Olivia Chang
Janav Sharma
Sudarshan Gopalakrishnan

COVID-19: Analysis on Spread, Risk and Policy

Abstract

The spread of COVID-19 in the United States is not just a test of the health of local and national medical infrastructure, but also an assessment of the efficacy of the state and federal governments to respond to unprecedented crises and the underlying bipartisan foundation of the American people. This report uses trend analytics to explore the pandemic's socio-political foundation, as well as inferencing techniques to understand the vulnerability of counties and successful precautionary models other states have used.

Introduction and Exploratory Data Analysis

Understanding the Spread

Figure 1 shows a 13-week analysis of COVID-19's spread across the US for a preliminary understanding of virus trends. After the first case in Washington on 1/22/2020, pockets of the West Coast, Midwest and Northwest started seeing infections. The virus spread through the Southwest one month later, with early break-out states seeing a four-fold increase in cases. However, the next six weeks were crucial as the pandemic started to ravage. On 3/11/20 (the day the travel ban was announced), all but six states started to see cases, and the nation then saw an exponential rise of cases. The administrative policies did not heed prior warnings, only making post-affect policies that failed to address the pandemic.

Analyzing Federal and State Policies

On 3/11, key port of entries—Washington, California, New York and Massachusetts—had just under 250,000 cases, justifying a travel ban from affected nations. However, delayed policy unfolding after the travel ban enabled tremendous growth of the virus count. As seen in Figure 2, by the time White House announced Federal Guidelines on 3/16, the virus epicenters saw an average 4.6 times increase in the number of cases, and the South and Midwest regions saw a more than 10-times increase in cases.

Figure 2a: Number of Cases on Travel Ban Date

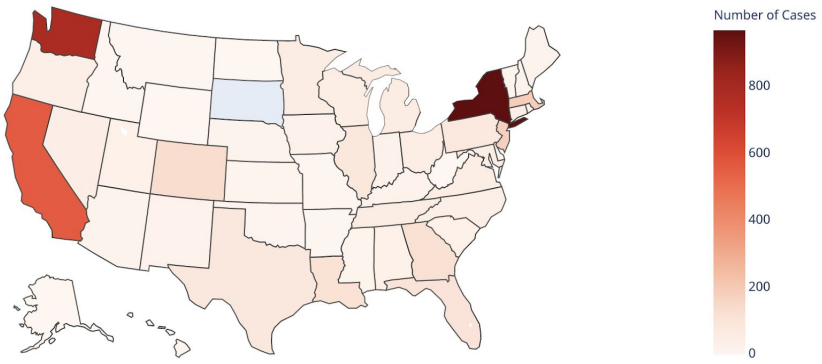
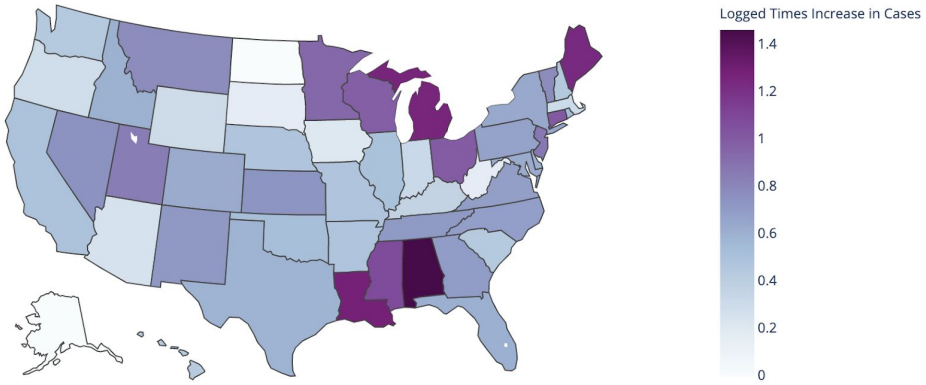


Figure 2b: Times Increase in COVID-19 Cases between Travel Ban and Federal Guidelines



Statewide policy implementations highlighted several geo-political qualms during times of crisis. We analyzed the case increase between the Federal Guideline and Public School Closures, Restaurant Dine-In, and Entertainment/Gym Closures policies. As seen in Figure 3, we notice that most major Democratic states (with the exception of New York) did not see a major increase in the compound of disease growth. Given the exponential growth nature of COVID-19 (Anderson), this indicates that they were quick to implement changes. On the contrary, major Republican centers saw a 5 to 10-times increase in the number of cases, suggesting that the federal guidelines were either not taken seriously, and/or state legislatures were late to implement policies.

Figure 3a: Times Increase in COVID-19 Cases between Federal Guidelines and Public School Closures in Democrat States

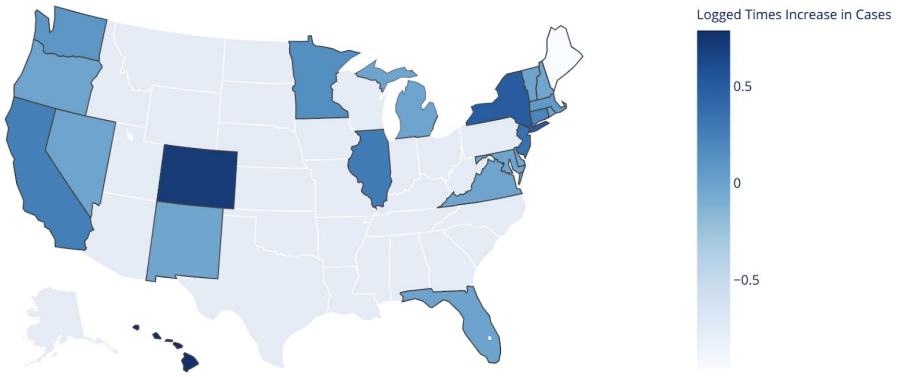


Figure 3b: Times Increase in COVID-19 Cases between Federal Guidelines and Public School Closures in Republican States

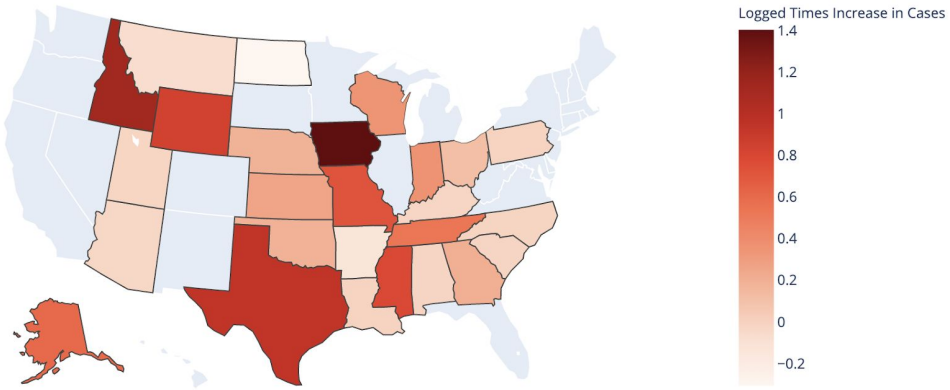


Figure 3c: Times Increase in COVID-19 Cases between Federal Guidelines and Restaurant Dine-In Policies in Democrat States

