Finding Similar Neighbourhoods Between City of Toronto and New York

Nitin Rajput

October 23, 2020

1. Introduction

1.1 Background

In previous modules, we explored New York City and the city of Toronto and segmented and clustered their neighbourhoods. Both cities are very diverse and are the financial capitals of their respective countries. One interesting idea is to compare the neighbourhoods of the two cities and determine how similar or dissimilar they are. Is New York City more like Toronto or not?

1.2 Problem

Now, why are we comparing the two cities? To solve the problem defined below: Let's consider an employee that is currently working in an organization in New York City and he is promoted but is given a position in the same organization but in the city of Toronto. So, thus he has to shift from one city to another. Now we all know how difficult or tedious it is to find a similar environment that you have been living in again. If this process is done manually it would take weeks of research and understanding of other cities, which concludes to be a very hectic task. So to ease this process of shifting and finding a similar neighbourhood in the city of Toronto we are going to use a k-means machine learning algorithm to cluster the neighbourhoods and Foursquare location data to explore a particular neighbourhood to solve this problem easily.

1.3 Interest

It is very clear that a person who wants to search for similar neighbourhood would definitely be interested in such a project.

Also, businesses that want to expand their presence also can benefit from this solution provided they get additional info like demographics, customer behaviour etc.

2. Data acquisition and cleaning

2.1 Data sources

So to solve this problem we are going to use location data provided by Foursquare. We converted addresses into their equivalent latitude and longitude values. Also, we have used the Foursquare API to explore neighbourhoods in Toronto and New York City. We have used the explore function to get the most common venue categories in each neighbourhood, and then used this feature to group the neighbourhoods into clusters.

1.) New York City

New York City has a total of 5 boroughs and 306 neighbourhoods. In order to segment the neighbourhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighbourhoods that exist in each borough as well as the latitude and longitude coordinates of each neighbourhood.

Luckily, this dataset exists for free on the web, here is the link to the dataset: https://geo.nyu.edu/catalog/nyu_2451_34572

2.) Toronto

Unlike New York, the neighbourhood data is not readily available on the internet. For the Toronto neighbourhood data, a Wikipedia page exists that has all the information we need to explore and cluster the neighbourhoods in Toronto. We will be required to scrape the Wikipedia page and wrangle the data, clean it, and then read it into a pandas data frame so that it is in a structured format like the New York dataset. So the data which we are for Toronto has 10 boroughs and 217 neighbourhoods.

Link to Wikipedia page containing data regarding city of Toronto : https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Also we will be using an csv file that contains the latitude and longitude values for all the neighbourhoods in city of Toronto.

Link to the csv file : https://cocl.us/Geospatial_data

3.) Foursquare Location Data

To explore each neighbourhood we are going to use Foursquare API to explore most common venues in each neighbourhood and then used this feature to group the neighbourhoods into clusters.

2.2 Data cleaning

Firstly, I scraped New York city data which was in JSON format. I grabbed relevant information regarding neighbourhood from this JSON file which was present in features key. Then this JSON formatted data was converted into pandas data frame named “neighbourhoods”. This data frame consisted of 5 boroughs and 306 neighbourhoods.

Then for the data pertaining to the city of TORONTO, here we will be using “read_html” function of the pandas library. Using this function, we pass in the url of webpage where our data is present. Now the “read_html” function goes through the webpage and if tables are present on the webpage it will return it in the form of list containing all the tables that were present. Since we only have one table, so we use list indexing to access it. The tables are returned or stored as data frame in the list. Now accessing only, the data frame present at the zeroth index of the list return, we get our desired data. We will ignore cells with a borough that is 'Not assigned’. But this data frame consists of only 3 columns that are: Postal Code, Borough, Neighbourhoods. So in order to explore each and every neighbourhood we need latitude and longitude values. So then I used ‘geo_data.csv’ which contains all the latitude and longitude values to append to the neighbourhoods data frame. So now finally we got our dataset that contains neighbourhoods and their relevant latitude and longitude values. I named it ‘‘toronto_data’’. So our final ‘’toronto_data’’ data frame contains 10 boroughs.

Finally, then I appended both the cities data frame together resulting in “result” data frame which contains 5 columns named “Borough”,” Neighbourhood”,” Latitude”,” Longitude”,” Postal Code” and 409 rows.

