Covid-19 study

Kevin Juandi

2024-03-22

In this work, I will try to understand how case numbers change over time and use this to predict the number of cases in the future.

Installing Libraries

```
#install.packages("tidyverse")
#install.packages("lubridate")
#install.packages("ggplot2")
#install.packages("repr")
```

Loading Libraries

```
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse, warn.conflicts = FALSE)
```

```
## -- 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   1.5.0
## 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(lubridate, warn.conflicts = FALSE)
library(ggplot2, warn.conflicts = FALSE)
library(repr, warn.conflicts = FALSE)
options(repr.plot.width=10, repr.plot.height=8)
options(dplyr.summarise.inform = FALSE)
options(warn = -1)
```

Loading the Data

Data Location: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series

This data is the list of COVID-19 cases by country and province. I will use three files consisting of data for confirmed cases, deaths, and recoveries.

```r
death_global_data =
  read.csv(file = paste(
    'https://github.com/CSSEGISandData/COVID-19',
    'raw/master/csse_covid_19_data/csse_covid_19_time_series',
    'time_series_covid19_deaths_global.csv',
    sep = '/'))
confirmed_global_data =
  read.csv(file = paste(
    'https://github.com/CSSEGISandData/COVID-19',
    'raw/master/csse_covid_19_data/csse_covid_19_time_series',
    'time_series_covid19_confirmed_global.csv',
    sep = '/'
  ))
recovered_global_data =
  read.csv(file = paste(
    'https://github.com/CSSEGISandData/COVID-19',
    'raw/master/csse_covid_19_data/csse_covid_19_time_series',
    'time_series_covid19_recovered_global.csv',
    sep = '/'
  ))
```

Preprocessing the data

This will take pretty long time due to the size of the data.

```r
tidy_data <- function(data) {
  result <- data %>%
    pivot_longer(cols = -c(`Province.State`, `Country.Region`, Lat, Long),
                 names_to = 'Date',
                 values_to = 'Cases') %>%
    select(-c(Lat, Long)) %>%
    mutate(Date = mdy(substring(Date, 2, length(Date)))) %>%
    group_by(Date) %>%
    summarise(
      Cases = sum(Cases),
    )
}

death_global_data_tidy <- tidy_data(death_global_data) %>%
  mutate(Deaths = Cases) %>%
  select(-Cases)
confirmed_global_data_tidy <- tidy_data(confirmed_global_data) %>%
  mutate(Confirmed = Cases) %>%
  select(-Cases)
recovered_global_data_tidy <- tidy_data(recovered_global_data) %>%
  mutate(Recovered = Cases) %>%
  select(-Cases)

global <- death_global_data_tidy %>%
  full_join(confirmed_global_data_tidy, by = 'Date') %>%
  full_join(recovered_global_data_tidy, by = 'Date')
```

Here I've converted the date columns to rows and then grouped by date. I've also added a column for the number of cases.

Three datasets were joined together to create a single dataset.

