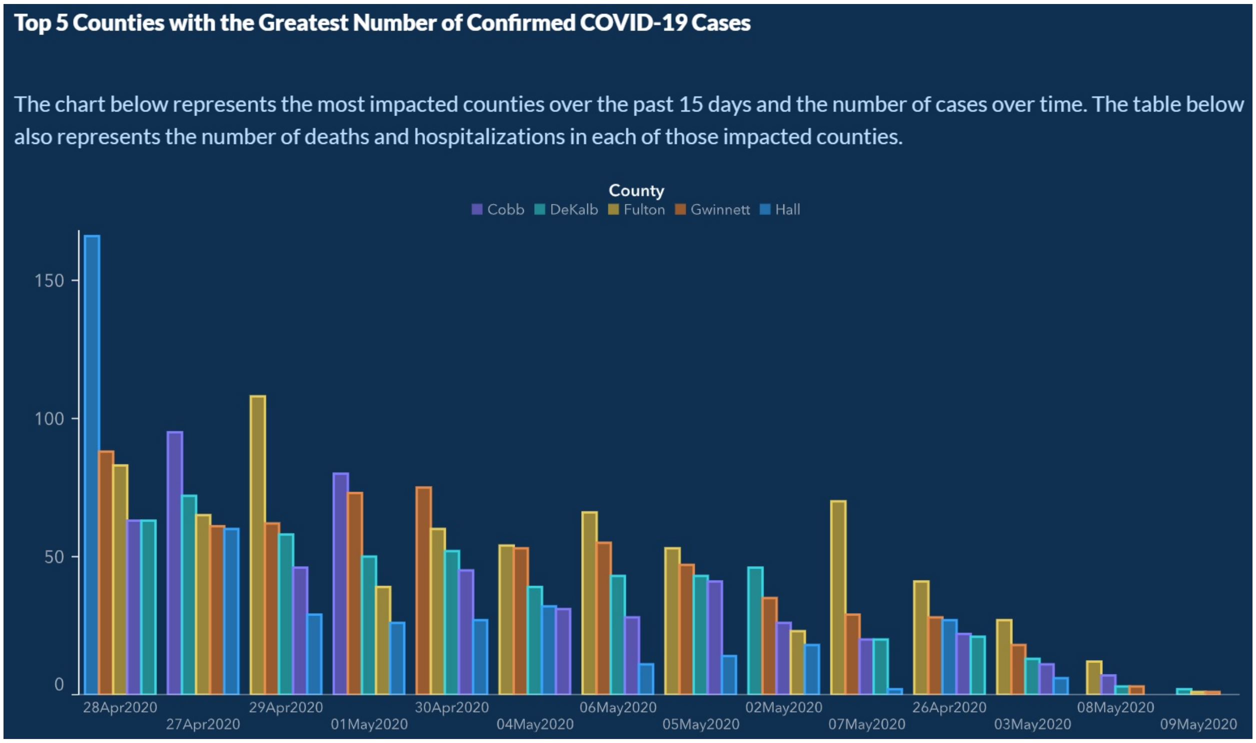
DATA 101: Home work 2

Christopher Hainzl

In May 2020, the Georgia Department of Public Health posted the following plot to illustrate the number of confirmed COVID-19 cases in their hardest-hit counties over two weeks.



Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.

The statistical community and several media outlets heavily criticized the plot for its deceptive portrayal of COVID-19 trends in Georgia. Whether the end result was malicious intent or poor judgment, it is incredibly irresponsible to publish data visualizations that obscure and distort the truth.

In this homework, we will "pretend" that we are data scientists tasked with making better COVID-19 visualizations.

We will use the *New York Times* COVID-19 data to get county-level information for Georgia. The code below reads in the the data through May 13, 2022.

```
library(readr)
library(dplyr)
library(ggplot2)
library(gridExtra)

us_counties = read_csv("https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv"
```

1. Create a new data frame called `georgia_counties` that only contains the data from Georgia. Add a new variable called `new_cases` that stores the number of new confirmed cases for each day at the

