Clustering and Comparing Venues in the Neighborhoods of New York City and Toronto

Part 2

(Data Description)

1-Introduction and Problem Statement

In this project, I assume that I am working for consultation firm and they need a list of major venues in the major cities around the world.

This will help in providing our client with right data which suits their business plans.

As a start I 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.

I 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 businesspeople to take better and more effective decisions regarding where to open their businesses.

New York City Brief:

New York City (NYC), also known as the City of New York or simply New York (NY), is the most populous city in the United States. With an estimated 2018 population of 8,398,748 distributed over a land area of about 302.6 square miles (784 km2).

A global power city,[14] New York City has been described as the cultural,[15][16][17][18][19] financial,[20][21][22] and media capital of the world,[23][24] and exerts a significant impact upon commerce,[22] entertainment, research, technology, education, politics, tourism, art, fashion, and sports.

Toronto City Brief:

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 2,954,024 as of July 2018.

Toronto is an international Center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

2-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:

A-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".

B- 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.

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.

A-Neighborhood Data

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

For 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.

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.

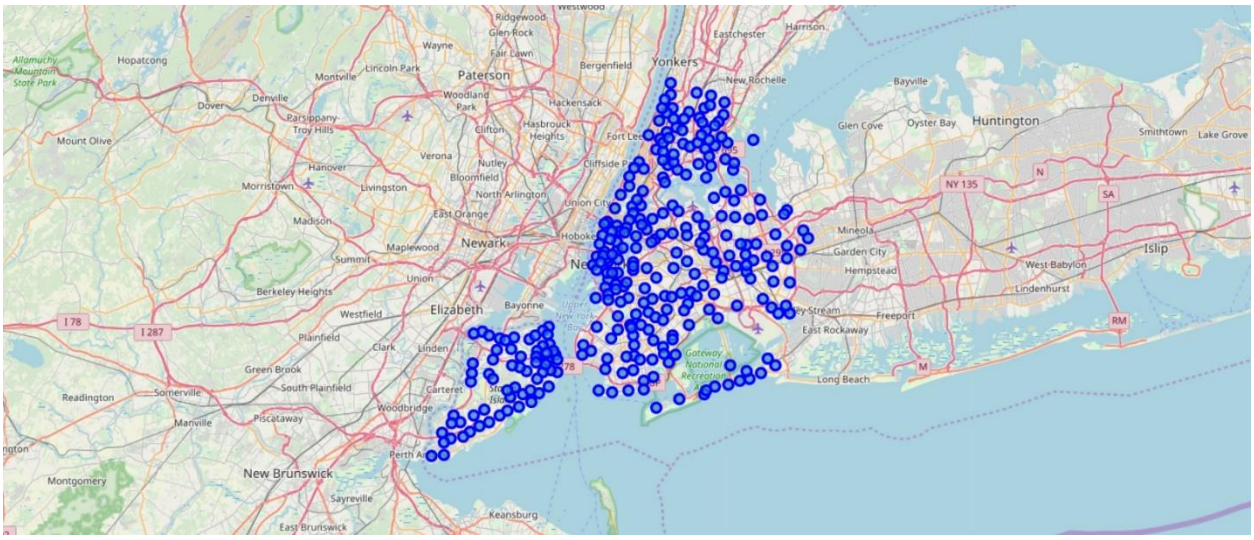
```
[5]: nyc_neighborhoods_data[0]

Out[5]: {'type': 'Feature',
         'id': 'nyu_2451_34572.1',
         'geometry': {'type': 'Point',
          'coordinates': [-73.84720052054902, 40.89470517661]},
         'geometry_name': 'geom',
         'properties': {'name': 'Wakefield',
          'stacked': 1,
          'annoline1': 'Wakefield',
          'annoline2': None,
          'annoline3': None,
          'annoangle': 0.0,
          'borough': 'Bronx',
          'bbox': [-73.84720052054902,
           40.89470517661,
           -73.84720052054902,
           40.89470517661]}}
```

```
[7]: nyc_neighborhoods.head()
```

Out[7]:

	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

Having data of the coordinates of NYC neighborhoods, it is possible to draw a map using Folium Python package of NYC and its neighborhoods. You can see this map below; each circle represents the location of one neighborhood.



For Toronto City:

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.

