Jordan Roberts

July 15, 2020

BAN 530

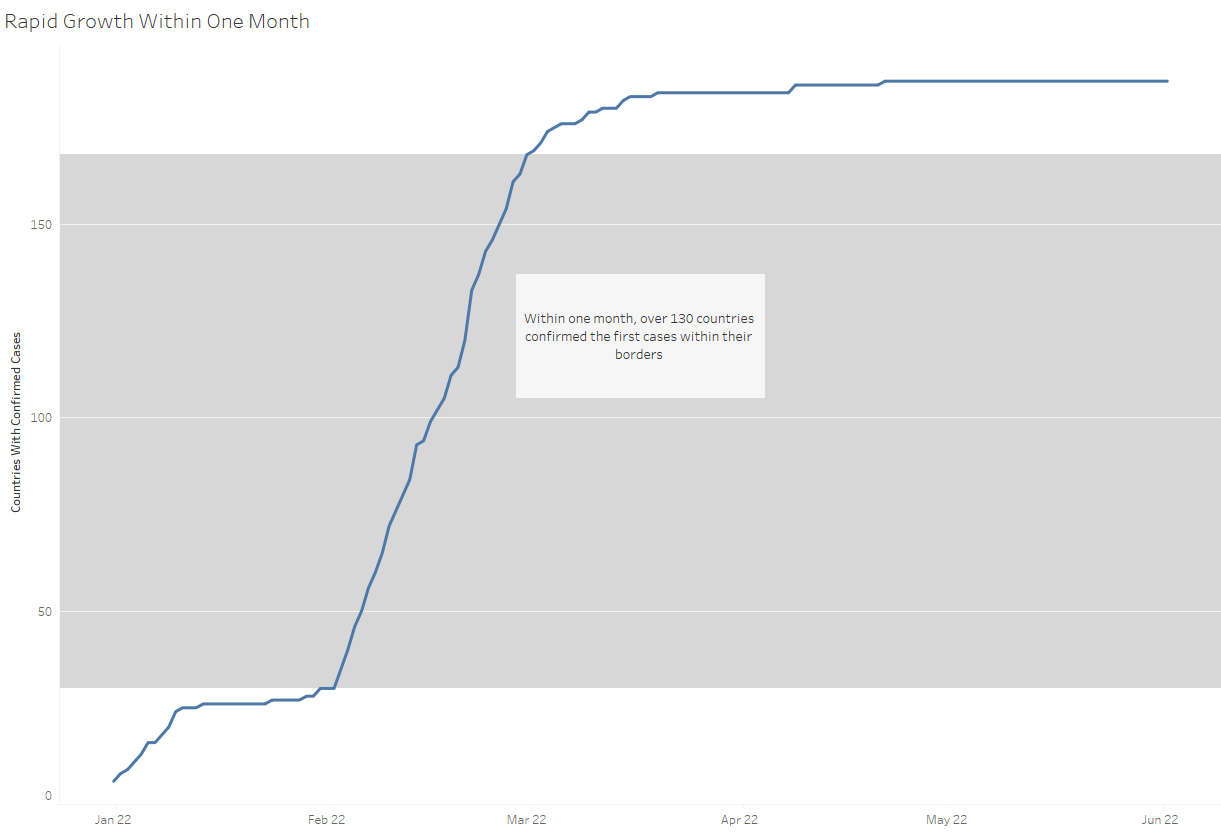
Data Cleaning/Descriptive Analysis

**Data Cleaning**

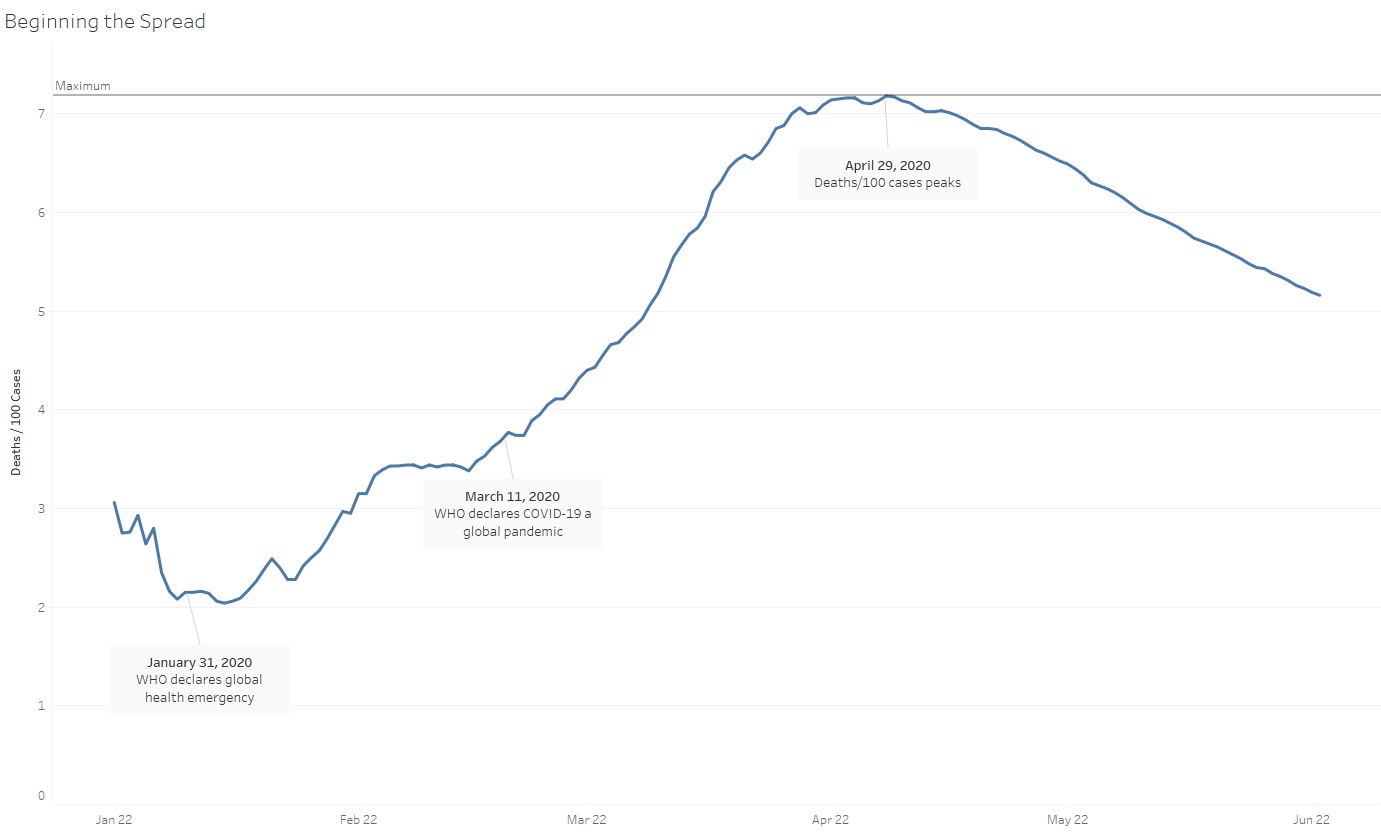
These six datasets are set up quite similarly, but there is still some modification needed for my purposes. For the US Counties dataset, I added two columns: Census Region (based on the US Census, either Northeast, South, Midwest, or West US) and Division (a further breakdown of each Region, also based on the US Census). I also included designations for the different US territories and the two cruise ships. US territories and cruise ships had blank fields in “Admin2”, which I renamed more appropriately to “County”. Otherwise, there were no blank fields in this dataset.

The “worldometer\_data” dataset had numerous blank fields. In most cases, it is assumed it was a stand-in for zero, but some countries had no associated WHO Region while others had no associated Continent (cruise ships). Entries with no WHO Region turned out to be island municipalities, technically under rule by foreign countries but far enough away that they do not quite fit in that region of the world. In the end, I did not use WHO Region from the Worldometer set in Tableau, as it was available in the Country-Latest dataset. I filled numerical entries with zeroes. Worldometer also excluded several countries such as China (global epicenter of the outbreak), so it was difficult to gather anything useful from the set.

