Study of the Neighborhoods of Toronto and New York

Sebastian Brandt

April 17, 2020

1. Introduction

1.1 Background

On this project I'm going compare the neighborhoods for the cities of Toronto and New York, deciding if is better to conduct a study of Toronto and New York, or Manhattan instead. The cities are compared each other taking in consideration their general features and the most common venue categories, like public spaces and stores.

1.2 Problem

When someone is looking to invest in a profitable business, first must look for opportunities, needs that have not been covered yet and discover the most successful cases, like how many are competing for the same market. Using analytics is possible to extract this information.

For this purpose, I'm using the machine learning clustering algorithm of k means. This allow to find groups of neighborhoods with similar features and some methods the find the optimal number of clusters, this with the help of visualization libraries as matplotlib and seaborn to better understand the results and perform the analysis.

1.3 Interest

The objective is to extract useful information pointing main differences between the clusters, cities, and opportunities, for people looking to open a business or even for instance, wants to move to a similar place where he can make the most of it. I chose to include Manhattan for testing the idea that could be more similar with Toronto and better for conducting this study.

2. Data acquisition and cleaning

2.1 Data sources

To achieve my goal I'm using the Foursquare API to get the most common venues categories for every neighborhoods, postal code data from Wikipedia for Toronto, geospatial coordinates of latitude and longitude to get the top venues of each neighborhoods and from the New York University spatial data repository published by the New York (City). Department of City Planning.

2.2 Data cleaning

The first obstacle I encountered was the table of postal codes for Toronto from Wikipedia having a different format than the one needed. To solve this issue, I used an older version that at the same time satisfy the conditions of the right amount of neighborhoods. Cells with Not assigned boroughs are ignored. Boroughs with not assigned neighborhoods, are filled with the borough name instead. Neighborhoods with same Postal code are combined into one row, with the neighborhoods separated with a comma. New York and Manhattan datasets are already prepared to perform analysis, not cleaning is needed.

2.3 Feature selection

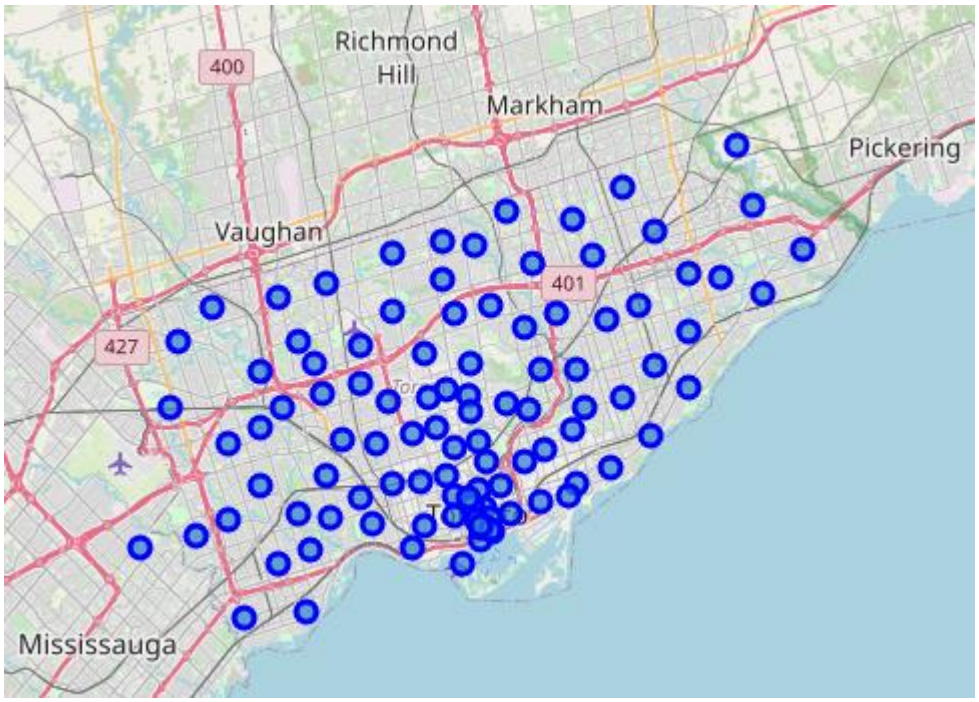
For the New York dataset, it is stored in a .json file with all the relevant data inside the feature key, containing the list of neighborhoods. All non-relevant data from the file then is discarded. As with Toronto the features of Borough, Neighborhood, Latitude and Longitude are used. Then using the Foursquare API the top 100 venues for every neighborhood within a radius of 500 meters are added, finally adding the 10 most common venues to the neighborhood.

3. Exploratory Data Analysis

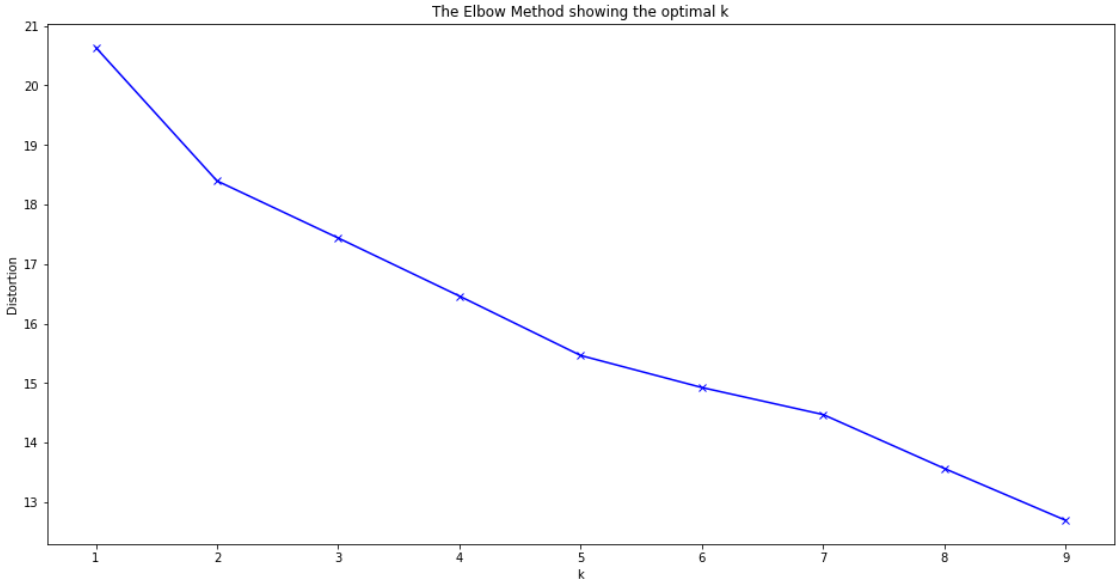
To perform the data analysis, I'm using the k-means algorithm to create groups of similar neighborhoods and then categorize them. It uses the most common venues in each neighborhood to assign them to a cluster. The best number of clusters is difficult to determine, but various techniques exist to help with this decision. First, to help visualizing these results later, I'm using the folium library for displaying them into a map.

