IBM/Coursera Data Science Capstone Project

Opening an Indian Restaurant in San Francisco

by Yash Kamat

September 2020

Introduction

The city of San Francisco is one of the biggest and most diverse cities in the United States. It is a melting pot of people from different backgrounds, nationalities and cultures. The result of this cosmopolitan nature of the city is the wide variety of cuisines available in the restaurants of SF. From popular American and European cuisines to the exotic foods of Asia and Africa, San Francisco boasts a great variety of options to dine.

Of these, Indian food is one of the most popular choices, largely due to the growing South Asian population in the city. In fact, the San Francisco-Oakland-Hayward area contained the fourth-largest population of Asian Indians in United States, according to the 2010 Census data (ProximityOne, n.d.). The growing Indian population in the San Francisco metropolitan area means that the demand for Indian food is increasing, which is a great opportunity for entrepreneurs and restauranteurs to open an Indian restaurant in the city.

Business Problem

The main question of my Capstone Project is quite simple - **if someone were to open an Indian restaurant in the city of San Francisco, where would should they open it?** In order for the restaurant to be successful, there have to be enough customers, while also being in a place without too many competitors in the neighborhood.

The primary audience of this study is meant to be entrepreneurs and restaurateurs in San Francisco. If they are new to the market, they would be interested in this study to scout potential locations to start their venture in. Or, if they already have existing establishments, to seek new locations to expand into.

Data

Data Requirements

To explore the business problem, the following information is required:

- List of neighborhoods in San Francisco. This defines the scope of the project which is confined to the city of San Francisco, located in Silicon Valley in California on the west coast of the USA.
- Latitude and longitude of the neighborhoods of San Francisco. This is to geolocate points of interest (especially Indian restaurant) and also to plot the map in the Jupyter notebook.
- Geolocation and venue data, specifically for Indian restaurants in San Francisco. This data will then be used for clustering and its inferences.

Data Sources

As a starting point I will use the Wikipedia page which contains the list of neighborhoods in San Francisco https://en.wikipedia.org/wiki/Category:Neighborhoods_in_San_Francisco). This page contains a list of neighborhoods in San Francisco, with over 90 distinct neighborhoods listed. Using a Python-based web scraper, I will extract the list of neighborhoods from the Wikipedia page. Based on this list, I will then get the individual geographical coordinates of the neighborhoods using a Python package which will return the latitude and longitude for each neighborhood.

After that, I will use the Foursquare API to get the geolocation data for those neighborhoods. Foursquare has a database of 100+ million places and is used developers all around the world. While the Foursquare API provides multiple categories of venue data, I am mostly interested in the Indian restaurant category for the purpose of this business problem.

Methodology

Firstly, I retrieved a list of neighborhoods in San Francisco by using the Wikipedia link https://en.wikipedia.org/wiki/Category:Neighborhoods_in_San_Francisco. Following this, I used a Python based web scraper which used a combination of Python requests, HTML parsing and the BeautifulSoup package to extract the names of all the neighborhoods and append them into a Python list. After making a list of the neighborhoods I aimed to find the geographical coordinates to use them with the Foursquare API for clustering and further analysis. I did this using the "geocoder" library that returns a latitude and longitude for a given address. After looping through the list of neighborhoods, I retrieved a list of their respective latitude and longitude coordinates and combined this data into a pandas DataFrame. Lastly, I plotted these coordinates on a Folium map to not only check the accuracy of the geocoder data, but also to get a good visual understanding of the area I was working with.

Next, I used the Foursquare API to get the top 100 venues within a radius of 500 meters for each neighborhood's coordinates. I restricted the radius to 500 meters to avoid an overlap of venues across two neighborhoods. I did this since there are almost 100 neighborhoods in a city of 121 km^2, and some of these neighborhoods are quite close to one another. After setting my Foursquare Client ID and Secret and the radius, I passed these arguments into a URL and sent requests for individual neighborhoods through a Python loop. The API returned neighborhood specific data for venues in each neighborhood in JSON. This included a venue's name, category and coordinates. With this data, I grouped the rows of each neighborhood, analyzed the frequency of each venue category and later filtered out to get data specifically for Indian restaurants.

