The Battle of Neighborhoods:

Boston vs San Francisco

AUTHOR: SANTIAGO CARULLO

1. Introduction/ Business Problem

Boston and San Francisco are two large cities in United States. Both cities being cities by the bay and both metro area sharing a similar population (Boston population around 4.9 million and San Francisco 4.7 million). Given that both cities are in opposite ends of the country, it would be interesting to compare both of their neighborhood, with the power of machine learning, to observe how the resemble each other or differ from each other.

This analysis is intended to show which areas of one city resemble those of the other: Boston and San Francisco. This could be helpful for different use case (and therefore different stakeholders:

- *People moving between cities* would like to live in a very specific type of neighborhood. This study could be useful to filter similar neighborhoods in both places.
- *Companies expanding from one city to the other* might also try to find a neighborhood to settle in first. They can use their experience from the original city and look for a fitting area in the second one.

2. DATA

For this project the Foursquare API will be used. A list of neighborhoods in Boston and San Francisco is downloaded and their respective location in longitude and latitude coordinates is obtained. The sources are the following:

- Boston neighborhoods: https://www.kaggle.com/yingzhou474/boston-neighborhoods-geojson
- San Francisco neighborhoods: https://www.kaggle.com/broach/san-francisco-neighborhood-maps

The data downloaded are the neighborhoods located in Boston and San Francisco. Moreover, their specific coordinates are merged. Then, the Foursquare API GET request is sent in order to acquire the surrounds venues that are within a radius of 500m. The data is formatted using one hot encoding with the categories of each venue. Then, the venues are grouped by neighborhoods computing the mean of each feature.

The similarities will be determined based on the frequency of the categories found in the neighborhoods.

3. Methodology

3.1 Feature Extraction

In order to obtain the features, One Hot Encoding is used in terms of categories. Each category represents a venue. After, each category is transformed into a binary classification, with the values being 1 or 0. Moreover, the data is grouped by the neighborhood they are obtained from, using the mean of each neighborhood feature to do calculations. This provides us a venue for each row and each column will contain the frequency of a certain category.

3.2 Exploratory Data Analysis

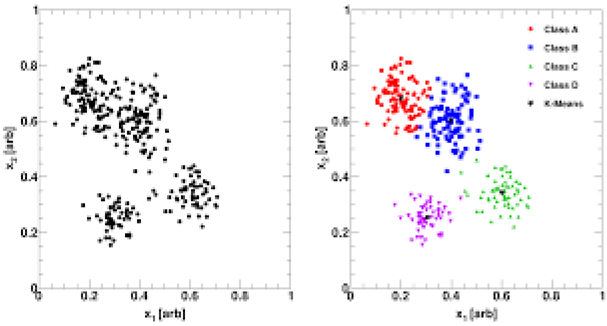
Analyzing graphs containing the most common venues in each city, and comparing between them

3.3Unsupervised Machine Learning

A clustering algorithm is implemented in order to obtain similar neighborhood in both cities. In this case, K-Means is used due to its simplicity and its similarity approach to found patterns.

- **K-Means:**

K-Means is a clustering algorithm. This algorithm search clusters within the data and the main objective function is to minimize the data dispersion for each cluster. Thus, each group found represents a set of data with a pattern inside the multi-dimensional features.

