Battle of Neighborhoods: Toronto vs New York

Velvizhi Viswalingam

April 23, 2020

1. Introduction

1.1 Background

Toronto is the provincial capital of Ontario and the most populous city in Canada. Its economy is highly diversified with strengths in technology, design, financial services, life sciences, education, arts, fashion, aerospace, environmental innovation, food services, and tourism. The diverse population of Toronto reflects its current and historical role as an important destination for immigrants to Canada.

New York is also the most densely populated major city in the United States. New York City has been described as the cultural, financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports.

On the outset, both the cities are densely populated and a major destination in both the countries. Therefore, it would be helpful to see how both cities are similarly distributed and their differences.

1.2 Problem Data

Data that might contribute to cluster them will be based on the neighborhoods that contain similar distribution of venues like Italian restaurants, coffee shops, and medical centers. This project aims to predict similarities and differences in the cities based on these data.

1.3 Interest

Obviously, Immigrants would be very interested to know about the cities before their move and decide a place according to their needs and interest. Others who are interested like travelers or visitors would also love to know about the city to choose their stay.

2. Data acquisition and cleaning

2.1 Data sources

Clustering will be based on the categories of venues in the neighborhoods. So, we need data that specifies the venues in the neighborhoods and their categories for New York City and Toronto.

Neighborhood data for New York: https://cocl.us/new_york_dataset

Neighborhood data for Toronto:
https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_H&oldid=945633050.

Geo Spatial data for Toronto: https://cocl.us/Geospatial_data

Foursquare API

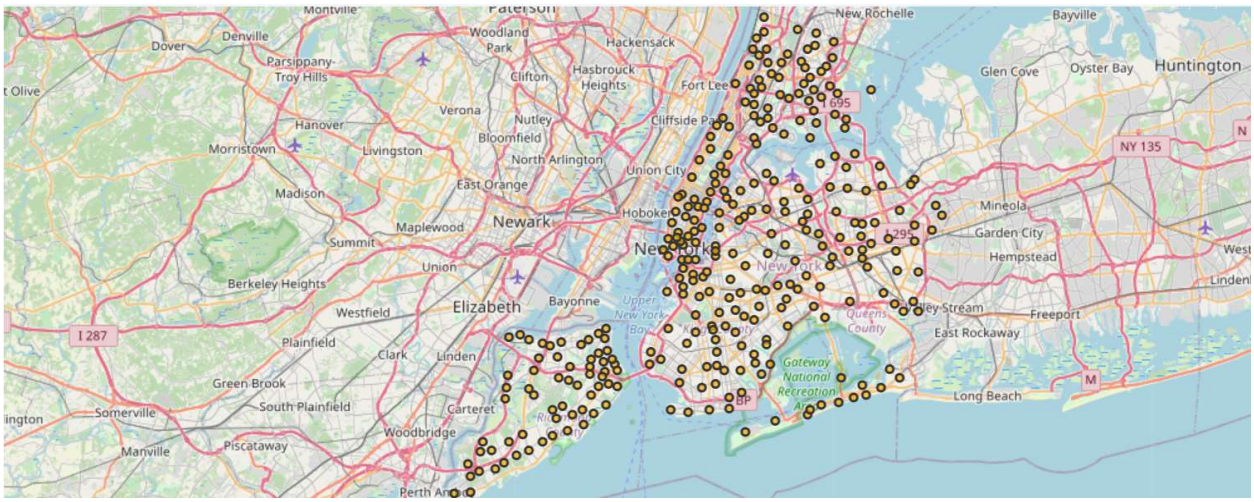
Foursquare is one of the world largest sources of location and venue data. To retrieve the venues and their categories in a given neighborhood, the coordinates—the latitude and the longitude—of the neighborhood are sent in the API request. We will using this API to get data on venues around these neighborhood in these locations to compare further.

2.2 Data cleaning

Data downloaded or scraped from multiple sources were combined into one table for New York and Toronto respectively.

