Arjun Srivastava
Daniel Qiang
CSE 163

Exploring Covid-19's Implications in the US

Research Questions:

1. Based on data up until the present, when can we expect the number of hospitalizations of Covid-19 in the United States to be equal to those of the flu in 2019?
    a. With our linear model, we can expect the number of hospitalizations to equal those of the flu in 2019 around October 15.
2. Did the killing of George Floyd in Minneapolis and the subsequent countrywide mass protesting/civil strife cause a spike in new Covid-19 cases? What would case numbers have looked like had Floyd not been killed?
    a. Based on both of our models trained with case data prior to Floyd's death, active cases in Minnesota and Washington would both be lower than they are currently by upwards of 15,000 average cases.
3. Has the rate of recovery of those afflicted by Covid-19 improved on average since January?
    a. After an initial downward spike, the recovery rate for Covid-19 in the US has steadily increased over time, with no indication of a peak or stall any time soon.
4. Has the percentage of young people (ages 18-30) contracting Covid-19 in the US increased since March 2020? If so, what groups decreased? If not, what groups increased?
    a. The percentage of young people (ages 18-30) contracting COVID-19 in the US has steadily increased since March 2020. Older age groups (50+) decreased in

proportion, whereas younger age groups (all age groups under 50) increased in proportion relative to the total number of COVID-19 cases per day.

5. Which states had the most success in reducing new Covid-19 cases per day? Which were the top 5?
    a. In order: Texas, New York, California, Michigan and Florida were the most successful in reducing the number of new COVID-19 cases per day. In one day, Texas reduced their daily COVID-19 cases by 5300 (compared to the previous day).

Motivation and Background:

The Covid-19 crisis is one of the most significant events of the modern era, with over 600 thousands deaths worldwide, economic fallouts, drastic policy shifts, and a polarizing sociopolitical environment rising from the damage. As the entire world continues to deal with the virus, different countries have seen varying amounts of success in terms of flattening their curves, recovery/death rates, financial compensation for those affected, and control over their respective populations. The United States in particular is currently a political battleground as citizens fight over mask policies, the reopening of businesses, public gatherings, social distancing, and more, with rates constantly changing due to a lack of cohesive strategy from federal and local governments. With a constant spread of misinformation on social media, the American public remains divided over how to best handle the crisis, resulting in strong-armed debates unheard of in many other more successful countries. In order to analyze the situation more objectively, we decided to conduct a research project examining active cases, death rates,

recoveries, demographics, and more. To see how the virus compares with the common flu in terms of spread and case numbers, we aim to compare the two and see how long it could take for the virus to reach the same level of infection among Americans. With such a heated conversation regarding social distancing legislation and policy, we will be looking at potential death rates for states had they not adopted these policies, as well as measuring which states have had the most success with flattening their curves. With this information, we will be able to make claims about the true effectiveness of legislation in order to combat the spread of the virus. Additionally, we aim to use objective data to truly see if the situation is improving country-wide, as so many online sources do not seem to agree. By exploring all of these questions and potential alternate outcomes, we can get a better understanding overall of the Coronavirus and its effects in our society.

Data:

- https://github.com/nytimes/covid-19-data
    - This repository contains case and death statistics for the US and its states
- https://ourworldindata.org/coronavirus-source-data
    - This dataset contains information about deaths, cases, and population
- https://catalog.data.gov/dataset/covid-19-cases-summarized-by-age-group-and-gender
    - This dataset contains information about cases by age
- https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases?force_layout=desktop

    - This link has data regarding recoveries throughout the country
- https://www.cdc.gov/flu/about/burden/2018-2019.html#:~:text=CDC%20estimates%20that%20the%20burden,from%20influenza%20(Table%201).
    - Flu information (not a dataset)
- https://covidtracking.com/api/v1/us/daily.csv
    - Contains comprehensive US data

Challenge goals:

- **Requests**
    - We used the requests library to make HTTP requests and pull our various data sets. This method ended up proving beneficial to ensure that our data was always up to date.
- **Multiple data sets**
    - Many different independent and official sources have been conducting Covid-19 research separately. In order to paint a fuller picture of the situation, we needed to pull data from many sources and combine them in the most productive ways. This was accomplished by merging our various DataFrames for numerous questions.