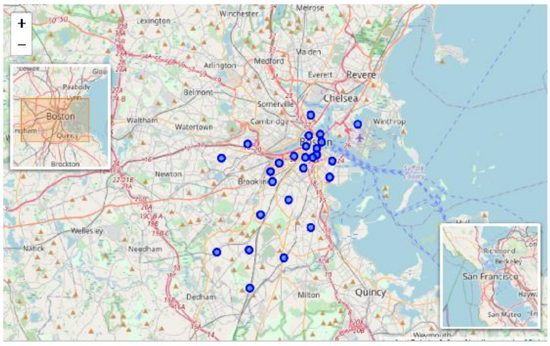


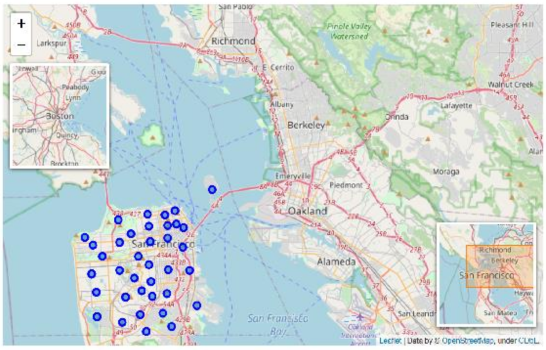
Source: https://freddyox.github.io/blog/Kmeans/

The number of clusters must be determined before running the algorithm, as it is one of the inputs. Therefore, in order to determine the optimal number of clusters, the elbow method will be utilized. This method consists of a chart that compares the mean square error vs number of clusters is done and the elbow is selected.

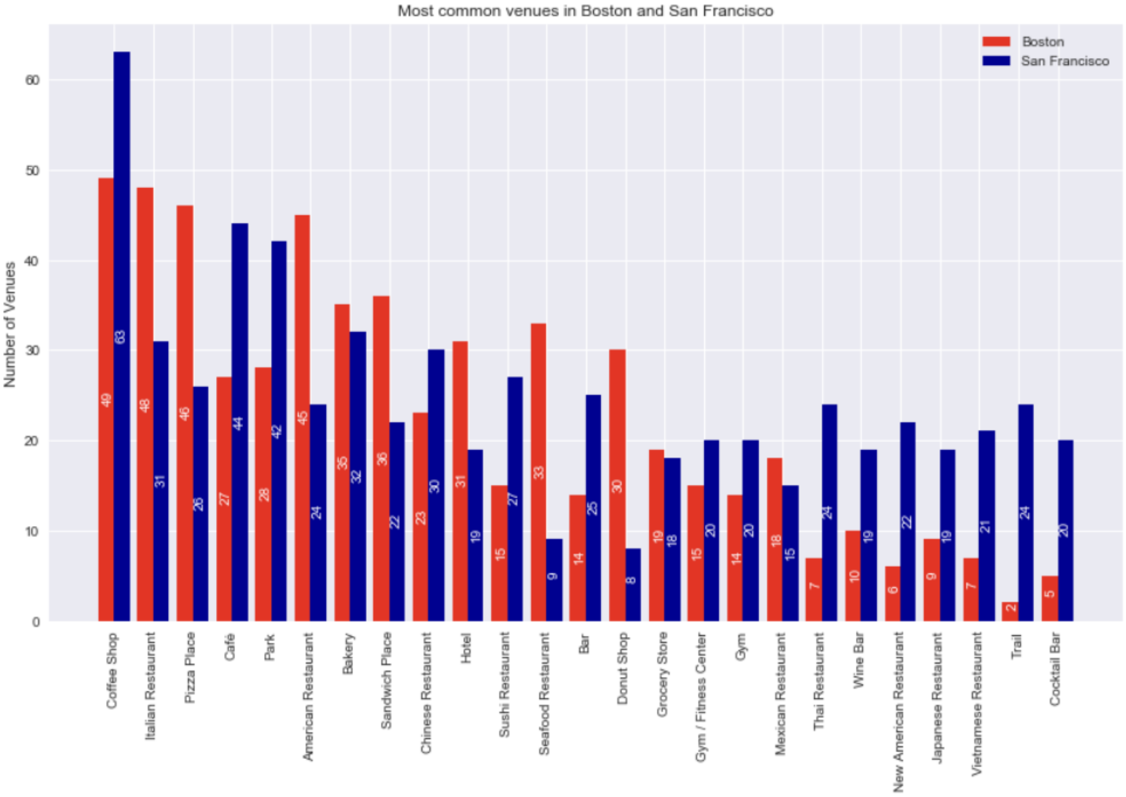
4. Analyzing Results

To begin with, data is plotted in a geographical map represent the data location. The following images present the neighborhoods in Boston and San Francisco that will be used to do the analysis.





