Covid19

2023-02-28

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.1      v purrr   1.0.0
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## 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
```

```
url_in<-"https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid
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",
  "time_series_covid19_recovered_global.csv"
)
urls<-str_c(url_in,filenames)
```

```
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.
```

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

```
## Rows: 289 Columns: 1144
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1142): 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.
```

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

```
## 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.
```

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

```
## Rows: 289 Columns: 1144
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1142): 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.
```

```
g_rec<-read_csv(urls[5])
```

```
## Rows: 274 Columns: 1144
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1142): 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.
```

```
g_cases<-g_cases %>%
  pivot_longer(cols=-c('Province/State','Country/Region', Lat, Long),
               names_to = "date",values_to = "cases")%>%
  select(-c(Lat,Long))

g_death<-g_death %>%
```

```
  pivot_longer(cols=-c('Province/State','Country/Region', Lat, Long),
               names_to = "date",values_to = "deaths")%>%
  select(-c(Lat,Long))

global <- g_cases %>%
  full_join(g_death) %>%
  rename(Country_Region='Country/Region',
         Province_State='Province/State')%>%
  mutate(date=mdy(date))
```

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

```
global <- global %>%filter(cases>0)

summary(global)
```

```
##  Province_State     Country_Region          date                cases
##  Length:305966      Length:305966      Min.   :2020-01-22   Min.   :        1
##  Class :character   Class :character   1st Qu.:2020-12-11   1st Qu.:     1309
##  Mode  :character   Mode  :character   Median :2021-09-14   Median :    20275
##                                        Mean   :2021-09-10   Mean   :  1029137
##                                        3rd Qu.:2022-06-13   3rd Qu.:   270739
##                                        Max.   :2023-03-06   Max.   :103655657
##      deaths
##  Min.   :      0
##  1st Qu.:      7
##  Median :    213
##  Mean   :  14378
##  3rd Qu.:   3649
##  Max.   :1122264
```

```
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_))


us_death <-us_death %>%
  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_death)
```

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

```
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))
```

```
## 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.
```

```
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_by_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()
```

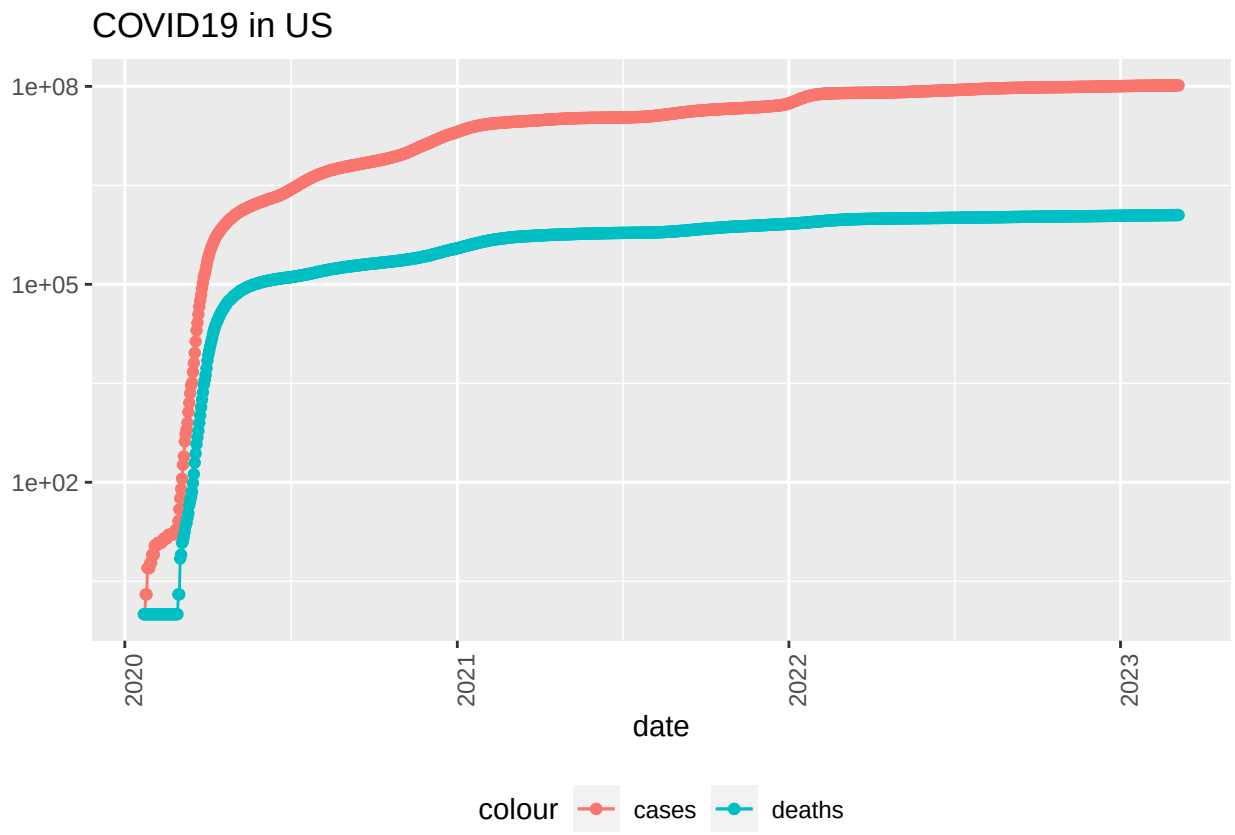
```
## `summarise()` has grouped output by 'Province_State', 'Country_Region'. You can
## override using the `.groups` argument.
```

