Capstone Project - The Battle of the Neighborhoods

1. Introduction

1.1. Description and Discussion of the Background

San Francisco, is the cultural, commercial, and financial center of Northern California. San Francisco is the 15th most populous city in the United States, and the fourth most populous in California, with 881,549 residents as of 2019.

San Francisco is the 12th-largest metropolitan statistical area in the United States by population, with 4.7 million people, and the fourth-largest by economic output, with GDP of $549 billion in 2018.

San Francisco is one of the important cities in the United States and where there is a high population density. Because of this, there are investors who want to invest their capital in cities like this. But one of the problems when starting a new project is the little information they have about the environment. This project will help in some way to locate his project in the neighborhoods of San Francisco. We will also see what businesses already exist in each neighborhood and thus guide investors in their new project.

1.2. Data Description

- To get the information of the neighborhoods of San Francisco I scraped this web page, http://www.healthysf.org/bdi/outcomes/zipmap.htm.

- To get the latitude and longitude of each neighborhood I used a library called uszipcode. This library helps you obtain latitude and longitude from the zip code of the neighborhood.

- I used Foursquare API to get the venues of every neighborhood in San Francisco.

2. Methodology

First, we are going to scrape the page indicated above to obtain the neighborhoods of San Francisco. After cleaning the data, we obtain the following table where we see the neighborhoods of San Francisco with its zip code. In total there are 20 neighborhoods.

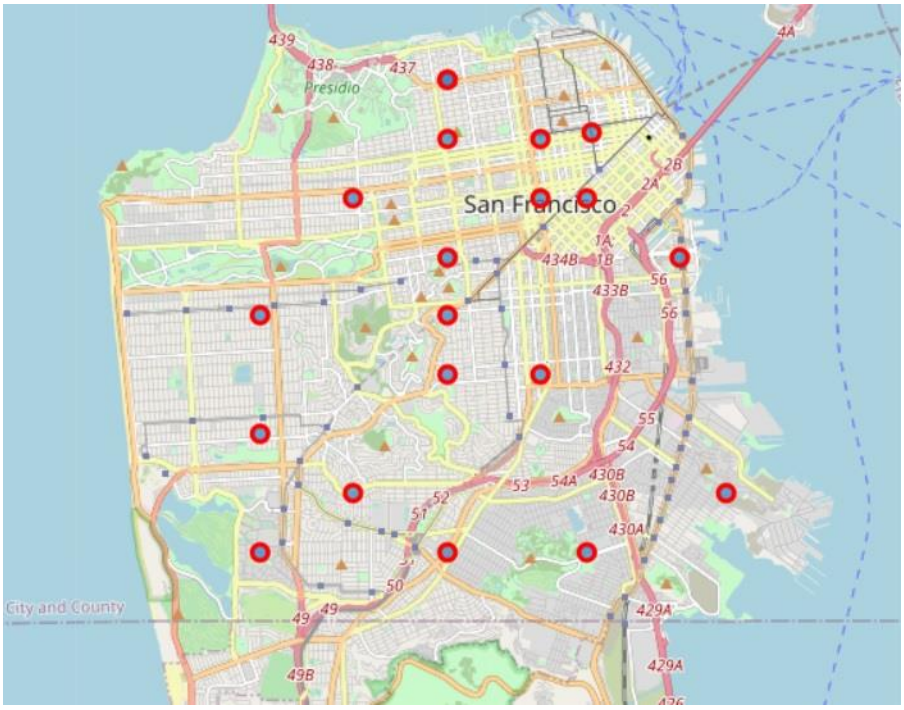
	Zip Code	Neighborhood
0	94102	Hayes Valley/Tenderloin/North of Market
1	94103	South of Market
2	94107	Potrero Hill
3	94108	Chinatown
4	94109	Polk/Russian Hill (Nob Hill)
5	94110	Inner Mission/Bernal Heights
6	94112	Ingelside-Excelsior/Crocker-Amazon
7	94114	Castro/Noe Valley
8	94115	Western Addition/Japantown
9	94116	Parkside/Forest Hill
10	94117	Haight-Ashbury
11	94118	Inner Richmond

To obtain the latitude and longitude I use a library that I found on the web called uszipcode, this library works as a database that through the zip code you can obtain the latitude and longitude of each neighborhood of San Francisco or another city in the United States.

After obtaining the latitude and longitude through the library we have the following table.

