Coursera

IBM Data Science Professional Certificate

Applied Data Science Capstone

Analysis of San Francisco Neighborhoods Using Python

Steven Young

March 17, 2021

Table of Contents

1 Introduction

1.1 Background

San Francisco is one of the United States' most well known cities. It is the cultural, commercial, and financial center of Northern California and is famous for its iconic Golden Gate Bridge and cable cars. As a popular tourist destination and the headquarters for numerous major international companies and banks, the city is an important global center that attracts people from far and wide to make a living there.

Before making the move to start a business, it is important to familiarize oneself with the areas and neighborhoods of a city as diverse as San Francisco. Prospective business owners must first understand the best location where their business will be successful. The neighborhood a small business is in will most certainly impact the number of customers it will have. In terms of location, three of the most important factors that will influence the success of a business are:

1. Finding the location that is best suited for the business's intended target market.
2. Understanding the location's area traffic.
3. Finding a location that provides easy accessibility and visibility.

Of course, there are various other important factors to consider, such as the cost of rent or the crime index of a location. However, this report will focus on the three factors that were listed above to provide some preliminary insight on where one should start a new business in San Francisco. It should also be noted that although this study is focused on the city of San Francisco, the same ideas and approach can be applied to other cities as well.

1.2 Business Problem: Where should JB open an Indian-Italian fusion cuisine restaurant in San Francisco?

A college friend of mine, JB, expressed interest in opening a new Indian-Italian fusion cuisine restaurant in San Francisco. She is drawn to the city by its lively multicultural atmosphere and wishes to contribute to this melting pot. However, since there are so many neighborhoods in San Francisco, each with their own distinct area traffic and demographics of people, she is puzzled by where to start looking first. As such, JB is interested in learning more about each neighborhood and has stated several facts about her intended restaurant to help with the analysis. Her restaurant:

- Should appeal to everyone (locals and tourists alike).
- Is unique and stands out.
- Will primarily serve lunch and dinner.
- Will be moderately priced (∼$$$).

This report will use data science methods to identify the most promising neighborhoods for JB to start exploring for her new restaurant. The study will consider JB's intended market criteria based on the three location factors as listed in the Background section.

2 Data

2.1 Description of the Data

The following data was used to study the neighborhoods of San Francisco:

1. **San Francisco GeoJSON:** Contains the neighborhoods that exist in San Francisco as well as their respective latitude and longitude coordinates. This data was sourced from: https://data.sfgov.org/Geographic-Locations-and-Boundaries/SF-Find-Neighborhoods/pty2-tcw4.

2. **Venues in San Francisco:** A collection of all the venues in San Francisco, named *venues_df*. The dataset was pulled using the Foursquare API and includes each venue's name, what neighborhood they're in, their respective latitude and longitude coordinates, and their respective venue category.

2.2 How the Data Will Be Used to Solve the Business Problem

Exploratory data analysis will be performed to understand the two datasets in order to solve the business problem.

The San Francisco GeoJSON data will be used to show the neighborhood shapes. Each neighborhood's "coordinates" of the geometry object is composed of a multipolygon which is a multidimensional array of positions. Modifications were performed on the GeoJSON data to transform it into a polygon format that is easier to use for further analysis. This step was done before loading the data into the Jupyter notebook.

The Venues in San Francisco data will be used for finding all of the commerce venues associated with each neighborhood. As we are trying to understand each neighborhood's area traffic and potential to be a suitable market location that satisfies JB's criteria, the commerce venues information is used to get an impression of the kinds of people that frequent each neighborhood.

In the analysis, the Python library *Folium* will be used to visualize San Francisco and the locations of each venue of interest. One hot encoding will be used to divide the commerce venues into: Business Offices, Colleges, Hotels, Public Areas, Restaurants, Shopping Areas, and Social Places. This subdivision will allow us to analyze what commerce types each neighborhood is known for so we can better determine the one that is best suited for JB's restaurant. Finally, K-means clustering will be used on the venues dataset to see what venue categories can be used to distinguish each neighborhood cluster.

3 Methodology

3.1 Data Preparation - San Francisco GeoJSON

To begin, we need the coordinates and spatial geometry of the neighborhoods in San Francisco. This information is contained within the San Francisco GeoJSON data. Figure 1 is a quick snapshot of what the raw GeoJSON data looks like after it was loaded.

