How County Features Have Impacted the Spread of COVID-19

Sabrina Chiang, Makena Wilcox & Susan Zhang

Abstract

How prepared was our United States healthcare system for COVID-19? We compared data from a dataset of COVID-19 confirmed cases per county, a dataset of COVID-19 deaths per county, and a dataset of a variety of features that might affect the likelihood of getting COVID-19 for each county in the United States. We cleaned our data and created visualizations to find which states had the most cases, which counties had the most cases, which states had the highest/lowest proportion of deaths per cases, and if the number of ICUs and the number of hospitals in each state affected the number of deaths. We used linear regression and trained our data to find the actual versus the predicted values and the error rates against the root mean squared equation. We found that while there is a correlation between the number of hospitals and ICU beds to percentage of deaths/cases, the correlation was positive. We further found that a majority of the cases of COVID-19 were found to be clustered in the east coast of the United States as compared to the west. Michigan had the highest percentage of deaths to cases with 7.6%, even though New York had the highest number of COVID-19 cases. New York was found to have 6.8% of cases to deaths and when the cases were normalized against the population New York had 1.24% of cases per population. The maximum percentage of cases per population was found to be 1.23%, while the lowest was Minnesota with 0.04%. Our findings supported our hypothesis, but didn't predict a negative correlation between hospital and ICU beds to the percentage of death to cases. While the United States healthcare has enough resources to support the growing number of cases, it doesn't solve the problem of preventing cases from continuing to grow.

Introduction

Coronavirus or COVID-19, is an infectious disease, first identified in a 55-year old individual in November of 2019, Wuhan, China. COVID-19 is known to be similar to SARS and MERS, a large family of enveloped RNA viruses. Symptoms can range from mild to severe illness, such as infections of the lower and upper respiratory tract. Incubation periods for symptoms range from 2 - 10 days, 2-14 days, and 10-14 days, where patients with asymptomatic transmission can spread the contagious virus. On March 11, 2020, COVID-19 was declared a pandemic. Compared to SARS-CoV with 812 total deaths and case fatality of 9.6%, and MERS with a total of 2,519 cases, 866 deaths, and 34.3% fatality rate, coronavirus has a lower case fatality, but higher infection rate (Hewings-Martin). Coronavirus currently has 4,424,119 cases, 297,683 deaths and 1,654,918 recovered as reported by worldometer as of May 13, 2020. Through analyzing datasets of hospital resources against growing cases of coronavirus per county in the United State, we wanted to assess if hospitals have enough resources to treat the growing number COVID-19 patients. Specifically, we want to understand: what are the "current" hotspots of COVID-19 cases, do the amount of deaths in each county have an even ratio to the number of total cases in the same area, and does the amount of hospitals/ICU beds in each state impact the ratio of total cases to deaths in each state. We predict the percentage of deaths to cases will decrease with the number of hospitals and ICU beds available per county increases.

