de Silveira during the 2005-2006 school year.

To focus on the effect of alcohol consumption on student learning we used the variables relating to alcohol consumption. On the other hand, to focus on the family circumstances in comparison to alcohol consumption we used the variables related to alcohol consumption. With these variables, we built regression models to decide which attributes were significant to our findings. In addition, we used proper visual models to display our findings in a visual manner.

Our model shows that having extra paid classes within the course subject of math plays a significant factor in whether or not one drinks more often. In addition, our data showed that increased alcohol consumption also showed an increase in the amount of failed courses.

## Introduction

There are many beliefs about the prevalence of student drinking and what sort of factors impact or are indicative of high levels of drinking as well as how drinking may impact individual academic performance. For our project, we will be studying two main questions with regard to alcohol consumption.

The first question is what family or external circumstances are drivers for high student alcohol consumption. Insight into these impact of these factor may allow for a better understanding for how family education, support, size, and others may be associated with binge drinking.

The second question we seek to answer is how student alcohol consumption correlates with student life whether it be through absences, class failures, marks received on exams, activities outside of school, or desire to attend university.

In order to conduct our study of these two questions, we will be working with a dataset we found on Kaggle. The data was collected for a research paper called "Using Data Mining to Predict Secondary School Student Performance." The authors describe how the surveys were developed and reviewed by school staff and students (Cortez et al 2). The surveys were then delivered to the students through paper questionnaires with predefined answer choices for 37 different questions, 33 of which are variables in the dataset (Cortez et a 2). The questionnaires were answered by all 395 math students in the Gabriel Pereira and Moushino da Silveira secondary public schools (Cortez et al 2). The secondary school students are what would be considered as "high school" in the United States, with age ranges from 15 - 19 years old, but with a few older individuals up to 22 years old. Additionally different from the United States is that the legal drinking age in Portugal is 18 years old, which is on the upper spectrum for secondary students. Thus, the drinking prevalence in this

data set may not correlate to prevalence found in American high schools. However, this data may mirror drinking at the university level better, as to the 21 year age requirement in the US is in the upper age range for university students as is 18 in the upper age range for Portuguese secondary school students.

## Overall Methodology

In order to assess the first question, we modeled a logistic regression with binge drinking being the respondent variable and familial and external variables serving as the predictor variables. Through this analysis, we intend to understand the significance of certain family and external factors in impacting whether or not a student is a binge drinker and interpret the severity of impact of those variables. More specifically, we will be investigating the student's residence, family size, parental separation, mother and father education levels, school support, family support, additional paid math classes, and attendance of nursery as the explanatory variables.

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

## Logistic Regression Methodology:

We conducted a logistic regression to determine what sort of family or external factors may be drivers for student alcohol consumption. We decided to pursue a logistic regression in this case due to the categorical nature of many of the variables. Furthermore, we wanted to do a logistic regression on top of a linear regression, where instead of quantifying the amount of alcoholic drinks, we can assess how these indicators are involved with impacting whether or not a student can be considered an "alcoholic" or "binge drinker."

Thus, in order to set up this logistic regression, we had to begin by transforming the daily alcoholic consumption value into a binary variable. This was done by using the mutate() function to create a new variable that said that an individual was considered a "binge drinker" or "binger" (variable name) based on their sex and drink count. Sex was an important indicator here since alcoholic consumption for intoxication can vary greatly based on sex. According to the National Institute on Alcohol Abuse and Alcoholism (NIAAA), binge drinking varies between males and females, where 4 or more drinks is for females and 5 or more drinks is for males (NIAAA). Furthermore, for youths, which is where we will categorize the students here, the drinks are 3 for girls and 3 to 5 for boys (NIAAA). Thus, from the NIAAA, we classified male student binge drinkers as consumers of 4 or more drinks and female student binge drinkers as consumers of 3 or more drinks.

Next, we had to select the predictor variables for our regression. Because we wanted to answer what sort of external or familial factors impact whether a student is a binge drinker, we used the following variables: residence, family size, parental separation, mother and father education level, educational support, family educational support, extra paid tutoring classes, and whether or not the student attended nursery school.

