Effects of Mobility on Spread Rate of Covid-19 During Quarantine in the U.S

Carlos Yu, Angel Zhou, Joshua Zhang

Research Questions

1. Which location in the community contributes the most to the spread of the virus?
   Given the mobility report, we are curious about how people in the US behave differently according to the stay-at-home order imposed on most of the states. And most importantly, we want to find the direct influence of quarantine on the prevention of infection.
   Result: we've found that among the 6 locations (parks, residentials, grocery and pharmacy, retail and recreation, transit stations, and workplaces), parks are the one that match the closest to the daily new confirmed cases, and its correlation also confirms that it is most related to daily new confirmed cases.
2. Do different states in the U.S have similar growth rates on daily confirmed cases? What about the factors of the growing number of cases?
   We will divide the states into two groups based on their positive cases. We will then plot a graph for each state and compare them within the group. For the early stages of the pandemic (Feb. & Mar.), we will try to find growth rates for each state since the early spread of the virus should follow an exponential model. We will then analyze the possible factors like public health coverage and mobility change that contribute to the bigger trend and smaller trend by more graphs.
   Result: The growth rates among the most states show similar trends that can be applied to both largely and lessly affected states. We found health coverage and population to be the critical factors when mobility change doesn't contribute to the increase of cases in the early stages.
3. How will the number of daily confirmed cases in the U.S change in the near future?
   We will use machine learning to predict the number of daily confirmed cases in the U.S based on the major influencing factors we found in question 1. We plan to predict until the end of this course and we could calculate the error between our prediction and the real number along the way.

Result: The daily new cases will oscillate slightly, but the general trend should ease for the next few weeks, which means the daily new cases wouldn't grow or descend too much if the current policy (mask, public gatherings, etc.) stays the same. In other words, the current daily new cases is at its stability.

Motivation and background

Starting from the beginning of 2020, Covid-19 is a serious problem to human health worldwide. Among all states in the USA that held quarantine since March, Washington state successfully tested potential cases and prevented the massive spread of Covid-19 when Snohomish County had the first case of this novel coronavirus and deaths. In August, no more than 100 cases are newly discovered in WA every day (https://coronavirus.1point3acres.com/). However, Covid-19 didn't stop in the USA because of the quarantine, while total Covid-19 confirmed cases are reaching 5 million. Why is quarantine work in WA but not some other areas? We hope to find the possible factors influencing the function of the quarantine such as population density and health coverage in the states. We want to find the most essential elements for public health in the background of a global virus crisis threatening humans. Ultimately, our goal is to use these information, including the daily confirmed cases and the biggest influencing factors, to build a model that could predict the trend of future number of Covid-19 cases in the US.

Dataset

A. COVID-19 Stats and Mobility Trends:

This dataset includes all changes in the Covid-19 daily spread rate and surrounding environmental factors (geography, population, public health coverage, etc.). The dataset includes information from 133 countries worldwide over time (2020/02/15 - still updating). It's valuable for finding how quarantine works (decrease on mobility) and their effects on spread rate.

URL:https://www.kaggle.com/diogoalex/covid19-stats-and-trends

B. COVID-19 in USA:

This dataset includes information about Covid-19 in the USA that includes daily testing rate and infection rate by states and cities. More importantly, it tells how many patients were in the hospitals and were using ventilators (an indicator of the severe cases). As a huge source of Covid-19 information, investigating and comparing its influences on individual states will locate the largest population and the fastest growth rate of Covid-19 in the USA.

URL:https://www.kaggle.com/sudalairajkumar/covid19-in-usa

C. **United States by Density by Population 2020:**

This dataset is given by US Census State Population Estimates and includes population, land area, and population density of 51 states (including Washington D.C.) in America in 2020.

URL:https://worldpopulationreview.com/state-rankings/state-densities

D. **United States Geodataframe:**

This GeoDataframe provides geometry information about each state. It's good for constructing USA map by states or counties

URL:https://alicia.data.socrata.com/Government/States-21basic/jhnu-yfrj

E. **USA Google Mobility Reports by states**

Dataset by google with mobility change by each state on visiting parks, grocery stores, transition station, etc.(February to April) All data is in percentage change form.

