US COVID-19 Cases,Trend and Heat Map Analysis by states

Chitra G

2024-05-29

```
knitr::opts_chunk$set(echo = TRUE)
#install all the below packages using install.package(pakage-name)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(readr)
library(lubridate)
library(ggplot2)
library(dplyr)
library(viridis)
```

```
## Loading required package: viridisLite
```

Source of COVID Data

The COVID-19 dataset used in this analysis is sourced from the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) repository. This repository provides comprehensive and up-to-date information on COVID-19 cases and deaths globally and in the United States.

For this analysis, we specifically used the following CSV files from the JHU CSSE GitHub repository: - "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 to the COVID-19 data files
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi
file_names <- c("time_series_covid19_confirmed_global.csv",
                "time_series_covid19_deaths_global.csv",
                "time_series_covid19_confirmed_US.csv",
                "time_series_covid19_deaths_US.csv")

urls <- str_c(url_in, file_names)
```

```r
# Read the data files
global_cases <- read_csv(urls[1])
```

```
## Rows: 289 Columns: 1147
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1145): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20,...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
global_deaths <- read_csv(urls[2])
```

```
## Rows: 289 Columns: 1147
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1145): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20,...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
US_cases <- read_csv(urls[3])
```

```
## Rows: 3342 Columns: 1154
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1148): UID, code3, FIPS, Lat, Long_, 1/22/20, 1/23/20, 1/24/20, 1/25/20...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
US_deaths <- read_csv(urls[4])
```

```
## Rows: 3342 Columns: 1155
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1149): UID, code3, FIPS, Lat, Long_, Population, 1/22/20, 1/23/20, 1/24...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
# View the structure of the US cases data
#str(US_cases)
#str(US_deaths)

# Data cleaning and transformation for US cases
```

```r
US_cases_long <- US_cases %>%
  pivot_longer(cols = -c(UID, iso2, iso3, code3, FIPS, Admin2, Province_State, Country_Region, Lat, Long
               names_to = "date",
               values_to = "cases") %>%
  mutate(date = mdy(date))

# Data cleaning and transformation for US deaths
US_deaths_long <- US_deaths %>%
  pivot_longer(cols = -c(UID, iso2, iso3, code3, FIPS, Admin2, Province_State, Country_Region, Lat, Long
               names_to = "date",
               values_to = "deaths") %>%
  mutate(date = mdy(date))


