An Analysis of the Impact of Student Alcohol Consumption

By:
Xiya Ma xiyma@ucdavis.edu
Yesuunee Erdenebat yerdenebat@ucdavis.edu

Contributions

Xiya Ma:

Worked on Q1. Made histograms, a correlation table and organized references . Analyzed and made comprehensive conclusions on the graphs. Also worked on methods part of the report.

Yesuunee Erdenebat：

Worked on Q2. Made box plots, model selection and correlation analysis. Analyzed and gave comprehensive analysis on the conclusion of the boxplots. Worked on the Intro and Background of the project.

Third Member：
(This person decided to choose the second grading option)

Worked on hypothesis tests, post-hoc tests, mean charts for Question 1 and 2. Also, gave a comprehensive analysis on the hypothesis tests and the charts. Also, contributed to methods and conclusion part of the report.

Introduction

We analyzed the dataset named "Student Alcohol Consumption", which contained a comprehensive examination of the possible factors that could lead to alcohol consumption amongst Math and Portuguese language class students in a secondary school. The main goal of the project was to identify the biggest factors that influence the given secondary school student's alcohol consumption and how each level (1-5) of alcohol consumption affects their school performance.

Since the secondary school dataset had two separate datasets involving the Math class and Language class, we decided to only focus on Language class. We were planning on using a linear regression model for our analysis. However, it was pointed out by Dr. Characiejus**(response to professor's suggestion)** that the dataset contained mostly binary and categorical data instead of quantitative data. Therefore, we used a model selection, correlation test, Chi-Square test, Kruskal-Wallis test, Dunn test, histograms, and boxplots to answer most of our questions on identifying the best fit model for the data, visualizing the data and testing the hypothesis.

Background of the Dataset

The "Student Alcohol Consumption" dataset is separated into two parts: Math class and Portuguese Language class. There are 33 same variables for each class with most of the variables either categorical or binary. Due to the space limit of the project report and time restraints our team have only decided to focus on Portuguese Language class dataset. We have also decided to analyze variables we think are most relevant to the proposed question. Therefore, amongst 33 variables, the variables we have taken account are: sex, family size, parent's cohabitation status, quality of family relationship, student's workday alcoholic consumption, student's weekend alcohol consumption, number of class failures, number of school absenses, and grades. In other words, we chose variables that can have more impact on students' alcohol consumption and variables we believe are affected by student's alcohol consumption.

Statistical Questions of Interest

The two key objectives of our project are to find and analyze the variables that lead students toward alcohol consumption and illustrate the effect of the alcohol usage on students' school performance. Therefore, we will ask the following questions.

1. What is the best fit model based on students' family background (variables including address type, family size, parent's cohabitation status, mother's education, father's education, mother's job, father's job, family size and family relationship)? How correlated are the independent variables to the dependent variables? Can you visualize the data and test your hypothesis?

2. What is the effect of a student's alcohol consumption on their grades, number of past class failures, and absences? How correlated are the independent variables to the dependent variables? Can you visualize the data and test your hypothesis ?

Method

For the most part of our dataset, the variables we were using were categorical variables so we had to change them with numerical values by using as.numeric() function in R. Then we did model selection using forward and backward AIC model selection methods to find our best fit model.

For question 1, first we used Kendall correlation test to see how strong the relationship is between each variable. Since our data is not normal (normality test is attached in the appendix) and is categorical, we used chi-square to test the independence between two variables. Chi square was used to test for the association between alcohol consumption and categorical independent variables including sex, cohabitation status, and family size. In Q1, Kruskal-Wallis test was also used to test family relationship, which is a non-categorical and numerical variable. For better visualizing differences, we used the library(ggplot2) package in R to make histograms in question 1 and include a table containing the means of our variables.

In question 2, we used a correlation test with the cor() function to figure out the Spearman (used for non-parametric data) correlation of these independent variables to the dependent variables. To better visualize question 2, we used the boxplot() function and, because our variables for Q2 were numerical and non categorical, we used the Kruskal-Wallis test with kruskal.test() function to test our hypothesis. We will also include a table with the means from our variables.

