Stat 198 Final Project

Introduction and Data:

Our group wanted to explore how the COVID-19 pandemic affected the movement patterns for people around the world. We wondered if people became more stationary and spent more time in their homes or if they became more exploratory and inclined to spend more time outside. Our team chose to focus on mobility trends around the world during the progression of the COVID-19 pandemic. Our central research question was the following:

How did the progression of the COVID-19 pandemic affect mobility trends in different countries?

We used two datasets to conduct a statistical analysis to answer this overarching question. The first dataset we used was the **time_series_covid19_confirmed_global.csv** file which contains COVID-19 epidemiological data from January 22, 2020 to July 24, 2020. This data was compiled by the Johns Hopkins University Center for Systems Science and Engineering from various sources including the World Health Organization, DXY.cn, BNO News, National Health Commission of the People's Republic of China, China CDC, Hong Kong Department of Health, Macau Government, Taiwan CDC, US CDC, Government of Canada, Australia Government Department of Health, European Centre for Disease Prevention and Control, Ministry of Health Singapore, and others. Each of these organizations collected the data in accordance with their country's data collection policies during the pandemic. The variables in this dataset are listed below with their corresponding description:

Province/state (value is name of province/state, blank if not applicable)

Country/region (value is name of country/region)

Lat (value is latitude number for location)

Long (value is longitude number for location)

1/22/20 (example of a date variable, columns exist for each date up until 7/24/20, and value is number of cases on specified date)

The second dataset used in this analysis was the **applemobilitytrends-2020-04-14-1.csv** file which contains information on the COVID-19 mobility trends in countries/regions/cities based on daily requests for directions in Apple Maps. The data shows a relative volume of directions requests per country/region/city compared to a baseline volume on January 13th, 2020. The dataset's timeline ranges from January 13th, 2020 to April 14th, 2020 and one day is defined as midnight-to-midnight, Pacific standard time. Data that is sent from users' devices to the Apple Maps service is associated with random, rotating identifiers, meaning Apple did not have a profile of an individuals' movements and searches. Additionally, Apple Maps has no demographic information about our users, meaning no statements can be made about the representativeness of usage against the overall population. The variables in this dataset are listed below with their corresponding description:

geo_type (value is either country/region or city)

region (value is name of country/region or city)

transportation_type (value is either driving, walking, or transit)

2020-01-13 (example of a date variable, columns exist for each date up until 2020-04-14, and value is mobility rate score relative to a baseline score of 100 on January 13th, 2020)

Methodology:

In order to address our research question about how the progression of COVID-19 affected mobility trends in different countries, we decided to focus on three different countries: The United States, South Korea, and New Zealand. These three countries were chosen based on relative news trends in the past few months and their differing policies during the pandemic and quarantining. We then chose to manipulate our raw datasets into two new datasets to use for our data visualizations and statistical tests. The original datasets were manipulated in Google sheets so that only the overlapping dates from the original two datsets were used (Jan 22nd - April 14th). This manipulation resulted in the datasets below:

Manipulated Dataset 1 - combined.csv:

A new dataset that includes the mobility score for each transportation type per day along with the number of covid cases per day for each country of interest.

Variables:

day (value is the date in the format of year-month-day)

driving_us (value is driving mobility score for U.S.)

walking_us (value is walking mobility score for U.S.)

transit_us (value is transit mobility score for U.S.)

cases_us (value is the number of cases for U.S.)

driving_korea (value is driving mobility score for South Korea)

walking_korea (value is walking mobility score for South Korea)

cases_korea (value is the number of cases for South Korea.)

driving_newzealand (value is driving mobility score for New Zealand)

walking_newzealand (value is walking mobility score for New Zealand)

transit_newzealand (value is transit mobility score for New Zealand)

cases_newzealand (value is number of cases for New Zealand)

Manipulated Dataset 2 - regressiondata.csv:

A new dataset that includes the mobility score for only driving and walking per day, the number of covid cases per day for each country of interest, and a combined mobility score variable for driving and walking per day.

Variables:

Driving_mobility_score (value is the mobility score for driving)

Walking_mobility_score (value is the mobility score for walking)

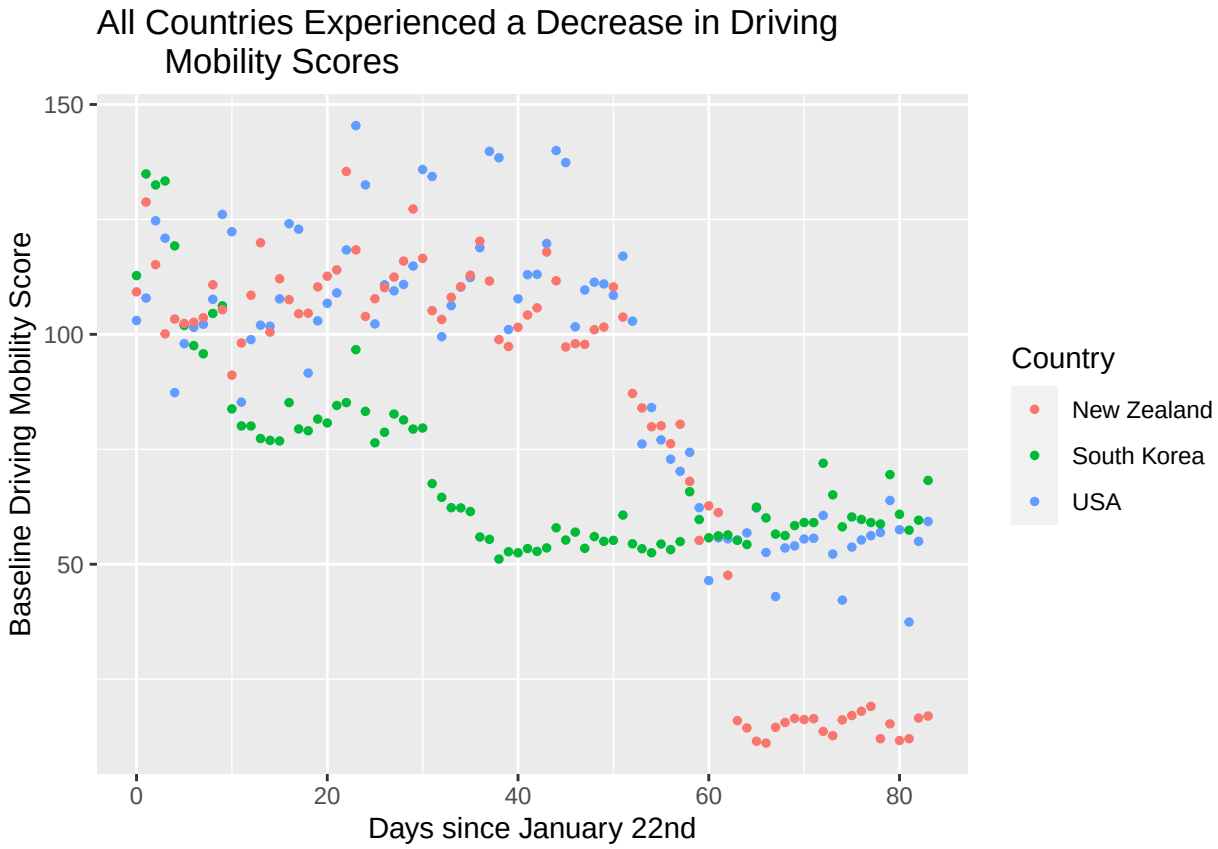
Number_of_cases (value is number of COVID-19 cases)