3.1 The City of Toronto

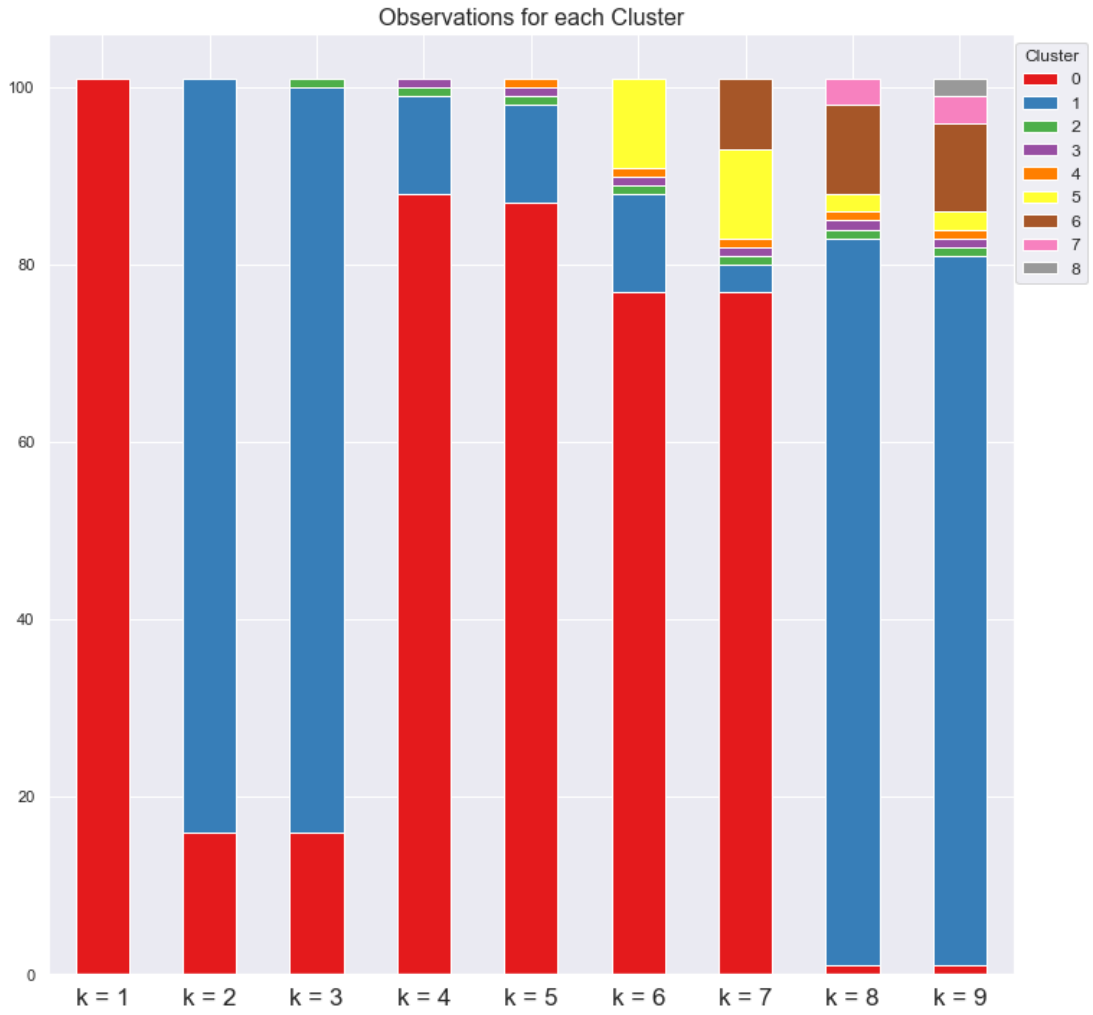
The neighborhoods from Toronto coordinates are plotted into a map pointing at the locations for each one. This map would be useful later to displaying the clusters and distribution of datapoints.



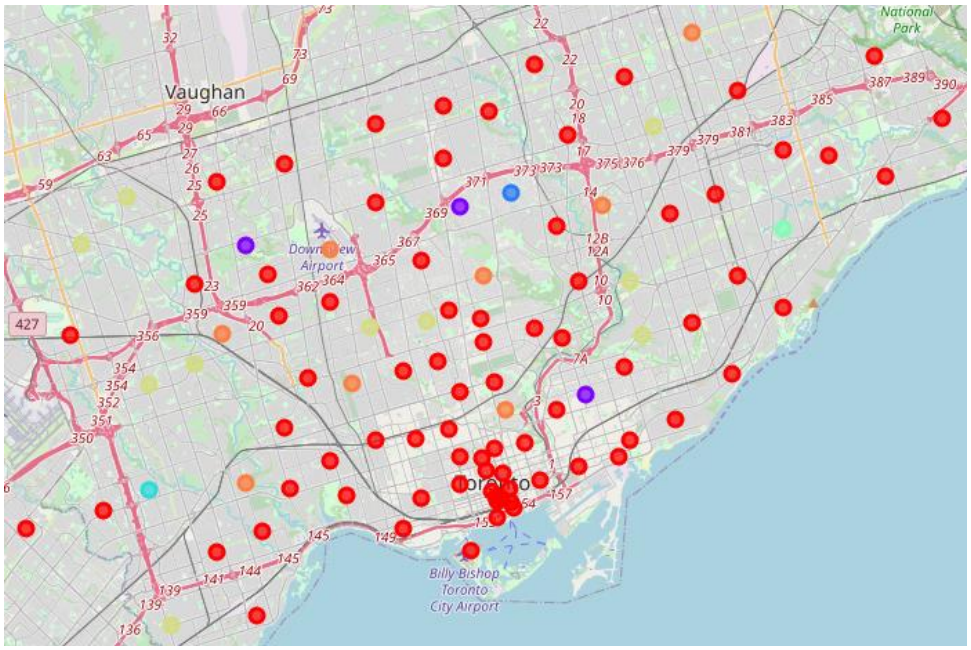
First, to find the optimal amount of K the model is tested with 1 to 10 clusters and the results are plotted, using the elbow method to find the best K. As clearly seen, there is not definitive answer for this question.