Result and Analysis

Statistical Question of Interest 1

Model Selection:

Since some of the variables we were considering were binary variables, we changed the data and used numerical values to represent them instead. Most importantly we used Forward and Backward AIC model selection methods to select for the variables that best fit the model. Because there were so many variables in the original dataset, we chose to model select from variables that represent students' background, including address, famsize, parents's status, Mother's Education, Father's Education, Family support, Family relationship, etc. From the model selection we concluded that family size, family relationship and parents' status have higher effect on Workday and Weekend alcohol consumption more than the other variables. The best fit model is:

$$\mathbf{Y_{(alcohol\ consumption\ workday/weekend)} = B0 + B1_{(family\ size)} + B2_{(family\ relationship)} + B3_{(parents'\ status)}}$$

Correlation:

Furthermore, we created a correlation chart to show how strongly correlated each of the variables are to one another. For the correlation chart, in addition to the 3 best fit variables we included the variable sex to see how student's gender has effect on their alcohol consumption. As we can see in the chart below, student sex, family size, and parents status have positive correlations; family relationship has negative correlation. However, none of the correlation values exceed 0.5, so we cannot say there is a strong correlation between any of these variables and workday and weekend alcohol consumption. The only notable correlation is students' sex, which is 0.27 for workday alcohol consumption and 0.269 for weekend alcohol consumption.

	sex	famsize	Pstatus	famrel	Dalc	Walc
sex	1.00000000	0.09820486	0.06469983	0.07449045	0.27027515	0.26931435
famsize	0.09820486	1.00000000	-0.23960842	0.01548151	0.07324998	0.07283022
Pstatus	0.06469983	-0.23960842	1.00000000	0.02755438	0.05651568	0.06984852
famrel	0.07449045	0.01548151	0.02755438	1.00000000	-0.08815870	-0.08723817
Dalc	0.27027515	0.07324998	0.05651568	-0.08815870	1.00000000	0.55540700
Walc	0.26931435	0.07283022	0.06984852	-0.08723817	0.55540700	1.00000000

Figure 1: Correlation between variables

Visualization using histogram & Hypothesis testing
Further visual analysis by using histograms was done to show each of our variables against workday and weekend alcohol consumption.

Cohabitation Status:

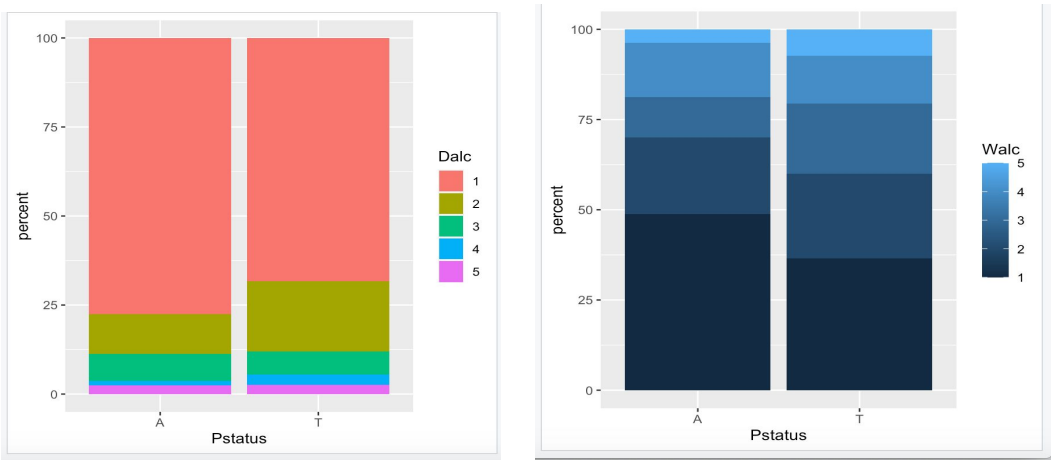


Figure 2: Percentage of Alcohol Consumption for Workday and Weekend Based on Parents' Cohabitation Status
A→ parents live apart T→ parents live together

Table: Mean Alcohol Consumption Based on Parents' Cohabitation Status

Parent's Cohabitation Status	**Away**	**Together**
Mean Workday Alcohol Consumption	1.4	1.52
Mean Weekend Alcohol Consumption	2.04	2.31

H_0: Student alcohol consumption is independent of the parent's cohabitation status

H_a:Student alcohol consumption is dependent of the parent's cohabitation status

The chi-square test yields the p-values of 0.3693 and 0.1374 for workday and weekend alcohol consumption, respectively. Since our p-values are greater than 0.05, we cannot reject the null hypothesis.

