COVID-19 Impact on Key Countries

Duarte 'Eddie' Costeira, Chia-Chun Lee, Dayong Lee, Jeff Lee

20 November 2020

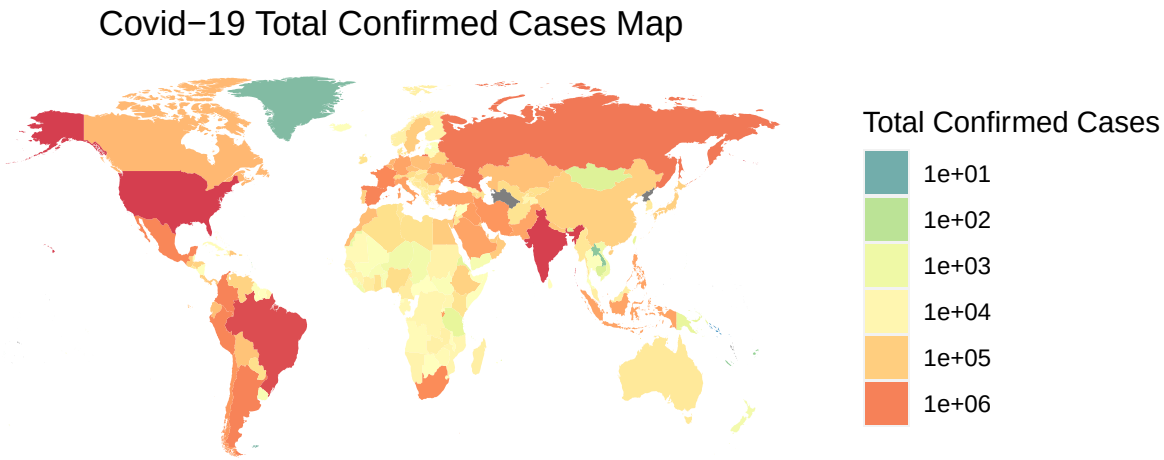


Figure 1: Total confirmed COVID-19 cases for all countries as of 15 October 2020, visualized on a logarithmic scale.

Introduction

Abstract

The COVID-19 Pandemic is undoubtedly the most significant event in recent human history, and the spread of the SARS-CoV-2 virus to every country on Earth has prompted extensive discussion over the effectiveness of different countries' responses to the virus. The effectiveness of these responses was analyzed, with a focus on the United States, using several methods. In order to accurately compare the COVID-19 situation in different countries, the primary metric used was *Positivity Rate*, measured as the number of confirmed positive cases over number of total tested. It was found that of the key countries considered, the United States did not have the largest *Positivity Rate* - that distinction belongs to the United Kingdom. Furthermore, a convincing correlation between *Positivity Rate* and Population Density was not found, which were fairly surprising results considering that the transmission vectors of SARS-CoV-2 require close proximity with infected persons. A Multivariable Linear Regression model the United State's *Positivity Rate* was built, and this model accurately predicted total deaths at a future date. Finally, the change in US Public Sentiment with the progression of the virus was analyzed, and a correlation between *Positivity Rate* and Public Sentiment towards the virus was not found.

Analysis Conducted

Our goal for this investigation was to determine the effectiveness of the United States's response to the COVID-19 pandemic, to model the spread of the virus through the United States, and to discover how public sentiment towards the virus changed as the pandemic progressed.

There are a multitude of factors that differentiate countries (including population size, population density, demographic breakdown, level of development, cultural differences, etc.) that make meaningful comparison difficult. In an attempt to standardize each country's response to the pandemic, the metric of *Positivity Rate* was chosen, which for a given time point and country is defined as:

$$\text{Positivity Rate} = \frac{\text{Total Number of Confirmed Cases}}{\text{Total Number of People Tested}}$$

To compare the United States's response to the COVID-19 Pandemic, *Positivity Rate* was plotted over time, along with four other key countries. Additionally, there was curiosity to see what, if any, change to *Positivity Rate* occurred in response to key time points in the Pandemic. As the primary vectors of transmission for SARS-CoV-2 are through aerosolized bodily fluids, an attempt to determine if there was a relationship between *Positivity Rate* and Population Density was made by plotting each country along those metrics and observing the results.

There was interest to see if the United States's response to COVID-19 could be modeled, and performed a multivariable linear regression in an attempt to produce a model that could predict future deaths with some degree of accuracy.

Finally, the response to COVID-19 has become a significant political issue, and has inspired many anti-shutdown protests across the United States. There was curiosity to see how the Public Sentiment towards COVID-19 changed, and if there was any relationship between Public Sentiment and *Positivity Rate*.

Data Sources

The data used was sourced primarily from the Google Cloud Platform repository of COVID-19 data, which is a repository of aggregated and curated info relating to COVID-19. This repository pulls data from Wikipedia, Eurostat, DataCommons, and other direct sources such as countries' ministries of health.

https://github.com/GoogleCloudPlatform/covid-19-open-data

To create our multivariable linear regression, a secondary dataset from OurWorldInData was used. This dataset was chosen because it factored the data into multiple variables, making it more convenient for building a regression model.

https://ourworldindata.org/coronavirus

To determine public sentiment, the text of several thousand articles published by the New York Times was scraped and the sentiment score of the wording of each article was analyzed. The positive and negative sentiment scores of articles over time were used as an analogue for Public Sentiment.

Positivity Rate Comparrison

Key Countries Chosen

It was determined that comparing the United States's COVID-19 response to every other country was not feasible, so four key countries were chosen: The United Kingdom, India, Japan, and South Korea. These countries were selected because they each display unique considerations:

- The United Kingdom's Prime Minister was hospitalized after contracting SARS-CoV-2, and the country did not impose travel restrictions.
- India's population is significantly larger and more dense than the United States.
- Japan had one of the earliest government responses to COVID-19 in terms of travel restrictions and economic shutdown.
- South Korea experienced it's first confirmed case on the same day as the United States, and the country did not impose any official economic shutdown measures.

Our goal is to compare these four countries in terms of *Positivity Rate*, and as can be seen in Fig.2, it appears that the United States accounts for about 50% of total counts for each variable plotted - however, the United

