Analysis of Covid-19 Case and Social Venue in the San Francisco Bay Area Cities

Yi Hong

August 9, 2020

1. Introduction: Business Problem

The San Francisco Bay Area is a metropolitan region surrounding the San Francisco Bay. According to the 2010 United States Census, the region has over 7.1 million inhabitants and approximately 6,900 square miles of land. The Bay Area consists of nine counties, including Alameda, Contra Costa, Marin, Napa, San Francisco, San Mateo, Santa Clara, Solano, and Sonoma.

Since the first Covid-19 case reported in January, the rate of cases has surged across California. As of August 3, California reported a total of 517,395 positive cases with 55,976 confirmed cases in the Bay Area.

In this project, we will explore the relationship between social venues and the spread of Covid-19 in some Bay Area cities. We will investigate cities in San Mateo county and Santa Clara county which show different case patterns ranging from 1,000 to 13. Other factors such as demographics, access to healthcare, and social economics status will not considered in this project due to the limited time and resource.

2. Data Description

Based on the description of our problem, we collected data source as below:

- Reported case data by the Bay Area county health departments and collected by San Francisco Chronicle
- List of cities and towns in the San Francisco Bay Area from Wiki
- The venues of city and towns from Foursquare API

The following methodology is used to extract information from the datasource:

- Number of venues from cities are obtained from Foursquare and aggregated by category and city.
- The population density are calculated
- The infection rate per 10K people are calculated.

3. Methodology

3.1 Data preparation

Data obtained from multiple sources were combined into one table.

First, basic information of cities in the Bay Area was scrapped from Wiki. From the dataset, we decided to study cities in the San Mateo and Santa Clara except San Jose because its size and population overweighted other cities in the region. Having dropped San Jose, there are 34 cities in our bay city dataset.

	Name	Type	County	Population (2010)[8][9]	sq mi	km2	Incorporated[7]
0	Alameda	City	Alameda	73812	10.61	27.5	April 19, 1854
1	Albany	City	Alameda	18539	1.79	4.6	September 22, 1908
2	American Canyon	City	Napa	19454	4.84	12.5	January 1, 1992
3	Antioch	City	Contra Costa	102372	28.35	73.4	February 6, 1872
4	Atherton	Town	San Mateo	6914	5.02	13.0	September 12, 1923

Table 1. List of cities and towns in the Bay Area from Wiki

Second, we retrieved the latitude and longitude data from Google Map's geocode API and combined them into the bay city table shown above.

	name	type	county	population	area_km2	lat	lng
0	Atherton	Town	San Mateo	6914	13.0	37.461327	-122.197743
1	Belmont	City	San Mateo	25835	12.0	37.520215	-122.275801
2	Brisbane	City	San Mateo	4282	8.0	37.680766	-122.399972
3	Burlingame	City	San Mateo	28806	11.4	37.577870	-122.348090
4	Colma	Town	San Mateo	1792	4.9	37.674904	-122.456153

Table 2. List of cities and towns with latitude and longitude data

Third, we obtained Covid-19 case data as of August 2 from San Francisco Chronicle and added it to our dataset, as shown below.

	name	type	county	population	area_km2	lat	lng	case_num
0	Atherton	Town	San Mateo	6914	13.0	37.461327	-122.197743	27
1	Belmont	City	San Mateo	25835	12.0	37.520215	-122.275801	83
2	Brisbane	City	San Mateo	4282	8.0	37.680766	-122.399972	17
3	Burlingame	City	San Mateo	28806	11.4	37.577870	-122.348090	110
4	Colma	Town	San Mateo	1792	4.9	37.674904	-122.456153	13

Table 3. List of cities and towns with Covid-19 case data

Fourthly, we requested the name, venue category, location for venues within 1,000 meters radius of the city center from certain venue category from Foursquare API. As we focused on studying the impact of social venues, we selected food, shop and service, transport and residence venues from Foursquare category database. Other categories may have impact on the spread of Covid-19 and can be subject of further study. Below is the example of residence venue data of Foster city.

