Capstone Project - The Battle of Neighborhoods

Comparing New York City with Toronto

04 May 2020

Harsha Stanislaus

Introduction

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.

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.

Objective

- As such, the goal in 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 our client's clients may value in their search for a property.

Intended use of results

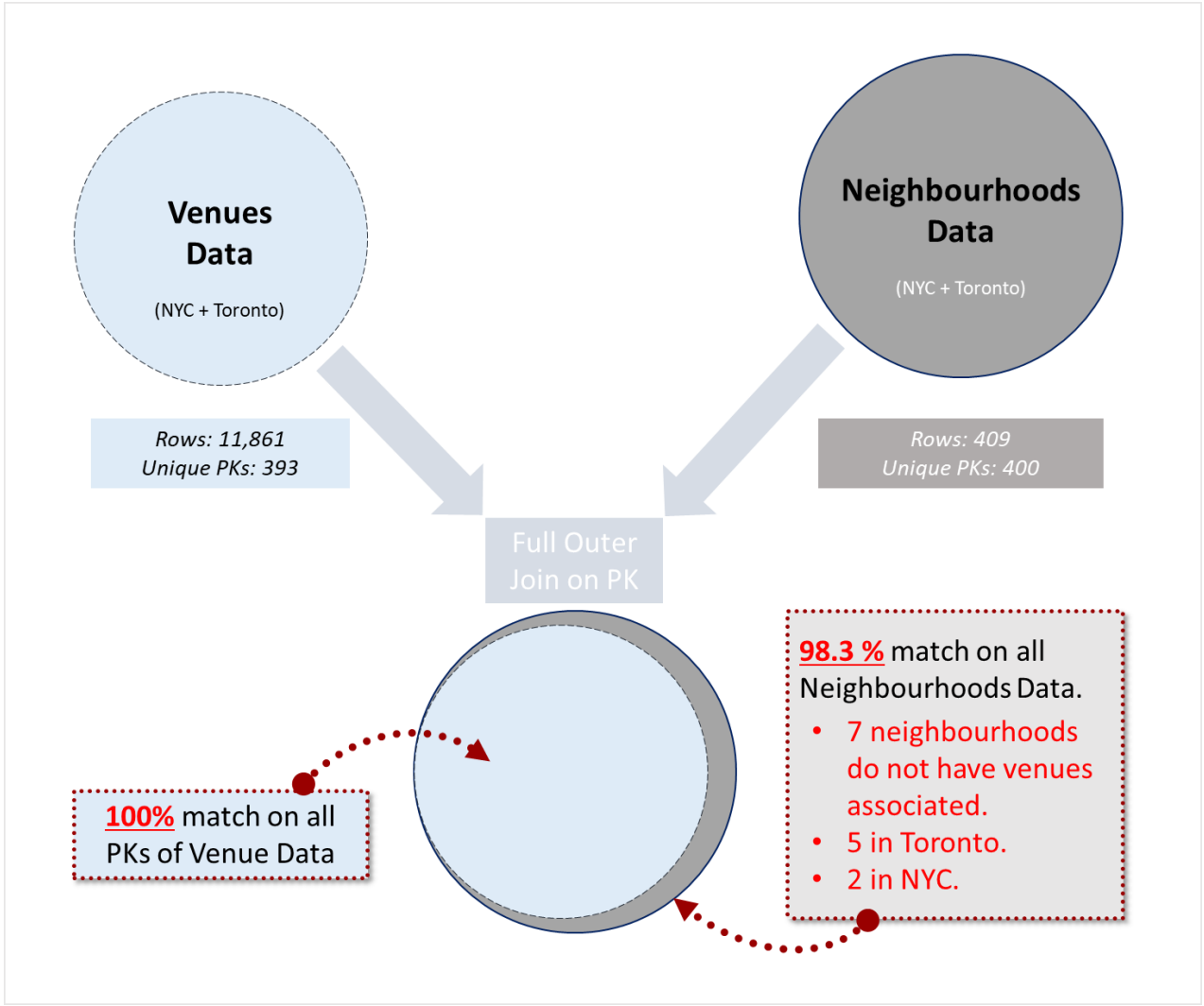
- The results of the analysis, will provide a good starting point for the marketing client, to pinpoint potential locations in Toronto that they could use to venture into. It will indicate which profiles are common between each city, and subsequently where they could leverage their previous expertise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

Data

- 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 value, with its longitudinal and latitudinal coordinates.
- 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 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 for the longitudinal and latitudinal coordinates of each neighbourhood from (1), within a radius of 500m.

