The Battle of Neighborhoods: New York City and Toronto

Soumya Ranjan Behera

11th February, 2021

Part I: Introduction

In this project, we will be studying about the neighborhoods of two of most multicultural and cosmopolitan cities in the world: New York City (NYC), United States of America and Toronto, Canada. We will be investigating on what kinds of businesses are common in both cities, what kinds of businesses are more common in one of the two cities than the other city, and what kinds of businesses are not common in both cities.

Doing this project will enable us to get a better understanding of similarities and differences between the two cities which will make it known to business people what types of businesses are more likely to thrive in both cities, what are the neighborhoods that are suitable for each type of business, and what types of businesses are not very desirable in each city. This allows business people to take better and more effective decisions regarding where to open their businesses.

Part II: Data for the Project

In this section, we will be discussing about the dataset used in this project. To be able to do this project, two types of data are needed:

- **Neighborhood Data:** Lists the names of the neighborhoods of NYC and Toronto and their latitude and longitude coordinates. We have this data provided through the "IBM Data Science Professional Certificate" course and also, we need to scrape some data from the internet.
- **Venues Data:** Lists the top 100 venues (restaurants, cafes, parks, museums, etc.) in each neighborhood of the two cities. The data should list the venues of each neighborhood with their categories. This data will be retrieved from Foursquare which is one of the world largest sources of location and venue data. Foursquare API will be utilized to get and download the data.

1. Neighborhood Data

For each city, data that describes the names of its neighborhoods and their coordinates is needed.

1.1 New York City

A dataset that specifies the neighborhood data for New York City was provided through the “Applied Data Science Capstone” course which is provided by IBM. The dataset is originally a JSON file that specifies the name of each neighborhood, its coordinates—latitude and longitude, its borough, and other data too. The below snippet shows a part of this JSON file.

Part of the JSON file containing NYC Neighborhood data

To be able to use the data of this JSON file in the later parts of this project, it would be processed and stored in a Pandas dataframe. In total, the JSON file contains data on 306 neighborhoods.

1.2 Toronto

For Toronto, there is no dataset that contains all the needed neighborhood data as was the case for NYC. In the “Applied Data Science Capstone” course, a dataset was provided mapping the Toronto postal codes to their respective latitude and longitude coordinates. Below is a snippet showcasing the same.

Postal Code	Latitude	Longitude
M1B	43.8066863	-79.1943534
M1C	43.7845351	-79.1604971
M1E	43.7635726	-79.1887115
M1G	43.7709921	-79.2169174
M1H	43.773136	-79.2394761
M1J	43.7447342	-79.2394761
M1K	43.7279292	-79.2620294
M1L	43.7111117	-79.2845772
M1M	43.716316	-79.2394761

Toronto's Postal Codes with their coordinates

We will be needing another dataset enlisting the names of the neighborhoods and their postal codes which can be used in combination with the above dataset to produce the desired results. There is a Wikipedia page titled "List of postal codes of Canada: M" containing the information about the postal codes with neighborhood and borough name associated with it. The postal codes in Canada that starts with the letter M are the postal codes of Toronto city. We will be extracting the relevant data from the webpage with the Pandas read_html() function. The URL for the Wikipedia page is 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'.

Postal Code	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront
M6A	North York	Lawrence Manor, Lawrence Heights

Toronto's Postal Codes with Neighborhood and Borough names from Wikipedia

In the above list on the Wikipedia page, there are 77 records out of 180 where the "Borough" variable has the value "Not assigned"; for these 77 records, the "Neighborhood" variable also has the value "Not assigned"; In the above figure, the first two rows shows some examples

of these records. Thus, these records will be deleted because they don't carry meaningful information regarding Toronto neighborhoods.

2. Venues Data

For both of the cities, we need data that provides information about the different venues in a particular neighborhood and the categories of these venues. Venues data will be retrieved from Foursquare which is a popular source of location data. Foursquare API service will be utilized to access and download venues data.

To retrieve data from Foursquare using their API, a URL should be prepared and used to request data related a specific location. An example URL is the following:

https://api.foursquare.com/v2/venues/search?&client_id=1234&client_secret=1234&v=20180605&ll=40.89470517661,-73.84720052054902&radius=500&limit=100

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 for Foursquare developer account, v indicates the API version to use (i.e, it should be the current date), 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.

