DSaaFFinalPt2

Kane

2025-12-09

Statement of Question/Interest

Questions: How has the number of CoVID cases changed over time? How has the deaths due to CoVID changed over time?

Description of Data

Source: Center for Systems Science and Engineering (CSSE) at Johns Hopkins University

Description: This data is the number of cases and deaths of CoVID from Jan 22, 2020 to March 10, 2023. This data includes the Country, Sate/Province, and the Lat/Lon of where the data is from. There is a subset of the Data that also includes the data broken up by reporting districts from the USA.

```
library(tidyverse)
library(zoo)
library(lubridate)
library(plyr)
library(ggplot2)
library(scales)
library(mgcv)
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi
file_names <- 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, file_names)

global_cases <- read_csv(urls[2])
global_deaths <- read_csv(urls[4])
us_cases <- read_csv(urls[1])
us_deaths <- read_csv(urls[3])
```

Global Cases and Global Case Change

```
head(global_cases)
```

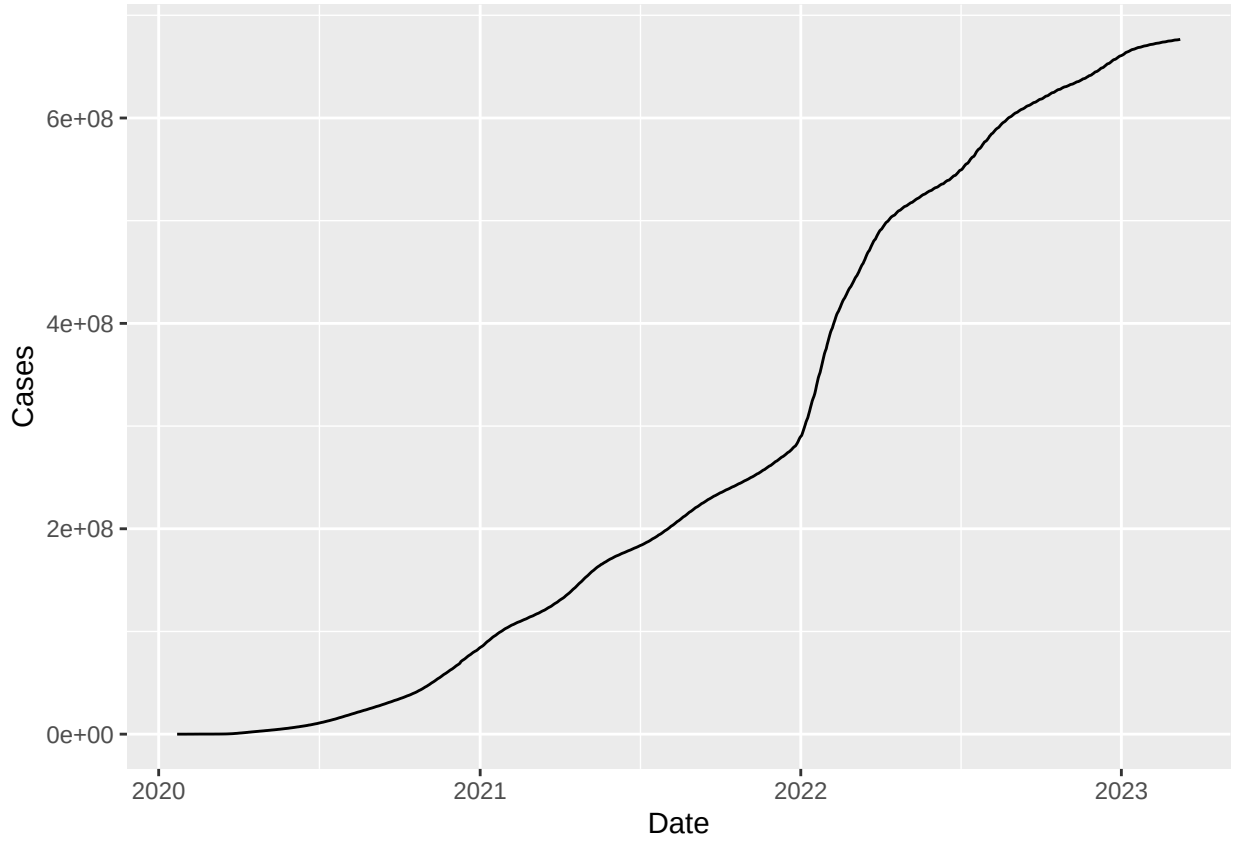
```
## # A tibble: 6 x 1,147
##   `Province/State` `Country/Region`   Lat  Long `1/22/20` `1/23/20` `1/24/20`
##   <chr>            <chr>            <dbl> <dbl>     <dbl>     <dbl>     <dbl>
```

```
## 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>             Antarctica       -71.9 23.3          0         0         0
## # i 1,140 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>, ...
```

