New Business Location Analysis in San Francisco

Joye

December 25, 2019

1. Introduction

1.1 Background

Members of Generation Z have gradually become the main force of the consumption market. They grew up in the era of technology and information explosion, and also witnessed the development of internet giants, like Google, Amazon, Facebook, etc. Obviously, their consumption preferences are different from former generations. They love cool things and love to share that in the internet. Therefore, for those companies which provide cool products and services, having the Generation Z consumers is of great significance to achieve their long-term development and increase market shares.

However, to win the Generation Z consumers is not an easy task. According to consumption characteristics, different companies should improve their products, services, marketing etc. While among all of those, location is a top priority, because a bad location will always discount consumption experiences and satisfaction. This is especially vital for a new opening business.

1.2 Audience

My target audience are those companies who want to win more market shares and have the plan to open several new businesses in cities. Since in this capstone project, we will choose San Francisco as our target city, so my audience are stakeholders who plan to open new businesses in San Francisco. To help them make a wise decision on location, I will take advantage of the Foursquare location data and provide valuable suggestions to them.

1.3 Why San Francisco

San Francisco is a great city, diverse, dynamic and entrepreneurial. With its economy humming, the city attracts many young people to live and work. After analyzing the location data of San Francisco, we can have a visual sight of the distribution of location, and then recommend some locations to some business.

2. Data

2.1 What data

To successfully finish this capstone project, we need the following kinds of data:

a) The neighborhoods in San Francisco, and their latitude and longitude coordinates.

b) The venue data of those neighborhoods.

2.2 Data sources

At first, we can get the list of neighborhoods from Wikipedia via the link: https://en.wikipedia.org/wiki/List_of_neighborhoods_in_San_Francisco using magic BeautifulSoup library. Then we can use geopy library to get the latitude and longitude values of those neighborhoods. With these data, we can create a map with neighborhoods superimposed on top.

Next, we can apply Foursquare API to get venue data of the above neighborhoods, which are the most important data for our analysis. After analyzing neighborhoods data, we use K-means method to cluster them. Based on this part's work, we present the findings and make the recommendation for some business.

3. Methodology

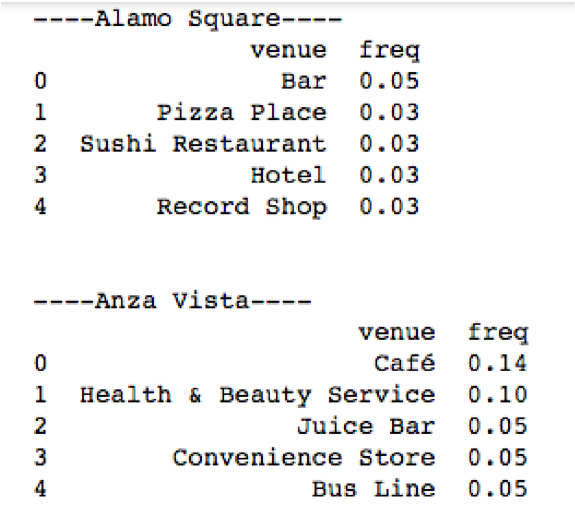
After data preprocessing, we will get the venue data and we can start our data analysis. The first step is to check the size of resulting data, so that we can have a clear sight of the analysis scope and depth. Obviously the more valid data we get, the more chances we can have significant findings. Next, we need to check the venue categories to see if data are diverse enough to cover most kinds of venues, and more importantly, it will tell us how many kinds of dummies we need to create.

Secondly, we convert the categorical data into numerical data, namely convert the venue categories into dummies. Therefore, with these numerical data, we can calculate the frequency of each category in all venues we got. Naturally, we can also get the frequency of each category in our concerned neighborhoods. This is an important part of our data analysis, since we can use the top 10 frequency venues in each neighborhood as our new attributes for clustering. At same time, we can also check top common businesses in each neighborhood and we can compare the differences among different areas. From the general review, we will have a good understanding of the next clustering and data analysis.

