Final Report

Almas K.

2020-03-16

Introduction:

Alcohol use has been linked with cognitive impairement in the short term in a variety of situations such as in operation of a motor vehicle. Numerous factors have been found to affect a student's performance in a class, from sleep to diet.One previous study has shown the negative affect of alcohol on academic achievement in a student dataset from the United States 1. Thus, it would be interesting to see if this affect on performance can be replicated in other datasets and whether time of alcohol consumption (weekend or weekday) makes a difference.

Data Description

The datasets are obtained from UCI and is originally from Fabio Pagnotta and Hossain Mohammad Amran. It contains survey data from Portugese highschool students in a Math and Portugese class and contains information on 33 attributes. Each class is its own .csv file, but I will be focussing on the attributes from the Portugese class dataset as it contains more students (649 students). Each student makes up each row. This was generated from a colon separated file I made from the original txt metadata file.

Below is the entire variable set:

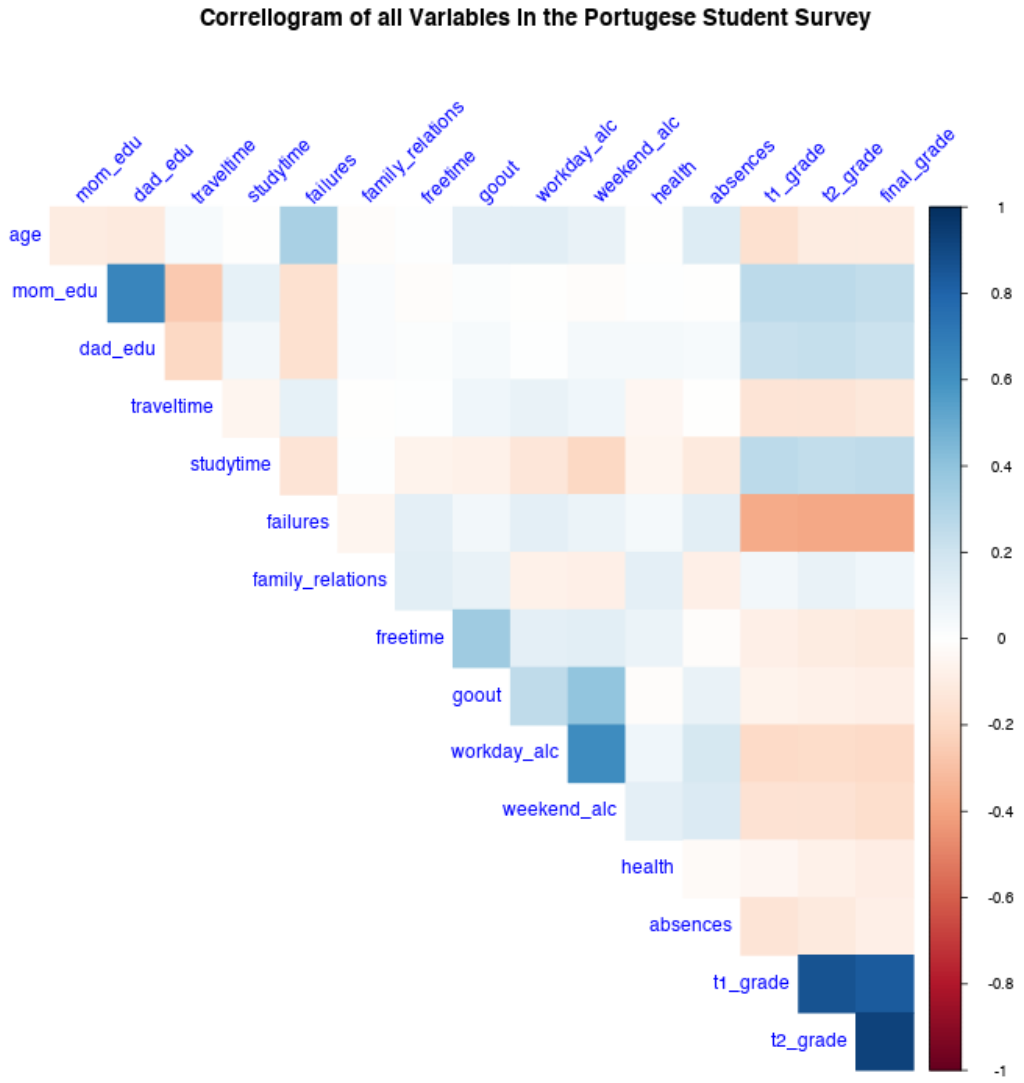
```
meta_dat <- read.delim((here("data","student_metadata.txt")),sep = ";", header=FALSE)
colnames(meta_dat) <- c("variable","description","type")
knitr:::kable(meta_dat, format="markdown")
```

