Stats 4714: Final Project

Alexander Nassif

Introduction:

A recent issue in politics, whether federal or local, has been the issue of education. Currently, our educational system is failing in many regards. A large reason behind this is the inability of school systems to address students' underlying problems. All students are different. Some students grow up in certain economic conditions that make it difficult to maintain their educational needs. This can come through an inability for their families to provide food on the table, dysfunctional family dynamics, or an inability to meet the child's specific hurdles and challenges. Whatever the case, these are societal influences that hurt children's education and are a major challenge for modern educators.

The schooling system is currently understaffed. There are too many students for every teacher, and current teachers are already overworked. This is why this project aims to measure the statistical significance of personal and familial conditions on a student's academic performance in school. The overall goal of finding relationships between conditions like the gender of a student, their parents' education level, and whether or not they're in a romantic relationship is to point out groups of students who might be more susceptible to falling through the cracks of the education system. The overall goal of this is to be able to tell teachers which students to specifically look out for in order to offer more assistance or aid to them.

Dataset:

The data used in this project comes from the UCI Machine Learning Repository - Student Performance Dataset. This dataset contains detailed information on secondary school student performance from Portugal. Included in the students' academic performance are a bunch of characterizing variables that include their academic performance, family background, personal characteristics, and living conditions. The academic performance includes the student's study time, number of absences, and final grades. The family background characteristics go over a student's parental education levels, parental jobs, family size, marital status of the parents, and whether or not their parents are currently married. The personal conditions go over a student's gender, age, romantic relationships, and the amount of alcohol a student consumes.

This data also gives data for two grading periods and an overall grading period. For this project, in order to limit the scope to be more manageable, only the final grading period will be looked at. This grading period provides an overall look at the grade of a student, encompassing both the first-period grade and the second-period grade. In addition, there are over 33 characteristics that the dataset outlines for each student. In order to provide a deeper understanding of the overall effects on a student's success, some characteristics will be disregarded. This will allow a deeper focus to be spent on the characteristics that don't get disregarded.

To determine which characteristics should be disregarded, blanket assumptions were made about a lot of the characteristics. For example, it is pretty intuitive that a student who has a lot of absences won't perform well in school. For these reasons, it doesn't seem valuable to look at the effect on the number of absences a student has and their academic performance in school. Similar assumptions can be made for if a student regularly drinks alcohol, wants to go to university, and many other variables. Some variables also aren't valuable or relevant to the overall discussion. For example, one of the variables regards which specific school in Portugal the student being surveyed went to. For the sake of this study, it is being

assumed that both schools are the same, and we are pulling students from both schools and assuming they are the same.

This is done to focus on several key variables. The variables that were chosen for this study are the parents' cohabitation status, their family size, the sex of the student, whether or not the student is in a romantic relationship, whether or not the student participates in extracurricular activities, how often the student goes out, and whether or not the student has internet access. These provide a varied picture of a student that goes over economic status, familial status, and personal characteristics in order to provide a fuller picture of the effects that cause a student to either fail or succeed in school.

Methods:

In order to display an understanding of the statistical concepts taught in this class, multiple statistical concepts that were taught throughout the semester need to be displayed in this project. Three were chosen overall.

The first method involved plotting the grade distributions of the final grades to obtain an overall picture of how the class performed. Then, the mean of the final grades was calculated. The grades were judged numerically from 1-20. The mean was calculated from a sample of every student. Then, conditional probabilities were calculated based on each of the variables. Conditional probabilities were calculated by assuming that a given variable was true and then comparing the probability that the student scored above the mean. For example, given that a student was a man, what was the probability that they exceeded the class average? The average was used as a benchmark to determine class performance. If a student performed above the average, they were determined to be performing well. To complement that, if a student performed below the average, they were determined to be performing poorly.

The second method involved generating bootstrap probabilities of final grades for the different demographics of the variables chosen above. The variables that seemed to have the most effect on academic performance from the first method were chosen for this step. As will be explained below, this meant that bootstrapping occurred for the frequency of going out and also whether or not the student had internet access. Then, a CLT behavior was shown for the sample means. The bootstraps helped to display how that specific variable was behaving.

The third method involved running a linear regression model to compare how likely the variable actually had an impact on academic performance. The way I did this was similar to HW5. To do this, I compared the grade performance with the frequency of a student going out and whether or not they have internet access. First, I did a least squares estimate of the data. I then performed a randomization test for the slope in order to determine the p-value. I then did a model comparison with anova and found the R^2 value for each data. I then plotted the Residual vs Fitted, Q-Q Residuals, Scale-Location, and Residuals vs Leverage graphs for the going out characteristic in order to determine its overall effect on grade performance.

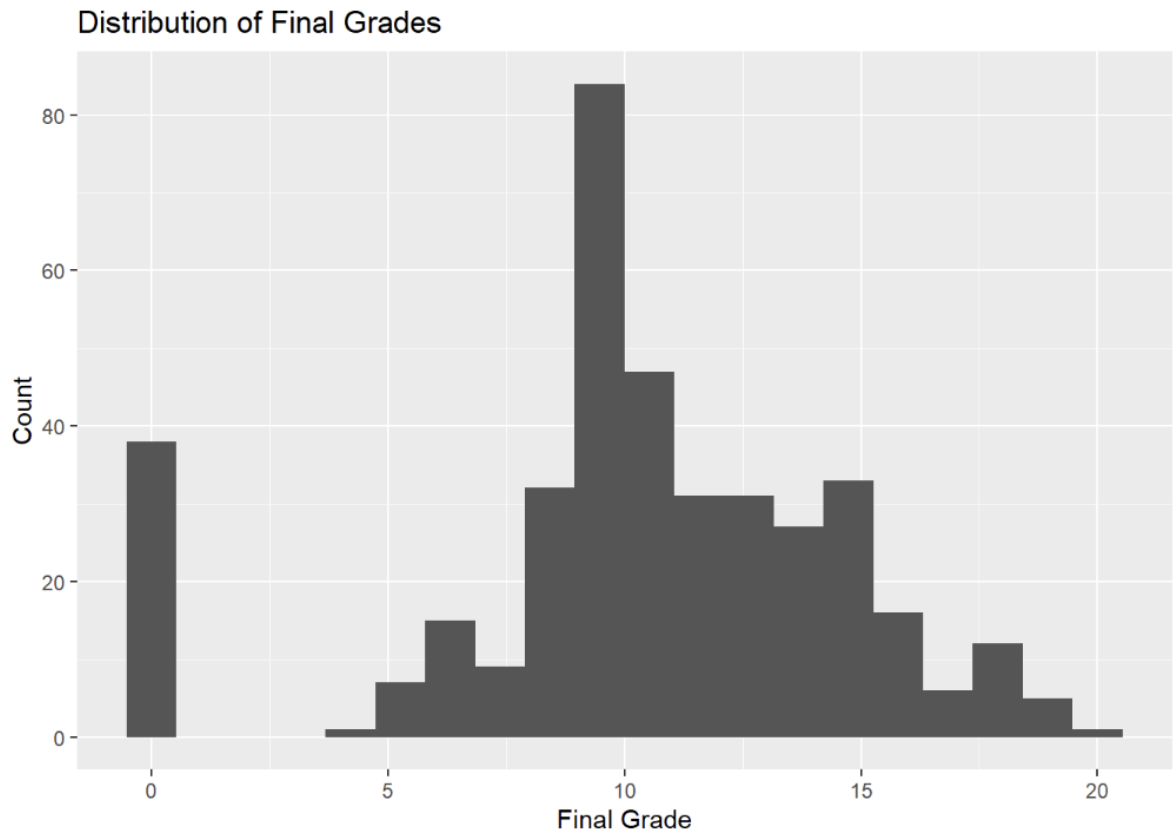
Results - Part 1:

To begin the data analysis, the data had to first be imported from the UCI dataset. To begin this, the data was first downloaded as a CSV file. From there, the data was then placed in the same file directory as the Rmd file used to manipulate and analyze the data. This allowed the Rmd file to read the data with the following command:

```
students <- read.csv("student-mat.csv", sep = ";")
```

From there, an overall distribution of the final grades according to the characteristic variable G3 was computed. Here is the histogram code, as well as the histogram that was plotted of the data.

```
ggplot(students, aes(x = G3)) +
geom_histogram(bins = 20) +
labs(title = "Distribution of Final Grades", x = "Final Grade", y = "Count")
```



Here, it can be seen that the grade data is slightly bimodal. The first peak is around the 9-11 grade mark. The other peak is greatly smaller than the first, but still great enough to mention. This peak is at the 0 grade mark. This showcases that a lot of students weren't performing well at all. From here, a boxplot of grade performance based on each variable chosen needed to be generated. Before doing this, R needs to know that some of the variables being looked at are, in fact, categorical and not numerical. Variables like whether or not a student has internet access are being answered with a "Y" or "N". For R to properly categorize and compute these variables, they need to be factored because they are categorical. The code to do this is shown below:

```
students <- students %>%
mutate(
    sex = factor(sex),
```

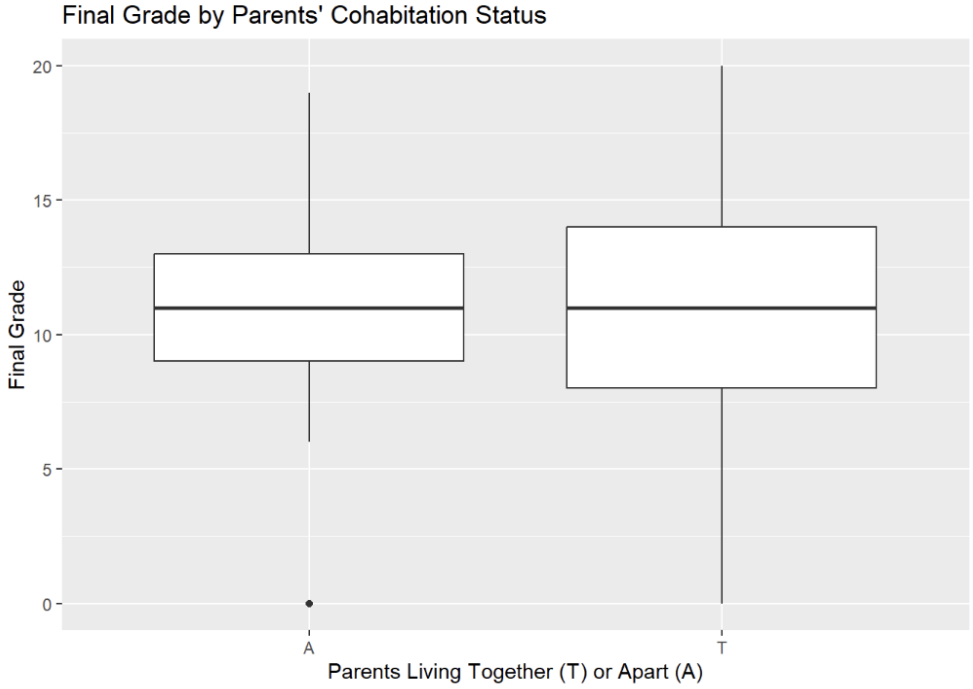
```
    Pstatus = factor(Pstatus),
    famsize = factor(famsize),
    romantic = factor(romantic),
    activities = factor(activities),
    internet = factor(internet)
)
```

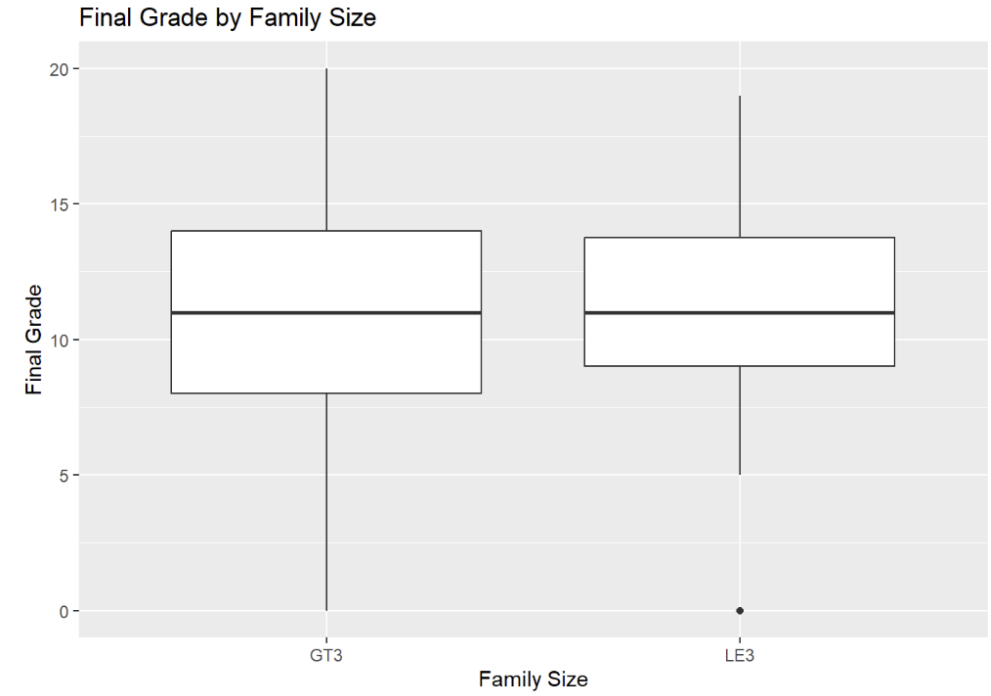
This code showcases each of the variables being looked at in the dataset being factored besides how often a student goes out. This is because the variable is answered numerically (it is answered 1-5, with 5 being going out very often, while 1 being not going out very often). Below is the code for creating the boxplots, as well as the actual boxplots that were generated.

