# John Hopkins Covid-19 Data Project

DM

2024-10-12

## Load Data

The data comes from John Hopkins from the COVID-19 pandemic and was archived on Mar 10, 2023.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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(lubridate)
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)
urls
```

```
## [1] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
## [2] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
## [3] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
## [4] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
```

```
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.
```

```
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.
```

```
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.
```

```
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.
```

## Cleaning and Transforming Data

```
#tidy data

global_cases <- global_cases %>%
  pivot_longer(cols = -c('Province/State','Country/Region', Lat, Long),
               names_to = "date",
               values_to = "cases") %>%
  select(-c(Lat,Long))

head(global_cases, n=10)
```

```
## # A tibble: 10 x 4
##    ‘Province/State‘ ‘Country/Region‘ date    cases
##    <chr>            <chr>            <chr>   <dbl>
##  1 <NA>             Afghanistan      1/22/20     0
##  2 <NA>             Afghanistan      1/23/20     0
##  3 <NA>             Afghanistan      1/24/20     0
##  4 <NA>             Afghanistan      1/25/20     0
##  5 <NA>             Afghanistan      1/26/20     0
##  6 <NA>             Afghanistan      1/27/20     0
##  7 <NA>             Afghanistan      1/28/20     0
##  8 <NA>             Afghanistan      1/29/20     0
##  9 <NA>             Afghanistan      1/30/20     0
## 10 <NA>             Afghanistan      1/31/20     0
```

```r
global_deaths <- global_deaths %>%
  pivot_longer(cols = -c('Province/State','Country/Region', Lat, Long),
             names_to = "date",
             values_to = "deaths") %>%
  select(-c(Lat,Long))

head(global_deaths, n=10)
```

```
## # A tibble: 10 x 4
##    ‘Province/State‘ ‘Country/Region‘ date    deaths
##    <chr>            <chr>            <chr>    <dbl>
##  1 <NA>             Afghanistan      1/22/20      0
##  2 <NA>             Afghanistan      1/23/20      0
##  3 <NA>             Afghanistan      1/24/20      0
##  4 <NA>             Afghanistan      1/25/20      0
##  5 <NA>             Afghanistan      1/26/20      0
##  6 <NA>             Afghanistan      1/27/20      0
##  7 <NA>             Afghanistan      1/28/20      0
##  8 <NA>             Afghanistan      1/29/20      0
##  9 <NA>             Afghanistan      1/30/20      0
## 10 <NA>             Afghanistan      1/31/20      0
```

```r
#transform data
#combine the cases into deaths per date into one variable called global
global <- global_cases %>%
  full_join(global_deaths) %>%
  rename(`Country_Region`='Country/Region',
        Province_State = 'Province/State') %>%
  mutate(date = mdy(date))
```

```
## Joining with ‘by = join_by(‘Province/State‘, ‘Country/Region‘, date)‘
```

```r
global <- global %>% filter(cases >0)
summary(global)
```

```
##  Province_State     Country_Region          date                 cases
##  Length:306827      Length:306827      Min.   :2020-01-22   Min.   :        1
##  Class :character   Class :character   1st Qu.:2020-12-12   1st Qu.:     1316
##  Mode  :character   Mode  :character   Median :2021-09-16   Median :    20365
##                                         Mean   :2021-09-11   Mean   :  1032863
##                                         3rd Qu.:2022-06-15   3rd Qu.:   271281
##                                         Max.   :2023-03-09   Max.   :103802702
##      deaths
##  Min.   :      0
##  1st Qu.:      7
##  Median :    214
##  Mean   :  14405
##  3rd Qu.:   3665
##  Max.   :1123836
```

```
#double check amount of cases close to maximum
global %>% filter(cases > 103000000)
```

```
## # A tibble: 23 x 5
##    Province_State Country_Region date            cases  deaths
##    <chr>          <chr>          <date>          <dbl>   <dbl>
##  1 <NA>           US             2023-02-15 103023231 1115741
##  2 <NA>           US             2023-02-16 103083910 1116851
##  3 <NA>           US             2023-02-17 103131898 1117572
##  4 <NA>           US             2023-02-18 103134605 1117589
##  5 <NA>           US             2023-02-19 103136077 1117590
##  6 <NA>           US             2023-02-20 103138119 1117663
##  7 <NA>           US             2023-02-21 103198669 1118025
##  8 <NA>           US             2023-02-22 103308832 1118886
##  9 <NA>           US             2023-02-23 103365511 1119521
## 10 <NA>           US             2023-02-24 103378408 1119573
## # i 13 more rows
```

