Covid 19 data

Ernesto Medina

2022-06-05

Covid-19, is significant the relation between cases and deaths?

To answer this question we'll be analyzing a data source with covid data from the beginning of reporting data.

The source of the data comes from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University: https://github.com/CSSEGISandData/COVID-19.

We will be importing the time series covid19 data.

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_cov
file_names <- c("time_series_covid19_confirmed_US.csv","time_series_covid19_deaths_US.csv","time_series_
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_
```

Import data

```
cases_us <- read_csv(urls[1])
deaths_us <- read_csv(urls[2])
global_cases <- read_csv(urls[3])
global_deaths <- read_csv(urls[4])
```

Data Description

In this document we are presenting covid 19 data from Johns Hopkins University gihub repo. We will go through a process of visual analysis and modeling. The data we import is as follows:

- Data collected daily and updated every day.
- The data consists of 3342,3342,285 and 285 rows in the datasets.
- Column names are Province/State|Country/Region|Lat/Long|and each day from 1/22/20.
- We won't need Lat Lon and will need to "tidy" the date columns to our needs.We will make each date column into a row.

Data Transformation

```
global_cases <- global_cases %>%
  pivot_longer(cols = -c(`Province/State`,`Country/Region`,Lat,Long),
               names_to = "date",
               values_to = "cases") %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat,Long))
global_deaths <- global_deaths %>%
  pivot_longer(cols = -c(`Province/State`,`Country/Region`,Lat,Long),
               names_to = "date",
               values_to = "deaths") %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat,Long))
cases_us <- cases_us %>%
  pivot_longer(cols = -c(UID:Combined_Key),
               names_to = "date",
               values_to = "cases") %>%
  mutate(date = mdy(date)) %>%
  select(Admin2:cases) %>%
  select(-c(Lat,Long_))
deaths_us <- deaths_us %>%
  pivot_longer(cols = -c(UID:Population),
               names_to = "date",
               values_to = "deaths") %>%
  mutate(date = mdy(date)) %>%
  select(Admin2:deaths) %>%
  select(-c(Lat,Long_))
```

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

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

```
us <- cases_us %>%
  full_join(deaths_us)
```

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

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

```
## Rows: 4317 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))
```

##Visualize Data

```
us_by_state <- us %>%
  group_by(Province_State,Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
                      Population = sum(Population)) %>%
  mutate(deaths_per_thou = deaths *1000/Population,
         cases_per_thou = cases *1000/Population) %>%
  select(Province_State, Country_Region, date,
         cases, deaths, deaths_per_thou, cases_per_thou, Population) %>%
  ungroup()
```

```
## ‘summarise()‘ has grouped output by ’Province_State’, ’Country_Region’. You can
## override using the ‘.groups‘ argument.
```

