Final Data Project

Xiaojia Ren

Introduction and Data

Research Question: How to Predict Secondary School Students' Weekend Alcohol Consumption Level Using Family, School and Personal Information?

Project Motivation: According to a CDC (2024) report about underage drinking, underage drinking is a significant public health problem in the U.S. The 2021 Youth Risk Behavior Survey found that underage drinking is more common than we think among high school students. In specific, among all high school students during the past month, 23% drank alcohol, 11% binge drank, 5% of drivers drove after drinking alcohol, and 14% rode with a driver who had been drinking alcohol. Underage Drinking has posed great risks to teenagers' health, safety, and school performance. The Substance Abuse and Mental Health Services Administration (2022) identified several reasons why teenagers might engage in underage drinking activities including stress from school, peer pressure to drink, life transitions such as breakups and school transfers, and family environment. Given the alarming statistics on underage drinking among high school students reported by the CDC, this project aims to predict secondary school students' weekend alcohol consumption levels using family, school, and personal information, which are potential reasons identified by Substance Abuse and Mental Health Services Administration for why teenagers start consuming alcohol.

Data Introduction: The data is obtained from the survey responses of students enrolled in Portuguese language courses in secondary schools in Portugal (UCI Machine Learning, 2016). It contains many variables including socioeconomic status, family status, and school performance information about the students. The key variables that I have chosen for the analysis purpose of the project are listed as follows:

Response Variable

Walc - weekend alcohol consumption (categorical: from 1 - very low, 2 - low, 3 - medium, 4 - high to 5 - very high)

Predictors

Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart) famsize - family size (binary: ‘LE3’ - less or equal to 3 or ‘GT3’ - greater than 3)

absences - number of school absences (numeric: from 0 to 93)

studytime - weekly study time (categorical: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

G3 - final grade of the Portuguese class (numeric: from 0 to 20)

romantic - whether the student is in a romantic relationship (binary: yes or no)

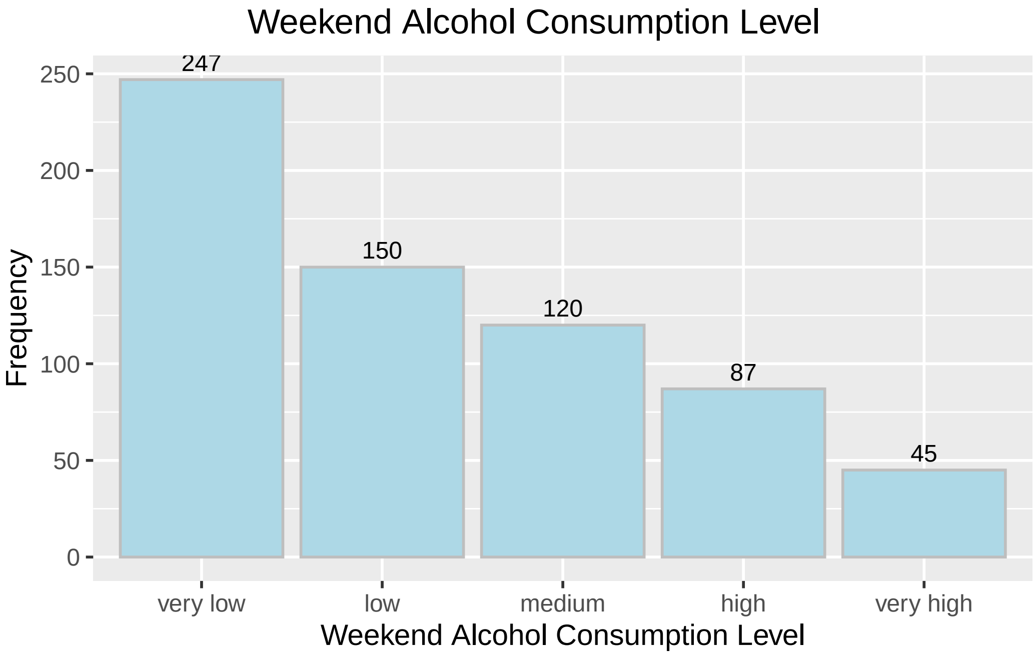
goout - going out with friends (categorical: from 1 - very low, 2 - low, 3 - medium, 4 - high to 5 - very high)

health - current health status (numeric: from 1 - very bad to 5 - very good)

