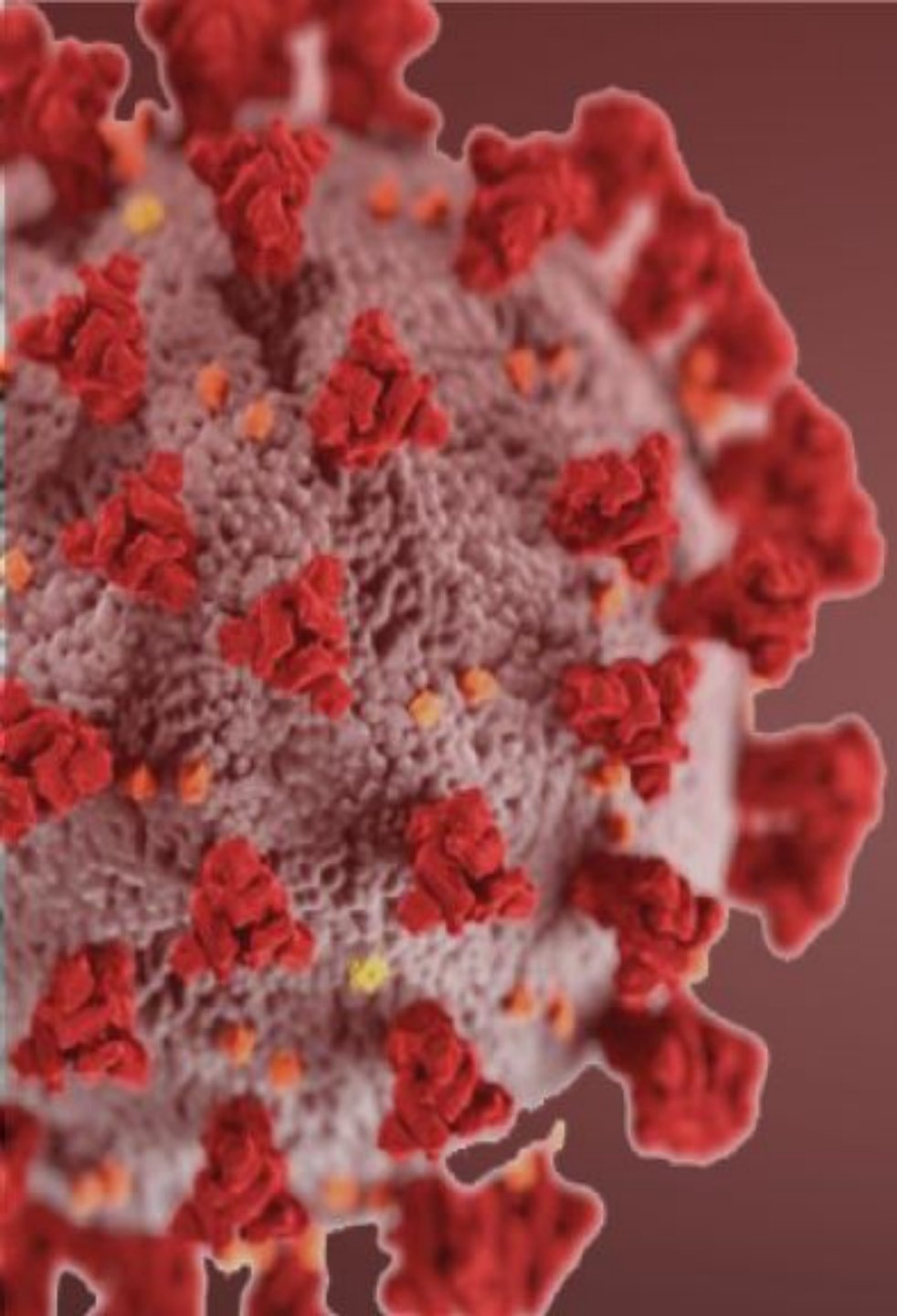


Color of COVID-19

MADS SIADS 591-592 Data Analysis Project

Presenter
Gaurav Vijaywargia
June 11, 2020

Image Source: https://www.shutterstock.com/

Table of Content

COVID-19 Is Affecting People Of Color The Most

Analysis Summary

Background

While plenty of coverage is given to spread of COVID-19, not much analysis is done to access its impact on racial and ethnic minorities.

Our first attempt tried to establish correlation between weather, population density and disease spread. With this analysis I shifted focus and tried to access impact of COVID-19 on people of color and ethnic minorities.

The key question I wanted to answer with this analysis -

Are people of color and ethnic minority are disproportionately affected by Covid-19?

Methodology

The novel coronavirus has claimed about 112,000+ American lives through June 9, according to officially reported statistics. We now know the race and ethnicity for 93% of these deaths.

After some research I decided to use data from following sources to complete this analysis.

1. County level demographic data from Census Bureau's 2018 ACS 5-Year estimates.
2. State level data compiled by The COVID Tracking Project team at The Atlantic
3. County level COVID-19 cases and death data compiled by New York Times
4. Several wikipedia articles that provides mapping between Core Based Statistical Area and Counties

The work was heavily inspired by analysis of The COVID Tracking Project and APM Research Labs.

