THE EFFECTS OF POLITICS BASED ON COVID-19 CASES

A PREPRINT

Christopher Chang
Department of Computer Science
Boston University
Boston, MA 02215
changc21@bu.edu

Kevin Sujanto
Department of Computer Science
Boston University
Boston, MA 02215
kevinsuj@bu.edu

December 18, 2020

ABSTRACT

This paper attempts to understand the relationship between politics on the state and county level and COVID-19 cases in the United States. We used linear regressions and multiple variable regressions to analyze how politics and other factors impact cases. This paper will argue that a more Republican state will have more cases than a Democratic state. It will also raise questions as to what else impacts COVID-19 numbers.

***K*eywords** COVID-19 · Politics · Cases · Death · Education · Income · Age

1 Problem Statement and Motivation

Throughout the pandemic, the United States has seen a wide discrepancy on how to handle COVID-19 and consequently, the number of cases across state and county lines. There are many factors that influence the spread of COVID-19, but this project aims to understand the relationship between US politics and the virus. We hypothesize that there is a correlation between politics and COVID-19 cases based on differing preventative measures, and in some cases the lack thereof, taken by Democratic and Republican states. We will also look at how several external factors affect these numbers. Through this analysis, we hope to gain a better understanding on how politics and these external factors impact COVID-19 numbers. The significance of our findings can be used to help formulate trends to better predict the future of states as this pandemic continues to plague our nation.

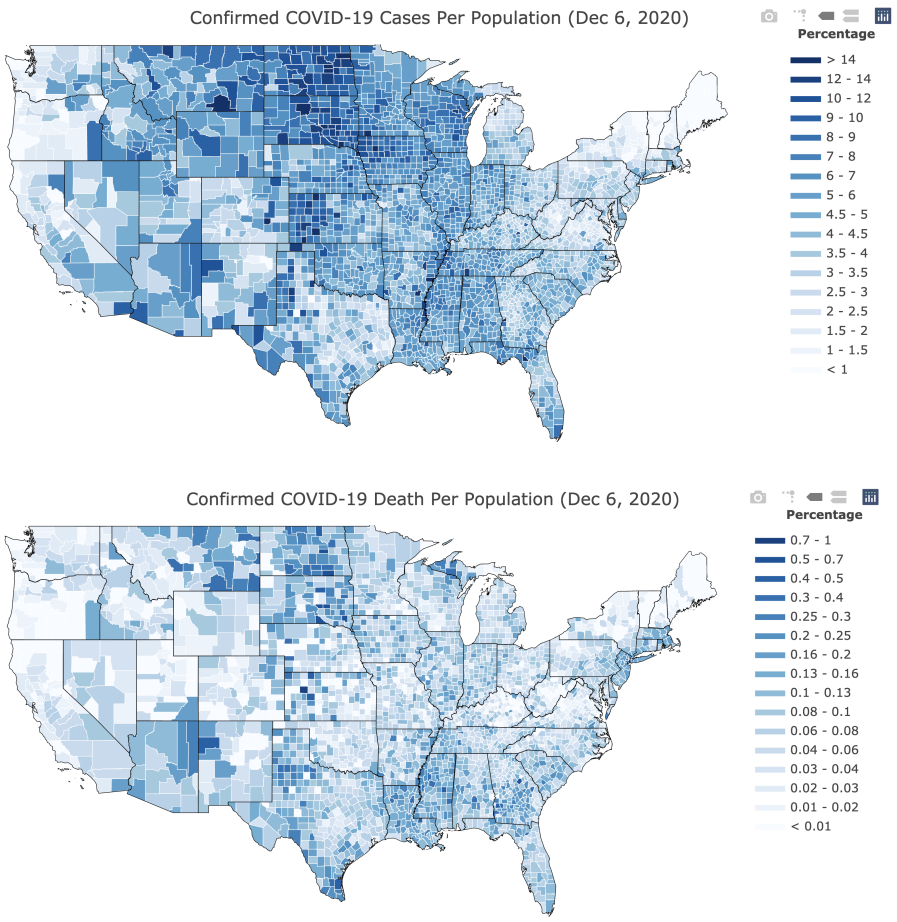
2 Datasets

For our project, we combined datasets from COVID-19, political and population to create a table with all the necessary information needed in our project. For county level, we used FIPS as the primary key or foreign key between datasets and was able to create a new table that consists of FIPS, county name, population, COVID-19 cases, total death, county political view, county average income, percentage of Democrats and percentage of voters. For state level, we used some data from the county level such as COVID-19 cases, total death, and county political view to create a state level data for them, used some state-level datasets for age, income and education level, to create a new table that consists of state name, population, COVID-19 cases, total death, COVID-19 cases per population, death per COVID-19, average age, average income and percentage of people with high school or college degrees.