Thus, we developed the following logistic regression model:

logit(pii) = log(pii/(1-pii)) = B0 + B1(urbani) + B2(famsmalli) + B3(parents_togetheri) + B4(mother_secondaryi) + B5(father_secondaryi) + B6(school_supporti) + B7(family_supporti) + B8(extra_tutoringi) + B9(attended nurseryi)

such that

urbani = 1 if the respondent lives in an urban environment, famsmalli = 1 if the family size is less than 3, parents_togetheri = 1 if the parents are together, mother_secondaryi = 1 if the mother's education level is secondary or above, father_secondaryi = 1 if the father's education level is secondary or above, school_supporti = 1 if the student is receiving school support, family_supporti = 1 if the student is receiving extra family support for education, extra_tutoringi = 1 if the student is receiving extra paid classes within math, and attended_nurseryi = 1 if the student attended nursery school before.

For these variables, we curated them so that for the values equal to 1, we hypothesized those values would be most involved in influencing a student to not be a binge drinker because we believe those to be positive

familial or external attributes in one's life that may influence one to avoid binge drinking.

For the parental education, since this value was not originally binary, we assessed that it would be most relevant to assess whether or not the education level was secondary or not since if the respondent's parents had reached the at least the same level of education as the respondent him/herself, then there would be some likelihood of greater parental support due to them having shared the same experiences. Regardless, as we will explain in the next section, switching the value to be higher than secondary education makes no impact.

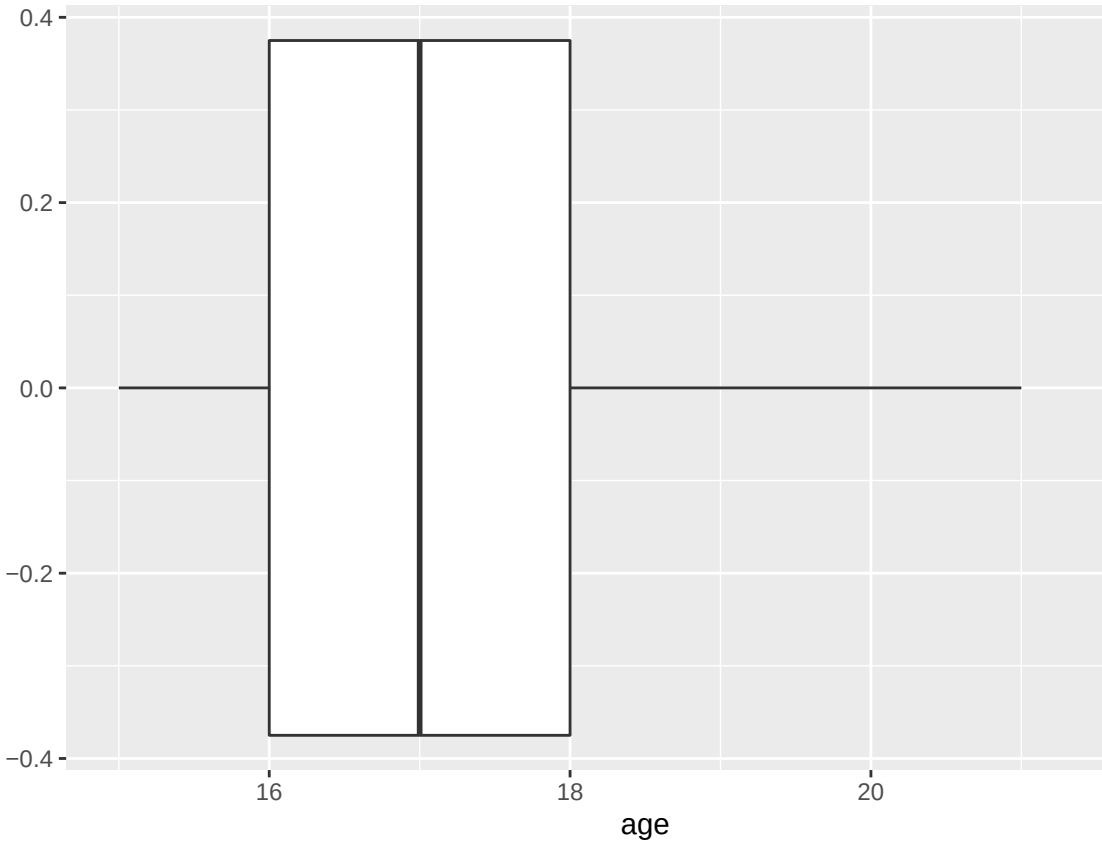
```
## # A tibble: 10 x 7
##    term              estimate std.error statistic p.value conf.low conf.high
##    <chr>                <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
##  1 (Intercept)        0.0608     0.981   -2.86   0.00429  0.00752     0.365
##  2 urban              0.949      0.532   -0.0986 0.921    0.357       2.99
##  3 famsmall           1.48       0.477    0.819  0.413    0.556       3.70
##  4 parents_together   0.678      0.681   -0.571  0.568    0.199       3.14
##  5 mother_secondary   1.38       0.554    0.584  0.559    0.470       4.21
##  6 father_secondary   1.03       0.516    0.0617 0.951    0.379       2.91
##  7 school_support     1.26       0.670    0.346  0.729    0.277       4.19
##  8 family_support     0.801      0.489   -0.455  0.649    0.313       2.17
##  9 extra_tutoring     4.32       0.520    2.81   0.00491  1.64       13.0
## 10 attended_nursery   0.391      0.515   -1.82   0.0682   0.145       1.12
```

