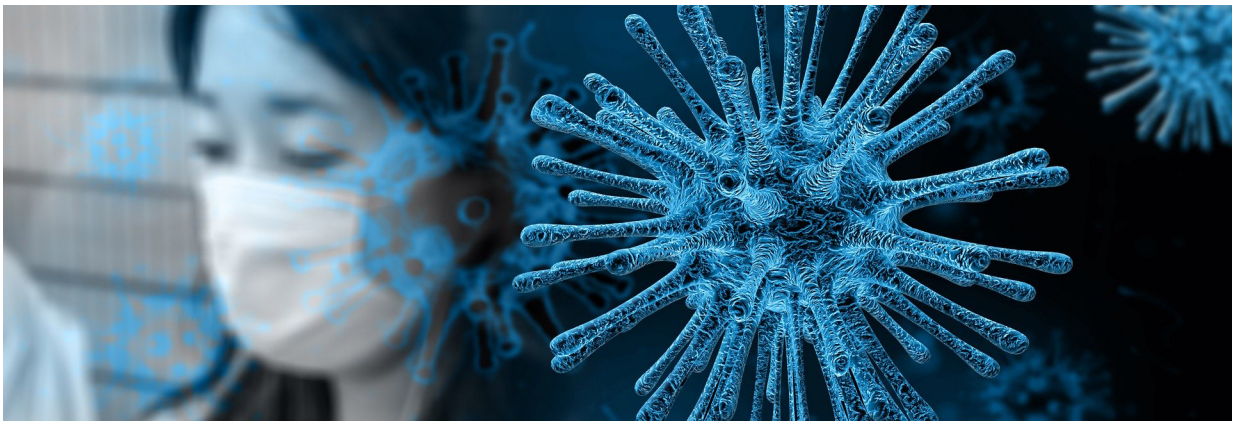
Final_Project

PRANJAL SRIVASTAVA(Net ID: ps4379)

1/22/2022

Synopsis



Coronavirus (COVID-19) is a highly contagious disease caused by the SARS-CoV-2 virus. The virus can be spread from an infected person's mouth or nose through small fluid particles when the person coughs, sneezes, talks, sings or breathes (1). Most people infected with the virus will experience mild to moderate respiratory illness and recover without requiring special treatment. However, some will become seriously ill and require medical attention. .

In step with Center for disease Control and Prevention (CDC) record, the primary community transmission of COVID-19 become detected in February, 2020 inside the united states(2). Since then the geographic version within the US nation is remarkable. This difference will be due to epidemiological and population degree versions. Well timed monitoring of COVID-19 instances and deaths is critical to understand the magnitude of the disease and to design records pushed techniques.

Detail information about the COVID-19 can be accessed from CDC website using this *link*

Question

In order to understand the current status of COVID-19 within the US, I will analyse the top 10 states with highest number of cases at the lastest date so that more resources and help can be applied towards these states in response to current situation. I will also analyse the mortality rate of various states in the United States to find out what mismanagement or ill administeration might have caused such a bizzare.

Packages Required

I. Install packages

To work through this data, we'll need to Install various packages and load them

```
## install and load the necessary libraries
#install.packages("tidyverse")
#install.packages("readr")
#install.packages("ggplot2")
#install.packages("janitor")
#install.packages("dplyr")
#install.packages("tidyr")
#install.packages("lubridate")
#intsall.packages("sqldf")
```

II. Load library

```
library(tidyverse)    ## To easily install and load the 'Tidyverse'
library(readr)        ## To easily read rectangular data.
library(tidyr)        ## To create tidy data
library(ggplot2)      ## For mapping and ploting the data
library(lubridate)    ## Dates and times made easy with lubridate
library(janitor)      ## For cleaning and examining data
library(dplyr)        ## Data Manipulation
library(sqldf)        ## To perform SQL selects on R data frames.
library(r02pro)
library(usmap)        ## To plot US State maps
```

Data Preparation

The most important part of this project is to import and clean the data as needed.The dataset contains the US COVID-19 data till 01/11/2022. The data is originally taken from The New York Times github repository: https://github.com/nytimes/covid-19-data

I. Importing data and creating the dataframe

First I set the working directory as I have downloaded the 'US_counties.csv' data already in my folder from Brightspace(University's academic platform)

```
setwd("/Users/pranjalsrivastava/Desktop/Intermediate_R/Final_Project")
```

After setting the working directory, we can load the data in 'csv' file.

```
us_counties_raw <- read_csv("us-counties.csv", show_col_types = FALSE)
```

