Spencer Marlen-Starr STAT 515 Fall Term 2, 2020
Final Project Applied Statistics and Visualization for Analytics GMU

An Exploratory Data Analysis of Deaths Caused by COVID-19 in the United States

The reason I chose to explore data on the novel Coronavirus pandemic for my final project in this class is because I also chose this topic for the final project in AIT 580 but came away from it feeling unsatisfied with how much I was able to learn from doing that project because of the single and limited dataset I chose to use for it. The dataset I used for all the analysis in that project had no breakdowns by state within the United States and no breakdowns by age brackets, both of which I wanted to investigate. So, when I found COVID-19 datasets on the CDC's website broken down by state, age, and other factors, I thought that I could use them to do the project I had intended to do last time.

The reason I went with a COVID-19 dataset for my final project in AIT 580 is because I wanted to kill two birds with one stone so to speak by satisfying my curiosity to learn more about the ongoing pandemic by looking at the data and practicing the data analysis skills I had learned during that class at the same time. So, because I acquired new abilities to create many different types of data visualizations in R in this course as well as learning several methods of statistical learning, I had the same motivation in mind this time around as well.

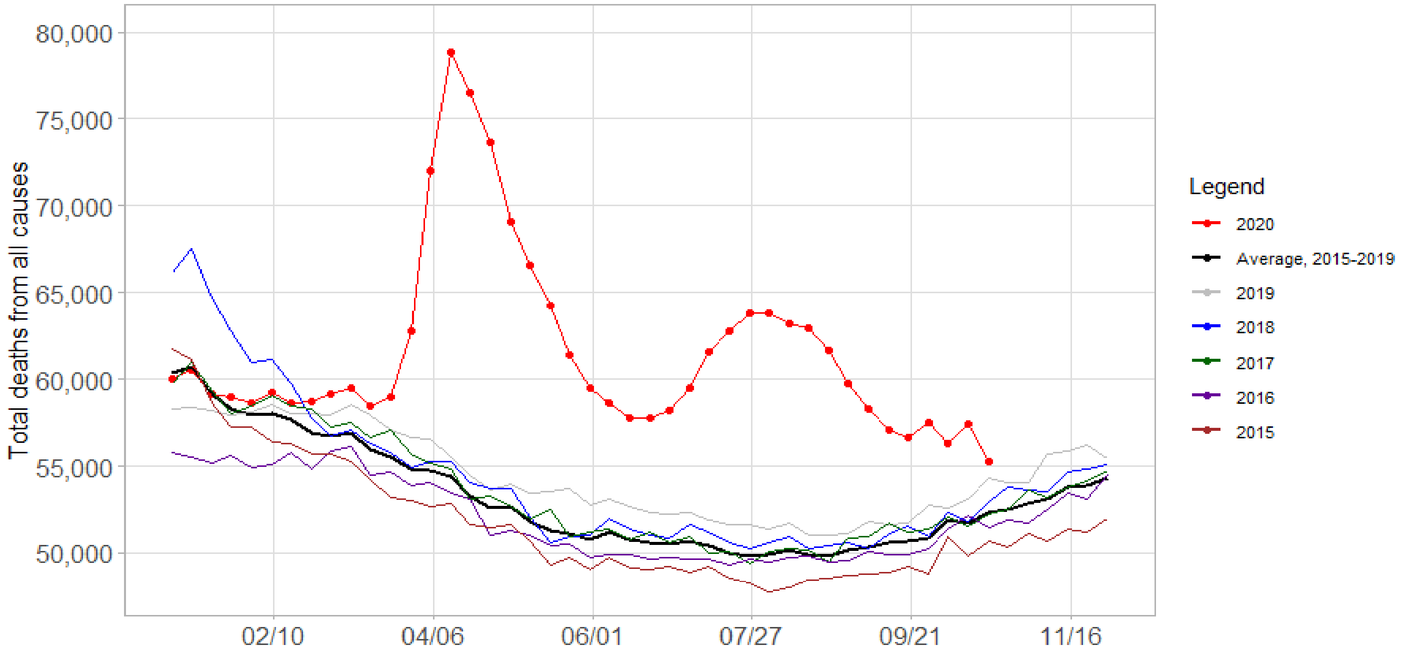
The datasets I chose to work with:
Provisional COVID-19 Death Counts by Sex, Age, and State from the CDC & NCHS. It contains data on cumulative deaths from COVID-19, total deaths from all causes, Pneumonia deaths, and Influenza deaths in the United States starting at 2/1/20, the beginning of the pandemic's arrival in this country and going until 11/18/20. The deaths are broken down by sex (male/female), age groups, and state.
Weekly Counts of Deaths by State and Select Causes, 2019-2020 from the CDC & NCHS. It contains data on weekly deaths in the US from all causes including COVID-19 from 1/5/19 until 11/7/20 which is also broken down by state.
Excess mortality raw death counts from Our World in Data's article on Excess Mortality. The third dataset contains the raw weekly counts of excess mortality for a multitude of different countries around the world starting on 1/5/20 and going until 10/18/20.
Excess mortality P-scores by age from Our World in Data's article on Excess Mortality. The fourth dataset has data on the P-scores of excess mortality in the US broken down by age buckets also starting on 1/5/20 and going until 10/18/20.
The 5th and final dataset I used is what Our World in Data calls their "complete COVID-19 dataset" which is updated daily. It has columns for total cases, total tests, total deaths, total cases per million, total tests per million, total deaths per million, new cases, new tests, new deaths, new cases per million, new tests per million, new deaths per million, smoothed versions of all the aforementioned items, and a large host of potential explanatory or predictive factors for most countries worldwide.