URL:https://www.kaggle.com/mikalainis/covid19-google-mobility-reports?select=COVID19_Google_Mobility_Report_US_State.csv

F. **USA Hospital Information**

Dataset of all USA hospitals in 2019 with information of addresses. Data sources are from US government(Homeland Infrastructure Foundation-Level Data (HIFLD))

URL:https://www.kaggle.com/carlosaguayo/usa-hospitals

Challenge Goals

1. **Multiple Datasets**: We will be using mainly three datasets (with three additional ones) for our project. The three main dataset are as follows: Dataset A will record different features that might contribute to the spread of virus and the total confirmed cases by country, which would allow us to verify our results of the U.S on other countries; Dataset

B records the daily-confirmed cases for each state so we could group them by population density and compare them separately; Dataset D has a population map of the U.S, which helps visualizing our results on Question 2. For question 2, we joined B and D to depict the level of seriousness by states. In addition, we used C, E, and F to further analyze the relationship between infection rate, number of hospitals, and mobility trend. On Question 3, we joined A and B to predict the U.S's future daily confirmed cases based on the previous numbers and the major influencing factors that we found from Question 1.

2. **Machine Learning**: We used a decision tree regressor to solve our third research question. We used the results from Research Question 1 as the features to predict the future daily confirmed cases in the U.S. Compared to the machine learning we did in class, this time we manually splitted the dataset into a training set and a testing set since the data are in chronological order, and it only makes sense to predict the future using the past data, not the other way around. Besides, we also manipulate two hyperparameters, the max tree depth and the number of features included, to optimize our model. We also tried different ratios of training/testing data to see how the predictions are different at different stages of the pandemic.

Methodology

Question 1: First, we read in and explore the dataset A, since it contains a lot of extra information, we need to filter it down to the rows and columns we need. To answer this question, we will plot the mobility trends over time for each location and compare it to the plot for daily new confirmed cases over time. We will also calculate the correlation between each mobility trend and the daily confirmed cases to see if they are actually related.

Question 2: The essential part of this question is filtering data for each Covid-19 related state in dataset B and finding the total COVID-19 case in each state from the latest date. Then the number for each state will be on a plot by Geodataframe from the dataset D to show the level of seriousness of these states. Knowing where the largest and lowest cases are, we will pick 10 states with the largest numbers and 10 states with the least and make two scatter plots for them by each day's cases (test only, not including death since we focus on infection rather than treatment). This is the place where we stop with the trend of cases and start looking for factors. We will divide the numbers of hospitals by population of the twenty chosen states from dataset F (in our estimation, fewer people per hospital means more choices on health care and less chance of infection) and the mobility change from the dataset E by averaging all types of mobility change. Population density from data frame C is also critical for considering. We will join these three sets of data and plot them into one bar graph. We will analyze these 6 graphs by our estimations. A state with a big infection trend should have fewer hospitals by population density and less mobility change when the one with a smaller trend does the opposite.

Question 3: Answering this question, we joined dataset B with a filtered dataset A. The model was built on the scikit-learn library. The features come from Dataset A and are mostly based on results from Q1, and the label will be the daily confirmed cases from Dataset B. Since the data are in chronological order, the training set is always the early, consecutive section of data instead of randomly chosen among the dates. We manipulate different aspects of this model, including hyperparameters like the max depth of the decision tree and inclusion/exclusion of some features, to optimize our model. We set a range for both hyperparameters and tried every combination of them to find the pair that yields the least error. Our output is a double-line graph showing the real data and the model's predictions of the future cases.

Result

Question 1: Which location in the community contributes the most to the spread of the virus?

By examining *Figure 1.* below, we can clearly see that after the surge of COVID-19 cases since day 30, the curve began to flatten out starting at day 50 until day 120. During that period of time, many state governments had imposed different kinds of stay-at-home orders depending on the situation in different states in attempts to contain the virus. So we suspect that quarantine is effective in combating COVID-19.

