AMS 572 Project

Jane Condon, Valentina Tillmann

2024-12-02

Influence of Social Life and Socioeconomic Plus Demographic Factors on a High School Student's Mathematics Grades

Introduction: Introducing and Preparing the Data for Analysis

Downloading R packages that we may need

```
install_if_needed <- function(package) {
  if (!requireNamespace(package, quietly = TRUE)) {
    install.packages(package)
  }
}

install_if_needed("tidyverse")
install_if_needed("dplyr")
install_if_needed("MASS")
install_if_needed("pscl")
install_if_needed("pheatmap")
install_if_needed("reshape2")
install_if_needed("mice")
install_if_needed("car")
install_if_needed("ggplot2")

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(pscl)
```

```
## Classes and Methods for R originally developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University (2002-2015),
## by and under the direction of Simon Jackman.
## hurdle and zeroinfl functions by Achim Zeileis.
```

```r
library(MASS)
```

```
##
## Attaching package: ’MASS’
##
## The following object is masked from ’package:dplyr’:
##
##     select
```

```r
library(pheatmap)
library(reshape2)
```

```
##
## Attaching package: ’reshape2’
##
## The following object is masked from ’package:tidyr’:
##
##     smiths
```

```r
library(mice)
```

```
##
## Attaching package: ’mice’
##
## The following object is masked from ’package:stats’:
##
##     filter
##
## The following objects are masked from ’package:base’:
##
##     cbind, rbind
```

```r
library(car)
```

```
## Loading required package: carData
##
## Attaching package: ’car’
##
## The following object is masked from ’package:dplyr’:
##
##     recode
##
## The following object is masked from ’package:purrr’:
##
##     some
```

```r
library(ggplot2)
```

Loading the dataset

```r
#Reading the csv file
math_data <- read.csv("student_math_clean.csv")
```

```r
#Displaying the dataset
head(math_data)
```

```
##   student_id school sex age address_type            family_size
## 1          1     GP   F  18        Urban          Greater than 3
## 2          2     GP   F  17        Urban          Greater than 3
## 3          3     GP   F  15        Urban Less than or equal to 3
## 4          4     GP   F  15        Urban          Greater than 3
## 5          5     GP   F  16        Urban          Greater than 3
## 6          6     GP   M  16        Urban Less than or equal to 3
##     parent_status              mother_education              father_education
## 1           Apart              higher education              higher education
## 2 Living together primary education (4th grade) primary education (4th grade)
## 3 Living together primary education (4th grade) primary education (4th grade)
## 4 Living together              higher education              5th to 9th grade
## 5 Living together           secondary education           secondary education
## 6 Living together              higher education           secondary education
##   mother_job father_job school_choice_reason guardian   travel_time
## 1    at_home    teacher               course   mother 15 to 30 min.
## 2    at_home      other               course   father      <15 min.
## 3    at_home      other                other   mother      <15 min.
## 4     health   services                 home   mother      <15 min.
## 5      other      other                 home   father      <15 min.
## 6   services      other           reputation   mother      <15 min.
##      study_time class_failures school_support family_support extra_paid_classes
## 1  2 to 5 hours              0            yes             no                 no
## 2  2 to 5 hours              0             no            yes                 no
## 3  2 to 5 hours              3            yes             no                yes
## 4 5 to 10 hours              0             no            yes                yes
## 5  2 to 5 hours              0             no            yes                yes
## 6  2 to 5 hours              0             no            yes                yes
##   activities nursery_school higher_ed internet_access romantic_relationship
## 1         no            yes       yes              no                    no
## 2         no             no       yes             yes                    no
## 3         no            yes       yes             yes                    no
## 4        yes            yes       yes             yes                   yes
## 5         no            yes       yes              no                    no
## 6        yes            yes       yes             yes                    no
##   family_relationship free_time social weekday_alcohol weekend_alcohol health
## 1                   4         3      4               1               1      3
## 2                   5         3      3               1               1      3
## 3                   4         3      2               2               3      3
## 4                   3         2      2               1               1      5
## 5                   4         3      2               1               2      5
## 6                   5         4      2               1               2      5
##   absences grade_1 grade_2 final_grade
## 1        6       5       6           6
## 2        4       5       5           6
## 3       10       7       8          10
## 4        2      15      14          15
## 5        4       6      10          10
## 6       10      15      15          15
```

```
#Viewing the types of variables we have to verify that we have both numerical and categorical variables
sapply(math_data, class)
```

```
##            student_id                school                   sex
##             "integer"           "character"           "character"
##                   age          address_type           family_size
##             "integer"           "character"           "character"
##         parent_status      mother_education      father_education
##           "character"           "character"           "character"
##            mother_job            father_job  school_choice_reason
##           "character"           "character"           "character"
##              guardian           travel_time            study_time
##           "character"           "character"           "character"
##        class_failures        school_support        family_support
##             "integer"           "character"           "character"
##    extra_paid_classes            activities        nursery_school
##           "character"           "character"           "character"
##             higher_ed       internet_access romantic_relationship
##           "character"           "character"           "character"
##   family_relationship             free_time                social
##             "integer"             "integer"             "integer"
##       weekday_alcohol       weekend_alcohol                health
##             "integer"             "integer"             "integer"
##              absences               grade_1               grade_2
##             "integer"             "integer"             "integer"
##           final_grade
##             "integer"
```

Our categorical variables are of the "character" type, so we must turn these into factor variables.

Data Cleaning and Preparation for Analysis

```
#Selecting only the variables that we are using in our analysis to be part of the dataframe
math_data <- math_data[, c("final_grade", "social", "weekend_alcohol","activities","romantic_relationshi
```

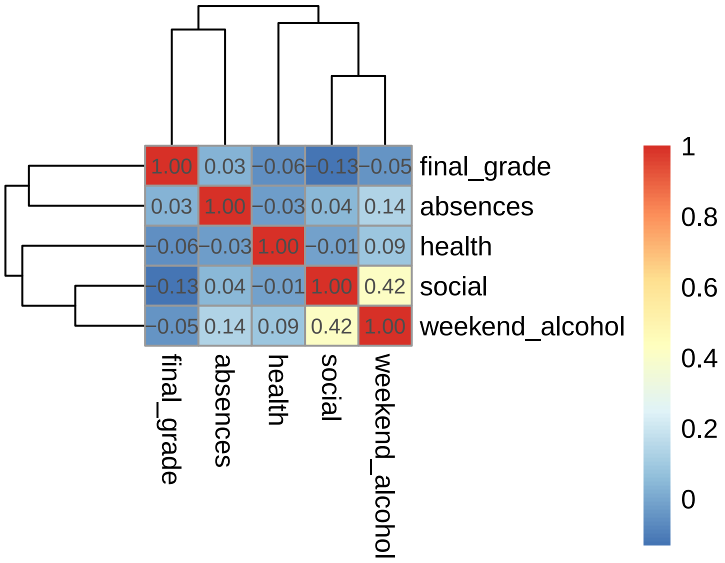
```
#Turning categorical variables into factor variables (variable with multiple levels)
math_data <- math_data %>%
  mutate(across(c(address_type,parent_status, family_support, extra_paid_classes, activities, internet_a

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

We transform mother's education, father's education, travel time, and study time, into ordered factor variables, as the levels of these variables do have a defined order (i.e., mother's education can be categorized as "none," "primary education," "5th to 9th grade," etc.). We set "secondary education" as the "baseline" for linear regression, as a high school education is the "standard" education level. All other categorical variables are nominal or binary, so we can factor these normally, with no defined order.

Exploratory Data Analysis

Data Visualization

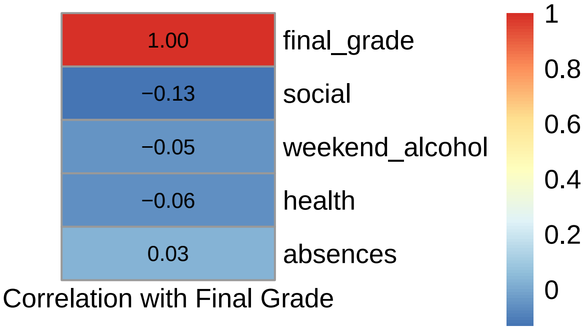
```r
#Creating a heatmap showing correlation between numeric variables
numeric_data <- math_data[sapply(math_data, is.numeric)]
cor_matrix <- cor(numeric_data, use = "complete.obs")
pheatmap(cor_matrix,display_numbers = TRUE,cellwidth=20,cellheight=15)
```



Shown above is the Heatmap describing the correlation between numeric variables in our data. Weekend alcohol and social score have a moderate correlation of .42. Other variables do not show a signficant pairwise correlation.

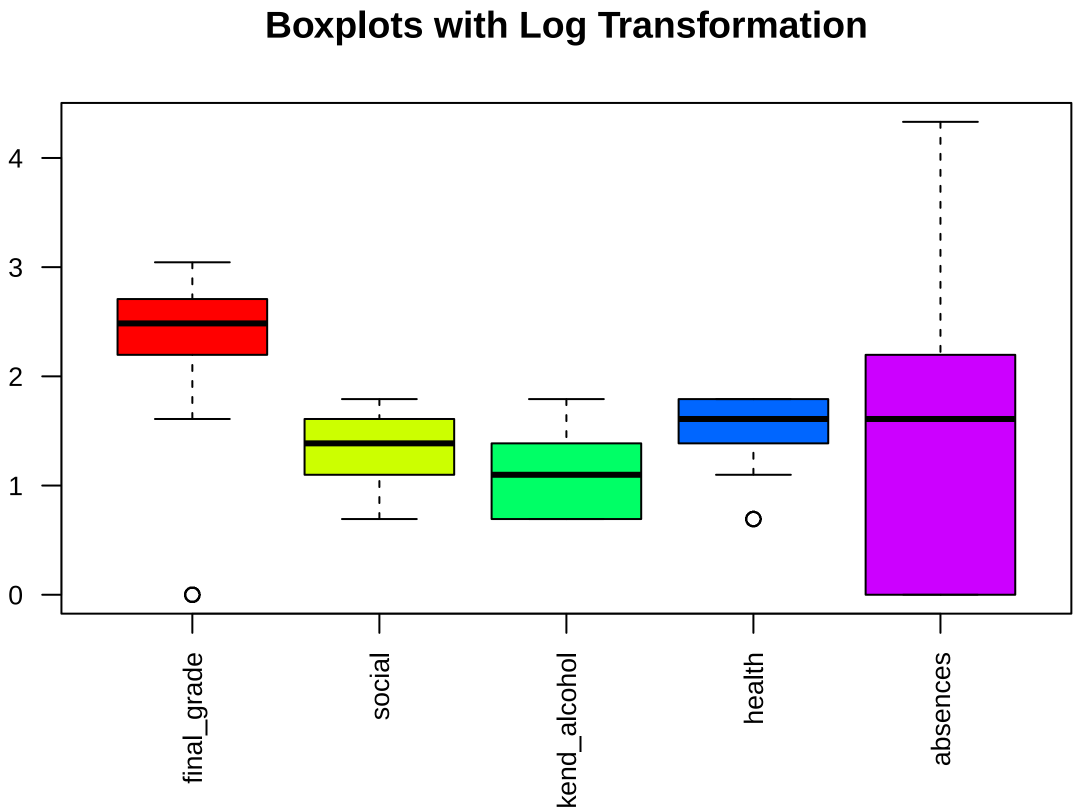
```r
#Creating heatmap showing correlation of numeric variables with final grade
target_variable <- "final_grade"
cor_with_target <- cor(numeric_data, numeric_data[[target_variable]], use = "complete.obs")
cor_df <- as.data.frame(cor_with_target)
colnames(cor_df) <- c("Correlation with Final Grade")
```

```r
row.names(cor_df) = names(numeric_data)
pheatmap(as.matrix(cor_df), display_numbers = TRUE, cluster_rows = FALSE, cluster_cols = FALSE,angle_col
```



We can see from the Heatmap above that alcohol consumption has a weak negative correlation with Final Grade.

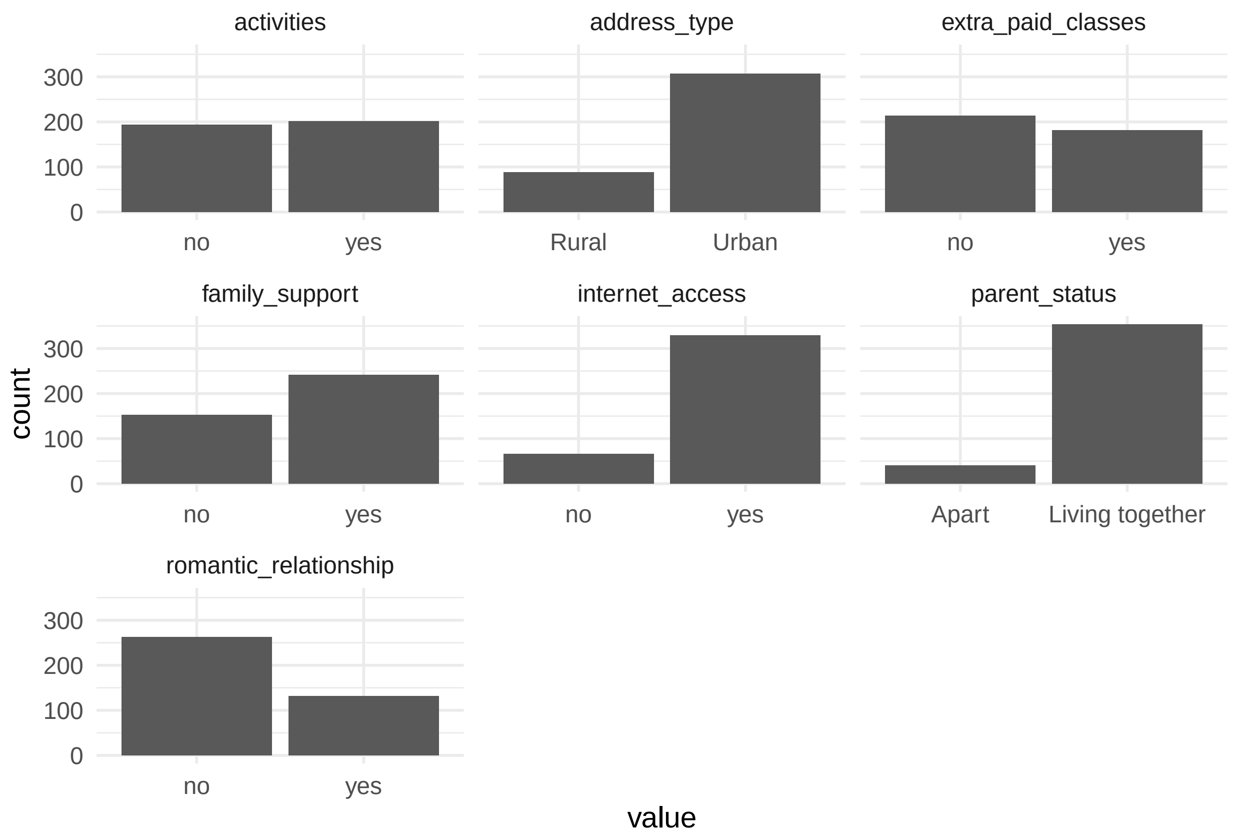
```r
#Creating side by side boxplots to show distributions of numeric variables
boxplot(log1p(numeric_data),
        main = "Boxplots with Log Transformation",
        las = 2,
        col = rainbow(ncol(numeric_data)),cex.axis = 0.8)
```



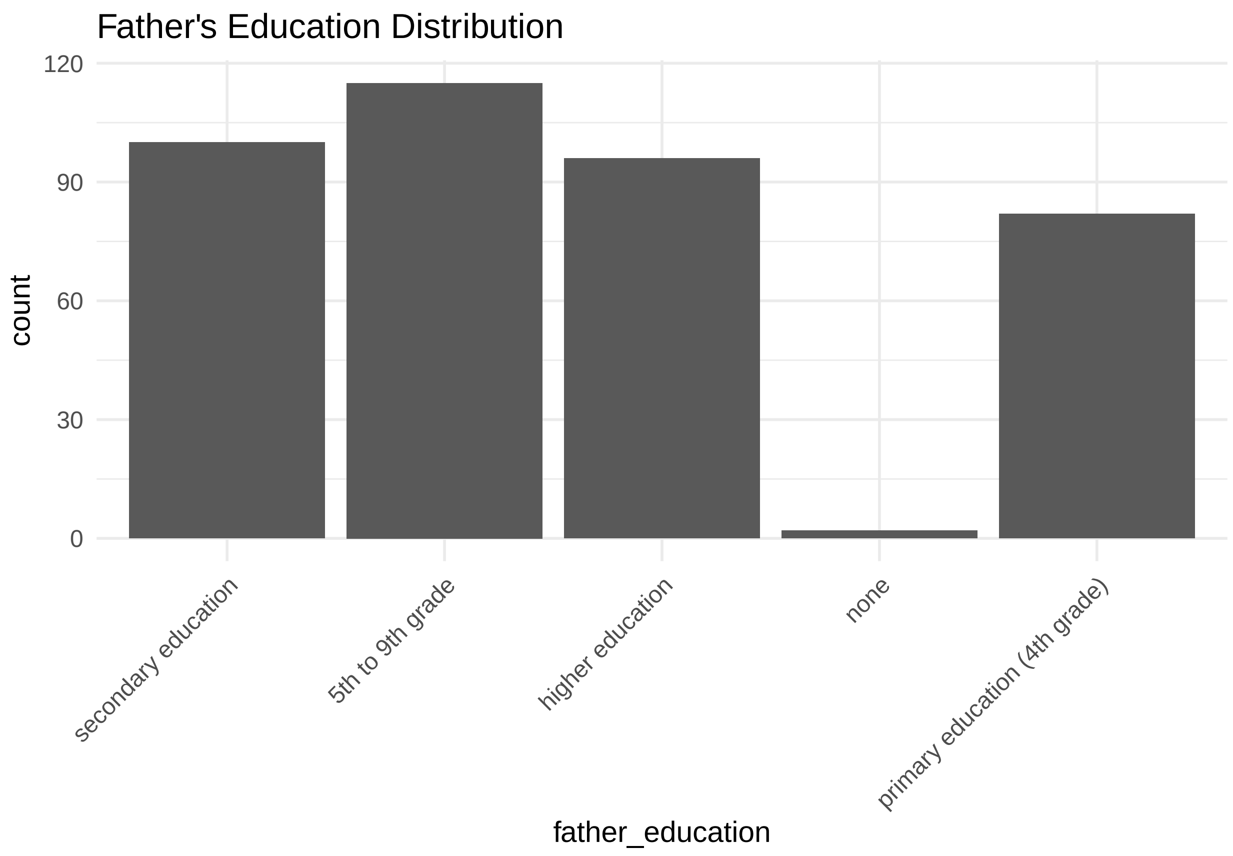
‘final_grade’ and ‘social’ appear approximately symmetric, while ‘weekend_alcohol’, ‘health’, and ‘absences’ do not. The log transformation is applied due to the severity of outliers in the ‘absences’ column.

Visualization of factor variables:

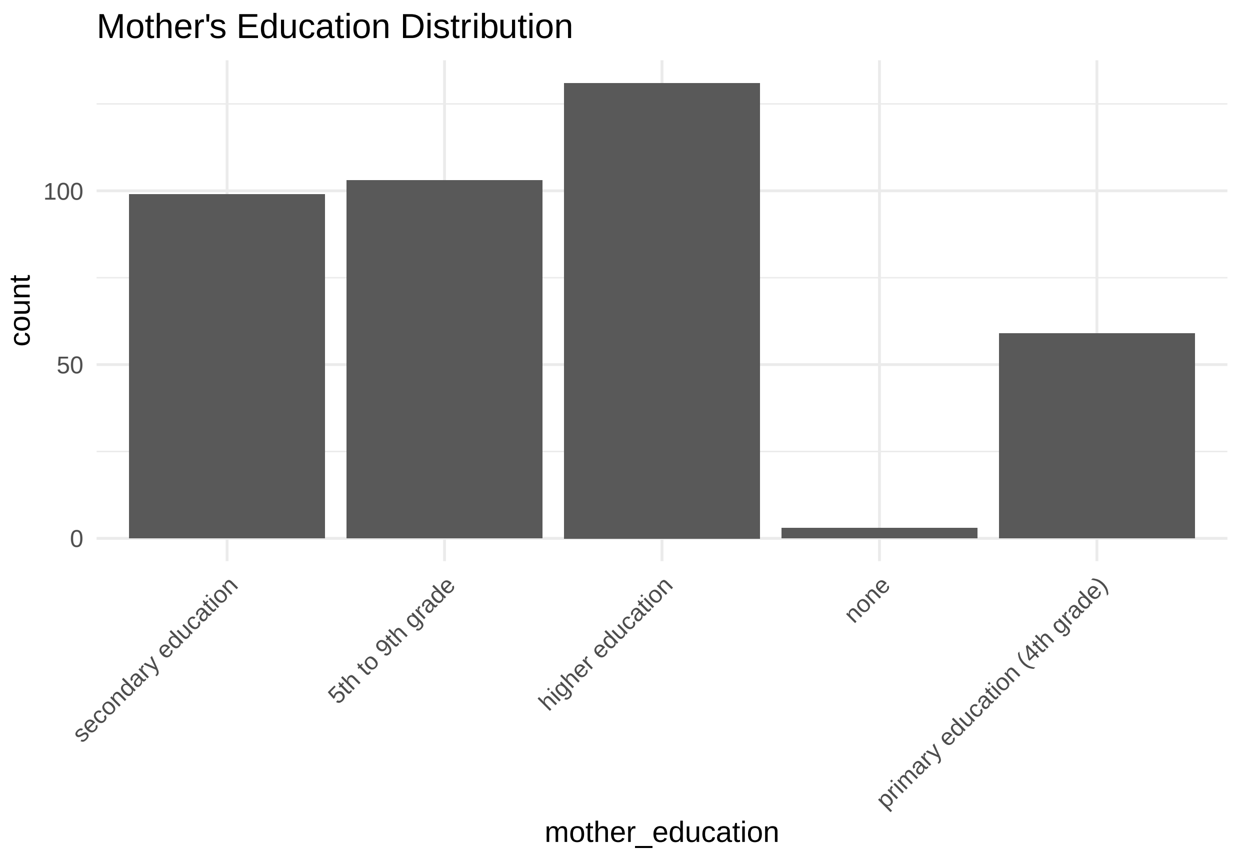
```r
# Filter out columns 'father_education' and 'mother_education'
factor_data <- math_data[sapply(math_data, is.factor)]
factor_data <- factor_data[, !(colnames(factor_data) %in% c("father_education", "mother_education"))]

