Project proposal

Fat Goldfish: Daniel Bolja, Elyse McFalls, Damla Ozdemir, Nathan Nguyen

October 9th, 2020

Section 1. Introduction

There has been a recent study that looked at changes in the drinking habits of Americans, which finds that "Americans Are Drinking 14% More Often During Pandemic"[1], and "Instances of heavy drinking among women, which for women was defined as four or more drinks within a couple of hours, spiked by 41%"[2]. Although these studies did not look at younger adults, we found this conclusion interesting, so we want to analyze data having something to do with alcohol consumption in a younger population. We want to investigate the relationship between alcohol consumption in younger age groups and its effects on academic performance. We referenced a study using various types of data to try and predict secondary school performance of Portugese students, with weekend and workday alcohol consumption being two important variables to consider. We want to determine if alcohol consumption is a significant predictor of academic success. Based on the variables of the dataset, we hypothesize that:

- Alcohol consumption will significantly predict first period/ second period/ final grades (depending on what grades we want to look at) after controlling for other variables. There will be a negative relationship between grades and alcohol consumption.
- Alcohol consumption will significantly predict weekly study times after controlling for other variables. There will be a negative relationship between weekly study times and alcohol consumption.
- Alcohol consumption will significantly predict absences after controlling for other variables. There will be a positive relationship between absences and alcohol consumption.
- Alcohol consumption will significantly predict the number of failed classes after controlling for other variables. There will be a positive relationship between the number of failed classes and alcohol consumption.

Section 2. Data description

We are using the data from the study of Cortez and Silva (2008), which focused on the study performance of secondary students based on their alcohol consumption. The data contains 649 observations which are secondary students enrolled in Portuguese languages from two public schools in Alenjeto region of Portugal during the 2005 - 2006 school year. Since at this time, the majority of Portuguese public schools' information systems remained very poor, the authors built the database from two sources: paper-based school reports with few attributes from grades and number of absences, and questionnaires about several demographic, social/emotional and school related variables that were expected to affect student performance. Data with lack of identification details was discarded.

The details of variables are shown below, with four last attributes taken from school reports.

Variables	Meaning
`school`	student's school ('GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
`sex`	student's sex ('F' - female or 'M' - male)
`age`	student's age (from 15 to 22)

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7369979/
[2] https://www.npr.org/2020/10/05/920437811/americans-are-drinking-14-more-often-during-pandemic-study-finds

Variables	Meaning
address	student's home address type ('U' - urban or 'R' - rural)
famsize	family size ('LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus	parent's cohabitation status ('T' - living together or 'A' - apart)
Medu	mother's education (0 to 4)
Fedu	father's education (0 to 4)
Mjob	mother's job ('teacher', 'health' care related, civil 'services' , 'at_home' or 'other')
Fjob	father's job ('teacher', 'health' care related, civil 'services' , 'at_home' or 'other')
reason	reason to choose this school (close to 'home', school 'reputation', 'course' preference or 'other')
guardian	student's guardian ('mother', 'father' or 'other')
traveltime	home to school travel time ($1 - 10$ hours)
studytime	weekly study time ($1 - 10$ hours)
failures	number of past class failures (n if $0 \leq n < 3$, else 4)
schoolsup	extra educational support
famsup	family educational support
paid	extra paid classes within the course subject
activities	extra-curricular activities
nursery	attended nursery school
higher	wants to take higher education
internet	Internet access at home
romantic	with a romantic relationship
famrel	quality of family relationships
freetime	free time after school
goout	going out with friends
Dalc	workday alcohol consumption
Walc	weekend alcohol consumption
health	current health status
absences	number of school absences (from 0 to 93)
G1	first period grade
G2	second period grade
G3	final grade

Note: Mother and father education (`Medu` and `Fedu`) were valued from 0 to 4 with: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education. For `famrel`, `freetime`, `goout`, `Dalc`, `Walc` and `health` variables, the values were taken from 1 (very low/bad) to 5 (very high/excellent). Variables `schoolsup`, `famsup`, `paid`, `activities`, `nursery`, `higher`, `internet`, and `romantic` variables are measured with yes or no. Grades variables were taken from 0 to 20.