A function is created that takes as input the names, latitudes, and longitudes of the neighborhoods, and returns a dataframe with information about each neighborhood and its venues. It creates an API URL for each neighborhood and retrieves data about the venues of that neighborhoods from Foursquare. Below is a screenshot of the data after retrieving venue data for each neighborhood from Foursquare.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
3	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
4	Regent Park, Harbourfront	43.65426	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant

Processed sample venue data for a neighborhood retrieved from Foursquare

2.1 New York City

From Foursquare, the data retrieved for the NYC contained more than 23,000 venues in it. For each venue, venue name, category, latitude, and longitude were retrieved. Different numbers of venues were found in different neighborhoods: For example, data about 81 venues were returned for Williamsburg, 86 venues for Yorkville neighborhood and 57 venues for Midtown. Each venue belongs to one of 578 unique categories.

2.2 Toronto

Similar to what has been done for NYC, data was retrieved from Foursquare describing the venues present in the different Toronto neighborhoods. We got information about 7,700 venues in Toronto. Different numbers of venues were found in different neighborhoods: for example, data about 84 venues were returned for Weston neighborhood and 70 venues in Lawrence Park neighborhood. Each venue belongs to one of 504 unique categories.

Part III: Analyzing the Data

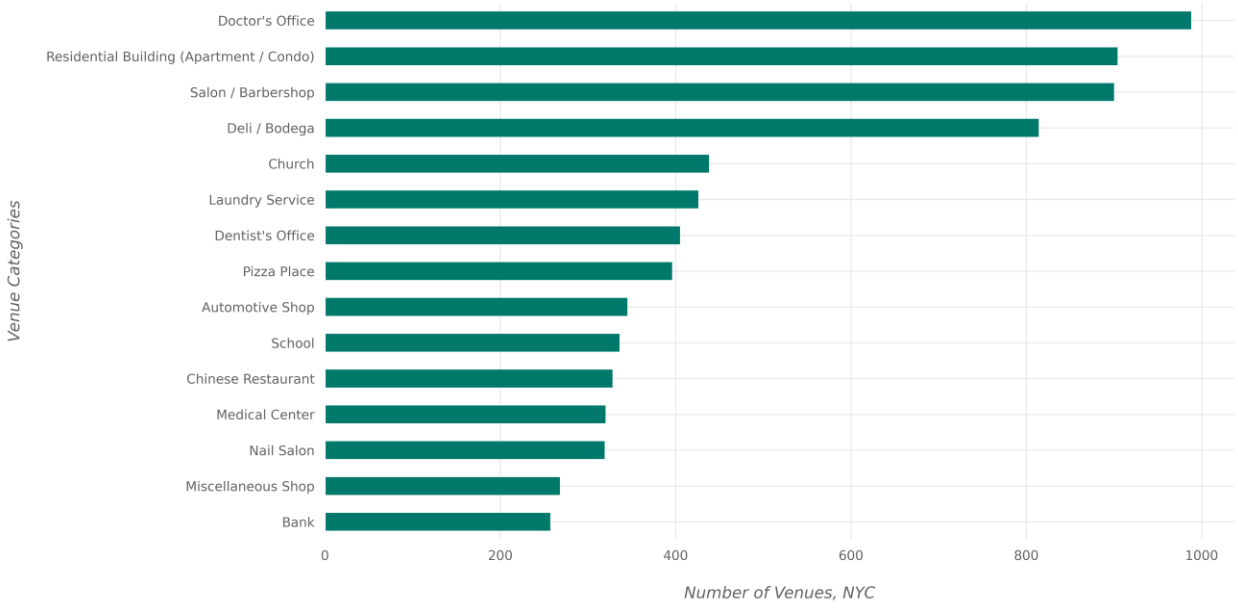
In this section, we will be exploring the data to understand it better through visualizations.

1. Most Common Venue Categories

What are the categories that have more venues than the others in NYC and Toronto? To answer this question, we will be plotting a bar chart to determine the popularity of the most common venue categories in each city.

