Capstone project
Finding the Best Locations for New Covid19 Testing Sites

1. Business Problem

As the coronavirus (Covid-19) outbreaks this spring 2020, total confirmed cases are over one million in the United States and total death is over 70K by May 3rd, 2020. San Diego is the 8th largest city in the US with population over 1.5 million. By May 3rd, the reported Covid-19 cases in San Diego has been over 4000 and the number keeps increasing. Covid-19 is an infectious disease. The Covid-19 virus spreads primarily through discharge from the nose when an infected person coughs or sneezes. Currently, there is no preventive vaccine for Covid-19.

Due to the rapidly increasing number of Covid-19 cases, city government of San Diego plans to setup new Covid-19 testing sites.

This project is to find the best locations for coronavirus testing sites in San Diego, California.

2. Data

In this project, data was acquired from multiply resources. First data set is the number of coronavirus cases by zip code in San Diego. The data was found at SanDiegoCounty.gov and was downloaded as a pdf file (COVID-19 Summary of Cases by Zip Code). Tabula library was imported to read data from the pdf file. From data, there is totally 4020 coronavirus cases in san Diego by May 3rd, 2020. The data summarized Covid19 cases and its zip code in San Diego area by zip code. "Count" is the total reported coronavirus cases at each zip code by May 3rd, 2020. "Rate_per_100K" is the rate of coronavirus in 100K people at each zip code.
In few areas, count number is lower than 3, and Rate_per_100K is "**" . Since these areas have low number of Covid19 cases, the demand for testing sites is low, so I decided to drop these missing data rows. After cleaning, the first data set is shown as table1 below.

	Zipcode	Count	Rate_per_100K
0	91902	31	178.4
1	91910	162	195.9
2	91911	211	249.3
3	91913	93	187.8
4	91914	20	117.2

Table1. Number of coronavirus cases by zip code (top 5 rows)

Second data set is San Diego zip code and its corresponding neighborhoods. The data was acquired from Superior Court of California Zipcode and Neighborhoods. Beautiful soup was

used to scrape the web data and summarized in table2 below. Table2 concludes “Zipcode” and its corresponding neighborhood names in “Name” column. “Venue” is not a useful information in this study and will been dropped.

	Zipcode	Name	Venue
0	91901	ALPINE	EAST
1	91902	BONITA	SOUTH
2	91903	ALPINE (POB)	EAST
3	91905	BOULEVARD	EAST
4	91906	CAMPO	EAST

Table2. Zip code and corresponding neighborhood

Then I combined table1 and table2 based on the same zip code, therefor table3 was created, it concludes Zip code, neighborhood name, count of Covid19 cases and rate of Cocid19. I also calculated population by (Count/Rate_per_100K * 100K).

	Zipcode	Count	Rate_per_100K	Name	Population
0	91902	31	178.4	BONITA	17376
1	91910	162	195.9	CHULA VISTA	82695
2	91911	211	249.3	CHULA VISTA	84636
3	91913	93	187.8	CHULA VISTA	49520
4	91914	20	117.2	CHULA VISTA	17064

Table3. Coronavirus cases by zip code

Third data set is latitude and longitude information. It was accessed by google geocode based on zip code and name of neighborhoods. Two coordinates columns were appended to table3 and renamed to table4.

	Zipcode	Count	Rate_per_100K	Name	Population	Zipcode_API	latitude	longitude	location_API
0	91902	31	178.4	BONITA	17376	91902	32.670358	-117.014674	Bonita, CA 91902, USA
1	91910	162	195.9	CHULA VISTA	82695	91910	32.638513	-117.061755	Chula Vista, CA 91910, USA
2	91911	211	249.3	CHULA VISTA	84636	91911	32.605974	-117.044101	Chula Vista, CA 91911, USA
3	91913	93	187.8	CHULA VISTA	49520	91913	32.616875	-116.997015	Chula Vista, CA 91913, USA
4	91914	20	117.2	CHULA VISTA	17064	91914	32.676234	-116.944031	Chula Vista, CA 91914, USA

