Covid-19 Analysis

Tittiwat Tonburinthip

9/7/2021

COVID-19

coronavirus 2019 also known as COVID-19 is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first known case was identified in Wuhan, China in December 2019. The disease has since spread worldwide, leading to an ongoing pandemic. In this analysis, I set the list of questions as followed. (https://en.wikipedia.org/wiki/COVID-19)

1. The *number of Covid-19* patients in each country, each day, and its trend
2. In the US, which has the states that have *top 10 cases* per thousand?
3. Which is the areas that have *top 10 dead rates* per thousand?
4. How many people get fully vaccinated in each state?
5. What are the *factors* that have an *impact* on the number of Covid-19 cases and dead rates?

Collect data

- The Covid-19 data in this analysis are from JOHN HOPKINS University of Medicine

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi
file_names <- c("confirmed_global.csv",
                    "deaths_global.csv",
                    "confirmed_US.csv",
                    "deaths_US.csv")
urls <- str_c(url_in,file_names)
```

There are 6 parts.

1. global_cases is the data of Covid-19 cases from all countries around the world.
2. global_deaths is the data of the number of dead Covid-19 patients from all countries around the world.
3. US_cases is the data of Covid-19 cases in the US.
4. US_deaths is the data of the number of dead Covid-19 patients in the US.
5. vac_data is the data of daily vaccination in the US from https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/vaccinations/us_state_vaccinations.csv
6. area is the data of area in each state in the US from https://raw.githubusercontent.com/jakevdp/data-USstates/master/state-areas.csv

```
global_cases <- read_csv(urls[1], show_col_types = FALSE)
global_deaths <- read_csv(urls[2], show_col_types = FALSE)
US_cases <- read_csv(urls[3],show_col_types = FALSE)
```

```
US_deaths <- read_csv(urls[4], show_col_types = FALSE)
vac_data<- read_csv('https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/vaccinatio
area<-read_csv('https://raw.githubusercontent.com/jakevdp/data-USstates/master/state-areas.csv')
```

Cleaning data

- Use pivot_longer with the global_cases dataframe to get date from the name of columns. Then, they were stored in the table as records and the column was named as "date".

```
global_cases <- global_cases %>%
    pivot_longer(cols=-c('Province/State','Country/Region','Lat','Long'),
                 names_to = 'date', values_to = 'cases')%>%
    mutate(date=parse_date(date, format='%m/%d/%y'))%>%
    select(-c(Lat,Long))
```

- Use pivot_longer with the global_deaths dataframe to get date from the name of columns. Then, they were stored in the table as records and the column was named as "date".

```
global_deaths=global_deaths%>%
    pivot_longer(cols=-c('Province/State','Country/Region','Lat','Long'),
                 names_to = 'date', values_to = 'deaths')%>%
    mutate(date=parse_date(date, format='%m/%d/%y'))%>%
    select(-c(Lat,Long))
```

- Use full join between the global_cases dataframe and global_deaths dataframe and named it as global_dc dataframe.

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

- Check all data in global_dc with summary function.

```
summary(global_dc)
```

```
##  Province_State     Country_Region          date                cases
##  Length:168516      Length:168516      Min.   :2020-01-22   Min.   :       0
##  Class :character   Class :character   1st Qu.:2020-06-20   1st Qu.:     143
##  Mode  :character   Mode  :character   Median :2020-11-18   Median :    2266
##                                        Mean   :2020-11-18   Mean   :  283710
##                                        3rd Qu.:2021-04-18   3rd Qu.:   50994
##                                        Max.   :2021-09-16   Max.   :41785903
##      deaths
##  Min.   :     0
##  1st Qu.:     1
##  Median :    35
##  Mean   :  6553
##  3rd Qu.:   833
##  Max.   :670000
```

- Use pivot_longer with the US_cases dataframe to get date from the name of columns. Then, they were stored in the table as records and the column was named as date.

```
US_cases<-US_cases%>%
  pivot_longer(cols=-c(UID:Combined_Key),
               names_to='date', values_to='cases')%>%
  select(Admin2:cases) %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat, Long_))
```

- Use pivot_longer with the US_deaths dataframe to get date from the name of columns. Then, they were stored in the table as records and the column was named as date.

```
US_deaths<-US_deaths%>%
  pivot_longer(cols=-c(UID:Population),
               names_to='date', values_to='deaths')%>%
  select(Admin2:deaths) %>%
  mutate(date = parse_date(date, format='%m/%d/%y'))%>%
  select(-c(Lat, Long_))
```

- Connect the US_cases and US_deaths with full join function.

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

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

- Get population data via "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/UID_ISO_FIPS_LookUp_Table.csv"

```
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))
```

- global_dc, which store the number of cases and deaths from Covid-19 connected with uid with left join function.

```
global_dc<- global_dc%>%
  left_join(uid, by=c("Province_State","Country_Region"))%>%
    select(-c(UID, FIPS))
```

Analyze the data

- create new columns which are "new_cases" and "new_deaths" because the data in "cases" column and "deaths" column store data as the cumulative number of cases and dead people.

