COVID 19 Data Analysis

Courtney

06/12/2021

In this report we are looking at a COVID-19 data from John Hopkins, in order to determine some global and US trends regarding COVID-19 cases, deaths and the differences in those trends across countries and states. Tidyverse, Lubridate and ggplot2 are used in order to clean, transform and visualize the data and the analysis leads to some very interesting outcomes. A few models are created of the model also to further understand the relationships between variables.

Read in Data

Start by reading in the data which is retrieved from GitHub. The data that is being used is from John Hopkins University Center for Systems Science and Engineering (CSSE). It is split up into files, two containing data for the US, split up into cases and deaths, and two for global cases split up the same way. All four of these data sets are read into separate variables and further investigated later on in this report. The URL for the main Github folder containing the data is given below, and from that page it is possible to navigate to each of the individual data sets.

The URL the data is found at is https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series

```
# LOCATE DATA
# Data is contained in four different files
# All coming from the same folder
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_cov

file_names <- c("time_series_covid19_confirmed_US.csv",
                "time_series_covid19_confirmed_global.csv",
                "time_series_covid19_deaths_US.csv",
                "time_series_covid19_deaths_global.csv"
)

urls <- str_c(url_in, file_names)

# READ IN THE DATA
global_cases <- read.csv(urls[2]) # imported 3342 rows, 675 column
us_cases <- read.csv(urls[1]) # imported 280 rows, 668 column
global_deaths <- read.csv(urls[4]) # imported 3342 rows, 675 column
us_deaths <- read.csv(urls[3]) # imported 280 rows, 668 column
```

Now that the data is loaded into our variables, we can start to examine and make sense of it. The data in its current form is not convenient for us to work with, or even display. There are a few steps we can take to resolve this issue, and those will be shown in detail in the next section.

Clean up the Data

Now we can look at the data and begin tidying the data and transforming it. The first issue addressed is that each day in the data has its own column. We want to pivot that so that every row repersents one day. Also we want to combine the global cases and deaths as well as the US cases and deaths. Additional tiding steps included removing Lat and Long columns as they will not be used, renaming a few variables, converting date variables to the proper class and some preliminary filtering.

Although similar steps are taken to clean each of the data frames, the are slight differences in the global and US data and is done separate for each. The data after being tidied and merged is shown below unter the Global and US headings.

Global data

```
## Joining, by = c("Province.State", "Country.Region", "date")
```

```
## # A tibble: 6 x 5
##   Province_State Country_Region date       cases deaths
##   <chr>          <chr>          <date>     <int>  <int>
## 1 ""             Afghanistan    2020-02-24     5      0
## 2 ""             Afghanistan    2020-02-25     5      0
## 3 ""             Afghanistan    2020-02-26     5      0
## 4 ""             Afghanistan    2020-02-27     5      0
## 5 ""             Afghanistan    2020-02-28     5      0
## 6 ""             Afghanistan    2020-02-29     5      0
```

US data

```
## Joining, by = c("Admin2", "Province_State", "Country_Region", "Combined_Key",
## "date")
```

```
## # A tibble: 6 x 8
##   Admin2  Province_State Country_Region Combined_Key date       cases Population
##   <chr>   <chr>          <chr>          <chr>        <date>     <int>      <int>
## 1 Autauga Alabama        US             Autauga, Al~ 2020-01-22     0      55869
## 2 Autauga Alabama        US             Autauga, Al~ 2020-01-23     0      55869
## 3 Autauga Alabama        US             Autauga, Al~ 2020-01-24     0      55869
## 4 Autauga Alabama        US             Autauga, Al~ 2020-01-25     0      55869
## 5 Autauga Alabama        US             Autauga, Al~ 2020-01-26     0      55869
## 6 Autauga Alabama        US             Autauga, Al~ 2020-01-27     0      55869
## # ... with 1 more variable: deaths <int>
```

Transform the Data

Looking at the tables above it is easy to compare the two data frames. You can see that there are more columns in the US data. Going back to the source or the data is is determined that Admin2 shows the county name, and the Combine_Key is both the county, state names and country put together. A combined_Key column can be created in the global data by merging the Province_State and Country_Region columns. Also the global data frame is missing population data for all the countries. This will be important as normalizing data with population will lead to more accurate comparisons. Population data is located on Github, also from John Hopkins University Center for Systems Science and Engineering, and added to our data in the following.

The URL for that data is "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/UID_ISO_FIPS_LookUp_Table.csv

```
## Joining, by = c("Country_Region", "date", "cases", "deaths", "Combined_Key")
```