# Convert to long format for plotting
factor_data_long <- tidyr::pivot_longer(factor_data, cols = everything(), names_to = "variable", values_

# Create the ggplot
ggplot(factor_data_long, aes(x = value)) +
  geom_bar() +
  facet_wrap(~ variable, scales = "free_x") +
  theme_minimal()
```



```
factor_data_father <- math_data["father_education"]
factor_data_mother <- math_data["mother_education"]

ggplot(factor_data_father, aes(x = father_education)) +
  geom_bar() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  ggtitle("Father's Education Distribution")
```



```
ggplot(factor_data_mother, aes(x = mother_education)) +
  geom_bar() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  ggtitle("Mother's Education Distribution")
```

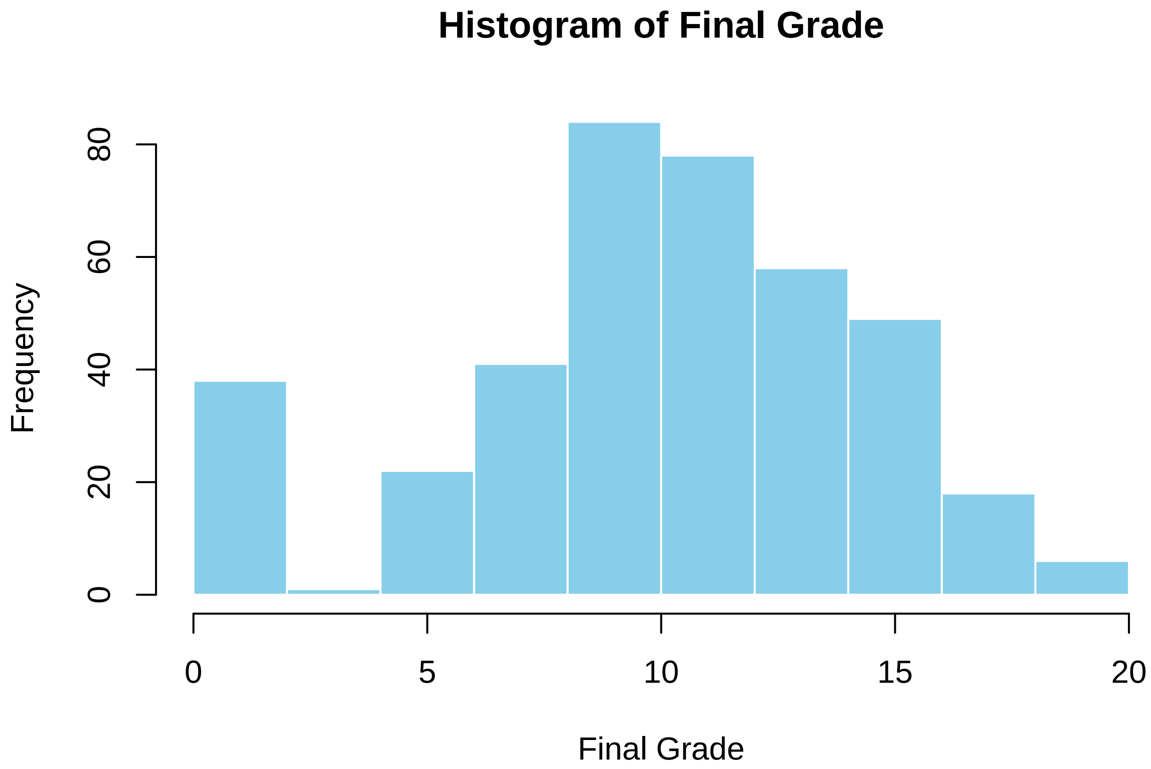


Summarizing 'final_grade":

```
summary(math_data$final_grade)
```

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

```
hist(math_data$final_grade,
     main = "Histogram of Final Grade",
     xlab = "Final Grade",
     col = "skyblue",
     border = "white",
     breaks = 10)
```



The grades do not appear normally distributed. There is an unusual amount of zeroes in the data.

Checking whether or not 'final_grade' follows a normal distribution:

```
shapiro.test(math_data$final_grade)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  math_data$final_grade
## W = 0.92873, p-value = 8.836e-13
```

The miniscule p-value confirms our assumption from the histogram. 'final_grade' is not normally distributed.

Hypothesis 1: Is Having a More Active Social Life Associated with a Better or Worse Mathematics Grade?

To explore this hypothesis, we will look at a variety of factors such as involvement in extracurricular activities, involvement in a romantic relationship, weekend alcohol consumption, how often a student engages in social outings, etc. Then, we will use multiple methods, including the Wilcoxon Rank Sum test (non-parametric alternative to two-sample t-test) and the Kruskal Wallis test (non-parametric alternative to ANOVA test), to determine whether the mean mathematics grade for the students who have a more active social life is less than that of students who have a less active social life.

Two Sample T-test (Wilcoxon Rank Sum test) to Test the Difference in Median Final Grade for Students Who Are in a Romantic Relationship vs Students Who Are Not

Checking the assumptions to determine which type of t-test to use Normality Assumption:

```
shapiro.test(math_data$final_grade[math_data$romantic_relationship == "no"])  # For group 0
```

```
##
##  Shapiro-Wilk normality test
##
## data:  math_data$final_grade[math_data$romantic_relationship == "no"]
## W = 0.9445, p-value = 2.009e-08
```

```
shapiro.test(math_data$final_grade[math_data$romantic_relationship == "yes"])  # For group 1
```

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

To test if the final mathematics grade variable follows a normal distribution, we use a Shapiro test for both 'groups,' with one group being the students who are in a relationship and the other group being the students who are not in a relationship. As evident by the output above, the p-value for both tests is far below our significance level of 0.05. Thus, we reject the null hypothesis and can conclude that the data does not follow a normal distribution for either group. Thus, we must use a Wilcoxon signed rank test, rather than the standard t-test.

Equal Variance Assumption:

```
leveneTest(final_grade ~ factor(romantic_relationship), data = math_data)
```

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

Since the data is not normally distributed, we use Levene's test to check for equal variance between groups, since Levene's test is more robust than Bartlett's test and less sensitive to departures from normality. Given a p-value of 0.2775, we cannot reject the null hypothesis and conclude that there is insufficient evidence that the variances are unequal at a significance level of 0.05. Thus, we can assume that the variances are equal across the two groups.

```
# Two-sided Wilcoxon Rank Sum test
w_test_result <- wilcox.test(final_grade ~ romantic_relationship, data = math_data, alternative = )
print(w_test_result)
```

Wilcoxon Test to Test the Difference in Median Final Grade for Students in a Romantic Relationship vs Students Who Aren't

```
##
##  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
```

```
# One-sided Wilcoxon Rank Sum test with alternative = "greater"
w_test_result_greater <- wilcox.test(final_grade ~ romantic_relationship, data = math_data, alternative
print(w_test_result_greater)
```

```
##
##  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
```

Since the normality assumption has been violated and the equal variance assumption appears to hold true, we use a Wilcoxon signed rank test. From this 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 insufficient evidence that there is a significant difference in median final grade of students who are in a romantic relationship versus those who are not. At 0.10, we can reject the null hypothesis and conclude that there is a significant difference in median final grade of students who are in a romantic relationship versus those who are not. Since the result is significant at alpha = 0.10, we also test to see if the median mathematics grade for students involved in a romantic relationship is greater than that of students not involved in a romantic relationship. From this test, we obtain a p-value of 0.03476. At a significance level of 0.05, we reject the null hypothesis and conclude that students who are involved in a romantic relationship tend to achieve higher mathematics grade than those who are not.

Two Sample T-test (Wilcoxon Rank Sum test) to Test the Difference in Median Final Grade for Students Involved in Extracurricular Activities vs Students Who Aren't

Checking the assumptions to determine which type of t-test to use Normality Assumption:

```
shapiro.test(math_data$final_grade[math_data$activities == "no"])  # For group 0
```

```
##
##  Shapiro-Wilk normality test
##
## data:  math_data$final_grade[math_data$activities == "no"]
## W = 0.93723, p-value = 1.909e-07
```

```
shapiro.test(math_data$final_grade[math_data$activities == "yes"])  # For group 1
```

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

To test if the final mathematics grade variable follows a normal distribution, we use a Shapiro test for both ‘groups,’ with one group being the students who are involved in extracurricular activities and the other group being the students who are not involved in extracurricular activities. As evident by the output above, the p-value for both tests is far below our significance level of 0.05. Thus, we reject the null hypothesis and can conclude that the data does not follow a normal distribution for either group. Thus, we must use a Wilcoxon Rank Sum test, rather than the standard two-sample t-test.

Equal Variance Assumption:

```
leveneTest(final_grade ~ factor(activities), data = math_data)
```

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

Since the data is not normally distributed, we use Levene’s test to check for equal variance between groups, since Levene’s test is more robust than Bartlett’s test and less sensitive to departures from normality. Given a p-value of 0.733, we cannot reject the null hypothesis and conclude that there is insufficient evidence that the variances are unequal at a significance level of 0.05. Thus, we can assume that the variances are equal across the two groups.

```
# Two sided Wilcoxon Rank Sum test
w_test_result <- wilcox.test(final_grade ~ activities, data = math_data, alternative = )
print(w_test_result)
```

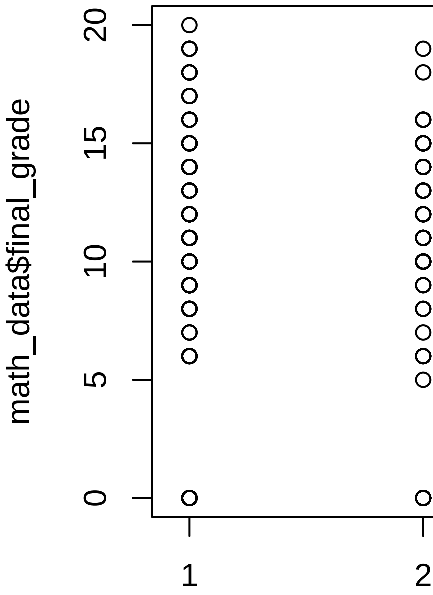
Wilcoxon Test to Test the Difference in Median Final Grade for Students Involved in Extracurricular Activities vs Those Who Aren’t

```
##
##  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
```

Since the normality assumption has been violated and the equal variance assumption appears to hold true, we use a Wilcoxon Rank Sum test. From this test, we obtain a p-value of 0.6049. At a significance level of 0.05, we fail to reject the null hypothesis and conclude that there is insufficient evidence that there is a significant difference in median final grade of students who are involved in extracurricular activities versus those who are not.

ANOVA Test (Kruskal Wallis Test) to Test the Relationship Between Final Grade and Weekend Alcohol Consumption

```
plot(math_data$weekend_alcohol, math_data$final_grade)
```



Why do we use an ANOVA test, rather than a Pearson's Correlation Test?:

As shown in the plot above, weekend alcohol consumption can be considered to be an ordinal categorical variable, since there are only 5 unique values, which can be thought of as "levels." So, we have 5 "groups" and we would like to analyze whether there is a difference in the mean final grade between the five groups. Thus, it is more appropriate to use an ANOVA test, rather than Pearson's correlation test.

Checking the ANOVA Assumptions Fitting an ANOVA model:

```
anova_model <- aov(final_grade ~ weekend_alcohol, data = math_data)
```

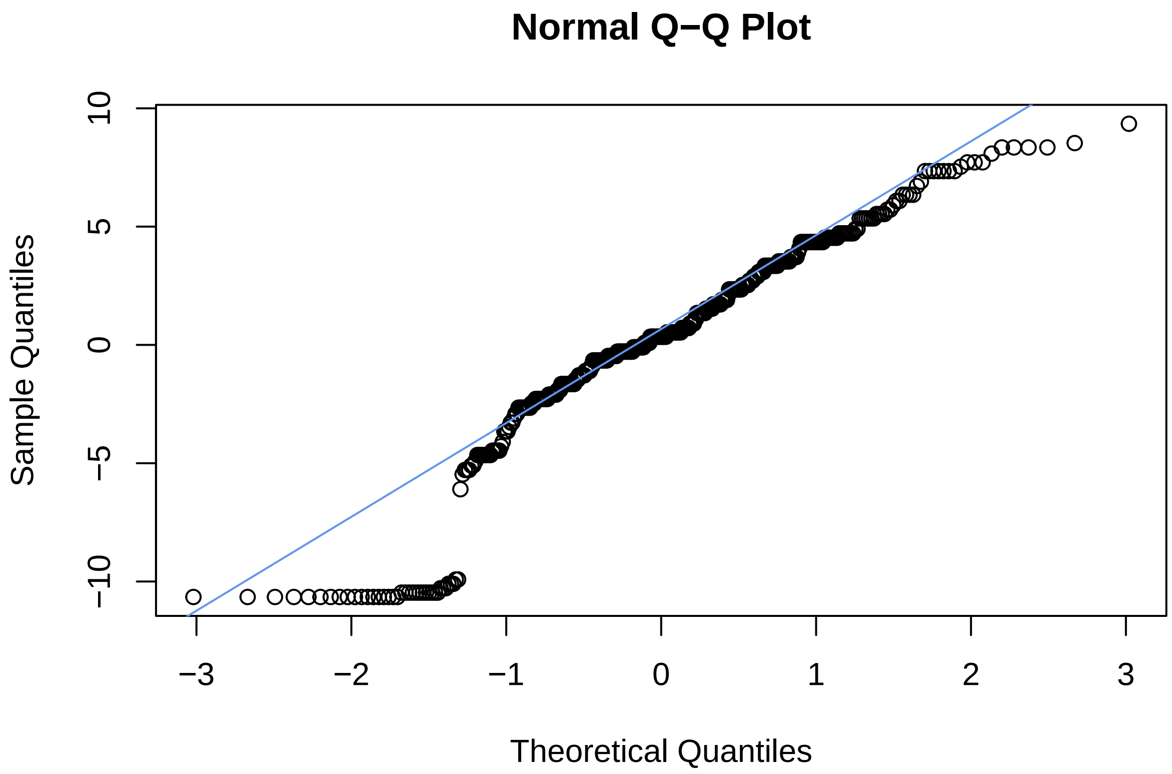
Checking for Normality:

```
# Checking for normality of residuals
aov_residuals <- residuals(anova_model)
shapiro.test(aov_residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.93045, p-value = 1.341e-12
```

Using a significance level of 0.05, we can reject the null hypothesis and conclude that there is sufficient evidence that the residuals do not follow a normal distribution. Thus, the normality assumption has been violated.

```
# QQ Plot to check for normality
qqnorm(aov_residuals)
qqline(aov_residuals, col = "cornflowerblue")
```



We reach a similar conclusion about the normality of the data from the QQ plot. It is clear that the residuals do not follow a normal distribution. Thus, the normality assumption has been violated and we should consider a nonparametric alternative to the ANOVA test, such as the Kruskal Wallis test.

Checking for homogeneity of variance:

```
leveneTest(final_grade ~ factor(weekend_alcohol), data = math_data)
```

```
## 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
```

Using a significance level of 0.05, we can reject the null hypothesis and conclude that there is sufficient evidence that the variance is not the same across all groups. Thus, the homogeneity of variances assumption has been violated.

Since both of the assumptions for ANOVA have been violated, we will use a nonparametric alternative to the ANOVA test, known as the Kruskal Wallis test.

```
kruskal_result <- kruskal.test(final_grade ~ factor(weekend_alcohol), data = math_data)
print(kruskal_result)
```

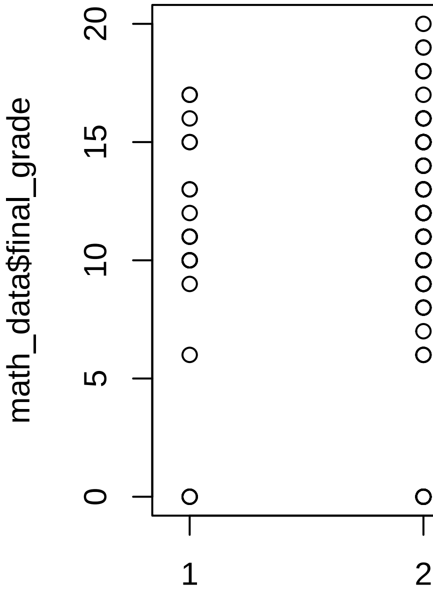
Kruskal Wallis Test to Test the Relationship Between Final Grade and Weekend Alcohol Consumption

```
##
##  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 a significance 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 difference in the medians across the five groups. There is not a significant relationship between weekend alcohol consumption and final grade. Since our results are insignificant, it is unnecessary to perform post-hoc analysis.

ANOVA Test (Kruskal Wallis test) to Test the Relationship Between Final Grade and Frequency of Social Outings

```
plot(math_data$social, math_data$final_grade)
```



Why do we use an ANOVA test, rather than a Pearson's Correlation Test?:

As shown in the plot above, frequency of social outings can be considered to be an ordinal categorical variable, since there are only 5 unique values, which can be thought of as “levels.” So, we have 5 “groups” and we would like to analyze whether there is a difference in the mean final grade between the five groups. Thus, it is more appropriate to use an ANOVA test, rather than Pearson’s correlation test.

