COVID-19 Analysis - Predicting Confirmed Cases in Counties and Observing Key Trends

(Authors: Yin Deng, Basant Apurva, Yatin Agarwal)

Abstract

This project aims to find and analyze key trends that have governed the COVID-19 situation in the United States. Using COVID-19 datasets of confirmed cases, county demographics, and recent vehicle movement data, we find several observations. We build a few different models for accurately predicting the percentage of newly confirmed cases in a county, study how the percentage of cases in a county varies with its policies, vehicle mobility, and county demographics, understand how cases and their growths have varied geographically.

I. Introduction & Question Framing

In the first half of 2020, the Coronavirus pandemic has drastically changed the world around us. For students such as ourselves, our classes have been moved online, social interaction has dropped to a minimum, and our prospects for the future have become more uncertain than ever. On a global scale, this pandemic has caused significant levels of death and unemployment, and threatens to fundamentally transform life as we know it.

Given the historic significance of these times, our group was naturally inclined to work with the COVID-19 dataset. Doing so allowed us to put the skills we learned in Data 100 to use in a highly relevant way, while also developing a deeper understanding of the Coronavirus pandemic.

Our analysis into the COVID-19 data set was guided by two main questions:

1) How can we better understand the Coronavirus in its current state?
    a) What factors have enabled the virus to spread? Which of these factors are the most/least important in the spread of the virus?
    b) How have the effects and response to the Coronavirus varied across the United States?
2) What is the future of the Coronavirus?
    a) What will the number of cases/deaths look like in the future?
    b) How does our response to the virus affect these numbers?

II. Data Description and Cleaning

1) Confirmed Dataset: This dataset contains daily confirmed cases for all counties in the United States from late January to early May.

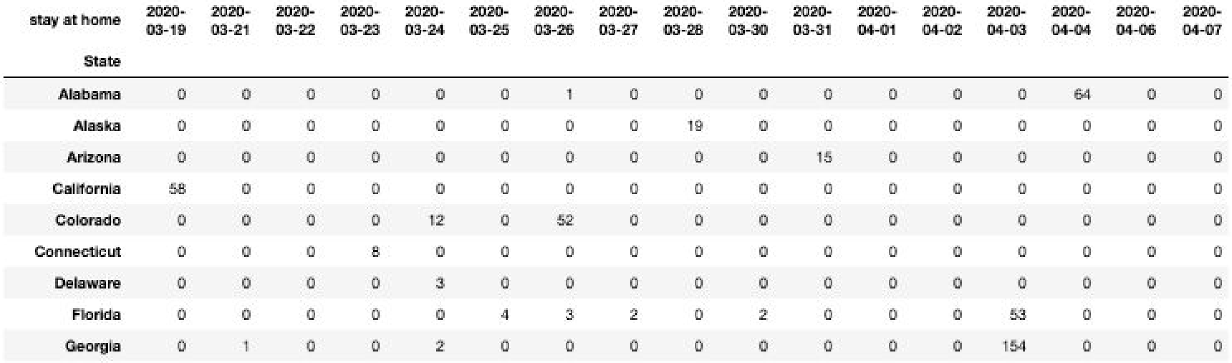
    a) We start by filtering out all columns except for the combined_key, which includes location, and confirmed cases per day from 1/22 - 5/2. The filtered columns contain irrelevant information as we are only interested in the number of confirmed cases each day at each location.
    b) We confirm that there are no missing values in this dataset
2) Demographics Dataset: This dataset contains demographic information for people in all counties in the United States regarding their age, gender, health, etc.
    a) First, we identify the unique states in the dataset and create a dictionary mapping their two-letter acronym to their full name. The two letter-acronyms are all fully present in the data while there are many missing values for the full state name. Using state name and county name, we generate a combined_key for each data entry.
    b) After looking over all the columns, we filter for six columns that are most relevant to our questions and drop rows with missing values in those six columns.
    c) Next, we create a new table to highlight when stay at home policies were adopted and standardize the formatting of all dates.
3) Merging Datasets
    a) We merge our demographics dataset with our confirmed cases and stay at home policies dataset all on the "combined_key" column, which represents an individual county in the US.
    b) We replace two of our columns relating to population with a new column that represents the ratio of individuals over 65 in a county to standardize data.
4) VMT Dataset: This dataset contains vehicle miles travelled information for counties in the United States from the beginning of March till the beginning of May.
    a) We apply pivot table to this dataset to be in a similar format as our previous datasets
    b) We create a combined_key for each entry using county name and state name.
    c) We check for and drop missing values in the dataset and merge it with the rest of our datasets to create one complete dataset.
5) State Dataset: This dataset contains information related to COVID-19 on a state level, such as hospitalization rate, testing rate, mortality rate, etc.
    a) Note: We opted to use the state dataset for mapping because we felt a state by state visualization would be more readable and familiar to the reader as compared to a county by county visualization. Additionally, because neighboring counties share similar statistics, we believed it was only relevant to highlight differences on a state by state level. However, the information provided by the state dataset is provided in greater detail in our merged dataset, which is why we opted to use this dataset solely for mapping.
    b) First, we filter for US states and check for and drop any states included in our dataset but not on our map (i.e. US territories)
    c) We then combine our state dataset with a US map '.shp' file which allows us to plot our data over a map.

