Table of Contents

I. ABSTRACT

COVID-19 is a global pandemic that has claimed the lives of thousands, causing countries to take drastic measures in hopes of "flattening the curve".[1] There has been extensive debate amongst researchers and scientists about which features have the greatest impact on the virus's spread, and which can slow it. This analysis seeks to identify these features by creating predictive models for confirmed case and death rates across U.S. states. Furthermore, the models are used to simulate case and death rates in the absence of particular features. After data cleaning, feature consolidation, model selection, cross validation, and outlier removal, a ridge regression model pointed towards Heart Disease Mortality, % Female in U.S. 2017, and Stroke Mortality as the best predictors (top 4 for both predictions) of confirmed cases and deaths given time-bound data from The Yu Group at UC Berkeley Statistics and EECS. When median age was increased in our model, predicted incidence rates increased as expected, which could be used to develop predictions for different populations under otherwise similar conditions under COVID-19 pandemic events in the future. Even though these results support the assertion that certain features are the most relevant, an important ethical consideration here is that socioeconomically disadvantaged groups are being systematically excluded in this model. The current prediction model we use only identifies impactful measures for people who can afford to get tested/go to the hospital. That being said, this ethical issue is also a limitation in our model given data availability, therefore the best we can do is acknowledge the problem when generating takeaways.

II. RESEARCH QUESTIONS

i. Which features best predict COVID-19 confirmed cases and deaths?

We are interested in determining the features that have the greatest influence on the rate of confirmed COVID-19 cases. Additionally, this can be extended to predict the rate of COVID-19 deaths. Analysis of these variables enable us to draw inferences about which measures are most effective in slowing the spread of the virus.

ii. How would confirmed cases and deaths change if certain features were different?

We can also utilize our model to answer some intriguing "what if…" questions. For example, what if California had a higher median age? Or what if there were more hospitals? How would this impact the predicted rate of confirmed cases or deaths? The results seen here could lead to insights about the demographics and available facilities in a state in relation to COVID-19 and could be useful for further research to prepare for the anticipated second wave of this virus.

III. EXPLORATORY DATA ANALYSIS

i. Data Cleaning

In order to build a model that was least affected by missing values and outliers, we chose to analyze the data of states (including DC and Puerto Rico), rather than counties. There were a few other territories, such as Guam, that we dropped because they did not have sufficient data available. Since many of the features we wanted to test in our model were provided in the counties dataset, a large portion of our cleaning process was dedicated to grouping and averaging these values amongst each state. For example, in order to calculate variables that were provided as proportions for counties, we first converted them into numbers by using the population count data, and then summed them when grouping by states, and then converted them to proportions once more, using state population data. Additionally, we reset all of our data that was in the form of counts by converting them to per capita counts.

We encountered null values in much of the data that related to stages of lockdown, indicating that these counties may not have mandated the lockdown at the time of data collection. In order to fill these null values without skewing the data, we examined the counties table and noted the latest date of these lockdowns, all of which seemed to be around the first week of April. Since the confirmed and deaths datasets' latest recorded date

[1] "Mayo Clinic Q&A Podcast: The Importance of Isolation to Flatten the Curve on COVID-19 (Coronavirus)." *Mayo Clinic*, Mayo Foundation for Medical Education and Research, newsnetwork.mayoclinic.org/discussion/mayo-clinic-qa-podcast-the-importance-of-isolation-to-flatten-the-curve-on-covid-19-coronavirus/.

was April 18th, we filled the null values with April 19th because it would demonstrate that these counties had not locked down these venues, at least until this date, which is still around two weeks behind the latest recorded date of lockdowns.

ii. Visualization

Since our main focus of this project was building a model for incident rate, in our exploration of the data, we decided to examine the incident rate for each state using a bar graph. Ppl