Checking the ANOVA Assumptions Fitting an ANOVA model:

```r
anova_model2 <- aov(final_grade ~ social, data = math_data)
```

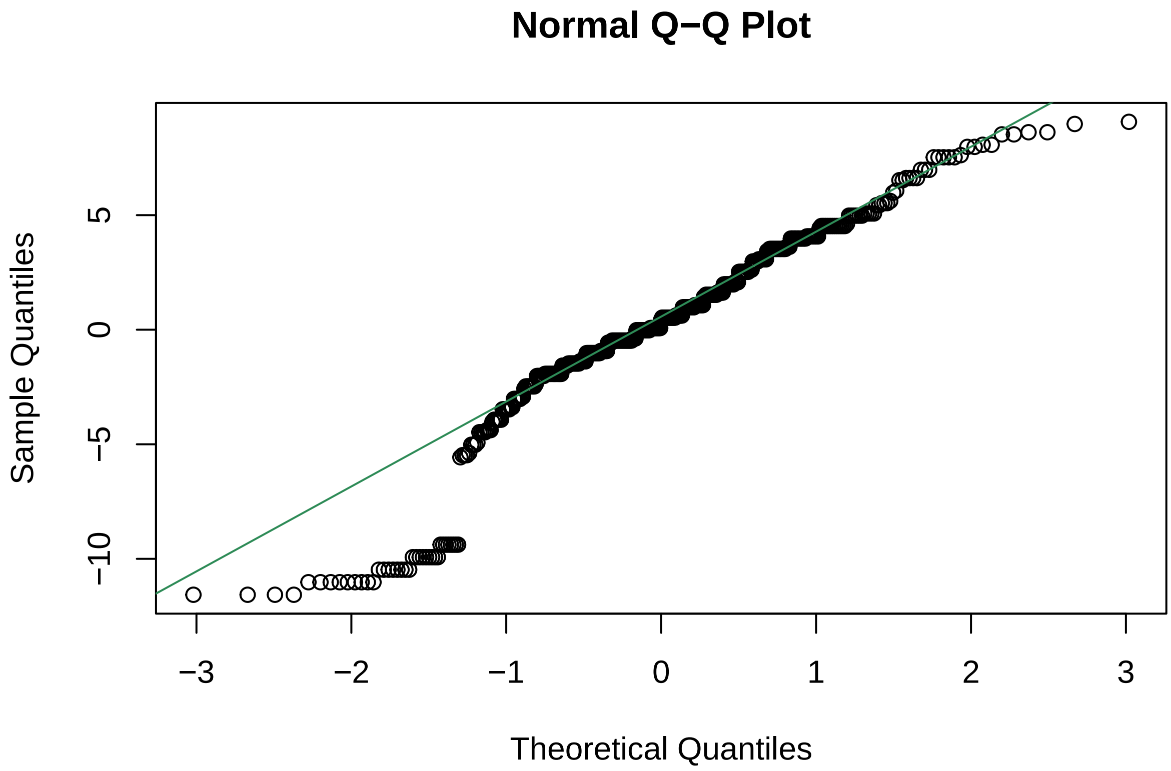
Checking for Normality:

```r
# Checking for normality of residuals
aov_residuals2 <- residuals(anova_model2)
shapiro.test(aov_residuals2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals2
## W = 0.93944, p-value = 1.328e-11
```

Using a significance level of 0.05, we can reject the null hypothesis and conclude that there is sufficient evidence that the residuals do not follow a normal distribution. Thus, the normality assumption has been violated.

```r
# QQ Plot to check for normality
qqnorm(aov_residuals2)
qqline(aov_residuals2, col = "seagreen")
```



We reach a similar conclusion about the normality of the data from the QQ plot. It is clear that the residuals do not follow a normal distribution. Thus, the normality assumption has been violated and we should consider a nonparametric alternative to the ANOVA test, such as the Kruskal Wallis test.

Checking for homogeneity of variances:

```
leveneTest(final_grade ~ factor(social), data = math_data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   4  0.6346 0.6381
##       390
```

Using a significance level of 0.05, we cannot reject the null hypothesis and conclude that there is insufficient evidence that the variance is not the same across all groups. Thus, the homogeneity of variances assumption appears to hold true.

Since the normality of residuals assumption has been violated, we will use a nonparametric alternative to the ANOVA test, known as the Kruskal Wallis test.

```


kruskal_result2 <- kruskal.test(final_grade ~ factor(social), data = math_data)
print(kruskal_result2)
```

Kruskal Wallis Test to Test the Relationship Between Final Grade and Frequency of Social Outings

```
##
##  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 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 social outing frequency and final grade. Next, we must conduct a post-hoc analysis to examine the relationship further.

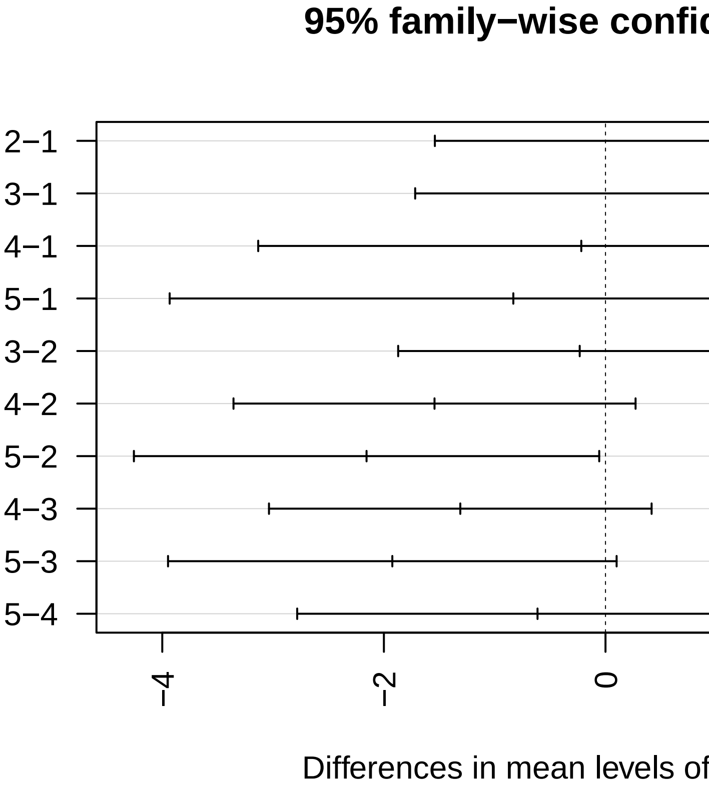
```
anova_model2 <- aov(final_grade ~ factor(social), data = math_data)
tukey_test <- TukeyHSD(anova_model2, conf.level=.95)
tukey_test
```

Pairwise Comparisons using Tukey HSD Test

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = final_grade ~ factor(social), data = math_data)
##
## $‘factor(social)‘
##           diff       lwr         upr     p adj
## 2-1  1.3246095 -1.539942  4.18916083 0.7114664
## 3-1  1.0919732 -1.717753  3.90169932 0.8243028
## 4-1 -0.2184024 -3.134177  2.69737229 0.9996036
## 5-1 -0.8318294 -3.933237  2.26957793 0.9481981
## 3-2 -0.2326363 -1.871118  1.40584517 0.9951370
## 4-2 -1.5430120 -3.357345  0.27132072 0.1373977
## 5-2 -2.1564389 -4.256150 -0.05672748 0.0408022
## 4-3 -1.3103757 -3.036849  0.41609750 0.2307900
## 5-3 -1.9238026 -3.948079  0.10047371 0.0714703
## 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 alpha = .05. There is a statistically significant difference between group 5 and group 3 at alpha = 0.10. This indicates that going out with friends ‘very often’ is associated with having a lower final grade, compared to ‘somewhat often’ and ‘not often’.

```
plot(tukey_test, las = 2)
```



Visualizing Confidence Intervals from Tukey Test

According to the plot above, we can be 95% confident that the true difference in mean between group 5 and group 2 is between -4 and -.05 (the mean math scores of social score = 5 is lower than the mean math scores of social score = 2.)

Hypothesis 2: What Kind of Socioeconomic and Demographic Factors Have the Strongest Effect on Final Grade?

To explore this hypothesis, we will use a multiple linear regression model. We will use mathematics final grade as the dependent variable, and use a variety of independent variables including address type (urban or rural), family support, health status, access to internet, mother's and father's education level, student's participation in extra paid classes, and parent status (parents living together vs parents living apart).

Fitting a Multiple Linear Regression Model

Final Grade = Address Type + Family Support + Health + Internet Access + Mother's Education Level + Father's Education Level + Extra Paid Classes + Parent Status + error.

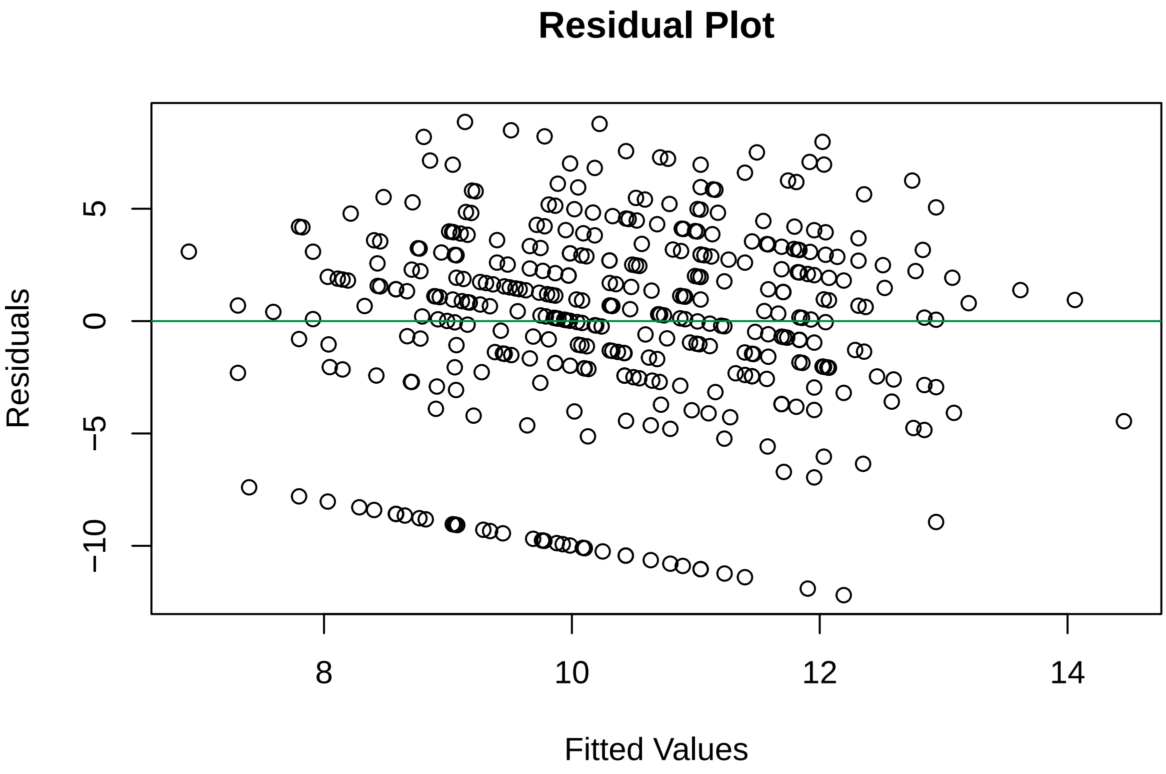
```r
# Simple multivariate linear regression model
model <- lm(final_grade ~ address_type + family_support + health + internet_access + mother_education +
summary(model)
```

```
##
## Call:
## lm(formula = final_grade ~ address_type + family_support + health +
##     internet_access + mother_education + father_education + extra_paid_classes + 
##     parent_status, data = math_data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -12.1937  -2.0680   0.6614   2.9567   8.8625 
## 
## Coefficients:
##                                               Estimate Std. Error t value
## (Intercept)                                    10.5713     1.2140   8.708
## address_typeUrban                               0.6058     0.5612   1.079
## family_supportyes                              -1.0086     0.4919  -2.050
## health                                         -0.1192     0.1655  -0.720
## internet_accessyes                              0.5154     0.6391   0.807
## mother_education5th to 9th grade               -0.5711     0.6620  -0.863
## mother_educationhigher education                1.1479     0.6509   1.764
## mother_educationnone                            2.6136     2.6492   0.987
## mother_educationprimary education (4th grade)  -1.3121     0.8329  -1.575
## father_education5th to 9th grade                0.1311     0.6488   0.202
## father_educationhigher education                0.3824     0.6784   0.564
## father_educationnone                            2.3130     3.2158   0.719
## father_educationprimary education (4th grade)  -0.2855     0.7724  -0.370
## extra_paid_classesyes                           0.7972     0.4830   1.650
## parent_statusLiving together                   -0.5794     0.7580  -0.764
##                                               Pr(>|t|)    
## (Intercept)                                     <2e-16 ***
## address_typeUrban                               0.2811    
## family_supportyes                               0.0410 *  
## health                                          0.4720    
## internet_accessyes                              0.4205    
## mother_education5th to 9th grade                0.3889    
## mother_educationhigher education                0.0786 .  
## mother_educationnone                            0.3245    
## mother_educationprimary education (4th grade)   0.1160    
## father_education5th to 9th grade                0.8400    
## father_educationhigher education                0.5733    
## father_educationnone                            0.4724    
## father_educationprimary education (4th grade)   0.7119    
## extra_paid_classesyes                           0.0997 .  
## parent_statusLiving together                    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
```

As illustrated by this model, most of the socioeconomic factors involved in this study do not have a significant effect on a student's final grade, holding all other factors constant. In this model, extra paid classes, mother's education, and family support had a significant effect on a student's final mathematics grade at a significance level of 0.10. That is, a student who has participated in extra paid classes or a student whose mother has obtained a bachelor's degree or higher, will achieve a higher mathematics grade. Interestingly, a student who has family support will achieve a lower mathematics grade. The predictive power of this model is very weak, with an adjusted R-squared value of 0.0508.

Checking if the Linear Regression Assumptions Have Been Violated Checking if the linearity assumption is violated:

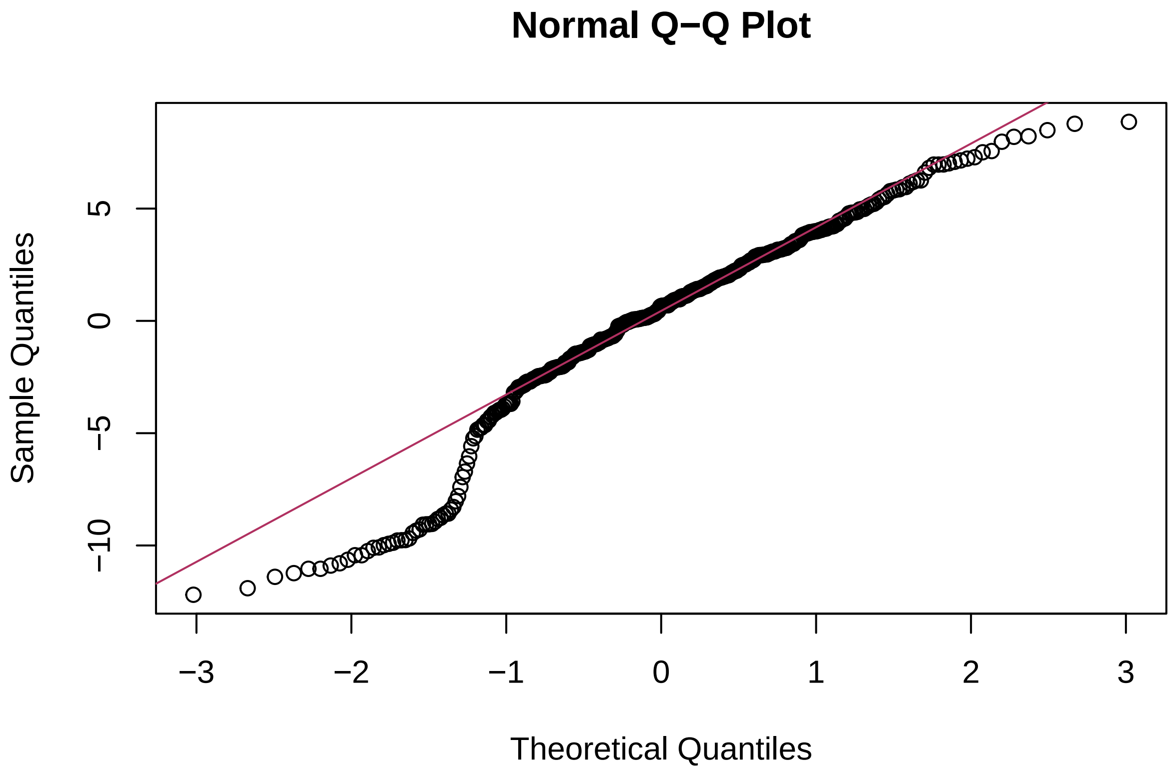
```
#Residual Plot
plot(model$fitted.values, residuals(model), main = "Residual Plot", xlab = "Fitted Values", ylab = "Res
abline(h = 0, col = "springgreen4")
```



Looking at the plot above, the linearity assumption appears to be violated, thus making our linear regression model invalid.

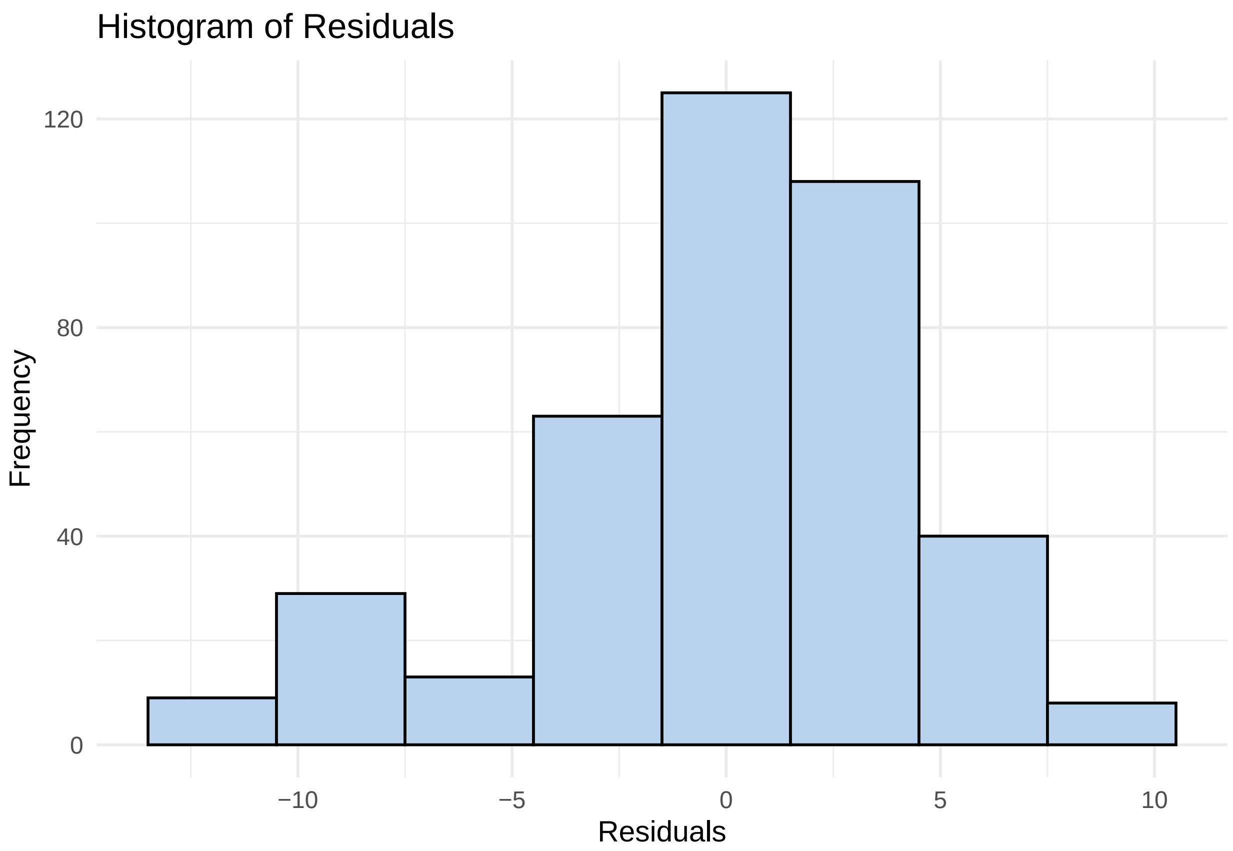
Checking if the normality assumption is violated:

```
#QQ Plot
qqnorm(residuals(model))
qqline(residuals(model), col = "maroon")
```



Based on this plot, it appears that the normality assumption has been violated.

```
#Histogram of residuals
ggplot(data.frame(residuals = residuals(model)), aes(x = residuals)) +
  geom_histogram(binwidth = 3, fill = "slategray2", color = "black") +
  labs(title = "Histogram of Residuals", x = "Residuals", y = "Frequency") +
  theme_minimal()
```



This histogram of the residuals indicates that the residuals are heavily skewed left, and do not follow a normal distribution.