```
#now clean and wrangle US cases
US_cases %>%
  pivot_longer(cols = -(UID:Combined_Key),
               names_to = "date",
               values_to = "cases")
```

```
## # A tibble: 3,819,906 x 13
##         UID iso2  iso3  code3  FIPS Admin2  Province_State Country_Region   Lat
##       <dbl> <chr> <chr> <dbl> <dbl> <chr>   <chr>          <chr>          <dbl>
##  1 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
##  2 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
##  3 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
##  4 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
##  5 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
##  6 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
##  7 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
##  8 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
##  9 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
## 10 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
## # i 3,819,896 more rows
## # i 4 more variables: Long_ <dbl>, Combined_Key <chr>, date <chr>, cases <dbl>
```

```
US_cases <- US_cases %>%
  pivot_longer(cols = -(UID:Combined_Key),
                names_to = "date",
                values_to = "cases") %>%
  select(Admin2:cases) %>%
  mutate(date=mdy(date)) %>%
  select(-c(Lat,Long_))

head(US_cases, n=10)
```

```
## # A tibble: 10 x 6
##    Admin2  Province_State Country_Region Combined_Key         date       cases
##    <chr>   <chr>          <chr>          <chr>                <date>     <dbl>
##  1 Autauga Alabama        US             Autauga, Alabama, US 2020-01-22     0
##  2 Autauga Alabama        US             Autauga, Alabama, US 2020-01-23     0
##  3 Autauga Alabama        US             Autauga, Alabama, US 2020-01-24     0
##  4 Autauga Alabama        US             Autauga, Alabama, US 2020-01-25     0
##  5 Autauga Alabama        US             Autauga, Alabama, US 2020-01-26     0
##  6 Autauga Alabama        US             Autauga, Alabama, US 2020-01-27     0
##  7 Autauga Alabama        US             Autauga, Alabama, US 2020-01-28     0
##  8 Autauga Alabama        US             Autauga, Alabama, US 2020-01-29     0
##  9 Autauga Alabama        US             Autauga, Alabama, US 2020-01-30     0
## 10 Autauga Alabama        US             Autauga, Alabama, US 2020-01-31     0
```

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

head(US_deaths, n=10)
```

```
## # A tibble: 10 x 7
##    Admin2  Province_State Country_Region Combined_Key      Population date
##    <chr>   <chr>          <chr>          <chr>                  <dbl> <date>
##  1 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-22
##  2 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-23
##  3 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-24
##  4 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-25
##  5 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-26
##  6 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-27
##  7 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-28
##  8 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-29
##  9 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-30
## 10 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-31
## # i 1 more variable: deaths <dbl>
```

```
#join two US datasets together
US <- US_cases %>%
  full_join(US_deaths)
```

```
## Joining with ‘by = join_by(Admin2, Province_State, Country_Region,
## Combined_Key, date)‘
```

```
#check join
head(US, n=10)
```

```
## # A tibble: 10 x 8
##    Admin2 Province_State Country_Region Combined_Key date       cases Population
##    <chr>  <chr>          <chr>          <chr>        <date>     <dbl>      <dbl>
##  1 Autau~ Alabama        US             Autauga, Al~ 2020-01-22     0      55869
##  2 Autau~ Alabama        US             Autauga, Al~ 2020-01-23     0      55869
##  3 Autau~ Alabama        US             Autauga, Al~ 2020-01-24     0      55869
##  4 Autau~ Alabama        US             Autauga, Al~ 2020-01-25     0      55869
##  5 Autau~ Alabama        US             Autauga, Al~ 2020-01-26     0      55869
##  6 Autau~ Alabama        US             Autauga, Al~ 2020-01-27     0      55869
##  7 Autau~ Alabama        US             Autauga, Al~ 2020-01-28     0      55869
##  8 Autau~ Alabama        US             Autauga, Al~ 2020-01-29     0      55869
##  9 Autau~ Alabama        US             Autauga, Al~ 2020-01-30     0      55869
## 10 Autau~ Alabama        US             Autauga, Al~ 2020-01-31     0      55869
## # i 1 more variable: deaths <dbl>
```

```
#add population data to global dataset

global <- global %>%
  unite("Combined_Key",
        c(Province_State, Country_Region),
        sep=", ",
        na.rm=TRUE,
        remove=FALSE)

head(global, n=10)
```

```
## # A tibble: 10 x 6
##    Combined_Key Province_State Country_Region date       cases deaths
##    <chr>        <chr>          <chr>          <date>     <dbl>  <dbl>
##  1 Afghanistan  <NA>           Afghanistan    2020-02-24     5      0
##  2 Afghanistan  <NA>           Afghanistan    2020-02-25     5      0
##  3 Afghanistan  <NA>           Afghanistan    2020-02-26     5      0
##  4 Afghanistan  <NA>           Afghanistan    2020-02-27     5      0
##  5 Afghanistan  <NA>           Afghanistan    2020-02-28     5      0
##  6 Afghanistan  <NA>           Afghanistan    2020-02-29     5      0
##  7 Afghanistan  <NA>           Afghanistan    2020-03-01     5      0
##  8 Afghanistan  <NA>           Afghanistan    2020-03-02     5      0
##  9 Afghanistan  <NA>           Afghanistan    2020-03-03     5      0
## 10 Afghanistan  <NA>           Afghanistan    2020-03-04     5      0
```

```
#retrieve lookup table URL from John Hopkins Website

uid_lookup_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/

uid <- read_csv(uid_lookup_url) %>%
  select(-c(Lat, Long_, Combined_Key, code3, iso2, iso3, Admin2))
```

```
## Rows: 4321 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): iso2, iso3, FIPS, Admin2, Province_State, Country_Region, Combined_Key
## dbl (5): UID, code3, Lat, Long_, Population
##
## 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.
```

```
#join uid onto global dataset

global <- global %>%
  left_join(uid, by=c("Province_State","Country_Region")) %>%
  select(-c(UID, FIPS)) %>%
  select(Province_State, Country_Region, date, cases, deaths, Population, Combined_Key)

head(global, n=10)
```

```
## # A tibble: 10 x 7
##    Province_State Country_Region date       cases deaths Population Combined_Key
##    <chr>          <chr>          <date>     <dbl>  <dbl>      <dbl> <chr>
##  1 <NA>           Afghanistan    2020-02-24     5      0   38928341 Afghanistan
##  2 <NA>           Afghanistan    2020-02-25     5      0   38928341 Afghanistan
##  3 <NA>           Afghanistan    2020-02-26     5      0   38928341 Afghanistan
##  4 <NA>           Afghanistan    2020-02-27     5      0   38928341 Afghanistan
##  5 <NA>           Afghanistan    2020-02-28     5      0   38928341 Afghanistan
##  6 <NA>           Afghanistan    2020-02-29     5      0   38928341 Afghanistan
##  7 <NA>           Afghanistan    2020-03-01     5      0   38928341 Afghanistan
##  8 <NA>           Afghanistan    2020-03-02     5      0   38928341 Afghanistan
##  9 <NA>           Afghanistan    2020-03-03     5      0   38928341 Afghanistan
## 10 <NA>           Afghanistan    2020-03-04     5      0   38928341 Afghanistan
```

