Covid19 Project

Me

2023-04-11

Abstract:

The following report uses Covid-19 data from John Hopkins Github, and various metrics were explored. The way data was cleaned and parsed, organized, and what questions were to be answered was included in the final report.

Introduction:

The Covid 19 Pandemic has had many effects on the world. And thanks to modern technology, since it began people have been tracking and trending this data. A great data set that is easily accessible would be one provided by John Hopkins university, which can be found at https://github.com/CSSEGISandData/COVID-19. The data set includes, dates of occurrence, geographical areas, locations, populations and other relevant information. And it poses many great questions. The question that will be addressed today is going to be. Have are there now more per capita cases in Alabama or in New York city over time. Having a larger population overall, a reasonable hypothesis would be that the State of New York would have a higher amount of cases per thousand or million people than the State of Alabama.

$$H_0 = \text{New York has more cases per capita.}$$

$$H_a = \text{Alabama has more cases per capita.}$$

Materials and methods

The libraries used: | dplyr | lubridate | tidyverse | stringr| ggplot2|

The Raw File location is below for reference :

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi
```

Each file name is shown below, as the files are combined:

```
file_names <- c("time_series_covid19_confirmed_global.csv","time_series_covid19_deaths_global.csv","time
library(tidyverse)
urls <- str_c(url_in,file_names)
```

When Each file is examined, there is a large amount of extra data, along with the data being not in the proper format to be worked with. For example, the dates are individual which makes it impossible to properly sort the data for the purposes of this report. It also makes the data set much larger than it needs to be.

```
library(knitr)
kable(US_1)
```

UID	iso2	iso3	code3	FIPS
84001001	US	USA	840	1001
84001003	US	USA	840	1003
84001005	US	USA	840	1005
84001007	US	USA	840	1007
84001009	US	USA	840	1009

```
# A tiny snip of the data. Note the meaningless columns.
```

In order to better utilize the data. A form of data cleaning was performed. The built in `Pivot_longer` method was utilized to help put the data into a more useful format. All extra columns were filtered out. Leaving dates, locations, populations, cases and a few columns for future use. The raw code is below.

To further assist in data cleaning the `Lubridate` library was also utilized, in order to change the date format to something R can use properly. As string variables take up much more space than numerical.

```
library(lubridate)
US_cases <-US_cases %>%
  pivot_longer(cols = -(UID:Combined_Key),
               names_to ="date",
               values_to = "cases") %>%
  select(Admin2:cases) %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat,Long_))

US_deaths <- US_deaths %>%
  pivot_longer(cols = -(UID:Population),
               names_to = "date",
               values_to ="deaths") %>%
  select(Admin2:deaths) %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat, Long_))

US <- US_cases %>%
  full_join(US_deaths)


US_1 <- US[1:6,1:6]
```

A more relevant slice of data is highlighted below.

```
library(knitr)
kable(US_1)
```

Admin2	Province_State	Country_Region	Combined_Key	date	cases
Autauga	Alabama	US	Autauga, Alabama, US	2020-01-22	0
Autauga	Alabama	US	Autauga, Alabama, US	2020-01-23	0
Autauga	Alabama	US	Autauga, Alabama, US	2020-01-24	0
Autauga	Alabama	US	Autauga, Alabama, US	2020-01-25	0
Autauga	Alabama	US	Autauga, Alabama, US	2020-01-26	0
Autauga	Alabama	US	Autauga, Alabama, US	2020-01-27	0

As a final step the data was grouped again, and filtered for when case counts were above 0. The deaths per million people metric was also calculated. And the formula for reference is below. Along with source code.

$$deaths.per.million = deaths * 1000000/population$$

```
US_by_state <- US %>%
  group_by(Province_State, Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
            Population = sum(Population)) %>%
  mutate(deaths_per_mill = deaths * 1000000/ Population) %>%
  select(Province_State, Country_Region, date,
         cases, deaths, deaths_per_mill, Population) %>%
  ungroup()

