Aditya Shah

May 8th 2020

Data 100

COVID-19 Data Project Final Report

Abstract:

COVID-19 has affected the lives of millions of Americans since its first occurrence around January 22nd 2020. The purpose of this study is to explore the various factors that affect the Mortality Rate due to COVID-19 in the United States. Mortality Rate is defined as the number of deaths per 1,000 individuals. Through Exploratory Data Analysis (EDA), it is concluded that Number of Hospitalizations is positively associated with Mortality Rate, naturally due to merely a higher number of cases in a particular state. One less apparent association is between the Undeserved Population, and the mortality rate, and there is a positive association between these variables. Furthermore, contrary to the initial claim that Mortality Rates of other disorders, including Heart Disease, Respiratory Disorders, and Stroke, are associated with overall statewide Mortality Rate, there is, in fact, no correlation. According to the data, the instance of these disorders is independent of the statewide mortality rate. To predict Mortality Rate, I built a Linear Regression Model, Regression Tree Model, and a Random Forest of Regression Trees Model(To capture the variance of overfitting regression trees effectively). To evaluate the performance of these models, 60% training, and 40% testing split as well as bootstrapped cross-validation accuracy were employed to capture the variability of samples for a relatively small dataset. In short, the Random Forest of Regression Trees performed the best with Mean Squared Error(MSE) of 0.28805 on the test set followed by the Regression Tree Model with MSE of 0.34619, and Linear Regression Model with MSE of 0.429169.

Introduction:

The spread of COVID-19, also known as the Coronavirus, has become a deadly global pandemic that has affected the lives of billions in just the past three months. According to CNN, the number of cases in the United States has surpassed 75,054 as of May 7th 2020. Given this dire scenario, I decided to further explore the factors that affect the mortality rate of Coronavirus in the United States. Mortality Rate is defined as the number of deaths per 1,000 individuals, and I initially assume that States with higher mortality rates have higher populations on average. Specifically, the focus question I would like to answer using Data Science inference techniques is, what factors on a

state by state basis are associated with Mortality Rate? Although one cannot necessarily conclude causation between these factors and the Mortality Rate, it is most appropriate to conduct research on factors that will serve as proper predictors in model development.

One such factor is age. The older an individual is, the more likely they are susceptible to catching the Coronavirus. According to STAT, "But the fatality rate was 14.8% in people 80 or older, likely reflecting the presence of other diseases, a weaker immune system, or simply worse overall health."(Begley 1). In other words, the older an individual is on average, they are frailer with regards to the susceptibility to multiple disorders. Thus, potentially, the population of older individuals, specifically Male and Females older than 75 years of age in a given state, would be essential features in predicting mortality rate. Additionally, children are at relatively row risk for Coronavirus: "Even cases among children and teens aged 10 to 19 are rare. As of February 11th, there were 549 cases in that age group, 1.2% of the total, China CDC found. Only one had died"(Begley 1). This means that using a feature which represents the number of Males and Females who are children will likely effectively underscore the differences between states with higher mortality rates and lower mortality rates, and a potential inverse correlation.

Another factor is underlying medical conditions such as Heart Disease, Instance of Strokes, and Respiratory Issues. STAT news website discusses that Men are more likely to have cardiovascular disease, and those with cardiovascular disease are more likely to experience the severe symptoms of the virus. Particularly, they concede that "People with pre-existing illness are more likely to get seriously ill from Covid-19, and men have a higher incidence of such chronic illnesses as cardiovascular disease"(1). With regards to other health conditions, STAT indicates overall that individuals with any pre-existing conditions such as diabetes, cancer, respiratory issues, etc., as well as smokers, have a significantly higher chance of developing symptoms that require hospitalization. Higher rates of hospitalization, as well as the number of individuals hospitalized in a given state, are correlated with mortality rate, especially for those with severe symptoms. Begley makes the research based claim: "After taking into account the patients' ages and smoking status, the researchers found that the 399 patients with at least one additional disease (including cardiovascular diseases, diabetes, hepatitis B, chronic obstructive pulmonary disease, chronic kidney diseases, and cancer) had a 79% greater chance of requiring intensive care."(1).

The factor of Incidence Rate, which is defined as according to Investopedia as "Incidence rate or "incidence" is numerically defined as the number of new cases of a disease within a time period, as a proportion of the number of people at risk for the disease"(Hargrave 1). Particularly, Incidence Rate is quite similar to the mortality rate; however, it is a proportion of the number of people at risk for the

disease, rather than the proportion of the total population. In the numerator of both the mortality rate and the incidence rate is the number of cases in a given time period. This means that Incidence Rate is a good predictor of mortality rate. The Under deserved feature is also an important one, as individuals below the poverty line are more likely to catch the virus because of living conditions.