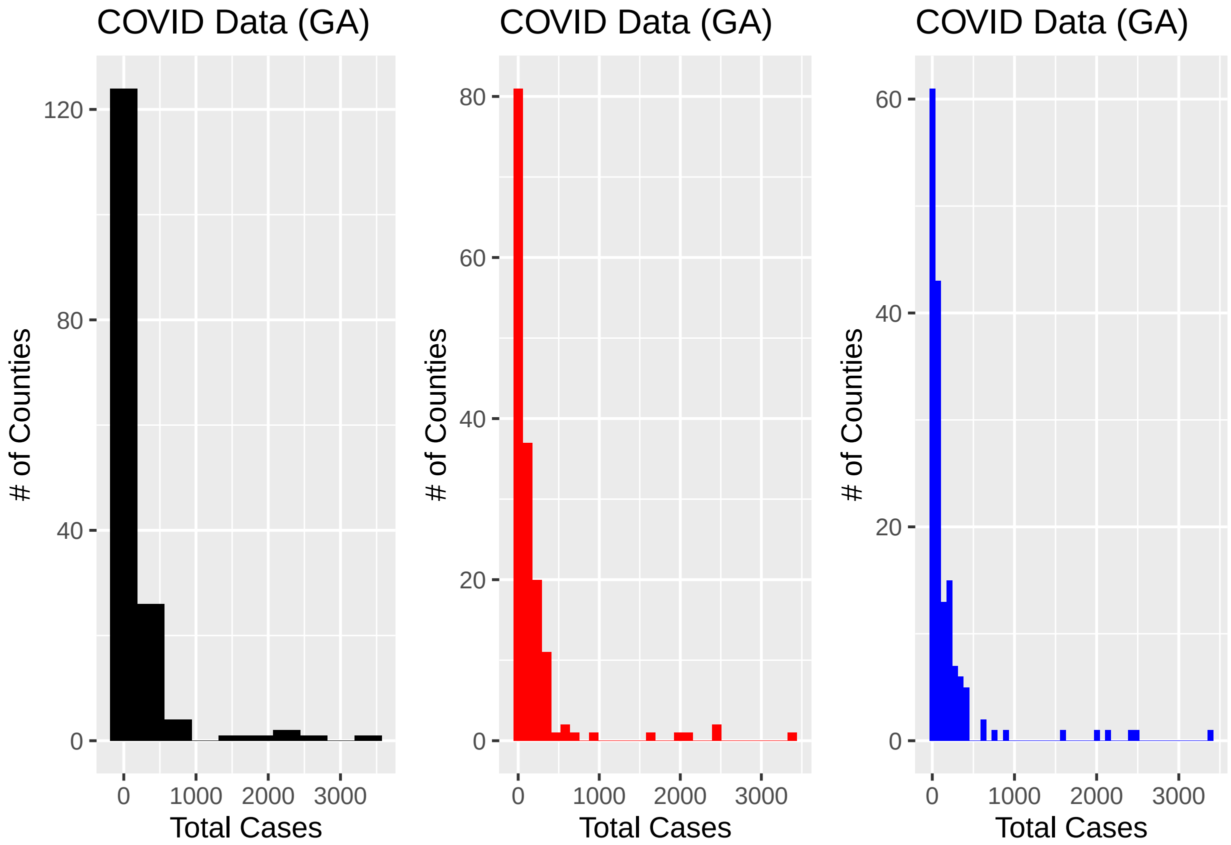
county level. Hint: the `lag` function returns the previous values in a vector.

```
georgia_counties <- us_counties %>%
  filter(state == "Georgia") %>%
  group_by(county) %>%
  arrange(date) %>%
  mutate(new_cases = cases - lag(cases)) %>%
  ungroup()
```

2. Assuming today is May 9th, 2020. We want to get a sense of today's distribution of the total number of confirmed cases in each county in Georgia. Make three histograms, one with 10 black bins, one with 30 red bins, and one with 50 blue bins. Include nice axis labels and titles. Use the `grid.arrange` function from the `gridExtra` package to place the three plots next to each other.