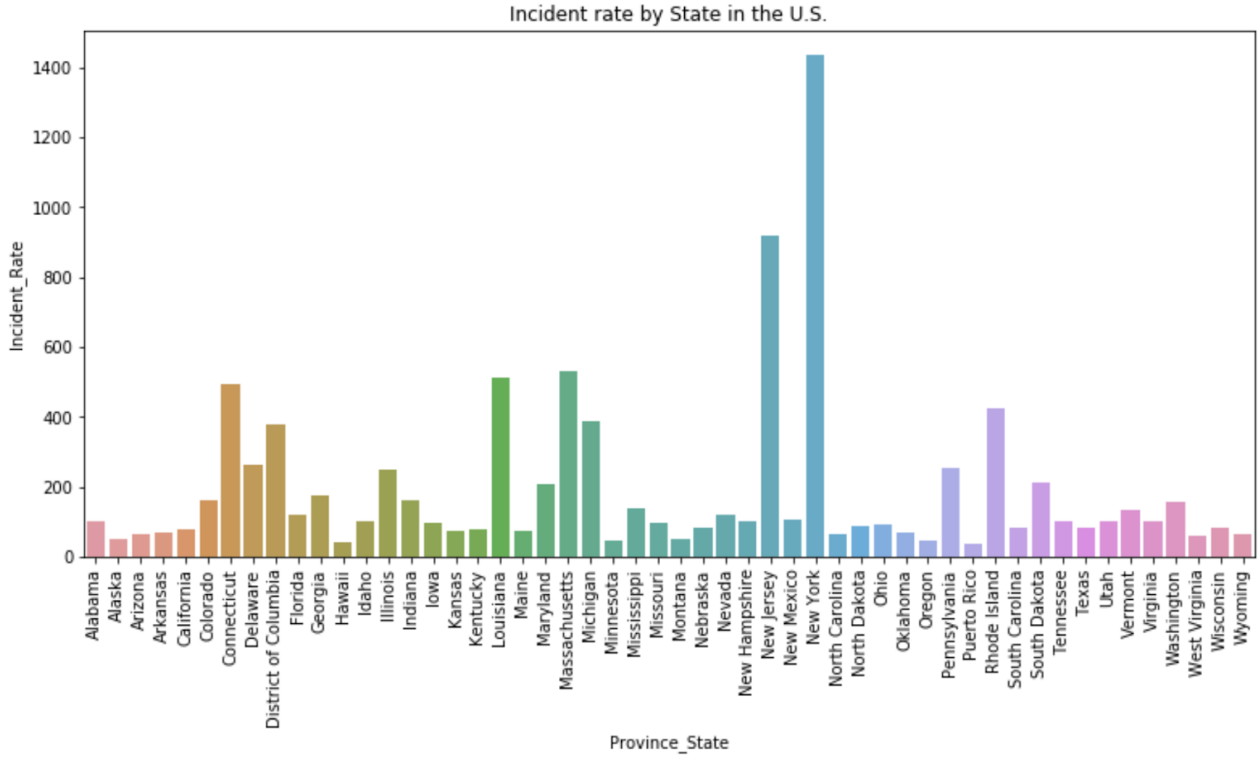

Image 1: Distribution of COVID-19 incident rate by state

This analysis helped us identify that New York was a major outlier, and should be kept in mind as we train and edit our model. Additionally, we wanted to look at how new cases have arisen, especially in relation to the stay-at-home ban, as this is currently a hot topic of discussion amongst the scientific and political community. We used a line plot to examine how state's new cases have changed over a time, while placing a vertical line at the point the state declared a stay-at-home order. An interesting trend emerged. States with earlier bans tended to have more cases per 100k. For example, as shown to the right and below, California and New York both had earlier bans than Texas, but experienced a higher peak in the number of cases. This seems counterintuitive, but it could be due to other demographics in these states. Regardless, this is an interesting relationship and would therefore be a good feature to include in our model.