Notice that all of the relevant data is in the "features" key, which is basically a list of the neighborhoods. After transforming this data of nested Python dictionaries into a *pandas* dataframe and calculating each neighborhood geometry's respective centroid coordinates (which is their average latitude and longitude), we get the following as shown in Figure 2.

```
{'type': 'FeatureCollection',
 'features': [{'type': 'Feature',
   'properties': {'link': 'http://en.wikipedia.org/wiki/Sea_Cliff,_San_Francisco,_California',
    'name': 'Seacliff'},
   'geometry': {'type': 'Polygon',
    'coordinates': [[[-122.49345526799993, 37.78351817100008],
      [-122.49372649999992, 37.78724665100009],
      [-122.49358666699993, 37.78731259500006],
      [-122.49360569399994, 37.78752774600008],
      [-122.49283007399993, 37.787882585000034],
      [-122.4927566799999, 37.78773917700005],
      [-122.48982906399993, 37.789482184000065],
      [-122.48899105699991, 37.78928318700008],
      [-122.4878640209999, 37.78958817900008],
      [-122.48736904899994, 37.78942984100007],
      [-122.48598032899991, 37.79080370600008],
      [-122.48581537399991, 37.79070384600004],
      [-122.48557750799989, 37.790559847000054],
      [-122.4850531269999, 37.79036813300007],
      [-122.4842660519999, 37.789411709000035],
      [-122.48407706799992, 37.78939909400009],
      [-122.4838230019999, 37.78928250300004],
      [-122.48370738599994, 37.788776950000056],
      [-122.4839269609999, 37.788315201000046],
      [-122.4839504329999, 37.78802775100007],
      [-122.48414271299993, 37.78777522900009],
      [-122.4841506649999, 37.787554653000086],
      [-122.48463982999994, 37.78753212700008],
      [-122.48464285299991, 37.787378785000044],
      [-122.48431022499994, 37.78735203400004],
      [-122.4841736059999, 37.78731086500005],
      [-122.48407980499991, 37.78579452900004],
      [-122.48728636499993, 37.78564884000008],
      [-122.48715071499993, 37.783785427000055],
      [-122.49345526799993, 37.78351817100008]]]}},
```


Figure 1: Structure of the raw GeoJSON data of San Francisco's neighborhoods.

	Neighborhood	Geometry	Centroid	Longitude	Latitude
0	Seacliff	[[[-122.49345526799993, 37.78351817100008], [-...	(-122.48727411196657, 37.78794365720007)	-122.487274	37.787944
1	Lake Street	[[[-122.48715071499993, 37.783785427000055], [...	(-122.48056858607683, 37.78615137738468)	-122.480569	37.786151
2	Presidio National Park	[[[-122.47758017099994, 37.81099311300005], [-...	(-122.46797365011281, 37.79926798262909)	-122.467974	37.799268
3	Presidio Terrace	[[[-122.47241052999993, 37.787346539000055], [...	(-122.46795748666659, 37.78692981188894)	-122.467957	37.786930
4	Inner Richmond	[[[-122.47262578999994, 37.786314806000064], [...	(-122.46875409127264, 37.7811122254546)	-122.468754	37.781112

Figure 2: The raw GeoJSON data converted into a pandas dataframe with each neighborhood's centroid coordinates.

3.2 Data Collection - Coordinates of San Francisco

Before we can start showing each neighborhood, the latitude and longitude coordinates of San Francisco needs to be obtained so that the Folium map can be centered over the city. Once we have a base map of San Francisco, we can then overlay the neighborhoods on top of the map as markers or a choropleth map. As such, the *geopy* library was used to get the latitude and longitude values of San Francisco which are (37.7790262, -122.4199061). Now, as an example, we can add circular markers on top of the San Francisco Folium map where the markers are the calculated centroid coordinates of each neighborhood.

