Analyzing the Severity of COVID-19 Across the US: A County Level Analysis

Bridget Cheng, Vatsal Bajaj, Poojan Shukla

I. Abstract

COVID-19 has caused a global pandemic and has cost the lives of thousands of Americans. Although we can easily see which counties have the most cases, deaths, and mortality rates little is known about the factors that predispose a county's population to be at greater risk for being particularly hard hit by COVID-19. In this report we ask and attempt to answer the following key question,

What variables best predict the severity of COVID-19 at the county level? We define severity in two ways, leading to two sets of closely-related analysis:
1. Number of cases
2. Deaths / Mortality Rates

In order to answer these questions, we built three inference models. The Principal Component Analysis to understand what county level features explain the most variance in the county level data. This analysis helps us make sense of the results from the Regression and Classification models. The Regression Model forecasts the number of cumulative cases in counties, and the classification model to predict whether a county lies within a 'low', 'medium' and 'high' mortality rate. Our results from out Random Forest Regressor showed that social distancing variables are extremely important in classification of mortality rates while healthcare features also demonstrate weight. Meanwhile, our Linear Regression demonstrates that parameters associated with larger county populations are key determinants of number of cases, but also that parameters that gauge the underlying health of populations and socioeconomic conditions matter as well. .

II. Introduction

The COVID-19 pandemic has highlighted many of the shortcomings of the United States' ability to respond to a highly infectious pathogen. To date, SARS-CoV2, the virus responsible for COVID-19 has claimed over 80,000 lives in the US alone. Understanding how the virus spreads and what steps can be taken to minimize the number of confirmed cases and deaths is a key priority for lawmakers. Here we describe the results of an investigation into existing data on the outbreak in the US. We analyze parameters at the county level with the goal of uncovering what attributes of a county make it more vulnerable to larger numbers of confirmed cases, as well as higher mortality rates. Our goal in this investigation is to identify "risk factors" that could predict outcomes for counties during this pandemic and those to come in the future so that resources may be deployed effectively to curtail the total of deaths. In this report we ask and attempt to answer the question,

What variables best predict the severity of COVID-19 at the county level? We define severity in two ways, leading to two sets of closely-related analysis:

1. Number of cases
2. Deaths / Mortality Rates

To answer these questions we build three sets of inference models for data as of April 18:

1. Principal Component Analysis to understand what county level features explain the most variance in the county level data. This analysis helps us make sense of the results from the Regression and Classification models.

2. Regression model to forecast the number of cumulative cases in counties:

- We use at the county-level features we engineer from the provided data to predict the number of cases as of April 18

2. Classification model to predict whether a county lies within a 'low', 'medium' and 'high' mortality rate:
- Again, we use at the county-level features we engineered predict which bucket of 'low', 'medium' and 'high' mortality rate does a county fall in.

The implication of results are as follows:
States and the federal governments can use the models above to obtain:
1. The predicted number of cumulative cases at the county at a point in time:
- This is especially useful for counties which have missing or unreliable information for cumulative cases and accordingly plan and prepare for the impact covid-19 .
- Understand the features that contribute to higher or lower cumulative cases, allowing governments to drive policies that can perhaps lower the cumulative cases.

2. Which counties fall within which bucket i.e. high, medium or low mortality:
- Understand which counties are at a high risk of adverse impact due to covid-19. This is especially useful for counties which have missing or unreliable information around death rates and cumulative cases
- This information will allow state and federal governments to understand what features may lead to a reduction in mortality

rates and design policies to achieve the same.

The model makes predictions based on county level features that governments readily have available

III. Description of Data

As part of this project, we had access to several .csv files. We used the files provided to us by the course so our data goes up to 4/18/2020.

1. "4.18states.csv": This file contains information on the 50 states as well as other US territories and other nations on parameters such as total number of confirmed cases, deaths, recovered, number of people tested, number of people hospitalized, hospitalization rate, mortality rate, and incidence rate. There is also additional data on latitude and longitude of states as well as information on when this data was last updated. We did not make extensive use of this data. As it did not have the granularity we are looking for in a county level analysis and does not contain columns that could be used as features to predict disease burden.

2. "abridged_couties.csv": This dataset lists all of the counties in the US and its territories. Key identifying parameters are included such as "FIPS" codes which are unique to each county and geographic information given by indicating the latitude and longitude coordinates. This dataset also contains lots of information about counties themselves. Such information includes demographic variables such as population estimates and age structures, mortality rates by age group, factors associated with socioeconomics and healthcare access disparities (all HPSA fields). The dataset also contains mortality attributable to specific diseases such as diabetes for instance. This dataset contained many missing values, for which the we had to impute values as described elsewhere in the report.

3. "time_series_covid19_cases_US.csv": This dataset provides key identifying information about each county such as FIPS codes, geographic location, and also list the cumulative number of confirmed cases of COVID-19 in the county every single day starting on 1/22/2020 to 4/18/2020.

4. "time_series_covid19_deaths_US.csv": This dataset provides key identifying information about each county such as FIPS codes, geographic location, and also list the cumulative number of deaths due to COVID-19 in the county every single day starting on 1/22/2020 to 4/18/2020.

IV. Description of Methods

A. Basic Data Cleaning: We used the following datasets: time_series_cases, time_series_death and counties. Since the counties dataset contains specific features about every county, we focused on cleaning this dataset by looking at Social Distancing Parameters, Health Outcomes, Health Resource Availability and Demographic variables for each county.

