THE BATTLE OF NEIGHBORHOODS

Clustering and Segmentation Project

Table of Contents

1. Background & Business Problem 2
2. Analytics Approach 2
3. Data Sourcing and Requirements 2
4. Data Collection 3
   4.1. Collecting Toronto Data – Web Scraping 3
   4.2. Collecting NEW York Data 5
5. Data Understanding and Preprocessing 7
   5.1. Descriptive Statistics (Exploratory Data Analysis) 7
   5.2. Data Preprocessing and Cleaning 7
6. Data Modelling (Predictive Model) 9
   6.1. Cluster Neighborhoods 9
      6.1.1. Model Selection 9
   6.1.2. Examine Clusters (Results and Discussion) 11
      Cluster 0 12
      Cluster 1 12
      Cluster 2 13
      Cluster 3 13
      Cluster 4 14
      Cluster 5 14
7. Conclusion 15

1. Background & Business Problem

New York city is America's hub for business and pleasure, and is one of the most visited cities across the world due to its touristic sites, landmarks and unique venues. Toronto on the other hand is the biggest Canadian, and also most visited city in Canada. Its full of Canada's richest landmarks, locations, and venues.

The goal of the business case here, is to understand the similarities and differences between 2 cities' venues (Specifically Downtown), and to be able to have a better insight of the demographic and decide on what neighborhoods will be the most suitable to open a specific venue downtown.

The Two cities' downtown neighborhoods will be compared to each other, based on the clusters they fall within. The City of NEW YORK will be compared to the city of TORONTO, to better understand the style of their venues, and how they are similar or dissimilar.

This business case is aimed towards new business venues owners to allow them to decide on what Neighborhoods are the most suitable for their new venue investment such as; Restaurants, coffee shops or other entertainment location.

2. Analytics Approach

K means clustering will be used to segment the cities' neighborhoods and give an idea of how some Neighborhoods are similar or dissimilar to others, based on the venues' categories that exist in each of these Neighborhoods.

3. Data Sourcing and Requirements

The first Data Set to be used is of the NEW YORK city - Including different cities, boroughs and Neighborhoods within NY – Imported from IBM data sets.

The second Data Set is from the Wikipedia page for Toronto city and its Neighborhoods, boroughs and postal codes. The third data set includes Latitude and Longitude of Toronto neighborhoods, provided by IBM data sets

The fourth is of the Foursquare Location Data – API Venues' data acquired using Foursquare account.

Both Data sets of New York and Toronto will utilize the Foursquare location Data, and all of the venues for each neighborhood will be displayed. Finally, both Datasets will be merged together within a bigger data frame that includes different cities (NY and Toronto) and their boroughs and Neighborhoods.

4. Data Collection

4.1. Collecting Toronto Data – Web Scraping

The first step was to acquire a Wikipedia page including all neighborhoods of Toronto. Using Pandas method for web scraping – all data was acquired.

Next step was to pre-process and Filter only valid Boroughs - Removing "Not Assigned" Boroughs. Then I have Included the Latitude and Longitude for each neighborhood.