	Zip Code	Neighborhood	Latitude	Longitude
0	94102	Hayes Valley/Tenderloin/North of Market	37.780	-122.420
1	94103	South of Market	37.780	-122.410
2	94107	Potrero Hill	37.770	-122.390
3	94108	Chinatown	37.791	-122.409
4	94109	Polk/Russian Hill (Nob Hill)	37.790	-122.420
5	94110	Inner Mission/Bernal Heights	37.750	-122.420
6	94112	Ingelside-Excelsior/Crocker-Amazon	37.720	-122.440
7	94114	Castro/Noe Valley	37.760	-122.440
8	94115	Western Addition/Japantown	37.790	-122.440
9	94116	Parkside/Forest Hill	37.740	-122.480
10	94117	Haight-Ashbury	37.770	-122.440
11	94118	Inner Richmond	37.780	-122.460
12	94122	Sunset	37.760	-122.480
13	94123	Marina	37.800	-122.440
14	94124	Bayview-Hunters Point	37.730	-122.380

To show the neighborhoods we are using the library learned in the course, the folium library, on the map you can see the city of San Francisco with the markers in each neighborhood. Altogether we show twelve neighborhoods in San Francisco.

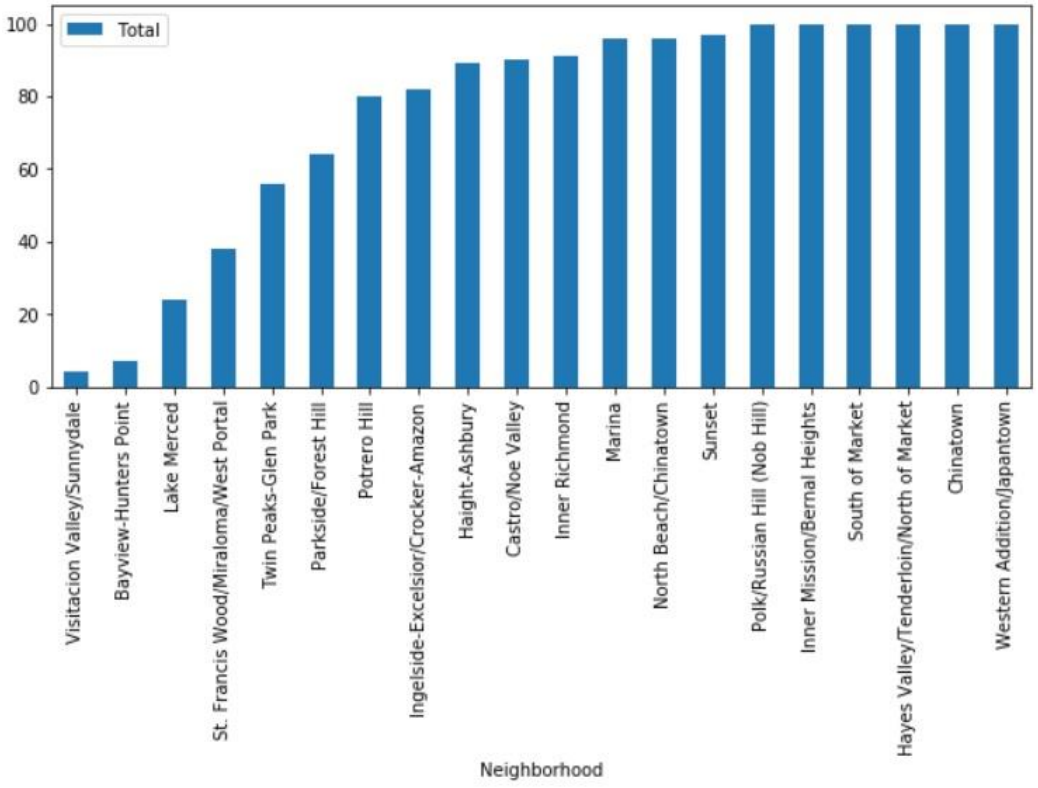


I used Foursquare API to get the venues of each neighborhood. I configure the search for a maximum of 100 venues for each neighborhood and a radius of 700 meters. These are the parameters that we must send to Foursquare apart from the latitude and longitude that we already have. In this table we see a part of the venues that Foursquare returns.