Therefore, next we will cluster the neighborhoods, so that we can discover the similarity among them. Here we use the unsupervised machine learning k-means clustering. It is an elegant approach to partition a data set into K distinct clusters via using the frequency data in last step. After clustering, we can visualize the neighborhoods on the map of San Francisco. In the final step, we will carefully review each cluster and its attributes, and use word cloud to visualize the most common business in clusters. From the view of neighborhood and cluster, we can draw our conclusion about the recommendation of new business location according to the balance of industry agglomeration and unique.

4. Results

From Wikipedia, we get a list of 119 neighborhoods, but only 90 of them can find their latitude and longitude coordinates. Viewing on the map, we notice that these neighborhoods are widely displayed in San Francisco, so we reduce the scope and the number decreases to 45. After applying Foursquare API we get 3428 rows of venue data, which contains 305 unique categories. Then we convert the value of "Venue Category" column into dummies 0 and 1, and then we can get its frequency. The following table shows the top 5 business categories and their frequency in Alamo Square and Anza Vista. Notice, the frequency here is global frequency, so it's comparable among all neighborhoods.

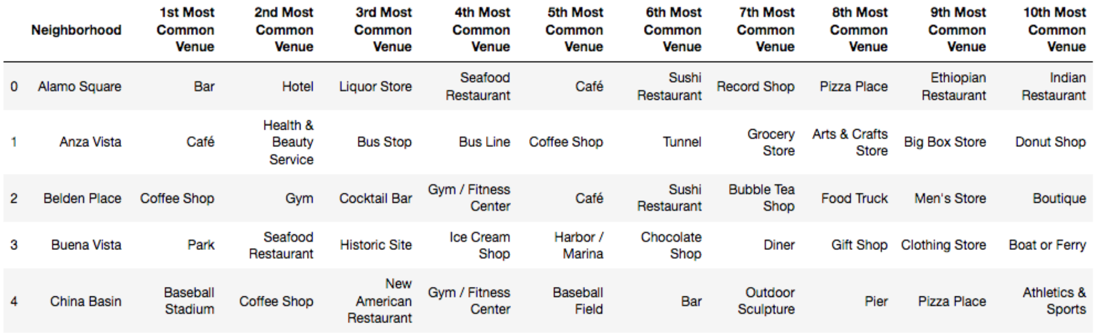


```
----Alamo Square----
              venue  freq
0               Bar  0.05
1       Pizza Place  0.03
2  Sushi Restaurant  0.03
3             Hotel  0.03
4       Record Shop  0.03


----Anza Vista----
                     venue  freq
0                     Café  0.14
1  Health & Beauty Service  0.10
2                Juice Bar  0.05
3        Convenience Store  0.05
4                 Bus Line  0.05
```

Table 1. The information of top business in neighborhoods

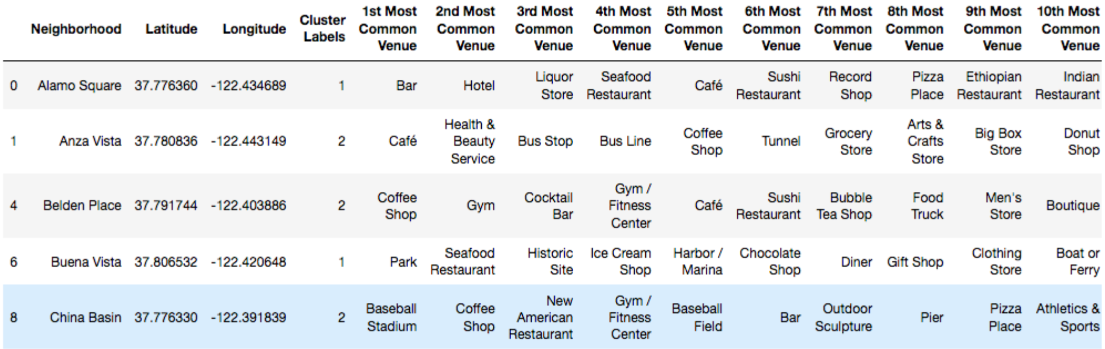
We can find out that except the top 1 or 2 have much better performance in frequency, the next several ones share similar frequency. Therefore, we use top 10 most common venue as 10 new columns of each neighborhood. The following table shows the new table structure of the neighborhood data.