2.2.1 New York

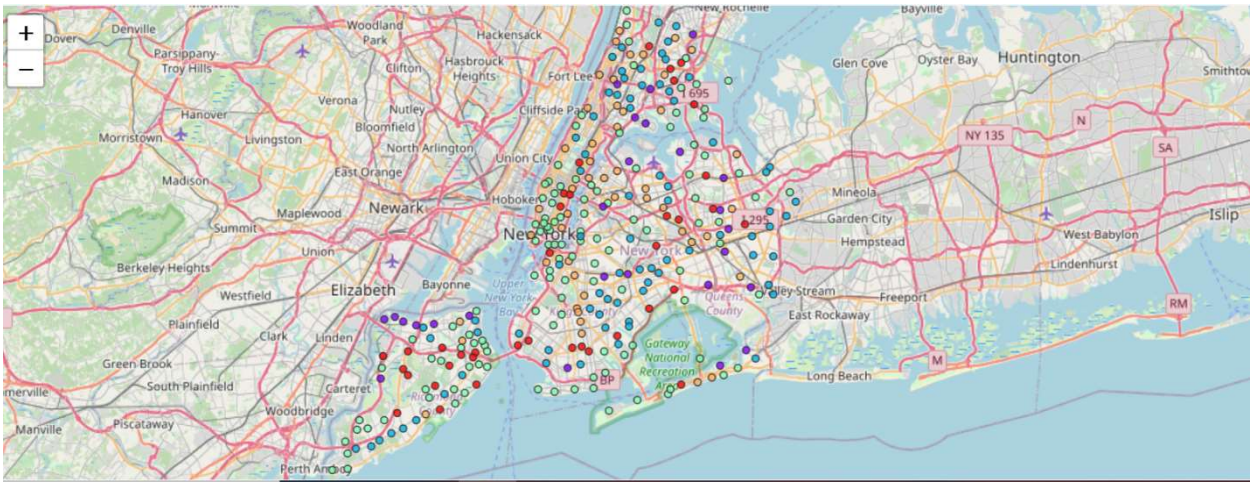
Firstly, the neighborhood, longitude and latitude data for New York was extracted from JSON to a data frame with 5 boroughs and 306 neighborhoods. We used the Geopy library to get the coordinates for New York City and mapped it as below.



Later, we used Foursquare API to get the nearby venues in the neighborhoods using their coordinates.

Dropped features	Reason for dropping features
'Building', 'Office', 'Bus Line', 'Bus Stop', 'Bus Station', 'Road'	It will not be helpful in checking similarities of the cities with these features like building, office, bus line etc…

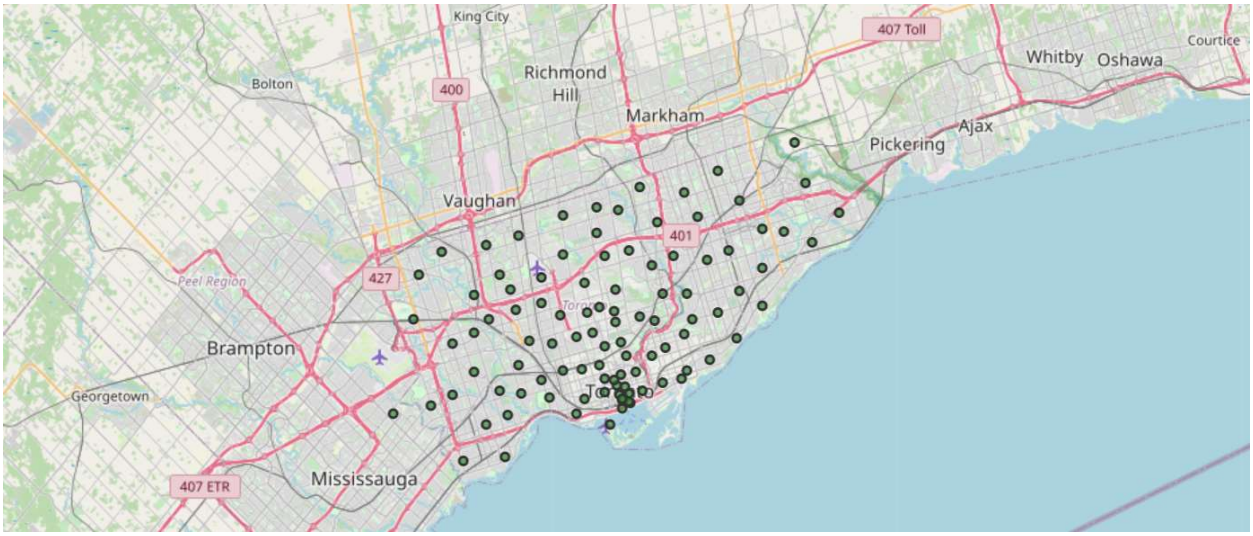
We had 580 unique categories and we are aggregating the values for each neighborhood so that they can be represented in one row. The aggregation will be done by grouping rows by neighborhood and by taking the mean of the frequency of occurrence of each category. We have mapped the data using Folium as below.



