# COVID-19 Final Project for 5301

## Introduction

In this report, we will analyze COVID-19 data for global and in the US. We will look at three main variables: cases, deaths and vaccine rates. We will look at how the data interacts and connects.

Let's import important libraries needed for this report.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tinytex)
library(ggplot2)
library(dplyr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

## Question and Quick Summary

As we all have experienced the COVID-19 (coronavirus) pandemic for over a year, it is important to start to look at possible trends and causes of case and death rates. Although there are many factors which can influence this, we will look at how the US data trends compare to the global trends. Then further on, we will analyze vaccine rates and their effects on cases and deaths. In our project we ask "What are the global case and death rates compared to the US rate and what factors could be influencing those differences and similarities?" After analysis of the data, we will summarize and give a conclusion for our finding.

## Import Data and Initial Look

First we need to import the data sets. We will import case and death data from the John Hopkins Github site. Then, we import US/global vaccine data from "Our World in Data" Github site, which will be used later on in the analysis and models.

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_cov
filenames <- c("time_series_covid19_confirmed_US.csv",
               "time_series_covid19_confirmed_global.csv",
               "time_series_covid19_deaths_US.csv",
               "time_series_covid19_deaths_global.csv")
urls <- str_c(url_in, filenames)
US_cases <- read_csv(urls[1])
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   iso2 = col_character(),
##   iso3 = col_character(),
##   Admin2 = col_character(),
##   Province_State = col_character(),
##   Country_Region = col_character(),
##   Combined_Key = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
global_cases <- read_csv(urls[2])
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   `Province/State` = col_character(),
##   `Country/Region` = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
US_deaths <- read_csv(urls[3])
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   iso2 = col_character(),
##   iso3 = col_character(),
##   Admin2 = col_character(),
##   Province_State = col_character(),
##   Country_Region = col_character(),
##   Combined_Key = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
global_deaths <- read_csv(urls[4])
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   `Province/State` = col_character(),
##   `Country/Region` = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
US_vacs <- read_csv('https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/vaccinatio
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   date = col_date(format = ""),
##   location = col_character(),
##   total_vaccinations = col_double(),
##   total_distributed = col_double(),
##   people_vaccinated = col_double(),
##   people_fully_vaccinated_per_hundred = col_double(),
##   total_vaccinations_per_hundred = col_double(),
##   people_fully_vaccinated = col_double(),
##   people_vaccinated_per_hundred = col_double(),
##   distributed_per_hundred = col_double(),
##   daily_vaccinations_raw = col_double(),
##   daily_vaccinations = col_double(),
##   daily_vaccinations_per_million = col_double(),
##   share_doses_used = col_double()
## )
```

```
global_vacs <- read_csv('https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/vaccina
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   location = col_character(),
##   iso_code = col_character(),
##   date = col_date(format = ""),
##   total_vaccinations = col_double(),
##   people_vaccinated = col_double(),
##   people_fully_vaccinated = col_double(),
##   total_boosters = col_logical(),
##   daily_vaccinations_raw = col_double(),
##   daily_vaccinations = col_double(),
##   total_vaccinations_per_hundred = col_double(),
##   people_vaccinated_per_hundred = col_double(),
##   people_fully_vaccinated_per_hundred = col_double(),
##   total_boosters_per_hundred = col_logical(),
##   daily_vaccinations_per_million = col_double()
## )
```

```
global_pop <- read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   UID = col_double(),
##   iso2 = col_character(),
##   iso3 = col_character(),
##   code3 = col_double(),
##   FIPS = col_character(),
##   Admin2 = col_character(),
##   Province_State = col_character(),
##   Country_Region = col_character(),
##   Lat = col_double(),
##   Long_ = col_double(),
##   Combined_Key = col_character(),
##   Population = col_double()
## )
```

Now that we have imported the data, we will take a preliminary look at what we have and what we will be looking at. This will help us identify what data is missing and what could be useful. We can also have a better understanding of the data sets to know how to organize, visualize, and later analyze it all.

```
head(global_cases)
```

```
## # A tibble: 6 x 618
##   `Province/State` `Country/Region`     Lat   Long `1/22/20` `1/23/20` `1/24/20`
##   <chr>            <chr>              <dbl>  <dbl>     <dbl>     <dbl>     <dbl>
## 1 <NA>             Afghanistan         33.9  67.7          0         0         0
## 2 <NA>             Albania             41.2  20.2          0         0         0
## 3 <NA>             Algeria             28.0   1.66         0         0         0
## 4 <NA>             Andorra             42.5   1.52         0         0         0
## 5 <NA>             Angola             -11.2  17.9          0         0         0
## 6 <NA>             Antigua and Barbu~  17.1 -61.8          0         0         0
## # ... with 611 more variables: 1/25/20 <dbl>, 1/26/20 <dbl>, 1/27/20 <dbl>,
## #   1/28/20 <dbl>, 1/29/20 <dbl>, 1/30/20 <dbl>, 1/31/20 <dbl>, 2/1/20 <dbl>,
## #   2/2/20 <dbl>, 2/3/20 <dbl>, 2/4/20 <dbl>, 2/5/20 <dbl>, 2/6/20 <dbl>,
## #   2/7/20 <dbl>, 2/8/20 <dbl>, 2/9/20 <dbl>, 2/10/20 <dbl>, 2/11/20 <dbl>,
## #   2/12/20 <dbl>, 2/13/20 <dbl>, 2/14/20 <dbl>, 2/15/20 <dbl>, 2/16/20 <dbl>,
## #   2/17/20 <dbl>, 2/18/20 <dbl>, 2/19/20 <dbl>, 2/20/20 <dbl>, 2/21/20 <dbl>,
## #   2/22/20 <dbl>, 2/23/20 <dbl>, 2/24/20 <dbl>, 2/25/20 <dbl>, 2/26/20 <dbl>,
## #   2/27/20 <dbl>, 2/28/20 <dbl>, 2/29/20 <dbl>, 3/1/20 <dbl>, 3/2/20 <dbl>,
## #   3/3/20 <dbl>, 3/4/20 <dbl>, 3/5/20 <dbl>, 3/6/20 <dbl>, 3/7/20 <dbl>,
## #   3/8/20 <dbl>, 3/9/20 <dbl>, 3/10/20 <dbl>, 3/11/20 <dbl>, 3/12/20 <dbl>,
## #   3/13/20 <dbl>, 3/14/20 <dbl>, 3/15/20 <dbl>, 3/16/20 <dbl>, 3/17/20 <dbl>,
## #   3/18/20 <dbl>, 3/19/20 <dbl>, 3/20/20 <dbl>, 3/21/20 <dbl>, 3/22/20 <dbl>,
## #   3/23/20 <dbl>, 3/24/20 <dbl>, 3/25/20 <dbl>, 3/26/20 <dbl>, 3/27/20 <dbl>,
## #   3/28/20 <dbl>, 3/29/20 <dbl>, 3/30/20 <dbl>, 3/31/20 <dbl>, 4/1/20 <dbl>,
## #   4/2/20 <dbl>, 4/3/20 <dbl>, 4/4/20 <dbl>, 4/5/20 <dbl>, 4/6/20 <dbl>,
## #   4/7/20 <dbl>, 4/8/20 <dbl>, 4/9/20 <dbl>, 4/10/20 <dbl>, 4/11/20 <dbl>,
## #   4/12/20 <dbl>, 4/13/20 <dbl>, 4/14/20 <dbl>, 4/15/20 <dbl>, 4/16/20 <dbl>,
## #   4/17/20 <dbl>, 4/18/20 <dbl>, 4/19/20 <dbl>, 4/20/20 <dbl>, 4/21/20 <dbl>,
```

```
## #   4/22/20 <dbl>, 4/23/20 <dbl>, 4/24/20 <dbl>, 4/25/20 <dbl>, 4/26/20 <dbl>,
## #   4/27/20 <dbl>, 4/28/20 <dbl>, 4/29/20 <dbl>, 4/30/20 <dbl>, 5/1/20 <dbl>,
## #   5/2/20 <dbl>, 5/3/20 <dbl>, ...
```

As we can see with the top first rows of data and the summary, we have many columns of data. We can see for global, we have these columns: Province/State, Country/Region, Lat, Long, and then hundreds of date input columns. We will repeat the process for all the data sets and make sure they are the same setup. Since there are so many date columns, I found it easier just to look at the head() of each data set for now.

```
## # A tibble: 6 x 618
##   `Province/State` `Country/Region`     Lat   Long `1/22/20` `1/23/20` `1/24/20`
##   <chr>            <chr>              <dbl>  <dbl>     <dbl>     <dbl>     <dbl>
## 1 <NA>             Afghanistan         33.9  67.7          0         0         0
## 2 <NA>             Albania             41.2  20.2          0         0         0
## 3 <NA>             Algeria             28.0   1.66         0         0         0
## 4 <NA>             Andorra             42.5   1.52         0         0         0
## 5 <NA>             Angola             -11.2  17.9          0         0         0
## 6 <NA>             Antigua and Barbu~  17.1 -61.8          0         0         0
## # ... with 611 more variables: 1/25/20 <dbl>, 1/26/20 <dbl>, 1/27/20 <dbl>,
## #   1/28/20 <dbl>, 1/29/20 <dbl>, 1/30/20 <dbl>, 1/31/20 <dbl>, 2/1/20 <dbl>,
## #   2/2/20 <dbl>, 2/3/20 <dbl>, 2/4/20 <dbl>, 2/5/20 <dbl>, 2/6/20 <dbl>,
## #   2/7/20 <dbl>, 2/8/20 <dbl>, 2/9/20 <dbl>, 2/10/20 <dbl>, 2/11/20 <dbl>,
## #   2/12/20 <dbl>, 2/13/20 <dbl>, 2/14/20 <dbl>, 2/15/20 <dbl>, 2/16/20 <dbl>,
## #   2/17/20 <dbl>, 2/18/20 <dbl>, 2/19/20 <dbl>, 2/20/20 <dbl>, 2/21/20 <dbl>,
## #   2/22/20 <dbl>, 2/23/20 <dbl>, 2/24/20 <dbl>, 2/25/20 <dbl>, 2/26/20 <dbl>,
## #   2/27/20 <dbl>, 2/28/20 <dbl>, 2/29/20 <dbl>, 3/1/20 <dbl>, 3/2/20 <dbl>,
## #   3/3/20 <dbl>, 3/4/20 <dbl>, 3/5/20 <dbl>, 3/6/20 <dbl>, 3/7/20 <dbl>,
## #   3/8/20 <dbl>, 3/9/20 <dbl>, 3/10/20 <dbl>, 3/11/20 <dbl>, 3/12/20 <dbl>,
## #   3/13/20 <dbl>, 3/14/20 <dbl>, 3/15/20 <dbl>, 3/16/20 <dbl>, 3/17/20 <dbl>,
## #   3/18/20 <dbl>, 3/19/20 <dbl>, 3/20/20 <dbl>, 3/21/20 <dbl>, 3/22/20 <dbl>,
## #   3/23/20 <dbl>, 3/24/20 <dbl>, 3/25/20 <dbl>, 3/26/20 <dbl>, 3/27/20 <dbl>,
## #   3/28/20 <dbl>, 3/29/20 <dbl>, 3/30/20 <dbl>, 3/31/20 <dbl>, 4/1/20 <dbl>,
## #   4/2/20 <dbl>, 4/3/20 <dbl>, 4/4/20 <dbl>, 4/5/20 <dbl>, 4/6/20 <dbl>,
## #   4/7/20 <dbl>, 4/8/20 <dbl>, 4/9/20 <dbl>, 4/10/20 <dbl>, 4/11/20 <dbl>,
## #   4/12/20 <dbl>, 4/13/20 <dbl>, 4/14/20 <dbl>, 4/15/20 <dbl>, 4/16/20 <dbl>,
## #   4/17/20 <dbl>, 4/18/20 <dbl>, 4/19/20 <dbl>, 4/20/20 <dbl>, 4/21/20 <dbl>,
## #   4/22/20 <dbl>, 4/23/20 <dbl>, 4/24/20 <dbl>, 4/25/20 <dbl>, 4/26/20 <dbl>,
## #   4/27/20 <dbl>, 4/28/20 <dbl>, 4/29/20 <dbl>, 4/30/20 <dbl>, 5/1/20 <dbl>,
## #   5/2/20 <dbl>, 5/3/20 <dbl>, ...
```

