IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

--+--

Capstone Project

The Battle of the Neighborhoods

by

Tieu My Trinh

10 Jun 2020

1. Introduction

San Francisco is the cultural, commercial, and financial center of Northern California, United States. As of 2020, San Francisco has the highest salaries, disposable income, and median home prices in the world at $1.7 million. In the United State, San Francisco ranked among the top ten highest income counties, which a personal income per capita is $130,696 [1]. Life in San Francisco is vibrant too, as there are a lot of restaurants, café, and shopping centers.

Given that prosperity, a client, which is a high-class jewelry maker, is looking for locations to open a chain of stores in San Francisco. The client has following criteria for the locations:

- In areas surrounded by restaurants, bars, cafés, shopping malls, etc. to get more visitors.
- Must be in safe areas as the inventory value in the store is very high
- Near high-income neighborhoods

This paper will analyze the relevant data and suggest locations which meet the client's criteria.

2. Data Description and Preparation

The relevant data is collected based on the criteria above:

- **List of neighborhoods**

  I download the geospatial data from DataSF [2] – an open data portal of San Francisco's government. I extract the list of the neighborhood from the geospatial JSON file.

- **Geographical coordinates of the neighborhoods**

  I use the library **geopy.geocoders** in Python to obtain the geographical coordinates of each neighborhood

- **Venues**

  I use **Foursquare API** to get 100 venues and their geographical coordinates within a radius of 500 meters from each neighborhood. In total, there are 2,014 venues, grouped in 290 unique categories.

- **Crime data**

  I download the San Francisco's "Police Incidents Report from 2018 to present" [3] from DataSF. There are 24,669 incidents to be consider. Each incident is reported with a Police district where it happened.

	Incident Category
Larceny Theft	111372
Other Miscellaneous	26399
Non-Criminal	21712
Malicious Mischief	21452
Assault	20773
Burglary	16368
Motor Vehicle Theft	13219
Warrant	12503
Lost Property	12421
Recovered Vehicle	10505
Fraud	10326
Drug Offense	8850
Robbery	8301

- **Income**

  I obtain the income (in thousand USD) from the report "Median household income by neighborhood" [4] by Statisticalatlas. Unit is thousand USD.

	Neighborhood	Income
0	Balboa Terrace	217.7
1	Sea Cliff	216.5
2	Monterey Heights	214.2
3	Forest Hill	207.4
4	St. Francis Wood	188.4

3. Methodology

First, using the package in Python, I could generate the list of 41 neighborhoods in San Francisco from geospatial JSON file. Below are a few neighborhood names:

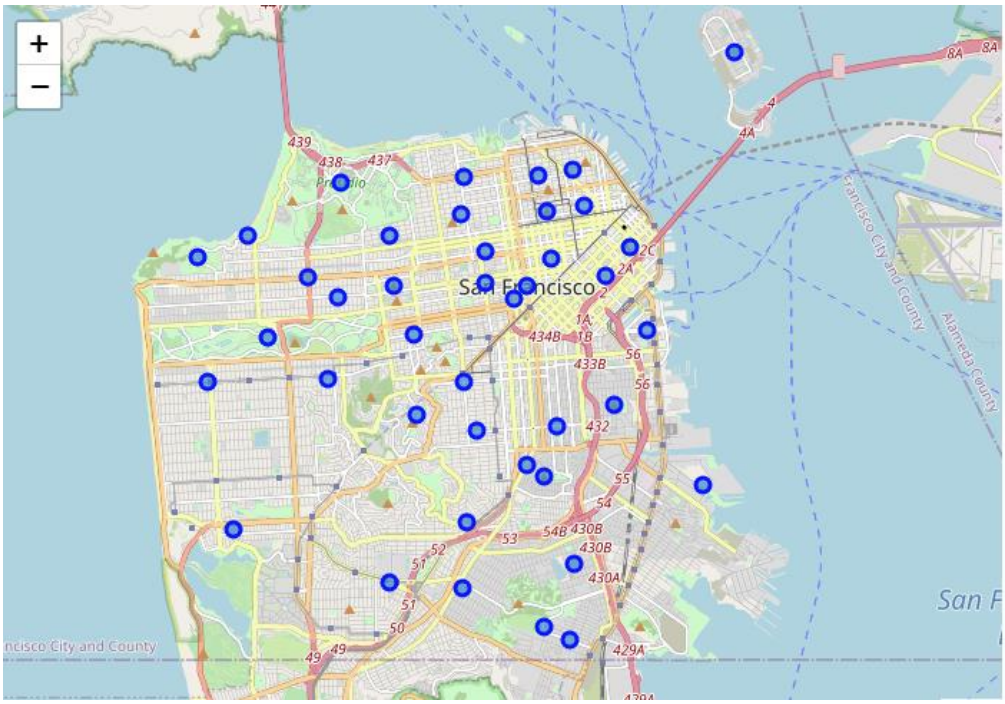
	Neighborhood
0	Bayview Hunters Point
1	Bernal Heights
2	Castro/Upper Market
3	Chinatown
4	Excelsior

Some names are modified to be able to get the geographical coordinates later; however, the new names should not lead to wrong results.

Second, I use the library **geopy.geocoders** in Python to obtain the geographical coordinates of each neighborhood. I add the geographical coordinates into the table of neighborhoods:

	Neighborhood	Latitude	Longitude
0	Bayview Hunters Point	37.7413	-122.378
1	Bernal Heights	37.743	-122.416
2	Castro/Upper Market	37.7609	-122.435
3	Chinatown	37.7943	-122.406
4	Excelsior	37.7218	-122.435

Then, I visualize the neighborhoods on the map of San Francisco thanks to the **folium** library in Python.



Third, with geographical coordinates for each neighborhood, I utilize **Foursquare API to** get 100 venues and their geographical coordinates within a radius of 500 meters from each neighborhood.

In total, there are 2,014 venues, grouped in 290 unique categories. For instance, the following table shows in Bayview Hunters Point neighborhoods, there are Heron's Head Park, Speakeasy Ales & Lagers Brewery, USPS Cafeteria, etc., together with their venue latitude and longitude.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bayview Hunters Point	37.741286	-122.377633	Heron's Head Park	37.739663	-122.375904	Park
1	Bayview Hunters Point	37.741286	-122.377633	Speakeasy Ales & Lagers	37.738468	-122.380874	Brewery
2	Bayview Hunters Point	37.741286	-122.377633	Bay Natives Nursery	37.740532	-122.376845	Garden Center
3	Bayview Hunters Point	37.741286	-122.377633	Hunter's Point Shoreline	37.738240	-122.376753	Waterfront
4	Bayview Hunters Point	37.741286	-122.377633	USPS Cafeteria	37.740744	-122.382309	Café

Then, I summarize the 10 most common venue categories in each neighborhood in the following table:

	Neighborhood	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	Bayview Hunters Point	Brewery	Waterfront	Café	Food Truck	Restaurant	Park	Garden Center	Fried Chicken Joint	French Restaurant	Frozen Yogurt Shop
1	Bernal Heights	Coffee Shop	Italian Restaurant	Grocery Store	Mexican Restaurant	Playground	Yoga Studio	Gourmet Shop	Food Truck	Bakery	Cocktail Bar
2	Castro/Upper Market	Gay Bar	Thai Restaurant	Coffee Shop	Yoga Studio	Indian Restaurant	Gym	Mediterranean Restaurant	New American Restaurant	Deli / Bodega	Pet Store
3	Central Richmond	Sushi Restaurant	Chinese Restaurant	Korean Restaurant	Massage Studio	Pizza Place	Bar	Coffee Shop	Skate Park	Gift Shop	Breakfast Spot
4	Chinatown	Italian Restaurant	Chinese Restaurant	Coffee Shop	New American Restaurant	Cocktail Bar	Bakery	Szechuan Restaurant	Hotel	Tea Room	Wine Bar

We can see many neighborhoods having the same common venue categories. Therefore, I use K-means algorithm – an unsupervised machine learning algorithm to cluster the neighborhoods into four clusters. K-means is one of the most popular method for clustering. By clustering the neighborhoods, I expect to eliminate the neighborhoods which do not seem to be commercial areas.

***Fourth*,** I check if the commercial neighborhood is a safe place to open a jewelry store. Using the Police Incidents Report from 2018 to present, I select the incidents which belongs to Burglary and Robbery categories. In result, there are 24,669 incidents, in which 16,368 is of burglary and 8,301 is of robbery. Each incident is reported with a Police district where it happened, so I can plot the incidents on the choropleth map. Consequently, I can eliminate the neighborhoods which are in the most unsafe areas.

Last, I use income data to narrow down to the neighborhoods which have high-income earners who may be potential customer of luxury jewelries.

4. Result and Discussion

The result of clustering neighborhoods shows that:

- **Cluster 0:** includes the neighborhood Seacliff where playground, beach, and fountain are the most popular venues.

	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
23	Seacliff	37.7885	-122.487	0	Playground	Beach	Fountain	Flower Shop	Food	Food & Drink Shop	Food Court	Food Stand	Food Truck	Yoga Studio

- **Cluster 1:** includes 32 neighborhoods which have a lot of coffee shops, restaurants, bars, etc.

	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Bernal Heights	37.743	-122.416	1	Coffee Shop	Mexican Restaurant	Playground	Grocery Store	Trail	Italian Restaurant	Food Truck	American Restaurant	Vietnamese Restaurant	Bakery
2	Castro/Upper Market	37.7609	-122.435	1	Gay Bar	Coffee Shop	Thai Restaurant	Yoga Studio	Gym	Convenience Store	Cosmetics Shop	Pet Store	Deli / Bodega	New American Restaurant
3	Chinatown	37.7943	-122.406	1	Chinese Restaurant	Italian Restaurant	Coffee Shop	Cocktail Bar	New American Restaurant	Bakery	Hotel	Szechuan Restaurant	Tea Room	Dive Bar
4	Excelsior	37.7218	-122.435	1	Mexican Restaurant	Bakery	Convenience Store	Pizza Place	Bank	Vietnamese Restaurant	Latin American Restaurant	Sandwich Place	Chinese Restaurant	Thai Restaurant
5	Financial District/South Beach	37.7866	-122.395	1	Coffee Shop	Café	Food Truck	Juice Bar	Bar	Seafood Restaurant	Salad Place	Cocktail Bar	Brazilian Restaurant	Gym / Fitness Center

- **Cluster 3:** includes 6 neighborhoods which has places for outdoor activities such as parks, scenic lookouts, lakes, some cafés, and shops

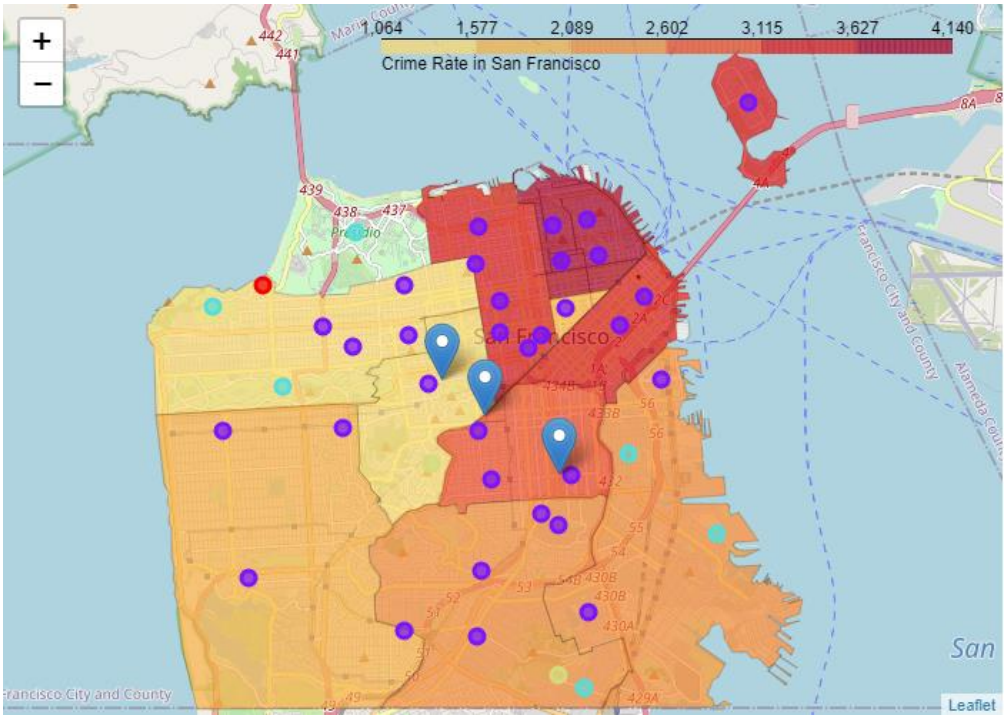
	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bayview Hunters Point	37.7413	-122.378	2	Brewery	Restaurant	Food Truck	Café	Park	Garden Center	Waterfront	Discount Store	Garden	Furniture / Home Store
8	Golden Gate Park	37.7694	-122.482	2	Park	Music Venue	Lake	BBQ Joint	Waterfall	Bus Stop	Harbor / Marina	Dog Run	Sculpture Garden	Track
16	Lincoln Park	37.7846	-122.499	2	Scenic Lookout	Trail	Monument / Landmark	Cafeteria	Park	Café	Sculpture Garden	Gift Shop	Golf Course	Art Museum
33	Potrero Hill	37.7566	-122.399	2	Grocery Store	Park	Deli / Bodega	Trail	Liquor Store	Hill	Playground	Plaza	Cosmetics Shop	Convenience Store
34	Presidio	37.7987	-122.465	2	Brewery	Art Gallery	Outdoor Sculpture	Trail	Tunnel	Bowling Alley	Park	General Entertainment	American Restaurant	Frozen Yogurt Shop
38	Visitacion Valley	37.7121	-122.41	2	Park	Vietnamese Restaurant	Chocolate Shop	Grocery Store	Yoga Studio	Food Stand	Flea Market	Flower Shop	Food	Food & Drink Shop

- **Cluster 4:** includes 2 neighborhoods where there are parks, trails, and scenic lookouts

	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
13	McLaren Park	37.7146	-122.416	3	Park	Scenic Lookout	Pool	Trail	Playground	Yoga Studio	Food Court	Flea Market	Flower Shop	Food
37	Twin Peaks	37.7546	-122.446	3	Trail	Scenic Lookout	Bus Station	Bus Stop	Reservoir	Hill	Tailor Shop	Frame Store	French Restaurant	Fountain

Comparing four clusters, Cluster 1 seems to have the most commercial areas which are suitable to open a jewelry store.

Next, I visualize all neighborhoods of four clusters on the crime map to eliminate the neighborhoods in unsafe areas.



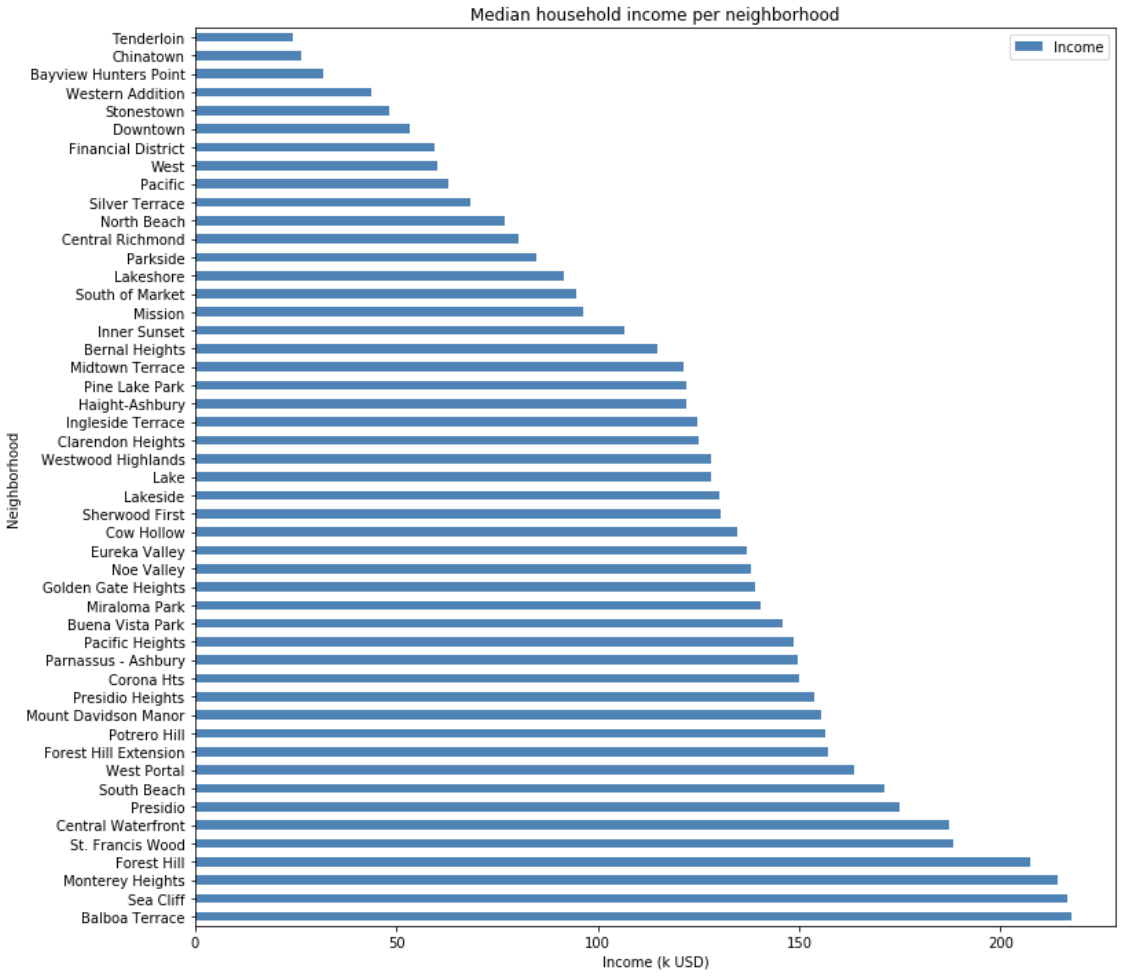
In the map above, neighborhoods of Cluster 1 are the blue circles. I can eliminate those in the red-shaded areas where burglary and robbery are more than 2,600 cases. In addition, I pin three existing Jewelry stores in the city. They also seem to avoid the unsafe areas.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
134	Castro/Upper Market	37.760856	-122.434957	D&H Sustainable Jewelry	37.763732	-122.433601	Jewelry Store
520	Haight Ashbury	37.770015	-122.446952	Braindrops	37.770402	-122.444094	Jewelry Store
1163	Mission	37.752498	-122.412826	Luz de Luna	37.752523	-122.415997	Jewelry Store

After this elimination, I shorten the list of potential neighborhoods as following:

Neighborhood
Bernal Heights
Excelsior
Glen Park
Richmond
Haight Ashbury
Inner Sunset
Lakeshore
Lone Mountain
Mission Bay
Ingleside
Sunset
Outer Mission
Outer Richmond
Portola
Presidio Heights

The last criterion is income. I plot the median household income per neighborhood on a bar chart.



I target the neighborhoods where median household income is more than $100,000 per year and combine with the list of neighborhoods above, I could identify five neighborhoods which meet the client's criteria:

Neighborhood
Haight Ashbury
Inner Sunset
Lakeshore
Ingleside
Presidio Heights

5. Conclusion

In this paper, I try to solve a business problem like a real scientist would have: Where to open a jewelry store in San Francisco, given certain criteria. I have used many Python libraries and functions to collect and analyze data such as locations, venues, crime data, and income data. Some highlighted techniques used in this analysis are:

- Foursquare API to explore the venues around the neighborhoods
- Python visualization libraries such as seaborn, matplotlib, and folium to plot the charts and maps
- K-means, an unsupervised machine learning algorithm, to cluster the neighborhoods

As a result, I could narrow down to a list of five most potential neighborhoods satisfying the client's criteria.

One drawback of using Foursquare API is that the result does not mean it inquires all the venues in the neighborhoods. Instead, it depends pretty much on the geographical coordinates, which is only one pair of latitude and longitude, the parameter of radius, and the database of Foursquare. However, it can be improved by adding more geographical information to the model.

6. References

[1] https://en.wikipedia.org/wiki/San_Francisco

[2] https://data.sfgov.org/api/geospatial/p5b7-5n3h?method=export&format=GeoJSON

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

[4] https://statisticalatlas.com/neighborhood/California/San-Francisco/Financial-District/Household-Income