Neighborhoods businesses Comparison Using K-Means Clustering, Case of New York vs Toronto

1 Introduction

1.1 Background

Cities around the World contain certain number of neighborhoods formed of venues naturally distributed in many different way. The historic need to identify city's driver factors such as financial incentives, quality of urbanisation, social and food diversities, crime level, housing, education, health, etc, often lead to the exploration of similarity or dissimilarity between cities.

Comparing venues within neighborhoods in cities helps the orientation of decision making in terms of business investment, immigration or relocation, job hunting and much more. The analysis in this work is applied the two most populous and diverse cities of the United State and Canada: New York vs Toronto.

1.2 Problem Definition

The Neighborhoods businesses comparison of two or more different cities is a type of unsupervised classification where venues have to be segmented and grouped into a certain number of dissimilar and non-overlapping clusters. These clusters should contain similar venues of common characteristics, without any internal structure or label. K-Means algorithms is one of the most popular tools of segmentation of unsupervised data that will be used here coupled with foursquare API to make essentially venues calls to retrieve needed information.

1.3 Stakeholders

This work is particularly targeting business people who need to invest by giving them insight views of market in neighborhoods. This work concern also at large job seekers to explore possibilities that offer one city over the other. The results of this work should also advice on people movement, immigration and relocation, because positive numbers on neighborhoods businesses in a city, will most probably influence the influx of people.

2 Data Source, Loading and Exploration

Data used in this work were collected from open sources on the internet:

- For New York, a total of 306 neighborhoods with a total of 5 boroughs in neighborhood as well as the latitude and the longitude of each neighborhood fount on the web under the link: https://geo.nyu.edu/catalog/nyu_2451_34572
- For Toronto, City neighborhoods were found on Wikipedia website under the following link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
  This link provided postal codes of venues that could be transform into needed data by using pandas library function. In case the geocoder package couldn't work, the geographical coordinates of postal code can be found under the link:
  http://cocl.us/Geospatial_data

These data will be cleaned and formatted accordingly to adequately feed the k-Means algorithm for convergent clustering.

2.1 Toronto Data

Raw data on Toronto city used here was scraped on Wikipedia.org as lxml file. A dataframe was builded after cleaning and arranging data….

	PostalCode	Borough
0	M1B	Scarborough
1	M1C	Scarborough
2	M1E	Scarborough
3	M1G	Scarborough
4	M1H	Scarborough
5	M1J	Scarborough
6	M1K	Scarborough
7	M1L	Scarborough
8	M1M	Scarborough
9	M1N	Scarborough
10	M1P	Scarborough
11	M1R	Scarborough

Table 1 Toronto Neighborhoods List Dataframe illustration

From these information in Table 1, we extract geographical coordinates of each neighborhood in Toronto city and merge all in one dataframe called df_toronto illustrated in **Table 2**.

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
6	M1K	Scarborough	Kennedy Park, Ionview, East Birchmount Park	43.727929	-79.262029
7	M1L	Scarborough	Golden Mile, Clairlea, Oakridge	43.711112	-79.284577
8	M1M	Scarborough	Cliffside, Cliffcrest, Scarborough Village West	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff, Cliffside West	43.692657	-79.264848
10	M1P	Scarborough	Dorset Park, Wexford Heights, Scarborough Town...	43.757410	-79.273304
11	M1R	Scarborough	Wexford, Maryvale	43.750072	-79.295849

Table 2 Toronto Neighborhoods Geospatial Coordinates

Geospatial coordinates can be mapped using visualisation libraries such as Folium to pin point location of each neighborhood by markers. Folium require the latitude and longitude of a location to represent it on the map by a marker. **Figure 1** show by blue makers 103 neighborhoods of the city of Toronto. The general observation is that the natural distribution of neighborhoods are spread as converging to a focal point in Central Toronto.

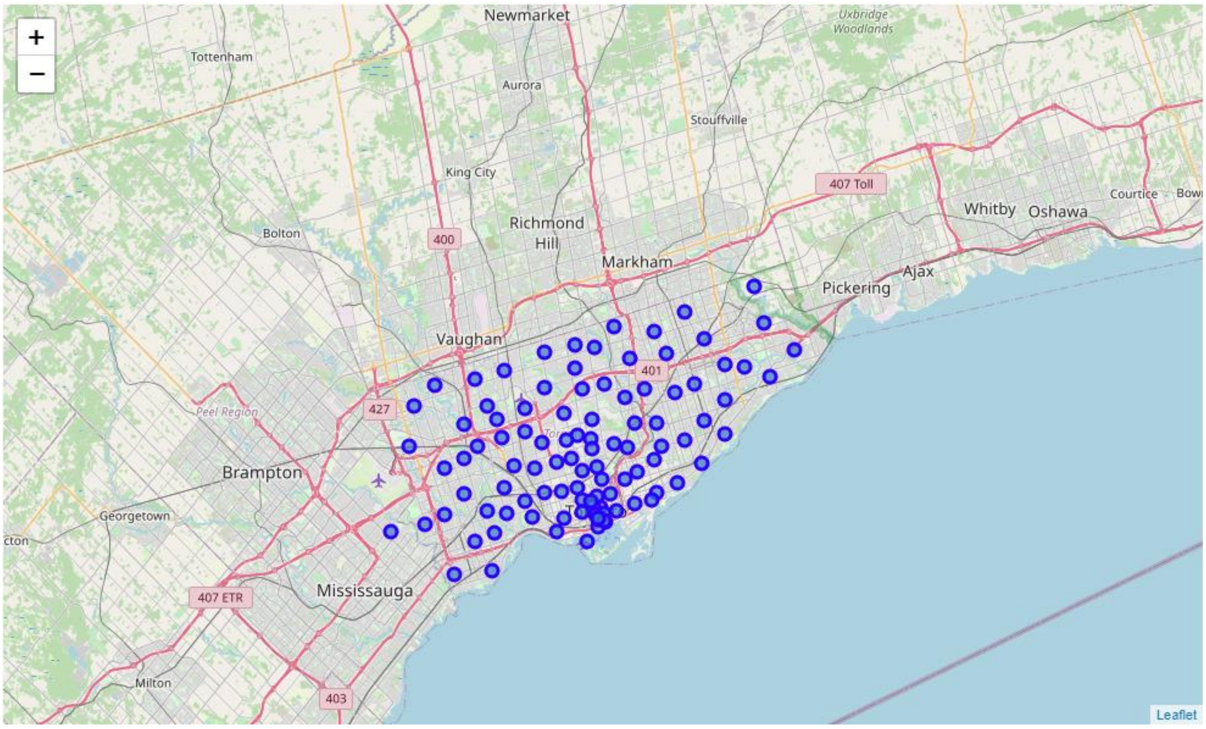

Figure 1 Toronto Neighborhoods marked in blue.

2.2 New York Data

Data of New York neighborhoods were downloaded from the web though the dataset link:

https://geo.nyu.edu/catalog/nyu_2451_34572

Json file were extracted and saved as newyork_data with the first 12 rows illustrated in **Table 3**. This dataframe has 5 boroughs and 306 neighborhoods with geospatial data.

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585
5	Bronx	Kingsbridge	40.881687	-73.902818
6	Manhattan	Marble Hill	40.876551	-73.910660
7	Bronx	Woodlawn	40.898273	-73.867315
8	Bronx	Norwood	40.877224	-73.879391
9	Bronx	Williamsbridge	40.881039	-73.857446
10	Bronx	Baychester	40.866858	-73.835798
11	Bronx	Pelham Parkway	40.857413	-73.854756

Table 3 New York Neighborhood data illustration

Using Folium library, New York neighborhoods were marked in yellow as shown in **Figure 2**.

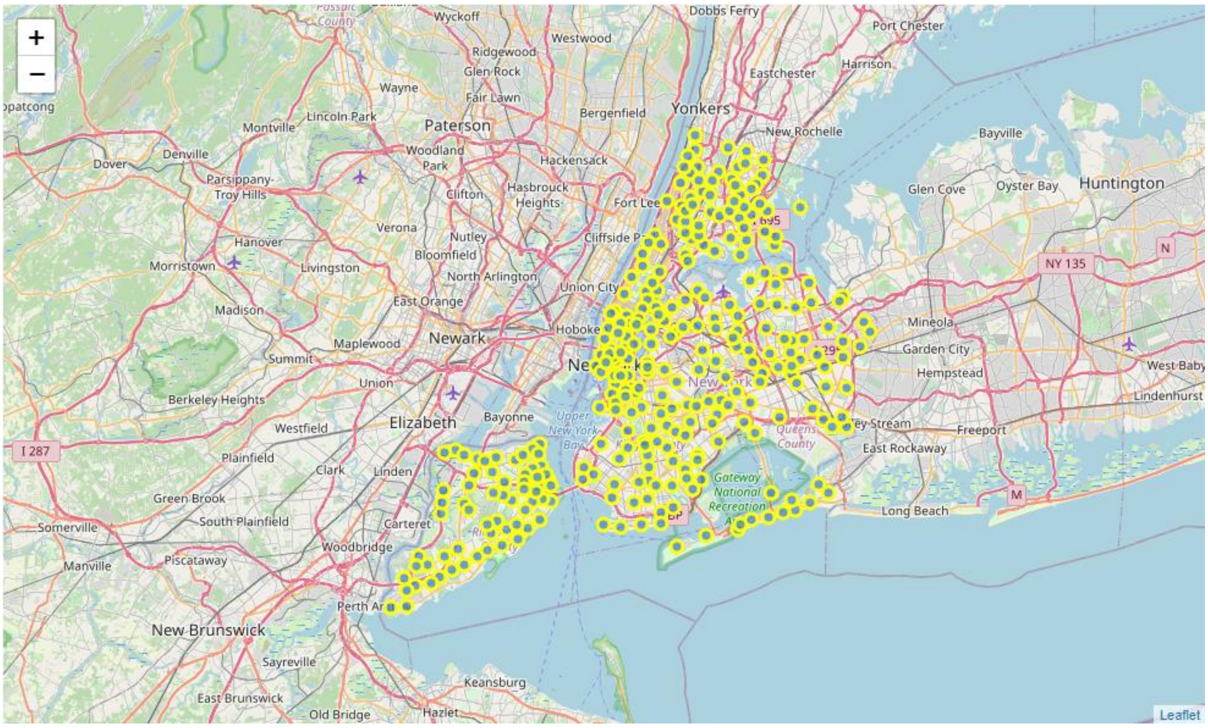

Figure 2 306 Neighborhoods of New York City

3 Methodology

Having neighborhoods of Toronto and New York, to find venues in these neighborhoods, Foursquare API is used. With explored, k-Means is used to segment venues into clusters. By comparing similar clusters of the two city, we will be able to conclude on the level of business activities into the city; this is not to confuse with the number of enterprises in each city.

3.1 k-Means Method

If k is given, the K-means algorithm can be executed in the following steps:

- Partition of objects into k non-empty subsets
- Identifying the cluster centroids (mean point) of the current partition.
- Assigning each point to a specific cluster
- Compute the distances from each point and allot points to the cluster where the distance from the centroid is minimum.
- After re-allotting the points, find the centroid of the new cluster formed.

Mathematical Formulation for K-means Algorithm:

D = {$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_m$} a data set of m records

$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ a each record is an n-dimensional vector

$$C_j = Cluster(X_i) = arg_i min \left\| X_i - \mu_j \right\|^2$$

$$Distortion = \sum_{i=1}^{m} \left(X_i - C_j\right)^2 = \sum_{j=1}^{k} \sum_{i \in (\mu_j)} \left(X_i - \mu_j\right)^2$$

The cluster centres are those that minimize the distortion. For any k clusters, the value of k should be such that even if we increase the value of k from after several levels of clustering the distortion remains constant. The achieved point is called the "Elbow" and the procedure is called Elbow Method.

Method:

For both New York and Toronto neighborhoods extracted data are cleaned and arranged using Pandas into dataframes: ***df_Newyork*** and ***df_toronto*** respectively. Foursquare API calls were passed to return venues in each neighborhood as *Json file* data. These venues data in neighborhoods of New York and Toronto cities were cleaned and rearranged as newyork_venues and Toronto_venues respectively, illustrated by **Tables 4A and 4B**.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern, Rouge	43.806686	-79.194353	Toronto Pan Am Sports Centre	43.790623	-79.193869	Athletics & Sports
1	Malvern, Rouge	43.806686	-79.194353	African Rainforest Pavilion	43.817725	-79.183433	Zoo Exhibit
2	Malvern, Rouge	43.806686	-79.194353	Toronto Zoo	43.820582	-79.181551	Zoo
3	Malvern, Rouge	43.806686	-79.194353	Polar Bear Exhibit	43.823372	-79.185145	Zoo
4	Malvern, Rouge	43.806686	-79.194353	Morningside Park	43.786546	-79.205322	Park
5	Malvern, Rouge	43.806686	-79.194353	Gorilla Exhibit	43.819080	-79.184235	Zoo Exhibit
6	Malvern, Rouge	43.806686	-79.194353	Lamanna's Bakery, Cafe & Fine Foods	43.797971	-79.148432	Bakery
7	Malvern, Rouge	43.806686	-79.194353	Orangutan Exhibit	43.818413	-79.182548	Zoo Exhibit
8	Malvern, Rouge	43.806686	-79.194353	Australasia Pavillion	43.822563	-79.183286	Zoo Exhibit
9	Malvern, Rouge	43.806686	-79.194353	Mona's Roti	43.791613	-79.251015	Caribbean Restaurant

Table 4A Toronto Venues Dataframe illustration

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
4	Wakefield	40.894705	-73.847201	Shell	40.894187	-73.845862	Gas Station
5	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop
6	Wakefield	40.894705	-73.847201	Subway	40.890468	-73.849152	Sandwich Place
7	Wakefield	40.894705	-73.847201	Central Deli	40.896728	-73.844387	Deli / Bodega
8	Wakefield	40.894705	-73.847201	Louis Pizza	40.898399	-73.848810	Pizza Place
9	Wakefield	40.894705	-73.847201	Koss Quick Wash	40.891281	-73.849904	Laundromat
10	Co-op City	40.874294	-73.829939	Capri II Pizza	40.876374	-73.829940	Pizza Place
11	Co-op City	40.874294	-73.829939	Rite Aid	40.870345	-73.828302	Pharmacy

Table 4B New York Venues Dataframe illustration

4 Results and Discussion

The natural diversity of cities is always trivial. The analysis of category of venues has shown that New York neighborhoods contain at least 431 unique categories, while Toronto neighborhoods have only 233 categories. This was clearly expected due to the fact that New York has been characterized as the world's premier financial centre (Business Insider, Inc. 2014).

Prior to proceed with clustering data need to be normalised. In this case dataframes of New York and Toronto passed by the grouping of rows by neighborhood and by calculating the mean of frequency of occurrence of each category. The outputs are shown by **Table 5A** and **5B**

	Neighborhood	Zoo Exhibit	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	American Restaurant	Aquarium
0	Agincourt	0.04	0.00	0.0	0.00	0.00	0.00	0.00
1	Alderwood, Long Branch	0.00	0.00	0.0	0.00	0.00	0.00	0.00
2	Bathurst Manor, Wilson Heights, Downsview North	0.00	0.00	0.0	0.01	0.00	0.00	0.00
3	Bayview Village	0.00	0.00	0.0	0.00	0.00	0.00	0.00
4	Bedford Park, Lawrence Manor East	0.00	0.01	0.0	0.00	0.00	0.00	0.00
5	Berczy Park	0.00	0.00	0.0	0.00	0.00	0.01	0.01
6	Birch Cliff, Cliffside West	0.00	0.01	0.0	0.00	0.00	0.01	0.00
7	Brockton, Parkdale Village, Exhibition Place	0.00	0.00	0.0	0.00	0.00	0.00	0.01

Table 5A Toronto Grouped Dataframe

	Neighborhood	Zoo Exhibit	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant
0	Allerton	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
1	Annadale	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
2	Arden Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
3	Arlington	0.0	0.0	0.0	0.0	0.0	0.0	0.200000
4	Arrochar	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
5	Arverne	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
6	Astoria	0.0	0.0	0.0	0.0	0.0	0.0	0.010000
7	Astoria Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.000000

Table 5B New York Grouped Dataframe

After Classification using k-Means clustering, considering k = 5, the resulting clusters are visualised using the Folium package by **Figure 3A** and **3B** respectively for New York City and Toronto City.