US_totals <- US_by_state %>%
  group_by(Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
            Population = sum(Population)) %>%
  mutate(deaths_per_mill = deaths * 1000000/ Population) %>%
  select(Country_Region, date,
         cases, deaths, deaths_per_mill, Population) %>%
  ungroup()
```

Analyzing data using visualizations

Then the number of cases and deaths VS time was plotted to provide a basic visualization.(plot1)

To continue analysis of data. The cases were filtered by New York, followed by Alabama. And the same plots were rendered again using the filtered data. (plo2,plot3)

Then both sets of data were overlaid in the same space in order to preliminary view them.(plot4,plot5)

Then a plot was shown to show deaths per 1000 individuals VS dates in each state (plot7) And that same data was plotted using a log transformation.

And a linear model was created using cases per thousand and deaths per thousand. It is very apparent that there is indeed a slight linear trend upward. Which also makes sense on just a basic level. In general, all things being equal, more cases would generally equal more deaths. (predic1)

Results/Conclusion and discussion:

As of 2023 the per ca-pita deaths of individuals was indeed higher in Alabama than in New York. And thus the decision was reached to reject the null hypothesis.

From the simple raw numbers a very straightforward conclusion was reached.

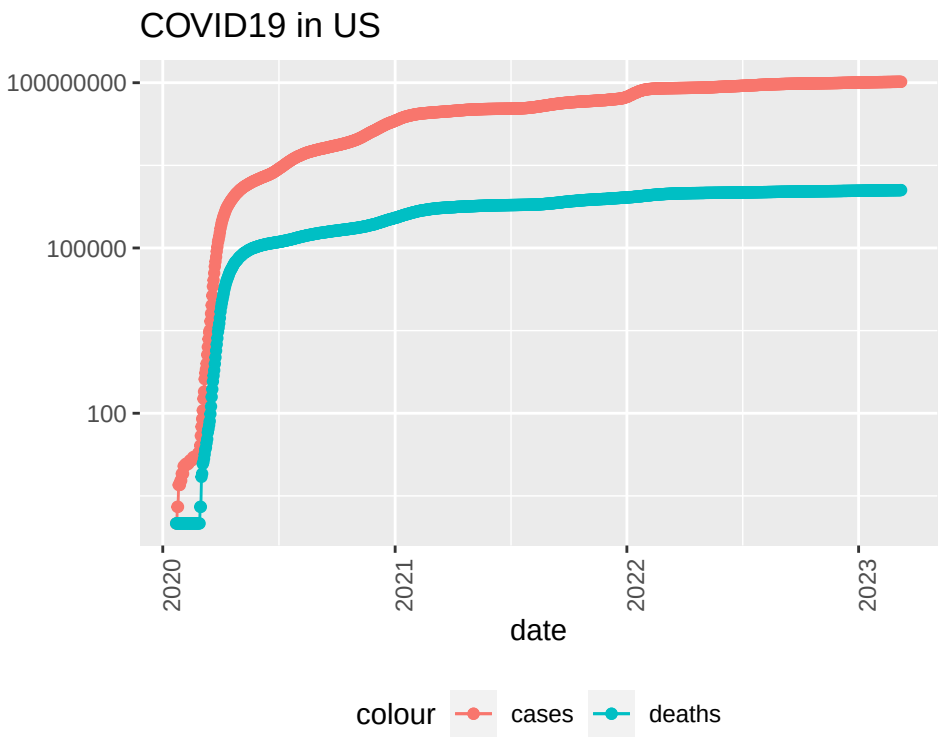
What was interesting and of note would be the cases and deaths log plots for each state. When looking at curves for data such as these (Please refer to plot 2 and plot 3) a squarish shape can mean the data reached saturation so to speak. To give an example (please humor me here) when measuring activity levels in an enzymatic reaction, if you have too much substrate often times the log plot creates this same squarish shape. This is because instead of proportionally reacting, the enzyme quickly uses up everything and saturates itself. Thank you for humoring me.

But what this tells me is there were either not enough tests or not enough labs to do the testing that was needed. While the testing log plot for Alabama looks more rounded, so not as saturated. Which may mean not enough testing. The deaths plot is incredibly depressed. So that means to me. Deaths were maybe under-reported, just a smidge. How can this conclusion be drawn? Well because of the low shape of the curve. If there is not enough substrate then the reaction will proceed only very slowly. And when the cases and deaths flattened out for both states, there was still a large gap between the two log curves and the Alabama curve still had some curve to go.