```
US_dc<- US_dc%>%
  mutate(new_cases=cases-lag(cases), new_deaths=deaths-lag(deaths))%>%
  na.omit()
```

- Group US_dc by country regions and date
- "cases" is the summation of "new_cases"
- "deaths" is the summation of "new_deaths"
- Create two new column
  - death_per_mill= deaths / population
  - case_per_mill=cases / population
- All data were stored in US_dc_1

```
US_dc_1<-US_dc%>%
  group_by(Country_Region,date)%>%
  summarise(cases=sum(new_cases), deaths=sum(new_deaths),
            Population=max(Population))%>%
  mutate(death_per_mill=deaths*1000000/Population,
         case_per_mill=cases*1000000/Population)
```

```
## Joining, by = "Province_State"
```

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

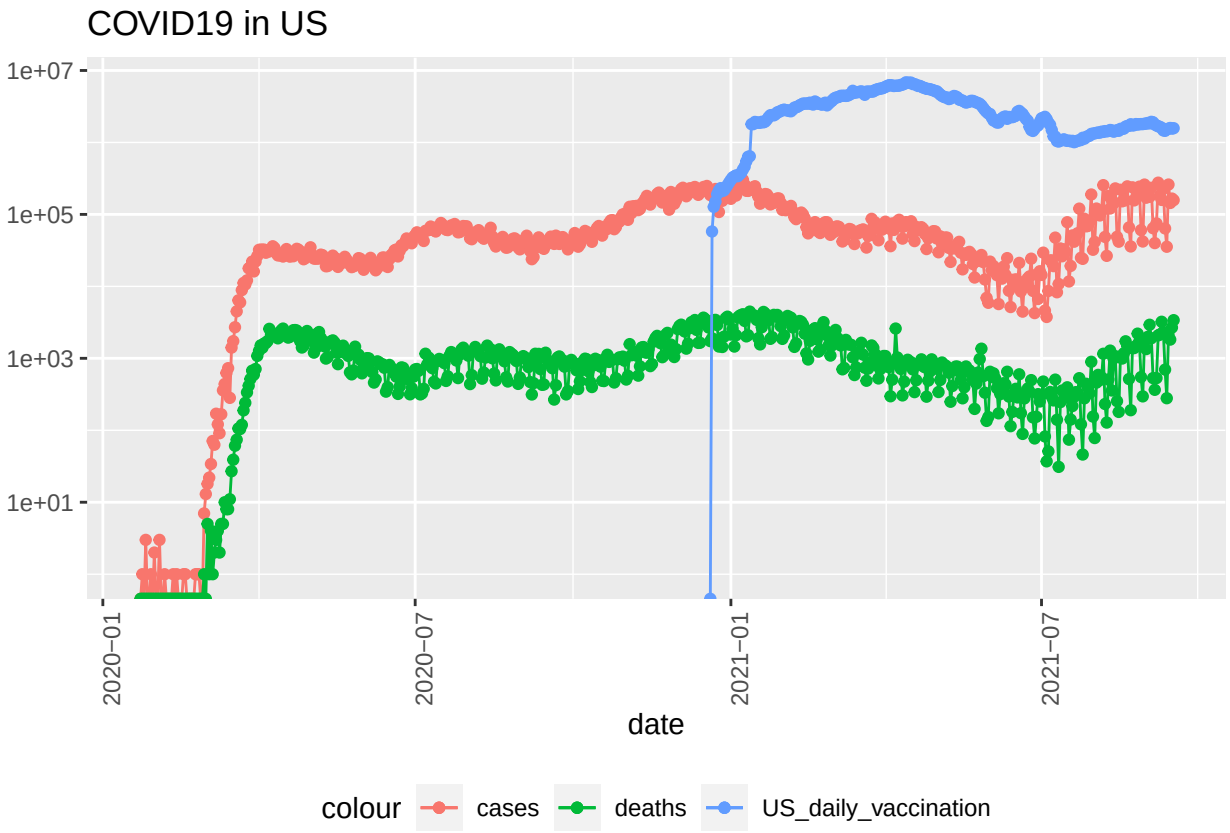
- Group vac_data that collect about Covid-19 vaccination by date
- Then, create new column named "us_daily_vaccinations" by calculating the total of daily vaccination in each day.

