COVID-19 Data Repository Report

Eric Yu

2024-06-25

Introduction

This report is an analysis of the COVID-19 Data Repository by John Hopkins University. Currently, the data-set contains recorded cases from 2020 to 2023.

Research Questions

- How do cases/deaths in the most populated states compared to the total in the US?
- Can we use the data to predicted the death-rate of COVID cases?

Importing Libraries and Data-set

The analysis will be performed using R libraries. These tools will be used to import the data-set of interest, clean/transform the data, and generate results/visualizations.

The COVID-19 data contains four time-series data-sets:

- US confirmed cases
- Global confirmed cases
- US deaths
- Global deaths

The US data-sets record the number of cases per date at the level of counties for every state. The global data-sets record the number of cases per date at the level of country/region. For this report, I will focus on COVID cases in the US.

Load R libraries

```r
library(tidyverse)
library(ggplot2)
library(tidyr)
library(lubridate)
library(maps)
library(mapproj)
```

Load data

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

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)

us_cases <- read.csv(urls[1])
global_cases <- read.csv(urls[2])
us_deaths <- read.csv(urls[3])
global_deaths <- read.csv(urls[4])
```

Tidy and clean data

In both the US and global data-sets, I want to view the cases and deaths by date (a row for each date). The default format of the dates is a string padded with an "X" character, i.e. "XMM.DD.YY" or "XM.DD.YY". To convert the strings into a date object, "X" characters will be removed using the built-in gsub() function to transform the dates into a format that can be interpreted by the lubridate library.

```
# Reformat data to have date on every row
us_cases <- us_cases %>%
  pivot_longer(cols = -(UID:Combined_Key),
               names_to = "date",
               values_to = "cases") %>%
  select(Admin2:cases) %>%
  mutate(date = mdy(gsub("X", "", date)))

global_cases <- global_cases %>%
  pivot_longer(cols = -c("Province.State",
                         "Country.Region", Lat, Long),
               names_to = "date",
               values_to = "cases") %>%
  mutate(date = mdy(gsub("X", "", date)))

us_deaths <- us_deaths %>%
  pivot_longer(cols = -(UID:Population),
               names_to = "date",
               values_to = "deaths") %>%
  select(Admin2:deaths) %>%
  mutate(date = mdy(gsub("X", "", date)))

global_deaths <- global_deaths %>%
  pivot_longer(cols = -c("Province.State",
                         "Country.Region", Lat, Long),
               names_to = "date",
               values_to = "deaths") %>%
  mutate(date = mdy(gsub("X", "", date)))
```

Join US data-sets

```r
us <- us_cases %>%
  full_join(us_deaths)

glimpse(us)
```

```
## Rows: 3,819,906
## Columns: 10
## $ Admin2         <chr> "Autauga", "Autauga", "Autauga", "Autauga", "Autauga", ~
## $ Province_State <chr> "Alabama", "Alabama", "Alabama", "Alabama", "Alabama", ~
## $ Country_Region <chr> "US", "US", "US", "US", "US", "US", "US", "US", "US", "~
## $ Lat            <dbl> 32.53953, 32.53953, 32.53953, 32.53953, 32.53953, 32.53~
## $ Long_          <dbl> -86.64408, -86.64408, -86.64408, -86.64408, -86.64408, ~
## $ Combined_Key   <chr> "Autauga, Alabama, US", "Autauga, Alabama, US", "Autaug~
## $ date           <date> 2020-01-22, 2020-01-23, 2020-01-24, 2020-01-25, 2020-0~
## $ cases          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ Population     <int> 55869, 55869, 55869, 55869, 55869, 55869, 55869, 55869,~
## $ deaths         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