An alternative method is used to visualize the amount of data points per cluster, for any given number of K (centroids) to form the clusters.

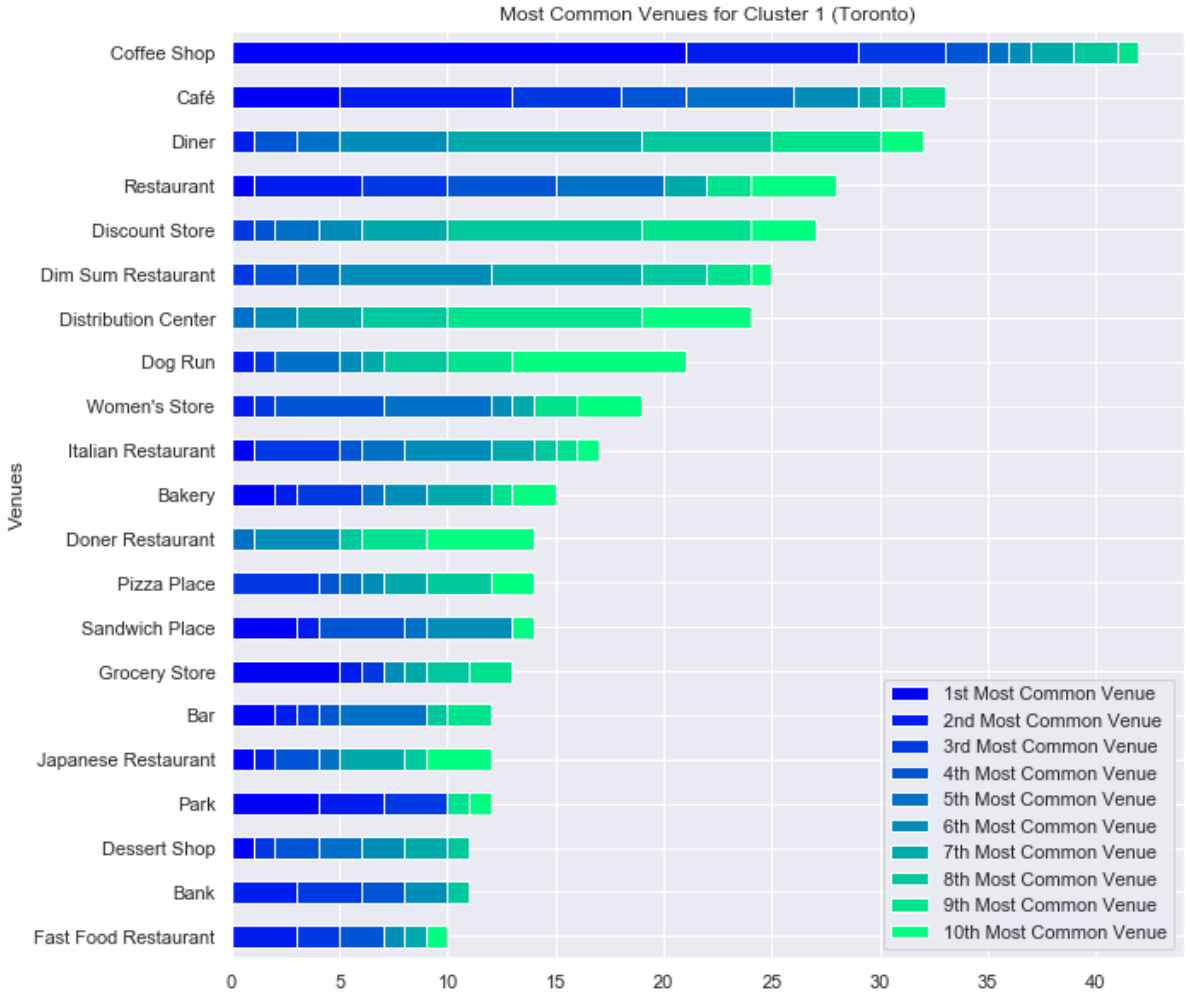


With 2 – 5 centroids, only 2 clusters are created, the smaller ones having only one data point. With 6 and 7 centroids, we have 3 and 4 clusters. Using 7 centroids is possible to find a good balance between the amount of clusters and data points in them.



