The San Francisco Bay Area City Segmentation

1. Introduction

1.1 Project background

The San Francisco bay area is one the most populous and diverse areas in the United States. It has many attractions. It is the home of a large number of businesses. It is a tourist hub. It attracts many people to settle here.

The San Francisco bay area physically is a big group of cities. It consists of about 100 cities [1], small or big. As a person living, I don't even know the names of some of these cities, not to mention their location, attraction and styles. It was a good exercise for me to know the bay area better. For people who are new to this area, this work could be a brief guidance. By knowing the cities, their venues and styles, tourists can choose the places they are interested in visiting, new settlers can select the towns they like as their initial residence.

1.2 Data description

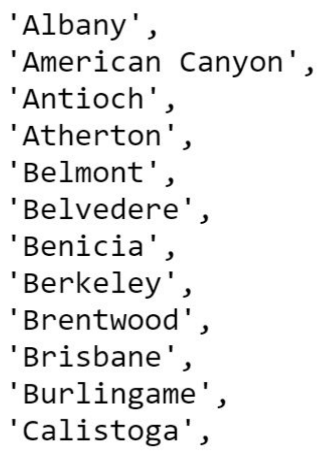
- Bay area city list was found from a table of a wikipedia web page [1]. The data of this web page was processed and a city list can be obtained, even the corresponding county each city belongs to can be acquired.
- Foursquare API was used to get the venues and venue categories around each city center.

2. Methodology

2.1 Web page scraping

Python beautifulsoup library was used to translate the webpage table to a city list.

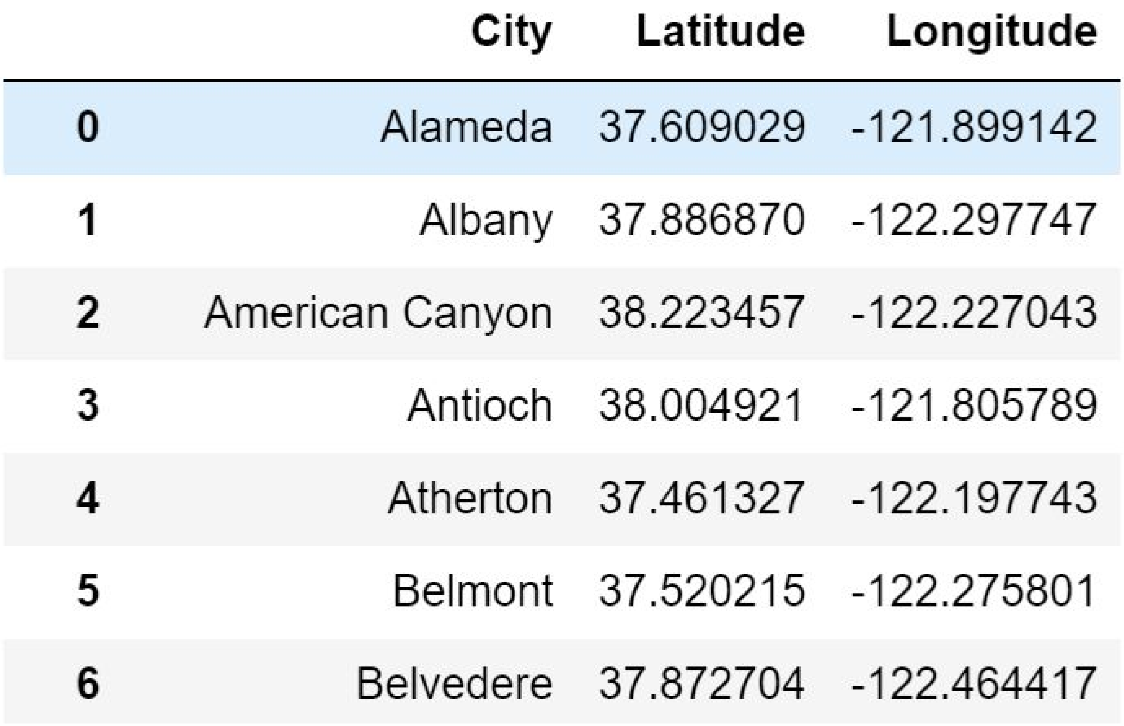
Add alt text



```
'Albany',
'American Canyon',
'Antioch',
'Atherton',
'Belmont',
'Belvedere',
'Benicia',
'Berkeley',
'Brentwood',
'Brisbane',
'Burlingame',
'Calistoga',
```

