Comparing and Clustering the Neighborhoods of New York City and Toronto

Introduction

In this post, I will talk about a project I've done in order to get IBM Data Science Professional Certificate. In this project, the neighborhoods of two important cities in the world—New York City (NYC) and Toronto—were clustered into groups of similar items. One can use the results of this analysis to understand the similarities and differences between the two cities neighborhoods.

The clustering will be based on the types (categories) of venues in the neighborhoods. Each of the resulting clusters should have neighborhoods that contain similar distribution of venues. For example, one cluster may contain neighborhoods that have many Italian restaurants, many coffee shops, and few medical centers; another cluster may contain neighborhoods with many residential buildings, many barbershops, and few Italian restaurants.

This post will explain the stages and results of this project broadly. For a more in-depth description, check the project report. The code used to perform this project is available as a Jupyter notebook.

Data Acquisition and Preparation

Since clustering will be based on the categories of venues in the neighborhoods, we need data that specifies the venues in the neighborhoods and their categories. Of course we, in the first place, need a list of the neighborhoods of NYC and Toronto.

The figure below shows a map of NYC with its neighborhoods represented as orange circles on the left side and a map of Toronto with its neighborhoods represented as green circles on the right side.

To acquire data on venues and their categories, Foursquare API is used. 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.

where search indicates the API endpoint used, client_id and client_secret are credentials used to access the API service and are obtained when registering a Foursquare developer account, v indicates the API version to use, ll indicates the latitude and longitude of the desired location, radius is the

maximum distance in meters between the specified location and the retrieved venues, and limit is used to limit the number of returned results if necessary.

The result of this data-acquisition-and-preparation stage is two tables (dataframes) that specify the neighborhoods and the venues of each of NYC and Toronto. Below is a part of the NYC table. In these tables, one neighborhood might take many rows depending on the number of retrieved venues for that neighborhood.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Shell	40.894187	-73.845862	Gas Station
1	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
2	Wakefield	40.894705	-73.847201	Pitman Deli	40.894149	-73.845748	Food
3	Wakefield	40.894705	-73.847201	Julio C Barber Shop 2	40.894165	-73.845748	Salon / Barbershop
4	Wakefield	40.894705	-73.847201	Public School 87	40.895331	-73.845918	School

Several data-preparation techniques were used to arrive at this table.