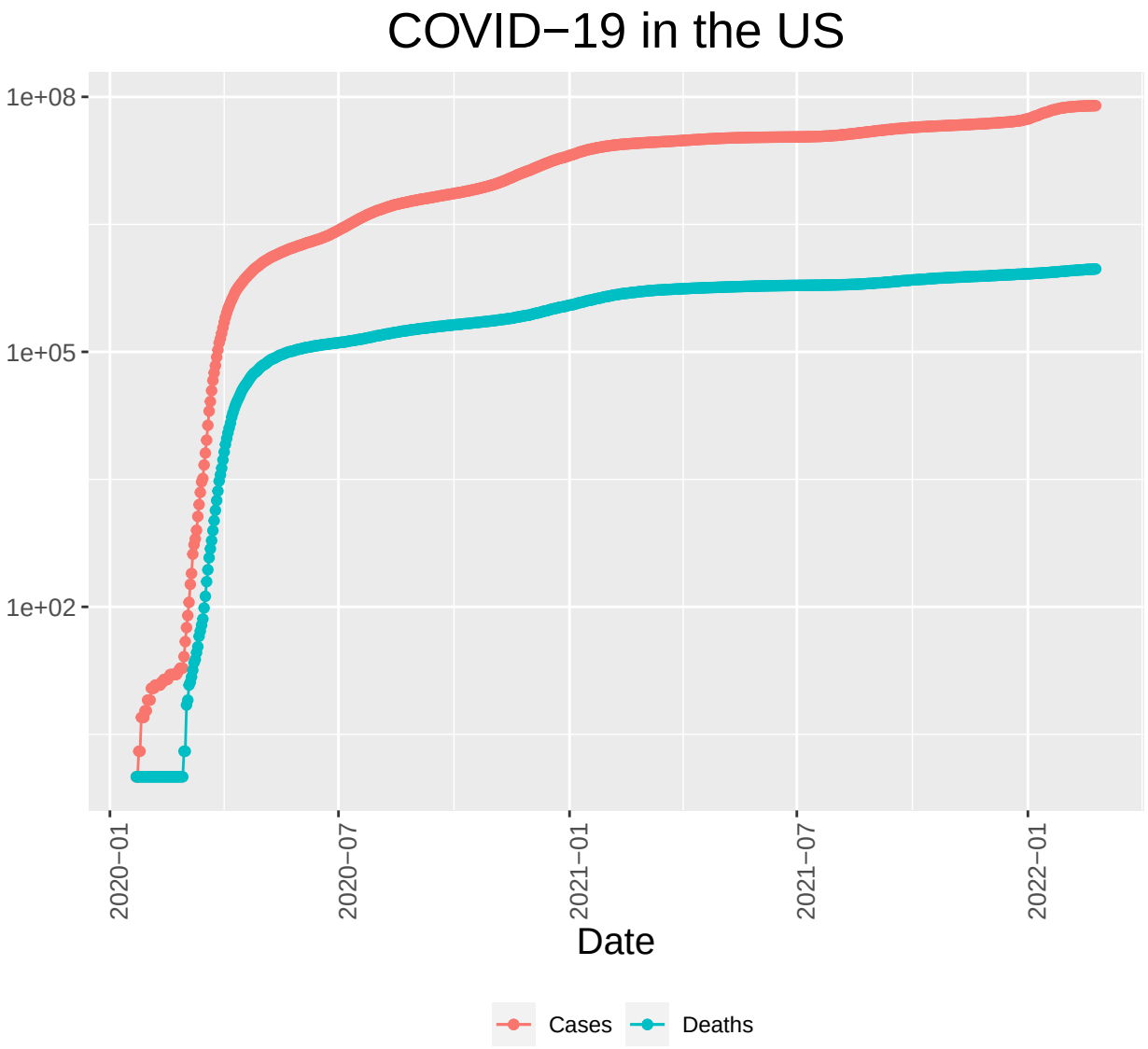
```
## # A tibble: 6 x 7
##   Country_Region date       cases deaths Population Combined_Key  Province_State
##   <chr>          <date>     <int>  <int>      <int> <chr>         <chr>
## 1 Afghanistan    2020-02-24     5      0   38928341 , Afghanistan ""
## 2 Afghanistan    2020-02-25     5      0   38928341 , Afghanistan ""
## 3 Afghanistan    2020-02-26     5      0   38928341 , Afghanistan ""
## 4 Afghanistan    2020-02-27     5      0   38928341 , Afghanistan ""
## 5 Afghanistan    2020-02-28     5      0   38928341 , Afghanistan ""
## 6 Afghanistan    2020-02-29     5      0   38928341 , Afghanistan ""
```

```
## # A tibble: 6 x 8
##   Admin2  Province_State Country_Region Combined_Key date       cases Population
##   <chr>   <chr>          <chr>          <chr>        <date>     <int>      <int>
## 1 Autauga Alabama        US             Autauga, Al~ 2020-01-22     0      55869
## 2 Autauga Alabama        US             Autauga, Al~ 2020-01-23     0      55869
## 3 Autauga Alabama        US             Autauga, Al~ 2020-01-24     0      55869
## 4 Autauga Alabama        US             Autauga, Al~ 2020-01-25     0      55869
## 5 Autauga Alabama        US             Autauga, Al~ 2020-01-26     0      55869
## 6 Autauga Alabama        US             Autauga, Al~ 2020-01-27     0      55869
## # ... with 1 more variable: deaths <int>
```

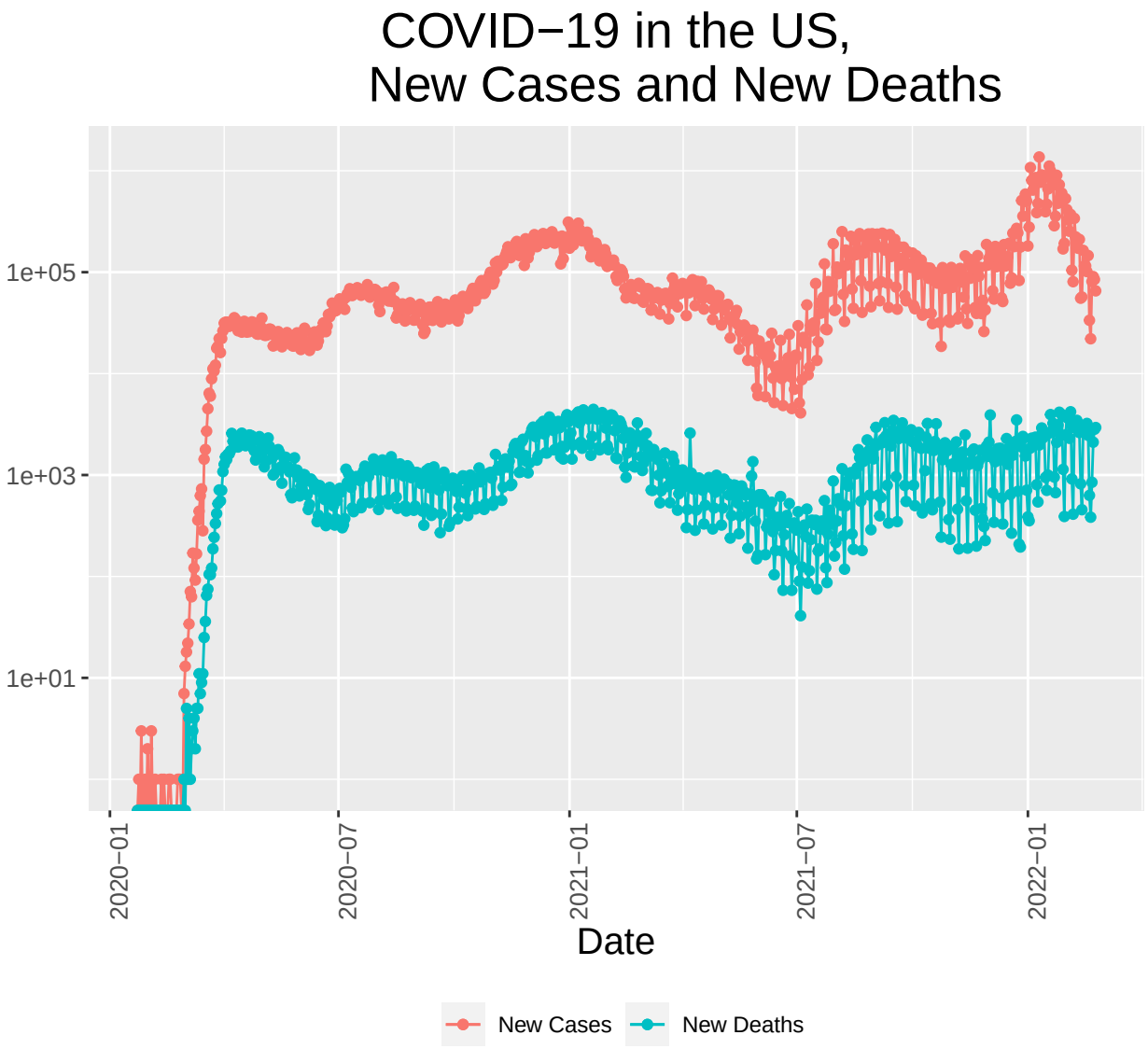
Now that both data frames are in similar formats, it will be alot easier in to analyze, visualize and model to pull out trends and correlations.

