Predicting U.S. COVID-19 Cases

Nicholas Archambault, Melissa Lu

31 October 2020

Introduction

The United States, a global leader in technology and innovation, has fared among the worst of all nations in mitigating the spread and impact of COVID-19. Cases are swelling toward a third peak as daily averages rise in four out of every five U.S. states, a resurgence driven by milder temperatures and a populace weary of the months of restrictions that social distancing and lockdown measures have imposed on their lives. Nine months into the pandemic, the country just recorded its worst week yet, with 528,927 new cases and a seven-day average of over 75,500, according to the *New York Times*. *USA Today* reports that a one-day record 88,521 cases were reported on October 29, equivalent to a new diagnosis every 0.976 seconds.

In addition to fighting a deadly pandemic, dealing with financial concerns and adhering to social distancing measures, the American public has been forced to confront a wave of disinformation regarding the virus and the proper steps our society must take to contain it. Science has become politicized to the extreme, in the moment we can least afford it to be. With an eye toward one of the most consequential and contentious elections in national memory, the president and other politicians have consistently downplayed the virus. Instead of formulating a sound response to the worst mass-casualty event in national history, they have prioritized undisguised political maneuvering to deepen the stark partisan divide among Americans over the costs and benefits of various public health measures.

Given the breadth of factors contributing to America's uniquely poor handling of the pandemic, any attempt to comprehensively analyze and track the spread of COVID-19 must consider the effects of influences beyond climbing case totals. This project integrates not just health data from the past six months, but also political preferences, social distancing and economic reopening procedures, and access to healthcare and resources in order to predict COVID-related deaths across the United States in the second week of November.

Data were collected from 11 diverse sources, substantially cleaned and manipulated, and combined into a dataframe that provides American states' individual totals not only for health-related statistics like positive cases and deaths, but also for engineered features such as population density, access to hospital beds per 100,000 civilians, and survey responses on mask-wearing practices.

Using this collection of traits and features germane to all the different angles of this pandemic, we attempt to predict state-by-state COVID-related deaths in the second week of November, 2020.

Data

The following section will provide a brief overview of all datasets collected and cleaned for this project, including their sources, important features, and the processes required to manipulate them into useful form.

Health Data by State

Our primary source for data on the spread of the pandemic itself was The COVID Tracking Project by *The Atlantic*, a repository of volunteer-collected data updated daily from every U.S. state and territory. The data is freely available to download at the project's website and contains state-by-state totals for every individual day going back to January 28 of this year.

The dataset initially held 13,213 rows and 43 different variables. We first eliminated all data prior to April 12 in order to provide uniformity with the other health dataset we analyzed from Johns Hopkins University.

We also restricted the scope of the project to the 50 U.S. states and dropped all data from non-state territories, including Washington D.C. In doing so, we pre-emptively avoided the possibility of having to deal with non-uniform datasets, some of which may contain territorial data and some of which may not.

Finally, we removed all variables with more than 1,000 NA values, leaving the dataset with 9,850 rows and 16 variables.

The other major data source on the spread of the pandemic was taken from the Johns Hopkins University Coronavirus Resource Center, widely considered a reliable authority for carefully tracking the virus' spread. The Github profile from which the data was pulled is attributed to the Center for Systems Science and Engineering at Johns Hopkins.

Similar to the *Atlantic* data, the JHU dataset provided data from each state and territory for each day between April 12 and October 25. Once more, we eliminated all non-states and all variables with a significant proportion of NA values. The final dataset contained 9,850 rows and 15 variables.

Colleges

Next, we read in the U.S. Department of Education College Scorecard, a large annual summary of information on every college in the nation collected by the federal government. This dataset initially contained over 6,000 rows of college listings and nearly 2,000 variables, but we kept only four: each college's state, undergraduate population, cost if private, and cost if public.

Our rationale for examining college data stemmed from our own experiences as students subject to rigorous travel restrictions and community conduct guidelines. The mass migration this autumn of so many young people from various geographic locations has led to an uptick in cases across the country. Colleges are fertile breeding grounds for disease, especially when schools are not uniformly restrictive in their public health procedures. We hypothesized that the number of colleges in each state and the total number of students attending them could be statistically significant predictors of COVID cases.

We initially included only undergraduate populations because data for total student populations, including graduate students, were not as reliable. After analyzing these data, however, we eliminated them, concluding that the neither total undergraduates per state nor average undergraduates per school were useful statistics since a college's likelihood to spread the virus is a direct function of its own particular dynamics rather than the aggregate characteristics of schools in each state. We also assessed the average tuition—not including books, fees, room and board, or other expenses—after the effects of the average institutional financial aid for colleges in each state. The rationale behind the inclusion of this variable was that schools of different costs have different demographic compositions which may lead to variations in not only school-enforced public safety policy, but also the prevalence of responsible student behavior.

The final college dataset, stored as '*college.data*', contains one row for each state, as well as columns listing the number and average cost of colleges in each state.

Politics

As outlined in the introduction, the nation has become so politically fragmented in its approach to fighting the virus that any sound analysis of the pandemic's spread must include consideration of states' varying political leanings and acknowledgment of the partisan divide over science. We decided to use the results of the 2016 presidential election as a proxy for states' general political preferences. The dataset we used contained the results of all federal elections, presidential and otherwise, since 1976. It was uploaded to Harvard University's Dataverse repository by the MIT Election and Data Science Lab. We filtered the data so that it contained results for either Donald Trump or Hillary Clinton from the 2016 election across only the 50 states. The resulting dataframe included the raw values of votes that each candidate received in each state, and the percentage of total voters that those raw values represented. From these initial variables, we engineered the difference in vote totals and percentages, assigning a winner to each state. Trump's totals were subtracted from Clinton's, and in the final form of the dataframe—stored as '*voting*'—all negative values represent states that Trump carried in 2016.

