COVID-19 Analysis

Student

2023-03-07

Questions of Interest 1.) Did Florida and New York differ in the amount of COVID-19 cases? 2.) What is the predicted increase in cases per month going forward?

First, I am going to declare my libraries.

```
library(tidyverse)
library(lubridate)
library(ggplot2)
```

Below is a link to the data used in this analysis. The data obtained is from Johns Hopkins University. It contains the the cumulative number of cases and deaths for each day starting 01-22-2020, when the first cases appeared in Washington State. Data is available for US and globally. For this analysis, we will focus on US data, both in aggragate and for specific states, specifically Florida and New York.

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

Next, we are going to load the data for US confirmed cases and deaths. The US data includes records by county and state. The US deaths data even includes the population by county.

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

urls <- str_c(url_in,file_names)

US_cases <- read_csv(urls[1])
US_deaths <- read_csv(urls[2])
```

Next we will clean the data and transform into 'tidy' data. Currently, each date is represented as column, but will make it 'tidy' by pivoting the data so that each date is in a row.

```
## Clean the US cases data and transform into 'tidy' data
US_cases <- US_cases %>%
  pivot_longer(cols = -c(UID:Combined_Key),
               names_to = "date",
               values_to = "cases") %>%
  select(Admin2:cases) %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat, Long_))

## Clean the US deaths data and transform into 'tidy' data
US_deaths <- US_deaths %>%
```

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

## Merge cases and deaths into one single file
US <- US_cases %>%
  full_join(US_deaths)
```

Now, we will clean the date further by creating to separate datasets to help answer our questions of interest. Since different areas have different populations, deaths per million people and cases per million people are calculated in order to standardize the values by population.

```
US_by_State <- US %>%
  group_by(Province_State,Country_Region,date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths), Population = sum(Population)) %>%
  select(Province_State, Country_Region, date,
         cases, deaths, Population) %>%
  filter(Population > 0) %>%
  mutate(deaths_per_mill = deaths * 1000000 / Population,
         cases_per_mill = cases * 1000000 / Population) %>%
  ungroup()

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

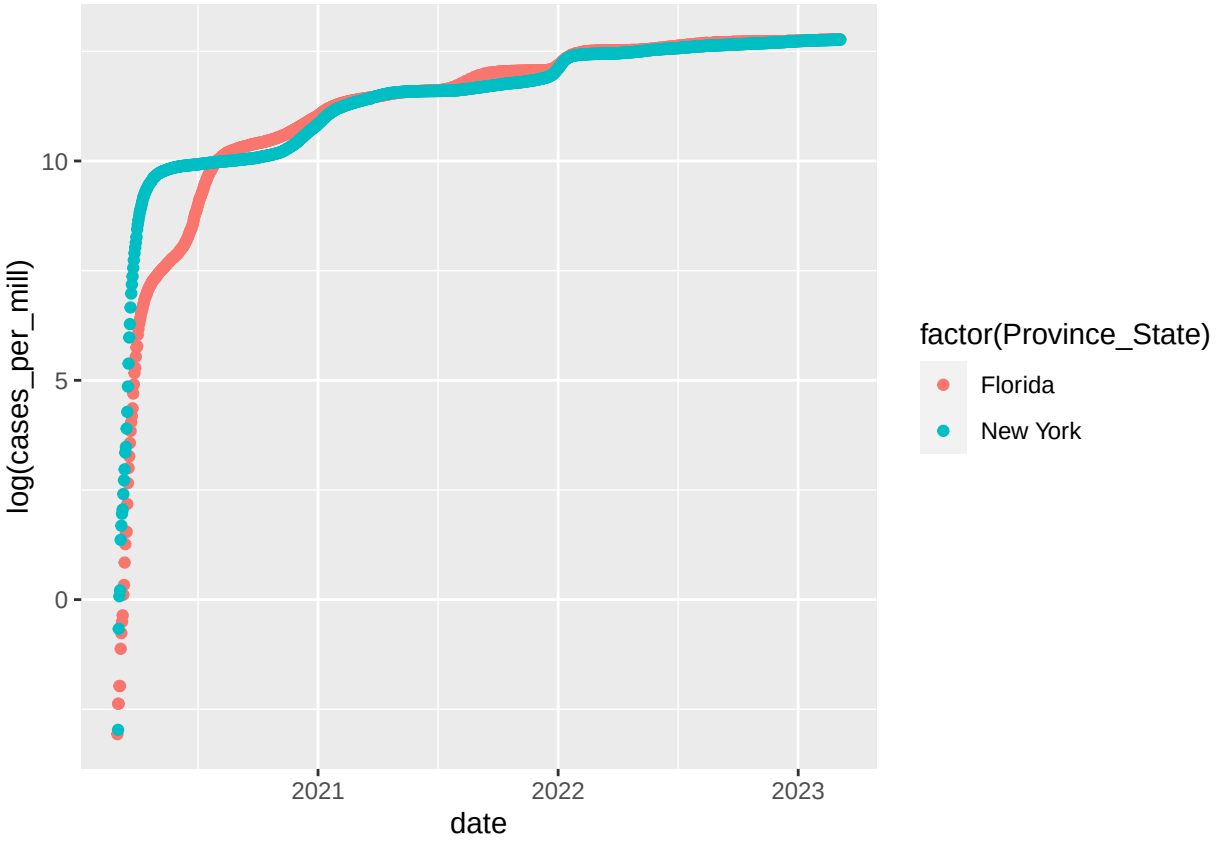
Now, we will compare Florida to New York. During the earlier stages of the pandemic, these two states were at the center of the conversation. New York was categorized as having very strict restrictions to reduce the spread of Covid-19 and Florida was categorized as having more relaxed restrictions. Let's see how their charts of cumulative cases per million people and deaths per million people compare.

First, we will filter the data to include only Florida and New York

