Data and Programming II Final Project: Reproducible Research

Daniel Chen 12214046

December 6, 2020

1 Essential Question

Hyper-partisanship has inevitably invaded public health in the United States. While Coronavirus cases continue to surge across the country with no end in sight, party loyalty has divided the nation into those who trust public health experts and those who follow the advice of party leaders.

With this in mind, I seek to answer: What affect, if any, does party identification have on Coronavirus cases throughout the United States? In general, it appears that Democrats are relatively more likely to follow public health guidelines while Republicans listen to elected officials who downplay the severity of the virus. Using data primarily from the New York Times and 2016 election data, I am interested in finding out whether there exists a true (versus perceived) impact on public health as it relates to COVID-19 from being either a Democrat or a Republican.

2 Understanding the Python Script

Figure 1 below outlines the overall structure of my script (The images included in this .pdf are large. Please maximize the size of your window so that images render clearly!). I scrape data and load in tabular data from various sources to create a large dataframe. Using the dataframe, I create line plots, choropleths, and run regressions.

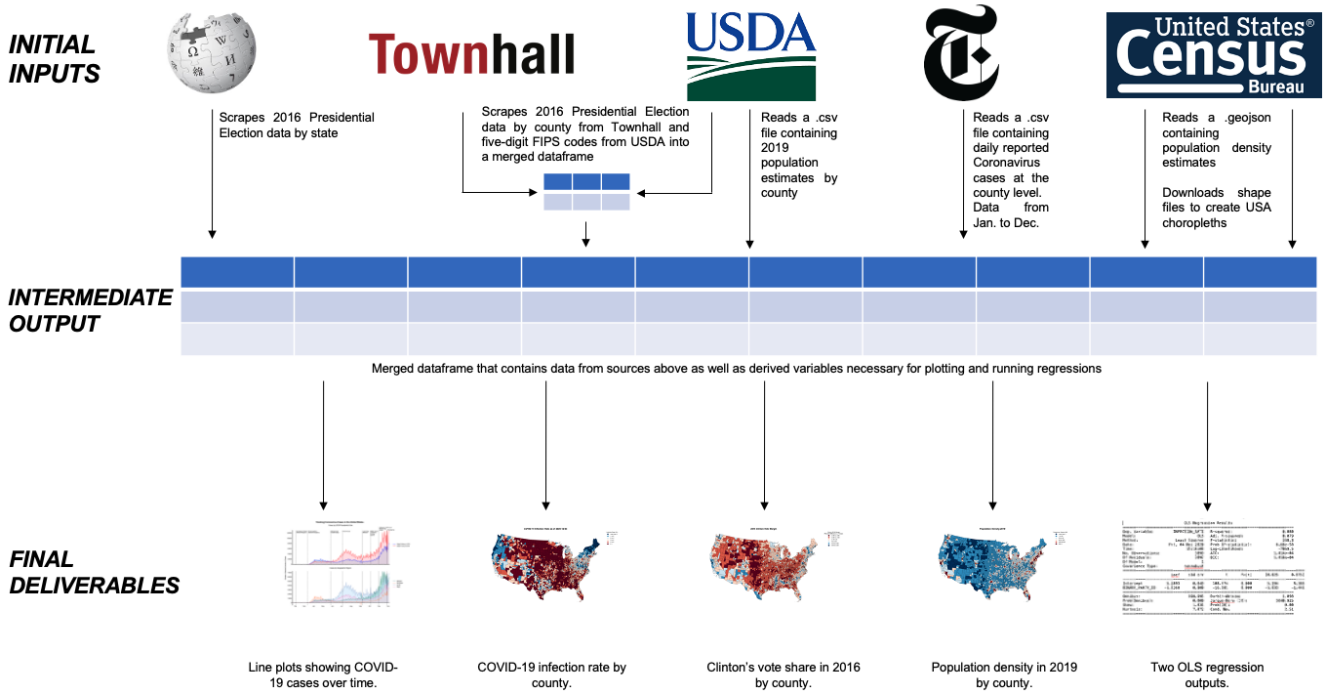

Figure 1

The script should take about two to three minutes to run. In the next section, I document all of the files that will be saved from running the script and attribute them to their respective sources.

