JHU COVID-19 Data Analysis Report

2023-08-14

Introduction

In this report we describe four datasets on COVID-19 obtained from the Johns Hopkins University CSSE COVID-19 github site. The datasets contain information on confirmed cases of COVID-19 and COVID-19 related deaths from either the USA only, or globally. The goal of this analysis is to look at the difference in case rates and death rates from one state to another with a focus on the states of New York, Alaska, and Arizona. In addition to the comparisons across states, we wanted ask the question, can we model death rates from case rates both in the US (across states) and globally (across countries).

Data Loading

To start, we load in the data from the four main files of time series data on COVID-19 from Johns Hopkins University. This data is obtained from the JHU CSSE COVID-19 Dataset hosted on github at the following url: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data

Read in the data and take a look at the structure.

```
## Rows: 289 Columns: 1147
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1145): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20,...
##
## 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.
## Rows: 289 Columns: 1147
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1145): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20,...
##
## 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.
## Rows: 3342 Columns: 1154
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1148): 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.
## Rows: 3342 Columns: 1155
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1149): 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.
## 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.
```

Table 1: global_cases data

Province/State	Country/Region	Lat	Long	1/22/20
NA	Afghanistan	33.93911	67.70995	0
NA	Albania	41.15330	20.16830	0
NA	Algeria	28.03390	1.65960	0
NA	Andorra	42.50630	1.52180	0
NA	Angola	-11.20270	17.87390	0
NA	Antarctica	-71.94990	23.34700	0

Table 2: global_deaths data

Province/State	Country/Region	Lat	Long	1/22/20
NA	Afghanistan	33.93911	67.70995	0
NA	Albania	41.15330	20.16830	0
NA	Algeria	28.03390	1.65960	0
NA	Andorra	42.50630	1.52180	0
NA	Angola	-11.20270	17.87390	0
NA	Antarctica	-71.94990	23.34700	0

Table 3: us_cases data

UID	iso2	iso3	code3	FIPS	Admin2	Province_State	Country_Region	Lat	Long_	Combined_Key	1/22/20	1/23/20
84001001	US	USA	840	1001	Autauga	Alabama	US	32.53953	-86.64408	Autauga, Alabama, US	0	0
84001003	US	USA	840	1003	Baldwin	Alabama	US	30.72775	-87.72207	Baldwin, Alabama, US	0	0
84001005	US	USA	840	1005	Barbour	Alabama	US	31.86826	-85.38713	Barbour, Alabama, US	0	0
84001007	US	USA	840	1007	Bibb	Alabama	US	32.99642	-87.12511	Bibb, Alabama, US	0	0
84001009	US	USA	840	1009	Blount	Alabama	US	33.98211	-86.56791	Blount, Alabama, US	0	0

UID	iso2	iso3	code3	FIPS	Admin2	Province_State	Country_Region	Lat	Long_	Combined_Key	1/22/20	1/23/20
84001011	US	USA	840	1011	Bullock	Alabama	US	32.10031	-85.71266	Bullock, Alabama, US	0	0

Table 4: us_deaths data

UID	iso2	iso3	code3	FIPS	Admin2	Province_State	Country_Region	Lat	Long_	Combined_Key	Population	1/22/20
84001001	US	USA	840	1001	Autauga	Alabama	US	32.53953	-86.64408	Autauga, Alabama, US	55869	0
84001003	US	USA	840	1003	Baldwin	Alabama	US	30.72775	-87.72207	Baldwin, Alabama, US	223234	0
84001005	US	USA	840	1005	Barbour	Alabama	US	31.86826	-85.38713	Barbour, Alabama, US	24686	0
84001007	US	USA	840	1007	Bibb	Alabama	US	32.99642	-87.12511	Bibb, Alabama, US	22394	0
84001009	US	USA	840	1009	Blount	Alabama	US	33.98211	-86.56791	Blount, Alabama, US	57826	0
84001011	US	USA	840	1011	Bullock	Alabama	US	32.10031	-85.71266	Bullock, Alabama, US	10101	0

Data Cleaning

After reading in the four datasets, the datasets need to be tidied up to put each variable in their own column in long format. This is obvious when looking at the tables above. Additionally, there is no need for the latitude and longitude for the purpose of the planned analysis so this can be dropped from the datasets.

Global Dataset

```
## Joining with `by = join_by(`Province/State`, `Country/Region`, date)`
```

The tidied "global" dataset looks much nicer now.

province_state	country_region	date	cases	deaths
NA	Afghanistan	2020-01-22	0	0
NA	Afghanistan	2020-01-23	0	0
NA	Afghanistan	2020-01-24	0	0
NA	Afghanistan	2020-01-25	0	0
NA	Afghanistan	2020-01-26	0	0
NA	Afghanistan	2020-01-27	0	0

```
##  province_state     country_region          date                 cases
##  Length:330327      Length:330327      Min.   :2020-01-22   Min.   :        0
##  Class :character   Class :character   1st Qu.:2020-11-02   1st Qu.:      680
##  Mode  :character   Mode  :character   Median :2021-08-15   Median :    14429
```

```
##                                         Mean   :2021-08-15   Mean   :   959384
##                                         3rd Qu.:2022-05-28   3rd Qu.:   228517
##                                         Max.   :2023-03-09   Max.   :103802702
##      deaths
##  Min.   :      0
##  1st Qu.:      3
##  Median :    150
##  Mean   :  13380
##  3rd Qu.:   3032
##  Max.   :1123836
```

```
## tibble [330,327 x 5] (S3: tbl_df/tbl/data.frame)
##  $ province_state: chr [1:330327] NA NA NA NA ...
##  $ country_region: chr [1:330327] "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ date          : Date[1:330327], format: "2020-01-22" "2020-01-23" ...
##  $ cases         : num [1:330327] 0 0 0 0 0 0 0 0 0 0 ...
##  $ deaths        : num [1:330327] 0 0 0 0 0 0 0 0 0 0 ...
```

It looks like there are probably a lot of rows where cases are equal to 0 so let's remove those rows.

```
##  province_state     country_region          date                 cases
##  Length:306827      Length:306827      Min.   :2020-01-22   Min.   :        1
##  Class :character   Class :character   1st Qu.:2020-12-12   1st Qu.:     1316
##  Mode  :character   Mode  :character   Median :2021-09-16   Median :    20365
##                                        Mean   :2021-09-11   Mean   :  1032863
##                                        3rd Qu.:2022-06-15   3rd Qu.:   271281
##                                        Max.   :2023-03-09   Max.   :103802702
##      deaths
##  Min.   :      0
##  1st Qu.:      7
##  Median :    214
##  Mean   :  14405
##  3rd Qu.:   3665
##  Max.   :1123836
```

```
## tibble [306,827 x 5] (S3: tbl_df/tbl/data.frame)
##  $ province_state: chr [1:306827] NA NA NA NA ...
##  $ country_region: chr [1:306827] "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ date          : Date[1:306827], format: "2020-02-24" "2020-02-25" ...
##  $ cases         : num [1:306827] 5 5 5 5 5 5 5 5 5 5 ...
##  $ deaths        : num [1:306827] 0 0 0 0 0 0 0 0 0 0 ...
```

```
## # A tibble: 9 x 5
##   province_state country_region date           cases  deaths
##   <chr>          <chr>          <date>         <dbl>   <dbl>
## 1 <NA>           US             2023-03-01 103533872 1120897
## 2 <NA>           US             2023-03-02 103589757 1121658
## 3 <NA>           US             2023-03-03 103648690 1122165
## 4 <NA>           US             2023-03-04 103650837 1122172
## 5 <NA>           US             2023-03-05 103646975 1122134
## 6 <NA>           US             2023-03-06 103655539 1122181
## 7 <NA>           US             2023-03-07 103690910 1122516
## 8 <NA>           US             2023-03-08 103755771 1123246
## 9 <NA>           US             2023-03-09 103802702 1123836
```

Here we only keep rows where cases are greater than zero and double check to make sure that the maximum values do not appear to be a typo and there is continuity in the dataset. It seems OK at this point for both cases and deaths for the “global” dataset.

