Evaluation of SF areas of COVID-19 risk

1. Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). SARS-CoV-2, is a positive-sense single-stranded RNA virus and quite contagious in humans. Once infected, people might develop symptoms including fever, cough, shortness of breath, diarrhea, sore throat, and abdominal pain. While the majority of cases result in mild symptoms, some progress to pneumonia and multi-organ failure.

The on-going pandemic of COVID-19 that has been designated a Public Health Emergency of International Concern on 30 January 2020 by the World Health Organization (WHO), and a pandemic on 11 March 2020. As to date, over 3 millions of people over the world have been infected, with 1 million of these cases in the United States. Thousands people lost their lives, while millions of people have to follow social distancing or stay-at-home orders to protect themselves.

It is widely believed that COVID-19 is not going away soon. So the modifications to Stay-at-home Order must be guided by health risk and a commitment to equity. San Francisco is the one of the most populous city in the United States. The denser the city, the more easily disease can spread. Residents in larger cities definitely have to be more careful and engage in more intense social distancing. In this project, I would like to use data sets obtained from DataSF.org to examine different areas in San Francisco. I will consider the COVID-19 confirmed case numbers, population, health care availability, as well as venue types to identify the most risky areas in San Francisco. I hope this can be an easy guide for people to choose where to go for essential businesses, such as grocery shopping. This information might also be helpful to small business owners (e.g. restaurant owners) to make decisions about whether it is too risky to still continue businesses based on their locations.

2. Data

DataSF.org provides a lot of open data sets about San Francisco, while Four Square API is resourceful of venue information. I will use the following data sources for this project: 1) Rate of COVID-19 cases by census ZIP code from DataSF.org, which contains data updated as of 04/25/2020 about confirmed COVID-19 cases in different regions of San Francisco. This file also provides population in different areas. 2) Health care facilities information from DataSF.org, which summarizes all the health care facilities available in San Francisco. 3) Information about venues in this region from the Four Square API.

Data downloaded were converted to data frames, cleaned, and converted to other relevant information if needed, as follows.

First, I collected the confirmed cases of COVID-19 and population in regions with different zip codes. This one is straightforward from the csv file. Coordinates for each zip codes were then obtained by geolocator. Some coordinates were wrong and had to be manually corrected by finding them from Google maps.

Next, I cleaned the health care facilities file. This data set included only longitude and latitude values for each facility, but did not indicate the corresponding neighborhood or zip code. So the coordinates were converted to zip code by geopy in order to join the other tables later. This step took some work as a couple of zip codes were incorrect or not acquired. Google maps were then used to find the zip codes by the facility name manually. I also identified health care facilities more relevant to fight against COVID-19, i.e. excluding facilities such as drug treatment facilities or sexual health clinics. One hot coding was then used to collect different types of health care facilities and then summarized for each zip code.

Venue information for each zip code area was collected from Four Square API and sorted. Then I used one hot coding to examine different categories and grouped them by zip code. The frequency of occurrence of each category was calculated and the top 10 common venue categories for each zip code region were displayed in a new data frame.

These steps left us with data frames that could be joined by the same zip codes and contain information about the COVID-19 confirmed case number/population, health care availability, as well as top venue categories.

3. **Methodology**

I started with area clustering using the K-Means method by comparing the top 10 common venue categories in each zip code region. A folium map was created and showed these different clusters with distinct colors.

I then looked at all the combined data. The confirmed case distribution was visualized in a bar graph and compare with population and other information. Correlation of confirmed case rate per 10k with population and health care availability was calculated and plotted. I also examined for differences in the region clusters and see if the difference correlated with COVID-19 spread.

Finally, I assigned a risk value to each region by giving certain weights to the following factors: population, current confirmed case rate per 10k, health care facility rate per 10k, and cluster label. The regions was then divided into four groups based on their risk values and visualized on a map.

4. **Results**

All regions were clustered into 5 different groups and shown in the folium map below (Figure 1). It is quite clear that the downtown financial areas were in one group, while the other more residential areas in the other. Treasure Island and Hunters Point are two unique regions and they were each in a cluster alone.