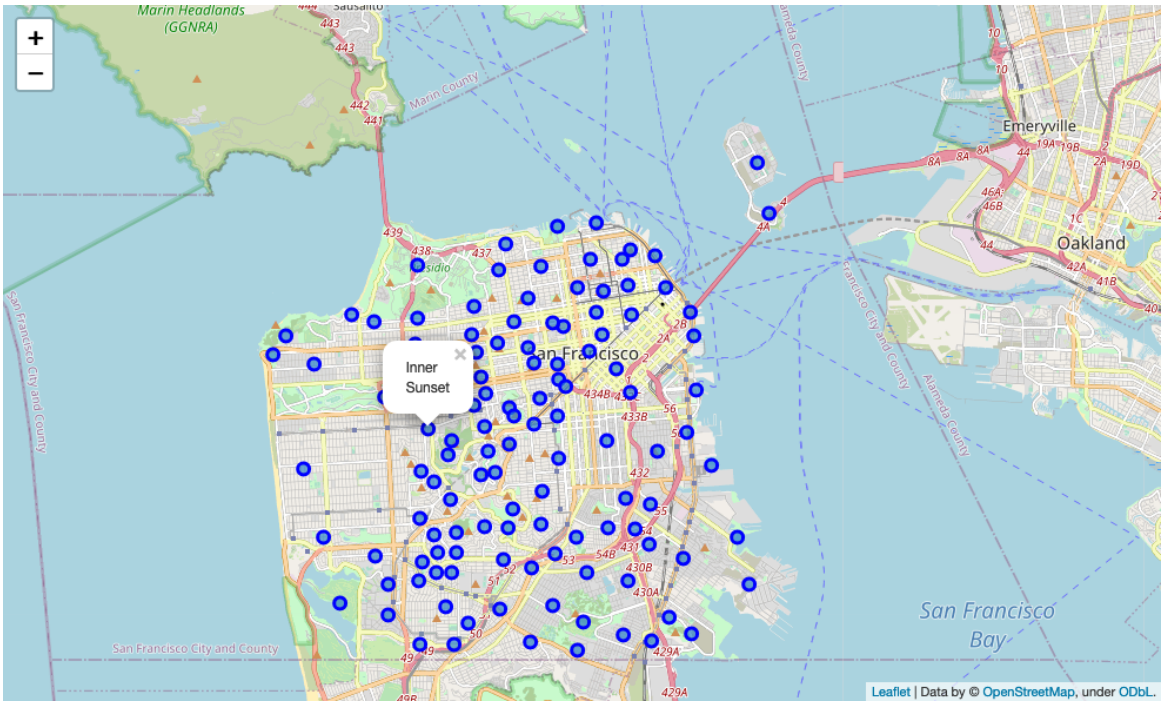


Figure 3: Map of the city of San Francisco with each neighborhood represented by a blue circular marker. The location of each marker is the centroid coordinates of each neighborhood.

3.3 Data Collection - Venues in San Francisco

Foursquare is a tech company that provides massive datasets of location data. The company uses crowd-sourcing to build their data where they have people use their app to construct their datasets, add venues, and complete any missing information. It is currently the most comprehensive repository of location data available to the public and its data is accurate enough that popular services like Apple Maps, Uber, Snapchat, Twitter, and many others are currently using it to power their own location data. It is also used by over 100,000 developers with this number continuing to grow [1].

The second dataset, containing all of the venues in San Francisco, was acquired from the Foursquare API. After pulling all of the different venues data, it is then stored in the dataframe called *venues_df*. Below is a small section of what the Foursquare venues data looks like.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Seacliff	37.787944	-122.487274	Baker Beach	37.793355	-122.483740	Beach
1	Seacliff	37.787944	-122.487274	China Beach	37.788090	-122.491186	Beach
2	Seacliff	37.787944	-122.487274	UC San Francisco	37.789087	-122.487971	College Administrative Building
3	Seacliff	37.787944	-122.487274	Lands End Coastal Trail	37.784183	-122.509022	Trail
4	Seacliff	37.787944	-122.487274	Lincoln Park Stairs	37.783496	-122.493597	Trail
5	Seacliff	37.787944	-122.487274	Pearl	37.783839	-122.483083	Restaurant
6	Seacliff	37.787944	-122.487274	MUNI Bus Stop #13839	37.783956	-122.484822	Bus Stop
7	Seacliff	37.787944	-122.487274	Muni 13838	37.783792	-122.485390	Bus Station
8	Seacliff	37.787944	-122.487274	Lobos Creek	37.786981	-122.485191	Other Great Outdoors

Figure 4: The Venues in San Francisco data, obtained from Foursquare, where every venue in each of San Francisco's neighborhoods are listed including their respective latitude and longitude coordinates and their venue category. This snippet shows some of the venues in the Seacliff neighborhood.

3.4 Exploratory Data Analysis

There are 529 unique venue categories in the *venues_df* dataframe. However, not every category is relevant to our study (like trails or yoga studios) so only those that are relevant to satisfying the 3 location factors and JB's criteria are used for further analysis.

3.4.1 Get the Relevant Venues

In order to decide which venues will be useful, we refer back to JB's list of criteria for her restaurant. She stated that her restaurant would cater to both locals and tourists. From this valuable information, we can narrow down our venues category list to those venues that are frequented by both locals and tourists alike.

First, we know that tourists must reside inside hotels, motels, or resorts so all of those venues will be grouped together into a Hotels category. Next, we know that both locals and tourists love to eat at restaurants, go out shopping, mingle in social places like bars, clubs, and other nightlife venues, and visit public spaces like parks, beaches, or San Francisco's piers. Therefore, we will find all of the venues that fall within the Restaurants, Shopping Areas, Social Places, and Public Areas categories. But now, what are the specific venues that only locals usually frequent? Of course, locals would go to any venue within their hometown but we know that office buildings and other business places are most usually frequented by locals who work in the city. We also can't forget about the university students who go to school and live inside the city. As such, we will include all venues that are related to both of those into a Business Offices and Colleges group, respectively. Great, now that we have a list of these "commerce venues", a dataframe called *commerce_venues_df* is created that contains all of the venues we just determined which will be useful in our analysis.