	name	city lat	city lng	city venue	venue lat	venue lng	type	category
3985	Foster City	37.558546	-122.271079	Harbor Cove Apartments	37.553401	-122.275317	Residential Building (Apartment / Condo)	residence
3986	Foster City	37.558546	-122.271079	The Lagoons Apartments	37.555003	-122.262600	Residential Building (Apartment / Condo)	residence
3987	Foster City	37.558546	-122.271079	The Plaza	37.565053	-122.270414	Miscellaneous Shop	residence
3988	Foster City	37.558546	-122.271079	Water's Edge	37.566855	-122.266883	Building	residence
3989	Atherton	37.461327	-122.197743	Matched Caregivers	37.459263	-122.193726	Home Service	residence

Table 4. Venue data of each city/town

Fifty, as we are interested in comparing the Covid-19 infection among cities, details of the venues such as location, sub-category are only necessary for high level of granularity analysis. So we aggregated the venue data by its city and venue category. The number of different venues within each category in each city are combined to dataset. The final dataset is shown below.

	name	type	county	population	area_km2	lat	lng	case_num	infection_rate	density	food	residence	shop	transport
0	Redwood City	City	San Mateo	76815	50.3	37.485215	-122.236355	1140	1484.085140	1527.137177	97.0	9.0	100.0	19.0
1	San Mateo	City	San Mateo	97207	31.4	37.562992	-122.325525	980	1008.157849	3095.764331	89.0	13.0	100.0	6.0
2	Daly City	City	San Mateo	101123	19.8	37.687924	-122.470208	817	807.926980	5107.222222	56.0	4.0	71.0	7.0
3	Gilroy	City	Santa Clara	48821	41.8	37.002983	-121.556637	615	1259.703816	1167.966507	50.0	1.0	93.0	5.0
4	East Palo Alto	City	San Mateo	28155	6.5	37.468827	-122.141075	569	2020.955425	4331.538462	17.0	NaN	56.0	4.0

Table 5. Cities/towns with venue characteristics

3.2 Exploratory Data Analysis

To illustrate the difference of those characteristics and their relation, we plotted the case number data along with other features of each city/town using bar charts in Fig.1.

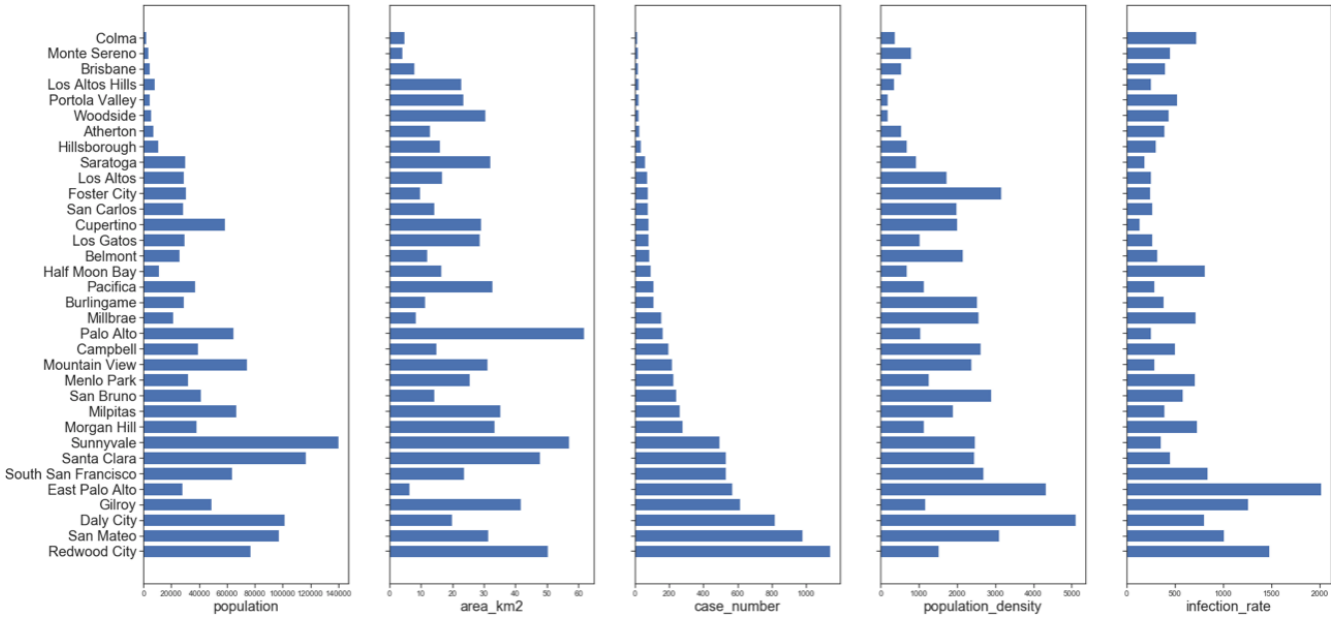

Figure 1. Comparison of population, size, Covid-19 case number, population density and Covid-19 infection rate for each city/town

