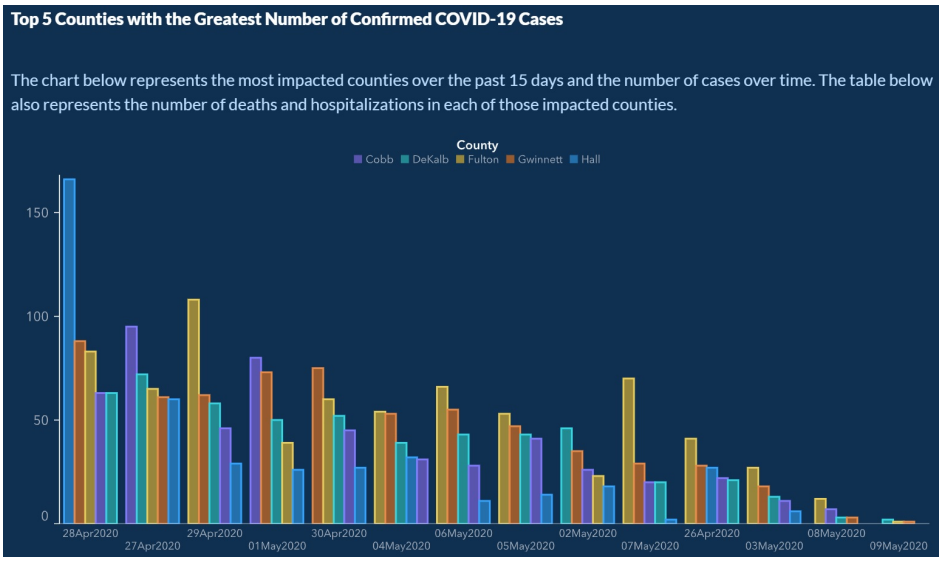
BST 270 Individual Project

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.



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. 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 (use `us-counties.csv`). Data for hospitalizations can be downloaded from The COVID Tracking Project. 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. 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.

Tips:

- You can extract the number of new cases from the case totals using the `lag` function. In this toy example, cases records the daily total/cumulative number of cases over a two-week period. By default, the lag function simply shifts the vector of cases back by one. The number of new cases on each day is then the difference between `cases` and `lag(cases)`.

```
cases = c(13, 15, 18, 22, 29, 39, 59, 61, 62, 67, 74, 89, 108, 122)
new_cases = cases - lag(cases)
new_cases
```

```
##  [1] NA  2  3  4  7 10 20  2  1  5  7 15 19 14
```

- You can write your own function to calculate a seven-day rolling average, but the `zoo` package already provides the `rollmean` function. Below, the `k = 7` argument tells the function to use a rolling window of seven entries. `fill = NA` tells `rollmean` to return `NA` for days where the seven-day rolling average can't be calculated (e.g. on the first day, there are no days that come before, so the sliding window can't cover seven days). That way, `new_cases_7dayavg` will be the same length as `cases` and `new_cases`, which would come in handy if they all belonged to the same data frame.

```
new_cases_7dayavg = rollmean(new_cases, k = 7, fill = NA)
new_cases_7dayavg
```

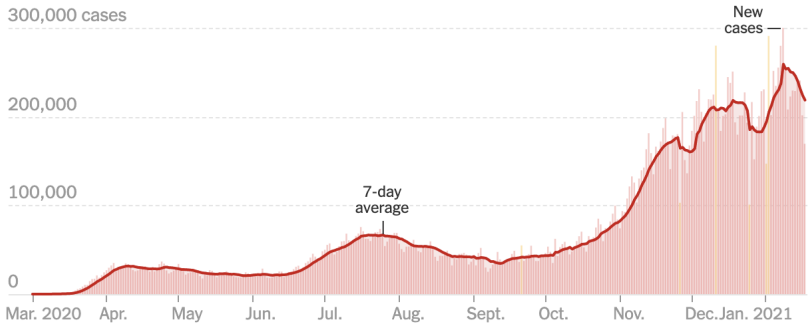
```
##  [1]       NA       NA       NA       NA 6.857143 6.714286 7.000000 7.428571
##  [9] 8.571429 9.857143 9.000000       NA       NA       NA
```

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

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

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

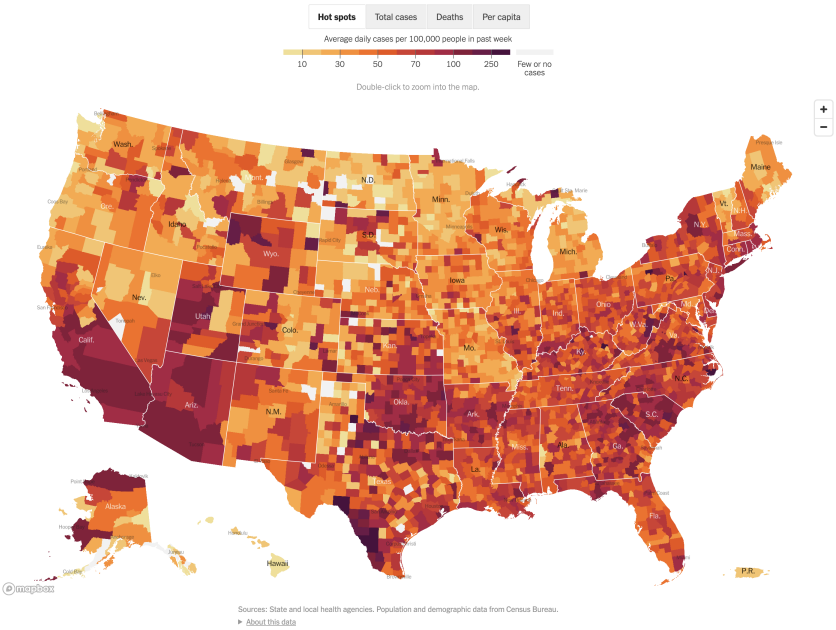
Leer en español



	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.

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



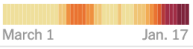
4. Table of cases by state - the second table on the page (do not need to include per 100,000 or per capita

columns)

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.

Cases | Deaths | Search counties

	TOTAL CASES	PER 100,000	DAILY AVG. IN LAST 7 DAYS	▼ PER 100,000	WEEKLY CASES PER CAPITA (FEWER – MORE; March 1 – Jan. 17)
+ 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