Data points for each Cluster

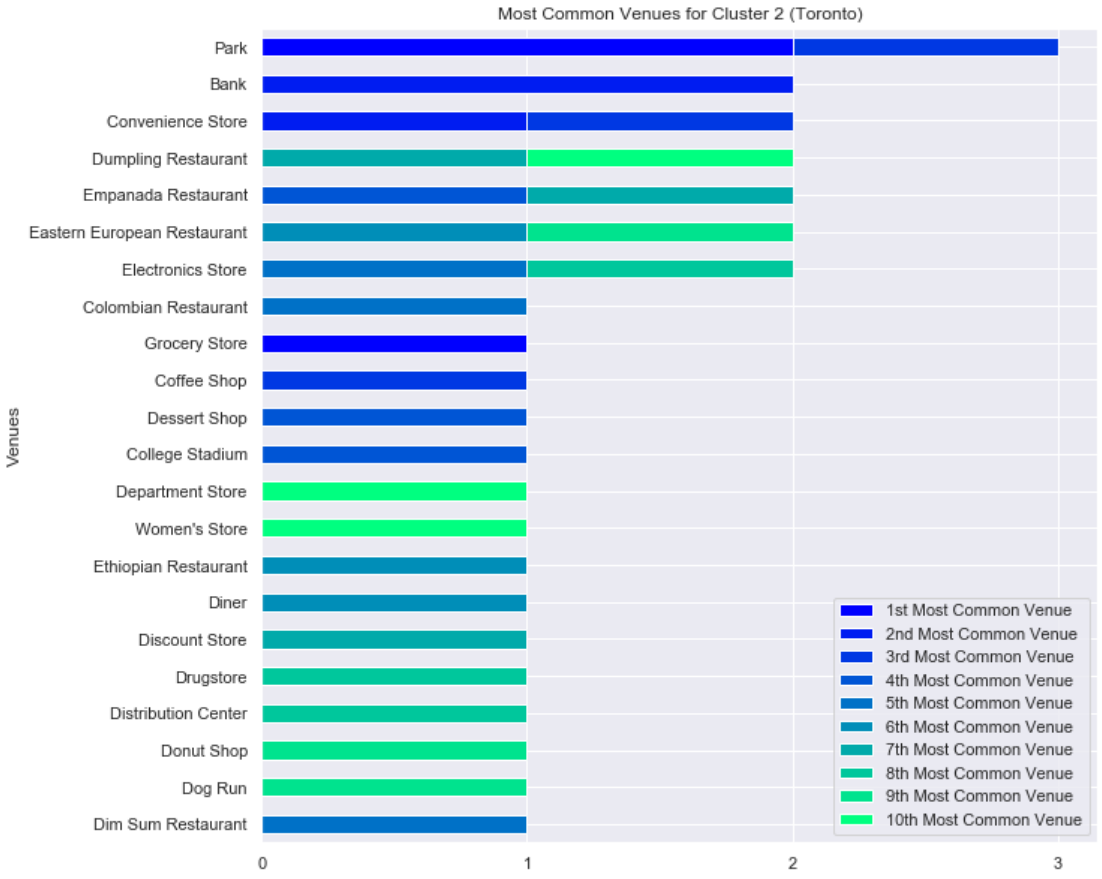
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Data Points	76.24%	2.97%	0.99%	0.99%	0.99%	9.90%	7.92%

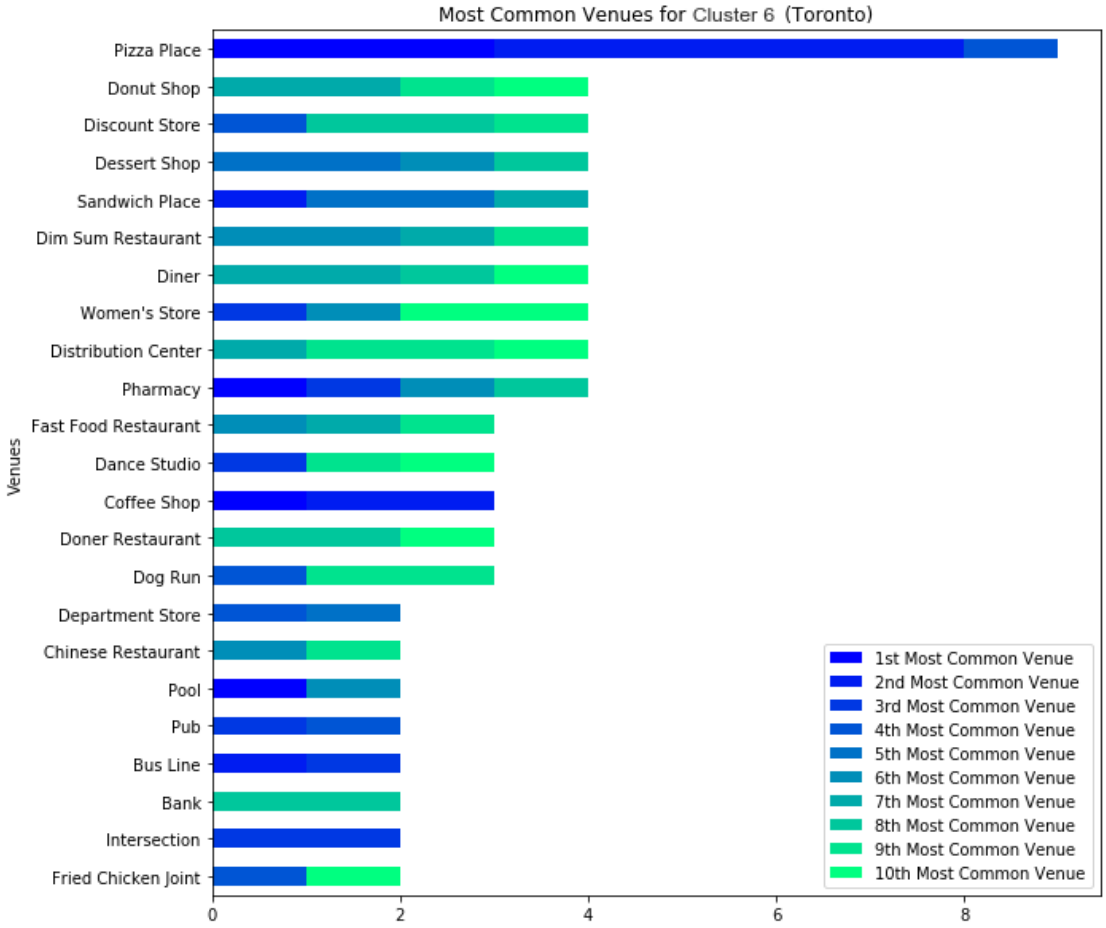


Cluster 1 with 76% of the Neighborhoods is the main cluster. Its dominated by Cafes (Coffee Shops and Cafés), in second place Restaurants (and Diners), including from Chinese (Dim Sum), Italian, Turkish (Doner) and of Japanese origin. Strong presence of discount stores, distribution centers and Women's stores. Dog Runs are of important to note, being very common for all the clusters studied, the same happens with Dim Sum restaurants.

Cluster 2 have 3% of the Neighborhoods. Its dominated by Parks, Banks, Convenience and Electronic stores. Empanadas (Hispanic), Dumpling (Chinese), Eastern European restaurants, as well Colombian, Ethiopian and Dim Sum (Chinese) restaurants.

