Capstone Project- The Battle of Neighbourhoods

Applied Data Science Capstone by IBM

Gábor Bence Fehér, October 2020

1. Introduction: Business Case

The aim of this project is to gather information which can help us choose between the cities of San Francisco, California state, United States of America and Toronto, Ontario state, Canada. The specific goal is to find data as to which city would be ideal to open an establishment, such as a Hungarian restaurant. After a city is chosen, a recommendation can also be made for the location inside the city in the form of a neighbourhood type.

For this we need to first identify the neighbourhoods themselves, then find out what the common establishments there are. This will help us decide if the neighbourhood is a good choice for a restaurant, based on demand and competition.

2. Data

Data for this notebook is scraped from websites or downloaded in file format.

A list of neighbourhoods of San Francisco was found on the webpage of the New York University in geojson file format. You can find the file through this URL: https://geo.nyu.edu/download/file/ark28722-s71k5h-geojson.json. The district names were extracted from this file and cross checked with the list available on Wikipedia on this page: https://en.wikipedia.org/wiki/List_of_neighborhoods_in_San_Francisco#Midtown_Terrace.

Location information of these districts were collected with Geopy's Nominatim API. During this exercise I discovered that the district names in both the .json file and on the Wikipedia page sometime can't be found through this API. To remedy this I consulted the wiki page's descriptions of the districts and selected alternative names for these districts, making sure that the geographical location stays in or around the same place - the relative centre of the district. This is described further below in the code.

Postal code information of Toronto was scraped from another Wikipedia page: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. This was further enriched with location information found here: https://cocl.us/Geospatial_data. The file contains the coordinates for each postal code.

Venue information was acquired from Foursquare's API, with a free account. https://developer.foursquare.com/

To goal is to use this data to determine the two cities' characteristics through the venues found in their districts.

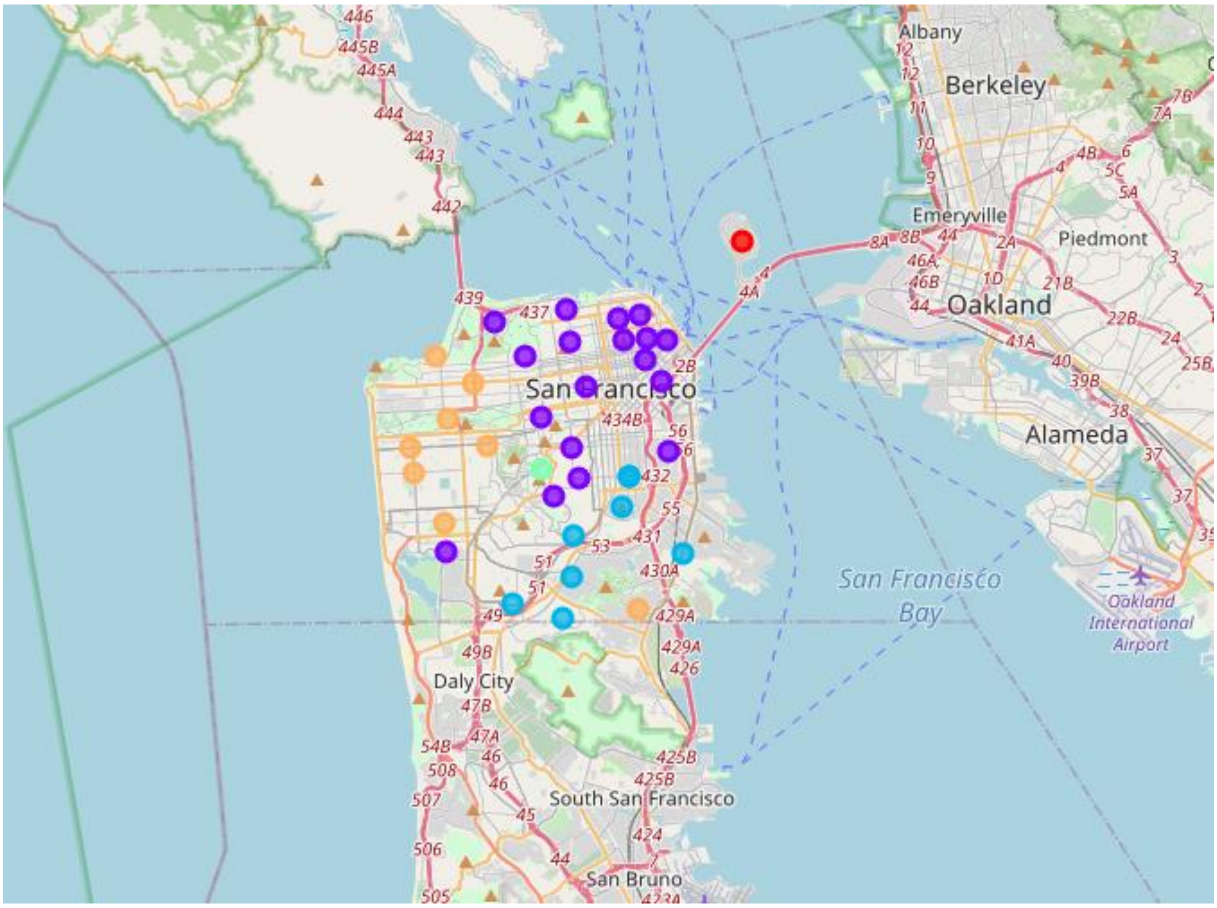
3. Methodology

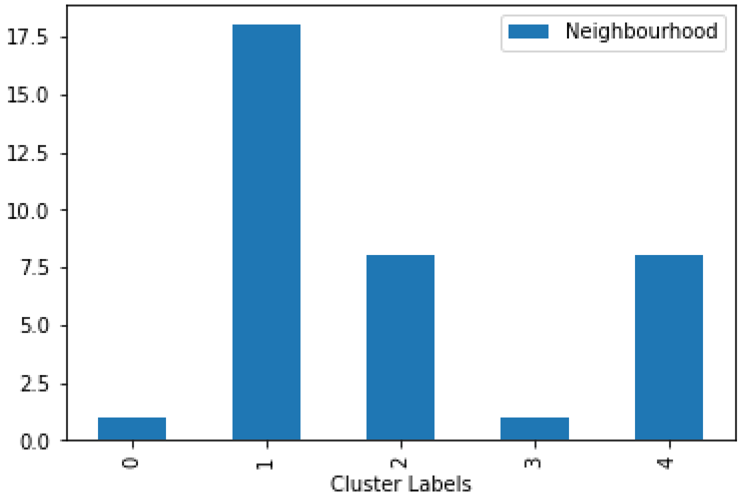
The goal is to identify neighbourhoods and categorise them based on the existing establishments. K-means clustering will help us identify clusters of neighbourhoods. To find the correct k number we will utilise the elbow point method. For a more detailed breakdown of this please review the notebook.