2.2 Acquire latitude/longitude for each city

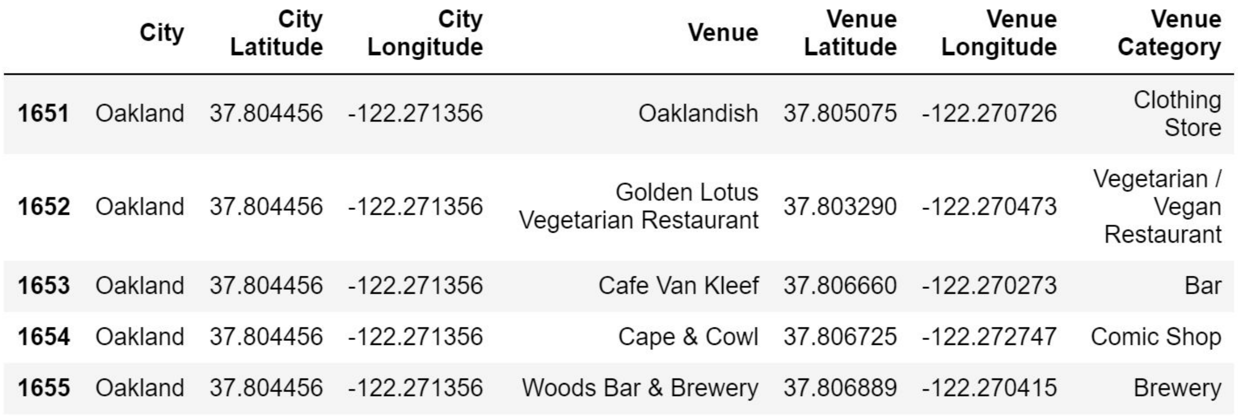
Python geopy package was used to acquire the latitude and longitude for each city. The name of each city in the list was fed in for its latitude and longitude. A dataframe was created to store these three features: City, Latitude and Longitude.



	City	Latitude	Longitude
0	Alameda	37.609029	-121.899142
1	Albany	37.886870	-122.297747
2	American Canyon	38.223457	-122.227043
3	Antioch	38.004921	-121.805789
4	Atherton	37.461327	-122.197743
5	Belmont	37.520215	-122.275801
6	Belvedere	37.872704	-122.464417

2.3 Acquire venues for each city

Foursquare API was utilized to acquire venue data for each city, by putting in latitude and longitude. Afterwards, a dataframe was built including city name, its latitude and longitude, venues name and its category type, like below:

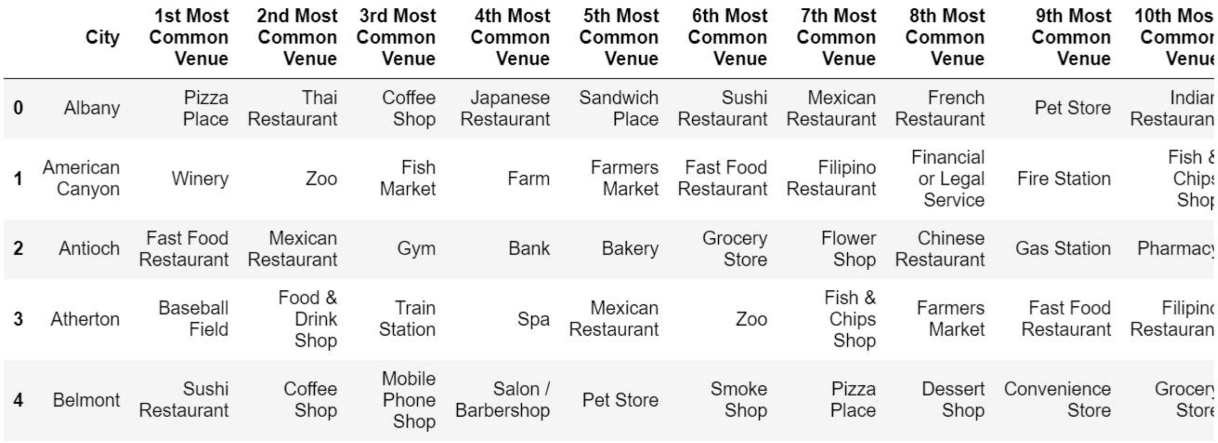


	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1651	Oakland	37.804456	-122.271356	Oaklandish	37.805075	-122.270726	Clothing Store
1652	Oakland	37.804456	-122.271356	Golden Lotus Vegetarian Restaurant	37.803290	-122.270473	Vegetarian / Vegan Restaurant
1653	Oakland	37.804456	-122.271356	Cafe Van Kleef	37.806660	-122.270273	Bar
1654	Oakland	37.804456	-122.271356	Cape & Cowl	37.806725	-122.272747	Comic Shop
1655	Oakland	37.804456	-122.271356	Woods Bar & Brewery	37.806889	-122.270415	Brewery

Some parameters were needed to instruct the Foursquare API. Here, 500 was used for the parameter of radius, meaning venues were searched within 500 meters range around the city center. 100 was used for the parameter of LIMIT. This controls the maximum number of returned venues for each search. In this study, it was noted that the number of venues is less than 100, for some cities like Belmont.