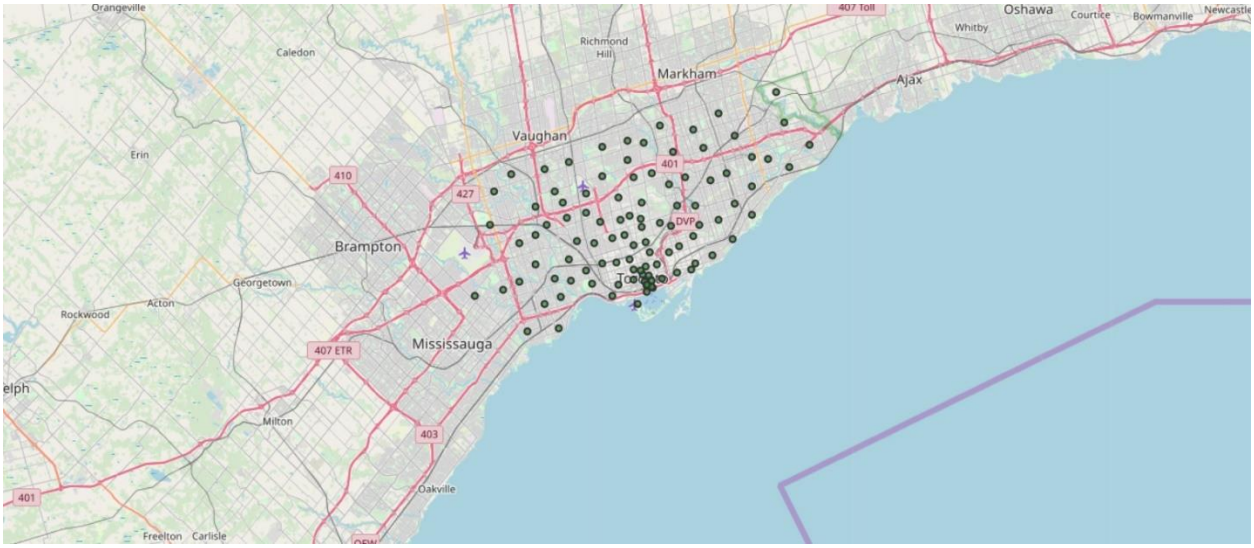
Another dataset that lists the neighborhoods and their postal codes, Latitudes, Longitudes should be used so the combination of the two datasets produces the desired results.

```
In [49]: tor_neighborhoods.head()
```

Out[49]:

	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

As with NYC, you can see a map of Toronto city and its neighborhoods; each circle represents the location of one neighborhood or a group of neighborhoods that share the same coordinates.



B- 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.

Then 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.
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.

For New York City

NYC neighborhood data retrieved data about more than **23,700** venues in NYC neighborhoods.

```
(23753, 7)
```

Out[18]:

	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	Pitman Deli	40.894149	-73.845748	Food
2	Wakefield	40.894705	-73.847201	Julio C Barber Shop 2	40.892648	-73.855725	Salon / Barbershop
3	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
4	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop

Different numbers of venues were found in different neighborhoods:

```
In [*]: nyc_venues.groupby('Neighborhood').size()
```

```
Out[19]: Neighborhood
Allerton                                  83
Annadale                                  77
Arden Heights                             69
Arlington                                 71
Arrochar                                  78
Arverne                                   80
Astoria                                   73
Astoria Heights                           63
Auburndale                                61
Bath Beach                                84
Battery Park City                         88
Bay Ridge                                 87
Bay Terrace, Queens                       81
Bay Terrace, Staten Island                76
Baychester                                79
Bayside                                   85
Bayswater                                 80
Bedford Park                              68
Bedford Stuyvesant                        85
```

For Toronto

Similar to what has been done for NYC, a dataframe that describes the venues of Toronto neighborhoods was created.

The dataframe contains data for more than **7,800** venues in Toronto.

```
(7870, 7)
```

Out[54]:

	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	Subway	43.801095	-79.200304	Sandwich Place
3	Rouge, Malvern	43.806686	-79.194353	Wendy's	43.802008	-79.198080	Fast Food Restaurant
4	Rouge, Malvern	43.806686	-79.194353	Tim Hortons	43.802000	-79.198169	Coffee Shop
5	Rouge, Malvern	43.806686	-79.194353	Tim Hortons / Esso	43.801863	-79.199296	Coffee Shop

Different numbers of venues were found in different neighborhoods

```
[55]: #Let's check how many venues were returned for each neighborhoo
      tor_venues.groupby('Neighborhood').size()
```

```
Out[55]: Neighborhood
Adelaide, King, Richmond                                                                    82
Agincourt                                                                                   75
Agincourt North, L'Amoreaux East, Milliken, Steeles East                                    58
Albion Gardens, Beaumond Heights, Humbergate, Jamestown, Mount Olive, Silverstone, South Steeles, Thistletown    95
Alderwood, Long Branch                                                                      81
Bathurst Manor, Downsview North, Wilson Heights                                             81
Bayview Village                                                                             80
Bedford Park, Lawrence Manor East                                                           86
Berczy Park                                                                                 68
Birch Cliff, Cliffside West                                                                 74
Bloordale Gardens, Eringate, Markland Wood, Old Burnhamthorpe                               73
Brockton, Exhibition Place, Parkdale Village                                                67
Business Reply Mail Processing Centre 969 Eastern                                           70
CFB Toronto, Downsview East                                                                 57
CN Tower, Bathurst Quay, Island airport, Harbourfront West, King and Spadina, Railway Lands, South Niagara    98
Cabbagetown, St. James Town                                                                 90
Caledonia-Fairbanks                                                                         84
Canada Post Gateway Processing Centre                                                       65
Cedarbrae                                                                                   77
Central Bay Street                                                                          88
Chinatown, Grange Park, Kensington Market                                                   92
Christie                                                                                    79
Church and Wellesley                                                                        81
Clairlea, Golden Mile, Oakridge                                                             72
Clarks Corners, Sullivan, Tam O'Shanter                                                     86
```