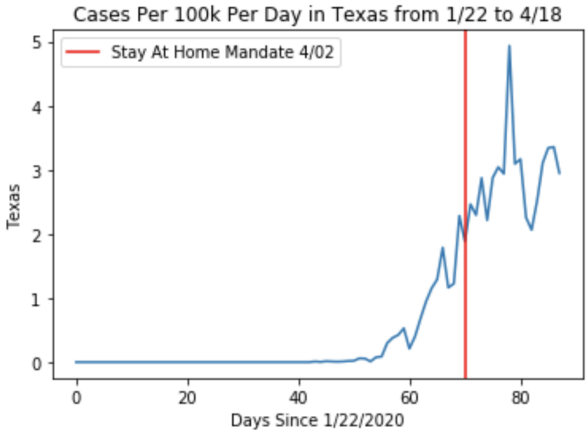

Image 2: Cases per 100k in Texas

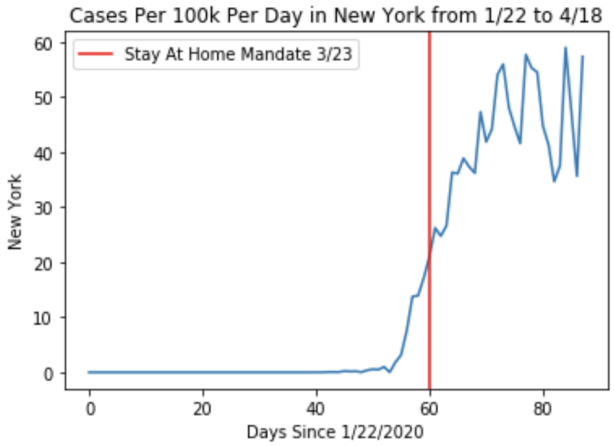

Image 3: Cases per 100k in New York

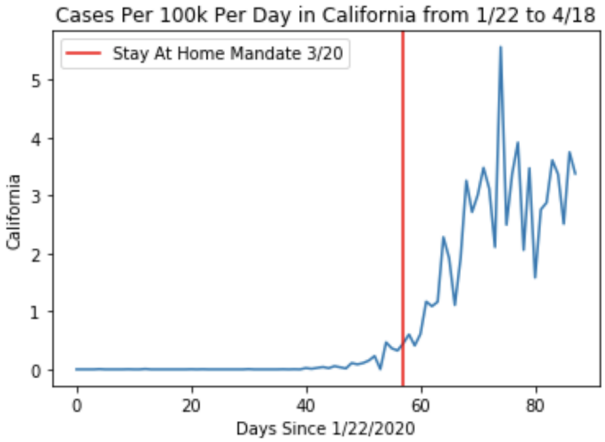

Image 4: Cases per 100k in California

IV. FEATURE SELECTION & ASSUMPTIONS

i. Features

Many of the data features in the COVID-19 datasets were co-dependent, which would be problematic if we were to use them directly to create linear models. Secondly, the abundance of features compared to the comparatively little amount of data entries (using 52 states) meant that the likelihood of overfitting to data was high. In order to optimize model accuracy, we chose a subset of available features after grouping by state. After removing dependent variables identified through GitHub descriptions, we leveraged a pairplot visualization to verify independence between features and correlation with COVID-19 case and death rates. All three of the features shown seem to have relatively positive association with incident rate. We noted that heart disease mortality and stroke mortality have some positive association, which could be a point of error in our model. Since we ended up using a ridge regression model, however, we felt that we could still include both features because the regularization process will eliminate any multicollinearity between the features that would lead to overfitting.

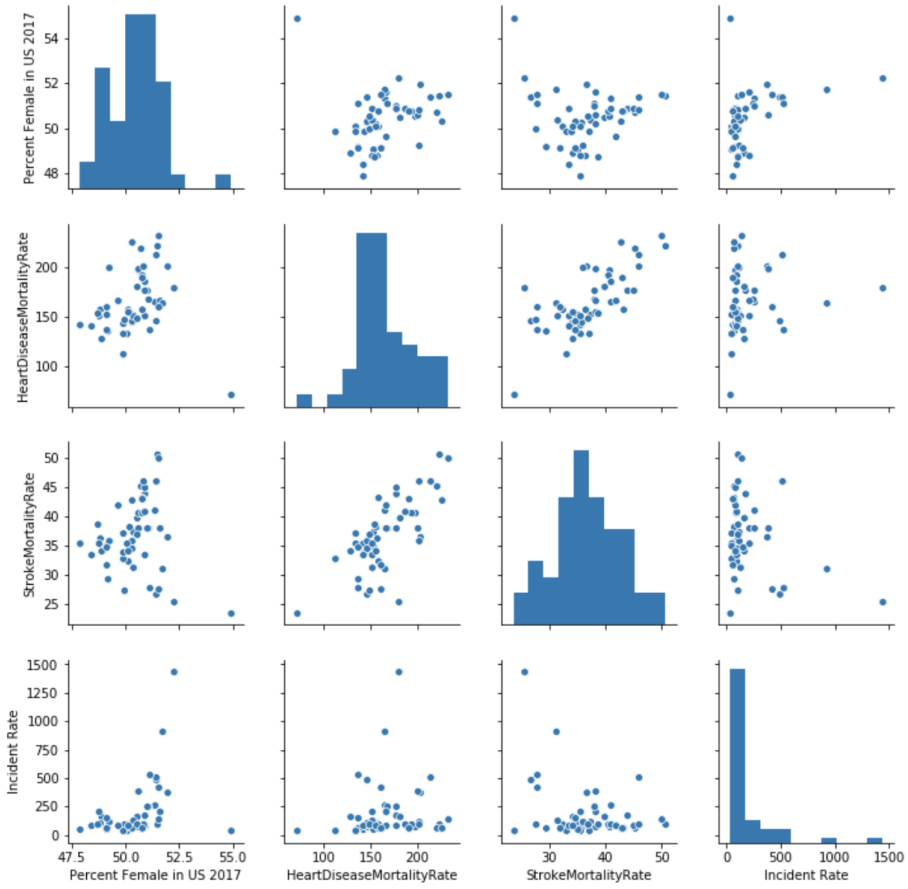

Image 5: Pairplot of select features and correlations

Selected features included characteristics identified as important though secondary research and changes made directly in response to the novel virus by government authorities[2]. A full list of these features can be found in our notebook.

