Data Science Final Project – San Francisco Data Analysis

Introduction

Background:

The purpose of this project is to help the people wanting to immigrate to major cities of the world with research on good neighborhoods and ultimately make a smart decision.

The settlement decision will be made based on crimes count and access to the colleges.

Essentially this project will provide a detailed information of the neighborhoods in a city including the various catergorical venues including yoga centers, coffee shops, restaurants etc.

Assuming that an immigrant wants to setup a restaurant business, this project will be providing with a list of neighborhoods with least number of restaurants for someone to setup as a business.

In totality it will help people get an awareness on a new city. It will provide a comparative analysis of various neighbor hoods for settlement and opening up a particular business.

Problem Statement:

The purpose of this project is to provide information on:

- Good neighborhoods with accessible venues
- Good secondary school location and information
- Good neighborhood to setup a restaurant

Location to be analyzed:

San Francisco, CA is a very popular destination for immigration for staying and setting up business. Therefore, for this project the SF city, its neighborhoods, school districts and various categorical venues will be used.

K-Means clustering:

The similarities or dissimilarities between two neighborhoods in a city could be visualized by segmenting them into various clusters utilizing the k-means clustering machine learning algorithm. So, this project aims at clustering and making sense of data obtained from this clustering technique

Location Data Analysis - Four Square API:

Forsquare is one of the location data providers. A developer account has been created and credentials have been obtained. Due to a limited number of accesses for this developer account, there will be some restrictions on the radius and count of venues search.

Data Analysis and Libraries:

Along with the Foursquare API, Python and its associated libraries and packages will be used for data analysis and visualization.

Pandas: Dataframe creation and manipulation

Numpy: Mathematical Analysis

Matplotlib: Python plotting module

Folium: Interactive leaflet map creation

Scikit Learn: Implementing k-means clustering algorithm

JSON: Handling JSON files

Geocoder: Retrieving location data

Data Section

The SF crimes data is obtained form:

https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783

The SF colleges data:

https://data.sfgov.org/Economy-and-Community/Colleges-in-San-Francisco-2011-/8r3f-pc6a

Also, the data set with SF neighborhood spatial data is obtained from:

https://data.sfgov.org/Geographic-Locations-and-Boundaries/Analysis-Neighborhoods/p5b7-5n3h

Data obtained from Four Square API:

Four Square us a location data provider with information on venues nearby a neighborhood. It provides:

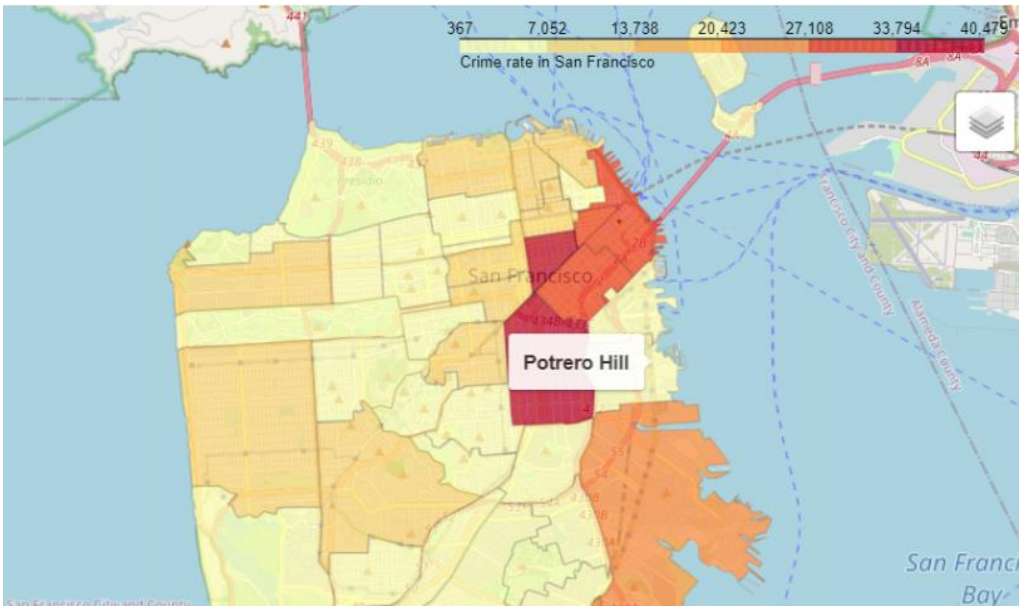
- Neighborhood
- Neighborhood Latitude
- Neighborhood Longitude
- Venue
- Name of the Venue
- Venue latitude
- Venue longitude
- Venue Category