2.2.2 Toronto

Like New York City dataset, we didn't have dataset available for Toronto. So we had to extract from Wikipedia and Geospatial data.

First, we extracted the data from Wikipedia with pandas to get the postal codes, borough and neighborhood data of Toronto. We dropped "not assigned" borough and neighborhood. We also merged neighborhood with the same postal code and same borough to remove multiple entries of postal cod. Then we extracted the longitude and latitude data for corresponding postal code with geospatial data. Finally we merged these data as one table for Toronto with 103 rows. We used the Geopy library to get the coordinates for Toronto and mapped it as below.

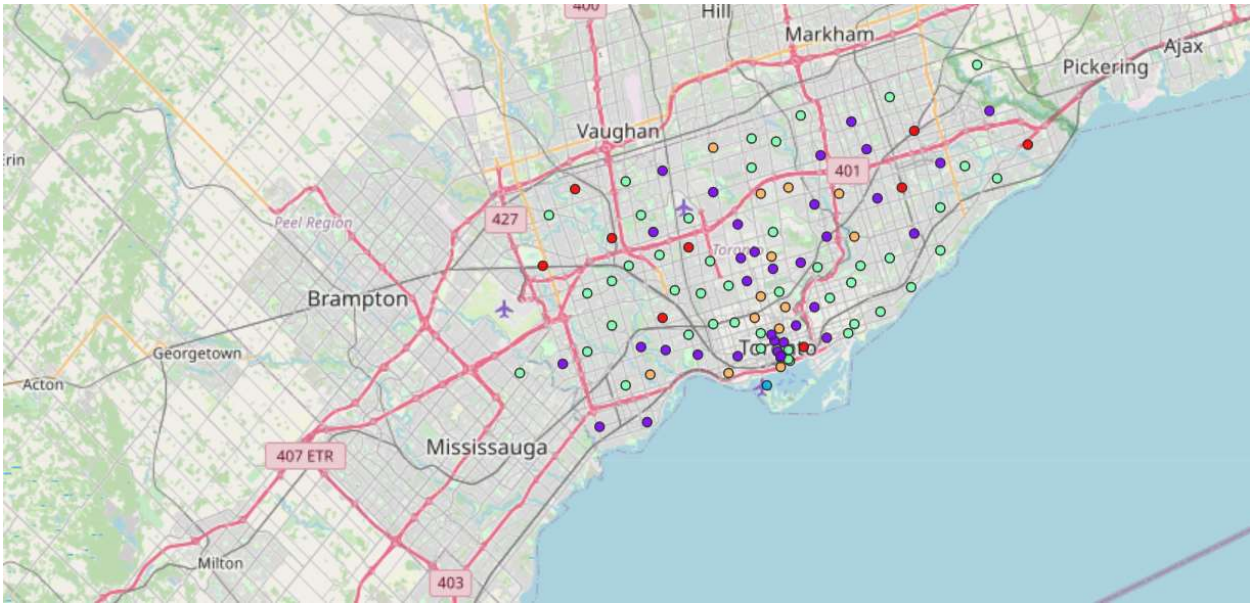


Later, we used Foursquare API to get the nearby venues in the neighborhoods using their coordinates.

