Statistical Analysis of Duke Student's GPA

Sophie Osborn

Introduction

Within this report, I will examine a data set detailing the habits and statistics of students at Duke University. I will use the data set to explore the effects that gender and hours of sleep per night could have on GPA. I was interested in these questions because of I wanted to test stereotypes surrounding men and women within education and either confirm or deny the claims that are often made based on these stereotypes. It is also prevalent for students to spend more time studying than sleeping so I was curious to see whether this benefits or harms performance in school. Before the exploration, I think it is important to display all elements of the data set, and conduct background research on the content of the data set that I will examine.

Data Set Breakdown

This dataset describes details of Duke students' lives both in and out of school as reported through a voluntary survey of students. While this data comes directly from the surveyors, I've received this data from OpenIntro. There are 5 variables and 55 observations within this data set. The variables include GPA, hours of study per week, hours of sleep per night, number of nights out per week, and gender of the person taking the survey. GPA, hours of sleep per night, and nights out per week are all numerical continuous variables. Hours of study per week is also numerical, but discrete rather than continuous. Gender is the only nominal categorical variable.

	GPA	Study hours/week	Sleep hours/night	Out	Gender
1	3.89	50	6	3	Female
2	3.9	15	6	1	Female
3	3.75	15	7	1	Female
4	3.6	10	6	4	Male
5	4	25	7	3	Female
6	3.15	20	7	3	Male
7	3.25	15	7	1	Female
8	3.925	10	8	3	Female
9	3.428	12	8	2	Female
10	3.8	2	8	4	Male
11	3.9	10	8	1	Female
12	2.9	30	6	2	Female
13	3.925	30	7	2	Female
14	3.65	21	9	3	Female
15	3.75	10	8.5	3.5	Female
16	4.67	14	6.5	3	Male
17	3.1	12	7.5	3.5	Male
18	3.8	12	8	1	Female
19	3.4	4	9	3	Female
20	3.575	45	6.5	1.5	Female

21	3.85	6	7	2.5	Female
22	3.4	10	7	3	Female
23	3.5	12	8	2	Male
24	3.6	13	6	3.5	Female
25	3.825	35	8	4	Female
26	3.925	10	8	3	Female
27	4	40	8	3	Female
28	3.425	14	9	3	Female
29	3.75	30	6	0	Female
30	3.15	8	6	0	Female
31	3.4	8	6.5	2	Female
32	3.7	20	7	1	Female
33	3.36	40	7	1	Female
34	3.7	15	7	1.5	Male
35	3.7	25	5	1	Female
36	3.6	10	7	2	Female
37	3.825	18	7	1.5	Female
38	3.2	15	6	1	Female
39	3.5	30	8	3	Male
40	3.5	11	7	1.5	Female
41	3	28	6	1.5	Female
42	3.98	4	7	1.5	Female
43	3.7	4	5	1	Male
44	3.81	25	7.5	2.5	Female
45	4	42	5	1	Female
46	3.1	3	7	2	Male
47	3.4	42	9	2	Male
48	3.5	25	8	2	Male
49	3.65	20	6	2	Female
50	3.7	7	8	2	Female
51	3.1	6	8	1	Female
52	4	20	7	3	Female
53	3.35	45	6	2	Female
54	3.541	30	7.5	1.5	Female
55	2.9	20	6	3	Female

Background Research

Before furthering my own analysis, it is important to look into what other research may have been done with this data, or with questions of this nature and what they resulted in. For

example, an exploration of this data set has been done by a person named Christina Pace who examined each variable and how it may have affected the students' GPAs. Pace's findings in reference to this specific data set seem to acknowledge that the data set may simply be too small to notice any significant variation in GPA based on the factors included, but ultimately found that hours of sleep per night had the most impact on overall GPA (Pace, 2018).

In examining research done on the correlation between gender and GPA performed by Dylan Conger and Mark C. Long shows that there is a relationship between the two. They found that women tend to have a higher GPA in college than men (Conger & Long). Long and Conger performed their own research surrounding many other factors of GPA outcome including focus of study, the college which the student attends, quality of student in High School, and non-cognitive differences between men and women. Though the factors of focus may be different than to variables in this study, it is significant to note that Long and Conger found a higher GPA in women than men (Conger & Long).

Research surrounding the correlation between hours of sleep per night and GPA have also taken place. A study by Creswell and colleagues monitored first month of students first of second semester of Freshman year and then observed their end of semester subsequent GPA (Creswell, et al., 2013). The study found that "… every additional hour of average nightly sleep duration early in the semester was associated with an 0.07 increase in end-of-term GPA" (Creswell, et al., 2013). They Also found that students who slept for less than 6 hours a night were more likely to have a below average GPA come semester end (Creswell, et al., 2013). This data leads me to believe that there may be a similar trend to be found in the data I am examining here.

From this research, I can conclude that there is significant reason for me to explore these topics within this data set, specifically to confirm or refute Pace's finding of the relationship between hours of sleep per night and GPA. Based on this research I will explore the questions and hypotheses included in the next section.

Hypotheses and Research Questions

Do male students have a higher GPA on average than female students?

Null Hypothesis: $\overline{x_{diff}} = 0$

There is not a significant difference between the average of male students GPA and female students GPA.

Alternative Hypothesis: $\overline{x_{diff}} \neq 0$

There is a significant difference between the average of male students GPA and female students GPA.

Does sleeping more every night guarantee a higher GPA?

Null Hypothesis: $\beta = 0$

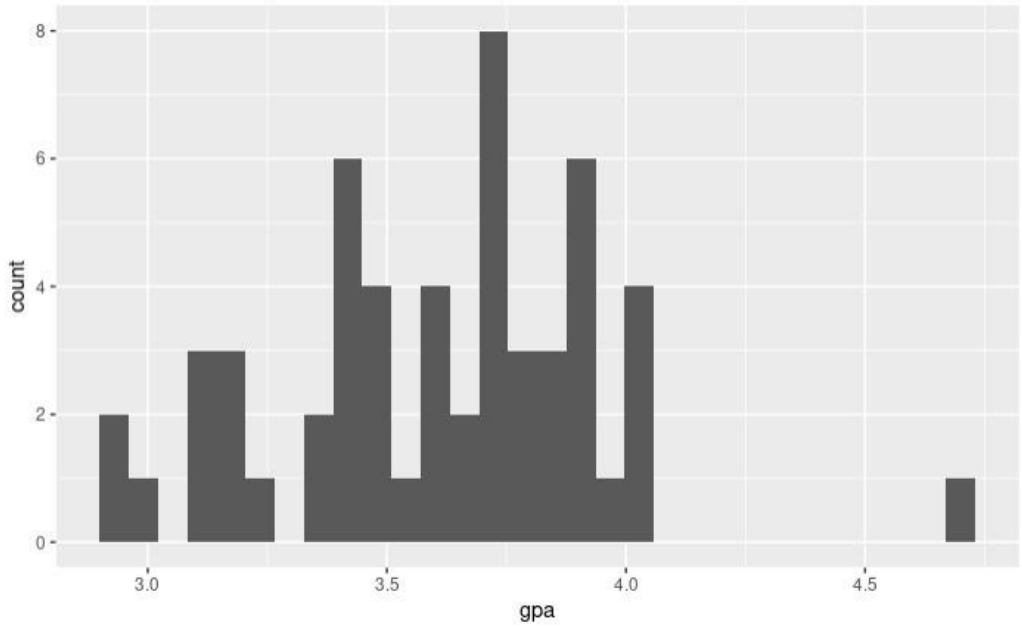
The slope of the linear regression line is 0; there is no relationship between hours of sleep per night and a student's GPA.