#Clearing Missing Data

```
##      school   sex   age address famsize Pstatus  Medu  Fedu  Mjob  Fjob reason
## [1,]  FALSE FALSE FALSE   FALSE   FALSE   FALSE FALSE FALSE FALSE FALSE  FALSE
## [2,]  FALSE FALSE FALSE   FALSE   FALSE   FALSE FALSE FALSE FALSE FALSE  FALSE
## [3,]  FALSE FALSE FALSE   FALSE   FALSE   FALSE FALSE FALSE FALSE FALSE  FALSE
## [4,]  FALSE FALSE FALSE   FALSE   FALSE   FALSE FALSE FALSE FALSE FALSE  FALSE
## [5,]  FALSE FALSE FALSE   FALSE   FALSE   FALSE FALSE FALSE FALSE FALSE  FALSE
## [6,]  FALSE FALSE FALSE   FALSE   FALSE   FALSE FALSE FALSE FALSE FALSE  FALSE
##      guardian traveltime studytime failures schoolsup famsup  paid activities
## [1,]    FALSE      FALSE     FALSE    FALSE     FALSE  FALSE FALSE      FALSE
## [2,]    FALSE      FALSE     FALSE    FALSE     FALSE  FALSE FALSE      FALSE
## [3,]    FALSE      FALSE     FALSE    FALSE     FALSE  FALSE FALSE      FALSE
## [4,]    FALSE      FALSE     FALSE    FALSE     FALSE  FALSE FALSE      FALSE
## [5,]    FALSE      FALSE     FALSE    FALSE     FALSE  FALSE FALSE      FALSE
## [6,]    FALSE      FALSE     FALSE    FALSE     FALSE  FALSE FALSE      FALSE
##      nursery higher internet romantic famrel freetime goout  Dalc  Walc health
## [1,]   FALSE  FALSE    FALSE    FALSE  FALSE    FALSE FALSE FALSE FALSE  FALSE
## [2,]   FALSE  FALSE    FALSE    FALSE  FALSE    FALSE FALSE FALSE FALSE  FALSE
## [3,]   FALSE  FALSE    FALSE    FALSE  FALSE    FALSE FALSE FALSE FALSE  FALSE
## [4,]   FALSE  FALSE    FALSE    FALSE  FALSE    FALSE FALSE FALSE FALSE  FALSE
## [5,]   FALSE  FALSE    FALSE    FALSE  FALSE    FALSE FALSE FALSE FALSE  FALSE
## [6,]   FALSE  FALSE    FALSE    FALSE  FALSE    FALSE FALSE FALSE FALSE  FALSE
##      absences    G1    G2    G3
## [1,]    FALSE FALSE FALSE FALSE
## [2,]    FALSE FALSE FALSE FALSE
## [3,]    FALSE FALSE FALSE FALSE
## [4,]    FALSE FALSE FALSE FALSE
## [5,]    FALSE FALSE FALSE FALSE
## [6,]    FALSE FALSE FALSE FALSE
```

As you can see from this quick check. There are no missing values in our data. Therefore we can move on with further analysis and no clearing of variables needs to be done. I put only the head of the data because it was too long to visually see the whole thing however it is all false.

