Finding Similar Neighbourhoods in city of Toronto and New York

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 neighborhood in the city of Toronto we are going to use **k-means** machine learning algorithm to cluster the neighborhoods and **Foursquare** location data to explore a particular neighborhood to solve this problem easily.

Why comparing the two cities is valuable?

- Generally, people tend to shift from one place to another. Therefore, finding places that are similar to earlier place is very crucial.
- Also comparing two cities on the basis of venues or amenities present in their respective neighbourhoods helps us to find similarity score or index between two cities. Which helps us to quantify our statement such as : Is **New York** City more like **Toronto** or Paris or some other multicultural city?
- It also helps us in answering the query such as : In a city of your choice, if someone is looking to open a restaurant, where would you recommend that they open it?

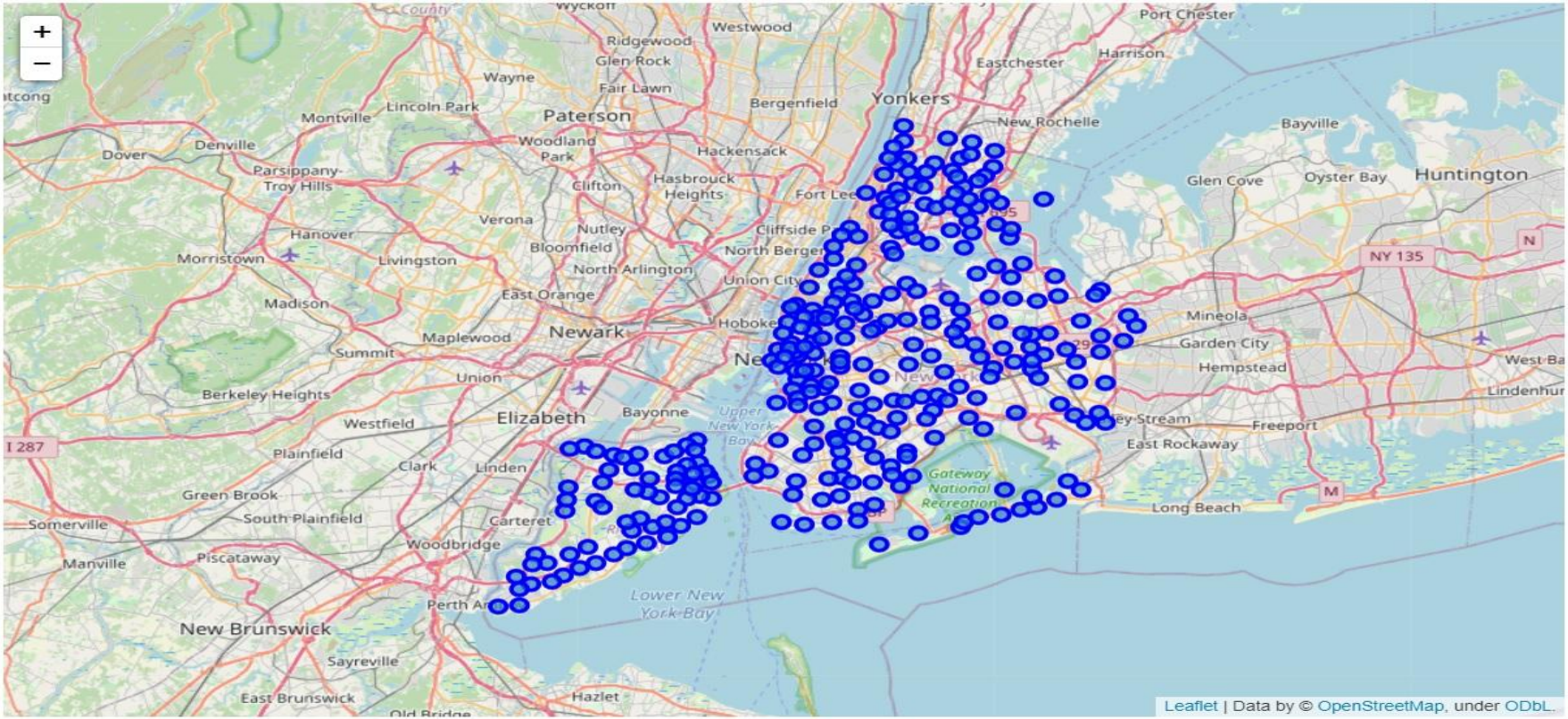
Data acquisition

- **New York** City Data containing **5 boroughs** and **306 neighborhoods** with latitude and longitude values scraped from https://rb.gy/8vwmlf
- **Toronto** City Data containing **10 boroughs** and **217 neighborhoods** scraped from https://rb.gy/ccpj06
- **Geospatial_Coordinates.csv** containing **latitude and longitude** values for neighborhoods belonging to **Toronto** present at https://rb.gy/l2ufq3
- **Foursquare Location Data** to explore most common venues in each neighborhood

