APPLIED DATA SCIENCE CAPSTONE – WEEK 4

ANALYZING THE OUTBREAK OF COVID-19 IN THE UNITED STATES AND USING MACHINE LEARNING ALGORITHMS TO PREDICT CONFIRMED CASES IN THE COMING DAYS

BACKGROUND

The coronavirus COVID-19 pandemic is the defining global health crisis of our time and the greatest challenge we have faced since the second World War. Since its emergence in Asia late last year, the virus has spread to every continent except Antarctica. COVID-19, short for "coronavirus disease 2019," is the official name given by the World Health Organization. As many as 213 countries and territories have registered COVID-19 cases, and the entire world is buzzing with uncertainty and questions. At the time of writing this report, there are over 4,713,026 confirmed cases of COVID-19 across the globe. The COVID-19 pandemic has been greatly affecting people's lives and the world's economy. Among many infection related questions, governments and people are most concerned with (i) when the COVID-19 outbreak will peak; (ii) how long the outbreak will last and (iii) how many people will eventually be infected.

Through this project, we will try to understand and analyze the trend of COVID 19 in different states in the US, predict its damage so that effective measures can be taken against it. The forecasts obtained from this project can help inform public health decision-making by projecting the likely impact in coming days. This understanding can help the government gauge the current level f preparedness and devote appropriate resources and medical services in regions requiring critical attention.

The stakeholders for the project are US Government and Health Department as this analysis would help them determine the resources and medical services required in the near future. This report is also beneficial to all the people who have been affected by the lockdown.

DATA

To perform the required analysis, data has been extracted from two sources, namely, Kaggle and Foursquare.
COVID-19 dataset has been taken from Kaggle. The features of the dataset are as follows :

- **Sno** - Serial number
- **ObservationDate** - Date of the observation in MM/DD/YYYY
- **Province/State** - Province or state of the observation (Could be empty when missing) **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

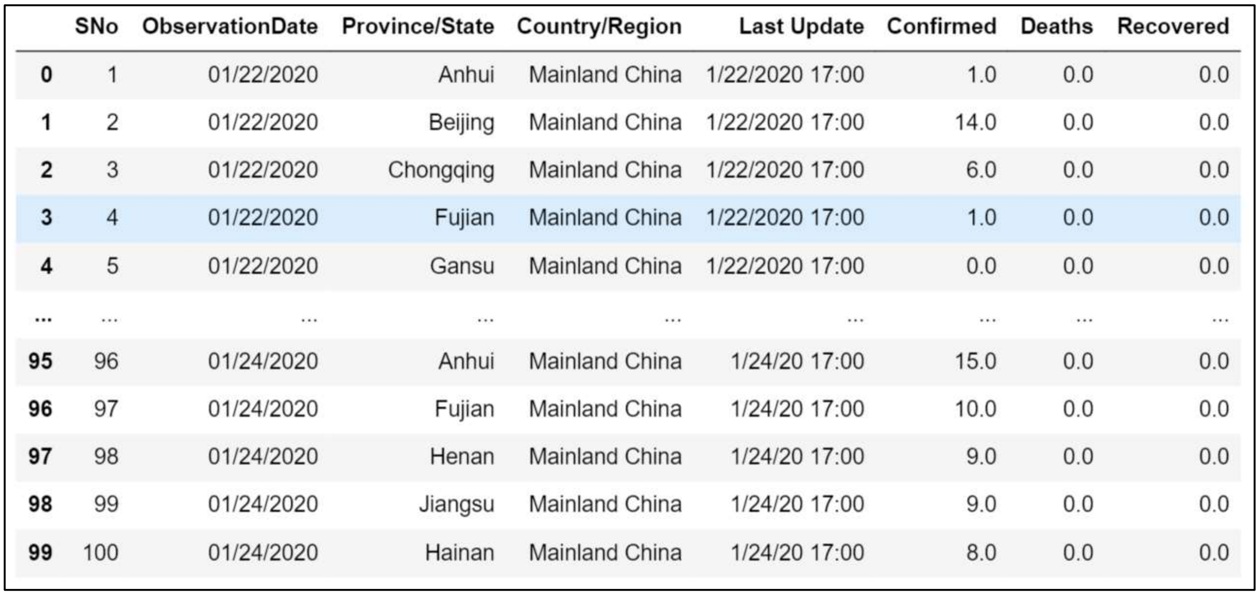
- **Recovered** - Cumulative number of recovered cases till that date

Datasource : https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

Foursquare Location Dataset

It provides data about different places all around the world.

Features and Examples: Analysis of different geographic locations including exploring places, finding trending locations near a venue, exploring users and their reviews about places can be done through Foursquare location dataset.



	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
0	1	01/22/2020	Anhui	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
1	2	01/22/2020	Beijing	Mainland China	1/22/2020 17:00	14.0	0.0	0.0
2	3	01/22/2020	Chongqing	Mainland China	1/22/2020 17:00	6.0	0.0	0.0
3	4	01/22/2020	Fujian	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
4	5	01/22/2020	Gansu	Mainland China	1/22/2020 17:00	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...
95	96	01/24/2020	Anhui	Mainland China	1/24/20 17:00	15.0	0.0	0.0
96	97	01/24/2020	Fujian	Mainland China	1/24/20 17:00	10.0	0.0	0.0
97	98	01/24/2020	Henan	Mainland China	1/24/20 17:00	9.0	0.0	0.0
98	99	01/24/2020	Jiangsu	Mainland China	1/24/20 17:00	9.0	0.0	0.0
99	100	01/24/2020	Hainan	Mainland China	1/24/20 17:00	8.0	0.0	0.0