COVID_19 Dataset

Shweta Govind

8/6/2021

#Import the dataset

```
library(stringr)
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi
file_names <-
c("time_series_covid19_confirmed_global.csv",

"time_series_covid19_deaths_global.csv",

"time_series_covid19_confirmed_US.csv",

"time_series_covid19_deaths_US.csv")
urls <- str_c(url_in, file_names)
```

#Read in the data

```
global_cases <- read_csv(urls[1])
global_deaths <- read_csv(urls[2])
US_cases <- read_csv(urls[3])
US_deaths <- read_csv(urls[4])
```

```
global_cases <- global_cases %>%
    pivot_longer(cols =
-c(`Province/State`,

`Country/Region`, Lat, Long),
                 names_to = "date",
               values_to = "cases") %>%

  select(-c(Lat,Long))
  global_cases
```

```
## # A tibble: 157,635 x 4
##    `Province/State` `Country/Region` date    cases
##    <chr>            <chr>            <chr>   <dbl>
##  1 <NA>             Afghanistan      1/22/20     0
##  2 <NA>             Afghanistan      1/23/20     0
##  3 <NA>             Afghanistan      1/24/20     0
##  4 <NA>             Afghanistan      1/25/20     0
##  5 <NA>             Afghanistan      1/26/20     0
##  6 <NA>             Afghanistan      1/27/20     0
##  7 <NA>             Afghanistan      1/28/20     0
##  8 <NA>             Afghanistan      1/29/20     0
##  9 <NA>             Afghanistan      1/30/20     0
## 10 <NA>             Afghanistan      1/31/20     0
## # ... with 157,625 more rows
```

```
global_deaths <- global_deaths %>%
    pivot_longer(cols =
-c(`Province/State`,

`Country/Region`, Lat, Long),
                 names_to = "date",
               values_to = "deaths") %>%

select(-c(Lat,Long))
global_deaths
```

```
## # A tibble: 157,635 x 4
##    `Province/State` `Country/Region` date    deaths
##    <chr>            <chr>            <chr>    <dbl>
##  1 <NA>             Afghanistan      1/22/20      0
##  2 <NA>             Afghanistan      1/23/20      0
##  3 <NA>             Afghanistan      1/24/20      0
##  4 <NA>             Afghanistan      1/25/20      0
##  5 <NA>             Afghanistan      1/26/20      0
```

```
##  6 <NA>             Afghanistan      1/27/20      0
##  7 <NA>             Afghanistan      1/28/20      0
##  8 <NA>             Afghanistan      1/29/20      0
##  9 <NA>             Afghanistan      1/30/20      0
## 10 <NA>             Afghanistan      1/31/20      0
## # ... with 157,625 more rows
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

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

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

```
global
```

```
## # A tibble: 157,635 x 5
##    Province_State Country_Region date       cases deaths
##    <chr>          <chr>          <date>     <dbl>  <dbl>
##  1 <NA>           Afghanistan    2020-01-22     0      0
##  2 <NA>           Afghanistan    2020-01-23     0      0
##  3 <NA>           Afghanistan    2020-01-24     0      0
##  4 <NA>           Afghanistan    2020-01-25     0      0
##  5 <NA>           Afghanistan    2020-01-26     0      0
##  6 <NA>           Afghanistan    2020-01-27     0      0
##  7 <NA>           Afghanistan    2020-01-28     0      0
##  8 <NA>           Afghanistan    2020-01-29     0      0
##  9 <NA>           Afghanistan    2020-01-30     0      0
## 10 <NA>           Afghanistan    2020-01-31     0      0
## # ... with 157,625 more rows
```

```
global %>% filter(cases > 35000000)
```

```
## # A tibble: 8 x 5
##   Province_State Country_Region date          cases deaths
##   <chr>          <chr>          <date>        <dbl>  <dbl>
## 1 <NA>           US             2021-08-01 35003417 613228
## 2 <NA>           US             2021-08-02 35131393 613679
## 3 <NA>           US             2021-08-03 35237950 614295
## 4 <NA>           US             2021-08-04 35330664 614785
## 5 <NA>           US             2021-08-05 35440488 615320
## 6 <NA>           US             2021-08-06 35695469 616493
## 7 <NA>           US             2021-08-07 35739551 616718
## 8 <NA>           US             2021-08-08 35763414 616829
```