Table4. Number of coronavirus cases by zip code with latitude and longitude

The last data set is medical related venues. Foursquare API was used to search nearby medical venues at each area based on coordinates information. Search query is set as “medical”,

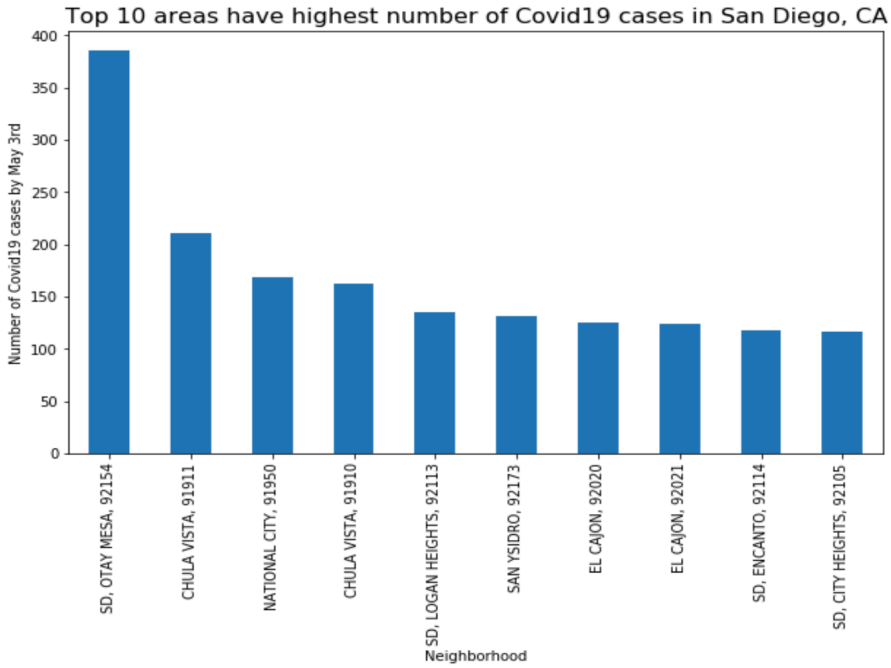
“clinic”, and “hospital”. I cleaned unrelated and repeated venues, such as pet hospitals. Medical venues is summarized in table below. Then I counted the number of medical venues in each zip code and it will be used for future plots.

	Address	Neighborhood Latitude	Neighborhood Longitude	Medical Center Name	latitude	longitude	Venue Category	Zipcode
0	BONITA, 91902	32.670358	-117.014674	Family & Preventive Medical Center	32.6468	-117.004	Doctor's Office	91910
3	CHULA VISTA, 91913	32.616875	-116.997015	Sharp Rees-Stealy Otay Ranch	32.6236	-116.996	Doctor's Office	91913
4	CHULA VISTA, 91913	32.616875	-116.997015	SBPMG - South Bay a Primary Medical Group	32.619	-117.02	Doctor's Office	91911
5	CHULA VISTA, 91913	32.616875	-116.997015	Sharp Laboratory Services at 765 Medical Cente...	32.6191	-117.02	Doctor's Office	91911

