Shiyi Wang

Final Project of Applied Data Science Capstone

A STUDY OF SAN FRANCISCO BAY AREA FOOD MAP

Shiyi Wang

7/25/2019

# INTRODUCTION

As part of the final project of Applied Data Science Capstone Course, this report is to investigate the neighborhood of San Francisco Bay Area in terms of restaurant or food distribution based on data science and algorithms. The main tool applied during the research includes FourSquare API, Python folium, Python Pandas Data Frame, Python Plots, k-means clustering method and etc. The main content of this report is organized in four parts, which are Problem Statement, Data Clean and Methodology, Results and Discussions, and finally Conclusions. The goal of this report is to provide recommendations for potential restaurant owners, food-lover and future residents who plan to move to Bay Area.

# Problem Statement

As one of the most diverse and energetic area, greater San Francisco Bay area has become an important hub of United States that attracts thousands of talented souls and merges various cultures from all of the world. Of course, those talented immigrants and explorers bring their customs and cuisines with them, which makes Bay Area, a desert centuries ago, become a heaven of food. In the meanwhile, in order to heritage the innovation spirit of Silicon Valley, it is highly required to apply data science to research the data of food, including the restaurant distribution and the “story behind it” that data would like to tell.

The main purpose of this report contains three points as below,

1. To recommend a suitable location or neighborhood to operate a restaurant for its manager
2. To recommend a community to live in for food-lovers who decide to move to Bay Area

# Data Clean and Methodology

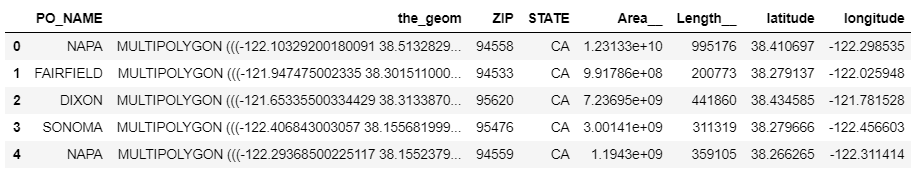
The data used in this project is mainly from two sources. The main raw data is downloaded from “data.sfgov.org”, which covers Greater San Francisco Area as in Table 1. This table contains city name, the multi-polygon information with latitude and longitude of each polygon points that encloses the region, zip-code, state, region area and length of region contour. Column “PO\_NAME”, “the\_geom” and “ZIP” values are used as input of FourSquare API. Particularly, pairs of city’s latitude and longitude are calculated based on the averaged value of full “MULTIPOLYGON” array.

Table Header of Greater San Francisco Area geographic raw data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| PO\_NAME | the\_geom | ZIP | STATE | Area\_\_ | Length\_\_ |
| NAPA | MULTIPOLYGON (((-122.10329200180091 38.51328299864645, -122.10348400246679 38.51190700043408, …))) | 94558 | CA | 1.23E+10 | 995176.2 |
| FAIRFIELD | MULTIPOLYGON (((-121.947475002335 38.30151100055196, -121.94717500296504 38.299311999636366, …))) | 94533 | CA | 9.92E+08 | 200772.6 |
| DIXON | MULTIPOLYGON (((-121.65335500334429 38.31338700062926, -121.69340200238318 38.313735999442926, …))) | 95620 | CA | 7.24E+09 | 441860.2 |
| SONOMA | MULTIPOLYGON (((-122.406843003057 38.1556819994161, -122.40757000042267 38.155718998957866, …))) | 95476 | CA | 3E+09 | 311318.5 |
| NAPA | MULTIPOLYGON (((-122.406843003057 38.1556819994161, -122.40757000042267 38.155718998957866, …))) | 94559 | CA | 1.19E+09 | 359104.6 |