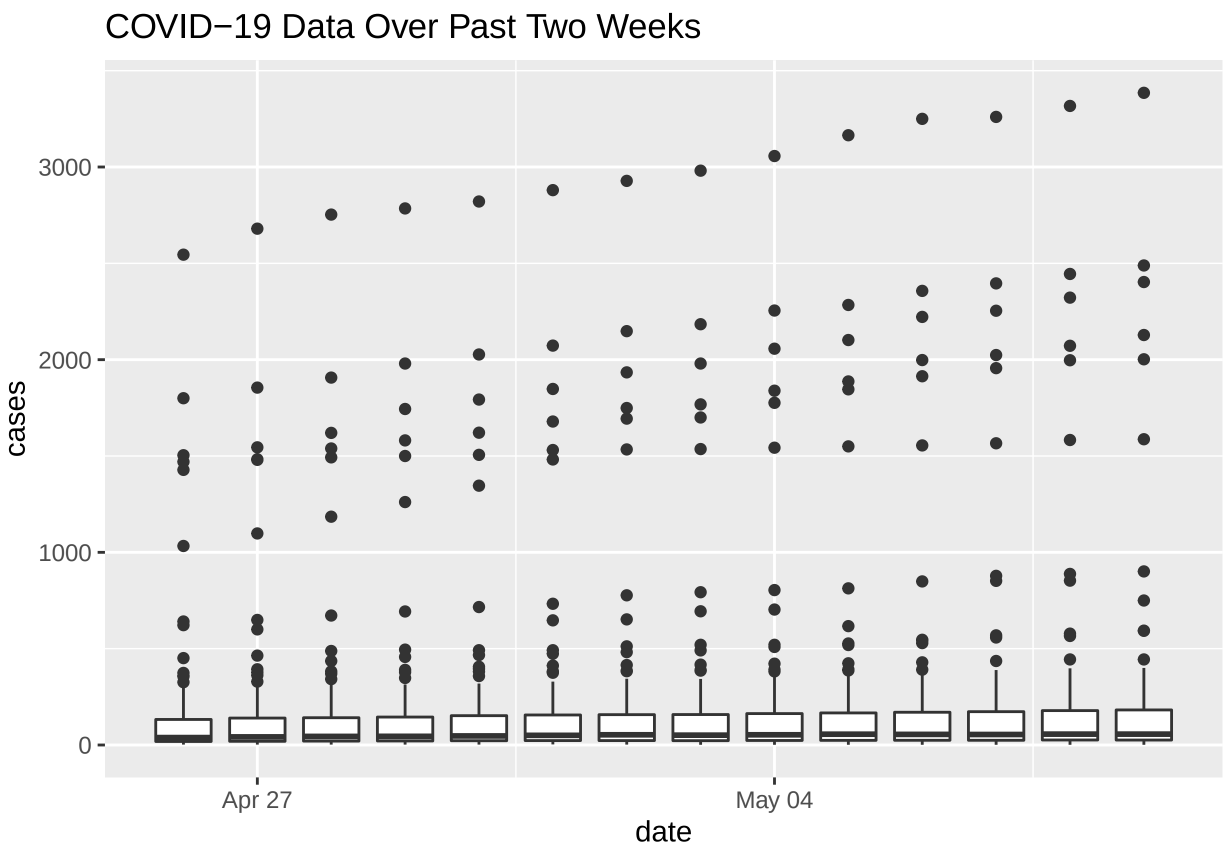
```
# Filter so we are working with the data regarding
# May 9th, 2020
mayninth <- georgia_counties %>% filter(date == "2020-05-09")
histogram <- function(number, bin_color){
  ggplot(data = mayninth,
             mapping = aes(x = cases)) +
    # we can get an idea of how the cases are distributed among all
    # the counties in Georgia by putting 'cases' on the x-axis
geom_histogram(bins = number, fill = bin_color) +
  labs(title =
      "COVID Data (GA)",
       x = "Total Cases",
       y = "# of Counties")
}
g1 <- histogram(10, "black")
g2 <- histogram(30, "red")
g3 <- histogram(50, "blue")

grid.arrange(g1, g2, g3, nrow = 1)
```