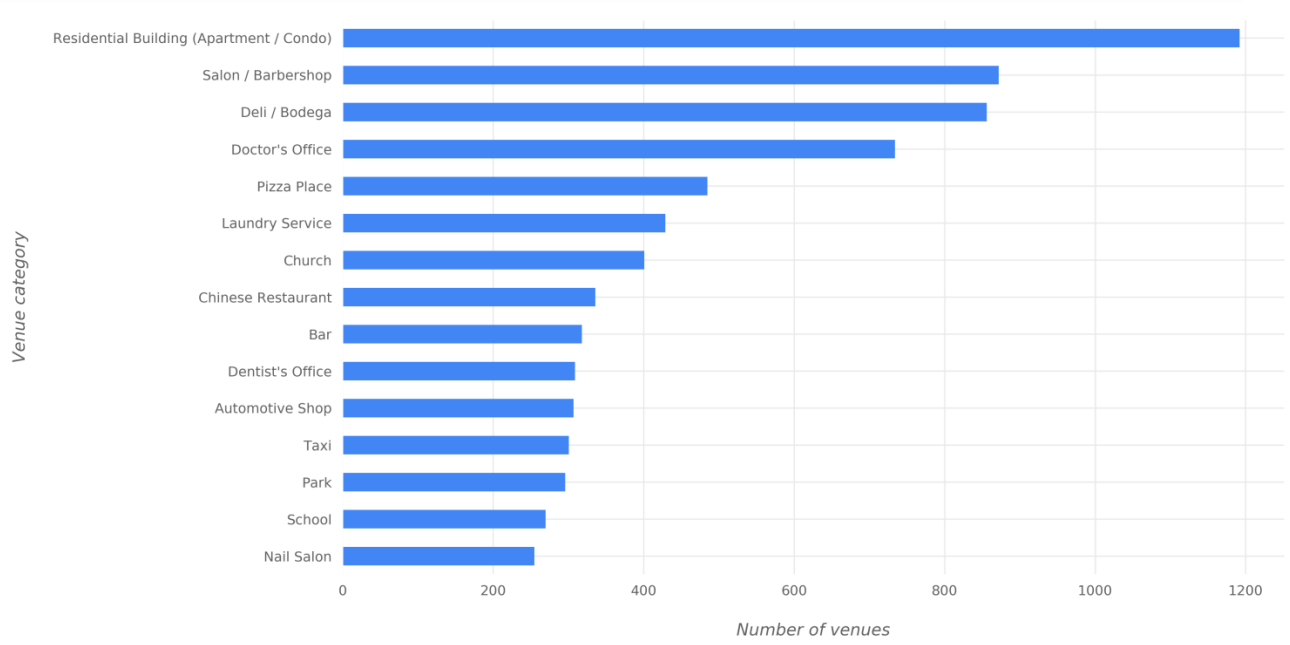
Exploratory Data Analysis

To get a better understanding of the venues data, we performed some exploratory analysis. In this analysis, we found the most common and the most widespread venue categories in NYC and Toronto.

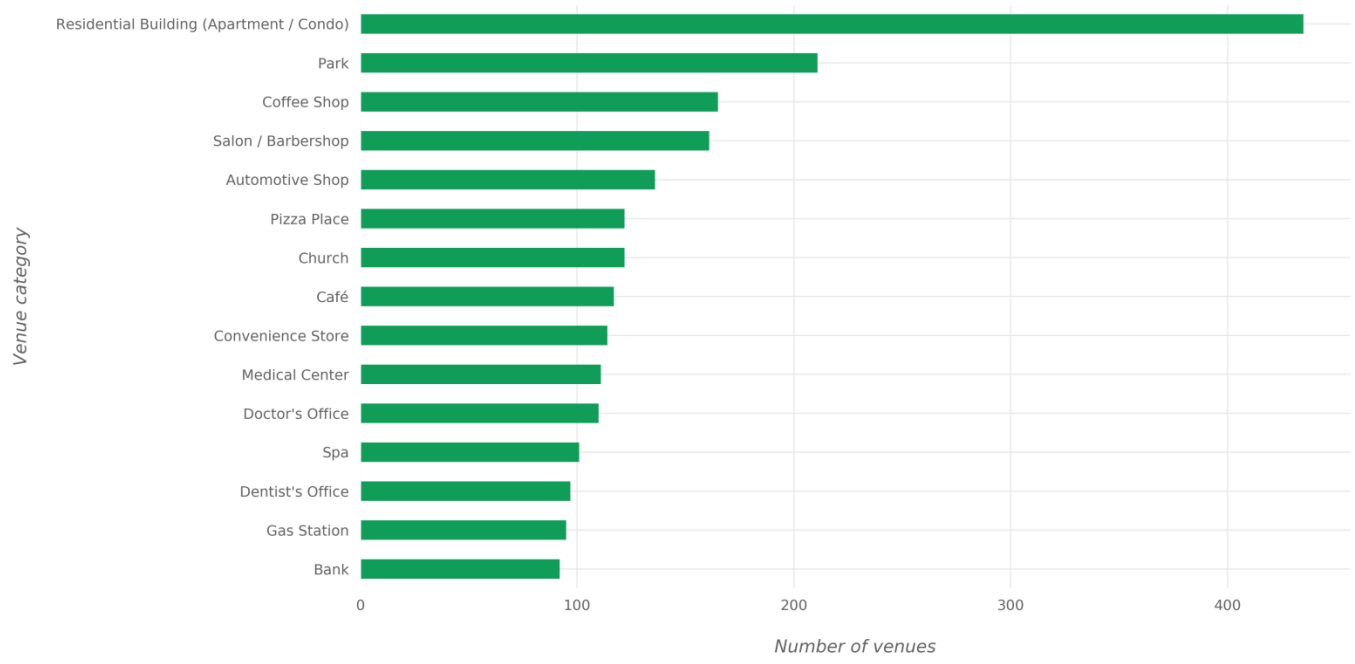
To explain the difference between “most common” and “most widespread”, we use an example: Suppose that there are 15 venues with the category “VR Games” and that these venues exist in 7 neighborhoods only out of 80 neighborhoods; also suppose that there are 10 venues with the category “Syrian Restaurant” and that these venues exist in 10 neighborhoods—each one of them in a different neighborhood. Then it can be said that the “VR Games” category is more com- mon than “Syrian Restaurant” category because there are more venues under this category, and it can be said that the “Syrian Restaurant” category is more widespread than the “VR Games” category because venues under this category exist in more neighborhoods than the other category.

