Covid

2024-03-03

#The purpose of this report is to analyze data used from the Johns Hopkins Covid-19 dataset provided in a github repository. This rmd file should be reproducible as the libraries used are noted at the top and the links to the data are from the github rather than my local PC. The aim was to analyze this data in any way we wanted, so I chose to analyze the data through looking at deaths and cases in both the US and globally as well as to compare these deaths and cases to see any patterns.

```
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   3.5.0     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 errors
```

```
library(ggplot2)
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

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

```
## [1] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_1
## [2] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_1
## [3] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_1
## [4] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_1
```

```r
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
global_cases
```

```
## # A tibble: 289 x 1,147
##    ‘Province/State‘  ‘Country/Region‘   Lat  Long ‘1/22/20‘ ‘1/23/20‘ ‘1/24/20‘
##    <chr>             <chr>            <dbl>  <dbl>     <dbl>     <dbl>     <dbl>
##   1 <NA>              Afghanistan       33.9  67.7          0         0         0
##  2 <NA>              Albania           41.2  20.2          0         0         0
##  3 <NA>              Algeria           28.0   1.66         0         0         0
##  4 <NA>              Andorra           42.5   1.52         0         0         0
##  5 <NA>              Angola           -11.2  17.9          0         0         0
##  6 <NA>              Antarctica       -71.9  23.3          0         0         0
##  7 <NA>              Antigua and Bar~  17.1 -61.8          0         0         0
##  8 <NA>              Argentina        -38.4 -63.6          0         0         0
##  9 <NA>              Armenia           40.1  45.0          0         0         0
## 10 Australian Capit~ Australia        -35.5 149.           0         0         0
## # i 279 more rows
## # i 1,140 more variables: ‘1/25/20‘ <dbl>, ‘1/26/20‘ <dbl>, ‘1/27/20‘ <dbl>,
## #   ‘1/28/20‘ <dbl>, ‘1/29/20‘ <dbl>, ‘1/30/20‘ <dbl>, ‘1/31/20‘ <dbl>,
## #   ‘2/1/20‘ <dbl>, ‘2/2/20‘ <dbl>, ‘2/3/20‘ <dbl>, ‘2/4/20‘ <dbl>,
## #   ‘2/5/20‘ <dbl>, ‘2/6/20‘ <dbl>, ‘2/7/20‘ <dbl>, ‘2/8/20‘ <dbl>,
## #   ‘2/9/20‘ <dbl>, ‘2/10/20‘ <dbl>, ‘2/11/20‘ <dbl>, ‘2/12/20‘ <dbl>,
## #   ‘2/13/20‘ <dbl>, ‘2/14/20‘ <dbl>, ‘2/15/20‘ <dbl>, ‘2/16/20‘ <dbl>, ...
```

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

```
## # A tibble: 330,327 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
## # i 330,317 more rows
```

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

```
## # A tibble: 330,327 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
## # i 330,317 more rows
```

#Seen above are the pivot longer statements we included in class in order to tidy the data and view the dates as rows rather than columns

```
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)`
```

```
global
```

```
## # A tibble: 330,327 x 5
##    Province_State Country_Region date       cases deaths
##    <chr>          <chr>          <date>     <dbl>  <dbl>
##  1 <NA>           Afghanistan    2020-01-22     0      0
##  2 <NA>           Afghanistan    2020-01-23     0      0
##  3 <NA>           Afghanistan    2020-01-24     0      0
##  4 <NA>           Afghanistan    2020-01-25     0      0
##  5 <NA>           Afghanistan    2020-01-26     0      0
##  6 <NA>           Afghanistan    2020-01-27     0      0
##  7 <NA>           Afghanistan    2020-01-28     0      0
##  8 <NA>           Afghanistan    2020-01-29     0      0
##  9 <NA>           Afghanistan    2020-01-30     0      0
## 10 <NA>           Afghanistan    2020-01-31     0      0
## # i 330,317 more rows
```

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

```
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_))
US_cases
```

```
## # A tibble: 3,819,906 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
## # i 3,819,896 more rows
```

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

```
## Warning: There was 1 warning in ‘mutate()‘.
## i In argument: ‘date = mdy(date)‘.
## Caused by warning:
## !  3342 failed to parse.
```

```
US <- US_cases %>%
full_join(US_deaths)
```

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