Section 3. Glimpse of data

```
alcohol <- read.csv("data/student-por.csv")
```

```
glimpse(alcohol)
```

```
## Rows: 649
## Columns: 33
## $ school     <chr> "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "G...
## $ sex        <chr> "F", "F", "F", "F", "F", "M", "M", "F", "M", "M", "F", "...
## $ age        <int> 18, 17, 15, 15, 16, 16, 16, 17, 15, 15, 15, 15, 15, 15, ...
## $ address    <chr> "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "...
## $ famsize    <chr> "GT3", "GT3", "LE3", "GT3", "GT3", "LE3", "LE3", "GT3", ...
## $ Pstatus    <chr> "A", "T", "T", "T", "T", "T", "T", "A", "A", "T", "T", "...
```

```
## $ Medu       <int> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4, 3, 3,...
## $ Fedu       <int> 4, 1, 1, 2, 3, 3, 2, 4, 2, 4, 4, 1, 4, 3, 2, 4, 4, 3, 2,...
## $ Mjob       <chr> "at_home", "at_home", "at_home", "health", "other", "ser...
## $ Fjob       <chr> "teacher", "other", "other", "services", "other", "other...
## $ reason     <chr> "course", "course", "other", "home", "home", "reputation...
## $ guardian   <chr> "mother", "father", "mother", "mother", "father", "mothe...
## $ traveltime <int> 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 2, 1, 1, 1, 3, 1,...
## $ studytime  <int> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3, 2, 1,...
## $ failures   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3,...
## $ schoolsup  <chr> "yes", "no", "yes", "no", "no", "no", "no", "yes", "no",...
## $ famsup     <chr> "no", "yes", "no", "yes", "yes", "yes", "no", "yes", "ye...
## $ paid       <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "n...
## $ activities <chr> "no", "no", "no", "yes", "no", "yes", "no", "no", "no", ...
## $ nursery    <chr> "yes", "no", "yes", "yes", "yes", "yes", "yes", "yes", "...
## $ higher     <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", ...
## $ internet   <chr> "no", "yes", "yes", "yes", "no", "yes", "yes", "no", "ye...
## $ romantic   <chr> "no", "no", "no", "yes", "no", "no", "no", "no", "no", "...
## $ famrel     <int> 4, 5, 4, 3, 4, 5, 4, 4, 4, 5, 3, 5, 4, 5, 4, 4, 3, 5, 5,...
## $ freetime   <int> 3, 3, 3, 2, 3, 4, 4, 1, 2, 5, 3, 2, 3, 4, 5, 4, 2, 3, 5,...
## $ goout      <int> 4, 3, 2, 2, 2, 2, 4, 4, 2, 1, 3, 2, 3, 3, 2, 4, 3, 2, 5,...
## $ Dalc       <int> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2,...
## $ Walc       <int> 1, 1, 3, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2, 1, 4,...
## $ health     <int> 3, 3, 3, 5, 5, 5, 3, 1, 1, 5, 2, 4, 5, 3, 3, 2, 2, 4, 5,...
## $ absences   <int> 4, 2, 6, 0, 0, 6, 0, 2, 0, 0, 2, 0, 0, 0, 0, 6, 10, 2, 2...
## $ G1         <int> 0, 9, 12, 14, 11, 12, 13, 10, 15, 12, 14, 10, 12, 12, 14...
## $ G2         <int> 11, 11, 13, 14, 13, 12, 12, 13, 16, 12, 14, 12, 13, 12, ...
## $ G3         <int> 11, 11, 12, 14, 13, 13, 13, 13, 17, 13, 14, 13, 12, 13, ...
```

Bibliography

1. P. Cortez and A. Silva. Using data mining to pre-dict secondary school student performance. 2008.
2. Student Alcohol Consumption dataset: https://www.kaggle.com/uciml/student-alcohol-consumption