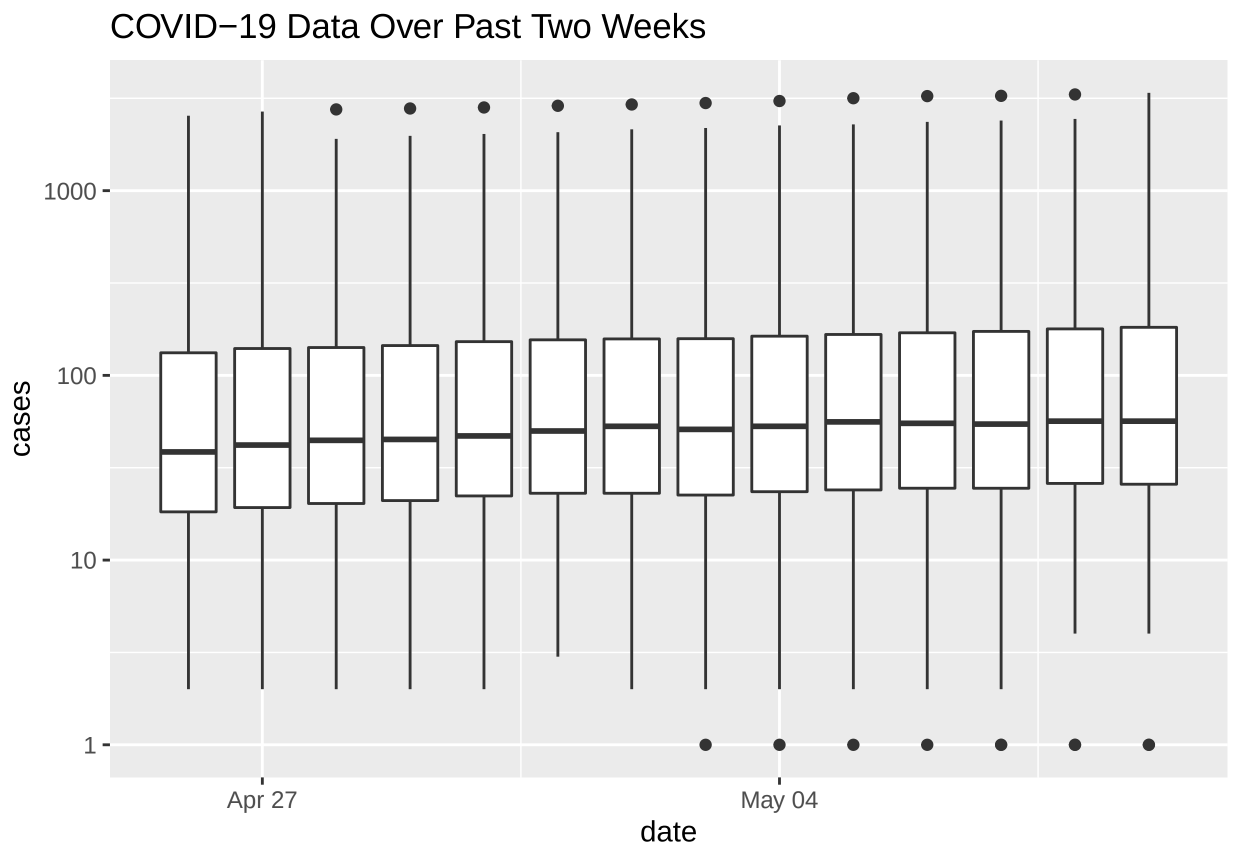


3. A single day doesn't tell the whole story, so you decide to look at the data from the past two weeks, April 26 to May 9, 2020. Boxplots can be easier to interpret than histograms when you are comparing the distributions of multiple groups. Draw boxplots of the total number of confirmed cases in each county by date. Try this with and without a log (base 10) transformation.

```
two_weeks <- georgia_counties %>% filter(date >= "2020-04-26", date <= "2020-05-09")
ggplot(data = two_weeks, aes(x=date, y=cases, group = date)) + geom_boxplot() +
  labs(
    title = "COVID-19 Data Over Past Two Weeks"
  )
```



```
ggplot(data = two_weeks, aes(x=date, y=cases, group = date)) +
  geom_boxplot() + scale_y_log10() +
  labs(
    title = "COVID-19 Data Over Past Two Weeks"
  )
```



4. From your plots in Questions 2 and 3, it is clear that there are some counties with a lot of cases! It might be useful to study them more closely. Identify the five most impacted counties, which we will take to be the counties with the highest case totals on May 9, 2020.

