Analysis of COVID-19 time series data

STAT632

Patricia Reynoso

May 22, 2020

The Johns Hopkins University, Center for Systems Science and Engineering, has compiled a global epidemiological data set on the Novel Corona Virus (COVID-19).1 You can use the following command to load the data into R:

```
covid <- read.csv("https://ericwfox.github.io/data/covid19confirmed.csv")
head(covid)
```

```
##   Province.State     Country Lat Long    Date Value
## 1                Afghanistan  33   65 3/25/20    84
## 2                Afghanistan  33   65 3/24/20    74
## 3                Afghanistan  33   65 3/23/20    40
## 4                Afghanistan  33   65 3/22/20    40
## 5                Afghanistan  33   65 3/21/20    24
## 6                Afghanistan  33   65 3/20/20    24
```

The data set gives the number of confirmed cases (Value column) since 22 January 2020 for each country. As a data processing step, we need to convert the Date column from a factor type in R, to a Date object. We can use the lubridate package to do this:

```
library(lubridate)
covid$Date <- mdy(covid$Date)
class(covid$Date)
```

```
## [1] "Date"
```

```
head(covid)
```

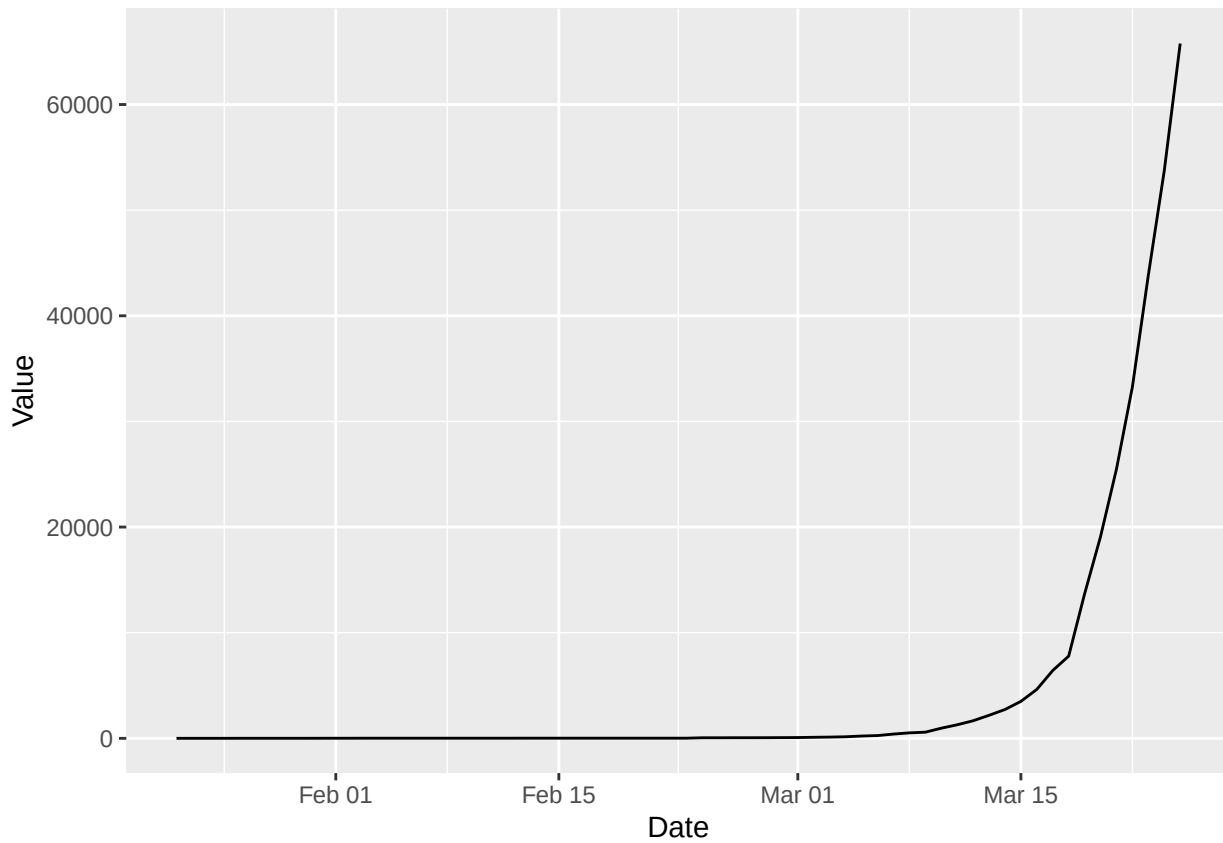
```
##   Province.State     Country Lat Long       Date Value
## 1                Afghanistan  33   65 2020-03-25    84
## 2                Afghanistan  33   65 2020-03-24    74
## 3                Afghanistan  33   65 2020-03-23    40
## 4                Afghanistan  33   65 2020-03-22    40
## 5                Afghanistan  33   65 2020-03-21    24
## 6                Afghanistan  33   65 2020-03-20    24
```

Let's start our analysis by using dplyr to subset the time series of confirmed cases in the United States:

```
library(dplyr)
covid_us <- covid %>% filter(Country == "US")
```

Question 1: Use ggplot2 to make a time series plot with the date on horizontal axis, and number of confirmed cases in the US on the vertical axis. Use geom line() when displaying the data. Next, let's investigate times series plots for other countries as well. The following command uses dplyr to subset data for Hubei province in China, Italy, South Korea, and the United States:

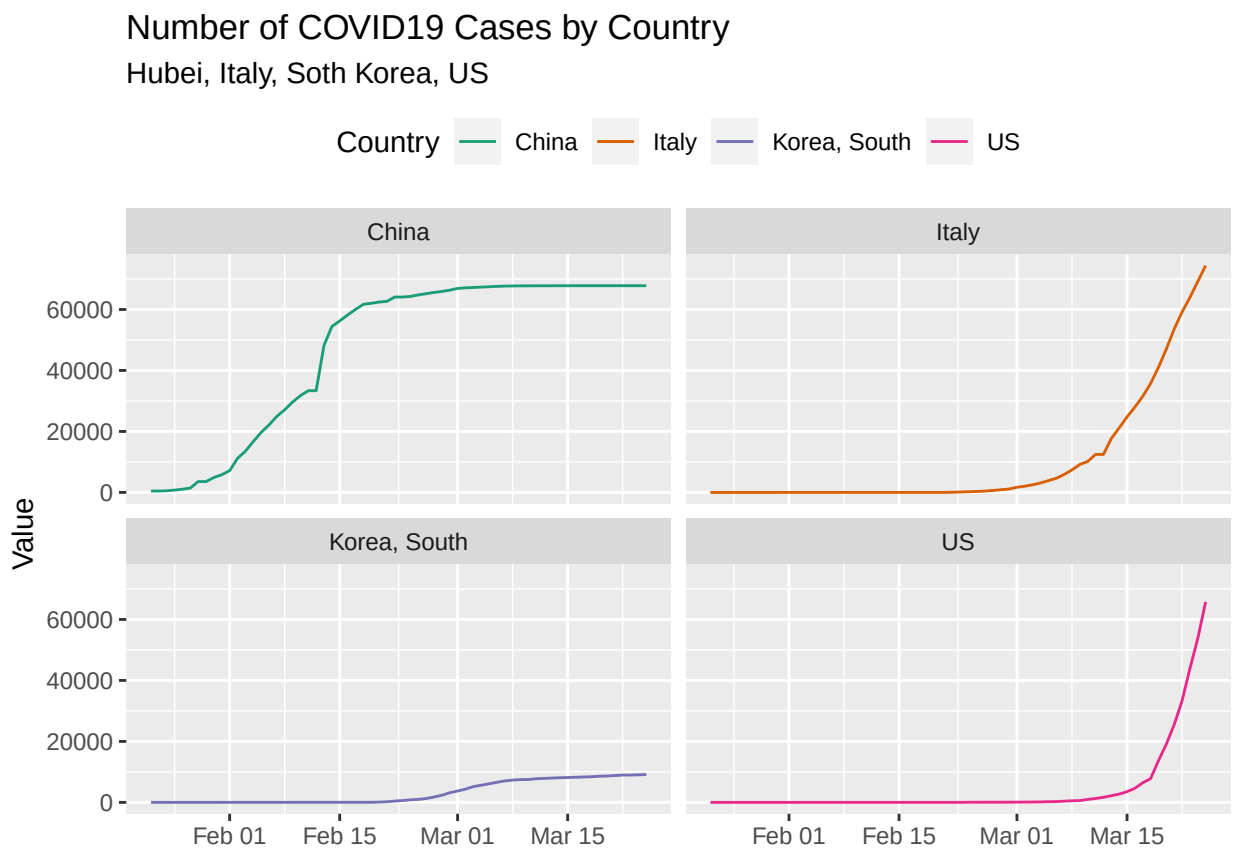
```r
library(ggplot2)

covid_us%>%
  ggplot(aes(x = Date, y = Value)) + geom_line()
```