```
US
```

```
## # A tibble: 3,823,248 x 7
##    Admin2  Province_State Country_Region Combined_Key    date       cases deaths
##    <chr>   <chr>          <chr>          <chr>           <date>     <dbl>  <dbl>
##  1 Autauga Alabama        US             Autauga, Alaba~ 2020-01-22     0      0
##  2 Autauga Alabama        US             Autauga, Alaba~ 2020-01-23     0      0
##  3 Autauga Alabama        US             Autauga, Alaba~ 2020-01-24     0      0
##  4 Autauga Alabama        US             Autauga, Alaba~ 2020-01-25     0      0
##  5 Autauga Alabama        US             Autauga, Alaba~ 2020-01-26     0      0
##  6 Autauga Alabama        US             Autauga, Alaba~ 2020-01-27     0      0
##  7 Autauga Alabama        US             Autauga, Alaba~ 2020-01-28     0      0
##  8 Autauga Alabama        US             Autauga, Alaba~ 2020-01-29     0      0
##  9 Autauga Alabama        US             Autauga, Alaba~ 2020-01-30     0      0
## 10 Autauga Alabama        US             Autauga, Alaba~ 2020-01-31     0      0
## # i 3,823,238 more rows
```

#Now that I have full joined the data for both deaths and cases for the US and global sets, visuals are ready to be created.

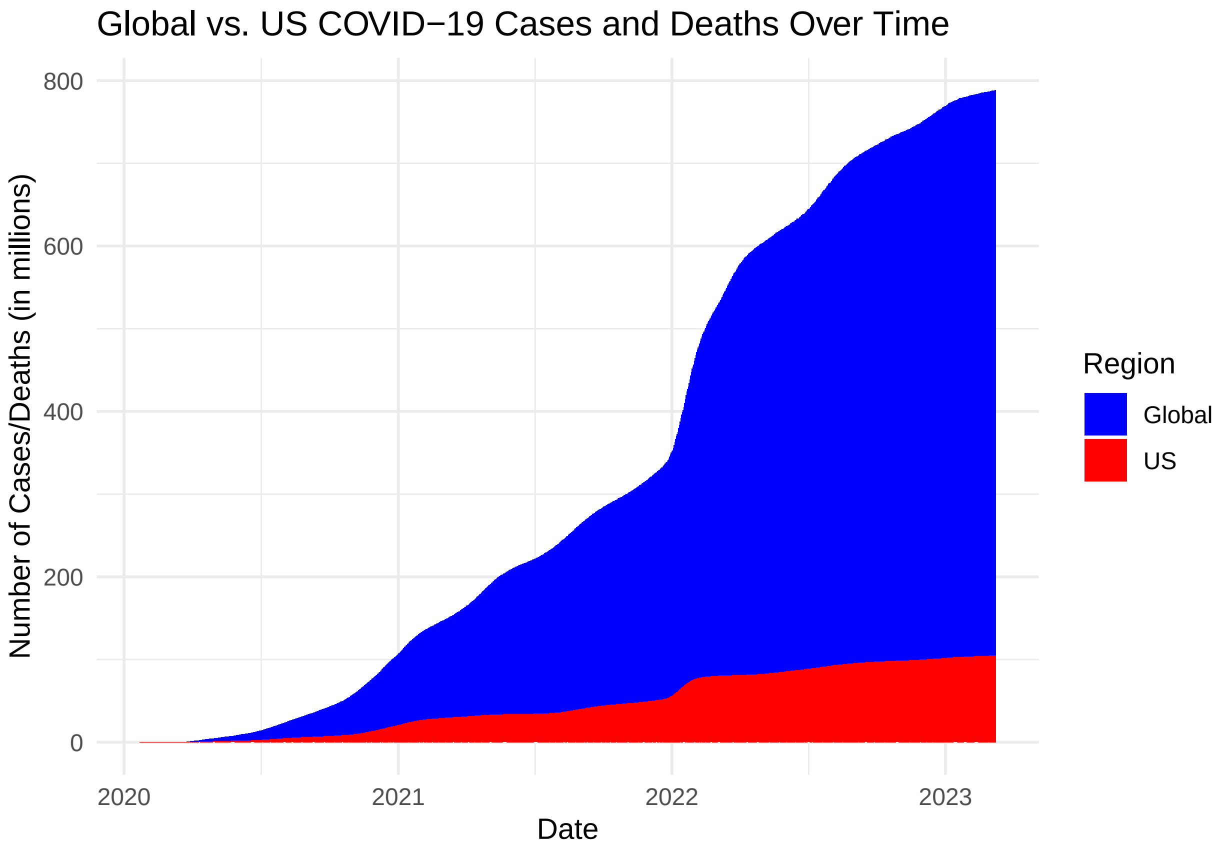
```
global_data_summary <- global %>%
  group_by(date) %>%
  summarize(total_cases = sum(cases, na.rm = TRUE),
            total_deaths = sum(deaths, na.rm = TRUE),
            region = "Global")

us_data_summary <- US %>%
  group_by(date) %>%
  summarize(total_cases = sum(cases, na.rm = TRUE),
            total_deaths = sum(deaths, na.rm = TRUE),
            region = "US")

combined_data <- bind_rows(global_data_summary, us_data_summary)

