Clustering and Comparing the Neighborhoods of New York City and Toronto

By: Utkarsh Joshi

Part I

Introduction and Problem Statement

In this project, we will study, analyze, cluster, and compare the neighborhoods of two important cities in the world: New York City which is located in United States of America and Toronto which is located in Canada. We will investigate 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.

New York City (NYC) is one of the most most populous city in the United States of America. Also, NYC is the most linguistically diverse city in the world: as many as 800 languages are spoken in it. Moreover, NYC plays an essential role in the economics of USA: if New York City were a sovereign state, it would have the 12th highest GDP in the world. New York City consists of five boroughs: Brooklyn, Queens, Manhattan, The Bronx, and Staten Island.

Figure 1: Left: A picture of New York City. Right: A picture of Toronto

The second city of interest in this project is Toronto. As with NYC in USA, Toronto is the most populous city in Canada. It's recognized as one of the most multicultural and cosmopolitan cities in the world. Toronto also is a very diverse city: over 160 languages are spoken in it. On the economic side, Toronto is an international centre for business and finance and it is considered the financial capital of Canada.

Part II

Data Acquisition and Preparation

In this section, the processes of acquiring, cleaning, and preparing each dataset used in this project for next stages will be specified. To be able to do this project, two types of data are needed:

- **Neighborhood Data**: datasets that lists the names of the neighborhoods of NYC and Toronto and their latitude and longitude coordinates. We have some of this data provided by the coordinators of "IBM Data Science Professional Certificate" and we also need to scrape some data from the internet.
- **Venues data**: data that describes 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. For example:

Venue	Category
Los Mariscos	Seafood
Julio C Barber Shop 2	Salon / Barbershop

Table 1: Example of the venues data

  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 by the organizers of “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. Figure 2 shows a part of this JSON file.

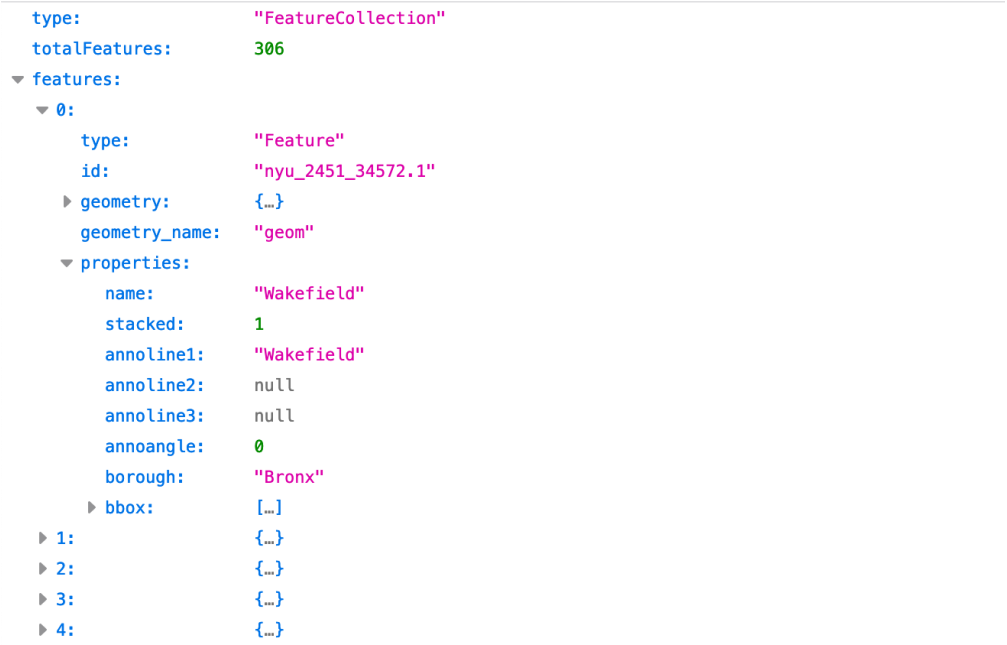

Figure 2: A part of the JSON file that describes NYC neighborhoods

To be able to use the data of this JSON file in the later parts of this project, it should be stored in a Pandas dataframe. Figure 3 shows the Python code used to process the JSON file data and store it in a dataframe named `nyc_neighborhoods`. Note that in the figure, the JSON file is stored in a variable named `nyc_neighborhoods_data`.

```
# define the dataframe columns
column_names = ['Borough', 'Neighborhood', 'Latitude', 'Longitude']

# instantiate the dataframe
nyc_neighborhoods = pd.DataFrame(columns=column_names)

for data in nyc_neighborhoods_data:
    borough = neighborhood_name = data['properties']['borough']
    neighborhood_name = data['properties']['name']

    neighborhood_latlon = data['geometry']['coordinates']
    neighborhood_lat = neighborhood_latlon[1]
    neighborhood_lon = neighborhood_latlon[0]

    nyc_neighborhoods = nyc_neighborhoods.append({'Borough': borough,
                                                  'Neighborhood': neighborhood_name,
                                                  'Latitude': neighborhood_lat,
                                                  'Longitude': neighborhood_lon}, ignore_index=True)
```

Figure 3: The code used to store NYC neighborhood data into a dataframe

Figure 4 shows the resulting dataframe which contains data on **306** neighborhoods.

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Figure 4: The NYC neighborhood-data dataframe

Having data of the coordinates of NYC neighborhoods, it is possible to draw a map using Folium Python package of NYC and its neighborhoods. Figure 5 shows this map; each orange circle represents the location of one neighborhood.