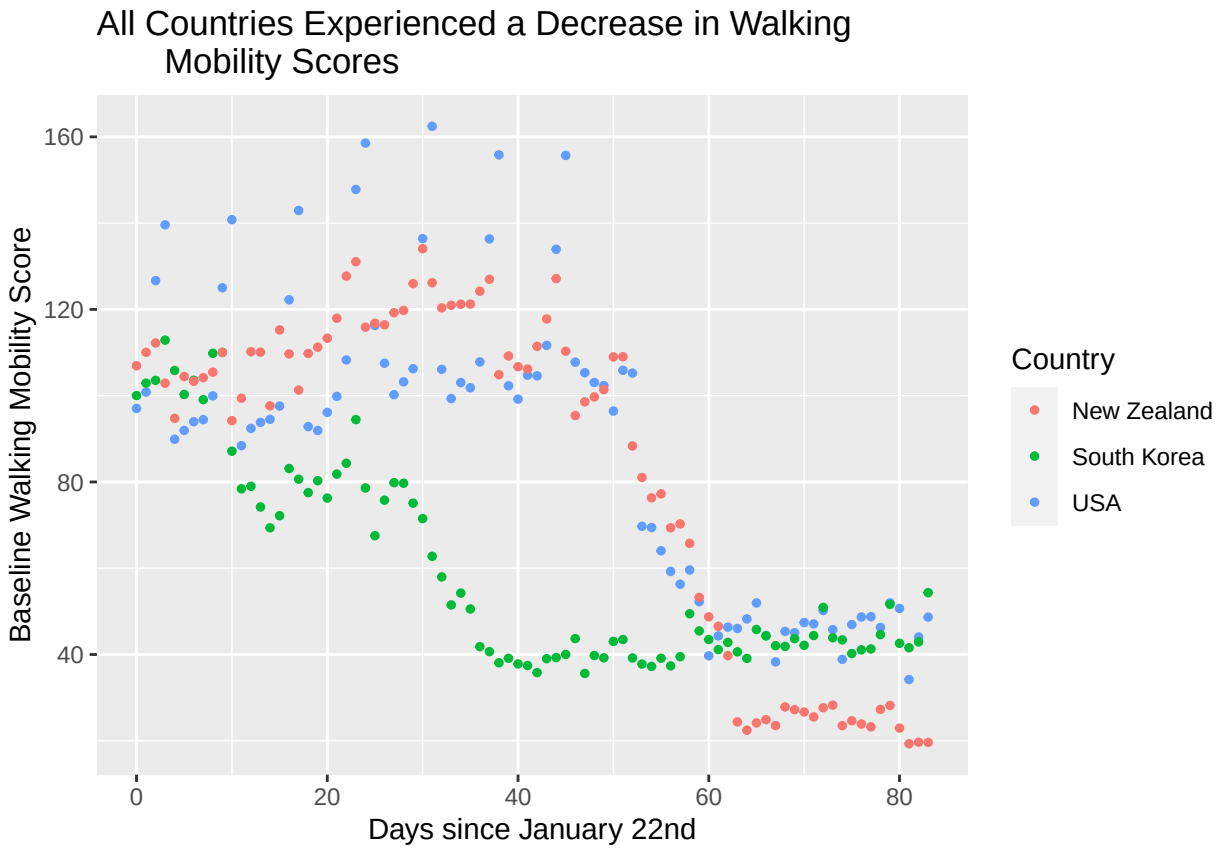
Country (value is name of country)

Days_since_January_22nd (value is number of days since January 22nd)

Our group then chose to create data visualizations for the U.S., South Korea, and New Zealand. This allowed us to analyze trends in the data and determine the next steps in conducting tests for our statistical analysis. We chose to use the mobility score for walking and driving to display the mobility trends across the three countries. It should be noted that public transit is controlled by governments, so public transit may not have been operational during the pandemic in a lot of area. Thus, this mobility score was not included in any of our analyses.

The scatterplots are shown below.





These scatterplots display the walking and driving mobility scores on different days in 2020, for our three countries of interest (United States, South Korea, and New Zealand). In the scatter/line plots, most countries seem to have somewhat steady mobility scores for a certain period of time, then undergo a steep drop off, and then once again become somewhat steady. Our group hoped to explore potential reasons for the shapes of these curves. Could it be because of COVID? Is there a certain date range where the populations begin to see a large drop off in mobility scores? To further explore this interesting trend, we performed various statistical tests.

