What Predicts COVID-19 Deaths

Authors:
Sunny Shen, Olivia Zhang, Jerome Chen

Abstract:

This project aims to build a model that predicts the number of COVID-19 deaths in different counties in the United States and finds the most predictive features. It explores the relationship between the number of deaths due to COVID-19 and demographic data in different counties. Based on our linear regression model with Lasso regularization, we find that age, the number of people who died of heart disease, population density, and confirmed cases are the most important features in estimating deaths due to COVID-19.

I. Introduction:

Thinking back to the time in the course where we learned about feature engineering and the different models to portray relationships within data, our group project wondered if it is possible to do the same with the COVID-19 data. One subject of curiosity was the number of deaths. However, we are more interested in finding what particular traits or characteristics have a bigger influence on the number of deaths than others, instead of the projected increase of deaths over time like some of the other models. For example, if a county has less of a certain kind of resource or more members of a particular age group, will that affect the number of deaths in that county in any significant way? This project will thus be our attempt in exploring the existence of such a relationship.

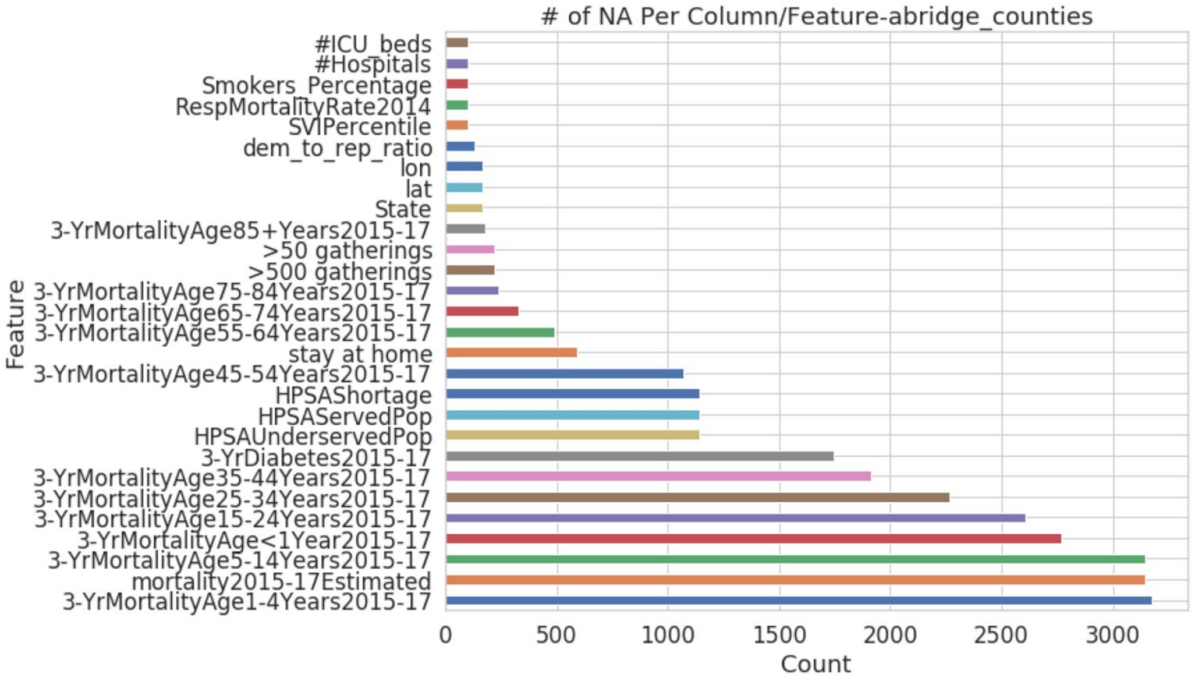
II. Description of Data:

Description of Datasets:

- Abridged_couties.csv: a dataset that contains location, demographic information, hospital information of counties in the United States.
- States.csv: a dataset that has all state names and their abbreviations.
- us-counties.csv: a dataset that has cumulative confirmed cases and cumulative deaths due to Covid-19 from 2020-01-20 to 2020-05-05 in counties in the United States.
- PovertyEstimates.csv: a dataset that has poverty information of counties in the United States.

