Data Wrangling Final Report

Introduction

For my final project, I decided to work with data relating to the current Coronavirus pandemic and inspect how the number of cases and the number of deaths reported correlates with headlines in the news over time, specifically the New York Times. I felt this would be an interesting problem to study since a lot of current controversy over how the virus is being handled in the United States deals with the questionable accuracy of reported figures, as well as the perceived response by the U.S. government to controlling the pandemic. With how much the pandemic has affected our daily lives, and with how reliant we are on the news, with the enforcement of social distancing and self-quarantining, I felt that using what I learned from this course could provide some potentially useful insight on the entirety of the pandemic and how the pandemic is being perceived and handled.

Data Description

The data comes from the NYT API used earlier in the semester, which provides two different data sources; one being the day-by-day number of cases and deaths in 2020 by state as a result of the virus, and the other source being the news articles published by the NYT since 1851. Since my primary focus is on how the virus and the news are affecting each other, I specifically limit the time frame of my data from January 2020 to the current day. For the articles, I am interested in the semantic significance of the pandemic on the articles being written and so for the sake of simplicity, I focus on the lead paragraphs of the archived articles. The reason I choose the lead paragraphs is because I noticed during my analysis that using the article abstract (which was provided) contained many useless semantics and words that otherwise would have been captured by the lead paragraph anyway. On the other hand, using only a title would make my words data susceptible to things such as clickbait, or exaggerated titles. I found that a leading paragraph, which would provide enough context to interest the reader while also drawing their attention, would be an appropriate source of words to use for the analysis.