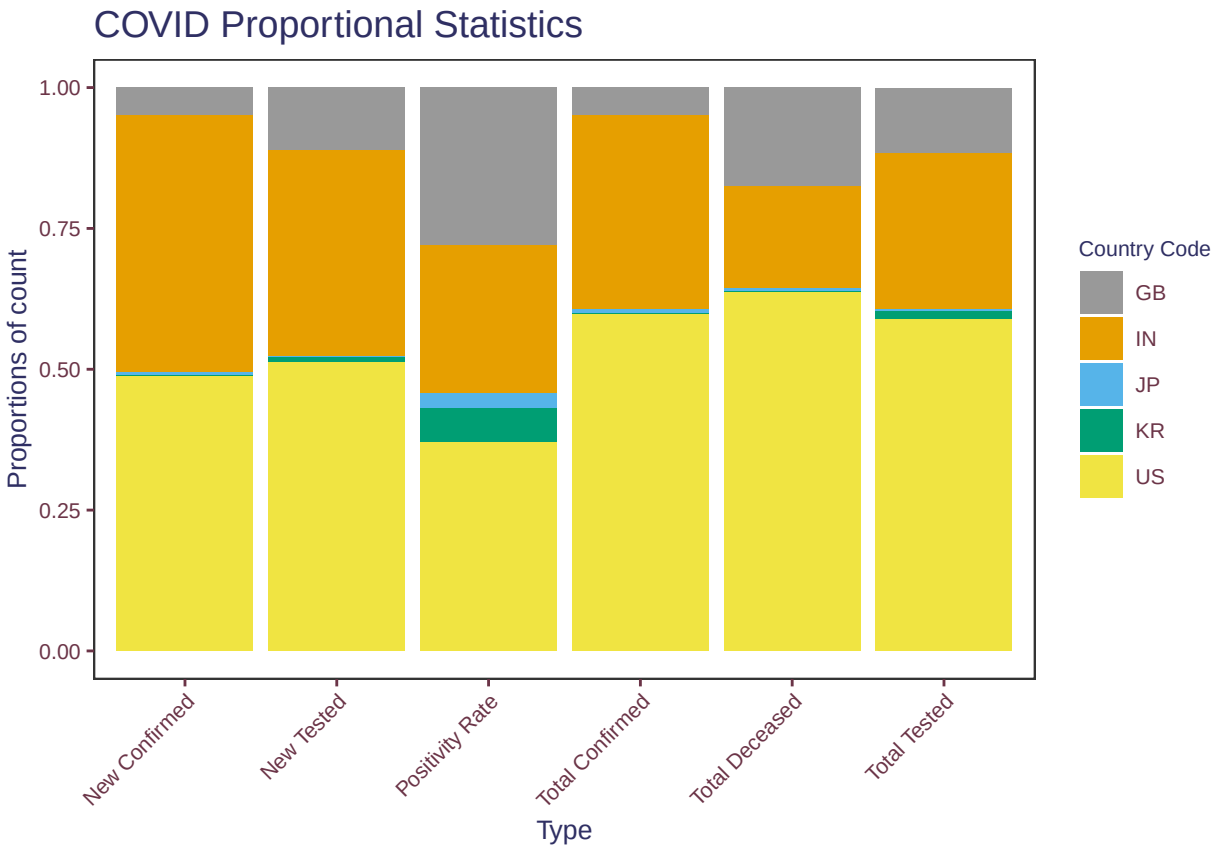


Figure 2: Proportinal comparison of the five key countries chosen, on several methods.

States also accounts for over 50% of *Total Tested*, so there was interest to see how this would impact *Positivity Rate* over time.

Positivity Rate Comparison between Key Countries

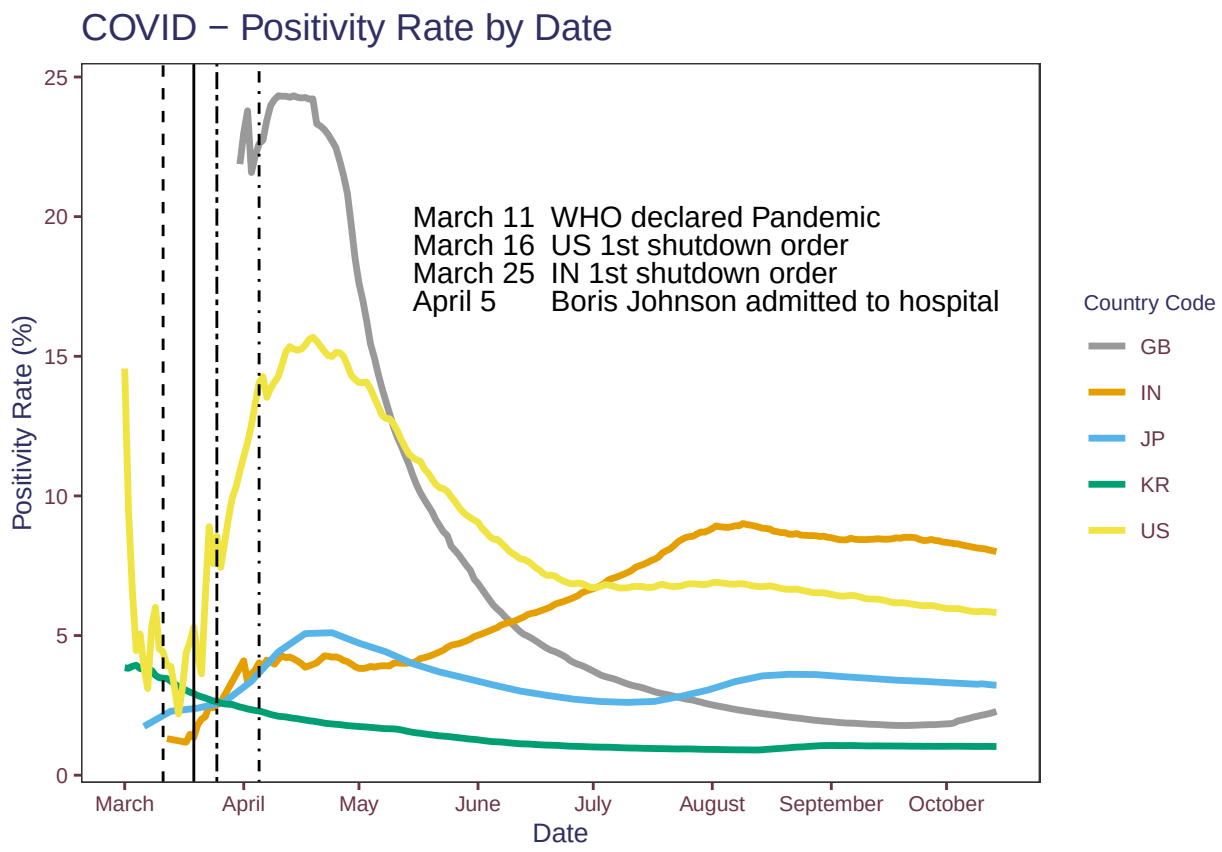


Figure 3: Plot of changes in Positivity Rate over time for the five key countries. Additionally, key dates have been marked with vertical lines.

Fig.3 shows a plot of the *Positivity Rate*s of the five key countries over time, and a few notable trends present themselves. Immediately, it can be seen that the United States did not experience the highest *Positivity Rate*, the country that did was the United Kingdom, and this occurred shortly after the UK Prime Minister Boris Johnson was hospitalized after infection with SARS-CoV-2. After this time point, UK *Positivity Rate* drops significantly.

Secondly, it can be observed that each country experienced a general decrease in *Positivity Rate* over time, even after temporary increases (as can be seen with Japan and United States), with the only exception being India, which experienced a steady increase. But what is interestingly unique to the United States is the lack of stability in *Positivity Rate* - whereas other countries experienced steady changes, the United States experienced very extreme changes in *Positivity Rate* in the first half of 2020, which suggest that the government response was not as well organized as that of the other four countries.

Finally, there was curiosity to see if any patterns emerged if key dates in the pandemic timeline were highlighted. The United States had the most pronounced responses to these milestone dates, responding to each date with a sharp drop in *Positivity Rate*. However, each response proves to be temporary and is accompanied by a sharp rise in *Positivity Rate* shortly afterwards.

