New York City and the city of Toronto - Similarities and Dissimilarities

Introduction

Deciding where to relocate, between New York and Toronto, is a tough choice. You need to consider several factors that may affect your living in any of these cities. Both are English speaking cities and the standard of living is very similar. Both cities have been developed to become a centre of attention for residential, job employment, tourism, education, shopping and sports activity.

According to https://en.wikipedia.org/wiki/New_York_City:

"The City of New York, often called New York City (NYC) or simply New York (NY), is the most populous city in the United States. With an estimated 2017 population of 8,622,698 distributed over a land area of about 302.6 square miles (784 km2), New York City is also the most densely populated major city in the United States."

"Toronto is the capital city of the province of Ontario and the largest city in Canada by population, with 2,731,571 residents in 2016. Current to 2016, the Toronto census metropolitan area (CMA), of which the majority is within the Greater Toronto Area (GTA), held a population of 5,928,040, making it Canada's most populous CMA."

According to https://versus.com/en/new-york-vs-toronto, New York City seems to be leading the race:

Has a seaside beach
3.8% lower unemployment rate (6.1% vs 9.9%)

1.4 year younger population (35.5 years vs 36.9 years)

4.2% lower VAT (8.8% vs 13%)

0.53$ lower average price of a litre of milk (1.02$ vs 1.55$)

0.11$ lower cost per litre of fuel (0.68$ vs 0.79$)

Is that so?

In this project, we are going to investigate the similarities and dissimilarities between these two cities, using data science tools, such as classification using Foursquare data and machine learning segmentation and clustering.

Data Sources

Based on the most common places captured from Foursquare, we are going to segment areas of Toronto and New York. Two randomly neighbourhoods will be picked and analysed the top 10 most common venues in each of those two neighbourhoods based on the number of visits by people in each of those places. K-mean clustering unsupervised machine learning algorithm will cluster the venues based on the place category such as restaurants, park, coffee shop, gym etc.

The following urls will be used in order to collect the data necessary for analysis:

- Toronto Postal Codes from Wikipedia:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'

- New York dataset that contains the 5 boroughs and the neighbourhoods that exist in each borough as well as the latitude and longitude coordinates or each neghbor5hood.

- Foursquare API: (https://foursquare.com/)
Performing location search, location sharing and details about a business. Photos, tips and reviews.

- Folium- Python visualization library used to visualize the neighbourhoods cluster distribution

Various libraries, such as pandas (for data analyses), json (to handle json files), geopy (to retrieve location data), sklearn (machine learning library), etc.

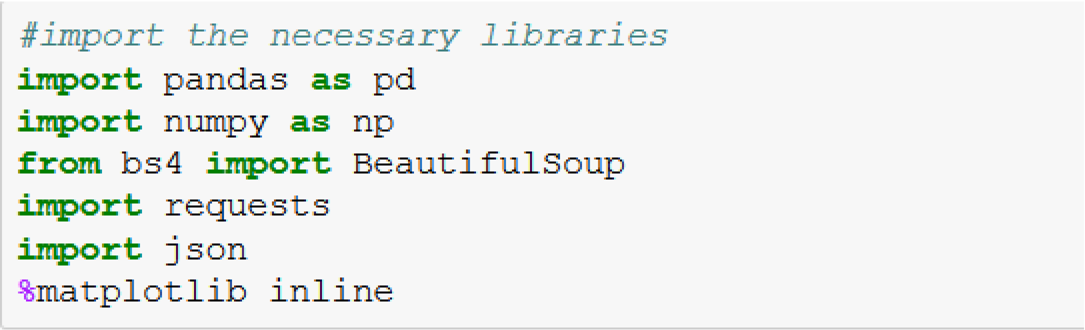
Methodology

In order to achieve our goal, we have followed the following stages:

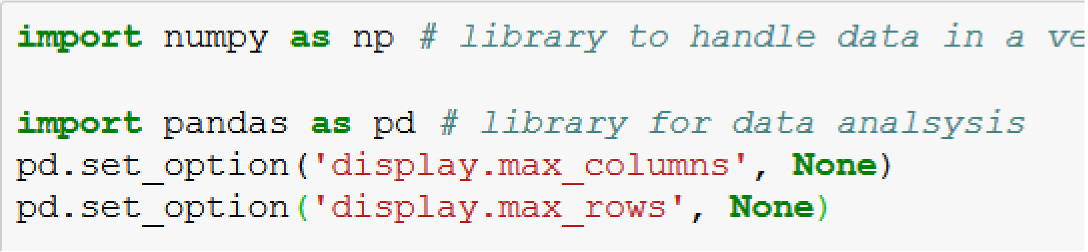
1. Collect Toronto and New York Data
2. Data preparation and preprocessing
3. Modeling, Visualization and Clustering

