

Comparing and Clustering the Neighborhoods of New York City and Toronto



Prepared by: Engr. Muhammad Bilal

25 September 2020

Introduction and Problem Statement

In this project, we will compare, cluster, and analyze the neighborhoods of New York City which is located in the United States of America and Toronto which is located in Canada. We will investigate what kinds of businesses are common and rare in each city, what kinds of businesses are more common in one of the two cities than the other city.

Doing this task will empower us to improve comprehension of likenesses and contrasts between the two urban areas which will make it known to financial specialists what sorts of organizations are bound to flourish in the two urban areas, what are the neighborhoods that are appropriate for each kind of business, and what kinds of organizations are not entirely attractive in every city. This permits financial specialists to take better and more successful choices with respect to where to start their businesses.

New York City (NYC) is one of the most crowded urban areas in the US of America. Additionally, NYC is the most etymologically different city on the planet: more than 800 dialects are spoken in it. Additionally, NYC assumes a fundamental part in the financial matters of the USA: if New York City were a sovereign state, it would have the twelfth most elevated Gross domestic product on the planet. New York City comprises five districts: Brooklyn, Staten Island, Manhattan, The Bronx, and Queens.

The second city of enthusiasm for this task is Toronto. As with New York City in the USA, Toronto is the most crowded city in Canada. It's perceived as one of the most multicultural and cosmopolitan urban communities on the planet. Toronto additionally is an exceptionally differing city: more than 160 dialects are spoken in it. On the monetary side, Toronto is a worldwide community for business and account and it is viewed as the budgetary capital of Canada.

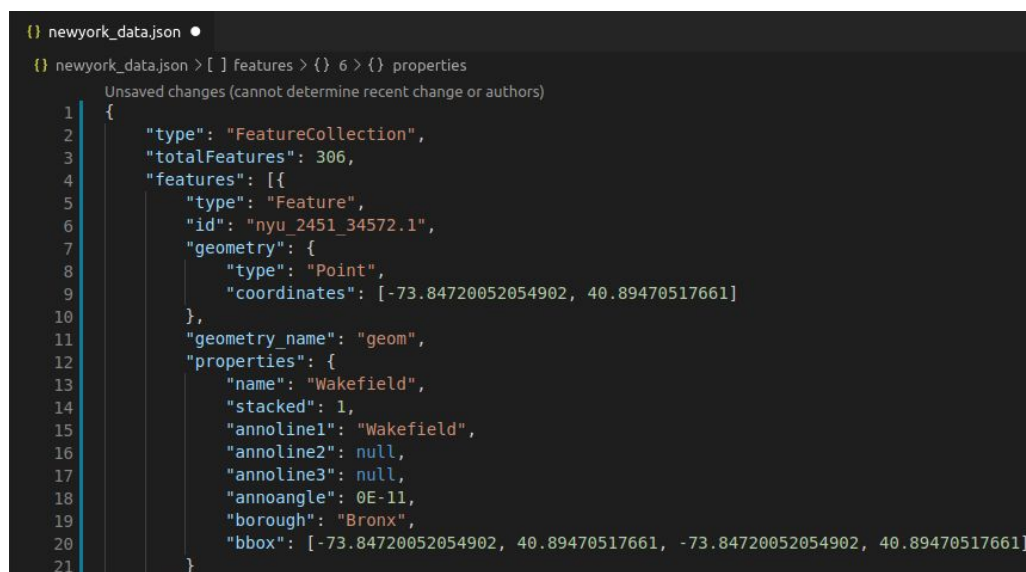
Data Acquisition and Preprocessing

In this part, the processes of obtaining, cleaning, and setting up each dataset utilized in this venture for next stages will be specified. To complete this project, two types of datasets are needed, Neighborhood and Venues data. **Neighborhood dataset** includes the names of the neighborhoods of NYC and Toronto and their latitude and longitude coordinates. We have a portion of this information given by the facilitators of "IBM Data Science Professional Certificate" and we additionally need to scratch some information from the web. **Venues dataset** describes the top 100 venues (bars, parks, restaurants, museums, etc.) in each neighborhood of the two cities. This dataset will be retrieved from Foursquare API which is one of the world largest sources of location and venue data. Then it will be preprocessed so that it should list the venues of each neighborhood with their categories.

