Final Project 2 Covid 19 Report

jz

2024-06-24

About COVID-19

COVID-19 (coronavirus disease 2019) is a disease caused by a virus named SARS-CoV-2. It can be very contagious and spreads quickly. Over one million people have died from COVID-19 in the United States. Also see https://www.cdc.gov/coronavirus/2019-ncov/your-health/about-covid-19.html.

This report is mainly focus on the worst and least impacted counties in California. It contains a model about cases per thousand and death per thousand for Los Angeles county. The data for the report comes from COVID 19 data hosted at Johns Hopkins Github site. In order to achieve reproducibility, this report shows each step including how to import, tidy and analyze COVID 19 data. The data contains COVID-19 global and US data sets as well as their related links and file names.

Step 0: Import Library

```
# install.packages("tidyverse")
library(tidyverse)
library(lubridate)
```

Step 1: Importing Data

```
## read_csv() reads comma delimited files, message=FALSE, warning=FALSE}
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_deaths_US.csv",
                "time_series_covid19_confirmed_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[4])
```

```
## 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[3])
```

```
## 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.
```

Step 2: Transform Covid-19 Data

Step 2.1: Transform global Covid-19 Data

```
# get global data
global_cases <-global_cases %>% pivot_longer(cols = -c(`Province/State`,`Country/Region`,Lat,Long),name

global_deaths <-global_deaths %>% pivot_longer(cols = -c(`Province/State`,`Country/Region`,Lat,Long),na

global <-global_cases %>% full_join(global_deaths) %>% rename(Country_Region=`Country/Region`,Province_
## Joining with ‘by = join_by(‘Province/State‘, ‘Country/Region‘, date)‘
```

```
head(global)
```

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

```
tail(global)
```

```
## # A tibble: 6 x 5
##   Province_State Country_Region date        cases deaths
##   <chr>          <chr>          <date>      <dbl>  <dbl>
## 1 <NA>           Zimbabwe       2023-03-04 264127   5668
## 2 <NA>           Zimbabwe       2023-03-05 264127   5668
## 3 <NA>           Zimbabwe       2023-03-06 264127   5668
## 4 <NA>           Zimbabwe       2023-03-07 264127   5668
## 5 <NA>           Zimbabwe       2023-03-08 264276   5671
## 6 <NA>           Zimbabwe       2023-03-09 264276   5671
```

```
summary(global)
```

```
##  Province_State     Country_Region          date                cases
##  Length:330327      Length:330327      Min.   :2020-01-22   Min.   :        0
##  Class :character   Class :character   1st Qu.:2020-11-02   1st Qu.:      680
##  Mode  :character   Mode  :character   Median :2021-08-15   Median :    14429
##                                        Mean   :2021-08-15   Mean   :   959384
##                                        3rd Qu.:2022-05-28   3rd Qu.:   228517
##                                        Max.   :2023-03-09   Max.   :103802702
##      deaths
##  Min.   :      0
##  1st Qu.:      3
##  Median :    150
##  Mean   :  13380
##  3rd Qu.:   3032
##  Max.   :1123836
```

Step 2.2 : Transform US Covid-19 Data

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

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

Step 2.3: Adding each country's polulation to global data

```
uid_lookup_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/
uid <- read_csv(uid_lookup_url)
```

```
## 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.
```

```
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)
```

Step 3: Visualize California data

```
CA <-US %>% filter(`Province_State`=="California") %>%
      group_by ( Province_State,Admin2) %>%
      summarize(cases = max(cases), deaths = max(deaths),Population = unique(Population))  %>%
      mutate(death_per_thou = deaths * 1000/Population ) %>%
      mutate(cases_per_thou = cases * 1000/Population ) %>%
      mutate(deaths_per_mill = cases * 100000/Population) %>%
      select(Province_State, Admin2,cases,deaths,death_per_thou,cases_per_thou,deaths_per_mill,Populati
      ungroup()
```

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

```
worst_CA <- CA %>% filter(death_per_thou >3.5 ) %>% filter(`Admin2` != "Unassigned")
g <- ggplot(data=worst_CA, mapping=aes(x = Admin2, y=death_per_thou)) +
  geom_point(size = 5) +
  geom_line(color="red")


best_CA <- CA %>% filter(death_per_thou <1.2)

g <- ggplot(data=best_CA, mapping=aes(x = Admin2, y=death_per_thou)) +
  geom_point(size = 5) +
  geom_line(color="red")
```