Determining the location to live

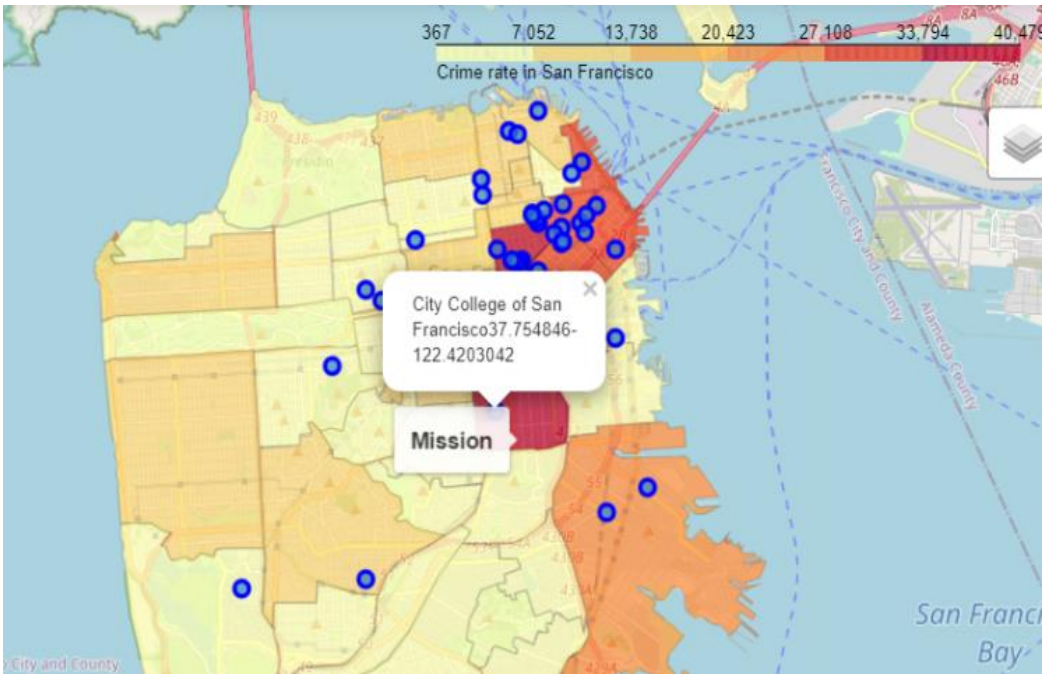
The immigrants look for lesser crimes and easy access to good colleges in the neighborhood they want to settle.

Based on crimes, the top 10 neighborhoods with least crimes is provided.

	Neighborhood	Count
0	Japantown	3653
1	Noe Valley	3447
2	Presidio Heights	2174
3	Twin Peaks	1846
4	Glen Park	1841
5	Treasure Island	1180
6	Presidio	822
7	Lincoln Park	436
8	Seacliff	408
9	McLaren Park	367



Then the college data is obtained and map visualizations are provided.



Based on the college locations and lesser crimes, the best neighborhoods for settlement is provided.

