A Sentiment Analysis on US Twitter Data in Relation to COVID-19

Daniel Varivoda

I. Abstract

The main purpose of this study is to closely gauge how the US population reacted to the COVID-19 pandemic using Twitter data and to learn more about the demographic of these tweeters. To gain a better sense of the sentiments, we focused on identifying words that were most commonly used in the expressed sentiments. Moreover, in hopes to better understand the demographics of these tweeters who are vocal about the pandemic, we investigated the location and readability of their tweets. Lastly, we explored the correlations between COVID-19 deaths and multiple predictor variables such as Flesch Reading Score, sentiment, number of airports, population, land area, and water area of a said state. With the variety of responses in different states, we hoped to find that negative sentiment (people who believe coronavirus is not serious), would correlate with greater occurrences of coronavirus deaths within their state. We also hoped to find that more complex tweets as measured by greater Flesch Reading Score, would be positively correlated with sentiment and negatively correlated with the number of COVID-19 deaths. Our analysis failed to find any conclusive link between our two target predictors and COVID-19 death, with state population, land area, and water area being our most influential predictors.

II. Introduction

COVID-19, also known as SARS-CoV-2, was first detected in Wuhan, Hubei Province, People's Republic of China in December 2019. In the early stages of the outbreak, President Donald Trump heavily downplayed the rampant virus while cases continued to grow in the US and elsewhere. For instance, when the Centers for Disease Control (CDC) advised that Americans wear face coverings in public to curb the spread, President Trump instead assessed it as voluntary. The president even stated that he himself would not wear any face covering, presenting Americans with mixed messages. This put thousands of American lives at risk as many continued to work and go about their daily lives, unknowingly exacerbating the spread. In just a matter of months, the US had reached an unfathomable death toll. Currently, the US holds the highest number of confirmed cases in the world.

Even though the US was well aware of the worsening condition of places like Wuhan and Italy that were in lockdown months ahead of the sweeping outbreak in its own soil, many Americans did not know exactly what to make of the spreading virus. China and Italy, countries that were most profoundly stricken at the outset, cautioned vigilance. Compared to South Korea and Taiwan who automatically took action and effectively subdued their battle against COVID-19, the foreboding reality of a pandemic was essentially minimized in the American media. Initial statements by Trump affirming Americans that the situation was "very much under control" and "will disappear" one day "like a miracle" misled many Americans to concede the pandemic as simply nominal. While the US's wavering response and delayed actions

may have been an attempt to first fully grasp the gravity of COVID-19 and prevent mass hysteria, ultimately, the US underestimated the spreading of this novel coronavirus disease.

By mid-March 2020, the virus intensified globally and the World Health Organization (WHO) declared the coronavirus outbreak a pandemic. Trump eventually declared a national emergency shortly afterward. In contrast to when the virus first devastated China and those who stocked up on food and other essential items in the US were ridiculed, ensuing these announcements, the entire country was inevitably in disarray. Markets were swamped with shoppers and shelves were emptied. For weeks, markets suffered substantial supply shortages. Peculiarly, many Americans were seen panic buying toilet paper and even fighting over other items they deemed essential. Upon all this, seemingly, the US became the subject of mockery and a cautionary tale to the rest of the world.

While the US government was slow in its response to the virus, it was especially quick to stimulate and revive its economy and calm investors. Months into 2020, the US stock market saw its worst crash since the Great Depression. Contradictory to recurring political claims of scarce funding for efforts that would ultimately improve the livelihood of everyday Americans, the Federal Reserve pumped trillions of dollars into the US economy. This stunt demonstrated the feasibility of worthwhile efforts repeatedly urged by Americans such as alleviating student debt, subsidizing SNAP benefits, and mitigating homelessness. To boot, while a strong economy is crucial for a sustainable society, the US seemed to prioritized the health of its economy over the lives of its people, sparking outrage. This led many Americans to flood the Twitter platform where they scrutinized the government and expressed their sentiments. Unlike the economy, it is not possible to revive lives that are lost. Treating lives affected by a virus with no known cure is not as simple as pumping money into the economy.

