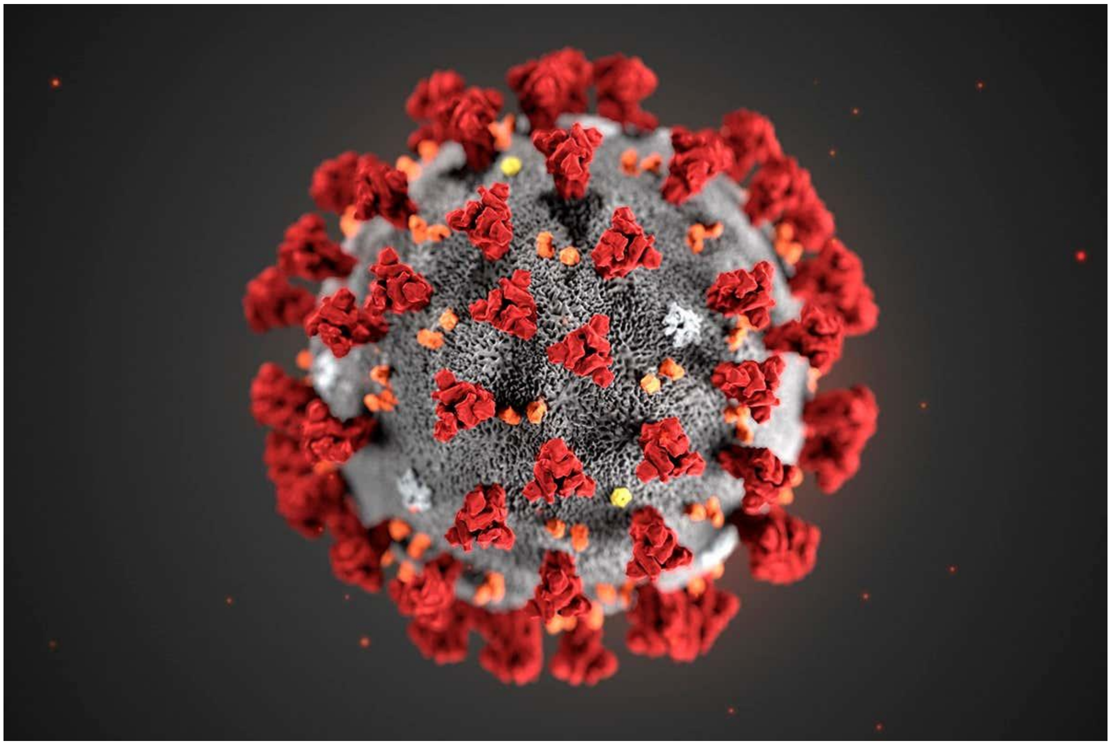
IBM Applied Data Science Capstone

IBM Data Science Professional Certificate

Analysing COVID-19 in the US by State

Shreeviknesh (Apr 2020)



1. Introduction

Coronavirus Disease (COVID-19) is an infectious disease caused by a newly discovered form of coronavirus. Most people infected with the COVID-19 virus will experience mild to moderate respiratory illness and recover without requiring special treatment. Older people, and those with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illness.

The best way to prevent and slow down transmission is be well informed about the COVID-19 virus, the disease it causes and how it spreads. Protect yourself and others from infection by washing your hands or using an alcohol-based rub frequently and not touching your face.

At this time, there are no specific vaccines or treatments for COVID-19. Therefore, it is essential to slow down the transmission of COVID-19 by tightening lockdown measures by introducing high frequency zones where there are the greatest number of recorded COVID-19 cases. Moreover, special camps and medical stations can be set up at these high-frequency “red zones” that can provide the initial diagnosis for patients, quarantine potential positives patients, and provide medical supplies for preventive measures. Setting up such “red zones” can prove to be the most efficient way of fighting against this global pandemic.

2. Business Problem

The objective of this capstone project is to analyse the state-wise cases of COVID-19 in the US to cluster them into zones, such that high-frequency zones can be identified and medical camps can be setup. Using data science methodology and machine learning techniques such as data wrangling and KMeans clustering, this project aims to answer the question: Which are the spots where setting up medical camps will be most efficient in the United States of America.

3. Target Audience

This project is particularly useful to medical institutions, hospitals, and other health organizations that are focussed on fighting back the global pandemic of COVID-19. This project is timely as US is currently suffering from the worst hit of the disease and is leading the world in the greatest number of positive cases and deaths.

