Searching for Good Location for a Chinese restaurant in San Francisco (SF)

Zijie Xia

Apr 30th, 2020

1. Introduction

1.1 Background

San Francisco (SF) is one of the major cities in the United States and one of the most diverse cities as well. The population of San Francisco is close to 900,000. With a diverse community and tourist industry, SF has a large number of restaurants featuring a wide range of exotic cuisines. Many factors need to be considered to open a Chinese restaurant in such a major city, even though delicious food and excellent services might be guaranteed. Location is a significant factor to consider when a new business is opened in the town. Analyzing each neighborhood in SF as a potential location for a new Chinese restaurant is advantageous for any potential business owners. For example, this information can be used to target available commercial rental spots.

1.2 Problem

This project aims to predict which neighborhood in SF is better to open a new Chinese restaurant. Data that might contribute to determining competition level and population might include the number of existing competitors and population data.

1.3 Interest

Any potential business owner would be interested in seeing which neighborhood would be suitable for open a restaurant. Others who are interested in rent their properties may also be interested.

2. Data acquisition and cleaning

2.1 Data sources

SF neighborhood data is downloaded from https://data.sfgov.org/Geographic-Locations-and-Boundaries/SF-Find- Neighborhoods/pty2-tcw4. The raw data set

contains 117 rows and 3 features. The 3 features are the link of the website containing the information of the neighborhoods, polygon geo-data, and neighborhood name. SF zip code data is scraped from http://www.healthysf.org/bdi/outcomes/zipmap.html. In total 21 rows and 3 features in the raw dataset for zip code-based SF neighborhood information. The 3 features are zip code, neighborhood, and population. SF latitude and longitude data are imported from a separate csv file. https://github.com/ZBerylX-lab/first-repository/blob/master/SF%20zip%20code%20LatLon.xlsx?raw=true. The csv file contains 21 rows and 3 features. The 3 features are zip code, latitude, longitude. The restaurants' information for each neighborhood in SF is obtained from the FourSquare API.

2.2 Data cleaning

Data downloaded or scraped from multiple sources were combined into one table that contains the neighborhood information of SF. There were no missing values. All the neighborhoods of SF are combined into zip code based neighborhood and location information. There is no missing data. There were some issues when the data frames got imported. The row summarizes total for the neighborhoods is deleted, and the column title has been renamed.

The category information of the restaurant is extracted from the json data file obtained from FourSquare. The number of popular types of Asian restaurants (not including Chinese restaurants) and the number of Chinese restaurants are summarized with the data and appended on the data frame that contains the location data of SF neighborhoods.

3. Exploratory data analysis/Methodology

3.1 Visualization SF neighborhood

The neighborhoods of SF are visualized with folium package with the geojson data.

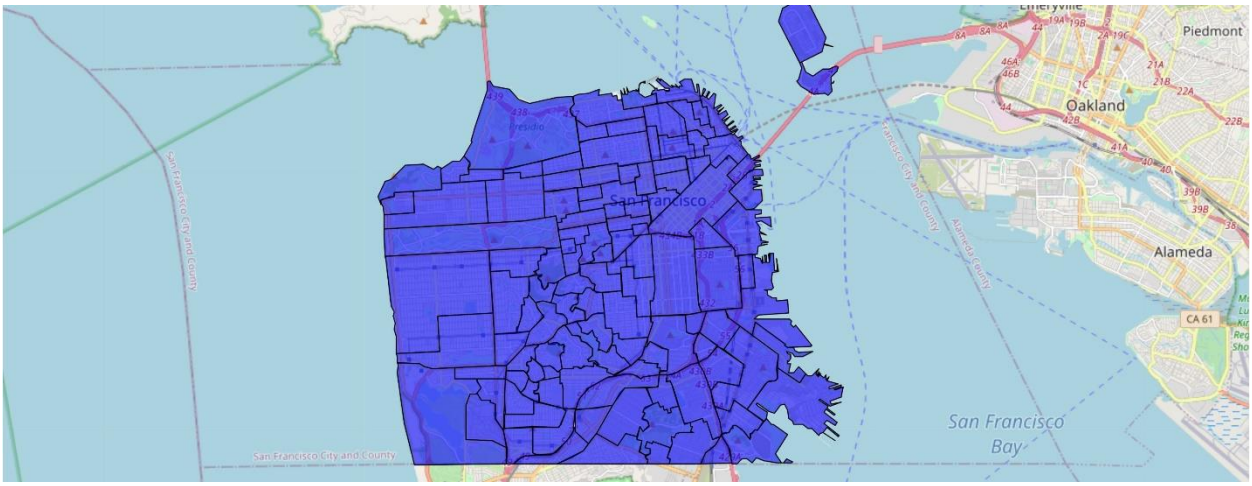

Figure 1: The neighborhoods of SF are imposed on the map.

