Minkwan Kim

DATA 100

Final Project

05 May 2020

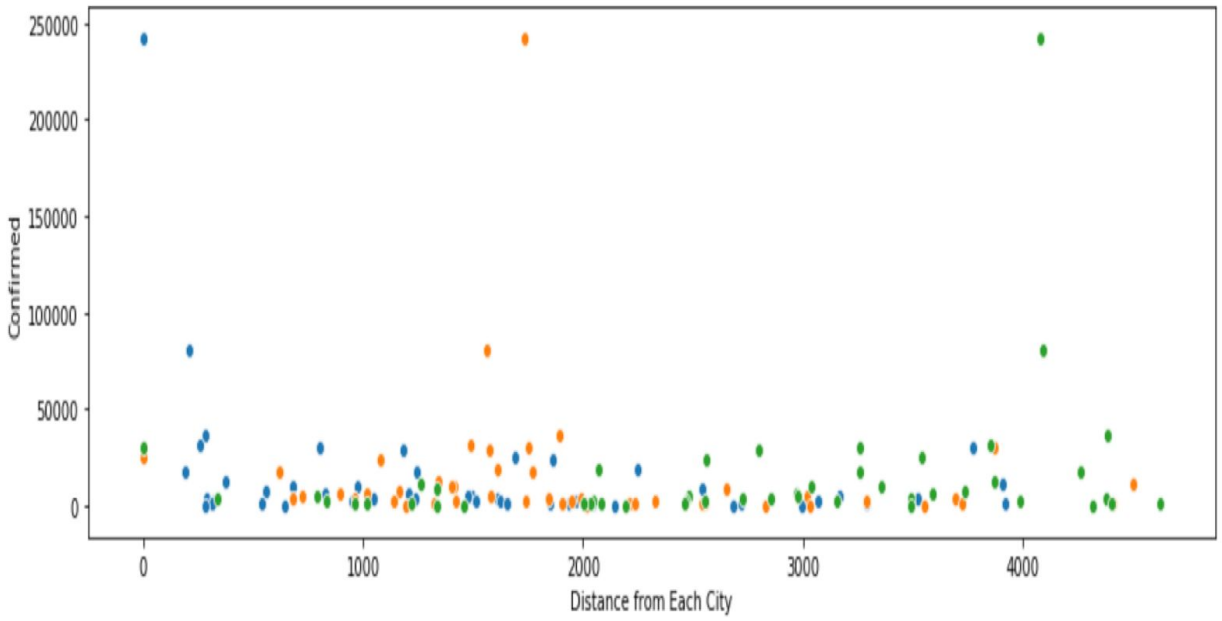
Contagion According to the Distance Between States

Research on covid19 has been actively conducted these days. These include predictions of the number of infected people in the future, and research on an increase in infectivity according to the total number of infected people. There were a few topics I think are interesting. Those are the number of deaths, confirmed, and checked according to the number of hospitals compared to the population, and infections caused by floating populations among states. While collecting data on such a topic, I focused on some news that has a certain topic, which is about a blockade. The disease, which began in a city in China, spread, and the Chinese government blocked the intercity movement. As a result, it has been reported that the number of infected people has decreased significantly. Countries that blocked borders earlier, such as Taiwan, or those with less Chinese influx, have relatively few infected people. When there is no infected person inside, if there is little external influx, this is natural because infectiousness increases if there are more infected people. To verify this with data, I would like to look at the relationship between the number of infected people of a state and the distance between the state and near states. If the predictions are correct, if a state has the shorter distance from states with a high percentage or number of infected people, the number of infected people would be high. Also, if the correlation is sufficient, we can create a reliable prediction model.