Moreover, private and government organizations interested in the fight against COVID-19 can take advantage of this project to help out the cause. This would help the world get rid of the pandemic while also increasing the organization’s public image. Therefore, I believe that many organizations can use this project and the analysis to an extent to fight against the cause.

4. Data

To solve the problem, we need the following data:

- List of states in the United States of America. This defines the scope of this project which is confined to the US.
- Latitude and Longitude coordinates of the states. This is required in order to plot the map with the number of active COVID-19 cases.
- Venue data for each of the positive COVID-19 cases recorded in each of the states.

Sources of data:

The main source of this data is from Kaggle. The dataset used is the **"Kaggle Novel Corona Virus 2019 Dataset."**

The details of the dataset:

- Sno – Serial Number
- ObservationDate – Date of the observation in MM/DD/YYYY
- Province/State – Province or state of the observation
- Country/Region – Country of observation
- Last Update – Time in UTC at which the row is updated for the given province or country.
- Confirmed – Cumulative number of confirmed cases till that date
- Deaths – Cumulative number of deaths till that date
- Recovered – Cumulative number of recovered cases till that date

Moreover, we will leverage the Foursquare API to fetch coordinates for the different US states.

5. Methodology

Initially, we need to read and clean the dataset that we're going to be using. As we can see from the above list, the dataset contains 8 attributes, out of which we only need 3: State, Country and Confirmed cases. Therefore, we will drop the other columns from the dataframe. A problem that is encountered while cleaning the data is that the data was not proper. In the dataset, there were a total of 199 different States entered for the US. As we know, US only has 50 states, so we have to remove all the wrong values.

Doing so will require a list of all the states in the US along with their abbreviation. Fortunately, there are many sources for this data, but I chose to scrape this data from Wikipedia using the BeautifulSoup python package. After collecting the list of states, every row with the country as US was retrieved and checked if the state of the row matched with the list of states.

Moreover, the dataframe contains entries for the number of COVID-19 cases from 22/01/2020 to 20/04/2020 but we are only interested in the latest count of confirmed cases and hence, we will drop all the rows for each State, Country except the last one. After doing so, we are left with a dataframe of length 50 that contains the details of the number of current COVID-19 cases for each of the 50 States.

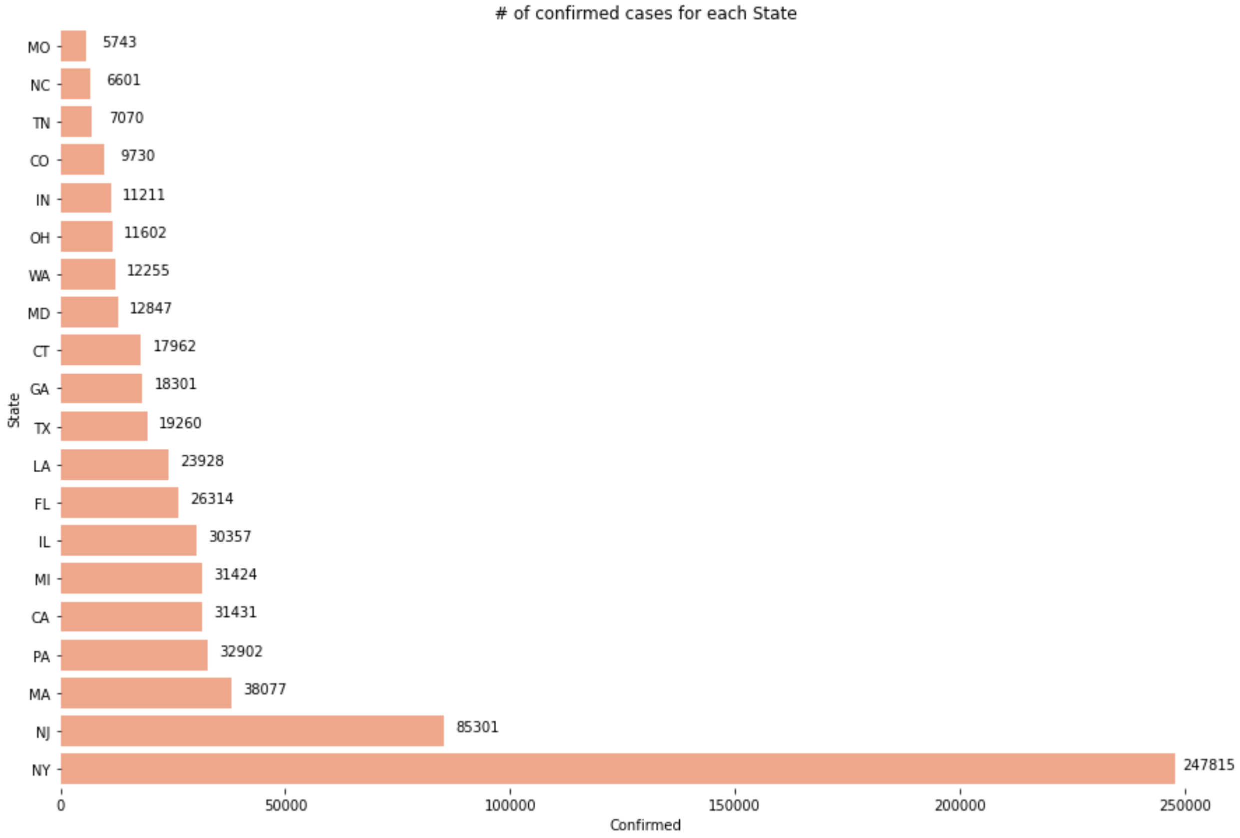
Next, we perform exploratory analysis of the data by plotting a barplot for the data and find the States with the greatest number of active COVID-19 cases. Then we hypothesize the reason for the same.

