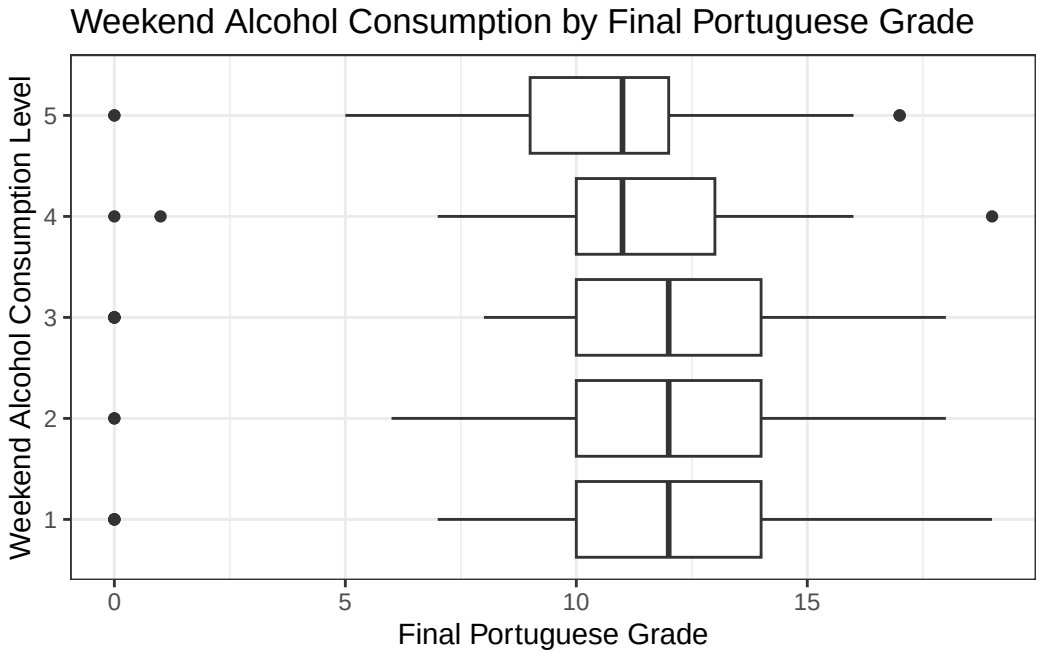
Final Data Project

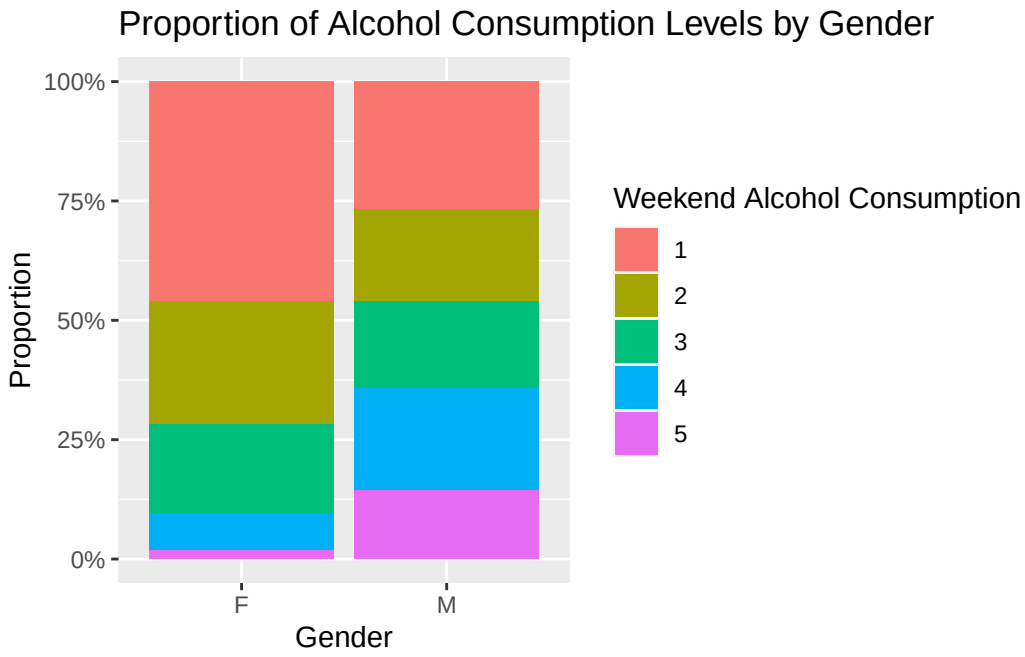
Xiaojia Ren

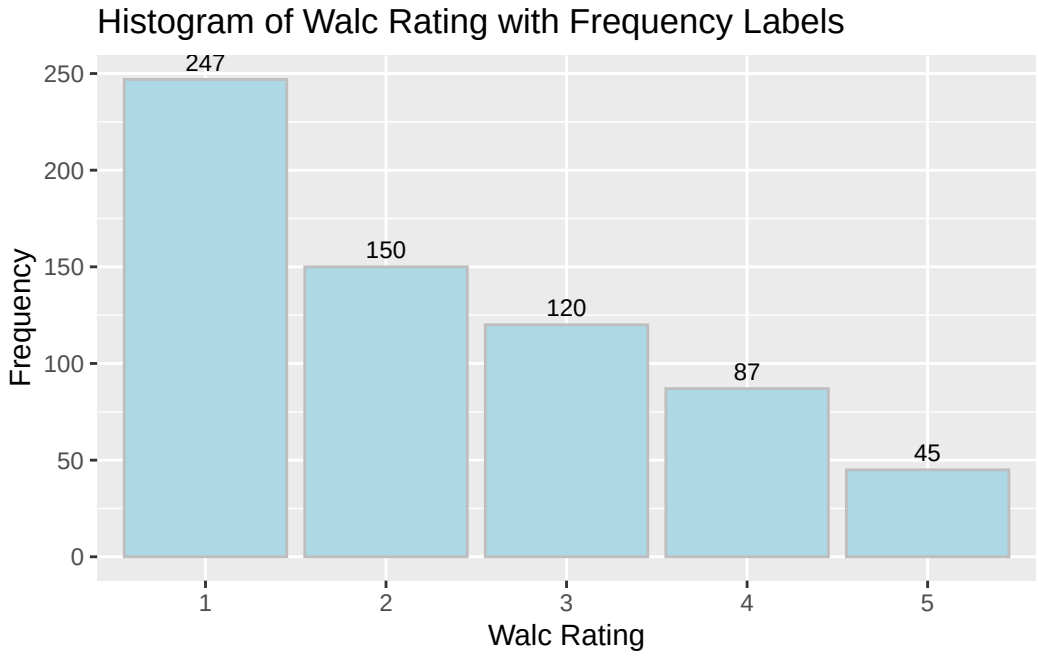
Introduction and Data

```
[1] "Summary Statistics of Weekend Alcohol Consumption"
```

```
     Category Frequency Proportion           Mode
[1,] "1"      "247"     "0.380585516178737"  "1"
[2,] "2"      "150"     "0.231124807395994"  "1"
[3,] "3"      "120"     "0.184899845916795"  "1"
[4,] "4"      "87"      "0.134052388289676"  "1"
[5,] "5"      "45"      "0.0693374422187981" "1"
```







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 among high school students, during the past 30 days, 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. Shocked by the statisitcs in the report, this project tries to predict secondary school students' weekend alcohol consumption level using family, school and personal information

Data Introduction: The data is obtained from the survey responses of students enrolled in the Portuguese language courses in secondary schools in Portugal. It contains many predictors including the socioeconomic 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 (caregorical: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

G3 - final grade of the Portugese 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)

Examing the graphs, we can see that male secondary school students seem to more alcohol than female students.

Citations: https://www.kaggle.com/datasets/uciml/student-alcohol-consumption

https://www.cdc.gov/alcohol/fact-sheets/underage-drinking.htm

Methodology

```
Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select


Re-fitting to get Hessian

Call:
polr(formula = factor(Walc) ~ Pstatus + famsize + absences +
    studytime + G3 + romantic + health + goout, data = alcohol)

Coefficients:
               Value Std. Error t value
PstatusT     0.60225    0.24289   2.480
famsizeLE3   0.50024    0.16383   3.053
absences     0.04972    0.01592   3.123
