Analyzing The Association Between Social Factors and Mathematics Performance

Jane Condon, Valentina Tillmann
Stony Brook University, Department of Applied Mathematics (AMS)
Dec 5, 2024

Abstract:

This study investigates the association between social factors and mathematics performance among high school students. Using data collected from Portuguese high schools, we analyzed the impact of variables such as social outing frequency, romantic relationships, extracurricular involvement, and parental education on final math grades. Employing statistical techniques such as Wilcoxon Rank Sum tests, Kruskal-Wallis tests, and multiple linear regression, we explored relationships and validated assumptions about the data's distribution and linearity. Individual social variables showed limited explanatory power in predicting final grades. Rather, it was the combined effect of these variables that yielded significant results. Our findings emphasize the nuanced role of social and demographic factors in academic achievement.

1. Introduction

Research suggests that numerous social factors impact a student's success in mathematics courses. For example, socioeconomic status, gender, and even parental attitudes towards the subject can affect a person's performance on mathematics exams (Szczygieł 2020). In addition, possessing self-efficacy, or the belief in one's competencies, leads to positive outcomes in mathematics learning (Tarkar et. al., 2022). A positive self-image in young people is typically associated with a healthy, active social life. We present an analysis examining the impact of a student's social life on their academic performance in math courses. Our dataset is drawn from two Portuguese high schools and was collected using school reports and questionnaires. We obtained the data from Kaggle.com. The data includes test scores, student demographics, social level, and parent-related information. It contains 34 columns and 395

rows, with 21 categorical variables and 13 numerical variables. We extracted the following relevant columns:

Column	Class	Description
final_grade	integer	final grade (numeric: from 0 to 20, output target)
activities	factor	extra-curricular activities (binary: yes or no)
family_support	factor	family educational support (binary: yes or no)
mother_education	factor	mother's education (ordinal: "none", "primary education (4th grade)", "5th to 9th grade", "secondary education" or "higher education")
parent_status	factor	parent's cohabitation status (binary: "Living together" or "Apart")
social	integer	going out with friends (numeric: from 1 - very low to 5 - very high)
romantic_relationship	factor	with a romantic relationship (binary: yes or no)
health	integer	current health status (numeric: from 1 - very bad to 5 - very good)
father_education	factor	father's education (ordinal: "none", "primary education (4th grade)", "5th to 9th grade", "secondary education" or "higher education")
absences	integer	number of school absences (numeric: from 0 to 93)
weekend_alcohol	integer	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
address_type	factor	student's home address type (binary: "Urban" or "Rural")
internet_access	factor	Internet access at home (binary: yes or no)
extra_paid_classes	factor	extra paid classes within the course subject (Math) (binary: yes or no)

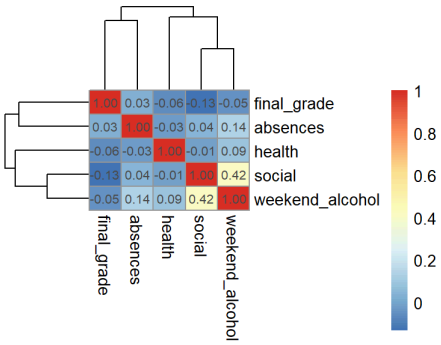
Our variables describe the level of stability the student experiences at home, as well as their involvement in activities outside of the classroom. We will conduct our analysis entirely using the statistical software R.

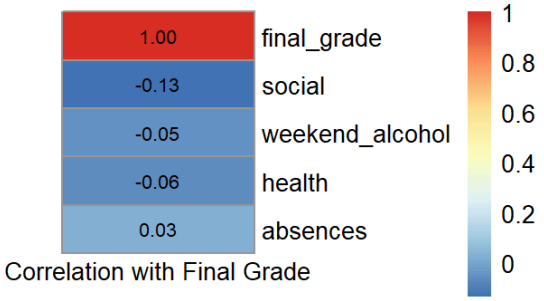
For our purposes, we converted the mother's education, father's education, travel time, and study time into ordered factor variables. Since all other variables are nominal or binary, we factor them with no defined order. In addition, we set “secondary education” as the baseline for linear regression, since a high school education is considered standard. Below is the relevant R code:

```{r}
#Turning categorical variables into factor variables (variable with multiple levels)
math_data <- math_data %>%
  mutate(across(c(address_type,parent_status, family_support,
extra_paid_classes, activities, internet_access, romantic_relationship,
mother_education,father_education), as.factor))

math_data$mother_education <- relevel(math_data$mother_education, ref =
"secondary education")
math_data$father_education <- relevel(math_data$father_education, ref =
"secondary education")
```

2. Exploratory Data Analysis

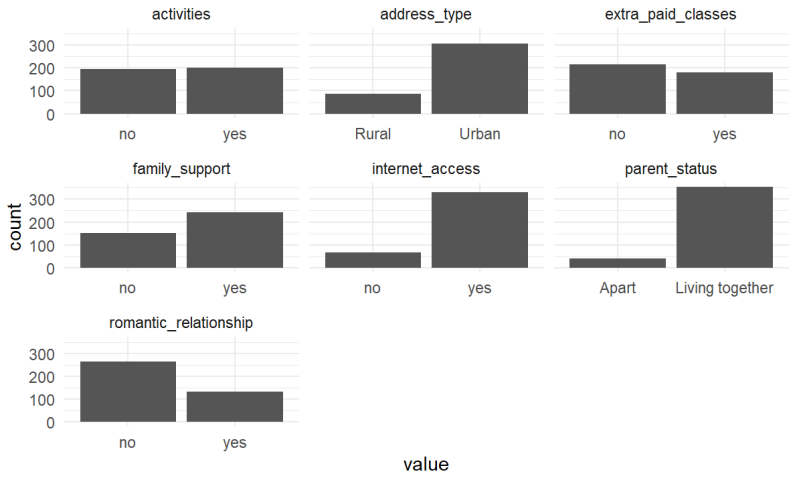
Shown below is the heatmap describing the correlation between numeric variables in our data. Weekend alcohol consumption and social score exhibit a moderate correlation of 0.42, suggesting a meaningful but not strong relationship between these two variables. Other variables do not show significant pairwise correlations, indicating that there are no notable linear relationships among the remaining numeric variables in the dataset. Considering these results, for this analysis we used multiple linear regression, as the combined effect of the variables can provide more meaningful insights.