Description of Data/Data Cleaning

I utilized the data from four tables: the first table(4.18states.csv) contains information about mortality rate, number of hospitalized, number of cases, and number of deaths etc. for each of the 50 U.S. states; the second table(abridgedcounties.csv) contains information for each county regarding male/female population, mortality rates for heart disease, diabetes percentage, smokers etc.; the third table(time_series_confirmed_cases_us.csv) contains time series data representing the spread of Coronavirus over time for each county in the United States with regards to confirmed cases; the fourth table(time_series_covid19_deaths_us.csv) contains time series data representing the spread of Coronavirus with regards to deaths.

I imputed null values using the median values for each of the columns, and created a standardized copy for each table and non-standardized(used for data visualization), for each table. I intended to predict the mortality rate for each of the states in the 4.18states.csv table, but I also wanted information where each record in the final joined table represented a state. I then aggregated the time_series_covid19_deaths_us.csv table as I extrapolated that the time series data(specifically relating to the death toll over time), would be important to demonstrate the exponential growth in number of cases due to the virus in a data visualization. I did the same for the time_series_covid19_confirmed_cases_us.csv table and joined it with the other table. Additionally, the abridgedcounties.csv contained data that served as good predictors of mortality rate(male/female population, mortality rates for heart disease, diabetes percentage, smokers etc., under deserved population), so I selected these columns of interest. I then joined these tables on two primary keys: the state name, and county name, and then grouped by state name, and aggregated the data using the mean. Finally, I joined this merged table with the 4.18states.csv table which has the crucial mortality rate label, with the other three merged tables.

Exploratory Data Analysis (EDA) / Description of Methods (Modeling Phase)

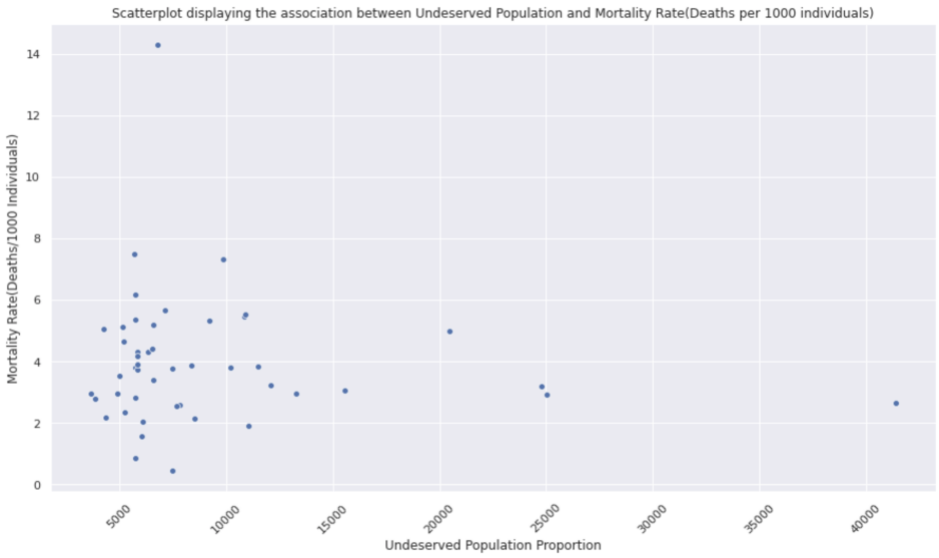

Figure 1. Scatterplot displaying the association between Underserved Population Proportion(explanatory variable) and Mortality Rate (response variable)

As the undeserved population proportion increases, the mortality rate increases as well. Although the association may appear to be fairly weak, it is important to note that for undeserved population proportion values larger than around 5000, not a single state had a mortality rate lower than 2 deaths per thousand.