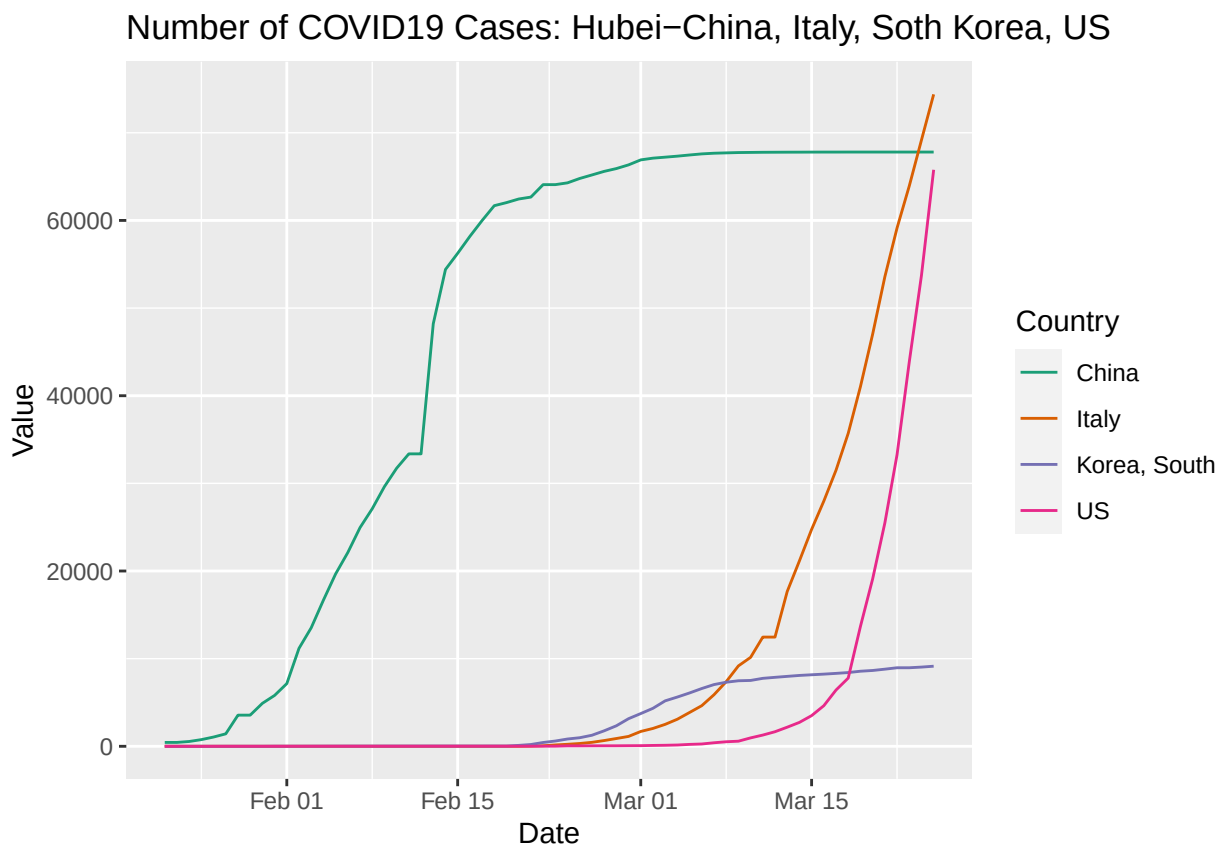

```r
covid_2 <- covid %>%
filter(Province.State == "Hubei" | Country == "Italy" |
Country == "Korea, South" | Country == "US")
```

Question 2: Use ggplot2 to make times series plots for the four countries that were subsetted. Use facet wrap() to display each plot in a separate panel. You can also try to create a "combined" plot with all four time series curves colored by country.

```r
covid_2%>%
  ggplot(aes(x=Date, y= Value, group=Country, color = Country)) + geom_line() +
  facet_wrap(~ Country, ncol =  2) +
  scale_color_brewer(type = "qual", palette = "Dark2")+
  labs(title = "Number of COVID19 Cases by Country",
  subtitle = "Hubei, Italy, Soth Korea, US",
  fill = "Country", color = "Country", x = NULL, y= "Value") +
  theme(legend.position = "top")
```




```
t <- ggplot(data = covid_2,
            aes(x = Date, y = Value, group = Country, color = Country))


t + geom_line() +
  scale_color_brewer(type = "qual", palette = "Dark2") +
  ggtitle("Number of COVID19 Cases: Hubei-China, Italy, Soth Korea, US")
```



Question 3: Make time series plots for several other countries that are of interest to you. Consider continuing your exploratory analysis, and investigate other aspects of the data set.

For the following analysis I used: https://github.com/nytimes/covid-19-data

```
library(skimr)
library(ggplot2)
library(dplyr)
library(lubridate)

us_states <- read.csv("us-states.csv")

tail(us_states)
```

```
##            date          state fips cases deaths
## 4354 2020-05-20 Virgin Islands   78    69      6
## 4355 2020-05-20       Virginia   51 32908   1074
## 4356 2020-05-20     Washington   53 20179   1045
## 4357 2020-05-20  West Virginia   54  1567     69
## 4358 2020-05-20      Wisconsin   55 13574    481
## 4359 2020-05-20        Wyoming   56   787     11
```

```
us_states <- transform(us_states, date = as.Date(date, "%Y-%m-%d"),
                       state = as.character(state))

cases1 <- us_states%>%
                  filter(date == "2020-05-20")%>%
                  mutate(cases = cases/1000)
```

```r
head(cases1)
```

