Analysis of market competition of San Francisco

Charith Herath

08/05/2020

1. Introduction

1.1 Background

In 2015, the United States Department of Commerce found that for the first time in history, Americans spent more money on dine-out food than on groceries; nowhere was that more evident than in the Bay Area, where new restaurants seem to open weekly. And pizza places are no stranger to this. With this rising economy and competition become restaurants, it has become really hard for new coming restaurants to become successful as big and more established restaurants and fast food chains such as Domino’s and Pizza hut dominates the entire market.

1.2 Problem

It is now clear than ever that the geographical location plays a crucial part in the success of new businesses due to the demand and competition of its location. It would be really helpful for new businesses to be successful if they can identify the best locations which are less competitive but has a decent demand.

1.3 Interest

Both upcoming business owners and more established restaurant owners will be interested in this as it helps them to either start a new pizza place or to expand their current chain.

2. Data acquisition and data cleaning

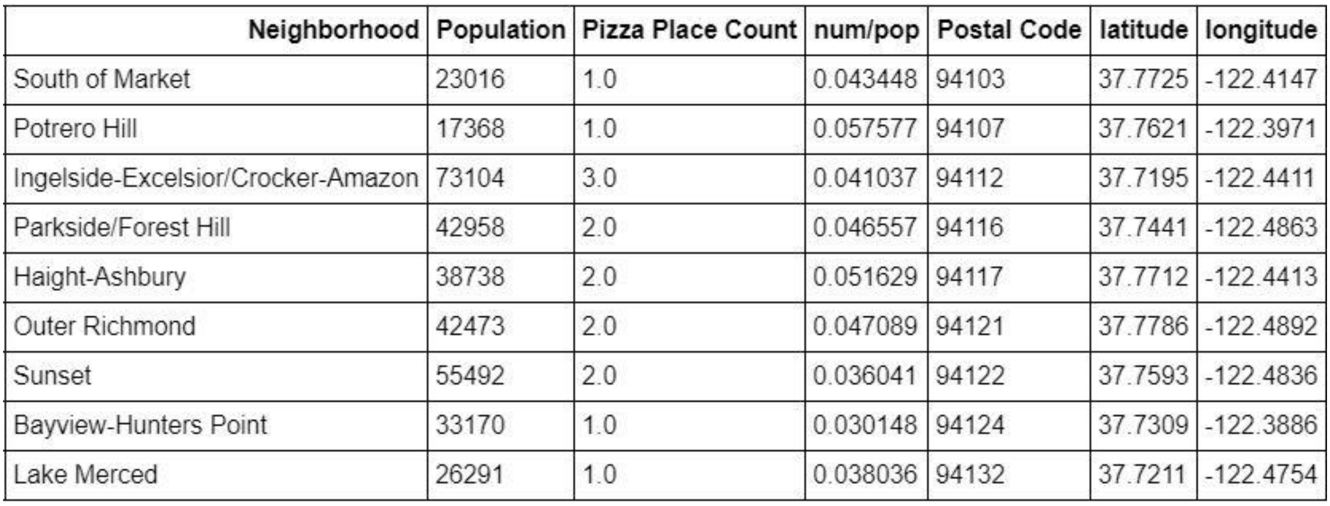
2.1 Data Source

The data related to neighborhoods in San Francisco is obtained from the dataset "San Francisco Neighborhoods as ZIP Codes" from San Francisco Burden of Disease & Injury Study. The dataset contains the zip codes and population of each district in San Francisco. The python library pgeocode is used to obtain latitude and longitude of each district by the zip code. Information related to venues in San Francisco neighborhoods is collected using the Foursquare api.

2.2 Data cleaning and preprocessing

The venues are filtered by their categories and the venues marked as pizza place, bakery and Italian restaurant are filtered and marked as the place that could potentially serve pizzas. And other venues are dropped. All the data obtained from each source is then combined to a single table which contains the neighborhood name, postal code, population, number of pizza places, latitude and longitude of each neighborhood. The data set contained no missing values thus no dropping data needed. For clustering purpose, a new data frame is created only containing the neighborhood name and ratio of population and number of pizza places.

