# Report

# Coffee shop clustering in San Francisco

Dimitar Kumenov

22.11.2019

## Introduction

## Definition of the Business problem

A business owner wants to open a new coffee shop in San Francisco. He wants to be recommended which is the most suitable neighborhood.

We will try to solve the problem using the foursquare location data in order to choose the most suitable neighborhood in the city for the purpose.

The purpose is to recommend a neighborhood in San Francisco which is most suitable for opening a new coffee shop.The right location of the shop is key to it's success. The solution is suitable to all kind of business owners which operates coffee shops or other different type of restaurants , bars, pubs and look for a location for the new one.

The proposed solution is to gather the needed data about neighborhoods in the city and the data about all coffee shops in them. After this we will cluster the neighborhoods and we will chose the best one based on a criteria.

## Data

The data that will be used is from two sources: a web page with information about the neighborhoods in San Francisco and Foursquare location data.

web page : http://www.healthysf.org/bdi/outcomes/zipmap.htm consisting the list of neighborhoods in San Francisco.

Forusquare.com : location data from Foursquare about the venues and coffee shops in specific.

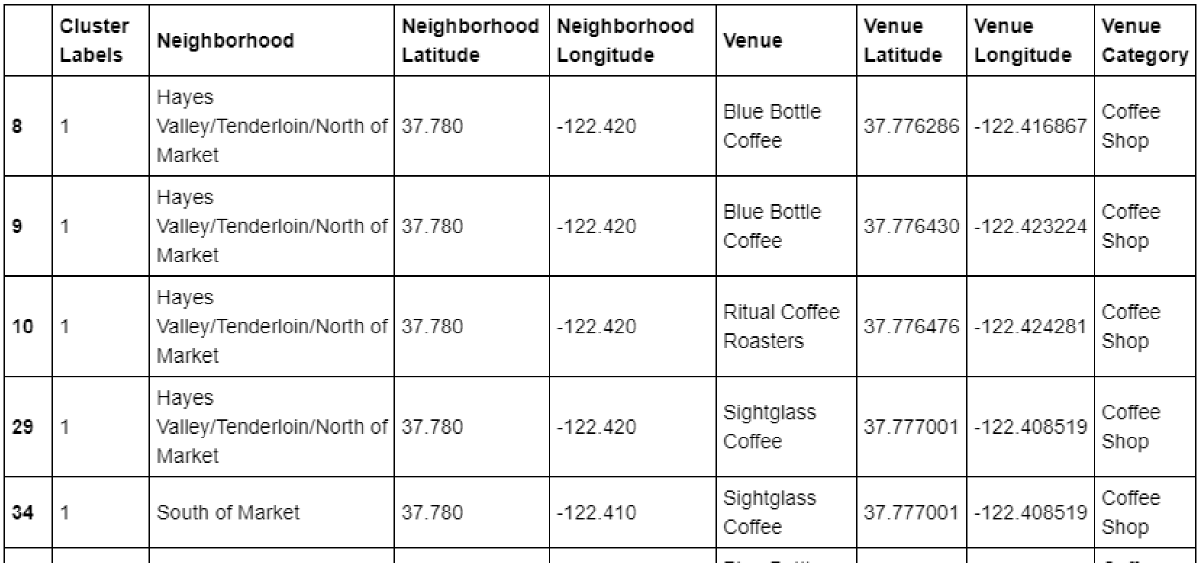
For the purpose of the project we will scrap the data about the neighborhoods in San Francisco from the web page. After this we will add the geographic coordinates for each neighborhood. After this for each neighborhood will be retrieved the location data from Foursquare in order to cluster the neighborhoods.

## Metodology

First we web scraping the data about the neighborhoods in San Francisco and pass them to a dateframe.in pandas Add to each neighborhood the corresponding geographic coordinates. For each neighborhood the Foursquare location data for the nearest venues is gathered and pass to the existing dataset.

Filtering only the records with Venue Category "Coffee shop" in order to create new dataset only with the coffee shops returned by Foursquare and the data about them. We cluster the new dataset in 5 clusters and add the cluster label for each row in our table with coffee shops.

The algorithm used for the project is k-means which cluster the data based on the similarity of the values in it. Using the algorithm counts each neighborhood and add it to a specific cluster which will help us to determine which cluster will be the most suitable to choose a neighborhood from.



| | Cluster Labels | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|---|
| **8** | 1 | Hayes Valley/Tenderloin/North of Market | 37.780 | -122.420 | Blue Bottle Coffee | 37.776286 | -122.416867 | Coffee Shop |
| **9** | 1 | Hayes Valley/Tenderloin/North of Market | 37.780 | -122.420 | Blue Bottle Coffee | 37.776430 | -122.423224 | Coffee Shop |
| **10** | 1 | Hayes Valley/Tenderloin/North of Market | 37.780 | -122.420 | Ritual Coffee Roasters | 37.776476 | -122.424281 | Coffee Shop |
| **29** | 1 | Hayes Valley/Tenderloin/North of Market | 37.780 | -122.420 | Sightglass Coffee | 37.777001 | -122.408519 | Coffee Shop |
| **34** | 1 | South of Market | 37.780 | -122.410 | Sightglass Coffee | 37.777001 | -122.408519 | Coffee Shop |

The idea is that we would recommend to open a new shop in a neighborhood that has a second largest number of coffee shops after the clustering is performed and the labels are added to a new dataframe.

# Results

The number of all coffee shops locations returned by Foursquare is 27. In the next step we will continue to work only with those neighborhoods which have a coffee shops.