Presumably, the US's lack of urgency, transparency, shrewd leadership, as well as its misplaced priorities, has evoked a range of sentiments among Americans. The US was overly unprepared for the implications that came along with COVID-19 such as its disruption to normal life and the establishment of a new norm. Compared to early February when the US had barely 20 cases of the virus, the US is now faced with sobering numbers. To prevent overwhelming the healthcare system and mitigate COVID-19, states encouraged social distancing and administered their own approaches to dealing with the virus. By all means, the divergent approaches and plans between each state prompts a unique challenge for the country in its struggle against flattening the curve. As an illustration, the governor of Florida refused to close beaches despite growing cases and misgivings from others. Undeterred by the virus, people crowded Florida beaches just days following WHO's announcement. This posed conflicting viewpoints among the population on how others behave and how the government is managing the pandemic.

All things considered, the aspirations for this study are to shed light on the different, changing perspectives of the American people, the most common words

used to express sentiment, and the demographic variables of US tweeters in relation to COVID-19 deaths during this time. Ultimately, we hope that through this study we can document and better understand how the pandemic has affected the lives of the American people from different and similar backgrounds.

III. Data Acquisition and Data Cleaning

For our study, we collected a number of datasets from various sources. Our main dataset is the Twitter dataset containing tweets from February 1, 2020 to April 10, 2020. To clean this dataset, we first removed all non-English tweets since we want to focus this study on the US. We also removed all non-ASCII characters, URLs, retweets, and empty white spaces from the text in order to get an accurate flesch reading score for each tweet text and assist with parsing user locations. The rest of the datasets we used required little to no cleaning and were acquired to assist us in exploring a greater interpretation of the tweets and finding correlations between different variables. Those datasets are as follows: COVID-19 deaths by week per state from the CDC website, and states and airports from Wikipedia. For the COVID-19 deaths by week, we excluded all dates before February 1, 2020, and after April 11, 2020.

IV. Calculating Flesch Reading Scores

Again, a focus of our project is to dig deeper into tweets and learn about what kind of individuals are publicly sharing their opinions on varied topics related to the pandemic. More specifically, we want to find out whether a tweet is considered more or less educated given the assumption that more educated people tend to use formal language and less educated people tend to use informal language. Based on the article "The Readability of Tweets and their Geographical Correlation with Education", we used a metric called "Flesch Reading Score" to determine the readability of the written content in each tweet.

$$RE = 206.835 - 1.015(\#Words/\#Sentences) - 84.6(\#Syllables/\#Words)$$

To accommodate Twitter's short format, we used the modified version of the standard Flesch formula above given in the article mentioned earlier. This metric bases its readability score on word and sentence length including the number of syllables in each word. It produces values ranging from 0 to 122. Higher RE scores indicate greater ease in readability while lower scores indicate otherwise. Values below 0 indicate very complex text.

The mean for the Flesch Reading Score for each tweet is approximately 50. This value falls closer in the low range of RE scores suggesting that tweets concerning the virus require more effort to read and comprehend. However, it is important to mention that the distribution of these scores has a high variance of 1014.89 which is expected since we are working with a dataset that is not only very geographically

diverse but in general, very diverse. Therefore, it is difficult to conclude any significant deductions here.

V. Creating a Dictionary for Sentiment Analysis

In order to perform any actual sentiment analysis on the Twitter data, we first had to generate a frequency table of words used in the tweets. Below is a snippet of the frequency table. We have only displayed a snippet of the frequency table as the full output is lengthy.

```
##            words Freq
## 7361         the 1432
## 7492          to  981
## 227            a  667
## 5184          of  651
## 529          and  573
## 3780          in  563
## 3975          is  516
## 1799 coronavirus  417
## 3017         for  356
## 3691           I  299
## 5235          on  286
## 7356        that  264
## 626          are  252
## 8366         you  244
## 1435       China  241
## 3994          it  218
## 3429        have  215
## 7422        this  215
## 1800 Coronavirus  213
## 7644       Trump  199
```

By doing this, we were able to get a sense of which words came up the most and use them to help us create a sentiment dictionary. After identifying the most frequently used words, we sorted them into two categories: big_deal and not_big_deal. Below is the small dictionary we created:

big_deal	pandemic, outbreak, epidemic, crisis, global, death, infected, quarantine, lockdown, bad, dead, dangerous, deadly, emergency, serious, spreading, killed, ban, not good
not_big_deal	down, flu, fine, support, control, l, vaccine, nothing, hoax, free, clear, Dems, open, open, safe, propaganda, healthy, MAGA, healthier, fake news

VI. Measuring Sentiment

To measure sentiment, we used our dictionary to gauge the level of concern for the virus. The category big_deal represents negative sentiments, specifically considerable anxiety and concern regarding COVID-19. Conversely, the category not_big_deal refers to sentiments that display little concern or interest amid the pandemic. For each common word that appears in each tweet, we assign either a positive or negative point based on the category it falls in. After doing this for each word in a tweet, we calculate the sum to get the sentiment. Essentially, a sum of 1 indicates that the tweeter perceives the virus as threatening and a sum of 0 indicates a nonthreatening perception of the virus.

