CAPSTONE PROJECT – THE BATTLE OF NEIGHBORHOODS

Jayrajsinh Sisodiya

MARCH 11, 2020
IBM DATA SCIENCE
Coursera

1. Introduction

1.1 Background

The city of San Francisco is one of the biggest cities in the United States. The city is diverse in culture and is considered the financial capital of the west coast. There is an enormous number of opportunities for business in San Francisco that has attracted many people to start a business. Besides, the city is also considered to be a hub for banking, finance, retailing, transportation, tourism, tourism, real estate, advertising and arts in the United States. Thus, it is highly competitive and risky to start your own business in San Francisco. Besides, it requires a large investment to start the business as San Francisco is a developed city. Thus, any new business or expansion needs careful consideration and analysis. The business insights generated through research and analysis will assist a business owner to strategize based on the market conditions. Later, it can help in risk reduction and can help you to generate a high return on the investment.

1.2 Business Problem

San Francisco is one of the highest visited cities in the United States. In 2018, there were approx. 26 million travelers across the country and the world, and this number have been growing for the last 9 years. Besides, the total spending by visitors was $10 billion that assisted in creating over 80,000 jobs in San Francisco. The hotels and Airbnb have an average occupancy of more than 82 % in 2018 with the average daily rates of $265 which is also expected to grow soon. All these statistics make San Francisco a great place for tourism and an opportunity for those who want to start their business in the tourism industry.

My client is willing to start a new hotel in San Francisco and the client wants to find the optimal location for the hotel. In the project, we will find the potential neighborhood based on the number of Airbnb properties in each neighborhood. Besides, the client's primary focus is to find a neighborhood that has a moderate number of Airbnb properties. The reason behind the condition is that the client wants a hotel location that neither has high competition from Airbnb nor he wants a hotel location with less number of Airbnb properties where there is a low return of investment. Overall, there is a great opportunity for our client in San Francisco and our job is to find the hotel location that attracts a high number of customers with a high return on investment.

For the scope of this project, we will only consider the competition with Airbnb properties and keep other hotel's information out of the scope.

1.3 Interest

This project is for all those business owners who want to start their new hotel business in San Francisco and exploring the neighborhoods of San Francisco with common venues around.

2. Data Acquisition

2.1 Data Acquisition

The data acquired for this project is a combination of data from 2 sources. The first data source of the project uses San Francisco Airbnb property listed as per June 2019 that shows number of properties in each region and neighborhood. The dataset originally has 7575 number of observations with 107 columns. Initially, we will remove all the unnecessary columns. After data cleansing , our new dataset will have 7575 observation with 10 columns as follow

San Francisco Airbnb listing Data

1) Sr. No.: Serial Number

2) id: Property ID

3) name: Property Name

4) host_is_superhost: Whether the host is super host or not

5) neighborhood: Neighborhood of the property

6) latitude: Latitude of the property

7) longitude: Longitude of the property

8) price in USD: Average price of the property for a day

9) guest_included: Number of guests allowed for a price

10) Region: Region of the property in San Francisco

The dataset has 5 regions and 55 neighborhoods. The San Francisco Airbnb dataset can be found from the following link:

Dataset URL: https://data.world/ajsanne/sf-airbnb-listings

The 2nd Dataset is the geographical co-ordinates data of San Francisco will be used for input from the Foursquare API which will be leveraged to provide venue information for the neighborhood. The Foursquare API will be used to explore the neighborhoods in San Francisco city.

3. Methodology

3.1 Exploratory Data Analysis

3.1.1 Number of properties in each region & neighborhood

The value_count() function is used to observe number of observations in each category. In our case, the function will return the number of properties in each region and number of properties in each neighborhood. Initially, we will observe number of properties in each region. *(Fig – 1)*

```
Southern              3302
Western Addition      1805
Downtown              1275
Outside Lands          659
North of Downtown      534
Name: Region, dtype: int64
```

Fig 1: No. of Properties in Each Region

Comparing all the regions with the number of properties based on June 2019 data, it is evident that "Southern" region has the highest number of properties followed by "Western Addition", "Downtown", "Outside Land" & "North of Downtown". *(Fig - 2)*