1. Collect Toronto Data

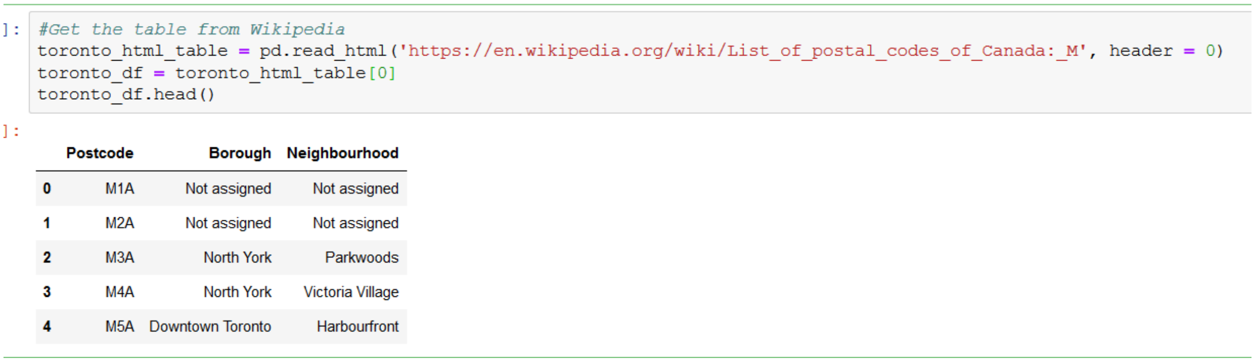
After importing the necessary libraries, we download the data from Wikipedia ("https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"):



```
#import the necessary libraries
import pandas as pd
import numpy as np
from bs4 import BeautifulSoup
import requests
import json
%matplotlib inline
```



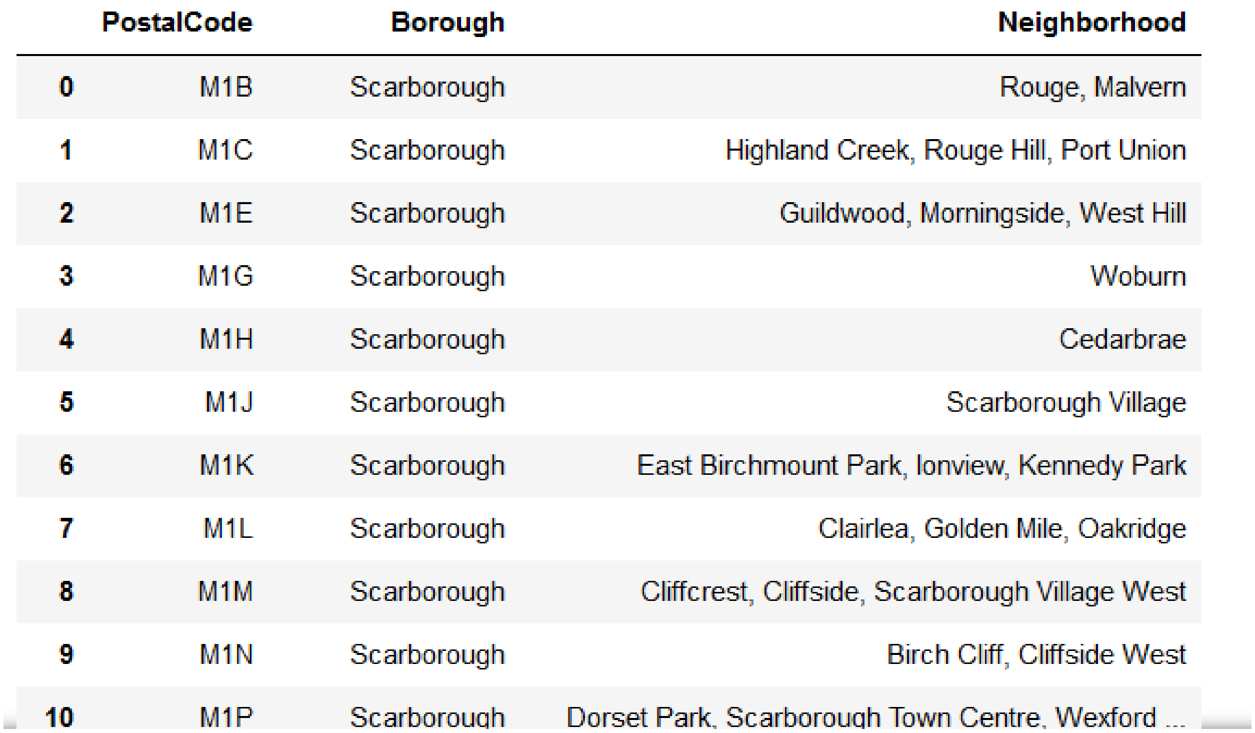
```
import numpy as np # library to handle data in a ve

import pandas as pd # library for data analsysis
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
```



```
]: #Get the table from Wikipedia
   toronto_html_table = pd.read_html('https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M', header = 0)
   toronto_df = toronto_html_table[0]
   toronto_df.head()
]:
```

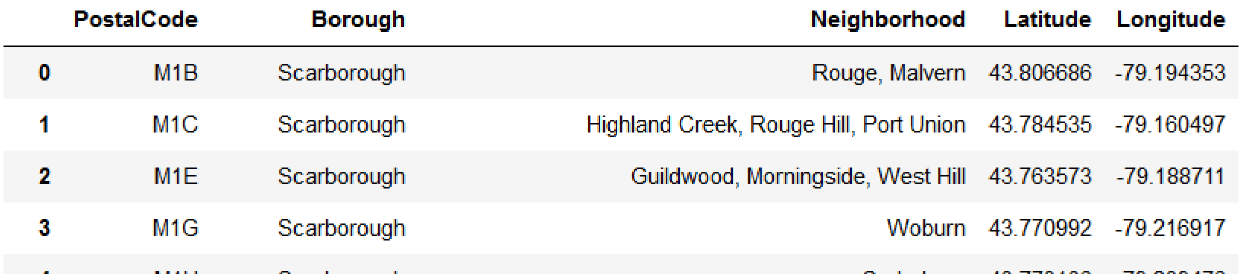
	Postcode	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront

Dataset had to be cleaned, removed null values, etc.



