Covid-19 Project

Siavash Keivani

2023-03-01

COVID-19 Data Analysis

In this analysis we will take a look at USA national covid-19 cases and deaths and compare them to trends in California. We will also create 2 models to asses the impact of cases per thousand on deaths per thousand, and to predict USA deaths per million in 2024 using the open source prophet library.

These data sets are provided by Johns Hopkins on Github. This data is gathered from different sources which are listed by Johns Hopkins on their github links below.

This data set includes location information for countries, and the number of covid cases and deaths.

Libraries Used

library("tidyverse") library("lubridate") library("dplyr") library("ggplot2") library("scales") library("readr") library("prophet") library("sf") library("ggspatial") library("plotly") library("maps") library("usmap")

Read in Data

```
urlfile1 <- 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_c
urlfile2 <- 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_c
urlfile3 <- 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_c
urlfile4 <- 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_c
```

```
us.cases <- read_csv(url(urlfile1))
global.cases <- read_csv(url(urlfile2))
us.deaths <- read_csv(url(urlfile3))
global.deaths <- read_csv(url(urlfile4))
```

Tidy Data

Tidy global.cases

```
# Pivot global cases and name columns
global.cases <- global.cases %>%
  pivot_longer(cols = -c('Province/State','Country/Region','Lat','Long'),
    names_to = "date",
    values_to = "cases")
```

```r
#remove lat long columns
global.cases<-select(global.cases,-c(Lat,Long))
```

Tidy global.deaths

```r
# pivot and give column names
global.deaths <- global.deaths %>%
  pivot_longer(cols = -c('Province/State','Country/Region','Lat','Long'),
    names_to = "date",
    values_to = "deaths")
```

```r
# remove lat and long columns
global.deaths<-select(global.deaths,-c(Lat,Long))
```

```r
# Merge global cases and deaths
global <- global.cases %>%
  full_join(global.deaths) %>%
  rename(Country_Region = 'Country/Region', Province_State = 'Province/State') %>%
  mutate(date = myd(date))
```

```r
summary(global)
```

```
##  Province_State     Country_Region          date                cases                deaths       
##  Length:329460      Length:329460      Min.   :2001-01-21   Min.   :        0   Min.   :      0  
##  Class :character   Class :character   1st Qu.:2008-07-20   1st Qu.:      674   1st Qu.:      3  
##  Mode  :character   Mode  :character   Median :2016-03-21   Median :    14336   Median :    149  
##                                        Mean   :2016-03-25   Mean   :   955749   Mean   :  13353  
##                                        3rd Qu.:2023-11-21   3rd Qu.:   227878   3rd Qu.:   3017  
##                                        Max.   :2031-12-22   Max.   :103655657   Max.   :1122264
```

```r
# only cases above 0
global<-global %>% filter(cases > 0)
```

Wrangling, cleaning, and merging us.deaths and us.cases

```r
us.cases <- us.cases %>%
  pivot_longer(cols = -(UID:Combined_Key),
               names_to = "date",
               values_to = "cases") %>%
  select(Admin2:cases) %>%
  mutate(date = mdy(date))%>%
  select(-c(Lat,Long_))
```

```r
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 <- us.cases %>%
  full_join(us.deaths)
```

Data wrangling for global data sets

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

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

```
uid <- read_csv(uid_lookup_url) %>%
  select(-c(Lat, Long_, Combined_Key, code3, iso2, iso3, Admin2 ))
```

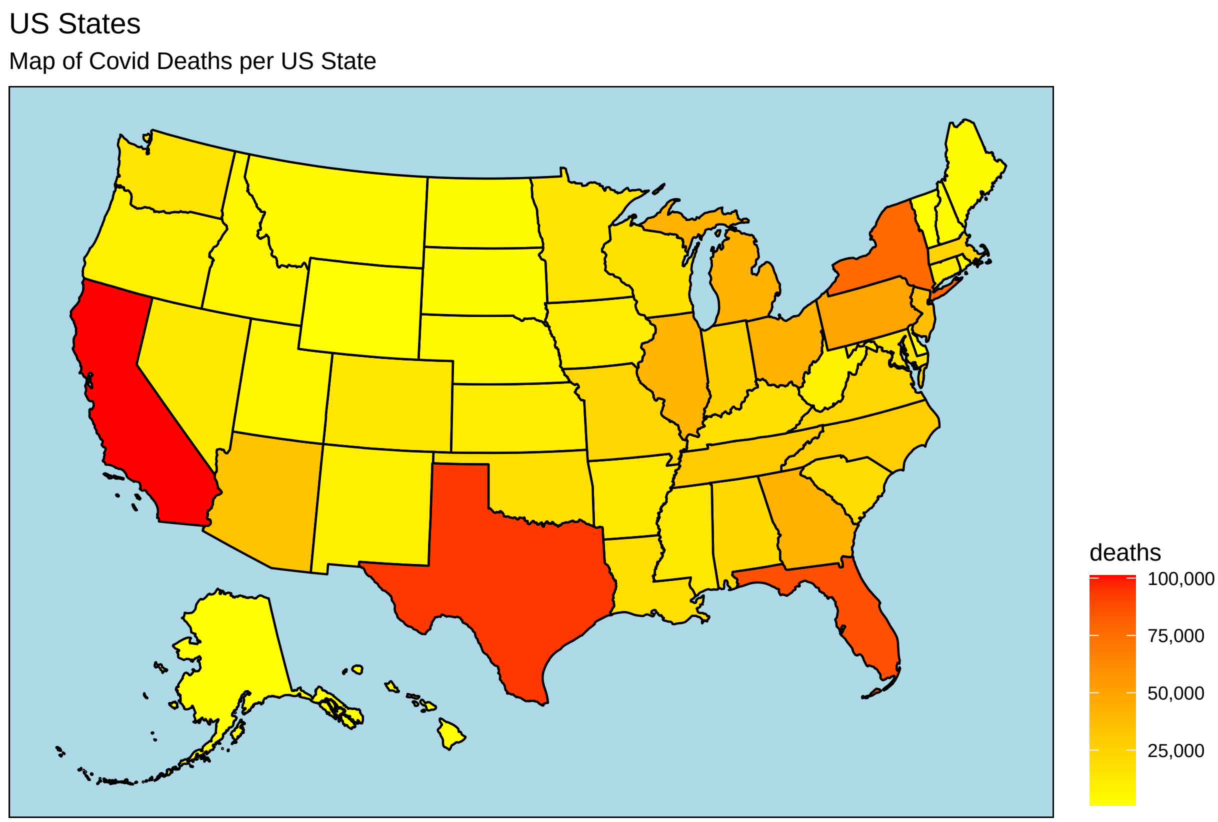
```
global <- global %>%
  left_join(uid, by = c("Province_State", "Country_Region")) %>%
  select(-c(UID, FIPS))%>%
  select(Province_State, Country_Region, date, cases, deaths, Population, Combined_Key)
```

```
us.state <- us %>%
  group_by(Province_State, Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths), Population = sum(Population)) %>%
  mutate(deaths_per_mill = deaths*1000000 / Population) %>%
  select(Province_State, Country_Region, date, cases, deaths, deaths_per_mill, Population) %>%
  ungroup()
```

