Project Proposal

due October 11, 2021 by 11:59 PM

Danielle Mensah, Haby Sow, and Colin Lee

10/11/2021

Load Packages

```
library(tidyverse)
library(tidymodels)
library(readr)
```

Load Data

```
student <- readr::read_csv("student-mat.csv")
```

Introduction and Data, including Research Questions

For our research project, we plan to study two questions with respect to alcohol consumption amongst students in math courses in secondary school. Although this is not a traditional "global health" dataset, there is merit to analyzing this data from a health perspective. Alcohol consumption impacts personal health and cognition, but is also impacted by various other factors, such as family and the environment, just like many other health conditions or illnesses. Thus, the first question is what family circumstances are correlated with student alcohol consumption, which may allow us to gain better insight into how family education, presence, or occupation may be associated with alcohol use. 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.

To accomplish this task, we will be utilizing a dataset we found on Kaggle, which 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. In Portuguese secondary schools, ages typically range from 15-19, but this dataset also includes a few individuals from 20 - 22 years old. Additionally, unlike the USA, the minimum drinking age is 18 years old, so drinking prevalence at this educational level might be less comparable to high schools in the USA, but since drinking is legal for students nearing graduation, this may be more parallel to drinking trends in US universities.

Although the dataset was made available by University of California Irvine (UCI) Machine Learning, the source itself was revealed in a paper called "Using Data Mining to Predict Secondary School Student Performance." In the paper, the authors state that the data was collected through paper questionnaires with predefined answers for most questions. This questionnaire was developed in conjunction with and reviewed by school staff and students (Cortez et al 2). The questionnaire was answered by all students attending a math course in two public schools, Gabriel Pereira or Mousinho da Silveira, which added up to 395 students (Cortez et al 2). The questionnaire collected information through 37 questions, of which 33 exist as variables in the dataset (Cortez et al 2). However, as of now, we plan to most closely assess the following variables for our first question regarding potential family factors correlated with alcohol consumption: family size, parental

cohabitation status, mother's education level, father's education level, mother's job, father's job, guardian, family relationship quality, and family educational support. For our second question regarding alcohol consumption in student life, we want to utilize the following variables: desire to pursue higher education, first, second, and final grades, number of failed courses, and student health level. Across both questions, we will investigate their correlation amongst the variables of reported daily and weekly alcohol consumption. # Glimpse

```
glimpse(student)
```

```
## Rows: 395
## Columns: 33
## $ school     <chr> "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP",~
## $ sex        <chr> "F", "F", "F", "F", "F", "M", "M", "F", "M", "M", "F", "F",~
## $ age        <dbl> 18, 17, 15, 15, 16, 16, 16, 17, 15, 15, 15, 15, 15, 15, 15,~
## $ address    <chr> "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U",~
## $ famsize    <chr> "GT3", "GT3", "LE3", "GT3", "GT3", "LE3", "LE3", "GT3", "LE~
## $ Pstatus    <chr> "A", "T", "T", "T", "T", "T", "T", "A", "A", "T", "T", "T",~
## $ Medu       <dbl> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4, 3, 3, 4,~
## $ Fedu       <dbl> 4, 1, 1, 2, 3, 3, 2, 4, 2, 4, 4, 1, 4, 3, 2, 4, 4, 3, 2, 3,~
## $ Mjob       <chr> "at_home", "at_home", "at_home", "health", "other", "servic~
## $ Fjob       <chr> "teacher", "other", "other", "services", "other", "other", ~
## $ reason     <chr> "course", "course", "other", "home", "home", "reputation", ~
## $ guardian   <chr> "mother", "father", "mother", "mother", "father", "mother",~
## $ traveltime <dbl> 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 2, 1, 1, 1, 3, 1, 1,~
## $ studytime  <dbl> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3, 2, 1, 1,~
## $ failures   <dbl> 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0,~
## $ schoolsup  <chr> "yes", "no", "yes", "no", "no", "no", "no", "yes", "no", "n~
## $ famsup     <chr> "no", "yes", "no", "yes", "yes", "yes", "no", "yes", "yes",~
## $ paid       <chr> "no", "no", "yes", "yes", "yes", "yes", "no", "no", "yes", ~
## $ activities <chr> "no", "no", "no", "yes", "no", "yes", "no", "no", "no", "ye~
## $ nursery    <chr> "yes", "no", "yes", "yes", "yes", "yes", "yes", "yes", "yes~
## $ higher     <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", "ye~
## $ internet   <chr> "no", "yes", "yes", "yes", "no", "yes", "yes", "no", "yes",~
## $ romantic   <chr> "no", "no", "no", "yes", "no", "no", "no", "no", "no", "no"~
## $ famrel     <dbl> 4, 5, 4, 3, 4, 5, 4, 4, 4, 5, 3, 5, 4, 5, 4, 4, 3, 5, 5, 3,~
## $ freetime   <dbl> 3, 3, 3, 2, 3, 4, 4, 1, 2, 5, 3, 2, 3, 4, 5, 4, 2, 3, 5, 1,~
## $ goout      <dbl> 4, 3, 2, 2, 2, 2, 4, 4, 2, 1, 3, 2, 3, 3, 2, 4, 3, 2, 5, 3,~
## $ Dalc       <dbl> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1,~
## $ Walc       <dbl> 1, 1, 3, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2, 1, 4, 3,~
## $ health     <dbl> 3, 3, 3, 5, 5, 5, 3, 1, 1, 5, 2, 4, 5, 3, 3, 2, 2, 4, 5, 5,~
## $ absences   <dbl> 6, 4, 10, 2, 4, 10, 0, 6, 0, 0, 0, 4, 2, 2, 0, 4, 6, 4, 16,~
## $ G1         <dbl> 5, 5, 7, 15, 6, 15, 12, 6, 16, 14, 10, 10, 14, 10, 14, 14, ~
## $ G2         <dbl> 6, 5, 8, 14, 10, 15, 12, 5, 18, 15, 8, 12, 14, 10, 16, 14, ~
## $ G3         <dbl> 6, 6, 10, 15, 10, 15, 11, 6, 19, 15, 9, 12, 14, 11, 16, 14,~
```

