J. Ian Stewart
Professor Majid Zamani
CSCI 4022
7 May 2024

Did politics affect COVID-19 cases?

COVID-19 was the pandemic of our era. It spread quickly because people could infect others before they realized they were infected. Many local governments eventually enacted full quarantine, some including mask mandates. In December 2020, the vaccine became available to the public. However, many people protested the mask and the vaccine for political reasons. One large group of protestors was a part of the Republican party, perhaps egged on by comments from right-wing media. I hypothesized that these comments, and the subsequent actions taken by supporters, caused states that voted more Republican to have more COVID-19 cases overall.

Data Sources

I used three different data sources: a COVID-19 dataset, a U.S. House votes dataset, and the U.S. Census for 2020. The COVID-19 dataset was sourced from healthdata.gov, provided by the U.S. Department of Health and Human Services. This dataset comprised state-wide COVID-19 statistics and has 135 columns and over 81 thousand rows, where each row is a daily report from a state. A few columns are the number of inpatient beds, patient admissions with confirmed COVID-19 cases, suspected COVID-19 admissions, and ICU bed utilization. However, most of the columns are adult statistics, such as previous day admissions of adults with confirmed COVID-19 cases, but split up into age 8 groups. The columns I will mention are the previous day's admissions of adults and pediatric patients with confirmed COVID-19 cases, which I combined into a “total_previous_day_cases” column. I decided to stick with confirmed cases, rather than use suspected because a respiratory illness like COVID-19 has a lot of similar symptoms to the Flu.

The U.S. House votes dataset was sourced from dataverse.harvard.edu, provided by MIT Election Data and Science Lab. This dataset has the number of votes for each candidate in every House election since 1976. However, I only used data for 2020 and 2022 because that is when COVID-19 happened. This dataset has 20 columns, but I used state, party, candidate votes, unofficial election, total votes, and district. I used the unofficial column to filter out unofficial election data and removed data before 2020. When doing my data analysis I grouped 3rd parties into one “other” party.

The final dataset is the U.S. Census for 2020. This dataset was each state’s resident population for 2020, including D.C., Puerto Rico, and the Virgin Islands, which I did not use as they do not vote in the U.S. House of Congress elections. Originally, I did not use the census data but brought it in because I realized I had to normalize the COVID-19 cases for each state as the discrepancy in the populations of states is massive. If I had just stuck with votes, the same issue would also occur with the house voting data. However, instead of normalizing with the census data, I decided to plot the number of votes for a political party in a state divided by all votes in that state, or the total proportion of votes cast for a political party. I normalized the political data as proportional rather than normalized with the census data because only about 50% of each state's population votes in House elections.

This question has been asked before. Numerous studies have been done on data similar to this about the impact of politics on COVID-19 cases. This study by Rongxiang Rui, Maozai Tian, and Wei Xiong, on the impact of political ideology disparity on COVID-19 transmission in the U.S., found that "individual-level disparity of political ideology has impacted the nationwide COVID-19 transmission…". This other study by Brian Neelon, Fedelis Mutiso, Noel T Mueller, John L Pearce, and Sara E Benjamin-Neelon, on governor political affiliation and COVID-19 cases, deaths, and testing "...suggest that governor political party affiliation may differentially impact COVID-19 incidence and death rates". These studies did not find conclusive evidence for their claims, but they did show that there is some correlation to investigate.

Methods

The data science techniques I used were simple linear regression and the k-means clustering algorithm. Simple linear regression was used to show general relationships in the data, with the equation being:

*the proportion of COVID-19 cases = m * proportion of votes for one party + b*,

where m is the slope and b is the constant from simple linear regression. I decided to use the k-means algorithm because the data points are states. I hypothesized that the clusters of states it would produce would be grouped by geographic location. If the data were geographically clustered, it would help to answer how COVID-19 cases were spread throughout the nation. Additionally, if they aren't geographically clustered, maybe there is some other reason states are clustered together.

I first grouped the COVID-19 data by year and state to organize my data and summed all the "total_previous_day_cases" columns. This gave me all the confirmed COVID-19 cases for each state for 2020 through 2023, which I divided by the census data for each state to get a normalized measure. For the political data, I grouped by year, state, and party and summed the votes for each party. American politics is mostly concerned with the two main parties, Democrat and Republican, so I decided that any third-party votes would be grouped into an "OTHER" party. I then divided each party's votes by the sum of all votes cast in that state's election to get the proportion of votes for each party. By grouping the data like this, I could make each state a point on the plot, where the x-coordinate was the proportion of political votes for a party and the y-coordinate was the normalized confirmed COVID-19 cases for that state. I then did this for each year starting with 2020.