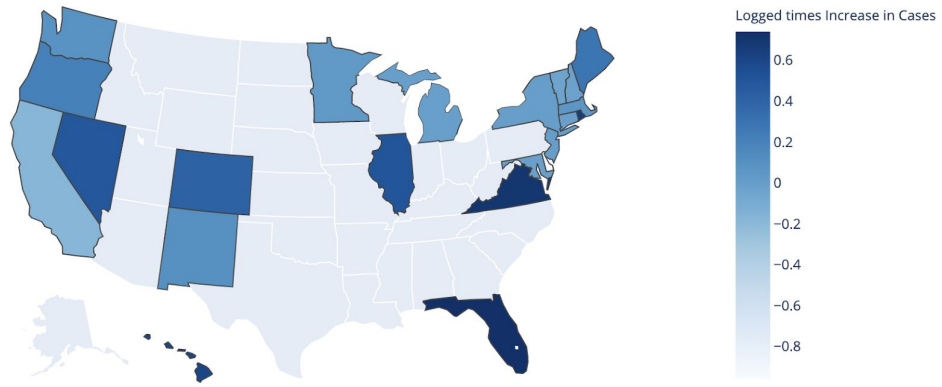
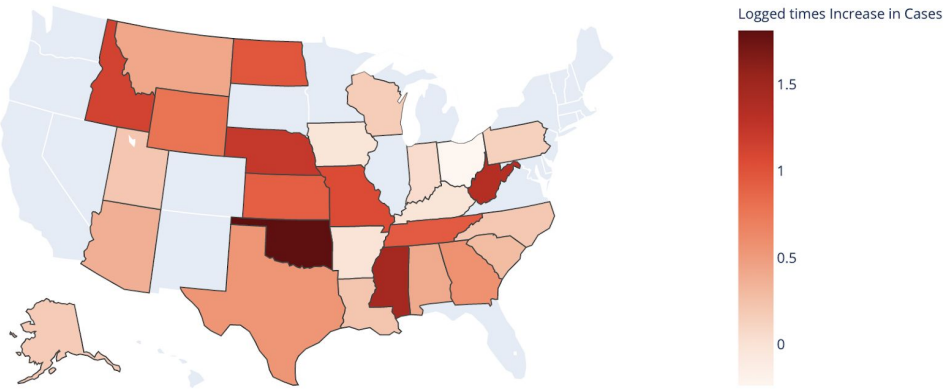


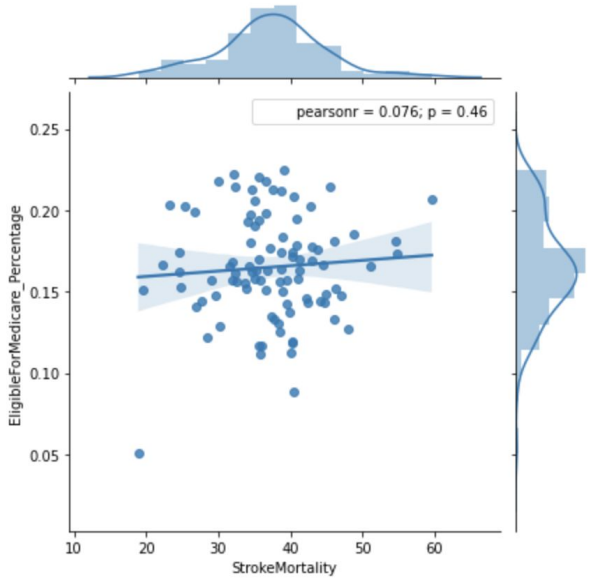
Figure 3d: Times Increase in COVID-19 Cases between Federal Guidelines and Restaurant Dine-In Policies in Republican States



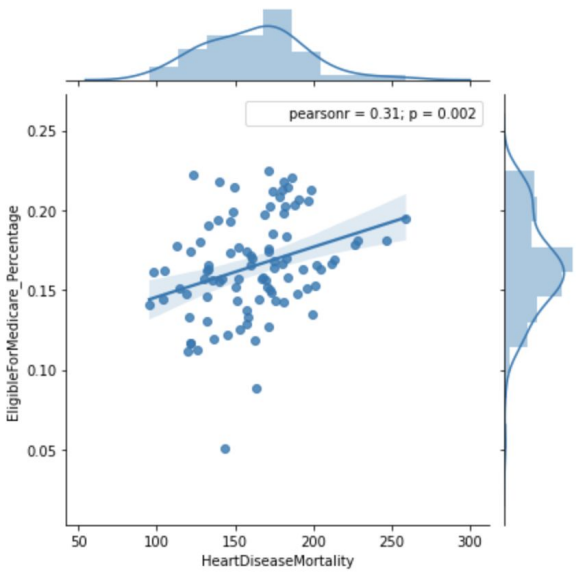
Analysis of the US Medical System

Medicare is a crucial backbone to the fight against COVID-19 for millions of Americans (“Medicare & Coronavirus”). We analyzed the demographics of US counties such as strokes, heart disease, diabetes, and smokers, each being risk populations or symptoms of COVID-19. As seen in Figure 4, we regressed between the four conditions (using aggregated data from 2015-2017) and the percentage-persons eligible for Medicare in order to understand whether the Medicare system accounts for the increased number of affected people in certain areas. The regressions produce a weak positive correlation (low r-value) between stroke, heart disease, diabetes, and smoker vs. the eligibility of Medicare. Although this regression does not give us a very accurate picture since we have not accounted for other pre-qualifications, this analysis provides a ball-park aggregate nationwide statistic.

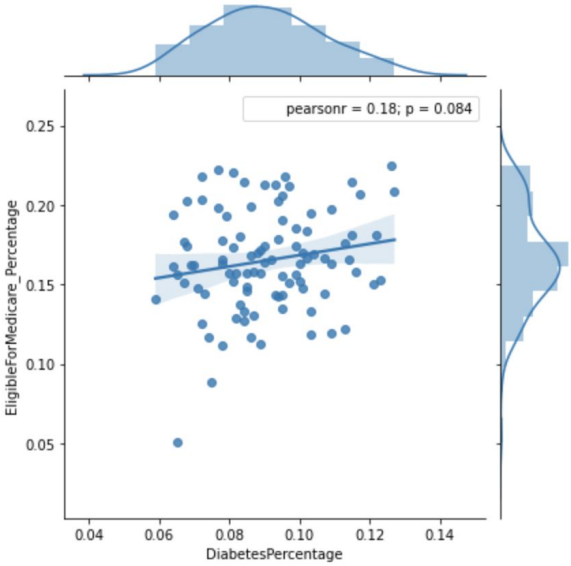
FIgure 4: Regressions between Health Conditions and Eligibility for Medicare



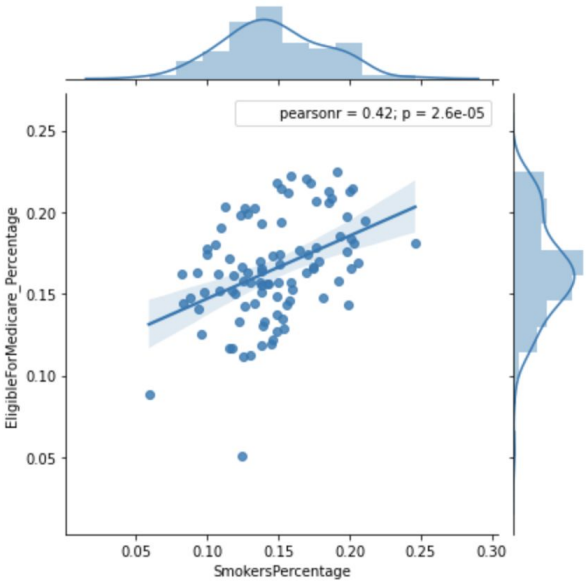
The regression produces a weak positive correlation (low r-value showing weak correlation) between stroke vs. the eligibility of Medicare. Although this regression does not give us a very accurate picture since we have not accounted for other pre-qualifications, this analysis provides a ball-park aggregate nationwide statistic.



