An Analysis on US Twitter Data in Relation to COVID-19 Pandemic

The US Response to COVID-19

Due to a lack of transparency, as well as misplaced priorities, by some officials the US response to COVID-19 was wavering and delayed.

- presented Americans with mixed messages that shaped:
  - moral intensity: the degree of feeling that a person has about the consequences of a moral choice
  - ethical decision-making i.e compliance to social distancing and other safety protocols

Formulation of Research Objective

Questions:

- What is the general public's level of concern for the spreading virus?
  - Are they threatened? Morally aware? Skeptical?
- While moral intensity plays a role in human mobility, is this sentiment an accurate predictor for COVID-19 deaths?

Hypothesis:

- More well-educated people tend to use more complex language and less educated people tend to use less complex language
- Morality is related to level of education

Potential Pain Points:

- Can we use the readability of the written content in each tweet to deduce number of coronavirus deaths?

The Objective

- To determine whether moral intensity and readability are significantly correlated to COVID-19 deaths and can be used to accurately predict future trends based on tweets in the US

Data Acquisition

Data was acquired from multiple sources and combined into a single dataframe for analysis.

Sources:

- Sample Twitter Data from 2/01 - 4/10
- COVID-19 deaths by week per state from the CDC website
- Table of U.S. States and Abbreviations from Wikipedia (Included Population, Water Area (mi^2), and Land Area (mi^2) for each State)
- Table of Airports by State from Wikipedia.

Data Cleaning

Twitter Data:

- Removed all non-english tweets from the dataframe
- Removed all non-ASCII characters, URLs, retweets, and empty white spaces from the text using text parsing functions in R (stringr)
- Removed non-ASCII characters from locations to make them easier to parse
- Removed extraneous data columns to create a more concise data frame

Covid-19 Death Data:

- Removed all data for weeks ending before 2/01 and after 4/11
- Removed all data not pertaining to US states
- Grouped all weekly COVID-19 deaths by state and found the sum to create a new dataframe of total Covid Deaths for time period by state

Sentiment Calculation

- Broke up every Twitter text into a vector of individual words and combined each new vector to a vector of total words in the text
- Created a frequency table of words using the total words vector to show most commonly used words
- Manually looked through the list to find words that would most commonly be associated with the view that COVID-19 was not a major threat (negative sentiment) and the view that COVID-19 was a major threat to public health (positive sentiment)
- Using these words created a small sentiment dictionary then parsed all twitter text to assign a sentiment score for each tweet.
    - Every word that was in the positive sentiment dictionary increased the score by one point and every word in the negative dictionary decreased the score by one point

Measuring Readability: Using A Flesch Reading Score & Getting User Locations

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

- # Sentences was set to 1 to accommodate the short text format of tweets

- Used cleaned Twitter text data to calculate a vector Flesch Reading Scores for each Twitter datapoint then added the vector as a column in the twitter data frame.
- Parsed the location column of the Twitter dataset using the data frame of state names and abbreviations and kept any matches, throwing out the rest.
- Each match was assigned a state number which was later used to match a state name for each remaining datapoint
  - Some double matches were found (state contained the name of another state like West Virginia or multiple states listed), however those were dealt with manually as there weren’t very many

Combining Data Frames

A Small Example:

- Each of our 3 final dataframes (tweets, airports by state, and Covid deaths by state) had a state variable attached to each datapoint
- Used the join family of R functions in the Dplyr library to join all three data frames by each state.
- The tweet data frame was used as the base dataframe and rows from other data frames were added based on the value of the state variable.