Visualize the Data

Now we can begin to plot the data in order to determine what we can learn from it. In order to visualize certian plots we can group data in different ways to look at different aspects of it.

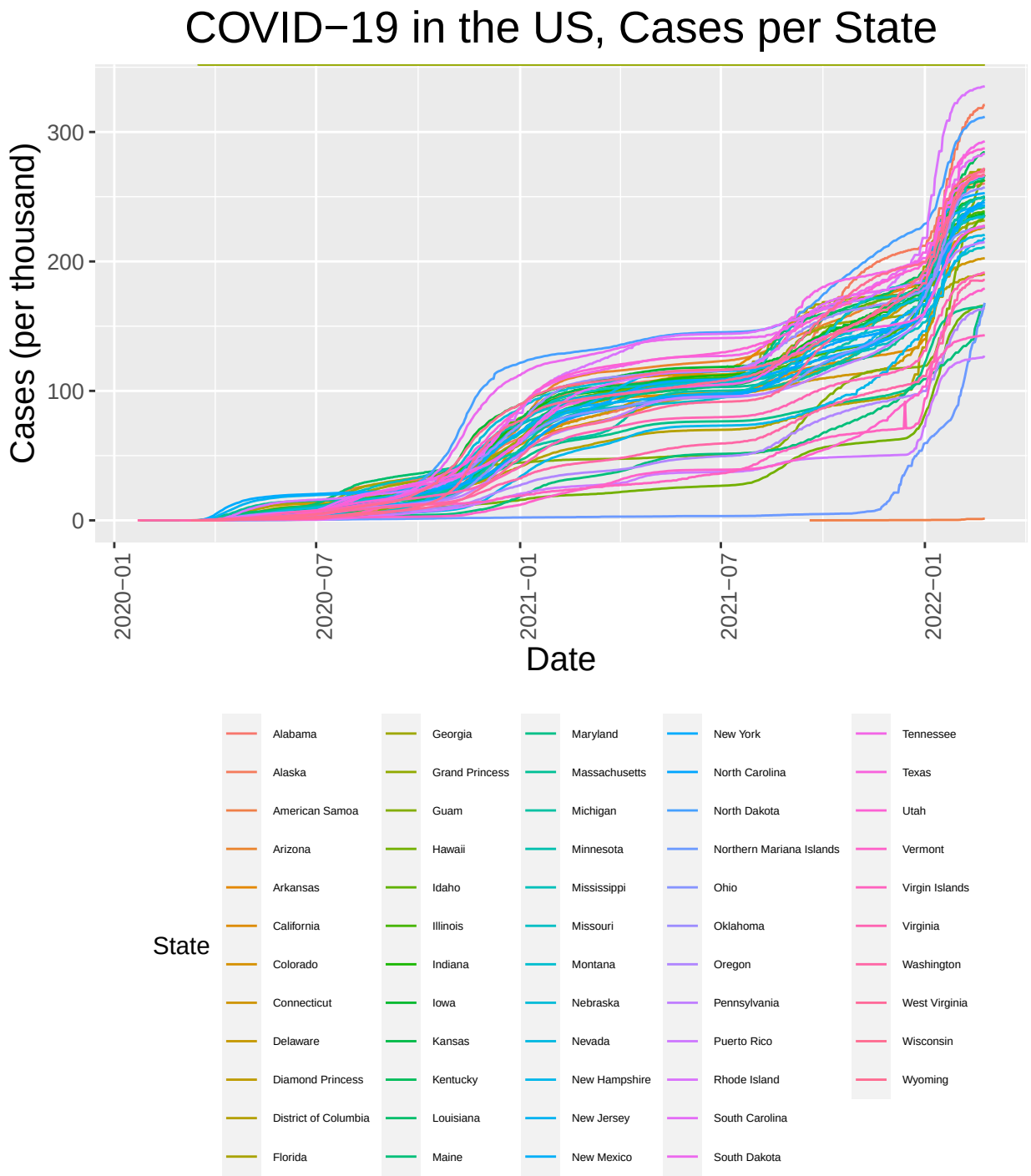


This plot shows the number of cases present in the US since COVID-19 first appeared there. Its quite interesting as you can see there is quite a steep increase in cases and deaths right at the beginning of the pandemic, however it looks like the increase in cases is getting smaller. This could be because of the way the plot is being shown on a logarithmic scale. The following graph adjusts the plot so that instead of showing the total number of cases and deaths, it shows the new cases and deaths for each day, which makes it easier to see trends over periods of time, and you can see clearly that there are still a large number of new cases and deaths every day.