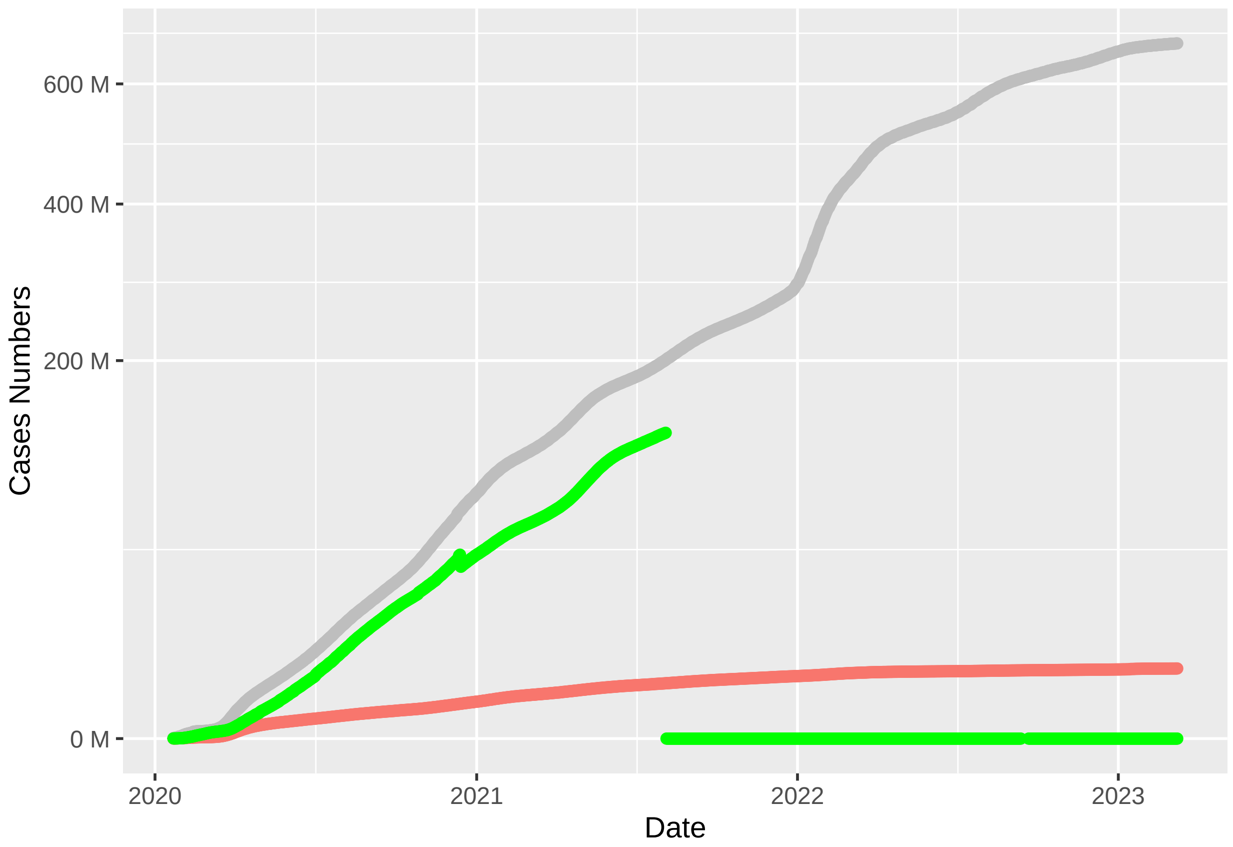
```
head(global, 10)
```

```
## # A tibble: 10 x 4
##    Date       Deaths Confirmed Recovered
##    <date>      <int>     <int>     <int>
##  1 2020-01-22     17       557        30
##  2 2020-01-23     18       657        32
##  3 2020-01-24     26       944        39
##  4 2020-01-25     42      1437        42
##  5 2020-01-26     56      2120        56
##  6 2020-01-27     82      2929        65
##  7 2020-01-28    131      5580       108
##  8 2020-01-29    133      6169       127
##  9 2020-01-30    172      8237       145
## 10 2020-01-31    214      9927       225
```

Visualization of data

Numbers of cases

```
global %>%
  ggplot() +
  ylab("Cases Numbers") +
  theme(legend.position = "none") +
  scale_y_sqrt(labels = scales::unit_format(unit = "M", scale = 1e-6)) +
  geom_point(aes(Date, Deaths, colour = 'Red')) +
  geom_point(aes(Date, Confirmed), colour = 'Gray') +
  geom_point(aes(Date, Recovered), colour = 'Green')
```

