

Battle of Neighborhoods: Boston vs San Francisco

IBM DATA SCIENCE CAPSTONE PROJECT

AUTHOR: Santiago Carullo

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).

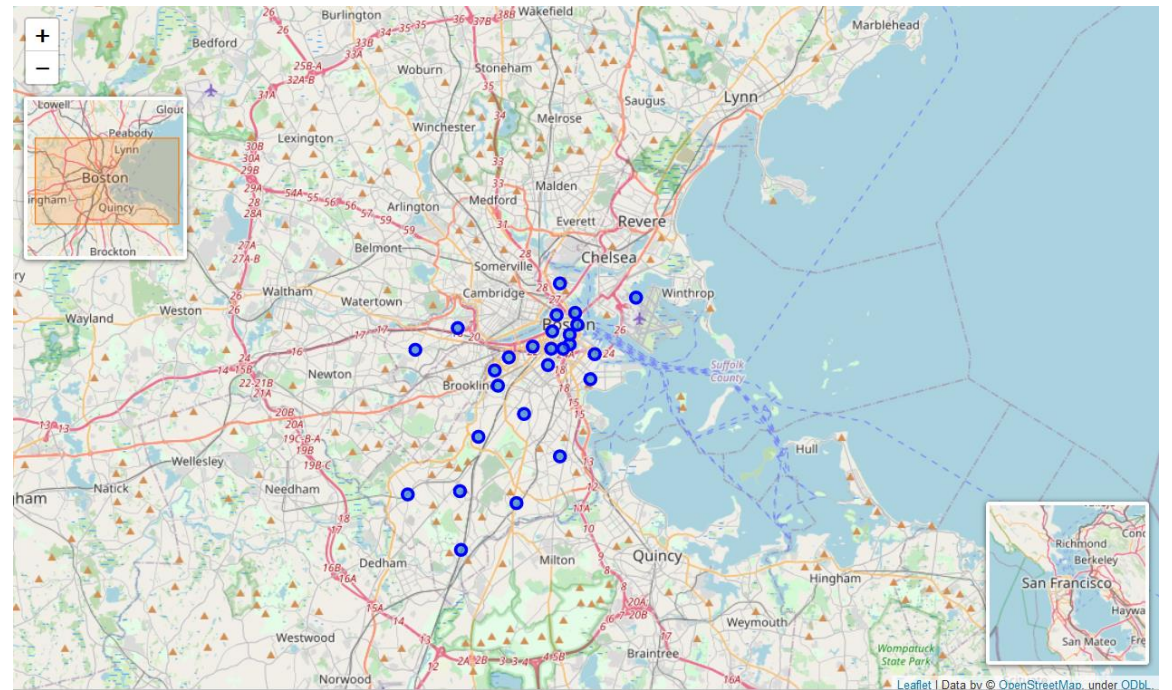
This analysis is intended to show which areas of one city resemble those of the other: **Boston and San Francisco**. This could be helpful for different use case (and therefore different stakeholders:

- *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.