Dropped features	Reason for dropping features
Not assigned	We cannot find data without neighborhood information

'Building', 'Office', 'Bus Line', 'Bus Stop', 'Bus Station', 'Road'	It will not be analytical in checking similarities of the cities with these features like building, office, bus line.

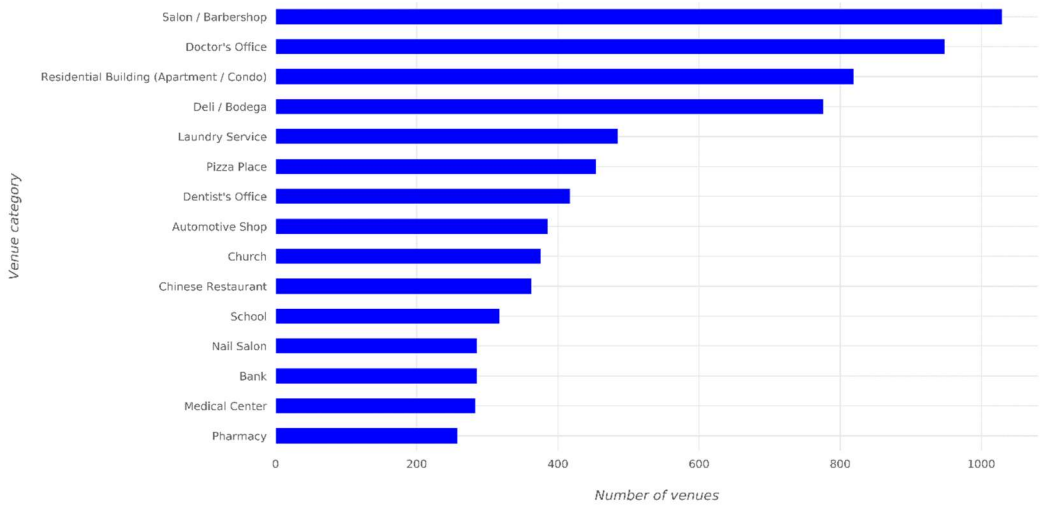
Finally, we had 502 unique categories and we are aggregating the values for each neighborhood so that they can be represented in one row. The aggregation will be done by grouping rows by neighborhood and by taking the mean of the frequency of occurrence of each category. We have mapped the data using Folium as below.