Another feature which the data set I chose for the previous course's final project lacked was any information on excess deaths aka excess mortality which was a major shortcoming of it, so that is the first thing I decided to make data visualizations on when I embarked on this project. Excess mortality is a figure which refers to the number of deaths from all causes during some event, an epidemic for example, which are above what would have been expected if that event had not occurred. In the datasets I chose to use, the "what would have been expected if that event had not occurred" part was represented by using the mean of the weekly deaths from all causes over the previous 5 years. So, for example, in these datasets, the raw number of excess deaths for week 31 of 2020 would be $Excess\ Deaths_{week31\ 2020} = Total\ Deaths_{week31\ 2020} - \overline{Total\ Deaths}_{week31\ 2015-2019}$ where the vertical line above total deaths indicates it is the mean of the deaths for that week over those years.

Excess mortality is believed by many public health researchers to be a superior of the total deaths caused by the COVID-19 pandemic than just using the total confirmed COVID-19 death count, the measurable but flawed case fatality rate, or the better but unmeasurable infection fatality rate. One reason is that excess mortality picks up deaths that were caused by COVID-19 but were misdiagnosed or misreported. My graph of the raw number of excess deaths in the United States since the start of the pandemic over time is below.

Excess Mortality in the United States during the COVID-19 pandemic:
Raw number of deaths from all causes compared to recent years

Shown is how the raw number of weekly deaths in 2020 differs from the number of deaths in the previous five years (2015-2019).

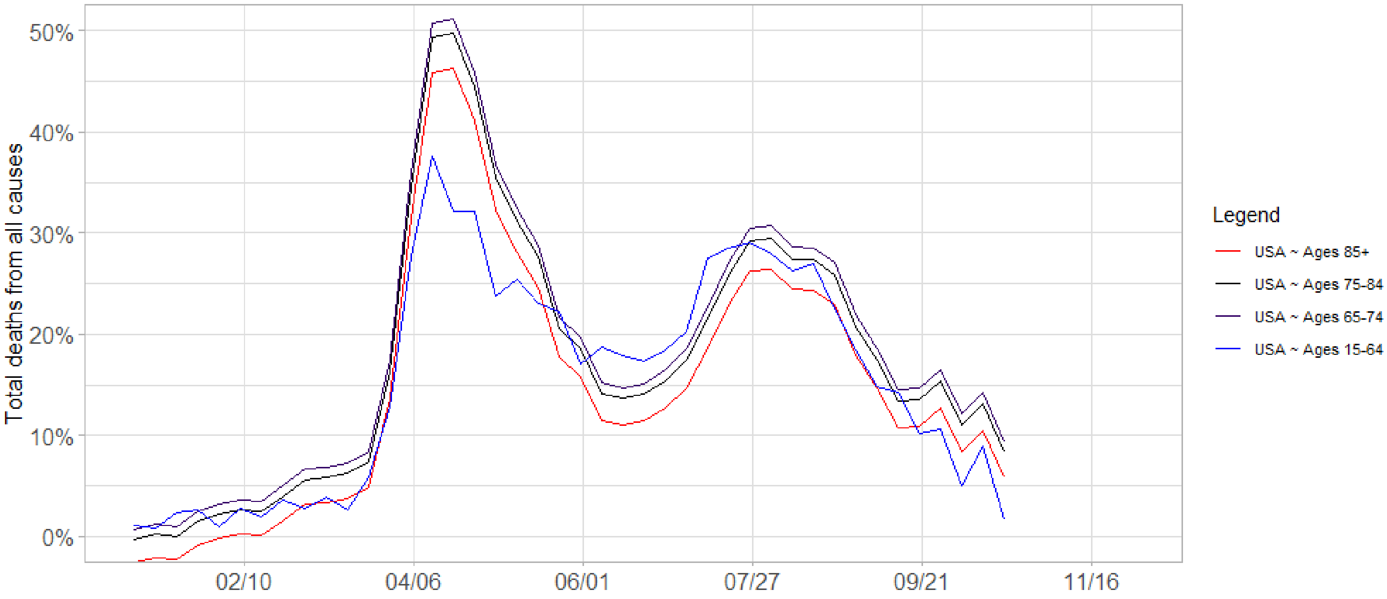


Sources: Human Mortality Database (2020), OurWorldinData.org

Excess mortality does not have to be measured in terms of the raw numbers though, it can also be measured as a percentage difference between the number of deaths during a given week this year versus the mean deaths in the previous five years over that same week. This alternative way to express excess mortality is called the P-score and it is a better figure to make international comparisons with. Below is a graph of the P-score of excess mortality in the United States over the same time frame, but this time, it is also broken down by 4 different age groups.

Excess Mortality in the United States during the COVID-19 pandemic:
Deaths from all causes compared to recent years, by age

Shown is how the weekly number of deaths in 2020, broken down by age brackets, differs as a percentage from the average number of deaths in the same week over the most recent five years (2015-2019); this metric is known as the P-score.

