I. Abstract

In the United States, COVID - 19 has become a wildly politicized issue. From mask mandates, to stay at home orders, citizens have received contradictory messages and dis-information regarding proper protection from and the severity of the virus. While some parts of the country have adopted effective policies to slow the transmission of the disease and have been able to reopen their economies, others have had a more difficult time both reopening and slowing the spread of COVID-19. With these diverging patterns and the political nature of the pandemic, we used a data science approach to evaluate if parts of the nation which voted for President Trump in 2016 had better outcomes regarding COVID-19 than areas that voted for Hilary Clinton. To do this, we mined publicly available data sources and compared COVID-19 death rates, positive cases per capita, and unemployment with the voting results at the state and county level and key co-variables. Spatial autocorrelation methods and random forest models revealed that political affiliation had little to do with how different counties fared with regards to COVID-19 outcomes.

II. Introduction/Background/Question/Methods Overview

There is no one person, institution or community in the United States that has not felt the impact of the devastating coronavirus outbreak taking place in this nation (Fig 1) and around the world (Figure 2).

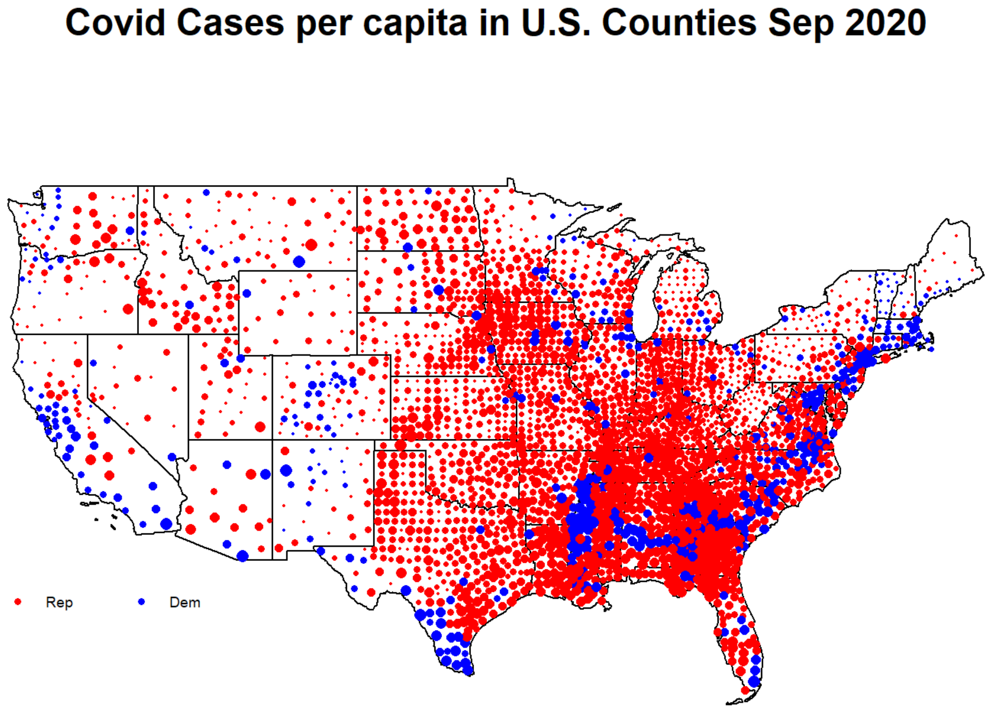


Figure 1 COVID -19 Cases per capita by county. Centroids are proportional to cumulative number of cases

Many historic events have taken place in the country during this unprecedented time, all in the midst of one of the most critical years for the American people - the year we elect a president for the next term. With COVID-19 being the overriding theme of political campaigns, bills have been passed, policies have been enforced and many statements have been made creating divisive political sentiment regarding the virus. One notable remark made by the U.S. president is

"If you take the blue states out, we're at a level that I don't think anybody in the world would be at. We're really at a very low level. But some of the states, they were blue states and blue-state-managed."