We've isolated data that satisfies one of JB's criteria where she wants to serve both locals and tourists. Next, let's move on to meeting her second requirement, the one where she wants her restaurant to be unique and stand out among other restaurants in the area. The first logical step is to take a look at all of the different restaurants and eateries that are in San Francisco. Then, because we know JB's restaurant will serve Indian-Italian fusion cuisine, we will study where each Indian and Italian restaurant in the city is located and eventually try to pick neighborhoods that do not have one of those restaurant types to maximize her restaurant's uniqueness in the area.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Seacliff	37.787944	-122.487274	Pearl	37.783839	-122.483083	Restaurant
1	Seacliff	37.787944	-122.487274	Pizzetta 211	37.783694	-122.482879	Pizza Place
2	Seacliff	37.787944	-122.487274	Little Sushi Bar	37.783873	-122.482761	Sushi Restaurant
3	Seacliff	37.787944	-122.487274	Ocean Side Mexican Grill	37.784411	-122.480892	Mexican Restaurant
4	Seacliff	37.787944	-122.487274	Angelina's Cafe	37.784074	-122.481809	Café
5	Lake Street	37.786151	-122.480569	Pearl	37.783839	-122.483083	Restaurant
6	Lake Street	37.786151	-122.480569	Bazaar Cafe	37.784050	-122.481320	Café
7	Lake Street	37.786151	-122.480569	Fiorella	37.781887	-122.484510	Italian Restaurant
8	Lake Street	37.786151	-122.480569	Angelina's Cafe	37.784074	-122.481809	Café
9	Lake Street	37.786151	-122.480569	Pizzetta 211	37.783694	-122.482879	Pizza Place

Figure 5: Dataframe, called *restaurants_df*, consisting of all the restaurants and other food-service venues in San Francisco. This dataset was pulled from the larger *commerce_venues_df* dataset.

3.4.2 Analyze Each Neighborhood

To give JB a better understanding of what each neighborhood is known for, we can look at the number of different commerce venues within each one. This can be done by applying One

Hot Encoding to the list of venue categories such that a new binary variable is added for each unique venue category per row, as shown in Figure 6, where the venue categories that exist in the neighborhood will be denoted by 1 while those that do not exist are denoted by 0.

	Neighborhood	Accessories Store	African Restaurant	American Restaurant	Antique Shop	Argentinian Restaurant	Arts & Crafts Store	Asian Restaurant	Automotive Shop	Baby Store	Bagel Shop	Bakery	Bar	Beach	Bed & Breakfast
0	Seacliff	0	0	0	0	0	0	0	0	0	0	0	0	1	0
1	Seacliff	0	0	0	0	0	0	0	0	0	0	0	0	1	0
2	Seacliff	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Seacliff	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Seacliff	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 6: Using one hot encoding, the venue categories that exist in the neighborhood are represented by 1's while those that do not exist are represented by 0's. This snippet shows that Seacliff has at least 2 beaches.

Next, the rows can be grouped by neighborhood and the frequency of occurrence of each category can be analyzed.

	Neighborhood	Accessories Store	African Restaurant	American Restaurant	Antique Shop	Argentinian Restaurant	Arts & Crafts Store	Asian Restaurant	Automotive Shop	Baby Store	Bagel Shop	Bakery	Bar	Beach	Bed & Breakfast
0	Alamo Square	0	0	0	1	0	0	0	0	0	0	0	0	0	0
1	Anza Vista	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Apparel City	0	0	1	1	0	1	0	9	0	0	0	0	0	0
3	Aquatic Park / Ft. Mason	0	0	0	0	0	2	0	0	0	0	0	0	2	0
4	Ashbury Heights	0	0	0	1	0	0	0	0	0	0	0	0	0	1
5	Balboa Terrace	0	0	1	1	0	1	0	0	0	0	1	1	0	0

Figure 7: Example view of dataset where the sum of each venue category is shown per neighborhood.

Sort all of the 247 different venue categories into their designated Business Offices, Colleges, Hotels, Public Areas, Restaurants, Shopping Areas, and Social Places main groups.

	Neighborhood	Business Offices	Colleges	Hotels	Public Areas	Restaurants	Shopping Areas	Social Places	Total
0	Alamo Square	2	1	1	2	4	2	1	13
1	Anza Vista	8	0	1	3	7	8	2	29
2	Apparel City	6	0	0	1	3	12	4	26
3	Aquatic Park / Ft. Mason	2	1	0	9	13	2	1	28
4	Ashbury Heights	2	0	1	4	4	6	3	20
5	Balboa Terrace	23	2	1	3	6	6	1	42
6	Bayview	2	2	0	3	23	6	3	39
7	Bernal Heights	1	1	0	1	23	12	2	40
8	Bret Harte	3	1	0	2	8	2	3	19
9	Buena Vista	3	0	0	9	3	0	2	17

Figure 8: This dataframe can easily tell us what each neighborhood is known for. For example, Balboa Terrace has by far more business offices than any other commerce venue groupings while Bayview has a majority in restaurants.

