Role of vaccinations in COVID19 pandemic: analysis of trends and performance by US States

(DTSC 5301-001 Assignment)

Spriha Awasthi

Contents

1 Problem statement

For the purpose of the assignment we want to analyze and answer the following questions:

1. Which country has performed best in terms of cases per million population?
2. Which US state has performed best in terms of cases per million population?
3. What is the role of vaccination on daily new cases?

2 Dataset attributes

2.1 Description and sources

We will be using Github pages of following official accounts and data provided:

1. *CSSEGISandData*: the link contains the data from Johns Hopkins as primary source. There are 5 different CSV files we will use. The data from each file we will use is as follows:

    a. *time_series_covid19_confirmed_global.csv* - we extract the timeseries by date of COVID19 cases for different countries from this file.
    b. *time_series_covid19_deaths_global.csv* - we extract the timeseries by date of deaths due to COVID19 for different countries from this file.
    c. *time_series_covid19_confirmed_US.csv* - here we extract the timeseries by date of COVID19 cases along with the total population for different states in the US.
    d. *time_series_covid19_deaths_US.csv* - we extract the timeseries of deaths due to COVID19 for different states in the US from this file.
    e. *UID_ISO_FIPS_LookUp_Table.csv* - the first file does not provide the population of the countries that we will need to compute cases/deaths per 1000. So we use this file to provide us the population by countries.

2. *BloombergGraphics*: this official account of Bloomberg covers vacccination data all U.S. states, territories and several countries, on a daily basis. Data has been gathered from government websites, official statements, Bloomberg interviews and third-party sources including the World Health Organization, Johns Hopkins University and Our World In Data.

    a. historical-usa-doses-administered.csv - contains the timeseries of daily total vaccinations achieved by date for different states. There are several dates missing so will need cleanup and filling.

3. *CivilServiceUSA*: This account maintains a variety of political data for US. We will use one table to get state names to codes mapping. We could have hard coded it but using this official dataset ensures its reproducible and adapts to future changes.

    a. us-governors.csv - the state codes and names are drawn from this file.

2.2 Dataset dimensions

Let us now load the dataset and observe the dimensions. Here we will also rename some columns to be more coherent across data files.

```r
# Some formatted strings to create URLs
base_cssegi_uri <- str_c("https://raw.githubusercontent.com/CSSEGISandData/",
                         "COVID-19/master/csse_covid_19_data/")
file_names <- c("csse_covid_19_time_series/time_series_covid19_confirmed_global.csv",
                "csse_covid_19_time_series/time_series_covid19_deaths_global.csv",
                "csse_covid_19_time_series/time_series_covid19_confirmed_US.csv",
                "csse_covid_19_time_series/time_series_covid19_deaths_US.csv")
urls <- str_c(base_cssegi_uri, file_names)

# Load cases and deaths globally and in US.
global_cases <- read.csv(urls[1], header = TRUE, check.names = FALSE)
global_deaths <- read.csv(urls[2], header = TRUE, check.names = FALSE)
```

```r
US_cases <- read.csv(urls[3], header = TRUE, check.names = FALSE)
US_deaths <- read.csv(urls[4], header = TRUE, check.names = FALSE)

# Load countries' populations.
uid_lookup_url <- str_c(base_cssegi_uri, "UID_ISO_FIPS_LookUp_Table.csv")
uid <- read.csv(uid_lookup_url, header = TRUE, check.names = FALSE) %>%
  select(-c(Lat, Long_, Combined_Key, code3, iso2, iso3, Admin2))

# Load vaccination data for US states.
us_vac_url <- str_c("https://raw.githubusercontent.com/BloombergGraphics/covid-vaccine-",
                    "tracker-data/master/data/historical-usa-doses-administered.csv")
us_vac <- read.csv(us_vac_url, header = TRUE, check.names = FALSE)
us_vac <- rename(us_vac, state_code = id, vaccinations = value)

# Load political affiliation data for US states.
us_political_url = str_c("https://raw.githubusercontent.com/CivilServiceUSA/us-governors/",
                         "master/us-governors/data/us-governors.csv")
us_states_codes_names = read.csv(us_political_url, header = TRUE, check.names = FALSE)
us_states_codes_names <- rename(us_states_codes_names, Province_State = state_name) %>%
  select(Province_State, state_code)

nrow(global_cases)
```

```
[1] 279
```

```r
ncol(global_cases)
```

```
[1] 613
```

```r
nrow(global_deaths)
```

```
[1] 279
```

```r
ncol(global_deaths)
```

```
[1] 613
```

```r
nrow(US_cases)
```

```
[1] 3342
```

```r
ncol(US_cases)
```

```
[1] 620
```

```r
nrow(US_deaths)
```

```
[1] 3342
```

```
ncol(US_deaths)
```

```
[1] 621
```

```
nrow(uid)
```

```
[1] 4196
```

```
ncol(uid)
```

```
[1] 5
```

```
nrow(us_vac)
```

```
[1] 16219
```

```
ncol(us_vac)
```

```
[1] 3
```

```
nrow(us_states_codes_names)
```

```
[1] 50
```

```
ncol(us_states_codes_names)
```

```
[1] 2
```

