Predicting which counties will be severely affected based on mortality rate and the SIR model

Katlyn Ho, Trevor Williams

New York, NY is the hardest hit county in the United States, but no one saw this coming. Preventative measures were not taken to prevent such a large infection rate. There was little to no data at the start of this pandemic, but now there's a few months of data available. There are a lot of questions that are unanswered. What makes this location so susceptible to the virus and it's spread? With the beginning of the re-opening of states, can we somehow predict which places are at risk? As it has been in the past, people with compromised immune systems are told to be wary and cautious of a virus, but do populations with a greater percentage of people with, say, asthma, report greater amounts of confirmed cases or deaths? Here, we seek to find answers to these questions by selecting certain features to use in a model to predict where we should be looking next.

INTRODUCTION

Coronavirus disease (COVID-19) is a highly infectious disease caused by a newly discovered coronavirus. It spreads very easily from person to person primarily through respiratory droplets, such as saliva or nose discharge that can come from coughs or sneezes. Currently, there are no vaccines or treatments for COVID-19; however, there are a multitude of researchers and clinical trials working to find potential treatments. Older people and those with compromised immune systems or underlying medical problems are more likely to develop serious illnesses and are told to be overly cautious when going about their daily lives. We have officials researching and operating at the state and national level, but maybe the answers are at the county level.

A lot of predictions and data seen on the news primarily involves large populations of people either at the state or national level. There is a lack of prediction and scope of how COVID-19 is spreading at the county level. Take for example, California. In this one state, there are 58 counties that range in geographic location (coastal, forest, inland, etc.), number of people, and demographics. In this project, we explore how stay at home orders and the implications that these features may pose in the spread of COVID-19 as well as predicting the spread in the future in US counties. To do so, we perform EDA and data cleaning to understand the data better and remove or fix unwanted values. Then we run the data on the models; the two that we use are the sklearn LinearRegression model and the SIR infectious disease model.

Fig. 1: *counties* table columns

DATA

The data we were given are 4.18states.csv (named as *provinces*), abridged_couties.csv (named as *counties*), time_series_covid19_confirmed_US.csv (named as *confirmed*), and time_series_covid19_deaths_US.csv (named as *deaths*).

To clarify, we used the updated versions of the time_series csv files so our most recent date would be 5/11/20.

The *provinces* table has general and average data pertaining to US states and global provinces as of 4/18/2020. Its columns include 'Province_State', 'Country_Region', 'Last_Update', 'Lat', 'Long_', 'Confirmed', 'Deaths', 'Recovered', 'Active', 'FIPS', 'Incident_Rate', 'People_Tested', 'People_Hospitalized', 'Mortality_Rate', 'UID', 'ISO3', 'Testing_Rate', and 'Hospitalization_Rate'.

The *counties* table has 87 columns that pertain to each county's ID code, name, location, population demographics, and lockdown/closure dates. A full list of column values can be seen in **Fig. 1**.

The *confirmed* and *deaths* tables have the same column labels: 'UID', 'iso2', 'iso3', 'code3', 'FIPS', 'Admin2', 'Province_State', 'Country_Region', 'Lat', 'Long_', 'Combined_Key', and each day from 1/22/20 - 5/11/20. The difference between
the two is that the actual values per date correspond to the cumulative number of confirmed cases or dead on that day, respectively.

METHODS

EDA & Data Cleaning

Before creating any data visualizations, we had to do data cleaning and transformations on the data in order to compare the values we wanted to on a plot. The questions we sought to answer led us to look for the data in the tables, which had a lot of invalid and non-relevant values.

The problems and questions we posed to answer pertain to COVID-19 data in the US only, thus, we filtered *provinces* for only the US as well as any rows with invalid data, such as a 'Province_State' named 'Recovered.' The *counties* table was cleaned to include valid counties with valid FIPS and Admin2 and transformed to have Timestamp dates rather than the original Gregorian ordinal of dates. This would make it easier when visualizing things over time with respect to these columns, such as the 'stay at home' column. The FIPS values for county, state, and state-county were not in a uniform type, so we had to convert some from floats to ints in order to perform merging on these tables later on. We merged *counties* with *confirmed* to create a table with all variables per county sorted in decreasing order by the last date in the time series; the same was done with the *deaths* table.