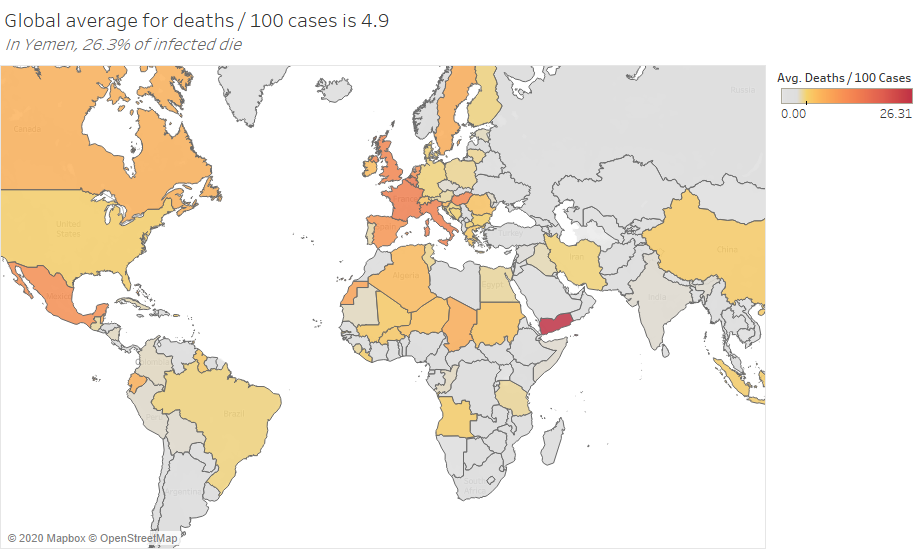
**Worldwide Data**



Cases started slowly spreading throughout the world until February 22. Within a month, 139 new countries had documented COVID-19 cases. Afterward, it climbed slowly to 187 countries by early May.

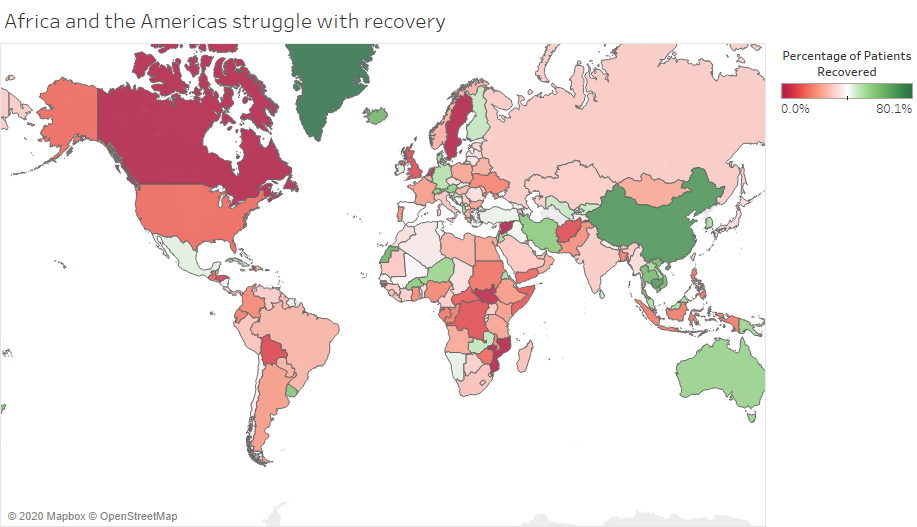


By January 31, 24 countries had infections. The WHO declared a global health emergency as cases and deaths began to rise. On the 31st, COVID was killing 2.15 people for every 100 cases. When the WHO declared COVID-19 a pandemic, that number had risen by 71%. By the end of April, it had almost doubled and almost every country was infected. After the 29th, deaths per 100 cases dropped off steadily. More cases were being reported than deaths as the rest of the world started to take notice.

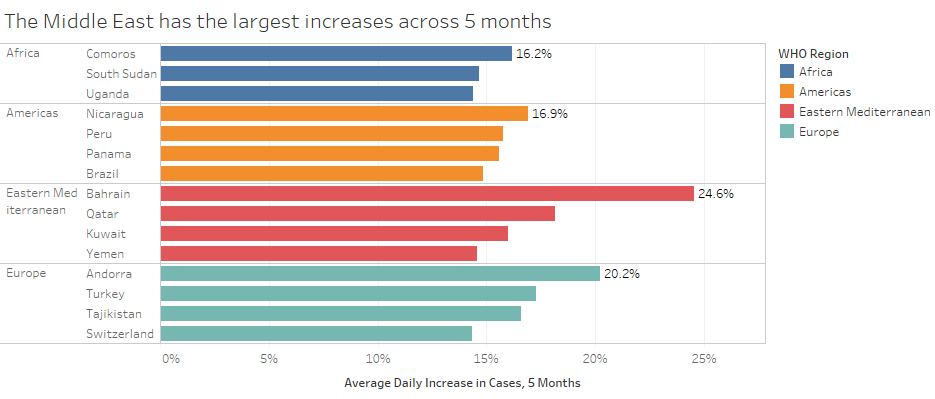


The global average for deaths was 4.9 on June 22, 2020. Amounts at the global average are mapped in gold, with Yemen and other red countries suffering higher fatality rates. In Yemen, 26.3 patients are dying with every 100 cases. Much of Western Europe, notably France, Belgium, and Italy, are also struggling. A large swath of Saharan African countries also struggles to keep patients alive, as well as China and Indonesia. Much of the Middle East and South America is at or below the global average.