```
## # A tibble: 6 x 625
##        UID iso2  iso3  code3  FIPS Admin2  Province_State Country_Region   Lat
##      <dbl> <chr> <chr> <dbl> <dbl> <chr>   <chr>          <chr>          <dbl>
## 1 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
## 2 84001003 US    USA     840  1003 Baldwin Alabama        US              30.7
## 3 84001005 US    USA     840  1005 Barbour Alabama        US              31.9
## 4 84001007 US    USA     840  1007 Bibb    Alabama        US              33.0
## 5 84001009 US    USA     840  1009 Blount  Alabama        US              34.0
## 6 84001011 US    USA     840  1011 Bullock Alabama        US              32.1
## # ... with 616 more variables: Long_ <dbl>, Combined_Key <chr>, 1/22/20 <dbl>,
## #   1/23/20 <dbl>, 1/24/20 <dbl>, 1/25/20 <dbl>, 1/26/20 <dbl>, 1/27/20 <dbl>,
## #   1/28/20 <dbl>, 1/29/20 <dbl>, 1/30/20 <dbl>, 1/31/20 <dbl>, 2/1/20 <dbl>,
## #   2/2/20 <dbl>, 2/3/20 <dbl>, 2/4/20 <dbl>, 2/5/20 <dbl>, 2/6/20 <dbl>,
```

```
## #   2/7/20 <dbl>, 2/8/20 <dbl>, 2/9/20 <dbl>, 2/10/20 <dbl>, 2/11/20 <dbl>,
## #   2/12/20 <dbl>, 2/13/20 <dbl>, 2/14/20 <dbl>, 2/15/20 <dbl>, 2/16/20 <dbl>,
## #   2/17/20 <dbl>, 2/18/20 <dbl>, 2/19/20 <dbl>, 2/20/20 <dbl>, 2/21/20 <dbl>,
## #   2/22/20 <dbl>, 2/23/20 <dbl>, 2/24/20 <dbl>, 2/25/20 <dbl>, 2/26/20 <dbl>,
## #   2/27/20 <dbl>, 2/28/20 <dbl>, 2/29/20 <dbl>, 3/1/20 <dbl>, 3/2/20 <dbl>,
## #   3/3/20 <dbl>, 3/4/20 <dbl>, 3/5/20 <dbl>, 3/6/20 <dbl>, 3/7/20 <dbl>,
## #   3/8/20 <dbl>, 3/9/20 <dbl>, 3/10/20 <dbl>, 3/11/20 <dbl>, 3/12/20 <dbl>,
## #   3/13/20 <dbl>, 3/14/20 <dbl>, 3/15/20 <dbl>, 3/16/20 <dbl>, 3/17/20 <dbl>,
## #   3/18/20 <dbl>, 3/19/20 <dbl>, 3/20/20 <dbl>, 3/21/20 <dbl>, 3/22/20 <dbl>,
## #   3/23/20 <dbl>, 3/24/20 <dbl>, 3/25/20 <dbl>, 3/26/20 <dbl>, 3/27/20 <dbl>,
## #   3/28/20 <dbl>, 3/29/20 <dbl>, 3/30/20 <dbl>, 3/31/20 <dbl>, 4/1/20 <dbl>,
## #   4/2/20 <dbl>, 4/3/20 <dbl>, 4/4/20 <dbl>, 4/5/20 <dbl>, 4/6/20 <dbl>,
## #   4/7/20 <dbl>, 4/8/20 <dbl>, 4/9/20 <dbl>, 4/10/20 <dbl>, 4/11/20 <dbl>,
## #   4/12/20 <dbl>, 4/13/20 <dbl>, 4/14/20 <dbl>, 4/15/20 <dbl>, 4/16/20 <dbl>,
## #   4/17/20 <dbl>, 4/18/20 <dbl>, 4/19/20 <dbl>, 4/20/20 <dbl>, 4/21/20 <dbl>,
## #   4/22/20 <dbl>, 4/23/20 <dbl>, 4/24/20 <dbl>, 4/25/20 <dbl>, 4/26/20 <dbl>,
## #   4/27/20 <dbl>, 4/28/20 <dbl>, ...

## # A tibble: 6 x 626
##        UID iso2  iso3  code3  FIPS Admin2  Province_State Country_Region   Lat
##      <dbl> <chr> <chr> <dbl> <dbl> <chr>   <chr>          <chr>          <dbl>
## 1 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
## 2 84001003 US    USA     840  1003 Baldwin Alabama        US              30.7
## 3 84001005 US    USA     840  1005 Barbour Alabama        US              31.9
## 4 84001007 US    USA     840  1007 Bibb    Alabama        US              33.0
## 5 84001009 US    USA     840  1009 Blount  Alabama        US              34.0
## 6 84001011 US    USA     840  1011 Bullock Alabama        US              32.1
## # ... with 617 more variables: Long_ <dbl>, Combined_Key <chr>,
## #   Population <dbl>, 1/22/20 <dbl>, 1/23/20 <dbl>, 1/24/20 <dbl>,
## #   1/25/20 <dbl>, 1/26/20 <dbl>, 1/27/20 <dbl>, 1/28/20 <dbl>, 1/29/20 <dbl>,
## #   1/30/20 <dbl>, 1/31/20 <dbl>, 2/1/20 <dbl>, 2/2/20 <dbl>, 2/3/20 <dbl>,
## #   2/4/20 <dbl>, 2/5/20 <dbl>, 2/6/20 <dbl>, 2/7/20 <dbl>, 2/8/20 <dbl>,
## #   2/9/20 <dbl>, 2/10/20 <dbl>, 2/11/20 <dbl>, 2/12/20 <dbl>, 2/13/20 <dbl>,
## #   2/14/20 <dbl>, 2/15/20 <dbl>, 2/16/20 <dbl>, 2/17/20 <dbl>, 2/18/20 <dbl>,
## #   2/19/20 <dbl>, 2/20/20 <dbl>, 2/21/20 <dbl>, 2/22/20 <dbl>, 2/23/20 <dbl>,
## #   2/24/20 <dbl>, 2/25/20 <dbl>, 2/26/20 <dbl>, 2/27/20 <dbl>, 2/28/20 <dbl>,
## #   2/29/20 <dbl>, 3/1/20 <dbl>, 3/2/20 <dbl>, 3/3/20 <dbl>, 3/4/20 <dbl>,
## #   3/5/20 <dbl>, 3/6/20 <dbl>, 3/7/20 <dbl>, 3/8/20 <dbl>, 3/9/20 <dbl>,
## #   3/10/20 <dbl>, 3/11/20 <dbl>, 3/12/20 <dbl>, 3/13/20 <dbl>, 3/14/20 <dbl>,
## #   3/15/20 <dbl>, 3/16/20 <dbl>, 3/17/20 <dbl>, 3/18/20 <dbl>, 3/19/20 <dbl>,
## #   3/20/20 <dbl>, 3/21/20 <dbl>, 3/22/20 <dbl>, 3/23/20 <dbl>, 3/24/20 <dbl>,
## #   3/25/20 <dbl>, 3/26/20 <dbl>, 3/27/20 <dbl>, 3/28/20 <dbl>, 3/29/20 <dbl>,
## #   3/30/20 <dbl>, 3/31/20 <dbl>, 4/1/20 <dbl>, 4/2/20 <dbl>, 4/3/20 <dbl>,
## #   4/4/20 <dbl>, 4/5/20 <dbl>, 4/6/20 <dbl>, 4/7/20 <dbl>, 4/8/20 <dbl>,
## #   4/9/20 <dbl>, 4/10/20 <dbl>, 4/11/20 <dbl>, 4/12/20 <dbl>, 4/13/20 <dbl>,
## #   4/14/20 <dbl>, 4/15/20 <dbl>, 4/16/20 <dbl>, 4/17/20 <dbl>, 4/18/20 <dbl>,
## #   4/19/20 <dbl>, 4/20/20 <dbl>, 4/21/20 <dbl>, 4/22/20 <dbl>, 4/23/20 <dbl>,
## #   4/24/20 <dbl>, 4/25/20 <dbl>, 4/26/20 <dbl>, 4/27/20 <dbl>, ...
```

By looking at these data sets, we get a pretty good idea of what we will do to clean and tidy the data. For each state and country we have the number of cases and deaths, dates, and some extra information which we will not use. We also notice for US data, we have the population.

## Clean Data and Tidy

First, we will make it easier to work with my rotating the columns down to rows, making each row a different date per state or country.

```
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))

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

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

After looking at the summary, we notice we have many rows with no cases, let's filter out those rows. We also want to know if the max numbers could be a typo or true data. By looking at the max numbers, dates and locations, we can see it looks about right, so we will leave that for now.

```
global_cases_deaths <- global_cases_deaths %>% filter(cases>0)
global_cases_deaths %>% filter(cases >28000000)
```

We will now repeat this process for the USA data sets. We notice there are many columns we don't need for our analysis. Let's delete those columns and tidy it up. Finally, we will have a look at our summary of data.

```
US_cases <- US_cases %>% pivot_longer(cols = -(UID:Combined_Key),
               names_to = "date",
               values_to = "cases")

US_cases[c('UID', 'iso2', 'iso3', 'code3', 'FIPS')] <- NULL

US_cases <- US_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_))

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

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

```
summary(US_cases_deaths)
```

We notice that the US has two extra columns than the global data. These are the "Combined_Key" and "Population." To be able to to more accurately compare the two data sets, countries and states, we will need to use population and have the data sets with the same columns. For population, we put the csv file above with the other import files to keep it organized. Below, we will now add the population data to the global data.

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

global_pop[c("UID", "iso2", "iso3", "code3", "FIPS", "Lat", "Long_", "Combined_Key", "Admin2")] <- NULL

global_cases_deaths <- global_cases_deaths %>%
  left_join(global_pop, by = c("Province_State", "Country_Region"))
```

We have now completed tidying the cases and deaths data sets to the US and global, we will now clean and tidy the vaccine data sets. To start, we will delete columns we won't be using in this analysis. If you look at the vaccine data set on Github, it is recommended to use the daily_vaccinations column (instead of the daily_vaccinations_raw). We also will work with the full data. For example, the vaccination data has many columns with "per_hundred" or "per_million". Since we want to calculate our own rates with our data, we will also get rid of these. As well as booster data. Again, we are interested in the overall picture of the data trends; what happens to deaths and cases as more people are vaccinated? We will target this question in our analysis later.

```
global_vacs[c("iso_code", "total_boosters", "daily_vaccinations_raw",
              "total_vaccinations_per_hundred", "people_vaccinated_per_hundred",
              "people_fully_vaccinated_per_hundred", "daily_vaccinations_per_million",
              "total_boosters_per_hundred")] <- NULL

global_vacs <- global_vacs %>% rename(Country_Region = "location")

global <- global_cases_deaths %>%
  left_join(global_vacs, by = c("Country_Region", "date"))

US_vacs[c("total_distributed", "people_fully_vaccinated_per_hundred",
          "total_vaccinations_per_hundred", "people_vaccinated_per_hundred",
          "distributed_per_hundred", "daily_vaccinations_raw", "daily_vaccinations_per_million",
          "share_doses_used")] <- NULL

US_vacs <- US_vacs %>% rename(Province_State = "location")

US <- US_cases_deaths %>%
  left_join(US_vacs, by = c("Province_State", "date"))
US <- US %>% rename(County = "Admin2")
```