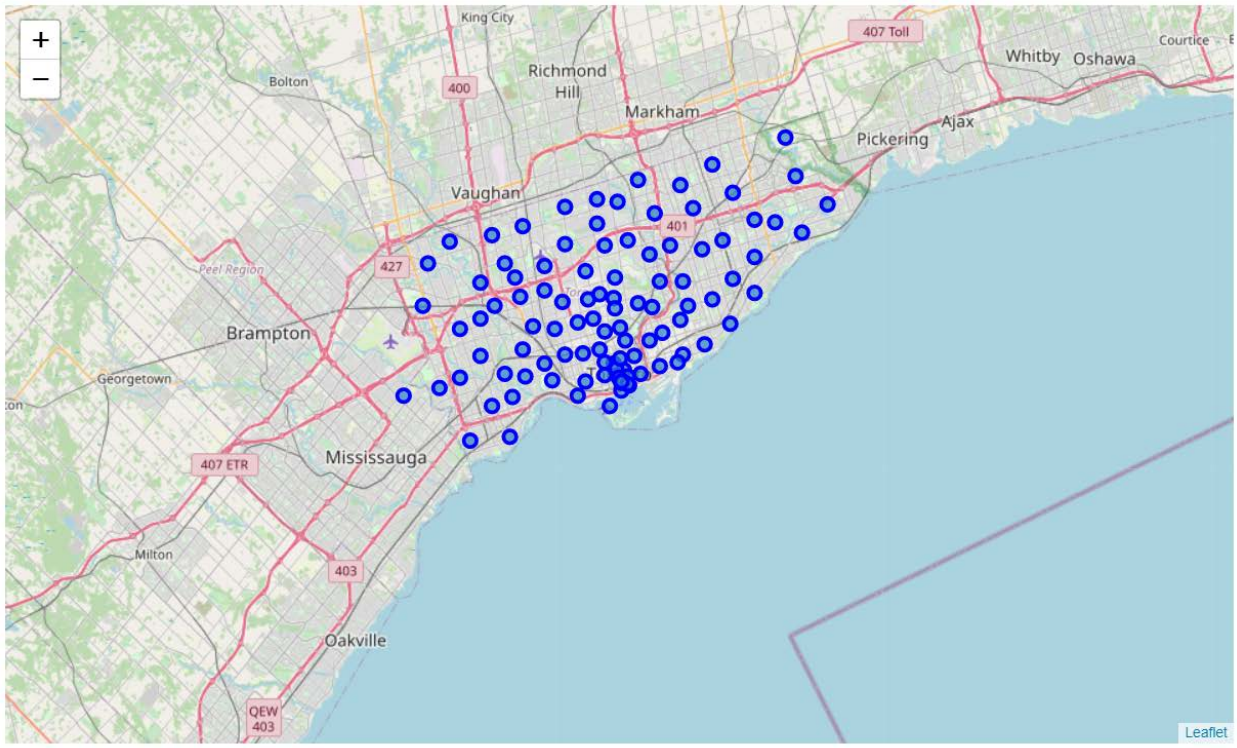
Note that getting Latitude and Longitude of Neighborhoods using Google Geocoder was an option but it is not working as intended - There is also ARC GIS Option but not used. So instead of acquiring the Latitude and longitude, IBM data frame was provided, and I have joined it to my original Toronto data set.

After joining Toronto Neighborhoods with their Latitudes and Longitudes, I was able to explore Neighborhoods and Venues in Toronto, realizing that there are 10 Unique boroughs and 103 Neighborhoods within Toronto. After some cleaning I realized that there is actually 99 unique Neighborhoods, which meant that 4 neighborhoods were repeated but were given different postal codes.

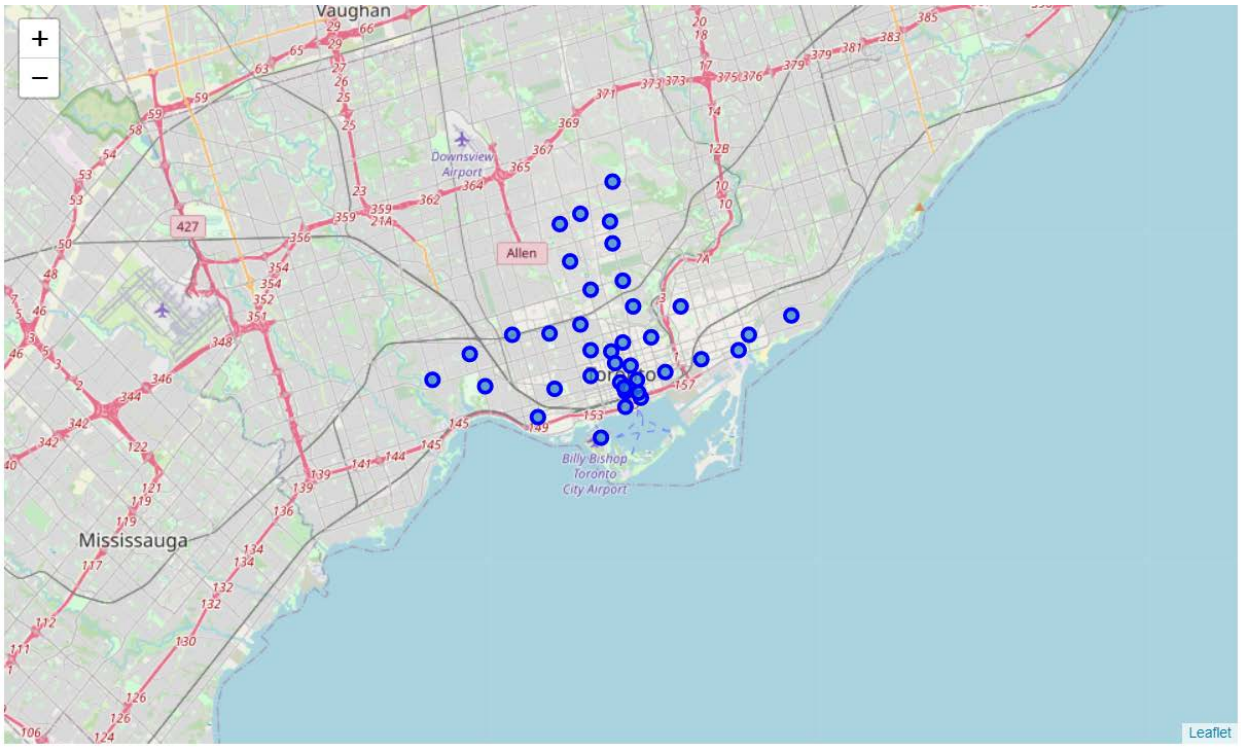
```
North York          24
Downtown Toronto    19
Scarborough         17
Etobicoke           12
Central Toronto      9
West Toronto         6
York                 5
East York            5
East Toronto         5
Mississauga          1
Name: Borough, dtype: int64
```