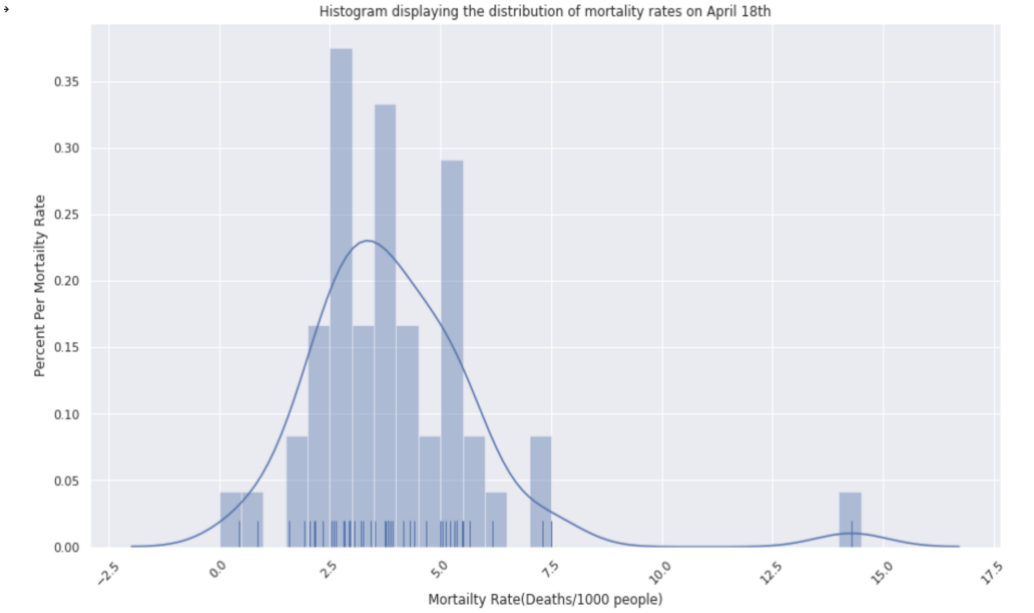

Figure 2. Histogram displaying the mortality rates of all the U.S. states and U.S. territories on April 18th 2020.

The histogram appears to be bimodal, and skewed to the right (skewed towards higher mortality rates). The histogram tells me that there is likely an outlier with a mortality rate of around 15 deaths/1000 people. I would assume that this corresponds to a state/territory with an especially large population density. However, most of the states/territories have morality rates between around 1.5 deaths/1000 people and 5 deaths/1000 people.

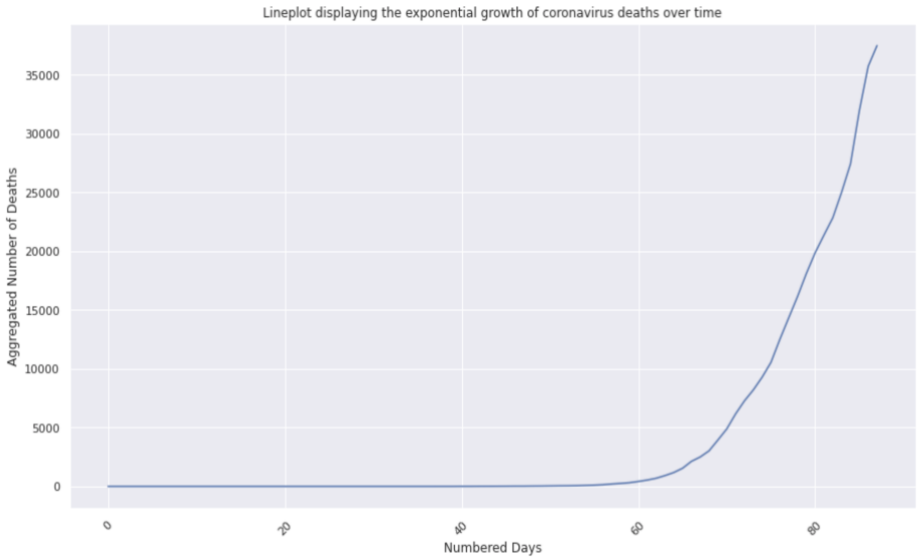

Figure 3. Lineplot displaying the cumulative distribution function of the aggregated number of deaths over time.

Around day 60 is when the number of deaths begins to grow exponentially, sometime around early March. This visualization indicates that we are still currently in the exponential growth portion of the logistic carrying capacity curve of the number of deaths due to Covid-19. The deaths curve has not leveled off as of 4/18/2020.