4. Results

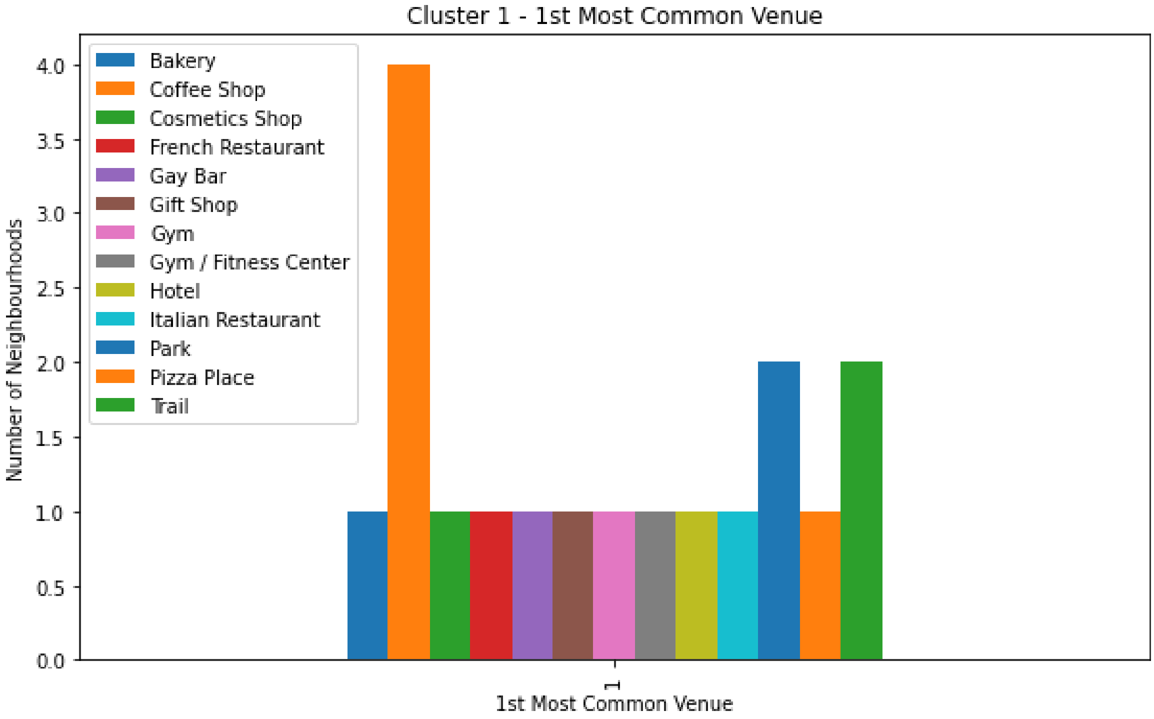
During our analysis we managed to identify 5 clusters of neighbourhoods based on venues in both cities.

In San Francisco most of the neighbourhoods got grouped in 3 major clusters, clusters 1, 2 and 4.

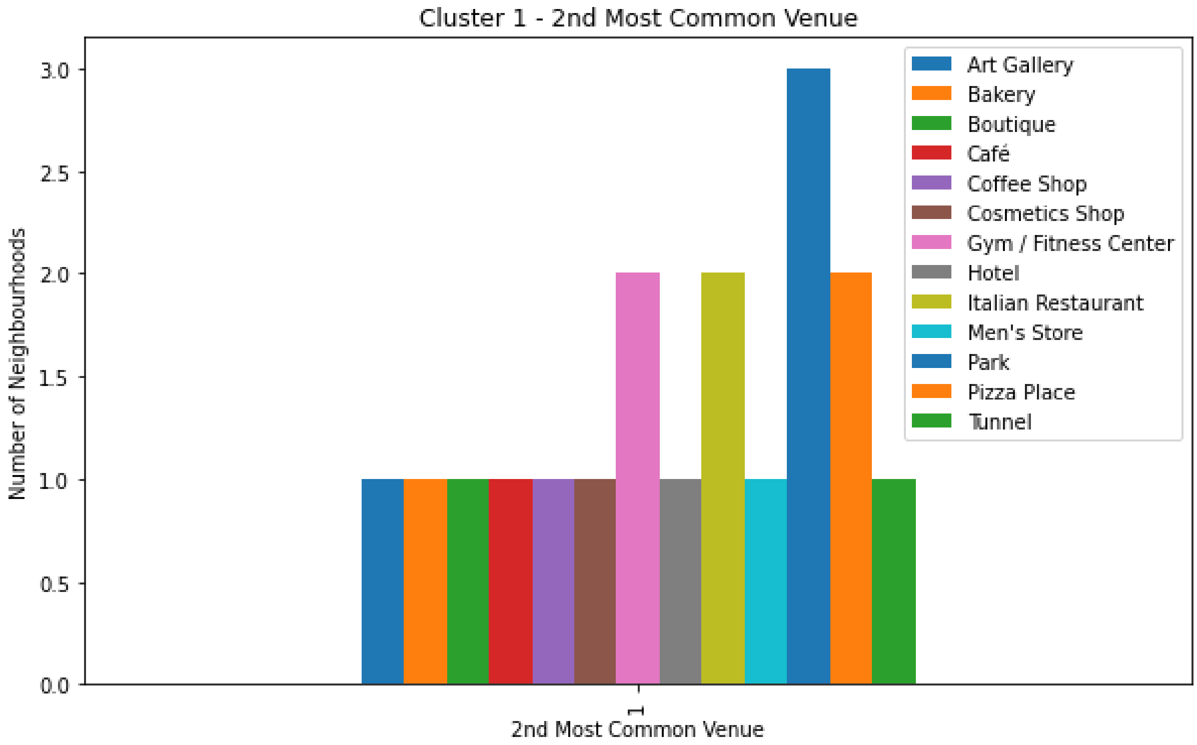




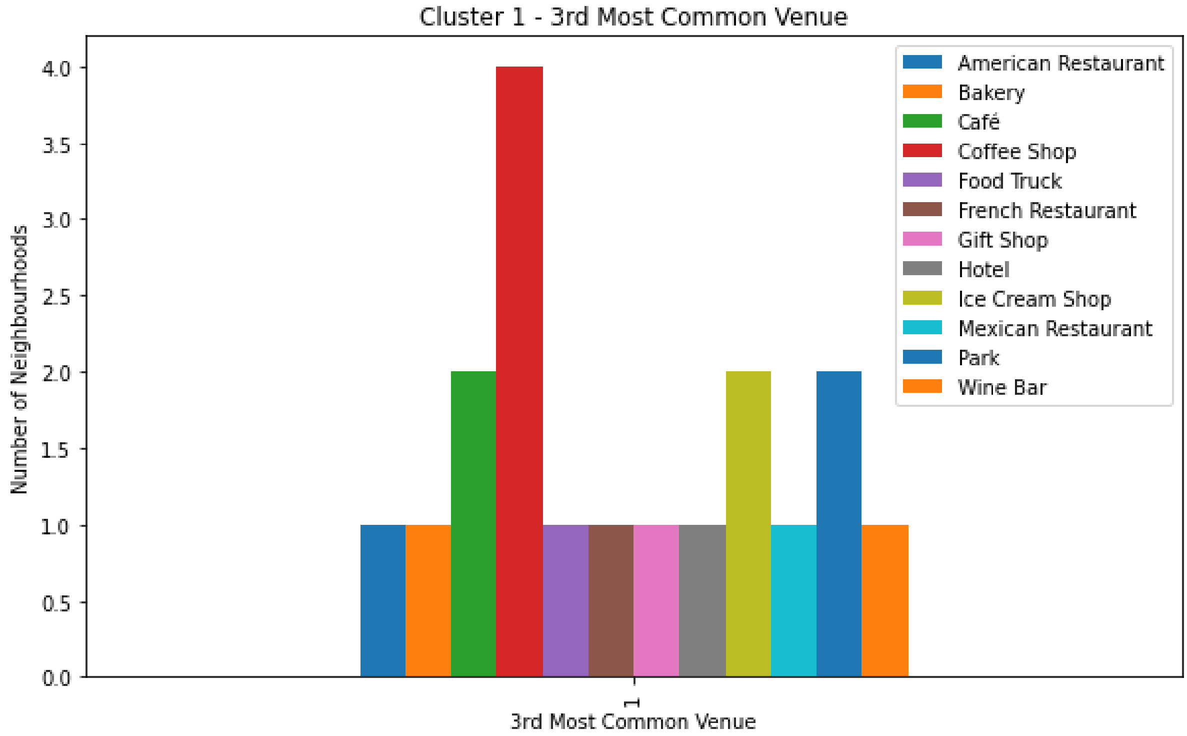
Cluster 1 is mostly made up of coffee shops, parks and trails.