US Dataset

Next, lets clean the “US” dataset.

```
## Joining with ‘by = join_by(Admin2, Province_State, Country_Region,
## Combined_Key, date)‘
```

The tidied “US” dataset looks much nicer now, but it has some additional columns which are not present in the “global” dataset. We will need to add these columns to the global dataset.

Table 6: data_structure_us

county	province_state	country_region	combined_key	date	cases	population	deaths
Autauga	Alabama	US	Autauga, Alabama, US	2020-01-22	0	55869	0
Autauga	Alabama	US	Autauga, Alabama, US	2020-01-23	0	55869	0
Autauga	Alabama	US	Autauga, Alabama, US	2020-01-24	0	55869	0
Autauga	Alabama	US	Autauga, Alabama, US	2020-01-25	0	55869	0
Autauga	Alabama	US	Autauga, Alabama, US	2020-01-26	0	55869	0
Autauga	Alabama	US	Autauga, Alabama, US	2020-01-27	0	55869	0

Data Transformation

Here we add new columns to the “global” dataset to ensure that both the “US” and “global” datasets have the same structure.

Adding columns to “global”

Now that we have the same columns in both datasets we can start the analysis.

Prior to starting our exploratory analysis we will need to create two subsets of data with summary statistics by state and summary statistics for the US totals.

Subsetting - US by State

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

Let’s check the structure of our subset.

Table 7: data_structure_by_state

province_state	country_region	date	cases	deaths	deaths_per_mill	population
Alabama	US	2020-01-22	0	0	0	4903185
Alabama	US	2020-01-23	0	0	0	4903185
Alabama	US	2020-01-24	0	0	0	4903185
Alabama	US	2020-01-25	0	0	0	4903185
Alabama	US	2020-01-26	0	0	0	4903185

province_state	country_region	date	cases	deaths	deaths_per_mill	population
Alabama	US	2020-01-27	0	0	0	4903185

Quality check - US by State

Now that we have our data by state, lets do a sense check on the population values to be sure everything is ok. The population of Alaska was reported to be around 731,158 in 2020 (source US Census Bureau).

```
## # A tibble: 1 x 2
##   province_state population
##   <chr>               <dbl>
## 1 Alaska             740995
```

Here we calculate of population of 740,995 which is relatively close to the census data found online.

Subsetting - US totals

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

Let’s check the structure of our subset.

Table 8: data_structure_us_totals

country_region	date	cases	deaths	deaths_per_mill	population
US	2020-01-22	1	1	0.0030041	332875137
US	2020-01-23	1	1	0.0030041	332875137
US	2020-01-24	2	1	0.0030041	332875137
US	2020-01-25	2	1	0.0030041	332875137
US	2020-01-26	5	1	0.0030041	332875137
US	2020-01-27	5	1	0.0030041	332875137

Quality check - US totals

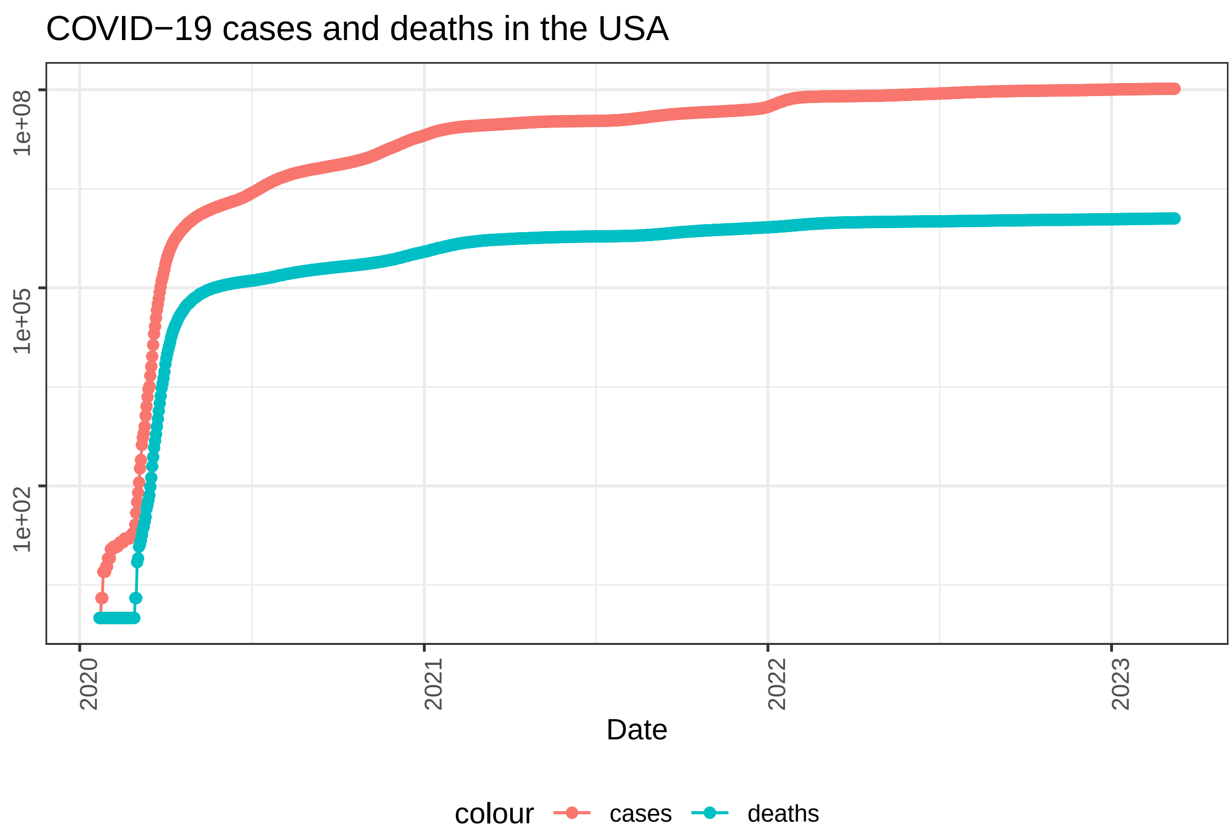
Now that we have the US totals, lets double check the population here as well. The projected total US population on January 1st, 2023 was 334,233,854 (source US Census Bureau). Our total of 332,875,137 is quite close, however, we see that the population for at the start of the pandemic is the same. As such, there may be some bias in the results knowing that the population data is static and does not change over time as it should in reality.

```
## # A tibble: 6 x 6
##   country_region date           cases  deaths deaths_per_mill population
##   <chr>          <date>         <dbl>   <dbl>           <dbl>      <dbl>
## 1 US             2023-03-04 103650837 1122172           3371.  332875137
## 2 US             2023-03-05 103646975 1122134           3371.  332875137
## 3 US             2023-03-06 103655539 1122181           3371.  332875137
## 4 US             2023-03-07 103690910 1122516           3372.  332875137
## 5 US             2023-03-08 103755771 1123246           3374.  332875137
## 6 US             2023-03-09 103802702 1123836           3376.  332875137
```