Description of Data

We named the abridged_counties.csv “features_data”, the time_series_covid19_confirmed_US.csv “confirmed_data”, and the time_series_covid19_deaths_US.csv “death_data”. First we cleaned the features_data by removing feature columns that were not relevant to the questions we were trying to answer. We kept the columns that had the county FIPS as the primary key, the columns for county names, state names, and state abbreviations so that we could group our data based on counties and states, the latitude and longitude columns so that we could plot each county in our visualizations of the United States map (visualizations 1, 2, 3), the column containing the population estimate of 2018 so that we could calculate what percent of each state got COVID-19, and the columns containing the number of hospitals and the number of ICU beds so that we could use these variables to train our linear regression model. Second, we noticed that a lot of states were missing their corresponding state name, so we filled those in. Third, we removed all the rows that do not include data from the 50 states in features data because this was not relevant to our analysis of the United States. We also dropped the last two rows because the format of the FIPS did not match the format of the rest of the other FIPS. Fourth, for both confirmed_data and death_data, we cleaned the data by only keeping the FIPS, County Name, State Name, Latitude, and Longitude columns for the same reason as features_data. We also kept the 4/18/20 column, so we could keep track of how many confirmed cases and death cases there were in each county. We need this to calculate the proportion of deaths per number of cases in each state as seen in visualization 4. We also dropped the bottom two rows because they did not have the data we wanted. Fifth, we found the total number of cases and deaths per state and put those in new data frames indexed by the state name called total_per_state and total_deaths_state respectively. After doing all this data cleaning, we were able to create our first three visualizations. Next, we merged the three datasets on the primary key, FIPS. We only keep the columns, ‘FIPS’ and ‘4/18/20’, in both confirmed_data and death_data so that we know how many confirmed cases and deaths there are for each FIPS in features_data. Confirmed_data and death_data have more rows than features_data because we did not clean them, but we did not have to clean them because when we merged on FIPS, the merged data frame only kept the rows that we needed. This new merged data frame includes all the features, the number of cases, and the number of deaths which we need to create Figs 4, 5, 6, and 7 below.

Description of Methods

We decided to create visualizations for the amount of cases and deaths per county to first have a better understanding of any hotspots in the United States, by having the information displayed on an United State map it was clear where the longitude and latitude values were located in the states. From recognizing the different clusters of confirmed cases and death count, we chose to create a bar plot of the death to confirmed cases ratio for each state. Learning the ratio led us to our next questions of whether the number of hospitals and the number of ICUs was correlated to the deaths per number of cases rate per state. We split our merged data set into 20% testing and 80% training sets. We trained our linear model on two different features: number of ICUs and number of hospitals. Then, we calculated a training error of 1.5669939888738655 and a testing error of 2.1669629442717846. To further assess if we had made a good model, we used 5-fold cross validation. We normalized the data and found the hyperparameter with the smallest cross-validation error. We found that the best alpha value was 0.5, and the cross validation error for this alpha value was 1.7777724172826972. We wanted to visualize our errors and test to see if a linear regression model was good for the data, so we created a residual plot. We then made a regression line to determine if there was correlation between the number of hospitals and the number of ICUs with COVID-19 death rate for each state.

Summary of Results

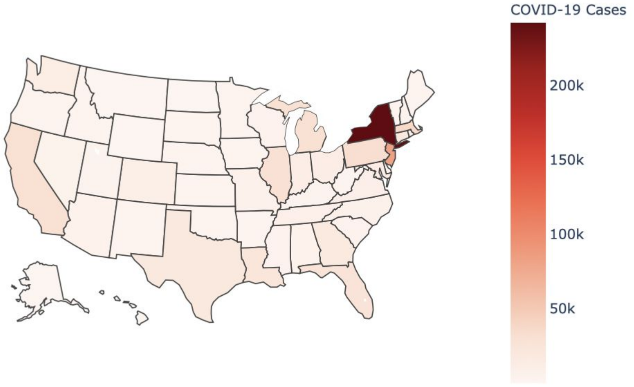

Fig. 1. US COVID-19 Cases per State

New York State had the most confirmed cases on April 18th, 2020 with over 241,000 cases. The state with the second most cases was New Jersey, however it only had roughly a third of the amount of cases compared to New York with 81,000. After these two states, The amount of total cases dropped to 30,000 and below (Fig.1).