```
#Shapiro test to test for normality of residuals
shapiro.test(residuals(model))
```

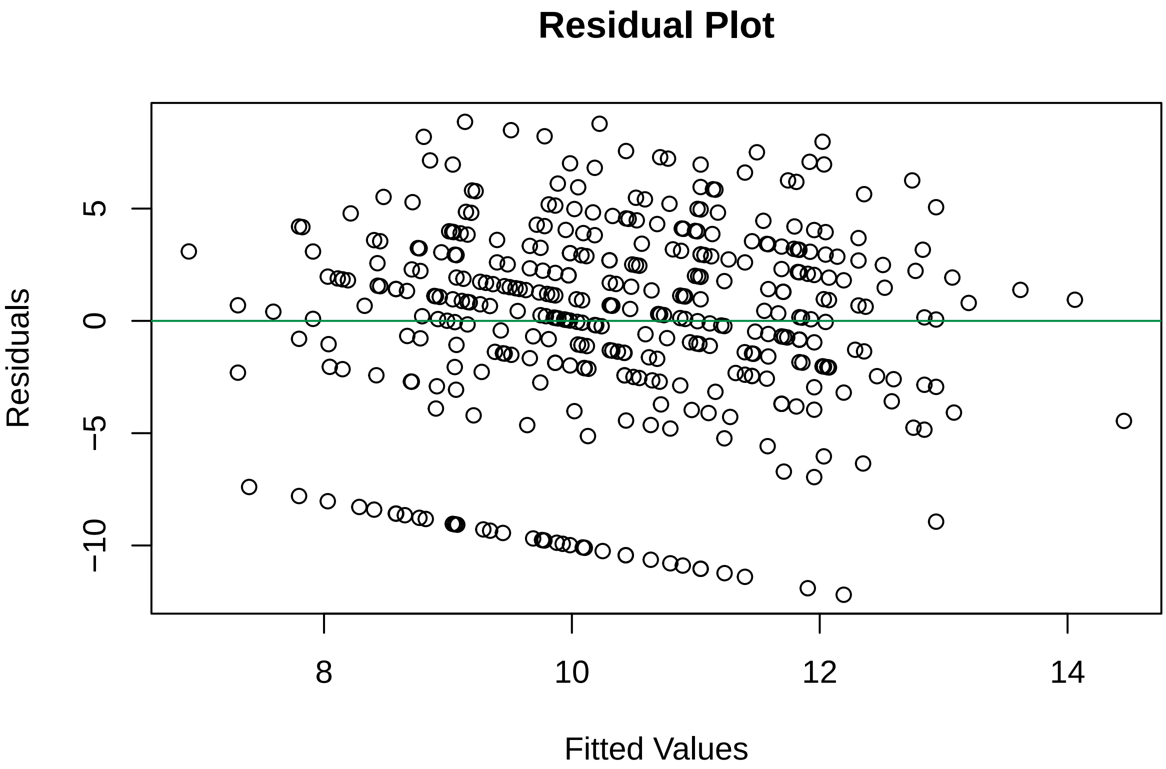
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model)
## W = 0.94986, p-value = 2.555e-10
```

Given a p-value of 2.555e-10, we can reject the null hypothesis and conclude that the residuals are not normally distributed.

Based on these three procedures, it's obvious that the normality assumption has been violated.

Checking if the homoscedasticity (equal variance) assumption has been violated:

```
#Residual plot
plot(model$fitted.values, residuals(model), main = "Residual Plot", xlab = "Fitted Values", ylab = "Res
abline(h = 0, col = "springgreen4")
```



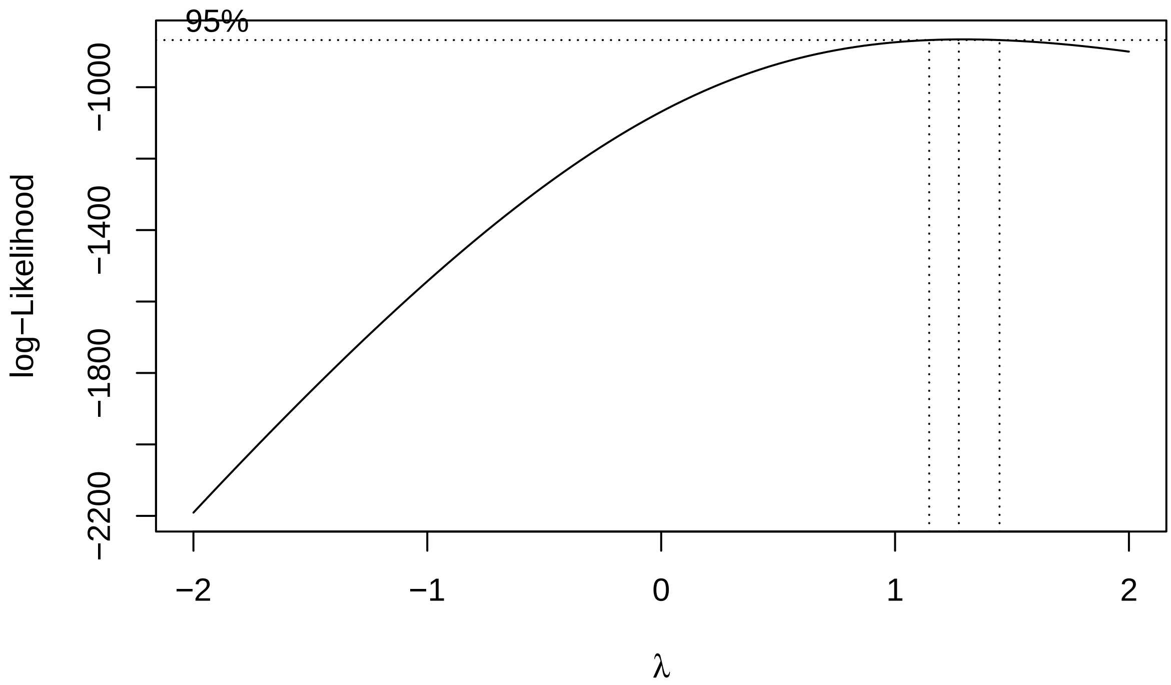
According to this residual plot, the data appears to follow the homoscedasticity/equal variance assumption. However, the other two assumptions have been violated, so this linear regression model is invalid.

Multiple Linear Regression with Box-Cox Transformation on Y

To address the issue of the normality assumption being violated, we will perform a Box-Cox transformation on the response variable, final grade. First, we must shift the values of the response variable by 1, since all of the values of the response variable must be greater than 0 in order to use the Box-Cox transformation.

```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
}
```

```r
#Box Cox Transformation
bc <- boxcox(final_grade_plus1 ~ address_type + family_support + health + internet_access + mother_educa
```



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

Next, we will construct a new multiple linear regression model, with the Box-Cox transformation applied to the response variable.

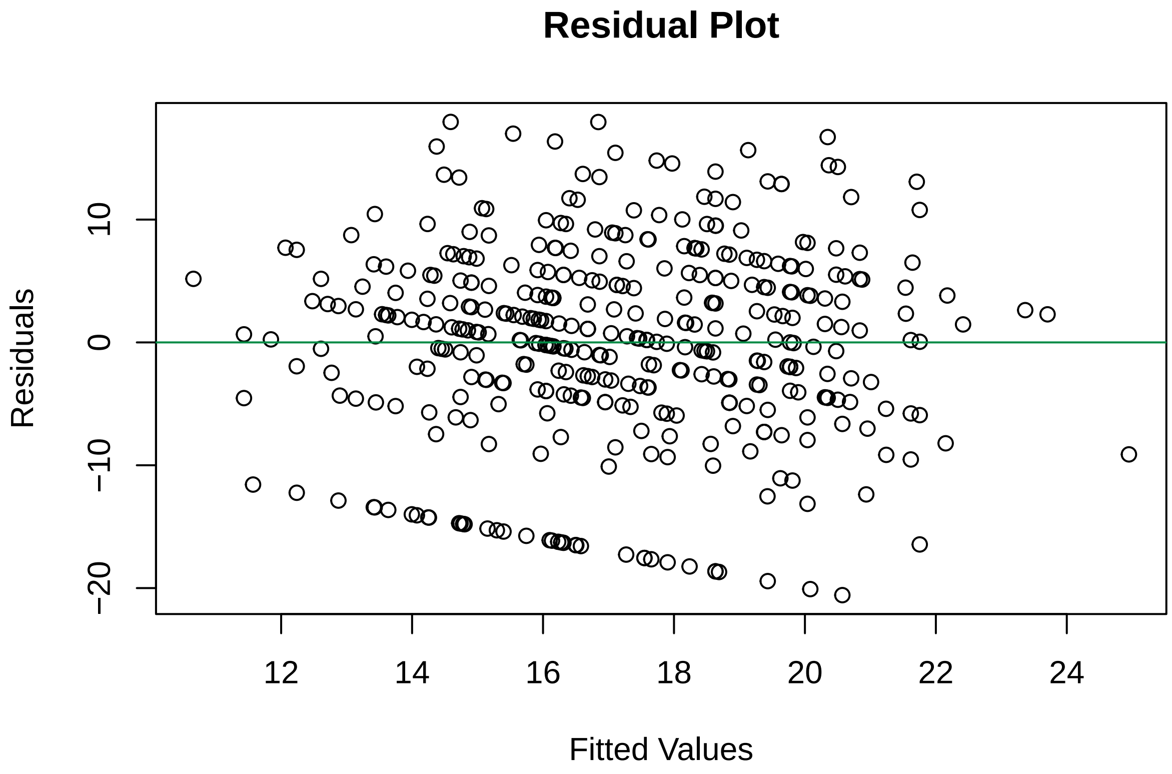
```
#New multiple linear regression model with Box-Cox transformation applied
boxcox_model <- lm(transformed_y ~ address_type + family_support + health + internet_access + mother_edu
summary(boxcox_model)
```

```
##
## Call:
## lm(formula = transformed_y ~ address_type + family_support +
##     health + internet_access + mother_education + father_education +
##     extra_paid_classes + parent_status, data = math_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.5720  -4.4922   0.6578   5.2507  17.9500
##
## Coefficients:
##                                               Estimate Std. Error t value
## (Intercept)                                    17.4679     2.1990   7.943
## address_typeUrban                               1.1768     1.0166   1.158
## family_supportyes                              -1.8439     0.8910  -2.070
## health                                         -0.2661     0.2998  -0.887
## internet_accessyes                              1.0378     1.1576   0.897
## mother_education5th to 9th grade               -1.0588     1.1991  -0.883
## mother_educationhigher education                2.1622     1.1790   1.834
## mother_educationnone                            4.7572     4.7987   0.991
## mother_educationprimary education (4th grade)  -2.5235     1.5087  -1.673
## father_education5th to 9th grade                0.4216     1.1752   0.359
## father_educationhigher education                0.9282     1.2289   0.755
## father_educationnone                            4.3351     5.8250   0.744
## father_educationprimary education (4th grade)  -0.4116     1.3992  -0.294
## extra_paid_classesyes                           1.1397     0.8750   1.303
## parent_statusLiving together                   -0.9646     1.3731  -0.703
##                                               Pr(>|t|)
## (Intercept)                                   2.25e-14 ***
## address_typeUrban                               0.2477
## family_supportyes                               0.0392 *
## health                                          0.3754
## internet_accessyes                              0.3705
## mother_education5th to 9th grade                0.3778
## mother_educationhigher education                0.0674 .
## mother_educationnone                            0.3221
## mother_educationprimary education (4th grade)   0.0952 .
## father_education5th to 9th grade                0.7200
## father_educationhigher education                0.4505
## father_educationnone                            0.4572
## father_educationprimary education (4th grade)   0.7688
## extra_paid_classesyes                           0.1935
## parent_statusLiving together                    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
```

The results of this model are slightly different from the original model. In this model, the independent variables that have a statistically significant relationship are family support (at alpha = 0.05), and mother's education (higher education and primary education) (at alpha = 0.10). That is, a student whose mother has achieved a bachelor's degree or higher is predicted to achieve a higher final grade, while a student whose mother has achieved only a primary education is predicted to achieve a lower mathematics grade. Similar to the original model, a student with family support is predicted to achieve a lower mathematics grade.

Checking if Linear Regression Assumptions Have Been Violated Checking if the linearity assumption is violated:

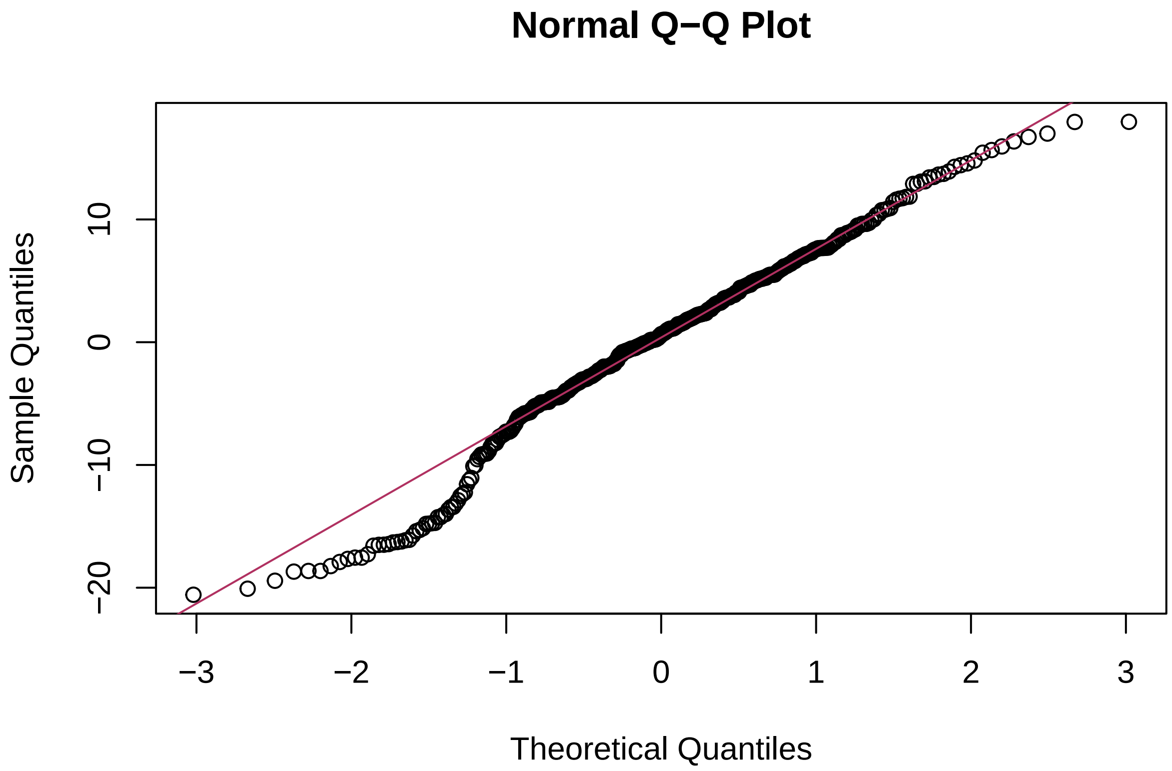
```
#Residual Plot
plot(boxcox_model$fitted.values, residuals(boxcox_model), main = "Residual Plot", xlab = "Fitted Values
abline(h = 0, col = "springgreen4")
```



According to this residual plot, the linearity assumption has been violated, even after applying the Box-Cox transformation.

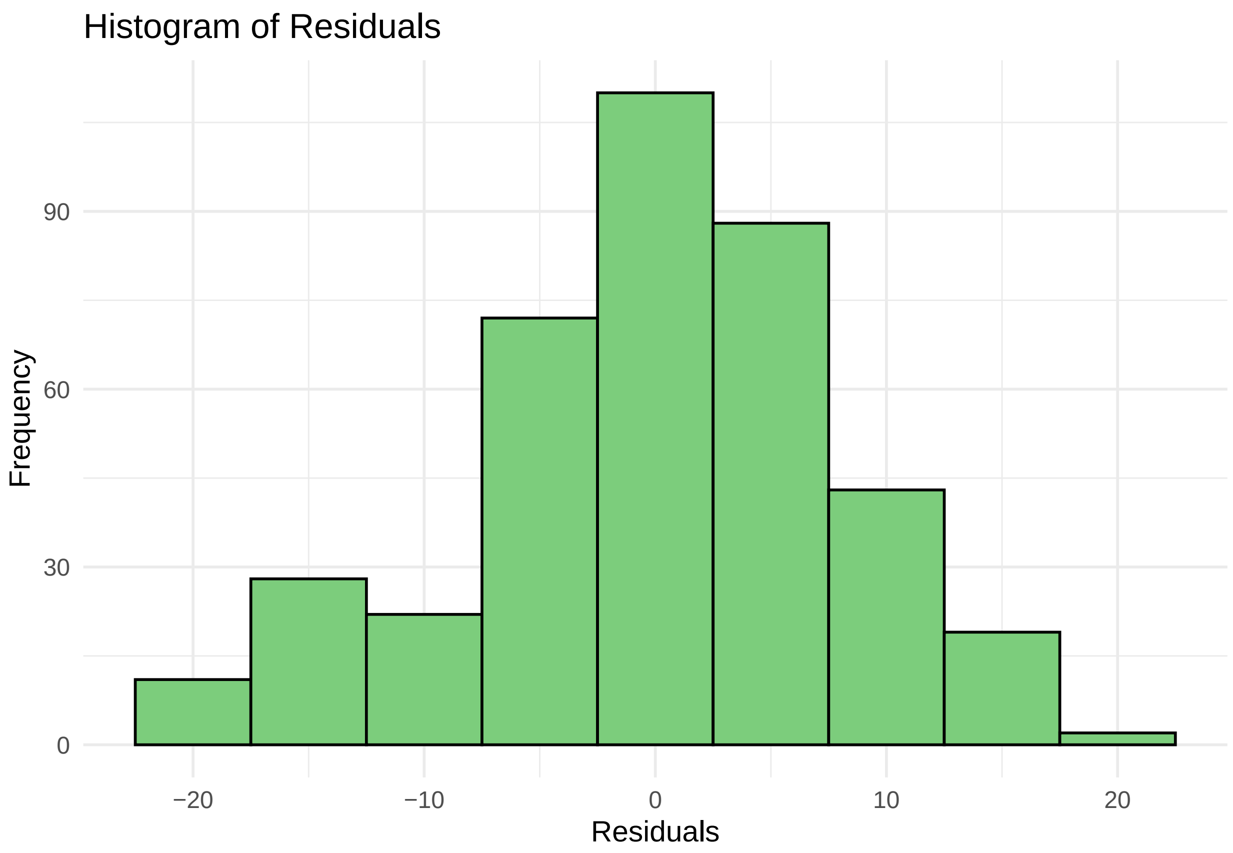
Checking if the normality assumption is violated:

```
#QQ Plot
qqnorm(residuals(boxcox_model))
qqline(residuals(boxcox_model), col = "maroon")
```



Based on this plot, it appears that the normality assumption has been violated.

```
#Histogram of residuals
ggplot(data.frame(residuals = residuals(boxcox_model)), aes(x = residuals)) +
  geom_histogram(binwidth = 5, fill = "palegreen3", color = "black") +
  labs(title = "Histogram of Residuals", x = "Residuals", y = "Frequency") +
  theme_minimal()
```



Based on this histogram, the distribution of the residuals still seems to be slightly skewed.

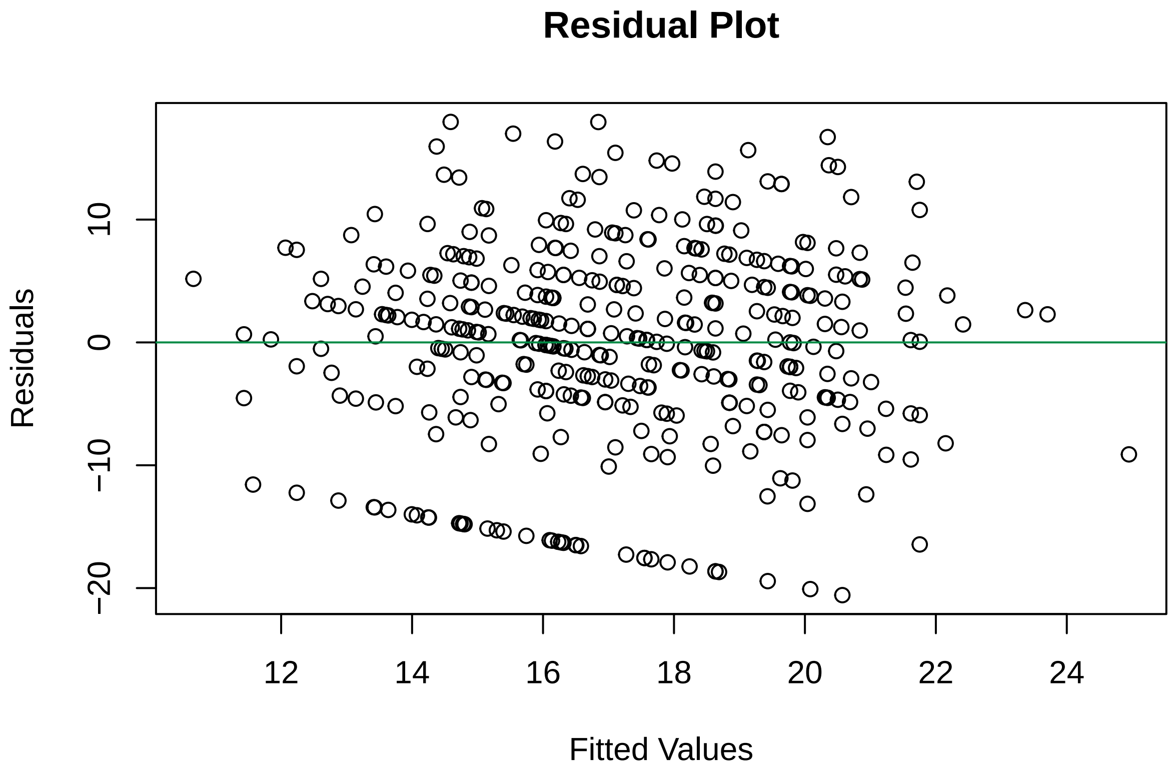
```r
#Shapiro test to test for normality of residuals
shapiro.test(residuals(boxcox_model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(boxcox_model)
## W = 0.97988, p-value = 2.631e-05
```

Given a p-value of 2.631e-05, we reject the null hypothesis and can conclude that the residuals are not normally distributed.