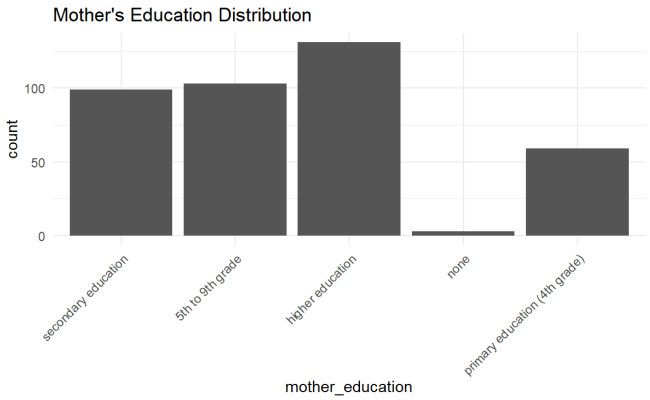


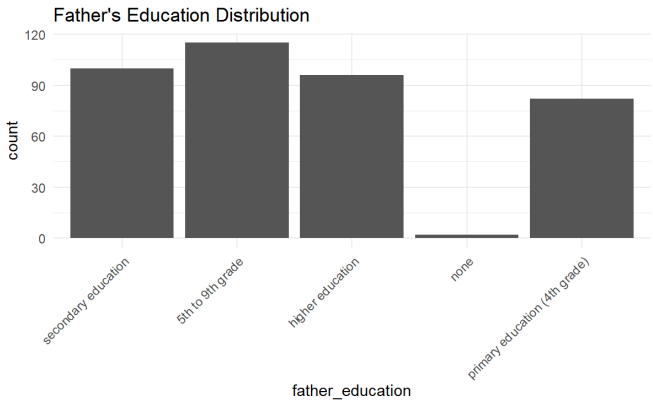
In addition, above is the heatmap illustrating the relationship between various numeric variables in the dataset and the final grade. Among the correlations, the most notable correlation, compared to other variables, is associated with the social score, which indicates a weak negative correlation.

We found that due to the severity of outliers in ‘absences’, we needed to apply a log transformation to visualize the distribution of our 5 numeric variables. As shown below, 'final_grade' and 'social' appear approximately symmetric, while 'weekend_alcohol', 'health', and 'absences' do not. ‘Final_grade’, ‘social’, ‘weekend_alcohol’, and ‘health’ have ranges similar in magnitude, but the ‘absences’ variable has a much larger spread. The nature of factor variables requires us to represent them on a different plot. Shown below are bar charts for our various two-level factor variables. Note that the counts for both categories are not significantly different for ‘activities’ and ‘extra_paid_classes’. However, ‘for address_type’ the data is skewed toward ‘Urban’ for address type, which could mean that the two schools are located in cities. Furthermore, ‘family_support’ is slightly skewed towards ‘yes’, ‘internet_access’ towards ‘yes’, ‘parent_status’ toward ‘Living Together’, and ‘romantic_relationship’ towards ‘no’.



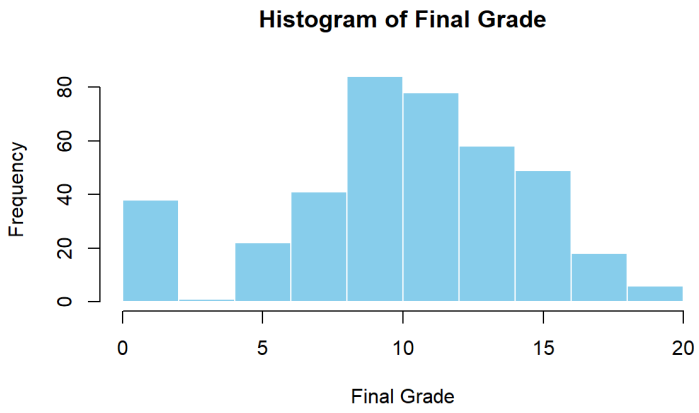
‘mother_education’ and ‘father_education’ are five-level factor variables. Their plots are shown below. There are more mothers with higher education than mothers with only high school, middle school, or primary education. There are slightly more fathers with a high school education than the other levels of education. For both mothers and fathers, there are very few that have no education whatsoever.





'fiinal_grade' is our dependent variable. Shown below is its distribution, which does not appear to be normally distributed. We confirmed this using a Shapiro-Wilk test for normality. Based on the histogram below, it appears to be skewed left.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00    8.00   11.00   10.42   14.00   20.00
```



3. Hypotheses of Interest

3.1 Hypothesis 1

For our first hypothesis, we want to study the relationship between a student's social life and their final grade in their mathematics course. To be more specific, we want to answer the following question: "Is having a more active social life associated with a better or worse mathematics grade?" To answer this question, we will conduct a series of statistical tests. We

will look at the following independent variables as a proxy for social life: romantic relationship status, involvement in extracurricular activities, weekend alcohol consumption level, and social outing frequency. Since our data is not normally distributed, we will use only nonparametric statistical tests in our analysis.

3.1.1 Wilcoxon Rank Sum Test to Test the Difference in Median Final Grade Between Students Involved in Romantic Relationships vs Those Who Are Not

First, we would like to know whether the median final grade for students involved in a romantic relationship is equal to that of their counterparts. In other words, do students involved in a romantic relationship tend to earn higher or lower grades in mathematics than students who are not involved in a romantic relationship? To answer this question, we must use a two-sample t-test. To determine which type of t-test is appropriate, we will check whether the assumptions of normality and homogeneity of variance hold true.

```
        Shapiro-Wilk normality test

data:  math_data$final_grade[math_data$romantic_relationship == "no"]
W = 0.9445, p-value = 2.009e-08


        Shapiro-Wilk normality test

data:  math_data$final_grade[math_data$romantic_relationship == "yes"]
W = 0.88655, p-value = 1.314e-08
```

To start with the normality assumption, we will conduct a Shapiro-Wilk test to test each group for normality. For both of the tests, we obtain a p-value less than 0.05, which indicates that we can reject the null hypothesis at a significance level of 0.05, and conclude that the data is not normally distributed for either group. Thus, we need to use a nonparametric test. Since our data is not normally distributed, we will then use Levene's test to test whether or not the equal variance assumption holds true, as Levene's test is more robust and less sensitive to departures from normality.

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  1.1824 0.2775
      393
```

Since we obtain a p-value of 0.2775, we fail to reject the null hypothesis at a significance level of 0.05, and can conclude that the homogeneity of variance assumption holds true.

