**Developing a Lasso Regression Model to Predict COVID-19 Mortality Rate Given the Policy Implementation Methods and Landscape Features of an American State**

Nicole E. Linden

UC Berkeley

Data 100: Principles and Techniques of Data Science

Professor Ani Adhikari & Joey Gonzalez

May 13, 2020

**Abstract**

Covid-19 is a highly contagious disease causing potentially fatal respiratory illness and, at this time, has no cure. Over 200,000 people have died from the virus and this number is only projected to increase. This study aims to identity influential factors in the spread of Covid-19 in the United States by training a lasso regression model to extract the most informative features of statewide data involving population statistics, environment features, and policy implementation. Through exploratory data analysis of county conditions and statewide Covid-19 statistics, a feature set was created, and a lasso regression model was developed to reduce overfitting and minimize the root-mean-squared error. Results show that the model was able to predict a state’s overall mortality rate, with small error, based heavily on the number of hospitals in a state’s network, the number of people tested, and the timeline of when both a ‘Stay at home’ order and when restrictions on businesses related to entertainment were introduced.

**Introduction**

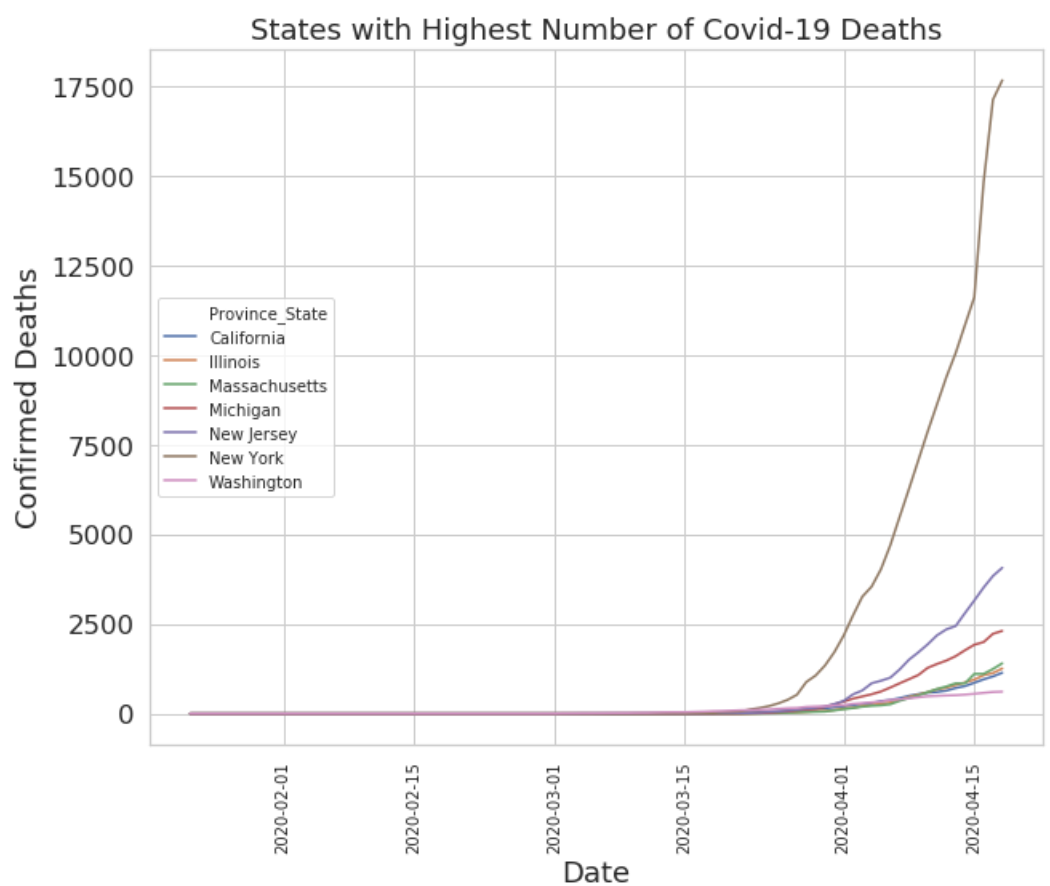
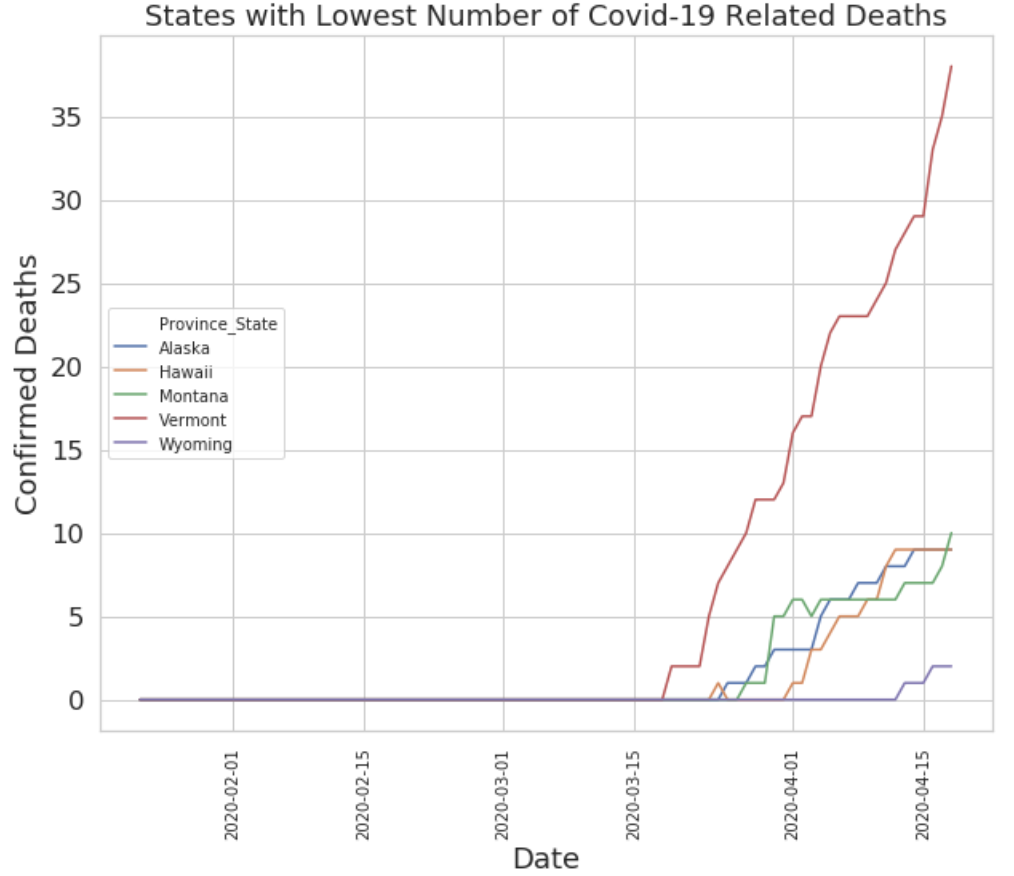
On December 31, 2019, a case of pneumonia with unknown origin was first detected in Wuhan, China and, by January 30 of 2020, it evolved into an international outbreak and Public Health Emergency. This highly infectious disease, named Covid-19, has had over 4.18 million confirmed cases worldwide and continues to spread at a rapid pace despite extreme protective measures put into place by governments around the world.[[1]](#footnote-1) In this paper, I aim to explore patterns related to the outbreak in the United States to test whether a state’s overall Covid-19 mortality rate has greater correlation with the implementation of policy or with environmental factors. In other words, this study aims to determine whether a state’s mortality curve was a product of government policy decisions (“Stay at Home” orders, restrictions on non-essential businesses, etc.) or a state’s unique landscape features (spread of population, living conditions, public transportation options, hospital resources, etc.).