```
vac_data_1<-vac_data%>%
  group_by(date)%>%
  summarise(us_daily_vaccinations=sum(daily_vaccinations,na.rm = TRUE))
```

Plot

The US case

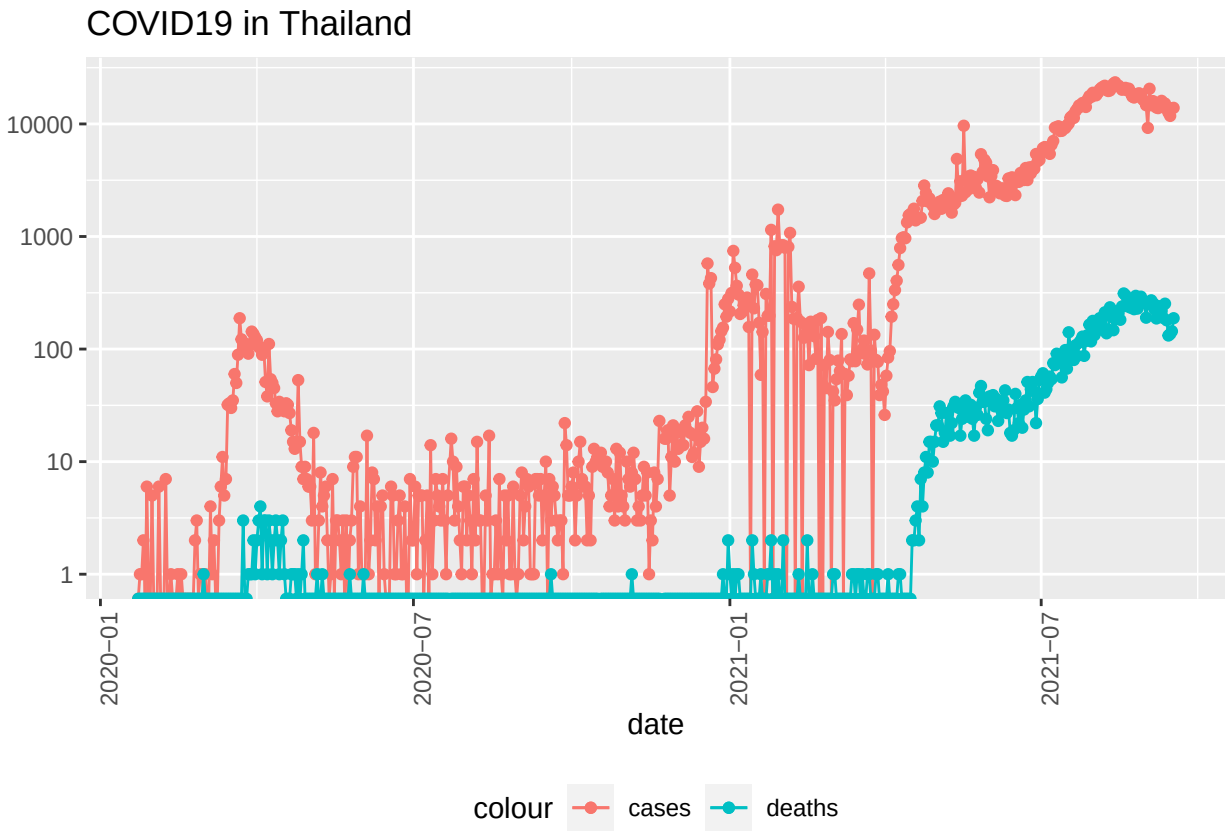
- To answer the first question, the data US_dc_1 was plotted in line, which its x-axis is date and its y-axis is the number of covid-19 cases.
- geom_line and geom_point were used.
- vac_data_1 was plotted as line as well. Its x-axis was date and y-axis was the number of daily vaccination.
- The data informed that there is an increase in Covid-19 cases and dead people since July 2021 due to a new variant of Covid-19, Delta.
- While the Covid-19 cases reduced in April 2021 - June 2021, the number of daily vaccination reduced as well.




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

Thailand case

- I compared the results of the US with my homecountry, Thailand.
- Use global_dc, filter only Thailand, group by date
- Plot in line, which its x-axis is date and y-axis are Covid-19 cases and the number of deaths.
- the slope of both cases and deaths are much steeper compared to US due to Delta variant.



US map plot for total Covid-19 cases, total dead people, and total fully vaccinated people

- US_dc was used.
- Group by province states and country region.
- "total_cases" is the summation of "new_cases".
- "total_deaths" is the summation of "new_deaths".
- "total_cases_per_thou" is total cases per thousand of population, which was "total_cases" divided by "Population".
- "total_deaths_per_thou" is total deaths per thousand of population, which was "total_deaths" divided by "Population".