```
us_totals <- us_by_state %>%
  group_by(Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
                      Population = sum(Population)) %>%
  mutate(deaths_per_thou = deaths *1000/Population,
         cases_per_thou = cases *1000/Population) %>%
  select(Country_Region, date,
         cases, deaths, deaths_per_thou, cases_per_thou, Population) %>%
  ungroup()
```

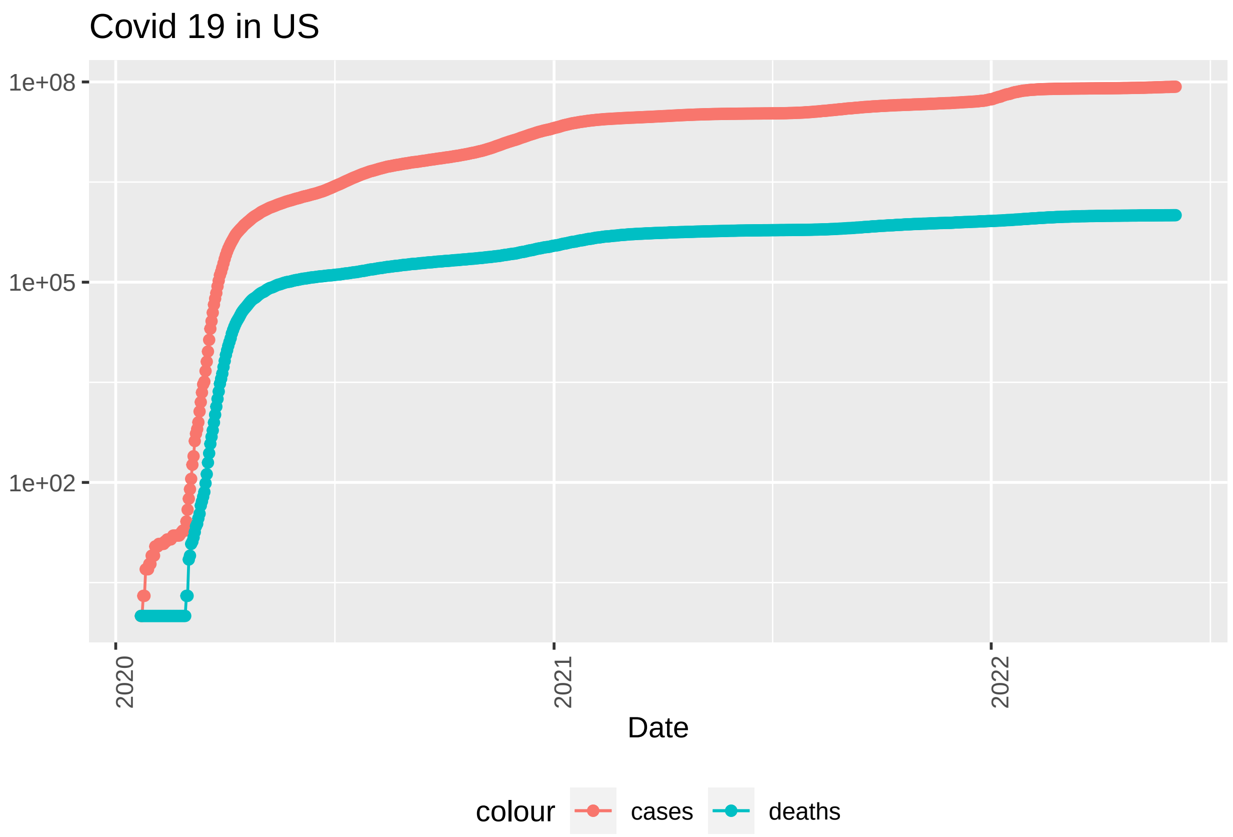
```
## ‘summarise()‘ has grouped output by ’Country_Region’. You can override using
## the ‘.groups‘ argument.
```