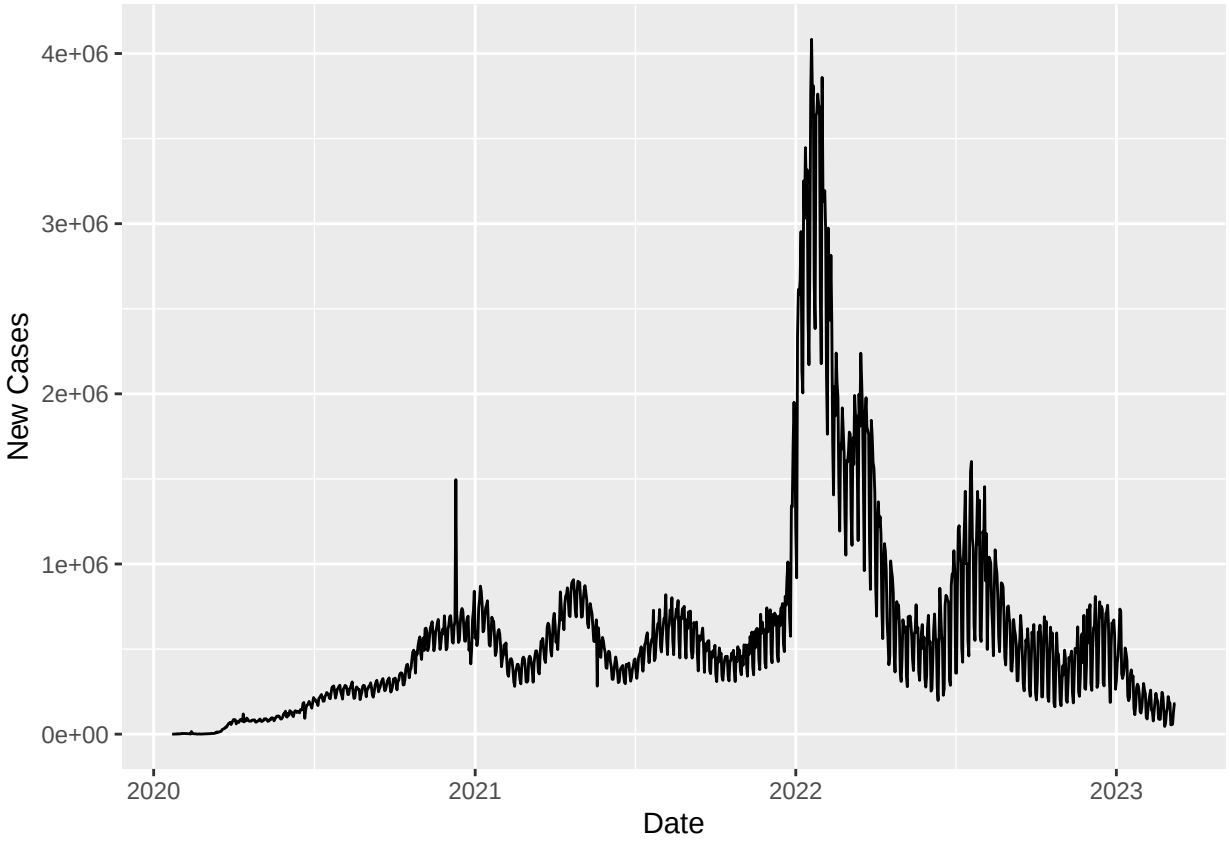
```
global <- as.data.frame(global_cases) %>%
    subset(select = c(-`Province/State`, -Lat, -Long)) %>%
    aggregate(. ~ `Country/Region`, FUN = sum) %>%
    pivot_longer(-`Country/Region`, names_to = "Date", values_to = "Cases") %>%
    transform(Date = as.Date(Date, format = "%m/%d/%y")) %>%
    subset(select = c(-`Country/Region`)) %>%
    aggregate(Cases ~ Date, FUN = sum)

newCases = c(0, global$Cases[2:nrow(global)] - global$Cases[1:(nrow(global) -
    1)])

global <- global %>%
    mutate(`New Cases` = newCases)

head(global)
```

```
##         Date Cases New Cases
## 1 2020-01-22   557         0
## 2 2020-01-23   657       100
## 3 2020-01-24   944       287
## 4 2020-01-25  1437       493
## 5 2020-01-26  2120       683
## 6 2020-01-27  2929       809
```

```
ggplot(aes(y = Cases, x = Date), data = global) + geom_line()
```




```
ggplot(aes(y = `New Cases`, x = Date), data = global) + geom_line()
```