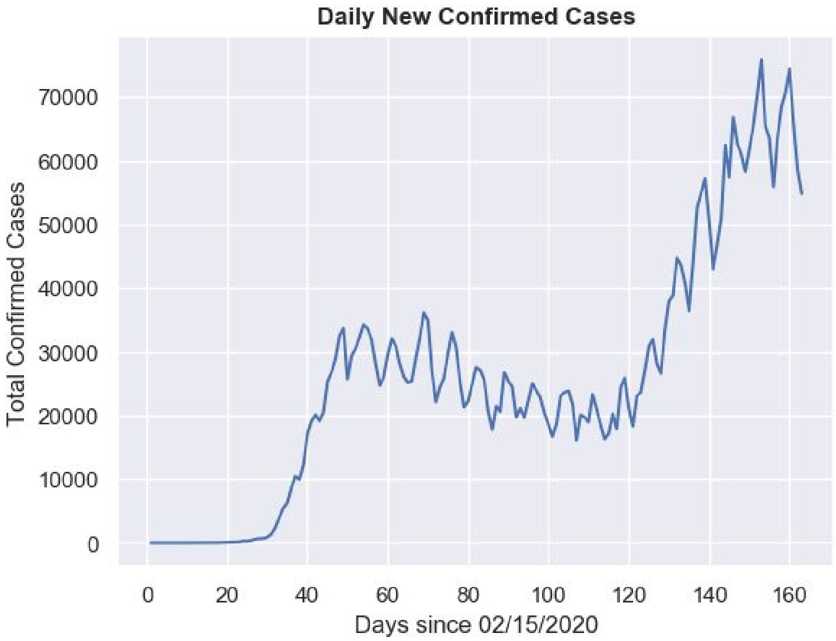

Figure 1: U.S. Daily New Confirmed Cases Over Time

We then want to further analyze how peoples' mobility changed at 6 different locations in response to the stay-at-home order and how that affects the number of cases. According to figure 2, we observe that there are some kinds of plunges to different degrees for all 6 locations at around day 20. Then the curves begin to rise again at around day 60. (Except for the residential graph, because higher percentage means staying at home more, so we should interpret it as if it is upside down) We also observe that Residential, Retail and Recreation, Transit Stations, and Workplace looks very similar in a sense that they all fluctuate a lot during its rise at around day 60, and the range of the rising portion is about 30 percent, which is relatively lower compared to Parks (about 120 percent). And they never return to the level before the plunge whereas the percentage for Parks surpasses the percentage before the plunge and keeps soaring at around day 100. For Grocery and Pharmacy, there was a sharp increase just before the plunge, then the number started to rise at a moderate pace (that's similar to the bottom 4 plots) at around day 60 but returned to the level before the plunge.