1.1 New York City

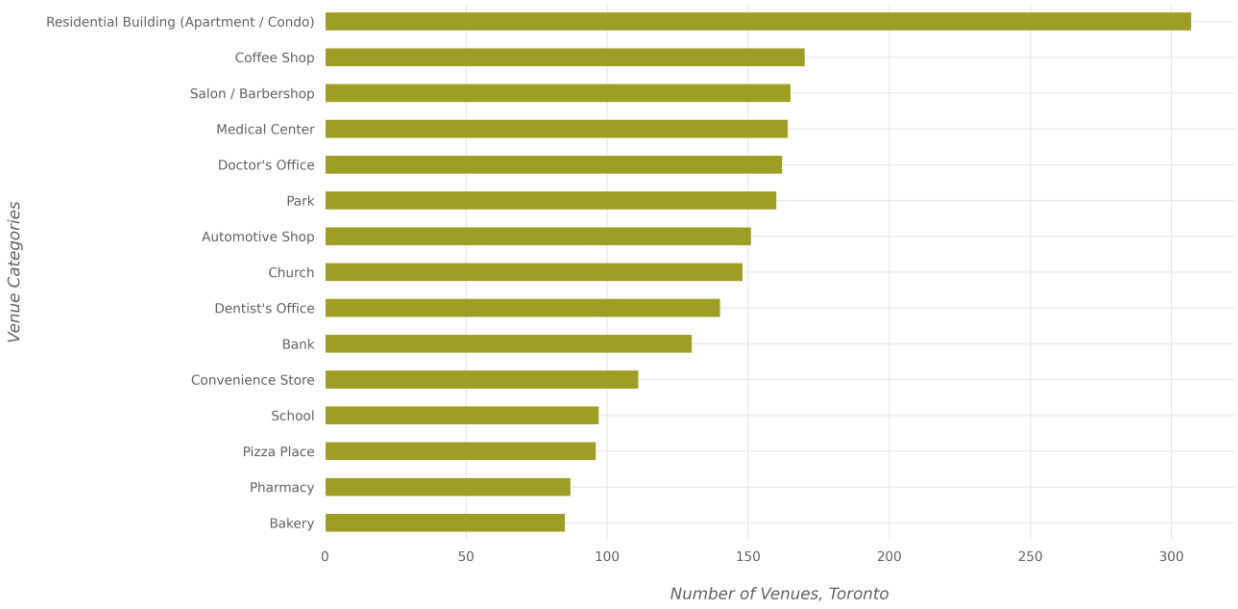
The below figure shows a bar plot of the most common venues in NYC. We can see that the most common category is “Doctor’s Office” with 988 venues in NYC. In the second rank, the category “Residential Building (Apartment / Condo)” appears with 904 venues. In the third rank comes the “Salon/Barbershop” category with a count of 900.



Most common venue categories in NYC

1.2 Toronto

The below figure shows a bar plot of the most common venues in Toronto. For Toronto, the most common venue category is “Residential Building (Apartment / Condo)” with 307 venues. Then comes “Coffee Shop” category with 170 venues. And in the third place appears “Salon/Barbershop” with around 165 venues.



Most common venue categories in Toronto

It can be seen that there are similarities between the most common categories in NYC and Toronto; we see many categories appearing in both the plots.

2. Most Widespread Venue Categories

Now another question is to be answered: What are the venue categories that exist in most number of neighborhoods?

This metric is different than the one mentioned in the earlier section. To explain the difference with an example, suppose that there are 15 venues with the category “Cafe” and that these venues exist in 7 out of

80 neighborhoods; also suppose that there are 10 venues with the category “Pharmacy” and that these venues exist in 10 out of the 80 neighborhoods. Then it can be said that the “Cafe” category is more commonly occurring than “Pharmacy” category because there are more venues under this category (15 vs. 10), and it can be said that the “Pharmacy” category is more widespread than the “Cafe” category because venues under this category exist in more neighborhoods (7 vs. 10) than the other.