For the data preparation, I used Google Colab (a free Jupyter Lab environment), Python programming language, Pandas as primary data manipulation and Altair as primary data visualization library.

Key Findings

Our analysis shows where the burden of this virus falls inequitably upon certain communities, especially Black and Hispanic.

Black people are dying at a rate nearly 2 times higher than their population share. (where race is known)

Further Consideration

1. We still do not have complete and consistent data from every US states for race/ethnicity. E.g. only 48 and 44 states/territories are reporting positive cases and deaths respectively
2. State-level statistics tell only part of the story, many US states are deeply segregated—meaning different counties in the same state can have vastly different breakdowns by race and ethnicity.

Data source: Subprime Percentage from Federal Reserve of NY and Equifax via FRED | 2019 Demographic Estimates from Gale Business DemographicsNow

Methodology | Data sources and technology

Key to any data analysis is to identify and access right data and have a clear goal.

For this analysis our goal was to do an Exploratory Data Analysis with rich visualization to access impact on COVID-19 on people of color. This goal influenced our choice of data sources. After careful evaluation, I landed on following data sources to complete this analysis -

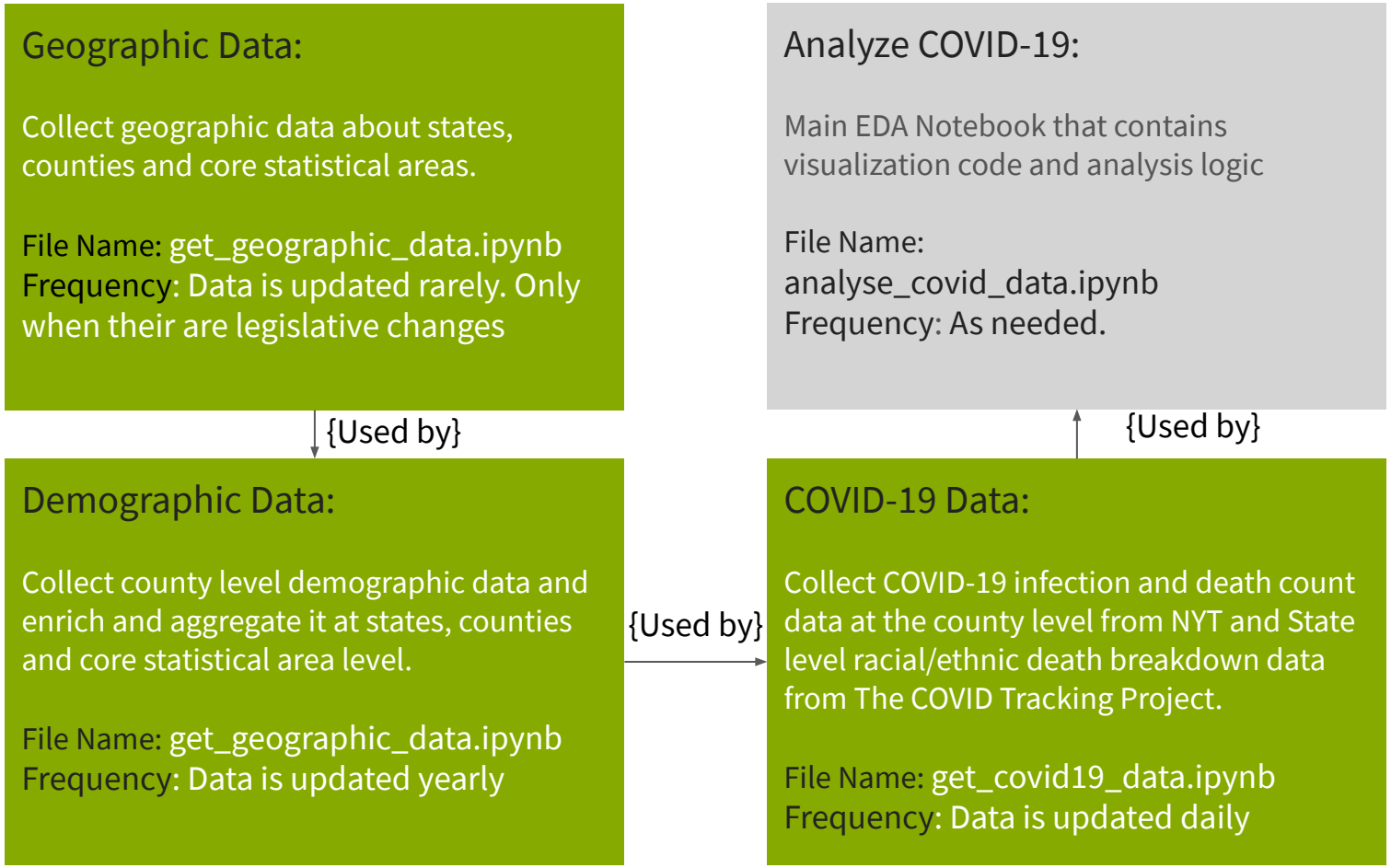
1. County level demographic data from Census Bureau's 2018 ACS 5-Year estimates: This dataset provides us percentage estimate for various racial and ethnic groups within each US county or equivalent. This data is of importance to us because it provides us baseline for racial/ethnic makeup of each county that can be rolled up to State and National level.
2. State level data compiled by The COVID Tracking Project team at The Atlantic: This dataset is of particular importance because racial/ethnic breakdown of COVID-19 infections and deaths are not consistently captured and reported by government agencies. As of this writing, only 48 and 44 states/territories are reporting positive cases and deaths respectively. Team and The Covid Tracking Project has done a great job compiling data from various State agencies and made this available to public.
3. County level COVID-19 cases and death data compiled by New York Times: NYT dataset gives us cumulative case and death total by each county or equivalent in US.
4. Several wikipedia articles that provides mapping between Core Based Statistical Area and Counties: I wanted to extend our Covid-19 spread tracking to Core Base Statistical Area, which is used by US Government's Office of Management and Budget to track economic activity. This data provides hook point for future Metro Statistical Area level analysis.