The end result was a dataframe containing each state's total vote and percent differences—as positive or negative numbers—as well as a binary indicator showing which party each state 'belonged to' in 2016.

Population Density

We next incorporated data from the World Population Review on state population and population density, two metrics essential to comparing variables across states because they allow for variables to be standardized per capita or per square mile. We specifically targeted population density because throughout the pandemic, striking disparities have been observed in the degree to which the virus ravages densely-populated cities, like New York, versus the way it impacts rural communities.

The final dataframe for this dataset, '*pop*', contained each state's population, land area in square miles, and population density.

Hospital Access

Geography plays a crucial role in determining not only the rate and severity of viral propagation, but also residents' ease of access to the hospitals, healthcare staff, and resources that could be the difference between death and recovery. To assess this geographic impact on healthcare amid a pandemic, we collected data on U.S. hospitals from the Homeland Infrastructure Foundation-Level Data repository, operated by the U.S. Department of Homeland Security.

After imputing a few negative or missing values, we engineered a number of additional variables with the help of the '*pop*' data. The final dataframe, '*hospital.totals*', contained each state's:

- number of hospitals
- number of hospital beds
- population
- number of hospitals per 100,000 state residents
- number of beds per 100,000 state residents
- total area
- number of hospitals per 1,000 square miles
- number of beds per 1,000 square miles

Adherence to Public Health Guidelines

Scientific research has concluded that proper adherence to public health guidelines, including wearing masks and limiting close contact within indoor spaces, can mitigate the spread of COVID-19. We determined that it would be useful to examine compliance with mask-wearing procedures and how it varies by state, since embracing safety precautions on a large scale has proven in other countries to be extremely effective at fighting the pandemic.

The relevant data assessed mask-wearing practices by county across each state. Collected by the *New York Times*, it measured 250,000 survey responses from July to the question, "How often do you wear a mask in public when you expect to be within six feet of another person?" Responses ranged across five levels: 'never', 'rarely', 'sometimes', 'frequently', and 'always'. The *Times'* Github page stated that raw survey responses were weighted by age and gender, and that respondents' locations were approximated from zip codes in order to extrapolate the data to county-by-county estimates.

The data contained the FIPS identification code for each county, along with the proportion of people in each county estimated to wear their mask in public at each of the five frequency levels. We faced the obstacle of identifying which FIPS numbers corresponded to which counties in which states, so we imported a dataset from the website of MDR Education which contained a key for FIPS codes from all American counties. By merging this FIPS dataset with the mask-wearing dataset, we were able to assign names and states to each county listed.

Our ultimate goal was to use these data to understand the proportion of people in each state who wear their masks at each of the five frequency levels. To do so, we imported another dataset from the U.S. Department of Agriculture consisting of population estimates for each U.S. county made after the 2010 national census. Armed with identifiable counties, their estimated 2020 populations, and their percentages of residents who adhere to public health guidelines, we calculated raw totals in each frequency level for each county, then aggregated the raw totals by state. The final dataframe,

'*mask.wearing*', contained the population and proportion, for each state, of people who wear masks 'never', 'rarely', 'sometimes', 'frequently', and 'always'.

Reopening Data

The political tension over the virus in states across the nation largely stems from disagreement over the delicate balance of two major initiatives—promoting public health and safety, and reopening an economy that continues to suffer catastrophic damage. We hypothesized that it would be useful and interesting to examine data pertaining to the reopening and public safety procedures enacted by each state, since such measures are bound to have critical impacts on a state's success in managing the virus.

Using data from the Kaiser Family Foundation and the National Academy for State Health Policy, we were able to get a sense of the policies of each state as they attempt to balance economic stimulation with public safety. The data required substantial textual cleaning to shorten variable names and entries, and the final dataset contained 26 categorical variables characterizing the nature of states' stay at home orders, face covering requirements, reopening status, and restrictions on restaurants, bars, travel, and group gatherings.

Additional Supplementary Data

Finally, we incorporated a U.S. Census dataset that categorized each state into four particular geographic regions: Northeast, Midwest, South, and West.

Merging and Cleaning the Data

Once all individual datasets had been whittled down to their most important elements, we combined them by state into one large dataframe, '*current*'. After including log transformation of 21 variables and per 100,000 resident transformations of 13 more, the '*current*' dataframe contained 50 rows, one for each state, and 82 variables. Of these variables, four were of character type, 14 were categorical, and the remaining 64 were numerical, some transformed and others standardized between 0 and 1.

The '*current*' data represented the most updated state of the pandemic across the nation on October 25, with cumulative values for metrics like deaths, cases, and people tested.

Analysis

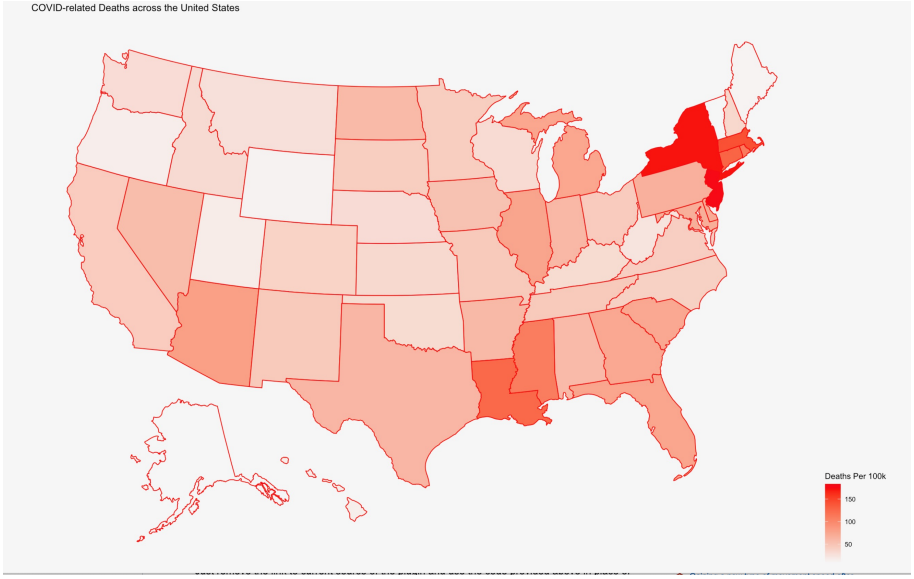
In an effort to explore the data visually and uncover latent trends between variables, we created a number of plots using variables from '*current*' as well as newly engineered variables.