II. Examining the dataframe

We'll use the head(), str(), and colnames() to find the details for the data frames.

```
head(us_counties_raw)
```

```
## # A tibble: 6 x 6
##   date       county    state      fips  cases deaths
##   <date>     <chr>     <chr>      <chr> <dbl>  <dbl>
## 1 2020-01-21 Snohomish Washington 53061     1      0
## 2 2020-01-22 Snohomish Washington 53061     1      0
## 3 2020-01-23 Snohomish Washington 53061     1      0
## 4 2020-01-24 Cook      Illinois   17031     1      0
## 5 2020-01-24 Snohomish Washington 53061     1      0
## 6 2020-01-25 Orange    California 06059     1      0
```

We could clearly note the types of every column, which are 1 date, 3 Characters, and 2 Double.

```
str(us_counties_raw)
```

```
## spec_tbl_df [2,105,876 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ date  : Date[1:2105876], format: "2020-01-21" "2020-01-22" ...
##  $ county: chr [1:2105876] "Snohomish" "Snohomish" "Snohomish" "Cook" ...
##  $ state : chr [1:2105876] "Washington" "Washington" "Washington" "Illinois" ...
##  $ fips  : chr [1:2105876] "53061" "53061" "53061" "17031" ...
##  $ cases : num [1:2105876] 1 1 1 1 1 1 1 1 1 1 ...
##  $ deaths: num [1:2105876] 0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "spec")=
```

```
##   .. cols(
##   ..   date = col_date(format = ""),
##   ..   county = col_character(),
##   ..   state = col_character(),
##   ..   fips = col_character(),
##   ..   cases = col_double(),
##   ..   deaths = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

There are 6 columns and 2,105,876 rows in the data set.The variables in the data are date, county, state, fips, cases and deaths. The columns Deaths and Cases have cumulative values according to the Date column.

III. Creating Tidy Data

Once the Raw basic data has been loaded, we need to pre-process it to get the data into a tidy format.

To continue the same, we first need to find out if there are any missing values and duplicate columns or rows.

```
# checking duplicates
us_counties_raw[duplicated(us_counties_raw),]
```

```
## # A tibble: 0 x 6
## # ... with 6 variables: date <date>, county <chr>, state <chr>, fips <chr>,
## #   cases <dbl>, deaths <dbl>
```

```
## There are no duplicates found in the data

# Checking missing data
colSums(is.na(us_counties_raw))
```

```
##   date county  state   fips  cases deaths
##      0      0      0  19437      0  48089
```

```
## There are 19,437 and 48089 missing value in fips and deaths respectively.

death_na_v1 <- sqldf("select distinct state from us_counties_raw where deaths is null")
str(death_na_v1) ## To know which state has missing value in deaths.
```

```
## 'data.frame':    1 obs. of  1 variable:
##  $ state: chr "Puerto Rico"
```

Now, when we know that the data has missing values and that the vaiable 'Fips'is not usefull for the project, we will remove "FIPS" from the data but will keep value NA as upon further analysis of data, we came to know that Deaths variable has missing values corresponding only to state "Puerto Rico". Removing rows with NA values will remove the whole data for state "Puerto Rico" which will effect our results. Rather I'll drop Puerto Rico when needed manually.

```
us_counties <- us_counties_raw %>%
     ##dropping the variable 'FIPS'
   select(-fips) %>%
   ## to make sure all the column names are inline with naming convention
      clean_names()
```