Regression produces a weak positive correlation (low r-value showing weak correlation) between heart disease mortality vs. the eligibility of Medicare. Although this regression does not give us a very accurate picture since we have not accounted for other pre-qualifications, this analysis provides a ball-park statistic.



The regression produces a weak positive correlation (low r-value showing weak correlation) between diabetes vs. the eligibility of Medicare. Although this regression does not give us a very accurate picture since we have not accounted for other pre-qualifications, this analysis provides a ball-park aggregate nationwide statistic.



The regression produces a weak positive correlation (low r-value showing weak correlation) between smokers vs. the eligibility of Medicare. Although this regression does not give us a very accurate picture since we have not accounted for other pre-qualifications, this analysis provides a ball-park aggregate nationwide statistic.

Our Hypothesis

Given our preliminary analysis, we hypothesize that there exists a relationship between political affiliation and the growth of COVID-19 confirmed cases in the US. We want to investigate the virus's county-wide changes in incident rates, predict the US's confirmed cases curve, analyze the role of state partisanship and policy, understand the risk factor that counties may be facing, and, finally, synthesize how all of the above affects state reopening policies.

Model Methods

Data Cleaning

Before conducting our EDA discussed above, we worked on sanitizing the data sets in order to simplify the EDA Process. This included renaming variables to more intuitive identifiers, dropping dependent/repeated columns, converting date columns like stay at home order date etc. into DateTime format, converting fields from percentage format to decimal format. We also converted certain quantitative fields (e.g. dem_to_rep_ratio) to Qualitative Labels for plotting and grouping for EDA and prediction analysis. As far as dealing with missing data is concerned, if we were performing State-wide EDA or predictions, we dropped counties that had vital data missing since many of them were smaller counties that would not affect the analysis much. Our EDA motivated us to ask interesting state and county level predictive questions that required us to combine certain metrics (using grouping and calculations) to produce sensible features. For Model 1, a lot of our data transformations involved combining multiple data frames, grouping columns by State and performing further analysis. For Model 2, a lot of our data transformations included relating features from our "demographics" dataset to the county specific features that were already provided. Many of our vital missing data was calculated through the combination of multiple datasets.

Model 1: Forecasting COVID-19 Confirmed Cases

Model 1 explores several forecasting and regression techniques to estimate the growth in the number of COVID-19 cases in the United States, based on the daily increase in confirmed and death case counts. We used the widely-accepted sigmoid curve to be the foundational understanding to model the spread of COVID-19. We attempted the use of several forecasting methods, ultimately, to understand the key failure points in the modeling process:

Model	**Why did the technique fail?**
Holt Winters Forecasting Algorithm	We attempted this technique with Exponential Smoothing only to see that we were unable to model the post-peak decline towards zero-growth. Although Holt-Winters models the rise-and-trough noisy behavior that we noticed with the time-series confirmed cases data, we realize that the modeling approach forecasts trends like stocks which do not approach permanent zero-growth like disease.
Markov Processes	Due to the lack of reliable/concrete spread statistics, developing a decently accurate process was difficult, and not any better than regressing over the time series data.

Apart from the failed ones mentioned above, we attempted two models: a SIR model and a polynomial-fit model. Although the SIR model has a mathematical foundation behind its functionality, we assume that preventive public policy measures do not reduce the population susceptible to vulnerability, and we assume a constant recovery rate (ignoring advancement in medication used to help COVID-19 patients recover). While the polynomial-fit model is not driven by established models, fitting to existing trends allows it to better forecast since the implied assumptions in the actual trends carry forward.

SIR Model Design (Link For Overview: Here)

We use this model to forecast the number of confirmed cases in a county. We used the confirmed cases, deaths, and recovery rate for a given state ($\frac{\#Recovered\ on\ 4/18}{\#Confirmed\ Cases\ on\ 4/18}$) to calculate the number of active cases using this formula:

$$Total\ Active\ Cases\ on\ Day\ X\ =\ (1\ -Recovery\ Rate)\ \times Total\ Conf.\ on\ Day\ X\ -\ Total\ Dead\ on\ Day\ X$$

Based on reports, we assumed the average recovery time to be 21 days. In order to calculate the susceptible population, we calculated a proportion of the county population based on senior citizens, children under 5, and people who died of respiratory illness.

Next, we calculated the expected number of people an infected person infects in a day through OLS. We used scipy.optimize.minimize to optimize a quadratic cost function that found the optimal beta that minimized the error between the actual infected counts and the predicted infected counts (calculated using an Iterative SIR). Using the optimal-beta, we forecast infected counts and confirmed cases counts.

Polynomial-Fit Model Design

We used a Huber regressor instead of linear regressors as it uses a trained combination of L1 and L2 regularization and is more robust to the outliers and noise. The model design is as follows:

- Developed a pipeline on top of scikit-learn's package in order to construct a model that is individually fit to the disease growth pattern for each state.
- The time-series data about confirmed cases was changed such that it represents the daily case growth number (Change in Confirmed Cases Per Day).
- A scikit-learn pipeline was devised to fit a Polynomial Function with Huber loss, and each state's growth trends were individually trained and stored.
- The training process was as follows:
    - 6 different pipelines were devised, where each state's growth trends were fitted to polynomials with 1 to 6 degrees of freedom.

- Calling predict on this model would trigger this: (1) use the 6 pipelines to collect prediction; (2) calculate and return the weighted (inverse of regression error) mean for each day's prediction made by the 6 pipelines.

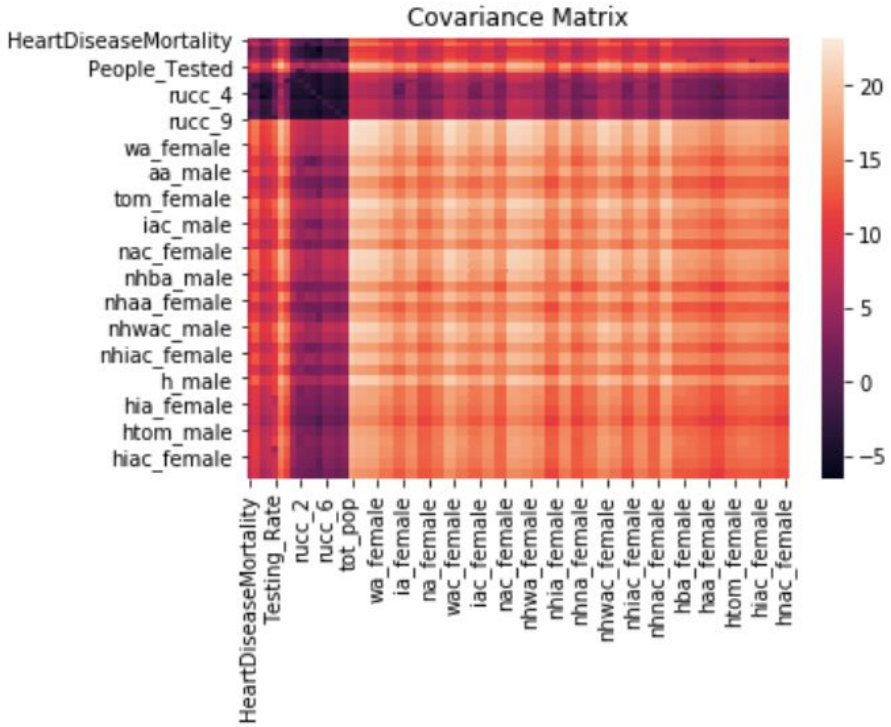
Since we do not know much about the disease spread statistics, fitting to provide growth trends was important. However, to provide flexibility with forecasting, lower degrees of freedom were also used in the pipeline. This resulted in the weighted system highlighted above.

