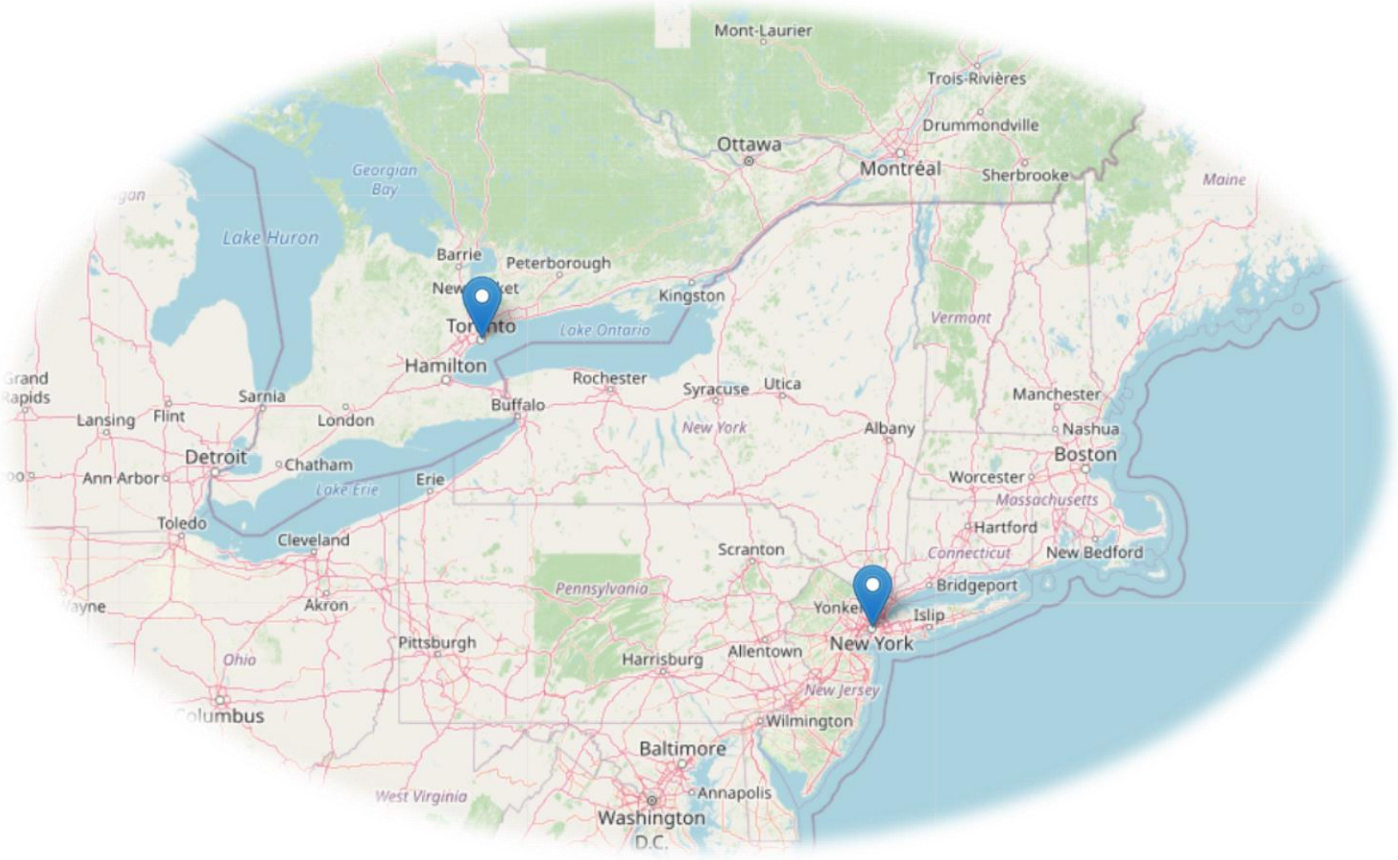
A JOURNEY TO TORONTO AND NYC

EXPLORING AND CLUSTERING NEIGHBORHOODS



MOSTAFA ELSEIDY

IBM APPLIED DATA SCIENCE CAPSTONE ON COURSERA

Contents

1. Introduction

A. Background

Toronto is the largest city of Canada and capital city of Ontario Province situated north-western shore of Ontario Lake.