Creating a Map of Deaths per US State

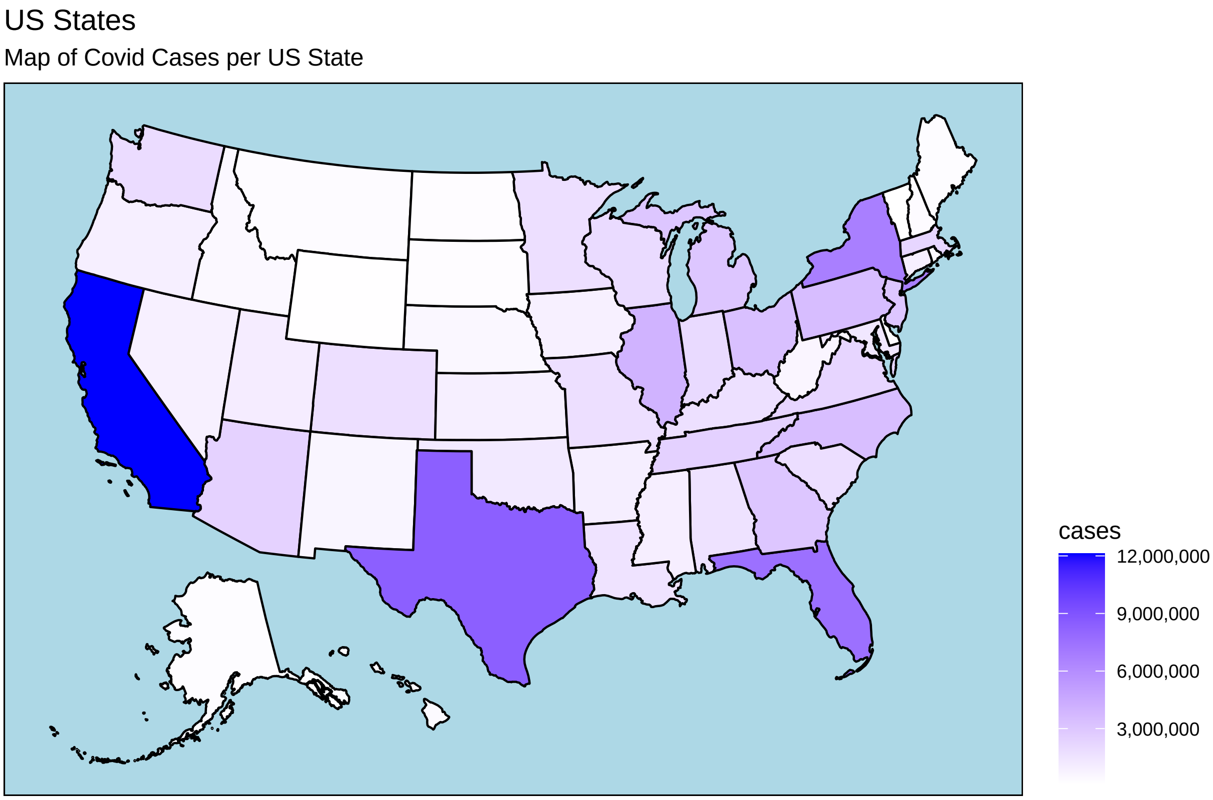
```
us.state.md <- us.state %>%
  select(Province_State,deaths,cases,Population) %>%
  rename(state = "Province_State") %>%
  group_by(state) %>%
  summarize(deaths = max(deaths),cases =  max(cases), Population = max(Population)) %>%
  mutate(deaths.pc=deaths/Population, cases.pc = cases/Population )
 # rename(state = "Province_State")
```

```
plot_usmap(data = us.state.md, values = "deaths", regions = "states") +

  labs(title = "US States",
       subtitle = "Map of Covid Deaths per US State ") +
  theme(panel.background = element_rect(color = "black", fill = "lightblue")) +
  scale_fill_continuous(
    low = "yellow", high = "red", name = "deaths", label = scales::comma) +
  theme(legend.position = "right")
```



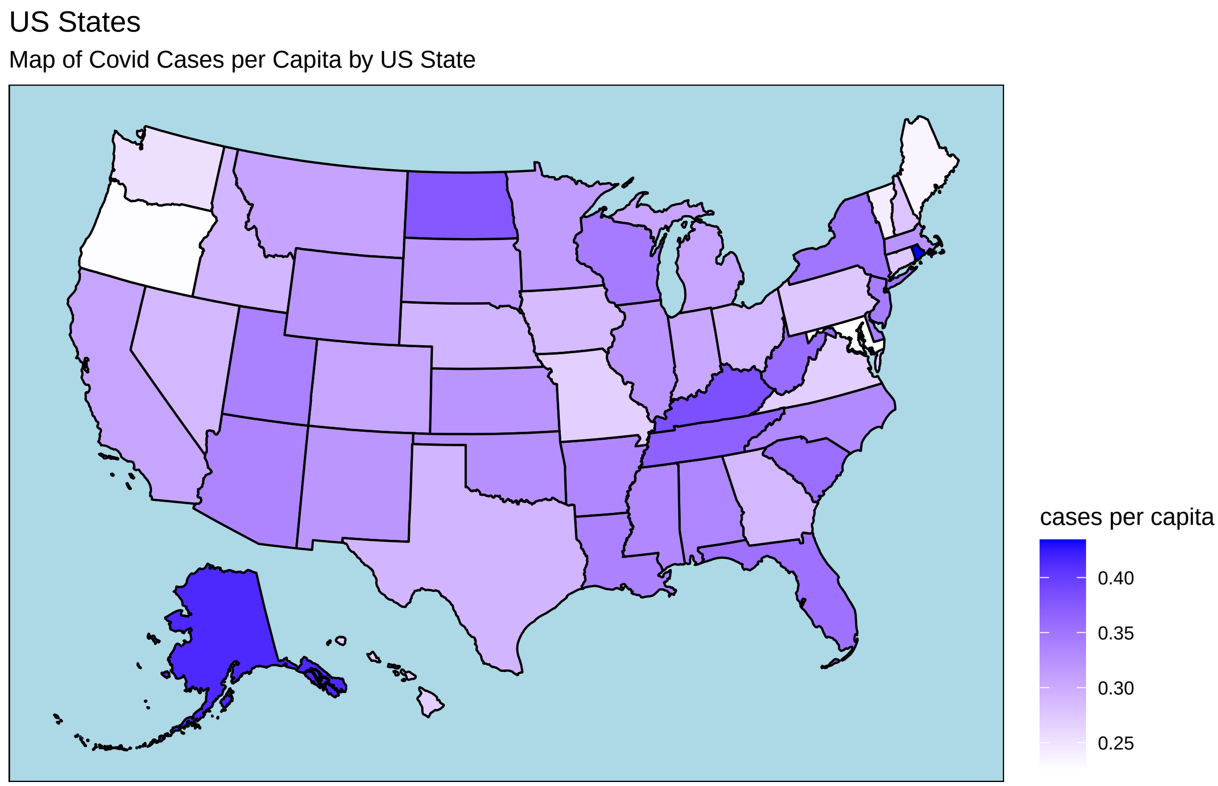
Map of Cases per US State

```
plot_usmap(data = us.state.md, values = "cases", regions = "states") +

  labs(title = "US States",
       subtitle = "Map of Covid Cases per US State ") +
  theme(panel.background = element_rect(color = "black", fill = "lightblue")) +
  scale_fill_continuous(
    low = "white", high = "blue", name = "cases", label = scales::comma) +
  theme(legend.position = "right")
```

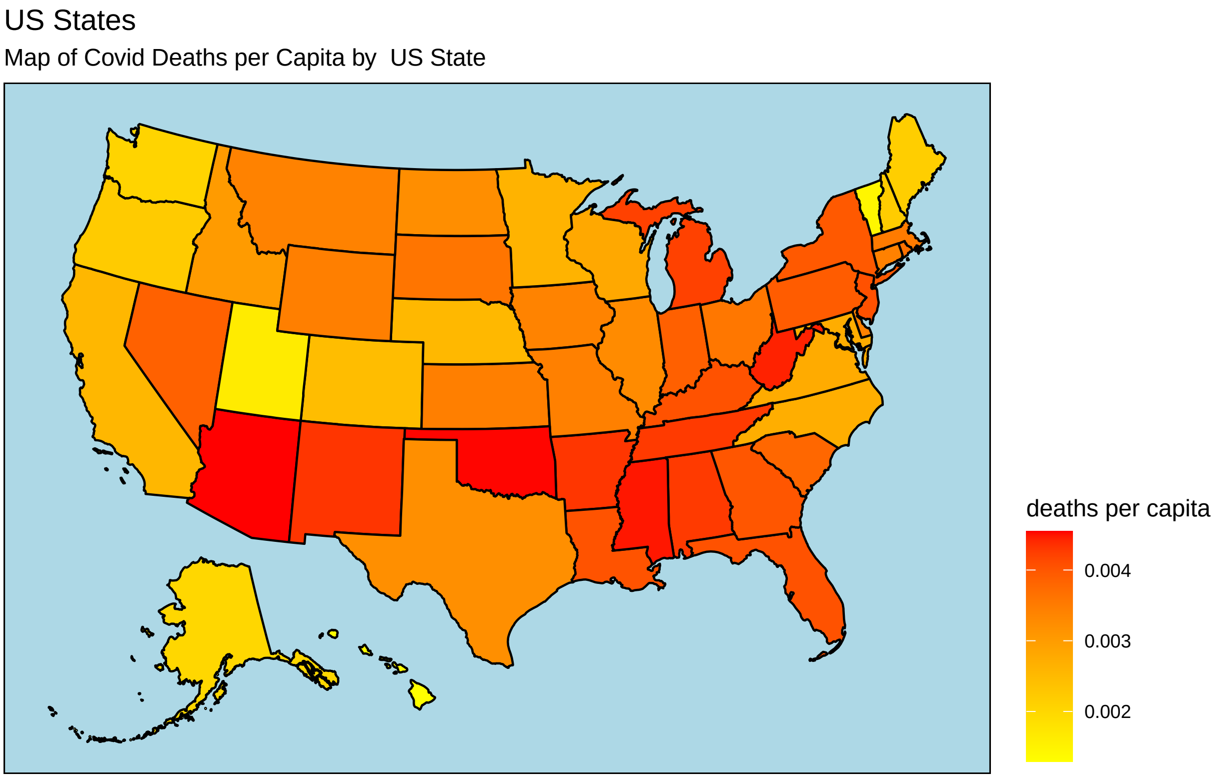


The maps above show California, Texas and Florida as being states with large amounts of cases and deaths. However a more accurate analysis might be to look at cases and deaths per capita. Lets take a look below.

```
plot_usmap(data = us.state.md, values = "cases.pc", regions = "states") +

  labs(title = "US States",
       subtitle = "Map of Covid Cases per Capita by US State ") +
  theme(panel.background = element_rect(color = "black", fill = "lightblue")) +
  scale_fill_continuous(
    low = "white", high = "blue", name = "cases per capita", label = scales::comma) +
  theme(legend.position = "right")
```