In the data cleaning and transformation, we removed unnecessary columns for our analysis like latitude and longitude. We combined the cases into deaths per date into one variable called global. Combined_key puts together the county and the state for US cases.

## Visualizing Data

```
#First, group US dataset by state, region, and date
US_by_state <- US %>%
  group_by(Province_State, Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
            Population = sum(Population)) %>%
  mutate(deaths_per_mill = deaths * 1000000 / Population) %>%
  select(Province_State, Country_Region, date, cases, deaths, deaths_per_mill, Population) %>%
  ungroup()
```

```
## ‘summarise()‘ has grouped output by 'Province_State', 'Country_Region'. You can
## override using the ‘.groups‘ argument.
```

```r
head(US_by_state, n=10)
```

```
## # A tibble: 10 x 7
##    Province_State Country_Region date       cases deaths deaths_per_mill
##    <chr>          <chr>          <date>     <dbl>  <dbl>           <dbl>
##  1 Alabama        US             2020-01-22     0      0               0
##  2 Alabama        US             2020-01-23     0      0               0
##  3 Alabama        US             2020-01-24     0      0               0
##  4 Alabama        US             2020-01-25     0      0               0
##  5 Alabama        US             2020-01-26     0      0               0
##  6 Alabama        US             2020-01-27     0      0               0
##  7 Alabama        US             2020-01-28     0      0               0
##  8 Alabama        US             2020-01-29     0      0               0
##  9 Alabama        US             2020-01-30     0      0               0
## 10 Alabama        US             2020-01-31     0      0               0
## # i 1 more variable: Population <dbl>
```

```r
#look at total for US, group by Country/Region and date

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

