California City Segmentation for Real Estate Industry

Rui Luo

May 12 2019

1. Introduction

1.1 Background

California state is the wealthiest and most populated state in the United State, and the real estate market in the California is very competitive. Housing is a necessary supply for during human's whole life; therefore, selecting an ideal place to live bothers everybody. The famous three-factor for buying a home, "location, location, location", tells everyone that location is the only most important factor that has to be considered before buying a home. The reason behind it is that location can determine where do people work, how long do people commute, how much do people earn, how much do people spend for living, and the life quality of people. The reason list will last because everybody has a different life so that location means different for everybody. Since location is so important for buying a home, what contribute to a location is necessary point to consider. It should be very easy to understand that what venues exist within a location forms the characteristics of a location. For example, if a location has luxury shopping mall, golf course, and high-end restaurants, this location tends to have a higher housing price. The venues exist within a location contribute the most to the characteristics of a location. Therefore, real estate market, one of the oldest and biggest market in the world, is playing with the characteristics of a location all the time. Real estate company often find and evaluate a location to build business strategy such as house pricing and home type. The nearby venues of a location have to be considered before building business strategy for a location. Furthermore, real estate company often expands its business to multiple locations such as different cities. Therefore, city segmentation becomes a very useful tools to help real estate company find similar

cities, build business strategy, and apply strategy to those cities that are similar. Thus, the venues within a city are very good and important elements for city segmentation.

1.2 Problem

This project aims to cluster a group of cities in California in order to split the cities into different groups so that the cities within a group are very similar type. Obtaining venues data for each city is the necessary process, and the cities will be grouped based on the most common ten venue types of each city.

1.3 Interest

Real estate companies would be very interested in city segmentation since they could utilize this to find similar cities and build business strategy that can be applied to these similar cities. This could reduce the cost and potentially increase the profits.

2. Data Acquisition and Cleaning

2.1 Data Sources

California cities data can be obtained from simplemaps.com, and the venues data for each city can be obtained from Foursquare. The data will have city name, 1st most common venue type, 2nd most common venue type, until 10th most common venue type for each city.

2.2 Data Cleaning

California cities data is downloaded from simplemaps.com, and it has city name, latitude, longitude, and population density for each city. I decided to use cities with 1000.0 or higher density because 1000.0 or higher density makes sure that the city has enough and significant data to be gathered.

In [120]:
```
# extract the California cities with densities that are higher than 1000.0
cities = data.loc[:,['city_ascii', 'state_id', 'lat', 'lng', 'density']].copy()
cities = cities.rename(index=str, columns={'city_ascii':'city','state_id':'state'})
cities_t = cities[cities['state'] == 'CA'].reset_index(drop=True)
CA_cities = cities_t[cities_t['density']>=1000.0].reset_index(drop=True).copy()

# explorer the dataset
print(CA_cities.shape[:])
CA_cities.head()
```

```
(605, 5)
```

Out[120]:

	city	state	lat	lng	density
0	Kensington	CA	37.9084	-122.2805	1849.8
1	Pleasant Hill	CA	37.9540	-122.0759	1909.0
2	Pomona	CA	34.0585	-117.7625	2572.0
3	Oakdale	CA	37.7616	-120.8470	1446.0
4	Knights Landing	CA	38.7979	-121.7176	1199.0

Secondly, there will be at most 100 venues obtained from each city through the Foursquare API.

In [134]:
```
# explorer the dataset
print(cities_venues.shape[:])
cities_venues.head()
```

```
(52166, 7)
```

Out[134]:

	city	city Latitude	city Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Kensington	37.9084	-122.2805	The Little Farm	37.909633	-122.264792	Farm
1	Kensington	37.9084	-122.2805	Blake Garden	37.912217	-122.281758	Garden
2	Kensington	37.9084	-122.2805	Indian Rock Park	37.892207	-122.273088	Park
3	Kensington	37.9084	-122.2805	Zachary's Chicago Pizza	37.891453	-122.278608	Pizza Place
4	Kensington	37.9084	-122.2805	Rivoli	37.891095	-122.286327	New American Restaurant

Thirdly, this dataset will be sorted into a new dataset that has city name, 1st most common venue type, 2nd most common venue type, until 10th most common venue type.