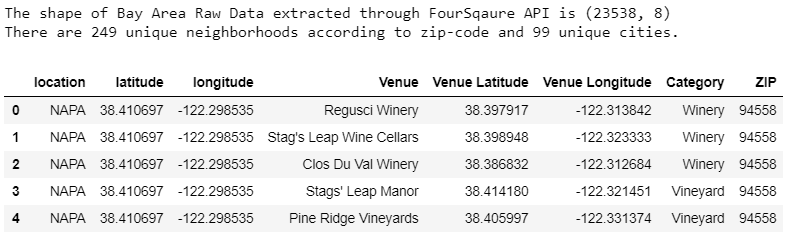
On the other hand, in order to complete the full bay area, a second dataset with San Jose geometric information is downloaded from “public.opendatasoft.com”. Similarly, city name, latitude/longitude, zip-code and state are extracted and combined to the main raw data through pandas.DataFrame manipulation. Table 2 displays the bay area geographic data after merging.

Table Header of Bay Area geographic data after merging



Furthermore, the merged raw data are passed into Foursquare API. As a valid input, this API needs CLIENT ID, CLIENT SECRET and VERSION as identity recognition, latitude/longitude/radius as geographic conditions, and LIMIT number as total returned result number. The returned values include “Venue”, “Venue latitude”, “Venue longitude” and its business type “Category”. By using pandas.Dataframe, the data can be easily re-organized and cleaned, and its final output is as shown in Table 3.

Table FourSquare Venue Information after data clean



Next, the cleaned venue table will be applied with one hot encoding. One hot encoding is a process by which categorical variables are converted into a form that could be provided to Machine Learning algorithms to do a better job in prediction. After data filtering, only the categories of restaurants and food remain in the table. Then according to the number of top venues, the final dataset looks like,

Table Bay Area Restaurants in the form of "Most Common Venue" (header of table)



The table is now ready for K-means clustering, which is a method of vector quantization, originally from signal processing, which is popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This also ends the data cleaning stage.

# Results and Discussions

Firstly, by using folium python package, the raw bay area location data can be easily visualized in the form of map as shown in Figure 1. Here, the blue circles represent the center of each neighborhood in terms of zip-code. The generated map is interactive as the user can adjust its size and click for details.

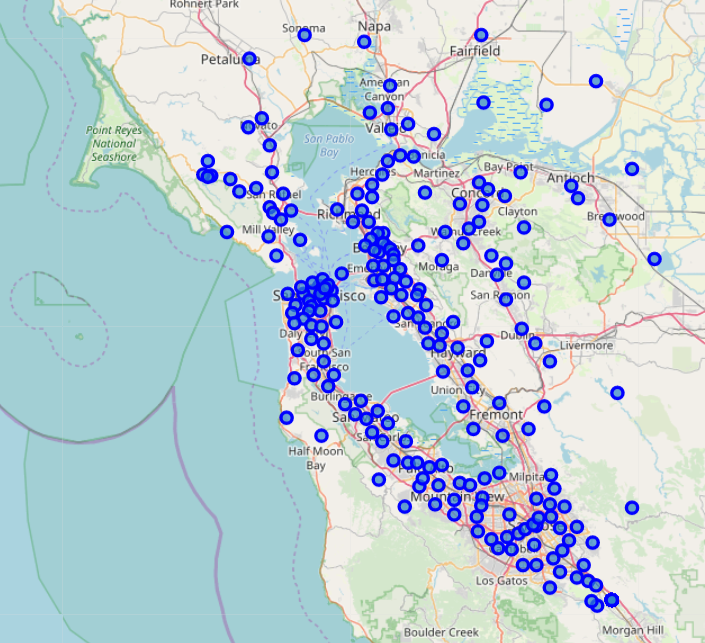
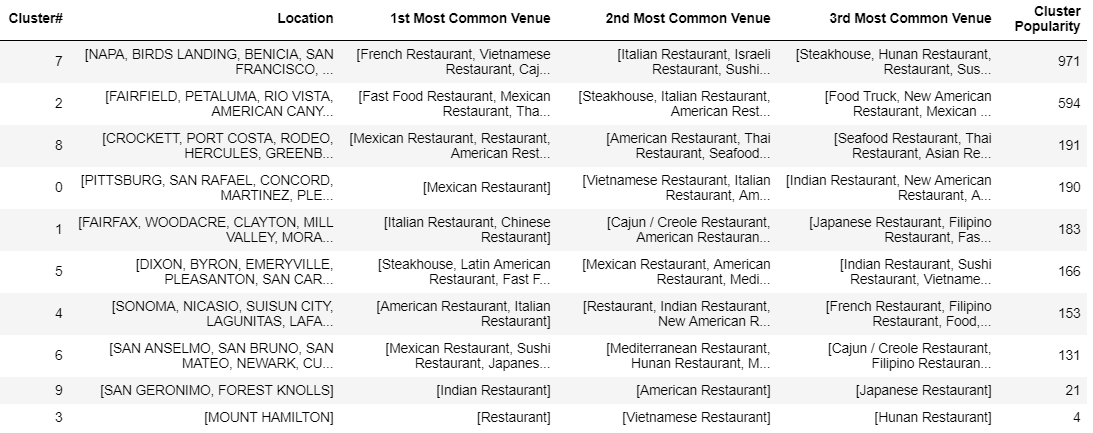