(please see plot 4 and plot 5 on page 6) And that is why the deaths in Alabama eventually eclipse Deaths in New York. (plot7) Because the New York Cases flatten completely. But the cases in Alabama do not. So there is still growth and increase. And eventually Alabama catches up and eclipses New York in the number of deaths. And incredibly dark tortoise and hare story if we may. This is easily proven when the log of data used for plot 7 is plotted. (plot 8) It helps to highlight and better illustrate the rate of change. When the data for both states was log transformed using deaths per million and total deaths, and this data was used to create a prediction, using a linear model, an interesting effect was noted. The early estimates for deaths for both states was negative. (plot 9) It is not possible to have negative deaths so the data needed to be fit to 0.The change did not make the model any more accurate.(plot11) In fact the prediction actually turned out worse than the actual results. (plot 12) But the trend still stands, with New York state having lower per ca-pita death due to Covid, VS Alabama. Plot 12, which is a table, shows the numerical outputs of each prediction. And the predictions themselves still follow the same trend.

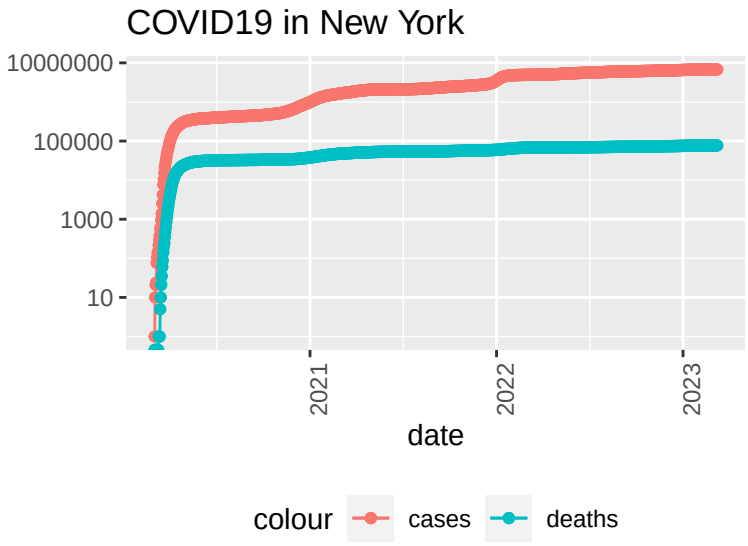
We as data scientists simply report what we see. Hopefully the data that is carefully gathered and curated by those in the John Hopkins repository gets put to good use. As a side note: A possible source of bias would have been the fact that cases were only tracked as a sum. Instead of daily cases as a number. Something like that would make it very easy to over estimate the severity of cases or the impact of Covid on an area. Or under estimate a large change in numbers. Or to see an area negatively, maybe not moving there or moving out. Another source of bias would be the linear model that was created showing a smaller estimation of current cases than was truly the case. If data like this is used to allocate resources, many places would not receive the funding they need or would not themselves devote enough resources.

Figures

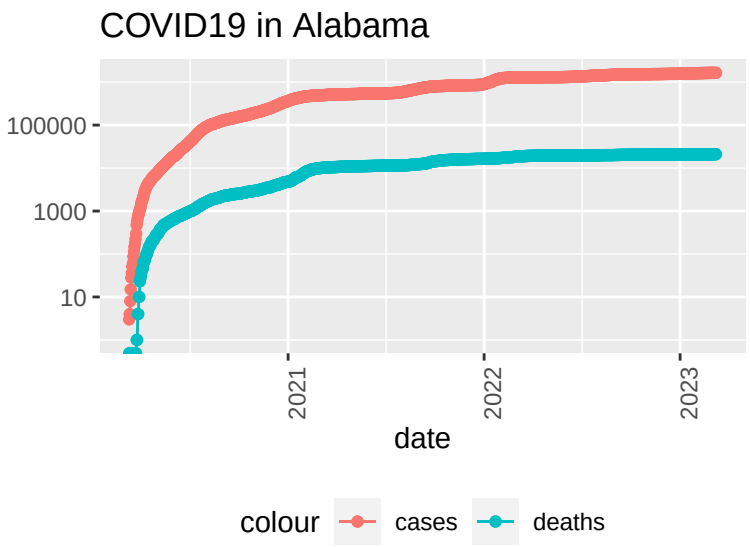
```
plot(plot1)
```



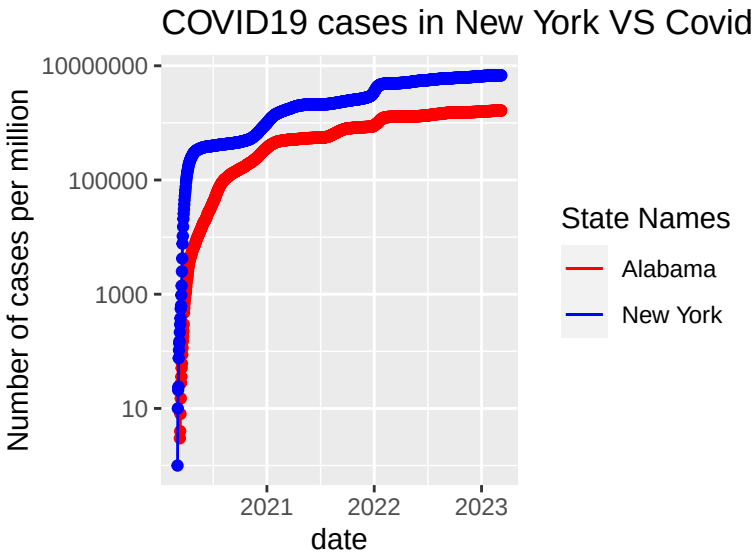