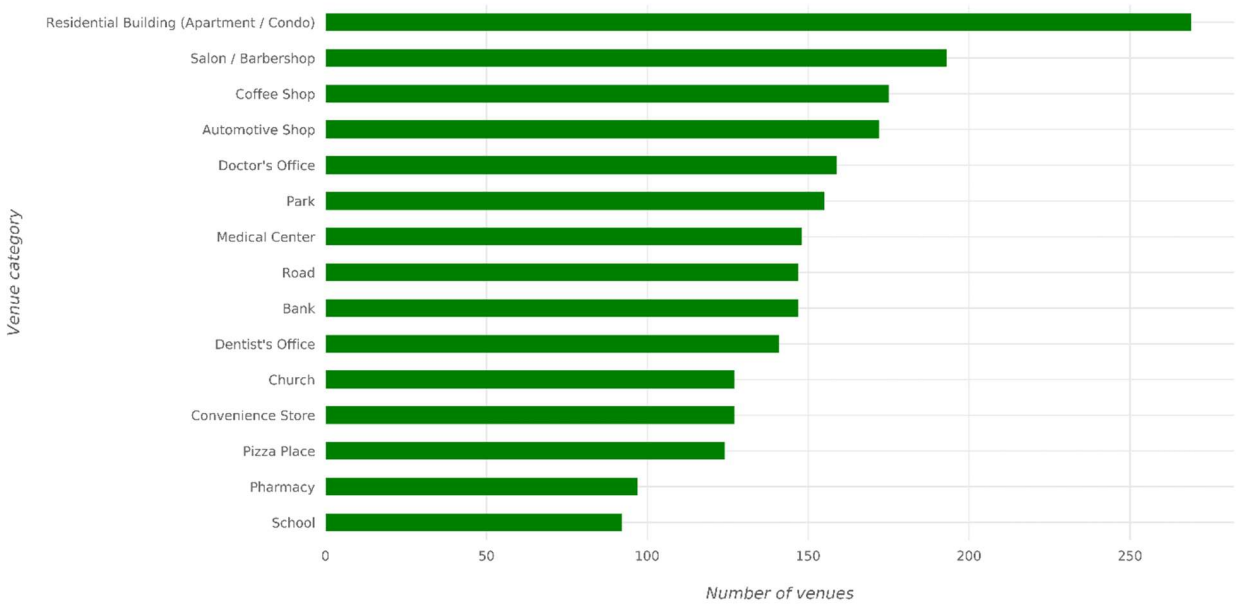
3. Exploratory Data Analysis

In this analysis, we found the most common and the most widespread venue categories in NYC and Toronto.

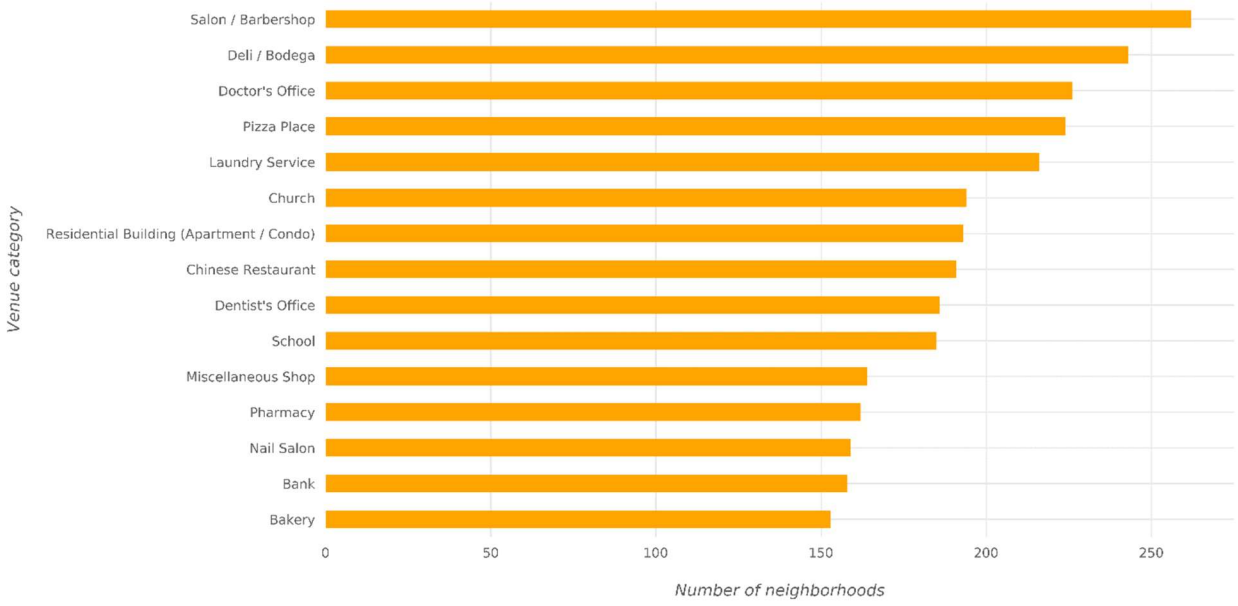
The figure below shows the most common venue categories in NYC.



