CAPSTONE PROJECT

An Overview of Healthcare System Capacity for Covid-19 Pandemic in Orange County, California, U.S.A

INTRODUCTION

Problem

The novel coronavirus (COVID-19) outbreak has been sweeping the United States like explosive wildfires and infected more than 1 million people as of April 30th, 2020. All states have been enforcing “social distancing” and “stay home order” to allow the healthcare system to function within its capacity. However, a growing number of admitted patients has overwhelmed the healthcare system nation-wide. Not mention the frontline health workers have to work non-stop to treat patients, but they are facing the shortage of personal protection equipment (PPE) to protect themselves properly.

Target Audience

Let’s shift our attention to Orange County, California. Government agencies need to know if the medical resources are sufficient enough to support the community. By evaluating the number of local hospitals and existing COVID-19 patients, the government agencies can perceive the general picture of the healthcare system preparedness in this area. If certain areas are in short of medical institutes, the government has to act immediately to provide financial as well as logistic support.

DATA

Data Sources

I refer to the statistics “Covid-19 Case Counts by Cities” posted on the orange county government website: https://occovid19.ochealthinfo.com/coronavirus-in-oc. The dataset contains the names of the cities in orange county, as well as populations and total diagonalized cases in each city.

In addition, I used Foursqure geolocation data to identify local hospitals. Combining the two datasets allows me to observe the distribution of medical

institutes and thus find out if the healthcare system is functioning properly in Orange County.

Data Preprocessing (COVID-19 Information)

The Orange County government website archives COVID-19 statistics data. I used Pandas HTML to scrape the dataset. As we can see from first five rows of the dataset, it gives us a general information of city names, population and total COVID-19 cases in each city. **Please note, the statements and graphics below were produced on April 30th. You may find data vary due to the growing number of patients confirmed in May.**

	City	Population	Total Cases
1	Aliso Viejo	51372	18
2	Anaheim	359339	316
3	Brea	45606	20
4	Buena Park	83384	76
5	Costa Mesa	115830	38

Next, we removed the “Population” column as it is irrelevant to my following analysis regarding the calculation of healthcare system capacity. The updated and re-indexed dataset is now shown below.

	City	Total Cases
0	Aliso Viejo	18
1	Anaheim	316
2	Brea	20
3	Buena Park	76
4	Costa Mesa	38

Using Geocoder API, we get geographic locations for the information obtained before. This process gave us latitudes and longitudes for each city in Orange County as it is shown here.

	City	Total Cases	City Latitude	City Longitude
0	Aliso Viejo	18	33.56964	-117.72691
1	Anaheim	316	33.83286	-117.91524
2	Brea	20	33.91759	-117.88899
3	Buena Park	76	33.86520	-117.99804
4	Costa Mesa	38	33.66389	-117.90239

Data Preprocessing (Hospital Information)

With preprocessed COVID-19 information in hand, now we need to retrieve local hospital information using Foursqure API.

Foursqure API is an open source for developers to retrieve geographic information. Each developer has unique Client ID and Client Secret to call API in Python. After entering my own developer credentials and search conditions in Python, I was able to gather a list of local hospital information through lines of code.

At first, I used Irvine geolocation to identify hospitals within 5000 meters range. The result is a JSON file, so I render the data to view it in form of Pandas data frame. According to the rendered table, we got a list of hospitals including animal hospitals and duplicated names. We need to program the code to get our expected results.

```
# Identify Animal hospitals and drop them off the dataframe.

def get_category_type(row):
    try:
        categories_list = row['categories']
    except:
        categories_list = row['venue.categories']

    if len(categories_list) == 0:
        return None
    else:
        return categories_list[0]['name']

hospital_irvine['categories'] = hospital_irvine.apply(get_category_type, axis=1)
hospital_irvine.columns = [column.split('.')[-1] for column in hospital_irvine.columns]

# Include only Hospitals and Medical Centers
hospital_irvine = hospital_irvine[(hospital_irvine.categories == 'Hospital')
                                  | (hospital_irvine.categories == 'Medical Center')]

# Exclude hospital names that contain animal, pet, etc.
hospital_irvine = hospital_irvine[~hospital_irvine['name'].str.contains('Animal')]
hospital_irvine = hospital_irvine[~hospital_irvine['name'].str.contains('Pet')]
hospital_irvine = hospital_irvine[~hospital_irvine['name'].str.contains('Veterinary')]
hospital_irvine = hospital_irvine.reset_index(drop = True)
print(hospital_irvine.shape)
hospital_irvine.head()
```