VII. Regression Model

To explore multiple predictor variables and their correlation to COVID-19 deaths by state, we used the following simple linear model:

```
lm(total_march_covid_deaths ~ flesch + sentiment + Land.area.mi.2. +
Water.area.mi.2. + Population, c_t)
```

After running diagnostic plots for this model shown in the next section, we found that it is not a valid model. Based on the assumptions of normally distributed error terms and constant variance of the error terms, the model is not valid. In an attempt to improve our model, we made some minor changes that slightly improved its accuracy. However, the data seems to be too cluttered for us to make any valid conclusions as it does not allow for a proper normal distribution. The improved linear model is shown below:

```
lm(sqrt(total_march_covid_deaths) ~ flesch + sentiment + Land.area.mi.2. +
Water.area.mi.2. + Population, c_t)
```

We also later added airports as a predictor into the model given the following model:

```
lm(sqrt(total_march_covid_deaths) ~ flesch + sentiment + num_Airports +
Land.area.mi.2. + Water.area.mi.2. + Population, c_t)
```

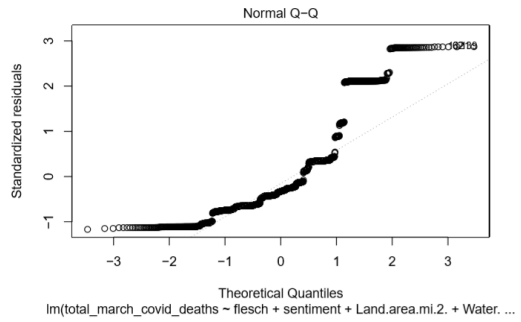
Lastly, we have the model that seems to be the best predictor of COVID-19 cases:

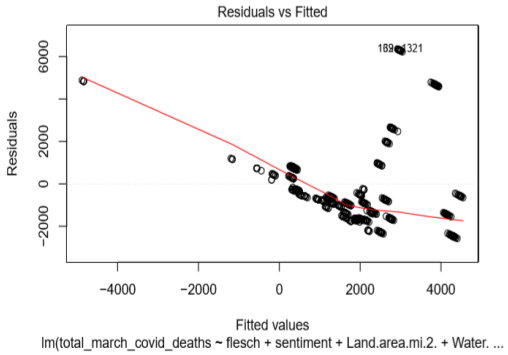
```
lm((total_march_covid_deaths)^(1/4) ~ Land.area.mi.2. + Water.area.mi.2. +
num_Airports + Population, c_t)
```

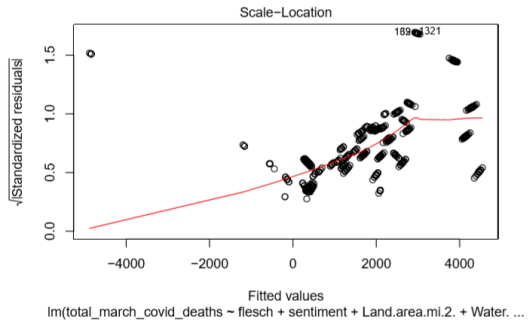
While even our transformed models were still not ideal, as seen by our diagnostic plots below, we opted not to use any power transformations on our predictors. Conducting power transformations on our predictors may create a better fitting model, however, it would destroy the interpretability for individual predictors, which was our main goal.