Based on the results of the Shapiro-Wilk tests and Levene's test, we have decided that the Wilcoxon Rank Sum test is the appropriate test to use. Our null and alternative hypotheses are as follows:

H_0 = Mathematics grade is equally distributed across the two groups, or there is no significant difference between the medians of the two groups.

H_0 = Mathematics grade is not equally distributed across the two groups, or there is a significant difference between the medians of the two groups.

```
        Wilcoxon rank sum test with continuity correction

data:  final_grade by romantic_relationship
W = 19293, p-value = 0.06953
alternative hypothesis: true location shift is not equal to 0
```

From the Wilcoxon Rank Sum test, we obtain a p-value of 0.06953. At a significance level of 0.05, we fail to reject the null hypothesis and conclude that there is not a significant difference between the two groups. However, we can reject the null hypothesis at a significance level of 0.10. To analyze further, we also test to see if the median final grade for students involved in a romantic relationship is higher than that of their counterparts, using the 'greater' alternative.

```
        Wilcoxon rank sum test with continuity correction

data:  final_grade by romantic_relationship
W = 19293, p-value = 0.03476
alternative hypothesis: true location shift is greater than 0
```

Here, we obtain a p-value of 0.03476, indicating that we can reject the null hypothesis at a significance level of 0.05, and conclude that the median final grade for students involved in a

romantic relationship is higher than that of students who are not involved in a romantic relationship.

3.1.2 Wilcoxon Rank Sum Test to Test the Difference in Median Final Grade Between Students Involved in Extracurricular Activities vs Those Who Are Not

Next, we would like to know whether the median final grade for students involved in extracurricular activities is equal to that of their counterparts. Do students who are involved in extracurricular activities earn higher mathematics grades than those who are not involved in extracurricular activities? Once again, we must check our assumptions of normality and homogeneity of variance assumptions to determine which test is appropriate to use in this situation. To check whether or not the normality assumption holds true, we will perform a Shapiro-Wilk test for each group. The results are as follows:

```
        Shapiro-Wilk normality test

data:  math_data$final_grade[math_data$activities == "no"]
W = 0.93723, p-value = 1.909e-07


        Shapiro-Wilk normality test

data:  math_data$final_grade[math_data$activities == "yes"]
W = 0.91818, p-value = 4.008e-09
```

Since we obtain a p-value less than 0.05 for both tests, we can reject the null hypothesis for both tests at a significance level of 0.05 and conclude that neither of the groups follows a normal distribution. This indicates that we must use a nonparametric test. To check the equal variance assumption, we will use Levene's test, for the same reason as indicated above in the previous section. We obtain the following result:

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  0.1165  0.733
      393
```

Given a very large p-value of 0.733, we fail to reject the null hypothesis at a significance level of 0.05 and can conclude that the variances are equal across the two groups.

According to the results of the Shapiro-Wilk tests and Levene's test, we decided that a Wilcoxon Rank Sum test would be the best choice to test this hypothesis. The null and alternative hypotheses are as follows:

H_0 = Mathematics grade is equally distributed across the two groups, or there is no significant difference between the medians of the two groups.

H_0 = Mathematics grade is not equally distributed across the two groups, or there is a significant difference between the medians of the two groups.

We observe the following results:

```
        Wilcoxon rank sum test with continuity correction

data:  final_grade by activities
W = 18912, p-value = 0.6049
alternative hypothesis: true location shift is not equal to 0
```

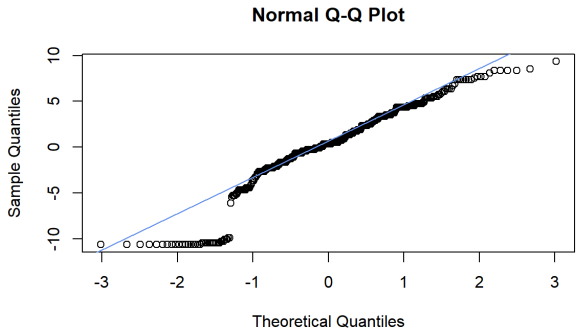
From this test, we obtained a p-value of 0.6049, which is far greater than our significance level of 0.05, so we failed to reject the null hypothesis at a significance level of 0.05. Thus, we can conclude that there is insufficient evidence to conclude that there is a significant difference in median final grade between students who are involved in extracurricular activities versus those who are not. In other words, whether or not a student is involved in extracurricular activities is unrelated to their mathematics grades.

3.1.3 Kruskal Wallis Test to Test the Difference in Median Final Grade for Different Levels of Weekend Alcohol Consumption

Our objective for this part of our analysis was to use an ANOVA test for the relationship between final math grade and weekend alcohol consumption. We considered this test to be more appropriate than a Pearson correlation test since weekend alcohol consumption can be considered an ordinal categorical variable; there are only 5 unique values, which can be thought of as 5 "levels" of alcohol consumption or 5 groups. To decide whether we should use the standard ANOVA test or a nonparametric alternative, we attempted to verify the following assumptions:

1.) Normality of Residuals (Shapiro Test)

```
        Shapiro-Wilk normality test

data:  aov_residuals
W = 0.93045, p-value = 1.341e-12
```



At α = .05, we reject the null hypothesis and conclude that the residuals do not follow a normal distribution. Thus, the normality assumption has been violated.

2.) Homogeneity of variance

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   4  3.1226 0.01509 *
      390
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

At α = .05, we can conclude there is sufficient evidence that the variances of all groups are not equal. Thus, both assumptions of the ANOVA test have been violated. As a result, we conducted the non-parametric Kruskal Wallis test. Unlike ANOVA, Kruskal Wallis does not assume a normal distribution of the underlying data.

H_0: *There is no difference in median math score among the five groups.*

H_a: *Not* H_0.