First, we noticed that in the "abridged_counties.csv" dataset, there were no values under "States" corresponding to "Alaska" or "Hawaii". These states were entered as "AK" and "HI" in the data under the "StateName" column, respectively. We changed all county and state names to lower-case to ease further analysis. Since there were no null values under the "StateName" column, we were able to use the two-letter state abbreviations to replace the NaNs with the full state name under "States".

B. Data Cleaning for Demographic Features: The demographic variables included population estimates for different age groups, party preferences, medical enrollment etc. The demographic estimates for each county were considerably clean and did not have many NaN values. "dem_to_rep_ratio" was a feature with multiple NaN values. We inferred that these values likely signified missing data, so we computed a weighted average based on the dem-to-rep ratios of all the counties in a state and took into account their populations to compute a dem_to_rep ratio to impute. All other missing demographic features were imputed in the same way. These features included "MedicareEnrollment,AgedTot2017", "EligibleforMedicare2018", "SVIPercentile".

C. Data Cleaning for Social Distancing and Mobility Features: For each column, the dates in the social distancing parameters represent when a certain policy was put into action. For example, for the stay at home column, the dates show when stay at home was implemented.We discovered that all of the dates in the columns for fields such as "federal travel ban", "federal guidelines", and other markers of government imposed social distancing measures were entered as ordinal dates. We used feature engineering to instead express these columns as the number of days from 4/18/2020 that certain measures had been in place. If measures were not enacted, they would receive a value of 0 (which is how we imputed NaN values). Some measures such as federal guidelines and foreign travel ban had the exact same date for every single county. Since these measures were not differentiated between counties, they play no role in expressing county-level differences in policy. Thus, we did not account for these measures.

D. Data Cleaning for Health Outcome Parameters: Health outcome parameters included mortality rates for different age groups, 3 year Diabetes, Stroke mortality and other similar features. 'DiabetesPercentage', 'HeartDiseaseMortality', 'StrokeMortality', 'Smokers_Percentage', and 'RespMortalityRate2014' all had NaN values which were imputed using weighted averages for the parameters within states.

Our decision for how to treat NaN values in the columns corresponding to age specific mortality were informed by data visualization of the number of these missing values over their respective age groups (Fig. 5). The above bar plot depicts the count of NaN values in each of the age group mortality columns. This simple visualization was made to examine whether there is any significance to the number of NaNs in the age group mortalities. After plotting the number of NaNs, it can be seen that the distribution of the number of NaNs resembles the tail of an exponential distribution.

The hue of the bar plot is the average mortality of each of the age groups, excluding the counties with NaN values. A darker red means that the age group has a higher mortality rate, and lighter shades represent a lower mortality rate. From looking at the plot, the age groups from 1 to 4 has and 5 to 14 have the lightest hue and thus the lowest mortality rate, which reflects the truth of age-based mortality rates. As age increases, older age groups are more robust in face of accidents, diseases,

and other sources of mortality. However as age continues to increase older age groups starting from age 45 again become more susceptible to various sources of mortality.

Since the groups with the lowest mortality rates also have the highest number of NaN values, it is reasonable to suggest that the NaNs were meant to represent an infinitesimal or zero mortality in some counties for certain age groups. As mentioned above in the data cleaning section, the fact that none of the entries had a mortality rate of zero also supports our assumption.

For the "mortality2015-2017Estimated" we noticed that all values in this column were computed using values from all of the other age specific mortality rates. However, we were not given and could not determine the manner in which this estimate was computed. This in conjunction with the fact that well over 3000 of roughly 3200 rows of the "abridged_couties.csv" dataset were missing a value for this feature motivated our decision to drop this column from further analysis.

E. Data Cleaning for Health Resource Availability Features: Another set of features that had several NaN values were "HPSAShortage", "HPSAUnderservedPop", and "HPSAServedPop". HPSA refers to health professional shortage area, and exactly 1078 counties were missing values for these features. We can reasonably infer that not all counties are going to be designated as shortage areas so we assumed that these 1078 do not have shortages. For that reason, all missing values for "HPSAShortage" and "HPSAUnderservedPop" were filled with 0's because in the absence of a shortage, 0 additional medical professionals are needed to alleviate the shortage and there are (theoretically) 0 individuals in the underserved population. Following this reasoning, we also decided to impute all "HPSAServedPop" NaN values with the population from that county.

F. Data Visualization:

Geoplot of Date of first cases in each state

We created a geoplot (fig. 6) from computing the date of the earliest case in each of the 50 states from the time series "time_series_covid19_confirmed_US.csv." We first compute the date of the first case on the county level, and later simply group by state and take the minimum of the dates. As seen in the graph, on the color spectrum from red to yellow, a brighter red indicates an earlier confirmed first case of Covid-19. According to the time series for confirmed cases, the first confirmed case in the US is from Washington on January 22nd. Thus, for Washington, along with other states with earlier first cases such as Illinois and California there is a bright red color in the region.

The motivation behind the creation of this geoplot is to examine the possibility of using the dates of the first cases as a feature in our model. This feature, which was engineered from the "time_series_covid19_confirmed_US.csv" dataset, seems like a useful feature for our model. Although it is also possible that the inclusion of this feature will introduce some redundancy into our feature set. This geoplot of first confirmed cases is also meant to be analyzed together with the geoplot below of mortality rates.

