Capstone Project Data Science

The Battle of Neighborhoods

New Italian Restaurant in San Francisco, United States of America

Business problem

San Francisco is the city of my dreams and I think it might be a good thing to open an Italian restaurant in that city.

Hence the idea of this project. I would like to determine the best possible location to open an Italian restaurant based on the different localities of the city, already established Italian restaurant in varios geographical location and ease of accessibility by maximum number of people so that the revenue from the latest venture can be maximized.

So I thought well to use an ML algorithm and Foursquare data to understand in which neighborhood to open the restaurant based on what people have done before him.

Data

This project will use data from:

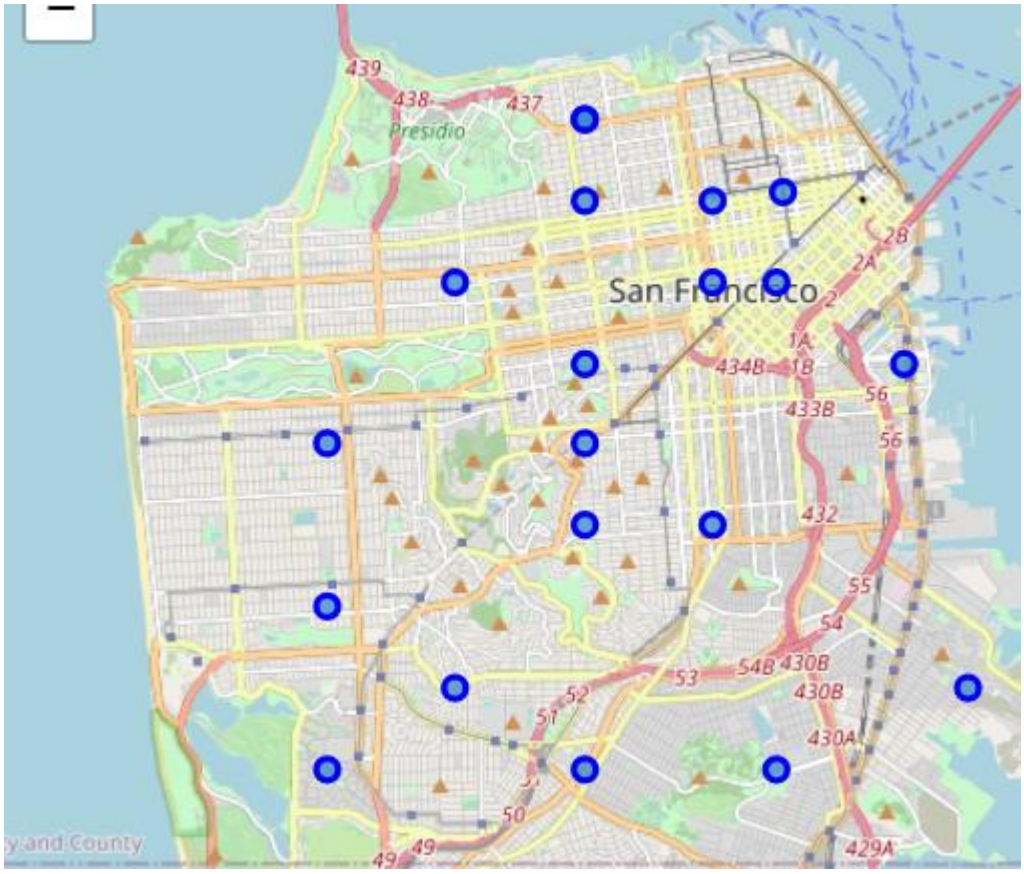
- http://www.healthysf.org/bdi/outcomes/zipmap.htm - For getting information about Neighborhoods in San Francisco.
- Geopy - For getting the co-ordinated of different locations.
- Foursquare API - To get the list of vanues and their details around a given location.

Zip Code	Neighborhood	Population (Census 2000)
94102	Hayes Valley/Tenderloin/North of Market	28991
94103	South of Market	23016
94107	Potrero Hill	17368
94108	Chinatown	13716
94109	Polk/Russian Hill (Nob Hill)	56322
94110	Inner Mission/Bernal Heights	74633
94112	Ingelside-Excelsior/Crocker-Amazon	73104
94114	Castro/Noe Valley	30574
94115	Western Addition/Japantown	33115
94116	Parkside/Forest Hill	42958
94117	Haight-Ashbury	38738
94118	Inner Richmond	38939
94121	Outer Richmond	42473
94122	Sunset	55492
94123	Marina	22903
94124	Bayview-Hunters Point	33170
94127	St. Francis Wood/Miraloma/West Portal	20624
94131	Twin Peaks-Glen Park	27897
94132	Lake Merced	26291
94133	North Beach/Chinatown	26827
94134	Visitacion Valley/Sunnydale	40134

Data Exploration

Zip Code	Neighborhood	Latitude	Longitude
94102	Hayes Valley/Tenderloin/North of Market	37.780	-122.420
94103	South of Market	37.780	-122.410
94107	Potrero Hill	37.770	-122.390
94108	Chinatown	37.791	-122.409
94109	Polk/Russian Hill (Nob Hill)	37.790	-122.420
94110	Inner Mission/Bernal Heights	37.750	-122.420
94112	Ingelside-Excelsior/Crocker-Amazon	37.720	-122.440
94114	Castro/Noe Valley	37.760	-122.440
94115	Western Addition/Japantown	37.790	-122.440
94116	Parkside/Forest Hill	37.740	-122.480
94117	Haight-Ashbury	37.770	-122.440
94118	Inner Richmond	37.780	-122.460
94121	Outer Richmond	37.800	-122.700
94122	Sunset	37.760	-122.480
94123	Marina	37.800	-122.440
94124	Bayview-Hunters Point	37.730	-122.380
94127	St. Francis Wood/Miraloma/West Portal	37.730	-122.460
94131	Twin Peaks-Glen Park	37.750	-122.440
94132	Lake Merced	37.720	-122.480
94133	North Beach/Chinatown	37.800	-122.440
94134	Visitacion Valley/Sunnydale	37.720	-122.410

After cleansing the data, the next step was to analyze it. We then created a map using Folium and colour-coded each Neighborhood.