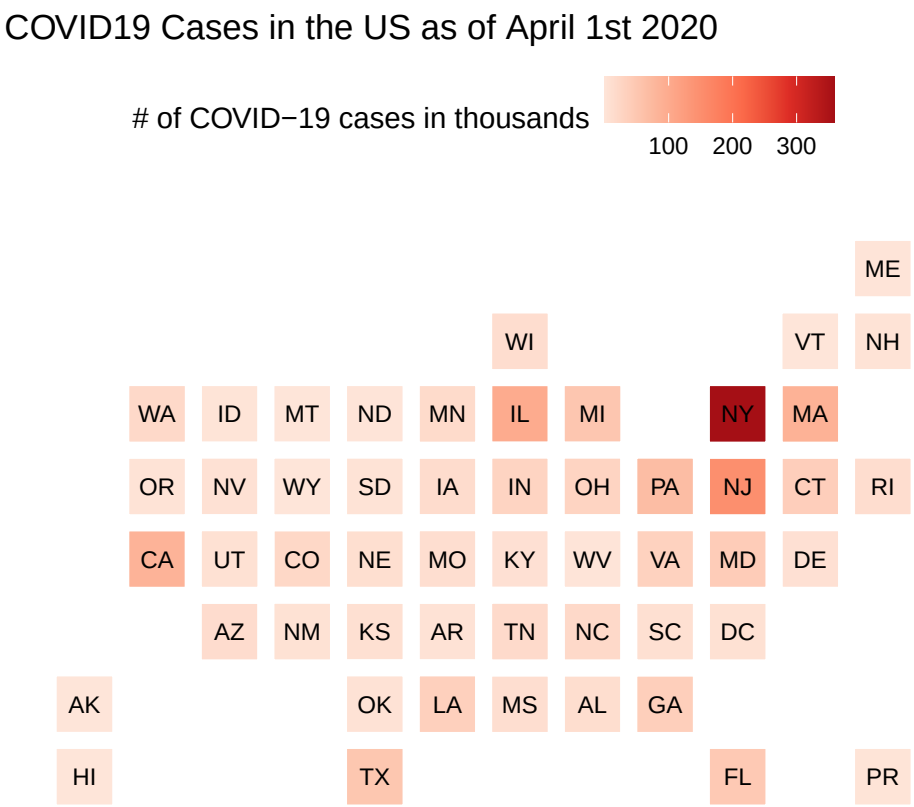
```
##         date      state fips  cases deaths
## 1 2020-05-20    Alabama    1 13.052    522
## 2 2020-05-20     Alaska    2  0.402      8
## 3 2020-05-20    Arizona    4 14.897    747
## 4 2020-05-20   Arkansas    5  5.003    107
## 5 2020-05-20 California    6 86.125   3514
## 6 2020-05-20   Colorado    8 22.769   1299
```

```r
#us_counties <- read.csv("us-counties.csv")
#us_counties <- transform(us_counties, date = as.Date(ï..date, "%m/%d/%Y"))

tail(us_states)
```

```
##            date          state fips cases deaths
## 4354 2020-05-20 Virgin Islands   78    69      6
## 4355 2020-05-20       Virginia   51 32908   1074
## 4356 2020-05-20     Washington   53 20179   1045
## 4357 2020-05-20  West Virginia   54  1567     69
## 4358 2020-05-20      Wisconsin   55 13574    481
## 4359 2020-05-20        Wyoming   56   787     11
```

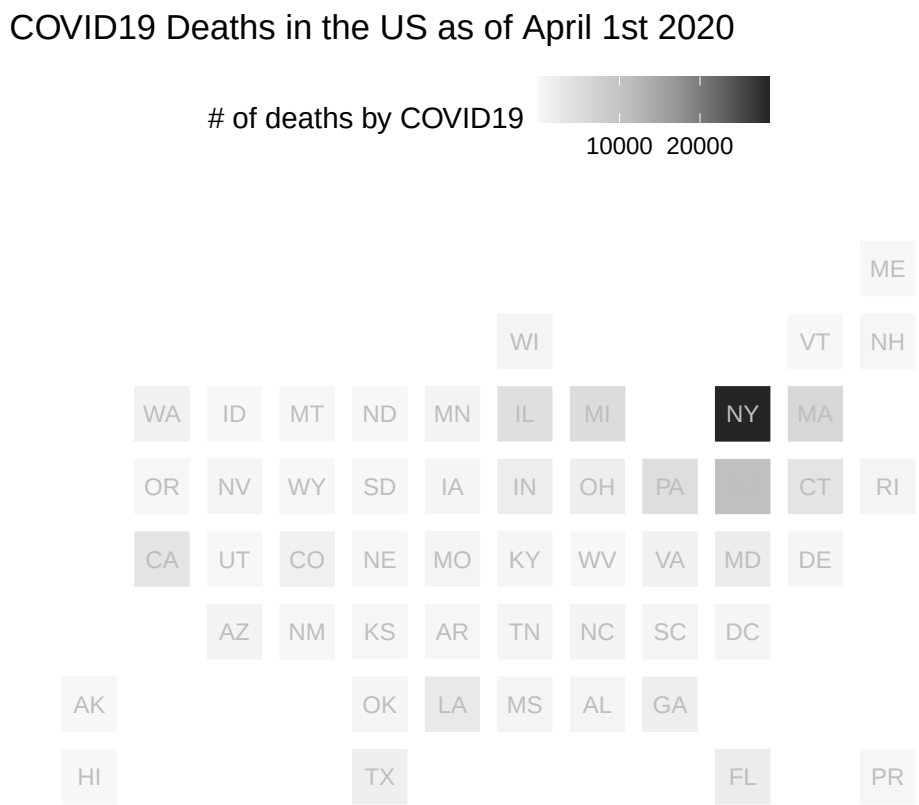
```r
#head(us_counties)
```

```r
library(statebins)
statebins_continuous(state_data = cases1, state_col = "state",
text_color = "black", value_col = "cases",
brewer_pal="Reds",
legend_title="# of COVID-19 cases in thousands",
plot_title = "COVID19 Cases in the US as of April 1st 2020", title_position = "top")
```




```r
statebins_continuous(state_data = cases1, state_col = "state",
text_color = "grey", value_col = "deaths",
brewer_pal="Greys", font_size = 3,
legend_title="# of deaths by COVID19",
plot_title= "COVID19 Deaths in the US as of April 1st 2020", title_position = "top")
```

```
## Warning in validate_states(state_data, state_col, merge.x): Found invalid state
## values: GuamNorthern Mariana IslandsVirgin Islands

## Warning: `show_guide` has been deprecated. Please use `show.legend` instead.
```




```r
cases_NY <- us_states%>%
                  filter(state == "New York")%>%
                  mutate(cases = cases/1000)
head(cases_NY)
```

```
##         date    state fips cases deaths
## 1 2020-03-01 New York   36 0.001      0
## 2 2020-03-02 New York   36 0.001      0
## 3 2020-03-03 New York   36 0.002      0
## 4 2020-03-04 New York   36 0.011      0
## 5 2020-03-05 New York   36 0.022      0
## 6 2020-03-06 New York   36 0.044      0
```

```r
cases_NY%>%
  filter(date=="2020-03-22")
```

```
##         date    state fips  cases deaths
## 1 2020-03-22 New York   36 15.188    142
```

```r
cases_CA <-us_states%>%
                  filter(state == "California")%>%
                  mutate(cases = cases/1000)

cases_CA%>%
  filter(date=="2020-03-19")
```

```
##         date      state fips cases deaths
## 1 2020-03-19 California    6 1.067     19
```

