# COVID-19 Data Report

Javel Williamson

March 4,2024

## Introduction

This report takes publicly available COVID-19 data and presents visualizations depicting the local (US) and global impact of the coronavirus pandemic. A model is presented at the end that explores whether there is a statistical relationship between COVID-19 vaccination rates and death rates in the US.

### COVID-19 Data Sources

- COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University
  - https://github.com/CSSEGISandData/COVID-19
- Coronavirus (COVID-19) Vaccinations
  - https://ourworldindata.org/covid-vaccinations

## Process

In this report, R is used to import, tidy, transform, visualize, and model COVID-19 data. The details of the process are given in the steps below.

### 1 Import R libraries and set up environment

```
# Import libraries and set options
library(tidyverse)
library(lubridate)
options(warn=-1)
options(dplyr.summarise.inform = FALSE)
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_knit$set(root.dir = getwd())
```

### 2 Download and import COVID-19 source data files

```
# Read .csv files for global and US COVID-19 cases and deaths
url_in <-
  'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_ti
filenames <- c('time_series_covid19_confirmed_global.csv', 'time_series_covid19_deaths_global.csv', 'ti
```

```
urls <- str_c(url_in, filenames)
global_cases <- read_csv(urls[1])
global_deaths <- read_csv(urls[2])
US_cases <- read_csv(urls[3])
US_deaths <- read_csv(urls[4])

# Read .csv file for global population lookup table
uid_lookup_url <-
  'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/UID_ISO_FIPS_Lool
uid <- read_csv(uid_lookup_url) %>%
  select(-c(Lat, Long_, Combined_Key, code3, iso2, iso3, Admin2))

# Read .csv file for US COVID-19 vaccinations
url_in <-
  'https://covid.ourworldindata.org/data/vaccinations/us_state_vaccinations.csv'
US_vaccinations <- read_csv(url_in) %>%
  select(-c(total_vaccinations, total_distributed, people_vaccinated, people_fully_vaccinated,
            daily_vaccinations_raw, daily_vaccinations, daily_vaccinations_per_million,
            share_doses_used, total_boosters)) %>%
  rename(Province_State = 'location')
```

### 3 Tidy and transform COVID-19 data in preparation for visualization and analysis

- Perform transformations on global COVID-19 case and death data

```
# For 'global_cases' df, make 'Province/State' and 'Country/Region' factors and pivot dates into rows
global_cases <- mutate_at(global_cases, vars('Province/State', 'Country/Region'), as.factor) %>%
    pivot_longer(cols = -c('Province/State', 'Country/Region', 'Lat', 'Long'),
                 names_to = 'Date',
                 values_to = 'Cases') %>%
    select(-c('Lat', 'Long'))

# For 'global_deaths' df, make 'Province/State' and 'Country/Region' factors and pivot dates into rows
global_deaths <- mutate_at(global_deaths, vars('Province/State', 'Country/Region'), as.factor) %>%
    pivot_longer(cols = -c('Province/State', 'Country/Region', 'Lat', 'Long'),
            names_to = 'Date',
            values_to = 'Deaths') %>%
    select(-c('Lat', 'Long'))

# Merge 'global_cases' df and 'global_deaths' df into 'global' df and rename columns
global <- global_cases %>%
    full_join(global_deaths) %>%
    rename(Country_Region = 'Country/Region',
           Province_State = 'Province/State') %>%
    mutate(Date = mdy(Date))

# Combine 'Province_State' and 'Country_Region' columns into one 'Combined_Key' column
global <- global %>%
    unite('Combined_Key',
        c(Province_State, Country_Region),
        sep = ', ',
        na.rm = TRUE,
```

```
        remove = FALSE)

# Join 'global' df with global population lookup table df and remove unneeded columns
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)

# Group 'global_cases_per_hundred' df by 'Country_Region', calculate 'Cases_per_hundred' variable
global_cases_per_hundred <- global %>%
  group_by(Country_Region, Population) %>%
  summarize(Cases = max(Cases), Population = max(Population, na.rm = T)) %>%
  mutate(Cases_per_hundred = (Cases/Population)*100) %>%
  arrange(desc(Cases_per_hundred)) %>%
  filter(Population > 0) %>%
  select(Country_Region, Population, Cases, Cases_per_hundred) %>%
  ungroup()

# Combine populations and case totals for countries with 'State_Province' factors
global_cases_per_hundred <- global_cases_per_hundred %>%
  group_by(Country_Region, Population) %>%
  summarize(Cases = sum(Cases), Population = sum(Population)) %>%
  summarize(Cases = max(Cases), Population = max(Population, na.rm = T)) %>%
  mutate(Cases_per_hundred = (Cases/Population)*100) %>%
  arrange(desc(Cases_per_hundred)) %>%
  filter(Population > 0) %>%
  select(Country_Region, Population, Cases, Cases_per_hundred) %>%
  ungroup()

# Group 'global_deaths_per_hundred' df by 'Country_Region', calculate 'Deaths_per_hundred' variable
global_deaths_per_hundred <- global %>%
  group_by(Country_Region, Population) %>%
  summarize(Deaths = max(Deaths), Population = max(Population, na.rm = T)) %>%
  mutate(Deaths_per_hundred = (Deaths/Population)*100) %>%
  arrange(desc(Deaths_per_hundred)) %>%
  filter(Population > 0) %>%
  select(Country_Region, Population, Deaths, Deaths_per_hundred) %>%
  ungroup()

# Combine populations and death totals for countries with 'State_Province' factors
global_deaths_per_hundred <- global_deaths_per_hundred %>%
  group_by(Country_Region, Population) %>%
  summarize(Deaths = sum(Deaths), Population = sum(Population)) %>%
  summarize(Deaths = max(Deaths), Population = max(Population, na.rm = T)) %>%
  mutate(Deaths_per_hundred = (Deaths/Population)*100) %>%
  arrange(desc(Deaths_per_hundred)) %>%
  filter(Population > 0) %>%
  select(Country_Region, Population, Deaths, Deaths_per_hundred) %>%
  ungroup()
```