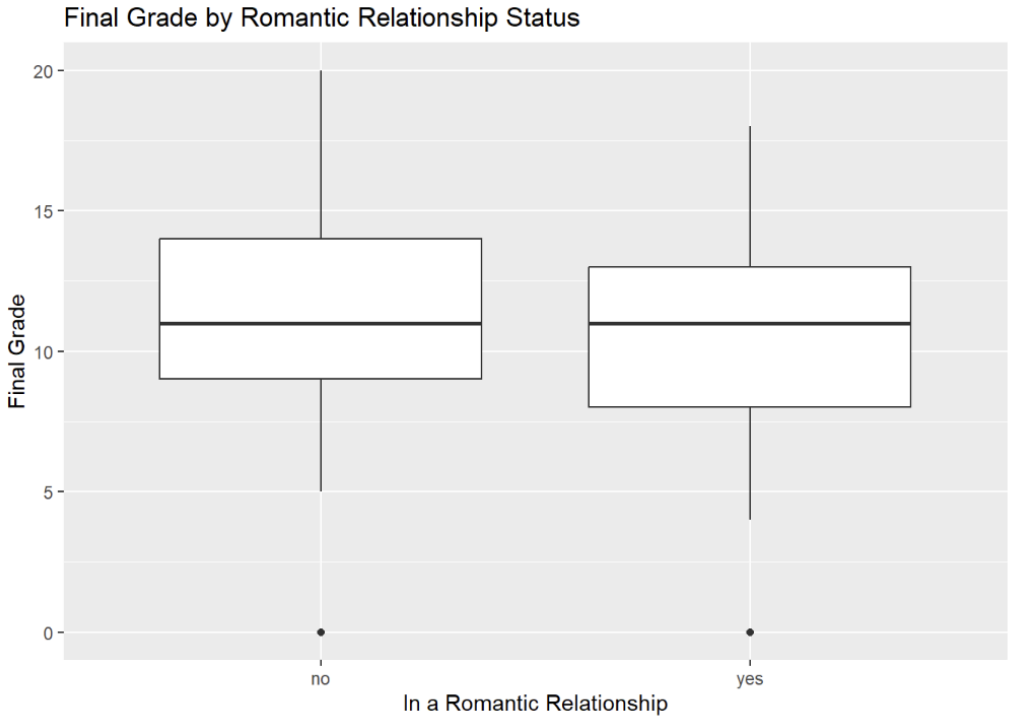
```
ggplot(students, aes(x = sex, y = G3)) +
  geom_boxplot() +
  labs(title = "Final Grade by Gender", x = "Gender", y = "Final Grade")

ggplot(students, aes(x = Pstatus, y = G3)) +
  geom_boxplot() +
  labs(title = "Final Grade by Parents' Cohabitation Status",
       x = "Parents Living Together (T) or Apart (A)",
       y = "Final Grade")

ggplot(students, aes(x = famsize, y = G3)) +
  geom_boxplot() +
  labs(title = "Final Grade by Family Size",
       x = "Family Size",
       y = "Final Grade")

ggplot(students, aes(x = romantic, y = G3)) +
  geom_boxplot() +
  labs(title = "Final Grade by Romantic Relationship Status",
       x = "In a Romantic Relationship",
       y = "Final Grade")

ggplot(students, aes(x = activities, y = G3)) +
  geom_boxplot() +
  labs(title = "Final Grade by Extracurricular Activities",
       x = "Participates in Activities",
       y = "Final Grade")

ggplot(students, aes(x = internet, y = G3)) +
  geom_boxplot() +
  labs(title = "Final Grade by Internet Access at Home",
       x = "Internet Access",
       y = "Final Grade")

ggplot(students, aes(x = factor(goout), y = G3)) +
  geom_boxplot() +
  labs(title = "Final Grade by Frequency of Going Out",
       x = "Going Out Frequency (1 = Low, 5 = High)",
       y = "Final Grade")
```

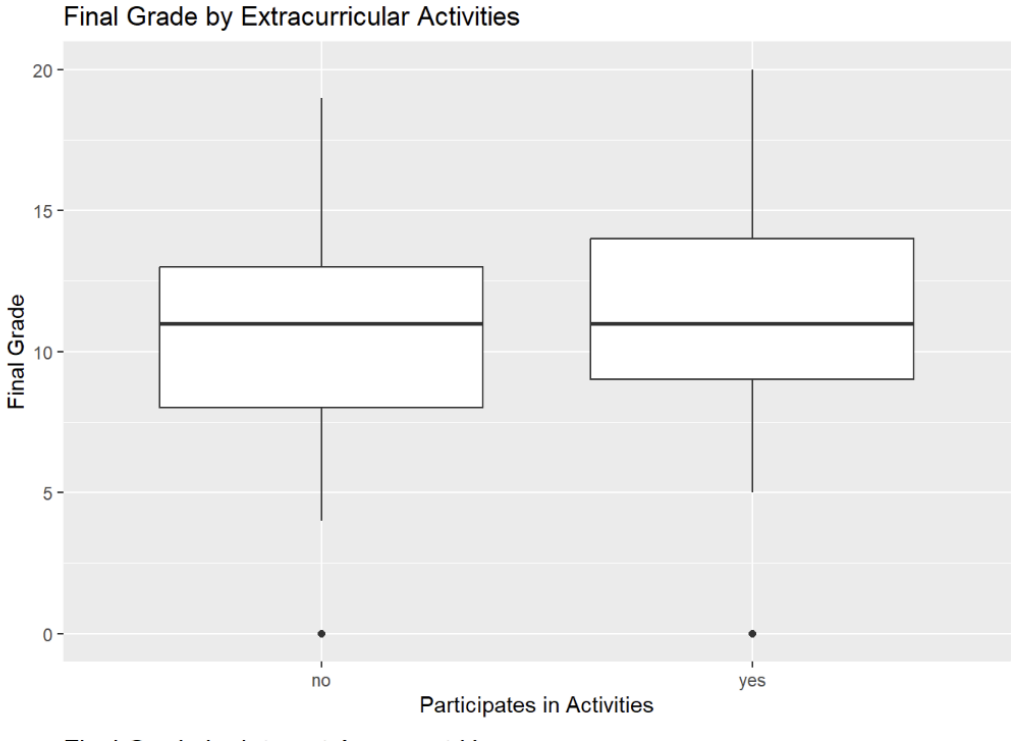
And here are the actual boxplots:

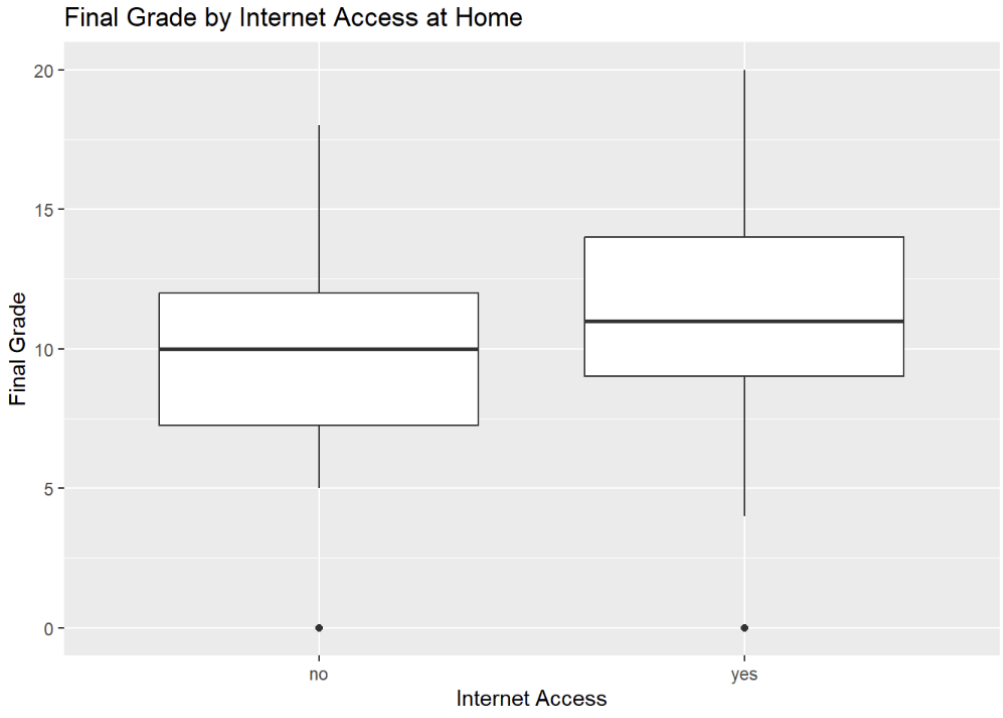

Final Grade by Gender
Final Grade
20
15
10
5
0
F
M
Gender

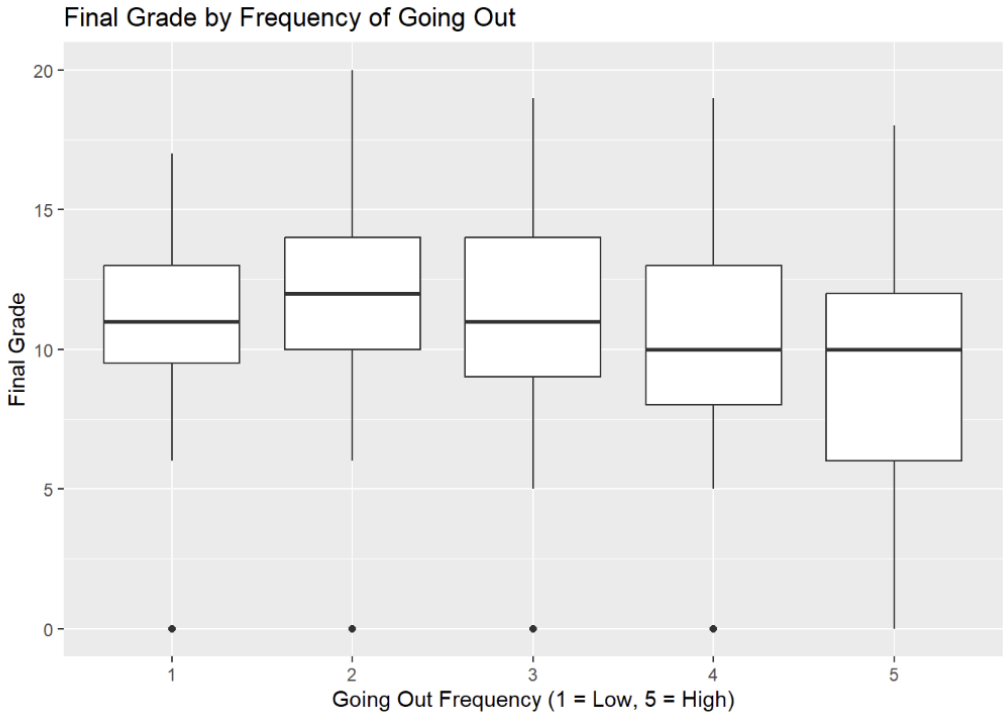

Final Grade by Parents' Cohabitation Status
Final Grade
20
15
10
5
0
A
T
Parents Living Together (T) or Apart (A)


Final Grade by Family Size
Final Grade
20
15
10
5
0
GT3
LE3
Family Size


Final Grade by Romantic Relationship Status
Final Grade
20
15
10
5
0
no
yes
In a Romantic Relationship


Final Grade by Extracurricular Activities
Final Grade
20
15
10
5
0
no
yes
Participates in Activities


Final Grade by Internet Access at Home
Final Grade
20
15
10
5
0
no
yes
Internet Access



These boxplots help to generate interesting visuals and helpful observations that give a bit of an understanding of how these variables affect the overall academic performance of a student. For one, it can be seen that there is a slight shift for men to perform better academically than women. This isn't a great trend, and it doesn't seem like this is statistically significant, but this shift exists. Additionally, it can be seen that if a student doesn't have access to the internet or goes out more frequently (a 4 or 5), their academic performance noticeably drops. There is also a greater range in families with greater than 3 members and families where the parents live apart than in families with fewer than 3 members and families that live together. These boxplots don't provide any numerical data, but so far they seem to suggest a positive correlation between having internet access and academic performance. It also seems to suggest a positive correlation between going out less and academic performance. To provide a more complete picture, the conditional probabilities need to be calculated.

The conditional probabilities were calculated by first calculating the overall mean of the dataset. In the code below, the overall mean of the final grade period is calculated. From there, a new column in the dataset is created called above_mean. If the student is above the mean, a true is placed there. If the student is below the mean, a false is placed there. This allows an easy way to tell if a student performed above the mean or not. The overall mean was calculated to be 10.41519 for the overall dataset. Additionally, the overall median was calculated to be 11 for the overall dataset. The median was calculated in order to determine if the overall distribution of data is relatively symmetrical. This will allow the analysis of the conditional probability to be more accurate.

```
mean_G3 <- mean(students$G3)
students <- students %>%
  mutate(above_mean = G3 > mean_G3)
mean_G3
```

```
## [1] 10.41519
```

```
median_G3 <- median(students$G3)
median_G3
```

```
## [1] 11
```

From there, conditional probabilities can be calculated. These were calculated for a given student using the code below.

```
cp_pstatus <- students %>%
  group_by(Pstatus) %>%
  summarize(
    prob_above_mean = mean(above_mean),
    n = n()
  )
cp_pstatus

cp_famsize <- students %>%
  group_by(famsize) %>%
  summarize(
    prob_above_mean = mean(above_mean),
    n = n()
  )
cp_famsize

cp_sex <- students %>%
  group_by(sex) %>%
  summarize(
    prob_above_mean = mean(above_mean),
    n = n()
  )
cp_sex

cp_romantic <- students %>%
  group_by(romantic) %>%
  summarize(
    prob_above_mean = mean(above_mean),
    n = n()
  )
cp_romantic

cp_activities <- students %>%
  group_by(activities) %>%
  summarize(
    prob_above_mean = mean(above_mean),
    n = n()
  )
cp_activities

cp_goout <- students %>%
```

```
  group_by(goout) %>%
  summarize(
    prob_above_mean = mean(above_mean),
    n = n()
  )
cp_goout

cp_internet <- students %>%
  group_by(internet) %>%
  summarize(
    prob_above_mean = mean(above_mean),
    n = n()
  )
cp_internet
```

This code splits the group by the individual variables that were chosen. It then counts up the number of trues in that column and divides it by the total count. This provides a probability that is functionally equivalent to the conditional probability. This is because it looks at only the rows where a given variable is true, and then it starts counting to see whether each row is above the mean and returns the probability that a given variable is going to lead to rows that are above the mean.

Parental Status (A = Apart, T = Together)

```
## # A tibble: 2 × 3
##   Pstatus prob_above_mean     n
##   <fct>             <dbl> <int>
## 1 A                 0.610    41
## 2 T                 0.520   354
```

Family Size (GT3 = Greater than 3 Members, LE3 = Less than or Equal to 3 Members)

```
## # A tibble: 2 × 3
##   famsize prob_above_mean     n
##   <fct>             <dbl> <int>
## 1 GT3               0.516   281
## 2 LE3               0.561   114
```

Sex (F = Female, M = Male)

```
## # A tibble: 2 × 3
##   sex   prob_above_mean     n
##   <fct>           <dbl> <int>
## 1 F               0.495   208
## 2 M               0.567   187
```

Romantic Relationship (No = Not in Relationship, Y = In Relationship)

```
## # A tibble: 2 × 3
##   romantic prob_above_mean     n
##   <fct>              <dbl> <int>
## 1 no                 0.532   263
## 2 yes                0.523   132
```

Participate in Extracurriculars (Yes = yes, No = no)

```
## # A tibble: 2 × 3
##   activities prob_above_mean     n
##   <fct>                <dbl> <int>
## 1 no                   0.526   194
## 2 yes                  0.532   201
```