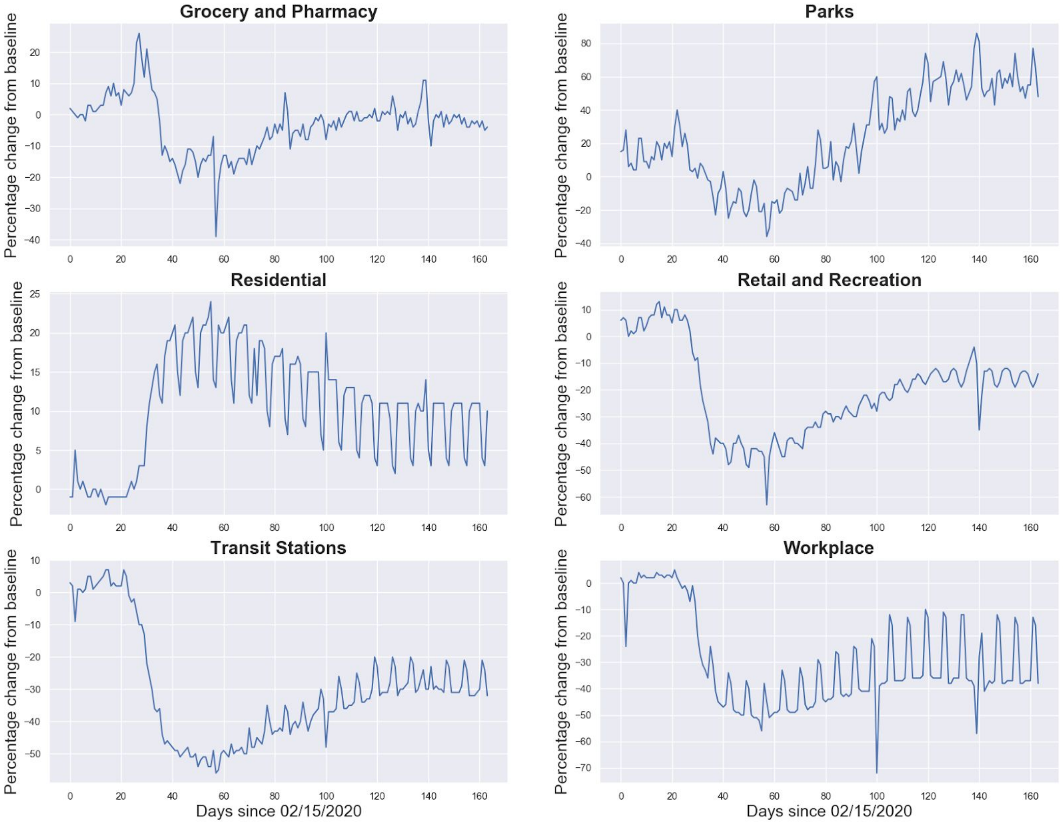

Figure 2: Mobility Change from Baseline Over Time for Various Locations

We also calculated the correlation between daily confirmed cases with the mobility percent change to see if they are actually related to each other. We get [('parks', 0.4809), ('residential', 0.2669), ('grocery', -0.2352), ('retail', -0.2533), ('transit', -0.3701), ('workplaces', -0.4337)]. Since Parks have the highest positive correlation coefficient, we can say that mobility changes in Parks are positively correlated with the daily new confirmed cases. However, high correlation does not imply causation, instead, it means the higher percentage change in mobility in Parks, the more daily new confirmed cases.

Question 2: Do different states in the U.S have similar growth rates on daily confirmed cases? What about the factors of growth rate?

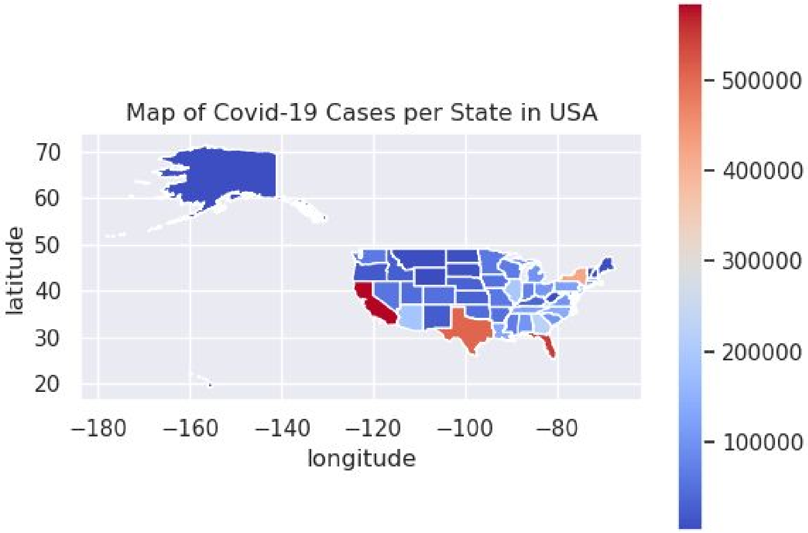

Figure 3: Map of Covid-19 Cases per State in USA

Considering the heat map with all positive cases in USA states on 8/11, we concluded that most states in eastern coast and states in the southern western coast had a serious level of Covid-19. However, only 4 states reported more than 30 thousands cases(50% point of the biggest number) and more states are around 10 thousands cases(25% point). Thus, we estimate different states have similar growth rates when there are factors like random beginning cases, and population density's influences .