variable	description	type
school	student's school	binary: GP for Gabriel Pereira or MS for Mousinho da Silveira
sex	student's sex	binary: F for female or M for male
age	student's age	numeric: from 15 to 22
address	student's home address type	binary: U for urban or R for rural
famsize	family size	binary: LE3 for less or equal to 3 or GT3 for greater than 3
Pstatus	parent's cohabitation status	binary: T for living together or A for apart
Medu	mother's education	numeric: 0 for none, 1 for primary education (4th grade), 2 for 5th to 9th grade, 3 for secondary education or 4 for higher education
Fedu	father's education	numeric: 0 for none, 1 for primary education (4th grade), 2 for 5th to 9th grade, 3 for secondary education or 4 for higher education
Mjob	mother's job	nominal: teacher, health care related, civil services (e.g. administrative or police), at_home or other
Fjob	father's job	nominal: teacher, health care related, civil services (e.g. administrative or police), at_home or other
reason	reason to choose this school	nominal: close to home, school reputation, course preference or other
guardian	student's guardian	nominal: mother, father or other

variable	description	type
traveltime	home to school travel time	numeric: 1 for <15 min., 2 for 15 to 30 min., 3 for 30 min. to 1 hour, or 4 for >1 hour
studytime	weekly study time	numeric: 1 for <2 hours, 2 for 2 to 5 hours, 3 for 5 to 10 hours, or 4 for >10 hours
failures	number of past class failures	numeric: n if 1<=n<3, else 4
schoolsup	extra educational support	binary: yes or no
famsup	family educational support	binary: yes or no
paid	extra paid classes within the course subject (Math or Portuguese)	binary: yes or no
activities	extra-curricular activities	binary: yes or no
nursery	attended nursery school	binary: yes or no
higher	wants to take higher education	binary: yes or no
internet	Internet access at home	binary: yes or no
romantic	with a romantic relationship	binary: yes or no
famrel	quality of family relationships	numeric: from 1 for very bad to 5 for excellent
freetime	free time after school	numeric: from 1 for very low to 5 for very high
goout	going out with friends	numeric: from 1 for very low to 5 for very high
Dalc	workday alcohol consumption	numeric: from 1 for very low to 5 for very high
Walc	weekend alcohol consumption	numeric: from 1 for very low to 5 for very high
health	current health status	numeric: from 1 for very bad to 5 for very good
absences	number of school absences	numeric: from 0 to 93
G1	first period grade	numeric: from 0 to 20
G2	second period grade	numeric: from 0 to 20
G3	final grade	numeric: from 0 to 20, output target

Data Exploration:

Correllogram

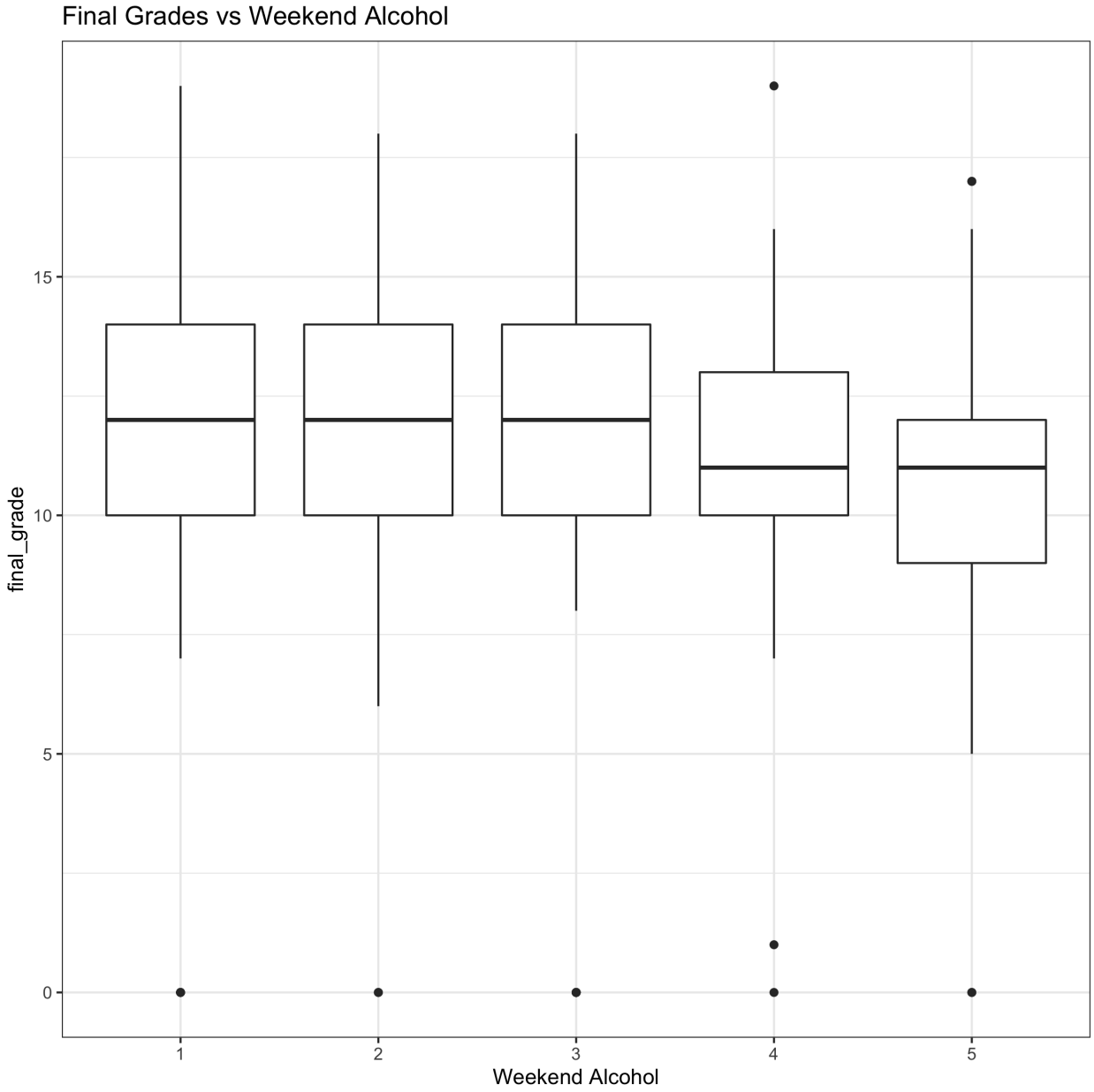


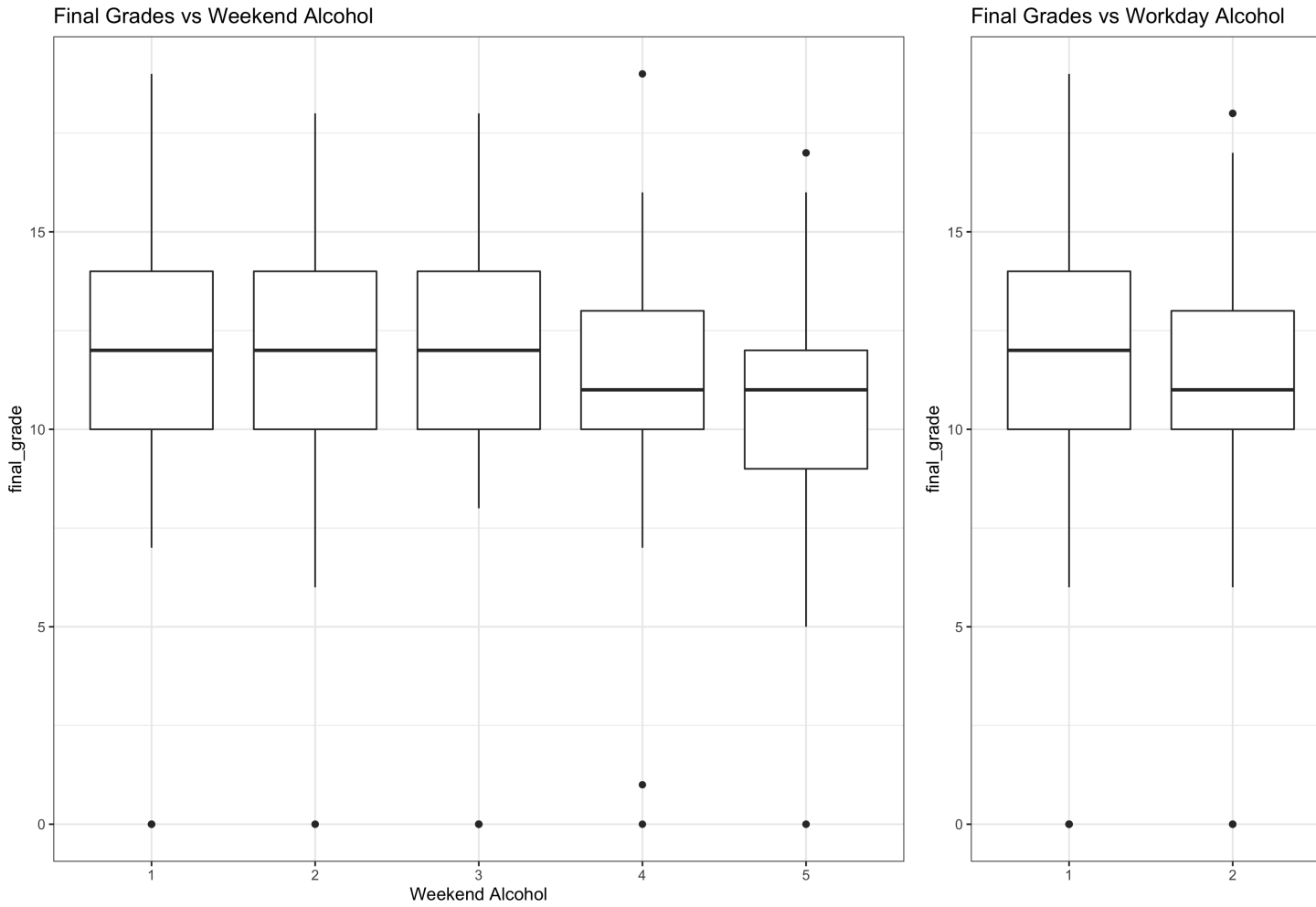
Correllogram of all Variables in the Portugese Student Survey