```
US_dc_2<-US_dc%>%
  filter(new_cases>=0, new_deaths>=0)%>%
  group_by(Province_State, Country_Region)%>%
  summarise(total_cases=sum(new_cases), total_deaths=sum(new_deaths),
            total_cases_per_thou=total_cases*1000/max(Population),
            total_deaths_per_thou=total_deaths*1000/max(Population))%>%
  rename(state=Province_State)
```

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

Covid-19 cases in the US

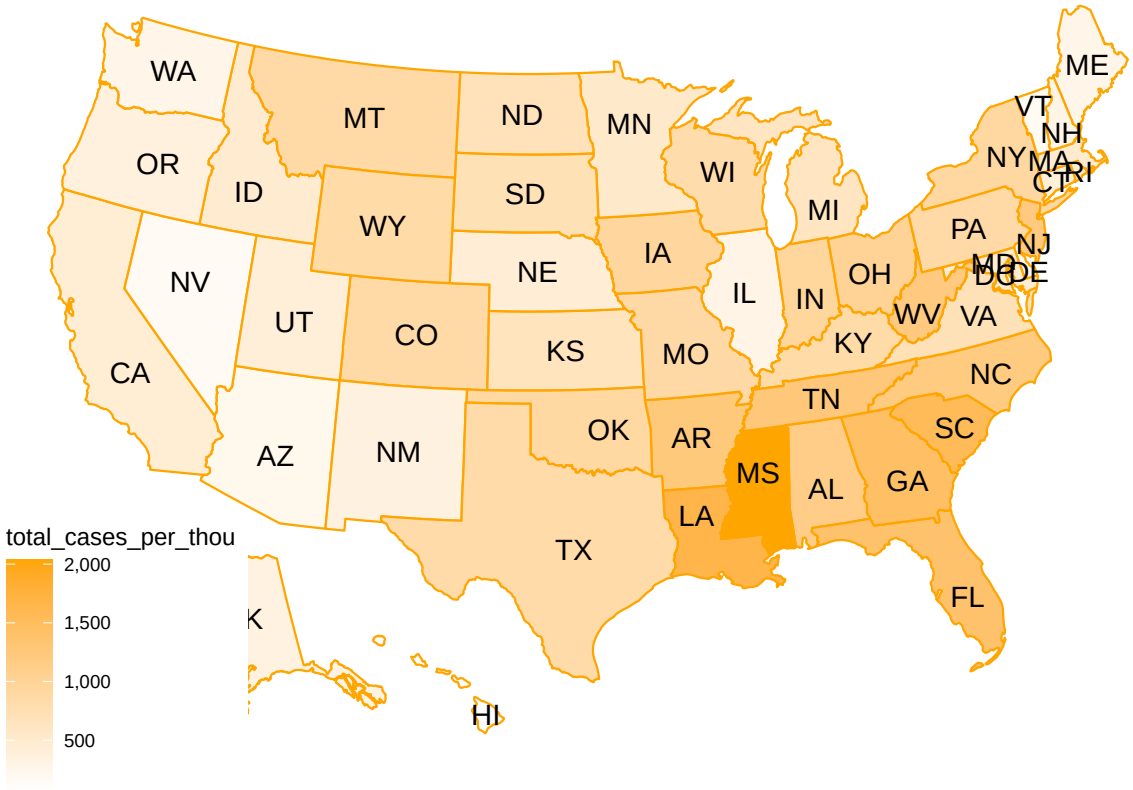
- Show top 10 max Covid-19 cases in the US.

Table 1: Table for top 10 max case per thousand

state	Country_Region	total_cases	total_deaths	total_cases_per_thou	total_deaths_per_thou
Mississippi	US	471,629	9,184	2,034.29	39.61
Louisiana	US	733,279	13,594	1,666.32	30.89
South Carolina	US	808,566	11,675	1,544.41	22.30
Georgia	US	1,533,894	25,024	1,441.72	23.52
Florida	US	3,748,191	48,797	1,379.56	17.96
Tennessee	US	1,188,312	14,968	1,267.98	15.97
Arkansas	US	484,060	7,701	1,235.13	19.65
New Jersey	US	1,134,304	28,437	1,216.80	30.51
West Virginia	US	215,750	3,610	1,211.23	20.27
North Carolina	US	1,330,727	15,891	1,196.95	14.29

US map plot for the cases in the US

- The data was shown in the US map plot.
- It's much clearer that the southern states such as Mississippi, Louisiana, and South Carolina have more cases than in the northern states.