Follium maps are a great way to visualize location data. Using all of my data variables including Latitude, Longitude, Boroughs, and Neighborhoods, I was able to display each point as a circle with a label on the map outlining different neighborhoods and boroughs in Toronto.

Using Follium maps, centered around Toronto, I explored all of the boroughs across Toronto.



Next stage was to decrease the size of the data frame and thus the cluster points, I have filtered only Downtown Boroughs in Toronto



Foursquare API database contains all the venues for different cities including ratings, trending location, and reviews etc. The next stage was to connect to the Foursquare API using my account and secret ID, and retrieve all venues for the downtown neighborhoods of Toronto from my dataset.

A function was created that would loop over each neighborhood, and retrieve a list of venues, with a limit of 100 venues per each neighborhood. The resulting data frame included each neighborhood and all their associated venues, including the venue name, Lat and Long coordinates, and the Venue Category, which is the most important metric, that will be used later for clustering.

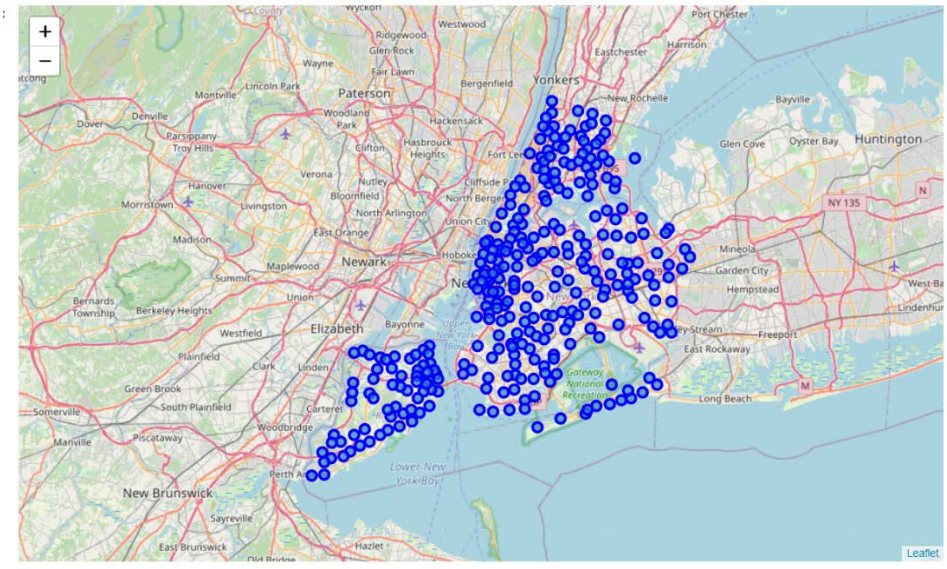
4.2. Collecting NEW York Data

The next objective was to acquire IBM JSON file for New York city data, including all the city information. Since it's a JSON file, the data had to be converted into a data frame. The resulting data frame contained boroughs and neighborhood of New York city. Using descriptive statistics, I realize there are 5 boroughs and 306 neighborhoods in New York.