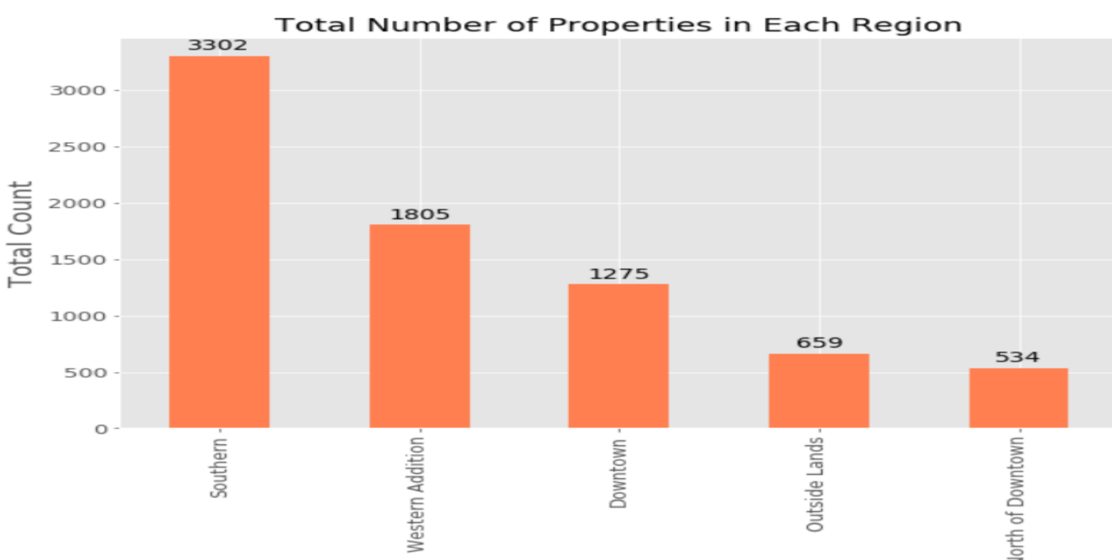

Fig 2 – Total number of properties in each region

Below is the figure that demonstrate the number of properties in each neighborhood.

```
Out[100]:  Mission District              722
           SoMa                          526
           Western Addition/NOPA         438
           Bernal Heights                394
           Noe Valley                    364
           Richmond District             362
           Outer Sunset                  361
           Downtown                      347
           The Castro                    313
           Nob Hill                      309
           Haight-Ashbury                210
           Potrero Hill                  204
           Pacific Heights               179
           Bayview                       170
           Inner Sunset                  150
           Union Square                  135
           Telegraph Hill                134
           Excelsior                     127
           Cole Valley                   125
           Duboce Triangle               115
           Russian Hill                  108
           Tenderloin                    108
           Marina                        105
           Crocker Amazon                103
           South Beach                   100
           Sunnyside                      95
           Lower Haight                   89
           Hayes Valley                   89
           Glen Park                      71
           Mission Terrace                70
           Twin Peaks                     67
           Chinatown                      66
           Financial District             65
           Alamo Square                   63
           Cow Hollow                     62
           Visitacion Valley              59
           Lakeshore                      53
           Ingleside                      51
           Fisherman''s Wharf             47
           Portola                        46
           Balboa Terrace                 45
           Oceanview                      44
           Parkside                       44
           Mission Bay                    44
           North Beach                    42
           Dogpatch                       36
           Presidio Heights               31
           Civic Center                   25
           Diamond Heights                19
           Forest Hill                    15
           West Portal                    11
           Daly City                       6
           Japantown                       6
           Sea Cliff                       3
           Fisherman''''s Wharf            1
           Presidio                        1
           Name: Neighborhood, dtype: int64
```

Fig 3 – Number of Properties in each neighborhood

3.1.2 Number of properties in Neighborhood

As per our client's primary focus is to find a neighborhood that has a moderate number of Airbnb properties. The reason behind the condition is that the client wants a hotel location that neither has high competition from Airbnb nor he wants a hotel location with a smaller number of Airbnb properties where there is a low return of investment. Therefore, we will select “Downtown” region for our client where there is a moderate competition with 1275 Airbnb properties as per Fig – 2. Following is the figure that visualize number of properties in San Francisco downtown. *(See Fig – 4)*