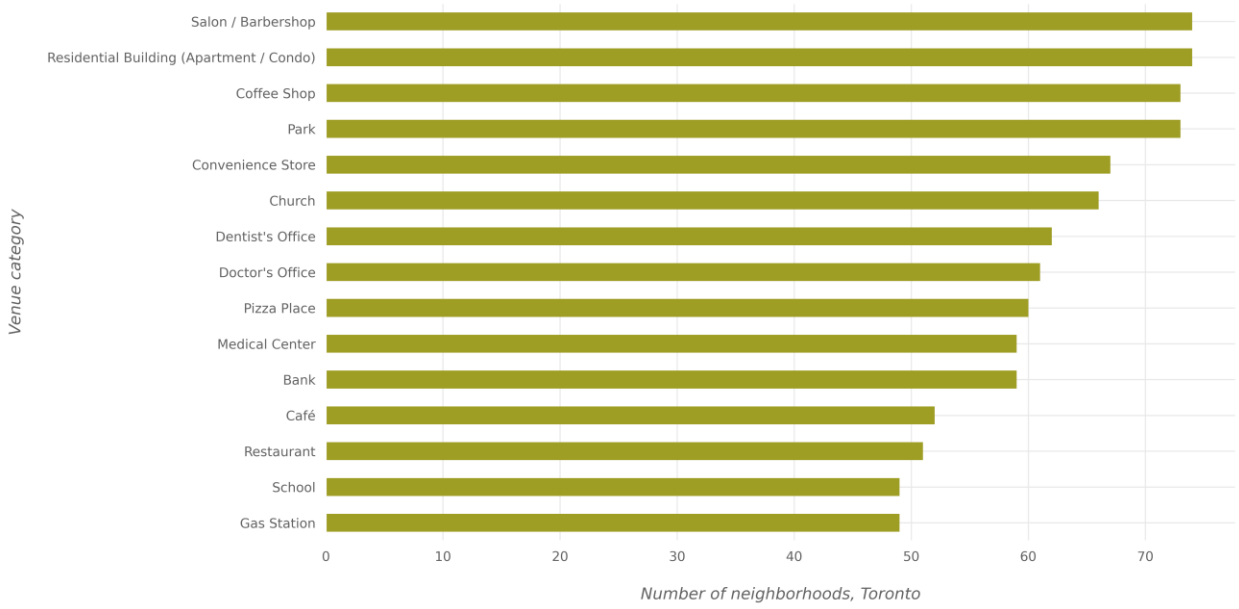
2.1 New York City

The below figure depicts the bar plot showcasing the most widespread venue categories in NYC. It can be seen that the order of categories in this figure is different than that of the most common categories. The most widespread category is “Salon / Barbershop”; Salons and barbershops exist in a total of 268 neighborhoods out of the 306 neighborhoods. After that comes the “Deli / Bodega” category with venues in a total of 246 neighborhoods also. In the third place comes the “Laundry Service” category with venues in 221 neighborhoods.

Most widespread venue categories in NYC

2.2 Toronto

The following bar plot shows the most widespread venue categories in Toronto. As with NYC, the order of the most-widespread-categories in Toronto differs than the order of the most common categories. For the first place, it is a tie between “Residential Building (Apartment / Condo)” and “Salon/Barbershop” with venues in a total of 74 out of 99 neighborhoods. Then comes “Coffee Shop” and “Park” with venues in 73 neighborhoods. The third most-widespread category is “Convenience Store” with venues in 67 neighborhoods.



Most widespread venue categories in Toronto

Part IV: Clustering of Neighborhoods

In this section, we will be applying clustering on the NYC and Toronto neighborhoods to find similar neighborhoods among them. For our purpose in particular, we will be using the K-means clustering algorithm of the scikit-learn python library. Before we apply our algorithm to find similar neighborhoods, the data collected till now has to be refined further to be considered ready.

1. Feature Selection

The goal of the clustering is to cluster neighborhoods based on the similarity of venue categories in the neighborhoods. This means that the two things of interest here are the neighborhood and the venue categories in the neighborhood. Thus, we will be selecting those two features from the dataset, i.e., “Neighborhood” and “Venue Category”. For the clustering algorithm to work, the categorical values needs to be converted into numerical features.

For that, one-hot encoding will be applied on the “Venue Category” feature and the result of the encoding will be used for our clustering algorithm. After applying one-hot encoding on NYC data, the resulting dataframe looks like the one shown below. For example, in the following figure we can see that for the last row corresponding to “Fox Hill” , the value for the column “African Restaurant” is 1 which tell us that the venue category for that particular row was “African Restaurant” in the previous dataframe before applying one-hot encoding; and the same applies for all the rows.