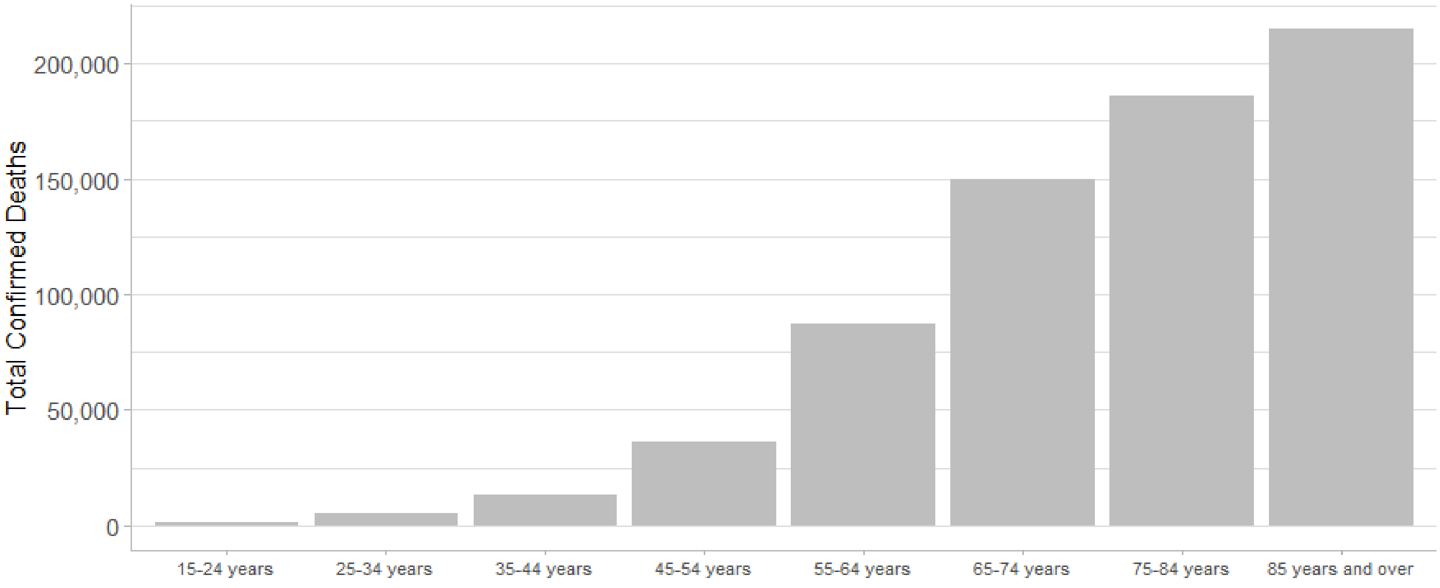


Sources: Human Mortality Database (2020), OurWorldinData.org

One drawback of the datasets I used to create both excess mortality graphs is the lack of mortality data in recent weeks so I could have a chance to compare the second wave to the initial wave. This is because as Our World in Data's article on excess mortality puts it, "Mortality data is incomplete in the weeks, and even months, after a death occurs because of delays in reporting."

The reason I took the age factor into account in the second excess mortality graph is because of the large amount I love information I have heard about the varying impacts of COVID-19 depending on the age of the person who catches it. That graph alone was not enough to satisfy my curiosity about the role that age plays in mortality from COVID-19, so I decided to follow it up by creating a bar chart of total confirmed covid deaths cumulatively in the United States broken down by 9-year age brackets, that graph is included below.

COVID-19 Deaths in the United States by Age

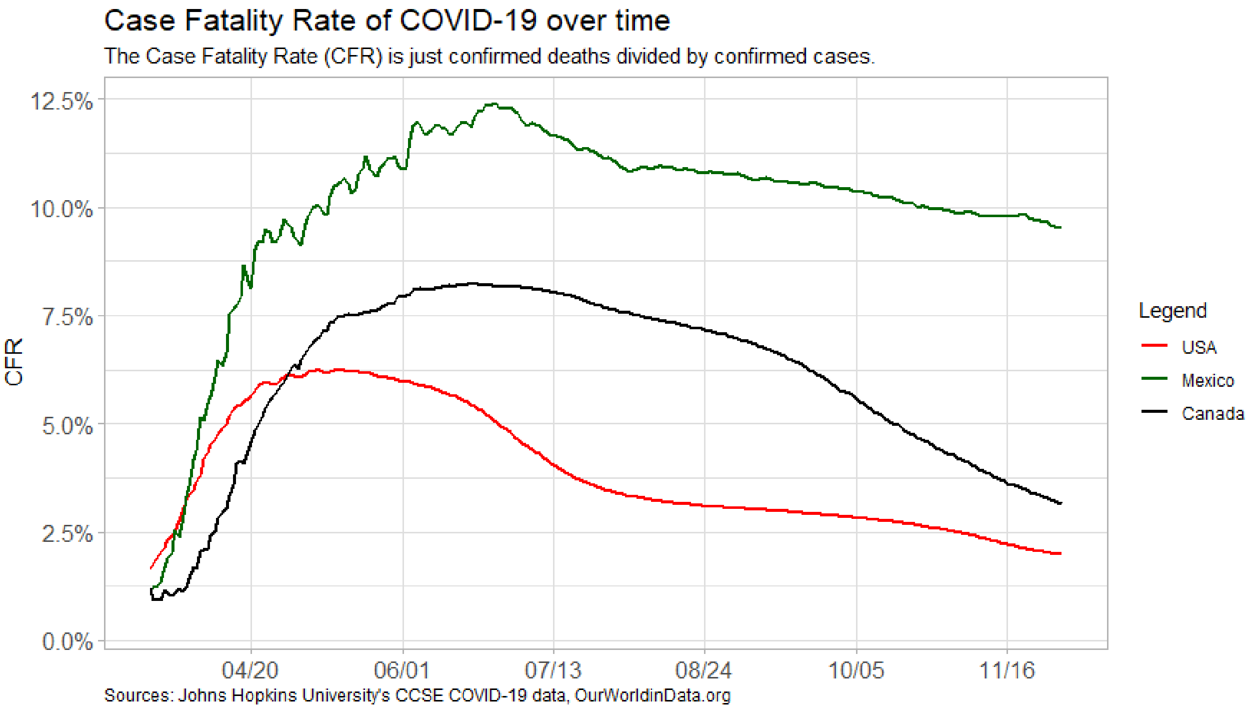


Source: Center for Disease Control (CDC)