```
## ‘summarise()‘ has grouped output by ’Country_Region’. You can override using
## the ‘.groups‘ argument.
```

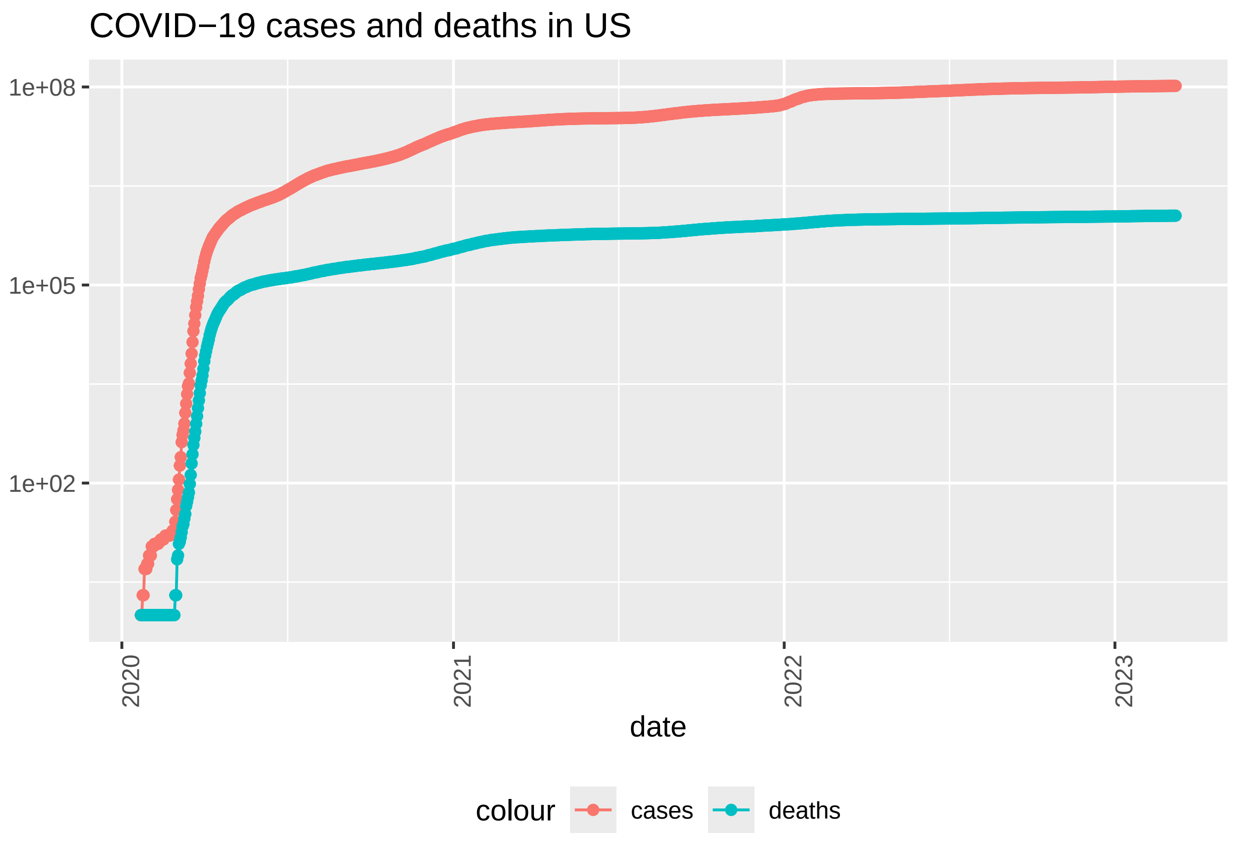
```r
head(US_totals, n=10)
```

