Capstone Project - The Battle of Neighborhoods

Yajie Wang

June 2021

1. Introduction

Irvine is the largest city in area in Orange County with 65.61 sq. miles, and 3rd largest in population with 212,375 people according to census in 2010.

Several corporations, particularly in the technology and semiconductor sectors, have their national or international headquarters in Irvine. A lot of them also have office in Bay Area (this analysis will focus on Santa Clara county). During the last couple of years, I had quite a few friends who relocated to Santa Clara county. One big task in the relocation is to find housing, and neighborhood is an important consideration.

Then I was wondering if I can take some machine learning approach and solve this question. Thus, the purpose of this project is to conduct clustering analysis on the 15 cities in Santa Clara county and find out which one is the most like Irvine.

2. Data acquisition and cleaning

2.1 Data sources

Data used in this analysis mainly came from two sources.

1) Population and demographic related data from Wikipedia pages extracted using web scaping approach:

https://en.wikipedia.org/wiki/Santa_Clara_County,_California

https://en.wikipedia.org/wiki/Irvine,_California

2) Venue information from foursquare API.

2.2 Data manipulation

In each Wikipedia page, there is a table called “Population (2010 Census)” that contains list of city names and population data gathered from 2010 census. I used web-scrapping tool BeautifulSoup in python to extract the information.

With the list of city names, I then used Nominatim to get latitude and longitude information.

According to the latitude and longitude gathered for the cities, I used foursquare API to pull the surrounding venues with LIMIT = 400 and radius=50000.

3. Exploratory Data Analysis

Before running the machine learning model, I want to take a brief look at the data.

There are altogether 12 cities selected in Santa Clara country. Including Irvine, there are 13 cities included in my analysis.

Using the parameters above, there are altogether 158 unique categories or kinds of venues.

Below shows top 5 venues for each city:

```
----Campbell----
                venue  freq
0  Italian Restaurant  0.07
1  Mexican Restaurant  0.07
2      Sandwich Place  0.05
3         Pizza Place  0.05
4    Recording Studio  0.02


----Cupertino----
                    venue  freq
0       Mobile Phone Shop  0.09
1             Coffee Shop  0.09
2                    Bank  0.07
3      Chinese Restaurant  0.07
4  Furniture / Home Store  0.05


----Gilroy----
                  venue  freq
0  American Restaurant  0.67
1       Shipping Store  0.33
2                  ATM  0.00
3        Movie Theater  0.00
4   Mexican Restaurant  0.00


----Irvine----
                   venue  freq
0               Fountain  0.25
1                   Park  0.25
2                  Trail  0.25
3      Baseball Stadium  0.25
4  Mongolian Restaurant  0.00
```

```
----Los Altos----
                venue  freq
0         Coffee Shop  0.06
1         Pizza Place  0.06
2  Italian Restaurant  0.06
3  Mexican Restaurant  0.04
4              Bakery  0.04


----Milpitas----
                    venue  freq
0      Indian Restaurant  0.09
1  Vietnamese Restaurant  0.07
2         Sandwich Place  0.07
3   Fast Food Restaurant  0.05
4      Korean Restaurant  0.05


----Monte Sereno----
                       venue  freq
0               Home Service   1.0
1              Movie Theater   0.0
2                Men's Store   0.0
3         Mexican Restaurant   0.0
4  Middle Eastern Restaurant   0.0


----Morgan Hill----
                    venue  freq
0     Italian Restaurant  0.09
1                Brewery  0.07
2  Vietnamese Restaurant  0.05
3     Mexican Restaurant  0.05
4           Burger Joint  0.05


----Mountain View----
                      venue  freq
0               Coffee Shop  0.09
1                      Park  0.07
2                    Bakery  0.05
3         Indian Restaurant  0.05
4  Mediterranean Restaurant  0.03


----Palo Alto----
            venue  freq
0            Café  0.06
1  Ice Cream Shop  0.06
2     Coffee Shop  0.04
3           Hotel  0.04
4             Spa  0.03
```

```
----San Jose----
                venue  freq
0  Mexican Restaurant  0.08
1      Sandwich Place  0.07
2         Cocktail Bar  0.06
3                 Pub  0.04
4                 Bar  0.03


----Sunnyvale----
                venue  freq
0         Coffee Shop  0.09
1       Grocery Store  0.07
2  Chinese Restaurant  0.07
3  Mexican Restaurant  0.05
4                Bank  0.05
```

4. Modeling Building

K-means clustering is adopted to complete the analysis.