How often a Student goes out (1 = Not often, 5 = Often)

```
## # A tibble: 5 × 3
##   goout prob_above_mean     n
##   <int>           <dbl> <int>
## 1     1           0.522    23
## 2     2           0.641   103
## 3     3           0.562   130
## 4     4           0.430    86
## 5     5           0.396    53
```

Access to Internet (Yes = Access to Internet, No = No Access to Internet)

```
## # A tibble: 2 × 3
##   internet prob_above_mean     n
##   <fct>              <dbl> <int>
## 1 no                 0.424    66
## 2 yes                0.550   329
```

Reading these tables is simple. The dbl column is the actual conditional probability. The fct column is the condition based on the assumption made. The n column is the number of people who fall under that condition. For example, given a student has internet access, there is a 55% chance that they are going to perform better than the class average. 329 students have access to the internet. Also, given a student doesn't have internet access, there is a 42.4% chance they are going to perform better than the class average, which is a lot worse. There are also only 66 students who don't have access to the internet. These results help to display which conditions seem to have an influence on student performance.

The conditional probabilities for all of the variables can be seen above. From these, one would expect the probability of being above the mean to be around 50% because the mean and median are so similar. Because the mean and median are so similar, conditional probability analysis can be used. To summarize the above data, the only data that appears to suggest a trend is that having no internet access and going out frequently hurts students' academic performance. These two characteristics have the strongest evidence supporting them that they are harming academic success. Having a smaller family size, being a man, and having your parents be apart also seems to imply a moderately suggestive trend that they help slightly towards boosting academic performance.

Overall, internet access and frequency of going out showed the clearest relationships with above-average final grades, while romantic relationships and extracurricular participation showed little effect. The

similarity between mean- and median-based analyses suggests that these findings are robust to distributional assumptions.

Results - Part 2:

From part 1, it can be seen that the only variables that suggest a trend between academic performance and personal factors are whether or not the student has internet access at home and the frequency with which the student goes out. Because these factors seemed to suggest a trend, further analysis was conducted on them. This involves generating a bootstrap in order to analyze the trends and effects of particular variables on academic performance.

```
set.seed(123)
B <- 5000
boot_yes <- numeric(B)
boot_no  <- numeric(B)
for (b in 1:B) {
  boot <- students %>% sample_frac(replace = TRUE)
  boot_yes[b] <- mean(boot %>% filter(internet == "yes") %>% pull(above_mean))
  boot_no[b]  <- mean(boot %>% filter(internet == "no")  %>% pull(above_mean))
}
# 95% confidence intervals
quantile(boot_yes, c(0.025, 0.975))
quantile(boot_no,  c(0.025, 0.975))
boot_diff_internet <- boot_yes - boot_no
quantile(boot_diff_internet, c(0.025, 0.975))
```

This code uses bootstrap resampling to estimate the conditional probability that a student scores above the mean final, given they have internet access, and given that they don't have internet access. The results of this code are below:

Have Internet Access

```
##     2.5%    97.5%
## 0.4953264 0.6041056
```

Don't Have Internet Access

```
##     2.5%    97.5%
## 0.3060965 0.5428571
```

Difference in Probabilities

```
##        2.5%       97.5%
## -0.002972267  0.255004121
```

These generate 95% bootstrap confidence intervals for:

P(G3 > mean | internet = yes) and P(G3 > mean | internet = no)

The interpretation of this data means that we are 95% confident that between 49.5% and 60.4% of students who have internet access score above the mean final grade. We are also 95% confident that between 30.6% and 54.3% of students who don't have internet access score above the mean final grade. The first interval is almost entirely above 50%, which suggests that students who have internet access are slightly more likely than average to perform above the mean. Additionally, the interval for students without internet access is lower and wider, which suggests more uncertainty and generally worse outcomes for grade performance. The difference in probabilities is the key result, however. The difference in probabilities, since it is positive, suggests that having internet access increases the probability of scoring above the mean final grade. It isn't definitive, but it is highly suggestive. The same methodology was repeated for the going out variable. The code for that is below.

```
set.seed(123)
B <- 5000
boot_go2 <- numeric(B)
boot_go4 <- numeric(B)
for (b in 1:B) {
  boot <- students %>% sample_frac(replace = TRUE)
  boot_go2[b] <- mean(boot %>% filter(goout == 2) %>% pull(above_mean))
  boot_go4[b] <- mean(boot %>% filter(goout == 4) %>% pull(above_mean))
}
quantile(boot_go2, c(0.025, 0.975))
quantile(boot_go4, c(0.025, 0.975))
boot_diff_goout <- boot_go2 - boot_go4
quantile(boot_diff_goout, c(0.025, 0.975))
```

These generate 95% bootstrap confidence intervals for:

P(G3 > mean | Going out = 2) and P(G3 > mean | Going out = 4)

Below are the results:

Going Out = 2 (Going out Less Frequently)

```
##      2.5%     97.5%
## 0.5471534 0.7319634
```

Going Out = 4 (Going out More Frequently)

```
##      2.5%     97.5%
## 0.3289429 0.5349045
```

Difference in Probabilities

```
##       2.5%      97.5%
## 0.07367103 0.34589878
```

The interpretation of this data means that we are 95% confident that between 54.7% and 73.2% of students who go out less frequently score above the mean final grade. We are also 95% confident that

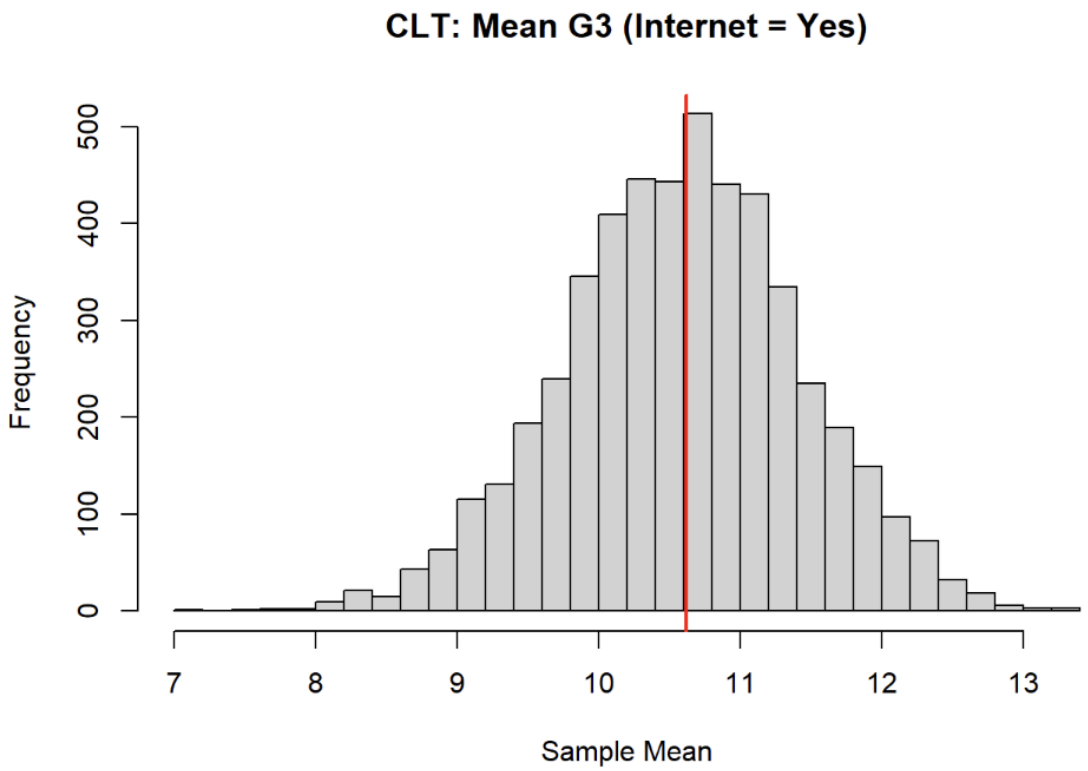
between 32.9% and 53.5% of students who go out more frequently score above the mean final grade. The first interval is entirely above 50%, which suggests that students who go out less are slightly more likely than average to perform above the mean. Additionally, the interval for students who go out more is far lower and almost entirely below 50%, which suggests more uncertainty and generally worse outcomes for grade performance. The difference in probabilities is the key result, however. The difference in probabilities, since it is entirely positive, suggests that going out less increases the probability of scoring above the mean final grade. It isn't definitive, but it is highly suggestive.