As shown above, I defined a function to filter the hospitals whose type is either “Hospital” or “Medical Center”. However, the result still gave us a list of animal hospitals. Hence, I applied few lines of command to exclude hospitals that is irrelevant to our findings. Now we have a nicely laid out data frame for later analysis.

	name	categories	lat	lng
0	Hoag Hospital Irvine	Hospital	33.660804	-117.772452
1	Kaiser Permanente Hospital	Hospital	33.657947	-117.774132
2	hospital	Hospital	33.665738	-117.769588
3	St Joseph Hospital	Hospital	33.682416	-117.807061
4	Orange Coast Memorial Hospital	Hospital	33.661222	-117.784409

By using a for loop in Python, we iterate the process to identify reachable local hospitals near each city. The result is tedious to read, so I made a table to understand the number of hospitals for each city.

	City	Number of Hospitals
0	Garden Grove	15
1	Santa Ana	14
2	Villa Park	13
3	Orange	13
4	Costa Mesa	12

METHODOLOGY

Data Exploration

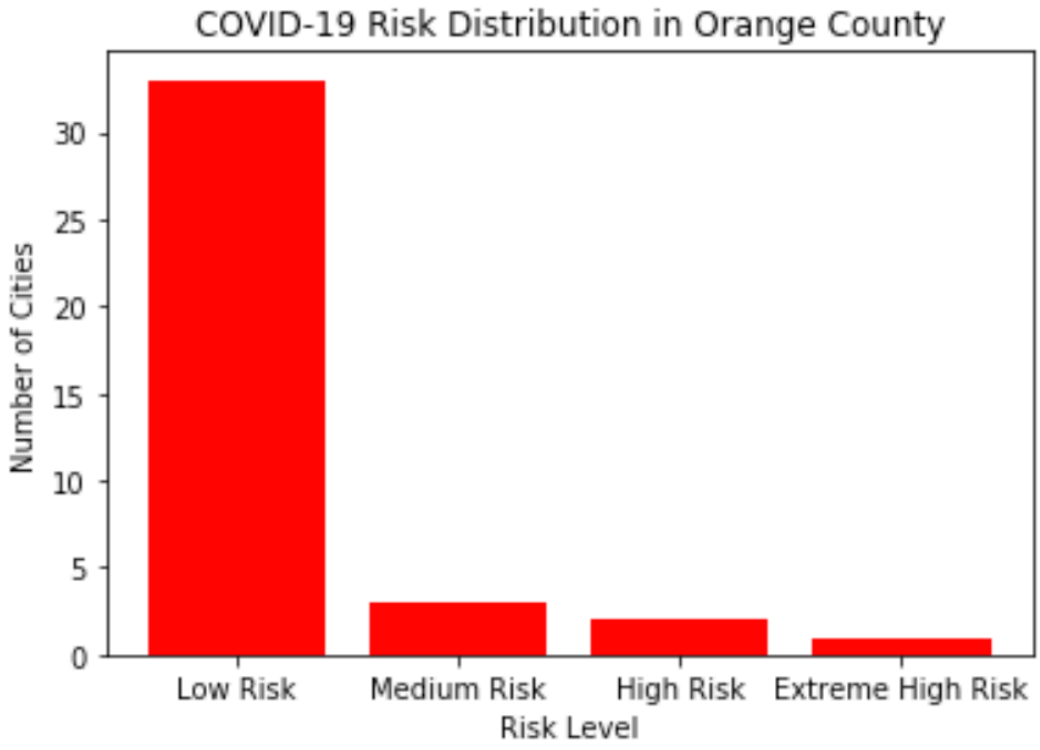
To understand the distributions of COVID-19, I use binning method to show the statics of existing cases. Therefore, I divide the data into four bins:

1. Low Risk Region: 5~82 existing cases
2. Medium Risk Region: 82~159 existing cases
3. High Risk Region: 159~236 existing cases
4. Extreme High-Risk Region: 236~313 existing cases

	City	Total Cases	Latitude	Longitude	Risk Level
0	Aliso Viejo	18	33.56964	-117.72691	Low Risk
1	Anaheim	316	33.83286	-117.91524	Extreme High Risk
2	Brea	20	33.91759	-117.88899	Low Risk
3	Buena Park	76	33.86520	-117.99804	Low Risk
4	Costa Mesa	38	33.66389	-117.90239	Low Risk