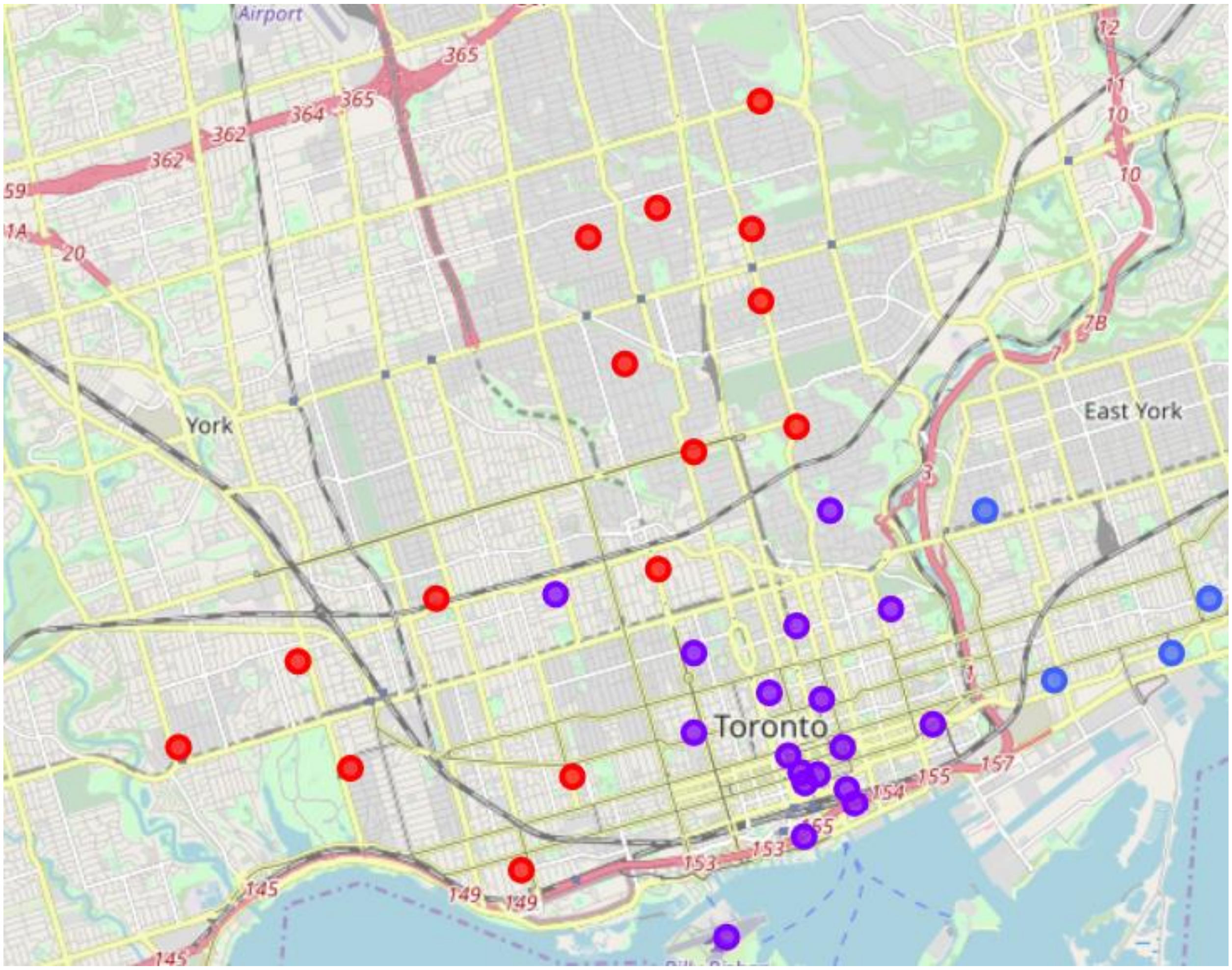
	PostalCode	Borough	Neighborhood
0	M1B	Scarborough	Rouge, Malvern
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae
5	M1J	Scarborough	Scarborough Village
6	M1K	Scarborough	East Birchmount Park, Ionview, Kennedy Park
7	M1L	Scarborough	Clairlea, Golden Mile, Oakridge
8	M1M	Scarborough	Cliffcrest, Cliffside, Scarborough Village West
9	M1N	Scarborough	Birch Cliff, Cliffside West
10	M1P	Scarborough	Dorset Park, Scarborough Town Centre, Wexford ...

We are getting the latitudes and longitudes for the postal codes.