Model 2 – Forecasting Incidence Rate and County Risk Factor

Model 2 combines data cleaning, feature engineering, and prediction methods to identify the best features and model for forecasting daily incident rates. We used all four of the provided datasets and sourced county population data from the National Bureau of Economic Research ("Census U.S. Intercensal County Population Data"). We combined all relevant features into a single "incident_features" dataframe by county FIPS codes and began data cleaning. We first one-hot encoded the "Rural-UrbanContinuumCode2013" to account for its categorical ordinal values, and removed cities, outdated counties, and Puerto Rico from our dataset to focus only on official US counties. Finally, we accounted for remaining null values by replacing them with the average feature value of counties in the same state with the same RUCC code, believing these counties could provide representative values.

We split the data into training (80%) and test sets (20%) and continuously analyzed each model's linear regression training error based on which features performed best. In the feature engineering process, we first considered our previous regression analysis on Medicare and smoker mortality rates, heart disease mortality rates, etc., alongside a covariance matrix to identify which features may work best and which had a high dependency. The covariance matrix for all features included in Model 2 is provided in Figure 5; x- and y-axis labels are unable to include all 80+ feature names. It is notable that the racial demographic features have high covariances because their data is dependent on each other. We then monitored the linear model's training error to determine how a select group of features performed.

Figure 5: Model 2 Covariance Matrix for Final Features.



One interesting feature we included was the RUCC13, something we were not previously aware of. We found that the RUCC13 data greatly increased the accuracy of our model, implying a correlation

between a county's ruralness and incident rates. We also initially only included data on races with large percentages of the US population (white, Hispanic, etc.), but found that increasingly adding minority data decreased error. A feature that we thought would be useful is a county's percentage of senior citizens ('PercentSeniorCitizen'); however, given the conclusions from our exploratory analysis and feature engineering process, we found that it showed high covariance with other features such as heart disease mortality and slightly decreased the model accuracy when included.

Our next step was testing different models (linear regression, Huber regression, cross validation on ridge regression, and cross validation on lasso regression) on our features to determine the most accurate for predicting the 4/18/20 incident rate. Using Linear or Huber Regression without regularization performed very poorly with a R2 Cross Validation best-score of -0.01 or -0.04 respectively, showing that fitting a constant value would have been superior. The RidgeCV had a best-score of 0.094 while Lasso had 0.054, showing that many of the features that we used still had some importance and could not be cancelled out.

Given that we were using a large number of features to make our predictions, and given that many of the features that we were using were indeed correlated to each other, we wanted to explore PCA as a tool to reduce dimensionality. We noticed that almost 90% of our data-set variance was explained by PC1, and hence tried to construct a Linear Regression (PCR) between PC1 and Incidence Rate. However, this resulted in an increased training error. Only after using all our PCs were we able to reduce error back to what RidgeCV offered. This allowed us to conclude that although there is correlation between some of our input features, we still need to use all of them.

As a result, we decided to move forward with the RidgeCV(cv=5) model; this makes sense, because ridge regression is useful when a dataset has high multicollinearity and is robust to outliers. We firstly trained and tested how well the model predicted only the 4/18/20 incident rate. Then, we identified the incident rate growth from 1/22/2020 to 4/18/2020 using the county features. Using this information and intuition from Model 1, we built a model that can forecast incidence rates beyond 4/18/20.

Model Validation and Results

Model 1: Forecasting COVID-19 Cases

We trained all our prospective models using the rate of change of confirmed cases daily. Our models forecast this rate for the future, and we use this data to integrate back to the predicted number of cases. In order to understand the performance of the Holt Winters Algorithm (Exponential Smoothing), we performed a train-test split and trained the model up to the estimated inflection point in the assumed sigmoid distribution of the confirmed case increase. We then tested on the rest of the available data.

In order to train our polynomial fit + L1 and L2 loss minimizer, we had to use the entire data set to maximize the amount of information available to forecast trends beyond the scope of the data available. Therefore, in this context, we incorporated, but did not incorporate, a proper train-test process. Similarly for our SIR model, we used our entire dataset to determine optimal beta.

Figure 6a illustrates the performance of the Holt Winters Exponential Smooth and the interpolator (labeled as poly predict), with respect to the original data. As suspected, we see that the Holt Winters approach fails to converge to trends that we see in the derivative of the sigmoid curve, following a pure exponential growth trend. Meanwhile, our polynomial fit was able to converge to the trends.

Our SIR model did not perform as well as intended. Unlike our test for NYC (shown in notebook), disease did not spread as much in non-Bay Area counties. The change in cases for smaller counties were rather high, causing overestimated beta values. Furthermore, we could not account for things like shelter-in-place properly, creating issues with the forecast. This would cause us to see a bad estimate of confirmed cases increase numbers overall. If we had more case stabilization in California counties, we

would have been able to make a better SIR prediction. On our test data, our MSE are as follows: 590,708 for Poly-fit; 655,398 for HW; 7,860,145 for SIR. Figure 6b illustrates our observations on how much increase in case our model projects by 7/26/20.

Figure 6a: Rate of Daily increase in Confirmed COVID-19 Cases in Texas
from 3/5 (date of first case) to 5/17 (30 days in future). Blue line is the training data for Holt Winters

Figure 6b.i: TImes Increase in Cases in Democrat States by July 26th (Poly-fit Model)

