Covid19 Data Analysis Workflow

Sarah Tang

May 2020

Abstract

This analysis examines covid-19 data to determine the growth of the virus in the United States across various states and counties. Here I utilize datasets of confirmed covid-19 cases and deaths alongside information about county demographics and state testing to predict the number of new covid-19 cases on a given day per county and to predict the number of total deaths per state. This analysis found that historical trends are important to understand the future number of covid-19 cases and that number of deaths, predicted off of number of people tested and confirmed cases, is a better indicator of the growth of covid-19 than the number of confirmed cases because of testing biases. Future research will examine percent change over time and incorporate more detailed hospitalization data to form a clearer picture of the covid-19 crisis.

Introduction

This analysis focuses on exploring and analyzing covid-19 data to determine the impact of the virus in the United States by identifying informative variables through EDA and using sklearn to build more complex linear models. Here, I examine cumulative and additional cases by state and percent change of cases to predict the number of covid-19 cases on 4/18/20 for each county and I analyze biases in the data before predicting total number of deaths per state. The main questions asked are:

1. Can we predict the number of new corona cases? What features impact corona virus growth? Specifically, why did New York experience such a high number of cases?
2. What factors contribute to total number of deaths per state?

Data

The state dataset consists of 140 records of various states and provinces around the world and includes information regarding the number of confirmed cases, deaths, people hospitalized, people tested, etc. as of 4/18/20. The dataset has 18 features.

The *covid_confirmed* and *covid_deaths* datasets contain 3,255 records, one for each county in the United States and include the cumulative number of covid-19 confirmed cases and deaths respectively by county from 1/22/20 until 4/18/20. The confirmed cases dataset has a total of 99 features while the covid-19 deaths dataset has a total of 100 features.

The abridged counties dataset has 3244 records for the counties in the United States and has 87 features including information about the demographics of a specific county (population total, ages, poverty, mortality rates, health issues) as well as information on lockdown dates and social distancing.

The last dataset used is the daily dataset, which has 3769 records and 27 features. This dataset provides daily information for each state about positive and negative cases, hospitalization, the number of people in the ICU or using a ventilator, and the amount of testing.

Exploratory Data Analysis

I started off by examining my home county, Westchester, New York for the total number of confirmed cases and deaths. Westchester has a high number of cases but relatively few deaths compared to other counties (Figure 1).

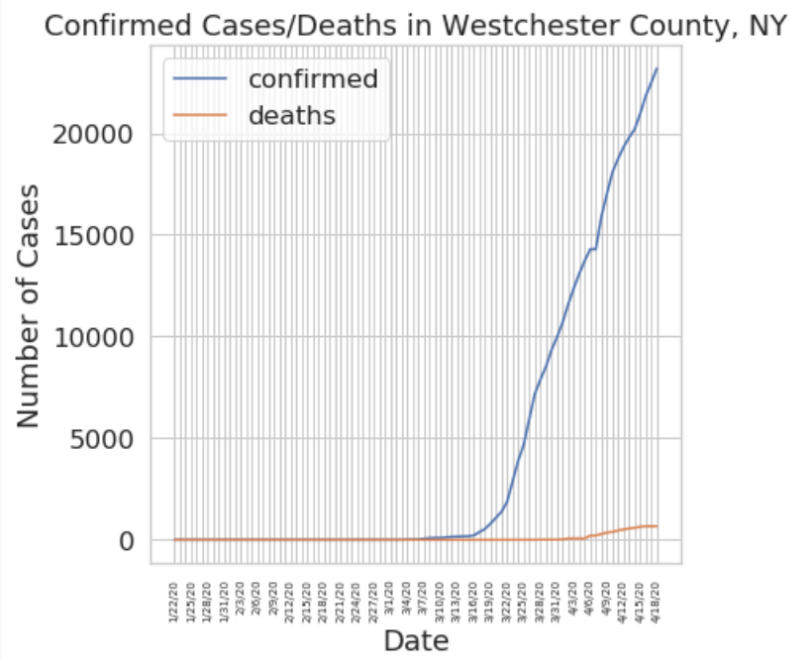


Figure 1: Cumulative confirmed cases and deaths in Westchester County, New York.

Next, I delved into the number of covid-19 cases by state and looked to see if there was any correlation with incident rate versus hospitalization and testing rates.

I found that New York had the highest incident rates and testing rates, but a smaller hospitalization rate than other states such as Kentucky. I decided to understand more about the correlation between these various rates as well as an additional case by case basis rather than cumulative cases. Examining the number of additional cases per day allowed me to see how covid-19 was growing and at what pace. (Figure 2, 3, 4).