```
        Kruskal-Wallis rank sum test

data:  final_grade by factor(weekend_alcohol)
Kruskal-Wallis chi-squared = 5.453, df = 4, p-value = 0.2439
```

At α = .05, we conclude that the medians are not equal across the groups. Thus, we can conclude that there is not a significant relationship between weekend alcohol consumption and

mathematics final grade. Since our results are insignificant, it is unnecessary to perform a post-hoc analysis.

3.1.4 Kruskal Wallis Test to Test the Difference in Median Final Grade for Different Levels of Social Outing Frequency

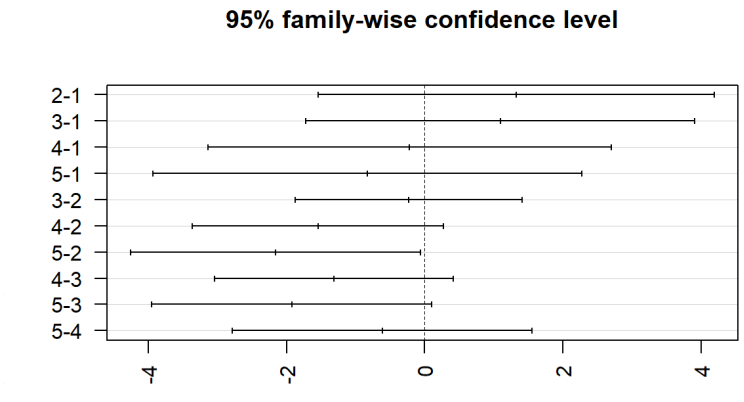
Our objective for this part of our analysis was to use an ANOVA test for the relationship between final grade and social outing frequency. We considered this test more appropriate since social outing frequency can be considered an ordinal categorical variable; there are only 5 unique values, which can be thought of as 5 "levels" of social outing frequency or 5 groups. We found that while the equal variance assumption held true, the residuals of the ANOVA model were not normally distributed. Thus, we conducted a non-parametric Kruskal Wallis test.

H_0: $There\ is\ no\ difference\ in\ median\ math\ score\ among\ the\ five\ groups.$

H_a: $Not\ H_0$.

```
        Kruskal-Wallis rank sum test

data:  final_grade by factor(social)
Kruskal-Wallis chi-squared = 14.697, df = 4, p-value =
0.005372
```

At a significant level of 0.05, we can reject the null hypothesis and conclude that there is sufficient evidence that the medians are not equal across the groups. Thus, we can conclude that there is a significant relationship between weekend alcohol consumption and final grade. Next, we must conduct a post-hoc analysis to examine the relationship further.




```
= math_data)
```

```
5-4 -0.6134269 -2.782500  1.55564572 0.9376881
```

As shown above, there is a statistically significant difference between group 5 and group 2 at α = .05. There is a statistically significant difference between group 5 and group 3 at α = 0.10. According to the plot above, we can be 95% confident that the true difference in the median between group 5 and group 2 is between -4 and -.05. This means that the median final grade of social score = 5 is higher than the median final grade of social score = 2. Essentially, students who go out with their friends very often receive lower mathematics grades than those who go out with their friends “not very often” or “somewhat often.”

3.2 Hypothesis 2

For our second hypothesis, we would like to know whether or not a student’s socioeconomic background affects their mathematics grades. To put it simply, we would like to answer the following question: “What kind of socioeconomic and demographic factors have the strongest effect on mathematics grade?” To answer this question, we will use a technique called multiple linear regression. We will include the following independent variables: address type, family support, health, internet access, mother’s education, father’s education, extra paid classes, and parent status. We obtain the following regression equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \epsilon$$

Since our data violates some of the linear regression assumptions, we will experiment with a few different models to create one that is unbiased and appropriate for our data.

3.2.1 Linear Regression Model to Determine if There is a Relationship Between Final Grade and Socioeconomic Factors

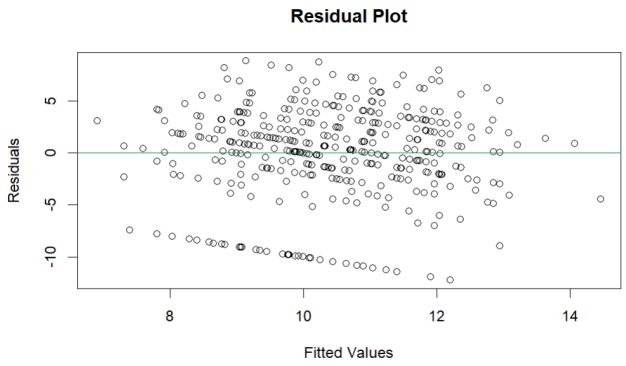
For our first multivariate regression model, we will use the standard model, whose equation is shown in the previous section. We obtain the following results:

```
Residuals:
     Min       1Q   Median       3Q      Max
-12.1937  -2.0680   0.6614   2.9567   8.8625