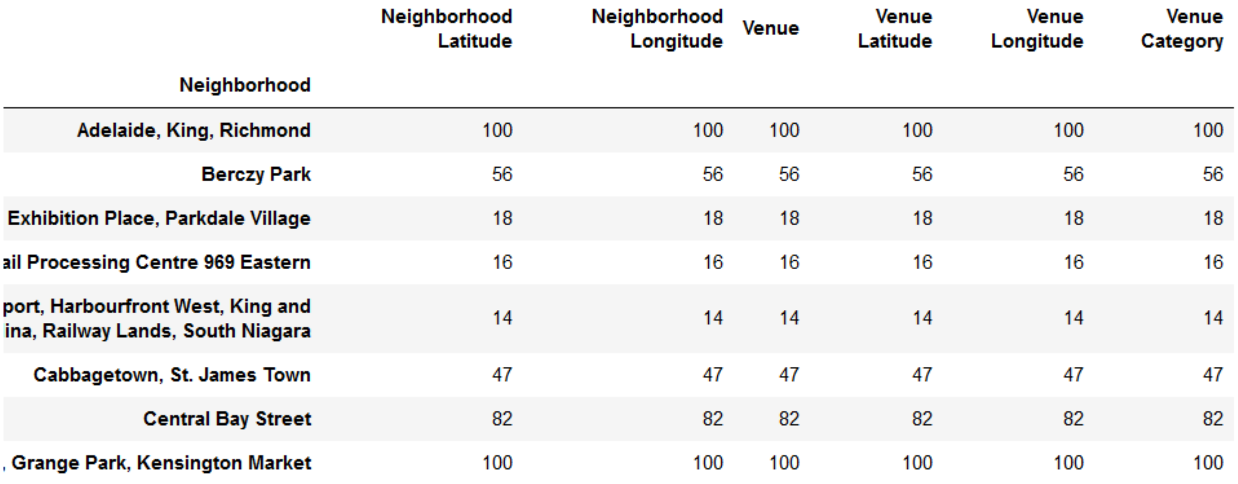


	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

We confirm Toronto has 11 boroughs and 103 neighborhoods and create a map using latitude and longitude values



At this stage we explore Toronto neighborhood with Foursquare API

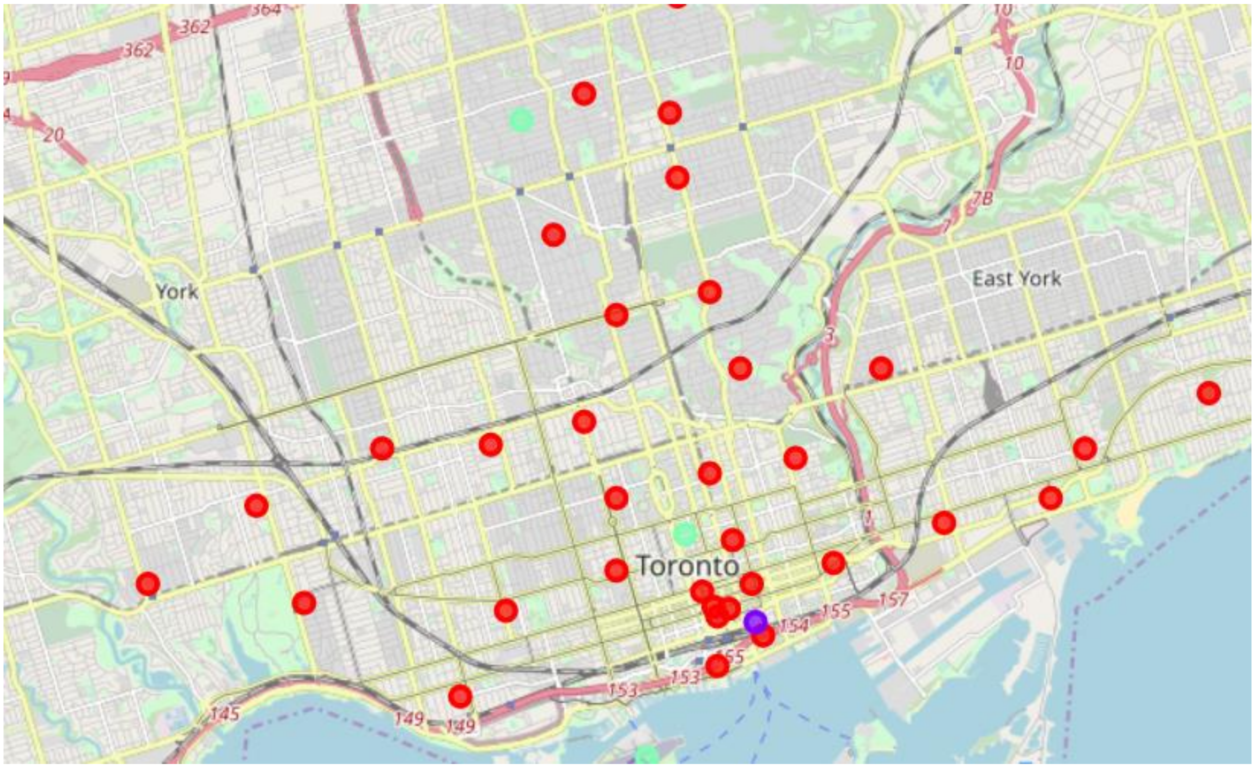


	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Adelaide, King, Richmond	100	100	100	100	100	100
Berczy Park	56	56	56	56	56	56
Exhibition Place, Parkdale Village	18	18	18	18	18	18
ail Processing Centre 969 Eastern	16	16	16	16	16	16
port, Harbourfront West, King and ina, Railway Lands, South Niagara	14	14	14	14	14	14
Cabbagetown, St. James Town	47	47	47	47	47	47
Central Bay Street	82	82	82	82	82	82
, Grange Park, Kensington Market	100	100	100	100	100	100