Figure Bay Area City Locations in terms of zip-code (baseline)

By importing clustering stage from sklearn package, we can apply k-means clustering for further data mining, whose result is shown as in Table 5. Also, the areas with different zip-codes but same city are also merged into one venue. The restaurant analysis result is ranked by popularity of each cluster.

Table Clustered Bay Area Restaurant Data (10 clusters are used)



Furthermore, the clustered result is also visualized through interactive map, which provides more convenience for analysis (Figure 2).

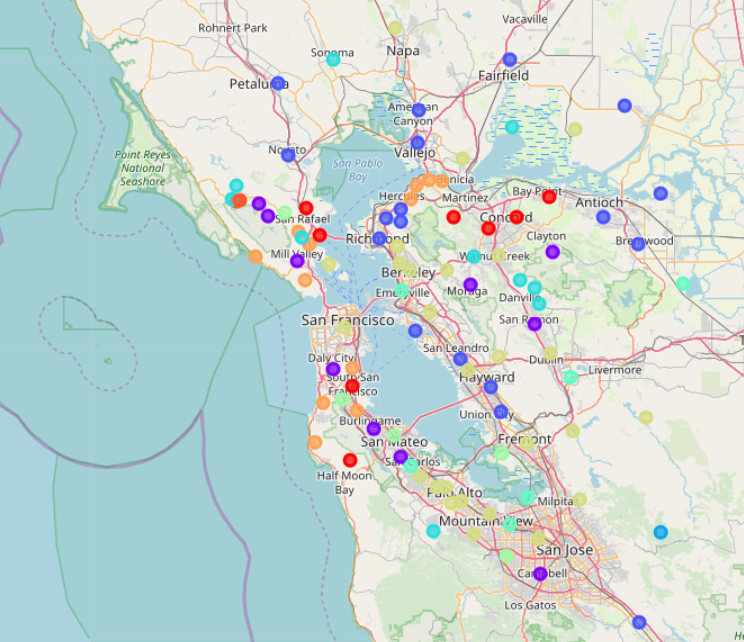


Figure Bay Area Restaurant Clusters shown in interactive map

By digging into the details of Table 5 (in Appendix), we can figure out the corresponding food preference according to the clustering result.

**Cluster 7:** This is the top 1 cluster with highest popularity (light yellow circle in Figure 2). The restaurant preference of Cluster 7 is very diverse, which includes French, Vietnamese, Japanese, Italian, American, Chinese and etc. For the food-lovers and restaurant managers, the cities that belong to this cluster seem to be welcome to all food categories. Meantime, it also means the competition in these area is expected to be intense. The typical cities from this cluster include San Francisco, Napa, Mountain View, Oakland, Los Altos, Palo Alto, Milpitas, Fremont and Santa Clara. We can see most of the cities in this cluster has a very dense population. This also concludes that the area with dense population and more diverse immigrants tends to have a more open food preference.