	Neighborhood_	ATM	Accessories Store	Acupuncturist	Adult Boutique	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Gate	Airport Service	Airport Terminal
26687	Fox Hills	0	0	0	0	0	0	0	0	0	0	0
26688	Fox Hills	0	0	0	0	0	0	0	0	0	0	0
26689	Fox Hills	0	0	0	0	0	0	0	0	0	0	0
26690	Fox Hills	0	0	0	0	0	0	0	0	0	0	0
26691	Fox Hills	0	0	0	0	0	0	1	0	0	0	0

The result of one-hot encoding on NYC data

The below figure shows the resulting dataframe for Toronto after applying the same one-hot encoding on the dataset.

	Neighborhood_	ATM	Accessories Store	Acupuncturist	Adult Boutique	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate
0	Malvern, Rouge	0	0	0	0	0	0	0	0	0	0
1	Malvern, Rouge	0	0	0	0	0	0	0	0	0	0
2	Malvern, Rouge	0	0	0	0	0	0	0	0	0	0
4	Malvern, Rouge	0	0	0	0	0	0	0	0	0	0
5	Malvern, Rouge	0	0	0	0	0	0	0	0	0	0

The result of one-hot encoding on Toronto data

The next step in refining the dataset is to aggregate the values for each neighborhood so that each neighborhood is represented by only one row in the dataframe. This is achieved by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each category. So for example, if the Allerton neighborhood has 15 venues and 4 of these venues are of the “Gas Station” category (i.e. the “Gas Station” column in the above defined one-hot encoded data for NYC has a value of 1 for four of Allerton rows), then Allerton row in the resulting aggregated dataframe will have the value 4/15 = 0.27 for the “Gas Station” column. The below two figures shows how the aggregated dataframe looks like for NYC Toronto.

	Neighborhood_	ATM	Accessories Store	Acupuncturist	Adult Boutique	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Gate
0	Allerton	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	Annadale	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Arden Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Arlington	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	Arrochar	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

A part of the aggregated dataframe for NYC

	Neighborhood_	ATM	Accessories Store	Acupuncturist	Adult Boutique	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate
0	Agincourt	0.0	0.0	0.0000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	Alderwood, Long Branch	0.0	0.0	0.0000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Bayview Village	0.0	0.0	0.0125	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	Bedford Park, Lawrence Manor East	0.0	0.0	0.0000	0.0	0.0	0.0	0.0	0.0	0.0	0.0

A part of the aggregated dataframe for Toronto

2. Combining NYC and Toronto Data

In the next step, we would be combining the two aggregated dataframes one each for NYC and Toronto generated in the previous section. However, in order to distinguish NYC neighborhoods from Toronto neighborhoods in the new dataframe, we will add a text string to the end of each neighborhood name before merging the dataframes, i.e., for NYC, we will append the string "_NYC" at the end of each neighborhood name and "_Toronto" for Toronto neighborhoods. Also, NYC and Toronto dataset don't necessarily have the same venue categories (i.e. some columns in the aggregated dataframe for NYC

won't exist in the corresponding dataframe for Toronto and vice versa). To deal with this issue before combining the dataframes, the columns of both dataframes are made the same by adding the columns that exist only in NYC dataframe to Toronto dataframe and vice versa; the newly added columns have a value of 0 for all the rows. The following figure shows a part of the dataframe that resulted from the combination of NYC and Toronto aggregated dataframes. This dataframes contains data on 405 neighborhoods in both NYC and Toronto. We will be using this resulting dataframe as our input data to the clustering algorithm, discussed in the later section.

	Neighborhood_	Accessories Store	Acupuncturist	Adult Boutique	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate
303	Woodrow_NYC	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
304	Woodside_NYC	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
305	Yorkville_NYC	0.0	0.011628	0.0	0.0	0.0	0.0	0.0	0.0	0.0
306	Agincourt_Toronto	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
307	Alderwood, Long Branch_Toronto	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
308	Bathurst Manor, Wilson Heights, Downsview Nort...	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0

The combination of NYC and Toronto aggregated dataframes

3. The Most Common Categories for Each Neighborhood

Using the combined NYC and Toronto aggregated dataframe we just created, another dataframe is created to specify the seven most common categories for each neighborhood in NYC and Toronto. This dataframe is created by retrieving the 7 categories with the largest values for each neighborhood in the previous dataframe, i.e., the mean of the frequency of the venue category count for each neighborhood. The below figure shows this dataframe.