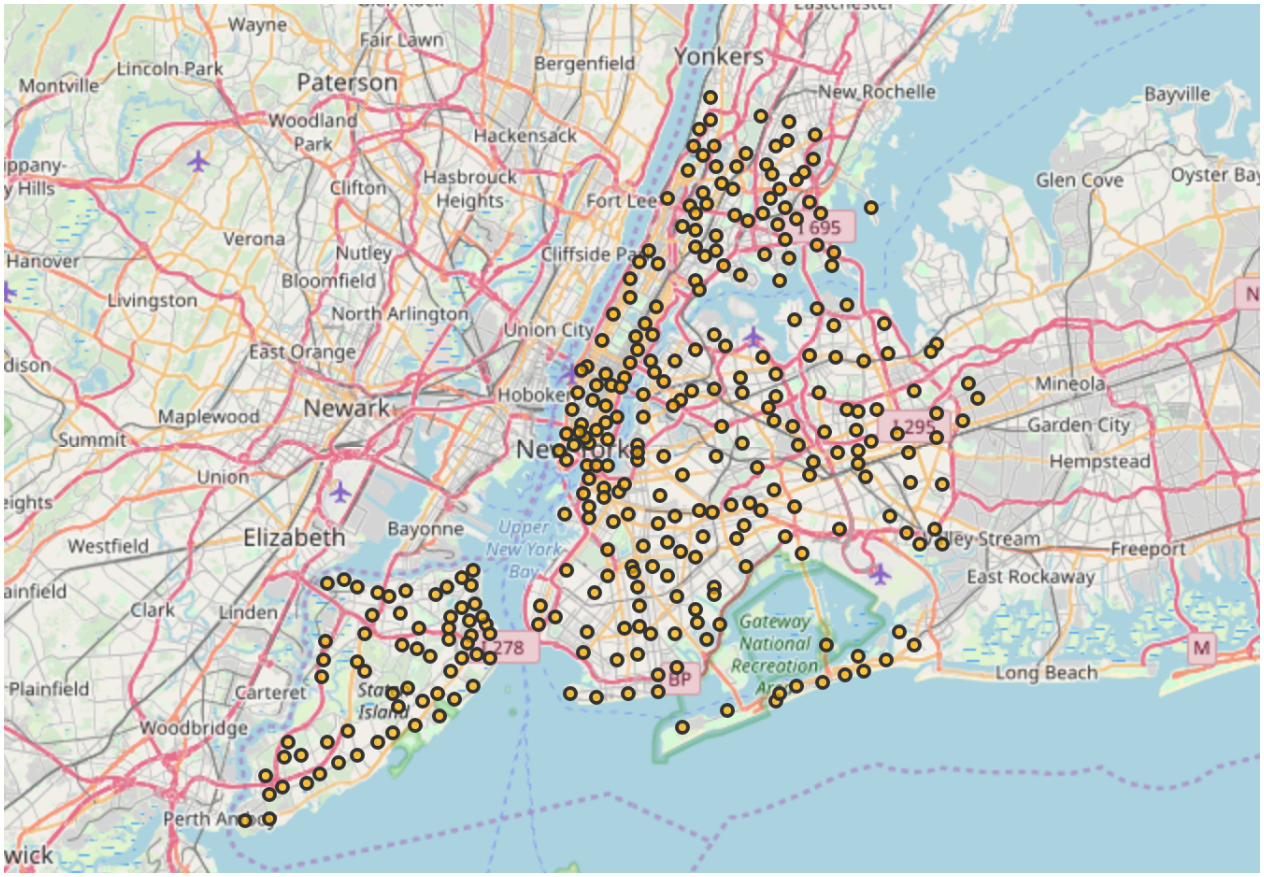

Figure 5: A map of NYC and its neighborhoods

1.2 Toronto

For Toronto, there is no dataset that contains all needed neighborhood data as was the case for NYC; only a dataset that maps Toronto postal codes to latitude and longitude coordinates was provided by the organizers of "Applied Data Science Capstone" course mentioned above; Figure

6 shows few rows of this dataset. Another dataset that lists the neighborhoods and their postal codes should be used so the combination of the two datasets produces the desired results.

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Figure 6: Toronto's postal codes with their coordinates

There is a Wikipedia page titled "List of postal codes of Canada: M". This page lists the postal codes in Canada that start with the letter M which are the postal codes of Toronto city; it lists the postal codes with the neighborhood and borough name associated with each postal code. To download this web page and extract the relevant data from it, Pandas `read_html() functions` can be used. It reads HTML tables on a web page in a list of dataframes. Figure 7 shows the first few rows of the dataframe extracted from that web page.

	PostalCode	Borough	Neighborhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront

Figure 7: Toronto's postal codes with neighborhood and borough names

In that dataframe, there are 77 records out of 288 where the "Borough" variable has the value "Not assigned"; for these 77 records, the "Neighborhood" variable also has the value "Not assigned"; Figure 7 shows some examples of these records. Thus, these records will be deleted because they don't carry meaningful information regarding Toronto neighborhoods. Also, there is one record in that dataframe where there is a valid value for the "Borough" variable but the value of the "Neighborhood" variable is "Not assigned". For this record, the borough name will be given to neighborhood. Figure 8 shows this record before and after. Moreover, there are many cases in the dataframe where multiple neighborhoods share the same postal code and the same borough and since the dataframe of Figure 6 specifies the coordinates based on the postal code, then the records where multiple neighborhoods share the same postal code and borough will be merged; Figure 9 shows a before-and-after example of this operation.