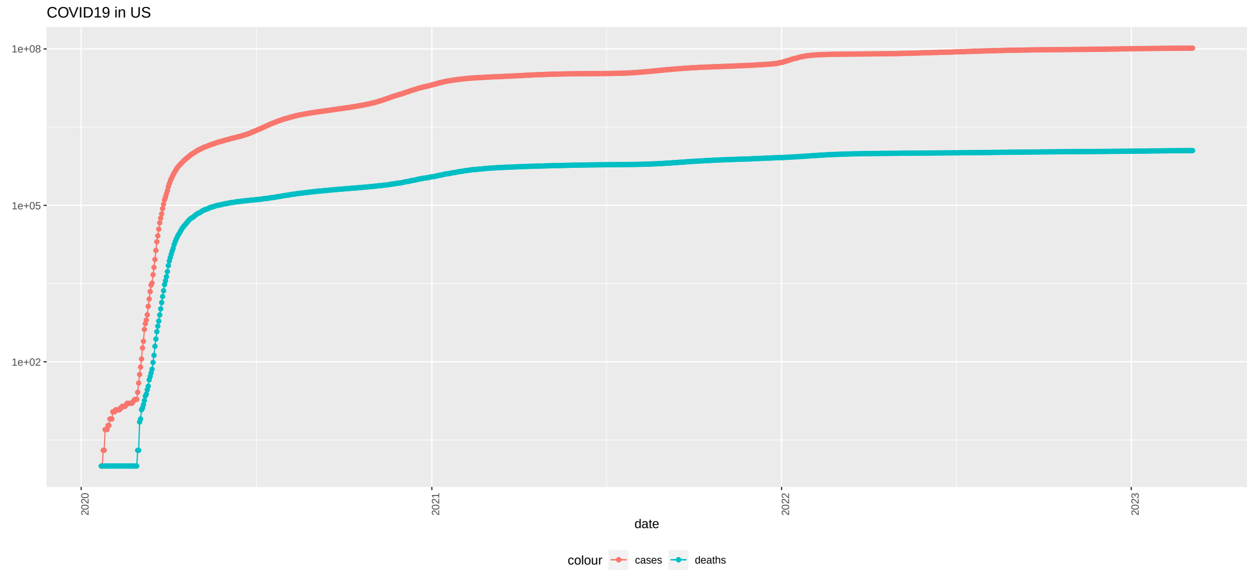

```
plot_usmap(data = us.state.md, values = "deaths.pc", regions = "states") +

  labs(title = "US States",
       subtitle = "Map of Covid Deaths per Capita by  US State ") +
  theme(panel.background = element_rect(color = "black", fill = "lightblue")) +
  scale_fill_continuous(
    low = "yellow", high = "red", name = "deaths per capita", label = scales::comma) +
  theme(legend.position = "right")
```



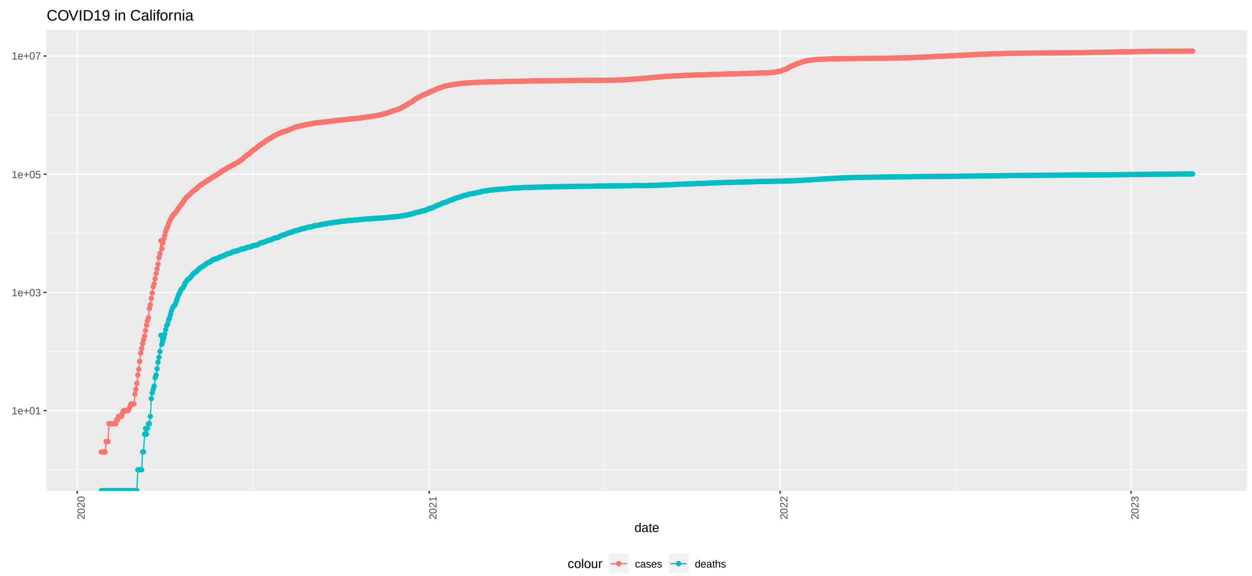
Cases and deaths per capita show a different story when it comes to CA, TX, and FL. California and Texas seem to now be in the middle of the pack in contrast to before when the maps showed them in the lead with cases and deaths.

```
us.totals <- us.state %>%
  group_by(Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths), Population = sum(Population)) %>%
  mutate(deaths_per_mill = deaths*1000000 / Population) %>%
  select(Country_Region, date, cases, deaths, deaths_per_mill, Population) %>%
  ungroup()
```

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




```
state <- "California"
us.state %>%
  filter(Province_State == state) %>%
  filter(cases > 0) %>%
  ggplot(aes(x = date, y = cases)) +
  geom_line(aes(color = "cases")) +
  geom_point(aes(color="cases")) +
  geom_line(aes(y = deaths, color = "deaths")) +
  geom_point(aes(y = deaths, color = "deaths")) +
  scale_y_log10()+
  theme(legend.position = "bottom", axis.text.x = element_text(angle=90)) +
  labs(title = "COVID19 in California", y = NULL)
```