In [146]:
```
num_top_venues = 10

indicators = ['st', 'nd', 'rd']

# create columns according to number of top venues
columns = ['city']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

# create a new dataframe
cities_venues_sorted = pd.DataFrame(columns=columns)
cities_venues_sorted['city'] = CA_grouped['city']

for ind in np.arange(CA_grouped.shape[0]):
    cities_venues_sorted.iloc[ind, 1:] = return_most_common_venues(CA_grouped.iloc[ind, :], num_
top_venues)

# explorer the sorted dataset that will be used for modeling
print(cities_venues_sorted.shape)
cities_venues_sorted.head()
```

```
(605, 11)
```

Out[146]:

	city	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 Co Ve
0	Agoura Hills	Deli / Bodega	Breakfast Spot	Brewery	American Restaurant	Hotel	Gym / Fitness Center	Mexican Restaurant	Fa Re
1	Alameda	Café	Beach	Coffee Shop	Mexican Restaurant	Grocery Store	Deli / Bodega	Sushi Restaurant	Wi
2	Albany	Coffee Shop	Pizza Place	Grocery Store	Trail	Flower Shop	Brewery	Mexican Restaurant	Ba
3	Alhambra	Chinese Restaurant	Mexican Restaurant	Convenience Store	Burger Joint	Italian Restaurant	Bakery	Dessert Shop	Pa
4	Aliso Viejo	Park	Pizza Place	Grocery Store	Burger Joint	Bakery	Sushi Restaurant	Breakfast Spot	Me Re
5	Alondra Park	Japanese Restaurant	Burger Joint	Convenience Store	Cosmetics Shop	Noodle House	Coffee Shop	Mediterranean Restaurant	Vie Re
6	Alpaugh	Fast Food Restaurant	Sandwich Place	Mexican Restaurant	Discount Store	Pizza Place	Convenience Store	Pharmacy	De

The sorted data set is the final dataset that will be used in city clustering.

