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 Problem

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. Also, due to the pandemic, the lockdown policy has been enforced. Around a month ago, the reopen has happened stage by stage. I want to investigate whether the reopen exacerbate the situation.

## 1.3 Interest

Apparently, the pandemic is closely related to wellbeing of each one. I hope my analysis can show some solid facts. In such way, people can have a better understanding about the outcomes about their decisions, such as whether to reopen dine-in, or whether to go to work physically instead of work from home.

# 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.

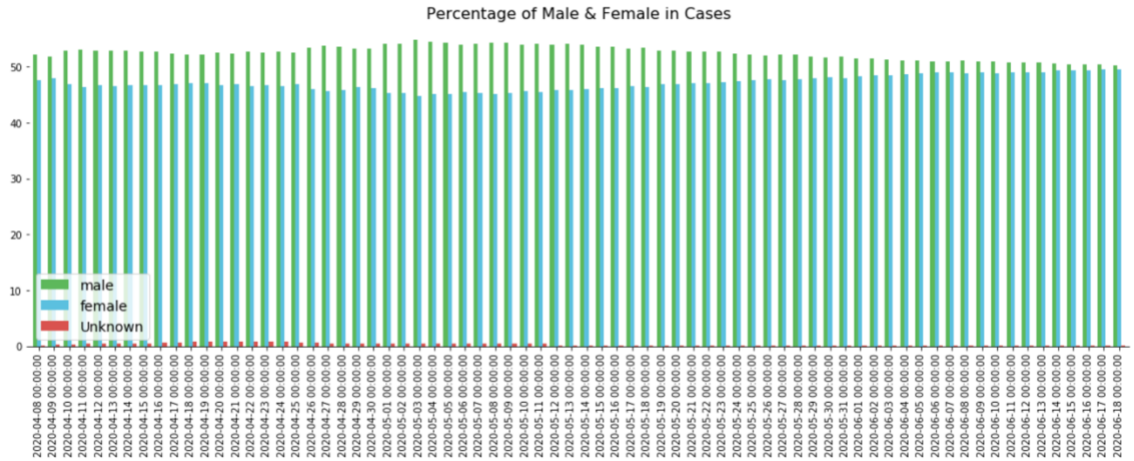
- df_city: columns are city, latitude, longitude, and population. I will use this dataset to locate the city and search the hospitals in cities
- df_cases: this is the dataset includes the detailed amount of cases in female, male, and different ages for each day from April 8th to June 18th. I will use 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.

# 3. Data Analysis

## 3.1 COVID 19 by Sex

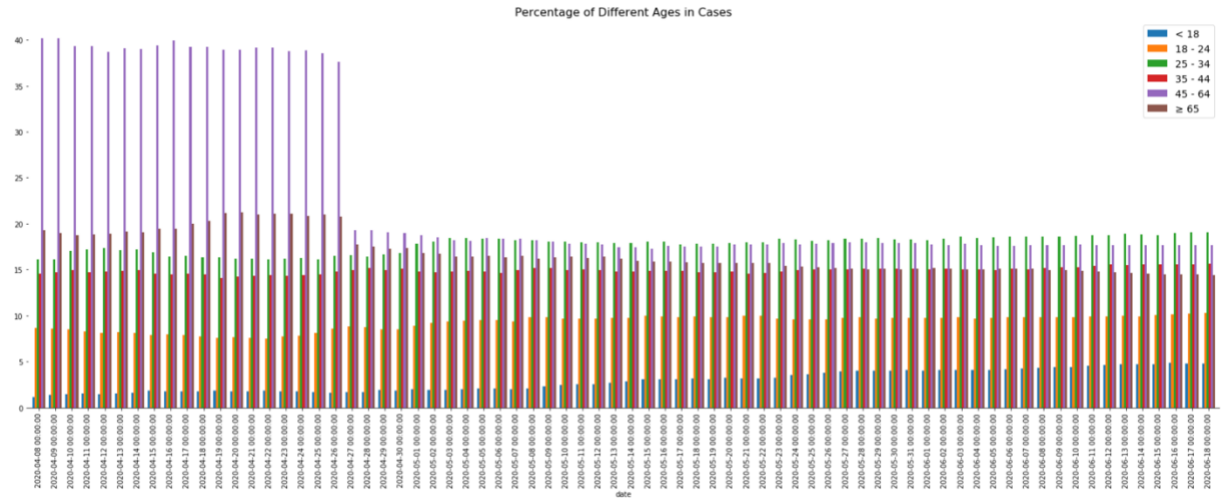
From the bar plot, we can see the significant difference between female and male. Therefore, male is more likely to be infected. As time passed and people got aware about the fact, the difference became less.