```
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_cases
```

```
## # A tibble: 1,888,230 x 6
##    Admin2  Province_State Country_Region Combined_Key         date       cases
##    <chr>   <chr>          <chr>          <chr>                <date>     <dbl>
##  1 Autauga Alabama        US             Autauga, Alabama, US 2020-01-22     0
##  2 Autauga Alabama        US             Autauga, Alabama, US 2020-01-23     0
##  3 Autauga Alabama        US             Autauga, Alabama, US 2020-01-24     0
##  4 Autauga Alabama        US             Autauga, Alabama, US 2020-01-25     0
##  5 Autauga Alabama        US             Autauga, Alabama, US 2020-01-26     0
##  6 Autauga Alabama        US             Autauga, Alabama, US 2020-01-27     0
##  7 Autauga Alabama        US             Autauga, Alabama, US 2020-01-28     0
##  8 Autauga Alabama        US             Autauga, Alabama, US 2020-01-29     0
##  9 Autauga Alabama        US             Autauga, Alabama, US 2020-01-30     0
## 10 Autauga Alabama        US             Autauga, Alabama, US 2020-01-31     0
## # ... with 1,888,220 more rows
```

```
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_deaths
```

```
## # A tibble: 1,888,230 x 7
##    Admin2  Province_State Country_Region Combined_Key      Population date
##    <chr>   <chr>          <chr>          <chr>                 <dbl> <date>
##  1 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-22
##  2 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-23
##  3 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-24
##  4 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-25
##  5 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-26
##  6 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-27
##  7 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-28
##  8 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-29
##  9 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-30
## 10 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-31
## # ... with 1,888,220 more rows, and 1 more variable: deaths <dbl>
```

```
US <- US_cases %>%
    full_join(US_deaths)
```

```
## 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)
global
```

```
## # A tibble: 157,635 x 6
##    Combined_Key Province_State Country_Region date       cases deaths
##    <chr>        <chr>          <chr>          <date>     <dbl>  <dbl>
##  1 Afghanistan  <NA>           Afghanistan    2020-01-22     0      0
##  2 Afghanistan  <NA>           Afghanistan    2020-01-23     0      0
##  3 Afghanistan  <NA>           Afghanistan    2020-01-24     0      0
##  4 Afghanistan  <NA>           Afghanistan    2020-01-25     0      0
##  5 Afghanistan  <NA>           Afghanistan    2020-01-26     0      0
##  6 Afghanistan  <NA>           Afghanistan    2020-01-27     0      0
##  7 Afghanistan  <NA>           Afghanistan    2020-01-28     0      0
##  8 Afghanistan  <NA>           Afghanistan    2020-01-29     0      0
##  9 Afghanistan  <NA>           Afghanistan    2020-01-30     0      0
## 10 Afghanistan  <NA>           Afghanistan    2020-01-31     0      0
## # ... with 157,625 more rows
```

```
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))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   UID = col_double(),
##   iso2 = col_character(),
##   iso3 = col_character(),
##   code3 = col_double(),
##   FIPS = col_character(),
##   Admin2 = col_character(),
##   Province_State = col_character(),
##   Country_Region = col_character(),
##   Lat = col_double(),
##   Long_ = col_double(),
##   Combined_Key = col_character(),
##   Population = col_double()
## )
```

```
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 )
global
```

```
## # A tibble: 157,635 x 7
##    Province_State Country_Region date       cases deaths Population Combined_Key
```

```
##    <chr>          <chr>          <date>     <dbl>  <dbl>      <dbl> <chr>
##  1 <NA>           Afghanistan    2020-01-22     0      0   38928341 Afghanistan
##  2 <NA>           Afghanistan    2020-01-23     0      0   38928341 Afghanistan
##  3 <NA>           Afghanistan    2020-01-24     0      0   38928341 Afghanistan
##  4 <NA>           Afghanistan    2020-01-25     0      0   38928341 Afghanistan
##  5 <NA>           Afghanistan    2020-01-26     0      0   38928341 Afghanistan
##  6 <NA>           Afghanistan    2020-01-27     0      0   38928341 Afghanistan
##  7 <NA>           Afghanistan    2020-01-28     0      0   38928341 Afghanistan
##  8 <NA>           Afghanistan    2020-01-29     0      0   38928341 Afghanistan
##  9 <NA>           Afghanistan    2020-01-30     0      0   38928341 Afghanistan
## 10 <NA>           Afghanistan    2020-01-31     0      0   38928341 Afghanistan
## # ... with 157,625 more rows
```

#Visualizing the Data

