City Clusters : A comparison of two major North American Cities

IBM Course: Applied Data Science Capstone Project

Ilyes Kabbourim

April 1st, 2019

Table of Content

1. Introduction
   1.1 General introduction
   1.2 Problem
   1.3 Interested parties
2. Data
3. Methodology
4. Results
5. Discussion
6. Conclusion

1. Introduction
   1.1 General introduction

Toronto and New York City are two major cities in North America. They are both their country's economic capitals, therefore there's, famously a big diverse population in each one of these cities, people come from all around the world, either to visit or to work in these economic centers. The fact that cities like that have a big diverse population, can represent a good subject of study, in fact, there's a large amount of data to be studied and mined. One of the questions that one might ask, is how can we identify and label each area of a big city like New York in order to better target a subset of the population. Or even What makes a certain part of the city more attractive than an other, and to which kind of population is the part of the city in question attractive to.

One way to identify the caracteristics of a city, is to take a look at the venues and the amenities that are available for its population. Therefore we can for example identify business areas by the type of facilities, or even more residential parts of the city by looking at the type of buildings and shops available.

1.2 Problem

One way to start this endeavor, is to find a way to scrape and retrieve the data necessary to this project. There are many ressources over the internet that contain the data and the knowledge necessary, however, we have to find a way to scrape the data and put it in manageable format, in order to apply what we have learned over the course of the IBM data science specialization.

Once the data collected, we are going to try and find a way to cluster each neighborhood according to the type of facilities and venues available, this will be done using the sickitlearn library, and the Kmeans function. There are other type of clustering algorithms, however we are going to focus on Kmeans, because of the simplicity and the unsupervized nature.

Once the clusters identified for each city, we are going to try and compare Toronto and New York using these clusters.

1.3 Interested parties

This study might be interesting for different types of parties, one example might be to help an application recommend the best areas for a tourist visiting the city, and that by looking at the interests of each individual, and trying to match it with the right cluster that might contain the venues that are more likely to be interesting for that person.

One other way this study can be interesting, is to help new businesses target a certain area that might like to type of service they are offering, for example this might be a great way to see where a certain type of restaurant is more likely to be a successful business, by looking at where the demand is more likely to be high.

City planners might be interested as well, in an effort of improving the day to day life of the citizens.

In a socio economic study, this project might be helpful in order to see where these two cities are comparable, an what makes them the business centers that they are, and what helps the success and the flow of capitals they are subject to.

2. Data

In order to start this endeavor, we are going to need, different types of datasets, mainly data containing the type of venues for each city, the different boroughs and neighborhoods.

For New York City, we are going to use a .json file containing all the boroughs and neighborhoods, this file can be found here : https://cocl.us/new_york_dataset.

For Toronto, we are going to use a combination of python fucntions, mainly the Pandas library and the Beautiful soup library in order to scrape the table of neighborhoods, boroughs and postal codes of Toronto, that can be found on wikipedia, and more precisely this link :

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

In order to retrieve the coordinates of each neighborhood, we are going to use the geocoder function of the Geopy python library.

In order to find the venues, their type and their location (geo coordinates and adress), we are going to use the FOURSQUARE API. This information will help us identify the most popular venues in each area, and therefore try and cluster together the neighborhoods that represent the most similarities.

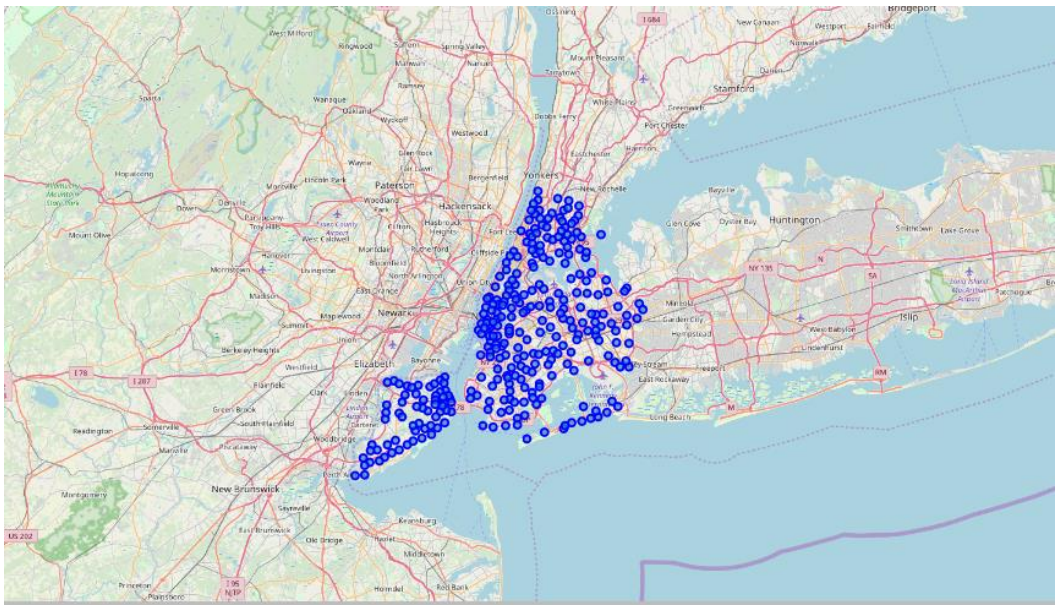
For example we can see a subset the data in the json file :