Neighborhood Dataset

New York City

We need to acquire and preprocess the dataset that specifies the neighborhood data for New York City along with its coordinates and some other data, provided by the facilitators of "IBM Data Science Professional Certificate" course as a JSON file.



```
{
  "type": "FeatureCollection",
  "totalFeatures": 306,
  "features": [
    {
      "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": null,
        "annoline3": null,
        "annoangle": 0E-11,
        "borough": "Bronx",
        "bbox": [-73.84720052054902, 40.89470517661, -73.84720052054902, 40.89470517661]
      }
    }
  ]
}
```

Figure 1: JSON file for getting Neighborhoods of NYC.

The JSON file is stored in a variable named `newyork_neighborhoods_data`. 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. In Figure 2 Python code is shown used to process the JSON file data and store it in a dataframe named `newyork_neighborhoods`. Also there is the output of the resulting data frame shown in the figure below.

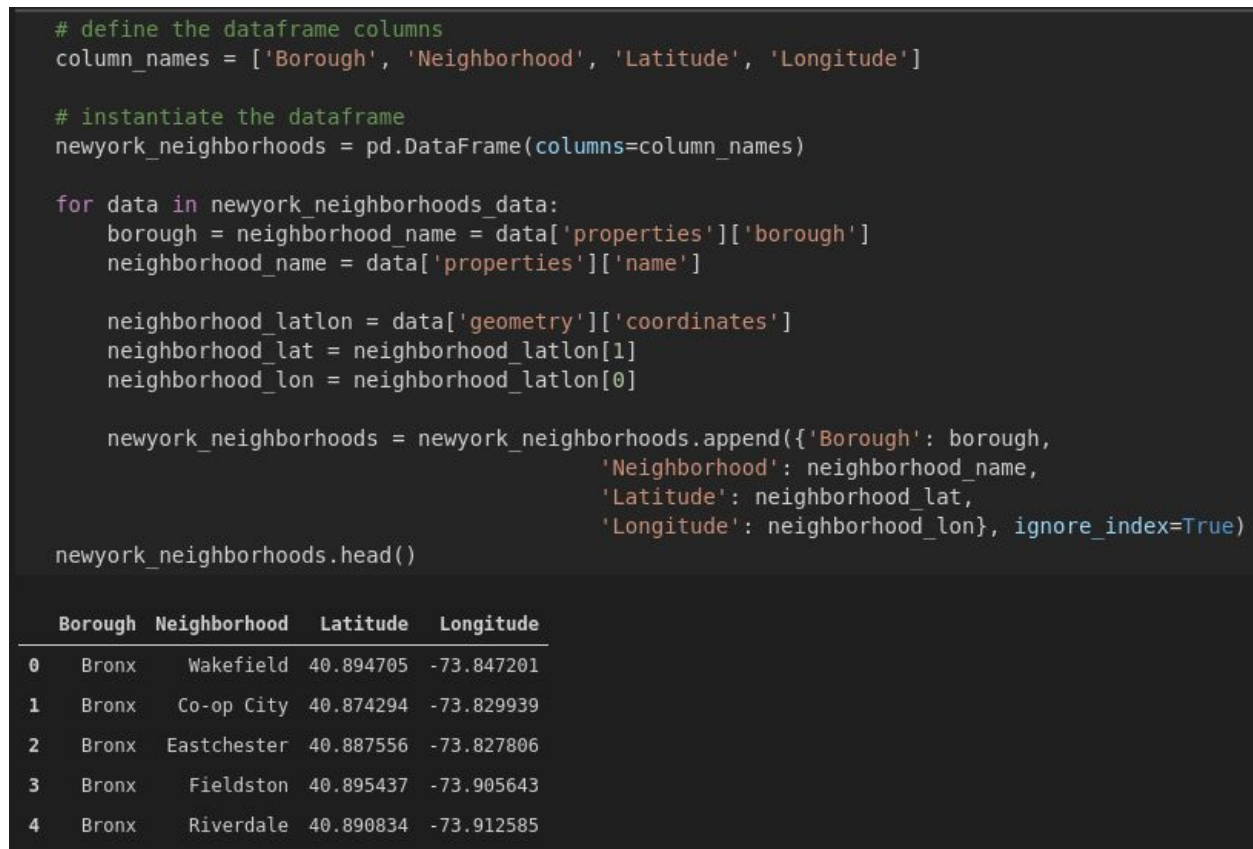


Figure 2: The NYC neighborhood data dataframe

Now we can draw a map using the **Folium** Python package of NYC and its neighborhoods, as we have their coordinates. Figure 3 shows the resulting map in which each blue circle represents the location of one neighborhood in Newyork City.

