US COVID-19 - Data Wrangling and Analysis

Hema Chandra Yedoti

07/05/2021

Objective:

COVID-19, started in early 2020, is the largest global pandemic since the 1918 Spanish Flu and the greatest global crisis since World War II. Countries are still suffering from this devastating pandemic with continued increase in cases and deaths even after an year. Since the start of the pandemic, we have seen countless dashboards and visualization plots tracking the COVID-19 cases across the world. In this project, I would like to take the United States COVID-19 data and make some plots using this data to see the data trends and gather some insights from the plots.

```
library(rvest)
library(dplyr)
library(tidyverse)
library(zoo)
library(tidyr)
library(reshape)
library(boot)
library(splines)
```

COVID-19 DATA: UNITED STATES

NY Times has been maintaining comprehensive datasets, in their github profile, which keeps track of COVID-19 cases and deaths country-wise, state-wise(US) and county-wise(US). For this project, I am using only the state-wise dataset.

Data Loading and Cleaning:

```
nyt_state_url = "https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-states.csv"

nyt_state_data = nyt_state_url %>%
  url() %>%
  read_csv()

dim(nyt_state_data)
```

```
## [1] 23719     5
```

```r
head(nyt_state_data)
```

```
## # A tibble: 6 x 5
##   date       state      fips  cases deaths
##   <date>     <chr>      <chr> <dbl>  <dbl>
## 1 2020-01-21 Washington 53        1      0
## 2 2020-01-22 Washington 53        1      0
## 3 2020-01-23 Washington 53        1      0
## 4 2020-01-24 Illinois   17        1      0
## 5 2020-01-24 Washington 53        1      0
## 6 2020-01-25 California 06        1      0
```

- Dimensions of Dataset: 23719 rows and 5 columns
- Data types for the variables are already taken care of.
- 'cases' & 'deaths' variable indicate the cumulative sum of cases & deaths in the state respectively.
- 'fips' is a unique identifier for each state and not of much importance for our analysis.

```r
# Removing 'fips' & 'state' variable to create a dataset for the country:
usa_data <- nyt_state_data %>%
  select(-c("state","fips")) %>%
  group_by(date) %>%
  summarise(cases = sum(cases, na.rm = TRUE), deaths = sum(deaths, na.rm = TRUE))

usa_state_data <- nyt_state_data %>%
  select(-c("fips"))
```

- I grouped the state-level data to get the national data for US.

```r
#Extracting and adding daily values from cumulative values for cases & deaths in national level data:
usa_data$date <- as.Date(usa_data$date)
usa_data <- usa_data %>%
  mutate(daily.cases = (cases - lag(cases)), daily.deaths = (deaths - lag(deaths)))
usa_data[is.na.data.frame(usa_data)] <- 0

#Extracting and adding daily values from cumulative values for cases & deaths in state-level data:
usa_state_data$date <- as.Date(usa_state_data$date)
usa_state_data <- usa_state_data %>%
    group_by(state) %>%
    arrange(date) %>%
    mutate(daily.cases = (cases - lag(cases)), daily.deaths = (deaths - lag(deaths)))
usa_state_data[is.na.data.frame(usa_state_data)] <- 0
```

```r
#Saving the tidy version of scraped US COVID-19 data:

write.csv(usa_data,"usa_data.csv", row.names = FALSE)
write.csv(usa_state_data,"usa_state_data.csv", row.names = FALSE)
```

Analysis:

```
head(usa_data)
```

```
## # A tibble: 6 x 5
##   date       cases deaths daily.cases daily.deaths
##   <date>     <dbl>  <dbl>       <dbl>        <dbl>
## 1 2020-01-21     1      0           0            0
## 2 2020-01-22     1      0           0            0
## 3 2020-01-23     1      0           0            0
## 4 2020-01-24     2      0           1            0
## 5 2020-01-25     3      0           1            0
## 6 2020-01-26     5      0           2            0
```

```
head(usa_state_data)
```

```
## # A tibble: 6 x 6
## # Groups:   state [3]
##   date       state      cases deaths daily.cases daily.deaths
##   <date>     <chr>      <dbl>  <dbl>       <dbl>        <dbl>
## 1 2020-01-21 Washington     1      0           0            0
## 2 2020-01-22 Washington     1      0           0            0
## 3 2020-01-23 Washington     1      0           0            0
## 4 2020-01-24 Illinois       1      0           0            0
## 5 2020-01-24 Washington     1      0           0            0
## 6 2020-01-25 California     1      0           0            0
```