```
us_totals <- us_by_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()
```

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

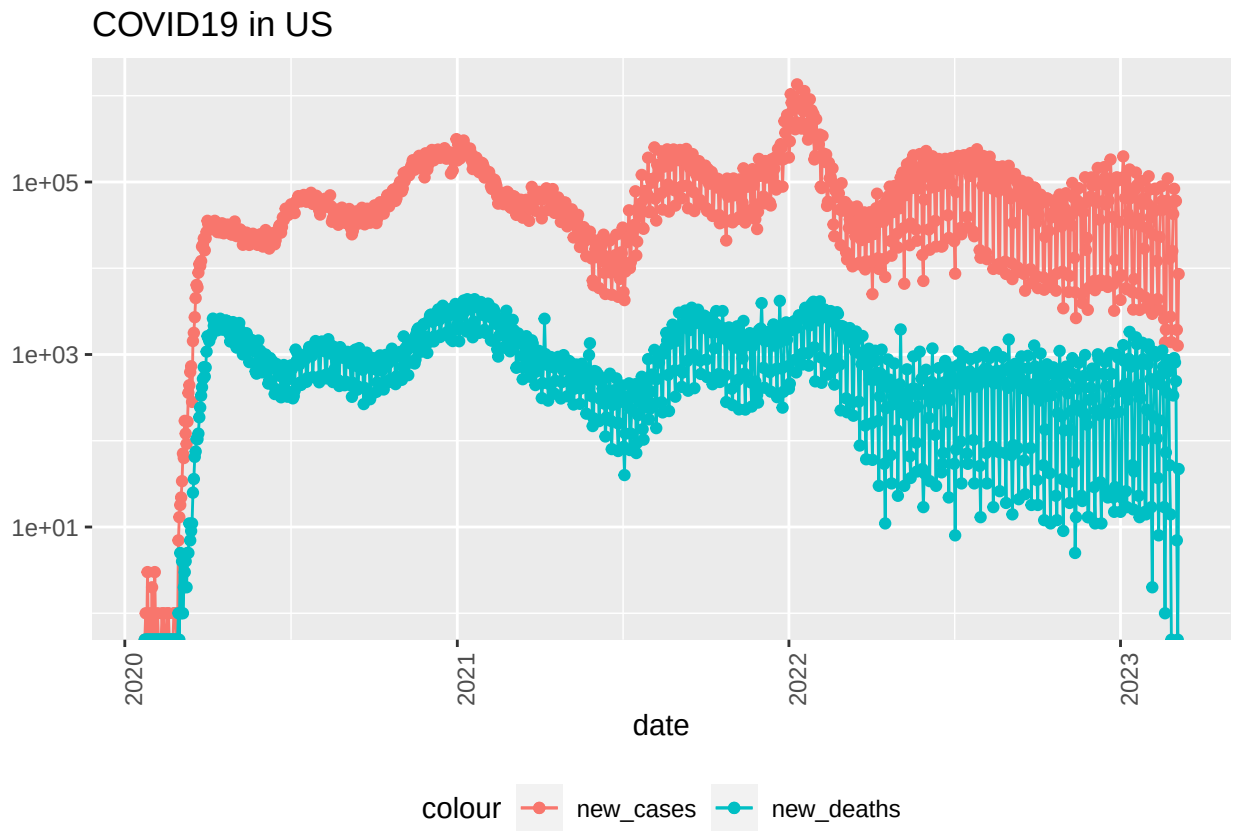
```
us_totals%>%
  ggplot(aes(x=date,y=cases))+
  geom_line(aes(color="cases"))+
  geom_point(aes(color="cases"))+
```