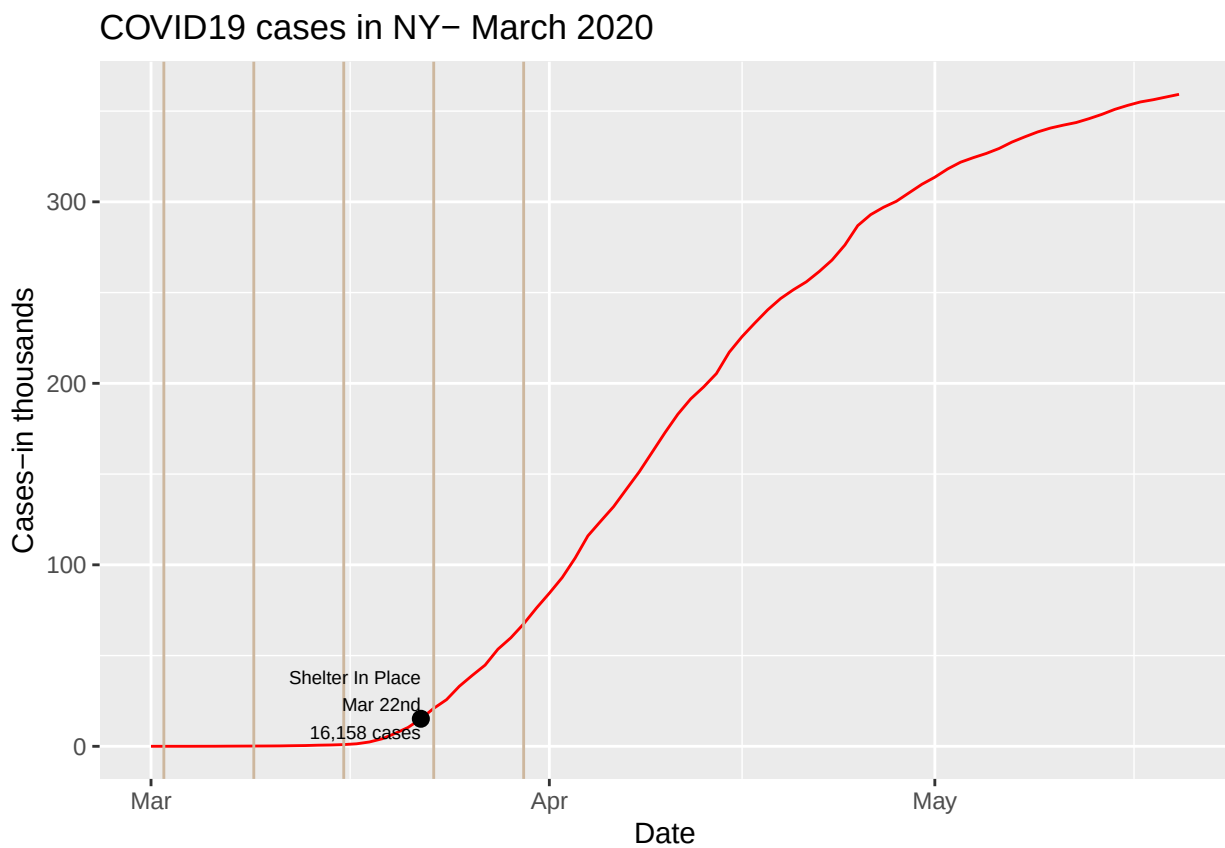
```r
text <- bind_rows(
  data.frame(x = as.Date("2020-03-22"), y = 23,
             label = paste("Shelter In Place\n", "Mar 22nd\n","16,158 cases"), adj = 0))

text
```

```
##            x  y                                       label adj
## 1 2020-03-22 23 Shelter In Place\n Mar 22nd\n 16,158 cases   0
```

```r
shelterInPlace <- data_frame(
                    when = as.Date("2020-03-22"),
                    cases = (15.168)
                    )

shelterInPlace
```

```
## # A tibble: 1 x 2
##   when       cases
##   <date>     <dbl>
## 1 2020-03-22  15.2
```

```r
weeks_march <- data_frame(
                    date = as.Date(c("2020-03-02","2020-03-9","2020-03-16",
                                     "2020-03-23", "2020-03-30")))

casesPlot <- cases_NY%>%
                 ggplot(aes(x = date, y = cases)) + geom_line(size = 0.2)+
                 geom_line(col = "red") +
                 ggtitle("COVID19 cases in NY- March 2020") + ylab("Cases-in thousands")+
                 xlab("Date") + geom_vline(data = weeks_march,
                 aes(xintercept = as.numeric(date)), col = "bisque3") +
                 geom_text(data = text, aes(x = x, y = y, label = label),
                           hjust = "right", size = 2.5) +
                 geom_point(data = shelterInPlace, aes(x = when, y= cases), size = 2.5)

casesPlot
```




```
text <- bind_rows(
  data.frame(x = as.Date("2020-03-19"), y = 1.8,
             label = paste("Shelter In Place\n", "Mar 19th\n","1,067 cases"), adj = 0))

text
```

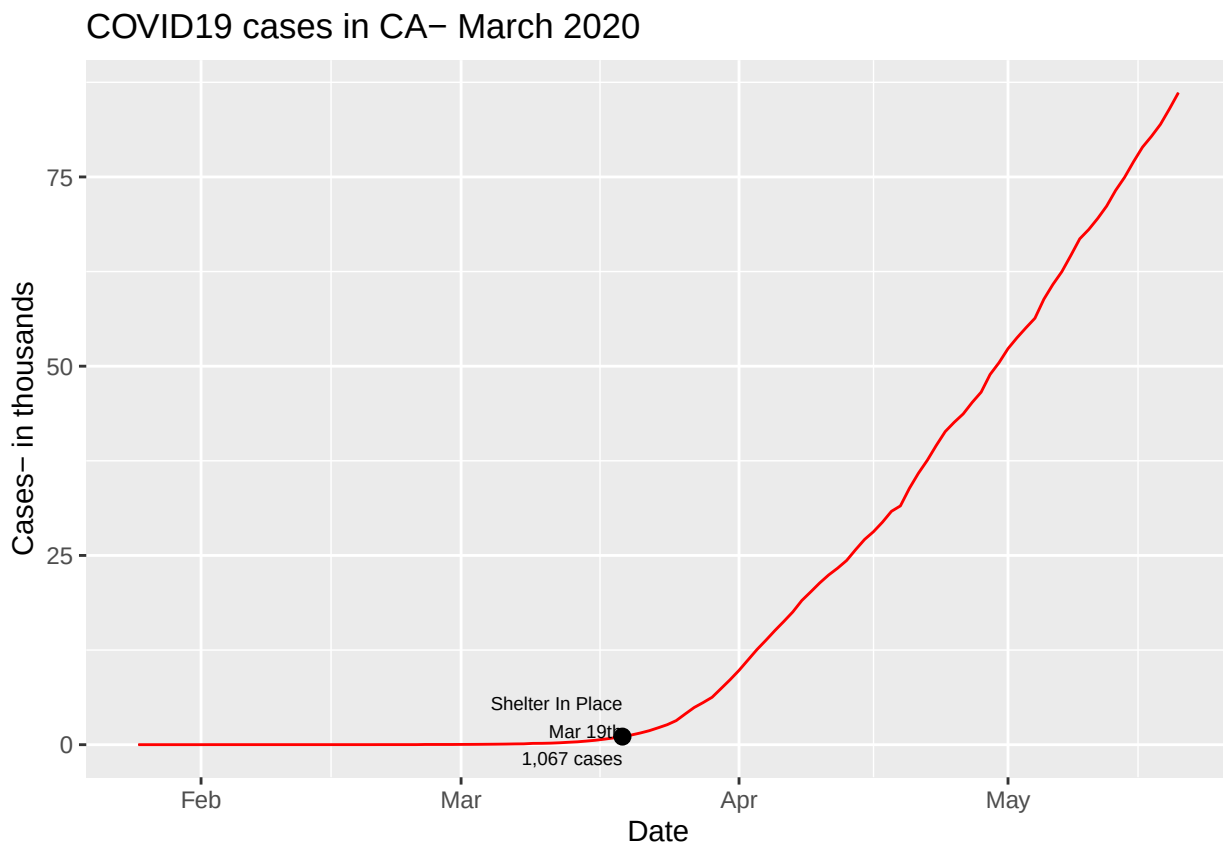
```
##            x   y                                      label adj
## 1 2020-03-19 1.8 Shelter In Place\n Mar 19th\n 1,067 cases   0
```

```
shelterInPlace1 <- data_frame(
                    when = as.Date("2020-03-19"),
                    cases = (1.067)
                    )

shelterInPlace1
```

```
## # A tibble: 1 x 2
##   when       cases
##   <date>     <dbl>
## 1 2020-03-19  1.07
```

```
casesPlot1 <- cases_CA%>%
                 ggplot(aes(x = date, y = cases)) + geom_line(size = 0.2)+
                 geom_line(col = "red") +
                 ggtitle("COVID19 cases in CA- March 2020") +
                 xlab("Date")  +  ylab("Cases- in thousands") +
                 geom_text(data = text, aes(x = x, y = y, label = label),
                           hjust = "right", size = 2.5) +
```

```
                 geom_point(data = shelterInPlace1, aes(x = when, y= cases), size = 2.5)

casesPlot1
```