Total No. of cases (Statewise - US):

```
# Most recent data entry:
max(usa_state_data$date)
```

```
## [1] "2021-05-07"
```

```
library(data.table)
usa_state_data_total <- setDT(usa_state_data)[ ,.SD[which.max(as.Date(date, format= "%y-%m-%d"))], by =
```

```
head(usa_state_data_total,5)
```

Top 5 states based on total number of cases:

```
##          state   cases deaths daily.cases daily.deaths
## 1: California 3755647  62210        2222           45
## 2:      Texas 2912023  50748        2930           58
## 3:    Florida 2262590  35634        4165           86
## 4:   New York 2065172  52077        2465           39
## 5:   Illinois 1355300  24524        3160           41
```

```r
usa_state_data_total %>% arrange(desc(deaths)) %>% head(5)
```

Top 5 states based on total number of deaths:

```
##           state   cases deaths daily.cases daily.deaths
## 1:   California 3755647  62210        2222           45
## 2:     New York 2065172  52077        2465           39
## 3:        Texas 2912023  50748        2930           58
## 4:      Florida 2262590  35634        4165           86
## 5: Pennsylvania 1174510  26547        2647           50
```

```r
usa_state_data %>% arrange(desc(daily.cases)) %>% select(state,daily.cases) %>% head(5)
```

Top 5 highest daily increases in cases in US states:

```
##         state daily.cases
## 1: California       64987
## 2: California       60941
## 3:      Texas       58256
## 4: California       52197
## 5: New Jersey       51092
```

```r
usa_state_data %>% arrange(desc(daily.deaths)) %>% select(state,daily.deaths) %>% head(5)
```

Top 5 highest daily increases in deaths in US states:

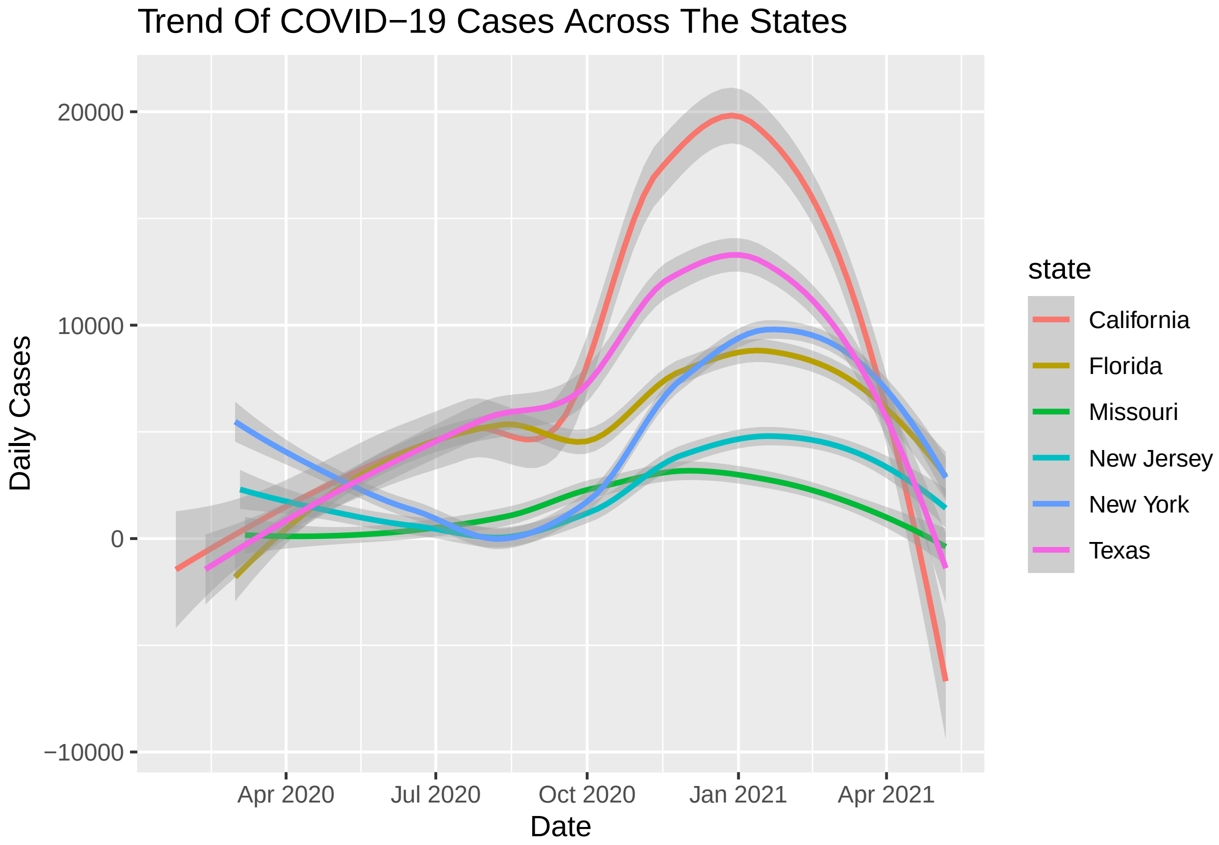
```
##         state daily.deaths
## 1:       Ohio         2559
## 2: New Jersey         1877
## 3:   Oklahoma         1716
## 4:    Indiana         1546
## 5:       Ohio         1204
```