3. Exploratory Data Analysis

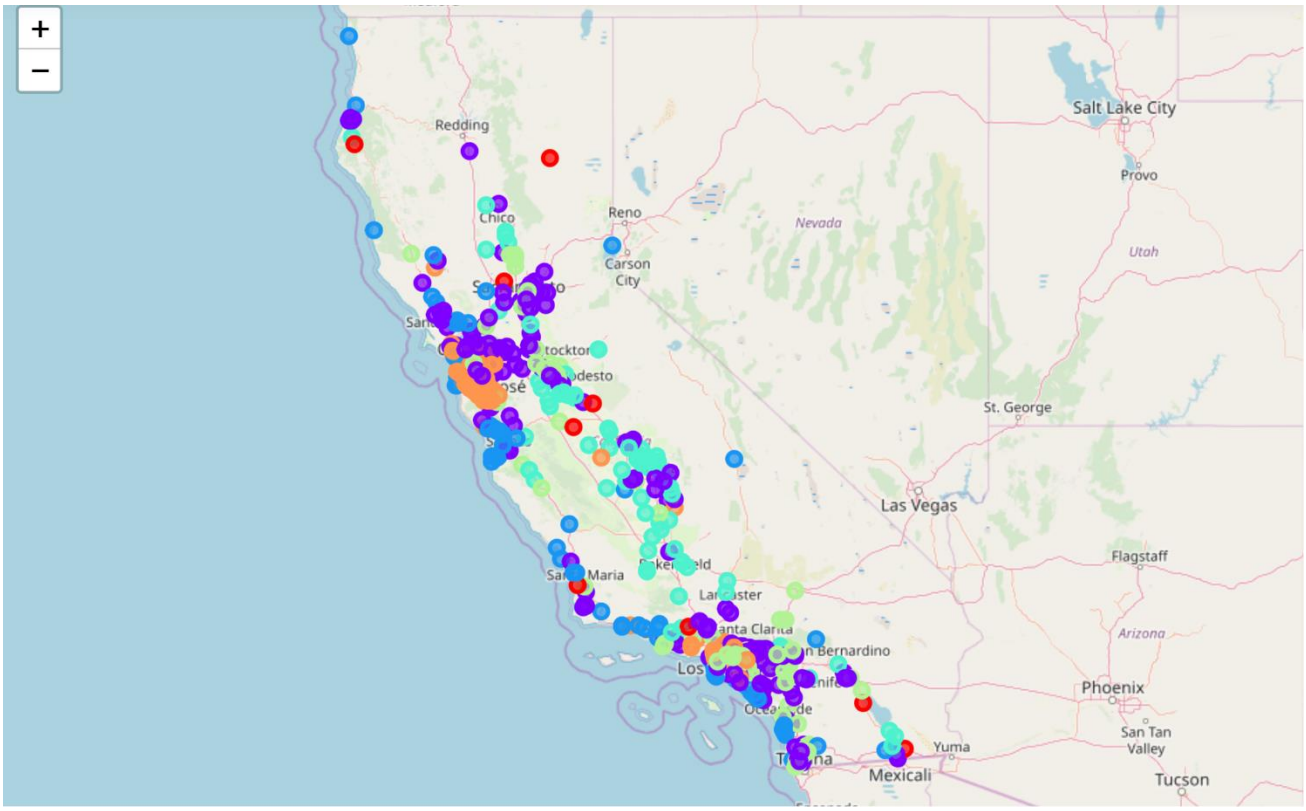
There are total 605 cities which have the 1000.0 or higher population density. By using the Foursquare API, 52166 venues were obtained from 605 cities. Most of the cities have the maximum 100 venues which are returned by the API, and some of them have less than 100 venues. The higher population density a city has, the more venues will be returned