#Data Wrangling There are two big questions that we want answered with this data set: whether a students average alcohol consumption is correlated with their family circumstances and whether alcohol consumption has an effect on student life. Lets first look at some geographics of our students and adjust our data as needed to




complete a through analysis.
We can see from the data that the average age of the students tested was about 17 and there was an out liar at age 22.

```
##
##  15  16  17  18  19  20  21  22
##  82 104  98  82  24   3   1   1
```

```
##
##  GP  MS
## 349  46
```

```
##
##   F   M
## 208 187
```

```
##
##   R   U
##  88 307
```

```
##
## GT3 LE3
## 281 114
```

```
##
##   A   T
##  41 354
```

```
##
##   0   1   2   3   4
##   3  59 103  99 131
```

```
##
##   0   1   2   3   4
##   2  82 115 100  96
```

```
##
##  at_home   health    other services  teacher
##       59       34      141      103       58
```

```
##
##     course       home      other reputation
##        145        109         36        105
```

```
##
## father mother  other
##     90    273     32
```

```
##
##  no yes
## 153 242
```

```
##
##  no yes
##  66 329
```

It is important to count some of the important variables in our data so we can see which variables had significant responses that need to be analyzed.

## Logistic Regression Analysis

Upon completing the logistic regression, we found that only one variable to be significant at a p-value of 0.05, which was whether or not a student had extra tutoring or not (B8, coefficient for extra_tutoringi). However, another variable, which was B9 (coefficient for attended_nurseryi), had a p-value of 0.0682, which is very, very close. All other estimates had much higher p-values, which were not close to being significant. Furthermore, we attempted to see how changing the level of education we identified as appropriate for our binary variables on parental education, but the p-value did not shift significantly (change from 0.56/0.95 to 0.87/0.22 for mother's education and father's education respectively).

Moving forward with the extra_tutoringi variable, the estimate came out to be 4.318 with a 95% CI between 1.64 and 12.96, which is an extremely wide range. Because the values are exponentiated, this indicates that the odds ratio for this value 4.318, which means that a student with extra paid math classes is 4.318 times as likely to be a binge drinker than someone who does not have extra paid math classes. This is quite interesting and contradicts our original thoughts regarding this variable that higher educational support through tutoring might indicate less likelihood for binge drinking. Some speculation as to why tutoring may be correlated with binge drinking is possibly due to a perceived need for tutoring for the respondent from the students' parents because of the drinking itself.

With regard to the attended_nurseryi variable, the estimate was 0.3911 with a 95% CI between 0.145 and 1.123. The CI is very wide and spans across 1. However, since the p-value of 0.0682 is fairly close to 1 and the upper bound for the CI is just above 1. Thus, we will proceed with caution on this variable as we analyze its results due to its closeness to significance. As this value is exponentiated, the odds ratio is 0.3911 and thus a student who has attended nursery may be 0.3911 times as likely to be a binge drinker. Although this value is not very significant, this is somewhat in line with our hypothesis that nursery attendance may be indicative of a family's prioritization of education and possible aversion to binge drinking for children.

Thus, in terms of external and familial factors impacting whether or not a student respondent is a binge drinker, we assess that there is only one significant variable, which is whether or not the student has extra paid math classes. However, we can also proceed with caution on the variable regarding whether the student has attended nursery school or not because of the low p-value.

```
## # A tibble: 3 x 7
##   term             estimate std.error statistic p.value conf.low conf.high
##   <chr>               <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 (Intercept)        0.0608     0.981     -2.86 0.00429  0.00752     0.365
## 2 extra_tutoring     4.32       0.520      2.81 0.00491  1.64       13.0
## 3 attended_nursery   0.391      0.515     -1.82 0.0682   0.145       1.12
```

#Linear Regression How does student life impact daily alcohol consumption?