Finally, a choropleth map showing the distribution of Indian and Italian restaurants within San Francisco along with each neighborhood's number of commerce venues will now be visualized. The color for the two types of restaurants are set as follows:

Restaurant	Color
Indian	green
Italian	orange

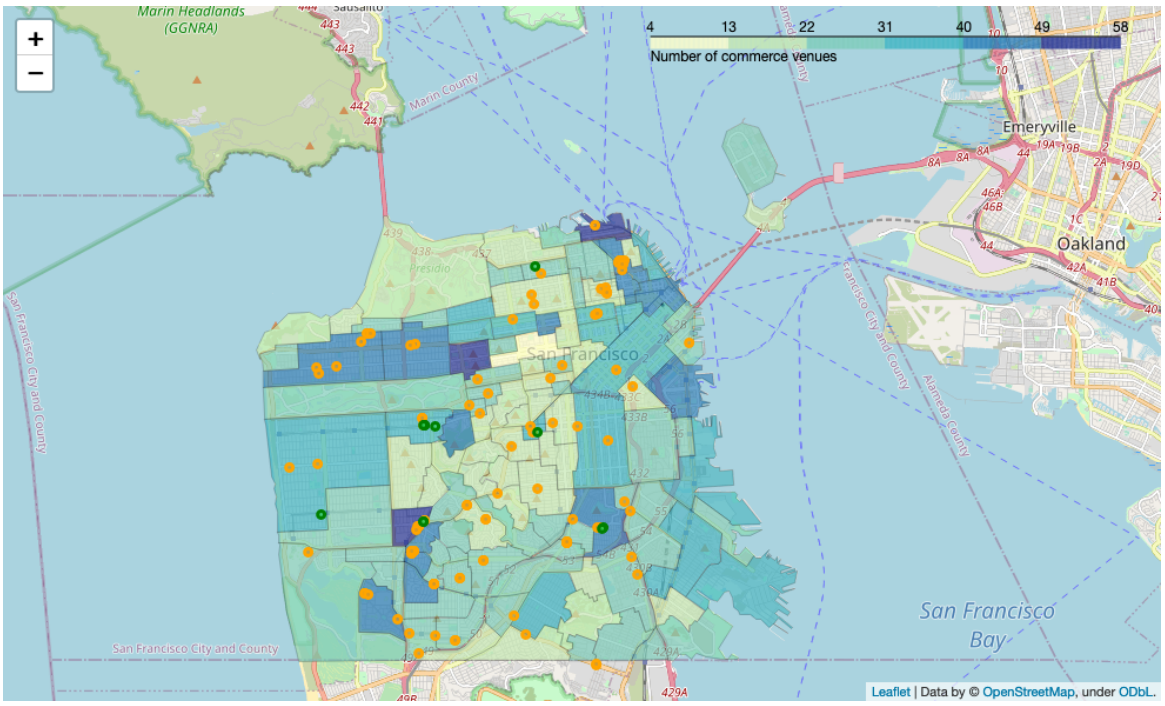


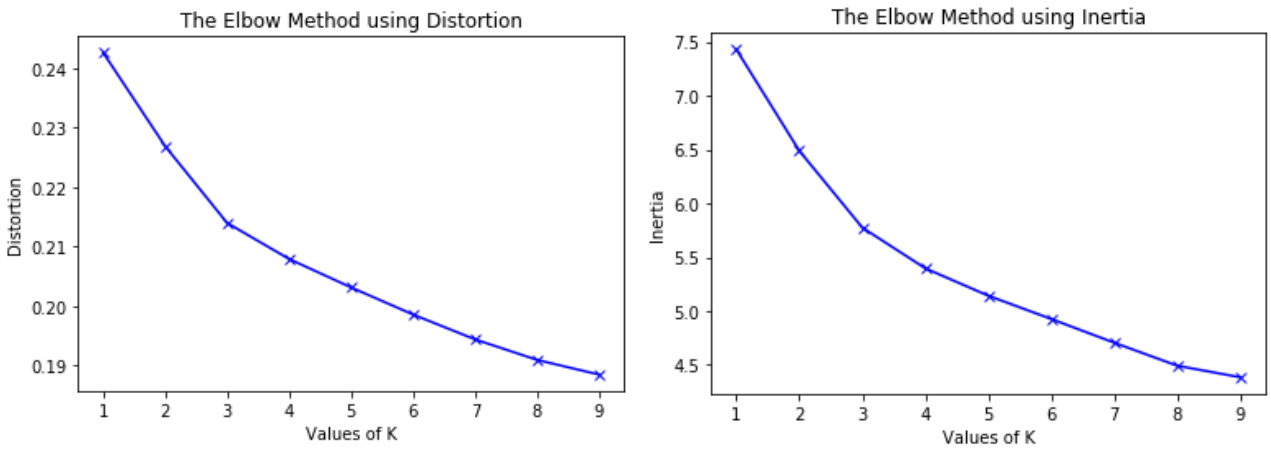
Figure 9: From the map, we can see where each Indian and Italian restaurant is located along with the density of commerce venues in each neighborhood.

3.5 Cluster the Neighborhoods

We can partition the neighborhoods into different groups that have similar common venues. An algorithm that can be used for segmenting the neighborhoods is K-Means Clustering. It is a popular unsupervised machine learning algorithm that will help to group the neighborhoods based on their similarity to each other. The algorithm divides the data into k non-overlapping subsets or clusters without any cluster internal structure or labels. Objects within a cluster are very similar to one another while objects across different clusters are very different or dissimilar to each other [2].