New York City is situated in the northeastern United States, in southeastern New York State, approximately halfway between Washington, D.C. and Boston.

They are one of the leading economic sectors including business services, finance, aerospace, telecommunications, media, transportation, arts, film, television, production, publishing, media research, education, software production, engineering, sports industries, and tourism in their respective countries.

Different geographical location of cities not always means different neighborhoods inside those cities.

B. Problem Description

Say you live in the city of Toronto in Canada. You love your neighborhood, mainly because of all the great amenities and other types of venues that exist in the neighborhood, such as gourmet fast food joints, pharmacies, parks, graduate schools and so on. Now say you decide to reside in the city of New York in USA (or vice versa).

Wouldn't it be great if you are able to determine neighborhoods in the other city that are the same as current neighborhood in your city, and if not, perhaps similar neighborhoods.

C. Objective

The aim of this report is to study and analyze the neighborhoods of Toronto city and Queens in New York city and group them into similar clusters and, to analyze those clusters to gather meaningful information. That information can be used to find out neighborhoods that are same as your current neighborhood or at least similar.

D. Target Audience

This information provided by this report would be useful for people who are interested in relocating to a different part of the city or another city and are interested in finding new neighborhoods that are similar to their existing neighborhood or to discover other neighborhoods.

2. Data Description

For this project, I will basically use 3 sources Coursera, Wikipedia and Foursquare.

- The first source comes in handy (I must take things easy for now as I am still learning) and it is the list previously provided by Coursera instructor: New York City data which contains list Boroughs, Neighborhoods along with their latitude and longitude.
- To get Toronto neighborhoods data I will use Wikipedia.
- To be able to reach the venues and their information I will use a foursquare API source.

A. Neighborhoods data

a. Neighborhoods in Toronto city from Wikipedia. The dataset contains the following columns:
   - Postal Code
   - Borough
   - Neighborhood Name

b. Neighborhoods in Queens New York city from New York data json file on Coursera. The dataset contains the following columns:
   - Borough
   - Neighborhood Name
   - Latitude
   - Longitude

B. geographical coordinates data

for each Neighborhood in Toronto from csv file on Coursera and merge it with neighborhoods data in Toronto

- Postal Code
- Borough
- Neighborhood Name
- Latitude
- Longitude

C. Venue data from Foursquare

for each neighborhood get the most common venues in radius 500 meters for both Toronto and Queens in NYC

- Neighborhood name, Latitude, Longitude
- Venue, Latitude, Longitude, Category

3. Data Preprocessing

- Collect the Toronto and Queens NYC data.
- Get for each neighborhood its geographical coordinates.
- Clean neighborhood data.
- Using Foursquare API, we will find all venues for each neighborhood.
- Combine and merge neighborhood data.
- Combine and merge venues data.

The data preparation for each of the 3 sources of data is done separately.