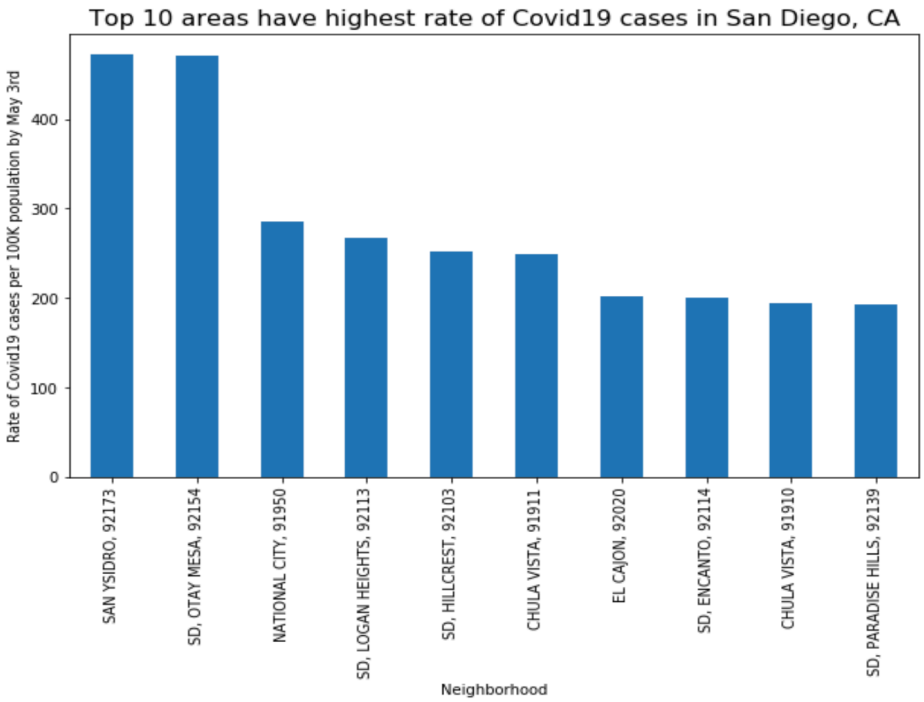
Table5. Medical related venues at each zip code

Data Analysis

Based on data from table3 Coronavirus cases by zip code, two bar plots were created**.**
First plot, Table3 data is sorted by reported cases (“Count” column) and the plot below is the top 10 areas in San Diego, which have highest number of Covid-19 cases. From the plot, Otay Mesa, 92154 has highest number of Covid19 cases.



Second Plot, Table3 data is sorted by rate of Covid19 cases and the plot below is the top 10 areas in San Diego, which have highest rate of Covid-19 cases. From the plot, San Ysidro, 92173 has highest rate of Covid19.



3. Methodology

Civid19 testing sites are built for potential patients. Both count and rate of Covid19 cases are important features while evaluating the testing site locations. Before modeling, standard scaler is used to normalize data with different magnitudes as the picture below.

```
from sklearn.preprocessing import StandardScaler
X = df_cluster.values[:]
X = np.nan_to_num(X)
Clus_dataSet = StandardScaler().fit_transform(X)
Clus_dataSet
```

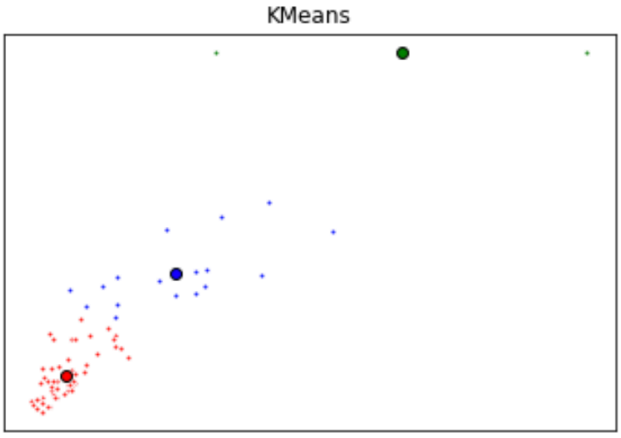
```
array([[-0.37083085,  0.7651684 ],
       [ 1.89081163,  0.96637515],
       [ 2.7367695 ,  1.58034318],
       [ 0.69956483,  0.87324517],
       [-0.56073976,  0.06151965],
       [ 0.02625142,  0.79736148],
```

After normalization, machine learning k-means clustering algorithm is applied to the unsupervised Covid19 data. Cluster number is set at 3.

```
clusterNum = 3
k_means = KMeans(init = "k-means++", n_clusters = clusterNum, n_init = 12)
k_means.fit(X)
labels = k_means.labels_
print(labels)
```

