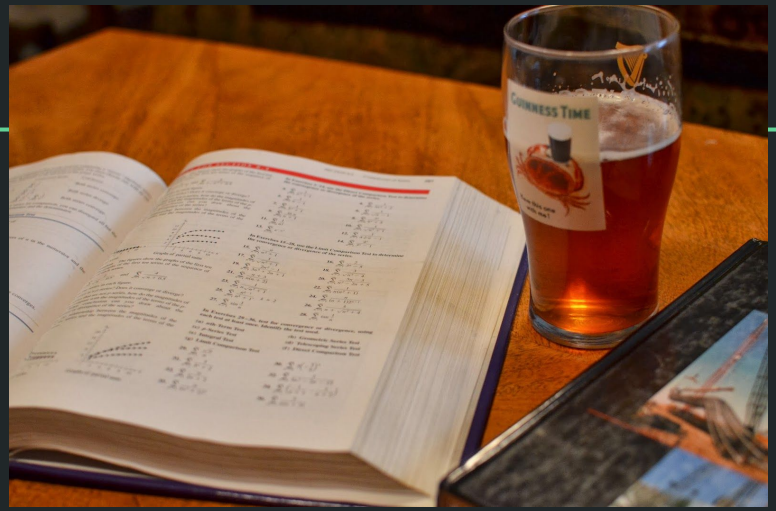
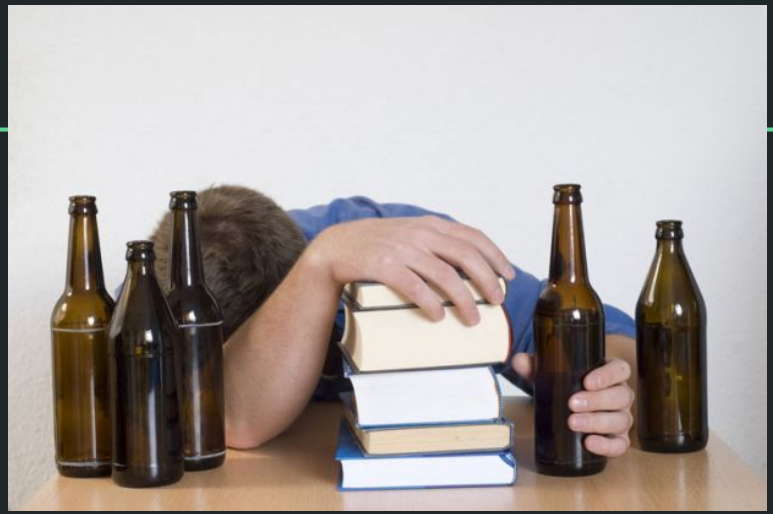
Analyzing the relationship between student alcohol consumption and school performance

By Daniel Bolja, Elyse McFalls, Nathan Nguyen, Damla Ozdemir





Introduce topic and motivation

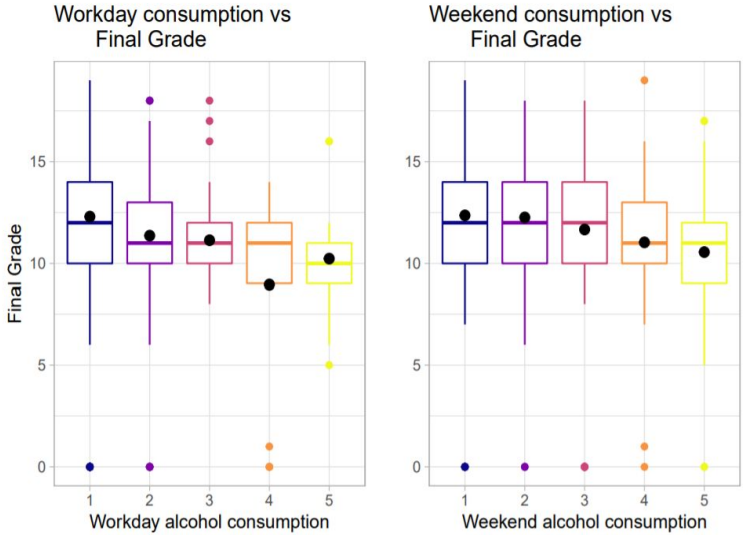
- Recent study that looked at changes in the drinking habits of Americans, which finds that “Americans Are Drinking 14% More Often During Pandemic”, and “Instances of heavy drinking among women, which for women was defined as four or more drinks within a couple of hours, spiked by 41%” ➜ contemporary relevance
- Studies did not look at younger adults, but we found this conclusion interesting, so we want to analyze data having something to do with alcohol consumption in a younger population.
- Investigation: 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 Portuguese 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, especially in grades, weekly study time, absences, and number of failed classes.**

Introduce the data

- The data was from the study of Cortez and Silva (2008), which **focused on the study performance of secondary students based on their alcohol consumption** (“Using data mining to predict secondary school student performance”)
- The data contains **649 observations** which are secondary students enrolled in Portuguese languages from two public schools in the Alenjeto region of Portugal during the 2005 - 2006 school year.
- The authors built the database from **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 were discarded.
- There are a total of 33 variables, where our main focus will be **workday (`Dalc`)** and **weekend (`Walc`) alcohol consumption**, **first**, **second**, **and third period grades (`G1`, `G2`, `G3`)**, **weekly study time (`studytime`)**, **absences (`absences`)** and the **number of failed classes (`failure`)**.