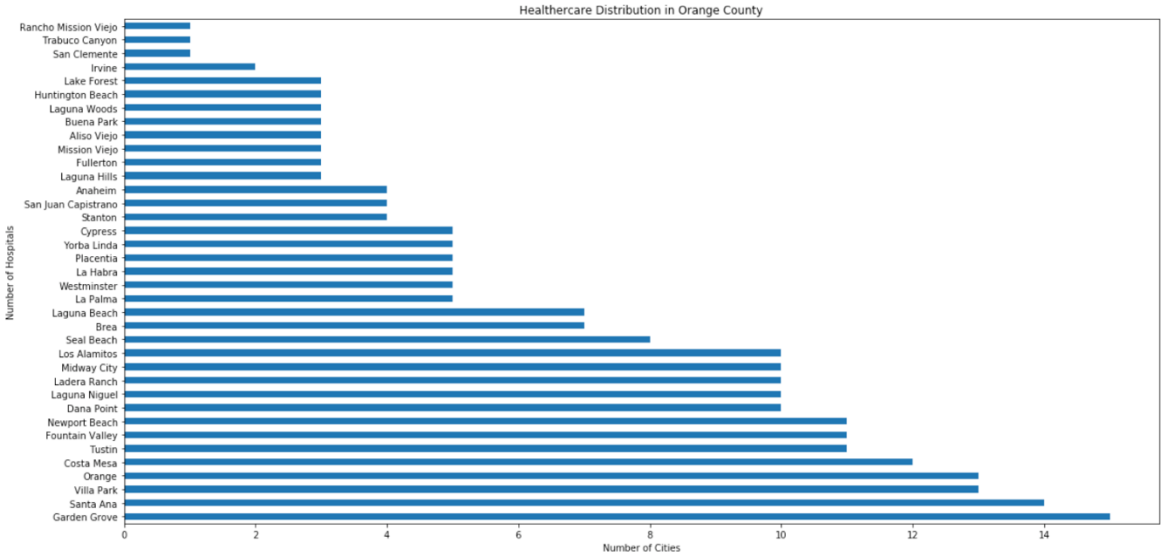
As shown above, we labeled each city with the bin they fall into to understand the risk level in every region.

However, a table like this may not be intuitive to understand, so I created a bar chart to visualize the risk distribution in Orange County. Its X-axis represents the risk level labels; the Y-axis represents the number of cities that fall into each category.