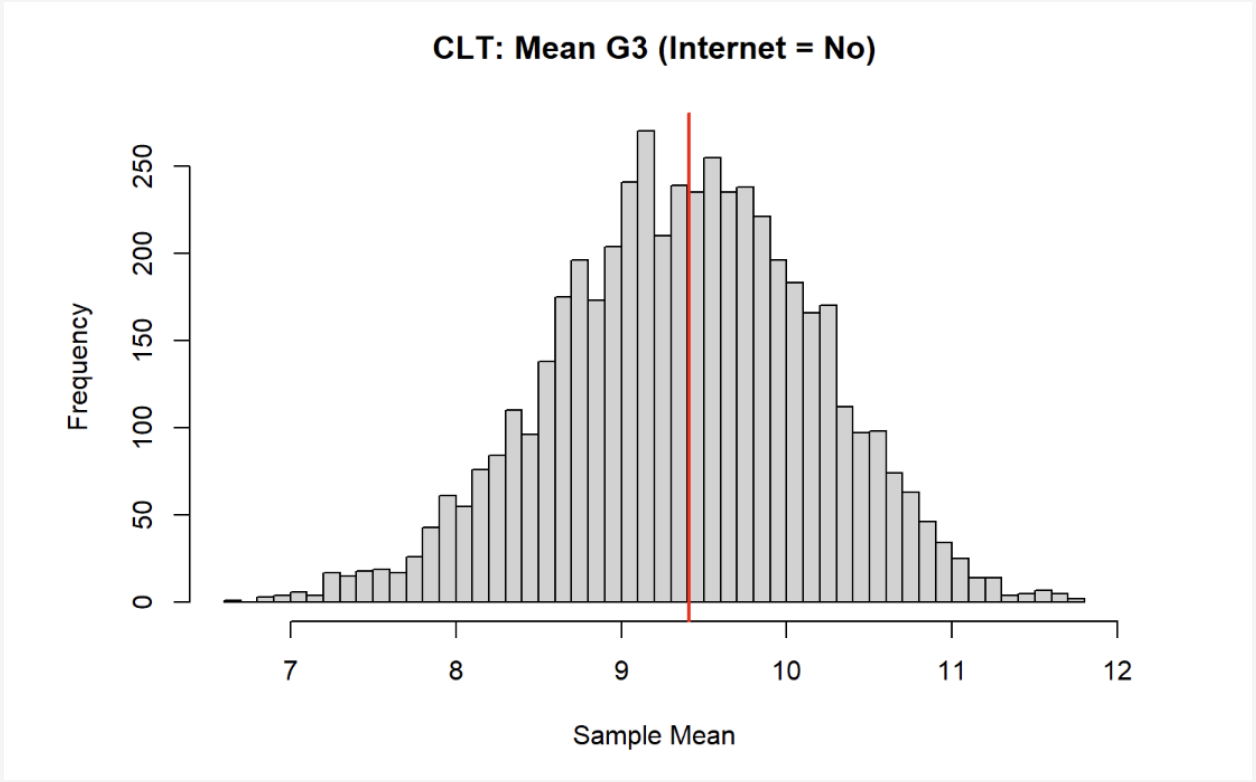
The CLT behavior was also shown for both. Once you resample and sample the dataset multiple times, you start to get a normal distribution, which can be seen with the code below and outputs. This code demonstrates the Central Limit Theorem, in which multiple resamples of the mean can eventually generate a normal distribution.

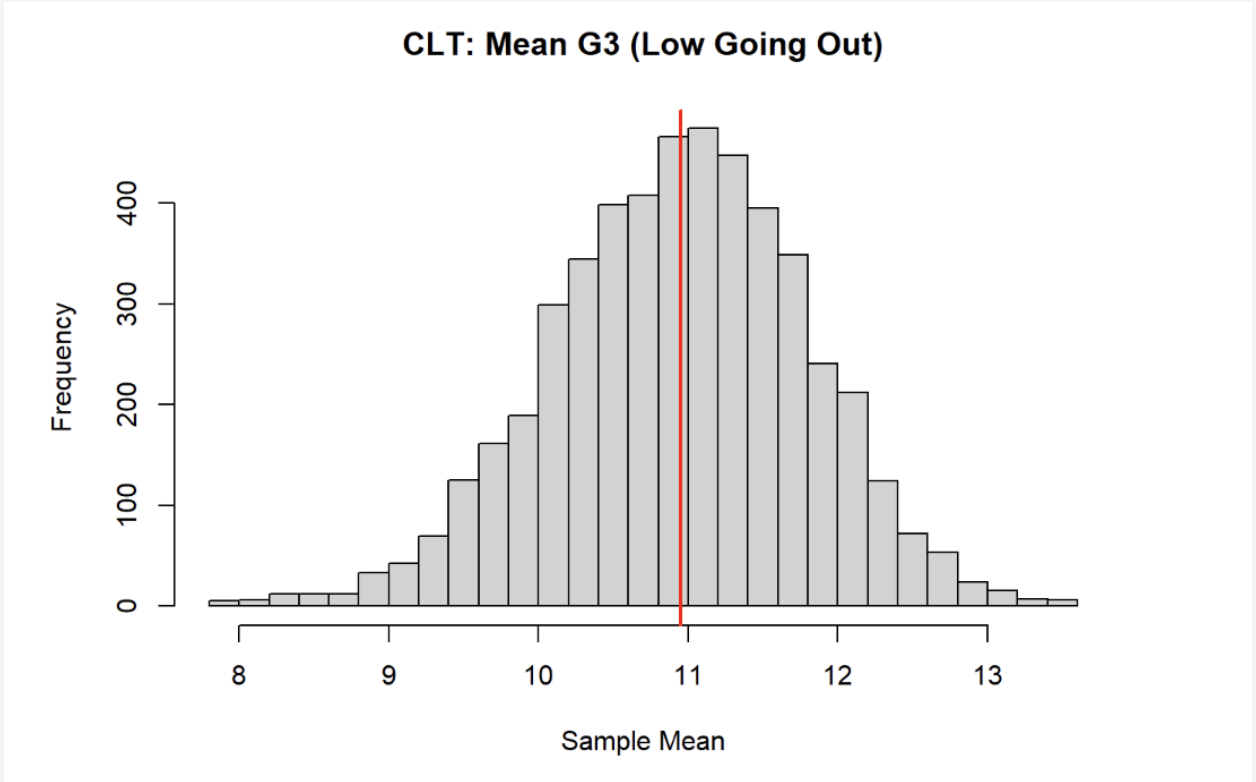
```
n <- 30
B <- 5000
internet_yes <- students %>% filter(internet == "yes") %>% pull(G3)
internet_no  <- students %>% filter(internet == "no")  %>% pull(G3)
clt_yes <- numeric(B)
clt_no  <- numeric(B)
for (b in 1:B) {
  clt_yes[b] <- mean(sample(internet_yes, n, replace = TRUE))
  clt_no[b]  <- mean(sample(internet_no,  n, replace = TRUE))
}
hist(clt_yes,
     breaks = 40,
     main = "CLT: Mean G3 (Internet = Yes)",
     xlab = "Sample Mean")
abline(v = mean(internet_yes), col = "red", lwd = 2)

hist(clt_no,
     breaks = 40,
     main = "CLT: Mean G3 (Internet = No)",
     xlab = "Sample Mean")
abline(v = mean(internet_no), col = "red", lwd = 2)
low_goout  <- students %>% filter(goout <= 2) %>% pull(G3)
high_goout <- students %>% filter(goout >= 4) %>% pull(G3)
clt_low <- numeric(B)
clt_high <- numeric(B)
for (b in 1:B) {
  clt_low[b]  <- mean(sample(low_goout,  n, replace = TRUE))
  clt_high[b] <- mean(sample(high_goout, n, replace = TRUE))
```