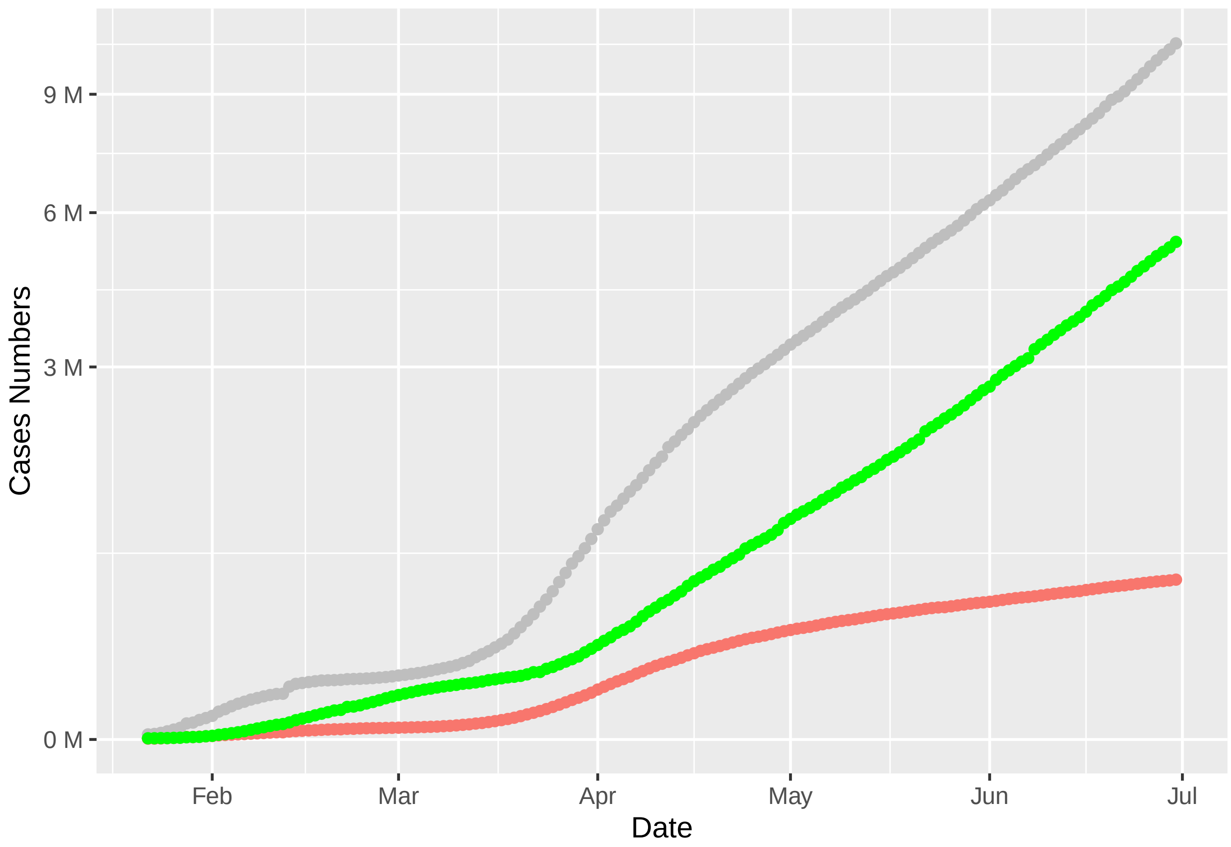


Red - deaths, Gray - confirmed, Green - recovered.

Here we can see how the number of cases change over time. Unfortunately, recovered cases data is missing after the mid of 2021.