Earlier it was noted that the United States had a very large number of *Total Tested* cases, and it is worth investigating to see if there is a relationship between *Positivity Rate* and *Total Tested*. Fig.4 plots the

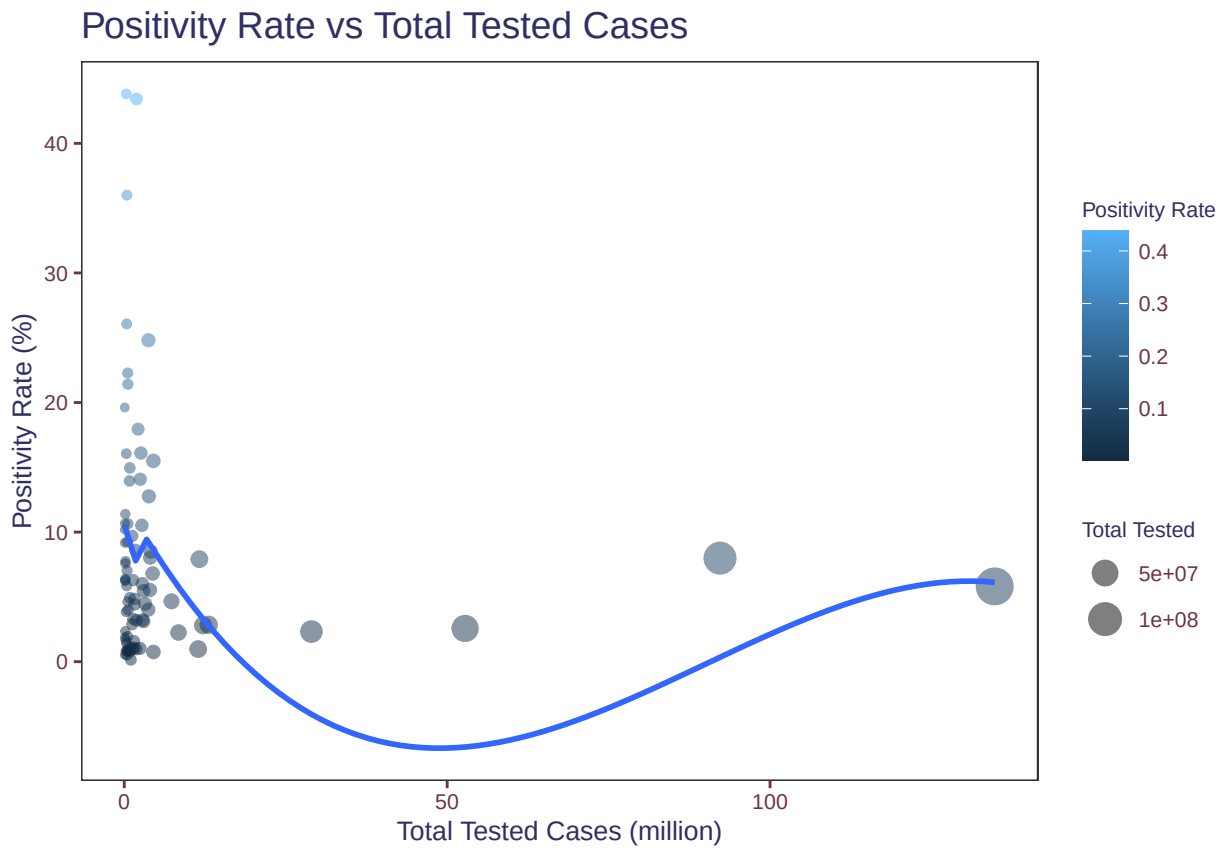


Figure 4: Positivity Rate vs Total Tested Cases, with a smoothed LOESS regression method curve using the default parameters of span = 0.75 and degree = 2

Positivity Rate vs *Total Tested* for ever country and adds a smooth curve using the LOESS regression method, it can be seen that the curve captures almost no data points, and thus a very weak relationship between these two metrics can be inferred.

Positivity Rate vs Population Density

One of the key countries chosen in our previous analysis, India, was picked because it had a population that is significantly larger and more dense than the United States. Population density was a metric worth considering, because as mentioned previously, the primary vectors of transmission for SARS-CoV-2 are through aerosolized bodily fluids. Thus, one would expect to see that countries with higher population densities would experience increased transmission of the virus.

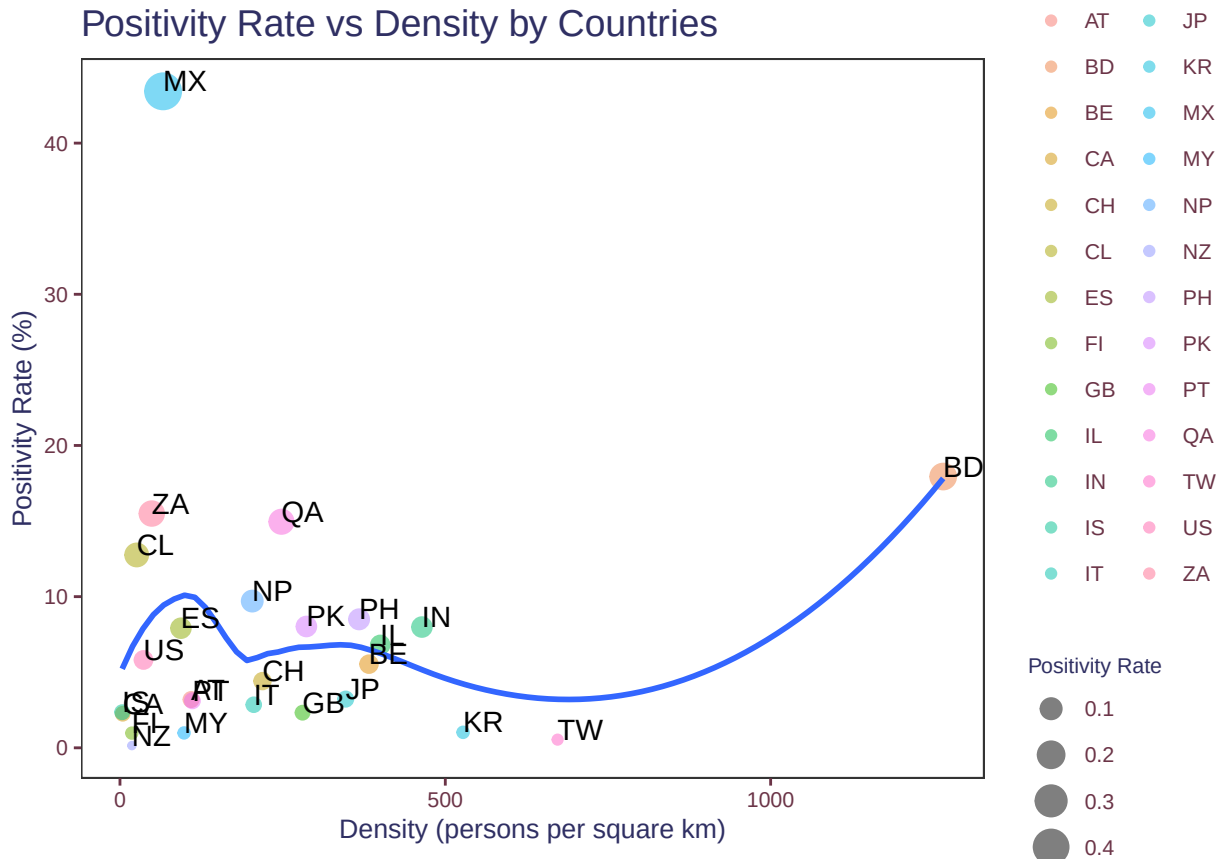


Figure 5: Positivity Rate plotted against Population Density, with a smoothed LOESS regression method curve using the default parameters of span = 0.75 and degree = 2

However, plotting these two variables (Fig.5) again fails to yield a convincing correlation. Plotting every country and adding a smooth curve using the LOESS regression method produces a curve that does not capture the distribution of data points. Two countries of note are Bangladesh and Mexico. Bangladesh has the highest population density of the countries analyzed, yet had a fairly low *Positivity Rate* in comparison to the country with highest. Conversely, Mexico displays the highest *Positivity Rate* but has a low population density.