- Perform transformations on US COVID-19 case and death data

```r
# For 'US_cases' df, create factors and pivot dates into rows, change 'Date' column to mdy
US_cases <- mutate_at(US_cases, vars(Admin2, Province_State, Country_Region), as.factor) %>%
  rename(County = 'Admin2') %>%
    pivot_longer(cols = -(UID:Combined_Key),
        names_to = 'Date',
        values_to = 'Cases') %>%
  filter(Cases >= 0) %>%
    select(County:Cases) %>%
    mutate(Date = mdy(Date)) %>%
    select(-c(Lat, Long_))

# For 'US_deaths' df, create factors and pivot dates into rows, change 'Date' column to mdy
US_deaths <- mutate_at(US_deaths, vars(Admin2, Province_State, Country_Region), as.factor) %>%
  rename(County = 'Admin2') %>%
    pivot_longer(cols = -(UID:Population),
        names_to = 'Date',
        values_to = 'Deaths') %>%
  filter(Deaths >= 0) %>%
    select(County:Deaths) %>%
    mutate(Date = mdy(Date)) %>%
    select(-c(Lat, Long_))

# Merge 'US_cases' df and 'US_deaths' df into 'US' df
US <- US_cases %>%
    full_join(US_deaths)

# For 'US_by_state' df, calculate sums of 'Cases', 'Deaths', and 'Population' variables by US state
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()

# For 'US_by_state_cases_deaths_per_day' df, calculate 'New_Cases' and 'New_Deaths' variables
US_by_state_cases_deaths_per_day <- US_by_state %>%
  group_by(Province_State) %>%
  mutate(New_Cases = Cases - lag(Cases),
         New_Deaths = Deaths - lag(Deaths)) %>%
  select(Province_State, Country_Region, Date, Cases, Deaths, Population,
         New_Cases, New_Deaths) %>%
  ungroup()

# Remove negative 'New_Cases' values from 'US_by_state_cases_deaths_per_day' df
index1 <- which(US_by_state_cases_deaths_per_day$New_Cases >= 0)
US_by_state_cases_deaths_per_day <- US_by_state_cases_deaths_per_day[index1,]

# Remove negative 'New_Deaths' values from 'US_by_state_cases_deaths_per_day' df
index2 <- which(US_by_state_cases_deaths_per_day$New_Deaths >= 0)
US_by_state_cases_deaths_per_day <- US_by_state_cases_deaths_per_day[index2,]

# Group 'US_by_state_cases_deaths_per_day' df by 'Province_State' and filter rows with population > 0
US_by_state_cases_deaths_per_day <- US_by_state_cases_deaths_per_day %>%
  group_by(Province_State, Date) %>%
```

```
  select(Province_State, Country_Region, Date, Cases, Deaths, Population,
         New_Cases, New_Deaths) %>%
  filter(Population > 0) %>%
  ungroup()

# Group by 'Province_State', record max in 'Cases' variable, and calculate 'Cases_per_hundred' variable
US_by_state_cases_per_hundred <- US_by_state %>%
  group_by(Province_State, Population) %>%
  summarize(Cases = max(Cases)) %>%
  mutate(Cases_per_hundred = (Cases/Population)*100) %>%
  arrange(desc(Cases_per_hundred)) %>%
  filter(Population > 0) %>%
  select(Province_State, Population, Cases, Cases_per_hundred) %>%
  ungroup()

# Group by 'Province_State', record max in 'Deaths' variable, and calculate 'Deaths_per_hundred' variab
US_by_state_deaths_per_hundred <- US_by_state %>%
  group_by(Province_State, Population) %>%
  summarize(Deaths = max(Deaths)) %>%
  mutate(Deaths_per_hundred = (Deaths/Population)*100) %>%
  arrange(desc(Deaths_per_hundred)) %>%
  filter(Population > 0) %>%
  select(Province_State, Population, Deaths, Deaths_per_hundred) %>%
  ungroup()
```