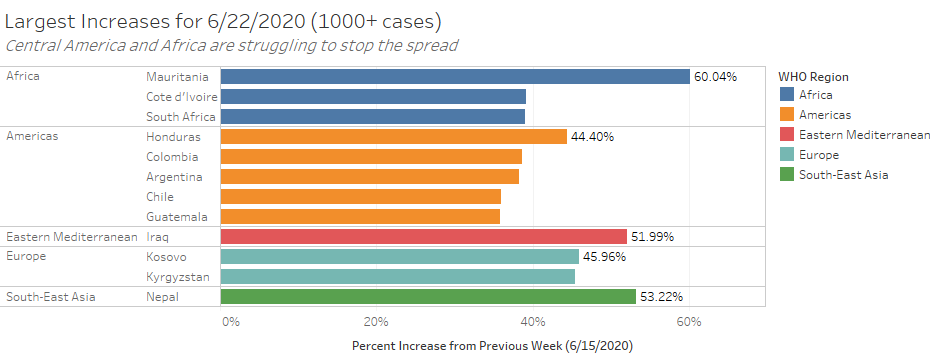
Adjusted for population, the Americas and Europe have the highest deaths overall. The top 15 countries for deaths relative to population are European or American. The US is the only non-European country in the top 10. San Marino had 1238 deaths with a population of 33930, a death rate of 3.6%.

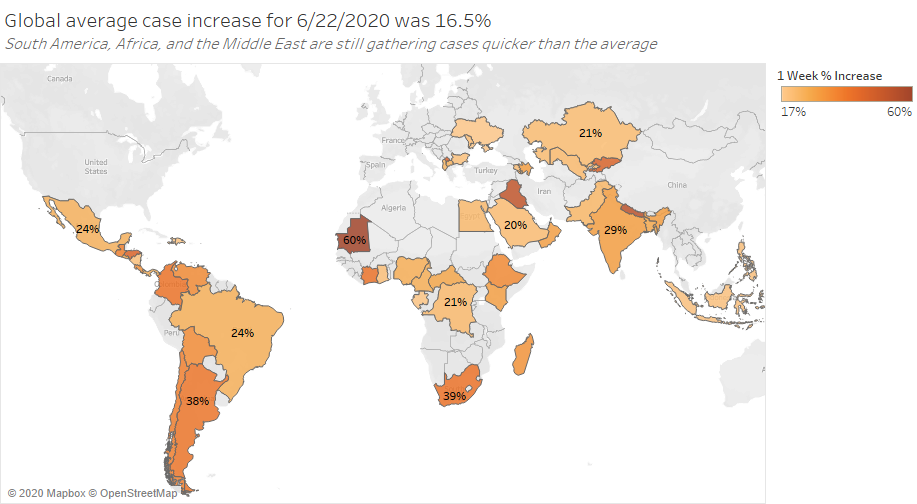


The Americas and Africa are struggling with recovery. No patients in Canada, Sweden, Mozambique, or Syria have recovered as of June 22. The UK has single-digit recovery records alongside Bolivia, Syria, and Afghanistan. Unlike with death/case ratios, Africa mainly suffers in patient recovery rates along the east and south regions of the continent.



Bahrain and Qatar lead the Middle East in case increases over the 5 months of data. Bahrain had case increases of 24.6% daily, on average. Andorra, Tajikistan, and Turkey round out the Mediterranean area. South America also had large increases.