ii. Assumptions/Decisions

Some of the assumptions we made in developing our model stem from our process of grouping the counties data into states data. For example, since we had lockdown dates for counties, we decided to take the mean date when grouping by states. We felt that this was a reasonable assumption because while most counties in a particular state had the same lockdown dates, taking the average ensures that even if there were some outliers, our data will be a good representation of the entire state's lockdown date. Additionally, we are assuming that the population data that we used to transform our counties' proportions data into states is accurate because our methodology is dependent upon this accuracy.

For our analysis, we started with linear regression to fit our data. We chose this method of fitting because it provides a simple way to analyze the individual effects of different parameters on confirmed COVID-19 cases and deaths. In our model testing, we did not use a training/test set due to the very small amount of data available. We substituted this train/test split with just using 5-fold Cross Validation to find the average CV RMSE to evaluate our model. Because we had so few points and still a substantial amount of relevant features we chose to use, we changed our model to a ridge regression so that the error coming from potential multicollinearity of our selected features and overfitting our data would be minimized.

V. MODELS

i. Linear Regression

We found our linear regression model for predicting confirmed cases (incident rate) to have an RMSE of 97, whereas a cross validation score using 5-fold CV of around 260. The difference between the RMSE and CV is expected, but could be an indicator of overfitting the data. However, this could also be a result of our small sample size. Since there are only 52 states, cross validating portions of these states is expected to produce substantial error because each subsection selected in the cross validation process is unlikely to be representative of the entire group. This limitation of our sample size is important to note, and is discussed in further detail later on.

ii. Ridge Regression

We chose to go a step further than linear regression by using ridge regression, which helped us feature-select our model. We also standardized all of our features in our pipeline, which allows every value to be on the same scale. This normalization allows us to compare the influence of each feature based on the values of the coefficients. We tested 1000 alpha values between 0 and 15 and cross validated each one, so that we could identify the alpha parameter that would result in the lowest cross validation score. As shown by our visualization of the scores associated with the alpha values tested, the best score falls at an alpha value of around 6.

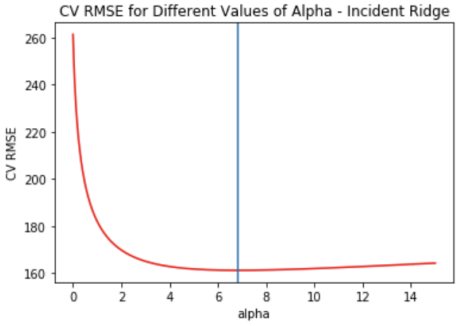

Image 6: RMSE alpha plot for Incident Ridge Model

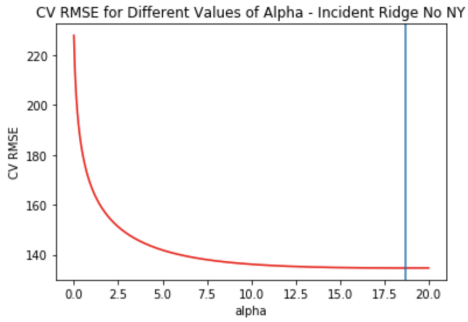

Image 7: RMSE alpha plot for Incident Ridge Model (No NY)

Ultimately, our ridge regression model using our best alpha value resulted in an RMSE of 109 and CV of 161. In order to analyze the strength of our model, we looked at the range of our feature variables, along with the standard deviation of y-variable in order to ensure that our RMSE was small enough in comparison. With a standard deviation of 238 for incident rate, our model seems to be performing fairly well.

Since we had noticed that New York was a major outlier during our EDA process, we decided to try fitting another ridge regression

[2] Roberts, Siobhan. "Flattening the Coronavirus Curve." *The New York Times*, The New York Times, 27 Mar. 2020, www.nytimes.com/article/flatten-curve-coronavirus.html.

model without New York. Using the same process as before to find the best alpha value, this model resulted in an RMSE of 104 and a CV of 134. Unsurprisingly, the RMSE was lower and the CV score was much closer to it because it was not being skewed by an outlier in its train/test splitting process.

The second part of our question related to mortality rate, rather than incident rate, so we ran another ridge regression with mortality rate as our y-values. We chose to use the same features, as the ridge regression minimized the effects of any nonoptimal collinearity between them. This resulted in an RMSE of 1.03 and CV of 1.45, which is reasonable compared to the SD of mortality rate which is 1.5.