```
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 `
```

```
US_by_state
```

```
## # A tibble: 32,770 x 7
##    Province_State Country_Region date       cases deaths deaths_per_mill
##    <chr>          <chr>          <date>     <dbl>  <dbl>           <dbl>
##  1 Alabama        US             2020-01-22     0      0               0
##  2 Alabama        US             2020-01-23     0      0               0
##  3 Alabama        US             2020-01-24     0      0               0
##  4 Alabama        US             2020-01-25     0      0               0
##  5 Alabama        US             2020-01-26     0      0               0
##  6 Alabama        US             2020-01-27     0      0               0
##  7 Alabama        US             2020-01-28     0      0               0
##  8 Alabama        US             2020-01-29     0      0               0
##  9 Alabama        US             2020-01-30     0      0               0
## 10 Alabama        US             2020-01-31     0      0               0
## # ... with 32,760 more rows, and 1 more variable: Population <dbl>
```

```
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()
```

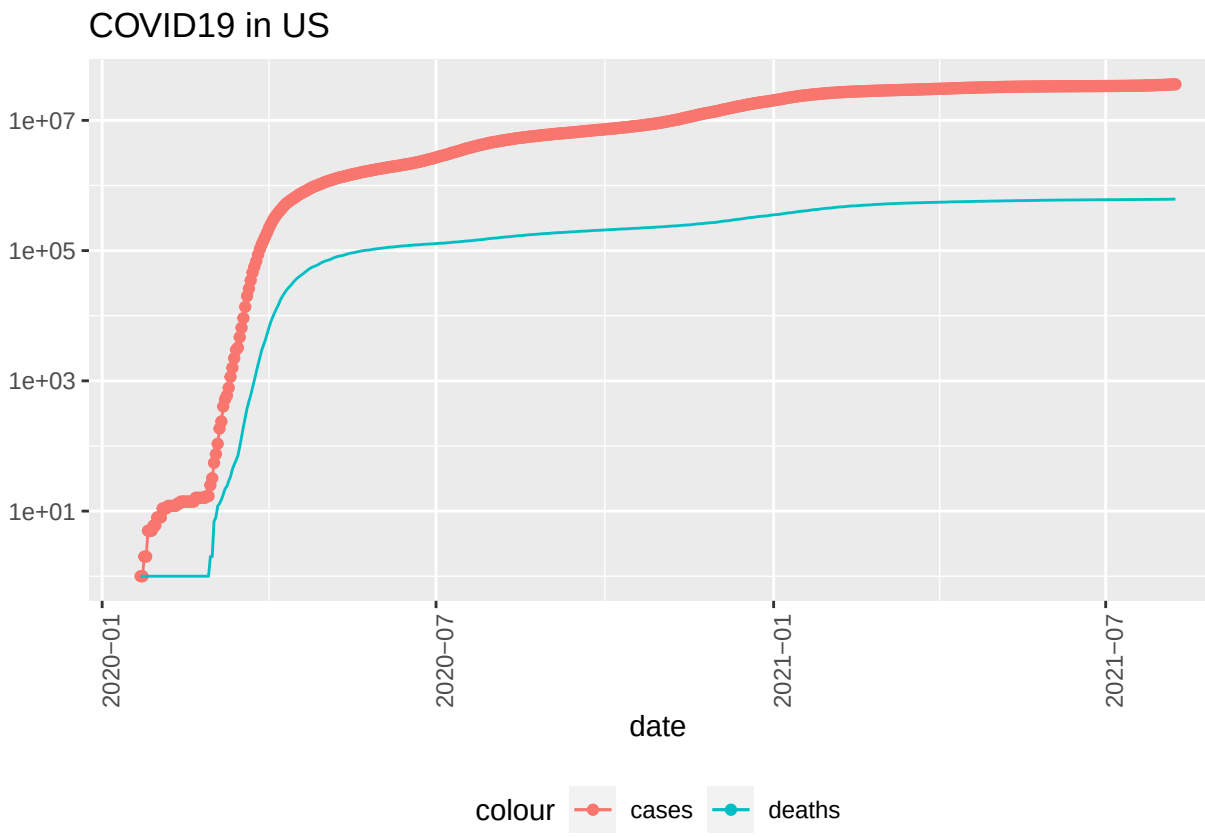
```
## `summarise()` has grouped output by 'Country_Region'. You can override using the `.groups` argument.
```