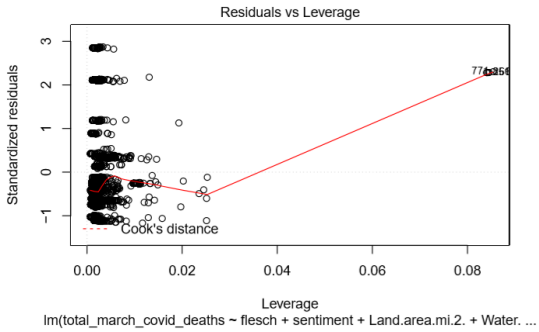
VIII. Model Plots

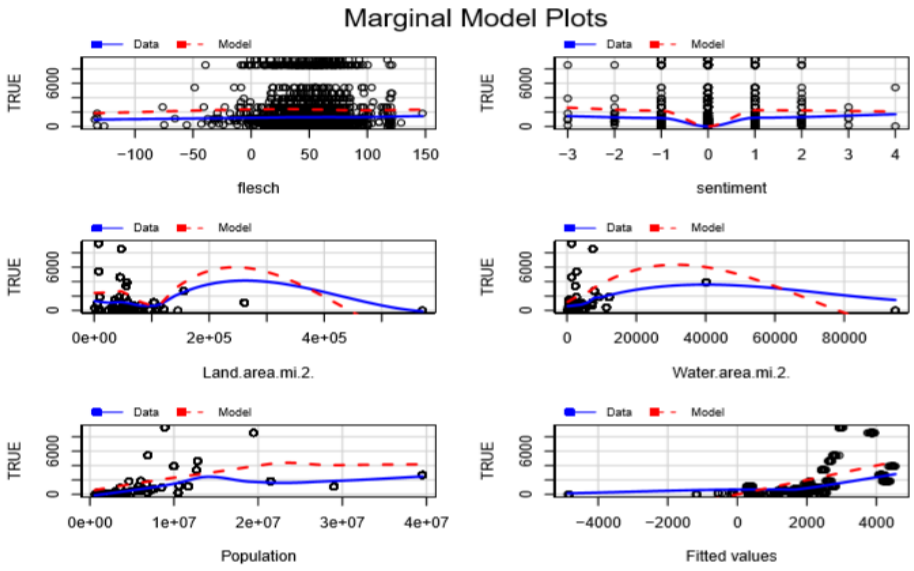
For the first model, we have displayed its diagnostic plots and summary output below.