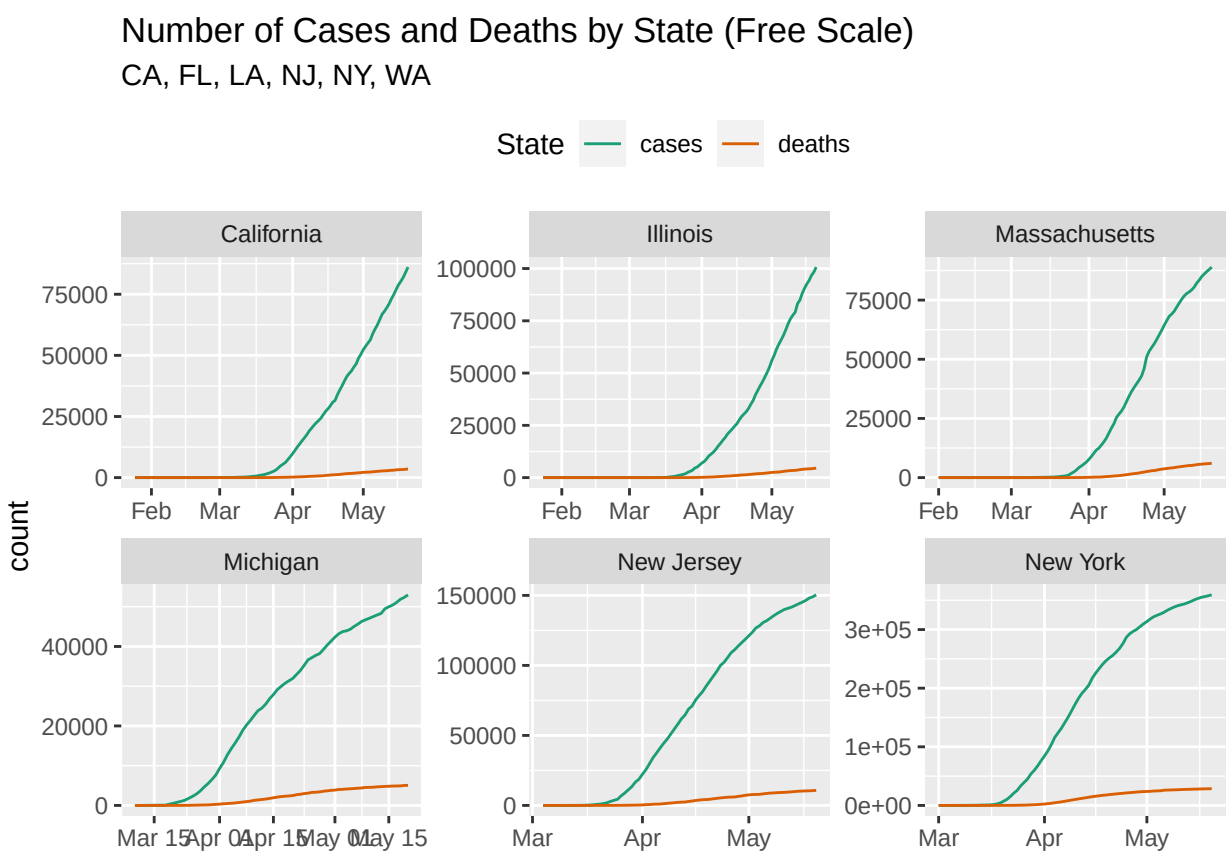



```
library(tidyr)

us_states_long <- us_states%>%
  pivot_longer(cols = cases:deaths, names_to = "Type", values_to = "count")%>%
  mutate(Type = as.factor(Type))


head(us_states_long)
```