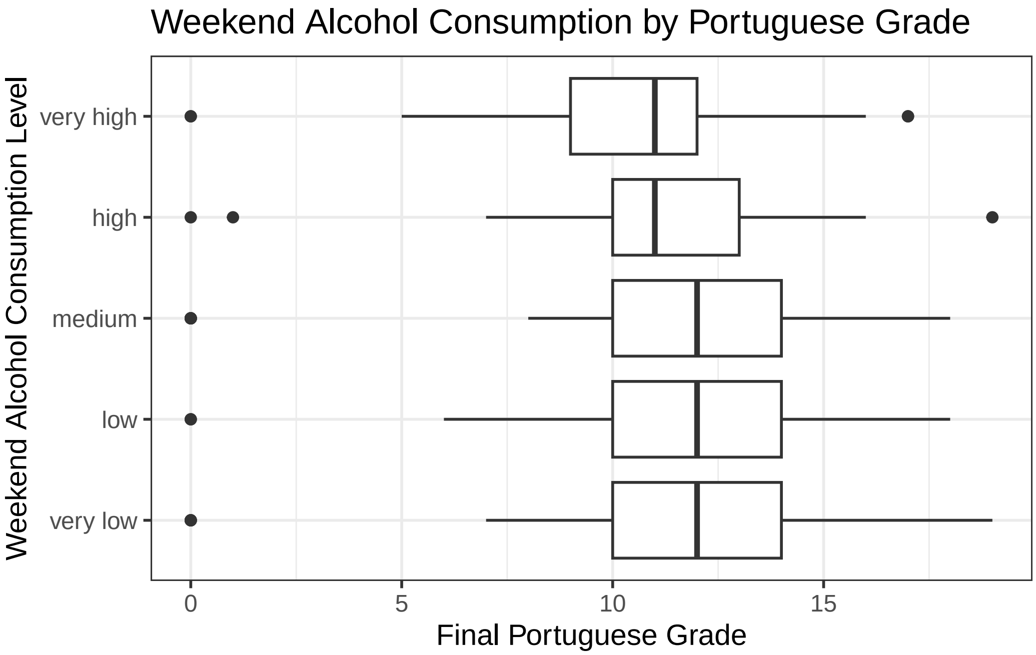
Data Cleaning: The outcome variable’s categories are renamed from 1 to very low, 2 to low, 3 to medium, 4 to high, and 5 to very high for better visualization and analysis purposes.