The first statistical test we chose to perform was a two-sample t-test. We wanted to determine whether there was evidence that suggested that the declaration of a state of emergency due to COVID-19 would affect mobility. We chose to focus on the U.S. when conducting this test, noting that March 1st, 2020 was the date where the U.S. declared a national emergency. Our question of interest was the following:

Did mobility trends change after March 1st in the U.S.?

We compared two means in this two-sample t-test, the first being the mean mobility score before March 1 and the second being the mean mobility score after March 1. The tests' summary statistics for the mean driving mobility score and mean walking mobility score in the U.S. are shown below.

```
##
##  Welch Two Sample t-test
##
## data:  driving_us_pre and driving_us_post
## t = 7.733, df = 65.571, p-value = 4.134e-11
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  29.21172      Inf
## sample estimates:
## mean of x mean of y
```

```
## 112.38775  75.13955
```

```
##
##  Welch Two Sample t-test
##
## data:  walking_us_pre and walking_us_post
## t = 7.8117, df = 77.492, p-value = 1.13e-11
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  34.94119      Inf
## sample estimates:
## mean of x mean of y
##  111.8118   67.4075
```

In order to account for outliers that may have influenced the mean mobility scores, such as holidays, we also decided to use a Wilcoxon rank-sum test to compare the median mobility scores before and after March 1 for driving and walking in the US. We focused on the same question of interest listed for the two-sample t-test. The tests' summary statistics for the median driving mobility score and median walking mobility score in the U.S. are shown below.

```
## Warning in wilcox.test.default(walking_us_pre, walking_us_post, alternative =
## "two.sided", : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  walking_us_pre and walking_us_post
## W = 1465, p-value = 1.65e-07
## alternative hypothesis: true location shift is not equal to 0
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  driving_us_pre and driving_us_post
## W = 1454, p-value = 6.533e-08
## alternative hypothesis: true location shift is not equal to 0
```

```
## [1] 102.66
```

```
## [1] 51.32
```

```
## [1] 109.24
```

```
## [1] 61.415
```

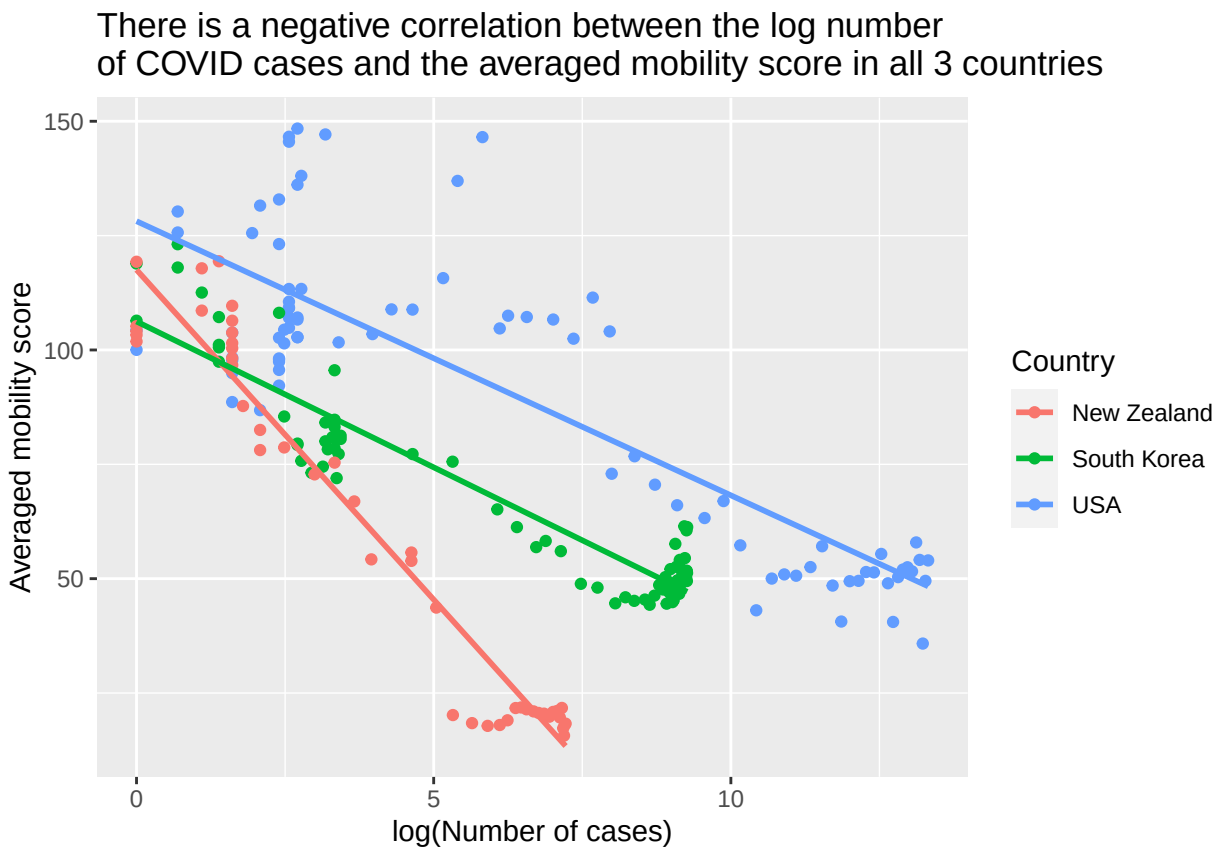
Based on our results from comparing the means and medians before and after March 1st in the U.S., we had reason to believe that mobility trends did change during the COVID-19 pandemic. This was also in line with what we saw in the scatterplots, where we saw a general decrease in mobility scores (although there were some initial increases). These t-tests allowed us to make further conclusions about this drop off in mobility scores, as we created two populations of mobility scores based on whether the data occurred before the national state of emergency in the US or afterwards. Still, the t-tests should be taken with a grain of salt as there may have been some issues with independence of samples. Because of this, our group decided to explore a linear regression model using COVID-19 cases as a predictor for mobility score to see if we can learn more about the relationship between mobility trends and COVID-19. More specifically, our group wanted to see how the combined mobility score variables changed as the number of cases progressed in different countries. For our linear regression, we realized when checking our assumptions that using the number of cases in our model as a predictor would not work, because the graph was not even close to linear. Thus, we performed a linear regression using the logarithm of the number of COVID-19 cases in a country as

the predictor. We chose to focus on our three selected countries: the U.S., South Korea, and New Zealand and focused on the following question of interest:

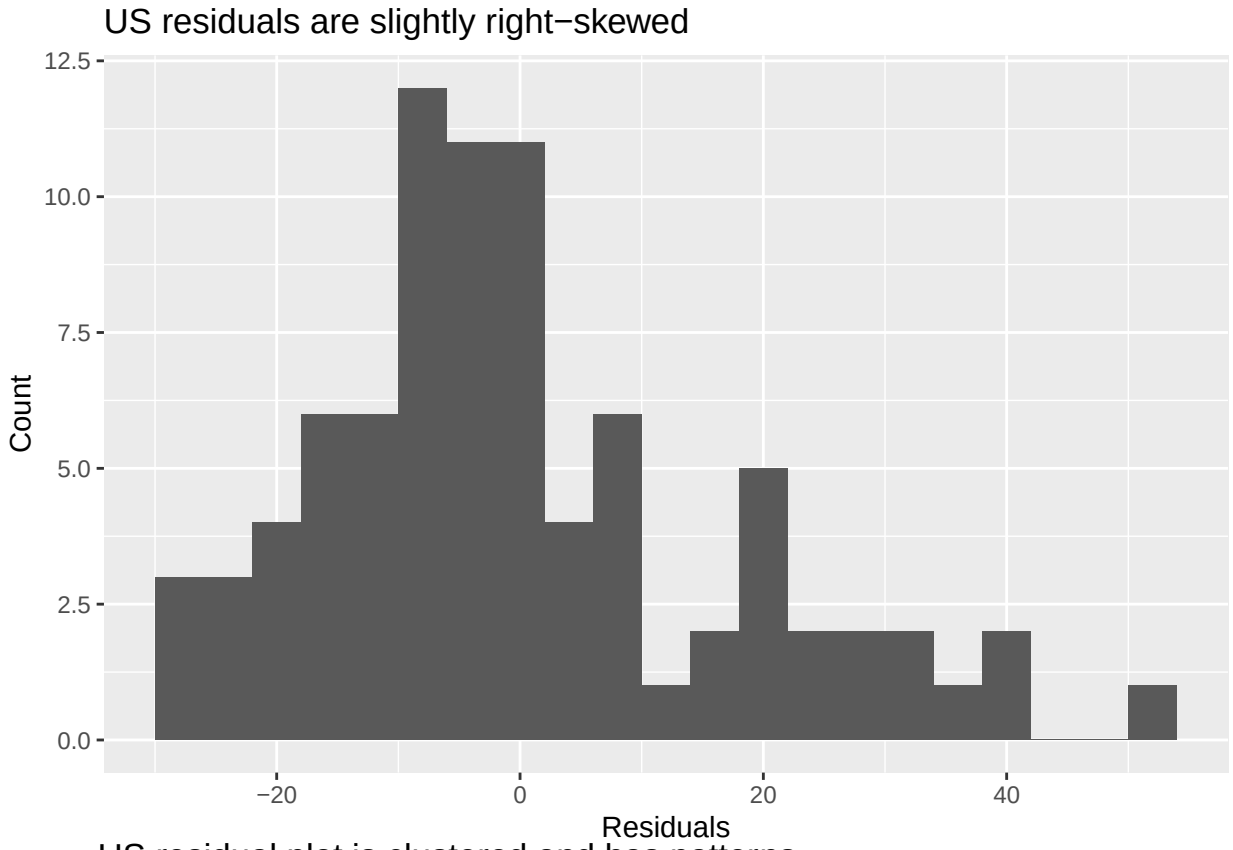
How did the number of COVID-19 cases affect mobility trends in the U.S., South Korea, and New Zealand?

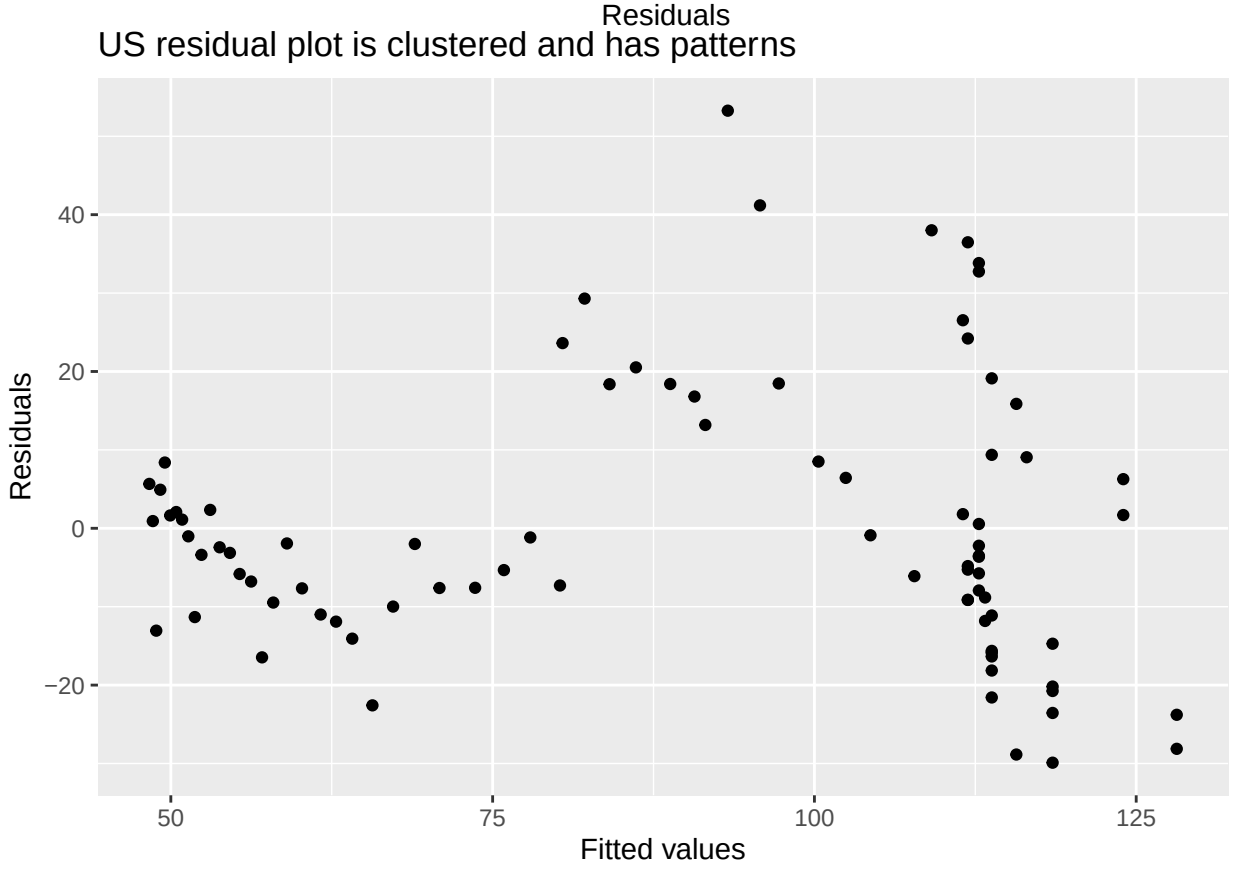
The regression visualization and assumptions charts are shown below.

