IBM Data Science Professional Certificate

Course No.9: Applied Data Science Capstone

The Battle of Neighborhoods (Weeks 4 and 5)

Final Report

1. Introduction/Business Problem

A restaurant business company would like open an Indian restaurant in the San Francisco Bay Area, California. There are a lot of Indians working in the Bay area tech companies. Also, considerable number of Indian students are studying in the universities and colleges located in the Bay Area. Along with Indians, people from other nations also like Indian food very much. Therefore, San Francisco Bay Area has good potential to open an Indian restaurant.

San Francisco Bay Area has made up of nine counties, namely, Santa Clara, Alameda, San Mateo, San Francisco, Contra Costa, Marin, Solano, Napa, and Sonoma. There are 186 ZIP codes and 97 boroughs in the Bay Area. Since there are a large number of boroughs in the Bay Area, it is important to understand which place is the best to start a new branch of an Indian restaurant.

Target Audience: Client is a restaurant business company. As a data analyst, we should help the restaurant business company that wants to start an Indian restaurant by recommending the best suitable borough and explain why this place is better than others in the Bay Area.

Success Criteria: The success criteria of the project will be a good recommendation of borough in the Bay area for an Indian restaurant business.

2. Data

The San Francisco Bay Area ZIP codes and PO_NAME are downloaded from DataSF open data website. The following csv file downloadable link is used to create a Pandas dataframe in Python. https://data.sfgov.org/api/views/f9wk-m4qb/rows.csv?accessType=DOWNLOAD
From this dataframe, a new dataframe is generated with ZIP codes and PO_NAME, and then PO_NAME is renamed to 'Borough'.

The geographical coordinates are retrieved for each borough in the dataframe using the geocoder and then leveraged the Foursquare AIP service to collect the location data, explore each borough of the Bay area and obtain information about different venue categories, especially restaurants, cafes, bars, pubs, etc. within 1000 m radius from the geographical coordinates. This data is used to understand better that how many eateries are available in the Bay area boroughs. Then the

dataset is analyzed systematically to identify best suitable boroughs to open an Indian restaurant business.

3. Methodology

Business Understanding: The project main goal is to find optimal boroughs in the San Francisco Bay Area for a new Indian restaurant business.

Analytic Approach: There are two steps to tackle this problem. In the first step, the objective is to find borough where there are a lot of eateries in the Bay area. Then these boroughs are used as "Center of Attraction" for dining with different food varieties. At the same time from these boroughs, exclude boroughs where there are Indian restaurants. This is to avoid competition with other Indian restaurants. But, it is supposed that an Indian restaurant can still attract a part of customers where there are a lot of restaurants with other kinds of dishes. Thus,

(i) Find , for example, top 25 boroughs with maximum number of restaurants, cafes, bars, pubs, etc.;
(ii) Find boroughs with Indian restaurants;
(iii) Exclude the boroughs with Indian restaurants from the top 25 borough with restaurants, cafes, etc.;
(iv) Find short-list of boroughs where there are a lot of restaurants but no Indian restaurant.

In the second step, the goal is to select clusters similar to boroughs with Indian restaurants. In any one of the boroughs where there is an Indian restaurant that continuous to work successfully, then we should find borough with similar conditions. Therefore, k-means clustering algorithm is used to find clusters.

Exploratory Data Analysis:

1. Download the csv-file from https://data.sfgov.org/api/views/f9wk-m4qb/rows.csv?accessType=DOWNLOAD and convert the csv-file to a dataframe using Pandas.

	PO_NAME	the_geom	ZIP	STATE	Area__	Length__
0	NAPA	MULTIPOLYGON (((-122.10329200180091 38.5132829...	94558	CA	1.231326e+10	995176.225313
1	FAIRFIELD	MULTIPOLYGON (((-121.947475002335 38.301511000...	94533	CA	9.917861e+08	200772.556587
2	DIXON	MULTIPOLYGON (((-121.65335500334429 38.3133870...	95620	CA	7.236950e+09	441860.201400
3	SONOMA	MULTIPOLYGON (((-122.406843003057 38.155681999...	95476	CA	3.001414e+09	311318.546326
4	NAPA	MULTIPOLYGON (((-122.29368500225117 38.1552379...	94559	CA	1.194302e+09	359104.646602

Fig. 1 A Part of the data frame containing PO_NAME, etc. from DataSF open data.