A. Scrape and Clean the data

a. Neighborhoods and geographical coordinates data

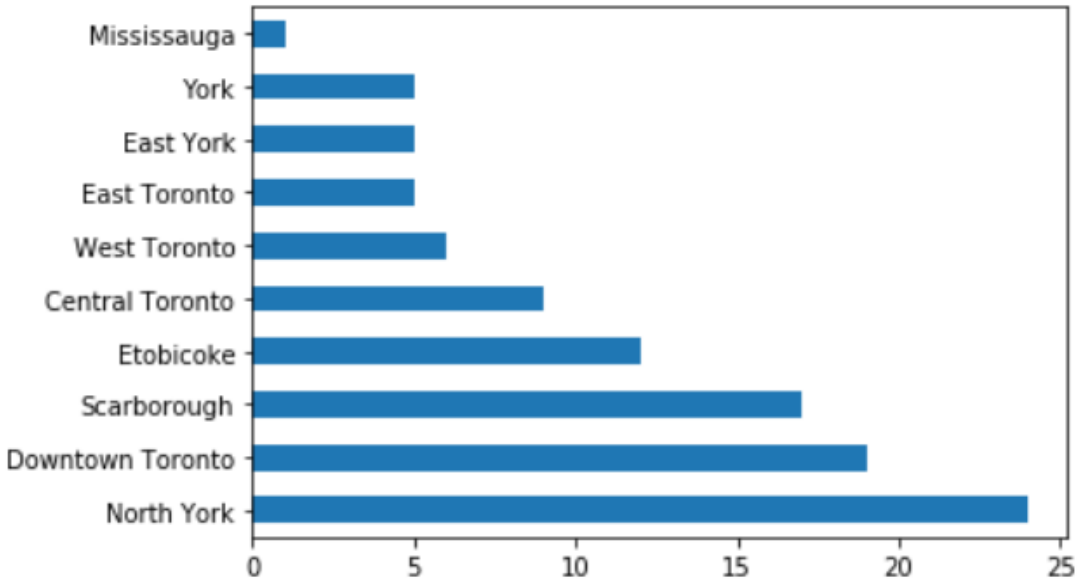
1. Toronto city:
    I. Scrape from Wikipedia
    II. Clean the data

	Postal Code	Borough	Neighborhood
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	Regent Park, Harbourfront

- Only process the cells that have an assigned borough. Ignore cells with a borough that is **Not assigned.**
- More than one neighborhood can exist in one postal code area.

- If a cell has a borough but a **Not assigned** neighborhood, then the neighborhood will be the same as the borough.

	PostalCode	Borough	Neighborhood
0	M5G	Downtown Toronto	Central Bay Street
1	M2H	North York	Hillcrest Village
2	M4B	East York	Parkview Hill, Woodbine Gardens
3	M1J	Scarborough	Scarborough Village
4	M4G	East York	Leaside
5	M4M	East Toronto	Studio District
6	M1R	Scarborough	Wexford, Maryvale
7	M9V	Etobicoke	South Steeles, Silverstone, Humbergate, Jamest...
8	M9L	North York	Humber Summit
9	M5V	Downtown Toronto	CN Tower, King and Spadina, Railway Lands, Har...
10	M1B	Scarborough	Malvern, Rouge
11	M5A	Downtown Toronto	Regent Park, Harbourfront



- Use only Borough that contains “Toronto”

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M4E	East Toronto	The Beaches	43.676357	-79.293031
1	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188
2	M4L	East Toronto	India Bazaar, The Beaches West	43.668999	-79.315572
3	M4M	East Toronto	Studio District	43.659526	-79.340923
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790