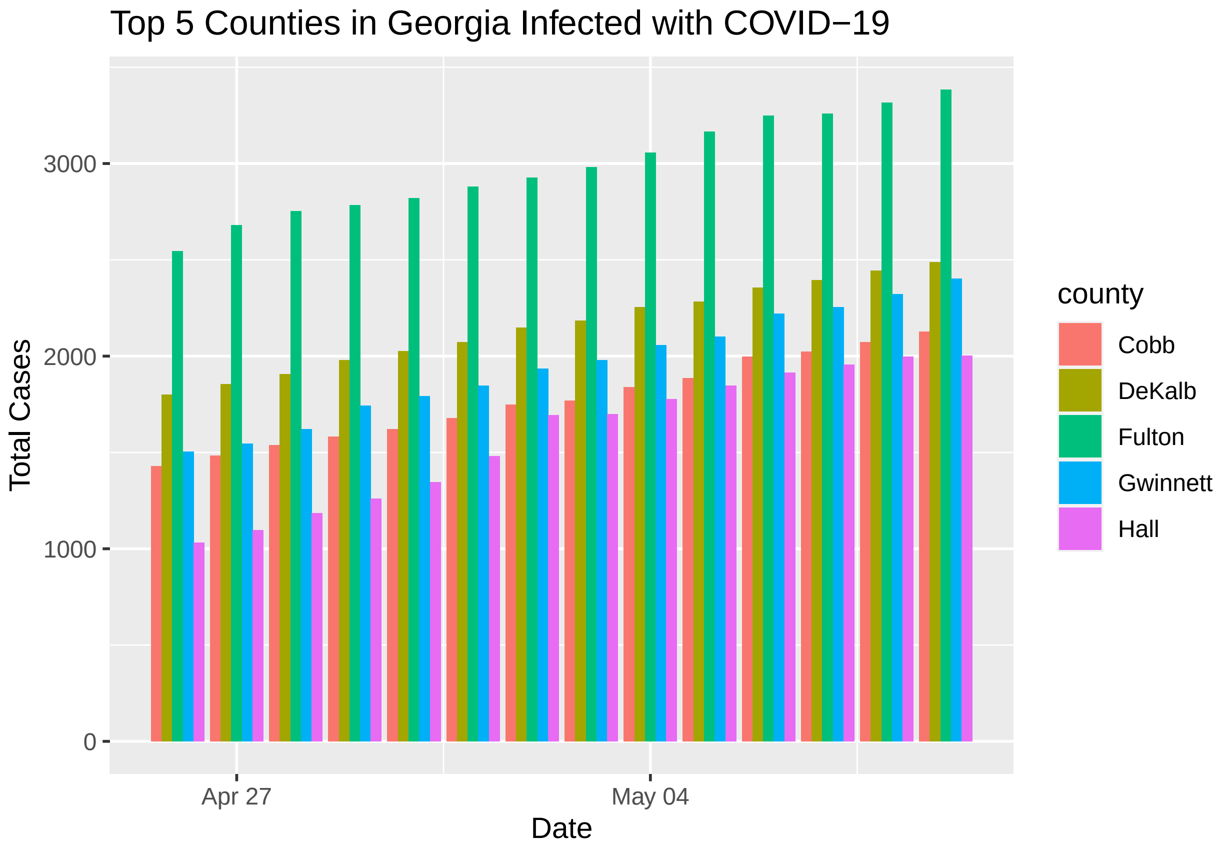
```
d1 <- two_weeks %>% filter(date == "2020-05-09") %>% slice_max(cases, n = 5)
d1
```

```
## # A tibble: 5 x 7
##   date       county   state   fips  cases deaths new_cases
##   <date>     <chr>    <chr>   <chr> <dbl>  <dbl>     <dbl>
## 1 2020-05-09 Fulton   Georgia 13121  3385    144        68
## 2 2020-05-09 DeKalb   Georgia 13089  2489     69        44
## 3 2020-05-09 Gwinnett Georgia 13135  2403     87        81
## 4 2020-05-09 Cobb     Georgia 13067  2128    116        56
## 5 2020-05-09 Hall     Georgia 13139  2002     28         5
```

Fulton, DeKalb, Gwinnett, Cobb, Hall

5. Make an improved clustered barplot of the new cases reported in the 5 most impacted counties between April 26 and May 9. Be sure to order the dates chronologically on the x-axis and maintain the order of the counties within each day's cluster of bars. Does your impression of the COVID-19 situation in Georgia change?