It is noted that this analysis is not considered to be complete. Further analysis is recommended to address the limitations of this plot, with the following considerations:

- Inclusion of change in *Positivity Rate* vs Population Density over time
- Population Density of specific cities or states, opposed to entire countries
- Removal of outlier countries like Mexico and Bangladesh

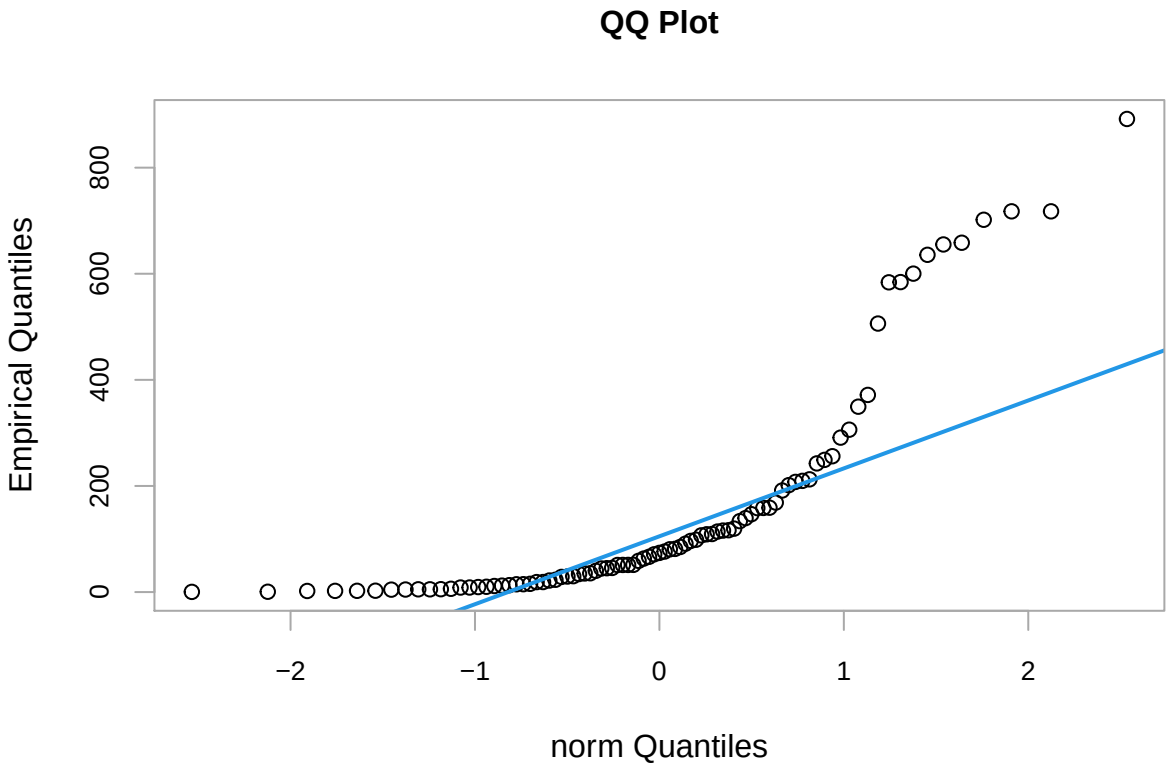
Multivariable Linear Regression

There was interest to see if it would be possible to model the Total COVID-19 Deaths in the United States. Unfortunately, while the primary data source contained extensive COVID-19 data for several countries, it did not contain enough variables for to use as predictors in a regression model. To rectify this, a secondary data source, from OurWorldInData was used, and the following predictors were chosen:

- stringency index
- diabetes prevalence
- life expectancy
- gdp per capita
- aged 70 or older, and
- positive rate

to form the initial regression model, assuming $\alpha = 0.05$.

Model Construction

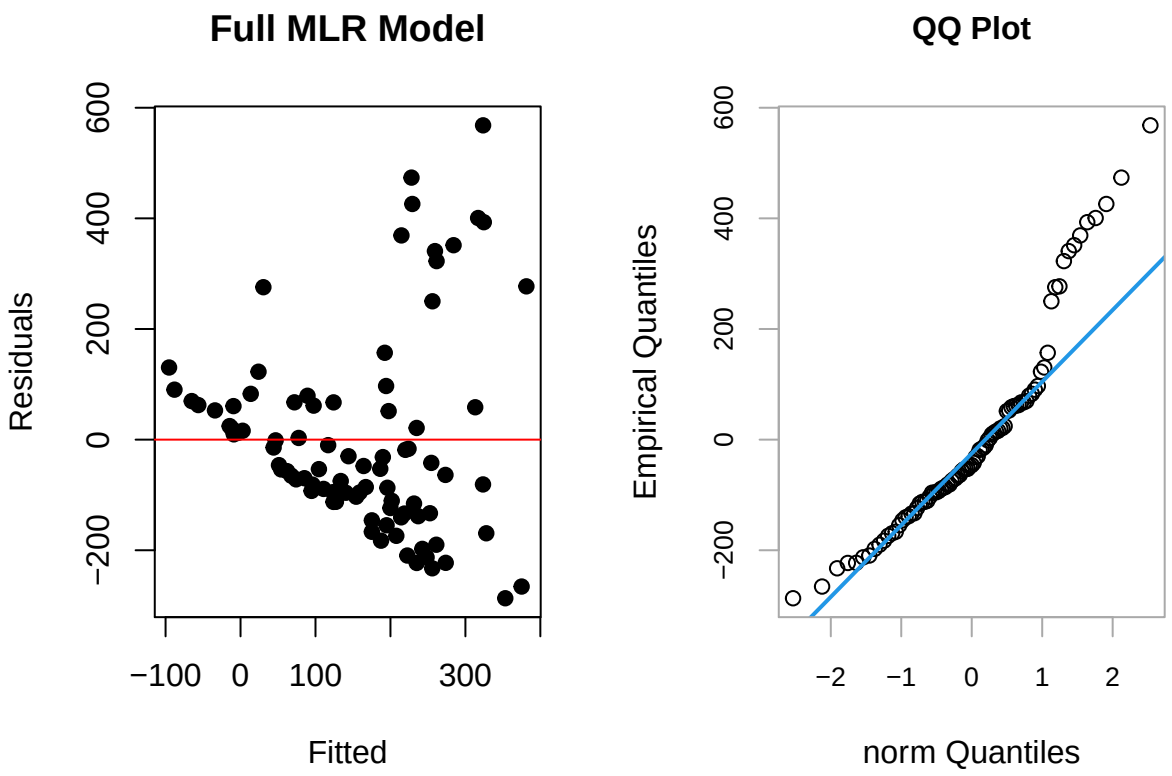


To create the initial model, a subset of the *Total Deaths* data was taken for the date "2020-10-15", omitting NA values. Initial tests for normality failed (Shapiro-Wilkes yielded a p-value of 7.993e-12), so an effort was made to determine which predictors could be dropped,

As can be seen from Table 1, it appears that only Stringency Index and Positive Rate are predictors worth keeping - however, further tests were conducted to see if transformation would yield better results:

Table 1: Initial Model

	Dependent variable:
	Total Deaths per Million
Diabetes Prevalence	−8.282
	(7.128)
Stringency Index	3.270**
	(1.534)
Life Expectancy	5.569
	(6.179)
GDP per Capita	0.001
	(0.001)
Aged 70 or Older	6.327
	(7.730)
Positive Rate	978.612***
	(314.521)
Constant	−525.083
	(373.076)
Observations	89
R^2	0.288
Adjusted R^2	0.236
Residual Std. Error	183.684 (df = 82)
F Statistic	5.530*** (df = 6; 82)
Note:	*p<0.1; **p<0.05; ***p<0.01