DEA and Data Cleaning

The bar chart below shows the total number of null values for each column (that has a value >100) in the abridge county dataset. We can see that certain groups of features such as 3-year mortality ages tend to have a large number of null values and thus we might consider taking the whole group out later for our model since it's hard to replace such large quantity of null values, and even when replaced, they are not good choices as features given the lack of information.



In the *Abridged Counties Dataset*, 2 rows aren't counties -- New York City and Kansas City. The reason is that NYC is divided into 5 counties, and the 5 counties are all included in the abridged county dataset. Kansas City lies in different counties, and all those different counties are included in the dataset as well. Therefore, we dropped these 2 rows to avoid double counting. Another reason to drop the rows is that they don't include any actual demographic data. The columns except for county and state name and IDs are all null so those rows may not be helpful to our analysis.

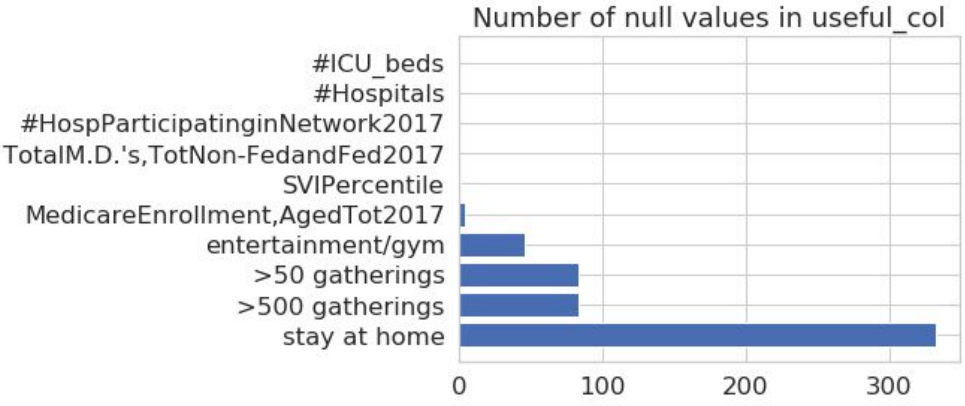
Based on the bar chart plotted in EDA, we see that there are some null values in the "State" column, which has the full names of the states, but there are no null values in "StateName", which has the abbreviations of states. Therefore, we import another data frame called states that has all states' full names and abbreviations. Then, we merge Abridged counties with states so that all counties have complete information on their state names. After we merge states and abridged counties, we name the new data frame *clean_counties*.

For the *county_death* Data frame, it includes the cumulative number of deaths in each county starting January 20th. We use the most recent data available at the time we obtain the data frame so we use the data up to May 5th. EDA shows that there are null values in FIPS columns, and further exploration shows that the rows with null FIPS columns also have "unknown" as county names. Therefore we cannot know which county the rows refer to, so we drop the rows with "Unknown" in the county column. We name the data frame *clean_death* afterward.

Then, we merge *clean_death* and *clean_counties* based on their unique key County FIPS ID and name the new data frame *merge_county_deaths*. The new data frame has information about 2572 counties, which should be a good amount for our analysis purpose. In *merge_county_deaths*, we drop columns that may not help answer our question. For example, we drop columns that have longitude and latitude data and Census Region Name or Census Division Name. Then, we also drop columns that have way too many null values such as "3-YrMortalityAge1-4Years2015-17" and "mortality2015-17 Estimated". We believe that it may not be good to replace null values with the average of the column since they may vary a lot by counties, and that the majority of rows have null values in those columns so

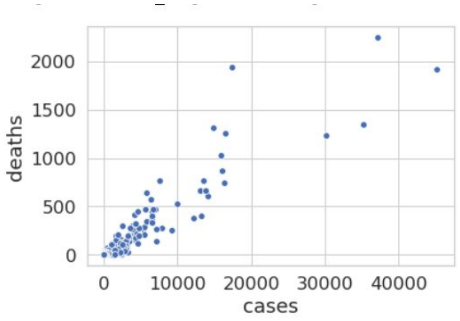
the average is not a good measure either. After dropping those columns, we name the new data frame *useful_col*, which contains 2572 rows and 60 columns.