4. Results

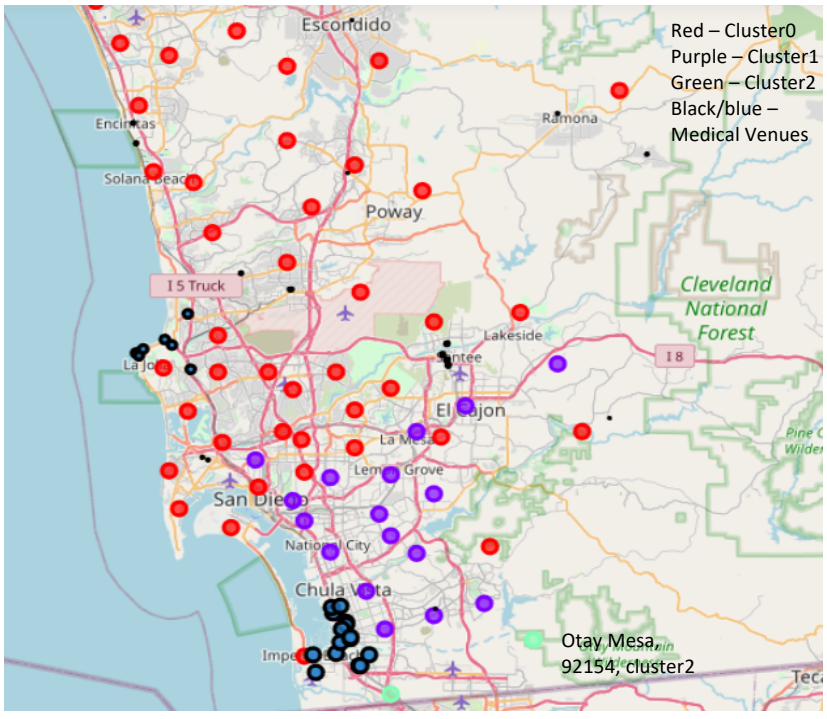
Covid19 data was grouped into three clusters. k-Means visual plot is created below. Covid19 data is well clustered to three groups.



Then I calculated the mean values of two features in each cluster. For cluster0, mean values are 28 for feature Count and 71 for feature Rate. For cluster1, mean values are 104 and 199. For cluster2, mean values are 258 and 472. So Cluster2 is the group has highest count and rate of Covid19 cases. Cluster2 only includes two areas, San Ysidro, 92173 and Otay Mesa, 92154. It is recommended to setup new Covid19 testing sites in cluster2 areas.

In addition, medical venues data has been acquired by Foursquare API. Since medical centers can provide test service and treatments, I plan to locate new testing sites in areas lacks of medical service.

After combining Covid19 data and medical venues data, a folium map is created as below. Three clusters are in red, purple, and green colors. Medical venues data is in black circle with blue filling, bigger size circle represents a greater number of medical venues in that area.



There are no medical venues near Otay Mesa, 92154. However, this is the area in cluster2 with high count and rate of Covid-19 cases. It is highly recommended to set up a new testing site in Otaytes Mesa, 92154.

5. Discussion

From Foursquare API, it only returns medical venues within search range of each zip code. Medical venues out of range are not counted.
Since the Covid-19 is highly infectious, drive-thru testing sites are recommended. With drive-thru sites, people do not come in contact with other patients, which limits opportunities for disease transmission.

6. Conclusion

In this project, I used k mean clustering to analyze Covid19 cases in San Diego area. Cluster2 is the group has highest count and rate of Covid19 cases. Furthermore, I used Foursquare API to search for medical venues in San Diego area and fount out there are no medical venues near Otay Mesa, 92154. However, Otay Mesa, 92154 is belong to cluster2, which is the area with high count and rate of Covid-19 cases. It is highly recommended to set up a testing site in Otaytes Mesa, 92154.