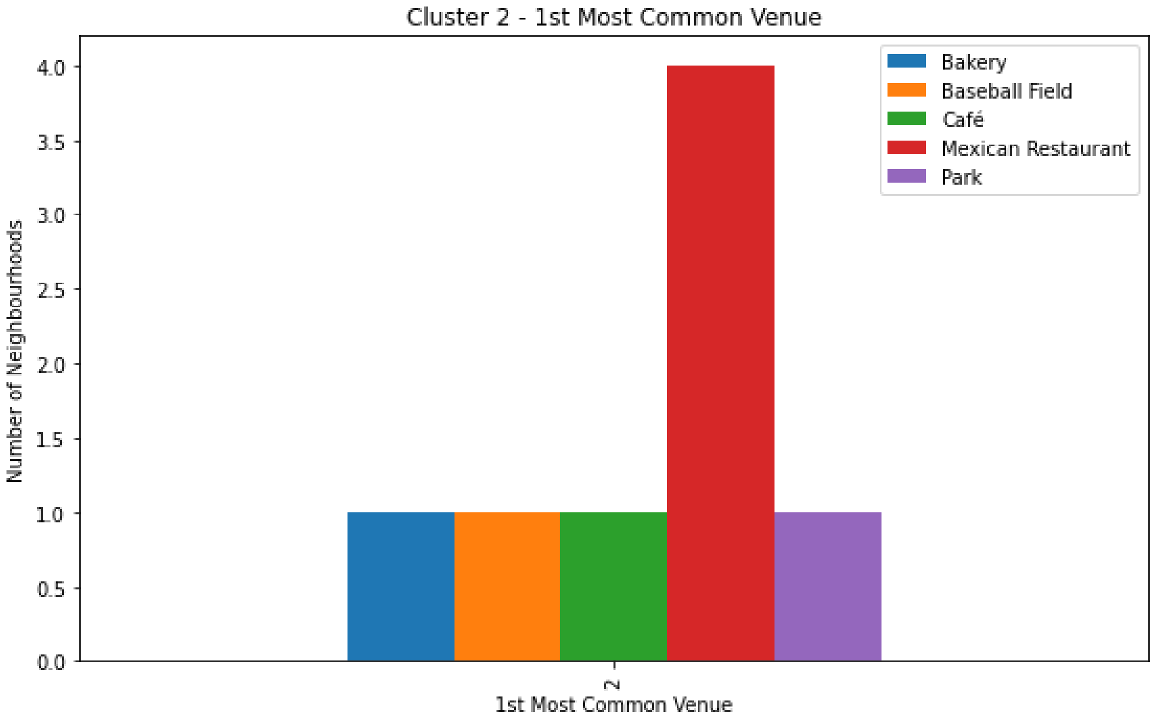
In addition gyms, Italian restaurants and pizza places are common as well. Cafés and ice cream shops can be often found here too.



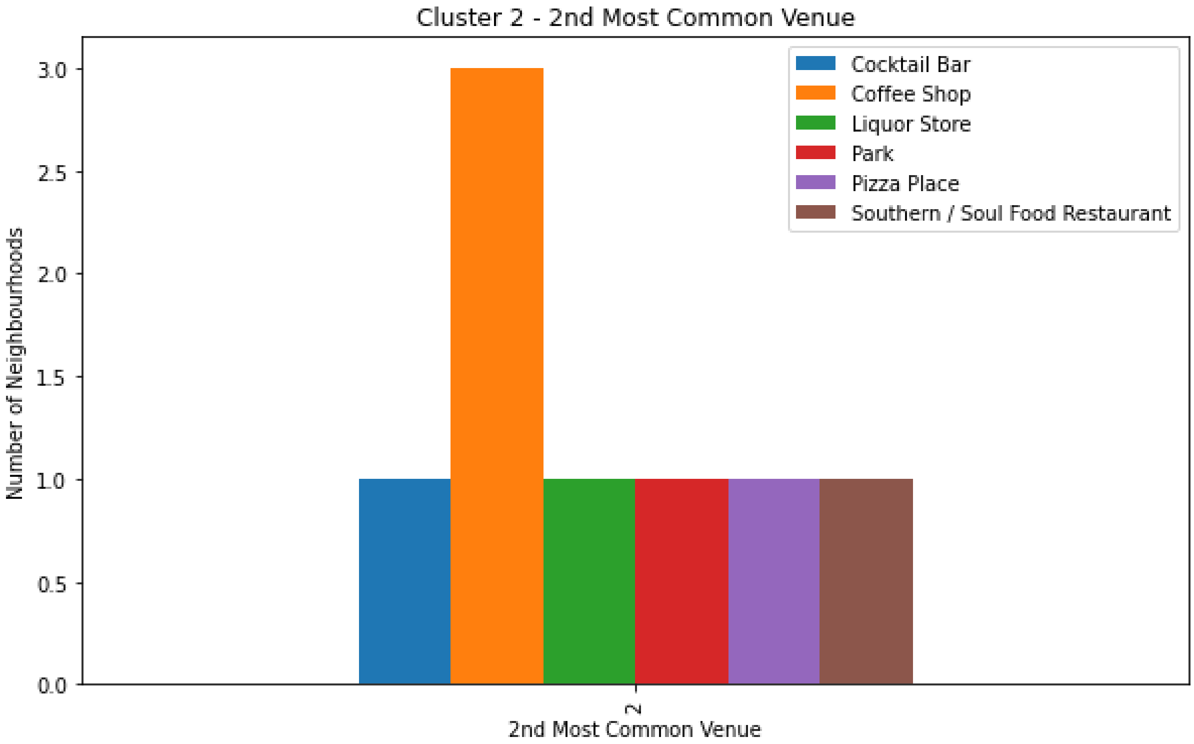
French, American and Mexican restaurants deserve a mention too.



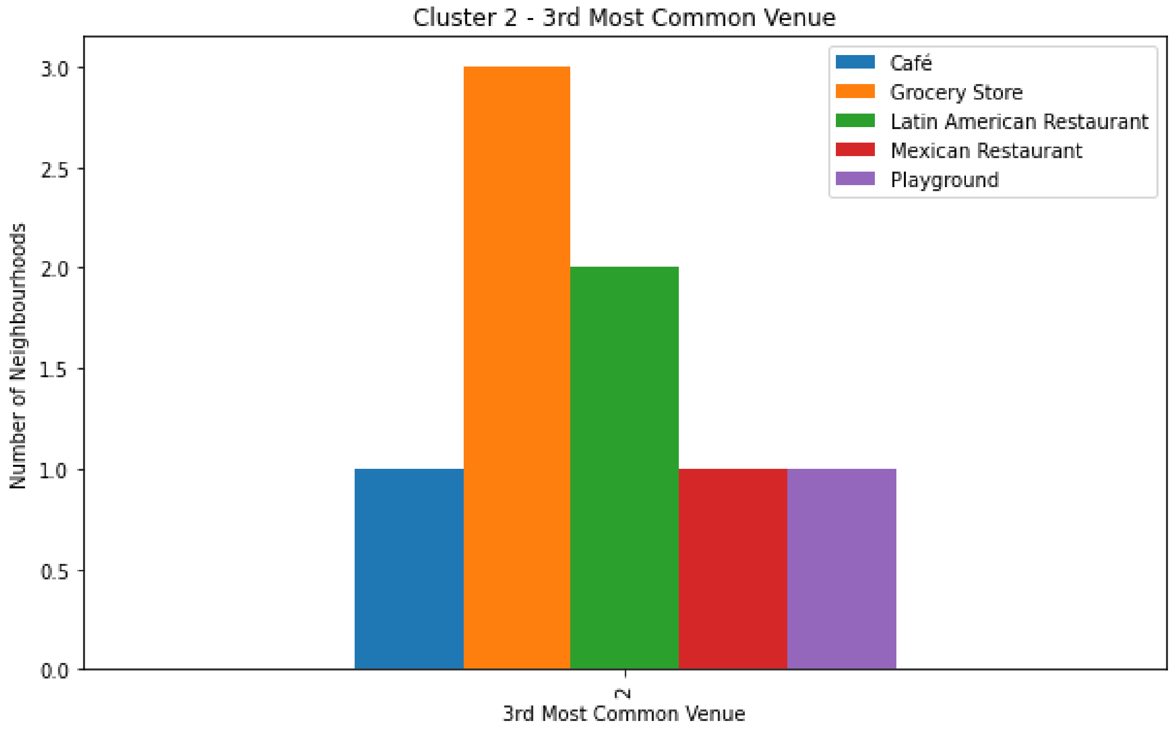
Cluster 2 differs in that that the most common venues found here are Mexican restaurants. We can also see baseball fields here.