County Level Geoplot for Mortality Rate

Figure 7 demonstrates a geospatial visualization for county level mortality rate where a lighter shade corresponds to a higher mortality rate. Since we believe that our mortality rate is meaningless if a county has less than 30 cases, we removed all the counties with less than 30 cases. This is because we are concerned that in counties with fewer than 30 confirmed cases, a few deaths or lack thereof could skew the mortality rate measure. This plot relates to our second question and helps us identify which counties have been able to deal better with the spread of COVID-19 in order to control the number of deaths due to the pandemic. Counties which have a high mortality rate saw more number of deaths relative to their confirmed cases than counties with a darker shade or lower mortality rate. This makes us wonder what factors about these counties lead to this difference in mortality rates.

The Geospatial plot for first cases demonstrates which states were first hit with coronavirus

Line Plots to Visualize COVID Cases and Deaths

The four sets of two subplots (figs 8-11) split up the 50 US states into four groups: the West, the Midwest, the South, and the Northeast. Each pair of the two subplots graphs the number of confirmed cases and deaths from the time series datasets for each of the four groups.

We did the splitting by region in order to observe any underlying relationship between the region that a state belongs to and its number of confirmed cases and mortality rate. Another motivation was just so the visualizations are more clear and comprehensible without having 50 cluttered lines in one line plot. Without the use of multiple separate plots, it was also hard to generate a good legend. The legend would be too long, and the need for 50 distinct colors makes some lines indistinguishable from the others.

However, a downside to this splitting is that the four groups of states are not graphed on the same scale anymore. In the group containing states of Northeast US, including New York, the number of deaths on the y-axis ranges from 0 to around 17500, whereas for states in the South US the y-axis only goes up to 1200 for the number of deaths. This makes it slightly more difficult to compare across the four groups/sets of subplots.

Feature Correlation Through Heatmaps

To inform our feature selection process, we created a multitude of heatmaps to assess the correlation within certain groups of features (such as social distancing, health outcomes, etc.) and then assessed correlations between multiple groups of features. The output of this process is displayed in Figure 1. Note that we had to remove "federal guidelines" and "foreign travel ban" from our analysis because all counties had the same value for this columns, resulting in a column full of NaN values when we attempted to standardize the column for the purposes of correlation. Many of the social distancing features have modest correlation with each other. 'Stay at home' has some of the highest correlation values with other features, which would indicate that when municipalities and other local governments announce orders to stay at home, they are also rolling out regulations against sizes of gatherings. 'Public schools' has low correlation values indicating local

governments' decision to close schools on a different timeline from other orders.

The plot of all the health outcome variables (fig. 3) reveals a high level of correlation between many of the age specific mortality features. However, the mortality rates due to specific diseases have low little to no correlation with age specific mortality but have a moderate level of correlation with each other. This could be because many of these diseases, including stroke, heart disease, and respiratory mortality could be associated with larger older populations.

As we would expect with the health resources features, parameters such as number of physicians, number of hospitals, and number of ICU beds are all highly correlated (fig. 4). Interestingly, HPSA served population also correlates highly with these features. This could be that as a result of our imputation since we replaced HPSAServed Population with county populations, the most populous counties may not be HPSA areas and so HPSAServed Population becomes a proxy for a measure of county population as are number of healthcare resources since we would expect there to be more doctors and hospitals where there are more people.

Many of the demographic features pertaining to age structure correlate very highly with each other (fig. 2). Rural-urban continuum codes correlate negatively with population measures since high values for these codes indicate more rural areas. SVI percentile seems to be a feature that doesn't correlate well with any of the other parameters and could be a very useful feature. We did not include a heatmap of all parameters because we felt that this would result in overplotting and would not be useful in determining what features to select for our model. Instead we turned to PCA to glean more insights as to the dimensionality of our data.

V. Summary of Results

A. Principal Component Analysis

PCA is a dimensionality reduction technique commonly used in EDA applications as well as classification problems. Here we performed PCA to understand which feature columns of our counties dataframe were most contributing to the variation we saw in county parameters. We performed PCA on all counties with a non-zero number of cases. The first two principal components explained about 80% of the variance in our data. Judging by the graph generated from the PC1 scores (fig. 13), and the actual PC1 vector itself (fig. 15), it appears that population centers have very high PC1 scores. This is corroborated by the fact that all of the parameters that contribute the most to PC1 scores are all proxies in one way or another for the size of a population. For instance, features such as age specific mortality rates or the number of males/females within a certain age group are going to be greatest in the counties with the most number of people. On the other hand, features such as rural-urban continuum codes have higher values for rural, less densely populated counties and so this feature has a negative value for the PC1 vector. For the second principal component, we can see that the features that are most emphasized are certain disease specific mortality rates, such as from heart disease, stroke, and respiratory illnesses as well as SVI percentile (fig. 16). On the map (fig. 14), we can see that the counties with the highest PC2 scores tend to be in the south and south eastern US. Social Vulnerability Index Percentile (SVI) is a measure of social vulnerability according to the CDC. This measure could indicate that many of the counties in the south are particularly vulnerable due to socioeconomic barriers to obtaining healthcare since several states with the lowest GDP per capita are in the south. Overall, the PCA tells us that roughly 80% of the variation we see between counties can be explained by parameters related to population sizes, and features related to social vulnerability as well as mortality rates from specific diseases. This provides us the insight that in addition to demographic variables tied to population size, features relating to the underlying medical problems afflicting populations of the county could predispose residents to a more severe COVID-19 outbreak.

B. Regression Model

After three trials of feature selection, we have selected a set of best performing features of size 38 as our final feature set. Out of the 38 final features, 22 of the 38 are features that are related to the overall population or the size of a demographic group. It turns out that the 22 population-related features are also the features that were assigned the largest weights in terms of magnitude.