# Combine cases and deaths data
US_data <- US_cases_long %>%
  left_join(US_deaths_long, by = c("UID", "iso2", "iso3", "code3", "FIPS", "Admin2", "Province_State",
str(US_data)
```

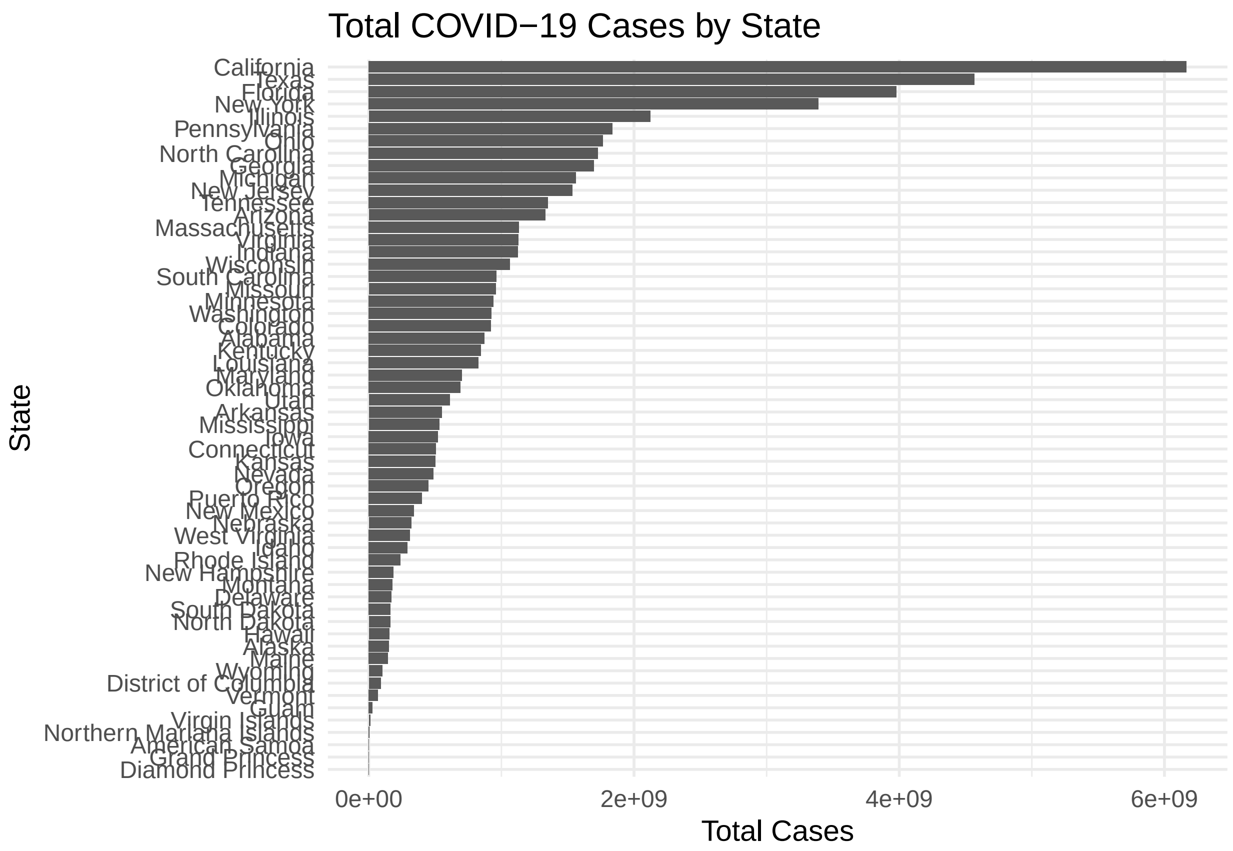
```
## tibble [3,819,906 x 15] (S3: tbl_df/tbl/data.frame)
##  $ UID           : num [1:3819906] 8.4e+07 8.4e+07 8.4e+07 8.4e+07 8.4e+07 ...
##  $ iso2          : chr [1:3819906] "US" "US" "US" "US" ...
##  $ iso3          : chr [1:3819906] "USA" "USA" "USA" "USA" ...
##  $ code3         : num [1:3819906] 840 840 840 840 840 840 840 840 840 840 ...
##  $ FIPS          : num [1:3819906] 1001 1001 1001 1001 1001 ...
##  $ Admin2        : chr [1:3819906] "Autauga" "Autauga" "Autauga" "Autauga" ...
##  $ Province_State: chr [1:3819906] "Alabama" "Alabama" "Alabama" "Alabama" ...
##  $ Country_Region: chr [1:3819906] "US" "US" "US" "US" ...
##  $ Lat           : num [1:3819906] 32.5 32.5 32.5 32.5 32.5 ...
##  $ Long_         : num [1:3819906] -86.6 -86.6 -86.6 -86.6 -86.6 ...
##  $ Combined_Key  : chr [1:3819906] "Autauga, Alabama, US" "Autauga, Alabama, US" "Autauga, Alabama, 
##  $ date          : Date[1:3819906], format: "2020-01-22" "2020-01-23" ...
##  $ cases         : num [1:3819906] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Population    : num [1:3819906] 55869 55869 55869 55869 55869 ...
##  $ deaths        : num [1:3819906] 0 0 0 0 0 0 0 0 0 0 ...
```

```r
# Identify states with the highest number of cases
US_state_cases <- US_data %>%
  group_by(Province_State) %>%
  summarize(total_cases = sum(cases, na.rm = TRUE), total_deaths = sum(deaths, na.rm = TRUE)) %>%
  arrange(desc(total_cases))