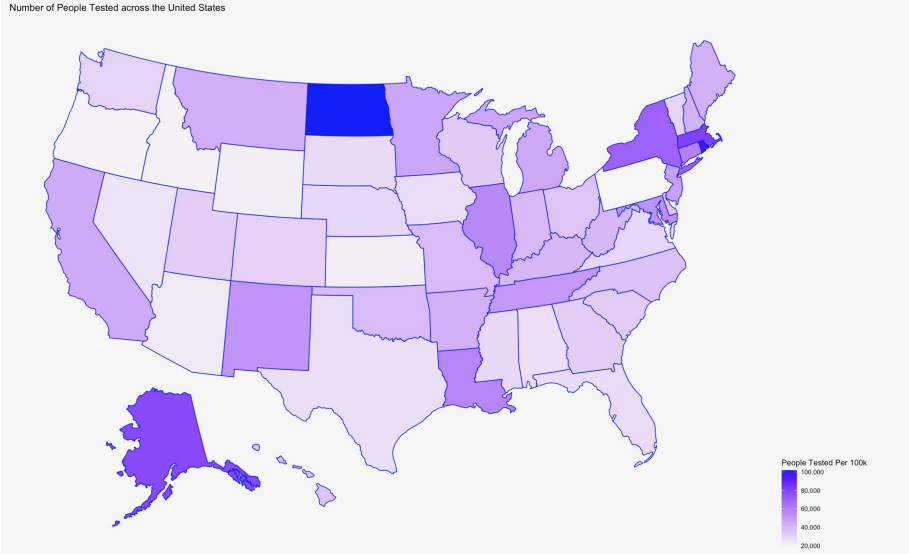
Maps

We first examined the state of the nation on a holistic level by visualizing deaths and testing totals per 100,000 civilians across all states. Darker colors represent higher totals

The map of deaths reflects spikes in New York, New Jersey, Louisiana, and parts of New England, areas that are well-known to have been hit hard by the virus. We see that more rural, Western states generally have lower numbers of deaths per 100,000.

The testing totals indicate zones of relative success in testing volume, especially North Dakota, Alaska, and Rhode Island. The former two states are somewhat surprising, given that they are majority Republican-leaning. It seems counter-intuitive for a Conservative state government and populace to embrace mass testing.



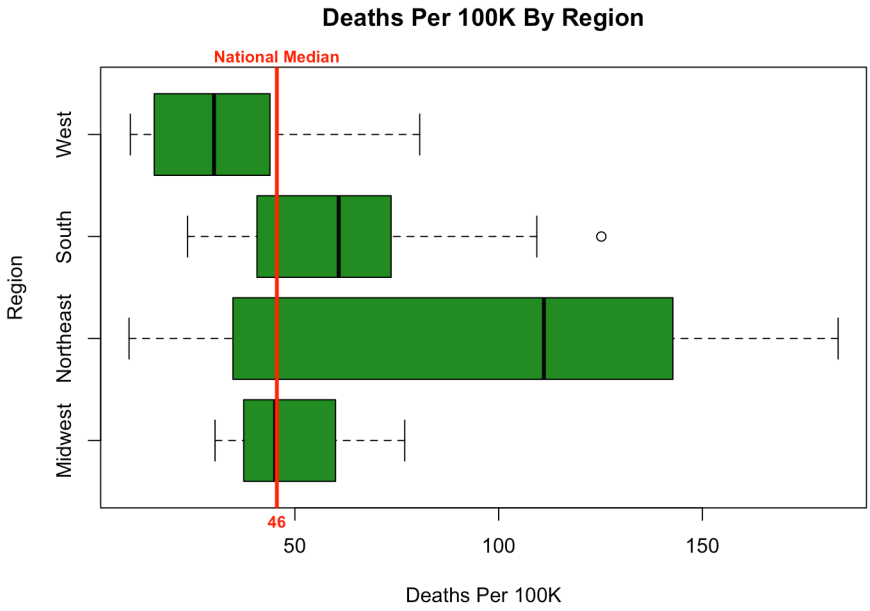


Boxplots

As a follow-up to the maps, we examined deaths and confirmed cases by geographic region using boxplots and the capabilities of base R. The boxplots represent the range, median, and IQR for all states in each of the distinct geographic regions, with the aggregated national median superimposed over the data.