Note: Before proceeding with exploratory data analysis and the subsequent steps, a data-preparation operation was performed: venues whose category is “Building”, “Office”, “Bus Line”, “Bus Station”, “Bus Stop”, or “Road” were excluded because they are not expected to add analytical value in this project.

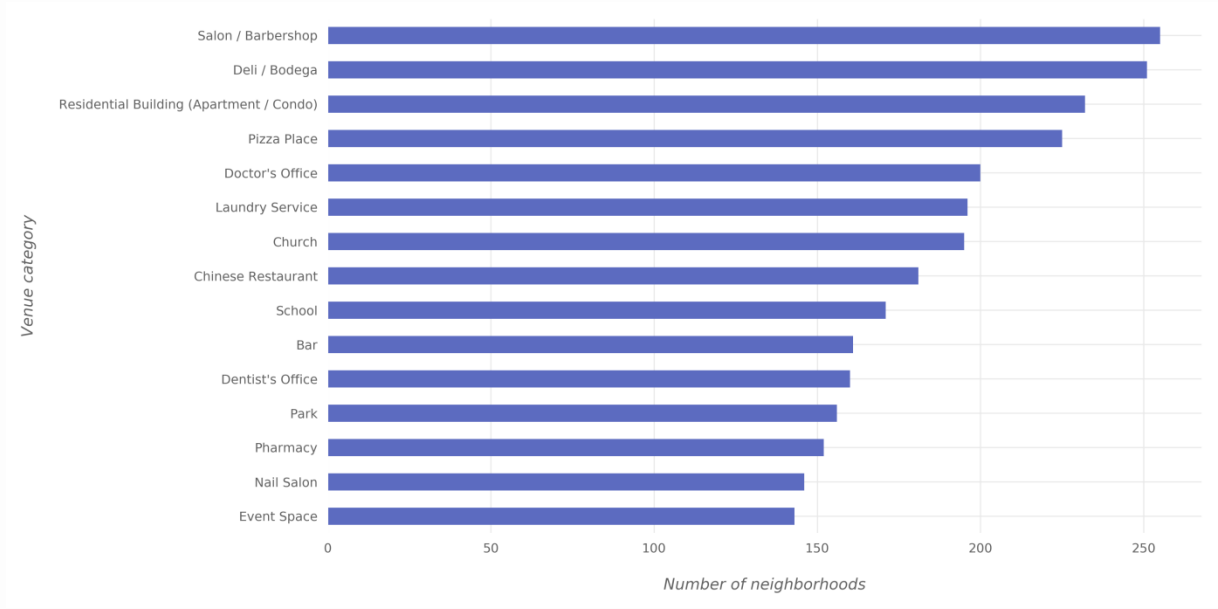
The figures below shows the most common venue categories in NYC.