Coefficients:
                                                Estimate Std. Error
(Intercept)                                      10.5713     1.2140
address_typeUrban                                 0.6058     0.5612
family_supportyes                                -1.0086     0.4919
health                                           -0.1192     0.1655
internet_accessyes                                0.5154     0.6391
mother_education5th to 9th grade                 -0.5711     0.6620
mother_educationhigher education                  1.1479     0.6509
mother_educationnone                              2.6136     2.6492
mother_educationprimary education (4th grade)    -1.3121     0.8329
father_education5th to 9th grade                  0.1311     0.6488
father_educationhigher education                  0.3824     0.6784
father_educationnone                              2.3130     3.2158
father_educationprimary education (4th grade)    -0.2855     0.7724
extra_paid_classesyes                             0.7972     0.4830
parent_statusLiving together                     -0.5794     0.7580
```

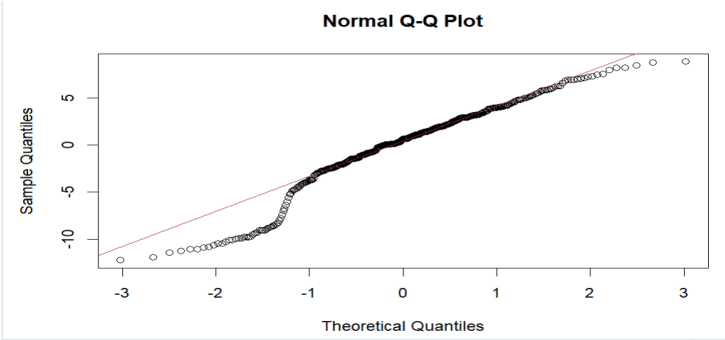
```
                                                t value Pr(>|t|)
(Intercept)                                       8.708   <2e-16 ***
address_typeUrban                                 1.079   0.2811
family_supportyes                                -2.050   0.0410 *
health                                           -0.720   0.4720
internet_accessyes                                0.807   0.4205
mother_education5th to 9th grade                 -0.863   0.3889
mother_educationhigher education                  1.764   0.0786 .
mother_educationnone                              0.987   0.3245
mother_educationprimary education (4th grade)    -1.575   0.1160
father_education5th to 9th grade                  0.202   0.8400
father_educationhigher education                  0.564   0.5733
father_educationnone                              0.719   0.4724
father_educationprimary education (4th grade)    -0.370   0.7119
extra_paid_classesyes                             1.650   0.0997 .
parent_statusLiving together                     -0.764   0.4451
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.464 on 380 degrees of freedom
Multiple R-squared:  0.08453,   Adjusted R-squared:  0.0508
F-statistic: 2.506 on 14 and 380 DF,  p-value: 0.002004
```

In this model, the only independent variable that has a statistically significant relationship with final grade at the 0.05 significance level is family support. Interestingly, receiving support from family in the subject of mathematics is associated with a lower mathematics grade, at least in this school. We hypothesize that this may be due to heightened academic pressure faced by students who receive family support. A student who receives family support in the subject of mathematics may be more likely to take more advanced mathematics courses, making it more difficult to achieve a high grade. Or, they may simply be adversely affected by academic stress due to their parents' high expectations and achieve poor grades for this reason alone. At the 0.10 significance level, mother's education level (higher education) and extra paid classes have a statistically significant relationship with final grade. A student whose mother has obtained a Bachelor's degree or higher, as compared to the standard high school education, is predicted to earn higher grades in mathematics. Finally, a student who participates in extra-paid mathematics classes outside of school is also predicted to earn higher grades in mathematics. It is important to note that this model has an adjusted R-squared value of 0.0508, which suggests that the socioeconomic factors in our model do not explain much of the variation in final grade.

To gauge whether these results are valid and unbiased, we will test whether our model follows the three assumptions necessary for linear regression: linearity, normality of residuals, and homogeneity of variance. First, we construct a residual plot to test both the linearity and homogeneity of variance assumptions:



The plot indicates that the linearity assumption has been violated, as the residuals do not appear to be randomly scattered around the horizontal axis. The equal variance assumption has not been violated, as the residuals appear to have a constant spread across all fitted values. Next, we check the normality of residuals assumption by looking at the QQ-normal plot of residuals:

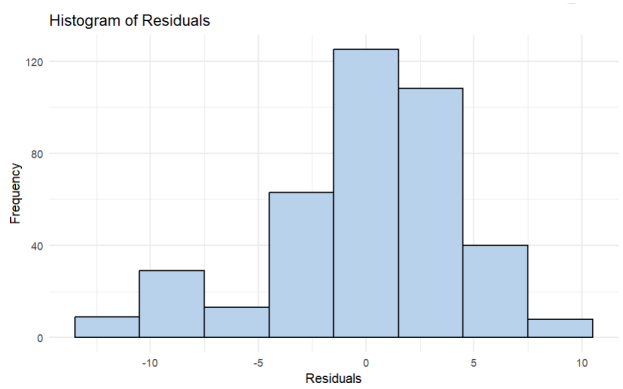


Based on the plot above, the residuals do not seem to follow a normal distribution. The points follow more of an “S” shape, rather than a straight line, indicating that our data has a heavy “tail,” i.e., it is skewed. We can also check the normality assumption using the Shapiro-Wilk test on the residuals:

```
        Shapiro-Wilk normality test

data:  residuals(model)
W = 0.94986, p-value = 2.555e-10
```

Given an extremely small p-value, we can reject the null hypothesis at a significance level of 0.05 and conclude that the residuals are not normally distributed. This is similar to the

conclusion that we made using the QQ plot of residuals. We can also look at a histogram of the residuals to check the normality assumption:



From the histogram above, we observe that the data is skewed, and subsequently does not follow a normal distribution. Based on these three methods to test for normality, it is clear that the normality assumption has been violated. Since two out of three of the assumptions necessary for linear regression have been violated, the results of our multiple linear regression model may be biased and are not valid.

3.2.2 Modified Linear Regression Models Accounting for Non-Normality of the Data

To account for the non-normality and non-linearity of our original data, we will apply two different transformations to the dependent variable, and use whichever model we believe to be more accurate. First, we will apply a Box-Cox transformation to the final grade variable. The Box-Cox transformation often helps to make the data more symmetric, bringing it closer to a normal distribution. Since the final grade column contains many ‘0’ values, we will first shift the data by 1, as performing the Box-Cox transformation requires all values of the response variable to be greater than zero.

```r
#Shifting the data to be positive so we can use Box-Cox transformation
min_value <- min(math_data$final_grade)
if (min_value <= 0) {
  math_data$final_grade_plus1 <- math_data$final_grade - min_value + 1
}
```

We apply the Box-Cox transformation as follows:

```r
#Box Cox Transformation
bc <- boxcox(final_grade_plus1 ~ address_type + family_support + health +
internet_access + mother_education + father_education + extra_paid_classes +
parent_status, data = math_data)

lambda1 <- bc$x[which.max(bc$y)]
math_data$transformed_y <- (math_data$final_grade_plus1^lambda1-1)/lambda1
```

Now, we will repeat the multiple linear regression model in the previous section, with our transformed dependent variable. We observe the results as follows:

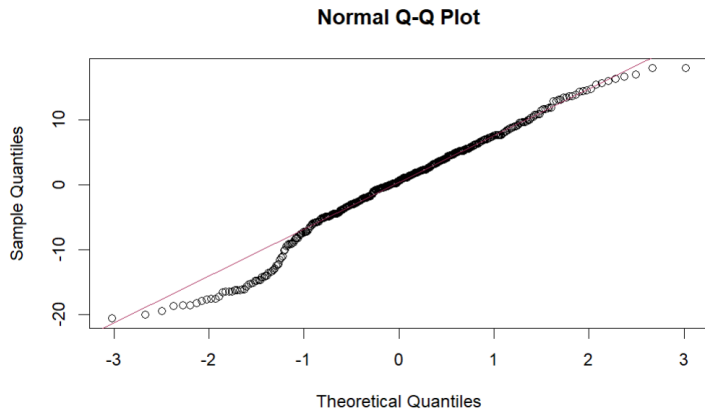
```
Residuals:
     Min       1Q   Median       3Q      Max
-20.5720  -4.4922   0.6578   5.2507  17.9500