6) Deaths Dataset
   a) Note this dataset was included in the 'Side Experiments' section of our project
   b) We standardize our combined key, sum up total deaths, and combine the data with our other merged data

III. Data Visualization

With visualizations, we wanted to understand relationships between different parameters that govern the pandemic and look at trends we could possibly use for modeling the total number of confirmed cases on a given day. We were able to observe the following -

1. Understand how counties functioned independently of the states to which they belonged in issuing stay at home orders -



stay at home	2020-03-19	2020-03-21	2020-03-22	2020-03-23	2020-03-24	2020-03-25	2020-03-26	2020-03-27	2020-03-28	2020-03-30	2020-03-31	2020-04-01	2020-04-02	2020-04-03	2020-04-04	2020-04-06	2020-04-07
State																	
Alabama	0	0	0	0	0	0	1	0	0	0	0	0	0	0	64	0	0
Alaska	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0
Arizona	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0
California	58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Colorado	0	0	0	0	12	0	52	0	0	0	0	0	0	0	0	0	0
Connecticut	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0
Delaware	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
Florida	0	0	0	0	0	4	3	2	0	2	0	0	0	53	0	0	0
Georgia	0	1	0	0	2	0	0	0	0	0	0	0	0	154	0	0	0

Figure 1: Table for number of counties for each state that imposed stay at home at the same time

For stay at home restrictions, we realized most of the counties belonging to the same state issued these restrictions at the same time, suggesting most 'stay at home' orders for households in America were ordered by their respective state governments rather than their individual counties. However, for some states such as Texas, Pennsylvania, and Colorado, some independent counties issued the 'stay at home' policy before it was declared for the whole state. This suggests how some counties tend to function independently of the states to which they belong in their policies.

2. Understand whether America's response to the virus was preventive or a reaction to when it got bad -

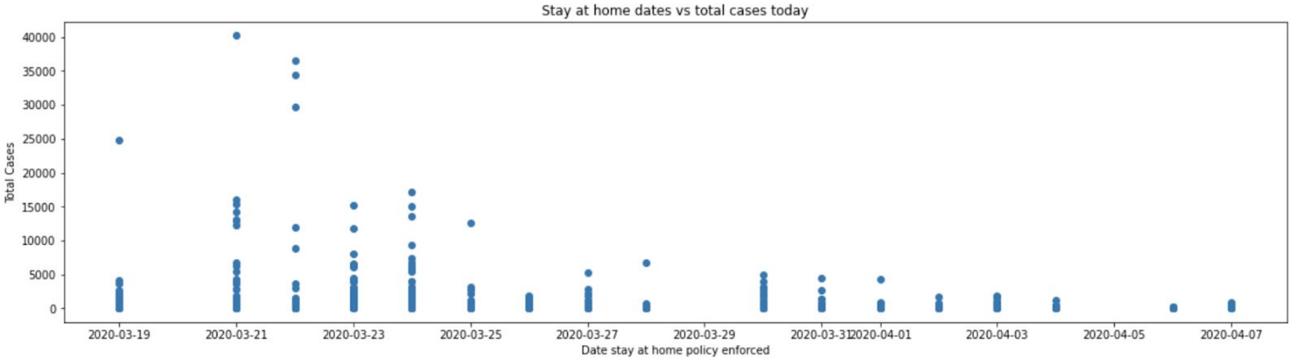


Figure 2: Scatter plot for number of cases today in counties vs when stay at home was imposed

