Covid-19 Report

Sergey Ostrovsky

5/31/2021

Contents

1 Introduction

The data Covid-19 contain region, data, number of cases and deaths, and population. My primary goal is to analyze the relationship between covid-19 cases and deaths and the percentage of cases and deaths based on population.

2 Importing Data

2.0.0.0.1 First, I will import the libraries to use for the report.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: ’lubridate’

## The following objects are masked from ’package:base’:
##
##     date, intersect, setdiff, union
```

```
options(warn=-1)
```

2.0.0.0.2 Now I can load Covid-19 Data from https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse__covid__19__data/csse__covid__19__time__series/ link.

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi
file_names <- c(’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 <- str_c(url_in, file_names)

US_cases <- read_csv(urls[1])
```

```
##
## -- 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(urls[2])
```

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

```
US_deaths <- read_csv(urls[3])
```

```
##
## -- 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_deaths <- read_csv(urls[4])
```

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

3 Tidying and Transforming Data

3.0.0.0.1 First I will transfer data rows into columns for global_cases and global_deaths tables

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

3.0.0.0.2 Next I will join global_cases an globas_deaths

```
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:137448      Length:137448      Min.   :2020-01-22   Min.   :       0
##  Class :character   Class :character   1st Qu.:2020-05-25   1st Qu.:      91
##  Mode  :character   Mode  :character   Median :2020-09-26   Median :    1394
##                                        Mean   :2020-09-26   Mean   :  195968
##                                        3rd Qu.:2021-01-29   3rd Qu.:   29793
##                                        Max.   :2021-06-02   Max.   :33307363
##      deaths
##  Min.   :     0
##  1st Qu.:     1
##  Median :    22
##  Mean   :  4645
##  3rd Qu.:   520
##  Max.   :595833
```

3.0.0.0.3 The last step for global is to select the only recods where cases are greater then zero.

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

```
##  Province_State     Country_Region          date                cases
##  Length:123156      Length:123156      Min.   :2020-01-22   Min.   :       1
##  Class :character   Class :character   1st Qu.:2020-06-25   1st Qu.:     257
##  Mode  :character   Mode  :character   Median :2020-10-19   Median :    2494
##                                        Mean   :2020-10-17   Mean   :  218709
##                                        3rd Qu.:2021-02-10   3rd Qu.:   43685
##                                        Max.   :2021-06-02   Max.   :33307363
##      deaths
##  Min.   :     0
##  1st Qu.:     2
##  Median :    45
##  Mean   :  5184
##  3rd Qu.:   742
##  Max.   :595833
```

3.0.0.0.4 Let's check see what data displayed

```
global %>% filter(cases > 28000000)
```

```
## # A tibble: 108 x 5
##    Province_State Country_Region date          cases deaths
##    <chr>          <chr>          <date>        <dbl>  <dbl>
##  1 <NA>           India          2021-05-30 28047534 329100
##  2 <NA>           India          2021-05-31 28175044 331895
##  3 <NA>           India          2021-06-01 28307832 335102
##  4 <NA>           India          2021-06-02 28441986 337989
##  5 <NA>           US             2021-02-19 28048511 498162
##  6 <NA>           US             2021-02-20 28120119 499981
##  7 <NA>           US             2021-02-21 28177280 501232
##  8 <NA>           US             2021-02-22 28233431 502556
##  9 <NA>           US             2021-02-23 28305709 504830
## 10 <NA>           US             2021-02-24 28380445 508005
## # ... with 98 more rows
```

3.0.0.0.5 Now we repeat the same procedure as above for the US_cases and US_deaths

```
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: 1,664,316 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
## # ... with 1,664,306 more rows
```

```
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_))
US_deaths
```

```
## # A tibble: 1,664,316 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
## # ... with 1,664,306 more rows, and 1 more variable: deaths <dbl>
```

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

```
## Joining, by = c("Admin2", "Province_State", "Country_Region", "Combined_Key", "date")
```

3.0.0.0.6 The difference between US and global is that global do not have poplulaton column. Thus, I will download extra table which contains population column and add it to global table.