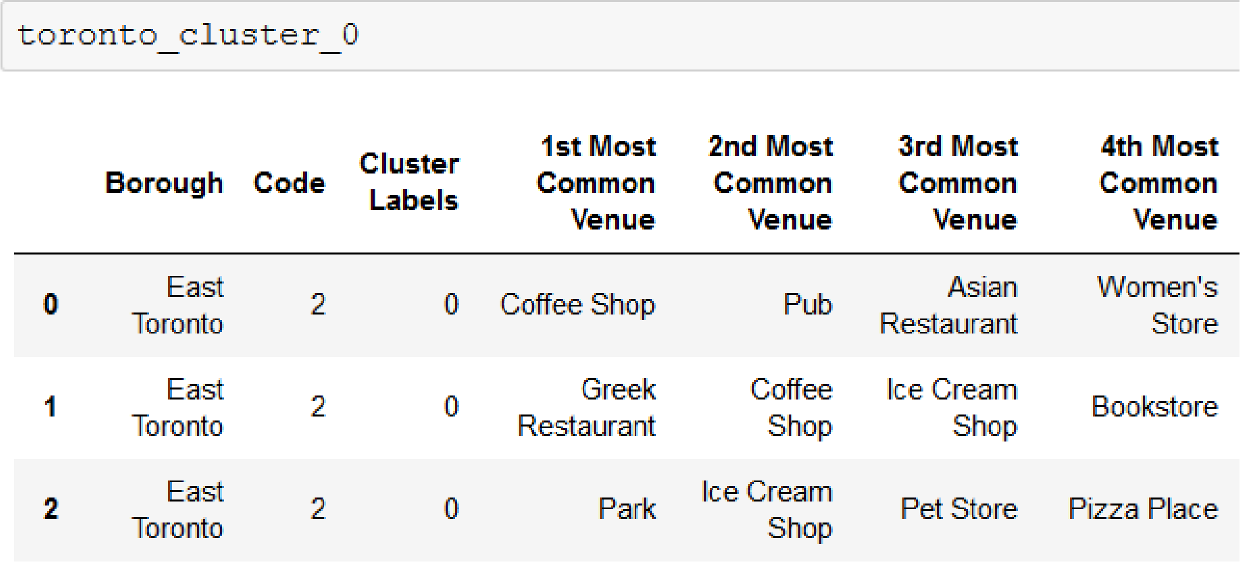
And print the top 5 most common venues

```
----Adelaide, King, Richmond----
                 venue  freq
0          Coffee Shop  0.06
1                 Café  0.05
2           Steakhouse  0.04
3      Thai Restaurant  0.04
4  American Restaurant  0.04


----Berczy Park----
            venue  freq
0     Coffee Shop  0.07
1      Restaurant  0.05
2    Cocktail Bar  0.05
3          Bakery  0.04
4  Farmers Market  0.04
```

We cluster the Neighborhoods using k-means and visualize the clusters



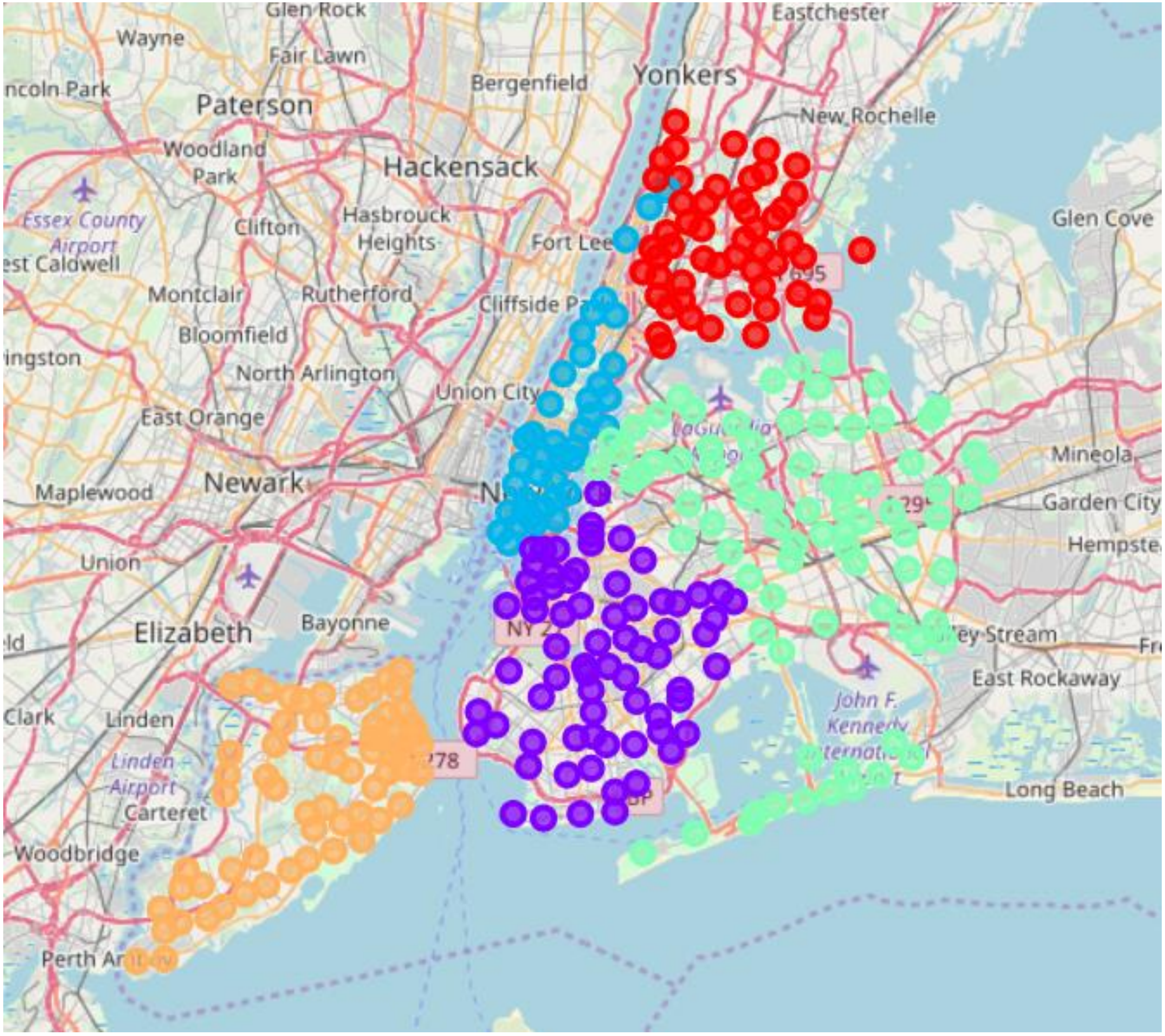
Examine each cluster and determine the discriminating venue categories that distinguish each cluster.