3 Files in the Repository

Running the Python script will save the following files:

1. *Votes by State in 2016.csv*: The Wikipedia table showing Clinton and Trump raw votes in each state.
2. *Votes by County in 2016.csv*: The Townhall data showing Clinton and Trump raw votes in each county across the country.
3. *FIPS codes.csv*: Five-digit FIPS codes along with their respective county and state from USDA.
4. *Reported Daily Coronavirus Cases.csv*: Daily reported Coronavirus case count from The New York Times' GitHub Repository.
5. *Poplation Estimates 2019.csv*: Estimated population counts by county in 2019 from The USDA.
6. *Population Density Estimates.csv*: People per square kilometer by county. Data from The U.S. Census.
7. *cb_2018_us_county_500k.cpg, cb_2018_us_county_500k.prj, cb_2018_us_county_500k.dbf cb_2018_us_county_500k.shx, cb_2018_us_county_500k.shp, cb_2018_us_county_500k.shp.iso.xml cb_2018_us_county_500k.shp.ea.iso.xml*: Shape files from The US Census required to create choropleths.
8. *Final Dataframe.csv*: The final table that is produced by merging all the aforementioned data. **NOTE: The final dataframe does not contain the columns from the shape files which are used to create the choropleths because it's about 10GB large and increases the run time of the script significantly. Unfortunately, some of the derived columns I created from the scraped had to be dropped as well because of the 100 MB file size limit imposed by GitHub. For example, I group each state into a regional bucket, but this column was ultimately dropped.**
9. *Lineplots.png*: Two subplots on a single figure. First plot shows the daily change in reported Coronavirus cases grouped by states who voted for Clinton and states who voted for Trump in 2016. Second plot shows daily change in reported Coronavirus cases across the country with states grouped regionally.
10. *Infection Choropleth.png*: Map of COVID-19 infection rate at the county level.
11. *Density Choropleth.png*: Map of people per square kilometer at the county level.
12. *Vote Choropleth.png*: Map displaying Clinton's vote margin at the county level in 2016. Margin is defined as the difference in percentage of votes that Clinton received less the percentage of votes that Trump received in a given county. If Clinton received 30% of the vote and Trump received 70% of the vote in Autauga County, Alabama, the county will appear as a relatively darker shade of red and result in Clinton's vote share being -40%. This analysis restricts Clinton and Trump as the only two candidates.
13. *Total Cases Regression.txt*: Regression output for model where the total number of cases are explained by partisanship and population density.
14. *Infection Rate Regression.txt*: Regression output for model where the infection rate is explained by partisanship.

4 Analysis

This section will be divided into two parts. The first part contains observational analysis where I utilize plots and choropleths to determine trends. Plots are saved as fairly large .png files, so they may initially appear pixelated before automatically adjusting resolution on screen. The second section is more "causal" where I run regressions to back-up my findings. This is not a research methods course, so my analysis should not be taken as causal under the standards of academic rigor. Moreover, in section two, I highlight limitations of my model.

4.1 Observational Analysis

As I mentioned earlier in my Essential Question section, I am interested in tracking Coronavirus cases along party lines. Media outlets have characterized a split between Democrats and Republicans where the former appears more likely to take precautions such as mask wearing while the latter has been relatively more carefree. It should go without saying that these are generalizations. I am not saying that every single Democrat wears a mask and every single Republican does not. However, I am curious to see if the empirical evidence to back up these generalizations exists. A simple method, shown below, would be to look at the total number of Coronavirus cases among states that voted for Hillary Clinton in 2016 compared to total cases among states that voted for Donald Trump in the same year where any given state's majority vote for a candidate in the 2016 presidential election would serve as a proxy for the overall partisan lean of that state.

As a side note, I limit my analysis to the days between January 21, 2020 and December 1, 2020 for practical considerations. The NYT provides records beginning in late January and updates data daily, but I restrict the end date to the first of December because GitHub cannot handle the size of a .csv that includes data for more dates.

From the first subplot in Figure 1 below, it's clear that the states who voted for Trump in 2016 have a much larger change in the daily reported number of cases relative to states that voted for Clinton - though cases have surged across the board since Halloween. This trend has held consistent for the larger part of 2020 with the exception of a period between when President Trump declared a national emergency and when George Floyd related protests begun. The regional subplot tells a similar story. In the South - where states are typically red - cases surged during the summer relative to the rest of the country. Additionally, cases have ballooned in the Autumn months for states in the Midwest. While there are a few notable blue states in this region, the majority are red or purple.