```
str(us_counties)
```

```
## tibble [2,105,876 x 5] (S3: tbl_df/tbl/data.frame)
##  $ date  : Date[1:2105876], format: "2020-01-21" "2020-01-22" ...
##  $ county: chr [1:2105876] "Snohomish" "Snohomish" "Snohomish" "Cook" ...
##  $ state : chr [1:2105876] "Washington" "Washington" "Washington" "Illinois" ...
##  $ cases : num [1:2105876] 1 1 1 1 1 1 1 1 1 1 ...
##  $ deaths: num [1:2105876] 0 0 0 0 0 0 0 0 0 0 ...
```

Now as the data is so big, we will first look for the date with the highest case recorded.

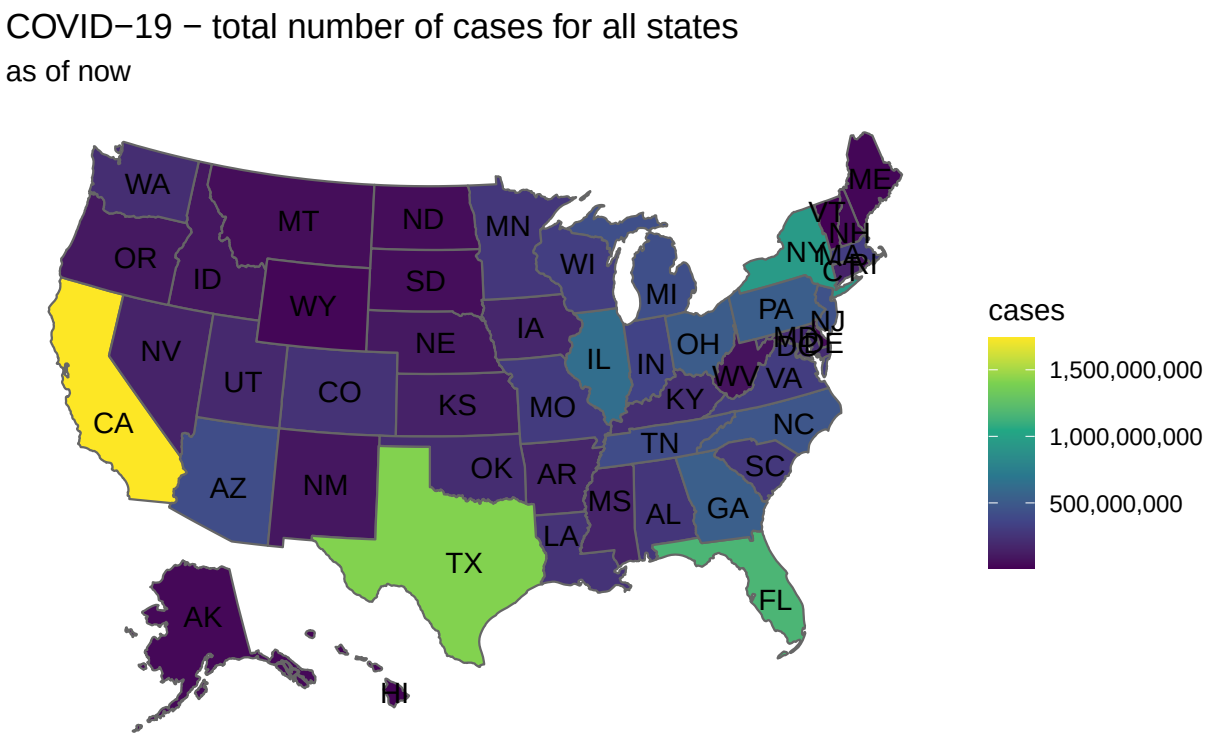
Exploratory Data Analysis

I. Exploring Data

Let us first see the scattering of total cases over the different states of USA.

```
us_counties_top <- us_counties %>%
  group_by(state) %>%
  summarize(cases_total_main = sum(cases),
            deaths_total_main = sum(deaths),
            average_cases=mean(cases),
            average_deaths=mean(deaths)) %>%
  arrange(desc(cases_total_main))


us_counties_top$state <- factor(us_counties_top$state,
            levels = us_counties_top$state[
              order(us_counties_top$cases_total_main)])

plot_usmap(data = us_counties_top,
           values = "cases_total_main",
           color = "grey40", labels = TRUE) +
  #change the gradient type, add comma on legend
scale_fill_continuous(type='viridis', label = scales:: comma) +
  #add title, subtitle, caption, and legend title
```

```
  labs(title = "COVID-19 - total number of cases for all states",
       subtitle = "as of now",
       fill = "cases") +
    theme_classic()+   #set the theme to classic (my go-to theme!)
      theme(panel.background = element_blank(),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.position = "right",
        axis.line = element_blank(),
        axis.ticks = element_blank(),
        axis.text.x =element_blank(),
        axis.text.y = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank())
```



Now let us filter the the top 10 states with the highest number of cases on the last date of data.