There is an interesting trend where the number of cases never decreases, this must be because this data is total number of confirmed cases and not number of current cases. So I adjusted the metric to be the derivative of total cases to the number of new cases where each data point is the difference between the previous and the current. It is interesting how there is a massive increase in the number of cases when 2020 hit, 4 milion new cases a day, and then it fell back down to a more consistent number of arround 500 thousand.

Global Deaths and Global Death Change

```
head(global_deaths)
```

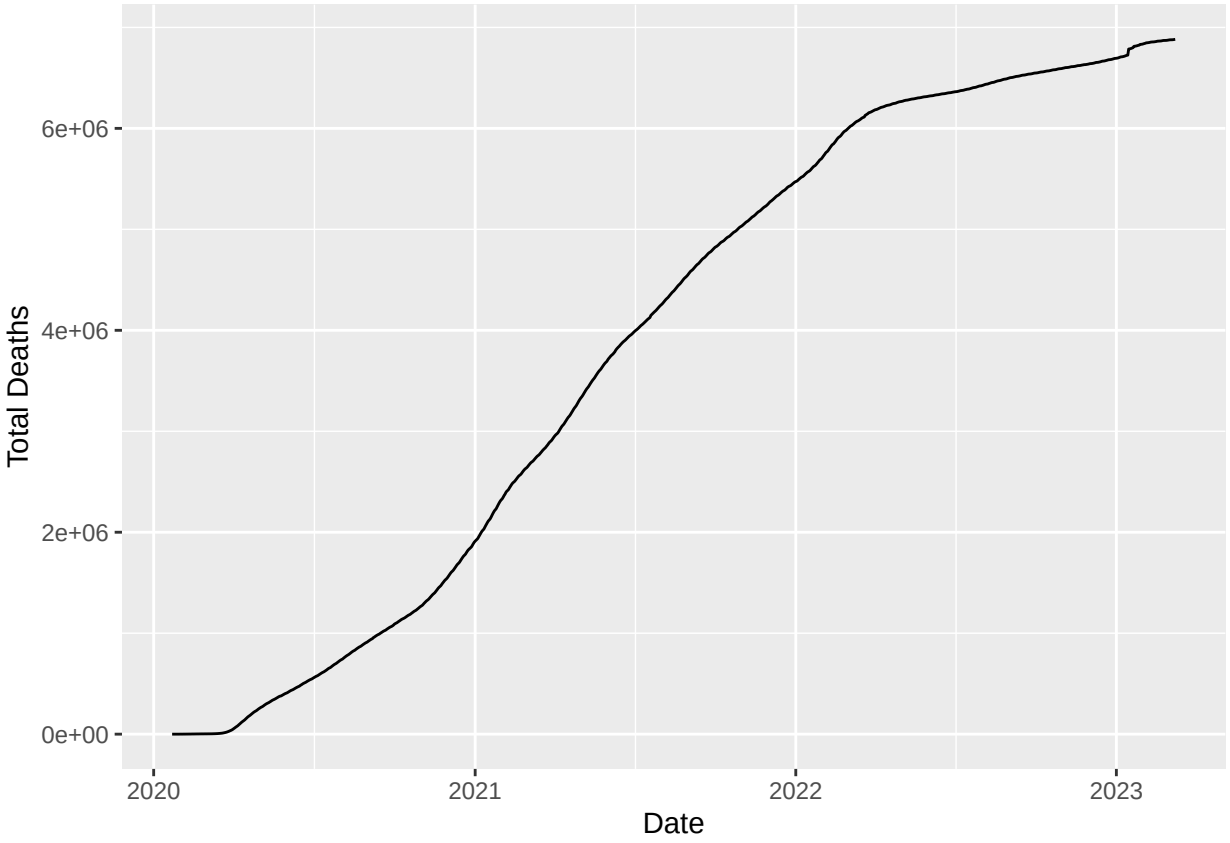
```
## # A tibble: 6 x 1,147
##   `Province/State` `Country/Region`   Lat  Long `1/22/20` `1/23/20` `1/24/20`
##   <chr>            <chr>            <dbl> <dbl>     <dbl>     <dbl>     <dbl>
## 1 <NA>             Afghanistan       33.9 67.7          0         0         0
## 2 <NA>             Albania           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>             Antarctica       -71.9 23.3          0         0         0
## # i 1,140 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>, ...
```