studytime2  -0.59119    0.17174  -3.442
studytime3  -0.79492    0.24064  -3.303
studytime4  -1.07864    0.38067  -2.834
G3          -0.05221    0.02439  -2.141
romanticyes -0.09635    0.15591  -0.618
health2      0.44041    0.29350   1.501
health3      0.55431    0.27147   2.042
health4      0.58675    0.27750   2.114
health5      0.67622    0.23986   2.819
goout2       0.85508    0.36083   2.370
goout3       1.40437    0.34663   4.052
goout4       2.01477    0.36261   5.556
goout5       2.80513    0.37374   7.506

Intercepts:
    Value   Std. Error t value
1|2  1.1835  0.5343     2.2151
2|3  2.3072  0.5406     4.2677
3|4  3.4049  0.5474     6.2201
4|5  4.8313  0.5643     8.5609
```

```
Residual Deviance: 1742.282
AIC: 1782.282
```

```
Installing package into '/home/guest/R/x86_64-pc-linux-gnu-library/4.3'
(as 'lib' is unspecified)
```

```
--------------------------------------------
Test for    X2  df  probability
--------------------------------------------
Omnibus     59.88   48  0.12
PstatusT    3.17    3   0.37
famsizeLE3  0.92    3   0.82
absences    0.9 3   0.83
studytime2  6.16    3   0.1
studytime3  6.06    3   0.11
studytime4  5.71    3   0.13
G3      1.79    3   0.62
romanticyes 2.22    3   0.53
health2     2.25    3   0.52
health3     3.01    3   0.39
health4     2.17    3   0.54
health5     0.82    3   0.84
goout2      2.21    3   0.53
goout3      3.37    3   0.34
goout4      0.34    3   0.95
goout5      0.21    3   0.98
--------------------------------------------
```

```
H0: Parallel Regression Assumption holds
```

The model that I fit here is an ordinal regression model. The reason why I chose to fit a 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 that I considered are from three perspective, respondent's family status, school performace, and personal factors. I carefully reviewed all of the columns in the dataset and picked out 2 predictors for family status and 3 for two other perspectives, a total of 8 predictors.