Alternative Hypothesis: $\beta \neq 0$

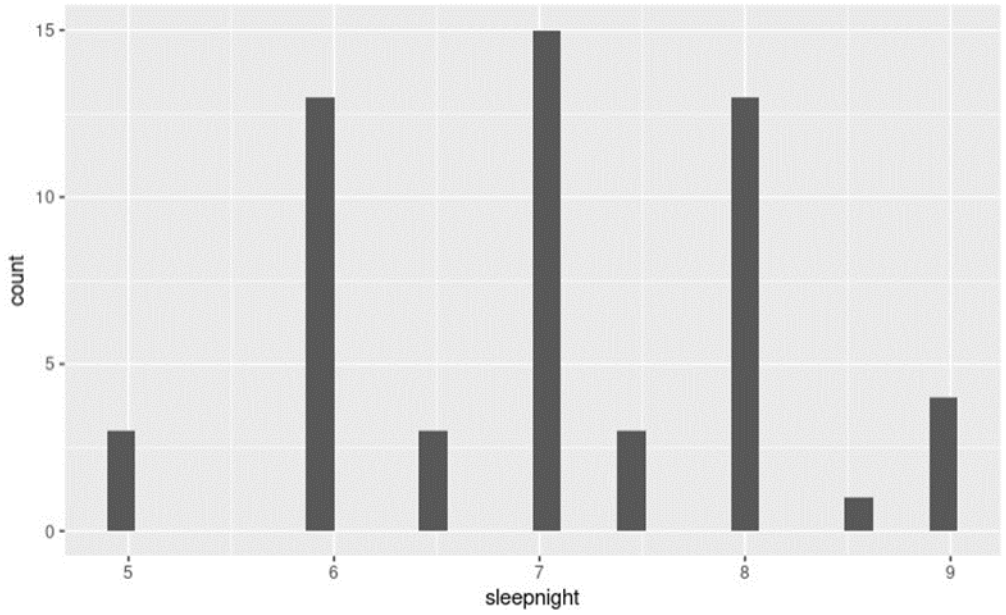
The slope of the linear regression line is not 0; there is a relationship between hours of sleep per night and a student's GPA.

Data Exploration

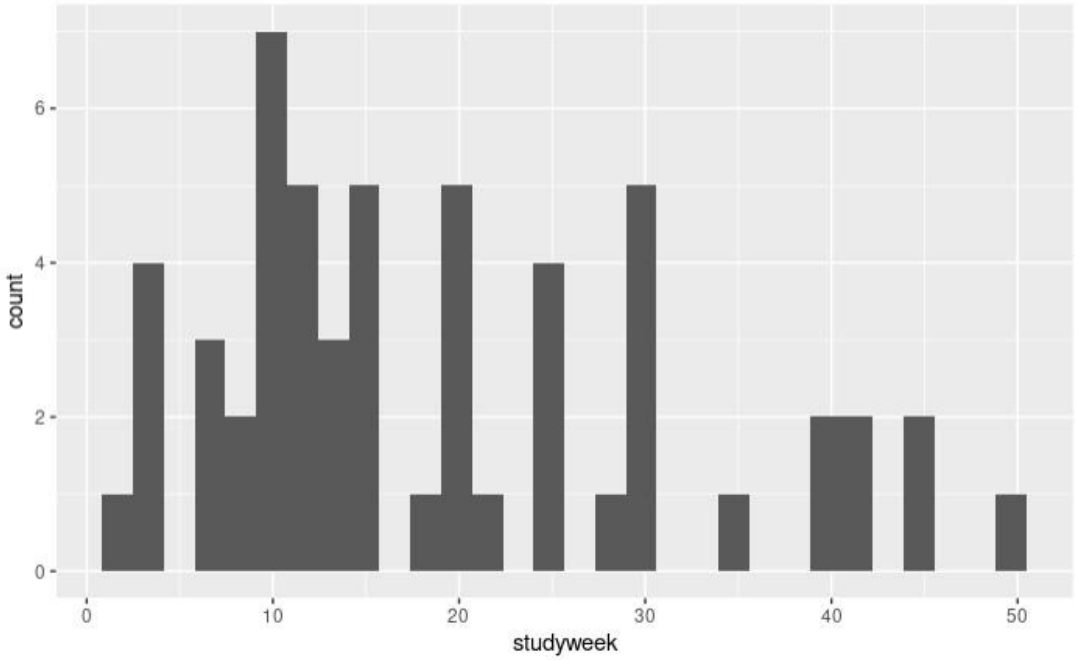
Once I formulated my research questions, I performed some initial analysis of the data to determine the behavior of the data. The first step to take is to look at each numerical variable's distribution to determine if it is normal or skewed.



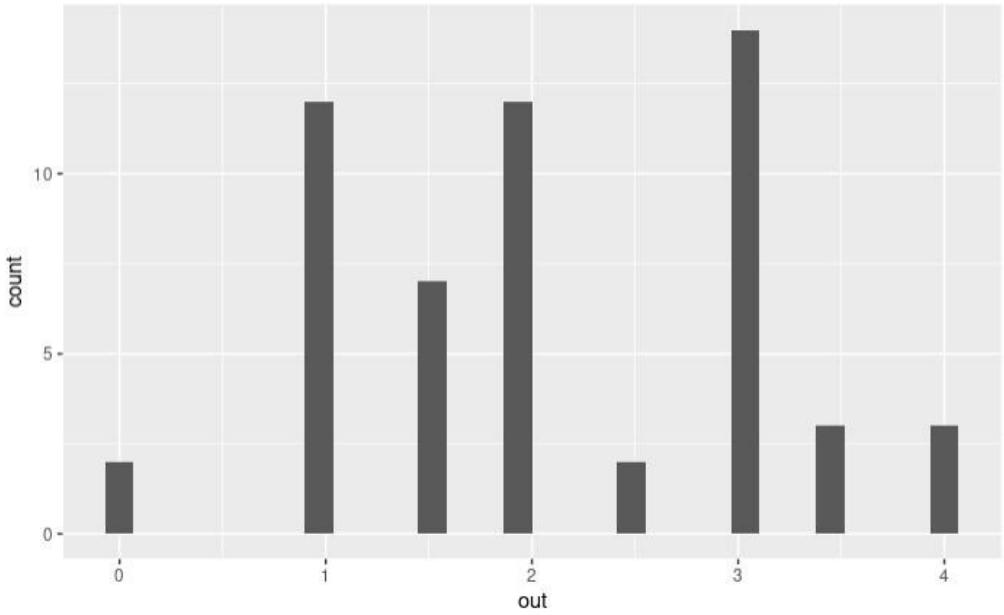
This histogram displays the distribution of the GPA variable. As you can see above it initially seems normally distributed, but taking a closer look it is clear that there is an outlier on the far-right end of the data distribution. With this in mind, the data seems to be skewed right.



The above figure shows the distribution of the variable describing the hours of sleep per night of each student. As you can see from examining the histogram, the data appears to behave relatively normally, but with some interesting outliers throughout.