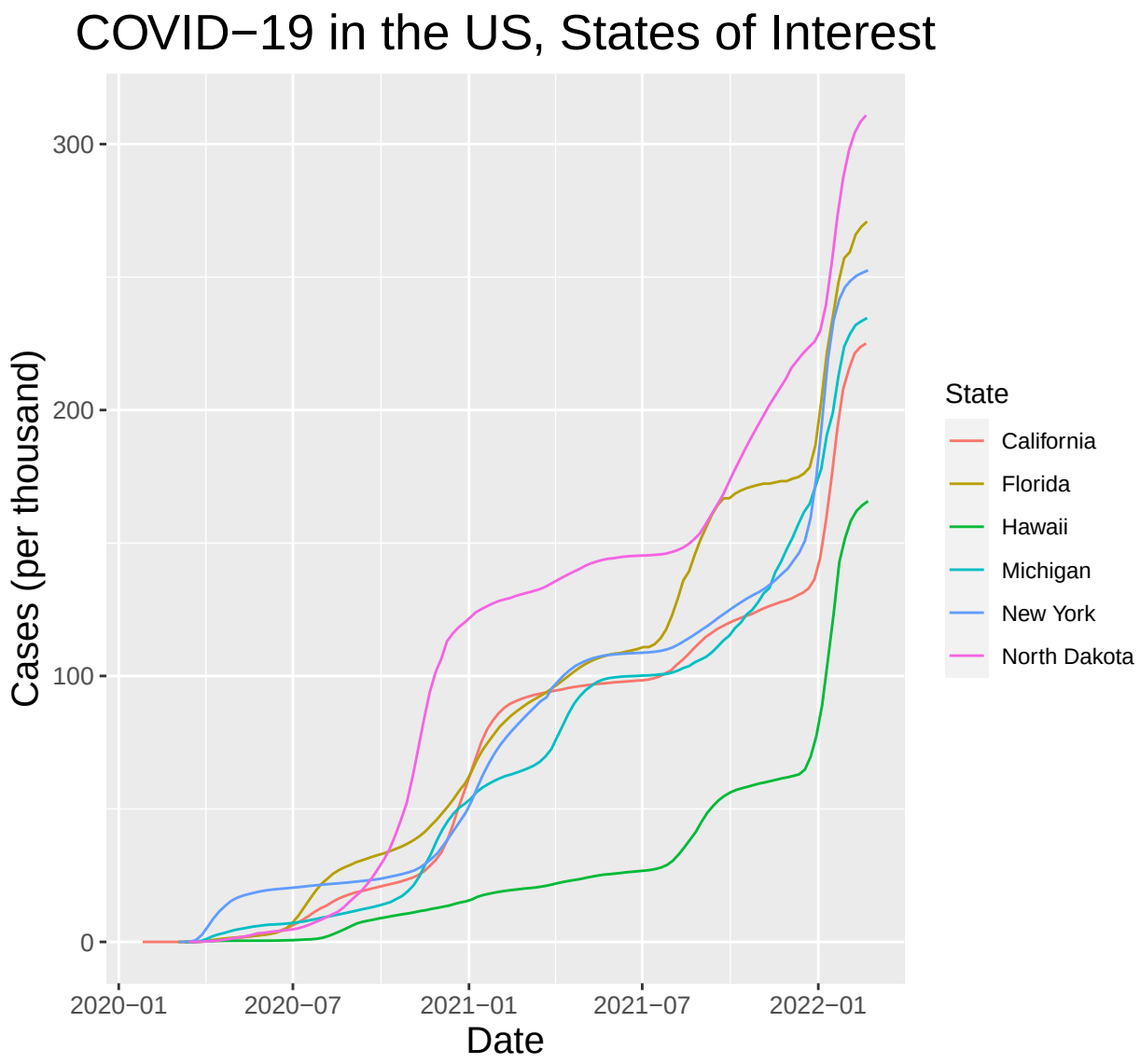


You can see from this graph that there has been periods over the course of the pandemic in which the number of cases and deaths per day has declined. It looks like July of 2021 saw the lowest number of new deaths and cases since the start of the pandemic, but that low period unfortunately didn't last for long as the numbers rose up again reaching close to record highs around September / October. On thing that is interesting to note is that the variability day to day has increased greatly in the recent months, where it was minimal closer to the beginning of the pandemic. It would be interesting to do further analysis on why this might be, and see if it's potentially related to vaccine administration. One last thing to note about this graph is that it is interesting to see that the trends in deaths closely follow the trends in cases.

The last two plots have shown the data for the United Sates as a whole. Next well break down the data by state to see whether there are similarities in how the pandemic played out between the states and there are any notable outlines. For this analysis we will be focusing on cases only, as this seems to be relatively indicative of trends in deaths.

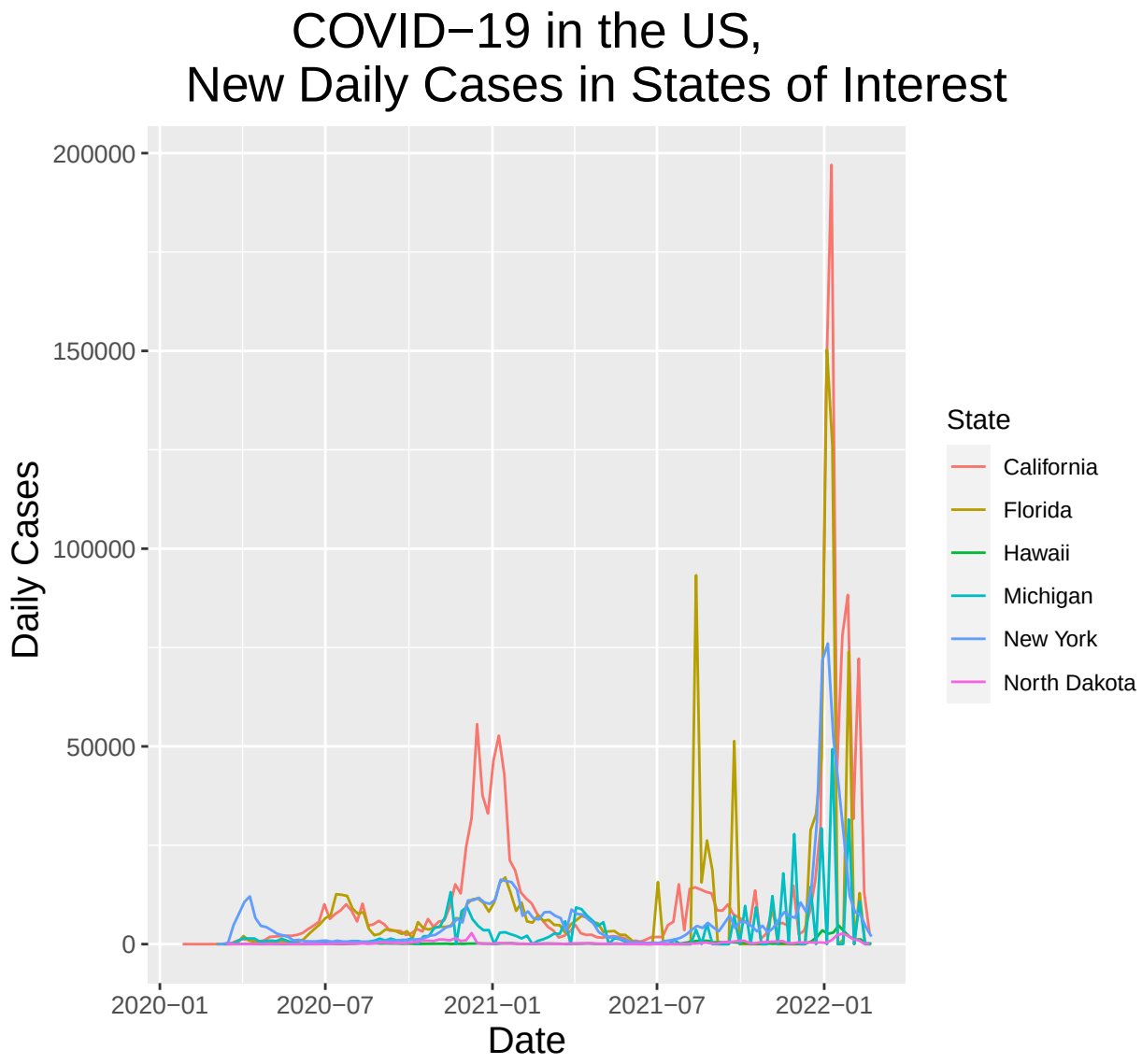


This plot shows the number of number of cases per 1000 in each of the states over the course of the pandemic,. This plot gives a good idea of how well each state did in preventing the spread of Covid over time, but there is too much going on to be able to discern in depth meaining from it. For this section of the analysis we are going to focus on only a few states. The states focused on are North Dakota which had the highest number or cases per million, Hawaii which had the least, Michigan who fell in the middle, and Florida, New York, and California as they got so much press coverage over the pandemic. The plot below is showing the trend of cases per million for each of these states.