Figure 4 below further proves our hypothesis on the growth rates. We calculate the difference between positive cases at the beginning of May(one month after most states announced quarantine) and the beginning of August. For ten states with the least cases, the average number of cases grows from about 15 hundreds to 5 thousands which is 149% each month. For states with the most cases, the average number of cases grows from about 75 thousands to 500 thousands which is 167% each month. The average growth rates between these states are only different by 12%. Most states follow similar trends of growing cases after quarantine.

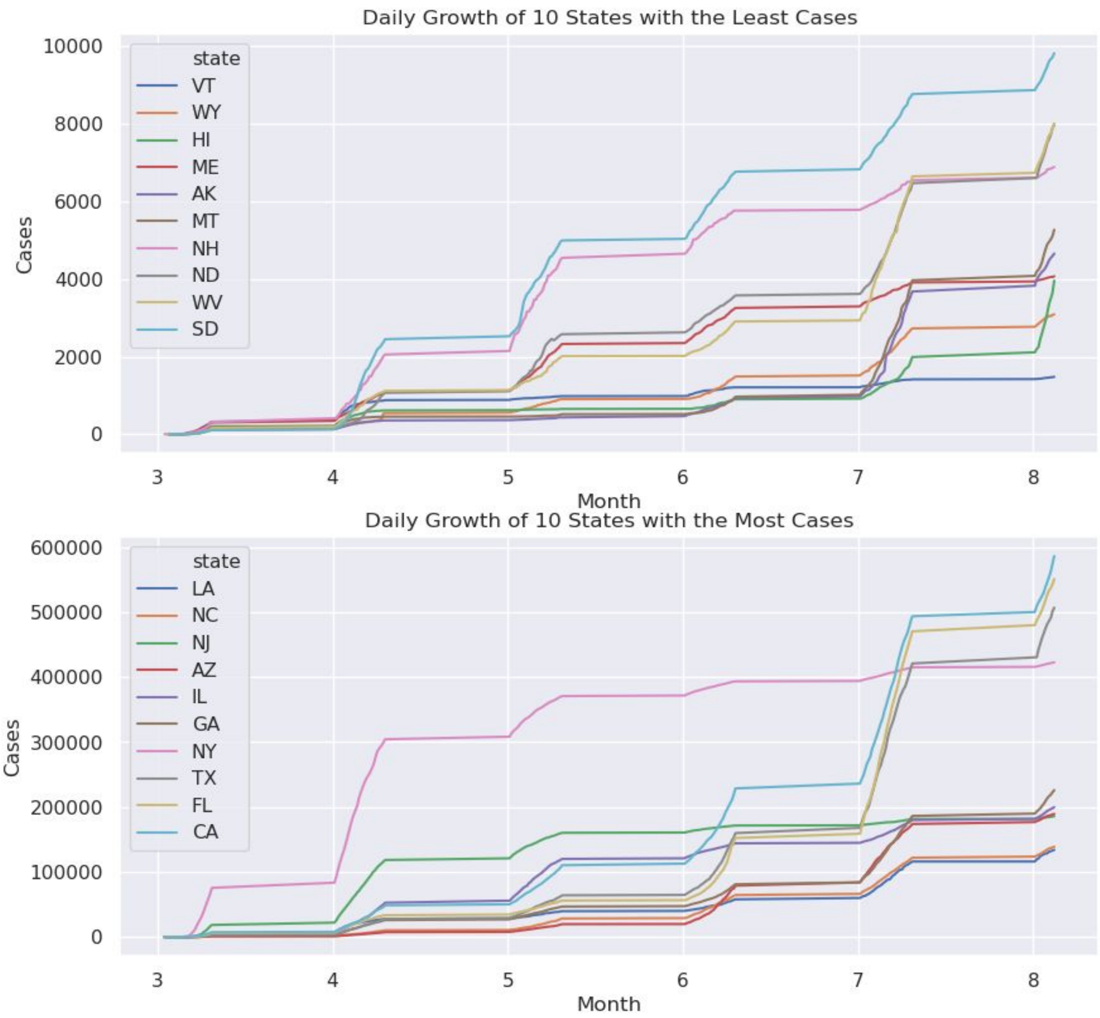

Figure 4: Daily Growth of 10 States with the Least and Most Cases

However, it's still strange to see the number of cases grow in such a trend when their beginning cases are varied largely because the spread of virus is not like bacteria growth in well prepared lab devices. People as the spreader of Covid-19 also reveal what are the reasons the cases grow in those ways. We look at factors such as population density of states, proportion of total population to number of hospitals, and mobility change during quarantine to be the major factors of increasing rate.