2.1 COVID-19 Data

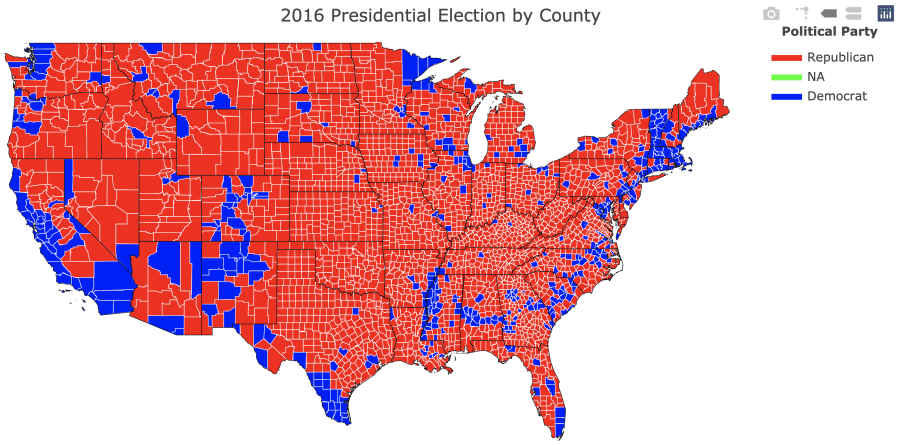
Our COVID-19 Data comes from *USA Facts* [1] and provides per county COVID-19 Cases, per county population numbers, and per county deaths as of December 6, 2020.

[1] https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/



2.2 Political Data

Our Political Data comes from *MIT* [2] and provides information such as year, state, county name, FIPS, candidate, political party and votes that helps us determine if a state or county is considered Republican or Democratic. Factored in our study is the party of the state's current governor, the parties of state's current Senators and 2016 and 2018 House Representatives, who states voted for in the 2012 and 2016 Presidential Election, and who counties voted for in the 2016 Presidential Election.



[2] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VOQCHQ

2.3 Population Data

Our Population Data comes from *Census* [3], *Kaggle* [4], *Corgis* [5] and provides information per state such as age, median household income, and education level respectively. This dataset is imperfect for several reasons. First, the age dataset gave us the amount of people at each age per state up to age 85. We found the average age of each state with this information but we are unsure if the dataset included people over the age of 85 and if so we don't know if they just added those people to the 85 age category. Not including this age group or including them as age 85 gave us an inaccurate number for average age of a state. The median household income dataset gave a median household income of various cities and towns throughout the US. At times, they gave multiple per county while others a county would be represented by only one town. Additionally, some median household incomes were 0 or exactly 300,000 which is obviously wrong so we omitted this data. We would then average all the median household incomes given for a state to give us our number which does not completely represent the actual median household income of a state since not every city and town was represented. Our education data set gave percentages of population of each state based on how many people had received a high school or higher education or a Bachelor's degree or higher so our charts do not completely show us what we wanted to know which is who had only received a high school education and not higher.

3 Hypothesis

We propose that politics is correlated with COVID-19 cases. Specifically, we think that more Republican states correlate with higher cases due to the fact that some have neglected to enact strict COVID-19 safety measures. We also think that states with a higher average age, higher income and higher education with have less COVID-19.

4 Methodology and Approach

We first aggregate political data associated with each county to determine how Republican or Democratic a county is. We then combined political data with COVID-19 cases per county. For each county, we looked at how many people voted in the 2016 presidential elections, and how many voted for democratic or republican, ignoring other political parties and also looked at how many COVID-19 cases and deaths there are in each state. Since we do not know if a person is republican or democrat when they are diagnosed with COVID-19, we decided to use the county's political results from the 2016 presidential election to estimate how many democrats or republicans got COVID-19. For example if a county is 30% democratic, and there are 10 COVID-19 cases, we added 3 cases to the democratic side and 7 cases to the republican side. We then use a linear regression to find any correlation between COVID-19 cases per population vs county average income and Deaths per COVID-19 cases vs county average income.