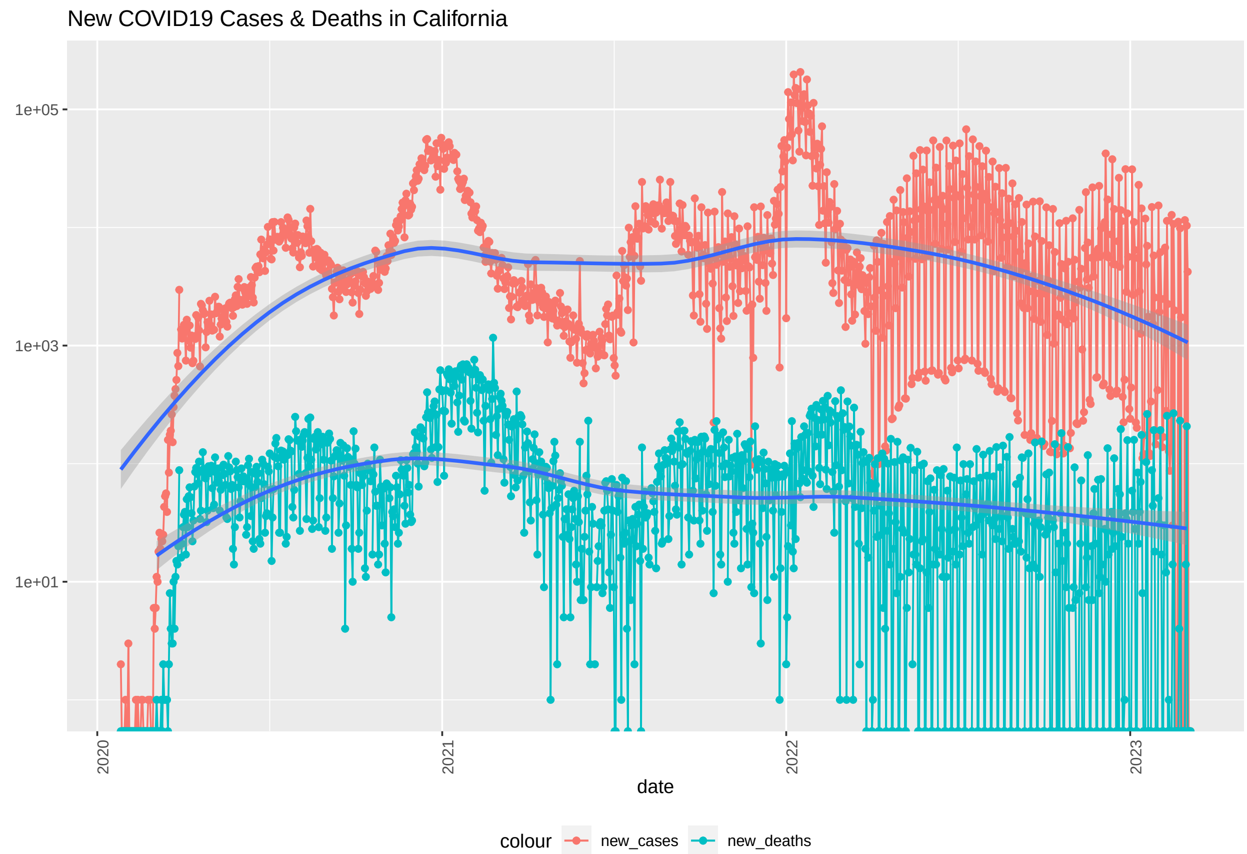
We see from the above visualizations that covid cases and deaths follow similar trends when comparing California to the USA as a whole. We see an exponential rise in both cases and deaths. then in 2022-2023 we see that cases and deaths plateu. Early on deaths were closer to cases but as vaccines and new treatments rolled out it seems that the gap between deaths and cases has increased with many more cases than deaths.

Taking a Look at New Cases and New Deaths

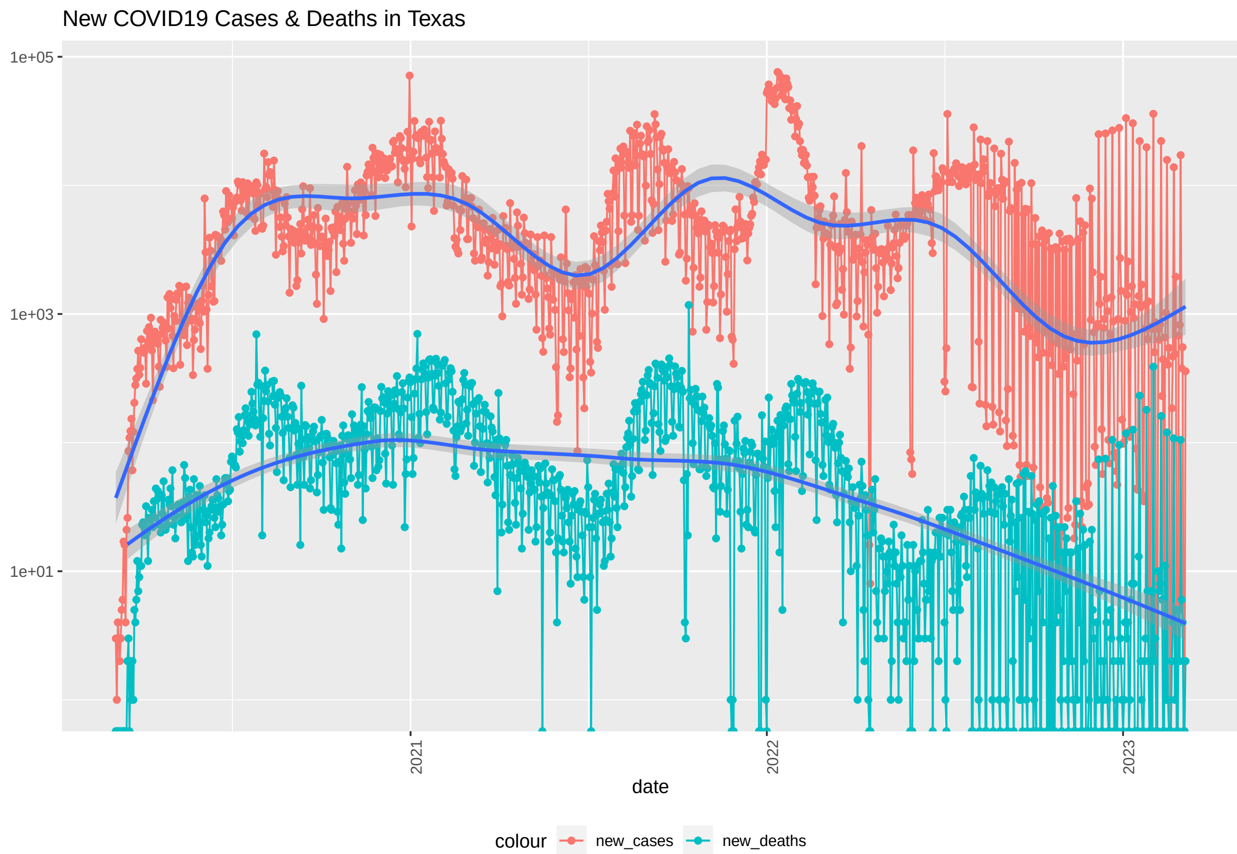
```
# creating new cases and deaths columns
us.state <- us.state %>%
  mutate(new_cases = cases - lag(cases), new_deaths = deaths - lag(deaths))
us.totals<- us.totals %>%
  mutate(new_cases = cases - lag(cases), new_deaths = deaths - lag(deaths))
```