```
max(tail(us_counties$date))  ## to find out the latest date of the data
```

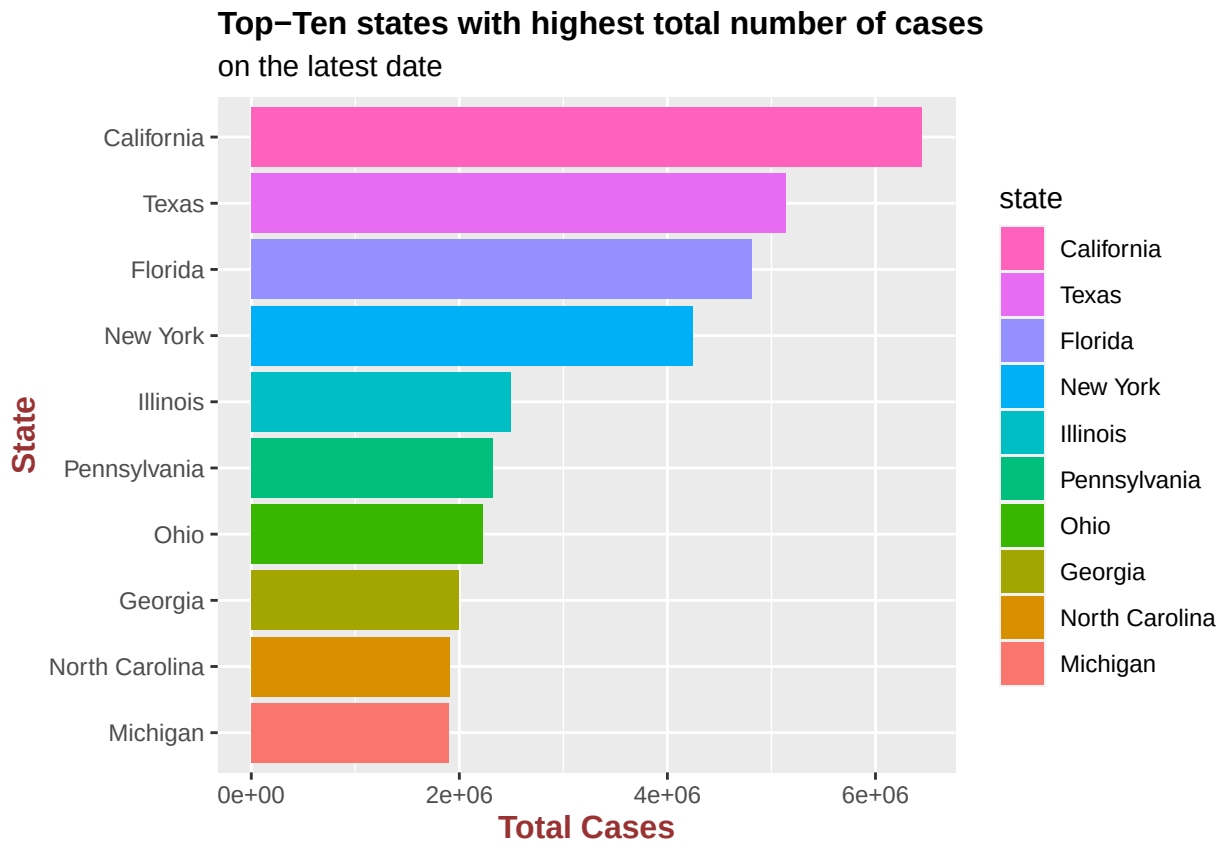
```
## [1] "2022-01-11"
```

```r
us_latest <- us_counties %>% filter(date == "2022-01-11")
## The new data frame for the latest date is ready

us_cases_top10 <- us_latest %>%
  group_by(state) %>%
### making new variable Cases_total_main with the highest number of cases for top sates.
  summarize(cases_total_main = sum(cases), average_cases=mean(cases)) %>%
  arrange(desc(cases_total_main)) %>%
  head(10)
us_cases_top10
```

```
## # A tibble: 10 x 3
##    state          cases_total_main average_cases
##    <chr>                     <dbl>         <dbl>
##  1 California              6447667       111167.
##  2 Texas                   5141662        20243.
##  3 Florida                 4806782        70688.
##  4 New York                4244155        73175.
##  5 Illinois                2492045        24195.
##  6 Pennsylvania            2321087        34643.
##  7 Ohio                    2226881        25021.
##  8 Georgia                 1999095        12494.
##  9 North Carolina          1904215        19042.
## 10 Michigan                1899802        22617.
```

Now that we have all top ten states with highest number of caes on the last date, we will go on to analyse these sates with the help of Bar garph.

```r
us_cases_top10$state <- factor(us_cases_top10$state,
          levels = us_cases_top10$state[
            order(us_cases_top10$cases_total_main)])