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

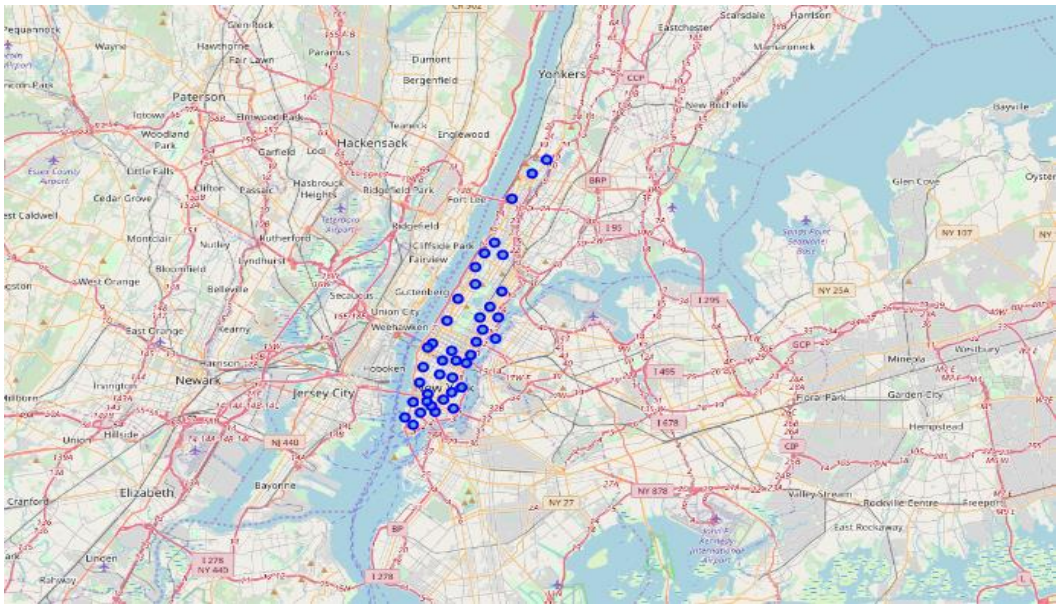
3. Methodology

First of all, we are going to visualize each neighborhood of New York City and Toronto on a map using the map function of the Folium Python Library :

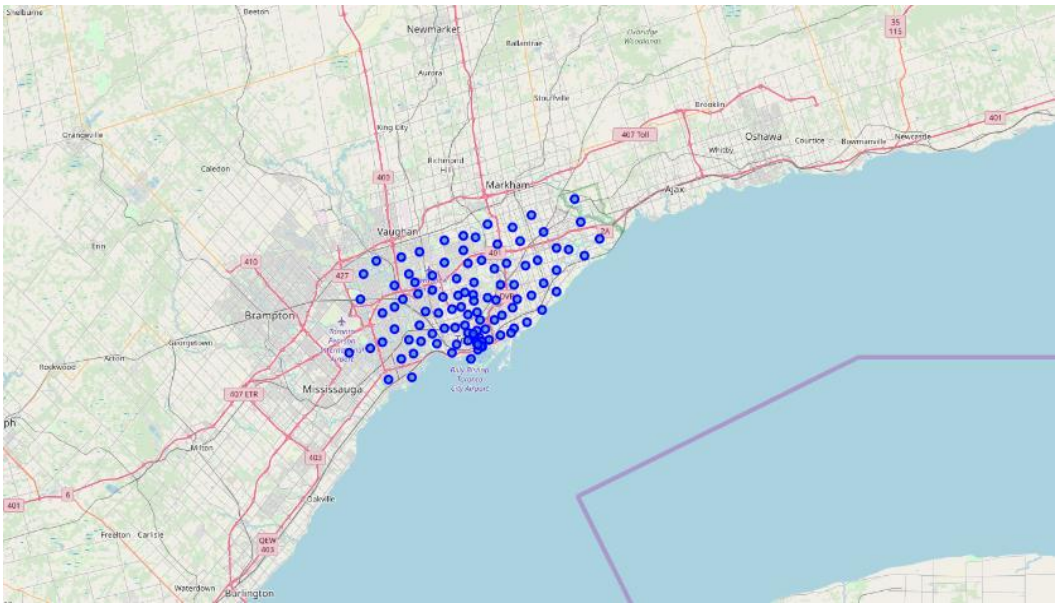
For New York City :



