Final Report

due November 16, 2021 by 11:59 PM

Danielle Mensah, Haby Sow, Colin Lee

11/12/2021

Abstract

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 workday 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 with workday alcohol consumption (made binary) as the response variable.

The methodologies and results for these analyses will be further described in their sections below.

```
## 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 self-reported drinking level. According to the National Institute on Alcohol Abuse and Alcoholism (NIAAA), binge drinking is defined as 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). However, for this dataset, the values 1-5 do not indicate the amount of drinks, but rather, a self-reported drinking level, where 1 is "very low" drinking and 5 is "very high" drinking. This is very arbitrary, but in order to create a more binary definition, we defined drinking levels of 4 and 5 (high and very high drinking) on workdays. We additionally focus on workday binge drinking as opposed to the other variable, which is weekend binge drinking, because workday binge drinking is more negatively stigmatized, and is a larger factor in a student's academic and family life than weekend binge drinking.

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)        0.0949     1.04    -2.27    0.0231   0.0104     0.636
```

```
##  2 urban              0.957      0.595   -0.0739  0.941    0.322      3.52
##  3 famsmall           1.42       0.544    0.645   0.519    0.459      4.03
##  4 parents_together   0.429      0.707   -1.20    0.232    0.118      2.06
##  5 mother_secondary   1.41       0.611    0.560   0.575    0.425      4.79
##  6 father_secondary   0.741      0.579   -0.518   0.604    0.237      2.34
##  7 school_support     1.85       0.698    0.879   0.380    0.391      6.62
##  8 family_support     0.896      0.557   -0.196   0.844    0.309      2.84
##  9 extra_tutoring     4.29       0.581    2.50    0.0123   1.45      14.7
## 10 attended_nursery   0.252      0.560   -2.46    0.0139   0.0834     0.775
```

Logistic Regression Analysis

Upon completing the logistic regression, we found two variables to be significant at a p-value of 0.05, which were whether or not a student had extra tutoring or not (B8, coefficient for extra_tutoringi) and whether a student had attended nursery or not (coefficient for attended_nurseryi). 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 0.58/0.60 to 0.66/0.80 for mother's education and father's education respectively).

Moving forward with the extra_tutoringi variable, the estimate came out to be 4.285 with a p-value of 0.0122 and 95% CI between 1.45 and 14.73. Because the values are exponentiated, this indicates that the odds ratio for this value 4.285, which means that a student with extra paid math classes is 4.285 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 0.252 with a p-value of 0.0139 and 95% CI between 0.08 and 077. As this value is exponentiated, the odds ratio is 0.252 and thus a student who has attended nursery may be 0.252 times as likely to be a binge drinker. This is more in line with our hypothesis that nursery attendance may be indicative of a positive external factor, as it is possible that 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 school/work day binge drinker, we assess that there are only two significant variables, which are whether or not the student has extra paid math classes and whether the student has attended nursery school or not.

```
## # 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)        0.0949     1.04      -2.27  0.0231   0.0104     0.636
## 2 extra_tutoring     4.29       0.581      2.50  0.0123   1.45      14.7
## 3 attended_nursery   0.252      0.560     -2.46  0.0139   0.0834     0.775
```

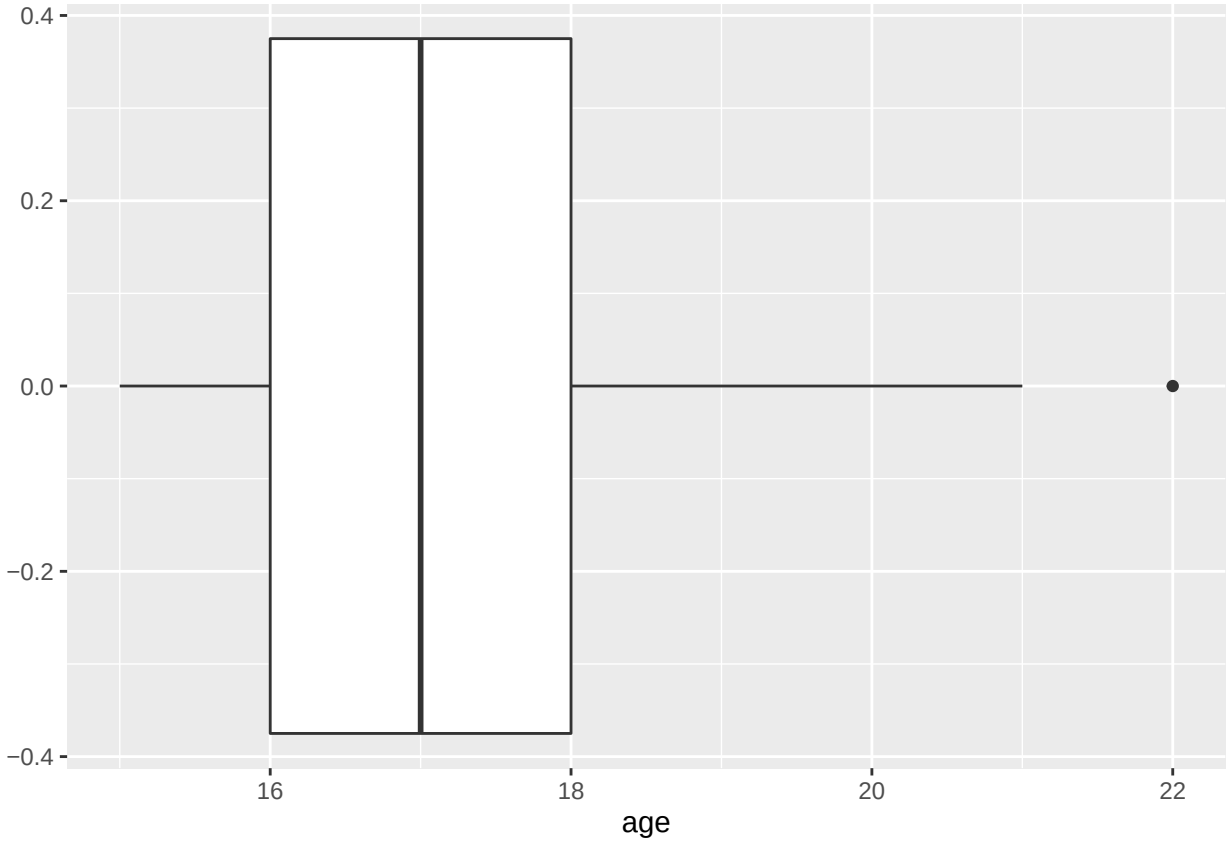
#Clearing Missing Data

```
##      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.



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
```

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.