

Determining the Impact of COVID-19 on the 2020 Presidential Election

By Jacob Johnston

December 22, 2020

# 

[**INTRODUCTION**](#_78ahqwxxg8sh) **2**

[Problem Statement](#_h7l377dxjmhs) 2

[**Motivation and Background**](#_8m4nu85eic37) **3**

[**Explanation of Datasets**](#_3pi9rav6lend) **3**

[**ANALYSIS**](#_dqj1x9xksqmf) **4**

[Data Cleaning and Tracking Changes in Voter Behavior from 2016 to 2020](#_nr6zzk6dyydl) 4

[Integrating Longitudinal COVID-19 Data](#_3k7512ei1xa7) 5

[What Do Statistics Say About the Relationship Between COVID-19 Infection Rates and the Presidential Election Results?](#_vvi4jbmya6cr) 5

[Mapping the Impact of COVID-19 on the 2020 Presidential Election](#_iey5fg2v6m4u) 8

[**DISCUSSION**](#_mbdv04xb38or) **9**

# 

# 

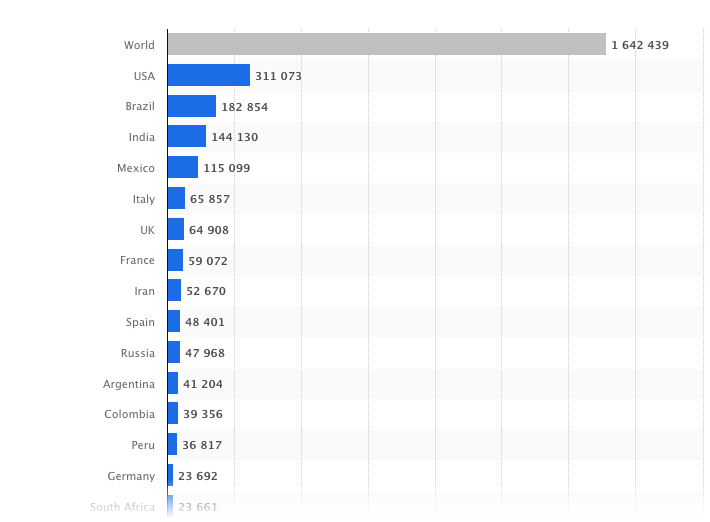
# 

# 

# INTRODUCTION

## Problem Statement

The novel coronavirus (COVID-19) has impacted economies and lives worldwide in 2020. At the time of writing, the United States leads the world in COVID-19 cases and deaths,[[1]](#footnote-1) and many attribute this tragedy to President Trump’s response to the pandemic.[[2]](#footnote-2) Given the prevalence of COVID-19 in the U.S., the assumption that the virus influenced the results of the 2020 Presidential Election is not an unreasonable one.



*Number of novel coronavirus (COVID-19) deaths worldwide as of December 16, 2020, by country*1

What exactly was the impact of COVID-19 on the 2020 election? Did the pandemic harm President Trump’s chances of being reelected? These are the types of questions I hope to answer in this paper, providing both statistically backed answers and intuitive visuals to display the findings.

# Motivation and Background

The COVID-19 pandemic has become a largely political issue in the United States. President Trump has been widely criticized for his response to the pandemic.[[3]](#footnote-3) President Trump has also downplayed the pandemic on many occasions as a ploy by Democrats to remove him from office, prompting many of his supporters to go so far as to claim the virus is a hoax.[[4]](#footnote-4) Did this ploy work? Did Trump’s response to the pandemic cost him voter support? If that were the case, then the voting behavior of those who live in areas most affected by COVID-19 would be the best indicator of whether infections and deaths from the pandemic really caused people to rally behind Democratic candidates.

Pandemics are rare events, especially during the same year as a presidential election. This is a unique opportunity for understanding how a pandemic and the government response can influence voter behavior.