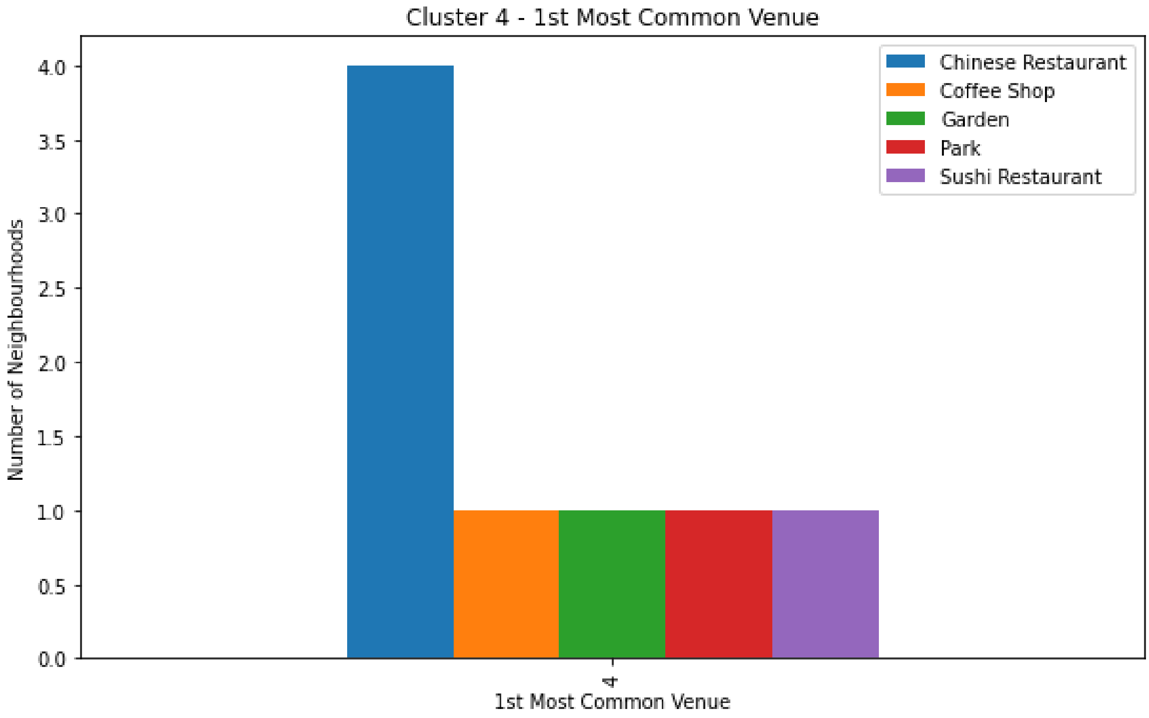
Coffee shops and pizza places are ever present, but southern / soul food restaurants make an appearance as a difference.



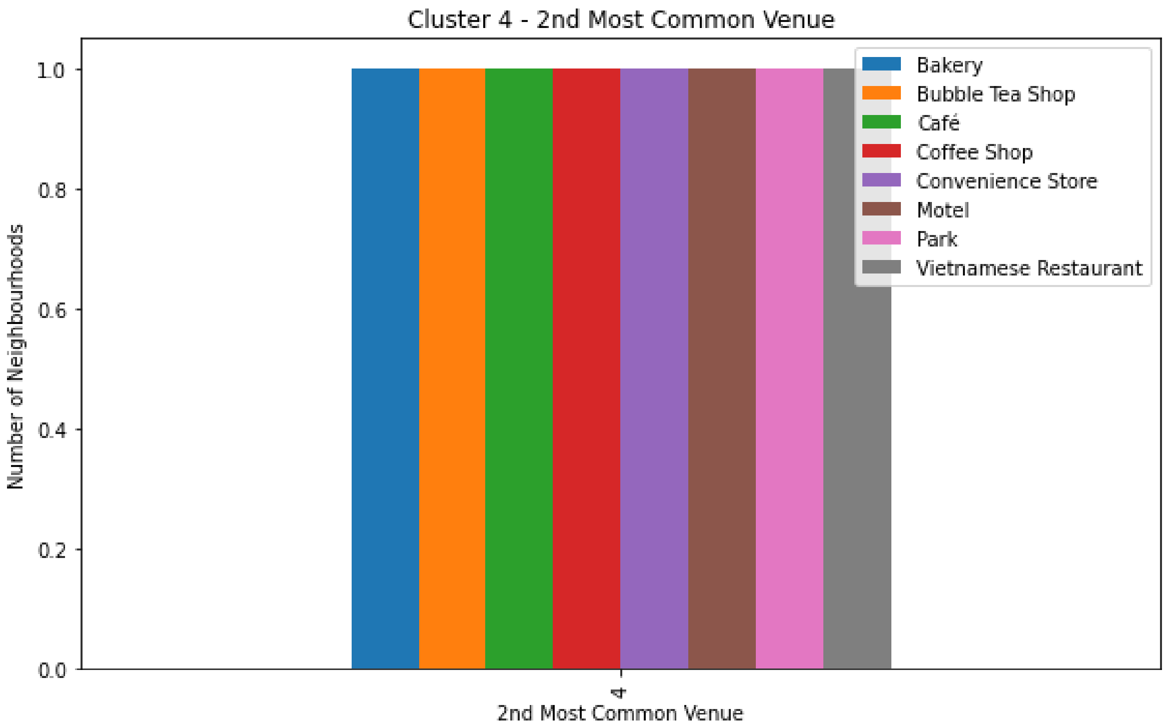
Next grocery stores, Latin American Restaurants and playgrounds prove to be common.