Table 2.1: The Final data frame



Neighborhood	Population	Pizza Place Count	num/pop	Postal Code	latitude	longitude
South of Market	23016	1.0	0.043448	94103	37.7725	-122.4147
Potrero Hill	17368	1.0	0.057577	94107	37.7621	-122.3971
Ingelside-Excelsior/Crocker-Amazon	73104	3.0	0.041037	94112	37.7195	-122.4411
Parkside/Forest Hill	42958	2.0	0.046557	94116	37.7441	-122.4863
Haight-Ashbury	38738	2.0	0.051629	94117	37.7712	-122.4413
Outer Richmond	42473	2.0	0.047089	94121	37.7786	-122.4892
Sunset	55492	2.0	0.036041	94122	37.7593	-122.4836
Bayview-Hunters Point	33170	1.0	0.030148	94124	37.7309	-122.3886
Lake Merced	26291	1.0	0.038036	94132	37.7211	-122.4754

3. Methodology : Clustering

Since, our main objective is to identify neighborhood which has a low competition and a good demand to open a new pizza place. Clustering a is good approach for the problem in our hand. The neighborhoods are divided into several clusters depending on the population and the number of pizza places in each neighborhood using the K-means clustering algorithm. The optimal number of clusters was found using the 'elbow joint method'.

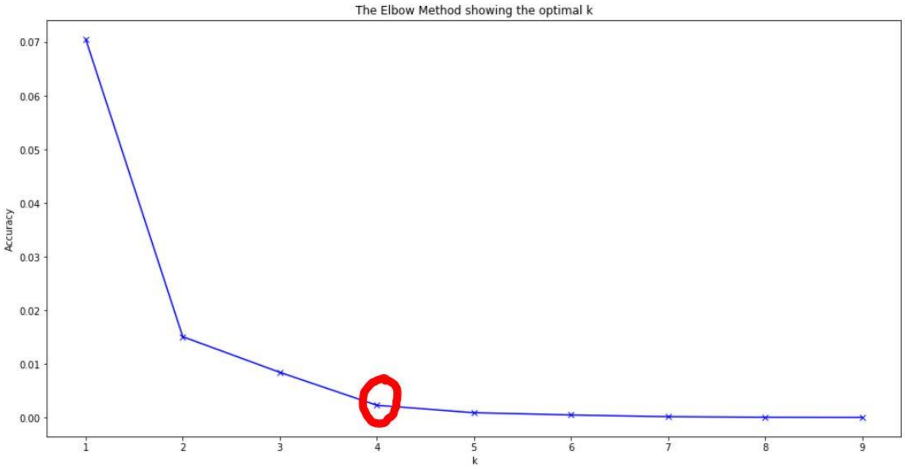


Figure 3.1: Variation of accuracy vs k

From the 'elbow joint method' it was found that the k=4 will give the best result after the testing k values from 2-10. Then, the dataset containing the ratio between the population and number of pizza places is used for clustering using K-means clustering algorithm with k set to 4.

4. Results

The neighborhood were divided into 4 clusters according to the population and number of pizza places. The 4 clusters can be marked as ‘no or very small competition’, ‘small competition’, ‘medium competition’ and ‘high competition’. Clustred neghborhoods are then displayed on the map of San Francisco using Folium.