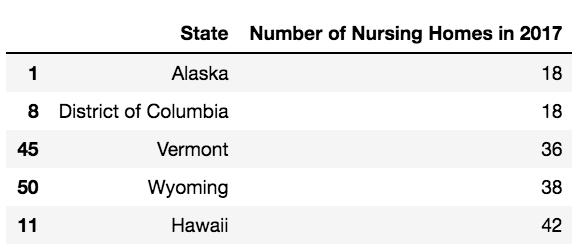
**Description of Data**

To evaluate the hypotheses presented in this paper, I will be merging datasets given to me by UC Berkeley’s Data 100 Course Staff[[2]](#footnote-2), including the Covid-19 daily report from April 18, 2020[[3]](#footnote-3) and American county data from the Yu Group’s abridged\_couties.csv. The Covid-19 daily report file contains cumulative counts for the United States up until April 18, 2020 of the total numbers of confirmed, recovered, and active state-wide Covid-19 cases as well as the number of Covid-19 related deaths. Additionally, it provides the statewide Covid-19 incident rate, testing rate, hospitalization and mortality rate-- metrics defined here.[[4]](#footnote-4) This dataset was merged with county information obtained from the Yu Group’s Covid-19 Severity Prediction file[[5]](#footnote-5), grouped by state in order to combine county data and ultimately analyze the state’s overall landscape and policy implementation decisions. To account for the variance in total number of counties within each state, I aggregated the numerical county data by state-wide averages. Columns that address policy-implementation (i.e. ‘stay at home’, ‘>50 gatherings’, etc.) contain the date that the policy was implemented statewide. All dates contained within the constructed training dataset are Gregorian ordinal timestamps. In addition, to address another state landscape feature I included the statewide total of certified nursing facilities in 2017 found at the Kaiser Family Foundation’s website.[[6]](#footnote-6) This feature was included in response to an observation I made in my home county. The hospital nearest to me has a statistically high incidence rate of Covid-19 and a feature of this community is that it has a statistically high number of certified nursing homes, which many have theorized could be influential to the area’s spread of the disease. To address this idea, I included the statewide count of certified nursing facilities and will see if developed models extract this feature as informative.