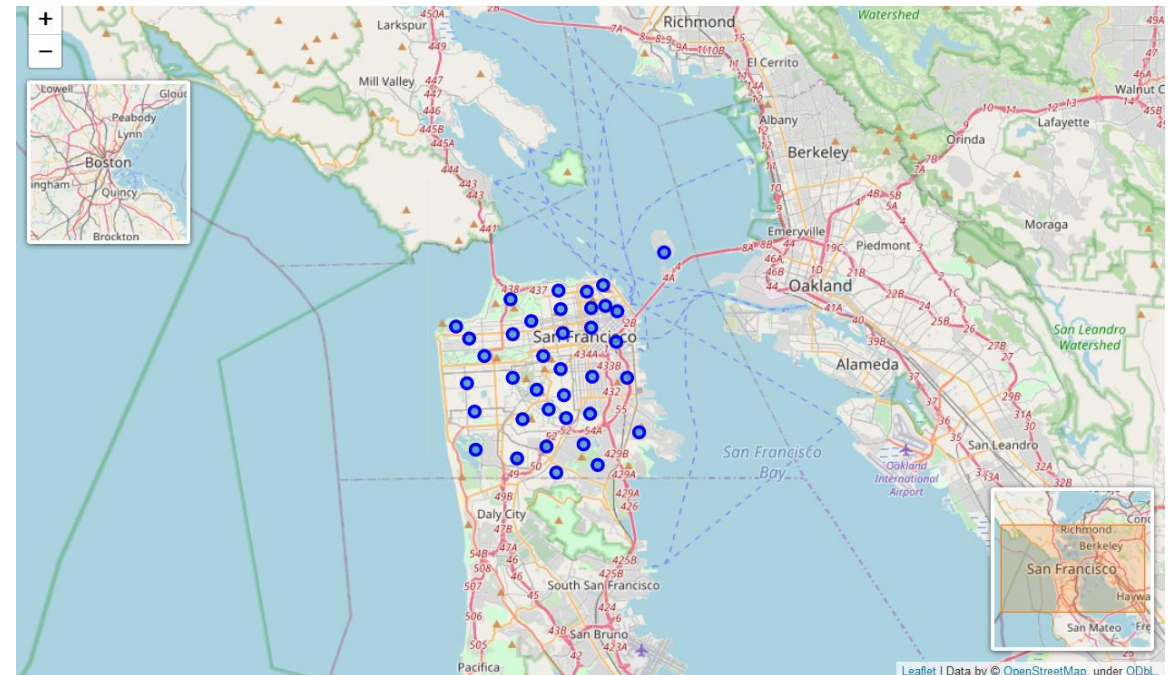
Data

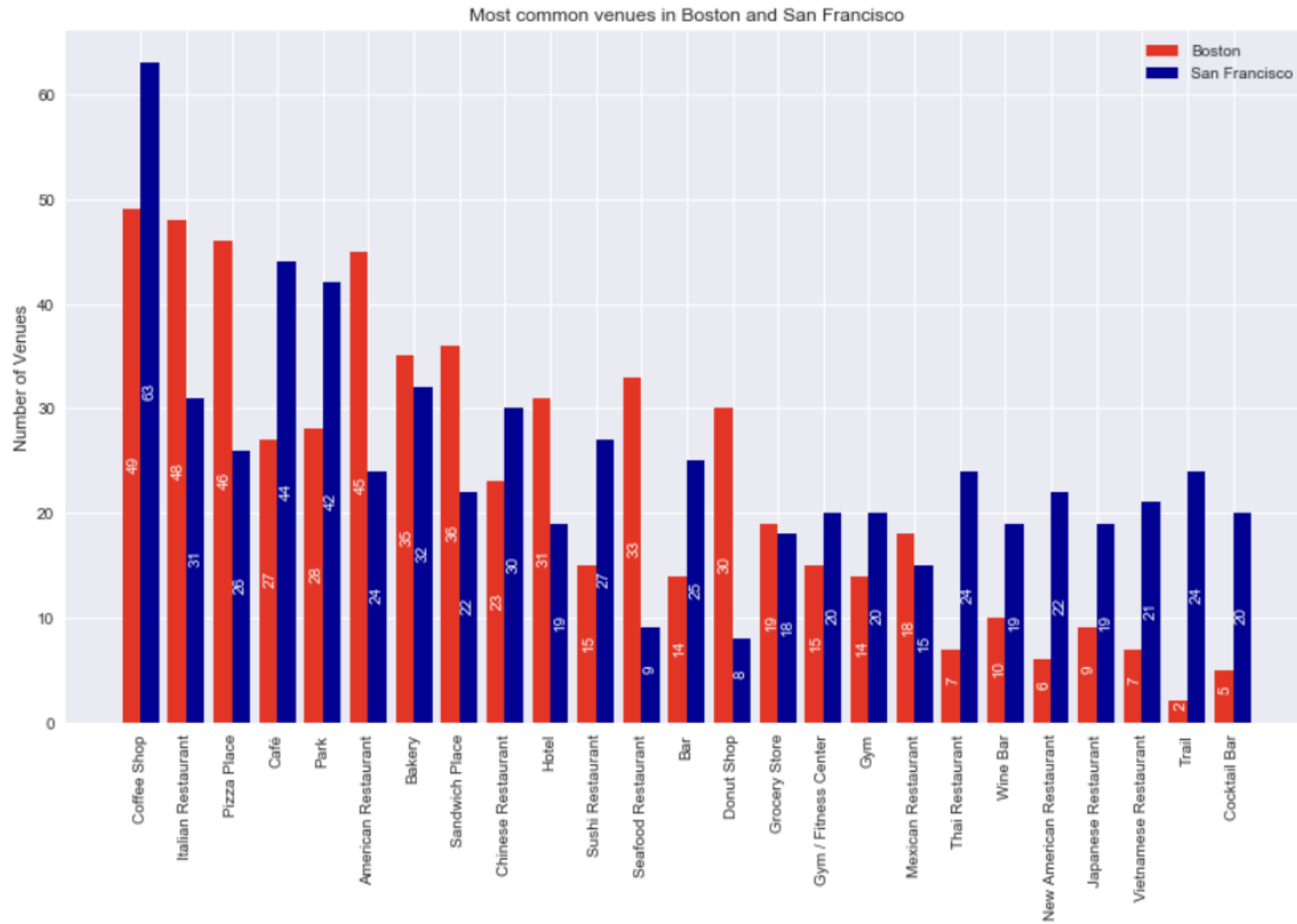
- Boston neighborhoods: <https://www.kaggle.com/yingzhou474/bostonneighborhoods-geojson>
- San Francisco neighborhoods: <https://www.kaggle.com/broach/sanfrancisco-neighborhood-maps>
- Foursquare API

