DATA 101: Home work 2

Anish Bhurtyal

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.

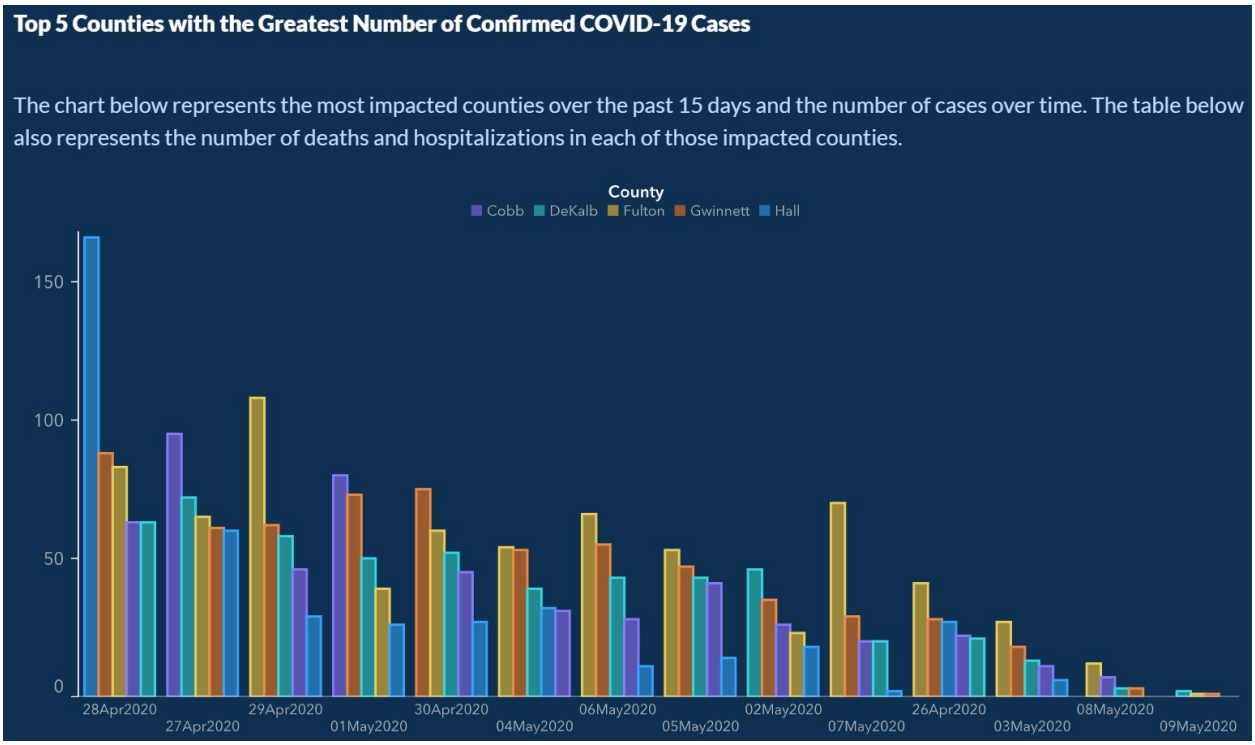


Figure 1: chart of top 5 country with hightest COVID cases

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(tidyverse)
library(gridExtra)

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

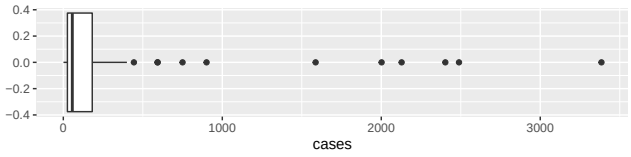
1. Create a new data frame called `georgia_counties` that only contains the data from Georgia. Add a new variable called `new_cases` that stores the number of new confirmed cases for each day at the county level. Hint: the `lag` function returns the previous values in a vector.

```
georgia_counties <- us_counties %>%
  filter(us_counties$state == "Georgia")

georgia_counties <- georgia_counties[order(georgia_counties$county, georgia_counties$date),]

georgia_counties$new_cases = lead(georgia_counties$cases,1) - georgia_counties$cases
georgia_counties
```

```
## # A tibble: 124,889 x 7
##    date       county  state   fips  cases deaths new_cases
##    <date>     <chr>   <chr>   <chr> <dbl>  <dbl>     <dbl>
##  1 2020-03-30 Appling Georgia 13001     2      0        -1
##  2 2020-03-31 Appling Georgia 13001     1      0         0
##  3 2020-04-01 Appling Georgia 13001     1      0         2
##  4 2020-04-02 Appling Georgia 13001     3      0         2
##  5 2020-04-03 Appling Georgia 13001     5      0         0
##  6 2020-04-04 Appling Georgia 13001     5      0         0
##  7 2020-04-05 Appling Georgia 13001     5      0         1
##  8 2020-04-06 Appling Georgia 13001     6      0         0
##  9 2020-04-07 Appling Georgia 13001     6      0         1
## 10 2020-04-08 Appling Georgia 13001     7      0         0
## # ... with 124,879 more rows
```

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.