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



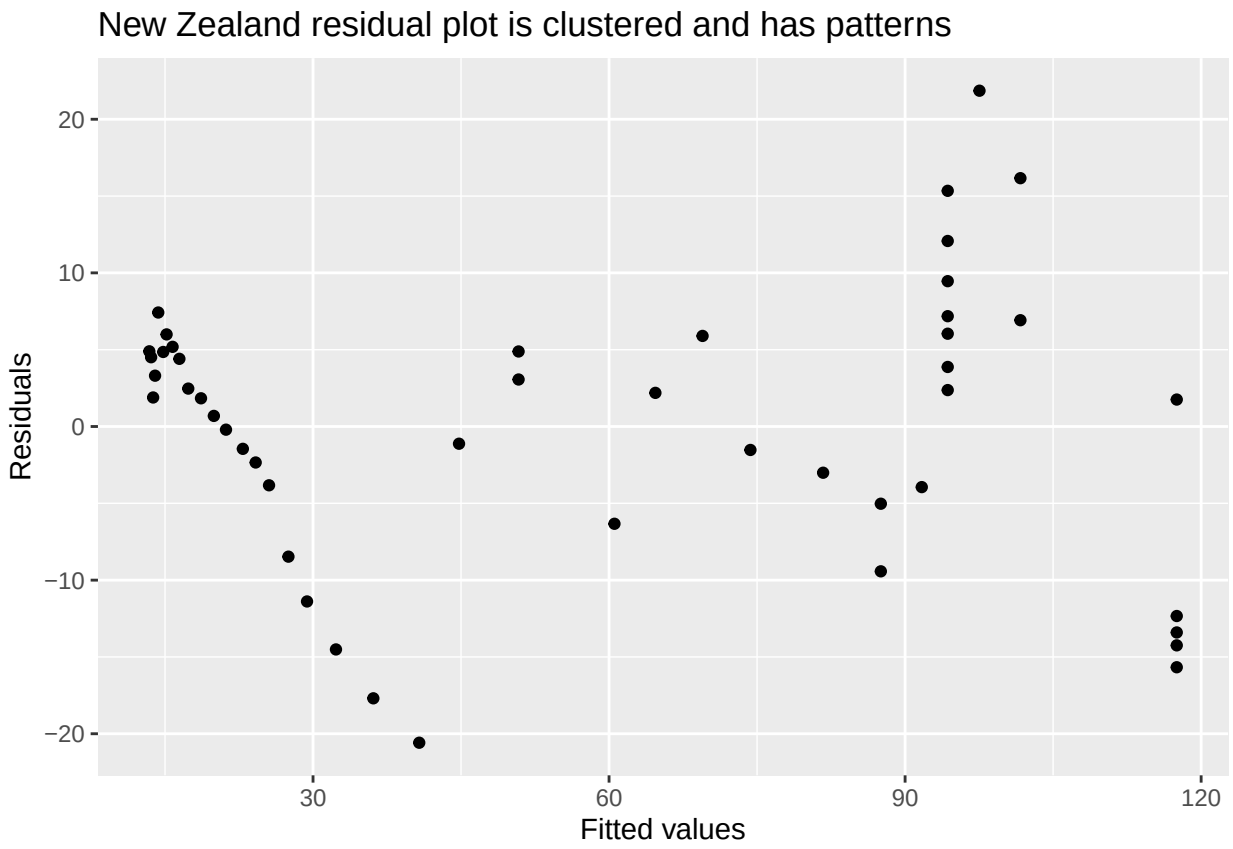

```
## # A tibble: 2 x 5
##   term                estimate std.error statistic  p.value
##   <chr>                  <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)            128.       3.25       39.4 4.74e-55
## 2 log(Number_of_cases)    -5.99     0.425     -14.1 1.28e-23
```