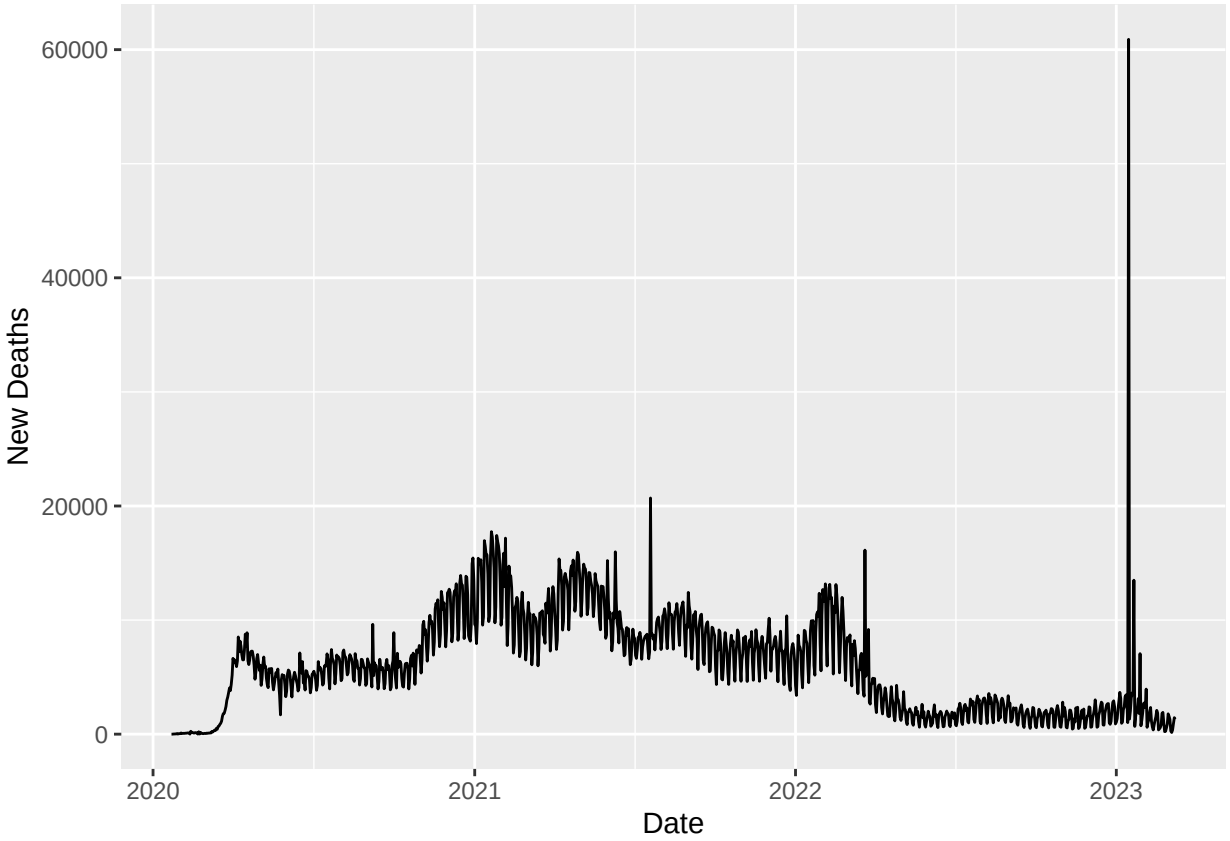
```
deaths <- as.data.frame(global_deaths) %>%
    subset(select = c(-`Province/State`, -Lat, -Long)) %>%
    aggregate(. ~ `Country/Region`, FUN = sum) %>%
    pivot_longer(-`Country/Region`, names_to = "Date", values_to = "Total Deaths") %>%
    transform(Date = as.Date(Date, format = "%m/%d/%y")) %>%
    subset(select = c(-`Country/Region`)) %>%
    aggregate(`Total Deaths` ~ Date, FUN = sum)

newDeaths = c(0, deaths$`Total Deaths`[2:nrow(deaths)] - deaths$`Total Deaths`[1:(nrow(deaths) -
    1)])

deaths <- deaths %>%
    mutate(`New Deaths` = newDeaths)

head(deaths)
```

```
##         Date Total Deaths New Deaths
## 1 2020-01-22           17          0
## 2 2020-01-23           18          1
## 3 2020-01-24           26          8
## 4 2020-01-25           42         16
## 5 2020-01-26           56         14
## 6 2020-01-27           82         26
```

```
ggplot(aes(y = `Total Deaths`, x = Date), data = deaths) + geom_line()
```




```
ggplot(aes(y = `New Deaths`, x = Date), data = deaths) + geom_line()
```