In order to simplify the study of New York Neighborhoods, we are going to focus on Manhattan :



For Toronto :



We are going to explore each neighborhood of Manhattan and Toronto, and by exploring, we mean, find out the venues in each neighborhood, with the radius of 500m.

And in order to do that, we are going to need to connect to The FOURSQUARE API and retrieve the venues :

Manhattan :

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Dunkin' Donuts	40.877136	-73.906666	Donut Shop
4	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop

Toronto :

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rouge, Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge, Malvern	43.806686	-79.194353	Interprovincial Group	43.805630	-79.200378	Print Shop
2	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Chris Effects Painting	43.784343	-79.163742	Construction & Landscaping
3	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	Swiss Chalet Rotisserie & Grill	43.767697	-79.189914	Pizza Place

Let’s analyze each neighborhood, by looking at the frequency of the type of venue in each one of those cities, and finding out the most common venues per neighborhood.

Manhattan :

	Neighborhood	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
0	Battery Park City	Park	Coffee Shop	Hotel	Gym	Food Truck	Italian Restaurant	Wine Shop	Sandwich Place	Clothing Store	Food Court
1	Carnegie Hill	Pizza Place	Café	Coffee Shop	Cosmetics Shop	Yoga Studio	Spa	Japanese Restaurant	Gym	French Restaurant	Bookstore
2	Central Harlem	African Restaurant	Chinese Restaurant	American Restaurant	French Restaurant	Gym / Fitness Center	Seafood Restaurant	Cosmetics Shop	Dessert Shop	Event Space	Library
3	Chelsea	Coffee Shop	Italian Restaurant	Ice Cream Shop	American Restaurant	Nightclub	Bakery	Seafood Restaurant	Hotel	Theater	Cupcake Shop
4	Chinatown	Chinese Restaurant	Dim Sum Restaurant	Bubble Tea Shop	American Restaurant	Cocktail Bar	Hotpot Restaurant	Salon / Barbershop	Bakery	Noodle House	Vietnamese Restaurant

Toronto :

	Neighborhood	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
0	Adelaide, King, Richmond	Coffee Shop	Steakhouse	Café	Thai Restaurant	Hotel	Bar	Sushi Restaurant	Bakery	American Restaurant	Burger Joint
1	Agincourt	Lounge	Skating Rink	Clothing Store	Breakfast Spot	Drugstore	Diner	Discount Store	Dog Run	Doner Restaurant	Donut Shop
2	Agincourt North, L'Amoreaux East, Milliken, St...	Park	Playground	Women's Store	Drugstore	Dim Sum Restaurant	Diner	Discount Store	Dog Run	Doner Restaurant	Donut Shop
3	Albion Gardens, Beaumond Heights, Humbergate, ...	Grocery Store	Pharmacy	Pizza Place	Coffee Shop	Fast Food Restaurant	Beer Store	Fried Chicken Joint	Sandwich Place	Women's Store	Doner Restaurant
4	Alderwood, Long Branch	Pizza Place	Athletics & Sports	Gym	Coffee Shop	Skating Rink	Pharmacy	Pub	Sandwich Place	Dog Run	Dessert Shop

We are going to use a version of these data frame in order to cluster together neighborhoods in each one of the cities, by using this script :

```
# set number of clusters
kclusters = 5

toronto_grouped_clustering = toronto_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=8, max_iter = 10000).fit(toronto_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_
```