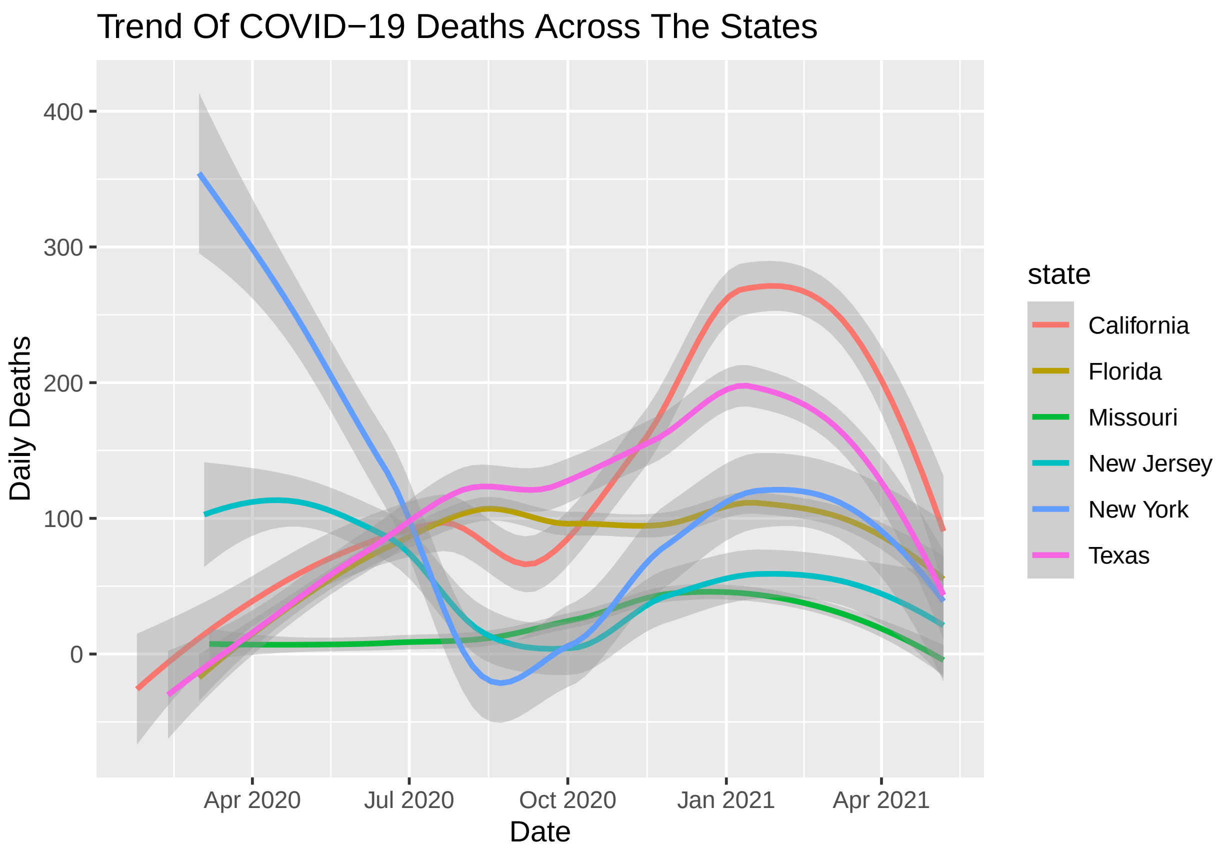
- 3 out of top 5 highest daily increase in cases: California
- 2 out of top 5 highest daily increase in deaths: Ohio

Now, lets look at top 5 states with each of their highest daily increases in cases and deaths:

```
## # A tibble: 5 x 2
## # Groups:   state [5]
##   state      daily.cases
##   <chr>            <dbl>
## 1 California       64987
## 2 Texas            58256
## 3 New Jersey       51092
## 4 Missouri         46263
## 5 Florida          31518
## # A tibble: 5 x 2
## # Groups:   state [5]
##   state      daily.deaths
##   <chr>             <dbl>
## 1 Ohio               2559
## 2 New Jersey         1877
## 3 Oklahoma           1716
## 4 Indiana            1546
## 5 Texas              1202
```

Let's plot the number of cases and deaths of these 5 states to get a better a idea of their situation.