A multiple linear regression was run to understand how student life impacts alcohol consumption. We use Dalc (daily alcohol consumption) as response variables and failures(number of past class failures ) , goout (going out with friends) , studytime (weekly study time ) , freetime (free time after school), famrel (quality of family relationships), G1(first period grade) , G2 (second period grade), G3 ( third period grade) and activities ( extra-curricular activities) as explanatory variables. The linear regression used is the following one: Dalc= b0 + b1 * failures +b2* goout + b3 * studytime +b4 * freetime +b5 * famrel+ b6 * activities + b7* G1 +b8* G2+ b9 * G3 In this regression output, only goout, studytime, freetime and famrel were significant at 0.05. The result tells us that an increase in daily alcohol consumption is positively correlated (correlation coefficient= 0.27) with going out with friends. The more frequently a student goes out with friends, the more the student drinks. A one unit increase in going out with friends increases daily alcohol consumption by 0.176.In this case, we can say that going out with friends impacts student daily alcohol consumption. Additionally, Daily alcohol consumption is negatively correlated (correlation coefficient= 0.2 ) with studytime. In other words, the more students dedicate their time to studying, the less they consume alcohol. A one unit increase in studytime decreases alcohol consumption by 0.146. In short, spending more time studying decreases daily alcohol consumption. Furthermore, daily alcohol consumption is negatively correlated(correlation coefficient= 0.08) with the quality of family relationships. The higher the quality is, the less alcohol consumption is. A one unit increase in the quality of family relationships decreases daily alcohol consumption by 0.1. Also, Daily alcohol consumption is positively correlated with freetime (correlation coefficient=0.21 ). The more students have free time, the more they consume alcohol. A one unit increase in freetime increase alcohol consumption by 0.13

#Linear regression model:

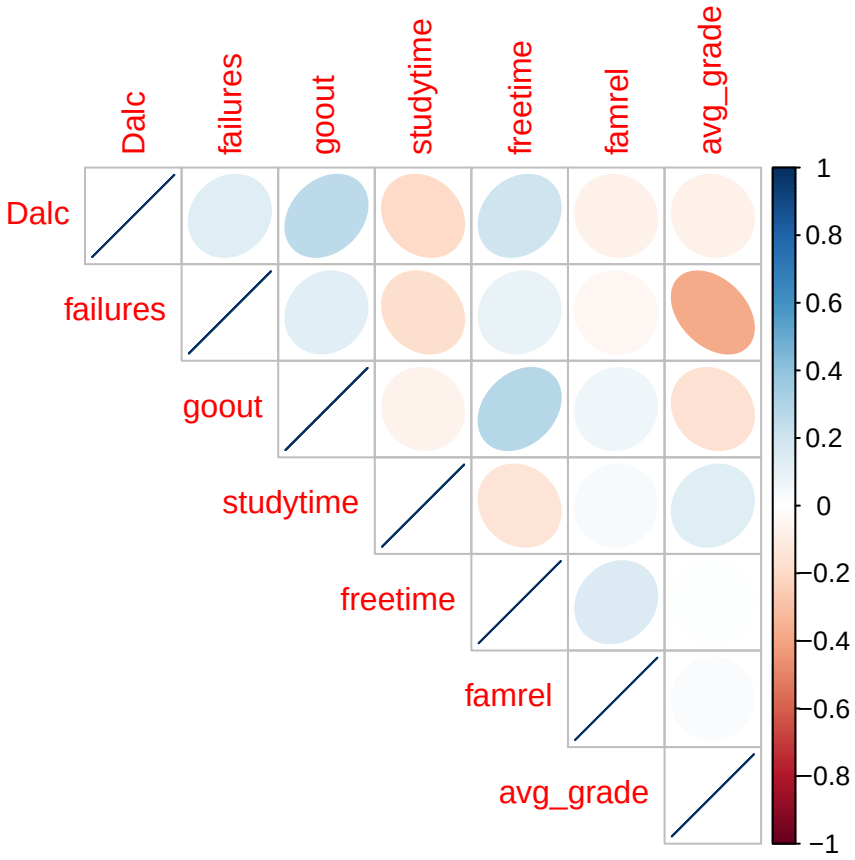
Dalc= 1.28 + 0.176 * goout + 0.146 * studytime + 0.1 * famrel + 0.13 *freetime

