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

DATA SCIENCE CAPSTONE PROJECT - IBM/COURSERA

YI HONG, AUGUST 9 2020

Executive Summary

- This report is a preliminary study to demonstrate how to explore geolocation data with other data source to gain insights about the spread of Covid-19 cases.

- We obtained and analyzed geolocation and Covid-19 case data of 34 cities and towns in the Peninsular and South Bay. By exploring the data with visualization and K-means clustering modeling, we identified three groups with different infection risk levels from high to low. The final results are presented in a map.

Outline

- Introduction: Problem
- Data Description
- Methodology
- Results and Discussion
- Conclusions

Introduction

- The San Francisco Bay Area is a metropolitan region surrounding the San Francisco Bay with over 7.1 million inhabitants and approximately 6,900 square miles of land.

- 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 explore the relationship between social venues and the spread of Covid-19 in some Bay Area cities. We will investigate cities in San Mateo and Santa Clara which show different spread patterns with case number ranging from 1,000 to 13.

Description of Data

- Reported case data by the Bay Area county health departments and collected by San Francisco Chronicle (https://projects.sfchronicle.com/2020/coronavirus-map/)
- List of cities and towns in the San Francisco Bay Area from Wiki (https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_the_San_Francisco_Bay_Area)
- The venues of city and towns from Foursquare API, more information can be found at https://developer.foursquare.com/docs/api-reference/venues/explore/

Data Processing Steps

1. List of cities/towns in the Bay Area is scrapped from Wiki.

2. 34 cities/towns are selected and the latitude and longitude data of those cities are obtained from Google Map.

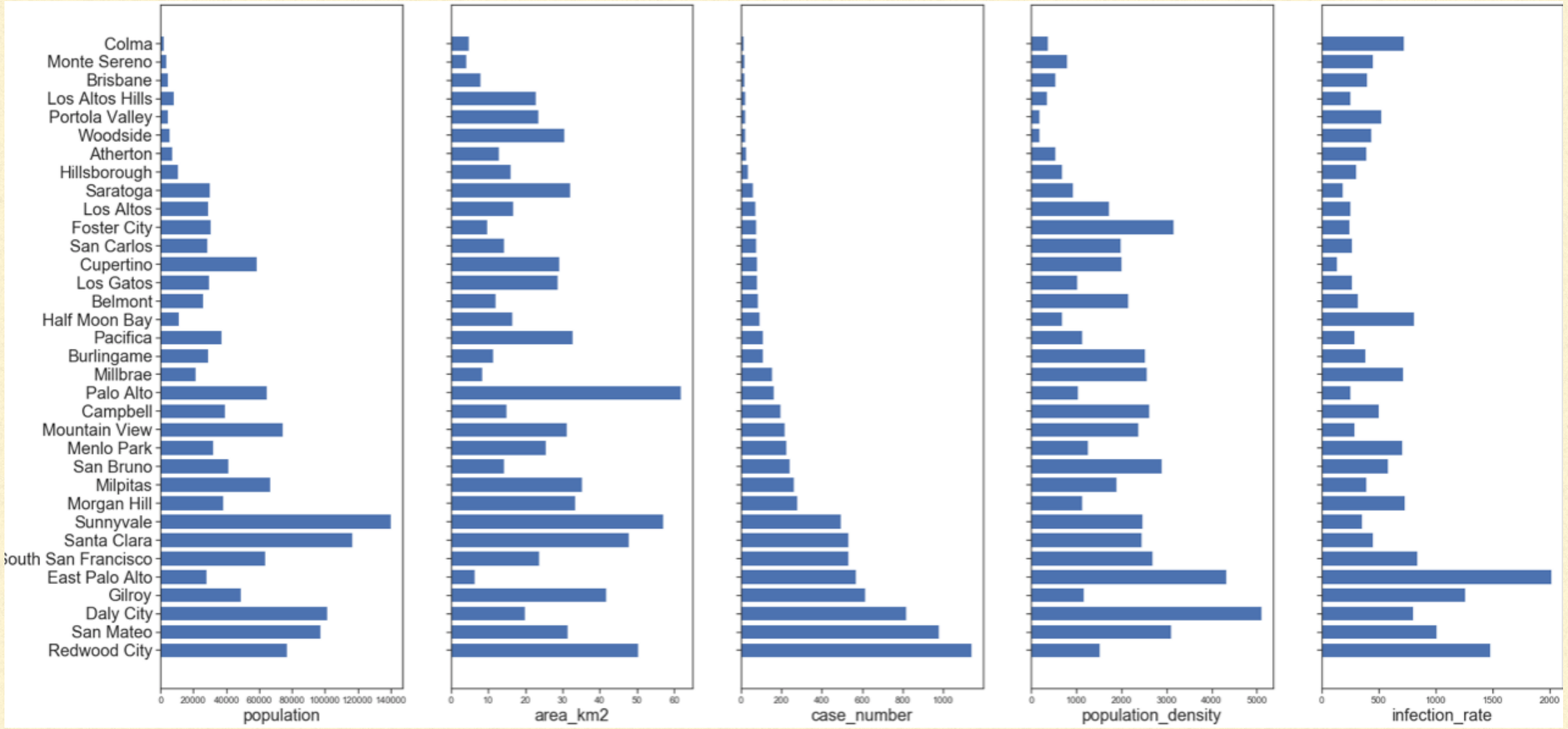
3. Covid-19 case data (as of August 2) is obtained from San Francisco Chronicle.