	PostalCode	Borough	Neighborhood
8	M7A	Queen's Park	Not assigned

	PostalCode	Borough	Neighborhood
8	M7A	Queen's Park	Queen's Park

Figure 8: Dealing with case where the neighborhood name is not assigned. Left: before; right: after.

	PostalCode	Borough	Neighborhood
4	M5A	Downtown Toronto	Harbourfront
5	M5A	Downtown Toronto	Regent Park

	PostalCode	Borough	Neighborhood
53	M5A	Downtown Toronto	Harbourfront, Regent Park

Figure 9: Merging records where multiple neighborhoods share the same postal code and borough. Left: before; right: after

Now by merging the dataframes of Figure 6 and Figure 7 after performing the aforementioned cleaning processes, the desired dataframe that contains neighborhood names and coordinates is formed; Figure 10 shows the head of this dataframe which contains data on **211** neighborhoods in **103** rows—due to the merging operations described above.

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Figure 10: The Toronto neighborhood-data dataframe

As with NYC, Figure 11 shows a map of Toronto city and its neighborhoods; each green circle represents the location of one neighborhood or a group of neighborhoods that share the same coordinates.

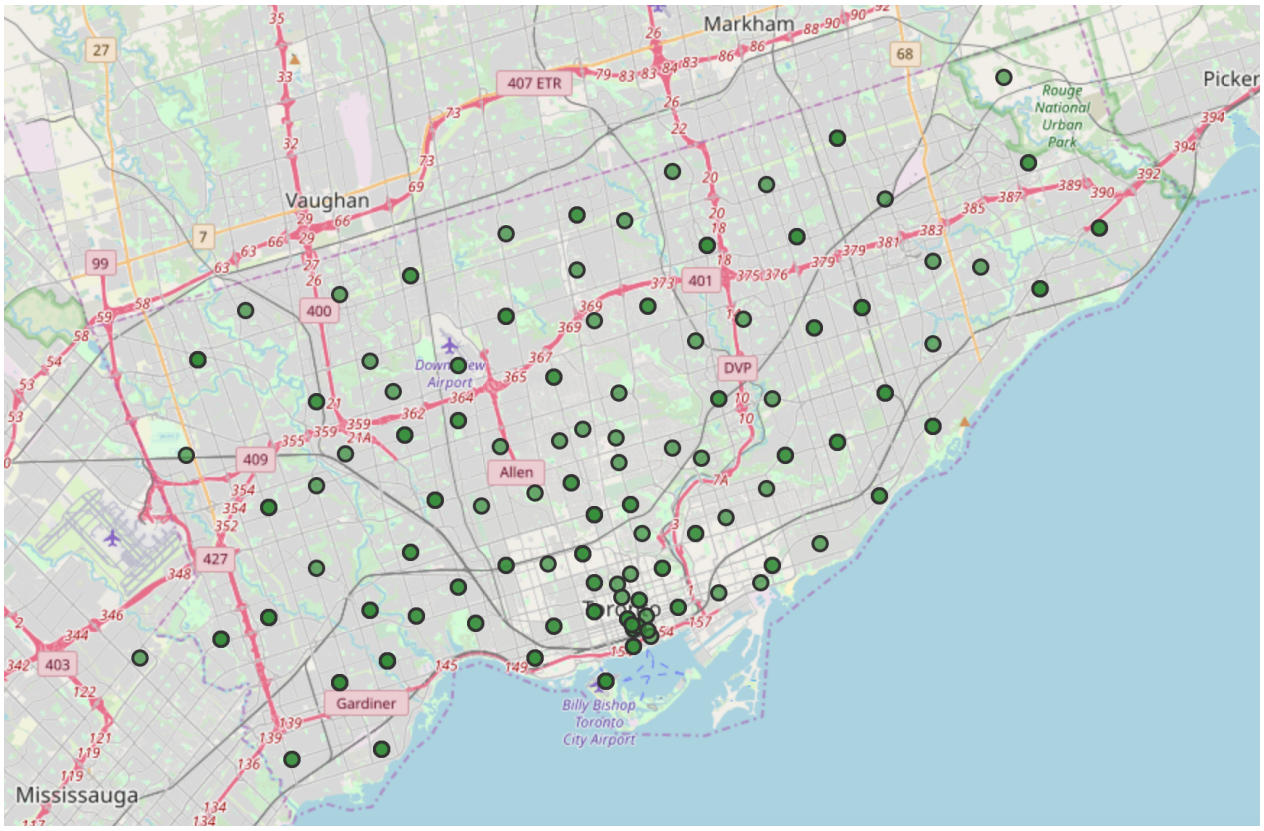

Figure 11: A map of Toronto and its neighborhoods

2 Venues Data

For each city, data that describes the venues of its neighborhoods and the categories of these venues is needed. Venues data will be retrieved from Foursquare which is a popular source of location and venue 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 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.