```
## Filtor the date to include only Florida and New York
FL_NY <- US_by_State %>%
  filter(cases > 0,
         Province_State == "Florida" |
         Province_State == "New York")
```

Next, we will plot the log of cumulative cases per million people for each state.

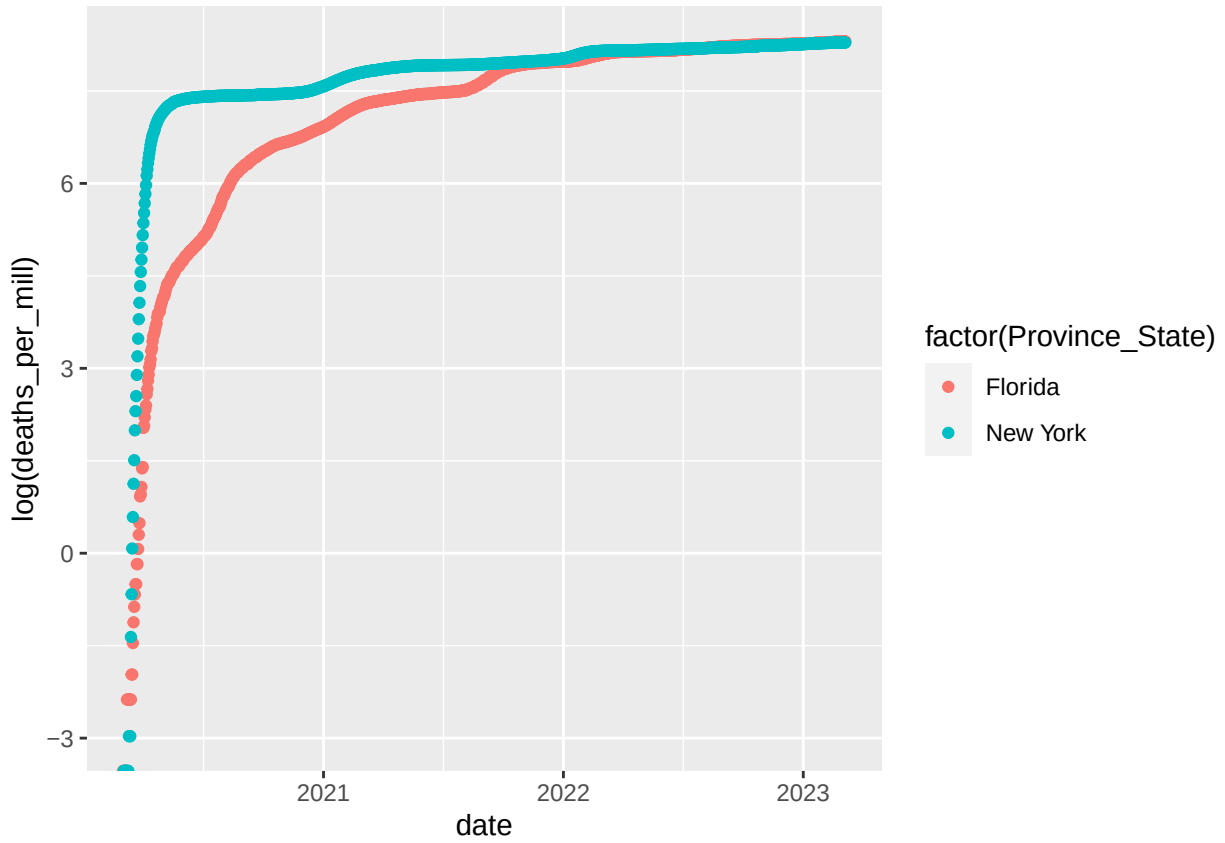
```
## Plot cumulative cases over time by state
ggplot(FL_NY, aes(x = date, y=log(cases_per_mill), colour = factor(Province_State))) +
  geom_point()
```



In this plot, log of the cases per million looks similar between the two states. There is some variation, particularly in the early part of the pandemic. But, in general, they look very similar.

Now, we will plot the log of cumulative deaths per million people for each state.