One of the questions we initially sought to address was the impact of stay at home orders on the spread of the virus; however, there was no one table that had this data. Thus, we had to merge *provinces* with *counties*. This allowed us to visually see how the number of confirmed cases changed over time for each county. There are too many counties to include them all on the plot, so we looked at the trends of the top 20 counties with the most confirmed cases (**Fig. 2**). The legend is comprised of 'countyFIPS', and the vertical dashed lines correspond to stay at home order dates (before 3/22, before 3/29, before 4/5, before 4/12), which were taken from this site: https://en.wikipedia.org/wiki/File:COVID-19_outbreak_USA_stay-at-home_order_county_map.svg

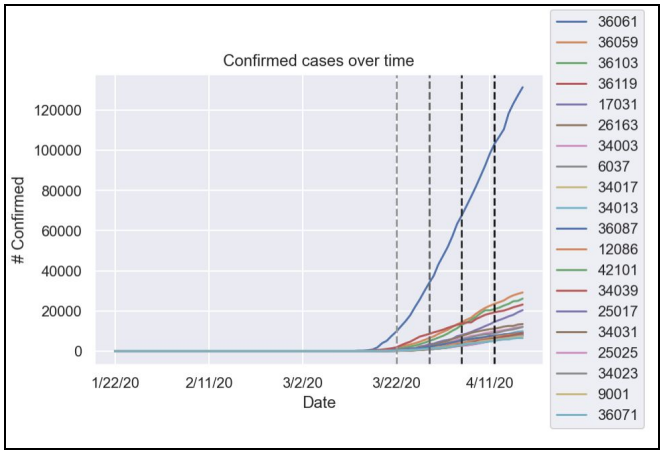


Fig. 2: top 20 counties confirmed over time

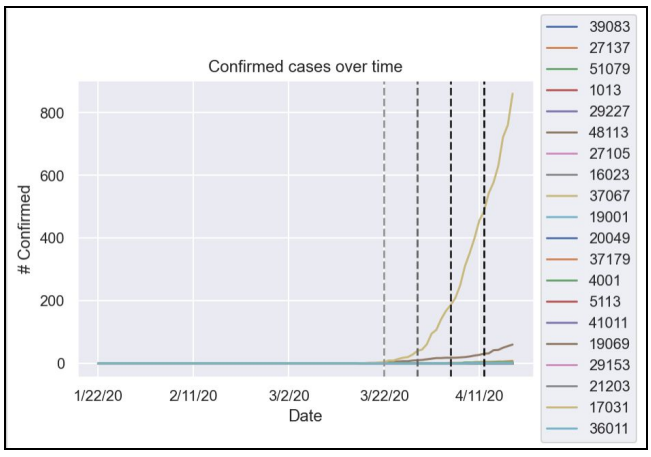


Fig. 3: random 20 counties confirmed over time

Using the same function to make the plots we can see how the number of confirmed cases changes over time for a random selection of 20 countries (**Fig. 3**) as well as the top 20 countries with the most deaths (**Fig 4**). The blue line is New York, NY, which is already known to have a high confirmed case count, and has a much greater slope than the other counties on the graph. This was also a way to help visualize if and/or how the stay at home orders affected COVID-19 spread and number of infections, and there are no apparent changes at any of the vertical lines.