	Neighborhood	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 Most Common Venue
0	Alamo Square	Bar	Hotel	Liquor Store	Seafood Restaurant	Café	Sushi Restaurant	Record Shop	Pizza Place	Ethiopian Restaurant	Indian Restaurant
1	Anza Vista	Café	Health & Beauty Service	Bus Stop	Bus Line	Coffee Shop	Tunnel	Grocery Store	Arts & Crafts Store	Big Box Store	Donut Shop
2	Belden Place	Coffee Shop	Gym	Cocktail Bar	Gym / Fitness Center	Café	Sushi Restaurant	Bubble Tea Shop	Food Truck	Men's Store	Boutique
3	Buena Vista	Park	Seafood Restaurant	Historic Site	Ice Cream Shop	Harbor / Marina	Chocolate Shop	Diner	Gift Shop	Clothing Store	Boat or Ferry
4	China Basin	Baseball Stadium	Coffee Shop	New American Restaurant	Gym / Fitness Center	Baseball Field	Bar	Outdoor Sculpture	Pier	Pizza Place	Athletics & Sports

Table 2. Neighborhoods with 10 new attributes

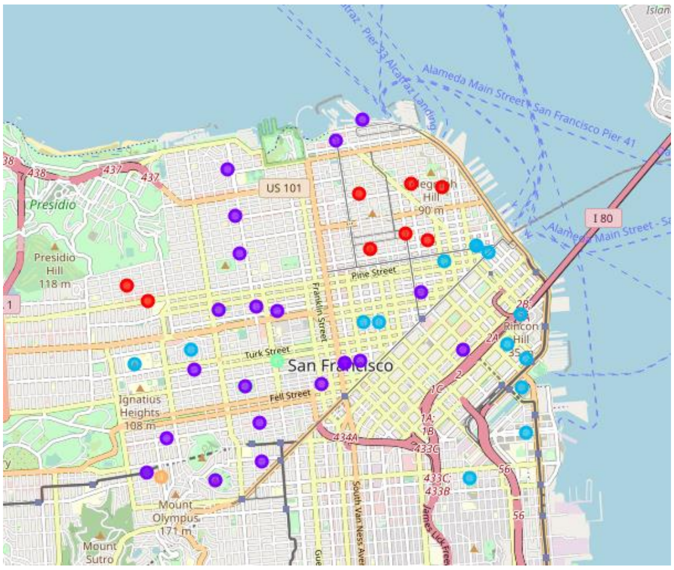
Then using the 10 new attributes as inputs we cluster neighborhoods into 5 groups.



	Neighborhood	Latitude	Longitude	Cluster Labels	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 Most Common Venue
0	Alamo Square	37.776360	-122.434689	1	Bar	Hotel	Liquor Store	Seafood Restaurant	Café	Sushi Restaurant	Record Shop	Pizza Place	Ethiopian Restaurant	Indian Restaurant
1	Anza Vista	37.780836	-122.443149	2	Café	Health & Beauty Service	Bus Stop	Bus Line	Coffee Shop	Tunnel	Grocery Store	Arts & Crafts Store	Big Box Store	Donut Shop
4	Belden Place	37.791744	-122.403886	2	Coffee Shop	Gym	Cocktail Bar	Gym / Fitness Center	Café	Sushi Restaurant	Bubble Tea Shop	Food Truck	Men's Store	Boutique
6	Buena Vista	37.806532	-122.420648	1	Park	Seafood Restaurant	Historic Site	Ice Cream Shop	Harbor / Marina	Chocolate Shop	Diner	Gift Shop	Clothing Store	Boat or Ferry
8	China Basin	37.776330	-122.391839	2	Baseball Stadium	Coffee Shop	New American Restaurant	Gym / Fitness Center	Baseball Field	Bar	Outdoor Sculpture	Pier	Pizza Place	Athletics & Sports

Table 3. Neighborhoods with cluster labels

Figure 1. Visualization of neighborhood with cluster labels



In the cluster examination step, we can see that there are 8 neighborhoods in red(cluster 0), 21 in purple(cluster 1), 14 in blue(cluster 2), 1 in green(cluster 3) and 1 in orange(cluster 4).

In the next, we will analyze the distribution of business in cluster 0, 1, 2 in detail. Cluster 3 and cluster 4 are exclude, since their data are so lack, we can hardly analyze the trend of surrounding and hence hardly provide wise suggestions for stakeholders.