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

```
## # A tibble: 123,156 x 6
##    Combined_Key Province_State Country_Region date       cases deaths
##    <chr>        <chr>          <chr>          <date>     <dbl>  <dbl>
##  1 Afghanistan  <NA>           Afghanistan    2020-02-24     1      0
##  2 Afghanistan  <NA>           Afghanistan    2020-02-25     1      0
##  3 Afghanistan  <NA>           Afghanistan    2020-02-26     1      0
##  4 Afghanistan  <NA>           Afghanistan    2020-02-27     1      0
##  5 Afghanistan  <NA>           Afghanistan    2020-02-28     1      0
##  6 Afghanistan  <NA>           Afghanistan    2020-02-29     1      0
##  7 Afghanistan  <NA>           Afghanistan    2020-03-01     1      0
##  8 Afghanistan  <NA>           Afghanistan    2020-03-02     1      0
##  9 Afghanistan  <NA>           Afghanistan    2020-03-03     2      0
## 10 Afghanistan  <NA>           Afghanistan    2020-03-04     4      0
## # ... with 123,146 more rows
```

```
uid_lookup_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/U
uid <- read_csv(uid_lookup_url) %>%
    select(-c(Lat, Long_, Combined_Key, code3, iso2, iso3, Admin2))
```

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

```
## # A tibble: 123,156 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     1      0   38928341 Afghanistan
##  2 <NA>           Afghanistan    2020-02-25     1      0   38928341 Afghanistan
##  3 <NA>           Afghanistan    2020-02-26     1      0   38928341 Afghanistan
##  4 <NA>           Afghanistan    2020-02-27     1      0   38928341 Afghanistan
##  5 <NA>           Afghanistan    2020-02-28     1      0   38928341 Afghanistan
##  6 <NA>           Afghanistan    2020-02-29     1      0   38928341 Afghanistan
##  7 <NA>           Afghanistan    2020-03-01     1      0   38928341 Afghanistan
##  8 <NA>           Afghanistan    2020-03-02     1      0   38928341 Afghanistan
##  9 <NA>           Afghanistan    2020-03-03     2      0   38928341 Afghanistan
## 10 <NA>           Afghanistan    2020-03-04     4      0   38928341 Afghanistan
## # ... with 123,146 more rows
```

4 Visualizing Data

4.0.0.0.1 Let’s visualize the data that shows number of cases and deaths per date in each country regiion

```
US_by_state <- US %>%
    group_by(Province_State, Country_Region, date) %>%
    summarise(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 ‘
```

```
US_by_state
```

```
## # A tibble: 28,884 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
## # ... with 28,874 more rows, and 1 more variable: Population <dbl>
```

```
US_totals <- US_by_state %>%
    group_by(Country_Region, date) %>%
    summarise(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.
```