Finally, I performed K-Means clustering on the data. K-Means clustering algorithm works by using a set 'k' number of centroids (and clusters), and then allocates every data point to its nearest cluster. The simple, yet powerful nature of K-Means clustering makes it a perfect fit for tackling our problem at hand, as this unsupervised machine learning technique helps clustering and categorize similar data points together quite efficiently. In my analysis, I chose different values of k to see how the data responded to them. Eventually, I decided to work with k = 3 clusters for my final analysis. This categorized the neighborhoods into varying levels of frequency of Indian restaurants, which was used for further analysis and recommendations.

Results

The K-Means clustering with k = 3 centroids effectively categorized each neighborhood on the basis of the number/frequency of Indian restaurants. Each cluster showed this frequency in the following way:

1. **Cluster 0** contained a list of no Indian restaurants within a 500-meter radius of these neighborhoods, according to the Foursquare API (frequency = 0).
2. **Cluster 1** contained a list of neighborhoods with a considerable number of Indian restaurants within a 500-meter radius (frequency > 0.1).
3. **Cluster 2** contained a list of neighborhoods with a small number of Indian restaurants within a 500-meter radius (frequency = 0.1).

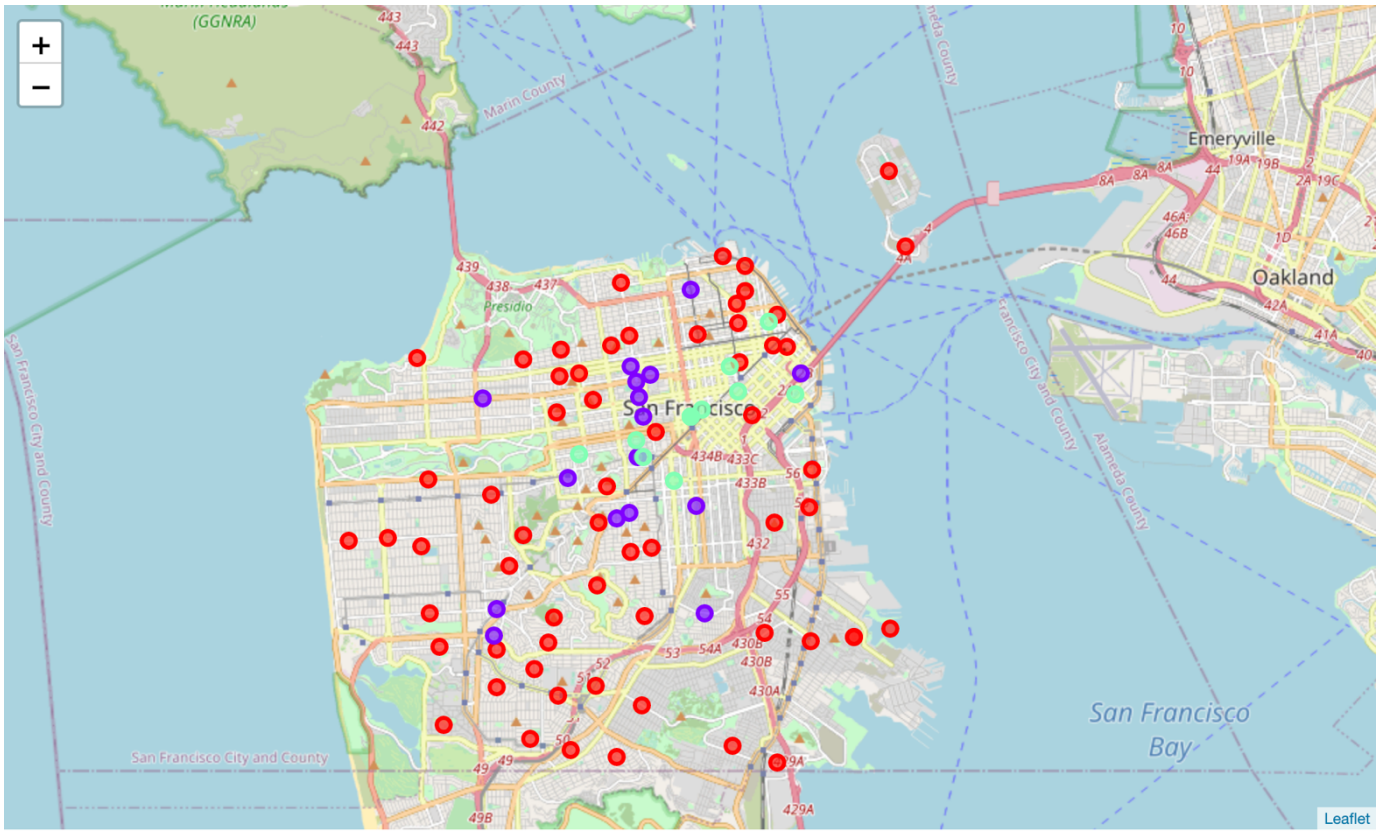

Figure 1 – Map of San Francisco with clustered neighborhoods (Cluster 0: Red, Cluster 1: Purple, Cluster 2: Mint)

NOTE: Each cluster and all the neighborhoods in them can be found in the Appendix section.

Recommendations

As the map in the Results section shows above, the majority of the neighborhoods fall under Cluster 0, which is the cluster of neighborhoods with no Indian restaurants. On the flip side, the second highest cluster is Cluster 1, which is the cluster with a considerable number of Indian restaurants (highest in SF). Cluster 2, which has the neighborhoods with a small number of Indian restaurants, thus, falls in between these two.

Based on this information, a few observations can be made. The first is that a large majority of neighborhoods do not have Indian restaurants, based on the Foursquare data. This shows that there is a large section of the city that where an interested party could be the one to open the first Indian restaurant in a neighborhood. However, while this provides opportunity, it also provides a risk, as there is no indication as to whether the restaurant will be successful. Perhaps the reason why these neighborhoods don't have Indian restaurants is because there is no demand for it from the residents of that neighborhood. The opposite is true for neighborhoods cluster 1, where there is certainly demand, but already a decent amount of competition, which could make it difficult to gain traction.

