Covid-19 Data

2022-11-18

INTRO

I will be going through the Covid-19 data. The primary question I am interested in is what was the effect of the Covid-19 vaccine on Covid-19 cases and deaths. I will be going through the Covid-19 data that was looked in class and Vaccine data from the site Listed below.

My hypothesis is that the number of deaths_per_thou should be lower in the states with the highest vaccination rates.

In order to build this RMD file we the following packages: library(tidyverse) library(lubridate)

Tidy the Data

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
## Loading required package: timechange
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
# Just assign a name to a variable. (str)
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_

# Same as above... create an array with 4 names.
file_names <- c("time_series_covid19_confirmed_US.csv", "time_series_covid19_deaths_US.csv" )

# concatinate each one with the url_in.
```

```
    urls <- str_c(url_in,file_names)

    urlvaccine <- "https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/vaccinations
    # read those and assign to variables.

    us_cases <- read_csv(urls[1])
```

```
## Rows: 3342 Columns: 1151
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1145): UID, code3, FIPS, Lat, Long_, 1/22/20, 1/23/20, 1/24/20, 1/25/20...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

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

```
## Rows: 3342 Columns: 1152
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1146): UID, code3, FIPS, Lat, Long_, Population, 1/22/20, 1/23/20, 1/24...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
    Vaccine_data <- read_csv(urlvaccine)
```

```
## Rows: 50148 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (1): location
## dbl  (14): total_vaccinations, total_distributed, people_vaccinated, people_...
## date  (1): date
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
    uid <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/UID_ISO

    popul <- read_csv(uid)
```

```
## Rows: 4321 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): iso2, iso3, FIPS, Admin2, Province_State, Country_Region, Combined_Key
## dbl (5): UID, code3, Lat, Long_, Population
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
# Now we are going to join the population in global.

# DO THE SAME THINGS WITH US CASES:
us_cases <- us_cases %>% pivot_longer(cols = -(UID:Combined_Key), names_to = "date", values_to = "ca
us_cases <- us_cases %>% select(Admin2:cases)
us_cases <- us_cases %>% mutate(date = mdy(date))

# DO THE SAME WITH US DEATHS (INCLUDING POPULATION)

us_deaths <- us_deaths %>% pivot_longer(cols = -(UID:Population), names_to = "date", values_to = "de
us_deaths <- us_deaths %>% select(Admin2:deaths)
us_deaths <- us_deaths %>% mutate(date = mdy(date))
us_deaths <- us_deaths %>% select(-c(Lat,Long_))

# Combine the two
US_by_state <- us_cases %>% full_join(us_deaths)
```

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

Transform

Lets add a few more columns and start to look at some things of interest.

```
# Started another one to now go through and visualize and analize some of the data above

# Group the data by the following:
US_by_state <- US_by_state %>% group_by(Province_State, Country_Region, date)

# add the following:
US_by_state <- US_by_state %>% summarise(cases = sum(cases), deaths = sum(deaths), Population = sum
```

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

```
# Add another column
US_by_state <- US_by_state %>% mutate(deaths_per_mill = deaths *1000000 / Population)
US_by_state <- US_by_state %>% mutate(cases_per_mill = cases *1000000 / Population)
# Ungroup
US_by_state <- US_by_state %>% ungroup()

US_state_totals <- US_by_state %>% group_by(Province_State) %>% summarise(deaths=max(deaths), cases=

# Lets add the Vaccine Data to the US_state_totals

VaccineTotals <- Vaccine_data %>% select(date,location, people_fully_vaccinated)

# Lets get the totals for each state

VaccineTotals <- VaccineTotals %>% group_by(location) %>% summarise(People_fully_vaccinated=max(peop
# rename location to Province_State
VaccineTotals <- VaccineTotals %>% rename(Province_State = location)

# Lets combine this with the US_State_totals

US_state_totals <- US_state_totals  %>% full_join(VaccineTotals)
```

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

```
# Lets add another column for percent

US_state_totals <- US_state_totals  %>% mutate(VaccinePercent = People_fully_vaccinated/population )

# Lets Cleanup Vaccine data as well.

Vaccine_data <- Vaccine_data %>% select(date, location, people_fully_vaccinated )
Vaccine_data <- Vaccine_data %>% rename(Province_State = `location`)
```

Interesting Charts

Now lets take a look at some interesting states.

```
# Here is a list of states with the most deaths.

US_state_totals %>% slice_max(deaths_per_thou, n=10)
```

```
## # A tibble: 10 x 8
##    Province_State deaths   cases population cases_per_~1 death~2 Peopl~3 Vacci~4
##    <chr>           <dbl>   <dbl>      <dbl>        <dbl>   <dbl>   <dbl>   <dbl>
##  1 Arizona         33076 2440294    7278717         335.    4.54 4809730   0.661
##  2 Oklahoma        17940 1287378    3956971         325.    4.53 2393713   0.605
##  3 Mississippi     13351  989282    2976149         332.    4.49 1597932   0.537
##  4 West Virginia    7960  642760    1792147         359.    4.44 1070043   0.597
##  5 New Mexico       9054  670301    2096829         320.    4.32 1577485   0.752
##  6 Arkansas        13001 1005930    3017804         333.    4.31 1716584   0.569
##  7 Alabama         21001 1642062    4903185         335.    4.28 2607186   0.532
##  8 Tennessee       29225 2510002    6829174         368.    4.28 3843967   0.563
##  9 Michigan        42096 3057222    9986857         306.    4.22 6225457   0.623
## 10 New Jersey      35995 3046838    8882190         343.    4.05 7031019   0.792
## # ... with abbreviated variable names 1: cases_per_thou, 2: deaths_per_thou,
## #   3: People_fully_vaccinated, 4: VaccinePercent
```

Let us take a look at some states with the lowest death rates:

```
US_state_totals %>% slice_min(deaths_per_thou, n=10)
```

```
## # A tibble: 10 x 8
##    Province_State          deaths   cases popul~1 cases~2 death~3 Peopl~4 Vacci~5
##    <chr>                    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 American Samoa              34 8.32e3   55641    150.   0.611   42467   0.763
##  2 Northern Mariana Islan~     41 1.37e4   55144    248.   0.744   43953   0.797
##  3 Virgin Islands             130 2.48e4  107268    231.   1.21    58951   0.550
##  4 Hawaii                    1834 3.80e5 1415872    268.   1.30  1154399   0.815
##  5 Vermont                    910 1.51e5  623989    243.   1.46   535103   0.858
##  6 Puerto Rico               5810 1.10e6 3754939    293.   1.55  2685604   0.715
##  7 Utah                      5287 1.09e6 3205958    340.   1.65  2141235   0.668
##  8 Alaska                    1486 3.07e5  740995    414.   2.01   476567   0.643
##  9 District of Columbia      1430 1.78e5  705749    252.   2.03   638828   0.905
## 10 Washington               15683 1.93e6 7614893    253.   2.06  5795079   0.761
## # ... with abbreviated variable names 1: population, 2: cases_per_thou,
## #   3: deaths_per_thou, 4: People_fully_vaccinated, 5: VaccinePercent
```