Join global data-sets

```r
global <- global_cases %>%
  full_join(global_deaths)

glimpse(global)
```

```
## Rows: 330,327
## Columns: 7
## $ Province.State <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", "",~
## $ Country.Region <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanist~
## $ Lat            <dbl> 33.93911, 33.93911, 33.93911, 33.93911, 33.93911, 33.93~
## $ Long           <dbl> 67.70995, 67.70995, 67.70995, 67.70995, 67.70995, 67.70~
## $ date           <date> 2020-01-22, 2020-01-23, 2020-01-24, 2020-01-25, 2020-0~
## $ cases          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ deaths         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

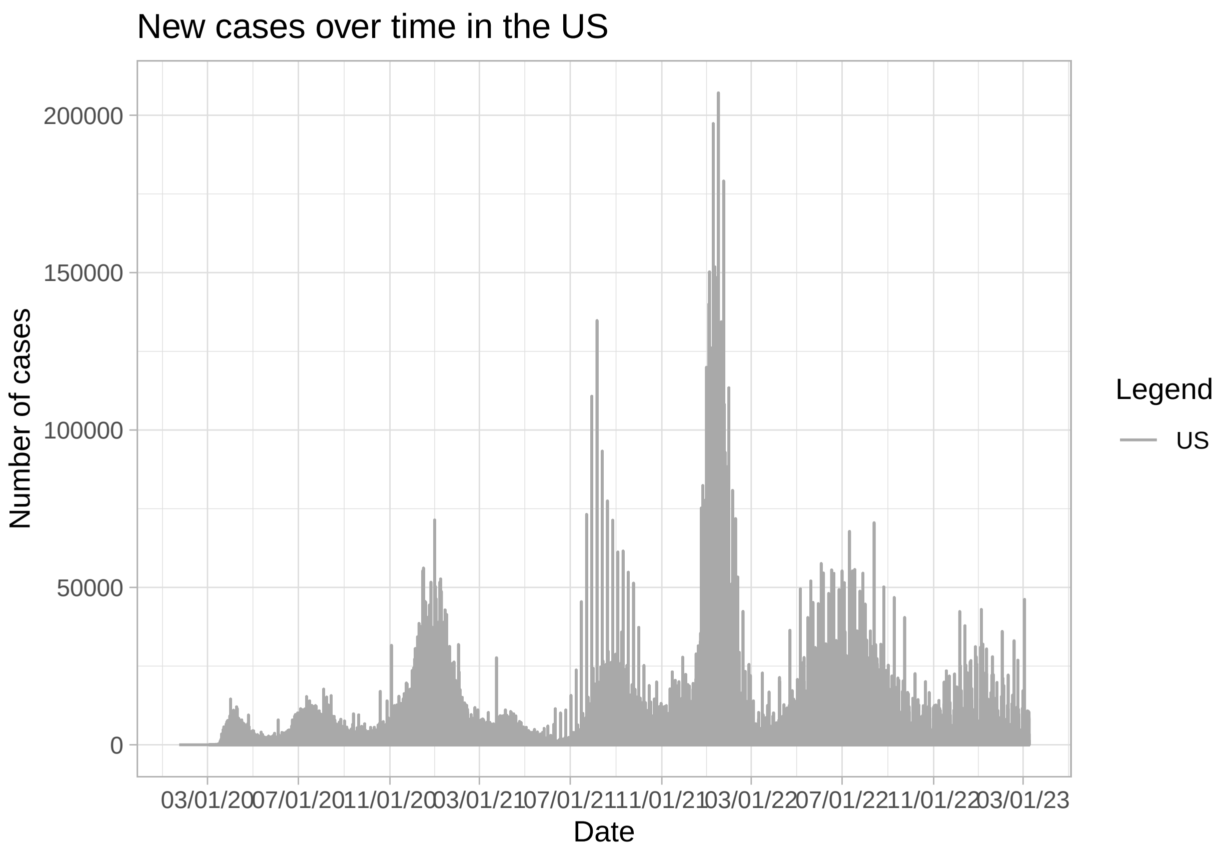