4. The name, category, location of food, shop, transport and residence venues within 1,000 meters radius of each city/ town center are obtained from Foursquare.

5. The venue data is aggregated by the city and venue category and combined into the dataset shown below for analysis

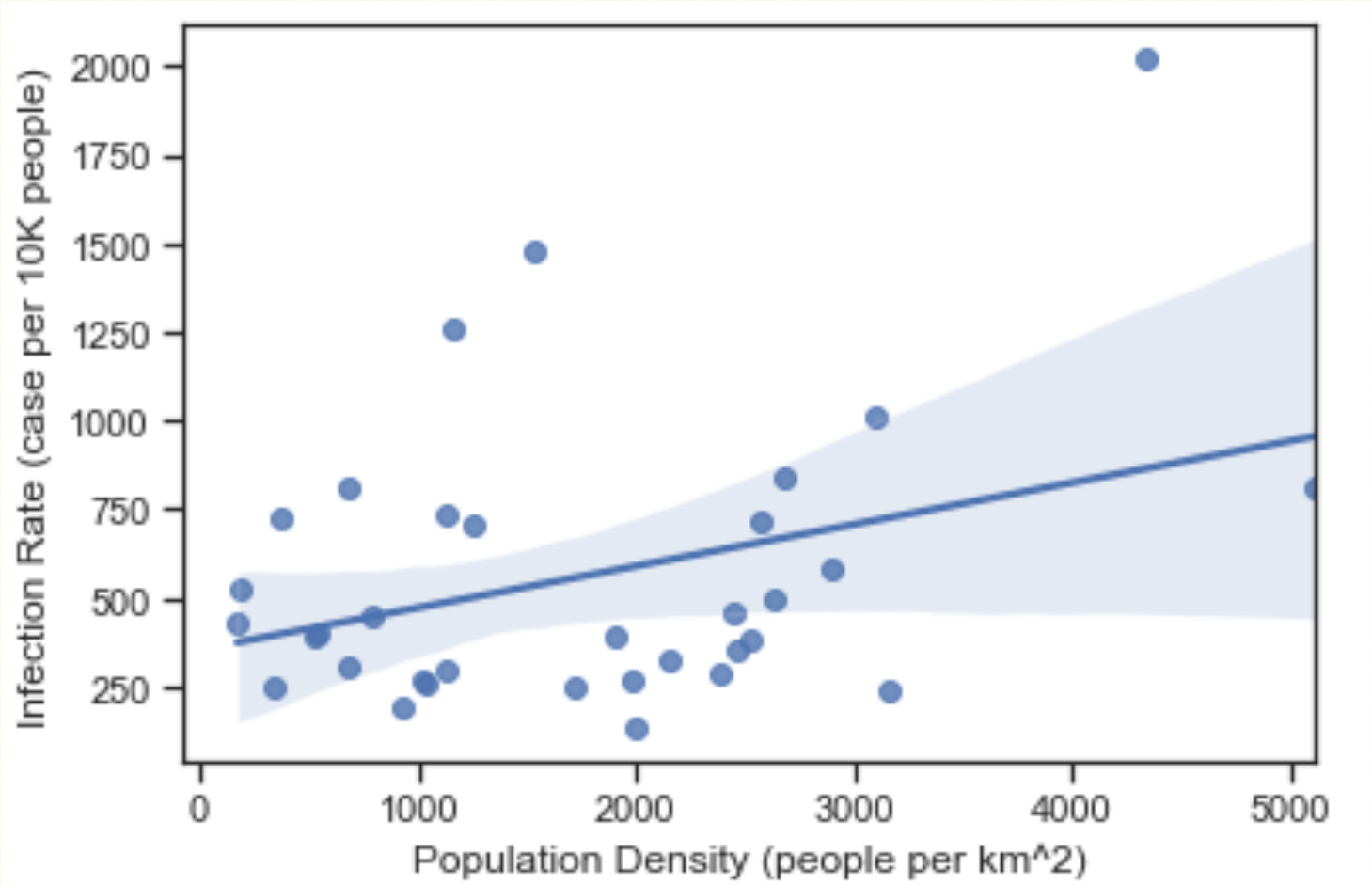
	name	type	county	population	area_km2	lat	lng	case_num	infection_rate	density	food	residence	shop	transport
0	Redwood City	City	San Mateo	76815	50.3	37.485215	-122.236355	1140	1484.085140	1527.137177	97.0	9.0	100.0	19.0
1	San Mateo	City	San Mateo	97207	31.4	37.562992	-122.325525	980	1008.157849	3095.764331	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

