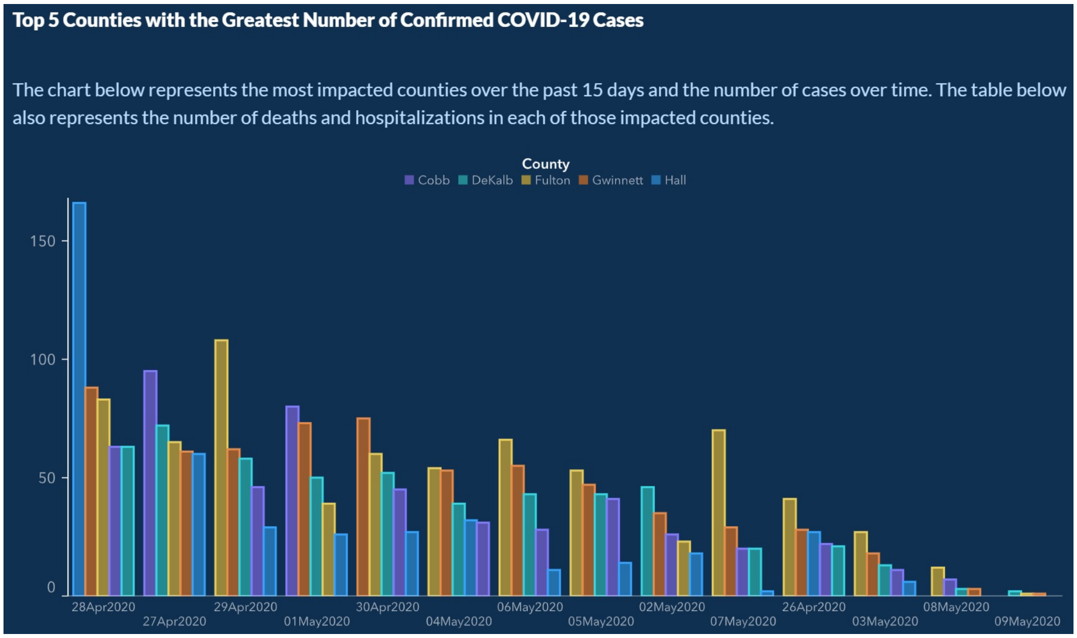
Joseph Bruch BST 270 Individual Project

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 a two-week period. Health officials claimed that the plot provided evidence that COVID-19 cases were decreasing and made the argument for reopening the state.



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

Data visualization is an incredibly powerful tool that can affect health policy decisions. Ensuring they are easy to interpret, and more importantly, showcase accurate insights from data is paramount for scientific transparency and the health of individuals. For this assignment you are tasked with reproducing COVID-19 visualizations and tables published by the New York Times (https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html). Specifically, you will attempt to reproduce the following for January 17th, 2021:

1. New cases as a function of time with a rolling average plot - the first plot on the page (you don't need to recreate the colors or theme)
2. Table of cases, hospitalizations and deaths - the first table on the page
3. The county-level map for previous week ('Hot spots') - the second plot on the page (only the 'Hot Spots' plot)
4. Table of cases by state - the second table on the page (do not need to include per 100,000 or per capita columns)

Data for cases and deaths can be downloaded from this NYT GitHub repository (https://github.com/nytimes/covid-19-data) (use `us-counties.csv`). Data for hospitalizations can be downloaded from The COVID Tracking Project (https://covidtracking.com/data). The project must be submitted in the form of a Jupyter notebook or RMarkdown file and corresponding compiled/knitted PDF, with commented code and text interspersed, including a **brief critique of the reproducibility of each plot and table**. All project documents must be uploaded to a GitHub repository each student will create within the reproducible data science organization (https://github.com/reproducibleresearch). The repository must also include a README file describing the contents of the repository and how to reproduce all results. You should keep in mind the file and folder structure we covered in class and make the reproducible process as automated as possible.