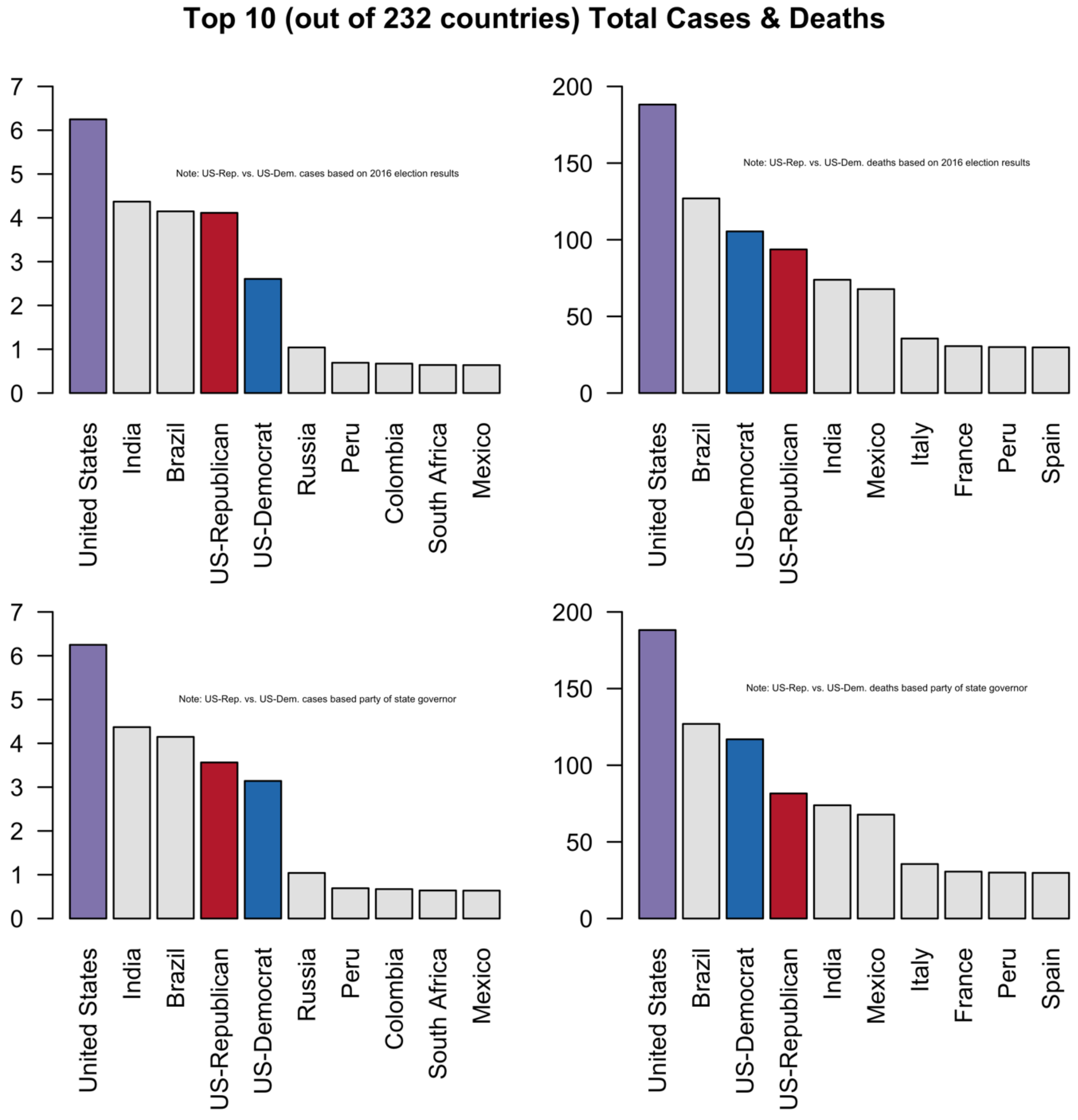


Figure 2 Country ranking of cumulative COVID -19 cases (left) and deaths (right) with the US broken down into cases by 2016 election results (top) state level party affiliation (bottom). Bar plots show there is little difference between red and blue counties

Although we may hope it was this simple, different states may have encountered varying effects of the virus based on a plethora of variables (e.g. co-morbidities, mask mandates, relaxed restrictions, etc.). In this study, we evaluate the validity of President Trump's statement by answering the question: "did areas of the nation that voted for President Trump in 2016 have better outcomes regarding COVID-19 than those which voted for Hilary Clinton?" Before answering this question, however, we needed to define what exactly we meant by "outcomes". Within the scope of this project, we define these to be cumulative numbers of cases, deaths, and unemployment (as a proxy for economic impacts on individuals). Higher counts of these outcomes indicated a poorer outcome. We investigate if dominant party affiliations among states and counties influenced the outcomes of COVID-19 by visually mapping and comparing to regional properties from geospatial, health and socio-economic variables, and run spatial autocorrelation methods and random forest models to test for significance of the effect of party affiliation on COVID-19 outcomes.