```
## # A tibble: 10 x 6
##    Country_Region date       cases deaths deaths_per_mill Population
##    <chr>          <date>     <dbl>  <dbl>           <dbl>      <dbl>
##  1 US             2020-01-22     1      1         0.00300  332875137
##  2 US             2020-01-23     1      1         0.00300  332875137
##  3 US             2020-01-24     2      1         0.00300  332875137
##  4 US             2020-01-25     2      1         0.00300  332875137
##  5 US             2020-01-26     5      1         0.00300  332875137
##  6 US             2020-01-27     5      1         0.00300  332875137
##  7 US             2020-01-28     5      1         0.00300  332875137
##  8 US             2020-01-29     6      1         0.00300  332875137
##  9 US             2020-01-30     6      1         0.00300  332875137
## 10 US             2020-01-31     8      1         0.00300  332875137
```

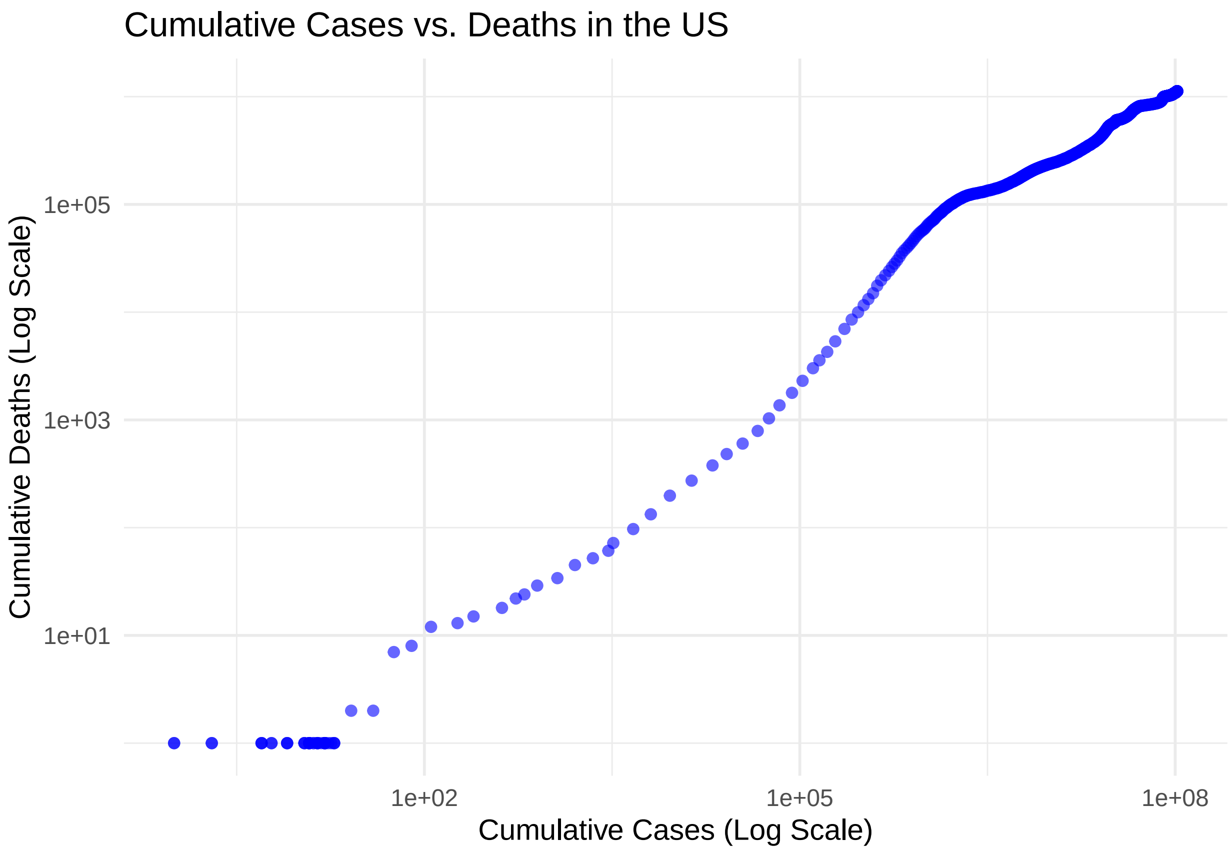
```r
tail(US_totals, n=10)
```