from the Foursquare API. There are total 488 unique categories.

4. Clustering

4.1 K-Mean Clustering

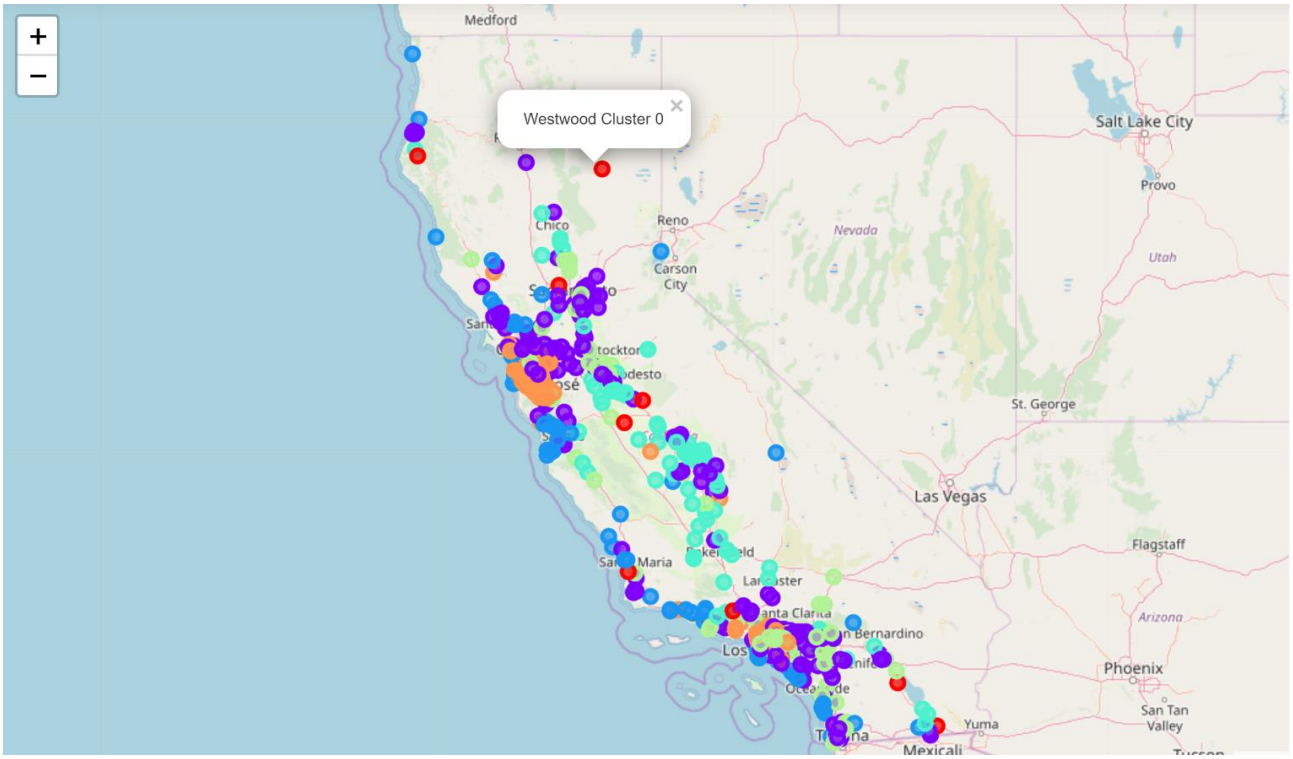
There are total 52166 venues and 488 categories, a new dataset is created to show how many different categories are in each city. From this dataset, the mean of categories for each city is also calculated. The final dataset which will be used in the clustering is created.