Table 1: Summary Statistics of Weekend Alcohol Consumption

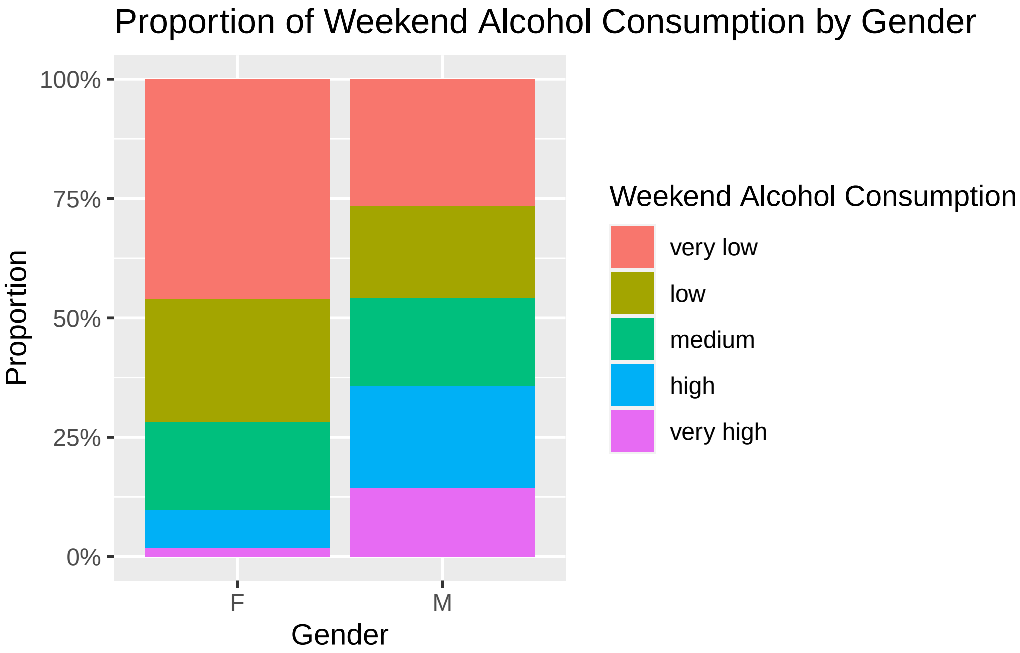
Category	Frequency	Proportion	Mode
very low	247	0.3806	1
low	150	0.2311	1
medium	120	0.1849	1
high	87	0.1341	1
very high	45	0.0693	1



From the summary statistics and the histogram of the weekend alcohol consumption level, we can conclude that the most common alcohol consumption level on weekends for secondary school students is “very low”, with a frequency of 247. There is also a decrease in frequency as the consumption level increases.



Examining the boxplot of weekend alcohol consumption by final Portuguese grade, we can see that the higher the weekend alcohol consumption level is, the lower the median final Portuguese grade is, potentially suggesting that worse academic performances at school are associated with higher weekend alcohol consumption levels.



Examining the proportion of weekend alcohol consumption levels by gender graph, we can see that male secondary school students seem to consume more alcohol than female students in general.

Methodology

Table 2: Brant-Wald Test Results for Proportional Odds Assumption

Test	X2	df	Probability
Omnibus	59.88	48	0.12
parents are living together	3.17	3	0.37
family size less or equal to 3	0.92	3	0.82
number of absences	0.9	3	0.83
weekly study time 2 to 5 hours	6.16	3	0.1
weekly study time 5 to 10 hours	6.06	3	0.11
weekly study time larger than 10 hours	5.71	3	0.13
final grade of the Portugese class	1.79	3	0.62

Test	X2	df	Probability
the student is in a romantic relationship	2.22	3	0.53
bad current health status	2.25	3	0.52
medium current health status	3.01	3	0.39
good current health status	2.17	3	0.54
very good current health status	0.82	3	0.84
time going out with friends is low	2.21	3	0.53
time going out with friends is medium	3.37	3	0.34
time going out with friends is high	0.34	3	0.95
time going out with friends is very high	0.21	3	0.98

In this study, an ordinal regression model was employed to analyze the weekend alcohol consumption levels among secondary school students. The reason to employ an ordinal regression model is that the response variable, Walc - weekend alcohol consumption, is an ordered and categorical variable, which satisfies the outcome requirement for an ordinal regression model. The predictors considered for this model are from three categories, respondent's family status, school performance, and personal factors. After a careful and thorough examination of the dataset, two predictors were selected to represent family status or environment, and three predictors were selected to represent school performance and personal factors, resulting in a total of eight predictors.

The proportional odds assumption, a critical assumption for the validity of ordinal regression, was tested using the Brant-Wald test. The results indicated that the p-values for Omnibus and all individual predictors are larger than 0.05. This outcome suggests that we fail to reject the null hypothesis, thereby supporting the assumption that the proportional odds condition is satisfied within this model (McNulty, 2021).