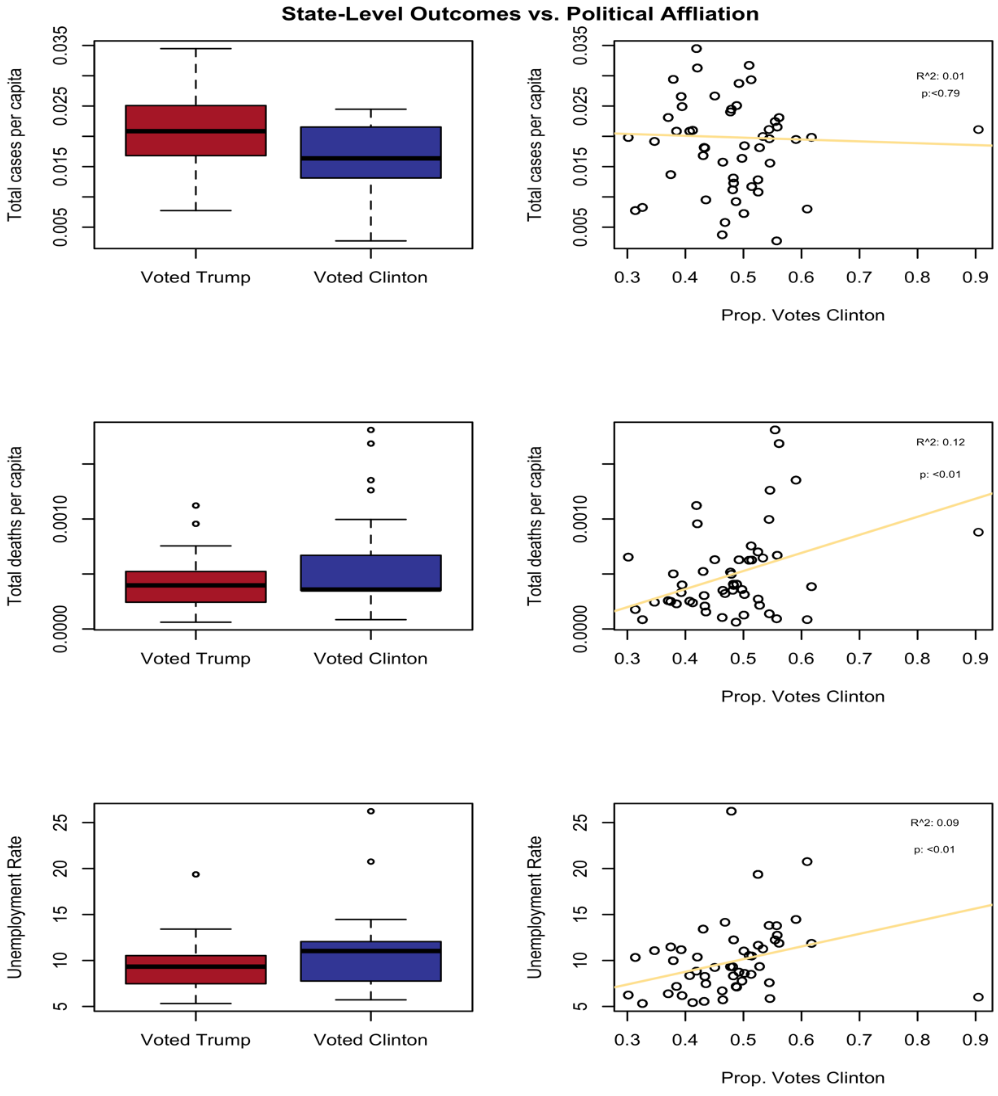


Figure 3 Box and scatter plots of COVID- 19 outcomes vs. party votes in the 2016 election. Box plots show little difference among means of outcomes while scatterplots show performance of the linear models.

III. Results and Discussion

After removing the cases and deaths attributed to blue states, the United States did not exhibit very low cases or deaths, and ranked 3-4 worldwide (figure 2). Box plots and regressions indicated no significant difference between the parties and increased unemployment, and significant, but weak positive relationships with COVID19 cases and deaths (Figures 3). All data were significant at the p < 0.05 level. significant relationships with the COVID19 outcomes at both the county and state level (Figure 4). Random Forest Regression of the predictors in Figure 4 identified that contributed most to COVID - 19 Case Counts, COVID - 19 Death Counts, and Unemployment at county scale.