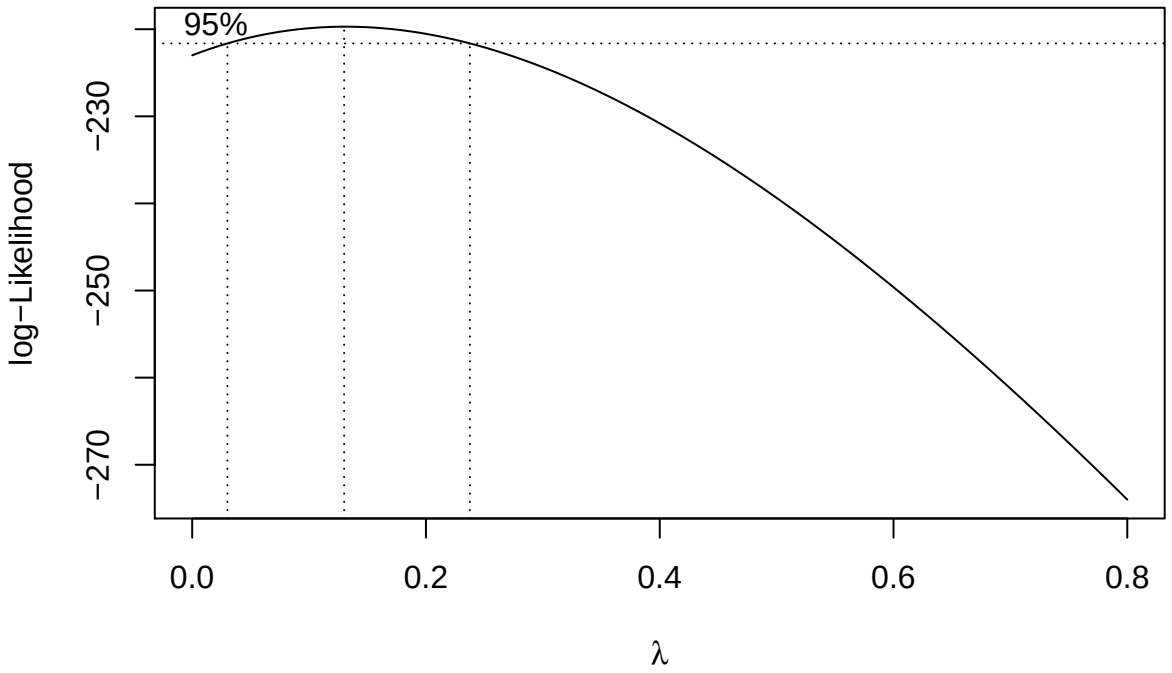


The triangular fanning pattern of the residual plot indicates the constant variance of errors is violated. The obvious linear trend on the bottom is also undesirable. The Q-Q plot confirms the distribution is not normal. To address this, a Box-Cox transformation was used.

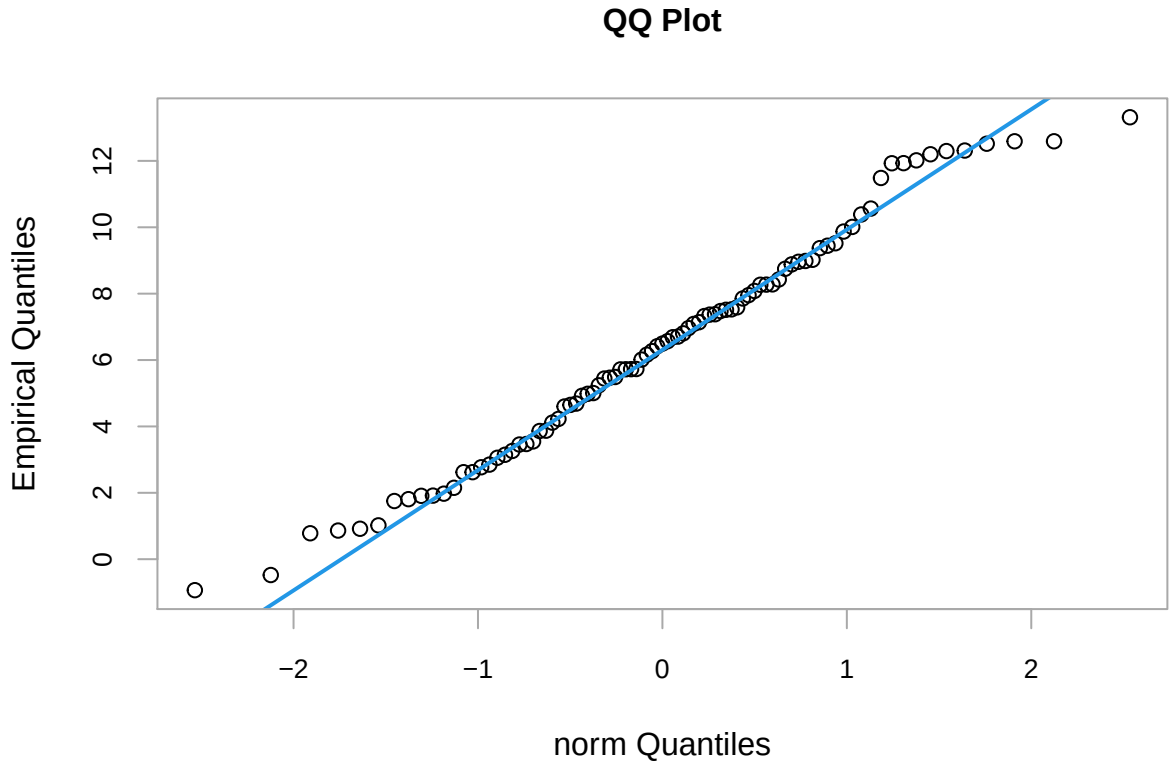
Transformation of Model Parameters

A Box-Cox Transformation transforms non-normal dependent variables into a normal distribution, and is summarized by the formula,

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases}$$



The resulting plot from the *boxcox* function indicates a value of $\lambda = 0.18$ is optimal.



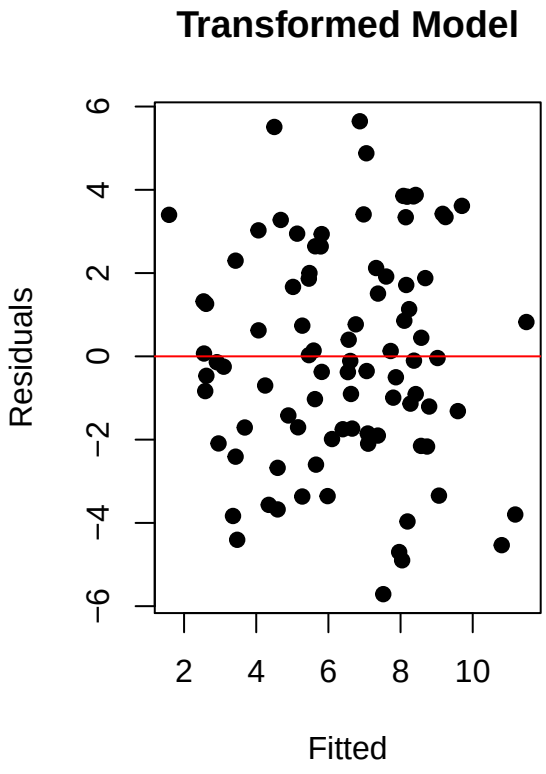
Repeating the Shapiro-Wilkes test yields a p-value of 0.2324 which is greater than the decided alpha value of 0.05, thus the null hypothesis that the transformed data is not normal is not rejected. A Q-Q Plot of the transformed data confirms it is normalized.