Checking if the homoscedasticity (equal variance) assumption has been violated:

```r
#Residual plot
plot(boxcox_model$fitted.values, residuals(boxcox_model), main = "Residual Plot", xlab = "Fitted Values
abline(h = 0, col = "springgreen4")
```



Based on the residual plot, it appears that the homoscedasticity/equal variance assumption holds true.

However, the other two assumptions have been violated, so this linear regression model is invalid.

Multiple Linear Regression Model, Ignoring Values of Y = 0

One problem with this model is that our dependent variable, final grade, is heavily skewed to the left with many values of '0'. Since a final grade of 0 in a course generally means that a student did not complete the course or had their score voided (due to cheating, absences, etc.), these scores are most likely irrelevant. Thus, we will remove all rows where final grade is equal to 0, and repeat the linear regression model.

```
#Taking subset of the data where final grade is above 0
over0 <- subset(math_data, final_grade > 0)
```

```
#Creating a multiple linear regression model with this new subset of the data
model_no0 <- lm(final_grade ~ address_type + family_support + health + internet_access + mother_educati
summary(model_no0)
```

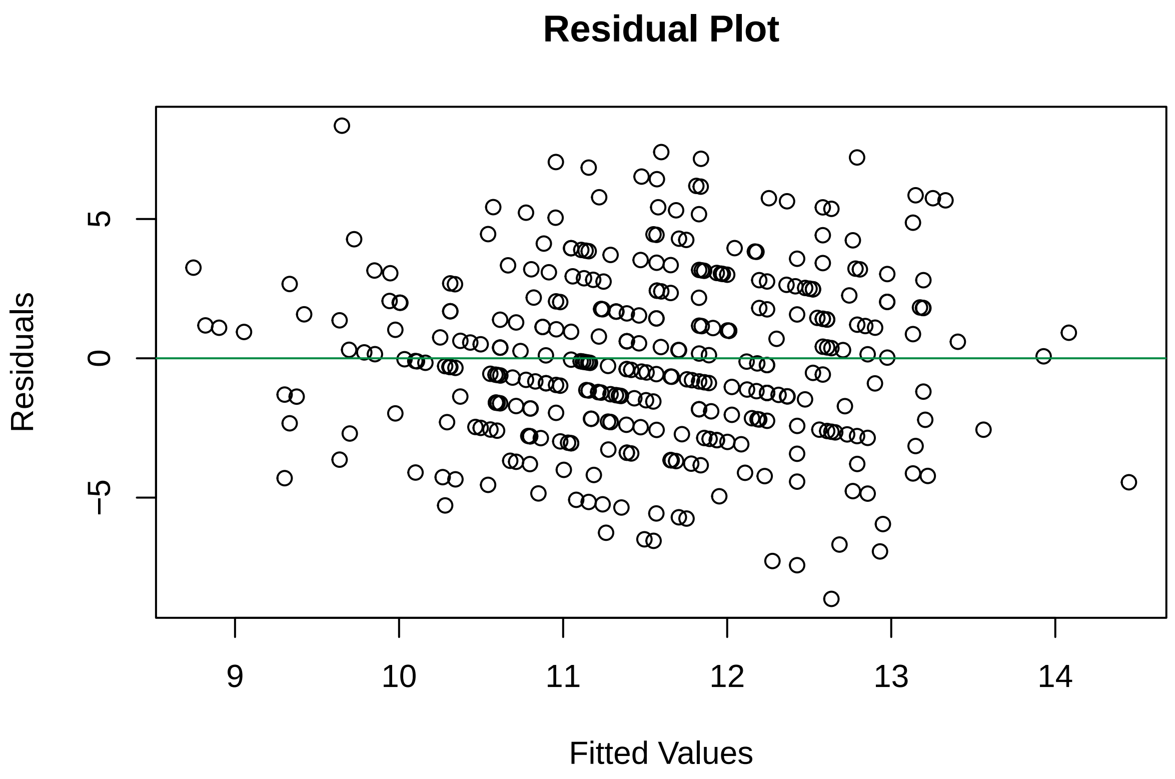
```
##
## Call:
## lm(formula = final_grade ~ address_type + family_support + health +
##     internet_access + mother_education + father_education + extra_paid_classes +
##     parent_status, data = over0)
##
## Residuals:
```

```
##     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
```

With zero values removed, the results slightly differ. The independent variables that have a statistically significant relationship with final grade at the 0.10 significance level are family support and father's education level. That is, a student whose father has obtained a Bachelor's degree or higher will achieve a higher final grade. Similar to the original model, a student with family support will achieve a lower final grade.

Checking if the Linear Regression Assumptions Have Been Violated Checking if the linearity assumption has been violated:

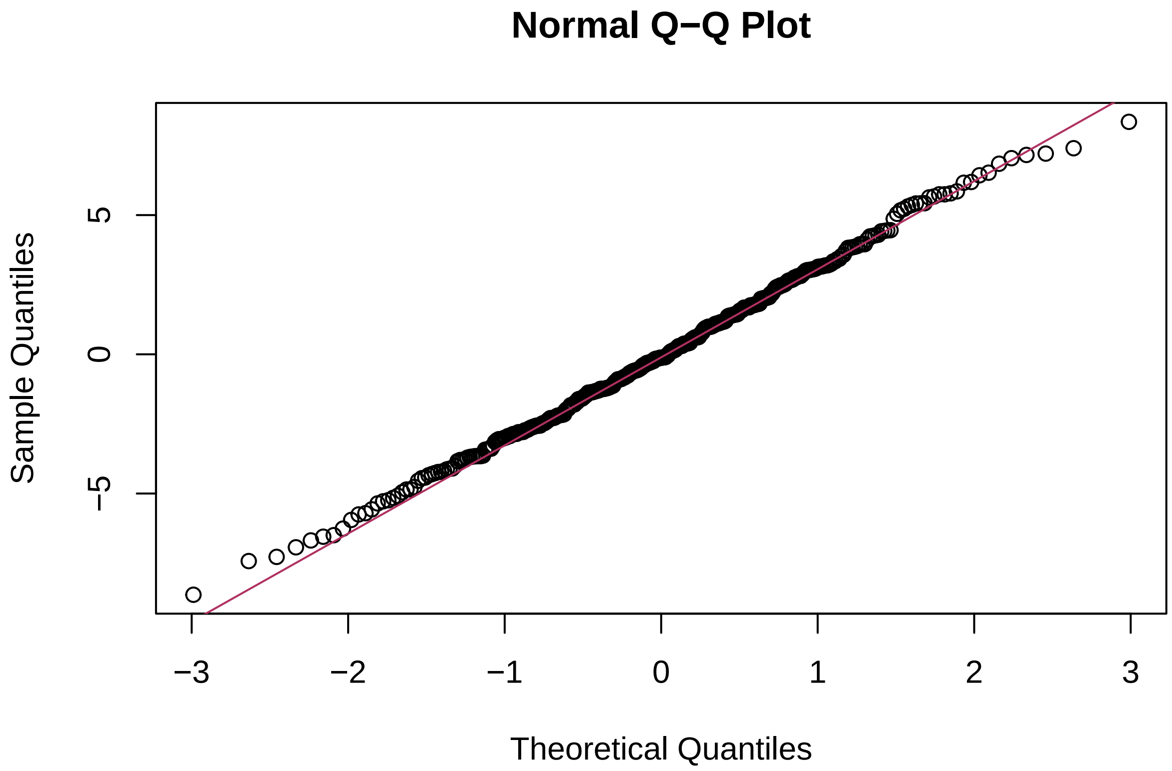
```
#Residual Plot
plot(model_no0$fitted.values, residuals(model_no0), main = "Residual Plot", xlab = "Fitted Values", yla
abline(h = 0, col = "springgreen4")
```



Based on this residual plot, we can assume that the linearity assumption holds true, as the data does not appear to follow any pattern.

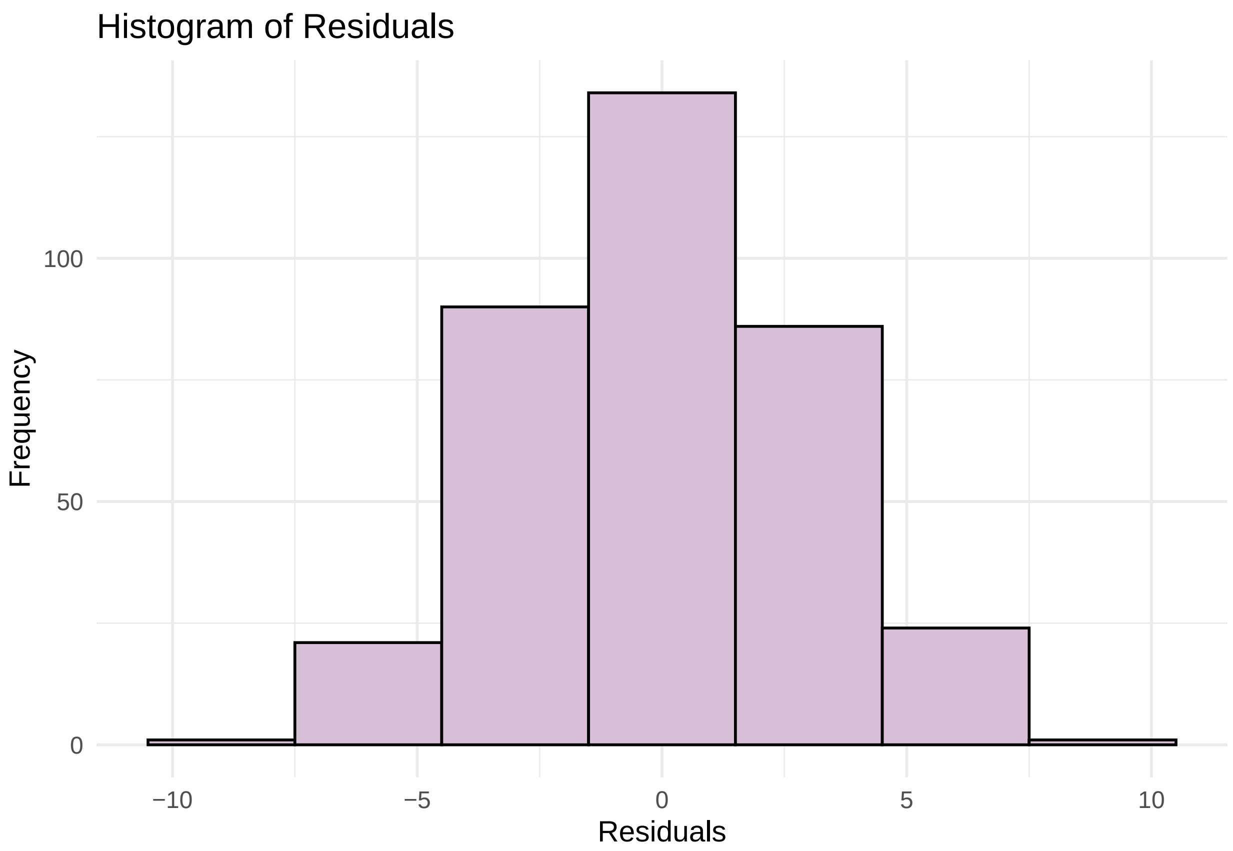
Checking if the normality assumption is violated:

```
#QQ Plot
qqnorm(residuals(model_no0))
qqline(residuals(model_no0), col = "maroon")
```



Based on this plot, it appears that the normality assumption holds true.

```
#Histogram of residuals
ggplot(data.frame(residuals = residuals(model_no0)), aes(x = residuals)) +
  geom_histogram(binwidth = 3, fill = "thistle", color = "black") +
  labs(title = "Histogram of Residuals", x = "Residuals", y = "Frequency") +
  theme_minimal()
```



Based on this histogram, the residuals appear to be normally distributed.

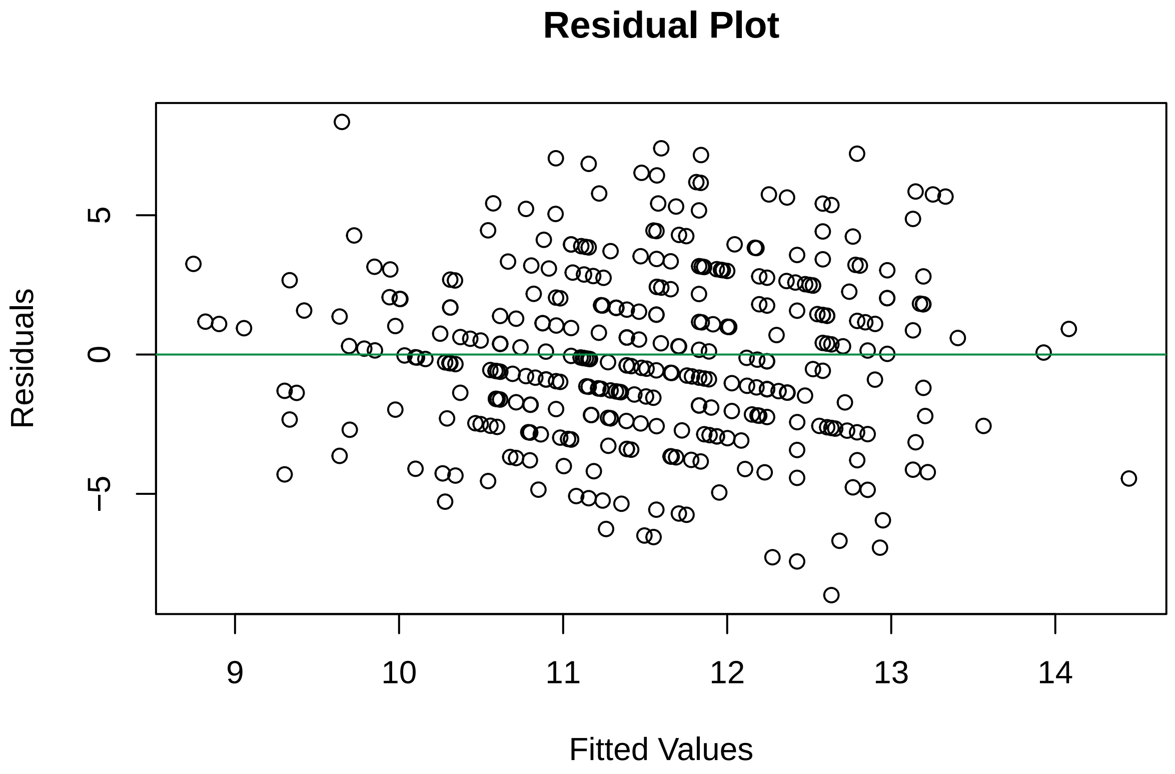
```
#Shapiro test to test for normality of residuals
shapiro.test(residuals(model_no0))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model_no0)
## W = 0.99728, p-value = 0.8223
```

Given a p-value of 0.8223, we fail to reject the null hypothesis and can conclude that the residuals are normally distributed.

Checking if the homoscedasticity (equal variance) assumption has been violated:

```
#Residual plot
plot(model_no0$fitted.values, residuals(model_no0), main = "Residual Plot", xlab = "Fitted Values", ylal
abline(h = 0, col = "springgreen4")
```



Based on the plot of residuals, it appears that the homoscedasticity/equal variance assumption holds true.

All three linear regression assumptions hold true, so we can assume that our results from this model are valid.

Effects of Missing Data on Our Analysis

Checking for missing data first:

```
sum(is.na(math_data)) #check for missing data
```

```
## [1] 0
```

There are no missing values in our data.

Simulating and Dealing With MCAR Data: Missing Values Selected Completely at Random

To deal with MCAR data, we will use listwise deletion, i.e., deleting any rows with missing values. Since we don't have any actual NA values in our dataset, we will simulate a situation where 20% of our data is composed of missing values. To do so, we will select a random subset of 80% of the rows, which is approximately equivalent to assigning an NA value to 20% of the rows and subsequently removing the rows with an NA value.

```
#Selecting a random sample containing 80% of the rows in our dataset
data_MCAR <- math_data[sample(nrow(math_data), ceiling(0.8*nrow(math_data))),]
```

Checking how many rows we have left after selecting a subset, to ensure that 20% of the data has been removed:

```
sum(complete.cases(data_MCAR))
```

```
## [1] 316
```

(316)*(0.2) = 395, so we have correctly removed 20% of the rows.

```


shapiro.test(data_MCAR$final_grade[data_MCAR$romantic_relationship == "no"])  # For group 0
```

Repeating Wilcoxon Rank Sum Tests using MCAR data

```
##
##  Shapiro-Wilk normality test
##
## data:  data_MCAR$final_grade[data_MCAR$romantic_relationship == "no"]
## W = 0.94411, p-value = 2.422e-07
```

```
shapiro.test(data_MCAR$final_grade[data_MCAR$romantic_relationship == "yes"])  # For group 1
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_MCAR$final_grade[data_MCAR$romantic_relationship == "yes"]
## W = 0.89659, p-value = 7.964e-07
```

With our new MCAR data, the data is still not normally distributed.

Test for unequal variance:

```
leveneTest(final_grade ~ factor(romantic_relationship), data = data_MCAR)
```

```
## Levene’s Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  4.8953 0.02765 *
##       314
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The p-value is greater than .05, therefore we cannot conclude that the variances are unequal. We will proceed with Wilcoxon Rank Sum test.

```
#Two sided Wilcoxon Rank Sum test using MCAR data for romantic_relationship
w_test_result_MCAR <- wilcox.test(final_grade ~ romantic_relationship, data = data_MCAR, alternative = 
print(w_test_result_MCAR)
```

```
## 
##  Wilcoxon rank sum test with continuity correction
## 
## data:  final_grade by romantic_relationship
## W = 12488, p-value = 0.0375
## 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 median final grade of students who are in a romantic relationship versus those who are not.

Moving onto the effect of extracurricular activities on final grade:

```
shapiro.test(data_MCAR$final_grade[data_MCAR$activities == "no"])  # For group 0
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  data_MCAR$final_grade[data_MCAR$activities == "no"]
## W = 0.93577, p-value = 1.695e-06
```

```
shapiro.test(data_MCAR$final_grade[data_MCAR$activities == "yes"])  # For group 1
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  data_MCAR$final_grade[data_MCAR$activities == "yes"]
## W = 0.91902, p-value = 8.553e-08
```

We conclude that the data is not normally distributed, due to p-value being less than .05.

Testing equal variance assumption:

```
leveneTest(final_grade ~ factor(activities), data = data_MCAR)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.0812 0.7759
##       314
```

The p-value is very high, so we assume equal variance.