##

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

Covid-19 death in the US

- Show top 10 max Covid-19 death in the US.

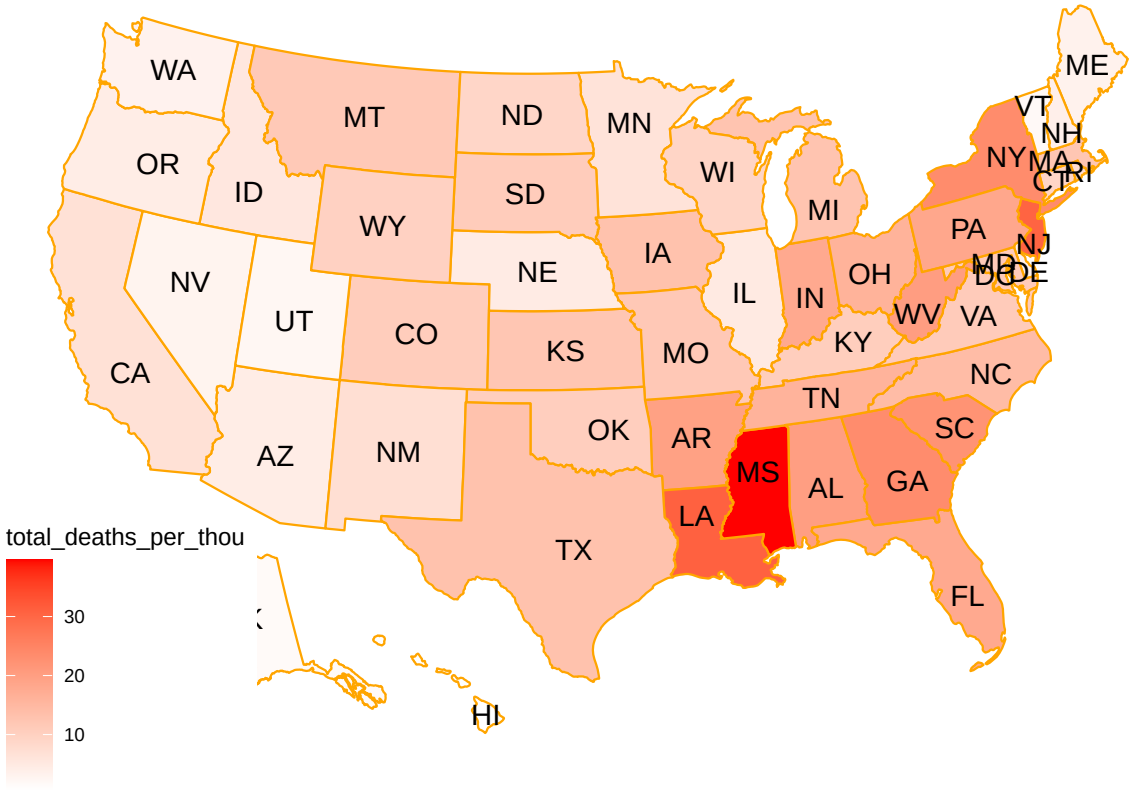
Table 2: Table for top 10 max death per thousand

state	Country_Region	total_cases	total_deaths	total_cases_per_thou	total_deaths_per_thou
Mississippi	US	471,629	9,184	2,034.29	39.61
Louisiana	US	733,279	13,594	1,666.32	30.89
New Jersey	US	808,566	11,675	1,544.41	30.51
New York	US	1,533,894	25,024	1,441.72	23.56
Georgia	US	3,748,191	48,797	1,379.56	23.52
South Carolina	US	1,188,312	14,968	1,267.98	22.30
West Virginia	US	484,060	7,701	1,235.13	20.27
Alabama	US	1,134,304	28,437	1,216.80	20.06
Arkansas	US	215,750	3,610	1,211.23	19.65
Pennsylvania	US	1,330,727	15,891	1,196.95	18.39

US map plot for the death in the US

- The data was shown in the US map plot.
- It's much clearer that the southern states such as Mississippi, Louisiana, and South Carolina has a higher death rate than in the northern states.
- However, New York and New Jersey still have high death rates in the same level as the southern state.

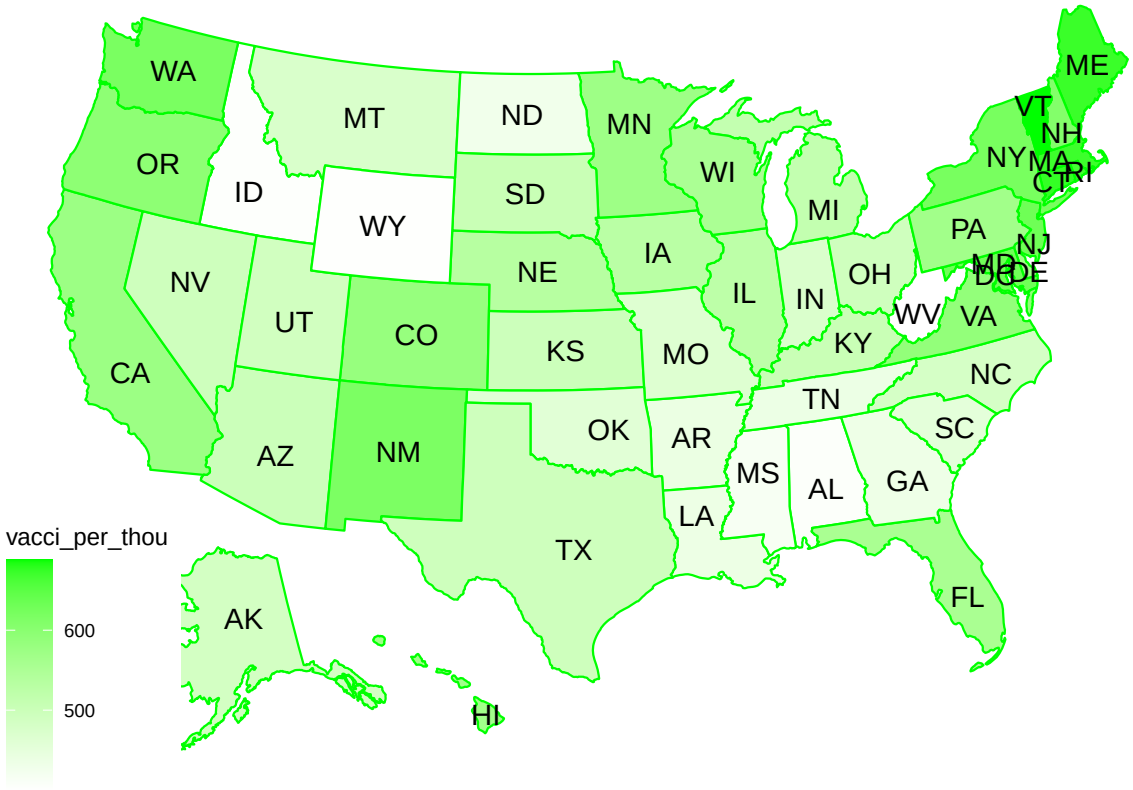
```
plot_usmap(data=US_dc_3, values='total_deaths_per_thou', color='orange', labels = TRUE)+
  scale_fill_continuous(low = "white", high = "red", name = "total_deaths_per_thou", label = scales::co