	Neighborhood	Latitude	Longitude	Count
0	Pacific Heights	37.795614	-122.423493	6169
1	Presidio Heights	37.791879	-122.446300	2174
2	Japantown	37.785335	-122.424687	3653
3	Potrero Hill	37.767240	-122.384870	5863
4	Lone Mountain/USF	37.782643	-122.442692	4390

Determining the best neighborhood to open restaurant business

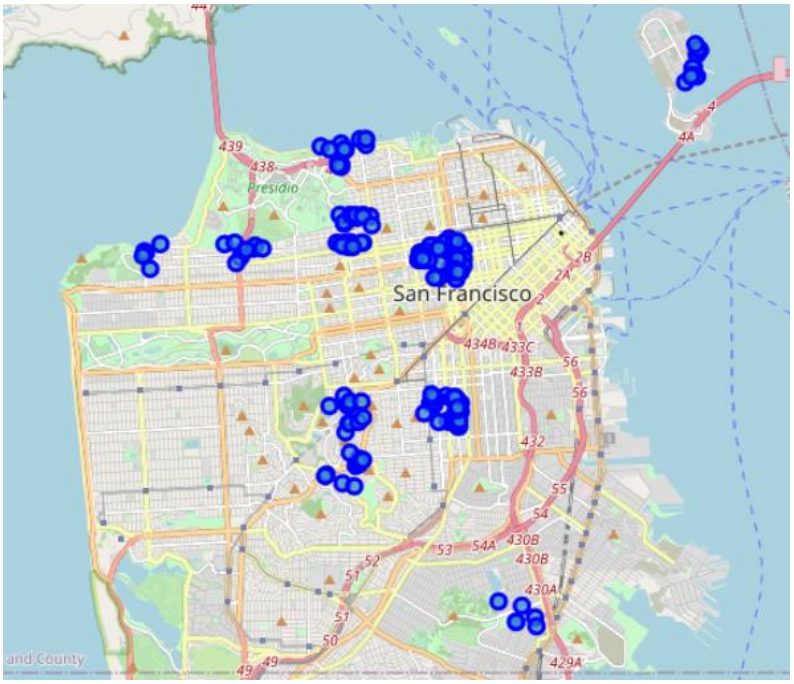
The decision is made on the basis that the restaurant is is not in top 3 venues for any particular neighborhood.

The top 10 neighborhoods based on crimes are considered to display the venues.

Foursquare API is used to get all the information of venues and locations around the neighborhoods.

Venue category is an important column for us to determine the neighborhoods without coffee shop as their top venues.

Then, a table is created with all the neighborhoods along with their top 5 venues.



	Neighborhoods	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	Glen Park	Bus Station	Scenic Lookout	Rental Car Location	Grocery Store	Park	Garden	Cosmetics Shop	Trail	Athletics & Sports	Performing Arts Venue
1	Japantown	Sushi Restaurant	Japanese Restaurant	Hotel	Korean Restaurant	Ramen Restaurant	Massage Studio	Shopping Mall	Sandwich Place	Gift Shop	Grocery Store
2	Lincoln Park	Trail	Scenic Lookout	Beach	Golf Course	Accessories Store	Music Store	Pharmacy	Pet Store	Performing Arts Venue	Park
3	McLaren Park	Garden	Art Gallery	Historic Site	Performing Arts Venue	Park	Music Venue	Accessories Store	Pizza Place	Pharmacy	Pet Store
4	Noe Valley	Indian Restaurant	Art Gallery	Ice Cream Shop	Park	Record Shop	Gift Shop	Coffee Shop	Music Venue	Playground	Bookstore
5	Presidio	Park	Beach	Snack Place	Donut Shop	Escape Room	Lighthouse	Athletics & Sports	Event Space	Outdoor Sculpture	Historic Site
6	Presidio Heights	Furniture / Home Store	Trail	Park	Scenic Lookout	Spa	Pet Store	Women's Store	Sculpture Garden	Indie Movie Theater	Hotel
7	Seacliff	Playground	Trail	Scenic Lookout	Park	Tennis Court	Bus Station	Intersection	Art Gallery	Dog Run	Pet Store
8	Treasure Island	Food Truck	Music Venue	Rugby Pitch	Food Stand	Harbor / Marina	Baseball Field	Brewery	Athletics & Sports	Performing Arts Venue	Nail Salon
9	Twin Peaks	Trail	Park	Hill	Tailor Shop	Scenic Lookout	Garden	Accessories Store	Miscellaneous Shop	Pet Store	Performing Arts Venue