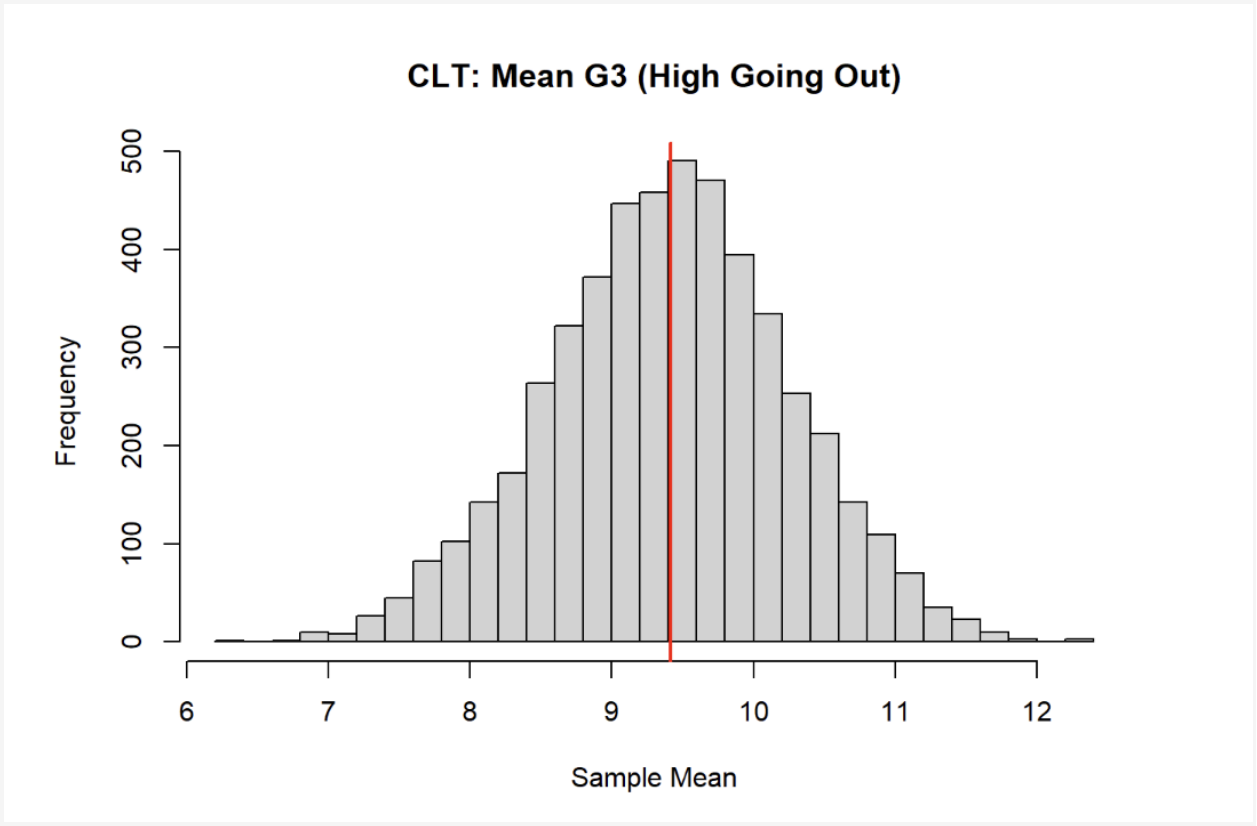
```
}
hist(clt_low,
     breaks = 40,
     main = "CLT: Mean G3 (Low Going Out)",
     xlab = "Sample Mean")
abline(v = mean(low_goout), col = "red", lwd = 2)

hist(clt_high,
     breaks = 40,
     main = "CLT: Mean G3 (High Going Out)",
     xlab = "Sample Mean")
abline(v = mean(high_goout), col = "red", lwd = 2)
```

Here are the histograms that were produced:




CLT: Mean G3 (Internet = No)
Frequency
0
50
100
150
200
250
7
8
9
10
11
12
Sample Mean


CLT: Mean G3 (Low Going Out)
Frequency
0
100
200
300
400
8
9
10
11
12
13
Sample Mean



As can be seen above, each histogram has a normal distribution that overall displays the CLT. When an individual sample is resampled multiple times (in this case, 5000), it eventually displays a normal distribution that is centered over the true population mean. Additionally, as can be seen from the histograms above, the sample means for going out less tend to be shifted lower than going out more. Additionally, the sample means for not having internet access are shifted lower than having internet access. This displays a trend that going out less and having more internet access can have a positive benefit on academic performance.

Results - Part 3:

The first thing to be done in this model is to provide a least squares estimation (using both manual estimation and using the lm function). This will provide us with the expected change in the final grade for a one-unit increase in the going-out frequency. It will also measure the difference in mean final grade between students with and without internet access. The internet access is different because this variable is categorical and not numerical. The code below conducts this.

```
x <- students$goout
y <- students$G3
b1 <- sum((x - mean(x)) * (y - mean(y))) / sum((x - mean(x))^2)
```

```
b0 <- mean(y) - b1 * mean(x)
c(b0 = b0, b1 = b1)
lm_goout <- lm(G3 ~ goout, data = students)
coef(lm_goout)
lm_internet <- lm(G3 ~ internet, data = students)
summary(lm_internet)
```

```
## (Intercept)       goout
##   12.114098   -0.546473
##
## Call:
## lm(formula = G3 ~ internet, data = students)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.617  -2.013   0.383   3.383   9.383
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.4091     0.5619  16.745   <2e-16 ***
## internetyes   1.2079     0.6157   1.962   0.0505 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.565 on 393 degrees of freedom
## Multiple R-squared:  0.009699,   Adjusted R-squared:  0.007179
## F-statistic: 3.849 on 1 and 393 DF,  p-value: 0.05048
```

The analysis for this shows a negative slope for people who go out. The negative slope is -0.546 units. This means that for every additional frequency of going out, you can expect a decrease of your final grade by 0.55 points. This shows how students who go out more tend to have slightly lower grades. The linear regression for the internet shows that the average grade of students without internet is 9.409, which is below the overall mean. It also shows that students with internet access score about 1.2 points higher than students without. The p-value is also right at the threshold of statistical significance, which suggests a marginally significant effect. The R^2 squared value is really low, though, which shows that it explains less than 1% of the overall variation in the final grades, which means that internet access tends to have a small effect overall.