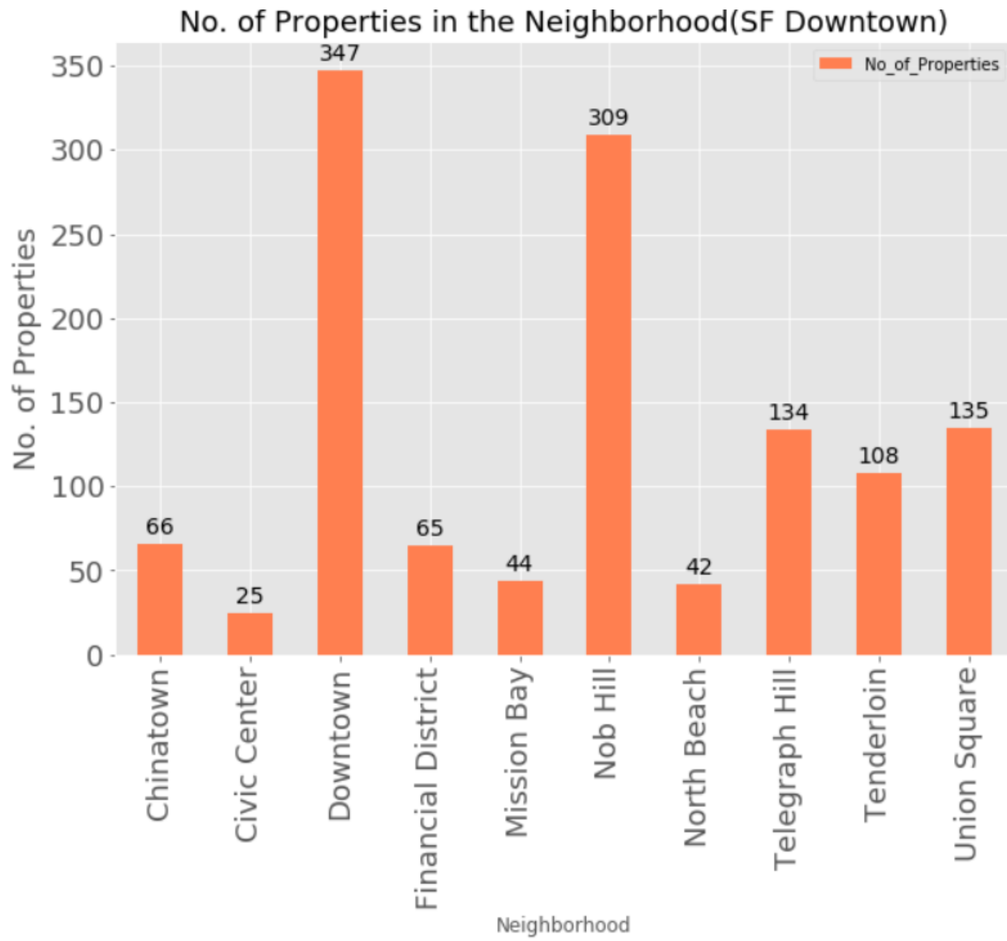

Fig 4 – No. of properties in San Francisco Downtown's Neighborhood

3.1.3 Neighborhoods in San Francisco Downtown

There are 9 neighborhoods in San Francisco Downtown, and they are visualized on a map using folium library from python.

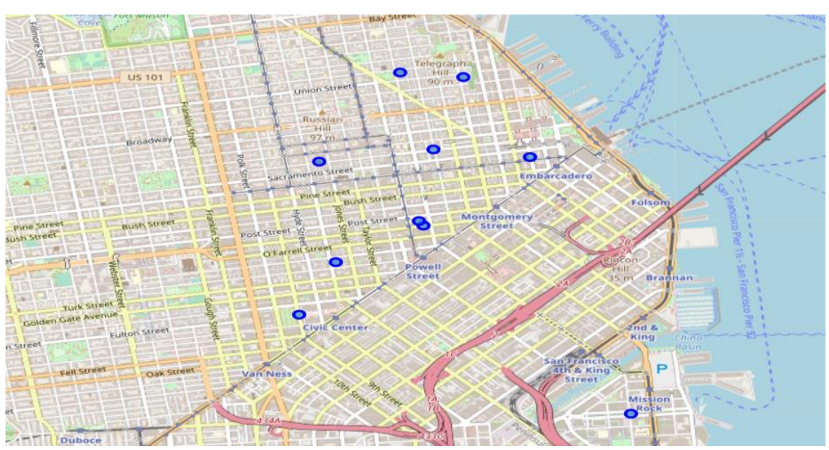

Fig 5 – Neighborhood in San Francisco Downtown

3.2 Modelling

Using final dataset that contains the neighborhoods of San Francisco downtown with respective latitude and longitude and we can find the venues within a 500-meter radius with a limit of 100 by connecting Foursqaure API. The command will return a json file with all venues in each neighborhood which is converted to pandas dataframe. Following figure has all venues with co-ordinates and category. *(See fig – 6)*

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Chinatown	37.794301	-122.406376	Blue Bottle Coffee	37.792771	-122.404833	Coffee Shop
1	Chinatown	37.794301	-122.406376	Hinodeya	37.794656	-122.404544	Ramen Restaurant
2	Chinatown	37.794301	-122.406376	Red Blossom Tea Company	37.794643	-122.406379	Tea Room
3	Chinatown	37.794301	-122.406376	Chapel Hill Coffee Co.	37.794041	-122.404247	Coffee Shop
4	Chinatown	37.794301	-122.406376	Mister Jiu's	37.793790	-122.406615	Chinese Restaurant

Fig 6 – Venue details of each neighborhood

One hot encoding is done on the venues data. For categorical variables where no such ordinal relationship exists, the integer encoding is not enough. In fact, using this encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results (predictions halfway between categories).In this case, a one-hot encoding can be applied to the integer representation. This is where the integer encoded variable is removed and a new binary variable is added for each unique integer value. The data of venues if grouped by neighborhood and the mean of venues are calculated and finally 10 common venues are calculated for each neighborhood.