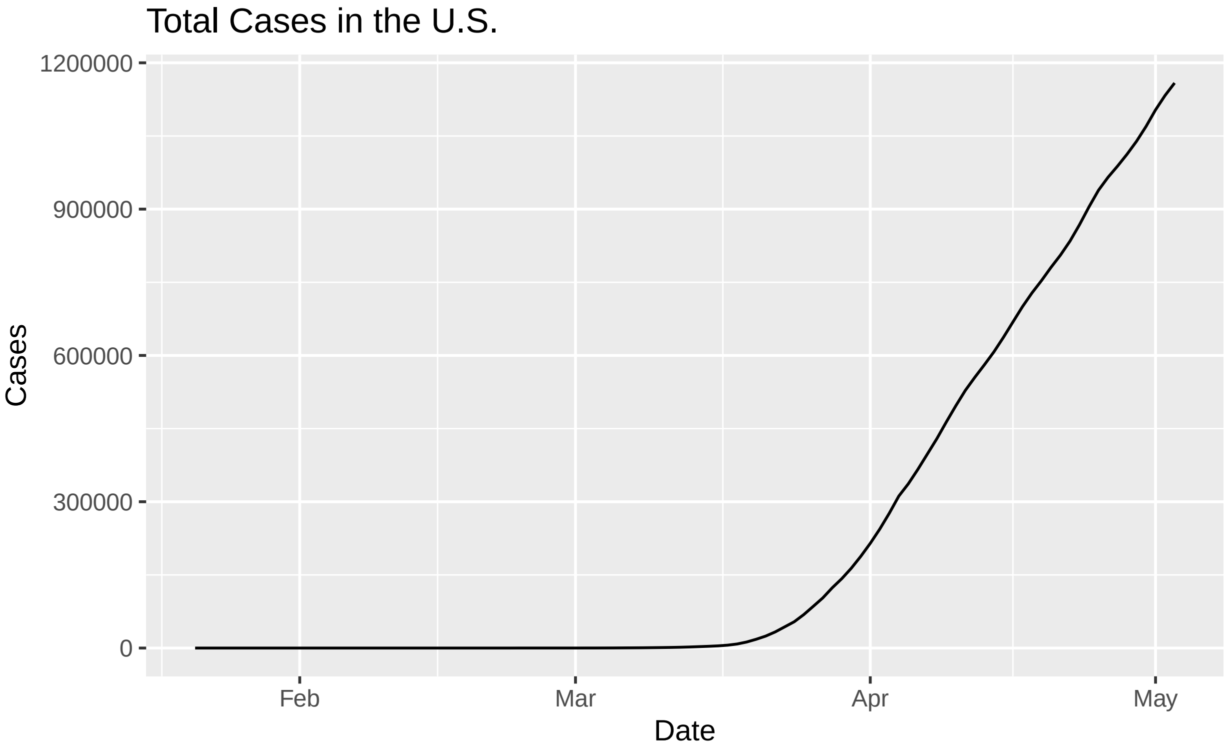
Coronavirus Cases/Deaths by Month

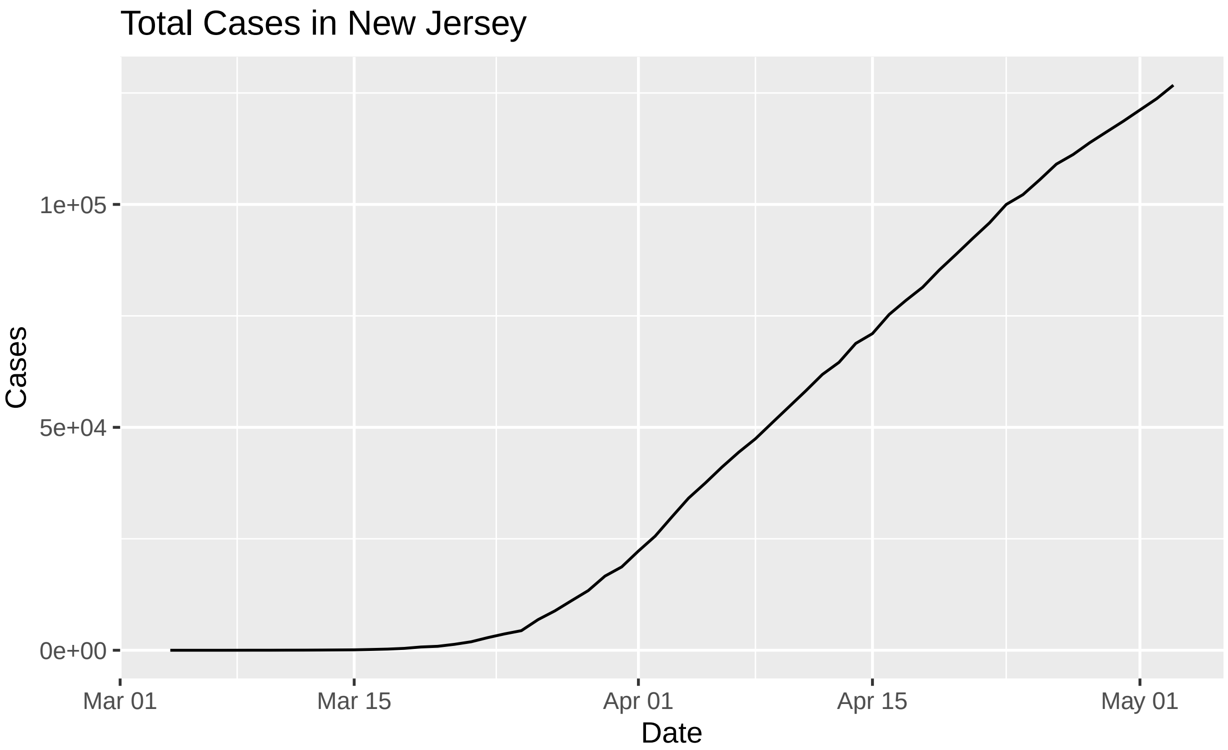
Once aggregated over all states in the country, we see that the number of reported cases by the New York Times seems to increase linearly after mid March. It is surprising to see this sharp increase occur in the middle of March, as opposed to February when articles were already being written about the coronavirus.

It is both interesting and concerning to see that the deaths reported in aggregate follow a similar trend as well. With many experts speaking about a plateau in the number of cases and deaths in other countries, it is surprising to see that the trend in the deaths and cases in the U.S. remains consistent.