To be specify, Santa Clara becomes an individual cluster if total cluster number increases, which shows a strong preference to Korean food.

**Cluster 2**: Ranked as No.2 cluster in popularity, this cluster shows strong preference to Mexican food and Vietnamese food followed by American food and fast food (blue circle in Figure 2). The main cities include San Jose, Union City, Hayward and etc. This makes perfect sense, since San Jose, for example, owns both Mexican community and Vietnamese community with large size.

**Cluster 8**: Ranked as No.3 cluster in popularity, this cluster shows a more diverse preference compared to Cluster 2, but the preferred restaurant categories are much fewer than Cluster 7. The favorite food category of this cluster is Mexican food, Vietnamese food, American food, Chinese food and Seafood.

**Cluster 0**: (Red circle in Figure 2) No.4 cluster in popularity. The residents from the areas under this cluster are solid Mexican food-lovers.

**Cluster 1**: No.5 cluster in popularity. The favorite food under this cluster is Chinese food and Italian food (purple circle in Figure 2). The key cities in this cluster include Campbell and Daly City.

**Cluster 5**: (Green blue in Figure 2) No.6 cluster in popularity. They love steakhouse, Latin restaurant, Mediterranean Restaurant and Vegetarian restaurant. The key city is Sunnyvale.

**Cluster 4**: (Light blue in Figure 2) No.7 cluster in popularity. They love American Restaurant and Italian Restaurant. These two category are also the most two common restaurants as in the nationwide range.

**Cluster 6**: (Yellow green in Figure 2) No.8 cluster in popularity. They love Sushi or Japanese Restaurant, and also some Mexican Restaurant. The key cities in cluster include Cities: San Anselmo, San Bruno, San Mateo, Newark and Cupertino.

Meanwhile, let’s also investigate what are the most popular food in bay area (Figure 3). In this figure, normalized popularity values are shown for each restaurant category. We can clearly see the top four restaurants are Mexican food, American food, Italian food and Vietnamese food. Then second popular group to bay area people includes fast food, Indian food and Japanese/Sushi food. Besides the shared popular food as in the nationwide range of United States, which are American food and Italian food, the data of bay area demonstrates a strong preference to Mexican food and Vietnamese. To help food-lovers find their target, Appendix provides detailed analysis results with all city names and the preferred food accordingly.

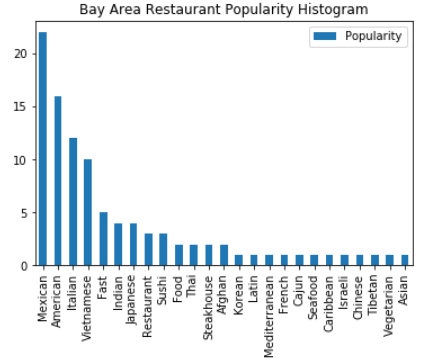


Figure Bay Area Restaurant Popularity

# Conclusions

In summary, the favorite cuisine of bay area people includes Mexican food, American food, Italian food and Vietnamese food. People has also shown their preference in the categories of Indian food, Japanese/Sushi food, Thai food and steakhouse. The food-lovers or restaurant managers can select the cities under Cluster 7 according to the clustered food map for a diverse environment, and also pick up the area with strong preferences under Cluster2, Cluster0, Cluster1, Cluster5 and Cluster4 based on the selector’s own favorite. This food map can also provide a community distribution, as the cluster with more food categories tend to have diverse residents with different backgrounds. On the other hand, the cluster with few food categories may also indicate a concentrated community with single background.

**APPENDIX**

\*\*\* Cluster #0 with Popularity of 190 loves Mexican Restaurant.