Methodology: Comparison of Population, Size, Covid-19 Case Number and Density



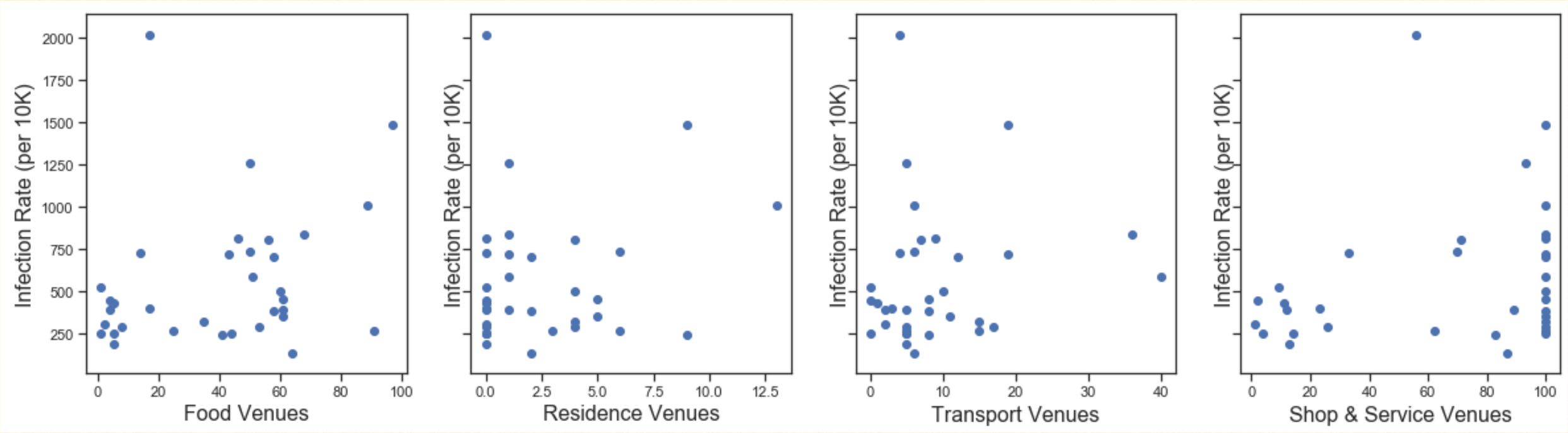
- Infection rate is the ratio of case number to population multiplied by 100K