There are 117 neighborhoods in SF. The sizes of neighborhoods vary significantly depending on the location. Many neighborhoods are not in a shape that can be easily characterized (Figure 1). FourSquare also has a difficult time to return a geolocator data based on the neighborhood name. For the project, grouping the neighborhoods by zip codes are more ideal.

3.2 Visualization SF neighborhood by neighborhood

The neighborhoods of SF is visualized with folium package with the latitude and longitude data.

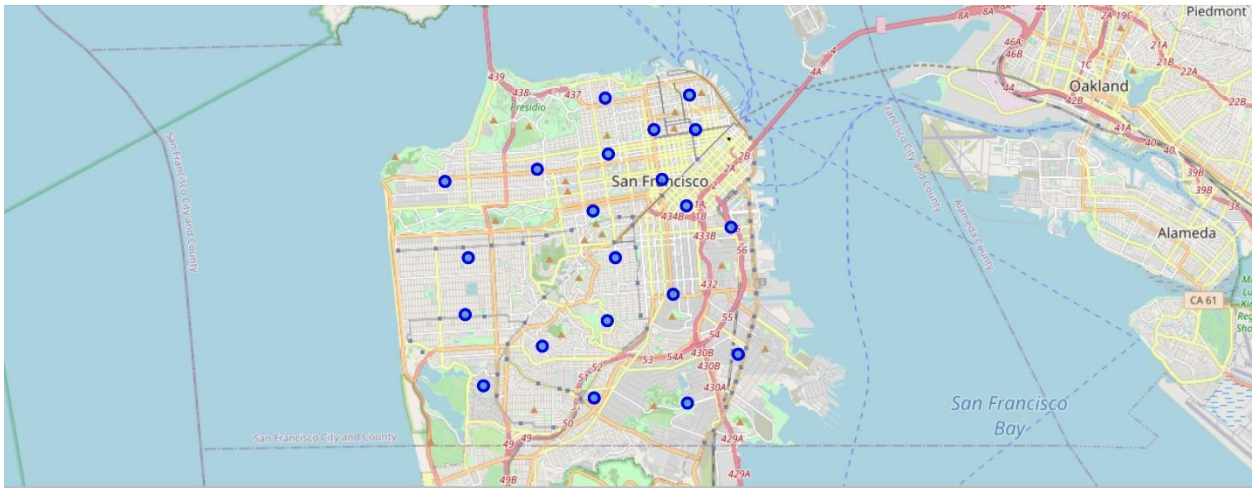

Figure 2: The neighborhoods of SF grouped by neighborhood are imposed on the map.

There are in total of 21 zip codes in SF. Each zip code covers a relatively consistent size of the area (Figure 2). The sizes of the neighborhood of West SF and South SF are slightly larger than Northeast SF.

3.3 Summary of the number of Asian and Chinese restaurants in SF neighborhoods

The number of Asian and Chinese restaurants are counted for each neighborhood of SF. FourSquare returns 100 venues under the food section for each neighborhood for a set radius (5000 meters). Therefore, there are no significant differences in the sampling for each neighborhood. The Asian restaurants are defined as favored types of Asian restaurants, such as sushi, noodle house, and so on, but do not include any Chinese restaurants. Chinese restaurants are defined as Chinese restaurants and any sub-category of Chinese restaurants. Doing statistical studies of the number of restaurants in SF neighborhoods can provide a general view of the competition of restaurants in SF.