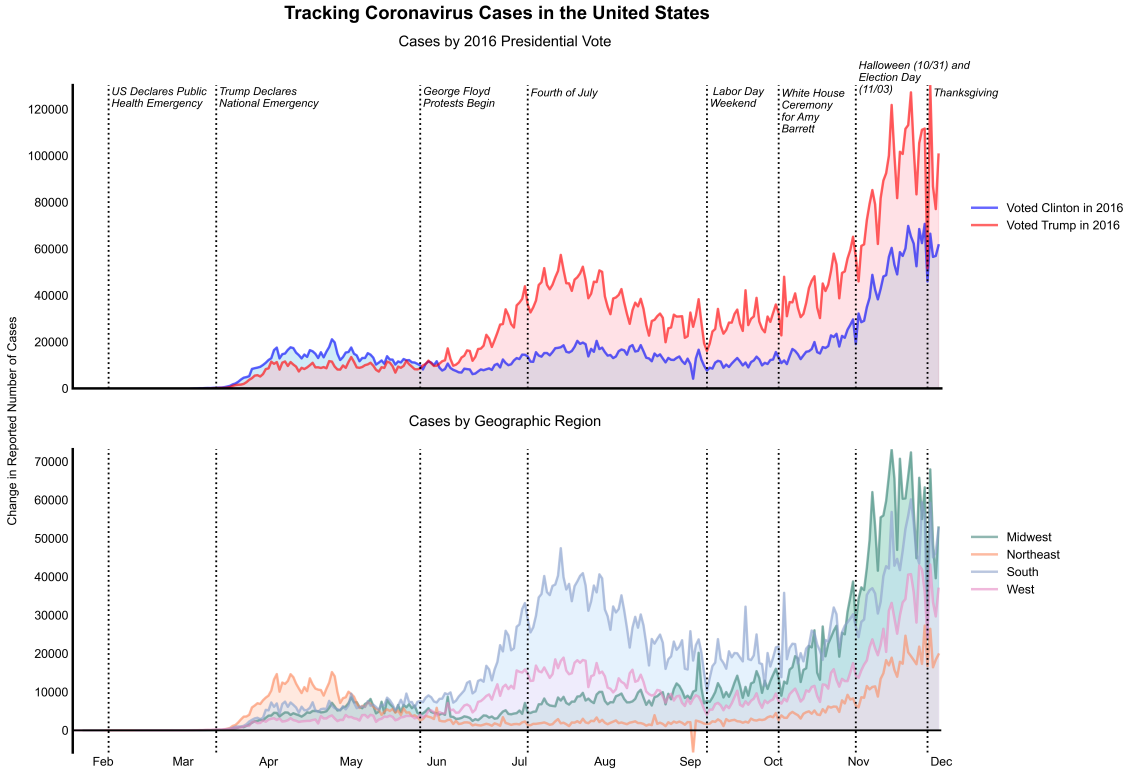

Figure 2 - Data from January 21, 2020 to December 1, 2020 provided by The New York Times

While it may be tempting to definitively claim that partisanship is tied to Coronavirus cases, this would fail to acknowledge two factors. First, states have different populations, and second, more states voted for Trump over Clinton four years ago. In other words, the larger change in daily reported cases of the virus in Republican states may be simply explained by a larger total population. With this caveat in mind, I turn to choropleths shown beginning on the next page.