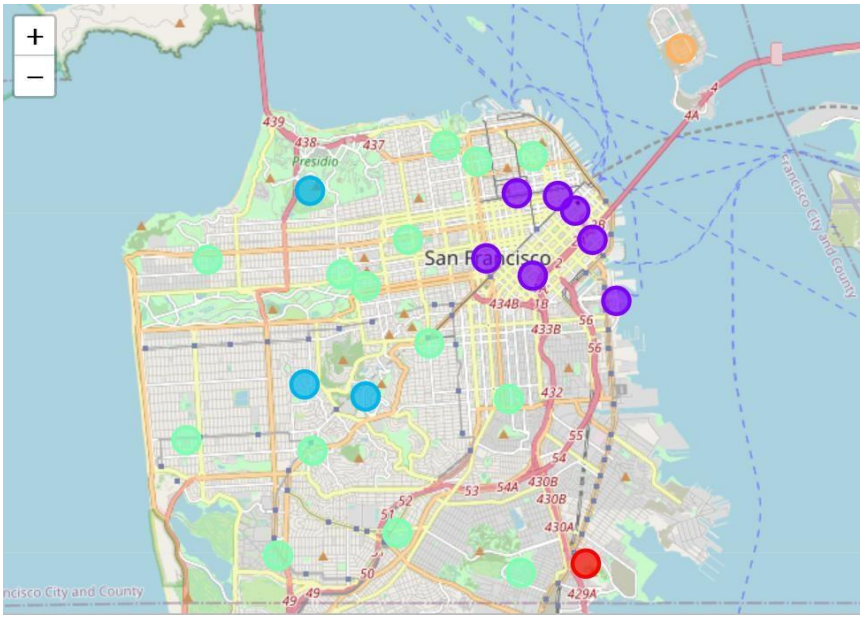

Figure 1. Folium map showing different clusters

I then visualized the confirmed case number and rate per 10k in San Francisco as shown in Figure 2. The latter is a better factor use to rule out the difference in population of each region. Zip code 94103 has so far the highest case rate per 10k and could be the most risky region.

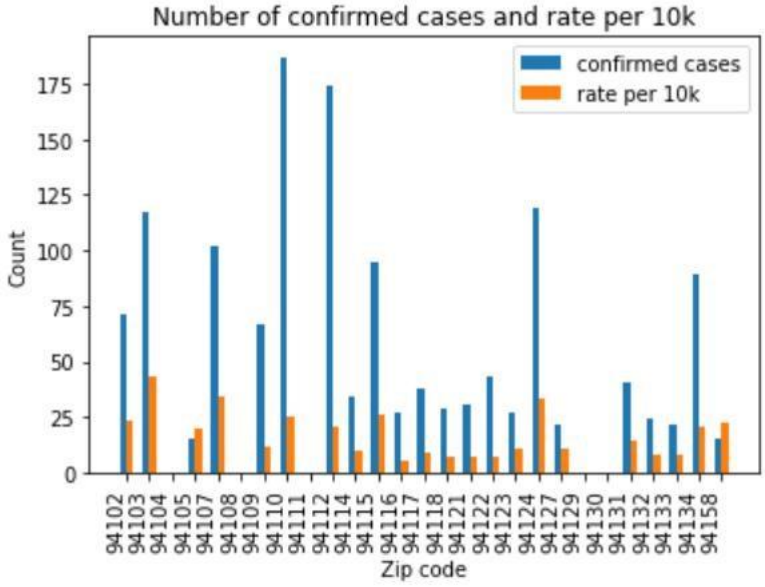

Figure 2. Confirmed case number and rate per 10k in different regions.

The health care facility rate per 10k was calculated for each region and shown in Figure 3 with the total facility numbers. Zip code 94104 showed a significant high rate compared to others. This is due to its very small size and population (~500). Thus, despite the fact that there is only one health care clinic in this region, the facility rate per 10k is rocket high. The other top regions with high facility rates are 94102 and 94158, which suggests that they have more capacity to fight again the disease. The case

rate and facility rate of each region were then displayed in Figure 4. The lower case rate and higher facility rate, the safer the region. Zip code 94103 indeed shows highest risk.