Technology Used

I used commonly applied EDA techniques to go through Data Engineering and visualization task of this analysis. The work was heavily inspired by analysis of The COVID Tracking Project and APM Research Labs. For the most part I tried to replicate and advance their analysis with this work.

For the data preparation, I used Google Colab (a free Jupyter Lab environment), Python programming language, Pandas as primary data manipulation and Altair as primary data visualization library.

Methodology | Solution Architecture



The primary goal of this solution architecture was to create a data pipeline that can be run repeatedly. The work was divided into logical Notebooks, so making change is easy.

Note about Data Processing -

Federal Information Processing Codes (FIPS) are used as key column to join various county level and state level data. This enables us to accurately join data between various files. Fortunately NYT dataset provides us data with FIPS codes.

Analysis Walkthrough

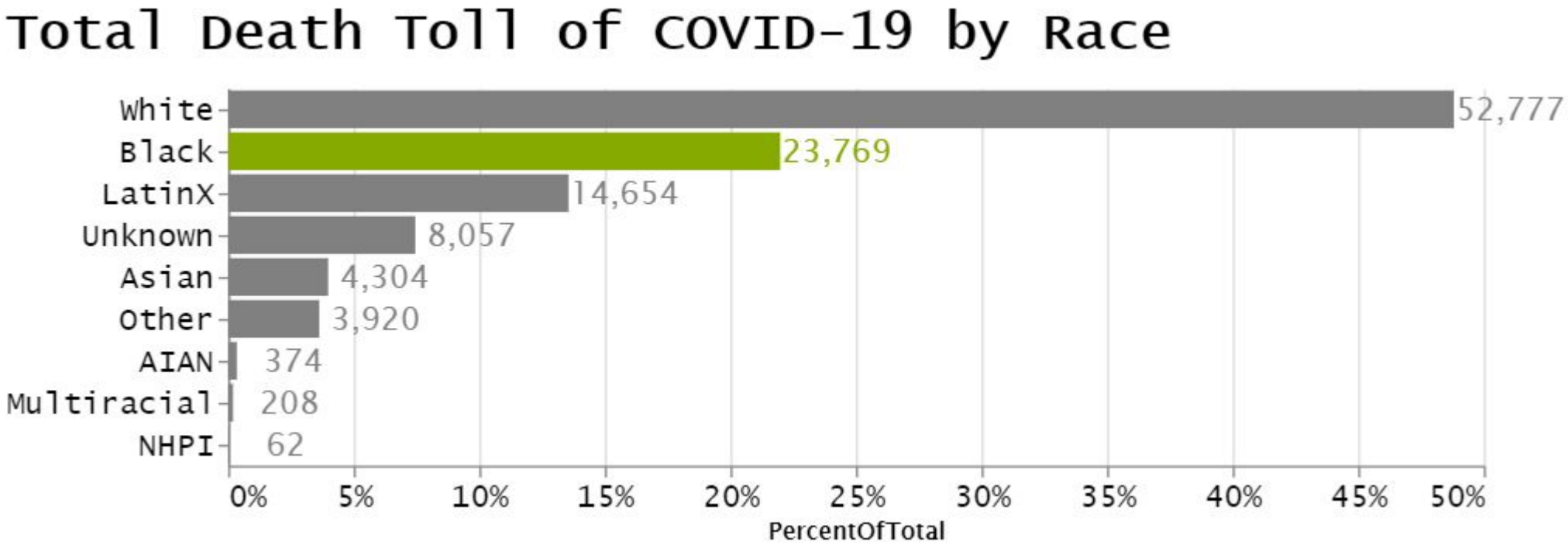
+ Key Finding # 1: Black people represent nearly 23% of Deaths in United States, while they makeup only 13% of Population.

+ Key Finding # 2: Number of Infection cases in top 20 US counties are proportional to US population, but we see clear sign of disproportionate impact on people of color when we look at number of Deaths.