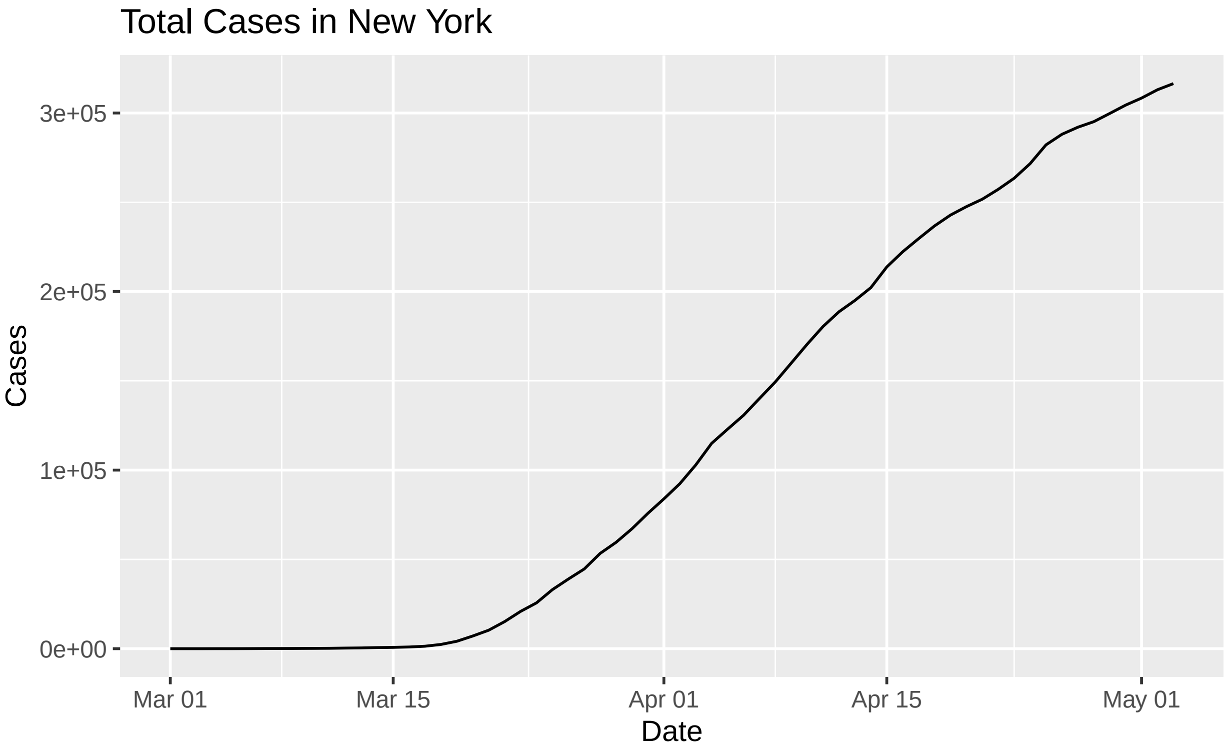
Finally, I felt that it would be worth seeing the trend of cases in New York, with most of the coverage being from New York, and also a majority of the cases and controversy stemming from the general tristate area. Unsurprisingly, the trend rises sooner than that of the overall U.S. cases trend. There also seems to be more of an approaching peak to the trend as well, whereas the overall cases seemed to continue increasing linearly.

```
## Parsed with column specification:
## cols(
##   date = col_date(format = ""),
##   state = col_character(),
##   fips = col_character(),
##   cases = col_double(),
##   deaths = col_double()
## )
```