```



#vaccination rate

US map plot for the vaccination in the US

- Total vaccination was plotted in US map.
- The results showed that the northern states have a much higher vaccination rate.



- Looking back to see the percentage of fully vaccinated people in the southern states.
- They have only 40% of total people.
- if more people are eager to have vaccination, the number of Covid-19 cases can be reduced.

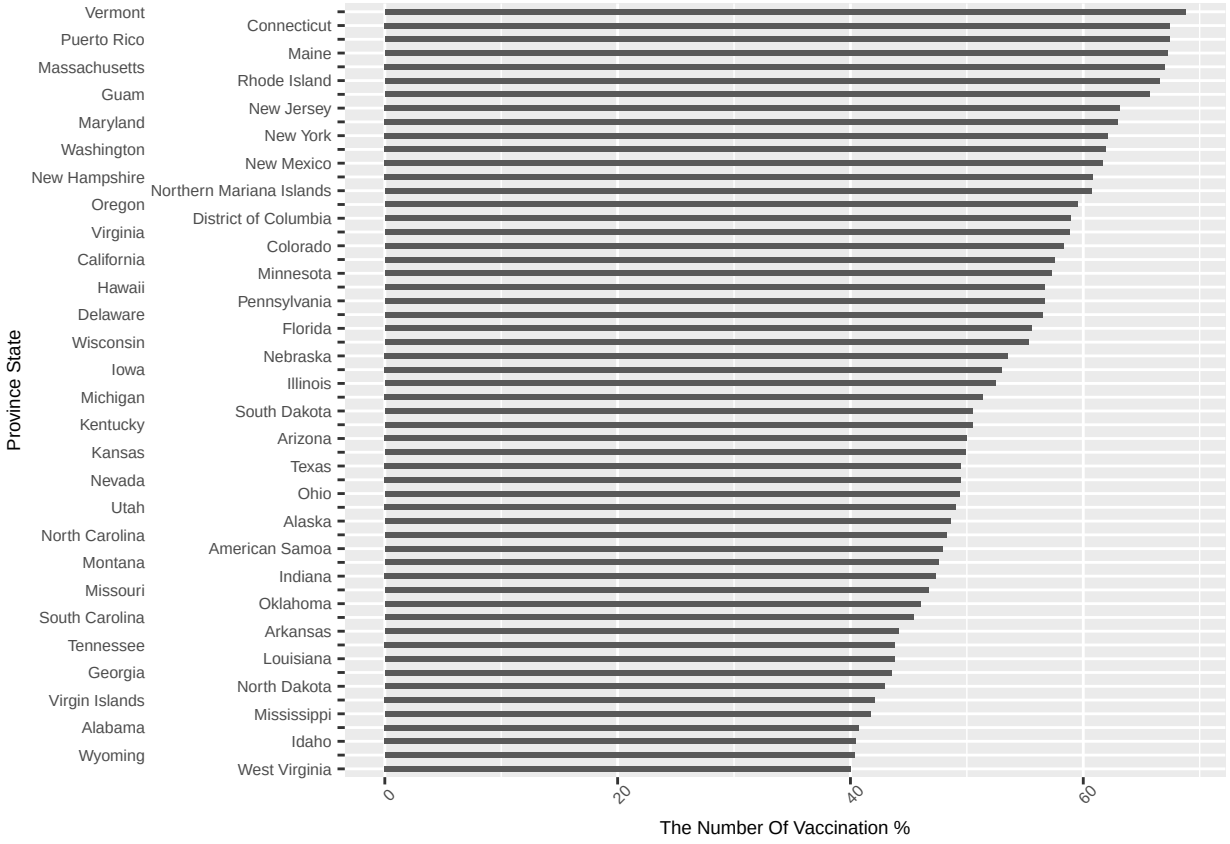
Table 3: Table for top10 min fully vaccinated people

Province_State	full_vaccination_percent	people_fully_vaccinated	population
West Virginia	40.05	717,702	1,792,147
Wyoming	40.36	233,559	578,759
Idaho	40.47	723,255	1,787,065
Alabama	40.74	1,997,364	4,903,185
Mississippi	41.71	1,241,211	2,976,149
Virgin Islands	42.11	45,167	107,268
North Dakota	42.93	327,120	762,062
Georgia	43.54	4,622,898	10,617,423
Louisiana	43.80	2,036,277	4,648,794
Tennessee	43.82	2,992,530	6,829,174

The percent of fully vaccinated people in each state

- The number of fully vaccinaed people in percent were arranged from high to low.
- In most high cases, the percent of vaccination is around 60% of population, while the low cases are 40% of population.

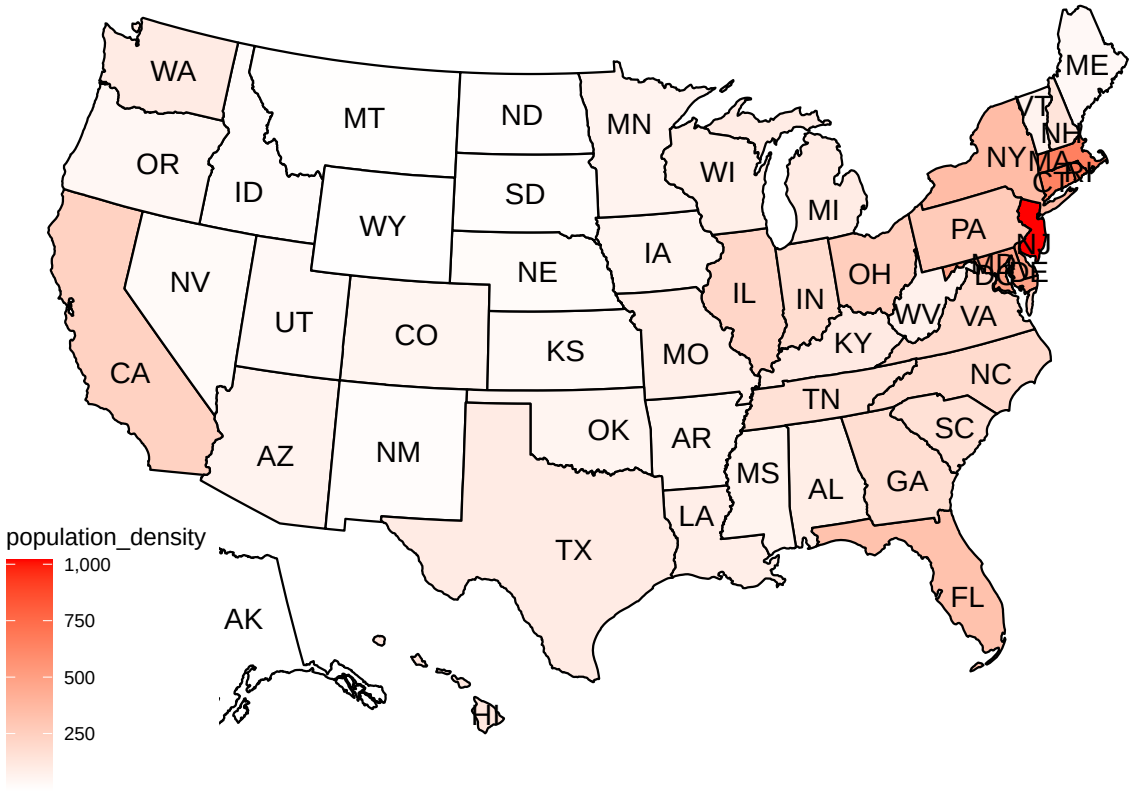
- There is a room to improve the immunity among the low vaccination states.