- Perform transformations on US COVID-19 vaccination data

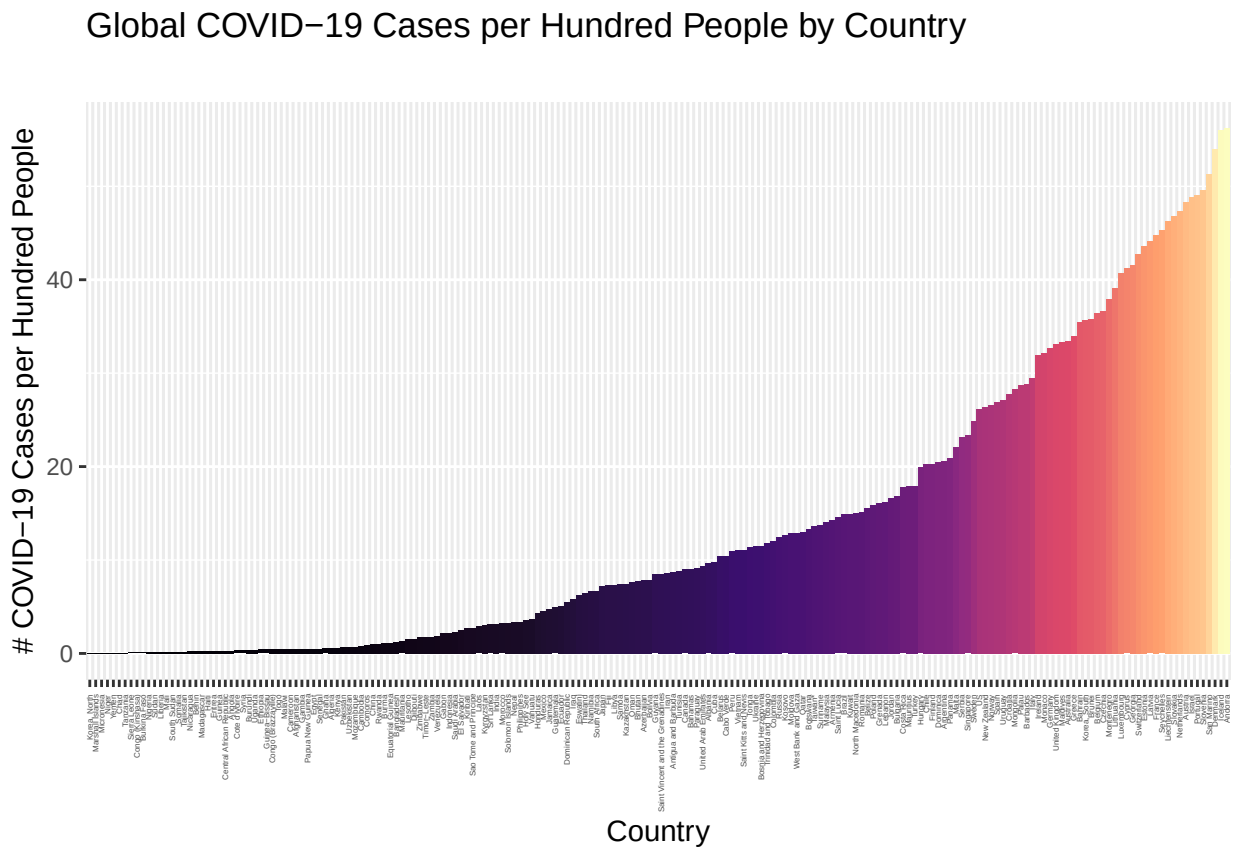
```
# Change 'Province_State' into factor
US_vaccinations <- mutate_at(US_vaccinations, vars('Province_State'), as.factor)

# Create 'US_by_state_vaccinations_per_hundred' df holding max vaccination rates per US state
US_by_state_vaccinations_per_hundred <- US_vaccinations %>%
  group_by(Province_State) %>%
  mutate(Province_State = fct_recode(Province_State,
    "New York" = "New York State")) %>%
  summarize(people_fully_vaccinated_per_hundred = max(people_fully_vaccinated_per_hundred, na.rm = T),
            total_vaccinations_per_hundred = max(total_vaccinations_per_hundred, na.rm = T),
            people_vaccinated_per_hundred = max(people_vaccinated_per_hundred, na.rm = T),
            distributed_per_hundred = max(distributed_per_hundred, na.rm = T),
            total_boosters_per_hundred = max(total_boosters_per_hundred, na.rm = T))

# Merge 'US_by_state_deaths_per_hundred' df and 'US_by_state_vaccinations_per_hundred' df
US_by_state_deaths_vaccinations_per_hundred <- US_by_state_deaths_per_hundred %>%
    full_join(US_by_state_vaccinations_per_hundred) %>%
  filter(Population > 0)
```