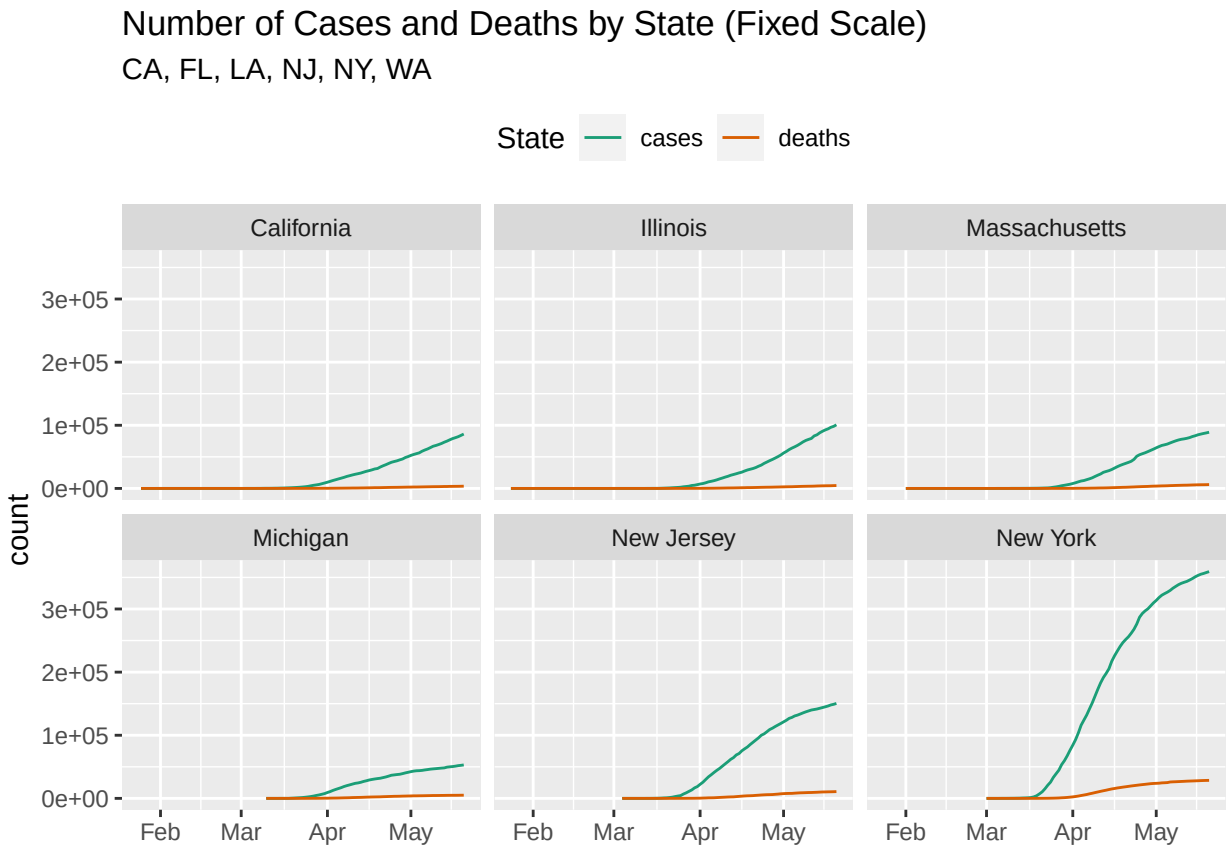
```
## # A tibble: 6 x 5
##   date       state       fips Type   count
##   <date>     <chr>      <int> <fct>  <int>
## 1 2020-01-21 Washington    53 cases      1
## 2 2020-01-21 Washington    53 deaths     0
## 3 2020-01-22 Washington    53 cases      1
## 4 2020-01-22 Washington    53 deaths     0
## 5 2020-01-23 Washington    53 cases      1
## 6 2020-01-23 Washington    53 deaths     0
```

```
p <- ggplot(data = subset(us_states_long, state %in% c("New York", "California",
                                                       "Michigan", "New Jersey",
                                                       "Massachusetts", "Illinois")),
mapping = aes(x = date, y = count, group = Type, color = Type))
```

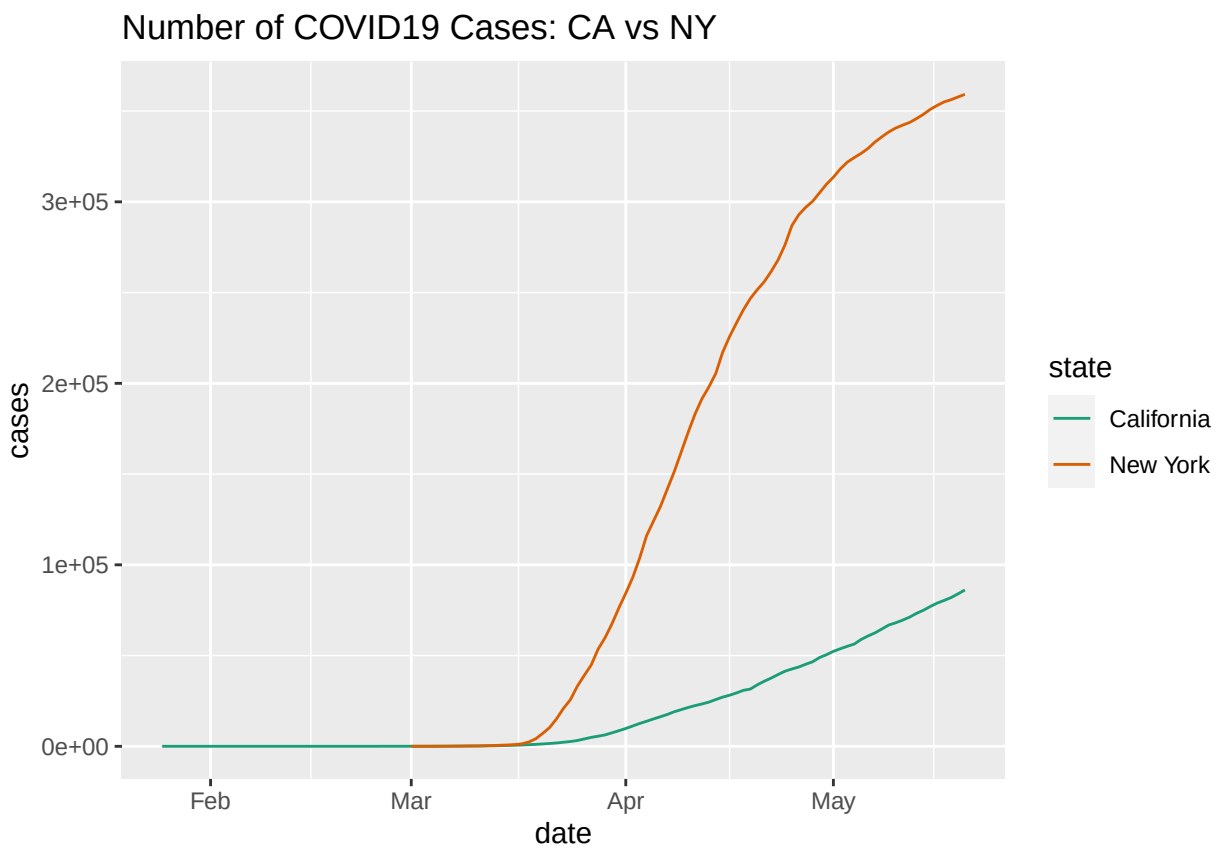
```r
p + geom_line()+
facet_wrap(~ state, ncol = 3, scales = "free") +
scale_color_brewer(type = "qual", palette = "Dark2") +
labs(title = "Number of Cases and Deaths by State (Free Scale)",
subtitle = "CA, FL, LA, NJ, NY, WA", fill = "state",
color = "State", x = NULL, y = "count") +
theme(legend.position = "top")
```



