SARS-CoV-2 U.S. Mortality Rate Prediction

By: Joyce Zheng & Cameron Hirsh

ABSTRACT

It is not an understatement to say that the entire planet has experienced the impact of the infamous COVID-19 outbreak of late 2019. The spread of this virus amongst global communities resulted in reverberations throughout nearly every sector of human society. The tangible global societal changes included but are not limited to: record levels of unemployment, closing and/or bankruptcy of massive corporations and small businesses alike, and a shortage of many household commodities. The overwhelmed healthcare systems struggled to manage the overwhelming number of ill people requiring medical treatment and hospital services. The virus infiltrates the respiratory system, causing a variety of symptoms and is extremely transmittable. This infection can lead to pneumonia and fatality in severe cases, especially when the host has any pre-existing medical conditions. We intend to use the data gathered thus far regarding the outbreak to gather insight that could be beneficial to the medical community.

As of May 2020, this pandemic continues to evolve and thus we are limited by the constant evolution of the virus as time progresses. The results we come to in this analysis can give direction to efforts regarding the response to this virus for the period we are investigating and into the near future. Further analysis will be necessary once dramatic changes to the dataset occur in the future.

INTRODUCTION

Amidst this healthcare crisis comes many questions and concerns amongst healthcare providers, governments, and families regarding the fatality and health risk the virus poses to those that contract it. As the virus spreads and the number of people contracting the virus increases, the death rate becomes more of a concern for many citizens. Given this, we ask: *what factors are the most effective in predicting the mortality rate of COVID-19?* To answer this question, we trained a model to predict the mortality rate of the virus as it spreads across U.S. counties using data collected about COVID-19. This insight could inform communities of the magnitude of deaths to be expected amongst their population and provide ease of mind to those that are feeling left wondering. The datasets we use in our calculations are gathered from the John Hopkins Center for System Science and Engineering repository as well as from the Yu group, based in Berkeley. These repositories are open for public use.

DESCRIPTION OF DATA

First, we took a look at the four imported dataframes. We used the github documentation as well as querying the columns for values and object types to figure out the meaning and significance of each dataframe and column. As an exploratory exercise, we created a visualization to compare the mortality magnitudes of different states over time using the `time_series_covid19_deaths_US` dataframe.

As an example, we used California, New York, New Jersey, and Massachusetts, though the states are customizable by the user.