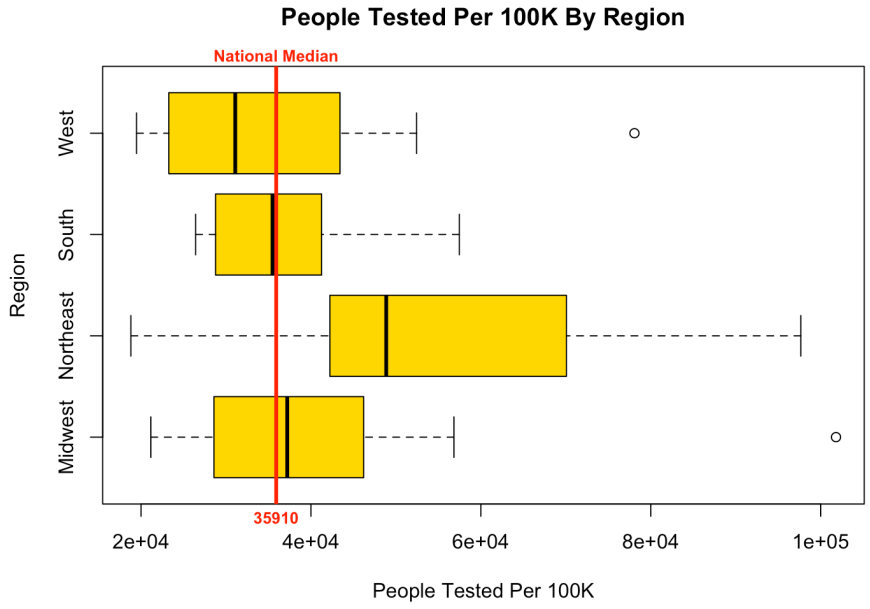
From the first plot, we observe that the West is the best region for mitigating COVID-related deaths: the 75th percentile of Western state data is lower than the national median of deaths per 100,000. We can make some guesses about why this may be the case. We know that California, a major western state, was among the first to impose strict social distancing guidelines and business closures upon its 39 million residents. These measures, undoubtedly, have played a substantial role in combating the virus and reducing deaths in that state. Relatedly, we know that the Seattle area was one of the nation's first hotspots, likely prompting tighter and longer lockdown protocols in Washington and neighboring states. Finally, we know that western states, such as Wyoming, Nevada, Montana and Alaska, are generally larger and less densely populated than states in other regions. Geography has likely played a role in the West's success.

In contrast, we observe that the death totals in the Northeast are substantially elevated over those of other regions, likely due to early and particularly severe outbreaks in New York City and Boston. The median and upper bound of the Northeast's elevated totals are biased by the conditions of these cities in spring and early summer, but it's also interesting to note that the spread of the Northeast data is by far the greatest among the regions, indicating that some states—notably



Vermont and Maine—are actually among the best in the country at reducing deaths.

In the second plot, we observe the number of tests conducted across all four regions, with the national median among all states once more superimposed. The trends are consistent with those of the previous plot: the median Northeast state is well above states in the other four regions, a trend that aligns with the acceleration of testing efforts to fight early New York and Massachusetts outbreaks.



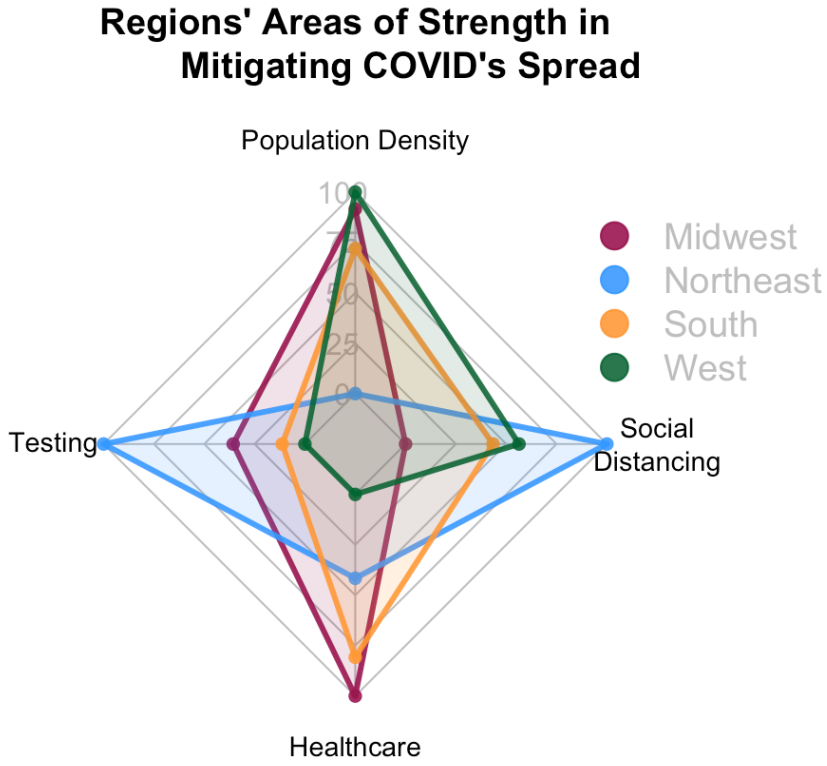
Radar Plot

Building off observations from the boxplots, we constructed the next plot in an attempt to understand regions' relative strengths and weaknesses in combatting the virus. To evaluate this comparison, we assessed the impacts of population density, testing protocols, access to healthcare, and civilians' willingness to adhere to public safety protocols across the four regions. Population density and testing per 100,000 residents already existed as variables. We created the proxy for

healthcare access by summing the values of hospitals and beds per 100,000 residents of each state, and we engineered the proxy for adherence to public safety procedures by summing the populations per 100,000 who 'frequently' and 'always' wear masks. We then standardized these values between 0 and 1 to compare the regions relative to one another.

	Population Density	Testing	Healthcare	Social\nDistancing
Midwest	0.9151737	0.3561809	1.0000000	0.0000000
Northeast	0.0000000	1.0000000	0.4158206	1.0000000
South	0.7195438	0.1126935	0.8076494	0.4345595
West	1.0000000	0.0000000	0.0000000	0.5635156

We see, for example, that of the four regions, the West has the highest score for Population Density, meaning its geographic conditions are the most conducive to limiting the spread of the virus. The Midwest and South fare quite well in comparison to the West, but the Northeast lags far behind due to its congested cities and small state sizes. The same type of relative scale is applied to the other variables, as well.