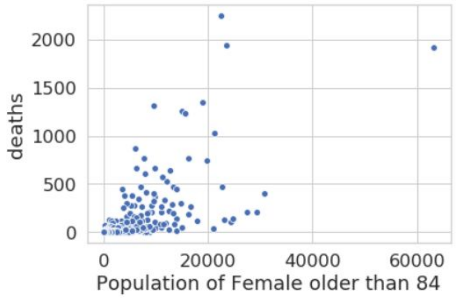
We explore the null values in the data frame *useful_col*, and we find that there are still null values of policy columns like "stay at home", ">500 gatherings", ">50 gatherings", etc. After some online research we find out that for counties that have null values in those columns, they do not have those policies restricting people's activities in their counties, and therefore we fill the values with the ordinal numbers of today's date, assuming the dates of the start of the policies would have an impact on deaths.

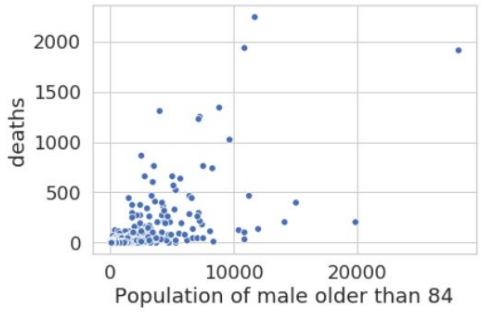


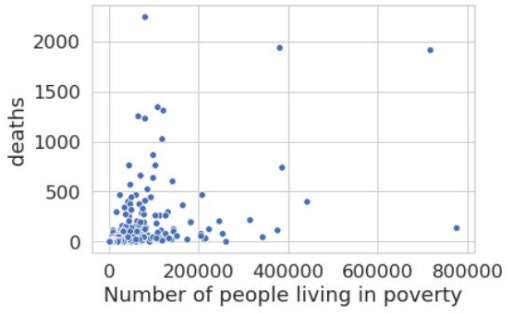
We are also curious about if the level of poverty would influence deaths due to COVID-19, so we imported the PovertyEstimates dataset which includes poverty data in each county. Then we merge useful_col with the poverty dataset and name it *pov_death_county* so that the new data frame has demographic data, COVID-related data, and poverty data of the counties. We fill the 4 null values in "MedicareEnrollment, AgedTot2017 " with national average enrollment rate multiplied by county population, and we fill the 1 null value in "SVIPercentile " with the national average. After that, our *pov_death_county* has no null values and is ready to be used for model building.

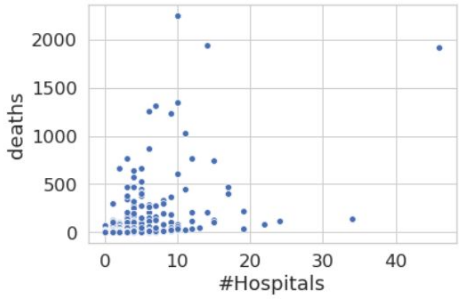
Intuitively, we think that there may be a positive relationship between deaths and features like the number of confirmed cases, the number of older people, or the number of people living in poverty. There may be a negative relationship between deaths and features like the number of hospitals, median income. Here are some scatter plots exploring the relationship among features.

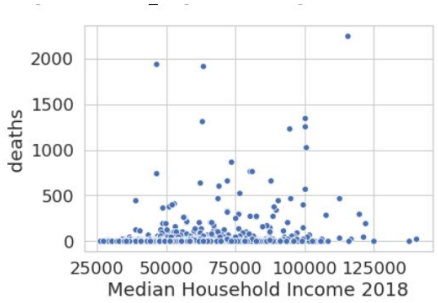












While there seems to be a linear relationship between cases and deaths, between the older population and deaths, between people living in poverty and deaths, we don't see an obvious relationship between median income and death. While we expected to see a negative relationship between the number of hospitals and deaths because there are more medical resources, there seems to be a slightly positive relationship. One explanation could be that counties with large numbers of hospitals also have a large population, so there may be more deaths in those counties.

III. Description of Methods:

Linear Regression:

We use Linear Regression to estimate the number of deaths in each county. We believe that there is a linear relationship between the number of deaths due to COVID-19 and demographic or poverty information of the county.