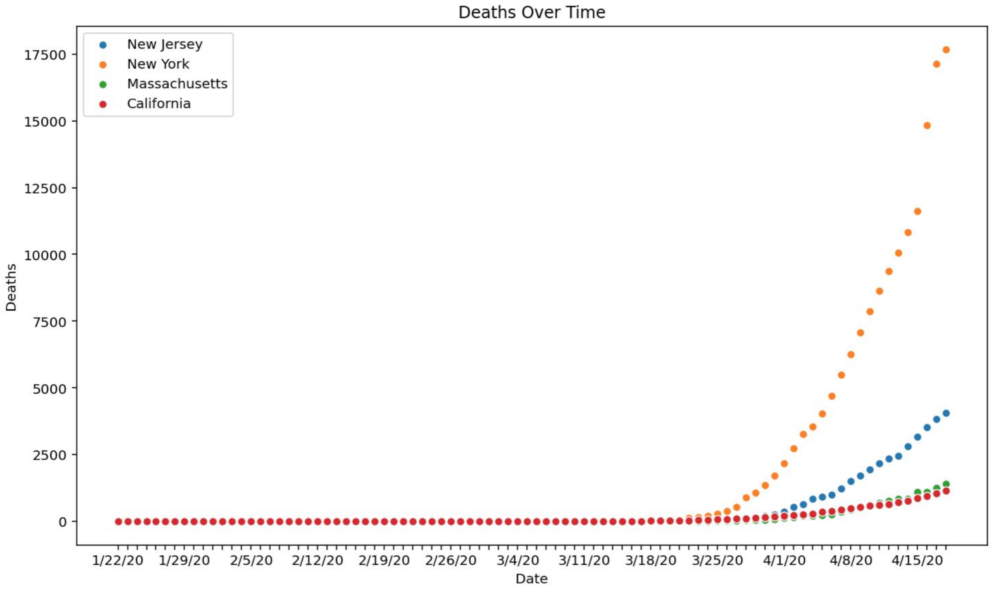
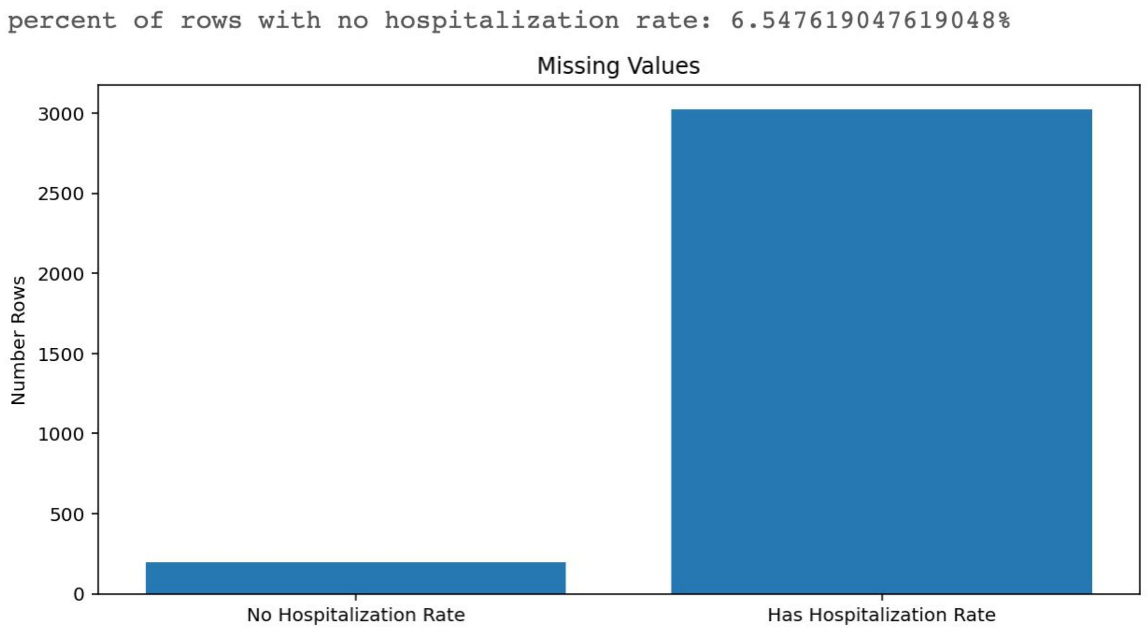


Figure 1
Number of deaths from 1/22/2020 - 4/18/2020 caused by COVID-19
for states New Jersey, New York, Massachusetts, and California

Realizing that we would need to combine the data from all the dataframes to generate an accurate predictor with sufficient features, we proceeded by merging the dataframes and proceeding to data analysis and cleaning of the resulting table.

During exploratory data analysis, we found that there were four columns with NaN values: ‘`stay at home`’, ‘`>500 gatherings`’, ‘`entertainment/gym`’, and ‘`Hospitalization_Rate`’. States Iowa, Nebraska, Arkansas, Oklahoma, South Dakota, Utah, and North Dakota had NaN values for the ‘stay at home’ column. After doing some research, we discovered that all of these states did not enact stay at home orders as of 4/18. We decided to set the ‘stay at home’ order date for these states to be 4/19/2020 and assumed that they did not implement such an order. We applied the same logic to the ‘`>500 gatherings`’ and ‘`entertainment/gym`’ columns, setting those NaN values to 4/19/2020 as well.

There were 198 rows where the ‘`Hospitalization_Rate`’ column had NaN values, where each row corresponded to a county.



```
percent of rows with no hospitalization rate: 6.547619047619048%
```

Figure 2
Bar plot displaying the number of rows with/without
values for the column ‘Hospitalization_Rate’

We decided to drop these rows because `‘Hospitalization_Rate’` data was only presented for states which provided cumulative hospital data; the states Nebraska, Indiana, and Nevada all did not provide such information and thus had NaN values. As a result, all three of these states were not represented in the dataset and models.

For numerical variable columns #3-22 (‘`lat`’ through ‘`TotalM.D.'s,TotNon-FedandFed2017`’) and #35-122 (‘`1/22/20`’ through ‘`4/18/20`’), we standardized their respective values.

DESCRIPTION OF METHODS

After our initial data analysis and cleaning, we proceed to creating our models for prediction. We decided to utilize linear regression to help answer our question. This method is appropriate because we believe that multiple factors must be considered in order to determine a county’s mortality rate from COVID-19. With linear regression, multiple features can be used to predict these mortality rates and the features can be compared by training various models and comparing their performances.