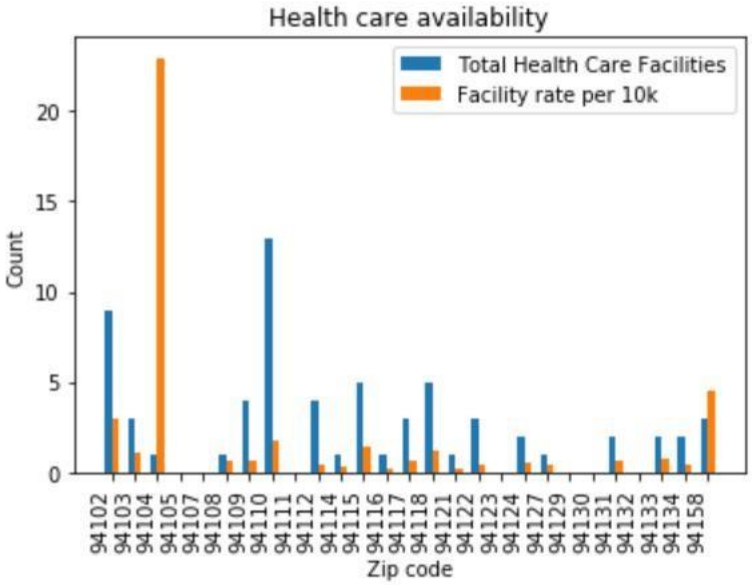

Figure 3. Total health care facilities and facility rate per 10k in all regions

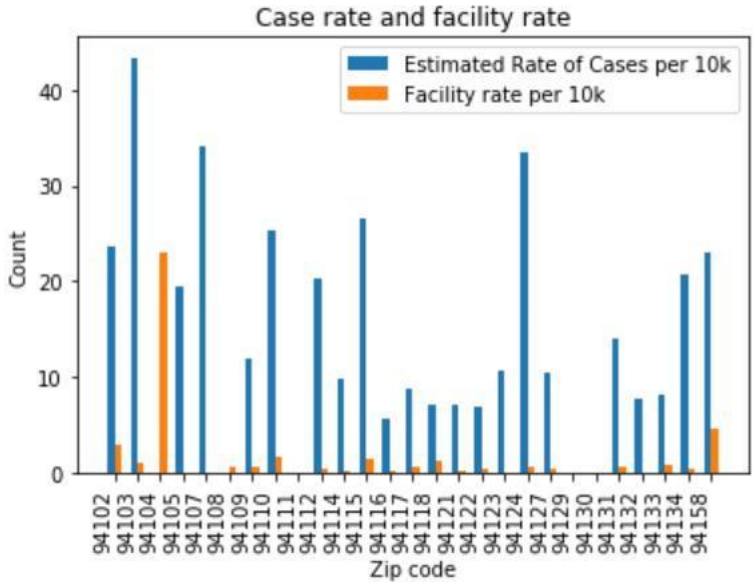

Figure 4. Case rate and facility rate per 10k in all regions

Next, I evaluated how strongly confirmed case rate is correlated with other factors, such as population and facility rate. The seaborn regression plot shows that population is very weakly correlated with confirmed case rate (Figure 5), with a correlation coefficient of 0.272794. The health care availability is negatively correlated with confirmed a case rate, as we expected, but also very weak as shown in Figure 6. The correlation coefficient is -0.150776. This means at this stage, it is not wise to use population and facility rate to make estimations of the future spread.

To explore if regions in different clusters show difference in confirmed case rate, I used scatter plot to visualize the results. Very interestingly, Cluster 2 and Cluster 4