However, there is a caveat to the COVID-19 data for 2020: not all states reported COVID-19 cases for the entire year of 2020. This was since COVID-19 started in a few states and eventually spread to all states by March. I decided to keep the data this way because the cases near the beginning of COVID-19 paled to the later numbers of COVID-19 cases.

Results

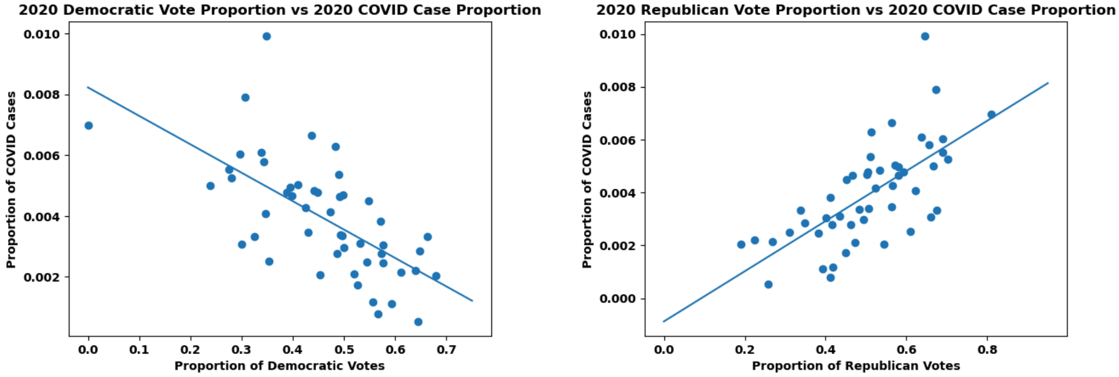


Figure 1 (left) and Figure 2 (right). *The above graphs show the plot of the proportion of the votes for Democrats (left) and Republicans (right) vs the proportion of COVID-19 cases for each state. The trend lines shown are simple linear regressions of the data. For the Democrats, there is an r^2 value of 42.5%, whereas for the Republicans there is an r^2 value of 46.6%. In the left graph, the data point with no democratic votes is South Dakota, which didn't have a Democratic candidate in the 2020 House elections.*

The trend line in Figure 1 suggests that in 2020, the proportion of COVID-19 cases was negatively correlated with the proportion of Democratic votes in the House 2020 elections. The trend line in Figure 2, however, suggests that the proportion of COVID-19 cases correlated with the proportion of Republican votes in the House 2020 elections. This trend was what I expected as the studies surrounding this topic agree that Republican-voting states had higher COVID-19 cases on average. To analyze this data more, I then decided to cluster this data, as each data point is a state. I hypothesized there might be some geographic explanation for this trend. So, first I generated the necessary elbow plots to choose the correct value for k.

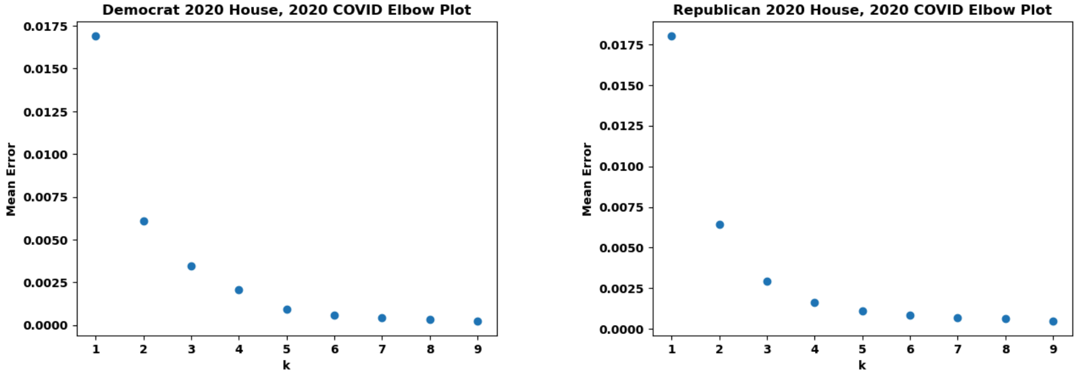


Figure 3 (left) and Figure 4 (right). *Shown above are the elbow plots for Democrats in 2020 (left) and Republicans in 2020 (right). For both elbow plots, I chose a cluster size of k=4 as it was the best cutoff before the mean error changed significantly.*