'CensusPopulation2010' is the feature with the heaviest weight of around -261352. This feature yields the population of counties from the 2010 census, and it is the feature that most directly describes the size of a county. It is not surprising that this feature has the heaviest weight as assigned by the model, since the response variable we're trying to predict is the number of confirmed cases in each county. Due to the fact that we are not predicting a proportional variable such as mortality rate, the population of a county should consequently have a large impact on our response variable. The features relating to the population of demographic groups also have larger weights in general. The large weights of these features indicate that the age groups sizes have large impacts on the number of confirmed cases. This makes sense in the context of Covid-19, since the older age groups have weaker immune systems in general and are more susceptible to the coronavirus.

It is also worth noting that the feature with the highest weight that's unrelated to population is 'dem_to_rep_ratio' with a weight of around -413. Although this weight is much smaller than the weights of the features mentioned above, it is still higher than the weights of 'MedianAge2010', 'SVIPercentile', 'stay at home', and 'HPSAUnderservedPop', which seem like more helpful features to use. Although a large generalization, this high weight could be due to the different beliefs and behaviors Democrats and Republicans have in the face of Covid-19.

With our final model, we're able to achieve a test root-mean-squared-error (RMSE) of around 998.39, which is quite close to our training RMSE of 950.61. This is expected as we tested out different models and examined their cross validation scores in multiple trials, to ensure that we do not over-fit to the training data. The test R-squared value is not as high as we'd expect from a training R-squared of 0.92, but at the least it is still a positive number with a value of around 0.46, meaning

that it fits our data well to some extent.

C. Classification Model

For counties with at least 30 cumulative cases as of April 18, we build a classification model to predict which counties fall within, high, medium and low range of mortality rates. This model allows us to predict, based on county level features, whether a county belongs in the high, medium or low risk category of mortality rates.

Mortality rates at the county level for counties with at least 30 cases range between 0 to 20%, with the average being around 4%. We decide the cut-offs for low, medium and high based on epidemiological studies we saw and the distribution of mortality rates in our data (Figure 22).

States and the federal governments can use this model to predict which counties fall within which bucket i.e. high, medium or low mortality rates among counties for which they don't have mortality rate estimates (say due to lack of information or because the epidemic hasn't yet started there i.e. less than 30 cases).

States and governments can also use the feature importance charts to understand which features contribute to a county being placed in a high vs. low mortality rate bucket, for example. This information will allow state and federal governments to understand what features may lead to a reduction in mortality rates and design policies to achieve the same.

First, we attempted a baseline logistic regression model with regularization and parameter tuning to achieve a model with an accuracy of 48%. This model does not seem to be predicting anything in the class, high: which is strange so we can't really trust this model. This baseline model has no regularization involved and no hyper parameters have been tuned. The model's accuracy is 51%, which is not very bad for baseline since we have 3 classes, where a coin-flip model would have a 33% accuracy. There seems to some imbalance in the classes so we look at Area Under the Curve (AUC), a metric that is agnostic to class-imbalance: An AUC of 59% for the baseline model is not bad at all.

We then tune the Logistic Regression Model using a GridSearchCV to reach our best model. Contrary to the baseline, the best model seems to be predicting all 3 classes: high, medium and low. More specifically, this best model had some regularization (c = 0.23) and worked with the l1 or lasso penalty which drops features that don't add any value to the model. The model's accuracy is 52%, which is a little better, while the AUC is still around the 58% - 59% mark.

In figure 19-21 we look at the feature importance charts for the best performing Logistic Regression Model. Here each chart corresponds to features important for one class. Here we can see that the longer it has been since people have been staying at home and restaurant-dine has ceased, the more likely the county is to be placed in the low mortality rate category; while fraction of male contributes strongly and positively to the high mortality rate category agreeing with analysis that says that men are more susceptible to covid-19 than women.

Following, this we tried out a Random Forest Classification Model which performed must better with an accuracy of 51% and an AUC of 61%. Therefore, the Random Forest Classifier had the best results.

In order to observe feature importance, we created a SHAP summary plot as shown in Figure 17. SHAP values allow us to learn feature importance and the directionality of importance in black-box models such as tree-based models by accounting for the marginal contribution of each feature to the model. We noticed the following trends:

1. The longer 'staying at home' policy has been in effect, the higher the chance that the county gets classified in the 'low or medium category' for mortality rates.
2. The higher the median age, the more likely it is that the county gets characterized in the 'medium' class.
3. Heart-disease mortality seems to be contributing more to the 'medium' and 'high' category than the 'low' category.
4. Stroke mortality seems to be equally important across all classes, which is a confusing result.
5. Diabetes % seems to be most important for the 'high' class which makes sense since people with underlying conditions are at a higher risk of death from coronavirus.

Following this we created a SHAP plot, Figure 18, which showed very interesting results:
1. Social distancing variables are extremely important:
- The longer 'staying at home' policy has been in effect, the higher the chance that the county gets classified towards the 'low' category.
- Similarly, the longer 'restaurant-dine-in' and 'entertainment / gym' has been shut, the higher the chance that the county gets classified towards the 'low' category.
2. Health-care features are important:
- 'Stroke mortality' has a very mixed effect, which is confusing with high rates corresponding to being categorized both towards the 'high' and 'low' categories.
- 'Heart disease mortality' similarly seems to be having the reverse effect, which does not make sense

3. Demographic features are also helpful:
- However, population density tells us that the lower the population density, the higher the chance that the county gets classified into a 'high' category.