Visualizing New Cases and New Deaths

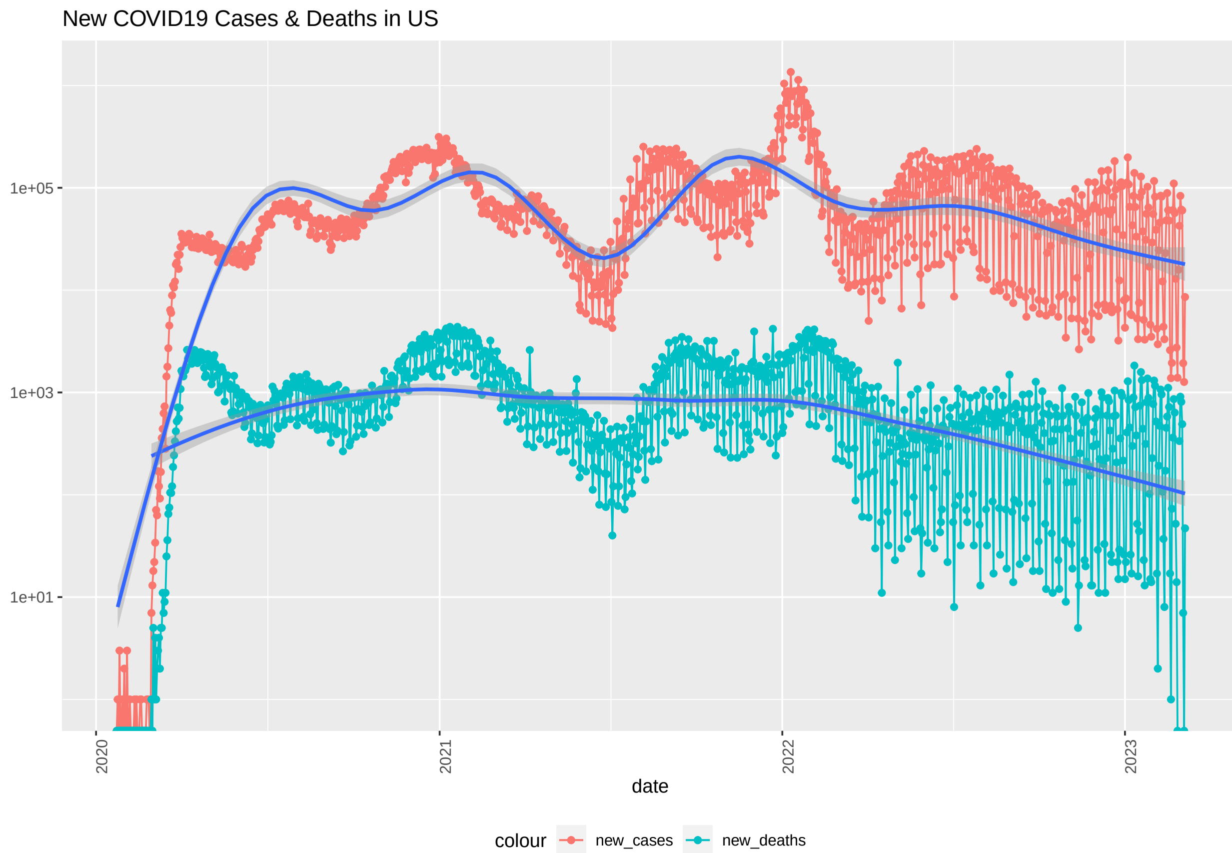
```
state <- "California"
us.state %>%
  filter(Province_State == state) %>%
  filter(cases > 0) %>%
  ggplot(aes(x = date, y = new_cases)) +
  geom_line(aes(color = "new_cases")) +
  geom_point(aes(color="new_cases")) +
  geom_line(aes(y = new_deaths, color = "new_deaths")) +
  geom_point(aes(y = new_deaths, color = "new_deaths")) +
  scale_y_log10()+
  theme(legend.position = "bottom", axis.text.x = element_text(angle=90)) +
  labs(title = "New COVID19 Cases & Deaths in California", y = NULL) +
  geom_smooth(aes(x=date,y=new_deaths),method = loess) +
  geom_smooth(aes(x=date,y=new_cases), method = loess)
```




```r
state <- "Texas"
us.state %>%
  filter(Province_State == state) %>%
  filter(cases > 0) %>%
  ggplot(aes(x = date, y = new_cases)) +
  geom_line(aes(color = "new_cases")) +
  geom_point(aes(color="new_cases")) +
  geom_line(aes(y = new_deaths, color = "new_deaths")) +
  geom_point(aes(y = new_deaths, color = "new_deaths")) +
  scale_y_log10()+
  theme(legend.position = "bottom", axis.text.x = element_text(angle=90)) +
  labs(title = "New COVID19 Cases & Deaths in Texas", y = NULL) +
  geom_smooth(aes(x=date,y=new_deaths), method = loess) +
  geom_smooth(aes(x=date,y=new_cases), methos = loess)
```




```
us.totals %>%
  filter(cases > 0) %>%
  ggplot(aes(x = date, y = new_cases)) +
  geom_line(aes(color = "new_cases")) +
  geom_point(aes(color="new_cases")) +
  geom_line(aes(y = new_deaths, color = "new_deaths")) +
  geom_point(aes(y = new_deaths, color = "new_deaths")) +
  scale_y_log10()+
  theme(legend.position = "bottom", axis.text.x = element_text(angle=90)) +
  labs(title = "New COVID19 Cases & Deaths in US", y = NULL) +
  geom_smooth(aes(x=date,y=new_deaths), method = loess) +
  geom_smooth(aes(x=date,y=new_cases), methos = loess)
```

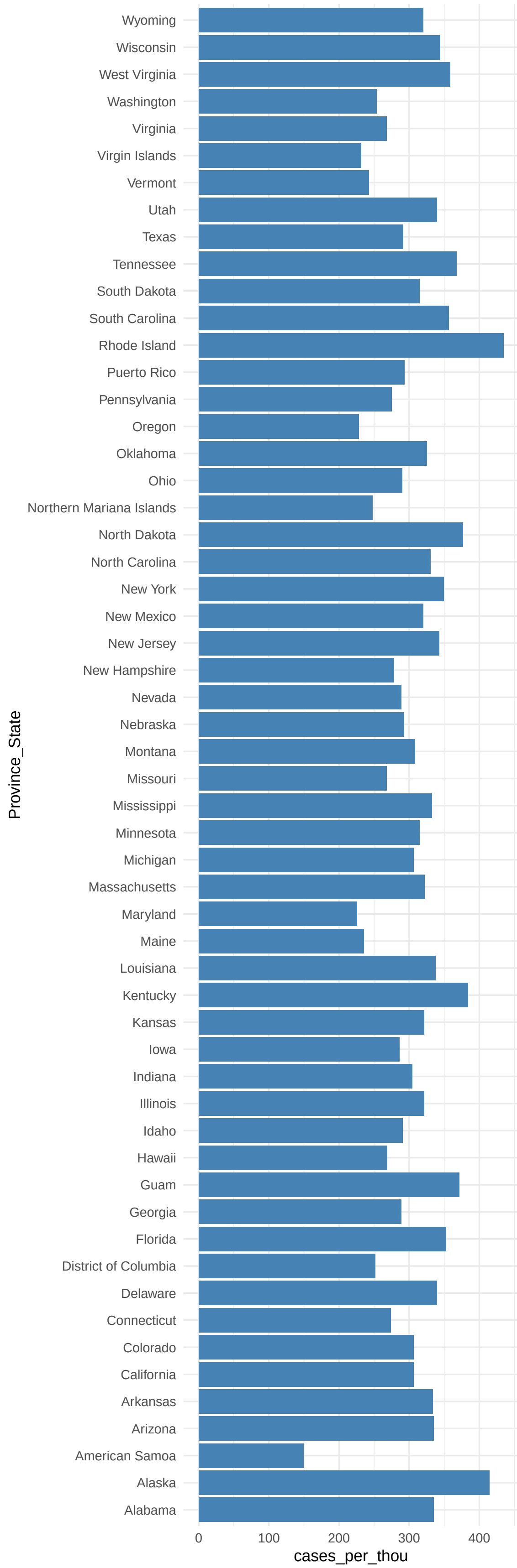


Data in 2022-2023 is showing a slight decline in new cases and new deaths. We have to think about why this might be the case. It is possible that this is actually the case however data reporting during this time might be a lot weaker.

During the height of the pandemic people were more likely to report their covid case. Now in a world of vaccines and boosters it is quite possible that people do not report when they test positive especially with the large amount of at home test available to the public.

```
us.state.totals <- us.state %>%
  group_by(Province_State)%>%
  summarize(deaths = max(deaths), cases = max(cases), population = max(Population),
            cases_per_thou = cases*1000/population, deaths_per_thou = deaths*1000/population) %>%
  filter(cases > 0, population > 0)
```

```
us.state.totals %>%
  ggplot(aes(x = Province_State, y = cases_per_thou))+
  geom_bar(stat="identity", fill='steelblue')+
  coord_flip()+
  theme_minimal()
```


Wyoming
Wisconsin
West Virginia
Washington
Virginia
Virgin Islands
Vermont
Utah
Texas
Tennessee
South Dakota
South Carolina
Rhode Island
Puerto Rico
Pennsylvania
Oregon
Oklahoma
Ohio
Northern Mariana Islands
North Dakota
North Carolina
New York
New Mexico
New Jersey
New Hampshire
Nevada
Nebraska
Montana
Missouri
Mississippi
Minnesota
Michigan
Massachusetts
Maryland
Maine
Louisiana
Kentucky
Kansas
Iowa
Indiana
Illinois
Idaho
Hawaii
Guam
Georgia
Florida
District of Columbia
Delaware
Connecticut
Colorado
California
Arkansas
Arizona
American Samoa
Alaska
Alabama
Province_State
0
100
200
300
400
cases_per_thou

A glance at cases per thousand show most states between 200 and 350.

Taking a look at the highest and lowest 10 states for deaths per thousand

```
us.state.totals %>%
  slice_min(deaths_per_thou, n =10)
```

```
## # A tibble: 10 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
##  2 Northern Mariana Islands     41   13666      55144           248.           0.744
##  3 Virgin Islands              130   24792     107268           231.           1.21
##  4 Hawaii                     1834  380098    1415872           268.           1.30
##  5 Vermont                     910  151477     623989           243.           1.46
##  6 Puerto Rico                5810 1100557    3754939           293.           1.55
##  7 Utah                       5287 1088853    3205958           340.           1.65
##  8 Alaska                     1486  307073     740995           414.           2.01
##  9 District of Columbia       1430  177714     705749           252.           2.03
## 10 Washington                15683 1928913    7614893           253.           2.06
```

```
us.state.totals %>%
  slice_max(deaths_per_thou, n =10)
```

```
## # A tibble: 10 x 6
##    Province_State deaths   cases population cases_per_thou deaths_per_thou
##    <chr>           <dbl>   <dbl>      <dbl>          <dbl>           <dbl>
##  1 Arizona         33076 2440294    7278717           335.            4.54
##  2 Oklahoma        17940 1287378    3956971           325.            4.53
##  3 Mississippi     13351  989282    2976149           332.            4.49
##  4 West Virginia    7960  642760    1792147           359.            4.44
##  5 New Mexico       9054  670301    2096829           320.            4.32
##  6 Arkansas        13001 1005930    3017804           333.            4.31
##  7 Alabama         21001 1642062    4903185           335.            4.28
##  8 Tennessee       29225 2510002    6829174           368.            4.28
##  9 Michigan        42096 3057222    9986857           306.            4.22
## 10 New Jersey      35995 3046838    8882190           343.            4.05
```

Linear model of deaths per thousand as a function of cases per thousand.

