# 

**SMITH ANALYTICS CONSORTIUM**

**Data Series Workshop**

**COVID-19 Tweets Analysis**

**Group 8**

**Alexander Binder, Jinxin Hou, Hsiaoting Ko,**

**Suchit Sanghvi, Neerja Singh**

**Date: 08/14/2020**

# Table of Contents

# Introduction and Background 2

# Methodology 3

# Analysis and Results 5

# Conclusions and Recommendations 9

# Appendices 11

# References 15

### Introduction and Background

In December, according to the U.S. Department of Health and Human Services and Centers for Disease Control and Prevention (CDC), Chinese authorities identified an outbreak caused by a novel—or new—coronavirus, SARS-CoV-2. Coronavirus causes mild to severe respiratory illness, known as Coronavirus Disease 2019 (COVID-19). While the outbreak began in Wuhan, Hubei Province, China, it since has spread to a growing number of countries worldwide including the United States. On March 11, 2020, the World Health Organization declared COVID-19 a pandemic.

As of August 2020, there have been over 5.17 million confirmed cases of COVID-19 in the United States since the first case of coronavirus in the U.S. was detected on January 20th, 20201. Since then, coronavirus has spread to all 50 states within the U.S. Currently, the U.S. has the 8th highest death rate per capita among countries with 50,000 or more reported cases at 47.9 deaths per 100,000 people. Additionally, the U.S. ranks 24th out of the highest case-fatality ratios. For every 100 people with COVID-19 in the U.S., three people have passed away2.

Although these statistics are notable, they do not paint the full picture about how effective the U.S. government's response has been to this global pandemic. The professional services company Deloitte has developed a whitepaper titled, 'Government's Response to COVID-19: From pandemic crisis to a better future'. In this report, the writers note that there are three main phases of government action throughout the pandemic: The Respond Phase, the Recover Phase, and the Thrive Phase. Figure 1, located in the appendix, displays a diagram of these phases.

We can see that each phase of action has a purpose. The purpose of the first phase, the Respond Phase, is to 'act to promote safety and continuity' by focusing on providing the U.S. population with essentials such as financial assistance and medical care. This should be done as quickly as possible. Additionally, the government may implement certain mandates to stop the spread of the virus such as limiting travel, reallocating resources, and closing non-essential businesses and institutions3. Oftentimes, these orders can be enacted on a state-by-state basis, with state governments choosing to implement a variety of procedures such as mandatory stay-at-home orders or providing financial safety nets for citizens and businesses. Because there have been a wide range of unique state responses, it is possible to evaluate each state government's response individually to determine if there are any notably effective state approaches to the pandemic.

We can evaluate state response to the pandemic by discerning the attitude of the public: positive, neutral, or negative. This can be done using sentiment analysis. Twitter, a popular social media platform based in the U.S., allows users to send out brief and succinct comments to their followers. Tweets can pertain to a variety of topics, and most significant for our purposes, often give individuals the ability to publicly express their opinion, allowing us to collect data on the collective reaction to each state government's response to the pandemic.

### Methodology

**2.1 Data and Preprocessing**

For this project we analyzed a dataset consisting of 20,620,442 rows, with each row containing a

tweet related to COVID-19. The dataset includes only tweets written between the end of March and mid-April and twitter users from the US and its territories. Each row consists of the ID of the twitter account that posted the tweet, the location where it was posted and the tweet itself. The tweets were already pre-cleaned. All stop-words as well as tagged accounts were removed and words were converted to lowercase.

The location data included the full state name for some tweets but abbreviations for others. We modified the data so that all tweets only had state abbreviations as their location. There were also some tweets which had a seven-digit number in the location column. Since we did not know the significance of the number, and the amount of these tweets compared to the number of tweets in the entire dataset was relatively small, we chose to exclude those tweets from further analysis.

A major decision we faced was how to address the retweets. While an argument could be made for keeping the retweets in the analysis, we ultimately chose to exclude retweets from our analysis. We did this for the following reasons: first, retweets do not always represent a user's opinion. A twitter user can choose to retweet a tweet while disagreeing with the content of the original tweet. In fact, many twitter users have explicitly stated that their retweets do not equal their endorsement. Therefore, we believe that adding retweets may induce noise and lead to incorrect conclusions about the data. Second, we were also concerned about the number of bots and fake accounts that retweet tweets of politicians and celebrities, and therefore amplify their opinion. A report from 2018 (Fishkin) suggests that 61% of president Trump’s twitter followers are bots, as well as 40.9% of Obama’s followers.4 From this, we assume that a large amount of retweets are generated by bots and therefore do not represent an opinion of a real user. Including these retweets in the analysis would highly bias the analysis towards the opinions of politicians, and users with many followers, which is undesirable given the objective of this report.