# View the states with the highest number of cases
head(US_state_cases)
```

```
## # A tibble: 6 x 3
##   Province_State total_cases total_deaths
##   <chr>                <dbl>        <dbl>
## 1 California      6166190335     65490302
## 2 Texas           4566537657     61302166
## 3 Florida         3978357707     51475342
## 4 New York        3392006819     58121236
## 5 Illinois        2122240785     28240376
## 6 Pennsylvania    1836846159     31912144
```

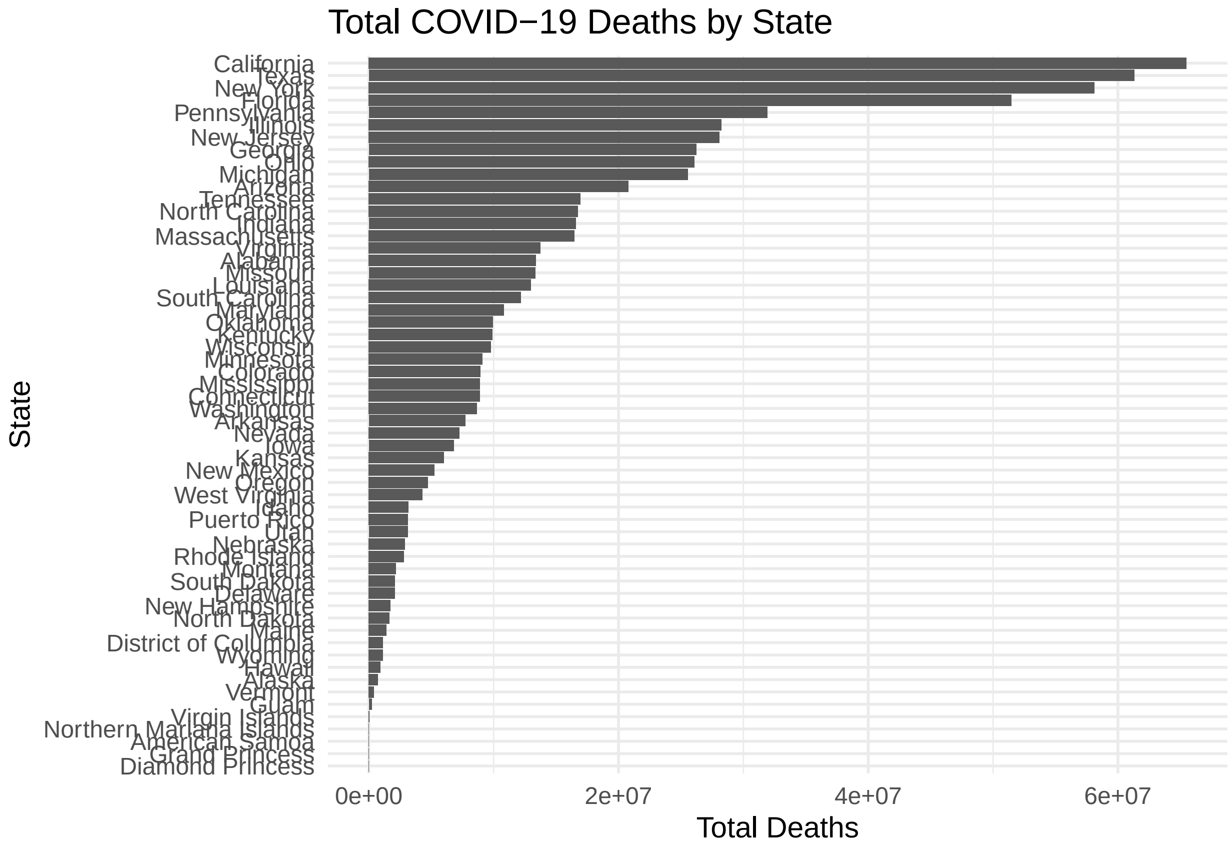
** Note California leads the covid cases Closely followed by Texas and Florida**