2.4 Select the most popular venue categories

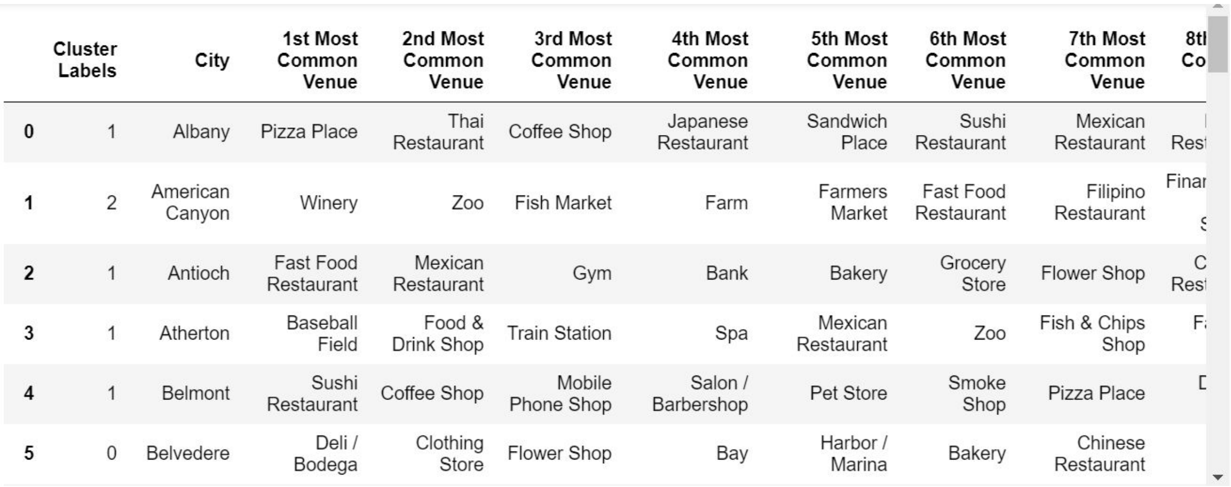
A total 315 venue categories were acquired when combining all the venue returns for all cities. It might be sensible to select the most popular venues for each city, especially for demonstration purposes.



	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 Most Common Venue	9th Most Common Venue	10th Mos Commor Venue
0	Albany	Pizza Place	Thai Restaurant	Coffee Shop	Japanese Restaurant	Sandwich Place	Sushi Restaurant	Mexican Restaurant	French Restaurant	Pet Store	Indiar Restauran
1	American Canyon	Winery	Zoo	Fish Market	Farm	Farmers Market	Fast Food Restaurant	Filipino Restaurant	Financial or Legal Service	Fire Station	Fish & Chips Shop
2	Antioch	Fast Food Restaurant	Mexican Restaurant	Gym	Bank	Bakery	Grocery Store	Flower Shop	Chinese Restaurant	Gas Station	Pharmacy
3	Atherton	Baseball Field	Food & Drink Shop	Train Station	Spa	Mexican Restaurant	Zoo	Fish & Chips Shop	Farmers Market	Fast Food Restaurant	Filipinc Restauran
4	Belmont	Sushi Restaurant	Coffee Shop	Mobile Phone Shop	Salon / Barbershop	Pet Store	Smoke Shop	Pizza Place	Dessert Shop	Convenience Store	Grocery Store