3 Prepare dataframes for feature modeling

3.1 Expanding and merging deaths/cases into one timeseries

One thing we noticed in previous section was high number of columns in the cases and deaths related files. This is because rows represent the region and columns represent the dates. So for each new date of data, a new column is added. To facilitate plotting a timeseries and comparing region to region on a given date we will use *pivot_longer* method to increase the rows but reduce the columns. This will map all date columns into a single column of *date* and the corresponding value in a new column for *cases* or *deaths*.

While expanding we also remove the columns we don't need and rename a few for better understanding and being coherent with others.

```
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))

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_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_))
```

Next let us merge the 2 global tiblles and 2 US tibbles into 1 which facilitates plotting later on to analyse. We will also filter the entries that have zero cases and zero population regions in US as it doesn't not add value and should be cleaned up. We will also map date string column values to date type objects for comparisons and consistency.

```
# Merge cases and deaths into one by joining and convert string into date type column
global <- global_cases %>%
  full_join(global_deaths) %>%
  rename(Country_Region = `Country/Region`,
         Province_State = `Province/State`) %>%
  mutate(date = mdy(date)) %>%
  filter(cases > 0)
```

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

```
# Join and add deaths/cases per million for comparison purposes.
global <- global %>%
  left_join(uid, by = c("Province_State", "Country_Region")) %>%
  filter(Population > 0) %>%
  select(-c(UID, FIPS)) %>%
  select(Province_State, Country_Region, date, cases, deaths, Population)

# Join deaths and cases into one table and remove all that have zero Population
US <- US_cases %>%
  full_join(US_deaths) %>%
  filter(Population > 0)
```

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

At this stage the *global* dataframe contains the timeseries of deaths and cases for countries. We will use this later in data visualization. Below is its summary printed out for understanding.

```
summary(global)
```

```
##    Province_State          Country_Region        date
##           :104660   China         : 20078   Min.   :2020-01-22
##  Anhui    :   609   Canada        :  7101   1st Qu.:2020-07-23
##  Beijing  :   609   France        :  6521   Median :2020-12-14
##  Chongqing:   609   United Kingdom:  6491   Mean   :2020-12-12
##  Fujian   :   609   Australia     :  4672   3rd Qu.:2021-05-04
##  Guangdong:   609   Netherlands   :  2777   Max.   :2021-09-21
##  (Other)  : 44061   (Other)       :104126
##      cases              deaths          Population
##  Min.   :       1   Min.   :     0   Min.   :8.090e+02
##  1st Qu.:     387   1st Qu.:     3   1st Qu.:9.775e+05
##  Median :    4597   Median :    69   Median :7.497e+06
##  Mean   :  322550   Mean   :  7431   Mean   :2.984e+07
##  3rd Qu.:   72722   3rd Qu.:  1290   3rd Qu.:3.102e+07
##  Max.   :42410607   Max.   :678407   Max.   :1.380e+09
##
```

3.2 Modeling country level and state level data for US

Next, we focus in generating state level data for our analysis. For our analysis we need a more comprehensive tied data of US by states which captures the cases, deaths, population and vaccinations. The vaccinations and data are obtained by joining the Bloomberg Organization's github page

```
# Group by state/country/date, sum relevant metric and add a new per million deaths and cases.
US_by_state <- US %>%
  group_by(Province_State, Country_Region, date) %>%
  summarize(cases = sum(cases),
            deaths = sum(deaths),
            Population = sum(Population)) %>%
  select(Province_State, Country_Region, date, cases, deaths, Population) %>%
  ungroup()
```

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

Before we join we will convert the date string column in vaccinations series into date object for smoothly joining differently formatted date in John Hopkin's data. We will also add cases and deaths per million population to be able to compare rates of cases and deaths.