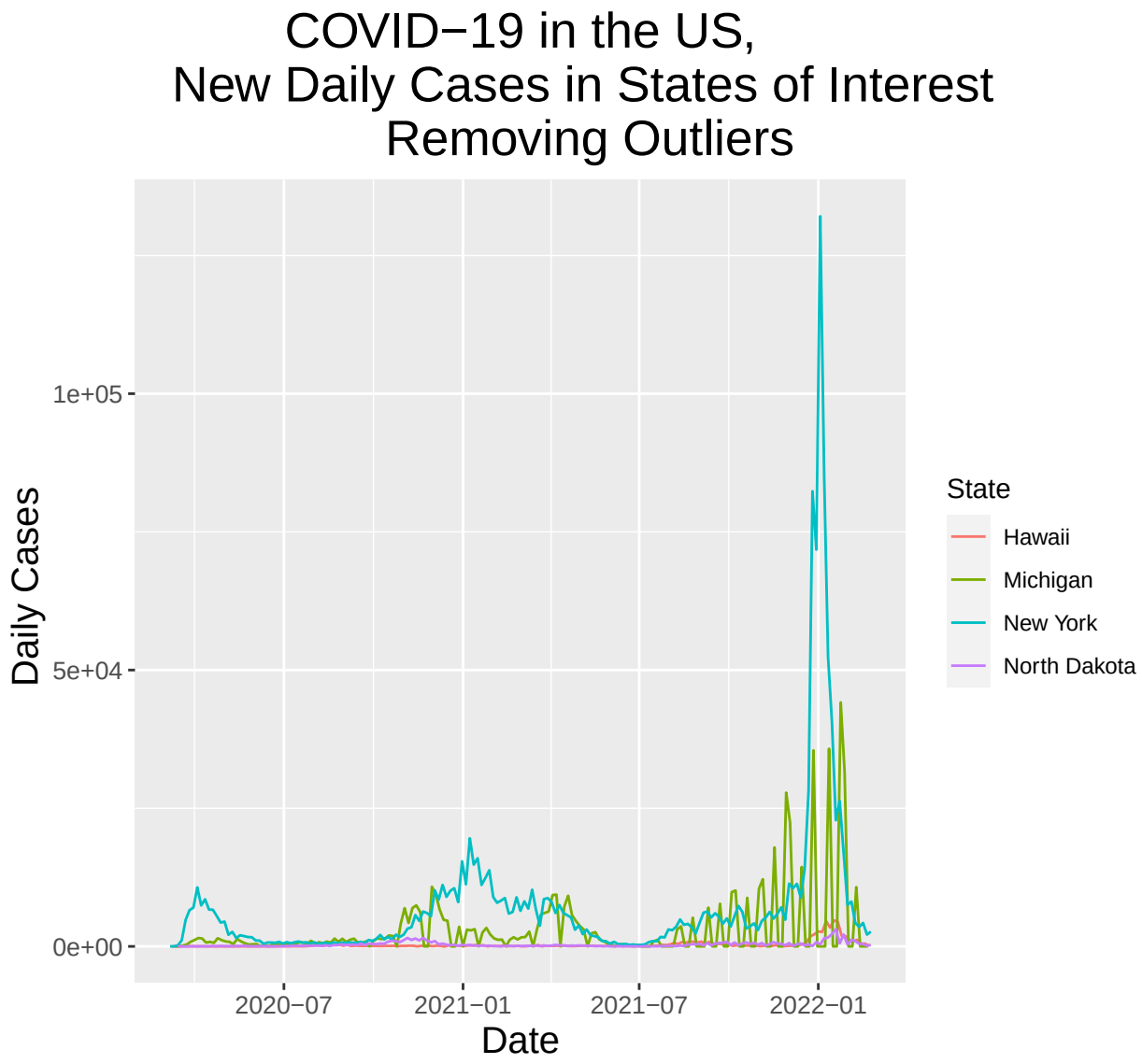


You can clearly see the North Dakota had the highest number of cases per million from pretty early on in the pandemic. Hawaii also from the start had less than the other five states, and didn't have any major increases until around September of this year. That may be because it is more isolated from the rest of the country. The sharp increase that happened recently might be related to more travel that has been permitted recently. Its also interesting to see that New York started out with a higher rate of cases and then leveled out to the middle of the pack. This could be because its such a major travel hub and saw large numbers of cases early on. Michigan clearly has stuck around the middle for almost the entire time. California, New York and Florida are all around the same rate, between Michigan and North Dakota, for the most part, but you can see that Florida had a big jump around September as well. It would be interesting in a deeper analysis to look at data on the implementation of lock downs, mask mandates and travel restrictions in each state to determine if those correlated with the number of cases and deaths.

Next we will look at the difference in new cases for these same states over the course of the pandemic.