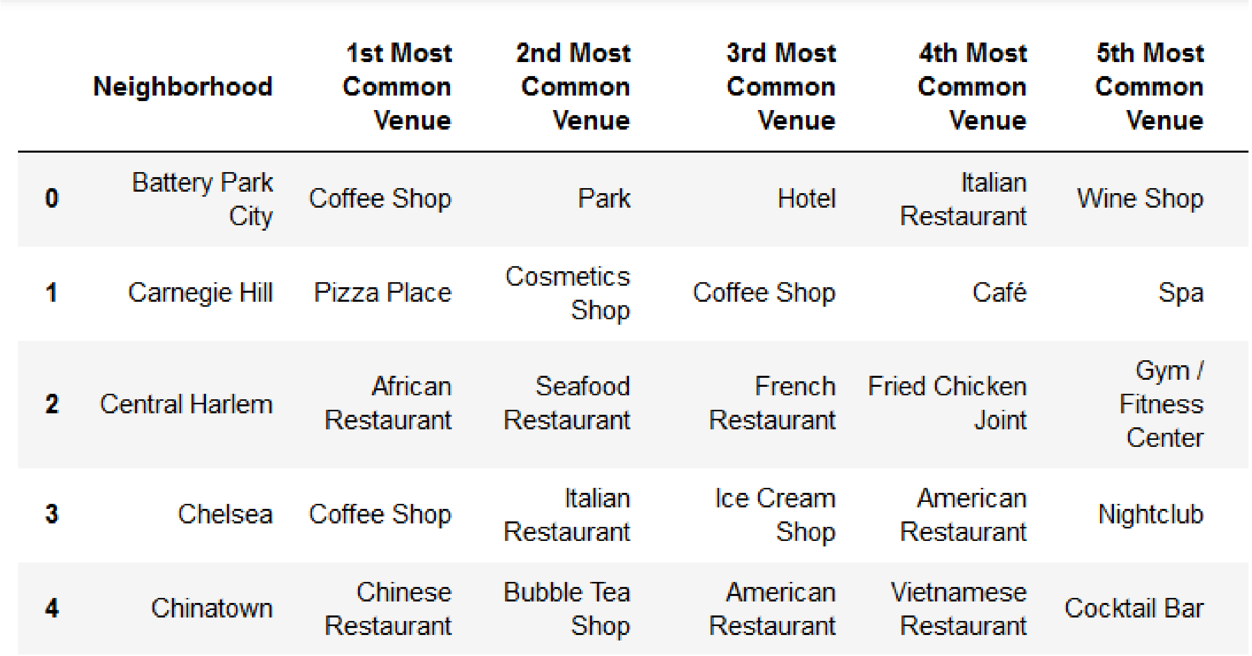
```
toronto_cluster_0
```

	Borough	Code	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	East Toronto	2	0	Coffee Shop	Pub	Asian Restaurant	Women's Store
1	East Toronto	2	0	Greek Restaurant	Coffee Shop	Ice Cream Shop	Bookstore
2	East Toronto	2	0	Park	Ice Cream Shop	Pet Store	Pizza Place

Similar approach is being performed with New York Data, creating a map of New York using latitude and longitude values with neighborhoods superimposed on top