	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

Data Exploration: Data Quality Checks



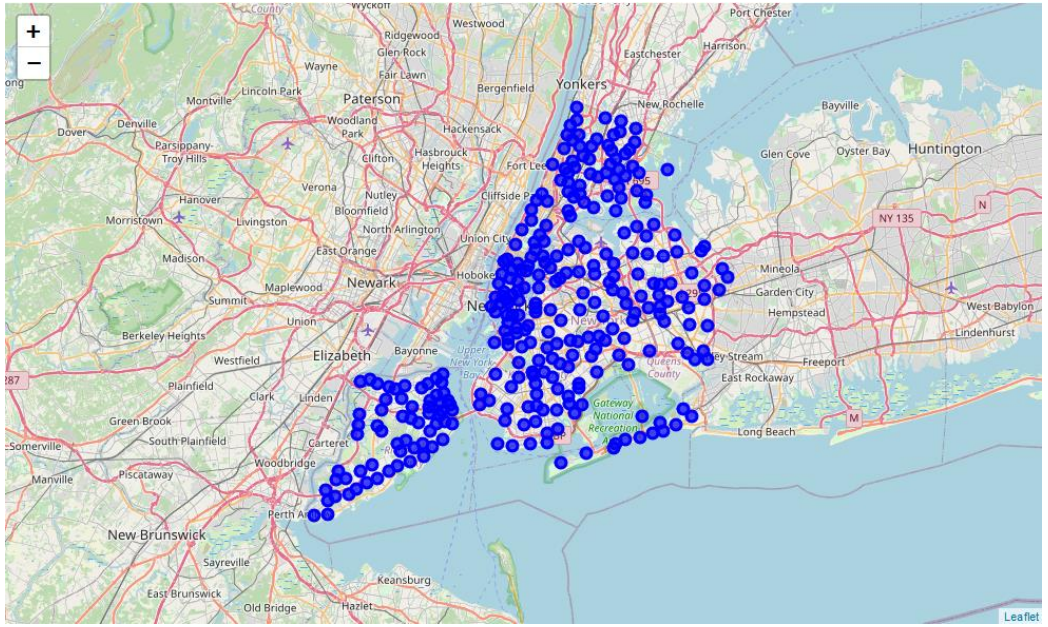
Venn diagram to show the join coverage between Neighbourhood data and Venue data on their Primary Key (PKs).

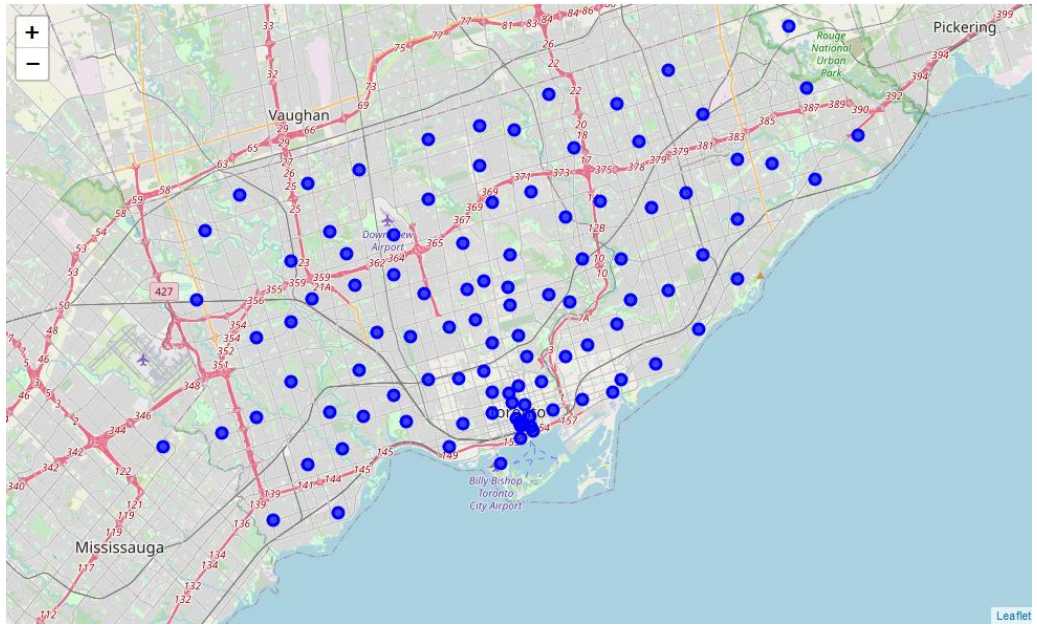
Neighbourhoods that do not have venues:

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

Otherwise, accuracy and completeness of data is good and sufficient quality for our purposes. Moreover, there are no duplicate records or PKs.