```
global_by_country <- global %>%
  group_by(Country_Region,date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
                      Population = sum(Population)) %>%
  mutate(deaths_per_thou = deaths *1000/Population,
         cases_per_thou = cases *1000/Population) %>%
  select(Country_Region, date,
         cases, deaths, deaths_per_thou, cases_per_thou, Population) %>%
  ungroup()
```

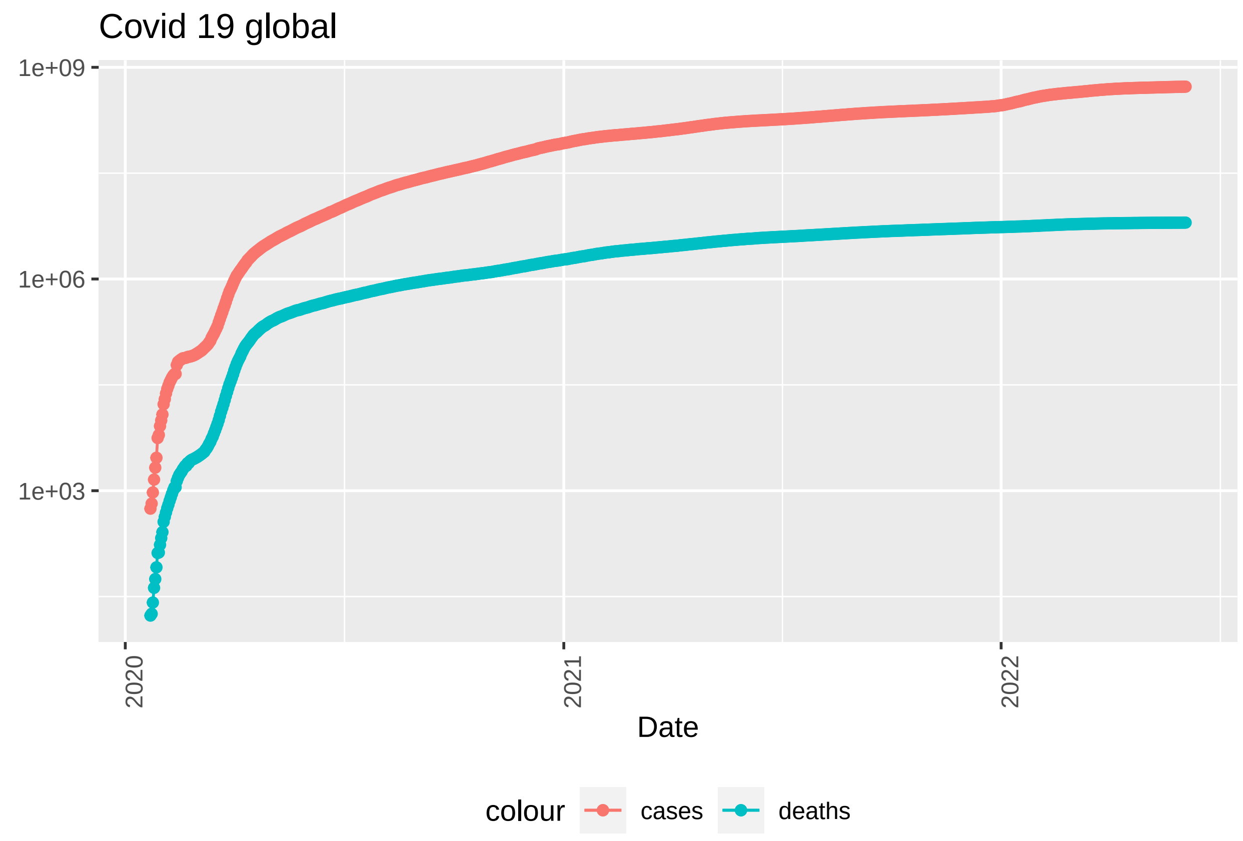
```
## ‘summarise()‘ has grouped output by ’Country_Region’. You can override using
## the ‘.groups‘ argument.
```

```r
global_total <- global %>%
  group_by(date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
                        Population = sum(Population)) %>%
  mutate(deaths_per_thou = deaths *1000/Population,
         cases_per_thou = cases *1000/Population) %>%
  select(date, cases, deaths, deaths_per_thou, cases_per_thou, Population) %>%
  ungroup()
us_state_totals <- us_by_state %>%
  group_by(Province_State) %>%
  summarize(cases = max(cases), deaths = max(deaths),
                        Population = max(Population),
            deaths_per_thou = deaths *1000/Population,
         cases_per_thou = cases *1000/Population) %>%
  filter(cases > 0, Population > 0)
```

```r
us_totals %>%
  ggplot(aes(x = date, y= cases)) +
  xlab("Date") +
  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 US", y=NULL)
```



```
global_total %>%
  ggplot(aes(x = date, y= cases)) +
  xlab("Date") +
  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 global", y=NULL)
```



## Analyzing data