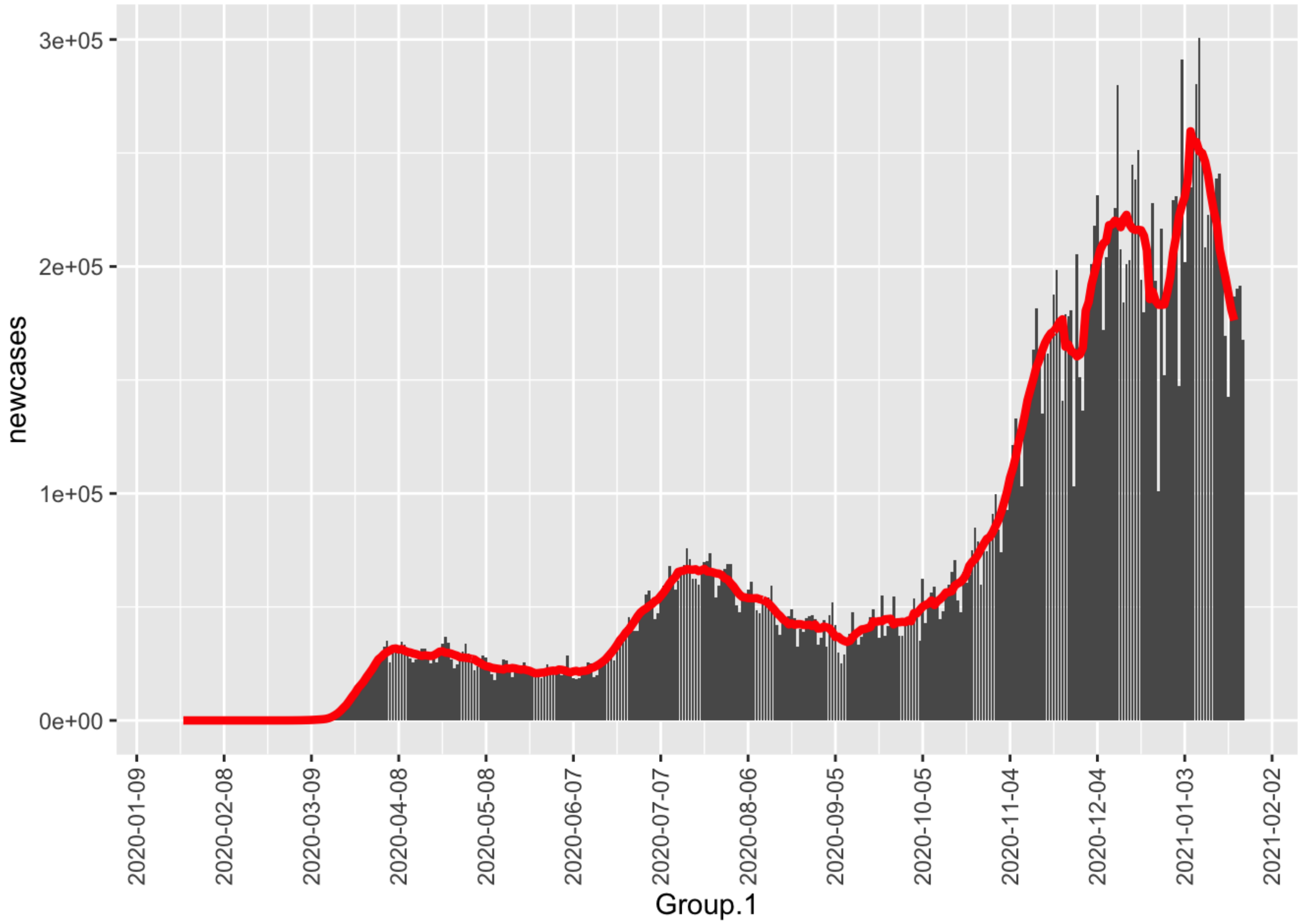
```
us_counties <- data <- read.csv("https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv")
#us_counties = us_counties[!us_counties$state=="Guam",]
#us_counties = us_counties[!us_counties$state=="Puerto Rico",]
#us_counties = us_counties[!us_counties$state=="Northern Mariana Islands",]