Exploratory Analysis

Visualization of total data in the USA

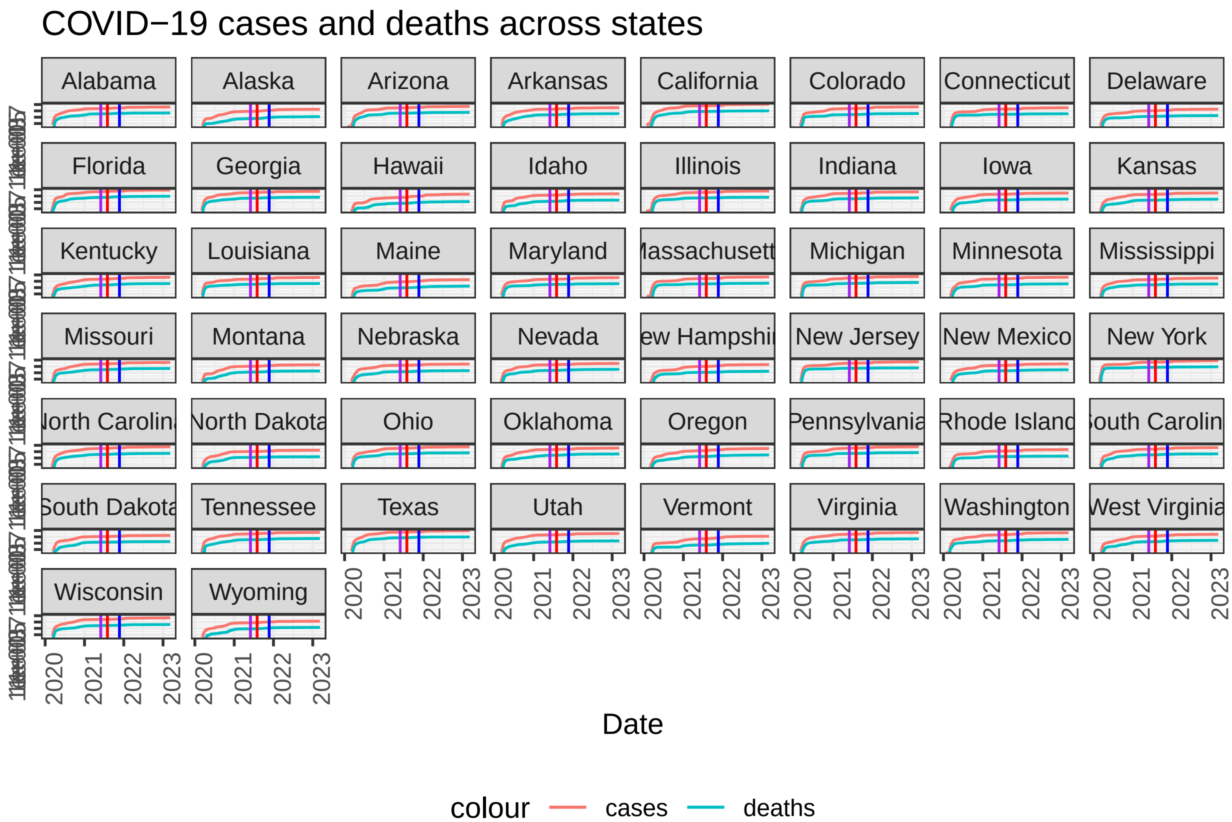
Now that we have checked the quality of both of our US data sets, let’s start by looking at our “us_totals” dataset and visualizing the data.



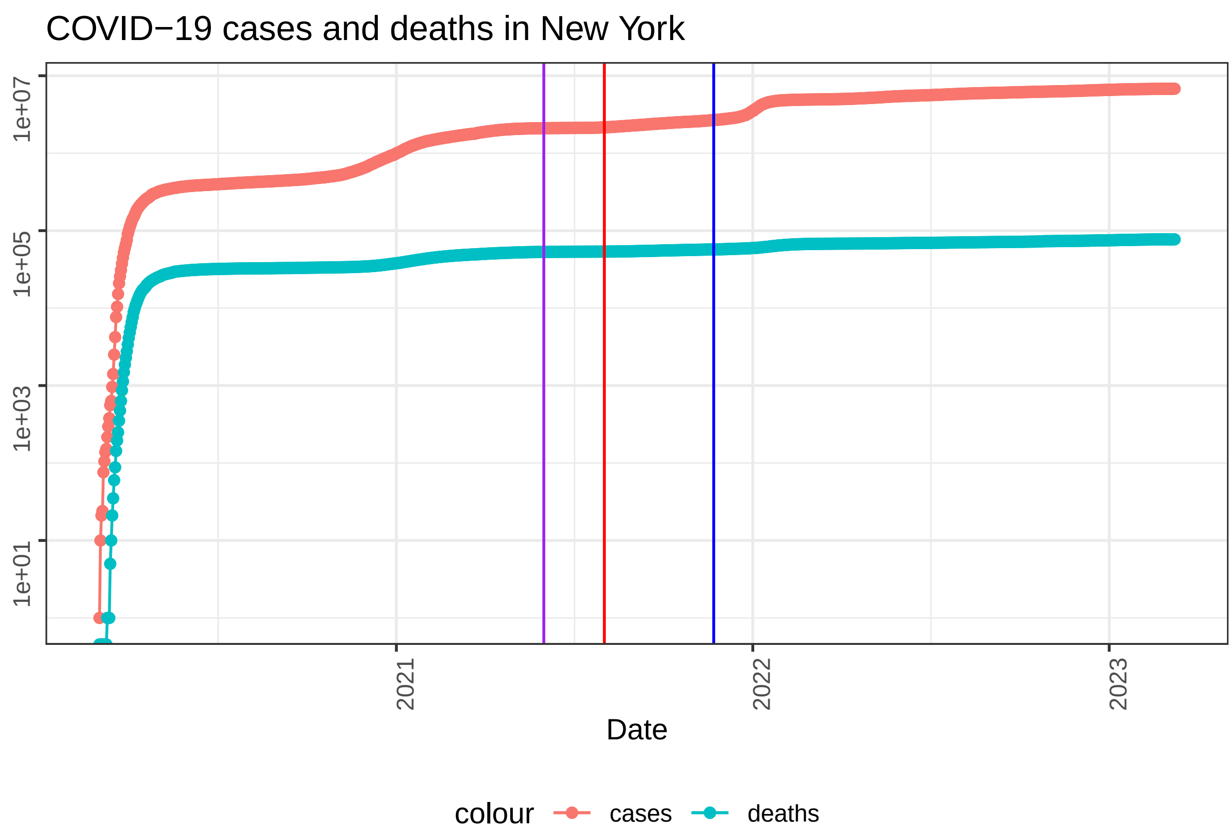
Shown above is a graph of the total number of cases and deaths (in red and blue respectively) by date since the start of the COVID-19 pandemic in the USA. The data are plotted on a logarithmic scale to facilitate reading of the graph.