First of all, we randomly split the *pov_death_county* data frame into training and testing sets, and since all our features are quantitative, we standardize the feature columns to train our linear model. We use root mean square error to judge the accuracy of our results, and we use 5-fold Cross-Validation to make sure that our model does not overfit. We use the absolute values of coefficients to judge the importance of features.

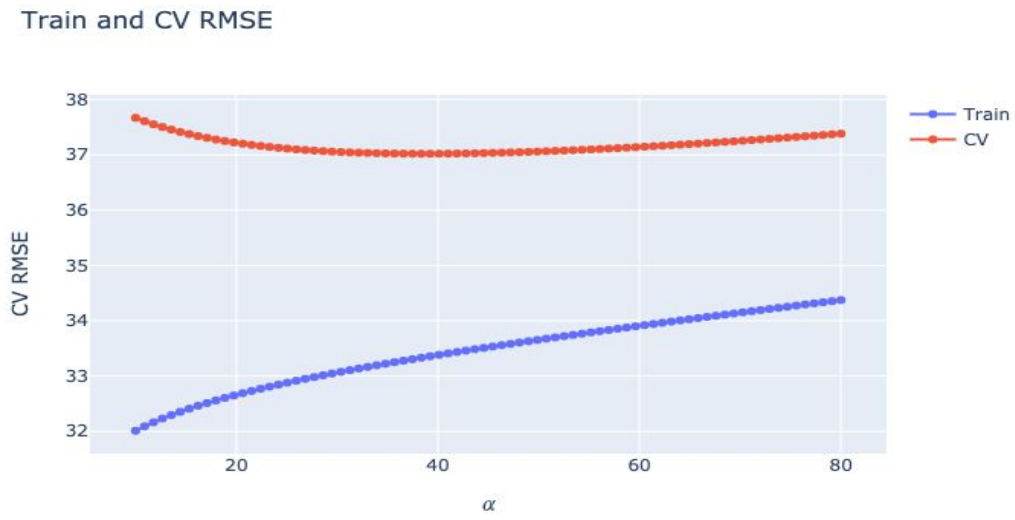
Because there are 58 features in the model, model would likely overfit, so we add both Ridge and Lasso regularization models.

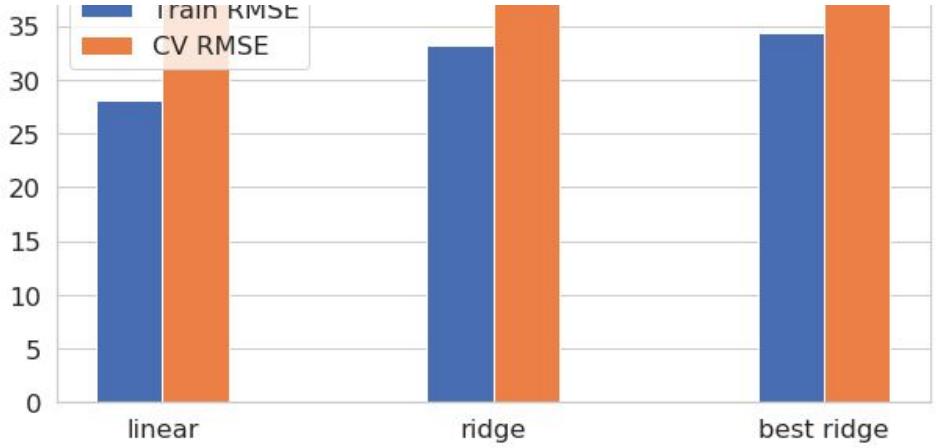
Regularization:

First of all, we trained our linear regression model without regularization. The Training RMSE is 28.03 while the average CV RMSE is 37.68, which indicates that the model overfits.