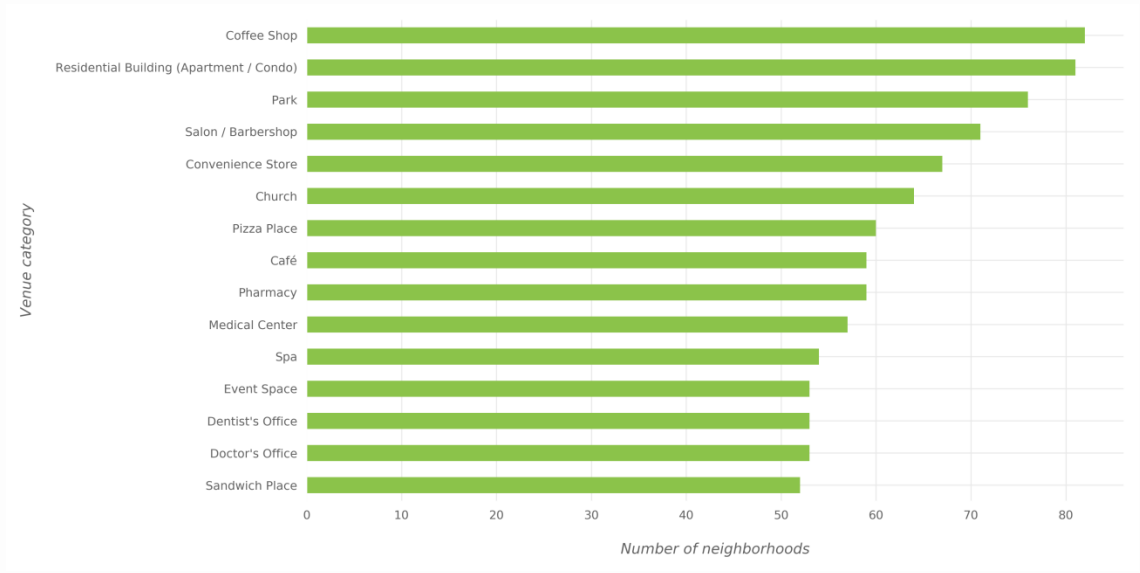
The figures below shows the most common venue categories in Toronto.



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



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



Clustering of Neighborhoods

To perform clustering, we need to feed the clustering algorithm with features in appropriate format. The data format that we saw above—show also below—is not suitable for the clustering algorithm.

	Neighborhood	Neighborhood Latitude	Neighborhoo
0	Wakefield	40.894705	
1	Wakefield	40.894705	
2	Wakefield	40.894705	
3	Wakefield	40.894705	
4	Wakefield	40.894705	

Since that we care most about the categories of the venues in the neighborhoods, one-hot encoding will be performed on the "Venue Category" field in the dataframe (table) shown above. A sample of the resulting dataframe of NYC is shown below.

	Neighborhood_	ATM	Accessories Store	Acupuncturist
0	Wakefield	0	0	0
1	Wakefield	0	0	0
2	Wakefield	0	0	0
3	Wakefield	0	0	0
4	Wakefield	0	0	0

The next step is aggregating the values for each neighborhood so that each neighborhood becomes represented by only 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. So for example, if the Fieldston neighborhood has 15 venues (i.e. 15 rows in the dataframe shown above) and 4 of these venues are of the "Sandwich Place" category, then Fieldston row in the aggregated dataframe will have the value 4/15 = 0.27 for the "Sandwich Place" column. The figure below shows the aggregated dataframe of NYC.

	Neighborhood_	ATM	Accessories Store	Acupuncturist
0	Allerton	0.0	0.0	0.0
1	Annadale	0.0	0.0	0.0
2	Arden Heights	0.0	0.0	0.0
3	Arlington	0.0	0.0	0.0
4	Arrochar	0.0	0.0	0.0

Now, we have two aggregated dataframes, one for NYC and another for Toronto. The next step is combining these two dataframes together while adding a suffix to the names of the neighborhoods to still be able to distinguish between NYC neighborhoods and Toronto neighborhoods. The figure below shows a sample of the combined dataframe.

	Neighborhood_	Accessories Store	Acupuncturist	Bo
303	Woodrow_NYC	0.0	0.0	
304	Woodside_NYC	0.0	0.0	
305	Yorkville_NYC	0.0	0.0	
306	Adelaide, King, Richmond_Toronto	0.0	0.0	
307	Agincourt_Toronto	0.0	0.0	
308	Agincourt North, L'Amoreaux East, Milliken, St...	0.0	0.0	

After that, clustering is applied. The clustering algorithm used is the K-means algorithm of Scikit-learn package and number of clusters is chosen to be 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.