2. Choose Bay area ZIP codes and PO_NAME columns from the dataframe and form a new dataframe using Pandas by renaming the 'PO_NAME' column to 'Borough'.

	ZIP	Borough
0	94558	NAPA
1	94533	FAIRFIELD
2	95620	DIXON
3	95476	SONOMA
4	94559	NAPA
5	94954	PETALUMA
6	94571	RIO VISTA
7	94535	TRAVIS AFB
8	94503	AMERICAN CANYON
9	94949	NOVATO
10	94945	NOVATO
11	94512	BIRDS LANDING

Fig. 2 A Part of the dataframe containing Bay area ZIP codes and Borough names.

3. Find latitude and longitude using Python geocoder library and add to the list of boroughs.

	ZIP	Borough	Latitude	Longitude
0	94558	NAPA	38.026885	-121.881924
1	94533	FAIRFIELD	38.249240	-122.044300
2	95620	DIXON	59.383710	-108.801550
3	95476	SONOMA	38.291880	-122.457280
4	94559	NAPA	38.026885	-121.881924
5	94954	PETALUMA	38.232470	-122.644440
6	94571	RIO VISTA	38.155540	-121.690430
7	94535	TRAVIS AFB	38.271530	-121.933030
8	94503	AMERICAN CANYON	38.168050	-122.252770
9	94949	NOVATO	38.106090	-122.567900
10	94945	NOVATO	38.106090	-122.567900
11	94512	BIRDS LANDING	38.132560	-121.875140

Fig. 3 A Part of the dataframe containing Boroughs with Latitude and Longitude.