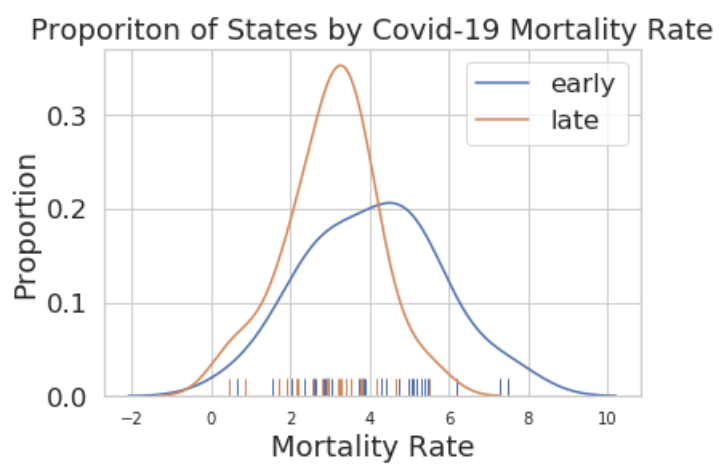
**Methods**

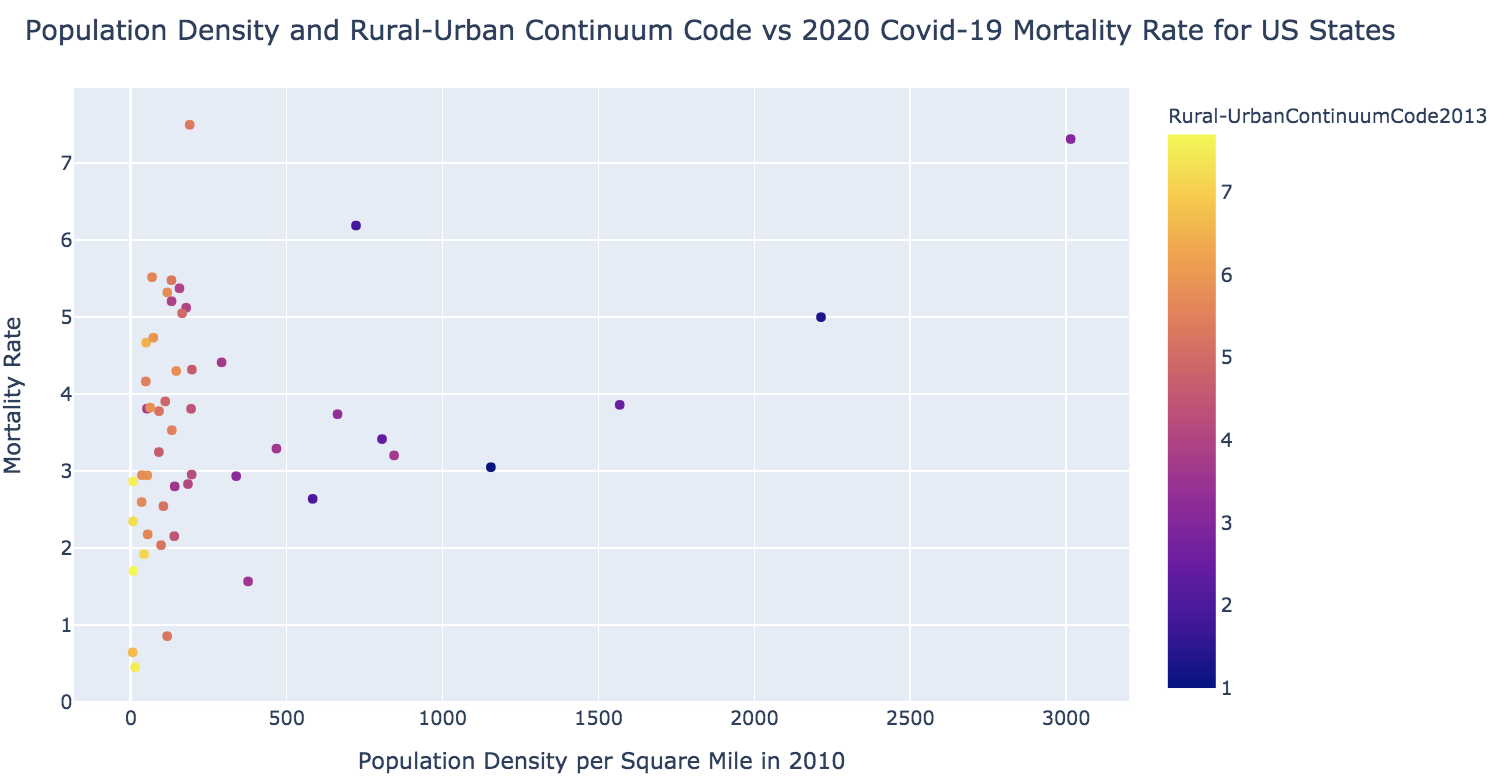
In my exploratory data analysis process, I began by addressing null or missing values in the datasets. I visualized each column’s count of null values and determined how to address these gaps by manually inspecting the column’s datatype, NaN value, and the original description of information it contains. For columns with over 30% of records missing, I dropped the feature entirely to avoid a high risk of skewing my model with an arbitrarily chosen filler variable. In other cases, I determined an appropriate value to replace the NaN value with. For example, in policy-related columns, a null value meant that the state or county had not implemented the designated policy before the date I accessed the dataset. To address this without dropping the record, I replaced the null with python’s datetime “now” timestamp to inform my model that at any point in the past, the state or county had not implemented the policy. Other null values were similarly addressed in a case by case basis found in this report’s corresponding Jupyter notebook. Other data cleaning methods I used included dropping both redundant features (i.e. FIPS codes, state abbreviations, etc.) and those that did not address the hypotheses I was interested in testing.

To determine an initial idea of Covid-19 cases by state, I visualized the curve of confirmed Covid-19 cases for the states with the highest and lowest counts:  



Interestingly, 4 out of 5 of the states with the lowest number of Covid-19 case numbers were also the states with the lowest number of certified nursing facilities, reinforcing my intuition to include this feature in the training data.