Like with the number of cases, the number of deaths appears to be total so I adjusted it to the change in deaths. There appears to be an outlier right at the start of 2023 I can only assume that that is a case of reporting lag. But the number of new deaths was relatively constant, hovering around 10 thousand deaths. After 2022 the number of new deaths dropped to around 1 thousand new deaths.

Global Death Rate

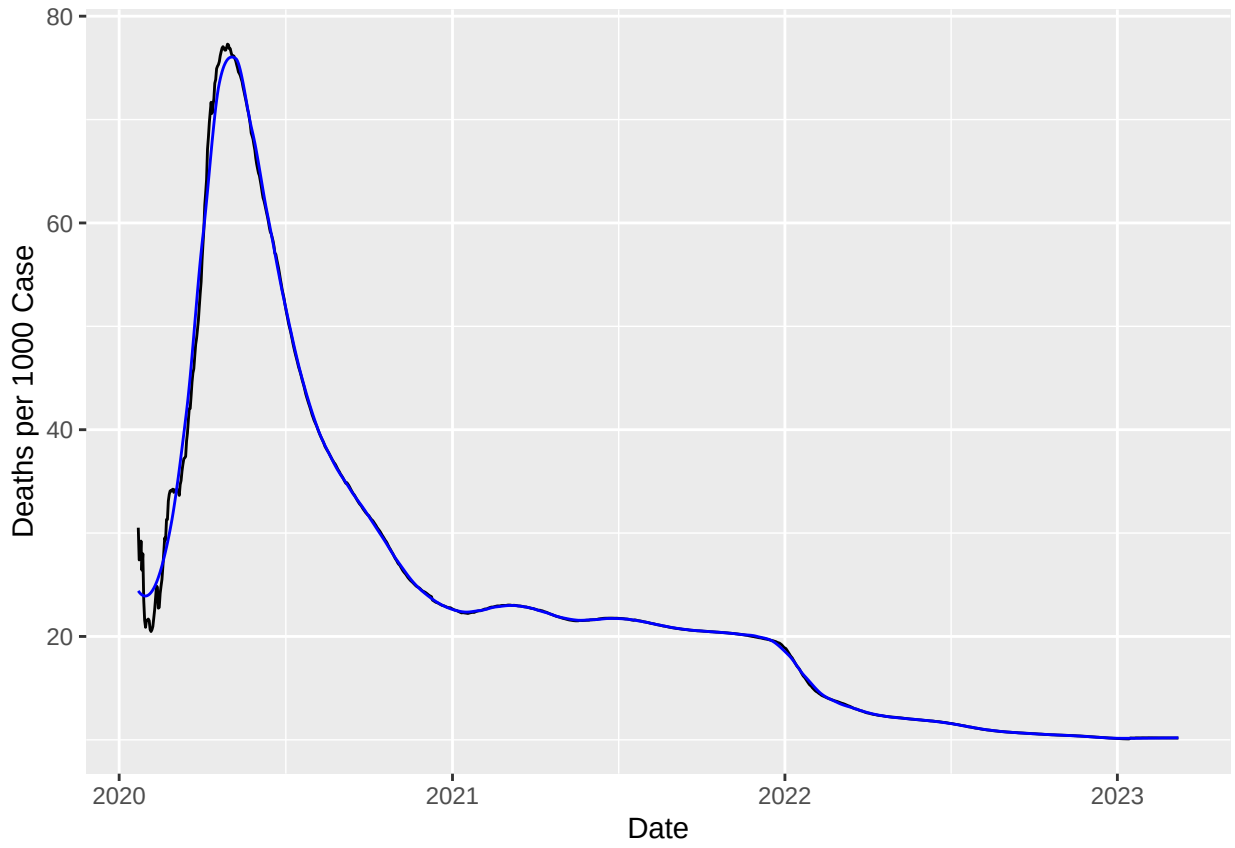
```
deathRate = deaths$`Total Deaths`/(global$Cases/1000)
deaths <- deaths %>%
    mutate(DeathRate = deathRate)

head(deaths)
```

```
##         Date Total Deaths New Deaths DeathRate
## 1 2020-01-22           17          0  30.52065
## 2 2020-01-23           18          1  27.39726
## 3 2020-01-24           26          8  27.54237
## 4 2020-01-25           42         16  29.22756
## 5 2020-01-26           56         14  26.41509
## 6 2020-01-27           82         26  27.99590
```

```
deathrate.model = loess(DeathRate ~ as.numeric(Date), data = deaths,
    span = 0.1)
summary(deathrate.model)
```

```
## Call:
## loess(formula = DeathRate ~ as.numeric(Date), data = deaths,
##     span = 0.1)
##
## Number of Observations: 1143
## Equivalent Number of Parameters: 28.69
## Residual Standard Error: 0.8361
## Trace of smoother matrix: 31.72  (exact)
##
## Control settings:
##   span     :  0.1
##   degree   :  2
##   family   :  gaussian
##   surface  :  interpolate      cell = 0.2
##   normalize:  TRUE
##  parametric:  FALSE
## drop.square:  FALSE
```

```
ggplot(aes(y = DeathRate, x = Date), data = deaths) + geom_line() +
    labs(y = "Deaths per 1000 Case") + geom_line(aes(y = deathrate.model$fitted),
    color = "blue")  #+ geom_smooth(formula=y~x, method='loess', span=0.1)