Numbers for first 6 month

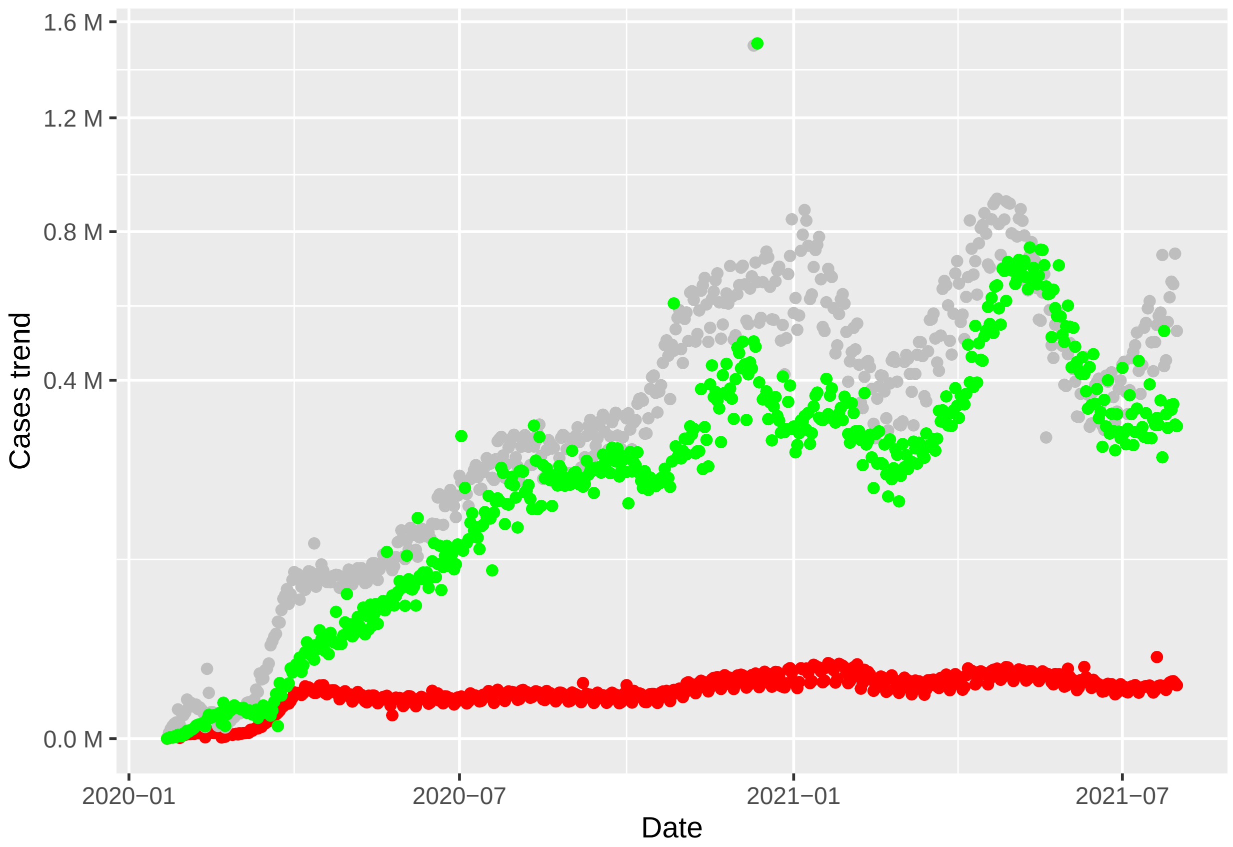
```
global %>%
  filter(Date < '2020-07-01') %>%
  ggplot() +
  ylab("Cases Numbers") +
  theme(legend.position = "none") +
  scale_y_sqrt(labels = scales::unit_format(unit = "M", scale = 1e-6)) +
  geom_point(aes(Date, Deaths, colour = 'Red')) +
  geom_point(aes(Date, Confirmed), colour = 'Grey') +
  geom_point(aes(Date, Recovered), colour = 'Green')
```



This graph shows first 6 month of pandemic.