	Neighborhood_	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
0	Allerton_NYC	Laundry Service	Doctor's Office	Deli / Bodega	Salon / Barbershop	Pizza Place	Dentist's Office	Pharmacy
1	Annadale_NYC	Salon / Barbershop	Pizza Place	Tattoo Parlor	American Restaurant	Nail Salon	Bakery	Doctor's Office
2	Arden Heights_NYC	Pool	Professional & Other Places	Dentist's Office	Salon / Barbershop	Doctor's Office	Food	Moving Target
3	Arlington_NYC	Church	Automotive Shop	Hardware Store	Professional & Other Places	Salon / Barbershop	Deli / Bodega	Laundry Service
4	Arrochar_NYC	Deli / Bodega	Laundry Service	Food Truck	Doctor's Office	Dry Cleaner	Liquor Store	Salon / Barbershop

Most common categories for each neighborhoods

4. Clustering and its Results

Now we are ready to apply the clustering algorithm to our dataset to find similar neighborhoods. The below snippet shows the code we have used to implement the K-means algorithm of scikit-learn python library. The variable named nyc_tor_grouped contains the combined dataframe containing NYC and Toronto aggregated data derived in the earlier section. We have dropped the “Neighborhood_” column from the dataframe before applying the clustering algorithm because the clustering algorithm doesn’t accept non-numerical columns as mentioned earlier. However, this column will be re-added later on.

```
# the number of clusters
kclusters = 5

nyc_tor_grouped_clustering = nyc_tor_grouped.drop('Neighborhood_', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(nyc_tor_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
2]: array([4, 3, 3, 0, 3, 3, 2, 3, 0, 3], dtype=int32)
```

Code used to perform K-means clustering

In the above snippet we can see the output the code executed, i.e., the clustering algorithm produced cluster-labels. These labels denote the cluster of each record (i.e. each neighborhood) in the data. Using these labels and the dataframe containing the seven most occurring venue categories for each neighborhood created in previous section, a dataframe is constructed to show the neighborhoods of NYC and

Toronto and the cluster to which each neighborhood belongs to. This dataframe can be seen in the below figure.

Neighborhood_	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
Wingate_NYC	4	Salon / Barbershop	School	Deli / Bodega	Food	Event Space	Caribbean Restaurant	Lounge
Woodhaven_NYC	4	Deli / Bodega	Salon / Barbershop	Laundry Service	Dentist's Office	Doctor's Office	Miscellaneous Shop	Chinese Restaurant
Woodlawn_NYC	4	Bar	Deli / Bodega	Pub	Salon / Barbershop	Food & Drink Shop	Pizza Place	Laundry Service
Woodrow_NYC	3	Pool	Grocery Store	Salon / Barbershop	School	Fish & Chips Shop	Physical Therapist	Art Gallery
Woodside_NYC	3	Bar	Salon / Barbershop	Platform	Thai Restaurant	Deli / Bodega	Mexican Restaurant	Miscellaneous Shop
Yorkville_NYC	2	Residential Building (Apartment / Condo)	Laundry Service	Spa	Flower Shop	Pharmacy	Gym	Construction & Landscaping
Agincourt_Toronto	0	Automotive Shop	Doctor's Office	Post Office	Storage Facility	Chinese Restaurant	Church	Hardware Store
Alderwood, Long Branch_Toronto	3	Gas Station	Dentist's Office	Bank	Medical Center	Salon / Barbershop	Conference Room	Convenience Store
Bathurst Manor, Wilson Heights, Downsview North_Toronto	1	Doctor's Office	Medical Center	Residential Building (Apartment / Condo)	Bank	Synagogue	Ice Cream Shop	Dentist's Office
Bayview Village_Toronto	3	Doctor's Office	Salon / Barbershop	Church	Residential Building (Apartment / Condo)	Park	Bank	Grocery Store

NYC and Toronto neighborhoods, their clusters, and the most common categories

The output of the clustering algorithm is five clusters with cluster labels 0, 1, 2, 3, and 4. Each cluster is expected to contain a group of similar neighborhoods based on the venue categories for each neighborhood. The clustering algorithm was run on a combined 405 neighborhoods in NYC and Toronto. The following table shows the number of neighborhoods in each cluster.