```
## # A tibble: 10 x 6
##    Country_Region date           cases  deaths deaths_per_mill Population
##    <chr>          <date>         <dbl>   <dbl>           <dbl>      <dbl>
##  1 US             2023-02-28 103443455 1119917           3364.  332875137
##  2 US             2023-03-01 103533872 1120897           3367.  332875137
##  3 US             2023-03-02 103589757 1121658           3370.  332875137
##  4 US             2023-03-03 103648690 1122165           3371.  332875137
##  5 US             2023-03-04 103650837 1122172           3371.  332875137
##  6 US             2023-03-05 103646975 1122134           3371.  332875137
##  7 US             2023-03-06 103655539 1122181           3371.  332875137
##  8 US             2023-03-07 103690910 1122516           3372.  332875137
##  9 US             2023-03-08 103755771 1123246           3374.  332875137
## 10 US             2023-03-09 103802702 1123836           3376.  332875137
```

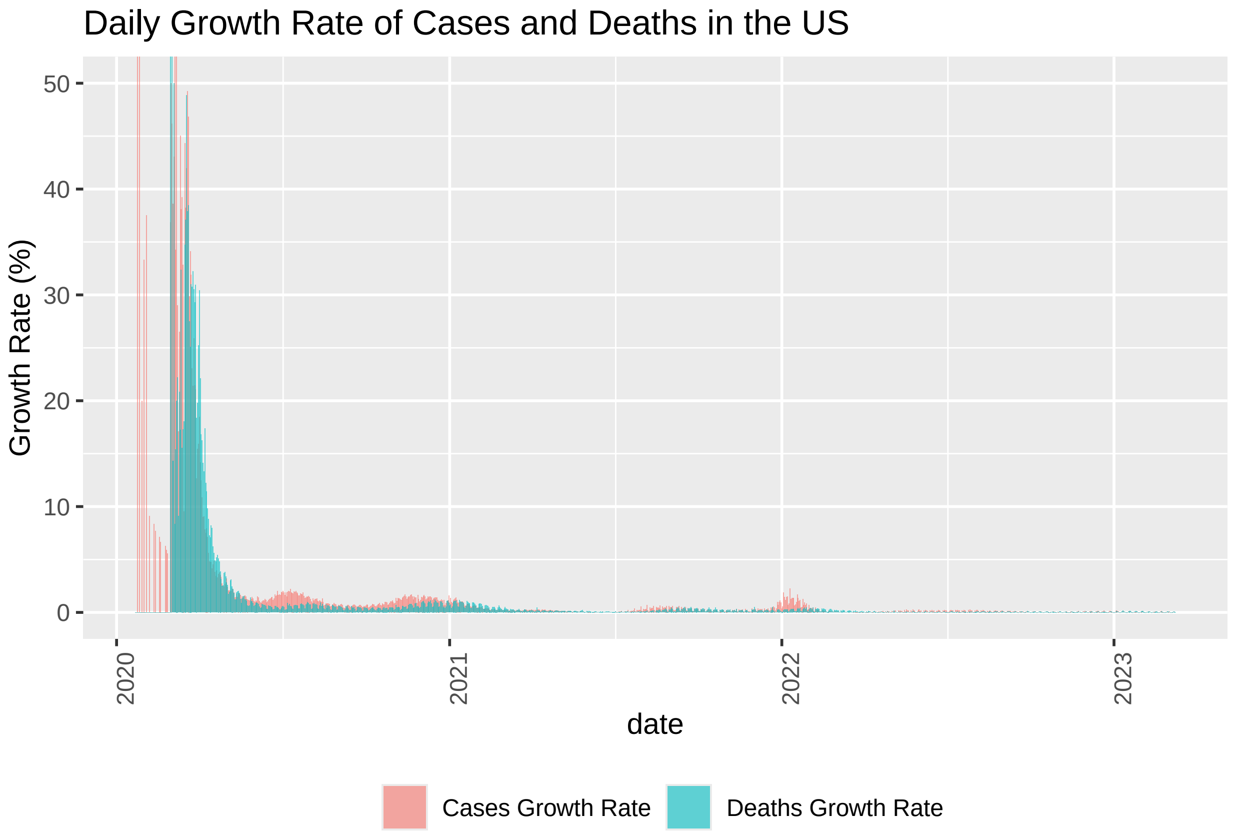
```
#plot US data

US_totals %>%
  filter(cases >0) %>%
  ggplot(aes(x=date, y=cases)) +
  geom_line(aes(color="cases")) +
  geom_point(aes(color="cases")) +
  geom_line(aes(y=deaths, color="deaths")) +
  geom_point(aes(y=deaths, color="deaths")) +
  scale_y_log10() +
  theme(legend.position="bottom",
        axis.text.x=element_text(angle=90)) +
  labs(title = "COVID-19 cases and deaths in US", y=NULL)
```



```
#New Visualization 1: Cumulative Cases vs. Deaths (Scatter Plot with Log Scale)
# This scatter plot will show the relationship between cumulative cases and cumulative deaths in the US
# We'll use a log scale to visualize a wide range of values effectively.

US_totals %>%
  filter(cases > 0, deaths > 0) %>%
  ggplot(aes(x = cases, y = deaths)) +
  geom_point(alpha = 0.6, color = "blue") +
  scale_x_log10() +
  scale_y_log10() +
  labs(title = "Cumulative Cases vs. Deaths in the US",
       x = "Cumulative Cases (Log Scale)",
       y = "Cumulative Deaths (Log Scale)") +
  theme_minimal()
```