All of our data should be combined into two tables, global and US. Each one should have cases, deaths, population and vaccine data. Let's take a look at each one.

```
summary(global)
```

```
##  Combined_Key       Province_State     Country_Region          date
##  Length:155280      Length:155280      Length:155280      Min.   :2020-01-22
##  Class :character   Class :character   Class :character   1st Qu.:2020-07-25
##  Mode  :character   Mode  :character   Mode  :character   Median :2020-12-17
##                                                           Mean   :2020-12-15
##                                                           3rd Qu.:2021-05-09
##                                                           Max.   :2021-09-26
##
##      cases              deaths         Population        total_vaccinations
##  Min.   :       1   Min.   :     0   Min.   :8.090e+02   Min.   :0.000e+00
##  1st Qu.:     366   1st Qu.:     3   1st Qu.:8.964e+05   1st Qu.:8.890e+05
##  Median :    4300   Median :    65   Median :7.276e+06   Median :7.501e+06
##  Mean   :  322694   Mean   :  7416   Mean   :2.983e+07   Mean   :1.841e+08
##  3rd Qu.:   71082   3rd Qu.:  1262   3rd Qu.:3.102e+07   3rd Qu.:5.408e+07
##  Max.   :42931354   Max.   :688032   Max.   :1.380e+09   Max.   :2.200e+09
##                                      NA's   :2154        NA's   :114636
##  people_vaccinated   people_fully_vaccinated daily_vaccinations
##  Min.   :0.000e+00   Min.   :1.000e+00       Min.   :       0
##  1st Qu.:4.281e+05   1st Qu.:2.913e+05       1st Qu.:    4742
##  Median :2.834e+06   Median :1.472e+06       Median :   44491
##  Mean   :1.608e+07   Mean   :1.294e+07       Mean   : 1295027
##  3rd Qu.:1.708e+07   3rd Qu.:8.514e+06       3rd Qu.:  344868
##  Max.   :1.101e+09   Max.   :1.022e+09       Max.   :22424286
##  NA's   :122404      NA's   :125903          NA's   :94437
```

```
head(global)
```

```
## # A tibble: 6 x 11
##   Combined_Key Province_State Country_Region date       cases deaths Population
##   <chr>        <chr>          <chr>          <date>     <dbl>  <dbl>      <dbl>
## 1 Afghanistan  <NA>           Afghanistan    2020-02-24     5      0   38928341
## 2 Afghanistan  <NA>           Afghanistan    2020-02-25     5      0   38928341
## 3 Afghanistan  <NA>           Afghanistan    2020-02-26     5      0   38928341
## 4 Afghanistan  <NA>           Afghanistan    2020-02-27     5      0   38928341
## 5 Afghanistan  <NA>           Afghanistan    2020-02-28     5      0   38928341
## 6 Afghanistan  <NA>           Afghanistan    2020-02-29     5      0   38928341
## # ... with 4 more variables: total_vaccinations <dbl>, people_vaccinated <dbl>,
## #   people_fully_vaccinated <dbl>, daily_vaccinations <dbl>
```

```
summary(US)
```

```
##     County          Province_State     Country_Region     Combined_Key
##  Length:2051988     Length:2051988     Length:2051988     Length:2051988
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##       date                cases           Population           deaths
##  Min.   :2020-01-22   Min.   :      0   Min.   :       0   Min.   :    0.00
##  1st Qu.:2020-06-23   1st Qu.:     32   1st Qu.:    9917   1st Qu.:    0.00
##  Median :2020-11-23   Median :    624   Median :   24892   Median :   11.00
```

```
##  Mean   :2020-11-23   Mean   :   5075   Mean   :   99604   Mean   :   96.62
##  3rd Qu.:2021-04-26   3rd Qu.:   2752   3rd Qu.:   64979   3rd Qu.:   52.00
##  Max.   :2021-09-26   Max.   :1454172   Max.   :10039107   Max.   :26013.00
##
##  total_vaccinations people_vaccinated  people_fully_vaccinated
##  Min.   :     880   Min.   :     401   Min.   :       5
##  1st Qu.: 1274838   1st Qu.:  816244   1st Qu.:  468285
##  Median : 3189476   Median : 1845302   Median : 1408899
##  Mean   : 5732741   Mean   : 3288677   Mean   : 2552329
##  3rd Qu.: 6910257   3rd Qu.: 4048946   3rd Qu.: 3144755
##  Max.   :50366708   Max.   :28248056   Max.   :23215969
##  NA's   :1232988    NA's   :1233430    NA's   :1236889
##  daily_vaccinations
##  Min.   :-52445
##  1st Qu.:  9151
##  Median : 18814
##  Mean   : 36367
##  3rd Qu.: 44626
##  Max.   :494575
##  NA's   :1210056
```

```
head(US)
```

```
## # A tibble: 6 x 12
##   County  Province_State Country_Region Combined_Key date       cases Population
##   <chr>   <chr>          <chr>          <chr>        <date>     <dbl>      <dbl>
## 1 Autauga Alabama        US             Autauga, Al~ 2020-01-22     0      55869
## 2 Autauga Alabama        US             Autauga, Al~ 2020-01-23     0      55869
## 3 Autauga Alabama        US             Autauga, Al~ 2020-01-24     0      55869
## 4 Autauga Alabama        US             Autauga, Al~ 2020-01-25     0      55869
## 5 Autauga Alabama        US             Autauga, Al~ 2020-01-26     0      55869
## 6 Autauga Alabama        US             Autauga, Al~ 2020-01-27     0      55869
## # ... with 5 more variables: deaths <dbl>, total_vaccinations <dbl>,
## #   people_vaccinated <dbl>, people_fully_vaccinated <dbl>,
## #   daily_vaccinations <dbl>
```

We notice there is a good number of NA's in the vaccine data. From outside knowledge, we know that the vaccine data was not recorded each day and/or there is limited availability to report data in some countries. In addition, we haven't had vaccines since day-one of the coronavirus outbreak (vaccines were created and then distributed almost a year after the first outbreak). Therefore, since global population is missing so few rows, we will delete these rows. Then for NA's in the vaccine data, we will populate them with zeros.

```
global_clean <- global[!is.na(global$Population), ]

global_clean <- mutate_at(global_clean, c("total_vaccinations", "people_vaccinated",
                                "people_fully_vaccinated", "daily_vaccinations"),
                          ~replace(., is.na(.), 0))
summary(global_clean)
```

```
##  Combined_Key       Province_State     Country_Region          date
##  Length:153126      Length:153126      Length:153126      Min.   :2020-01-22
##  Class :character   Class :character   Class :character   1st Qu.:2020-07-25
##  Mode  :character   Mode  :character   Mode  :character   Median :2020-12-17
```

```
##                                                        Mean   :2020-12-14
##                                                        3rd Qu.:2021-05-08
##                                                        Max.   :2021-09-26
##      cases              deaths         Population        total_vaccinations
##  Min.   :       1   Min.   :     0   Min.   :8.090e+02   Min.   :0.000e+00
##  1st Qu.:     395   1st Qu.:     3   1st Qu.:8.964e+05   1st Qu.:0.000e+00
##  Median :    4696   Median :    70   Median :7.276e+06   Median :0.000e+00
##  Mean   :  327231   Mean   :  7520   Mean   :2.983e+07   Mean   :4.877e+07
##  3rd Qu.:   74150   3rd Qu.:  1320   3rd Qu.:3.102e+07   3rd Qu.:2.500e+04
##  Max.   :42931354   Max.   :688032   Max.   :1.380e+09   Max.   :2.200e+09
##  people_vaccinated   people_fully_vaccinated daily_vaccinations
##  Min.   :0.000e+00   Min.   :0.000e+00       Min.   :       0
##  1st Qu.:0.000e+00   1st Qu.:0.000e+00       1st Qu.:       0
##  Median :0.000e+00   Median :0.000e+00       Median :       0
##  Mean   :3.398e+06   Mean   :2.452e+06       Mean   :  513602
##  3rd Qu.:0.000e+00   3rd Qu.:0.000e+00       3rd Qu.:   13306
##  Max.   :1.101e+09   Max.   :1.022e+09       Max.   :22424286
```

```
US_clean <- mutate_at(US, c("total_vaccinations", "people_vaccinated",
                            "people_fully_vaccinated", "daily_vaccinations"),
                      ~replace(., is.na(.), 0))
summary(US_clean)
```

```
##     County          Province_State     Country_Region     Combined_Key
##  Length:2051988     Length:2051988     Length:2051988     Length:2051988
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##       date                cases           Population           deaths
##  Min.   :2020-01-22   Min.   :      0   Min.   :       0   Min.   :    0.00
##  1st Qu.:2020-06-23   1st Qu.:     32   1st Qu.:    9917   1st Qu.:    0.00
##  Median :2020-11-23   Median :    624   Median :   24892   Median :   11.00
##  Mean   :2020-11-23   Mean   :   5075   Mean   :   99604   Mean   :   96.62
##  3rd Qu.:2021-04-26   3rd Qu.:   2752   3rd Qu.:   64979   3rd Qu.:   52.00
##  Max.   :2021-09-26   Max.   :1454172   Max.   :10039107   Max.   :26013.00
##  total_vaccinations people_vaccinated  people_fully_vaccinated
##  Min.   :       0   Min.   :       0   Min.   :       0
##  1st Qu.:       0   1st Qu.:       0   1st Qu.:       0
##  Median :       0   Median :       0   Median :       0
##  Mean   : 2288081   Mean   : 1311885   Mean   : 1013846
##  3rd Qu.: 2143727   3rd Qu.: 1320748   3rd Qu.:  888344
##  Max.   :50366708   Max.   :28248056   Max.   :23215969
##  daily_vaccinations
##  Min.   :-52445
##  1st Qu.:     0
##  Median :     0
##  Mean   : 14921
##  3rd Qu.: 13893
##  Max.   :494575
```

Our data is looking much cleaner and seems ready to analyze. Let's visualuize the data and see how patterns look.

## Analyze Data - Visualizations

First let's take an overall look at the US data to see where we want to go with the data. First we will group by state then visualize it.

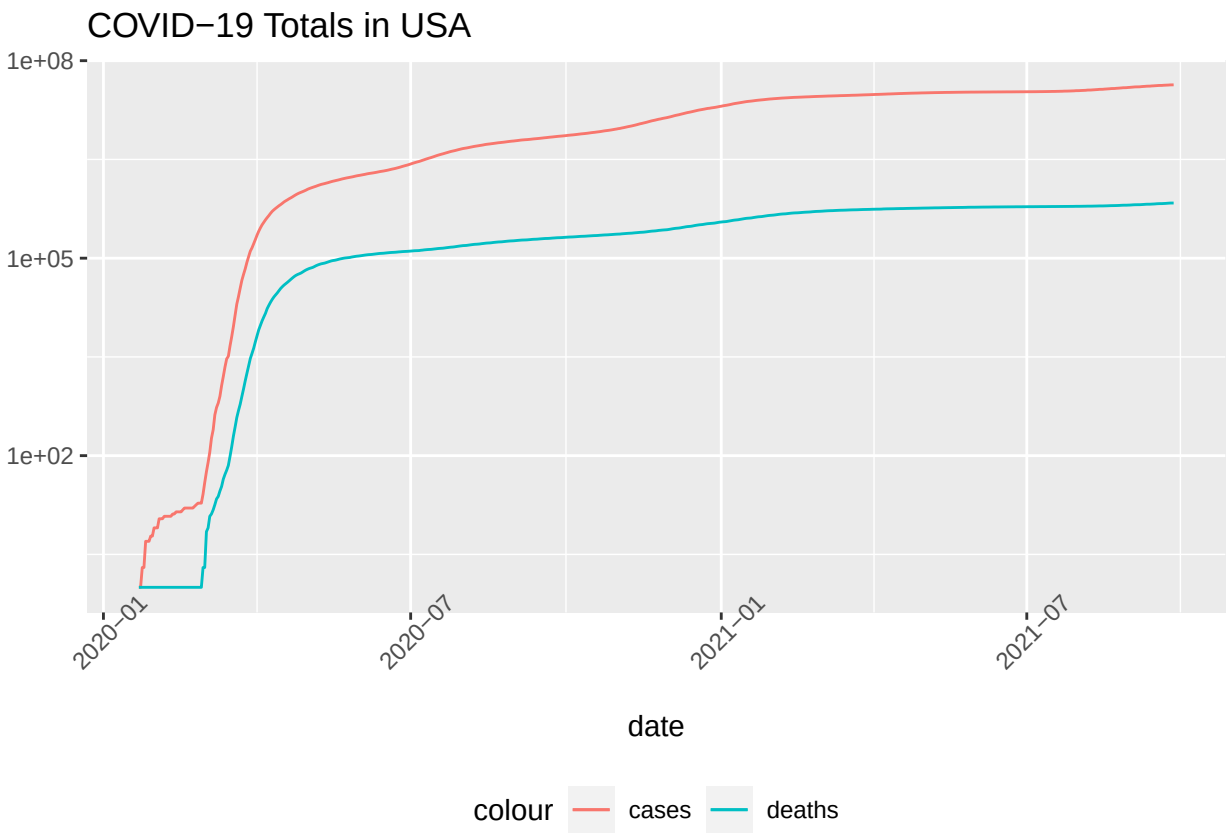
```
US_by_state <- US_clean %>% group_by(Province_State, Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths), Population = sum(Population),
            total_vacs = sum(total_vaccinations)) %>%
  mutate(deaths_per_hundred = deaths*100/Population) %>%
  select(Province_State, Country_Region, date, cases, deaths, deaths_per_hundred, Population, total_vacs
  ungroup()
```

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

```
US_totals <- US_by_state %>% group_by(Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths), Population = sum(Population),
            total_vacs = sum(total_vacs)) %>%
  mutate(deaths_per_hundred = deaths*100/Population) %>%
  select(Country_Region, date, cases, deaths, total_vacs, deaths_per_hundred, Population) %>%
  ungroup()
```