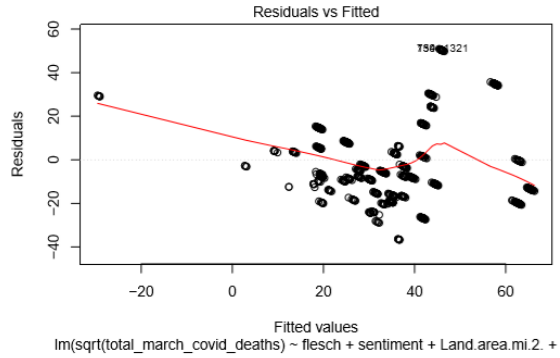


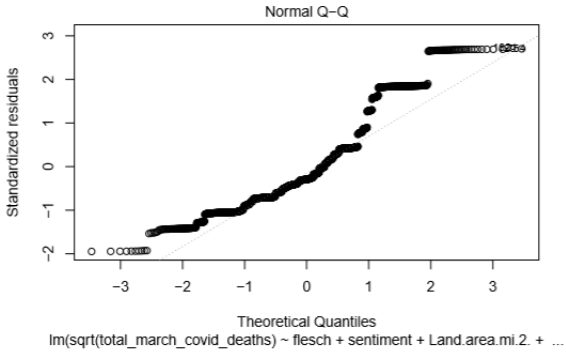


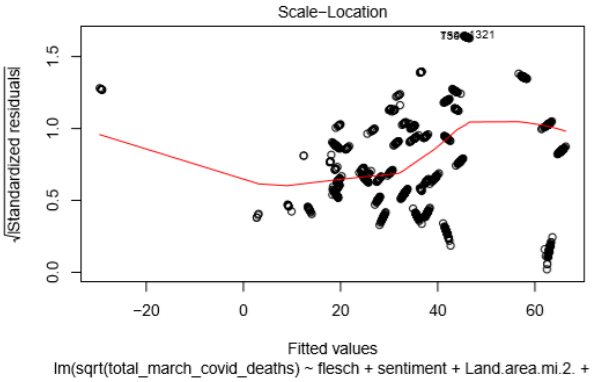


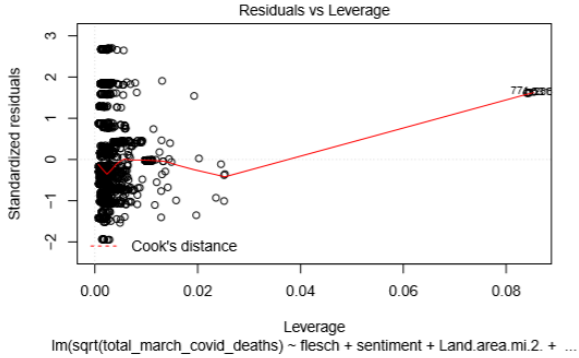

```
## lm(formula = total_march_covid_deaths ~ flesch + sentiment +
##     Land.area.mi.2. + Water.area.mi.2. + Population, data = c_t)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -2572  -1434   -718    757   6353
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.856e+03  1.305e+02  14.216   <2e-16 ***
## flesch            1.478e-01  1.866e+00   0.079    0.937
## sentiment         3.829e+01  7.616e+01   0.503    0.615
## Land.area.mi.2.  -2.242e-02  8.620e-04 -26.004   <2e-16 ***
## Water.area.mi.2.  6.324e-02  6.065e-03  10.427   <2e-16 ***
## Population        1.332e-04  5.419e-06  24.577   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2215 on 1894 degrees of freedom
## Multiple R-squared:  0.3141, Adjusted R-squared:  0.3123
## F-statistic: 173.5 on 5 and 1894 DF,  p-value: < 2.2e-16
```

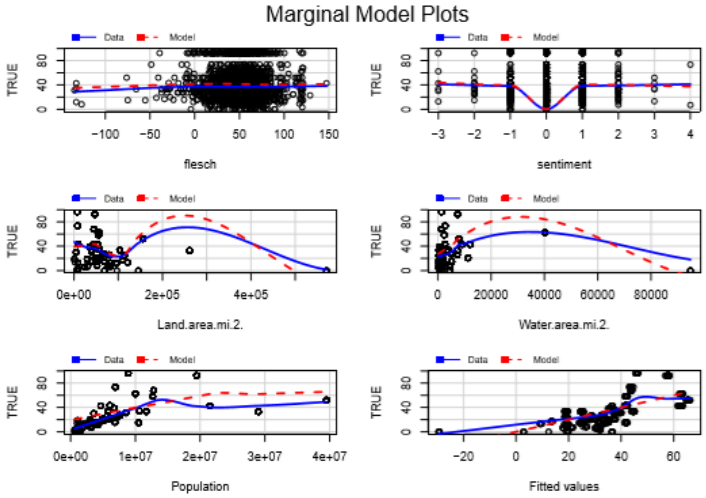
For our second model, we have also displayed its diagnostic plots and ANOVA output below.


Residuals vs Fitted
Residuals
Fitted values
lm(sqrt(total_march_covid_deaths) ~ flesch + sentiment + Land.area.mi.2. + ...


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(sqrt(total_march_covid_deaths) ~ flesch + sentiment + Land.area.mi.2. + ...


Scale−Location
√|Standardized residuals|
Fitted values
lm(sqrt(total_march_covid_deaths) ~ flesch + sentiment + Land.area.mi.2. + ...


Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
lm(sqrt(total_march_covid_deaths) ~ flesch + sentiment + Land.area.mi.2. + ...

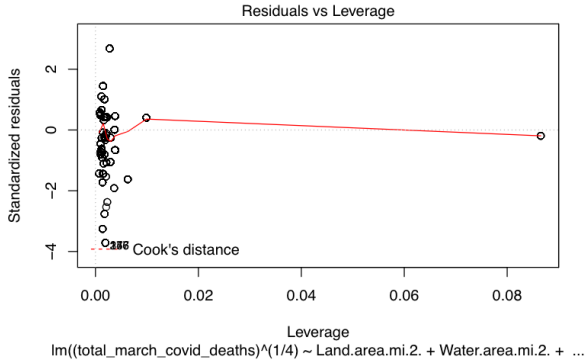

Marginal Model Plots
Data
Model
TRUE
flesch
sentiment
Land.area.mi.2.
Water.area.mi.2.
Population
Fitted values