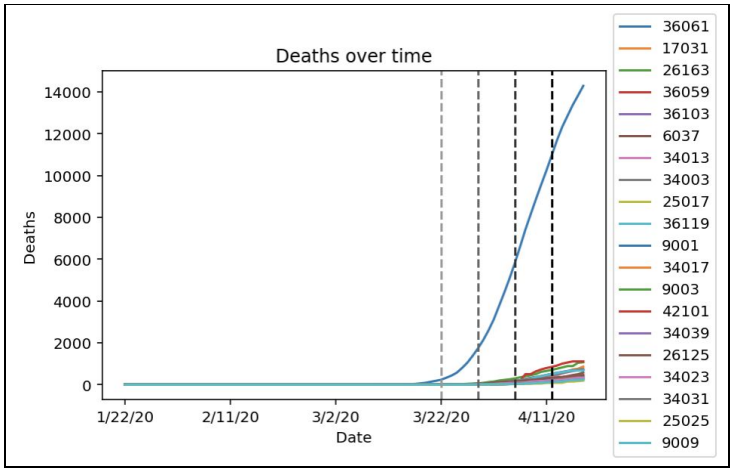


Fig. 4: top 20 counties deaths over time

To get a better idea of how the previously listed stay at home order dates have affected spread, we first made a boxplot of deaths versus all stay at home order dates (**Fig 5**), but nothing could be drawn from it because of the scale. Thus, we made four different line plots for mortality rate vs each of the four stay at home order dates. Each one has the countries with the corresponding date. Unfortunately, from the plots, we can see that the stay at home order date has no large impact on the number of deaths over time, so this feature was not considered for use in the model.