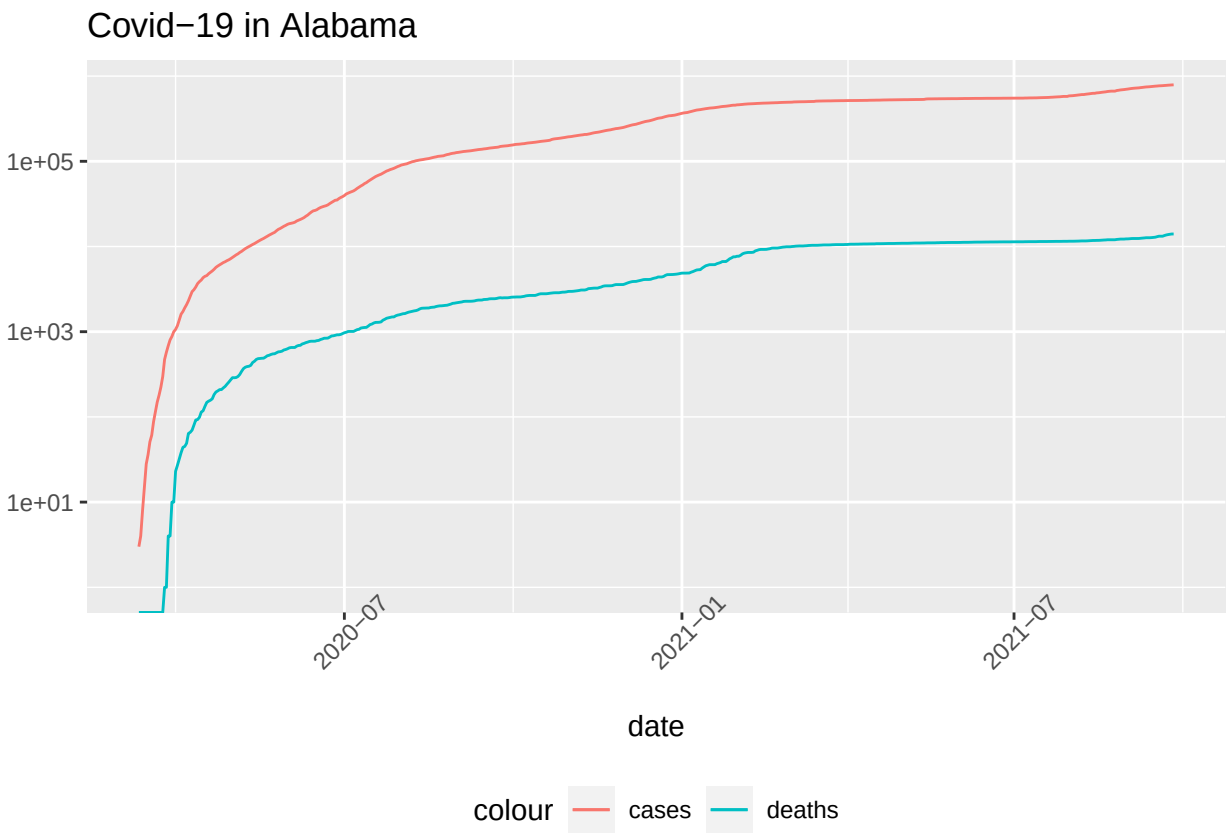
```
## `summarise()` has grouped output by 'Country_Region'. You can override using the `.groups` argument.
```

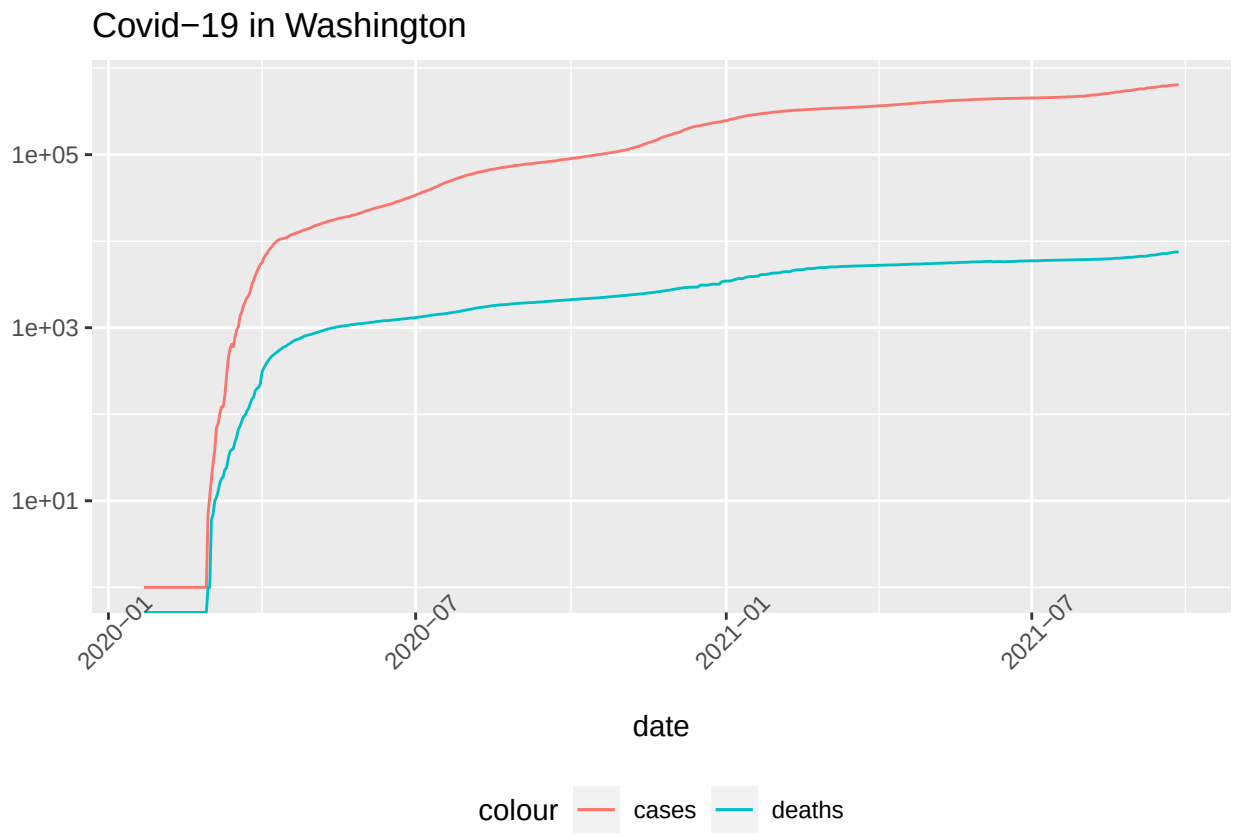
We can see that when we group the data by state, the cases, deaths and eventually vaccines increase. The same happens for grouping it by all of the US (therefore by looking at day-to-day data). But wait, shouldn't the cases and deaths fall as more vaccines are administered? We would hope so. But it is hard to tell exactly what the data is doing in table form. Therefore, we will visualize it with some graphs. We will first scale with log scale so that we get the trend and don't lose any detail.




As we can see, the cases and deaths had a sharp increase at the beginning then as time went on, the totals start to plateau suggesting that there was a quick increase of cases and deaths then a slower rate of cases and deaths, but still increasing.

This is interesting but it doesn't tell us very much about state to state. To get a better idea, lets focus on just two states. Randomly choosing two states, we will look at Alabama and Washington.


Covid−19 in Alabama
1e+05
1e+03
1e+01
2020−07
2021−01
2021−07
date
colour
cases
deaths




This is interesting but also expected. Washington and Alabama both had covid cases start around the same time (as most states did). Therefore we would assume their graphs to look quite similar, which they do. Let's now look at specific data to see where peaks might be.

```
max(US_totals$date)
```

```
## [1] "2021-09-26"
```

```
max(US_totals$cases)
```

```
## [1] 42931354
```