```r
# Visualize total cases by state
US_state_cases %>%
  ggplot(aes(x = reorder(`Province_State`, total_cases), y = total_cases)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Total COVID-19 Cases by State", x = "State", y = "Total Cases") +
  theme_minimal()
```

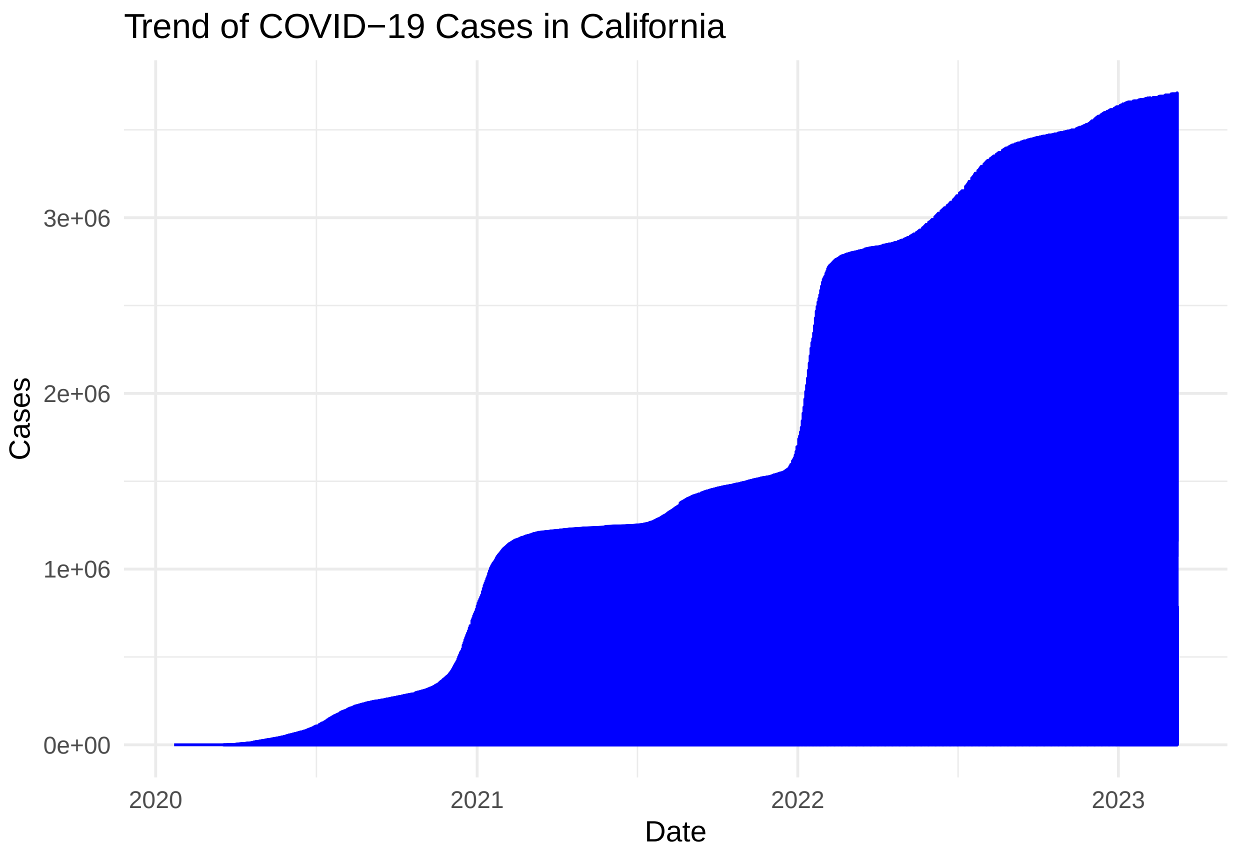


** Note California leads the Death cases Closely followed by Texas and New York The deaths are consistent with number of cases too.**

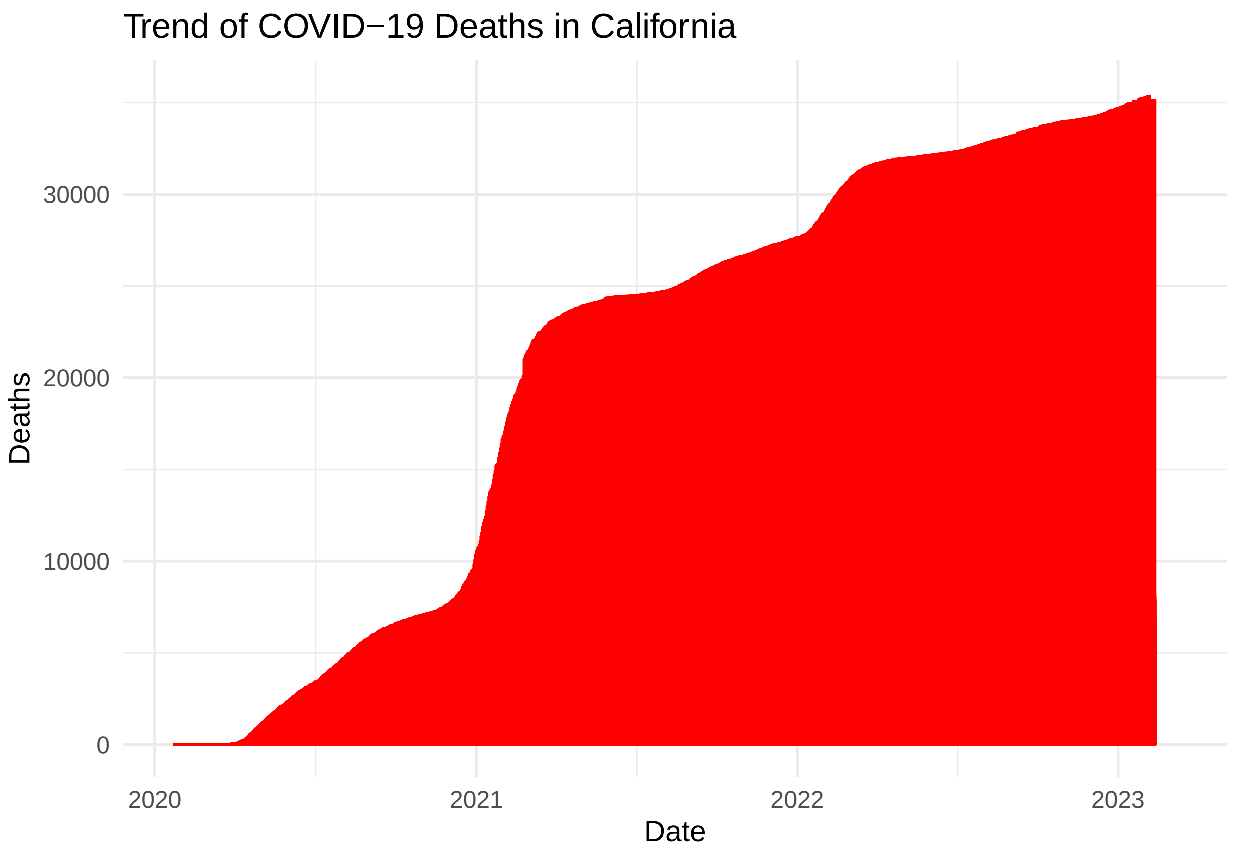