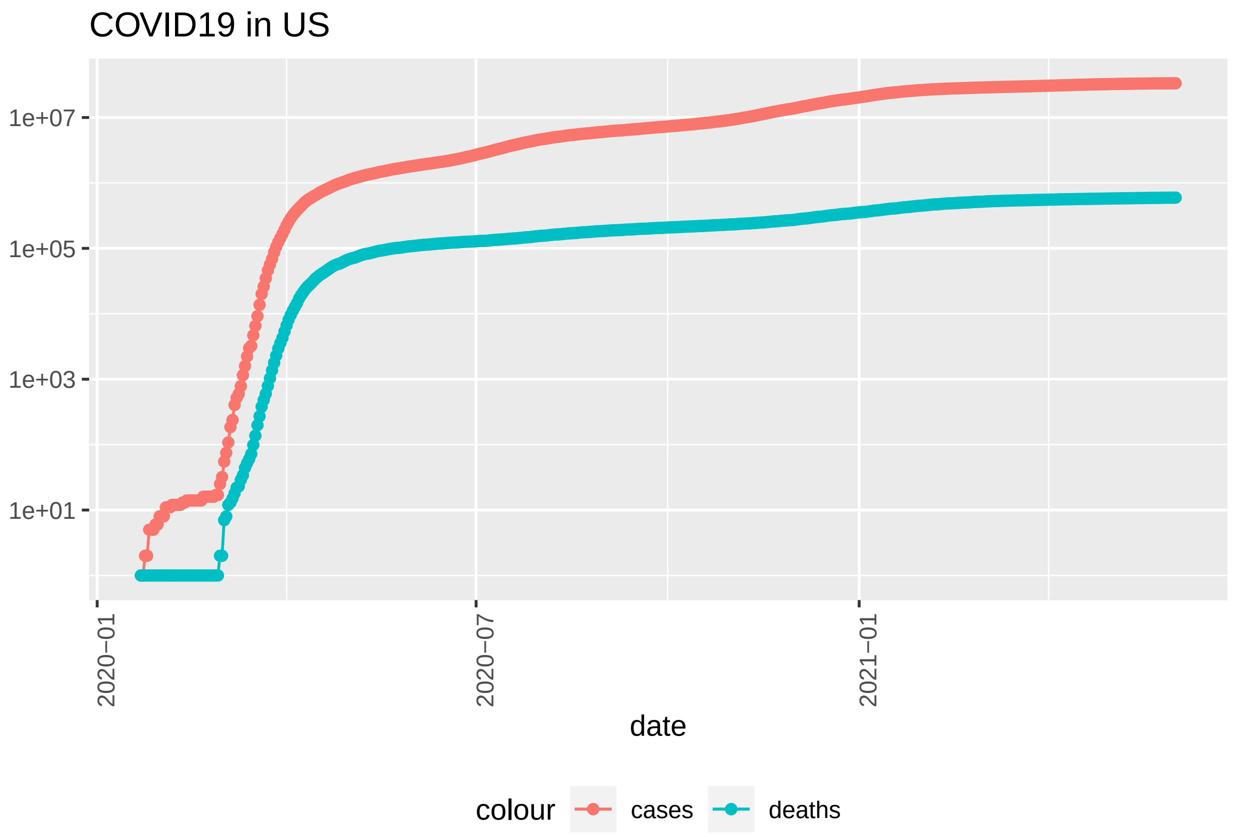
```
US_totals
```

```
## # A tibble: 498 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
## # ... with 488 more rows
```

```
tail(US_totals)
```

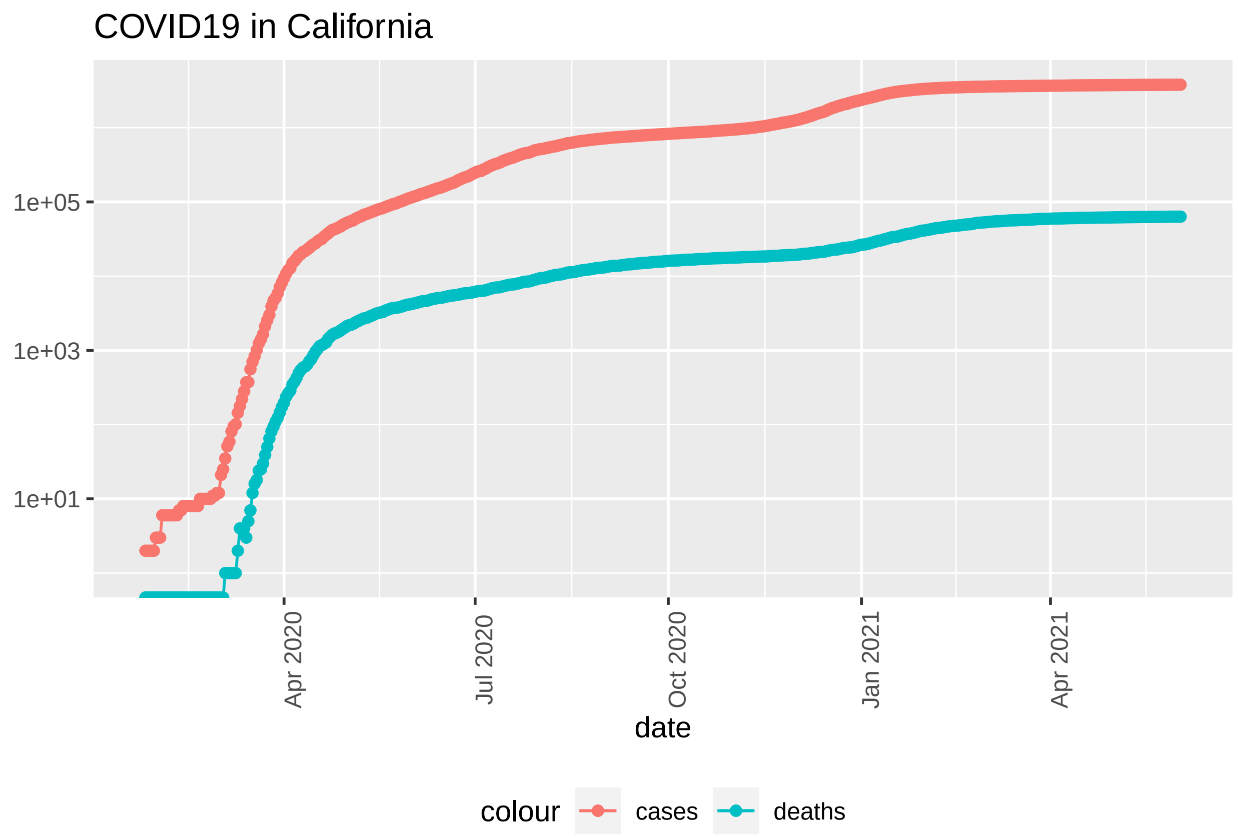
```
## # A tibble: 6 x 6
##   Country_Region date          cases deaths deaths_per_mill Population
##   <chr>          <date>        <dbl>  <dbl>           <dbl>      <dbl>
## 1 US             2021-05-28 33242999 594976           1784.  332875137
## 2 US             2021-05-29 33254998 594319           1785.  332875137
## 3 US             2021-05-30 33261731 594443           1786.  332875137
## 4 US             2021-05-31 33267507 594585           1786.  332875137
## 5 US             2021-06-01 33290450 595223           1788.  332875137
## 6 US             2021-06-02 33307363 595833           1790.  332875137
```

```
US_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 = "COVID19 in US", y = NULL)
```