The infection rate is the ratio of case number to population multiplied by 100K. It is a measure of infection risk. The bigger the infection rate, the higher infection risk. As observed from the bar charts above, infection rate appears to be correlating with population density. Therefore, we plotted the scatter plot of infection rate as a function of population density for each city/town, as shown in Fig. 2 below.

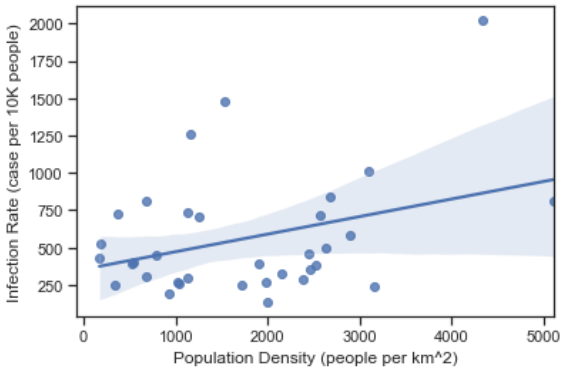

Figure 2. Relation of the infection rate with population density for different cities/towns

Having decided the metric of infection risk, we compared the difference of the number social venues and their relation with regard to infection rate using bar charts and scatter plots, as shown in the Fig. 3 and Fig.4 below.

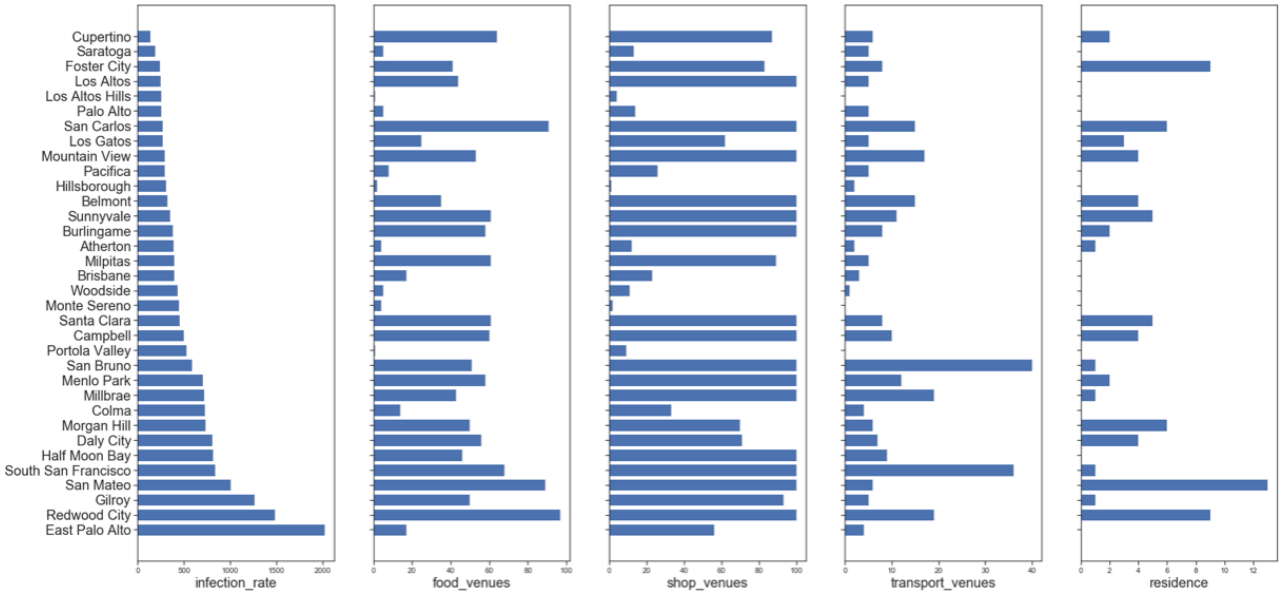

Figure 3. Comparison of Covid-19 infection rate, the number of food venues, shop and service venues, transport venues, residence venues for each city/town