cases<-aggregate(us_counties$cases, by=list(us_counties$date), FUN=sum,na.rm=TRUE)
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

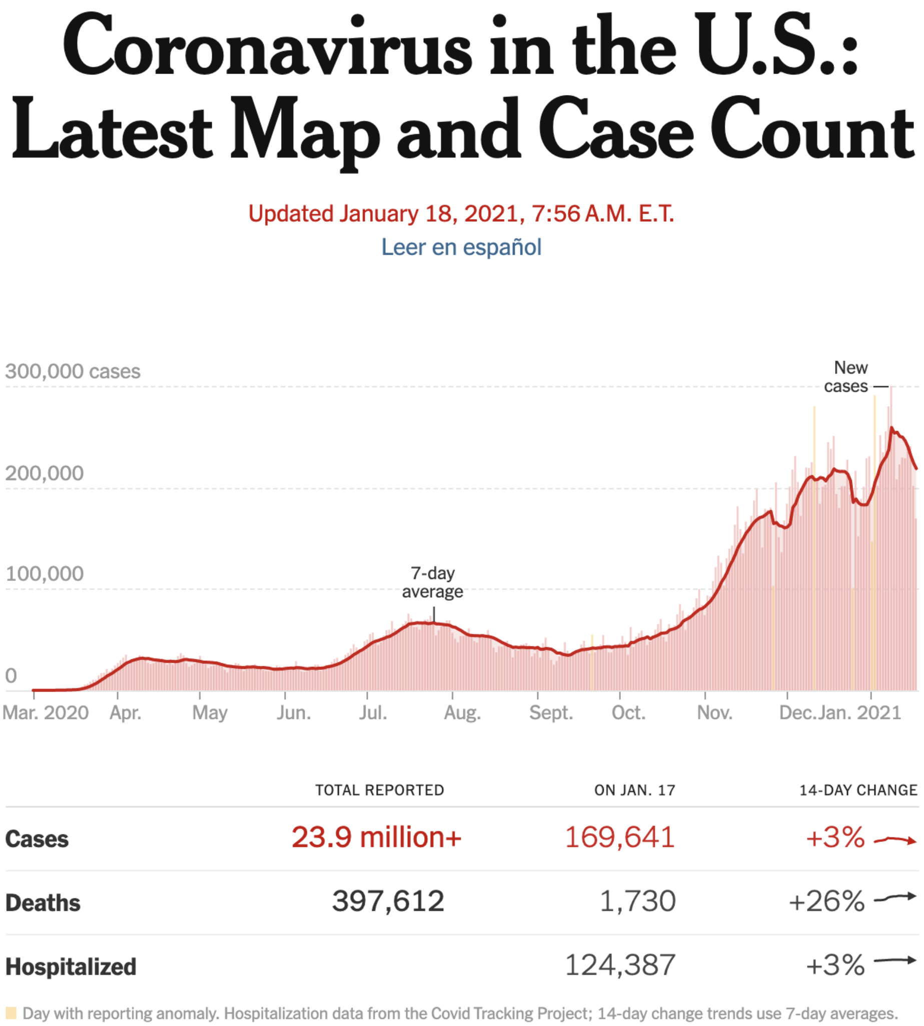
```
cases$lag<-shift(as.numeric(cases$x))
cases$newcases <- as.numeric(cases$x) - cases$lag