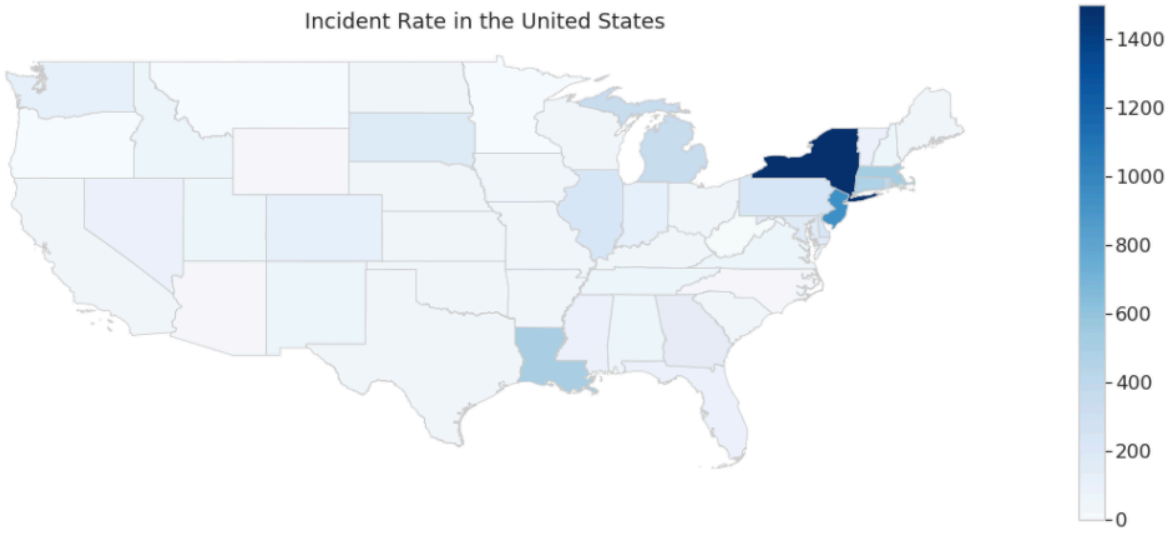


Figure 2: Incident rate by state in the United State

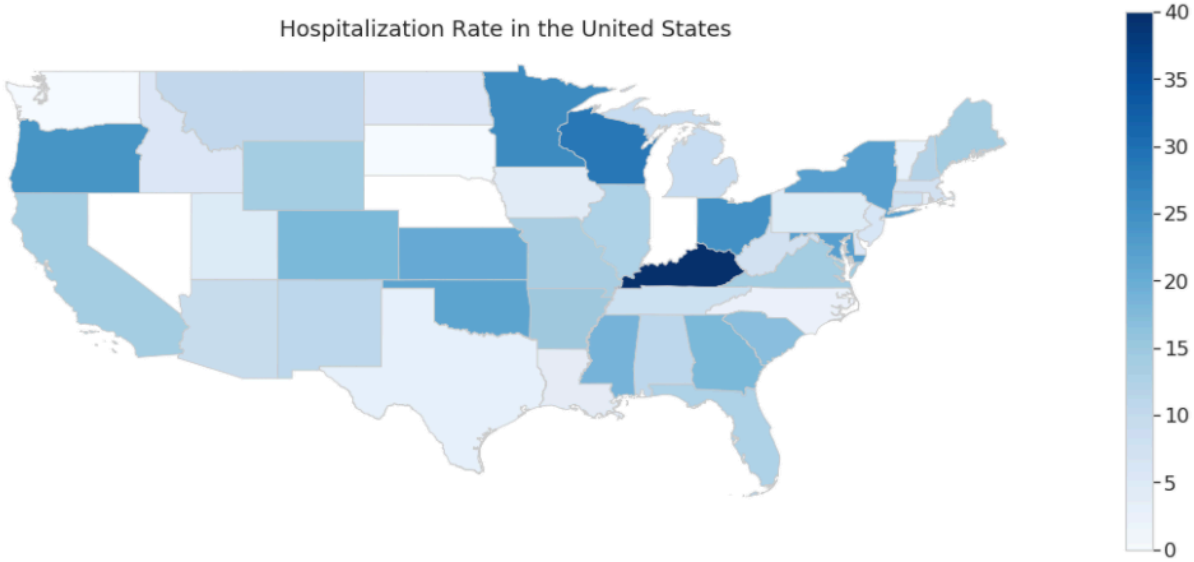


Figure 3: Hospitalization rate by state in the United States

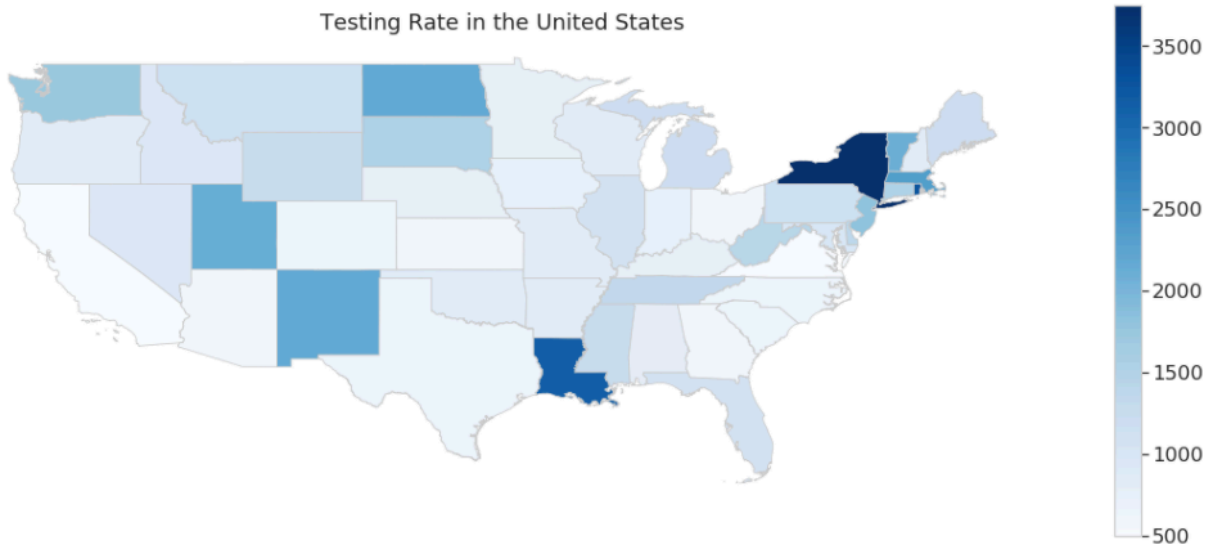


Figure 4: Testing rate by state in the United States

And examining the number of covid-19 cases per day shows similar trends of high initial growth before the number of cases levels out (Figure 5). This is seen across multiple states. Additionally, many states have either not yet reached the max number of new cases or have just started to level out (Figure 6).