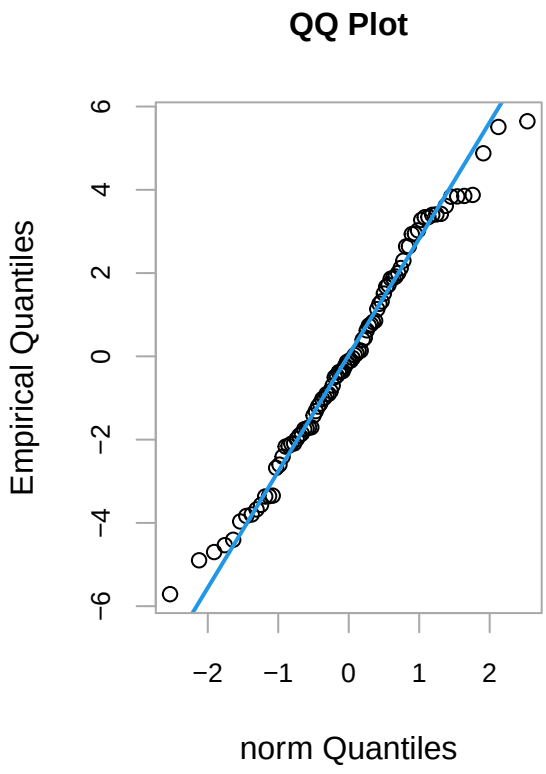
Table 2 represents a finalized summary of data after performing the Box-Cox transformation and using step-wise regression (forward and reverse) to remove the predictors Diabetes Prevalence and Life Expectancy from the data. While the low R^2 value of 0.415 is noted, the prediction of a test set will proceed.

Table 2: Box-Cox Transformed Model

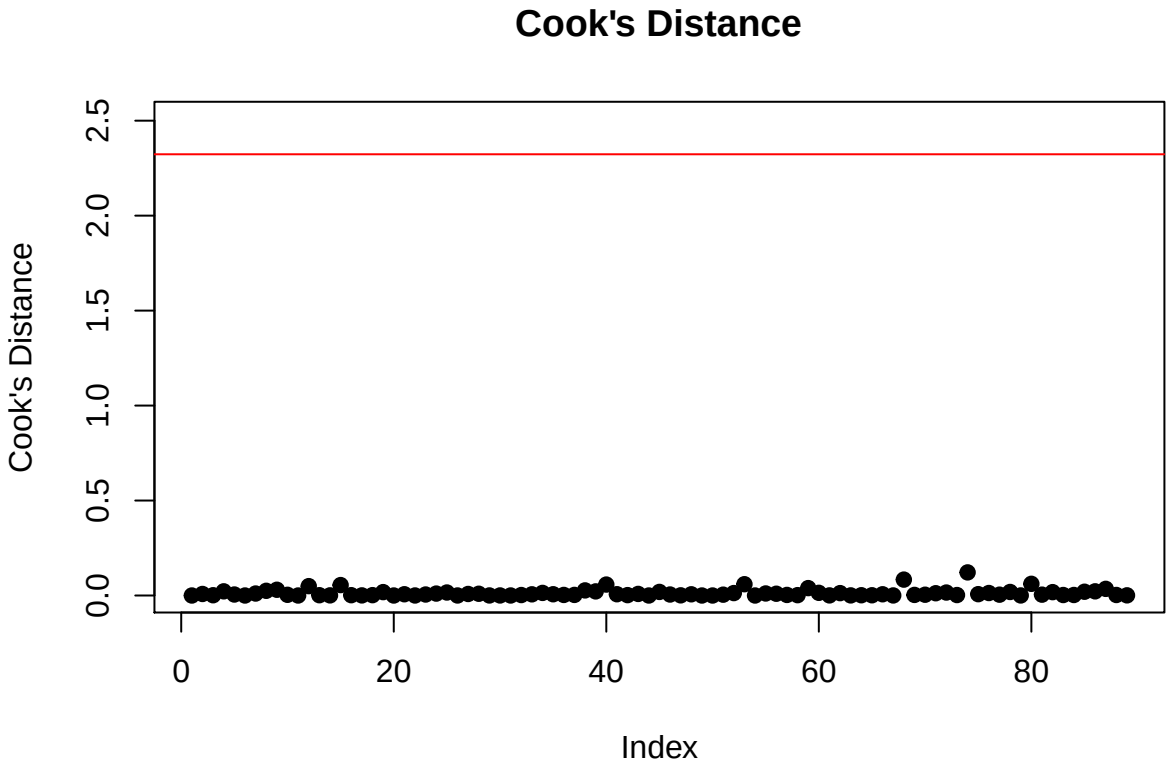
	Dependent variable:
	Total Deaths per Million
Stringency Index	0.044**
	(0.021)
Life Expectancy	0.00004***
	(0.00001)
GDP per Capita	0.225***
	(0.068)
Positive Rate	21.313***
	(4.511)
Constant	−0.165
	(1.310)
Observations	89
R^2	0.415
Adjusted R^2	0.387
Residual Std. Error	2.681 (df = 84)
F Statistic	14.879*** (df = 4; 84)
Note:	*p<0.1; **p<0.05; ***p<0.01

Model Evaluation





Post-transformation, a marked improvement can be seen in both residuals and normal Quantiles.



A check was conducted to detect outliers that could affect the data set, but the Cook's Distance plot shows no outliers above the F-statistic (red line), which indicates that there are no influential outliers.

The next step was to subset the original dataset as before, but for the date "10-27-2020".