Further, we collect the latitude and longitude data for the states of the US using the Foursquare API and store it in a dataframe. We then plot the data of the number of cases in each state using a variable radius circle proportional to the number of cases. This is plotted as a map using Folium.

Finally, we perform clustering to determine the optimal locations where medical camps can be setup which will maximize out chances at fighting against COVID-19. The KMeans clustering is done using Scikit-Learn python package. We then plot a circle for each of the clusters on the same map using Folium.

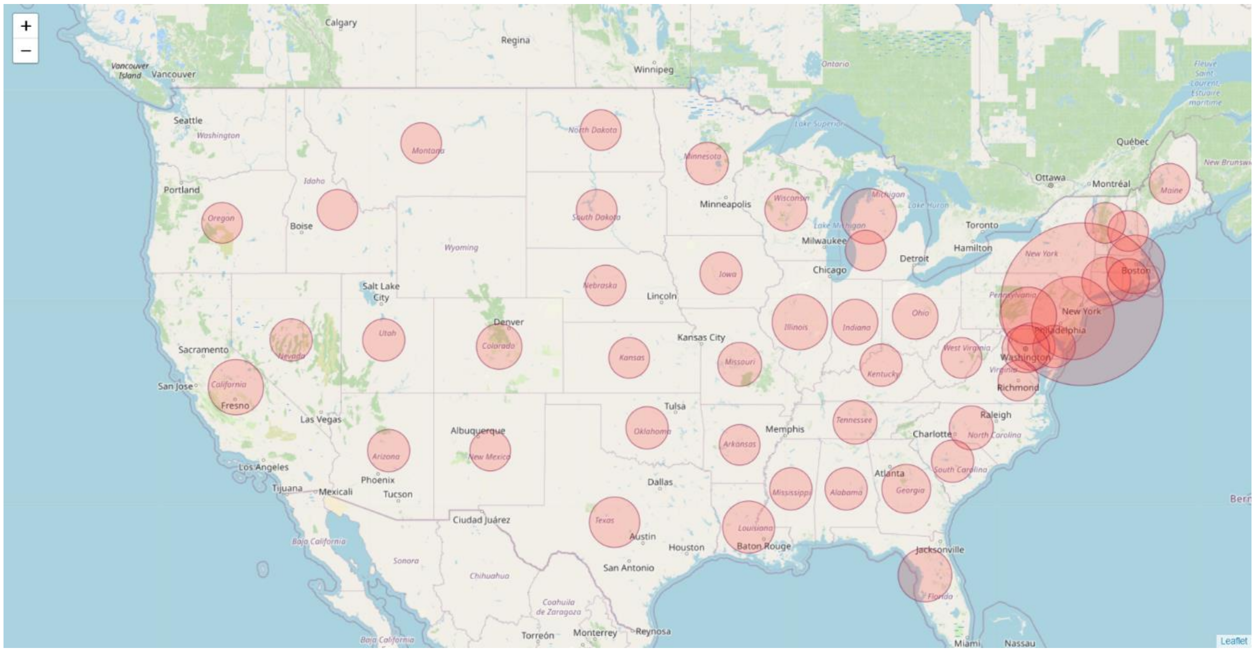
6. Results

The results from the exploratory analysis of the data is shown below.