In the first step, I loaded ‘4.18states.csv’ from the covid dataset. The data includes information about states in different countries and the number of confirmed, recovered, and active people in each state. Using data from all countries, there will be too many variables, which in turn will yield inappropriate correlations and prediction models with the variables. Therefore, I will limit information to states within the United States. Next, I loaded ‘abridged_couties’ from the dataset because I thought not only the number of infected people,

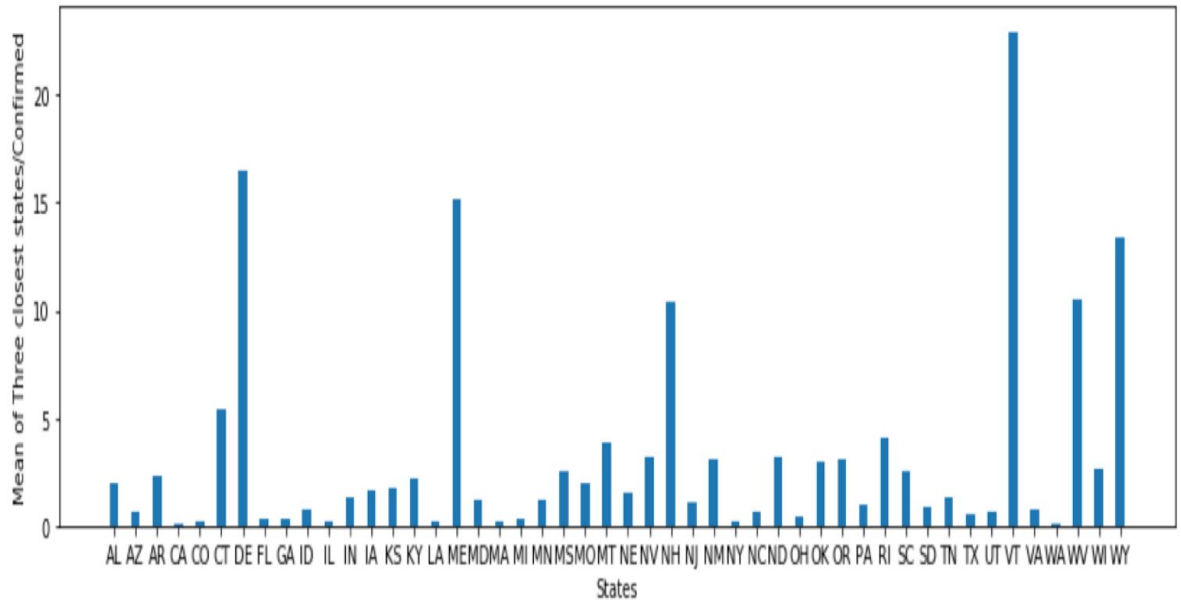
but also the proportion of infected people is important information. ‘Abridged_couties’ contains data about each state’s city and its information like estimated population.

I sorted the ‘Confirmed’ values in the highest order, and selected three states that were separated by distance (‘California’, ‘New york’, ‘Florida’). Then, I got the longitude and latitude of each state, and the differences between the longitude and latitude of each state were set to variables. I created a new data frame with those variables and population, and another data frame with the number of confirmed people. I thought the linear model using the differences of longitudes and latitudes, and the state’’s population would make great predictions. However, it doesn’t seem to work well (rmse was 30000). I needed to produce more suitable variables.