# Explanation of Datasets

For this project, I will be using data from three different datasets. The main dataset I use provides county level voter data for both the 2016 and the 2020 presidential elections, along with some demographic information, latitude and longitude, and the total cumulative number of COVID-19 deaths and infections for each county as of the day of the election (November 2, 2020). In total, the dataset consisted of 51 variable columns and 4,867 rows. This dataset is available on Kaggle.[[5]](#footnote-5)

The second dataset I used was a longitudinal dataset that provided the day-to-day cumulative number of COVID-19 cases and deaths for each county. The county appears in the dataset on the day of the first reported case and the cumulative progress of reported cases and deaths is then tracked each day for the rest of the year. The earliest reported U.S. case was on January 21, 2020, so that is the first row of the dataset. In total, the dataset contains 6 columns (date, county, state, zip, cases, deaths) and more than 836,000 rows. This dataset is also available through Kaggle (or The New York Times) and is updated daily.[[6]](#footnote-6)

The last dataset I used is a shapes file (and the index file) provided by the U.S. Census Bureau.[[7]](#footnote-7) This file simply contains geometric data for mapping the U.S. states and territories by latitude and longitude on a coordinate plane.

# ANALYSIS

## Data Cleaning and Tracking Changes in Voter Behavior from 2016 to 2020

My first step in the analysis was to clean the data from the primary dataset. I started by removing all the extra “counties” that did not contain information on voter behavior of COVID cases. This mostly removed rows that were not actually counties or county equivalents - the dataset came with 4,867 rows of “counties” but there are only 3,143 counties or county-equivalents in the U.S.[[8]](#footnote-8) After removing rows with missing information, I was left with 3,046 rows. These rows included county-level data for the District of Columbia and every state except for Alaska.

Part of the data cleaning also involved changing the values of the columns labeled “percentage” to actually be percentages instead of proportions. This was necessary for the calculation of the new variables for tracking the change in voter behavior from 2016 to 2020. I created two new variables: the first to track the raw difference in votes that Trump received from 2016 to 2020, the second to track the change in the percentage of votes that Trump received from 2016 to 2020. Ultimately, I made the decision to use the change in percentage of votes as the target variable because it is more comparable across counties with large differences in population.

## Integrating Longitudinal COVID-19 Data

In human psychology, recent events are more salient and have a greater influence on judgments and decisions.[[9]](#footnote-9) Given this information, I decided that looking at the number COVID-19 cases and deaths during the month leading up to the election would be more relevant in understanding how COVID-19 impacted voter behavior. I imported a second longitudinal dataset with COVID information by date. I filtered this dataset to only include data from November 2, 2020 (election day) and October 2, 2020 (the month before the election) and merged it with my primary dataset. Then I subtracted the end point from the start to get the number of new cases and deaths that occurred in the month of October for each county. I converted this information into percentages for comparability across counties.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Mean** | **Median** | **Max** |
| **Cases in October as Percent of Population** | 0.913140 | 0.700 | 13.35 |
| **Deaths in October as Percent of Population** | 0.013544 | 0.010 | 0.61 |
| **Total Cases as Percent of Population** | 2.858674 | 2.615 | 20.46 |
| **Total Deaths as Percent of Population** | 0.057354 | 0.040 | 0.64 |

## What Do Statistics Say About the Relationship Between COVID-19 Infection Rates and the Presidential Election Results?

Although Joe Biden received the greatest number of votes in the election, Donald Trump won more voter counties:

Votes Received Counties Won

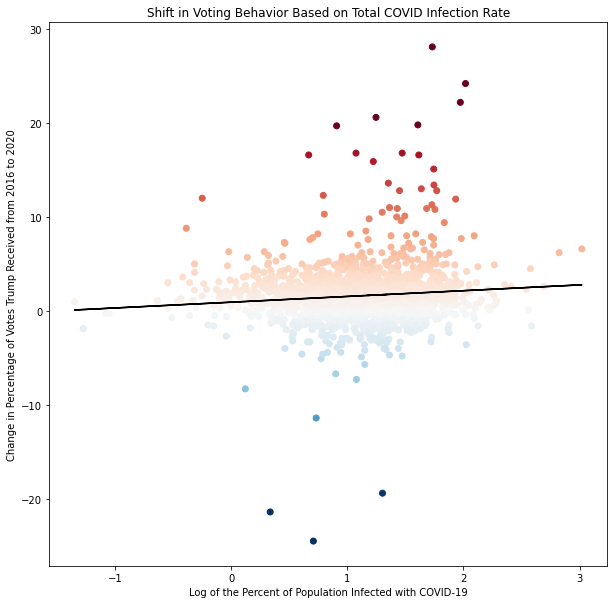
Joe Biden 71672423.0 470

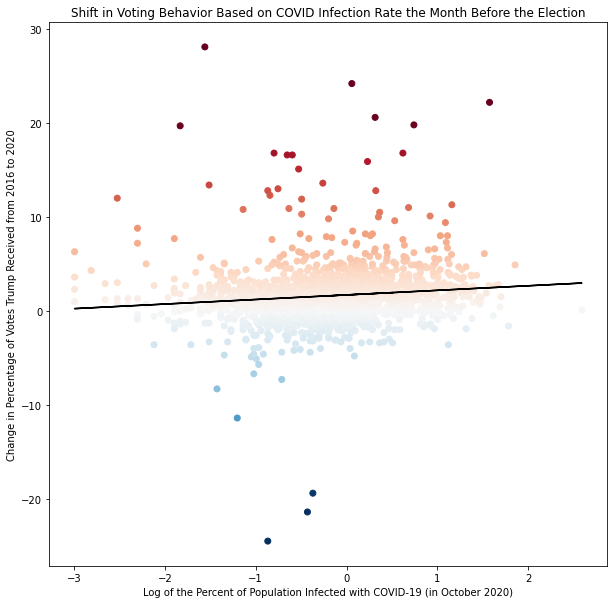
Donald Trump 68958451.0 2576

Number of counties where Trump gained support from 2016 to 2020: 2267

Number of counties where Trump lost support from 2016 to 2020: 740

After running these basic summary statistics, I ran linear regressions on rates of infection and death in each county against how their voting behavior shifted from 2016 to 2020. Log-transforming the data on cases and deaths was necessary to meet the requirement of homoscedasticity. These regression lines showed a slight but significant positive correlation between the percentage of the population infected and increased support for Donald Trump. When I ran the same regressions with the rates of infection and death during the month before the election, the strength of this correlation increased. The regression lines for deaths show the same trends as those for cases but are less illustrative, so I will only include the models for cases (both total and only the month before) in the visuals below.