+ Key Finding # 3: More black deaths relative to their share of population

Analysis Walkthrough | Key Finding #1

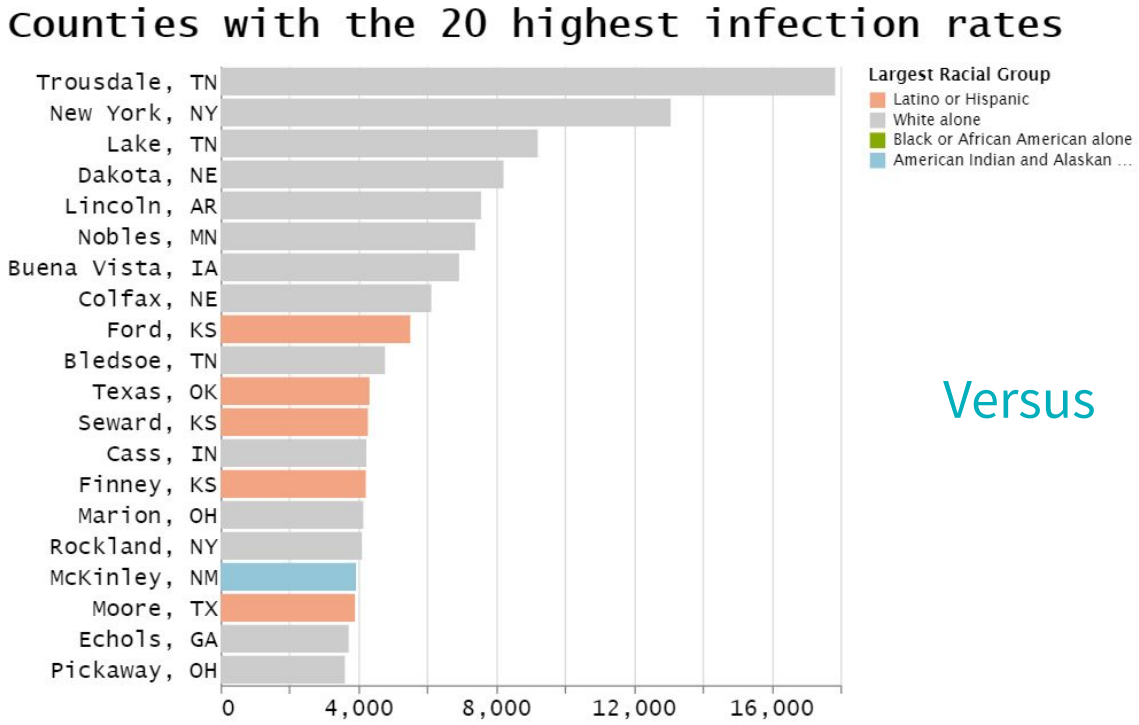
The very first thing I wanted to figure out at the national level was if number of death is proportional to communities overall population.
As we see here Black people represent nearly 23% of Deaths in United States, while they makeup only 13% of Population.



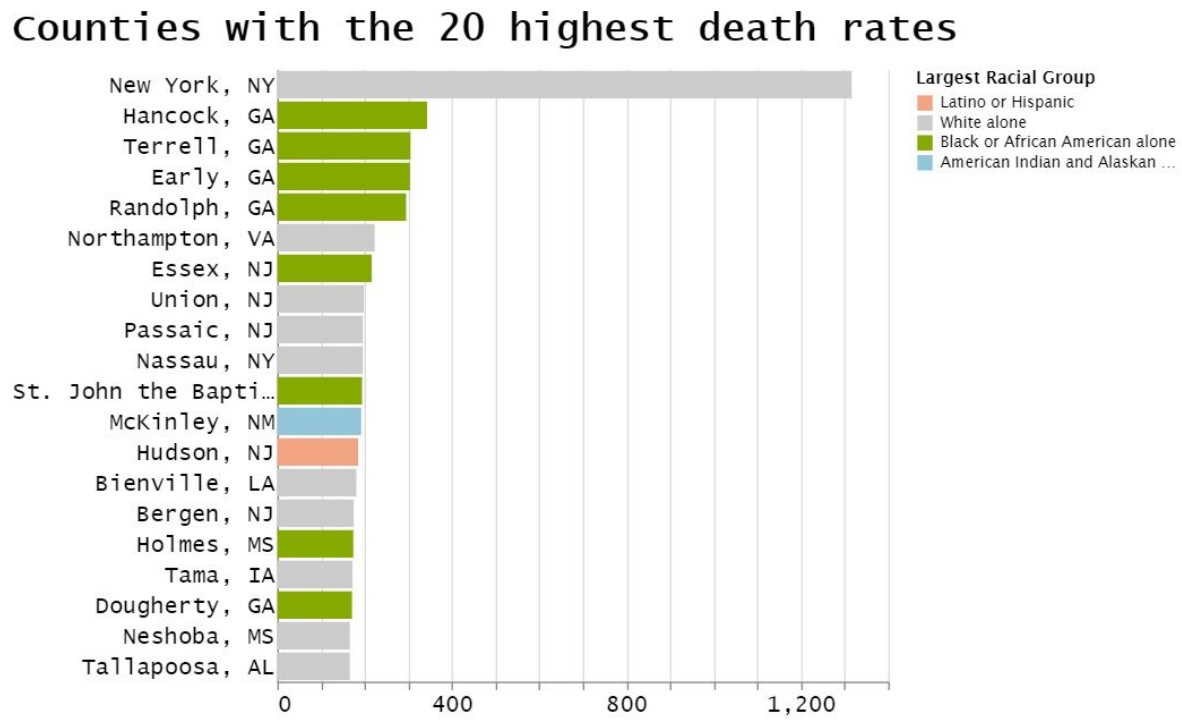
Data Source: New York Times | Covid Tracking Project from The Atlantic. Results are statistically significant as one sample z-test of proportion shows The z-score of -16.35 which results in very low p-value.