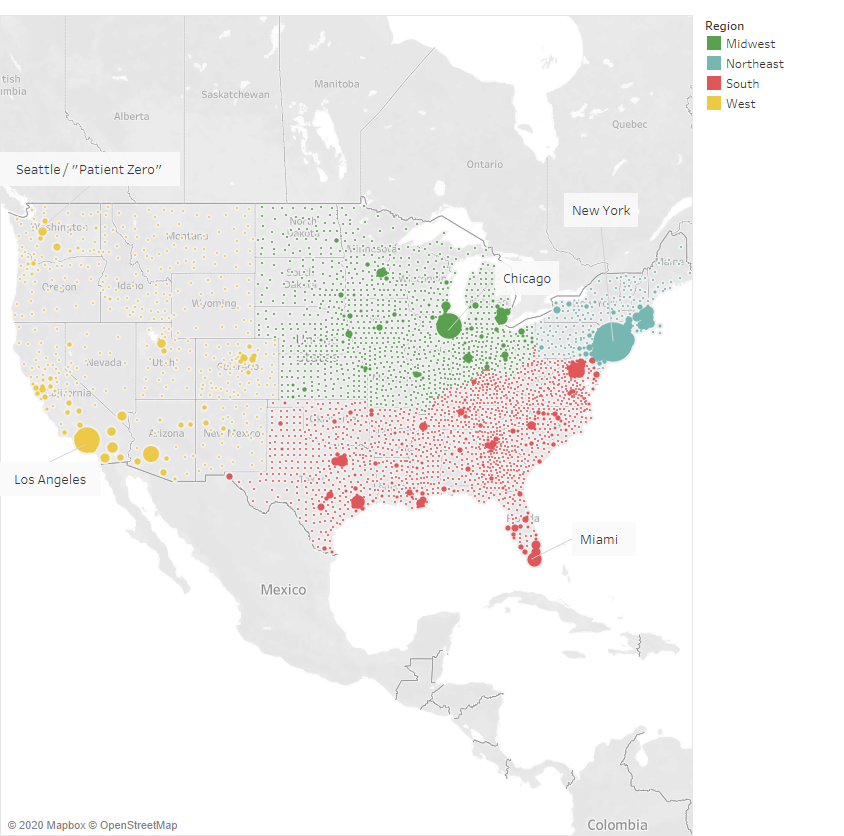




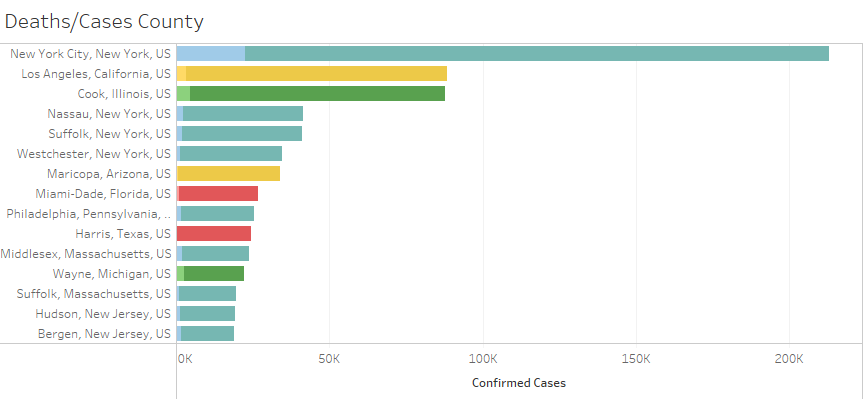
On the last day recorded, these countries were still obtaining large amounts of cases. We see South and Central America taking a large increase in the final week, with most countries above the global average. The Middle East is also above average, but much of Africa seems to be doing well. India and surrounding West Asian countries outside of China are also taking on more cases.

Overall, South America, the Middle East, and Central Africa seem to be suffering the most in dealing with the pandemic.