To help business owners find similar neighborhood in the moderate competition, we will utilize K-means clustering algorithm which is a unsupervised machine learning algorithm that clusters data based on the predefined cluster size. Later, we will cluster 9 neighborhoods into 5 cluster. The reason for K-means clustering is to cluster neighborhoods with similar venues together so that business owners can find location for the hotel based on the preferences.

4. Results

After running K-means clustering, we can observe cluster created to see which neighborhoods are assigned to the clusters. First cluster has the following neighborhoods (See fig 7)

	Neighborhood	Region	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	Downtown	Downtown	37.787514	-122.407159	0	Boutique	Jewelry Store	Clothing Store	Hotel	Theater
9	Union Square	Downtown	37.787936	-122.407517	0	Boutique	Hotel	Jewelry Store	Clothing Store	Theater

Fig 7 – Cluster 1

The cluster 1 has 2 neighborhoods out of 9. The cluster consist common venue such as Boutique, Jewelry Store, Hotel, Clothing Store, Jewelry Store, Theater etc.

Looking into other neighborhoods, we can see some of the cluster has only 1 neighborhood which means that these neighborhoods has unique venues and they can not be clustered into similar neighborhoods. (See fig 8,9,10 & 11)

	Neighborhood	Region	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
5	Nob Hill	Downtown	37.793262	-122.415249	1	Italian Restaurant	Hotel	Bar	Café	Coffee Shop

Fig 8 – Cluster 2

The cluster 2 has one neighborhood that consist venues such as Italian Restaurant, Hotel, Bar, Cafe, Coffee Shop.

	Neighborhood	Region	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
6	North Beach	Downtown	37.801175	-122.409002	2	Italian Restaurant	Pizza Place	Bakery	Café	Cocktail Bar
7	Telegraph Hill	Downtown	37.800785	-122.404091	2	Italian Restaurant	Pizza Place	Cocktail Bar	Café	Coffee Shop

Fig 9 – Cluster 3

The cluster 3 has 2 neighborhoods out of 9 and cluster consist of venues such as Italian restaurant, Pizza place, bakery, cocktail bar and Cafe etc.

	Neighborhood	Region	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Chinatown	Downtown	37.794301	-122.406376	3	Coffee Shop	Chinese Restaurant	Bakery	New American Restaurant	Men's Store
1	Civic Center	Downtown	37.779594	-122.416794	3	Café	Coffee Shop	Cocktail Bar	Vietnamese Restaurant	Beer Bar
3	Financial District	Downtown	37.793647	-122.398938	3	Coffee Shop	Food Truck	Café	Gym	Salad Place
8	Tenderloin	Downtown	37.784249	-122.413993	3	Vietnamese Restaurant	Coffee Shop	Thai Restaurant	Theater	Cocktail Bar

Fig 10 – Cluster 4

The cluster 4 has the highest number of neighborhoods in the cluster. It has 4 neighborhoods out of 9 and cluster consist of venues such as coffee shop, Cafe, Vietnamese restaurant, Chinese Restaurant, Food Truck, Bakery, gym, Beer Bar etc.

	Neighborhood	Region	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
4	Mission Bay	Downtown	37.770774	-122.391171	4	Food Truck	Gym	Coffee Shop	Pharmacy	Park

Fig 11 – Cluster 5

Cluster 5 has one neighborhood which has venues such as food truck, gym, coffee shop, pharmacy and park etc.

Visualizing the clustered neighborhoods on a map using folium library. *(See fig – 12)*

Each cluster is color coded for our ease. We can observe majority of the neighborhoods are in the “Light Green” clusters which is the 4th cluster. 2 of the neighborhoods have their own cluster and these are cluster 2 & cluster 5 colored “Blue” & “Orange” respectively. Whereas Cluster 1 is colored red that has 2 neighborhoods and Cluster 3 is colored pink that has 2 neighborhoods as well.

