IBM Data Science Capstone

COVID 19 in Orange County, California

Name: Yuyan Shi

1. Introduction

## 1.1 Background

The COVID 19 has impacted tremendously on economy, people’s lives, and politics. Millions of people lost their jobs. Since I had my undergraduate studies in University of California, Irvine, I am interested in analyzing the current situation of the pandemic.

## 1.2 Interest

I try to find the vulnerable groups to the pandemic. By analyzing number of cases and amount of hospitals, I cluster the cities to several groups and find the least and most serious cities. I also try to find the outcomes of the early reopen stages from the number.

****2. Data Acquisition and Cleaning****

## 2.1 Data Source

Data I used can be split into two parts: one is the data of COVID 19 cases in Orange County, downloaded from Kaggle (<https://www.kaggle.com/shubhamkulkarni01/orange-county-covid19-data>), the other is the hospital locations in each city of Orange County, using Foursquare API.

## 2.2 Data Cleaning

The number of confirmed cases for each day is for dates between March 6th and June 18th. However, since there are not a lot of cases found in the early several days. Until April 8th, there aren’t clear data for cases for different sex, age, and city. Therefore, the data I used to analyze is from April 8th to June 18th. However, there are some numbers not available in population of some cities. I searched it using google. Some cases are unknown sex. I have to leave it alone since the percentage of unknown is minor.

## 2.3 Datasets

There are eight datasets I set up.

* df\_city: columns are city, latitude, longitude, and population. I will use this dataset to locate the city and search the hospitals in cities
* df\_cases: this is the dataset includes the detailed amount of cases in female, male, and different ages for each day from April 8th to June 18th. I will use it to analyze the vulnerable groups to the COVID 19.
* df\_deaths: this is the dataset includes the detailed amount of deaths in female, male, and different ages for each day from April 8th to June 18th. I will use it to analyze the vulnerable groups to the COVID 19 as well.
* df\_hospitalized\_ICU: it has the number of hospitalization and ICU.
* df\_city\_cases: this is the accumulated cases in each city. I will use this dataset and Foursquare to analyze the relationship between pandemic severity to the amount of medical resources.
* df\_min\_per\_ten\_thous: columns are city, cases per ten thousand on June 18th.
* df\_case\_by\_sex: percentage of male and female for each day from April 8th to June 18th.
* df\_case\_by\_sex: percentage of male and female for each day from April 8th to June 18th.