The graph shows that workday and weekend alcohol consumption of students whose parents are apart seem to have lower level alcohol drinking . However if we do a hypothesis test, it seems that there is no significant difference in workday and weekend alcohol consumption between students whose parents are living together or living apart. Therefore, we can say that parent's cohabitation status does not have a significant impact on students' consumption of alcohol.

Sex:

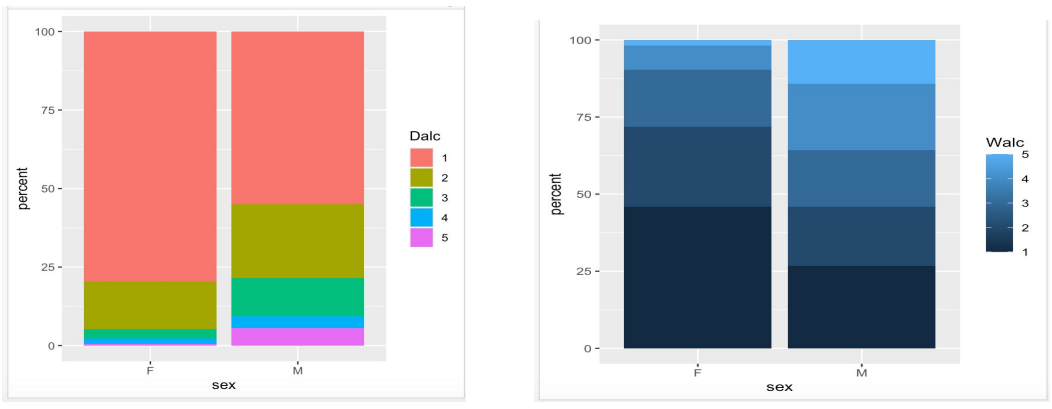


Figure 3: Percentage of Alcohol Consumption for Workday and Weekend Based on Sex

Table: Mean Alcohol Consumption Based on Students' Gender

Student Sex	**Female**	**Male**
Mean Workday Alcohol Consumption	1.28	1.82
Mean Weekend Alcohol Consumption	1.94	2.77

H_0: Student alcohol consumption is independent of the student sex

H_a:Student alcohol consumption is dependent of student sex

The p-value for workday and weekend are 8.513e-12 and 1.886e-15, therefore we have enough evidence to reject our null hypotheses.

From those histograms above, we see that more male students than female students consume at medium to high levels(3-5)on working days and weekends. On the other hand, female students' alcohol consumption level is generally very low, however on the weekends they consume at medium-high level compared to workday. Since we reject the null hypothesis, we conclude that there is an association between sex and student alcohol consumptions.

Family Relationship

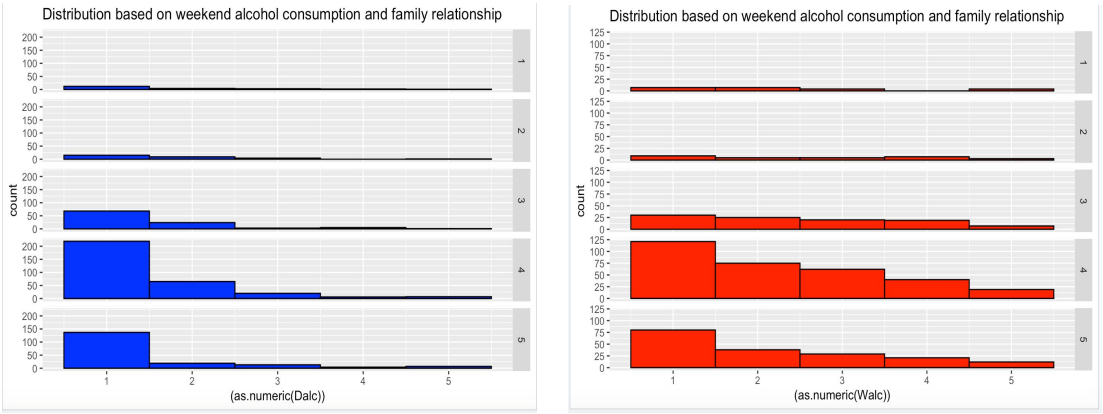


Figure 4: Scale of Workday and Weekend alcohol consumption based of Family Relationship

Table: Mean Alcohol Consumption Based on Family Relationship Levels

Family Relationship Levels	1	2	3	4	5
Mean Workday Alcohol Consumption	1.91	1.72	1.49	1.48	1.47
Mean Weekend Alcohol Consumption	2.41	2.66	2.49	2.25	2.15