I chose “Mortality Rate” as the variable for my model to ultimately predict because this metric will be easier to compare between states of different size and population. Mortality rate is defined statewide as the number of recorded deaths multiplied by 100 and divided by the total number of confirmed cases. To begin testing my hypothesis on whether policy implementation or landscape had a greater effect on the state-wide Covid-19 mortality rate, I categorized states as early or late in comparison to the median date American states chose to implement the ‘Stay at Home’ order and visualized the distribution of their mortality rates. Interestingly enough, the distribution of states who implemented this order earlier had a higher average mortality rate, demonstrating a negative correlation between the two variables. This finding corroborates the theory that states could have other, more influential, features that correlate to the increase in Covid-19 related deaths. So, next, I decided to look at whether there exists a potential correlation between state's categorization on the rural-urban continuum and their overall Covid-19 mortality rate.



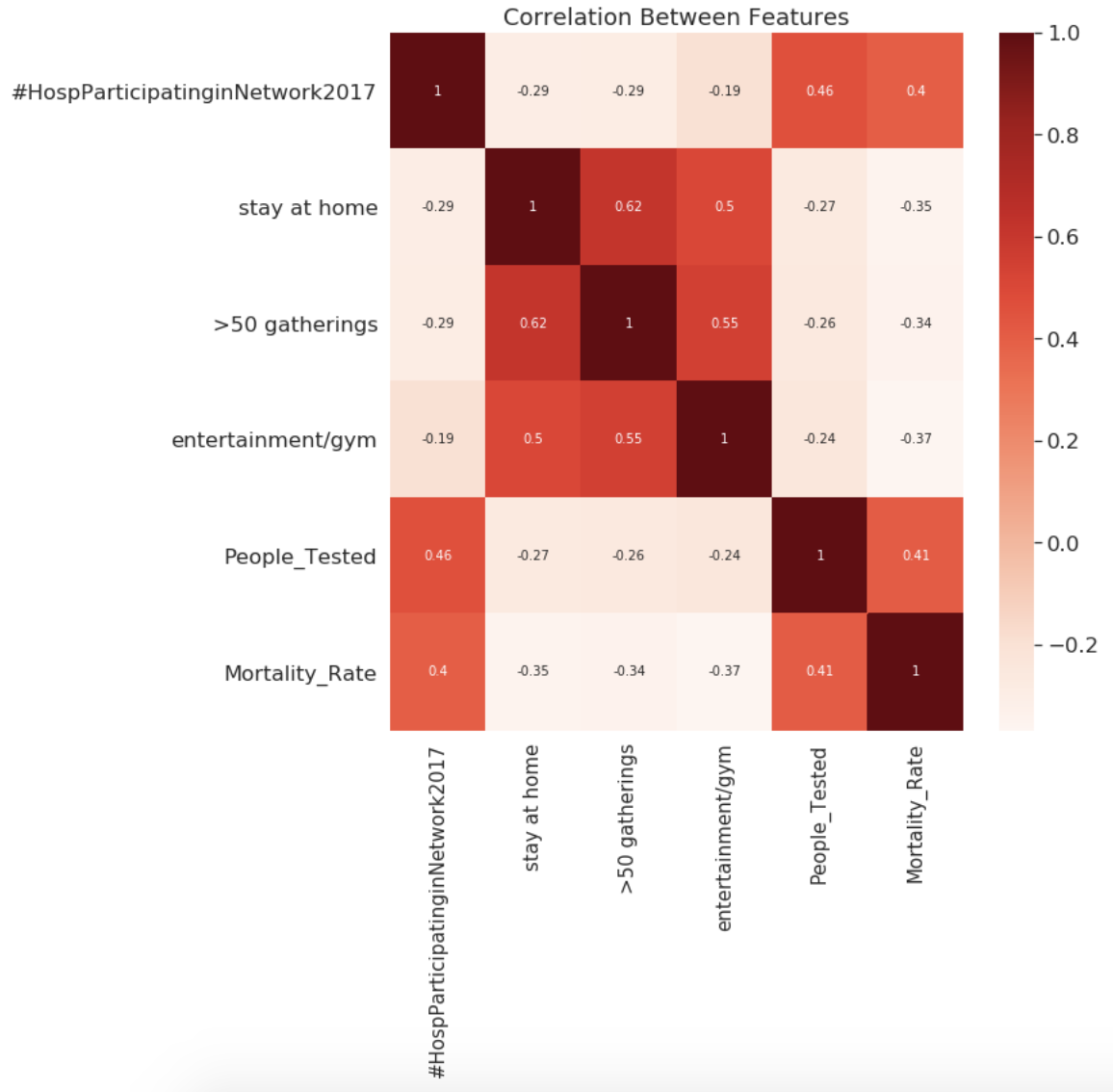
The visualization above depicts a scatterplot of states with their population density per square mile on the horizontal axis versus their overall Covid-19 Mortality Rate on the vertical axis. The color of each point depicts a state’s average Rural-Urban Continuum Code (the mean of all county codes within a state), a classification scheme that distinguishes counties by the population size of their metro and non-metro areas by ‘degree of urbanization’. This plot shows a direct relationship between the Rural-Urban Continuum Code and population density (states that are less dense on average have a similarly low continuum code). Additionally, this plot indicates a slightly positive correlation between the population density and mortality rate. In comparison, in the plot below, I created a similar graph visualizing the date the state’s ‘Stay at Home’ order and was implemented with points representing each state and colored by the state’s overall incident rate of Covid-19. There is not as clear of a relationship between the three variables-- states who implemented this policy earlier versus later do not have a distinct difference in distribution of mortality rate or incident rate. However, it does show that most states implemented the state at home order within a similar time frame so the data to conclude the effects of introducing a ‘stay at home’ policy later than most could be too limited to draw any concrete conclusions from.