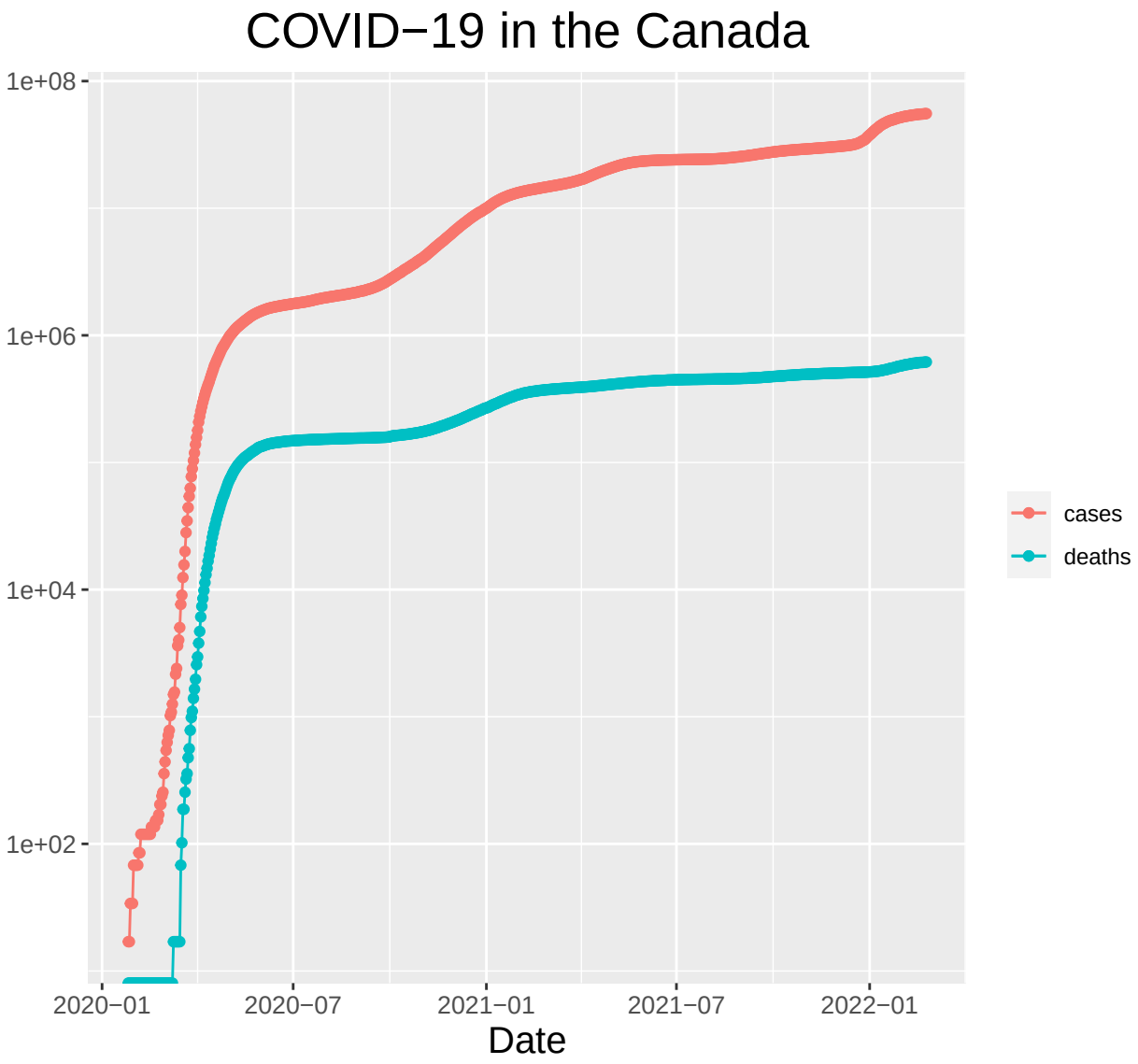


This plot is interesting as it shows some instances of huge amounts of cases being reported in California, with the highest around 45000 per day, and Florida, wit around 70000. These outlines make it difficult to see smaller, monthly trends, and data from states that did not have these huge numbers of new cases. This might explain however why these areas were receiving so much media coverage, and it might be interesting to look into the situations surrounding them. The next plot zooms in on the states without these extreme outliers.

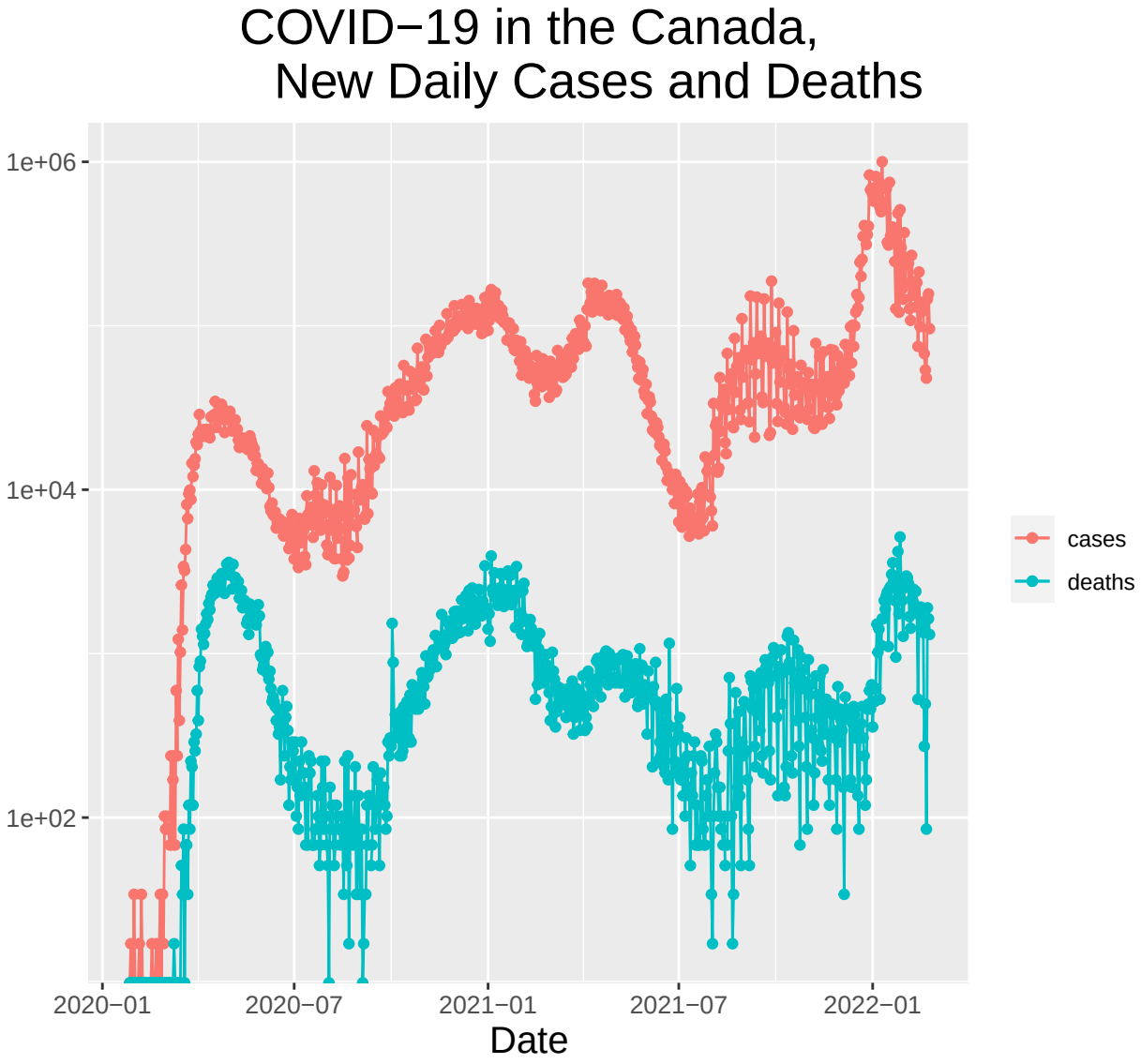


We can gather a few surprising things from this plot. You can see that although the number of cases per thousand is highest for North Dakota, there are states which are seeing much greater number of new cases each day. Hawaii however being the lowest per capital cases also seems to have the lowest number of new cases for most of the pandemic. It seems that the number of new cases is very closely tied to the population of each stats. One thing that is interesting to not is that the number of cases more recent months have been a lot more variable, with large differences in the numbers day to day. I would be interesting to see if there were any ties between this and vaccine administration.