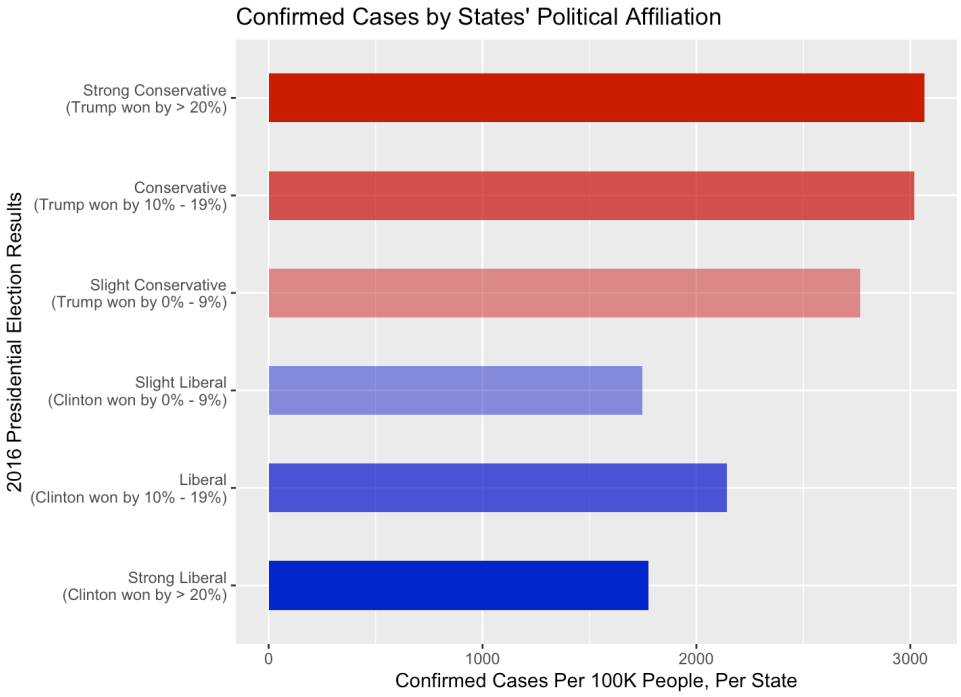


On this plot, greater distance toward the edge represents strength in a particular area, and greater total plot breadth represents well-rounded success. The plot shows that the Northeast excels relative to other regions in implementing testing procedures and adhering to public health guidelines; no other region is even half as proficient at testing. In contrast, the Northeast is hindered by high population density and mediocre access to healthcare. The striking horizontal and vertical dynamics of this chart make it easy to create a 'story': the Northeast has learned from its past outbreaks and become a leader in testing and social distancing, while the Midwest and South both benefit from advantageous population density and healthcare access, but fail to supplement these strengths with high performance in other areas.

Bar Plot

Next, we sought to understand how politics have influenced case numbers. Based on the percentages by which Clinton or Trump won each state in the 2016 election, we created six categories of political sentiment: Strong Liberal, Liberal, Slight Liberal, Slight Conservative, Conservative, and Strong

Conservative. We then aggregated states' confirmed cases per 100,000 across each group and plotted the results.



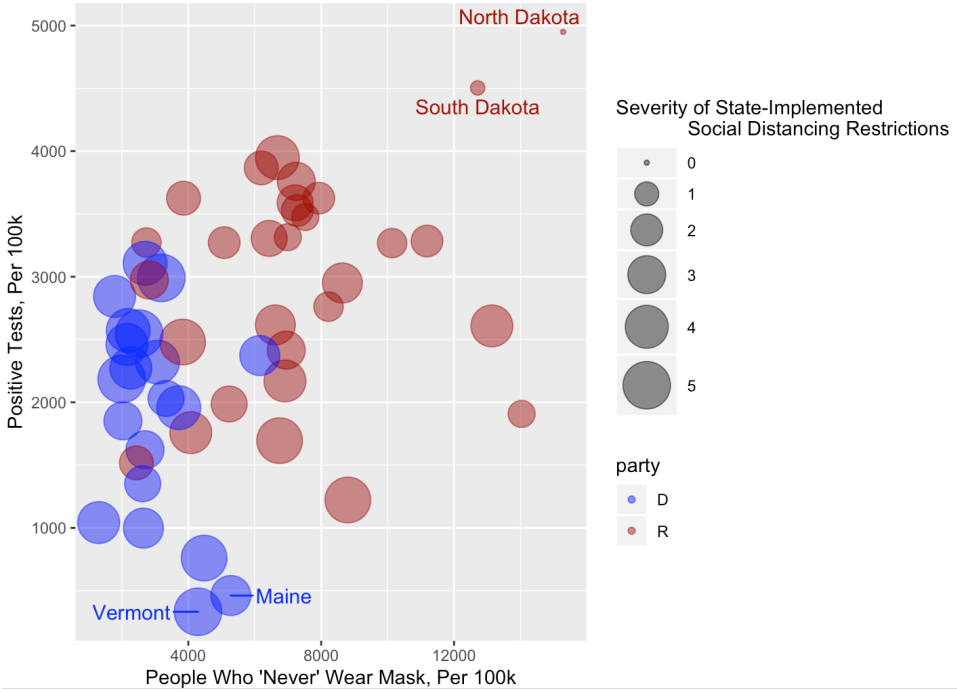
The plot is striking. We observe a clear trend: states carried by Trump have overwhelmingly more confirmed COVID cases per 100,000 people. In 'Strong Conservative' states, the case numbers are nearly double what they are in 'Strong Liberal' states. It has been demonstrated over the past few months that conservative state leaders are more likely to follow the president's urges to open the state economy and refrain from instituting restrictive business closures or social distancing measures. On the other hand, more liberal leaders have been inclined to rebuke the president and forge ahead with the policies that will best keep the public safe. This plot is reflective of the months-long divide between conservatives and liberals over the seriousness of the virus and the merits of instituting and cooperating with public safety measures.