The figure below shows the most common venue categories in Toronto.



The figure below shows the most widespread venue categories in NYC.



The figure below shows the most widespread venue categories in Toronto.

4. Cluster Neighborhoods

Before we cluster the neighborhoods, we need to aggregate the data of the both the cities and adding a suffix to the names to the neighborhoods to distinguish them. The figure below shows a sample of the combined dataframe.

	Neighborhood_	Accessories Store	Acupuncturist	Adult Boutique	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate
303	Agincourt_Toronto	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
304	Agincourt North, L'Amoreaux East, Milliken, St...	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
305	Albion Gardens, Beaumond Heights, Humbergate, ...	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
306	Alderwood, Long Branch_Toronto	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
307	Bathurst Manor, Downsview North, Wilson Height...	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
308	Bayview Village_Toronto	0.0	0.012195	0.0	0.0	0.0	0.0	0.0	0.0	0.0

The clustering algorithm used here is the K-means algorithm of Scikit-learn package and number of clusters chosen is 5 clusters. The output of clustering is a label for each neighborhood indicating to which cluster this neighborhood belongs. The figure below shows a sample of a dataframe created with the cluster labels.

	Cluster Labels	1st Most Common Category	2nd Most Common Category	3rd Most Common Category	4th Most Common Category	5th Most Common Category	6th Most Common Category	7th Most Common Category
Neighborhood_								
Woodside_NYC	3	Bar	Salon / Barbershop	Mexican Restaurant	Platform	Thai Restaurant	Miscellaneous Shop	Deli / Bodega
Yorkville_NYC	4	Residential Building (Apartment / Condo)	Laundry Service	Spa	Gym	Salon / Barbershop	Flower Shop	Pharmacy
Adelaide, King, Richmond_Toronto	0	Café	Coffee Shop	Food Court	Vegetarian / Vegan Restaurant	Ballroom	Restaurant	Indian Restaurant
Agincourt_Toronto	1	Automotive Shop	Chinese Restaurant	Doctor's Office	Church	Coffee Shop	Storage Facility	Furniture / Home Store
Agincourt North, L'Amoreaux East, Milliken, Steeles East_Toronto	0	School	Chinese Restaurant	Road	BBQ Joint	Medical Center	Doctor's Office	Pizza Place
Albion Gardens, Beaumond Heights, Humbergate, Jamestown, Mount Olive, Silverstone, South Steeles, Thistletown_Toronto	3	Salon / Barbershop	Movie Theater	Pizza Place	Farm	Art Gallery	Bakery	Clothing Store

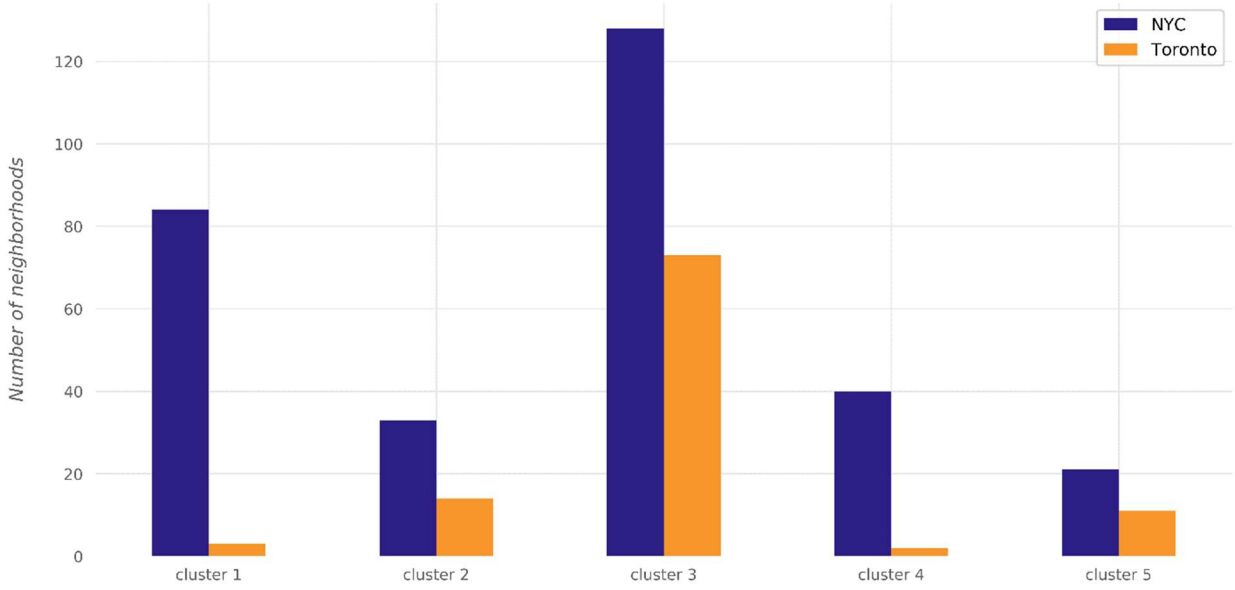
We added columns that show the most common venue categories in each neighborhood along with its cluster label.