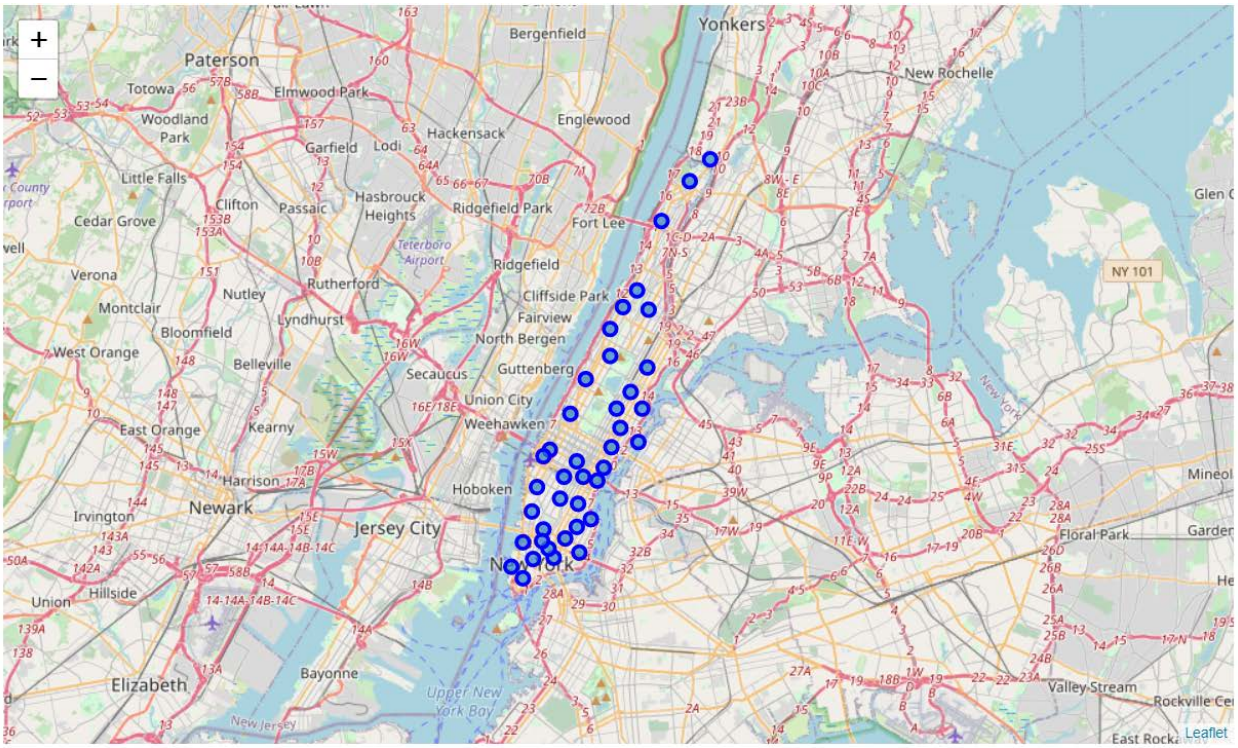
```
Queens           81
Brooklyn         70
Staten Island    63
Bronx            52
Manhattan        40
Name: Borough, dtype: int64
```

Then geopy library was used to get the latitude and longitude values of New York City to center the Follium map around New York.

Using all of the variables in the data frame, including Latitude, longitude, Boroughs and Neighborhoods, I was able to create a map of New York with neighborhoods superimposed on top, with labels indicating the neighborhood and its borough.



To decrease the size of the data frame and thus the cluster of neighborhoods, I filtered by Manhattan as the only borough to focus on, and viewed Manhattan on Follium maps.



Next stage was to again acquire Follium venues using their API, by defining the user name and the secret ID for Foursquare API.

A similar function to Toronto, was created to acquire all venues for each neighborhood in Manhattan, just like I did with Toronto. The resulting Data Frame for New York contained each neighborhood and their associated venues names, and categories, only for neighborhoods within Manhattan.

The last stage of data collection was to merge the two data frames of the two cities vertically (New York & Toronto), having exactly the same features, including the City name, boroughs, Neighborhoods, Venues names, and Venues categories etc.

5. Data Understanding and Preprocessing

5.1. Descriptive Statistics (Exploratory Data Analysis)

As shown above, the first exploratory analysis done was to display the location data points for each of Toronto city, and New York city, to get an idea of how each the downtown neighborhoods within each city are displayed across.

The next descriptive analysis was to count how many venues exist per each neighborhood within each city (The resulting data frame is long but can be viewed in the Notebook).

5.2. Data Preprocessing and Cleaning

After having some preprocessing issues, I realized that there is a Venue category called “Neighborhood” which would conflict with the actual Neighborhoods of each Venue, so I decided to remove the Venue Category labeled as Neighborhood to avoid confusion later.

So now we have a data frame containing two cities and their downtown boroughs, neighborhoods, venues names, and venues categories. The objective is to think of a metric that can be used as a clustering feature.

The most suitable feature in which neighborhoods can be clustered according to, is the Venue Category. The Venue Category describes the category of each venue, which is associated with a neighborhood and a location data, that can show exactly clusters of similar venues.

The first stage of dealing with Venue Category feature was to create a one hot encoded data frame, which displays data horizontally and provides a 1 or 0 when the specific venue category matches the neighborhood in which it lies in. The resulting data frame contained each neighborhood name from each city and their associated venues categories that exist there. Note that in this data frame, neighborhoods are duplicated, since each neighborhood contains multiple venues.