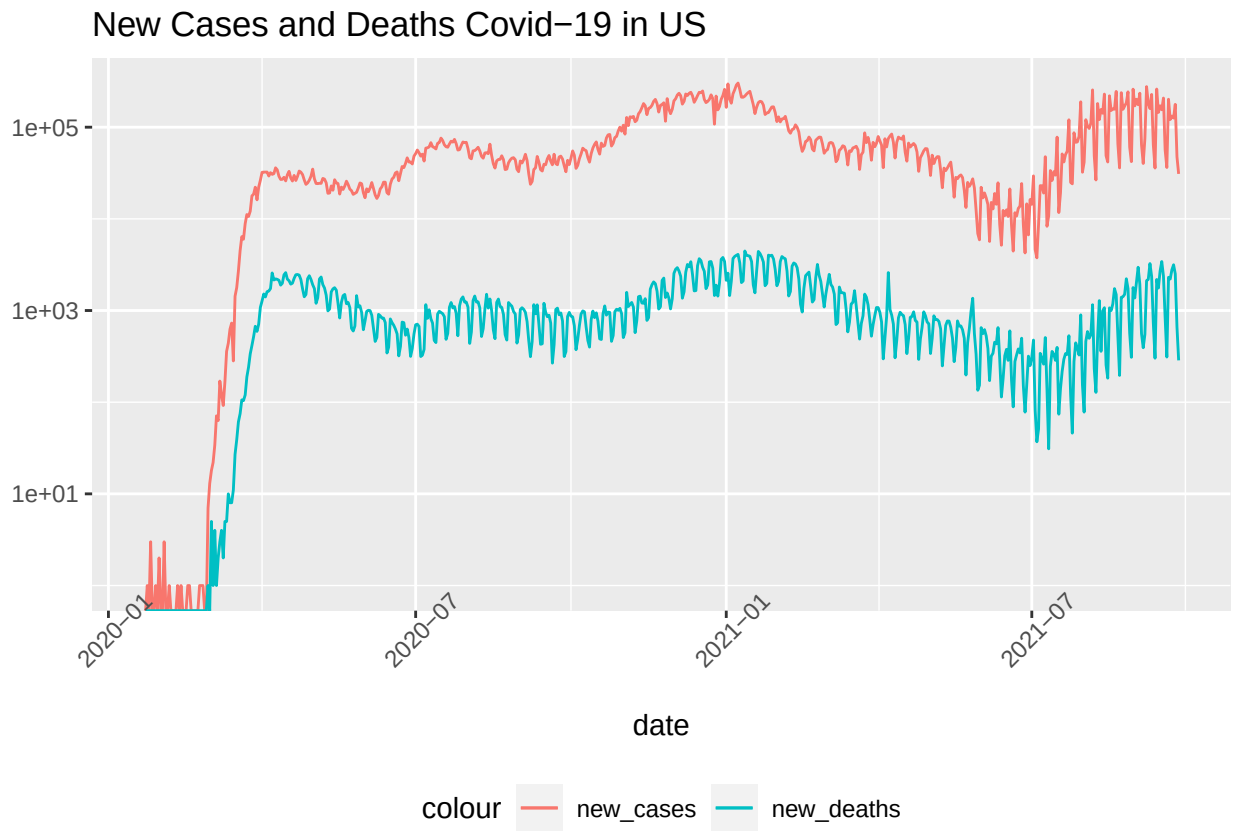
```
max(US_totals$deaths)
```

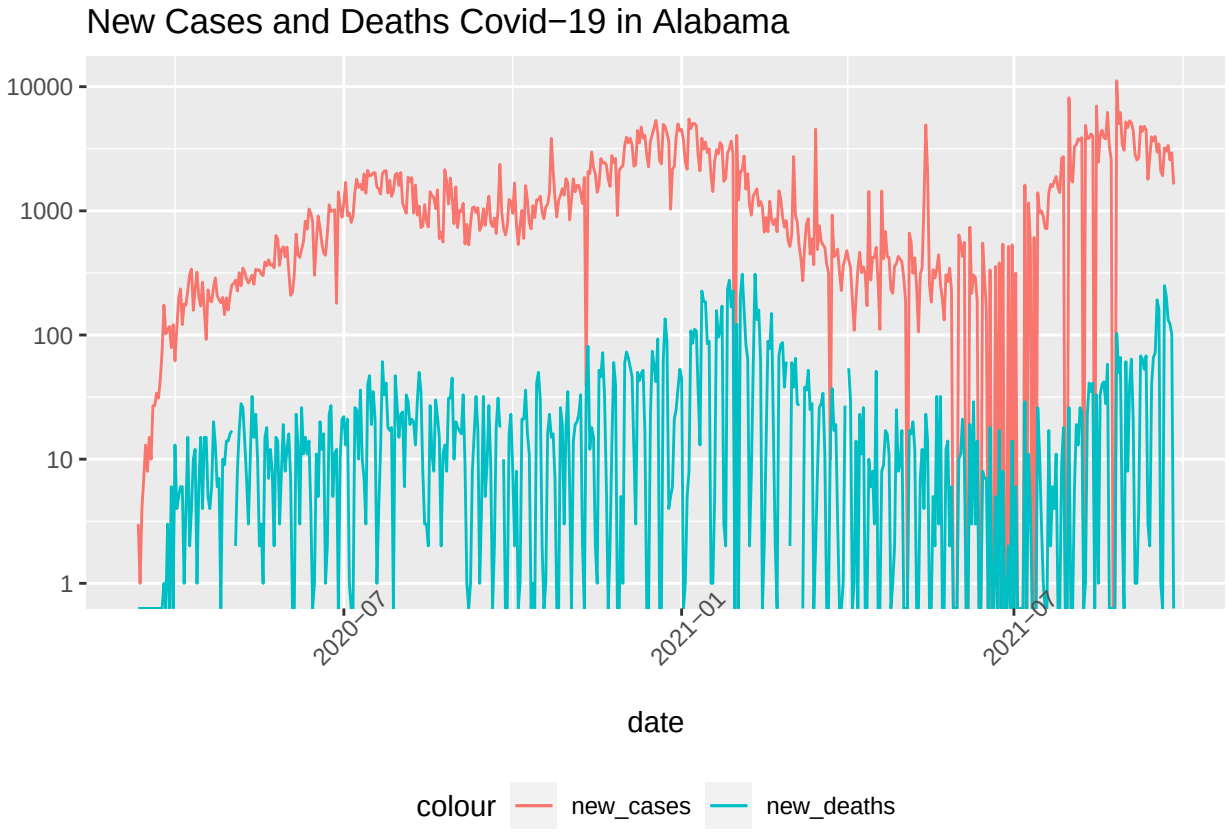
```
## [1] 688032
```

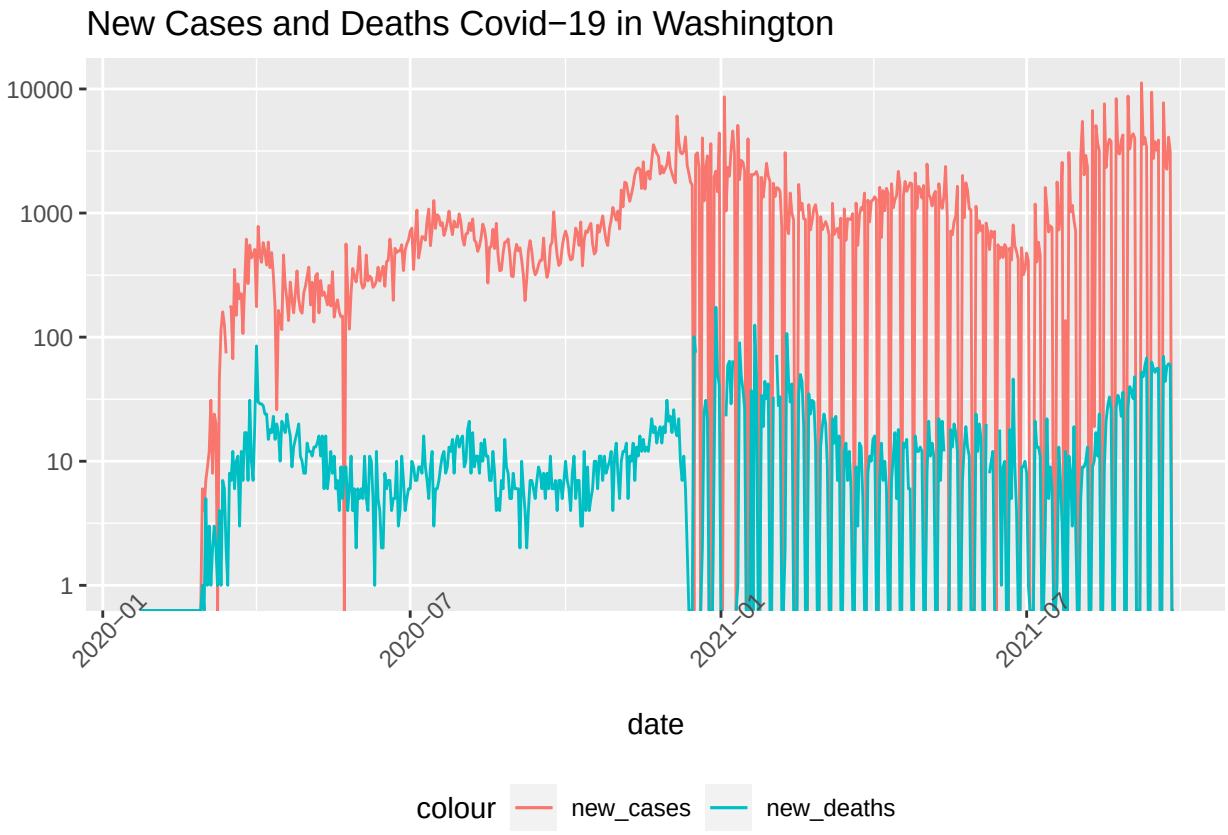
The peaks also fall in line with our experience with coronavirus and the graphs. By looking at the graphs above, we might think the cases and deaths are more steady but maybe that is just the way we charted it. For this reason, we will return to our data, reorganize and transform it some more then visualize it again.

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

First we will look at only the new cases and deaths per day for the US. That is, we will look at the change of cases and deaths as time passes. This will help us see a more accurate pattern of the data.




New Cases and Deaths Covid−19 in Alabama
10000
1000
100
10
1
2020−07
2021−01
2021−07
date
colour
new_cases
new_deaths




By looking at the overall US data, we see a similar trend of a quick start to the pandemic with rapid increase of cases and deaths, then as time passes people continue contract and die from covid, but just not at the extreme rates we saw at the beginning. Note also, these charts have an oscillating feature. We can best describe this with reporting tactics and human nature. What is meant to say is, we must take into account that many countries only reported cases and deaths on one day of the week each week. In addition, humans often will wait until Monday to go to the hospital instead of the weekend which could be contributing to the oscillating effect on the data. With such few data (cases and deaths only) further investigation would be needed to pin-point this exactly. We will discuss further in the conclusion and biases section.

Let's dive a bit deeper into our analysis. We will transform the data again to see the max numbers.

```
## # A tibble: 10 x 6
##    Province_State      deaths_per_thous cases_per_thous deaths  cases population
##    <chr>                          <dbl>           <dbl>  <dbl>  <dbl>      <dbl>
##  1 Northern Mariana I~           0.0363            4.81      2    265      55144
##  2 Vermont                       0.497            52.9     310  33031     623989
##  3 Hawaii                        0.535            55.2     757  78149    1415872
##  4 Virgin Islands                0.643            61.8      69   6631     107268
##  5 Alaska                        0.718           145.      532 107679     740995
##  6 Maine                         0.754            64.9    1013  87192    1344212
##  7 Puerto Rico                   0.834            48.1    3132 180666    3754939
##  8 Oregon                        0.873            76.1    3682 320990    4217737
##  9 Utah                          0.895           156.     2869 500698    3205958
## 10 Washington                    0.984            84.1    7494 640496    7614893

## # A tibble: 10 x 6
```

```
##    Province_State deaths_per_thous cases_per_thous deaths   cases population
##    <chr>                     <dbl>           <dbl>  <dbl>   <dbl>      <dbl>
##  1 Mississippi                3.17            162.   9425  482902    2976149
##  2 New Jersey                 3.08            129.  27328 1147199    8882190
##  3 Louisiana                  2.96            158.  13741  734524    4648794
##  4 Alabama                    2.86            161.  14022  789054    4903185
##  5 New York                   2.84            124.  55188 2408593   19453561
##  6 Arizona                    2.72            149.  19812 1084369    7278717
##  7 Massachusetts              2.69            116.  18541  802829    6892503
##  8 Rhode Island               2.66            161.   2823  170700    1059361
##  9 Arkansas                   2.52            163.   7590  492650    3017804
## 10 Florida                    2.49            167.  53580 3582807   21477737
```

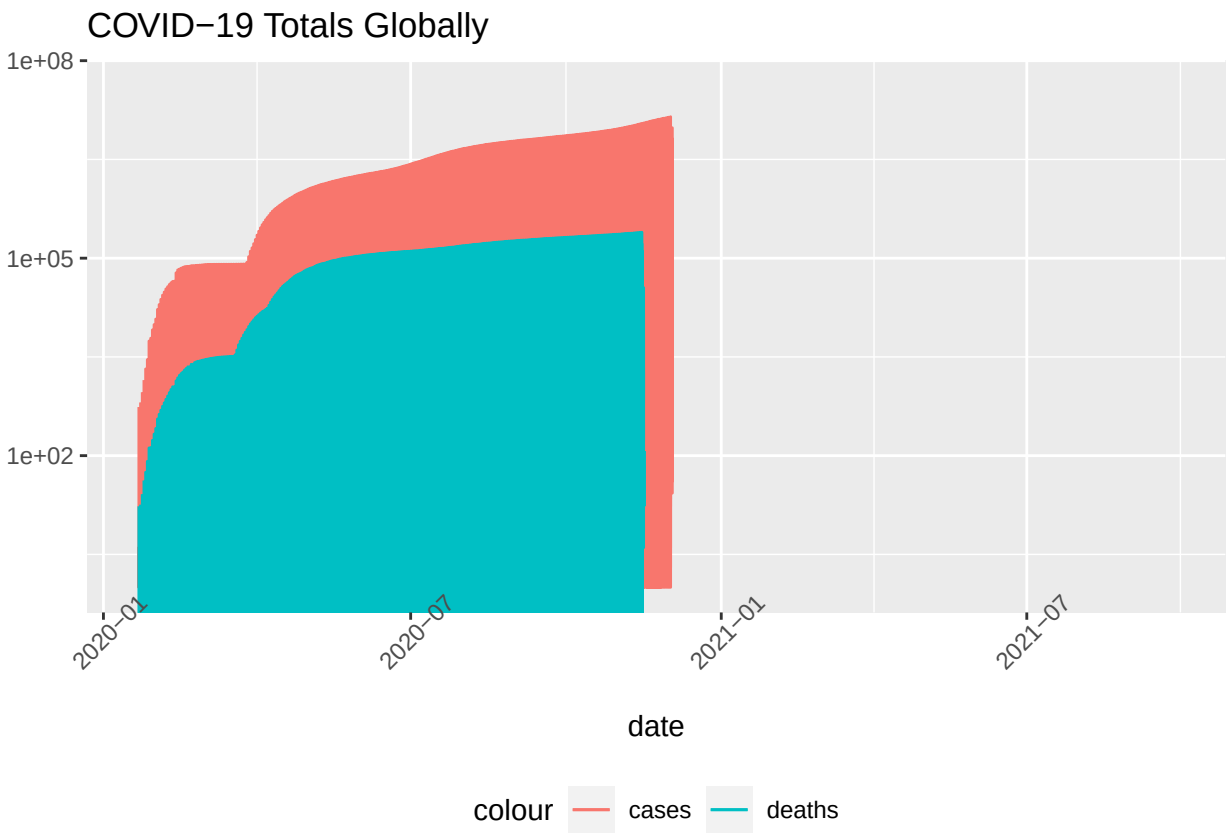
We really like what information we got so far from the US data. Since we want to compare it to the global data, let's transform the global data like we did with the US data.