The trends for cases numbers

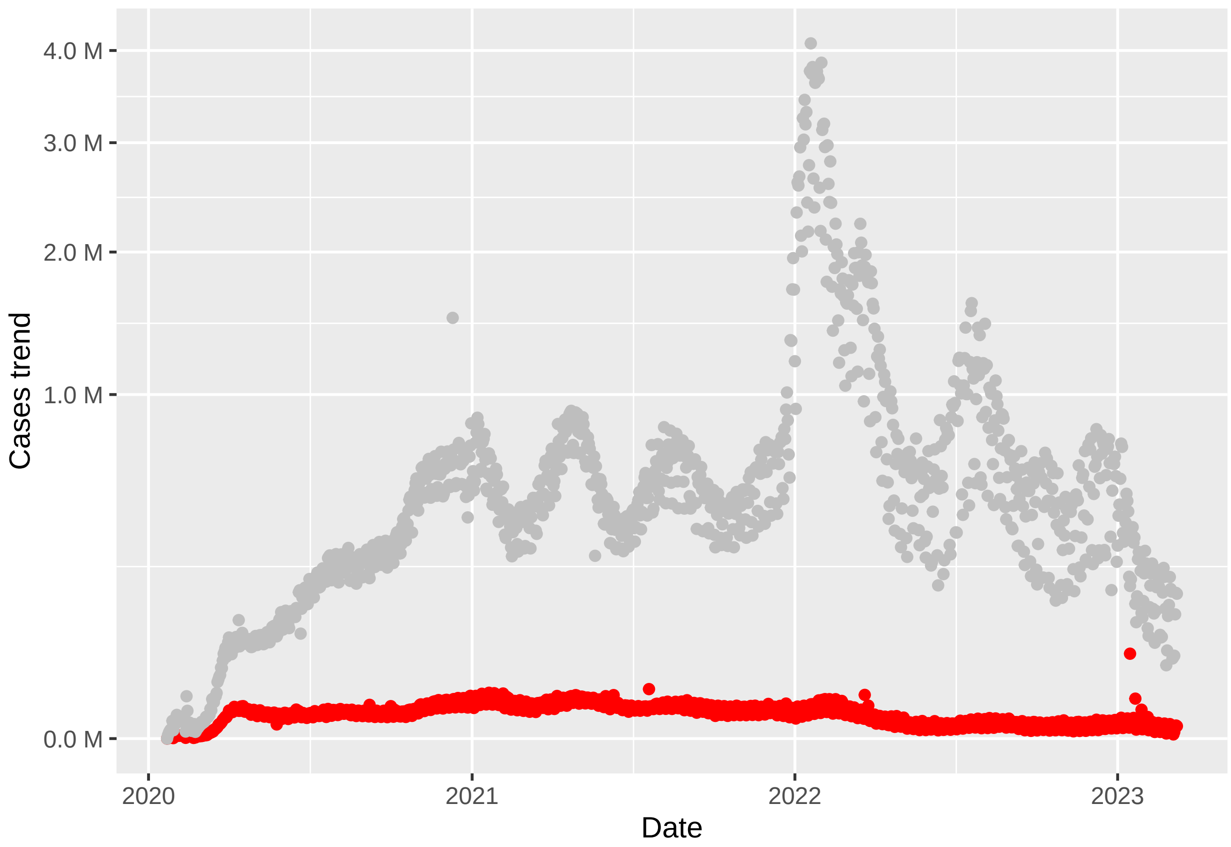
```
global %>%
  filter(Date < '2021-08-01') %>%
  mutate(trendDeaths = ifelse(is.na(lag(Deaths)), 0, Deaths - lag(Deaths))) %>%
  mutate(trendConfirmed = ifelse(is.na(lag(Confirmed)), 0, Confirmed - lag(Confirmed))) %>%
  mutate(trendRecovered = ifelse(is.na(lag(Recovered)), 0, Recovered - lag(Recovered))) %>%
  ggplot() +
  theme(legend.position = "none") +
  ylab("Cases trend") +
  scale_y_sqrt(labels = scales::unit_format(unit = "M", scale = 1e-6)) +
  geom_point(aes(Date, trendDeaths), colour = 'Red', na.rm = TRUE) +
  geom_point(aes(Date, trendConfirmed), colour = 'Gray', na.rm = TRUE) +
  geom_point(aes(Date, trendRecovered), colour = 'Green', na.rm = TRUE)
```



Red - deaths trend, gray - confirmed trend, green - recovered trend.

This graph show the trend of new cases over time until our last recovered data. Let's take a look of the entire data for confirmed cases and confirmed deaths