To figure this out we plotted a scatter plot of how the total number of cases in a particular county on May 2nd related to how early they'd enforced 'stay at home.' Surprisingly we saw that most of the counties that started the 'stay at home' policy for their citizens early tended to have a greater number of cases today. This indicates a more precautionary response of state/county policies towards the Virus rather than a reactionary one, as the states realized what states were going to get worse than other states, and shut themselves down early. Of course, this could have been even better if they'd shut themselves down even before.

3. Understand how total populations and their densities related to the total confirmed cases/percentage of population confirmed with the virus -

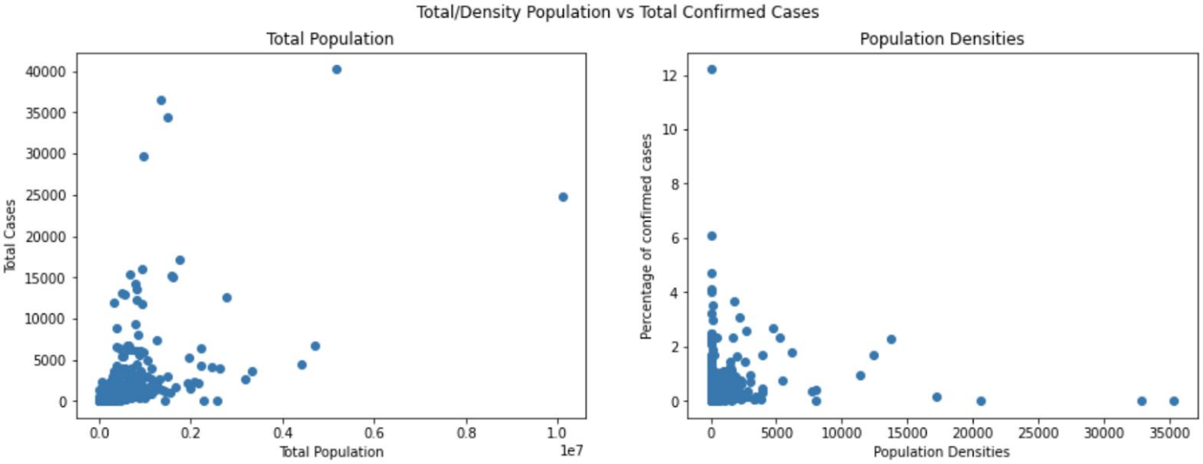


Figure 3: Scatter plots for the number of cases today in counties vs total population (left) and for the percentage of confirmed cases in counties vs population densities

We'd expected a strong linear relationship between population densities and the percentage of confirmed cases in that county, but surprisingly this wasn't the case. In fact the linear relation was a little more apparent in the case of total population vs total number of cases than for population densities with percentage cases.

4. How demographic/health conditions related to percentage of population infected -

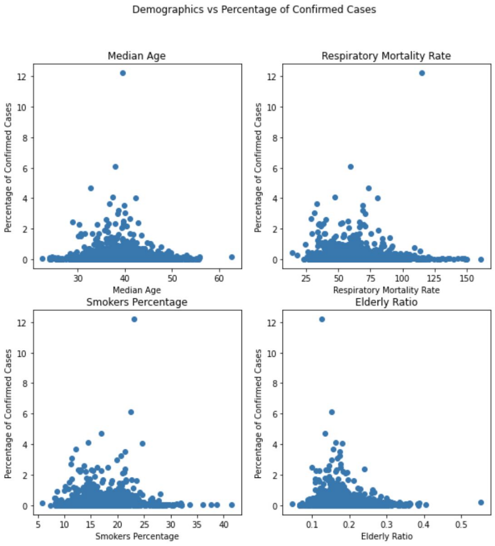


Figure 4: Scatter plots for the percentage of confirmed cases in counties vs median age, respiratory mortality rate, smoker's percentage, and elderly ratio

Similarly we'd expected a slight linear relation between the total number of cases and demographic/health conditions such as proportion of people aged over 65, median age, smoker's percentage, and respiratory mortality rate. However, this relation wasn't really visible and it complemented the observation that when we tried to later model solely based on such demographic information our model was really inaccurate.