The next stage was to group by Neighborhoods by taking the average of the Venue Categories - Frequency of Occurrence – Using the group by method (Split, Apply, Combine). So now we have unique neighborhoods against, the frequency of occurrence of each venue category for this neighborhood. So, a neighborhood with 0.5 in restaurants, and 0.2 in coffee shops, has a better average for having more restaurants etc. The resulting data frame can be used for a Clustering Model, having different features.

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	F
0	Battery Park City	0.000000	0.00	0.000000	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.0000	0.000000	0.000000	0.000000	
1	Berczy Park	0.000000	0.00	0.000000	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.0000	0.000000	0.000000	0.000000	
2	Brockton, Parkdale Village, Exhibition Place	0.000000	0.00	0.000000	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.0000	0.000000	0.000000	0.000000	
3	Business reply mail Processing Centre, South C...	0.000000	0.00	0.000000	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.0000	0.000000	0.000000	0.000000	
4	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	0.00	0.000000	0.000000	0.0625	0.0625	0.0625	0.125	0.1875	0.0625	0.000000	0.000000	0.000000	
5	Carnegie Hill	0.000000	0.00	0.000000	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.0000	0.000000	0.000000	0.000000	
6	Central Bay Street	0.000000	0.00	0.000000	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.0000	0.000000	0.000000	0.000000	
7	Central Harlem	0.000000	0.00	0.000000	0.068182	0.0000	0.0000	0.0000	0.000	0.0000	0.0000	0.045455	0.000000	0.000000	

Following the same logic and displaying the most occurrent venue category for each neighborhood, a function was created to print each neighborhood along with the top 5 most common venues

```
----Battery Park City----
           venue  freq
0           Park  0.09
1          Hotel  0.07
2            Gym  0.06
3    Coffee Shop  0.06
4  Boat or Ferry  0.04


----Berczy Park----
            venue  freq
0     Coffee Shop  0.09
1  Farmers Market  0.04
2          Bakery  0.04
3            Café  0.04
4     Cheese Shop  0.04


----Brockton, Parkdale Village, Exhibition Place----
               venue  freq
0               Café  0.14
1     Breakfast Spot  0.09
2        Coffee Shop  0.09
3  Convenience Store  0.05
4         Restaurant  0.05
```

The final pre-processioning stage included creating a data frame containing the 10 most common venues for each neighborhood, but displayed horizontally

	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	Hotel	Gym	Coffee Shop	Boat or Ferry	Memorial Site	Playground	Shopping Mall	Sandwich Place	Gourmet Shop
1	Berczy Park	Coffee Shop	Seafood Restaurant	Bakery	Café	Farmers Market	Cocktail Bar	Cheese Shop	Beer Bar	Restaurant	French Restaurant
2	Brockton, Parkdale Village, Exhibition Place	Café	Coffee Shop	Breakfast Spot	Convenience Store	Gym	Restaurant	Italian Restaurant	Performing Arts Venue	Nightclub	Intersection
3	Business reply mail Processing Centre, South C...	Yoga Studio	Gym / Fitness Center	Comic Shop	Restaurant	Park	Skate Park	Smoke Shop	Burrito Place	Brewery	Farmers Market
4	CN Tower, King and Spadina, Railway Lands, Har...	Airport Service	Airport Lounge	Boutique	Sculpture Garden	Coffee Shop	Harbor / Marina	Rental Car Location	Boat or Ferry	Airport Terminal	Bar

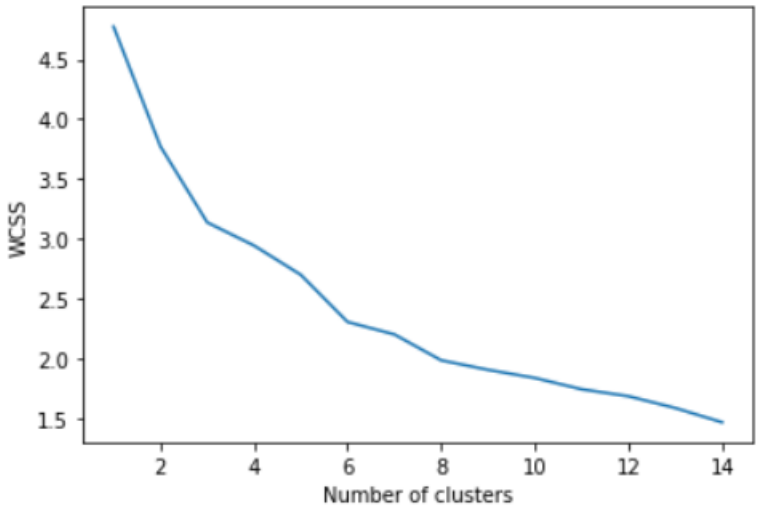
6. Data Modelling (Predictive Model)

6.1. Cluster Neighborhoods

6.1.1. Model Selection

I have decided to utilize the use of the K means Clustering model from SKLEARN.