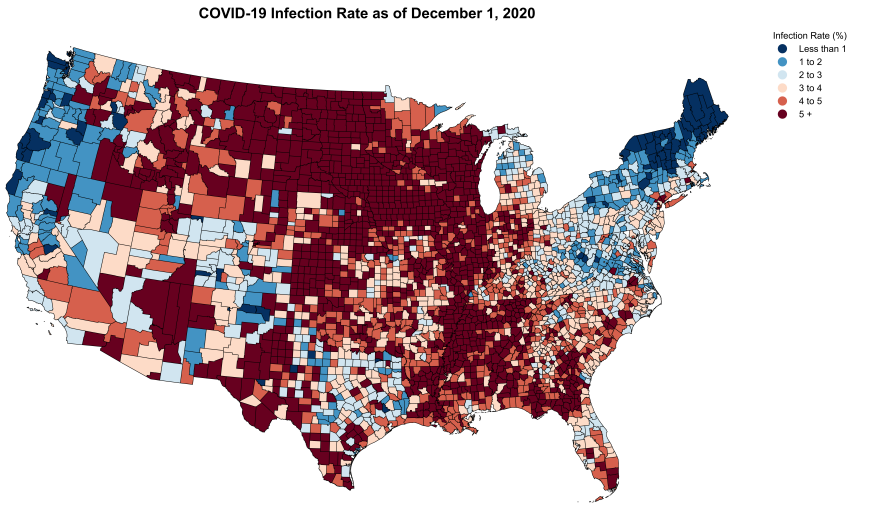

Figure 3

Figure 3, above, displays the infection rate on a county basis. I have purposefully limited the number of bins to six to more easily compare metrics across choropleths. However, it is worth noting that some counties have infection rates above 15%. I suspect that population may play a role. I create a population density choropleth, and I expect that it should resemble Figure 2 conditional on using an identical color scheme and the same number of categorical buckets. In other words, I would would expect the counties with the highest rates of infection to line up with the counties that are the most heavily populated. This rationale makes sense given CDC recommendations to socially distance and wear a mask. People living in crowded areas likely face greater rates of transmission. However, this is not the case when looking at Figure 4 below which illustrates population density.

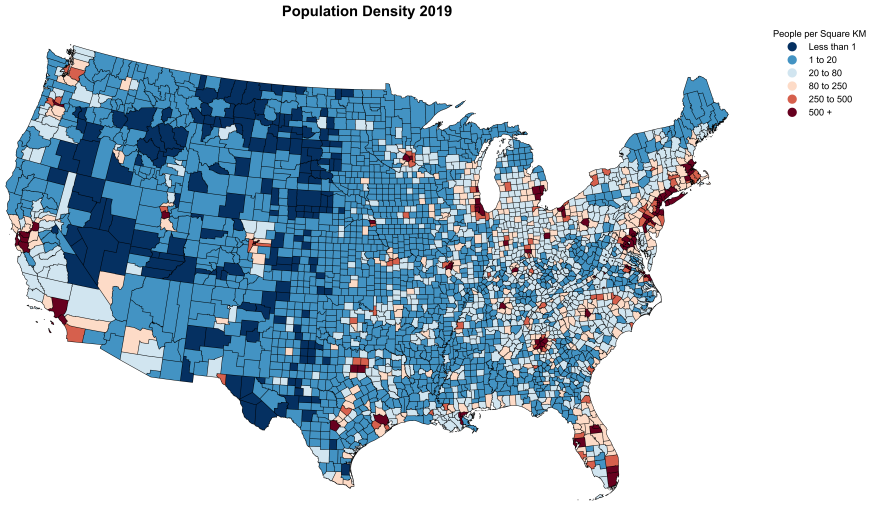

Figure 4

While there are certainly similarities between the two maps especially in looking at dense urban locations including - but not limited to - New York City, Chicago, Southern Florida, and Southern California. With these exceptions, the two choropleths are otherwise different. Large parts of the central United States, such as North Dakota, South Dakota, Kansas, and Oklahoma do not line up. These states have populations spread thin relative to coastal cities, and yet, I observe higher rates of infections. To synthesize, population may not be the *only* predictor of cases, but it still plays an important role. To explore another possible predictor, I plot Clinton's vote margin in 2016 shown in figure 5 below.