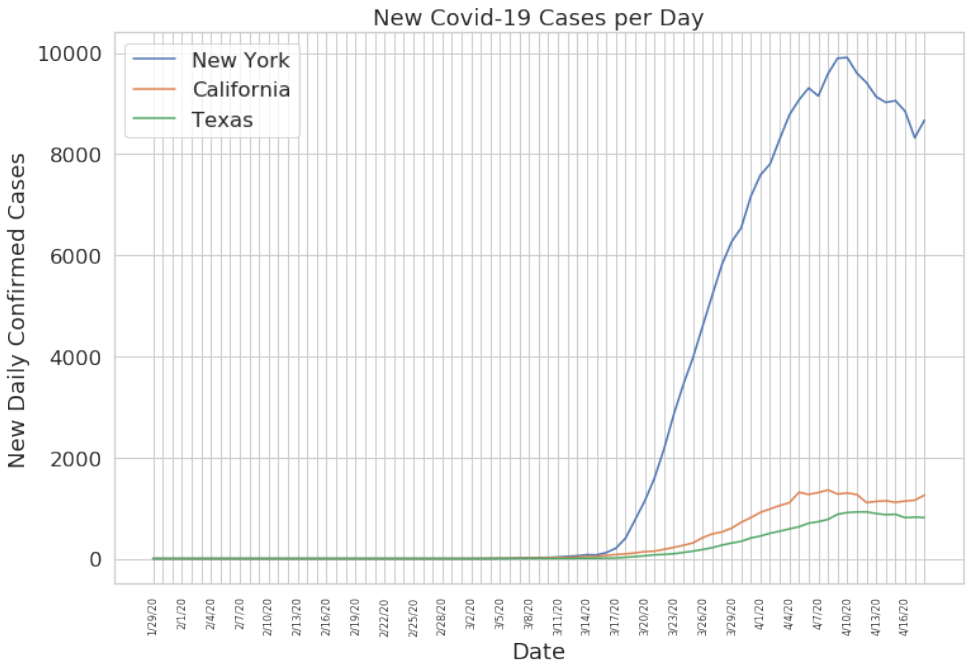


Figure 5: Number of new covid-19 cases per day from 1/22/20 to 4/18/20 in New York, California, and Texas.

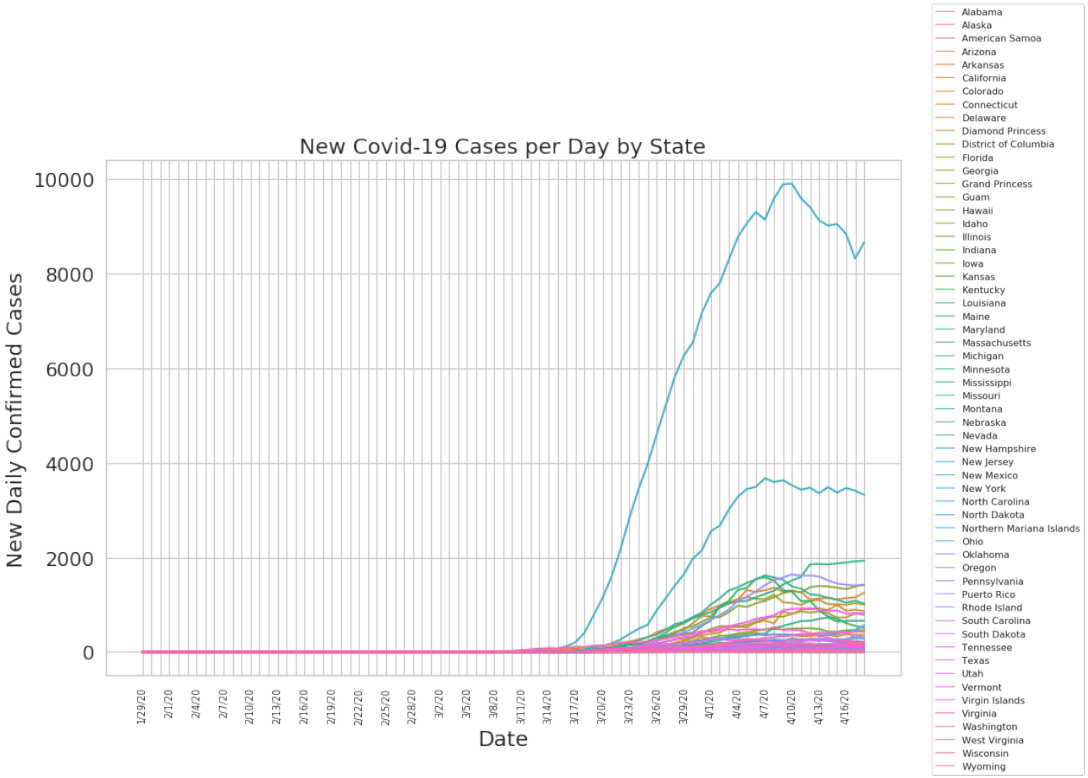


Figure 6: Number of new covid-19 cases per day form 1/22/20 to 4/18/20 by state

Additional exploratory data analysis conducted examined the relationship between mortality rate and hospitalization rate and testing rate. There was no major clustering between hospitalization rate and mortality rate. However, mortality rate and testing rate showed a little clustering near testing rates of 1000. This may be because of the limited capacities for testing (Figure 7).

