# Covid-19 Data Analysis

JamesYildiz

June 20, 2021

## Data Source

This covid 19 data has been retrieved from Johns Hopkins Universities github repository. It show the total covid 19 cases and deaths worldwide since the start of the pandemis until present time.

## Importing Libraries

```
# The following libraries will be needed for this study


library(tidyverse)
library(lubridate)
library(ggplot2)
```

## Importing Data

```
us_cases <-"https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_co

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

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

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



us_cases <- read_csv(us_cases)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   iso2 = col_character(),
##   iso3 = col_character(),
##   Admin2 = col_character(),
##   Province_State = col_character(),
##   Country_Region = col_character(),
```

```
##   Combined_Key = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
us_deaths <- read_csv(us_deaths)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   iso2 = col_character(),
##   iso3 = col_character(),
##   Admin2 = col_character(),
##   Province_State = col_character(),
##   Country_Region = col_character(),
##   Combined_Key = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
global_cases <- read_csv(global_cases)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   `Province/State` = col_character(),
##   `Country/Region` = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
global_deaths = read_csv(global_deaths)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   `Province/State` = col_character(),
##   `Country/Region` = col_character()
## )
## i Use `spec()` for the full column specifications.
```

## Cleaning Global 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))

global_cases
```

```
## # A tibble: 177,444 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
## # ... with 177,434 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))

head(global_deaths)
```

```
## # A tibble: 6 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
```

## Combining global data and formating date

```
global <- global_cases %>%
  full_join(global_deaths) %>%
  rename(Country_Region = `Country/Region`,
         Province_State = `Province/State`) %>%
  mutate(date = mdy(date))
```

```
## Joining, by = c("Province/State", "Country/Region", "date")
```

```
summary(global)
```

```
##  Province_State     Country_Region          date                cases
##  Length:177444      Length:177444      Min.   :2020-01-22   Min.   :       0
##  Class :character   Class :character   1st Qu.:2020-06-28   1st Qu.:     156
##  Mode  :character   Mode  :character   Median :2020-12-04   Median :    2586
##                                        Mean   :2020-12-04   Mean   :  311757
```

```
##                                       3rd Qu.:2021-05-12   3rd Qu.:   58993
##                                       Max.   :2021-10-18   Max.   :45050910
##      deaths
##  Min.   :     0
##  1st Qu.:     1
##  Median :    41
##  Mean   :  7088
##  3rd Qu.:   980
##  Max.   :725835
```

Filter out days where cases are equal to zero

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

summary(global)
```

```
##  Province_State     Country_Region          date                cases
##  Length:161373      Length:161373      Min.   :2020-01-22   Min.   :       1
##  Class :character   Class :character   1st Qu.:2020-07-31   1st Qu.:     393
##  Mode  :character   Mode  :character   Median :2020-12-29   Median :    4752
##                                        Mean   :2020-12-26   Mean   :  342805
##                                        3rd Qu.:2021-05-25   3rd Qu.:   77668
##                                        Max.   :2021-10-18   Max.   :45050910
##      deaths
##  Min.   :     0
##  1st Qu.:     3
##  Median :    70
##  Mean   :  7794
##  3rd Qu.:  1369
##  Max.   :725835
```

### Adding population data of the countries from Johns Hopkins population file.

```
population_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/U


population_data <- read_csv(population_url)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   UID = col_double(),
##   iso2 = col_character(),
##   iso3 = col_character(),
##   code3 = col_double(),
##   FIPS = col_character(),
##   Admin2 = col_character(),
##   Province_State = col_character(),
##   Country_Region = col_character(),
##   Lat = col_double(),
##   Long_ = col_double(),
```

```
##   Combined_Key = col_character(),
##   Population = col_double()
## )
```

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


head(global)
```

```
## # A tibble: 6 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
```

## Filter totals by country

Use the max function to filter the current cases and deaths

```
global_by_country <- global %>%
  group_by(Country_Region) %>%
  summarize(cases = max(cases), deaths = max(deaths))


head(global_by_country)
```

```
## # A tibble: 6 x 3
##   Country_Region       cases deaths
##   <chr>                <dbl>  <dbl>
## 1 Afghanistan         155776   7246
## 2 Albania             178188   2829
## 3 Algeria             205364   5873
## 4 Andorra              15367    130
## 5 Angola               63012   1670
## 6 Antigua and Barbuda   3918     95
```

## Death to case ratio

This data shows the ratio of the deaths and number of cases.