4. Create a map of all boroughs using Folium.

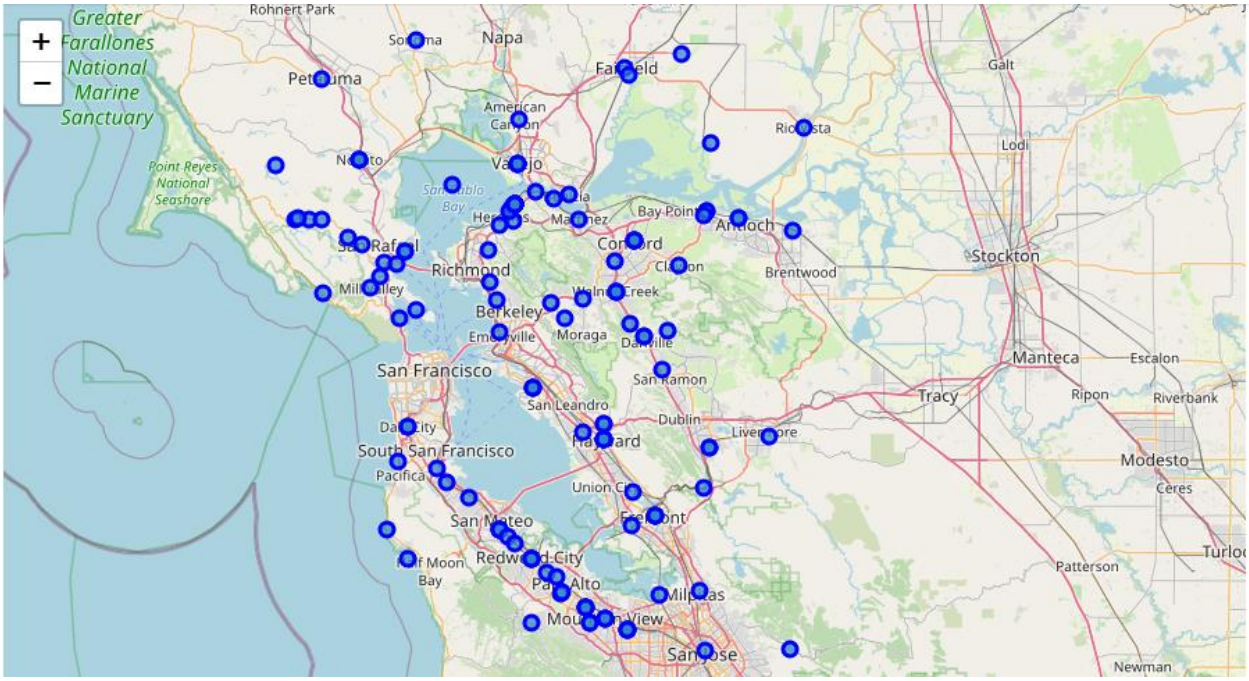

Fig. 4 Map of all Boroughs of San Francisco Bay Area.

5. Find nearby venues for every borough by using Foursquare API with radius 1000 m from the geographical coordinates of each borough. It is found that there are 353 unique venue categories.

	Borough	Borough Latitude	Borough Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	NAPA	38.026885	-121.881924	Steeltown Coffee & Tea	38.030414	-121.884043	Café
1	NAPA	38.026885	-121.881924	E.J. Phair Brewing Co.	38.033092	-121.882404	Brewery
2	NAPA	38.026885	-121.881924	Lumpy's Diner	38.031060	-121.883959	Diner
3	NAPA	38.026885	-121.881924	Pittsburg Farmers' Market	38.031399	-121.884030	Farmers Market
4	NAPA	38.026885	-121.881924	New Mecca Cafe	38.033207	-121.883478	Mexican Restaurant
5	NAPA	38.026885	-121.881924	Laveranda Ristorante Italiano	38.032210	-121.883890	Italian Restaurant
6	NAPA	38.026885	-121.881924	CreAsian	38.031041	-121.884045	Vietnamese Restaurant
7	NAPA	38.026885	-121.881924	Pittsburg City Park	38.022315	-121.889527	Park
8	NAPA	38.026885	-121.881924	Dad's BBQ	38.029681	-121.884373	BBQ Joint
9	NAPA	38.026885	-121.881924	Sassy Sweets	38.029886	-121.884450	Dessert Shop
10	NAPA	38.026885	-121.881924	La Raza Market (The Store With The Chinese Fo...	38.027287	-121.884758	Liquor Store
11	NAPA	38.026885	-121.881924	My Generator Guy	38.026083	-121.878897	Construction & Landscaping

Fig. 5 A Part of the dataframe containing nearby venues for each borough.

6. Find the boroughs with restaurants, cafés, bars, pubs, cuisines within the radius 1000 m from the geographical coordinates of each borough. It is found that there are a total of 2273 different kinds of eateries in the boroughs of San Francisco Bay Area.

	Borough	Borough Latitude	Borough Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	NAPA	38.026885	-121.881924	Steeltown Coffee & Tea	38.030414	-121.884043	Café
1	NAPA	38.026885	-121.881924	New Mecca Cafe	38.033207	-121.883478	Mexican Restaurant
2	NAPA	38.026885	-121.881924	Laveranda Ristorante Italiano	38.032210	-121.883890	Italian Restaurant
3	NAPA	38.026885	-121.881924	CreAsian	38.031041	-121.884045	Vietnamese Restaurant
4	NAPA	38.026885	-121.881924	Skyview Noodle & Tea	38.032825	-121.882040	Asian Restaurant
5	NAPA	38.026885	-121.881924	La Auora	38.029298	-121.890066	Mexican Restaurant
6	NAPA	38.026885	-121.881924	Waterfront Grill & Cafe	38.035077	-121.883607	American Restaurant
7	NAPA	38.026885	-121.881924	KiteCafe	38.035492	-121.883458	Café
8	FAIRFIELD	38.249240	-122.044300	Chez Soul	38.249759	-122.046145	Southern / Soul Food Restaurant
9	FAIRFIELD	38.249240	-122.044300	China Palace	38.249492	-122.045125	Chinese Restaurant
10	FAIRFIELD	38.249240	-122.044300	Rustwater Kitchen & Taproom	38.249426	-122.043313	Beer Bar
11	FAIRFIELD	38.249240	-122.044300	Koong Jyun	38.249259	-122.042890	Korean Restaurant

Fig. 6 A Part of the dataframe containing the boroughs with restaurants, cafés, etc.