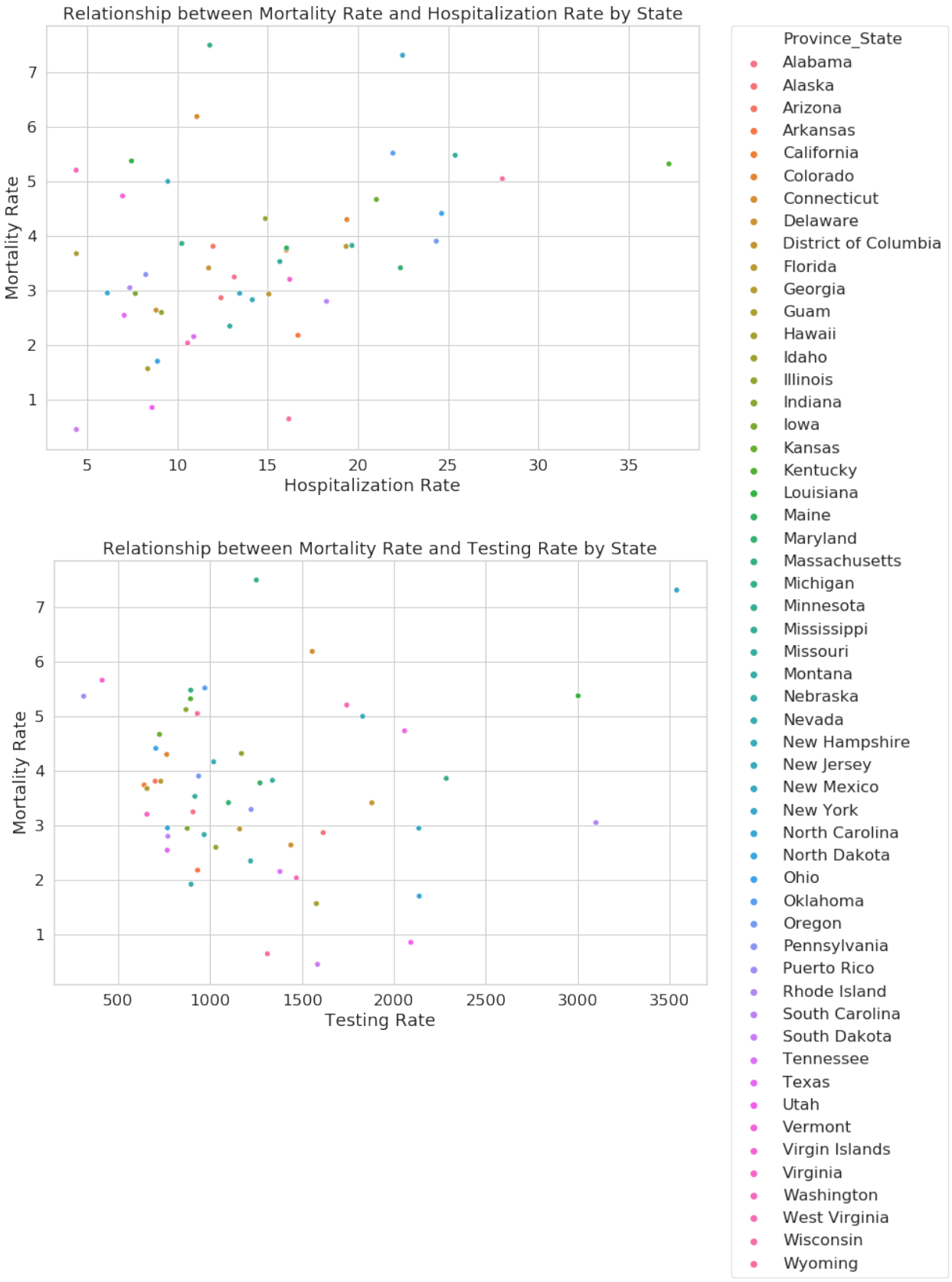


Figure 7: Relationships between mortality rate hospitalization rate/testing rate by state

After understanding the overall picture, I delved more in depth into the state of Washington and used the daily dataset and county dataset to examine the number of confirmed cases, deaths, tests, and additional cases, deaths, and tests per day. Washington being one of the first states to have a case in the United States had a relatively earlier timeline than other states, making it interesting to look at the progression over time.

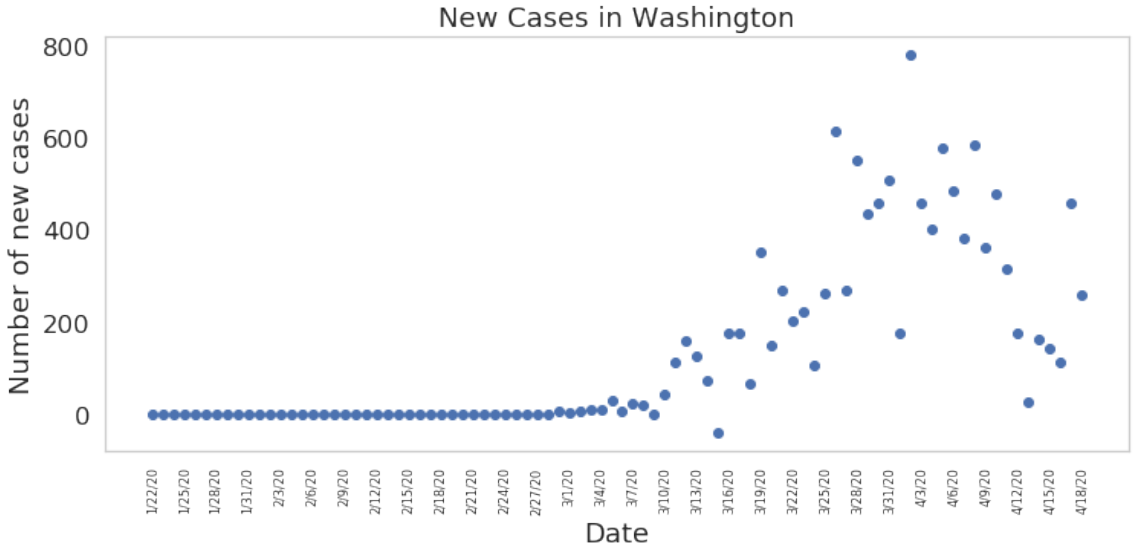


Figure 8: Number of new cases per day from 1/22/20 to 4/18/20 in Washington state

Washington showed a lot of irregularity in the number of new cases starting in March (Figure 8). I attribute this to the difference in cases "discovered" during the week and holidays versus the weekends when more people go in for testing. To fix this, I look instead at the rolling average of new cases over 7 days (Figure 9).