```
death_to_case <- round(((global_by_country$deaths/global_by_country$cases)*100), 2)

global_by_country <- global_by_country %>%
```

```
  mutate(death_ratio = death_to_case)

global_by_country[rev(order(global_by_country$death_ratio)),] %>% head(13)
```

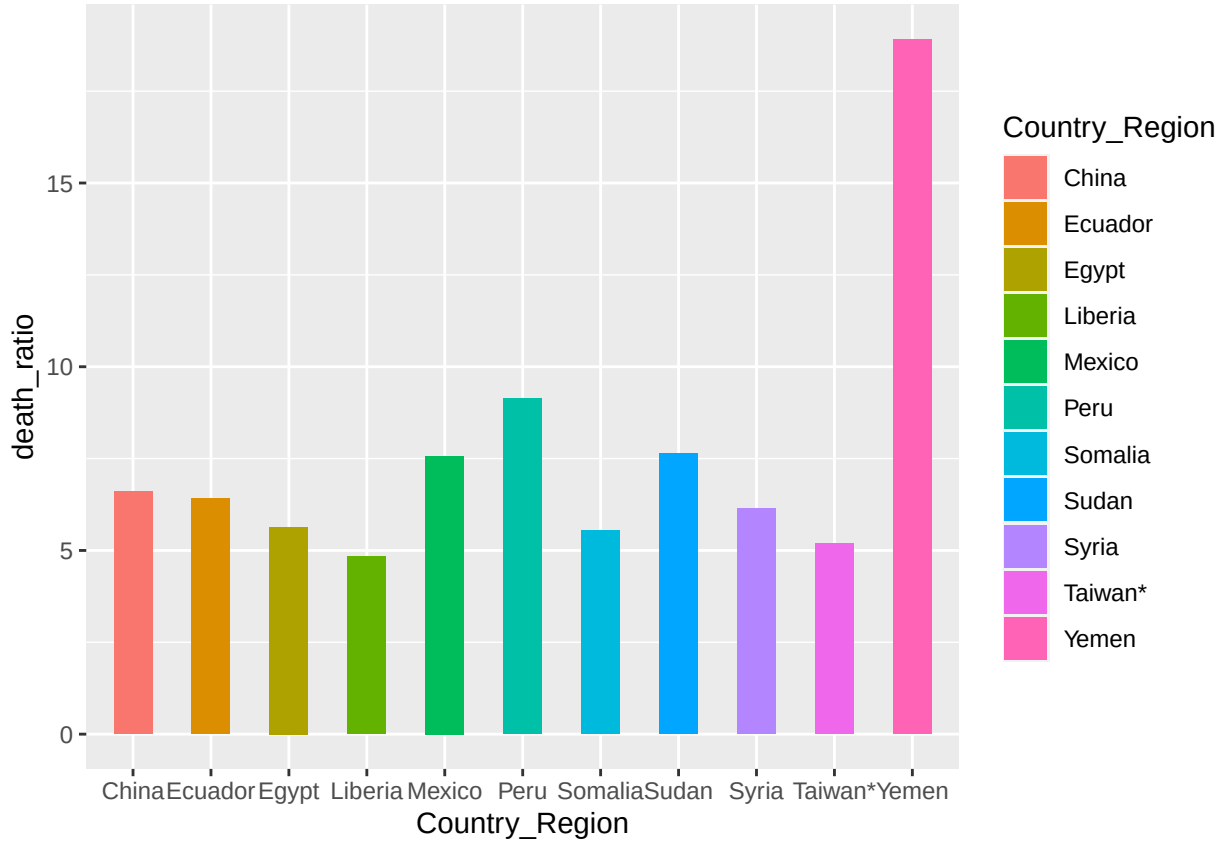
```
## # A tibble: 13 x 4
##    Country_Region   cases deaths death_ratio
##    <chr>            <dbl>  <dbl>       <dbl>
##  1 Vanuatu              4      1       25
##  2 MS Zaandam           9      2       22.2
##  3 Yemen             9556   1807       18.9
##  4 Peru           2190396 199882        9.13
##  5 Sudan            39839   3038        7.63
##  6 Mexico         3758469 284477        7.57
##  7 China            68303   4512        6.61
##  8 Ecuador         513026  32899        6.41
##  9 Syria            39488   2429        6.15
## 10 Egypt           319339  18015        5.64
## 11 Somalia          21269   1180        5.55
## 12 Taiwan*          16337    846        5.18
## 13 Liberia           5915    286        4.84
```

```
global_by_country_10 <- global_by_country %>% filter(20 > death_ratio, death_ratio > 4.7)

global_by_country_10[rev(order(global_by_country_10$death_ratio)),] %>% head(10)
```