5. Visualize growth of virus, and change in amount of vehicle mobility across counties -

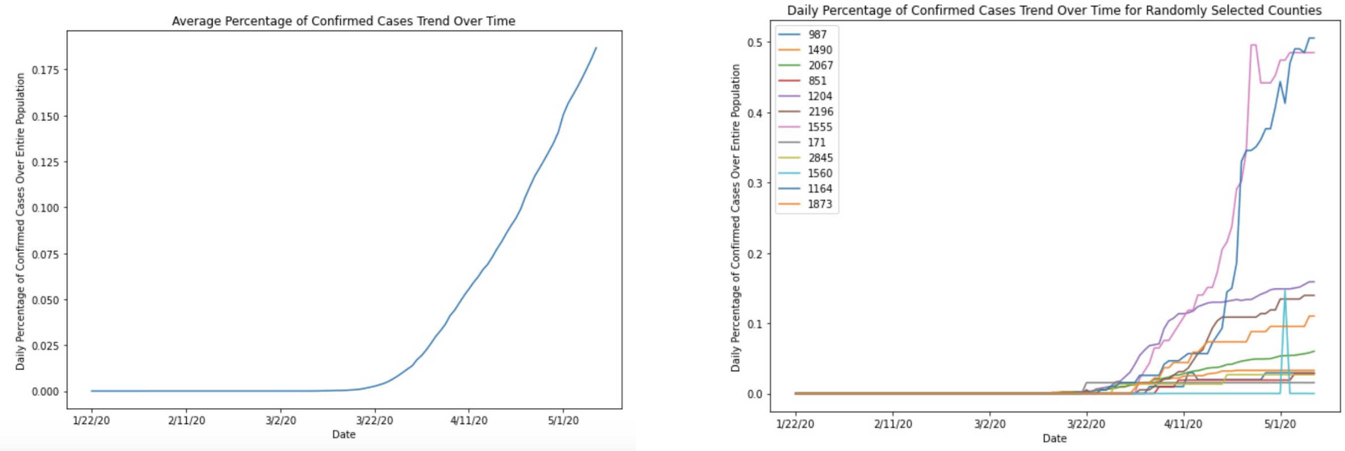


Figure 5: Line plots for avg. percentage of confirmed cases across counties (left), and daily percentage of confirmed cases for counties (right) over time

It was really interesting to see as the number of confirmed cases grew exponentially, the average miles traveled for vehicles in counties decreased significantly as well. There was a strong relation as when we later included the VMT data as a feature in our model, the accuracy increased significantly.