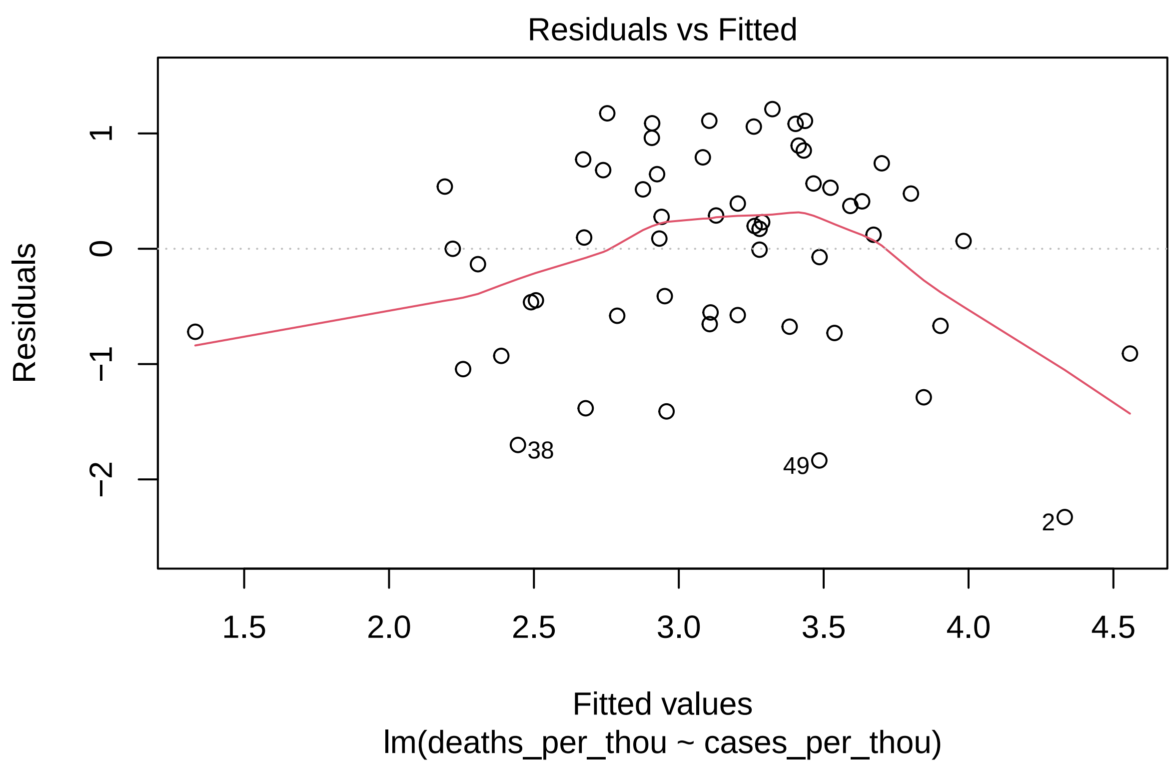
```
lm1 <- lm(deaths_per_thou ~ cases_per_thou, data = us.state.totals)
summary(lm1)
```

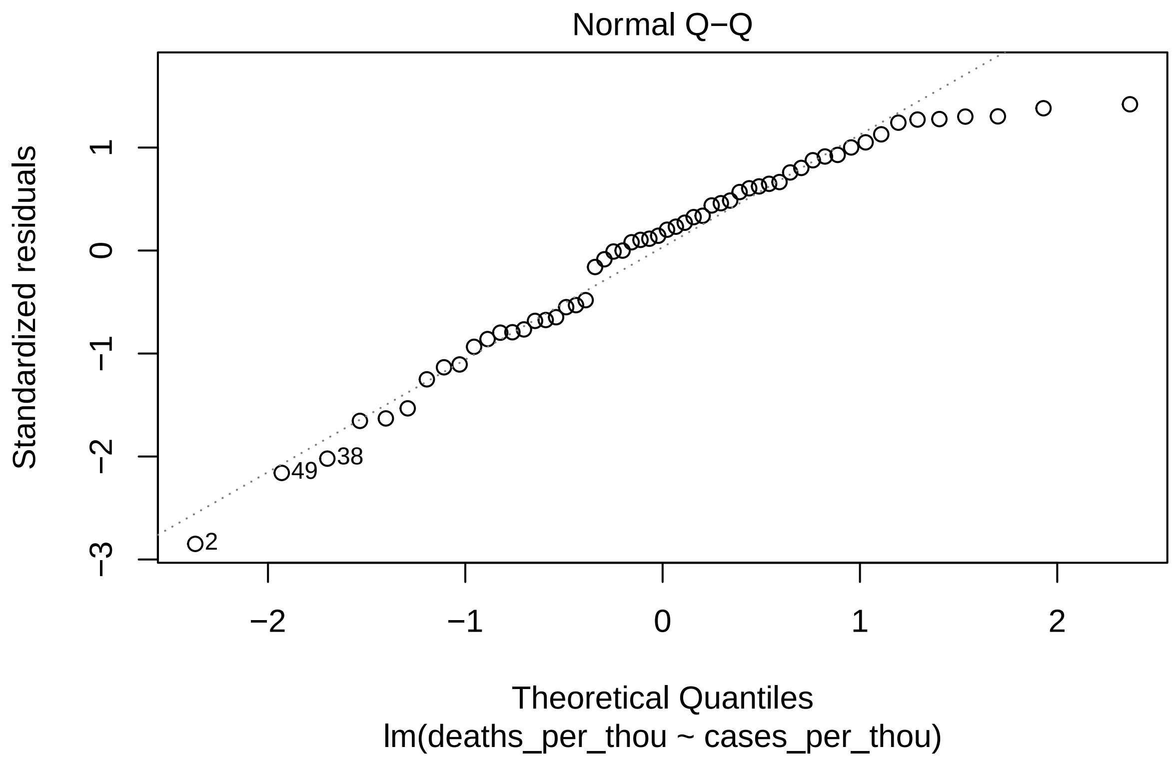
```
##
## Call:
## lm(formula = deaths_per_thou ~ cases_per_thou, data = us.state.totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3267 -0.5992  0.1470  0.6554  1.2107
##
```

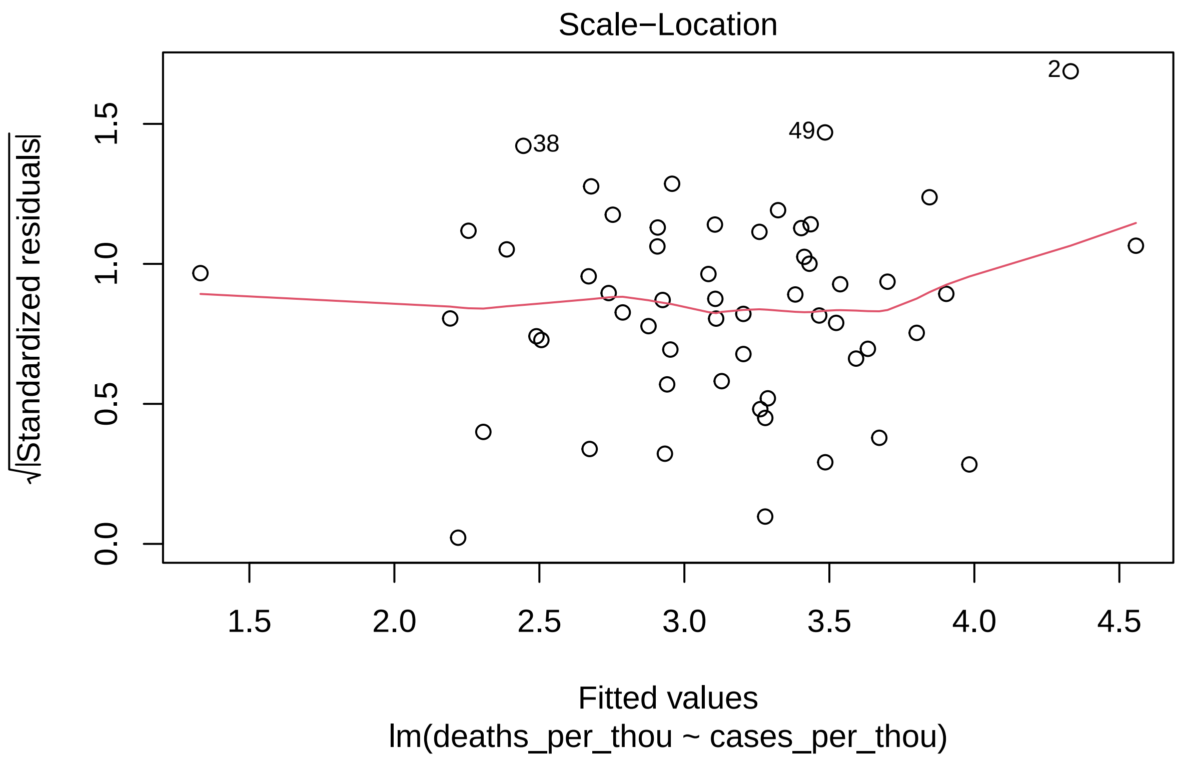
```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.36304    0.72369  -0.502    0.618
## cases_per_thou  0.01133    0.00232   4.883 9.69e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8611 on 54 degrees of freedom
## Multiple R-squared:  0.3063, Adjusted R-squared:  0.2935
## F-statistic: 23.84 on 1 and 54 DF,  p-value: 9.685e-06
```