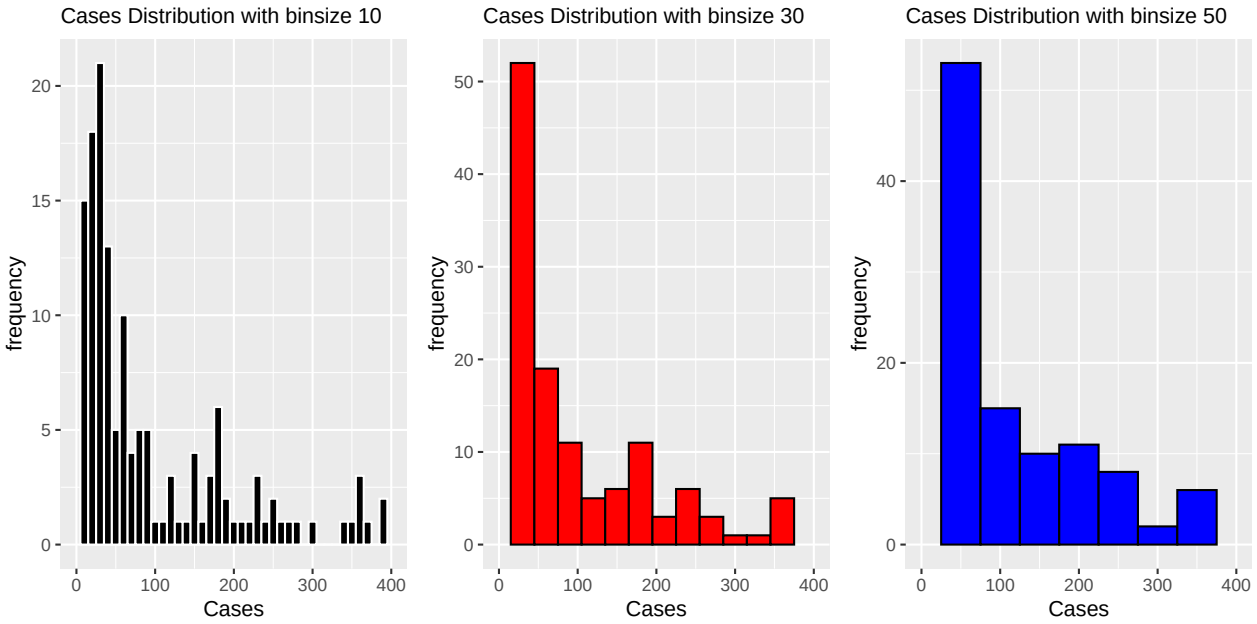
```
may_9_dist <- georgia_counties%>%
  filter(georgia_counties$date == "2020-05-09")

  ggplot(may_9_dist, aes(x = cases)) +
  geom_boxplot()
```



Here we observe that the outliers are counties with cases values >400 cases, so we will see the distribution better way in histogram with limiting values (0,400)

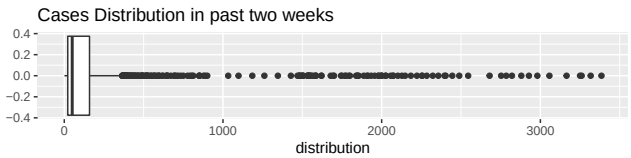
```
library(gridExtra)
  bin10 <- ggplot(may_9_dist, aes(x=cases))+
    geom_histogram(binwidth = 10,fill="black", col = 'white')+
    xlim(0,400)+
    labs(subtitle = "Cases Distribution with binsize 10",
         x = "Cases" ,
         y= "frequency")

 bin30 <- ggplot(may_9_dist, aes(x=cases ))+
    geom_histogram(binwidth = 30, fill="red", col = 'black')+
```

```
    xlim(0,400)+
   labs(subtitle = "Cases Distribution with binsize 30",
         x = "Cases" ,
         y= "frequency")

 bin50 <- ggplot(may_9_dist, aes(x=cases))+
    geom_histogram(binwidth = 50, fill = "blue", col = 'black')+
    xlim(0,400)+
   labs(subtitle = "Cases Distribution with binsize 50",
         x = "Cases" ,
         y= "frequency")

grid.arrange(bin10,bin30,bin50, ncol = 3)
```