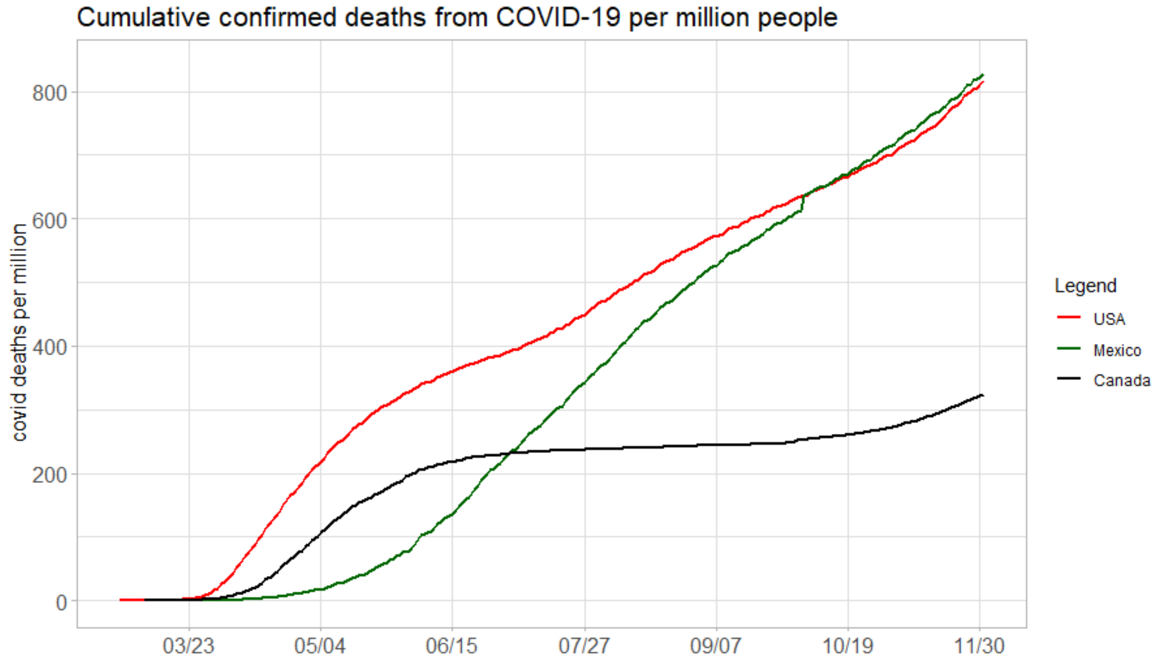
Compared to the P-scores of excess mortality graph from the previous page which did not fit the story of age playing a huge factor very cleanly, this column chart very clearly indicates how big of a role age has played in mortality during this pandemic, at least in the US. Deaths from COVID-19 appears to be a monotonic function by age group in this graph.

International comparisons

In order to see how the United States has faired so far in comparison to other countries in terms of deaths from this pandemic, I decided to create some line charts of various other crude mortality metrics comparing the US to its two closest and largest neighboring countries, Mexico and Canada. The first of these was a comparison of what is known as the case fatality rate of COVID-19 over time in these three countries. The case fatality rate (CFR) of a disease is simply total deaths divided by total cases and expressed as a percentage, so $CFR_{covid} = \frac{total\ deaths_{covid}}{total\ cases_{covid}} \times 100.$



As you can see, somewhat surprisingly, according to this crude measure of mortality, the United States has fared better than its two closest and largest neighbors throughout the pandemic thus far. But this is only one imperfect measure of mortality, so I also decided to look at how the United States stacks up against Mexico and Canada in terms of total deaths from COVID-19 per million persons of population over time. That graph is included on the next page.