Visualisation of data by state

Next let's look at the data across all states. For the purpose of this graph, we removed provinces/territories and only look at the 50 "official" states in the USA.



From this graph we can see that the death rates and case rates vary significantly from state to state. With some states experiencing high numbers of cases much later in the pandemic than others. Several important dates in the pandemic are marked on the graph. In June 2021, the delta sub variant of COVID-19 became the dominant strain and is marked in purple. The date of August 2nd 2021 is marked in red to highlight the date where the vaccination goal of 70% of the US population vaccinated with at least one dose of the COVID-19 vaccine was met. In November 2021, the omicron sub variant of COVID-19 became was identified and is marked in blue (https://www.cdc.gov/museum/timeline/covid19.html).



Shown above is a graph of the total number of cases and deaths (in red and blue respectively) by date since the start of the COVID-19 pandemic in the state of New York. The data are plotted on a logarithmic scale to facilitate reading of the graph. Several important dates in the pandemic are marked on the graph. In June 2021, the delta sub variant of COVID-19 became the dominant strain and is marked in purple. The date of August 2nd 2021 is marked in red to highlight the date where the vaccination goal of 70% of the US population vaccinated with at least one dose of the COVID-19 vaccine was met. In November 2021, the omicron sub variant of COVID-19 became was identified and is marked in blue (https://www.cdc.gov/museum/timeline/covid19.html).

At the time the preparation of this report, the latest information was from 2023-03-09 and the total number of deaths since the start of the pandemic in the USA has sadly reached 1.123836×10^6.

Analysis - New cases over time

After looking at the data, we are lead the the question of "Have the number of new cases leveled off?". To answer that question, we will need to go back to our data and transform it again by creating two new variables "new_cases" and "new_deaths".

Data Transformation - new cases and new deaths

Now that we have created the new variables, lets check what they look like for both the totals and by_state datasets.

Table 9: us_by_state with new variables

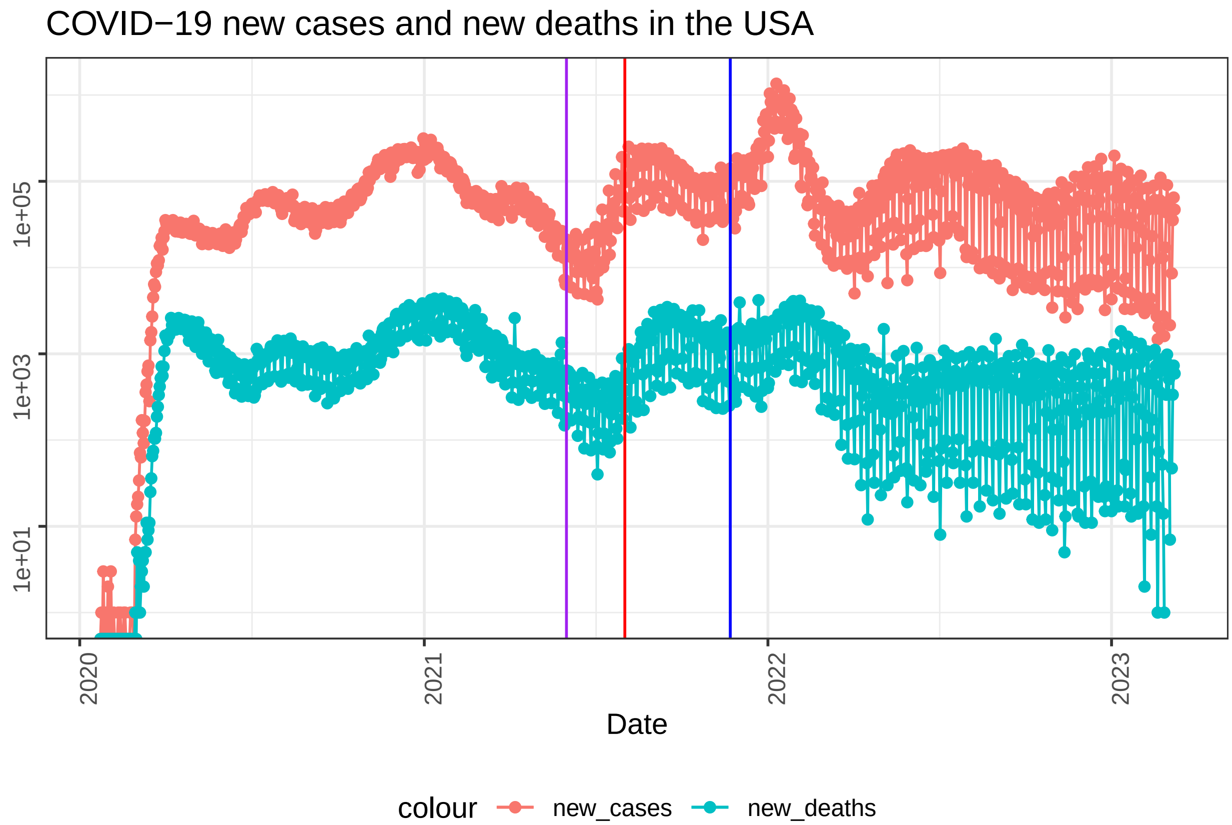
new_cases	new_deaths	province_state	country_region	date	cases	deaths	deaths_per_mill	population
0	0	Wyoming	US	2023-03-04	185159	2002	3459.125	578759
0	0	Wyoming	US	2023-03-05	185159	2002	3459.125	578759
0	0	Wyoming	US	2023-03-06	185159	2002	3459.125	578759
226	2	Wyoming	US	2023-03-07	185385	2004	3462.581	578759
0	0	Wyoming	US	2023-03-08	185385	2004	3462.581	578759
0	0	Wyoming	US	2023-03-09	185385	2004	3462.581	578759