7. From the dataframe obtained in step 6, find the boroughs with Indian restaurants, cafés and cuisines. It is found that there are a total of 52 Indian restaurants.

	Borough	Borough Latitude	Borough Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	FAIRFIELD	38.249240	-122.044300	Saffron Indian Cuisine & Bar	38.249067	-122.042989	Indian Restaurant
1	SONOMA	38.291880	-122.457280	Taste of the Himalayas	38.292343	-122.455789	Indian Restaurant
2	SONOMA	38.291880	-122.457280	delhi belly	38.291558	-122.457757	North Indian Restaurant
3	AMERICAN CANYON	38.168050	-122.252770	All Spice Indian Restaurant	38.166265	-122.254110	Indian Restaurant
4	NOVATO	38.106090	-122.567900	Batika India Bistro	38.107328	-122.568293	Indian Restaurant
5	BENICIA	38.052850	-122.153510	Aroma Indian Cuisine	38.050871	-122.157720	Indian Restaurant
6	CONCORD	37.981650	-122.025830	Naan n Curry	37.977223	-122.035039	Indian Restaurant
7	LAGUNITAS	38.014210	-122.696570	Arti Cafe	38.012457	-122.701456	Indian Restaurant
8	FAIRFAX	37.985170	-122.590380	Cafe Lotus	37.987427	-122.588553	Indian Restaurant
9	FAIRFAX	37.985170	-122.590380	Avatars Punjabi Burrito	37.986032	-122.583616	Indian Restaurant
10	WALNUT CREEK	37.901620	-122.061890	Sargam Indian Cuisine	37.907998	-122.058379	Indian Restaurant
11	LAFAYETTE	37.890770	-122.128130	T's Fire House	37.889019	-122.131360	Indian Restaurant

Fig. 7 A Part of the dataframe containing the boroughs with Indian restaurants.

8. Create a sorted list of boroughs with places similar to restaurants, cafes, etc.

Top 25 Boroughs with restaurants and cafés:

Borough	
HAYWARD	180
REDWOOD CITY	176
SUNNYVALE	160
FREMONT	124
WALNUT CREEK	117
SAN FRANCISCO	81
NOVATO	78
MOUNTAIN VIEW	72
PLEASANTON	66
CONCORD	64
PALO ALTO	60
SAN MATEO	60
SAN JOSE	53
CASTRO VALLEY	48
SAN CARLOS	46
LOS ALTOS	44
ALBANY	39
DALY CITY	38
SAN BRUNO	36
MENLO PARK	34
MILLBRAE	32
LAFAYETTE	32
SONOMA	30
PLEASANT HILL	30
VALLEJO	28

9. One of our goals as described earlier is to find boroughs where there are a lot of restaurants except Indian restaurants.

Use data from steps 7 and 8 to create a short list of good boroughs with lot of restaurants and no Indian restaurant.

Borough	
SAN FRANCISCO	81
PALO ALTO	60
SAN MATEO	60
LOS ALTOS	44
DALY CITY	38
MENLO PARK	34
PLEASANT HILL	30
VALLEJO	28

10. The top 5 most common venues in each borough from short list:

----DALY CITY----

	venue	freq
0	Sandwich Place	0.08
1	Fast Food Restaurant	0.08
2	Pizza Place	0.06
3	Fried Chicken Joint	0.05
4	Gas Station	0.05