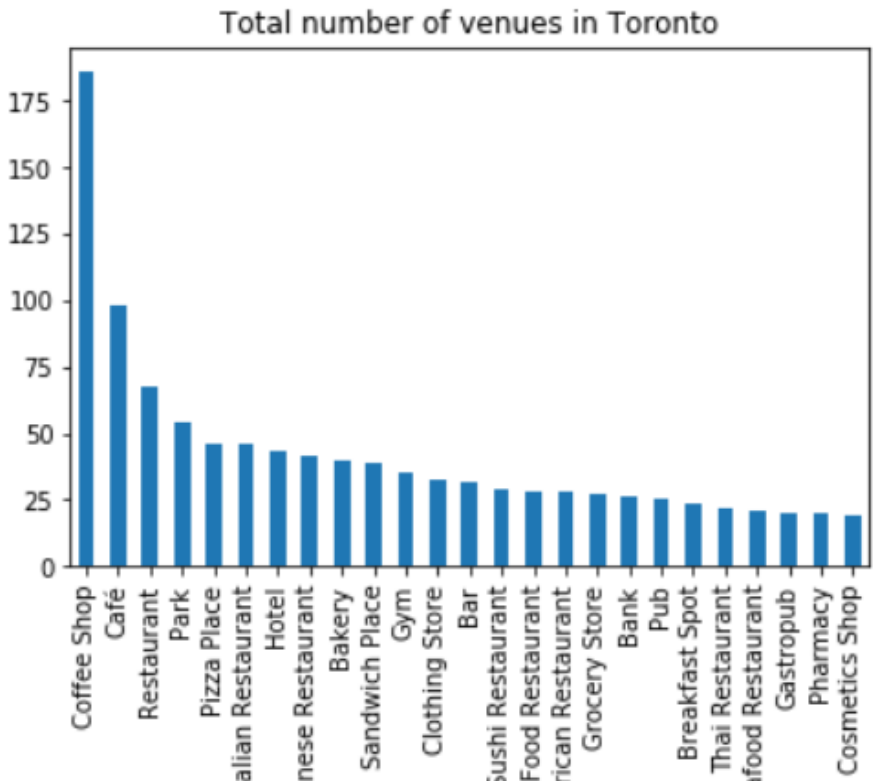
Data Exploration: Geolocation data sense checks