Exploring the Data

Examine new cases in the US over time

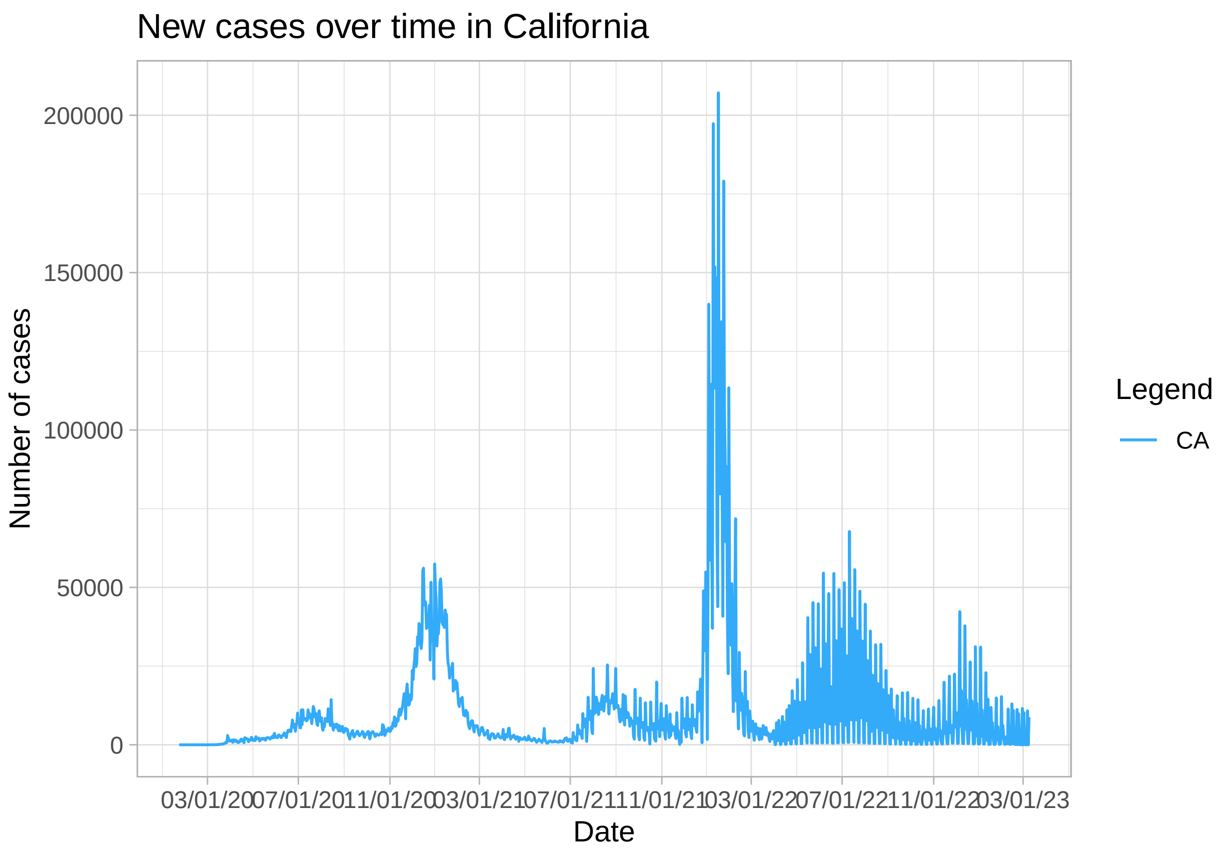
For this analysis, I aggregated the data to get the number of new cases and deaths by state in the US. This is distinct from the total cumulative cases/deaths, which would be the sum of all cases/deaths observed. Specifically, new cases for an individual date only counts cases observed on that date, while cumulative cases include all prior cases. For comparison, I plotted the data for the three most populated states in the US (California, Texas, Florida) in addition to the US as a whole.

