2020

Capstone Project - The Battle of Neighborhoods

Comparing New York City with Toronto

04 May 2020

Harsha Stanislaus

1. Introduction

1.1. *Background*

1.1.1. Target Audience

A marketing company, who has had previous experience in selling real-estate and rental properties in and around New York City, have asked me to perform a 'quick and dirty' analysis to help them understand how similar Toronto is to NYC. In particular, they are interested in the similarity of public venues, local amenities and other businesses available in each neighbourhood.

1.1.2. Business Problem

It is their intention to assess whether (and also where) they can use their various NYC marketing campaigns, in potential future real-estate deals and bids within the Toronto market.

1.1.3. Objective

As such, the goal of this study, is to compare the commonality between the two cities - Toronto and New York City. To do this, at a sufficient/initial level, I will conduct analysis on geolocation data, comparing the relative composition and frequency of local venues (e.g. restaurants, coffee shops, gyms, etc.) within each neighbourhood. Thereby building a set of clusters - or profiles - of distinct features that the client's clients may value in their search for a property.

1.1.4. Intended use of results

The results of the analysis, will provide a good starting point for the marketing client, to pinpoint potential locations in Toronto to venture into. It will indicate which profiles are common between each city, and subsequently where they could leverage previous expertise and marketing material to maximize their value proposition within Toronto.

2. Data

2.1. Data sources

I am aiming to determine the high-level similarities between a series of data points. Given that the scope of this report will not extend to performing a deep analysis of micro-level features of venues (e.g. user recommendations, facilities, etc.), nor will I analyze other external data (such as incomes, wages, property values, etc.) – I will limit the analysis to frequency of venue types between locations. Thus, the report will focus on results from two types of data:

1. Neighbourhoods and geographical data – static neighbourhood data, sourced from Wikipedia (for Toronto data) and an earlier IBM lab (for NYC data). These will be html/xml and json formats, which will be converted into data frames for the analysis. Each row will contain a unique neighbourhood [1] value, with its longitudinal and latitudinal coordinates.

2. Venue and venue category data [Foursquare location data] - each row will contain a unique venue name, category and its longitudinal and latitudinal coordinates, as well as its neighbourhood (with which we can merge this onto the data in item (1)). This will be sourced via the Foursquare API and imported as a json file, which will be converted into a data frame for the analysis. The call being made to the API, requests the top 100 venues within a radius of 500m, for the longitudinal and latitudinal coordinates of each neighbourhood centroid from (1).

2.2. Analytical approach

Since we are looking at the frequency of venue categories to groups which have not been labelled, the most suitable analytical approach for this type of assessment will be k-means clustering. This is an unsupervised approach, which is relatively quick to perform and iteratively generates groupings to build similar clusters upon. It has the additional factor of being relatively simple to explain to a non-data/IT-literate audience.

2.3. Tooling, output and automation

I will perform the data extraction, transformations, clustering analysis and visualisation using Python, since this will perform these tasks relatively quickly, with a presentation summarising the findings for the client. The code is fully automatable. The data sources lend themselves to being extracted and rerun whenever the analysis needs to be repeated. Since the Wikipedia website and NYC json files for neighbourhood data is slow moving, we do not expect these to change frequently, and can effectively remain as static reference datasets. Nevertheless, they can also be rerun and re-extracted quickly and easily, where required. The FourSquare data can be extracted through their API quickly – the download takes around 2-3 mins and is part of the python code that performs the analysis. It should be noted, the code requires FourSquare user credentials for a free (but credit card-verified) account, which allows 99,500 Regular API Calls per day and 500 Premium API Calls per day (at the time of writing this). The data in this analysis is currently classed as a Regular API call, and is well below the daily limit (see record counts section).

[1] ***Terminology Note:*** for the purposes of explaining the data, multiple neighbourhoods within the same postal code area will be referred to as a single 'neighbourhood', to improve the ease of discussion in this report – since the data is structured in this manner.

3. Methodology

3.1. Data quality exploratory checks

3.1.1. Data completeness and accuracy

Data in the neighbourhood datasets/Wikipedia table, appears up-to-date and complete. Neighbourhoods in large established cities are fairly slow moving and can be viewed here as static information. The venue data is more changeable. However, since this is being extracted from the live FourSquare database, a community-sourced real-time global database, I have a high degree of confidence that this is a largely accurate and complete repository of information. The analysis will be performed, however, on a recent snapshot of the data, extracted on 1st May 2020.