Determining the best neighborhood to open restaurant business and settle

Based on the analysis on crimes, colleges and venues data, we can come to a conclusion that:

The best neighborhoods to settle:

	Neighborhood	Latitude	Longitude	Count
1	Presidio Heights	37.791879	-122.446300	2174
2	Japantown	37.785335	-122.424687	3653

The best neighborhoods for restaurant business:

	Neighborhood	Latitude	Longitude	Count
0	Glen Park	37.746482	-122.447375	1841
1	Lincoln Park	37.787743	-122.493018	436
2	McLaren Park	37.719215	-122.406665	367
3	Presidio	37.806892	-122.448129	822
4	Presidio Heights	37.791879	-122.446300	2174
5	Seacliff	37.787234	-122.472352	408
6	Treasure Island	37.820871	-122.363583	1180
7	Twin Peaks	37.756550	-122.446950	1846

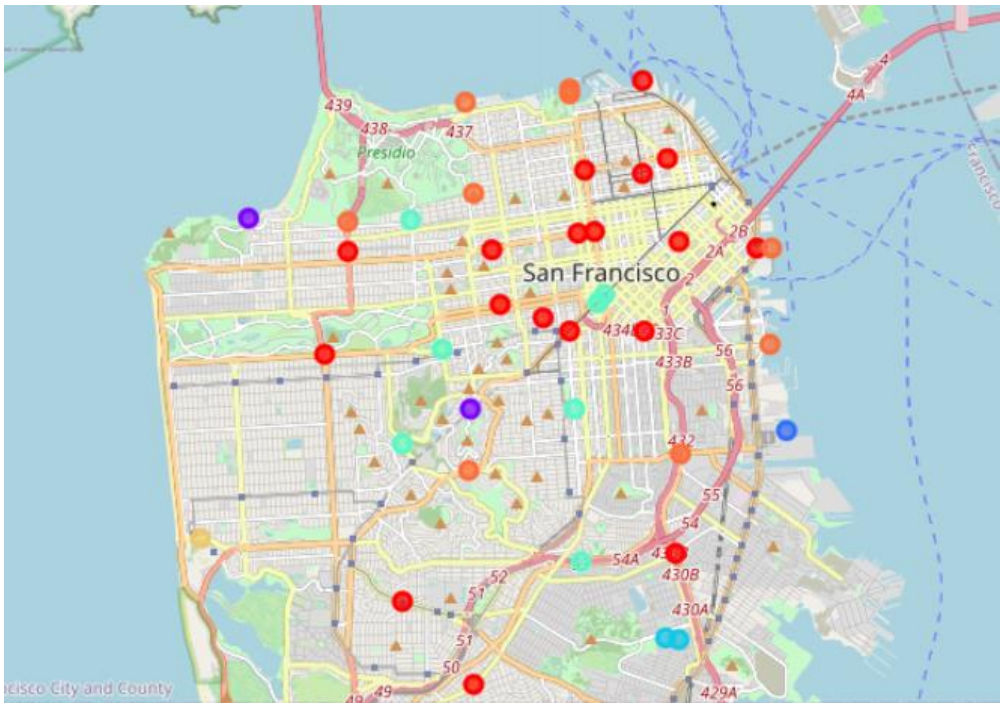
K-Means Clustering – Analyzing the Neighborhoods and Venues using Machine Learning

First, the cluster labels are created and cluster labels are added to the table with neighborhood and venues information.

Based on this cluster label we can map them to visualize the location distinction.

Then we can analyze them label by label to understand how the algorithms has segregated the city data.