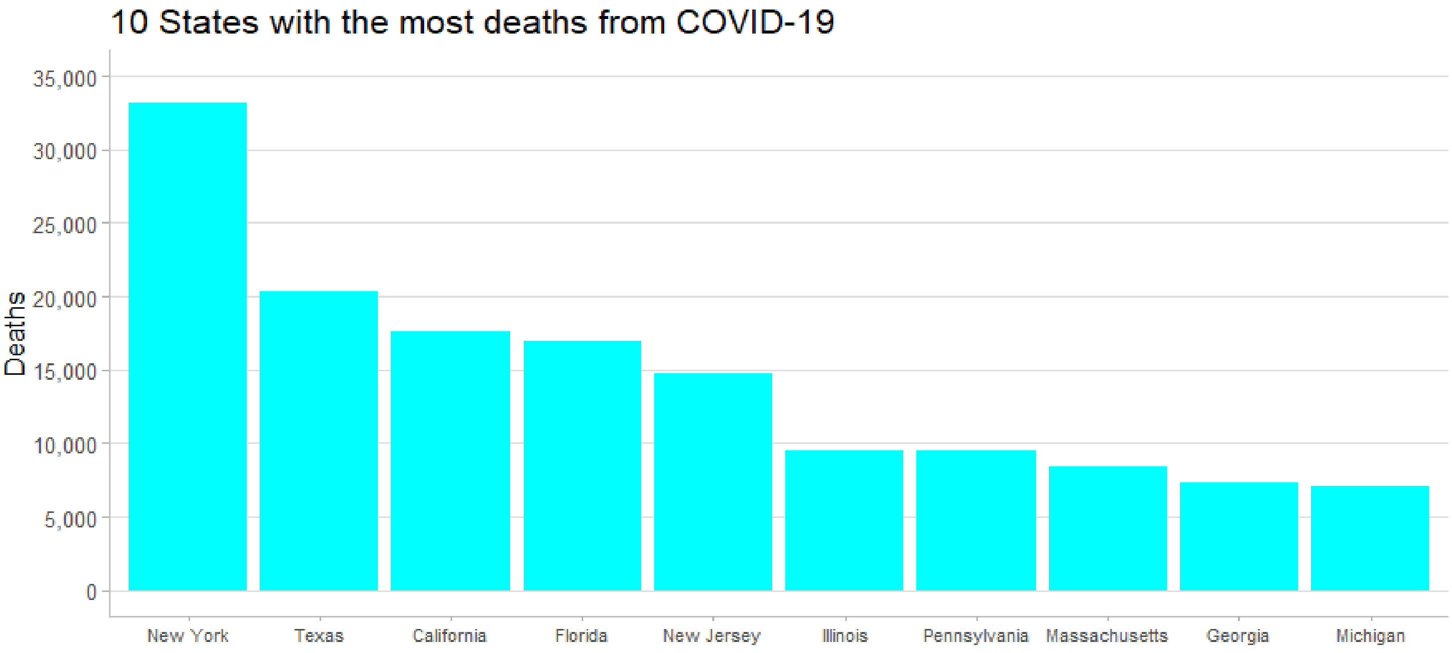


Sources: Johns Hopkins University's CCSE COVID-19 data, OurWorldinData.org

From inspection of the graph above, it is clear to see that the results have changed from the previous graph, but not as dramatically as I had anticipated. It is also clear that when the pandemic really started to strike the US and Canada was far earlier than it really struck Mexico which is interesting. Even when looking at this alternative metric of mortality from the CFR however, it is still the case that the US has not fared worst in international comparisons when attempting to keep geographical factors constant by just looking at its neighbors.

Interstate comparisons

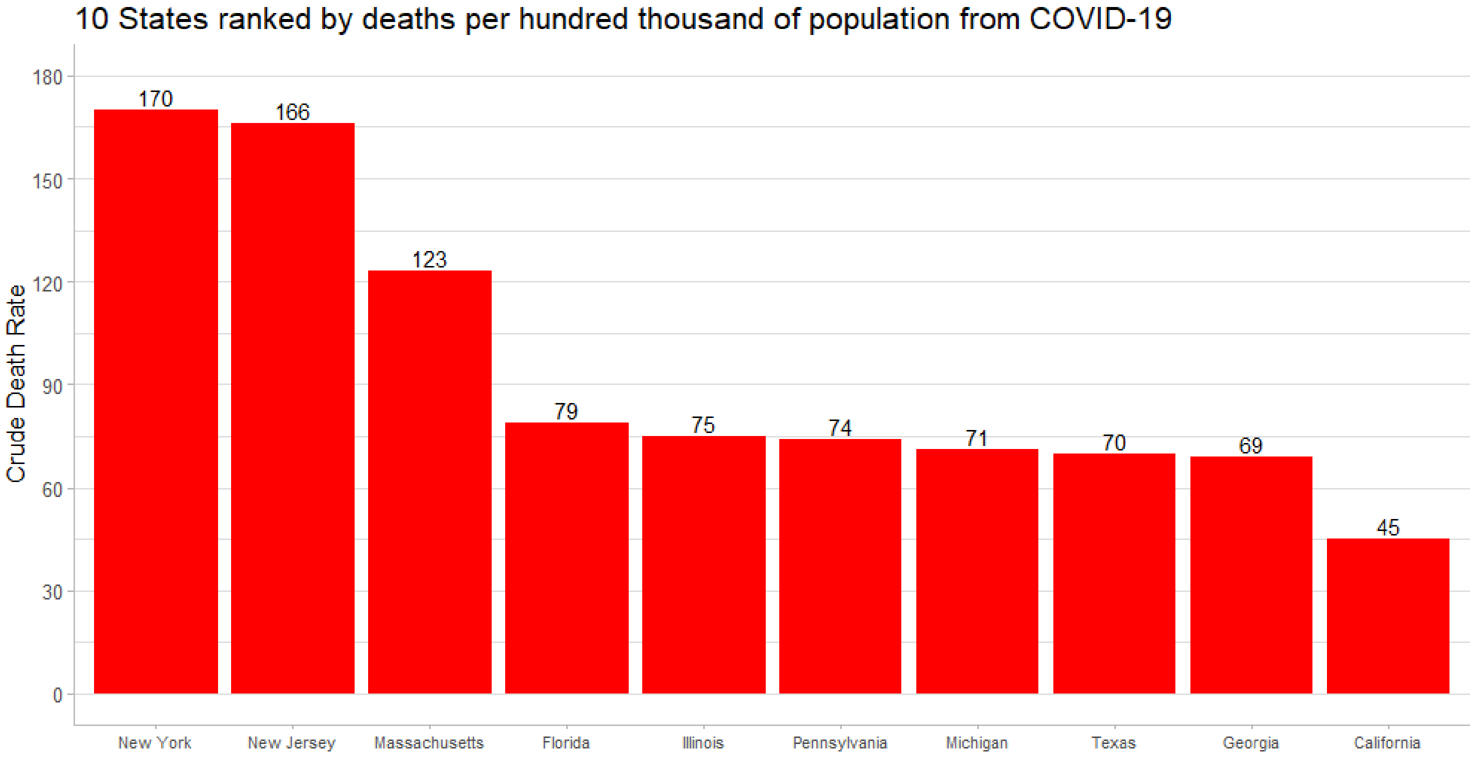
After comparing the outcomes in the United States overall to presumably similar countries due to geographical and size considerations, I turned my attention to interstate comparisons of COVID-19 outcomes within the United States. To me, the most logical place to start my exploratory analysis of Covid outcomes within the country was by creating a data visualization of the top 10 states in terms of most total deaths from Covid, so that is what I did, and the result is included below.



Source: Center for Disease Control (CDC)

This top 10 bar chart is sad for me because I currently live in California and while I knew because of its large population that it would likely make the top 10, I was hoping it would be somewhere after fifth place, third most deaths is very high indeed. Once you take population into account, it is very unfortunate for the people living there that the state of New York has had the most deaths by such a large margin when you consider that California has over twice the total population of New York. Before I created this graph, I had sort of expected Texas or Florida to have the most deaths from COVID-19 because of their more lenient policies towards combating the pandemic, but that prediction was falsified by the data. Perhaps the factors effecting the path and impacts of this virus are more complicated and less under our control than I had thought previously.