In this correllogram, we see a variety of factors having an association with final grades . The colour scheme shows all positive correlations as blue, and all negative correlations as red.Term 1 grades(t1_grades) and term 2 grades(t2_grades) having the highest correlation with final_grades makes sense here, as earlier term grades are correlated with later term grades. We will mainly focus on the alcohol (workday and weekend), which show negative correlation.

Boxplots

Let's look at weekend alcohol and workday alcohol use's spread.

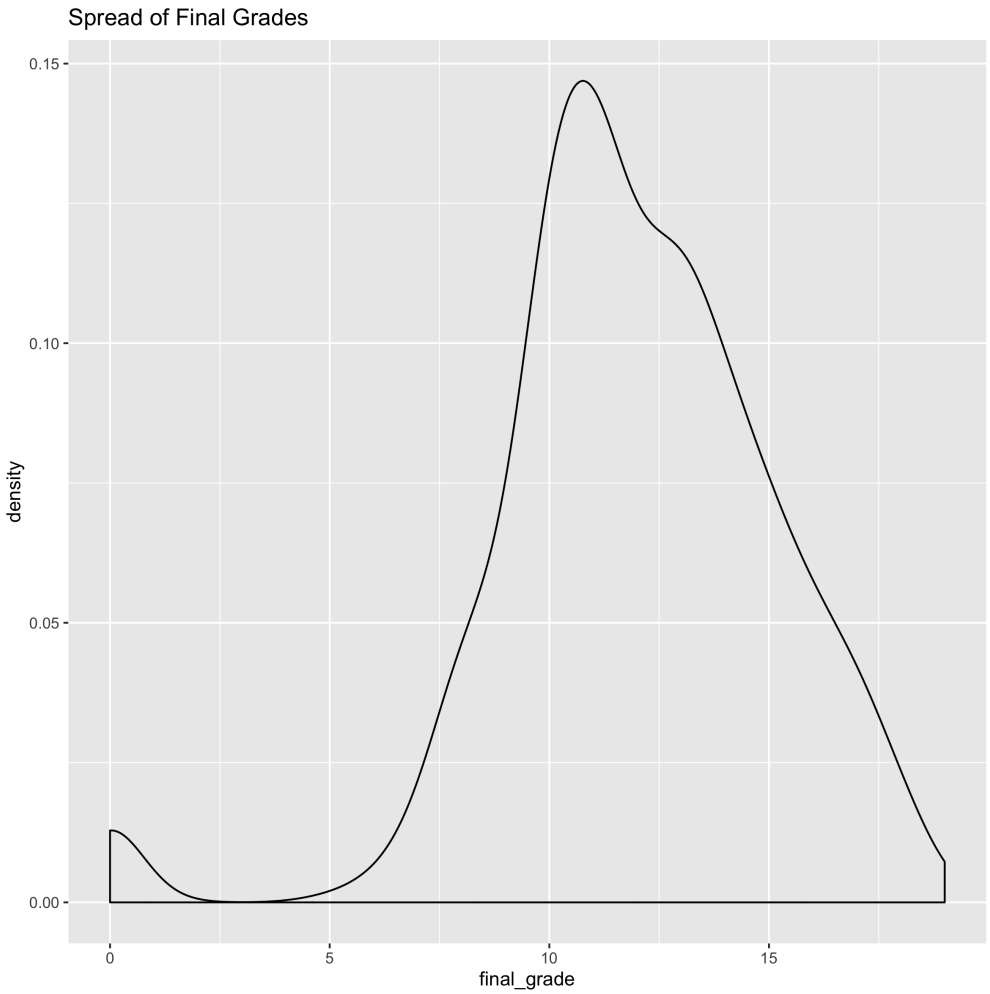




We see differences in the spread from the very low(1) to very high (5) consumption, with a general decrease in the mean as the amount of alcohol consumption increases increases, especially in the workday consumption.

Density Plots

Let's look at the distribution of grades.