Analysis Walkthrough | Key Finding #2

Number of Infection cases in top 20 US counties are proportional to US population, but we see clear sign of disproportionate impact on people of color when we look at number of Deaths



Versus

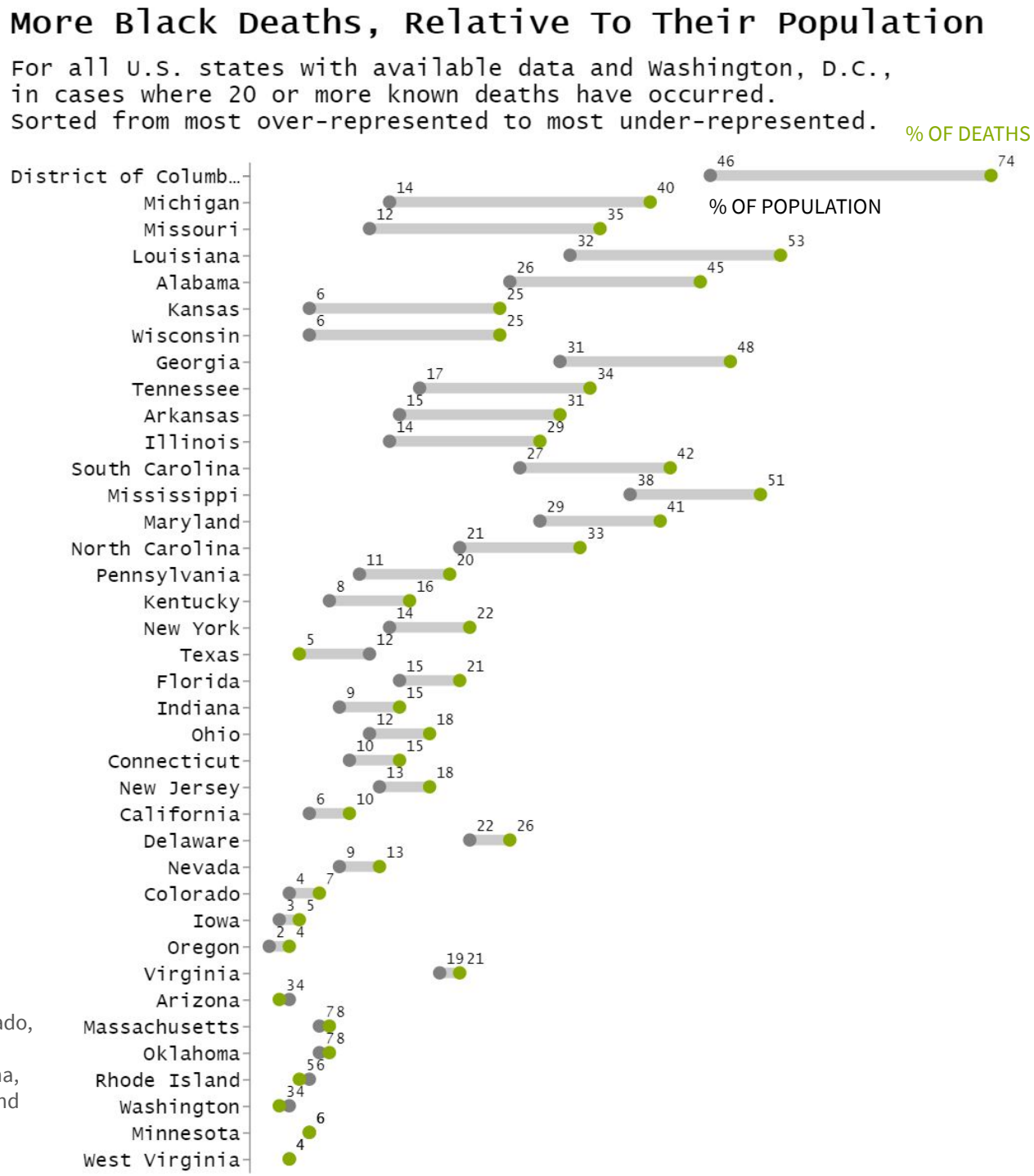


State-level statistics tell part of the story, but many US states are also deeply segregated—meaning different counties in the same state can have vastly different breakdowns by race and ethnicity.

Race and ethnicity data for COVID cases isn't widely available at the county level, so we're using two numbers we do have: the latest infection and death rates for each county, from a New York Times dataset, paired with the largest racial or ethnic group in that county, based on the Census Bureau's 2018 ACS 5-Year estimates. The results are staggering.

Analysis Walkthrough |
Key Finding #3

The purpose of this visualization is to showcase disparity between % of deaths experienced by Black community vs. their share of Population. This very powerful yet simple chart shows states where this disparity is very evident.



Includes data from Washington, D.C., and the 35 states of Alabama, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Georgia, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Maryland, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, Nevada New Jersey, New York, North Carolina, Ohio, Oklahoma, Pennsylvania, Rhode Island, South Carolina, Tennessee, Texas, Virginia, Washington and Wisconsin.