Wilcoxon Rank Sum test to test for difference in medians between students involved in extracurricular acitivites vs students not involved in extracurricular activities:

```
#Two-sided Wilcoxon Rank Sum test using MCAR data for activities
w_test_result_MCAR <- wilcox.test(final_grade ~ activities, data = data_MCAR, alternative = )
print(w_test_result_MCAR)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  final_grade by activities
## W = 11788, p-value = 0.3928
## 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 median final grade of students who are involved in activities versus those who are not.

Repeating Kruskal Wallis tests using MCAR data First, we will repeat the Kruskal Wallis test for weekend alcohol consumption.

We check the assumptions for ANOVA below:

```r
#Fitting an ANOVA model
anova_model_MCAR <- aov(final_grade ~ weekend_alcohol, data = data_MCAR)
```

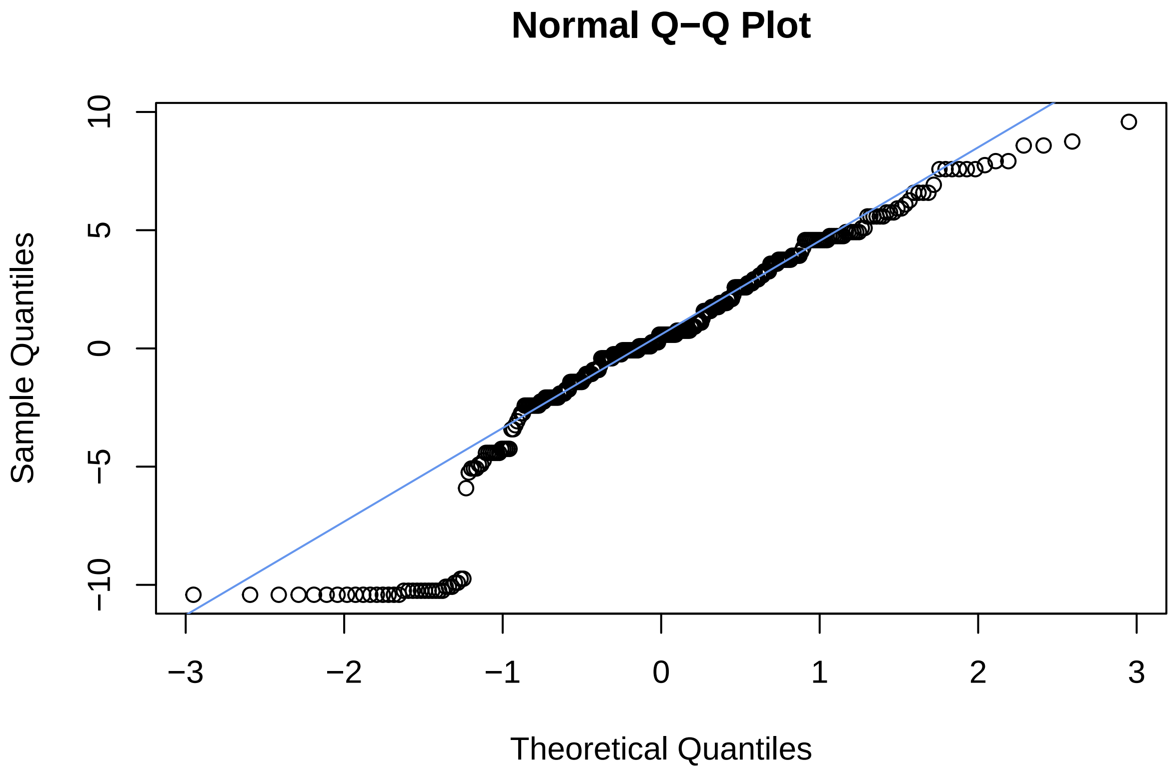
Checking for Normality:

```r
# Checking for normality of residuals
aov_residuals_MCAR <- residuals(anova_model_MCAR)
shapiro.test(aov_residuals_MCAR)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals_MCAR
## W = 0.93004, p-value = 5.029e-11
```

Using a significance level of 0.05, we can reject the null hypothesis and conclude that there is sufficient evidence that the residuals do not follow a normal distribution. Thus, the normality assumption has been violated.

```r
# QQ Plot to check for normality
qqnorm(aov_residuals_MCAR)
qqline(aov_residuals_MCAR, col = "cornflowerblue")
```



We reach a similar conclusion about the normality of the data from the QQ plot. It is clear that the residuals do not follow a normal distribution. Thus, the normality assumption has been violated and we should consider a nonparametric alternative to the ANOVA test, such as the Kruskal Wallis test.

Checking for homogeneity of variances:

```
leveneTest(final_grade ~ factor(weekend_alcohol), data = data_MCAR)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)
## group   4  3.6971 0.005875 **
##       311
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using a significance level of 0.05, we can not reject the null hypothesis and conclude that there is insufficient evidence that the variance is not the same across all groups. Thus, the homogeneity of variances assumption has not been violated.

Since the normality of residuals assumption has been violated, we will use a nonparametric alternative to the ANOVA test, known as the Kruskal Wallis test.

```
kruskal_result <- kruskal.test(final_grade ~ factor(weekend_alcohol), data = data_MCAR)
print(kruskal_result)
```

```
##
```

```
##  Kruskal-Wallis rank sum test
##
## data:  final_grade by factor(weekend_alcohol)
## Kruskal-Wallis chi-squared = 3.7463, df = 4, p-value = 0.4414
```

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 post-hoc analysis.

Next, we will repeat the Kruskal Wallis test for social outing frequency.

```
anova_model_MCAR_2 <- aov(final_grade ~ social, data = data_MCAR)
```

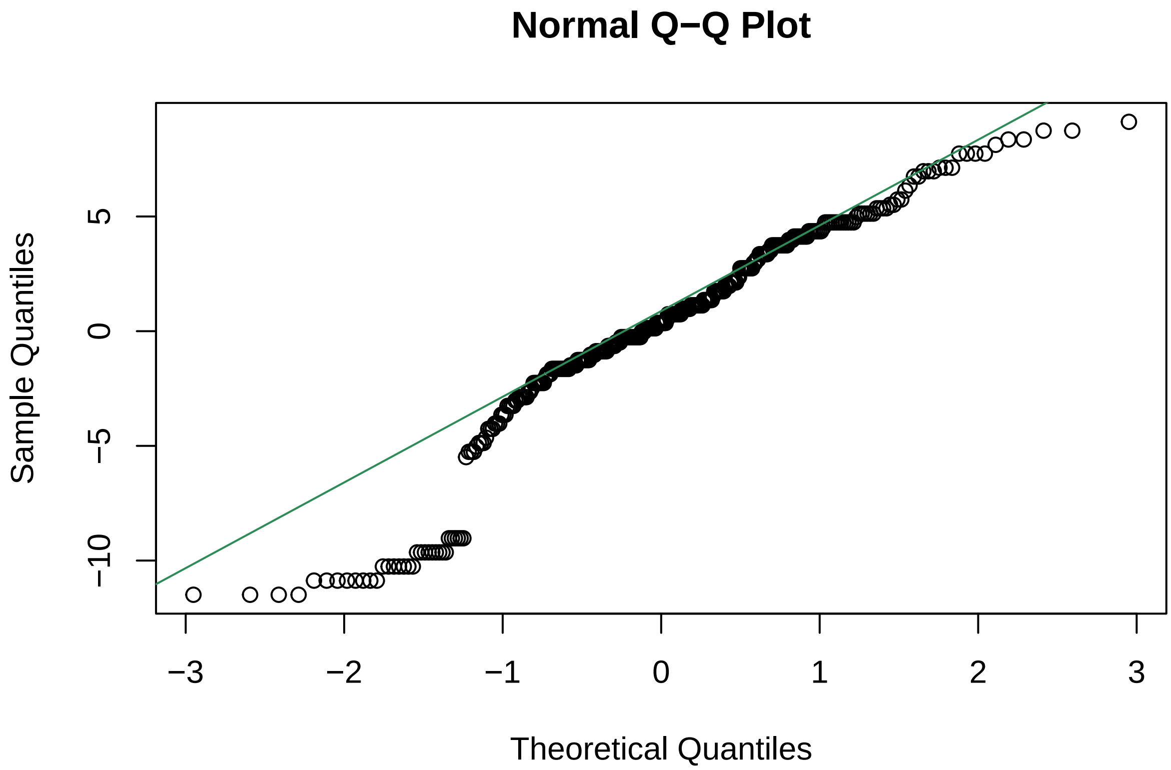
Checking for Normality:

```
# Checking for normality of residuals
aov_residuals_MCAR_2 <- residuals(anova_model_MCAR_2)
shapiro.test(aov_residuals_MCAR_2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals_MCAR_2
## W = 0.93819, p-value = 3.314e-10
```

Using a significance level of 0.05, we can reject the null hypothesis and conclude that there is sufficient evidence that the residuals do not follow a normal distribution. Thus, the normality assumption has been violated.

```
# QQ Plot to check for normality
qqnorm(aov_residuals_MCAR_2)
qqline(aov_residuals_MCAR_2, col = "seagreen")
```



We reach a similar conclusion about the normality of the data from the QQ plot. It is clear that the residuals do not follow a normal distribution. Thus, the normality assumption has been violated and we should consider a nonparametric alternative to the ANOVA test, such as the Kruskal Wallis test.

Checking for homogeneity of variances:

```
leveneTest(final_grade ~ factor(social), data = data_MCAR)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   4  0.2597 0.9036
##       311
```

Using a significance level of 0.05, we cannot reject the null hypothesis and conclude that there is insufficient evidence that the variance is not the same across all groups. Thus, the homogeneity of variances assumption appears to hold true.

Since the normality of residuals assumption for ANOVA has been violated, we will use a nonparametric alternative to the ANOVA test, known as the Kruskal Wallis test.

```
kruskal_result_MCAR_2 <- kruskal.test(final_grade ~ factor(social), data = data_MCAR)
print(kruskal_result_MCAR_2)
```

```
##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  final_grade by factor(social)
## Kruskal-Wallis chi-squared = 15.306, df = 4, p-value = 0.004106
```

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 social outing frequency and final grade. Next, we must conduct a post-hoc analysis to examine the relationship further.

```
anova_model_MCAR_2 <- aov(final_grade ~ factor(social), data = data_MCAR)
tukey_test_MCAR <- TukeyHSD(anova_model_MCAR_2, conf.level=.95)
tukey_test_MCAR
```

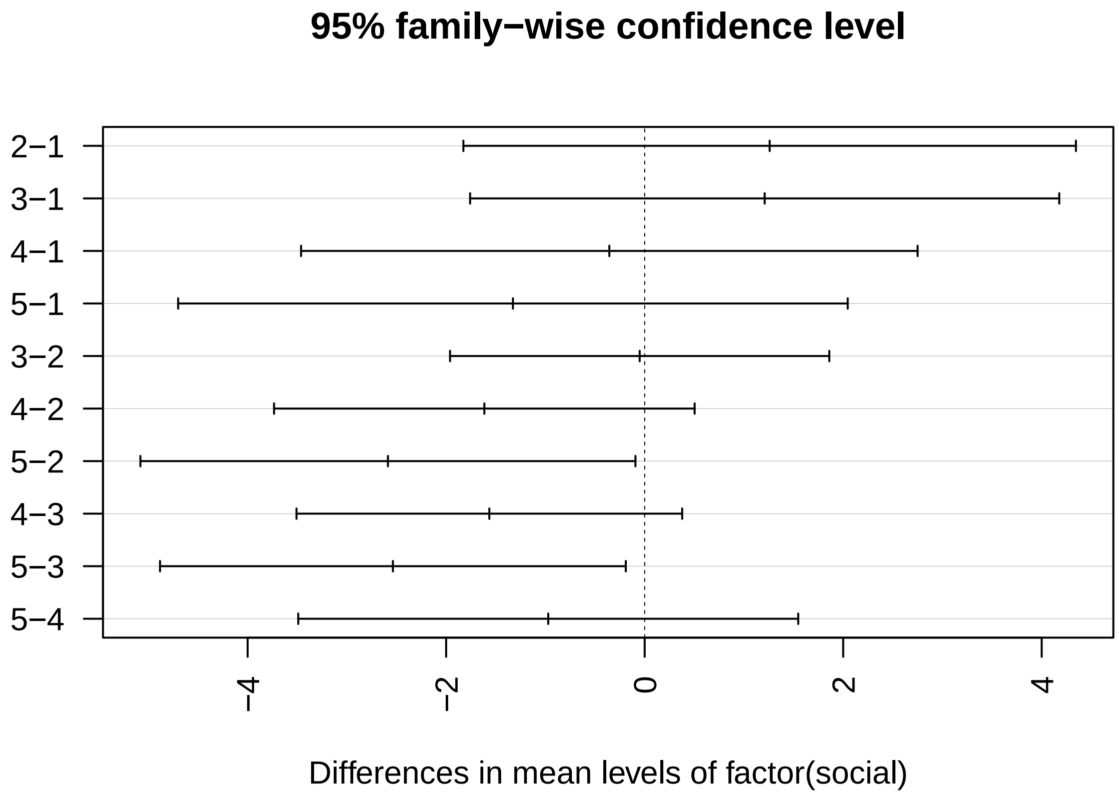
Pairwise Comparisons using Tukey HSD Test

```
##   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  1.25921376 -1.826435  4.34486242 0.7960629
## 3-1  1.20909091 -1.758591  4.17677276 0.7970395
## 4-1 -0.35584416 -3.461631  2.74994244 0.9978740
## 5-1 -1.32727273 -4.700088  2.04554263 0.8168855
## 3-2 -0.05012285 -1.960569  1.86032315 0.9999939
## 4-2 -1.61505792 -3.733685  0.50356870 0.2262491
## 5-2 -2.58648649 -5.080189 -0.09278406 0.0377617
## 4-3 -1.56493506 -3.507739  0.37786875 0.1786151
## 5-3 -2.53636364 -4.882522 -0.19020513 0.0267285
## 5-4 -0.97142857 -3.490006  1.54714931 0.8276312
```

There is a significant difference in medians between group 5 and group 2, and group 5 and group 3 at the 0.05 significance level.

Visualizing Confidence Intervals from Tukey Test:

```
plot(tukey_test_MCAR, las = 2)
```



Based on the plot above, we observe that students who go out with friends “very often” achieve lower final grades than students who go out with friends “not often” or “somewhat often”.

```
#Taking subset of the data where final grade is above 0
over0_MCAR <- subset(data_MCAR, final_grade > 0)
```

```
#Creating a multiple linear regression model with this new subset of the data
model_no0_MCAR <- lm(final_grade ~ address_type + family_support + health + internet_access + mother_edu
summary(model_no0_MCAR)
```

Multiple Linear Regression model (Final model, ignoring values of 0) using MCAR data

```
##
## Call:
## lm(formula = final_grade ~ address_type + family_support + health +
##     internet_access + mother_education + father_education + extra_paid_classes +
##     parent_status, data = over0_MCAR)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4056 -2.1311 -0.0209  2.0874  7.9165
```

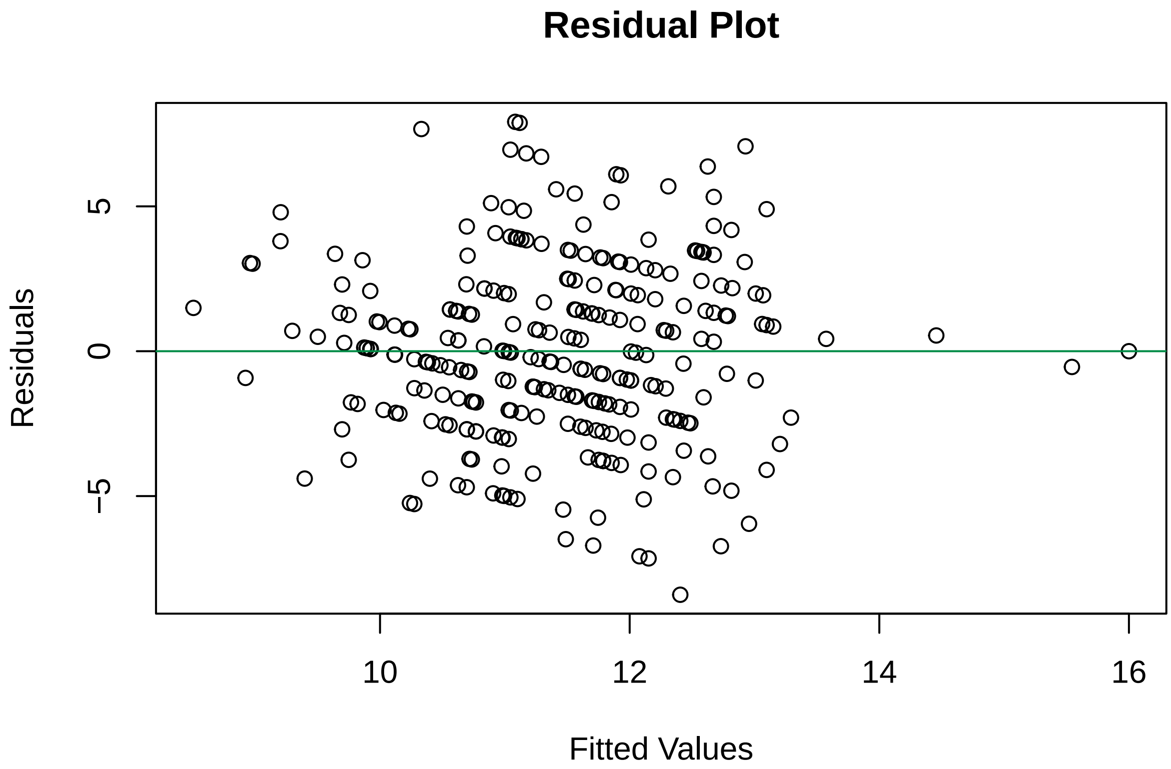
```
##
## Coefficients:
##                                               Estimate Std. Error t value
## (Intercept)                                   10.75477    1.03776  10.363
## address_typeUrban                              1.01636    0.47595   2.135
## family_supportyes                             -0.33602    0.42714  -0.787
## health                                        -0.14109    0.13979  -1.009
## internet_accessyes                             0.03591    0.57107   0.063
## mother_education5th to 9th grade              -0.34910    0.56424  -0.619
## mother_educationhigher education               0.87781    0.54430   1.613
## mother_educationnone                           3.07609    2.32471   1.323
## mother_educationprimary education (4th grade) -1.35228    0.70471  -1.919
## father_education5th to 9th grade               0.41708    0.54050   0.772
## father_educationhigher education               0.50529    0.57431   0.880
## father_educationnone                           6.44271    3.26674   1.972
## father_educationprimary education (4th grade)  0.47399    0.63865   0.742
## extra_paid_classesyes                         -0.66357    0.40333  -1.645
## parent_statusLiving together                   0.52538    0.65104   0.807
##                                               Pr(>|t|)
## (Intercept)                                     <2e-16 ***
## address_typeUrban                               0.0336 *
## family_supportyes                               0.4322
## health                                          0.3137
## internet_accessyes                              0.9499
## mother_education5th to 9th grade                0.5366
## mother_educationhigher education                0.1080
## mother_educationnone                            0.1869
## mother_educationprimary education (4th grade)   0.0561 .
## father_education5th to 9th grade                0.4410
## father_educationhigher education                0.3797
## father_educationnone                            0.0496 *
## father_educationprimary education (4th grade)   0.4586
## extra_paid_classesyes                           0.1011
## parent_statusLiving together                    0.4204
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.175 on 267 degrees of freedom
## Multiple R-squared:  0.1021, Adjusted R-squared:  0.05498
## F-statistic: 2.168 on 14 and 267 DF,  p-value: 0.009279
```

The MCAR data yields slightly different results than the original. The only independent variable that has a statistically significant relationship with final 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.

Checking if the Linear Regression Assumptions Have Been Violated:

Checking if the linearity assumption has been violated:

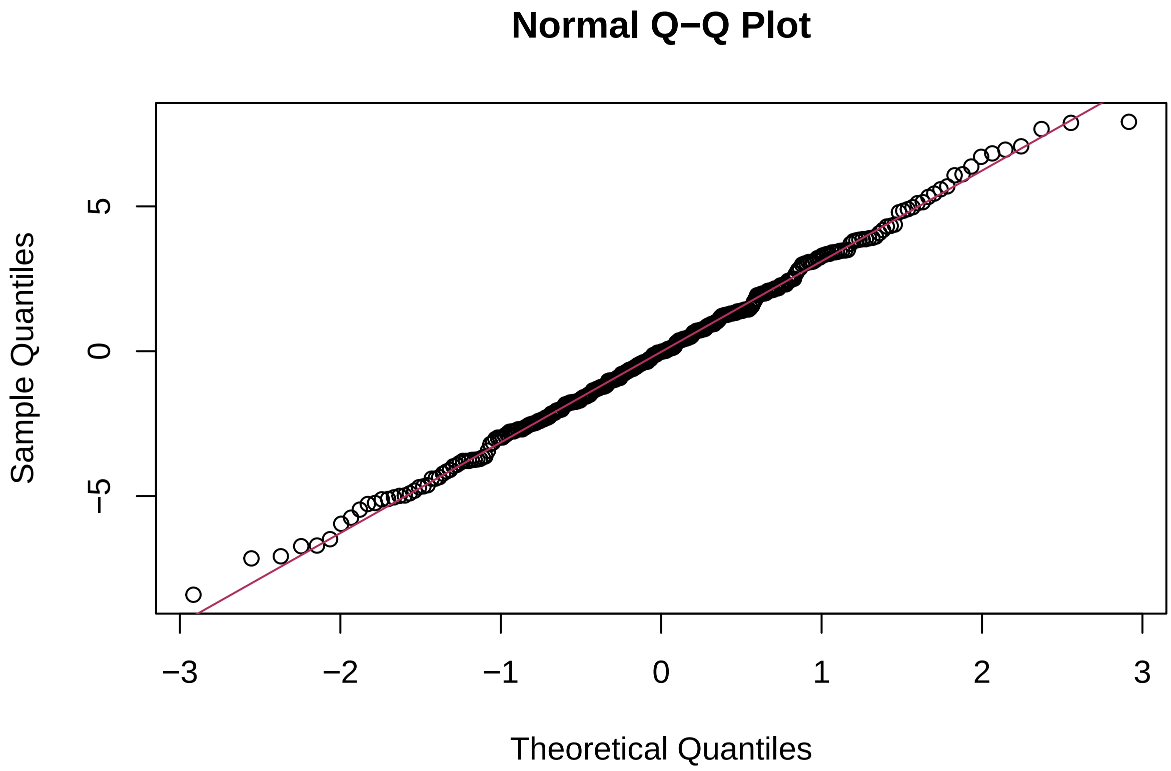
```r
#Residual Plot
plot(model_no0_MCAR$fitted.values, residuals(model_no0_MCAR), main = "Residual Plot", xlab = "Fitted Va
abline(h = 0, col = "springgreen4")
```



Based on this residual plot, we can assume that the linearity assumption holds true, as the data does not appear to follow any pattern.