Boston



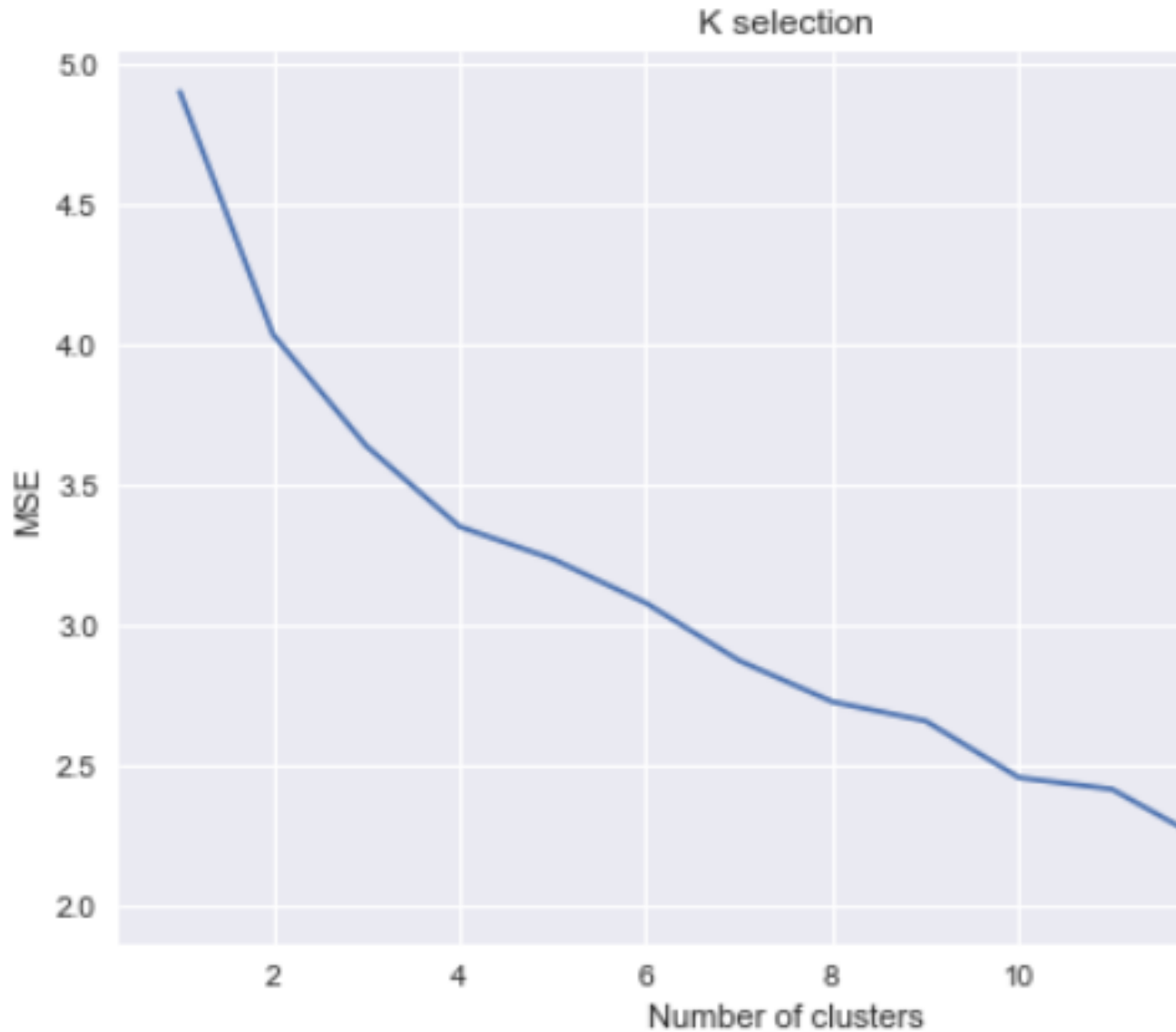
San Francisco





Comparing most common venues

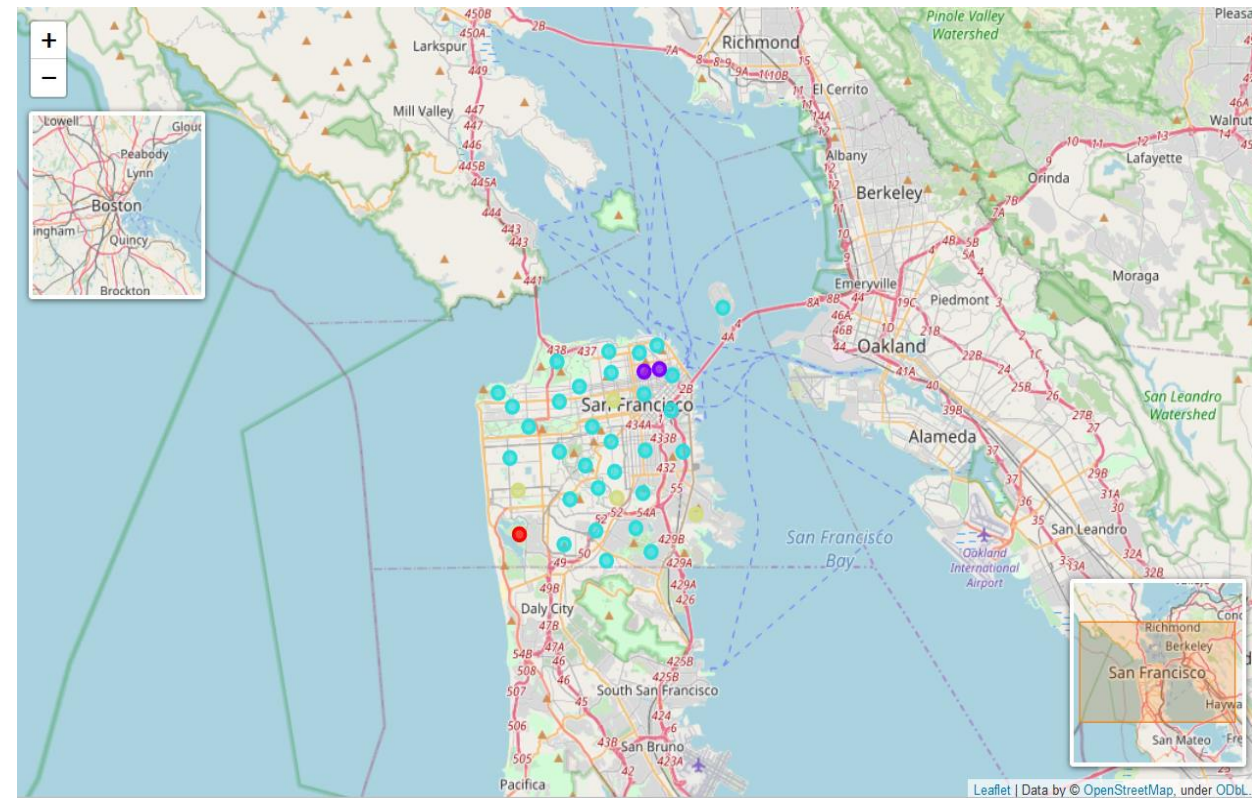
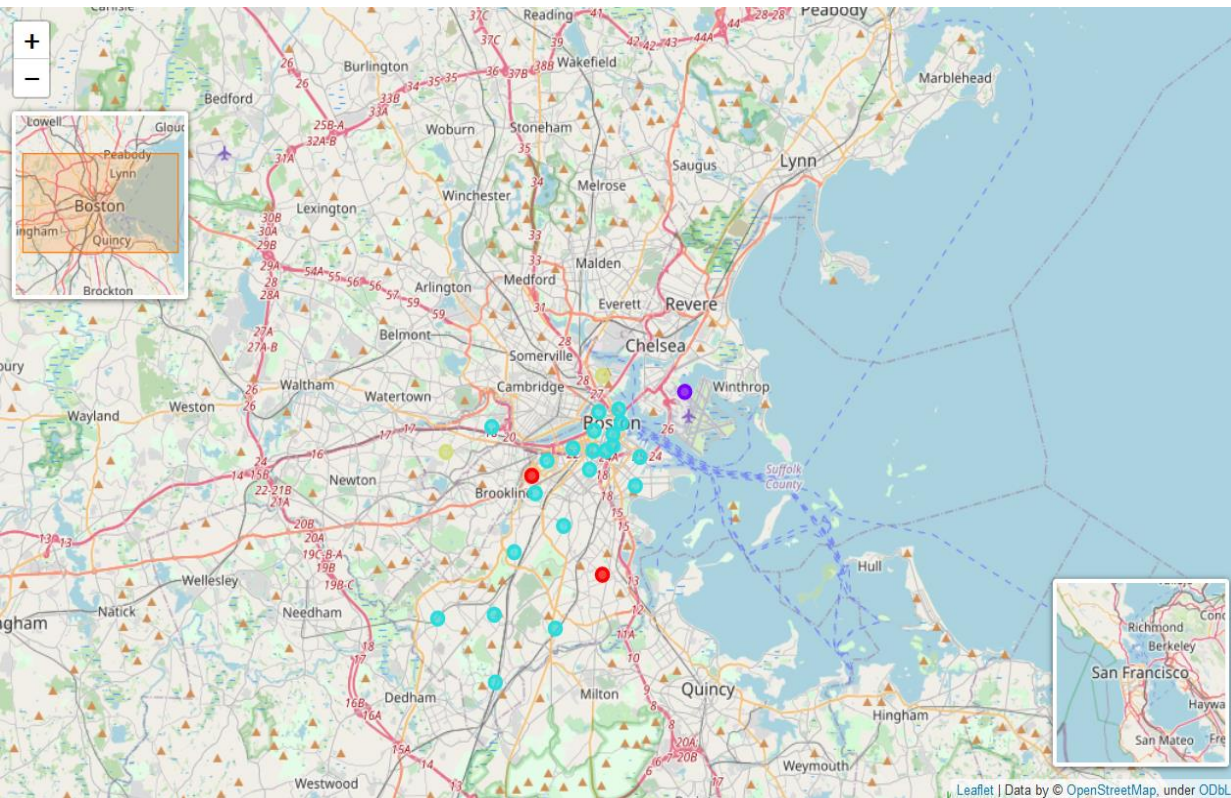
- Coffee shops are the most common venue in both cities. Although in San Francisco there are more. Some notable differences are:
- Comparing the types of restaurants, San Francisco appears to have a larger amount of international cuisine whereas Boston has more American, Italian Restaurant and Pizza Place.
- The amount of parks in San Francisco is greater than Boston ones.