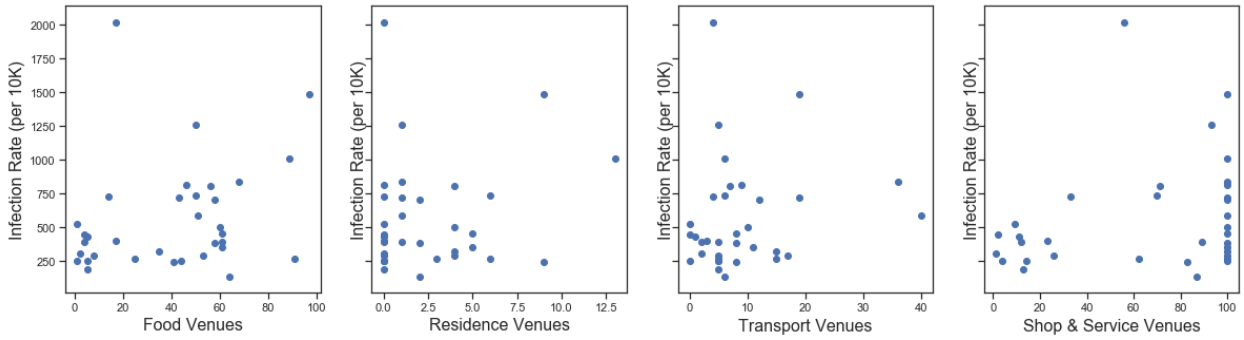

Figure 4. Relation between the infection rate and the number of food venues, shop and service venues, transport venues, residence venues for each city/town

Some may note that the number of shop venues levels at 100 for some cities. It is due to the limitation of returned results from Foursquare.

3.3 Feature Selection

From Fig. 4, we can see a corresponding increase of infection rate from food venues. We also observed a similar trend in the number of shop & services venues. For transport venues, the relation is not so clear. There are a large number of transport venues concentrates in South San Francisco and San Bruno and the rest are dispersed in other cities. We decided to discard this feature. Therefore those features including the number of shop & service venues, food venues and residence venues, population density and infection rate were used for clustering.

3.4 Clustering Modeling

After data normalization, we used K-Means algorithm to cluster the data. K-Means is one of common cluster algorithms for unsupervised learning. We used it to cluster the cities into 3 clusters because we'd like to categorize them into low, medium and high risk groups. Below is a table that combines the cluster labels with city dataset

	name	type	county	population	area_km2	lat	lng	case_num	infection_rate	density	food	residence	shop	transport	labels
0	East Palo Alto	City	San Mateo	28155	6.5	37.468827	-122.141075	569	2020.955425	4331.538462	17.0	0.0	56.0	4.0	2
1	Redwood City	City	San Mateo	76815	50.3	37.485215	-122.236355	1140	1484.085140	1527.137177	97.0	9.0	100.0	19.0	0
2	Gilroy	City	Santa Clara	48821	41.8	37.002983	-121.556637	615	1259.703816	1167.966507	50.0	1.0	93.0	5.0	0
3	San Mateo	City	San Mateo	97207	31.4	37.562992	-122.325525	980	1008.157849	3095.764331	89.0	13.0	100.0	6.0	1
4	South San Francisco	City	San Mateo	63632	23.7	37.654656	-122.407750	533	837.628866	2684.894515	68.0	1.0	100.0	36.0	1

Table 6. Dataset of cities/towns combined with corresponding cluster label

4. Results and Discussion

The results of clustering are plotted in pair shown below in Fig. 5 where cluster group 0, 1, 2 are colored by blue, orange, and green, respectively.

There is clear distinction between cities with high infection rates and population density, i.e. group 2, and other groups. Especially, East Palo Alto in group 2 demonstrates the highest infection rate in those infection rate plots against other features. Furthermore, it is noted that group 2 is different from other groups when looking into the relation between population density and those venue features. For example, for the same number of food, shop and residence venues, East Palo Alto and Daly show higher population density. It suggest that residents in those cities are likely to be underserved and need more movement to obtain the necessity of life, which could increase the risk of infection. Therefore, we conclude that group 2 has the highest infection risk given that other conditions are the same.

For group 1 and group 0, the distinction is not so great as compared to group 2. But the infection rate is relatively higher for group 0 than group 1 for the same population density, and other venues characteristics. So it is valid to conclude that infection risk in group 0 is higher than group 1.