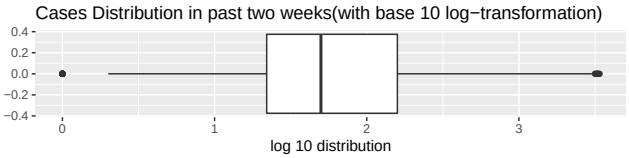


3. A single day doesn't tell the whole story, so you decide to look at the data from the past two weeks, April 26 to May 9, 2020. Boxplots can be easier to interpret than histograms when you are comparing the distributions of multiple groups. Draw boxplots of the total number of confirmed cases in each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

```
past_two_weeks <- georgia_counties%>%
  filter(georgia_counties$date >= "2020-04-26" & georgia_counties$date <= "2020-05-09" )

ggplot(past_two_weeks, aes( x=cases)) +
  geom_boxplot()+
  labs(title = "Cases Distribution in past two weeks",
         x = "distribution" ,
         y= "")
```



```
past_two_weeks$logCases = log10(past_two_weeks$cases)
ggplot(past_two_weeks, aes(x = logCases)) +
  geom_boxplot()+
  labs(title = "Cases Distribution in past two weeks(with base 10 log-transformation)",
         x = "log 10 distribution" ,
         y= "")
```



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.

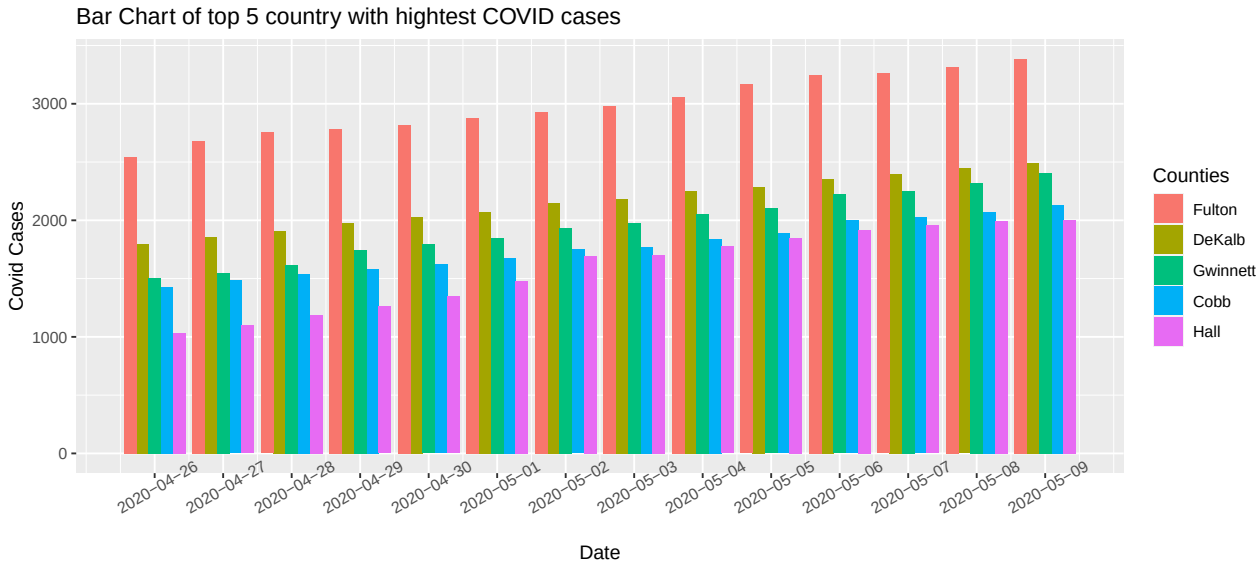
```
top_5 <- may_9_dist %>%
  arrange(desc(cases)) %>%
  slice(1:5)

top_5[[2]]
```

```
## [1] "Fulton"   "DeKalb"   "Gwinnett" "Cobb"     "Hall"
```