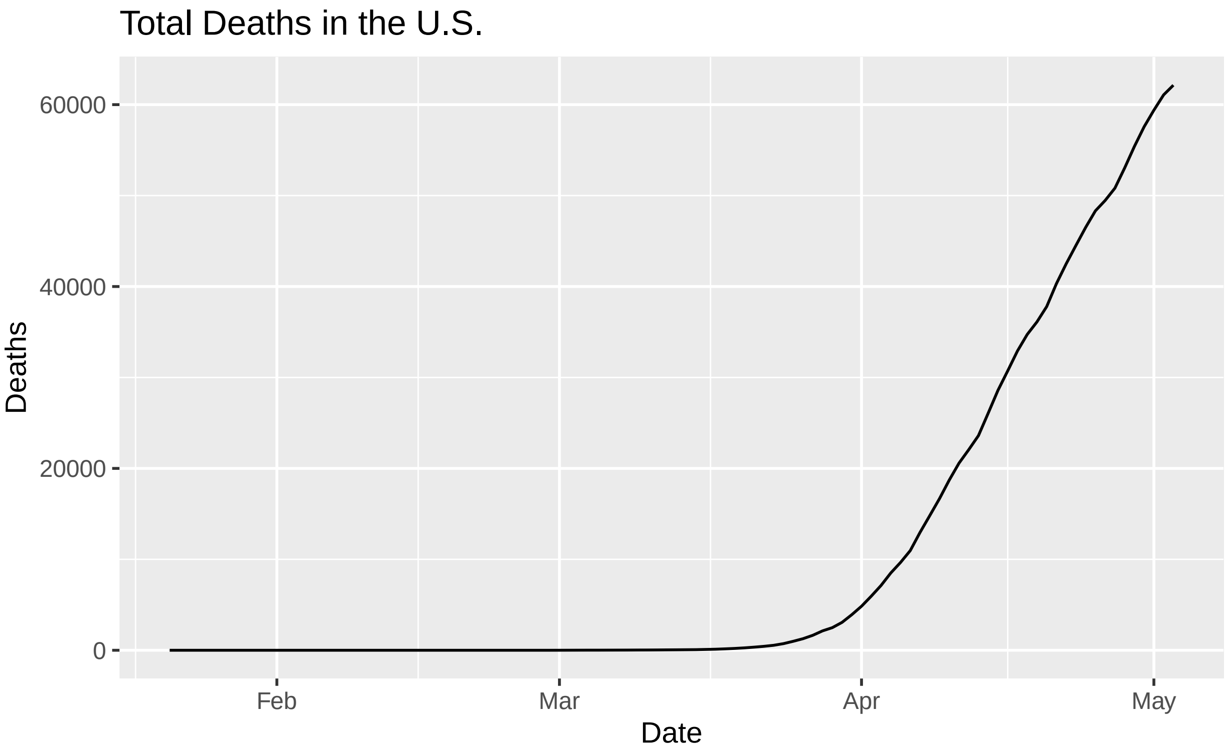




Total Cases in New Jersey



Total Cases in New York



2020 Month-by-Month Cases/Deaths

In this section, I look at the top 10 non-stop words used in the leading paragraphs of the articles of each month.