Methodology - Feature Selection: Infection rate, Population Density



- Infection rate as a measure of infection risk increases with population density

Methodology - Feature Selection: Food, Residence, Shop & Services, and Transport



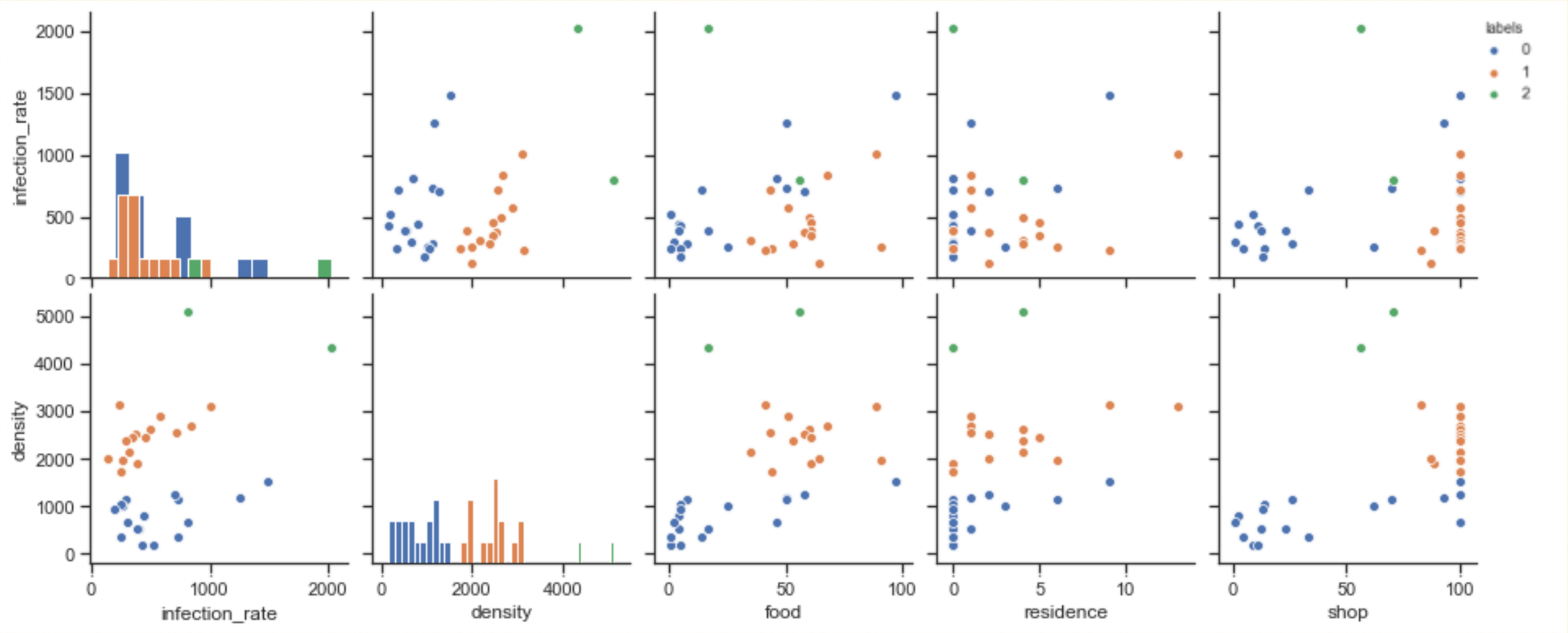
- Increase of infection rate with the number of food venues, shop & services venues.
- Transport is not selected because the correlation appears to be weak and the venues are disproportionally dispersed in the region.