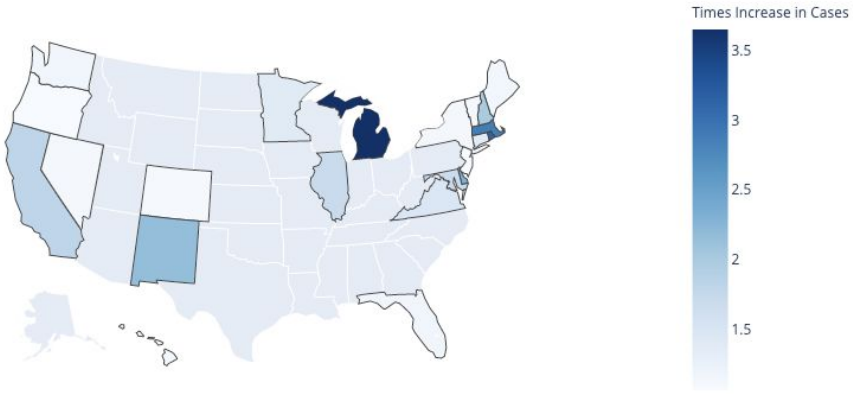
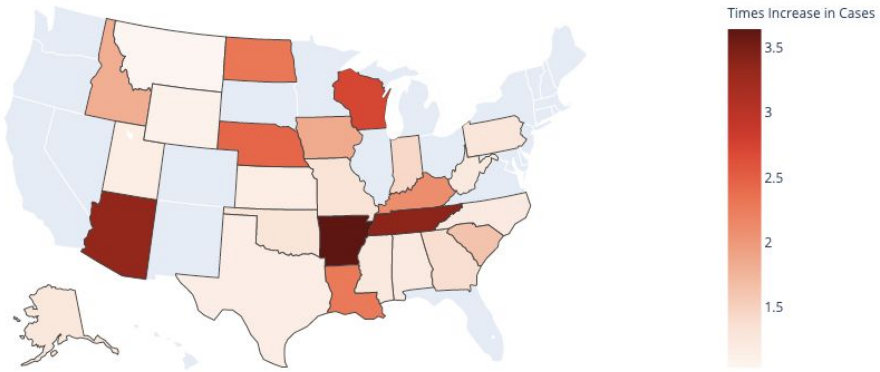


Figure 6b.ii: TImes Increase in Cases in Republican States by July 26th (Poly-fit Model)



Through the above figures, we are able to see that Republican states would be seeing a larger increase in the number of confirmed cases over the next few months. This provides us an interesting insight extending from our EDA. We originally noticed that Republican states saw larger spikes of cases after the announcement of federal guidelines, indicating a combination of delayed state measures and lack of heed for federal guidelines. The figures above could hint at continued lack of heed to social distancing measures, and hence a risk analysis of individual counties (as done in Model 2) might provide deeper insights about the same.

Challenges and Future Revisions

For the Poly-fit Model, the biggest assumption that is made with regards to our data is that we use the highest observed peak (i.e. in the case of certain states where the highest daily increase is already observed and the curve is flattening) to be that the "worst has passed." In reality, this is not necessarily true and can lead us to making faulty predictions. The key idea of having more observations will allow us to make more accurate forecasting for the future. Furthermore, having certain probabilistic spread statistics will allow us to build more powerful forecasts (combine interpolation with output from a Markovian model).

Model 2: Forecasting Incidence Rate and County Risk Factor

As mentioned in the methods, we chose RidgeCV(cv=5) as our final model. We performed predictions for 4/18/20's incidence rates and observed a final test mse of 150.5654.

As mentioned in the previous section, we then trained the RidgeCV on our selected features to predict incident rates from 1/22/2020 to 4/18/2020. Interestingly, the mean squared error significantly decreased to 48.9057 when our model was trained to predict the entire time series, as opposed to a single day. We initially inferred that the model was able analyze the relationship between all of the incident rates and therefore make better predictions. We contribute this lower error to the fact that we saw larger incidence rates only until late March. Therefore, since most of the time series was small numbers, MSE would naturally be small.

Our next step was to contextualize and understand this predicted time series data. First, we calculated the incident rate's rate of increase for every county on every day. Next, we identified the maximum rate of increase for every column and normalized this data so the values would fall between 0 and 1. We chose to identify each county's maximum rate of increase, or the greatest risk of a daily increase in confirmed cases, making the assumption that this metric would be a strong indicator of where each county is in its curve. Normalizing the data also enabled easy comparison of the maximum increase across all counties. We used the dem_to_rep_ratio column in the original states dataset to assign counties a "Republican" or "Democrat" label and created two choropleth maps showing the counties' normalized max incident rate for each political affiliation.

Figure 7: Normalized Maximum Incident Rate (0 to 1) for US Democratic Counties (1/22/20 to 4/18/20)

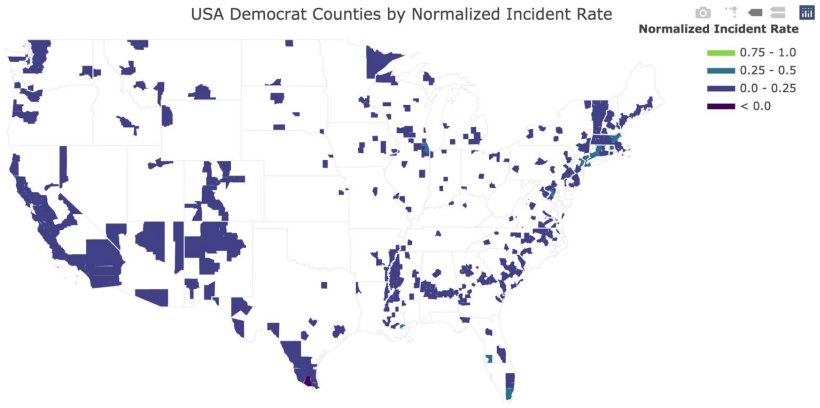
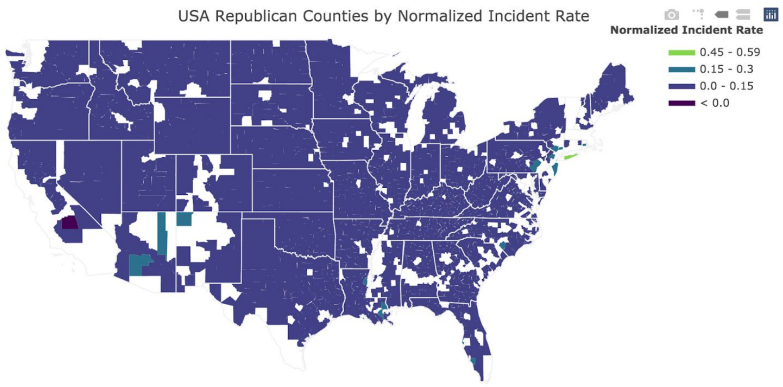


Figure 8: Normalized Maximum Incident Rate (0 to 1) for US Republican Counties (1/22/20 to 4/18/20)



Comparison of the two choropleths shows that, overall, the Democrat counties have higher incidence rates overall (Figure 7). Specifically, the normalized incident rates of Republican counties did not come close to 1 (scale of 0 to 1), instead peaking between 0.5 - 0.7 (Figure 8). However, more Democratic counties showed incident rates located at the higher end of the 0-1 scale. Furthemore, in Figure 9, we created projections of how the risk factor for counties will change over the next 3 weeks. We see that rural counties in south California, Midwest and Southwest are seeing risk decline from 0.25 to 0.

This data suggest that the more rural Republican counties are facing roughly a 30% less risk compared to the urban Democratic centers. Comparing this with our Model 1 analysis, we can see that even though certain “Republican” states might be seeing a stark increase in future cases, it might be through the Democratic-urban centers within those states. Switching to partisanship and current events, it might make sense as to why certain Republican governors are pushing for statewide reopening; this alludes to a need for more robust policies that control the disease within urban centers, but also addresses the concerns of low-risk rural counties.

Challenges and Future Revisions

One of the hardest challenges in developing Model 2 was deciding what features to include and exclude. Making these decisions put us in an ethical bind where we had to exclude data such as socio-economic status of communities or certain races. This was due to lack of information or simply because the covariance between many of these features was too high. Furthermore, the data provided was from the 2010 census because data from the 2020 census has not been released; each county’s demographic has likely changed in the past decade, and we may have not accounted for these changes because the data available is not fully up to date. As a result, we inevitably make the faulty assumption that US county demographics have not significantly changed in a decade.