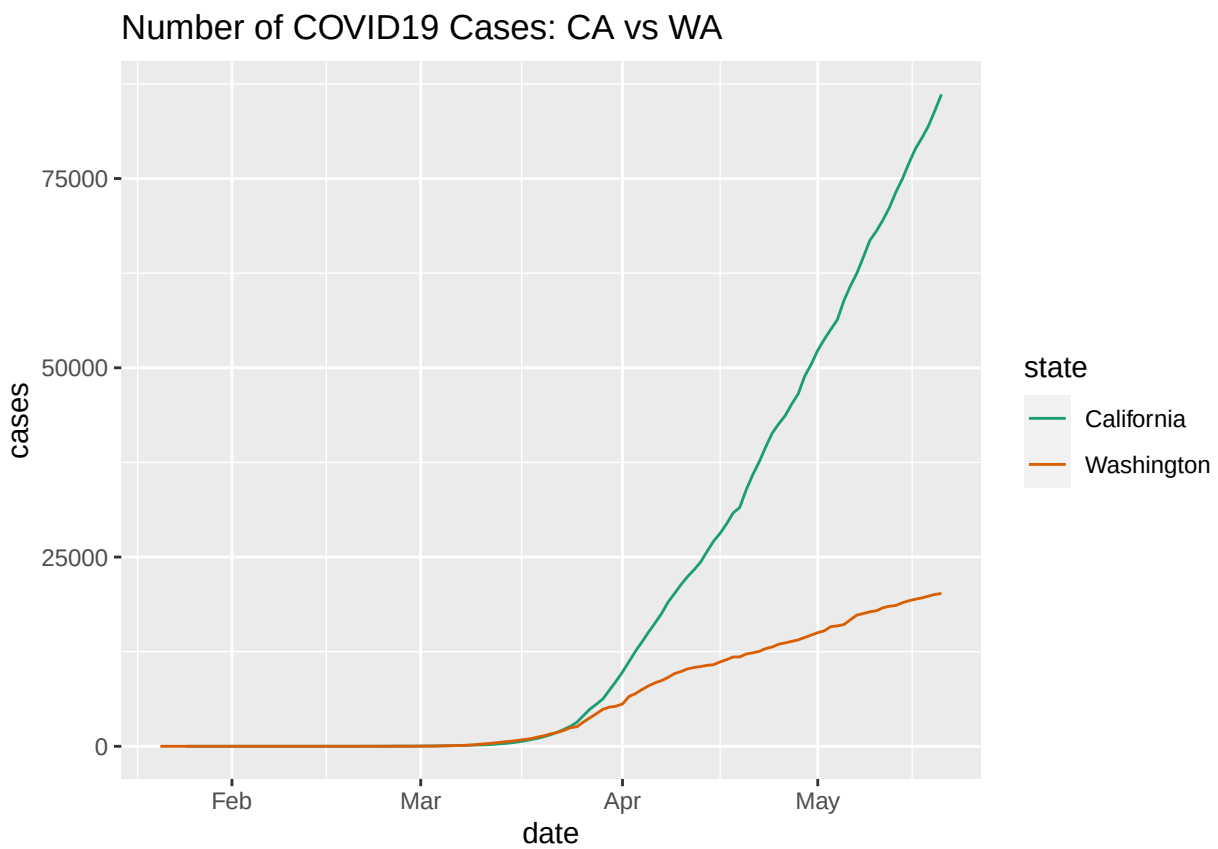

```r
p + geom_line()+
facet_wrap(~ state, ncol = 3) +
scale_color_brewer(type = "qual", palette = "Dark2") +
labs(title = "Number of Cases and Deaths by State (Fixed Scale)",
subtitle = "CA, FL, LA, NJ, NY, WA", fill = "state",
color = "State", x = NULL, y = "count") +
theme(legend.position = "top")
```



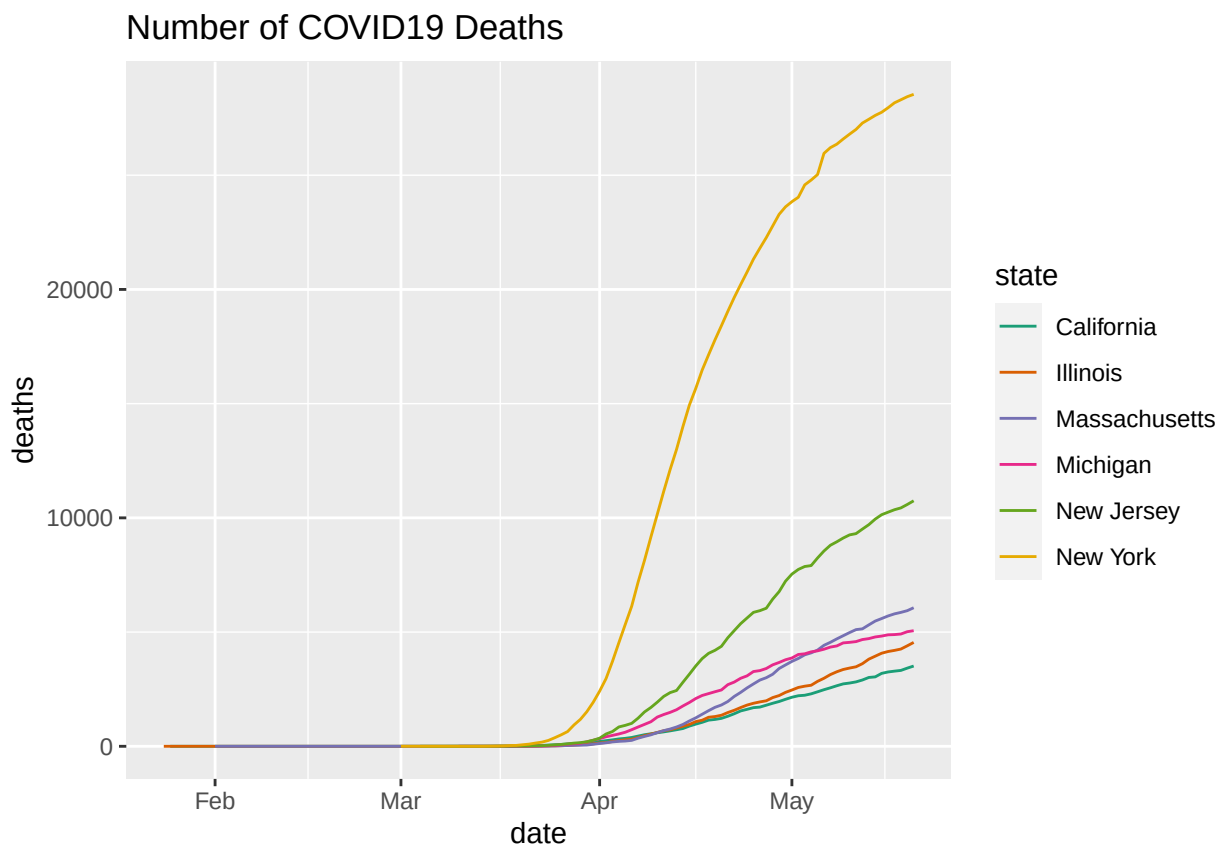

```
q <- ggplot(data = subset(us_states, state %in% c( "California",
                                                   "New York")),
            aes(x = date, y = cases, group = state, color = state))


q + geom_line() +
  scale_color_brewer(type = "qual", palette = "Dark2") +
  ggtitle("Number of COVID19 Cases: CA vs NY")
```