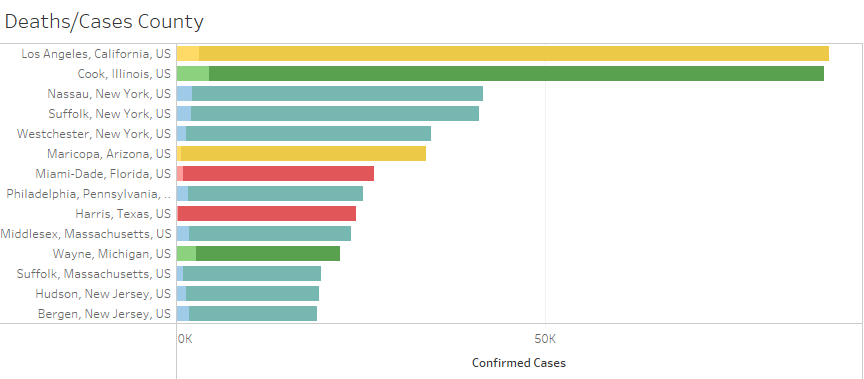
**US County Data**

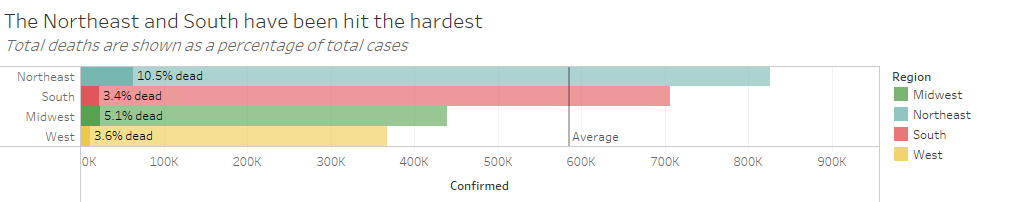


As of June 22, this is the state of the country, with larger circles representing more cases.

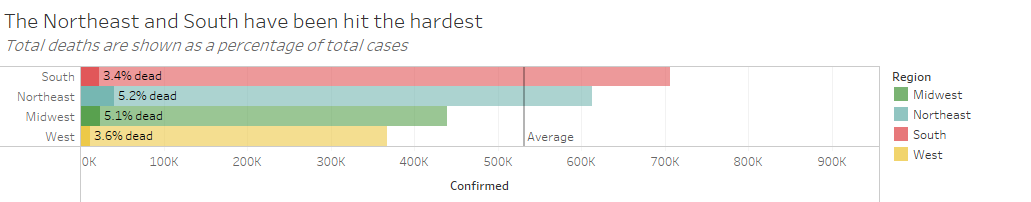


New York City bears the brunt of US cases by a huge margin. NYC has more cases than the next two cities combined, which each have more than double any other county’s case numbers. New York is a huge outlier in this case but can still be useful to see disease paths. Still, it is interesting to look at data without New York’s totals:

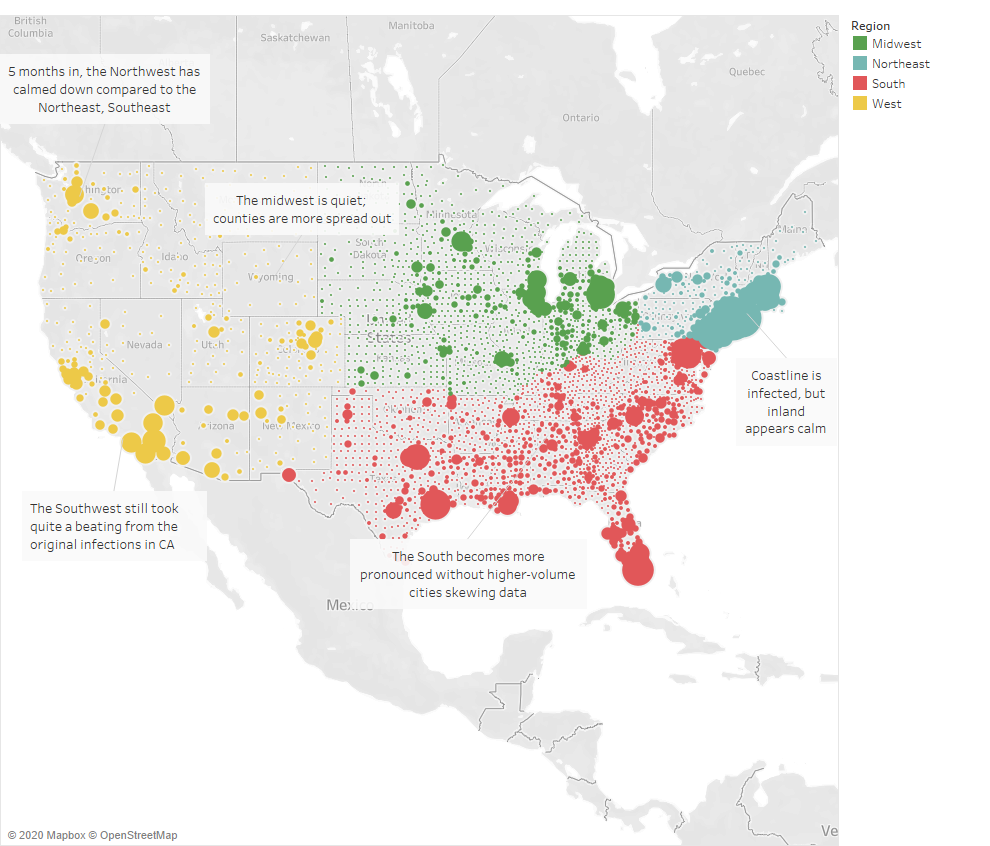
  
Even without NYC, New York state is still a hotbed of virus activity. The Northeast in general seems to be heavily hit.