Figure 12 shows the code used to create a function 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.

```
def getNearbyVenues(names, latitudes, longitudes, radius=500, LIMIT=100):
    """
    A function that retrieves information about venues in each neighborhood.
    It takes as input a list of the names of the neighborhoods, a list of
    their latitudes, and a list of their longitudes.
    It returns a dataframe with information about each neighborhood and its venues.
    """

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print('*', end='')

        # create the API request URL
        url = ('https://api.foursquare.com/v2/venues/search?&client_id={}&client_secret={}'
               '&v={}&ll={},{}&intent=browse&radius={}&limit={}'
               .format(CLIENT_ID, CLIENT_SECRET, VERSION, lat, lng, radius, LIMIT))

        # make the GET request
        results = None
        while results is None:
            try:
                results = requests.get(url).json()["response"]["venues"]
            except:
                print('X', end='')
                results = None

        # return only relevant information for each nearby venue
        venues_list.append([(name, lat, lng, v['name'], v['location']['lat'],
                             v['location']['lng'], v['categories'][0]['name'])
                            for v in results if len(v['categories']) > 0])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighborhood', 'Neighborhood Latitude', 'Neighborhood Longitude',
                             'Venue', 'Venue Latitude', 'Venue Longitude', 'Venue Category']

    return(nearby_venues)
```

Figure 12: Code used to build a venues dataframe for a city neighborhoods

After retrieving the venue data, 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.

2.1 New York City

Using the function in Figure 12 with NYC neighborhood data retrieved data about more than **23,000** venues in NYC neighborhoods. For each venue, venue name, category, latitude, and longitude were retrieved. The head of the dataframe returned by the function for NYC is shown in Figure 13. We can see that each row in the dataframe contains data about one venue: the venue name, coordinates (latitude and longitude), and category in addition to the neighborhood in which the venue is located and the coordinates of the neighborhood.

Different numbers of venues were found in different neighborhoods: for example, data about 81 venues were returned for Allerton neighborhood, 77 venues for Annadale neighborhood, and 51 venues for Belmont neighborhood.

As mentioned above, data on more than 23,000 venues was returned. Each venue of them belongs to one of 574 unique categories.

	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

Figure 13: Venue dataframe for NYC

2.2 Toronto

Similar to what has been done for NYC, a dataframe that describes the venues of Toronto neighborhoods was created; Figure 14 shows a part of it. The dataframe contains data for more than **7,700** venues in Toronto.

Different numbers of venues were found in different neighborhoods: for example, data about 78 venues were returned for Agincourt neighborhood and 64 venues for Berczy Park neighborhood.

As mentioned above, data on more than 7,700 venues was returned. Each venue of them belongs to one of 500 unique categories.[1]

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1	Rouge, Malvern	43.806686	-79.194353	Shell	43.803227	-79.192414	Gas Station
2	Rouge, Malvern	43.806686	-79.194353	Centennial Park	43.786257	-79.148776	Park
4	Rouge, Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
5	Rouge, Malvern	43.806686	-79.194353	Rouge Park - Woodland Trail	43.801782	-79.200427	Trail
6	Rouge, Malvern	43.806686	-79.194353	Auto Camping	43.807896	-79.199733	Automotive Shop

Figure 14: Venue dataframe for Toronto

Part III

Exploratory Data Analysis

In this section, the datasets produced in the previous section will be explored via effective visualizations to understand the data better.

[1] In fact, no venue data was returned for the postal-code areas that contain these neighborhoods. Remember that previously the records that represent neighborhoods that share the same postal code were merged together.

1 Most Common Venue Categories

What are the categories that have more venues than the others in NYC and Toronto? To answer this question, the number of occurrences is counted for each venue category in Figure 13 (for NYC) and Figure 14 (for Toronto). After doing so, a bar plot can be used to visualize the popularity of the most common venue categories in each city.

1.1 New York City

Figure 15 shows a bar plot of the most common venues in NYC. We can see that the most common category is "Residential Building (Apartment / Condo)" with ~1200 venues in NYC; this means that there are ~1200 residential buildings in NYC. In the second rank, the category "Salon / Barbershop" appears with ~900 venues. In the third rank comes the "Deli / Bodega" category; according to Oxford dictionary, a deli is a shop selling cooked meats, cheeses, and unusual or foreign prepared foods; and a bodega is a small grocery shop, especially in a Spanish-speaking neighborhood.

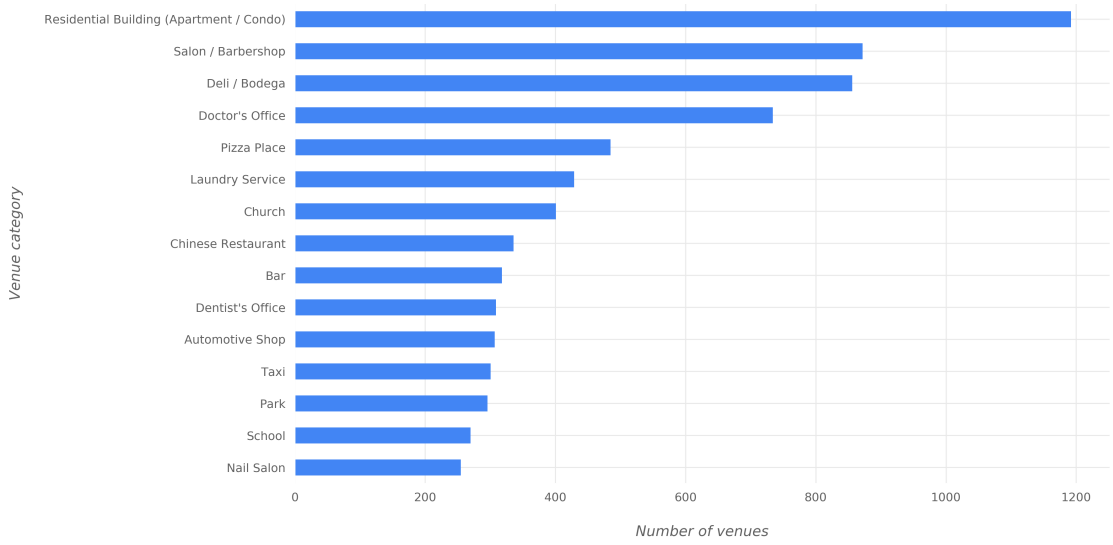

Figure 15: Most common venue categories in NYC

1.2 Toronto

Figure 16 shows a bar plot of the most common venues in Toronto. For Toronto, the most common venue category is also "Residential Building (Apartment / Condo)" with around 450 venues. Then comes "Park" category with ~200 venues. And in the third place appears "Coffee Shop" with around 175 venues.