For our state data, we calculated a composite political score where a more positive number designates a more Republican State and a more negative number designates a more Democratic State. We then look at how factors such as age, median household income, and education levels play into cases per population percentage. We then use a linear regression to find any correlation between COVID-19 cases per population percentage vs our composite political score. We then look at how factors such as age, average household income, and education levels play into cases per population percentage.

5 Terminology

For this paper, we are going to be using the following terminology to help make things easier to understand.

1. R-value: The correlation coefficient, denoted by r, is a measure of the strength of the straight-line or linear relationship between two variables. [6]
2. R^2-value: R^2 is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. [7]
3. Correlation Strength (based on r-value): Very weak (< 0.1), weak (~ 0.3), moderate (~ 0.5)

[3] https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-detail.html#par_textimage_673542126
[4] https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations
[5] https://corgis-edu.github.io/corgis/csv/state_demographics/
[6] http://www.dmstat1.com/res/TheCorrelationCoefficientDefined.html
[7] https://www.investopedia.com/terms/r/r-squared.asp

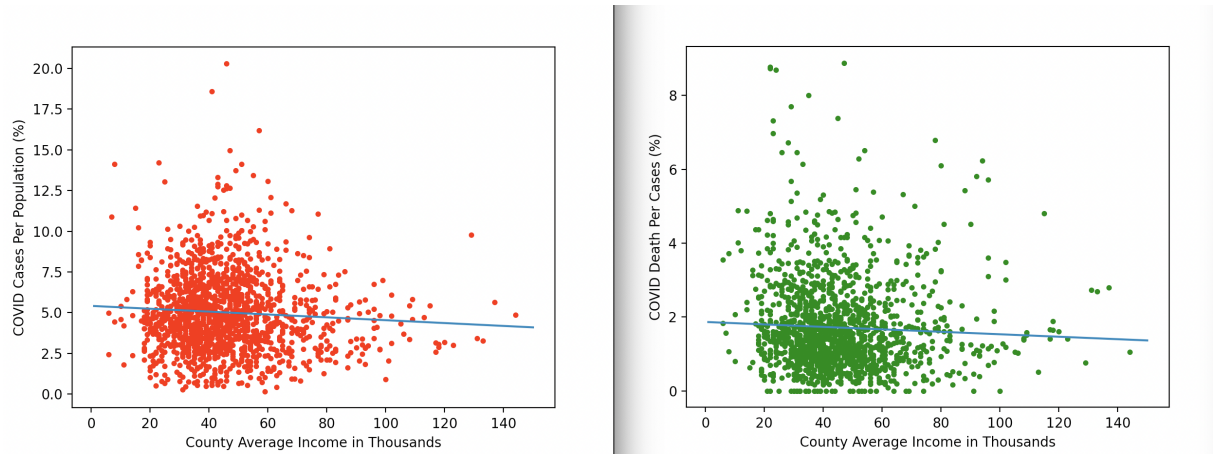
6 Results

6.1 Politics vs COVID-19

```
Democrat
------------
Population:  65350187
COVID-19:  2793444
Death:  57534
Covid/Population:  4.2746
Death/Covid:  2.0596
------------
------------
Republican
------------
Population:  62786743
COVID-19:  2786059
Death:  49861
Covid/Population:  4.4373
Death/Covid:  1.7897
```

Our first result is based on how many people voted democratic or republican in the 2016 presidential election and how many COVID-19 cases there are for each political party using our speculation method mentioned above. Here, we can see that although there were more democrat voters compared to republican voters, the republicans won more states and thus won the election. In our figure below, we can see see that democrats have a lower COVID-19 per population percentage and have a higher Death per COVID-19 cases compared to republicans.

6.2 COVID-19 Cases and Deaths vs County Income



Our second result is the linear regression between COVID-19 cases per population percentage vs county average income and Deaths per COVID-19 cases vs county average income. We can see that there is a weak correlation where counties that earn more money tend to have lower COVID-19 cases and lower deaths from COVID-19.