Lets take a look at the Vaccination rates

```
US_state_totals %>% slice_max(VaccinePercent, n=10)
```

```
## # A tibble: 10 x 8
##    Province_State          deaths  cases popul~1 cases~2 death~3 Peopl~4 Vacci~5
##    <chr>                    <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 District of Columbia      1430 1.78e5  705749    252.   2.03   638828   0.905
##  2 Guam                       420 6.10e4  164229    371.   2.56   144432   0.879
##  3 Rhode Island              3865 4.60e5 1059361    434.   3.65   929813   0.878
##  4 Vermont                    910 1.51e5  623989    243.   1.46   535103   0.858
##  5 Massachusetts            24249 2.22e6 6892503    322.   3.52  5810671   0.843
##  6 Maine                     2921 3.17e5 1344212    236.   2.17  1120822   0.834
##  7 Connecticut              12196 9.76e5 3565287    274.   3.42  2961528   0.831
##  8 Hawaii                    1834 3.80e5 1415872    268.   1.30  1154399   0.815
##  9 Maryland                 16513 1.36e6 6045680    226.   2.73  4821895   0.798
## 10 Northern Mariana Islan~     41 1.37e4   55144    248.   0.744   43953   0.797
## # ... with abbreviated variable names 1: population, 2: cases_per_thou,
## #   3: deaths_per_thou, 4: People_fully_vaccinated, 5: VaccinePercent
```

Lets Take a look at the states with the minimum Vaccine Rates

```
US_state_totals %>% slice_min(VaccinePercent, n=10)
```

```
## # A tibble: 10 x 8
##    Province_State deaths   cases population cases_per_~1 death~2 Peopl~3 Vacci~4
##    <chr>           <dbl>   <dbl>      <dbl>        <dbl>   <dbl>   <dbl>   <dbl>
##  1 Wyoming          2002  185159     578759         320.    3.46  306951   0.530
##  2 Alabama         21001 1642062    4903185         335.    4.28 2607186   0.532
##  3 Mississippi     13351  989282    2976149         332.    4.49 1597932   0.537
##  4 Missouri        22825 1773865    6626371         268.    3.44 3625649   0.547
##  5 Virgin Islands    130   24792     107268         231.    1.21   58951   0.550
##  6 Louisiana       18738 1570671    4648794         338.    4.03 2558088   0.550
##  7 Tennessee       29225 2510002    6829174         368.    4.28 3843967   0.563
##  8 Idaho            5399  519850    1787065         291.    3.02 1009286   0.565
##  9 Arkansas        13001 1005930    3017804         333.    4.31 1716584   0.569
## 10 Georgia         42427 3065390   10617423         289.    4.00 6086702   0.573
## # ... with abbreviated variable names 1: cases_per_thou, 2: deaths_per_thou,
## #   3: People_fully_vaccinated, 4: VaccinePercent
```

ANALYSIS

From the above date lets take a look at the following states:

I picked Massachusetts as it was a state with a pretty large population.. compared to the rest of the states.

Massachusetts 24249 2220962 6892503 322.2287 3.5181704 5810671

I thought it was important to pick something that was quite a bit further away from the Northeast.

Hawaii 1834 380098 1415872 268.4551 1.2953148 1154399

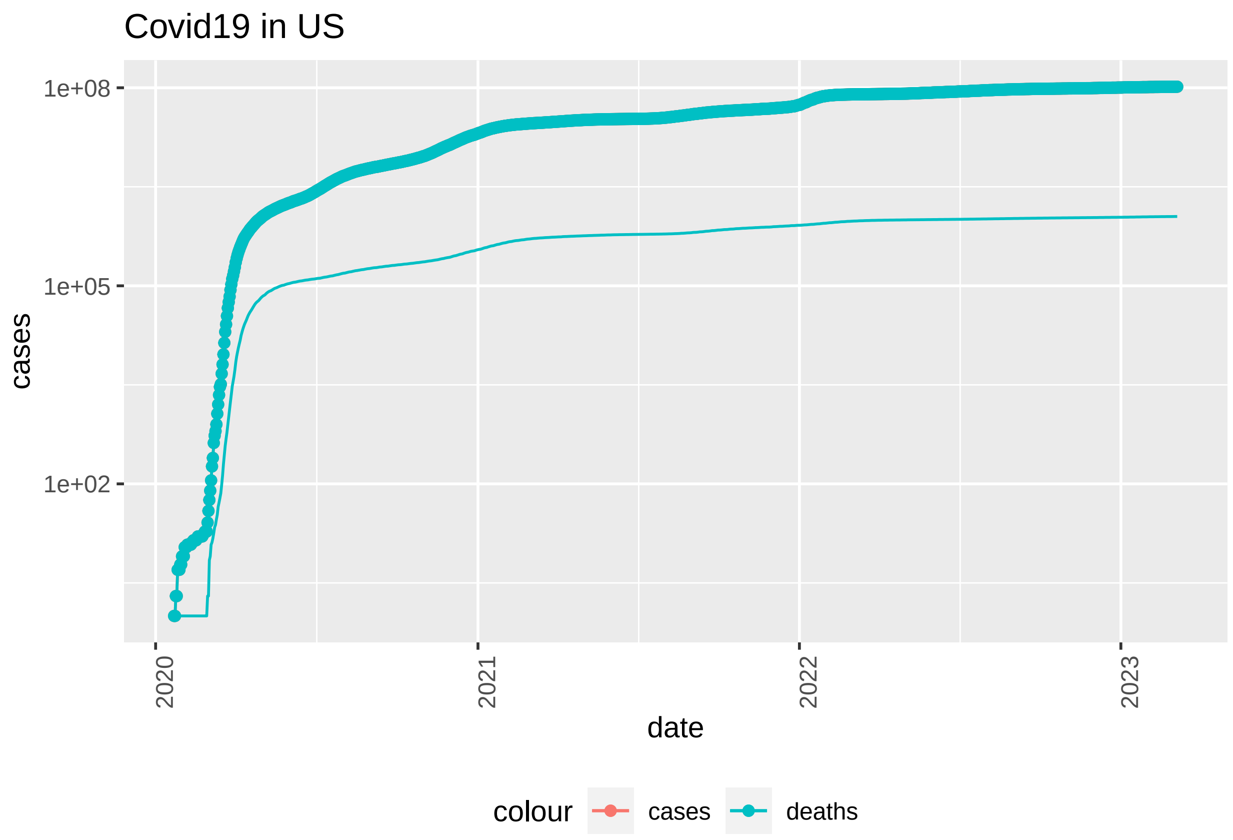
I picked these as these were the two states with the lowest vaccination rates. They also had population over one Million.