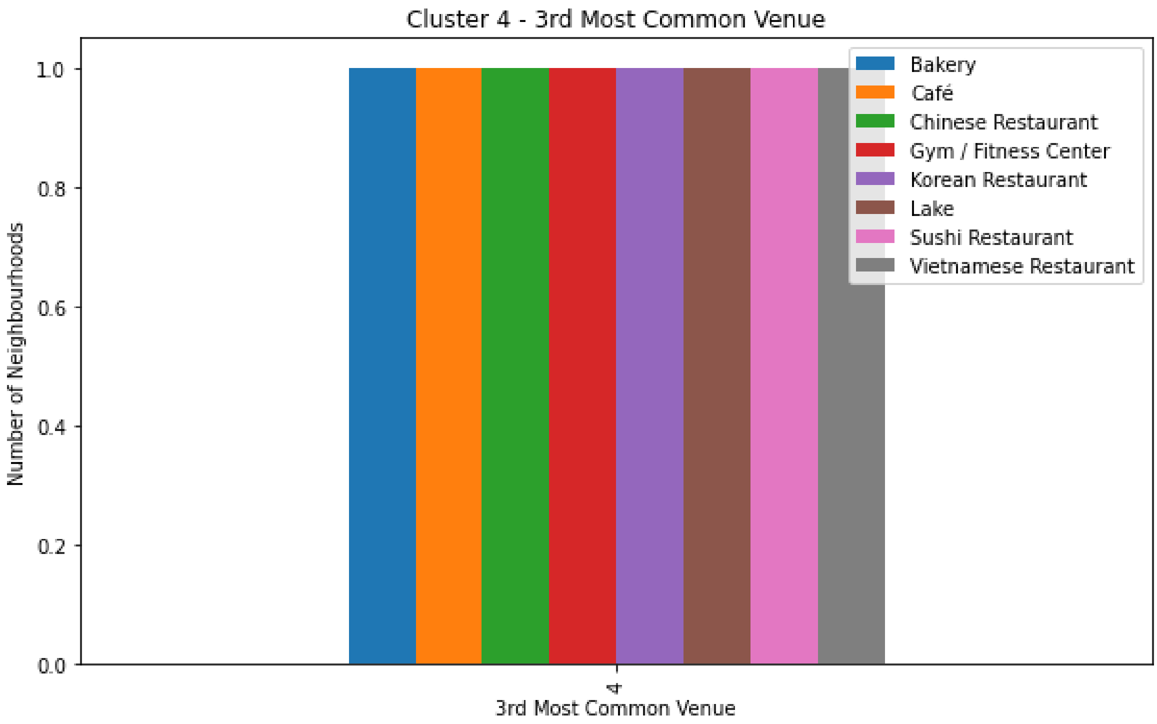


As always in cluster 4 we can find coffee shops and parks. However it is completely dominated by Chinese Restaurants.

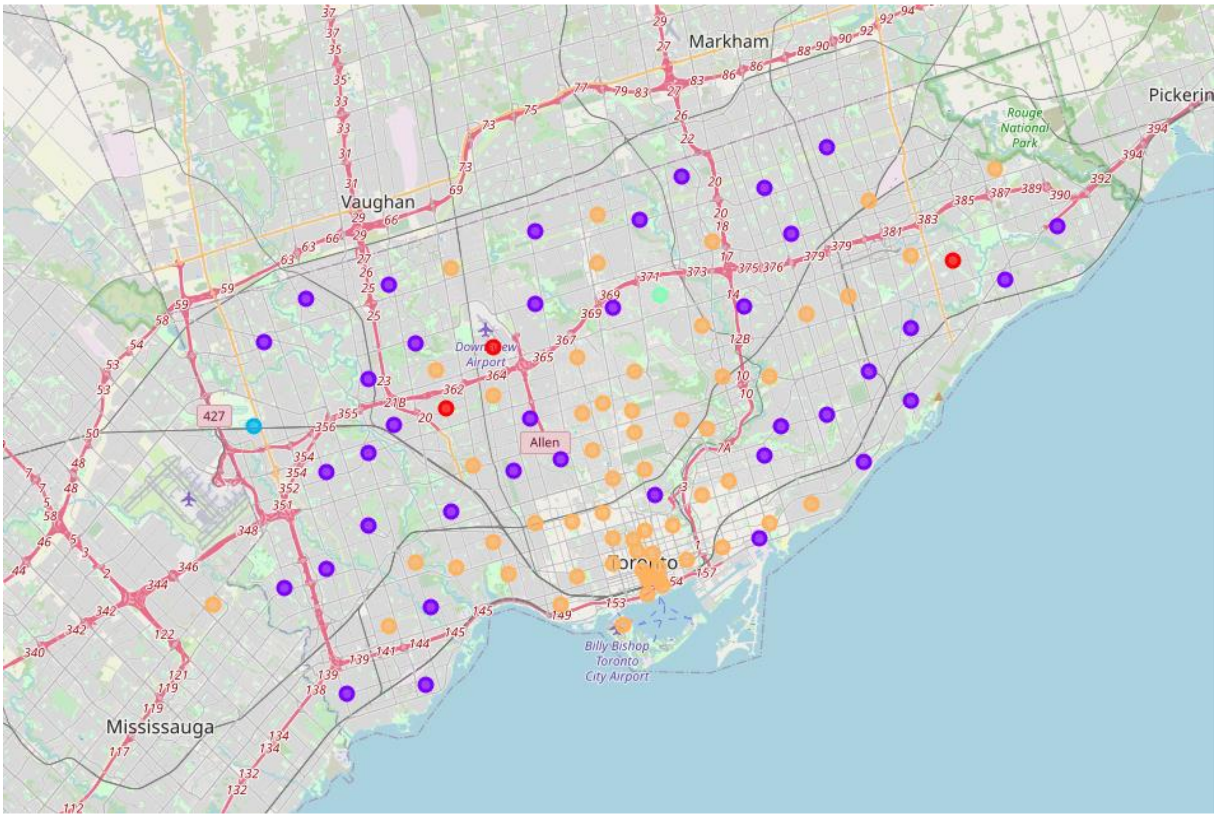


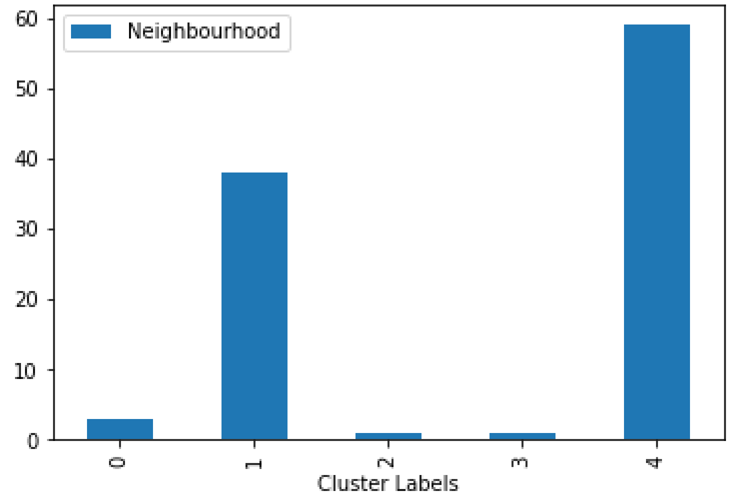
Sushi restaurants are very common too. Otherwise it is a mix bag of the usual bakeries, cafés and coffee shops. What makes the cluster different is the common appearance of Asian restaurants, such as Chinese, Korean, Sushi and Vietnamese restaurants.



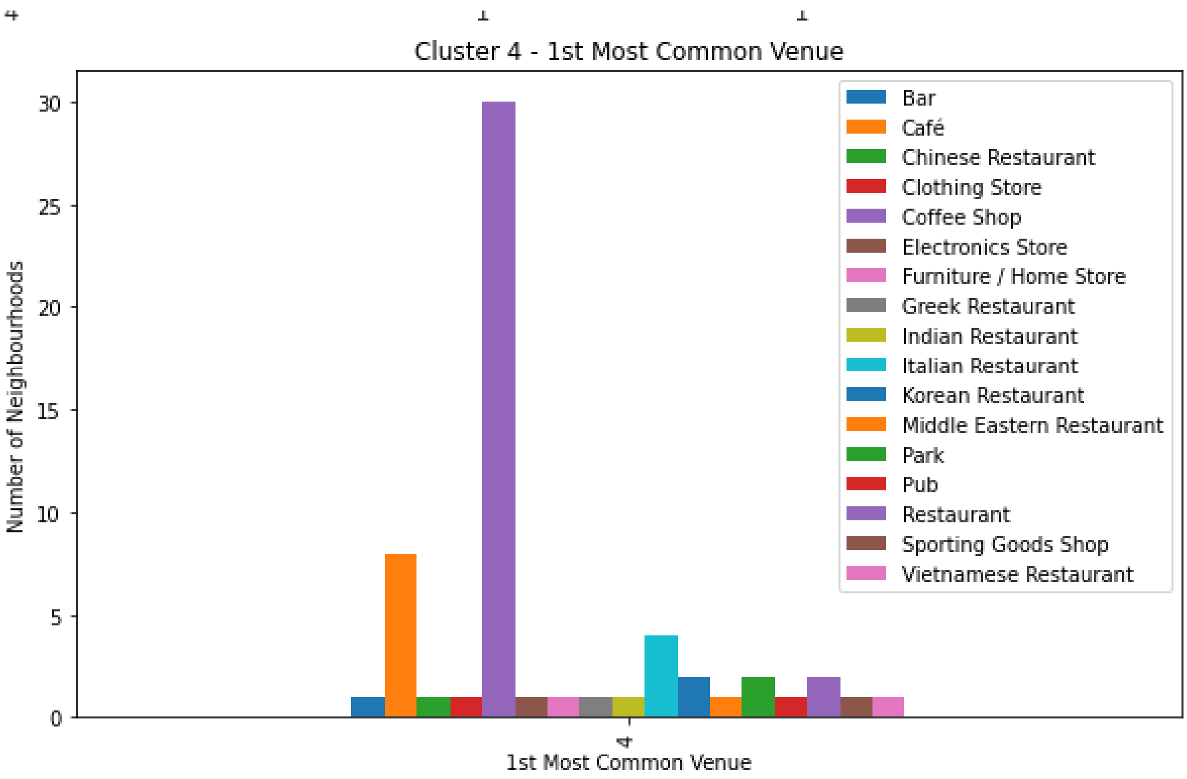


In Toronto two big clusters contain almost all of the establishments.