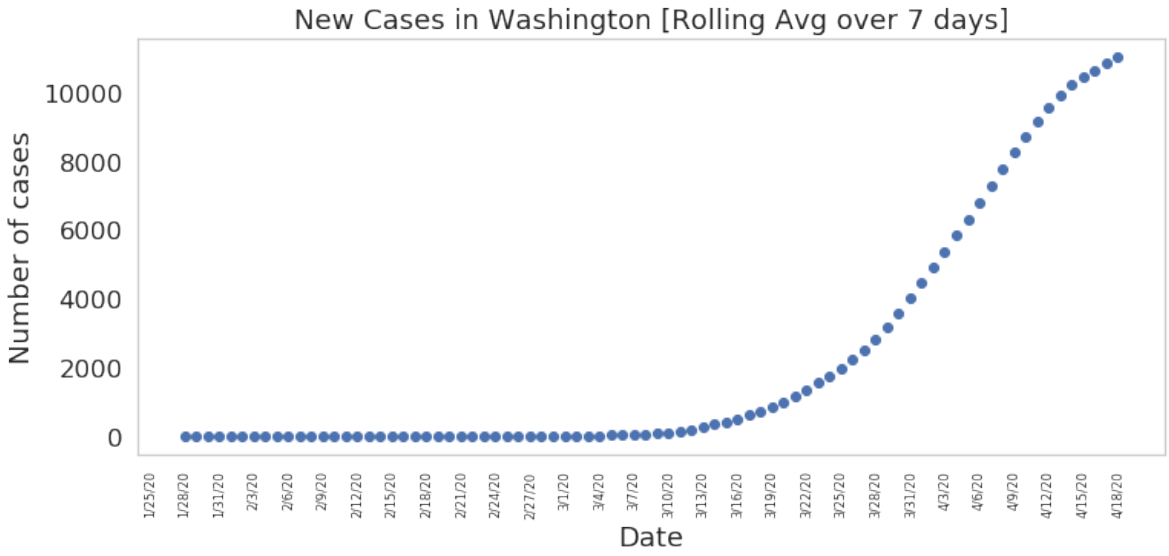


Figure 9: Number of new cases averaged over 7 days from 1/22/20 to 4/18/20 in Washington state

This graph definitely has reduced anomalies, yet the number of cases is probably still correlated strongly with the number of tests performed. As we can see below, in Washington there was a large increase in testing in the middle of March (Figure 10).

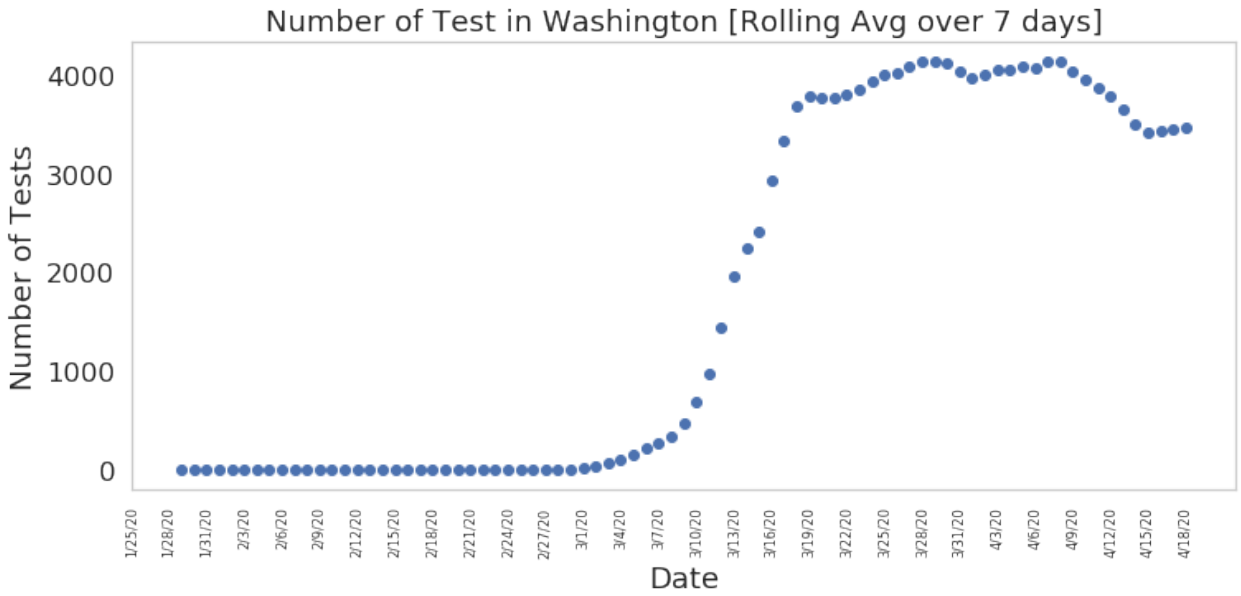


Figure 10: Number of new tests per day averaged over 7 days from 1/22/20 to 4/18/20 in Washington state

Testing biases cannot be eliminated as testing methodologies vary widely across the country. Additionally, current testing is much more widely available with testing of larger populations rather than only people that demonstrate symptoms. This bias shifted the focus of this project to examine the growth of covid-19 as a function of the number of deaths and hospitalizations (Method 2) rather than the number of confirmed cases (Method 1).

Methods

1: Predicting the number of covid-19 cases on 4/18/20 by count:

My first objective was to create linear models to produce estimates of the number of covid-19 cases on 4/18/20 by county and to determine what factors in a county are beneficial in predicting coronavirus cases and how the virus spreads in an area.