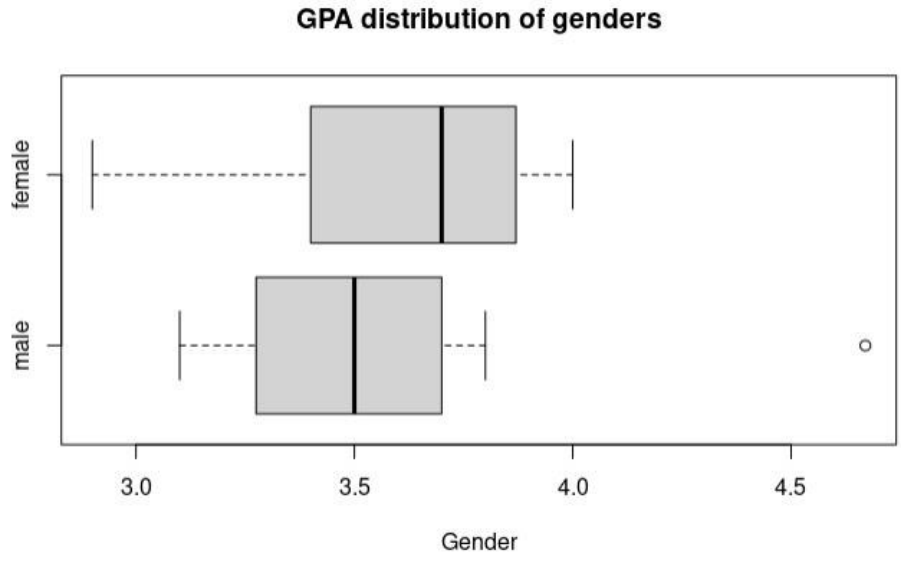
In the above histogram, the spread of the study hours per week variable is examined. The spread of this variable appears to be almost normal distribution with a skew on the left side.



Finally, the last numerical variable of nights out per week is examined. The resulting spread is like the hours of sleep per night variable in the sort of sporadic nature, but with less normal behavior.

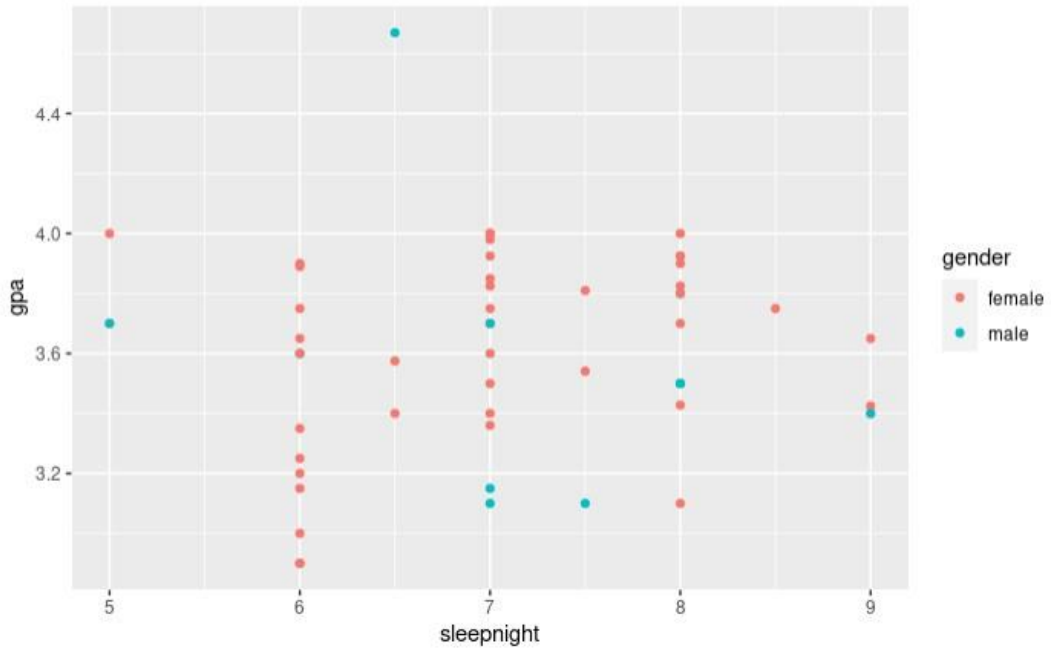
Now that the behavior of the numerical observations is clear, I will move on to the preliminary relationships between GPA and the variable present in the research question to visually assess the relationships present between GPA and the variables in question.

Do male students have a higher GPA on average than female students?



In this box plot the spread of GPA held by each gender is displayed. The higher overall GPA and higher average GPA is clearly seen with the female students surveyed, but there is also a greater spread of GPAs across the female students in comparison to the male students. The greater spread in distribution could be a result of there being far more female participants in the survey than male participants which could skew the data.

Does sleeping more every night guarantee a higher GPA?

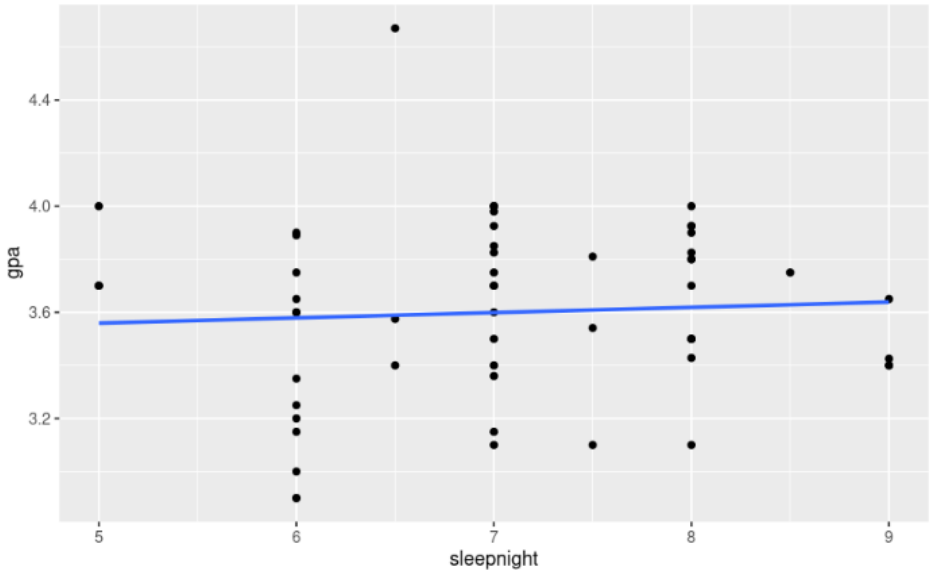


The scatter plot above shows the distribution of GPA compared to the distribution of hours of sleep per night. Aside from the outliers on either end of the hours of sleep per night, it is clear from this plot that the correlation between hours of sleep per night and respective GPA are not clear in this data set and need further analysis.

Methods

In order to determine whether the null hypotheses are correct or incorrect I will need to perform two different tests on the data. The first test I performed was a T-test for difference of two means to determine the relationship between gender and GPA. The data is suitable for this kind of test as it is a random sample and there are no extreme outliers. First, I calculated the mean GPA of both genders which is 3.61 for the female responses, and 3.56 for the male. Then I calculated the standard deviation present in each group which ended up being 0.31 for female, and 0.42 for the male GPA. Then I calculated the point estimate for the test using difference of the two means and got 0.051, and the standard error of my estimate which was 0.131. Using this information, I was able to find a t-score of 0.39. Then, using the t-score I was able to calculate a p-value of 0.7026. As a final test for this data, I used a confidence level of 0.95 and subsequent alpha of 0.05 to determine a confidence interval with an upper bound of 0.339 and a lower bound of –0.237.