----LOS ALTOS----

	venue	freq
0	Pizza Place	0.06
1	Italian Restaurant	0.06
2	Mexican Restaurant	0.04
3	Breakfast Spot	0.04
4	Coffee Shop	0.04

```
----MENLO PARK----
                 venue  freq
0          Coffee Shop  0.06
1  Japanese Restaurant  0.04
2     Asian Restaurant  0.04
3       Sandwich Place  0.03
4                Hotel  0.03

----PALO ALTO----
                 venue  freq
0                Hotel  0.10
1                 Park  0.07
2      Sushi Restaurant  0.04
3         Grocery Store  0.04
4  Gym / Fitness Center  0.04

----PLEASANT HILL----
                 venue  freq
0          Pizza Place  0.04
1  Fast Food Restaurant  0.04
2    Chinese Restaurant  0.04
3   American Restaurant  0.03
4    Mexican Restaurant  0.03

----SAN FRANCISCO----
                 venue  freq
0   Convenience Store  0.08
1        Home Service  0.08
2  Mexican Restaurant  0.08
3    Sushi Restaurant  0.08
4       Grocery Store  0.08

----SAN MATEO----
                venue  freq
0         Pizza Place  0.05
1  Mexican Restaurant  0.05
2      Breakfast Spot  0.03
3                Park  0.03
4              Bakery  0.03

----VALLEJO----
                 venue  freq
0         Coffee Shop  0.10
1  Chinese Restaurant  0.07
2       Grocery Store  0.03
3                 Bar  0.03
4               Trail  0.03
```

These boroughs have good infrastructure and we can treat them as “a center of attraction”.

11. Create a dataframe and display the top 10 venues for each borough:

	Borough	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	ALAMEDA	Coffee Shop	Beach	Pool	Skating Rink	Gastropub	Grocery Store	Marijuana Dispensary	Gym / Fitness Center	Arts & Crafts Store	Harbor / Marina
1	ALAMO	Coffee Shop	Pharmacy	Pizza Place	American Restaurant	Sandwich Place	Bank	Fast Food Restaurant	New American Restaurant	Gas Station	Chinese Restaurant
2	ALBANY	Thai Restaurant	Coffee Shop	Pizza Place	Sporting Goods Shop	Indian Restaurant	Pet Store	Mexican Restaurant	Café	Bar	Brewery
3	ALVISO	Food Truck	Hotel	Mexican Restaurant	Hotel Bar	Coffee Shop	Bar	Electronics Store	Café	Gym	Trail
4	AMERICAN CANYON	Park	ATM	Pharmacy	Pizza Place	Chinese Restaurant	Shopping Mall	Convenience Store	Hotel	Brewery	Thrift / Vintage Store

Fig. 8 A Part of the dataframe containing top 10 venues for each borough.

12. Run k-means clustering, find 5 clusters of similar boroughs and create a map.

The red circles in the graph are the best places for Indian restaurant since they have a lot of center of attractions. However, we have to remember that we want to exclude boroughs that are having Indian restaurants, found in step 7, from the boroughs shown in step 8, because we don’t want to compete with the existing Indian restaurants.

If we compare the short listed boroughs from step 10 and Fig. 9 we can recommend DALY CITY, MENLO PARK and PLEASANT HILL are the best boroughs for a new Indian restaurant business.