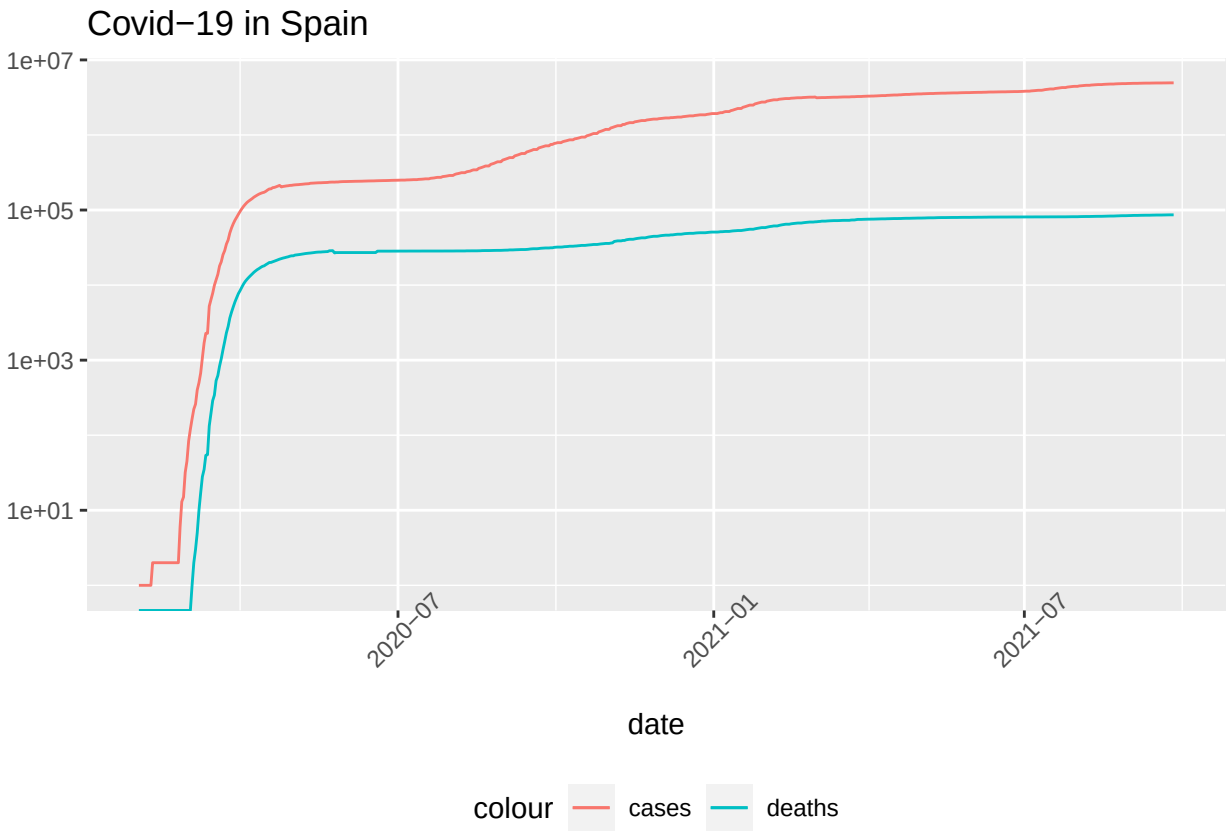
```
## `summarise()` has grouped output by 'Province_State', 'Country_Region'. You can override using the `
```

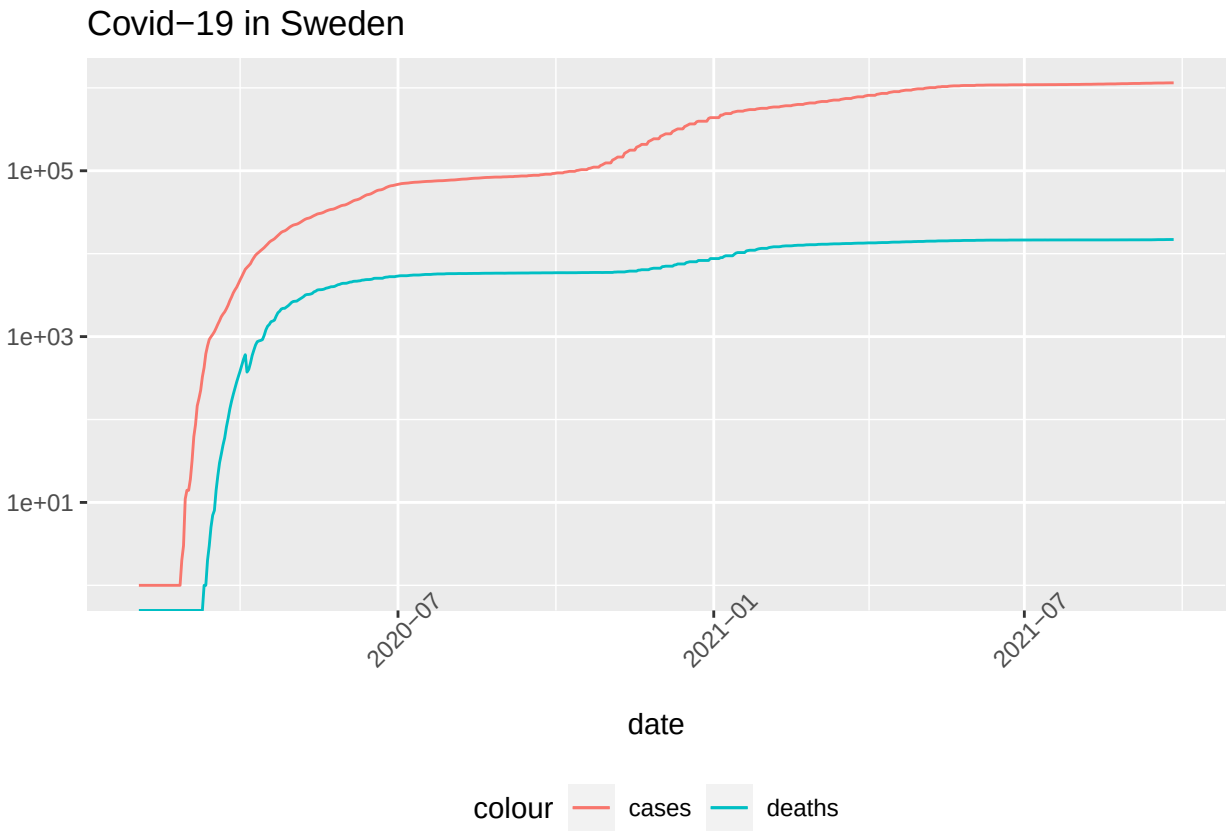
```
## # A tibble: 6 x 9
##   Province_State Country_Region date       cases deaths deaths_per_thous
##   <chr>          <chr>          <date>     <dbl>  <dbl>            <dbl>
## 1 Alberta        Canada         2020-03-06     1      0                0
## 2 Alberta        Canada         2020-03-07     2      0                0
## 3 Alberta        Canada         2020-03-08     4      0                0
## 4 Alberta        Canada         2020-03-09     7      0                0
## 5 Alberta        Canada         2020-03-10     7      0                0
## 6 Alberta        Canada         2020-03-11    19      0                0
## # ... with 3 more variables: cases_per_thous <dbl>, Population <dbl>,
## #   total_vacs <dbl>
```

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

```
## # A tibble: 6 x 8
##   Country_Region date       cases deaths total_vacs deaths_per_thous
##   <chr>          <date>     <dbl>  <dbl>      <dbl>            <dbl>
## 1 Afghanistan    2020-02-24     5      0          0                0
## 2 Afghanistan    2020-02-25     5      0          0                0
## 3 Afghanistan    2020-02-26     5      0          0                0
## 4 Afghanistan    2020-02-27     5      0          0                0
## 5 Afghanistan    2020-02-28     5      0          0                0
## 6 Afghanistan    2020-02-29     5      0          0                0
## # ... with 2 more variables: cases_per_thous <dbl>, Population <dbl>
```


COVID−19 Totals Globally
1e+08
1e+05
1e+02
2020−01
2020−07
2021−01
2021−07
date
colour
cases
deaths


Covid−19 in Spain
1e+07
1e+05
1e+03
1e+01
2020−07
2021−01
2021−07
date
colour
cases
deaths




We notice there is not as prominent of an oscillating effect with the global data. Again, with the data sets that we are given and using, it is impossible to pin-point what is causing these trends. For our analysis we are just looking at the broad overall picture. With that said, we see that the cases globally and in two countries, Spain and Sweden, we notice a similar trend as in the US. We chose to look at Spain becasue at the start of the pandemic it was the epicenter of outbreak for many weeks. We wanted to see how their case and death trends compared. Also, we looked at Sweden because they did very little at the start of their pandemic such as no primary lock-down like which was done in Spain and other European countries. It is interesting to see that all graphs we charted show similar trends.

## Analyze Data - Modeling

Now lets do some modeling of the data. We will first look at the US model and global model and compare them.

In this study, we will use a linear model to compare two variables.

```
deaths_cases_US_model <- lm(deaths_per_thous ~ cases_per_thous, data=US_state_totals)
summary(deaths_cases_US_model)
```

```
##
## Call:
## lm(formula = deaths_per_thous ~ cases_per_thous, data = US_state_totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4370 -0.3085 -0.0268  0.2748  1.1559
```

```
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.050033   0.246940   0.203     0.84
## cases_per_thous 0.014485   0.001923   7.533 6.27e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5145 on 53 degrees of freedom
## Multiple R-squared:  0.5171, Adjusted R-squared:  0.5079
## F-statistic: 56.74 on 1 and 53 DF,  p-value: 6.269e-10
```

```
#best case/deaths numbers
US_state_totals %>% slice_min(cases_per_thous, n=10) %>%
  select(Province_State,  cases_per_thous, deaths_per_thous, everything())
```

```
## # A tibble: 10 x 6
##    Province_State      cases_per_thous deaths_per_thous deaths  cases population
##    <chr>                         <dbl>            <dbl>  <dbl>  <dbl>      <dbl>
##  1 Northern Mariana I~            4.81           0.0363      2    265      55144
##  2 Puerto Rico                   48.1            0.834    3132 180666    3754939
##  3 Vermont                       52.9            0.497     310  33031     623989
##  4 Hawaii                        55.2            0.535     757  78149    1415872
##  5 Virgin Islands                61.8            0.643      69   6631     107268
##  6 Maine                         64.9            0.754    1013  87192    1344212
##  7 Oregon                        76.1            0.873    3682 320990    4217737
##  8 Washington                    84.1            0.984    7494 640496    7614893
##  9 District of Columb~           85.3            1.66     1172  60205     705749
## 10 New Hampshire                 86.4            1.08     1472 117454    1359711
```

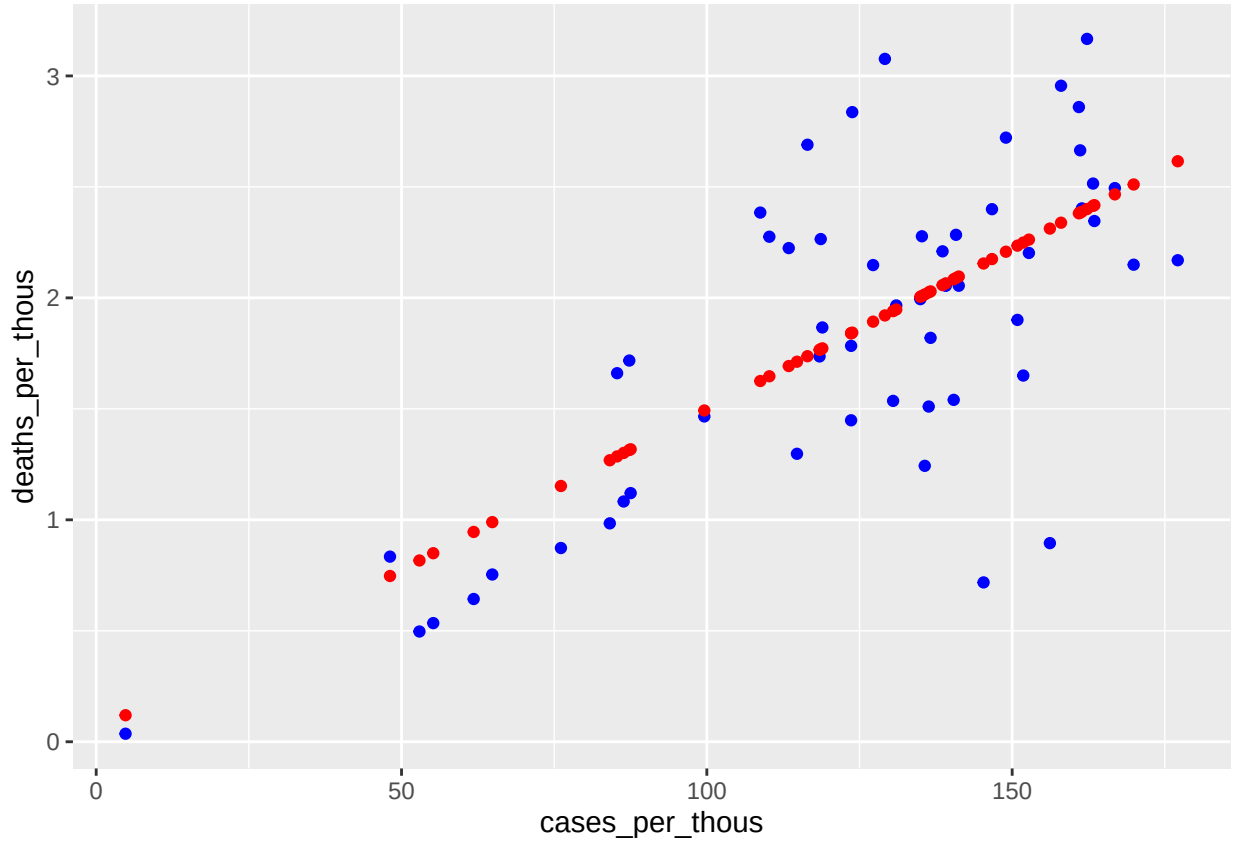
```
#worst case/deaths/numbers
US_state_totals %>% slice_max(cases_per_thous, n=10) %>%
  select(Province_State,  cases_per_thous, deaths_per_thous, everything())
```

```
## # A tibble: 10 x 6
##    Province_State cases_per_thous deaths_per_thous deaths   cases population
##    <chr>                    <dbl>            <dbl>  <dbl>   <dbl>      <dbl>
##  1 Tennessee                 177.             2.17  14817 1209568    6829174
##  2 North Dakota              170.             2.15   1638  129472     762062
##  3 Florida                   167.             2.49  53580 3582807   21477737
##  4 South Carolina            163.             2.35  12080  841600    5148714
##  5 Arkansas                  163.             2.52   7590  492650    3017804
##  6 Mississippi               162.             3.17   9425  482902    2976149
##  7 South Dakota              161.             2.40   2126  142800     884659
##  8 Rhode Island              161.             2.66   2823  170700    1059361
##  9 Alabama                   161.             2.86  14022  789054    4903185
## 10 Louisiana                 158.             2.96  13741  734524    4648794
```

```
#make a new df with predictions and raw totals
US_totals_predictions <- US_state_totals %>% mutate(deaths_pred = predict(deaths_cases_US_model))

US_totals_predictions %>% ggplot() +
```

```
  geom_point(aes(x=cases_per_thous, y=deaths_per_thous), color="blue") +
  geom_point(aes(x=cases_per_thous, y=deaths_pred), color="red")
```




This is a great start. We can see that cases and deaths are strongly connected and most likely impact each other, as we can see in the model summary. This intuitively makes sense as well, more cases probably leads to more deaths. By looking at the model plot, we can see at the lower end (fewer cases) the model is quite good at guessing the number of deaths, however as the cases increase, the accuracy is not as close (but still pretty good mostly). This suggests there are probably other factors influencing the deaths more than just cases.