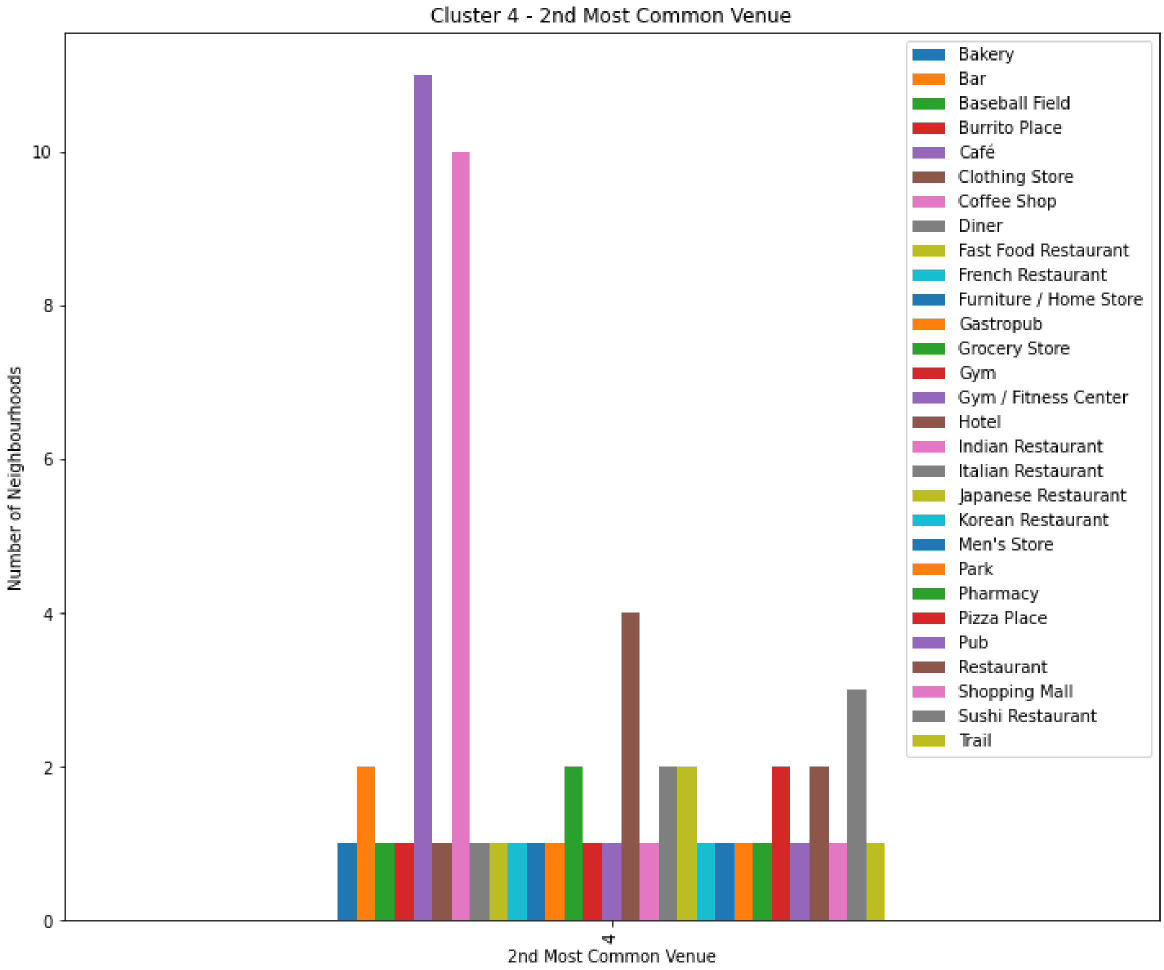




Cluster 4 is the biggest one. It’s absolutely dominated by coffee shops that proved to be the most common venues by far. Cafés are a distant second.

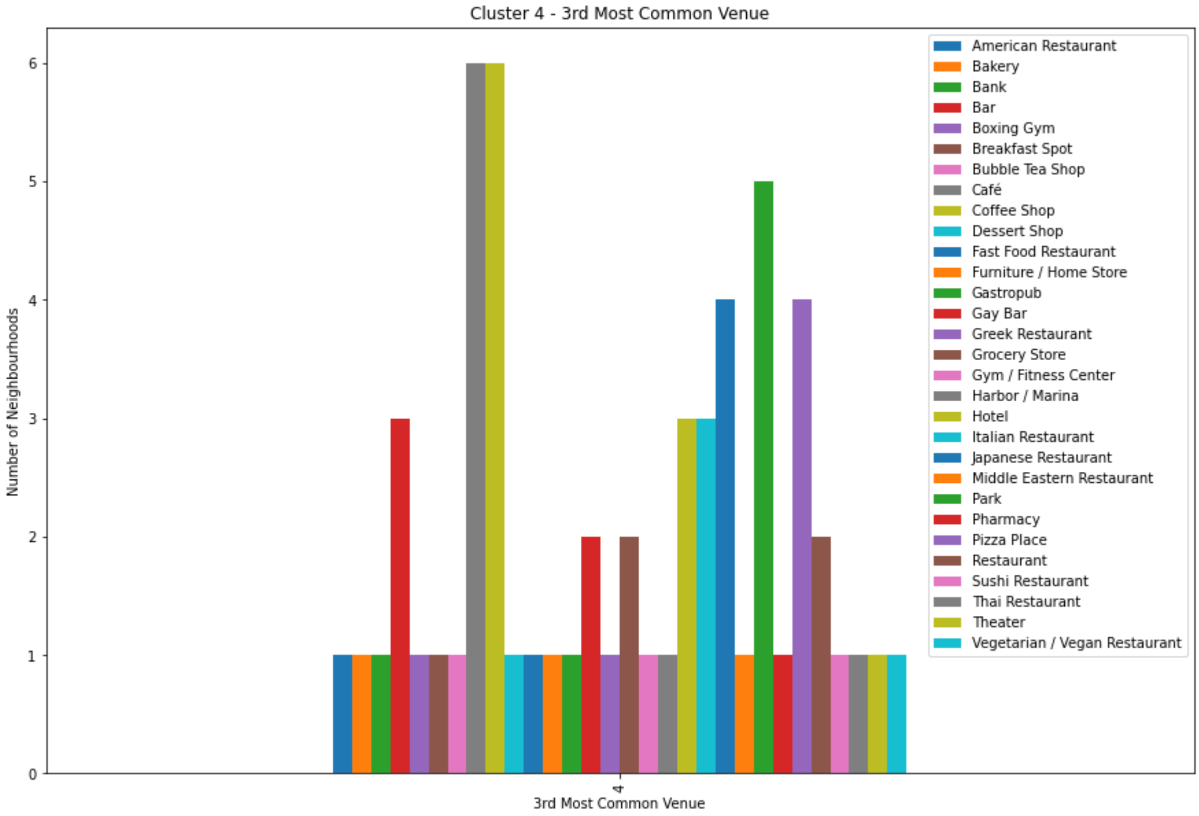


In terms of restaurants Italian is the most common, Chinese, Greek, Indian, Korean, Middle Eastern and Vietnamese restaurants all make an appearance along with local restaurants.