Data Analysis Plan

To answer the research question : What family circumstances are correlated with student alcohol consumption, which may allow us to gain better insight into how family education, presence, or occupation may be associated with alcohol use , we use Dalc (Weekday alcohol consumption) and Walc(Weekend alcohol consumption) as response variables and we will use famrel(Quality of family relationships , Pstatus(Parent's cohabitation) , Medu (Mother's education) , Fedu (Father's education), romantic (With a romantic relationship) Mjob(Mother's job) , Fjob(Father's job) , guardian (Student's guardian) , famsup (Family support) as explanatory variables. To answer the research question: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? , we will use Dalc (Weekday alcohol consumption) and Walc(Weekend alcohol consumption) as response variables and we will use failures(Number of past class) , activities(Extra-curricular activities) , goout (Going out with friends) , absences(Number of absences) and G1, G2, G3 which are respectively First period grade, Second period grade Final period grade , studytime (Weekly study time). We will compare alcohol consumption among different groups of age , sex and family size. We will try to find if there is a difference in the means alcohol consumption between the sex male and the sex female. We will study the difference in means of alcohol consumption among ages like students younger and students aged of 18 or greater. We will also compare alcohol consumption between students whose family size is less or equal than 3 and students whose family size is greater than 3. We will make a scatter plot visualization and the correlation matrix to visualize the correlations. between the response variable and the explanatory variables. The scatter plot also will help us to see if the regression model fits the data or to test if our model is good. We will use the summary statistic to find the p_value. For example, from the summary and using the single regression model analysis, we see that the number of absences is highly associated with weekday alcohol consumption with a p value less than 0.05.
The statistical method we will use is multiple regression analysis with anova test, chi square test for categorical variables to find the relationship between a response variable Y and explanatory variables Xi. The p values we obtain indicate if we reject or accept our null hypothesis, that is finding if there is a relationship between the response variable and the explanatory variables.

```
linear_reg() %>%
  set_engine("lm") %>%
  fit(Dalc ~failures, data = student) %>%
  tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    1.43     0.0488     29.3  9.79e-101
## 2 failures       0.163    0.0599      2.72 6.77e-  3
```

```
cor.test(student$Dalc, student$failures,
                    method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  student$Dalc and student$failures
## t = 2.7223, df = 393, p-value = 0.006771
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.03788446 0.23160887
## sample estimates:
##       cor
## 0.1360469
```