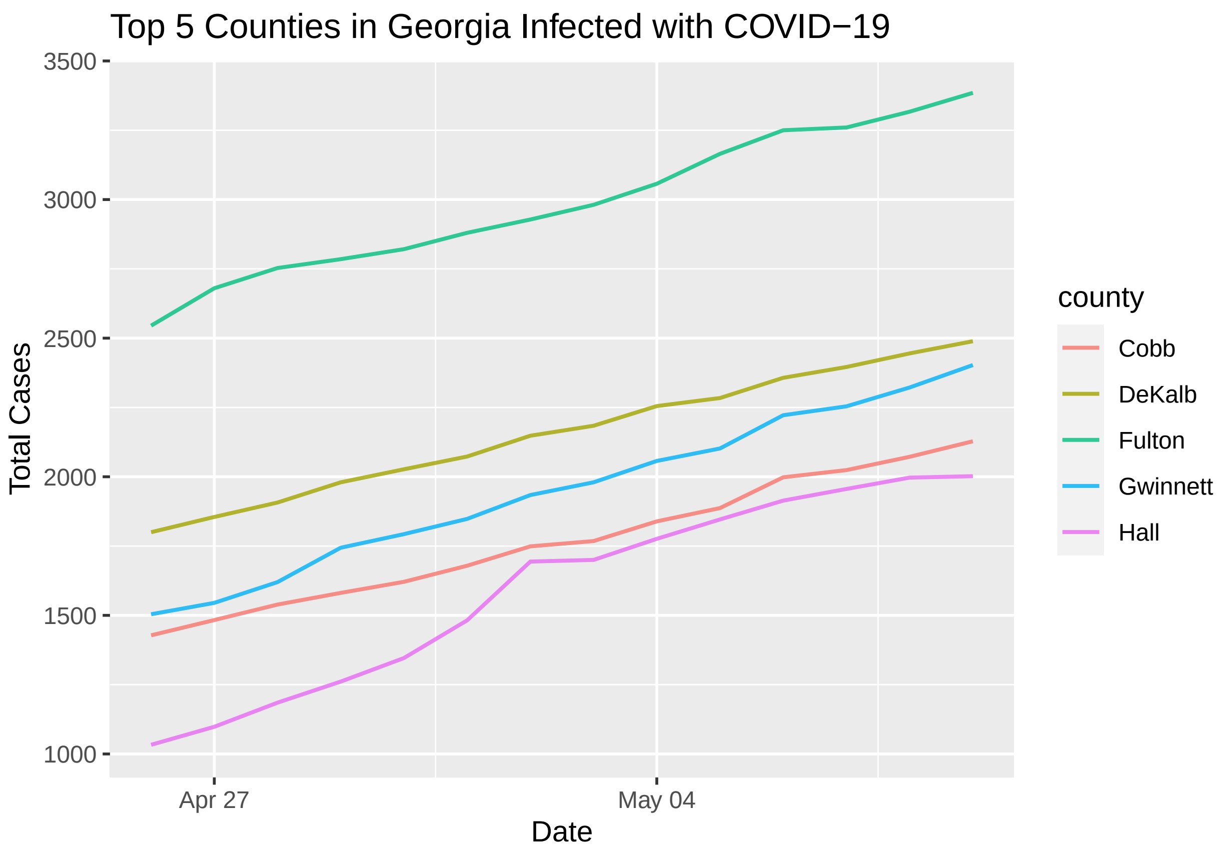
```
two_weeks_filtered <- two_weeks %>% filter(county %in% c("Fulton", "DeKalb", "Gwinnett", "Cobb", "Hall"
ggplot(data = two_weeks_filtered, mapping = aes(x = date,
      y = cases, fill = county)) +
  geom_bar(position = "dodge", stat = "identity") +
  labs(
    title = "Top 5 Counties in Georgia Infected with COVID-19",
    x = "Date",
    y = "Total Cases"
  )
```



Yes, my impression of the COVID-19 situation in Georgia does change with this visualization. We can tell the visualization was done correctly this time since it does not follow a similar pattern to what was officially provided by the Georgia Department of Public Health.

6. While much improved, the clustered barplot still makes it difficult to compare trends over time in the five counties. Present the data as a line plot with the date on the x-axis, the number of new cases on the y-axis, and each county plotted as a separate line.

```
ggplot(data = two_weeks_filtered, mapping = aes(x = date,
      y = cases, color = county)) +
  geom_line(size = 0.7, alpha = 0.8) +
  labs(
    title = "Top 5 Counties in Georgia Infected with COVID-19",
    x = "Date",
    y = "Total Cases"
  )
```


Top 5 Counties in Georgia Infected with COVID−19
3500
3000
2500
2000
1500
1000
Total Cases
Apr 27
May 04
Date
county
Cobb
DeKalb
Fulton
Gwinnett
Hall