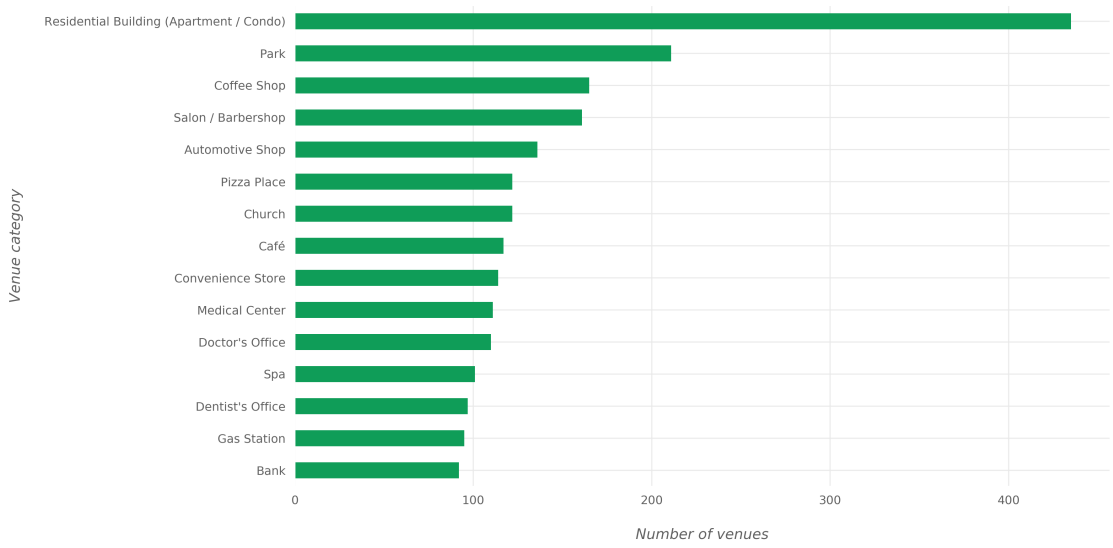

Figure 16: Most common venue categories in Toronto

It can be seen that there are similarities between the most common categories in NYC and in Toronto: we see many categories appearing in both plots of Figure 15 and Figure 16.

2 Most Widespread Venue Categories

Now another question is to be answered: What are the venue categories that exist in more neighborhood? This question is different than the one mentioned in 1. To explain the difference with 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 common 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.

2.1 New York City

Figure 17 shows the most widespread venue categories in NYC. It can be seen that the order of categories this time is different than that of the most common categories (Figure 15). The most widespread category is "Salon / Barbershop"; Salons and barbershops exist in ~250 neighborhoods out of the 306 neighborhoods. After that comes the "Deli / Bodega" category with venues in ~250 neighborhoods also. In the third place comes the "Residential Building (Apartment / Condo)" category with venues in ~230 neighborhoods.

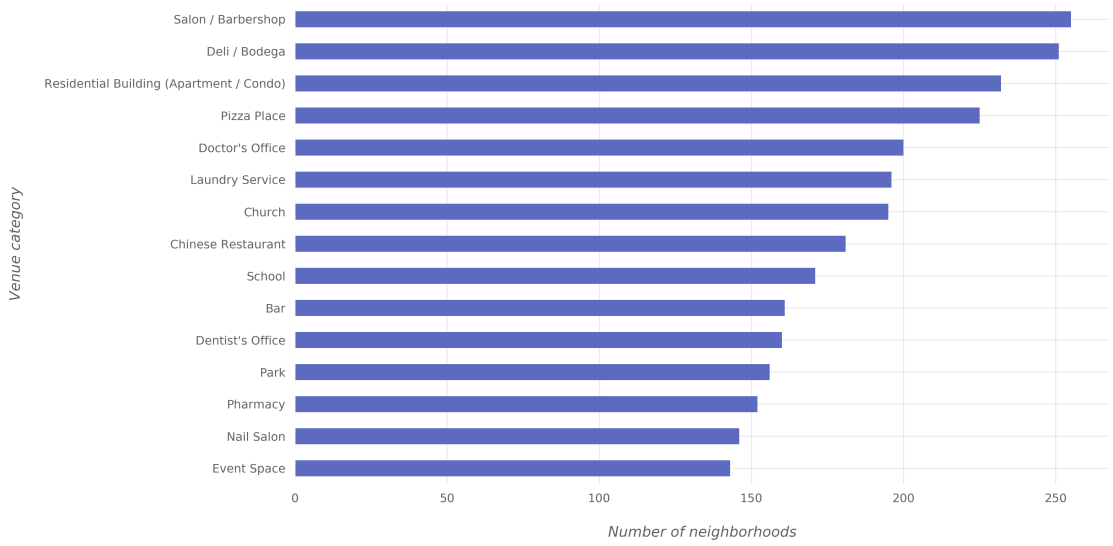

Figure 17: Most widespread venue categories in NYC

2.2 Toronto

Figure 17 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. In the first place comes the "Coffee Shop" category with venues in ~80 neighborhoods[2]. Then comes "Residential Building (Apartment / Condo)" category with venues in ~80 neighborhoods. And the third most-widespread category is "Park" with venues in ~75 neighborhoods.