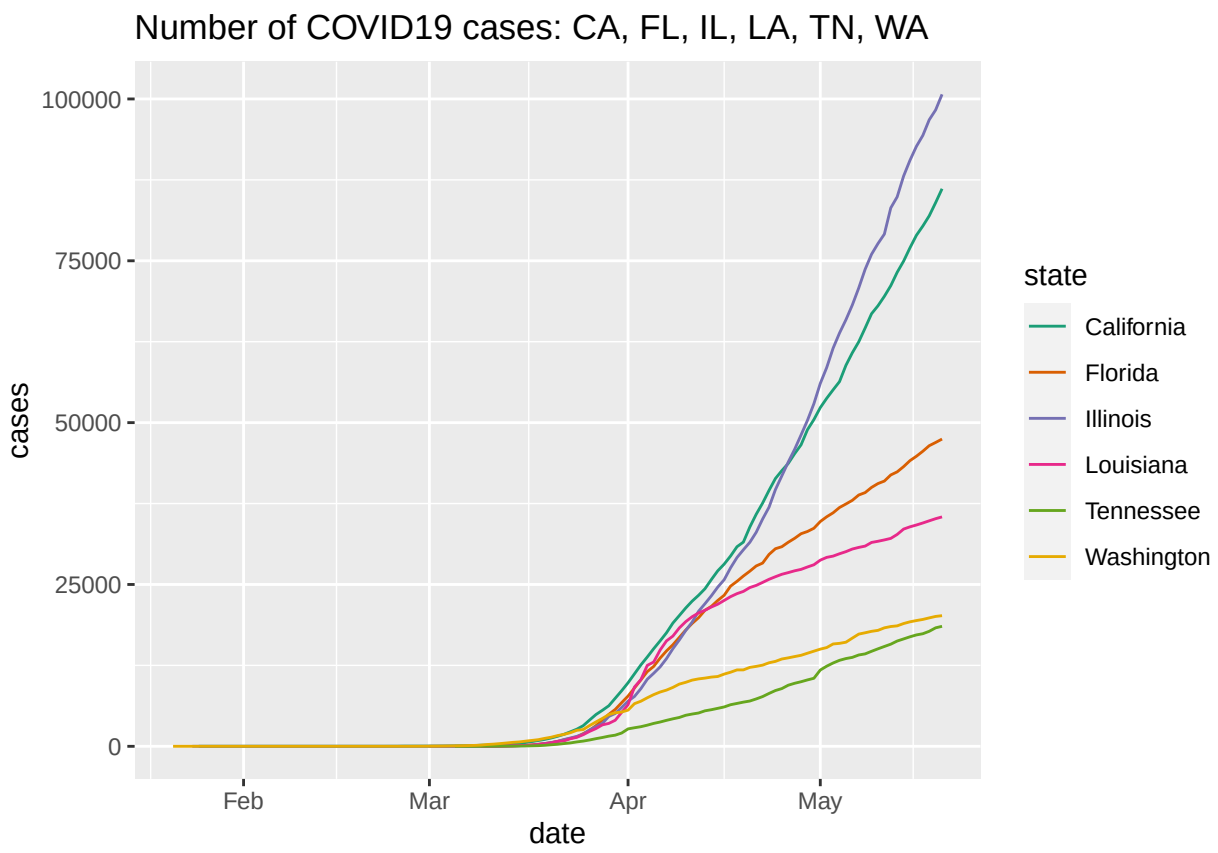

```
q <- ggplot(data = subset(us_states, state %in% c( "California",
                                                   "Washington")),
            aes(x = date, y = cases, group = state, color = state))


q + geom_line() +
  scale_color_brewer(type = "qual", palette = "Dark2") +
  ggtitle("Number of COVID19 Cases: CA vs WA")
```



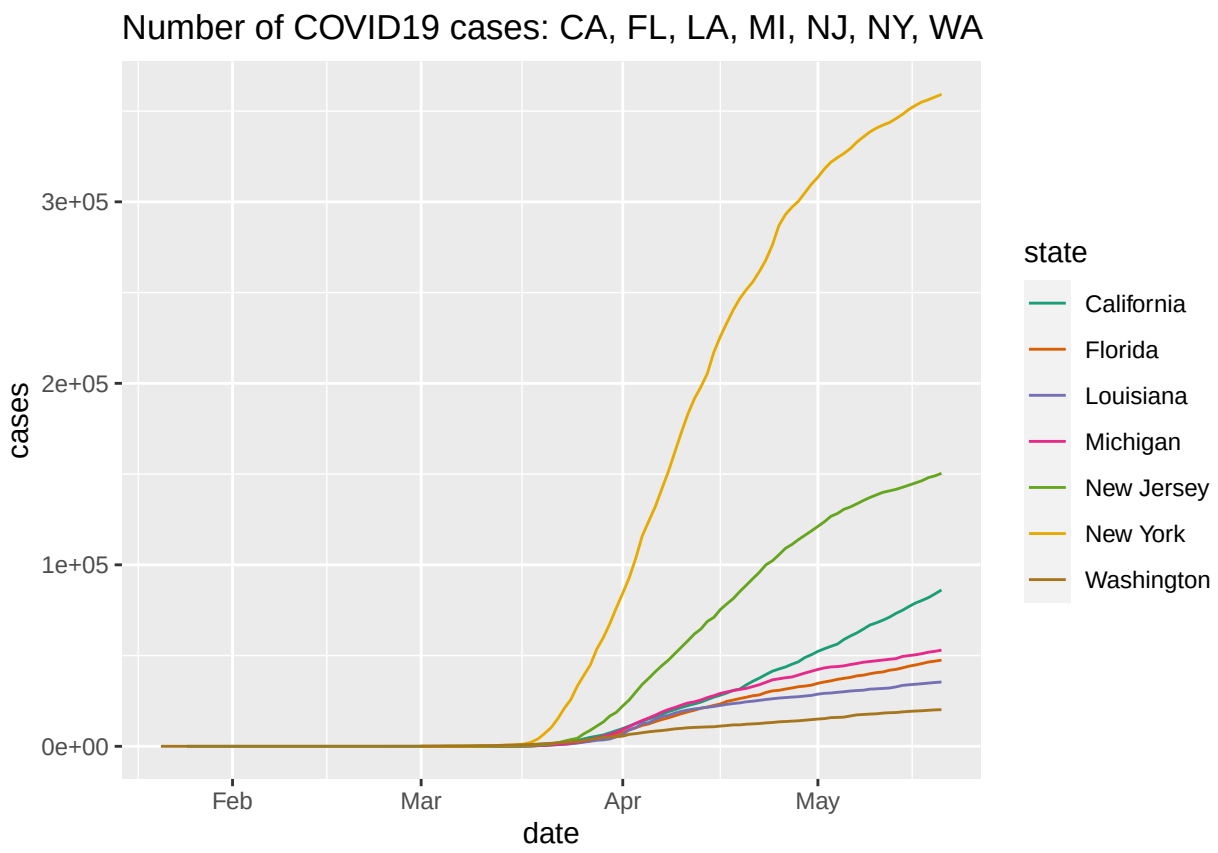

```
t <- ggplot(data = subset(us_states, state %in% c("New York", "California", "New Jersey", "Michigan", "
            aes(x = date, y = deaths, group = state, color = state))


t + geom_line() +
  scale_color_brewer(type = "qual", palette = "Dark2") +
  ggtitle("Number of COVID19 Deaths  ")
```




```
q <- ggplot(data = subset(us_states, state %in% c("Tennessee", "California",
                                                  "Washington", "Illinois",
                                                  "Louisiana", "Florida")),
            aes(x = date, y = cases, group = state, color = state))


q + geom_line() +
  scale_color_brewer(type = "qual", palette = "Dark2") +
  ggtitle("Number of COVID19 cases: CA, FL, IL, LA, TN, WA")
```




```
s <- ggplot(data = subset(us_states, state %in% c("New York", "California",
                                                  "Washington", "New Jersey",
                                                  "Louisiana", "Florida", "Michigan")),
            aes(x = date, y = cases, group = state, color = state))


s + geom_line() +
  scale_color_brewer(type = "qual", palette = "Dark2") +
  ggtitle("Number of COVID19 cases: CA, FL, LA, MI, NJ, NY, WA")
```


Number of COVID19 cases: CA, FL, LA, MI, NJ, NY, WA
cases
3e+05
2e+05
1e+05
0e+00
Feb
Mar
Apr
May
date
state
California
Florida
Louisiana
Michigan
New Jersey
New York
Washington