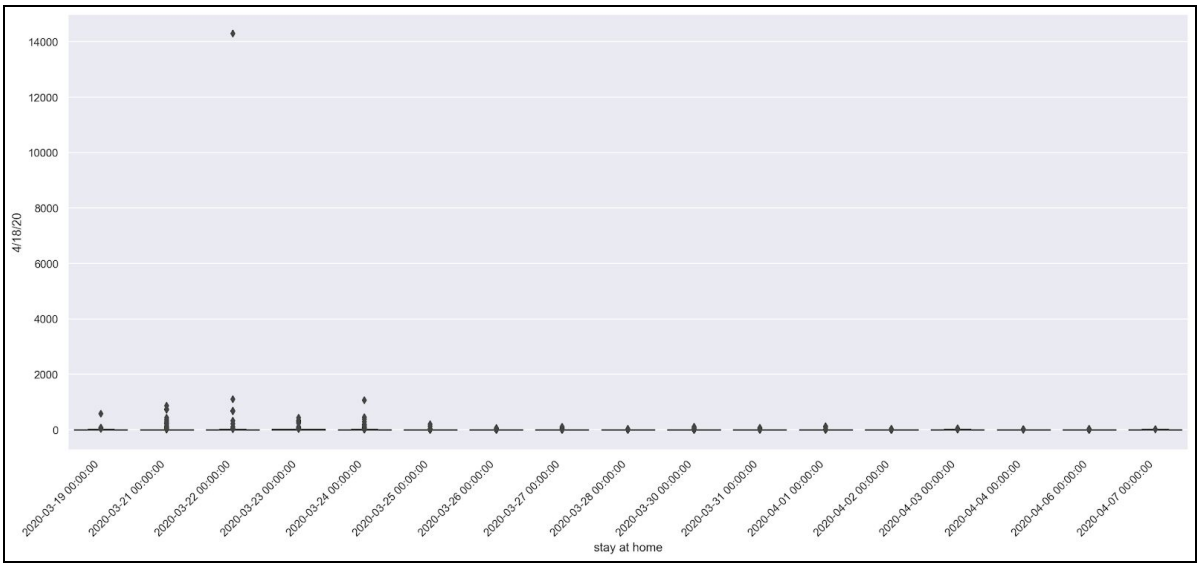


Fig. 5: deaths vs stay at home order date

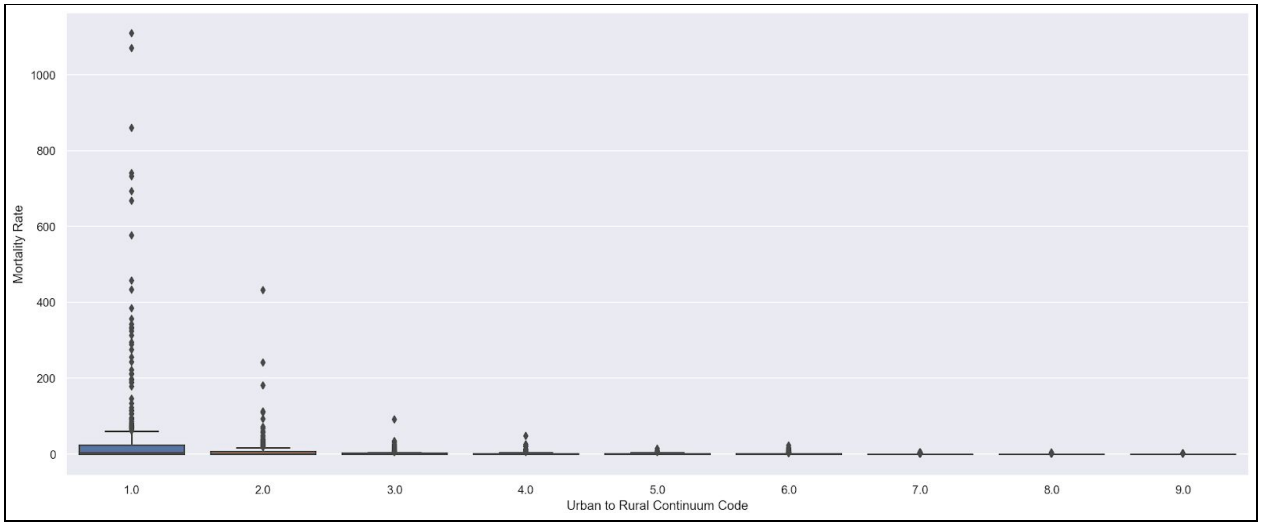


Fig. 6: Mortality Rate and Urban-Rural Continuum Codes

After analyzing our SIR model, we wanted to better understand what features of a county contributed to its death rate so that we could know ahead of time if a specific county would be more sensitive to an outbreak. In order to create a model that represented how vulnerable a county was to a coronavirus outbreak, in terms of death rates, we needed to better understand how each county was affected. To do this, we first started by reformatting our data. We merged the confirmed cases by counties data frame (variable *confirmed_new*), the counties data frame and the deaths by county data frame (variable *deaths_new*) to create a new data frame *confirmed_with_info* and added features such as death rate. We then found the counts of each county's death rate on 5/11/2020 and plotted them, as shown in **Fig 7**.