```
## # A tibble: 10 x 4
##    Country_Region   cases deaths death_ratio
##    <chr>            <dbl>  <dbl>       <dbl>
##  1 Yemen             9556   1807       18.9
##  2 Peru           2190396 199882        9.13
##  3 Sudan            39839   3038        7.63
##  4 Mexico         3758469 284477        7.57
##  5 China            68303   4512        6.61
##  6 Ecuador         513026  32899        6.41
##  7 Syria            39488   2429        6.15
##  8 Egypt           319339  18015        5.64
##  9 Somalia          21269   1180        5.55
## 10 Taiwan*          16337    846        5.18
```

```
ggplot(global_by_country_10, aes(x=Country_Region, y=death_ratio, fill = Country_Region)) + geom_col(wi
```




## Discussion

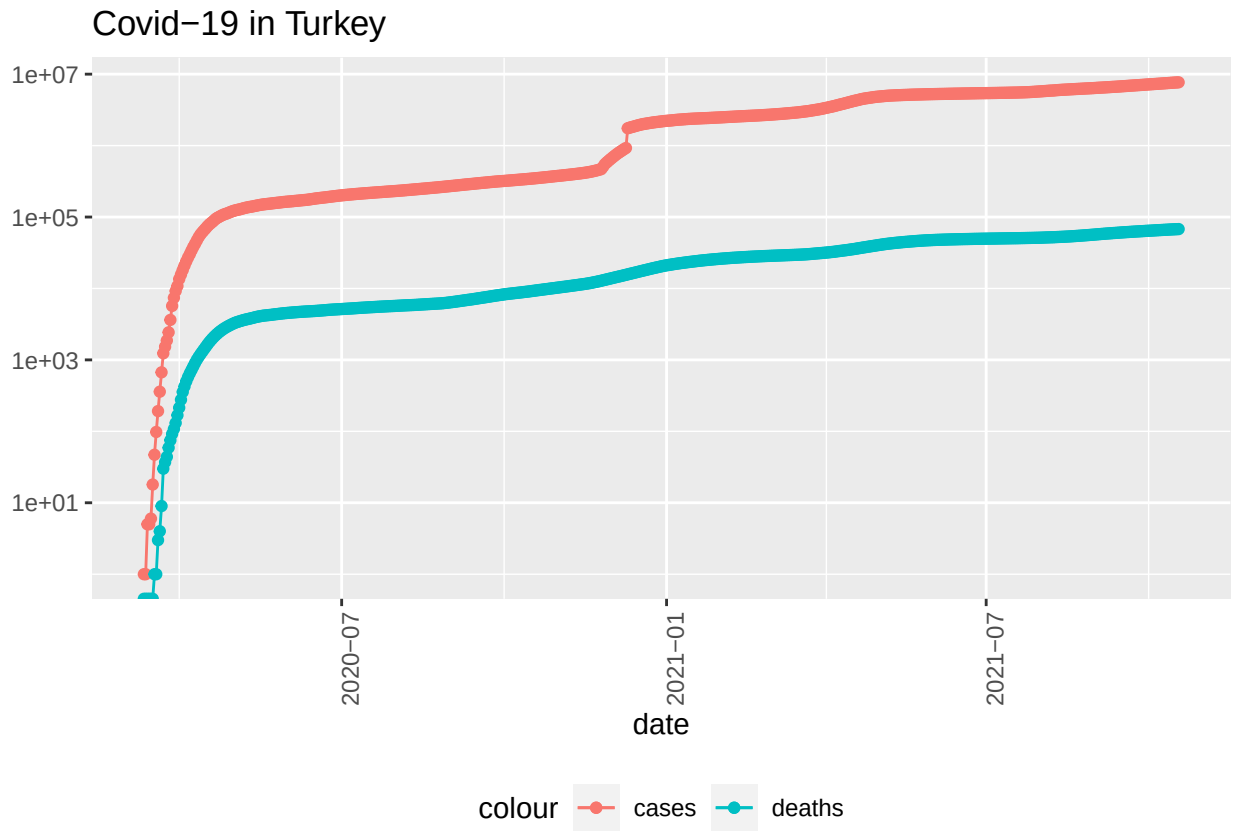
The above data shows that the countries with an ongoing civil war or limited helthcare facilies have higher death to cases ratio. The accuracy of the data collected from countries like Yemen, Somalia, and Syria can be debatable. It is definitely worth looking deeper into this data

## This section takes a deeper look at the data from Turkey (where I am from)

### Total Cases