**2.2 Assumptions**

In order to extract insights from the data we had to make a few assumptions. First, we assumed that tweets sent from a specific location are about that location. This might not always be the case, but in general, we believe that people are more likely to comment on the situation close to them rather than in other parts of the world. Second, we assume that the sentiment of tweets is directly related to the government’s response. Lastly, we assume that opinions on twitter can be seen as an indicator of the public opinion in a state, and we understand that this might lead to bias, which we try to account for in our final recommendations.

**2.3 Word Cloud**

The quickest way to understand what words twitter users utilize frequently is to create an image incorporating the prevalence of words within tweets. The cleaned dataset, without retweets, was processed in one giant string file then into a WordCloud function. WordCloud identifies the top 100 most important words from the entire data set (the individual words that appear the most frequently). The expectation was to see the most sensitive words that related to COVID-19 appear in WordCloud, such as "covid", "19", "covid19", "corona", "virus", "coronavirus". Another WordCloud was created by omitting the COVID-19 related sensitive words to reveal the significance of other words pertaining to coronavirus. To conduct the specific analysis, the COVID-19 sensitive keywords were added as stop words in the WordCloud attribute.

**2.4 Sentiment Analysis**

For the sentiment analysis we used the python package Vader. We calculated the sentiment score for each tweet. In the first step we aggregated the sentiment by twitter account. This avoids overweighting the opinion of people who tweet more than others in our analysis. We then aggregated the scores for each US state and recorded the states with the highest (Figure 8) and lowest(Figure 9) mean sentiment score. We did not include any US territories in our analysis. Additionally, we also classified the tweets into very positive (sentiment score over 0.5), positive (sentiment score between 0.05 and 0.5), neutral (sentiment score between 0.05 and -0.05), negative (sentiment score between -0.05 and -0.5) and very negative (sentiment score below -0.5) to gain a better understanding of the tweets.

**2.5 Cluster Analysis**

We applied k-means clustering to identify groups of tweets with common themes. We used the elbow method to determine the optimal number of clusters. After the number of clusters were set, we assigned each tweet to a cluster. We tried to find a theme for each cluster by analyzing the 50 most popular words in each cluster. Additionally, we calculated the sentiment for each cluster to gain further insights.

### Analysis and Results

**3.1 Word Cloud**

As expected, the first iteration of our WordCloud showed other terms for COVID-19 such as covid, corona, and coronavirus the most (Figure 2). To understand what significant words other than COVID-19 sensitive keywords, the additional stop words were added in our second iteration of the WordCloud (Figure 3). The new WordCloud provided more insight into what users talk about. The tier one words correspond to the overall situation at that time, such as "New Cases", "test", "death" which the death and tested new case raised to the peak (March to mid-April) with the virus outbreak in many cities. The public leveraged the key data based on tier one words to evaluate the safety of their location.

The tier two words concerned the topics of the changes after the virus outbreak in March, such as "lockdown", "work", "health", "home", "pandemic", and "trump." President Trump's name appears frequently suggesting that twitter users frequently tweeted about him. The initial assumption was that the announcement from President Trump had a significant impact on user's tweets. After evaluating the sentiment score of each state, we broke down and analyzed the WordCloud from each state. The states with the most negative sentiment score showed the keyword "trump" as tier one or tier two keywords compared to states with the most positive sentiment score did not contain it.

## 3.2 Sentiment Analysis

## What is the overall sentiment of the U.S. population to the government’s COVID-19 pandemic response?

To analyze sentiment, we classified tweet sentiment using Sentiment Intensity Analyzer from the package vaderSentiment. To begin the sentiment analysis, we removed retweets resulting in a dataset containing 7119562 tweets. We then conducted several different analyses. First, we checked the sentiment score for each tweet in our dataset. The range of possible scores were from -1 to +1. With -1 representing the most negative possible score a tweet can have, 0 as a neutral score, and +1 representing the most positive score a tweet can have. We then labeled each tweet as positive, neutral, and negative. Overall, tweet sentiment was positive as 2,859,767 tweets were labeled as positive. Figures 4 and 5, located in the appendix, display the classification for each label based on score as well as the number of tweets within each classification.