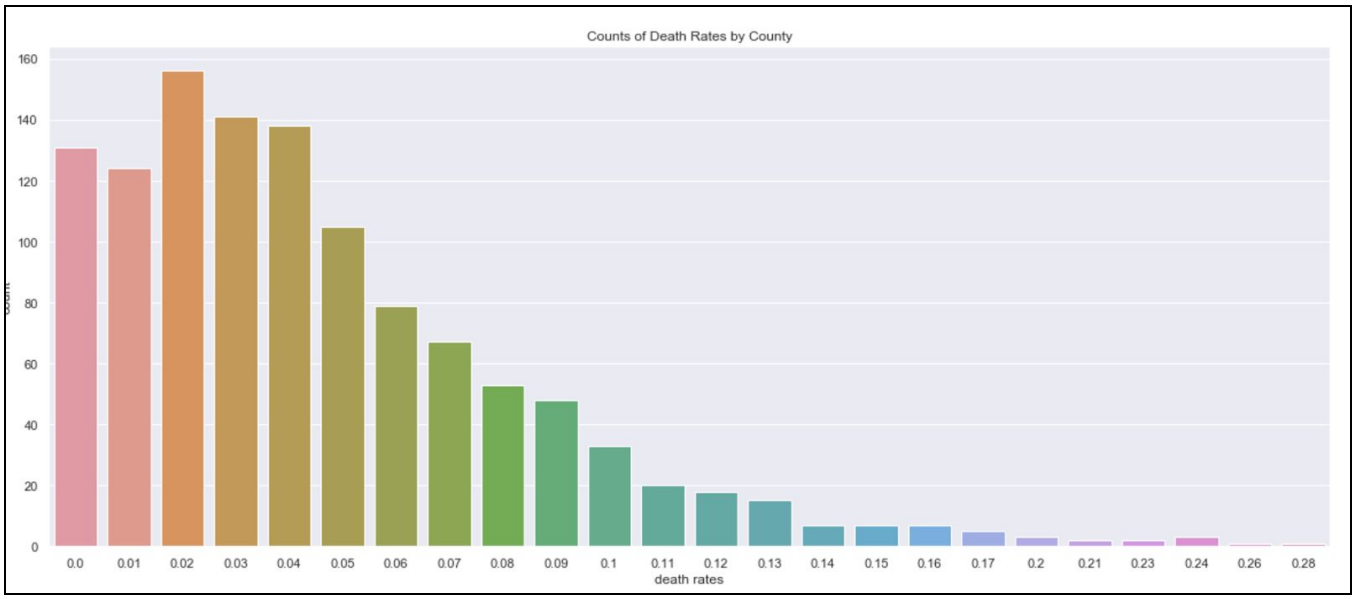


Fig. 7: Distribution of Mortality Rates

The mean of this data was 0.045 and the standard deviation was 0.04. From there, we classified each county as either 'good', 'fair', 'bad' and 'severe' based off of how high their death rate was on 5/11/2020. Counties within the first standard deviation were deemed 'fair', those 1 standard deviation below fair were

deemed ‘good’, those 1 standard deviation above ‘fair’ were deemed ‘bad’ and everything above ‘bad’ was deemed ‘severe’.

As we were creating a new model, our data needed to be re-cleaned to reflect our new needs. Since we were trying to predict which category each county would fall into, we decided it would be best to insert the average of a column, normalized by a county’s population for columns with missing values. We chose this over replacing each NaN with some constant because we believed that doing so would not be very representative of that county. This method was used for features such as the number of people eligible for medicare and the stroke and heart disease mortality counts. We also needed to find a way to encode the Gregorian time given to us in columns like ‘stay at home’. To clean this data, we converted these columns to the number of days since January 1st, 2020 that the Gregorian time represented and filled in NaNs with 366, to represent that the order had not been enacted.

As we were trying to get our model to classify a county into 1 of 4 categories, we opted for a Random Forest classifier, as we had already used one in lab. Since we were trying to predict how a county would be affected in the future, we included the death and confirmed cases counts from 1 week prior. From there, we picked and chose different features to add to our model until we got the highest cross-validation score we could get, using features such as Heart disease mortality, Diabetes percentage and the number of ICU beds.

SIR + DS Model

Interesting features: deaths

Ineffective features: stay at home

Challenges: no recovered/susceptible data over time

The purpose of this model is to predict what could happen to counties at a certain time. This model was adapted from the note posted on Piazza with some additional variables and features. An overview of the model can be seen in the figure to the right (**Fig 8**). The first attempt at creating this model, we followed the guidelines on Piazza and did not include any additional variables. We tried holding the sir data within lists, but found that to be inefficient considering that it had to be for every county in the table. Using lists would only work if we were looking at one county, so we made each variable (besides N) separate data frames that would accommodate lists of values per county and it’s the same kind of data for multiple counties.

S = susceptible, I = infected, R = recovered, DS = dead
N = population size

$R0 = 3.8 - 8.9$ https://en.wikipedia.org/wiki/Basic_reproduction_number
$R0 = \beta * D$

β = expected num of people an infected person infects per day
D = length of period when a person is ill and can infect others
https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-manage

$\gamma = 1/D$ (rate of recovery)
α = mortality rate
ρ = rate at which people die (1 / days from infected until death)

$$S(t+1) = S(t) - \beta * I(t) * S(t)/N$$
$$I(t+1) = I(t) + \beta * I(t) * S(t)/N - (\gamma * I(t)) - (\alpha * \rho * I(t))$$
$$R(t+1) = R(t) + \gamma * I(t) * (1 - \alpha)$$
$$DS(t+1) = D(t) + \alpha * \rho * I(t)$$

Fig. 8: SIR Model Overview

The error was still relatively large, and after realizing that it was due to the insufficient susceptible and recovered time-series data, we decided to include deaths in the equation. We had a time-series table that was the exact same as *confirmed* except that it tracked deaths. The SIR model does not account for deaths, so we had to add it in and make the corresponding adjustments to the current equations. Including deaths also means we had to include new variables to account for death rates and how long it takes for an infected person to die. The idea for this stemmed from an altered SIR model for coronavirus made at this site: https://towardsdatascience.com/infectious-disease-modelling-beyond-the-basic-sir-model-216369c584c4

We tried seeing how the model would change if we subtracted deaths from N, but realized that that no longer follows the model. One of the ideas of SIR is that population size is constant and S + I + R = N throughout, so including deaths (DS) shouldn’t change the equation. The new equation is S + I + R + DS = N. The output of

this model comprises a complete time-series table for SIR and DS, including the dates and values from train and test. **Fig 9** shows a line plot of the data over time. It does not look like a normal SIR graph because of the limitation of the susceptible and recovered data. The last four tables in the notebook answer the question of: "Which countries will have the most infections or deaths according to this model? What if we exclude NY counties since we already know they are the most affected?"