Overall, the results of the models are only as good as the data itself. We have engineered features to build models with practical-use cases and the ability to provide meaningful explanations as well: these provide pretty interesting insights such as the impact of social distancing features being very strong; however also provide confusing insights, which don't make sense such as those for health-care features and demographic features. However, this would make sense since there is not much variability in these features in the first place for the model to play around with. Regardless, the take home message is that social distancing features are extremely important.

VI. Discussion

Here we outlined our data cleaning process which allowed us to perform PCA on counties with a non-zero number of cases, linear regression using the county specific features

to predict the number of cases in a county, and finally a classification model that allows us to categorize counties in low, medium and high categories of mortality rates. The features that carried the greatest deal of predictive weight for our linear model included some markers of the total population in a county such as 'PopulationDensityPerSquareMile2010' and 'CensusPopulation2010'. Other features of importance that were not directly related to population size included "dem_to_rep_ratio" and "SVIPercentile". The democrat to republican ratio within a county could be a proxy of assessing how rural or urban a county is, since urban counties typically favor democrats. The most interesting features we encountered were 'SVIPercentile' and '3-YrMortalityAge¡1Year2015-17'. These features were interesting to us because based on the way we imputed age specific mortality rates we would expect the mortality for age ¡1 to be very sparse but it still represented strong predictive power. SVIPercentile is an interesting feature because it's the CDC's measure of how vulnerable the population of a county is in terms of healthcare outcomes. We were surprised how strongly this one feature influenced our model, especially since there were other disease specific mortality rates in our feature selection process which did not make the cut. We thought ICU beds or other measures of healthcare capacity would be good for predicting lower mortality. However, these features ended up not being very useful in our linear regression analysis. These results for the linear regression model are in contrast somewhat with our findings as a result of classification analysis which shows that features such as how early counties implemented stay at home orders and other social distancing measures strongly predict mortality rates. Underlying descriptors of the county's health such as disease specific mortality rates are also important considerations. But yet again, healthcare capacity is not a strong feature for classification or case number prediction. One of the biggest challenges we faced was in the process of data cleaning. There were many columns with lots of NaN values which required us to examine each column individually to determine how to treat NaN values and often required us to make several assumptions about when data was missing. Furthermore, we had to drop one of our columns due to the high number of NaN values and the lack of a good way to impute missing values. Our model made the following key assumptions. We limited our analysis to 50 US States only (this does not include District of Columbia). For our model built on mortality rate: we only train and predict on counties on at least 30 cases since that gives us reasonable and reliable mortality rate estimates. For the model built on cumulative cases: we train and predict on all counties that have a non-zero number of cases that come out of the data manipulation / preparation. We decided to exclude counties with zero cases as we feared we would be training a model to predict low values for cumulative number of cases. Further assumptions around data manipulation and modelling such as missing value imputation, were outlined previously in this report. For the purposes of linear regression, we also assumed that there exists some linear relationship between the features we chose and our observations. We further assumed that with the features used in our model, we could predict various observations reliably. The ethical issues with the datasets we used largely revolve on public health record keeping for rural counties. We found that rural counties often had the greatest number of missing values for the columns in the dataset. For age specific mortality and a few other features, we imputed missing values as 0's, which is akin to making the assumption that missing values for things like child mortality indicate that the event didn't happen. It could very well be the case that due to a lack of robust reporting, these events do occur in counties, the data could be overlooking actual health disparities or predictors of greater COVID-19 severity. This could certainly impact resource allocation and becomes an ethical issue at the level of resource distribution to disadvantaged counties. Furthermore, there are ethical problems inherent in the questions we are trying to answer. In trying to assess which counties have it the worse when dealing with COVID-19, analyses could unintentionally implicate the most populous counties are needing the most resources. However, it could be the case that rural counties and those with larger minority populations are actually the least equipped to deal with pandemics, and thus need more resources. Alternative, our analysis could identify socioeconomically disadvantaged counties are being forecast to have a severe pandemic which could lead to the imposition of greater lockdown orders instead of addressing underlying unmet medical needs in these communities. To address these inherent ethical questions, great care has to be made before making recommendations and all analyses should make the effort to include other datasets with socioeconomic data so that inequalities that could shape the course of the pandemic are not paved over. In terms of next steps and what additional data or hypotheses we would consider moving forward, fitting an SIR model to our time series data could help us better gauge pandemic severity. To construct such a model, we would need more time series data as well as additional data on specific counties pertaining to the efficacy of social distancing orders. In general, we would like to improve our classification and linear regression models. To this affect, we think that more data on specific county features would aid us in building better models. Not just more features, but more complete data that does not contain as many NaN values. We would also like to expand our analyses to the realm of predicting "vulnerability" to severe COVID-19 outbreaks. Our efforts here revolve around trying to describe the current outbreak in terms of mortality or number of cases, but we could also determine which counties are most likely to be hit by severe pandemics and allocate resources ahead of time. We also could test hypotheses using the latitude and longitude values of states and counties, since proximity to other states and counties with lots of infections could be a risk factor for sever pandemics.