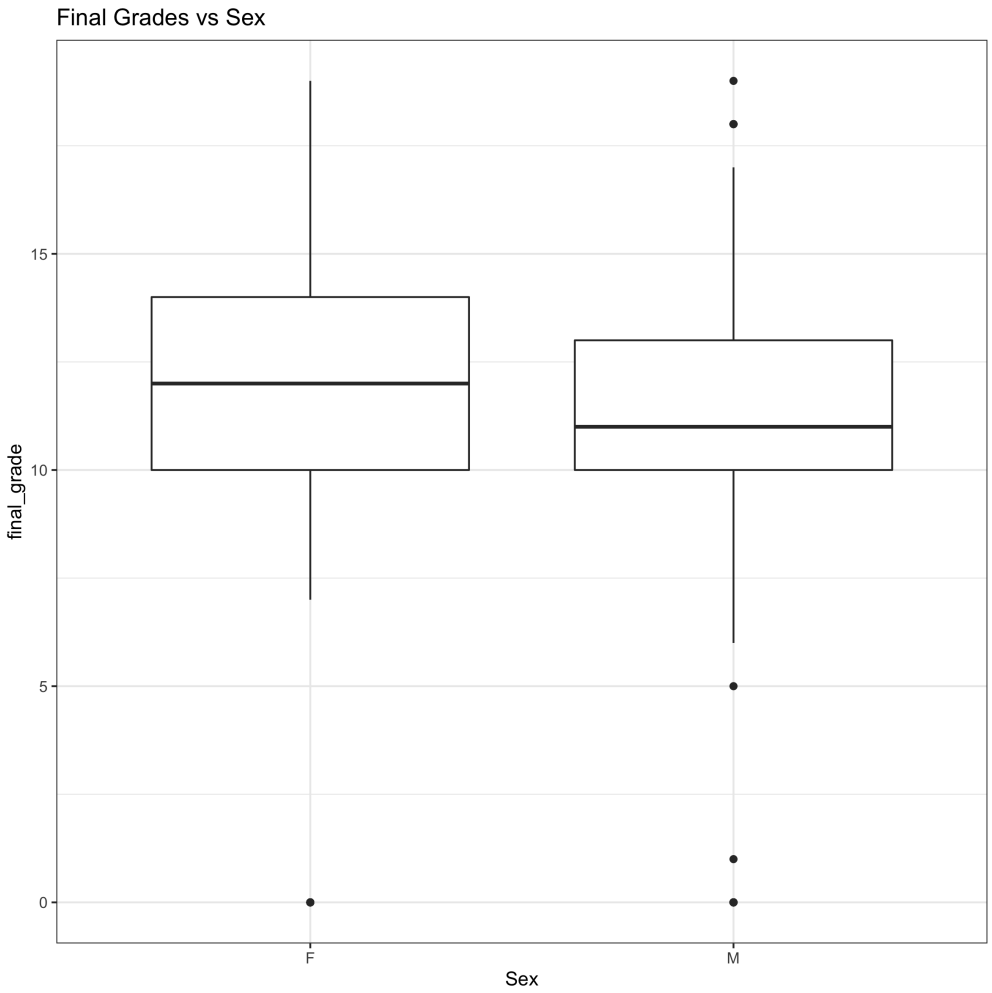


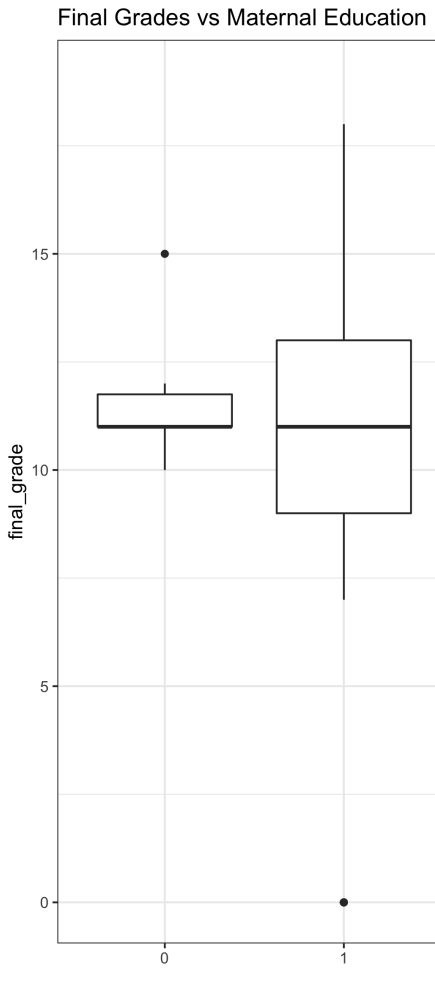
The distribution of grades appear to be a bit left skewed.