Now after cleaning our dataset I started to use Foursquare API to explore the neighbourhoods and segment them.

I created a function “getNearbyVenues” that will take a single record from our “result” data frame at a time and then get the top 100 venues near that neighbourhood, then convert the JSON formatted data that is received to a pandas data frame and all this data is stored into “venues” data frame. During this process we have used Foursquare API to send request and to get data from Foursquare Location Database.

The “venues” data frame consists of 12225 rows and 7 columns.
Then I found out how many unique venue categories can be curated from all the returned ven ues. There are 459 unique categories.
Looking good. So now we have all the venues in our respective neighbourhoods.

This concludes the data gathering phase - we're now ready to use this data for analysis to produce the report on similar neighbourhoods.

2.3 Feature selection

After data cleaning the data frame now we are going to create new features so as we can input this data into the machine learning algorithm.

Then we are going to analyse each and every neighbourhood, so that we can find top ten venues pertaining to each neighbourhood. Here we are going to use one hot encoding technique.

Then we grouped rows by neighbourhood and took the mean of the frequency of occurrence of each category. So, finally we have our resulting dataset which will be used as an input to the machine learning algorithm, this data frame consists of 395 rows and 459 columns. Take a look at the data frame:

	Neighborhood	New American Restaurant	Newsstand	Nightclub	Nightlife Spot	Non-Profit	Noodle House	North Indian Restaurant	Office	Opera House	Optical Shop	Organic Grocery	Other Great Outdoors	O Nigh
0	Agincourt	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
1	Alderwood, Long Branch	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
2	Allerton	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
3	Annadale	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
4	Arden Heights	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
5	Arlington	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
6	Arrochar	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
7	Arverne	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
8	Astoria	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
9	Astoria Heights	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
10	Auburndale	0.000000	0.00	0.000000	0.0	0.000000	0.045455	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
11	Bath Beach	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.020408	0.000000	0.000000	
12	Bathurst Manor, Wilson Heights, Downsview North	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
13	Battery Park City	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
14	Bay Ridge	0.012658	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.012658	0.000000	0.000000	
15	Bay Terrace	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
16	Baychester	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
17	Bayside	0.000000	0.00	0.000000	0.0	0.000000	0.013333	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
18	Bayswater	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
19	Bayview Village	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
20	Bedford Park	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
21	Bedford Park, Lawrence Manor East	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	

3. Methodology

In this project we are directing our efforts on finding the neighbourhoods that are similar to each other. Which will solve our problem of finding similar neighbourhoods to move to in different cities.

Here we are comparing two cities: New York and Toronto

First we collected the data regarding every neighbourhood in the city plus also collected their respective latitude and longitude values, so as to explore those neighbourhoods and to get top venues near them, on basis of which the entire clustering process will be performed.

Second step in our analysis will be to analyse each neighbourhood, then to group rows by neighbourhood and to find top 10 venues pertaining to each neighbourhood by taking the mean of the frequency of occurrence of each category.

In the third and final section, we are going to use k-means clustering algorithm to segment and group neighbourhoods. Reasons why we are using k-means clustering algorithm are:

- k-means is one of the simplest algorithm which uses unsupervised learning method to solve known clustering issues.
- It works really well with large datasets.
- Guarantees convergence.
- Can warm-start the positions of centroids.
- Easily adapts to new examples.
- Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

Then we are going to visualize neighbourhoods on the basis of clusters they are assigned to. Then finally we are going to share some insights or observations that were made through clustering to expedite our learning.

3.1 K-means Clustering

K - means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters; the more homogeneous (similar) the data points are within the same cluster.