Next, we used the Foursquare API to get a list of all the Venues in San Francisco which included Parks, Schools, Café Shops, Asian Restaurants etc. Getting this data was crucial to analyzing the number of Italian Restaurants all over San Francisco. There was a total of 2 Italian Restaurants in Toronto. We then merged the Foursquare Venue data with the Neighborhood data which then gave us the nearest Venue for each of the Neighborhoods.

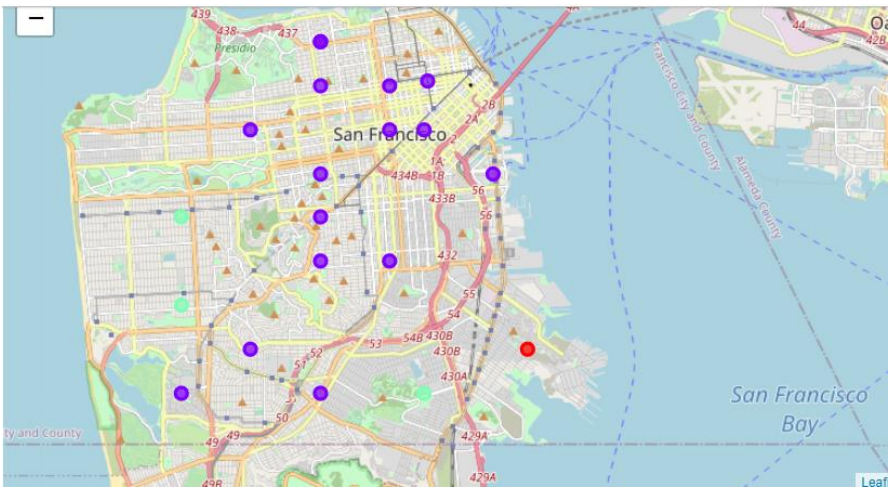
ML Clustering

Then to analyze the data we performed a technique in which Categorical Data is transformed into Numerical Data for Machine Learning algorithms. This technique is called One hot encoding. For each of the neighbourhoods, individual venues were turned into the frequency at how many of those Venues were located in each neighbourhood.

Then we grouped those rows by Neighborhood and by taking the average of the frequency of occurrence of each Venue Category.

To make the analysis more interesting, we wanted to cluster the neighbourhoods based on the neighbourhoods that had similar averages of Italian Restaurants in that Neighborhood. To do this we used K-Means clustering. To get our optimum K value that was neither overfitting or underfitting the model, we used the Silhuettes Technique.

After that in following images we can see the cluster on the map and on table



Discussion and Conclusion

Most of the Italian Restaurants are in cluster 2 represented by the purple clusters.

Also, the analysis does not take into consideration of the Italian population across neighbourhoods as this can play a huge factor while choosing which place to open a new Italian restaurant. This concludes the optimal findings for this project and recommends the entrepreneur to open an authentic Italian restaurant in these locations with little to no competition.

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	10th Most Common Venue
Hayes Valley/Tenderloin/North of Market	1.0	Coffee Shop	French Restaurant	Vietnamese Restaurant	Sushi Restaurant	Dessert Shop	Thai Restaurant	Cocktail Bar	Theater	Wine Bar	Bakery
South of Market	1.0	Coffee Shop	Bakery	Marijuana Dispensary	Vietnamese Restaurant	Gym / Fitness Center	Women's Store	Sandwich Place	Pizza Place	Speakeasy	Beer Bar
Potrero Hill	1.0	Coffee Shop	Baseball Stadium	Park	Café	Food Truck	Gym	Sandwich Place	Outdoor Sculpture	New American Restaurant	Pier
Chinatown	1.0	Hotel	Coffee Shop	Boutique	Gym	Sushi Restaurant	Speakeasy	Men's Store	Cocktail Bar	French Restaurant	Clothing Store
Polk/Russian Hill (Nob Hill)	1.0	Gym / Fitness Center	Wine Bar	Grocery Store	Vietnamese Restaurant	Coffee Shop	Sushi Restaurant	American Restaurant	Café	Bar	Art Gallery
Inner Mission/Bernal Heights	1.0	Mexican Restaurant	Coffee Shop	Cocktail Bar	Pizza Place	Italian Restaurant	Breakfast Spot	Yoga Studio	Dive Bar	Tea Room	Deli / Bodega
Ingelside-Excelsior/Crocker-Amazon	1.0	Mexican Restaurant	Latin American Restaurant	Pizza Place	Coffee Shop	Sandwich Place	Chinese Restaurant	Pharmacy	Vietnamese Restaurant	Baseball Field	Bakery
Castro/Noe Valley	1.0	Gay Bar	Coffee Shop	Thai Restaurant	Park	Trail	Hill	Scenic Lookout	Wine Bar	New American Restaurant	Yoga Studio
Western Addition/Japantown	1.0	Bakery	Cosmetics Shop	Sandwich Place	Salon / Barbershop	Ice Cream Shop	Spa	Gym / Fitness Center	Park	Boutique	American Restaurant