Alabama 21001 1642062 4903185 334.8970 4.283134 2607186 Missouri 22825 1773682 6626371 267.6702

```
# Lets look at the us totals as well.  This was done in class

    us_totals <- US_by_state %>% group_by(Country_Region, date) %>% summarize(cases=sum(cases), deaths=s
```

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

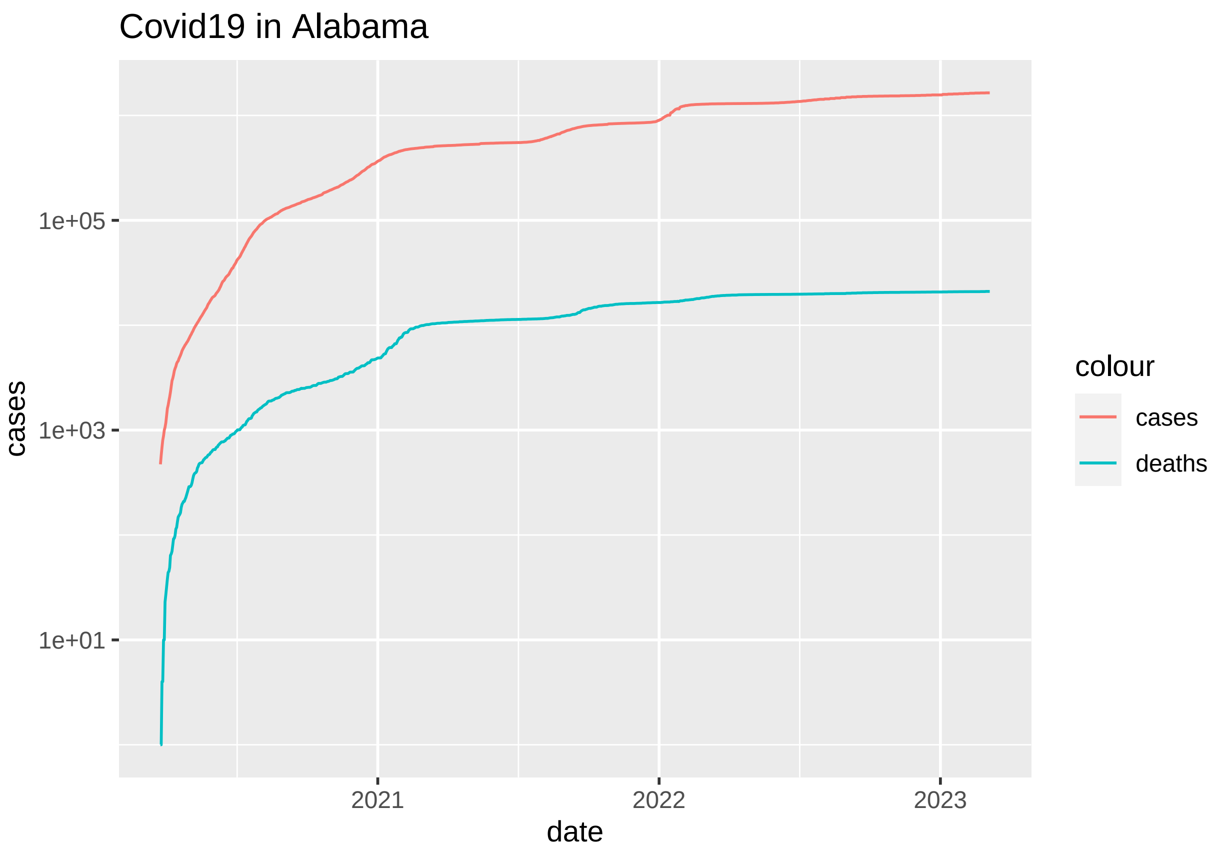
```
    # Ungroup them.
    us_totals <- us_totals %>% ungroup()

    # SHOW THE GRAPH OF DEATHS VS. CASES

    us_totals %>% filter(cases> 0) %>% ggplot(aes(x=date, y=cases)) +geom_line(aes(color = "cases")) +ge
```



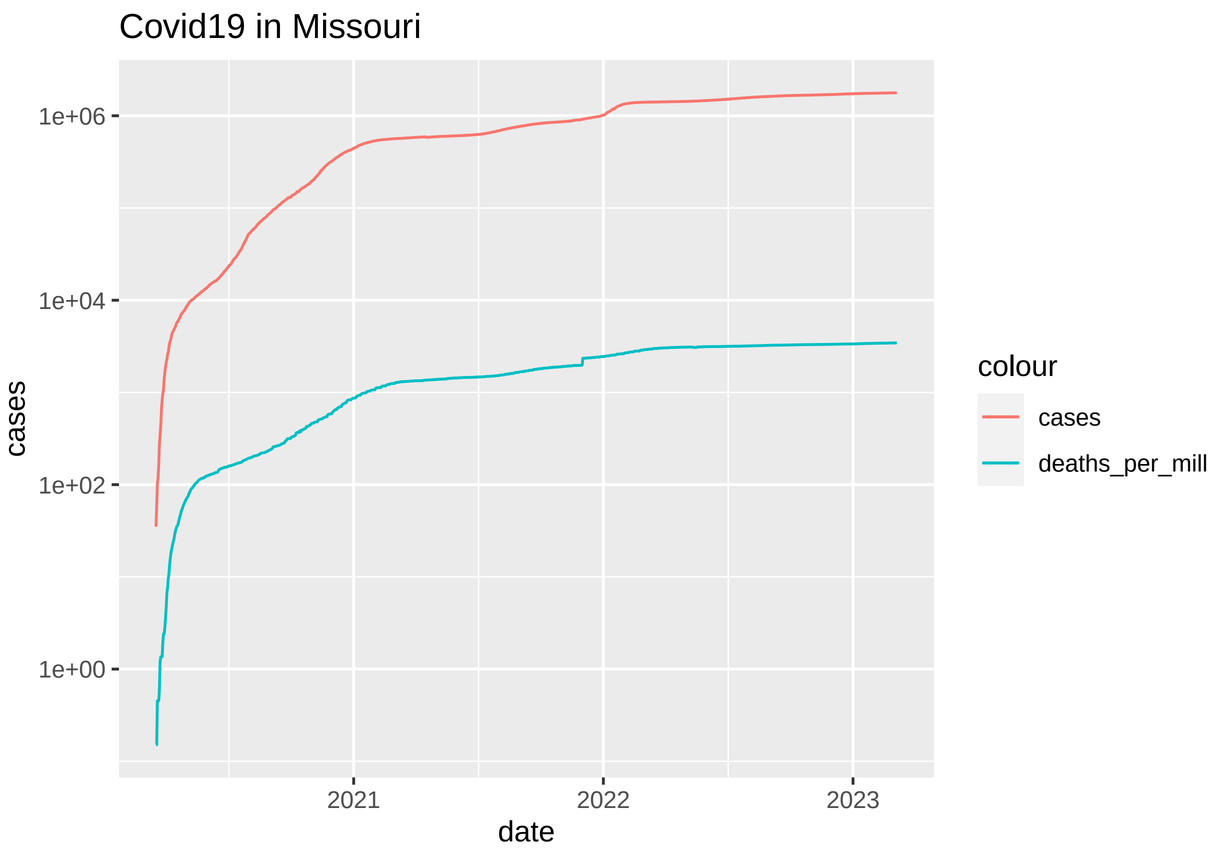
Visualize

It looks like overall the cases and death rates are flat. This was what we would expect from looking at the US Data above.