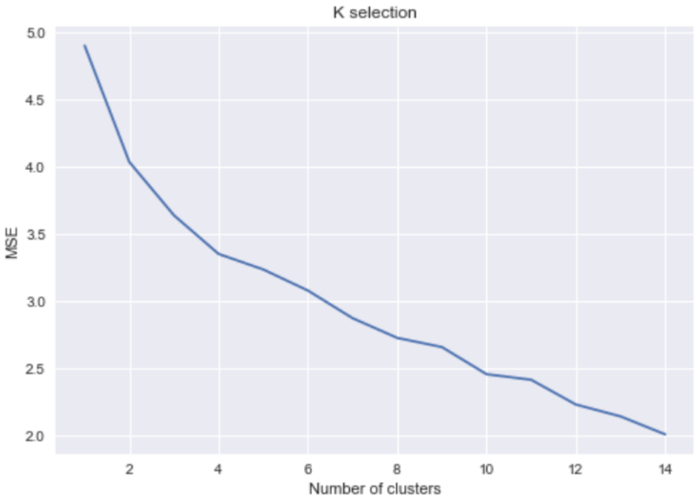
In the first place, both cities are compared at a city level. This is presented in the graph below, where the venues most common in both cities are compared with the amount within each city.



As it can be observed, coffee shops are the most common venue in both cities. Although in San Fracisco there are more. Some notable differences are:

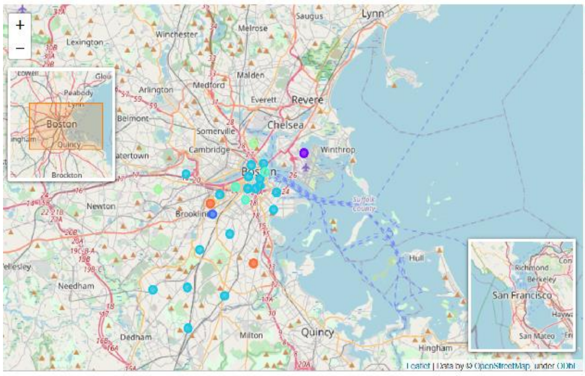
- Comparing the types of restaurants, San Francisco apears to have a larger amount of international cuisine whereas Boston has more American, Italian Restaurant and Pizza Place.
- The amount of parks in San Francisco is greater than Boston ones.

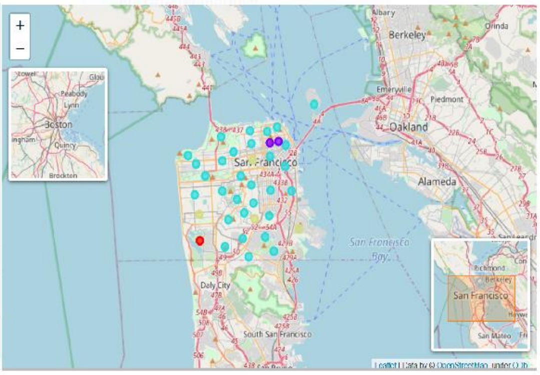
Moving forward, the cluster algorithm is implemented. The mean squared error (MSE) is plotted vs the number of clusters in order to obtain the optimal number of clusters. The number of clusters start with a value of 1 increasing until a value of 15. This graph is shown in the image below.



As expected, the MSE will decrease as the number of clusters increases. In this case, it is possible to see that the elbow is found around N=4. The MSE found below this number shows little changes rather than big ones. Finally, once the number of clusters is selected, the clustering algorithm is repeated through samples and each neighborhood is labeled according to the clusters found.