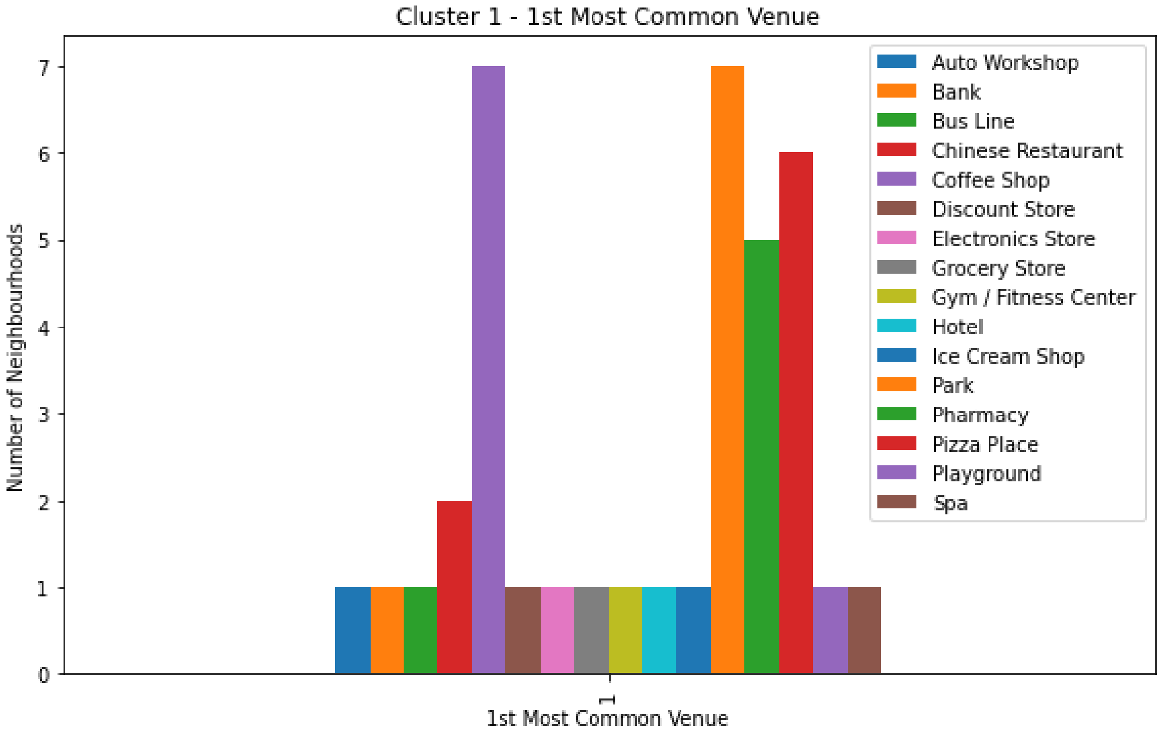


To a lesser extent we can find burrito places, diners, fast food restaurants, French restaurants, gastropubs, Japanese restaurants and Sushi restaurants too. Thai and vegetarian ones are also common.