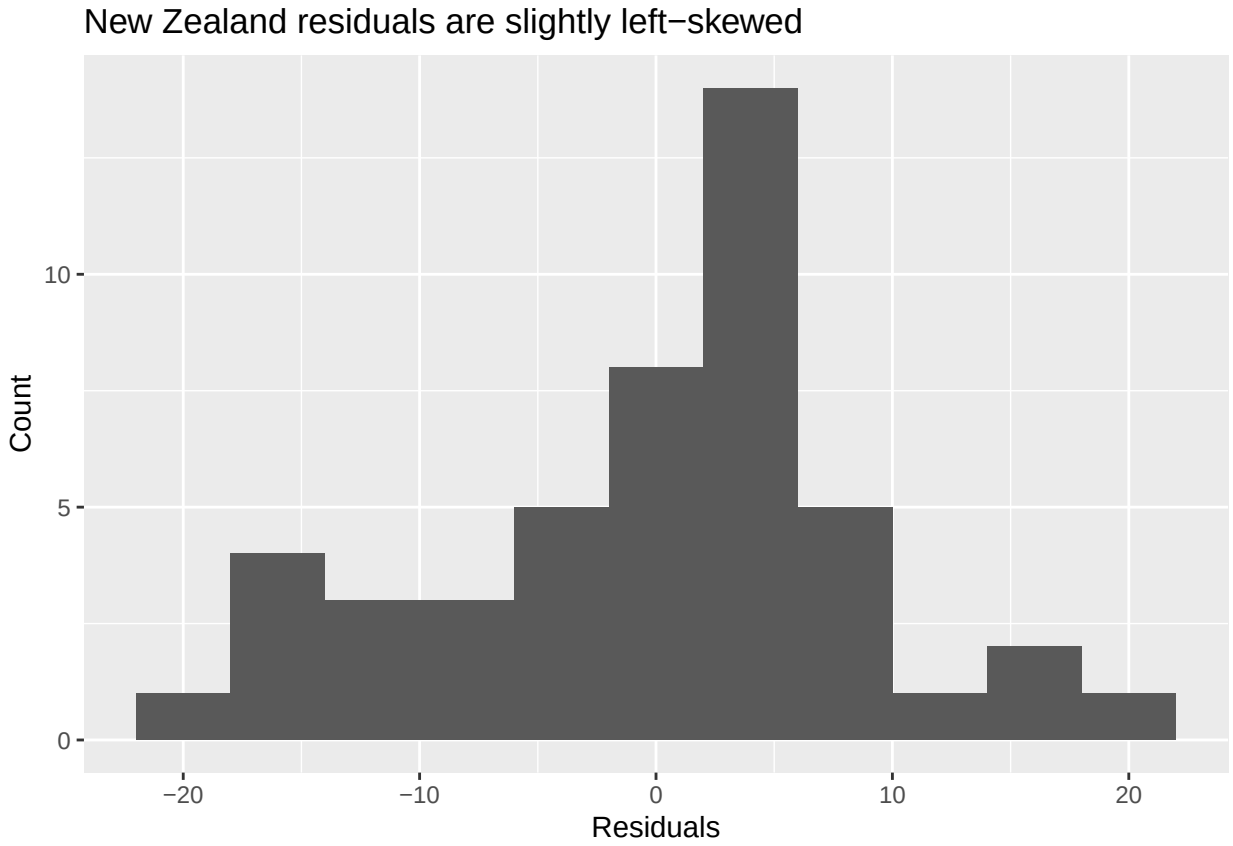

US residuals are slightly right−skewed
Count
Residuals
12.5
10.0
7.5
5.0
2.5
0.0
−20
0
20
40


US residual plot is clustered and has patterns
Residuals
Fitted values
40
20
0
−20
50
75
100
125

```
## # A tibble: 2 x 5
```

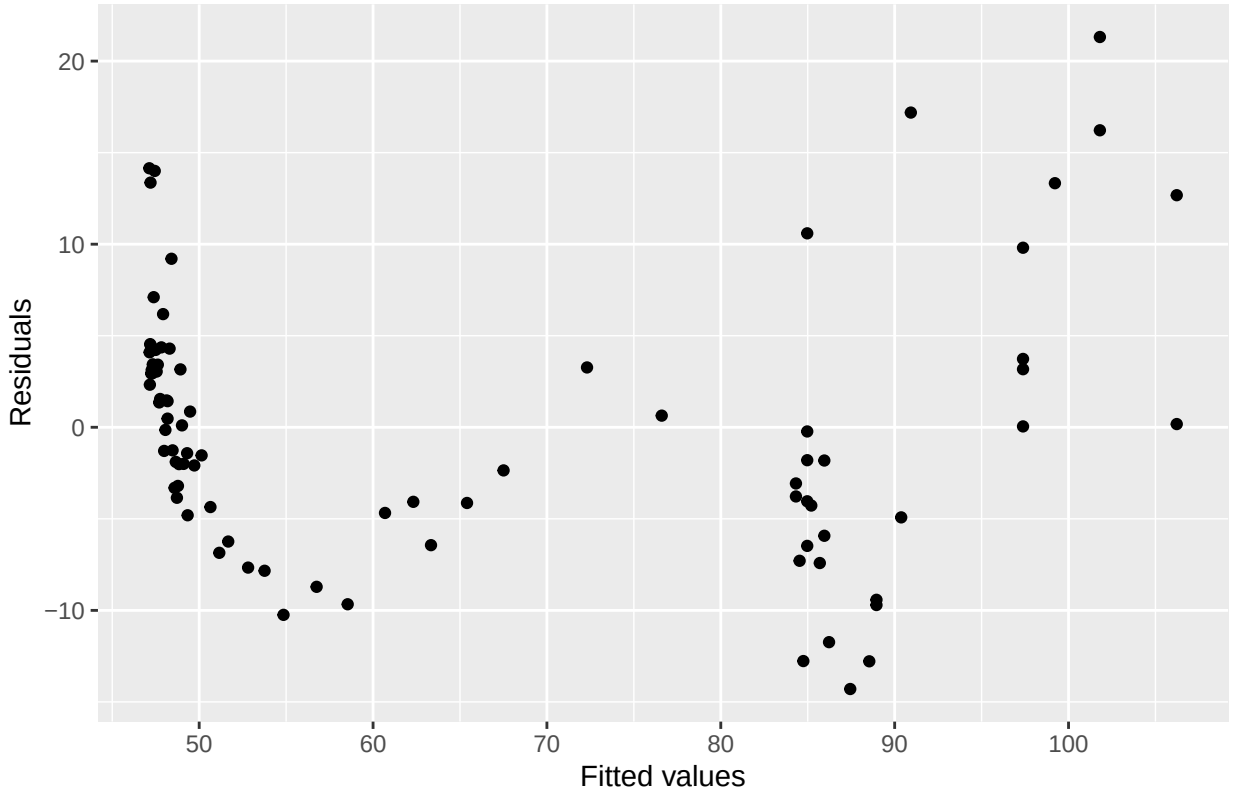
```
##   term                 estimate std.error statistic  p.value
##   <chr>                   <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)             118.      2.50       47.1 6.19e-40
## 2 log(Number_of_cases)    -14.4     0.516     -28.0 4.44e-30
```