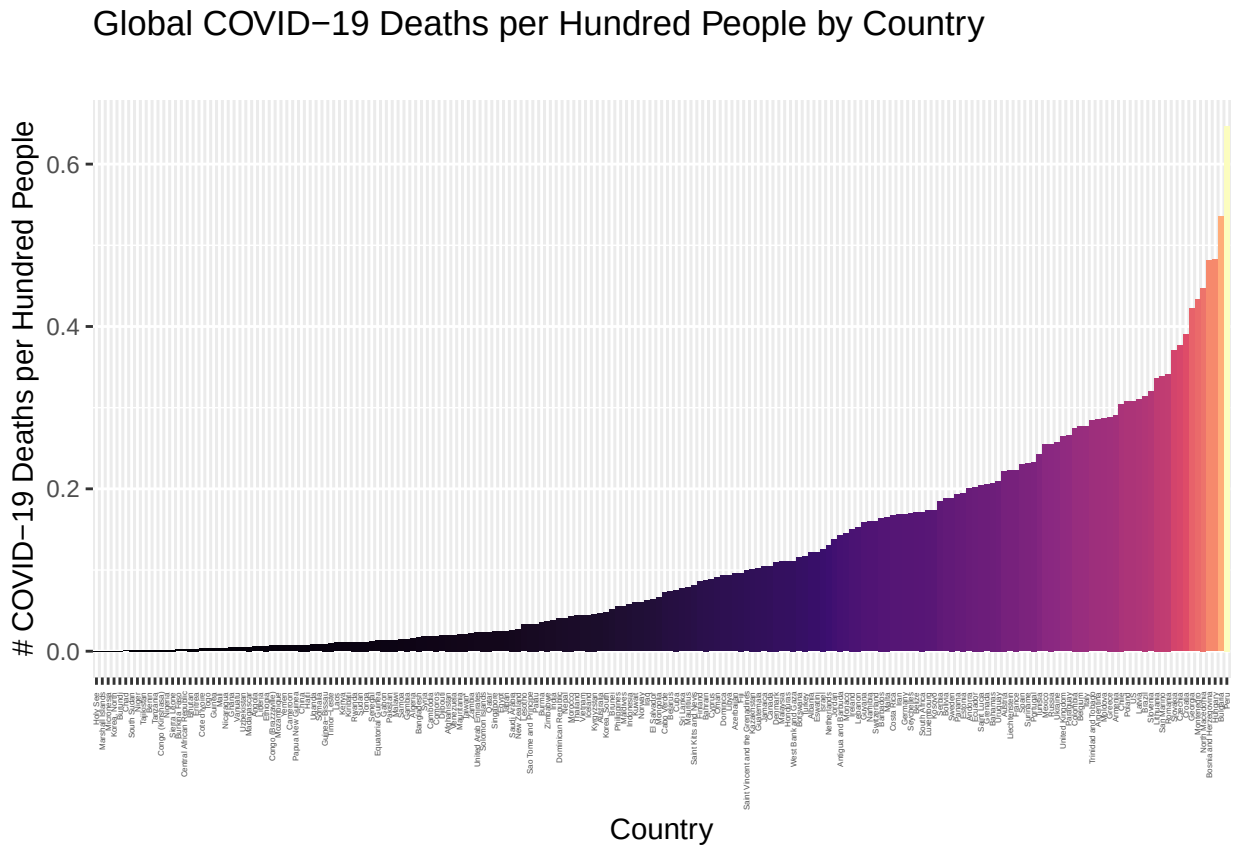
## 4 Data Visualization

```
ggplot(data = global_cases_per_hundred, aes(x = reorder(Country_Region, +Cases_per_hundred),
                                            y = Cases_per_hundred,
                                            fill = Cases_per_hundred)) +
  scale_fill_viridis_c(option = "magma") +
  geom_bar(stat = "identity") +
  labs(x = "Country",
       y = "# COVID-19 Cases per Hundred People",
       title = "Global COVID-19 Cases per Hundred People by Country",
       subtitle = "") +
  theme(legend.position="none", axis.text.x=element_text(angle=90, hjust=.98, vjust = .5, size=3))
```