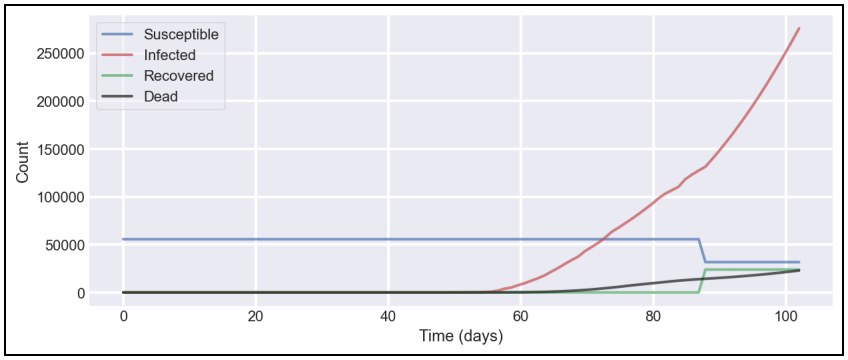


Fig. 9: SIR + DS Model Graph

Mortality Rate Model

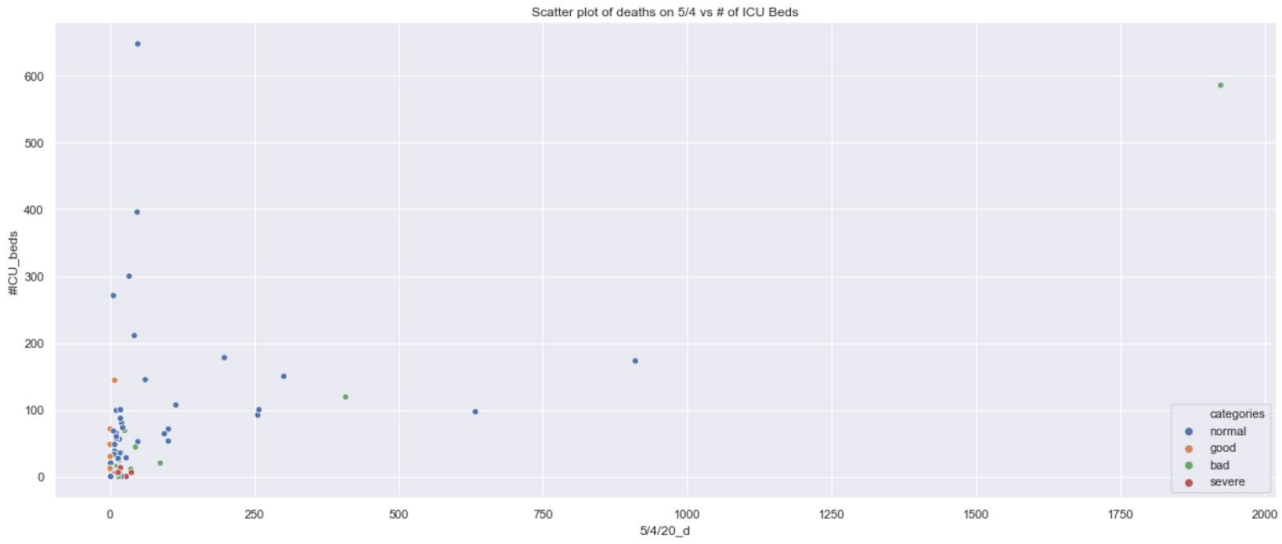
Interesting features: Death counts from the week prior, Diabetes percentage, Number of ICU Beds
Ineffective: stay at home order date and entertainment/gym closure date
Challenges: not enough data on other relevant features

After running different types of models and trying different features, we decided to use a Random Forest classifier to classify our counties. It ended up predicting the correct death rate category about 84% of the time. From this we wanted to see if we could predict the status of a county a week in advanced, as in the times of a pandemic, 1 week can make an extremely large difference in how many total cases, and therefore, total deaths there are.