```
rank_us <- us_state_totals %>%
  mutate(rank = rank(-cases), rank_deaths = rank(-deaths), rank_mill = rank(-cases_per_thou),
         rank_deaths_mill = rank(-deaths_per_thou))
rank_us <- rank_us[order(rank_us$rank),]
rank_by_death_us <- rank_us[order(rank_us$rank_deaths),]
rank_us_thousand <- rank_us[order(rank_us$rank_mill),]
rank_by_death_us_thousand <- rank_us[order(rank_us$rank_deaths_mill),]
```

The US State with more total cases: California, 9661436, 91502, 39512223, 2.31578972410639, 244.51765217057, 1, 1, 36, 39

The top 5 list:

Province_State	cases	Population	cases_per_thou	rank	rank_mill
California	9661436	39512223	244.5177	1	36
Texas	6974280	28995881	240.5266	2	38
Florida	6240440	21477737	290.5539	3	9
New York	5466873	19453561	281.0217	4	14
Illinois	3318982	12671821	261.9183	5	26

US state with more cases/thousand: Rhode Island, 394573, 1059361, 372.463211313235, 41, 1

The top 5 cases/thousand list:

Province_State	cases	Population	cases_per_thou	rank	rank_mill
Rhode Island	394573	1059361	372.4632	41	1
Alaska	262071	740995	353.6745	47	2
North Dakota	245476	762062	322.1208	48	3