We will look at one last group of visualization, which will be related to what we focus the model that created on. This part of the analysis will look at the country data, and more specifically at Canada, which is the country I live in. The fist plot that will be created will be similar to the fist one we looked at for the United States. It will show the number of cases and number of deaths that have occurred in Canada over the course of the pandemic on a log scale.



You can see that similarly in Canada the deaths are quite closely related to the total numbers of cases, though it looks like in the last year the number of cases has risen by more than the deaths. You can also see that over all there has been less cases in Canada then there was in the United States. We can also normalize this graph like we did for the US case to show the new cases and new deaths each day, in order to better see trends. That plot is shown next.



Comparing this plot to the US version, you can see that Canada has seen much more variance in the number of cases and deaths per day over the course of the pandemic. One thing to note is the population of Canada is quite a bit lower than that of the states, but it would be interesting to see if there are other differences between the way the two countries dealt with the pandemic and how they relate to the data. The Canadian data seems to have two equally low dips in their graph, where America only had the one. Also it seems that there is more variability in the numbers earlier on, around August and September of 2020, in addition to the variance around the same time the following year that was seen on the US graph as well. Now that we have a good idea of what the data looks like visually, we can begin to create models to determine quantitative relationships between different variables.

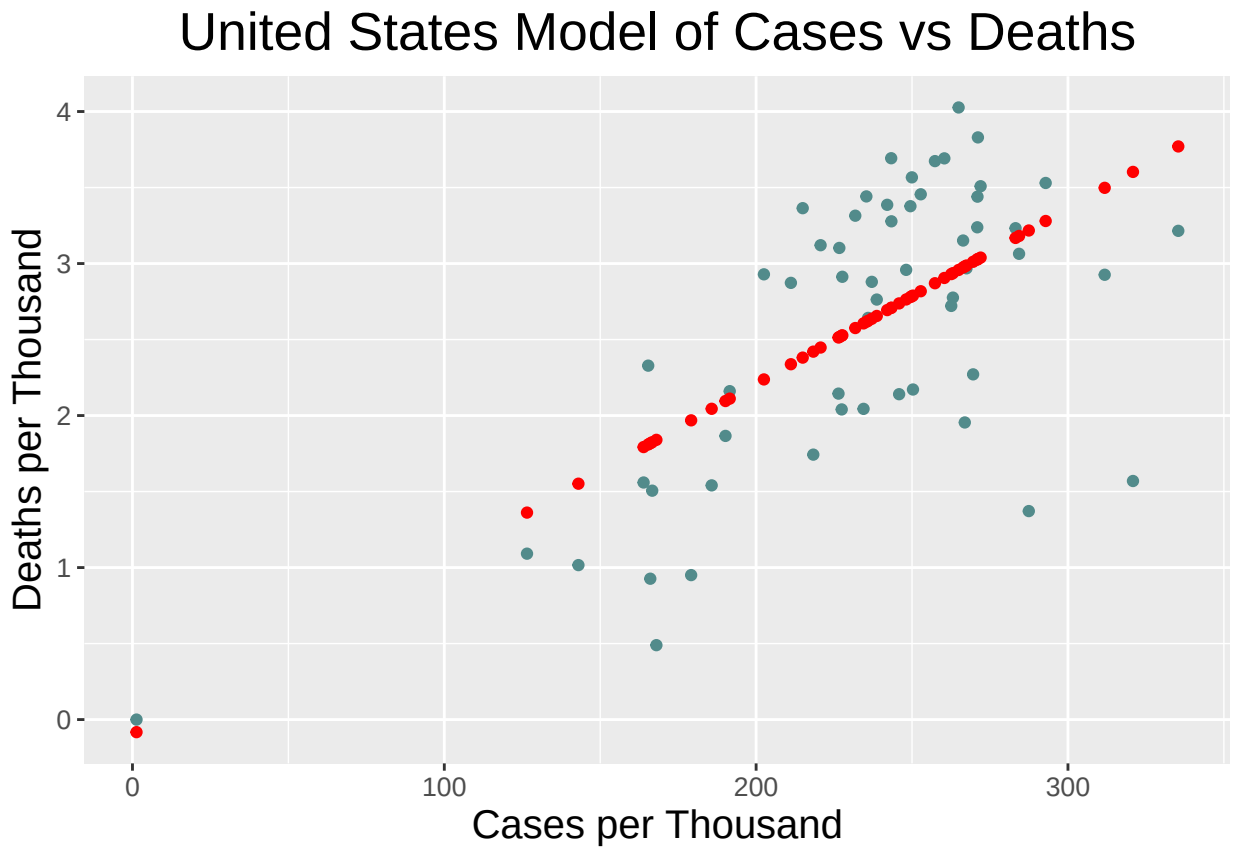
Modeling the Data

The models that are created look at the relationship between the number of cases and number of deaths. For this, the comparison of Canada and the US will be continued, as models of cases versus deaths will be created for both countries. These models will look specifically at the cases per thousand and deaths per thousand for each of the states, for the US model, and the provinces for the Canadian model.

The first model, for the US, is created and a summary is shown below. It is then plotted against the original

data points for each state.

```
##
## Call:
## lm(formula = deaths_per_thou ~ cases_per_thou, data = US_state_totals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.03328 -0.51182  0.07338  0.58886  1.06886
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.097625   0.425018  -0.230    0.819
## cases_per_thou  0.011533   0.001779   6.483 2.84e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7102 on 54 degrees of freedom
## Multiple R-squared:  0.4377, Adjusted R-squared:  0.4273
## F-statistic: 42.03 on 1 and 54 DF,  p-value: 2.841e-08
```