```r
# Visualize total deaths by state
US_state_cases %>%
  ggplot(aes(x = reorder(Province_State, total_deaths), y = total_deaths)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Total COVID-19 Deaths by State", x = "State", y = "Total Deaths") +
  theme_minimal()
```



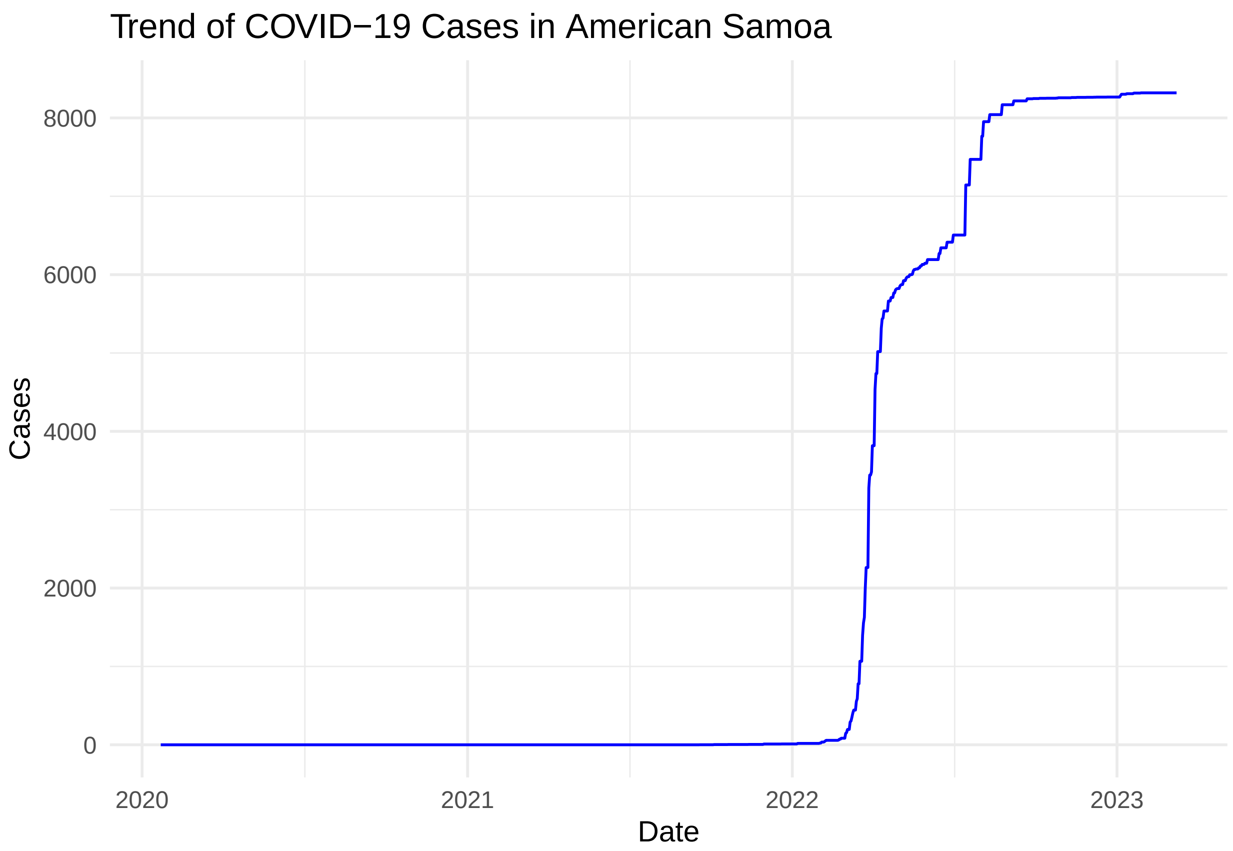
```r
# Trend analysis for COVID-19 cases and deaths over time