VI. FINDINGS & CHALLENGES

i. Best Predictor for Incident Rate with full data set

A ridge regression model as described above identified that Testing Rate (followed by Stroke Mortality Rate and % Female in U.S. 2017) is the leading predictor of the COVID-19 Incident Rate. Note that Incident Rate is measured as the number of incidences per 100k people. This translates to a takeaway that states with more COVID-19 testing experience higher rates of confirmed coronavirus cases. While this may sound trivial, it affirms the importance of scaling testing resources so that the public can get a more accurate picture of the spread of COVID-19. While developing this model, we were challenged by the fact that male and female population features were in fact collinear once we noticed the female population had a high coefficient value, or impact, on the model. Through a trial and error process we decided to convert the male/female group from population counts to proportions and remove the male column to preserve independence among predictor variables. Another challenge with these results is that our secondary research shows that males are more associated with COVID-19 instances, which contradicts our results[3]. On another note, we originally expected the stay at home order date for larger gatherings (>50 & >500) to have a large impact on both COVID-19 incidence and mortality, but it did not in fact have a significant coefficient value for Incident Rate. The complete distribution of absolute coefficient values for features against both incidence and mortality models (without outlier removal) can be found below.

ii. Best Predictor for Incident Rate without NY

Similar to the above subsection, we used a ridge regression model and identified two of the three same features as the leading predictors of COVID-19 Incident Rate (Testing Rate, % Female in U.S. 2017, Heart Disease Mortality Rate), all with positive associations in the model. Beyond the insights already mentioned, by attaining 2/3 same key features without the New York outlier data entry, we can increase our confidence in having more testing and increased caution around identified population groups. A challenge we had when assessing this specific model's validity however, was that the New York data entry had the highest population as expected, which meant that it represented a larger proportion of the already limited data as well. For this reason, our comparison horizontal bar chart and "what if…?" analysis in subsequent sections build off the ridge models predicting incidence and mortality without NY outlier removal. The implications of this decision is possibly skewed data, however we made this decision in light of limited data availability to begin with.

iii. Best Predictor for Mortality Rate with full data set

Here we saw that the strongest predictor for mortality rate using the full data set was % Female in US 2017. Just as in the incident rate model, this feature had a positive effect on mortality rate, with a coefficient of .61. This means if the percentage of females increased by one standard deviation, the mortality rate increased by .61. Interestingly, Testing Rate was only the 10th strongest predictor for COVID-19 mortality rate, despite it being the strongest predictor for incident rate. Features that proved to be far more relevant in predicting mortality rate were Stroke Mortality Rate and Heart Disease Mortality Rate. This result seems reasonable, because regions with higher serious illness mortality statistics may signify that the population is generally less healthy and would expect to have more people die from a virus like COVID-19. It is important to note that as shown in our EDA section above, these two features are somewhat correlated with each other and there may be overlap in our model, despite our regularization. Additionally, we expected Median Age to be important in predicting the rate of deaths, as there is much research showing that COVID-19 affects older individuals more

[3] Greenfieldboyce, Nell. "The New Coronavirus Appears To Take A Greater Toll On Men Than On Women." *NPR*, NPR, 10 Apr. 2020, www.npr.org/sections/goatsandsoda/2020/04/10/831883664/the-new-coronavirus-appears-to-take-a-greater-toll-on- men-than-on-women.

severely; however, this feature proved to have the smallest significance in this model. Below, you can see this relationship and others in a side-by-side comparison of the absolute coefficients for the incident rate and mortality rate models.