cases$new_cases_7dayavg = rollmean(cases$newcases, k = 7, fill = NA)



cases$Group.1 <- as.Date(cases$Group.1)

ggplot(aes(x=Group.1, y= newcases), data = cases) +
  geom_bar(stat = 'identity', position = 'dodge') +
  scale_x_date(breaks = '30 days') +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))+
geom_line(aes(x = Group.1, y = new_cases_7dayavg), size = 1.5, color="red", group = 1)
```



Critique: I was able to replicate the NYT results. I did attempt to try the color scheme that they created but was unsuccesful. Additionally, I found it difficult to discern whether we should include January 18th, 2021 on onward. I decided to include these days after going back and forth. Additionally, I did not know whether to include Marina Islands, Puerto Rico, and Guam. I kept them in but couldn't figure it out on the NYT website.

```
all_states_history<-read.csv("https://covidtracking.com/data/download/all-states-history.csv")
library(data.table)
eTable = data.frame(matrix(
nrow=3,              # number of rows
ncol=3))      # fill matrix by rows

colnames(eTable) <- c("Total Reported","On Jan. 17","14-Day Change")
rownames(eTable) <- c("Cases","Deaths","Hospitalized")
eTable[1,1]<-("23.9 million+") #I couldnnt calculate a plus
eTable[1,2]<-cases$newcases[cases$Group.1=="2021-01-17"]
eTable[1,3]<-paste((1-(cases$new_cases_7dayavg[cases$Group.1=="2021-01-18"]/cases$new_cases_7dayavg[cases$Group.1=="2021-01-03"]))*100*-1,"%")


deaths<-aggregate(us_counties$deaths, by=list(us_counties$date), FUN=sum,na.rm=TRUE)

deaths$lag<-shift(as.numeric(deaths$x))
deaths$newdeaths <- as.numeric(deaths$x) - deaths$lag

deaths$new_death_7dayavg = rollmean(deaths$newdeaths, k = 7, fill = NA)



eTable[2,1]<-sum(deaths$newdeaths,na.rm=TRUE) #I couldnnt calculate a plus
eTable[2,2]<-deaths$newdeaths[cases$Group.1=="2021-01-17"]
eTable[2,3]<-paste((deaths$new_death_7dayavg[cases$Group.1=="2021-01-17"]/deaths$new_death_7dayavg[cases$Group.1=="2021-01-03"]-1)*100,"%")

eTable[3,2]<-sum(all_states_history$hospitalizedCurrently[all_states_history$date=="2021-01-17"],na.rm=TRUE)


hospitalized<-aggregate(all_states_history$hospitalizedCurrently, by=list(all_states_history$date), FUN=sum,na.rm=TRUE)

hospitalized$lag<-shift(as.numeric(hospitalized$x))
hospitalized$newhospitalized <- as.numeric(hospitalized$x) - hospitalized$lag

hospitalized$new_hospitalized_7dayavg = rollmean(hospitalized$newhospitalized , k = 7, fill = NA)
eTable[3,3]<- paste((hospitalized$new_hospitalized_7dayavg[hospitalized$Group.1=="2021-01-17"]/hospitalized$new_hospitalized_7dayavg[hospitalized$Group.1=="2021-01-03"]+1)*100,"%")