Arjun Srivastava
Daniel Qiang
CSE 163

Methodology:

Pre-processing of the data:

In order to ensure our data is constantly up to date upon running our program, we pulled data from remote repositories using HTTP requests. When necessary, we cleaned up and reformatted data accordingly, joining multiple datasets on numerous equations.

Question 1:

First, we pulled a dataset that contains Covid-19 hospitalization data in the US over time. After formatting the data, we used scikit-learn to produce a linear regression model based on the data up to the present. Then, we extrapolated and found the intercept of our model and the number of flu hospitalizations in 2019. To visualize the results, we plotted the trends over time.

Question 2 (contains testing):

Using active case data in US states, we decided to focus on Washington and Minnesota for our model due to the states' large-scale protests. We trained both linear and polynomial regression models of cases over time only using data up until Floyd's death on May 25. Then, we plotted these models against the true cases data. For testing, we used a local copy of the data set to ensure our model would work if the HTTP request failed.

Question 3:

After being unable to find a singular data set with both recovery and death numbers by date, we decided to pull multiple data sets and merge them. To visualize the rate of recovery / deaths trend, we plotted this calculation over time from the beginning of the pandemic to the present. As the resulting graph shows, there is a generally positive trend.

Question 4:

Arjun Srivastava
Daniel Qiang
CSE 163

Using a dataset of new COVID-19 cases per day for individual age groups per state in the US, we calculated the percentage of new cases belonging to each age group and plotted a time series of this ratio per day for each age group in a line graph. As the initial month did not have data points for all age groups, we assumed that this meant there were no new confirmed cases for that age group at that time and set the ratio to 0. Once there was data for all the age groups (~April 2020), we saw a general upward trend indicating that more 18-30 year olds in the US were getting infected relative to the rest of the age groups as time went on.

Question 5:

Using a dataset containing cumulative COVID-19 cases per day by state, we calculated the number of new COVID-19 cases per day by state. We then found the 5 states with the largest one-day drop in case count, and visualized them by plotting these top 5 states on a line graph. To create the US map of case reductions, we wrote two methods to update given data and combine it with a US states geodataframe. Afterwards, we added the case reduction DataFrame to the US states DataFrame and plotted the map with GeoPandas.

Testing:

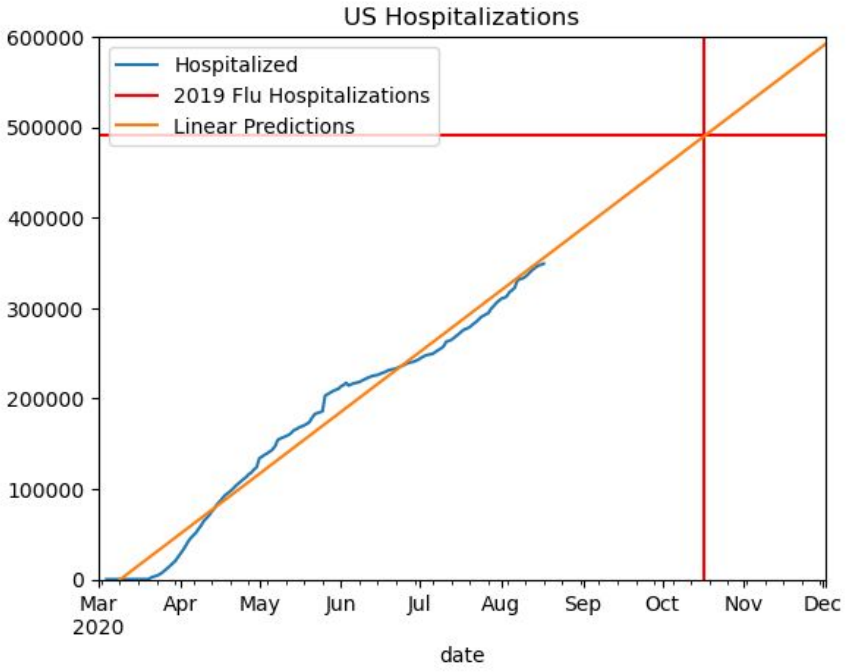
In order to ensure that our data collection methods do not fully rely on internet connection and the ability to make HTTP requests, we added a testing clause in our q2() method (Question #2). Testing this question correctly requires the entire repository to be downloaded. In order to run the test, simply disable internet connection on the local machine and run main.py. While all other questions should fail, Question #2 should still produce the same results as expected.