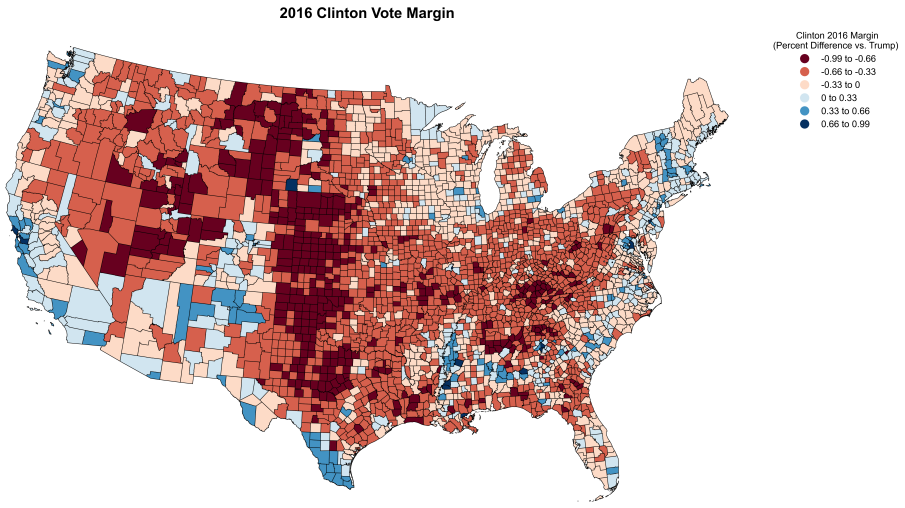

Figure 5

Figure 5 and Figure 3 are not carbon copies, but in comparing two maps, the former more closely resembles the latter as opposed to comparing Figure 2 and Figure 3. Though the intensities are not exactly identical, on the surface level it appears that counties that voted for Trump to a greater degree are also counties that exhibit greater rates of infection. In other words, the margin in which Trump wins does not perfectly predict the intensity of infection, but the two appear to be somewhat related in general. In the next section, I turn to regression analyses to examine whether or not there is a "causal" relationship.

4.2 "Causal" Analysis

In this section, I put causal in quotes because though I find statistically significant results in one of my models, the model is likely biased by many correlated factors which better explain the results. I explore these factors at the the end of this document.

In the first model I run, the dependent variable is the total number of cases and the independent variables are party identification and population estimate. Note, that in the final .txt file outputted by the Python script the Party variable is labeled as "BINARY_PARTY_ID" and the Population variable is referred to as "POP_EST_2019". Table 1 on the next page shows the regression results. The number of reported cases in Democratic counties is less than the number reported in Republican counties by a difference of about 822. This finding is statistically different from zero when controlling for the population of a county. However, in this model I worry about the high r-squared value. Moreover, Python returns a multicollinearity warning. I suspect that the population variable poses a potential issue in predicting cases too well.

As a result, I attempt to standardize the measure of spread before running the regression by looking at the infection rate per county as the new dependent variable in my second model. Table 2, also on the next page, summarizes the findings. Here I derive the infection rate by taking the total number of cases in a given county, and dividing that value by the estimated total number of people living in that county in the year 2019 in hopes of drawing a fairer comparison between places like Los Angeles County, California and Fairfield County, Connecticut where, by construction, the former is more likely to face a significantly larger number of cases because its population is much larger. In this second model, Democratic counties have a smaller rate of infection by the magnitude of roughly half a percentage point than their Republican counterparts. This difference is statistically meaningful at the 99% level of significance. Taken at face value, the model suggests that party identification has some predictive power towards the infection rate.

Table 1: Dependent Variable: Total Cases

Variable	**OLS**
Party	-822.442***
	(236.230)
Population	0.042***
	(0.000)
Intercept	88.778
	(280.501)
N	3099
R^2	0.905

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 2: Dependent Variable: Infection Rate

Variable	**OLS**
Party	-0.470***
	(0.120)
Intercept	4.815***
	(0.047)
N	3099
R^2	0.05

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

However, drawing such a conclusion would be naïve. In looking at the R-Squared value, party identification in model 2 only explains about 5% of the variation in the infection rate. Put differently, party identification is most likely correlated with other observables excluded from the model that better explain infection rates. Party identification is likely tied to race and income. I suspect income to be a primary driver of infection because individuals with higher incomes are better equipped to follow public health guidelines. They have the capital to purchase personal protective equipment and are likely privileged to work remotely where they are less exposed to human interaction compared to a low wage working individual in the gig economy who is required to serve restaurant patrons or deliver food to those who ordered online.

Moreover, income is tied to other extraneous factors such as access to health care and one's individual well-being. Higher income areas arguably have better access to health care services that mitigate the risk of infection. Additionally, higher-income earners are likely better able to afford a healthy lifestyle. They are likely to be better off in terms of personal health and wellness. Lower-income earners may face more health hardships and pre-existing conditions and compromised immune systems likely increase susceptibility to the novel virus.

The primary goals of this project were to retrieve data programmatically and leverage Python skills for data analysis. Because research methodology took a back seat to these first order goals, I was able to "explore" a relatively insignificant driver - party identification - of a serious disease that has taken over 270,000 lives. With that in mind, I would like to end on a more serious note. It's easy to engage in political divisiveness, but at the end of the day, we as citizens need to behave more intentionally to prevent the spread of the virus. Our elected officials are on the hook too to deliver economic relief and stimulus alongside affordable health coverage as they urge the nation to stay at home.