ggplot(us_cases_top10, aes(cases_total_main,state,fill = state)) +
  geom_bar(stat = "identity") +
  labs(title = "Top-Ten states with highest total number of cases",
       subtitle = "on the latest date",
       x = "Total Cases",
       y = "State") +
  theme(
    plot.title = element_text(color = "black", size = 12, face= "bold"),
    axis.title.x = element_text(color="#993333", size=12, face="bold"),
    axis.title.y = element_text(color="#993333", size=12, face="bold")) +
  guides(fill = guide_legend(reverse = TRUE))
```

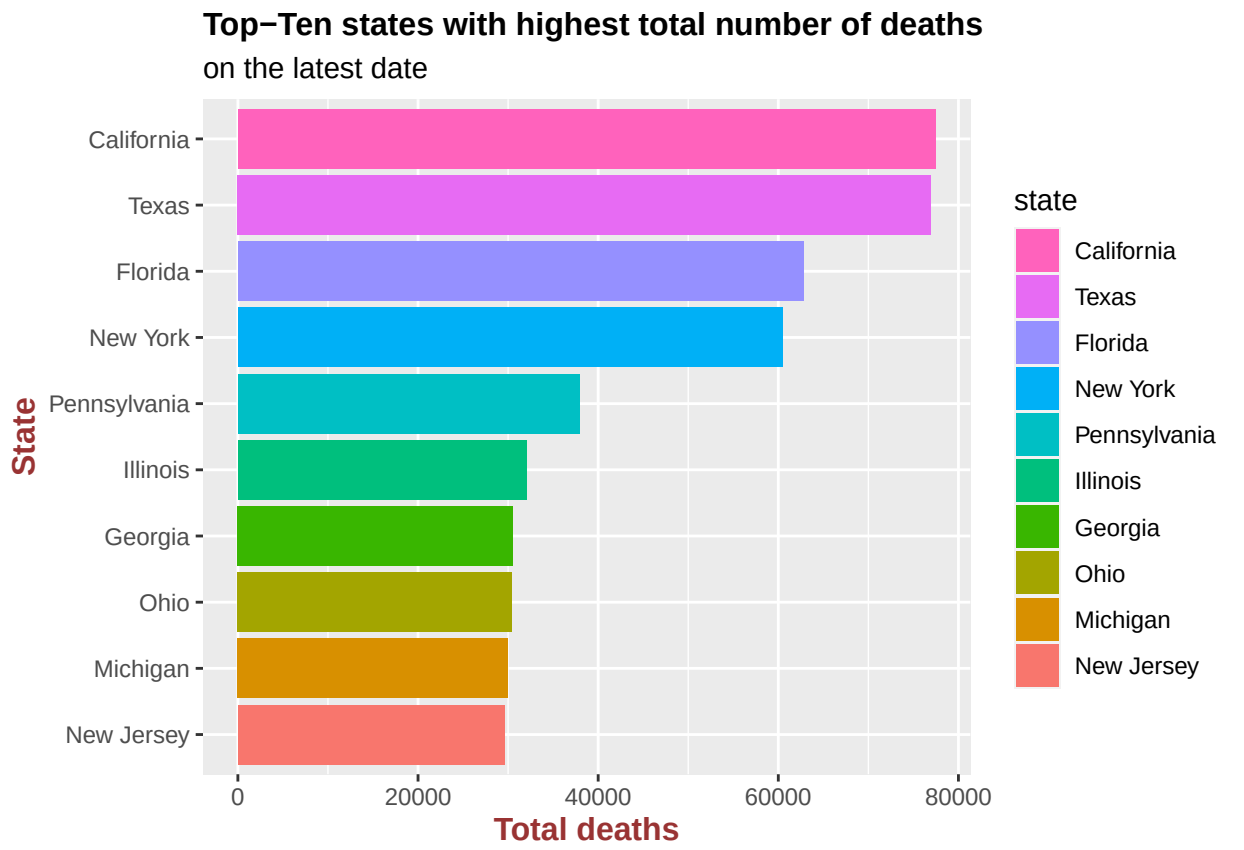


The highest total number of cases were found in California followed by Texas and Florida

let us find the state with highest total death recorded till the date when the USA had the highest case recorded in a single day in Florida.

```
us_deaths_top10 <- us_latest %>%
  group_by(state) %>%
  summarize(deaths_total=sum(deaths),
            average_cases=mean(cases),
            average_deaths=mean(deaths), na.rm = TRUE) %>%
  arrange(desc(deaths_total)) %>%
  head(10)


us_deaths_top10$state <- factor(us_deaths_top10$state,
                  levels = us_deaths_top10$state[order(us_deaths_top10$deaths_total)])

ggplot(us_deaths_top10, aes(deaths_total,state,fill=state)) +
  geom_bar(stat = "identity")+
  labs(title = "Top-Ten states with highest total number of deaths",
       subtitle = "on the latest date",
       x = "Total deaths",
       y = "State",)+
  theme(
```

```
    plot.title = element_text(color="black", size=12, face="bold"),
    axis.title.x = element_text(color="#993333", size=12, face="bold"),
    axis.title.y = element_text(color="#993333", size=12, face="bold")
  ) +
  guides(fill = guide_legend(reverse = TRUE))
```



So from the above Graph it is clear that California also had the highest total number of death at the last date.