First and using the Elbow Method, I wanted to know the suitable number of Clusters - using “Within Cluster Sum of Squares” WCSS



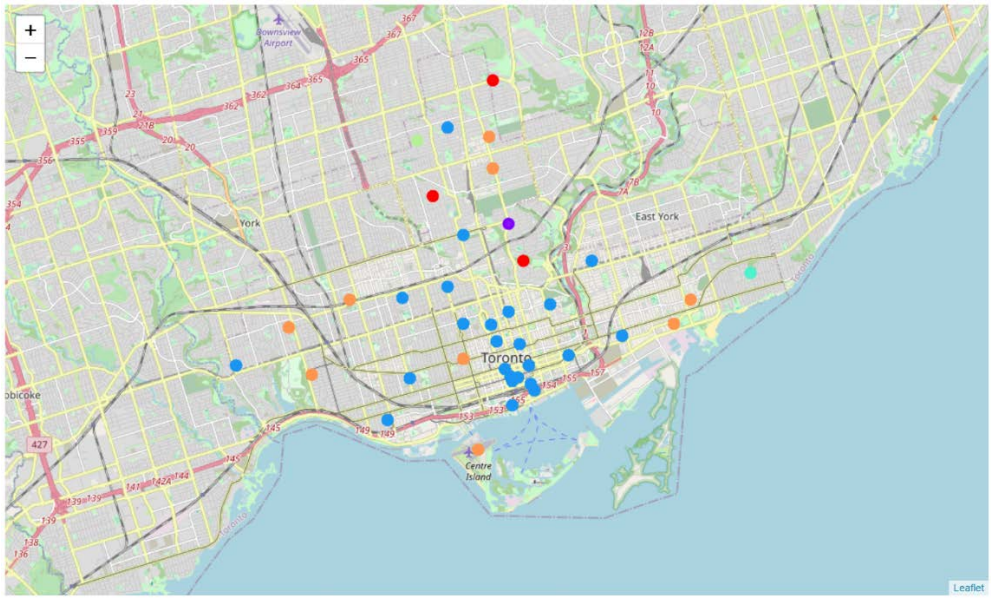
I have decided to cluster based on 6 clusters, using an init of K means++, random state of 1, and a convergence number of 12 iterations.

The resulted cluster labels from the trained model, can now be added to their associated neighborhoods in the last data frame we had. Now the data frame has all information about both cities’ boroughs, Neighborhoods, venues, and the number of clusters their neighborhoods are associated to.

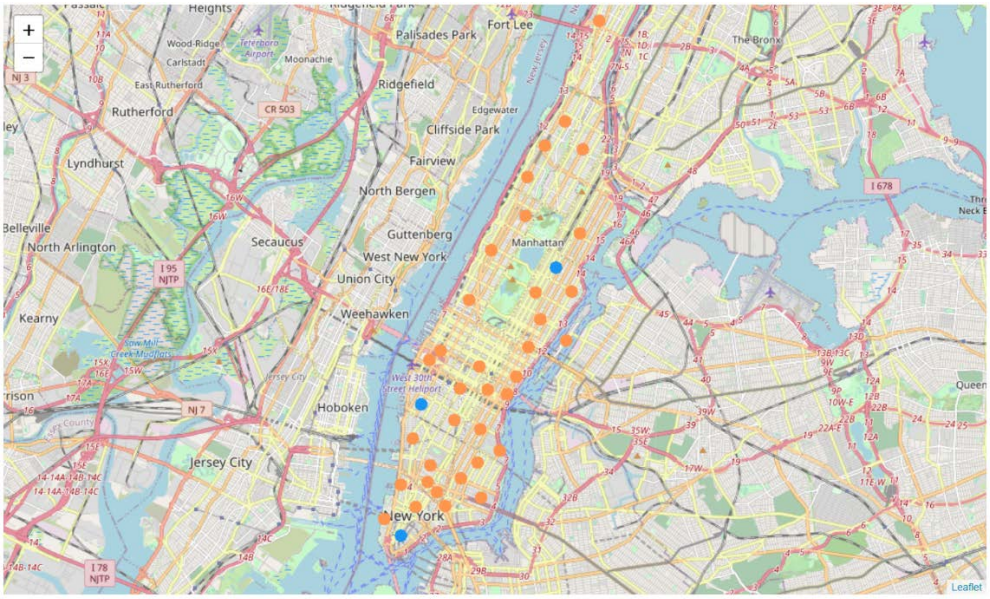
	City	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	TORONTO	Downtown Toronto	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery	0	Coffee Shop	Pub	Bakery
1	TORONTO	Downtown Toronto	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop	0	Coffee Shop	Pub	Bakery
2	TORONTO	Downtown Toronto	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center	0	Coffee Shop	Pub	Bakery
3	TORONTO	Downtown Toronto	Regent Park, Harbourfront	43.65426	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant	0	Coffee Shop	Pub	Bakery
4	TORONTO	Downtown Toronto	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa	0	Coffee Shop	Pub	Bakery