Arjun Srivastava
Daniel Qiang
CSE 163

Results

Question 1:

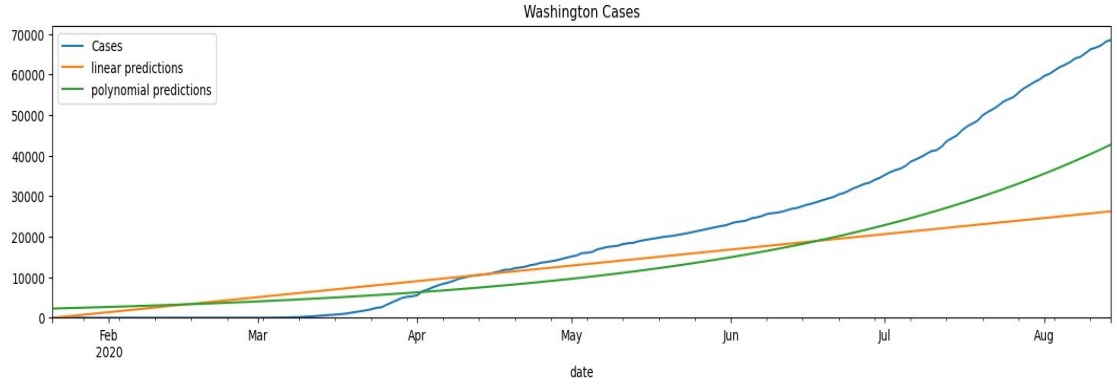
Based on our linear model of US Covid-19 cumulative hospitalization data, we will reach the 2019 number of flu hospitalizations (490,561) by ~October 15. Due to Covid-19's increased lethality when compared to the common flu, these are troubling results. With over 166,000 Covid-19 deaths already versus only ~34,000 flu deaths in 2019, these results present us with a daunting reality: we need to figure out more efficient and effective treatment methods if we want to halt this trend and decrease the rate of future hospitalizations.

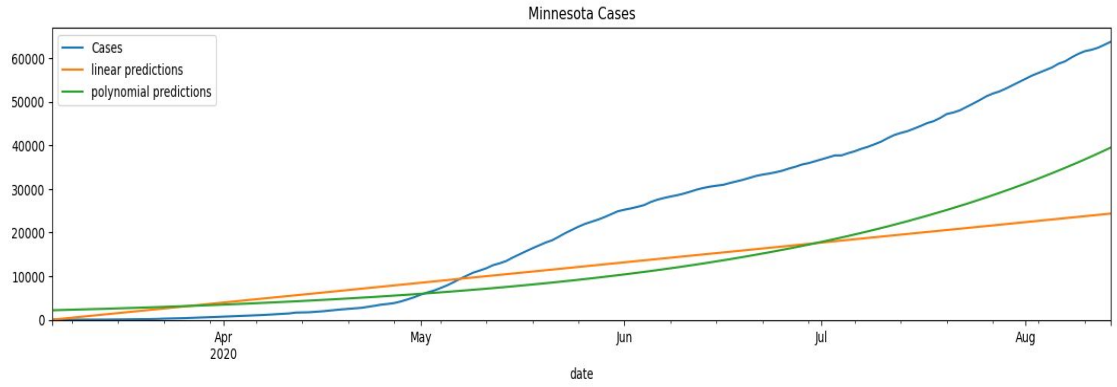


Arjun Srivastava
Daniel Qiang
CSE 163

Question 2:

Results from our regression models confirmed our hypothesis that case rates might have been lower had George Floyd never been killed and the subsequent protests never happened. Both our linear and polynomial models trained with pre-death case data resulted in a generally lower amount of average cases by around 20,000 and 15,000 for Minnesota and Washington respectively. Although our model observed a lower trend, however, we cannot make any claims about the protests having definitively increased cases based on this data alone.

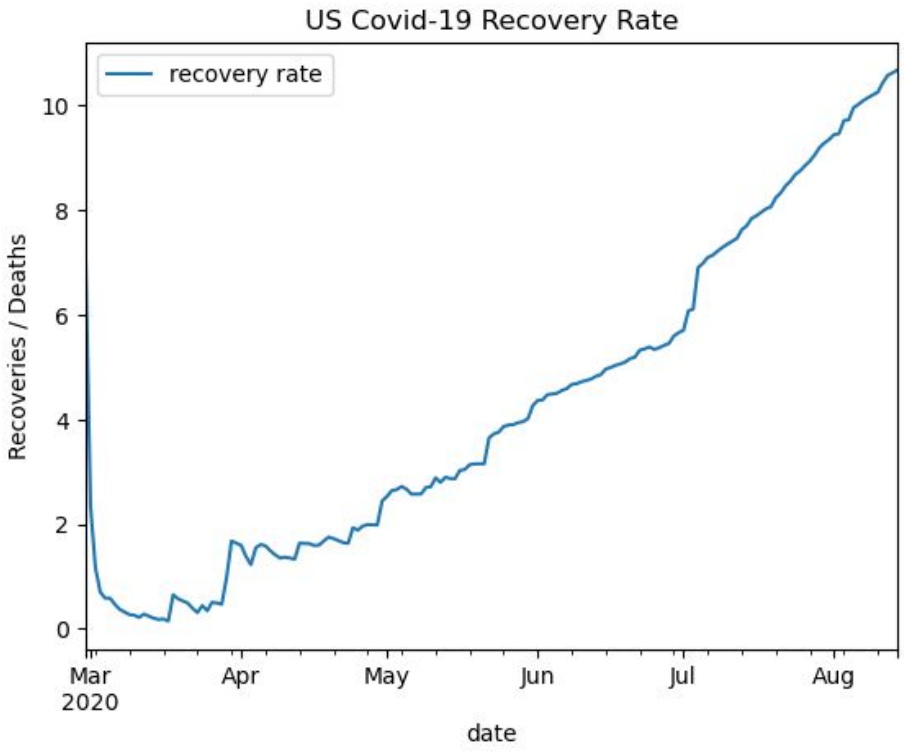




Arjun Srivastava
Daniel Qiang
CSE 163

Question 3:

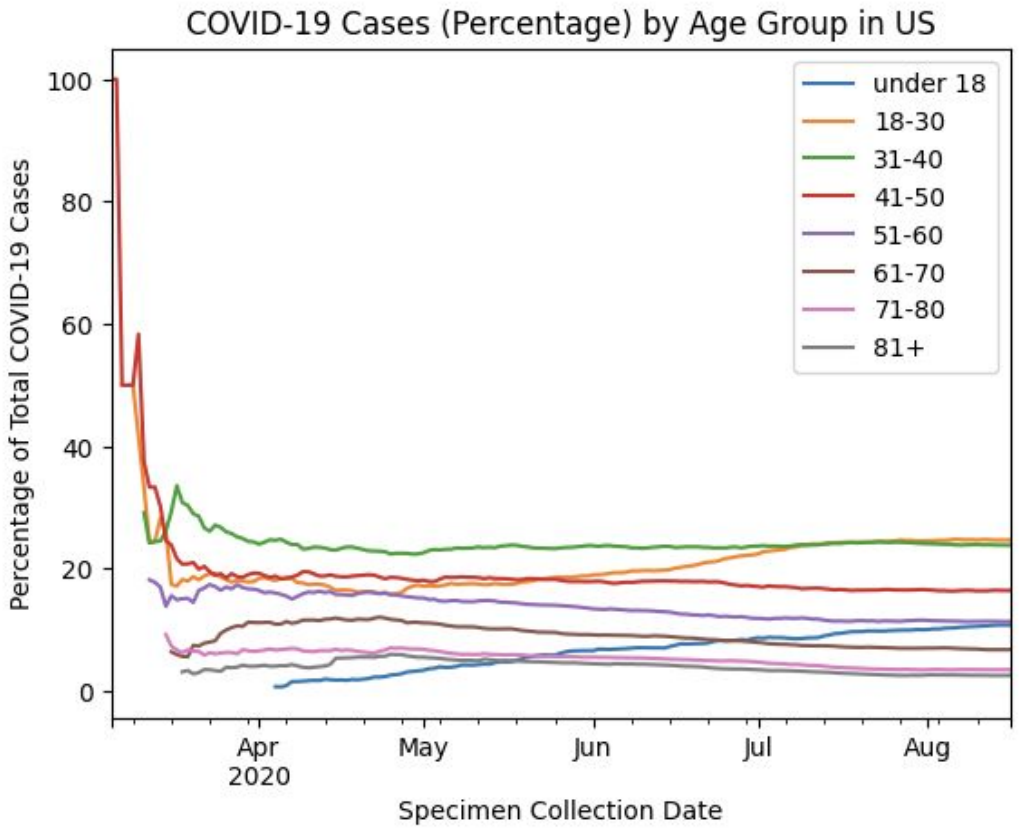
After plotting our definition of recovery rates against time for the US as a whole, we discovered that Americans are gradually recovering at a faster rate than they are dying from Covid-19. At first, the rates were more turbulent and unpredictable, but over time they have become increasingly positive. Ideally, we will reach an even steeper positive slope, indicating a decline in deaths and rise in recoveries



Arjun Srivastava
Daniel Qiang
CSE 163

Question 4:

The following visualization shows that the ratio of COVID-19 cases belonging to the 18-30 age group in the US increased gradually but steadily over time:

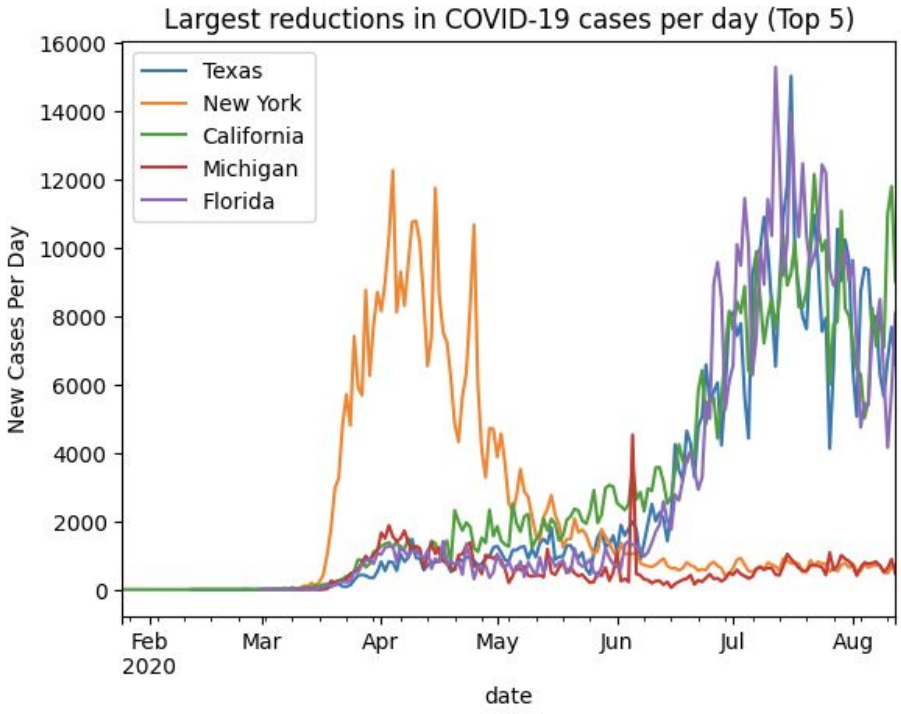


The ratio of cases belonging to the under 18 and 31-40 age group also increased, whereas the ratio of cases for older age groups all decreased. This confirms our hypothesis that younger people are likely to make up more of the body of infections over time, as the majority of COVID-19 casualties are older and/or have compromised immune systems/pre-existing conditions; thus, since the pool of older people decreases in size over time, COVID-19 cases belonging to younger individuals make up a larger proportion of the total population over time.

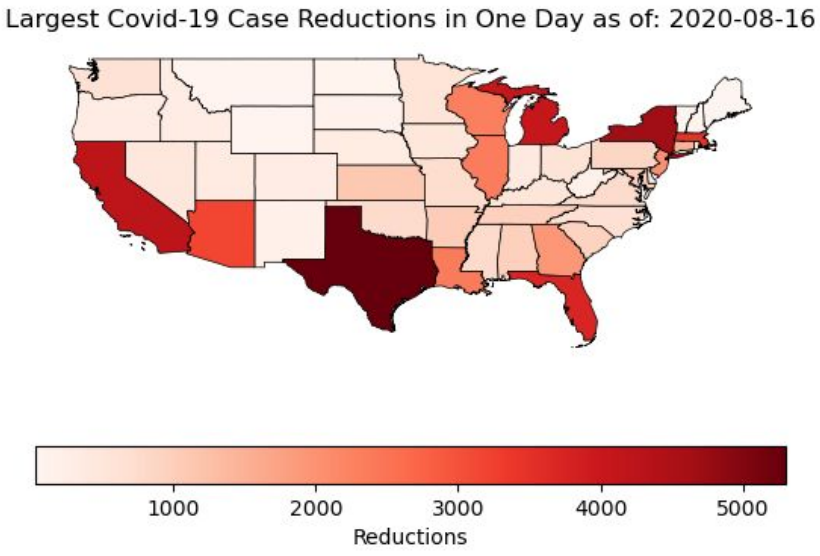
Arjun Srivastava
Daniel Qiang
CSE 163

Question 5:

The top 5 states for reducing COVID-19 cases per day were (in order): Texas, New York, California, Michigan and Florida. This can be seen in the following line graph:



And the following heat map:



Arjun Srivastava
Daniel Qiang
CSE 163

To identify the states with the most success in reducing COVID-19 cases per day, we calculated the largest reduction for each state between any two days over the 7 months of data in the dataset, and used this number to determine which states had the most success in reducing the number of cases. With this, we found that the states that had the most success in reducing COVID-19 cases necessarily had to have performed the worst in the initial month(s) of the infection (February, March, etc.), as many states did not have many COVID-19 cases to begin with. Additionally, the states that were marked as being the most successful in reducing the number of COVID-19 cases had both large populations and large oscillations between days, as opposed to a consistent downward trend; this is likely due to limitations/errors in data collection, as opposed to the actual number of new cases oscillating continuously. A better approach to measuring states' success in reducing COVID-19 cases may have been to first normalize the number of new COVID-19 cases against the population of each individual state and use those data points, instead of using the raw data.