We do observe here that the data from Bloomberg did not repeat dates if the new data was not available. So we need to fill the missing values such that for each state the missing value on date is filled with previous date's value. After that we replace all NA values with zero as the vaccinations did not start much later than COVID19 cases started.

```
# Convert for joining correctly
us_vac <- us_vac %>% mutate(date = ymd(date))

US_by_state <- US_by_state %>%
  left_join(us_states_codes_names) %>%
  left_join(us_vac) %>%
  group_by(Province_State) %>%
  fill(vaccinations) %>%
  ungroup() %>%
  mutate_at(c("vaccinations"), ~replace(., is.na(.), 0))
```

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

```
## Joining, by = c("date", "state_code")
```

Now that we have data by state we can compute the total for the US by each date through grouping on {country, date} and summing on each group to generate the cases, deaths, vaccinations and populations. We will also use *lag* method to generate new columns for US_by_state and US_totals to generate new cases, deaths, vaccinations each date.

```
US_totals <- US_by_state %>%
  group_by(Country_Region, date) %>%
  summarize(cases = sum(cases),
            deaths = sum(deaths),
            vaccinations = sum(vaccinations),
            Population = sum(Population)) %>%
  select(Country_Region, date, vaccinations, cases, deaths, Population) %>%
  ungroup()
```

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

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

US_totals <- US_totals %>%
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths),
         new_vaccinations = vaccinations - lag(vaccinations))

summary(US_by_state)
```

```
##         Province_State  Country_Region      date                 cases
##  Alabama       :  609   US:34104       Min.   :2020-01-22   Min.   :      0
##  Alaska        :  609                  1st Qu.:2020-06-22   1st Qu.:   6907
##  American Samoa:  609                  Median :2020-11-21   Median :  88276
##  Arizona       :  609                  Mean   :2020-11-21   Mean   : 296246
##  Arkansas      :  609                  3rd Qu.:2021-04-22   3rd Qu.: 358439
##  California    :  609                  Max.   :2021-09-21   Max.   :4651285
##  (Other)       :30450
##      deaths          Population         state_code     vaccinations
##  Min.   :    0.0   Min.   :   55144   AK     :  609   Min.   :       0
##  1st Qu.:  117.8   1st Qu.: 1355836   AL     :  609   1st Qu.:       0
##  Median : 1556.0   Median : 3855955   AR     :  609   Median :       0
##  Mean   : 5622.7   Mean   : 5944199   AZ     :  609   Mean   : 1686318
##  3rd Qu.: 6814.0   3rd Qu.: 6989056   CA     :  609   3rd Qu.: 1197182
##  Max.   :68087.0   Max.   :39512223   (Other):27405   Max.   :49800281
##                                       NA’s   : 3654
##    new_cases          new_deaths        new_vaccinations
##  Min.   :-4651285   Min.   :-68087.00   Min.   :-49800281
##  1st Qu.:      13   1st Qu.:     0.00   1st Qu.:        0
##  Median :     301   Median :     3.00   Median :        0
##  Mean   :       3   Mean   :     0.03   Mean   :       15
##  3rd Qu.:    1147   3rd Qu.:    17.00   3rd Qu.:     6956
##  Max.   :  151765   Max.   :  4448.00   Max.   :   805477
##  NA’s   :1          NA’s   :1           NA’s   :1
```

```
summary(US_totals)
```

```
##  Country_Region      date             vaccinations          cases
##  US:609         Min.   :2020-01-22   Min.   :        0   Min.   :       1
##                 1st Qu.:2020-06-22   1st Qu.:        0   1st Qu.: 2293285
##                 Median :2020-11-21   Median :        0   Median :12102675
##                 Mean   :2020-11-21   Mean   : 94433830   Mean   :16589761
##                 3rd Qu.:2021-04-22   3rd Qu.:215546146   3rd Qu.:31707897
##                 Max.   :2021-09-21   Max.   :378517141   Max.   :41997742
##
##      deaths         Population          new_cases        new_deaths
##  Min.   :     1   Min.   :332875137   Min.   :     0   Min.   :-1516.0
##  1st Qu.:120532   1st Qu.:332875137   1st Qu.: 24256   1st Qu.:  387.5
##  Median :254921   Median :332875137   Median : 46479   Median :  860.0
##  Mean   :314874   Mean   :332875137   Mean   : 69075   Mean   : 1076.6
##  3rd Qu.:563328   3rd Qu.:332875137   3rd Qu.: 79853   3rd Qu.: 1465.0
##  Max.   :654580   Max.   :332875137   Max.   :317448   Max.   : 5071.0
##                                       NA’s   :1        NA’s   :1
##  new_vaccinations
##  Min.   :      0
##  1st Qu.:      0
##  Median :      0
##  Mean   : 622561
##  3rd Qu.: 981376
##  Max.   :4566360
##  NA’s   :1
```

Finally, we have generated the data frames needed for our analysis and we will plot some visualizations for analysis.

4 Data visualizations and analysis

Next we answer the questions we set out in the beginning:

4.1 Which country has performed best in terms of cases per million population?

We will group by coutry and compute the top 3 best performing countries by this metric.