Further, we saw that features such as the total population, the percentage of people who have diabetes and the amount of deaths a week prior helped bring our accuracy up.

Additionally, we understand that 84% may seem a bit low for accuracy. We believe this is because there are many other factors involved in the spread of a pandemic that are not seen in the data given to us or that we could derive from the data we had. Things such as how full hospitals are at a given time, the initial viral load a person gets when contracting the disease, how easy it is to physically get to a hospital and other pre existing health conditions were not found in the data given that could potentially have a large impact on the death rate of a county.

To visualize this data, we created a scatterplot of the number of ICU beds vs the Deaths on 5/4/2020, seen below.



In general, the 'normal' counties were ones that had high ICU bed counts and death counts, while the 'good' counties had fewer ICU beds and fewer deaths, indicating that those cities were smaller in general. It is also seen that most 'severe' cities had similar amounts of ICU beds as 'good' cities, but just with more deaths.

SUMMARY

Cross-Validation and Test Data

To test the SIR model, we could not do a normal train-test split like we had done before. Because this followed a strict time-series, the split had to be in time order. The test set consists of the dates that we want to predict. To determine how well the model performed, we calculate the RMSE for the infected and deaths tables using the dates in the test; we can't calculate it for susceptible and recovered since there are no actual values to compare with. The output is a list of RMSE values from predicted day 1 to day 14, with RMSE increasing as the number of days predicted increases. This is to be expected since the model is calculating values that are farther away from the true values recorded.

For the Random Forest Model, we were able to do a test-train split. For this we used sklearn's built in function to separate our data for us. We also used sklearn's cross-validation function, using the model's accuracy. To understand how our model was doing when we added or subtracted features, we looked at the cross validation score and occasionally the test score to see how we were doing. The cross validation score was very high, while the test accuracy was much lower, as that is how a Random Forest model usually works.

Limitations (of the models)

A major limitation of the SIR model is that we did not have time-series recovered data for every county. We only had a total amount recovered per state as of 4/18/20. We tried our best to make the data work, but we understand that it's not 100% accurate. Being able to have this would improve the model and its predictions. Another limitation is that R0 varies greatly (3.8 - 8.9). I assume that it's because COVID-19 can spread so quickly depending on its environment. In order to minimize the error, we had to choose the smallest R0 value, which is not correct for every county.

A big limitation of the Random Forest model was that many of the counties had very limited data, in terms of confirmed covid cases. This meant that in general, counties with smaller populations would usually have lower confirmed case amounts. If one of those counties had a single death, it would disproportionately affect the death rate of that county.

DISCUSSION

Ethical Dilemmas and Concerns

One dilema we had when building our models was that predicting how a county will do in the future is not always very accurate. As this pandemic is a very fluid situation, no model can perfectly predict what will happen in the future. If we entirely rely on models like these to decide how to respond to future events, we run the risk of being very wrong, as has happened in the past. In today's world, we tend to assume that everything we see online or on TV is the truth, often without cross checking facts. If our model were to be wrong and it were to be seen and used widely, this could contribute to the spread of false information. A model like this is not a simple guide book to how to solve a problem, only a tool.

Surprising Discoveries

Something surprising when performing EDA was that there was no obvious change in the data due to stay at home orders. Considering that the entire nation is on lockdown, we expected it to have a much larger impact on the data that we saw, and we saw almost no abrupt difference in confirmed cases and deaths. This might be

because it takes 2-14 days from exposure to develop symptoms and the approximately two months that the US has been in lockdown is not sufficient to see an abrupt change. It might also be because there is more testing available and faster tests so more people are being tested positive.

Future Work

Something that we were considering was finding a way to include a different model that would take other features into account, similar to the Mortality Rate linear regression model, but in relation to the SIR model. So this model would be able to take into account more features that will help predict values at a certain time. This would definitely make the model more accurate, but we weren't able to figure out how to do that at this time.