In the next test I performed a T-test for a simple linear regression to determine the relationship between hours of sleep per night of the students and their GPA's. I used this test and the ordinary least squares method because the data consists of independent random samples, the residuals of a linear approximation are both normal and homoscedastic, and the explanatory variable, GPA, is independent from the residuals. Once I determined that this was the correct test for the hypotheses, I found the ordinary least squares statistics including the slope of the linear regression, about 0.198, the intercept, 2.199, and the R^2 to be 0.372. While I would hope that the R^2 value would be a bit higher, a linear regression would still be a good fit for this data. From the values I calculated I can define the equation of the regression to be: $y = 2.199 + 0.198x$, where y is the explanatory variable, GPA, and x is the response variable of hours of sleep per night. Using this equation, I fit the linear regression to the data in the graph seen below:



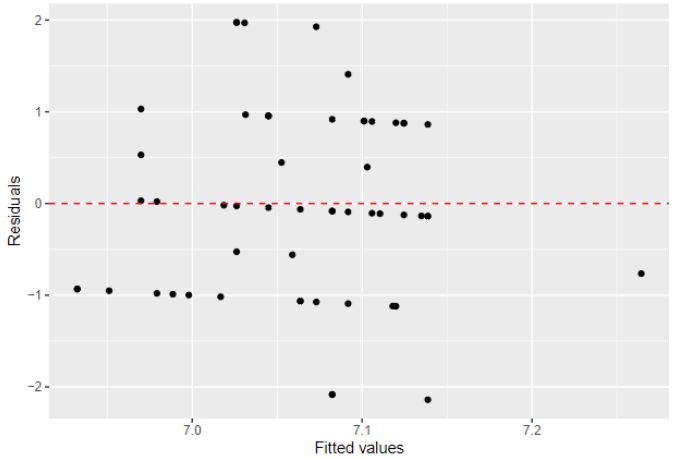
From there I performed the t-test and found a p-value of 0.13. Using an alpha of 0.05 and a significance level of 0.95 I calculated a confidence interval from -0.335 to 0.732.

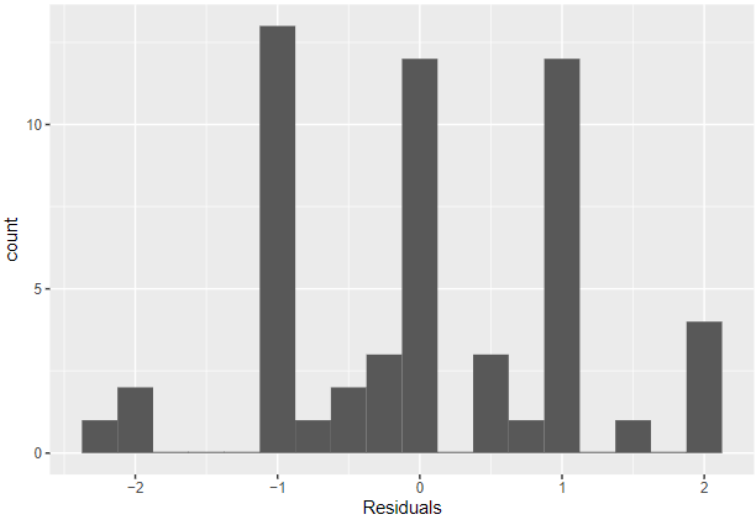
Results and Discussions

From the first hypothesis test of the relationship between average GPA's of male and female respondents I was able to conclude several things. First was that the error rate was relatively high compared to what I would hope. Then I noticed that the p-value is also very high. In well fit hypothesis test I would expected a p-value much lower. Because of this p-value relative to the 0.05 alpha value I believe I committed a type 1 error and failed to reject the null. Then, from my confidence interval including the null I was able to conclude that there is no significant relationship between gender and GPA. This leads me to believe that any variation between the two gender's average GPAs is most likely due to chance. I think this test could more accurately be completed with a few changes including a close to equal number of female and male responses to the survey, as well as a larger sample size. The potential response validity of a voluntary survey like the one used for this set of data is another component that could lead to a

bias of results. Both issues could potentially be remedied if the information was gathered directly from the school. The surveyors could approach the school and ask the school to allow students to give permission for their data to be anonymously used in a study such as this. This would also help to separate any other possible forms of error which could skew results such as a student's major, their class standing, and whether they had credits carried over from a high school course. Even though the research I conducted prior to performing any hypothesis test lead me to believe that I would find a difference between the average GPA of male student and female students, I am not surprised that my result is different from this research because there was for more factors considered in this conclusion compared to my own.

In the second hypothesis test, I also arrived at a few interesting conclusion regarding the relationship between hours of sleep per night and GPA. One of the first things I noticed was the relatively low R squared value of 0.37. This lead me to believe that the linear model may not be well fit initially, but I do know that this is also not the worst R squared value for a linear model, so I continued my test. I also noticed from the linear regression line that there was a slight positive linear slope, though it was very small compared to the spread of the data. In performing the t-test and finding a p-value of 0.13 I was relatively satisfied. Though, a 0.13 p-value when compared to an alpha value of 0.05 leads me to believe I committed a type 1 error and failed to reject the null, I was happy with this relatively low p-value. When confirming that I failed to reject the null using a confidence interval I found that the null value of 0 is within this interval meaning that the results are not statistically significant, which agrees with the p-value. From these results I can conclude that there is no significant linear relationship between hours of sleep per night and GPA, and that the slight positive linear slope of the regression line is most likely by chance. This result can also be confirmed by the visualization of the residuals of the hours of sleep per night variable. These graphs show that the residuals are not normally distributed and appear to be heteroscedastic. Though, it is important to note that plenty of other studies have been done which report that more sleep results in a better GPA, as I stated previously in my report. From this I think the surveying methods could be changed for the same reasons as I mentioned when analyzing the first test.



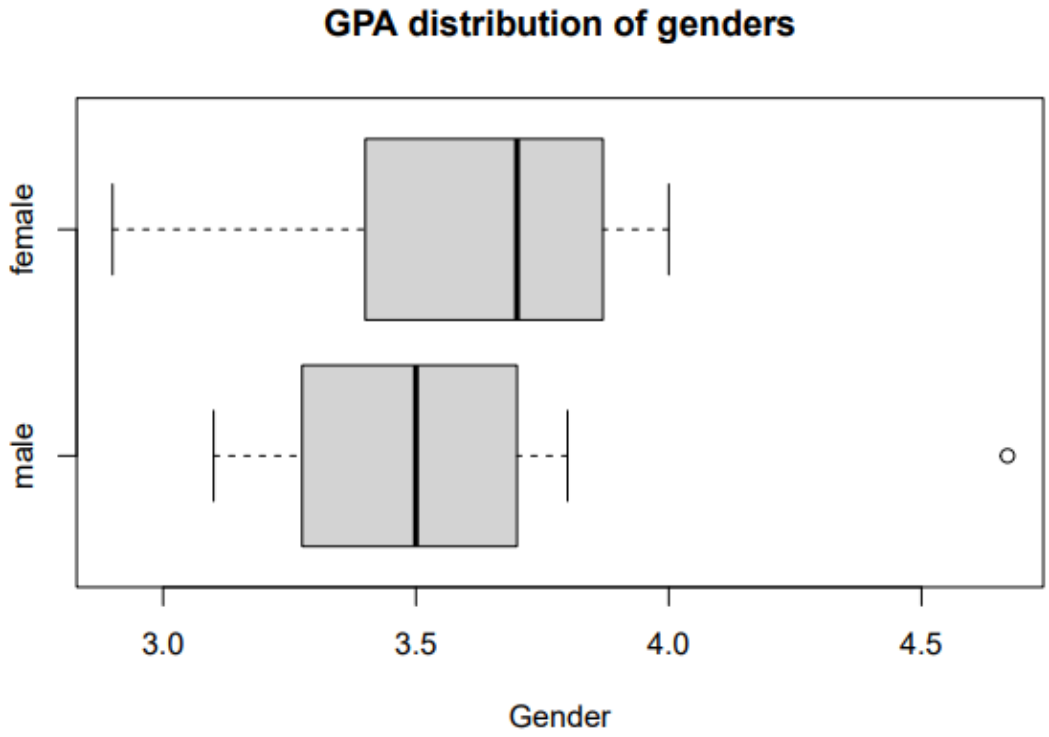


Conclusions

Throughout this report I examined the Openintro “gpa” data set which took a volunteer survey of Duke students and their statistics surrounding gender, GPA, hours of sleep per night, hours of study per week, and average nights out per week. Upon research done about this data I found that there has been conclusion reached which links a relationship between both gender and hours of sleep per night and GPA. Using this research, I formulated research questions and hypotheses examining whether a relationship between GPA and these two variables existed. Despite my initial research leading me to believe that there would be relationships present in both cases, I was only able to conclude that any relationship present in the data was most likely due to chance. Though, it is important to note that this was not only a volunteer survey which is most likely biased, but the responses were not well distributed across many scenarios which would lead to skewed results. I propose a larger data set with a more anonymous survey in order to perform a more accurate hypothesis test.