Let's do this again with the global data.

```
deaths_cases_global_model <- lm(deaths_per_thous ~ cases_per_thous, data=global_totals)
summary(deaths_cases_global_model)
```

```
##
## Call:
## lm(formula = deaths_per_thous ~ cases_per_thous, data = global_totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3279 -0.0608 -0.0529 -0.0138  4.9574
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.359e-02  1.332e-03   40.23   <2e-16 ***
```

```
## cases_per_thous 1.568e-02  3.854e-05  406.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3774 on 107437 degrees of freedom
## Multiple R-squared:  0.6063, Adjusted R-squared:  0.6063
## F-statistic: 1.655e+05 on 1 and 107437 DF,  p-value: < 2.2e-16
```

```
#best case/deaths numbers
global_totals %>% slice_min(cases_per_thous, n=10) %>%
  select(Country_Region, cases_per_thous, deaths_per_thous, everything())
```

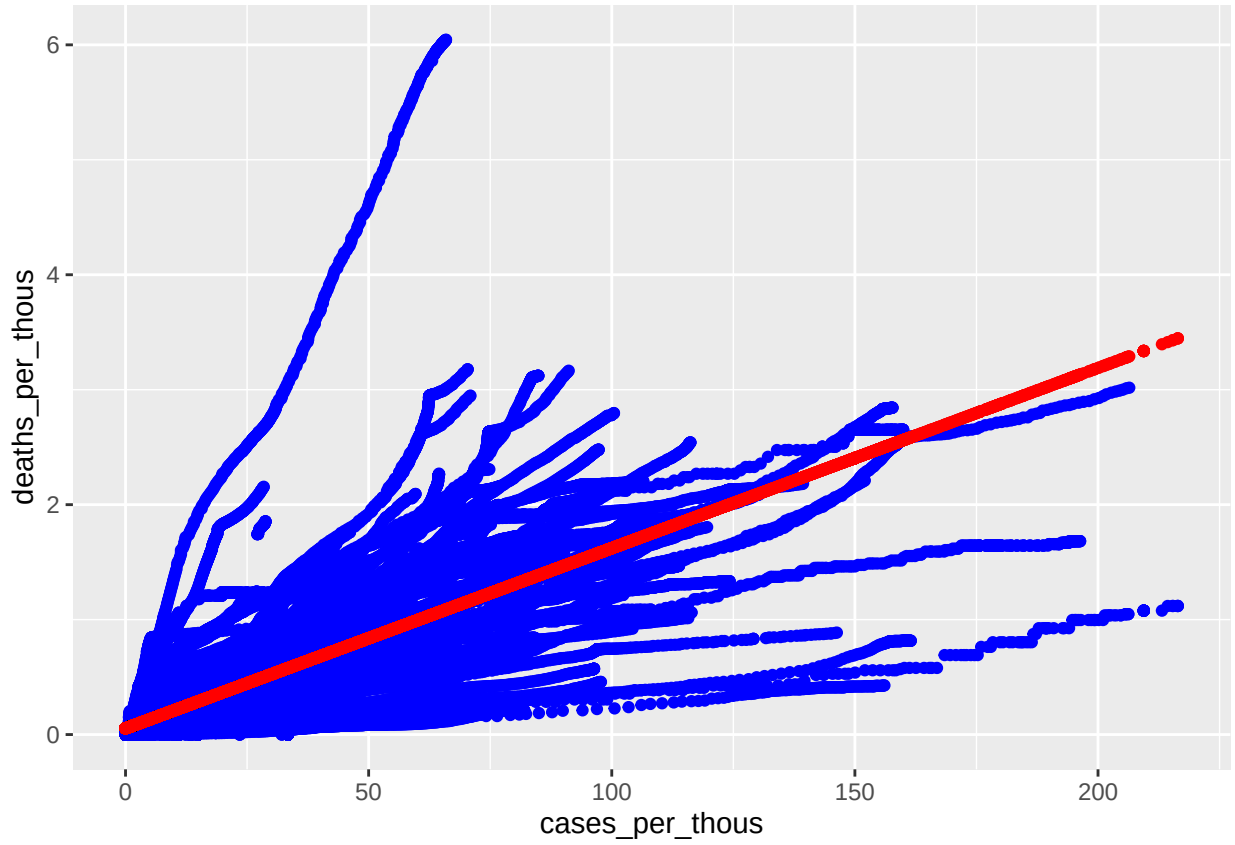
```
## # A tibble: 32 x 8
##    Country_Region cases_per_thous deaths_per_thous date       cases deaths
##    <chr>                   <dbl>            <dbl> <date>     <dbl>  <dbl>
##  1 India              0.000000725                0 2020-01-30     1      0
##  2 India              0.000000725                0 2020-01-31     1      0
##  3 India              0.000000725                0 2020-02-01     1      0
##  4 India              0.00000145                 0 2020-02-02     2      0
##  5 India              0.00000217                 0 2020-02-03     3      0
##  6 India              0.00000217                 0 2020-02-04     3      0
##  7 India              0.00000217                 0 2020-02-05     3      0
##  8 India              0.00000217                 0 2020-02-06     3      0
##  9 India              0.00000217                 0 2020-02-07     3      0
## 10 India              0.00000217                 0 2020-02-08     3      0
## # ... with 22 more rows, and 2 more variables: total_vacs <dbl>,
## #   Population <dbl>
```

```
#worst case/deaths/numbers
global_totals %>% slice_max(cases_per_thous, n=10) %>%
  select(Country_Region, cases_per_thous, deaths_per_thous, everything())
```

```
## # A tibble: 18 x 8
##    Country_Region cases_per_thous deaths_per_thous date       cases deaths
##    <chr>                   <dbl>            <dbl> <date>     <dbl>  <dbl>
##  1 Seychelles               216.             1.12 2021-09-24 21281    110
##  2 Seychelles               216.             1.12 2021-09-25 21281    110
##  3 Seychelles               216.             1.12 2021-09-26 21281    110
##  4 Seychelles               215.             1.12 2021-09-22 21180    110
##  5 Seychelles               215.             1.12 2021-09-23 21180    110
##  6 Seychelles               214.             1.12 2021-09-21 21077    110
##  7 Seychelles               213.             1.08 2021-09-20 20961    106
##  8 Seychelles               209.             1.08 2021-09-09 20593    106
##  9 Seychelles               209.             1.08 2021-09-10 20593    106
## 10 Seychelles               209.             1.08 2021-09-11 20593    106
## 11 Seychelles               209.             1.08 2021-09-12 20593    106
## 12 Seychelles               209.             1.08 2021-09-13 20593    106
## 13 Seychelles               209.             1.08 2021-09-14 20593    106
## 14 Seychelles               209.             1.08 2021-09-15 20593    106
## 15 Seychelles               209.             1.08 2021-09-16 20593    106
## 16 Seychelles               209.             1.08 2021-09-17 20593    106
## 17 Seychelles               209.             1.08 2021-09-18 20593    106
## 18 Seychelles               209.             1.08 2021-09-19 20593    106
## # ... with 2 more variables: total_vacs <dbl>, Population <dbl>
```

```
#make a new df with predictions and raw totals
global_predictions <- global_totals %>% ungroup()
global_predictions <- global_predictions %>%
  mutate(deaths_pred = predict(deaths_cases_global_model))

global_predictions %>% ggplot() +
  geom_point(aes(x=cases_per_thous, y=deaths_per_thous), color="blue") +
  geom_point(aes(x=cases_per_thous, y=deaths_pred), color="red")
```




These two models are really interesting. Notice in both models, the US and global models, that the number of cases is highly predictive of the number of deaths (notice the three stars in the model summaries). However, we can see that the global model did much better at predicting deaths for lower number of cases while the US data seems, visually, not as accurate. Notice also that the global data has many different types of trends by country. Some are more steady while others have higer deaths rates. While slight, this is also the case if we look at the summaries again. We notice the p-values and square-errors are higher for the US data than the global data, suggesting lower predictability for the US data, noting again that it is a slight difference.

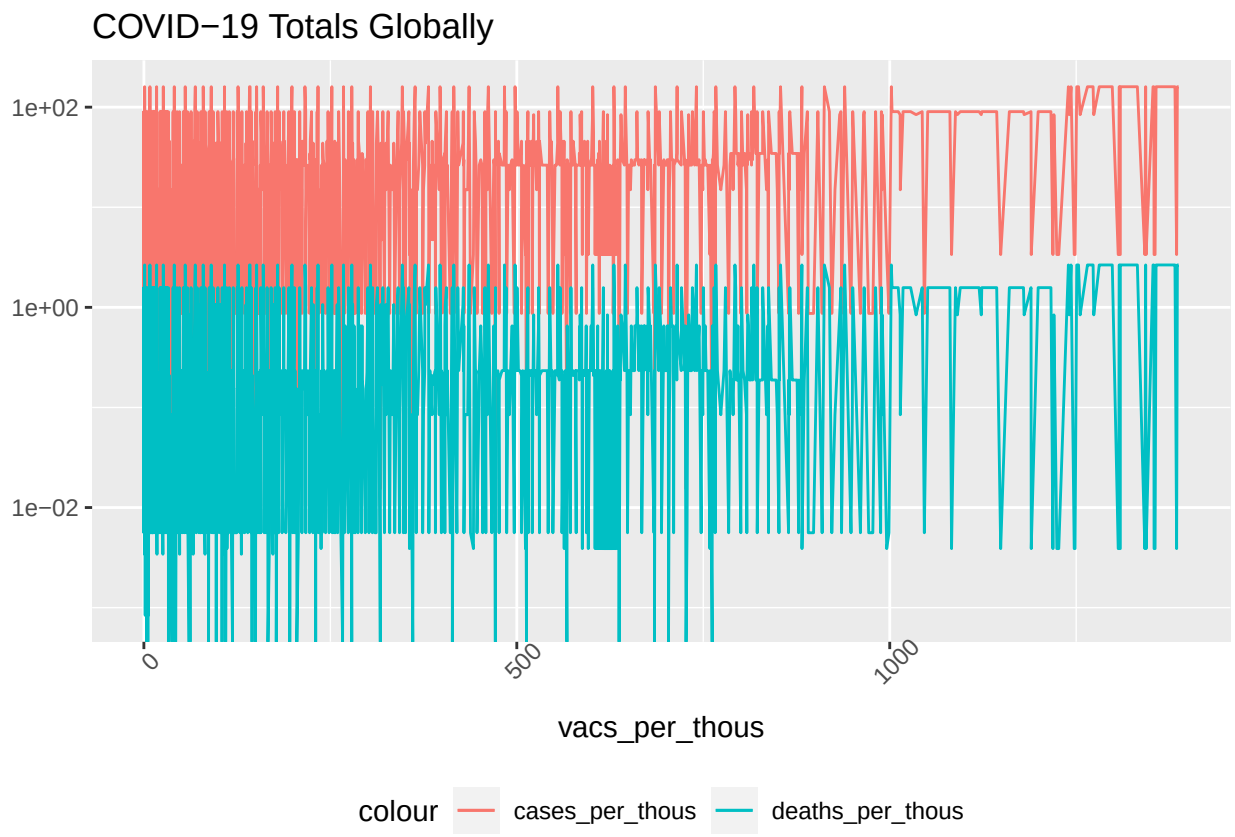
These models, as well as the graphs we visualized earlier, lead us to further questions. For example, what could be causing a slowing down of cases and deaths? Is it weather or location, culturally related, country wealth, vaccines, or a combination of these and/or more? The truth is that with our data it is impossible to know. Therefore we will try to get closer to some reasoning by introducing new data. We have imported above vaccine data above to help us try to understand the data more. Since we knew we were probably going to analyze the vaccine data, we went ahead and tidied, cleaned and organized the vaccine data right along with the rest of our data. In our global_totals and US_totals data sets we already have our vaccine data ready to plot and model. Let's take a look first at what we're working with here.