Venue	Venue Latitude	Venue Longitude	Venue Category
Herbst Theater	37.779548	-122.420953	Concert Hall
War Memorial Opera House	37.778601	-122.420816	Opera House
San Francisco Ballet	37.778580	-122.420798	Dance Studio
Louise M. Davies Symphony Hall	37.777976	-122.420157	Concert Hall
War Memorial Court	37.779042	-122.420971	Park

Now let's look at a bar chart to see how the places are distributed by each neighborhood of San Francisco. We see that Japantown, Chinatown, Hayes Valley, South of Market, Bernal Heights, and Russian Hill have the most venues in San Francisco.

While Visitation Valley, Bayview-Hunters Point and Lake Merced have the fewest venues in San Francisco. More venues could appear in each neighborhood if we increase the maximum number of places and the search radius.

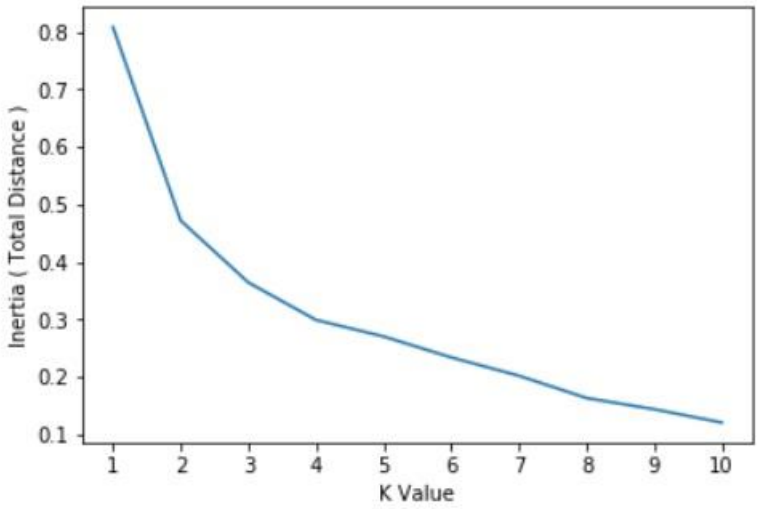


Let's see what categories the places in each neighborhood have. In the result that Foursquare returns we have a total of 244 categories, this is normal since San Francisco has a very dense population. Now I show you a part of the table that contains the 10 most common venues in each neighborhood.

	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	Motorcycle Shop	Coffee Shop	Spa	Food & Drink Shop	Deli / Bodega	Park	Home Service	Business Service	Farmers Market	Fast Food Restaurant
1	Castro/Noe Valley	Gay Bar	Park	Thai Restaurant	Coffee Shop	Scenic Lookout	Trail	Cosmetics Shop	Playground	Monument / Landmark	Indian Restaurant
2	Chinatown	Hotel	Coffee Shop	Cocktail Bar	Bubble Tea Shop	Art Gallery	Gym / Fitness Center	Italian Restaurant	Boutique	Spa	Speakeasy
3	Haight-Ashbury	Coffee Shop	Park	Scenic Lookout	Dog Run	Ice Cream Shop	Boutique	Liquor Store	Mediterranean Restaurant	Tennis Court	Thrift / Vintage Store
4	Hayes Valley/Tenderloin/North of Market	Coffee Shop	Theater	Thai Restaurant	Clothing Store	Wine Bar	Sandwich Place	Beer Bar	Cocktail Bar	Southern / Soul Food Restaurant	Juice Bar

We are going to group each neighborhood taking into account the similarity of venues that each one presents. For this we are going to use one of the techniques we learned in this specialization, I mean K-Means.

Let's first see how many clusters are needed. For this we are going to see a curve that gives us an idea of how many groups we need. The curve does not change much in the values of k equal 5 or 6. For this reason, for the number of clusters we are going to choose a value of k equal to 5.