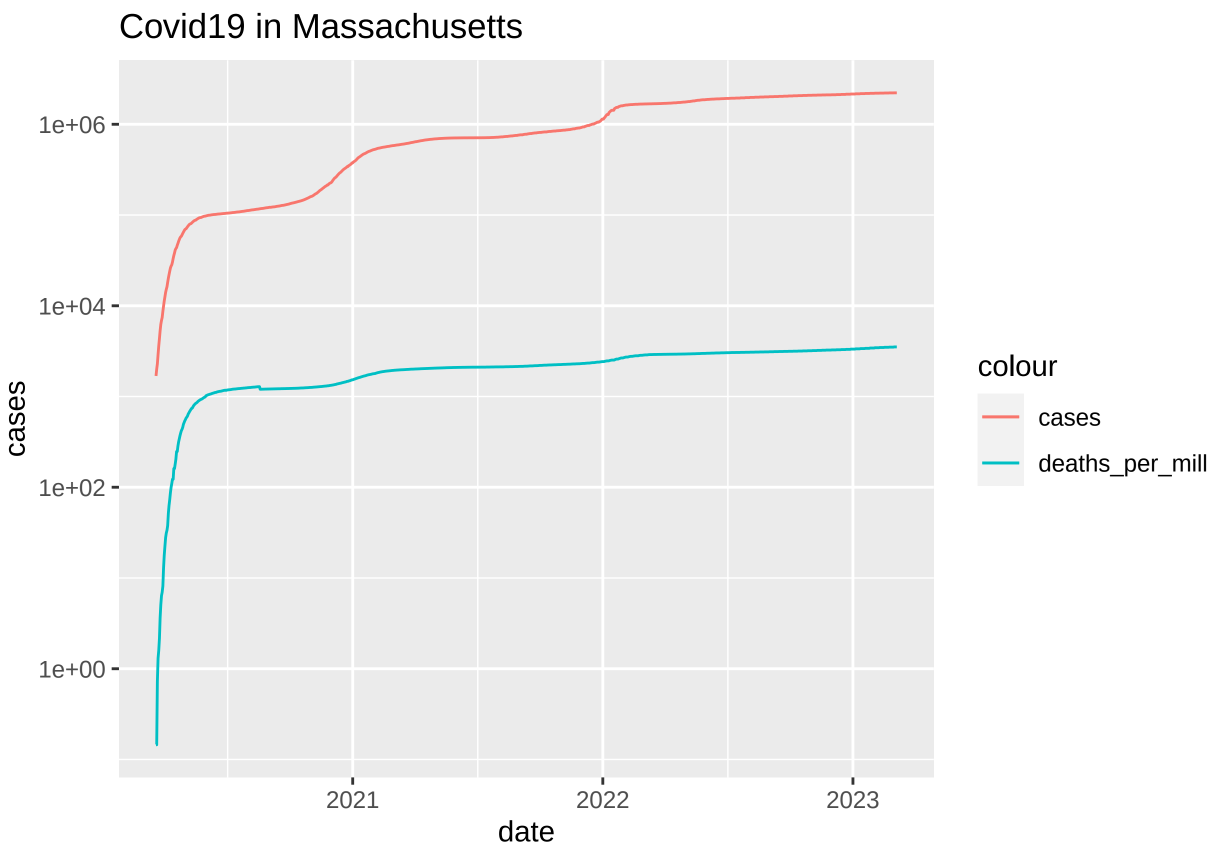
```
# Lets take a look at the 4 states we picked above.  Two from the lowest catagory and two from the high

    US_by_state %>% filter(cases > 0, deaths > 0, Province_State == "Alabama") %>% ggplot(aes(x=date, y=
```



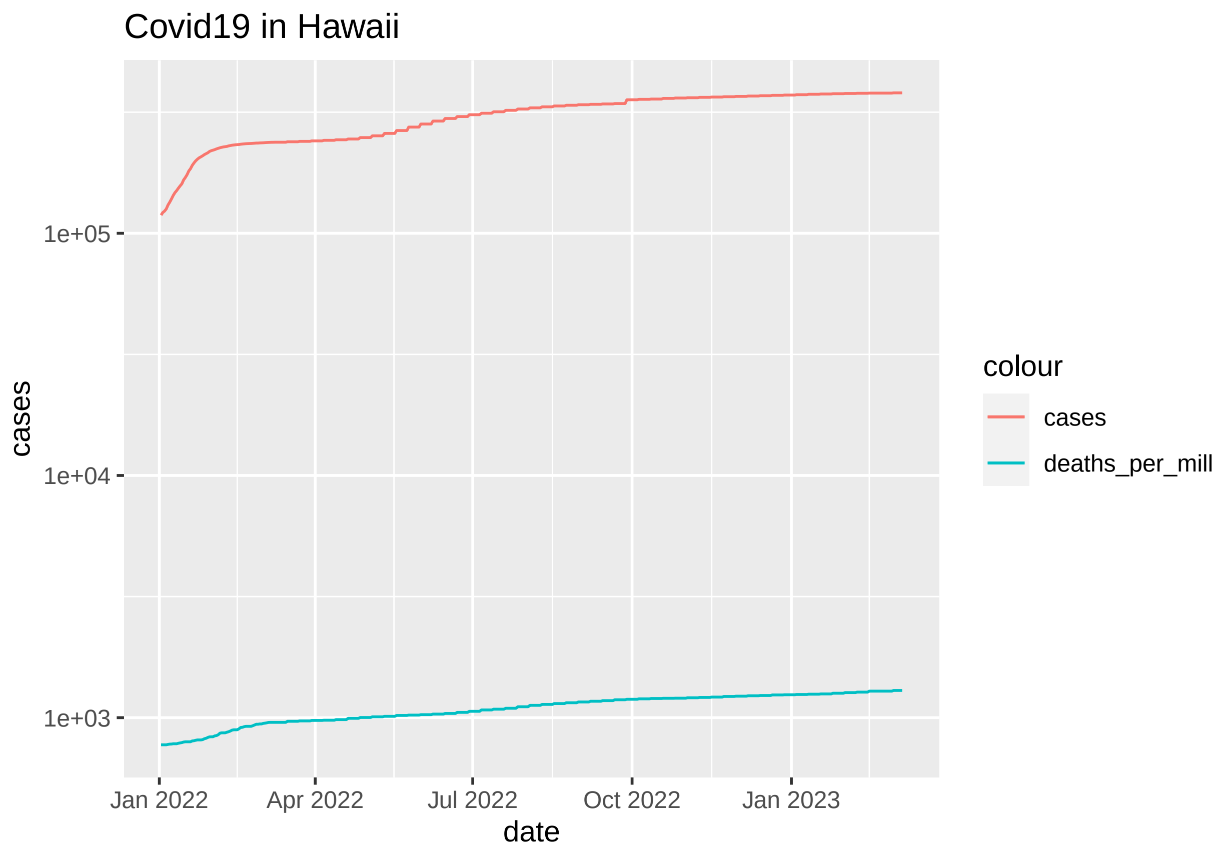
```
US_by_state %>% filter(cases > 0, deaths > 0, Province_State == "Missouri") %>% ggplot(aes(x=da
```



```
US_by_state %>% filter(cases > 0, deaths > 0, Province_State == "Massachusetts") %>% ggplot(aes
```