4.0.0.0.2 Next let's see the result just for California state

```
state <- "California"
US_by_state %>%
    filter(Province_State == state) %>%
    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 = str_c("COVID19 in ", state), y = NULL)
```



```
max(US_totals$deaths)
```

```
## [1] 595833
```

5 Analyzing Data

5.0.0.0.1 Let's see the relationship between US covid-19 cases and deaths

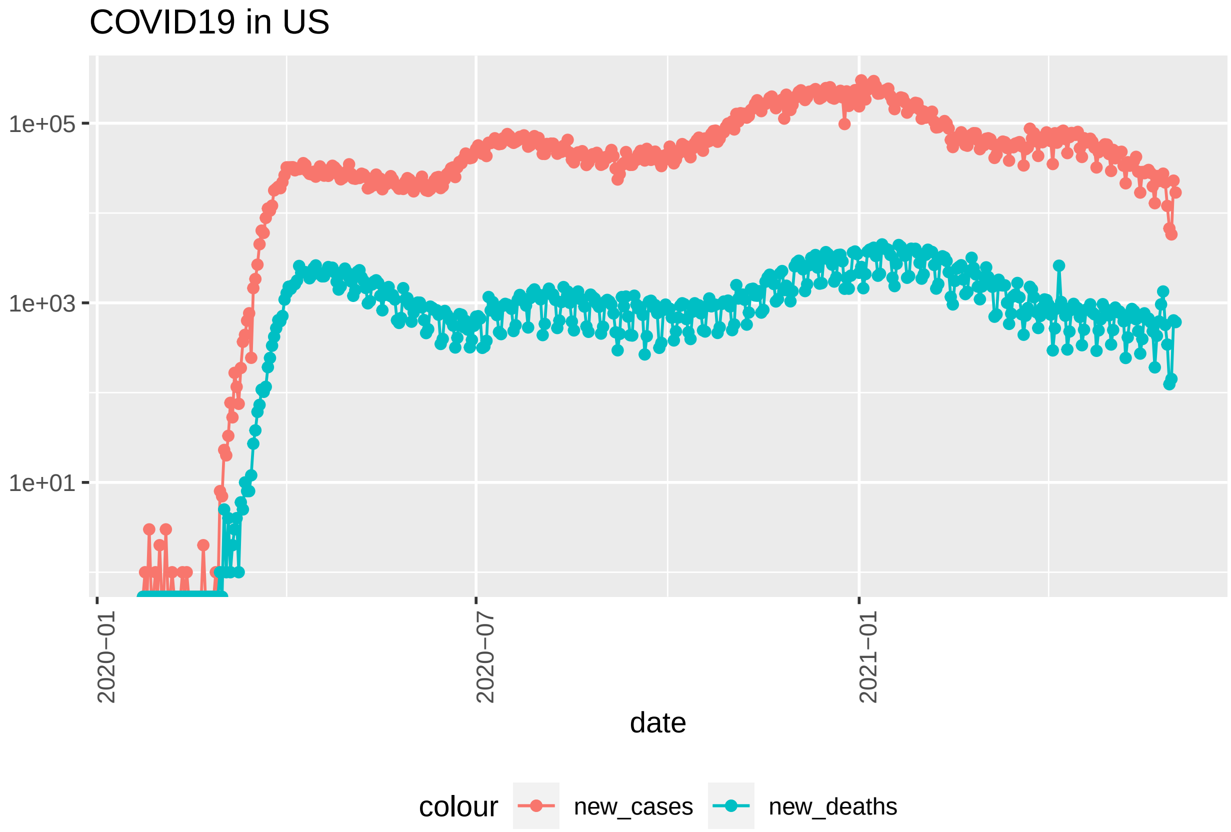
```
US_by_state <- US_by_state %>%
    mutate(new_cases = cases - lag(cases),
           new_deaths = deaths -lag(deaths))
US_totals <- US_totals %>%
    mutate(new_cases = cases - lag(cases),
           new_deaths = deaths -lag(deaths))
tail(US_totals)
```

```
## # A tibble: 6 x 8
##   Country_Region date          cases deaths deaths_per_mill Population new_cases
##   <chr>          <date>        <dbl>  <dbl>           <dbl>      <dbl>     <dbl>
## 1 US             2021-05-28 33242999 593976           1784.  332875137     21858
## 2 US             2021-05-29 33254998 594319           1785.  332875137     11999
## 3 US             2021-05-30 33261731 594443           1786.  332875137      6733
## 4 US             2021-05-31 33267507 594585           1786.  332875137      5776
## 5 US             2021-06-01 33290450 595223           1788.  332875137     22943
## 6 US             2021-06-02 33307363 595833           1790.  332875137     16913
## # ... with 1 more variable: new_deaths <dbl>
```

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

```
## # A tibble: 6 x 8
##   new_cases new_deaths Country_Region date          cases deaths deaths_per_mill
##       <dbl>      <dbl> <chr>          <date>        <dbl>  <dbl>           <dbl>
## 1     21858        567 US             2021-05-28 33242999 593976           1784.
## 2     11999        343 US             2021-05-29 33254998 594319           1785.
## 3      6733        124 US             2021-05-30 33261731 594443           1786.
## 4      5776        142 US             2021-05-31 33267507 594585           1786.
## 5     22943        638 US             2021-06-01 33290450 595223           1788.
## 6     16913        610 US             2021-06-02 33307363 595833           1790.
## # ... with 1 more variable: Population <dbl>
```