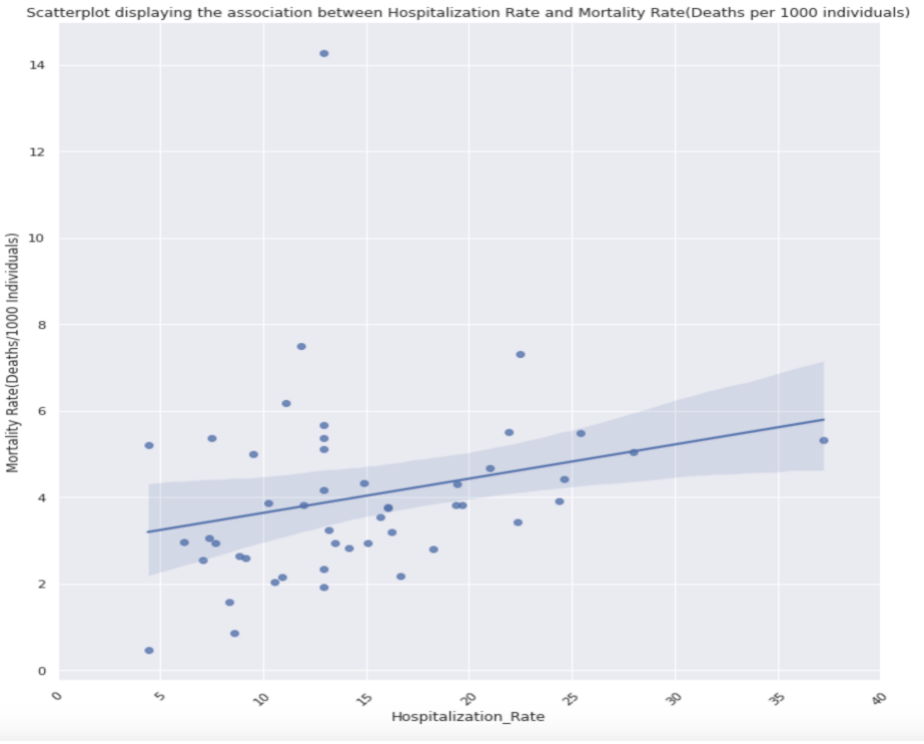

Figure 4. Scatterplot displaying the association between Hopspitalization Rate (explanatory variable) and Mortality Rate(response variable)

Visually, it can be observed that for the given data that we have, as the Hospitalization Rate increases, the mortality rate increases as well. This makes sense since, the more confirmed cases there are in a given state, the more are hospitalized, and the greater the mortality rate. While we cannot conclude causation between these two variables, they do appear to be associated, so it is likely that the Undeserved population is likely to be a good predictor of the mortality rate.

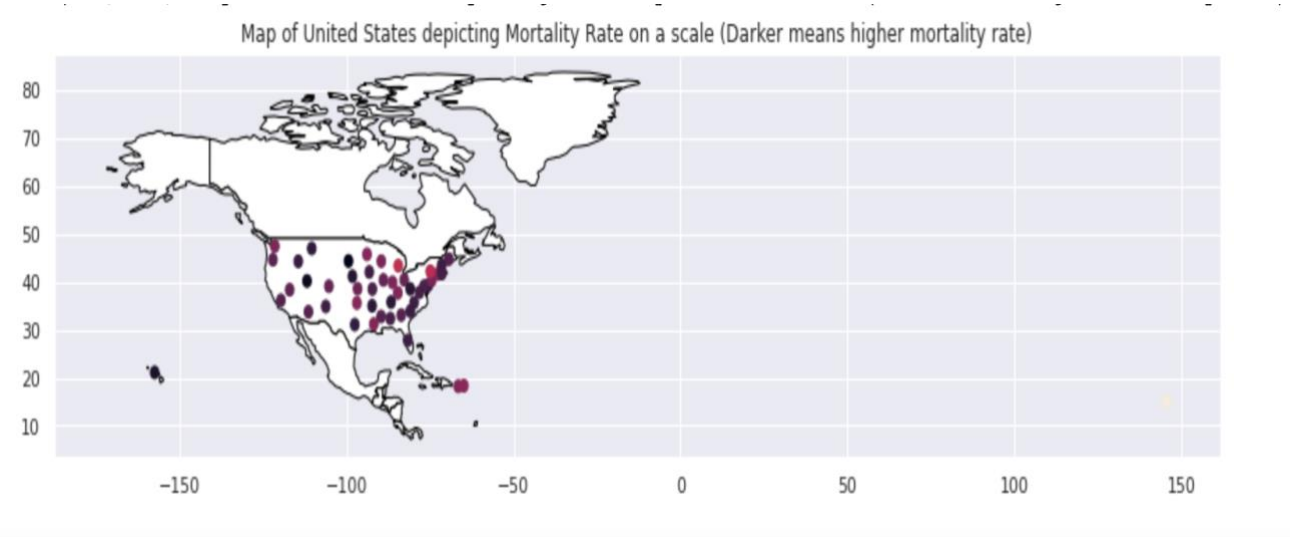

Figure 5. Plot of North America displaying the magnitude of the mortality rate among all States in the U.S. Visually, it can be observed that the darker the color(black), the higher the mortality rate. This may be contrary to what is expected, as we would expect midwest states with less population density to have lower mortality rates. This is compared to states on the coasts which are much more populous. However, mortality rate is agnostic of population, as larger states have more cases, but smaller less populous states have fewer cases but high mortality rate.