Table 10: us_totals with new variables

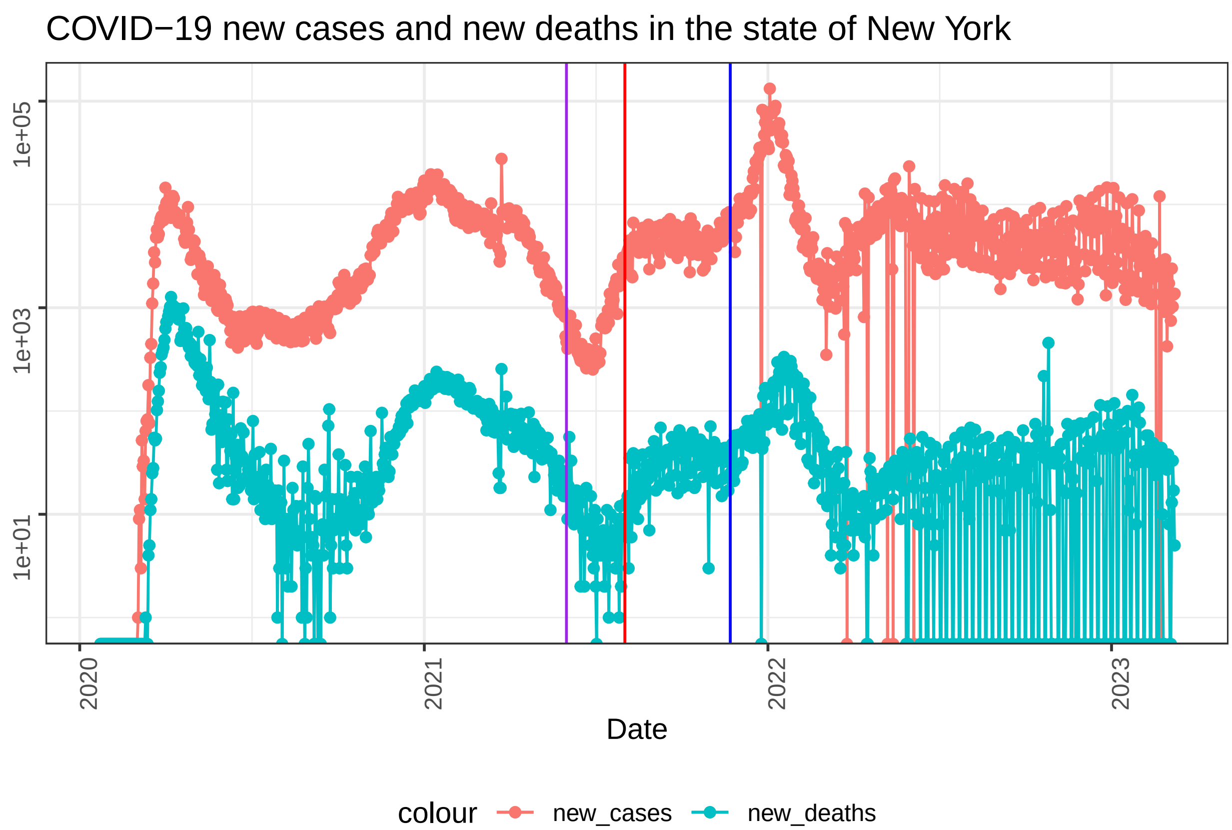
new_cases	new_deaths	country_region	date	cases	deaths	deaths_per_mill	population
2147	7	US	2023-03-04	103650837	1122172	3371.150	332875137
-3862	-38	US	2023-03-05	103646975	1122134	3371.036	332875137
8564	47	US	2023-03-06	103655539	1122181	3371.177	332875137
35371	335	US	2023-03-07	103690910	1122516	3372.183	332875137
64861	730	US	2023-03-08	103755771	1123246	3374.376	332875137
46931	590	US	2023-03-09	103802702	1123836	3376.149	332875137

Visualising New Cases and New Deaths

Let’s graph the new cases and new deaths over time for the US totals.



The graph above shows the number of new cases and new deaths (in red and blue respectively) in the USA since the start of the COVID-19 pandemic. Several important dates in the pandemic are marked on the graph. In June 2021, the delta sub variant of COVID-19 became the dominant strain and is marked in purple. The date of August 2nd 2021 is marked in red to highlight the date where the vaccination goal of 70% of the US population vaccinated with at least one dose of the COVID-19 vaccine was met. In November 2021, the omicron sub variant of COVID-19 became was identified and is marked in blue (https://www.cdc.gov/museum/timeline/covid19.html).



The graph above shows the number of new cases and new deaths (in red and blue respectively) in the state of New York since the start of the COVID-19 pandemic. Several important dates in the pandemic are marked on the graph. In June 2021, the delta sub variant of COVID-19 became the dominant strain and is marked in purple. The date of August 2nd 2021 is marked in red to highlight the date where the vaccination goal of 70% of the US population vaccinated with at least one dose of the COVID-19 vaccine was met. In November 2021, the omicron sub variant of COVID-19 became was identified and is marked in blue (https://www.cdc.gov/museum/timeline/covid19.html).

Data Transformation - Case and Death rates

After analyzing the new cases and deaths, we wanted to ask, which states were the worst in terms of case rate and death rate per population. To do so, we need to go back to our data and create some new variables.

Analysis of highest and lowest rates

Table 11: Ten lowest case rate states

deaths_per_thou	cases_per_thou	province_state	deaths	cases	population
0.6110602	149.5300	American Samoa	34	8320	55641
2.7364995	225.8302	Maryland	16544	1365297	6045680
2.2222818	228.4552	Oregon	9373	963564	4217737
1.2119178	231.3178	Virgin Islands	130	24813	107268
2.1782278	236.6665	Maine	2928	318130	1344212
1.4888083	244.5844	Vermont	929	152618	623989
0.7435079	247.8239	Northern Mariana Islands	41	13666	55144
2.0290500	252.1364	District of Columbia	1432	177945	705749

deaths_per_thou	cases_per_thou	province_state	deaths	cases	population
2.0595168	253.3080	Washington	15683	1928913	7614893
3.4513612	268.2283	Missouri	22870	1777380	6626371

Table 12: Ten lowest death rate states

deaths_per_thou	cases_per_thou	province_state	deaths	cases	population
0.6110602	149.5300	American Samoa	34	8320	55641
0.7435079	247.8239	Northern Mariana Islands	41	13666	55144
1.2119178	231.3178	Virgin Islands	130	24813	107268
1.3002588	268.8153	Hawaii	1841	380608	1415872
1.4888083	244.5844	Vermont	929	152618	623989
1.5507575	293.3387	Puerto Rico	5823	1101469	3754939
1.6525482	340.0999	Utah	5298	1090346	3205958
2.0054116	415.1917	Alaska	1486	307655	740995
2.0290500	252.1364	District of Columbia	1432	177945	705749
2.0595168	253.3080	Washington	15683	1928913	7614893