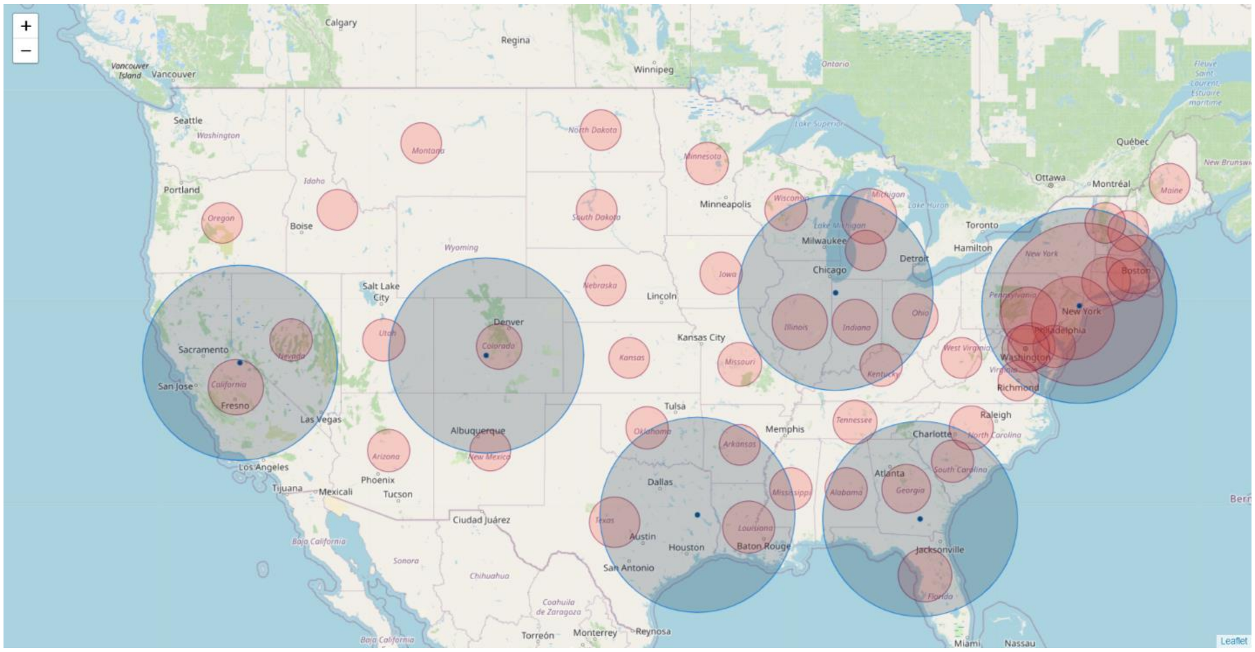


From the above plot, we can see that New York has almost 3 times as many cases as New Jersey (which has the second greatest number of confirmed cases). This is an astounding fact but it is justified by the fact that the population density of NY is almost **22 times** that of NJ! Moreover, NYC is one of the top tourist destinations of the US and it has the highest influx of foreign population from various foreign countries.

The plot of the number of COVID-19 cases in each state with the radius of the circle being proportional to the number of cases is shown below.



The results of the KMeans algorithm show the optimal spots where setting up medical camps would be most effective and efficient. The plot of the results of the clustering is shown below.



From the above plot, it is clear that setting up medical camp in the above 7 spots would be most optimal. Moreover, it is hard to see, but there are two cluster centres in NY and therefore, NY requires 2 medical camps to be set up and most attention should be focussed on NY.

7. Discussion

As noted in the Results section, setting up 7 medical camps in the 7 spots (2x NY, Indiana, Atlanta, Texas, Denver and California) would be most optimal to fight against COVID-19.

These spots are the spots that are at the centre of all the chaos and stopping COVID-19 in these spots would most definitely help flatten the curve and prevent future cases.

8. Conclusion

In this report, I analyse the COVID-19 data for the United States of America and suggest k (7) spots where setting up medical camps and “red zones” would prove most effective to fight against the pandemic. I used data collected from Kaggle, different python libraries and Wikipedia to base my comments and results. I used a KMeans clustering model with k = 7 to predict the 7 places to set up medical camp. This model and analysis could be very helpful to various public/private medical or health organizations. Moreover, this analysis not only predicts the places to set up medical camp at, but also could hint at other possibilities to fight the pandemic.