I used the Brant-Wald test to test proportional odds assumption and we can see from the output that the p value for Omnibus and all of the other predictors are larger than 0.05, this means that fail to reject H0: proportional odds assumtpion holds, suggesting that proportional odds assumption does hold.

Citations: https://peopleanalytics-regression-book.org/ord-reg.html#wald

```
Re-fitting to get Hessian

[1] "P values for all coefficients and intercepts"

    PstatusT   famsizeLE3     absences   studytime2   studytime3   studytime4
1.315645e-02 2.262194e-03 1.792861e-03 5.765920e-04 9.551868e-04 4.603479e-03
          G3  romanticyes      health2      health3      health4      health5
3.229552e-02 5.365932e-01 1.334781e-01 4.116076e-02 3.448171e-02 4.812825e-03
      goout2       goout3       goout4       goout5          1|2          2|3
1.780111e-02 5.087641e-05 2.756765e-08 6.128431e-14 2.675291e-02 1.975133e-05
         3|4          4|5
4.968956e-10 0.000000e+00

[1] "Exponentiation of Coefficients"

   PstatusT  famsizeLE3    absences  studytime2  studytime3  studytime4
  1.8262297   1.6491213   1.0509724   0.5536694   0.4516184   0.3400588
         G3 romanticyes     health2     health3     health4     health5
  0.9491285   0.9081460   1.5533369   1.7407472   1.7981273   1.9664380
     goout2      goout3      goout4      goout5
  2.3515650   4.0729748   7.4989664  16.5291729
```

Results

The p-value for all of the coefficients is presented above. We can see that the p value for all of the coefficients 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.

We cannot conclude that better or tighter family relationships or smaller family sizes will not hinder secondary school students from drinking on weekends since we can see from the output that secondary students whose parents are living together have approximately 1.84 the odds of having the next higher weekend alcohol consumption category (e.g., 4 vs. 3, or 2 vs. 1, etc.)

compared to secondary students whose parents are living apart and secondary students whose family size is less or equal to 3 have approximately 1.66 the odds of having the next higher weekend alcohol consumption category (e.g., 4 vs. 3, or 2 vs. 1, etc.) compared to secondary students whose family size is greater than 3, while controlling for other variables.

We can conclude that secondary school students with better school performances are more likely to consume less alcohol during weekends. As the output data shows, 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, and for every 1 point increase in G3 (final grade of the Portuguese class), the odds of being in the next higher weekend alcohol consumption category is multiplied by approximately 0.95 , while controlling for other variables. We can also see that as study time increases, the odds of being in the next high weekend alcohol consumption category decreaseas, suggesting that better school performance is linked with less weekend alcohol consumption.

We can conclude secondary school students who are in romantic relationships are less like consume alcohol during weekends. We can see from the output that secondary students whose are in relationships have approximately 0.91 the odds of having the next higher weekend alcohol consumption category (e.g., 4 vs. 3, or 2 vs. 1, etc.) compared to secondary students whose are single.

We can also conclude that secondary school students who are healthy and go out more often are morely likely to consume more alcohol during weekends as we can see that as health status and go out time increases, the odds of being in the nexr high weekend alcohol consumption category as increases, suggesting that students in good health conditions and who go out more often are linked with higher weekend alcohol consumption.

Discussion

In summary, for secondary school students, better family relationships does not necessarily guarantee a low weekend alcohol consumption level, however, better school performances are usually associated with lower weekend alcohol consumption. Secondary school students are taken are less likely to consume alcohol during weekends while students who are in good health status and go out more often are linked with higher weekend alcohol consumption levels.

One of the biggest limitations of my analysis is that I have only looked at weenkend alcohol consumption levels while there is another categorical variables in the dataset Dalc, weekday alcohol consumption that I did not include into the analysis. As a result, I failed to analyze a secondary school student's weekly alcohol consumption level. Considering ways that the analysis could be improved is that I can create a new response variable that includes both weekday alcohol consumption levels and weekend alcohol consumption levels and name it Weekalc, setting my own categories of the variable. This could also be an idea for future work, examing weekly alcohol consumption levels. In addition, there are many more socioeconomic information in the dataset that was not used in this analysis. More work including new

predictors could be done to predict the weekend alcohol consumption level of a secondary school student.