After this, a randomization test was conducted in order to test for the slope. It was conducted with the following code:

```
set.seed(702)
x <- students$goout
y <- students$G3
fit_obs <- lm(y ~ x)
b1_obs <- coef(fit_obs)[2]
B <- 5000
b1_null <- numeric(B)
```

```
for (b in 1:B) {
  y_perm <- sample(y)
  b1_null[b] <- coef(lm(y_perm ~ x))[2]
}
p_val <- mean(abs(b1_null) >= abs(b1_obs))
c(observed_slope = b1_obs, p_value = p_val)

set.seed(702)
x_int <- students$internet
y <- students$G3
fit_obs_int <- lm(y ~ x_int)
b1_obs_int <- coef(fit_obs_int)[2]
B <- 5000
b1_null_int <- numeric(B)
for (b in 1:B) {
  y_perm <- sample(y)
  b1_null_int[b] <- coef(lm(y_perm ~ x_int))[2]
}
p_val_int <- mean(abs(b1_null_int) >= abs(b1_obs_int))
c(observed_difference = b1_obs_int, p_value = p_val_int)
```

After this, the p-value for each could be constructed. For the going out variable, the slope was found to be the same as the least squares estimation, being around -0.55. The p-value was then 0.008. This suggests that the correlation that was found was statistically significant. This means that there is a real negative association between going out and academic performance. The same was done for internet access. The slope for that one was found to be 1.207, and the p-value was 0.05. This meant that there was also a statistically significant positive correlation between having internet access and performing better in school by 1 point.

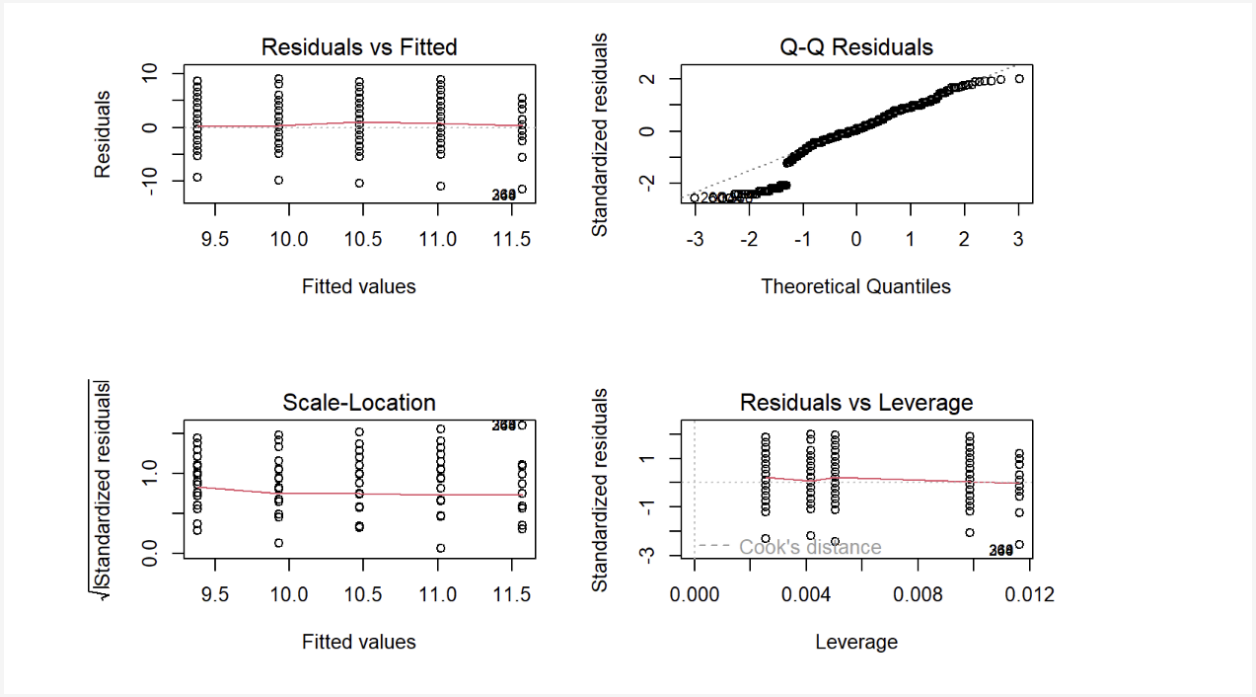
```
anova(lm_internet, lm_goout)

## Analysis of Variance Table
##
## Model 1: G3 ~ internet
## Model 2: G3 ~ goout
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1    393 8189.7
## 2    393 8124.1  0    65.618

summary(lm_internet)$r.squared

## [1] 0.009698974
```

```
summary(lm_goout)$r.squared
```

```
## [1] 0.01763358
```

Above, the code for generating the ANOVA table is shown. The ANOVA table shows that the internet access variable explains slightly less variation in grade performance than the going out variable. It should be noted, however, that it seems both effects are small, though. The R^2 values are also shown. The internet accounts for ~1% of the variance explained, while the going out variable accounts for ~1.76% of the variance. Finally, to summarize, the diagnostics are shown for the variable of going out because this was a numerical variable and also displayed more variation in the overall grade performance.

```
par(mfrow = c(2,2))
plot(lm_goout)
```



These diagnostic plots showcase linearity mostly. It also slightly showcases normality with some slight deviation in the tails. It also helps to show that the model is not dominated by outliers. This helps to showcase that the lm(G3~goout) model is reasonably well-behaved and that we can trust the slope estimate and p-value that were given to us.

Discussion:

Overall, the three statistical methods that were used showcase that grade performance is negatively tied to going out and the inability to access the internet at home. Initially, the conditional probabilities of

multiple different characteristics were found. From these conditional probabilities, the two with the most impact on the actual grade performance were chosen. These were the frequency of going out and whether or not the student had access to the internet. From there, a bootstrap analysis was performed to further analyze the effect of these characteristics on grade performance. These provided confidence intervals for mean differences between groups. These helped to indicate that going out more and not having internet hurt academic performance. These intervals also quantify the uncertainty in the differences and help to show that these variables have a real impact on grades. Then, the CLT behavior was analyzed. It can be seen that even for small sample sizes (n=30), the Central Limit Theorem is still observed. Repeated sampling led to a normal distribution of mean grades that appeared to center around the true mean of the group. The CLT validated the reliability of the sample means as an estimate of the true population mean, which justifies the use of linear models and confidence intervals. Finally, from there, linear regression was performed. The p-values were small, which indicated a real, measurable impact on the actual grades. Additionally, the linear regression showed that having interned increases grades slightly, and going out more is associated with a lower grade performance. The R^2 values, however, were unfortunately very small. This meant that these variables only explain a very small fraction of the overall variance in the grades. This is not surprising, though. This is because there were many factors that were ignored due to the scope of this project. This means that although they did impact the grades, they weren't the only factor impacting the grade performance. The findings are also consistent throughout all three statistical methods taken throughout this report. It's important for schools to know about this impact as it can help them be aware of the students they need to look out for. It should also be communicated to these schools that this isn't the only thing that they should look out for, and multiple other factors need to be studied to get a comprehensive overview of what impacts grade performance in students.