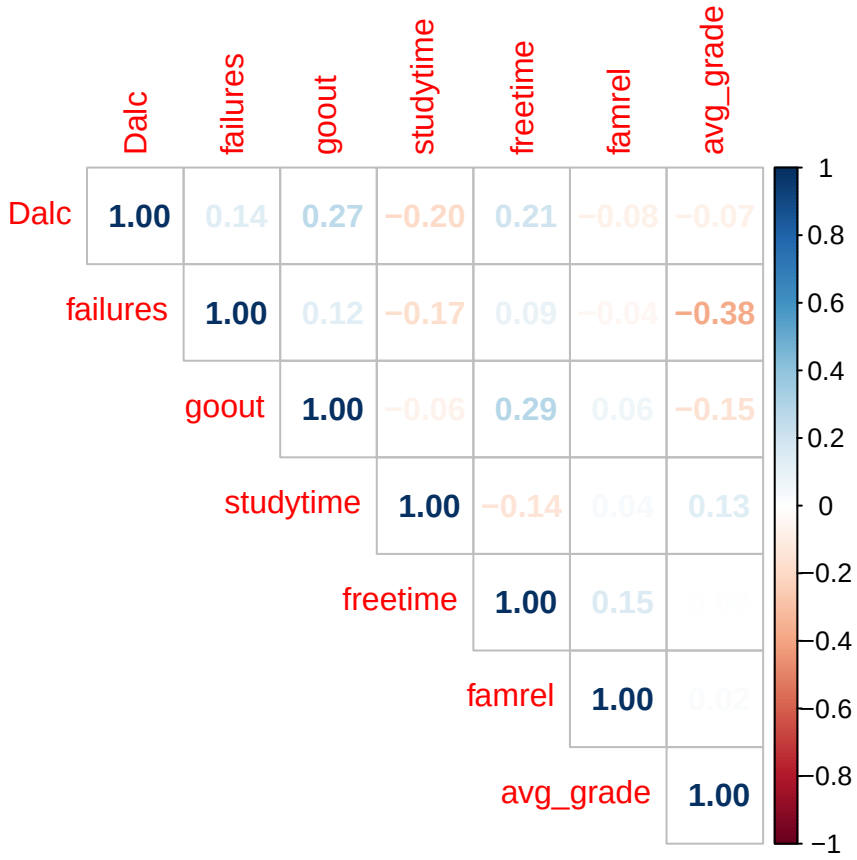
```
## # A tibble: 10 x 5
##    term          estimate std.error statistic   p.value
##    <chr>            <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)    1.28       0.308      4.16  0.0000397
##  2 failures       0.0767     0.0622     1.23  0.218
##  3 goout          0.176      0.0401     4.39  0.0000147
##  4 studytime     -0.146      0.0522    -2.80  0.00541
##  5 freetime       0.128      0.0453     2.83  0.00484
##  6 famrel        -0.0991     0.0483    -2.05  0.0409
##  7 activitiesyes -0.120      0.0855    -1.41  0.160
##  8 G1            -0.0265     0.0247    -1.07  0.284
##  9 G2             0.0196     0.0309     0.634 0.526
## 10 G3             0.00390    0.0223     0.175 0.861
```

#Showing relationship using correlation matrix

We create a new variable that shows that the average grade of the student using G1, G2, G3.

Then we made a visualization of the coefficient correlation between Dalc and the other predictors.


Dalc
failures
goout
studytime
freetime
famrel
avg_grade
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1




#Seocond linear model:

failures= 0.11 * Dalc + 0.16

With a p value of 0.006 less than 0.05, there is enough enough evidence to support that the true coefficient is not 0.

With this regression model, the result shows us that a one unit increase in daily alcohol consumption increases the number past failures. So, daily alcohol consumption impacts student average grade. The more students daily consume alcohol, the less , their average grade is.

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)    0.166    0.0721      2.30 0.0218
## 2 Dalc           0.114    0.0417      2.72 0.00677
```

Citations:

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

“Underage Drinking.” National Institute on Alcohol Abuse and Alcoholism, U.S. Department of Health and Human Services, https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/underage-drinking.