Since k-mean clustering is used for city segmentation, finding the best k for the model must be performed in order to get the best result. By using elbow method, the best k of k-mean clustering is 6. The clustering label will be added to the dataset, and a final dataset that will be used to create the map is created. This final dataset has city name, latitude, longitude, clustering label, and the top 10 common venue categories.

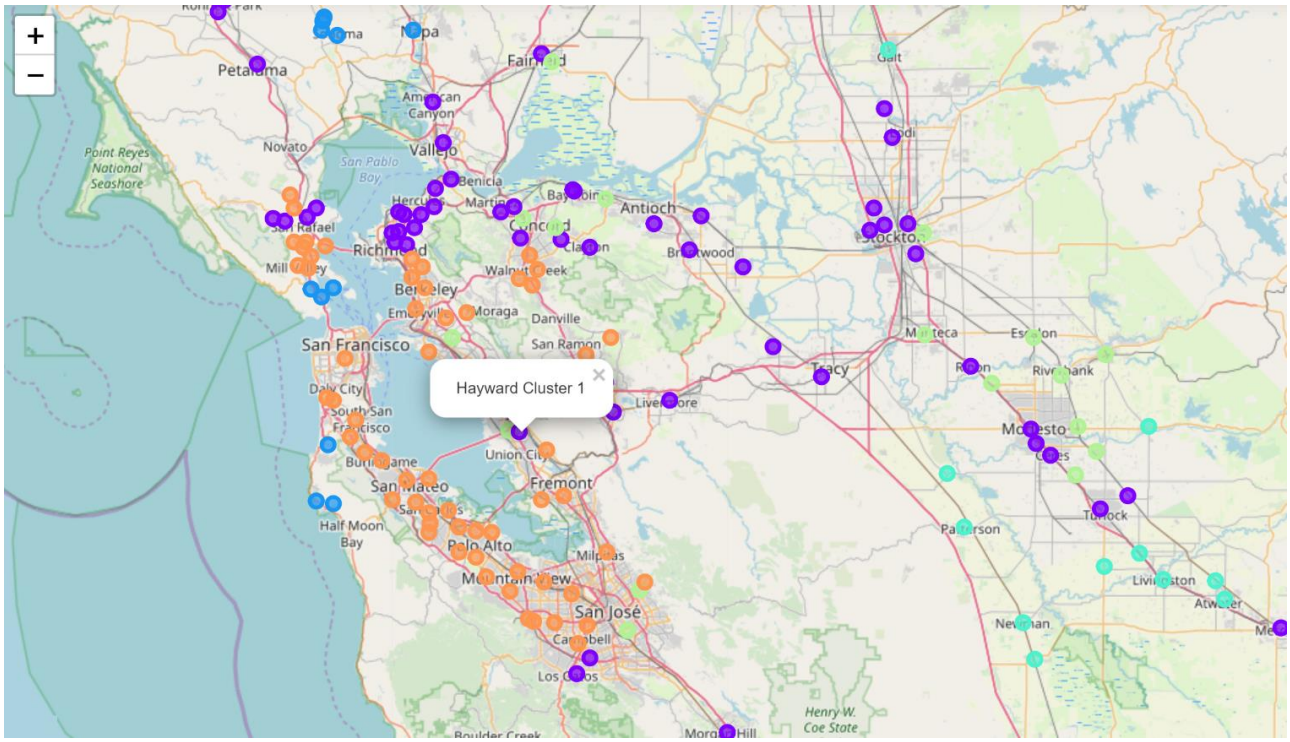


The above map is created based on the final dataset, and it shows 6 different colors that represent 6 different groups.