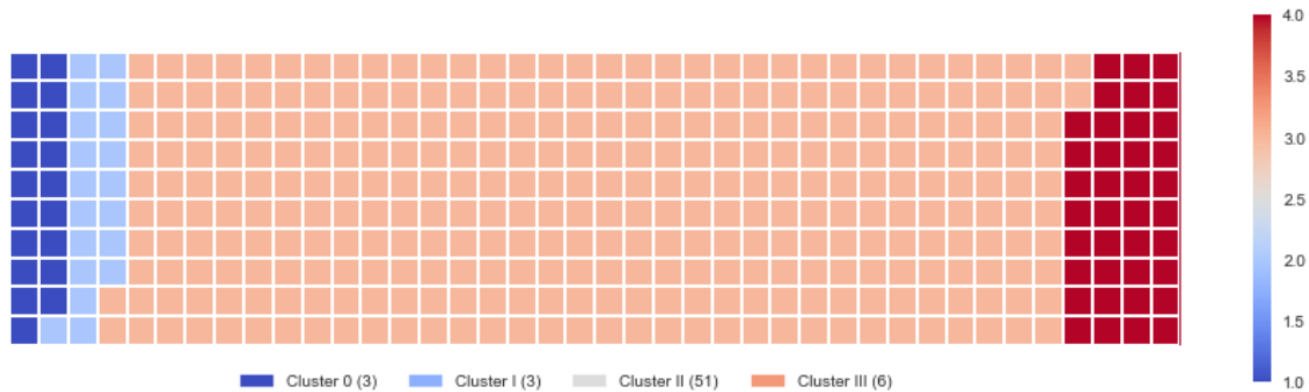
Number of Clusters

- Used the elbow method
- Compares mean square error with number of clusters
- K is determined when little information is gained by adding another cluster
- K=4 in this case