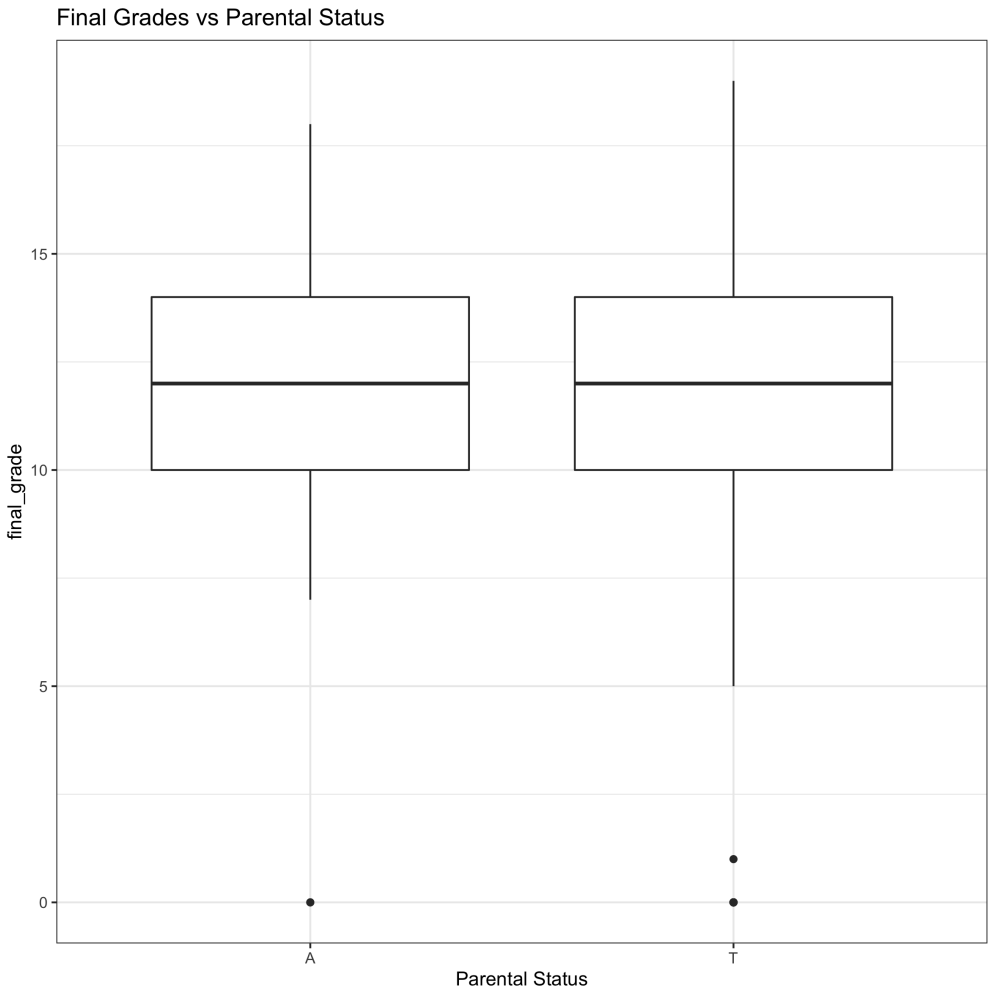
Other variables that may affect data

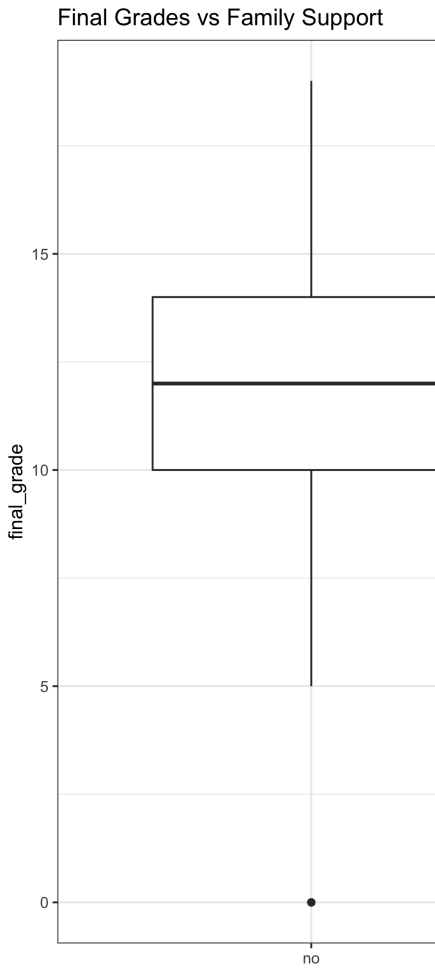
Let's look at potential confounding factors like sex of the student, parental status and family support and their spread in average final grades