[2]Remember that it is possible to be more than 45 neighborhoods because earlier in this project, records that represent neighborhoods that share the same postal code were merged together. So it is more accurate to say here that this category has venues in 45 postal-code areas.

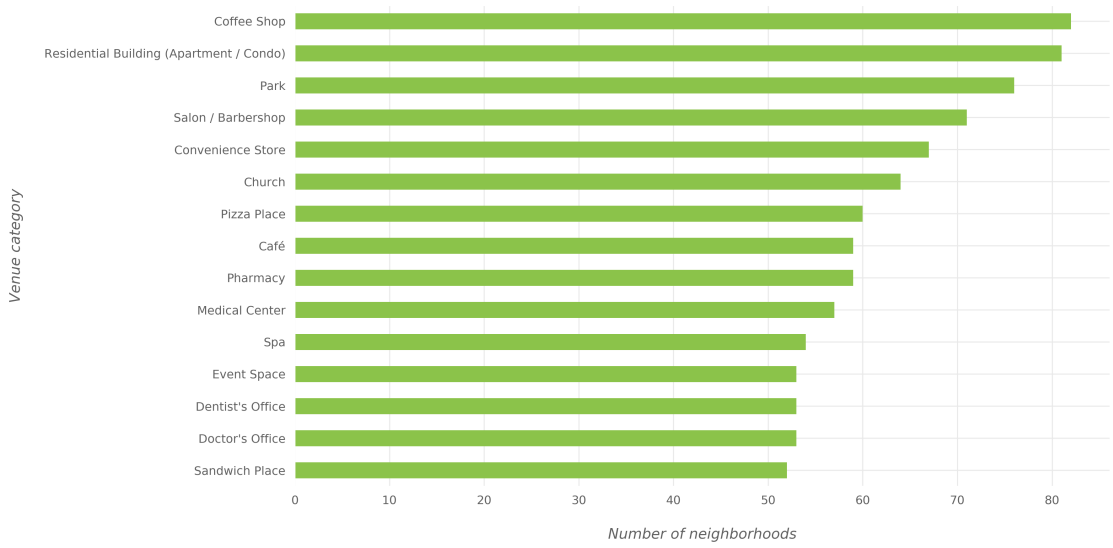

Figure 18: Most widespread venue categories in Toronto

Part IV

Clustering of Neighborhoods

In this section, clustering will be applied on NYC and Toronto neighborhoods to find similar neighborhoods in the two cities. Clustering is the process of finding similar items in a dataset based on the characteristics (features) of items in the dataset. In particular, K-means clustering algorithm of the Scikit-learn Python library will be used. To be able to perform clustering, a dataset suitable for clustering is needed; the datasets described in Figure 13 and Figure 14 are not ready to be used with clustering algorithms.

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, the following two features will be selected out of the dataframes of Figure 13 and Figure 14: "Neighborhood" and "Venue Category". But still after that, the data is not ready for the clustering algorithm because the algorithm works with 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 clustering. One-hot encoding will be applied on the NYC data and on Toronto data then, as will be explained later, the data of the two cities will be combined.

After applying one-hot encoding on NYC data, the resulting dataframe looks like the one shown in Figure 19. For example, when looking at the second row in Figure 13, it can be seen that the venue category for that row is "Dessert Shop", so the column whose value is 1 for the second row in Figure 19 is the "Dessert Shop" column; and the same applies for all rows.

	Neighborhood_	ATM	Accessories Store	Acupuncturist	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Design Studio	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store
0	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0
1	Wakefield	0	0	0	0	0	0	0	0	1	0	0	0
2	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0
3	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0
4	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0

Figure 19: The result of one hot encoding on NYC data

Figure 20 shows the resulting dataframe for Toronto after applying the same operations.

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

Figure 20: The result of one hot encoding on Toronto data

Note that in Figure 19 and Figure 20, the column that contains neighborhood names is named "Neighborhood_" instead of just "Neighborhood". This was done because there is a venue category called "Neighborhood" so the neighborhood-names columns was given the name "Neighborhood_" to avoid having two columns with the same name.

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 Wakefield neighborhood has 15 venues (i.e. 15 rows in the dataframe of Figure 19) and 4 of these venues are of the "Sandwich Place" category (i.e. the "Sandwich Place" column in Figure 19 has a value of 1 for 4 of Wakefield rows), then Wakefield row in the aggregated dataframe will have the value $4/15 = 0.27$ for the "Sandwich Place" column. Figure 21 shows how the aggregated dataframe looks like for NYC and Figure 22 for Toronto.

	Neighborhood_	ATM	Accessories Store	Acupuncturist	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Gate	Airport Lounge	Airport Terminal	Airport Tram	Alternative Healer	American Restaurant
0	Allerton	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
1	Annadale	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.038961
2	Arden Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
3	Arlington	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.028169
4	Arrochar	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000

Figure 21: 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	Airport Lounge	Airport Service	Airport Terminal
0	Adelaide, King, Richmond	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Agincourt North, L'Amoreaux East, Milliken, St...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Albion Gardens, Beaumond Heights, Humbergate, ...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 22: A part of the aggregated dataframe for Toronto

2 Combining NYC and Toronto Data

After producing the aggregated dataframes for each of NYC and Toronto, these dataframes should be combined before applying the clustering algorithm. However, in order to distinguish NYC neighborhoods from Toronto neighborhoods in the new dataframe, a text string is added to the end of each neighborhood name before merging the dataframes: for NYC, the string to be added is "_NYC" and "_Toronto" for Toronto.