```
turkey_totals <- global %>% filter(Country_Region == "Turkey")


turkey_totals %>%
  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 in Turkey", y=NULL)
```




**Daily Cases**

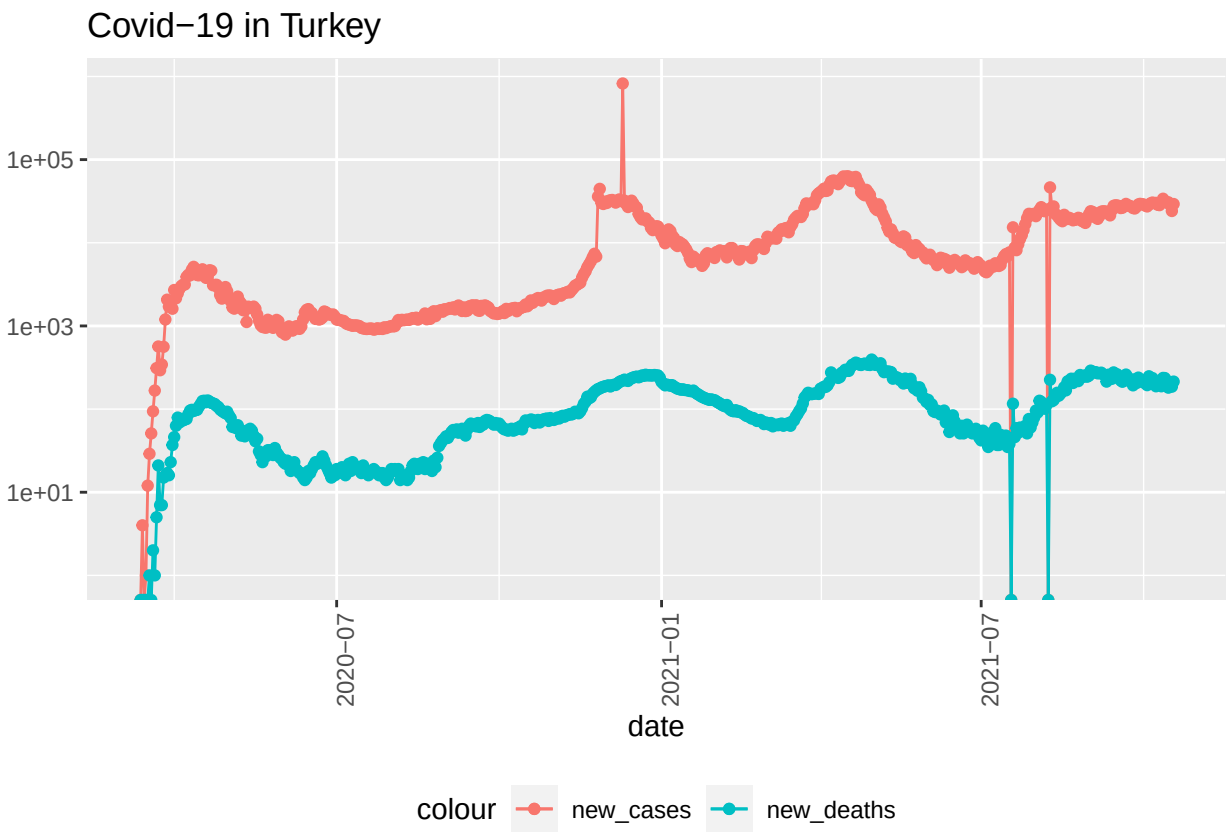
```
turkey_daily <- turkey_totals %>%
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths))

tail(turkey_daily %>% select(new_cases, new_deaths, everything()))
```

```
## # A tibble: 6 x 9
##   new_cases new_deaths Province_State Country_Region date       cases deaths
##       <dbl>      <dbl> <chr>          <chr>          <date>     <dbl>  <dbl>
## 1     31248        236 <NA>           Turkey         2021-10-13 7540193  66841
## 2     30709        203 <NA>           Turkey         2021-10-14 7570902  67044
## 3     30694        181 <NA>           Turkey         2021-10-15 7601596  67225
## 4     28537        212 <NA>           Turkey         2021-10-16 7630133  67437
## 5     24114        186 <NA>           Turkey         2021-10-17 7654247  67623
## 6     29240        214 <NA>           Turkey         2021-10-18 7683487  67837
## # ... with 2 more variables: Population <dbl>, Combined_Key <chr>
```

```
turkey_daily %>%
  ggplot(aes(x = date, y = new_cases)) +
  geom_line(aes(color = "new_cases")) +
  geom_point(aes(color = "new_cases")) +
```

```
  geom_line(aes(y = new_deaths, color = "new_deaths")) +
  geom_point(aes(y = new_deaths, color = "new_deaths")) +
  scale_y_log10() +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Covid-19 in Turkey", y=NULL)
```




**Discussion**

Turkey went into two national lockdown in mid-December and mid-April. I wanted to see how the daily numbers have changed during those timem. As it can be seen on the chart, the numbers have dropped significantly (consider the log scale) during and after the national lockdowns.

This global data set may contain bias from different sources. Some countries may have reported lower numbers due to concerns with losing public support for the government. In other countries, reported cases and deaths may be significantly lower due to limited testing and the way the deaths are reported.