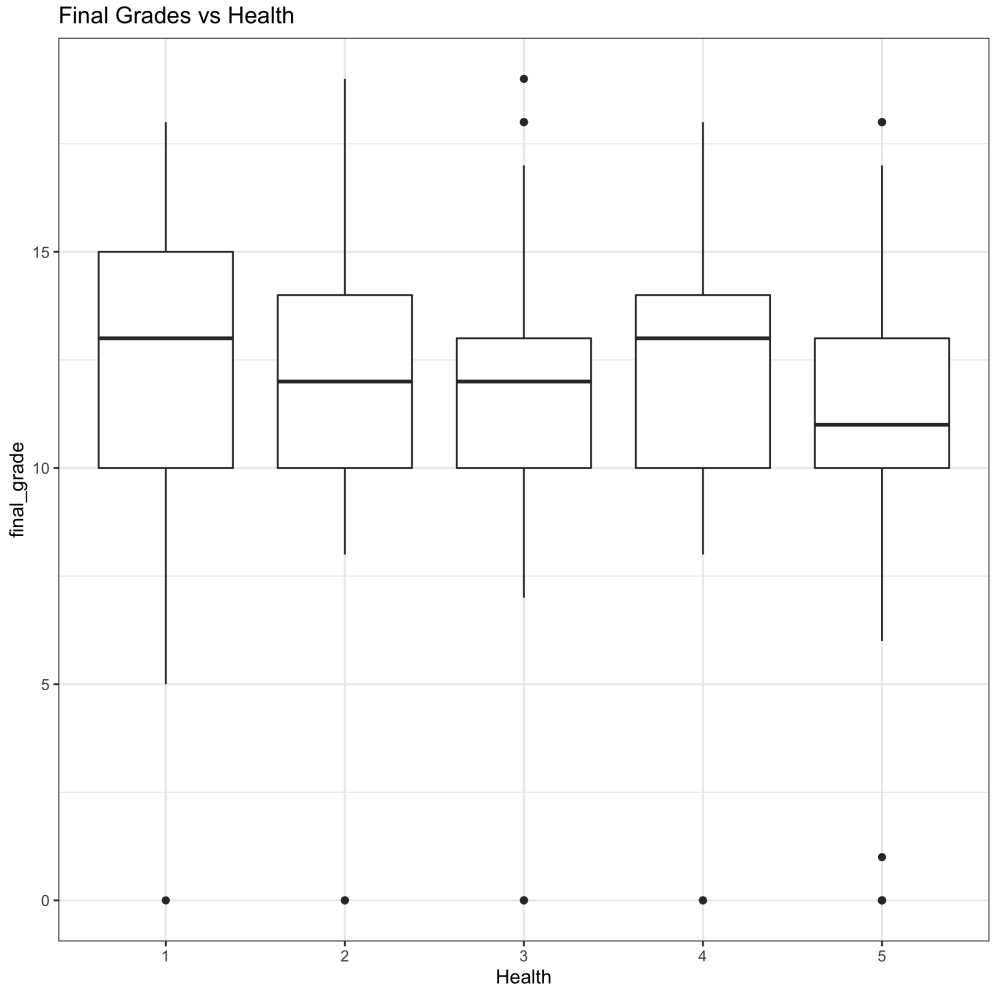
#####"Potential Confounding factors and grades"


Final Grades vs Sex
final_grade
15
10
5
0
F
M
Sex


Final Grades vs Maternal Education
final_grade
15
10
5
0
0
1


Final Grades vs Parental Status
final_grade
15
10
5
0
A
T
Parental Status


Final Grades vs Family Support
final_grade
15
10
5
0
no



It doesn't look like there is a huge difference between the grades in males compared to females. Males have a slighly lower average, but overall are similar. This is good because it will not be a huge confound in the data.Also family support and parental status have similar average values.

Research Question:

In this analysis, I will use linear regression to determine the relationship between alcohol use, either weekend, weekday (workday) or both and final grades (G3) for students. I chose the final grades as a output variable because it is more resistant to short term effects because it depends on work throughout the term.

Plan of Action:

I will remove those with very bad health status (1), as to reduce confounds in the data. My main focus is on the alcohol use categories and final grades, so I will probably ignore the other factors.I will then perform linear regression analysis and plot a regression line using the relevant variables.

Methods

I performed a multivariate simple linear regression using the lm package, after removing the very bad health status(1). I used workday alcohol and weekend alcohol as covariates and looked at interaction between these 2 as well.

```
lm_model <- readRDS(here("data","lm_model_alc.RDS"))
```

Results

Let's look at our linear model results.

```
tidy(lm_model)
```

```
## # A tibble: 4 x 5
##   term                                  estimate std.error statistic  p.value
##   <chr>                                    <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                             13.7       0.619     22.2  1.21e-78
## 2 dataset$workday_alc                     -1.16      0.478     -2.44 1.52e- 2
## 3 dataset$weekend_alc                     -0.370     0.208     -1.78 7.61e- 2
## 4 dataset$workday_alc:dataset$weekend_alc  0.161     0.117      1.38 1.69e- 1
```