As such, neighborhoods in cluster 2 could, perhaps, be the target neighborhoods for parties interested in opening an Indian restaurant in San Francisco. The reason for this simple – the existence of Indian restaurants indicates that there is demand for the cuisine, but the low frequency suggests that it might not be too saturated and thus, perfect for a new player to enter the market.

Project Limitations and Further Scope

The primary limitation of this project is that it considers the frequency of Indian restaurants as the sole factor in deciding the which neighborhoods to open a new restaurant. This is, of course, not true in the real world where a variety of different factors come in. These could include the ethnic makeup of a neighborhood (since South Asians are more likely to order Indian food), the average age of a neighborhood (since young people tend to eat out more (Reiter, 2017)) and the proximity to other restaurants in the neighborhood, among others. By getting this data and adding these dimensions to the dataframe, the clustering algorithm will be able to provide more detailed clustered and will yield more refined results for further analysis.

Another potential issue could have been the data retrieved from the Foursquare API. The radius, for this study, had been set to 500 meters to prevent overlapping data in neighborhoods that are close to one another. However, for larger neighborhoods, this prohibited the API from returning data about the entire neighborhood, thus, preventing us from seeing the full picture. This could be tackled by increasing the search radius and also by grouping certain close-by neighborhoods together.

Conclusion

This project aimed to tackle one business problem – "*if someone were to open an Indian restaurant in the city of San Francisco, where would should they open it?*". This was done by collecting a list of neighborhoods in the city, extracting their coordinates, retrieving location-specific data and clustering the neighborhoods into one of three clusters on the basis of the frequency of Indian restaurants in each neighborhood. Based on the clustering, the primary recommendation made was that neighborhoods in cluster 2 provided the best locations for to the relevant stakeholders to open an Indian restaurant as they indicated demand for the product without there being much competition.