Table 13: Ten highest case rate states

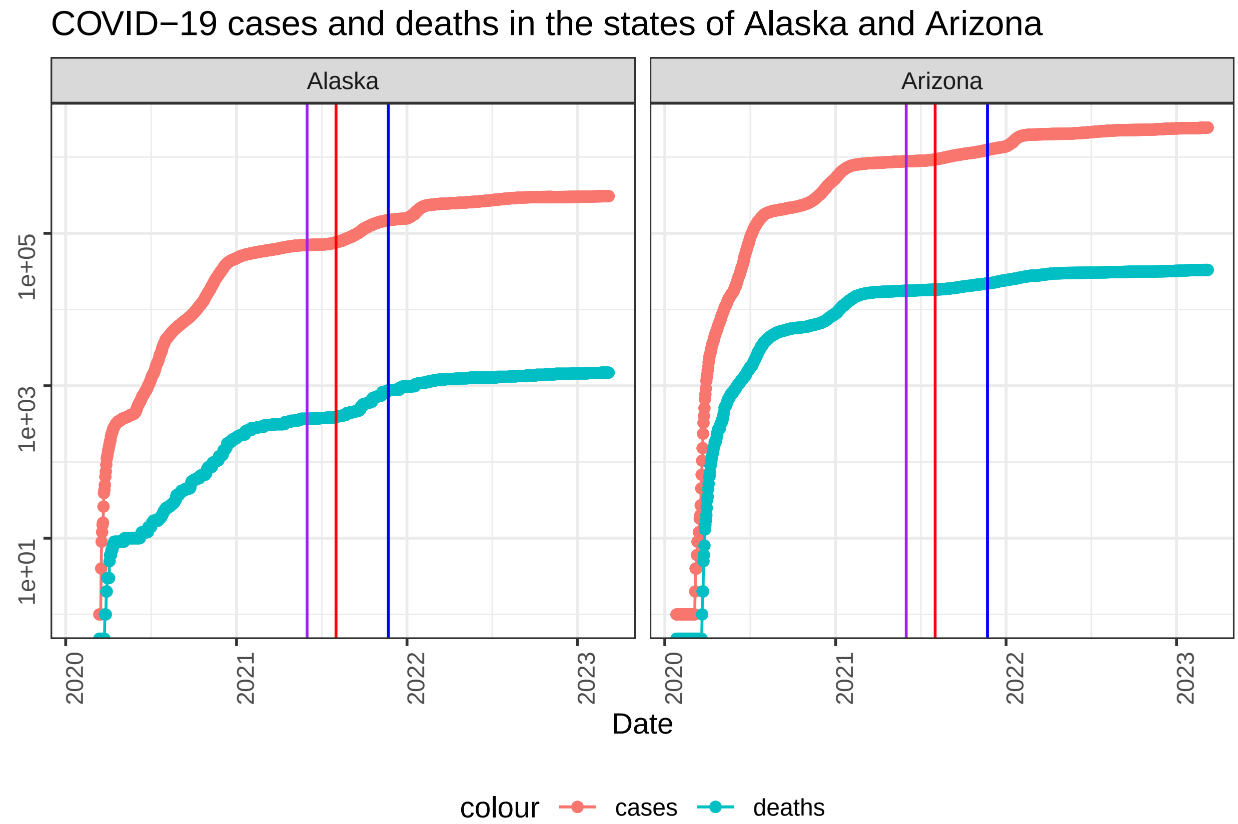
deaths_per_thou	cases_per_thou	province_state	deaths	cases	population
3.653146	434.8820	Rhode Island	3870	460697	1059361
2.005412	415.1917	Alaska	1486	307655	740995
4.058041	384.6457	Kentucky	18130	1718471	4467673
3.241206	376.5442	North Dakota	2470	286950	762062
2.557405	371.5970	Guam	420	61027	164229
4.284998	368.2920	Tennessee	29263	2515130	6829174
4.441600	358.6536	West Virginia	7960	642760	1792147
3.806776	356.7042	South Carolina	19600	1836568	5148714
4.043722	352.6717	Florida	86850	7574590	21477737
3.966215	349.2799	New York	77157	6794738	19453561

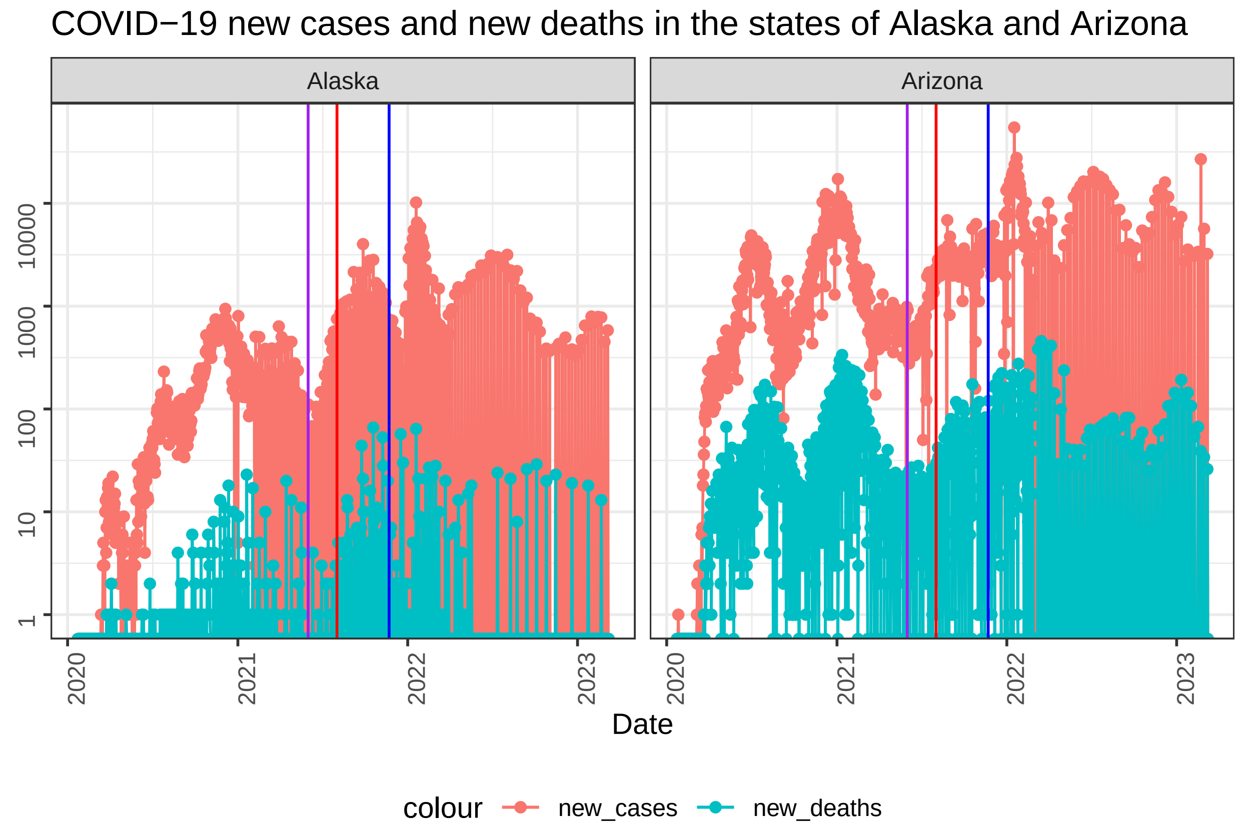
Table 14: Ten highest death rate states

deaths_per_thou	cases_per_thou	province_state	deaths	cases	population
4.547779	335.7067	Arizona	33102	2443514	7278717
4.541858	326.2417	Oklahoma	17972	1290929	3956971
4.492383	332.8987	Mississippi	13370	990756	2976149
4.441600	358.6536	West Virginia	7960	642760	1792147
4.321287	319.9732	New Mexico	9061	670929	2096829
4.314395	333.6476	Arkansas	13020	1006883	3017804
4.289457	335.4010	Alabama	21032	1644533	4903185
4.284998	368.2920	Tennessee	29263	2515130	6829174
4.226054	306.8157	Michigan	42205	3064125	9986857
4.058041	384.6457	Kentucky	18130	1718471	4467673

Analysis of Case and Death Rates - Alaska vs. Arizona

An interesting case is that of Alaska where there is a high case rate, but a relatively low death rate. Perhaps it would be interesting to visualize the cases in Alaska vs the cases in Arizona which had a lower case rate, but a higher death rate than Alaska to better understand why that may be.