```r
# Get cases and deaths in the US by state
states <- us %>%
  group_by(Province_State, date) %>%
  summarize(total_cases = sum(cases),
            total_deaths = sum(deaths),
            total_pop = sum(Population)) %>%
  ungroup() %>%
  mutate(new_cases = total_cases - lag(total_cases),
         new_deaths = total_deaths - lag(total_deaths))

# Remove NA entries
states <- states %>%
  filter(!is.na(new_cases),
         !is.na(new_deaths),
         new_cases >= 0,
         new_deaths >= 0)

# Get cases from CA
ca <- states %>%
  filter(Province_State == "California")

# Get cases from TX
tx <- states %>%
  filter(Province_State == "Texas")

# Get cases from FL
fl <- states %>%
  filter(Province_State == "Florida")
```

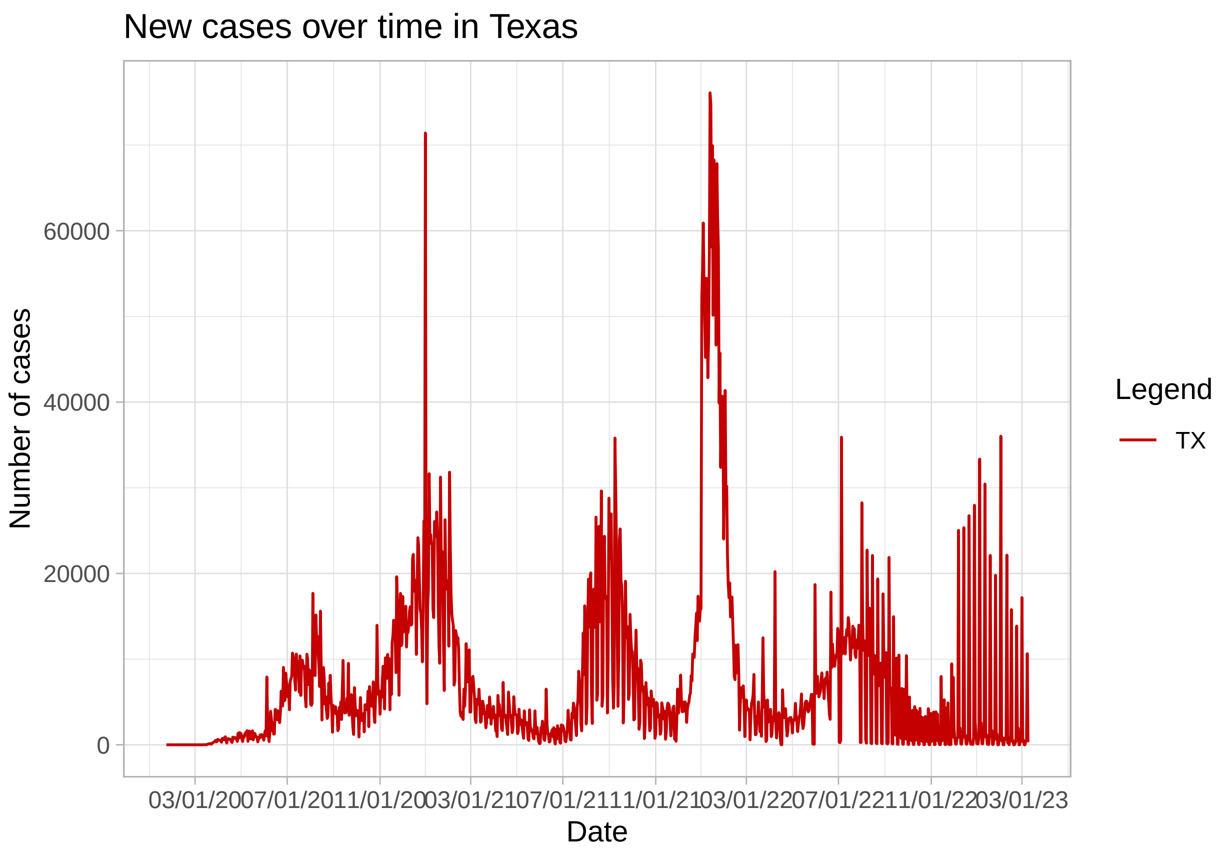
```r