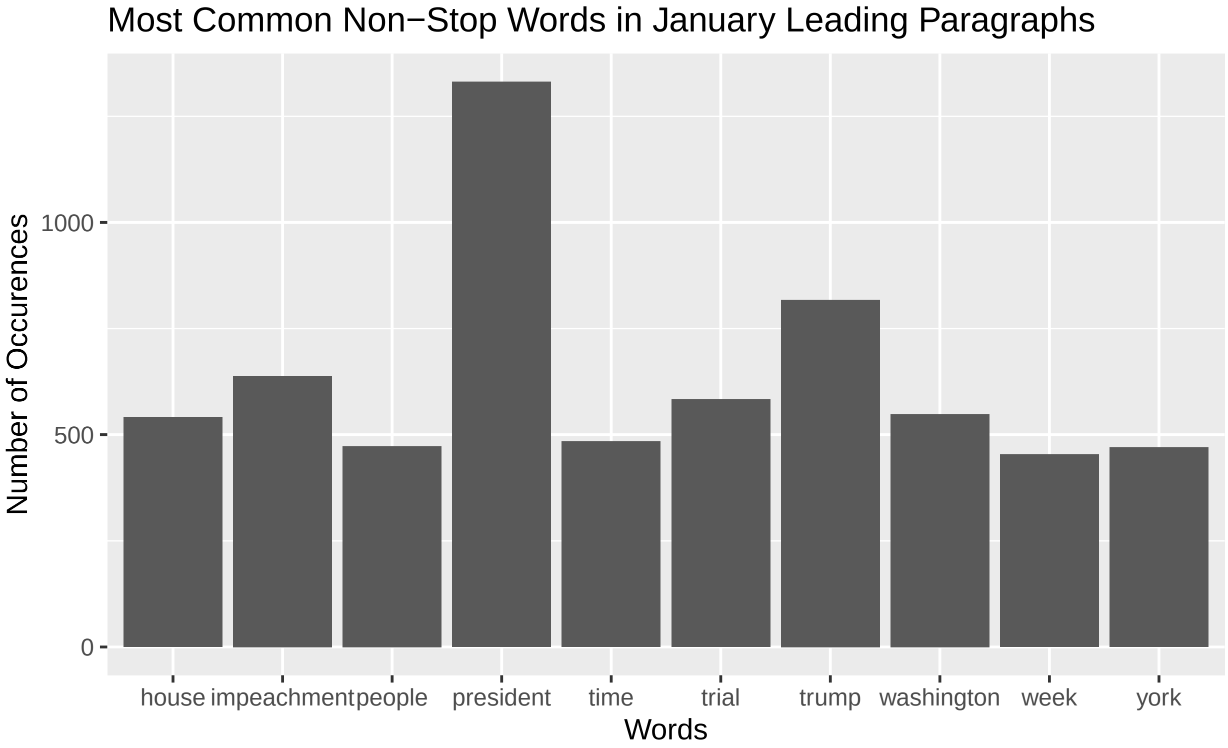
`In taking a preliminary look at the top ten non-stop words for January of 2020, it is`

interesting to note that most of the keywords are related to the news, impeachment, and the president. This is sensible, as the coronavirus still had nearly no coverage around the world, and meanwhile, the impeachment of President Trump was also taking place primarily during this month. This explains why keywords “president” and “trump” are the most common even compared to the other words.

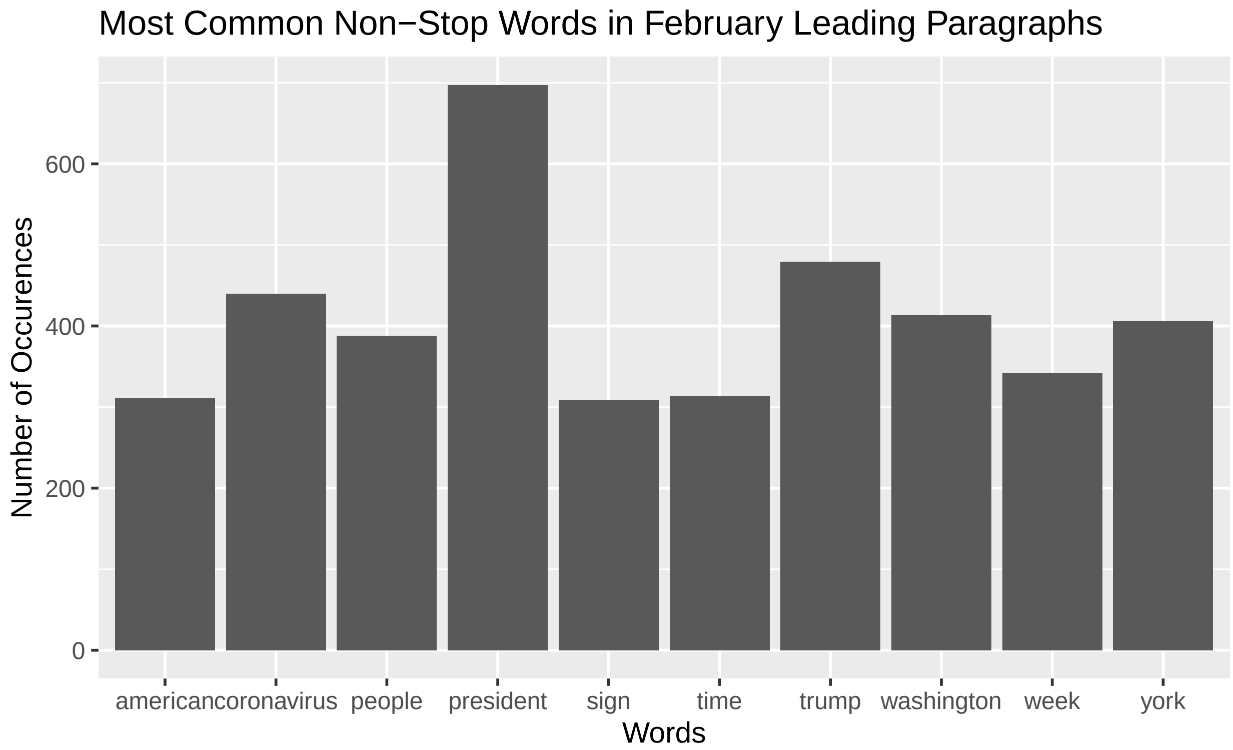
We note that for February, the word “coronavirus” is now in the top 10 most common words, even while being less common than “president” and “trump” still. This is an indicator of the coming pandemic, as well as the conclusion of President Trump’s impeachment trial compared to the occurrences of January. Something interesting to note is that the relative distribution of the occurrences is similar between January and February, with the primary difference being that January contains the word “impeachment” while February contains “coronavirus”, with both having similar occurrences of both “president” and “trump”. This could be indicative that the frequency of “president” and “trump” could be due to media attention to his response to early stages of coronavirus as much as his impeachment trial. Even so, there are still far more occurences of “president” and “trump” in January.