Cluster Label	Total Neighborhoods
0	31
1	46
2	42
3	204
4	82

Number of Neighborhoods in each Cluster

For examples, the following two figures shows a part of the first and third cluster, respectively. The details of all the neighborhoods

belonging to the five clusters can be accessed in the Jupyter notebook of this project.

Neighborhood_	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
Arlington_NYC	0	Church	Automotive Shop	Hardware Store	Salon / Barbershop	Professional & Other Places	Residential Building (Apartment / Condo)	Laundry Service
Auburndale_NYC	0	Automotive Shop	Train	Deli / Bodega	Nail Salon	Lawyer	Athletics & Sports	Residential Building (Apartment / Condo)
Blissville_NYC	0	Gas Station	Deli / Bodega	Hardware Store	Automotive Shop	Tech Startup	Factory	Donut Shop
Broadway Junction_NYC	0	High School	Automotive Shop	Metro Station	Deli / Bodega	Sandwich Place	Rental Car Location	Playground
College Point_NYC	0	Automotive Shop	Bank	Salon / Barbershop	Doctor's Office	Deli / Bodega	Church	Sandwich Place
Eastchester_NYC	0	Automotive Shop	Caribbean Restaurant	Auto Dealership	Deli / Bodega	Gas Station	Factory	Salon / Barbershop
Gravesend_NYC	0	Automotive Shop	Deli / Bodega	Salon / Barbershop	Pharmacy	Bakery	Bank	Pizza Place
Howland Hook_NYC	0	Boat or Ferry	Automotive Shop	Factory	Food	Bar	Storage Facility	Harbor / Marina
Hunts Point_NYC	0	Automotive Shop	Factory	School	Bank	Storage Facility	Hardware Store	Spanish Restaurant
Longwood_NYC	0	Automotive Shop	Train	Gas Station	Mexican Restaurant	Salon / Barbershop	Food	Church

Some records that belong to the first cluster

Neighborhood_	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
The Annex, North Midtown, Yorkville_Toronto	2	Residential Building (Apartment / Condo)	Speakeasy	Gym	Café	Playground	Sandwich Place	General Entertainment
Victoria Village_Toronto	2	Residential Building (Apartment / Condo)	Automotive Shop	Auto Dealership	Dentist's Office	Event Space	Government Building	School
Willowdale, Willowdale West_Toronto	2	Residential Building (Apartment / Condo)	Park	Medical Center	Bank	Synagogue	Doctor's Office	Pizza Place
York Mills West_Toronto	2	Residential Building (Apartment / Condo)	Church	Medical Center	Park	Government Building	Dentist's Office	Rental Car Location
York Mills, Silver Hills_Toronto	2	Park	Residential Building (Apartment / Condo)	Church	High School	Pool	School	Synagogue

Some records that belong to the third cluster

5. Cluster Analysis

The clustering algorithm grouped the neighborhoods of NYC and Toronto into five clusters based on the similarity between their venues. Now, we will deep dive a little into these clusters to see the most common categories in each of them. The below figure shows the seven most common venue categories in each cluster and their corresponding percentage of occurrence in the neighborhoods of that particular cluster.

Cluster 1

Category	% of venues
Automotive Shop	11.031579
Deli / Bodega	2.610526
Salon / Barbershop	2.273684
Church	2.189474
Gas Station	2.147368
Factory	1.894737
Pizza Place	1.852632

Cluster 2

Category	% of venues
Doctor's Office	14.406533
Dentist's Office	4.111986
Residential Building (Apartment / Condo)	3.645378
Medical Center	3.178769
Salon / Barbershop	2.799650
Deli / Bodega	2.303879
Laundry Service	1.574803

Cluster 3

Category	% of venues
Residential Building (Apartment / Condo)	14.831905
Doctor's Office	3.394858
Salon / Barbershop	2.669743
Deli / Bodega	2.570864
Laundry Service	2.406065
Dentist's Office	2.175346
Park	1.977587

Cluster 4

Category	% of venues
Salon / Barbershop	2.724830
Doctor's Office	2.143616
Residential Building (Apartment / Condo)	2.018624
Deli / Bodega	1.762390
Dentist's Office	1.649897
Medical Center	1.574902
Bank	1.518655