```
# To find total number of states in the data

n_distinct(us_counties$state)
```

```
## [1] 56
```

```
# To find the names of the states
unique(us_counties['state'])
```

```
## # A tibble: 56 x 1
##    state
##    <chr>
##  1 Washington
##  2 Illinois
##  3 California
##  4 Arizona
```

```
##  5 Massachusetts
##  6 Wisconsin
##  7 Texas
##  8 Nebraska
##  9 Utah
## 10 Oregon
## # ... with 46 more rows
```

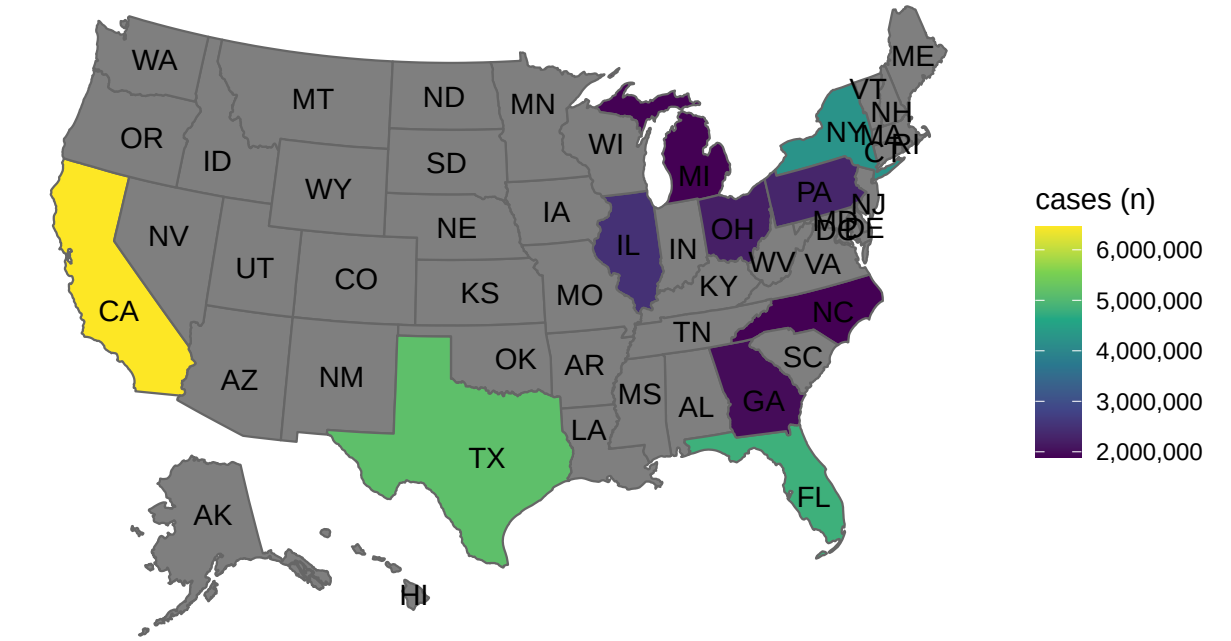
```
#To find the number of counties involved
n_distinct(us_counties$county)
```

```
## [1] 1930
```

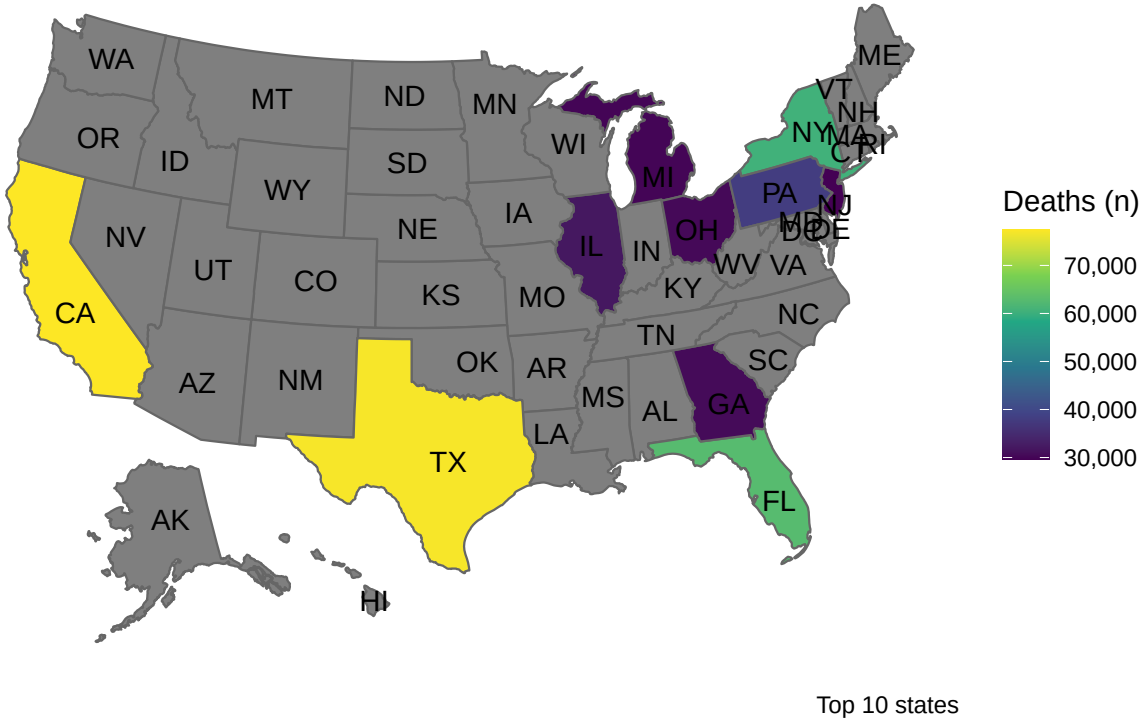
There are 56 states and 1930 Counties we'll be studying on.

I am going to depict the highest total number of cases for top 10 states on date "2022-01-11" using US-Map plotting.