```
## Joining, by = "state"
```

Population density

- Use vac_data and uid.
- Use left join between them.
- Use area from area table as mentioned before.
- Population density is population in each state divided by area.
- High population density areas are in New York, New Jersey, California, and Florida.

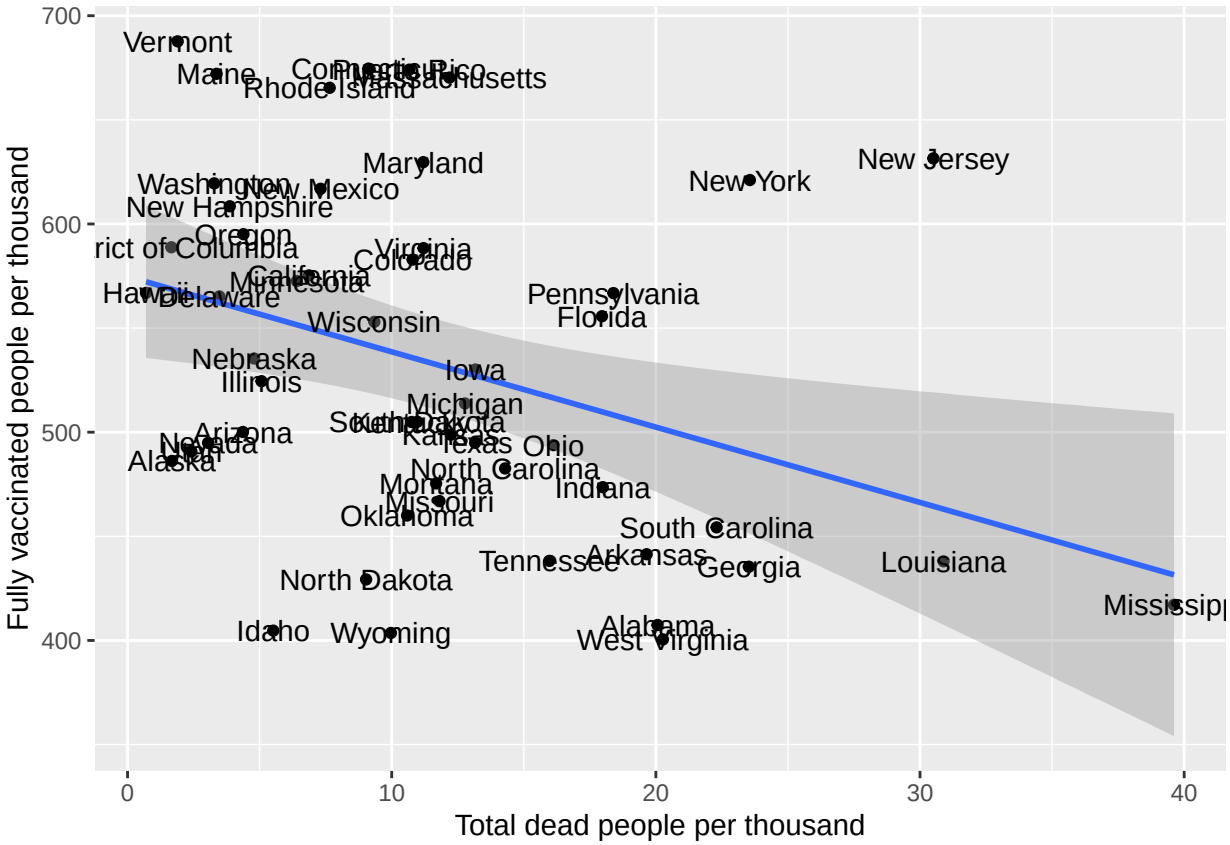


Model

- Using linear regression model to predict the results between total deaths and fully vaccinated people.
- Using the geom_smooth(method='lm')
- The result shows that the more fully vaccinated people, the less death rate.
- New York and New Jersey are outliers. The reason is that their population density is so high.