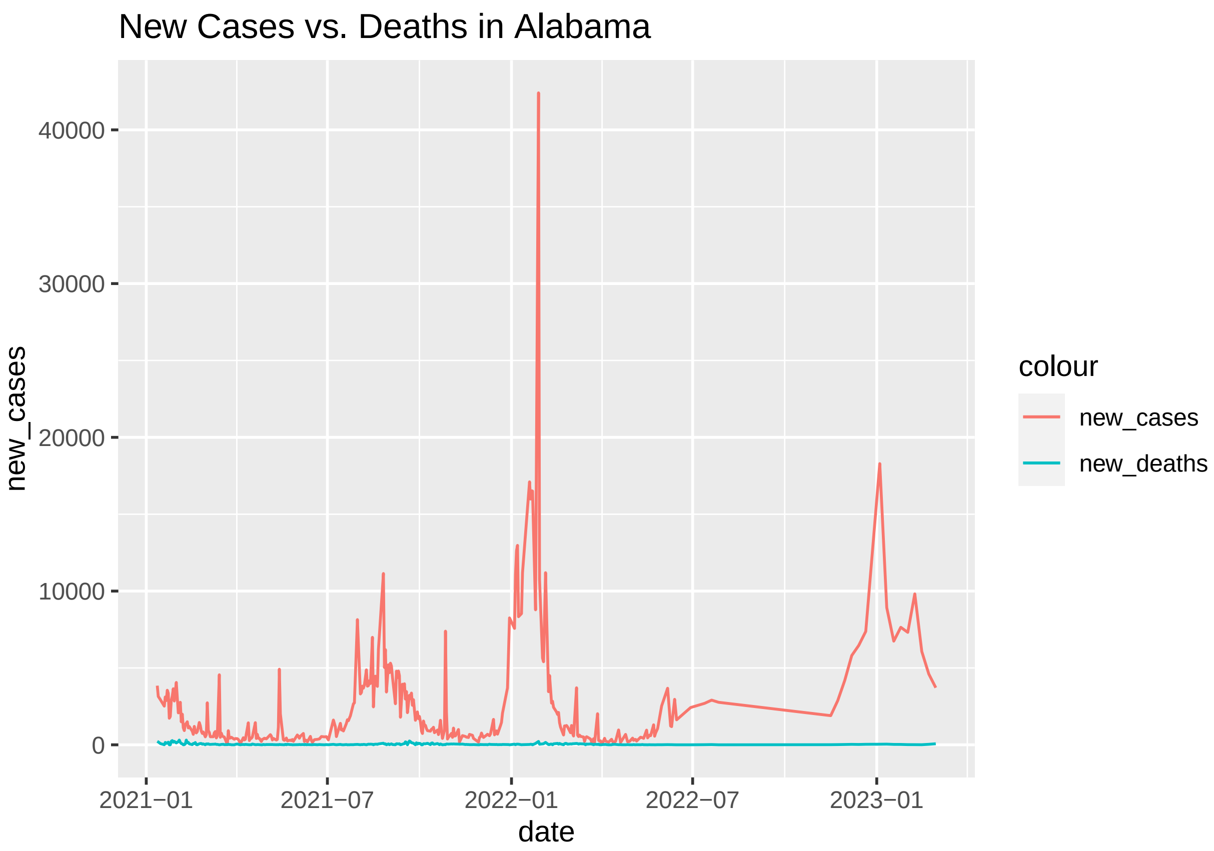
```
US_by_state %>% filter(cases > 0, deaths > 0, Province_State == "Hawaii", date > "2022-01-01") %>% g
```



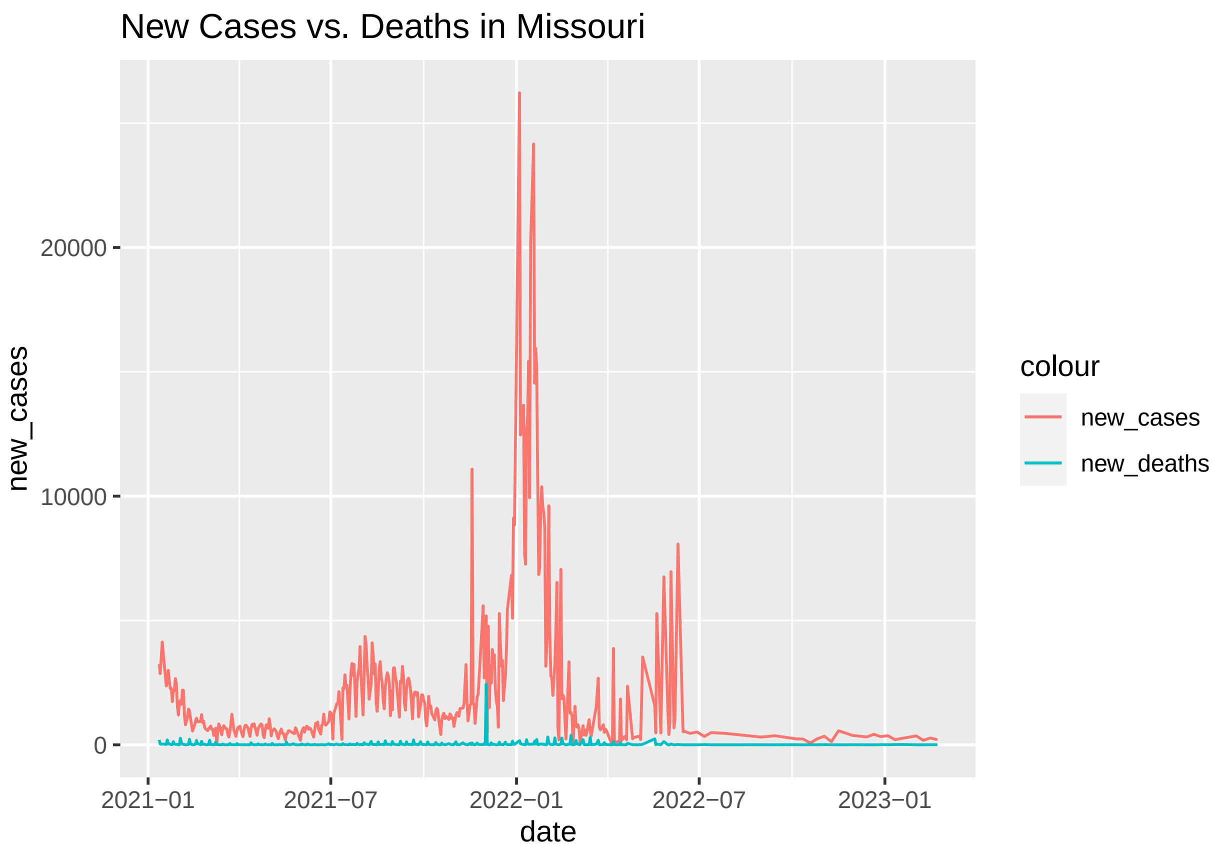
# VISUALIZE

Now lets look at the Vaccine data and new Deaths and New cases.