```
country_totals <- global %>%
  group_by(Country_Region) %>%
  summarize(deaths = max(deaths), cases = max(cases),
            population = max(Population),
            cases_per_mil = 1000000 * cases / population) %>%
  filter(cases > 0, population > 0)
country_totals %>%
  slice_min(cases_per_mil, n = 5)
```

```
## # A tibble: 5 x 5
##   Country_Region deaths cases population cases_per_mil
##   <fct>           <int> <int>      <int>         <dbl>
## 1 Micronesia          0     1     113815          8.79
## 2 Vanuatu             1     4     292680         13.7
## 3 Samoa               0     3     196130         15.3
## 4 Kiribati            0     2     117606         17.0
## 5 Tanzania           50  1367   59734213         22.9
```

The findings show that best performing countries are Pacific Ocean island nations. A possible explanation could be that they are not easily connected and were able to contain. The best nation in this metric is Micronesia. However there could be misrepresentations as well and the actual number could be higher. Tanzania's rank in top 5, however, raises some questions as it is well connected and should have higher cases.

4.2 Which US state has performed best in terms of cases per million population?

```
US_state_totals <- US_by_state %>%
  group_by(Province_State) %>%
  summarize(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 = 5)
```

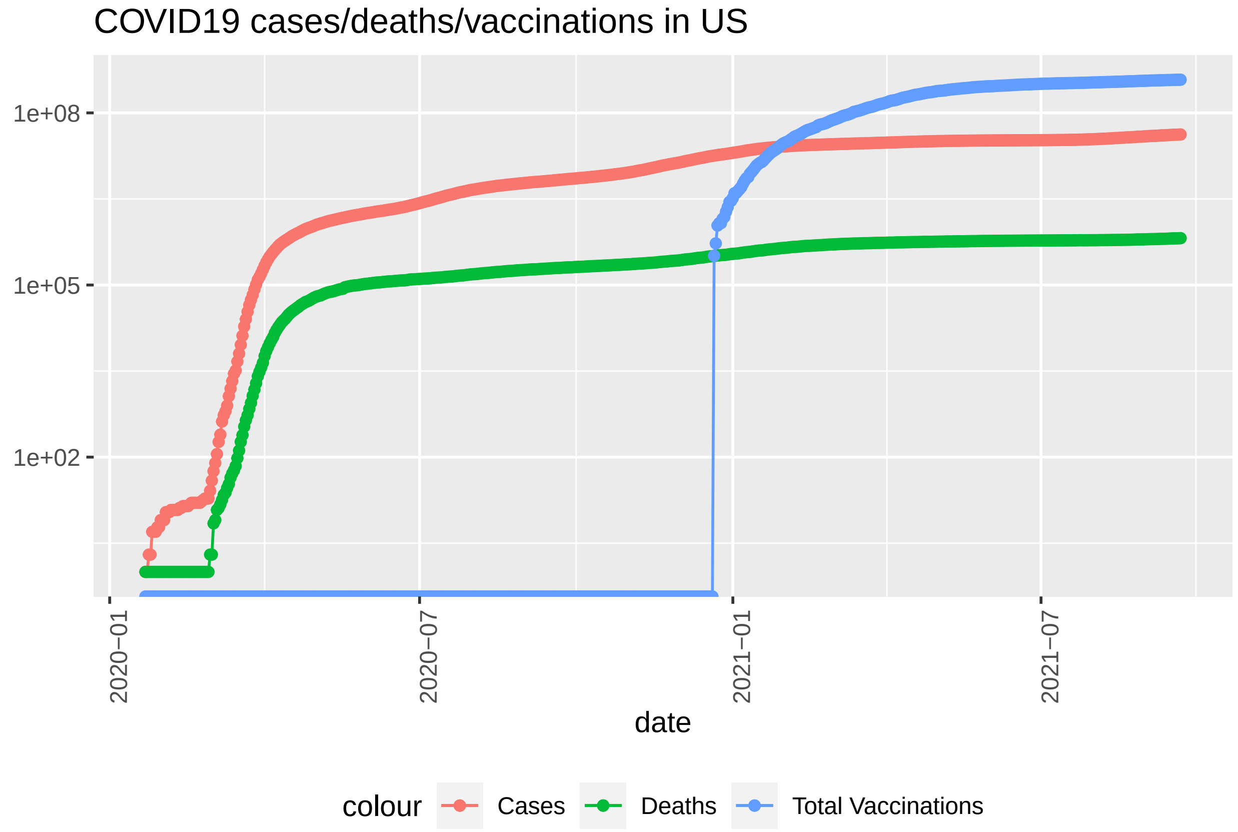
```
## # A tibble: 5 x 6
##   Province_State           deaths  cases population cases_per_thou deaths_per_thou
##   <fct>                     <int>  <int>      <int>          <dbl>           <dbl>
## 1 Northern Mariana Islands      2    265      55144           4.81          0.0363
## 2 Puerto Rico                1787 175489    3754939          46.7           0.476
## 3 Vermont                     301  31890     623989          51.1           0.482
## 4 Hawaii                      709  73841    1415872          52.2           0.501
## 5 Utah                       1898 367668    3205958         115.            0.592
```

Similar to analysis of countries, the US states that performed best were isolated and island states. The best province is Northern Mariana Islands and best mainland state was Vermont.

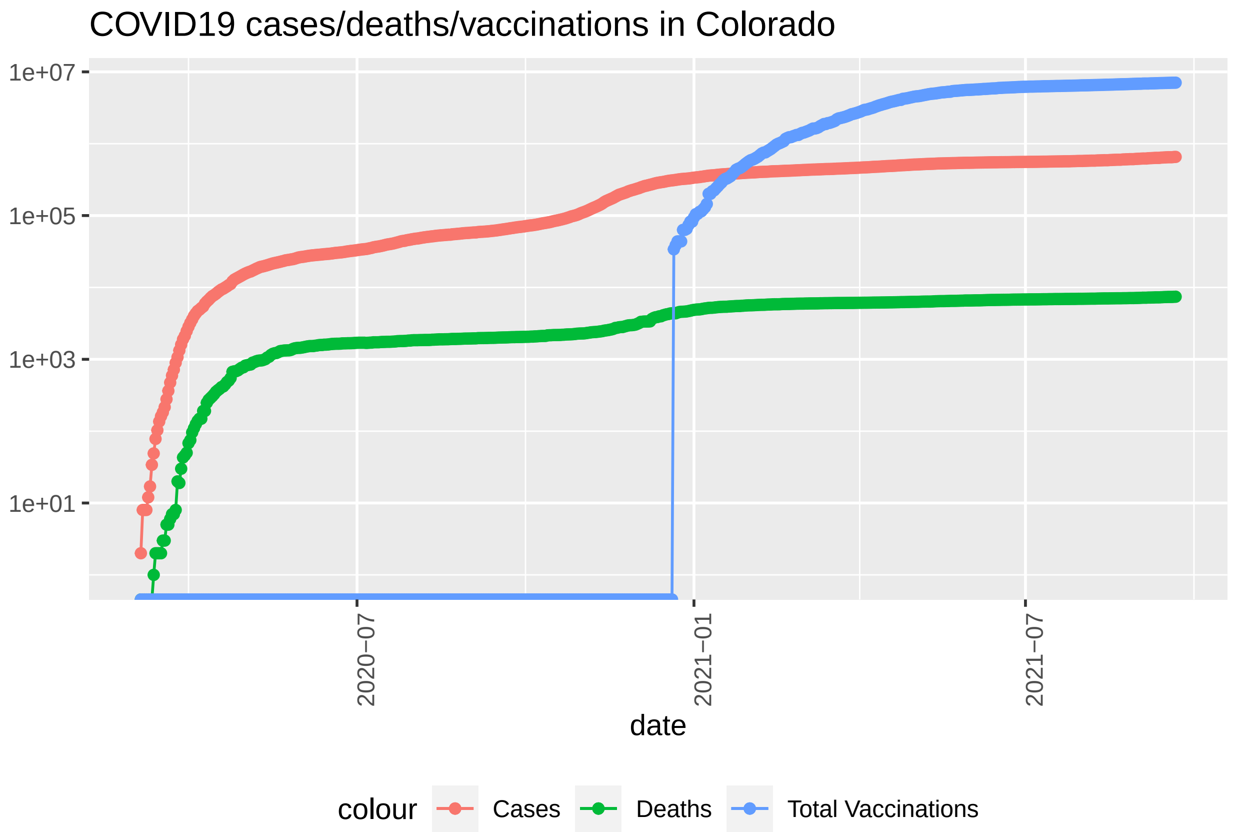
4.3 What is the role of vaccination on daily new cases?

Now that we have created and loaded our data into tibbles we will plot some visualizations to observe the progress of cases globally and in US.