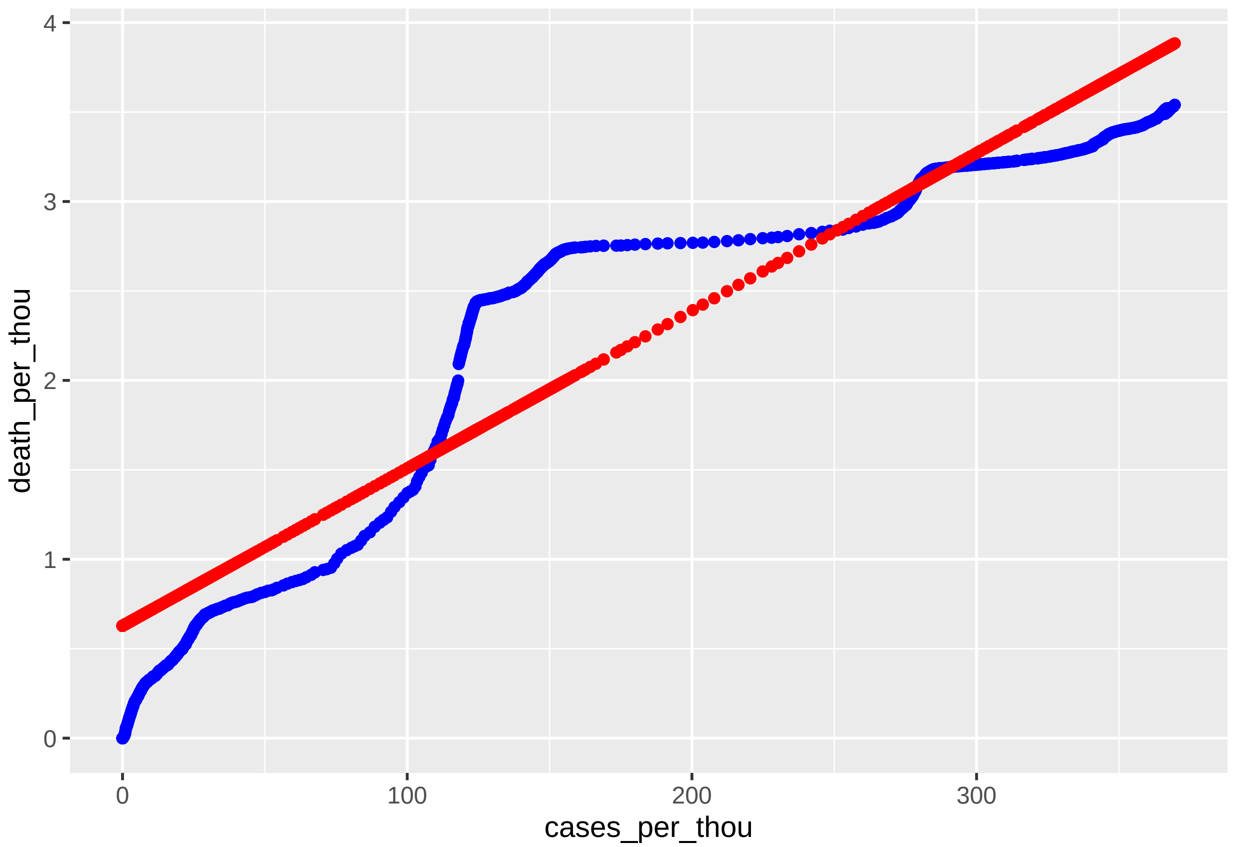
Step 3.1: Modling Los Angeles data

```
LA_daily <-US %>% filter(`Admin2` =="Los Angeles") %>%
      mutate(death_per_thou = deaths * 1000/Population) %>%
      mutate(cases_per_thou = cases * 1000/Population ) %>%
      ##select Admin2,cases,deaths,death_per_thou,cases_per_thou,Population) %>%
      ungroup()


mod <- lm(death_per_thou  ~ cases_per_thou, data=LA_daily )
summary(mod)
```

```
##
## Call:
## lm(formula = death_per_thou ~ cases_per_thou, data = LA_daily)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6282 -0.3372 -0.1970  0.5840  0.7390
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.6281169  0.0220877   28.44   <2e-16 ***
## cases_per_thou 0.0088095  0.0001045   84.27   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4563 on 1141 degrees of freedom
## Multiple R-squared:  0.8616, Adjusted R-squared:  0.8615
## F-statistic:  7102 on 1 and 1141 DF,  p-value: < 2.2e-16
```

```
Us_tot_w_pred <- LA_daily  %>% mutate (pred =predict(mod))
Us_tot_w_pred %>% ggplot() +
geom_point(aes(x = cases_per_thou, y= death_per_thou), color ="blue")+
geom_point(aes(x = cases_per_thou, y= pred), color ="red")
```



Step 4: Identify Bias

I would like to address two areas of biases. One is the data. Another is interpretation of the data. From JH site https://ictr.johnshopkins.edu/wp-content/uploads/CROWN-Registry-Bead-20200512.pdf,it has a section of "Potential Biases/issues - associated generally with registrie" which helps to understand data related biases such as measurement error, missing data and much more stated in the article.

Another is the analysis focus of this report: California and Los Angeles area. I am more familiar with LA and wanted to have a close look on LA data to understand the areas worst impacted by Covid 19. To eliminate my own bias and not to fall into my person experience, the analysis starts from California and get two groups of counties with two opposite characters.

This way I can have more balanced views. I have also viewed different articles about the LA area COVID 19 analysis and had cross-referenced on their results. It seems my result is close to the other sources as LA county.

Additional Resources

- LA County COVID-19 Data
- Los Angeles County, California coronavirus cases and deaths
- Even in 2022, L.A. COVID death rate is worse than car crashes. Here's why
- COVID-19 Mortality Rates in Los Angeles County Among People Experiencing Homelessness, March 2020–February 2021