To understand tweet sentiment on a more detailed level, we then re-labeled each tweet as very positive, positive, neutral, negative, or very negative based on its sentiment score. Figures 6 and 7 display the classification for each label based on score has been shown below as well as the number of tweets within each classification. From this analysis, we can see that the highest number of tweets were classified as neutral, the second highest being positive, and the third highest classification as very positive. Tweet sentiment distribution does not seem to be significantly different on a more detailed level than the overall summary.

## 3.3 Cluster Analysis

To form a better understanding of the public’s sentiment regarding COVID-19, we performed a cluster analysis. Based on the elbow method we split the data into six different clusters. This section will discuss the unique characteristics of each cluster.  
  
Cluster one is the smallest cluster and consists of only 3.7% of all tweets. It is also the cluster with the most positive average sentiment of 0.23, far higher than the cluster with the second highest sentiment (cluster six, 0.07). Based on the word count frequency, it can be inferred that tweets in cluster one are opinions, based on the popular keywords like “feel”, “think”, “say” and “like”.

Cluster two is the biggest cluster with 44.6% of all tweets which makes it more difficult to find a common theme. The sentiment of the cluster can be classified as neutral (0.03) which further suggests that it is difficult to put a label on this cluster. Unique popular words in this cluster are “work”, “help” and “government” which suggests that unemployment is a common theme in this cluster.

The most popular words in cluster three (case, death, test, report, positive) imply that tweets in this cluster are mostly about numbers and facts around COVID-19. This hypothesis is further supported by the neutral sentiment of this cluster.

In cluster four, the most popular words include emojis, which leads to the conclusion that tweets in this cluster are more emotional. The most popular words in this cluster also include swear words, which further supports this hypothesis.

Cluster five is very similar to cluster three, with similar keywords that suggest that tweets in this cluster mostly discuss facts. However, the cluster has a more negative sentiment than cluster three. This, together with the high occurrence of the name Trump and the words “government” and “country” lead to the conclusion that a considerable number of tweets blame leadership for the bad coronavirus numbers.

Tweets in cluster six seem to focus on community. It is the only cluster with the words “support”, “help”, “thank” and “community”. A slightly positive sentiment (0.07, second highest) supports the notion that these tweets focus on some positive aspects of the crisis.

In conclusion, we found that there are six main topics that people discuss pertaining to COVID-19: personal opinions, work and unemployment, case and death numbers, emotions regarding COVID-19, the role of the government in the rising numbers and help and support during the crisis. Knowing this, governments should prioritize these topics in their response to the crisis.

## 3.4 Summary of the Data Analysis

*“Public conversation can help the world learn faster, solve problems better and realize we’re all in this together. Facing a devastating global pandemic really brings that, and Twitter’s role, to light.” - Jack Dorsey, Twitter CEO*

Using Twitter, our data analysis explores the sentiment of the public conversation on the unprecedented COVID-19 pandemic and the US government response to the crisis.

### From Sentiment Analysis

The five states with highest mean sentiment scores were: Rhode Island, North Dakota, New Hampshire, Vermont, and Nebraska. The five States with the lowest mean sentiment scores were: Texas, Mississippi, Nevada, Hawaii, and Louisiana. We observed that states with the lowest scores had less confirmed cases between mid-March and April than the states with highest scores, with an exception of Hawaii1 (Figure 10). While several factors come into play when determining potential causes more cases in some states, we were able to identify some best practices and actions taken by states with a relatively lower number of cases in the response period of Mid-march to April.

### From Cluster Analysis

To further our efforts in researching how different state governments dealt with the crisis and their response, we decided to perform cluster analysis to identify topics surrounding the situation and potential pain points for governments to address.

During our analysis we identified six clusters. The discussion within these clusters spanned various areas including but not limited to *personal opinions, work and unemployment, case and death numbers, emotions regarding COVID-19, the role of the government in the rising numbers and help and support during the crisis.*

Our insights from the sentiment and cluster analyses form the basis for our recommendations on how governments should navigate optimally, according to public response, through the Respond Phase of a pandemic.