# Filter data for a specific state (e.g., California)
#Replace state_name with any state that you want to analyse
state_name <- "California"

state_data <- US_data %>%
  filter(Province_State == state_name)

# Plot trend of cases over time
ggplot(state_data, aes(x = date, y = cases)) +
  geom_line(color = "blue") +
  labs(title = paste("Trend of COVID-19 Cases in", state_name),
       x = "Date",
       y = "Cases") +
  theme_minimal()
```



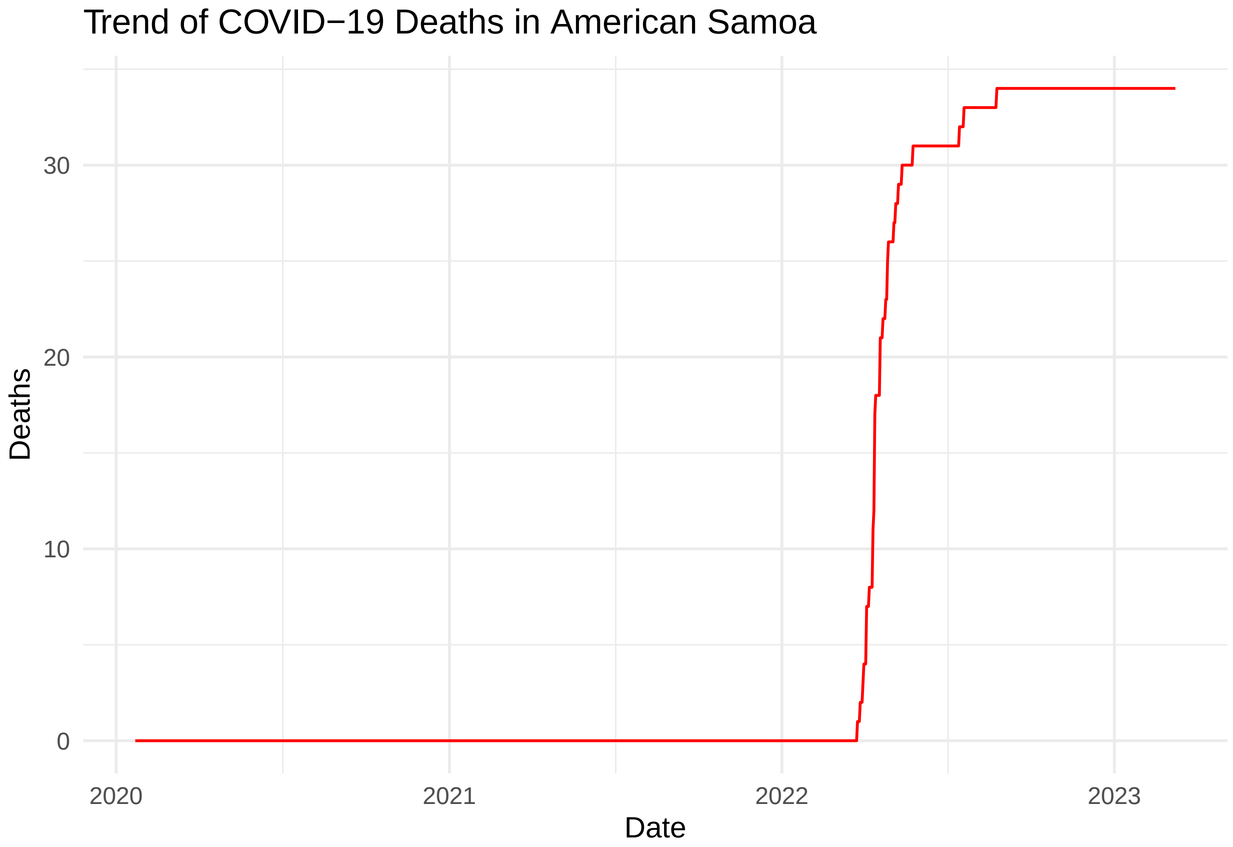
```r
# Plot trend of deaths over time
ggplot(state_data, aes(x = date, y = deaths)) +
  geom_line(color = "red") +
  labs(title = paste("Trend of COVID-19 Deaths in", state_name),
       x = "Date",
       y = "Deaths") +
  theme_minimal()
```



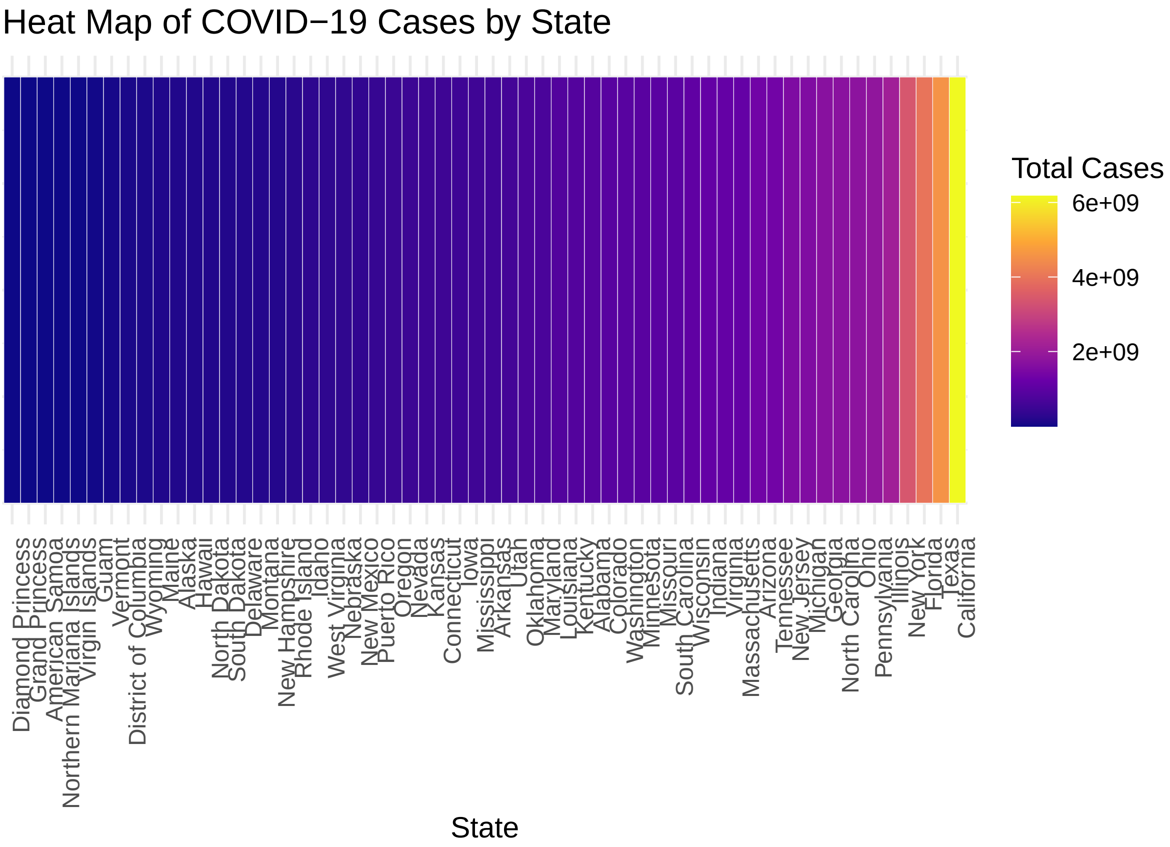
```r
state_name <- "American Samoa"

