Applied Data Science Capstone

Data Science Capstone: The Best and safest Neighborhood in San Francisco for Opening a Coffee Shop

By, Marcos Cifuentes

Julio 2021

1. Introduction

(Data initial by wikipedia)

California has a population of 39.5 million people (est. 2019) making it the most populous U.S. state. The state capital is Sacramento; the largest city is Los Angeles. Other major cities are San Francisco, San Diego, San Jose, and Long Beach. San Francisco is a cultural, commercial, and financial center in Northern California. San Francisco is the 16th most populous city in the United States, and the fourth most populous in California, with 881,549 residents as of 2019. It covers an area of about 46.89 square miles (121.4 square kilometers),mostly at the north end of the San Francisco Peninsula in the San Francisco Bay Area, making it the second most densely populated large U.S. city, and the fifth most densely populated U.S. county, behind only four of the five New York City boroughs.

With its diverse culture, comes diverse alternatives business relationship with food items,beverages and cafe in safest Neighborhoods. So as part of this project , we will list and visualize all options to create "businnes altenative safest" relations with cafe in San Francisco (CA).

This document contains some basic assumptions, data sets, and analysis that can inform your decision when selecting the optimal neighborhood in San Francisco for opening a coffee shop.

1.1. Business Problem

> As a business idea, the goal is to open a coffee shop in a safe and central place in San Francisco. Based on data science and data provided by the police department along with current business locations, optimal recommendations should be indicated to a set of investors before opening stores.

1.2. Target Audience

> Anyone who wants to buy or build a coffee shop in San Francisco, or anyone in San Francisco just looking for a nice area to take a cup of coffee.

Applied Data Science Capstone

2. Data and libraries

To retrieve information, we need two datasets for San Francisco: Crime Data and Registered Business Data. Both using from the https://datasf.org/academy/

a- For Crime Rate in San Francisco Section [San Francisco Crime Data]

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

pd.read_csv('Police_Department_Incident_Reports__2018_to_Present2.csv')

b- For Data Set with locations of businesses [San Francisco Registered Business Data]

with urllib.request.urlopen("https://data.sfgov.org/resource/wg3w-h783.json") as url:francisco = json.loads(url.read().decode())

c- Foursquare Data

The Places API offers real-time access to Foursquare's global database of rich venue data and user content to power your location-based experiences in your app or website.

d- geopy

python-geoip is a library that provides access to GeoIP databases. Currently it only supports accessing MaxMind databases.

e- folium

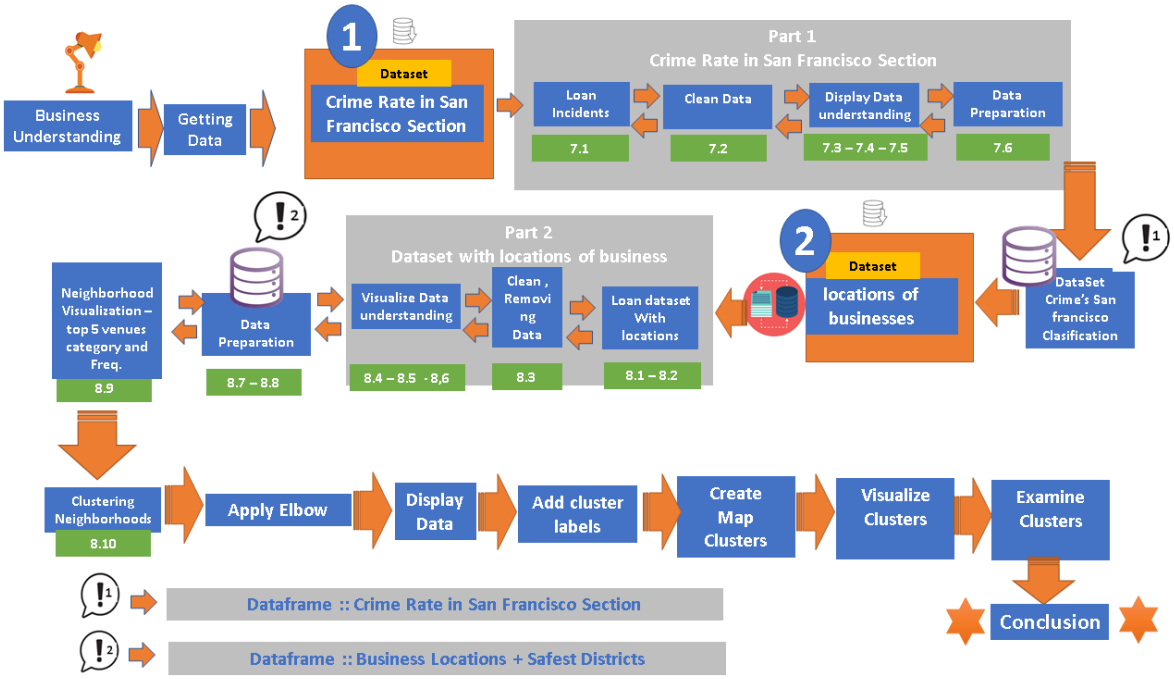
Folium is a Python library used for visualizing geospatial data. It is easy to use and yet a powerful library. Folium is a Python wrapper for Leaflet.js which is a leading open-source JavaScript library for plotting interactive maps.

IBM Developer
SKILLS NETWORK

Applied Data Science Capstone

3.Methodology

General Overview



On the graph above section we can see the main steps executed with brief descriptions.

To start will pull two datasets, one for data crime and another one for location of business. The url's: https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783 and ("https://data.sfgov.org/resource/wg3w-h783.json") respectively. In the two previous data sets, for each one we are going to clean data, remove duplicates, represent and leave unique sets of unrepeatable data. As a result, we must to merge in one dataset with the safest venues optimal to develop business.

Next, For each neighborhood we will find the venues locations using "api.foursquare.com" (coord. Lat and Long.). As a result, we will save a dataframe with each Neighborhood and 100 closest places to 500 meters. We will to grouped Neighborhood classifying each category of venues. the last table, already merged and normalized, contains the information of the business locations by categories. Each row contains the neighborhoods and the columns each category of locality with normalized values. The last table, already merged and cleaned, is normalized and categorized by neighborhoods and business types.

Next, we specify the data structure that we use to feed the K means algorithm. As you can see, we only load the closest locations that the search for "api.foursquare.com" returned.

Next, we will be clustering neighborhood and applying elbow method to determine optimal "K" in the K-means execution algorithm.

finally, we will show the label's clustering and the conclusions.

4. Approach

a) We can pull two datasets:

1) list of every business registered in San Francisco from the last couple of decades from the data SF website.

2) a list of date crime in San francisco (since 2018 - present)

b) Using FourSquare API to find all venues for each safe neighborhood.

c) Filter out all venues that are nearby by locality.

d) Using aggregative rating for each Coffee Shop to find the best places.

e) Visualize the Ranking of neighborhoods using folium library(python)

f) Apply Kmeans to find best candidates to open a new coffee shop.

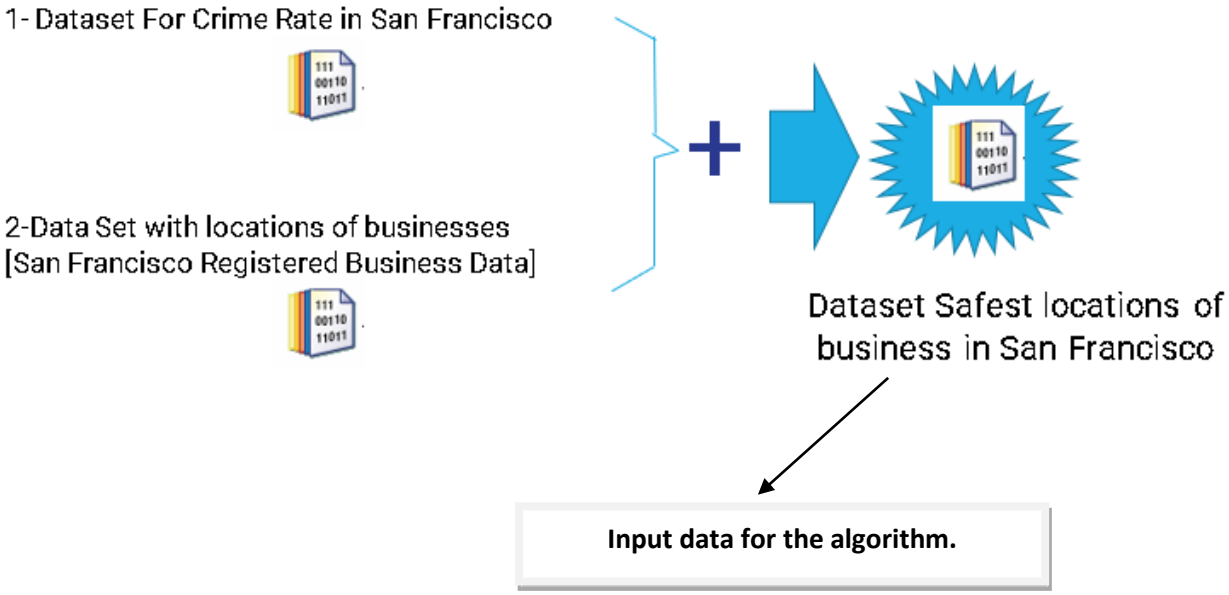
5. Predictive Modeling

Our modeling is based on the K Means clustering algorithm.

Steps:

1-We define the input

We concrete the data structure by mixing crime data entries and business locations to feed the algorithm.



IBM Developer
SKILLS NETWORK

[Applied Data Science Capstone]

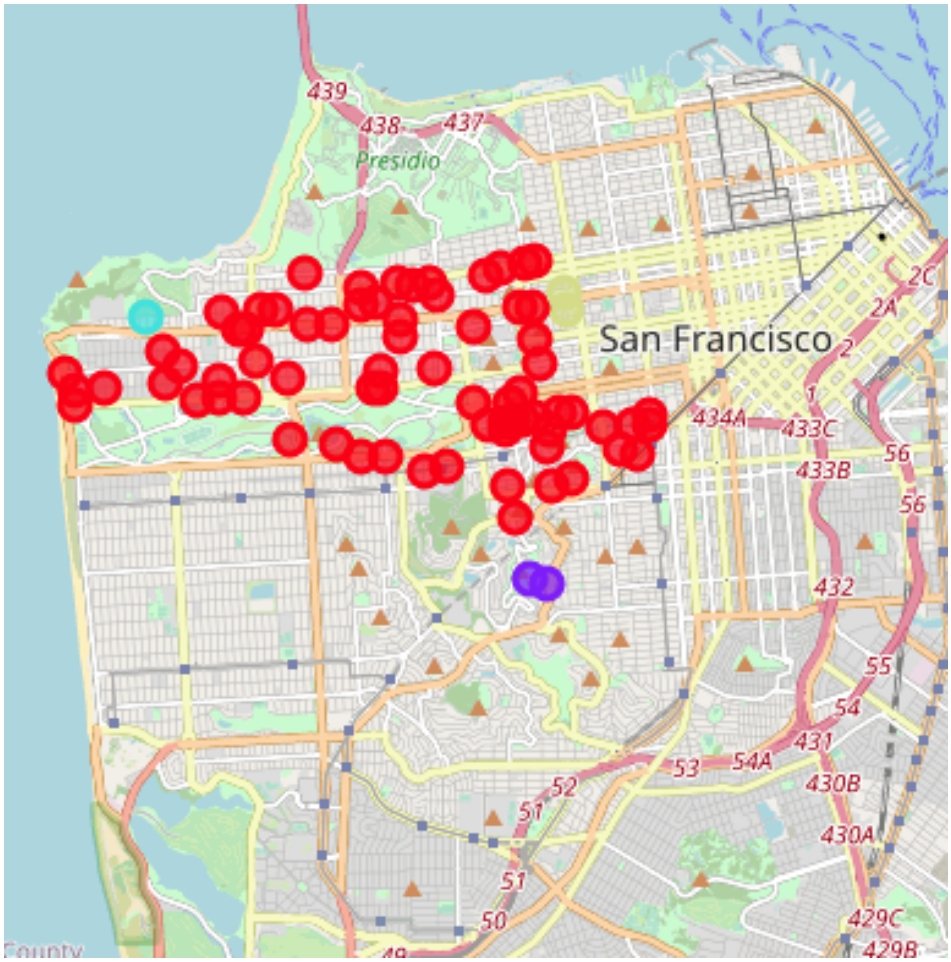
2-Obtain the K value

Find the value of K by making a graph and trying to find the "elbow point"

3-Execute K-Means

Run the algorithm for 4 clusters and obtain the labels and centroids.

Here we can see that the K-Means Algorithm with K = 4



Follows the graphics for clusters and labels::

IBM Developer
SKILLS NETWORK

[Applied Data Science Capstone]

Cluster 0

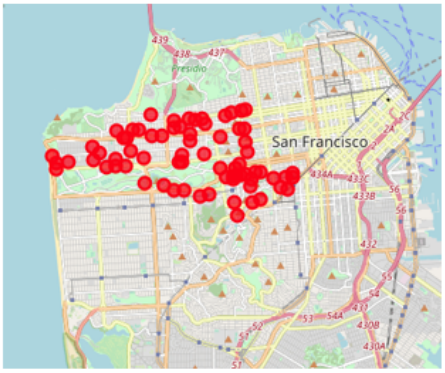
According to the results for the first group (labels=0)
The most popular categories are:

- Boutique
- Coffee Shop
- Thrift / Vintage Store
- Clothing Store
- Chinese Rest.
- Garden
- Park
- Café
- Bakery

	Neighborhood	Latitude	Longitude	police_district	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Haight Ashbury	37.762519	-122.449296	Park	0	Boutique	Coffee Shop	Thrift / Vintage Store	Park	Café
1	Haight Ashbury	37.760729	-122.449454	Park	0	Boutique	Coffee Shop	Thrift / Vintage Store	Park	Café
2	Haight Ashbury	37.760944	-122.443356	Park	0	Boutique	Coffee Shop	Thrift / Vintage Store	Park	Café
3	Haight Ashbury	37.769307	-122.431570	Park	0	Boutique	Coffee Shop	Thrift / Vintage Store	Park	Café
4	Haight Ashbury	37.770210	-122.445345	Park	0	Boutique	Coffee Shop	Thrift / Vintage Store	Park	Café
5	Inner Sunset	37.764272	-122.460958	Park	0	Coffee Shop	Garden	Sushi Restaurant	Bakery	Science Museum
6	Inner Sunset	37.759091	-122.460090	Park	0	Coffee Shop	Garden	Sushi Restaurant	Bakery	Science Museum
7	Lone Mountain/USF	37.778837	-122.445347	Park	0	Café	Coffee Shop	Cosmetics Shop	Boutique	Bank
8	Lone Mountain/USF	37.772790	-122.447497	Park	0	Café	Coffee Shop	Cosmetics Shop	Boutique	Bank
9	Haight Ashbury	37.769058	-122.435794	Park	0	Boutique	Coffee Shop	Thrift / Vintage Store	Park	Café

It is recommended to open a coffee shop

Cluster 0
Graphic and Table



	Neighborhood	Latitude	Longitude	police_district	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Haight Ashbury	37.762519	-122.449296	Park	0	Boutique	Coffee Shop	Thrift / Vintage Store	Park	Café
1	Haight Ashbury	37.760729	-122.449454	Park	0	Boutique	Coffee Shop	Thrift / Vintage Store	Park	Café
2	Haight Ashbury	37.760944	-122.443356	Park	0	Boutique	Coffee Shop	Thrift / Vintage Store	Park	Café
3	Haight Ashbury	37.769307	-122.431570	Park	0	Boutique	Coffee Shop	Thrift / Vintage Store	Park	Café
5	Haight Ashbury	37.770210	-122.445345	Park	0	Boutique	Coffee Shop	Thrift / Vintage Store	Park	Café
6	Inner Sunset	37.764272	-122.460958	Park	0	Coffee Shop	Garden	Sushi Restaurant	Bakery	Science Museum
7	Inner Sunset	37.759091	-122.460090	Park	0	Coffee Shop	Garden	Sushi Restaurant	Bakery	Science Museum
8	Lone Mountain/USF	37.778837	-122.445347	Park	0	Café	Coffee Shop	Cosmetics Shop	Boutique	Bank
9	Lone Mountain/USF	37.772790	-122.447497	Park	0	Café	Coffee Shop	Cosmetics Shop	Boutique	Bank
10	Haight Ashbury	37.769058	-122.435794	Park	0	Boutique	Coffee Shop	Thrift / Vintage Store	Park	Café
11	Castro/Upper Market	37.762553	-122.442983	Park	0	Gay Bar	Park	Coffee Shop	Grocery Store	Gym
12	Castro/Upper Market	37.765996	-122.431018	Park	0	Gay Bar	Park	Coffee Shop	Grocery Store	Gym
13	Lone Mountain/USF	37.776017	-122.444760	Park	0	Café	Coffee Shop	Cosmetics Shop	Boutique	Bank

It is recommended to open a coffee shop

Cluster 1

According to labels=1 The most popular categories are:

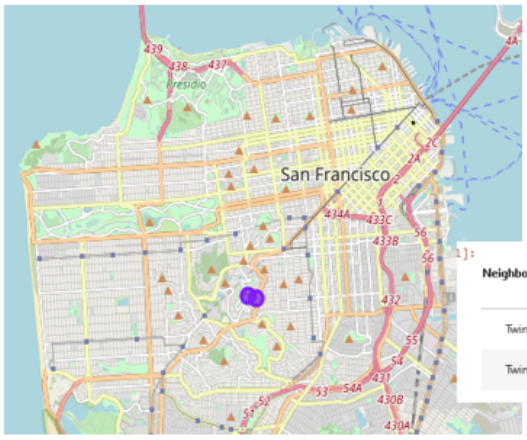
- Trail
- Scenic Lookout Prak
- Park
- Hill
- Tennis Court

1]:

Neighborhood	Latitude	Longitude	police_district	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Twin Peaks	37.751799	-122.443548	Park	1	Trail	Scenic Lookout	Park	Hill	Tennis Court
Twin Peaks	37.752352	-122.445984	Park	1	Trail	Scenic Lookout	Park	Hill	Tennis Court

It is not recommended to open a coffee shop

Cluster 1
Graphic and Table



1]:

Neighborhood	Latitude	Longitude	police_district	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Twin Peaks	37.751799	-122.443548	Park	1	Trail	Scenic Lookout	Park	Hill	Tennis Court
Twin Peaks	37.752352	-122.445984	Park	1	Trail	Scenic Lookout	Park	Hill	Tennis Court

It is not recommended to open a coffee shop

Cluster 2

According to labels=2 The most popular categories are:

- Cosmetic Shop
- Gym / Fitness Center
- Sandwich Place
- Deli / Bodega
- Park
- Sports Bar
- Yoga Studio
- Chinese Restaurant

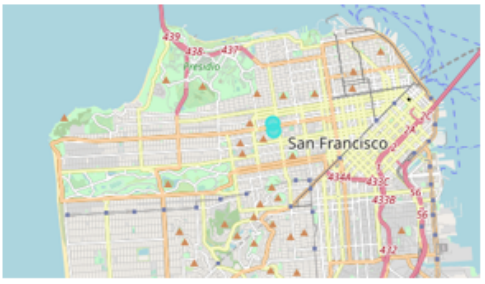
ighborhood	Latitude	Longitude	police_district	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Western Addition	37.782239	-122.440963	Park	2	Cosmetics Shop	Gym / Fitness Center	Sandwich Place	Deli / Bodega	Park
Japantown	37.784107	-122.441349	Richmond	2	Chinese Restaurant	Yoga Studio	Gym / Fitness Center	Salon / Barbershop	Sports Bar

It is not recommended to open a coffee shop

Cluster 2
Graphic and Table

According to labels=2 The most popular categories are:

- Cosmetic Shop
- Gym / Fitness Center
- Sandwich Place
- Deli / Bodega
- Park
- Sports Bar
- Yoga Studio
- Chinese Restaurant



It is not recommended to open a coffee shop

Cluster 3

According to labels=3 The most popular categories are:

- Pharmacy
- Park
- Gold Course
- Café
- Cafeteria
- Bus Stop

	Neighborhood	Latitude	Longitude	police_district	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Lincoln Park	37.781327	-122.499871	Richmond	3	Pharmacy	Park	Golf Course	Café	Cafeteria

It is recommended to open a coffee shop

Cluster 3
Graphic and table

According to labels=3 The most popular categories are:

- Pharmacy
- Park
- Gold Course
- Café
- Cafeteria
- Bus Stop

It is recommended to open a coffee shop

5. Answers to Initial Questions

- **Which all areas have more number of coffee shops?**

According to reports are: Lincoln Park, Castro/Upper Market, Haight Ashbury, Hayes Valley, Inner Sunset, Lone Mountain, Inner Richmond, Outer Richmond, Presidio Heights , Western Addition

- **Which all areas have less number of coffeeshops?**

  According to reports are: Inner Richmond, JapanTown, Twin Peaks

- **Which districts have the neighborhoods safest ?**

  According to reports are: Park and Richmond

- **Which districts Neighborhoods The Most Unsafe Districts ?**

According to reports are: Marina, South of Market, Tenderloin, District/ south Beachark, BayView and misión.

6. Conclusion

According to the final clustering aplying modeling with k-means and elbow=4, it is recommende to open a **Coffee Shop** on the venues corresponding to Lincoln Park (corresponding to Richmond) and / or Lincoln Park, Castro/Upper Market, Haight Ashbury, Hayes Valley, Inner Sunset, Lone Mountain, Inner Richmond, Outer Richmond, Presidio Heights , Western Addition (corresponding to Park).