3.5.1 K-Means Clustering

K-means, will be used to find clusters of neighborhoods based on similar most common venues. But before we can begin to study the neighborhood clusters, the optimal number of clusters should be found first, using the Elbow Method. We will look at two different metrics for the elbow method: Distortion and Inertia. Distortion is the average of the squared distances from the cluster centers of the respective clusters and, typically, the Euclidean distance is used. Inertia is the sum of squared distances of samples to their closest cluster center. To determine the optimal number of clusters, we have to select the value of k at the "elbow" or the point after which the distortion/inertia start decreasing in a linear fashion [5].



(a) Number of clusters vs. Distortion

(b) Number of clusters vs. Inertia

Figure 10: Using the Elbow Method to determine the optimal number of clusters to use for the K-Means analysis. The optimal k seems to be 3 for both distortion and inertia.

Although somewhat hard to tell, we can estimate that for both distortion and inertia that $k = 3$ clusters looks to be the elbow point so the optimal number of clusters for the data is 3. Now, we can run the K-means to cluster the neighborhoods into 3 clusters and then visualize the result using a choropleth map.

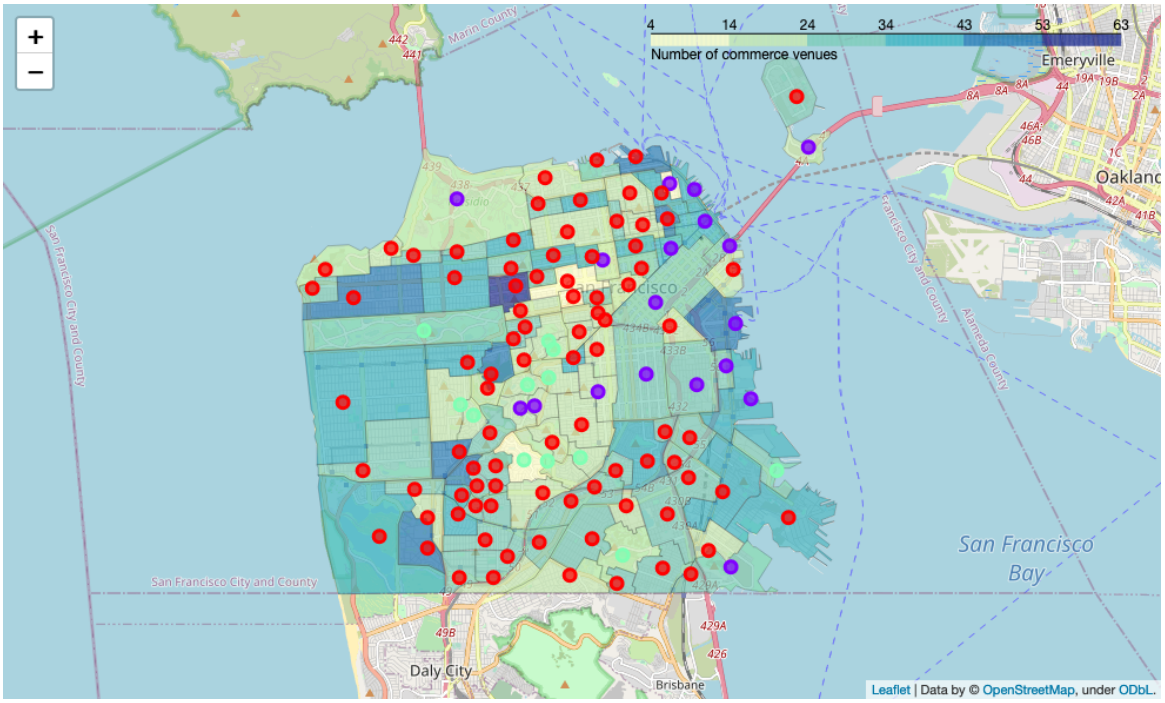


Figure 11: hello

The color code for the clusters are:

Cluster	Color
0	red
1	purple
2	green

3.5.2 Examine the Clusters

Now, we can examine each cluster to determine the discriminating venue categories that distinguish each cluster.