Bubble Plot

Finally, we sought to further understand the impact of states' reopening procedures on case totals. To do so, we created a 'score' measuring the restrictiveness of each state's public safety guidelines. For seven different categorical variables assessing the status of reopening plans—including travel bans, face-mask requirements, stay at home orders, and other aspects—we converted their various factor levels to numerical scores based on the levels' severity. If a variable had five factor levels, for example, the most severe level of restriction for that factor received a score of 5, the next most severe received a 4, and so on. By summing each state's total score across these seven variables and then squaring the result, we got a sense of how restrictive the overall nature of states' public health and safety measures have been.

We then created a bubble plot that examined the relationship between the number of people who 'never' wear masks and the number of positive COVID cases, per state, per 100,000 residents. The size of the bubble is given by the state's 'score' on a scale of 1 to 10, and the bubble are colored by the party that won that state in 2016.

Once more, the plot tells a striking story. There is a clear divide between red and blue states in terms of the number of positive tests, the number of people who claim to 'never' wear a mask, and the severity of protective public health guidelines. Blue states occupy the lower left corner of the graph, indicating lower positive test numbers and more people who regularly wear masks. There is only one red state with less than 4,000 'never' mask-wearers and less than 2,000 positive tests per 100,000. In contrast, the three other quadrants of the graph are solely occupied by small-bubbled red states, representative of the partisan split over public safety and the way such a split precipitates divergent attitudes toward the virus.



Modeling Process

Armed with the insights gleaned from visualizing the relationships between some crucial variables, we began to test different types and parameters of linear models in order to identify which one predicts log deaths with the lowest error.

The first step in this exploration process was to create 'lagged' data, or time series data that had been shifted backward in time by a certain number of days. The use of lagged variables is an important step of a dynamic model like ours. The number of deaths in a given state is more dependent on past trends of variables, such as confirmed cases and people tested, than it is on same-day behavior of those variables. Using lagged values also prevents us from biasing predictions by mistakenly incorporating future data instead of exclusively using past data. A model must not predict results based on data to which it can't possibly have access.

Our dataset contained three versions of the same variables: a raw version, log transformed version, and per 100,000 civilian version. Since these variables are collinear, we had to choose one of these three types to consistently use as regressors. We settled on the log transformed variables, as they were the most normally distributed. We created versions of these log transform variables lagged at one, two, seven, and 14 days, and enfolded them within 'current'.

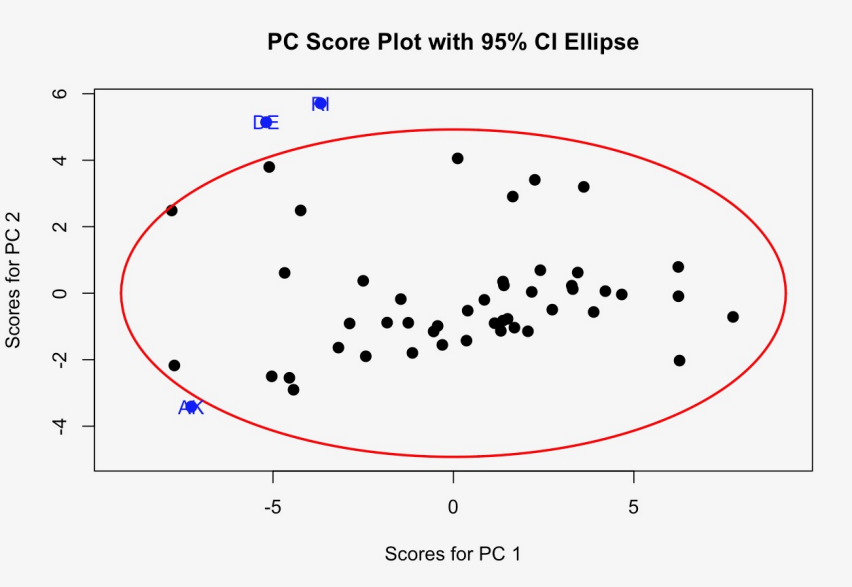
In pursuit of the most accurate model possible, we split the six months' worth of time series data into training and testing sets, the latter consisting of the fourteen days prior to October 25, the final day captured by our dataset.

We then created three functions to measure RMSE, mean absolute error, and AIC as error metrics. Comparison of subsequent models' performance across these three error types would inform our understanding of which model is most successful.

We chose the regressors for our initial model based upon the results of principal component analysis in conjunction with our own intuition about the forces that most significantly impact deaths in a given state. After checking multivariate normality using a Chi-Squared test, the first two principal components explain roughly 83 percent of the variance, as confirmed in the plot below, and help distinguish which variables, besides 'state', were most crucial to predicting deaths.

The final selection of variables included the lagged versions of all log transform variables besides deaths, as well as political party, population density, access to healthcare, geographic regions, reopening status, and number of colleges. We implemented backwards and both-ways stepwise regression, creating two versions of each model: one that used 'state' as a regressor and one that did not. Unsurprisingly, the model that included 'state' as a parameter yielded substantially lower error.

We then tested two predictive models under two disparate assumptions. The first, simpler



model assumed that as the length of the pandemic progressed, all variables other than deaths remained consistent with their values on October 25. The second iteration assumed that all time series variables continued progressing past October 25 at the unique rates at which they had developed over the prior 90 days. The inclusion of just 90 days, rather than the entire six-month span of the data, was chosen because some states had not yet experienced deaths prior to that point in the pandemic.

Results

Our final results examined two different models: one with 'state' as a predictor, and one without it.

State Model

A summary of the model is shown below.

```
Call:
lm(formula = log_deaths ~ state + lagged_log_confirmed + lagged_log_deaths +
    lagged_log_active + lagged_log_people_tested + lagged_log_negative +
    lagged_log_positive + lagged_log_totaltestresults, data = covid.obs)

Residuals:
     Min       1Q   Median       3Q      Max
-23.4221  -0.0512   0.0266   0.0918   5.7042

Coefficients:
```

As expected, state is overwhelmingly the predictor with the most explanatory power for log deaths in this model. Aligning with our expectations, we observe a positive relationship between confirmed cases and deaths. A one-unit increase in 'lagged.log.confirmed' results in a 32 percent increase in deaths–a frightening statistic indicating that a substantial percentage of confirmed cases end in death.