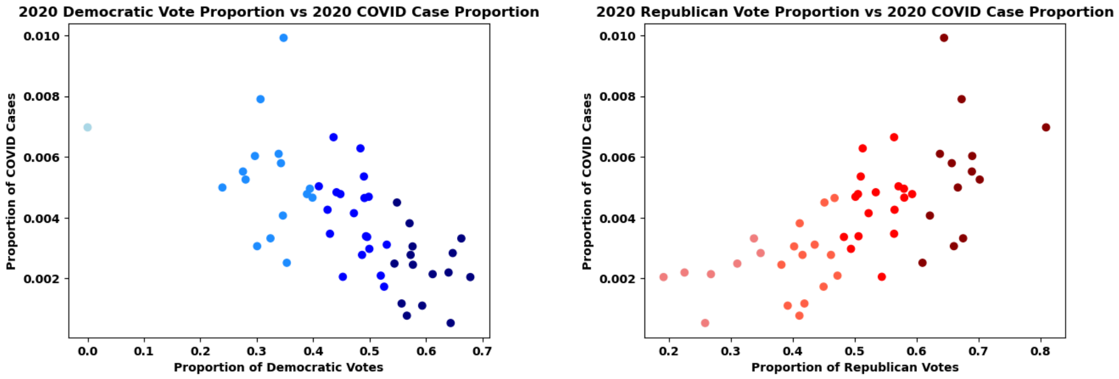
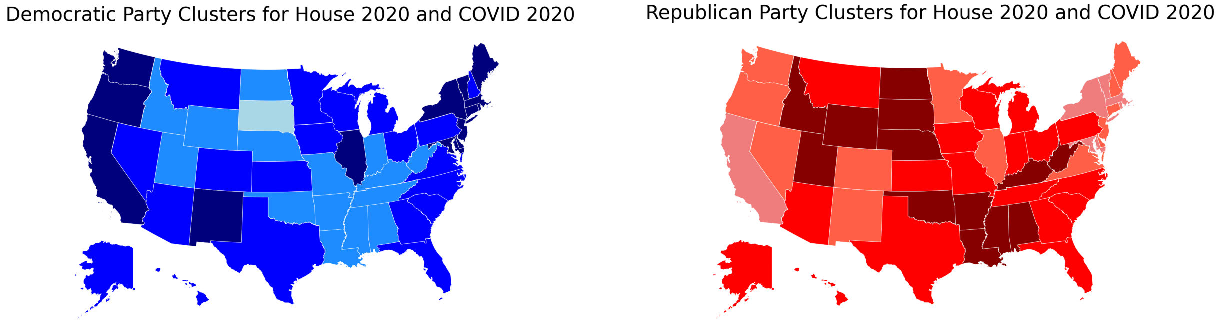


Figure 5 (left) and Figure 6 (right). *This is the clustering assignments for the data shown in Figures 1 and 2. Democratic clusters are shown on the left, while Republican clusters are shown on the right.*

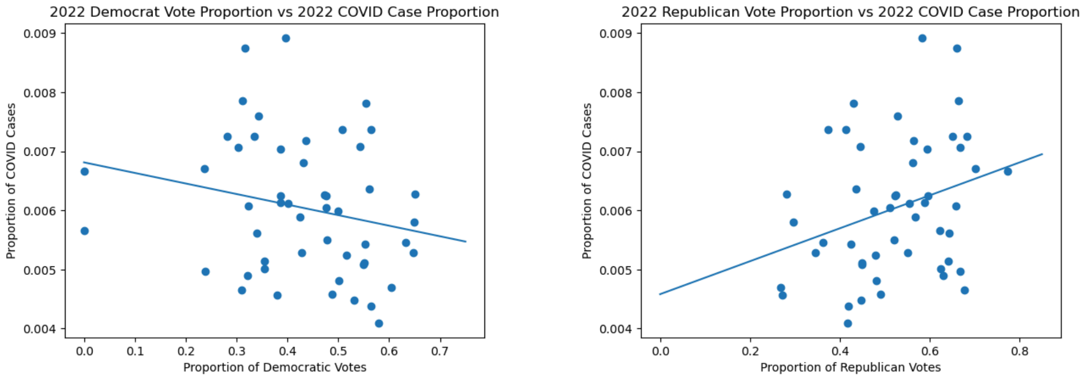
These clustering assignments shown in Figures 5 and 6 are difficult to interpret just by looking at the clustering assignment and labels would become overbearing fast as there are 50 data points on the one graph. So, to visualize this I used GeoPandas to plot the color assignments to a map of the United States.