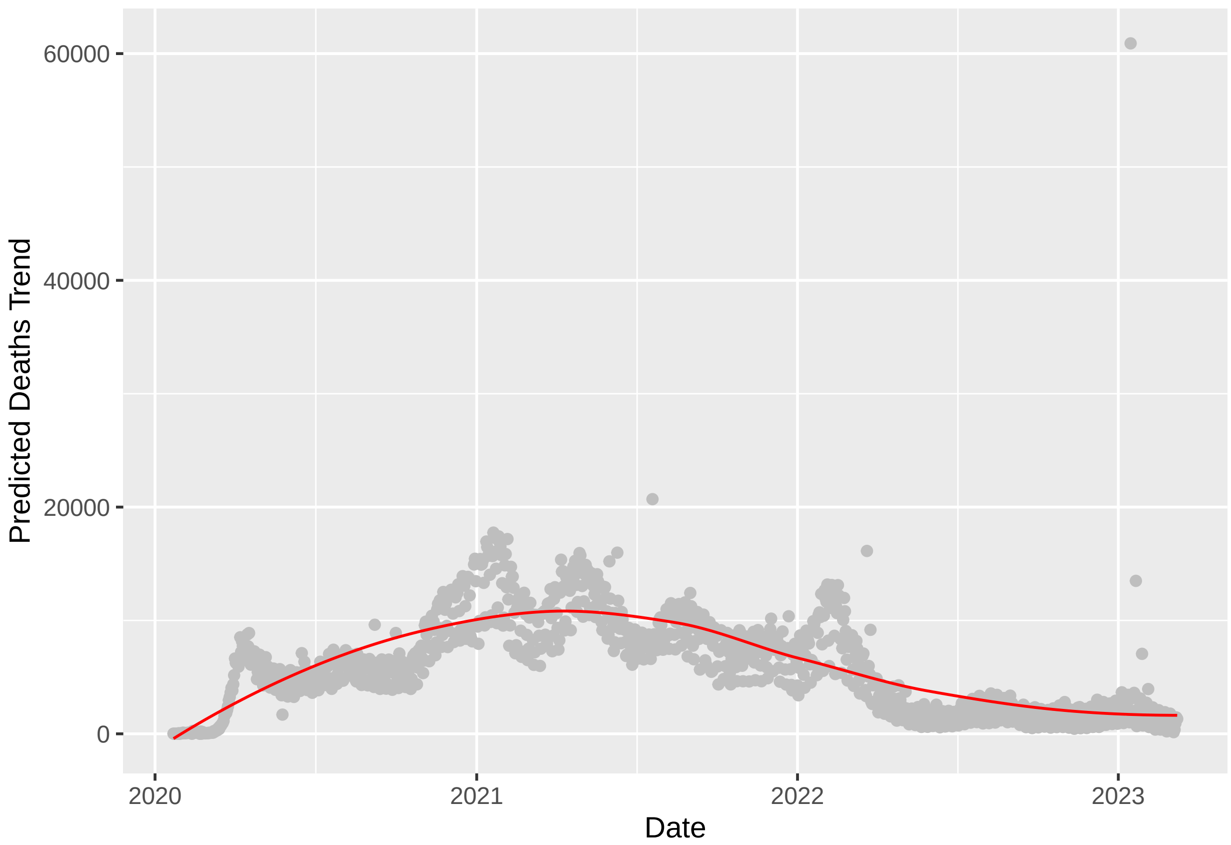
```
global %>%
  mutate(trendDeaths = ifelse(is.na(lag(Deaths)), 0, Deaths - lag(Deaths))) %>%
  mutate(trendConfirmed = ifelse(is.na(lag(Confirmed)), 0, Confirmed - lag(Confirmed))) %>%
  ggplot() +
  theme(legend.position = "none") +
  ylab("Cases trend") +
  scale_y_sqrt(labels = scales::unit_format(unit = "M", scale = 1e-6)) +
  geom_point(aes(Date, trendDeaths), colour = 'Red', na.rm = TRUE) +
  geom_point(aes(Date, trendConfirmed), colour = 'Gray', na.rm = TRUE)
```



Modeling covid-19 deaths trend

```
globalDeathsTrend <- global %>%
  mutate(trendDeaths = ifelse(is.na(lag(Deaths)), Deaths, Deaths - lag(Deaths)))
mod <- loess(trendDeaths ~ as.numeric(Date), globalDeathsTrend)
globalDeathsTrend <- globalDeathsTrend %>%
  mutate(predictedDeathsTrend = predict(mod))

globalDeathsTrend %>%
  ggplot() +
  ylab("Predicted Deaths Trend") +
  theme(legend.position = "none") +
  geom_point(aes(Date, trendDeaths), colour = 'Gray') +
  geom_line(aes(Date, predictedDeathsTrend), colour = 'Red')
```



Red line - deaths trend from the model, gray points - deaths trend from the data.

Conclusion

The model of the death trend clearly shows that although the number of deaths has always increased, the trend is going down. This is likely due to better understanding of the pandemic and vaccine availability, reducing the number of new cases.

Biases identified

It is unclear if the data collection in different countries followed the same rules/standards.