4.1 Cluster Analysis

The table below shows the number of neighborhoods in each cluster.

Cluster	No. of neighborhoods
1	190
2	36
3	45
4	86
5	48

And the plot below shows the number of NYC neighborhoods and the number of Toronto neighborhoods in each cluster.



The differences between the clusters can be seen from the figure; each cluster distinguishably has different distribution of common venue categories than other clusters. Some of the observations that can be made are:

Cluster 1:

Category	% of venues
Salon / Barbershop	2.53212
Doctor's Office	2.29907
Residential Building (Apartment / Condo)	1.91484
Pizza Place	1.73847
Deli / Bodega	1.68808
Dentist's Office	1.6188
Bank	1.55581

Cluster 2:

Category	% of venues
Residential Building (Apartment / Condo)	13.8705
Doctor's Office	3.33813
Salon / Barbershop	2.47482
Laundry Service	2.35971
Deli / Bodega	2.27338
Dentist's Office	2.10072
Park	2.01439

Cluster 3:

Category	% of venues
Automotive Shop	10.5263
Gas Station	2.42915
Church	2.26721
Deli / Bodega	2.10526
Pizza Place	2.02429
Factory	1.98381
Salon / Barbershop	1.94332

Cluster 4:

Category	% of venues
Doctor's Office	15.0976
Dentist's Office	4.25384
Residential Building (Apartment / Condo)	3.90516
Medical Center	3.20781
Salon / Barbershop	2.82427
Deli / Bodega	2.16179
Laundry Service	1.88285

Cluster 5:

Category	% of venues
Salon / Barbershop	8.63053
Deli / Bodega	4.80742
Laundry Service	2.72468
Doctor's Office	2.33951
Pizza Place	2.32525
Residential Building (Apartment / Condo)	2.26819
Church	2.11127

- While residential buildings constitute ~2% of venues in the neighborhoods of the first cluster, they constitute ~13% of the venues in the second cluster, ~4% of the venues in the fourth cluster, and 2% of the venues in the fifth cluster and completely missing in the third cluster.
- Salon/Barbershop appear in the most common category in all the clusters.
- Automotive shops appear in the most common category of the third cluster only also the most popular category in that cluster.
- Doctor and dentist offices constitute ~15% of fourth-cluster venues while they constitute only 2% to 3% of each of the first, second and fifth-cluster venues.

Other differences can also be observed.

5. Conclusions

In this project, the neighborhoods of NYC and Toronto were clustered in multiple groups based on the categories of the venues in these neighborhoods. The results shows that there are venue categories that are most common in some cluster than the others also the most common venue categories differ from one cluster to the other. If a deeper analysis is performed, it might result in discovering different style in each cluster based on the most common categories in the cluster.