Democratic Party Clusters for House 2020 and COVID 2020
Republican Party Clusters for House 2020 and COVID 2020

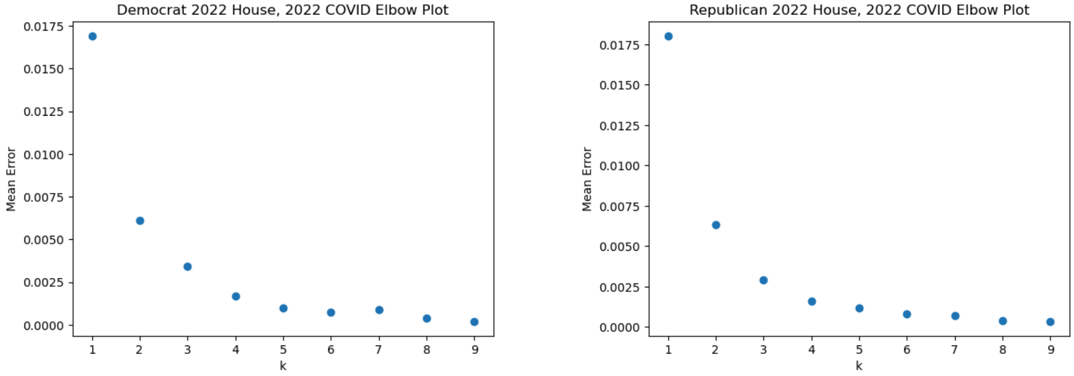
Figures 7 (left) and 8 (right): *The colors shown correspond to those in figures 5 and 6. Note that the darker the color, the higher the proportion of votes for that particular party in that state.*

Unfortunately, the geographic clusters I hypothesized were not to be. In Figure 7, the cluster that hosts states with the largest proportions of Democratic votes is split into two main clusters, one on the West Coast and another in the Northeast of the contiguous United States. These individual clusters are what I was expecting each actual cluster to look like. This means the clusters shown are based on state-level political affiliation, rather than geographic clustering. There is still some geographic clustering with highly republican clusters seen in Figure 8 in the South and North West and Figure 7 with the clusters I mentioned previously, but these clusters are not fully connected.

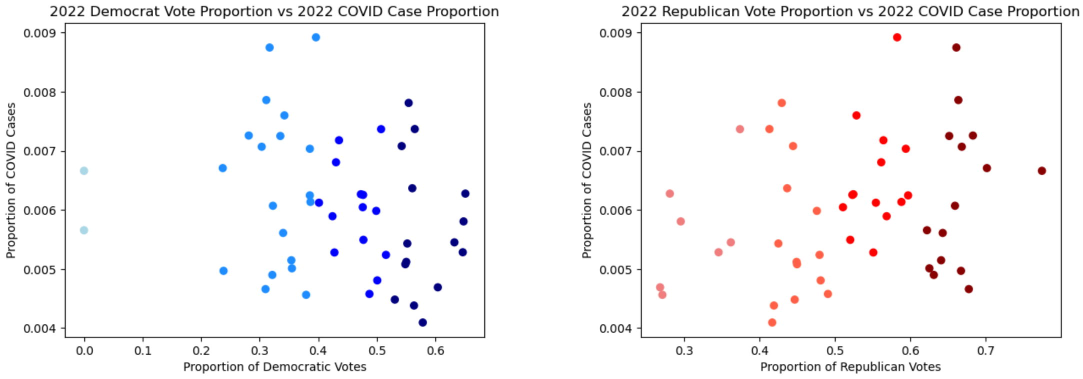
Figures 1 through 8 all show the year 2020 for the Republican and Democratic parties, which showed a general trend that the more Republican a state voted, the more COVID-19 cases the state was expected to have. The opposite was true for the Democratic party in 2020, but I wondered if this trend continued. To do this I repeated the previous analysis for the year 2022.