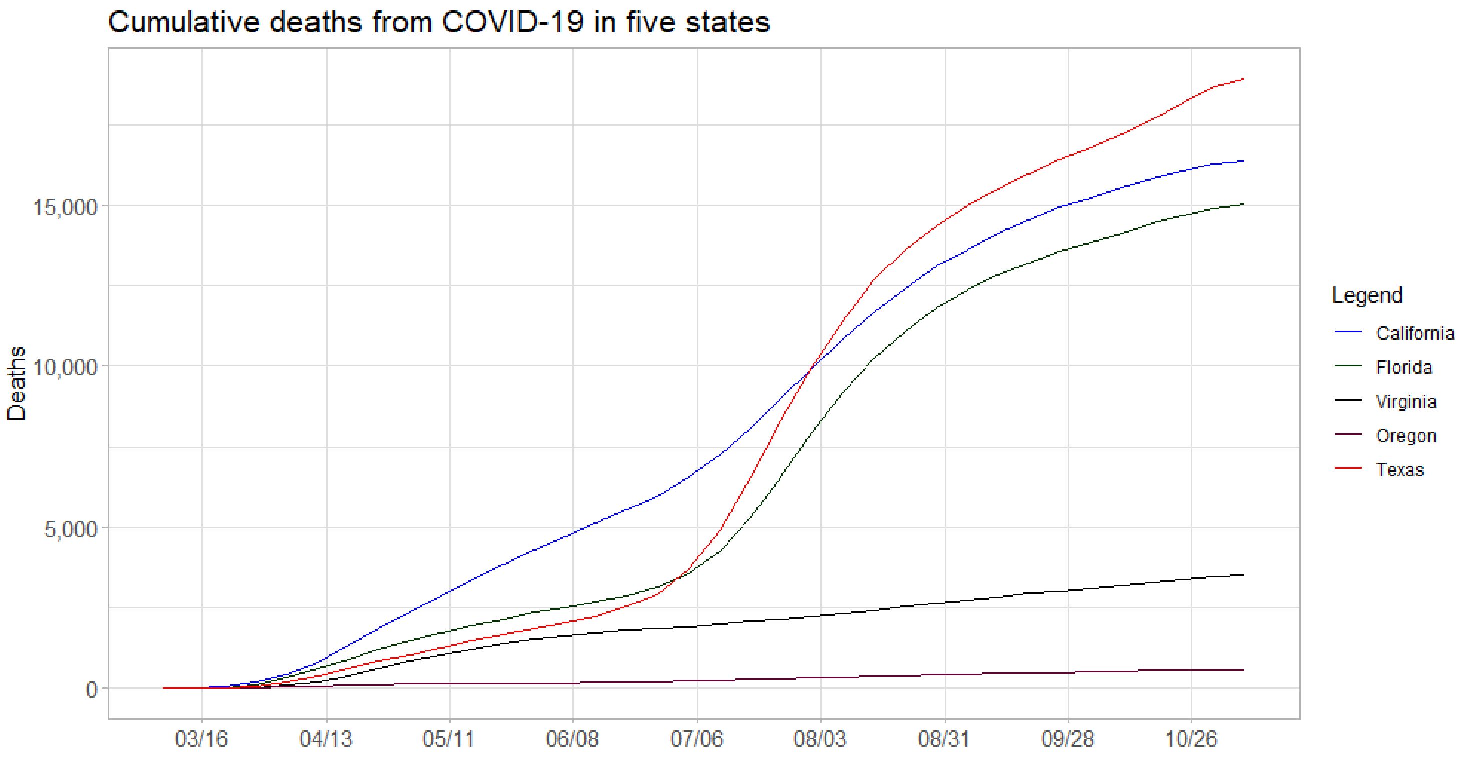
There is a big problem with the top 10 bar chart I created however, and that problem is that it makes no attempt to control for the population of each state. In order to compare those same 10 states by what is commonly referred to as a crude death rate [1], I created a second chart with those same 10 states but this time, I divided their total deaths from COVID-19 by their population and multiplied by 100,000 to get total deaths per hundred thousand population for each of those states. That chart is included on the next page.



Source: Center for Disease Control (CDC)

The ordering of which states have fared the worst when taking population into account changes quite a bit which is why it was necessary for me to create this graph to add more context to help interpret the information in the previous graph. From this graph which takes population into account, it appears California and Texas are not doing as bad while New York, New Jersey and Massachusetts are all doing worse than the previous graph portrayed. But both of the previous two column charts are static comparisons at a snapshot in time in an ongoing pandemic, so taking a look at how several specific states have fared over time might help us to gain a fuller picture of how the impacts of the SARS-COV2 virus on them has changed over time.

In order to further explore comparisons between states within the United States but in a dynamic manner this time, I then decided to plot the cumulative deaths over time in 5 different states. For the 5 states, I chose 3 from the top 10 chart (California, Florida, and Texas), Virginia just because that is where George Mason University is located, and Oregon because it is right next to California which as I already stated previously, is the state I live in. The following graph which does this will be the last graph included in this final project report.



Source: Center for Disease Control (CDC)

The line chart above took the most data preprocessing work to complete by far because the weekly death counts dataset from the CDC's website did not have a cumulative deaths column, so I had to create one by using a running total formula for each of the 50 states individually which took me well over an hour of typing, dragging, and scrolling to complete for all 50 states in that dataset in the csv file itself. And because of how long that took me, I only did so for the column displaying weekly counts of deaths with Covid as the only underlying cause of death and not for the column containing weekly counts of deaths with Covid as one of multiple underlying causes.