## Conclusions and Recommendations

## 4.1 Recommendations for future pandemic response

While it is obvious that the US government’s response to the coronavirus outbreak had some major failures of judgement and inaction, analyzing the public conversation from that time can help inform and shape the governments’ response to the next pandemic.5 With the appropriate decision-making, though challenging, governments can control the rapid and silent spread of a novel virus through speedy calculated actions during the early days of the outbreak.

Using the findings from our analysis, we identified some key topics that governments should address early on while in the response phase of the pandemic.

* Communication: From our analysis, we discovered that people expressed their opinions and emotions often leading to high negative sentiment overall. Through high frequency of words like “help” and “need” and indication of their feelings about the government, it can be said that there was a lack of trust in the institution leading the emergency response. Hence, during the early days of the pandemic, active communication between the government and the public can help manage people’s fear and expectations. Misleading and inaccurate conversations can lead to mistrust during the ongoing pandemic response.
* Unemployment: The economic impact of the outbreak has left millions of Americans without a job. Workers from a variety of industries were laid off as businesses suffered. While steepening the economy curve will take time, during the early days of the pandemic, the government should focus on providing immediate financial support. The US Congress passed a $2 trillion coronavirus relief bill that provided survival and support to individuals and businesses. This will help most affected by the pandemic and, in the long-term, restart the economy.
* Flattening the curve: Slowing the spread of the virus will allow healthcare institutions to maintain the necessary amount of medical supplies to provide for the public. The government can implement various measures to decrease the burden on healthcare. The closure of public areas early on, such as schools and parks, can have a meaningful and major impact.
* Stay-at-home orders: In addition to adequate social distancing and closure of public spaces, stay at home orders that mandate positive cases to self-quarantine can prove to effectively curb the spread of the virus. It is also important to accurately assess the situation when thinking about reopening as this can escalate the number of cases with an increase in public outings.
* Testing and contact tracing: Widespread testing are key to monitoring the spread of the virus. Inability to do so can lead to a lack of accurate perception of the degree of harm the pandemic is causing and, in turn, lead to an ineffective response. At the onset of the pandemic, the government should begin facilitating the immediate production and distribution of test kits. This should be as flawless and quick as possible. Pairing adequate testing with contact tracing in the early days of a pandemic can help health officials to identify and isolate potentially infected people and can help the government strike the right balance between stay-at-home and civil liberty needs.

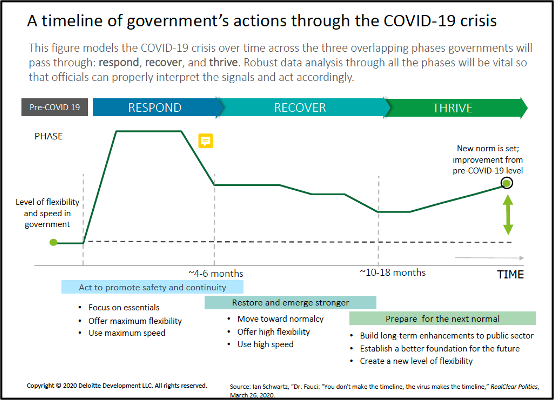
Public conversations can be a very informative source of data to evaluate errors in the government’s response to the pandemic. Hence, taking quick and decisive action based on these insights will prove to be an effective way to navigate through the response phase of the pandemic.

## 4.2 Limitations and further research

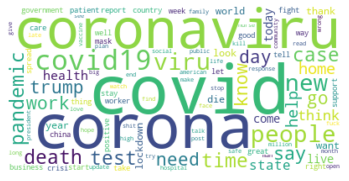
As already mentioned in the assumptions section, tweets can only be seen as an indication for the public opinion as twitter users are not a representative sample of the population. This was noted when generating our recommendations. Additionally, this is the first time a vast majority of the population has experienced a pandemic. Reactions to other pandemics in other parts of the world may be different, if people are more familiar with the situation. Future research might take these points into consideration to further refine our recommendations.