```
  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)
```




```
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))

us_totals%>%
  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 = "COVID19 in US", y=NULL)
```

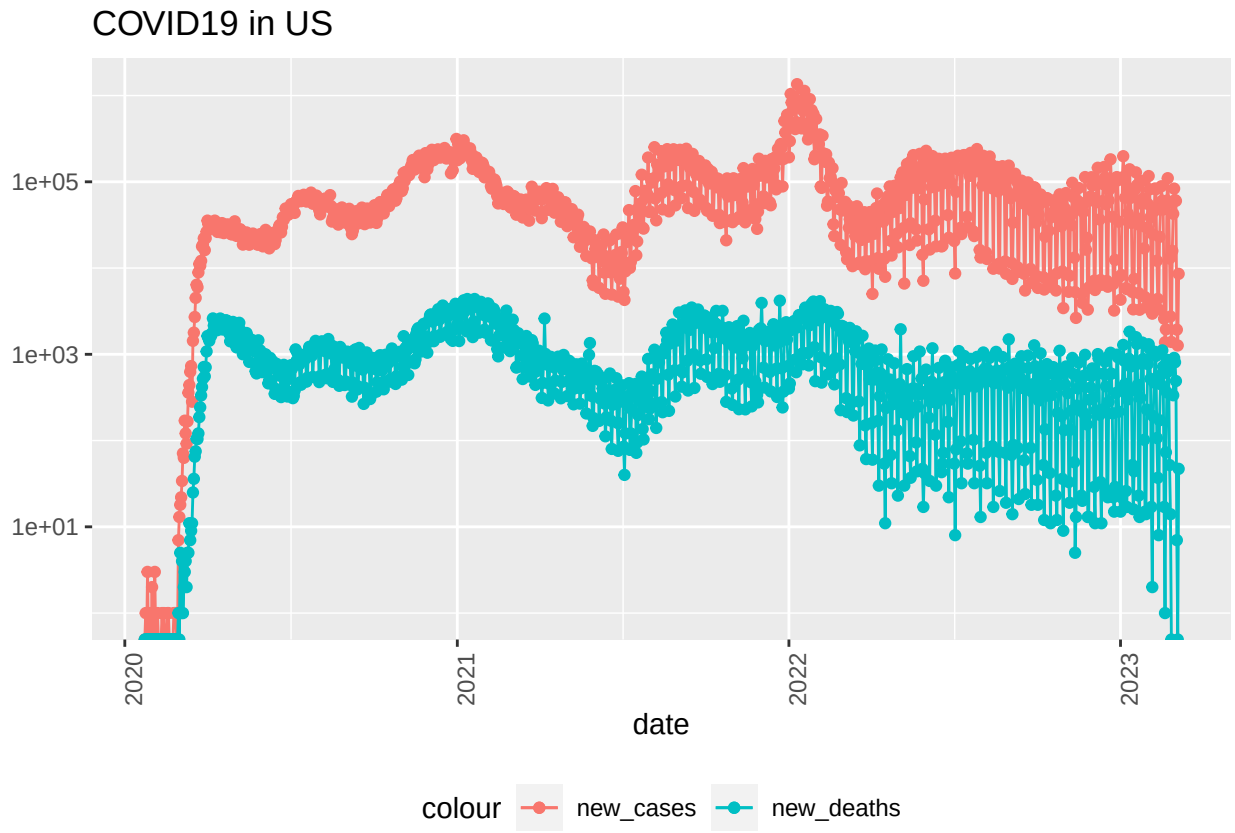
```
## Warning: Transformation introduced infinite values in continuous y-axis
## Transformation introduced infinite values in continuous y-axis

## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 1 row containing missing values (`geom_line()`).

## Warning: Removed 1 rows containing missing values (`geom_point()`).

## Warning: Removed 1 row containing missing values (`geom_line()`).

## Warning: Removed 3 rows containing missing values (`geom_point()`).
```




```
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))

us_totals%>%
    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 = "COVID19 in US", y=NULL)
```

```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 1 row containing missing values (`geom_line()`).