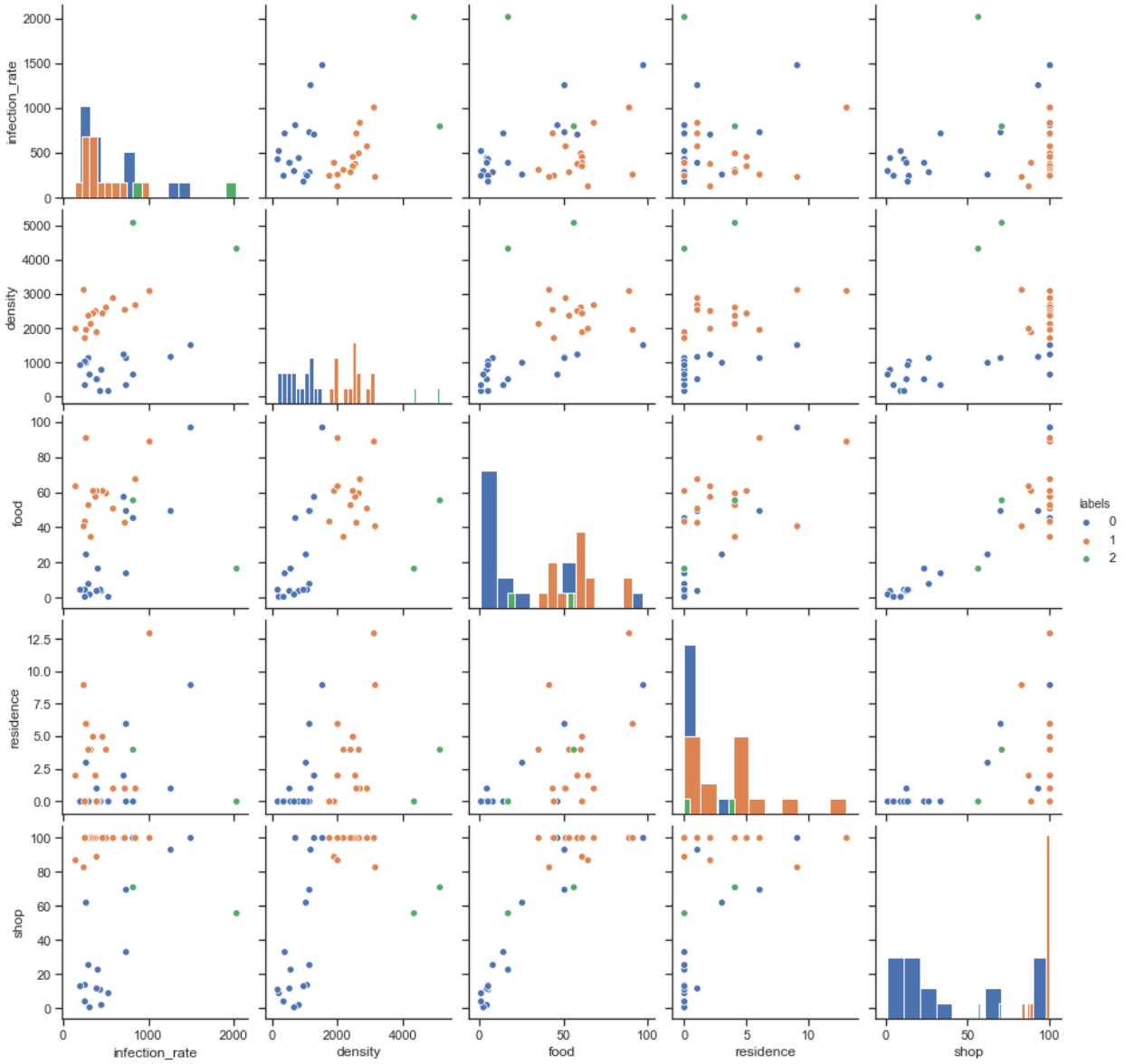

Figure 5. Pair relation between the infection rate and the number of food, shop and service, transport, residence venues for each city/town. The cluster 0, 1, 2 are colored by blue, orange, and green, respectively.

In summary, cities in group 2, group 0, group 1 are ranked from high to low with regards to Covid-19 infection risk based on our analysis.