We trained our linear regression model with Ridge regularization and default alpha value of 0.5. The training RMSE is 28.97 while the average CV RMSE is 38.02. The model still overfits. Then, we

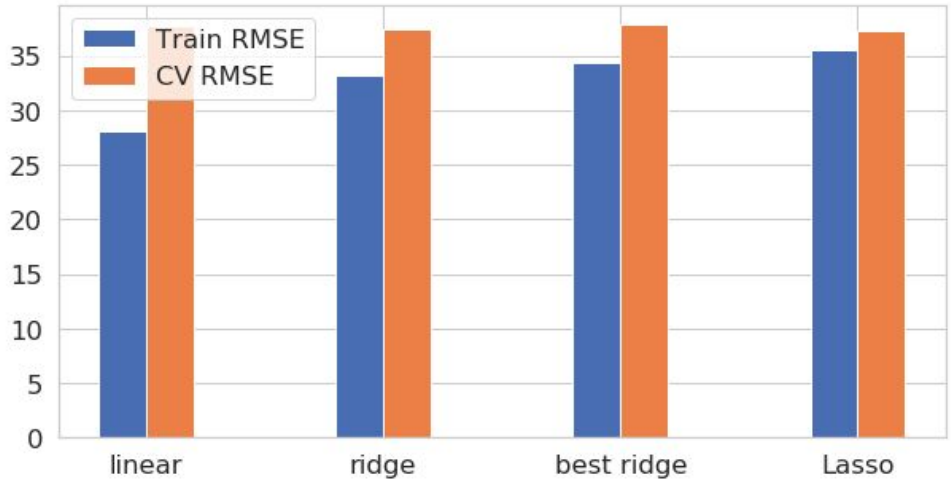
tried alpha values from 10 to 80 and plotted the resulting training and cv RMSE. We find that the CV RMSE is minimized when alpha equals to 38.35. With the best alpha value, we get a training RMSE of 34.32 and a CV RMSE of 37.80.





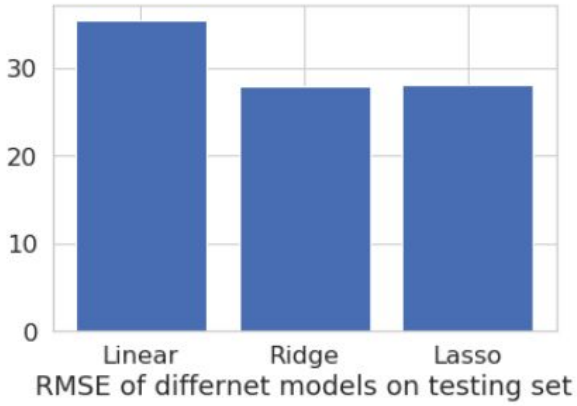
While the ridge model does not overfit, the purpose of our analysis is to find the most predictive features of the number of deaths in a county, so we use Lasso regularization, which is good for finding the best features, to train our model. With the Lasso model, our training RMSE is 35.58 and our CV RMSE is 37.32, which further reduces the possibility of overfitting.

IV. Summary of Results



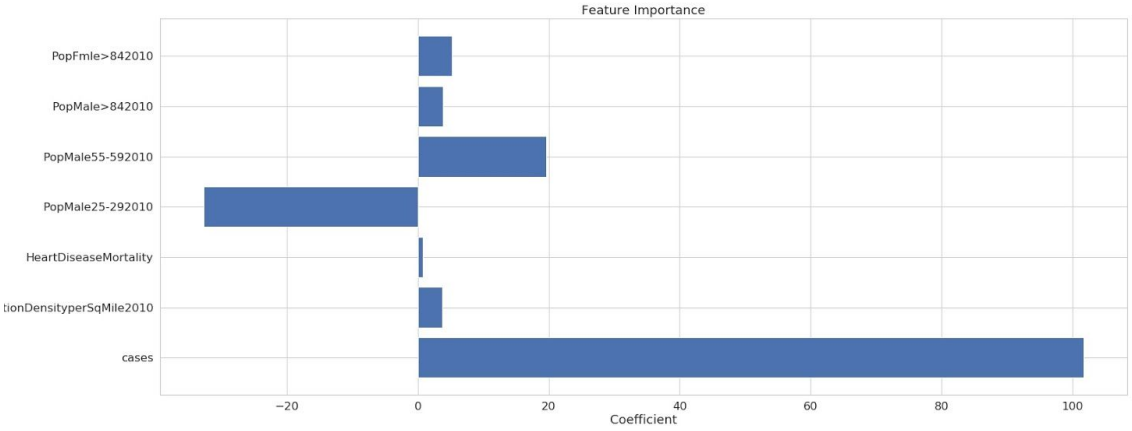
After trying linear regression without regularization and adding Ridge and Lasso regularization, we decide that the Lasso model is the most suitable for our project because it does not overfit and it can tell us what the most predictive features are based on the coefficients.