```
plot(plot2)
```



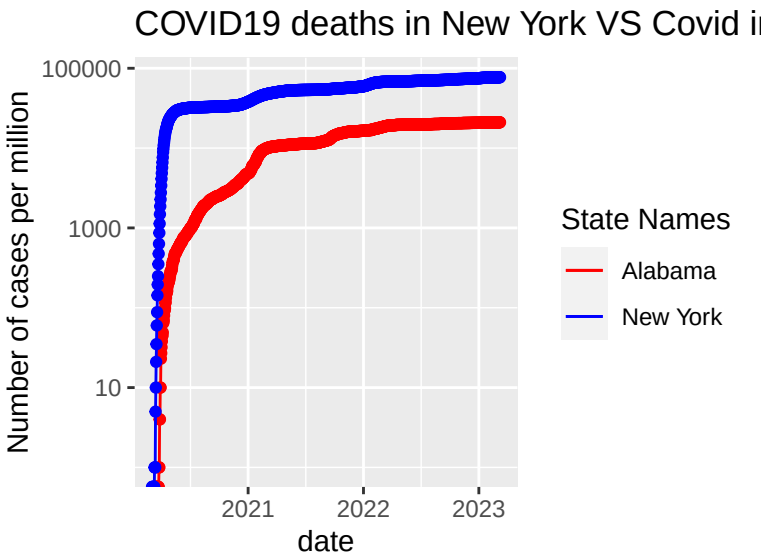
```
plot(plot3)
```



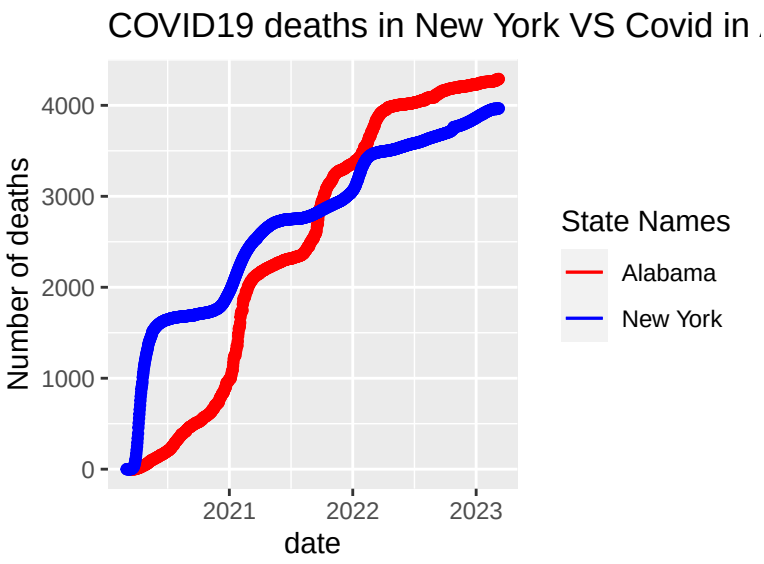