Finally, we clustered those cities/towns to create a map where the risk ranking are colored by red, orange and green for group 2, group 0 and group 1, respectively.

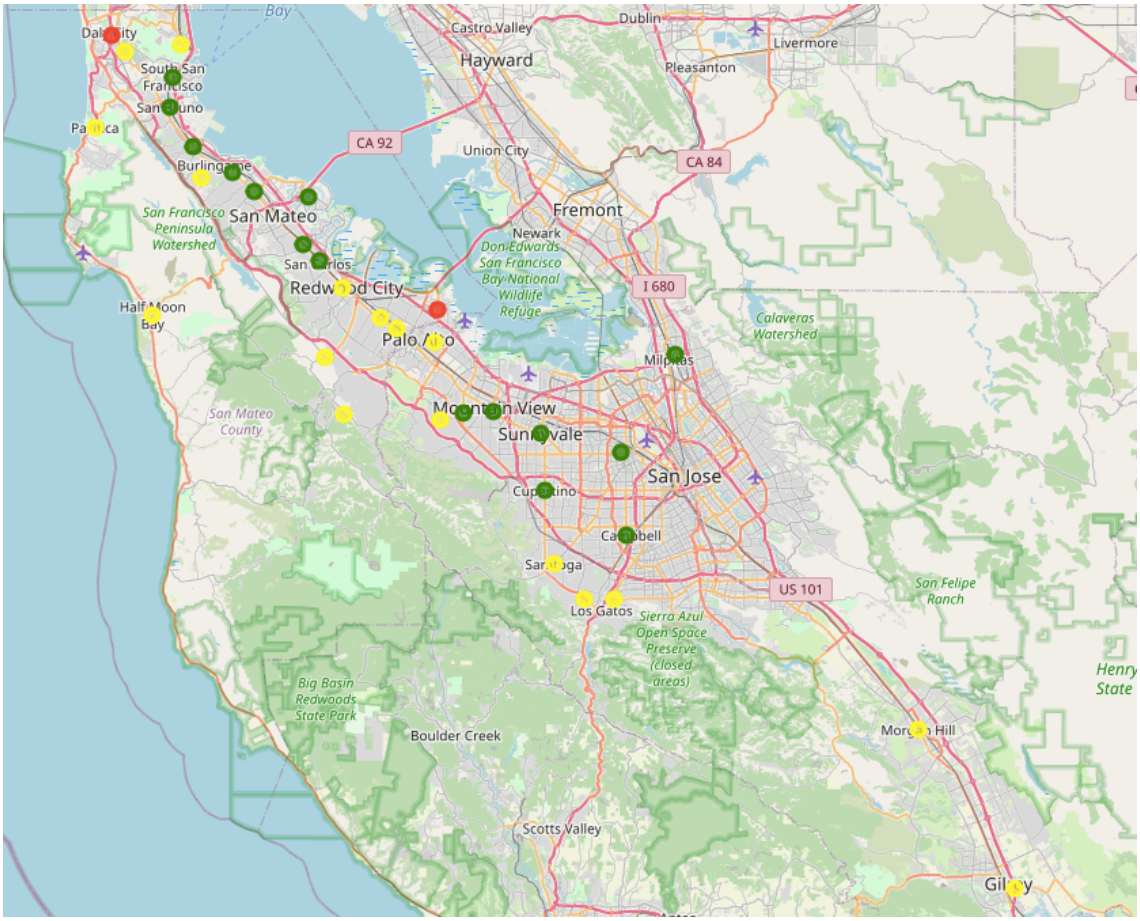

Figure 6. Cities in the Peninsular and South Bay Area with different Covid-19 infection risk levels from high, medium to low are colored by red, yellow and green, respectively.

5. Conclusion

The purpose of this project was to study the relation between geolocation data and Covid-19 infection data. We obtained geolocation data of 34 cities and towns in the Peninsular and South Bay from Google Map and Foursquare, and Covid-19 case data from San Francisco Chronicle. By exploring the data with visualization method and K-means clustering modeling, we identified three groups with different infection risk level form high to low. The final results are presented in a map.

This report is a preliminary study to demonstrate how to explore geolocation data and other data source to gain insights about the spread of Covid-19 cases. Due to the limitation in time and resource, other factors such as migration of population, access to healthcare and social economics status are not considered in this project. Those are important factors in a greater picture when addressing the challenges posed by the Covid-19 pandemic.