Finally, looking at March and April (both shown, however April looks similar), we see that the occurrences of “coronavirus” far exceed any other word, with the addition of “pandemic” and “home”. These words are once again indicative of the effect of the pandemic. It is also still interesting to note that “president” and “trump” are both still in the top ten occurences, implying that his involvement in the U.S’s response to the pandemic keeps him very relevant in the news.

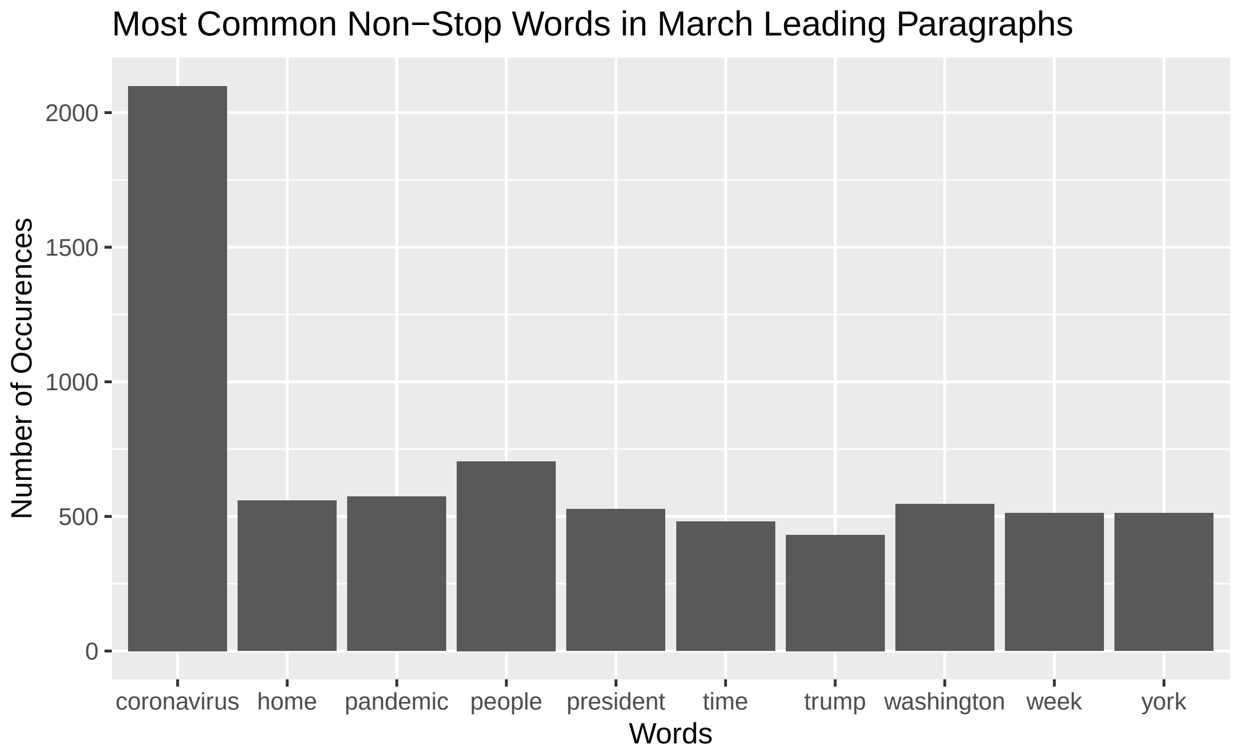
```
## Joining, by = "word"
```



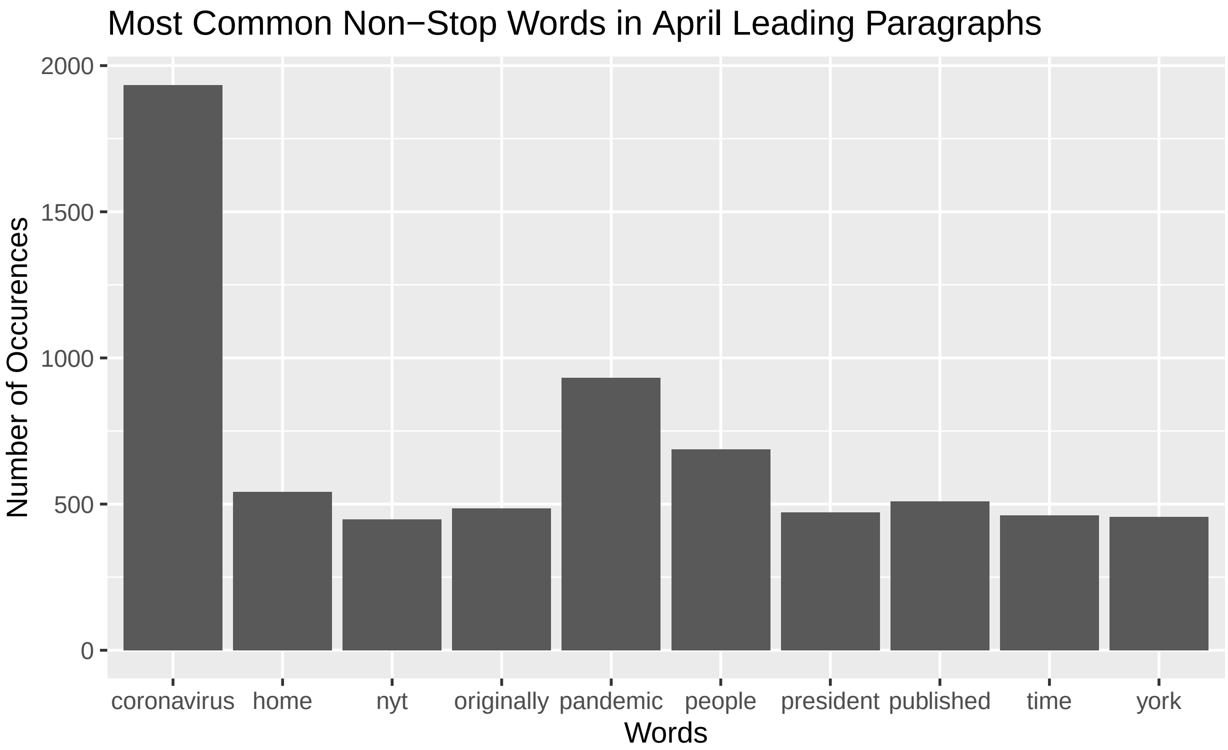
```
## Joining, by = "word"
```


Most Common Non−Stop Words in February Leading Paragraphs

```
## Joining, by = "word"
```


Most Common Non−Stop Words in March Leading Paragraphs

```
## Joining, by = "word"
```



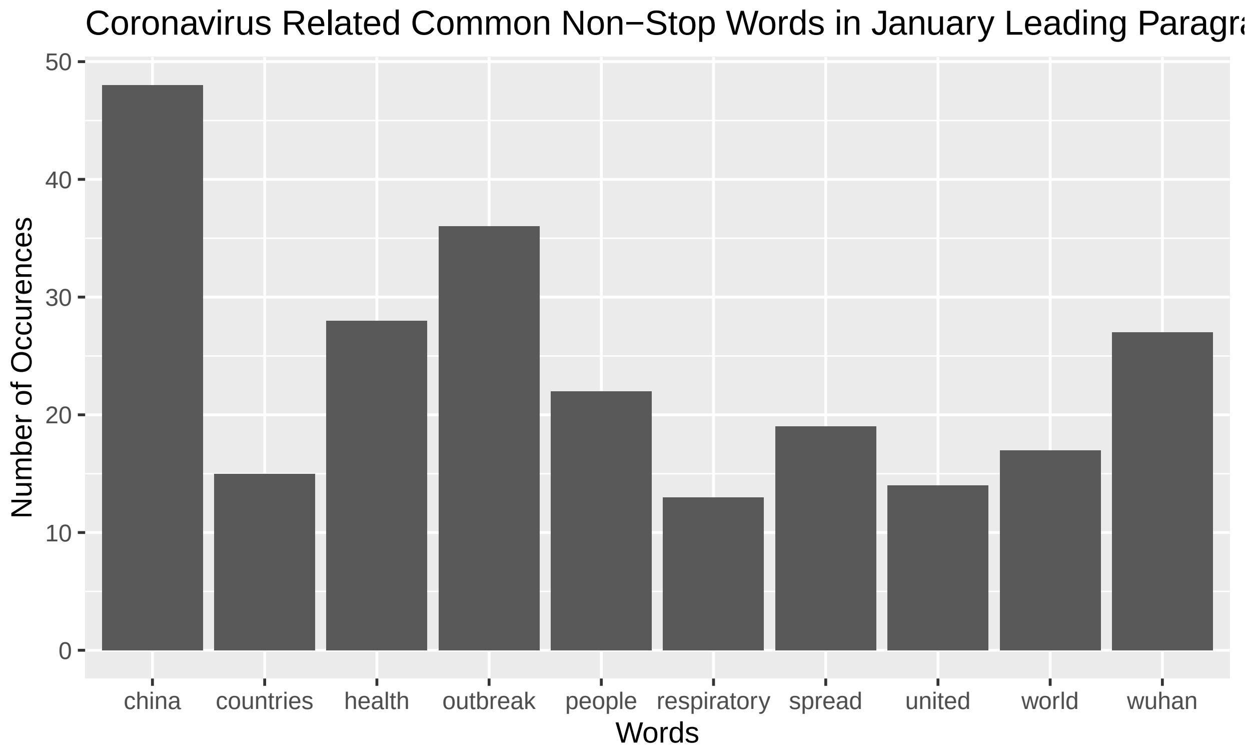
Coronavirus Related Articles Over Time

Next, since part of my focus is on how the perception of coronavirus has shifted over time, I decided to look at the articles about coronavirus and see if the most common words in these articles specifically would reveal anything interesting over the months.