Coefficients:
                                          Estimate Std. Error
(Intercept)                                17.4679     2.1990
address_typeUrban                           1.1768     1.0166
family_supportyes                          -1.8439     0.8910
health                                     -0.2661     0.2998
internet_accessyes                          1.0378     1.1576
mother_education5th to 9th grade           -1.0588     1.1991
mother_educationhigher education            2.1622     1.1790
mother_educationnone                        4.7572     4.7987
mother_educationprimary education (4th grade) -2.5235     1.5087
father_education5th to 9th grade            0.4216     1.1752
father_educationhigher education            0.9282     1.2289
father_educationnone                        4.3351     5.8250
father_educationprimary education (4th grade) -0.4116     1.3992
extra_paid_classesyes                       1.1397     0.8750
parent_statusLiving together               -0.9646     1.3731
                                          t value Pr(>|t|)
(Intercept)                                 7.943 2.25e-14 ***
address_typeUrban                           1.158   0.2477
family_supportyes                          -2.070   0.0392 *
health                                     -0.887   0.3754
internet_accessyes                          0.897   0.3705
mother_education5th to 9th grade           -0.883   0.3778
mother_educationhigher education            1.834   0.0674 .
mother_educationnone                        0.991   0.3221
mother_educationprimary education (4th grade) -1.673   0.0952 .
father_education5th to 9th grade            0.359   0.7200
father_educationhigher education            0.755   0.4505
father_educationnone                        0.744   0.4572
father_educationprimary education (4th grade) -0.294   0.7688
extra_paid_classesyes                       1.303   0.1935
parent_statusLiving together               -0.703   0.4828
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.085 on 380 degrees of freedom
Multiple R-squared:  0.08877,   Adjusted R-squared:  0.0552
F-statistic: 2.644 on 14 and 380 DF,  p-value: 0.001088
```

According to the results of this model, family support is the only variable that has a statistically significant relationship with final grade at a significance level of 0.05, similar to our original model. It is a negative relationship, once again. Mother’s education has a statistically significant relationship with final grade at the 0.10 significance level. To clarify, a student whose mother has achieved a Bachelor’s degree or higher (as opposed to the standard education of a high school diploma) is predicted to achieve higher mathematics grades, while students whose mother has only completed primary education are predicted to receive lower mathematics grades. We notice that the adjusted R-squared value is very low in this model as well, suggesting that the independent variables in this model do not explain much of the variation in final grade.

We check again to see if the normality assumption has been satisfied after applying the Box-Cox transformation:



While the residuals look “more normal” than the residuals of our original model, they still follow an “S” shape rather than a straight line. Thus, the normality assumption has been violated, making the results of our model invalid.

We consider one final “transformation,” which involves removing the ‘0’ values from our final grade variable. While this may not be a standard statistical procedure, we do so because our data contains an unusually large number of ‘0’ values for final grade, which we believe to be irrelevant to our analysis. When a student receives a zero for a final course grade, it often indicates that a student has either withdrawn from the course or has had their score voided (due to academic dishonesty, incomplete work, etc.). It is nearly impossible for a student to complete a course and receive a grade of zero. Since the next highest grade is a 4 out of 20, these students most likely did not truly earn a grade of zero. Since final grade is not a strong indicator of mathematics performance for these particular students, it is appropriate to remove them in this scenario. We remove all rows where final grade is equal to zero. We repeat the multiple linear regression model with our ‘transformed’ data, and obtain the following results:

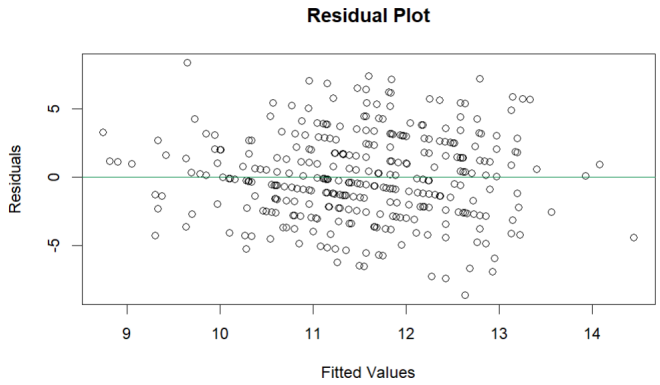
```
Residuals:
    Min      1Q  Median      3Q     Max
-8.6353 -2.2431 -0.1212  2.0249  8.3484

Coefficients:
                                                Estimate Std. Error t value
(Intercept)                                     11.17917    0.87560  12.767
address_typeUrban                                0.67487    0.41778   1.615
family_supportyes                               -0.61332    0.36423  -1.684
health                                          -0.18322    0.12127  -1.511
internet_accessyes                               0.58258    0.47339   1.231
mother_education5th to 9th grade                -0.10729    0.49046  -0.219
mother_educationhigher education                 0.67321    0.47722   1.411
mother_educationnone                             1.42415    1.86683   0.763
mother_educationprimary education (4th grade) -0.90521    0.62066  -1.458
father_education5th to 9th grade                 0.53994    0.47494   1.137
father_educationhigher education                 0.95396    0.49802   1.915
father_educationnone                             1.47391    2.26515   0.651
father_educationprimary education (4th grade) -0.07328    0.57235  -0.128
extra_paid_classesyes                           -0.33976    0.35521  -0.957
parent_statusLiving together                     0.04843    0.55106   0.088
                                                Pr(>|t|)
(Intercept)                                       <2e-16 ***
address_typeUrban                                 0.1071
family_supportyes                                 0.0931 .
health                                            0.1318
internet_accessyes                                0.2193
mother_education5th to 9th grade                  0.8270
mother_educationhigher education                  0.1592
mother_educationnone                              0.4461
mother_educationprimary education (4th grade)     0.1456
father_education5th to 9th grade                  0.2564
father_educationhigher education                  0.0563 .
father_educationnone                              0.5157
father_educationprimary education (4th grade)     0.8982
extra_paid_classesyes                             0.3395
parent_statusLiving together                      0.9300
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.138 on 342 degrees of freedom
Multiple R-squared:  0.09208,   Adjusted R-squared:  0.05491
F-statistic: 2.477 on 14 and 342 DF,  p-value: 0.002352
```

In this model, none of the independent variables have a statistically significant relationship with final grade at the 0.05 significance level. At the 0.10 significance level, family support and father's education (higher education) each have a statistically significant relationship with final grade. To provide further explanation, a student whose father has earned a Bachelor's degree or higher (as opposed to the standard education of a high school diploma) is predicted to achieve higher mathematics grades. We can interpret the relationship with family support in the same way that we have for the previous two models. Note that the adjusted R-squared value is equal to 0.05491, which indicates that the independent variables in this model do not explain much of the variation in final grade. The p-value of the F-test for this model is extremely small, at 0.002352, indicating that the model as a whole is statistically significant. While most of the independent variables do not have a significant effect on the final grade alone, they do have a highly significant joint effect. Therefore, socioeconomic status as a whole does have a significant impact on a student's final grade in a mathematics course.

To determine whether our model is valid and unbiased, we will once again check the validity of the three linear regression assumptions. First, we check the linearity assumption and equal variance assumption using a residual plot:



This time, the residuals are scattered randomly around the horizontal axis and do not appear to follow any sort of pattern, indicating that our data is linear. In addition to that, the residuals appear to have a constant spread across all fitted values, indicating that the equal variance assumption holds true. Finally, we will check the normality assumption using the Shapiro-Wilk test on the residuals of the model:

```
        Shapiro-Wilk normality test

data:  residuals(model_no0)
W = 0.99728, p-value = 0.8223
```

We obtain a very large p-value of 0.8223, so we can reject the null hypothesis at a significance level of 0.05 and conclude that the residuals follow a normal distribution. This new model passes all three of the assumptions necessary to perform linear regression. This means that the results can be considered to be valid and unbiased, and this will be our final multiple linear regression model.

4. Effect of Missing Data on Our Analysis

4.1 Missing Values Completely At Random (MCAR)

Data is considered to be MCAR, or “missing completely at random,” when the missing data is completely unrelated to the values of the variables for which the data is missing or other observed variables in the dataset. In other words, it’s random. There are generally two requirements that missing data must meet to be considered MCAR, as opposed to a different

type of missing data, such as MAR or MNAR. First, the missingness must be unrelated to the values of the missing data. To put it simply, the reason that a data point is missing should not depend on the value of the missing observation itself. Second, the missingness is independent of our observed data. This means that the likelihood of a particular value being missing is not influenced by other variables or observations in the dataset. Since the data was collected using a questionnaire distributed to students, many opportunities arise for MCAR data to show up in the dataset. In a survey with approximately twenty questions, a student may likely experience what is known as “survey fatigue” and unintentionally forget to answer some of the questions. If the survey was conducted electronically, there could also have been technical issues that could result in MCAR data being recorded, such as the survey skipping over a question or not allowing a student to select an answer for a particular question. To our surprise, we did not find any MCAR values in our dataset, so we simulated it by removing 20% of our data at random. To do so, we used the following method:

```
```{r}
data_MCAR <- math_data[sample(nrow(math_data),
ceiling(0.8*nrow(math_data))),]
```
```

The most common way of dealing with MCAR data when the sample size is large is generally to use listwise deletion, i.e., simply remove the rows where missing data is present. This can be done in R using the na.omit() function. Rather than assigning missing values to certain rows and subsequently deleting them, we decided to take a shortcut and take a random sample containing 80% of our data. This is approximately equivalent to randomly assigning at least one missing value to 20% of the rows in our dataset, and then deleting those rows using the na.omit() function.

4.1.1 Effect of MCAR Data on Our Analysis

To start, there was no effect on the conclusion of the Wilcoxon Rank Sum test for the effect of romantic relationship on final grade.

```
        Wilcoxon rank sum test with continuity correction

data:  final_grade by romantic_relationship
W = 12665, p-value = 0.05283
alternative hypothesis: true location shift is not equal to 0
```

Overall, our conclusion was the same. At a significance level of .05, we were unable to conclude that there is a significant difference in median final grade of students who are in a romantic relationship versus those who are not.

In addition, there was no effect on the conclusion of the Wilcoxon Rank Sum test for extracurricular activities and final grade.

```
       Wilcoxon rank sum test with continuity correction

data:  final_grade by activities
W = 12251, p-value = 0.7798
alternative hypothesis: true location shift is not equal to 0
```

The p-value is greater than .05, so we conclude that there is not a significant difference in the median final grade of students who are involved in extracurricular activities versus those who are not. We notice that the p-value is greater than that of the test for the original data, due to the lower power of the test, as the sample size is smaller than that of our original data due to the removal of MCAR values.

There was no effect on the conclusion of the Kruskal Wallis test for weekend alcohol consumption and final grade.

```
        Kruskal-Wallis rank sum test

data:  final_grade by factor(weekend_alcohol)
Kruskal-Wallis chi-squared = 5.3228, df = 4, p-value = 0.2558
```

At a significant level of 0.05, we fail to reject the null hypothesis and conclude that there is insufficient evidence that the medians are not equal across the groups. Thus, we can conclude that there is not a significant relationship between weekend alcohol consumption and mathematics final grade. Since our results are insignificant, it is unnecessary to perform a post-hoc analysis.

There was no effect on the conclusion of the Kruskal Wallis test for social outing frequency and final test score.

```
        Kruskal-Wallis rank sum test

data:  final_grade by factor(social)
Kruskal-Wallis chi-squared = 17.412, df = 4, p-value =
0.001607
```

At a significance level of 0.05, we can reject the null hypothesis and conclude that there is sufficient evidence that the medians are not equal across the groups. Thus, we can conclude that there is a significant relationship between weekend alcohol consumption and mathematics final grade. Next, we must conduct a post-hoc analysis to examine the relationship further.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = final_grade ~ factor(social), data = data_MCAR)

$`factor(social)`
           diff       lwr         upr     p adj
2-1  2.29824561 -0.818247  5.41473819 0.2571853
3-1  1.54504049 -1.515675  4.60575634 0.6377087
4-1  0.06015038 -3.113310  3.23361097 0.9999983
5-1 -0.16329285 -3.595463  3.26887738 0.9999343
3-2 -0.75320513 -2.552849  1.04643861 0.7805337
4-2 -2.23809524 -4.223439 -0.25275189 0.0182599
5-2 -2.46153846 -4.838623 -0.08445442 0.0382525
4-3 -1.48489011 -3.381478  0.41169729 0.2025483
5-3 -1.70833333 -4.011806  0.59513885 0.2517973
5-4 -0.22344322 -2.674740  2.22785332 0.9991337
```

None of the above p-values are less than .05, indicating that none of the pairwise differences are significant.

For the final multiple linear regression model, ignoring values of y=0, the MCAR data yields slightly different results than the original. The only independent variable that has a statistically significant relationship with final mathematics grade at the 0.10 significance level is father's education level. That is, a student whose father has obtained a bachelor's degree or higher will achieve a higher mathematics grade.