```
#line color = grey
usmap::plot_usmap(data = us_cases_top10,
                  values = "cases_total_main",
                  color = "grey40", labels = TRUE) +

#change the gradient type, add comma on legend
  scale_fill_continuous(type='viridis', label = scales:: comma) +

#add title, subtitle, caption, and legend title
  labs(title = "COVID-19 - total number of cases in top 10 states",
       subtitle = "on 2022-01-11",
       fill = "cases (n)") +

       theme_classic()+
  theme(panel.background = element_blank(),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.position = "right",
        axis.line = element_blank(),
        axis.ticks = element_blank(),
        axis.text.x =element_blank(),
        axis.text.y = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank())
```




```
#line color = grey
usmap::plot_usmap(data = us_deaths_top10,
                  values = "deaths_total",
                  color = "grey40", labels = TRUE) +

#change the gradient type, add comma on legend
  scale_fill_continuous(type='viridis', label = scales:: comma) +

#add title, subtitle, caption, and legend title
  labs(title = "COVID-19 - total number of deaths in top 10 states",
       subtitle = "on 2022-01-11",
       caption = "Top 10 states",
       fill = "Deaths (n)") +
       ## removing all the axes from the plotting.
       theme_classic()+
  theme(panel.background = element_blank(),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.position = "right",
        axis.line = element_blank(),
        axis.ticks = element_blank(),
        axis.text.x =element_blank(),
        axis.text.y = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank())
```



Seeing the USmap Plotting for both cases and deaths, it is clear that even when the Texas has lower number of cases as compared to California, it has very high number of deaths approximating to California in comaprison to other states.

Now that we are clear on the part that which are the top 10 states in the USA with the highest number of total and average number of cases and deaths and as we have clearly analysis these data through visualizations, lets move on to find out the worst mortality rates in the USA at the last day of the data and which are the top 10 states to have highest mortality rate.

To continue, we'll have to first calculate the mortality rate(Deaths per Case) for Top 10 states.But we know that the variable 'Deaths' has some missing values, we'll set those values to zero an will move further.Also for the latest dates some cases are 0 which will return mortality rate as inf which we will remove to process further.

```
## Assigning 0 to all missing value of variable "Death"
us_latest[is.na(us_latest)] = 0