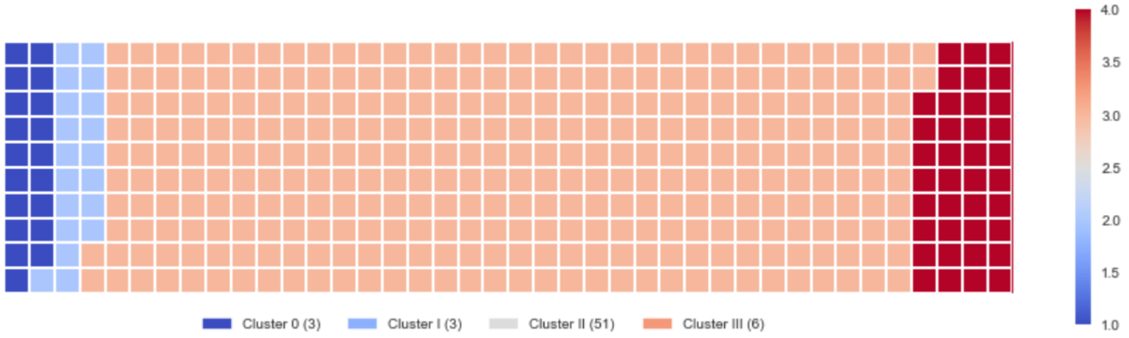
For visualization purposes, the geographical data is again plotted but with different colors. Each color represents the cluster for which that neighborhood belongs. This image is shown below.



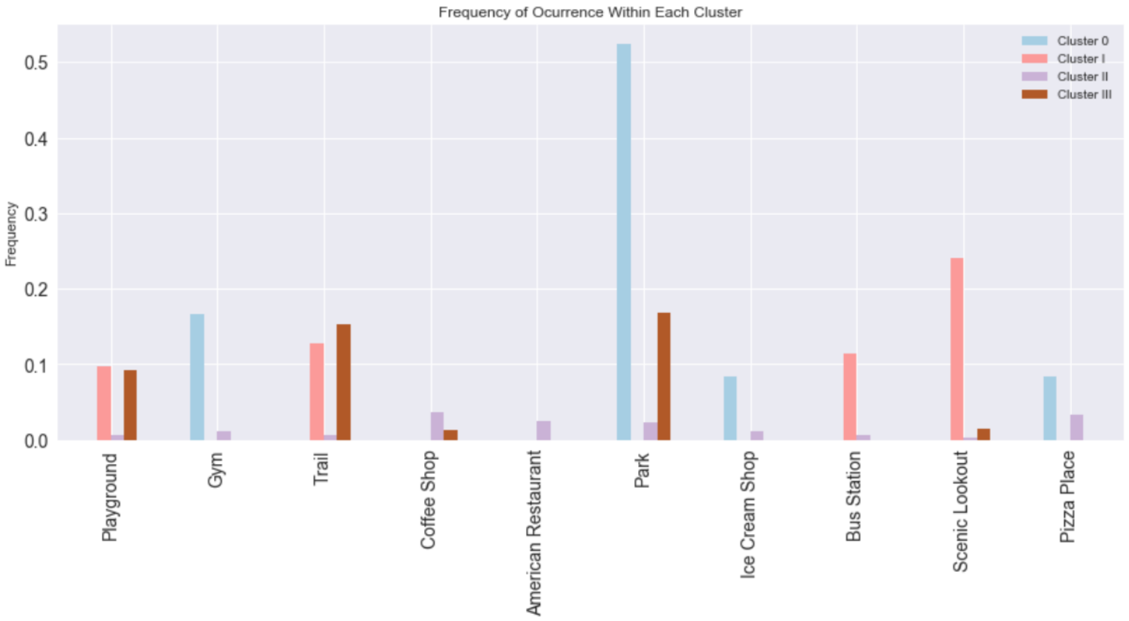


The images prove that cluster algorithm is not segmenting the neighborhoods based on location areas. Meaning that it is not true that geolocation of neighborhoods is correlated with the categories of the venues around each neighborhood. Yet, it is possible to see which neighborhoods Boston are like the neighborhoods in San Francisco. Neighborhoods considered to be similar are in the same cluster. Hence, they have the same color in the image above.