6.2.1 COVID Cases Per Population Percentage vs County Average Income in Thousands (Red Graph)

Linear Regression:

$$y = -0.008791130903384407x + 5.41764111796073$$

R-value: -0.06726658796615899, R^2-value: 0.004525

6.2.2 COVID Deaths Per Cases Percentage vs County Average Income in Thousands (Green Graph)

Linear Regression:

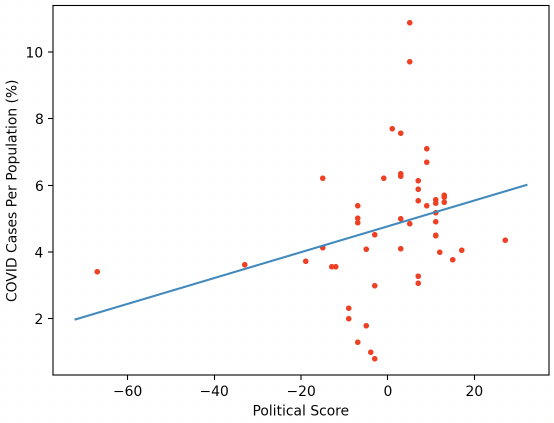
$$y = -0.003335049454034002x + 1.8672687407894062$$

R-value: -0.050119652465397097, R^2-value: 0.002512

6.3 COVID Cases Per Population vs Political Score

Our third result is comparing COVID cases per population percentage vs our political score. We can see that California is very democratic and by removing California, the slope value between the two graphs increased by around 38%. We can also see that the R-value improved slightly when California is excluded.

6.3.1 California Included

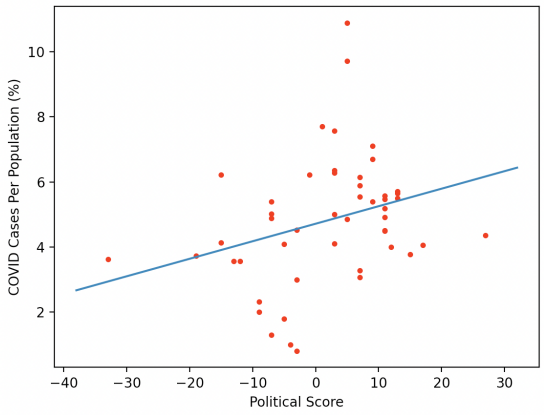


Linear Regression:

$$y = 0.03880502235880623x + 4.7670916532205965$$

R-value: -0.2877601345292418, R^2-value: 0.082806

6.3.2 California Excluded

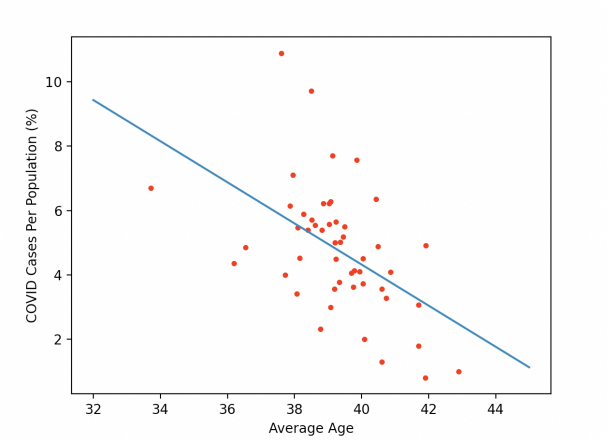


Linear Regression:

$$y = 0.05384830537612272x + 4.713345903110137$$

R-value: -0.2981101192118442, R^2-value: 0.088869

6.4 COVID-19 Cases vs Average Age

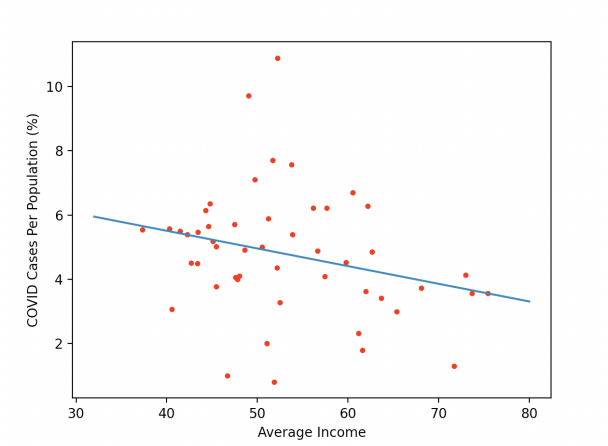


Linear Regression:

$$y = 0.638818718431507x + 19.873288528436333$$

R-value: 0.07889835247667554, R^2-value: 0.062249

6.5 COVID-19 Cases vs Income

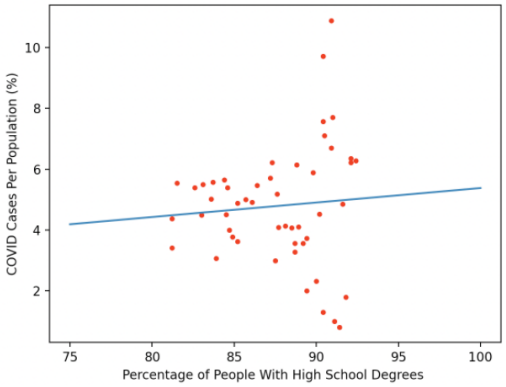


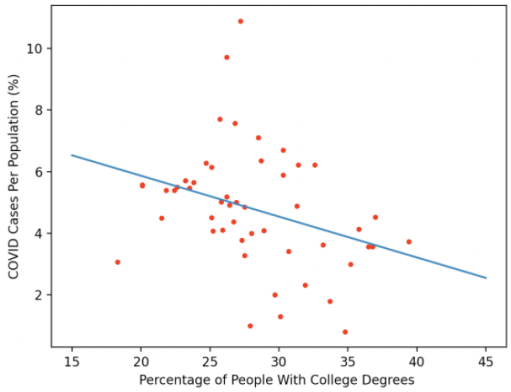
Linear Regression:

$$y = -0.055037389876353045x + 7.709478436743664$$

R-value: -0.2672591662968057, R^2-value: 0.071427

6.6 COVID-19 Cases vs Education





6.6.1 Statistics for High School Degrees

Linear Regression:

$$y = 0.047687612093757596x + 0.6095362211077733$$

R-value: 0.07889835247667554, R^2-value: 0.006225

6.6.2 Statistics for College Degrees

Linear Regression:

$$y = -0.13262252596440857x + 8.516370084623027$$

R-value: -0.3275027481939927, R^2-value: 0.107256

6.7 Regression on Multi Variables (Age, Average Income, Education, Political Score)

	coef	std err	t	P>|t|	[0.025	0.975]
const	17.6455	8.147	2.166	0.036	1.227	34.064
age	-0.6734	0.150	-4.478	0.000	-0.976	-0.370
avg_income	-0.0480	0.044	-1.102	0.276	-0.136	0.040
highschool	0.2260	0.084	2.696	0.010	0.057	0.395
college	-0.1298	0.083	-1.563	0.125	-0.297	0.038
political_score	-0.0140	0.021	-0.669	0.507	-0.056	0.028

R-value: 0.66, R^2-value: 0.436, Adjusted R^2-value: 0.372

Our Coefficients Table shows us an example where an increase in age, average household income, Composite Political score, and having a Bachelor's degree decreases the chance of catching COVID-19 while an increase in High School degree increases the chance of catching COVID-19.

With a multiple regression made up of several independent variables, the R^2-value must be adjusted since it is possible that an incorrectly high value of R-squared is obtained, even when the model actually has a decreased ability to predict. [8]

[8] https://www.investopedia.com/terms/r/r-squared.asp

7 Preliminary Conclusion

Our first conclusion is that democrats tend to have lower chance to get COVID-19 and tend to have higher chance of dying from COVID-19. However, this conclusion is not statistically significant where we can make a claim since we don't exactly know how many people who got COVID-19 are exactly democrats or republicans.

Our second conclusion is that there is a weak correlation where counties with higher average incomes are less likely to get COVID-19 and die from COVID-19. Although the data for this is more accurate compared to our first conclusion, the correlation is still very weak.

Our third conclusion is that there is a strong correlation between COVID-19 cases per population and our composite political score. However, in our hypothesis, we claimed that states that are more republican will have more COVID-19 cases and states that are more democratic will have less COVID-19 cases. Although our hypothesis is somewhat correct, we also found out from the graph that states that are slightly more republican than democratic tend to be above the regression line while states that are very republican tend to be below the regression line. Similarly, states that are slightly more democratic than republican tend to be below the regression line while states that are very democratic, in this case California, is above the regression line.