The way k - means algorithm works is as follows:

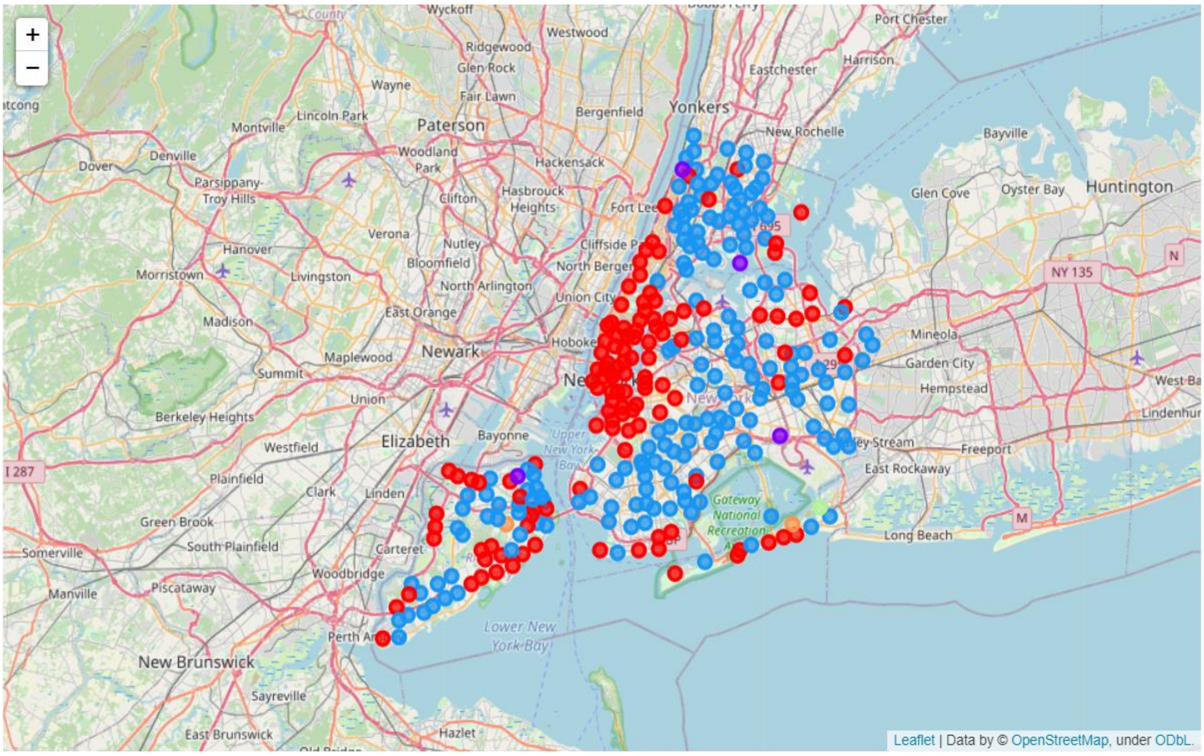
1. Specify number of clusters K.
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e. assignment of data points to clusters isn't changing.
4. Compute the sum of the squared distance between data points and all centroids.
5. Assign each data point to the closest cluster (centroid).
6. Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

So thus I used K-means algorithm with k value = 6 and got the cluster labels for all the neighbourhoods. Then I created a new data frame which consisted of all the info pertaining to a neighbourhood with their cluster labels and top ten venues in the respective neighbourhood. The final data frame looks like this:

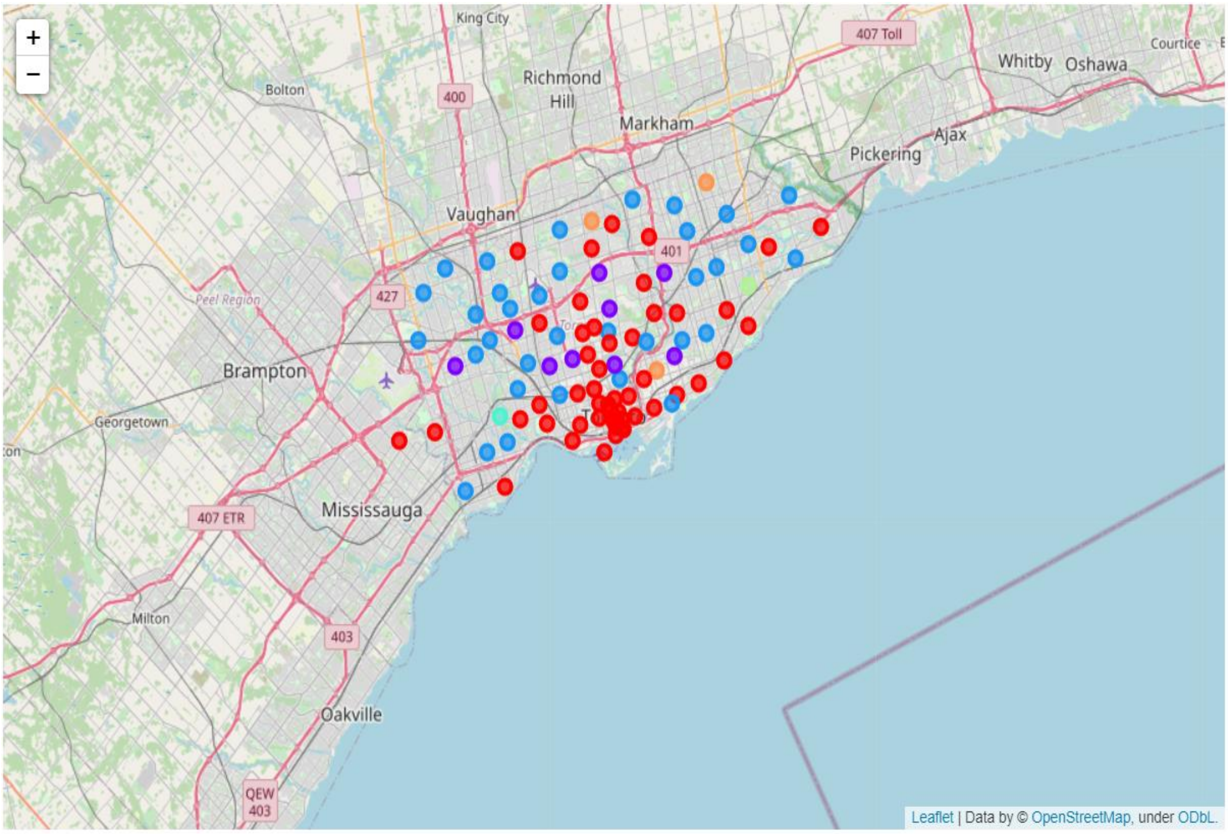
	Borough	Neighborhood	Latitude	Longitude	Postal Code	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th M Comr Ve
0	Bronx	Wakefield	40.894705	-73.847201	NaN	2	Pharmacy	Gas Station	Pizza Place	Laundromat	Dessert Shop	Sandv P
1	Bronx	Co-op City	40.874294	-73.829939	NaN	2	Baseball Field	Restaurant	Deli / Bodega	Pizza Place	Pharmacy	Grocery S
2	Bronx	Eastchester	40.887556	-73.827806	NaN	2	Bus Station	Caribbean Restaurant	Diner	Deli / Bodega	Bus Stop	Bowling A
3	Bronx	Fieldston	40.895437	-73.905643	NaN	2	River	Medical Supply Store	Bus Station	Plaza	Nail Salon	Veterina
4	Bronx	Riverdale	40.890834	-73.912585	NaN	2	Bus Station	Park	Medical Supply Store	Home Service	Bank	Food Ti
5	Bronx	Kingsbridge	40.881687	-73.902818	NaN	2	Pizza Place	Bakery	Bar	Sandwich Place	Latin American Restaurant	Liquor S
6	Manhattan	Marble Hill	40.876551	-73.910660	NaN	0	Coffee Shop	Sandwich Place	Gym	Discount Store	Supplement Shop	Donut S
7	Bronx	Woodlawn	40.898273	-73.867315	NaN	2	Pizza Place	Deli / Bodega	Food & Drink Shop	Pub	Playground	Ita Restau
8	Bronx	Norwood	40.877224	-73.879391	NaN	2	Pizza Place	Bank	Park	Pharmacy	Burger Joint	Grocery S
9	Bronx	Williamsbridge	40.881039	-73.857446	NaN	0	Nightclub	Caribbean Restaurant	Dance Studio	Bar	Soup Place	Nail S
10	Bronx	Baychester	40.866858	-73.835798	NaN	2	Donut Shop	Pet Store	Mexican Restaurant	Cosmetics Shop	Sandwich Place	Disc S
11	Bronx	Pelham Parkway	40.857413	-73.854756	NaN	2	Italian Restaurant	Pizza Place	Bus Station	Food	Bank	Sandv P

3.2 Visualizing resulting clusters

3.2.1 Visualizing clusters in New York City



3.2.2 Visualizing clusters in Toronto city



So this concludes our analysis and finally we have clustered neighbourhoods in the city of Toronto and New York. We have used most common venues data within 500 meters of respective neighbourhoods to cluster them. We have clustered them into 6 clusters. Also during our analysis, we found that we did not get location data regarding few neighbourhoods, so we had to drop those neighbourhoods in our final analysis.