Since the data frame has the two cities in 1 data frame, it is possible to see all of the clusters at once, however since Toronto is far from New York, Follium maps will have to be centered in between the two cities, and the clusters will not be properly display. So, I have decided to view the same map from Toronto Downtown point of view, and then view the same map from New York Manhattan point of view, to see the distribution of clusters in Downtown Toronto Vs Downtown New York

1- Toronto



2- New York



From Initial Inspection we can see that most of the Toronto Neighborhoods exists in the same cluster where most of NY Neighborhoods are as well - Marked orange

6.1.2. Examine Clusters (Results and Discussion)

To continue understanding how the clusters of Toronto compare to New York Manhattan, I have decided to merge the resulted clusters Data Frame with the categories frequency of Occurrence as shown below

	City	Neighborhood	Cluster Labels	Venue	Venue Category	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant
0	TORONTO	Regent Park, Harbourfront	2	Roselle Desserts	Bakery	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	TORONTO	Regent Park, Harbourfront	2	Tandem Coffee	Coffee Shop	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	TORONTO	Regent Park, Harbourfront	2	Cooper Koo Family YMCA	Distribution Center	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

The shown data frame contains the Neighborhoods, the Cluster Labels, and the one hot encoded data from the frequency of occurrence data frame.

Then I filtered that data frame twice, once with Toronto as the city, and the second time as New York as the city, and grouped by Cluster labels for each city. **This is important as it can show what's inside each cluster for each city.**

Toronto Clusters

```
toronto_clusters = two_cities_merged_with_onehot[two_c
toronto_clusters.head(5)
```

Cluster Labels	0	1	2	3	4	5
Accessories Store	0.0	0.0	0.000000	0.0	0.0	0.000000
Adult Boutique	0.0	0.0	0.000000	0.0	0.0	0.000000
Afghan Restaurant	0.0	0.0	0.000718	0.0	0.0	0.000000
African Restaurant	0.0	0.0	0.000000	0.0	0.0	0.000000
Airport	0.0	0.0	0.000000	0.0	0.0	0.004545

New York Clusters

```
ny_clusters = two_cities_merged_with_onehot[two_cities
ny_clusters.head(5)
```

Cluster Labels	2	5
Accessories Store	0.0	0.001026
Adult Boutique	0.0	0.000342
Afghan Restaurant	0.0	0.000000
African Restaurant	0.0	0.001026
Airport	0.0	0.000000

TORONTO Neighborhood Venues fall within 5 clusters from 0 to 5, whereas NEW YORK MANHATTAN Neighborhood Venues fall within only 2 clusters, cluster 2, and 5

The final task was to analyze each cluster at a time, and understand how each city performs in that cluster. Since most of the data points of Toronto, and ALL the data points or neighborhoods of New York fall under clusters 2 and 5, They are the main cluster of interest for the restaurant investors.

For each cluster, same analysis was completed, all venues from the two cities were displayed together, Neighborhoods from both cities were outlined together, and most importantly, top 10 Venues from Toronto vs Top 10 Venues from New York Manhattan were displayed

Cluster 0

Cluster 0 is for outside of Downtown Toronto entertainment venues such as parks, trails and playgrounds - Doesn't contain any Manhattan venues

3) TORONTO VENUES CATEGORIES IN CLUSTER 0

```
toronto_clusters.iloc[:,0].sort_values(ascending=False).head(10)
```

```
Park                0.363636
Trail               0.181818
Jewelry Store       0.090909
Playground          0.090909
Sushi Restaurant    0.090909
Swim School         0.090909
Bus Line            0.090909
Yoga Studio         0.000000
Drugstore           0.000000
Dry Cleaner         0.000000
Name: 0, dtype: float64
```

Cluster 1

Similar to Cluster 0, Cluster 1 is common with Entertainment venues such as Playgrounds, trails, yoga studios which are common outside the Toronto downtown area