```
plot(plot4)
```



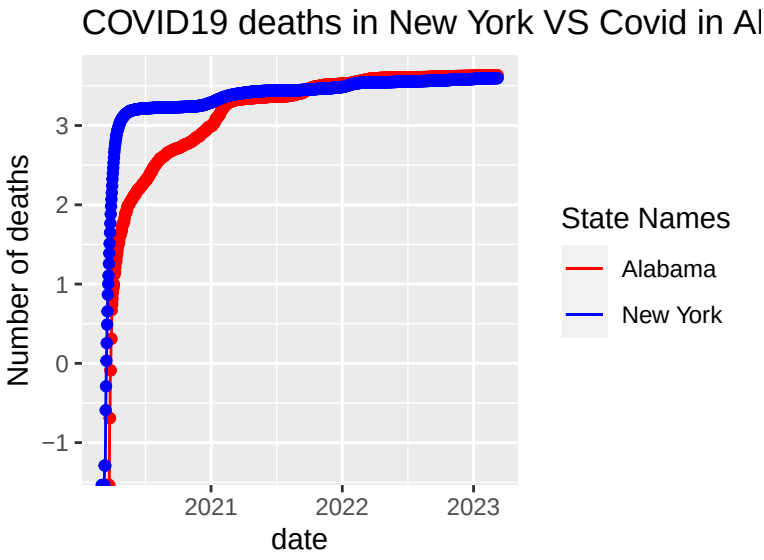

```
plot(plot5)
```


COVID19 deaths in New York VS Covid i
Number of cases per million
100000
1000
10
2021
2022
2023
date
State Names
Alabama
New York

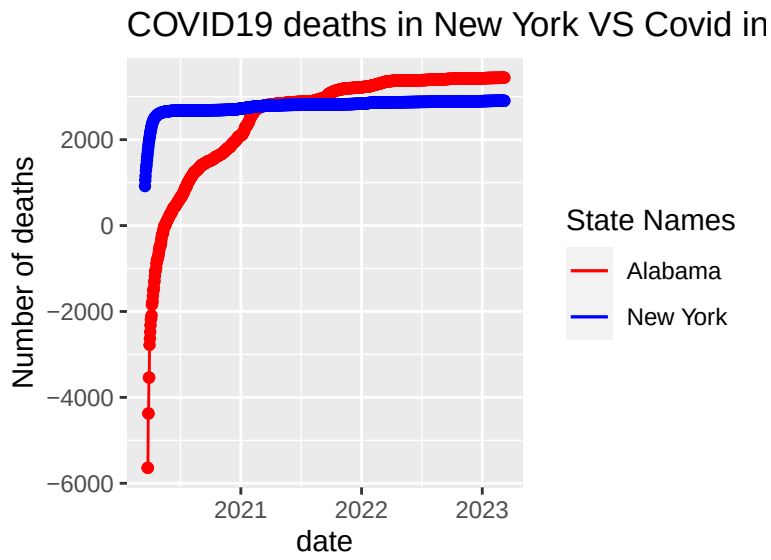
```
plot(plot7)
```