More data on the socio-economic backgrounds of counties and their respective citizens might be able to increase our model’s forecasting accuracy for not only the incident rates on these counties, but also on points in time where they might face supply shortages for masks or other medical equipment.

Summary and Discussion of Results

Model 1 and Model 2 suggest interesting insights about how partisanship could have affected COVID-19’s spread across the US. Although we thought Republican states would be hit hard, we saw our results tell a different story.

Predicting and conducting analysis on incident rates over time provided insight on the relationship between a county’s demographics and political affiliation and their COVID-19 incident rates. Given how Figures 5 and 6 show COVID-19 has impacted urban areas harder than rural communities, it is clear that

local governments' efforts to slow the outbreak follow the nation's political divisions. Because larger metropolitan areas are mostly lean Democrat and the more rural areas lean Republican, these figures provide the partisan rationale for how states are responding to this crisis. In the cited Wall Street Journal article, Zitner discusses the findings we realized to further explain why certain Republican states were hesitant to announce stay-at-home orders and follow policies that were introduced in the more widely affected Democratic regions.

However, it is important to acknowledge the ethical implications of such conclusions. It is important to re-emphasize that this does not exclusively imply that Democrat policy-making was at fault; many other external variables may have caused this difference in incident rates. For example, counties in the Bay Area, Los Angeles, New York, Washington, and Chicago all have Democratic affiliations, but also are massive travel and tourist hubs, implying more travel in and out of these cities and a higher probability of spreading the disease. Furthermore, these cities' population densities are higher than the average county, making the spread of the virus a lot easier, especially since the virus is airborne. On the contrary, we noticed that a large portion of the rural, Republican America was rather at a lower risk than Democrat centers. Policies that many Democrat-aligned counties have imposed to slow the spread of the virus should not be taken lightly. If no action is taken in republic-aligned counties, it is inevitable that COVID-19 will start to have higher incident rates in these areas as well.

The conclusions taken from our analysis is not that COVID-19 is a partisan game. Adversarial interpretation of our analysis might lead to faulty implications about what next steps should be taken. The purpose of our analysis is to provide insight on our initial hypothesis, whether partisanship and demographics may have played a factor in the US spread of COVID-19; however, this is still a retrospective analysis, where the key learning for next steps is to mitigate the spread as a nation, rather than blaming it on partisanship.

Appendix

References

Anderson, Meg. "U.S. Sees Exponential Growth In Coronavirus Death Toll." *NPR*, NPR, 29 Mar. 2020, www.npr.org/sections/coronavirus-live-updates/2020/03/29/823497607/u-s-sees-exponential-growth-in-coronavirus-death-toll.

"Census U.S. Intercensal County Population Data by Age, Sex, Race, and Hispanic Origin." *The National Bureau of Economic Research*, 17 Oct. 2016, data.nber.org/data/census-intercensal-county-population-age-sex-race-hispanic.html. The CSV file used in our model contained data on county intercensal population by age, sex, and race from 2010 to 2015.

"Medicare & Coronavirus." *Medicare.org*, Medicare.org, 2020, www.medicare.gov/medicare-coronavirus.

Zitner, Aaron, et al. "How Coronavirus Is Breaking Down Along Familiar Political Lines." *The Wall Street Journal*, Dow Jones & Company, 4 Apr. 2020, www.wsj.com/articles/how-coronavirus-is-breaking-down-along-familiar-political-lines-11586001600.

Figures Referenced In Report

Figure 1a: 13-week analysis of the spread of the disease across the US (Weeks 1 - 8)

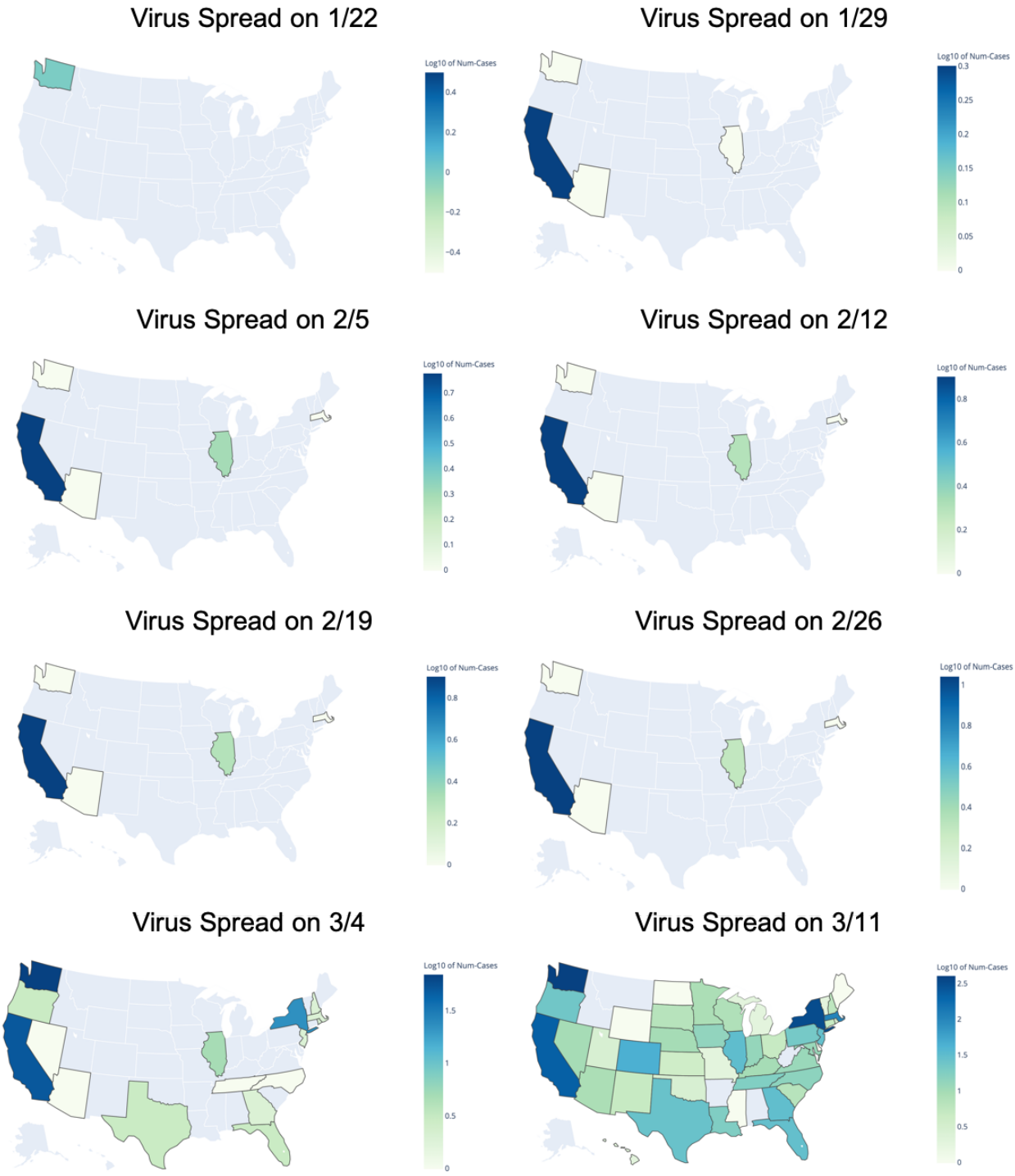


Figure 1b: 13-week analysis of the spread of the disease across the US (Weeks 9 - 11)