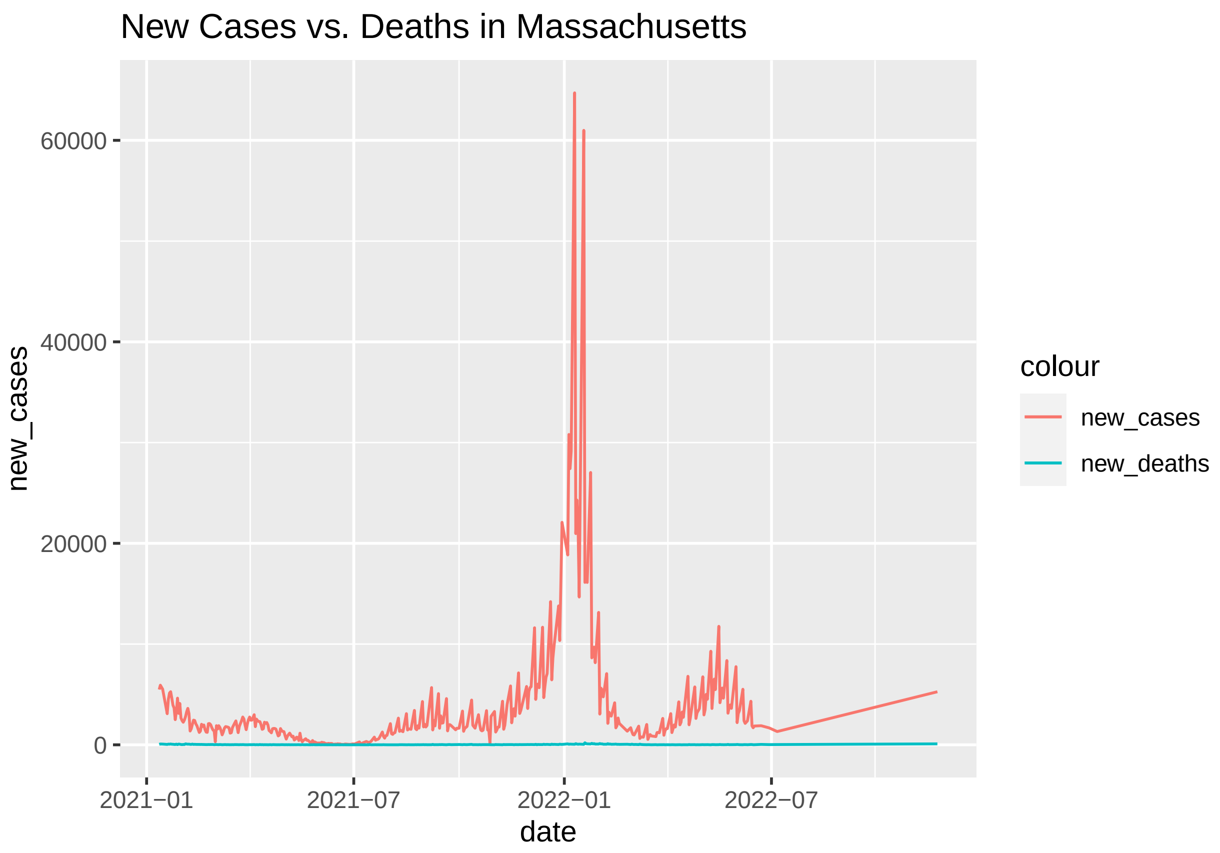
```
# Since all we care about is people who have been fully_vaccinated.

    #Vaccine_data <- Vaccine_data %>% select(date, location, people_fully_vaccinated )
    #Vaccine_data <- Vaccine_data %>% rename(Province_State = `location`)

# Now lets go ahead and combine this with the states total.

    US_by_state_vaccine <- US_by_state %>% full_join(Vaccine_data,by=c("Province_State","date"))

    # Now lets just look at the new cases
    US_by_state_vaccine <- US_by_state_vaccine %>% mutate(new_cases = cases - lag(cases), new_deaths=dea

    US_by_state_vaccine <- US_by_state_vaccine %>% mutate(Percent_deaths = new_deaths/new_cases *100)

    US_by_state_vaccine <- US_by_state_vaccine %>% mutate(vaccine_per_hundred = 1000 * people_fully_vacc

    US_by_state_vaccine <- US_by_state_vaccine %>% mutate(cases_per_hundred = 1000 * cases/Population)

    US_by_state_vaccine %>% filter(new_deaths > 0, people_fully_vaccinated > 0, new_cases > 0, Province_
```



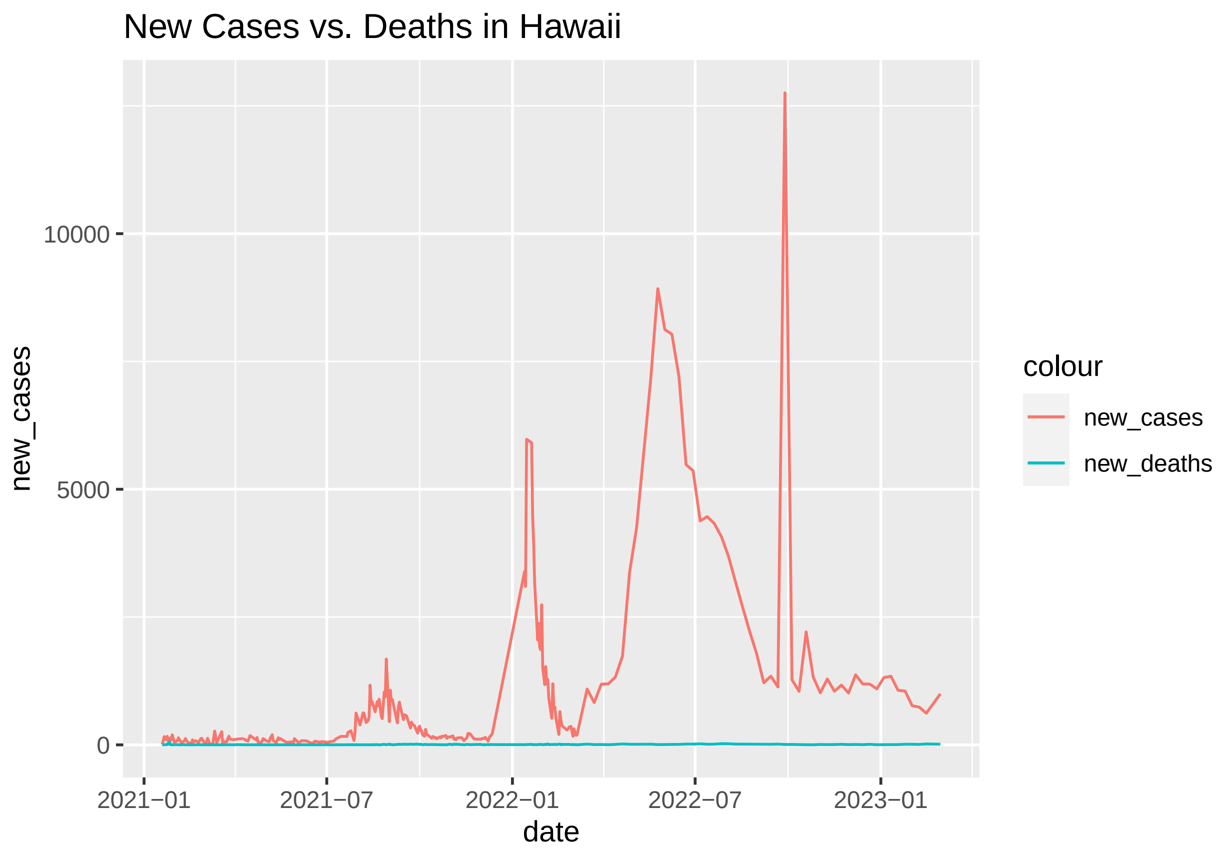
```
US_by_state_vaccine %>% filter(new_deaths > 0, people_fully_vaccinated > 0, new_cases > 0, Province_
```



```
US_by_state_vaccine %>% filter(new_deaths > 0, people_fully_vaccinated > 0, new_cases > 0, Province_
```



```
US_by_state_vaccine %>% filter(new_deaths > 0, people_fully_vaccinated > 0, new_cases > 0, Province_
```