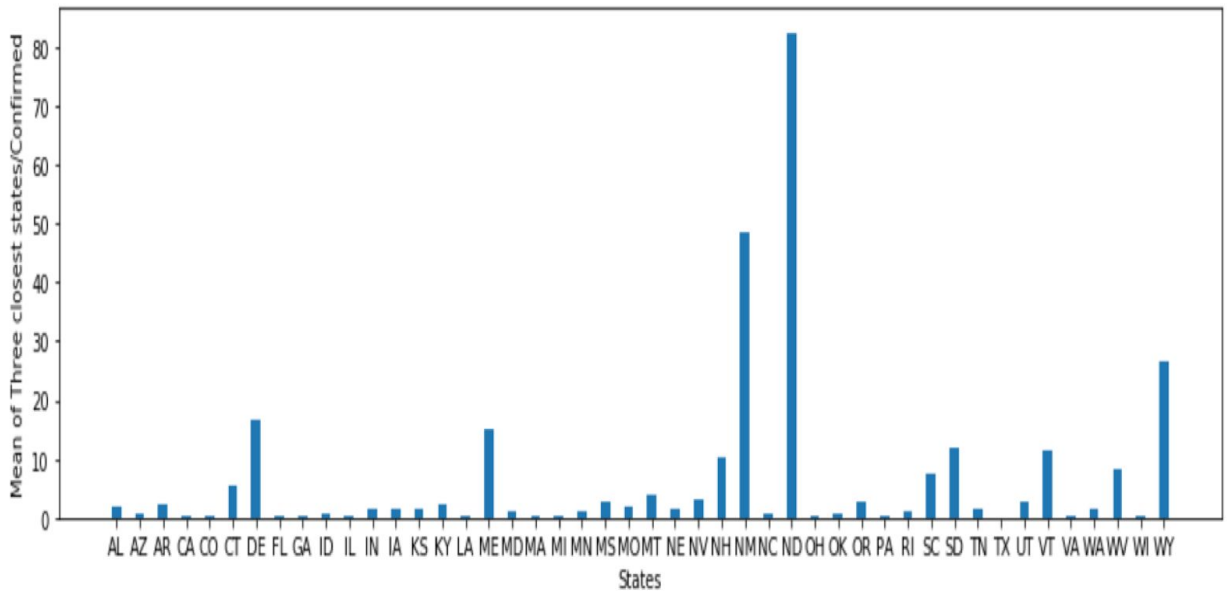


It shows the number of confirmed people by distance from three states I have set up (California, New York, Florida). The dots at zero are the states (California, New York, Florida), and if dots with the same colors are at a close distance and have a similar confirmed number, then my assumption was correct. However, the graph looks ambiguous.

The variables of distance from three designated states seem to tend to interfere with each other. Therefore, I wanted to try out a predictive model based on three states that have the shortest distance from one state. The three states' distance and the number of confirmed people would affect the one state, so with those variables, we can predict the state''s number of confirmed people. When comparing the mean values of three neighboring states with the 'Confirmed' of one state, some exceptions were found. It was that if there is a large number of confirmed people in the states (like New York and New Jersey), the mean value rises tremendously, meaning the comparison would be lost. I had no choice but to drop the two states. This means that even if later models can predict reliably, it would not be a good predictive model.



This shows a comparison of the mean value of each state and the confirmed values of each state (mean/confirmed). Some states make a big difference, and the states adjacent to New York and New Jersey have very high values.



This is a graph comparing the three states' mean values and each state's 'Confirmed' value without New York and New Jersey. There are cases where there is a high percentage. Those have very low 'Confirmed' value, and I will note how this will affect my prediction model.

```
Y_predicted - Y_confirmed1
```

```
array([  5059.36979847,   6126.09986558,   7351.9337786 , -18969.51895248,
         2026.13756696, -10203.28414581,   8121.97841137, -16725.06925881,
        -6245.16147187,  10272.35962519, -18638.48300979,  -3211.50676352,
         7065.02579798,   5898.21831153,   8691.26044548, -11315.23188498,
         7286.83460952,    170.56021908, -25955.61034645, -21889.27018436,
         7718.54389358,   4774.5725301 ,   2013.61453218,  11450.23857943,
         8809.86333686,   2666.4121071 ,   3900.87659276,   8659.46596989,
         5996.94007638,   9092.82997003,   1463.14147492,  10589.13919246,
        10493.38314682, -20885.82342803,   -910.92194097,   4497.08298984,
        11336.54796105,   4212.5803338 , -15243.11723035,   9024.18852275,
         4985.7697499 ,   2295.31128577,    804.33150369,  11571.96712244,
         1382.64256471,  11662.29990692])
```

I created a model to predict one state's 'Confirmed' value using the 'confirmed' values of three closest states to the state. The previous model had an rmse value of 30000, but this time I reduced it to 10000. However, when looking at the predicted values, those are significantly different from the actual values. This means that a model that fits the average value was created, not a model that predicts the value.

At the start of this project, I aimed to graphically show correlation between distance from states and contagiousness, and build a reliable predictive model using those variables. However, I could not show well. Looking at the reality, it seems clear that there was an epidemic between states, and I still think it will have a big impact when there is a large number of confirmed people in a nearby city. I don't think it was wrong to find three nearby states and predict the number of infected people in a state by the number of confirmed people in those states. I think the distance between states is too far than I thought, so the impact may not be great, and the movement between states may be less than I thought. If I used data about state-to-state floating population, road conditions, or variables such as 'incident_rate' and the number of hospitals per a person, I might have created a more realistic prediction model. If it is possible to collect data, more microscopic, I would like to show the correlation between the number of infected people in one city and other cities.