We used scikit learn to split the data into training and testing sets, using the training data to compare models composed of different features then choosing the best to use on the test data.
We created a total of four models to compare, each successive model adding new features that were not present in the previous model with model 4 containing the most features. After training and fitting the

models, we chose the one that performed the best and used it on the testing data. Out of curiosity, we also ran the other three models on the testing dataset to see how they would perform. We computed the root mean squared error (RMSE) of the predicted results to measure the spread.

SUMMARY OF RESULTS

Based on the results from our four models, we can come to many conclusions.

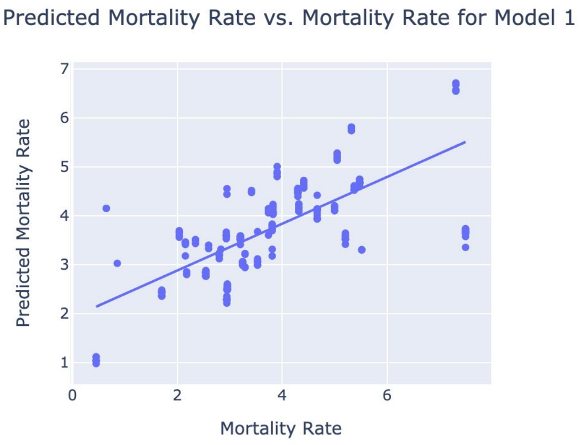


Figure 3
Scatter plot of the predicted mortality rate vs. actual mortality rate for model 1. Slope of OLS line is 0.4779

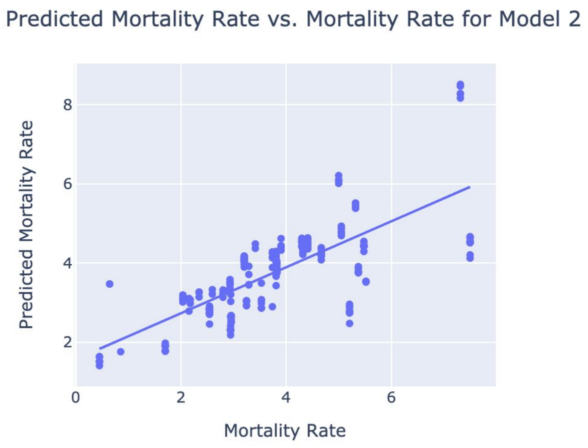


Figure 4
Scatter plot of the predicted mortality rate vs. actual mortality rate for model 2. Slope of OLS line is 0.5811

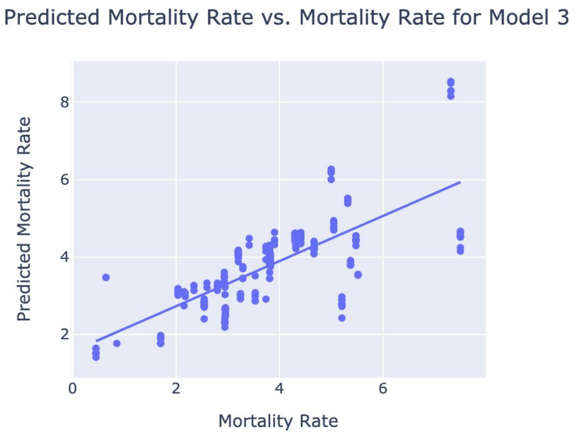


Figure 5
Scatter plot of the predicted mortality rate vs. actual mortality rate for model 3. Slope of OLS line is 0.5836

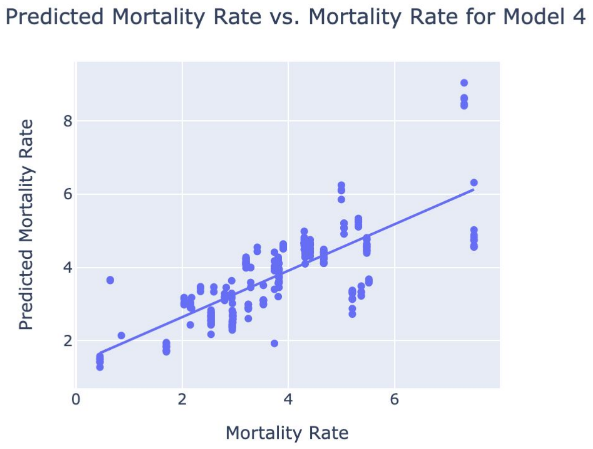


Figure 6
Scatter plot of the predicted mortality rate vs. actual mortality rate for model 4. Slope of OLS line is 0.6347