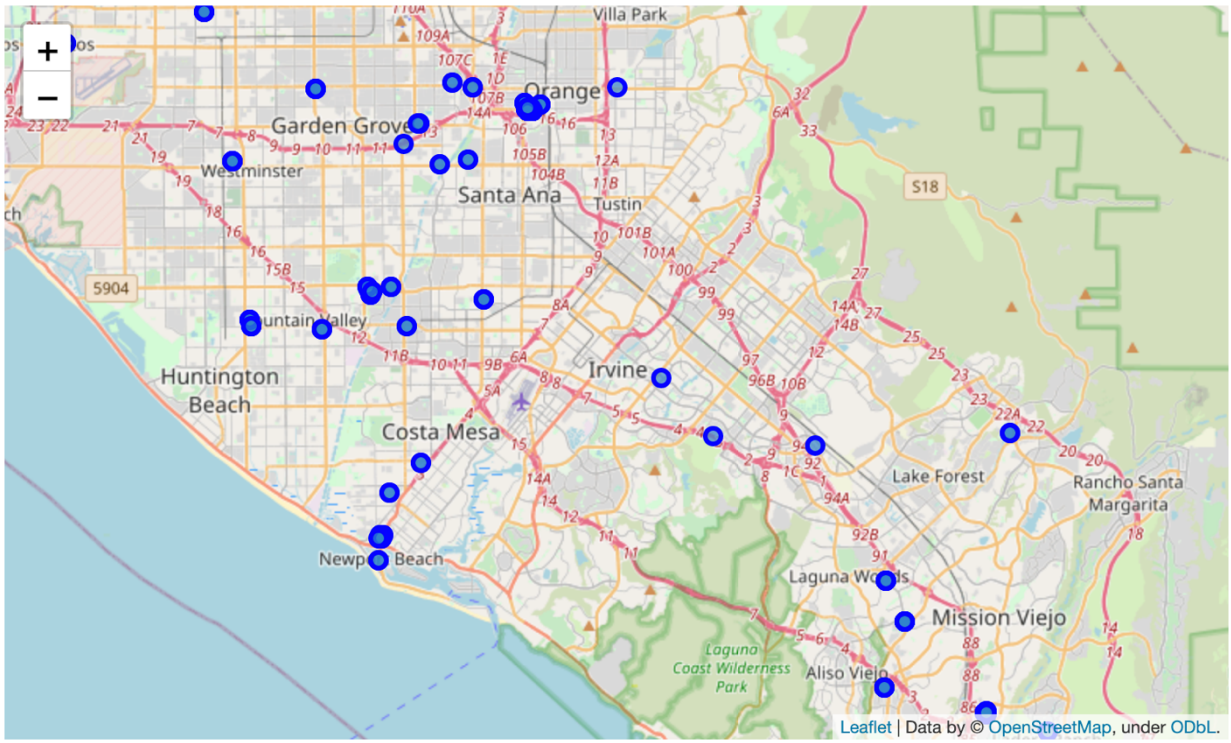


We can learn from the graph that the COVID-19 outbreak in Orange County is mild comparing to other states like New York and New Jersey. 34 cities in Orange County fall into low risk categories. However, we still have 5 cities confirmed more than 82 existing patients. **Please note, the statements and graphics above were produced on April 30th. You may find data vary due to the growing number of patients confirmed in May.**

Recall the table we made earlier to identify the number of local hospitals in Orange County, it's better to visualize it as well for further analysis.



In addition, I introduced Folium library to present all local hospitals on a map to show the distribution of healthcare resources in Orange County. They are marked as blue popups, as shown below.



Data Clustering

We have all the information we need – the COVID-19 case counts and local hospital information in Orange County. In order to study the healthcare system capacity and make constructive suggestions to government agencies, we need to wrangle the data and present the findings. Instead of applying machine learning to cluster the final dataset, I decided to calculate “Hardship Index” using obtained data and then group each region into labeled clusters.

First, since we have a wide range data distribution, I normalized the “Total Cases” and “Number of Hospitals” columns using Z-Score method. Next, I divided “Total Cases” column by “Number of Hospitals”, which gave us numerical results as we call them the “Hardship Index”. The hardship index allows us to measure if healthcare system in a city is overwhelmed with patients. Higher the hardship index, worse the healthcare system is functioning. Therefore, we may come to a conclusion that such city is reaching its maximum medical capacity and is in need of urgent supports.

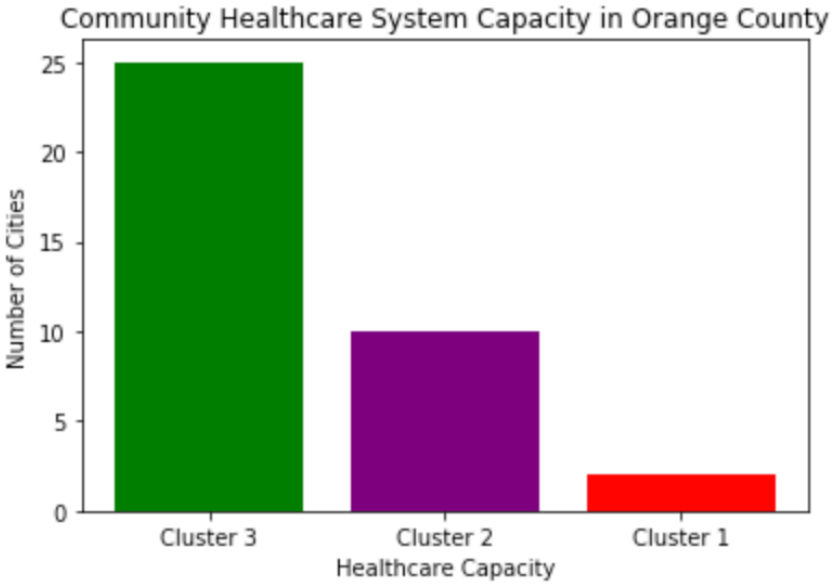
	City	Total Cases	City Latitude	City Longitude	Risk Level	Number of Hospitals	Total_z	Hospitals_z	Hardship Index
0	Aliso Viejo	18	33.56964	-117.72691	Low Risk	3	-0.526390	-0.883578	0.595748
1	Anaheim	382	33.83286	-117.91524	Extreme High Risk	4	3.291285	-0.628169	-5.239489
2	Brea	20	33.91759	-117.88899	Low Risk	7	-0.505414	0.138059	-3.660851
3	Buena Park	89	33.86520	-117.99804	Low Risk	3	0.218266	-0.883578	-0.247025
4	Costa Mesa	41	33.66389	-117.90239	Low Risk	12	-0.285163	1.415106	-0.201514