Kentucky	1361744	4467673	304.7994	23	4
Tennessee	2062239	6829174	301.9749	13	5

The US State with more total deaths: California, 9661436, 91502, 39512223, 2.31578972410639, 244.51765217057, 1, 1, 36, 39

The top 5 deaths list:

Province_State	deaths	Population	deaths_per_thou	rank_deaths	rank_deaths_mill
California	91502	39512223	2.315790	1	39
Texas	88390	28995881	3.048364	2	28
Florida	74667	21477737	3.476484	3	19
New York	69124	19453561	3.553283	4	14
Pennsylvania	45254	12801989	3.534919	5	16

US state with more deaths/thousand: Mississippi, 810484, 12470, 2976149, 4.18997839153886, 272.326419140977, 31, 28, 18, 1

The top 5 deaths/thousand list:

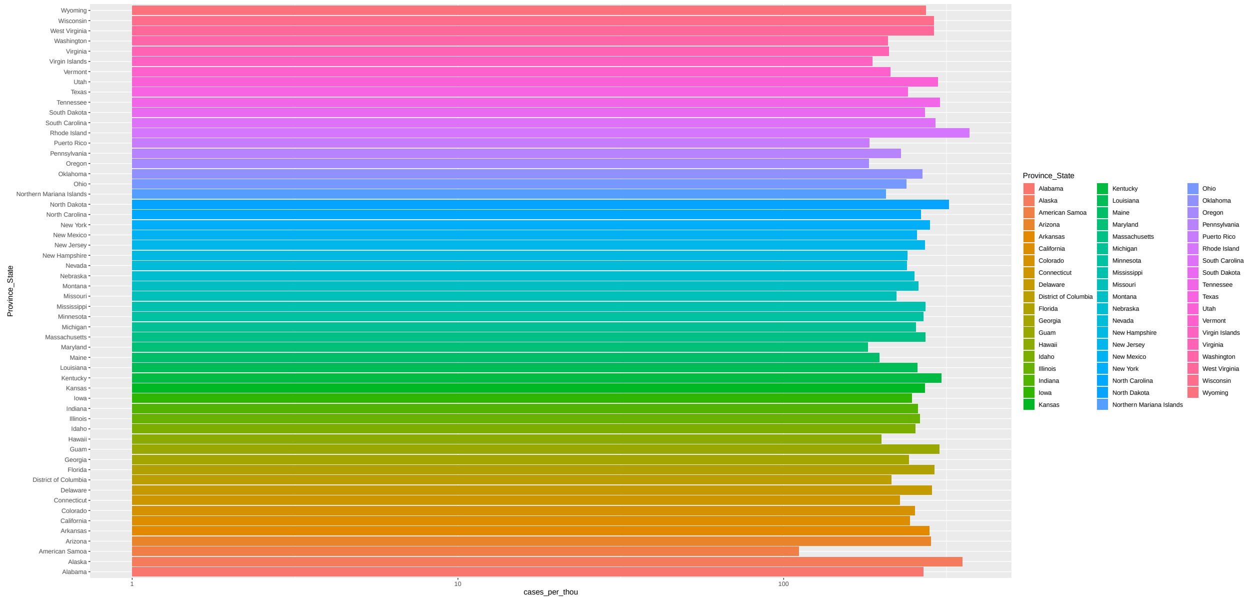
Province_State	deaths	Population	deaths_per_thou	rank_deaths	rank_deaths_mill
Mississippi	12470	2976149	4.189978	28	1
Arizona	30332	7278717	4.167218	11	2
Oklahoma	16127	3956971	4.075592	21	3
Alabama	19664	4903185	4.010454	18	4
West Virginia	6974	1792147	3.891422	36	5

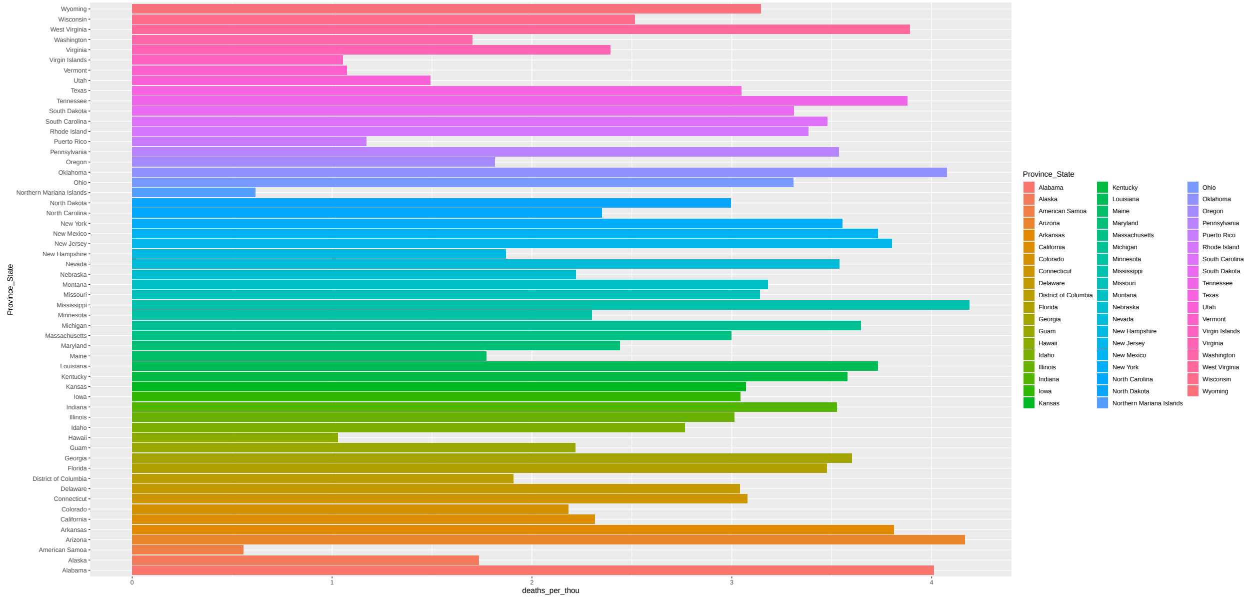
- The relation between deaths per thousand and cases per thousand don't seem obvious:

```
a <- us_state_totals %>%
     ggplot() +
     scale_y_log10() +
     geom_col(aes(x = Province_State, y = cases_per_thou, fill=Province_State)) +coord_flip()

b <- us_state_totals %>%
     ggplot() +
     geom_col(aes(x = Province_State, y = deaths_per_thou, fill=Province_State)) +coord_flip()

grid.arrange(a,b)
```





##Data Modeling

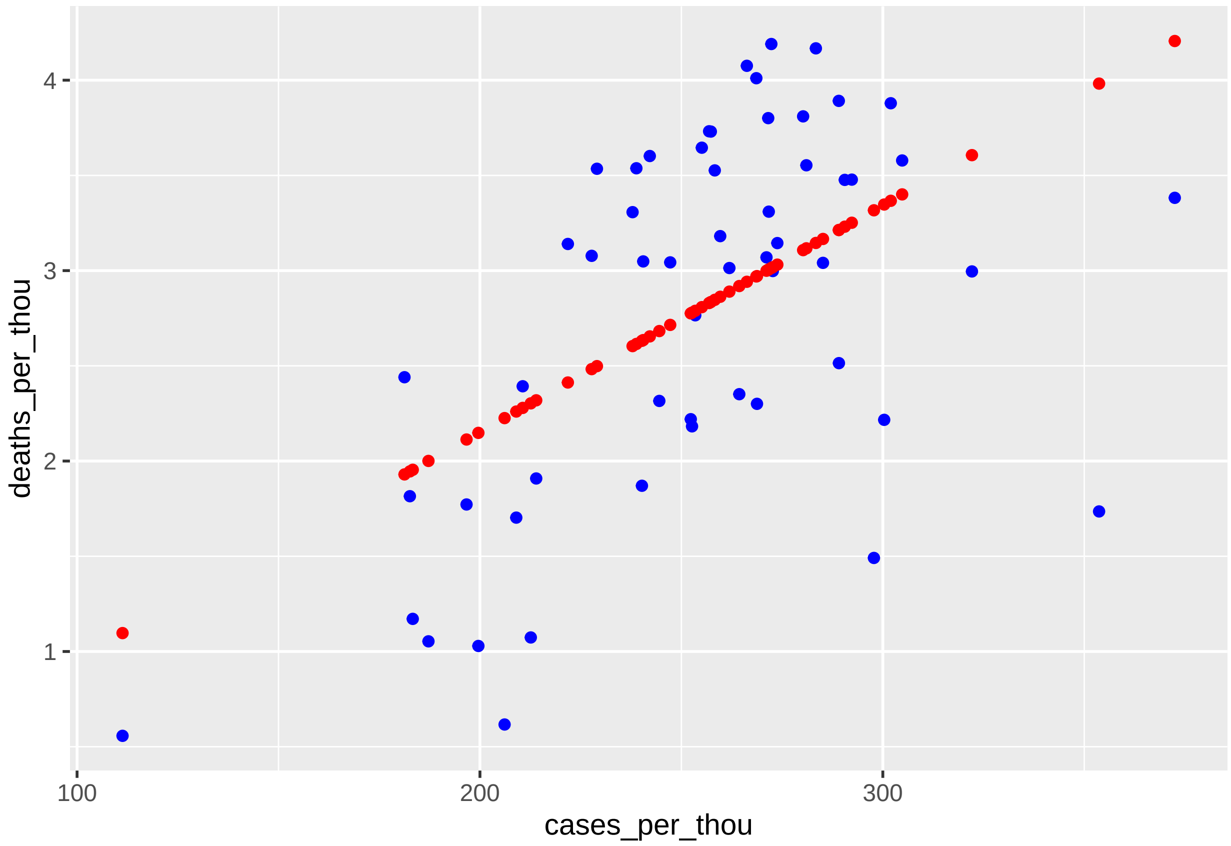
- To see if there is a relation between cases and deaths per thousand we'll use a linear model:

```
modl <- lm(deaths_per_thou~cases_per_thou, data = us_state_totals)
summary(modl)
```

```
##
## Call:
## lm(formula = deaths_per_thou ~ cases_per_thou, data = us_state_totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2466 -0.5751  0.1188  0.6854  1.1763
##
```

```
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.228082   0.637988  -0.358    0.722
## cases_per_thou  0.011904   0.002476   4.808 1.26e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.815 on 54 degrees of freedom
## Multiple R-squared:  0.2997, Adjusted R-squared:  0.2868
## F-statistic: 23.11 on 1 and 54 DF,  p-value: 1.262e-05
```

```
us_state_totals_pred <- us_state_totals %>% mutate(pred = predict(modl))
us_state_totals_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")
```



Conclussion

We've imported a dataset of Covid data from the beginning of the reported cases. By transforming the data and computing values we are able to makes plots that gives us a visual understanding of the evolution of the COVID Pandemic. We've seen how the cases grew exponentially in the beginning and get flatter along the way. There is a statistically significant relationship between cases and deaths per thousand (p-value: 1.262e-05).

Bias

I have chosen a path that directs me on the direction of Us data as i leave in the United States instead of analizing the global dataset.