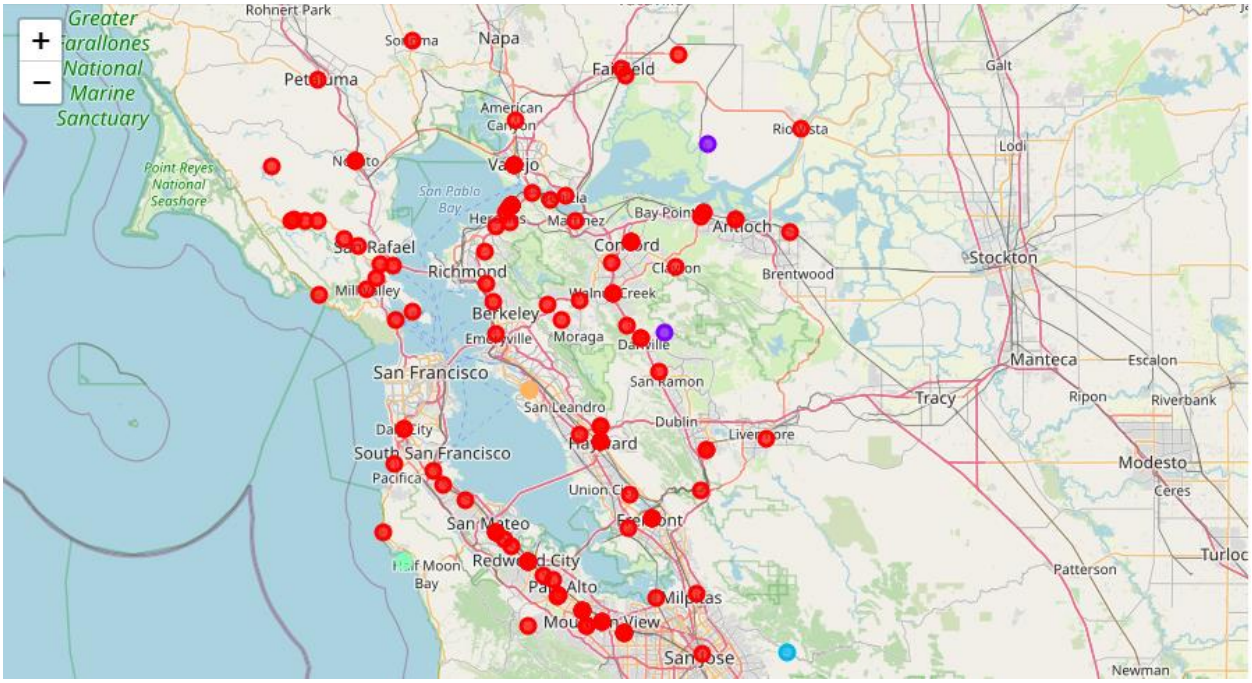

Fig. 9 Map of similar borough clusters in San Francisco Bay Area.

4. Results

The following are the main outcomes of the project work:

i. DALY CITY, MENLO PARK and PLEASANT HILL are the best boroughs for a new Indian restaurant business in the San Francisco Bay Area.
ii. DALY CITY, MENLO PARK and PLEASANT HILL boroughs are similar to boroughs where existing Indian restaurants are located.

5. Discussion

The main results are presented in the previous section. We can recommend DALY CITY, MENLO PARK and PLEASANT HILL as the best boroughs to open an Indian restaurant. These boroughs have a lot other restaurants, cafes, and shopping malls what can be considered as ‘a center of attraction’. At the same time these boroughs don’t have direct competitors (Indian restaurants), which is important for business development.

We should avoid boroughs such as FAIRFIELD, AMERICAN CANYON, BENICIA, LAGUNITAS, FAIRFAX, EL CERRITO, MILL VALLEY, SAUSALITO, EMERYVILLE, SAN RAMON, SAN LORENZO, UNION CITY, BURLINGAME, NEWARK, MILPITAS, SONOMA, NOVATO, CONCORD, WALNUT CREEK, LAFAYETTE, ALBANY, CASTRO

VALLEY, PLEASANTON, HAYWARD, SAN BRUNO, MILLBRAE, FREMONT, REDWOOD CITY, SAN CARLOS, SAN JOSE, SUNNYVALE, MOUNTAIN VIEW etc. to run an Indian restaurant because they have already Indian restaurants and these restaurants are successfully running the business.

6. Conclusion

The main aim of the project is to identify the best borough in San Francisco Bay Area to start an Indian restaurant business. In order to find the best borough, we use criteria's such as (i) "a center of attraction" for dining, (ii) minimum Indian restaurants, and (iii) k-means clustering to find similar boroughs to the boroughs where Indian restaurants are located. As a result, we can recommend DALY CITY, MENLO PARK and PLEASANT HILL are the best boroughs to open an Indian restaurant. For the further improvement of the project, it is a good idea to consider a large radius value i.e. larger than 1000 m and collect more information like income, population, rent payments, crime level, etc. It can help to cover some more good places within the boroughs and predict more accurate results.