state_data <- US_data %>%
  filter(Province_State == state_name)

# Plot trend of cases over time
ggplot(state_data, aes(x = date, y = cases)) +
  geom_line(color = "blue") +
  labs(title = paste("Trend of COVID-19 Cases in", state_name),
       x = "Date",
       y = "Cases") +
  theme_minimal()
```



```r
# Plot trend of deaths over time
ggplot(state_data, aes(x = date, y = deaths)) +
  geom_line(color = "red") +
  labs(title = paste("Trend of COVID-19 Deaths in", state_name),
       x = "Date",
       y = "Deaths") +
  theme_minimal()
```



```
# Prepare data for heat map
state_cases <- US_data %>%
  group_by(Province_State) %>%
  summarise(total_cases = sum(cases, na.rm = TRUE)) %>%
  arrange(desc(total_cases))

# Create a data frame for the heat map
heat_map_data <- data.frame(state = state_cases$Province_State, total_cases = state_cases$total_cases)

# Plot heat map
ggplot(heat_map_data, aes(x = reorder(state, total_cases), y = 1, fill = total_cases)) +
  geom_tile(color = "white") +
  scale_fill_viridis_c(option = "C") +
  labs(title = "Heat Map of COVID-19 Cases by State",
       x = "State",
       y = "",
       fill = "Total Cases") +
  theme_minimal() +
  theme(axis.text.y = element_blank(),  # Remove y-axis labels
        axis.ticks.y = element_blank(), # Remove y-axis ticks
        axis.text.x = element_text(angle = 90, hjust = 1)) # Rotate x-axis labels for better readabilit
```