Above is a plot of the number of reported COVID-19 cases in the countries of the world as of 2022-06-19. Each bar represents cases per one hundred members of the population in that country. You may need to zoom in to read the country names. These reported case numbers are strongly influenced by socio-economic and health infrastructure factors (such as the availability of COVID-19 tests).

```
ggplot(data = global_deaths_per_hundred, aes(x = reorder(Country_Region, +Deaths_per_hundred),
                                             y = Deaths_per_hundred,
                                             fill = Deaths_per_hundred)) +
  scale_fill_viridis_c(option = "magma") +
  geom_bar(stat = "identity") +
  labs(x = "Country",
       y = "# COVID-19 Deaths per Hundred People",
       title = "Global COVID-19 Deaths per Hundred People by Country",
       subtitle = "") +
  theme(legend.position="none", axis.text.x=element_text(angle=90, hjust=.98, vjust = .5, size=3))
```




This plot is similar to the previous one except that each bar represents the number of COVID-19 deaths per hundred people in the country's population. Note that the top countries in this graph are generally not the same as the top countries in the previous graph.

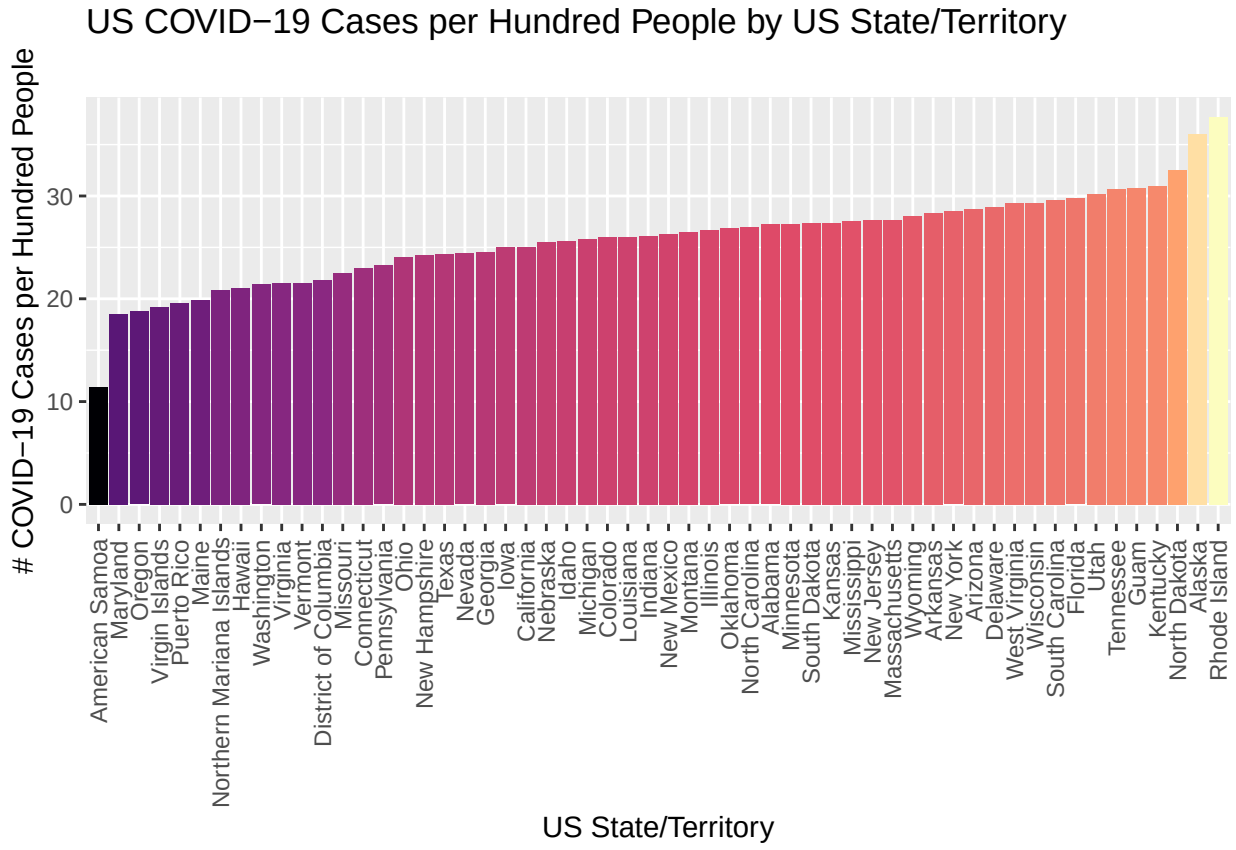
```
ggplot(data = US_by_state_cases_deaths_per_day, aes(x = Date, y = New_Cases,
                                                    color = Province_State)) +
    geom_line() +
    facet_wrap(~Province_State) +
    labs(x = "", y = "# of COVID-19 Cases", title = "COVID-19 Cases in the US",
         subtitle = "By State/Territory") +
    theme(legend.position="none")
```