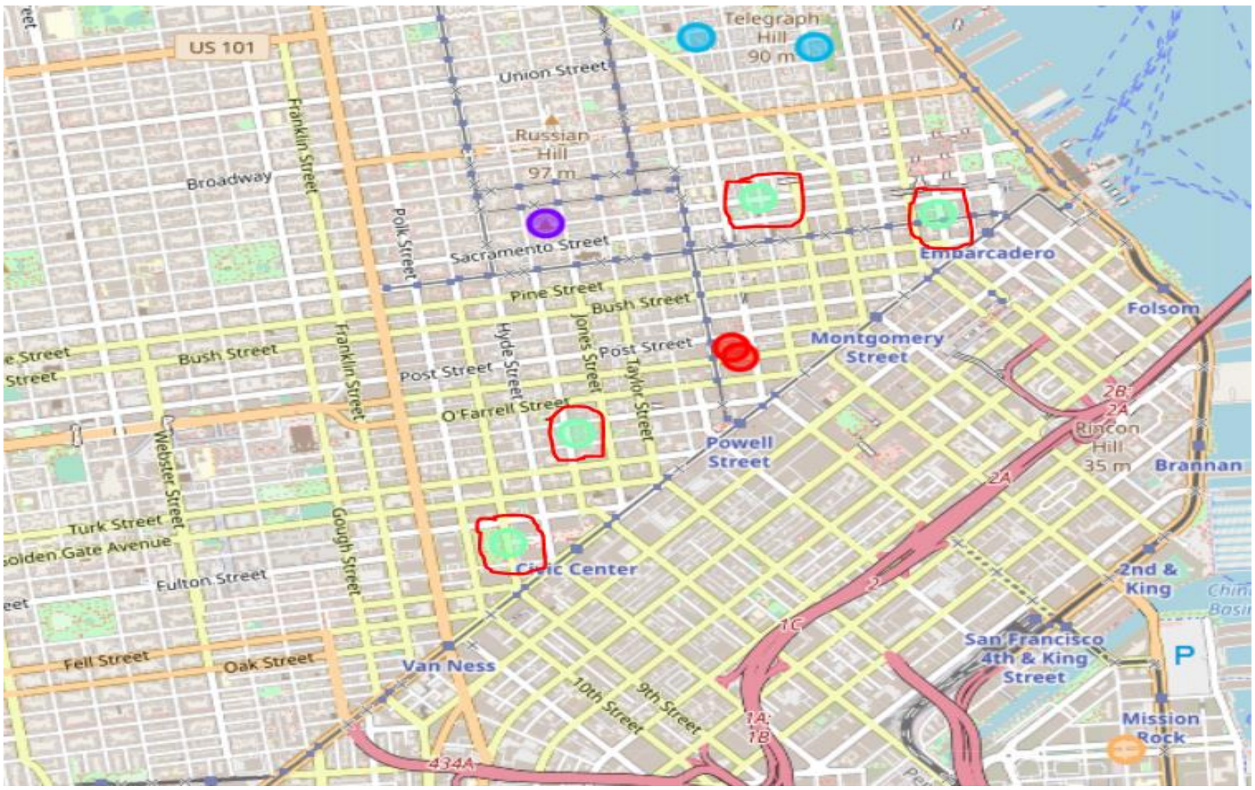

Fig 12 – Clustered neighborhoods in San Francisco Downtown

5. Discussion

The aim of the project is to help business owners to find the best place in San Francisco downtown. The business owners can choose the location based on the most common venues around. As an example, if a business owner wants a hotel that has shopping venues around such as boutique, jewelry store, clothing store then they can choose a neighborhood from cluster 1. In cluster 2,3 & 4 most of the places are food places. Such as if business owner wants a hotel near Italian Restaurant, Hotel, Bar, Cafe, Coffee Shop then he/she would choose cluster 2. The neighborhood in cluster 3 has proximity to Italian restaurant, Pizza place, bakery, cocktail bar and Café. Cluster 4 that has the highest neighborhoods and scattered based on Fig – 12, the business owner will choose such neighborhood if the preference is in favor of coffee shop, Cafe, Vietnamese restaurant, Chinese Restaurant, Food Truck, Bakery, gym, Beer Bar. Business owner will choose cluster 5 where there is only one neighborhood if he/she wants a hotel near places such as food truck, gym, coffee shop, pharmacy and park.

6. Conclusion

This project aims to help business owners to have a better understanding of neighborhoods in comparison with most common places around neighborhood. It is essential to use the technology to gain advantage in business such as knowing more about location before starting the business in the region. In this project, the competition with Airbnb properties has been considered. The future scope of this project can include the competition with other hotel, price offered by other business owner, safety in the neighborhood.