```
US_totals
```

```
## # A tibble: 565 x 6
##    Country_Region date       cases deaths deaths_per_mill Population
##    <chr>          <date>     <dbl>  <dbl>           <dbl>      <dbl>
##  1 US             2020-01-22     1      1         0.00300  332875137
##  2 US             2020-01-23     1      1         0.00300  332875137
##  3 US             2020-01-24     2      1         0.00300  332875137
##  4 US             2020-01-25     2      1         0.00300  332875137
##  5 US             2020-01-26     5      1         0.00300  332875137
##  6 US             2020-01-27     5      1         0.00300  332875137
##  7 US             2020-01-28     5      1         0.00300  332875137
##  8 US             2020-01-29     6      1         0.00300  332875137
##  9 US             2020-01-30     6      1         0.00300  332875137
## 10 US             2020-01-31     8      1         0.00300  332875137
## # ... with 555 more rows
```

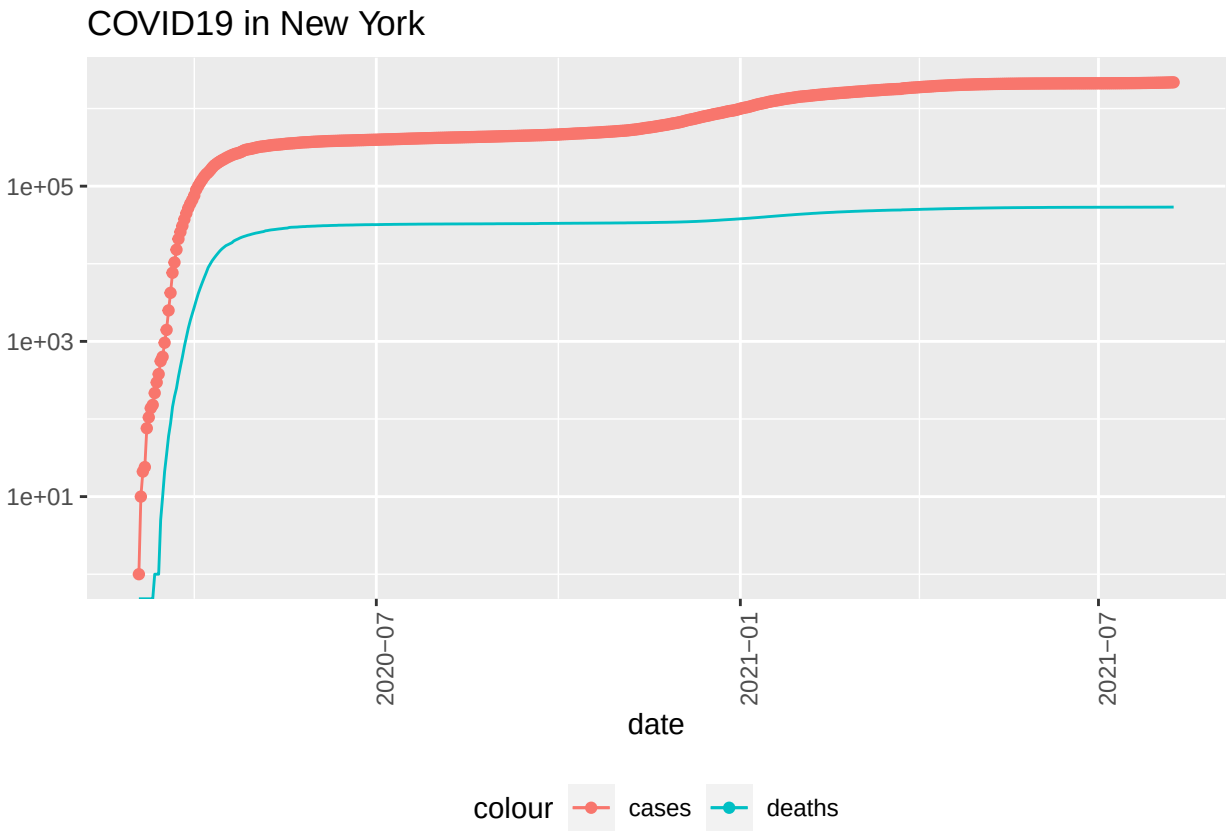
```
tail(US_totals)
```

```
## # A tibble: 6 x 6
##   Country_Region date          cases deaths deaths_per_mill Population
##   <chr>          <date>        <dbl>  <dbl>           <dbl>      <dbl>
## 1 US             2021-08-03 35237950 614295           1845.  332875137
## 2 US             2021-08-04 35330664 614785           1847.  332875137
## 3 US             2021-08-05 35440488 615320           1849.  332875137
## 4 US             2021-08-06 35695469 616493           1852.  332875137
## 5 US             2021-08-07 35739551 616718           1853.  332875137
## 6 US             2021-08-08 35763414 616829           1853.  332875137
```



#Plotting the Data

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