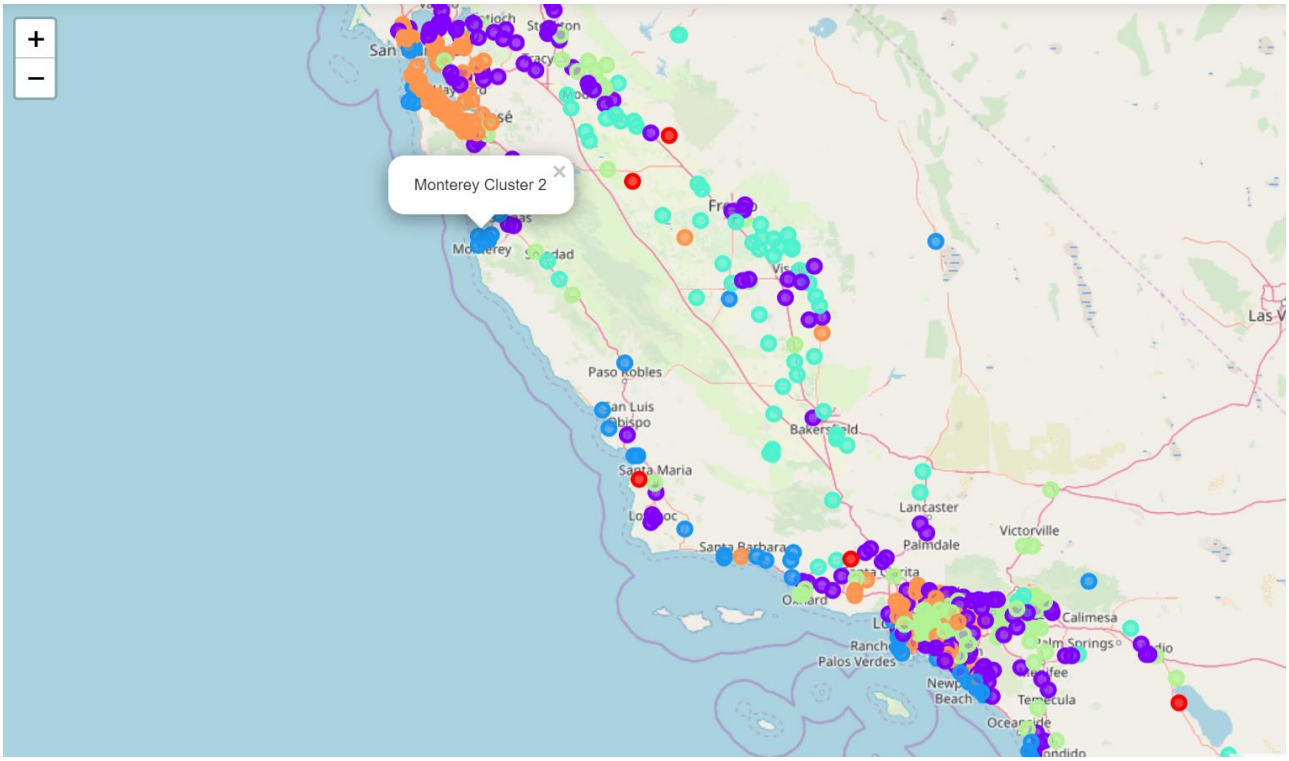
4.2 Clustering Results



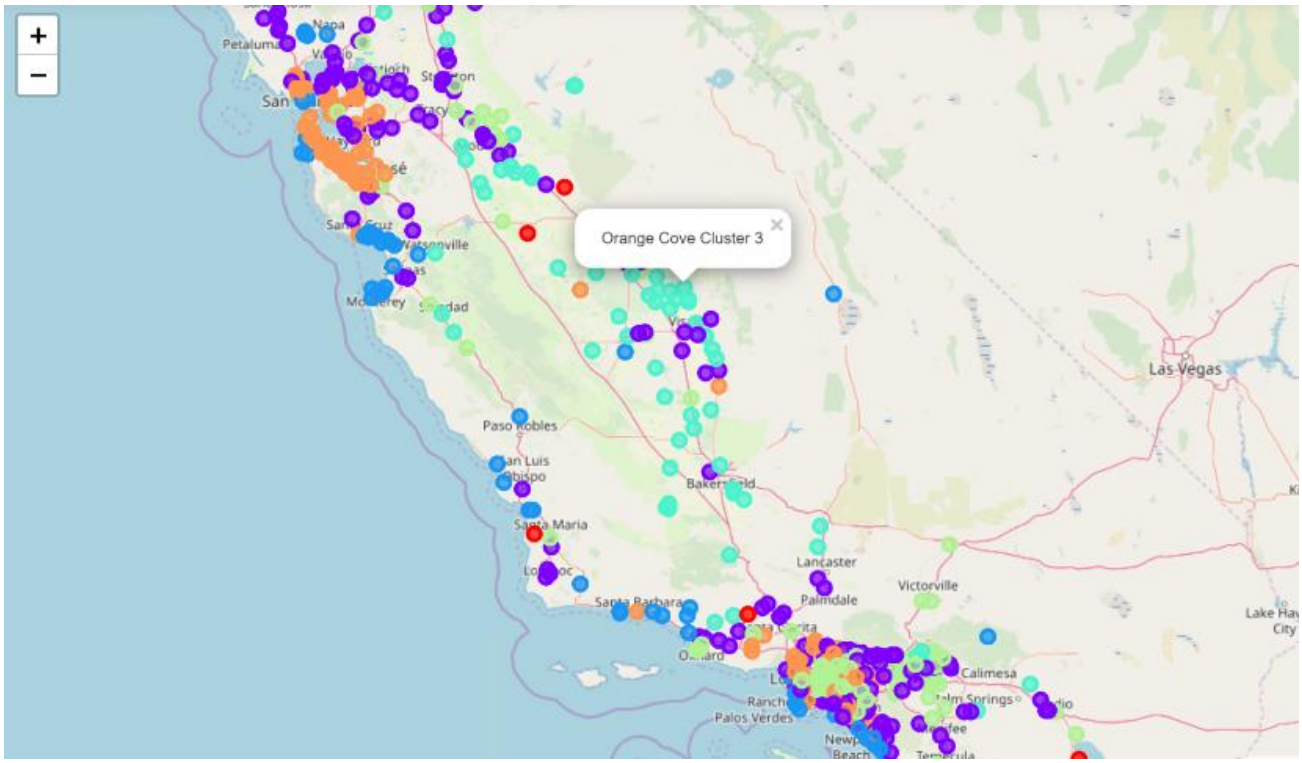
The cluster 0 represents some highly populated rural cities, and they are a very special group since they are also different with most of the rural citi