Record counts were extracted on the datasets, prior to the analysis and are listed below:

	Neighbourhood Data		*Venue Data*	
	Total / Distinct Neighbourhoods	Distinct Boroughs	Total Venue Categories	Distinct Venue Categories
New York City	306	5	9,751	423
Toronto	103	10	2,110	262

3.1.2. Data quality checks

The data is of good quality. Fields do not appear to contain spelling errors or incongruent variants of the same word, nor do they contain erroneously duplicate values, which would otherwise require cleansing and validating. Moreover, the numbers of distinct boroughs and neighbourhoods appear to be in line with a reasonable expectation of these.

3.1.3. Primary Join key checks

It should be noted, that while the information in these datasets do not contain erroneous duplicates, there is one duplicate Neighbourhood value between Toronto and New York City: “Rosedale”. The two datasets can therefore not be joined on their Neighbourhood field alone. I will ensure that any join performed (where neighbourhood is its expected granularity), is done so using a concatenation of City and Neighbourhood fields as its primary join key, to eliminate unexpected join results.

3.1.4. Coverage of Primary Join key checks

Using the PK (City concatenated with Neighbourhood) I tested the coverage of a full outer join between the Neighbourhood and Venue data. The results can be seen in the Venn diagram below. It shows 7 neighbourhood records do not have any venue data associated with them. 5 of these are in Toronto and the remaining two are in NYC. These are listed below.

City	Boroughs	Neighbourhoods
Toronto	Scarborough	Upper Rouge
Toronto	North York	Willowdale , Newtonbrook
Toronto	Etobicoke	Islington Avenue
Toronto	Etobicoke	West Deane Park , Princess Gardens , Martin Grove , Islington, Cloverdale
Toronto	North York	Humber Summit
NYC	Staten Island	Port Ivory
NYC	Staten Island	Howland Hook

What this means, is that there were no venues returned from FourSquare for their coordinates. As such, I will negate these neighbourhoods from the analysis and note it, as an assumption at this point.