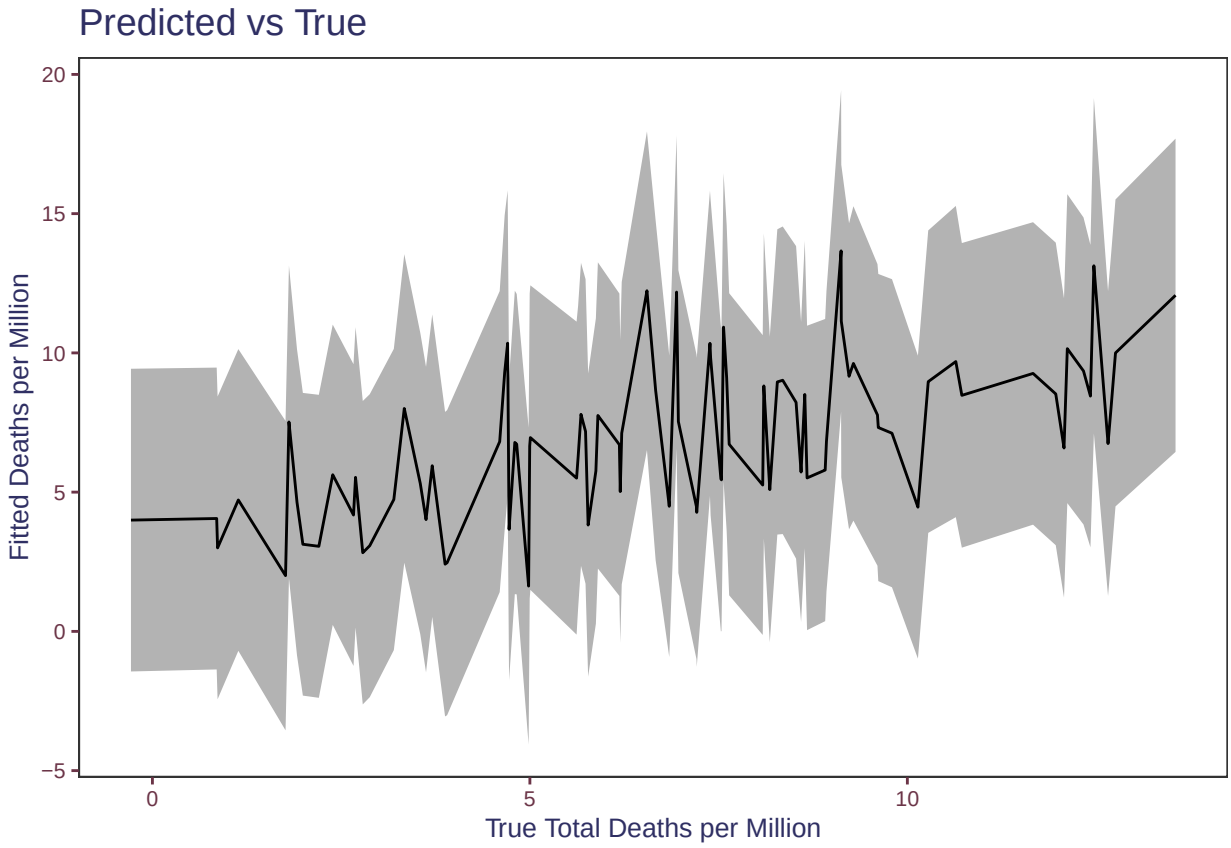
Which yields Mean-Squared Error Values of:

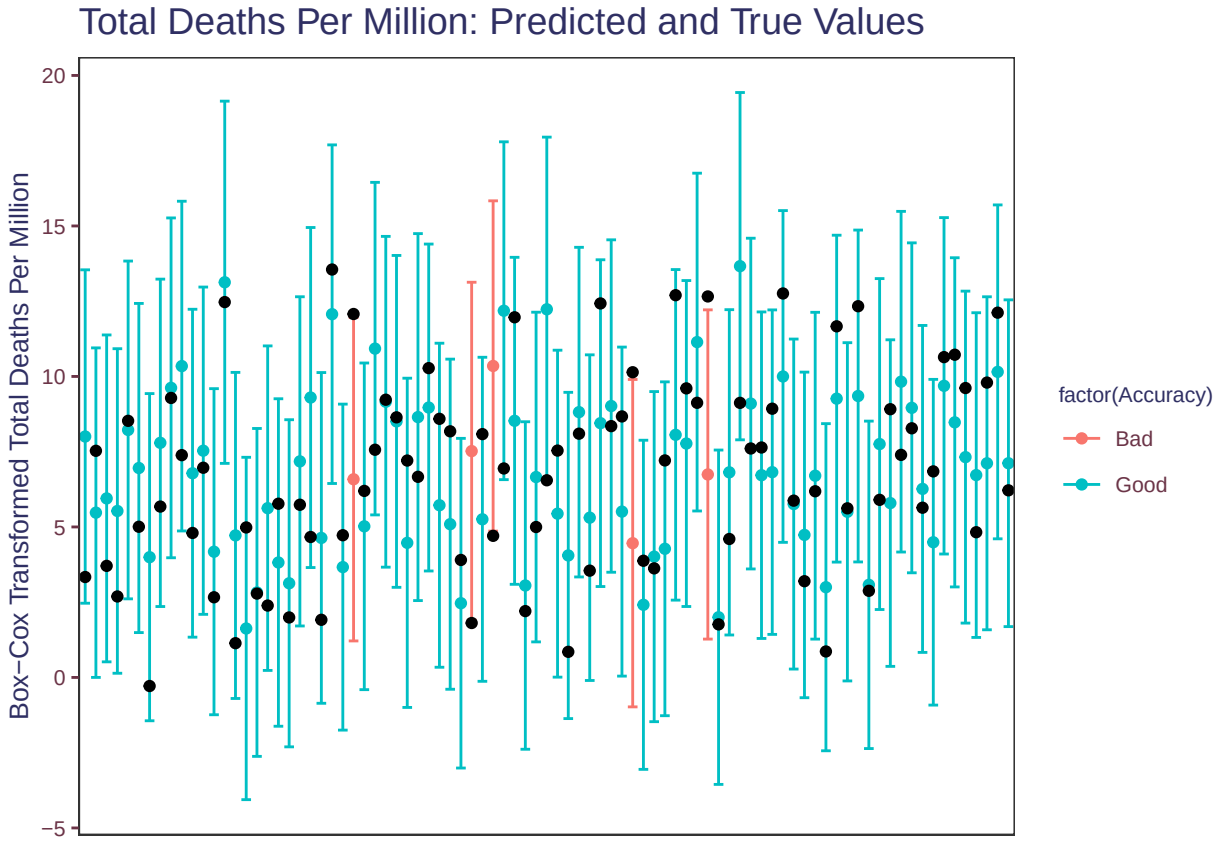
$$\text{Accuracy of MLR for Test Data via RMSE: } = 2.7234$$

$$\text{Accuracy of MLR for Training Data via RMSE: } = 2.6048$$

The Root Mean-Squared Error values are fairly large compared to the Box-Cox-transformed range of the data, and so they are a valid cause for concern and an issue to address in future refinements of this analysis. For now, the prediction will proceed.

Prediction of Total Deaths per Million





The end result of the Multivariable Linear Regression was a model that predicted future Total COVID-19 Deaths per million using predictors:

- stringency index
- life expectancy
- gdp per capita
- aged 70 or older, and
- positive rate

Again, it should be mentioned that some statistical values (R^2, RMSE) were less than satisfactory, so room for improvement certainly exists.

This MLR model was constructed through a Box-Cox transformation of the data set and step-wise forward and reverse regression. There were no significant outliers that leveraged the data and thus, a prediction could be made with some test data (this prediction of course assumes that the world remains the same without the discovery of a vaccine) with an expected accuracy of 94.2529.

US Public Sentiment on COVID-19

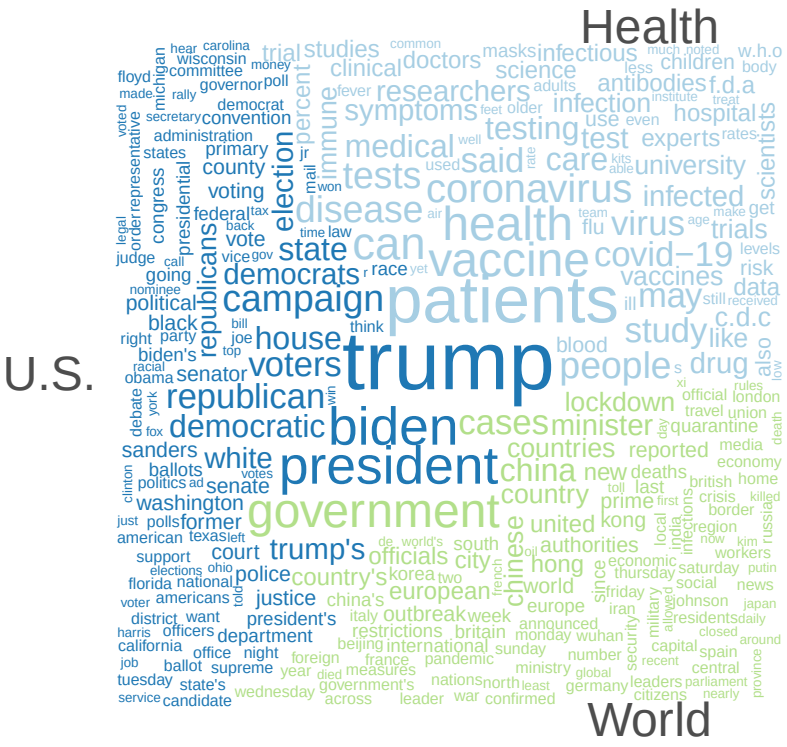


Figure 6: Word Cloud of most common words pertaining to the topics of 'U.S.', 'Health', and 'World', generated using text scraped from NYT articles related to COVID-19