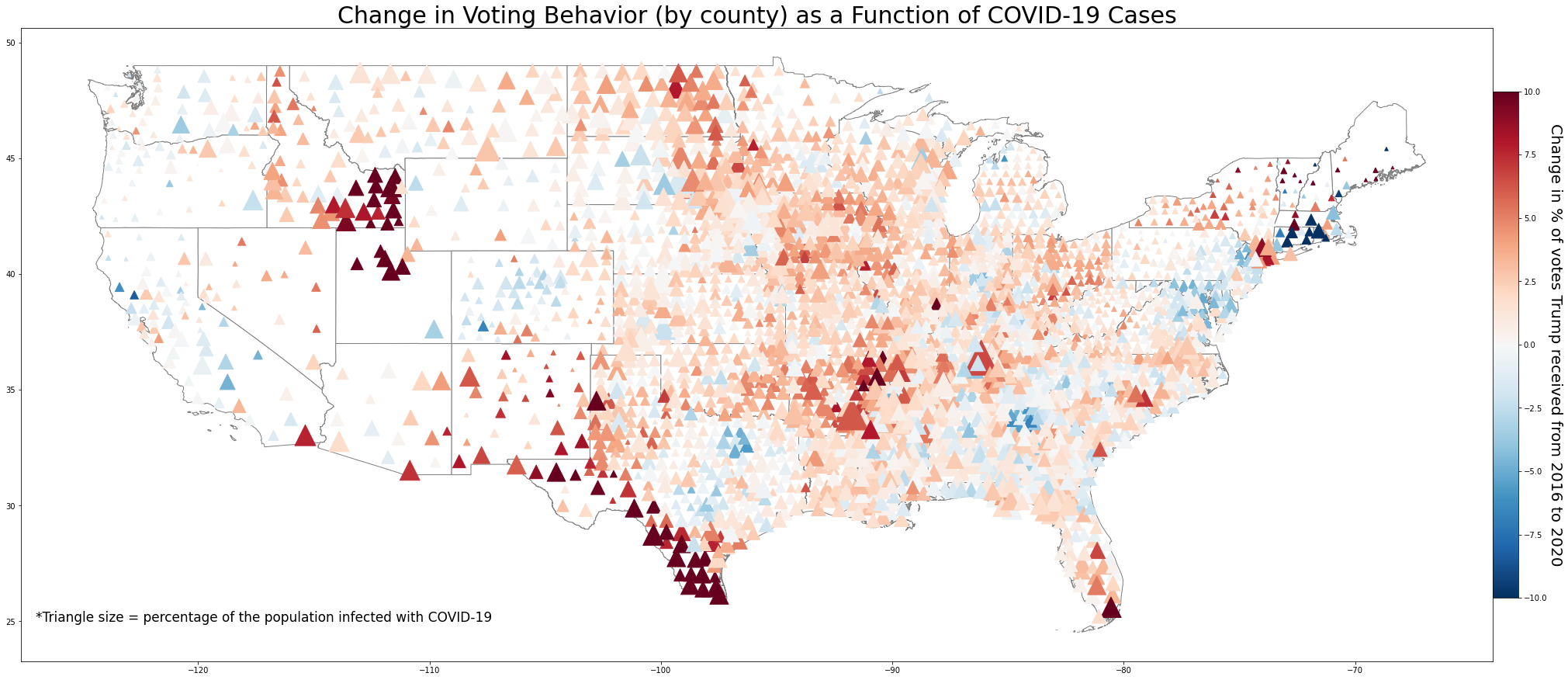


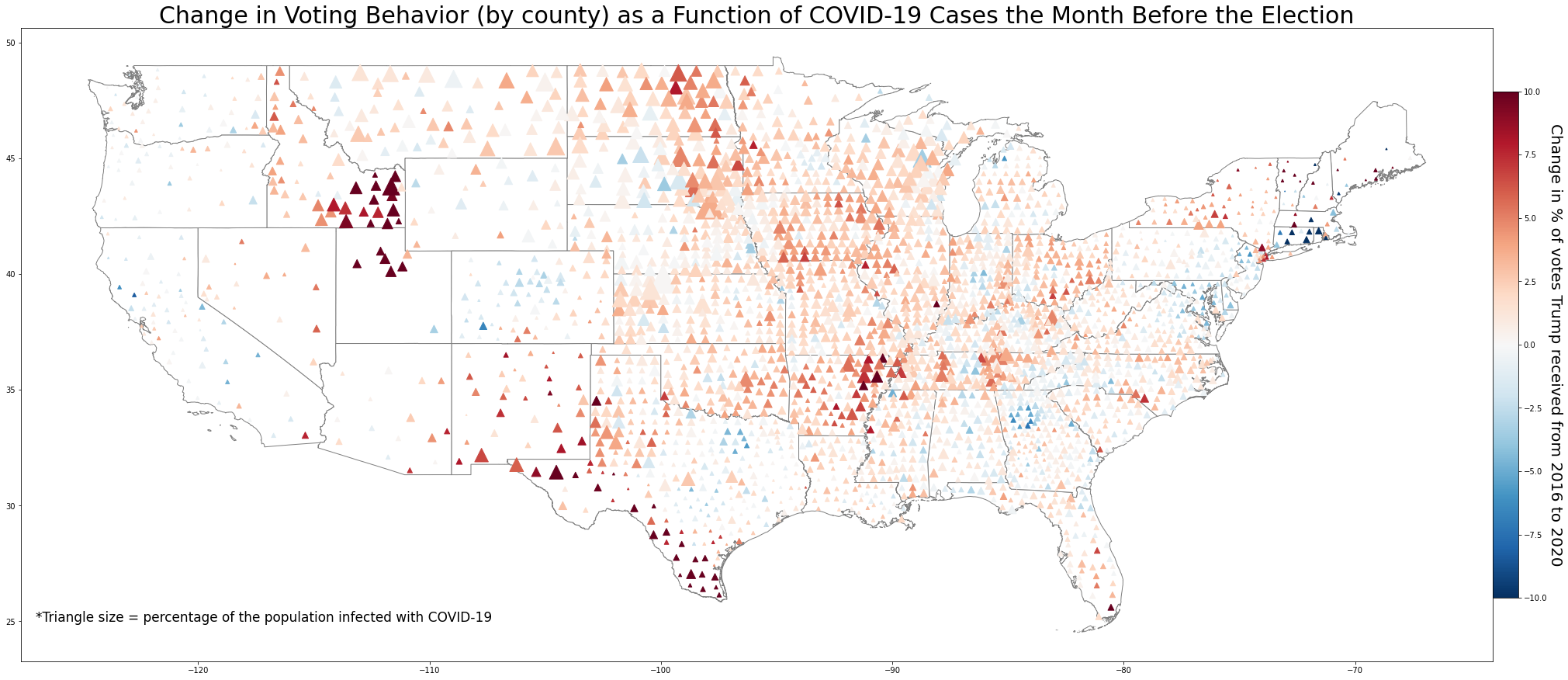


These regression models are interesting and highly significant (p<.0001). However, they must be taken with a grain of salt as other assumptions for linear regression have not been met (such as independence and normality). Because these are observational data, we also cannot be sure of the direction of causality. Another plausible explanation is that counties with increased support for Donald Trump were less likely to take precautions to prevent the spread of COVID-19 and, thus, have more cases and deaths.

## Mapping the Impact of COVID-19 on the 2020 Presidential Election

Finally, I integrated the state shapes file from the U.S. Census Bureau in order to create a geospatial visualization of the data. I made the decision to remove the U.S. territories, Hawaii, and Alaska in order to focus on the continental U.S. and create a clearer visual. This decision only removes 4 counties from the usable dataset, so there is little change in the results. As can be seen from the maps, the counties with the highest infection rates also tend to show increased support for Donald Trump in the 2020 election. This holds true for both the total COVID-19 infection rate and the infection rate for the month before the election.





# DISCUSSION

Despite the supposed Democratic ploy and the Trump Administration’s purportedly poor response to the pandemic, Trump did not lose voter support in areas that were most affected by COVID-19. In fact, those counties with the highest rates of infection among their populations are also those where Trump gained the greatest percentage of voter support from 2016 to 2020. A linear regression model reveals this relationship to be statistically significant.

One potential interpretation of these findings is that those who live in the most-affected counties see Trump’s COVID response to truly be helpful. Another interpretation could be that areas where Trump is gaining support also tend to downplay the seriousness of the pandemic as he has, not taking precautions to prevent the spread of the virus and, thus, having a larger number of infections. Due to the data collection method, the causal direction cannot be determined.