## Appendices

#### Figure 1. A timeline of government’s actions through the COVID-19 crisis



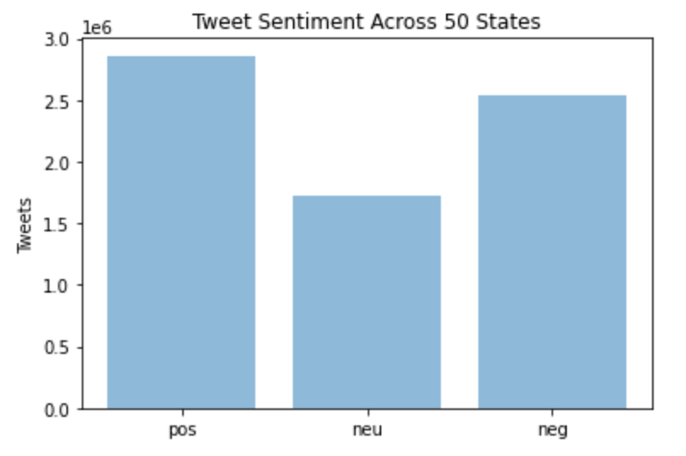
#### Figure 2. Initial WordCloud

****

#### Figure 3. WordCloud without COVID-19 sensitive words

****

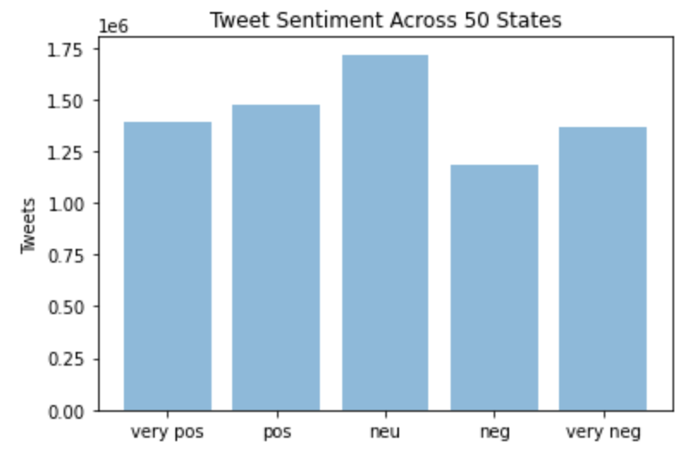
#### Figure 4. Tweet Sentiment Across 50 States (Positive, Neutral, Negative)

****

#### Figure 5. Number of Tweets for Each Sentiment (Positive, Neutral, and Negative)

|  |  |  |
| --- | --- | --- |
| Sentiment Score | Sentiment Label | Number of Tweets |
| .05 to 1 | Positive | 2,859,767 |
| -.05 to .05 | Neutral | 1,716,620 |
| -1 to .05 | Negative | 2,543,175 |

#### Figure 6: Tweet Sentiment Across 50 States (Very Positive, Positive, Neutral, Negative, Very Negative)



#### Figure 7. Number of Tweets for Each Sentiment (Very Positive, Positive, Neutral, Negative, Very Negative)

|  |  |  |
| --- | --- | --- |
| Sentiment Score | Sentiment Label | Number of Tweets |
| .50 to 1 | Very Positive | 1,389,157 |
| .06 to .49 | Positive | 1,470,602 |
| -.06 to .05 | Neutral | 1,716,628 |
| -.49 to - .05 | Negative | 1,180,976 |
| -1 to -.50 | Very Negative | 1,362,199 |