References

ProximityOne. (n.d.). *Asian Population Demographics | Largest Asian Growth*. Retrieved September 13, 2020, from ProximityOne: http://proximityone.com/asian_demographics.htm

Reiter, A. (2017, January). *Millennials Eat Out More — and Spend More When They Do — Than Non-Millennials*. Retrieved September 21, 2020, from Food Network: https://www.foodnetwork.com/fn-dish/news/2015/06/millennials-eat-out-more-and-spend-more-when-they-do-than-non-millennials

Appendix

Cluster 0:

Neighborhood	Frequency of Indian Restaurants	Cluster #
Lone Mountain	0	0
North Beach	0	0
Noe Valley	0	0
Nob Hill	0	0
Mount Davidson	0	0
Mission Bay	0	0
Mid-Market	0	0
Merced Manor	0	0
Marina District	0	0
Manilatown	0	0
Westwood Park	0	0
Little Hollywood	0	0
Lincoln Manor	0	0
Laurel Heights	0	0
Jackson Square	0	0
Irish Hill	0	0
Oceanview	0	0
Outer Mission	0	0
Outer Sunset	0	0
Pacific Heights	0	0
Westwood Highlands	0	0
Visitacion Valley	0	0
Union Square	0	0
Twin Peaks	0	0
Treasure Island	0	0
Telegraph Hill	0	0
Sunset District	0	0
International Settlement	0	0
South of Market	0	0
Sea Cliff	0	0

Rancho Las Camaritas	0	0
Presidio Terrace	0	0
Presidio Heights	0	0
Potrero Hill	0	0
Parkside	0	0
Parkmerced	0	0
Silver Terrace	0	0
Inner Sunset	0	0
Yerba Buena Island	0	0
Ingleside	0	0
Central Sunset	0	0
Chinatown	0	0
Ingleside Terraces	0	0
Corona Heights	0	0
Crocker-Amazon	0	0
Diamond Heights	0	0
Dogpatch	0	0
Bayview-Hunters Point	0	0
Bayview	0	0
Central Embarcadero Piers Historic District	0	0
Balboa Park	0	0
Excelsior District	0	0
Balboa Terrace	0	0
Cathedral Hill	0	0
Fisherman's Wharf	0	0
Anza Vista	0	0
India Basin	0	0
Forest Hill	0	0
Hunters Point	0	0
Forest Knolls	0	0
Glen Park	0	0
Fillmore District	0	0
Alta Plaza	0	0
Hayes Valley	0	0

Cluster 1:

Neighborhood	Frequency of Indian Restaurants	Cluster #
Rincon Hill	0.02272727	1
Upper Fillmore	0.02439024	1
Castro District	0.025	1
Bernal Heights	0.02777778	1
St. Francis Wood	0.03846154	1
West Portal	0.03225806	1
Russian Hill	0.02083333	1
Western Addition	0.02941176	1
Cole Valley	0.02272727	1
Alamo Square	0.04545455	1
Japantown	0.02597403	1
Lower Pacific Heights	0.02	1
Duboce Park Landmark District	0.02	1
Eureka Valley	0.02	1
Mission District	0.03	1
Richmond District	0.01960784	1

Cluster 2:

Neighborhood	Frequency of Indian Restaurants	Cluster #
Haight-Ashbury	0.01052632	2
Lower Haight	0.01538462	2
Financial District	0.01030928	2
Barbary Coast	0.01086957	2
Civic Center	0.01612903	2
Tendernob	0.01	2
Mission Dolores	0.01	2
Dumpville	0.01	2
Duboce Triangle	0.01	2
South Park	0.01098901	2
Carville	0.01	2
Outside Lands	0.01	2
Compton's Transgender Cultural District	0.01	2
Theatre District	0.01	2
Terrific Street	0.01	2