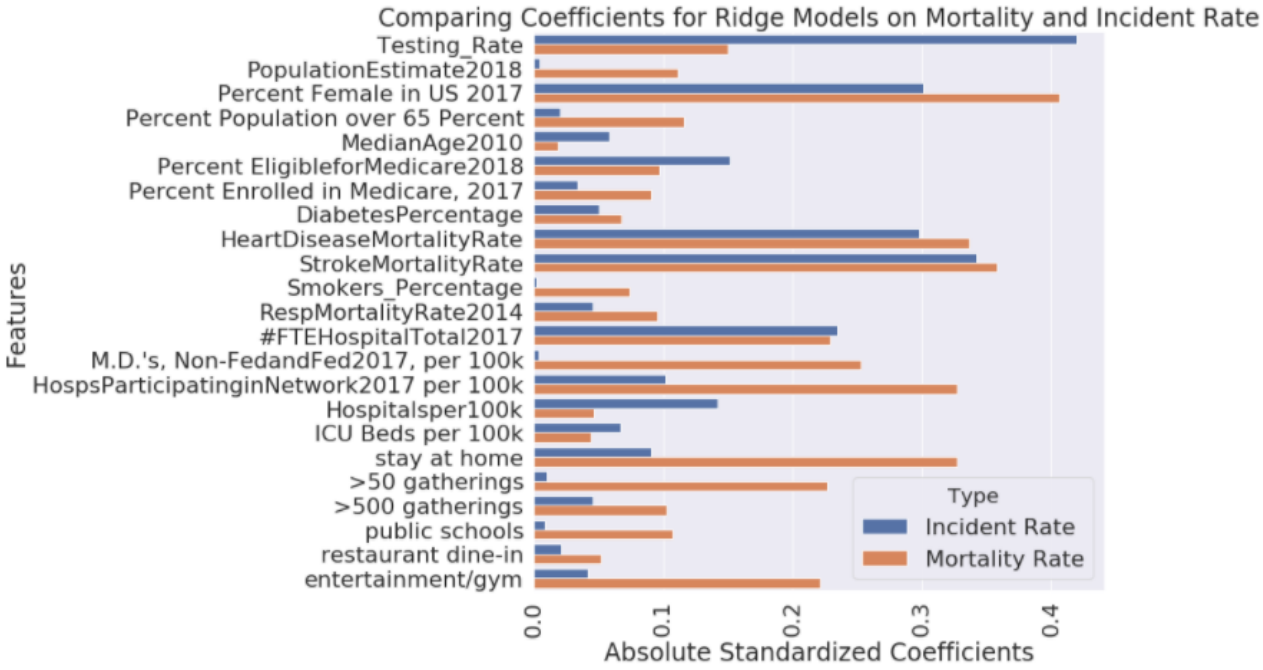

Image 8: Standardized coefficient comparison between ridge models

iv. "What if..." Cases

Another way to interpret our model is by answering some intriguing "what if…?" questions. For example, we chose to investigate what the incident rate of COVID-19 cases would have been in some states if the average median age in 2010 had been 5 years above its recorded age. As seen by the graph below, the predicted incident rate rises notably had the population been older than it was. Specifically, California had a predicted incident rate of 219 given its true median age, whereas it's predicted incident rate was 251 given a higher median age.

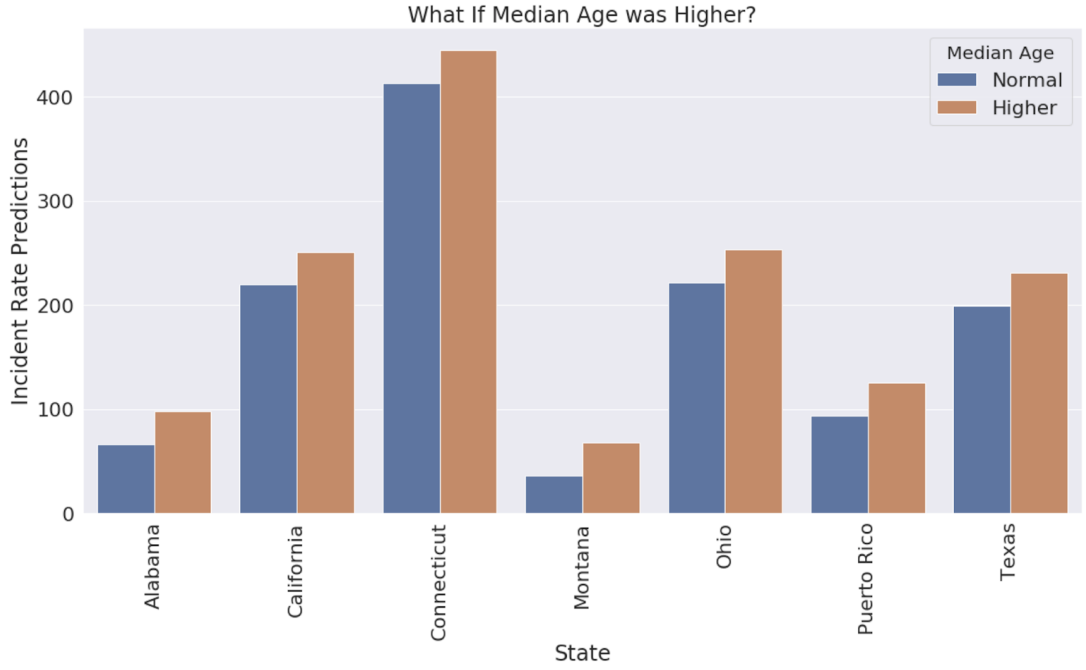

Image 9: Model predictions for COVID-19 Incident Rate with Median Age raised by 5 years