III. Show on map with Folium

2. Queens NYC:

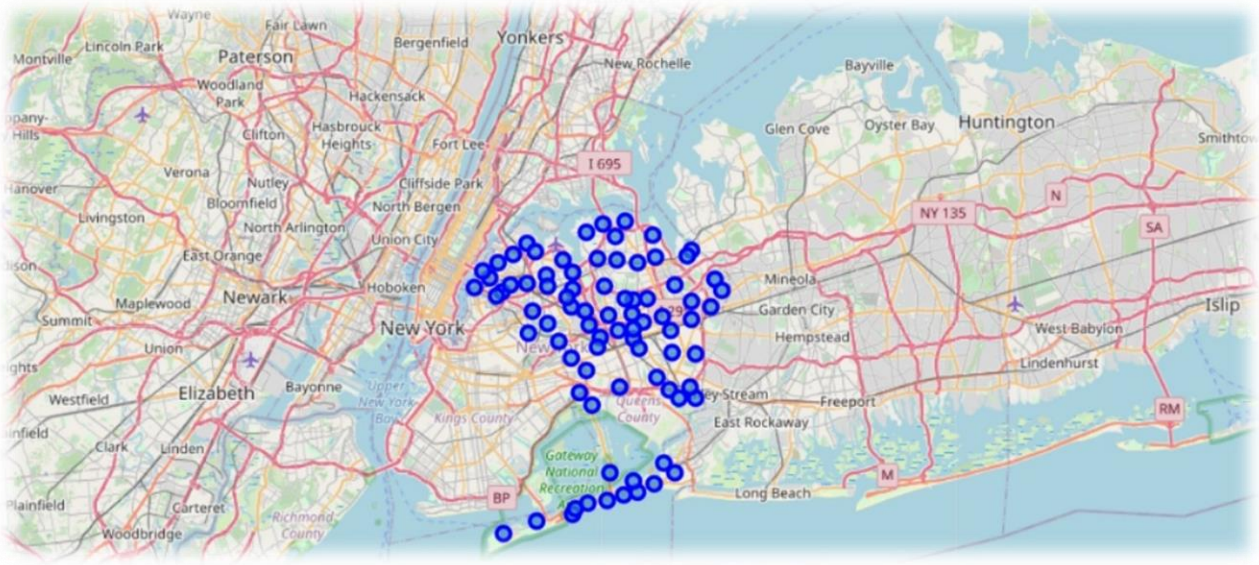
I. Get data from json file

II. Clean the data

- Only process the cells that have an assigned borough. Ignore cells with a borough that is Not assigned.
- More than one neighborhood can exist in one postal code area.
- If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.
- Use only Borough that contains “Queens”

	Borough	Neighborhood	Latitude	Longitude
0	Queens	Astoria	40.768509	-73.915654
1	Queens	Woodside	40.746349	-73.901842
2	Queens	Jackson Heights	40.751981	-73.882821
3	Queens	Elmhurst	40.744049	-73.881656
4	Queens	Howard Beach	40.654225	-73.838138

III. Show on map with Folium



b. Venue data from Foursquare

1. Toronto city

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
1	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
2	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub
3	The Beaches	43.676357	-79.293031	Upper Beaches	43.680563	-79.292869	Neighborhood
4	The Danforth West, Riverdale	43.679557	-79.352188	MenEssentials	43.677820	-79.351265	Cosmetics Shop

2. Queens NYC

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Astoria	40.768509	-73.915654	Favela Grill	40.767348	-73.917897	Brazilian Restaurant
1	Astoria	40.768509	-73.915654	Orange Blossom	40.769856	-73.917012	Gourmet Shop
2	Astoria	40.768509	-73.915654	Simply Fit Astoria	40.769114	-73.912403	Gym
3	Astoria	40.768509	-73.915654	CrossFit Queens	40.769404	-73.918977	Gym
4	Astoria	40.768509	-73.915654	Titan Foods Inc.	40.769198	-73.919253	Gourmet Shop

B. Analyze each neighborhood

We use One Hot Encoding, use the neighborhood to group data, and find out the top ten venues present in each neighborhood.

	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	Arverne	Surf Spot	Metro Station	Sandwich Place	Coffee Shop	Pizza Place	Beach	Donut Shop	Bus Stop	Café	Restaurant
1	Astoria	Middle Eastern Restaurant	Bar	Greek Restaurant	Indian Restaurant	Mediterranean Restaurant	Pizza Place	Seafood Restaurant	Hookah Bar	Bakery	Food Truck
2	Astoria Heights	Hostel	Gourmet Shop	Playground	Plaza	Business Service	Bus Station	Burger Joint	Shopping Mall	Bowling Alley	Supermarket
3	Auburndale	Italian Restaurant	Furniture / Home Store	Mobile Phone Shop	Korean Restaurant	Bar	Toy / Game Store	Discount Store	Supermarket	Train	American Restaurant
4	Bay Terrace	Clothing Store	Women's Store	Donut Shop	Kids Store	Lingerie Store	American Restaurant	Mobile Phone Shop	Shoe Store	Cosmetics Shop	Movie Theater

4. Methodology

Apply machine learning

Visualize neighborhoods clusters using folium library

A. Modeling and Machine Learning

We have some common venue categories in the neighborhoods. We use the unsupervised learning K-means algorithm to cluster the neighborhoods. K-Means algorithm is one of the most common method for clustering in unsupervised learning.