```
#New Visualization 2: Cases and Deaths Growth Rate (Bar Chart)
#This bar chart tracks the daily percentage growth rates of both cases and deaths, helping to highlight
US_totals %>%
  filter(cases > 0, deaths > 0) %>%
  mutate(growth_rate_cases = (cases - lag(cases)) / lag(cases) * 100,
         growth_rate_deaths = (deaths - lag(deaths)) / lag(deaths) * 100) %>%
  replace_na(list(growth_rate_cases = 0, growth_rate_deaths = 0)) %>%
  ggplot(aes(x = date)) +
  geom_bar(aes(y = growth_rate_cases, fill = "Cases Growth Rate"), stat = "identity", alpha = 0.6) +
  geom_bar(aes(y = growth_rate_deaths, fill = "Deaths Growth Rate"), stat = "identity", alpha = 0.6) +
  coord_cartesian(ylim = c(0, 50)) +  # Adjust y-axis limits to zoom in
  labs(title = "Daily Growth Rate of Cases and Deaths in the US",
       y = "Growth Rate (%)",
       fill = "") +
  theme(axis.text.x = element_text(angle = 90),
        legend.position = "bottom")
```



In the first plot of 'COVID-19 cases and deaths in the US', we see that cases and deaths follow similar trends with a large increase around March/April 2020 and continue to increase but somewhat level-out through the beginning of 2023. In the new visualization, Cumulative Cases vs Deaths in the US, it shows the relationship between cumulative cases and cumulative deaths in the US which seems to be a positive linear relationship. In the second new visualization, Daily Growth Rate of Cases and Deaths in the US, this bar chart tracks the daily percentage growth rates of both cases and deaths, helping to highlight periods of rapid growth or slowdown. We see large growth rates in the beginning of 2020 and a few other spikes in the beginning of 2021 and 2022.

## Further Transformation and Analysis

```
US_by_state <- US_by_state %>%
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths))

US_totals <- US_totals %>%
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths))

#transform to group by state and deaths and cases per thousand
US_state_totals <- US_by_state %>%
  group_by(Province_State) %>%
  summarize(deaths = max(deaths), cases = max(cases),
            population = max(Population),
            cases_per_thou = 1000 * cases / population,
            deaths_per_thou = 1000 * deaths / population) %>%
  filter(cases >0, population >0)

#Lets look at the 10 states that had the least amount of deaths per thousand
US_state_totals %>%
  slice_min(deaths_per_thou, n=10) %>%
  select(deaths_per_thou, cases_per_thou, everything())
```

```
## # A tibble: 10 x 6
##    deaths_per_thou cases_per_thou Province_State        deaths  cases population
##              <dbl>          <dbl> <chr>                  <dbl>  <dbl>      <dbl>
##  1           0.611           150. American Samoa            34 8.32e3      55641
##  2           0.744           248. Northern Mariana Isl~     41 1.37e4      55144
##  3           1.21            231. Virgin Islands           130 2.48e4     107268
##  4           1.30            269. Hawaii                  1841 3.81e5    1415872
##  5           1.49            245. Vermont                  929 1.53e5     623989
##  6           1.55            293. Puerto Rico             5823 1.10e6    3754939
##  7           1.65            340. Utah                    5298 1.09e6    3205958
##  8           2.01            415. Alaska                  1486 3.08e5     740995
##  9           2.03            252. District of Columbia    1432 1.78e5     705749
## 10           2.06            253. Washington             15683 1.93e6    7614893
```

```
#Here are the 10 states that had the most amount of deaths per thousand
US_state_totals %>%
  slice_max(deaths_per_thou, n=10) %>%
  select(deaths_per_thou, cases_per_thou, everything())
```