Then, we use our Lasso model on the testing set and get an RMSE of 28.02. To further justify the use of the Lasso model, we also use the ridge model and the linear model without regularization on the testing set. The ridge model gives us an RMSE of 27.98, which is very close to the result of the Lasso model. The linear model without regularization gives us an RMSE of 35.39, which again proves that this model overfits and justifies the use of regularization.



The features with coefficients that are not close to 0 are the features that are important in estimating death numbers due to COVID-19. We see that the population of both males and females older than 84, the population of male between 55-59 years old, the population of male between 25-29 years old, people who died of heart disease, population density, and confirmed cases are the most important features in estimating deaths. We see that there is a very strong positive relationship between the number of confirmed cases and the number of deaths, and there is a positive relationship between the number of older people and the number of deaths while there is a negative relationship between the number of young

people and deaths. We also see a positive relationship between the number of deaths in COVID-19 and population density & heart disease mortality.



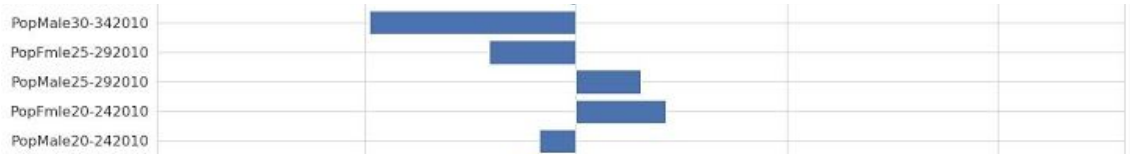
V. Discussion

(i) One of the interesting features that we chanced upon while deciding on the best model was the number of diabetes individuals. This appeared during the testing of the linear model without regularization, where the coefficients revealed to us that there seems to be a negative relationship (a subpart of the graph of coefficients is attached below) between the number of diabetic people and the number of deaths from COVID-19. It was to our surprise as one would assume that individuals that are already ill to other conditions may be more at risk of additional infections and conditions. However, it seems to be slightly to the contrary according to the county dataset. Although the linear model was not our final model, this was still a surprising feature that we would like to address.

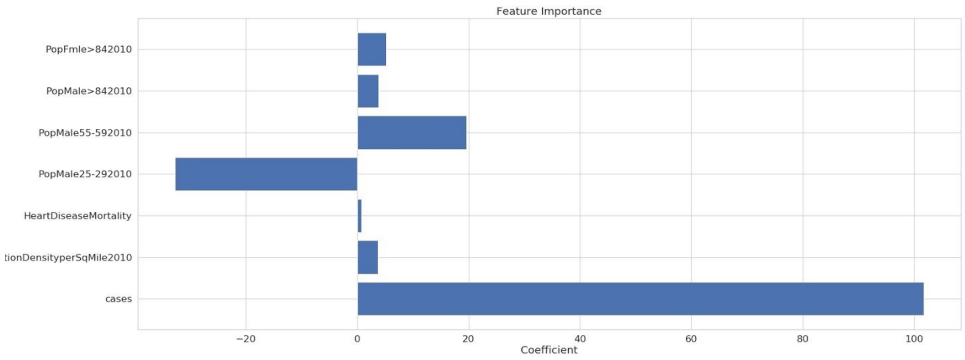
Part of the coefficients of the Linear model without regularization

Another interesting feature is the population group of males that are between 25-29. What makes it interesting is the change in correlation that it had when we changed the model type used. During our initial trial with the linear regression model without regularization, the model coefficients first told us this feature has a positive correlation with the death number. Yet, when we attempted the modeling with the Lasso regression model, there is now a negative correlation between the population group of males that are between 25-29 and the number of deaths.