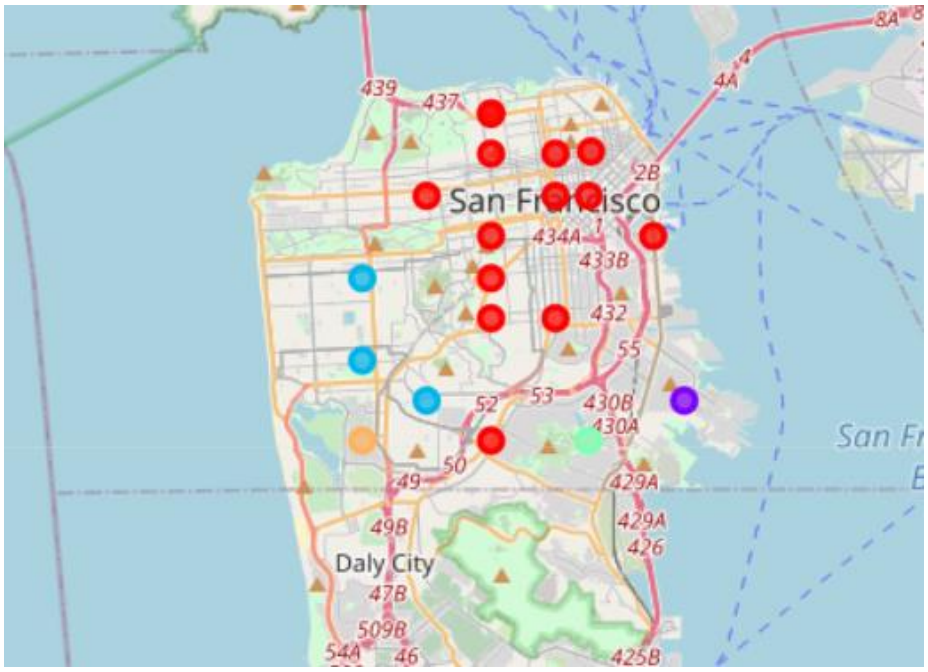
After using K-Means with k equal to 5, to a clean table that we make in this process we add the labels that the algorithm returned to us.

	Zip Code	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
0	94102	Hayes Valley/Tenderloin/North of Market	37.780	-122.420	0	Coffee Shop	Theater	Clothing Store	Sandwich Place	Wine Bar
1	94103	South of Market	37.780	-122.410	0	Coffee Shop	Bakery	Marijuana Dispensary	Beer Bar	Sandwich Place
2	94107	Potrero Hill	37.770	-122.390	0	Food Truck	Gym	Park	Coffee Shop	Café
3	94108	Chinatown	37.791	-122.409	0	Hotel	Coffee Shop	Gym / Fitness Center	Cocktail Bar	Bubble Tea Shop
4	94109	Polk/Russian Hill (Nob Hill)	37.790	-122.420	0	Cocktail Bar	Grocery Store	Coffee Shop	Italian Restaurant	Mexican Restaurant
5	94110	Inner Mission/Bernal Heights	37.750	-122.420	0	Mexican Restaurant	Grocery Store	Coffee Shop	Café	Cocktail Bar
6	94112	Ingelside-Excelsior/Crocker-Amazon	37.720	-122.440	0	Pizza Place	Bus Station	Bakery	Mexican Restaurant	Latin American Restaurant

We see that in some neighborhoods in the first places we see a coffee shop or food stores, this should not surprise us because in most densely populated cities this pattern is repeated.

3. Result

You can also see a clustered map boroughs of San Francisco in the below. As you can see the majority belongs to cluster 0 with red color, they are the neighborhoods that are in the north of San Francisco. In the second place is the cluster with label 2 with color blue, are the neighborhoods that are in the west of San Francisco.



Now let's see how each neighborhood in the cluster was labeled 0. This cluster is the one with the most neighborhoods among all the clusters, what predominates most in most of these clusters are coffee shops and then restaurants with other types of shops come. This distribution in the location of venues is something that is repeated in most large cities, you can see the analysis we made of New York City in the course and you will realize what I am telling you.

	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	Hayes Valley/Tenderloin/North of Market	Coffee Shop	Theater	Clothing Store	Sandwich Place	Wine Bar	Cocktail Bar	Thai Restaurant	Beer Bar	Boutique	Dessert Shop
1	South of Market	Coffee Shop	Bakery	Marijuana Dispensary	Beer Bar	Sandwich Place	Bar	American Restaurant	Theater	Mexican Restaurant	Cocktail Bar
2	Potrero Hill	Food Truck	Gym	Park	Coffee Shop	Café	Pharmacy	Pizza Place	Harbor / Marina	Dog Run	Performing Arts Venue
3	Chinatown	Hotel	Coffee Shop	Gym / Fitness Center	Cocktail Bar	Bubble Tea Shop	Gym	Art Gallery	Spa	Boutique	Speakeasy
4	Polk/Russian Hill (Nob Hill)	Cocktail Bar	Grocery Store	Coffee Shop	Italian Restaurant	Mexican Restaurant	Wine Bar	Massage Studio	Bar	Café	Steakhouse
5	Inner Mission/Bernal Heights	Mexican Restaurant	Grocery Store	Coffee Shop	Café	Cocktail Bar	Pizza Place	Breakfast Spot	Art Gallery	Deli / Bodega	Fish Market
6	Ingelside-Excelsior/Crocker-Amazon	Pizza Place	Bus Station	Bakery	Mexican Restaurant	Latin American Restaurant	Vietnamese Restaurant	Coffee Shop	Light Rail Station	Sandwich Place	Café
7	Castro/Noe Valley	Gay Bar	Park	Coffee Shop	Scenic Lookout	Thai Restaurant	Playground	Monument / Landmark	Indian Restaurant	Hill	Grocery Store