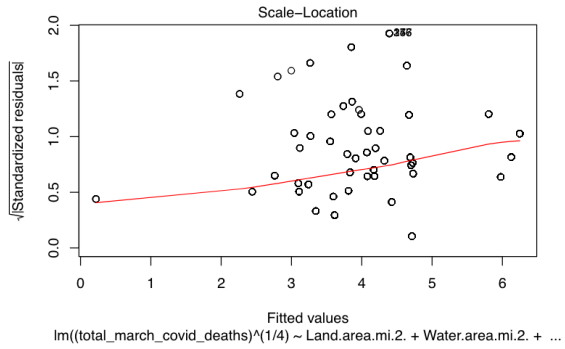
```
## Analysis of Variance Table
##
## Response: sqrt(total_march_covid_deaths)
##                    Df Sum Sq Mean Sq   F value Pr(>F)
## flesch              1      5       5    0.0134 0.9079
## sentiment           1     16      16    0.0438 0.8343
## Land.area.mi.2.     1  33990   33990   95.4318 <2e-16 ***
## Water.area.mi.2.    1  67493   67493  189.4984 <2e-16 ***
## Population          1 412415  412415 1157.9292 <2e-16 ***
## Residuals        1894 674579     356
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

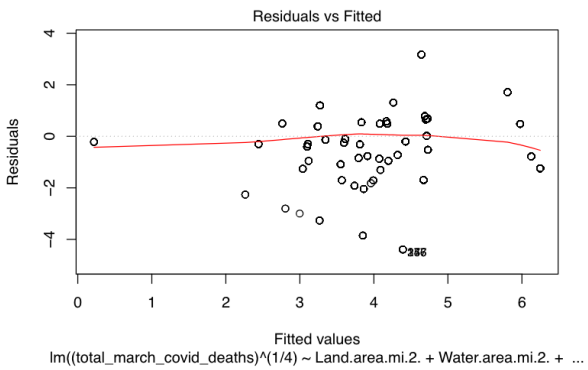
The following ANOVA output is from our third model which included airports as a predictor variable:

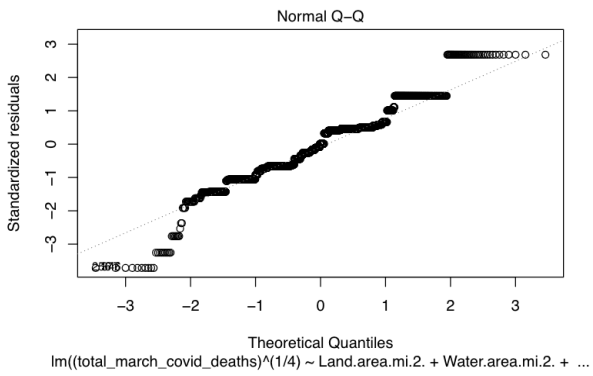
```
## Analysis of Variance Table
##
## Response: sqrt(total_march_covid_deaths)
##                    Df Sum Sq Mean Sq  F value  Pr(>F)
## flesch              1     13      13    0.0351 0.85137
## sentiment           1     55      55    0.1522 0.69647
## num_Airports        1  95969   95969 266.3043 < 2e-16 ***
## Land.area.mi.2.     1 322036  322036 893.6175 < 2e-16 ***
## Water.area.mi.2.    1   2458    2458   6.8214 0.00908 **
## Population          1  89290   89290 247.7711 < 2e-16 ***
```

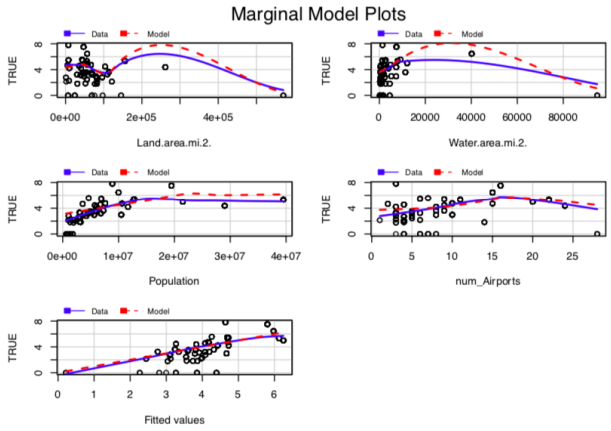
The following outputs are from our last and best predictor model:











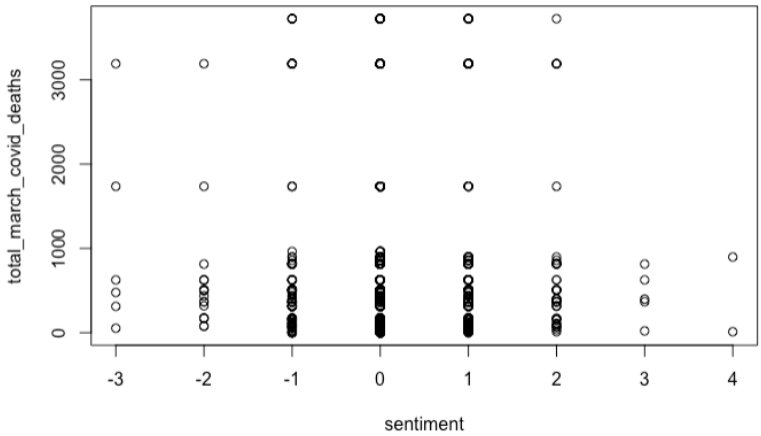