Further Consideration

We still do not have complete and consistent data from every single US states for race/ethnicity. E.g. only 48 and 44 states/territories are reporting positive cases and deaths respectively.

State-level statistics tell only part of the story, many US states are deeply segregated—meaning different counties in the same state can have vastly different breakdowns by race and ethnicity. We do not have county level data with racial and ethnic breakdown. A much more detailed analysis can be conducted once we have county level race/ethnic data available.

My analysis results are the first attempt to portray COVID-19 mortality by race, with a lens on inequitable deaths.

References

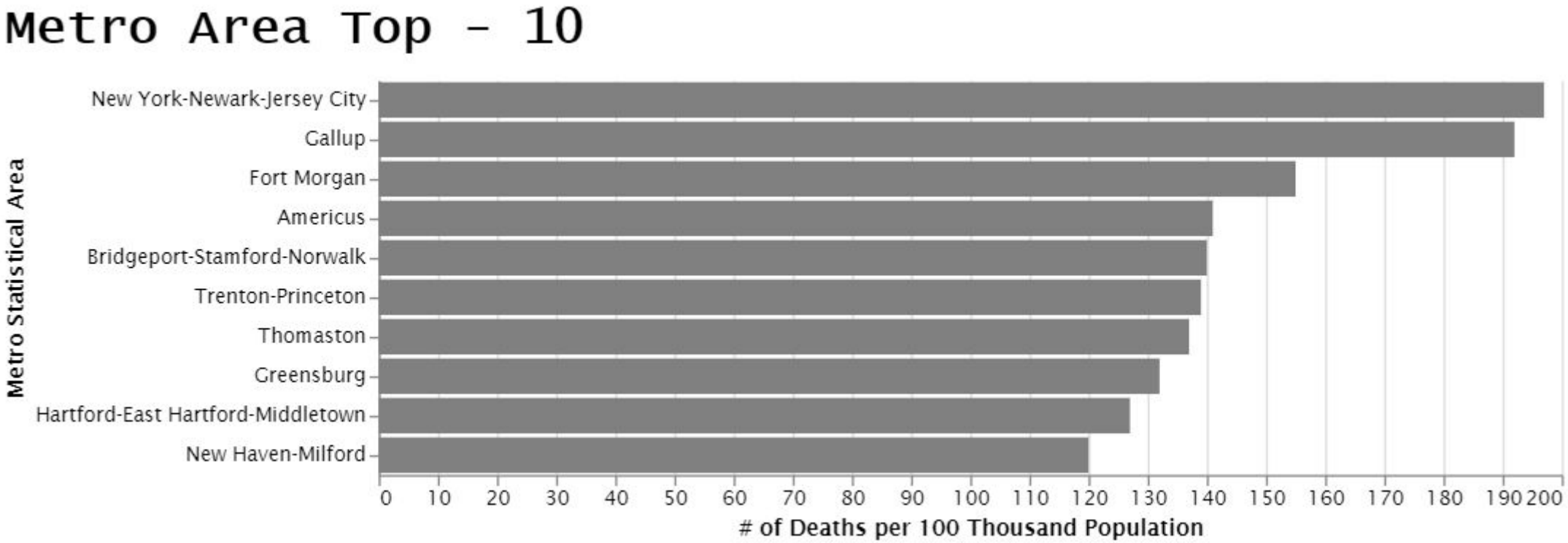
The work was heavily inspired by analysis of The COVID Tracking Project and APM Research Labs.

“American Community Survey.” 2018. *5 year estimates* https://www.census.gov/.