Creating a Linear Model

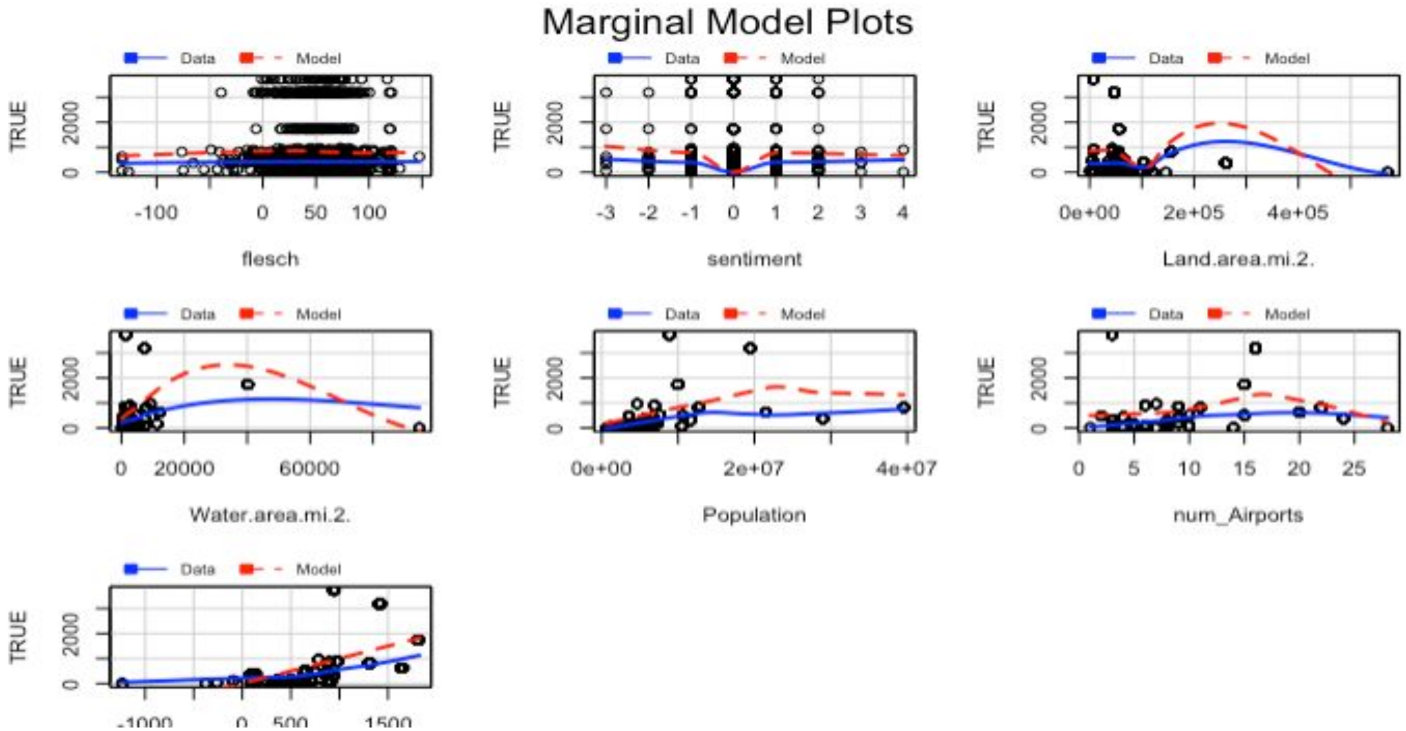
- First created a model using all of the data available to predict COVID-19 Deaths (FRE, Sentiment, Population, Airports, Water Area, & Land Area)
- Used diagnostic and marginal model plots to assess validity of the dataframe and found some major issues with the Residuals Plot, Scale Location, and QQPlot
  - This indicated issues with the trend, normality, and constant variance of conditions of our model
- Used an inverse response plot to find the best transformation for the response variable (covid deaths), which produced a lambda of $0.263725 \sim$ ¼
- Transformed the model using (covid_deaths)^(¼) as our response variable

Finding a Linear Model (cont’d)

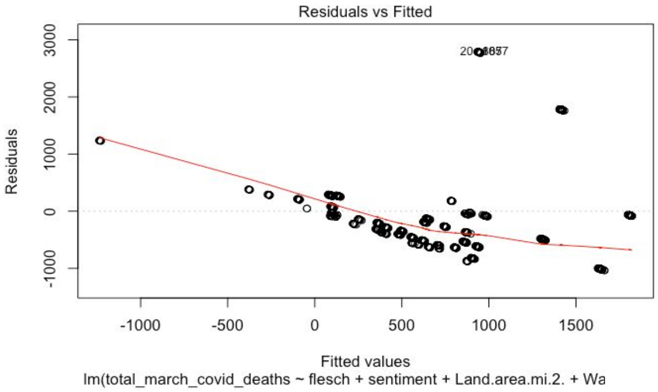
- Even in our new model, FRE and Sentiment was not significantly correlated with COVID-19 Deaths, therefore we cannot definitively say there is any relationship between the variables
- Next used exhaustive stepwise regression to find the best model, and selected a combination of number of airports, population, land area, and water area as the best predictor for COVID-19 Deaths as our best model
- Our final model can be seen here:

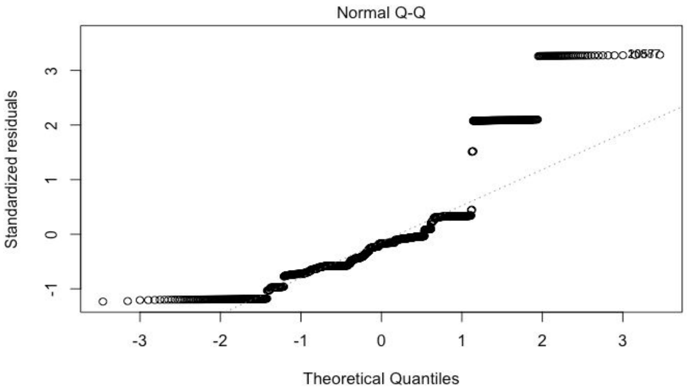
```
lm(formula = (total_march_covid_deaths)^(1/4) ~ Land.area.mi.2.
    Water.area.mi.2. + Population + num_Airports, data = c_t)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4391 -0.7337 -0.0093  0.5367  3.1690

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.876e+00  5.488e-02  70.633  < 2e-16 ***
Land.area.mi.2.  -1.429e-05  4.839e-07 -29.533  < 2e-16 ***
Water.area.mi.2.  3.621e-05  3.732e-06   9.702  < 2e-16 ***
Population        7.403e-08  5.424e-09  13.648  < 2e-16 ***
num_Airports      5.523e-02  1.041e-02   5.306 1.25e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.13 on 1850 degrees of freedom
  (45 observations deleted due to missingness)
Multiple R-squared:  0.4665,	Adjusted R-squared:  0.4654
F-statistic: 404.5 on 4 and 1850 DF,  p-value: < 2.2e-16
```

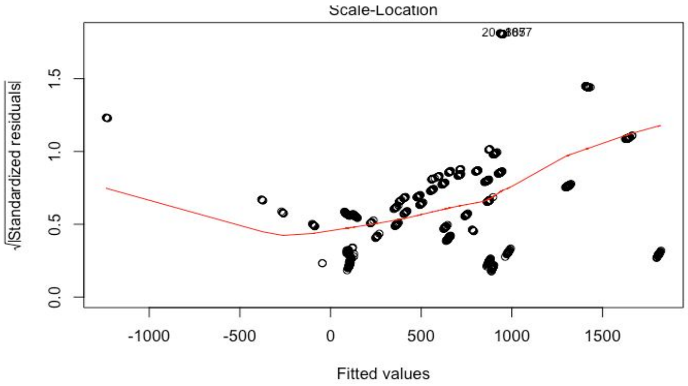
Original Model Plots



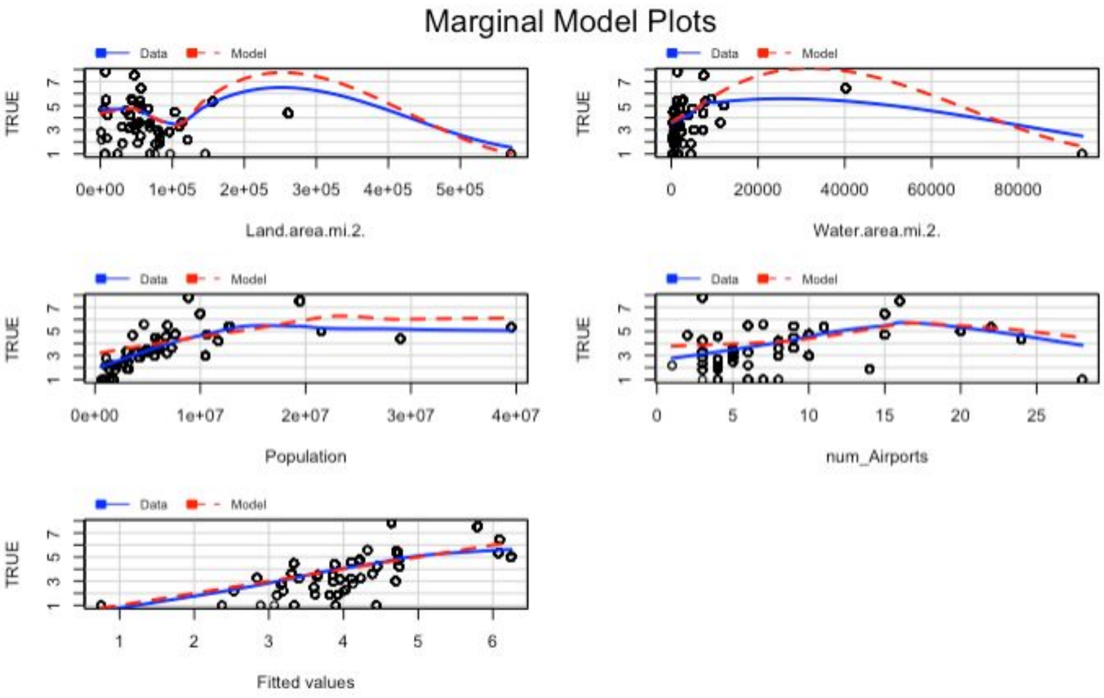
Original Model Plots (cont'd)