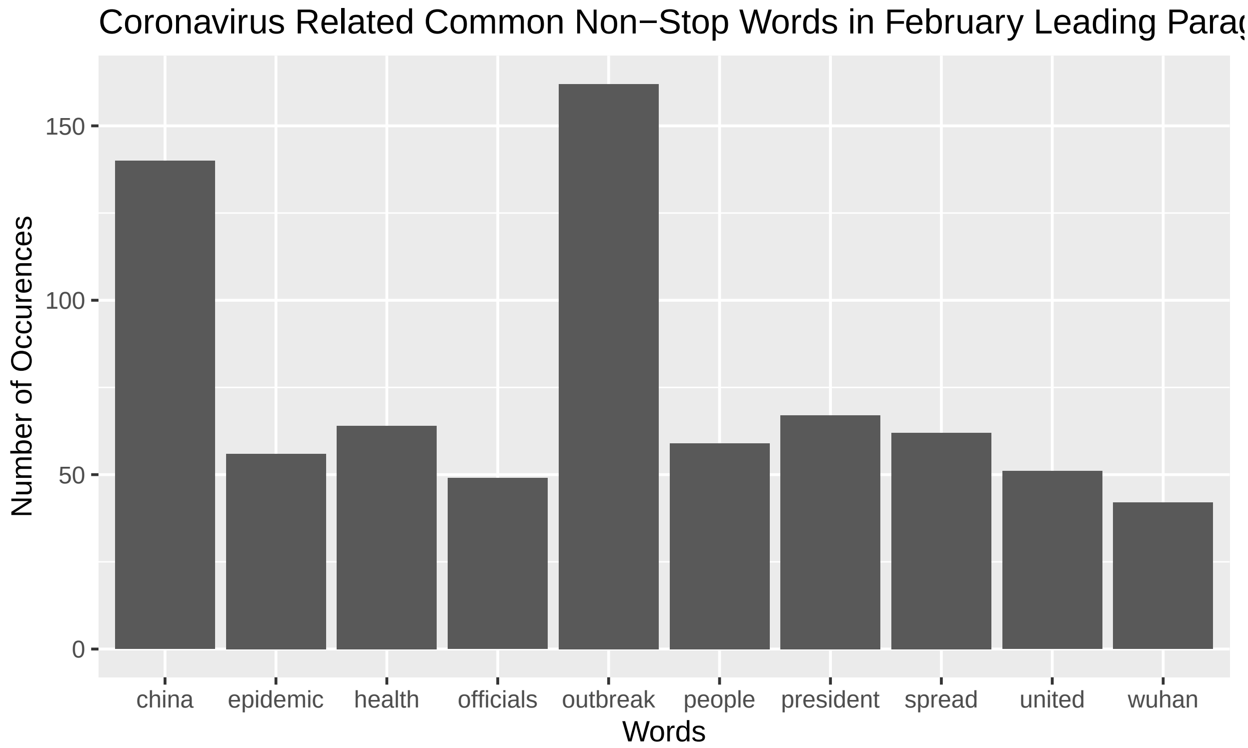
We see that the majority of the articles in January that were focused on coronavirus, or had the coronavirus keyword in the leading paragraph, were also primarily emphasizing key words "china" and "wuhan". I find this expected but also interesting in that our previous analysis suggested very little coverage of the virus in January specifically due to the impeachment trial. With other keywords such as "united", "world" and "spread", I found this to be an indicator that the NYT focuses more heavily on domestic news than international news, based on the connection between the earlier analysis and this one.

We see that in contrast with the January stop words, the distribution and general theme of the words is different, with the exclusion of "china" and "wuhan" and the increase in occurrences of "pandemic". Another observation that can be of note is how the trend of words has shifted to be more political in April while those of January were more around international coverage of world health. For instance, words like "health" and "respiratory" are now replaced with words like "washington" and "york", indicating that news in January was more focused on educating the public about the virus, while news in April was focused more on the response by President Trump.