Results

Table 3: Regression Results: Coefficients

Predictors	Value	std.error	t_value
parents are living together	0.6023	0.2429	2.4800
family size less or equal to 3	0.5002	0.1638	3.0530
number of absences	0.0497	0.0159	3.1230
weekly study time 2 to 5 hours	-0.5912	0.1717	-3.4420
weekly study time 5 to 10 hours	-0.7949	0.2406	-3.3030
weekly study time larger than 10 hours	-1.0786	0.3807	-2.8340
final grade of the Portuguese class	-0.0522	0.0244	-2.1410
the student is in a romantic relationship	-0.0964	0.1559	-0.6180

Predictors	Value	std.error	t_value
bad current health status	0.4404	0.2935	1.5010
medium current health status	0.5543	0.2715	2.0420
good current health status	0.5866	0.2775	2.1140
very good current health status	0.6762	0.2399	2.8190
time going out with friends is low	0.8550	0.3608	2.3700
time going out with friends is medium	1.4044	0.3466	4.0520
time going out with friends is high	2.0148	0.3626	5.5560
time going out with friends is very high	2.8051	0.3737	7.5060

Table 4: Regression Results: Intercepts

Term	Value	std.error	t_value
very low|low	1.1835	0.5343	2.2151
low|medium	2.3072	0.5406	4.2677
medium|high	3.4049	0.5474	6.2201
high|very high	4.8313	0.5643	8.5609

Table 5: P-values for all coefficients and intercepts

Predictors	p_value
parents are living together	0.01316
family size less or equal to 3	0.002262
number of absences	0.001793
weekly study time 2 to 5 hours	0.0005766
weekly study time 5 to 10 hours	0.0009552
weekly study time larger than 10 hours	0.004603
final grade of the Portuguese class	0.0323
the student is in a romantic relationship	0.5366
bad current health status	0.1335
medium current health status	0.04116
good current health status	0.03448
very good current health status	0.004813
time going out with friends is low	0.0178
time going out with friends is medium	5.088e-05
time going out with friends is high	2.757e-08
time going out with friends is very high	6.128e-14
very low|low	0.02675
low|medium	1.975e-05

Predictors	p_value
medium|high	4.969e-10
high|very high	0

Table 6: Exponentiation of Coefficients

Predictors	Exp_Coeff
parents are living together	1.826
family size less or equal to 3	1.649
number of absences	1.051
weekly study time 2 to 5 hours	0.5537
weekly study time 5 to 10 hours	0.4516
weekly study time larger than 10 hours	0.3401
final grade of the Portuguese class	0.9491
the student is in a romantic relationship	0.9081
bad current health status	1.553
medium current health status	1.741
good current health status	1.798
very good current health status	1.966
time going out with friends is low	2.352
time going out with friends is medium	4.073
time going out with friends is high	7.499
time going out with friends is very high	16.53

The p-values for all of the coefficients are presented above. We can see that the p-values for all of the coefficients, except for the student who is in a romantic relationship and bad current health status, are all smaller than 0.05.

If we set the following, Null Hypothesis: The coefficients equal 0, which implies that the predictors have no effect on the log odds of the outcome variable being in a particular category or above versus all lower categories combined. Alternative Hypothesis: The coefficients do not equal 0, which implies that the predictors do have an effect on the log odds of the outcome variable being in a particular category or above.

This suggests that the results are statistically significant and we reject the null hypothesis except that the student is in a romantic relationship and bad current health status.

In terms of family environment and family relationships, we can conclude that better or tighter family relationships or smaller family sizes will not hinder secondary school students from drinking on weekends. They are more likely to consume more alcohol on weekends. In fact, secondary school students whose parents are living together have approximately 1.83 the odds of having the next higher weekend alcohol consumption category compared to secondary students whose parents are living apart. In addition, secondary students whose family size is less or equal to 3 have approximately 1.65 the odds of having the next higher weekend alcohol consumption category compared to secondary students whose family size is greater than 3.