A screenshot of a computer

Description automatically generated

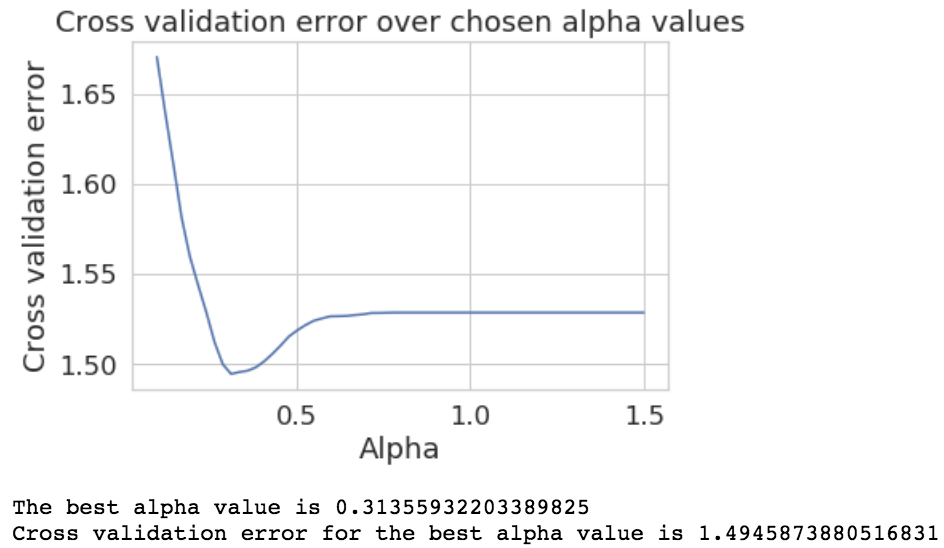
After this initial exploratory data analysis and motivated data cleaning, I chose feature columns with data related to the state’s overall landscape and to the state’s policy implementation while designating my prediction variable to be the state’s Covid-19 mortality rate. I then randomly split my data into a training set containing 90% of records to train my model and a test set with the remaining 10%. I chose this split because I only have 50 records of state data, so I prioritized a more diverse amount of training data to help build a stronger model. Next, I normalized my training data by standardizing each column (transforming each column Series so that it had a mean of 0 and a Standard Deviation of 1) to aid comparison and protect my model from placing greater importance on columns with larger values.