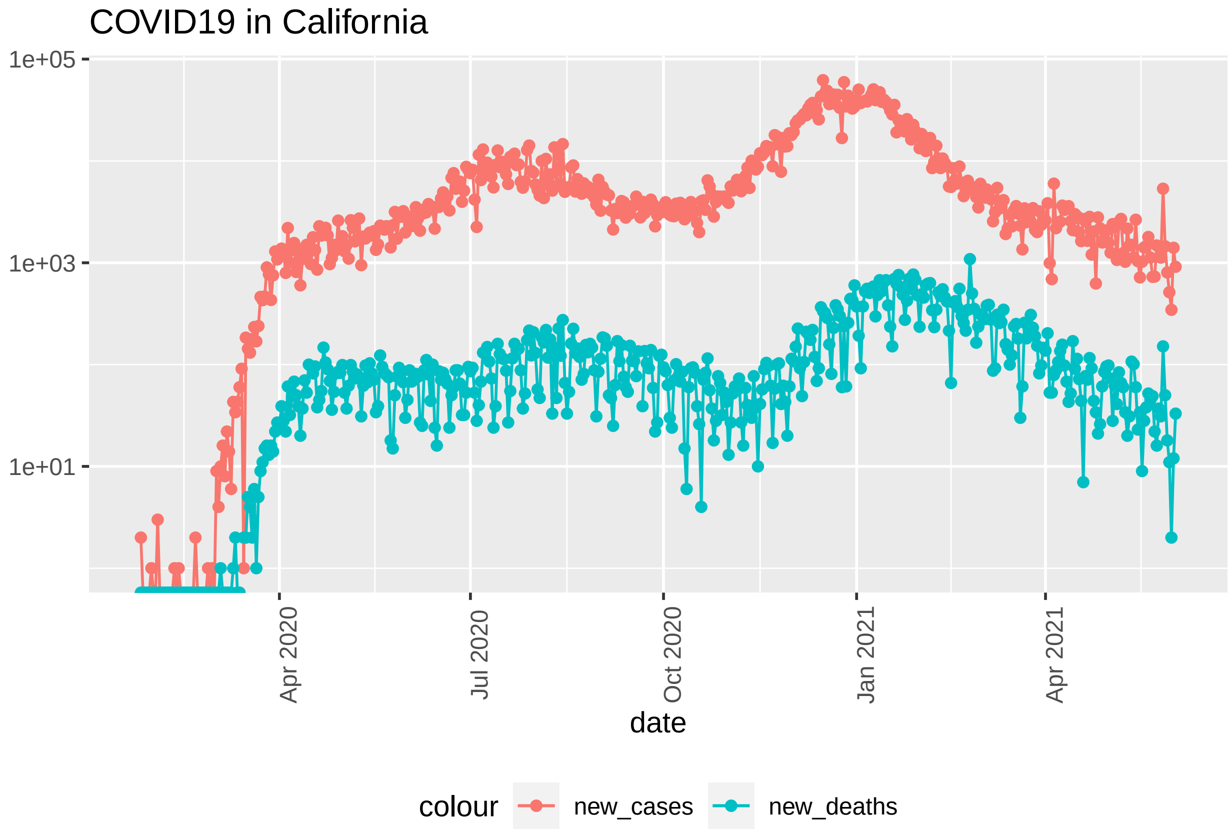
```
US_totals %>%
    ggplot(aes(x = date, y = new_cases)) +
    geom_line(aes(color = "new_cases")) +
    geom_point(aes(color = "new_cases")) +
    geom_line(aes(y = new_deaths, color = "new_deaths")) +
    geom_point(aes(y = new_deaths, color = "new_deaths")) +
    scale_y_log10() +
    theme(legend.position = "bottom",
          axis.text.x = element_text(angle = 90)) +
    labs(title = "COVID19 in US", y = NULL)
```



5.0.0.0.2 The above graph show that number of cases is proportional to number of deaths.

5.0.0.0.3 Let’s see how number of cases and deaths interact in California

```
state <- "California"
US_by_state %>%
    filter(Province_State == state) %>%
    filter(cases > 0) %>%
    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 = str_c("COVID19 in ", state), y = NULL)
```



```
US_state_totals <- US_by_state %>%
    group_by(Province_State) %>%
    summarise(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)

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          0.0363           3.32 Northern Mariana Isl~      2    183      55144
##  2          0.261           32.7  Virgin Islands            28   3512     107268
##  3          0.353           25.7  Hawaii                   500  36357    1415872
##  4          0.409           38.8  Vermont                  255  24232     623989
##  5          0.498           94.9  Alaska                   369  70355     740995
##  6          0.615           50.5  Maine                    827  67881    1344212
##  7          0.634           47.9  Oregon                  2676 201998    4217737
##  8          0.669           37.0  Puerto Rico             2512 138799    3754939
##  9          0.719          127.   Utah                    2305 406482    3205958
## 10          0.762           57.5  Washington              5801 437677    7614893
```

```
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            2.96          114.  New Jersey      26247 1016763    8882190
##  2            2.74          108.  New York        53338 2103269   19453561
##  3            2.59          103.  Massachusetts   17886  707265    6892503
##  4            2.56          143.  Rhode Island     2712  151895    1059361
##  5            2.46          107.  Mississippi      7322  317856    2976149
##  6            2.42          121.  Arizona         17648  882369    7278717
##  7            2.31           97.5 Connecticut      8247  347678    3565287
##  8            2.28          140.  South Dakota     2019  124227     884659
##  9            2.28          102.  Louisiana       10595  472304    4648794
## 10            2.28          111.  Alabama         11167  544598    4903185
```