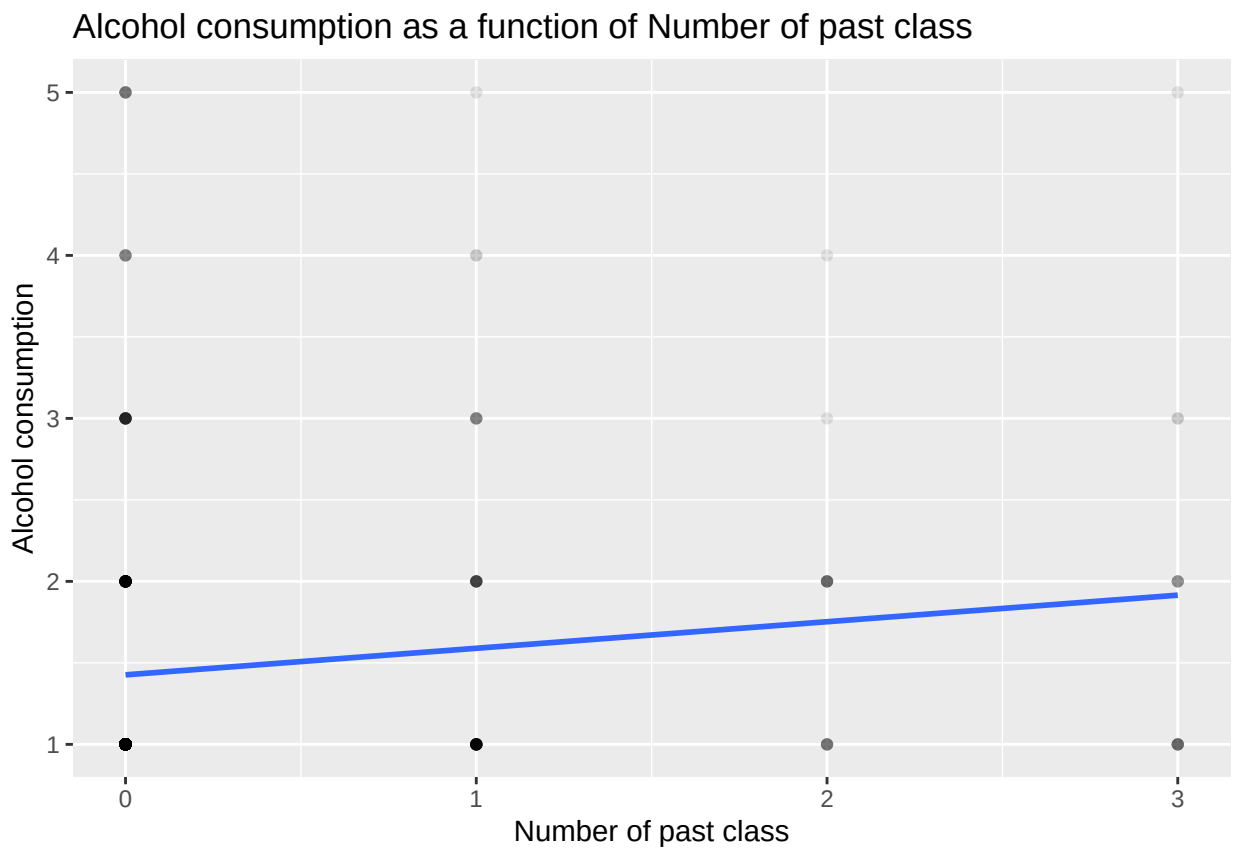
From the summary, we see that there is a positive relationship between alcohol consumption and Number of past class with a p_value of 6.770813e-03. In other words, alcohol consumption and Number of past class are correlated. In this model , Dalc= 6.770813e-03 *faiures + 1.4265564.

The correlation test also shows that the correlation between Dalc and failures is not 0 with a p_value of 0.006771.

```
ggplot(data = student, aes(x = failures, y = Dalc)) +
  geom_point(alpha=.1) +
  geom_smooth(method = "lm",
              se = FALSE) +
```

```
  labs(
    title = "Alcohol consumption as a function of Number of past class",
    x = "Number of past class",
    y = "Alcohol consumption"
  )
```

```
## `geom_smooth()` using formula 'y ~ x'
```



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.