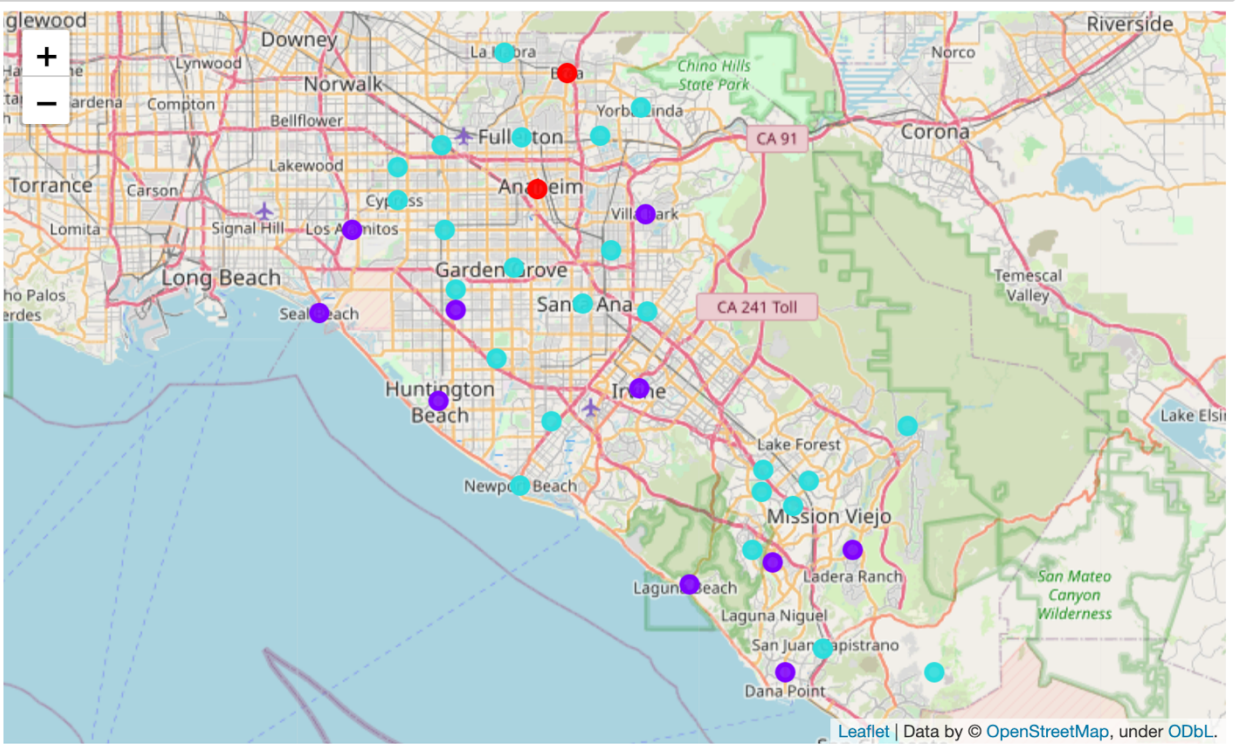
After we got a list of hardship index, we may divide them into three clusters:

- Cluster 1: Overwhelmed regions that need immediate attention
- Cluster 2: Average healthcare resources available to local residents.
- Cluster 3: Abundant medical resources in the city.

Data Visualization

Finally, I created a bar chart and map with Matplotlib and Folium Library to visualize our final dataset. Red bar and red dots represent the first cluster, purple bar and purple dots represent the second cluster, and green bar and blue-green dots represent the third cluster.





RESULTS

Data visualizations give us intuitive information on the healthcare capacity distribution in Orange County. According to obtained graphics, Anaheim and Brea are currently overflooded with COVID-19 patients and they need immediate attention from the government. Financial support as well as medical support should be implemented right away to assist local medical institutions in both regions.

Even though there are 10 cities are functioning normally within their healthcare capacity, hospitals should be well-prepared for next waves of outbreak. They should ensure enough personal protection equipment supplies and medical devices to perform medical operations on admitted patients.

More than 24 cities in orange county are in good shape to help local residents. Their flexible medical resources and health workers should be allocated to help adjacent cities to treat patients. It takes collaboration to recover in Orange County.

CONCLUSION

Government plays a crucial role in combatting pandemic outbreak like COVID-19. The agencies need to be informed with the right information to perform executive decisions. For this reason, data collection, data exploration, data analysis and data presentation are the cornerstones in the process. Government workers need to understand the cause, analyze the current data, and predict the future development, so that they can make sure healthcare system is functioning properly within its capacity to better protect their citizens at this challenging time.

REFERENCES

1. https://occovid19.ochealthinfo.com/coronavirus-in-oc
2. https://foursquare.com/