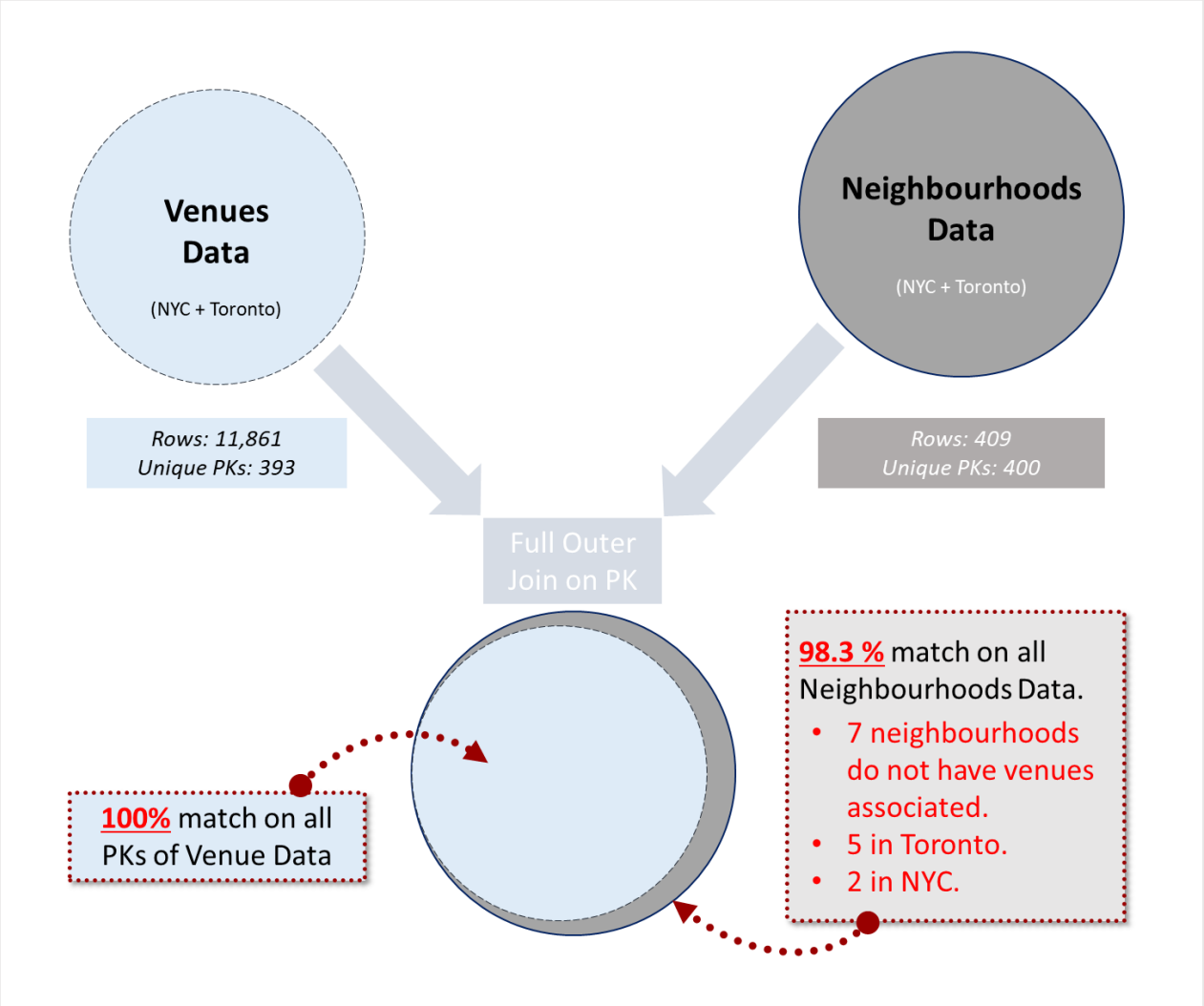


Fig. 1: Venn diagram to show the join coverage between Neighbourhood data and Venue data on their Primary Key (PKs).

3.2. Geolocation data sense checks

3.2.1. Plotting the latitudes and longitudes of each neighbourhood

By plotting the latitude and longitude values in both NYC and Toronto datasets, I can check whether I see the correct locations and areas covered on these. As can be viewed from the below plots, there is a good coverage and spread of neighbourhood centroids on the maps, in the expected city locations.

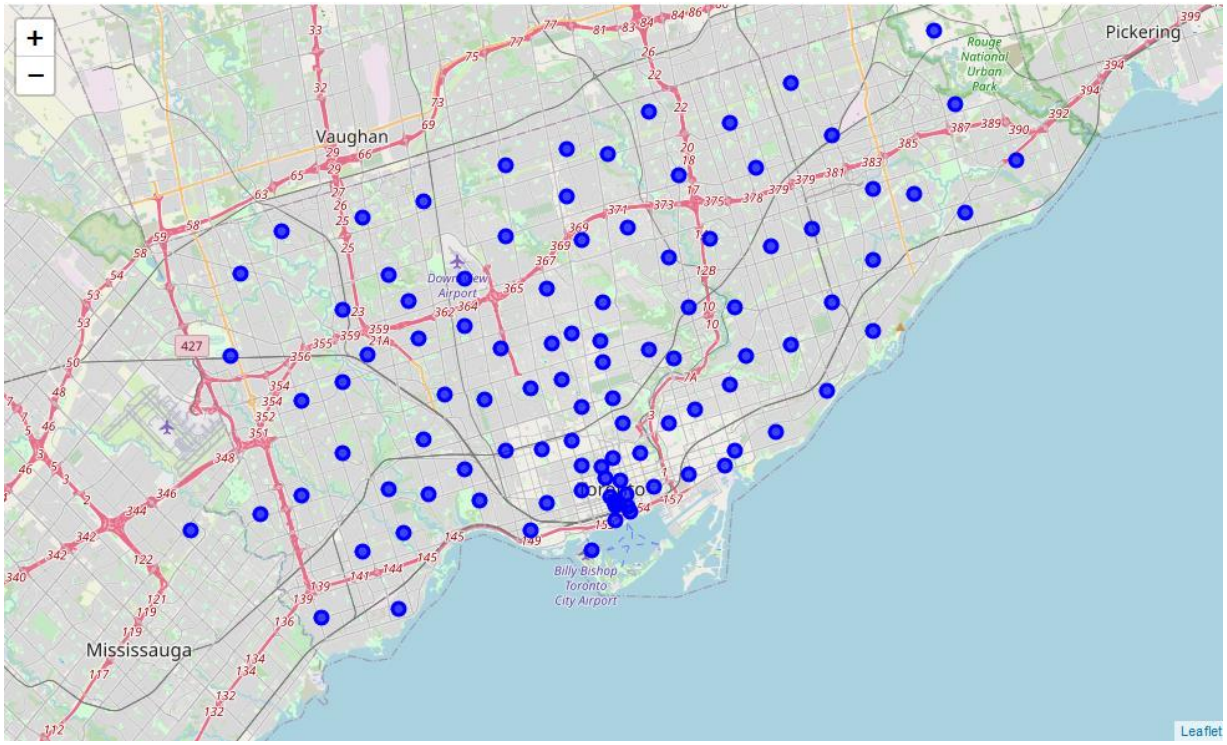


Fig. 2: Toronto map, showing the neighbourhood centroid locations in blue marks.