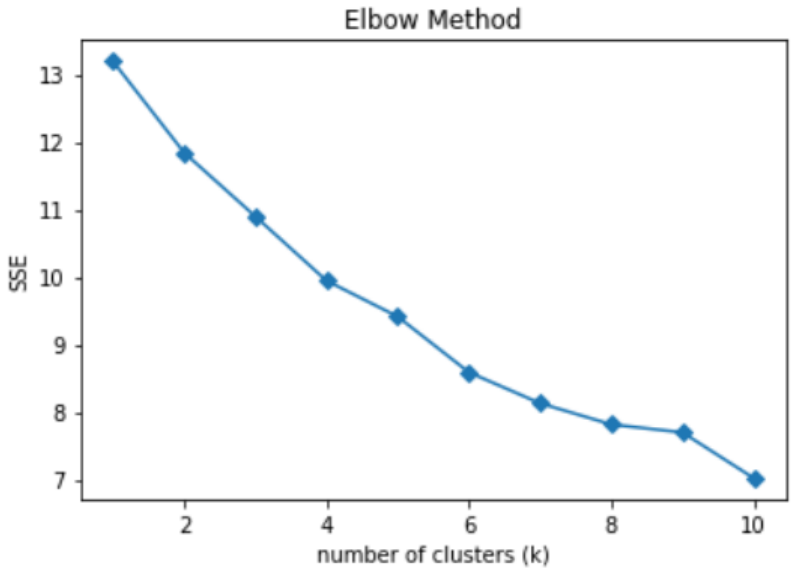
a. K-Means

- K-means clustering is an unsupervised machine learning algorithm that creates clusters within your data, which can help you to discover categories or groups that you might not have seen on your own.
- To implement the k-means, it is very important to determine the optimal number of clusters ‘K’.

Determine the value of optimal k with the Elbow Method and Silhouette Score.

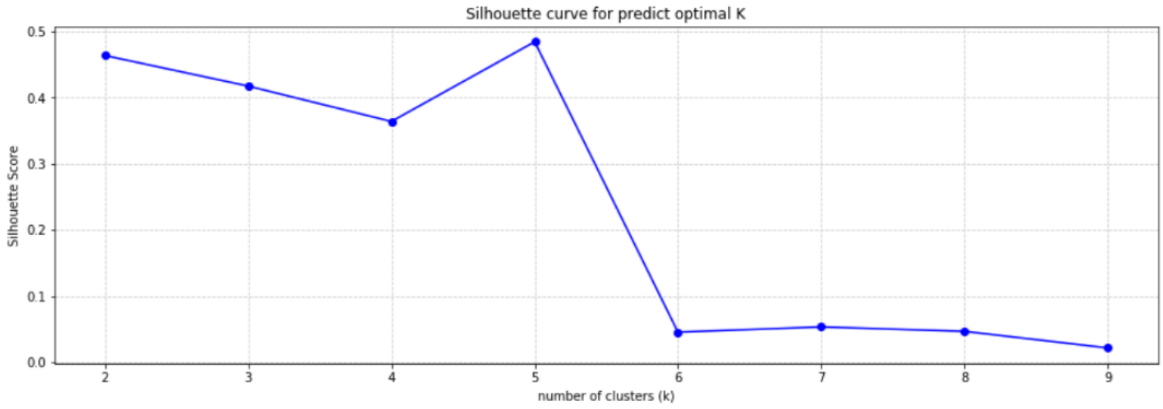
b. The Elbow Method

- Calculates the sum of squared distance of samples to their closest cluster center for different values of “K”.
- The optimal number of clusters is the value after which there is no significant decrease



C. Silhouette Score

- The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).