In the graphs above, the same important dates are marked as in previous graphs.

There are plenty of factors that may have contributed to the higher death rate in Arizona. I would hypothesize that both the difference in age demographics and population density between Alaska and Arizona are key contributing factors, but I would need to have the age of patients for each case to test this hypothesis. Nevertheless, 13.9% of the population of Alaska is over 65 years of age vs. 18.8% in Arizona (source US Census Bureau). Considering the fact that persons over 65 years of age have a higher risk of mortality from COVID-19 infection, this may explain some of the difference in death rates seen between the two states.

Additionally, we can see that the number of cases in Alaska remained relatively low early in the pandemic. This may be partially due to the relative isolation of Alaska and the low population density compared to Arizona. This my be a key factor contributing the the difference in death rate as the severity of disease in later variants of COVID-19 such as Delta and Omicron decreased significantly.

Modeling COVID-19 deaths from number of cases

US Data

Let's start of modeling our COVID-19 Data by creating a simple linear model of deaths per thousand predicted by cases per thousand using the US state totals.

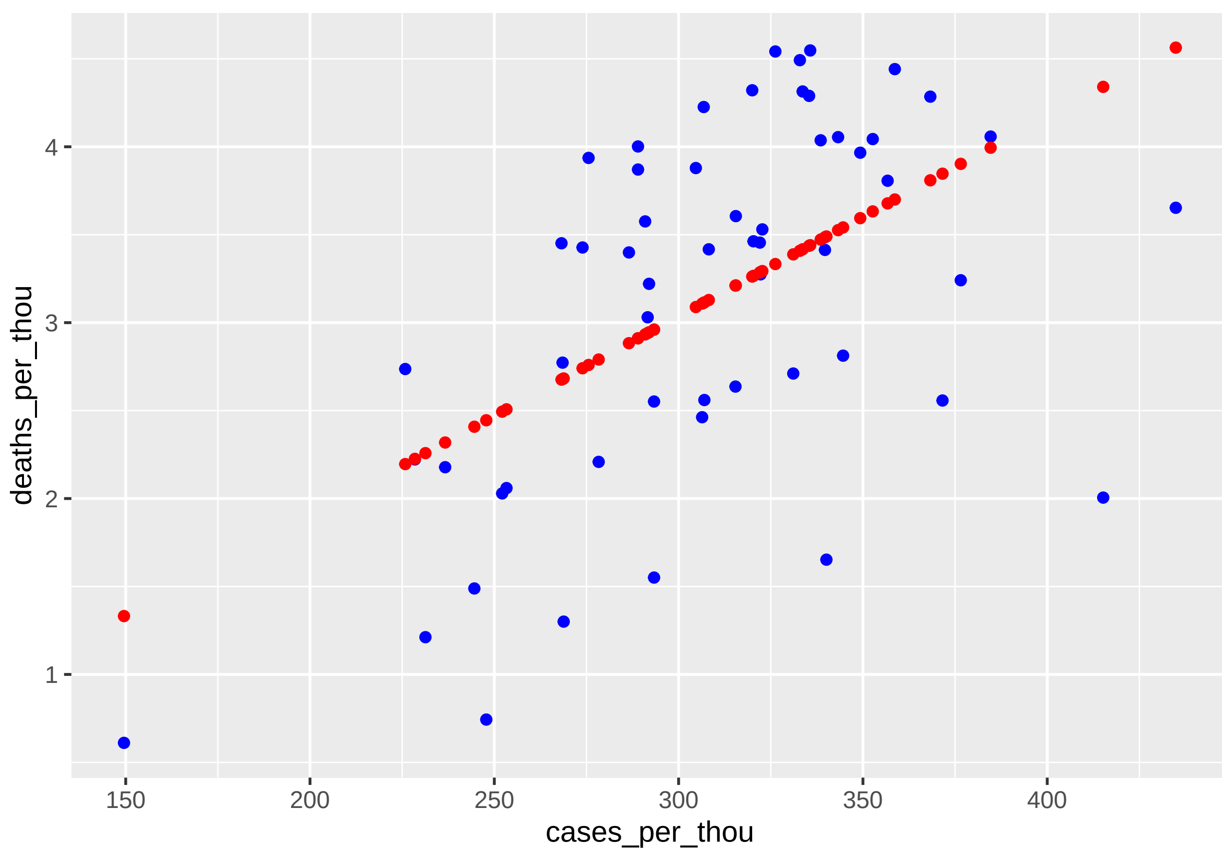
```
##
## Call:
## lm(formula = deaths_per_thou ~ cases_per_thou, data = us_state_totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3352 -0.5978  0.1491  0.6535  1.2086
```

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.36167    0.72480  -0.499     0.62
## cases_per_thou  0.01133    0.00232   4.881 9.76e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8615 on 54 degrees of freedom
## Multiple R-squared:  0.3061, Adjusted R-squared:  0.2933
## F-statistic: 23.82 on 1 and 54 DF,  p-value: 9.763e-06

## # A tibble: 1 x 6
##   province_state deaths cases population cases_per_thou deaths_per_thou
##   <chr>           <dbl> <dbl>      <dbl>          <dbl>           <dbl>
## 1 American Samoa     34  8320      55641           150.           0.611

## # A tibble: 1 x 6
##   province_state deaths  cases population cases_per_thou deaths_per_thou
##   <chr>           <dbl>  <dbl>      <dbl>          <dbl>           <dbl>
## 1 Rhode Island     3870 460697    1059361           435.            3.65
```

Now that we have created a model and added the predicted values to a new dataset, we can visualise the predicted values versus the actual values.



We can see from the data that cases per thousand is a predictor of deaths per thousand, however, there are clearly other factors leading to differences between one state and another.

Global data

Let's look at how this differs on a global scale. To do so, we will need to create a global totals with the deaths and cases per thousand variables.

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

## # A tibble: 6 x 6
##   country_region      deaths  cases population cases_per_thou deaths_per_thou
##   <chr>                <dbl>  <dbl>      <dbl>          <dbl>           <dbl>
## 1 Afghanistan           7896 209451   38928341           5.38          0.203
## 2 Albania               3598 334457    2877800         116.            1.25
## 3 Algeria               6881 271496   43851043           6.19          0.157
## 4 Andorra                165  47890      77265         620.            2.14
## 5 Angola                1933 105288   32866268           3.20          0.0588
## 6 Antigua and Barbuda    146   9106      97928          93.0           1.49
```