```
## [1] "2021-08-08"
```

#Analyzing the Data

```
max(US_totals$date)
```

```
## [1] "2021-08-08"
```

```
max(US_totals$deaths)
```

```
## [1] 616829
```

```
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))
tail(US_totals)
```

```
## # A tibble: 6 x 8
##   Country_Region date          cases deaths deaths_per_mill Population new_cases
##   <chr>          <date>        <dbl>  <dbl>           <dbl>      <dbl>     <dbl>
## 1 US             2021-08-03 35237950 614295           1845.  332875137    106557
## 2 US             2021-08-04 35330664 614785           1847.  332875137     92714
```

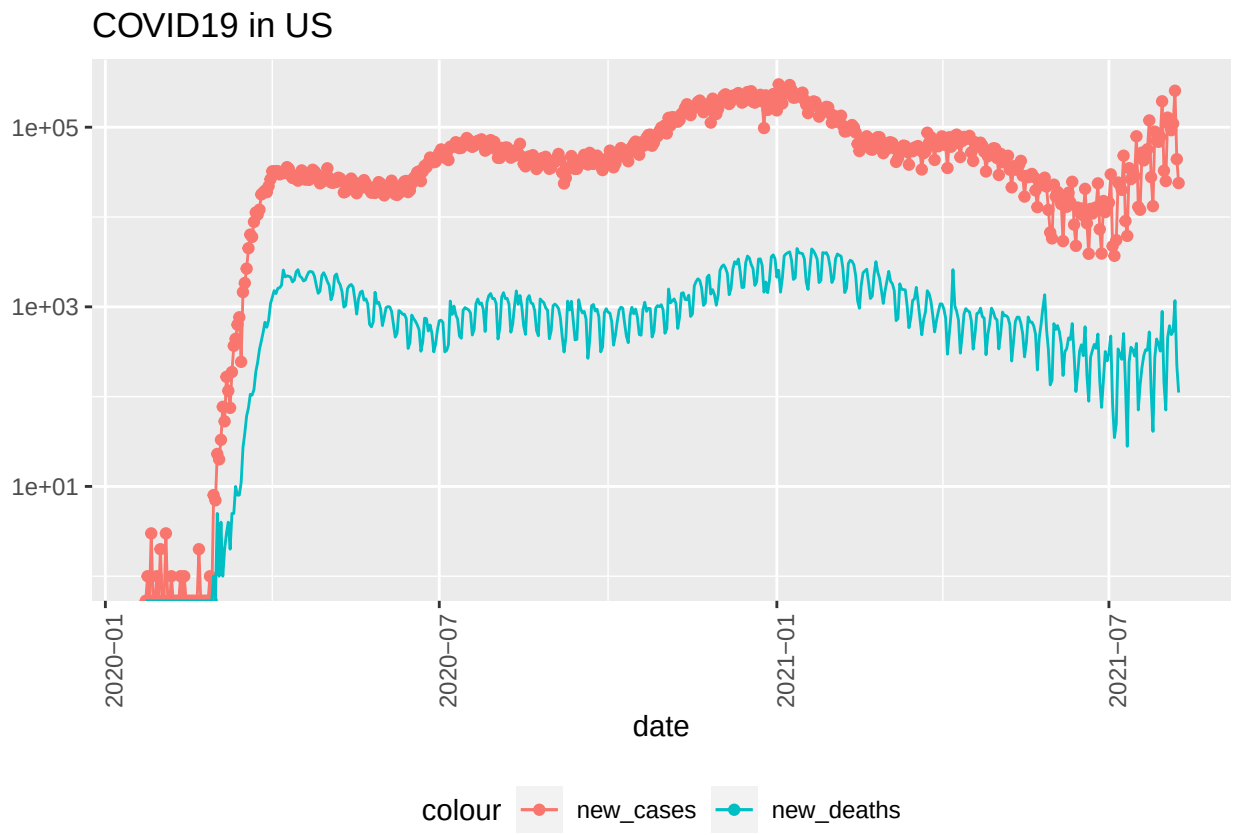
```
## 3 US             2021-08-05 35440488 615320           1849.  332875137    109824
## 4 US             2021-08-06 35695469 616493           1852.  332875137    254981
## 5 US             2021-08-07 35739551 616718           1853.  332875137     44082
## 6 US             2021-08-08 35763414 616829           1853.  332875137     23863
## # ... with 1 more variable: new_deaths <dbl>
```