Data cleaning & feature engineering:

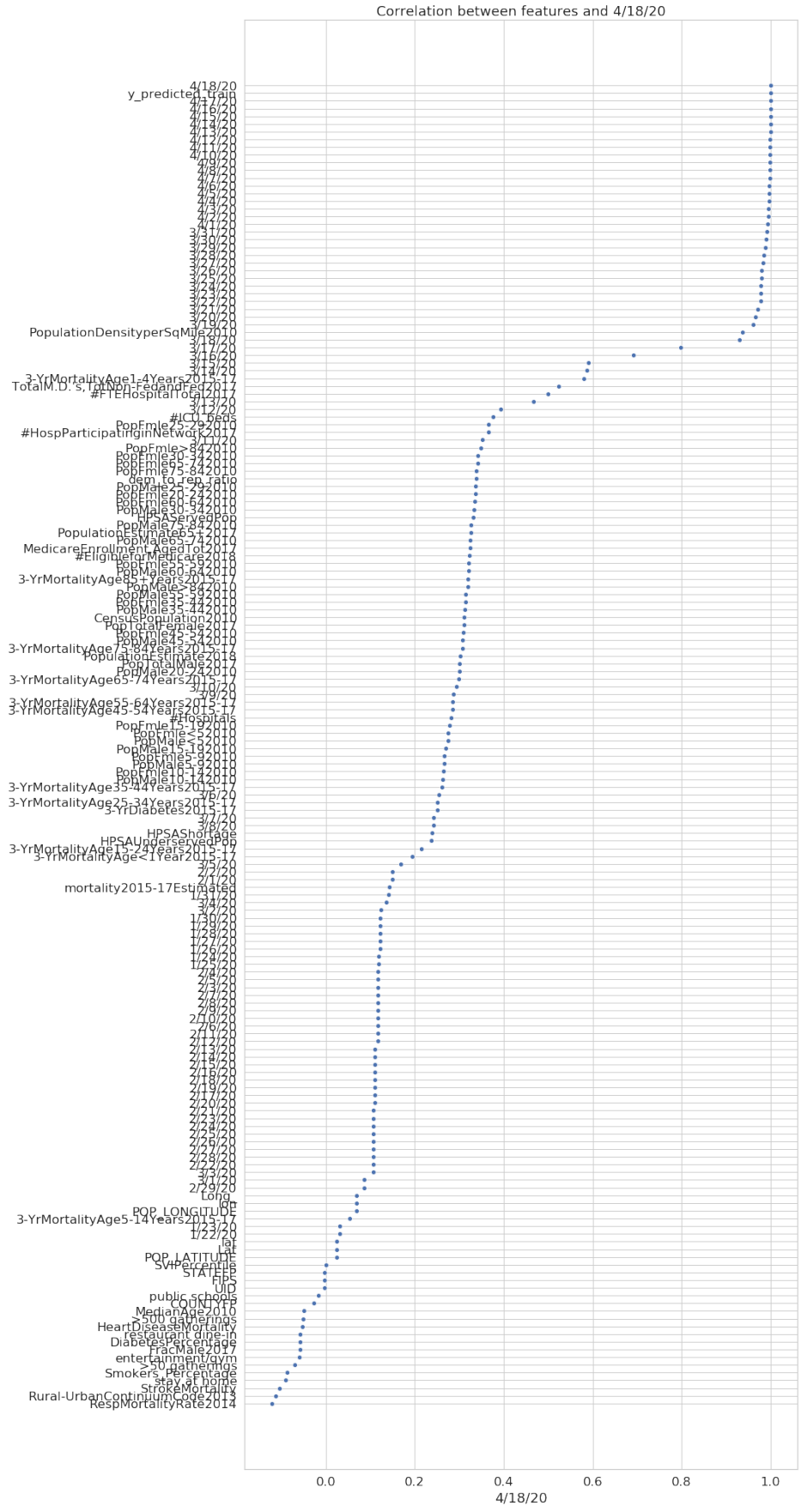
First, I merged the county data with the number of confirmed cases to include both county information and prior coronavirus case numbers as features for my model. Additionally, I created an additional feature by determining the percent change between days per county to see where covid-19 was growing rapidly and slowly. Next, I decided to add the max percent change for each county as the feature in predicting the number of cases on 4/18/20. This is because through my exploratory data analysis, many counties have not yet experienced a decline in cases as of 4/17/20 and those counties are still experiencing rapid growth.

Train/Test split and data processing:
After adding the max percent change feature to the data, I split the county data into a training and test set with a training size of 2108 records and test size of 528 records, and 80%/20% split respectively. In my initial model, I select columns on the dates from 4/1/20 up to 4/17/20 as well as population density, max percent change, and number of hospitals and ICU beds.

I performed linear regression on this data and got a training RMSE of 16.86. Interestingly, cross validation had an RMSE of 66.73.

To improve the features of the model used, I created a visualization that demonstrated the correlation of features and the number of cases on 4/18/20 (Figure 11).

Figure 11 (next page): Correlation between features and number of cases on 4/18/20