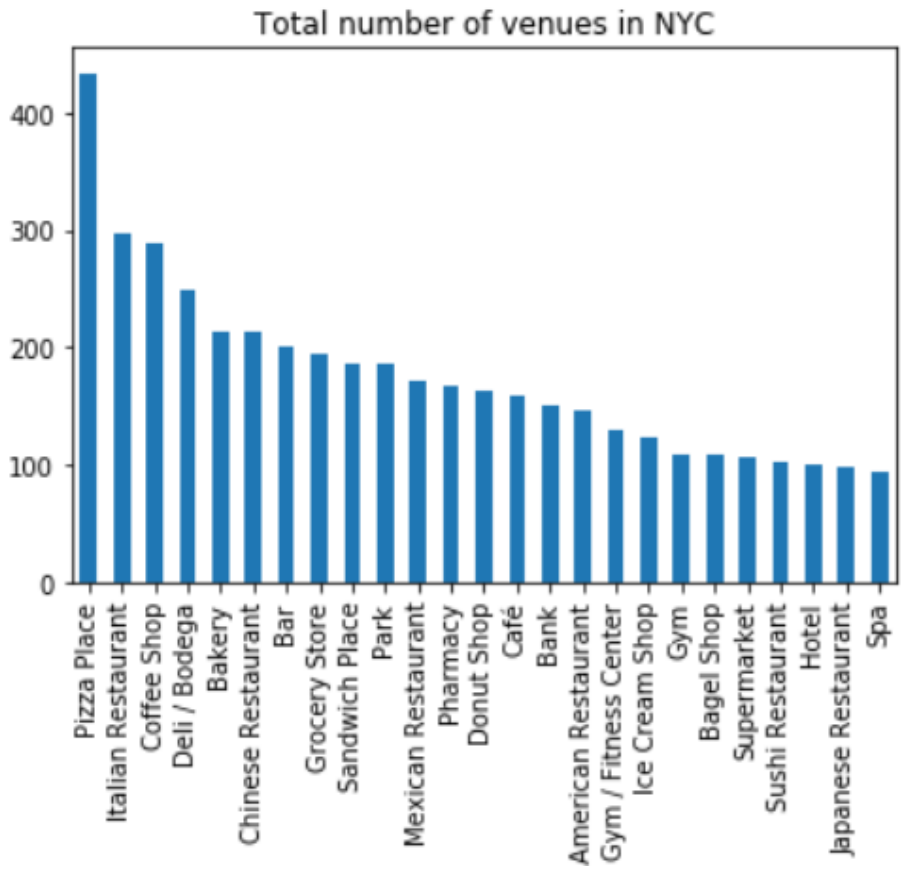




New York City map (left) and Toronto map (right), showing the neighbourhood centroid locations in blue marks.