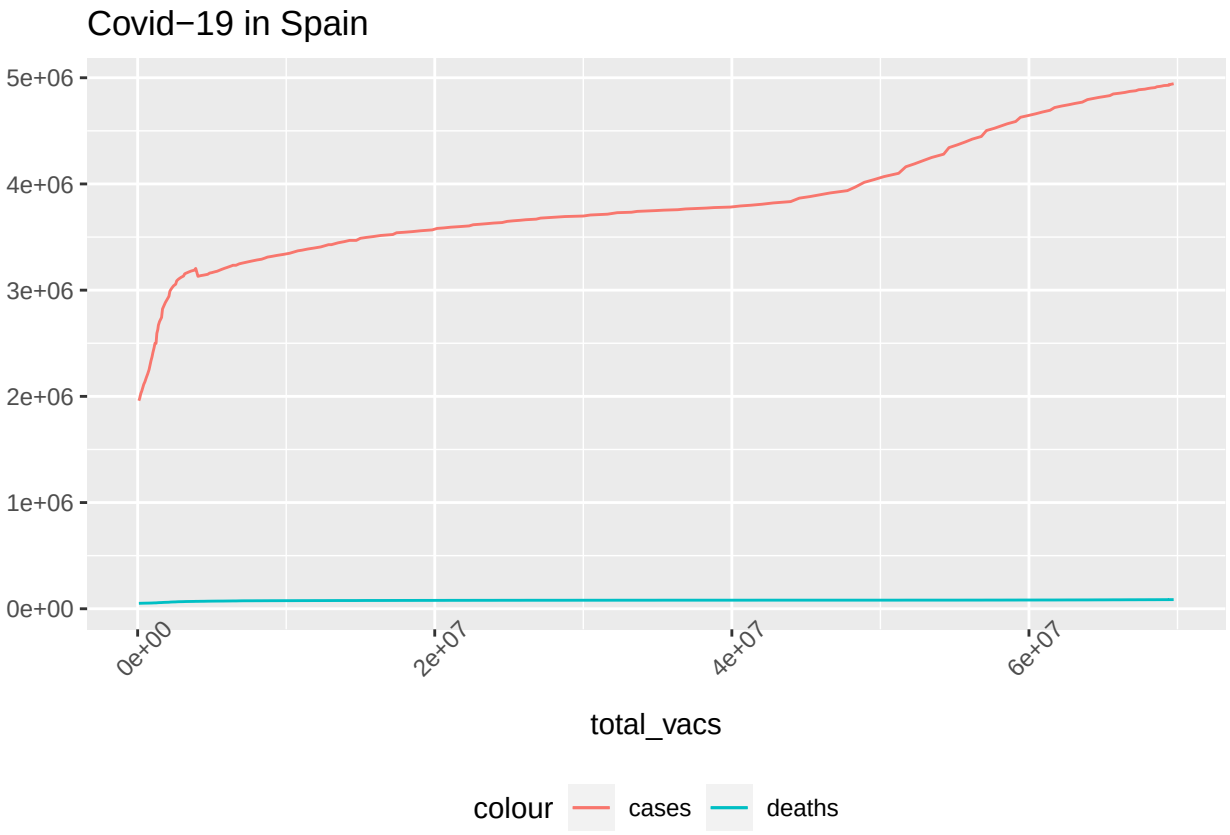
```
#global vaccine data
global_totals <- global_by_country %>%
  group_by(Country_Region) %>%
  summarize(deaths = max(deaths), cases = max(cases),
            population = max(Population),
            cases_per_thous = 1000*cases/population,
            deaths_per_thous = 1000*deaths/population,
            vacs_per_thous = 1000*total_vacs/population) %>%
  filter(cases>0, population>0, vacs_per_thous>0)
```

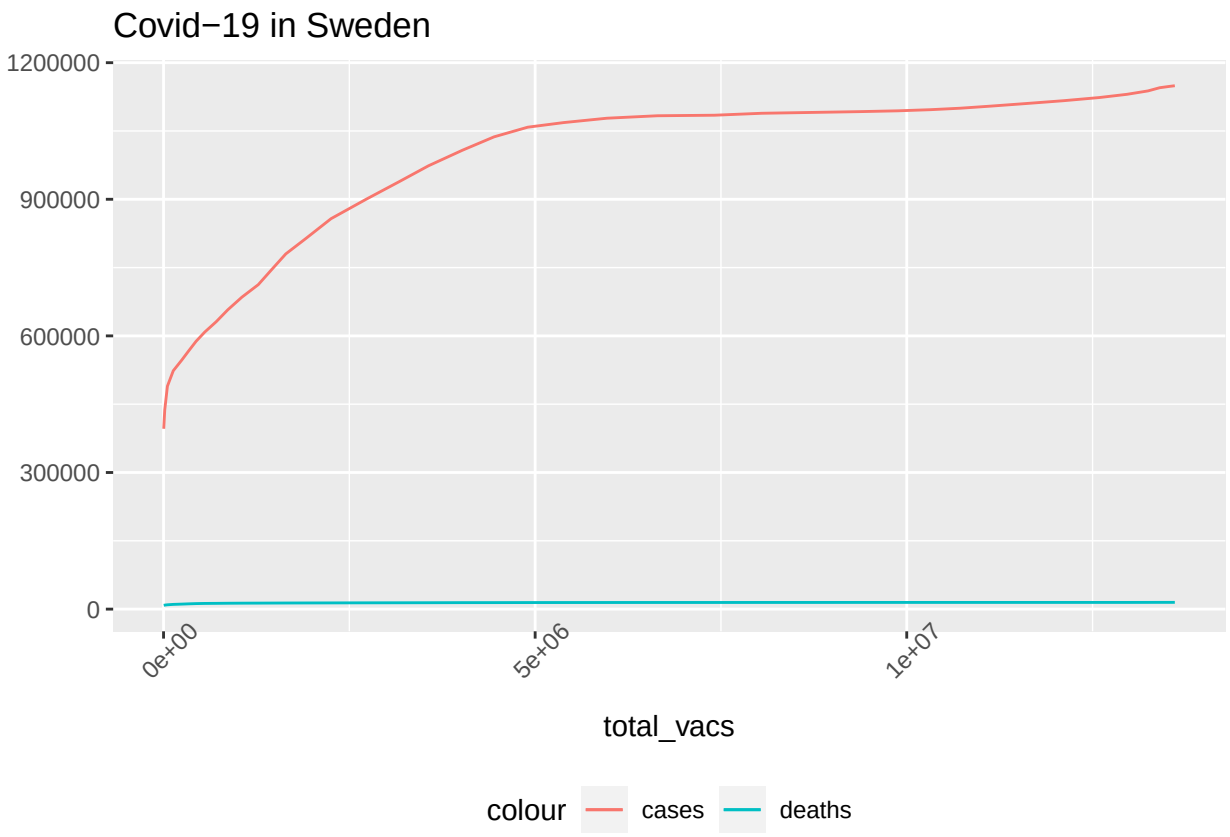
```
## `summarise()` has grouped output by 'Country_Region'. You can override using the `.groups` argument.
```

```
#US vaccine data
US_state_totals <- US_by_state %>%
  group_by(Province_State) %>%
  summarize(deaths = max(deaths), cases = max(cases),
            population = max(Population),
            cases_per_thous = 1000*cases/population,
            deaths_per_thous = 1000*deaths/population,
            vacs_per_thous = 1000*total_vacs/population) %>%
  filter(cases>0, population>0)
```

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




Covid−19 in Spain
5e+06
4e+06
3e+06
2e+06
1e+06
0e+00
0e+00
2e+07
4e+07
6e+07
total_vacs
colour
cases
deaths




First, we notice again that oscillating feature in the global data with regards to the vaccine rates, cases, and deaths. Notice this plot is different because we plotted the x-axis as the vaccine rate per thousand. We did this to see the change as vaccines when up, how did the cases and deaths fair. We would expect that the cases and, most importantly deaths, would curve and significantly slow down but we don't see a huge change with our data. There are many factors which could be impacting this. As we know, India had a huge increase in cases and deaths over the summer which definitely could have impacted the numbers. We also know that as weather is nicer, people like to go out more and mingle, leading to more cases and therefore deaths. Lastly, many countries lifted restrictions which could have lead to more cases and deaths, even as vaccines were increased. There are probably even more reasons which are beyond the scope of this analysis.

With regards to Spain and Sweden, we see that Spain actually had an increase of cases as more vaccines were administered, but the death rate stayed steady. This is promising, the vaccines were keeping people alive! Because I live in Spain, I know first hand that Spain has lifted almost all their restrictions which has allowed people to travel and interact more and therefore leading to more spread. However, the hospitals are not overwhelmed with patients and people aren't dying as often. For Sweden, we see a steady increase with not a significant change of cases or deaths. Again, this could be due to many outside factors which we will discuss later.

```
deaths_vacs_global_model <- lm(deaths_per_thous ~ vacs_per_thous, data=global_totals)
summary(deaths_vacs_global_model)
```

```
##
## Call:
## lm(formula = deaths_per_thous ~ vacs_per_thous, data = global_totals)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -1.1580 -0.7686 -0.1149  0.6473  4.9454
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.158e+00  4.820e-03  240.30   <2e-16 ***
## vacs_per_thous -7.781e-05  1.028e-06  -75.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8687 on 39891 degrees of freedom
## Multiple R-squared:  0.1256, Adjusted R-squared:  0.1256
## F-statistic:  5733 on 1 and 39891 DF,  p-value: < 2.2e-16
```

```
cases_vacs_global_model <- lm(cases_per_thous ~ vacs_per_thous, data=global_totals)
summary(cases_vacs_global_model)
```

```
##
## Call:
## lm(formula = cases_per_thous ~ vacs_per_thous, data = global_totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.541 -37.074  -3.345  39.296 157.450
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.556e+01  2.417e-01  271.23   <2e-16 ***
## vacs_per_thous -4.378e-03  5.153e-05  -84.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.56 on 39891 degrees of freedom
## Multiple R-squared:  0.1532, Adjusted R-squared:  0.1532
## F-statistic:  7217 on 1 and 39891 DF,  p-value: < 2.2e-16
```

By looking at the summary of the model, we see clearly that vaccines are connected to the cases and the number of deaths.

We have so much to think about now. Let's discuss further in the conclusion below.

## Conclusions

For our conclusion, we can confidently say that deaths, cases, and vaccine rates are all strongly connected, which answers our primary question. We notice that as vaccine rates increases, deaths stayed steady. Case numbers fluctuated more. We conclude based on our study and on prior knowledge, that the coronavirus vaccines are helping to curve the death rates globally.

However, our report is incomplete because there must be more factors at play which should be analyzed further on. For example, we could look at mask wearing trends, country wealth, access to vaccines, access to medical help, transportation systems, vaccines bought and administered, weather trends, and so much more. We must also question how often the data was reported and if the reported data is trustworthy? Are the countries reporting their data accurately? Are cases/deaths/vaccines reported in the same way in each country? These are import questions which should be persued in further analysis.

## Biases

Prior to starting this report, I did have some biases which are important to identify now. I thought that the wealthier countries will have higher rates of vaccines because they can afford them. However, many countries have had a high rate of people that refused to get vaccinated. Lower income countries might have lower rates of vaccines and/or higher rates of infection due to lack of vaccines and lack of medical records/access to medical assistance. Also, as someone from a medical family, I think vaccines are important and work. Therefore, I was hoping and expecting that vaccines would greatly decrease the rates of infection. While looking at the data, to try to prevent my biases from creeping in, I made sure to do multiple data analyzes. I also made sure to manipulate the data to look at it in multiple ways instead of just taking the first plot and moving on. Hopefully, this report is conclusive enough to get an idea of trends without incorporating biases into it.

## Appendix

- Cases and deaths data from John Hopkins Github site to get the COVID-19 Data sets: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series (https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/ was used as "url-in")
- World and state vaccine data from Our World in Data on Github: https://github.com/owid/covid-19-data/tree/master/public/data/vaccinations