5.0.0.0.4 The result for California is very similar to the result of the USA.

6 Modeling Data

6.0.0.0.1 To see a better picture I would like to see correlation between deaths an cases.

```
mod <- lm(deaths_per_thou ~ cases_per_thou, data = US_state_totals)
summary(mod)
```

```
##
## Call:
## lm(formula = deaths_per_thou ~ cases_per_thou, data = US_state_totals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.39513 -0.22236 -0.02912  0.19287  1.04787
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.016156   0.209062  -0.077    0.939
## cases_per_thou  0.016802   0.002105   7.980  1.2e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4616 on 53 degrees of freedom
## Multiple R-squared:  0.5458, Adjusted R-squared:  0.5372
## F-statistic: 63.69 on 1 and 53 DF,  p-value: 1.202e-10
```

```
US_state_totals %>% slice_min(cases_per_thou)
```

```
## # A tibble: 1 x 6
##   Province_State          deaths cases population cases_per_thou deaths_per_thou
##   <chr>                    <dbl> <dbl>      <dbl>          <dbl>           <dbl>
## 1 Northern Mariana Islan~      2   183      55144           3.32          0.0363
```

```
US_state_totals %>% slice_max(cases_per_thou)
```

```
## # A tibble: 1 x 6
##   Province_State deaths  cases population cases_per_thou deaths_per_thou
##   <chr>           <dbl>  <dbl>      <dbl>          <dbl>           <dbl>
## 1 North Dakota     1543 110045     762062           144.            2.02
```