4. Results

Finally, we have obtained our clustered dataset 'final_df' which contains each neighbourhood with their corresponding most common venue categories on basis of which our neighbourhoods are clustered.

So this data frame provides answer to our query or question that was to which neighbourhood or location should a person move provided the neighbourhood he/she is moving to has similar amenities or locations as of earlier neighbourhood where he/she is residing.

Thus on the basis of this data frame a recommendation can be made to a person who wants to shift from one neighbourhood to another.

5. Discussion

Now, we can examine each cluster and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, we can then assign a name to each cluster.

Also creating visual plots for each cluster.

Custer 1

Most of the neighbourhoods are grouped in cluster 1 and services or venues that are most common in cluster 1 are basically pharmacy, bus station, restaurants, eating places, gyms etc. So it can have concluded that neighbourhoods in cluster 1 are very much suitable for a person to shift as these neighbourhoods contains all the basic amenities or services that a person requires.

	Borough	Neighborhood	Latitude	Longitude	Postal Code	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th M Comm Ve
0	Manhattan	Marble Hill	40.876551	-73.910660	NaN	0	Coffee Shop	Sandwich Place	Gym	Discount Store	Supplement Shop	Donut S
1	Bronx	Williamsbridge	40.881039	-73.857446	NaN	0	Nightclub	Caribbean Restaurant	Dance Studio	Bar	Soup Place	Nail Sa
2	Bronx	City Island	40.847247	-73.786488	NaN	0	Thrift / Vintage Store	Seafood Restaurant	Deli / Bodega	Smoke Shop	History Museum	Arts & Cr S
3	Bronx	Throgs Neck	40.815109	-73.816350	NaN	0	American Restaurant	Sports Bar	Asian Restaurant	Pizza Place	Coffee Shop	
4	Bronx	Belmont	40.857277	-73.888452	NaN	0	Italian Restaurant	Deli / Bodega	Pizza Place	Bakery	Grocery Store	Sandv Pl
5	Bronx	Edgewater Park	40.821986	-73.813885	NaN	0	Italian Restaurant	Pizza Place	Liquor Store	Japanese Restaurant	Ice Cream Shop	Donut S
6	Brooklyn	Bay Ridge	40.625801	-74.030621	NaN	0	Italian Restaurant	Pizza Place	Spa	Bar	Grocery Store	Ameri Restau
7	Brooklyn	Greenpoint	40.730201	-73.954241	NaN	0	Bar	Pizza Place	Coffee Shop	Cocktail Bar	French Restaurant	Yoga Stu
8	Brooklyn	Brighton Beach	40.576825	-73.965094	NaN	0	Restaurant	Russian Restaurant	Eastern European Restaurant	Beach	Bank	S Restau
9	Brooklyn	Sheepshead Bay	40.586890	-73.943186	NaN	0	Turkish Restaurant	Dessert Shop	Sandwich Place	Outlet Store	Buffet	D
10	Brooklyn	Windsor Terrace	40.656946	-73.980073	NaN	0	Diner	Grocery Store	Park	Café	Deli / Bodega	Pl
11	Brooklyn	Prospect Heights	40.676822	-73.964859	NaN	0	Bar	Mexican Restaurant	Thai Restaurant	Wine Shop	Cocktail Bar	Bal
12	Brooklyn	Williamsburg	40.707144	-73.958115	NaN	0	Coffee Shop	Bar	Taco Place	Bagel Shop	Clothing Store	Ita Restau
13	Brooklyn	Bushwick	40.698116	-73.925258	NaN	0	Bar	Mexican Restaurant	Coffee Shop	Deli / Bodega	Bakery	Th Vintage S
14	Brooklyn	Bedford Stuyvesant	40.687232	-73.941785	NaN	0	Deli / Bodega	Coffee Shop	Bar	Café	Pizza Place	Tiki
15	Brooklyn	Brooklyn Heights	40.695864	-73.993782	NaN	0	Park	Italian Restaurant	Deli / Bodega	Yoga Studio	Bakery	Cosme S

Cluster 2

Most common venue in neighbourhoods that are grouped in cluster 2 include Parks, Train Satiation and Trains as common venues.