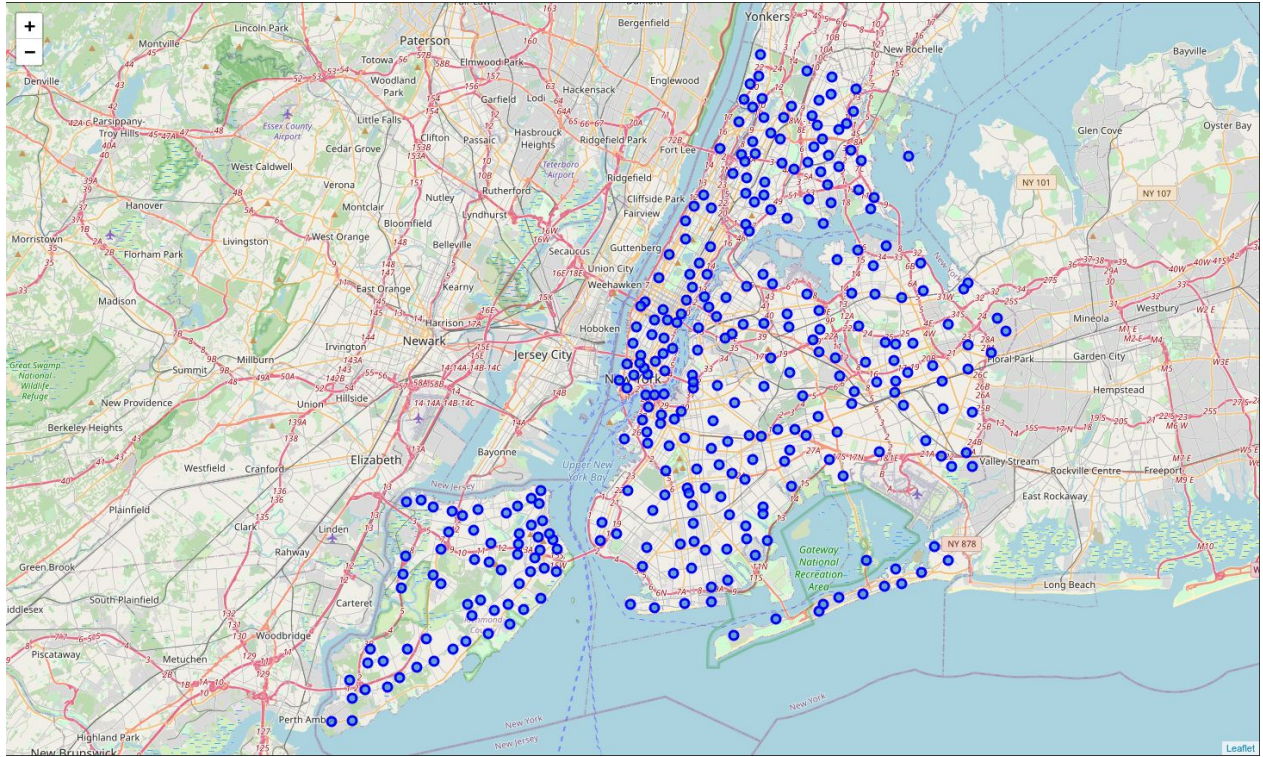


Figure 3: Map of New York City and its neighborhoods

Toronto City

To get the neighborhood and borough data for Toronto city we will parse the Wikipedia page with url "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M". it lists the postal codes with the neighborhood and borough name associated with each postal code. We can use Pandas read_html() functions to download this web page and extract the relevant data from it. Figure below shows the first few rows of the dataframe extracted from this web.

```
toronto_neighborhoods.head()
```

	Postal Code	Borough	Neighbourhood
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	Regent Park, Harbourfront

Figure 4: Postal codes of Toronto City with neighborhoods and boroughs

In the above dataframe we need to do cleaning and preprocessing. There are some records where the "Borough" and "Neighborhood" variable has the value "Not assigned", these records will be deleted as they don't have any useful information regarding Toronto neighborhoods. Also, borough names will be given to neighborhoods where there is a valid value for the "Borough" variable but the value of the "Neighborhood" variable is "Not assigned".