The model does, however, provide some hope for the national outlook. The coefficient of -0.337 for 'lagged.log.people.tested' implies that a single unit increase in testing reduces deaths by nearly 34 percent, indicating that testing has been an effective method to manage viral spread and reduce deaths across the country.

```
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.123732   0.077059 -14.583  < 2e-16 ***
stateAL                  1.320850   0.053046  24.900  < 2e-16 ***
stateAR                  0.682263   0.047847  14.259  < 2e-16 ***
stateAZ                  1.578830   0.056603  27.893  < 2e-16 ***
stateCA                  1.851708   0.058453  31.678  < 2e-16 ***
stateCO                  1.683511   0.057206  29.429  < 2e-16 ***
stateCT                  2.169190   0.061863  35.064  < 2e-16 ***
stateDE                  1.209205   0.052355  23.096  < 2e-16 ***
stateFL                  1.760081   0.057944  30.376  < 2e-16 ***
stateGA                  1.717263   0.058040  29.588  < 2e-16 ***
stateHI                  0.118007   0.042331   2.788  0.00532 **
stateIA                  1.072333   0.051814  20.696  < 2e-16 ***
stateID                  0.628475   0.048117  13.061  < 2e-16 ***
stateIL                  2.020893   0.060338  33.493  < 2e-16 ***
stateIN                  1.660980   0.058794  28.251  < 2e-16 ***
stateKS                  0.935452   0.049902  18.746  < 2e-16 ***
stateKY                  1.275830   0.050094  25.469  < 2e-16 ***
stateLA                  1.735308   0.057730  30.059  < 2e-16 ***
stateMA                  2.082806   0.063053  33.033  < 2e-16 ***
stateMD                  1.782852   0.059673  29.877  < 2e-16 ***
stateME                  0.683639   0.044484  15.368  < 2e-16 ***
stateMI                  2.089639   0.059749  34.974  < 2e-16 ***
stateMN                  1.267961   0.051651  24.549  < 2e-16 ***
stateMO                  1.443588   0.053683  26.891  < 2e-16 ***
stateMS                  1.386171   0.054562  25.406  < 2e-16 ***
stateMT                  0.286966   0.041580   6.901 5.49e-12 ***
stateNC                  1.250041   0.051467  24.288  < 2e-16 ***
stateND                  0.378606   0.049703   7.617 2.85e-14 ***
stateNE                  0.648251   0.049542  13.085  < 2e-16 ***
stateNH                  1.074672   0.050765  21.170  < 2e-16 ***
stateNJ                  2.459368   0.066911  36.756  < 2e-16 ***
stateNM                  1.217216   0.048699  24.995  < 2e-16 ***
stateNV                  1.311486   0.052262  25.094  < 2e-16 ***
stateNY                  2.663557   0.069331  38.418  < 2e-16 ***
stateOH                  1.703095   0.055698  30.578  < 2e-16 ***
stateOK                  0.980858   0.048029  20.422  < 2e-16 ***
stateOR                  0.958259   0.046394  20.655  < 2e-16 ***
statePA                  2.016464   0.062246  32.395  < 2e-16 ***
stateRI                  1.571828   0.057571  27.302  < 2e-16 ***
stateSC                  1.328764   0.052877  25.129  < 2e-16 ***
stateSD                  0.245810   0.048413   5.077 3.90e-07 ***
stateTN                  0.911792   0.048386  18.844  < 2e-16 ***
stateTX                  1.498995   0.055932  26.800  < 2e-16 ***
stateUT                  0.379326   0.045851   8.273  < 2e-16 ***
stateVA                  1.506077   0.054753  27.507  < 2e-16 ***
stateVT                  0.591498   0.044396  13.323  < 2e-16 ***
stateWA                  1.562317   0.053122  29.410  < 2e-16 ***
stateWI                  1.067740   0.049471  21.583  < 2e-16 ***
stateWV                  0.672386   0.042983  15.643  < 2e-16 ***
stateWY                 -0.171289   0.043754  -3.915 9.12e-05 ***
lagged_log_confirmed     0.321633   0.132473   2.428  0.01521 *
lagged_log_deaths        0.361513   0.013245  27.295  < 2e-16 ***
lagged_log_active       -0.126353   0.008793 -14.369  < 2e-16 ***
lagged_log_people_tested -0.337485   0.055492  -6.082 1.24e-09 ***
lagged_log_negative      0.210366   0.049552   4.245 2.20e-05 ***
lagged_log_positive      0.191369   0.132079   1.449  0.14740
lagged_log_totaltestresults  0.131320   0.062471   2.102  0.03557 *
---
```

Non-State Model

It is useful to fit a model that does not consider state as a predictor, since the conditions in individual states so dominate the other variables in terms of predictive power. The non-state model performs worse by error metrics than the state model, but its results are nonetheless interesting and informative.

A summary of predictors, listed by predictive power, is shown below. All variables are significant at the 0.05 level besides one.

We observe, sensibly, that the variable with the most influence is 'lagged.log.deaths': the previous trends in death for a given state. We also observe that precautionary measures to protect against the virus hold substantial weight in this model. The coefficient for testing frequency is, as expected, negative, indicating that more testing drives death totals down. Curiously, the coefficient for 'sometimes' wearing a mask is positive at 0.27. This is counter-intuitive, but perhaps indicates that in areas where deaths are relatively low, people don't feel that need to wear masks as frequently.