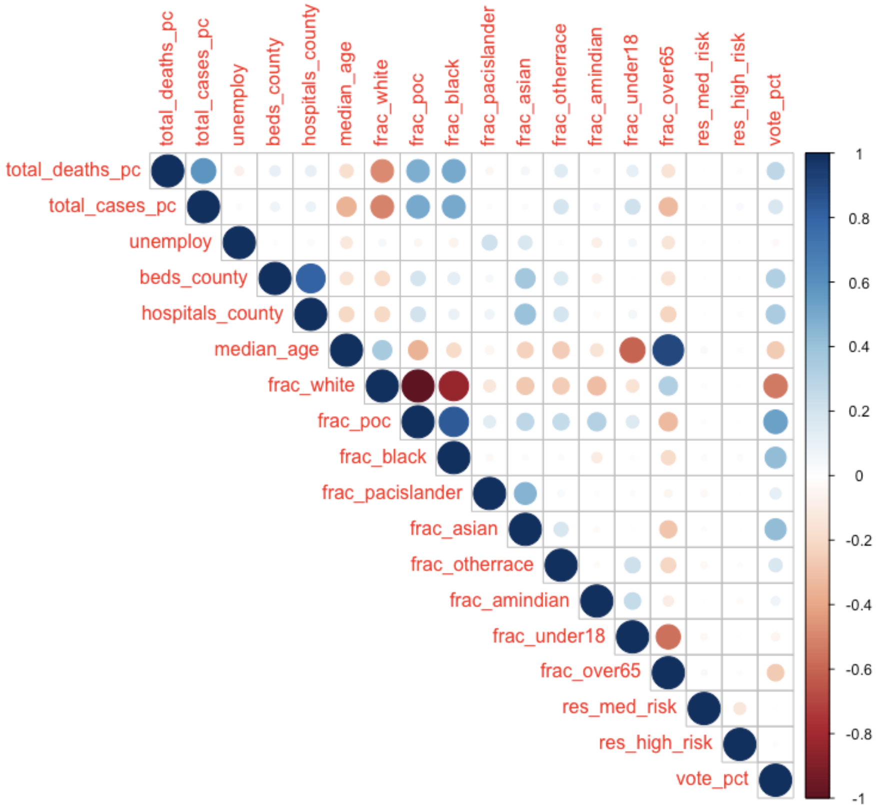
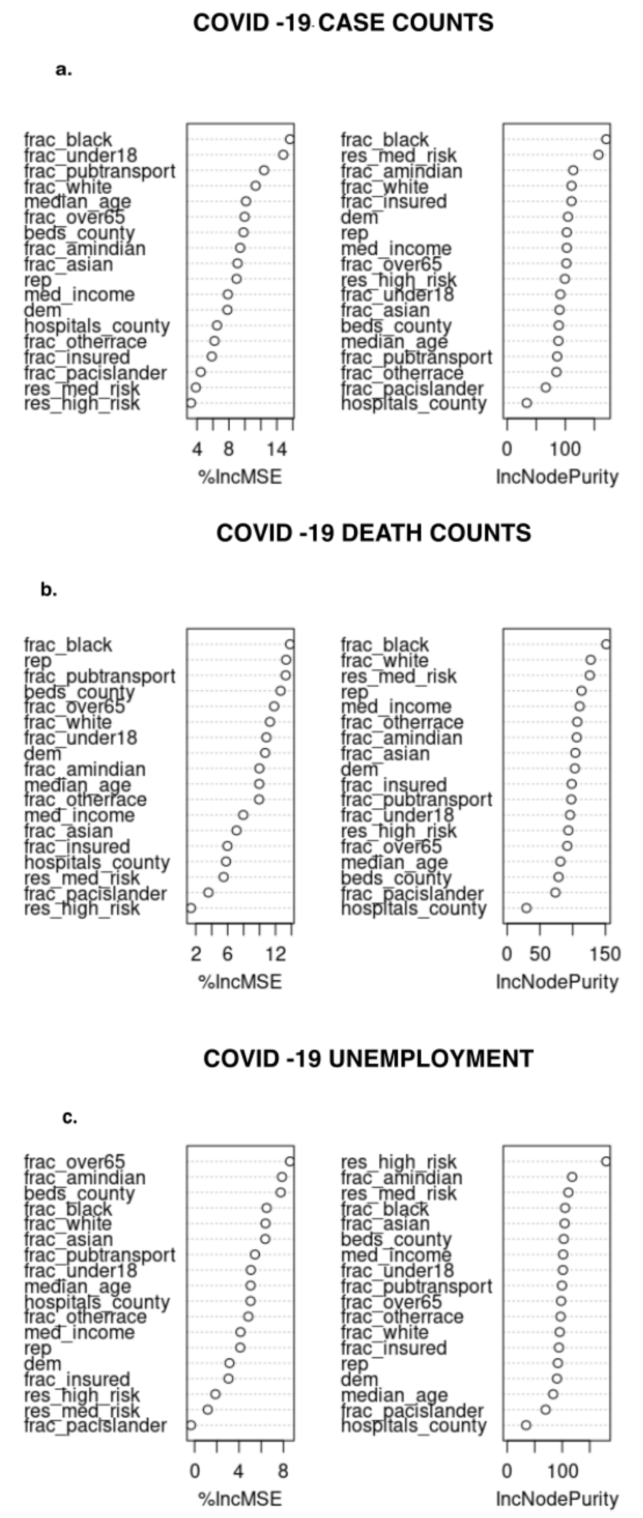