	Neighborhood	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
0	Seacliff	0	Beach	Grocery Store	Mexican Restaurant	College Academic Building	Resort	Sushi Restaurant	Karaoke Bar	Salon / Barbershop	Bike Shop
1	Lake Street	0	Automotive Shop	Dentist's Office	Café	Grocery Store	Office	Mexican Restaurant	Pizza Place	Sushi Restaurant	Resort
3	Presidio Terrace	0	Office	Salon / Barbershop	Park	Doctor's Office	Coffee Shop	Café	Real Estate Office	Fast Food Restaurant	Sporting Goods Shop
4	Inner Richmond	0	Chinese Restaurant	Dentist's Office	Automotive Shop	Bar	Asian Restaurant	Salon / Barbershop	Korean Restaurant	Pizza Place	Food & Drink Shop
5	Sutro Heights	0	Beach	American Restaurant	Dentist's Office	Salon / Barbershop	Café	Parking	Park	National Park	Doctor's Office

Figure 12: Cluster 0

	Neighborhood	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
2	Presidio National Park	1	Office	Park	Beach	American Restaurant	Women's Store	Food Truck	National Park	Public Art	Café
17	Telegraph Hill	1	Office	Parking	Business Service	Fabric Shop	Park	Doctor's Office	Comfort Food Restaurant	Public Art	Seafood Restaurant
18	Downtown / Union Square	1	Office	Clothing Store	Salon / Barbershop	Tailor Shop	Women's Store	Doctor's Office	Food Truck	Jewelry Store	Shoe Store
29	Rincon Hill	1	Office	Harbor / Marina	Clothing Store	Jewelry Store	Sandwich Place	Burger Joint	Dive Bar	Parking	Corporate Coffee Shop
31	South of Market	1	Office	Nightclub	Mexican Restaurant	Cocktail Bar	Café	Karaoke Bar	Furniture / Home Store	Cosmetics Shop	Miscellaneous Shop

Figure 13: Cluster 1

	Neighborhood	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
8	Golden Gate Park	2	Park	Mexican Restaurant	Automotive Shop	Office	Bar	Chinese Restaurant	Shoe Store	Doctor's Office	Salon / Barbershop
43	Golden Gate Heights	2	Park	Office	Korean Restaurant	Bakery	Frame Store	Men's Store	Food	Thai Restaurant	Doctor's Office
44	Forest Hill	2	Park	Office	Food	Japanese Restaurant	Bakery	Doctor's Office	Burger Joint	Sushi Restaurant	Frame Store
46	Clarendon Heights	2	Park	Office	Real Estate Office	Coffee Shop	Breakfast Spot	Food Truck	Tailor Shop	Caribbean Restaurant	Italian Restaurant
58	Fairmount	2	Park	Office	Bed & Breakfast	Candy Store	Public Art	Cosmetics Shop	College Lab	Music Store	Dive Bar

Figure 14: Cluster 2

After examining each cluster:

- Cluster 0 looks to be all the neighborhoods with varying venues as their first and second most common venues.
- Cluster 1 neighborhoods have offices as their most common venues and varying venues as their second most common venues.
- Cluster 2 neighborhoods have parks as their most common venues and either offices or some kind of edible goods retailer/service as their second most common venues.

4 Results & Discussion

The results of the San Francisco neighborhood analysis will be discussed in this section and a list of recommended neighborhoods will be provided to answer the three location factors and satisfy JB's criteria for her restaurant.

As a recap, the three location factors that will influence the success of a business are:

1. Whether or not the location is best suited for the business's target market.
2. The location's area traffic.
3. Whether or not the location offers easy accessibilty and visibility.

4.1 Satisfying the Three Location Factors

4.1.1 Should Appeal to Everyone

JB's target customers are both locals and tourists. Throughout the analysis, we have been systematically trying to satisfy the first factor in that we were slowly narrowing down the list of neighborhoods that have both locals and tourists. So now, let's finally identify the neighborhoods that have the greatest diversity of commerce venues.

	Neighborhood	Business Offices	Colleges	Hotels	Public Areas	Restaurants	Shopping Areas	Social Places	Total
0	Alamo Square	2	1	1	2	4	2	1	13
20	Cow Hollow	3	1	7	2	5	5	3	26
50	Laurel Heights / Jordan Park	12	1	1	1	6	8	2	31
55	Lower Nob Hill	12	1	3	3	11	3	1	34
58	McLaren Park	1	2	1	4	6	2	2	18
66	Mission Dolores	3	2	1	2	4	4	1	17
114	Westwood Highlands	6	2	1	2	9	2	2	24

Figure 15: The list of neighborhoods with at least one of each commerce venue group.

And here we are. The list of neighborhoods with every type of commerce venue, as shown in Figure 15, will help us to fulfill all three location factors. The existence of business offices, colleges, and hotels in the neighborhood will guarantee that there are always hungry locals and tourists in the area searching for food while the public areas, restaurants, shopping areas, and social places ensure the neighborhood is a lively area to be in. There will be enough foot and car traffic in those neighborhoods that JB will not have to worry those districts will ever be deserted.

According to the K-means cluster analysis, every neighborhood in the list, except for McLaren Park, are in cluster 0 which means they have a wide variety of venues. However, even though McLaren Park is in cluster 2, it also features offices, college class rooms, night-clubs, and various ethnic restaurants as some of its top ten most common venues. This further reinforces the three location factors. Those neighborhoods have enough venues of different kinds to attract locals and tourists to visit them, thereby increasing the people traffic. Having a sizable population of locals and tourists frequenting the area would provide higher probabilities for JB's restaurant to be noticed because there is good chance that her restaurant will be seen by passersby (assuming that it won't be established in a dark alleyway).