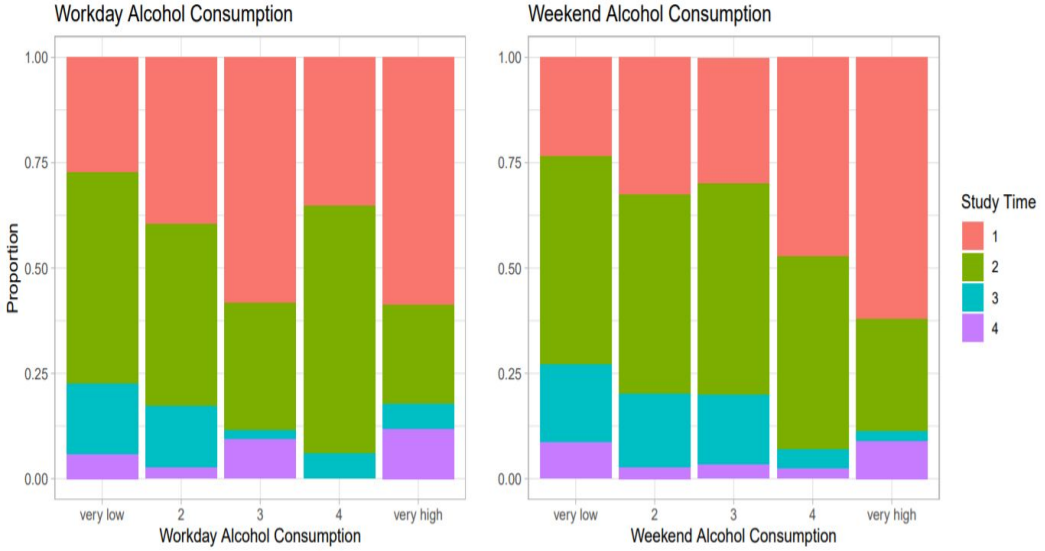
Highlights from Exploratory Data Analysis

Relationship between Final Grades and Alcohol Consumption



Final grades appear to decrease with an increase in alcohol consumption. This relationship is more clearly seen with weekend alcohol consumption.

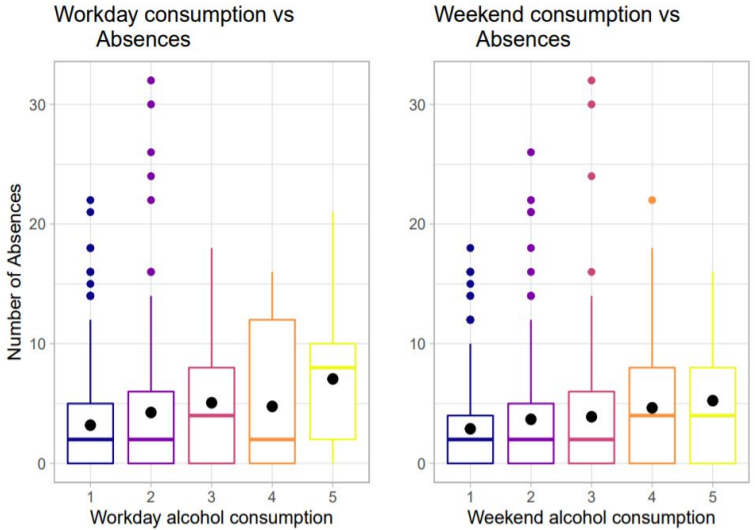
Relationship between Weekly Study Times and Alcohol Consumption



The proportion of students spending less time studying noticeably increases with an increase in weekend alcohol consumption. However, this trend isn't clearly seen with workday alcohol consumption