We also addressed what the predicted mortality rate would have been if certain states had more hospitals per 100,000 people. We found hospitals to be positively associated with mortality rate, meaning that more hospitals meant there would be more COVID-19
deaths. While this may seem counterintuitive, because one would assume that more hospitals would mean that more people would be treated and therefore less would die, it is likely due to the fact that we are dealing with reported deaths here, not the true amount of coronavirus deaths. If a state has more hospitals, then they can accommodate more people and therefore may have higher rates of recorded deaths because it is difficult to record deaths of individuals that did not have medical attention.

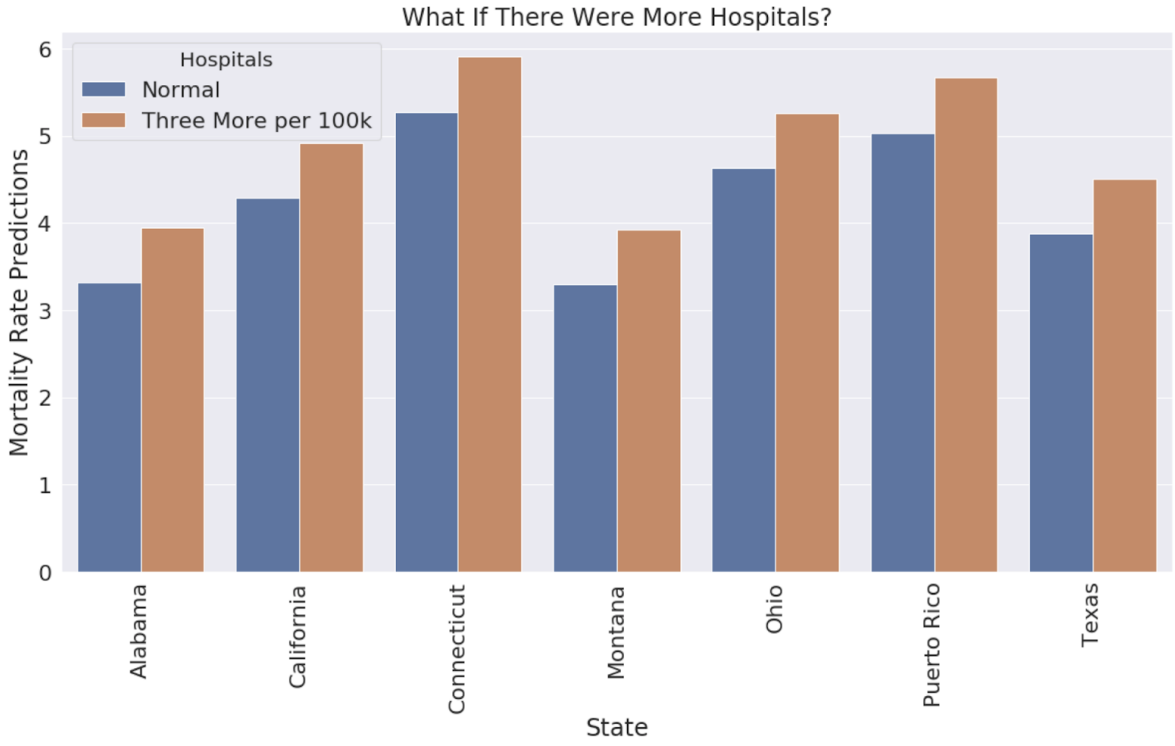

Image 10: Mortality rate predictions with varying hospital numbers

v. Limitations

There are consequential limitations of our model, many of which stem from the data itself. One of the questions we hoped to answer was "which features have the largest influence on the rate of COVID-19 confirmed cases and deaths?" Given the data we have access to, however, we cannot attribute any sort of causation, but rather just association. Also, some of the data we have could actually be a result of an area being greatly affected by the virus, rather than it being a factor in the rate of cases/deaths. For example, states that were affected more severely by the virus may have imposed stay-at-home bans earlier, and therefore our model shows that later stay-at-home bans are associated with lower incident rates. Furthermore, we are basing our results off of reported cases and deaths; however, there is a nationwide shortage of COVID-19 tests and hospital space available, so our inference is not reflective of how many people were actually infected, but rather just how many people were tested and treated. Of course, we are also relying on the accuracy of the data in our model.