4.2 MNAR Data

Data is considered to be MNAR, or “missing not at random,” when the missing data is non-ignorable. If not dealt with appropriately, MNAR values present in the dataset may lead us to draw biased or inaccurate conclusions. This is because for MNAR values, missingness is directly related to the missing data itself. In the context of our dataset, there could be MNAR values present for the “final grade” variable for a variety of reasons. We hypothesize that the value for final grade is more likely to be missing when that value is low. There are two reasons for this. First, if a student is likely to receive a failing grade in the course, they may be more inclined to withdraw from the course or not complete their work, thus resulting in a missing value for the final grade column. Second, a student with a low final grade may be ashamed of their poor academic performance and decline to answer the survey question that asks for their final grade. In this case, the likelihood of the value for final grade being missing is directly tied to the unobserved value of final grade itself, thus making it MNAR data. Since we did not have any missing values that we perceived to be MNAR, we simulated it. To do so, we did the following:

```r
#Finding the bottom 20% of mathematics grade
bottom20 <- quantile(data_MNAR$final_grade, 0.2)
```

```r
#Creating new final_grade column where the bottom 20% of values are
missing
data_MNAR$final_grade <- ifelse(data_MNAR$final_grade <= bottom20, NA,
data_MNAR$final_grade)
```

We selected the values of the final grade variable that were in the bottom 20% and assigned an NA value to each one of them, thus simulating a situation where 20% of our data

contains MNAR missing values. Unfortunately, dealing with MNAR data is much more complex than dealing with MCAR data, as we need to reduce bias as much as possible. For this project, we use a technique called multiple imputation to handle missing values considered to be MNAR, using the “mice” package in R. Although multiple imputation is technically considered to be biased for MNAR missing values, we modify it to reduce the bias by including a “missingness indicator.” This is shown below:

```r
#Creating an indicator for missingness
data_MNAR$missing_indicator <- ifelse(is.na(data_MNAR$final_grade), 1, 0)
```

```r
#Creating the correct predictor matrix
predictor_matrix <- matrix(0, nrow = ncol(data_MNAR), ncol =
ncol(data_MNAR))

colnames(predictor_matrix) <- colnames(data_MNAR)
rownames(predictor_matrix) <- colnames(data_MNAR)

predictor_matrix["final_grade", "missing_indicator"] <- 1
```

The missingness indicator is a binary variable that is equal to 1 if the value for the final grade variable is missing, and it is equal to 0 if the value for the final grade variable is not missing. By including the missingness indicator, we are “informing” the imputation process that there is a pattern to the missing data so that the imputation process can account for the fact that the missingness is not random. The imputation model will include the missingness indicator as a predictor and adjust the imputations accordingly. We performed multiple imputation as follows:

```r
#Performing multiple imputation
imputed_data <- mice(data_MNAR, method = "pmm", m = 5, predictorMatrix =
predictor_matrix)
```

```r
complete_data <- complete(imputed_data)
```

Should we wish to handle the MNAR missing data more accurately, it may require more advanced techniques, such as pattern-mixture models or selection models. These types of models explicitly model the relationship between the missing data and the unobserved values, rather than assuming that missingness can be explained by the observed data. However, these

types of models tend to be computationally expensive and time-consuming. They also tend to be a bit difficult to interpret, so for the sake of our analysis, we chose to stick with the multiple imputation method.

4.2.1 Effect of MNAR Data on Our Analysis

There was no effect on the conclusion of the Wilcoxon Rank Sum test for the effect of romantic relationship on final grade.

```
        Wilcoxon rank sum test with continuity correction

data:  final_grade by romantic_relationship
W = 9771, p-value = 0.4319
alternative hypothesis: true location shift is not equal to 0
```

The p-value is greater than .05, so we conclude there is insufficient evidence that there is a significant difference in the final grade of students who are in a romantic relationship versus those who are not.

There was no effect on the conclusion of the Wilcoxon test for the effect of involvement in extracurricular activities on final grade.

```
        Wilcoxon rank sum test with continuity correction

data:  final_grade by activities
W = 10929, p-value = 0.7535
alternative hypothesis: true location shift is not equal to 0
```

The p-value is greater than .05, so we conclude that there is not a significant difference in the median final grade of students who are involved in activities versus those who are not.

The MNAR data significantly altered the results of the Kruskal Wallis test for weekend alcohol consumption.

```
        Kruskal-Wallis rank sum test

data:  final_grade by factor(weekend_alcohol)
Kruskal-Wallis chi-squared = 12.309, df = 4, p-value = 0.0152
```

At a significant level of 0.05, we reject the null hypothesis and conclude that there is sufficient evidence that the medians are not equal across the groups. Thus, we can conclude that there is a significant relationship between weekend alcohol consumption and final grade. Since our results are significant, we will perform a post-hoc analysis.

According to the results of the Tukey-HSD test, there is a significant difference in final grade between weekend_alcohol = 4 and weekend_alcohol = 1. Students with a “high” weekend alcohol consumption earn lower mathematics grades than students with “very low” alcohol consumption.

The results of our Kruskal Wallis test for social outing frequency were also significantly altered.

```
        Kruskal-Wallis rank sum test

data:  final_grade by factor(social)
Kruskal-Wallis chi-squared = 2.7134, df = 4, p-value = 0.6069
```

At a significance level of 0.05, we cannot reject the null hypothesis as there is insufficient evidence that the medians are not equal across the groups. Thus, we cannot conclude that there is a significant relationship between social outing frequency and mathematics final grade.

For the multiple linear regression model, ignoring values of y=0, the MNAR data yields slightly different results than the original. The independent variables that have statistically significant relationships with final mathematics grade at the 0.10 significance level are mother's education level and health.

5. Conclusion/Summary of Results

Our analysis highlights the complex relationship between social factors and performance on mathematics exams. A key finding is that, while individual variables such as romantic relationships and involvement in extracurricular activities do not produce a statistically

significant change in mathematics scores, the collective impact of social variables is significant. Specifically, the p-value of the F-test for our multiple linear regression model is very small (0.002352), emphasizing the overall statistical significance of the model. This finding underscores that, although individual aspects of a student's social life may not have a notable influence on their math performance, their combined effect is impactful.

We conclude that social life, as a whole, does affect math performance. Future analyses should aim to disentangle these complex relationships further by considering geographical and cultural factors. We acknowledge that our data is limited to a Portuguese population, which may restrict the generalizability of our findings. Exploring these dynamics in different cultural contexts could yield valuable insights. Such research could guide educators and mental health professionals in developing tailored strategies to improve student outcomes and academic success.

References

Szczygieł, M. (2020). When does math anxiety in parents and teachers predict math anxiety and math achievement in elementary school children? The role of gender and grade year. Social Psychology of Education, 23(4), 1023-1054.

Tarkar, A., Matalka, B., Cartwright, M., & Kloos, H. (2022). Student-Guided Math Practice in Elementary School: Relation among Math Anxiety, Emotional Self-Efficacy, and Children's Choices When Practicing Math. Education Sciences, 12(9), 611.