VII. Figures

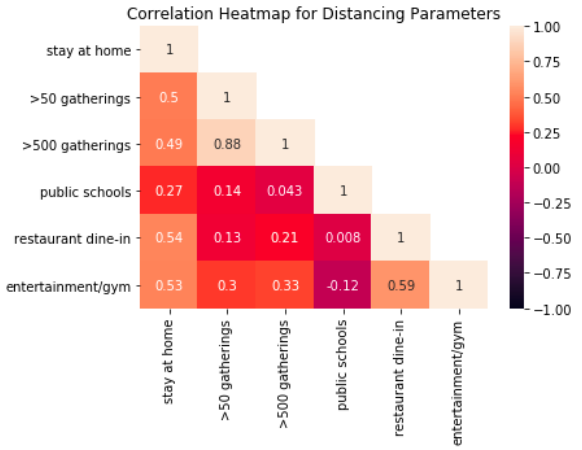

Fig. 1. Heatmap of Social Distancing parameters

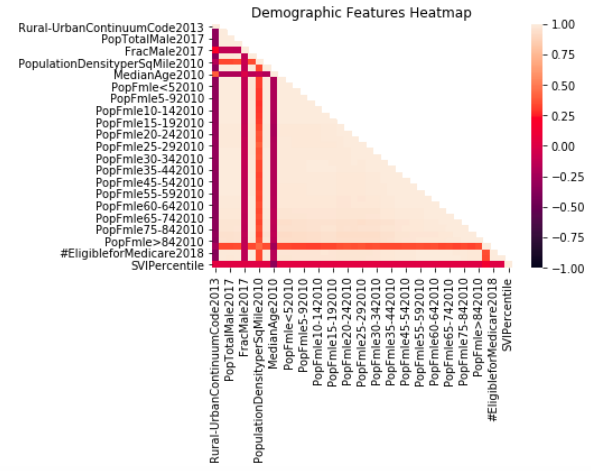

Fig. 2. Heatmap of Demographic Parameters

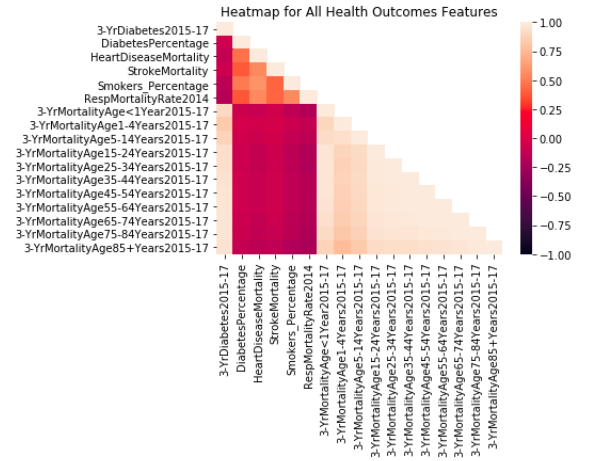

Fig. 3. Heatmap of Health Outcomes Parameters

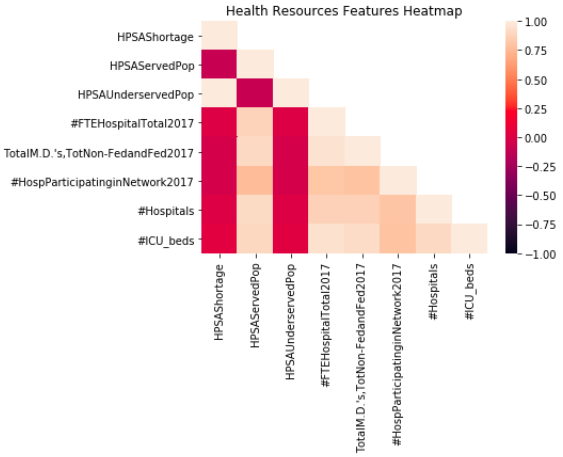

Fig. 4. Heatmap of Health Resources Parameters

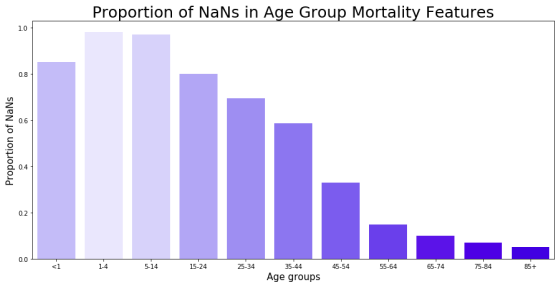

Fig. 5. Proportion of Column that is NaN for Age Specific Mortality Rates

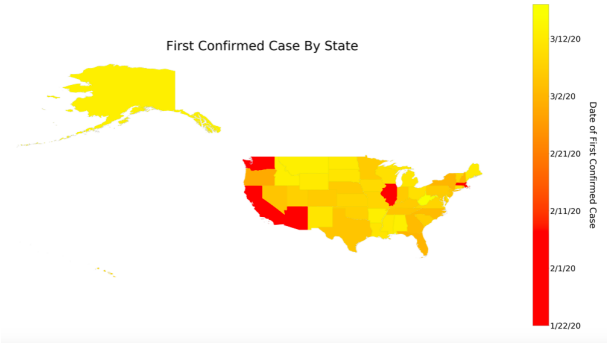

Fig. 6. Map of First Cases by State in US

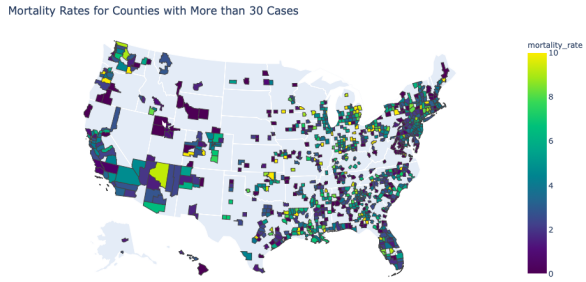