This cluster includes Cities:

PITTSBURG, SAN RAFAEL, CONCORD, MARTINEZ, PLEASANT HILL, SAN QUENTIN, SOUTH SAN FRANCISCO, HALF MOON BAY

\*\*\* Cluster #1 with Popularity of 183 loves Italian Restaurant, Chinese Restaurant.

This cluster includes Cities:

FAIRFAX, WOODACRE, CLAYTON, MILL VALLEY, MORAGA, SAN RAMON, DALY CITY, BURLINGAME, BELMONT, CAMPBELL

\*\*\* Cluster #2 with Popularity of 594 loves Fast Food Restaurant, Mexican Restaurant, Thai Restaurant, Vietnamese Restaurant, American Restaurant.

This cluster includes Cities:

FAIRFIELD, PETALUMA, RIO VISTA, AMERICAN CANYON, NOVATO, VALLEJO, OAKLEY, ALAMEDA, ANTIOCH, BRENTWOOD, PINOLE, RICHMOND, SAN PABLO, EL SOBRANTE, SAN LEANDRO, HAYWARD, UNION CITY, SAN JOSE

\*\*\* Cluster #3 with Popularity of 4 loves Restaurant.

This cluster includes Cities:

MOUNT HAMILTON

\*\*\* Cluster #4 with Popularity of 153 loves American Restaurant, Italian Restaurant.

This cluster includes Cities:

SONOMA, NICASIO, SUISUN CITY, LAGUNITAS, LAFAYETTE, LARKSPUR, ALAMO, DIABLO, DANVILLE, PORTOLA VALLEY

\*\*\* Cluster #5 with Popularity of 166 loves Steakhouse, Latin American Restaurant, Fast Food Restaurant, Afghan Restaurant, Vegetarian / Vegan Restaurant, Mediterranean Restaurant.

This cluster includes Cities:

DIXON, BYRON, EMERYVILLE, PLEASANTON, SAN CARLOS, ALVISO, SUNNYVALE

\*\*\* Cluster #6 with Popularity of 131 loves Mexican Restaurant, Sushi Restaurant, Japanese Restaurant, Fast Food Restaurant.

This cluster includes Cities:

SAN ANSELMO, SAN BRUNO, SAN MATEO, NEWARK, CUPERTINO

\*\*\* Cluster #7 with Popularity of 971 loves French Restaurant, Vietnamese Restaurant, Cajun / Creole Restaurant, Italian Restaurant, Food Truck, Mexican Restaurant, Sushi Restaurant, Tibetan Restaurant, Caribbean Restaurant, American Restaurant, Thai Restaurant, Afghan Restaurant, Japanese Restaurant, Indian Restaurant, Israeli Restaurant, Korean Restaurant.

This cluster includes Cities:

NAPA, BIRDS LANDING, BENICIA, SAN FRANCISCO, WALNUT CREEK, OAKLAND, EL CERRITO, BERKELEY, ORINDA, ALBANY, BELVEDERE TIBURON, LIVERMORE, CASTRO VALLEY, DUBLIN, SAN LORENZO, SUNOL, FREMONT, REDWOOD CITY, ATHERTON, PALO ALTO, MENLO PARK, MILPITAS, MOUNTAIN VIEW, STANFORD, LOS ALTOS, SANTA CLARA

\*\*\* Cluster #8 with Popularity of 191 loves Mexican Restaurant, Restaurant, American Restaurant, Vietnamese Restaurant, Asian Restaurant, Seafood Restaurant.

This cluster includes Cities:

CROCKETT, PORT COSTA, RODEO, HERCULES, GREENBRAE, CORTE MADERA, STINSON BEACH, SAUSALITO, BRISBANE, PACIFICA, MILLBRAE, MOSS BEACH

\*\*\* Cluster #9 with Popularity of 21 loves Indian Restaurant.

This cluster includes Cities:

SAN GERONIMO, FOREST KNOLLS