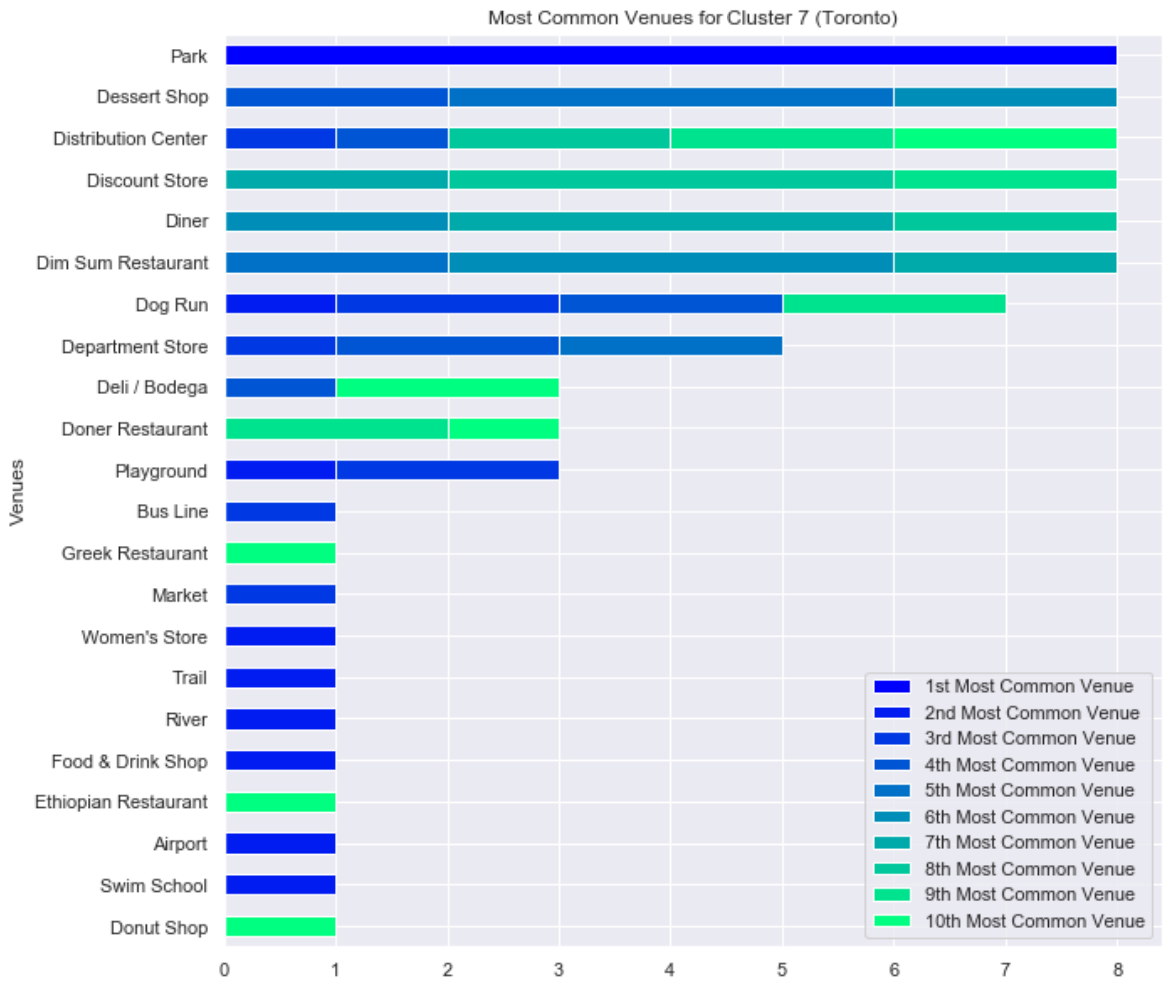
Clusters 3, 4 and 5 are very similar, each one with 1% or one data point. Cluster 3 is dominated by Cafeterias, Cluster 4 Golf Courses and Cluster 5 Playgrounds. Then followed by Women's Stores, Doner Restaurants, Dim Sum (Chinese) restaurants, Diner and Discount Stores.


Most Common Venues for Cluster 2 (Toronto)
Venues
Park
Bank
Convenience Store
Dumpling Restaurant
Empanada Restaurant
Eastern European Restaurant
Electronics Store
Colombian Restaurant
Grocery Store
Coffee Shop
Dessert Shop
College Stadium
Department Store
Women's Store
Ethiopian Restaurant
Diner
Discount Store
Drugstore
Distribution Center
Donut Shop
Dog Run
Dim Sum Restaurant
0
1
2
3
1st Most Common Venue
2nd Most Common Venue
3rd Most Common Venue
4th Most Common Venue
5th Most Common Venue
6th Most Common Venue
7th Most Common Venue
8th Most Common Venue
9th Most Common Venue
10th Most Common Venue


Most Common Venues for Cluster 6 (Toronto)
Venues
Pizza Place
Donut Shop
Discount Store
Dessert Shop
Sandwich Place
Dim Sum Restaurant
Diner
Women's Store
Distribution Center
Pharmacy
Fast Food Restaurant
Dance Studio
Coffee Shop
Doner Restaurant
Dog Run
Department Store
Chinese Restaurant
Pool
Pub
Bus Line
Bank
Intersection
Fried Chicken Joint
0
2
4
6
8
1st Most Common Venue
2nd Most Common Venue
3rd Most Common Venue
4th Most Common Venue
5th Most Common Venue
6th Most Common Venue
7th Most Common Venue
8th Most Common Venue
9th Most Common Venue
10th Most Common Venue