```
## # A tibble: 10 x 6
##    deaths_per_thou cases_per_thou Province_State deaths   cases population
##              <dbl>          <dbl> <chr>           <dbl>   <dbl>      <dbl>
##  1            4.55           336. Arizona         33102 2443514    7278717
##  2            4.54           326. Oklahoma        17972 1290929    3956971
##  3            4.49           333. Mississippi     13370  990756    2976149
##  4            4.44           359. West Virginia    7960  642760    1792147
##  5            4.32           320. New Mexico       9061  670929    2096829
##  6            4.31           334. Arkansas        13020 1006883    3017804
##  7            4.29           335. Alabama         21032 1644533    4903185
##  8            4.28           368. Tennessee       29263 2515130    6829174
##  9            4.23           307. Michigan        42205 3064125    9986857
## 10            4.06           385. Kentucky        18130 1718471    4467673
```

In this analysis, we wanted to look at which states had the most and least deaths per thousand residents. The state/territory with the least amount of deaths per thousand residents was American Samoa and the state territory with the most amount of deaths per thousand residents was Arizona.

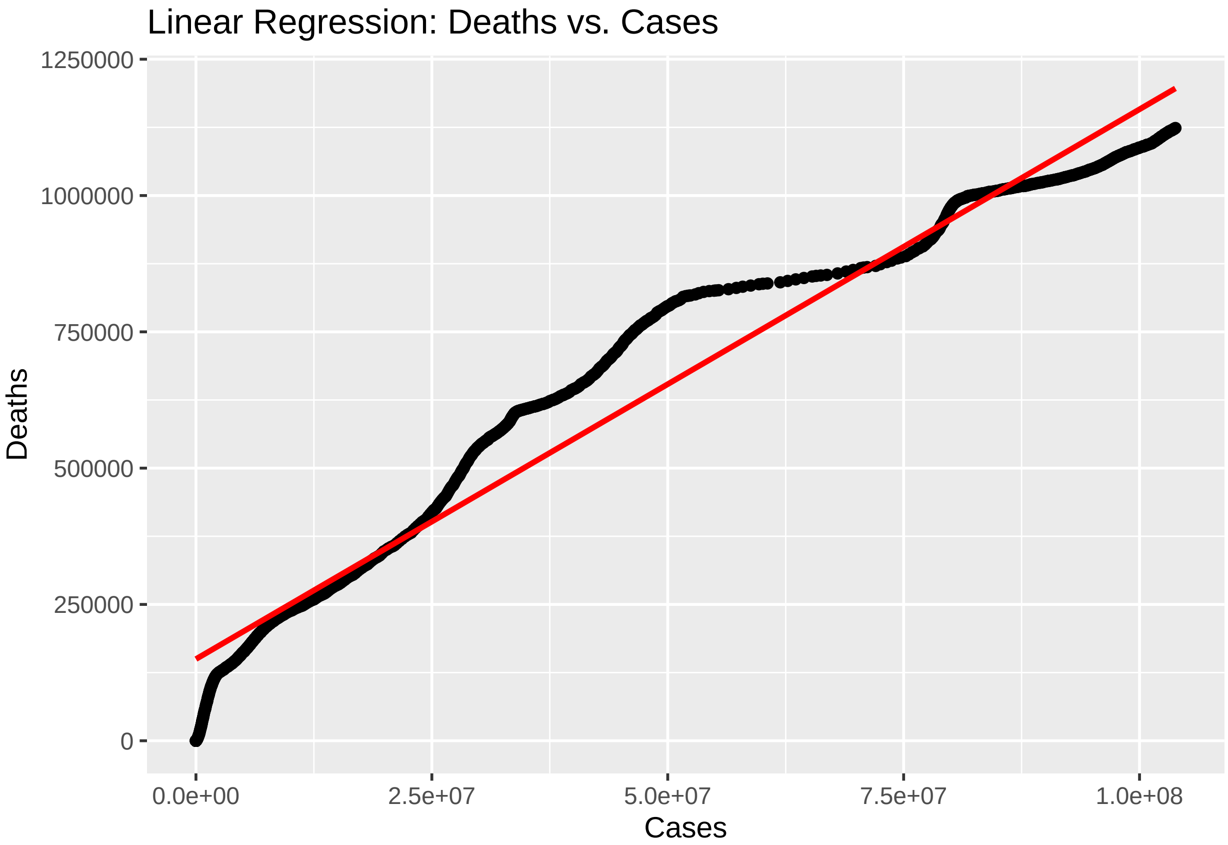
## Modeling the Data

```r
#Simple Linear Regression Model to predict deaths based on the number of cases in the US.

# Filter data to only include rows where both cases and deaths are greater than 0
US_lm_data <- US_totals %>%
  filter(cases > 0, deaths > 0)

# Fit a simple linear regression model
model <- lm(deaths ~ cases, data = US_lm_data)

# Summary of the model
summary(model)
```

```
##
## Call:
## lm(formula = deaths ~ cases, data = US_lm_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -149805  -62016  -13343   89500  143891
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.498e+05  3.869e+03   38.72   <2e-16 ***
## cases       1.008e-02  6.498e-05  155.19   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 80090 on 1141 degrees of freedom
## Multiple R-squared:  0.9548, Adjusted R-squared:  0.9547
## F-statistic: 2.408e+04 on 1 and 1141 DF,  p-value: < 2.2e-16
```

```r
# Visualize the regression model
ggplot(US_lm_data, aes(x = cases, y = deaths)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Linear Regression: Deaths vs. Cases", x = "Cases", y = "Deaths")
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



The modeled data is a simple linear regression model to predict deaths based on the number of cases in the US. The data was filtered to only include rows where both cases and deaths are greater than 0. We see that based on the total data given between early 2020 and March 2023, the model is predicting an increase in deaths. There is a positive linear relationship between the amount of cases and deaths. We can also see from the summary output that the p-value for the predictor variable, cases, is significant ($< 0.05$).