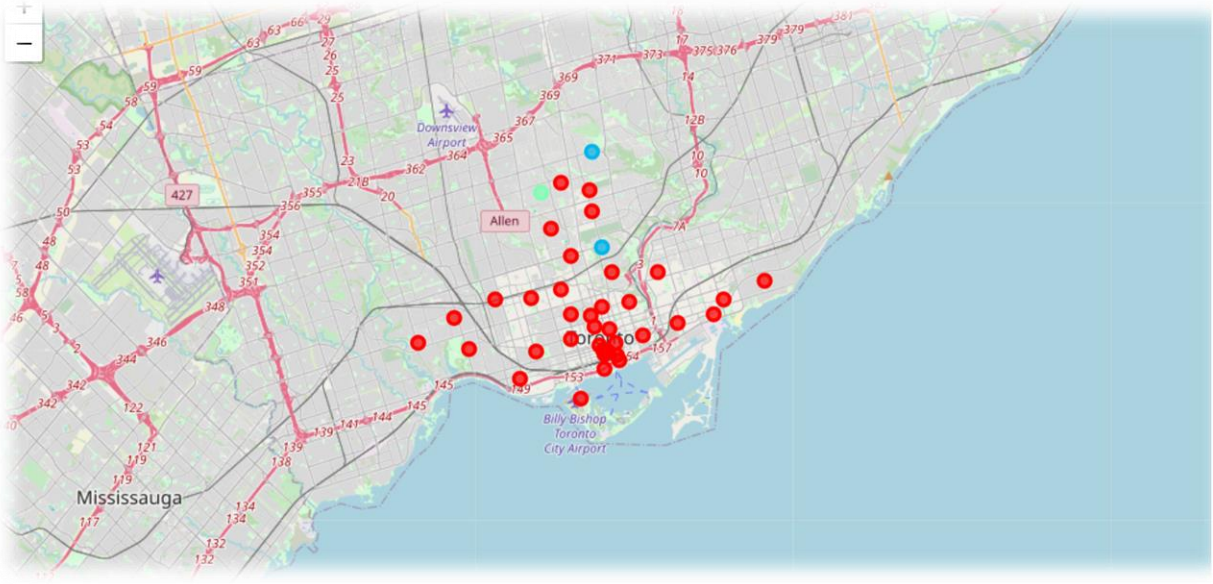
We use a k-cluster value of 5 to split the neighborhoods into 5 different clusters based on the similarity they have concerning the venues they contain.

5. Results

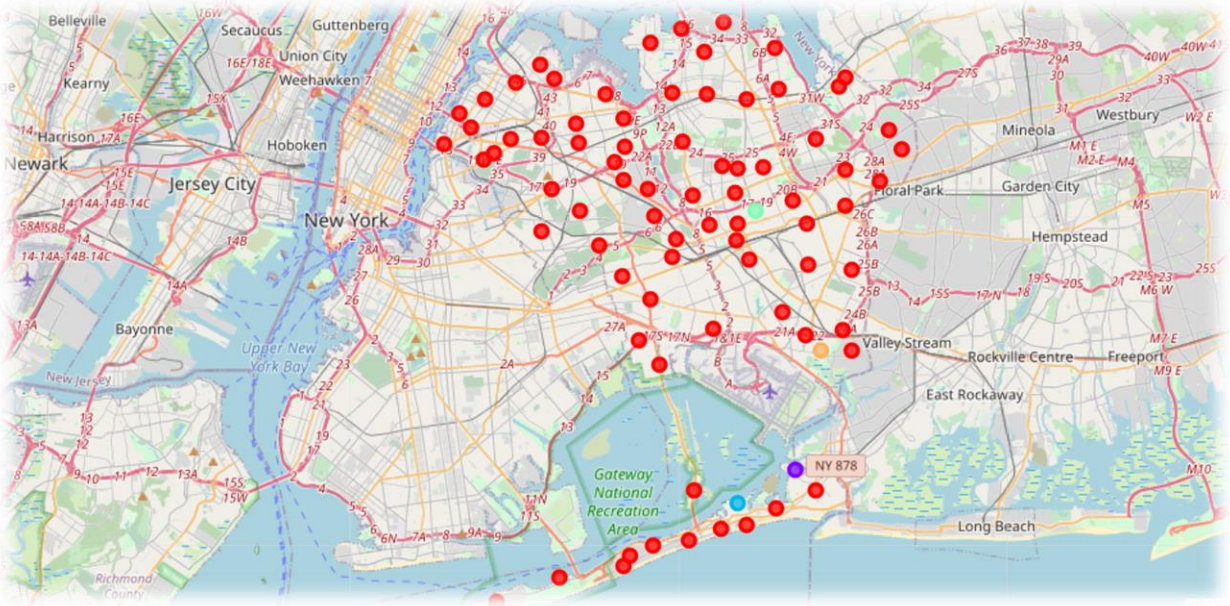
- Adding the Cluster Labels to the Venue Data
- Visualizing the resulting Clusters