ggplot(combined_data, aes(x = date, y = total_cases + total_deaths, fill = region)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(labels = function(x) format(x / 1e6, scientific = FALSE, big.mark = ",")) +
  labs(title = "Global vs. US COVID-19 Cases and Deaths Over Time",
       x = "Date",
       y = "Number of Cases/Deaths (in millions)",
       fill = "Region") +
  scale_fill_manual(values = c("Global" = "blue", "US" = "red")) +
  theme_minimal()
```

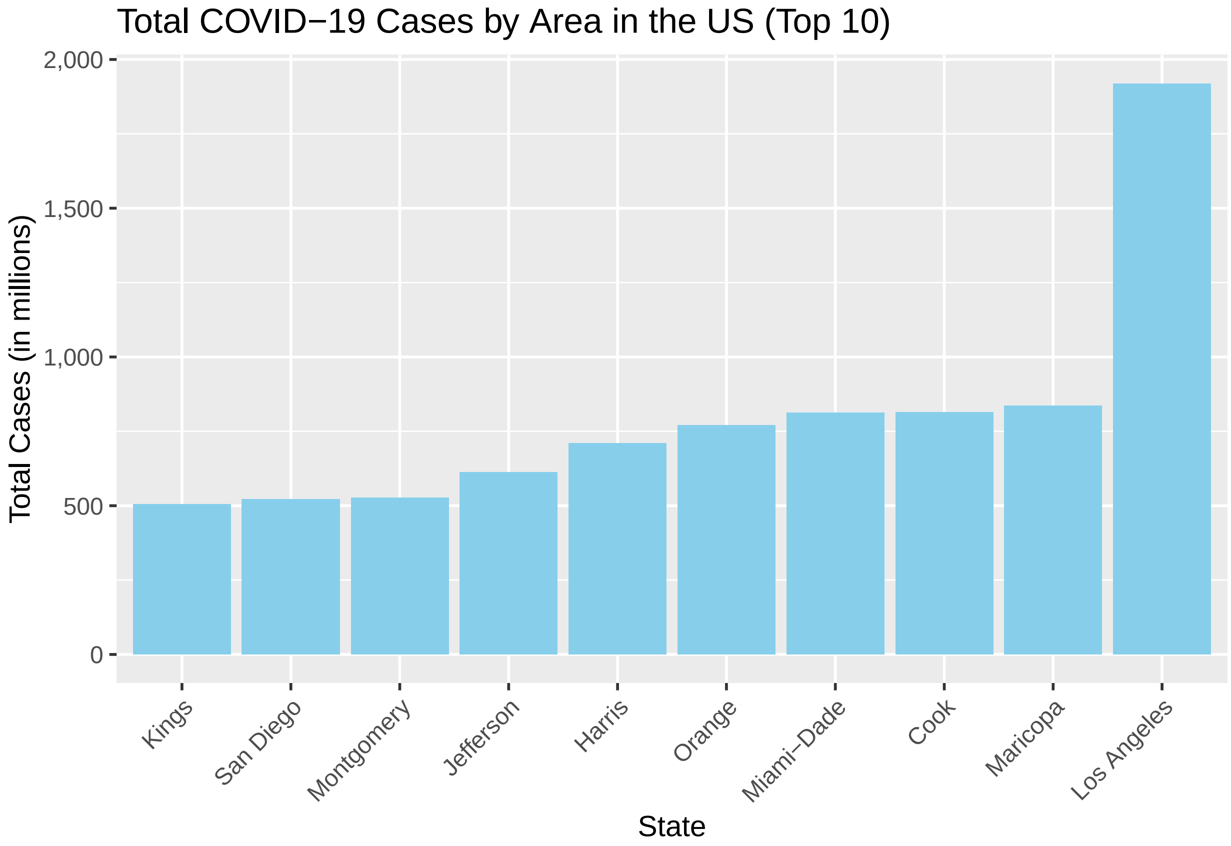
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_bar()‘).
```



#In this first visual, I combined the global and US data in order to create a plot to recognize the severity of the outbreak globally versus just in the US.

```r
us_data_summary <- US %>%
  group_by(Admin2) %>%
  summarize(total_cases = sum(cases, na.rm = TRUE)) %>%
  top_n(10, total_cases)  # Keep only the top 10 states

ggplot(us_data_summary, aes(x = reorder(Admin2, total_cases), y = total_cases)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  scale_y_continuous(labels = function(x) format(x / 1e6, big.mark = ",")) +
  labs(title = "Total COVID-19 Cases by Area in the US (Top 10)",
       x = "State",
       y = "Total Cases (in millions)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



This second plot shows which areas in the US were most affected by the virus. I originally wanted to display each area by cases, but ran into expected issues with overcrowded data.

#Regarding bias, one type of bias present in my analysis is data-processing bias. This is present because when cleaning the data to make it easier and more intuative to use, I removed certain data that will lead to a bias. Additionally, when the data was collected, there could very well have been reporting bias. Reporting bias would have came from certain issues that some states may have in terms of reporting their cases. Also, there are many people who do not have access to healthcare, so it is possible that these people were not in a hospital at the time of death and were not even reported to have covid, so their deaths were not counted in the report. Similarly, it is possible that some people had different conditions that caused death while they were simultaneously dealing with a covid infection. All of these reasons can lead to a bias in the dataset we worked with.