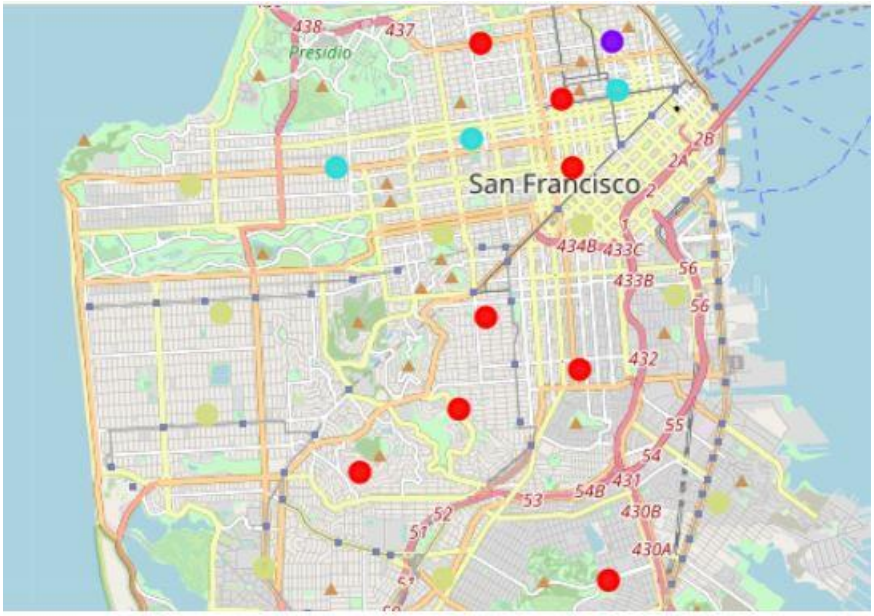


Figure 4.1: Map of San Francisco with clustered neighborhoods

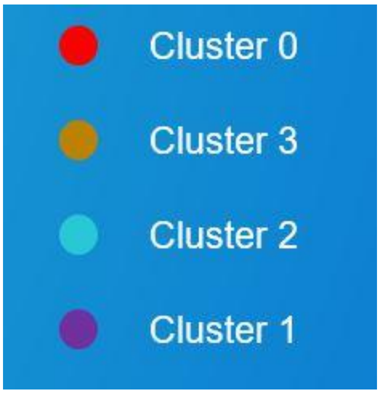
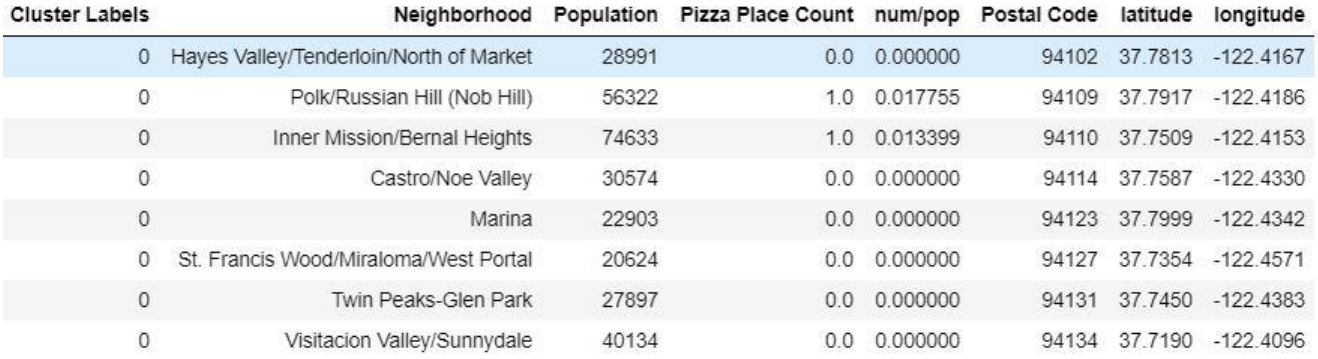


Table 4.1: Definitons of clusters

Cluster label	Definition
0	No or very small competition
1	High competition
2	Medium competition
3	Small competition

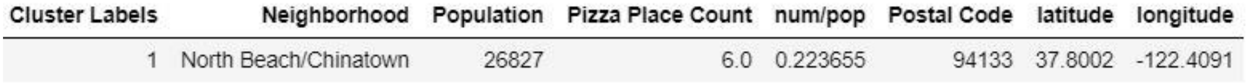
A detailed information of each cluster is given below.

Table 4.2: Cluster 0



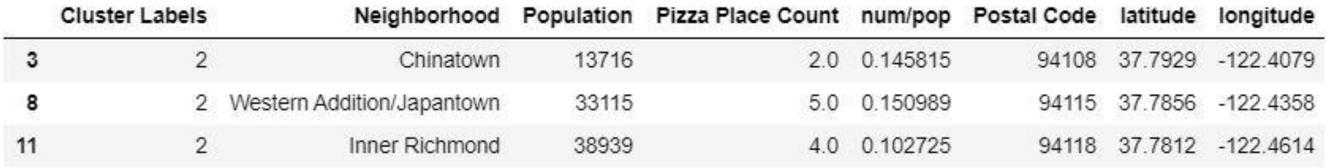
Cluster Labels	Neighborhood	Population	Pizza Place Count	num/pop	Postal Code	latitude	longitude
0	Hayes Valley/Tenderloin/North of Market	28991	0.0	0.000000	94102	37.7813	-122.4167
0	Polk/Russian Hill (Nob Hill)	56322	1.0	0.017755	94109	37.7917	-122.4186
0	Inner Mission/Bernal Heights	74633	1.0	0.013399	94110	37.7509	-122.4153
0	Castro/Noe Valley	30574	0.0	0.000000	94114	37.7587	-122.4330
0	Marina	22903	0.0	0.000000	94123	37.7999	-122.4342
0	St. Francis Wood/Miraloma/West Portal	20624	0.0	0.000000	94127	37.7354	-122.4571
0	Twin Peaks-Glen Park	27897	0.0	0.000000	94131	37.7450	-122.4383
0	Visitacion Valley/Sunnydale	40134	0.0	0.000000	94134	37.7190	-122.4096