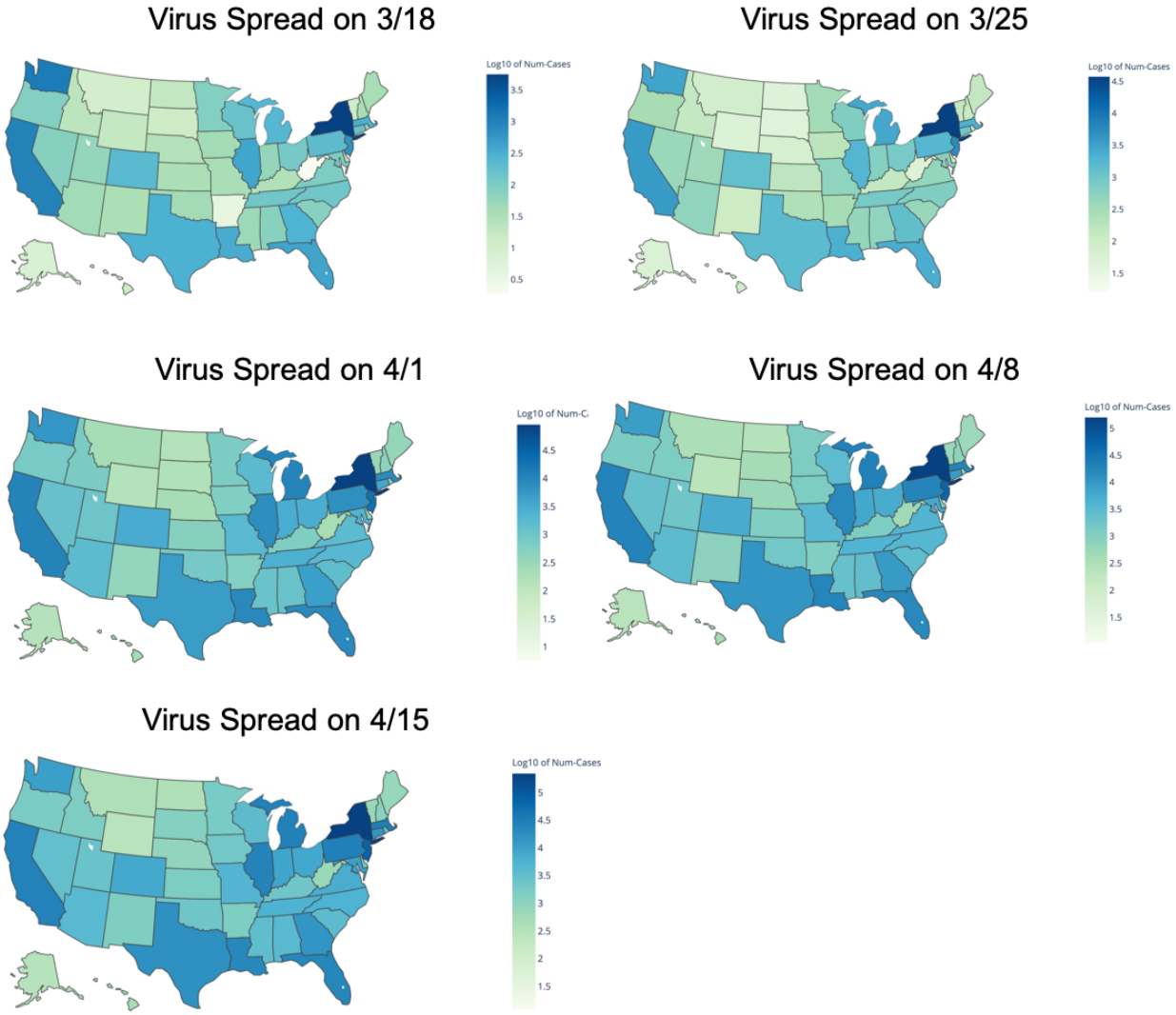
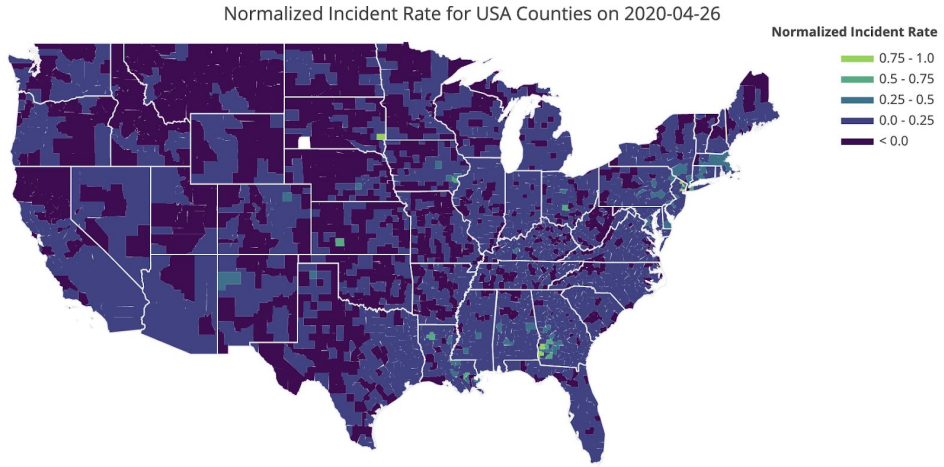
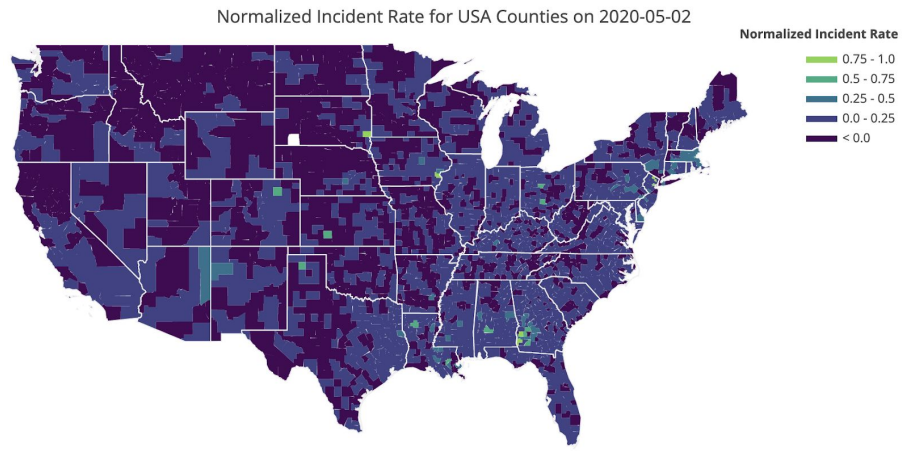
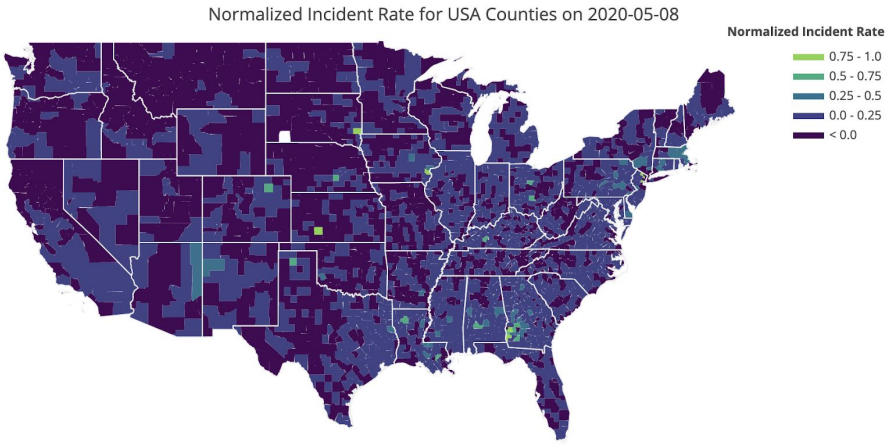