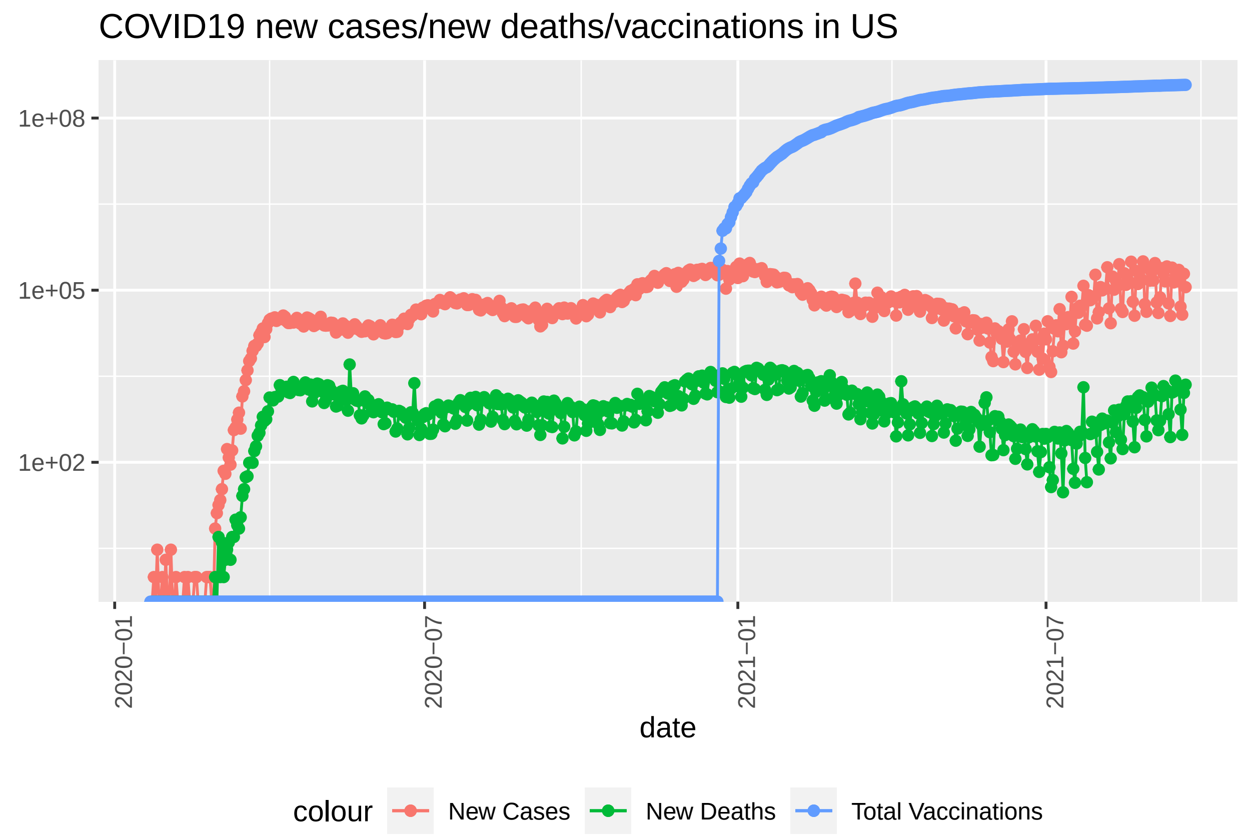
```
US_totals %>%
  filter(cases > 0) %>%
  ggplot(aes(x = date, y = cases)) +
  geom_line(aes(color = "Cases")) +
  geom_point(aes(color = "Cases")) +
  geom_line(aes(y = deaths, color = "Deaths")) +
  geom_point(aes(y = deaths, color = "Deaths")) +
  geom_line(aes(y = vaccinations, color = "Total Vaccinations")) +
  geom_point(aes(y = vaccinations, color = "Total Vaccinations")) +
  scale_y_log10() +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "COVID19 cases/deaths/vaccinations in US", y= NULL)
```



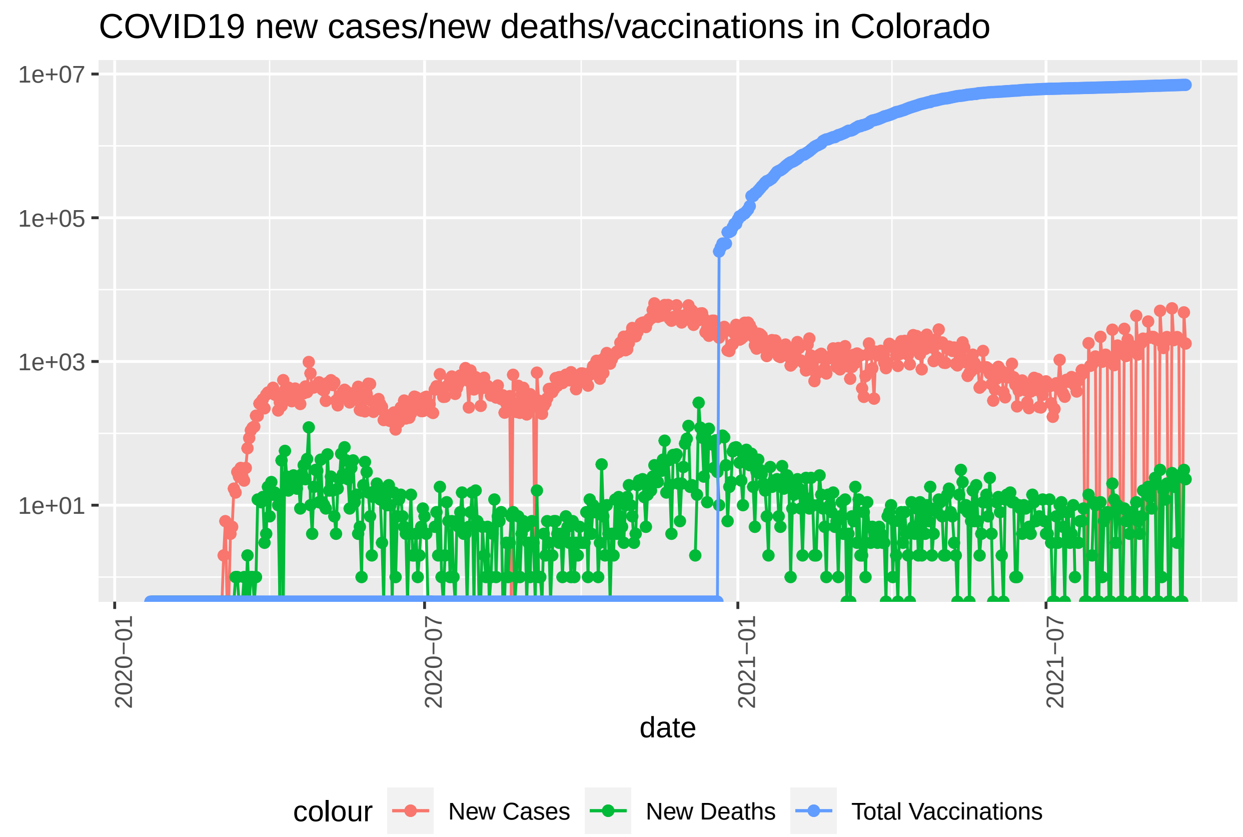
```
state <- "Colorado"
US_by_state %>%
  filter(Province_State == state) %>%
  filter(cases > 0) %>%
  ggplot(aes(x = date, y = cases)) +
  geom_line(aes(color = "Cases")) +
  geom_point(aes(color = "Cases")) +
  geom_line(aes(y = deaths, color = "Deaths")) +
  geom_point(aes(y = deaths, color = "Deaths")) +
  geom_line(aes(y = vaccinations, color = "Total Vaccinations")) +
  geom_point(aes(y = vaccinations, color = "Total Vaccinations")) +
  scale_y_log10() +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = str_c("COVID19 cases/deaths/vaccinations in ", state), y= NULL)
```