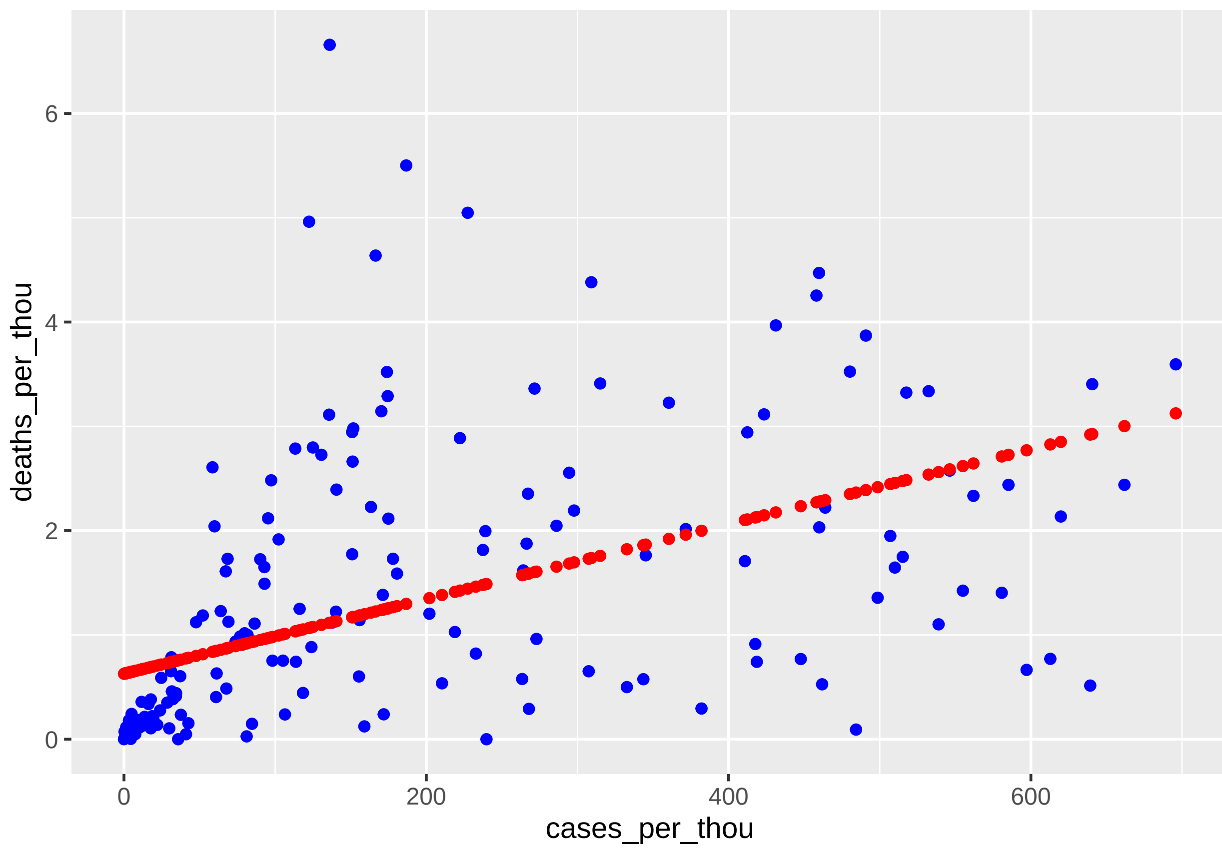
Now that we have created out global totals dataset, we can create a linear model and look at the summary.

```
##
## Call:
## lm(formula = deaths_per_thou ~ cases_per_thou, data = global_totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4069 -0.6090 -0.3898  0.4689  5.5423
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.6277935  0.1116605   5.622 6.57e-08 ***
## cases_per_thou 0.0035877  0.0004393   8.167 4.13e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.151 on 192 degrees of freedom
## Multiple R-squared:  0.2578, Adjusted R-squared:  0.254
## F-statistic:  66.7 on 1 and 192 DF,  p-value: 4.127e-14

## # A tibble: 1 x 6
##   country_region deaths cases population cases_per_thou deaths_per_thou
##   <chr>           <dbl> <dbl>      <dbl>          <dbl>           <dbl>
## 1 Korea, North        6     1   25778815      0.0000388        0.000233

## # A tibble: 1 x 6
##   country_region deaths cases population cases_per_thou deaths_per_thou
##   <chr>           <dbl> <dbl>      <dbl>          <dbl>           <dbl>
## 1 San Marino        122 23616      33938           696.            3.59
```

Let's visualise the predictions on the global linear model.



Here we see again that cases per thousand remains a good predictor of deaths per thousand, however, there are definitely other factors leading to the differences observed from one country to another. As mentioned previously in the Arizona vs. Alaska example, age demographics may play a role here, however, there may also be issues with how data are reported from one country to another. This leads to our final section on bias identification.

Bias identification and Conclusions

Within the COVID-19 dataset, there may be significant differences from how deaths related to COVID-19 are recorded from one country/state to another as well as how frequently cases are recorded. This makes it challenging to compare data from one country to another and may result in bias in the datasets.

Additionally, different countries have had different access to vaccines throughout the pandemic. This may impact the relationship between cases and deaths, specifically in years following the roll out of the vaccine which was approved for emergency use in December 2020. To complicate matters further, vaccine uptake rates, mask mandates, and other COVID related restrictions varied wildly from one country/state to another.

Furthermore, following mass vaccination programs and decreasing cases globally, the quality of data on COVID-19 and the frequency of reporting has decreased overall. Data in 2023 may be relatively unreliable. With new subvariants showing a high number of mutations and possible vaccine evasion, there may be an additional wave of COVID-19 coming over the northern winter season with the need to update COVID-19 vaccines to provide protection to those who need it most.

In summary, in analyzing the US and global data sets joined by aggregating the data from JHU, on 20/08/2023 the total number of cases in the USA was 2023-03-09 and the total number of deaths since the start of the pandemic in the USA has sadly reached 1.123836×10^6. Globally, we have seen a total of 6.6705286×10^8 cases and sadly 6.729347×10^6 deaths since the start of the pandemic.

When comparing differences between the states, I chose the states of Alaska and Arizona for two reasons. One, I was noted that Alaska had a very low death rate despite having a high case rate overall. Arizona had a similarly high case rate, but a much higher death rate than Alaska. Two, I have family that has lived in both of those states and I thought it would be interesting to compare the two.

The differences between the two states may be due to a variety of factors, but population density, differences in age demographics, the geographical isolation and the timeline of COVID-19 in Alaska are the factors that I would hypothesize have significantly contributed to these differences.

Following the comparison of case and death rates, I then asked the question of can death rates be modeled by case rates. I built two models, one for the US only by state and one for the global dataset by country. Differences between one country and another may be due to the reasons stated above with respect to bias in the data amongst others.

As mentioned previously, COVID-19 is here to stay with a potential new wave to come this year as new variants with high levels of mutations arise. Hopefully we can learn from this pandemic to be better prepared for the next zoonotic transmission and pandemic that arises without the loss of 7 million lives globally.