## Creating a variable "Mortality rate".
us_latest$Mortality_rate <- (us_latest$deaths/us_latest$cases)

## Removing Infinite values
us_latest$Mortality_rate[!is.finite(us_latest$Mortality_rate)] <- 0

Highest_mort10 <- us_latest %>% arrange(desc(Mortality_rate)) %>%
  head(10)
```

```
Highest_mort10
```

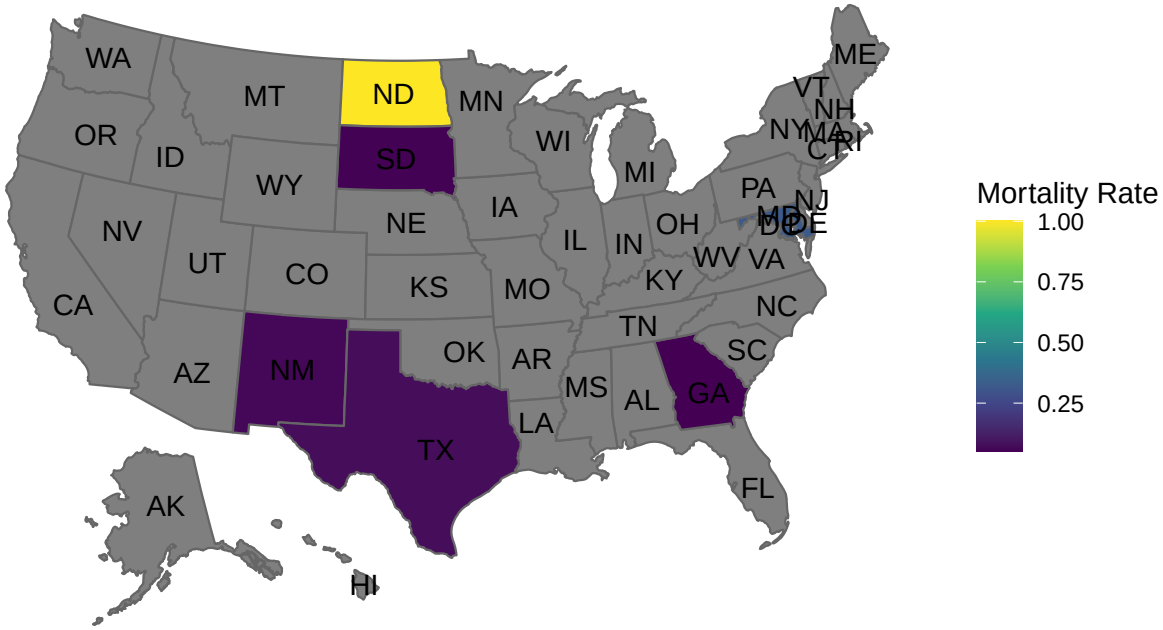
```
## # A tibble: 10 x 6
##    date       county   state        cases deaths Mortality_rate
##    <date>     <chr>    <chr>        <dbl>  <dbl>          <dbl>
##  1 2022-01-11 Unknown  North Dakota    45     45         1
##  2 2022-01-11 Unknown  Maryland       856    278         0.325
##  3 2022-01-11 Unknown  Puerto Rico  12551   3393         0.270
##  4 2022-01-11 Sabine   Texas         1040     79         0.0760
##  5 2022-01-11 Harding  New Mexico      60      4         0.0667
##  6 2022-01-11 McMullen Texas          130      8         0.0615
##  7 2022-01-11 Hancock  Georgia       1364     83         0.0609
##  8 2022-01-11 Jerauld  South Dakota   344     19         0.0552
##  9 2022-01-11 Knox     Texas          408     22         0.0539
## 10 2022-01-11 Candler  Georgia       1713     91         0.0531
```

We now have top 10 states with highest mortality rate on date "2022-01-11" . North Dakota being the state with the highest Mortality rate among others followed by Maryland and Puerto Rico.

```
#line color = grey
usmap::plot_usmap(data = Highest_mort10,
                  values = "Mortality_rate",
                  color = "grey40", labels = TRUE) +

#change the gradient type, add comma on legend
  scale_fill_continuous(type='viridis', label = scales:: comma) +

#add title, subtitle, caption, and legend title
  labs(title = "Mortality Rate for Top 10 States",
       subtitle = "on 2022-01-11",
       fill = "Mortality Rate") +

   ## removing all the axes from the plotting.
       theme_classic()+
  theme(panel.background = element_blank(),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.position = "right",
        axis.line = element_blank(),
        axis.ticks = element_blank(),
        axis.text.x =element_blank(),
        axis.text.y = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank())
```

Mortality Rate for Top 10 States

on 2022−01−11



Summary

Consequently, our analysis found that California, Texas and Florida had the highest total number of cases and deaths in the USA on the last date of the data. Others among the top 10 being New York, Illinois, Pennsylvania, Ohio, Georgia, North Carolina Michigan. So we can clearly say that these states acheived the Peak during the latest period. Also, the data analysis also showed us that the Top 10 states with the highest number of deaths were California, Texas, Florida, New York, Pennsylvania, Illinois, Georgia, Ohio, Michigan, New Jersey. So, clearly this list varied than the Top 10 states for cases list.

Secondly, we also saw that data for average number of cases and deaths for top 10 states varied from each other. As, the average number of deaths in Texas were closer to those in California but average number of cases in California were higher than those in Texas.

Lastly, we learned some facts about Mortality rate. North Dakota became the top state with highest mortality rate followed by "North Dakota", "Maryland", "Puerto Rico", "Texas", "New Mexico", "Texas", "Georgia", "South Dakota", "Texas", "Georgia" . This analysis depicts that these states failed to handle the Covid 19 cases and that mishandling and mismanagement of cases led to such a high number of deaths per cases.

This Exploratory data analysis of the data "US_Counties" for Covid 19 cases helped us understanding the bad situation of some states and their management, so that the organisations can improve the situation for future cases if any and the pandemic.

Reference

1.https://www.who.int/health-topics/coronavirus#tab=tab_1. Date accessed(01-26-2021)

2.https://www.cdc.gov/mmwr/volumes/69/wr/mm6915e4.htm. Date accessed(01-27-2021)