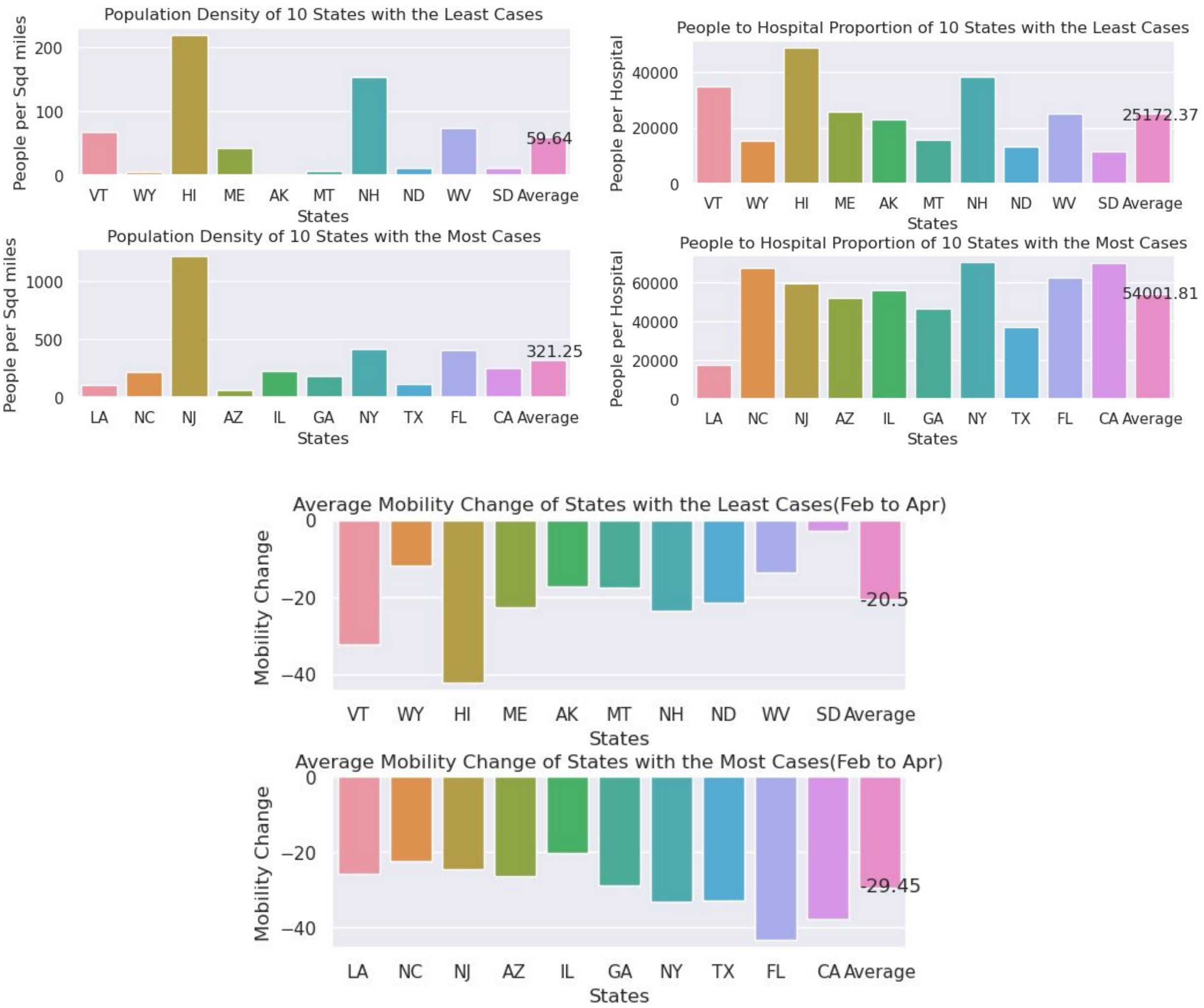

Figure 5, 6, 7: Different Factors on Growth Rates of Different States

Population density shows extremely different features among two groups of states. Least cases group's average is 59.64 people per square miles and the most cases group is 321.25 people per square miles. The difference is 439%. Population density definitely influences the growth rate by bigger sizes of communications and cross infection. In addition, smaller population density means fewer people will have to share health care resources. Least cases group's average proportion is only 46.6% of the average proportion of most cases group. Each hospital in least cases can hold more percentages of patients for treatments and decreases the chances of infection since most patients are separated correctly. Mobility change is 30% different when the least cases group has the lower difference in the period of massive quarantine(Middle March and April) which is against our estimation. This result can mean two possibilities:

Covid-19 has been spread largely before quarantine when these patients were not diagnosed until they had already spread the virus; quarantine policies that cannot be forced in states with high population densities are less efficient than mild level policies in low population density states.

Question 3: How will the number of daily confirmed cases in the U.S change in the near future?