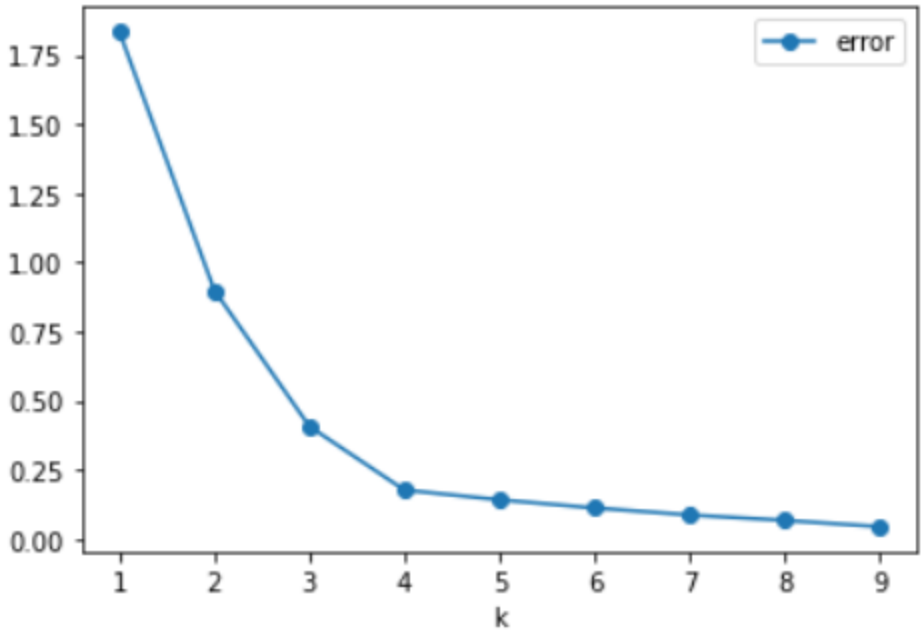
I pivoted the venue categories and summarized frequency of each kind of venue in the cities and got the table below.

	Neighborhood	ATM	Accessories Store	American Restaurant	Andhra Restaurant	Art Gallery	Art Museum	Asian Restaurant	Automotive Shop	BBQ Joint	...	Thrift / Vintage Store	Toy / Game Store	Trail	Train Station	Used Bookstore	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Wine Bar	W
0	Campbell	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.023256	0.000000	0.023256	0.00	0.000000	0.000000	
1	Cupertino	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.022727	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.022727	0.022727	
2	Gilroy	0.000000	0.000000	0.666667	0.000000	0.00	0.00	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	
3	Irvine	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.250000	0.000000	0.000000	0.00	0.000000	0.000000	
4	Los Altos	0.014493	0.000000	0.043478	0.000000	0.00	0.00	0.000000	0.000000	0.000000	...	0.014493	0.014493	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	
5	Milpitas	0.000000	0.000000	0.000000	0.023256	0.00	0.00	0.023256	0.023256	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.069767	0.000000	
6	Monte Sereno	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	
7	Morgan Hill	0.000000	0.000000	0.045455	0.000000	0.00	0.00	0.022727	0.000000	0.022727	...	0.022727	0.000000	0.000000	0.022727	0.000000	0.00	0.045455	0.000000	
8	Mountain View	0.000000	0.000000	0.034483	0.000000	0.00	0.00	0.000000	0.017241	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.017241	0.017241	
9	Palo Alto	0.000000	0.014706	0.014706	0.000000	0.00	0.00	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.029412	
10	San Jose	0.000000	0.000000	0.010000	0.000000	0.01	0.01	0.010000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.01	0.010000	0.000000	
11	Sunnyvale	0.000000	0.000000	0.023256	0.000000	0.00	0.00	0.023256	0.000000	0.000000	...	0.023256	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	

In order to decide what is the optimal number of clusters needed. I tried different numbers of clusters from 1 to 9 and looked at the trend of decreasing distance.

	k	error
0	1	1.834187
1	2	0.898774
2	3	0.406493
3	4	0.175955
4	5	0.140734
5	6	0.111050
6	7	0.086000
7	8	0.066129
8	9	0.043622

The graph will give a better view of trend.