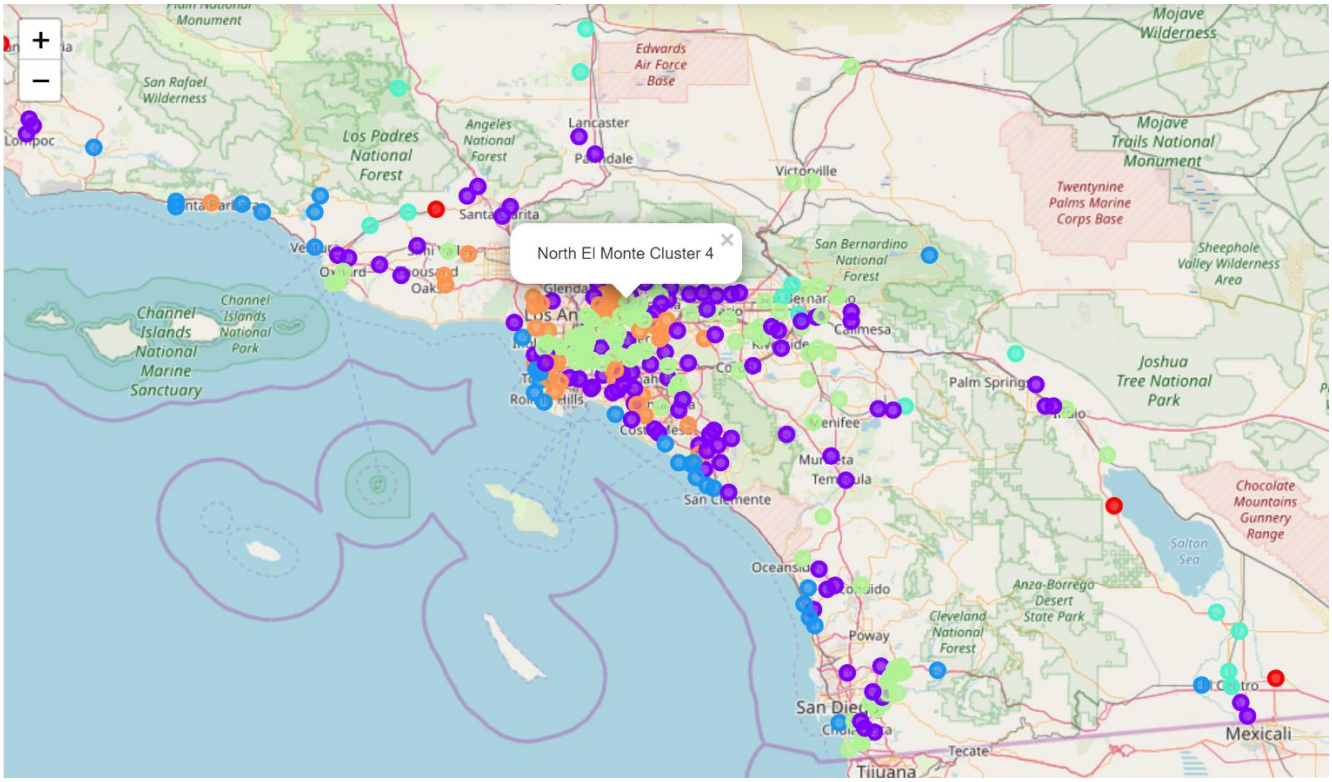
The cluster 1 represents the cities that are less developed and mostly locate at the edge of the highly developed area.