Fig. 7. Map of Mortality Rate by County in Counties with More than 30 Cases

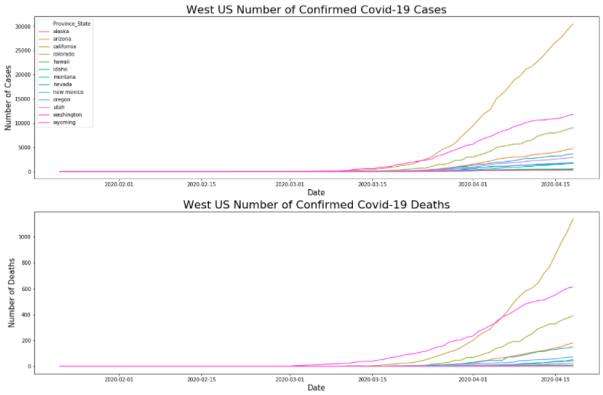

Fig. 8. Cases Over Time in the Western US

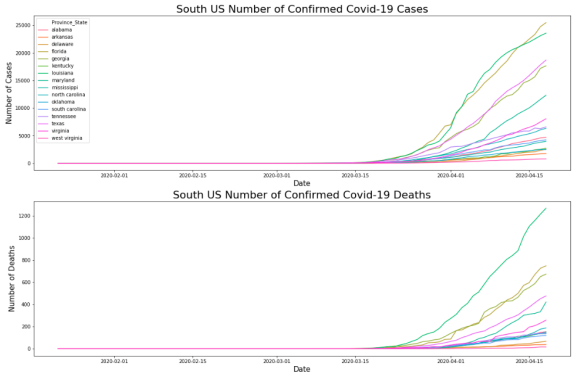

Fig. 9. Cases Over Time in the Southern US

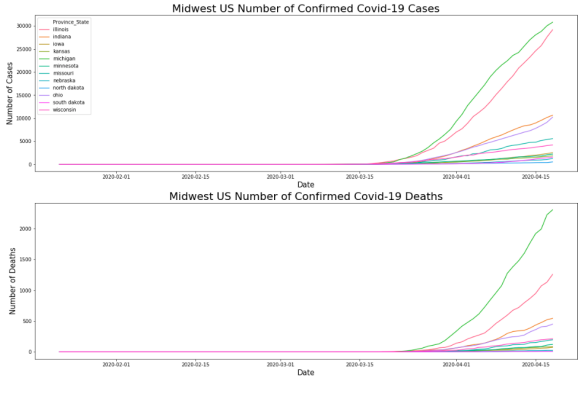

Fig. 10. Cases Over Time in the Midwestern US

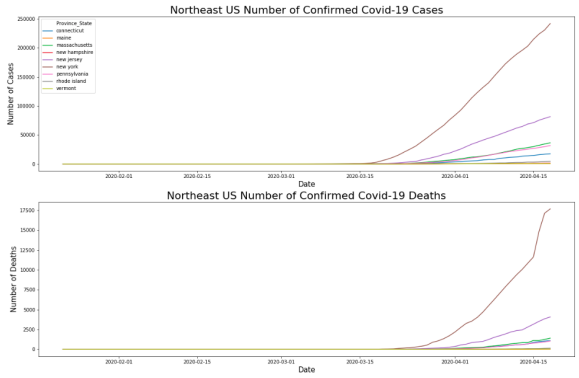

Fig. 11. Cases Over Time in the Northeastern US

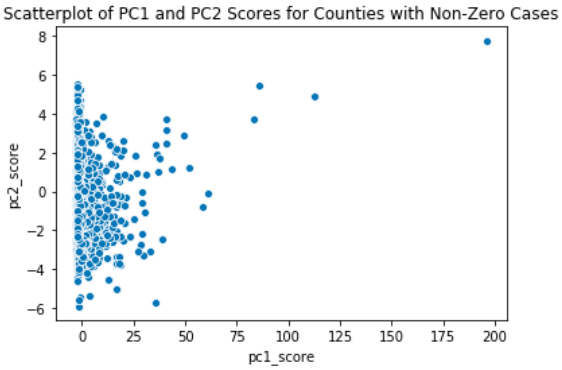

Fig. 12. Scatterplot of pc1 and pc2 scores for all counties.

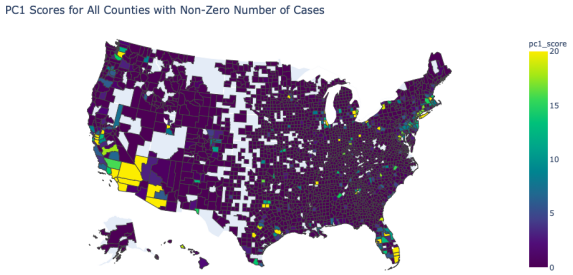

Fig. 13. Map of all PC1 Scores by County in counties with non-zero number of cases.