Figure 4 Correlation Matrix for simple linear regressions

Some challenges with this model were that the test and validation errors were high, possibly due to the use of Median Absolute Error opposed to Mean Squared Error. Additionally, our test Coefficient of Determination (COD) was low, but the validation COD was strong, indicating that this model performed adequately. For each random forest model the COD was 0.64,0.66, and 0.67 for COVID-19 cases, deaths, and unemployment respectively, revealing that the model was moderately predictive for all the variable causes for the outcome variables. Median

Absolute Error was calculated, and was high, at 0.26, 0.26, and 0.31. This error measure was chosen because it is more robust than other measures. However, because we used 70% of the data for training purposes, it's possible that the model may have been overfitted. Despite this uncertainty, all three models accounted for spatial relationships as follows: COVID -19 case Counts were not significantly related to party affiliation with a p-value of almost 0 to 0.63.

The COVID -19 deaths model yielded a low significant relationship with a p-value of almost 0 to 0.046. A better model should account for more of the spatial variance, but a spatial model might be needed. Regardless, the 0.05 significance value is highly debated due to false discovery rate and many corrections lower this to 0.01. The unemployment model had an insignificant relationship with a p-value of 0.18. The feature importance plots (Figure 5, left) show political affiliation to have little impact on predicting COVID-19 cases, deaths or unemployment numbers. This point is reinforced by the even smaller differences between each predictor variables node purities (Figure 5, right). None of the models proved very strong, suggesting that political party affiliation does not affect COVID -19 counts or deaths. The random forest error was high and the COD was very low on the test data. Because of the small differences in predictive ability of party affiliation within the model and the higher ranking of other factors such as race and community resilience, we interpret this to mean that there are other factors such as social and racial inequities which have a greater influence on the outcomes than politics. The fact that the models only resolve a portion of the variance seen in the data suggests that the correlations seen may be capturing latent variation beyond what the variables directly measure, such as inequality, stress, and health disparities. Furthermore, the residuals of our final random forest model were insignificant informing us that spatial autocorrelation had been accounted for within the model. These findings would suggest that in the face of future health


COVID -19 CASE COUNTS

COVID -19 DEATH COUNTS

COVID -19 UNEMPLOYMENT

Figure 5 Feature importance (left) and node purities (right) for random forest models of COVID -19 cases (top), deaths (middle), and unemployment numbers (bottom). All models show red and blue parties to be very similar in their predicting capabilities.

crises, we as a nation should seek to dissolve these inequities to increase our resilience.

IV. Challenges and Limitations

This 2-day project presented a great challenge: collecting, harmonizing, and analyzing, and synthesizing diverse data and results about a complex multi-faceted issue that is continuously evolving, with very limited resources to address the framing question. This challenge emerged particularly in dealing with the temporal aspects of the data, which resulted in several variables like domestic and international flights, sentiment data, and non-pharmaceutical response measures not being included in our final analyses. This relevance and variability in temporal and spatial resolutions meant that setting up and running models was quite challenging. We note that the omission of temporal differences in outcomes hinders us from evaluating outcomes over-time as the world learns more about the virus with regards to treatment, transmission and safety measures.

V. Conclusions

Although some of our models were significant, we believe there is no concrete evidence that supports President Trump’s statement. Many of our models found no significant differences between outcomes and party affiliation, indicating that COVID-19 cases and deaths are not predicted by just party affiliation. We believe there are many co-variates that we were unable to capture within our model, and COVID-19 isn’t a virus that discriminates based on party affiliation but rather makes visible underlying vulnerabilities and inequalities that may exist in a given society. Given further resources and time, our modeling efforts would seek to incorporate additional relevant data such as the degree of success of response measures like social distancing and mask wearing, as well as a temporal perspective to help describe the spatiotemporal variability of virus spread and peaks in relation to timing of response measures.