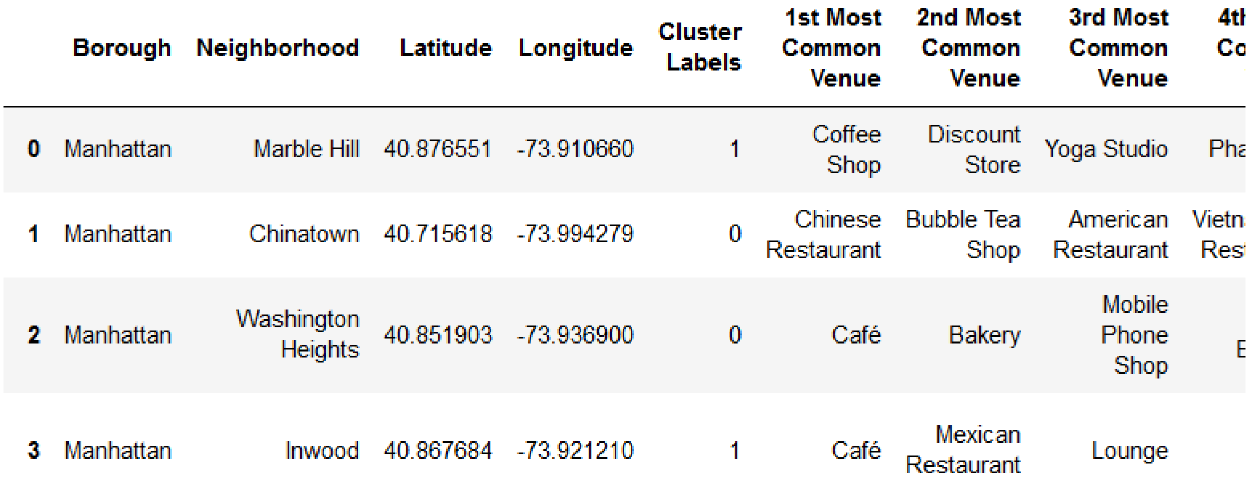


Using Foursquare API , we display the top ten venues for each neighborhood



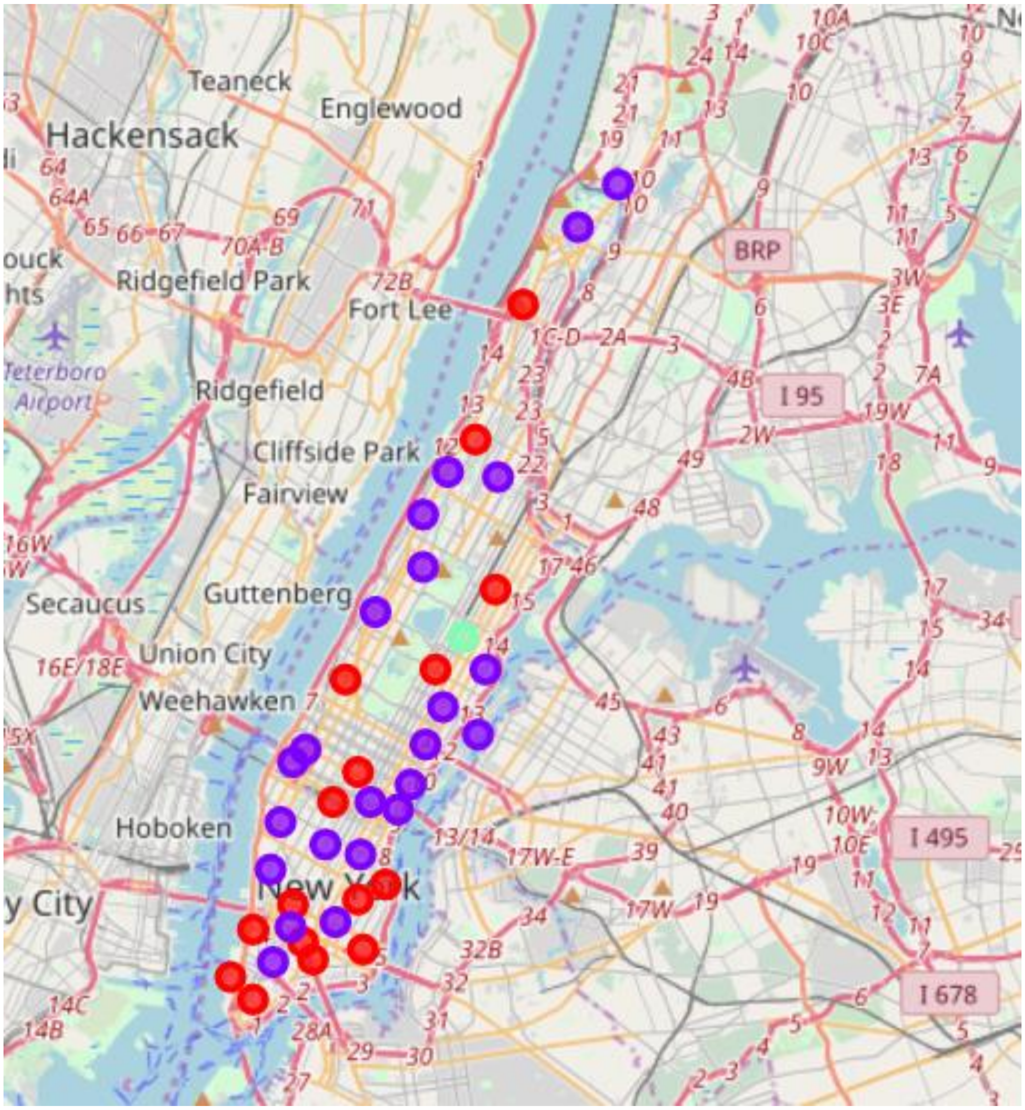
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Battery Park City	Coffee Shop	Park	Hotel	Italian Restaurant	Wine Shop
1	Carnegie Hill	Pizza Place	Cosmetics Shop	Coffee Shop	Café	Spa
2	Central Harlem	African Restaurant	Seafood Restaurant	French Restaurant	Fried Chicken Joint	Gym / Fitness Center
3	Chelsea	Coffee Shop	Italian Restaurant	Ice Cream Shop	American Restaurant	Nightclub
4	Chinatown	Chinese Restaurant	Bubble Tea Shop	American Restaurant	Vietnamese Restaurant	Cocktail Bar