A. Visualize the clusters on map

- Map of Toronto



- Map of Queens NYC



B. Examine neighborhoods segmentation

Cluster 1 (Red)

```
In [78]: cluster1 = qu_tor_merged.loc[qu_tor_merged['Cluster Labels'] == 0, qu_tor_merged.columns[[0, 1] + list(range(5, qu_tor_merged.sh
         print(cluster1.shape)
         cluster1
```

	Borough	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	Queens	Astoria	Middle Eastern Restaurant	Bar	Greek Restaurant	Indian Restaurant	Mediterranean Restaurant	Pizza Place	Seafood Restaurant	Hookah Bar	Bakery	Food Truck
1	Queens	Woodside	Grocery Store	Thai Restaurant	Bakery	Latin American Restaurant	Filipino Restaurant	American Restaurant	Bar	Pub	Pizza Place	Donut Shop
2	Queens	Jackson Heights	Latin American Restaurant	Peruvian Restaurant	South American Restaurant	Bakery	Mobile Phone Shop	Thai Restaurant	Mexican Restaurant	Supermarket	Diner	Kids Store
3	Queens	Elmhurst	Thai Restaurant	Mexican Restaurant	Chinese Restaurant	South American Restaurant	Vietnamese Restaurant	Bubble Tea Shop	Argentinian Restaurant	Snack Place	Colombian Restaurant	Malay Restaurant
4	Queens	Howard Beach	Bagel Shop	Italian Restaurant	Bank	Fast Food Restaurant	Sandwich Place	Pharmacy	Diner	Donut Shop	Supermarket	Bar

Cluster 2 (Violet)

```
qu_tor_merged.loc[qu_tor_merged['Cluster Labels'] == 1, qu_tor_merged.columns[[0, 1] + list(range(5, qu_tor_merged.shape[1]))]]
```

	Borough	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
79	Queens	Bayswater	Playground	Women's Store	Eastern European Restaurant	Distribution Center	Dive Bar	Dog Run	Doner Restaurant	Donut Shop	Dosa Place	Dry Cleaner

Cluster 3 (Light Blue "Indigo")

```
qu_tor_merged.loc[qu_tor_merged['Cluster Labels'] == 2, qu_tor_merged.columns[[0, 1] + list(range(5, qu_tor_merged.shape[1]))]]
```

	Borough	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
63	Queens	Somerville	Park	Women's Store	Dumpling Restaurant	Distribution Center	Dive Bar	Dog Run	Doner Restaurant	Donut Shop	Dosa Place	Dry Cleaner
85	Central Toronto	Lawrence Park	Park	Bus Line	Swim School	Women's Store	Dumpling Restaurant	Dog Run	Doner Restaurant	Donut Shop	Dosa Place	Dry Cleaner
89	Central Toronto	Moore Park, Summerhill East	Park	Trail	Women's Store	Diner	Distribution Center	Dive Bar	Dog Run	Doner Restaurant	Donut Shop	Dosa Place