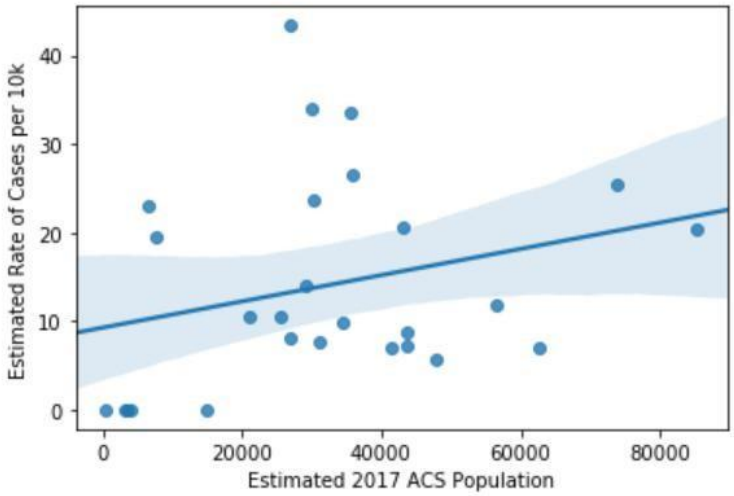

Figure 5. Regression plot of Case rate per 10 k vs. population

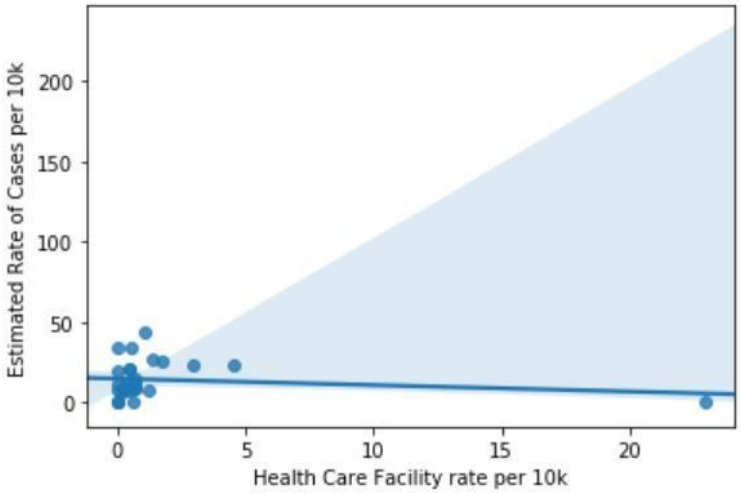

Figure 6. Regression plot of Case rate per 10 k vs. Facility rate per 10k

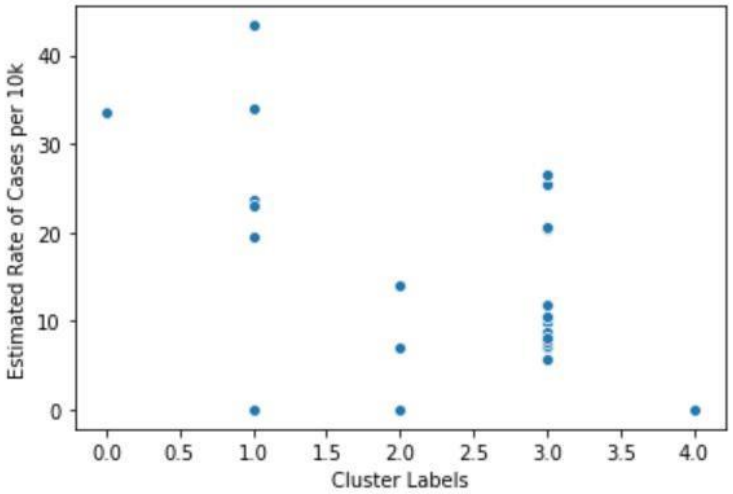

Figure 7. Scatter plot of Case rate per 10 k vs. Cluster labels

have relative low confirmed case rates. The closer examination of these two clusters (Table 1 and 2) revealed that they have open space like trails, sports field, parks, and

mountains as the top common venues. Unlike the downtown areas, where the top common venues are primarily in-doors, e.g. restaurants and bars.

Table 1. Top common venues in Cluster 2

ZIP Code	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
94122	37.753120	-122.467508	2	Trail	Chinese Restaurant	Bus Stop	Park	Dessert Shop	Bakery	Garden	Light Rail Station
94129	37.792799	-122.466140	2	Trail	Golf Course	Park	Sushi Restaurant	Café	American Restaurant	Intersection	Art Gallery
94131	37.750497	-122.451510	2	Trail	Scenic Lookout	Light Rail Station	Mountain	Tailor Shop	Sushi Restaurant	French Restaurant	Burrito Place

Table 2. Top common venues in Cluster 4

ZIP Code	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
94130	37.821707	-122.369931	4	Food Truck	Athletics & Sports	Music Venue	Baseball Field	Harbor / Marina	Rugby Pitch	Bus Station	Brewery

With the information at hand, I assigned a risk value to each zip code with following weights: 1) weights for population were 1/10k; 2) weights for current case rate per 10k were just 1; 3) weights for health care facility rate per 10k were also -1; 4) weights for Cluster 2 and 4 were -1, and 0 for the rest of Clusters. The risk values were calculated and divided into four groups (0-3). The areas with the highest risk are shown in Table 3. Therefore, 94103, 94107, 94124, 94110 are the most risky areas in San Francisco. Finally, I used a folium map to show the regions with their risk displayed in different colors (Figure 8).

Table 3. Regions with highest risk

	ZIP Code	Risk	Risk Group	Latitude	Longitude	Estimated 2017 ACS Population	Estimated Rate of Cases per 10k	Health Care Facility rate per 10k	Cluster Labels
22	94115	28.746538	2	37.782586	-122.440715	35751	26.57	1.398562	3.0
23	94110	30.970677	3	37.749984	-122.414547	73737	25.36	1.763023	3.0
24	94124	36.515693	3	37.716300	-122.394562	35492	33.53	0.563507	0.0
25	94107	37.082000	3	37.782740	-122.392789	29920	34.09	0.000000	1.0
26	94103	44.937477	3	37.775364	-122.408251	26990	43.35	1.111523	1.0

5. Discussion

The analysis showed that there was weak correlation of confirmed cases with population and also some extent to health care availability. As San Francisco is not a severe center of the pandemic, health care/hospital demands have not reached its full capacity yet, thus not showing a strong correlation. The case in New York City might be a completely different situation.