Neighborhood_
Wingate_NYC
Woodhaven_NYC
Woodlawn_NYC
Woodrow_NYC
Woodside_NYC
Yorkville_NYC
Adelaide, King, Richmond_Toronto
Agincourt_Toronto
Agincourt North, L'Amoreaux East, Milliken, Steeles East_Toronto
Albion Gardens, Beaumond Heights, Humbergate

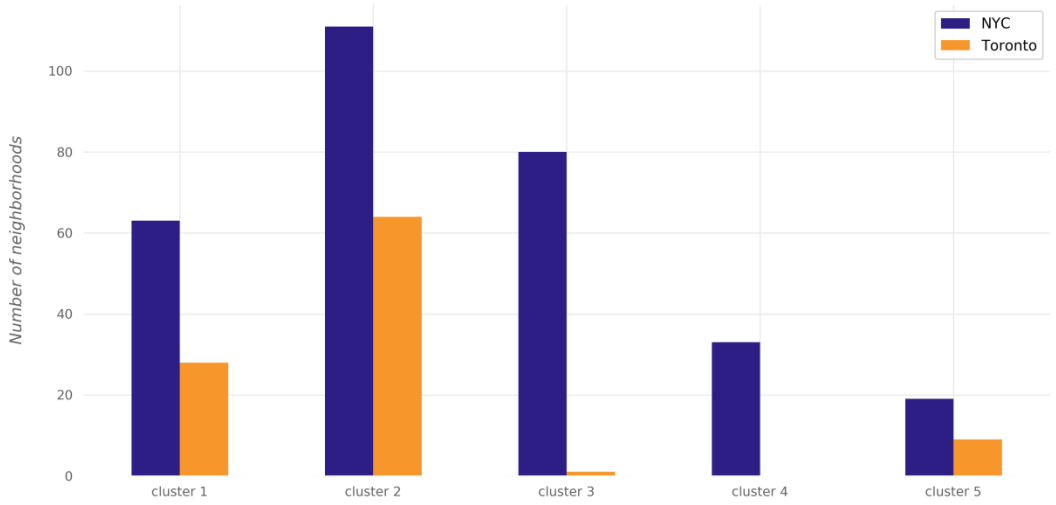
We added columns that show the most common venue categories in each neighborhood along with its cluster label.

Cluster Analysis

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

Cluster	Number of neighborhoods
0	91
1	175
2	81
3	33
4	28

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



The tables below show the most common venue categories in the neighborhoods of each cluster with the percentage of venues for each category.

Cluster 1:

Category	% of venues
Residential Building (Apartment / Condo)	9.50947
Deli / Bodega	2.92599
Salon / Barbershop	2.66781
Taxi	2.45267
Park	1.92197
Laundry Service	1.74986
Church	1.69248

Cluster 2:

Category	% of venues
Automotive Shop	2.19311
Salon / Barbershop	1.98353
Residential Building (Apartment / Condo)	1.88623
Pizza Place	1.86377
Park	1.85629
Doctor's Office	1.78144
Deli / Bodega	1.51198

Cluster 3:

Category	% of venues
Salon / Barbershop	7.48731
Deli / Bodega	5.5203
Laundry Service	2.61739
Pizza Place	2.60152
Church	2.58566
Residential Building (Apartment / Condo)	2.36358
Chinese Restaurant	2.22081

Cluster 4:

Category	% of venues
Doctor's Office	14.0034
Residential Building (Apartment / Condo)	4.4244
Dentist's Office	3.90893
Deli / Bodega	3.17869
Salon / Barbershop	2.70619
Medical Center	2.36254
Pizza Place	2.14777

Cluster 5:

Category	% of venues
Residential Building (Apartment / Condo)	21.7391
Doctor's Office	2.97732
Deli / Bodega	2.31569
Salon / Barbershop	2.22117
Park	2.22117
Laundry Service	1.93762
Dentist's Office	1.74858

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:

- While residential buildings constitute ~9% of venues in the neighborhoods of the first cluster, they constitute ~2% of the venues in the second and third clusters, ~4% of the venues in the fourth cluster, and 21% of the venues in the fifth cluster.
- Pizza places appear in the most common categories of the second, third, and fourth clusters only.
- Chinese restaurants appear in the most common categories of the third cluster only.
- Automative shops appear in the most common categories of the second cluster only; moreover, “Automative Shop” is the most popular category in that cluster.

- Doctor and dentist offices constitute ~18% of fourth-cluster venues while they constitute only 2% of each of the second-cluster and fifth-cluster venues.

Other differences can also be observed.