## 3.2 COVID 19 by Ages

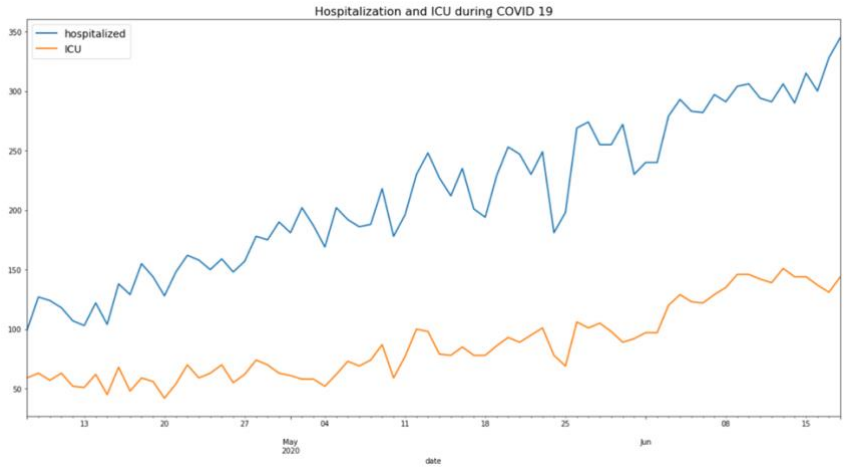
During the first two months, the two largest group is age from 45 to 64 and above 65, and people whose ages are below 18 are the least. Gradually, the percentages of people whose ages are from 45 to 64 and above 65

reduced by two times. Currently, the group that takes up the most percentage is people aging from 25 to 34.




## 3.3 Hospitalization and ICU during COVID 19

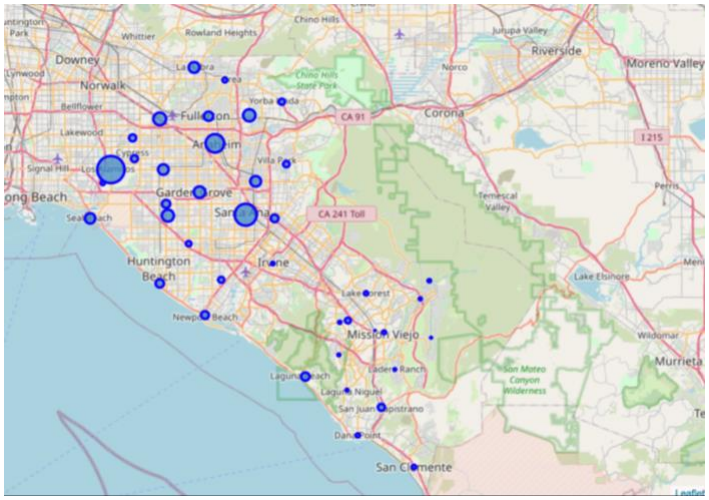
From the scatter plot, we can see several spikes, which are May 20, , , , and . Coincidentally, several days before these dates, they were either the reopen stage enforcement or the protests, which are crowded activities. This facts remind us to consider cautiously about the reopening and other gathering activities.




# 4 Model Selection

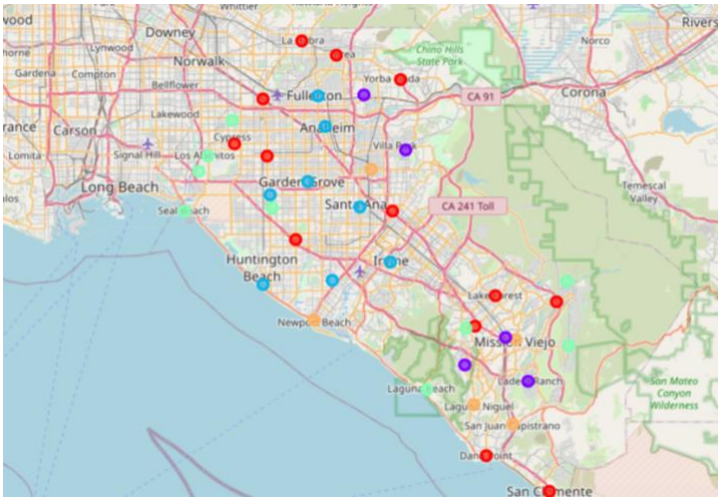
## 4.1 Different Severity in Different Cities

From the figure below, the bigger the circle size is, the greater number of cases per ten thousand people is. The next question coming up is what factors is related with the severity of the pandemic. We noticed that the more serious pandemic happens in the northwest of the orange county.




## 4.2 Relationship between Severity, Population, and Medical Resources

I choose to use K-Means Cluster methods. The figure below is the fitted output. From the map, we barely have any information gained. More details will be discussed in the discussion section.



# 5 Discussion

I used the number of hospitals in each city as proxy of the medical resources. The population and the medical resources have three levels. The same color represents the same cluster.

- Cluster 0: cities with medium population and medium medical resources. This cluster is the second serious.
- Cluster 1: cities with medium population and less medical resources. This cluster is the third serious.
- Cluster 2: cities with medium population and less medical resources. This cluster is the first serious.
- Cluster 3: cities with low population and less or medium medical resources. This cluster is the fourth serious.
- Cluster 4: cities with medium or high population and more medical resources. This cluster is the least serious.

# 6 Conclusion

From analyzing the data, we can find that the old people and male are more vulnerable, which we maybe should pay more attention or resources on them. Additionally, the sudden increase happened several days after the reopen and tests. Therefore, more cautious consideration should be brought.

The severity of the pandemic highly depends on the medical resources. Cities with more medical resources tends to have lower number of cases per ten thousand. People whose tests are positive should have more medical resources if they live in cities of Cluster 0, 3, or 4.