```



The ratio of deaths to new cases is indicative of how fast a disease is killing / how deadly it is. It can be seen that towards the start of the data that CoVID was more deadly killing around 75 people per 1000 cases and then it dropped to a steady 20 people per 1000 cases and then in 2022 it dropped all the way to 1 or $<$ 1 people per 1000 cases. The vaccine was released in 2021 and may indicate why there the data leveld out.

Conclusion

The number of new CoVID cases has gone though ups and downs with steady increase in 2020-2021 then in 2021-2022 the vaccine was released and the number of new cases flucuated regularly then in 2022 something happened, maybe Chinese censorship, maybe the lift of lockdown, and the number of new cases significantly increased. Then came back down to a steady state of around 500 thousand new cases globaly.

The number of deaths due to CoVID has not changed much over time, durning the beginning there was a low number of deaths then the number came up to around 5 thousand deaths, then the number gradually moved up to 10 thousand, then in 2022 the number of new deaths stated falling to around 1 thousand. In 2023 there is an anomoly which is probably caused by a sudden reporting of data all at once.

The death rate of covid has come down over time and leveld out once the vaccine was distributed in 2021 and has since dropped in 2022 down to very low < 1 death per 1000 cases.

Bias

This is a pos-hoc analysis and as such all there is an increase in type I error. This was not data I was particularly interested in and as such I may be looking too little and causing TYPE II error. I was expecting to see a decrease in the severity of the data over time.