We set the number of clusters to be 5.The clustering algorithm labels each neighborhood to a specific cluster and after this this information is added to the dataset.

In the resulting section after the coffee shops in each cluster are counted we chose the cluster number 3 because it has 5 coffee shops in It.

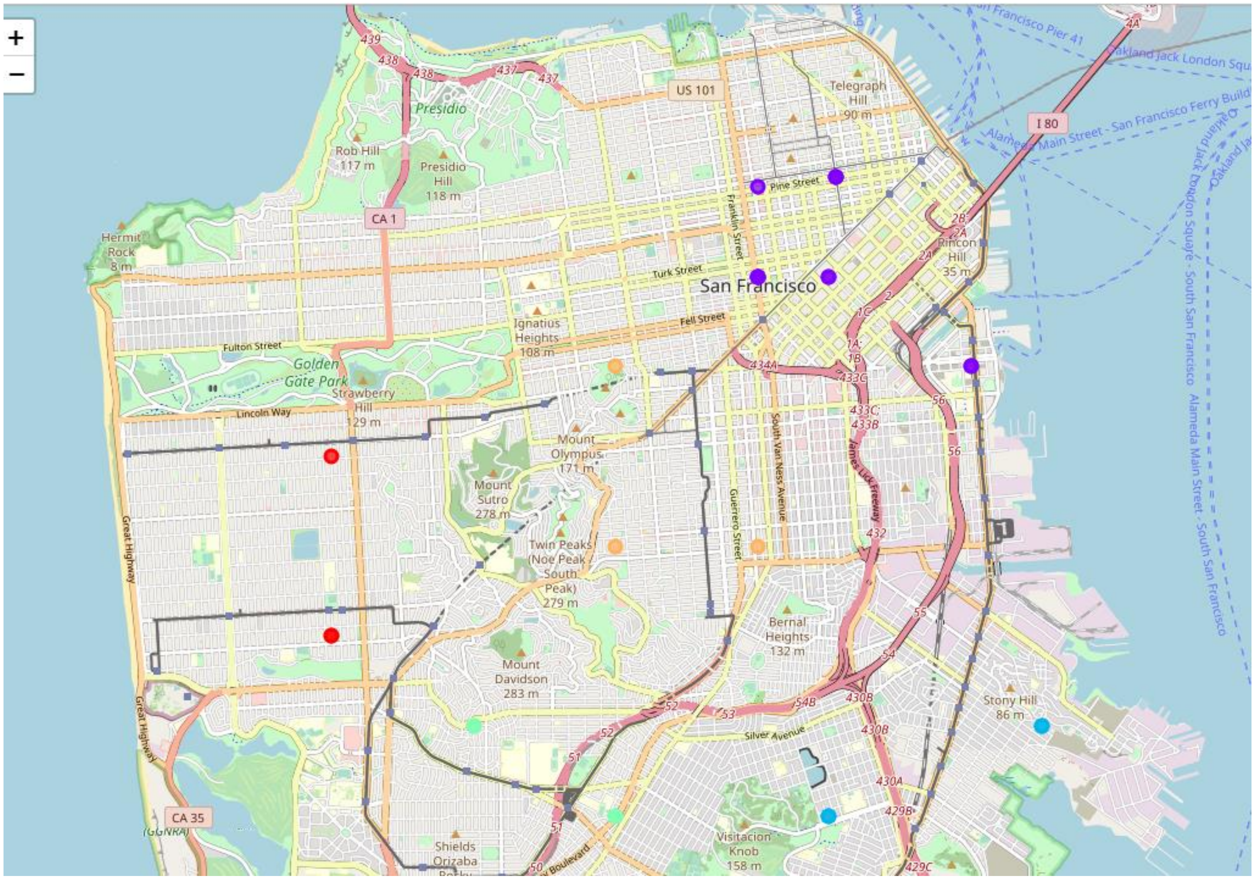
# Disclusion

Based on the returned numbers of coffee shops we assume that the cluster with high number of coffee shops would be to competitive for opening e new location. We suggest that the cluster with second largest number of coffee shops would be more suitable to choose a neighborhood from. In it there are currently coffee shops but not as many as in the other cluster and the competition wll be less.

# Conclusion

It must be noted that the performance of the explained above algorithm and methodology highly depends of the returned data from Foursquare. The project is showing the methodology of the data science approach for solving a problem using Foresquare location data.

We suggest to the client that based on the algorithm and it’s Performance he should open the new store in St. Francis Neighborhood which is labeled in cluster 3.


+
−
Presidio
Rob Hill 117 m
Presidio Hill 118 m
Hermit Rock 8 m
CA 1
US 101
I 80
Telegraph Hill 90 m
Pine Street
Franklin Street
Turk Street
San Francisco
Fell Street
Rincon Hill 35 m
Ignatius Heights 108 m
Fulton Street
Golden Gate Park
Strawberry Hill 129 m
Lincoln Way
Mount Olympus 171 m
Mount Sutro 278 m
Twin Peaks (Noe Peak South Peak) 279 m
Great Highway
South Van Ness Avenue
Guerrero Street
James Lick Freeway
Bernal Heights 132 m
Mount Davidson 283 m
Stony Hill 86 m
Silver Avenue
Visitacion Knob 158 m
Shields Orizaba
CA 35
(GGNRA)
San Francisco Pier 41
Oakland Jack London Square
Alameda Main Street - San Francisco Ferry Building
Alameda Main Street - South San Francisco