Table 4.3: Cluster 1



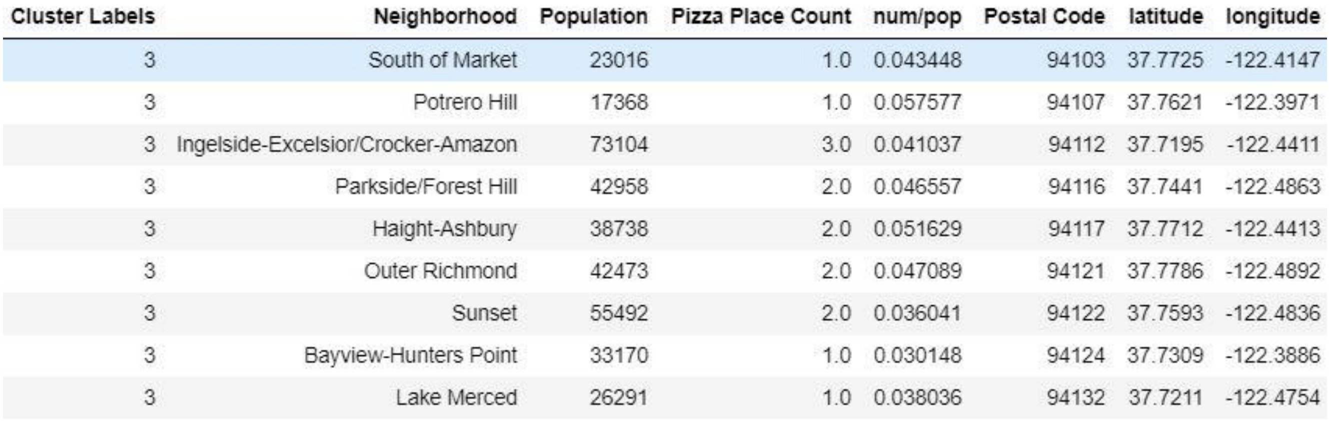
Cluster Labels	Neighborhood	Population	Pizza Place Count	num/pop	Postal Code	latitude	longitude
1	North Beach/Chinatown	26827	6.0	0.223655	94133	37.8002	-122.4091

Table 4.4: Cluster 2



	Cluster Labels	Neighborhood	Population	Pizza Place Count	num/pop	Postal Code	latitude	longitude
3	2	Chinatown	13716	2.0	0.145815	94108	37.7929	-122.4079
8	2	Western Addition/Japantown	33115	5.0	0.150989	94115	37.7856	-122.4358
11	2	Inner Richmond	38939	4.0	0.102725	94118	37.7812	-122.4614

Table 4.5: Cluster 3



Cluster Labels	Neighborhood	Population	Pizza Place Count	num/pop	Postal Code	latitude	longitude
3	South of Market	23016	1.0	0.043448	94103	37.7725	-122.4147
3	Potrero Hill	17368	1.0	0.057577	94107	37.7621	-122.3971
3	Ingelside-Excelsior/Crocker-Amazon	73104	3.0	0.041037	94112	37.7195	-122.4411
3	Parkside/Forest Hill	42958	2.0	0.046557	94116	37.7441	-122.4863
3	Haight-Ashbury	38738	2.0	0.051629	94117	37.7712	-122.4413
3	Outer Richmond	42473	2.0	0.047089	94121	37.7786	-122.4892
3	Sunset	55492	2.0	0.036041	94122	37.7593	-122.4836
3	Bayview-Hunters Point	33170	1.0	0.030148	94124	37.7309	-122.3886
3	Lake Merced	26291	1.0	0.038036	94132	37.7211	-122.4754

5. Discussion

As shown in the above results it can be stated that the neighborhoods in the cluster 0 are the best options for a new pizza place to start only considering the population and existing number of pizza places in the particular neighborhood. Whereas, cluster 2 is very competitive thus, not the best option for a new business but suitable for an established food chain to expand their business.

6. Conclusion

The clustering is done only using the population and number of pizza places in each neighborhood. But in reality, there are more factors that affects stating of a new business such as the GDP of the neighborhood, the cost of living, taxes and people’s interest etc. Also, the used data source is from few years back thus the current situation could have been changed. In conclusion it can be stated that the work carried out in this project can be really useful for small business looking for new opportunities and this can be improved drastically by adding new features and updated data.