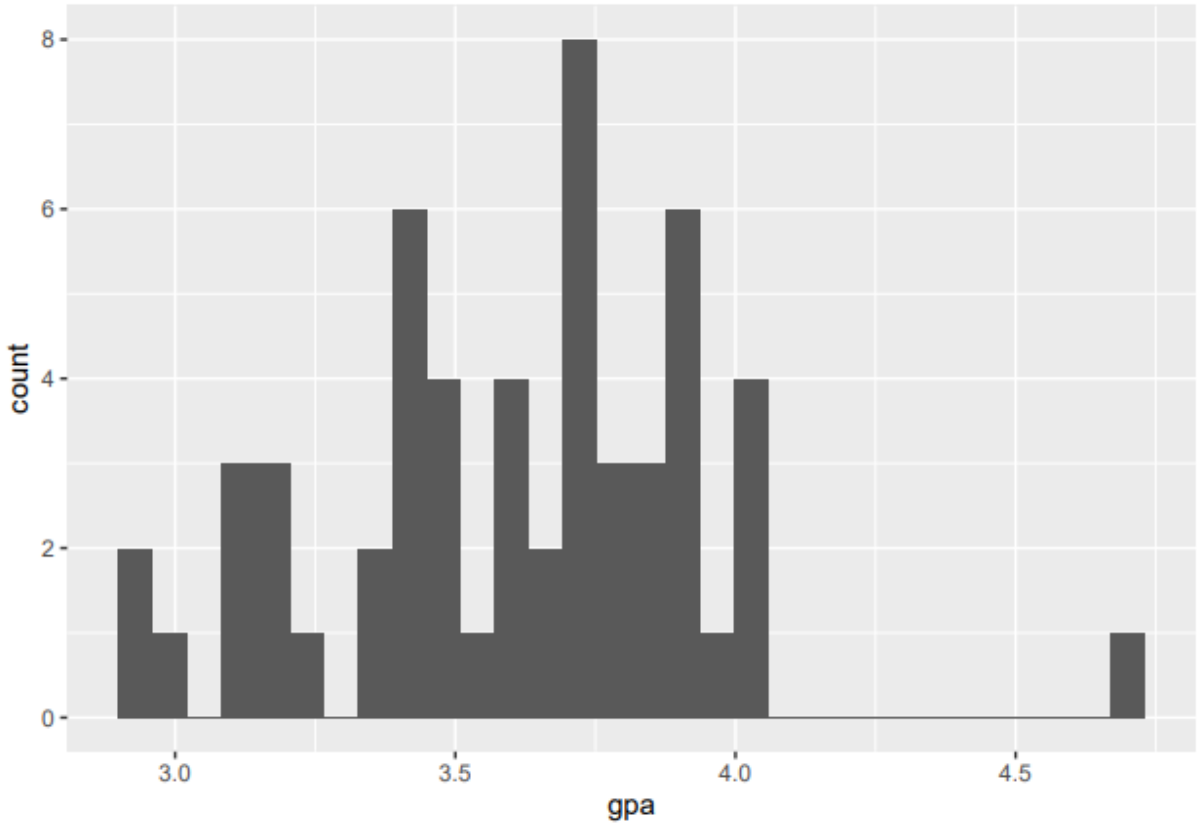
Appendix

```
# box plot of female gpa vs male gpa

female <- gpa %>%
  filter(gender == "female")
male <- gpa %>%
  filter(gender == "male")
boxplot(male$gpa, female$gpa,
        horizontal = TRUE,
        names = c("male","female"),
        xlab = "Gender",
        main = "GPA distribution of genders")
```