Our final conclusion is when we combined all the factors together: age, income, education, political score. We can see from the table that there is an inverse correlation between COVID-19 cases per population percentage and age, average household income, college degree and political score while there is a direct correlation between COVID-19 cases per population percentage and high-school degree which is similar to our intial conclusions.

From our research, we can conclude the following (strength of correlation):

1. Democrats are less likely to get COVID-19 (very weak)
2. Democrats are more likely to die from COVID-19 (very weak)
3. Counties with higher incomes are less likely to get COVID-19 per population (very weak)
4. Counties with higher incomes are less likely to die from COVID-19 per population (very weak)
5. States that are slightly more republican are more likely to get COVID-19 per population (moderate)
6. States that are slightly more democratic are less likely to get COVID-19 per population (weak)
7. States with a higher average age are less likely to get COVID-19 per population (moderate)
8. States with a higher average income are less likely to get COVID-19 per population (weak)
9. States with a higher high school education are more likely to get COVID-19 per population (very weak)
10. States with a higher college education are less likely to get COVID-19 per population (weak)

8 Limitations

We are doing an observational study so we cannot make any causal claims.

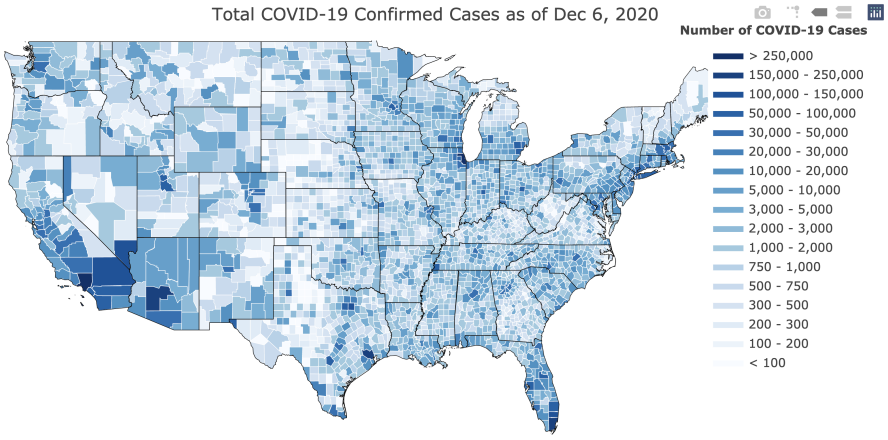
The Political Score is an imperfect measurement that does not take into account the 2020 election and deals in absolutes when it comes to Republican or Democrat, ignoring policy.

It is not possible to find out whether a person having COVID-19 is a democrat or republican since COVID-19 testing and presidential voting are two independent events.

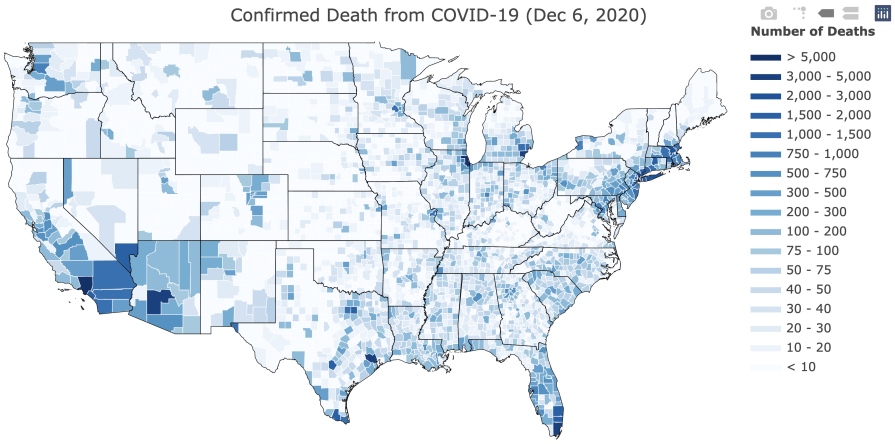
It is difficult to find sufficient data on the county level. Therefore, it is hard to portray how politics affects COVID-19 numbers locally.