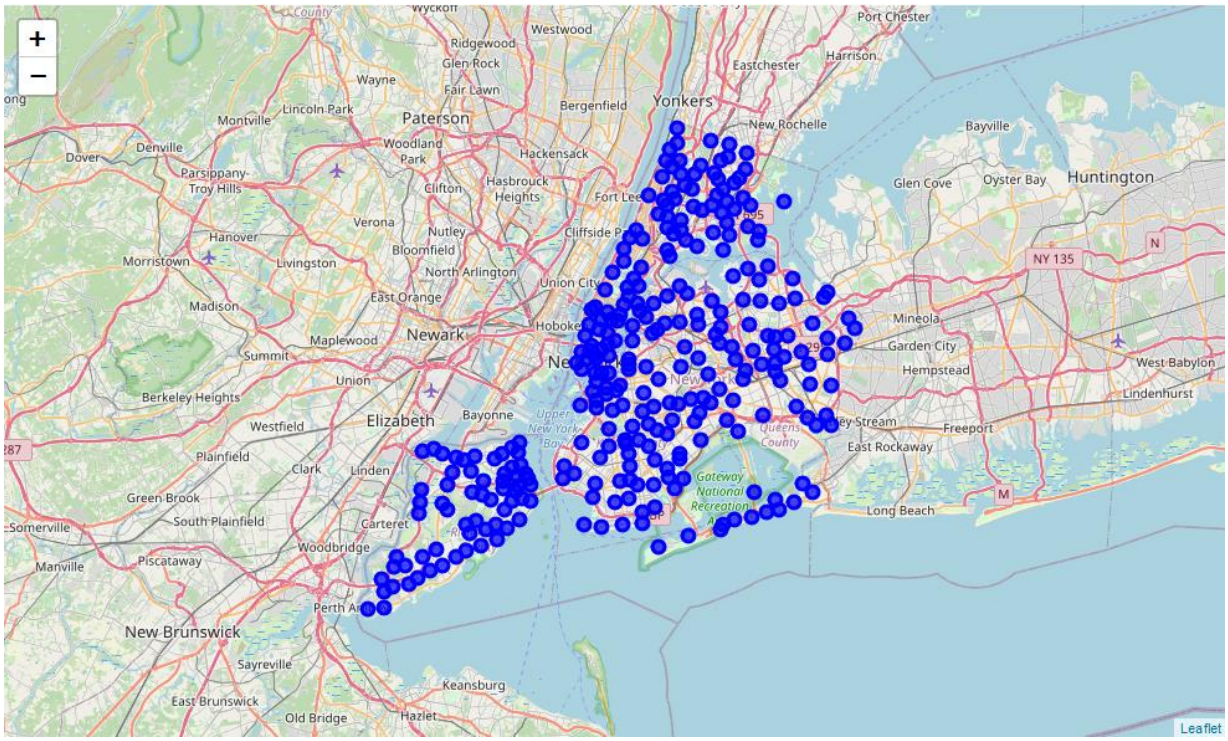
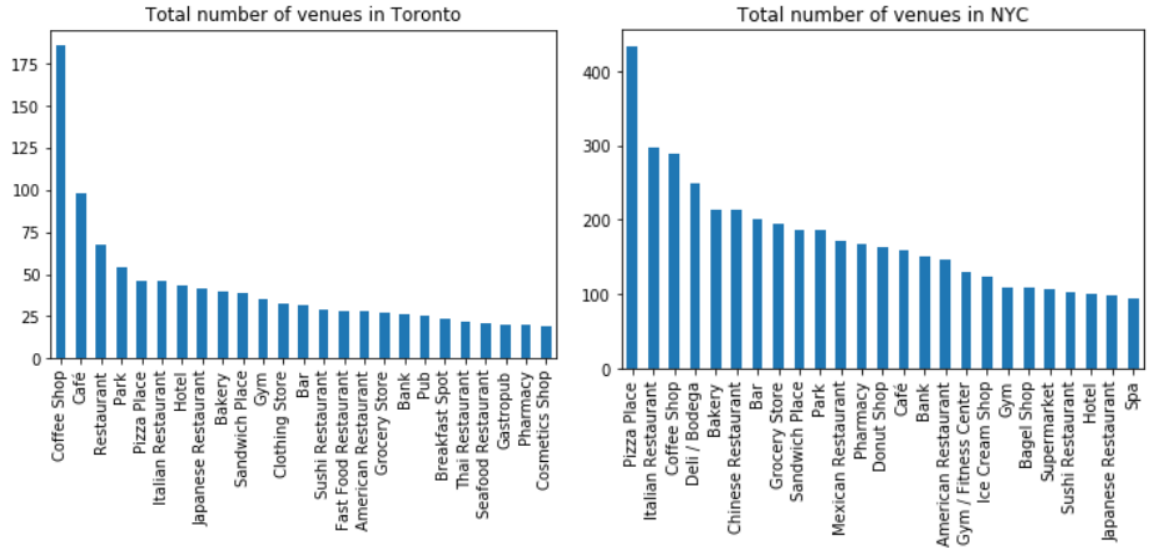