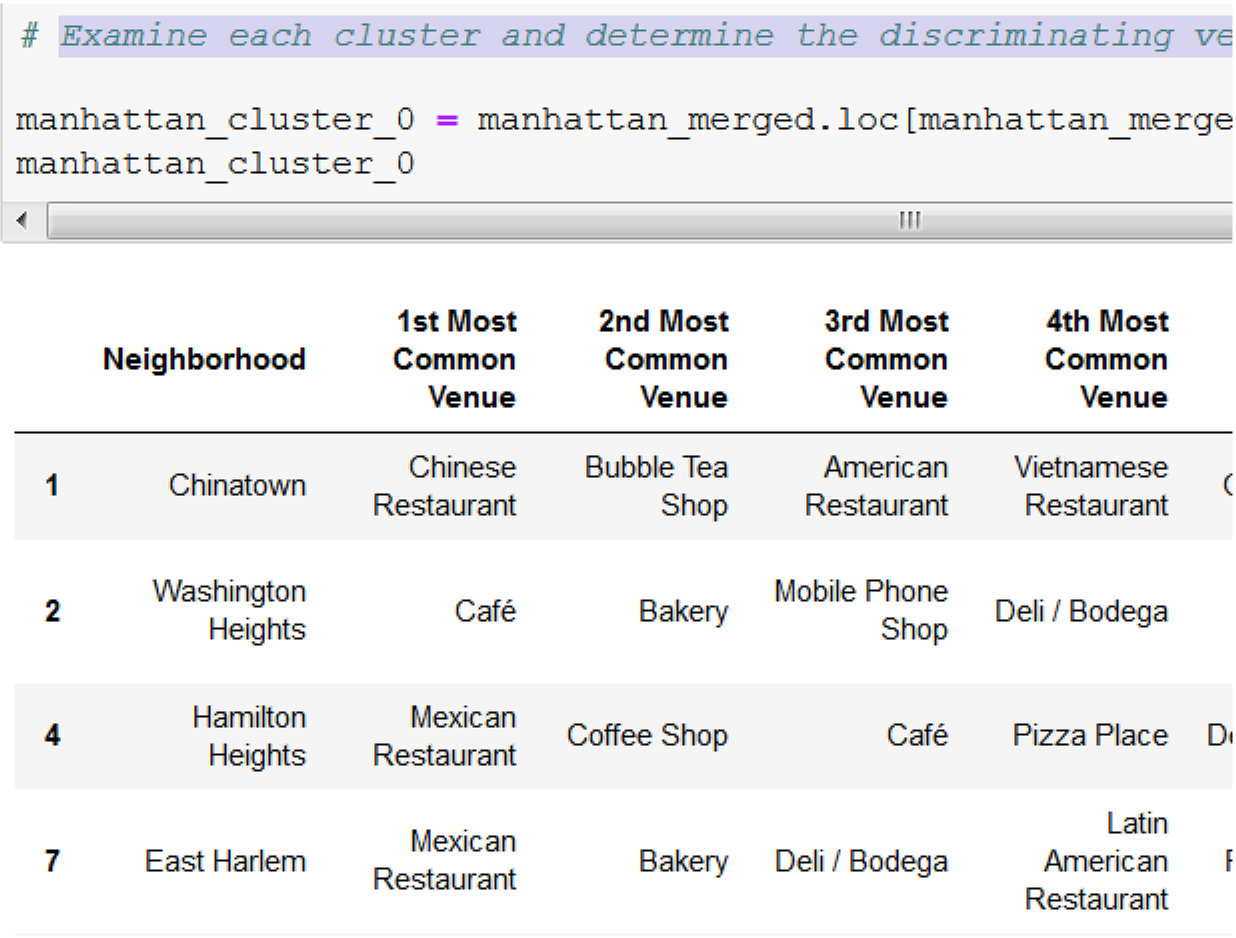
And after clustering the Neighborhoods using k-means



	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4tl Co
0	Manhattan	Marble Hill	40.876551	-73.910660	1	Coffee Shop	Discount Store	Yoga Studio	Pha
1	Manhattan	Chinatown	40.715618	-73.994279	0	Chinese Restaurant	Bubble Tea Shop	American Restaurant	Vietn Res
2	Manhattan	Washington Heights	40.851903	-73.936900	0	Café	Bakery	Mobile Phone Shop	E
3	Manhattan	Inwood	40.867684	-73.921210	1	Café	Mexican Restaurant	Lounge	

Using Foursquare API , we display the top ten venues for each neighborhood and examine each cluster and determine the discriminating venue categories that distinguish each cluster.


Teaneck
Englewood
Hackensack
Ridgefield Park
Fort Lee
Teterboro Airport
Ridgefield
Cliffside Park
Fairview
Secaucus
Guttenberg
Union City
Weehawken
Hoboken
New York
BRP
I 95
I 495
I 678



```
# Examine each cluster and determine the discriminating ve

manhattan_cluster_0 = manhattan_merged.loc[manhattan_merge
manhattan_cluster_0
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	
1	Chinatown	Chinese Restaurant	Bubble Tea Shop	American Restaurant	Vietnamese Restaurant	C
2	Washington Heights	Café	Bakery	Mobile Phone Shop	Deli / Bodega	
4	Hamilton Heights	Mexican Restaurant	Coffee Shop	Café	Pizza Place	D
7	East Harlem	Mexican Restaurant	Bakery	Deli / Bodega	Latin American Restaurant	F

Results and Discussion

Analyzing our results according to the clusters we have produced for Toronto and New York we notice all clusters praise an optimal range of facilities and amenities.

However, there are some particularities regarding the two cities. It seems that in Toronto the most common venue are the Coffee Shop., while in New York the most common venue are restaurants, in particular, Chinese, Mexican, Italian, etc

Conclusion

To sum up, both cities presents a variety of events that would accommodate everyone that would choose to live in any of the cities.