## Warning: Removed 1 rows containing missing values (`geom_point()`).

## Warning: Removed 1 row containing missing values (`geom_line()`).

## Warning: Removed 3 rows containing missing values (`geom_point()`).
```




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

us_state_totals%>%
  slice_min(deaths_per_thou,n=10)%>%
  select(Province_State,deaths_per_thou,cases_per_thou,everything())
```

```
## # A tibble: 10 x 6
##    Province_State           deaths_per_thou cases_per_thou deaths  cases popul~1
##    <chr>                              <dbl>          <dbl>  <dbl>  <dbl>   <dbl>
##  1 American Samoa                     0.611           150.     34 8.32e3   55641
##  2 Northern Mariana Islands           0.744           248.     41 1.37e4   55144
##  3 Virgin Islands                     1.21            231.    130 2.48e4  107268
##  4 Hawaii                             1.30            268.   1834 3.80e5 1415872
##  5 Vermont                            1.46            243.    910 1.51e5  623989
##  6 Puerto Rico                        1.55            293.   5810 1.10e6 3754939
##  7 Utah                               1.65            340.   5287 1.09e6 3205958
##  8 Alaska                             2.01            414.   1486 3.07e5  740995
##  9 District of Columbia               2.03            252.   1430 1.78e5  705749
## 10 Washington                         2.06            253.  15683 1.93e6 7614893
## # ... with abbreviated variable name 1: population
```

```
us_state_totals%>%
  slice_max(deaths_per_thou,n=10)%>%
 select(Province_State,deaths_per_thou,cases_per_thou,everything())
```

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

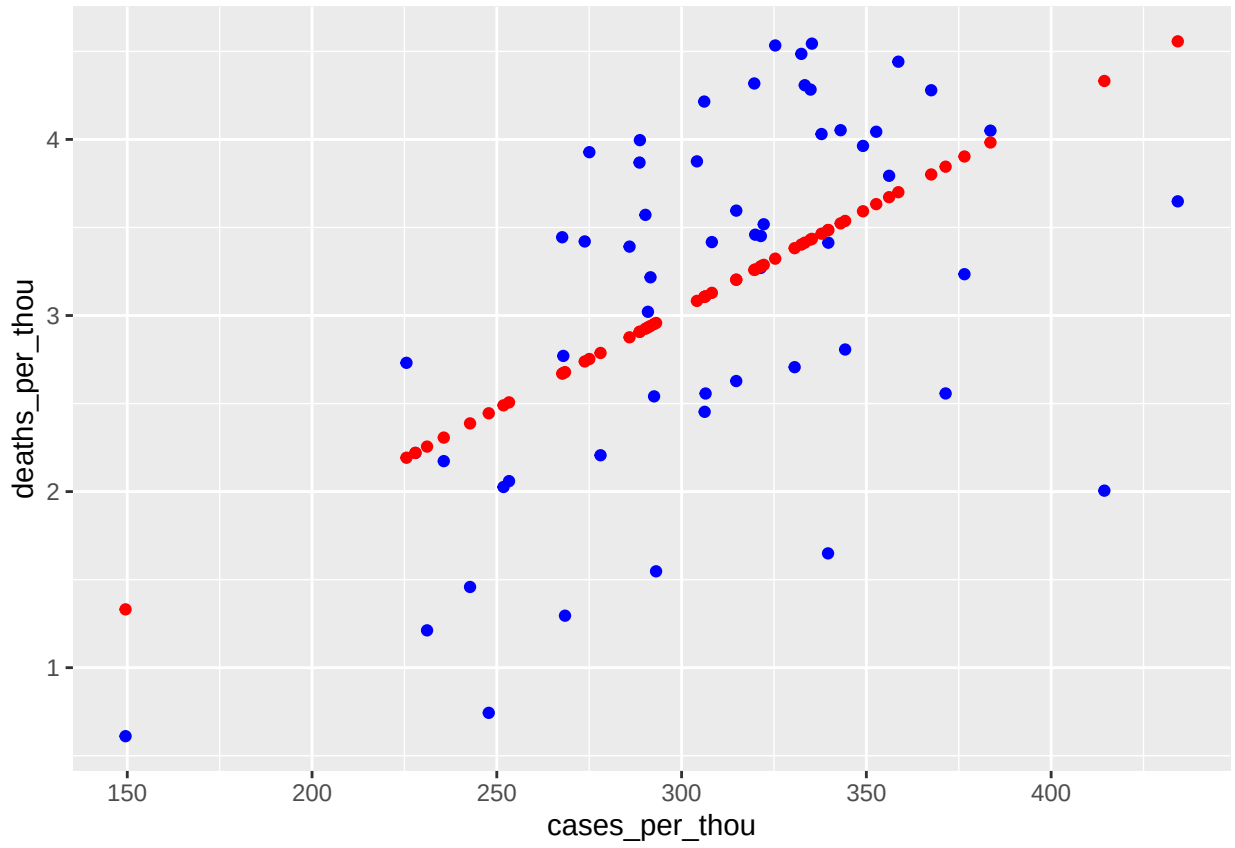
```
mod<-lm(deaths_per_thou ~ cases_per_thou,data=us_state_totals)
summary(mod)
```

```
##
## 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
```

```
us_state_totals %>% slice_max(cases_per_thou)
```