Now shifting focus to the US, this graph shows a time-based plot of the number of COVID-19 cases recorded in each US state/territory between 01-23-2020 and the present.

```
ggplot(data = US_by_state_cases_deaths_per_day, aes(x = Date, y = New_Deaths,
                                                    color = Province_State)) +
    geom_line() +
    facet_wrap(~Province_State) +
    labs(x = "", y = "# of COVID-19 Deaths", title = "COVID-19 Deaths in the US",
         subtitle = "By State/Territory") +
    theme(legend.position="none")
```

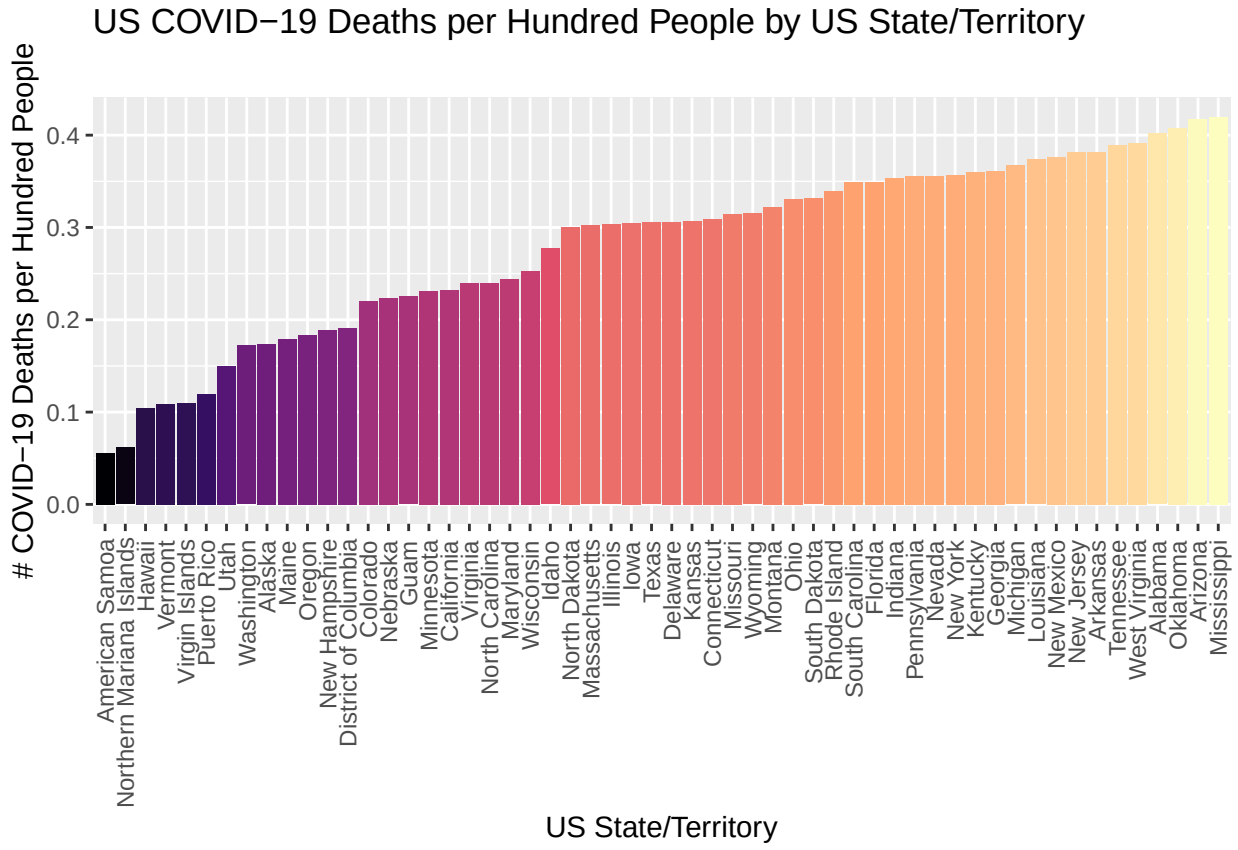

Shown here are time plots of COVID-19 deaths in each US state/territory between 01-23-2020 and the present. Notice the brief spikes in the number of deaths in some states. This is due to delays in data entry.

```
ggplot(data = US_by_state_cases_per_hundred, aes(x = reorder(Province_State, +Cases_per_hundred),
                                                 y = Cases_per_hundred,
                                                 fill = Cases_per_hundred)) +
  scale_fill_viridis_c(option = "magma") +
  geom_bar(stat = "identity") +
  labs(x = "US State/Territory",
       y = "# COVID-19 Cases per Hundred People",
       title = "US COVID-19 Cases per Hundred People by US State/Territory",
       subtitle = "") +
  theme(legend.position="none", axis.text.x=element_text(angle=90, hjust=.98, vjust = .5))
```




This plot shows the number of COVID-19 cases per one hundred members of the population in each US state/territory. Despite the variation in case numbers seen from state to state in the time plot on page 8, the plot above reveals that the per-state-population incidence of COVID-19 cases is relatively consistent across states in the US. For example, the number of COVID-19 cases per one hundred population members in Maryland is about half that in Rhode Island.