A dataset that maps Toronto postal codes to latitude and longitude coordinates was provided by the facilitators of "IBM Data Science Professional Certificate" course. The csv file for this data is available at "https://cocl.us/Geospatial_data". In figure below this dataset is shown.

```
toronto_neighborhoods_geo_coor = pd.read_csv("https://cocl.us/Geospatial_data")
toronto_neighborhoods_geo_coor.head()
```

	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: Postal codes of Toronto City with their coordinates

Now by merging the both dataframes and performing the cleaning processes, the desired dataframe that contains neighborhood names and coordinates is formed, as shown in figure below.

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	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 7: Neighborhood dataframe for Toronto City

After plotting 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, as shown in figure below.

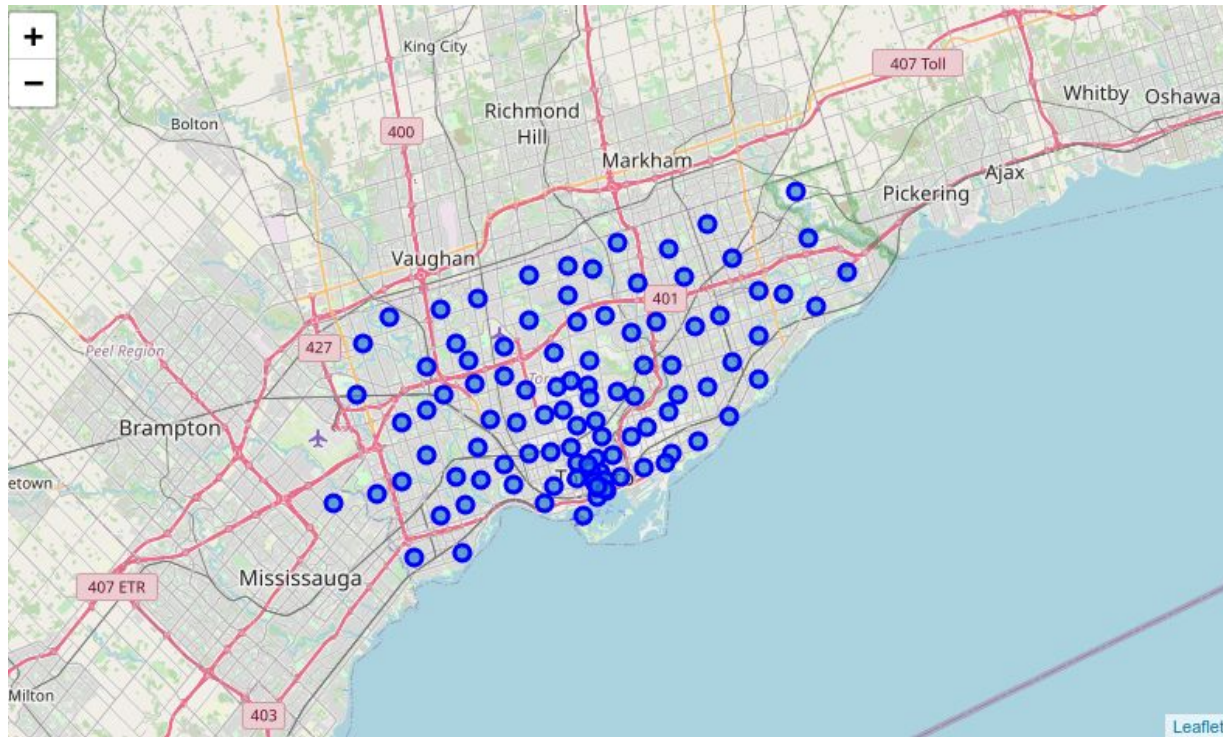


Figure 8: A map of neighborhoods of Toronto City

Venues Dataset

Venues data will be retrieved from Foursquare which is a popular source for getting location and venue data. Foursquare API service will be used to access and download venues data. To retrieve data a URL should be prepared, which in turn is used to request data related to a specific location. For preparing url we need the CLIENT_ID, CLIENT_SECRET and VERSION, you can get those by creating the account on Foursquare website. You can also set the RADIUS and LIMIT of the nearby venues. In this project we are using LIMIT=100 and RADIUS=500.