	Neighborhood	Latitude	Longitude	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
0	Bayview Hunters Point	37.753070	-122.381578	2	Dance Studio	Rental Service	Accessories Store	Outdoor Sculpture	Persian Restaurant	Performing Arts Venue
1	Bernal Heights	37.749337	-122.403613	7	Fabric Shop	Office	Brewery	Cocktail Bar	Liquor Store	Skate Park
2	Castro/Upper Market	37.769485	-122.426555	0	Sushi Restaurant	Cocktail Bar	Clothing Store	Boutique	Gym / Fitness Center	Coffee Shop
3	Chinatown	37.797559	-122.406226	0	Italian Restaurant	Coffee Shop	Pizza Place	Chinese Restaurant	Café	Men's Store
4	Excelsior	37.731552	-122.423982	4	Latin American Restaurant	Dog Run	Wine Shop	Farm	Food Truck	Convenience Store



Each color dot represents one cluster, for this example only 8 clusters are used.

Analyzing each cluster separately.

Cluster 0 represents the neighborhoods with restaurants as their most common venues. None of the neighborhoods to setup the business are in this cluster.

```
sfmerge.loc[sfmerge['Cluster Labels'] == 0]
```

	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	
2	Castro/Upper Market	37.769485	-122.426555	0	Sushi Restaurant	Cocktail Bar	Clothing Store	Boutique	Gym / Fitness Center	
3	Chinatown	37.797559	-122.406226	0	Italian Restaurant	Coffee Shop	Pizza Place	Chinese Restaurant	Café	
5	Financial District/South Beach	37.782797	-122.387525	0	Baseball Stadium	Sandwich Place	Coffee Shop	Scenic Lookout	American Restaurant	
8	Golden Gate Park	37.773635	-122.440922	0	Coffee Shop	Bar	Pizza Place	Indian Restaurant	Burrito Place	B
9	Haight Ashbury	37.771431	-122.431998	0	Coffee Shop	Record Shop	Dive Bar	Cocktail Bar	Grocery Store	
12	Japantown	37.785335	-122.424687	0	Sushi Restaurant	Hotel	Japanese Restaurant	Ramen Restaurant	Shopping Mall	R
17	Lone Mountain/USF	37.782643	-122.442692	0	Gym / Fitness Center	Sandwich Place	Café	Coffee Shop	Health & Beauty Service	
20	Mission	37.769433	-122.410959	0	Nightclub	Gay Bar	Bar	Furniture / Home Store	Clothing Store	
22	Nob Hill	37.795061	-122.411472	0	Chinese Restaurant	Bakery	Italian Restaurant	Hotel	Dim Sum Restaurant	

The top neighborhoods to open a restaurant business are in other clusters.

Neighborhood	Latitude	Longitude	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
Glen Park	37.746482	-122.447375	7	Bus Station	Scenic Lookout	Cosmetics Shop	Trail	Rental Car Location	Garden
Lincoln Park	37.787743	-122.493018	1	Trail	Beach	Golf Course	Scenic Lookout	Organic Grocery	Pedestrian Plaza
McLaren Park	37.719215	-122.406665	3	Garden	Performing Arts Venue	Music Venue	Historic Site	Art Gallery	Park
Presidio	37.806892	-122.448129	7	Park	Beach	Harbor / Marina	Lighthouse	Snack Place	Donut Shop
Presidio Heights	37.791879	-122.446300	7	Scenic Lookout	Furniture / Home Store	Trail	Spa	Park	Pet Store
Seacliff	37.787234	-122.472352	7	Intersection	Bus Station	Dog Run	Trail	Park	Art Gallery
Treasure Island	37.820871	-122.363583	5	Food Truck	Music Venue	Rugby Pitch	Athletics & Sports	Baseball Field	Brewery
Twin Peaks	37.756550	-122.446950	1	Trail	Garden	Tailor Shop	Park	Hill	Scenic Lookout

Conclusion

The San Francisco neighborhood data, crimes data and college data is collected and analyzed to decide on the neighborhood to settle. The venues data obtained from foursquare API along with the settlement analysis has given the best location to open a restaurant business.

The KMeans clustering is used to prove the hypothesis correct regarding the restaurant business.