```
## Joining, by = "state"
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
##
## Call:
## lm(formula = people_fully_vaccinated_per_thou ~ total_deaths_per_thou,
##     data = vac_data_4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -150.12  -57.56  -18.41   49.33  167.06
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            574.769     18.963  30.309  < 2e-16 ***
## total_deaths_per_thou   -3.617      1.327  -2.725  0.00883 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.22 on 50 degrees of freedom
## Multiple R-squared:  0.1293, Adjusted R-squared:  0.1119
## F-statistic: 7.427 on 1 and 50 DF,  p-value: 0.008832
```

```
##        state people_fully_vaccinated_per_thou Country_Region total_cases
## 1    Alabama                            407.4             US      752775
## 2     Alaska                            486.2             US       99596
## 3    Arizona                            500.1             US     1037878
## 4   Arkansas                            441.4             US      484060
```

```
## 5 California                            575.3             US     4579170
## 6   Colorado                            582.9             US      621525
##   total_deaths total_cases_per_thou total_deaths_per_thou   predlm      fit
## 1        13213             1143.0396              20.063076 502.2095 502.2095
## 2          484              345.8194               1.680556 568.6913 568.6913
## 3        19594              231.3896               4.368382 558.9706 558.9706
## 4         7701             1235.1274              19.649869 503.7039 503.7039
## 5        68934              456.1332               6.866547 549.9358 549.9358
## 6         7848              854.6694              10.791916 535.7394 535.7394
##        lwr      upr
## 1 471.0787 533.3402
## 2 534.1810 603.2017
## 3 529.6650 588.2762
## 4 473.3500 534.0578
## 5 524.5950 575.2766
## 6 513.8109 557.6678
```

Conclusion

1. The number of Covid-19 patients in the US slightly increased since July 2021 because of the new variant of Covid-19, Delta. However, the results in US compared to Thailand, the US has lower increases in Covid-19 cases than Thailand.

2. The top 10 of Covid-19 patients are in the southern states of the US. The reason is that they have a lower percentage of fully vaccination compared to the northern states.

3. Top top 10 of death rate conformed with the Covid-19 cases. The more cases the US has, the more death rates the US has to deal with.

4. The southern states which has low pecentage of full vaccinated people only 40% of the population, while the northern states have 60%.

5. According to the model, the more full vaccination the US has, the lower death rate the US gets. Although New York and New Jersey have high vaccinated rates, they still have high death rates because of their high population density.

Bias

1. The percentage of tested people in each area is not the same, it will directly lead to wrong cases.

2. The number of facilities such hospitals and medical equipment is not the same in each area, which can lead to high death rate.

3. The system used to count the number of cases and death rate is different in each state.

4. The number of immigrants can have an impact on the death rate and Covid-19 cases because immigrants will face many obstacles to access hospital services.