Below is the function used for getting the nearby Venues for neighborhoods.


```
def getNearbyVenues(names, latitudes, longitudes, radius=500):

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

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

        # make the GET request
        results = requests.get(url).json()["response"]["groups"][0]["items"]

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

    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 9: Python Function for getting venues dataframe of neighborhoods

Now using the function **getNearbyVenues**, we will get the venues for Toronto and Newyork City, by providing the neighborhood names and coordinates as an argument. The first five rows of the resulting dataframe of Newyork City are shown in the figure below.

```
print(newyork_venues.shape)
newyork_venues.head()
```

(10131, 7)

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

Figure 10: Venue dataframe for Newyork City

Methodology and Data Analysis

This section of report will include the exploratory data analysis of the previously formed pandas dataframes and datasets. And talk about the methodology used in performing this analysis.

Common Venue Categories

We can find the most common categories by counting the number of venue categories for each city. Figure below includes 10 most common categories of Toronto and Newyork City.

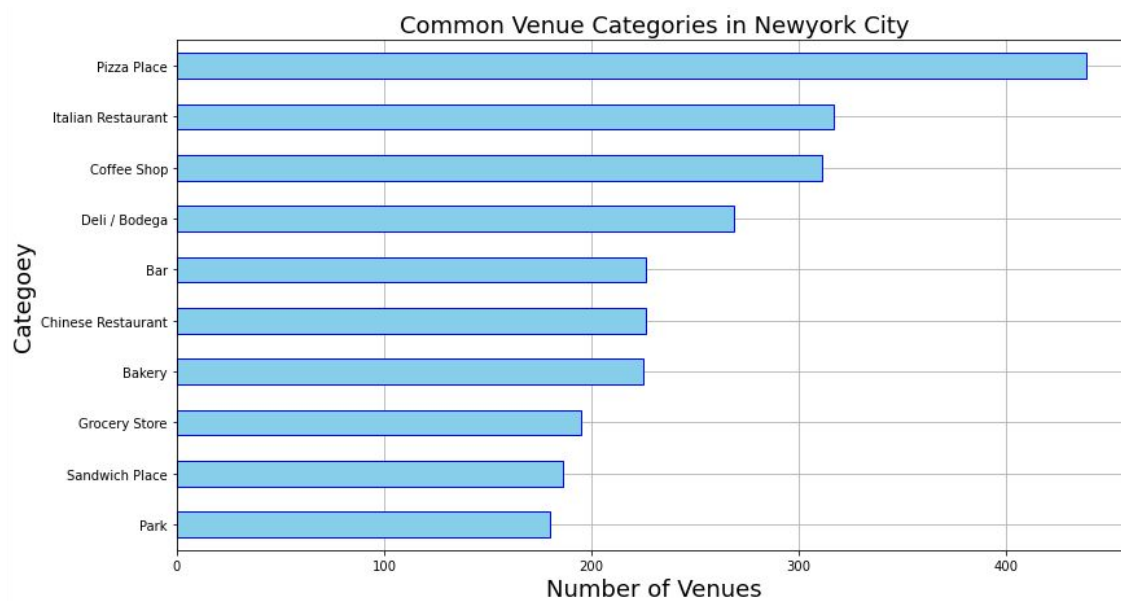


Figure 11: Common Venues for Newyork City

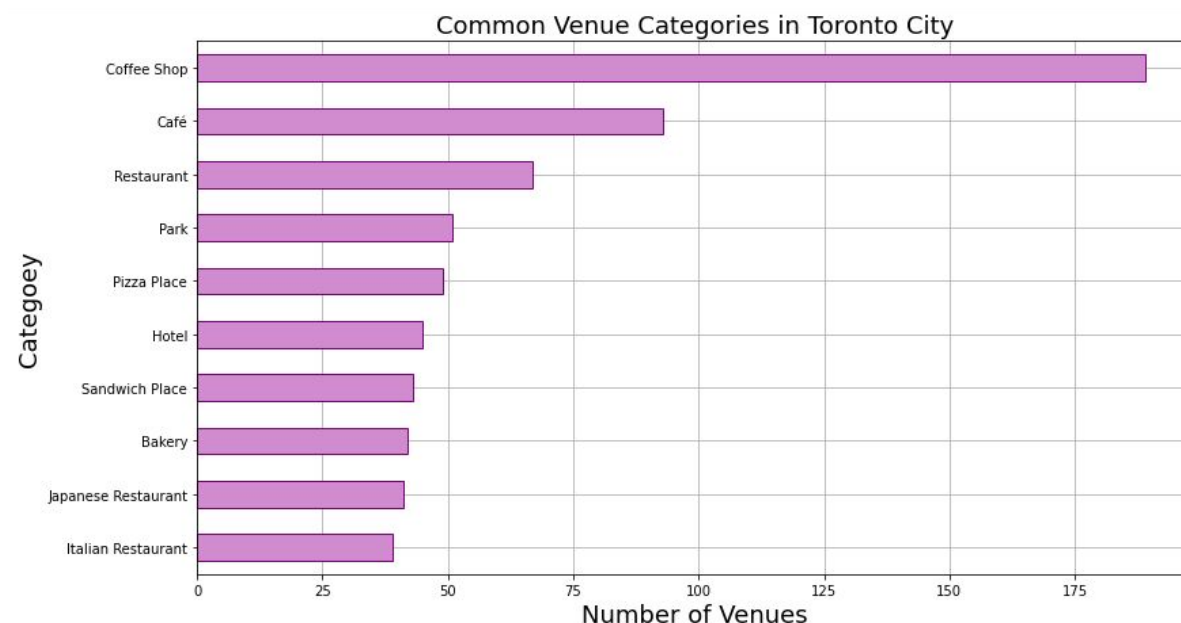


Figure 12: Common Venues for Toronto City

We can see that the most common categories for Newyork city are Pizza place, Italian Restaurant, Cafe and Parks whereas the most common categories for Toronto city are Coffee shops, Cafe and Restaurants.

Top 10 Widespread Venue Categories

Now someone might be interested to know the most widespread categories in Both cities i.e. the venue categories which exist in most neighborhoods of the city. To compute this we need to count those categories which exist in more neighborhoods of the city. To present this useful information we have used the horizontal bar plots for showing the Categories and on the x-axis we have the Number of Venues, as shown in below Figures.

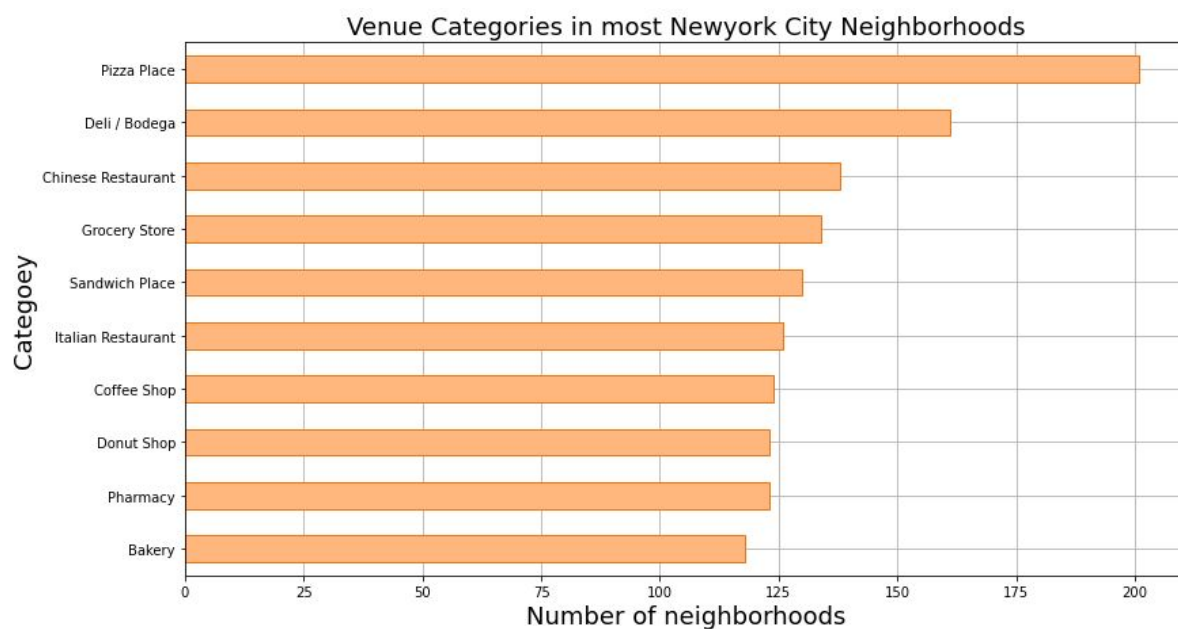


Figure 13: Most widespread venue categories for Newyork city

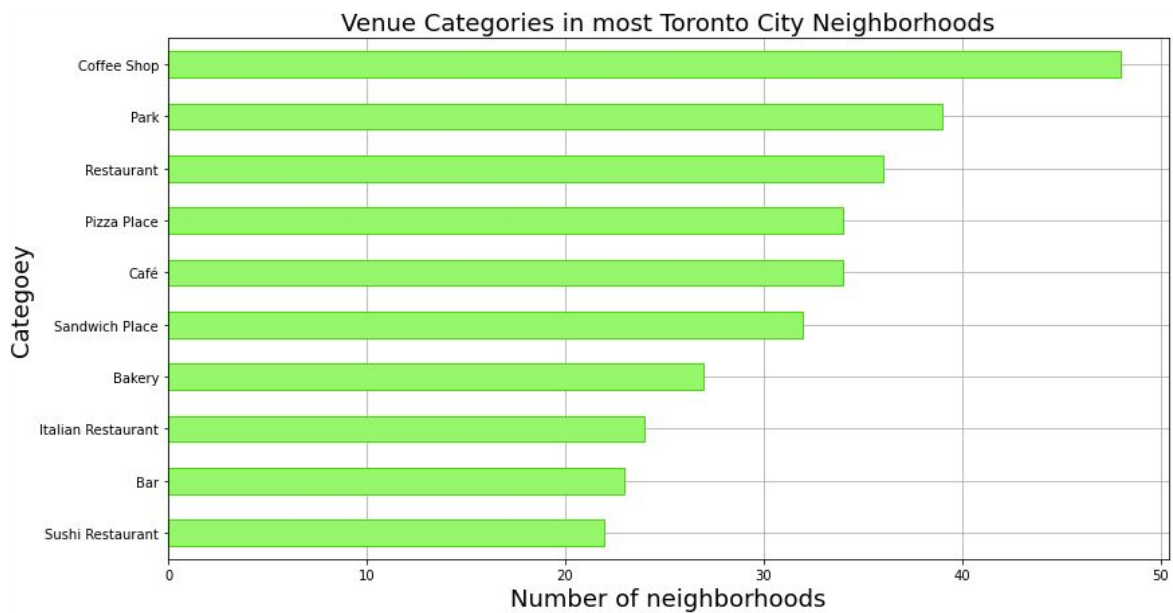


Figure 14: Most widespread venue categories in Toronto city

Methodology and clustering