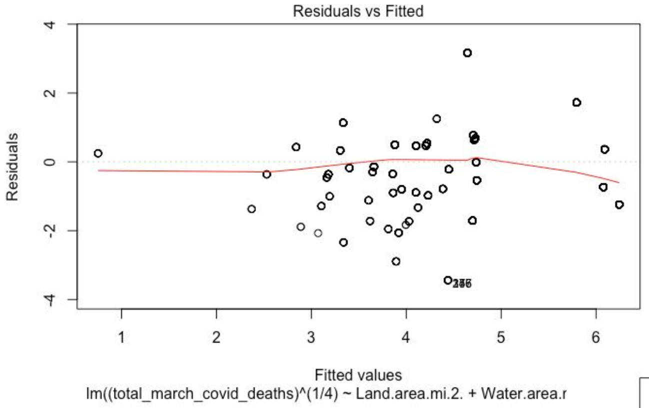


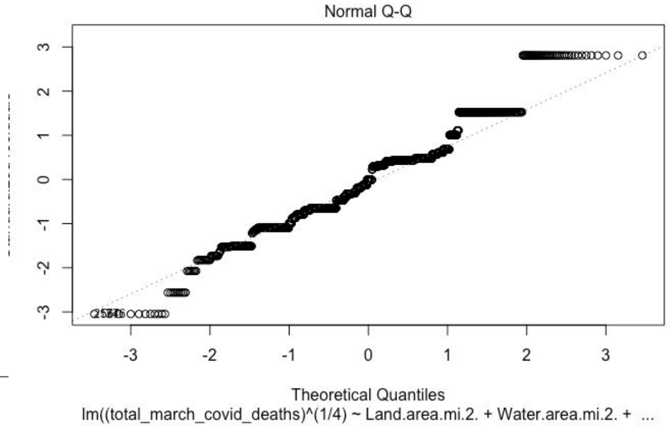


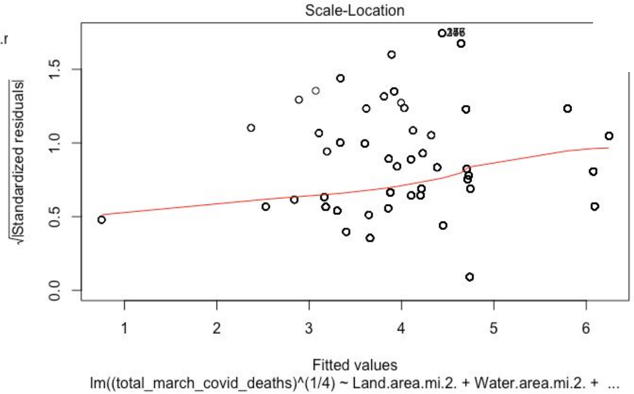
Final Model Plots



Final Model Plots (cont’d)







Conclusion

Based on the results from this study, we were unable to find any significant correlation between our target measurements and number of Covid deaths.

- This could very well be due to issues with our data:
  - All of our models had troublesome QQ Plots, which indicates that our data was not normally distributed.
    - This could be due to trouble with the location parser, as it only recognized state names and abbreviations, severely limiting the location data available.
    - It could be that certain states were underrepresented due to our simplistic parser leading to a skew in our data.
- Due to time constraints, I created only 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
  - Further, the number of COVID-19 deaths were treated as static for simplicity when it was in fact changing week by week