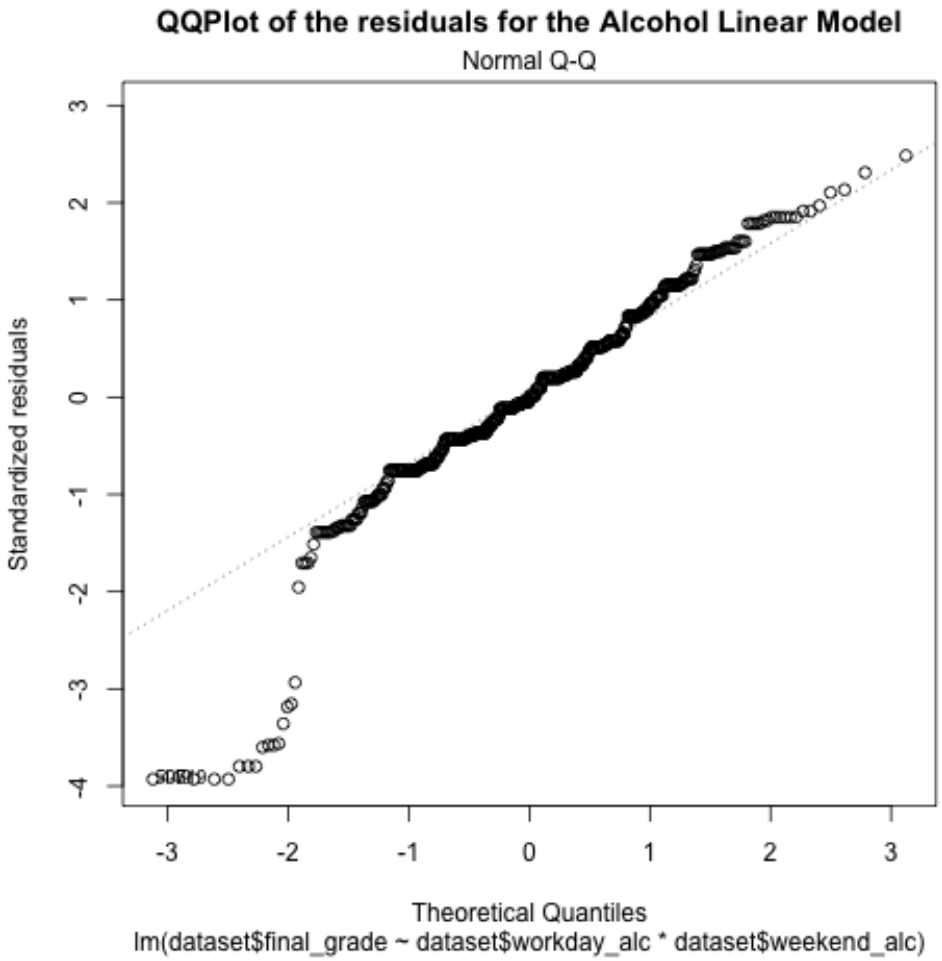
```
glance(lm_model)
```

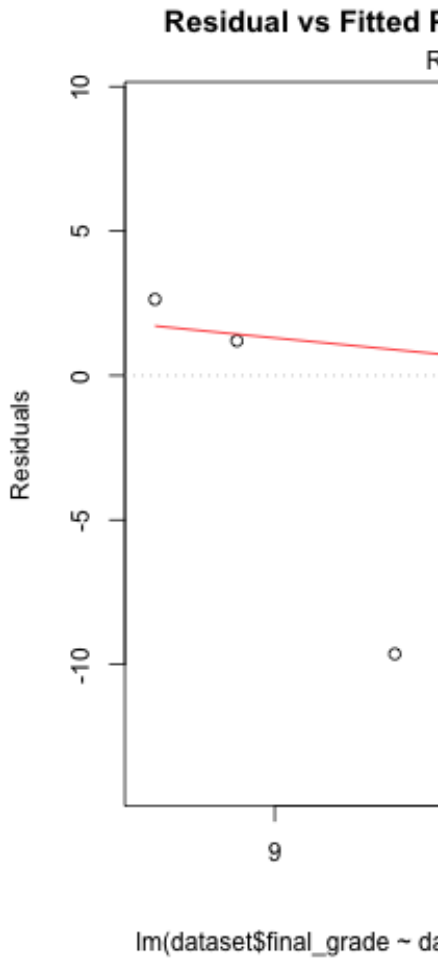
```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
## 1    0.0437        0.0385  3.16      8.45 1.70e-5     4 -1434. 2877. 2899.
## # ... with 2 more variables: deviance <dbl>, df.residual <int>
```

The only result that seems to be significant is the workday alcohol with grades. The interaction term is not significant, thus we can point that workday alcohol affects grades as a main effect. There appprea

Residual Plots:

Let's look at qqplots of the residulat and residual vs fitted plots.





The residuals do not all fall onto the qqplot and thus are not fully normally distributed.

Discussion/Conclusions

The only predictor variable that was significant was workday alcohol which had a negative association with final grades. This is in line with Balsa et al.'s study, which saw a significant, but small negative association with alcohol and grades, specifically for males. In my case, I did not separate by gender, which could be a future analysis. Also, I think including other covariates like family support in the future would be a good idea. Finally, given the qqplot, it would be best to potentially change the model from a simple linear regression that treats the predictor of alcohol use as a numeric, into a more complex model that treats this predictor as a categorical and uses dummy variables.