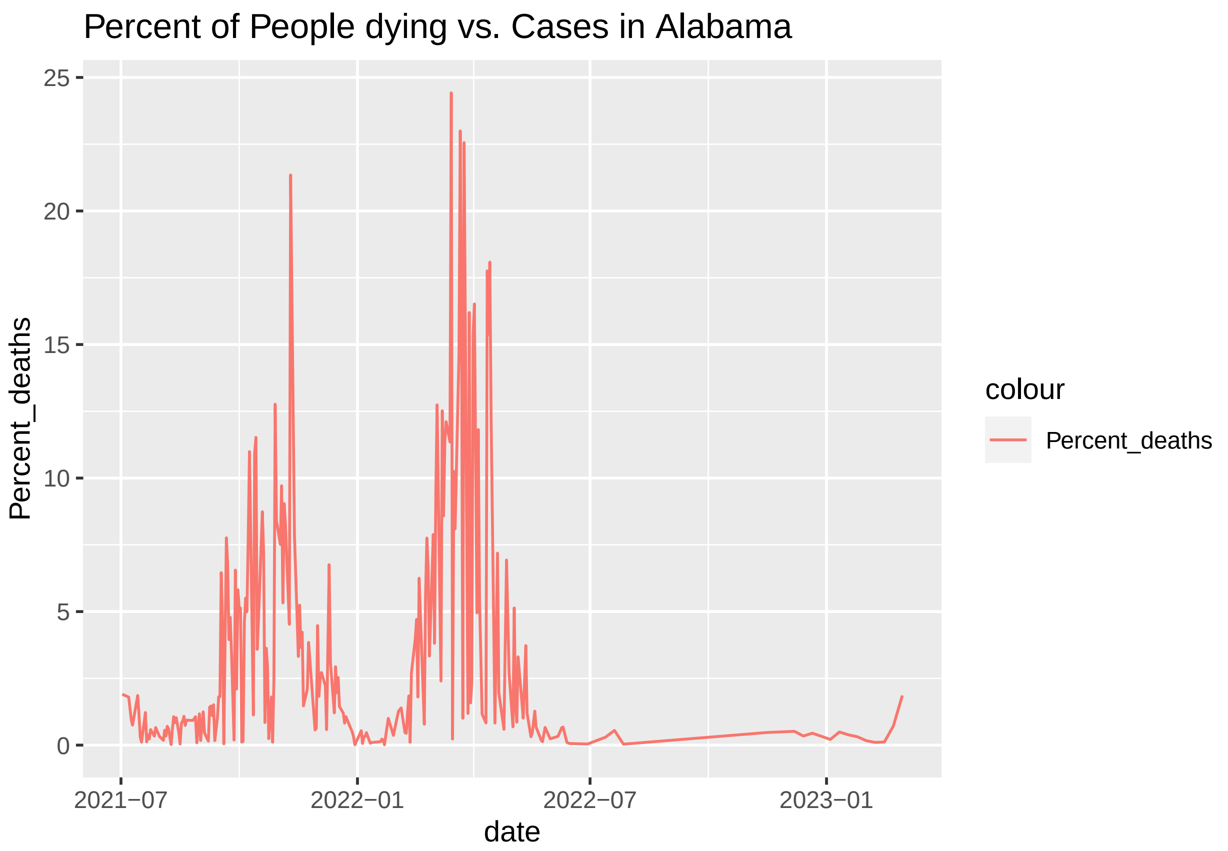


MORE VISUALS

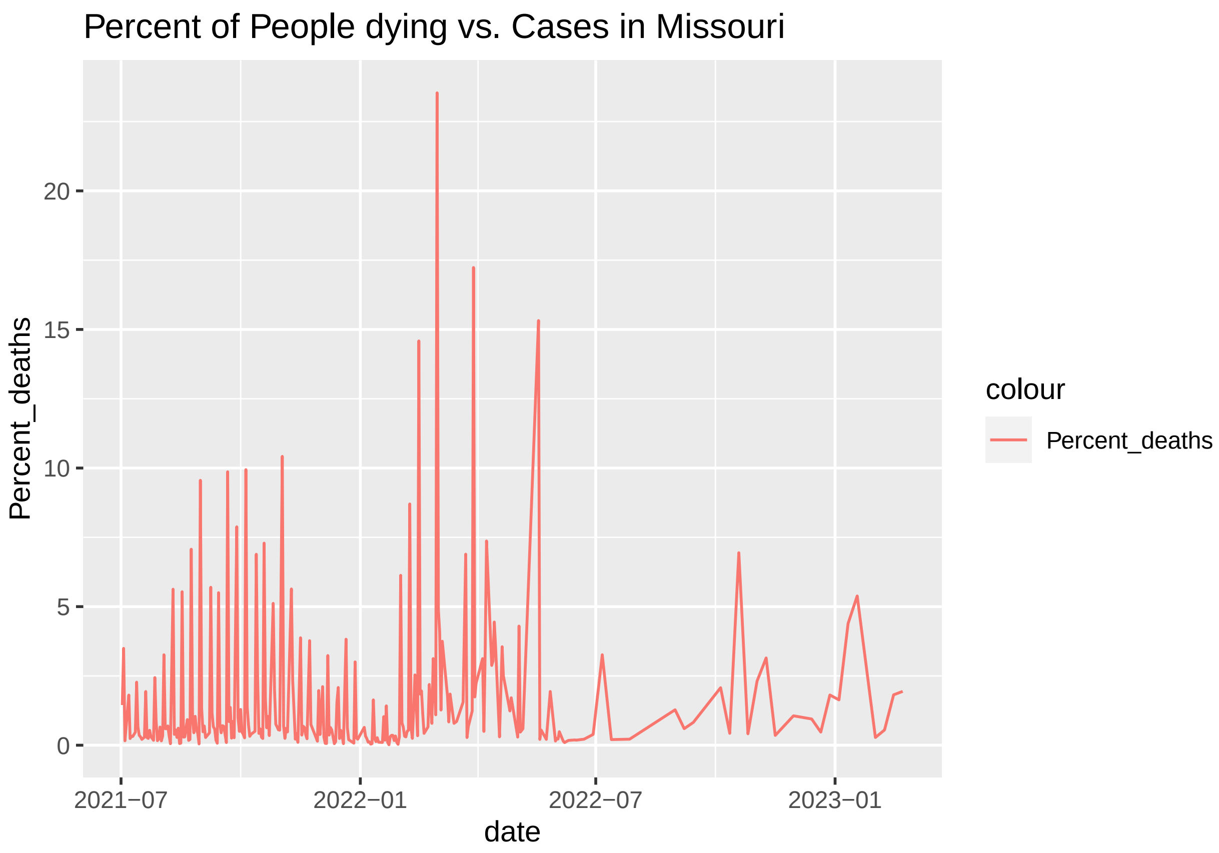
Lets see if the percent of deaths vs. new cases in each of the above states.

So our hypothesis was that states with the greater number of vaccinated residents should have lesser number of residents dying. So we are going to take a look at each of the states and see if that's true.