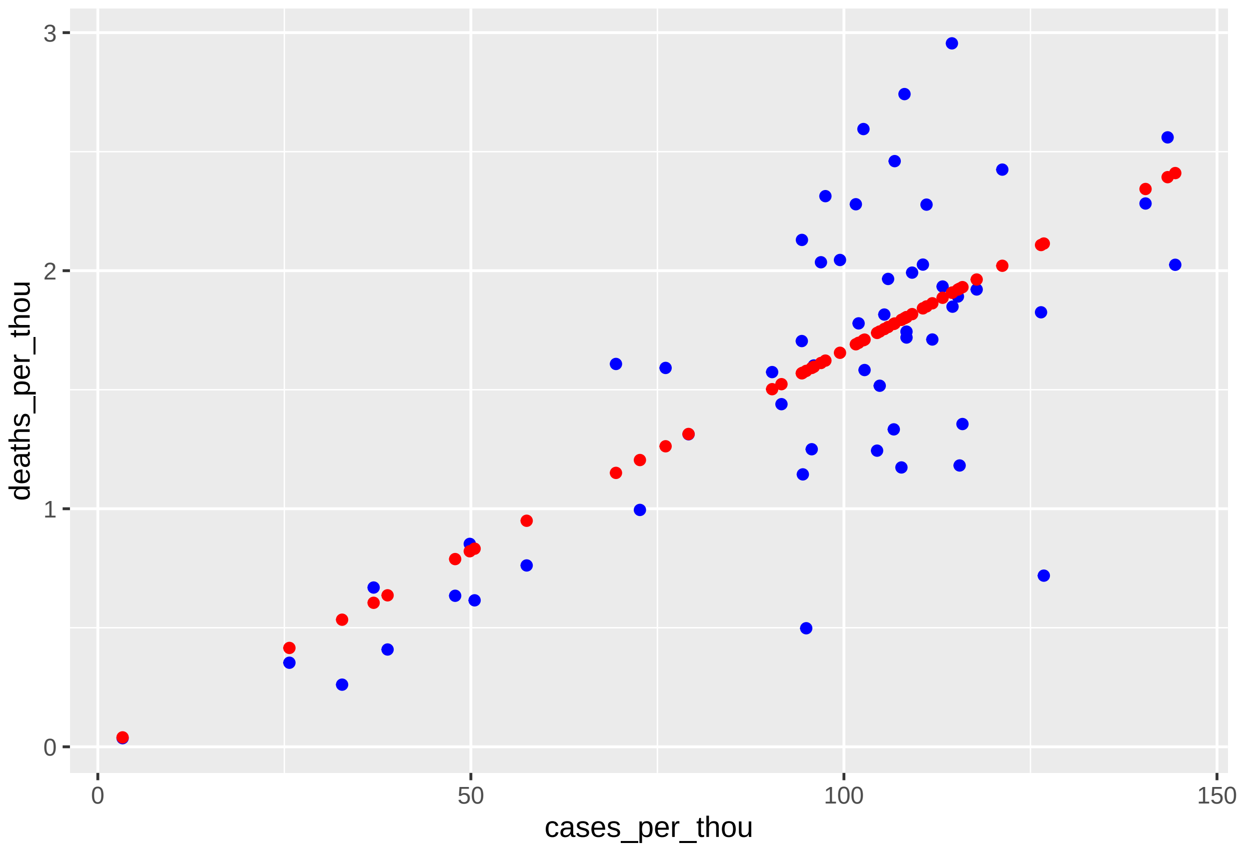
```
US_state_totals %>% mutate(pred = predict(mod))
```

```
## # A tibble: 55 x 7
##    Province_State  deaths  cases population cases_per_thou deaths_per_thou  pred
##    <chr>            <dbl>  <dbl>      <dbl>          <dbl>           <dbl> <dbl>
##  1 Alabama          11167 5.45e5    4903185          111.            2.28   1.85
##  2 Alaska             369 7.04e4     740995           94.9           0.498  1.58
##  3 Arizona          17648 8.82e5    7278717          121.            2.42   2.02
##  4 Arkansas          5835 3.42e5    3017804          113.            1.93   1.89
##  5 California       63294 3.79e6   39512223           96.0           1.60   1.60
##  6 Colorado          6590 5.44e5    5758736           94.5           1.14   1.57
##  7 Connecticut       8247 3.48e5    3565287           97.5           2.31   1.62
##  8 Delaware          1666 1.09e5     973764          112.            1.71   1.86
##  9 District of Co~   1135 4.90e4     705749           69.4           1.61   1.15
## 10 Florida          36924 2.33e6   21477737          108.            1.72   1.80
## # ... with 45 more rows
```

```
US_tot_w_pred <- US_state_totals %>% mutate(pred = predict(mod))
US_tot_w_pred
```

```
## # A tibble: 55 x 7
##    Province_State  deaths  cases population cases_per_thou deaths_per_thou  pred
##    <chr>            <dbl>  <dbl>      <dbl>          <dbl>           <dbl> <dbl>
##  1 Alabama          11167 5.45e5    4903185          111.            2.28   1.85
##  2 Alaska             369 7.04e4     740995           94.9           0.498  1.58
##  3 Arizona          17648 8.82e5    7278717          121.            2.42   2.02
##  4 Arkansas          5835 3.42e5    3017804          113.            1.93   1.89
##  5 California       63294 3.79e6   39512223           96.0           1.60   1.60
##  6 Colorado          6590 5.44e5    5758736           94.5           1.14   1.57
##  7 Connecticut       8247 3.48e5    3565287           97.5           2.31   1.62
##  8 Delaware          1666 1.09e5     973764          112.            1.71   1.86
##  9 District of Co~   1135 4.90e4     705749           69.4           1.61   1.15
## 10 Florida          36924 2.33e6   21477737          108.            1.72   1.80
## # ... with 45 more rows
```

```
US_tot_w_pred %>% ggplot() +
    geom_point(aes(x = cases_per_thou, y = deaths_per_thou),
               color = "blue") +
    geom_point(aes(x = cases_per_thou, y = pred),
               color = "red")
```



6.0.0.0.2 The graph above shows that prediction of number of deaths based on number of cases.

7 Conclusion and Bias

The analysis above shows that the number of cases plays a primary role in the number of deaths, although some points are far away from prediction. The bias of this analysis could be that it is very questionable if covid-19 caused the deaths or some other factors. Many people were tested positive, but not so many died. The deaths collected in the data source may be bais because it is possible that not Covid-19 played the primary role for the death but some prior condition of the body.