- We can see very clearly that in states like New York, New Jersey, the number of cases and deaths have peaked at the start of the pandemic.
- But all the states went on a downward trend (after the first lockdown) for a while and then increased (Second Wave) with California witnessing the maximum peak.
- Texas & New Jersey didn't go through much of a downward trend until recently.

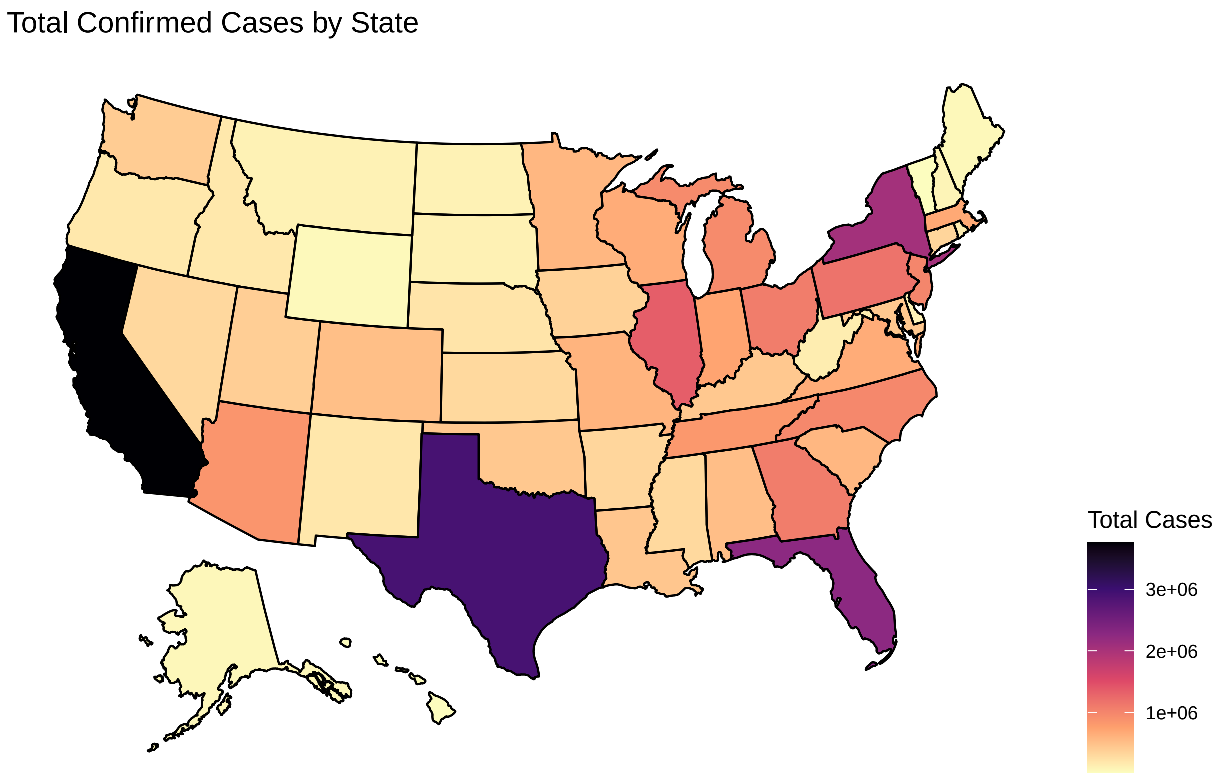
```
suppressPackageStartupMessages(library(usmap))
```

```
## Warning: package 'usmap' was built under R version 4.0.5
```

```
# Loading the US map:
state_map <- us_map(regions = "states")
county_map <- us_map(regions = "counties")
```