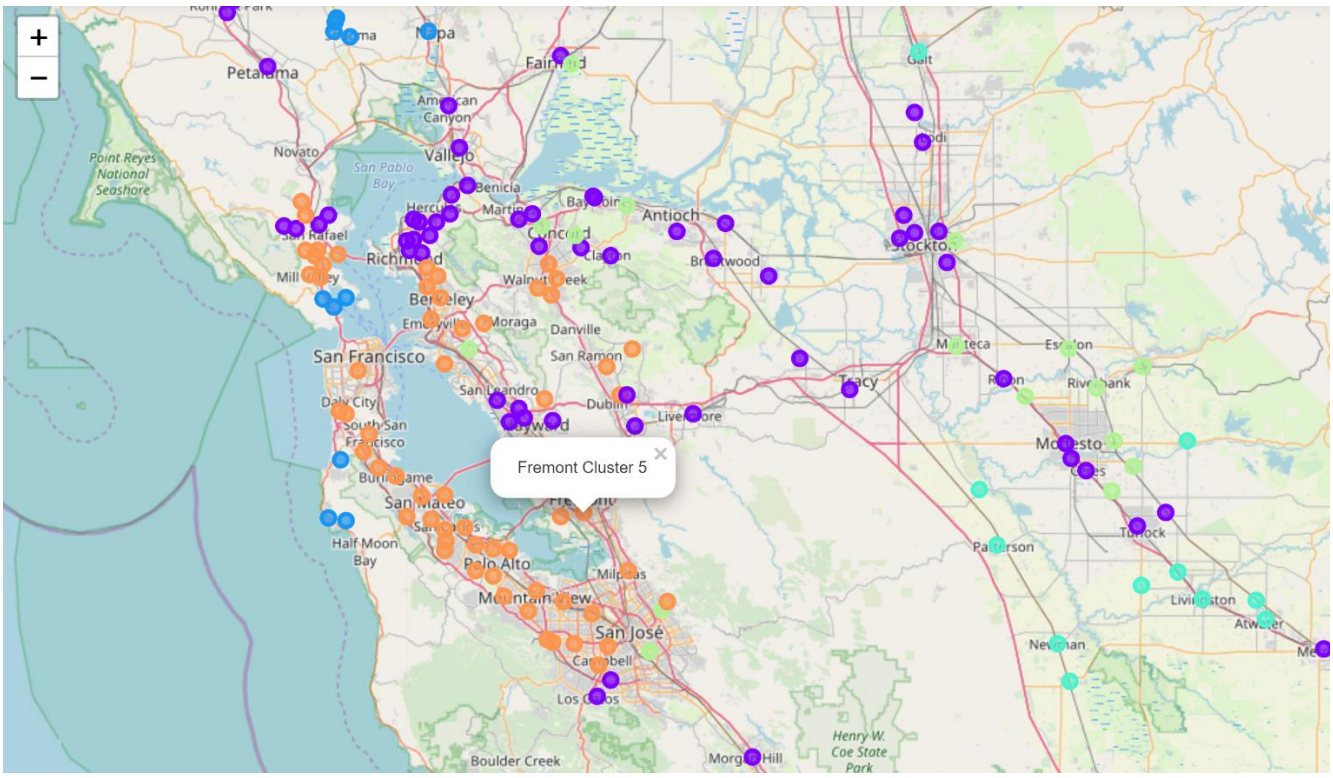
The cluster 2 represents most of the recreation cities that mostly locate near the sea.



The cluster 3 represents the typical less developed rural cities around the California.



The cluster 4 represents most of the less developed Los Angeles cities.



The Cluster 5 represents the typical highly developed cities in San Francisco Bay Area and Los Angeles.

5. Conclusion

In this project, I acquired California city data from Simplemaps and extracted the cities with 1000.00 or higher population density. And then, I retrieved venues of each city from FourSquare API and integrated them with city data. By using the integrated dataset, I performed k-means machine learning technique and obtained 6 clusters of cities. The 6 different groups of cities basically represent most of the types of California cities. Real Estate company could use the results to apply strategy to the similar cities.