The figure below contains two lines, the orange line shows the program's prediction of the daily new Covid-19 cases in the U.S, while the blue line, representing the real trend, was added to the graph after we received the data recently. As can be seen from the graph, our prediction doesn't quite match the real data in shape, but subtly lies in the middle of the violently-oscillating blue curve. If we stretch the timespan long enough, these oscillations will become minor noises, making the trend roughly a line slightly tilting downwards (just like the period from 04/06 to 06/20). In terms of the overall trend (or pattern), we are confident that our model predicts the trend pretty accurately.

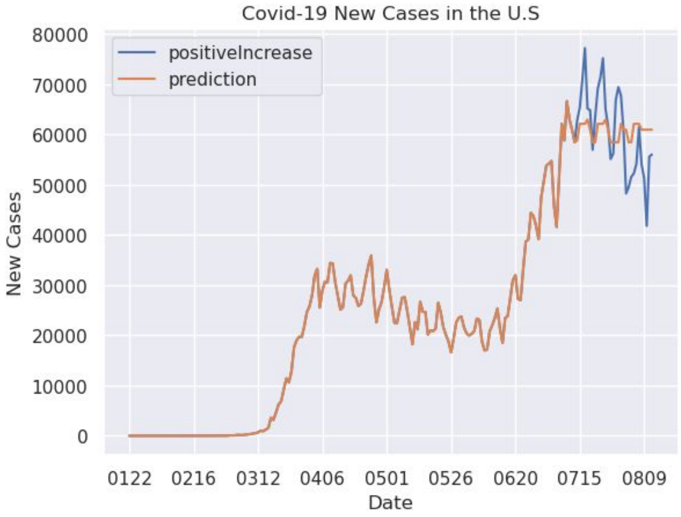

Figure 8: Daily New Cases in the U.S.(85% data)

But note that in the above graph, the amount of prediction made is relatively small compared to the total timespan since we are only predicting 15% of the data. We are interested in knowing how the model will behave when the prediction starts at the middle of an increasing interval (i.e. the period between 06/20 and 07/15). Therefore, we decreased the volume of our training data from 85% to 77.5%, which makes the prediction start in the designated interval. The resulting figure is below. Note that either time, we optimize the model so it will have the decision tree depth and the number of included features that produce the least mean squared error.