ggplot() +
  geom_line(data=states, aes(x=date, y=new_cases, color="US")) +
  scale_color_manual(values=c("US"="#A9A9A9")) +
  scale_x_date(date_labels="%m/%d/%y", date_breaks= "4 month") +
  theme(plot.title = element_text(hjust=0.5)) +
  labs(x="Date", y="Number of cases",
       title="New cases over time in the US",
       color="Legend") +
  theme_light()
```



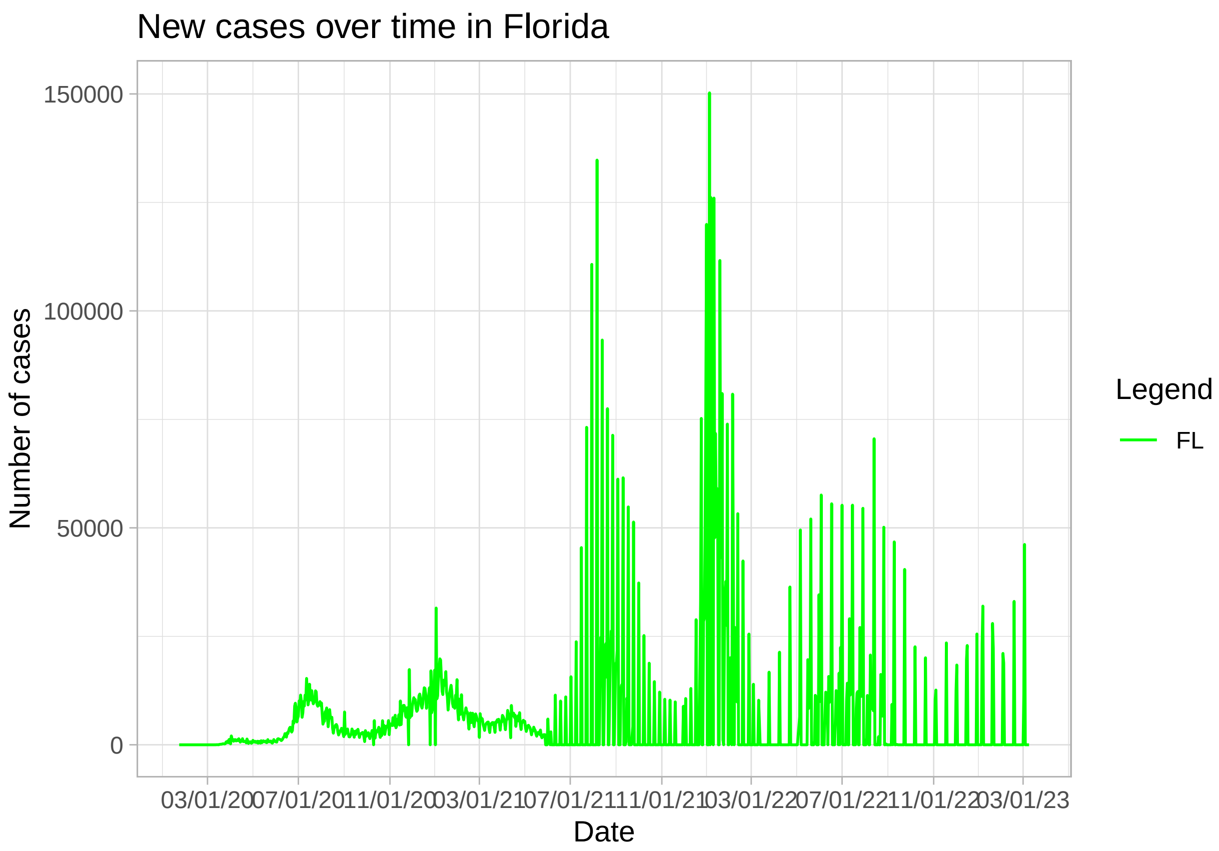
```
ggplot() +
  geom_line(data=ca, aes(x=date, y=new_cases, color="CA")) +
  scale_color_manual(values=c("CA"="#33ABF9")) +
  scale_x_date(date_labels="%m/%d/%y", date_breaks= "4 month") +
  theme(plot.title = element_text(hjust=0.5)) +
  labs(x="Date", y="Number of cases",
       title="New cases over time in California",
       color="Legend") +
  theme_light()
```