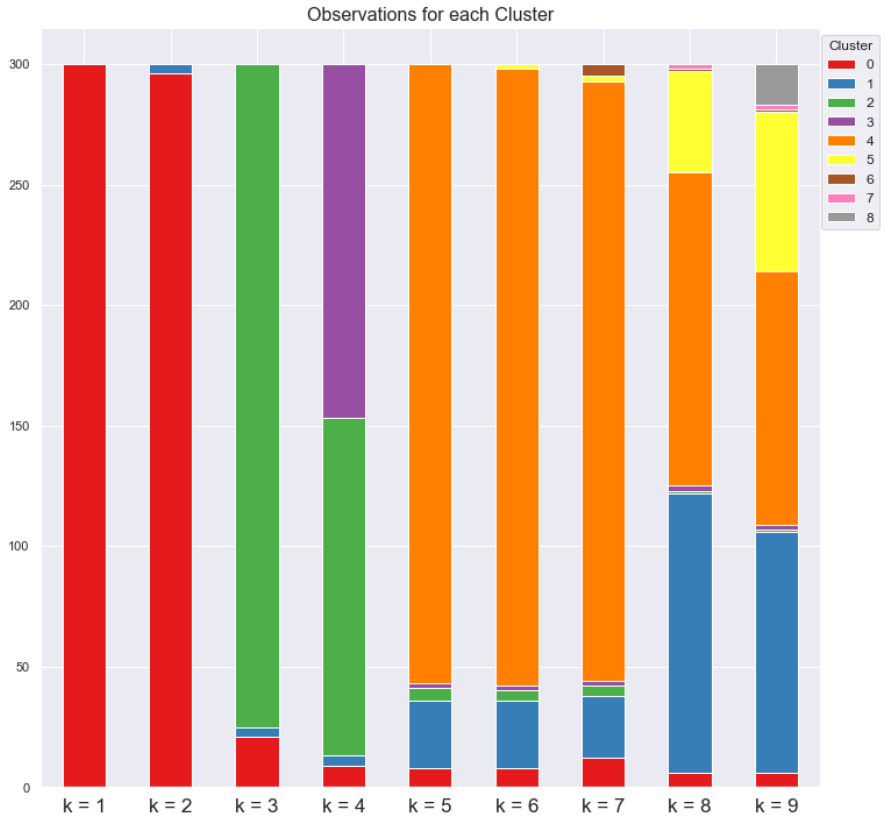
Cluster 6 represents 9% of the Neighborhoods and unlike previous clusters, the most common venue is much more prominent than in the others: pizza places. Followed by pharmacies, women's stores, dim sum (Chinese) restaurants, diner, sandwich, donuts, dessert and discount stores.

Cluster 7 represents 8% of the Neighborhoods Dominated by Parks and dessert shops. Followed by Distribution centers, discount stores, diner, dog runs, Dim Sum (Chinese) restaurants and department stores.

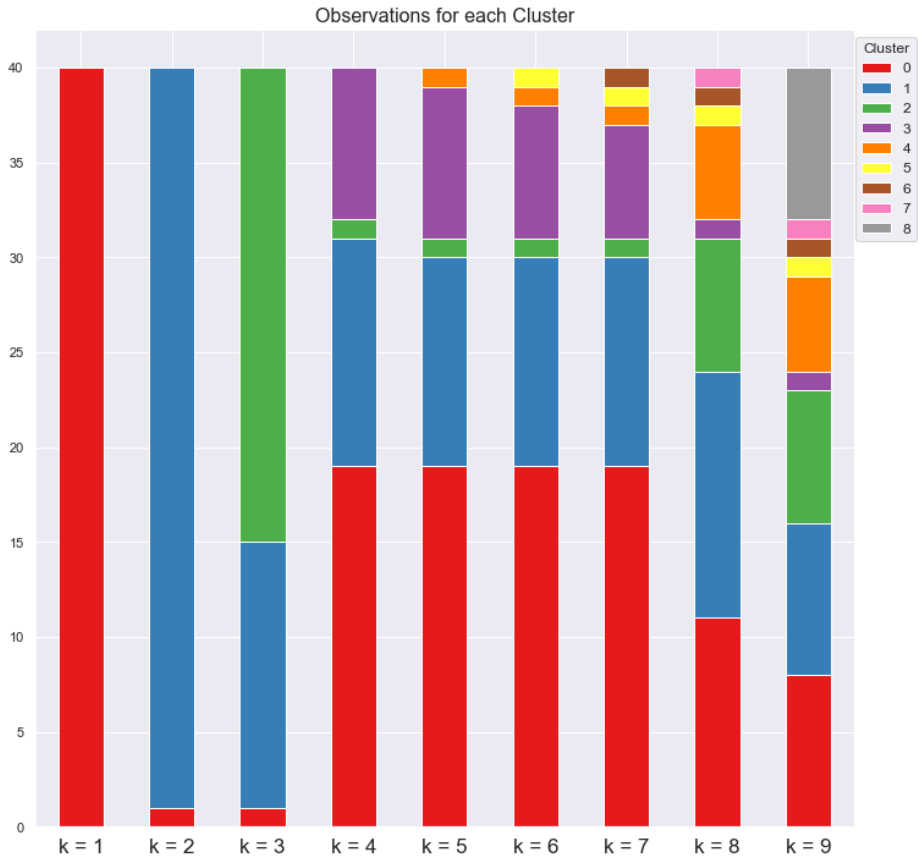


3.2 All of New York or Manhattan

When creating clusters for New York and Manhattan, I concluded that Manhattan and Toronto are more suited to perform comparatives. Manhattan serves as New York economic and administrative center, cultural identifier, and historical birthplace. Has been described as the cultural, financial, media, and entertainment capital of the world, hosting the United Nations Headquarters. Similarly Toronto is an international center of business, finance, arts, and culture. Its home to the Toronto Stock Exchange, the headquarters of Canada's five largest banks. (Source: Wikipedia)