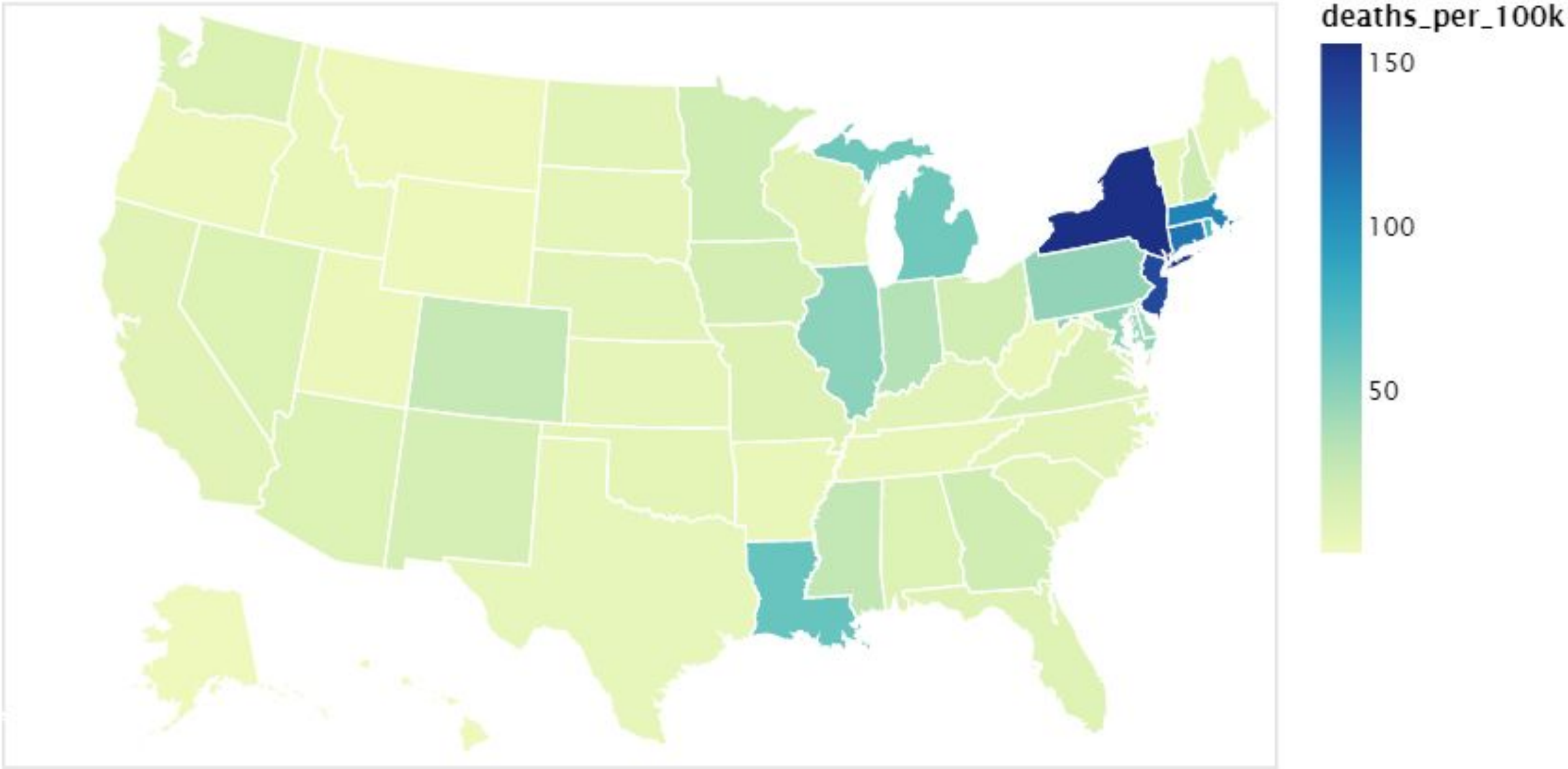
“List of United States counties and county equivalents”. Wikipedia Accessed June 8, 2020. https://en.wikipedia.org/wiki/List_of_United_States_counties_and_county_equivalents

“New York Times”. Covid-19 Data. https://github.com/nytimes/covid-19-data

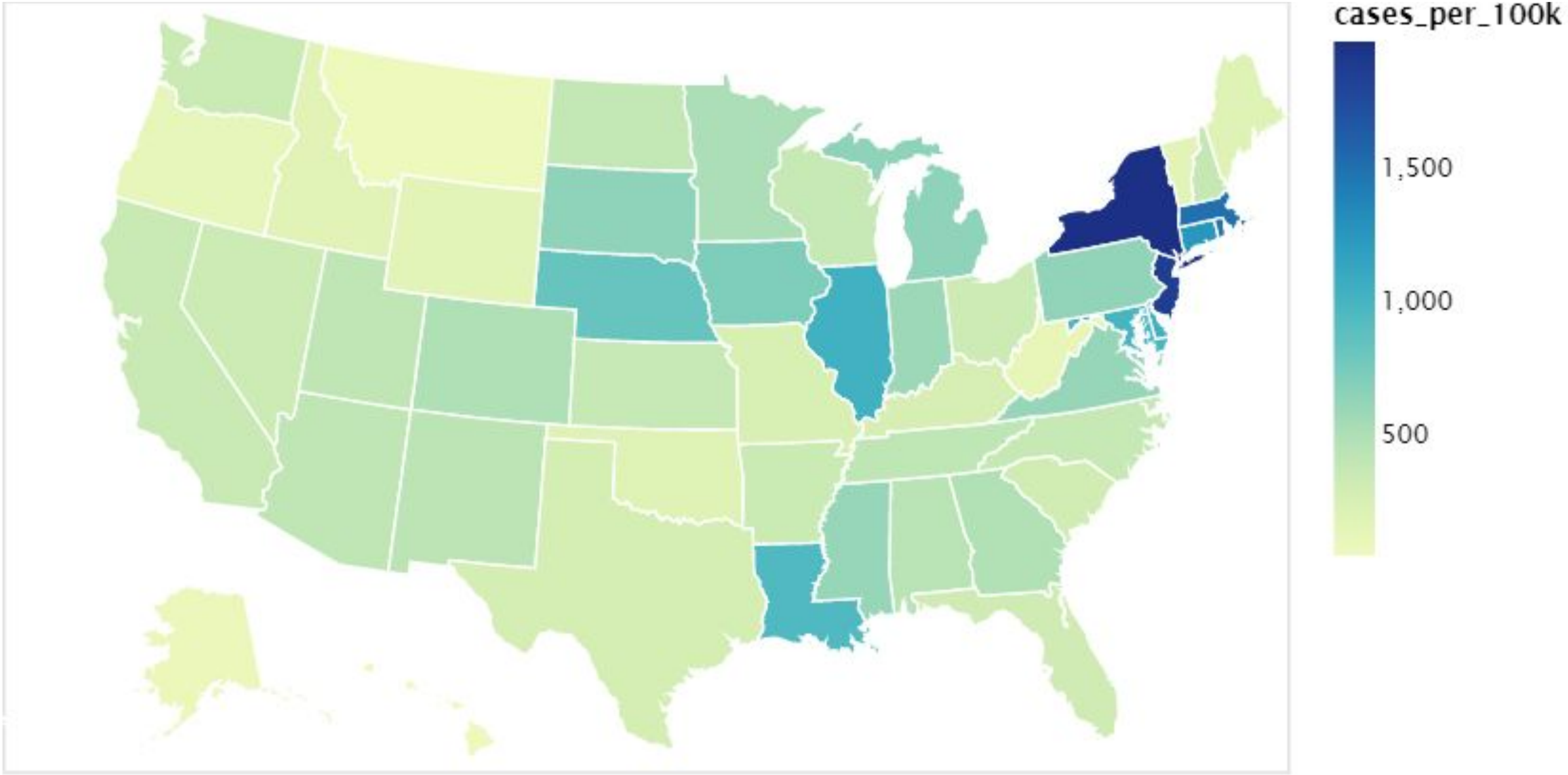
Appendix | Metro Statistical Area death Toll