```
ggplot() +
  geom_line(data=tx, aes(x=date, y=new_cases, color="TX")) +
  scale_color_manual(values=c("TX"="#C40000")) +
  scale_x_date(date_labels="%m/%d/%y", date_breaks= "4 month") +
  theme(plot.title = element_text(hjust=0.5)) +
  labs(x="Date", y="Number of cases",
       title="New cases over time in Texas",
       color="Legend") +
  theme_light()
```



```r
ggplot() +
  geom_line(data=fl, aes(x=date, y=new_cases, color="FL")) +
  scale_color_manual(values=c("FL"="#00FF00")) +
  scale_x_date(date_labels="%m/%d/%y", date_breaks= "4 month") +
  theme(plot.title = element_text(hjust=0.5)) +
  labs(x="Date", y="Number of cases",
       title="New cases over time in Florida",
       color="Legend") +
  theme_light()
```

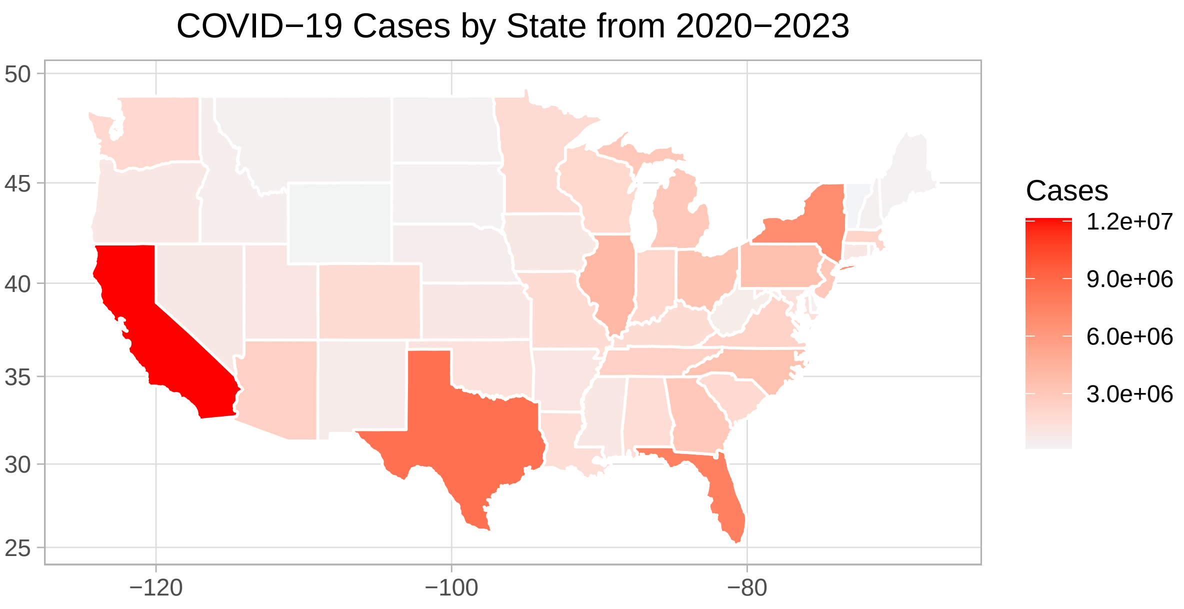


One observation I made is that the plots for the individual states take a similar shape to the plot for the US. The plot for California has the greatest similarity to the US plot, and this is logically sound because a large proportion of COVID cases are concentrated there. This can be observed more obviously using a heat map.

Heat map of COVID-19 cases in the US by state

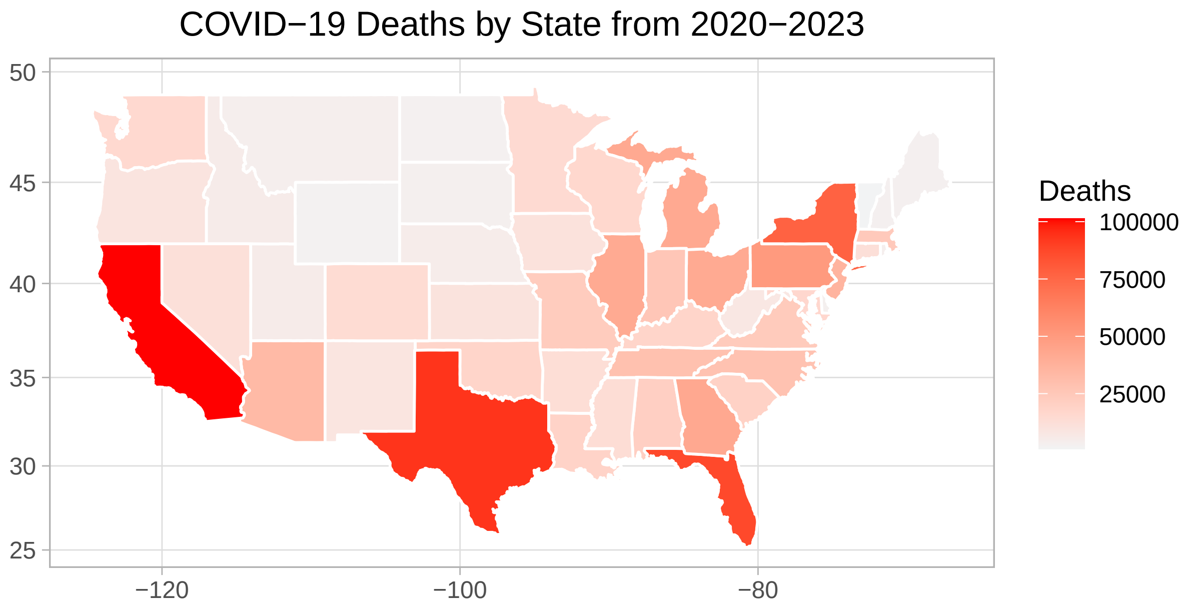
```
# Aggregate data by state
totals <- us %>%
  group_by(Province_State, date) %>%
  summarize(total_cases = sum(cases),
            total_deaths = sum(deaths),
            total_pop = sum(Population)) %>%
  ungroup() %>%
  group_by(Province_State) %>%
  summarize(total_cases = max(total_cases),
            total_deaths = max(total_deaths),
            total_pop = max(total_pop))

# Create heat map of the US
map_cases <- map_data("state")

# Lowercase states to match map
totals$Province_State <- tolower(totals$Province_State)

# Filter to only include regions in the map
```

```r
state_cases <- totals[totals$Province_State %in%
                           map_cases$region, ]

# Rename column to merge with map
colnames(state_cases)[colnames(state_cases) == "Province_State"] <- "region"

heatmap_cases <- merge(map_cases,
                       state_cases,
                       by="region",
                       all.x=TRUE)

# Plot map
ggplot(heatmap_cases, aes(x=long,
                          y=lat,
                          group=group,
                          fill=total_cases)) +
  geom_polygon(color="#FFFFFF") +
  scale_fill_gradient(low="#F2F3F4",
                      high="#FF0000",
                      na.value="#000000",
                      name="Cases") +
  theme_light() +
  theme(plot.title = element_text(hjust=0.5)) +
  labs(title="COVID-19 Cases by State from 2020-2023",
       x="", y="") +
  coord_map()
```



Heat map of COVID-19 deaths in the US by state