As the model complexity increases (model 1 → model 4), the scatter plot of the predicted vs. actual mortality rate becomes more linear. Data points become more clustered together. As a result, it appears that model 4 does the best at predicting mortality rates. However, we can see that our models are not perfect. We do see a positive linear relationship between the true mortality rates and predicted mortality rates. However, we do not see a line of slope 1, which would indicate the best possible relationship between the true mortality rates and predicted mortality rates. In terms of outliers, we see a couple of points with a true mortality rate greater than 6% that are relatively far off from the OLS line. They are consistently outliers in all four scatter plots.

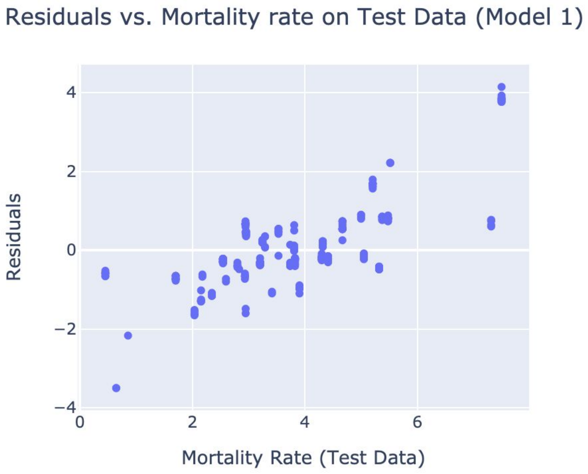


Figure 7
Plot of actual mortality rates against the residuals of model 1 for the test data

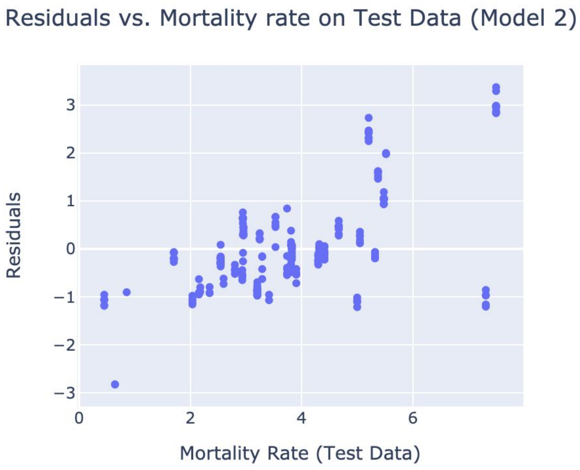


Figure 8
Plot of actual mortality rates against the residuals of model 2 for the test data

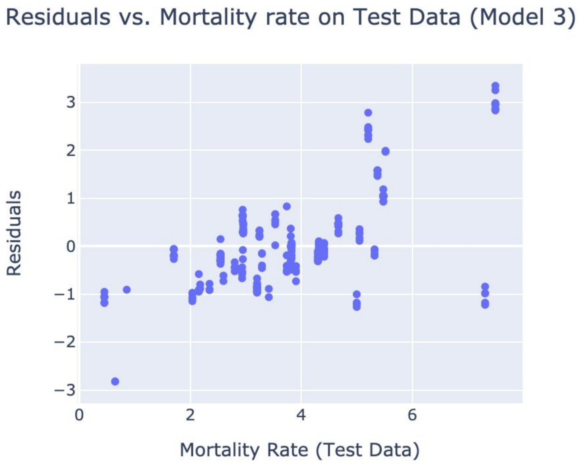


Figure 9
Plot of actual mortality rates against the residuals of model 3 for the test data

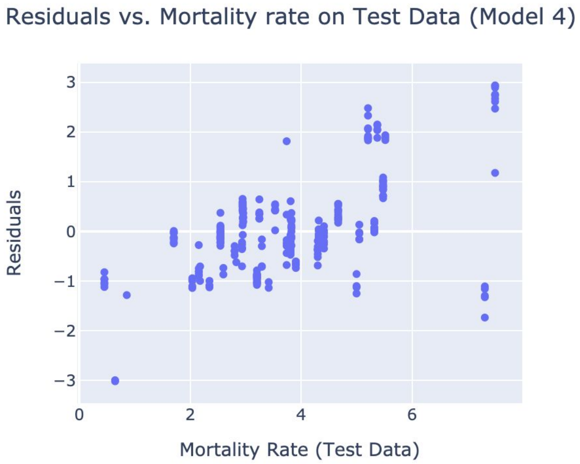


Figure 10
Plot of actual mortality rates against the residuals of model 4 for the test data

As the model complexity increases, the residual plots become less linear in shape and more equally/randomly spaced around the horizontal axis. Ideally, we want to see a horizontal line of points at 0 that would indicate a perfect prediction. The next best thing would be a homogenous set of points centered at 0. According to the residual plot from model 1, a nonlinear model would be more appropriate for the data. However, the residual plot from model 4 suggests that a linear model is appropriate and reasonable. Once again, model 4 appears to be the best at predicting mortality rates.