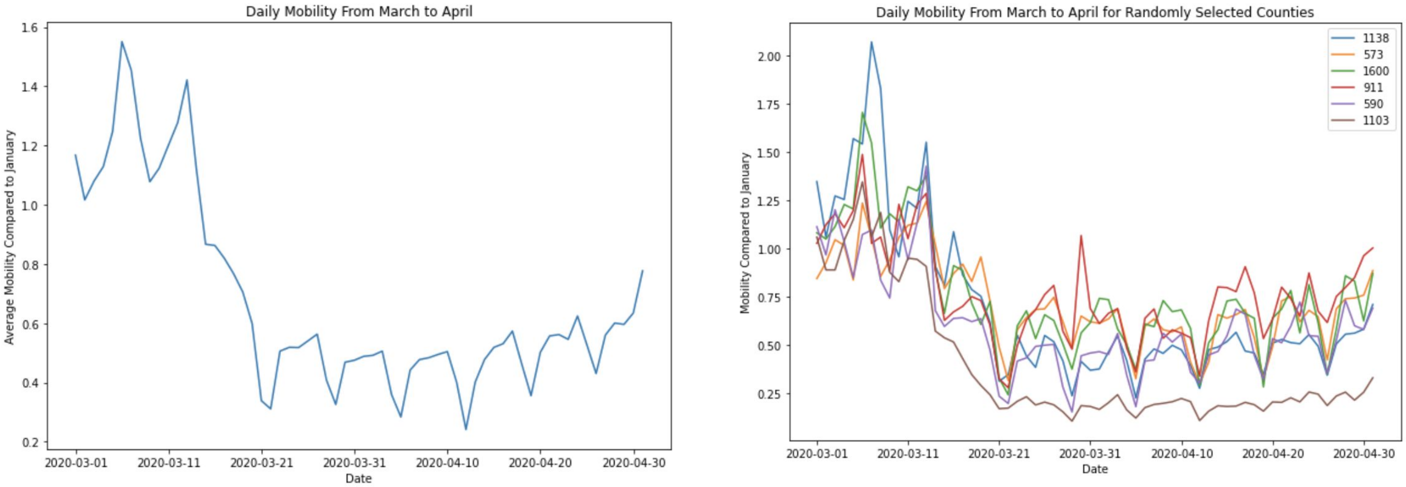


Figure 6: Line plots for avg. daily mobility compared to January across counties (left), and daily mobility compared to January for counties (right) over time

6. Visualize geographic spread of various virus related features across the country -

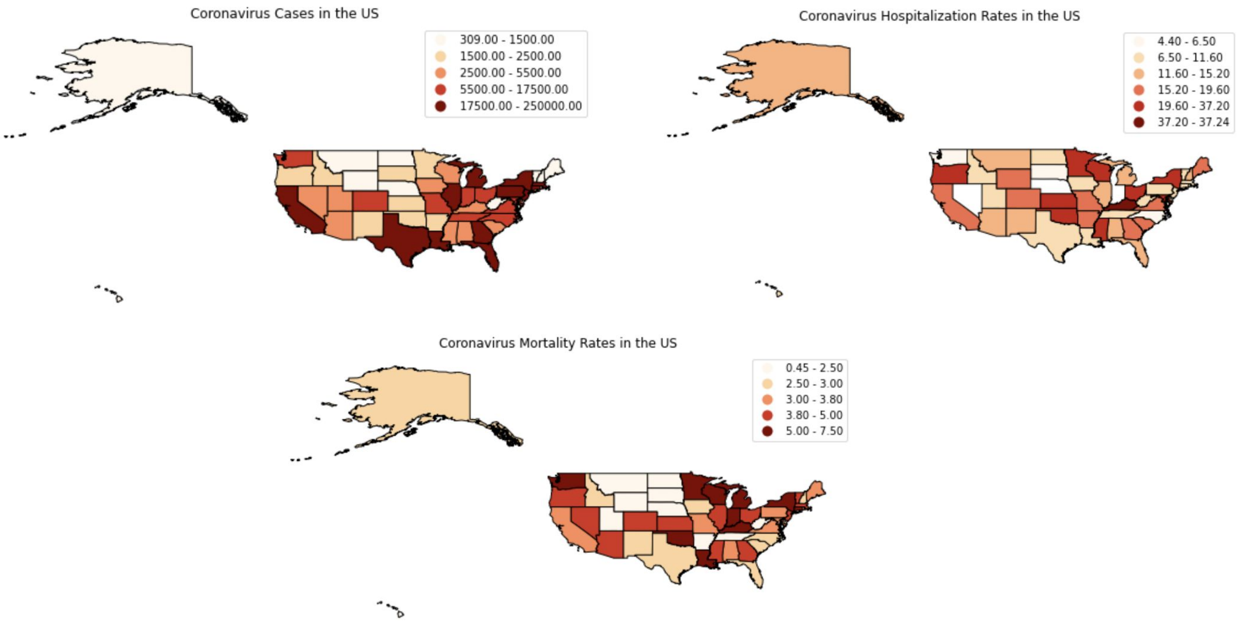


Figure 7: Heat-maps for total confirmed cases, hospitalization rates, and mortality rates

We wanted to easily understand how the virus has spanned across the country, we created a heatmap for each state's total number of confirmed cases, hospitalization rates, and mortality rates. It was interesting to visualize how although some states had a lot of cases, their hospitalization and mortality rates were relatively lower.