A significant limitation to this report is my personal lack of statistical knowledge. I know that, due to the failure to meet certain criteria in the data, a linear regression is likely not the best way to test the significance of the studied relationship. However, I do not know how to run a non-parametric version of a linear regression using Python. If I had more time, I would have spent longer researching and running more appropriate statistical methods for the analysis.

Despite these limitations, I believe this report, particularly the geospatial visualizations, provides enough data to intuitively identify the connection between COVID cases and voter behavior. Further research investigating how Donald Trump has managed to win the support of voters most affected by the pandemic would be interesting for understanding voter beliefs in America. This could prove to be highly beneficial information for future political candidates.

1. <https://www.statista.com/statistics/1093256/novel-coronavirus-2019ncov-deaths-worldwide-by-country/> [↑](#footnote-ref-1)
2. <https://coronavirus.house.gov/news/reports/select-subcommittee-releases-new-staff-report-investigations-trump-administration> [↑](#footnote-ref-2)
3. <https://coronavirus.house.gov/news/reports/select-subcommittee-releases-new-staff-report-investigations-trump-administration> ; <https://www.washingtonpost.com/graphics/2020/politics/trump-covid-pandemic-dark-winter/> ; <https://www.nature.com/articles/d41586-020-03035-4> [↑](#footnote-ref-3)
4. <https://www.washingtonpost.com/opinions/2020/10/26/trumps-last-gasp-pandemic-isnt-real-everythings-fine/> ; <https://www.theguardian.com/us-news/2020/oct/03/donald-trump-base-stays-loyal-president-fights-covid-19> [↑](#footnote-ref-4)
5. <https://www.kaggle.com/etsc9287/2020-general-election-polls?select=county_statistics.csv> [↑](#footnote-ref-5)
6. <https://www.kaggle.com/fireballbyedimyrnmom/us-counties-covid-19-dataset> [↑](#footnote-ref-6)
7. <https://www.census.gov/geographies/mapping-files/time-series/geo/cartographic-boundary.html> [↑](#footnote-ref-7)
8. <https://en.wikipedia.org/wiki/County_(United_States)#:~:text=As%20of%202016%2C%20there%20were,states%20and%20District%20of%20Columbia>. [↑](#footnote-ref-8)
9. <https://en.wikipedia.org/wiki/Recency_bias> [↑](#footnote-ref-9)