Methodology- Results of K-Means Clustering (N=3)

	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

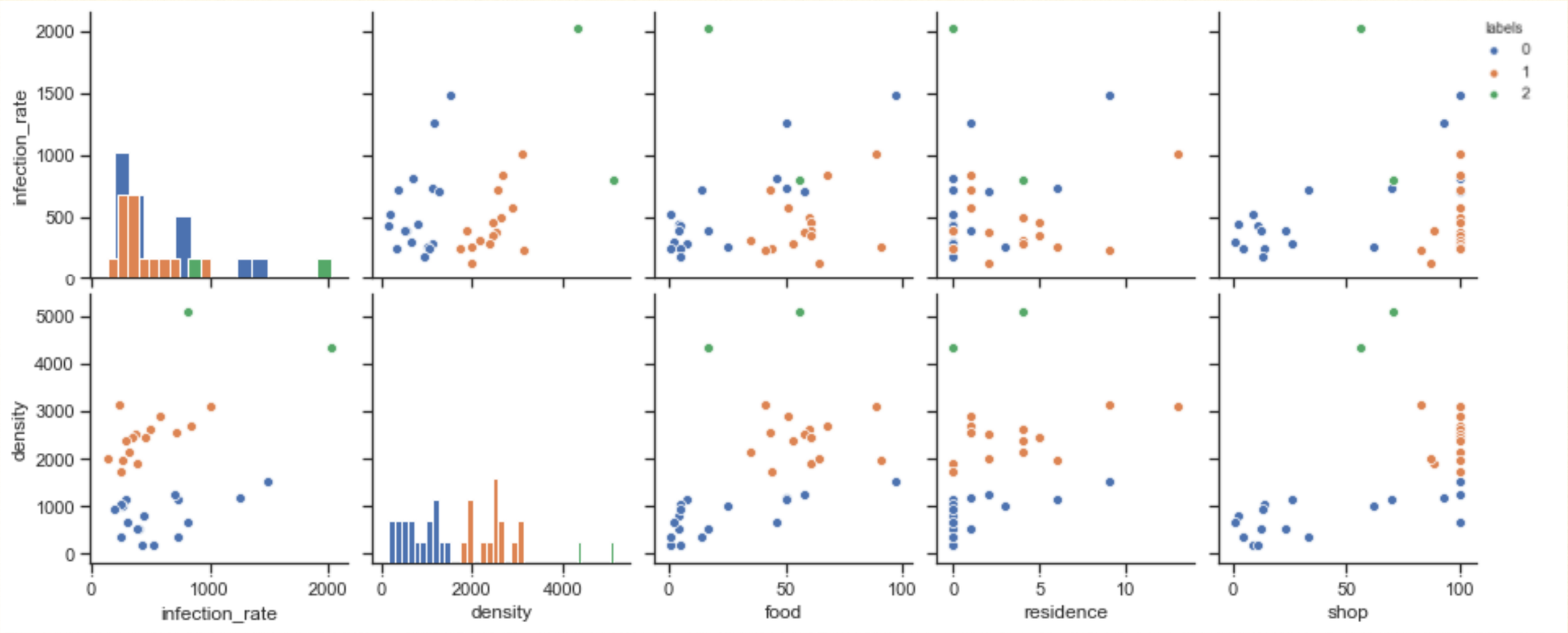
- After data normalization, K-Means algorithm is used to cluster the data.
- Clusters= 3 for K-Means modeling in order to categorize those cities in terms of infection risk from low, medium and high risk groups.
- The excerpt of the dataset with cluster labels

Results: Cluster 2 Shows Highest Infection Risk and Lowest Venue Availability by Population