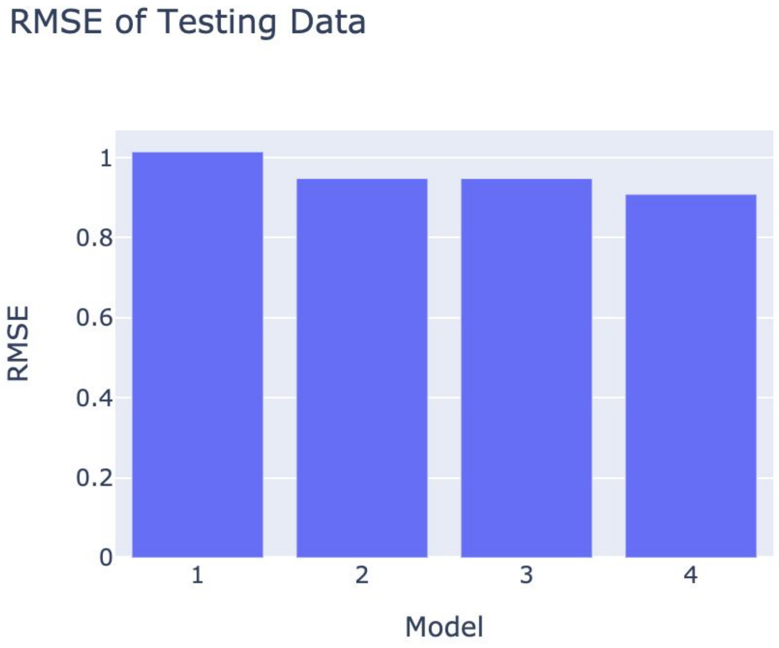


Figure 11
Root mean squared error (RMSE) of the predicted responses for each of the four models. Model 1: 1.015, Model 2: 0.9484, Model 3: 0.9482, Model 4: 0.9086

As the model complexity increases, the RMSE decreases. From model 1 to model 4, the RMSE decreased by 0.1064. As the model complexity increased, the residuals became less spread out. The testing data becomes more concentrated around the line of best fit. Since model 4 had the lowest RMSE error, it is the best model at predicting mortality rates.

According to model 4, the following features are useful in predicting mortality rates from COVID-19: 'lat', 'lon', 'PopulationEstimate2018', 'Rural-UrbanContinuumCode2013', 'PopulationDensityperSqMile2010', 'MedianAge2010', '#EligibleforMedicare2018', 'PopMale55-592010', 'PopFmle55-592010', 'PopMale60-642010', 'PopFmle60-642010', 'PopMale65-742010', 'PopFmle65-742010', 'PopMale75-842010', 'PopFmle75-842010', 'dem_to_rep_ratio', '#ICU_beds', '#Hospitals', '#FTEHospitalTotal2017', "TotalM.D.'s,TotNon-FedandFed2017", 'stay at home', '>500 gatherings', 'public schools', 'restaurant dine-in', 'entertainment/gym', 'foreign travel ban', 'Incident_Rate', 'Testing_Rate', ‘Hospitalization_Rate’, and number of deaths from 1/22/2020 - 4/18/2020 caused by COVID-19. Despite model 4 being our best model at predicting mortality rates of counties, it does have some flaws. It does not perfectly predict all of the test data set (as seen in the predicted vs. true mortality rates scatter plot), and only slightly drops the RMSE compared to models 2 and 3 (see bar chart above). Further modeling will have to be conducted to determine if there are other factors that are important in predicting mortality rates.

DISCUSSION

There were a few interesting features that we came across for our question. One feature was ‘Incident_Rate’. We knew that the confirmed number of cases per 100,000 persons was going to be important in determining mortality rate, but did not expect it to drop the RMSE by a large amount. Another feature we found interesting was ‘Rural-UrbanContinuumCode2013’. When we started constructing our models, we figured that the degree of urbanization would be important in determining how fast COVID-19 would spread in a county, which would thus affect confirmed cases and mortality rates. It was surprising to find that there was a column with such data in the given datasets.