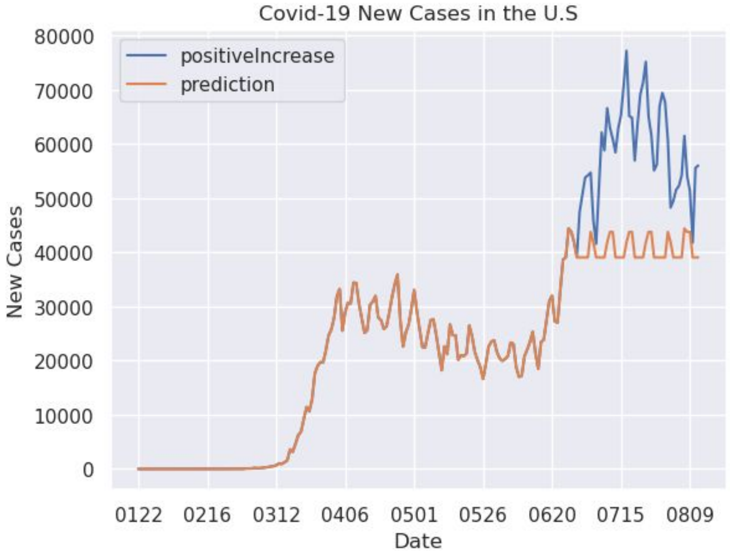

Figure 8: Daily New Cases in the U.S.(77.5% data)

Initially, we were expecting our model to continue growing at a high rate instead of leveling out at the peak, since there isn't any indication that the second 'easing' period would occur. But instead, the prediction curve immediately levels out and starts oscillating. This is probably due to the decision tree regressor that we used to predict the data. It would take into account all the previous curves' behavior, and since the 'easing' periods took up the majority of the graph, it makes sense that the regressor would try to make the prediction consistent with the previous graph.

From the blue curve (real data), we are seeing an exponential growth from June to mid July that's similar to the initial outbreak of the pandemic. This happened after the reopening of some states, which leads to the increase in crowd gathering in public spaces (parks, grocery, etc.). Based on the results from Research Question 1, these are the primary factors contributing to the spread of the virus. However, from our prediction which used the data during the middle of this growth, we could see that if the reopening ceases in the middle, meaning gatherings or so in public spaces are no longer encouraged, the growth curve should ease and won't continue growing exponentially. Both graphs are reflecting this behavior, so we could conclude that NOT reopening the states when the pandemic is still at a high level will ease the situation.

Work Plan

- Loading, parsing, and cleaning data (Estimated 2 hr | Actual 2 hr)
    - Create a script that would retrieve the data, parse it, and save each dataset into local Dataframe/Geodataframe variables.
- Creating plots and calculations (Q1 & Q2) (Estimated 6 hr | Actual 7 hr)
    - Further filtering the data into only a few columns
    - Making different types of plots for the first two questions from massive data to find correlations.
- ML modeling (Q3) (Estimated 5 hr | Actual 7 hr)
    - Filtering and splitting the data for training and testing
    - Plotting the predicted data
    - Altering hyperparameters and features to optimize the model
- Testing (Estimated 2 hr | Actual 2 hr)
    - Q1 & Q2: use assert_equals to check if the filtered data is in the correct size
    - Q3: newly reported daily cases will be compared with our predictions. Plot difference will be used to determine the error.
    - Each group member works on one question in one python file. After finishing, importing and testing them together in a single main.py file.
- Analyzing result (Estimated 2 hr | Actual 2.5hr)
    - Each group member will write a report for his or her research questions based on graphs and models.

Evaluation:

We follow the work plan very closely except for underestimating the level of difficulties for some tasks. For example, when creating plots, it takes time to experiment with different types of plots and their parameters. Analyzing the result and putting everything together also takes more time than we expected. We find it hard to organize the result and the graphs in a way that would make the most sense to our audience. We also make adjustments to our plan along the way, for example, we originally planned to test the first two questions by comparing our graphs with external sources, however, we could not find a way to test graphs. Therefore, we decide to instead test if the DataFrames we used to plot are in the correct size.

Testing

For the first two research questions, we mostly produce graphical analysis, so it is hard to find a way to test graphs. Instead, we used assert_equals to make sure data filtering is done correctly so that we use the right DataFrame to plot. Since there are only a few columns left from the filtered data, we could simply list the expected columns out and compare it with the actual data frame. For research question 3, since the machine learning code is relatively standard, our tests mainly focused on data filtering and model optimizing. In addition to a data filtering test similar to the previous ones, we also tested that our combination of hyperparameters produce the least error. We feed the model with some different hyperparameter values in the range and test it on whether the yielded error is always larger than the optimized one.

Collaboration

We did not collaborate with anyone else other than the course staff. We did look up online (e.g. StackOverFlow) for some python documentations and special uses of functions that are not covered in class.