H_0: There is no difference in the median alcohol consumption among different levels of family relationship

H_a: At least one of the median alcohol consumption is different

We obtained the p-value of 0.05234 for workdays, so we cannot reject the null hypothesis. However, we have the p-value of 0.09982 for weekends, so we reject the null hypothesis that there is no difference in the median weekend alcohol consumption amount for different levels of family relationship.
The graph shows that there are a large number of students who have excellent relationships with their parents drink less alcohol during both workday and weekend. However, students who have bad relationships with their parents are also not drinking a lot during both workday and weekend.

Our test yields the p-values of 0.05234 for workdays, so we do not have significant evidence to reject the null hypothesis that the median workday alcohol consumption is different between different family relationship levels. Likewise, we have the p-value of 0.09982 for weekends, so we cannot reject the null hypothesis or conclude that the median weekend alcohol consumption is different between family relationship levels. Therefore, family relationships do not affect weekend or workday alcohol consumption.

Family Size

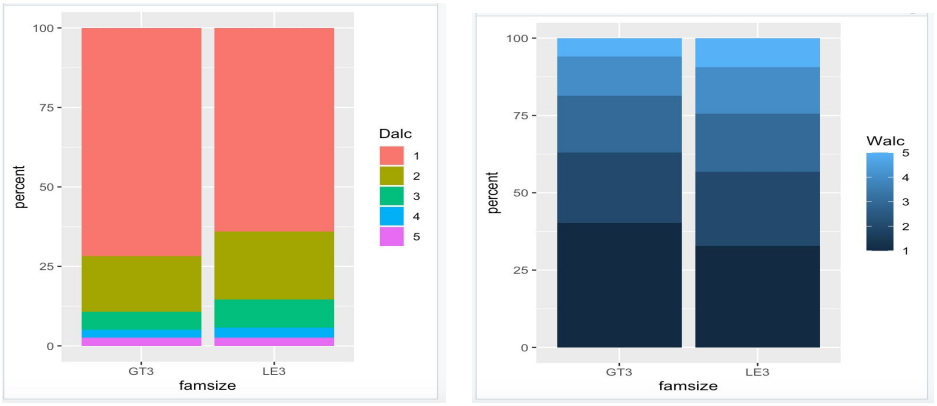


Figure 5: Percentage of Alcohol Consumption for Workday and Weekend Based on family size
GT3 → Greater than 3, LE3 → Less than 3

Table: Mean Alcohol Consumption Based on Family Size

Family Size	**Greater than 3**	**Less than 3**
Mean Workday Alcohol Consumption	1.47	1.59
Mean Weekend Alcohol Consumption	2.21	2.44

H_0: Student alcohol consumption is independent of the family size

H_a:Student alcohol consumption is dependent of the family size

The p-value for workday and weekend are 0.3377 and 0.2876.

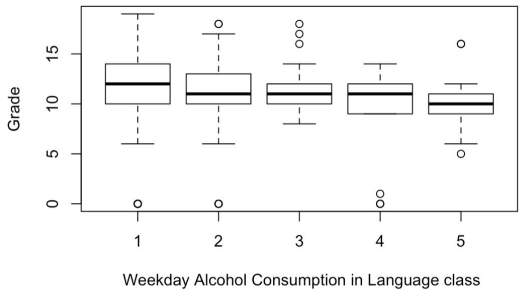
A large percentage of students who have more than three family members have lower drinking levels. However, the weekend alcohol consumption seems to vary in each level (level 1-5). According to the hypothesis testing, we do not have significant evidence to reject the null hypothesis that student alcohol consumption is independent of the family size. Therefore, we conclude that workday and weekend alcohol consumption is unaffected by family size.

Statistical Question of Interest 2
Determining the effect of student alcohol consumption on grades, number of past class failures, and absences

Our next objective is to determine if student alcohol consumption has an effect on school performance, especially grades. We have also taken a number of past class failures and absences into consideration as part of school performance. We utilized Spearman's correlation test to determine the correlation between the language class alcohol consumption rate with each of our three variables. We tried a linear model for visualizing correlation but since most of data is categorical or binary, we decided that a boxplot would better help us visualize the association between variables. Lastly, we will do a hypothesis test to determine if the medians are significantly different and a Dunn test to show which pairwise are different. Please note that all correlations are less than 0.5 so it is hard for us to say that there is a strong correlation.