Here are the 5 most impacted counties on May 9 2020

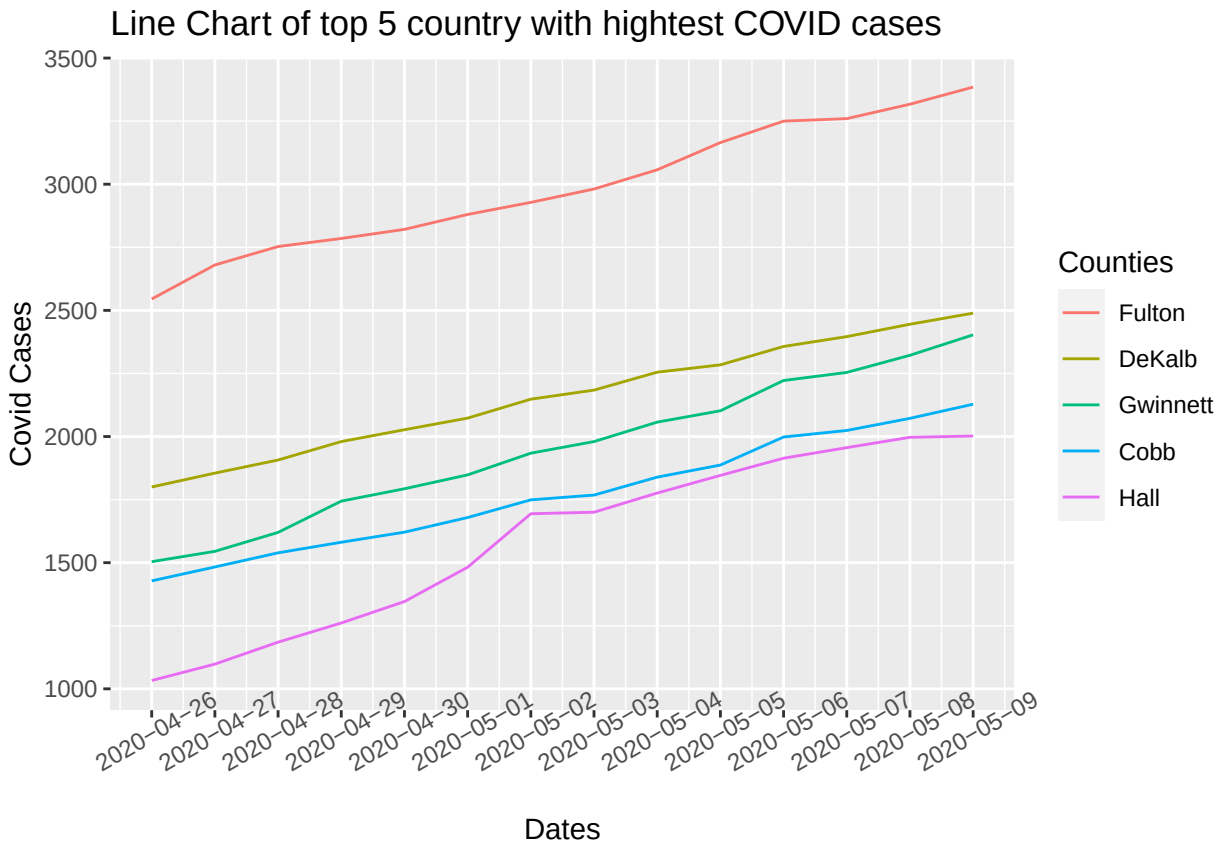
5. Make an improved clustered barplot of the new cases reported in the 5 most impacted counties between April 26 and May 9. Be sure to order the dates chronologically on the x-axis and maintain the order of the counties within each day's cluster of bars. Does your impression of the COVID-19 situation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

```
datebreaks <- seq(as.Date("2020-04-26"), as.Date("2020-05-09"), by = "1 day")

past_two_weeks%>%
  #filter(county == top_5$county) %>%
  filter(county ==  "DeKalb" |county ==  "Gwinnett" | county =="Cobb" |
           county == "Fulton" | county ==    "Hall") %>%
  #arrange(cases) %>%
  ggplot( aes(x= date, y= cases, fill= reorder(county, -cases)))+
  geom_bar(position = "dodge", stat = "identity")+
  scale_x_date(breaks = datebreaks)+
  theme(axis.text.x = element_text(angle = 30))+
  labs(
    y = "Covid Cases",
    x = "Date",
    fill = "Counties",
    title = "Bar Chart of top 5 country with hightest COVID cases"
  )
```



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.

```
past_two_weeks%>%
  #filter(county == top_5$county) %>%
  filter(county ==  "DeKalb" |county ==  "Gwinnett" | county =="Cobb" |
           county == "Fulton" | county ==    "Hall") %>%
  #arrange(cases) %>%
  ggplot( aes(x= date, y= cases, color = reorder(county, -cases)))+
  geom_line()+
  scale_x_date(breaks = datebreaks)+
  theme(axis.text.x = element_text(angle = 30))+
  labs(
    y = "Covid Cases",
    x = "Dates",
    color = "Counties",
    title = "Line Chart of top 5 country with hightest COVID cases",
  )
```


Line Chart of top 5 country with hightest COVID cases
Covid Cases
3500
3000
2500
2000
1500
1000
2020−04−26
2020−04−27
2020−04−28
2020−04−29
2020−04−30
2020−05−01
2020−05−02
2020−05−03
2020−05−04
2020−05−05
2020−05−06
2020−05−07
2020−05−08
2020−05−09
Dates
Counties
Fulton
DeKalb
Gwinnett
Cobb
Hall