Other observations include the 11.1 percent positive uptick in deaths caused by lifted stay at home orders. We also note that the Northeast region has a positive coefficient of 0.10, a result that aligns with our visualizations. The region is know to have been hit harder than other areas

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.6159760013	1.992499e-01	-8.1102993	5.705653e-16
lagged_log_deaths	0.7228598297	9.741101e-03	74.2071995	0.000000e+00
lagged_log_people_tested	-0.3644511854	5.338466e-02	-6.8268900	9.233960e-12
log_sometimes	0.2711055244	4.488313e-02	6.0402550	1.599173e-09
lagged_log_confirmed	0.2387106555	1.611195e-02	14.8157533	4.328579e-49
lagged_log_totaltestresults	0.2210234721	4.280616e-02	5.1633571	2.477027e-07
partyR	-0.1223654164	1.443728e-02	-8.4756545	2.703153e-17
stay.at.home.orderlifted	0.1142178397	1.791479e-02	6.3756159	1.911088e-10
stay.at.home.orderhigh risk	0.1113929023	2.616809e-02	4.2568221	2.094390e-05
regionnortheast	0.1008796064	2.231737e-02	4.5202278	6.255540e-06
lagged_log_negative	0.0904634087	4.147585e-02	2.1811103	2.920083e-02
status.of.reopeningreopened	-0.0777427540	1.113230e-02	-6.9835298	3.080856e-12
log_frequently	-0.0704025414	3.213163e-02	-2.1910664	2.847232e-02
regionsouth	-0.0673776080	1.340298e-02	-5.0270603	5.075707e-07
log_rarely	-0.0635027324	2.082944e-02	-3.0487010	2.304948e-03
regionwest	-0.0539812837	1.633864e-02	-3.3039040	9.571764e-04
log_num_colleges	0.0416659154	2.199538e-02	1.8943031	5.821654e-02
status.of.reopeningpaused	0.0360040357	1.801681e-02	1.9983584	4.570759e-02
stay.at.home.orderstatewide	0.0271551846	4.375421e-02	0.6206302	5.348585e-01
lagged_log_active	-0.0260633820	6.540808e-03	-3.9847340	6.808411e-05
density	0.0002519716	3.048849e-05	8.2644851	1.600252e-16

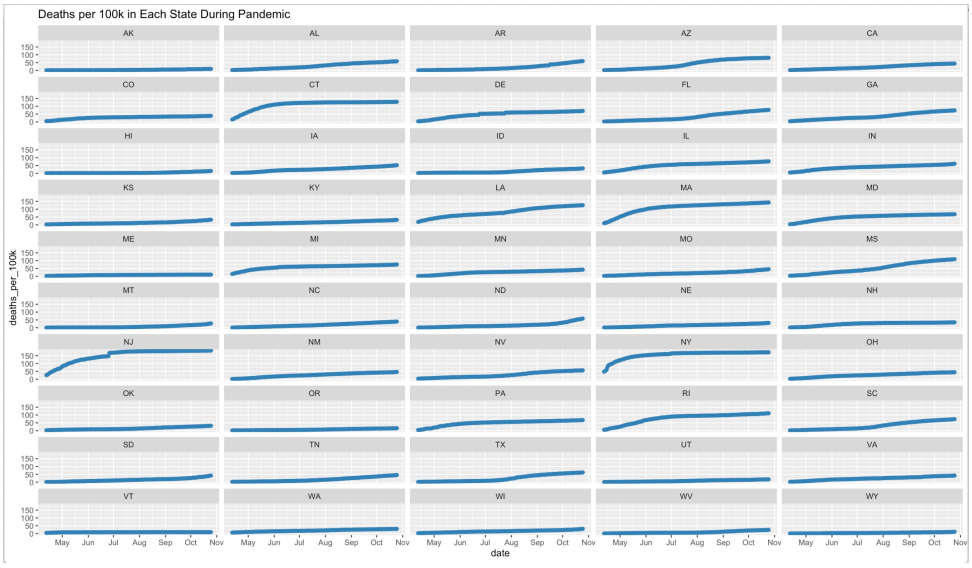
by the virus, so we expect a positive association between it and deaths–and we find one.

The greatest surprise is a negative coefficient for Republican-majority states. These states, our model indicates, are more likely to have 12.2 percent less deaths. Given the party's rhetoric and public safety inaction over the course of the pandemic, we expected a positive relationship between Republican states and increased deaths.

Conclusion

In this project, we combined a variety of data pertaining to health, politics, education, public safety measures, and healthcare access to predict COVID-related deaths across the United States in the first two weeks of November.

After careful data manipulation and model selection processes, we created two models to predict the log of deaths for the two weeks following October 25. One was heavily influenced by individual states, which–as can be seen below–have followed markedly different paths over the course of the pandemic in terms of deaths.



The state model provided us with an understanding of the virus' deadly nature and the ways it can be combated: although there is positive coefficient of 0.32 between confirmed cases and deaths, there is a negative coefficient of -0.337 between testing and deaths.

The model that didn't include states yielded a less accurate but more informative picture of the virus, indicating a few surprising relationships, including positive coefficients between deaths

and relatively infrequent mask-wearing as well as deaths and Republican-majority states.

Seeking the most accurate predictions for deaths in the upcoming weeks, we decided to move forward with the more accurate model that incorporates state as a predictor. This choice is in acknowledgement of the truly massive role individual states' combinations of conditions–from healthcare infrastructure, to governmental policy, to population density–play in determining those states' responses to the pandemic.

Appendix: Final Results

This model projected cumulative death totals across all 50 states for each day between 10/26/2020 and 11/8/2020. Consulting the Johns Hopkins official data for that time period and comparing it with the model's projections, we found that the model performed reasonably well, with an average per-state RMSE of 904 deaths. This value was high; however, we knew that the pandemic took a particularly nasty turn across the United States right at the end of October, so we expected to have underestimated death totals – somewhat severely in certain states, like Texas.