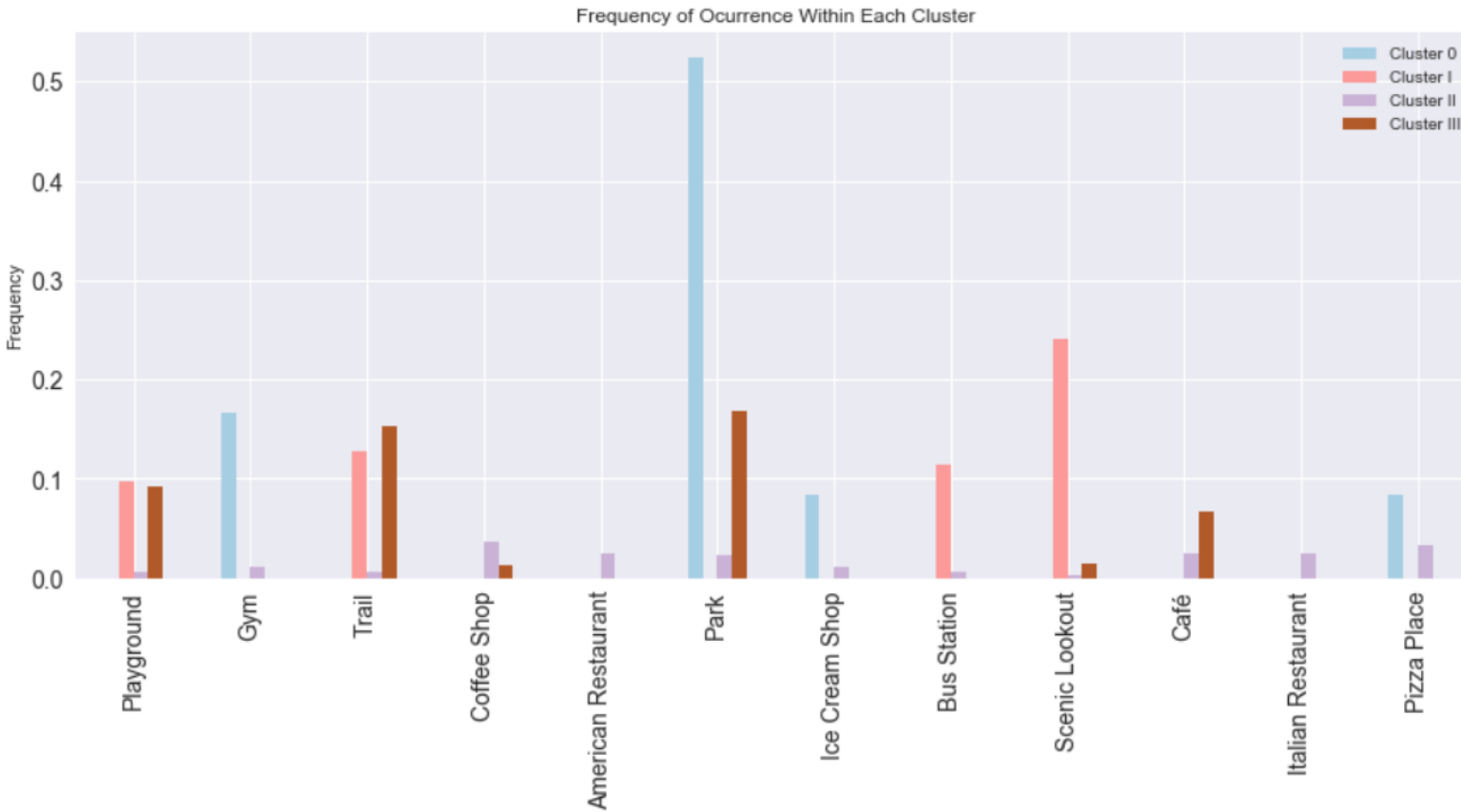
Plotting clusters in a map



Proportion of neighborhoods per cluster

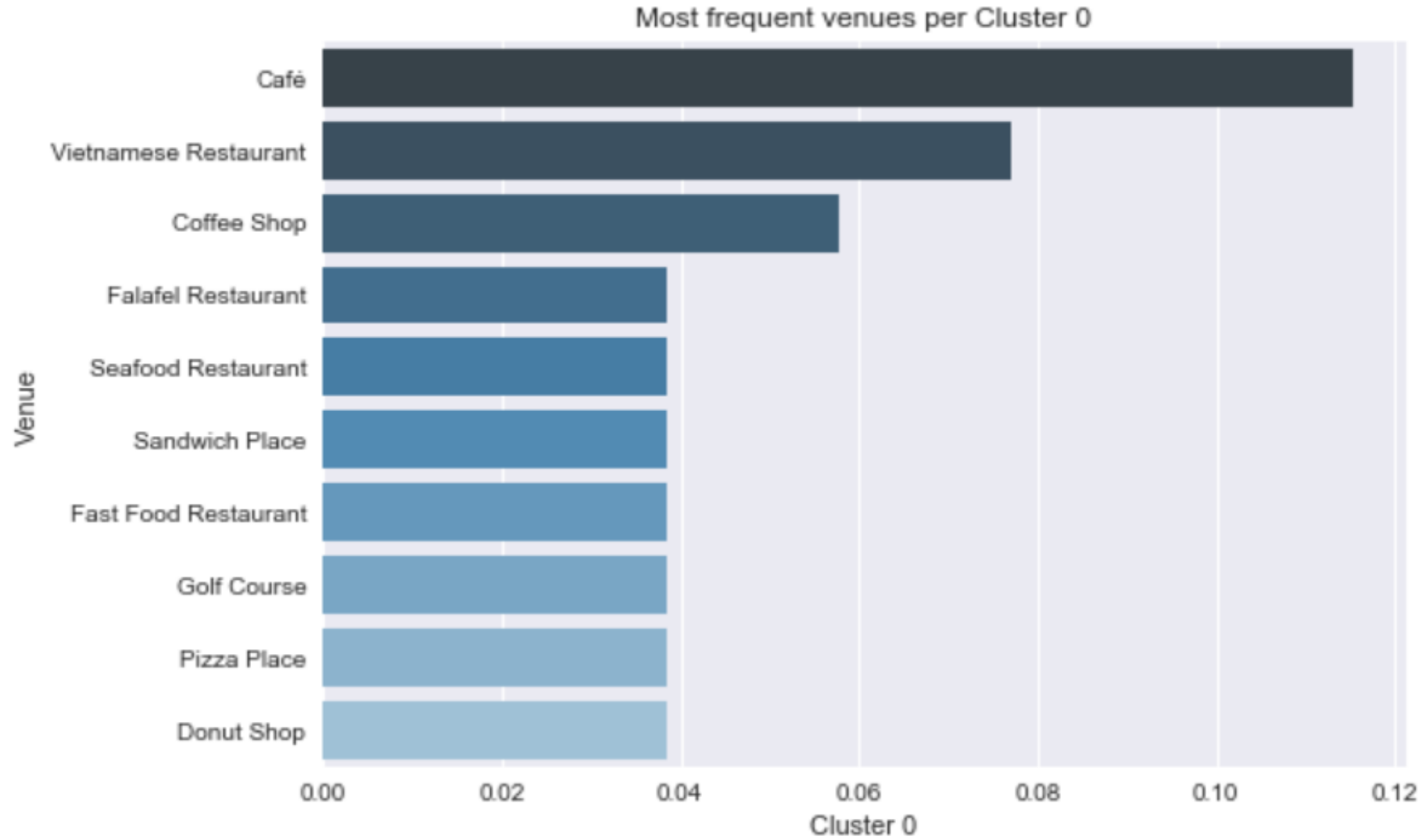


Frequency of occurrence within each cluster

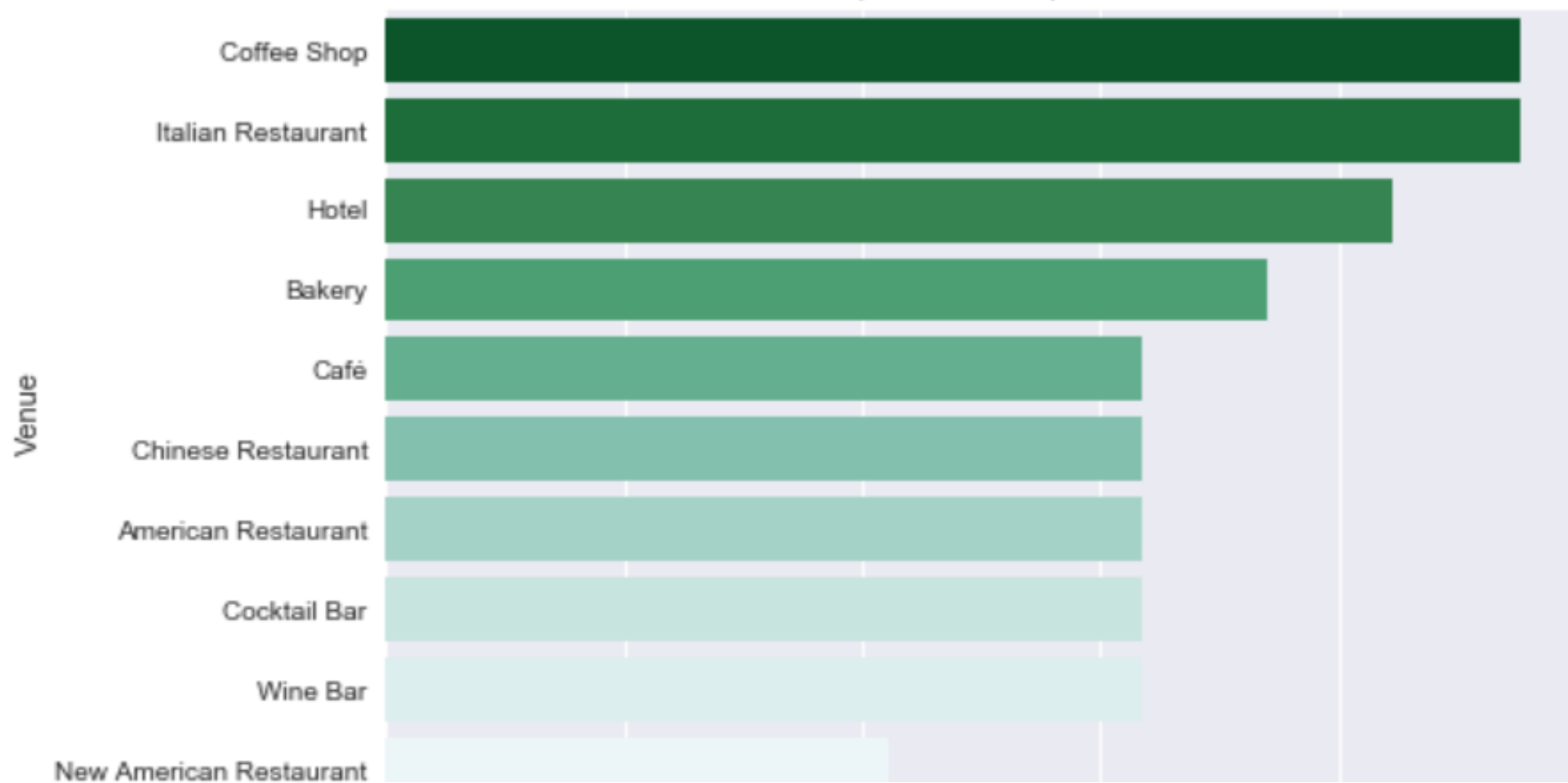


- **Cluster 0:** parks, gym and pizza places.
- **Cluster I:** trails, playgrounds and scenic lookout.
- **Cluster 2:** no strong frequency for its venues.
- **Cluster 3:** trails, playgrounds and parks