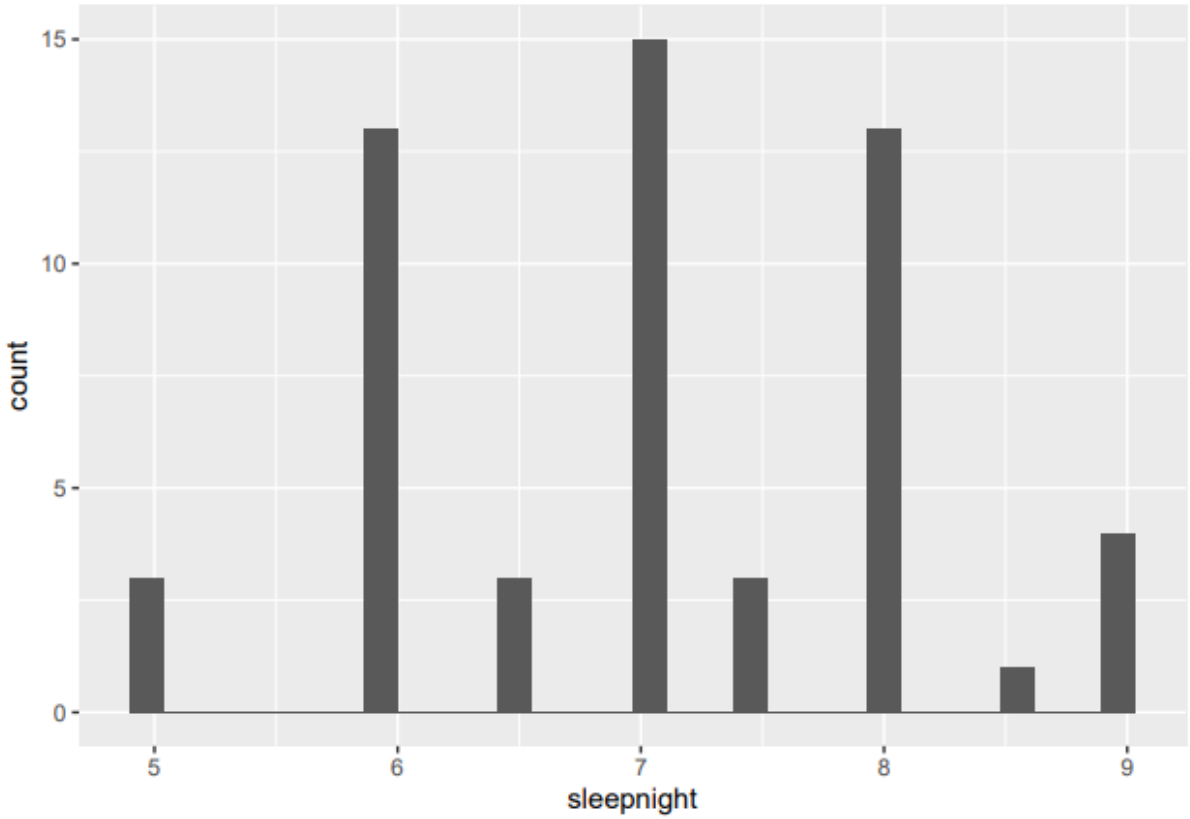

```
# histograms
ggplot(gpa, aes(x=gpa)) +geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



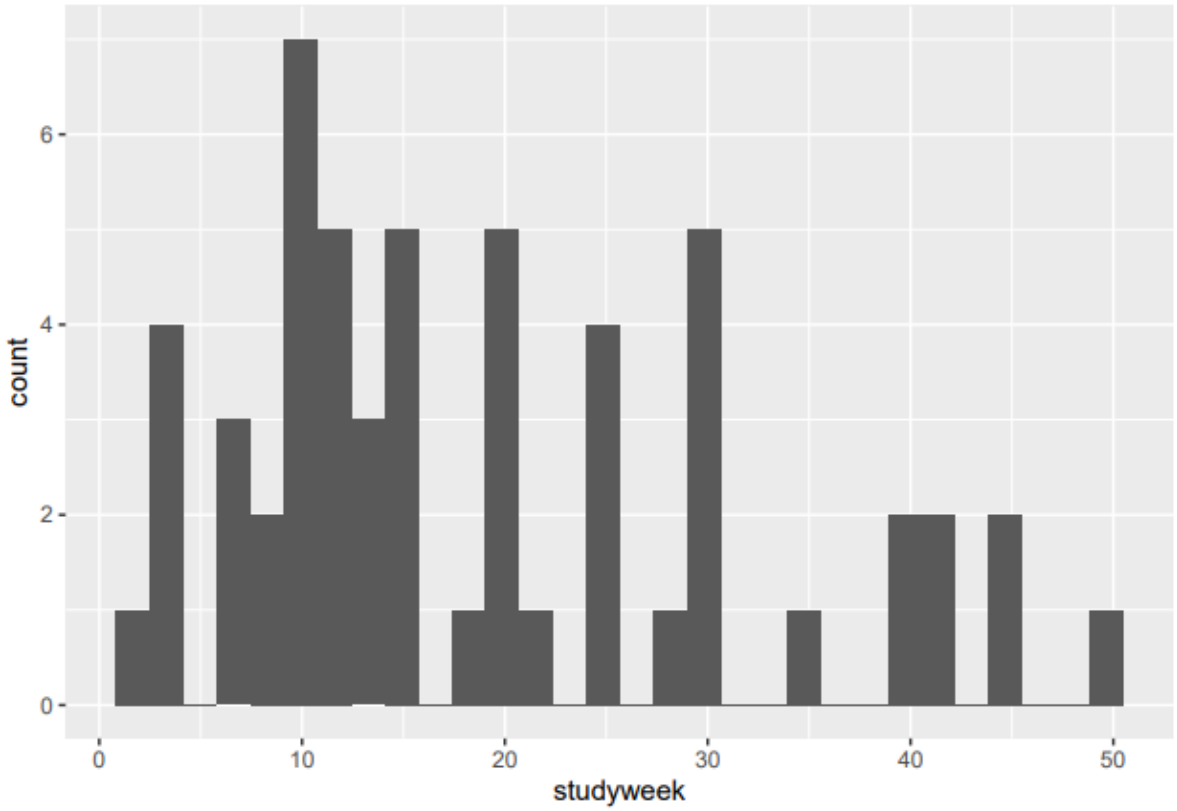

```
ggplot(gpa, aes(x=sleepnight)) +geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



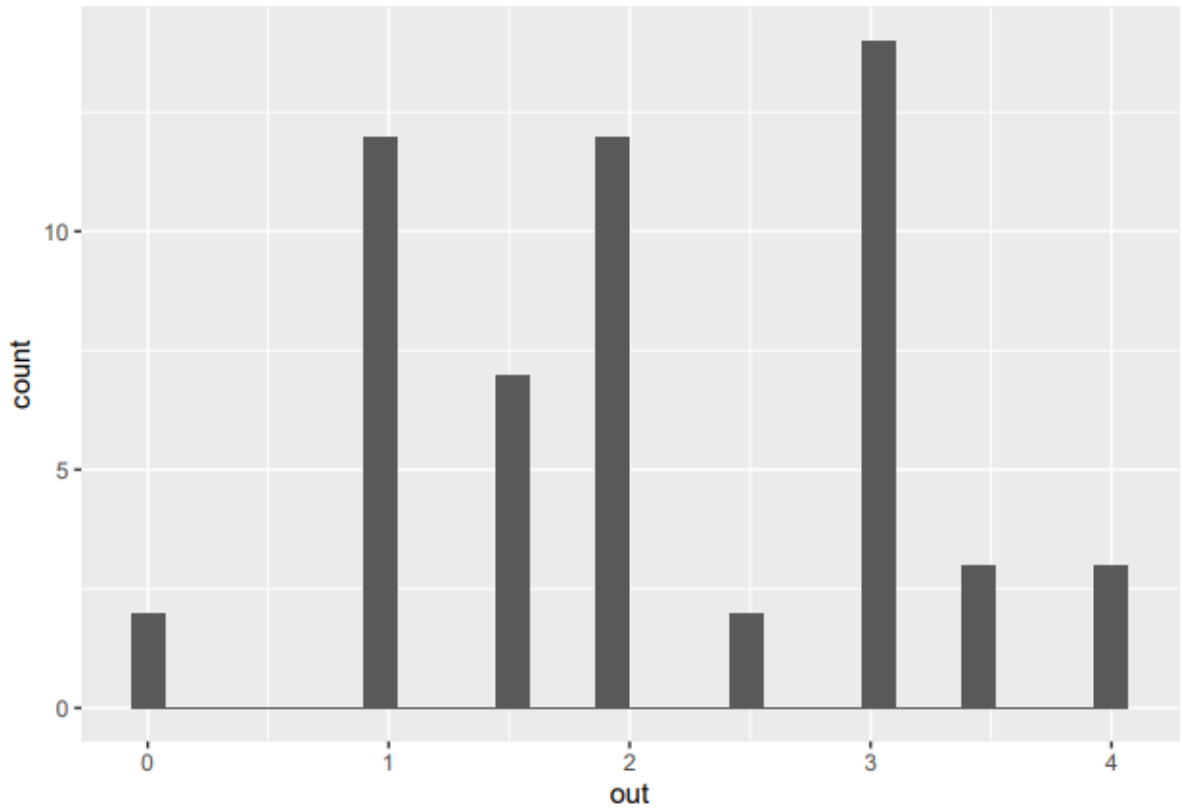

```
ggplot(gpa, aes(x=studyweek)) +geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

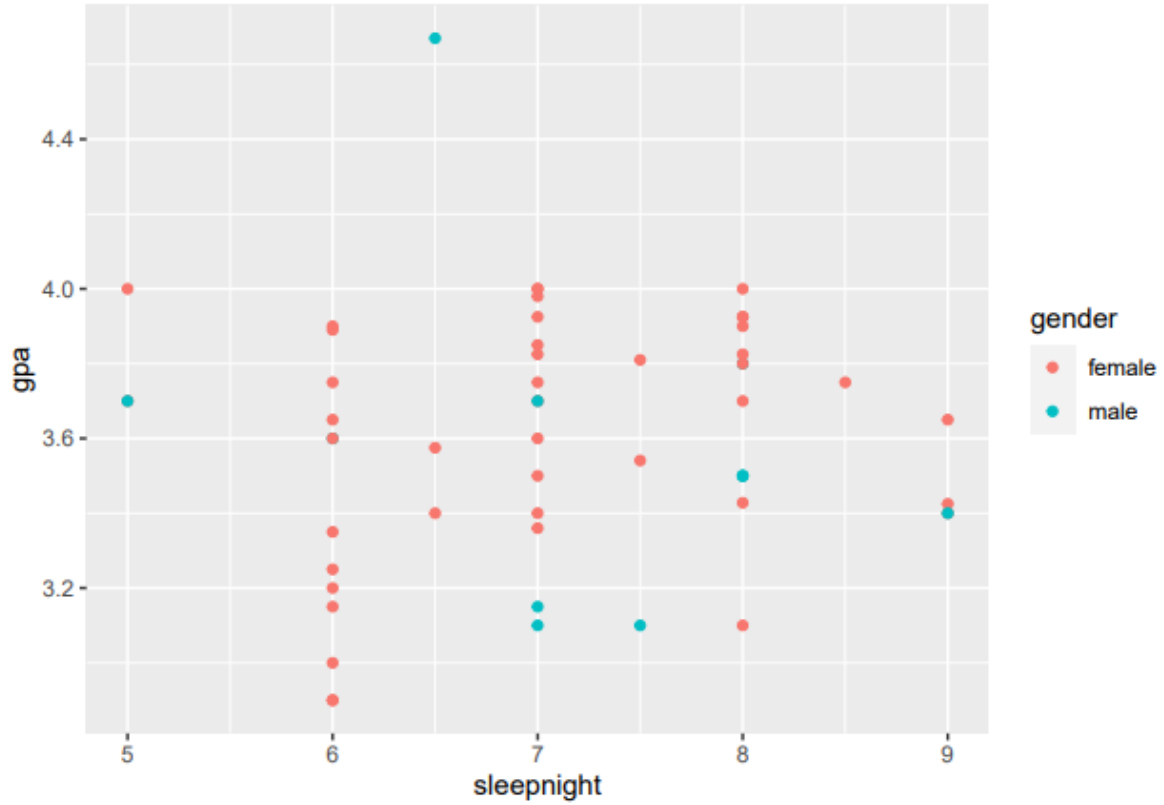