```
tail(US_totals %>% select(new_cases, new_deaths, everything()))
```

```
## # A tibble: 6 x 8
##   new_cases new_deaths Country_Region date          cases deaths deaths_per_mill
##       <dbl>      <dbl> <chr>          <date>        <dbl>  <dbl>           <dbl>
## 1    106557        616 US             2021-08-03 35237950 614295           1845.
## 2     92714        490 US             2021-08-04 35330664 614785           1847.
## 3    109824        535 US             2021-08-05 35440488 615320           1849.
## 4    254981       1173 US             2021-08-06 35695469 616493           1852.
## 5     44082        225 US             2021-08-07 35739551 616718           1853.
## 6     23863        111 US             2021-08-08 35763414 616829           1853.
## # ... with 1 more variable: Population <dbl>
```

```
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")) +
    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: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 1 row(s) containing missing values (geom_path).

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

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

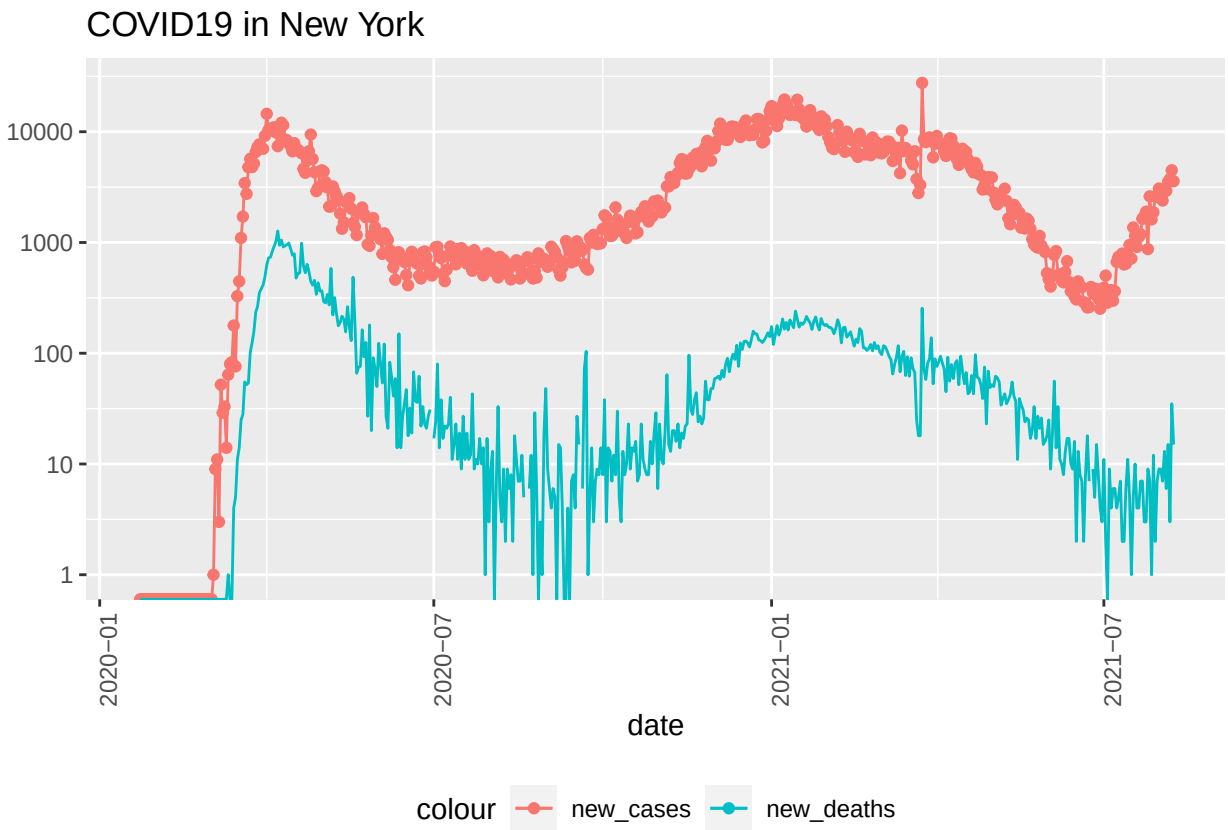



```
state <- "New York"
US_by_state %>%
    filter(Province_State == state) %>%
    ggplot(aes(x = date, y = new_cases)) +
    geom_line(aes(color = "new_cases")) +
    geom_point(aes(color = "new_cases")) +
    geom_line(aes(y = new_deaths, color = "new_deaths")) +
    scale_y_log10() +
    theme(legend.position="bottom", axis.text.x = element_text(angle = 90)) +
  labs(title = str_c("COVID19 in " , state), y = NULL)
```

