**Capstone Project - The Battle of the Neighborhoods**

1. **Introduction**
   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. **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.

1. **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.

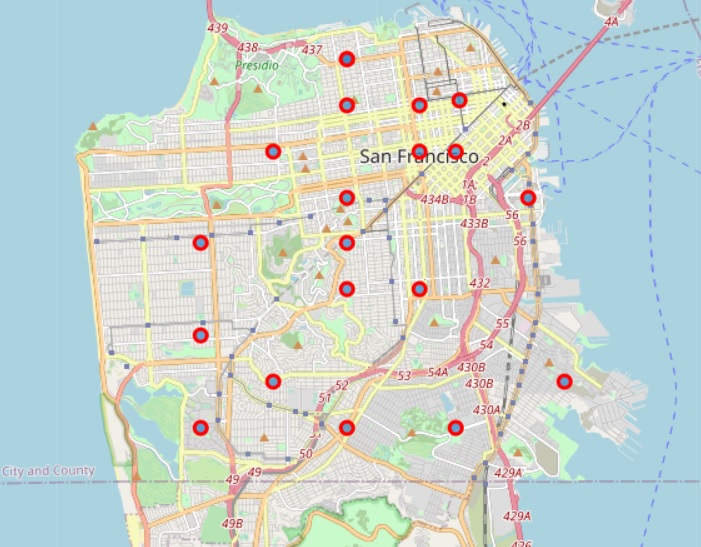


To obtain the latitude and longitude I use a library that I found on the web called uszipcode, this library works as a database that through the zip code you can obtain the latitude and longitude of each neighborhood of San Francisco or another city in the United States.

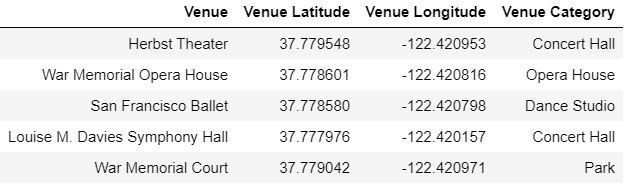
After obtaining the latitude and longitude through the library we have the following table.



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.

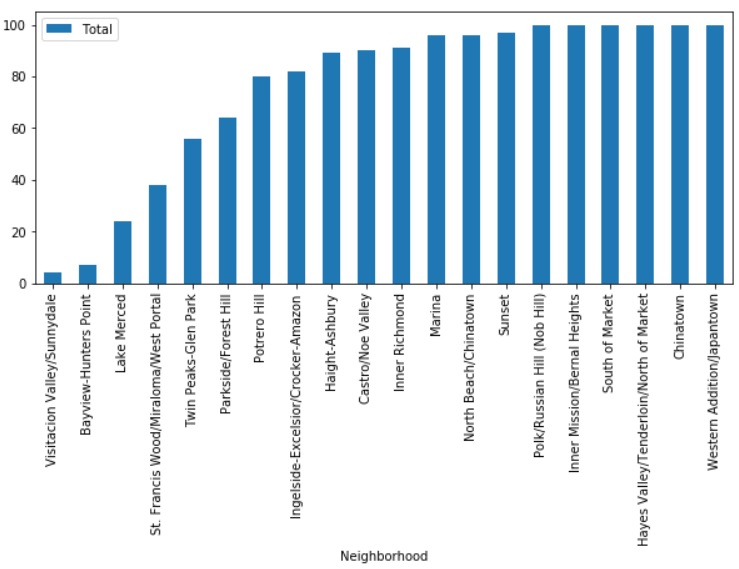


I used Foursquare API to get the venues of each neighborhood. I configure the search for a maximum of 100 venues for each neighborhood and a radius of 700 meters. These are the parameters that we must send to Foursquare apart from the latitude and longitude that we already have. In this table we see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



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.

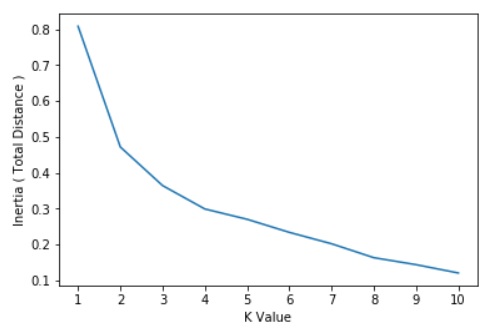


Let's see what categories the places in each neighborhood have. In the result that Foursquare returns we have a total of 244 categories, this is normal since San Francisco has a very dense population. Now I show you a part of the table that contains the 10 most common venues in each neighborhood.



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.



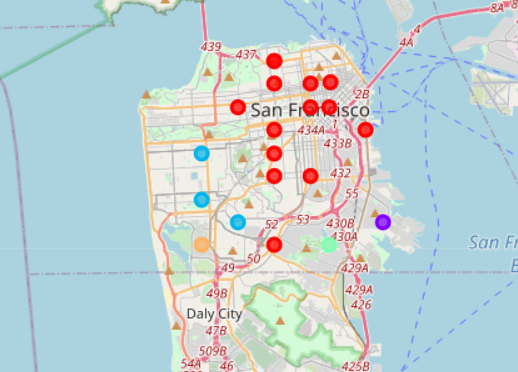
After using K-Means with k equal to 5, to a clean table that we make in this process we add the labels that the algorithm returned to us.



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.

1. **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.



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.

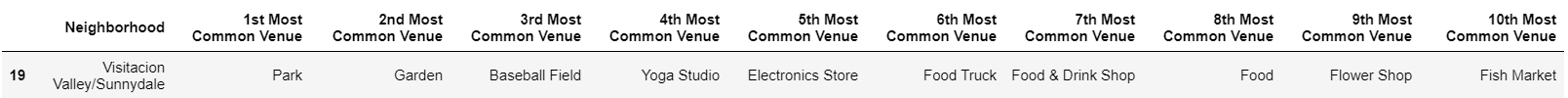


This is our third cluster with blue color made up of only three neighborhoods. In which we see that in the three neighborhoods the first place is occupied by Chinese restaurants. In addition, other types of restaurants occupy the rest of the list. We can conclude what characterizes this cluster from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



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.



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.



1. **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.

1. **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.

1. **Reference**

* [San Francisco (California)](https://es.wikipedia.org/wiki/San_Francisco_(California))
* [Foursquare API](https://developer.foursquare.com/)
* [San Francisco Neighborhoods](http://www.healthysf.org/bdi/outcomes/zipmap.htm)