15th of June, 2020

IBM Capstone Report

Comparative analysis of New York City, Toronto and Brussels

A. INTRODUCTION

New York City, Toronto and Brussels are three representative financial, economical and administrative capitals in their own geographic zone, respectively, the US, the Canada, and the European Union.

Despite some obvious differences due to their respective anchor areas, they share commun features such as multiculturality, cultural and touristic attractiveness, population density, economic indicators, etc ... New York City and Toronto are highly intensive financial places while Brussels hosts the European Community administration and the operational center of many international companies.

The following table illustrates some rough indicators of these 3 cities.

	Area (squared Km)	Population (million habitants)	Density (habitants per squared Km)	GDP per habitant (k€/habitant)
New York City	784	8.399	7101	89222
Toronto	630	2.731	4336	37771
Brussels	161	1.223	7582	65000

As exemplified by these figures, the cities exhibit high population density and high GDP/habitant figures. In this context, the business question we would like to answer in this exercice is the following. In the previous exercice, a separate caracterisation of the cities of New York and Toronto has been made on the basis of data gathered from the Four Square APIs, in practice, data about the most popular and common venues, irrespective of any categorization constraints.

It would be useful for commuters, for HR decisioners in multinational companies, and many other « movers » to be able to identify the main differences between the two and to benchmark them with a European city (in this case, Brussels). This is precisely what we will try to do in this short study.

As a reminder, this clusterisation has put in relief the similarities and differences between the neighboroughs of the two cities, on a certain number of dimensions and features of the most appreciated venues of the cities.

However, the exercice has been made separately for each of the two cities which a priori have quite comparable positionings in their respective countries, as being very diverse and financial capitals in their own country. A first objective of this work will be to compare the two cities on the basis of these same dimensions.

A second step will be to extend the comparison with a European representative city. We have chosen Brussels for its multicutural dimension and its central positioning with respect to European institutions and some international companies.

The added value of this exercice is to bring light on significant similarities and differences between the three cities, and this can be useful as showcases of elements to be taken into account for mobilities between these cities (and in particular between (New York, Toronto on one side, Brussels on the other side).

B. DATA DESCRIPTION AND GENERAL ASSUMPTIONS

In order to perform the exercice, we have considered the following sources of data, which we briefly describe synthetically hereafter, with some simplifying assumptions :

- The New York list of the boroughs with associated neighboroughs, latitudes and longitudes is provided by the following URL (used in a previous exercice) : https://cocl.us/new_york_dataset. This URL provides with a JSON file

containing 306 lines. We have selected from it data pertaining to the Manhattan borough. This choice reduces the number of lines to consider in the subsequent operations (from 306 lines to **40** lines)

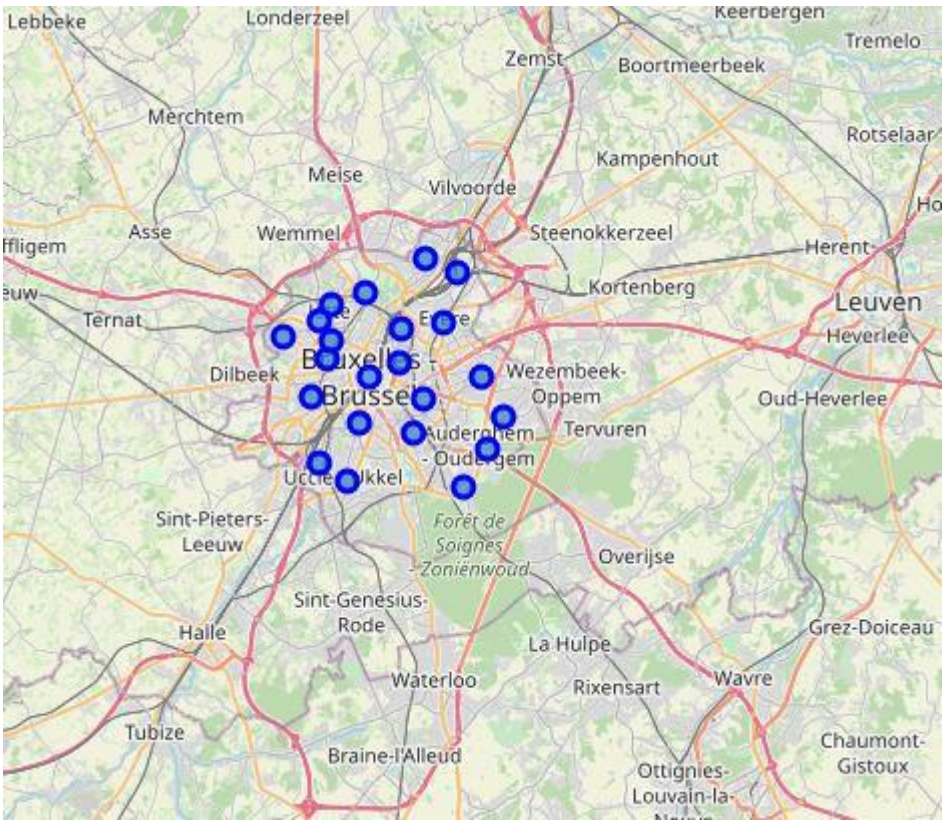
- The Brussels list has been built from a CSV file of all « cities and villages » from Belgium, provided by the National Post (at : https://raw.githubusercontent.com/jief/zipcode-belgium/master/zipcode-belgium.csv). We have filtered it on the postal codes in order to extract the Brussels zones composed of **22** neighborhoods.
- The Toronto data has been gathered from a wikipedia web page, at https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. Some scraping operations have been needed so as to come up with a dataframe in the right format (such as removing NaN values and useless lines). The result is a dataframe with **103** lines.
- The Foursquare API has been used so as to gather general information about venues of any type (we put no restriction on these), as the goal is to obtain a general characterization of the attractiveness of the cities. We used 2 parameters : a radius of one kilometer with the 50 most appreciated venues.
- Ultimately we used the Folium library to vizualize the geographic details of each city with their respective neighborhoods.
- These data have been regrouped in a single table called « Neighborhoods », containing 165 lines and whose structure looks like the following extract :

	Neighborhood	Latitude	Longitude
0	Marble Hill	40.876551	-73.910660
1	Chinatown	40.715618	-73.994279
2	Washington Heights	40.851903	-73.936900
3	Inwood	40.867684	-73.921210
4	Hamilton Heights	40.823604	-73.949688

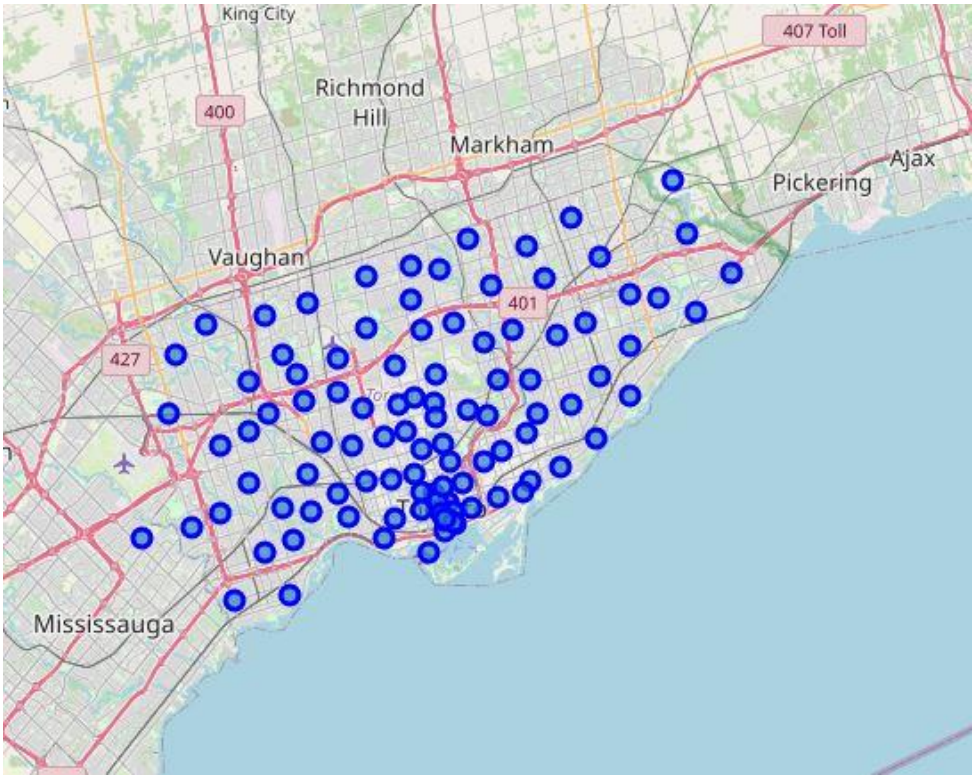
C. DATA VIZUALIZATION

The studied cities with their respective neighborhoods have been positioned on the Folium map, with the following results:

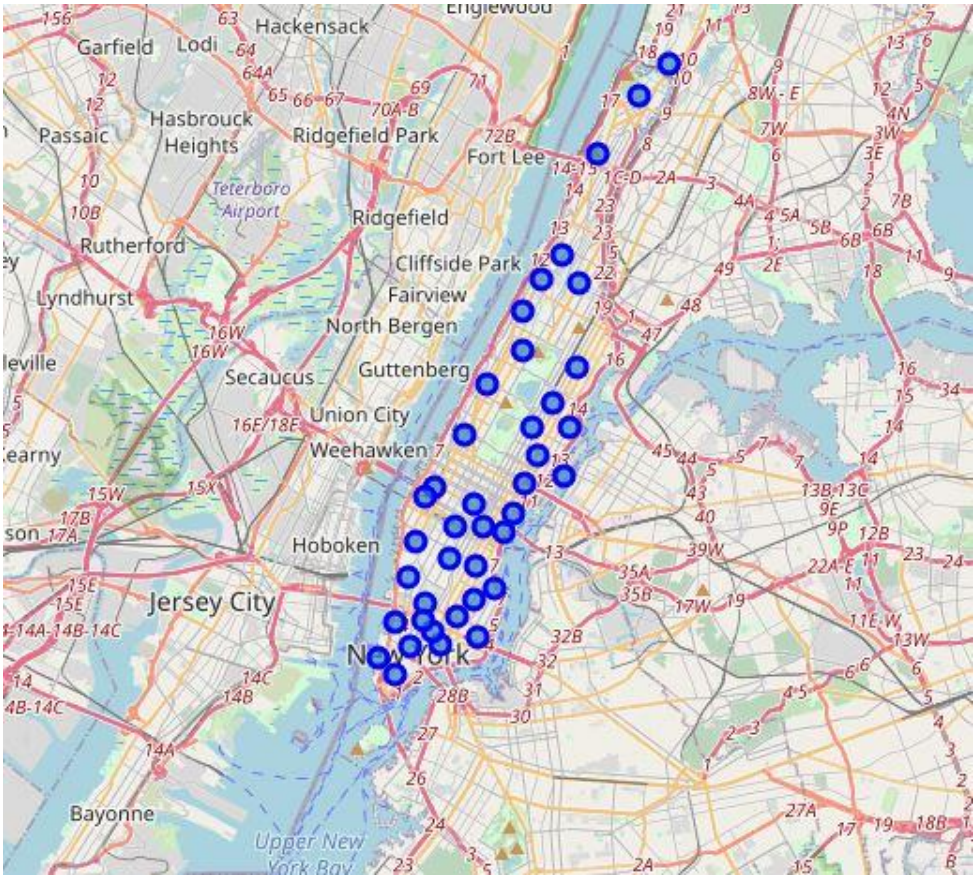
For Brussels, we see a circular and concentrated list of neighborhoods around the center of Brussels, corresponding to what is named "The Brussels periphery".



For Toronto, the spreading is different and covers a wider surface than for Brussels. Please note that the number of neighborhoods considered is much higher.



Lastly, for New York we have restricted the studied perimeter to Manhattan, as exemplified below.



-

We also positioned the 3 cities on a world map to have a kind of telescopic view of what we are talking about. See the following map.

D. METHODOLOGY

In order to perform a comparative analysis, we have inserted all neighborhoods in **one single KNN exercice**, with 2 pre-defined number of clusters : 3 and 10. The first number aims at a very global categorization, while the second allows us to fine grain the analysis « inside each city ».

Combined with the data gathered from Foursquare, we obtained the first following table with the features of all considered neighborhoods.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Agincourt	45	45	45	45	45	45
Alderwood, Long Branch	26	26	26	26	26	26
Anderlecht	50	50	50	50	50	50
Auderghem	50	50	50	50	50	50
Bathurst Manor, Wilson Heights, Downsview North	30	30	30	30	30	30
Battery Park City	50	50	50	50	50	50
Bayview Village	15	15	15	15	15	15
Bedford Park, Lawrence Manor East	43	43	43	43	43	43
Berchem-Sainte-Agathe	50	50	50	50	50	50
Berczy Park	50	50	50	50	50	50
Birch Cliff, Cliffside West	12	12	12	12	12	12

One point is worth mentioining : each neighborhood is treated equally with any other, on the basis of the precise same characterization from the one unique source of data.

On this basis, classical operations on venues data are performed, leading us to a one-hot encoding table (6406 X 419), a regrouping of data on the basis of the mean of frequency of venues per neighborhood (160 X 419), as well as the list of the top-ten most common venues per neighborhood. Please note the decrease from 165 to 160 neighborhoods, 5 bieng considered as statistically unsignificant. This last dataframe is used as the reference table for the clustering exercice.

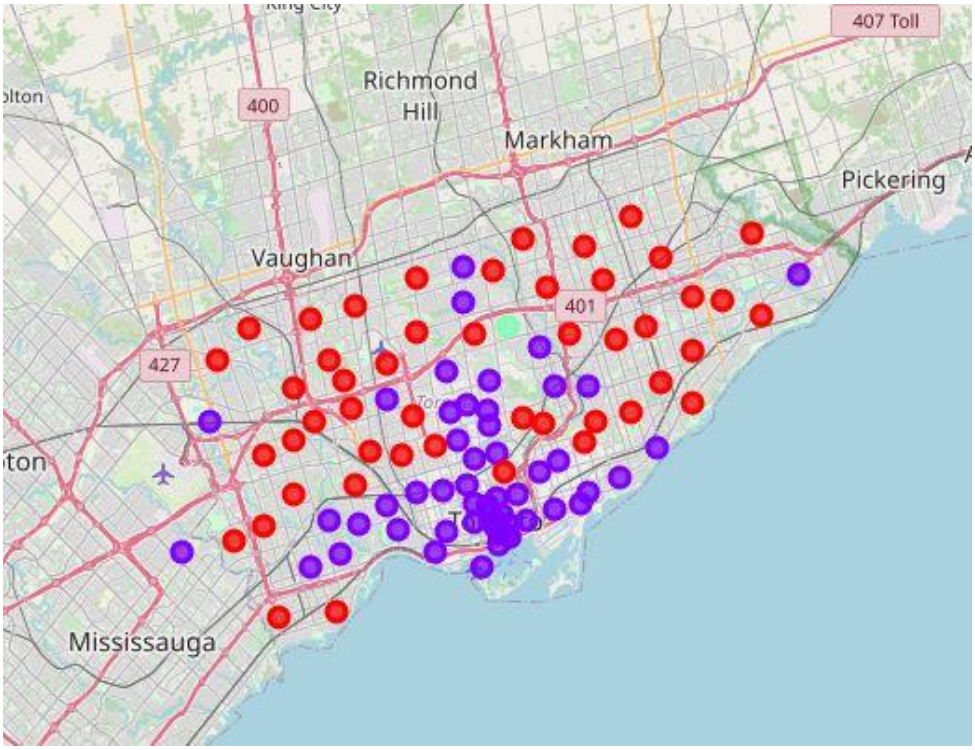
]2]:

	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	Agincourt	Chinese Restaurant	Shopping Mall	Sandwich Place	Caribbean Restaurant	Bakery	Pizza Place	Coffee Shop	Bubble Tea Shop	Bank	Latin American Restaurant
1	Alderwood, Long Branch	Discount Store	Pharmacy	Pizza Place	Convenience Store	Park	Moroccan Restaurant	Grocery Store	Athletics & Sports	Dance Studio	Shopping Mall
2	Anderlecht	Bar	Supermarket	Plaza	Bakery	Convenience Store	Snack Place	Sandwich Place	Restaurant	Greek Restaurant	Park
3	Auderghem	Italian Restaurant	Bakery	Fast Food Restaurant	French Restaurant	Bar	Belgian Restaurant	Thai Restaurant	Sushi Restaurant	Park	Middle Eastern Restaurant
4	Bathurst Manor, Wilson Heights, Downsview North	Pizza Place	Bank	Coffee Shop	Trail	Diner	Gas Station	Sandwich Place	Dog Run	Fried Chicken Joint	Sushi Restaurant

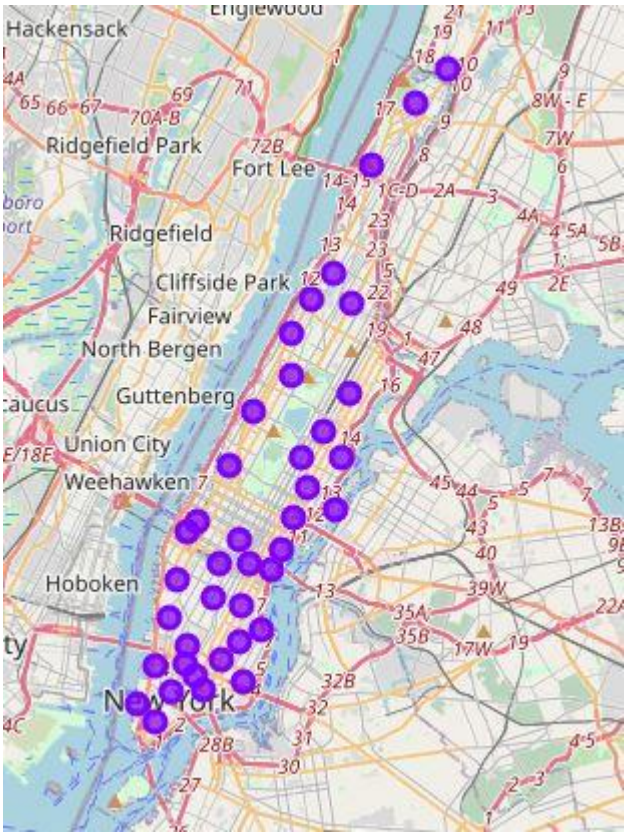
E. RESULTS

The first clustering exercice fixes the number of clusters to 3. We reproduce below the colored configuration for each city.

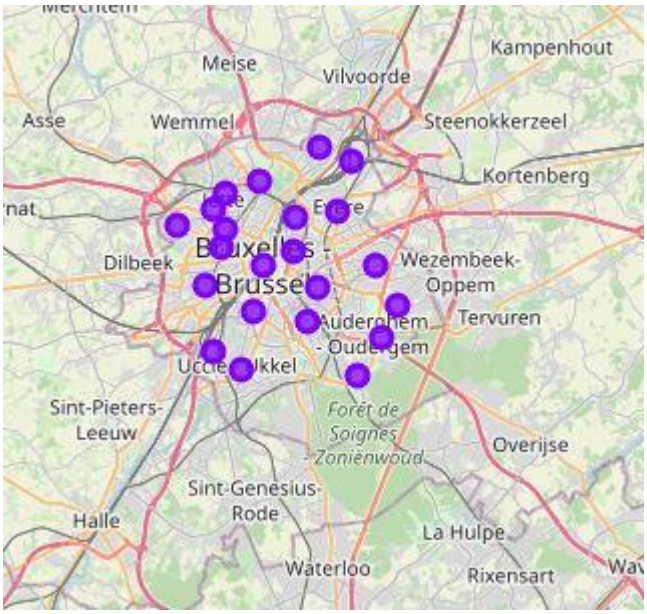
For Toronto :



For Manhattan :



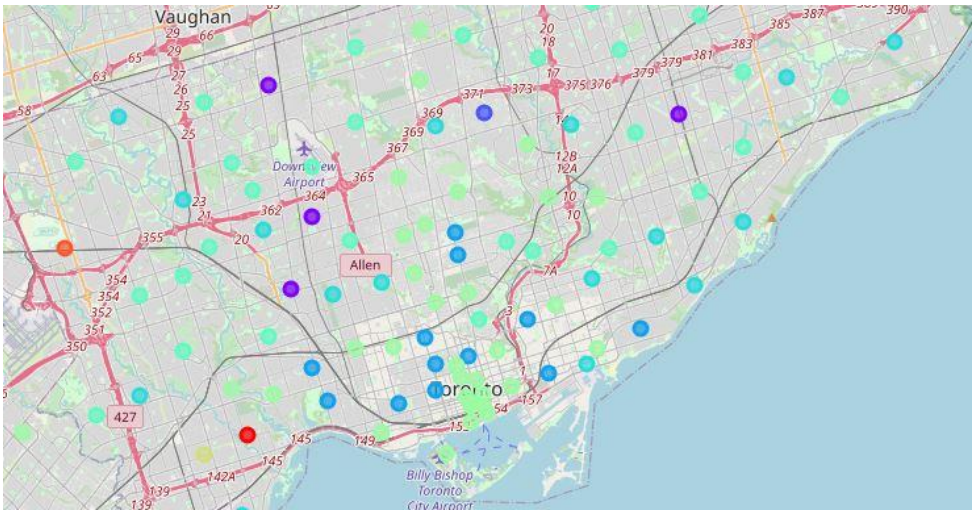
And lastly, for Brussels :



At this level of granularity, the figures show a striking similarity between Manhattan (NYC) and Brussels. Toronto is divided between a similar category and a very different one.

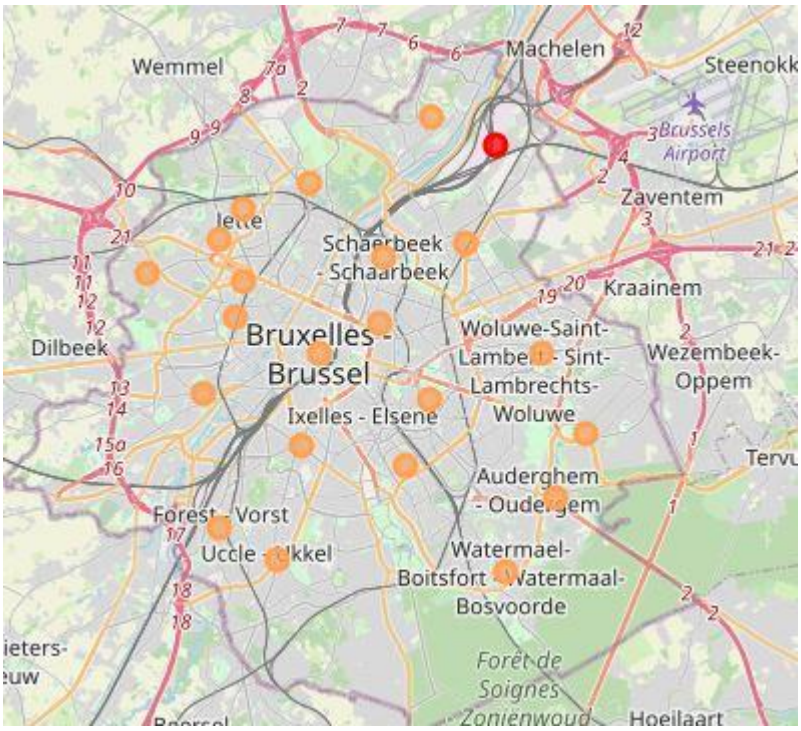
Are these insights confirmed with a more fine-grained analysis ? Let us see.

For Toronto :



The variety inside Toronto itself is popping more clearly obviously. What about the comparison with Manhattan ?

We see that the diagnosis is somehow more nuanced. Differences between Toronto and Mnahattan remain but they are superseeded by the internal heterogeneity of each city. We end-up with Brussels.



In this view 2 facts immediately pop-up : the homogeneity of Brussels as a whole and its relative dissimilarity with Toronto, even the proximity with Manhattan remains.

F. CONCLUSIONS

Clustering is a very powerful tool so as to examine the similarities and differences between use cases. In this specific use case, we saw that 3 cities with symbolic cities of the economy have stringent features which differentiate them from one another. The level of granularity chosen is also a powerful mean to refine diagnosis.