Cluster 0

In this cluster, there are 8 neighborhoods, six of them are very close to each other. Then we treat the top 10 venues from all 8 neighborhoods as a group, and find their distribution as Figure 2.

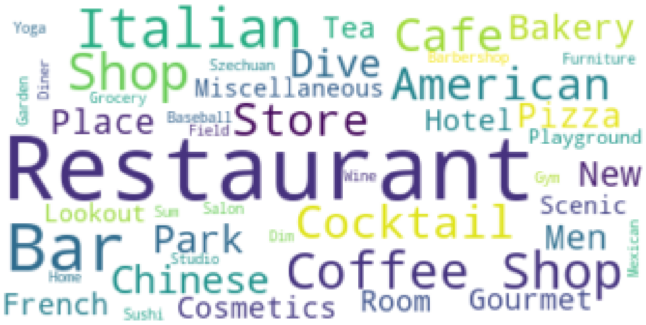


Figure 2. The word cloud of venues in cluster 0

From the analysis of this cluster, we find about 70% venues are about eating or drinking, and in restaurant category, 15 out of 32 are foreign taste. Therefore, from the location data analysis, we recommend innovative catering services to open their new business here.

Cluster 1

This is the cluster has most 21 neighborhoods, and they connect each other to be a quite wide area.

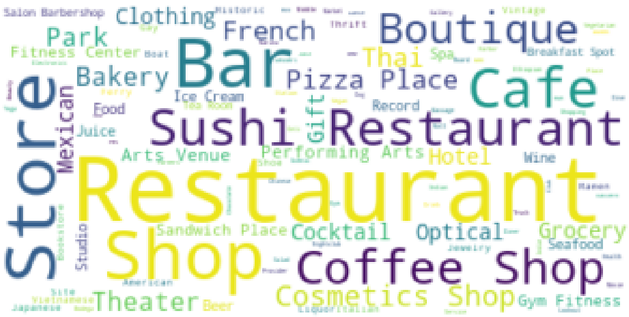


Figure 3. The word cloud of venues in cluster 1

This cluster has many restaurants, coffee shops, and stores. In this area, venues are more diverse, except restaurant and café we also recommend store or shop business in the big area of cluster 1 neighborhood.

Cluster 2

There are 14 neighborhoods in this cluster. From the map we can see that they are not clustered around each other, and they just shape like oblique “V”.

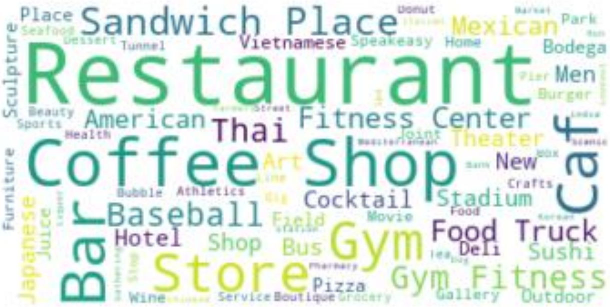


Figure 4. The word cloud of venues in cluster 2

Besides restaurant and coffee shop, it’s noticeable to see gym, fitness and baseball in this area. Therefore, this area will more attractive to those value exercises. Healthy food, organic shop, and sports shop can be suitable here.

5. Discussion

The new business location analysis in this capstone is mainly from the view of venues, we assume that industry agglomeration plays an important role. While, in reality the situation is much more complex, and more factors can affect the operation of new business. In future, we should include more data, like demographic statistics and financial data.

6. Conclusion

Since our neighborhoods are near the downtown of San Francisco, it’s not surprised that restaurants and coffee shops are the most common venue. Therefore, it’s fine to choose cluster 0, 1 and 2 area to start catering business. From the analysis, we notice that cluster 0, 1 and 2 have their own characteristics. Clusters 0 is more to office zone, clusters 1 is more to shopping, entertainment and living, and cluster 2 is more to gym and fitness.

In sum, cluster 0 zone welcomes special restaurant, café and bar business. Cluster 1 zone welcomes catering business and living related shops. Cluster 2 welcomes catering services, healthy industry and sports shops.