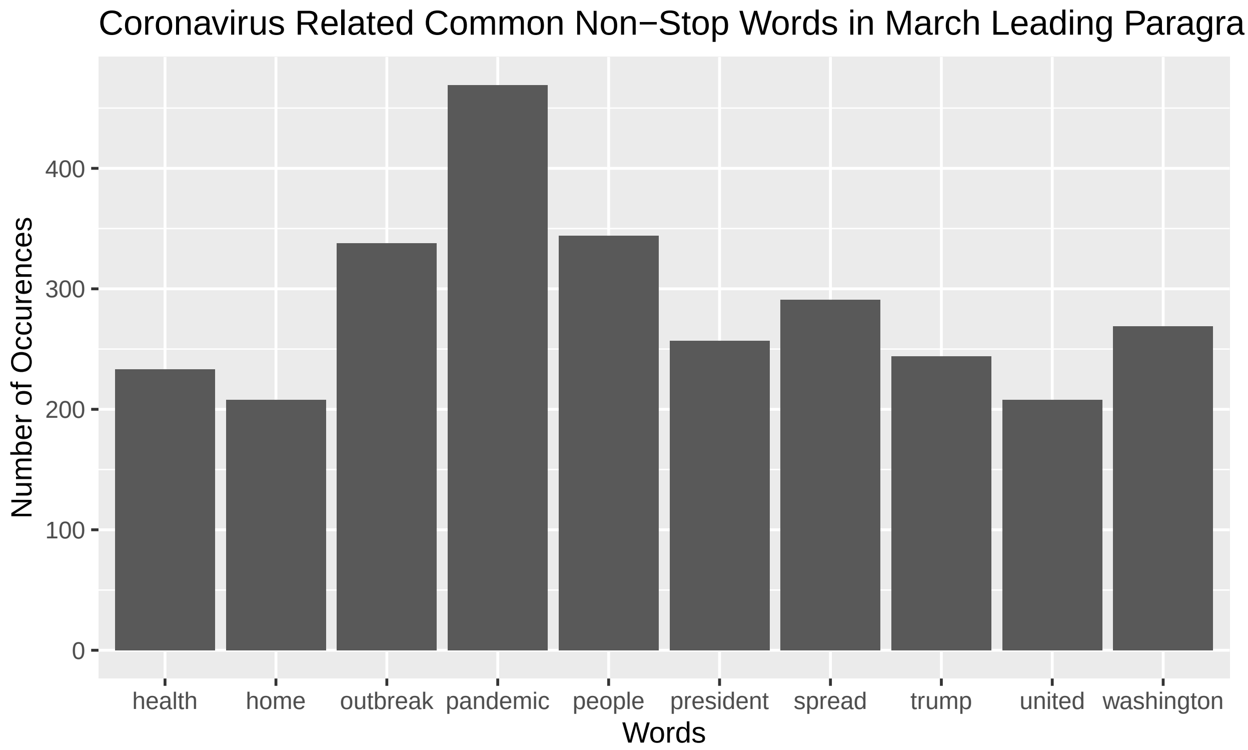
```
## Joining, by = "word"
```



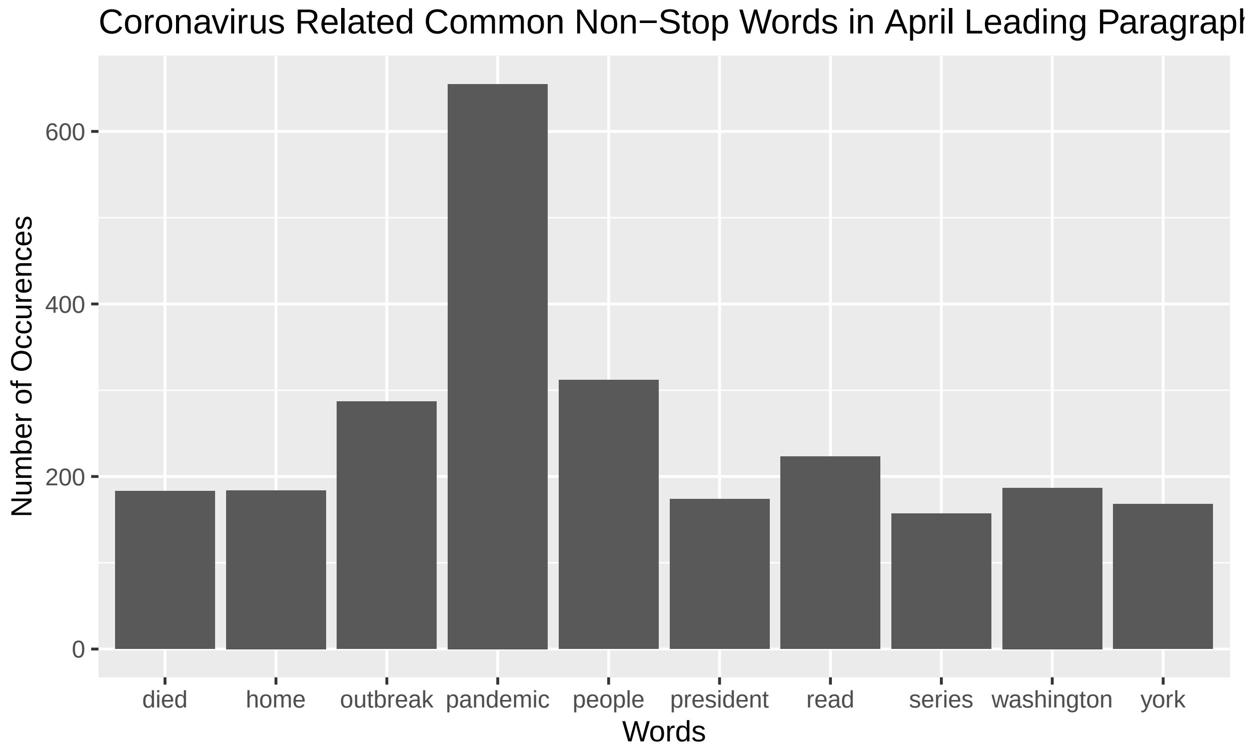
```
## Joining, by = "word"
```



```
## Joining, by = "word"
```



```
## Joining, by = "word"
```



Conclusions

Overall, there were several interesting observations to make about the progression of the words used in articles alongside the rising cases and deaths of the virus. Firstly, we noted that there was little coverage with the virus in the earlier months such as January and early February due to most coverage being on President Trump's impeachment trial. However, by analyzing articles in those months that did mention coronavirus, we saw based on the next most common words in those articles that much of the content was focused on educating the public about the virus, as it was still relatively new to the world. We noticed a dramatic shift in the theme of the articles as the cases rose in March and April, as articles began mentioning President Trump, safety measures such as staying at home and referring to the virus as a pandemic. Something small to notice was how early on, the word "epidemic" was used, while "pandemic" was used more frequently later, implying that the general public perceived the virus to be under control during early months.

Issues with Cleaning

Primarily, my issues with working with the dataset stemmed from the fact that most of the data provided from the API was stored in data frames when read into R. This made it such that I needed to extract further even after reading in my data. However, the API data format provided was surprisingly straightforward and compliant with R data frames, and so I did not have to scrape as much for the articles data.