Figure 3A New York City Clusters

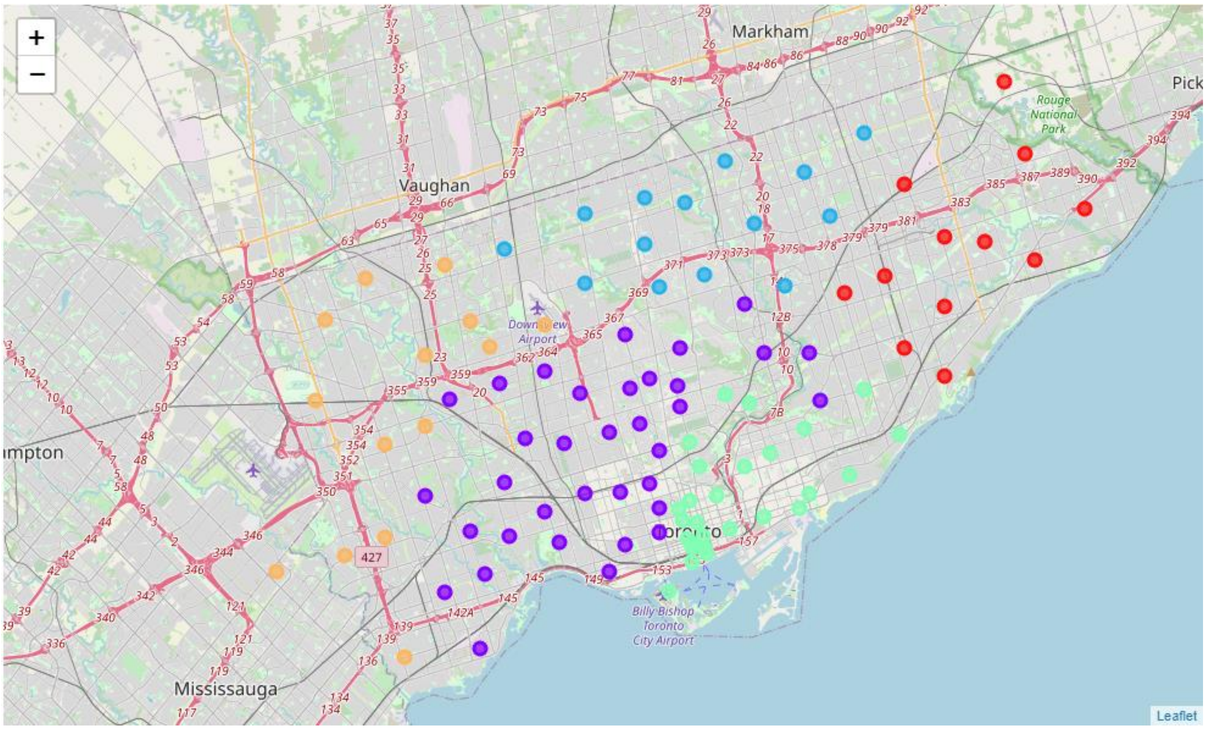

Figure 3B Toronto City Clusters

K-Means clustering method applied, in the same conditions, to the venues explored from neighborhoods of New York and Toronto into 10 most common clusters has shown different level of convergence. New York clustering converged at a third iteration while Toronto iterations went over a fifth rounds. It trended that New York has far more diverse groups of business that Toronto.

The results could be interpreted that for a qualified job seekers New York could be a better option because of multiple opportunities that Toronto, assuming that all political and public administrative influences were neglected. But for small and medium enterprises investors, the method shows that Toronto could be a better option, considering that competition level in Toronto should be less aggressive than in New York City.

5 Conclusion

The analysis conducted here was conclusive assuming that all other factors: immigration policies, political, public administration etc, have no influence. This general comparison of business between New York and Toronto indicated the point was not to look at a certain "size" or "graduation" but to explore the preponderance of groups of activities on either sides.

6 References

1. IBM Data Science Professional Certificate, Capstone Project Labs 2020
2. "Top 8 Cities by GDP: China vs. the U.S." Business Insider, Inc. July 31, 2011. Retrieved July 29, 2014.