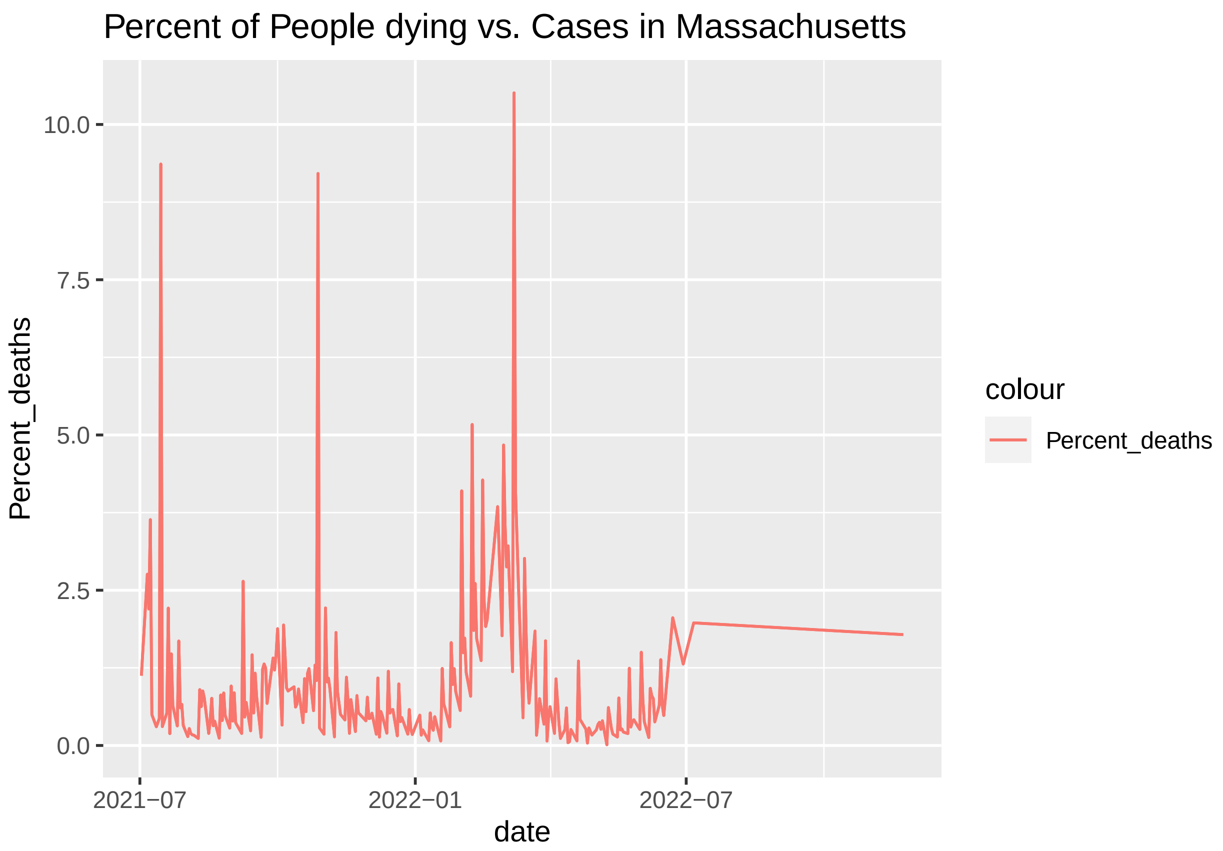
```
# Some things to note.  I took a look at all numbers after 2021-07-01.  Mostly because we started vacci

# Lets take a look at Alabama First.
US_by_state_vaccine %>% filter(new_deaths > 0, people_fully_vaccinated > 0, new_cases > 0, Province_Sta
```



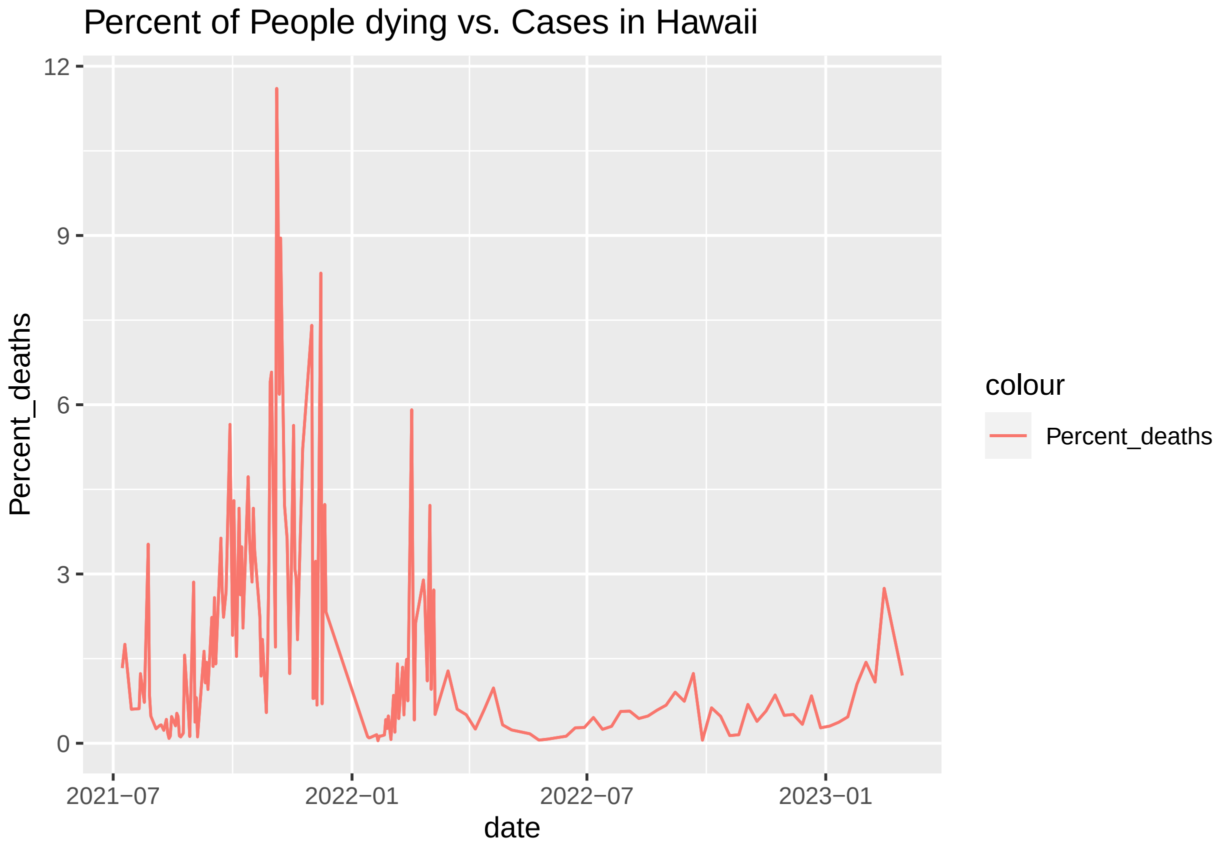
```
# Lets take a look at Missouri.

US_by_state_vaccine %>% filter(new_deaths > 0, people_fully_vaccinated > 0, new_cases > 0, Province_Sta
```



```
US_by_state_vaccine %>% filter(new_deaths > 0, people_fully_vaccinated > 0, new_cases > 0, Province_Sta
```



```
US_by_state_vaccine %>% filter(new_deaths > 0, people_fully_vaccinated > 0, new_cases > 0, Province_Sta
```



# ANALYSIS

As we can clearly see the death rate in states with high vaccination is quite a bit. Lets just take a look at Alabama and Massachusetts and you can see the chances of you dying from Covid-19 goes down dramatically.

BIAS

Personal Bias I think I picked Massachusetts as I have ties to that state.

Now sure how each state are collecting there data.

Demographics can also play a big role.