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

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

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

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

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




```
US_state_totals <- US_by_state %>%
    group_by(Province_State) %>%
    summarize(deaths = max(deaths), cases = max(cases), population = max(Population), cases_per_thou = 
    filter(cases > 0, population > 0)
US_state_totals %>%
    slice_min(deaths_per_thou, n = 10) %>%
select(deaths_per_thou, cases_per_thou, everything())
```

```
## # A tibble: 10 x 6
##    deaths_per_thou cases_per_thou Province_State        deaths  cases population
##              <dbl>          <dbl> <chr>                 <dbl>  <dbl>      <dbl>
##  1          0.0363           3.32 Northern Mariana Isl~     2    183      55144
##  2          0.373           46.2  Virgin Islands           40   4960     107268
##  3          0.383           32.8  Hawaii                  542  46503    1415872
##  4          0.417           40.6  Vermont                 260  25320     623989
##  5          0.537          105.   Alaska                  398  77586     740995
##  6          0.670           53.0  Maine                   901  71306    1344212
##  7          0.685           53.8  Oregon                 2889 226899    4217737
##  8          0.695           40.6  Puerto Rico            2611 152343    3754939
##  9          0.778          137.   Utah                   2494 438479    3205958
## 10          0.810           64.2  Washington             6168 488640    7614893
```

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

```
## # A tibble: 10 x 6
##    deaths_per_thou cases_per_thou Province_State deaths   cases population
##              <dbl>          <dbl> <chr>           <dbl>   <dbl>      <dbl>
##  1            3.00           118. New Jersey      26650 1049222    8882190
##  2            2.76           112. New York        53744 2176658   19453561
##  3            2.63           105. Massachusetts   18095  726395    6892503
##  4            2.59           147. Rhode Island     2743  155825    1059361
##  5            2.56           120. Mississippi      7621  358149    2976149
##  6            2.53           130. Arizona         18388  946054    7278717
##  7            2.41           123. Louisiana       11210  573903    4648794
##  8            2.37           124. Alabama         11624  607209    4903185
##  9            2.33           100. Connecticut      8296  358076    3565287
## 10            2.32           142. South Dakota     2052  125599     884659
```

#Modeling the Data

```
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
## -1.43555 -0.22623 -0.00472  0.21291  1.09042
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.013039   0.217993  -0.060    0.953
## cases_per_thou  0.016279   0.002055   7.922 1.49e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4696 on 53 degrees of freedom
## Multiple R-squared:  0.5421, Adjusted R-squared:  0.5335
## F-statistic: 62.75 on 1 and 53 DF,  p-value: 1.493e-10
```

```
US_state_totals %>% slice_min(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 Northern Mariana Islan~      2   183      55144           3.32          0.0363
```

```
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 North Dakota     1573 112336     762062           147.            2.06
```