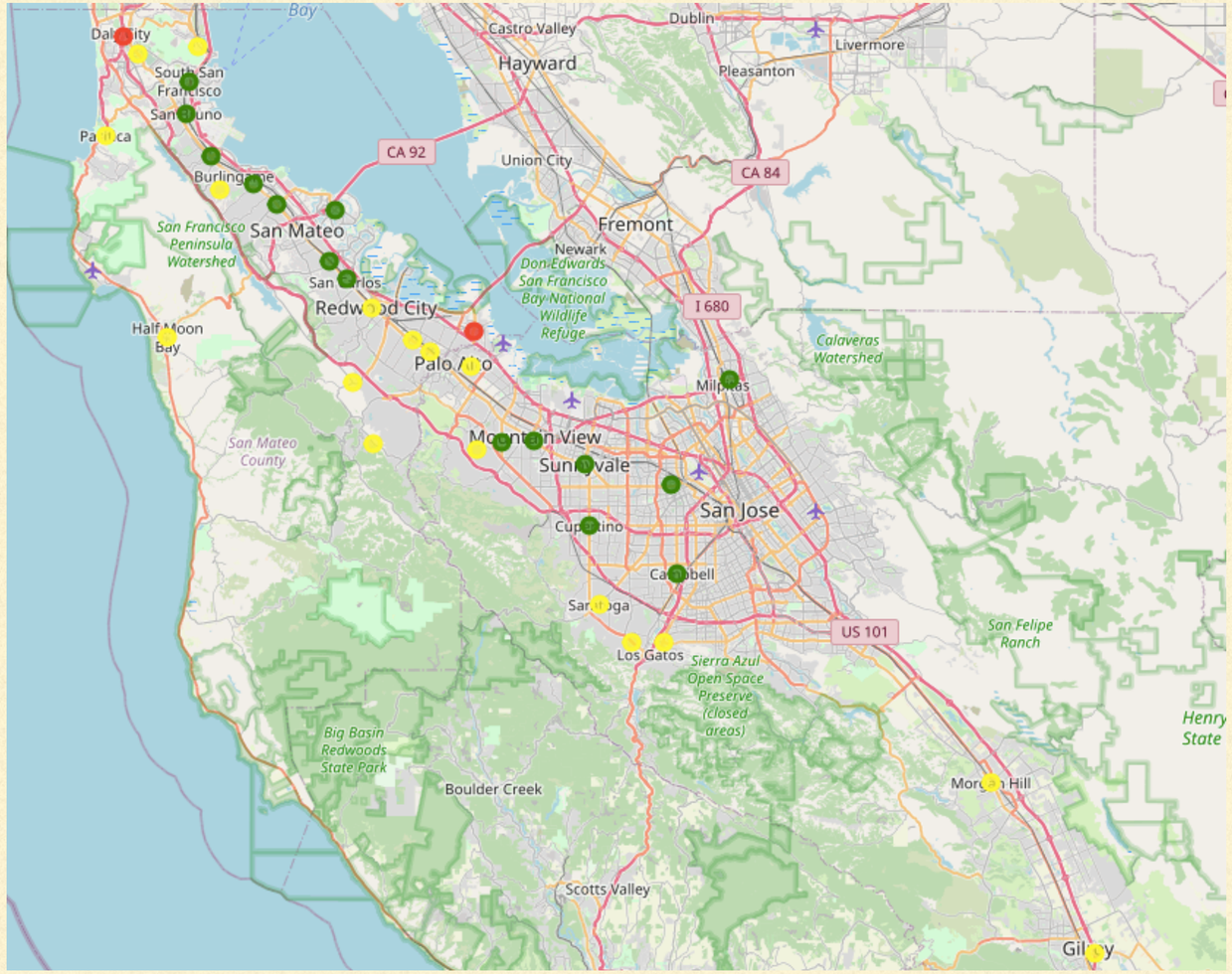
Lower venue availability by population indicating residents in cluster 2 are more likely to move for living

Results: Cluster 0 Shows Relatively Higher Infection Rate Than Cluster 1



Despite its higher venue availability, cluster 0 demonstrates relatively higher infection rate than cluster 1.

Results: Cities in The Peninsular and South Bay Area with Different Covid-19 Infection Risk Levels



- The 34 cities/towns which the infection risk ranking from high, medium, low are colored by RED, YELLOW, and GREEN respectively.

Executive Summary

- This report is a preliminary study to demonstrate how to explore geolocation data with other data source to gain insights about the spread of Covid-19 cases.

- We obtained and analyzed geolocation and Covid-19 case data of 34 cities and towns in the Peninsular and the San Francisco South Bay. By exploring the data with visualization 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.

- Other factors such as migration of population, access to healthcare and social economics status are not considered in this project. Those are also important factors in a bigger picture when addressing the challenges posed by the Covid-19 pandemic