Part of the coefficients of the Linear model without regularization



Coefficients of the Linear model with Lasso regularization



(ii) One feature that we initially believed would hold some degree of impact was the social mobility restrictions, such as limits on social gatherings, stay at home orders, and travel bans. We had believed that since these policies played a role in how those who were and weren't affected could interact with one another, it would thus also play a role in the number of deaths that could be recorded in each county. However, as seen from our EDA and modeling process, it was instead to the contrary. The coefficients that were obtained from the Lasso model in the final stages were all zeros for those particular features. Our reasoning for such a development is that the mobility restrictions have more of an indirect relationship with the death toll than a direct one, and what our model is attempting to locate are the features that have a direct relation. So for the mobility limitations, it more affects the number of observed cases, which in turn affects the number of deaths.

(iii) A major problem that we faced with our dataset was the large presence of missing values for many of the hospital-related columns. Some of these include the number of HPSAServedPop, HPSA Shortage, CU Beds, number of hospitals, and hospitals that are in the network. Most noticeably were the HPSAServedPop, HPSA Shortage, and HPSAUnderservedPop columns. These columns had a large number of missing values from numerous counties in the nation, and this missing data could be a source of bias or error in our final model. Even with using the most recent up to date data on the situation, there were still many missing values that would have been beneficial in helping model the threat of the virus. However, as the situation is still quite new to our scientists and ever-changing, there is little to be done regarding the amount of data available. Something only time and further investigation can help alleviate.

(iv) One major limitation of our work is that our analysis is purely based on the past data provided to us, which means our analysis does not have live updates that reflect the current situations of COVID-19 and thus may be easily outdated given the current rapid evolvement of the pandemic. We also made several key assumptions about the data: 1) no feature other than the ones included in the used datasets has

a significant impact on the death number (e.g. the virus mutated and there is a significant difference in the deadliness of the different strands; there is certain medical trait/life pattern/habit that dramatically increases/decreases the likelihood of dying from the virus shared by the entire group/neighborhood/county) 2) there is a lack of hospital-related data (e.g. number of certain hospital resources) and we assumed that these potential hospital-variables do not vary much across the counties 3) all of the data provided to us is accurate. We recognize that our analysis could be outdated/inaccurate if any of the above mentioned assumptions were proven to be wrong. For example, we believe there is a high chance that the death number is inaccurate for many counties since there might be people who died without getting tested (especially during the initial stage of the outbreak).

(v) An ethical concern that our group has is in regard to the collection of this data. This particular area isn't as much of a problem for us as these data are mainly logistical data regarding counties and hospital characteristics (i.e. does not involve much individual-identifying data). In addition, it has also been noted that these are data that are collected from public informational sources, sources that news sites like USAFacts and NY Times could access normally. However, this is with the assumption that the hospitalized individuals have given their consent in letting the hospital publish or to include them as part of the public COVID-19 database.

A more major concern that we possess is in the use of this data. Like the Pima Indian Diabetes Dataset, this one has the potential to be reused without explicit consent for completely unrelated projects/research. One example is by politicians in their election campaigns. Office runners who are using universal healthcare as part of their platforms could use these statistics for the betterment of their campaign credibility. But although this is beneficial to this particular politician and those who are also in support of attaining universal healthcare, this would be a violation of privacy and ownership if the people of whom the data is about did not give their consent to the politician for their data to be used in the campaign. This one of many possible examples in which this data could be used for other people's benefit without the required consent.

(vi) We believe data on the effectiveness (not just the date of implementation) of COVID-19 policies (e.g. stay-in-home, travel ba), other countries' COVID-19 data, as well as more data on hospital resources (e.g. the average number of ventilators) would strengthen our analysis and even potentially allow us to test out more detailed/specific hypothesis such as are some counties/countries low death rate due to the policies, abundant hospital resources or other (combinations) of factors.

(vii) One major ethical concern of our study might be the interpretation and use of our results. For example, if we were to find that a certain group with some shared medical characteristics/other traits are more vulnerable/immune to the current pandemic, it might lead to potential intentional/unintentional discrimination and related issues.

Future Work: Based on the limitations discussed above of our current work, one major area for improvement would be to obtain more data to examine the factors that influence the death toll in more detail and make sure our aforementioned assumptions are correct. Another area of major improvement can be to figure out a way to update our work with the current evolving situations. Perhaps a model that can automatically take in the live data and generate the most up-to-date predictions would be more useful in real-life applications.