```
##
## Call:
## lm(formula = (total_march_covid_deaths)^(1/4) ~ Land.area.mi.2. +
##     Water.area.mi.2. + Population + num_Airports, data = c_t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3921 -0.7855  0.0130  0.5822  3.1713
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.839e+00  5.750e-02  66.765  < 2e-16 ***
## Land.area.mi.2.  -1.477e-05  5.070e-07 -29.126  < 2e-16 ***
## Water.area.mi.2.  3.392e-05  3.910e-06   8.676  < 2e-16 ***
## Population        7.881e-08  5.683e-09  13.869  < 2e-16 ***
## num_Airports      5.481e-02  1.091e-02   5.026 5.49e-07 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.184 on 1850 degrees of freedom
##   (45 observations deleted due to missingness)
## Multiple R-squared:  0.4589, Adjusted R-squared:  0.4577
## F-statistic: 392.2 on 4 and 1850 DF,  p-value: < 2.2e-16
```

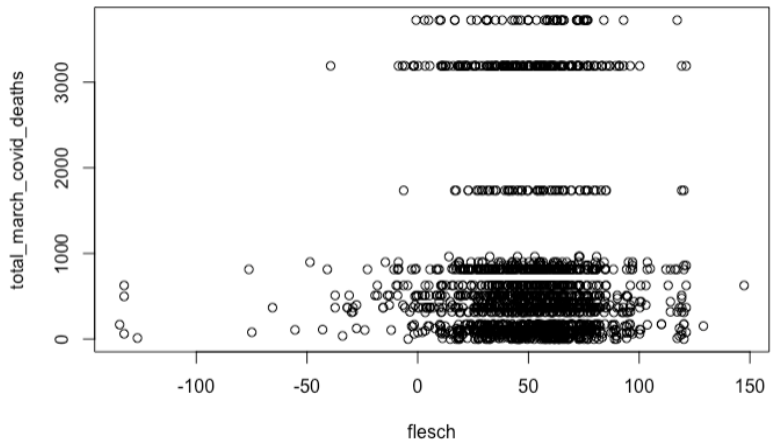
As can be seen above, many of the issues found in our residuals and scale location plots were fixed, however, there are still issues with the model indicating issues with the data.

IX. Results

From our models, we could see that there was no significant relationship on our predictor variable, COVID-19 Deaths, with Flesch Reading Score or Sentiment. Since our diagnostic plots showed some issues with our models, we also decided to check direct correlations between the variables.



The calculated correlation for sentiment is -0.01459617.



The calculated correlation for Flesch is -0.007971669

```
#no correlation between our 2 predictors either
cor(c_t$flesch, c_t$sentiment)
```

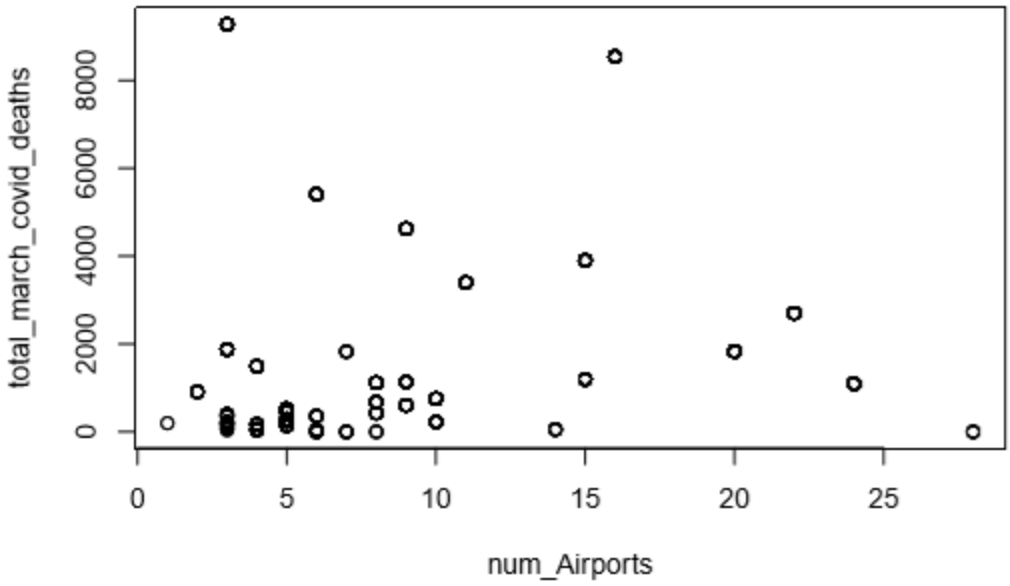
```
## [1] 0.009991998
```