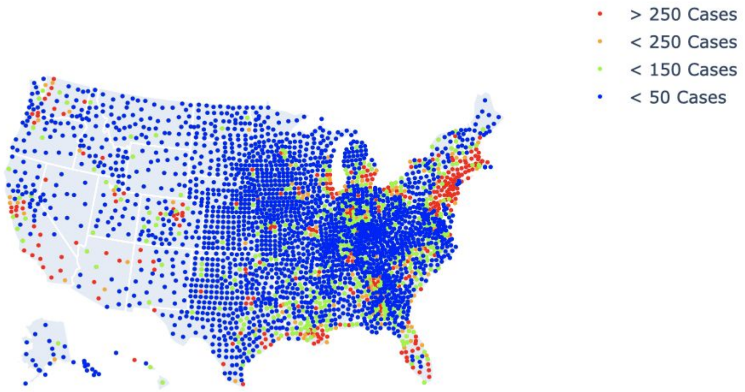

Fig. 2. US COVID-19 Cases per County

This visualization provides a color classification for each county depending on how confirmed cases are documented. From this we can recognize New York City has the highest number of cases compared to the other documented counties from the dot being colored red. One city in Southern California and a few others in the East Coast have a light blue color indicating they also have a higher number of cases. It is important to note that the amount of dots in each state are determined by the data provided in the CSV files, they do not show a density of cases. With having more data for the states in the eastern half of the United States the total number of cases per state will be more accurate due to having more information of the state's counties (Fig. 2).

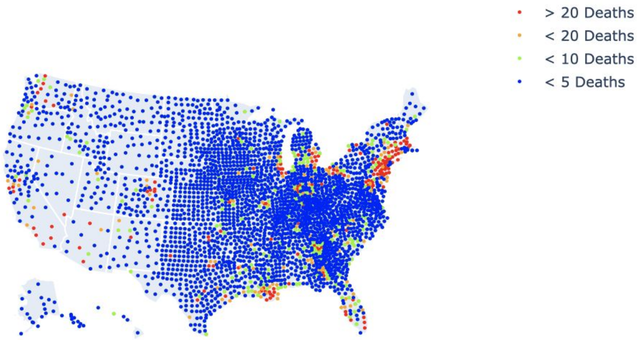

Fig. 3. US COVID-19 Deaths per County

There are similar hotspots with higher deaths compared to the Fig. 2 mapping of counties total number of cases. If a state's counties are marked with red and orange dots, we can conclude that the state overall will have a higher number of deaths, even though the dot color ranges aren't large (Fig. 3).

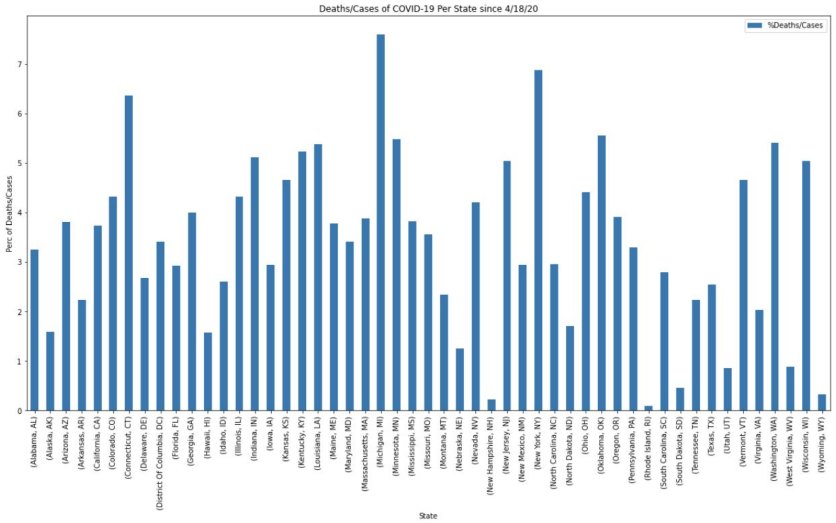

Fig. 4. Bar Plot of State's Proportion of Deaths vs. Cases

We plotted the proportion of the number of deaths vs. the number of cases of the coronavirus for each of the 50 states. We were shocked to find that Michigan had the highest proportion of around 7%. New York also had a higher death rate of COVID-19 compared to most of the other states. This was not shocking though as we know from our first visualization that New York has the most cases of the coronavirus by far. We were shocked to see that. We also found that Rhode Island had the smallest proportion of deaths vs. cases. Wyoming and New Hampshire also had a surprisingly low proportion (Fig. 4).