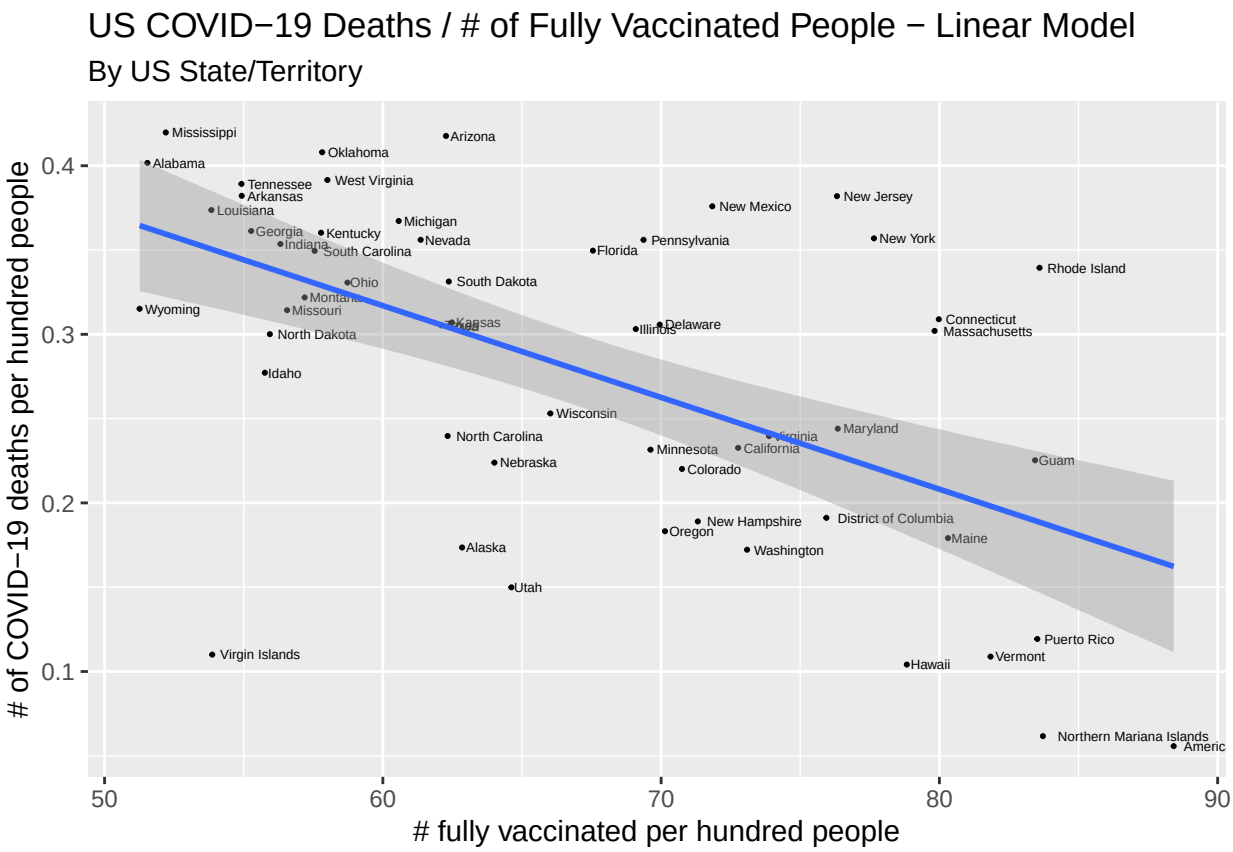
```
ggplot(data = US_by_state_deaths_per_hundred, aes(x = reorder(Province_State, +Deaths_per_hundred),
                                                  y = Deaths_per_hundred,
                                                  fill = Deaths_per_hundred)) +
  scale_fill_viridis_c(option="magma") +
  geom_bar(stat = "identity") +
  labs(x = "US State/Territory",
       y = "# COVID-19 Deaths per Hundred People",
       title = "US COVID-19 Deaths per Hundred People by US State/Territory",
       subtitle = "") +
  theme(legend.position="none", axis.text.x=element_text(angle=90, hjust=.98, vjust = .5))
```




Here the bar for each state represents the number of COVID-19 deaths per hundred people in that state. Whereas there was moderate variation from state to state on page 10 in the number of COVID-19 cases per 100 population members, in the plot above there is a wider variation across the states in the number of deaths per 100 population members. For example, the number of COVID-19 deaths per one hundred people in Hawaii is about one fourth that in Mississippi.