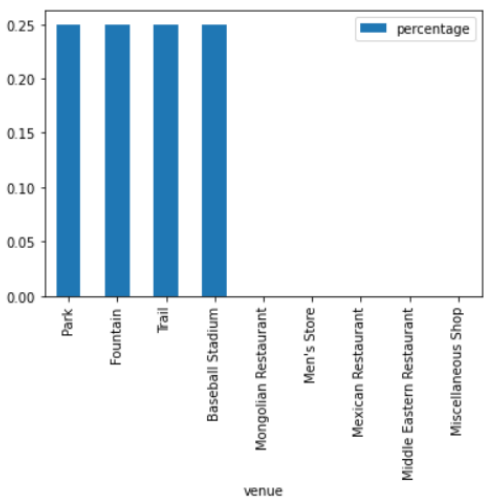


We can see that the curve flattened significantly after 4. Thus, 4 is the optimal number of clusters we want to have here.

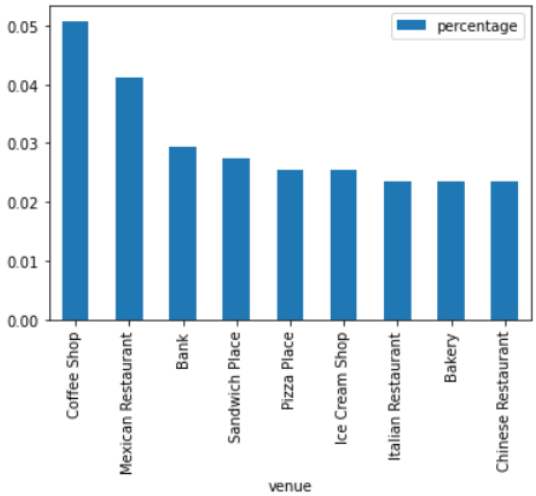
After running k-means clustering with 4 clusters, the cities are separated into 4 different groups. Most Santa Clara cities fall into group 1, and there are 3 cities each formed their own group.

Unfortunately, Irvine is one of them.

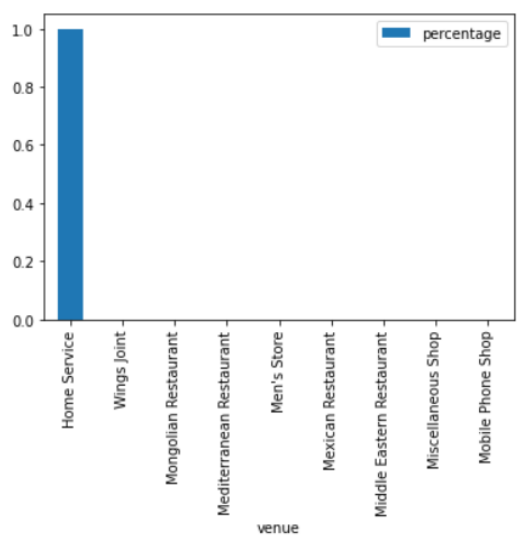
I looked at what are the most popular venues in each of the clusters.



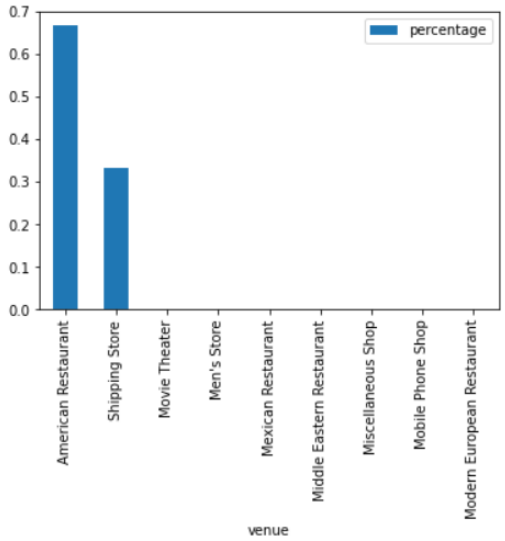
Cluster 0



Cluster 1



Cluster 2



Cluster 3

Irvine is cluster 0, and most Santa Clara cities fall into cluster 1. It appears that Irvine focuses more on family activities while most Santa Clara cities focus more on food and drink shops.

5. Conclusion

Unfortunately, Irvine is one of the 3 cities that formed their own clusters, which means there is no city in Santa Clara county that is like it according to the analysis so far.

However, the data density is an issue, as there are not a lot of venue tracked by Foursquare API in Irvine.

Further studies will be helpful using different APIs and extract more venue information.