9 Tables and Figures

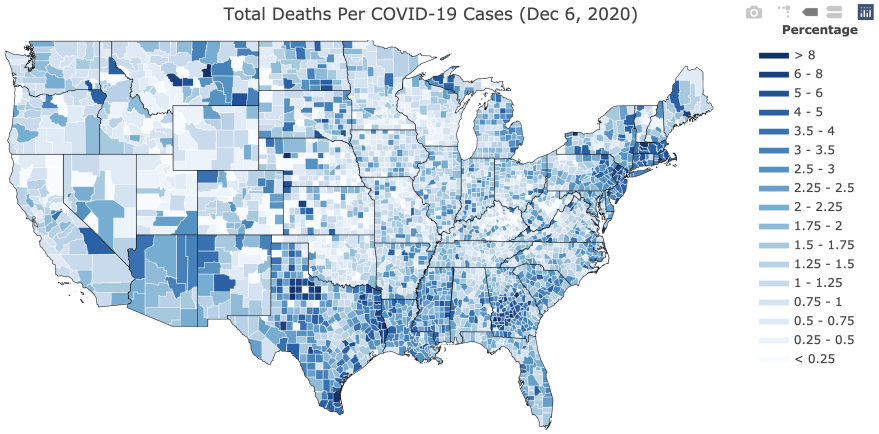
9.1 Total COVID-19 Cases County Map



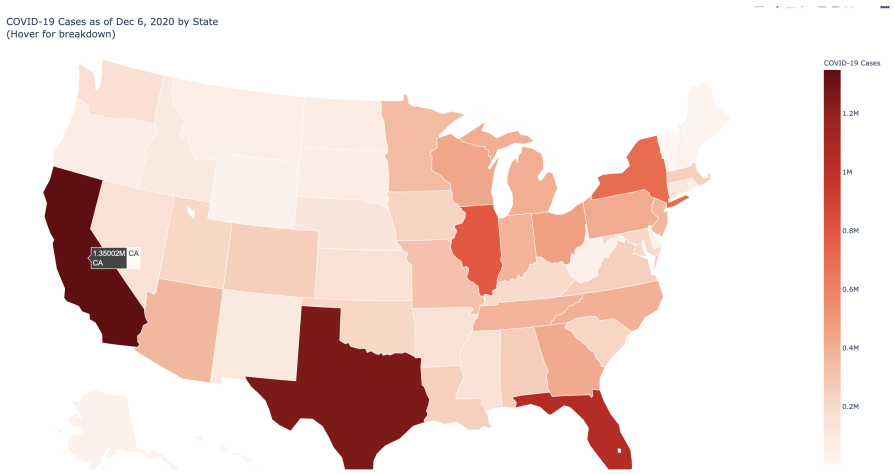
9.2 Total COVID-19 Confirmed Death County Map



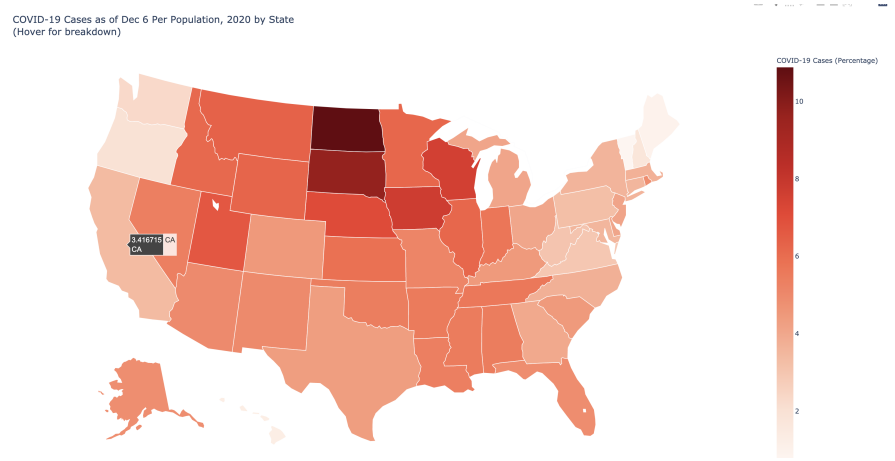
9.3 Total Deaths Per COVID-19 Cases County Map



9.4 COVID-19 Cases State Map



9.5 COVID-19 Cases Per Population State Map



References

[1] https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

[2] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VOQCHQ

[3] https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-detail.html#par_textimage_673542126

[4] https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations

[5] https://corgis-edu.github.io/corgis/csv/state_demographics/

[6] http://www.dmstat1.com/res/TheCorrelationCoefficientDefined.html

[7] https://www.investopedia.com/terms/r/r-squared.asp