Cluster 4 (Light Green)

```
qu_tor_merged.loc[qu_tor_merged['Cluster Labels'] == 3, qu_tor_merged.columns[[0, 1] + list(range(5, qu_tor_merged.shape[1]))]]
```

	Borough	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
54	Queens	Jamaica Estates	Antique Shop	Intersection	Discount Store	Dive Bar	Dog Run	Doner Restaurant	Donut Shop	Dosa Place	Dry Cleaner	Dumpling Restaurant
103	Central Toronto	Roselawn	Garden	Music Venue	Women's Store	Dumpling Restaurant	Dive Bar	Dog Run	Doner Restaurant	Donut Shop	Dosa Place	Dry Cleaner

Cluster 5 (Orange)

```
qu_tor_merged.loc[qu_tor_merged['Cluster Labels'] == 4, qu_tor_merged.columns[[0, 1] + list(range(5, qu_tor_merged.shape[1]))]]
```

	Borough	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
64	Queens	Brookville	Deli / Bodega	Women's Store	Eastern European Restaurant	Dive Bar	Dog Run	Doner Restaurant	Donut Shop	Dosa Place	Dry Cleaner	Dumpling Restaurant

6. Observations

- Cluster 1: (**Red**)
  - 113 neighborhoods
  - Most of the neighborhoods fall into Cluster 1 where the Airport, mostly business areas and Food and Drink Places like restaurants with different cuisines, Café, Supermarkets etc.
- Cluster 2: (**Violet**)
  - 1 neighborhood
  - **Playground**
- Cluster 3: (Light Blue "**Indigo**")
  - 3 neighborhoods
  - Cluster of **Parks** and Dog Run
- Cluster 4: (**Light Green**)
  - 2 neighborhoods
  - **Garden**
- Cluster 5: (**Orange**)
  - 1 neighborhood
  - food and store

7. Discussion

The intent with which analysis was carried out was to find out similar neighborhoods for a person relocating in other neighborhood Toronto city or Queens NYC.

As we analyze the results section, we can analyze the clusters and see similar neighborhoods in different parts of the city. For example, if we compare the different neighborhoods clustered in cluster 3.

Cluster 3 (Light Blue "Indigo")

```
qu_tor_merged.loc[qu_tor_merged['Cluster Labels'] == 2, qu_tor_merged.columns[[0, 1] + list(range(5, qu_tor_merged.shape[1]))]]
```

	Borough	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
63	Queens	Somerville	Park	Women's Store	Dumpling Restaurant	Distribution Center	Dive Bar	Dog Run	Doner Restaurant	Donut Shop	Dosa Place	Dry Cleaner
85	Central Toronto	Lawrence Park	Park	Bus Line	Swim School	Women's Store	Dumpling Restaurant	Dog Run	Doner Restaurant	Donut Shop	Dosa Place	Dry Cleaner
89	Central Toronto	Moore Park, Summerhill East	Park	Trail	Women's Store	Diner	Distribution Center	Dive Bar	Dog Run	Doner Restaurant	Donut Shop	Dosa Place

As seen in the table above, if a person wished to move from Queens to Toronto. If a person's current location were in the Neighborhood of Somerville in Queens, which has venues like Parks, Dog Runs, stores and Restaurants nearby, the person, would like to relocate to a neighborhood like Lawrence Park or Moore Park, Summerhill East in Central Toronto which also has venues like Parks, Dog Runs and Restaurants. This is just one example of how our data analysis can help people relocate from one city to another which similar to their current localities or other neighborhood in the same city.

8. Conclusion

In a fast-moving world, there are many real-life problems or scenarios where data can be used to find solutions to those problems. Like seen in the example above, data was used to cluster neighborhoods in Toronto and Queens NYC based on the most common venues in those neighborhoods. Similarly, data can also be used to solve other problems, which most people face in metropolitan cities.

9. References

- CSV for Coordinate data: http://cocl.us/Geospatial_data
- Foursquare API
- Notebooks for the course “IBM Applied Data Science Capstone” Coursera
- Wikipedia content: https://en.wikipedia.org/wiki/Toronto