IV. Method and Experiments

1. The first question we want to address is how to predict the percentage of newly confirmed cases in each county based on demographic information such as population density, the proportion of elderly people, respiratory disease mortality rate, etc.
   - We select these features based on what we have learned about COVID-19 from our preliminary research. COVID-19 is more fatal towards the elderly, so we include the proportion of elderly people and the median age in our models. COVID-19 attacks people's respiratory systems, so we include respiratory disease mortality rate and smoker percentage in our models. COVID-19 can transmit between people, so we include population density and vehicle mobility data in our models.
   - We create two linear regression models to predict the percentage of newly confirmed cases in a population on 05/02/2020, one without regularization (left) and one with regularization (right), and their residual plots are listed below.

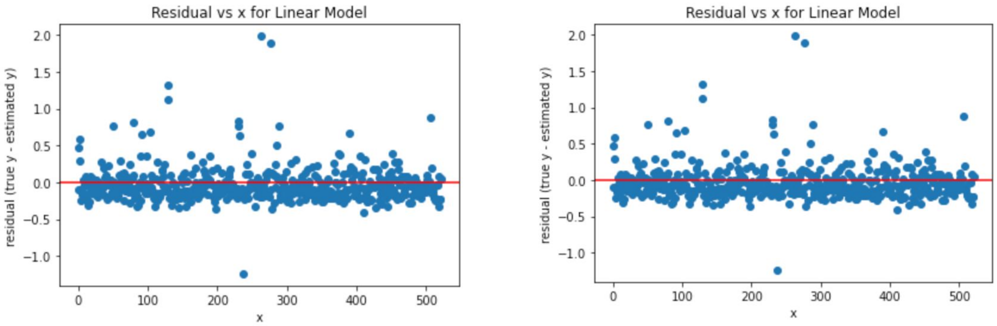


Figure 8: Residual plots for our prediction model with demographics data and VMT data, without regularization (left) and with regularization (right)

   - From the residual plots, we can see these two models share similar performance, but neither of them is good. The percentage of daily confirmed cases in a population is often around 0.05%. However, the error is often around 0.02%, which renders this model useless.
   - We think this happens because even though we know certain features of COVID-19 from our preliminary research, none of them are a strong indicator of the actual number of confirmed cases in a county. These features might show up in the long run or on a larger scale, but they are not very helpful here.

2. Since our first attempt to predict the percentage of confirmed cases in a population on 05/02/2020 fails, we decide to add more features.
   - Besides features we already include in our first attempt, we also include the daily percentage of newly confirmed cases in a population from January to May. Since these data are more directly related to what we want to predict, we believe they could boost the performance of our models.
   - We create two linear regression models to predict the percentage of confirmed cases in a population on 05/02/2020, one without regularization (left) and one with regularization (right), and their residual plots are listed below.

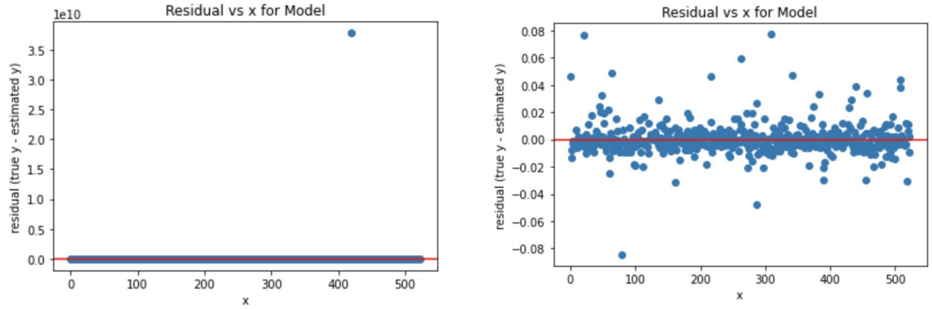


Figure 9: Residual plots for our prediction model with earlier cases data, demographics data and VMT data, without regularization (left) and with regularization (right)