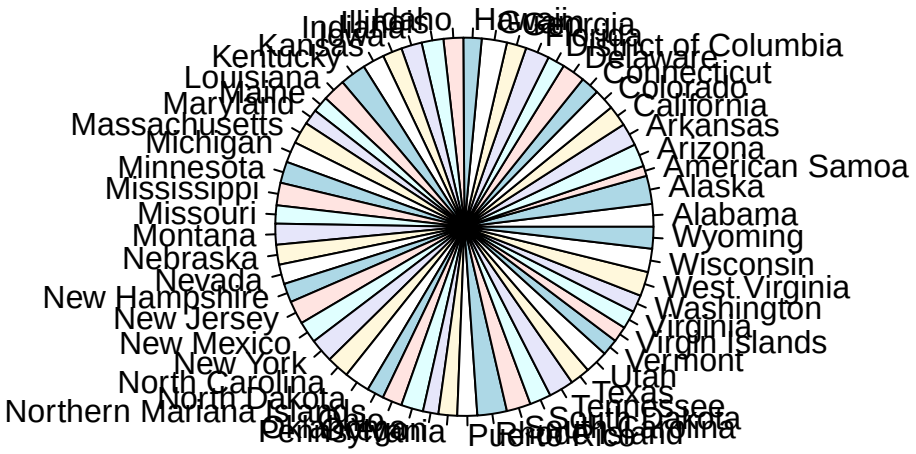
```
## # A tibble: 1 x 6
##   Province_State deaths  cases population cases_per_thou deaths_per_thou
##   <chr>           <dbl>  <dbl>      <dbl>          <dbl>           <dbl>
## 1 Rhode Island     3865 460045    1059361           434.            3.65
```

```
x_grid<-seq(1,450)
new_df<-tibble(cases_per_thou=x_grid)
us_state_totals_pred<-us_state_totals%>%mutate(pred=predict(mod))
us_state_totals_pred%>% ggplot()+
  geom_point(aes(x=cases_per_thou, y=deaths_per_thou), color="blue")+
  geom_point(aes(x=cases_per_thou, y=pred),color="red")
```

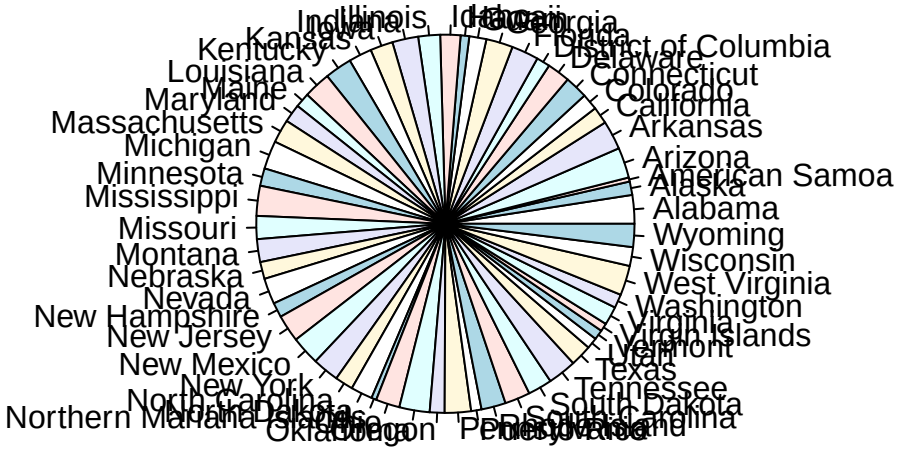


This are my additions:

```
pie(us_state_totals$cases_per_thou, labels = us_state_totals$Province_State,
    main = "Cases per thousands by state")
```