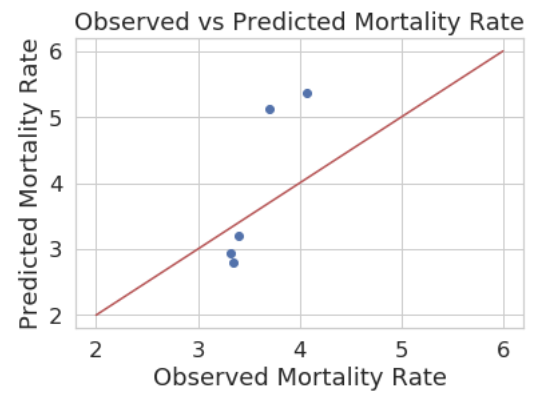
Because the variable I wanted my model to predict is continuous, I tested different regression models using Scikit Learn’s linear\_model module. For my loss function, I defined a root-mean squared error function and to select my final model, I performed cross validation with each model to simulate its performance on unseen data and test for potential overfitting. After these comparisons, I found that Scikit Learn’s Lasso Regression had the best performance because it addresses model bias, reduces overfitting, and performs feature selection. To determine the optimal alpha value, the regularization hyper-parameter, I performed cross-validation with a range of different alphas and found which one minimized the average root mean squared error. After setting the regularization hyper-parameter to its optimal 0.3135, my Lasso model extracted the ‘#HospParticipatinginNetwork2017', 'stay at home', '>50 gatherings', 'entertainment/gym', and 'People\_Tested' features as most informative for the prediction. I have included a correlation feature heatmap to visualize the approximate correlation between the features my final model extracted the state’s Covid-19 Mortality Rate.



**Summary of Results**

I have developed a regression model that takes in a set of landscape and policy-related features and outputs a state’s Covid-19 ‘Mortality Rate’ prediction, a continuous variable defined as the number of deaths multiplied by 100 and divided by the total number of confirmed cases. My final model was chosen based on its ability to minimize the root-mean-squared error loss function, which measures the difference between observed and predicted values. To avoid overfitting to training data, I performed cross-validation (to simulate testing the model on unseen data) and found an optimal regularization hyper-parameter (alpha=0.3236) that minimizes the model’s tendency to overfit. My final model extracted the recorded number of hospitals participating in its network in 2017, the number of people tested, and the dates when a state implemented both “Stay at Home” orders and entertainment/gym restrictions as the most informative features. Finally, to visualize my model’s performance on the test set, I plotted the observed mortality of our test dataset versus the model’s predicted value with an overlaid identity line to get a better visual of the residuals (the difference between the observed and predicted value). The model performed slightly better on states with lower mortality rate but, due to a lack of substantial data, this conclusion cannot be made confidently and requires further study.





**Discussion**

Some of the most interesting features I came across in the analysis of the data and the development of my model was the relationship between policies (the ‘Stay at Home’ order and restrictions on entertainment businesses), and a state’s overall Covid-19 mortality rate. I found that these features were negatively correlated which contrasted with my initial intuition. In other words, states that implemented these restrictions earlier tend to have, on average, higher mortality rates. I noticed this to be the case in my initial exploratory data analysis which led me to believe that a state’s landscape features (which I consider to be population density, classification as rural or urban, etc.) would be extracted as the most informative features in my final model. However, my model still extracted the dates when these policies were put into effect but assigned these features negative coefficients. This leads me to believe that the most informative features for a state’s coronavirus outbreak (features that prompted state government to quickly enforce strict policies) was not specifically included in my training dataset. Thus, my future studies will have to include a wider range of features. I believe that a possible variable for future studies would be a social network feature which would allow my prediction model to predict the likely spread of the disease given a particular member of a network was infected. These networks could be determined by a user’s interaction history on social media cites and their hometown.

Another extreme limitation of my analysis was the size of my training dataset. Because we don’t have access to historical trends of Covid-19 it is difficult to make any strong conclusions about the diseases’ spread and effect on the American community at large. Even more so, there are ethical concerns with Covid-19 data in general because it only represents those that are able to access testing or hospitals at all. This is a major concern because most of the world does not have access to essential health services.[[7]](#footnote-7) Furthermore, the lack of substantial test kits and treatment resources across the world result in skewed data because the sickest patients are those who receive testing and are hospitalized, potentially skewing overall mortality rates to reflect only the worst cases. To address this concern, I believe it is imperative that testing reach a wider scope of the international community and should be a priority in government’s next steps to address the Covid-19 Public Health Emergency.

**References**

“Coronavirus Disease (COVID-19) - Events as They Happen.” *World Health Organization*, World Health Organization, 2020, [www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen](http://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen).