<u>Workday</u> Alcohol Consumption

1. Effect of Workday alcohol consumption on Grades




```
        Spearman's rank correlation rho

data:  d2$G3 and d2$Dalc
S = 55054191, p-value = 8.432e-08
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
-0.208394
```

Figure 6: Median Language Class Grades Based on Workday Alcohol Consumption and their correlation

H_0: There is no difference in the median grades among the different levels of alcohol consumption

H_a: At least one of the median grades is different

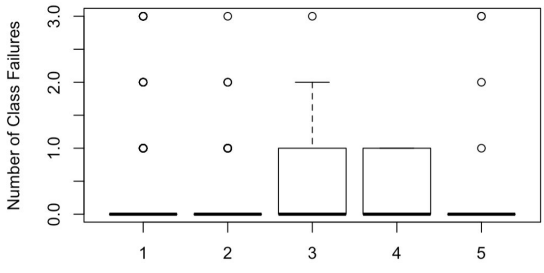
Table: Mean value of Grades based on each Workday Alcohol Consumption level

Workday Alcohol Consumption Levels in Portuguese	1	2	3	4	5
Mean Grades	12.3	11.4	11.1	8.94	10.2

Based on the figures above, it appears that median grades are highest when student alcohol consumption rate is low (with value equal to 1). Also, the correlation between alcohol consumption is negatively correlated, which shows that if the Workday alcohol consumption increases, the grades will decrease. We ran a Kruskal-Wallis test and a Dunn test for those with a p-value of less than 0.05.

Our p-value is 3.524e-05, thus we can reject the null hypothesis and conclude that at least one of the median grades is different. A Dunn test shows that the median grades are significantly different between workday alcohol consumption rate at 1 and 4, 2 and 4, 1 and 5, and 2 and 5. There are significant differences in median grades between low workday alcohol consumption (1 and 2) and high workday alcohol consumption (4 and 5). Therefore, looking at group difference(boxplot) and the hypothesis test we can conclude that students with lower workday alcohol consumption will have higher grades and students with higher workday alcohol consumption will have lower grades.

2.Effect of Workday Alcohol consumption on Class Failures




```
        Spearman's rank correlation rho

data:  d2$failures and d2$Dalc
S = 40600081, p-value = 0.0055
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1088617
```

Figure 7: Median Language Class Failure Based on Workday Alcohol Consumption and their Correlation

Table: Mean number of Class Failures based on each Workday Alcohol Consumption level

Workday Alcohol Consumption Levels in Portuguese	1	2	3	4	5
Mean Number of Class Failures	0.197	0.198	0.372	0.353	0.529

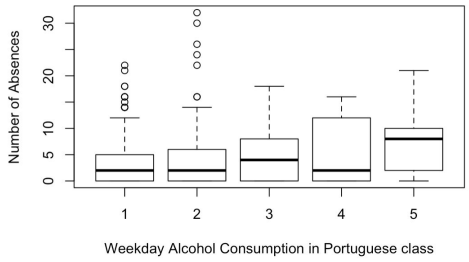
H_0: There is no difference in the median class failures among the different levels of alcohol consumption

H_a: At least one of the median class failures is different

Based on the above graph, median class failures are 0 for all levels of alcohol consumption. For alcohol consumption levels 3 and 4 the boxplots have less outliers than the others. The unusual distribution could mean that these students have very different numbers of failures and the data is heavily skewed. We will use the Kruskal-Wallis test to address the non-normality of our data. Our p-value is 0.02627, signifying that we have enough evidence to reject the null hypothesis and reject that there are no differences in the medians. A Dunn test shows that the median class failure is significant between workday alcohol consumption levels 1 and 4. There is a significant difference in median failures between low workday alcohol consumption (1) and high workday alcohol consumption (4).

Also, the correlation between class failure and alcohol consumption is positive. If alcohol consumption increases, the class failure will also increase. Therefore, from the group difference(boxplot) and the hypothesis test, we can conclude that the level of alcohol workday consumption does have an effect on the number of class failures.