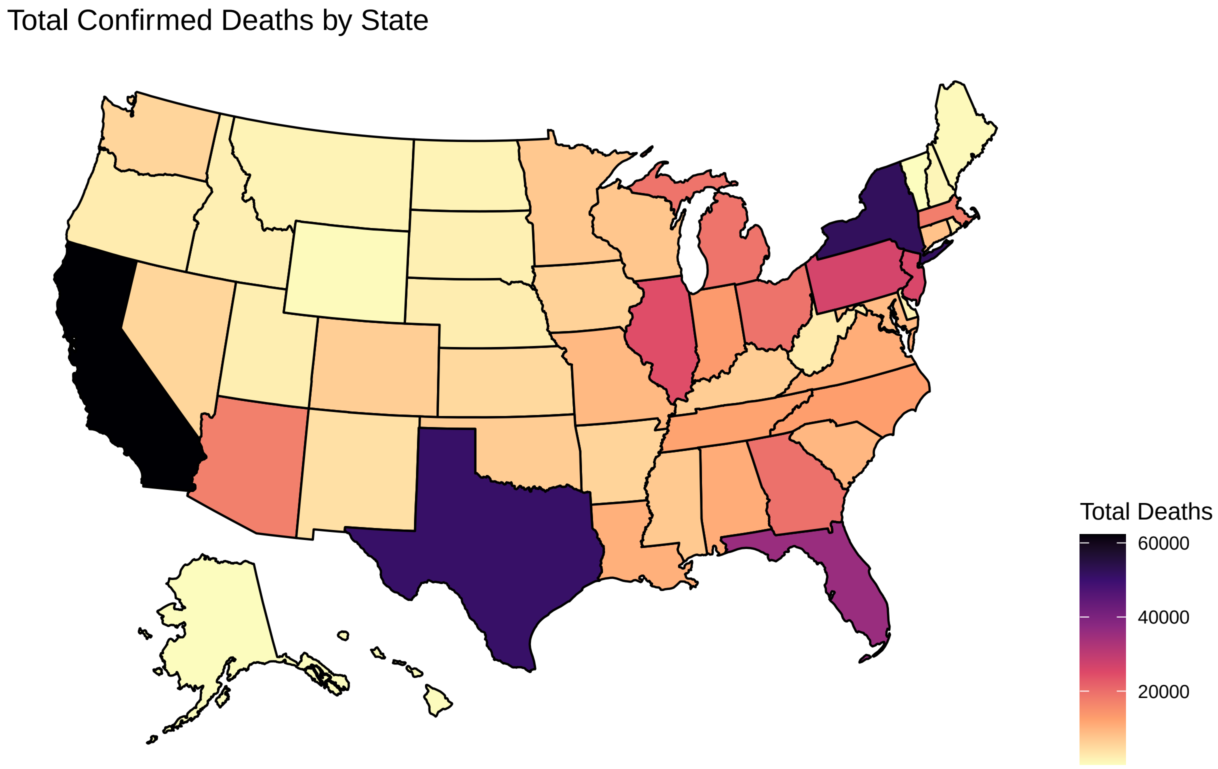
```
colnames(statepop)[colnames(statepop)=="full"] <- "state"

# Summarizing the data for map:
# For Cases:

usa_state_data %>%
  select(state, daily.deaths, daily.cases) %>%
  group_by(state) %>%
  summarize(TOTAL_CASES = sum(daily.cases)) -> CASES
# For Deaths:
```

```
usa_state_data %>%
  select(state, daily.deaths, daily.cases) %>%
  group_by(state) %>%
  summarize(TOTAL_DEATHS = sum(daily.deaths)) -> DEATHS


# Converting the sumarized data to data frames:
CASES <- data.frame(CASES)
DEATHS <- data.frame(DEATHS)

# Merging the data:
ALL_CASES <- left_join(CASES, statepop, by="state")
ALL_DEATHS <- left_join(DEATHS, statepop,  by="state")

# Map showing the distribution of COVID-19 cases across the US :
plot_usmap(data= ALL_CASES, values="TOTAL_CASES", regions = "state") +
  scale_fill_viridis_c(option = "A",direction=-1) +
  theme(legend.position = "right") +
  labs(fill="Total Cases") +
  ggtitle("Total Confirmed Cases by State")
```



```
plot_usmap(data=ALL_DEATHS, values="TOTAL_DEATHS", regions = "state") +
  scale_fill_viridis_c(option = "A",direction=-1) +
  theme(legend.position = "right") +
  labs(fill="Total Deaths") +
```

```r
ggtitle("Total Confirmed Deaths by State")
```



Further Research:

- I would like to include population of each state to check the most infected states.
- Also, since this is a time-series data, a animated plot which shows the varied number of cases or deaths along with time might be a better visualization plot. So, I would like to build such plot.

Thank You!