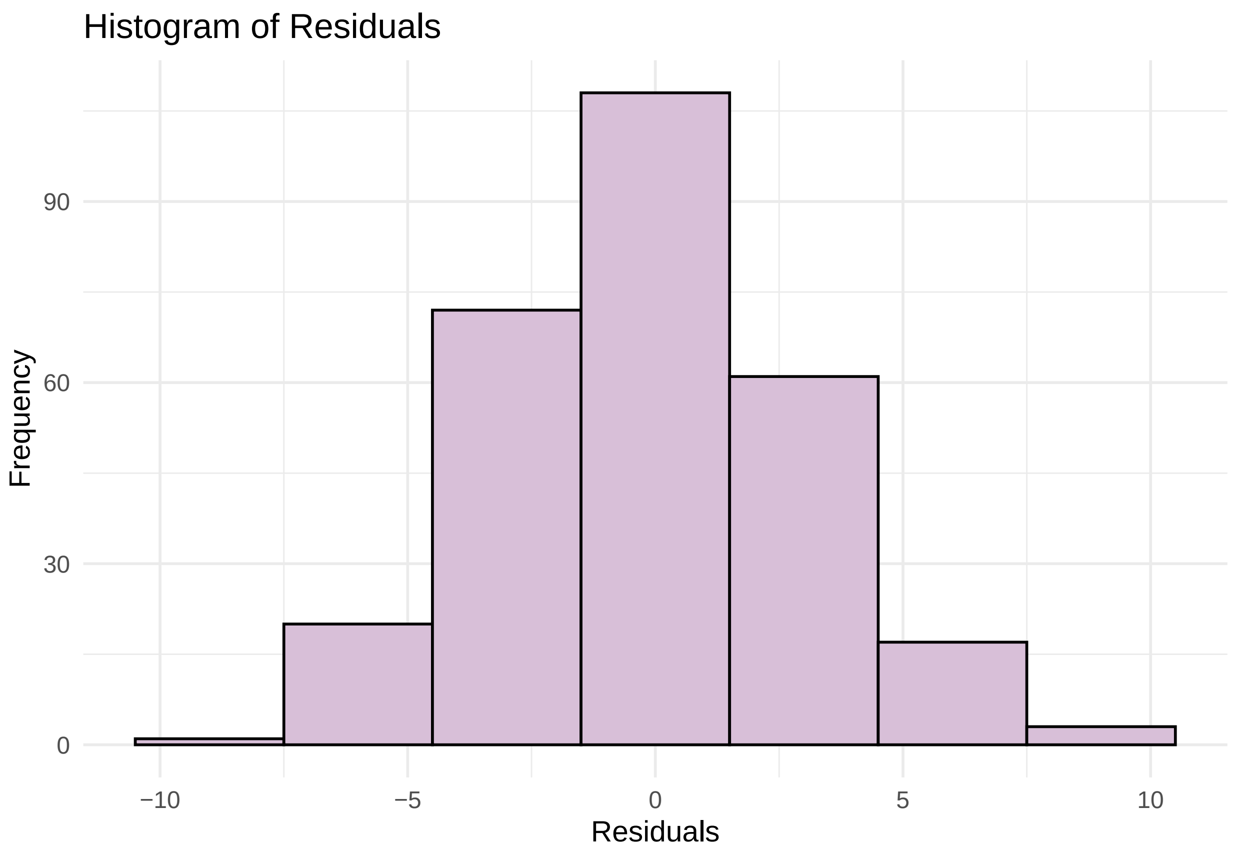
Checking if the normality assumption has been violated:

```
#QQ Plot
qqnorm(residuals(model_no0_MCAR))
qqline(residuals(model_no0_MCAR), col = "maroon")
```



Based on this plot, it appears that the normality assumption holds true.

```
#Histogram of residuals
ggplot(data.frame(residuals = residuals(model_no0_MCAR)), aes(x = residuals)) +
  geom_histogram(binwidth = 3, fill = "thistle", color = "black") +
  labs(title = "Histogram of Residuals", x = "Residuals", y = "Frequency") +
  theme_minimal()
```



Based on this histogram, the residuals appear to be normally distributed.

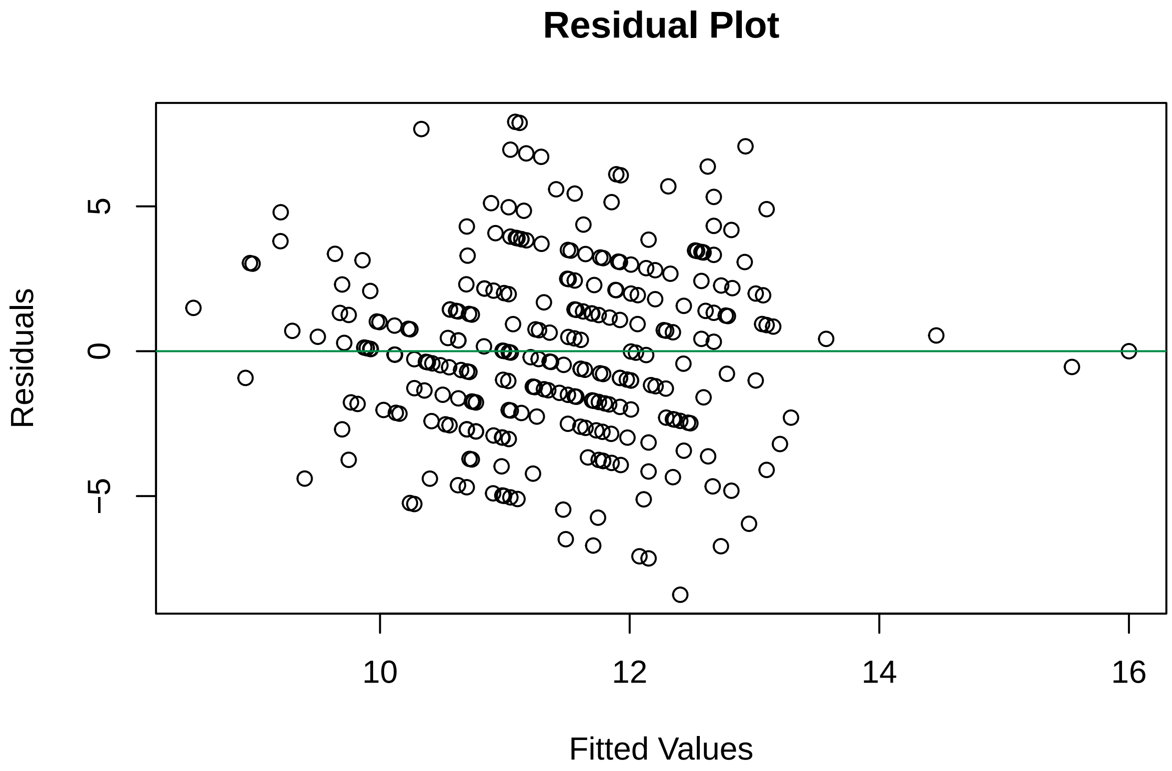
```
#Shapiro test to test for normality of residuals
shapiro.test(residuals(model_no0_MCAR))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model_no0_MCAR)
## W = 0.9968, p-value = 0.8474
```

Given a p-value of 0.8253, we fail to reject the null hypothesis and can conclude that the residuals are normally distributed.

Checking if the homoscedasticity (equal variance) assumption has been violated:

```
#Residual plot
plot(model_no0_MCAR$fitted.values, residuals(model_no0_MCAR), main = "Residual Plot", xlab = "Fitted Val
abline(h = 0, col = "springgreen4")
```



Based on the plot of residuals, it appears that the homoscedasticity/equal variance assumption holds true.

All three linear regression assumptions hold true, thus we can assume that our results from this model are valid.

Simulating and Dealing With MNAR Data: Missing Values Not at Random

To simulate MNAR data, we will use the following scenario: We will select the bottom 20% of final grades and set them to "NA". We do this because students who are likely to achieve a very low grade in the course are more likely to withdraw from the course or not show up for their exams, since they already know that they are going to fail the course, resulting in a missing value for their final course grade.

```
#Creating a copy of the data
data_MNAR <- math_data
```

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

```
#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)
```

To deal with the MNAR data that we have created, we will use the multiple imputation method using the mice package. Although multiple imputation is technically considered to be biased for MNAR data, we will modify it a bit to reduce the bias by adding a "missingness indicator." By including this missingness indicator, we are essentially "telling" the imputation process that there is a pattern to the missing data, so the imputation model 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. While this does not eliminate bias entirely, it will allow the imputation model to impute the missing values more effectively and reduce bias.

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

```
#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
```

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

```
## Warning: Number of logged events: 25
```

```
#Complete data after multiple imputation
complete_data <- complete(imputed_data)
```

If we wanted to address MNAR data more accurately, it may require more advanced techniques like pattern-mixture models or selection models. These models explicitly model the relationship between the missing data and the unobserved values, rather than assuming that missingness can be fully explained by the observed data. However, these models are often very complex and computationally expensive/time consuming. They also tend to be difficult to interpret, so for the sake of our analysis, we will stick with the multiple imputation approach.

Repeating Wilcoxon Rank Sum test using MNAR data: First, we repeat the Wilcoxon Rank Sum tests. Below is the test for the "romantic_relationship" variable.

Checking the assumptions:

Testing for Normality:

```
shapiro.test(data_MNAR$final_grade[data_MNAR$romantic_relationship == "no"])  # For group 0
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_MNAR$final_grade[data_MNAR$romantic_relationship == "no"]
## W = 0.91733, p-value = 3.478e-09
```

```
shapiro.test(data_MNAR$final_grade[data_MNAR$romantic_relationship == "yes"])  # For group 1
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_MNAR$final_grade[data_MNAR$romantic_relationship == "yes"]
## W = 0.94562, p-value = 0.0007805
```

With the MNAR data, the data is still not normally distributed.

Testing for unequal variance:

```
leveneTest(final_grade ~ factor(romantic_relationship), data = data_MNAR)
```

```
## Levene’s Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  6.1159 0.01397 *
##       291
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The p-value is less than .05, therefore we conclude that the variances are unequal. Although the Wilcoxon Rank Sum test generally assumes equal variance, we will proceed with this test and assume equal variance as there are not many other options for us to use since our data is not normally distributed.

```
#Two sided Wilcoxon Rank Sum test for romantic_relationship
w_test_result_MNAR <- wilcox.test(final_grade ~ romantic_relationship, data = data_MNAR, alternative = 
print(w_test_result_MNAR)
```

```
##
##  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 median final grade of students who are in a romantic relationship versus those who are not.

Moving onto the effect of activities on final grade:

Checking the assumptions:

Testing for normality:

```
shapiro.test(data_MNAR$final_grade[data_MNAR$activities == "no"])  # For group 0
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_MNAR$final_grade[data_MNAR$activities == "no"]
## W = 0.92605, p-value = 1.222e-06
```

```
shapiro.test(data_MNAR$final_grade[data_MNAR$activities == "yes"])  # For group 1
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_MNAR$final_grade[data_MNAR$activities == "yes"]
## W = 0.92655, p-value = 4.289e-07
```

We conclude that the data is not normally distributed, due to p-value being less than .05.

Testing for unequal variance:

```
leveneTest(final_grade ~ factor(activities), data = data_MNAR)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.7561 0.3853
##       291
```

The p-value is very high, so we assume equal variance.

```
#Two-sided Wilcoxon Rank Sum test for activities
w_test_result_MNAR <- wilcox.test(final_grade ~ activities, data = data_MNAR, alternative = )
print(w_test_result_MNAR)
```

```
##
##  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 median final grade of students who are involved in activities versus those who are not.

Repeating Kruskal Wallis Tests for MNAR Data First, we repeat the Kruskal Wallis test for weekend alcohol consumption.

Checking the assumptions:

```
#Fitting an ANOVA model using MNAR data
anova_model_MNAR <- aov(final_grade ~ weekend_alcohol, data = data_MNAR)
```

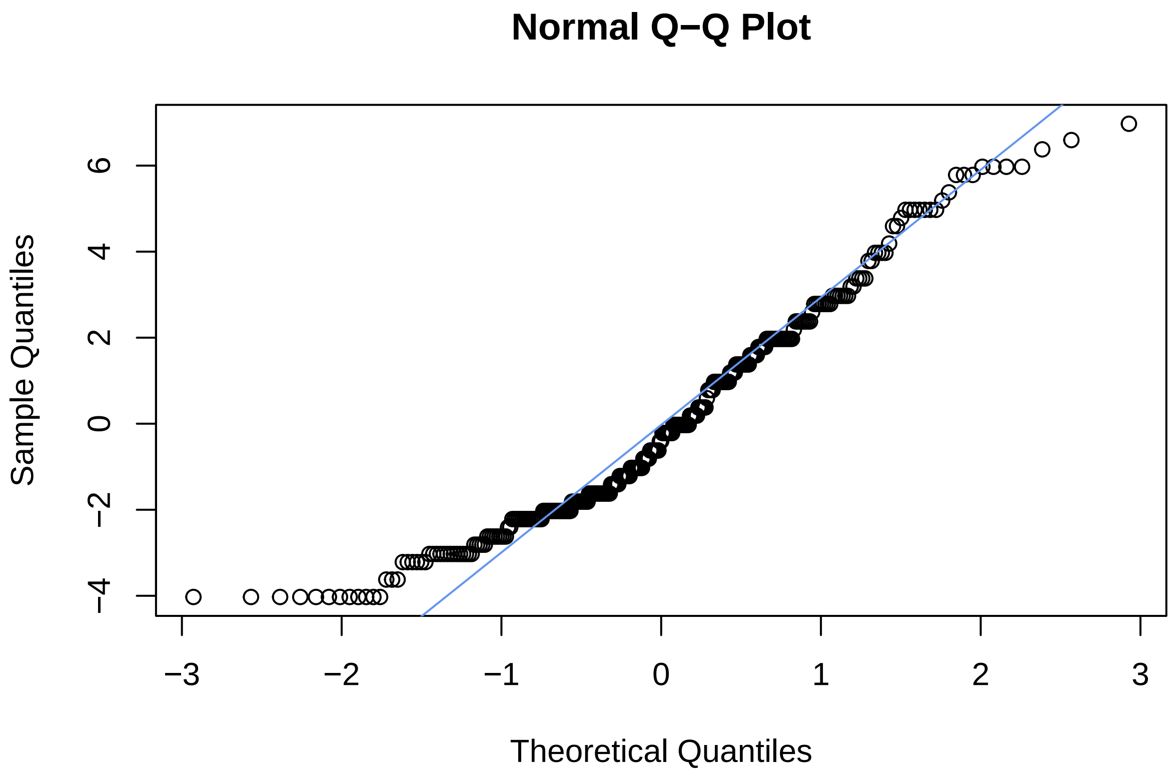
Checking for Normality:

```
# Checking for normality of residuals
aov_residuals_MNAR <- residuals(anova_model_MNAR)
shapiro.test(aov_residuals_MNAR)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals_MNAR
## W = 0.95479, p-value = 7.135e-08
```

Using a significance level of 0.05, we can reject the null hypothesis and conclude that there is sufficient evidence that the residuals do not follow a normal distribution. Thus, the normality assumption has been violated.

```r
# QQ Plot to check for normality
qqnorm(aov_residuals_MNAR)
qqline(aov_residuals_MNAR, col = "cornflowerblue")
```



The residuals do not appear to be normally distributed. For consistency, we will perform the Kruskal Wallis test.

Checking for homogeneity of variances:

```r
leveneTest(final_grade ~ factor(weekend_alcohol), data = data_MNAR)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   4  3.0157 0.01844 *
##       288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using a significance level of 0.05, we can reject the null hypothesis and conclude that there is sufficient evidence that the variance is not the same across all groups. Thus, the homogeneity of variances assumption has been violated.

Since both of the assumptions for ANOVA have been violated, we will use a nonparametric alternative to the ANOVA test, known as the Kruskal Wallis test.

```r
#Kruskal Wallis test for weekend alcohol consumption using MNAR data
kruskal_result <- kruskal.test(final_grade ~ factor(weekend_alcohol), data = data_MNAR)
print(kruskal_result)
```

```
##
##  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 mathematics final grade. Since our results are significant, we will perform post-hoc analysis.

```r


anova_model_MNAR <- aov(final_grade ~ factor(weekend_alcohol), data = data_MNAR)
tukey_test_MNAR <- TukeyHSD(anova_model_MNAR, conf.level=.95)
tukey_test_MNAR
```

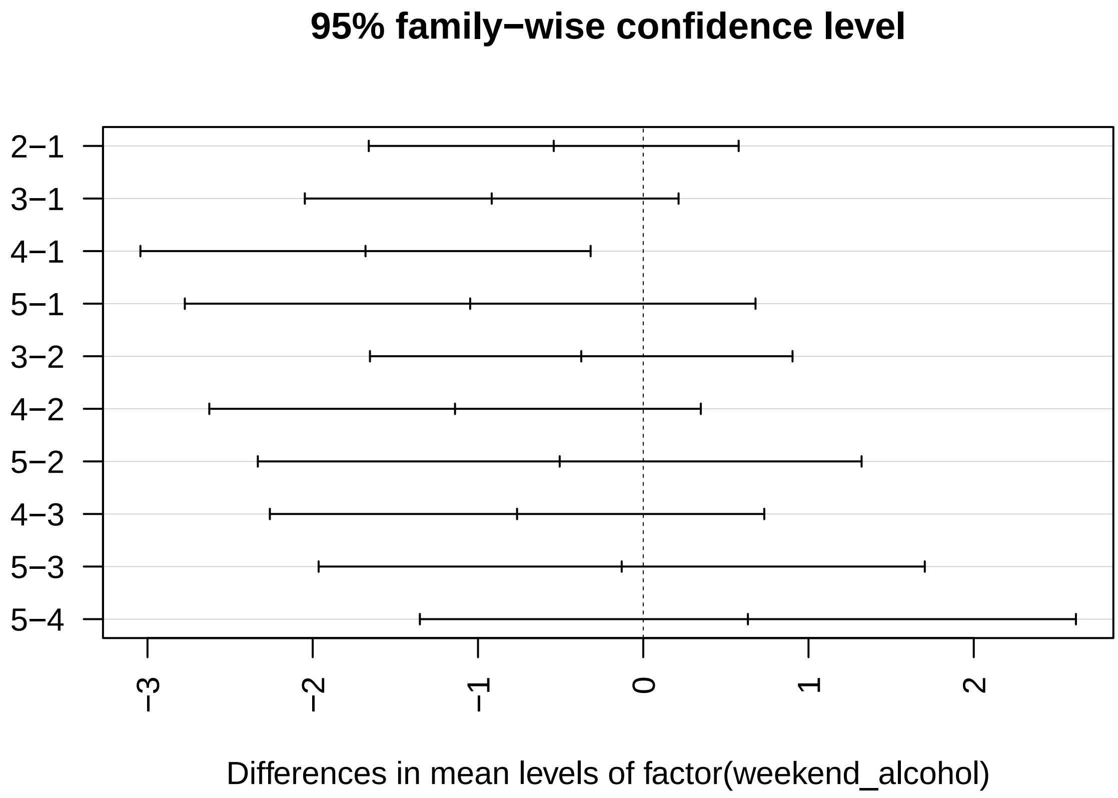
Pairwise Comparisons using Tukey-HSD test

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = final_grade ~ factor(weekend_alcohol), data = data_MNAR)
##
## $`factor(weekend_alcohol)`
##           diff       lwr        upr     p adj
## 2-1 -0.5417896 -1.660944  0.5773651 0.6733896
## 3-1 -0.9170173 -2.047890  0.2138556 0.1730580
## 4-1 -1.6806785 -3.042894 -0.3184625 0.0071503
## 5-1 -1.0473451 -2.774037  0.6793469 0.4572754
## 3-2 -0.3752277 -1.653781  0.9033259 0.9287561
## 4-2 -1.1388889 -2.625986  0.3482083 0.2218520
## 5-2 -0.5055556 -2.332381  1.3212699 0.9418155
## 4-3 -0.7636612 -2.259597  0.7322747 0.6272846
## 5-3 -0.1303279 -1.964356  1.7036998 0.9996755
## 5-4  0.6333333 -1.351715  2.6183815 0.9056015
```

According to the results of the Tukey test, there is a significant difference in test score between weekend_alcohol = 4 and weekend_alcohol = 1 at a significance level of 0.01.