Correlation between features and 4/18/20
4/18/20
y_predicted_train
4/17/20
4/16/20
4/15/20
4/14/20
4/13/20
4/12/20
4/11/20
4/10/20
4/9/20
4/8/20
4/7/20
4/6/20
4/5/20
4/4/20
4/3/20
4/2/20
4/1/20
3/31/20
3/30/20
3/29/20
3/28/20
3/27/20
3/26/20
3/25/20
3/24/20
3/23/20
3/22/20
3/21/20
3/20/20
3/19/20
PopulationDensityperSqMile2010
3/18/20
3/17/20
3/16/20
3/15/20
3/14/20
3-YrMortalityAge1-4Years2015-17
TotalM.D.'s,TotNon-FedandRes2017
#FTEHospitalTot2017
3/13/20
3/12/20
#ICU_beds
PopFmle75-292010
#HospParticipatinginNetwork2017
3/11/20
PopFmle>842010
PopFmle30-342010
PopFmle65-742010
PopFmle75-842010
dem_to_rep_ratio
PopMale25-292010
PopFmle20-242010
PopFmle60-642010
PopMale30-342010
HPSAServedPop
PopMale75-842010
PopulationEstimate65+2017
PopMale65-742010
MedicareEnrollment,AgedTot2017
#EligibleforMedicare2018
PopFmle55-592010
PopMale60-642010
3-YrMortalityAge85+Years2015-17
PopMale>842010
PopMale55-592010
PopFmle35-442010
PopMale35-442010
CensusPopulation2010
PopTotalFemale2017
PopFmle45-542010
PopMale45-542010
3-YrMortalityAge75-84Years2015-17
PopulationEstimate2018
PopTotalMale2017
PopMale20-242010
3-YrMortalityAge65-74Years2015-17
3/10/20
3/9/20
3-YrMortalityAge55-64Years2015-17
3-YrMortalityAge45-54Years2015-17
#Hospitals
PopFmle15-192010
PopFmle<52010
PopMale<52010
PopMale15-192010
PopFmle5-92010
PopMale5-92010
PopFmle10-142010
PopMale10-142010
3-YrMortalityAge35-44Years2015-17
3/8/20
3-YrMortalityAge25-34Years2015-17
3-YrDiabetes2015-17
3/7/20
3/6/20
HPSAShortage
HPSAUnderservedPop
3-YrMortalityAge15-24Years2015-17
3-YrMortalityAge<1Year2015-17
3/5/20
2/2/20
2/1/20
2/4/20
mortality2015-17Estimated
1/31/20
3/4/20
3/3/20
1/30/20
1/29/20
1/28/20
1/27/20
1/26/20
1/24/20
1/25/20
2/4/20
2/5/20
2/3/20
2/2/20
2/7/20
2/8/20
2/9/20
2/10/20
2/6/20
2/11/20
2/12/20
2/13/20
2/14/20
2/15/20
2/16/20
2/18/20
2/19/20
2/17/20
2/20/20
2/21/20
2/23/20
2/24/20
2/25/20
2/26/20
2/29/20
2/28/20
2/27/20
3/3/20
3/1/20
2/29/20
Long
lon
POP_LONGITUDE
3-YrMortalityAge5-14Years2015-17
1/23/20
1/22/20
Lat
lat
POP_LATITUDE
SVIPercentile
STATEFP
FIPS
UID
public schools
COUNTYFP
MedianAge2010
>500 gatherings
HeartDiseaseMortality
restaurant dine-in
DiabetesPercentage
FracMale2017
entertainment/gym
>50 gatherings
Smokers_Percentage
stay at home
StrokeMortality
Rural-UrbanContinuumCode2013
RespMortalityRate2014
0.0
0.2
0.4
0.6
0.8
1.0
4/18/20

I then adjusted my model to include additional features about the population demographics (number of males and females between 75-84 and the number of males and females older than 84) and additional dates starting from 3/19/20 until 4/17/20. I then re-processed the train and test data before performing linear regression. The training RMSE was 14.79 and the test RMSE was 65.55. The cross validation test RMSE was 39.94.

Interestingly, I found that processing the data with fewer dates had a higher training RMSE of 16.61 but a smaller test RMSE of 44.07. Additionally, the cross validation test RMSE was slightly lower at 39.85.

2: Predicting total deaths by state

Objective and questions:

While the original goal of this method was to predict the total number of deaths and hospitalizations as a measure of covid-19 growth, the limited amount of hospitalization data in the daily dataset meant that this analysis focused on the total number of deaths. Specifically, can we predict the number of deaths based on hospitalization rate or testing rate? What other features are important?

Training/Test split:

To do so, I examined the states data and split it into a training and test set of 43 and 11 records respectively. My initial model took the mean number of deaths from the training set and used that value as a stand-in for the number of deaths in the test set. Doing so received a RMSE of 660.83.

Linear regression and correlation:

I then conducted linear regression with one feature, the number of people tested, to determine the number of deaths. This had a RMSE of 668.11, higher than the simple model above. Next, I added an additional feature of the number of confirmed cases and ran linear regression. This achieved a training RMSS of 359.98 and test RMSE of 160.61.

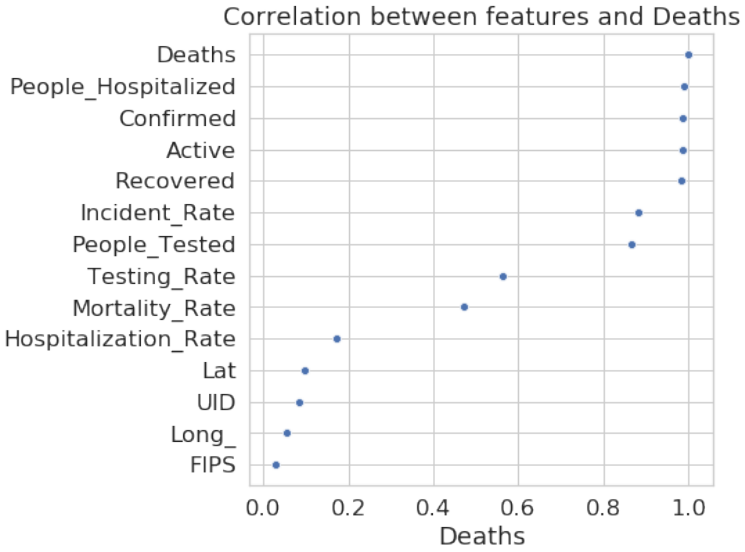


Figure 12: Correlation between features and number of deaths

After determining which features had high correlation with number of deaths (Figure 12), I included these additional features into my model, specifically incident rate and testing rate. Doing so changed the training RMSE to 271.34 and the test RMSE to 255.50.

Results

For predicting the number of covid-19 cases on 4/18/20, the linear model with the additional dates going back until 3/19 had a smaller test RMSE than the initial simple linear model dating back to 4/1/20. Conducting ridge regression with built-in cross-validation on the complex data processing had the highest RMSE. Thus, the most effective model to predict the number of covid-19 cases for 4/18/20 includes dates going back a month and uses linear regression (Figure 13).