In this part, clustering will be applied on Toronto and New York City neighborhoods to discover comparative neighborhoods in the two urban areas. Clustering is the process of finding similar items in a dataset based on the characteristics of items in the dataset. We have used the **K-mean clustering methods** of the **Scikit-learn Python library** and form a 5 clusters of the Toronto and Newyork city of Neighborhoods usings the nearby venues data. For clustering we needed to convert the previous dataset to a desired dataframe using the technique of one hot encoding. Figures below show the resulting data frame, after applying the one-hot encoding on the Newyork and Toronto Dataset.

	Neighborhood_	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Antique Shop	Arcade
0	Wakefield	0	0	0	0	0	0	0	0
1	Wakefield	0	0	0	0	0	0	0	0
2	Wakefield	0	0	0	0	0	0	0	0
3	Wakefield	0	0	0	0	0	0	0	0
4	Wakefield	0	0	0	0	0	0	0	0

Figure 15: one hot encoding on Newyork data

	Neighborhood_	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium
0	Malvern, Rouge	0	0	0	0	0	0	0	0	0	0	0
1	Rouge Hill, Port Union, Highland Creek	0	0	0	0	0	0	0	0	0	0	0
2	Rouge Hill, Port Union, Highland Creek	0	0	0	0	0	0	0	0	0	0	0
3	Guildwood, Morningside, West Hill	0	0	0	0	0	0	0	0	0	0	0
4	Guildwood, Morningside, West Hill	0	0	0	0	0	0	0	0	0	0	0