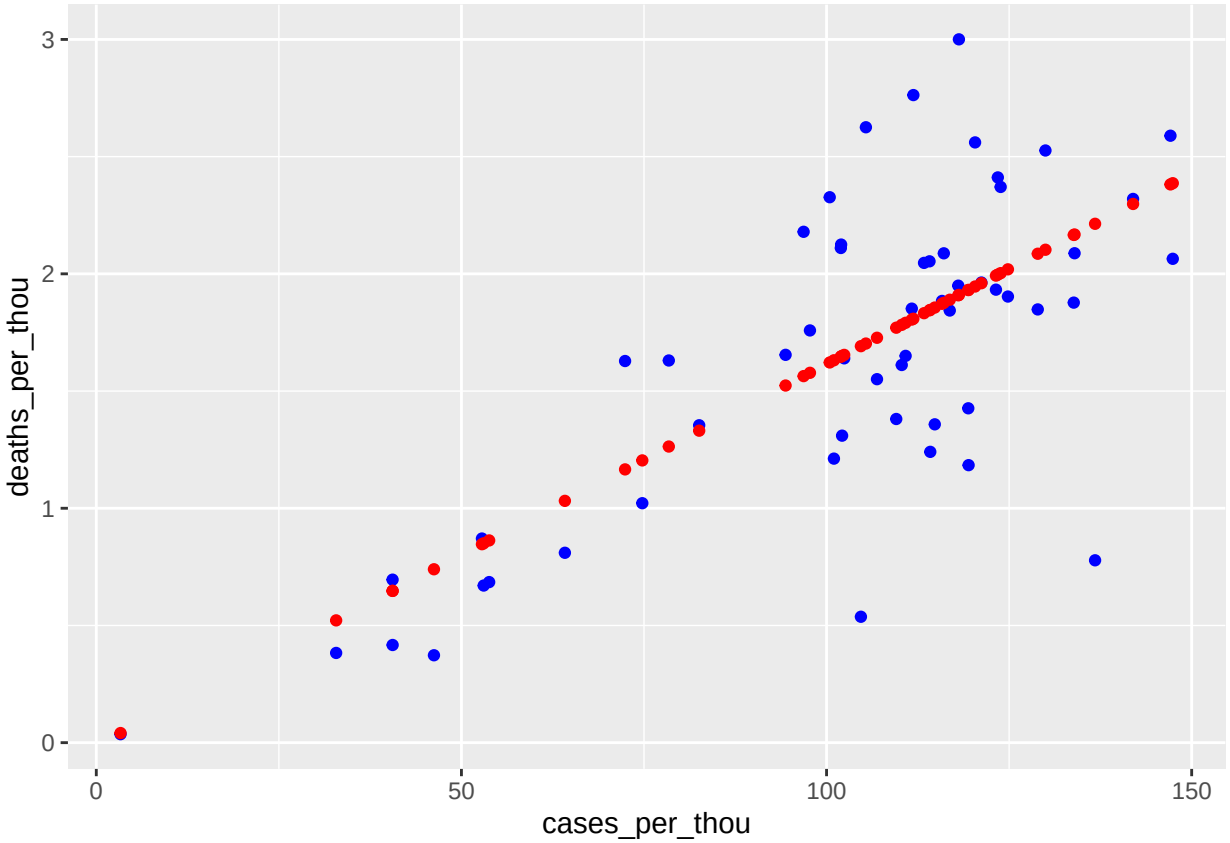
```
x_grid <- seq(1, 151)
new_df <- tibble(cases_per_thou = x_grid)
US_state_totals %>% mutate(pred = predict(mod))
```

```
## # A tibble: 55 x 7
##    Province_State  deaths  cases population cases_per_thou deaths_per_thou  pred
##    <chr>            <dbl>  <dbl>      <dbl>          <dbl>           <dbl> <dbl>
##  1 Alabama          11624 6.07e5    4903185          124.            2.37   2.00
##  2 Alaska             398 7.76e4     740995          105.            0.537  1.69
##  3 Arizona          18388 9.46e5    7278717          130.            2.53   2.10
##  4 Arkansas          6301 4.04e5    3017804          134.            2.09   2.17
##  5 California       64784 4.05e6   39512223          102.            1.64   1.65
##  6 Colorado          6978 5.82e5    5758736          101.            1.21   1.63
##  7 Connecticut       8296 3.58e5    3565287          100.            2.33   1.62
##  8 Delaware          1835 1.13e5     973764          116.            1.88   1.87
##  9 District of Co~   1149 5.11e4     705749           72.4           1.63   1.17
## 10 Florida          39695 2.77e6   21477737          129.            1.85   2.09
## # ... with 45 more rows
```

```
US_total_w_pred <- US_state_totals %>% mutate(pred = predict(mod))
US_total_w_pred
```

```
## # A tibble: 55 x 7
##    Province_State  deaths  cases population cases_per_thou deaths_per_thou  pred
##    <chr>            <dbl>  <dbl>      <dbl>          <dbl>           <dbl> <dbl>
##  1 Alabama          11624 6.07e5    4903185          124.            2.37   2.00
##  2 Alaska             398 7.76e4     740995          105.            0.537  1.69
##  3 Arizona          18388 9.46e5    7278717          130.            2.53   2.10
##  4 Arkansas          6301 4.04e5    3017804          134.            2.09   2.17
##  5 California       64784 4.05e6   39512223          102.            1.64   1.65
##  6 Colorado          6978 5.82e5    5758736          101.            1.21   1.63
##  7 Connecticut       8296 3.58e5    3565287          100.            2.33   1.62
##  8 Delaware          1835 1.13e5     973764          116.            1.88   1.87
##  9 District of Co~   1149 5.11e4     705749           72.4           1.63   1.17
## 10 Florida          39695 2.77e6   21477737          129.            1.85   2.09
## # ... with 45 more rows
```

```
US_total_w_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")
```


deaths_per_thou
cases_per_thou
0
1
2
3
0
50
100
150