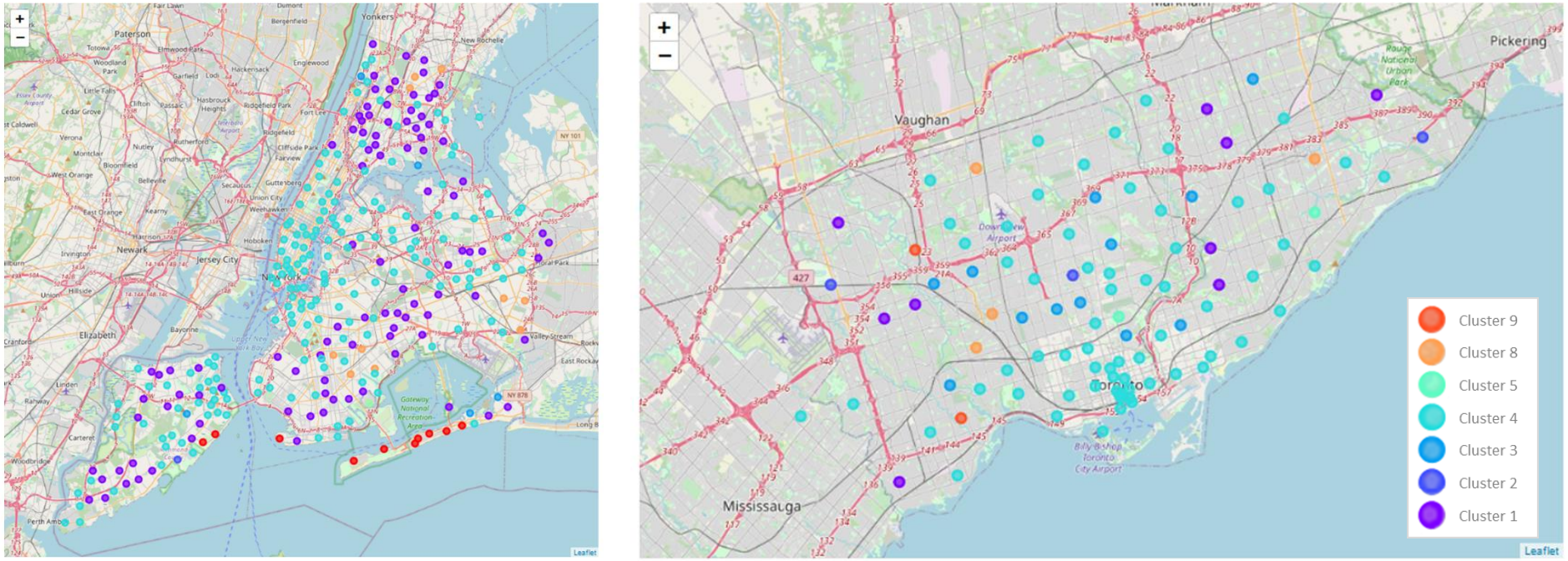
Data Exploration: Top 25 venues by city





Coffee shops, and pizza and Italian restaurants are top of both charts

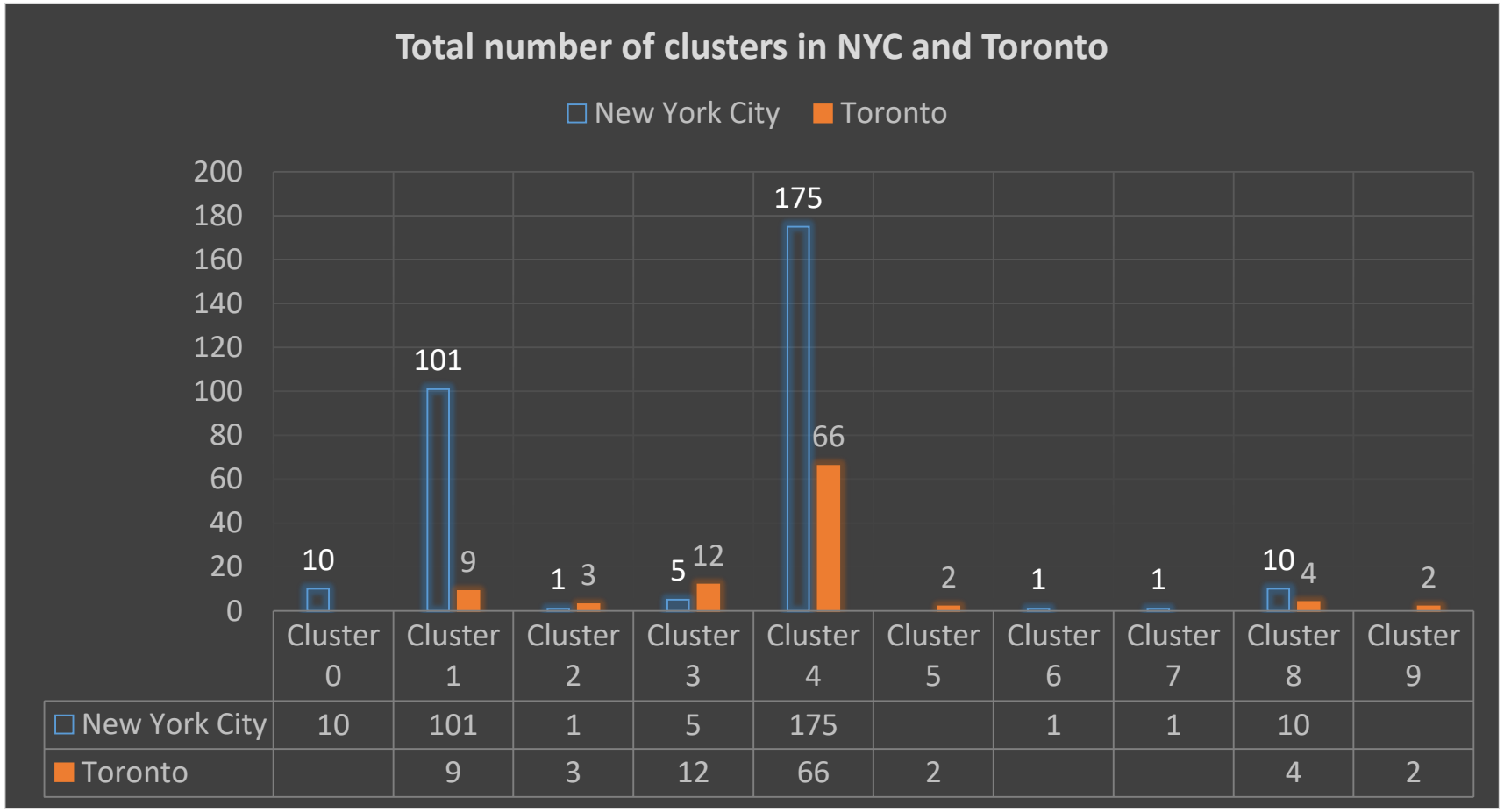
Results: K-Means Clustering (K=10)



Map of New York City clusters (left) and Toronto clusters (right)

Good similarity of clusters between the two cities.
Let's explore which ones stand out the most and what features defines them

Results: K-Means Clustering (K=10)



Most prominent clusters in Toronto, are: 4, 3 and 1

Cluster 4 has the most potential, as a good candidate for our client, since it can be seen NYC has its highest number of neighbourhoods also in this cluster.

Results: K-Means Clustering (K=10)

Venue types and counts per cluster

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

Common Cluster Profiles (between NYC and Toronto)

Cluster No.	Profile Characteristics	Diversity of Venues	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 Cluster Profiles (within 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

Conclusion and next steps

- Client to map out the areas in NYC they have the most experience and existing marketing materials to leverage.

- Then compare which neighbourhoods in Toronto, map most closely with these areas (using the cluster maps in this deck). The comparison should be done both against the colour coded areas of the maps, but also against the business knowledge the client has relative to the descriptions of the cluster profiles.

- 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.