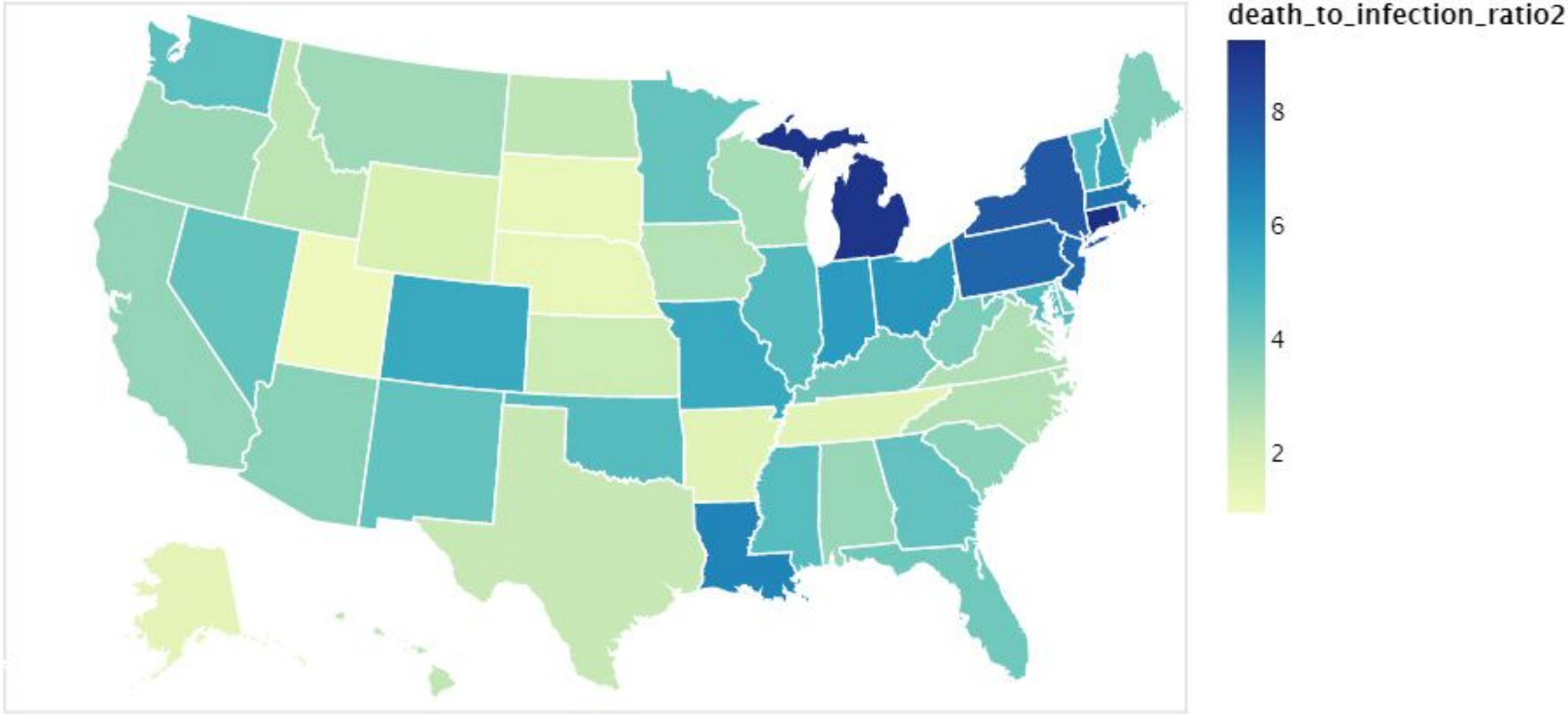
Appendix | Death Rate by State



Appendix | Infection Rate by State



Appendix | Death to Infection Ratio



Appendix | Death per 100K Population by County