We also have to admit that this is a very simple model to assess areas in San Francisco for their risk of COVID-19 spread. In reality, more factors can come into

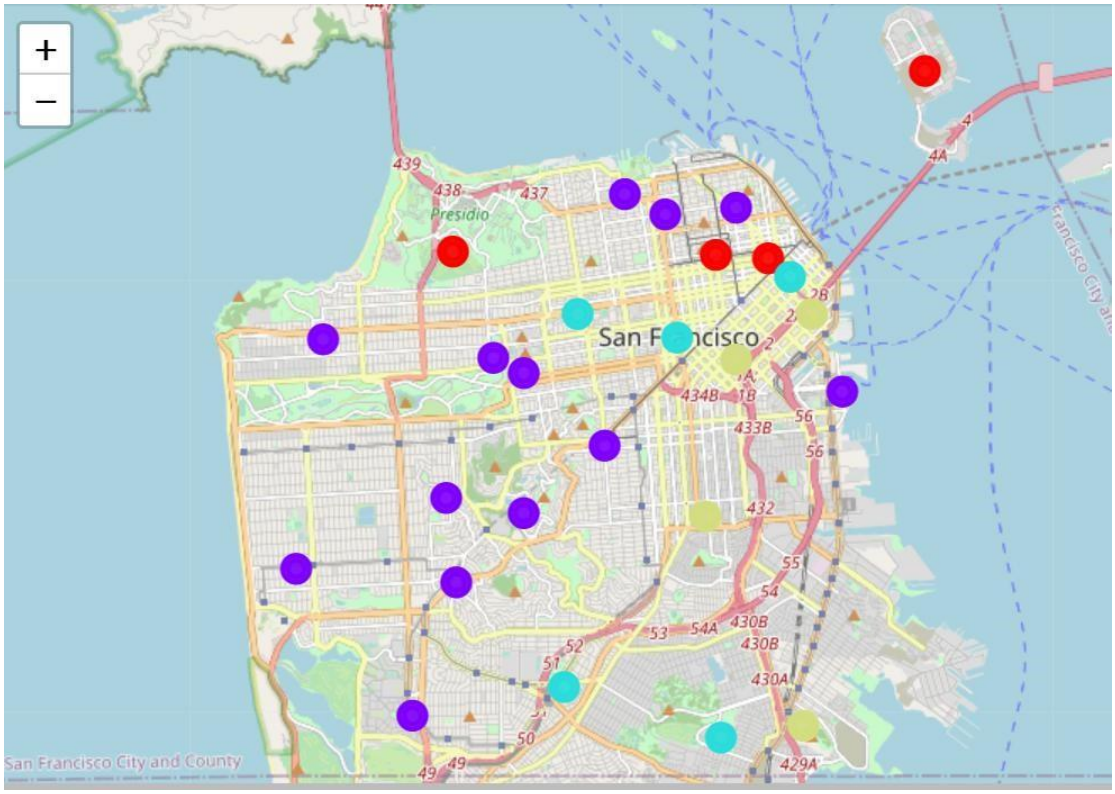

Figure 8. Folium map showing areas with risk in different colors

play and contribute to the spread in opposite directions. For instance, in-door venues like supermarkets and open space such as parks have indeed very different weights in terms of disease spread, as infection can much easily occur in the former. But we currently only assigned a label and weight to the clusters, can did not really go into details.

Besides, at the very beginning of this pandemic when people were not well aware of its existence, people have much more in-person interactions with each other. But with the enforcement of social distancing and stay-at-home orders, those interactions have been significantly reduced. Many venues such as cinemas and bars have been also forced to close, which has further limited the spread. Thus, sectioning the number of confirmed cases based on dates might help render a better illustration of the real situation. In the future, we can take those into account and improve the model for evaluation.

6. Conclusions

In summary, we have evaluated different areas in San Francisco during the current COVID-19 pandemic. We found that confirmed COVID-19 case numbers are related to population, the type of neighborhood cluster in terms of venue categories, and very weakly to health care facility rate per population. Region clustering were performed by using the K-Means method and showed that regions with more open space venues have lower COVID-19 case rate per 10k.

The purpose of this project is to identify regions with higher risk of disease spread so that people could avoid these areas for their essential business, if necessary. **The**

results suggested the region with zip code 94103 has the highest risk. In the near future, when the situation improves, policy makers could consider opening business area by area, from the safest to the most risky ones.