It is striking how similar the dynamic trends in cumulative deaths caused by the coronavirus in the states of Texas and Florida. If the true cause of the huge increase in deaths in Texas around early July was that state lifting its initial stay-at-home orders on April 30th [2], you would have expected the increase to have started 3 or 4 weeks after that which means in late May, not in late June or early July. Furthermore, if that had been the cause, because the governor of Texas reimposed lockdowns on June 26th as stated in that same article, you would have expected the rate of increase in deaths in Texas to have started to level off in July rather than in August.

As for Florida, their initial stay-at-home orders began phase 1 of being lifted on May 18th [3] which could be interpreted as lining up with the increase in the first derivative of the green Florida line in the graph at what appears to be about a month later. California has a very different trend line from the aforementioned states which could be for any number of different reasons, but this is mainly an exploratory analysis searching for interesting things to consider, it is not intended as a place to test previously developed hypotheses because I had none coming in.

Concluding remarks

I feel I learned a lot from doing this project, so in some ways, I am glad I chose the topic I did. However, I really should have taken the advice in the instructions of limiting the scope of the project more seriously than I did. I thought that choosing to only look at deaths from COVID-19 in the United States would limit my scope sufficiently to be able to learn a lot about that topic, but I should have known better. Had I known that my initial two datasets I thought would be sufficient to complete this project would balloon up to the five datasets I ended up needing, I would have chosen a different subject for my final project in this course.

I do not have any background in virology, epidemiology, public health, or anything else that could be directly relevant, so I hesitate to make any strong inferences from the many visualizations I have created while working on this project. One thing I will say though is that there does not appear to be much of a consistent pattern at all in terms of different behaviors and policies leading to different outcomes. There was no single robust finding.

References and Citations

[1] "CRUDE DEATH RATE." *New Jersey Department of Health*, March 16, 2009. https://www-doh.state.nj.us/doh-shad/view/sharedstatic/CrudeDeathRate.pdf

[2] Linnane, Ciara. "Texas becomes the first state to reimpose restrictions after lifting stay-at-home order on April 30th." *Market Watch*, June 26, 2020. https://www.marketwatch.com/story/texas-becomes-the-first-state-to-reimpose-restrictions-after-lifting-stay-at-home-order-on-april-30-2020-06-26

[3] Flores, Rosa and Amir Vera. "Floridians rejoice as malls, barbershops and gyms reopen across the state." *CNN*, May 19, 2020. https://www.cnn.com/2020/05/18/us/florida-reopen-roundup/index.html

Appendix

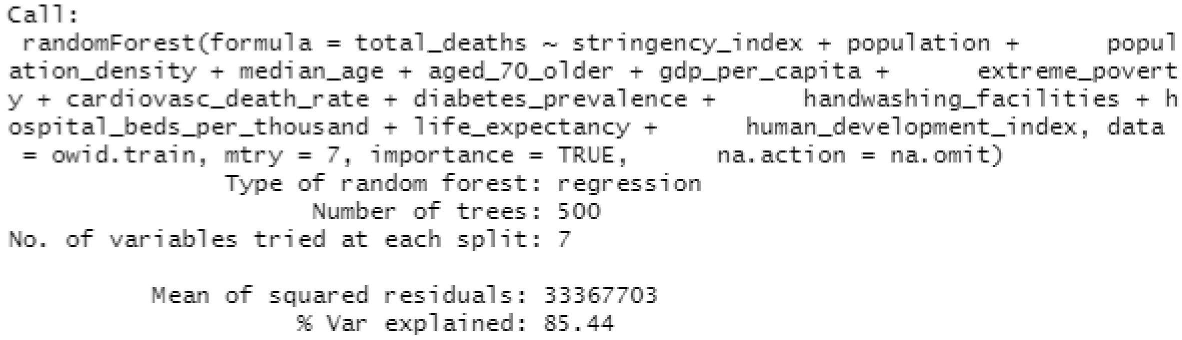
Running a statistical learning model: random forest regression

I decided to run a random forest regression model on the overall international coronavirus dataset from Our World in Data because random forest was the statistical learning method I found most interesting during this course. Because of the nature of that dataset with so many other columns that are potential alternative dependent variables rather than potential regressors, I had to manually enter all the included regressors into the random forest function in RStudio. This is that function:

```
set.seed(1)
split = sample.split(owid$total_deaths, SplitRatio = 0.7)
owid.train = subset(owid, split == TRUE)
owid.test = subset(owid, split == FALSE)

modelRF = randomForest(total_deaths ~ stringency_index+population+
population_density+median_age+aged_70_older+gdp_per_capita+extreme_poverty+
cardiovasc_death_rate+diabetes_prevalence+handwashing_facilities+
hospital_beds_per_thousand+life_expectancy+human_development_index,
      data = owid.train, mtry = 7, importance = TRUE, na.action = na.omit)
```

The results of this random forest regression model ran on the training dataset were



```
Call:
 randomForest(formula = total_deaths ~ stringency_index + population +      popul
ation_density + median_age + aged_70_older + gdp_per_capita +      extreme_povert
y + cardiovasc_death_rate + diabetes_prevalence +      handwashing_facilities + h
ospital_beds_per_thousand + life_expectancy +      human_development_index, data
 = owid.train, mtry = 7, importance = TRUE,      na.action = na.omit)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 7

          Mean of squared residuals: 33367703
                    % Var explained: 85.44
```

The reason I did not include this regression in the main text of my report is because I could never get my code for the test set mean squared error to run. Whenever I ran the following code:

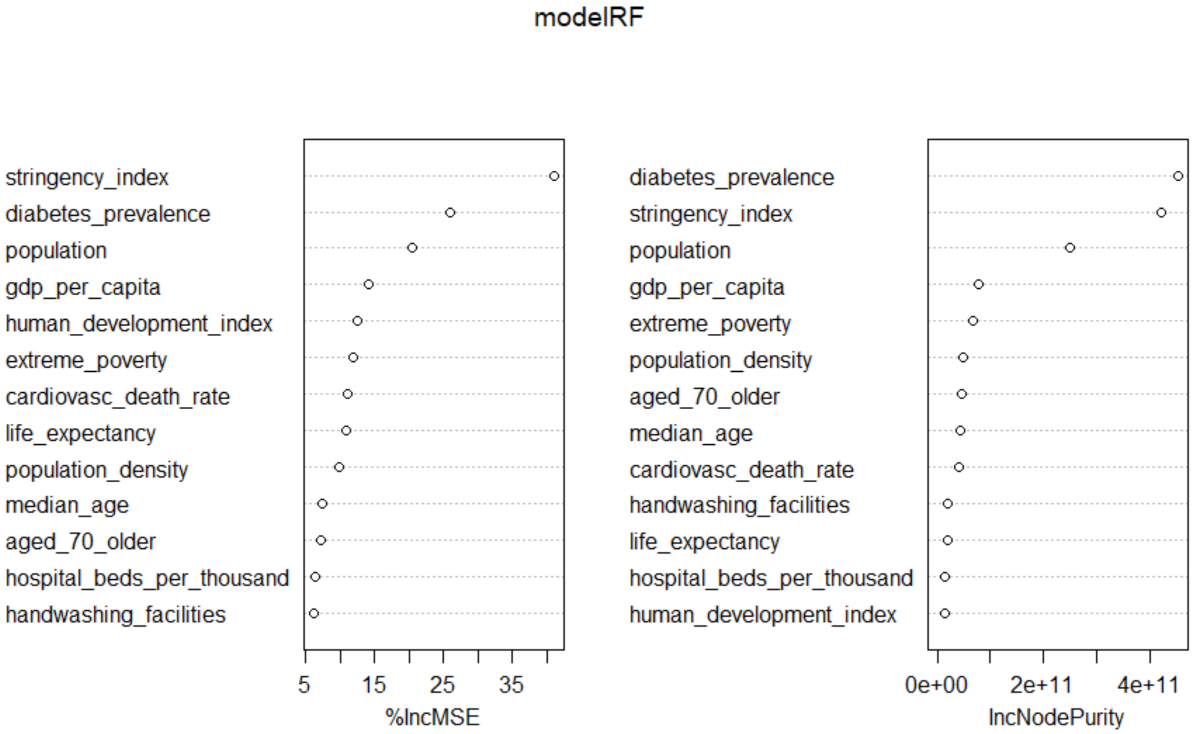
```
yhat.modelRF = predict(modelRF, newdata = owid.test)
mean((yhat.modelRF - owid.test)^2)
```

I would always get back the following error in the console:

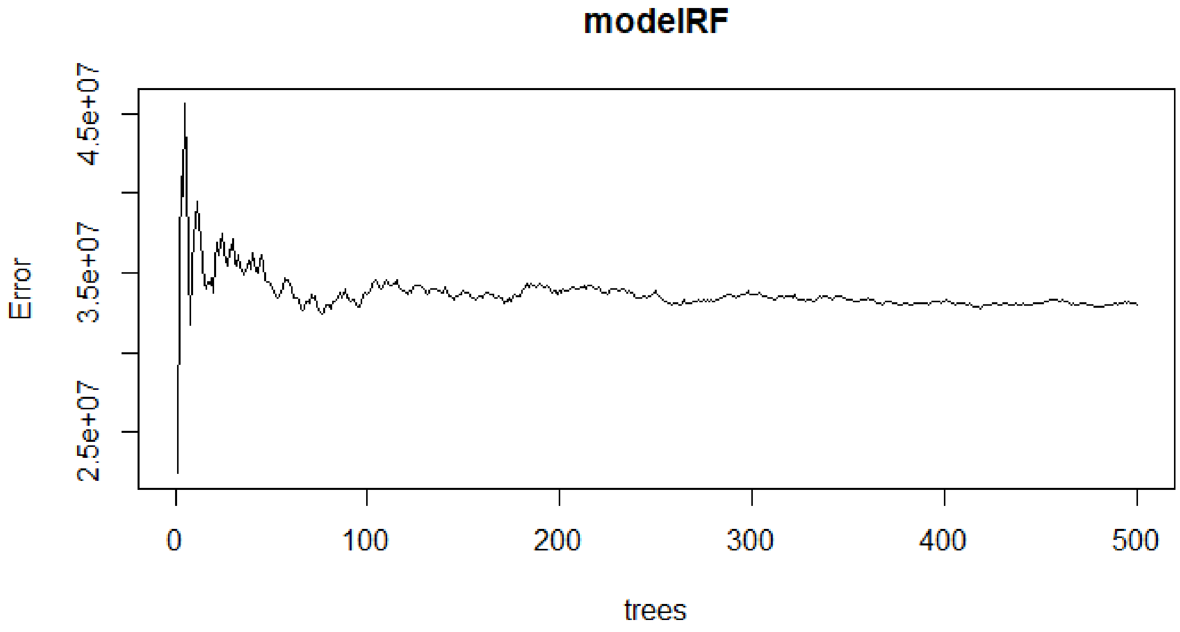
```
> mean((yhat.modelRF - owid.test)^2)
Error in FUN(left, right) : non-numeric argument to binary operator
```

So, after trying to solve it for multiple hours during consecutive days, I decided to relegate it to the appendix.

The variable importance plot for the results of the RF regression run on the training dataset:



The plot of the RF regression model itself:


modelRF
Error
2.5e+07
3.5e+07
4.5e+07
trees