Cluster 2

Cluster 2 has all of the downtown most common venues as shown below - Both Toronto and New York Manhattan have venues such as coffee shops, Italian restaurants and Bars - Toronto is unique for its downtown Japanese restaurants while New York Manhattan has American Restaurants

3) TORONTO VENUES CATEGORIES IN CLUSTER 2

```
toronto_clusters.iloc[:,2].sort_values(ascending=False).head(10)
```

```
Coffee Shop            0.106963
Café                   0.053841
Restaurant             0.037330
Hotel                  0.028715
Japanese Restaurant    0.022254
Italian Restaurant     0.021536
Bakery                 0.019383
Bar                    0.017229
Gym                    0.016511
Park                   0.015793
Name: 2, dtype: float64
```

4) NEWYORK VENUES CATEGORIES IN CLUSTER 2

```
ny_clusters.iloc[:,0].sort_values(ascending=False).head(10)
```

```
Coffee Shop            0.093103
Café                   0.031034
Bar                    0.027586
Pizza Place            0.027586
American Restaurant    0.024138
Hotel                  0.024138
Cocktail Bar           0.024138
Gym                    0.024138
Italian Restaurant     0.024138
Park                   0.020690
Name: 2, dtype: float64
```

Cluster 3

Cluster 3 is about out of town pubs, trails, and health food stores

3) TORONTO VENUES CATEGORIES IN CLUSTER 3

```
toronto_clusters.iloc[:,3].sort_values(ascending=False).head(10)
```

```
Pub                            0.333333
Trail                          0.333333
Health Food Store              0.333333
Dog Run                        0.000000
Donut Shop                     0.000000
Drugstore                      0.000000
Dry Cleaner                    0.000000
Dumpling Restaurant            0.000000
Duty-free Shop                 0.000000
Eastern European Restaurant    0.000000
Name: 3, dtype: float64
```

Cluster 4

Cluster 4 is about entertainment venues such as music venues, event spaces, and gardens

3) TORONTO VENUES CATEGORIES IN CLUSTER 4

```
toronto_clusters.iloc[:,4].sort_values(ascending=False).head(10)
```

```
Garden                          0.5
Music Venue                     0.5
Event Space                     0.0
Drugstore                       0.0
Dry Cleaner                     0.0
Dumpling Restaurant             0.0
Duty-free Shop                  0.0
Eastern European Restaurant     0.0
Electronics Store               0.0
Empanada Restaurant             0.0
Name: 4, dtype: float64
```

Cluster 5

Cluster 5 just like cluster 2 has all of the downtown most common venues as shown below - Both Toronto and New York have venues such as coffee shops, Italian restaurants and Bars

3) TORONTO VENUES CATEGORIES IN CLUSTER 5

```
toronto_clusters.iloc[:,5].sort_values(ascending=False).head(10)
```

```
Bar                  0.040909
Café                 0.040909
Coffee Shop          0.040909
Park                 0.040909
Pizza Place          0.031818
Dessert Shop         0.027273
Mexican Restaurant   0.027273
Bakery               0.022727
Brewery              0.022727
Italian Restaurant   0.022727
Name: 5, dtype: float64
```

4) NEWYORK VENUES CATEGORIES IN CLUSTER 5

```
ny_clusters.iloc[:,1].sort_values(ascending=False).head(10)
```

```
Italian Restaurant     0.041368
Coffee Shop            0.037949
Café                   0.026325
Bakery                 0.023590
American Restaurant    0.023590
Pizza Place            0.023248
Park                   0.021197
Hotel                  0.019145
Bar                    0.017778
Mexican Restaurant     0.017778
Name: 5, dtype: float64
```

7. Conclusion

The two cities data have been combined into 1 data frame - Containing each city Boroughs, Neighborhoods, and Venues - With the focus on downtown boroughs from each city (Downtown NY Manhattan VS Downtown Toronto)

After Segmenting the entire data frame containing the two cities downtown regions - we can see that Downtown New York neighborhoods all fall within 2 clusters (2 and 5) whereas most of Toronto downtown neighborhoods/venues fall within the same clusters, and the rest of the neighborhoods scattered around downtown Toronto, are within the other clusters

Cluster 2 and 5 has all of New York downtown Neighborhood venues, and most of Toronto Downtown Neighborhood Venues

Cluster 2 and 5 shows similarity between downtown Network and Downtown Toronto in the Type of venues categories available - With common venues such as:

- Expensive Italian restaurants
- Expensive Japanese Restaurants
- Coffee shops
- Hotels
- Parking spots
- Bars
- Bakeries

For clients looking to invest into any of the above venues, it would be a great idea to invest downtown - with creative venues that offer similar categories but with different flavors - such as expensive but foreign restaurant style venues etc