```
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")) +
  geom_line(aes(y = vaccinations, color = "Total Vaccinations")) +
  geom_point(aes(y = vaccinations, color = "Total Vaccinations")) +
  scale_y_log10() +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "COVID19 new cases/new deaths/vaccinations in US", y= NULL)
```



```
US_by_state %>%
  filter(Province_State == state) %>%
  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")) +
  geom_line(aes(y = vaccinations, color = "Total Vaccinations")) +
  geom_point(aes(y = vaccinations, color = "Total Vaccinations")) +
  scale_y_log10() +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = str_c("COVID19 new cases/new deaths/vaccinations in ", state), y= NULL)
```



Take note of two observations above:

1. The vaccines are launched in December 2020 in the US. Colorado is one of the early adopters of vaccines as it maps to first date of vaccinations in US.
2. As the total number of vaccinations have increased, daily new cases have reduced till July 2022. They start increasing gradually after this due to the Delta variant outbreak. However, the rate of increase could have been catastrophic in the absence of vaccinations like in early pandemic.

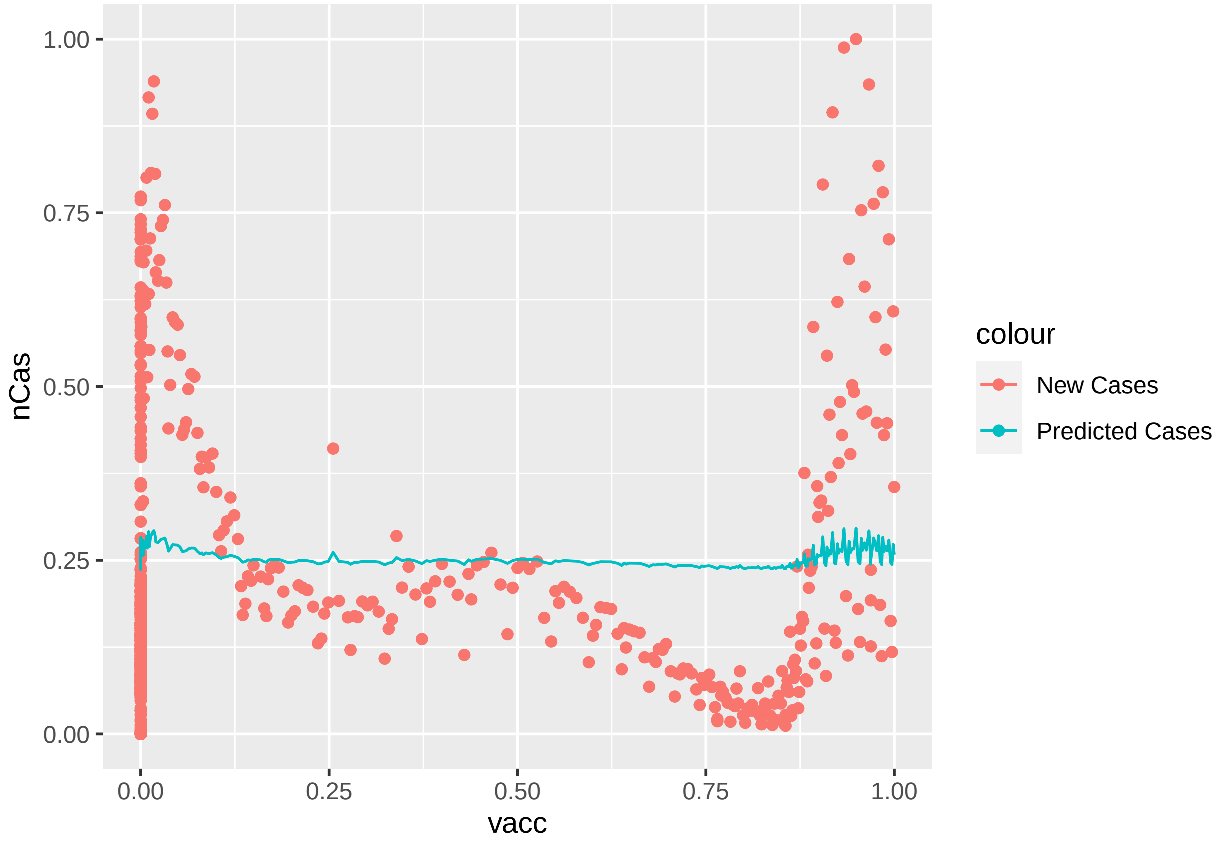
To establish a quantitative relationship between daily new cases and vaccinations, we will build a linear model on total vaccinated vs number of daily new cases. We want to model total vaccinations instead of daily new vaccinations as the new cases are likely impacted by total vaccinated population and not just by new vaccinations.We also need to normalize the columns as there is considerable order difference between the 2 quantities.