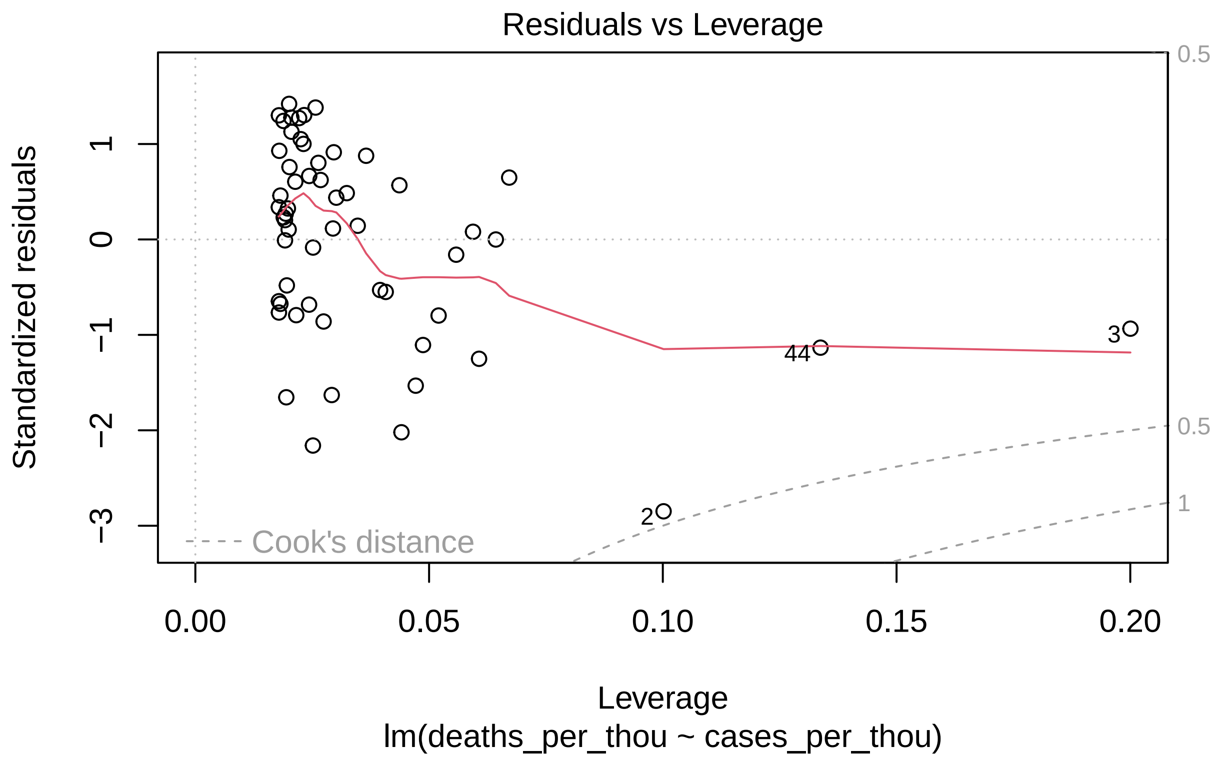
The low p value shows that cases per thousand is statistically significant when predicting deaths per thousand in the united states.

```
plot(lm1)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(deaths_per_thou ~ cases_per_thou)
2
49
38
−3
−2
−1
0
1
2

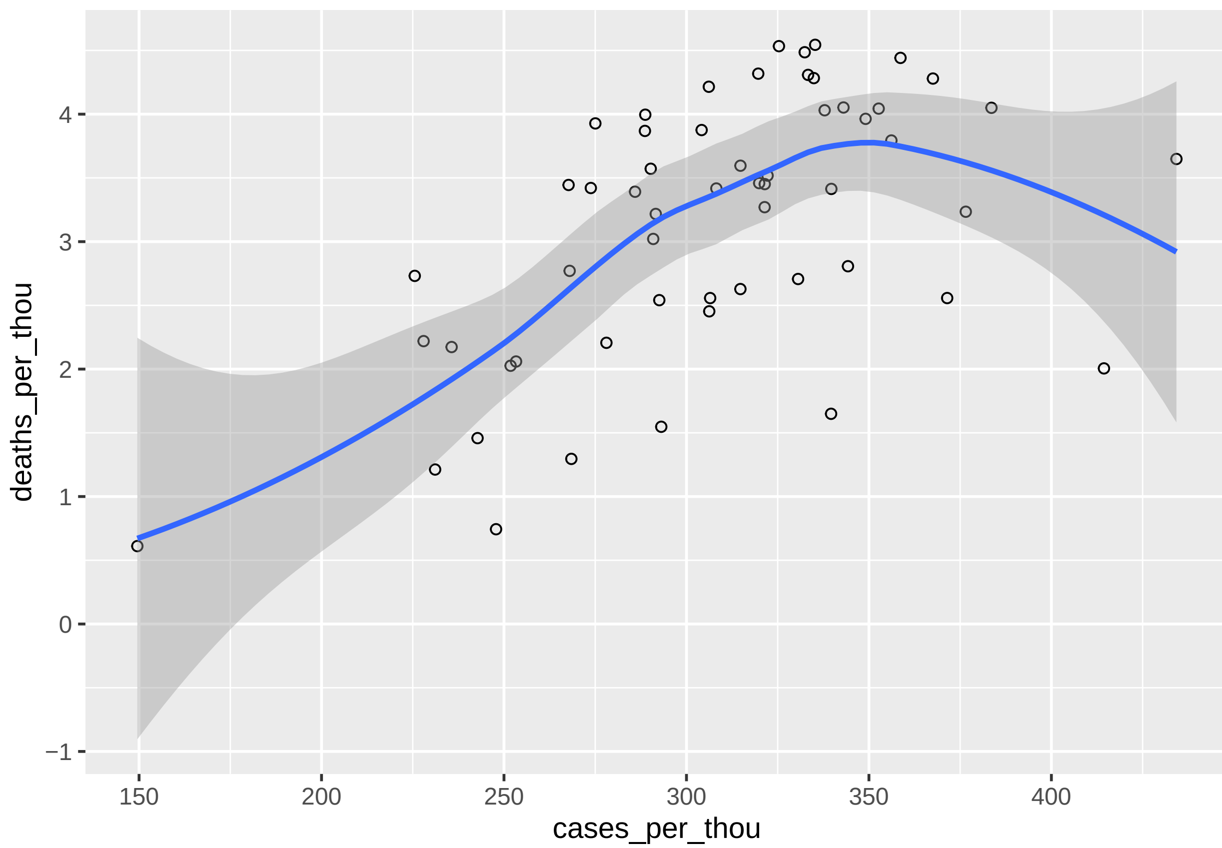

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
1.5
2.0
2.5
3.0
3.5
4.0
4.5
2
38
49
Fitted values
lm(deaths_per_thou ~ cases_per_thou)



Our residual vs fitted plot indicates that the data might not be linear however this does not mean we cant gain any insight from our model.

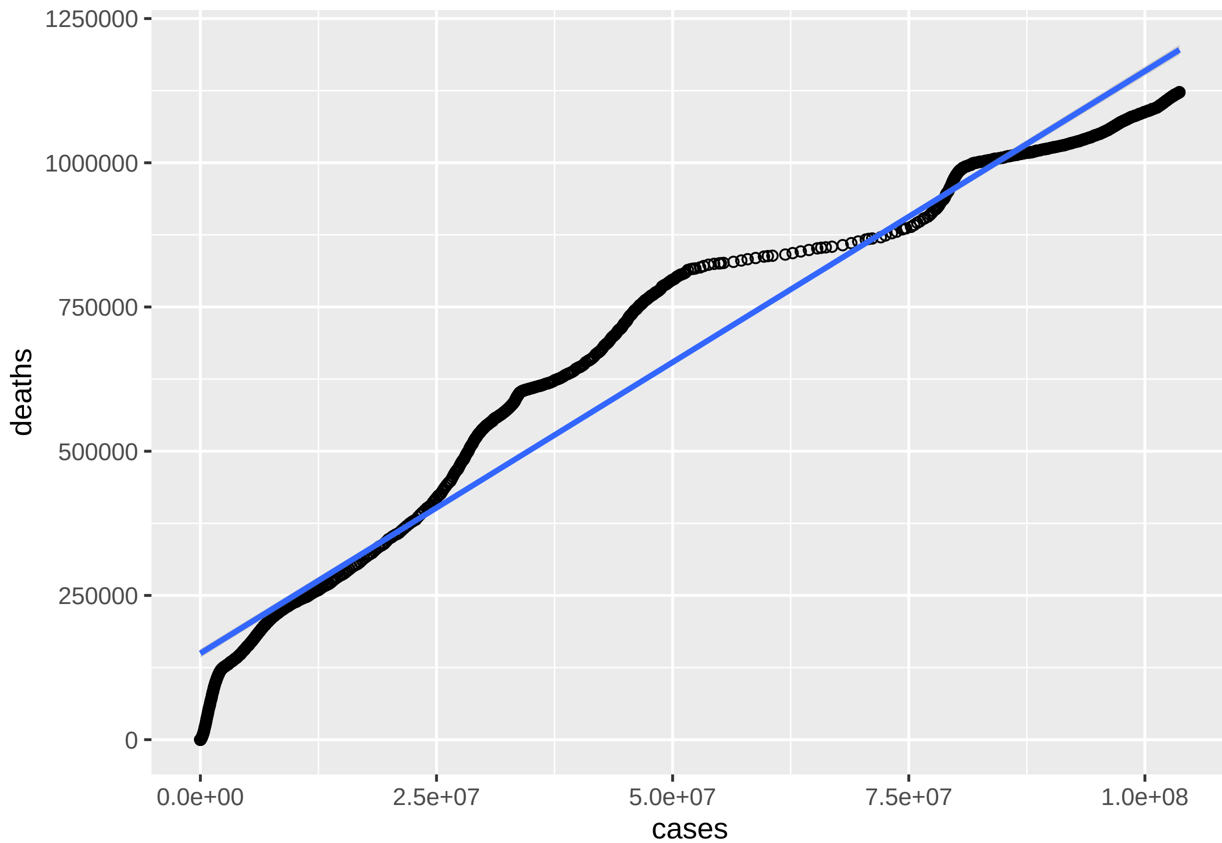
Visualizing cases per thousand and deaths per thousand