New York Times Article Scraping

For the final component of this analysis, the change in public sentiment towards COVID-19 in the United States over time was analyzed. COVID-19 has become a very politicized issue as a result of misleading statements by then US President Donald Trump, anti-China prejudice, widespread conspiracy theories of the virus being a hoax, and other factors. This has prompted several anti-mask and anti-shutdown protests, and many hypothesize that these sentiments affected (possibly hampered) the United States's response to COVID-19. There was curiosity to see how the Public Sentiment towards COVID-19 changed over time.

Tthe most prominent newspaper in the United States, the New York Times, was chosen as an indicator for Public Sentiment. Using the NYT API and the R package *rvest*, a scraping of the text of all NYT articles that were a hit for a keyword search of *"coronavirus"* between the dates of January 20 and October 26 was conducted. The resulting list of articles were filtered to only include articles marked as *News* and *News Analysis* (and exclude *opinion*, *briefing*, etc.). The text of the resulting list of 14,972 articles was analyzed.

To process the text, the collection of articles was converted into a *corpus* object, and then tokenized into individual words. The R package *quanteda* was then used to remove stopwords, punctuation, abbreviations, numbers, symbols, urls, and other irrelevant text, and all words were converted to lowercase.

Instead of using a more complex positional (string-or-words) approach, a non-positional (bag-of-words) approach to text analysis was taken. For each article, the number of "Positive Words" and the number of "Negative Words" were counted, and each article was assigned a *Sentiment Score*:

$$\text{Sentiment Score} = \frac{\text{\# of positive words - \# of negative words}}{\text{\# of positive words + \# of negative words}}$$

To determine which words were positive and which were negative, the Lexicoder Sentiment Dictionary (2015) was used. Some examples of positive and negative words are:

Negative:

- Abandon
- Abnormal
- Accursed
- ...

Positive:

- Ability
- Abundant
- Accomplish
- ...

US Public Sentiment on COVID-19 vs Positivity Rate

Plotting the *Sentiment Score* of NYT articles over time (Fig.7), a clear initial upward trend can be seen following a negative initial value, then stabilization above positive. However, a pattern does not present itself when US *Positivity Rate* is superimposed. *Sentiment Score* increased as *Positivity Rate* decreased, suggesting an inverse correlation, but this was not observed when *Positivity Rate* increased in mid-March. However, it is interesting to note that this increase in *Positivity Rate* was accompanied by a tighter spread of *Sentiment Score* - when *Positivity Rate* was peaking, *Sentiment Scores* had the least variability. This interesting observation would be a good candidate for future analysis.

Summary of Key Findings

- The United States did not experience the largest *Positivity Rate* of the countries analyzed, but did experience the most erratic changes in *Positivity Rate* over time.
- *Positivity Rate* in the United States did respond to key dates in the COVID-19 Pandemic timeline, but responses were temporary.
- A correlation between *Positivity Rate* and Population Density was not found.
- A Multivariable Linear Regression Model for Total COVID-19 Deaths was built that could predict Total Deaths at a future date with 94% accuracy.
- *Public Sentiment* in the United States towards COVID-19 was very low initially, but increased over time, and did not correlate with *Positivity Rate*.

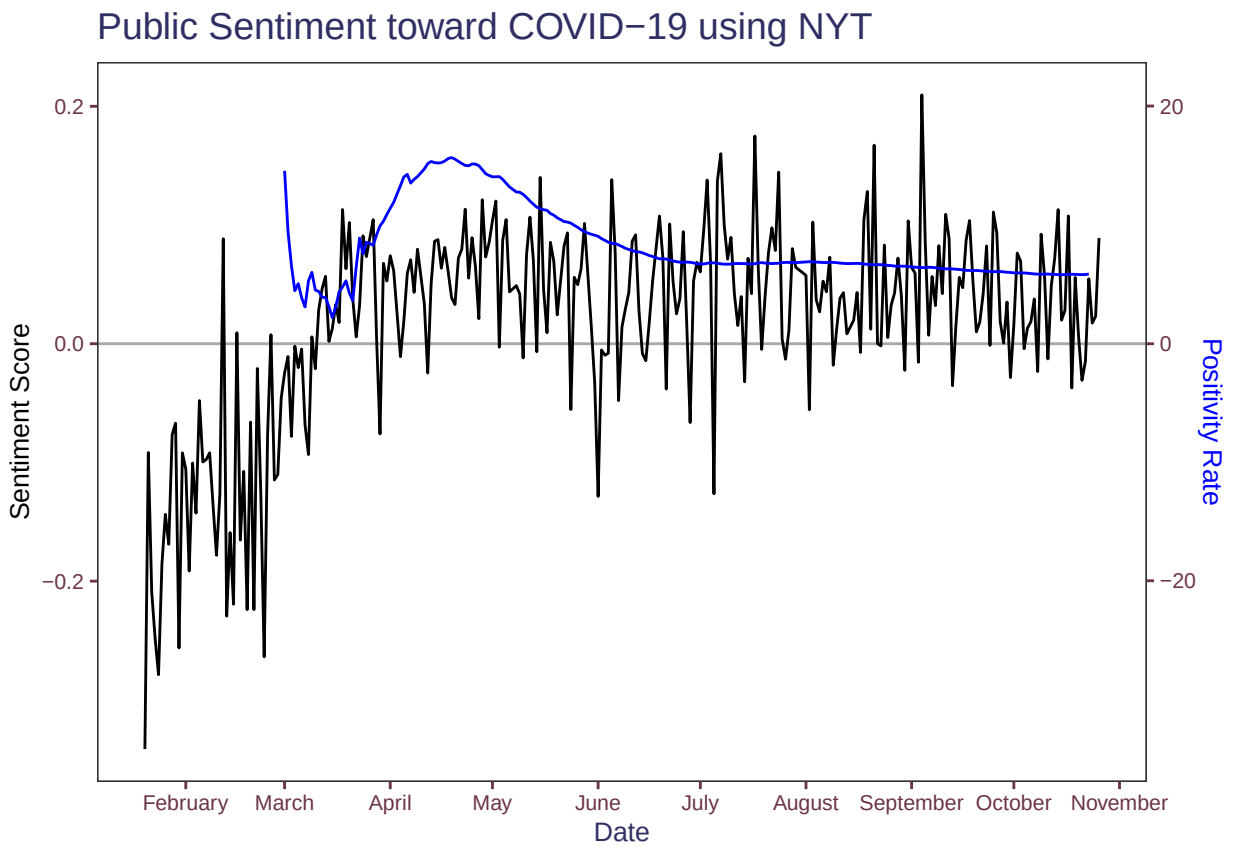


Figure 7: Sentiment Scores of NYT articles about COVID-19 over time, with US Positivity Rate superimposed.