```
vaccinations_data <- US_totals %>%
  filter((!is.na(vaccinations)) & (!is.na(new_cases))) %>%
  mutate(vacc = (vaccinations - min(vaccinations))/(max(vaccinations) - min(vaccinations)),
         nCas = (new_cases - min(new_cases))/(max(new_cases) - min(new_cases))) %>%
  select(vacc, nCas)

lmodel_cases <- lm(vacc ~ nCas, data = vaccinations_data)
summary(lmodel_cases)
```

```
##
## Call:
```

```
## lm(formula = vacc ~ nCas, data = vaccinations_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2827 -0.2456 -0.2377  0.3226  0.7529
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.23704    0.02092   11.33   <2e-16 ***
## nCas         0.05907    0.06870    0.86     0.39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3607 on 606 degrees of freedom
## Multiple R-squared:  0.001218,   Adjusted R-squared:  -0.0004298
## F-statistic: 0.7392 on 1 and 606 DF,  p-value: 0.3903
```

```
vaccinations_data <- vaccinations_data %>%
  mutate(pred_cases = predict(lmodel_cases))

vaccinations_data %>% ggplot(aes(x = vacc)) +
  geom_point(aes(y = nCas, color = "New Cases")) +
  geom_line(aes(y = pred_cases, color = "Predicted Cases"))
```



The graph shows the partially negative slope of new cases with vaccines increase and hence establishes their importance in containing COVID19 pandemic. The graph, however, has kinks and irregularities due to data getting skewed after the delta variant outbreak.

5 Bias

The findings above align with expectations but its worth calling out that there are biases involved at several places. Some of these are as follows:

1. Data collection is done from diverse sources. Accuracy of sources, specially international values are untrustworthy.
2. The number of cases in some places might be under reported because of political reasons
3. The number of deaths may not be exact as the death of the people suffering from prior health conditions may not be reported as a COVID19 death.
4. The Parameters for reporting of the cases, deaths and vaccines for different countries might not be same
5. The events may not be reported on the day of its occurrence.

6 Conclusion

We have performed analysis above to find best country and best US state in terms of COVID cases per million population. Micronesia is the best performing country in this metric and within US, the best province is Northern Mariana Islands with best mainland state being Vermont. The findings on role of vaccine makes it clear how they help contain pandemic. We saw daily new cases reducing with more vaccines being distributed till the delta outbreak. We have also discussed some biases in the data involved.