Fig. 3: New York City map, showing the neighbourhood centroid locations in blue marks.

3.3. Summary statistics on venue data

3.3.1. Top 25 Venues

A summary of the top 25 venues by city is shown below. This gives us a sense about the overall types of venues to expect to be pushed into the final profiles. Topping both lists are coffee shops, and pizza and Italian restaurants. So we can see there is a ranging degree of similarities and dissimilarities in both cities, moreover the clustering will help to evaluate the specific areas of direct comparability and overlap.



3.4. Analytical Model: Unsupervised learning application

3.4.1. K-means clustering

I will be applying a K-means clustering methodology on the data. The reason this is suitable, is that the objective here is to create a series of comparable groups of data within the neighbourhoods in NYC and Toronto. We do not know the groups, and this algorithm is particularly useful for this kind of application. The result could either show very dissimilar groups or a spectrum similar groups between the two cities. In which case, these overlaps will give a good indication as to where the client could potentially operate with the most success – note while this is only the first step of their process, these indicators will likely be a helpful start. I will use a k-value of 10, which will ask the model to generate 10 clusters (0-9).

4. Results

4.1. Running and outputting results from the model

The clustering algorithm used the venues of the each neighbourhood to build unassigned groups, depending on their similarity to each other, based on the top 100 venues in each neighbourhood. The 10 clusters and the top venues can be seen in the below tables:

Cluster 0 (partial...)	
Venue Category	Count
Beach	31
Deli / Bodega	7
Bagel Shop	4
Bus Stop	4
Donut Shop	3
Arepa Restaurant	2
BBQ Joint	2

Cluster 1 (partial...)	
Venue Category	Count
Deli / Bodega	92
Chinese Restaurant	83
Bank	72
Pizza Place	55
Bakery	54
Donut Shop	50
Grocery Store	46

Cluster 2	
Venue Category	Count
Bar	3
Bus Station	1
Drugstore	1
Garden	1
Ice Cream Shop	1
Lawyer	1
Rental Car Location	1

Cluster 3 (partial...)	
Venue Category	Count
Park	26
Convenience Store	3
Playground	3
Pool	3
Trail	3
Bank	2
Bakery	1
Boat or Ferry	1
Bus Line	1
Bus Stop	1
Construction & Landscaping	1
Field	1

Cluster 4 (partial...)	
Venue Category	Count
Bar	200
Bakery	197
Coffee Shop	175
Cafe	166
American Restaurant	156
Bank	98
Chinese Restaurant	90
Bagel Shop	79
Art Gallery	68
Burger Joint	64
Deli / Bodega	58
Bookstore	57

Cluster 8 (partial...)	
Venue Category	Count
Caribbean Restaurant	29
Deli / Bodega	7
Fast Food Restaurant	7
Fried Chicken Joint	7
Donut Shop	4
Bar	3
Bus Station	3
Chinese Restaurant	3
Discount Store	3
Bank	2
Breakfast Spot	2
Coffee Shop	2

Cluster 5	
Venue Category	Count
Playground	2
Convenience Store	1

Cluster 6	
Venue Category	Count
Pool	1

Cluster 7	
Venue Category	Count
Deli / Bodega	1

Cluster 9	
Venue Category	Count
Baseball Field	2
Food Service	1

We can immediate see that some clusters are more useful than others. Some, like clusters 1,3,4,and 8 contain many venue types with with a high degree of frequency associated with them. While other Clusters only show a single venue category, and at this stage look less useful for the objective.