	Borough	Neighborhood	Latitude	Longitude	Postal Code	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	Bronx	Clason Point	40.806551	-73.854144	NaN	1	Park	Boat or Ferry	Bus Stop	Recording Studio	Pool	Grocery Store	South American Restaurant	Co
1	Bronx	Spuyten Duyvil	40.881395	-73.917190	NaN	1	Park	Pharmacy	Tennis Court	Scenic Lookout	Bank	Thai Restaurant	Tennis Stadium	I
2	Queens	South Ozone Park	40.668550	-73.809865	NaN	1	Park	Deli / Bodega	Fast Food Restaurant	Bar	Donut Shop	Hotel	Sandwich Place	
3	Staten Island	Randall Manor	40.635630	-74.098051	NaN	1	Pizza Place	Park	Playground	Bus Stop	Nail Salon	Veterinarian	Train	Tr
4	North York	Parkwoods	43.753259	-79.329656	M3A	1	BBQ Joint	Food & Drink Shop	Park	Pool	Wine Bar	Wine Shop	Trail	
5	East York	Woodbine Heights	43.695344	-79.318389	M4C	1	Skating Rink	Curling Ice	Park	Beer Store	Venezuelan Restaurant	Track	Trail	
6	York	Humewood-Cedarvale	43.693781	-79.428191	M6C	1	Trail	Park	Hockey Arena	Field	Nail Salon	Vegetarian / Vegan Restaurant	Track	
7	York	Caledonia-Fairbanks	43.689026	-79.453512	M6E	1	Park	Women's Store	Pool	Nail Salon	Veterinarian	Train	Train Station	
8	North York	North Park, Maple Leaf Park, Upwood Park	43.713756	-79.490074	M6L	1	Construction & Landscaping	Bakery	Park	Nail Salon	Venezuelan Restaurant	Trail	Train	Tr
9	Central Toronto	Lawrence Park	43.728020	-79.388790	M4N	1	Swim School	Park	Bus Line	Nail Salon	Veterinarian	Trail	Train	Tr
10	North York	York Mills West	43.752758	-79.400049	M2P	1	Convenience Store	Electronics Store	Park	Track	Trail	Train	Train Station	
11	Etobicoke	Kingsview Village, St. Phillips, Martin Grove ...	43.688905	-79.554724	M9R	1	Sandwich Place	Pizza Place	Park	Bus Line	Veterinarian	Train	Train Station	
12	Central Toronto	Moore Park, Summerhill East	43.689574	-79.383160	M4T	1	Tennis Court	Restaurant	Park	Video Game Store	Train	Train Station	Turkish Restaurant	

Cluster 3

Most common venues belonging to neighbourhoods that are clustered together in cluster 3 are Bakery, Bar and Restaurants.