Moreover, the proportion of the neighborhoods assigned to each cluster can be depicted with a waffle chart. It can be observed below there is one major cluster containing most of the neighborhoods and three smaller ones.



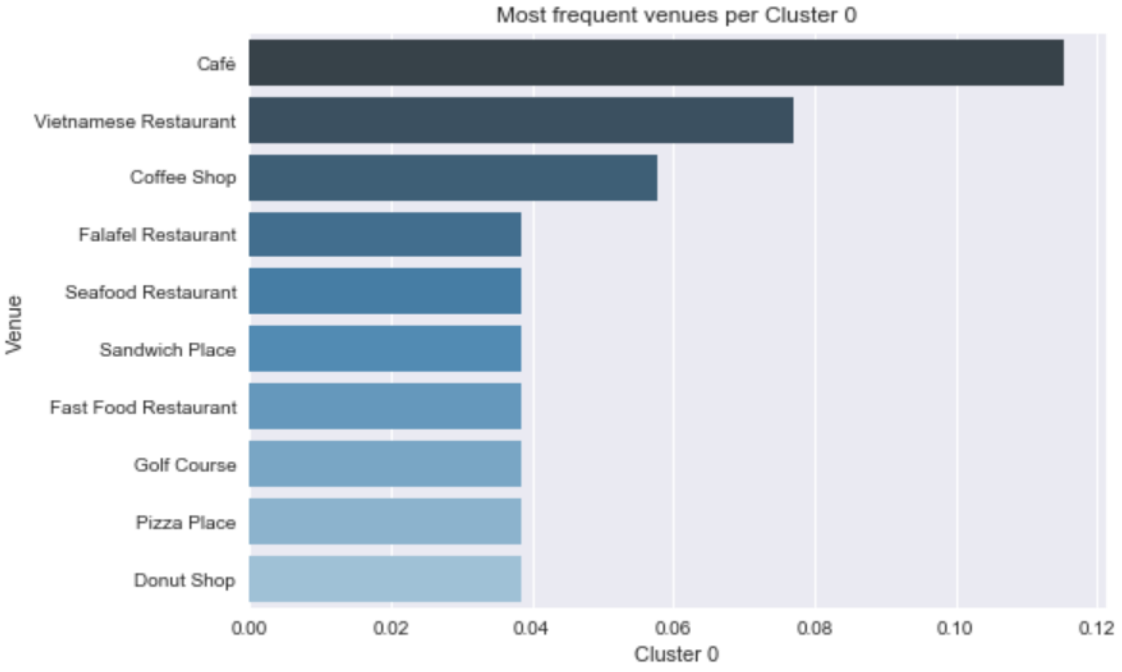
For further analysis, the following bar chart is employed in order to explore insights within the clusters. The graph shown below, portrays the features with higher frequency in the centroids.