One feature that we thought would be useful was ‘HPSAServedPop’. It is the estimated total population served by the full-time equivalent (FTE) health care practitioners within a (HPSA). Knowing how much of the population is accounted for by health practitioners sounded like it would improve our model, but the feature turned out to be ineffective. It increased our test RMSE for the first three models by almost 0.1 and the fourth model by 0.5. Due to this drastic increase in RMSE and decrease in prediction accuracy, we decided to remove this feature from our models.

We faced some challenges with our data. It was difficult to decide what to do with rows of the states Nebraska, Indiana, and Nevada that had NaN values for ‘Hospitalization_Rate’, one of the features that was in our model. We had two options: fill in these null values or remove all these rows from our dataset. After careful analysis, we realized that these missing values were not random; any row of these three states had a NaN value for ‘Hospitalization_Rate’. We did not have any data from these three states that would have enabled us to calculate, for example, a mean hospitalization rate that we could fill in these NaN values with. We thus deemed it safe to remove these rows from the dataset. Another challenge that we faced was with NaN values in the columns ‘stay at home’, ‘>500 gatherings’, and ‘entertainment/gym’. Null values in these columns were due to such orders never being implemented in certain states. We could only assume that these states did not implement such an order and assigned the date to be 4/19/2020. It did not seem reasonable to remove such a large chunk of the data from the dataset.

Our analysis was limited in that the data we worked with went up to 4/18/2020. COVID-19 numbers have gotten worse since then. States that previously did not have stay at home or >500 gatherings orders may have enacted such orders since 4/18. Having access to this updated data potentially would have enabled us to create a better predictor and strengthen our analysis. There is a reason why these states did not enact orders early on, and our model cannot reflect them. Another limitation is the lack of data for Nebraska, Indiana, and Nevada. These states were not represented in our model, so we do not know whether or not their data would have helped our model prediction accuracy. Having additional data for these three states would have enabled us to include them in our models and hopefully would have strengthened our analysis. Predicting mortality rates for counties from these states may not be very accurate. Some assumptions that we made include the relative importance of age groups. We assumed that young ages would not greatly affect the mortality rate while older ages would and thus included population sizes of older age groups in our models. This could be proven incorrect, as younger age groups are more likely to be carriers of COVID-19 and ongoing research continues to raise concerns about susceptibility of younger age groups to the virus.

Overall, our approach was somewhat effective. We were able to relatively accurately predict mortality rates for counties. Adding more features enabled us to generate a better model. However, requiring the usage of such a large number of features to create only a somewhat effective predictor does raise some concerns. It is possible that we were selecting for features that did not greatly affect mortality rate. Features that are important in predicting mortality rate may not have been provided to us in the given datasets. Another method/modeling type might be more suitable to help us answer our question.

It was surprising to see how much the ‘`Incident_Rate`’ feature affected our model. Adding more features that were related to COVID-19 statistics (i.e. ‘`Hospitalization_Rate`’ and ‘`Testing Rate`’) greatly improved our models compared to other features. This suggests that in future work, it would be useful to focus more on COVID-19 data such as transmission rates to create a better model. Further work should also take into consideration better ways to represent the data in the columns. In our work, we standardized numerical variable columns #3-22 (‘`lat`’ through ‘`TotalM.D.'s,TotNon-FedandFed2017`’) and #35-122 (‘`1/22/20`’ through ‘`4/18/20`’). Looking back, we should have converted timestamp columns (i.e. ‘`stay at home`’ and ‘`>500 gatherings`’) so that these columns, for example, indicated how many days it has been since 1/22/2020. This would have enabled us to standardize these columns, which hopefully would have helped improve our model predictions. The values that are currently in these columns are relatively large compared to values in the other columns (huge difference in range of values).

Data scientists should always have the ethicality of the data collected in mind when exploring datasets. In our case, we must recognize that the deaths we are analyzing and modeling were the lives of real people and treat them as such. The purpose of our particular experiment was to answer a question we had about the data, but we must always put into ethicality the purpose of our work. If the purpose were to, for example, decide which counties should be receiving more aid/resources depending on risk factors, this could be a questionable process. Everyone is equal and the health, safety, and privacy of the people the data is representing should always be the main concern when proceeding with a data analysis process.