3.4 K-means Clustering for SF neighborhoods

K-means Clustering is performed on the SF neighborhood dataset. Three features are used to label SF neighborhoods into four categories. The population of the neighborhood, the number of Chinese restaurants in the neighborhood, and the number of Asian restaurants in the neighborhood are the three features. All the numbers are normalized before the K-means clustering algorithm being applied. The 21 neighborhoods are divided into 4 clusters. Since the number of the neighborhoods is not big, it does not make sense to cluster them into smaller groups. The k-means clustering should provide a general view of how competitive and how many potential customers that each neighborhood is or has.

4. Results

4.1 Data of the number of restaurants in SF neighborhoods

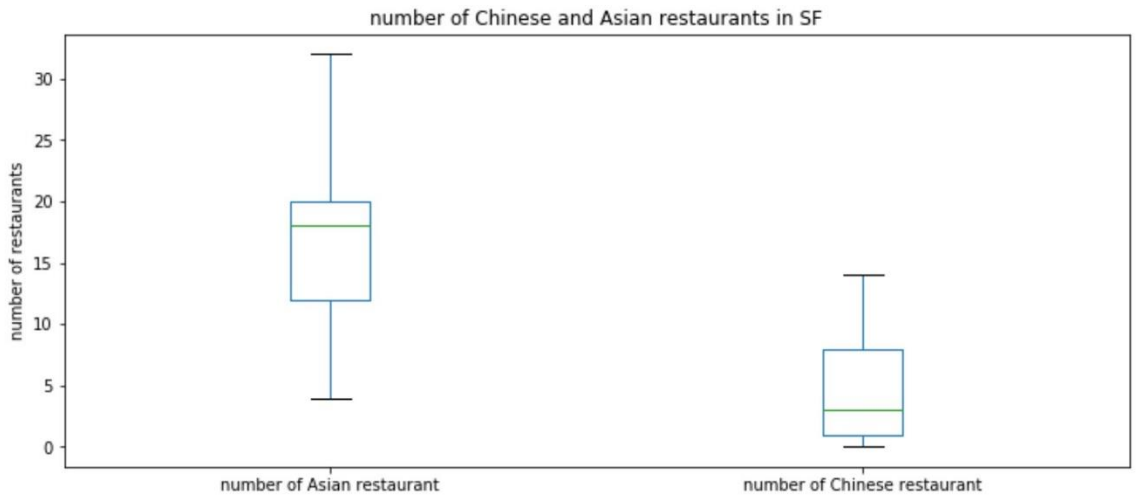

Figure 3: The box plot that summarizes the number of Chinese restaurants and the number of Asian restaurants in SF neighborhoods.