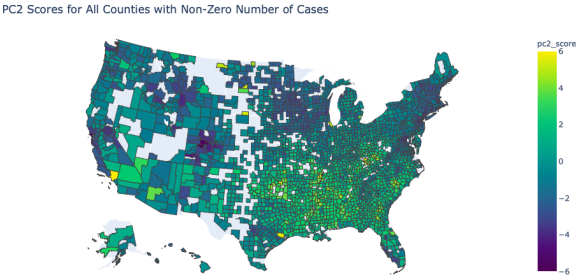

Fig. 14. Map of PC2 Scores by County in counties with non-zero number of cases

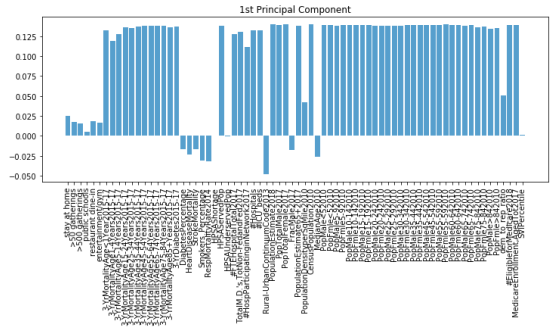

Fig. 15. First Principal Component

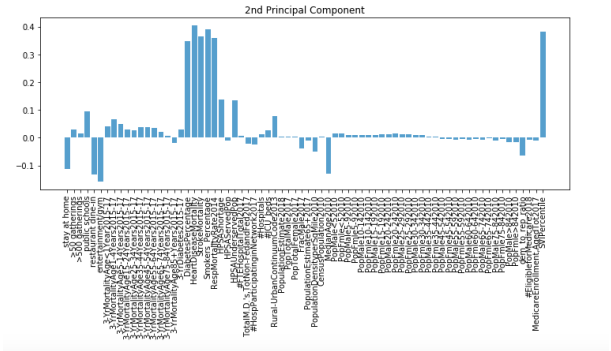

Fig. 16. Second Principal Component

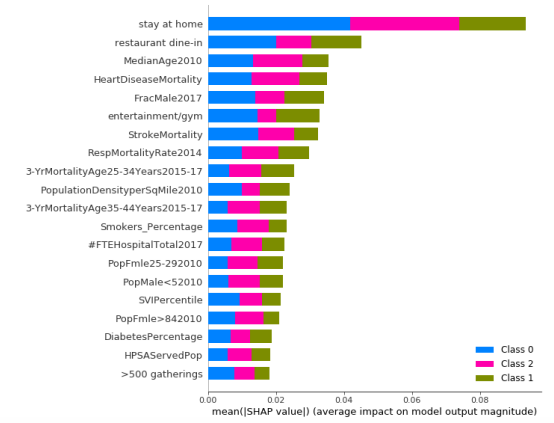

Fig. 17. Random forest classifier analysis

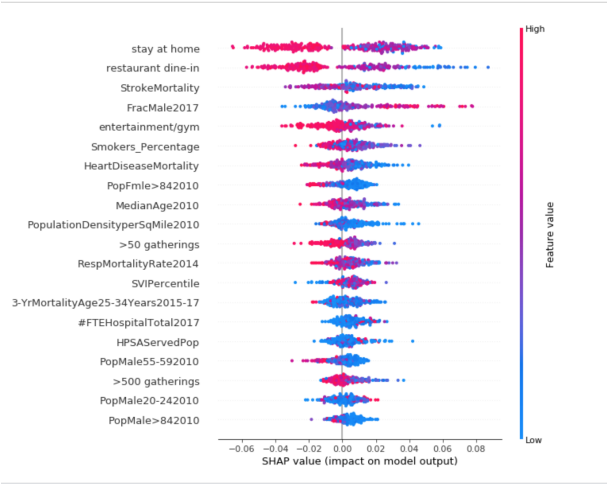

Fig. 18. SHAP plot

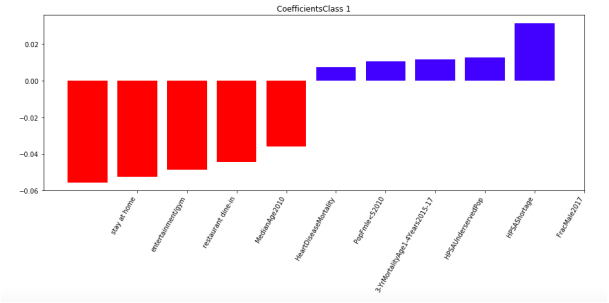

Fig. 19. Class 1 Coefficients for second logistic regression attempt.

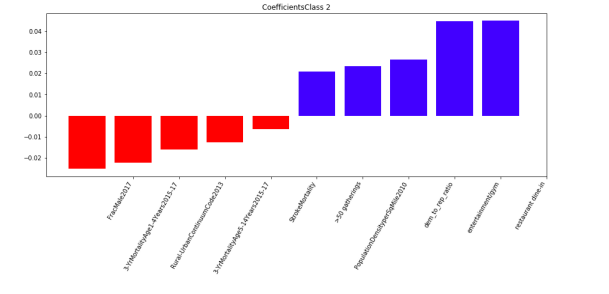

Fig. 20. Class 2 Coefficients for second logistic regression attempt.

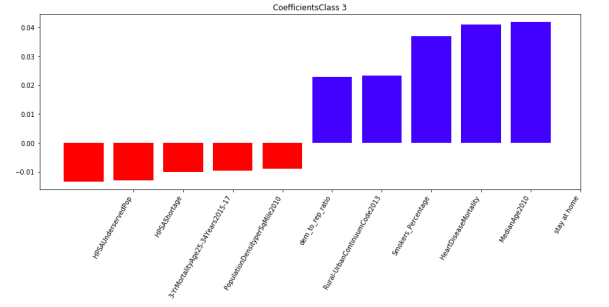

Fig. 21. Class 3 Coefficients for second logistic regression attempt.

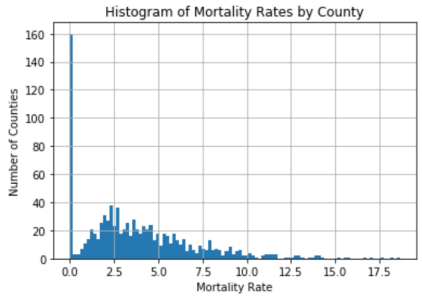

Fig. 22. Mortality Rates histogram.