In terms of school performance, we can conclude that secondary school students with better school performances and longer study time are more likely to consume less alcohol during weekends. In specific, as the weekly study time for secondary school students increases, the odds of them being in the next higher weekend alcohol consumption category decreases. For every 1-point increase in the final grade of the Portuguese class, the odds of being in the next higher weekend alcohol consumption category is multiplied by approximately 0.95. However, for every 1 unit more increase in absences, the odds of being in the next higher weekend alcohol consumption category is multiplied by approximately 1.05, suggesting that poorer school performances are linked with more weekend alcohol consumption while better school performance are associated with lower weekend alcohol consumption levels.

In terms of personal relationships and health conditions, we can conclude that secondary school students who are healthy and go out with friends more often are more likely to consume more alcohol during weekends compared to those who are in very bad health conditions and spend a very little amount of time going out with friends. We can see that as health status improves and go-out time increases, the odds of being in the next high weekend alcohol consumption category also increase, suggesting that students in good health conditions and who go out more often are linked with higher weekend alcohol consumption.

Table 7: Model Accuracy

Metric	Value
Accuracy	0.3875

The predictive accuracy of the ordinal regression model, as assessed by the overall classification accuracy, is 0.3875 (38.75%). This metric was calculated by comparing the predicted alcohol consumption levels against the actual weekend alcohol consumption levels observed in the dataset.

Discussion

For secondary school students, better family relationships do not necessarily guarantee a low weekend alcohol consumption level. In fact, secondary school students who have a smaller family size and tighter family relationships might consume more alcohol during weekends. However, better school performances are usually associated with lower weekend alcohol consumption. In addition, students who are in good health status and go out with friends more often are linked with higher weekend alcohol consumption levels.

In summary, high levels of weekend alcohol consumption levels for secondary school students are linked with a better family environment, tighter family relationships, smaller family size, worse school performance, shorter study time, better health conditions, and more time spent going out with friends.

One of the biggest limitations of this analysis is that it has an exclusive focus on weekend alcohol consumption levels neglecting the weekday alcohol consumption data (Dalc) available in the dataset. As a result, this analysis failed to comprehensively analyze a secondary school student's weekly alcohol consumption level. Considering ways that the analysis could be improved, a new response variable named Weekalc that includes both weekday alcohol consumption levels and weekend alcohol consumption levels can be created to represent the weekly alcohol consumption level of secondary school students.New categories of the variable can be set, such as very low for secondary school students whose weekend alcohol consumption and weekday alcohol consumption are both very low. This could also be an idea for future work, examining weekly alcohol consumption levels. Furthermore, the dataset contains extensive socioeconomic information that remains unexplored in the current analysis. The relatively low accuracy indicates potential areas for improvement as well. It suggests that the model might be underfitting or lacking in relevant predictors that capture the complexities of alcohol consumption behavior among secondary school students. Both of the problems can be solved by including new predictors in the model to predict the weekend alcohol consumption level of a secondary school student. Incorporating these additional predictors could enhance the model's ability to predict alcohol consumption patterns more accurately.

Citations

CDC. (2024, April 16). Underage drinking. Centers for Disease Control and Prevention. https://www.cdc.gov/alcohol/fact-sheets/underage-drinking.htm

McNulty, K. (2021b, January 1). Proportional Odds Logistic Regression for Ordered Category Outcomes: Handbook of Regression Modeling in people analytics: With examples in R, python and julia. https://peopleanalytics-regression-book.org/ord-reg.html#wald

SAMHSA. (2022). Why Your Child Might Start Drinking Alcohol.

https://www.samhsa.gov/sites/default/files/why-child-might-start-drinking_0.pdf

UCI Machine Learning. (2016, October 19). Student Alcohol Consumption. Kaggle. https://www.kaggle.com/datasets/uciml/student-alcohol-consumption