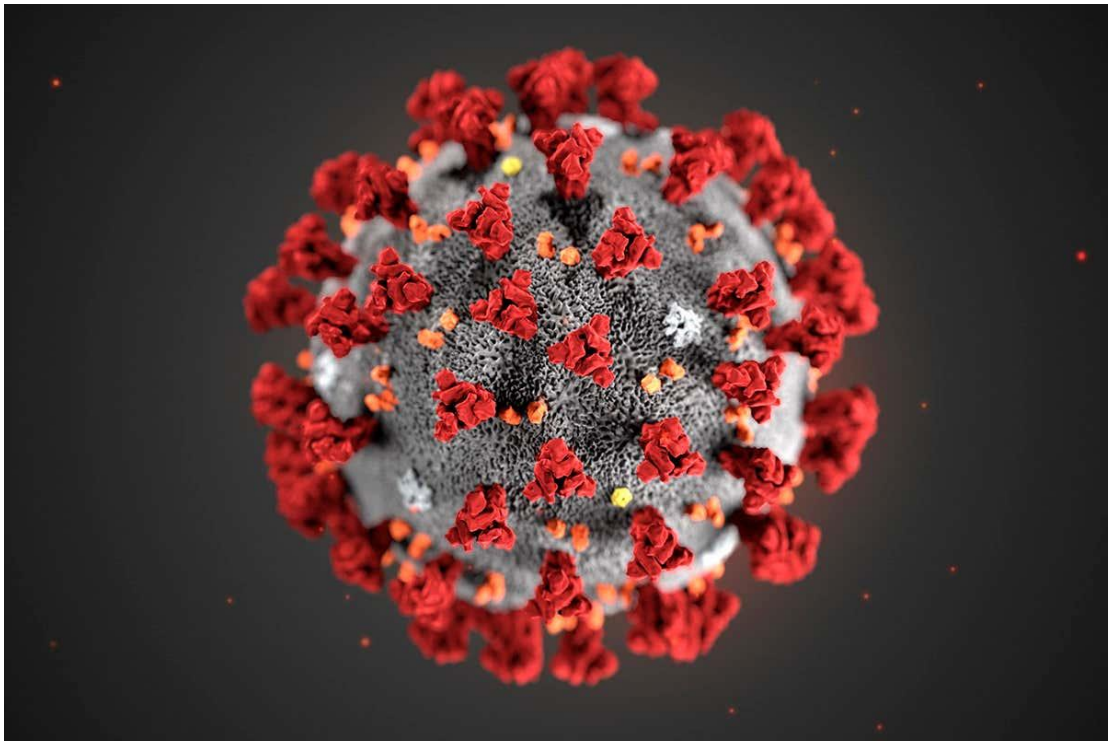
# COURSERA CAPSTONE PROJECT

IBM Applied Data Science Capstone

# 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 K-NN (K Nearest Neighbor) 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.