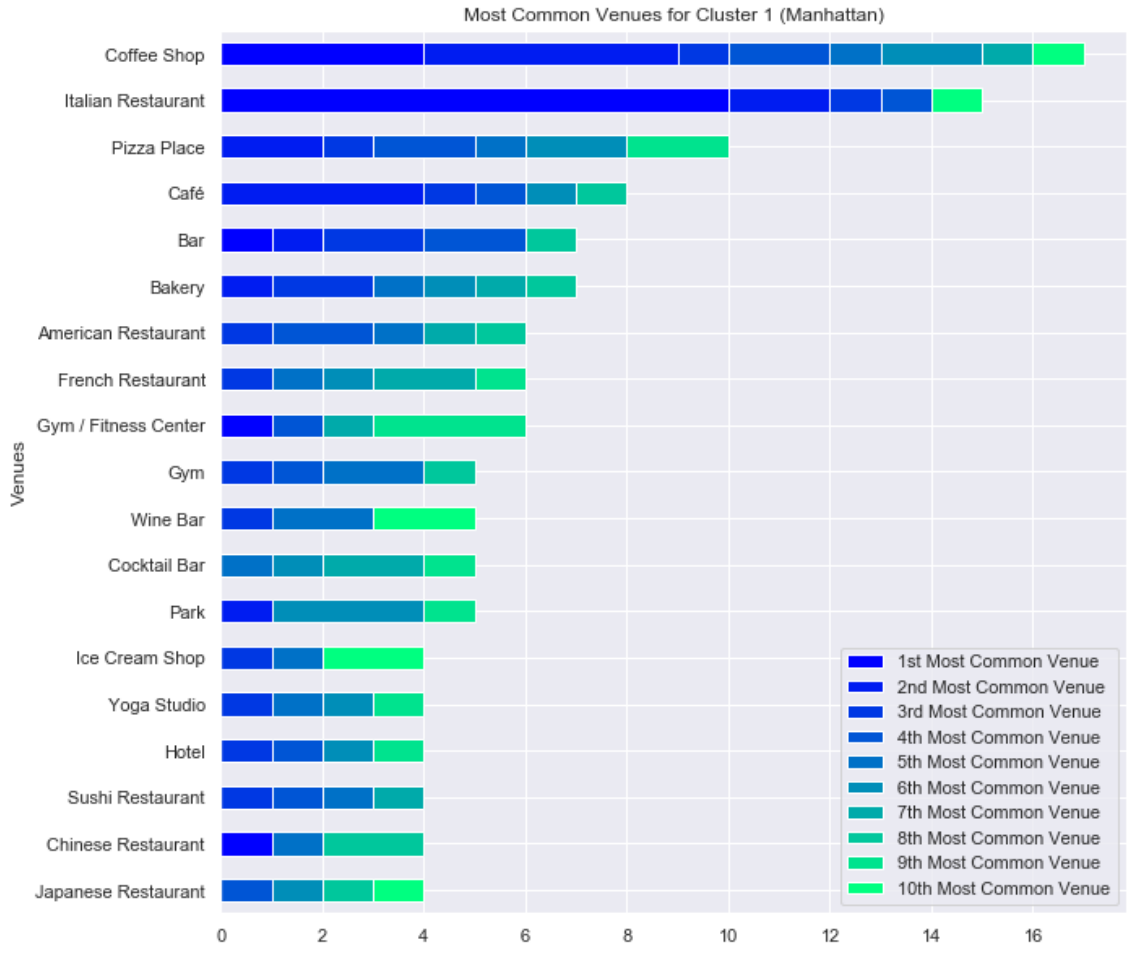
With Toronto 7 centroids where used, but only was possible to find 4 clusters, the 3 other clusters having 1 data point or neighborhood. New York have 300 neighborhoods vs 100 of Toronto, having enough data points to create clusters with no problems.



Instead Manhattan have 40 Neighborhoods and as with Toronto, being harder to find the right number of centroids. Was Possible to distinguish 2 clusters with 2 – 3 centroids, and 3 clusters with 4 – 7 centroids. With 8 and 9 centroids more clusters where created. I chose 4 centroids, finding a good balance between the amount of clusters and data points in them.

Data points for each Cluster

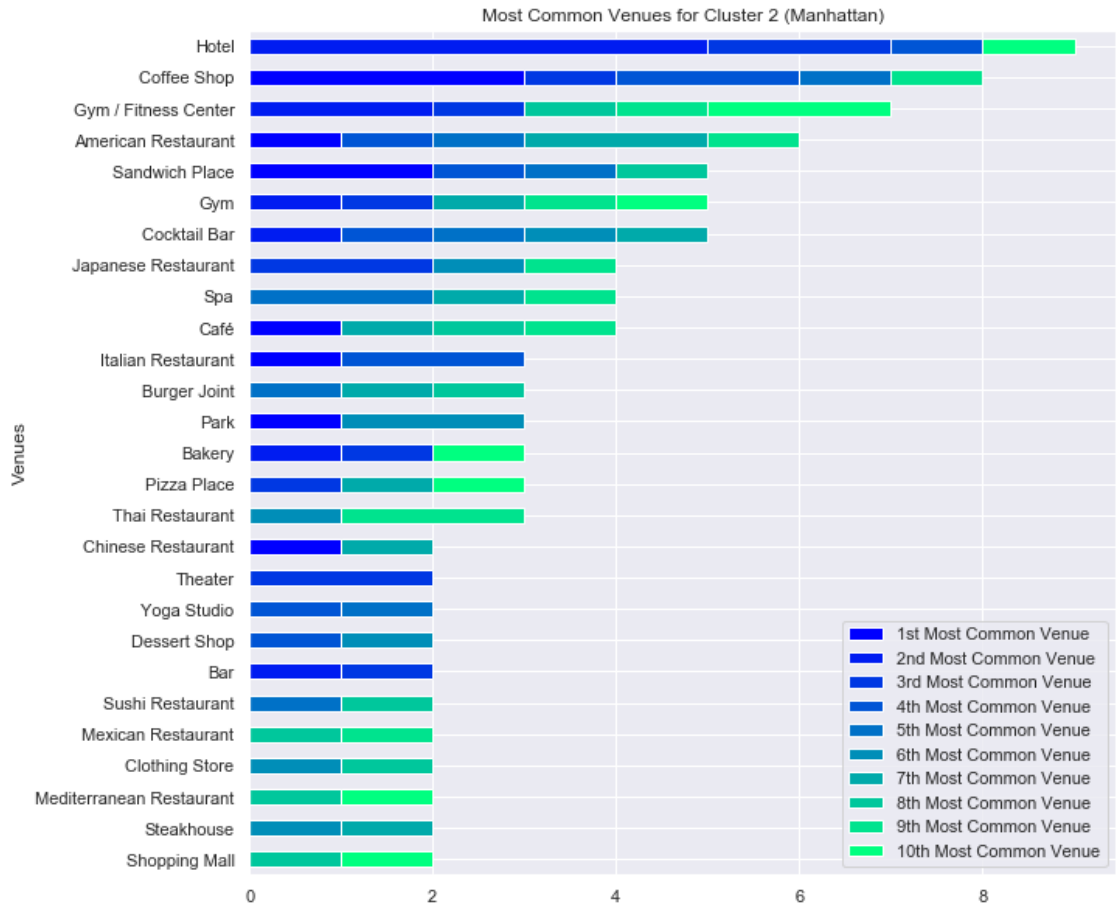
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Data Points	47.5%	30.0%	2.5%	20.0%