Cluster 5

Category	% of venues
Salon / Barbershop	8.802441
Deli / Bodega	5.491991
Laundry Service	2.807018
Doctor's Office	2.608696
Residential Building (Apartment / Condo)	2.349352
Church	2.334096
Chinese Restaurant	2.181541

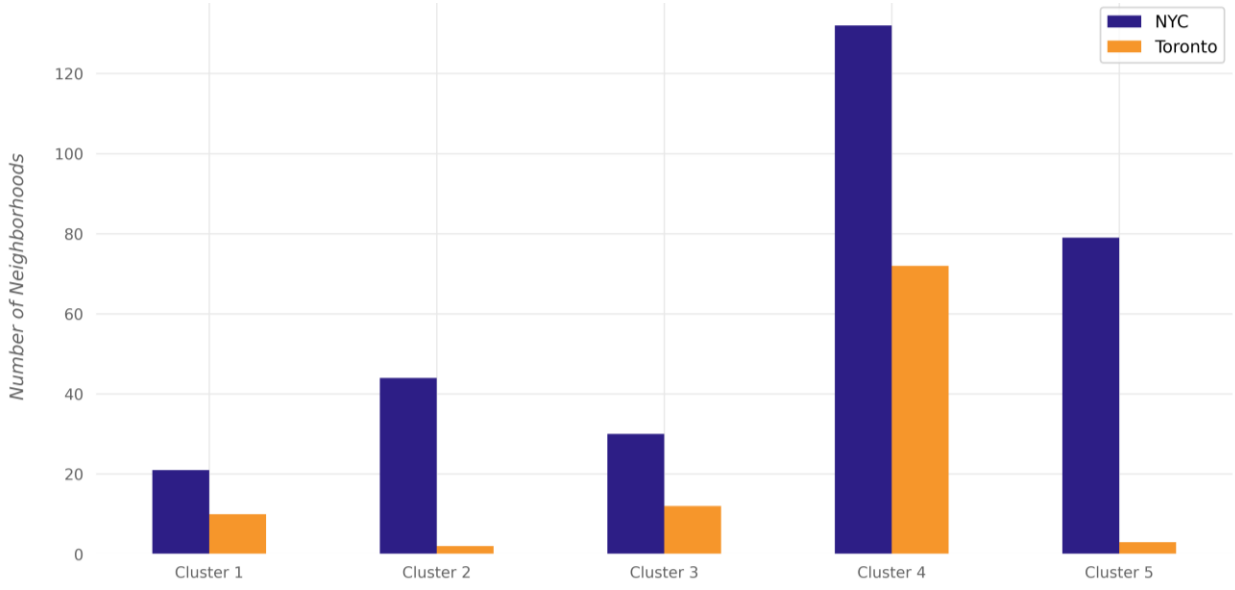
Most common venue-categories in each of the five clusters

The differences between the clusters can be seen from the above data; each cluster distinguishably has different distribution of common venue

categories than the other clusters. Some of the observations that can be made from the data are:

- While residential buildings constitute 14.83% of venues in the neighborhoods of the third cluster, they constitute 3.6% of the venues in the second cluster, 2.01 % in the third cluster and 2.34% of the venues in the fourth cluster.
- Salon/Barbershops appears at the top most common venue categories of the fourth and fifth clusters; along with appearing in the top seven categories of the rest of the clusters.
- Doctor's Office and Dentist's Office constitutes nearly 18 of the venue categories in the second cluster taking up the top two slots; while their percentages of occurrences in other clusters is not so high comparatively.
- Automotive shops is the most popular category for the first cluster; moreover, Automotive Shop appears only in the first cluster.

The following figure shows the number of NYC neighborhoods and the number of Toronto neighborhoods in each of the five resulting clusters.



The number of NYC and Toronto Neighborhoods in each cluster

Part V: Conclusion

In this project, the neighborhoods of New York City and Toronto were clustered into multiple groups based on the categories (types) of the venues in these neighborhoods. The results showed that there are venue categories that are more common in some cluster than the others; the most common venue categories differ from one cluster to the other. If a deeper analysis—taking more aspects into account—is performed, it might result in discovering different style in each cluster based on the most common categories in the cluster.