Figure 9: Projected Risk (Normalized Incidence Rates) for USA Counties

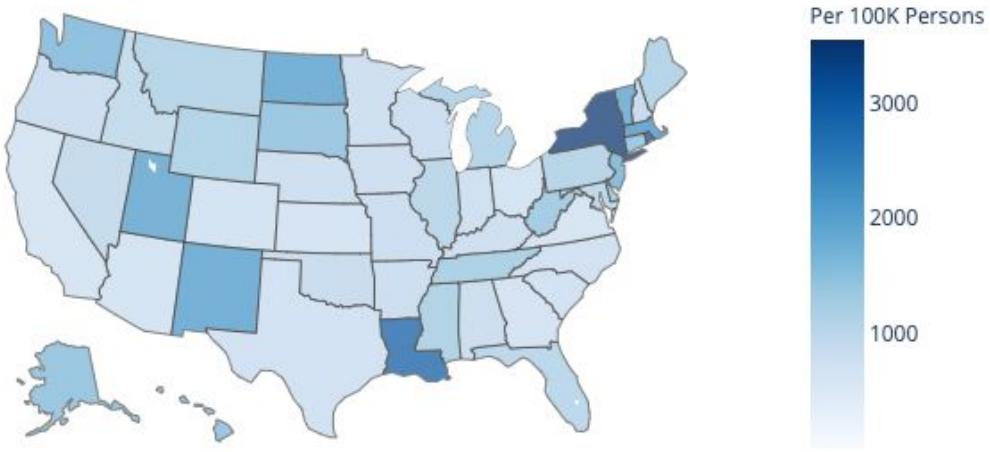



Normalized Incident Rate for USA Counties on 2020-05-02
Normalized Incident Rate
0.75 - 1.0
0.5 - 0.75
0.25 - 0.5
0.0 - 0.25
< 0.0


Normalized Incident Rate for USA Counties on 2020-05-08
Normalized Incident Rate
0.75 - 1.0
0.5 - 0.75
0.25 - 0.5
0.0 - 0.25
< 0.0

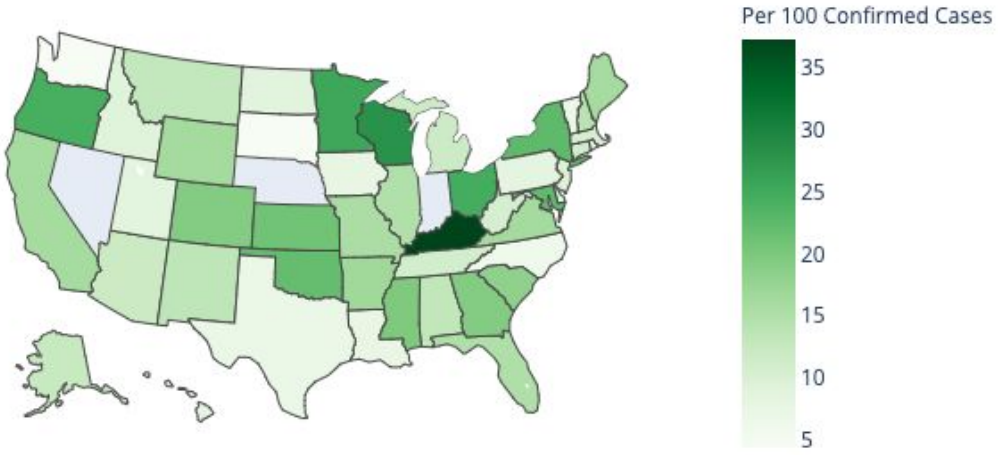
Auxiliary Analysis of Key Metrics

Figure A1: COVID-19 Testing Data on 4/18/2020



We see that there are only a very few states that are testing at very high rates. Based on prior information, we can tell that it is the first-hit states that have high testing rates. Furthermore, we see mild partisan trends on testing rates as well.

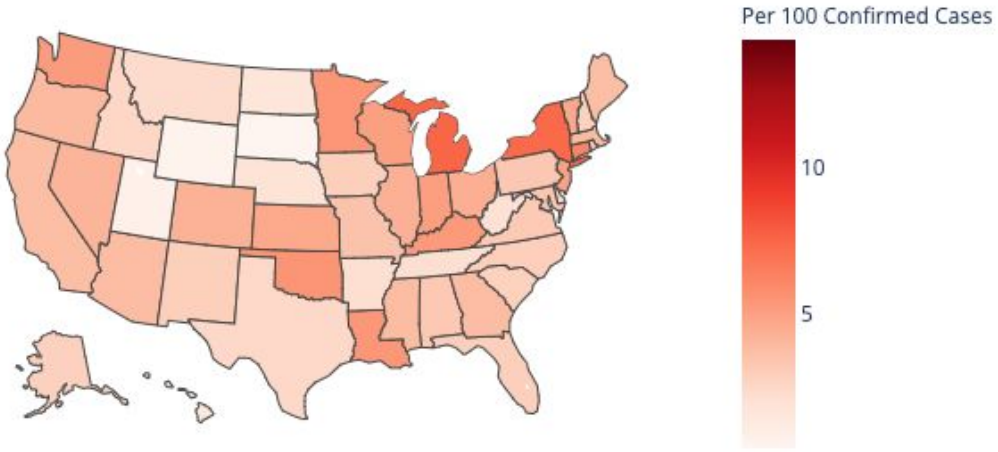
Figure A2: COVID-19 Hospitalization Data on 4/18/2020



We see some latitudinal trends with respect to hospitalization rates. Colder regions seem to have higher hospitalization rates. Furthermore, we can also see that states with higher population densities (North East) have higher hospitalization rates as well.

However beyond this information, hospitalization rates are not serving as a very useful metric since it seems a bit haphazard at the moment. We can try to correlate the insights we get from this with age, average county population density and immigration trends to understand the effect of COVID related hospitalization rates. If possible, we should also try to understand if some of the anomaly states have other regional/seasonal disease outbreaks as well.

Figure A3: COVID-19 Mortality Rate on 4/18/2020



The mortality rate in the states is indicating that some of the highly affected states include New York and its neighboring states, Washington, Oklahoma Louisiana, Great-Lakes region, and mildly in California+neighbors as well. They were the states that saw initial cases due to high foriegn travel.

Furthermore, some of these states also have high immigrant populations, and some states like Louisiana and Oklahoma are poor and underprivileged minority-dominated as well, indicating that high travel rates or income levels in combination with race might play a factor in the spread of the disease.