Also, NYC and Toronto don't necessarily have the same venue categories (i.e. some columns in the dataframe of Figure 21 don't exist in the dataframe of Figure 22 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.

Figure 23 shows a part of the dataframe that resulted from the combination of NYC and Toronto aggregated dataframes. This dataframes contains data on 408 neighborhoods in both NYC and Toronto.

	Neighborhood_	Accessories Store	Acupuncturist	Adult Boutique	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	Airport Tram	Alternative Healer	American Restaurant
303	Woodrow_NYC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.012658
304	Woodside_NYC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
305	Yorkville_NYC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
306	Adelaide, King, Richmond_Toronto	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.011111
307	Agincourt_Toronto	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.012821
308	Agincourt North, L'Amoreaux East, Milliken, St...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000

Figure 23: The combination of NYC and Toronto aggregated dataframes

3 The Most Common Categories for Each Neighborhood

Using the dataframe of Figure 23, another dataframe is created to specify the 5 most common categories for each neighborhood in NYC and Toronto. This dataframe is created by retrieving the 5 categories with the largest values for each neighborhood in Figure 23. Figure 24 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	Deli / Bodega	Salon / Barbershop	Gas Station	Pharmacy	Pizza Place	Spa
1	Annadale_NYC	Pizza Place	Salon / Barbershop	Tattoo Parlor	Nail Salon	American Restaurant	Pet Store	Gas Station
2	Arden Heights_NYC	Professional & Other Places	Park	Food	Doctor's Office	Pool	Dentist's Office	Moving Target
3	Arlington_NYC	Church	Residential Building (Apartment / Condo)	Boat or Ferry	Deli / Bodega	Hardware Store	Playground	Laundry Service
4	Arrochar_NYC	Deli / Bodega	Laundry Service	Nail Salon	Pizza Place	Dance Studio	Doctor's Office	Salon / Barbershop

Figure 24: Most common categories for each neighborhoods

4 Clustering and its Results

By obtaining the dataframe of Figure 23, it is now possible to apply the clustering algorithm. Figure 25 shows the code used to perform clustering using the K-means algorithm of Scikit-learn library. The variable named `nyc_tor_grouped` contains the dataframe of Figure 23. Notice that the "Neighborhood_" column was dropped before applying the clustering algorithm (i.e. the clustering algorithm was applied on all columns except that column); this was done because the clustering algorithm doesn't accept non-numerical columns as mentioned earlier. However, this column will be re-added as will be explained soon.

```
# 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]
```

```
array([2, 2, 1, 0, 2, 0, 4, 2, 1, 2], dtype=int32)
```

Figure 25: Code used to perform K-means clustering

As can be seen in Figure 25, 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 of Figure 24, a dataframe is constructed to show the neighborhoods of NYC and Toronto, the cluster to which each neighborhood belongs, and the most common venue categories in each neighborhood. This dataframe can be seen in Figure 26.

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	2	Salon / Barbershop	School	Food	Nightclub	Deli / Bodega	Residential Building (Apartment / Condo)	Event Space
Woodhaven_NYC	2	Salon / Barbershop	Deli / Bodega	Laundry Service	Chinese Restaurant	Pizza Place	Hookah Bar	Doctor's Office
Woodlawn_NYC	2	Bar	Deli / Bodega	Salon / Barbershop	Garden	Food & Drink Shop	Residential Building (Apartment / Condo)	Pizza Place
Woodrow_NYC	1	Pool	Salon / Barbershop	School	Grocery Store	Cocktail Bar	Art Gallery	Dentist's Office
Woodside_NYC	2	Salon / Barbershop	Bar	Platform	Thrift / Vintage Store	Miscellaneous Shop	Deli / Bodega	Thai Restaurant
Yorkville_NYC	4	Residential Building (Apartment / Condo)	Laundry Service	Pharmacy	Taxi	Lounge	General Entertainment	Art Gallery
Adelaide, King, Richmond_Toronto	1	Coffee Shop	Hotel Bar	Café	Italian Restaurant	Vegetarian / Vegan Restaurant	Food Court	Hotel
Agincourt_Toronto	1	Automotive Shop	Spa	General Entertainment	Chinese Restaurant	Residential Building (Apartment / Condo)	Salon / Barbershop	Event Space
Agincourt North, L'Amoreaux East, Milliken, Steeles East_Toronto	1	Chinese Restaurant	School	Doctor's Office	BBQ Joint	Medical Center	Church	Bakery
Albion Gardens, Beaumond Heights, Humbergate, Jamestown, Mount Olive, Silverstone, South Steeles, Thistletown_Toronto	2	Salon / Barbershop	Movie Theater	Spiritual Center	Pizza Place	Farm	Art Gallery	Bakery

Figure 26: NYC and Toronto neighborhoods, their clusters, and their most common categories

The output of the clustering operation is 5 clusters with cluster labels 0, 1, 2, 3, and 4. Each cluster is expected to contain a group of similar neighborhoods based on the categories of the venues in each neighborhood. The clustering algorithm was run on 406 neighborhoods in NYC and Toronto. Table 2 shows the number of neighborhoods in each cluster.