2.5 Cluster cities

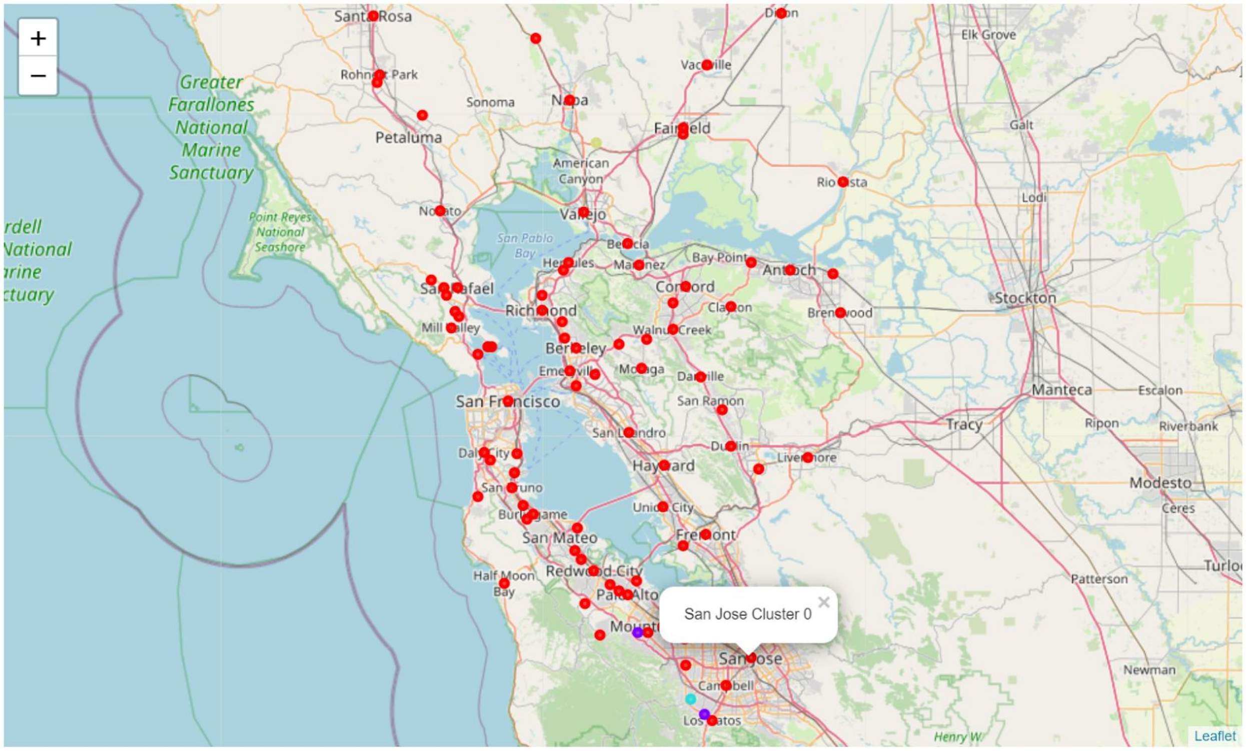
Here the cities were clustered into 4 groups. Kmeans algorithm was used.



	Cluster Labels	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	8tl Co
0	1	Albany	Pizza Place	Thai Restaurant	Coffee Shop	Japanese Restaurant	Sandwich Place	Sushi Restaurant	Mexican Restaurant	I Resi
1	2	American Canyon	Winery	Zoo	Fish Market	Farm	Farmers Market	Fast Food Restaurant	Filipino Restaurant	Finar S
2	1	Antioch	Fast Food Restaurant	Mexican Restaurant	Gym	Bank	Bakery	Grocery Store	Flower Shop	C Resi
3	1	Atherton	Baseball Field	Food & Drink Shop	Train Station	Spa	Mexican Restaurant	Zoo	Fish & Chips Shop	F
4	1	Belmont	Sushi Restaurant	Coffee Shop	Mobile Phone Shop	Salon / Barbershop	Pet Store	Smoke Shop	Pizza Place	D
5	0	Belvedere	Deli / Bodega	Clothing Store	Flower Shop	Bay	Harbor / Marina	Bakery	Chinese Restaurant	

3. Results

The Python folium package was used to show clustered cities in the map. Each cluster was labeled in a color.



- Cluster 0: populous city at least populous city center, with variety of venues, like restaurant, grocery store, and park, Mexican and Asian food is populous;
- Cluster 1: less populous city center, with less restaurants;
- Cluster 2: similar with Cluster 1, working out venues standing out;
- Cluster 3: the least populous city type, with winery

The results demonstrated that most of bay area cities bear the similar city style. The cities have a certain number of different types of restaurant, although Mexican and Asian food is pretty popular. The high population density could be one of the reasons, and many of bay area cities are very attractive to tourists. Bay area is the home of the wine country, winery is one of the attractions here. The results also shows to some extent the healthy life style of the bay area people, because in some cities the working out venues stand out.

4. Discussion

In this work, the city venues and their categories were used. Apparently very different features can be studied and possibly yield some high quality and interesting results.
From the map above, it was noticed that most of the bay area cities fall into one cluster. were clustered into one category. Kmeans algorithm was utilized to carry out clustering, and 4 was set as the k values. It is worth trying out other K values in the future to see if it will change the segmentation. Also the radius and LIMIT values of API affect the data to train the algorithm. They are also worthy of further investigation.

5. Conclusion

The San Francisco bay area is a populous and diverse area. By using the venue information of each city provides some information to people who are new to this area. They can get some hints on this area when they either want to visit for leisure or move into the San Francisco bay area.