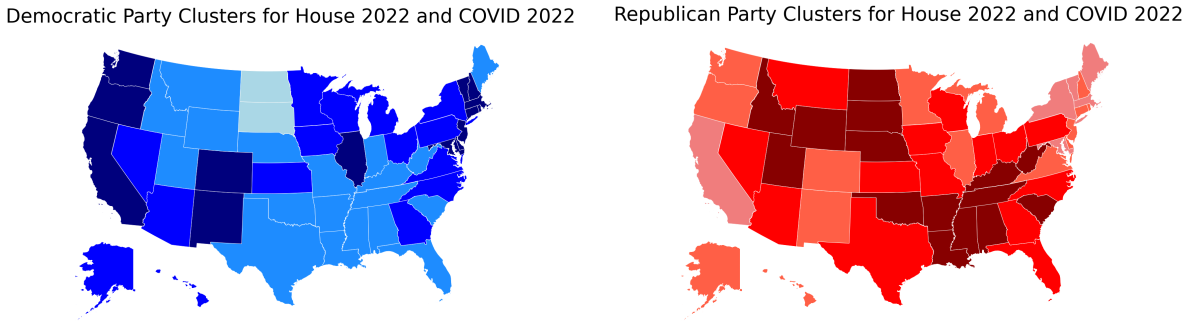
Figures 9 (left) and 10 (right): *The above plots use the same axes as Figures 1 and 2 respectively, but House and COVID-19 data for 2022. The trend line shown is a simple linear regression of the data. The Democratic trendline has an* r^2 *value of 4.91% and the Republican trendline has an* r^2 *value of 9.0%. In Figure 9, the two points with no Democratic party votes are North and South Dakota, which did not have a Democratic candidate for their 2022 House elections.*



Figures 11 (left) and 12 (right): *Shown above are the elbow plots for the Democratic party (left) and the Republican party (right) for the year 2022. For both elbow plots, I chose k=4 clusters as the optimal cutoff before the mean error started changing insignificantly.*



Figures 13 (left) and 14 (right): *The clustering assignments shown for the Democratic party (left) and the Republican party (right) in 2022.*


Democratic Party Clusters for House 2022 and COVID 2022
Republican Party Clusters for House 2022 and COVID 2022

Figures 15 (left) and 16 (right): *The colors shown correspond to colors in Figures 13 and 14 respectively.*

The trend that was seen in 2020 is almost nonexistent in the year 2022. There is still a minor correlation between the proportion of votes and the number of COVID-19 cases, but it is now negligible. However, while the trend changed, the clusters did not change much. There were some changes, such as Texas, Florida, and North Dakota, but these were minor changes and could be because these states are on the fringes of being in either cluster. The Republican clusters also did not change significantly in 2022. Once again, this is because the clustering is mostly based on state-political affiliation rather than geographical location.

Conclusions

This change in the trend from 2020 to 2022 could be due to a variety of factors. COVID-19 would be almost 3 years old by this time and there would have been a lot more information available to people, compared with 2020. This could have influenced their decision on quarantine and mask recommendations. Additionally, the vaccine would have been widely distributed by this time, leading to fewer COVID-19 cases. This can be seen in the comparison between Figures 1 and 9 and 2 and 10, where the maximum proportion of a state with COVID-19 cases dropped from 1% to 0.9%. This might seem insignificant, but on the scale of states, a 0.1% decrease means hundreds of people are not getting sick.

This was a general analysis of this data and the relationship between COVID-19 and politics. K-means turned out not to help uncover relationships between COVID-19 and politics as it revealed which states were more politically affiliated with one another. However, it showed a positive correlation between the number of Republican votes and the number of COVID-19 cases a state had in 2020. The opposite was true of the Democratic party, where a negative correlation was found between the number of votes and number of COVID-19 cases. However, this trend did not continue into 2022, potentially because of the distribution of the vaccine. Because this is a state-level analysis of politics, any results are very generalized. Most states have large populations of voters for both parties, so a more specific analysis could be done on the counties of each state, which would reveal more specific trends between COVID-19 and votes for political parties.

References

Rui, R., Tian, M. & Xiong, W. Exploration of the impact of political ideology disparity on COVID-19 transmission in the United States. BMC Public Health 22, 2163 (2022). https://doi.org/10.1186/s12889-022-14545-3

Neelon B, Mutiso F, Mueller NT, Pearce JL, Benjamin-Neelon SE. Associations between governor political affiliation and COVID-19 cases, deaths, and testing in the United States. medRxiv [Preprint]. 2021 Jan 6:2020.10.08.20209619. doi: 10.1101/2020.10.08.20209619. Update in: Am J Prev Med. 2021 Jul;61(1):115-119. PMID: 33106818; PMCID: PMC7587838.