eTable
```

```
##              Total Reported On Jan. 17      14-Day Change
## Cases         23.9 million+     169641 -18.416369059603 %
## Deaths               417390       1730  13.703073487185 %
## Hospitalized           <NA>     124387 -6.0242624758754 %
```



Coronavirus in the U.S.: Latest Map and Case Count

Updated January 18, 2021, 7:56 A.M. E.T.

Leer en español

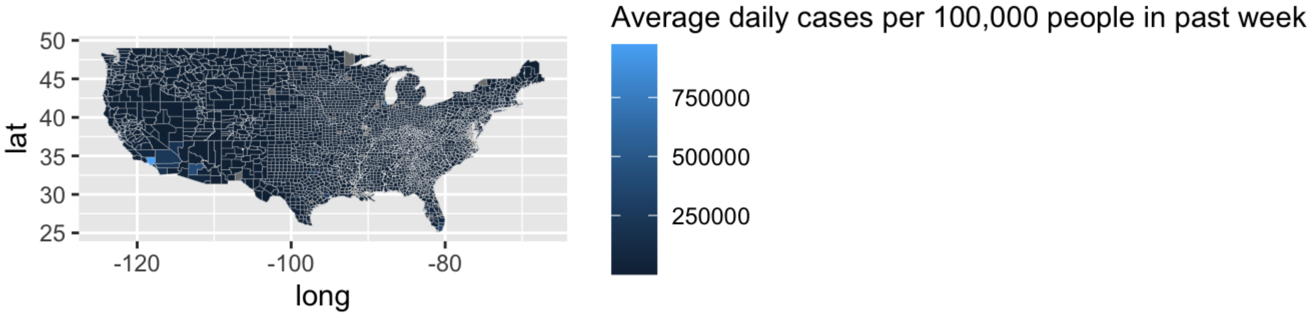
	TOTAL REPORTED	ON JAN. 17	14-DAY CHANGE
Cases	23.9 million+	169,641	+3% →
Deaths	397,612	1,730	+26% →
Hospitalized		124,387	+3% →

Day with reporting anomaly. Hospitalization data from the Covid Tracking Project; 14-day change trends use 7-day averages.

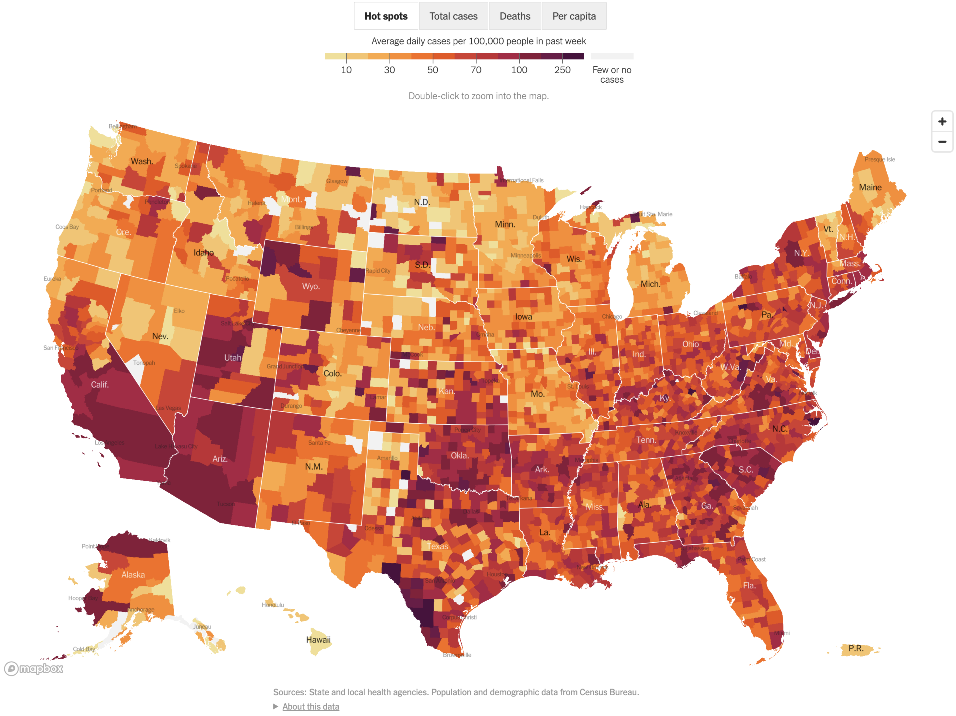
Critique: Most of the values are the same.I was unable to get a rounding average with + for the cases total reported, but I got a similar value unrounded. Unfortunately, I was not sure whether to include several of the locations (Puerto Rico, Guam,and Marina Islands), so I think some of the values were slightly different. My January 17th numbers (col 2) confirm to the NYT's. However, the 14-day average is different. I don't see how they see positive increases since the graph shows that Jan 18th rolling average is lower than the 4th.

3. The county-level map for previous week ('Hot spots') - the second plot on the page (only the 'Hot Spots' plot)

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

library(dplyr)

# Retrieve the states map data and merge with crime data
states_map <- map_data("county")
states_map$state<-states_map$region
states_map$county<-states_map$subregion

# Create the map
states_map$state<-states_map$region
states_map$county<-states_map$subregion
us_counties$state <- tolower(us_counties$state)
us_counties$county<-tolower(us_counties$county)
states_map$state <- tolower(states_map$state)
states_map$county<-tolower(states_map$county)



us_counties<-subset(us_counties,us_counties$date=="2021-01-17"|
us_counties$date=="2021-01-16"|
us_counties$date=="2021-01-15"|
us_counties$date=="2021-01-14"|
us_counties$date=="2021-01-13"|
us_counties$date=="2021-01-12"|
us_counties$date=="2021-01-11")

us_counties<-aggregate(us_counties$cases, by=list(us_counties$state,us_counties$county),
FUN=mean,na.rm=TRUE)

colnames(us_counties) <- c("state","county","cases")


states_map<-states_map %>% left_join(us_counties, by=c("state","county"))
p <- ggplot(data = states_map,
            mapping = aes(x = long, y = lat,
                          fill = cases,
                          group = group))

p1 <- p + geom_polygon(color = "gray90", size = 0.05) + coord_equal()

p2<-p1 + labs(fill = "Average daily cases per 100,000 people in past week")

p2
```