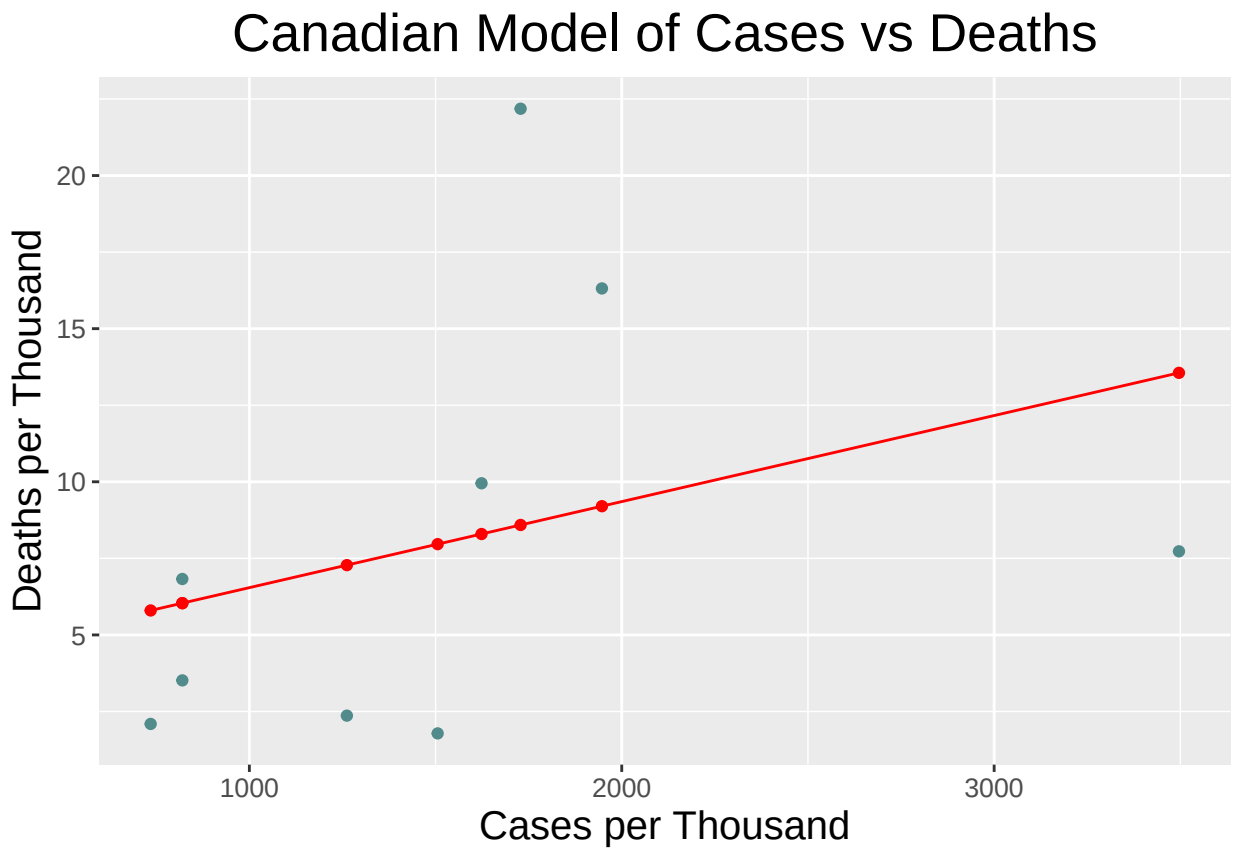


This model, seen in red, seems to do a pretty good job of fitting the data, although at the higher numbers of cases and deaths the data points are quite spread out, and therefor harder to match. At the lower number of cases, the model predicts that there will be more deaths than there acually is, and shows a higher mortality rate. It provides us with a pretty good prediction based on data that confirms what was observed from earlier visualizations, that the number of deaths is positiovly correlated to the number of cases.

To create the Canadian model an additional step is nessesary as the data we are working with does not have populations for the provinces. This is solved easily as that data was found from the site below, and transformed so that it could be merged with our data. Once that data is added, the model is created and is displayed below.

Provience population data: "https://raw.githubusercontent.com/tommy321/Canadian_Population_Density/master/2016_census_data/T1901EN.CSV"

```
##
## Call:
## lm(formula = deaths_per_thou ~ cases_per_thou, data = proviences_with_pop)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.179 -4.914 -2.522  1.657 13.592
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.733273   5.158865   0.724    0.493
## cases_per_thou 0.002810   0.002958   0.950    0.374
##
## Residual standard error: 7.122 on 7 degrees of freedom
## Multiple R-squared:  0.1142, Adjusted R-squared:  -0.01233
## F-statistic: 0.9025 on 1 and 7 DF,  p-value: 0.3737
```



The first obvious thing we see from this graph is that there are less data points, and therefor the model is not going to fpredict relationships as well. This is a tough situation as there are simply fewer provinces in

Canada then there are in the states. On solution for this could be to base this on cities cases and deaths per population, as that would provide many more data points. None the less, this plot shows that there is the same positive correlation in Canada between the number of cases and deaths due to COVID 19. This model, despite its weaknesses seems to do a better job at making predictions around the lower number of cases. This could be because most of the data from Canadian provinces seem to exhibit lower numbers of cases per thousand.

Conclusion

This analysis allowed us to pull alot of interesting and useful information out of the data. We were also able to practice skills in data analysis including finding good data, cleaning the data, transforming and visualizing the data, and then transforming it some more, and modeling data.

From this analysis we saw that the general tends in the number of deaths related to COVId-19 closely follow the trends in the number of cases, in both the United States and Canada, and very likely in other countries as well. This correlation was confirmed in our models of data, and the predictions based off these models, which used data from both the US and Canada. We were also able to compare the number of cases over the course of the pandemic between the US and Canada, as well as the changes in new daily cases and deaths. From this we were able to infer that maybe the differences in the numbers are a result of the different ways they handled the pandemic, in regards to restrictions and mandates. Further analysis would be required to say for sure.

We also looked at the US data by state, and found that there was vast differences in number of total cases, cases per one thousand, and daily new cases. Some states that we looked at even had extreme outliers, for example Florida with new daily cases of around 70000 at some points. It was interesting to see as well how the cases per thousand changed for states over time, and to make hypotheses on why that might have been.

In this report we were able to pull some interesting info out of the data, but there is the potential for alot more to be discovered from it, in a few areas identified throughout.

Recognition of Bias

There were obviously many potential situations for bias in this investigation of the data. Looking at just the data, it is being recorded by people all across the world, in each country for the global data. We hope there isn't, but its very possible that there could be huge differences in the way that data is collected and recorded. For example, in poorer countries people may not have access to healthcare or doctors, so when they get sick they may just stay home, ride it ou, and never report cases to the organization collecting data. Even within the united states there may be regions which reporting and testing is done differently meaning that the numbers are not consistent between states. Also there was additional population data that was introduced at two points in the analysis, which further opens up the report to biases.

This is just one example of possible bias. There is bias in the way that this report was put together. Not on purpose and hopefully not with negative consequences, but the topics I choose to focus on and the decisions I made in visualizing and modeling data were guided by my personal interest and views. This is something that is important to consider and be aware of in other peoples work, as well as your own. It's important to take steps during analysis to recognize your own bias and make an effort to counter in your own work. An example from this report is the fact that I am from Canada, and so when dealing with Canadian data I made an effort not to let my feeling about my country come through in my analysis. I put a lot of thought into this issue, and hope that I was successful in not conveying any personal biases, at least without pointing them out.