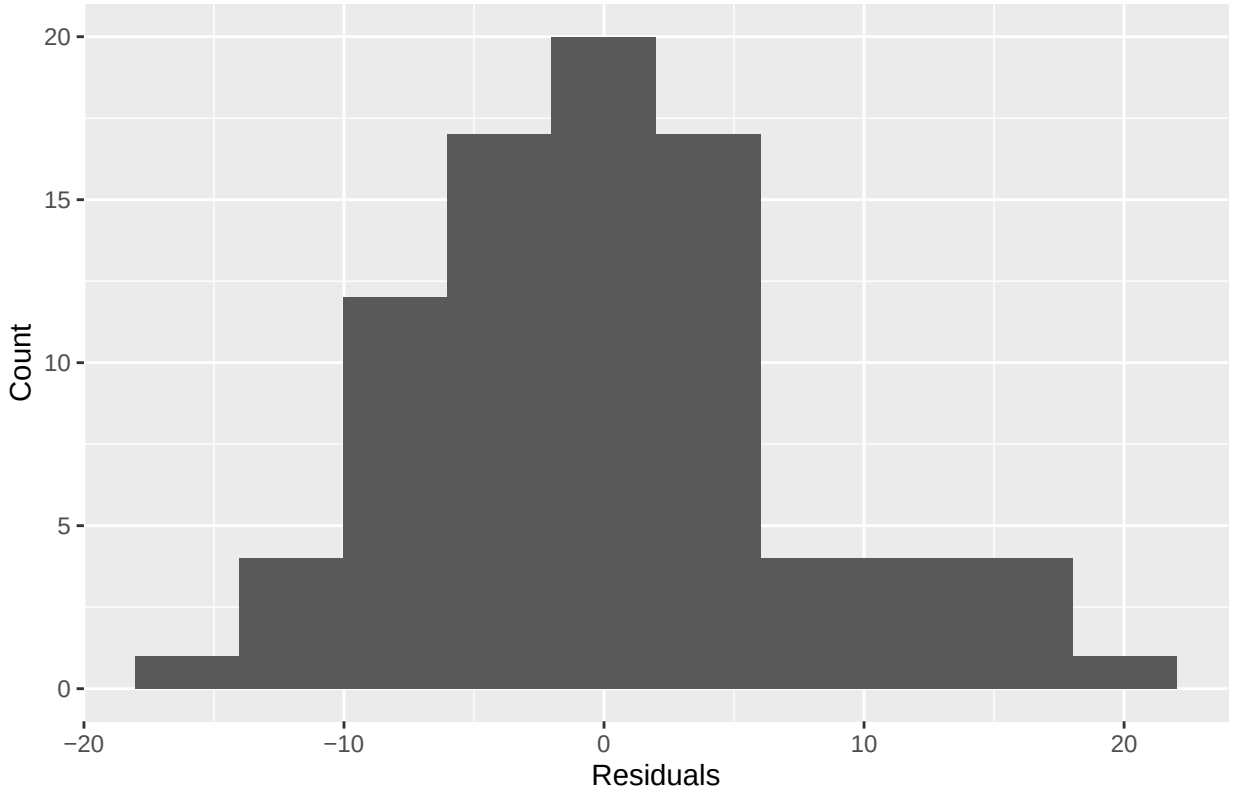



```
## # A tibble: 2 x 5
##   term                 estimate std.error statistic  p.value
##   <chr>                   <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)            106.       1.83       58.1 2.08e-68
## 2 log(Number_of_cases)    -6.38     0.254     -25.1 3.26e-40
```

South Korea residual plot is clustered and has patterns



South Korea residuals are roughly symmetric, but not normal



Results:

T-test results:

Null Hypothesis: the population mean of driving mobility trends before a national state of emergency was declared due to COVID-19 is equal to the population mean of mobility trends after a national state of emergency was declared due to COVID-19

Alternative Hypothesis:the population mean of driving mobility trends before a national state of emergency was declared due to COVID-19 is NOT equal to the population mean of mobility trends after a national state of emergency was declared due to COVID-19

Null Hypothesis: the population mean of walking mobility trends before a national state of emergency was declared due to COVID-19 is equal to the population mean of mobility trends after a national state of emergency was declared due to COVID-19

Alternative Hypothesis: the population mean of walking mobility trends before a national state of emergency was declared due to COVID-19 is NOT equal to the population mean of mobility trends after a national state of emergency was declared due to COVID-19

To conduct our T-test, we created new populations: pre-state of national emergency in the US (before Mar. 1) and after state of national emergency in the US (Mar 2. onwards). We found that there was a statistically significant difference in both the mean walking and driving mobility scores between these two populations. With an alpha of 0.025 (using a Bonferroni correction) and p-values of 2.26e-11 and 4.134e-11 for walking and driving respectively, we were able to reject the null hypotheses. Due to the significance of our p-values, there is evidence that pre-March 1st mean was lower than the post-March 1st mean for both driving and walking scores. However, it is important to note that our assumptions for these tests might not have been entirely met, namely independence. This could impact the validity of this test.