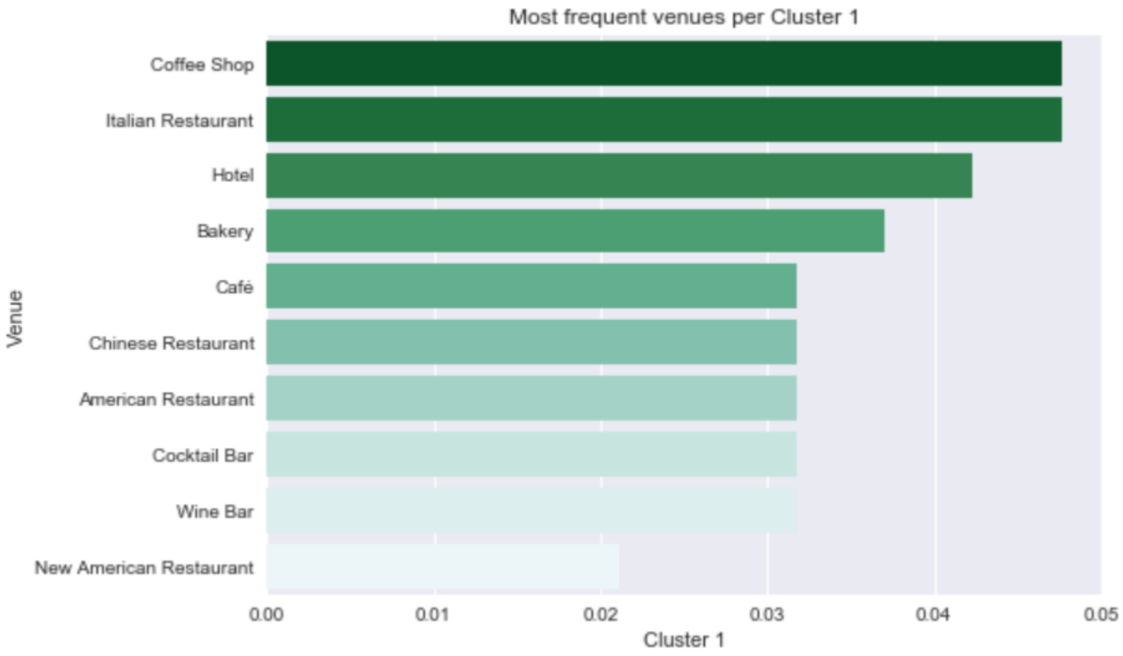
It is possible to see that Cluster 0 focuses on neighborhoods that have around parks, gym and pizza places. In addition, Cluster I portray neighborhoods that have around trails, playgrounds and scenic lookout. Furthermore Cluster 2, which is the biggest cluster, does not present any strong frequency for its venues. The last cluster, Cluster 3 focuses on similar aspects as Cluster 1 neighborhoods without de scenic lookout but with more parks.

For in depth study on the clusters, here are the top 10 venues per cluster:

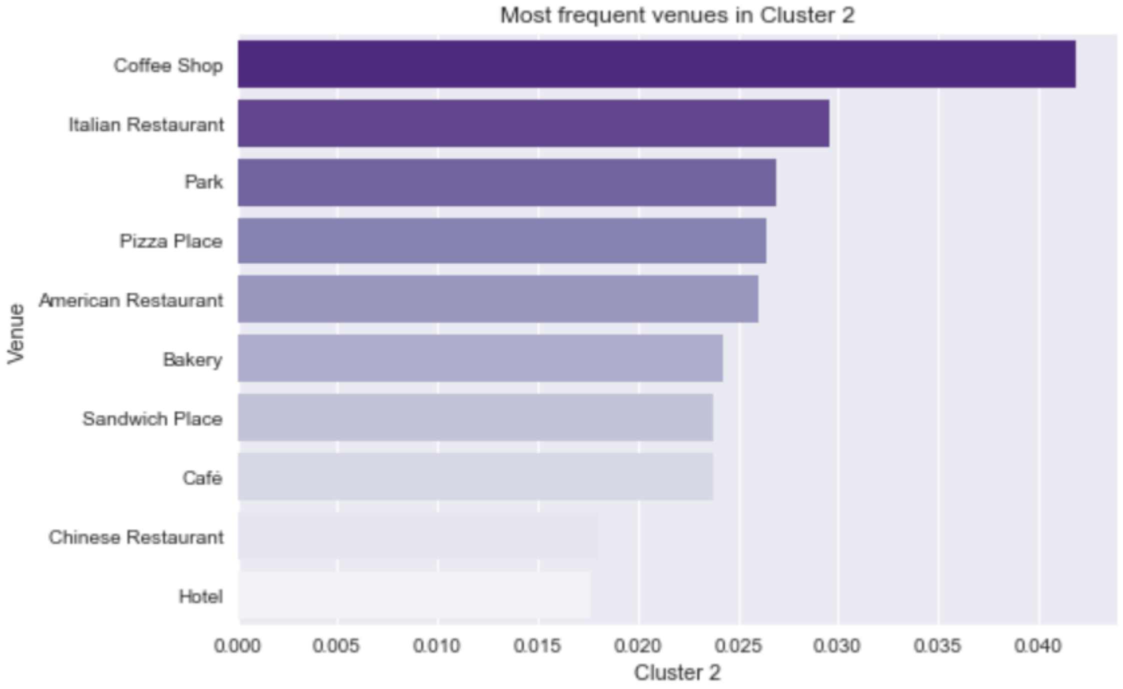
- Cluster 0



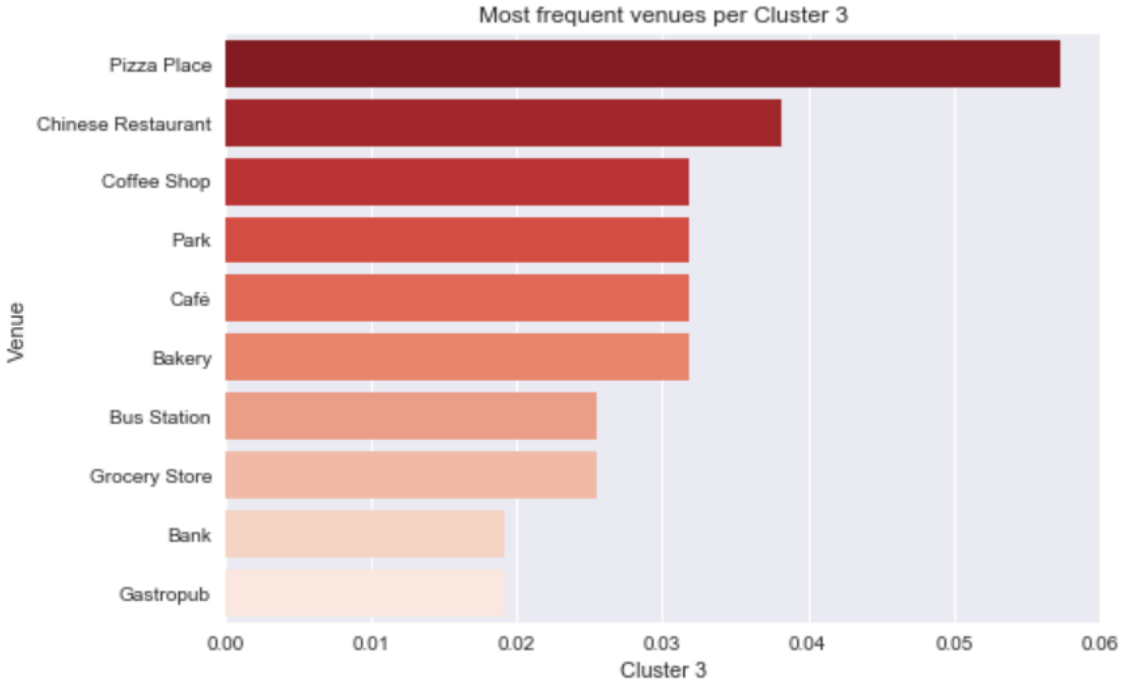
- Cluster 1



- Cluster 2



- Cluster 3



5. Conclusion

In this project a classification of two different cities is done. The cities involved are Boston, MA and San Francisco, CA. This is done by grouping venues by neighborhoods, which are the features for the clustering.

The algorithm could divide neighborhoods of both cities in for different clusters. Where in each cluster there are at least 3 neighborhoods in it. With different aspects for each cluster.

This analysis is not perfect due to the limitations of using Foursquare API without a pro account, limiting the calls one can do.

However, this project, is useful for differentiating neighborhoods within a city and finding similar ones in another.