```
plot(plot8)
```



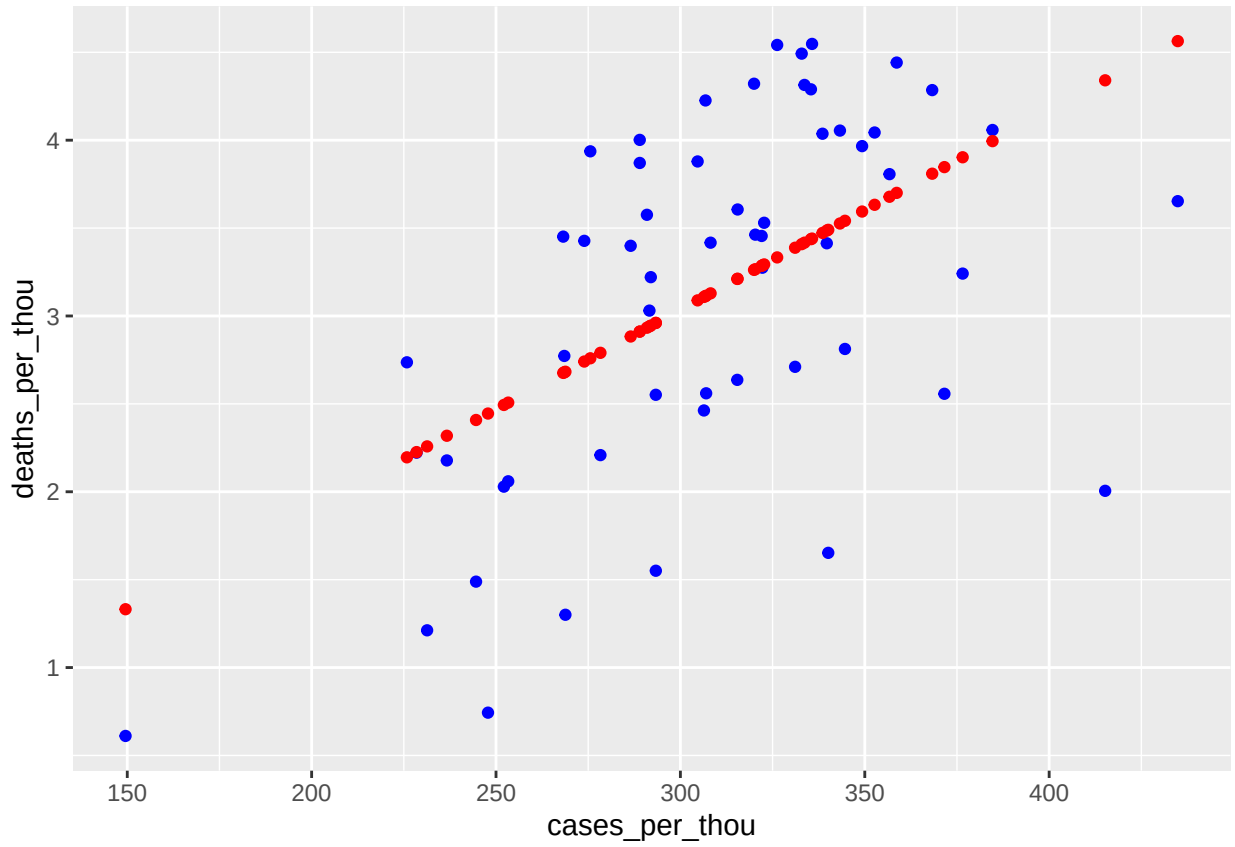
```
library(knitr)
library(dplyr)

plot(plot9)
```




```
kable(plot12)
```

Actual deaths AL	predictied deaths AL	predictied deaths AL 0 adjust	Actual deaths NY	predictied deaths NY	predictied deaths NY 0 adjust
4283.134	3441.952	2961.102	3962.308	2904.366	2904.366
4283.134	3441.952	2961.102	3962.308	2904.366	2904.366
4283.134	3441.952	2961.102	3962.308	2904.366	2904.366
4283.134	3441.952	2961.102	3957.425	2904.048	2904.048
4289.457	3443.298	2961.540	3958.967	2904.149	2904.149
4289.457	3443.298	2961.540	3960.355	2904.239	2904.239

```
plot(predic1)
```




```
mod
```

```
##
## Call:
## lm(formula = deaths_per_thou ~ cases_per_thou, data = US_state_totals)
##
## Coefficients:
##    (Intercept)  cases_per_thou
##       -0.36167         0.01133
```