```
ggplot(gpa, aes(x=out)) +geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# scatter plot of sleep per night vs gpa
ggplot(data = gpa, aes(x = sleepnight, y = gpa,
                       color = gender)) +
  geom_point()
```




```
#bar plot
#barplot(gpa,
        #main = "Grouped barchart",
        #xlab = "gender", ylab = "count",
        #col = c("gpa", "studyweek", "sleepnight", "out"),
        #legend.text = rownames(gpa),
        #beside = TRUE)
```

```
# difference of two means t-test of gender related to GPA
x_bar_female <- mean(female$gpa)
x_bar_male <- mean(male$gpa)
x_bar_diff <- x_bar_female-x_bar_male
n_female <- 43
n_male <- 12
female <- gpa %>%
  filter (gender == "female")
male <- gpa %>%
  filter (gender == "male")
S_female <- sd(female$gpa)
S_male <- sd(male$gpa)
df = n_male-1
SE <- sqrt((((S_female)^2)/n_female)+(((S_male)^2)/n_male))
t <- (x_bar_diff)/SE
pval <- 2*(1-pt(t,df))
t
```

```
## [1] 0.391907
```

```
pval
```

```
## [1] 0.7026131
```

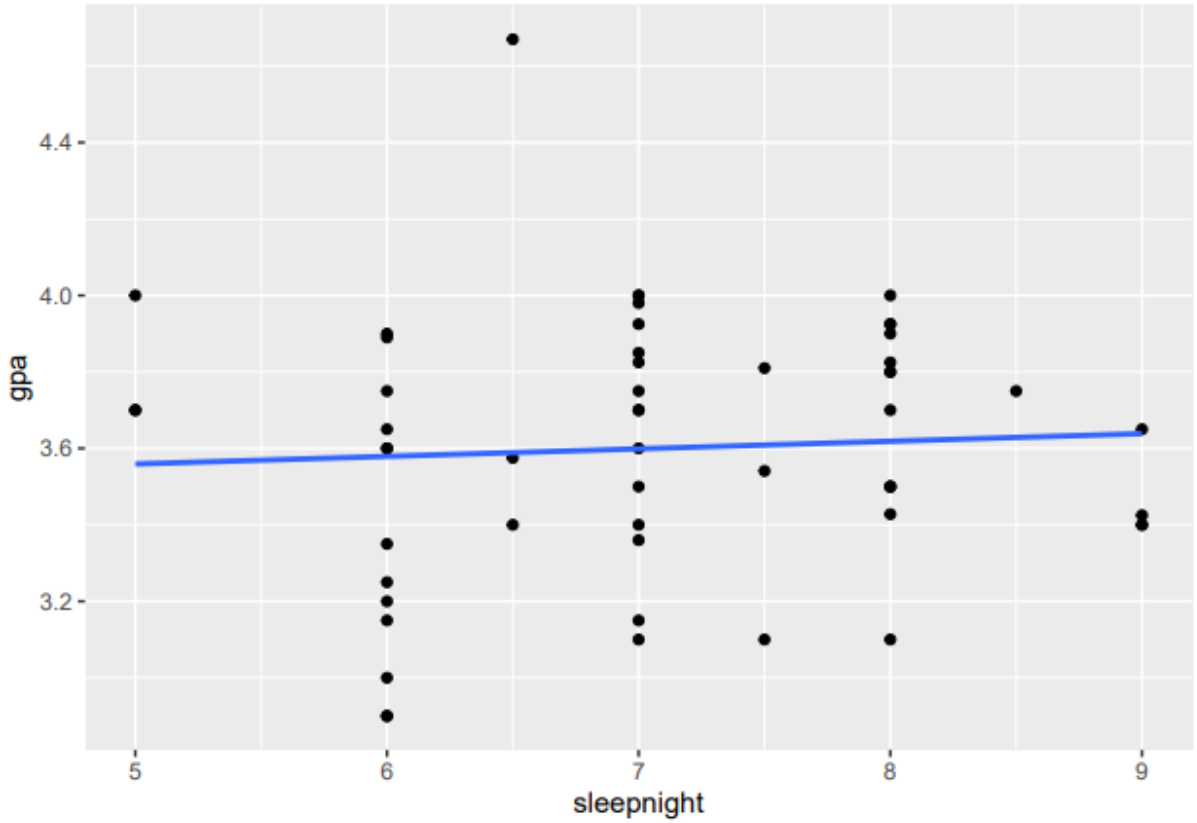
```
# confidence interval gender related to GPA
alpha <- 0.05
cl <- 0.95
t_star <- qt(cl+alpha/2,df)
CI_upper <- x_bar_diff + t_star*SE
CI_lower <- x_bar_diff - t_star*SE
```