```
# Create heat map of the US
map_deaths <- map_data("state")

# Filter to only include regions in the map
state_deaths <- totals[totals$Province_State %in%
                             map_deaths$region, ]

# Rename column to merge with map
colnames(state_deaths)[colnames(state_deaths) == "Province_State"] <- "region"

heatmap_deaths <- merge(map_deaths,
                        state_deaths,
                        by="region",
                        all.x=TRUE)

# Plot map
ggplot(heatmap_deaths, aes(x=long,
                          y=lat,
                          group=group,
                          fill=total_deaths)) +
  geom_polygon(color="#FFFFFF") +
  scale_fill_gradient(low="#F2F3F4",
                      high="#FF0000",
                      na.value="#000000",
                      name="Deaths") +
  theme_light() +
  theme(plot.title = element_text(hjust=0.5)) +
  labs(title="COVID-19 Deaths by State from 2020-2023",
       x="", y="") +
  coord_map()
```



The heat-maps show that California has the highest number of total cases and deaths from 2020 to 2023, followed by Texas, Florida, and New York. A possible explanation for why COVID cases/deaths were concentrated in these states is that they have the highest populations in the US. Along with being on coastlines, the high population and population densities of these states serve as potential catalysts for spreading infection.

Analysis

Linear regression

I will use a linear regression to predict the death rate in the US based on the number of confirmed cases and population. For this model, I will consider the population of each state, the total number of confirmed cases, and the death rate (total deaths / total cases from 2020-2023).

```
model_data <- us %>%
  group_by(Province_State, date) %>%
  summarize(total_cases = sum(cases),
            total_deaths = sum(deaths),
            total_pop = sum(Population)) %>%
  ungroup()

model_data <- model_data %>%
  group_by(Province_State) %>%
  summarize(total_cases = max(total_cases),
            total_deaths = max(total_deaths),
            total_pop = max(total_pop),
            death_rate = max(total_deaths)/max(total_cases))

log_reg <- lm(death_rate ~ total_cases + total_pop,
              data = model_data)

summary(log_reg)
```

```
##
## Call:
## lm(formula = death_rate ~ total_cases + total_pop, data = model_data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0098193 -0.0020220  0.0003697  0.0022706  0.0193070
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.819e-03  6.851e-04  14.332   <2e-16 ***
## total_cases -1.569e-10  2.177e-09  -0.072    0.943
## total_pop    1.186e-10  6.822e-10   0.174    0.863
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.00406 on 55 degrees of freedom
## Multiple R-squared:  0.01556,    Adjusted R-squared:  -0.02024
## F-statistic: 0.4347 on 2 and 55 DF,  p-value: 0.6497
```

Interpreting Results

The population of a state and number of confirmed cases do not appear to be significantly associated with death rate. Specifically, the regression model shows `p-value = 0.943` for `total_cases` and `p-value = 0.863` for `total_pop`. These values are both greater than 0.05, indicating that they are not statistically significant. To predict death rate, there are likely several other factors to consider (which are not available in the data) such as patient demographics.

Bias

My source of bias is the reasoning behind choosing certain states. I chose states based on high population, but there can be several reasons for why states are affected by COVID differently. When evaluating results, it is important to consider facts that may not be available in the data, such as vaccine availability and travel activity by state.

Conclusion

My analysis showed that the highest concentration of COVID cases were in states with the largest population. However, this does not necessarily equate to large states being more susceptible to COVID outbreaks. Number of COVID cases and population do not appear to be significantly correlated to death rates due to COVID. There are likely many other factors to consider that are beyond the scope of this data-set.

Bibliography

- https://github.com/CSSEGISandData/COVID-19
- https://www.britannica.com/topic/largest-U-S-state-by-population