This is our cluster with label 1 with purple color, the first place is occupied by a construction site, this is not very normal, this place must be an important site in this neighborhood. The second place is occupied by a motorcycle store, this is also not normal. The third place is occupied by a coffee shop, this is more common as we saw in the first cluster.

The other positions include fewer common venues such as a market, a spa, among others. The first places that we have mentioned in this cluster are not found in the majority in the first cluster, that is what makes it different from the first cluster.

	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
14	Bayview-Hunters Point	Construction & Landscaping	Motorcycle Shop	Coffee Shop	Spa	Food & Drink Shop	Deli / Bodega	Home Service	Park	Event Space	Farmers Market

This is our third cluster with blue color made up of only three neighborhoods. In which we see that in the three neighborhoods the first place is occupied by Chinese restaurants. In addition, other types of restaurants occupy the rest of the list. We can conclude what characterizes this cluster from the previous ones we saw is the number of restaurants from other origins.

Another characteristic of this cluster is that we also find coffee shops as in cluster 1. We must also note that these places on the map are almost together in the west of the city.

	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
9	Parkside/Forest Hill	Chinese Restaurant	Park	Sushi Restaurant	Pizza Place	Optical Shop	Dumpling Restaurant	Bubble Tea Shop	Sandwich Place	Café	Korean Restaurant
12	Sunset	Chinese Restaurant	Vietnamese Restaurant	Bubble Tea Shop	Bakery	Japanese Restaurant	Coffee Shop	Dumpling Restaurant	Thai Restaurant	Bar	Sushi Restaurant
15	St. Francis Wood/Miraloma/West Portal	Chinese Restaurant	Yoga Studio	Asian Restaurant	Café	Japanese Restaurant	Pharmacy	Shipping Store	Coffee Shop	Pool Hall	Convenience Store

This in our fourth cluster with green color, we only found a single neighborhood in this cluster. The first place is occupied by a park that we did not see in the previous clusters, that is why it is not found in the previous clusters that we saw. We also find other places like gardens, baseball fields, yoga studio among other places, these places are what make it different from the other clusters we saw.

	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
19	Visitacion Valley/Sunnydale	Park	Garden	Baseball Field	Yoga Studio	Electronics Store	Food Truck	Food & Drink Shop	Food	Flower Shop	Fish Market

In our last cluster with yellow color, we noticed that we see a gym ranks first, followed by a coffee shop and then followed by a pizza shop. Unlike the previous clusters, here we find other types of places such as nightclub and dog run. Those places we mention are what make this cluster different from the others.

	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
17	Lake Merced	Gym	Coffee Shop	Pizza Place	Mexican Restaurant	Sandwich Place	Café	Sushi Restaurant	Light Rail Station	Nightclub	Dog Run

4. Discussion

Throughout this process we analyze the city of San Francisco looking at the neighborhoods it owns. Through data analysis we saw the similarities and differences in terms of the venues presented by each neighborhood.

To get the neighborhoods of San Francisco I scraped the data from a website, then based on its postal code I obtained the geographic coordinates of each neighborhood. After obtaining the coordinates with the help of Foursquare I obtained the venues of each neighborhood. The process of obtaining venues can vary depending on the parameters sent to Foursquare. Finally use K-Means to cluster the neighborhoods of San Francisco.

One last note these results that we obtained may vary over time since the Foursquare database is constantly updated.

5. Conclusion

As we see San Francisco is a great city to make investments because it is a very dense city. If you decide to invest in a common venue for you to do well, you should have a capital comparable to other businesses in the same category.

But if you want to invest in a new business you could see the similarities and differences between the neighborhood where you want to invest and other neighborhoods and thus decide what to invest in. Keep in mind that this study cannot be validated as time passes by because the information we collect from Foursquare is changing because its database is updated with new venues.

6. Reference

- San Francisco (California)
- Foursquare API
- San Francisco Neighborhoods