3. Effect of Workday Alcohol consumption on Class Absences




```
        Spearman's rank correlation rho

data:  d2$absences and d2$Dalc
S = 40808826, p-value = 0.007844
alternative hypothesis: true rho is not equal to 0
sample estimates:
    rho
0.10428
```

Figure 8: Median Language Class Absences Based on Workday Alcohol Consumption and their Correlation

Table: Mean number of Absences based on each Workday Alcohol Consumption level

Workday Alcohol Consumption Levels in Language	1	2	3	4	5
Mean Number of Absences	3.20	4.26	5.07	4.76	7.06

H_0: There is no difference in the median class absences among the different levels of alcohol consumption

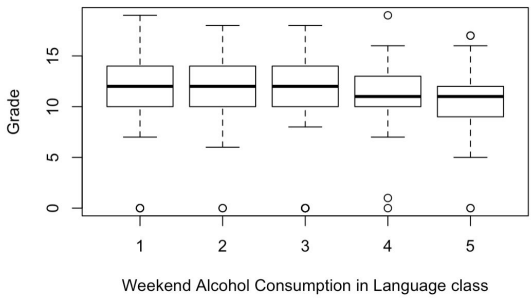
H_a: At least one of the median class absences is different

From the above graphs, median class absence tends to increases as the alcohol consumption increases. Median class absences also seem to increase with alcohol consumption, though median alcohol consumption on level 4 deviates from the assumption. The correlation between the median class absence and the workday alcohol consumption is positive but very weak.

The resulting p-value of 0.02007 allows us to reject the null hypothesis and conclude that at least one of the workday median numbers of absences is different. A Dunn test shows that the median is different for alcohol consumption rates between 1 and 5. There are significant differences in median absences between low workday alcohol consumption (1) and high workday alcohol consumption (5). Therefore, we can conclude that the alcohol consumption level has a varying effect on the number of absences in the class. Students who have the lowest workday alcohol consumption levels have the least absences while students with the highest workday alcohol consumption have the highest absences.

<u>Weekend</u> Alcohol Consumption

1. *Effect of Workday alcohol consumption on Grades*




```
        Spearman's rank correlation rho

data:  d2$G3 and d2$Walc
S = 53345753, p-value = 1.201e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1708952
```

Figure 9: Median Language Class Grade on Weekend Alcohol Consumption and their correlation.

Table: Mean values of Grade based each Weekend Alcohol Consumption level

Weekend Alcohol Consumption Levels in Language	1	2	3	4	5
Mean Grade	12.4	12.3	11.7	11.0	10.6

H_0: There is no difference in the median grades among the different levels of alcohol consumption

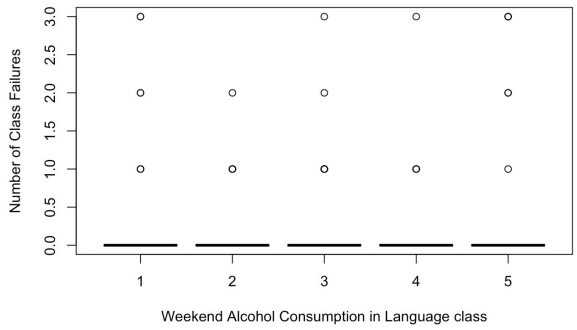
H_a: At least one of the median grades is different

Based on the figures above, it appears that median grades are highest when student alcohol consumption rate is low (with value equal to 1). Also, the correlation between alcohol consumption is negatively correlated, which shows that if the Weekend alcohol consumption increases, the grades will decrease. We ran a Kruskal-Wallis test and a Dunn test for Kruskal-Wallis tests with a p-value less than 0.05.

With the p-value 0.01982, we reject the null hypothesis and conclude that at least one of the median grades is significantly different. A Dunn test shows that median grades are different between weekend alcohol consumption levels 1 and 4, 1 and 5, 2 and 4, and 2 and 5. There are significant differences in median grades between low weekend alcohol consumption (1 and 2) and high weekend alcohol consumption (4 and 5).

Therefore, looking at group difference(boxplot) and the hypothesis test we can conclude that as Weekend alcohol consumption increases, grades will decrease.