   - From the residual plot on the left, we can see that linear regression without regularization overfits the training set, so there is a huge error for one of the entries in the test set. From the residual plot on the right, we can see that linear regression with regularization fits the test set pretty well, and the mean square error for the training set is 0.000196% and the mean square error for the test set is only 0.000142%. Therefore, we know the model on the right succeeds at predicting the percentage of newly confirmed cases on 05/02/2020.
   - Even though we reach the conclusion that features from our first attempt are not particularly helpful, we still include them in our second attempt because we believe those factors could show up in the long run.

V. Analysis and Inferences

1. What were two or three of the most interesting features you came across for your particular question?

   - In order to predict the percent of newly confirmed cases in a population, one interesting feature we came across was the daily vehicle miles traveled data since it tells us how much people go out. There are a lot of fluctuations in the data, but we see a decreasing trend from March to mid-April although it has bounced back recently.
   - In order to understand how COVID-19 spread in different states, features like COVID-19 mortality rates and COVID-19 hospitalization rates are also interesting. By making some heatmaps, we observe that the COVID-19 mortality rate is high in New York State and Washington State, and COVID-19 hospitalization rates are high in Kentucky.

2. Describe one feature you thought would be useful, but turned out to be ineffective.
   - One feature we thought would be useful was the proportion of the elderly in a population since we know COVID-19 is more fatal for the elderly. However, in EDA, we did not find any strong correlation between the proportion of the elderly and the number of confirmed cases in a county, so that feature was not effective.
3. What challenges did you find with your data? Where did you get stuck?
   - One challenge we have is the huge number of NaNs in abridged_couties.csv when we clean the data. If we just naively drop all the rows which contain NaNs, we might lose more than half of the entries in the dataset. Therefore, we first select a few columns that we think are useful for our analysis and only drop rows who have NaNs in those selected columns. By doing this, we only lose a couple hundred entries and still have enough entries left for further analysis.
4. What are some limitations of the analysis that you did? What assumptions did you make that could prove to be incorrect?
   - A lot of the demographic information from abridged_couties.csv is outdated. For example, lots of information comes from the last census which was 10 years ago. The mortality rate data also comes from a couple of years ago. We made the assumption that the outdated data is actually not too far away from what the population looks like today, and this could be proven wrong by the 2020 census.
5. What ethical dilemmas did you face with this data?
   - One ethical dilemma we faced was building a single model that could predict the percentage of newly confirmed cases in a population for all counties. By doing this, we might have ignored the differences among counties. Maybe there are some fundamental differences between a county on the west coast and a county on the east coast, which were not captured by our model.
6. What additional data, if available, would strengthen your analysis, or allow you to test some other hypotheses?
   - Data from the 2020 census will definitely strengthen our analysis since it has the latest demographic information on the entire United States population.

- More data on the number of confirmed cases and vehicle miles traveled could also be helpful. Although we see the number of newly confirmed cases increasing for the past two months, it might plateau out at some point and then start decreasing. Having more data on the number of confirmed cases can help us adjust our model. Also, we see the vehicle miles traveled bounce back at the end of April, so we are not sure if that is a new trend or just some random variation. Having more data on vehicle miles traveled can help us confirm which case it is.

7. What ethical concerns might you encounter in studying this problem? How might you address those concerns?
    - We think it could be helpful to have demographic information on individual patients who are diagnosed with COVID-19 so that we can build a model with stronger correlation. However, we also realize that in a time like these, singling out these patents might cause them to be ostracized by society. Therefore, we decide to settle with the aggregated data we have and study those instead.

Conclusion

We were able to use COVID-19 datasets to accurately model and predict the percentage of cases in a county on a given day. We observed key trends such as a strong dependence of Covid-19 cases on vehicle mobility and the number of cases in the previous days, and understood how several county demographics and policies have affected the response. We also built interesting visualizations that helped us understand the geographical spread and changes in total number of cases and mobility over the course of the last months.