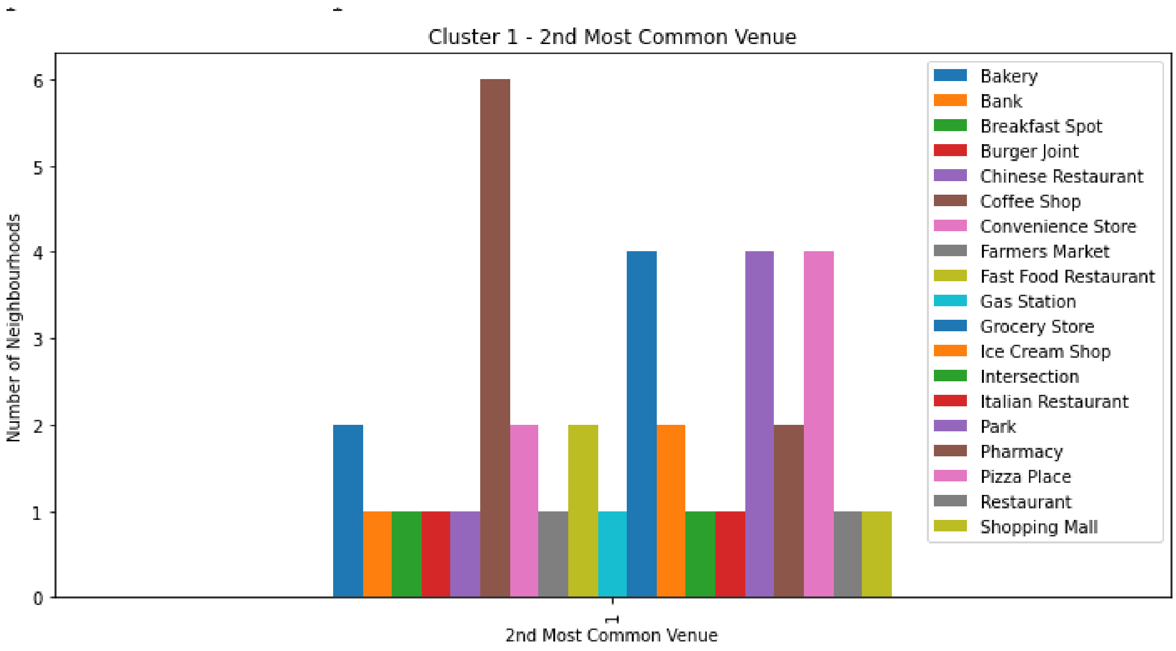
```
[1 rows x 29 columns]
```

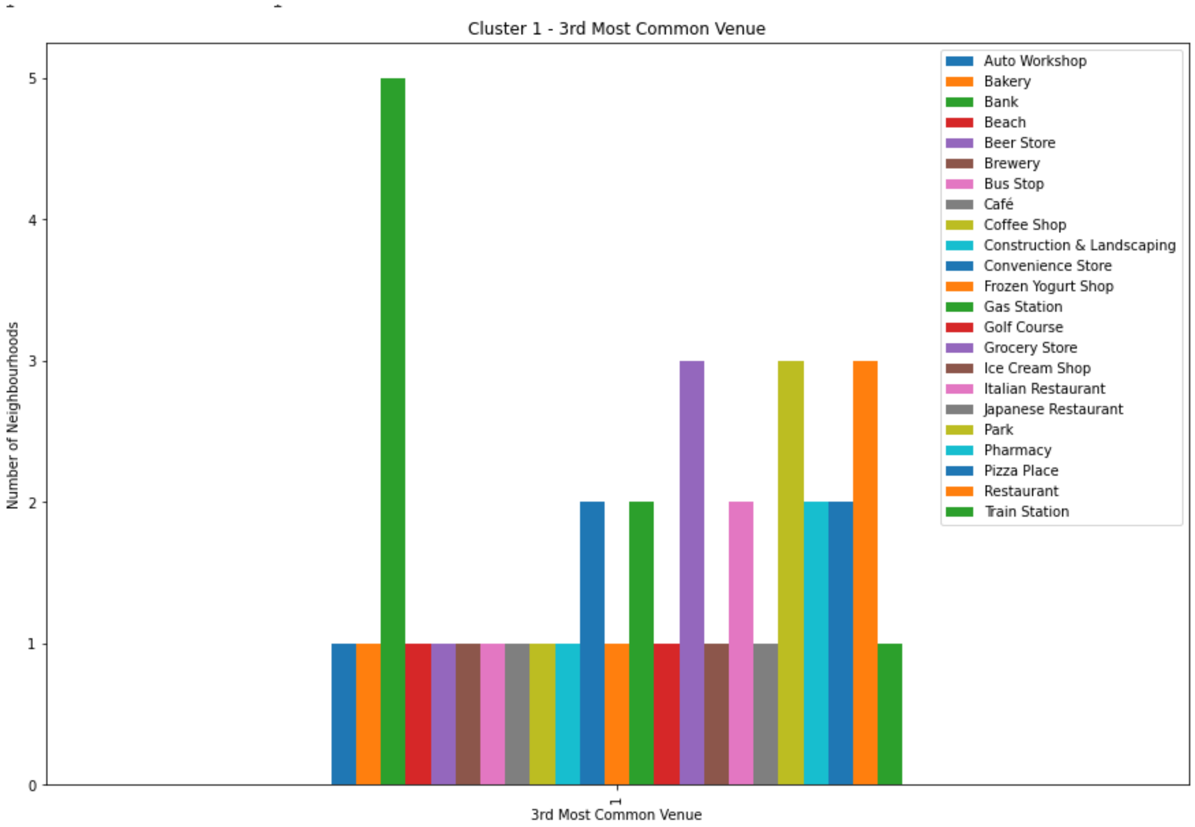


In cluster 1 we find other types of venues such as parks and pharmacies more often. Coffee shops are the most common along with parks.



In terms of restaurants Chinese ones and pizza places are rather common. Burger places and fast food restaurants along with Italian and local ones can be found as well.





We can observe that even though there are overlaps, the make up of the clusters of the two city are quite different in what is most frequent and where. The big exceptions are coffee shops that seem to be incredibly common and popular everywhere.

5. Discussion

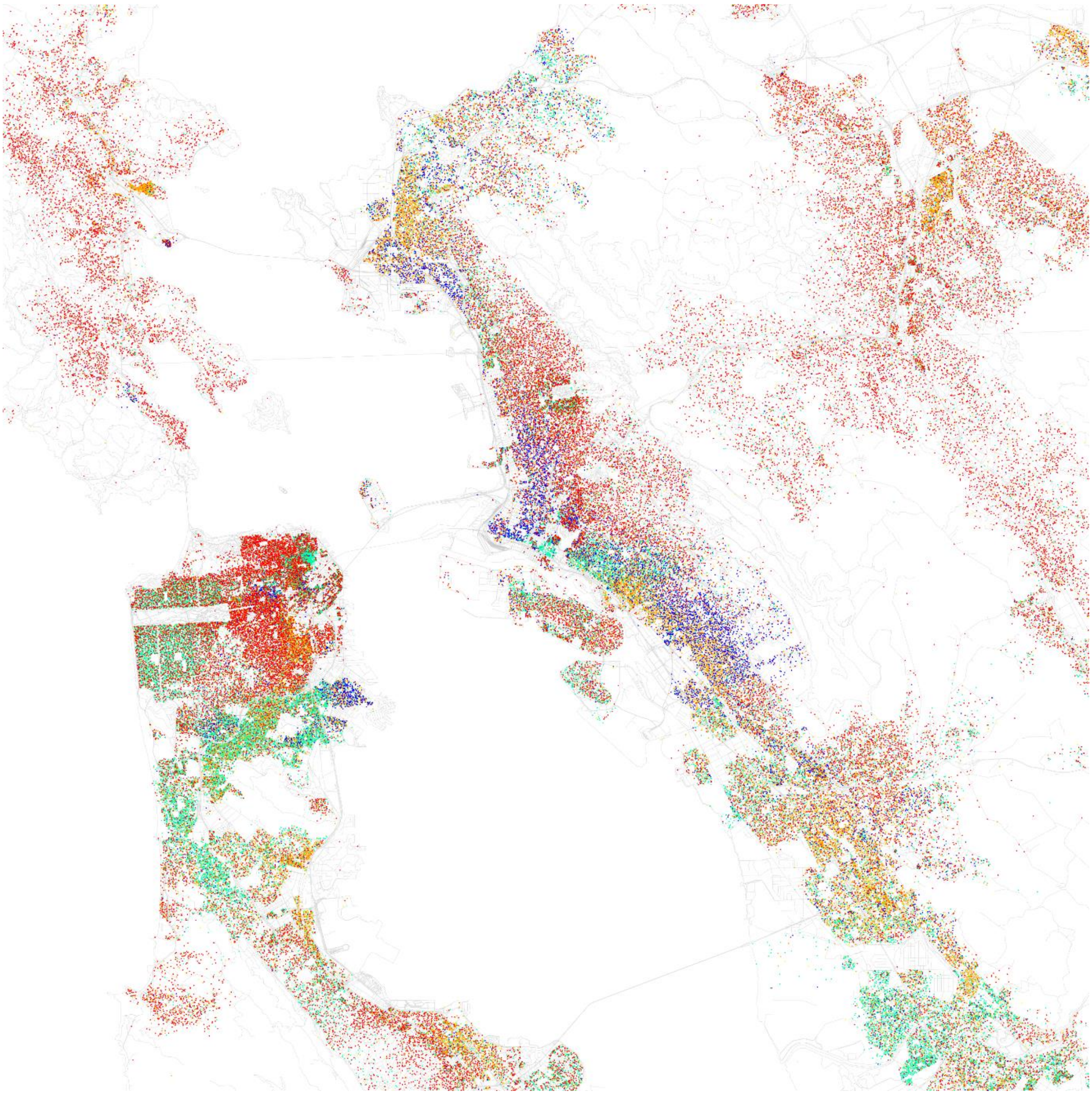
What is relatively apparent from the initial analysis is that San Francisco compared to Toronto has a smaller variety of venues to offer. The clusters seem to be separated more clearly. The former may be easily explained with the difference in physical size between the two cities: according to

Wikipedia San Francisco together with its county is smaller than the single-tier size of Toronto (about 600 square kilometers vs 630). San Francisco’s population is estimated to be around 880 thousand, while Toronto boasts a population three times as large at 2,7 million.

Still looking at San Francisco, we have to notice the well separated blocks of clusters. This is very interesting and from looking at what kinds of venues are popular in the clusters, we can tell that cluster 1 is generally business and European oriented in their offerings. Cluster 4 is quite more focused on providing for the Asian populace. Cluster 2 seems to be oriented towards the Latin populace.

This is supported by the 2010 census map of San Francisco, found behind this URL: https://upload.wikimedia.org/wikipedia/commons/a/ae/Race_and_ethnicity_2010-_San_Francisco%2C_Oakland%2C_Berkeley_%285560477152%29.png

As a legend for the racial distribution: red is white, black is blue, green is asian, hispanic is orange, black is other. The clustering map above is not as granular but the outlines are more or less there.



Based on this population I'd recommend that if the choice falls to San Francisco, establishing a Hungarian restaurant would make most sense in cluster 1 first and cluster 2 second.

Looking at Toronto we can recognise a different type of layout. It seems to be more like a downtown/business area/centre – residential split. Both of these clusters could be a good location for the restaurant. The choice depends on other factors: what the target audience of the restaurant would be, the crime and wealth details of the neighbourhoods, etc.. This could be the subject of further analysis.

At the end of our discussion I would think that Toronto as a city, and any one of its big clusters we established would be the better choice for our restaurants, due to how the population seems to blend in with each other. This could potentially mean a bigger and diverse customer base, contrary to the conclave like situation of San Francisco.

Above mentioned information would have to be researched for a well rounded recommendation and decision, which weren't in scope for this project.

6. Conclusion

In this project we looked at two great cities and tried to derive information based on venue data. We were able to establish clusters of neighbourhoods using machine learning and had findings about them for both towns. We made recommendations as to which parts of the cities and which city general would make a good choice for our Hungarian restaurant. We also made recommendations for further analysis of this business case to make a well informed and well rounded decision.