I was able to replicate the graph in the NYT. However, because the county population data was not supplied, I did not adjust for size as shown in the graph. Even though it was not graded, I found this actually very useful since I have for a while been procrastinating in learning how to make county-level ggplots.



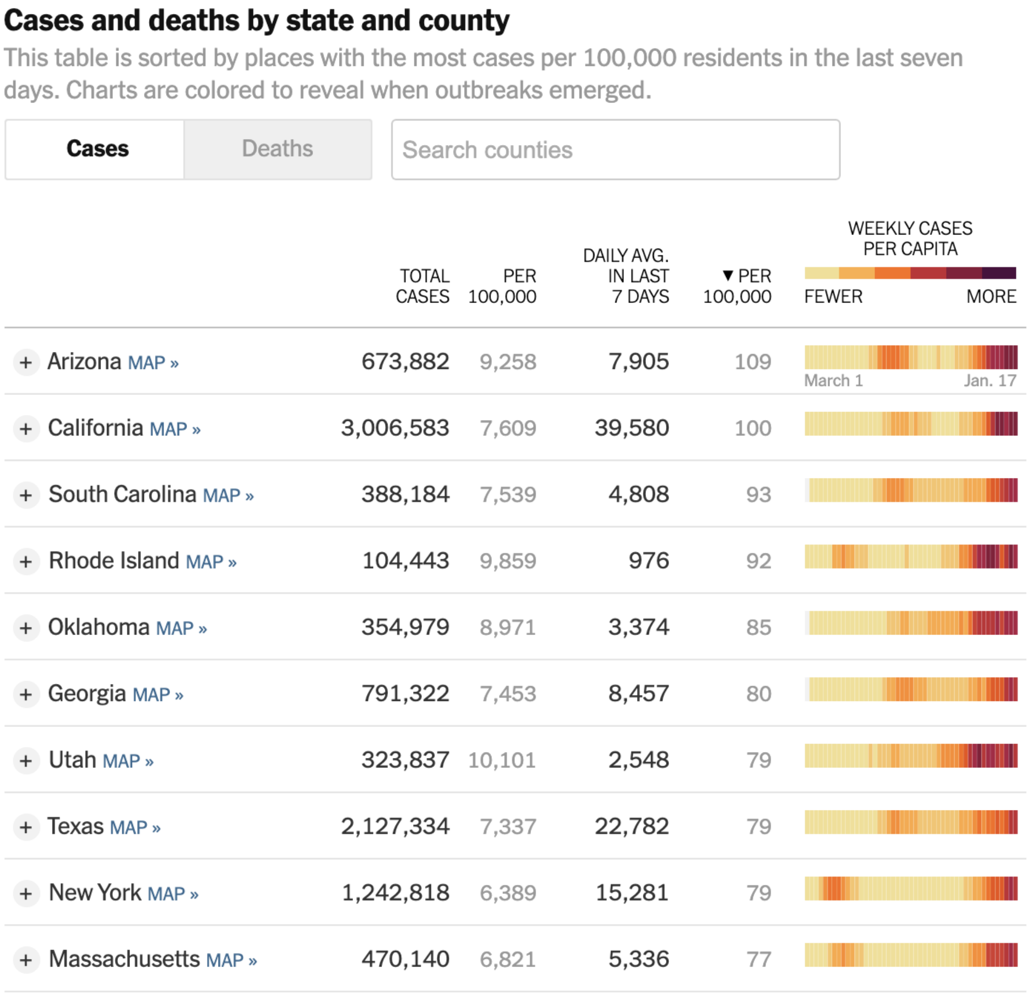
Sources: State and local health agencies. Population and demographic data from Census Bureau.