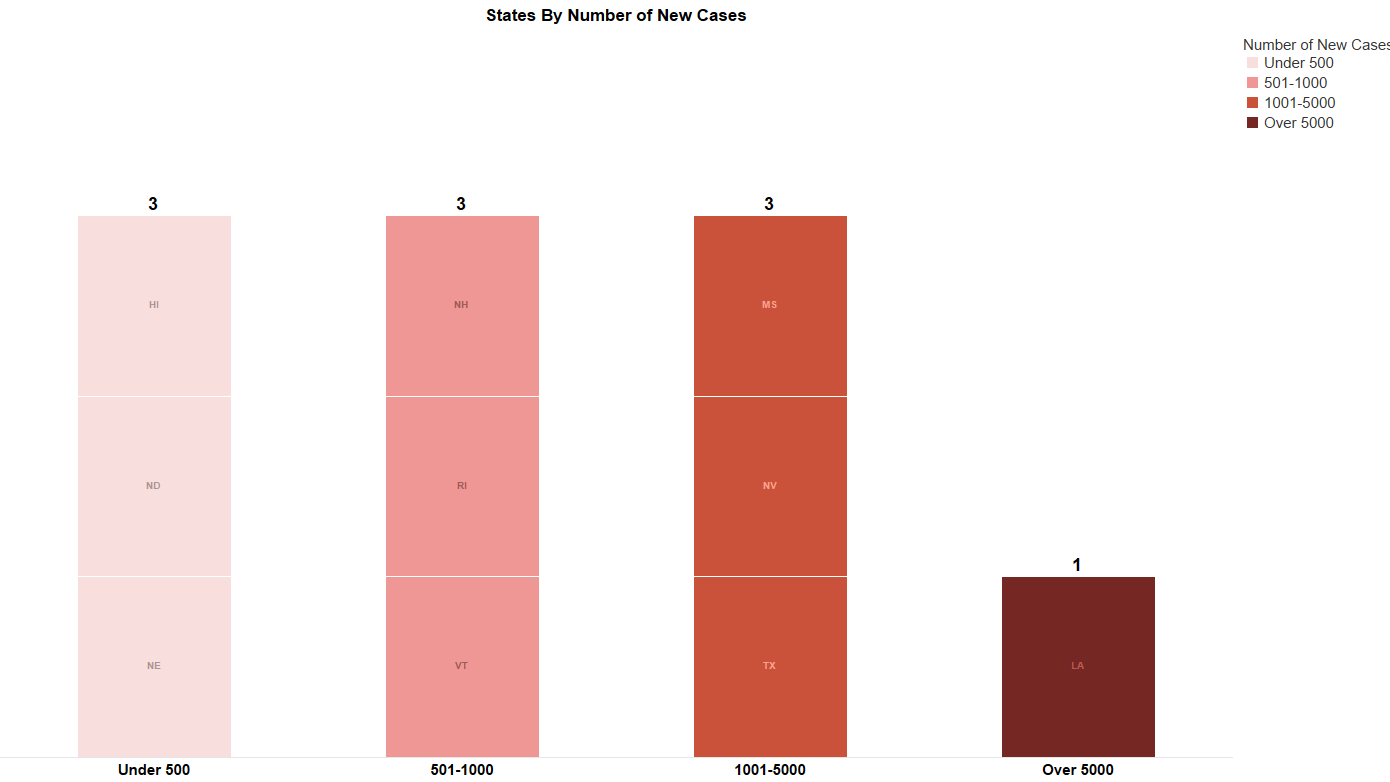
#### Figure 8: States with the Highest Mean Sentiment Scores

|  |  |
| --- | --- |
| State | Mean Overall Score |
| Rhode Island | .194 |
| North Dakota | .171 |
| New Hampshire | .167 |
| Vermont | .157 |
| Nebraska | .150 |

#### Figure 9: States with the Lowest Mean Sentiment Scores

|  |  |
| --- | --- |
| State | Mean Overall Score |
| Texas | .039 |
| Mississippi | .041 |
| Nevada | .050 |
| Hawaii | .051 |
| Louisiana | .054 |

#### Figure 10: States by Number of Cases between 03/15 to 04/01. *Can be seen that states with higher sentiment scores overall have lower numbers of cases.*



**References**

1. COVID-19 Map. (n.d.). Retrieved August 5, 2020, from https://coronavirus.jhu.edu/map.html

2. Craig, J. (2020, August 5). Charts: How the U.S. Ranks On COVID-19 Deaths Per Capita - And By Case Count Retrieved August 5, 2020 from https://www.npr.org/sections/goatsandsoda/2020/08/05/899365887/charts-how-the-u-s-ranks-on-COVID-19-deaths-per-capita-and-by-case-count

3. Eggers, W. (2020, April 16). Government’s response to COVID-19: From pandemic crisis to a better future: Retrieved August 5, 2020 from https://www2.deloitte.com/us/en/insights/economy/COVID-19/governments-respond-to-COVID-19.html

4. Fishkin, R. (2018, October 09). We Analyzed Every Twitter Account Following Donald Trump: 61% Are Bots, Spam, Inactive, or Propaganda. Retrieved August 10, 2020, from https://sparktoro.com/blog/we-analyzed-every-twitter-account-following-donald-trump-61-are-bots-spam-inactive-or-propaganda/

5. Zurcher, A. (2020, April 01). Coronavirus: Things the US has got wrong - and got right. Retrieved August 14, 2020, from https://www.bbc.com/news/world-us-canada-52125039