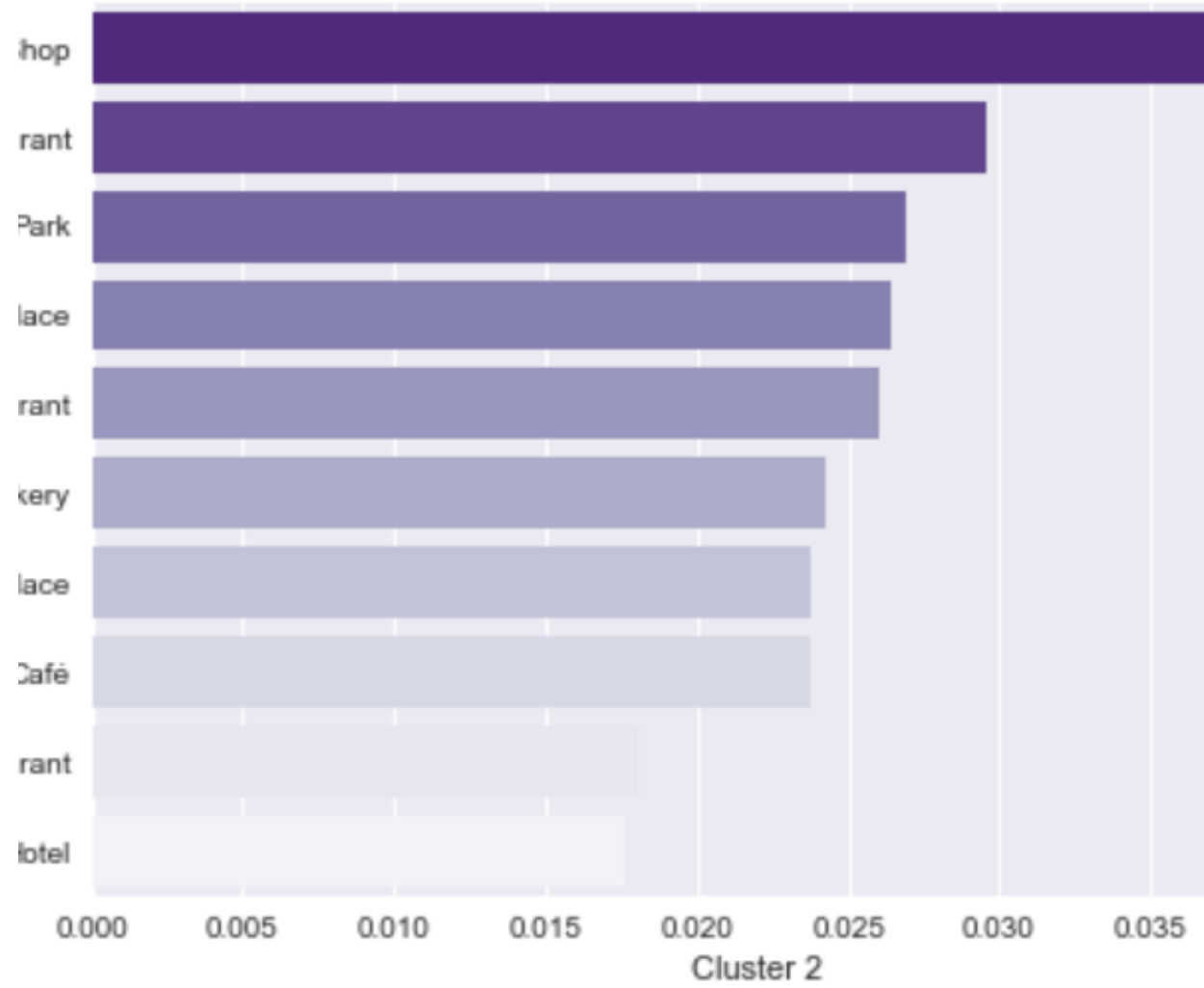
Most frequent
venues per Cluster

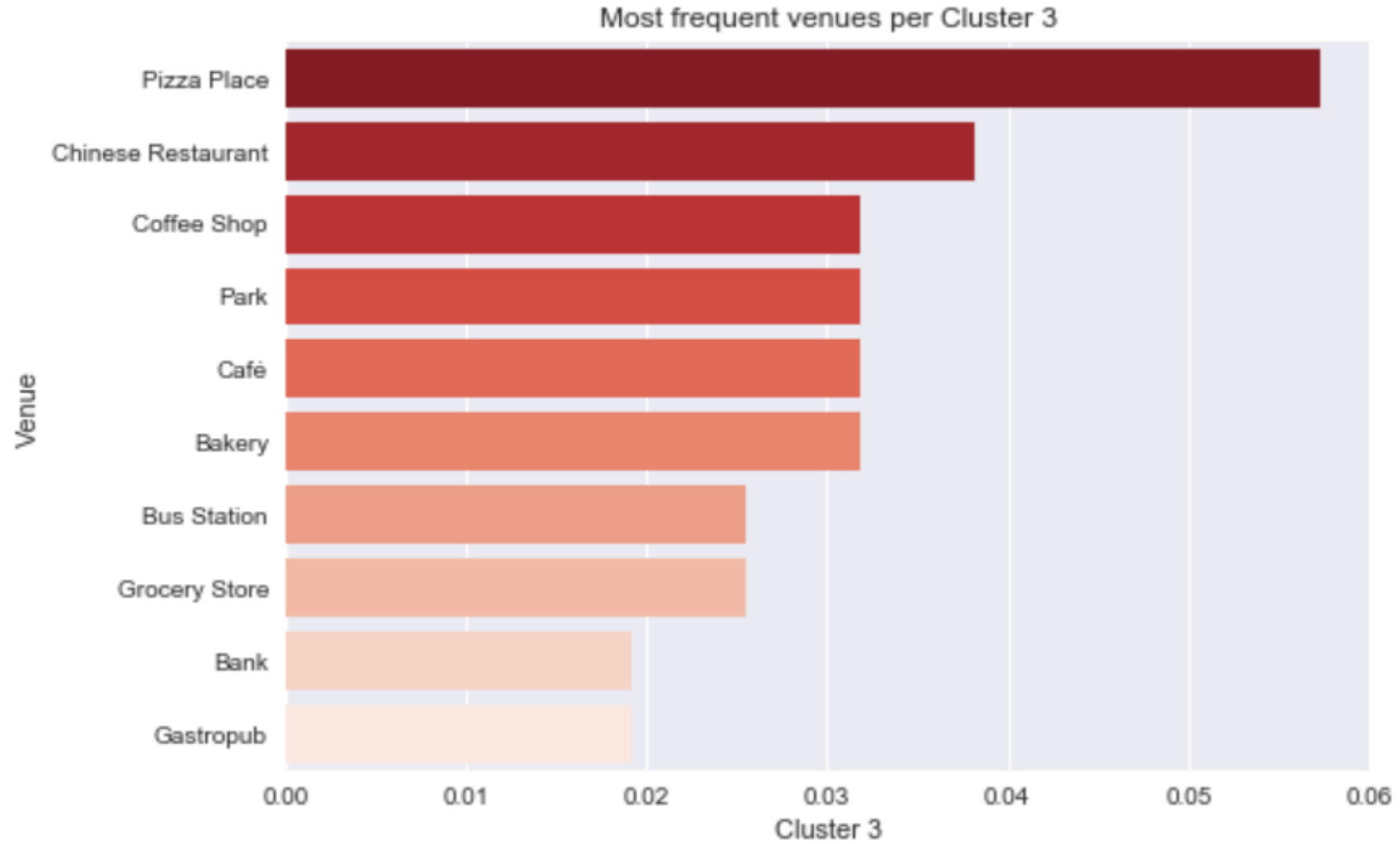


Most frequent venues per Cluster 1



Most frequent venues in Cluster 2





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 still useful for differentiating neighborhoods within a city and finding similar ones in another.