▸ About this data

4. Table of cases by state - the second table on the page (do not need to include per 100,000 or per capita columns)

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

# Retrieve the states map data and merge with crime data
us_states<-subset(us_counties,us_counties$date=="2021-01-17")
us_states<-aggregate(us_states$cases, by=list(us_states$state), FUN=sum,na.rm=TRUE)

#Now I get daily average


us_states2<-subset(us_counties,us_counties$date=="2021-01-17"|
us_counties$date=="2021-01-16"|
us_counties$date=="2021-01-15"|
us_counties$date=="2021-01-14"|
us_counties$date=="2021-01-13"|
us_counties$date=="2021-01-12"|
us_counties$date=="2021-01-11"|
us_counties$date=="2021-01-10"|
us_counties$date=="2021-01-09")



us_states2<-aggregate(us_states2$cases, by=list(us_states2$state,us_states2$date), FUN=s
um,na.rm=TRUE)


colnames(us_states2) <- c("state","date","cases")

us_states2<-reshape(us_states2, idvar = "state", timevar = "date", direction = "wide")

us_states2$avgchange<-((us_states2$`cases.2021-01-17`-
  us_states2$`cases.2021-01-16`)+
  (us_states2$`cases.2021-01-16`-
  us_states2$`cases.2021-01-15`)+
  (us_states2$`cases.2021-01-15`-
  us_states2$`cases.2021-01-14`)+
  (us_states2$`cases.2021-01-14`-
  us_states2$`cases.2021-01-13`)+
  (us_states2$`cases.2021-01-13`-
  us_states2$`cases.2021-01-12`)+
  (us_states2$`cases.2021-01-12`-
  us_states2$`cases.2021-01-11`)+
  (us_states2$`cases.2021-01-11`-
  us_states2$`cases.2021-01-10`))/7

us_states2<-us_states2 %>% select(avgchange, state)
colnames(us_states) <- c("State","Total Cases")
colnames(us_states2) <- c("DAILY AVG. IN LAST 7 DAYS","State")
us_states2[,c("State")]<-NULL
etable4<-cbind(us_states,us_states2)
head(etable4)
```

```
##          State Total Cases DAILY AVG. IN LAST 7 DAYS
## 1      Alabama      422598                 2956.8571
## 2       Alaska       51630                  242.1429
## 3      Arizona      673882                 7905.1429
## 4     Arkansas      271154                 2296.8571
## 5   California     3006583                39579.7143
## 6     Colorado      376921                 1986.0000
```



Cases and deaths by state and county

This table is sorted by places with the most cases per 100,000 residents in the last seven days. Charts are colored to reveal when outbreaks emerged.

	TOTAL CASES	PER 100,000	DAILY AVG. IN LAST 7 DAYS	▼ PER 100,000	WEEKLY CASES PER CAPITA
+ Arizona MAP »	673,882	9,258	7,905	109	
+ California MAP »	3,006,583	7,609	39,580	100	
+ South Carolina MAP »	388,184	7,539	4,808	93	
+ Rhode Island MAP »	104,443	9,859	976	92	
+ Oklahoma MAP »	354,979	8,971	3,374	85	
+ Georgia MAP »	791,322	7,453	8,457	80	
+ Utah MAP »	323,837	10,101	2,548	79	
+ Texas MAP »	2,127,334	7,337	22,782	79	
+ New York MAP »	1,242,818	6,389	15,281	79	
+ Massachusetts MAP »	470,140	6,821	5,336	77	

Critique: I got the same results for each of the values and as instructed did not include per capita values. I found this challenging since I was not sure how to use the zoo package to calculate average changes within a state. This was generally, however, doable using my own code and not utilizing a function.