Next, the model outputs these clusters onto a map (see figures 4 and 5 below). These show a direct comparison of similar neighbourhood clusters between NYC and Toronto. Figure 6, shows that Toronto

has 7 of the 10 model clusters associated with it. The majority of these neighbourhoods fall into clusters 4, 3 and 1 (and to a lesser extent 8 and 2).

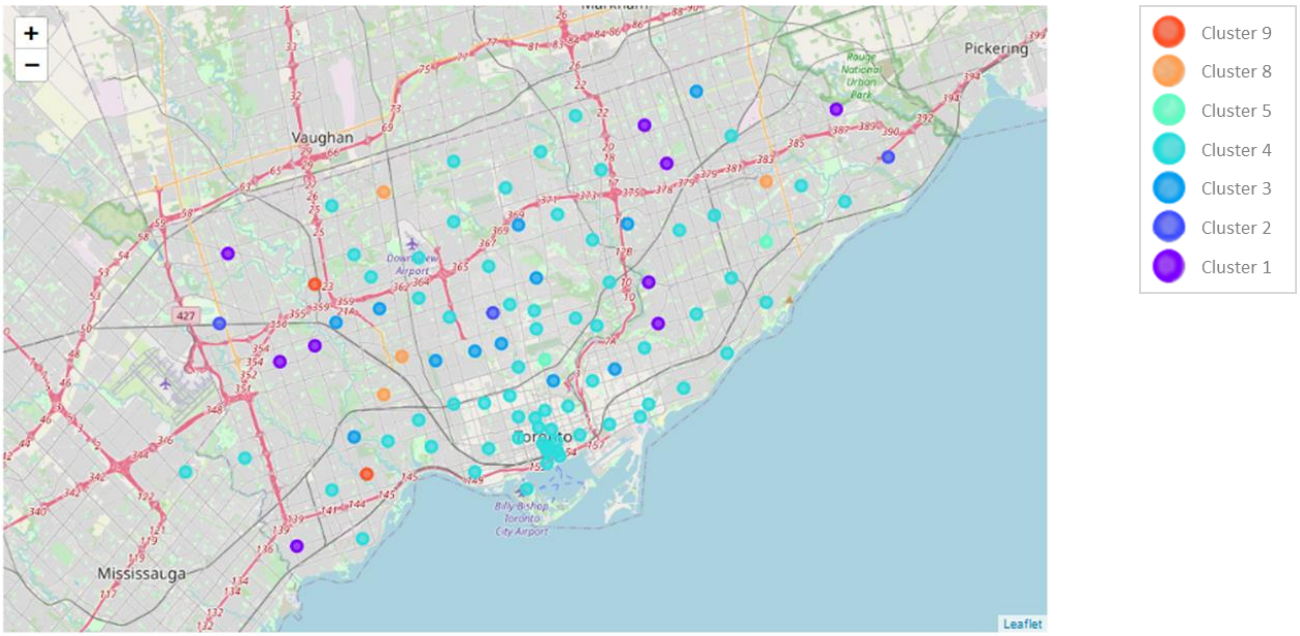


Fig. 4: Map of Toronto clusters

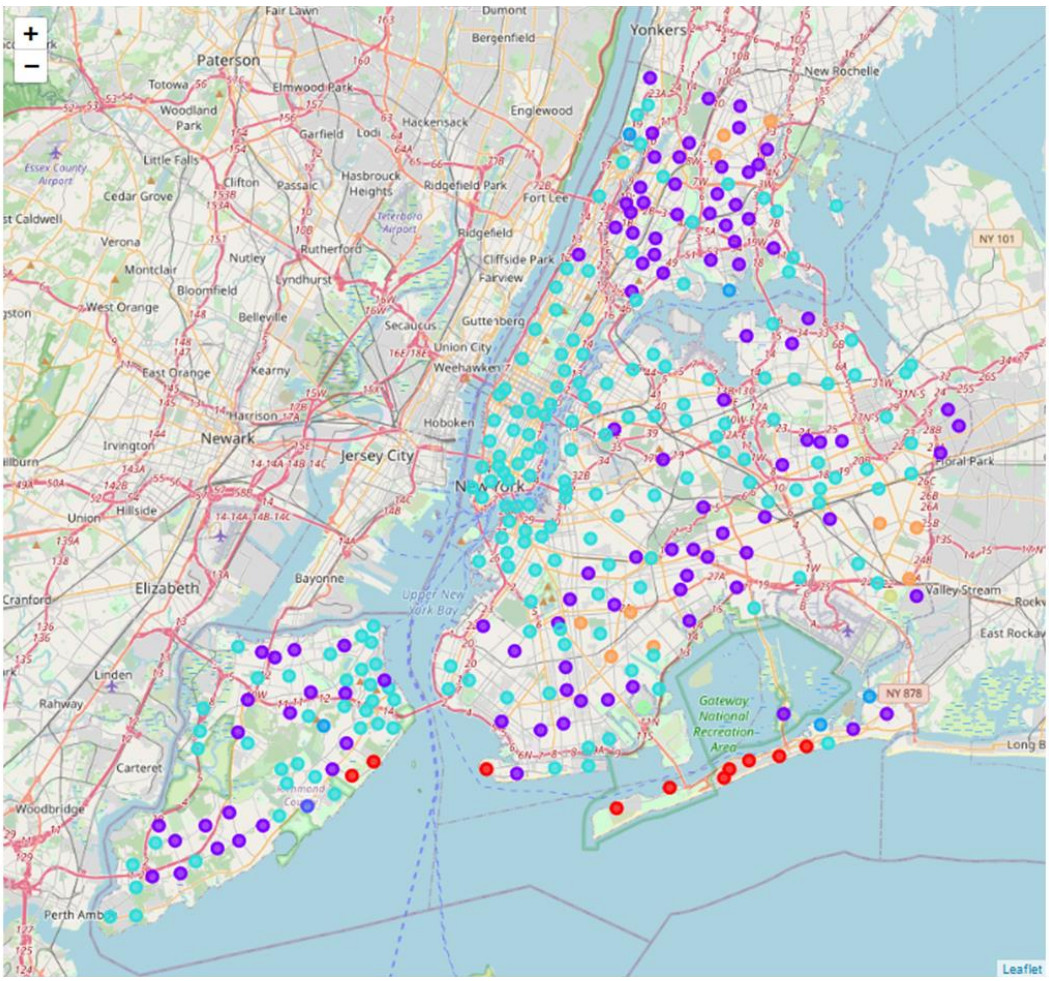


Fig. 5: Map of New York City clusters

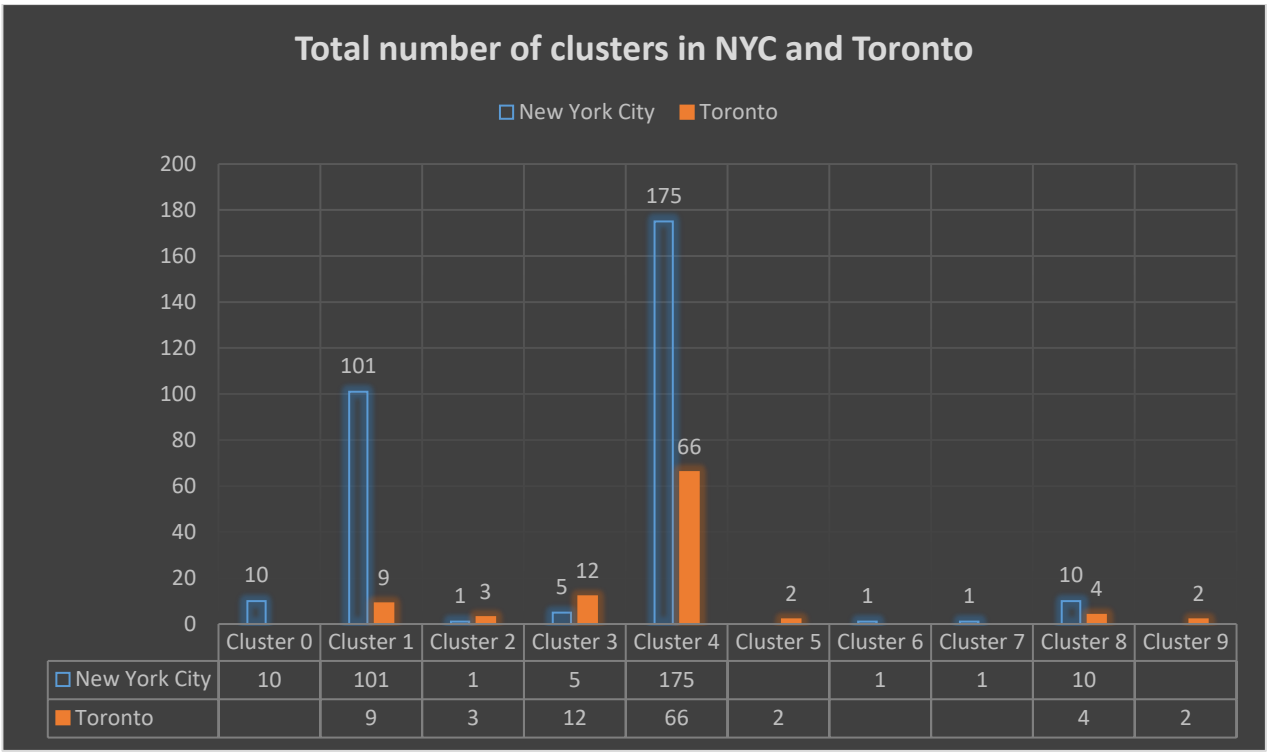


Fig.6: Counts of Cluster by City

Figure 6 also shows that clusters 0, 5, 6, 7, and 9 are uncommon between both cities, so are too dissimilar to be useful candidates for our client to infer that they would be able to replicate business from one city to the other.

5. Discussion

5.1. Assessment of cluster profiles

Below are two tables, outlining summary information of the 10 clusters produced from the analysis. They highlight some broad characteristics of each cluster. Diversity of venues has been included as a column. It is an assessment of the number and range of venue types within each neighbourhood.

The tables have been separated into clusters which exist in both NYC and Toronto Vs those that only exist in one of the two cities. The latter will therefore be less important for the client, given the assumptions of this analysis. (Furthermore, it can be seen that these clusters contain very low counts and therefore representation across either city. They also score poorly on venue diversity, with the exception of cluster 0, which is unique by its NYC beach locations).