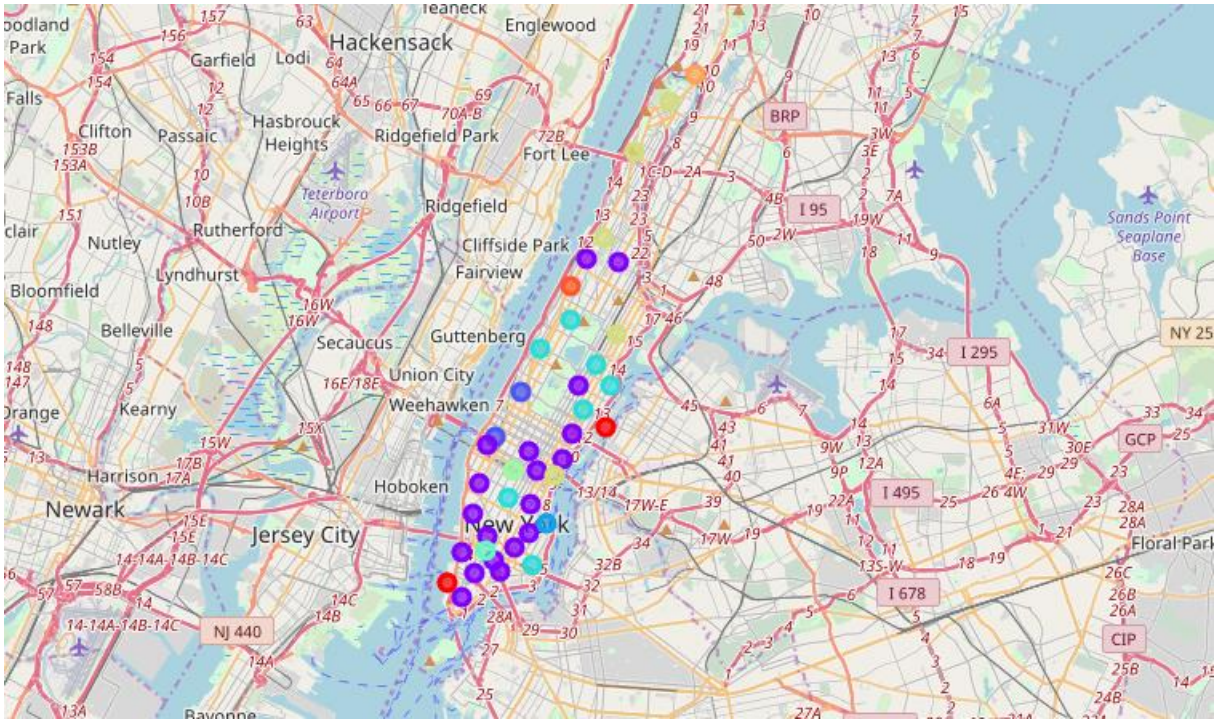
And the data frames containing the frequency if each type of venue :

	Neighborhood	Yoga Studio	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium
0	Adelaide, King, Richmond	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.03	0.0	0.0
1	Agincourt	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0
2	Agincourt North, L'Amoreaux East, Milliken, St...	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0
3	Albion Gardens, Beaumond Heights, Humbergate, ...	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0
4	Alderwood, Long Branch	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0