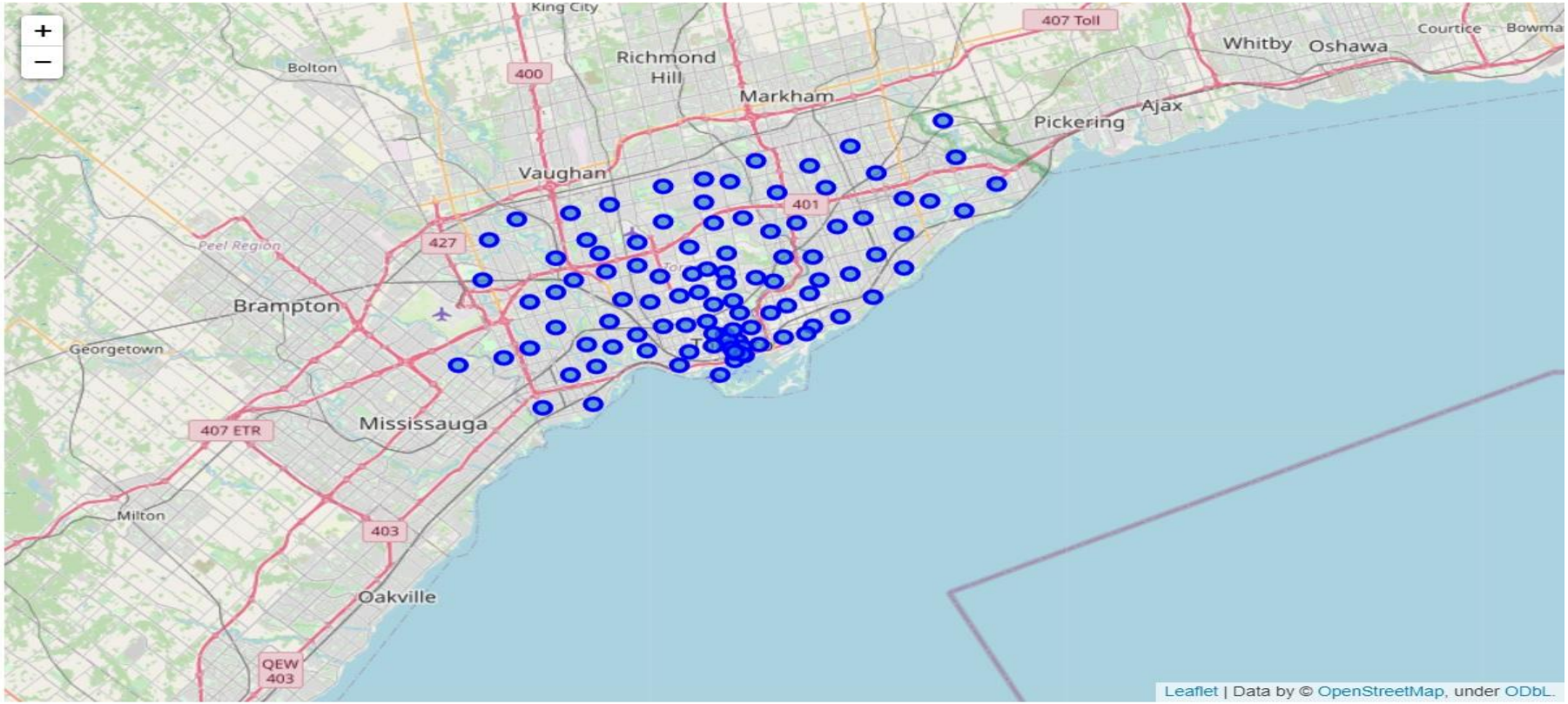
Data Cleaning

- Only processing the cells that have an **assigned borough**. Ignore cells with a borough that is **Not assigned**
- Merging of dataframes to add **latitude and longitude** values for neighborhoods in **Toronto**
- Appending dataframe and dropping any rows containing null values
- Raw dataset contains **409 rows** and **5 columns**
- Appending new columns and transforming dataframe as required
- Final dataframe consists of **408 rows** and **16 columns**

Map of New York city with neighbourhoods superimposed on top



Map of Toronto with neighborhoods superimposed on top



Methodology

- First we collected the data regarding every neighborhood in the city plus also collected their respective latitude and longitude values ,so as to explore those neighborhoods 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 analyze each neighborhood, then to group rows by neighborhood and to find top **10 venues** pertaining to each neighborhood 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 neighborhoods.
- Then we are going to visualize neighborhoods on the basis of clusters they are assigned to.

Foursquare Location Data

- Here we are using **Foursquare Location Data** to explore the neighbourhoods and to get top venues near them , on the basis of which entire clustering process will be performed.
- So to leverage the facility of **Foursquare Location Data** we have to use **Foursquare API** to send request to the server with our **Foursquare credentials**.

Example: url that is used to send request

```
'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'
```

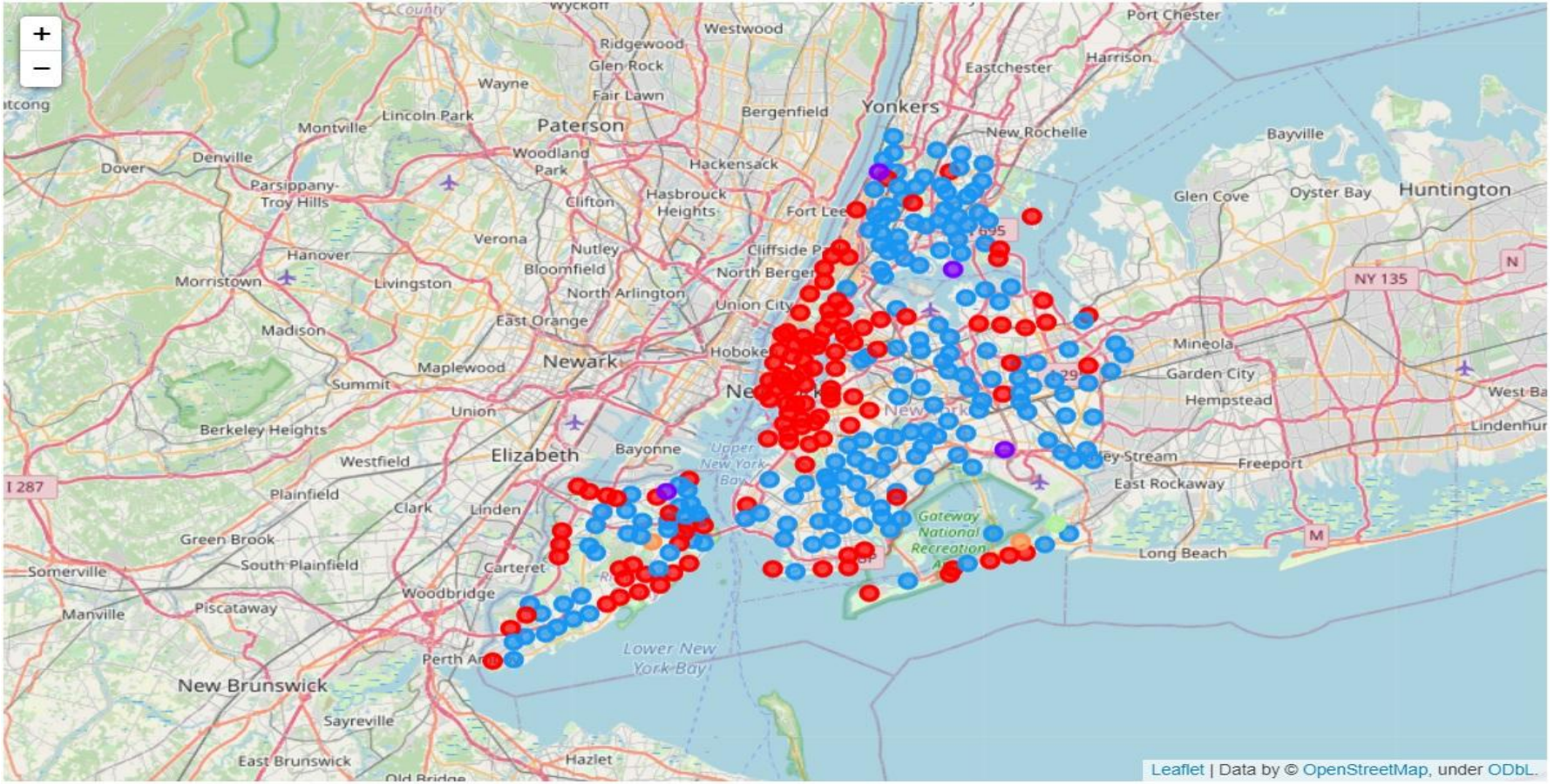
Dataset that is used for clustering

	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	Other Nightlife
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	0.0
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	0.0
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	0.0
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	0.0
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	0.0
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	0.0
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	0.0
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	0.0
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	0.0
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	0.0
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	0.0
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	0.0
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	0.0
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	0.0
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	0.0
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	0.0
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	0.0
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	0.0
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	0.0
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	0.0

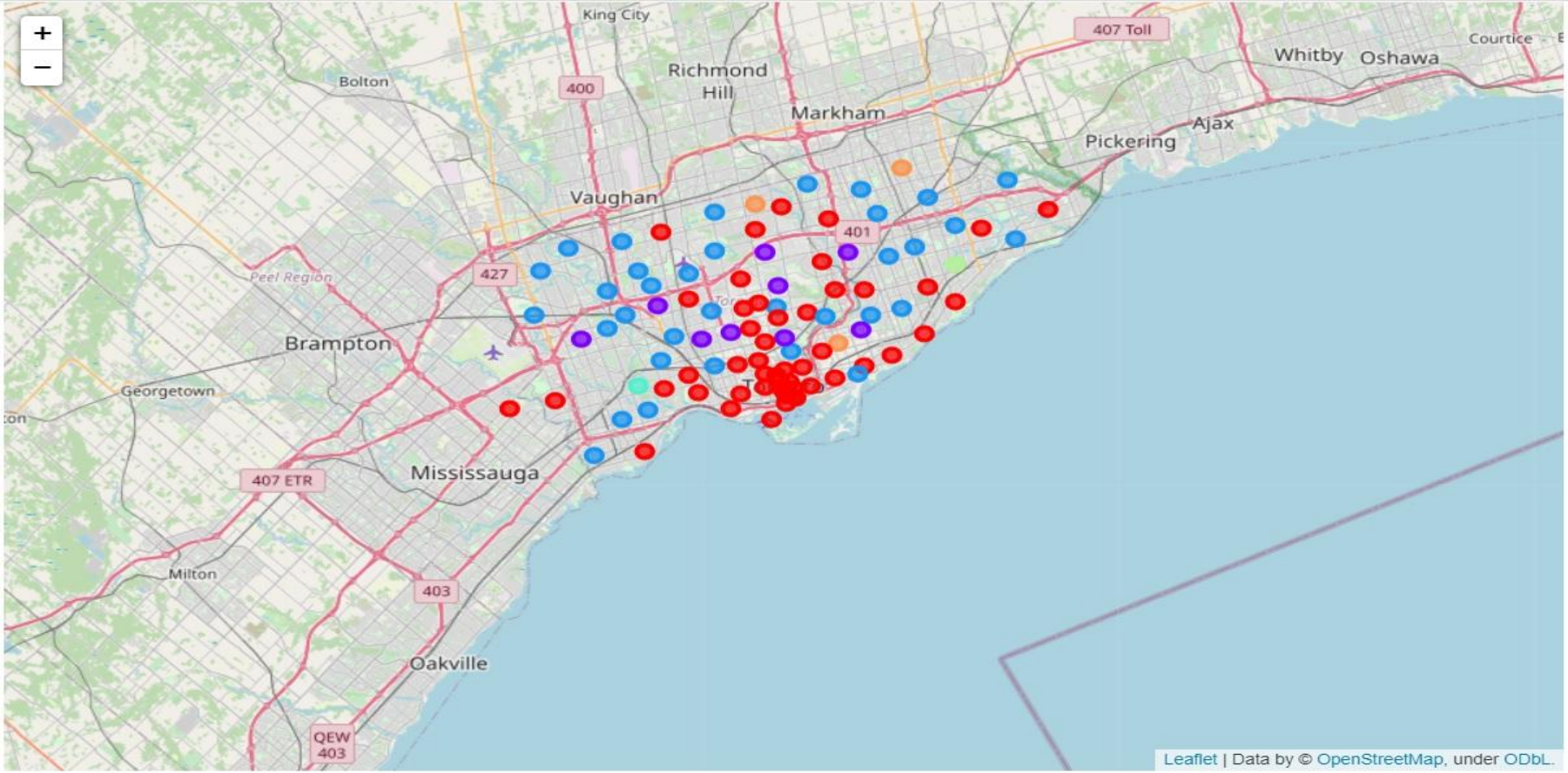
Final dataset consisting neighbourhoods with cluster labels and top 10 most common venues

	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
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	Manhattan	Marble Hill	40.876551	-73.910660	NaN	0	Coffee Shop	Sandwich Place	Gym	Discount Store	Supplement Shop	Donut Shop
7	Bronx	Woodlawn	40.898273	-73.867315	NaN	2	Pizza Place	Deli / Bodega	Food & Drink Shop	Pub	Playground	Italian Restaurant
8	Bronx	Norwood	40.877224	-73.879391	NaN	2	Pizza Place	Bank	Park	Pharmacy	Burger Joint	Grocery Store
9	Bronx	Williamsbridge	40.881039	-73.857446	NaN	0	Nightclub	Caribbean Restaurant	Dance Studio	Bar	Soup Place	Nail Salon
10	Bronx	Baychester	40.866858	-73.835798	NaN	2	Donut Shop	Pet Store	Mexican Restaurant	Cosmetics Shop	Sandwich Place	Discount Store
11	Bronx	Pelham Parkway	40.857413	-73.854756	NaN	2	Italian Restaurant	Pizza Place	Bus Station	Food	Bank	Sandwich Place

Visualizing clusters in New York City



Visualizing clusters in Toronto City



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.

Conclusion

- Finally we segmented and clustered all the neighbourhoods belonging to the two cities.
- We used **k-means clustering algorithm** , with **k = 6** ,i.e. all the neighbourhoods are clustered into six clusters on the basis of most common venues in the neighbourhoods.
- Now we can select the **neighbourhood** that is similar to our current neighbourhood easily.
- Final **decision on optimal neighborhood** will be made by end user based on **specific characteristics of neighborhoods** and locations in every recommended neighborhood 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 neighborhood etc.**