## 5 Data Modeling

```
ggplot(data = US_by_state_deaths_vaccinations_per_hundred, aes(x = people_fully_vaccinated_per_hundred,
                                                               y = Deaths_per_hundred,
                                                               label = Province_State)) +
    geom_point(size = .4) +
    geom_text(size = 1.7, vjust = .5, hjust = -.1) +
    geom_smooth(method = "lm") +
    labs(x = "# fully vaccinated per hundred people",
         y = "# of COVID-19 deaths per hundred people",
         title = "US COVID-19 Deaths / # of Fully Vaccinated People - Linear Model",
         subtitle = "By US State/Territory")
```




For this model, a simple linear regression is performed on the data for COVID-19 vaccination rates and death rates in the US to get some insight into the following question: ***"Is there a statistically significant relationship between the number of people fully vaccinated for COVID-19 per US state/territory and the number of COVID-19 deaths per US state/territory?"***

Looking at the graph above, the somewhat diffuse scattering of data points does nonetheless indicate a downward trend between the predictor variable (# of fully vaccinated people per 100 members of a state/territory population) and the outcome variable (# of deaths per 100 members of a state/territory population). As noted on page 11, there is a relatively large variation in the # of deaths per population in states across the US, and this is evident both in the death rate difference between states/territories within the 95% confidence bands (like Louisiana and Maine), and among outliers (like the Virgin Islands and Rhode Island).

```
fit <- lm(Deaths_per_hundred ~ people_fully_vaccinated_per_hundred,
          data = US_by_state_deaths_vaccinations_per_hundred)
summary(fit)
```

```
##
## Call:
## lm(formula = Deaths_per_hundred ~ people_fully_vaccinated_per_hundred,
##     data = US_by_state_deaths_vaccinations_per_hundred)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.240349 -0.052765  0.005194  0.045047  0.153791
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          0.643554   0.071245   9.033 2.20e-12 ***
## people_fully_vaccinated_per_hundred -0.005443   0.001057  -5.149 3.77e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07975 on 54 degrees of freedom
## Multiple R-squared:  0.3293, Adjusted R-squared:  0.3169
## F-statistic: 26.51 on 1 and 54 DF,  p-value: 3.773e-06
```

To determine if there is a statistically significant relationship in the US between the number of people in a state that are fully vaccinated for COVID-19 and the number of COVID-19 deaths in that state, the `summary()` function in R gives summary statistics for the linear regression model plotted on page 12.

In the summary statistics above, the model's `intercept` value is calculated as `.643554` and the coefficient of the `people_fully_vaccinated_per_hundred` variable is `-0.005443`. The intercept and the coefficient have p-values of `2.20e-12` and `3.77e-06`, respectively, suggesting both the intercept and the coefficient are statistically significant. With both p-values below the conventional threshold of .05 for statistical significance, we can reject the null hypothesis that a greater `t value` (i.e, `Estimate/Std. Error`) would be obtained were the the intercept and/or coefficient zero. This is corroborated by the `F-Statistic` value of `26.51` with a p-value of `3.773e-06`, suggesting that the predictor variable (`people_fully_vaccinated_per_hundred`) has a statistically significant effect on the outcome variable (`Deaths_per_hundred`).

However, the `Adjusted R-squared` value of `0.3169` suggests that there is only approximately 32% less variation around the regression line than there would be around the mean value of the `Deaths_per_hundred` variable with no regression line. Qualitatively, this can be seen in the dispersion of the data about the regression line.

## Potential bias in this report

This simple linear regression model does not take into account the multitude of factors that would affect COVID-19 death rates.

1) COVID-19 vaccines were not available until December 2020 in the US. If you look at the timeline for COVID-19 deaths in states like New York or New Jersey on page 9, a large portion of COVID-19 deaths in these states occurred before vaccines were available. This vaccine release-date factor would apply to all states to varying degrees.

2) Aside from vaccinations, other factors that would affect the COVID-19 death rate in states would be, in no particular order:

- Population density
- Availability of health care
- Age demographics
- Public health demographics
- Occupational exposure (i.e., “essential” workers)
- Socio-economic status

## References

https://github.com/CSSEGISandData/COVID-19
https://ourworldindata.org/covid-vaccinations
https://towardsdatascience.com/understanding-linear-regression-output-in-r-7a9cbda948b3
https://towardsdatascience.com/statistics-for-machine-learning-r-squared-explained-425ddfebf667
https://www.ajmc.com/view/a-timeline-of-covid-19-vaccine-developments-in-2021