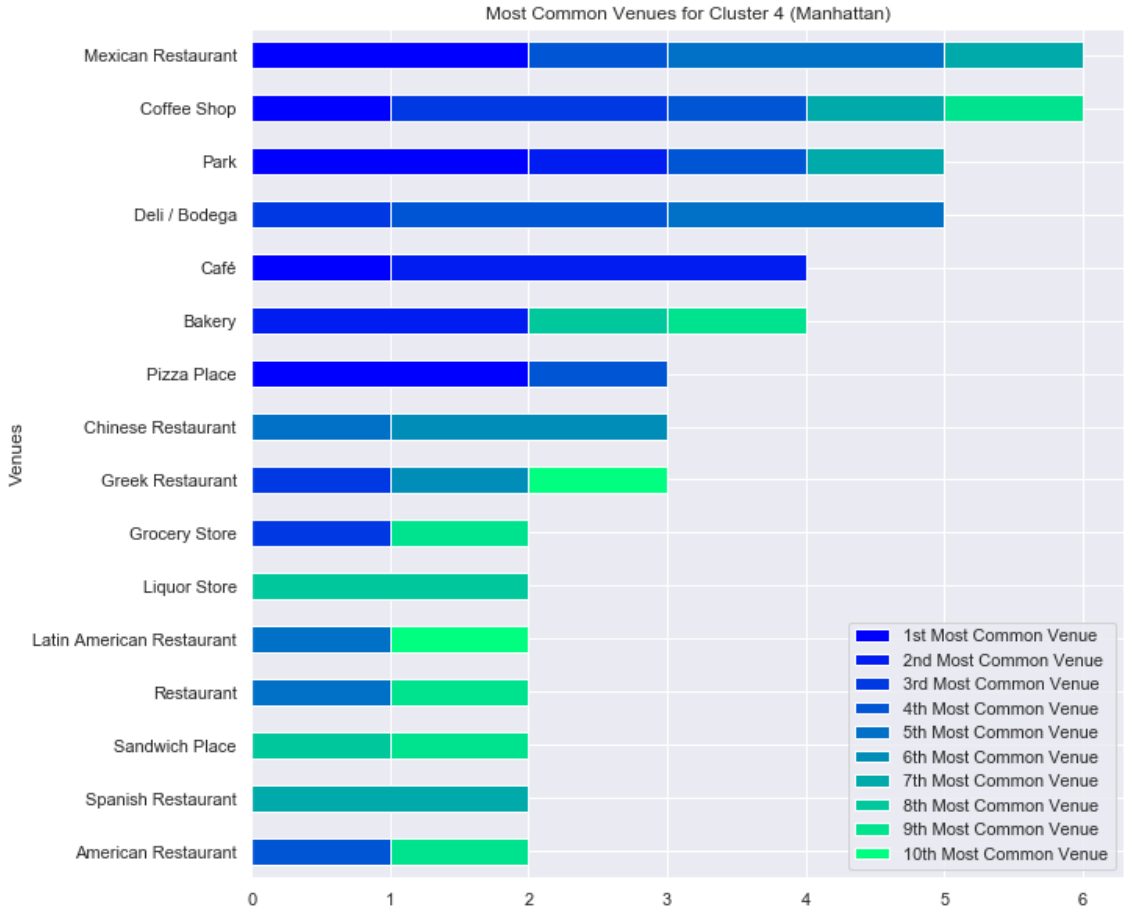


With 48% of the neighborhoods, Cluster 1 is the main cluster for Manhattan. Dominated by Coffee shops and Italian Restaurants. Followed by Pizza places, cafes, bars, bakery, gyms/yoga, American, French, Japanese and Chinese restaurants.

With 30% of the neighborhoods, Cluster 2 is the second main cluster. Its dominated by Hotels, coffees, gyms, cocktail bars, American, Japanese, Italian restaurants and spa.

Cluster 4 consist of 20% of the neighborhoods and is dominated by Mexican Restaurants, cafes, parks, bodegas, bakery, Pizza places. Followed by Chinese, Greek, Latin American, Spanish and American restaurants.


Most Common Venues for Cluster 2 (Manhattan)
Venues
Hotel
Coffee Shop
Gym / Fitness Center
American Restaurant
Sandwich Place
Gym
Cocktail Bar
Japanese Restaurant
Spa
Café
Italian Restaurant
Burger Joint
Park
Bakery
Pizza Place
Thai Restaurant
Chinese Restaurant
Theater
Yoga Studio
Dessert Shop
Bar
Sushi Restaurant
Mexican Restaurant
Clothing Store
Mediterranean Restaurant
Steakhouse
Shopping Mall
0
2
4
6
8
1st Most Common Venue
2nd Most Common Venue
3rd Most Common Venue
4th Most Common Venue
5th Most Common Venue
6th Most Common Venue
7th Most Common Venue
8th Most Common Venue
9th Most Common Venue
10th Most Common Venue


Most Common Venues for Cluster 4 (Manhattan)
Venues
Mexican Restaurant
Coffee Shop
Park
Deli / Bodega
Café
Bakery
Pizza Place
Chinese Restaurant
Greek Restaurant
Grocery Store
Liquor Store
Latin American Restaurant
Restaurant
Sandwich Place
Spanish Restaurant
American Restaurant
0
1
2
3
4
5
6
1st Most Common Venue
2nd Most Common Venue
3rd Most Common Venue
4th Most Common Venue
5th Most Common Venue
6th Most Common Venue
7th Most Common Venue
8th Most Common Venue
9th Most Common Venue
10th Most Common Venue

As for Cluster 3, it consists of only one neighborhood, with Boats or Ferry the most common venue, followed by Parks, Gas Stations, Baseball Fields and Harbor/Marina.

4. Conclusions

Chinese Restaurants, Coffee shops, Dog Runs and Women Stores are very characteristic of Toronto. For Manhattan Coffee Shops, Gyms, Pizzas and Bakery. This information denotes preferences, the cultural differences and similarities between these two great financial centers of their respective nations. Distribution center and discount stores have great success in Toronto given the amount of them. Opening stores directed to dog owners can be a big opportunity as well. Coffee is very good option in both Manhattan and Toronto.

The number of venues for all this business means that they are great opportunities to invest in them, especially in cluster with fewer number of them, with the potential to be easier to compete and dominate the market. With respect to Coffee business, it can be the best opportunity to sell in both cities, as its clear they are doing very well.