Visualizing Confidence Intervals from Tukey-HSD Test:

```r
plot(tukey_test_MNAR, las = 2)
```



Next, we repeat the Kruskal Wallis test for social outing frequency using MNAR data:

Checking the assumptions:

```
#Fitting ANOVA model using MNAR data
anova_model_MNAR_2 <- aov(final_grade ~ social, data = data_MNAR)
```

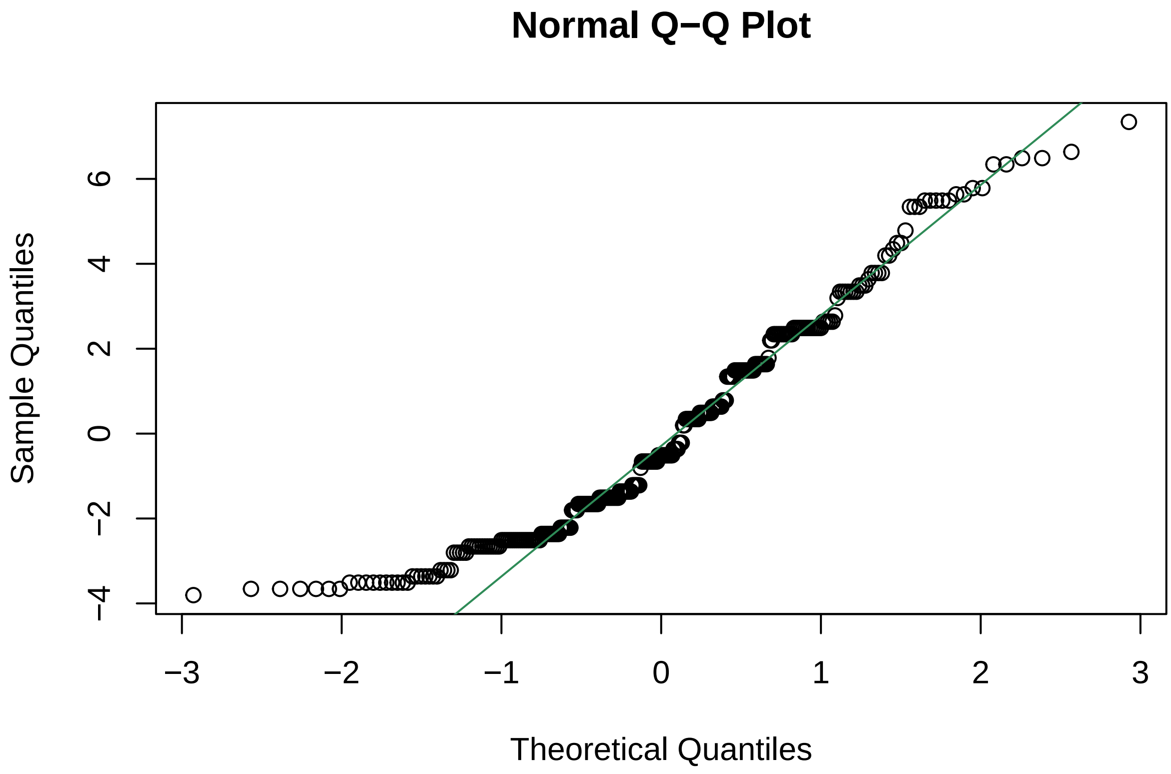
Checking for Normality:

```
# Checking for normality of residuals
aov_residuals_MNAR_2 <- residuals(anova_model_MNAR_2)
shapiro.test(aov_residuals_MNAR_2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals_MNAR_2
## W = 0.93996, p-value = 1.542e-09
```

Using a significance level of 0.05, we can reject the null hypothesis and conclude that there is sufficient evidence that the residuals do not follow a normal distribution. Thus, the normality assumption has been violated.

```
# QQ Plot to check for normality
qqnorm(aov_residuals_MNAR_2)
qqline(aov_residuals_MNAR_2, col = "seagreen")
```



The residuals do not appear to be normally distributed. For consistency, we will perform the Kruskal Wallis test.

Checking for homogeneity of variances:

```
leveneTest(final_grade ~ factor(social), data = data_MNAR)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   4   0.098  0.983
##       288
```

Using a significance level of 0.05, we cannot the null hypothesis and conclude that there is insufficient evidence that the variance is not the same across all groups. Thus, the homogeneity of variances assumption appears to hold true.

Since the normality of residuals assumption for ANOVA has been violated, we will use a nonparametric alternative to the ANOVA test, known as the Kruskal Wallis test.

```
#Kruskal Wallis test for social outing frequency using MNAR data
kruskal_result_MNAR_2 <- kruskal.test(final_grade ~ factor(social), data = data_MNAR)
print(kruskal_result_MNAR_2)
```

```
##
##  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 significant 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 final grade.

```
#Taking subset of the data where final grade is above 0
over0_MNAR <- subset(data_MNAR, final_grade > 0)
```

```
#Creating a multiple linear regression model with this new subset of the data
model_no0_MNAR <- lm(final_grade ~ address_type + family_support + health + internet_access + mother_edu
summary(model_no0_MNAR)
```

Multiple Linear Regression model (final model with 0 values removed) using MNAR data

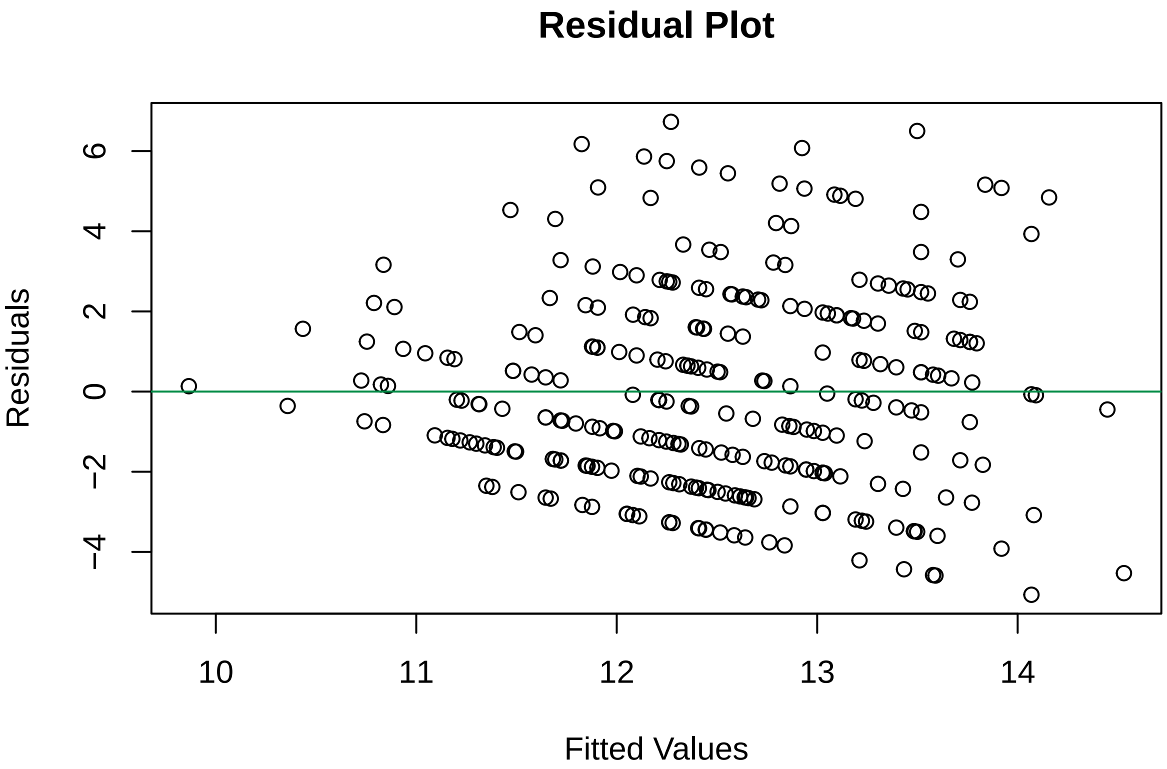
```
##
## Call:
## lm(formula = final_grade ~ address_type + family_support + health +
##     internet_access + mother_education + father_education + extra_paid_classes +
##     parent_status, data = over0_MNAR)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0688 -1.9456 -0.2484  1.7668  6.7294
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                   12.4339     0.7929  15.681
## address_typeUrban                              0.2909     0.3919   0.742
## family_supportyes                             -0.1954     0.3291  -0.594
## health                                        -0.1835     0.1096  -1.674
## internet_accessyes                             0.4653     0.4416   1.054
## mother_education5th to 9th grade              -0.5295     0.4516  -1.172
## mother_educationhigher education               0.6166     0.4296   1.435
## mother_educationnone                           0.2762     1.5368   0.180
## mother_educationprimary education (4th grade) -1.3907     0.5881  -2.365
## father_education5th to 9th grade               0.6042     0.4293   1.407
## father_educationhigher education               0.7655     0.4422   1.731
## father_educationnone                           0.8480     1.8593   0.456
## father_educationprimary education (4th grade)  0.5690     0.5448   1.044
## extra_paid_classesyes                         -0.4908     0.3208  -1.530
## parent_statusLiving together                   0.0590     0.4909   0.120
##                                              Pr(>|t|)
## (Intercept)                                    <2e-16 ***
## address_typeUrban                              0.4585
## family_supportyes                              0.5531
## health                                         0.0953 .
## internet_accessyes                             0.2930
## mother_education5th to 9th grade                0.2420
## mother_educationhigher education                0.1523
## mother_educationnone                            0.8575
## mother_educationprimary education (4th grade)   0.0187 *
## father_education5th to 9th grade                0.1604
## father_educationhigher education                0.0845 .
## father_educationnone                            0.6487
## father_educationprimary education (4th grade)   0.2972
## extra_paid_classesyes                           0.1272
## parent_statusLiving together                    0.9044
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.562 on 278 degrees of freedom
## Multiple R-squared:  0.1007, Adjusted R-squared:  0.05539
## F-statistic: 2.223 on 14 and 278 DF,  p-value: 0.007328
```

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 is mother's education level (primary), father's education level (higher education), and health.

Checking the linear regression assumptions:

Checking if the linearity assumption has been violated:

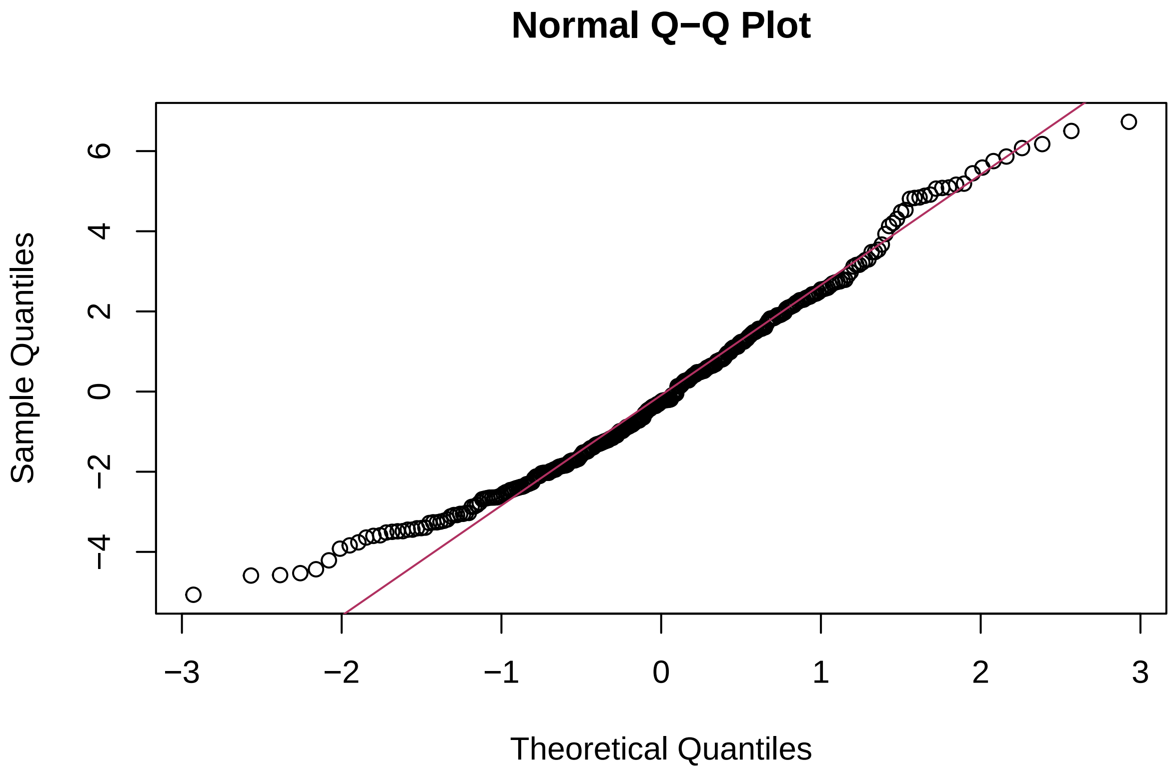
```
#Residual Plot
plot(model_no0_MNAR$fitted.values, residuals(model_no0_MNAR), main = "Residual Plot", xlab = "Fitted Va
abline(h = 0, col = "springgreen4")
```



Based on this residual plot, we can assume that the linearity assumption holds true, as the data does not appear to follow any pattern.

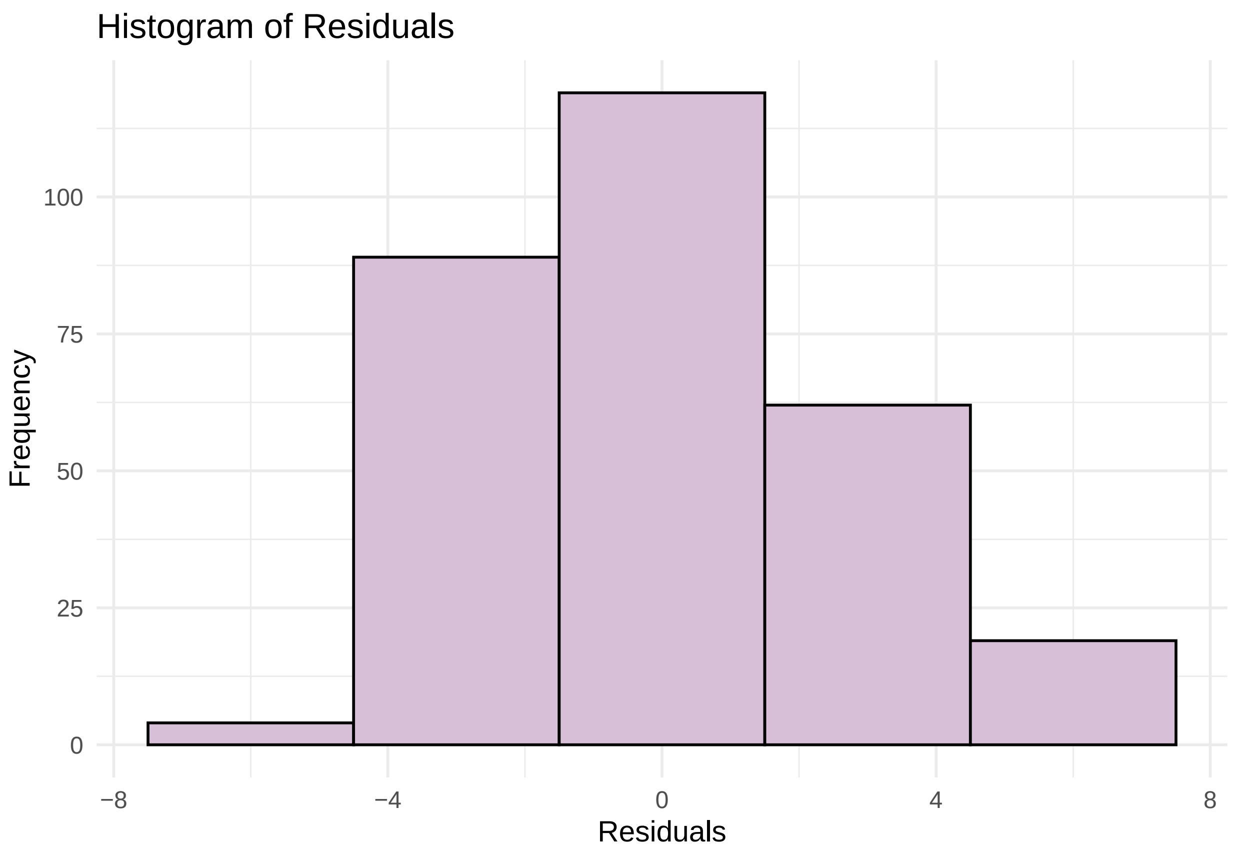
Checking if the normality assumption has been violated:

```
#QQ Plot
qqnorm(residuals(model_no0_MNAR))
qqline(residuals(model_no0_MNAR), col = "maroon")
```



Based on this plot, it appears that the normality assumption does not hold true.

```
#Histogram of residuals
ggplot(data.frame(residuals = residuals(model_no0_MNAR)), aes(x = residuals)) +
  geom_histogram(binwidth = 3, fill = "thistle", color = "black") +
  labs(title = "Histogram of Residuals", x = "Residuals", y = "Frequency") +
  theme_minimal()
```



Based on this histogram, the residuals do not appear to be normally distributed.

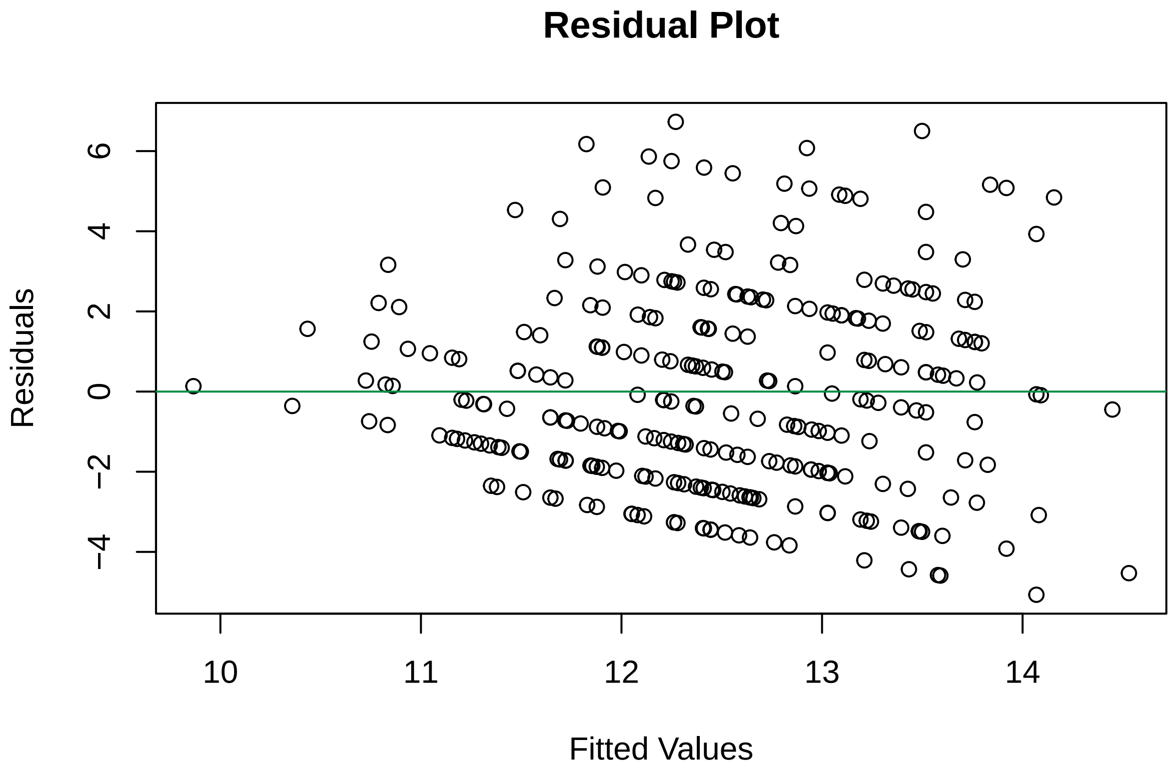
```
#Shapiro test to test for normality of residuals
shapiro.test(residuals(model_no0_MNAR))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model_no0_MNAR)
## W = 0.97555, p-value = 6.649e-05
```

Given that the p-value is less than .05, we can conclude that the residuals are normally distributed.

Checking if the homoscedasticity (equal variance) assumption has been violated:

```
#Residual plot
plot(model_no0_MNAR$fitted.values, residuals(model_no0_MNAR), main = "Residual Plot", xlab = "Fitted Va
abline(h = 0, col = "springgreen4")
```



Based on the plot of residuals, it appears that the variance is increasing.

All three linear regression assumptions do not hold true, thus we cannot assume that our results from this model are valid.