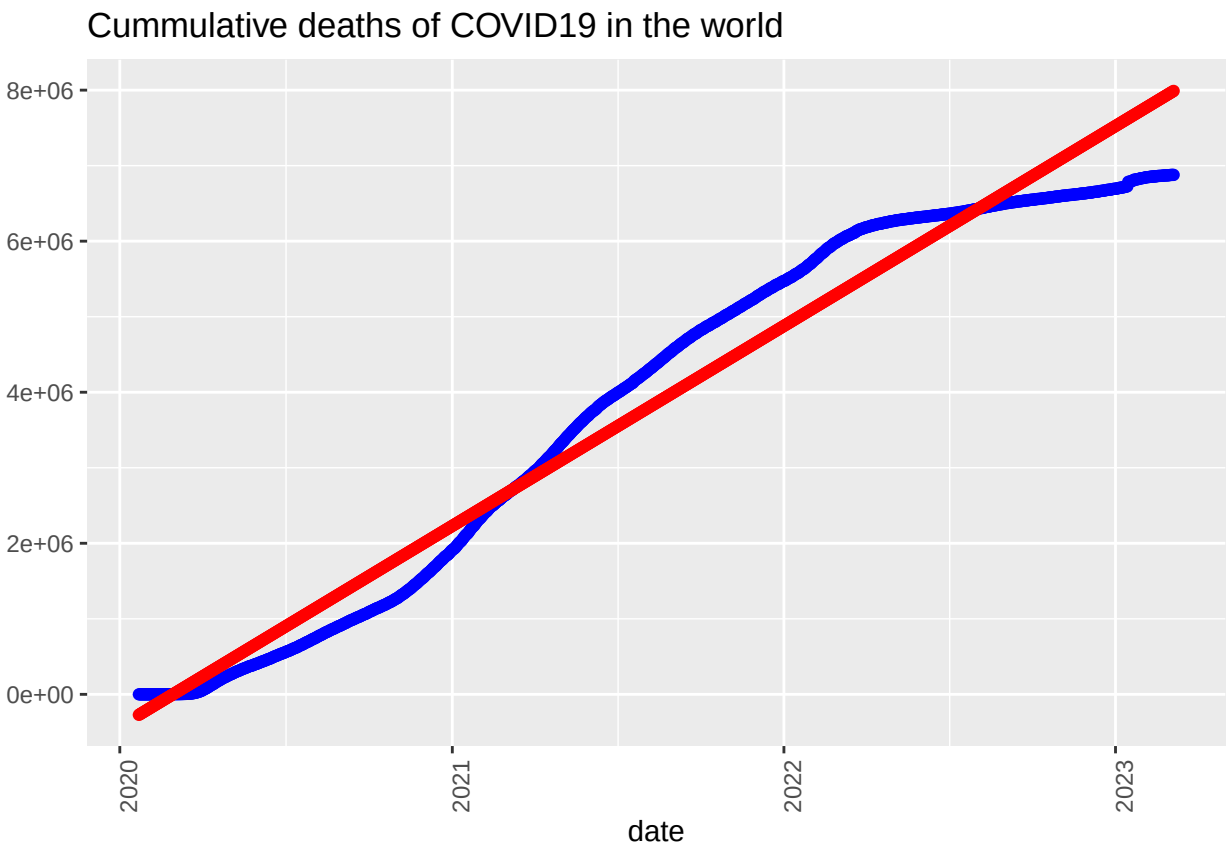

```
pie(us_state_totals$deaths_per_thou, labels = us_state_totals$Province_State,
    main = "Deaths per thousands by state")
```




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

global_totals<-global%>%
 group_by(date)%>%
 summarize(deaths=sum(deaths),cases=sum(cases))
```

I know because of the news and because of my understand of how viruses propagate that I was bias towards the belief that small states with many people, such us New York, would be leading the charge in terms of cases. However I very surprise to see that states that have very little population and are relatively big like Wyoming and Alaska have proportionately to their population so many cases. I must also note that after seen the number of cases I am not surprised to see Wyoming also has a high death count because it is my belief and my bias that young people tend to leave that state. This is also probes to me that the measures taken by states like New York were propagation is more likely were quite good.

```
mod2<-lm(deaths ~ date,data=global_totals)
global_pred<-global_totals%>%mutate(pred=predict(mod2))
global_pred%>% ggplot()+
  geom_point(aes(x=date, y=deaths), color="blue")+
  geom_point(aes(x=date, y=pred),color="red")+
  theme(legend.position = "bottom",
          axis.text.x = element_text(angle=90))+
    labs(title = "Cummulative deaths of COVID19 in the world", y=NULL)
```



Regarding the global data I'm just happy that the number of deaths is barely growing, I was biased towards that belief but the graph seems to show undeniably that.