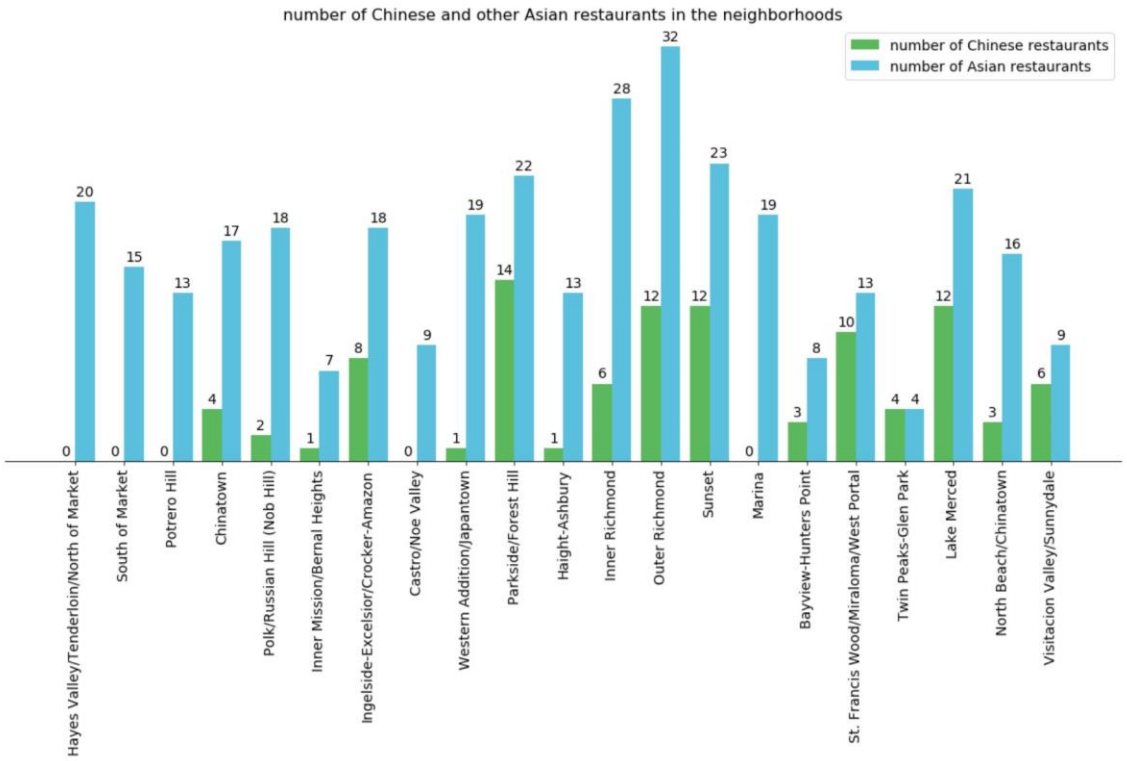

Figure 4: The bar plot that shows the number of Chinese restaurants and the number of Asian restaurants in SF neighborhoods

On average, there are 18 Asian restaurants and 3 Chinese restaurants in SF neighborhoods (Figure 3). There are several neighborhoods in SF having zero Chinese restaurants while they have around an average or above-average number of Asian restaurants. On the other hand, in one neighborhood that has four Asian restaurants, it also has four Chinese restaurants. This specific neighborhood has the highest ratio of the number of Chinese restaurants to the number of Asian restaurants. The majority of the neighborhoods has way more Asian restaurants than Chinese restaurants (Figure 4). Also, both the number of Chinese restaurants and the number of Asian restaurants have large variability in SF.

4.2 Data on the number of restaurants in SF neighborhoods

Zip Code	Neighborhood	Population	Latitude	Longitude	number of Chinese restaurant	number of Asian restaurant	label
94102	Hayes Valley/Tenderloin/North of Market	28991	37.77933	-122.419	0	20	0
94103	South of Market	23016	37.77233	-122.411	0	15	0
94107	Potrero Hill	17368	37.76653	-122.396	0	13	0
94108	Chinatown	13716	37.79268	-122.408	4	17	0
94109	Polk/Russian Hill (Nob Hill)	56322	37.79278	-122.422	2	18	2
94110	Inner Mission/Bernal Heights	74633	37.74873	-122.415	1	7	2
94112	Ingelside-Excelsior/Crocker-Amazon	73104	37.72093	-122.442	8	18	2
94114	Castro/Noe Valley	30574	37.75843	-122.435	0	9	0
94115	Western Addition/Japantown	33115	37.78613	-122.437	1	19	0
94116	Parkside/Forest Hill	42958	37.74338	-122.486	14	22	1
94117	Haight-Ashbury	38738	37.77094	-122.443	1	13	0
94118	Inner Richmond	38939	37.78203	-122.462	6	28	1
94121	Outer Richmond	42473	37.77873	-122.493	12	32	1
94122	Sunset	55492	37.75838	-122.485	12	23	1
94123	Marina	22903	37.80103	-122.438	0	19	0
94124	Bayview-Hunters Point	33170	37.7328	-122.393	3	8	3
94127	St. Francis Wood/Miraloma/West Portal	20624	37.73496	-122.46	10	13	3
94131	Twin Peaks-Glen Park	27897	37.7418	-122.438	4	4	3
94132	Lake Merced	26291	37.72423	-122.48	12	21	1
94133	North Beach/Chinatown	26827	37.80188	-122.41	3	16	0
94134	Visitacion Valley/Sunnydale	40134	37.71958	-122.411	6	9	3

Table 1: Summarized clustering results for SF neighborhood.