After Exploratory data analysis, I decided to build a Linear Regression Model, Regression Tree Model, and Random Forest of Regression Trees Model. I defined a mean squared error function, as well as a cross validation accuracy bootstrapped function. Since the dataset is quite small representing the U.S. states and territories, bootstrapping the sample each time before fitting the model to the training data and evaluating the model on the validation error ensured that the models would capture errors generated by natural sampling variability. As part of model development, I bootstrapped each model thirty times in the cross validation bootstrapped functions to achieve a good amount of natural sampling variability. I used a 60% training and 40% testing split. Initially, I selected a set of 16 features, 4 of which corresponded to populations of 5 year old and 10 year old males and females, the other 4 corresponding to populations of male and female 65 year olds and 75 year olds respectively, percentages of smokers and diabetes, mortality rate of heart disease and strokes, as well as mortality

rate. These features yielded me accuracies of as part of cross validation accuracy of 2.549, 0.414, and 0.265 on the Linear Regression Model, Regression Tree Model, and Random Forest of Regression Trees Model respectively. I then created a pairplot to visually depict the associations between each of these variables to narrow down on only the associated features. I discovered that heart disease mortality, stroke mortality and resp mortality rates features were not significantly correlated with the mortality rate. Furthermore, I realized that the populations of particular age group features appear to be redundant. Thus, I decided to only include the male and female populations under age 5 as well as those above 75 years of age to underscore the differences in mortality associated with differences in age.

After, I created another pairplot including features closely related to mortality rate. These features are Long, Confirmed, Deaths, Incident_Rate, and People_Tested features which are associated with Mortality_Rate. Amongst the associated features, Long_, People_Tested, and Incident_Rate are the most strongly associated, so these features will be added to the original list of features. Thus I narrowed down on these features: Mortality_Rate, PopMale5-92010, PopFmle5-92010, PopMale75-842010, PopFmle75-842010 , Diabetes Percentage , Smokers_Percentage , Long_ , Hospitalization_Rate , People_Tested , and the Incident_Rate. Using these features yielded a 67.05% reduction in the cross validation error for the Linear Regression Model, 18.46% reduction in error for the Regression Tree Model, and 14.31% reduction in error for the Random Forest of Regression Trees Model.In conclusion, the Linear Regression Model, Regression Tree Model, and the Random Forest of Regression Trees Model acheived low mean squared errors of 0.34619, 0.429169, 0.28805 respectivel y on the test set. This is for predicting the mortality rate of a given state in the United States.

Seven Specific Questions about the Project Answers:

1. The most interesting features I felt were the incident rate(which is a transformed mortality rate), and the longitude (I didn't expect that it would be a useful feature so that surprised me).
2. I thought the Heart Disease Mortality Rate would be a useful feature, however it turned out to not be associated with the Mortality Rate.
3. One of the biggest challenges with the data was dealing with the null values in each column. I finally built a function to replace the null values with the mean of the respective column if the distribution of non-null values was roughly normal, else replace it with the median.
4. One of the limitations is the scope of my analysis. For the purposes of understanding whether certain factors are significant predictors of the mortality rate this is appropriate, but since my analysis was limited to the United States, it may not generalize well to different countries.

5. As I did not consider ethnicity as a potential predictor of mortality rate in my analysis, it is possible that my model may not have captured all the data necessary in regards to such predictors. I decided not to include such data as it may not be equally representative.
6. Exact same data but from other countries, would make these models better predictors.
7. When examining if race/ethnicity is a predictor of mortality rate, generating such insights may raise ethical concerns, as it is dealing with the sensitive topic of race. I would need to ensure that the data represent all social classes, and races equally for a given geographical region as much as possible.

Bibliography:

Begley, Sharon, et al. “Who Is Getting Sick? A Look at Coronavirus Risk by Age, Gender, and More.” *STAT*, 10 Mar. 2020, www.statnews.com/2020/03/03/who-is-getting-sick-and-how-sick-a-breakdown-of-coronavirus-risk-by-demographic-factors/.

Hargrave, Marshall. “What Does the Incidence Rate Measure?” *Investopedia*, Investopedia, 6 Feb. 2020, www.investopedia.com/terms/i/incidence-rate.asp.

Hayes, Mike, et al. “US Coronavirus Update: Latest on Cases, Deaths and Reopening.” *CNN*, Cable News Network, 7 May 2020, www.cnn.com/uslive-news/us-coronavirus_update-05-07-20/index.html.