4. Results

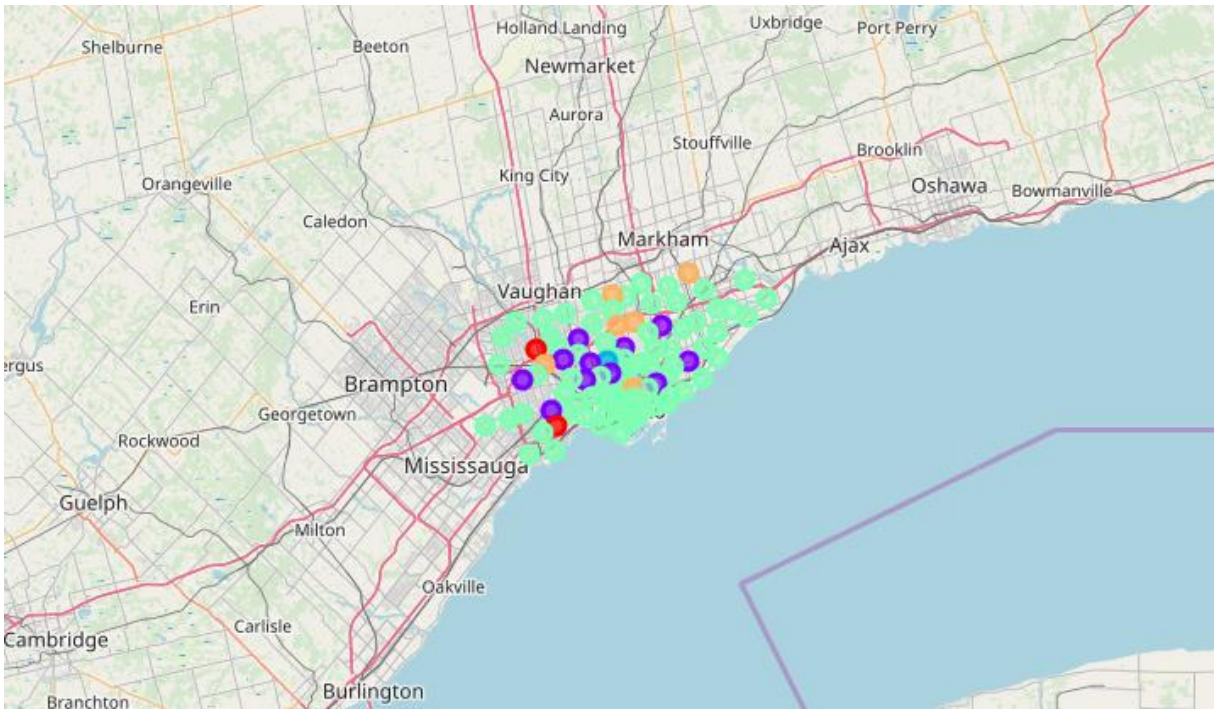
Manhattan

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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
0	Manhattan	Marble Hill	40.876551	-73.910660	8	Coffee Shop	Discount Store	Clothing Store	Big Box Store	Supplement Shop	Steakhouse	Spa
1	Manhattan	Chinatown	40.715618	-73.994279	1	Chinese Restaurant	Dim Sum Restaurant	Bubble Tea Shop	American Restaurant	Cocktail Bar	Hotpot Restaurant	Salon / Barbershop
2	Manhattan	Washington Heights	40.851903	-73.936900	7	Café	Bakery	Mobile Phone Shop	Shoe Store	Grocery Store	Mexican Restaurant	Latin American Restaurant
3	Manhattan	Inwood	40.867684	-73.921210	7	Mexican Restaurant	Café	Pizza Place	Lounge	Deli / Bodega	Restaurant	Frozen Yogurt Shop
4	Manhattan	Hamilton Heights	40.823604	-73.949688	7	Mexican Restaurant	Coffee Shop	Café	Deli / Bodega	Pizza Place	Liquor Store	Indian Restaurant



Toronto

	Postcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353	3.0	Fast Food Restaurant	Print Shop	Department Store	Dim Sum Restaurant	Diner	Discount Store
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	3.0	Bar	Construction & Landscaping	Women's Store	Discount Store	Dog Run	Doner Restaurant
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	3.0	Rental Car Location	Electronics Store	Pizza Place	Spa	Intersection	Breakfast Spot
3	M1G	Scarborough	Woburn	43.770992	-79.216917	3.0	Coffee Shop	Korean Restaurant	Insurance Office	Women's Store	Drugstore	Diner
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	3.0	Athletics & Sports	Hakka Restaurant	Fried Chicken Joint	Bakery	Caribbean Restaurant	Bank



5. Discussion

We can see in the figures above, the different types of neighborhoods according to the venues available in them. We can identify what type of clusters are given to us by looking at the most popular venues :

For example let’s look at the first cluster for each city :

Cluster 1 Manhattan :

	Neighborhood	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
11	Roosevelt Island	Park	Coffee Shop	Sandwich Place	Liquor Store	Train	Gym	Outdoors & Recreation	Bus Line	Dry Cleaner	Playground
28	Battery Park City	Park	Coffee Shop	Hotel	Gym	Food Truck	Italian Restaurant	Wine Shop	Sandwich Place	Clothing Store	Food Court

Cluster 1 Toronto :

	Borough	Cluster Labels	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	1
91	Etobicoke	0.0	Baseball Field	Eastern European Restaurant	Discount Store	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Women's Store	Dim Sum Restaurant	E
97	North York	0.0	Baseball Field	Furniture / Home Store	Falafel Restaurant	Event Space	Ethiopian Restaurant	Empanada Restaurant	Electronics Store	Eastern European Restaurant	Dim Sum Restaurant	Dumpling Restaurant	

We can see that the most popular venues in the first Manhattan Cluster are Parks and Coffee Shop.

The most popular venues in the first Toronto Cluster are Baseball Fields and ethnic restaurants.

6. Conclusion :

We can extend this study to a larger number of neighborhoods in order to identify the types of neighborhoods and the type of venues that can be more successful in each city. We can see what makes a neighborhood more attractive than an other, we can even use the Foursquare api in order to find what kind of rating is given to each venue, and try to find out what are the people that are more likely to visit the same types of venues.

We can use a different type of clustering method in order to be a little bit more precise and try to use a combination of methods to compare these cities.

This study can be used by city planners of developping cities in order to copy or follow the developping pattern of a major successful metropolis such as New York City or Toronto.