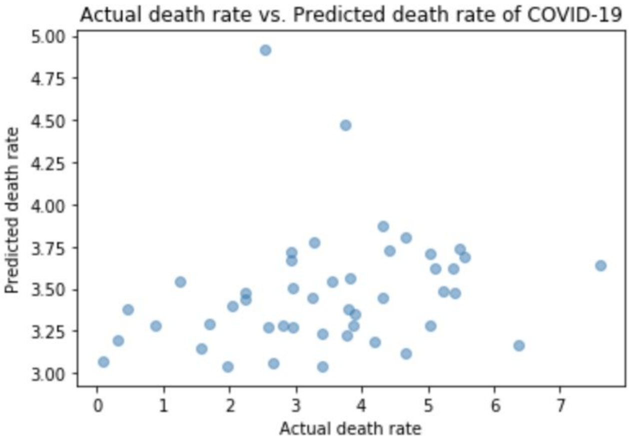

Fig. 5. Actual vs. Predicted

The model describes the relationship between the death rate of COVID-19 and two features: number of hospitals and number of ICUs per state. We found from the scatter plot of actual vs. predicted death rates, our model is not perfect. If it were perfect, we would see the identity line. We found that there are a few outliers, and the scatter plot has a weak positive correlation. Without the outliers, the slope is around .125 (Fig. 5).

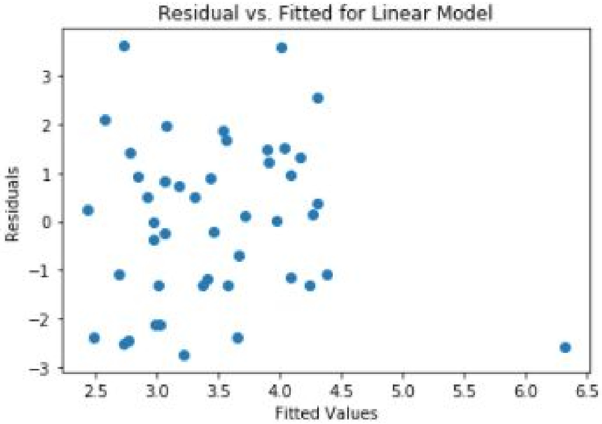

Fig 6. Residual Plot

The residual plot does not show a random pattern with most of the data points all clumped on the left side; therefore, it is not a "good" residual plot. If the outlier on the right was removed, this would be “good” because the plot would show a random pattern and similar vertical spread throughout. This would tell us that our linear model is a good fit for the data. However, because we have the outlier, we found that our model was not a good model to predict the death rate(Fig. 6).

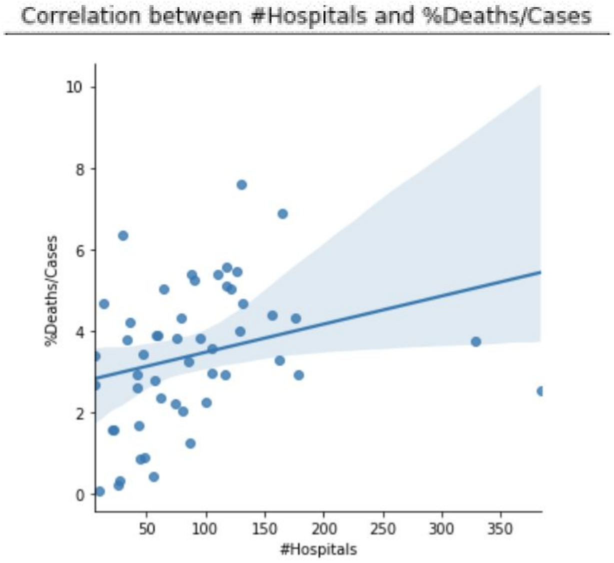

Fig. 7. Regression Line

After training our model on the two features: number of hospitals and number of ICUs per state, we were curious if the amount of hospitals available in each state impacted their percentage of deaths to cases. This residual plot shows a weak positive correlation between the death rate and the covariates. This is surprising because we expected the regression line to display a negative slope; indicating the percentage of deaths decreased as the amount of medical resources increased. Most states had the max amount of hospitals reaching just under 200 facilities. The positive slope indicates that there are other factors that determine the survival rate of a COVID-19 patient (Fig. 7).