Wilcoxon rank sum test:

Null Hypothesis: the population median of driving mobility trends before a national state of emergency was declared due to COVID-19 is equal to the population median of mobility trends after a national state of emergency was declared due to COVID-19

Alternative Hypothesis:the population median of driving mobility trends before a national state of emergency was declared due to COVID-19 is NOT equal to the population median of mobility trends after a national state of emergency was declared due to COVID-19

Null Hypothesis: the population median of walking mobility trends before a national state of emergency was declared due to COVID-19 is equal to the population median of mobility trends after a national state of emergency was declared due to COVID-19

Alternative Hypothesis: the population median of walking mobility trends before a national state of emergency was declared due to COVID-19 is NOT equal to the population median of mobility trends after a national state of emergency was declared due to COVID-19

In order to see if median mobility scores had changed during the COVID-19 pandemic, we conducted 2 Wilcox rank sum tests on the walking and driving mobility scores in the US, using the same populations as we did in the t-tests. We found that there was a statistically significant difference in both the median walking and driving mobility scores. With a alpha of 0.025 (using a Bonferroni correction) and p-values of 1.65e-7 and 6.533e-8 for walking and driving respectively, we were able to reject the null hypotheses. Due to the significance of our p-values, there is evidence that pre-March 1st median was lower than the post-March 1st median for both driving and walking scores. Similar to our t-tests, the assumption of independence was not met, possibly impacting the validity of this test.

Linear Regression:

Since we used a logarithmic transformation, some of our data could not be used, as log(0) is undefined. So, the data used for these regression starts when a country records its first case. The purpose of this linear

regression was to see how the number of COVID-19 cases has affected different countries' mobility scores. So, our analysis does not require observations when the number of cases in a country is equal to 0 since we already have this baseline established (as our dataset uses a pre-COVID baseline) In this test, we averaged a country's walking and driving mobility scores for each day. Our model found that all three countries exhibited a negative correlation between the logarithm of the number cases and their respective mobility scores. New Zealand's mobility score was the most responsive to their logarithm of their number of cases, as our model would expect New Zealand's mobility score to decrease by 14.4 points for an increase of 1 in the logarithm of the number of cases. The USA's mobility score was the least responsive to their logarithm of their number of cases, as our model would expect USA's mobility score to decrease by 5.99 points for an increase of 1 in the logarithm of the number of cases. While we did find y-intercepts for all three regressions, they are not very meaningful for the purpose of our analysis, as they predict the mobility score when a country records its first case. Lastly, none of the assumptions required for our regression models were entirely satisfied, which would likely impact our model's accuracy and/or validity.

Discussion:

As we tried to answer our overall research question, we first looked to see if walking and driving mobility scores had changed after the declaration of the national emergency in the US. Using a two-sample t-test comparing mobility scores before and after March 1st in the US, we found that there is a statistically significant difference in the mean mobility score for both driving and walking in the US. In order to account for outliers that may have influenced the mean mobility score, such as holidays, we also decided to use a Wilcoxon rank-sum test to compare the median mobility scores in the US. We found that there is a statistically significant difference in the median mobility scores for both driving and walking. After coming to the conclusion that mobility trends had changed during the COVID-19 pandemic, we wanted to see how different countries mobility scores changed as the number of cases progressed. To do this, we ran three linear regressions on data from the US, New Zealand, and South Korea. Since we knew that both walking and driving mobility scores had decreased in the US, we decided to average them into a new mobility score we used for these regressions. Overall, we found that all three countries had a negative correlation between the log(number of cases) and their mobility scores, as the regression produced negative coefficients for all three countries. Out of the three countries, New Zealand's regression model saw the largest expected decrease in average mobility score for an increase of 1 in the log(number of cases). The USA's regression model had the smallest expected decrease in average mobility score for an increase of 1 in the log(number of cases).

Although we were successfully able to perform a variety of statistical tests in order to answer our research questions, there were definitely certain limitations regarding the validity of the data and the results. First off, when considering our datasets, the one that was named "**time_series_covid19_confirmed_global.csv** did not specify how the data was collected. We don't know how representative the data was or if there was bias in the data which could affect our results. In regards to our first question, which dealt with mobility trends for walking and driving before and after a national state of emergency was declared in the US, we performed a series of two sample t-tests and Wilcoxon Rank Sum tests to examine if there was a statistical difference between pre and post national emergency. The assumptions for both of these tests were not fully satisfied; we concluded that the independence assumption was not satisfied due to the fact that our dataset used subsequent days in each sample. For example, maybe two days near each other have similar weather, and this similar weather affects how much people walk and/or drive in the given region. In addition, different policies put in place in different parts of the US could have affected the mobility trends in a certain area. This obviously alerted us to proceed with caution. In our linear regression models, the necessary assumptions were not fully satisfied. Our residuals were roughly mound-shaped, but often showed some sort of skew. Additionally, the assumption of homogeneity of variance was not satisfied, as there was clustering and patterns within all of the residual plots. Linearity was not satisfied either, as none of the residual plots exhibited symmetry along the x-axis. However, it is important to note that our transformation from the number of cases to the log(number of cases) in our regression made our scatter plots more symmetric and our residual distributions closer to normal. Lastly, for reasons similar to our t-test (e.g government policies, weather), the assumption of independence was not entirely satisfied.

If we were to do this project differently, one approach we would take is collecting more transportation data, as we were only able to find data from Apple Maps. This dataset is not representative of the entire population because it only consists of observations from Apple Maps users. Ownership of Apple devices may be correlated to wealth, which could reasonably impact mobility trends due factors like occupation, as some lower-wager jobs (retail and service for example) can not work remotely. Additionally, our Apple mobility dataset was limited because all of the data collected was based on daily requests for directions in the app. Directions would not apply to certain types of mobility, such as driving to the local grocery store or going on a walk in the neighborhood. A different approach we could take is trying to find data that is more representative of overall mobility. If we were to continue to work on this project, we would investigate how mobility trends are affected by the recovery from the COVID-19 pandemic. For example, we would conduct a t-test comparing the means of the mobility trends before and after the pandemic has taken its course. For example, we could compare a country's last month before a confirmed COVID-19 case to a country's first month with 0 COVID-19 cases. Obviously, this would not be possible for some countries right now, as virtually all countries are still dealing with the pandemic, but this could be a possibility in the distant future as vaccines are developed or the number of cases approaches 0. Based on the results of our analysis that mobility has decreased due to the pandemic, another interesting extension could have been to look into what people are doing in their homes. This could range from analyzing google search trends for cooking to gathering screen time usage data and performing further tests.

Works Cited: