The Battle of the Neighborhoods

1. Introduction

1.1 Background

Moving from one city to another is not an easy decision to make. In the world, there is a large number of cities and all of them have something that makes them unique and different than the rest. In order to move there are several factors to consider and one of them could be how similar is the other city and their people, compared to the place we live. It can be said that if both share similar kind of places then they are similar and therefore their people have similar preferences for certain kind of venues and this implies that they have the same habits. Therefore, to have this information in hand will be helpful for a person in this situation because it will be a big contributing factor at the time to make a such important decision.

1.2 Problem

A French family residing in Paris has to make a decision about whether to move to New York or Toronto, being both cities very diverse and the financial capitals of their respective countries. In addition, the family head has received two very similar job offers from the aforementioned cities and the decision to where to move will be based in how similar or dissimilar they are compared to Paris.

1.3 Interest

This project will be of interest of any person who is in similar situation as the people described above or simply wants to explore in order to find similarities among different cities,

2. Data acquisition and cleaning

2.1 Data sources

Data was obtained from different websites:

- Toronto's neighborhoods were obtained with Pandas by scrapping Wikipedia URL: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Toronto's neighborhoods geographical locations were imported by reading a CSV file from the URL: https://cocl.us/Geospatial_data
- New York's neighborhoods and coordinates were obtained from a JSON file located in URL: https://cocl.us/new_york_dataset
- Paris's neighborhoods and their geolocations were obtained using JSON read from the URL: https://opendata.paris.fr/explore/dataset/arrondissements/download/?format=json&timezone=Asia/Dubai
- The venues of the neighborhoods for each city were obtained and explored using FOURSQUARE.

2.2 Data cleaning

The format of the acquired data (either scrapped or downloaded) for this project was quite different and therefore different workflows were used for each particular case in order to obtain a resulting dataframe for each city to start working with that mainly consisted of the following main columns 'Neighborhood', 'Latitude' and 'Longitude'.

2.3 Neighborhood's Venues Information

Foursquare API credentials and the explore function were used on each neighborhood create a dataframe containing the most venues information. The resulting dataframe obtained for New York is shown on Table 1. Similar tables were obtained for Toronto and Paris neighborhoods.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	KFC	43.754387	-79.333021	Fast Food Restaurant
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

Table 1. Neighborhoods most common venues Dataframe (5 first rows) for New York city.

3. Exploratory Data Analysis

3.1 Analyzing Each Neighborhoods

The resulting most common venues in the dataframes obtained as described in Section 2 above were one-hot encoding and then the rows were grouped by 'neighborhood' and by taking the mean of the frequency of occurrence of each 'Most common venue' category in order to create a new dataframe and display the top 10 venues for each neighborhood using the 'return_most_common_venues' pre-defined function. The resulting dataframe for Toronto city neighborhoods is shown below in Table 2. Similar tables were obtained for New York and Paris cities.

	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	Adelaide, King, Richmond	Coffee Shop	Café	Thai Restaurant	Steakhouse	Bar	Gym	Breakfast Spot	Asian Restaurant	American Restaurant	Restaurant
1	Agincourt	Lounge	Clothing Store	Breakfast Spot	Skating Rink	Drugstore	Discount Store	Dive Bar	Dog Run	Doner Restaurant	Donut Shop
2	Agincourt North, L'Amoreaux East, Milliken, St...	Park	Playground	Donut Shop	Dim Sum Restaurant	Diner	Discount Store	Dive Bar	Dog Run	Doner Restaurant	Drugstore
3	Albion Gardens, Beaumond Heights, Humbergate, ...	Grocery Store	Pizza Place	Fast Food Restaurant	Beer Store	Sandwich Place	Fried Chicken Joint	Coffee Shop	Pharmacy	Comfort Food Restaurant	Dim Sum Restaurant
4	Alderwood, Long Branch	Pizza Place	Coffee Shop	Skating Rink	Dance Studio	Pharmacy	Pub	Sandwich Place	Gym	Airport Service	Dessert Shop

Table 2. Dataframe (5 first rows) displaying the top 10 most common venues in Toronto city.

3.2 Clustering Neighborhoods

The neighborhoods of each city were segmented and clustered into 5 clusters using the 'Most common venue' feature and the '***k-means***' clustering algorithm. Then, geopy library was used to get the map latitude and longitude reference values for Toronto, New York and Paris cities City and afterwards, the emerging clusters were mapped on it by using '***Follium***' library to visualize them.

The produced maps for Toronto, New York and Paris cities are represented in Figures 1 to 3 respectively. A rapid examination of Toronto map, Figure 1, show that the most predominant cluster is the one represented with red color dots. A similar inspection on Figure 2 allow us to identify two predominant clusters for New York city represented in red and pink colors. Finally, we can observe that for Paris, the largest cluster is represented in blue color circles

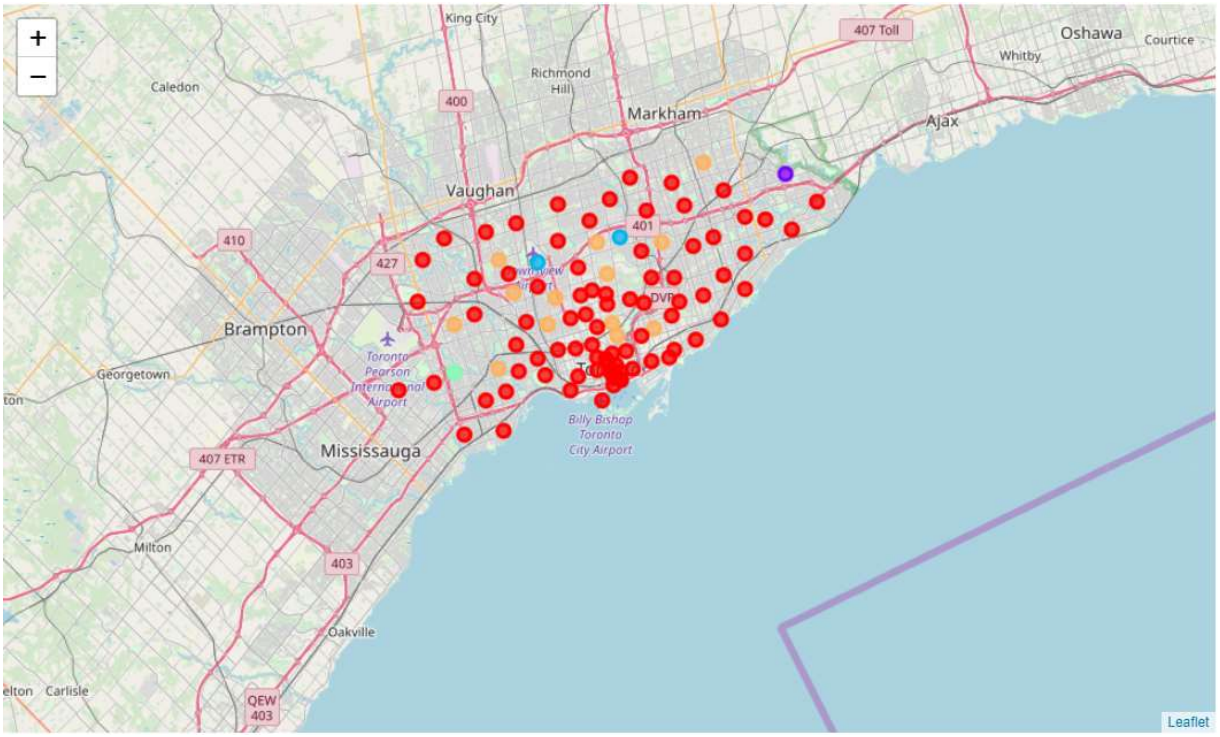

Figure1. Toronto map displaying Neighborhood clustering based on 'Most common venues' feature

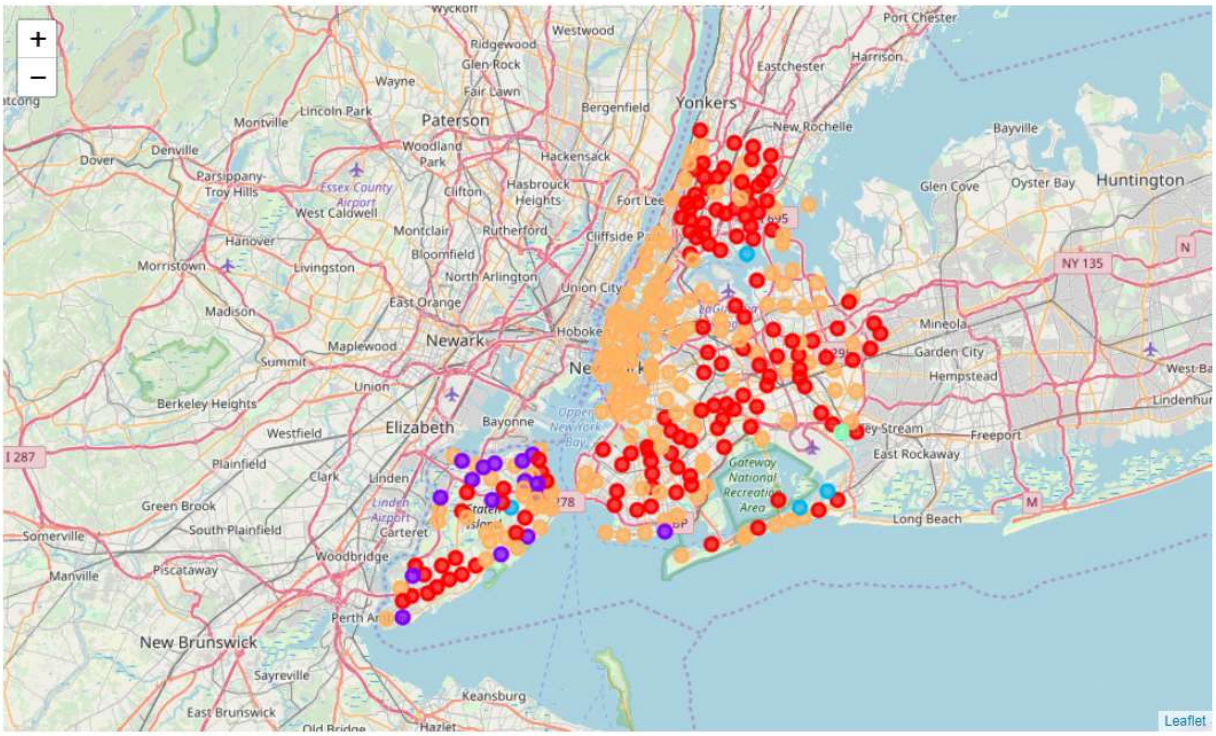

Figure 2. New York map displaying Neighborhood clustering based on ‘Most common venues’ feature

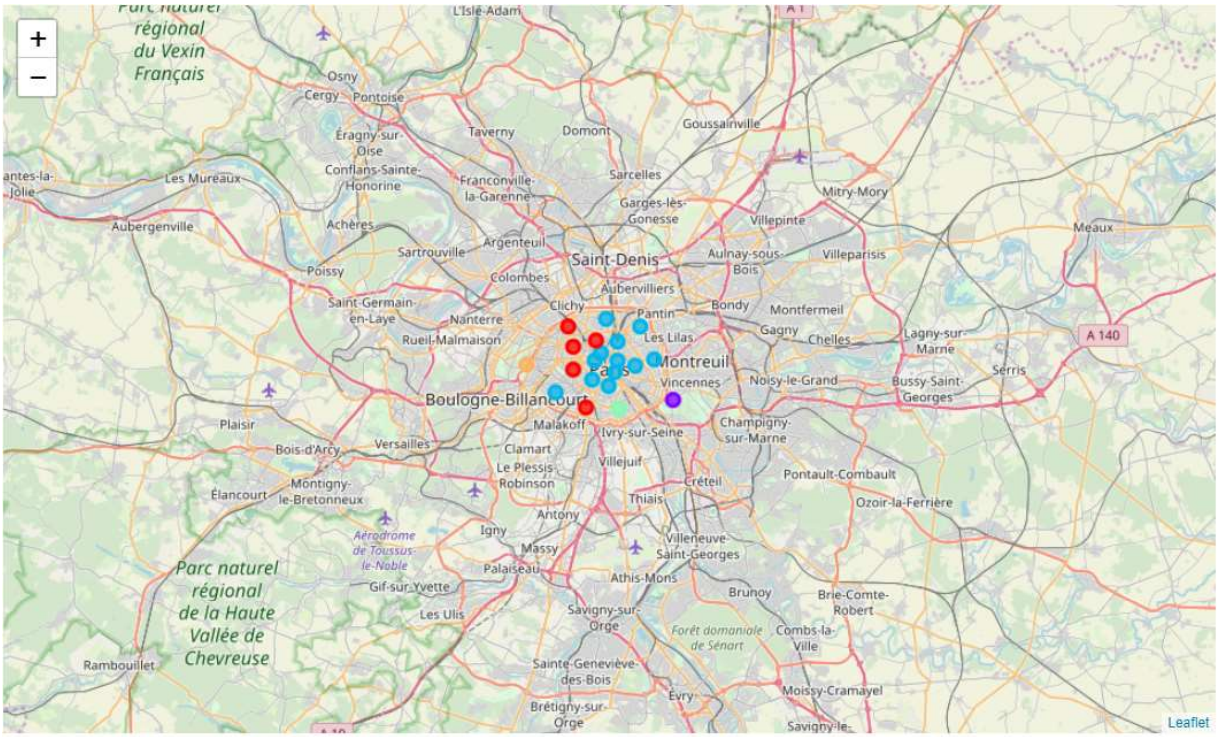

Figure 3. Paris map displaying Neighborhood clustering based on ‘Most common venues’ feature

3.3 Examining Most Representative Clusters

An examination the clusters dataframes for each city allowed us to identify and select the clusters with more neighborhoods as the most representative for each city. Thus, Cluster #1(red dots in Figure 1) with 82 neighborhoods and 820 top-ten 'Most common venues' observations resulted as the most representative for Toronto. Similarly, Cluster #5 (pink dots in Figure 2) with 159 neighborhoods and 1590 top-10 'Most common venues' observations' was identified for New York and finally Cluster #3 (blue dots in Figure 3) with 12 neighborhoods and 120 top-10 'Most common venues' for Paris.

A further inspection and transformation of the above selected clusters dataframes consisted in grouping the neighborhood's most common venues and count their occurrence on each selected cluster and subsequently transform these occurrences into their equivalent relative frequency expressed in percentage, obtaining a new dataframe for each city. These resulting counting and frequencies for Paris city are shown in Tables 3 and 4 respectively as example. The resulting total number of 'Most common venue' categories Toronto, New York and Paris were: 172,238 and 43 respectively.

Finally, bar charts of the calculated venue occurrence frequencies were produced for Toronto, New York and Paris cities and they are shown in Figures 4 to 6.

	Most common venue	Frequency
0	Art Gallery	1
1	Art Museum	1
2	Asian Restaurant	1
3	Bakery	5
4	Bar	6

Table 3. Most common venue occurrence counting (absolute frequency) for Paris city (first 5 rows)

	Most common venue	Frequency
0	Art Gallery	0.833333
1	Art Museum	0.833333
2	Asian Restaurant	0.833333
3	Bakery	4.166667
4	Bar	5.000000

Table 4. Most common venue frequency (%) for Paris city

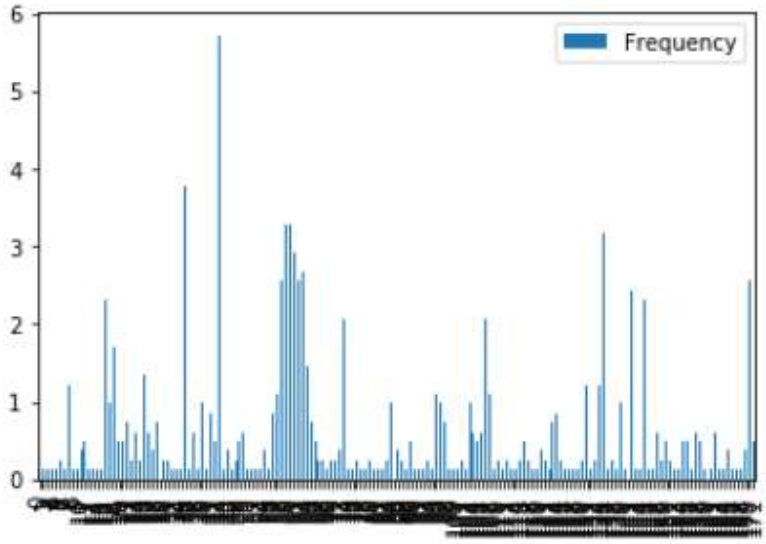

Figure 4. Toronto Cluster #1 Most common venue categories frequency bar chart.

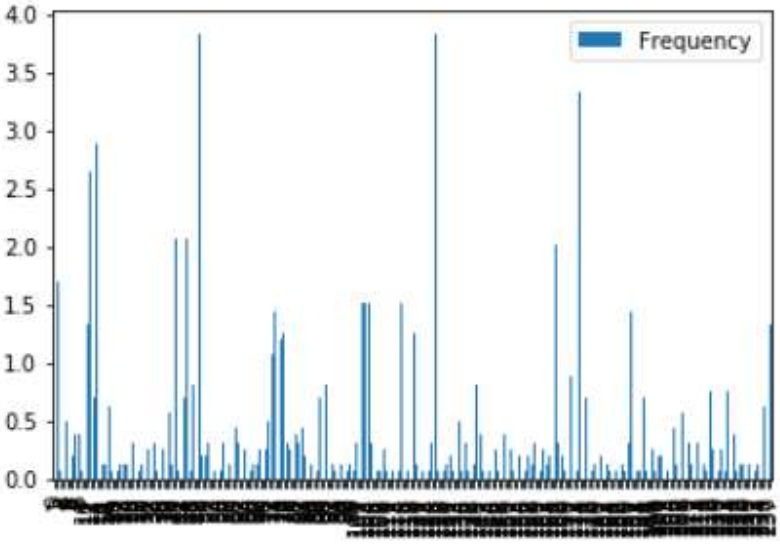

Figure 5. New York Cluster #5 Most common venue categories frequency bar chart.

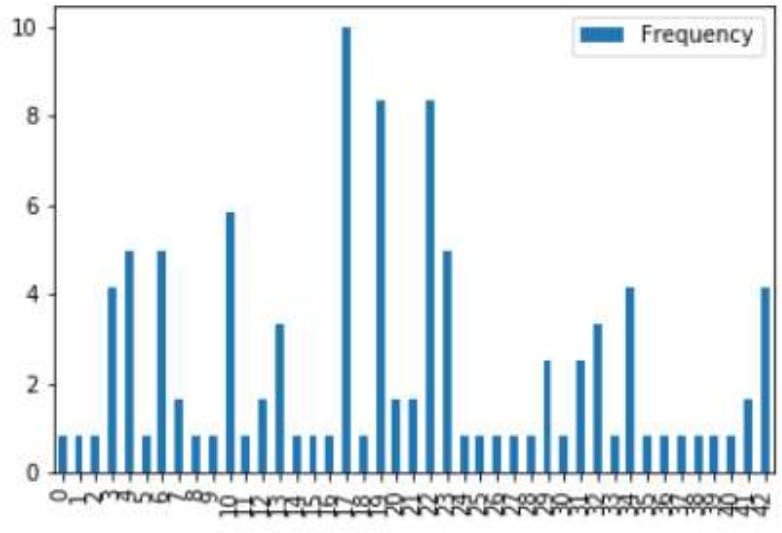

Figure 6. Paris Cluster #3 Most common venue categories frequency bar chart.

4. Comparative Modeling

in order to determine how similar or dissimilar the cities are then a comparison among the ‘Most common venues’ existing in the most representative clusters of each city needs to be done. Therefore, the dataframes obtained in Section 3.3 need to be merged but before doing that a transformation of the new dataset needs to be done.

First of all, a visual inspection of Figures 4 to 6 shows that it is clear that there are venues with more frequency than others and they are the ones we are interested in for comparison purposes. Therefore, a 1% frequency as minimum cut-off will be used to ignore rows data with a frequency less or equal than this cut-off. After applying the cut-off on the three dataframes these are reduced to 24, 29 and 19 ‘Most common venue’ categories on each dataframe for Toronto, New York and Paris respectively.

Second, a new column named 'city' is added to the dataframes, this column contains the name of the city. Then the 03 dataframes are merged into a single one containing 72 rows. Later, six columns populated with zero values were added into this new dataframe. The first three new columns were populated with frequency values filtered by city and the last three columns were populated with a normalized frequency based on the new size of the datasets after the cut-off.

Third, the new dataframe was grouped by venue grouping resulting in 48 ‘Most common venue’ categories. This resulting final dataframe is shown in Table 5 and was used for comparing the ‘most common venue; feature in the 03 cities subject to study.

Most common venue	Frequency	Toronto	New_York	Paris	Toronto_norm	New_York_norm	Paris_norm
American Restaurant	2.917625	1.219512	1.698113	0.000000	2.207506	3.195266	0.000000
Bakery	9.125249	2.317073	2.641509	4.166667	4.194260	4.970414	5.208333
Bar	9.600399	1.707317	2.893082	5.000000	3.090508	5.443787	6.250000
Breakfast Spot	1.341463	1.341463	0.000000	0.000000	2.428256	0.000000	0.000000
Café	11.689293	3.780488	2.075472	5.833333	6.843267	3.905325	7.291667
Coffee Shop	12.901519	5.731707	3.836478	3.333333	10.375276	7.218935	4.166667
Dim Sum Restaurant	1.097561	1.097561	0.000000	0.000000	1.986755	0.000000	0.000000
Diner	2.560976	2.560976	0.000000	0.000000	4.635762	0.000000	0.000000
Discount Store	3.292683	3.292683	0.000000	0.000000	5.960265	0.000000	0.000000
Dive Bar	3.292683	3.292683	0.000000	0.000000	5.960265	0.000000	0.000000
Dog Run	2.926829	2.926829	0.000000	0.000000	5.298013	0.000000	0.000000
Doner Restaurant	2.560976	2.560976	0.000000	0.000000	4.635762	0.000000	0.000000
Donut Shop	3.877895	2.682927	1.194969	0.000000	4.856512	2.248521	0.000000
Drugstore	1.463415	1.463415	0.000000	0.000000	2.649007	0.000000	0.000000
Fast Food Restaurant	3.331032	2.073171	1.257862	0.000000	3.752759	2.366864	0.000000
Grocery Store	2.606995	1.097561	1.509434	0.000000	1.986755	2.840237	0.000000
Italian Restaurant	14.242982	2.073171	3.836478	8.333333	3.752759	7.218935	10.416667
Japanese Restaurant	6.097561	1.097561	0.000000	5.000000	1.986755	0.000000	6.250000
Park	3.232091	1.219512	2.012579	0.000000	2.207506	3.786982	0.000000
Pharmacy	1.219512	1.219512	0.000000	0.000000	2.207506	0.000000	0.000000
Pizza Place	9.004065	3.170732	3.333333	2.500000	5.739514	6.272189	3.125000
Restaurant	6.605691	2.439024	0.000000	4.166667	4.415011	0.000000	5.208333
Sandwich Place	3.763614	2.317073	1.446541	0.000000	4.194260	2.721893	0.000000

Table 5. Final dataframe (23 out of 48 rows) containing data for ‘Most common venues’ comparison

Finally, two bar chart plots were prepared, one displaying the frequency and the other displaying the normalized frequency (just for better visualization purposes) for the ‘Most common venue’ categories in Toronto, New York and Paris. These mentioned charts are shown in Figures 7 and 8 respectively.

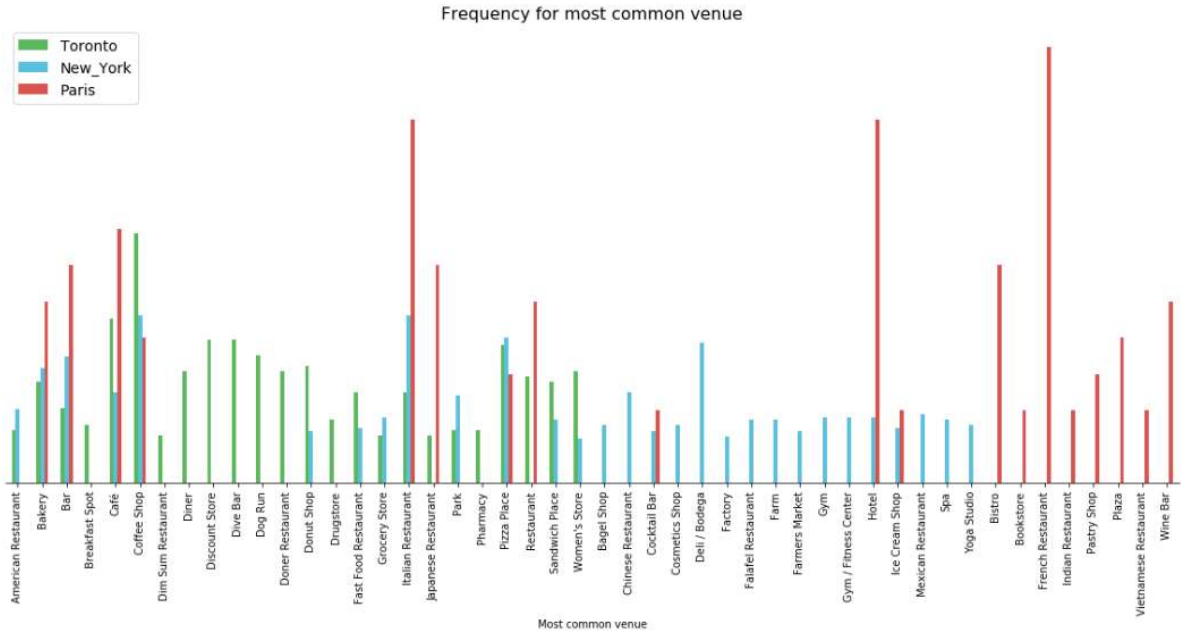

Figure 7. Frequency of 'Most common venue' categories comparison.