	Borough	Neighborhood	Latitude	Longitude	Postal Code	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	Bronx	Wakefield	40.894705	-73.847201	NaN	2	Pharmacy	Gas Station	Pizza Place	Laundromat	Dessert Shop	Sandwich Place
1	Bronx	Co-op City	40.874294	-73.829939	NaN	2	Baseball Field	Restaurant	Deli / Bodega	Pizza Place	Pharmacy	Grocery Store
2	Bronx	Eastchester	40.887556	-73.827806	NaN	2	Bus Station	Caribbean Restaurant	Diner	Deli / Bodega	Bus Stop	Bowling Alley
3	Bronx	Fieldston	40.895437	-73.905643	NaN	2	River	Medical Supply Store	Bus Station	Plaza	Nail Salon	Veterinarian
4	Bronx	Riverdale	40.890834	-73.912585	NaN	2	Bus Station	Park	Medical Supply Store	Home Service	Bank	Food Truck
5	Bronx	Kingsbridge	40.881687	-73.902818	NaN	2	Pizza Place	Bakery	Bar	Sandwich Place	Latin American Restaurant	Liquor Store
6	Bronx	Woodlawn	40.898273	-73.867315	NaN	2	Pizza Place	Deli / Bodega	Food & Drink Shop	Pub	Playground	Italian Restaurant
7	Bronx	Norwood	40.877224	-73.879391	NaN	2	Pizza Place	Bank	Park	Pharmacy	Burger Joint	Grocery Store
8	Bronx	Baychester	40.866858	-73.835798	NaN	2	Donut Shop	Pet Store	Mexican Restaurant	Cosmetics Shop	Sandwich Place	Discount Store
9	Bronx	Pelham Parkway	40.857413	-73.854756	NaN	2	Italian Restaurant	Pizza Place	Bus Station	Food	Bank	Sandwich Place
10	Bronx	Bedford Park	40.870185	-73.885512	NaN	2	Diner	Mexican Restaurant	Pizza Place	Chinese Restaurant	Deli / Bodega	Sandwich Place
11	Bronx	University Heights	40.855727	-73.910416	NaN	2	Pizza Place	African Restaurant	Convenience Store	Laundromat	Donut Shop	Pharmacy
12	Bronx	Morris Heights	40.847898	-73.919672	NaN	2	Pharmacy	Bank	Spanish Restaurant	Plaza	Recreation Center	Deli / Bodega
13	Bronx	Fordham	40.860997	-73.896427	NaN	2	Mobile Phone Shop	Shoe Store	Bank	Fast Food Restaurant	Spanish Restaurant	Donut Shop

Cluster 4
It can be seen that river is the only discriminating venue category for cluster 4 neighbourhood. That is the reason why cluster 4 only has one neighbourhood.

	Borough	Neighborhood	Latitude	Longitude	Postal Code	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
0	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944	M8X	3	River	Nail Salon	Veterinarian	Trail	Train	Train Station	Turkish Restaurant	Udon Restaurant

Cluster 5
Cluster 5 includes neighbourhoods that have venue categories as follows:

- Train
- Train Station
- Playground

	Borough	Neighborhood	Latitude	Longitude	Postal Code	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 Cor \
0	Queens	Bayswater	40.611322	-73.765968	NaN	4	Playground	Nail Salon	Veterinarian	Trail	Train	Train Station	Turkish Restaurant	Resta
1	Scarborough	Scarborough Village	43.744734	-79.239476	M1J	4	Spa	Playground	Video Game Store	Train	Train Station	Turkish Restaurant	Udon Restaurant	Usec Deal

Cluster 6
Most common venues in neighbourhoods in Cluster 6 are Parks, Nail Salon and Train Station etc.

	Borough	Neighborhood	Latitude	Longitude	Postal Code	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	8t Cc
0	Queens	Somerville	40.597711	-73.796648	NaN	5	Park	Nail Salon	Veterinarian	Trail	Train	Train Station	Turkish Restaurant	Res
1	Staten Island	Todt Hill	40.597069	-74.111329	NaN	5	Park	Nail Salon	Veterinarian	Trail	Train	Train Station	Turkish Restaurant	Res
2	East York	East Toronto, Broadview North (Old East York)	43.685347	-79.338106	M4J	5	Park	Convenience Store	Veterinarian	Trail	Train	Train Station	Turkish Restaurant	Res
3	North York	Willowdale, Newtonbrook	43.789053	-79.408493	M2M	5	Park	Nail Salon	Veterinarian	Trail	Train	Train Station	Turkish Restaurant	Res
4	Scarborough	Milliken, Agincourt North, Steeles East, L'Amo...	43.815252	-79.284577	M1V	5	Park	Playground	Nail Salon	Veterinarian	Train	Train Station	Turkish Restaurant	Res

6. Conclusion

Purpose of this project was to identify similar neighbourhoods in the city of New York and Toronto in order to aid people in narrowing down the search for optimal neighbourhood. We started by exploring each neighbourhood in our dataset and finding locations or venues around it in range of 500 meters. Then we sorted our neighbourhoods on the basis of most common venues that are present. Clustering of those neighbourhoods was then performed in order to create major zones of interest (containing greatest number of potential locations) and then this data is used to recommend most similar neighbourhood according to the required condition.

Final decision on optimal neighbourhood will be made by end user based on specific characteristics of neighbourhoods and locations in every recommended neighbourhood, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighbourhood etc.