The caveat to this list, however, is that there are other lively and bustling neighborhoods that could possibly be better for JB's restaurant which are not included. For example, Excelsior is also one of the most ethnically diverse districts in San Francisco. The diversity of people

there, in terms of cultures, would likely provide JB with a good pool of hungry and adventurous customers. But, as this analysis simply acts as a jumping point for JB and also satisfies the three location factors, we will conclude with this list and allow JB to decide.

4.1.2 Is Unique and Stands Out

To attract customers, JB wants her restaurant to be unique in the area that it is in. As such, we should suggest to her neighborhoods that do not have either Indian or Italian restaurants within them to maximize JB's restaurant's uniqueness.

	Neighborhood	Business Offices	Colleges	Hotels	Public Areas	Restaurants	Shopping Areas	Social Places	Total
0	Alamo Square	2	1	1	2	4	2	1	13
1	Cow Hollow	3	1	7	2	5	5	3	26
2	Laurel Heights / Jordan Park	12	1	1	1	6	8	2	31
4	McLaren Park	1	2	1	4	6	2	2	18

Figure 16: Neighborhoods without Indian or Italian restaurants.

Now, from Figure 16, we see all of the previously recommended neighborhoods that do not have either an Indian or Italian restaurant. This list of four neighborhoods will be great for JB to start exploring because they are bustling neighborhoods (as described in section 4.1.1) and the lack of Indian and Italian restaurants will maximize JB's restaurant's uniqueness in the area.

4.1.3 Will Primarily Serve Lunch and Dinner

Although our data does not have information that can explicitly answer this criterion, we can deduce from the list of recommended neighborhoods in section 4.1.2 that those neighborhoods are busy throughout the day. This is because all of the commerce venues are obviously openned at least during mornings and afternoons. There would be employees from the business offices that would want to order lunch on weekdays while the students from the colleges and anyone else who are living, shopping, or socializing in the area will no doubt be searching for food throughout the day, on any day.

4.1.4 Will Be Moderately Priced

Once again, our data does not have any information that can directly address this fourth criterion but we can make educated assumptions that people who live in, work in, and/or frequent the recommended neighborhoods will most likely have enough money to eat at JB's restaurant. For example, according to Wikipedia, the demographics of Alamo Square include many young people and upper-middle-class homeowners, Laurel Heights is an upper-middle-class suburban neighborhood, and Cow Hollow is a generally affluent neighborhood. Therefore, it wouldn't be unusual for JB to open a moderately priced restaurant in the list of recommended neighborhoods.

5 Conclusion

The purpose of this project was to identify several neighborhoods in San Francisco for JB to start exploring for her Indian-Italian fusion cuisine restaurant. Only the three location factors as described in the Background section of this notebook were used to determine the list of recommended neighborhoods. In the analysis, we systematically checked off each of JB's

criteria while also satisfying the three location factors. For example, we found the locations of every Indian and Italian restaurant in San Francisco to search for neighborhoods that can maximize JB's restaurant's uniqueness. Each commerce venue was grouped into the larger commerce venue categories of: Business Offices, Colleges, Hotels, Public Areas, Restaurants, Shopping Areas, and Social Places. This subdivision allowed us to analyze what commerce types each neighborhood is generally known for. Finally, we narrowed the list of neighborhoods to the four most promising ones for JB which include Alamo Square, Cow Hollow, Laurel Heights, and McLaren Park. Although these four neighborhoods are the recommended ones, it should be noted that there are other neighborhoods that may just be as fitting for JB to look into. Ultimately, this analysis will act as a starting point for JB and it will be up to her to make the final decision.

References

[1] Aklson, Alex. (2019, December 23). *Introduction to Foursquare.* [Lecture video]. Coursera. https://www.coursera.org/learn/applied-data-science-capstone/lecture/rD4tX/introduction-to-foursquare

[2] Aklson, Alex. (2019, December 23). *k-means Clustering.* [Lecture video]. Coursera. https://www.coursera.org/learn/applied-data-science-capstone/lecture/jpedY/k-means-clustering

[3] DataSF. (2019). *SF Find Neighborhoods.* [Data set]. https://data.sfgov.org/Geographic-Locations-and-Boundaries/SF-Find-Neighborhoods/pty2-tcw4

[4] Foursquare. (2019). [Data set]. https://foursquare.com/

[5] Gupta, Alind. (2019, June 6). *Elbow Method for optimal value of k in KMeans.* GeeksforGeeks. https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/

[6] San Francisco Panorama w the Golden Gate Bridge. [Online image]. (2020). CheapTickets. https://www.cheaptickets.com/blog/2020/02/cheap-san-francisco-hotels-golden-views-of-san-francisco-by-seaplane/san-francisco-panorama-w-the-golden-gate-bridge/

[7] *Why the Location of Your Restaurant is So Important.* TigerChef. https://www.tigerchef.com/why-the-location-of-your-business-is.html