```
## Plot cumulative deaths over time by state
ggplot(FL_NY, aes(x = date, y=log(deaths_per_mill), colour = factor(Province_State))) +
  geom_point()
```



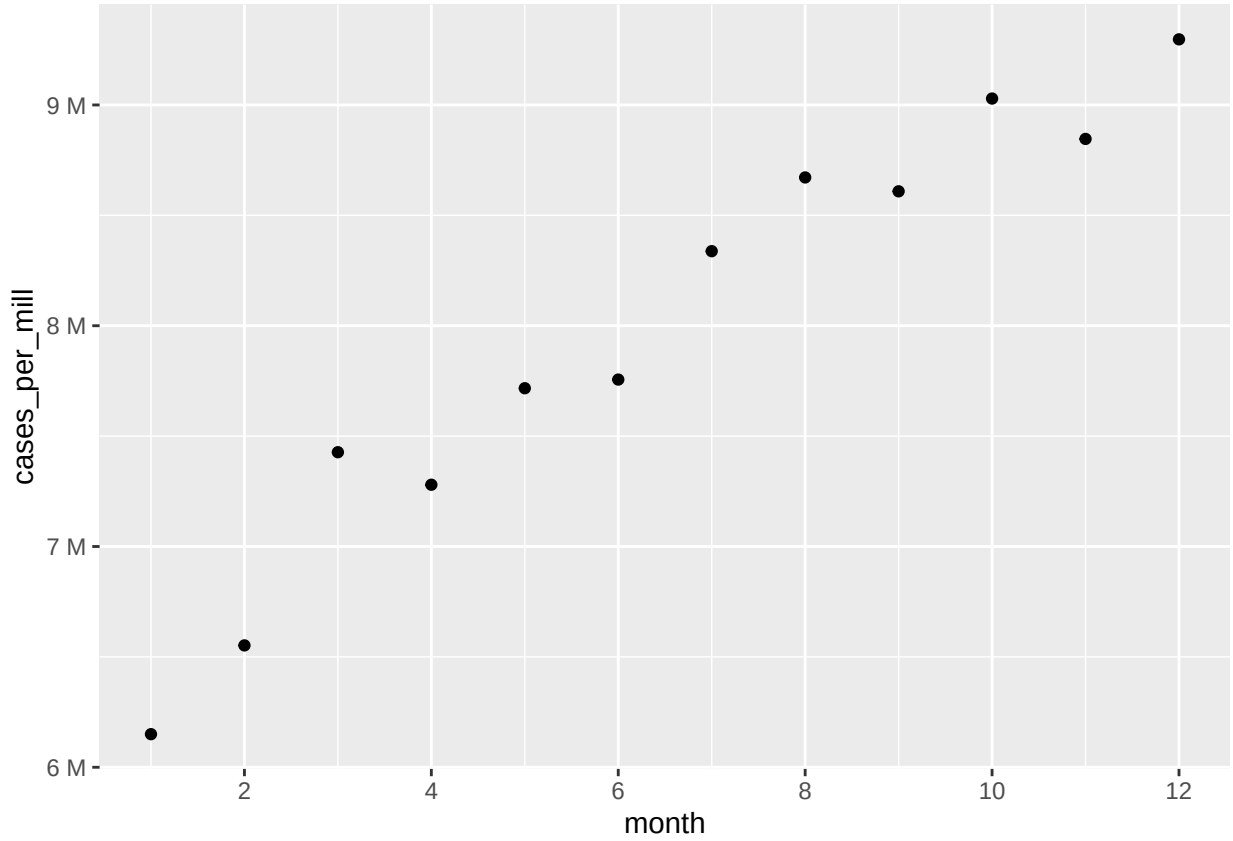
These charts look similar as well. New York had more deaths in the early part of the pandemic. But, at that time, everyone was just trying to figure things out.

My conclusion, just looking at these two graphs, is that, despite having very different approaches to handling the pandemic, New York and Florida had very similar results in terms of cases and deaths. However, this analysis contains bias that need to be explored. First, these two states have very different weather and population density. In is not unreasonable to assume that if they both handled Covid-19 in exactly the same way, then Florida would have greatly outperformed New York due to other factors. Also, these two states were cherry-picked due their prevalence in the national conversation. I would assume that it is possible to pick two separate states with different approaches to Covid-19 and draw a very different conclusion. In short, while at first glance, it would appear that the additional Covid-19 restrictions did not greatly effect outcomes, this is analysis is way too simple and would require much deeper investigation in order to come up with a convincing argument to that affect.

Now, let look at the increase in Covid-19 cases across the entire USA.