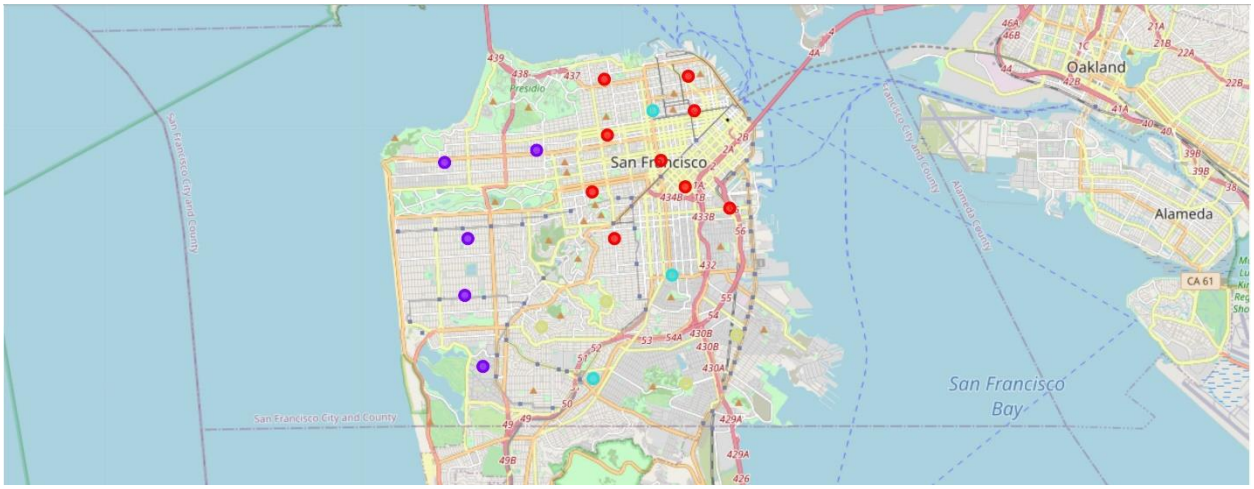

Figure 5: The map shows the cluster of neighborhoods in SF based on populations and the numbers of Chinese and Asian restaurants.

There are nine neighborhoods labeled 0 (red), five neighborhoods labeled 1 (purple), three neighborhoods labeled 2 (green-blue), and four neighborhoods labeled 3 (yellow) in Table 1 and Figure 5. All the neighborhoods labeled 0 are in Northeast SF, and all the neighborhoods labeled 1 are in West SF. The neighborhoods labeled 2 and 3 are more spread out and intertwined.

5. Discussion

The four clusters of SF neighborhoods corresponding to the following types: 1). Label 0 corresponds to neighborhoods with little Chinese restaurants and low populations 2). Label 1 corresponds to neighborhoods with average populations and with high numbers of Chinese and Asian restaurants 3). Label 2 corresponds to the neighborhoods with an average number of Chinses and Asian restaurants and high populations 4). Label 3 corresponds to the neighborhoods with an average number of Chinese and Asian restaurants and average populations.

Ideally, for a new restaurant, locations with average to high populations are more likely to succeed. Also, a restaurant is more likely to be successful if there are fewer competitions in the area. Based on the clustering results, the neighborhood with the fewest competition also has a low population, which makes it less ideal. However, the neighborhood labeled with label 2 has an average number of Chinese and Asian restaurants and relatively high populations. Although the competition is higher but the potential customers are way more. Based on Figure 5, there are only three neighborhoods out of 21 neighborhoods in SF that are labeled Label 2. One of them is in Northeast SF, and two of them are in South SF. Based on the street density in Figure

5, the location in Northeast SF is in the densest area of SF, which should attract most of the foot traffic. Therefore, that might be an appropriate candidate.

6. Conclusion

In this study, k-means clustering is used to group SF neighborhoods into four groups based on populations and the number of Asian and Chinese restaurants. Based on the clustering results, we find neighborhoods that have high populations but an average number of competitors for a new Chinese restaurant. Based on the street density, we find the best neighborhood is in Northeast SF that potentially can attract most of the foot traffic but not have the most competitors.

7. Future Direction

Many factors are not considered in this study. The number of features can be further expanded. For example, the average neighborhood income or the ratio of each ethnicity race of each neighborhood or average rent of restaurants can be added to the study to get a more thorough review of the neighborhoods. Another factor that can be considered is whether the neighborhood is in a business district or a residential district. This can influence the decision of whether to focus on customers that are looking for a quick meal or a fancy meal. More models can also be used in this study to determine ideas, such as which factor has a more substantial influence on the foot traffic in a neighborhood.