Work Plan:

High-level Tasks (1-7):

1. **Processing (1 hr)**
    a. We will utilize requests to download the necessary data from the internet and convert it to CSV format (20 min to retrieve all necessary data)

b. **Evaluation**: This proved to be the simplest aspect of our project. We wrote a method that uses requests to download a .csv file and then convert it into a pandas DataFrame, which took only around 15 minutes.

2. **Question 1 (1 hr):**
    - Based on cumulative hospitalization data from the beginning of the pandemic to the present, we will generate a linear regression model and find the intercept of our model and the number of flu hospitalizations in 2019. We will then plot the model against the true rates.
    - **Evaluation:** This task ended up taking far longer than expected due to surprising complications when plotting. Extending a regression line in scikit-learn beyond the data is not a functionality in matplotlib, so we were forced to improvise. We ended up having to manually extend the regression line far enough, which was not a trivial task. Dealing with x-values in regression when datetime objects are on the x-axis caused many unforeseen complications, and this problem ended up taking over 3 hours.
3. **Question 2 (1-2 hrs):**
    - Train regression model with active case data up until March 25 (Floyd's death) and then plot the results against the true case data after March 25
    - **Evaluation:** Being the first question we completed, this task ended up taking the longest. We had to first reformat our data in order to do linear regression on a specific subset of the data. It took some time for us to figure out how to make a simple polynomial regression model from the same subset of data, and we ran into

difficulties with plotting the correct information. Overall, the question took upwards of 3 hours to answer.

4. **Question 3 (3-5 hrs):**
    - Join necessary data sets. Create new column in dataframe of (# of recoveries on x day / # of deaths on x day). Plot results to see if slope is positive or negative for recovery rates. Plot recovery rates on map for more visualization
    - **Evaluation:** Merging and reformatting the necessary data sets was slightly tedious, but creating the new column of recovery rates was simple. Having already plotted data numerous times, this question was fairly easy to complete, taking around 30 minutes as opposed to our predicted 3 hours.
5. **Question 4 (2-3 hrs):**
    - Create a new column of the ratio of daily cases in the age group (18-30) over the total cases on that day for both March and July per state
    - Plot results on US map
    - **Evaluation:** Calculating the ratio of daily cases to total cases was more complex than expected since the case data was a time series; I had to merge/restructure the data. Outside of that, the question was not too difficult to answer and took about 2 hours total.
6. **Question 5 (2-3 hrs):**
    - Create new column of new cases per day per state
    - Generate visualizations to show the states with the largest singular decrease, signifying that they have been most successful in reducing cases

- Plot the states on a US map that shows how successful each state is
- **Evaluation:** Computing the largest singular decrease between any two days for every state and then plotting based on those states was difficult to do in an elegant manner; I ended up using a hashmap from state names to slices of the modified DataFrame in order to avoid code duplication. Outside of that, creating the visualization/answering the question was straightforward; it took about 2 hours total.

7. **Report (15-20 hrs each)**
    - For the write-up, we will plan the outline together, and then split up individual sections of the report equally. After completing our respective sections, we will reconvene and look over each other's work to make any final edits
    - Afterwards, we will physically add all the necessary graphs/visualizations to the report together, and finalize the entire project
    - **Evaluation:** Working both together and separately when necessary proved efficient, and we were able to complete our work goals on time.

Group work (coding):

- We will be utilizing pair programming for the bulk of the coding section of the project. With both of us on a zoom call, we can take turns coding in real time. When there is a need to work separately, we have created a Git repository and will be using it accordingly

- **Evaluation:** After utilizing pair programming to answer our first research question, we split the other four evenly for us to complete alone. This process was simple and worked without complications.

Collaboration:

All work on this project was completed by Arjun Srivastava and Daniel Qiang. Aside from online resources such as stackoverflow and documentation for the various libraries we used, no outside help was utilized in the data analysis itself.