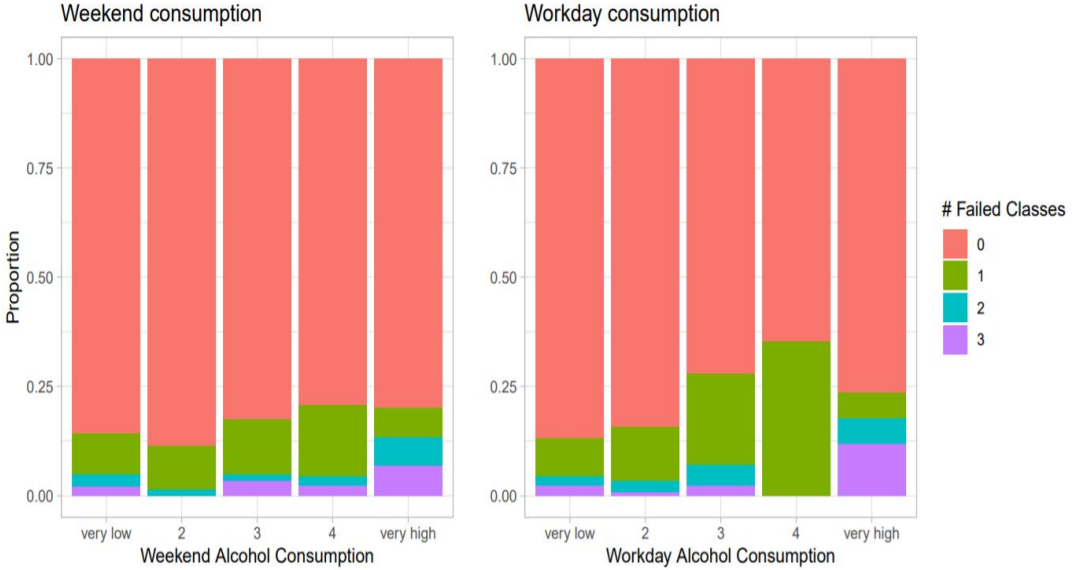
Highlights from Exploratory Data Analysis (cont.)

Relationship between the Absences and Alcohol Consumption



There appears to be a positive relationship between the number of absences and the severity of one's alcohol consumption during the week and on the weekends. However, this trend seems more stable when dealing with weekend alcohol consumption

Relationship between the Number of Failed Classes and Alcohol Consumption



The proportion of students failing at least two classes appears to increase with an increase in weekend and workday alcohol consumption

Hypothesis testing

- ***Grades***: Simulation-based testing for the negative correlation between **workday** and **weekend alcohol consumption** and **final grade (G3)** give very small p-values (nearly 0). Linear models are used to check relationship between **alcohol consumption** and **final grade (G3)** and give:

$$\widehat{\mathrm{G3}} = 12.36 - 0.100 \times \mathrm{Walc2} - 0.694 \times \mathrm{Walc3} - 1.326 \times \mathrm{Walc4} - 1.805 \times \mathrm{Walc5}.$$

$$\widehat{\mathrm{G3}} = 12.30 - 0.936 \times \mathrm{Dalc2} - 1.160 \times \mathrm{Dalc3} - 3.358 \times \mathrm{Dalc4} - 2.06 \times \mathrm{Dalc5}.$$

- ***Weekly study times***: Chi-square tests between **weekly study times** and **alcohol consumption** give small p-values (nearly 0). Simulation-based null distributions of correlation between **workday** and **weekend alcohol consumption** and **weekly study times** give very small p-values (nearly 0).
- ***Absences***: Linear model predicting **the number of absences** from **workday and weekend alcohol consumption** give positive slopes of both daily (1.433) and weekly (0.594) consumption. Hypothesis testing for positive correlation of **absences and alcohol consumption** gives small p-values (both 0), even if we consider students lived “closed” to their school. (0.0181 and 0.0009 for workday and weekend, respectively)
- ***Number of failed classes***: Chi-square test between **# failed classes** and **workday alcohol consumption** give a really small p-value (0.0033). However, the chi-square test for **weekend alcohol consumption** and **the number of failed** classes has a large p-value (0.134). Simulation-based testing for positive correlations between **alcohol consumption** and **number of failed classes** give two p-values of 0.0081 and 0.0233 for **workday** and **weekend** alcohol consumption, respectively, which is small.

Conclusion and future work

- It is safe to conclude that increased alcohol consumption is correlated with a lower level of academic performance.
    - Seen through our visualizations (i.e. **alcohol consumption and average grades**, the **relationship between study time and alcohol consumption**, and through the **hypothesis tests** conducted where a strong statistical connection between alcohol consumption and academic performance was observed.
- Reliability of data collection is questionable: 2 schools is adequate on small scale, but not enough to make a large-scale generalization of students in the Portuguese public school system.
- We don't know the accuracy of the students' answers.
- We lack some backgrounds of the data variables to have some more in-depth conclusions.
    - For example, the original study did not give any contexts about the failing threshold of the grading system in those schools. Therefore, we can only conclude significant negative correlations between grades and alcohol consumption, but we can't make further claims about whether students tend to have failed grades or not, given their alcohol consumption.
- Many of the variables are collected on a 1 to 5 scale and the information can be misleading.
- **WHAT TO CHANGE**: Analyze more about the relationship of the rest of the variables and see if we could get a better comprehensive conclusion based on more conditions (family status, internet, love status,...).
    - Around 33 variables in our data set and we only used about 10 in our analysis.
- Conduct a different statistical test (ANOVA) because the linear regression model is not as effective with our data as other tests could be.