```
ggplot(us.state.totals, aes(x=cases_per_thou, y=deaths_per_thou)) +
  geom_point(size=1.5, shape=1)+
  geom_smooth(method='loess', formula= y~x)
```



While it may seem that the relationship is not linear we have to ask why that might be? As stated before it is possible that as time passes, and the vaccines and treatments are rolled out, that people are less likely to report a positive test, especially given the abundance of at home test.

```
ggplot(us.totals, aes(x=cases, y=deaths)) +
  geom_point(size=1.5, shape=1) +
  geom_smooth(method='lm', formula= y~x)
```



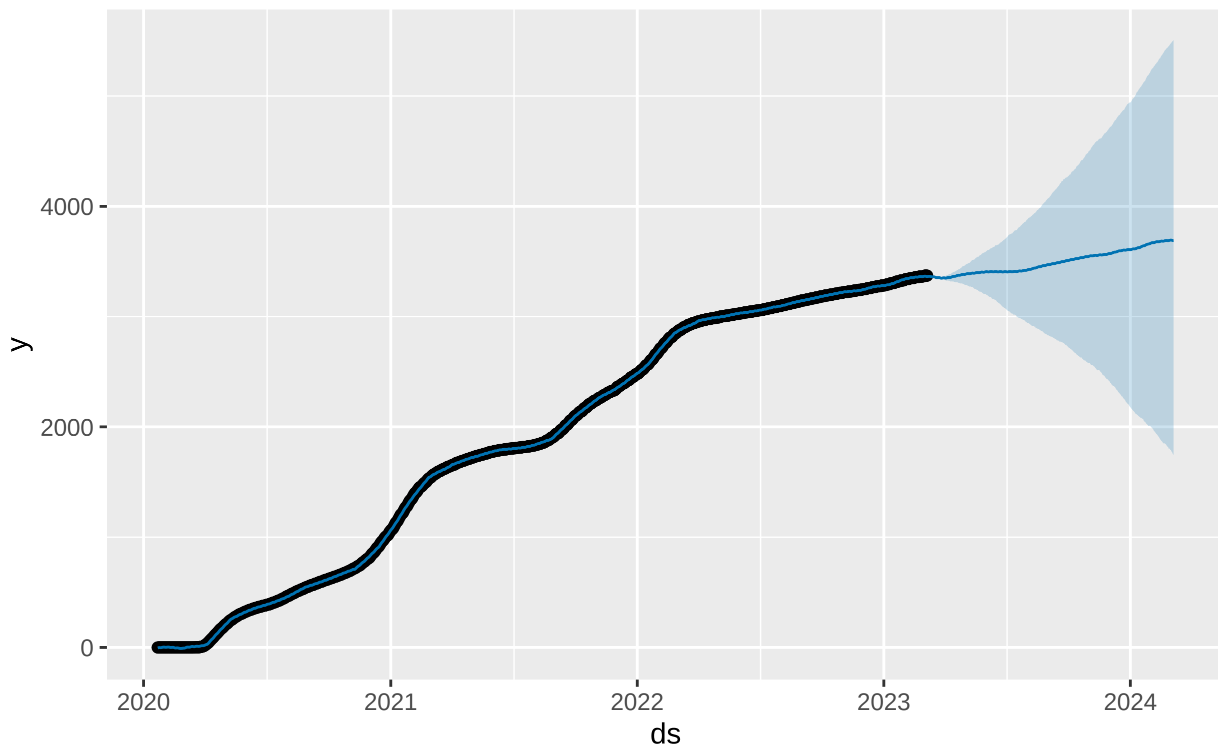
Creating a Model to Predict Deaths per Million in the USA

using an open source library called prophet which generates a model used to predict future outcomes.

```
# format data for use in the prophet function
us.totals.m <- us.totals %>%
  select(date, deaths_per_mill)%>%
  rename(
    ds = date,
    y = deaths_per_mill
    )
```

```
library(prophet)
mod <- prophet(us.totals.m, yearly.seasonality = TRUE, daily.seasonality = TRUE)
```

```
pred <- make_future_dataframe(mod,periods = 365)
fc<-predict(mod,pred)
plot(mod,fc)
```



This model seems to predict a rise in deaths per million in 2024 as compared to previous years. However, this is not accounting for better treatments, and more effective vaccines / booster.

Bias Identification

COVID19 was a very hot political topic in the USA and as a US citizen I certainly had some bias regarding this topic. I had initially thought states like California, Oregon, New York, and Washington would have the worst performance when it comes to controlling the spread of the virus and the death counts. Looking at raw state deaths and cases initially confirmed by bias, however a further analysis using maps and per capita calculations showed that in fact those mentioned states were not the worst performing.

Another potential source of bias is in the data capturing process. With the potential for less self reporting as time goes on, the data that is reported might only be of those whos cases were severe enough to be hospitalized or see a doctor. As we know, a very large portion of folks who are vaccinated and otherwise healthy will most likely not reach this state and therefore their case might not be captured. It is hard to tell exactly how many cases this would be but it is likely enough to have an effect on conclusions made based on our calculations.

Conclusion

Our analysis showed that new cases and deaths seem to have platued across the USA. Comparing this to California we can see a similar trend. We also were able to determine that while TX, CA, and FL were the highest in raw deaths and cases, TX and CA specifically were actually lower when it came to cases and deaths per capita when compared to the majority of US states.

We were able to show using a linear model that cases per thousand was statistically significant when predicting deaths per thousand.

Our 2nd model predicted a potential increase in deaths per million in the USA in 2024 when compared to 2023. This is with the caveat that this cannot factor in the potential for more effective treatments and vaccines/boosters.