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

Table 2: Number of neighborhoods in each cluster

For examples, Figure 27 shows a part of the first cluster and Figure 28 shows a part of the third cluster. It's hard to show all the clusters with all of their neighborhoods since there are 406 neighborhoods in the five clusters. However, they 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
Vinegar Hill_NYC	0	Art Gallery	Residential Building (Apartment / Condo)	Event Space	General Entertainment	Factory	Bike Rental / Bike Share	Coworking Space
West Village_NYC	0	Taxi	Residential Building (Apartment / Condo)	Laundry Service	Cocktail Bar	Café	Salon / Barbershop	Event Space
Williamsburg_NYC	0	Residential Building (Apartment / Condo)	Pizza Place	Salon / Barbershop	Park	Bar	American Restaurant	Wedding Hall
Bathurst Manor, Downsview North, Wilson Heights_Toronto	0	Residential Building (Apartment / Condo)	Spa	Doctor's Office	Bank	Bar	Ice Cream Shop	Medical Center
Bayview Village_Toronto	0	Residential Building (Apartment / Condo)	Church	Doctor's Office	Dog Run	Salon / Barbershop	School	Asian Restaurant
Berczy Park_Toronto	0	Residential Building (Apartment / Condo)	Parking	General Entertainment	Bar	Government Building	Gym	Laundry Service
Caledonia-Fairbanks_Toronto	0	Residential Building (Apartment / Condo)	Salon / Barbershop	Bar	Convenience Store	Park	Latin American Restaurant	Bakery
Cloverdale, Islington, Martin Grove, Princess Gardens, West Deane Park_Toronto	0	Residential Building (Apartment / Condo)	Park	Café	Tech Startup	School	Miscellaneous Shop	Wine Shop
Downsview Northwest_Toronto	0	Residential Building (Apartment / Condo)	Caribbean Restaurant	Hotel	Gas Station	Mobile Phone Shop	Clothing Store	Church
Downsview West_Toronto	0	Residential Building (Apartment / Condo)	Park	Vietnamese Restaurant	Elementary School	Salon / Barbershop	Doctor's Office	Bank

Figure 27: 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
Wingate_NYC	2	Salon / Barbershop	School	Food	Nightclub	Deli / Bodega	Residential Building (Apartment / Condo)	Event Space
Woodhaven_NYC	2	Salon / Barbershop	Deli / Bodega	Laundry Service	Chinese Restaurant	Pizza Place	Hookah Bar	Doctor's Office
Woodlawn_NYC	2	Bar	Deli / Bodega	Salon / Barbershop	Garden	Food & Drink Shop	Residential Building (Apartment / Condo)	Pizza Place
Woodside_NYC	2	Salon / Barbershop	Bar	Platform	Thrift / Vintage Store	Miscellaneous Shop	Deli / Bodega	Thai Restaurant
Albion Gardens, Beaumond Heights, Humbergate, Jamestown, Mount Olive, Silverstone, South Steeles, Thistletown_Toronto	2	Salon / Barbershop	Movie Theater	Spiritual Center	Pizza Place	Farm	Art Gallery	Bakery

Figure 28: Some records that belong to the third cluster

5 Cluster Analysis

The clustering algorithm grouped neighborhoods of NYC and Toronto in 5 clusters based on the similarity between their venues. Now, these clusters will be investigated to see the most common categories in each of them. Figures 29 show the most common 7 venue categories in each cluster; for each common category, the percentage of venues of that category in the neighborhoods of the cluster is shown also.

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

Figure 29: Most common venue-categories in each of the 5 clusters

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 from the tables of Figure 29 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 be observed in Figure 29.

Figure 30 shows the number of NYC neighborhoods and the number of Toronto neighborhoods in each cluster of the five resulting clusters.

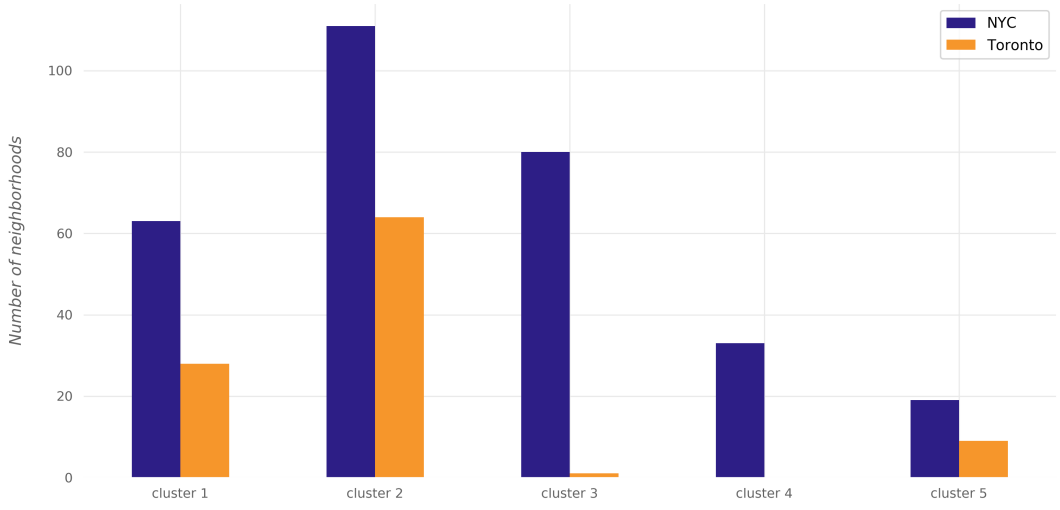

Figure 30: The number of NYC and Toronto neighborhoods in each cluster

Part V

Conclusions

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.