CSSEGISandData. “CSSEGISandData/COVID-19.” *GitHub*, CSSE at Johns Hopkins University, 18 Apr. 2020, github.com/CSSEGISandData/COVID-19/blob/master/csse\_covid\_19\_data/csse\_covid\_19\_daily\_reports\_us/04-18-2020.csv.

“Total Number of Certified Nursing Facilities.” *The Henry J. Kaiser Family Foundation*, WordPress.com VIP, 3 June 2019, [www.kff.org/other/state-indicator/number-of-nursing-facilities/?currentTimeframe=0&selectedDistributions=number-of-nursing-facilities&sortModel=%7B%22colId%22%3A%22Location%22%2C%22sort%22%3A%22asc%22%7D](http://www.kff.org/other/state-indicator/number-of-nursing-facilities/?currentTimeframe=0&selectedDistributions=number-of-nursing-facilities&sortModel=%7B%22colId%22%3A%22Location%22%2C%22sort%22%3A%22asc%22%7D).

Yu-Group. “Yu-Group/covid19-Severity-Prediction.” *GitHub*, 12 May 2020, github.com/Yu-Group/covid19-severity-prediction.

“World Bank and WHO: Half the World Lacks Access to Essential Health Services, 100 Million Still Pushed into Extreme Poverty Because of Health Expenses.” *World Health Organization*, World Health Organization, 13 Dec. 2017, www.who.int/news-room/detail/13-12-2017-world-bank-and-who-half-the-world-lacks-access-to-essential-health-services-100-million-still-pushed-into-extreme-poverty-because-of-health-expenses.

1. “Coronavirus Disease (COVID-19) - Events as They Happen.” *World Health Organization*, World Health Organization, 2020, [www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen](http://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen). [↑](#footnote-ref-1)
2. CSSEGISandData. “CSSEGISandData/COVID-19.” *GitHub*, CSSE at Johns Hopkins University, 18 Apr. 2020, github.com/CSSEGISandData/COVID-19/blob/master/csse\_covid\_19\_data/csse\_covid\_19\_daily\_reports\_us/04-18-2020.csv. [↑](#footnote-ref-2)
3. Yu-Group. “Yu-Group/covid19-Severity-Prediction.” *GitHub*, 12 May 2020, github.com/Yu-Group/covid19-severity-prediction. [↑](#footnote-ref-3)
4. CSSEGISandData. “CSSEGISandData/COVID-19.” *GitHub*, CSSE at Johns Hopkins University, 18 Apr. 2020, github.com/CSSEGISandData/COVID-19/blob/master/csse\_covid\_19\_data/csse\_covid\_19\_daily\_reports\_us/04-18-2020.csv. [↑](#footnote-ref-4)
5. Yu-Group. “Yu-Group/covid19-Severity-Prediction.” *GitHub*, 12 May 2020, github.com/Yu-Group/covid19-severity-prediction. [↑](#footnote-ref-5)
6. “Total Number of Certified Nursing Facilities.” *The Henry J. Kaiser Family Foundation*, WordPress.com VIP, 3 June 2019, [www.kff.org/other/state-indicator/number-of-nursing-facilities/?currentTimeframe=0&selectedDistributions=number-of-nursing-facilities&sortModel=%7B%22colId%22%3A%22Location%22%2C%22sort%22%3A%22asc%22%7D](http://www.kff.org/other/state-indicator/number-of-nursing-facilities/?currentTimeframe=0&selectedDistributions=number-of-nursing-facilities&sortModel=%7B%22colId%22%3A%22Location%22%2C%22sort%22%3A%22asc%22%7D). [↑](#footnote-ref-6)
7. “World Bank and WHO: Half the World Lacks Access to Essential Health Services, 100 Million Still Pushed into Extreme Poverty Because of Health Expenses.” *World Health Organization*, World Health Organization, 13 Dec. 2017, www.who.int/news-room/detail/13-12-2017-world-bank-and-who-half-the-world-lacks-access-to-essential-health-services-100-million-still-pushed-into-extreme-poverty-because-of-health-expenses. [↑](#footnote-ref-7)