```
#calculating linear regression statistics
gpa %>%
  summarise(S_x = sd(sleepnight),
            S_y = sd(gpa),
            r = cor(sleepnight,gpa),
            x_bar = mean(sleepnight),
            y_bar = mean(gpa))
```

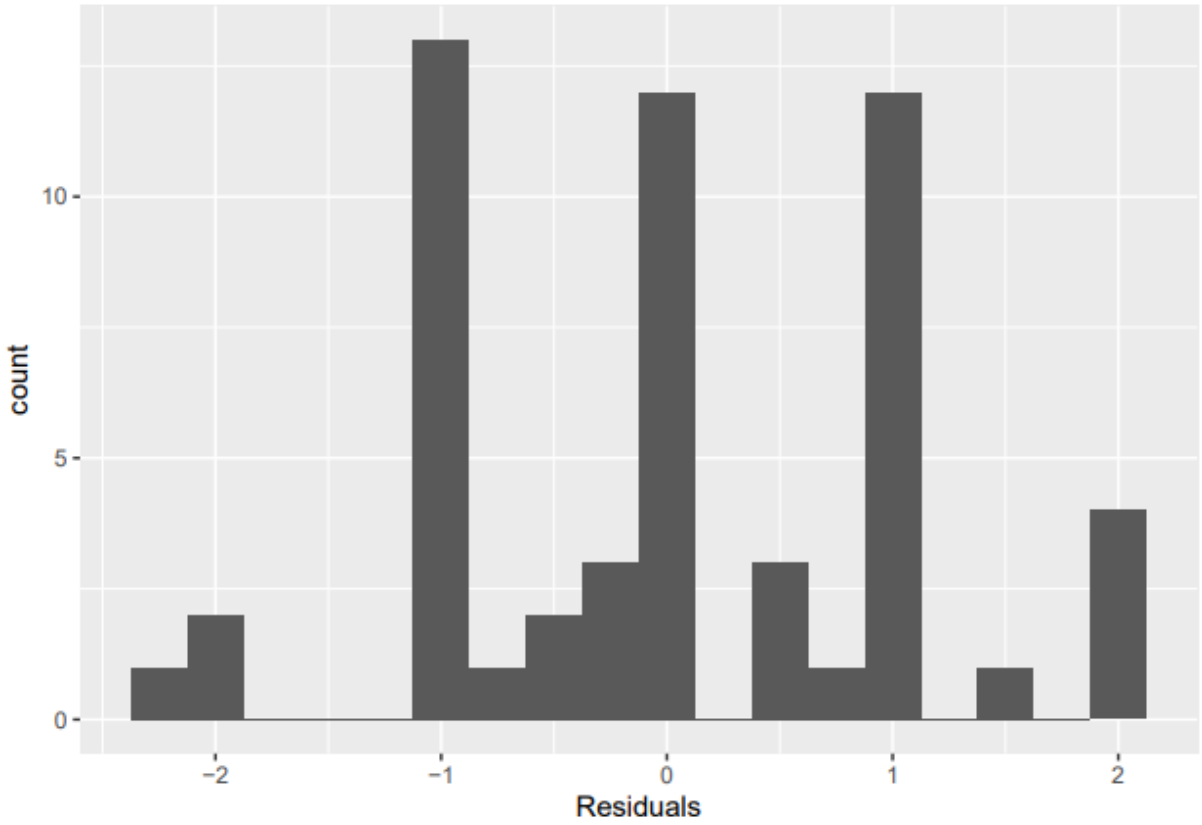
```
## # A tibble: 1 x 5
##     S_x   S_y      r x_bar y_bar
##   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1  1.03 0.336 0.0610  7.06  3.60
```

```
#compute slope, intercept, and R^2
S_x <- 1.032143
S_y <- 0.3356183
r <- 0.6098308
x_bar <- 7.063636
y_bar <- 3.600073
b_1 <- (S_y/S_x)*r
b_0 <- y_bar-b_1*x_bar
R_squared <- r^2

#plot the data with the regression line
ggplot(gpa, aes(sleepnight, gpa)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

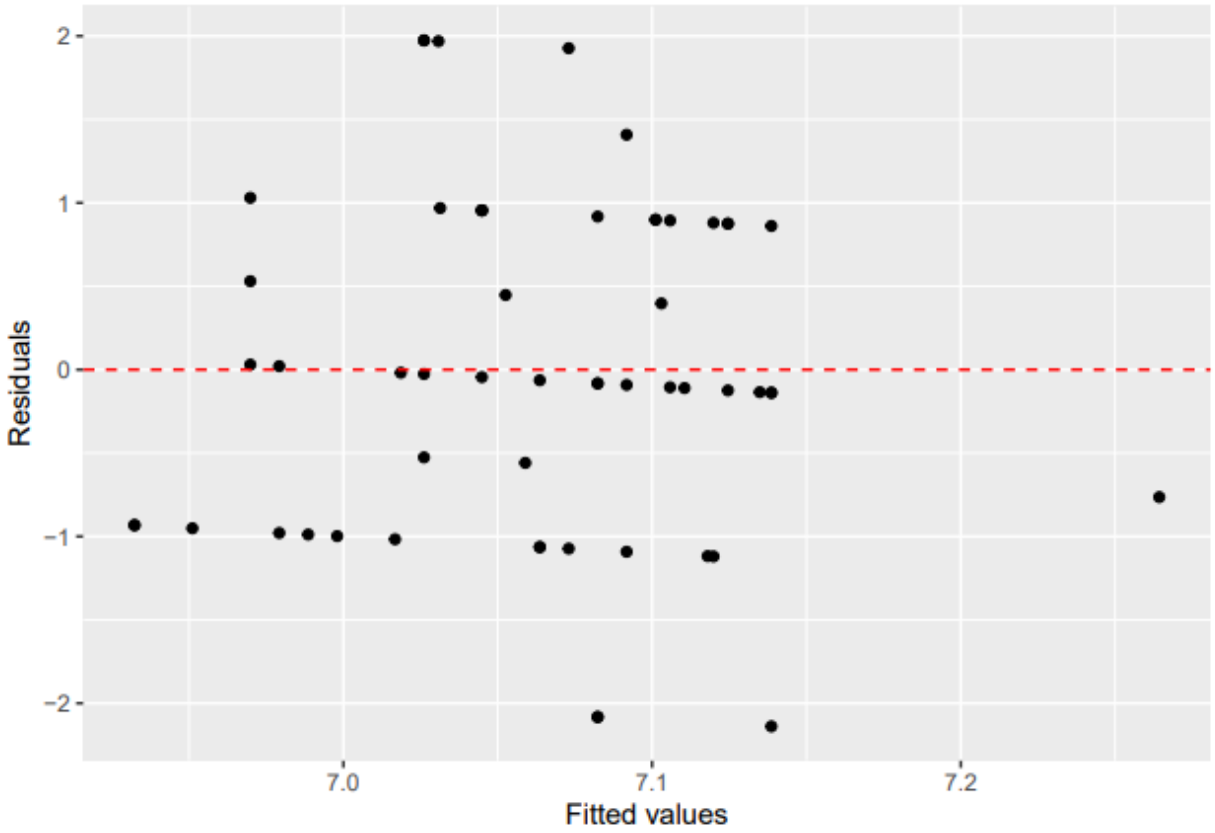
```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
# preform t-test of linear regression, and determine p-value
n <- 55
DF <- n-2
se <- sqrt(1-R_squared)*S_y
T <- b_1/SE
p_value <- 2*(1-pt(T,DF))

# confidence intervel
CL <- 0.95
alpha1 <- 0.05
T_star <- qt(CL+(alpha1/2),DF)
ci_upper <- b_1+T_star*se
ci_lower <- b_1-T_star*se

#residual visualization
m <- lm(sleepnight ~ gpa, data = gpa)

m_aug <- augment(m)
ggplot(data = m_aug, aes(x = .resid)) +
  geom_histogram(binwidth = 0.25) +
  xlab("Residuals")
```




```
ggplot(data = m_aug, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  xlab("Fitted values") +
  ylab("Residuals")
```


Residuals
2
1
0
−1
−2
7.0
7.1
7.2
Fitted values

References

Conger, D., & Long, M. C. (n.d.). *Why Are Men Falling Behind? Gender Gaps in College Performance and Persistence.* Pennsylvania State University.

Creswell, J. D., Tumminia, M. J., Price, S., Sefidgar, Y., Cohen, S., Ren, Y., . . . Lovett, M. C. (2013). Nightly Sleep Duration Predicts Grade Point Average in the First Year of College. *PNAS*.

Pace, C. (2018). *Are There Consequences to the Behaviros of a College Kid?* Retrieved from RPubs: https://rpubs.com/Chrissy_P/387065