2. *Effect of Weekend alcohol consumption on Class Failures*




```
        Spearman's rank correlation rho

data:  d2$failures and d2$Walc
S = 42609998, p-value = 0.09936
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.06474571
```

Figure 10: Median Language Class Failure Based on Weekend Alcohol Consumption and Correlation

Table: Mean number of Class Failures based each Weekend Alcohol Consumption level

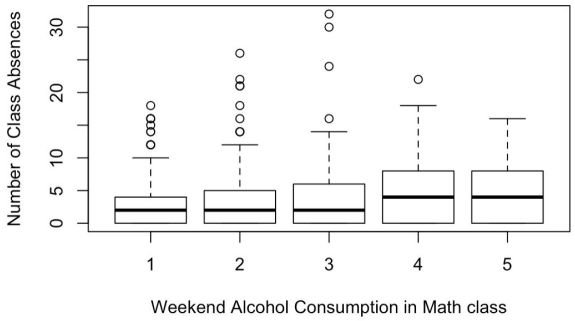
Weekend Alcohol Consumption Levels in Language	1	2	3	4	5
Mean Number of Class Failures	0.211	0.127	0.258	0.276	0.4

H_0: There is no difference in the median class failures among the different levels of alcohol consumption

H_a: At least one of the median class failures is different

Based on the above graph, median class failures are 0 for all levels of alcohol consumption. There are a lot of outliers in the box plot, which means our data might not be normal and highly skewed to one direction. The above figures show that the correlations are positive, but we will implement the nonparametric Kruskal-Wallis test to determine if there is a difference in the median class failures. Our p-value at 0.3426 does not provide significant evidence for us to reject the null hypothesis. Therefore, we conclude that the level of weekend alcohol consumption does not have an effect on the class failures.

3. *Effect of Weekend alcohol consumption on Class Absence*




```
        Spearman's rank correlation rho

data:  d2$absences and d2$Walc
S = 38948848, p-value = 0.000208
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1451049
```

Figure 11: Median Language Class Absences on Weekend Alcohol Consumption and their Correlation

Table: Mean number of Absences based each Weekend Alcohol Consumption level

Weekend Alcohol Consumption Levels in Language	1	2	3	4	5
Mean Number of Absences	2.89	3.69	3.9	4.64	5.24

H_0: There is no difference in the median class absences among the different levels of alcohol consumption

H_a: At least one of the median class absences is different

From the above figures, median class absence increases as the alcohol consumption increases. However, our box plot also shows many outliers, especially with the first three levels of weekend alcohol consumption. The correlation between the median class absences and the weekend alcohol consumption are positive.

We once again used the Kruskal-Wallis test and our resulting p-value is 0.1511. We cannot reject the null hypothesis or reject that there is no difference in our median class absences. Therefore, we can conclude that the alcohol consumption level has no effect on the number of absences.

Conclusion

Through forward and backward AIC model selection we found that the best fit model for our data has the variables parents' living status, family relationship, and family size. Furthermore, we made a correlation analysis on the chosen variables plus sex variable and concluded that while there is a correlation, there are no strong correlation between the independent variables and our dependent variable alcohol consumption. To understand the variables more, we built histograms and did hypotheses on each of the best fit model variables. The conclusion we derived from each of these analyses is that parents' cohabitation status and family relationship do not affect workday/weekend alcohol consumption, and family size only affects weekend alcohol consumption. It seems that the most defining difference in weekend/workday alcohol is the student's gender. According to our analysis, male students seem to drink more than female students during the weekend and workday.

Moreover, in the second part we also used box plots and the Kruskal-Wallis test to test whether or not alcohol consumption affects students' grades, class failures and absences. We also did an individual correlation test, which showed that there is a correlation between the independent and dependent variables, but similar to the correlation test in Q1, Q2 variables also didn't have a strong correlation. From our analysis we concluded that workday and weekend alcohol consumption has a negative effect on grades and only workday alcohol consumption has an effect on absences and failures in class since weekend alcohol consumption does not influence class absence or failures.

To conclude from both part 1 and 2, gender has the highest impact on the difference between a student's workday/weekend alcohol consumption and workday/weekend alcohol consumption negatively impacts class grade and only workday alcohol consumption has an impact on the number of absences and failures. Therefore, if students decrease their alcohol consumption during the workday, it can lead to higher grades and lower absences and failures.

References

1. https://en.wikipedia.org/wiki/Multinomial_logistic_regression
2. https://en.wikipedia.org/wiki/Chi-squared_test#Example_chi-squared_test_for_categorical_data
3. http://www.r-tutor.com/gpu-computing/correlation/kendall-tau-b
4. https://www.datanovia.com/en/lessons/kruskal-wallis-test-in-r/
5. https://rpubs.com/aaronsc32/post-hoc-analysis-tukey
6. https://rpubs.com/aaronsc32/spearman-rank-correlation
7. http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/