Additionally, our sample size of 52 is very small in relation to the 23 features that we are using, which makes it harder to create a model that aligns well with the data. This leads into our next limitation of our model, in that it is only relevant for the U.S. because other countries may have vastly different feature values. This presents a dilemma because we are observing and drawing inferences from a limited region, despite that this pandemic is global and the effects of its spread may be greatly intertwined between countries. We also must note that our model can only be used to predict the rates on April 18th, but not future cases because it does not account for the pandemic's growth factor.

VII. DISCUSSION & NEXT STEPS

i. Ethical Questions

COVID-19 and its impact is greatly dependent on the medical records of a region. The implications of building a model and drawing inference from data that may be systematically biased raises important ethical questions about our investigation. A large portion of our model depends directly on recorded medical data, such as heart disease mortality and diabetes percentage, but this data is only recorded when an individual has visited a doctor or medical facility, and therefore may be systemically ignoring individuals with lower socioeconomic status or that live in rural areas, who may not have access to these facilities. According to the American Psychological Association, African-Americans and Latinos live in poverty around double the rate of Caucasion-Americans, and "Racial and ethnic minorities have worse overall health than that of White Americans."[4] Given that we do not have access to data from these groups, it is necessary to understand that our model may not be representative of all Americans, and should not be publicized as such.

As we move further with our investigation into COVID-19, it is important to note that the results from our previous analysis may not be representative of which factors are the most influential in spreading the virus. As discussed earlier, there can only be association drawn from these variables, not causation. Placing excessive importance on certain features, such as heart disease mortality and entertainment/gym closure dates, while glossing over those that didn't have much of an effect in our model, such as eligibility for medicare and smoker prevalence could prove to be very dangerous, as there is much research that shows these factors are in fact important in preventing the virus[5]. In order to address these concerns while studying COVID-19, it is integral that we take into account the many factors that may lead to the results we see. It is also necessary for us to be clear that our results should not dictate future prevention steps, but rather guide further research.

ii. Next Steps

For our data we see that the strongest predictors for both mortality and incident rate were the percentage of females in the US in 2017, heart disease mortality, and stroke mortality. The former speaks to how there may be differences in how COVID-19 affects different genders, and the latter two speak to the overall public health of a state. As mentioned earlier, the only thing we can really comment on is the association between these variables and their respective incident and mortality rates.

More advanced models with this data can utilize factors such as the growth rate of the virus, different climates in specific geographic areas, racial demographics, and different modeling processes all together. One big limitation of this model is it works with a smaller subset of the COVID-19 data. Between now and the last data point for the provided datasets, we have seen many more cases around the country, giving us more data to work with, potentially making our models stronger and more accurate. Additional features, such as demographic information relating to race would allow us to make an even more sophisticated model, especially as there is much current research indicating that race may be a key player in the rate of cases and deaths.

What this model does provide us with is a framework for further questions to be asked. Our modeling indicated specific variables as having strong effects. These variables can be the focus of further study, seeing how they truly affect COVID-19 cases and can lead to predictions of future cases. It gives a strong starting point for further research and experimentation.

[4] "Ethnic and Racial Minorities & Socioeconomic Status." *American Psychological Association*, American Psychological Association, www.apa.org/pi/ses/resources/publications/minorities.

[5] "Assessing Risk Factors for Severe COVID-19 Illness." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 23 Apr. 2020, www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/assessing-risk-factors.html.

Works Cited

"Assessing Risk Factors for Severe COVID-19 Illness." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 23 Apr. 2020, www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/assessing-risk-factors.html.

"Cases, Data, and Surveillance." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 30 Apr. 2020, www.cdc.gov/coronavirus/2019-ncov/cases-updates/index.html.

"Ethnic and Racial Minorities & Socioeconomic Status." *American Psychological Association*, American Psychological Association, www.apa.org/pi/ses/resources/publications/minorities.

Greenfieldboyce, Nell. "The New Coronavirus Appears To Take A Greater Toll On Men Than On Women." *NPR*, NPR, 10 Apr. 2020, www.npr.org/sections/goatsandsoda/2020/04/10/831883664/the-new-coronavirus-appears-to-take-a-greater-toll-on-men-than-on-women.

Roberts, Siobhan. "Flattening the Coronavirus Curve." *The New York Times*, The New York Times, 27 Mar. 2020, www.nytimes.com/article/flatten-curve-coronavirus.html.

"Mayo Clinic Q&A Podcast: The Importance of Isolation to Flatten the Curve on COVID-19 (Coronavirus)." *Mayo Clinic*, Mayo Foundation for Medical Education and Research, newsnetwork.mayoclinic.org/discussion/mayo-clinic-qa-podcast-the-importance-of-isolation-to-flatten-the-curve-on-covid-19-coronavirus/.