Common Clusters existing in both NYC and Toronto:

Cluster No.	Profile Characteristics	Diversity of Venues[2]	Cluster Counts	
			NYC	Toronto
8	***International cuisine, low priced shopping:*** Caribbean/Chinese/Asian Restaurants, Fast Food, Delis/Bodegas, Cafes, Discount Stores	High	10	4
4	***Business district, large variety of shops and restaurants:*** Bars, Banks, Coffee Shops, Bakeries, All Types of Restaurant, Galleries, Gyms, Pubs, Hotels	High	175	66
3	***Green areas:*** Playgrounds, Parks, Pharmacies, Fitness	Low	5	12
2	Bars, limited shops	Low	1	3
1	***Suburban areas, high variety of restaurants, good transport:*** Banks, Coffee Shops, Bus Stops, All Types of Restaurant, Bakeries Big Stores	High	101	9

Uncommon Clusters existing exclusively in NYC or Toronto:

Cluster No.	Profile Characteristics	Diversity of Venues	Cluster Counts	
			NYC	Toronto
9	Baseball / Food only	Low	0	2
7	Deli/Bodegas only	Low	1	0
6	Pools only	Low	1	0
5	Playgrounds and Convenience Stores only	Low	0	2
0	Beach locations (incl. sports and monuments)	Medium	10	0

[2] Diversity of clusters are defined as those with significantly more than 2 types of venue in their neighbourhoods

The most frequent profiles in Toronto, are cluster numbers 4, 3 and 1 (these rows are highlighted in green in the previous tables, and are the purple/blue/aqua coloured bubbles in figure 4). These areas are most similar to those in NYC but also have a good spread and hence representation in NYC as well. It could be assumed that this will enable the client to leverage much of its existing marketing material from operations in NYC, and apply it to many areas in Toronto. Figures 4 and 5 should be referenced to help drill down into any specific NYC locations that the client feels its operations were particularly successful or transferable from, and can infer they will be successful in conducting similar business in those matching clusters of Toronto.

6. Conclusion

In this study, I have analysed the venues in NYC and Toronto and ran a k-means clustering algorithm to help group and label similar neighbourhoods together between the two cities. By mapping this out it can be seen which neighbourhoods in Toronto are largely similar to those in NYC.

Given the initial brief, the next step, would be for the client to map out the areas in NYC they have the most experience and existing marketing materials to leverage. As such, the client should then compare which neighbourhoods in Toronto, map most closely with these areas (using the cluster maps in figure 4 and 5). The comparison should be done both against the colour coded areas of the maps in Section 4, but also against the business knowledge the client has relative to the descriptions in the cluster profiles of Section 5.

This will act as a good indicator for the client to understand where they can proceed with next steps, in their assessment of conducting business at these locations.