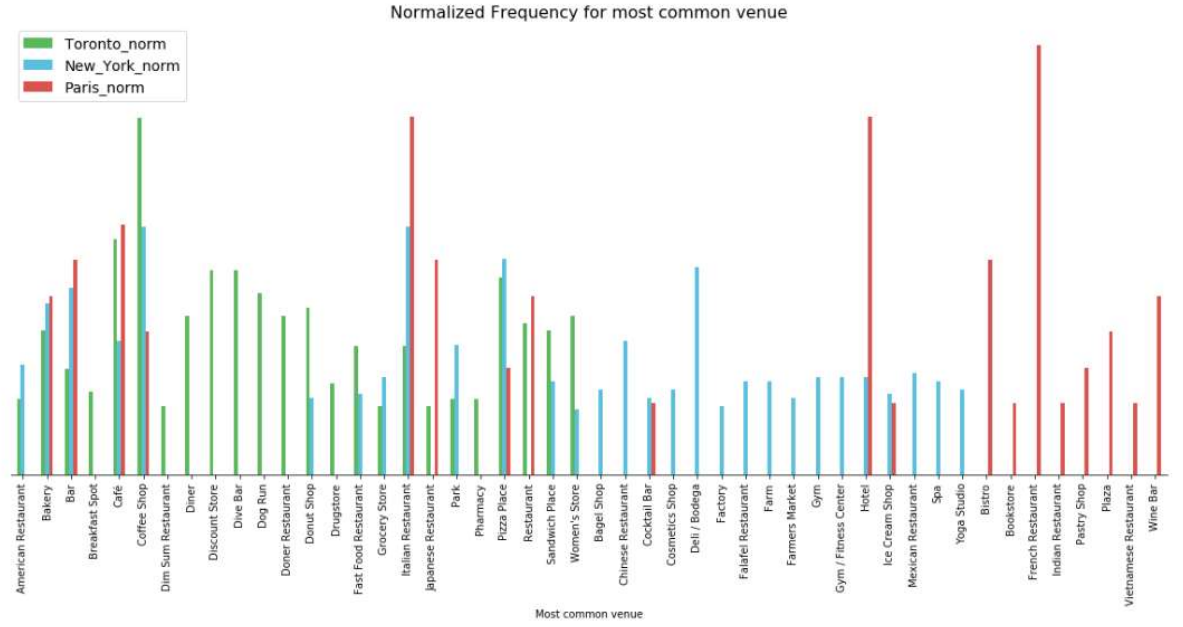

Figure 8. Normalized Frequency of 'Most common venue' categories comparison.

5. Results & Conclusions

The results are very interesting. It is clear that even though there are some similarities among these three multi-cultural cities, there are also differences among them. The plots show that out of the resulting 19 most common venues filtered for Paris, only 06 type of venues are most common simultaneously on these three cities. Besides, Paris shares 03 type of venues exclusively with New York and 02 type of venues exclusively with Toronto. Finally, there are 08 type of common venues (Bistro, Bookstore, French Restaurant, Indian Restaurant, Pastry Shop, Vietnamize Restaurant and Wine Bar) which represents 42%, this difference makes Paris distinctive from New York and Toronto. However, there are also similarities because Paris shares 09 (47%) common venues with New York and 08 (42%) with Toronto.

6. Way Forward

Moving from one city to another always represents a challenge. In this case, the analysis shows that people from New York shares more similarities with Paris than Toronto in terms of most common existing venues but the degree of similarity is only 47%. However, Toronto is not far with 42% on this field and therefore more study including other variables need to be taken account. For instance, French is a language that it is spoken widely in Toronto but not in New York. Similarly, parameters like weather, living cost, etc. need to be considered.

With regards to the workflow used to compare the cities, this can be improved, specially in the last stage where only visual analysis of the comparative bar chart was used to determine the degree of similarity between the cities. This will be very helpful when analyzing simultaneously more number of cities and larger amount of ‘Most common venue’ categories.