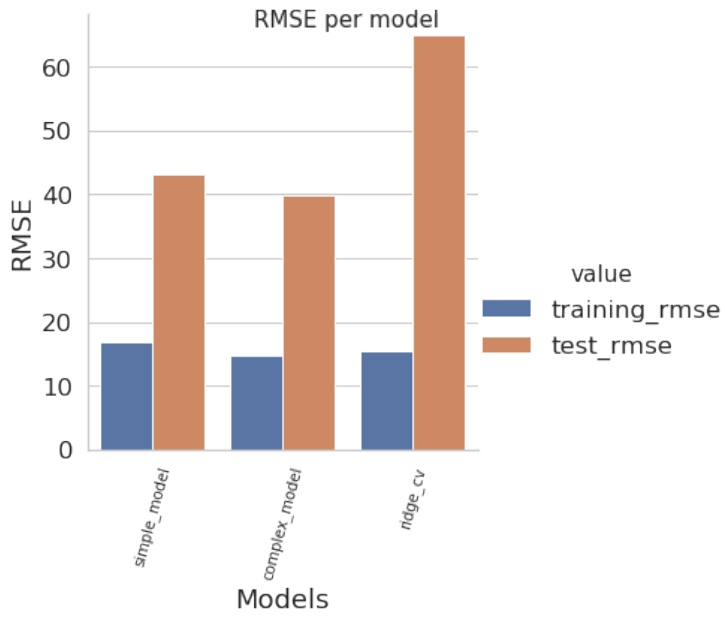


Figure 13: RMSE of the training set and test set for three models. 1) Linear regression on features including dates from 4/1/20 to 4/17/20. 2) Linear regression on features including dates from 3/19/20 to 4/17/20. 3) Ridge regression with cross-validation on features including dates from 3/19/20 to 4/17/20.

For predicting the total deaths by state, both linear regression models had much lower RMSE's than predicting the mean of the training set. Interestingly, the data containing the number of people tested and number of confirmed cases had a higher training RMSE but a lower test RMSE than the processed data that also included the testing and incident rate. This may be because the model with the testing and incident rate over fit the training data and thus performed worse on the test data (Figure 14).

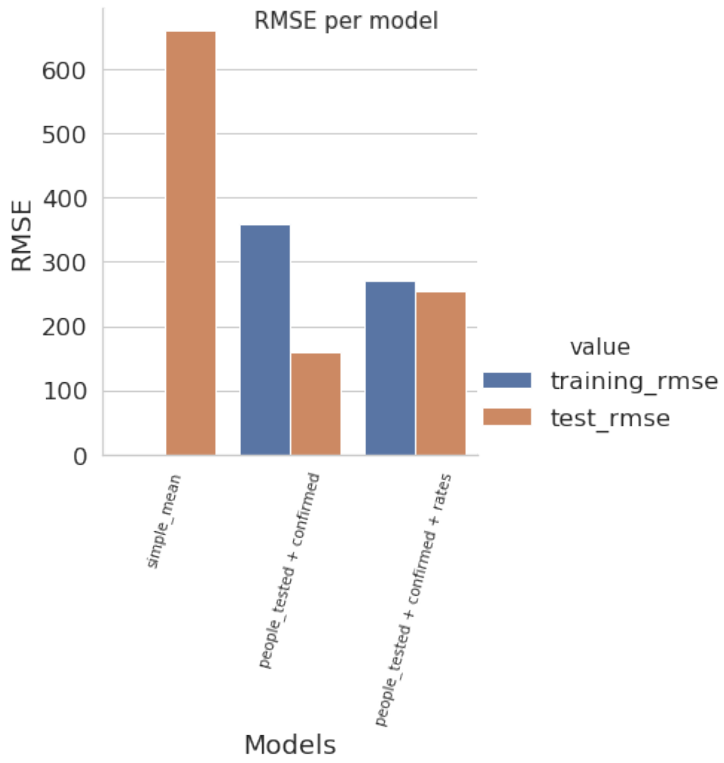


Figure 14: RMSE of the training set and test set for three models. 1) The mean number of deaths from the training set. 2) Features included are the number of people tested and number of confirmed cases. 3) Features included are the features from model 2 and the testing rate and incident rate.

Discussion

Overall, historical trends of corona virus cases are important when considering and predicting the number of new cases on a given day. Additionally, predicting deaths is a more accurate estimation of the changing nature and growth of the virus compared to the number of confirmed cases, which is limited by testing ability.

The features that I found most interesting when predicting the number of covid-19 cases on 4/18/20 was the population density per square mile in a county as well as the total number of medical doctors in a county. The first is an aspect of how corona virus spreads while the latter considers how healthcare professionals matter in stopping its spread. For my other question, predicting the number of deaths, I thought that including the testing and incident rate would be helpful to know the number of deaths due to covid-19. Interestingly, this was ineffective and rather over fit the data.

A challenge that I initially had with the data was determining how to best deal with the time-series aspect of the confirmed cases and deaths data set. I decided to look at percent change as means of measuring growth as well as using the dates as features for my model. Doing so is a potential limitation, as there are additional ways I could have worked with the time series

aspect of the data to create a better model. Another limitation is that I used the maximum percent change as a feature, assuming that it would be an accurate indicator of the growth of the virus in a county. In the future, I want to also include the average percent change.

An ethical dilemma I faced was considering the outcome of my results. Being able to predict the number of covid-19 cases in different areas around the country could help provide needed PPE to those areas. However, it also prioritizes certain areas over others, potentially areas that are primarily people of minority or low income backgrounds and that have to continue to work. To address these concerns, additional analysis should be done that examines the places of high risk and their demographics. Additionally, this research aims to guide decisions regarding coronavirus. An ethical dilemma is to what extent such work could be used as the sole decision-maker to ration critical care resources and support.

Additional data that could be included is more information regarding hospitalizations and the severity of hospitalizations, such as the number of patients in the ICU and on ventilators. This could help support the data analysis stream of determining the growth of covid-19 not just on confirmed cases but on a combination of hospitalizations and deaths.