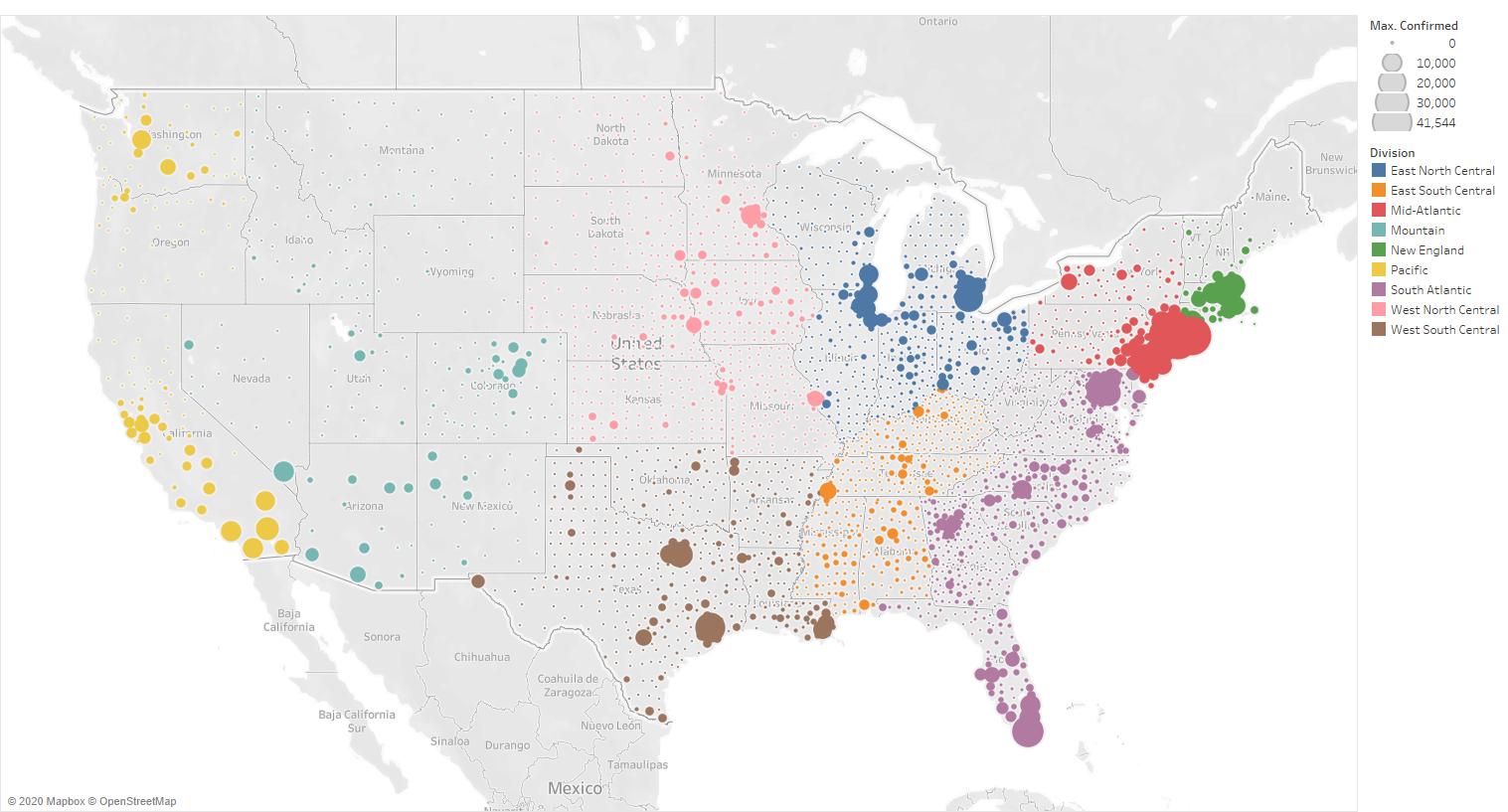
With New York present, the Northeast has 825,629 cases by June 22. The South has 706,734. Without NYC, however:



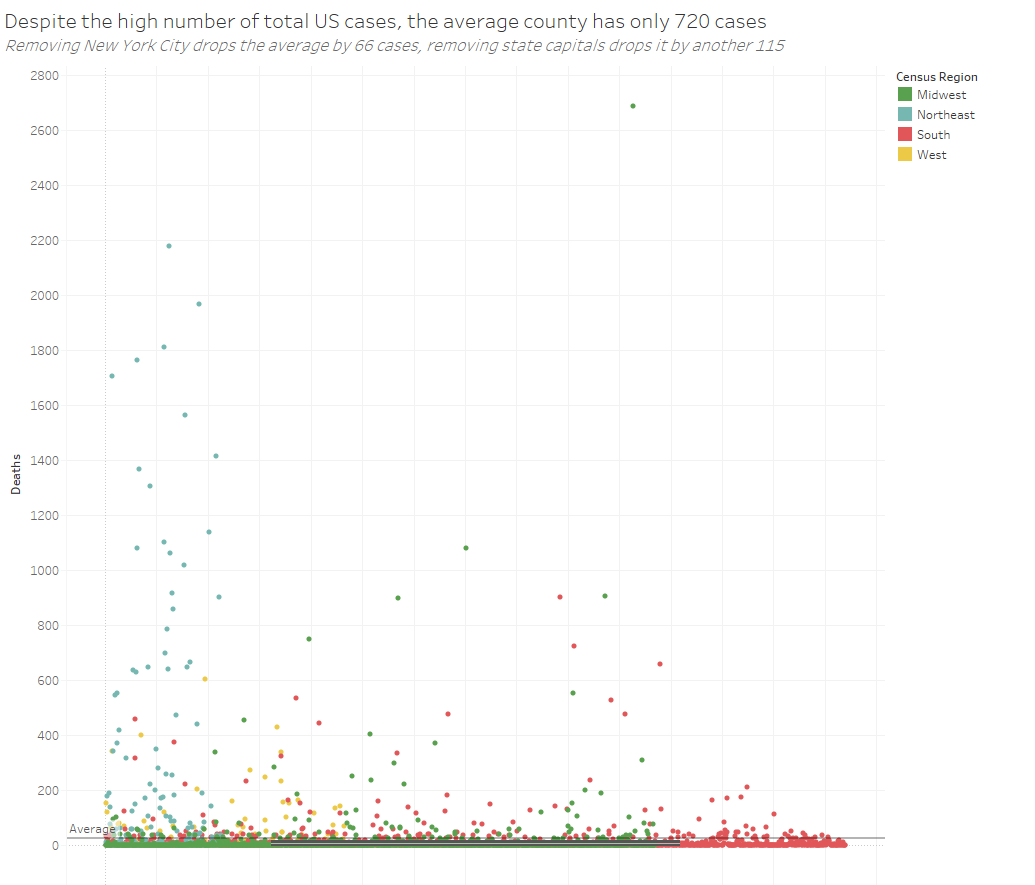
The South pulls ahead as the Northeast shrinks greatly. Even without NYC, the Northeast has the highest death rate of any region. The West has the lowest death rate and lowest case amount, despite being the origin of the US side of the pandemic (Seattle, Washington).



Here is our June 22 case distribution without our top three counties for cases: Cook, Los Angeles, and NYC. Also removed are all counties containing a state capital. Major cities appear to be hotbeds for the virus. All around NYC in the Northeast we still see immense numbers of cases. Same for Los Angeles and Cook (Chicago, Illinois). The South and Midwest have large concentrations across the region, but the West is quite quiet. We can see that many counties in the West, including in Wyoming and Montana, are spread quite far from each other and have few cases. Denver, Colorado and Salt Lake City, Utah have large bubbles radiating outward, but otherwise very little activity.



Here we see the same data split up by Division instead of merely Region. The Mid-Atlantic (around NYC) is the heaviest hit by far, and some appears to be spilling into New England. From NYC down to Miami we see a fairly strong string of cases along the South Atlantic. The area around Chicago (East North Central) is also teeming with the virus, but Mountain and West North Central are quiet.



In this distribution of total cases by county, we can see that most counties have a relatively low amount of cases, with a few larger cities bearing the brunt of the virus. The average county in the US has 720 cases. However, when removing only NYC, this drops to 654, and removing state capitals alongside the top three case spots reduces the national average to 539 per county.

Overall, we see that the Northeast US handled COVID the worst, with the most cases and deaths in the entire country. New York bore the brunt of the outbreak, then, like with many major cities, cases spilled into surrounding areas. The West fared well with more spread-out counties. The South ended up almost catching up to the Northeast toward the start of summer. Around Chicago, the Midwest has had a rough time, but otherwise there has been little activity in the region.