```
#no correlation between clesch and favorite count or sentiment
cor(c_t$favorite_count, c_t$flesch)
```

```
## [1] -0.03419367
```

```
cor(c_t$favorite_count, c_t$sentiment)
```

```
## [1] -0.003735875
```



Our final few models have indicated that land area, water area, state population, and the number of airports in a state, are all significant predictors of the number of coronavirus deaths occurring, while Flesch-Reading Score and sentiment were unable to be linked as significant predictors. The correlations between our two target predictors and COVID-19 deaths were also very low, suggesting no relationship.

X. Conclusion and Further Study

Based on the results from our study, we were unable to find anything significant. We recognize that there were limitations in our study that may have played a role here. For instance, all of our models had troublesome QQ Plots, which indicates that our data was not normally distributed. This could be due to trouble with our location parser, as it only recognized state names and abbreviations, severely limiting the location data we were able to get. For example, if someone listed their location as Waikiki, a neighborhood in Honolulu, our parser would discard that datapoint as it could not identify the state. While we still found a large dataset using the current parser, it could be that certain states were underrepresented leading to the skew in our data. Another issue with our data was our approach for sentiment analysis. Due to time constraints, we were able to create a rudimentary sentiment library, however, counting the number of "bad" and "good" is not an accurate measure of sentiment as it doesn't take into account the complexities of the English language. For example, if there were two tweets, with one saying "We should reopen" and the other saying"We shouldn't reopen", the dictionary parser would count both as negative sentiment due to the presence of the word reopen. Another problem with our approach that could have led to issues in our data is that we treated the number of COVID-19 deaths as static when it was in fact changing week by week. To get our

number of deaths we just summed the deaths by state over the tweet time period for ease of analysis. This means that if there were few deaths early on the timeframe in a given state leading to a “negative/not a big deal” twitter sentiment and then many deaths later in the time period leading to a “positive/big deal” twitter sentiment, our model would be unable to tell the difference as both would be counted with the same number of deaths. Our suggestions for future research in this area would be the use of more sophisticated parsers for location and Flesch reading scores, as well as more sophisticated linguistic analyses for sentiment using machine learning approaches to analyze the overall connotation of a tweet rather than the individual words. We also would advise using the number of coronavirus deaths sorted by the closest time period to when the tweet was sent, rather than an average for the overall time period, which would allow for the model to take into account change in sentiment for changing death tolls. Even though we failed to find any significant results linked to the Flesch Reading Score or sentiment, we anticipate that they could very plausibly exist and significant results could be found using more sophisticated analysis methods. Nevertheless, we hope that our study can help provide readers with some insight.

References

Davenport, et al. “The Readability of Tweets and Their Geographic Correlation with Education.” *ArXiv.org*, 23 Jan. 2014

Long, Heather. “The Federal Reserve Has Pumped $2.3 Trillion into the U.S. Economy. It's Just Getting Started.” *The Washington Post*, WP Company, 29 Apr. 2020, www.washingtonpost.com/business/2020/04/29/federal-reserve-has-pumped-23-trillion-into-us-economy-its-just-getting-started/.

Prasad, Ritu. “Coronavirus: Why Is There a US Backlash to Masks?” *BBC News*, BBC, 5 May 2020, www.bbc.com/news/world-us-canada-52540015.

Watson, Kathryn .“A Timeline of What Trump Has Said on Coronavirus.” *CBS News*, CBS Interactive, www.cbsnews.com/news/timeline-president-donald-trump-changing-statements-on-coronavirus/.