```
US_2022 <- US_Totals %>%
  filter(year(date) == 2022) %>%
  group_by(month = month(date)) %>%
  summarize(cases_per_mill = sum(cases_per_mill)) %>%
  arrange(month)

ggplot(US_2022, aes(x = month, y = cases_per_mill)) +
  geom_point() +
  scale_x_continuous(breaks = scales::pretty_breaks()) +
  scale_y_continuous(labels = scales::label_number(suffix = " M", scale = 1e-6))
```



There appears to be linear relationship. However, it is curious, that there is drop in cases in April. I would expect the date to be strictly increases. Something to explore at a later date.

Finally, let's run a linear regression model on this data.

```
lm_cases = lm(cases_per_mill ~ month, US_2022)
summary(lm_cases)
```

```
##
## Call:
## lm(formula = cases_per_mill ~ month, data = US_2022)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -350197 -164997  -28421  167149  391416
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6232438     155701   40.03 2.26e-12 ***
## month         267726      21156   12.65 1.77e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 253000 on 10 degrees of freedom
## Multiple R-squared:  0.9412, Adjusted R-squared:  0.9354
## F-statistic: 160.2 on 1 and 10 DF,  p-value: 1.769e-07
```

The model looks like a good fit, showing an increase in cases of about 270,000 cases per month. However, it is only 12 records. So, it is a very small sample. There is bias in the model as well. I limited the data to only 2022. The analysis could have started at anytime in the data, but a lot has happened over last three years, including the distribution of a vaccine. Changing the start and end date would likely result in a very different model.