```r
# Summarize the data, This is the model
summary(US_data)
```

```
##       UID                iso2               iso3               code3      
##  Min.   :      16   Length:3819906     Length:3819906     Min.   : 16.0  
##  1st Qu.:84018105   Class :character   Class :character   1st Qu.:840.0  
##  Median :84029206   Mode  :character   Mode  :character   Median :840.0  
##  Mean   :83429923                                         Mean   :834.5  
##  3rd Qu.:84046119                                         3rd Qu.:840.0  
##  Max.   :84099999                                         Max.   :850.0  
##                                                                          
##       FIPS          Admin2          Province_State     Country_Region    
##  Min.   :   60   Length:3819906     Length:3819906     Length:3819906    
##  1st Qu.:19077   Class :character   Class :character   Class :character  
##  Median :31012   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :33043                                                           
##  3rd Qu.:47130                                                           
##  Max.   :99999                                                           
##  NA's   :11430                                                           
##       Lat              Long_          Combined_Key            date           
##  Min.   :-14.27   Min.   :-174.16   Length:3819906     Min.   :2020-01-22  
##  1st Qu.: 33.90   1st Qu.: -97.81   Class :character   1st Qu.:2020-11-02  
##  Median : 38.01   Median : -89.49   Mode  :character   Median :2021-08-15  
##  Mean   : 36.72   Mean   : -88.64                      Mean   :2021-08-15  
##  3rd Qu.: 41.58   3rd Qu.: -82.31                      3rd Qu.:2022-05-28  
##  Max.   : 69.31   Max.   : 145.67                      Max.   :2023-03-09  
```

```
##
##      cases           Population           deaths
##  Min.   :  -3073   Min.   :       0   Min.   :  -82.0
##  1st Qu.:    330   1st Qu.:    9917   1st Qu.:    4.0
##  Median :   2272   Median :   24892   Median :   37.0
##  Mean   :  14088   Mean   :   99604   Mean   :  186.9
##  3rd Qu.:   8159   3rd Qu.:   64979   3rd Qu.:  122.0
##  Max.   :3710586   Max.   :10039107   Max.   :35545.0
##
```

Conclusion: The Deaths due to covid is more in states where the cases are more. The summary statistics reveal that the dataset contains a substantial amount of data, with over 3.8 million records. The cases variable ranges from negative values to over 3.7 million, indicating the presence of both decreases and increases in reported cases. Similarly, the deaths variable ranges from negative values to over 35,000, reflecting variations in reported deaths.

The mean values for cases, Population, and deaths are 14,088, 99,604, and 186.9 respectively, suggesting relatively high average numbers of cases and deaths across the dataset. The presence of negative values in cases and deaths may require further investigation to understand the data quality and integrity.

Additionally, the presence of NA values in some columns, such as FIPS, indicates potential missing or incomplete data that may need to be addressed during analysis.

Overall, the dataset provides valuable insights into the spread and impact of COVID-19 in the United States, but further exploration and analysis are warranted to fully understand the dynamics and patterns within the data.

- **Bias identification**
  - *Testing Bias: Differences in who gets tested for COVID-19 can affect the number of reported cases. For example, areas with more testing may find more cases, while areas with limited testing may have fewer reported cases.*
  - *Reporting Bias: How and when COVID-19 cases are reported can vary, leading to inconsistencies. Delays or errors in reporting can make the data incomplete or inaccurate.*
  - *Severity Bias: COVID-19 cases that are more severe, such as those requiring hospitalization, may be more likely to be reported. Mild cases or cases with no symptoms may go unreported, leading to an incomplete picture of the disease's spread.*
  - *Population Density Bias: Areas with more people may have more cases reported simply because there are more people to test. This can make it seem like the disease is more widespread in densely populated areas.*
  - *Access Bias: People with better access to healthcare may be more likely to get tested and have their cases reported. Areas with poorer access to healthcare may have fewer reported cases, even if the disease is present.*
  - *Demographic Bias: Certain groups of people may be more or less likely to get tested or have their cases reported, leading to disparities in the data. This could be due to factors like race, income, or immigration status.*

These biases can affect the accuracy of COVID-19 data and make it harder to understand the true extent of the pandemic. It's important to be aware of these biases when interpreting COVID-19 data and making decisions based on it.