Starting with an overview visualization of confirmed cases and deaths we were able to have a better understanding of what information we had access to. After each visualization we would take note of information that stands out and analyze that part to further understand what could be concluded by the data. For limitations, Fig. 2 and Fig. 3 showed that our datasets did not have equal representation for each state; not all counties were accounted for and some counties had incomplete data. This missing data could have potentially impacted the accuracy of a state's information when calculating the number of hospitals and ICU beds available.

At first, we were surprised that the increase in hospital beds did not directly decrease the death rates of COVID-19 patients (Fig. 7). Having a well resourced health care system is not the leading factor in helping patients with COVID-19, rather any preexisting health conditions play a role in the success of recovery.

Discussion

Before analysing the different datasets we were curious if the United State healthcare system was prepared to fight against a health crisis. We were expecting that the states with more hospitals and ICU beds were going to have fewer deaths indicating that more health care facilities made the state more prepared, however surprisingly there was no beneficial correlation between increasing available hospitals and decreasing death percentages.

Our original plan was to look at the features in the abridged_couties.csv including: the percentage of people with diabetes, number of people in different age ranges, percentage of smokers, and other health related variables to classify which features would someone are higher risk of caching COVID-19. However, this was ineffective because the data provided was for the state county as a whole and not individualized for each person. To have successfully conducted our first idea we would've needed data for individuals with a confirmed diagnosis of COVID-19 and others who hadn't been infected by the virus. If we had additional data for individuals we could've looked into how certain features affected people's recovery from COVID-19, such as pre-existing health conditions.

Besides not having information available for individual persons, the datasets provided didn't have complete data for each state. The majority of the county data represented to Eastern half of the United States. The total counts for the Western half could be underestimated due to having less counties recorded. Also, county data varied in completeness of recorded points, leaving some features left blank.

Challenges we had with our data was that we wanted to further use the cases and deaths by county to predict how likely certain factors would cause a person to be affected by COVID-19. However we were limited by datasets provided to us. We also found it difficult to interpret certain data provided in the tables, such as HPSA and 3-YrMortalityAge. We had to sort through the data to figure out which information would be valuable when predicting the adequacy of the healthcare systems in the United States to the percentage of deaths to cases. We also didn't want to fill in columns without data with 0's as it would impact the averages and sums of our information when comparing states between each other. Furthermore we struggled with figuring out how to test and train our data against our actual values.

Ethical dilemmas we faced when analyzing the datasets was that we didn't want to be biased towards certain states with more county data or states that had data for hospital and ICU beds, but also didn't contain data for COVID-19 cases and deaths. Thus, when we merged our tables we only took county data that was present across all tables. Other ethical concerns we had when we wanted to initially predict features that made it more likely for a person to get COVID-19 are issues with regards to privacy concerns and access to personal health care data. If we had access to this data, we would've removed patient names in order to keep it confidential and made sure to make it transparent to patients what we were using their data for.

Works Cited

1. “COVID-19 CORONAVIRUS PANDEMIC.” Worldometer, 13 May 2020, www.worldometers.info/coronavirus/.
2. Hewings-Martin, Yella. “How Do SARS and MERS Compare with COVID-19?” Medical News Today, MediLexicon International, 10 Apr. 2020, www.medicalnewstoday.com/articles/how-do-sars-and-mers-compare-with-covid-19#SARS.