Figure 16: one hot encoding on Toronto data

The next set is to take the mean of frequency of occurrence of every category after grouping them by neighborhoods. This will transform the data frame as shown in below figures.

	Neighborhood_	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Antique Shop	Arcade
0	Allerton	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
1	Annadale	0.0	0.0	0.0	0.0	0.0	0.181818	0.0	0.0
2	Arden Heights	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
3	Arlington	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
4	Arrochar	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0

Figure 17: Average or mean dataframe of Newyork City

	Neighborhood_	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
1	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
2	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
3	Bayview Village	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
4	Bedford Park, Lawrence Manor East	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.045455	0.0	0.0

Figure 18: Average or mean dataframe of Toronto City

Below figures shows the map of both cities along with neighborhoods clustered into 5 categories.

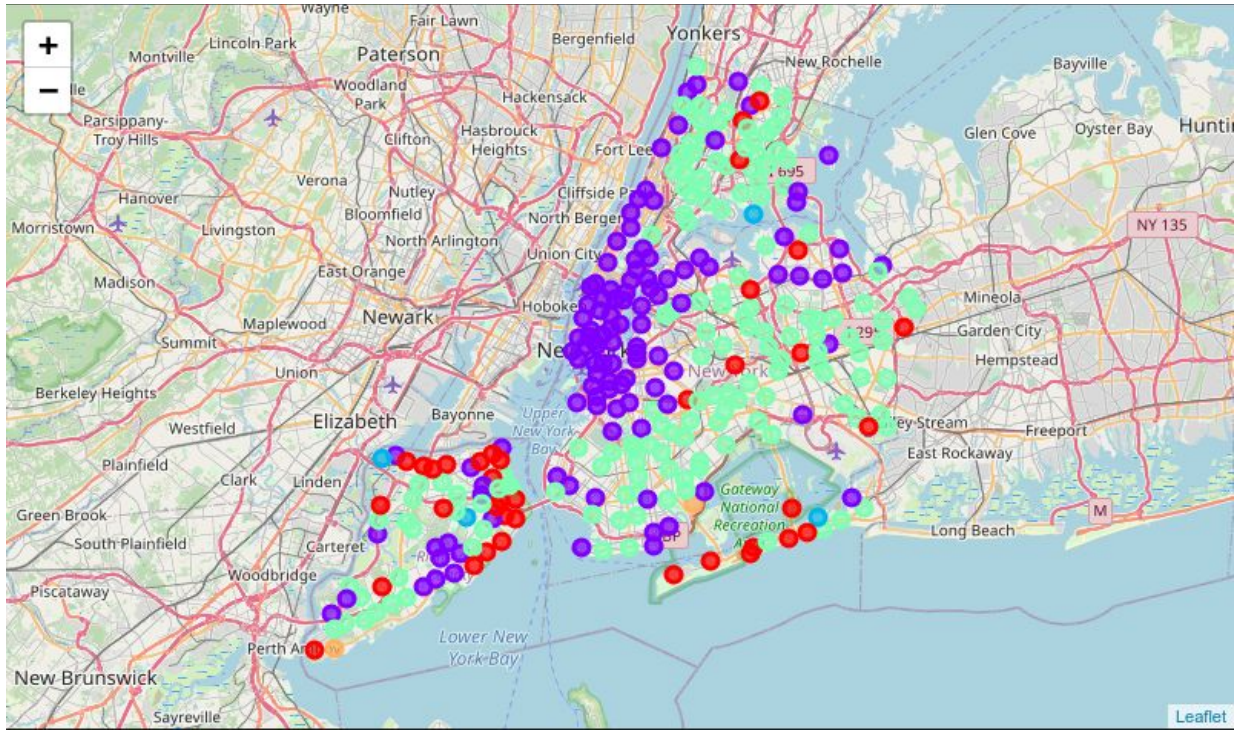


Figure 19: Clusters of Newyork City neighborhoods

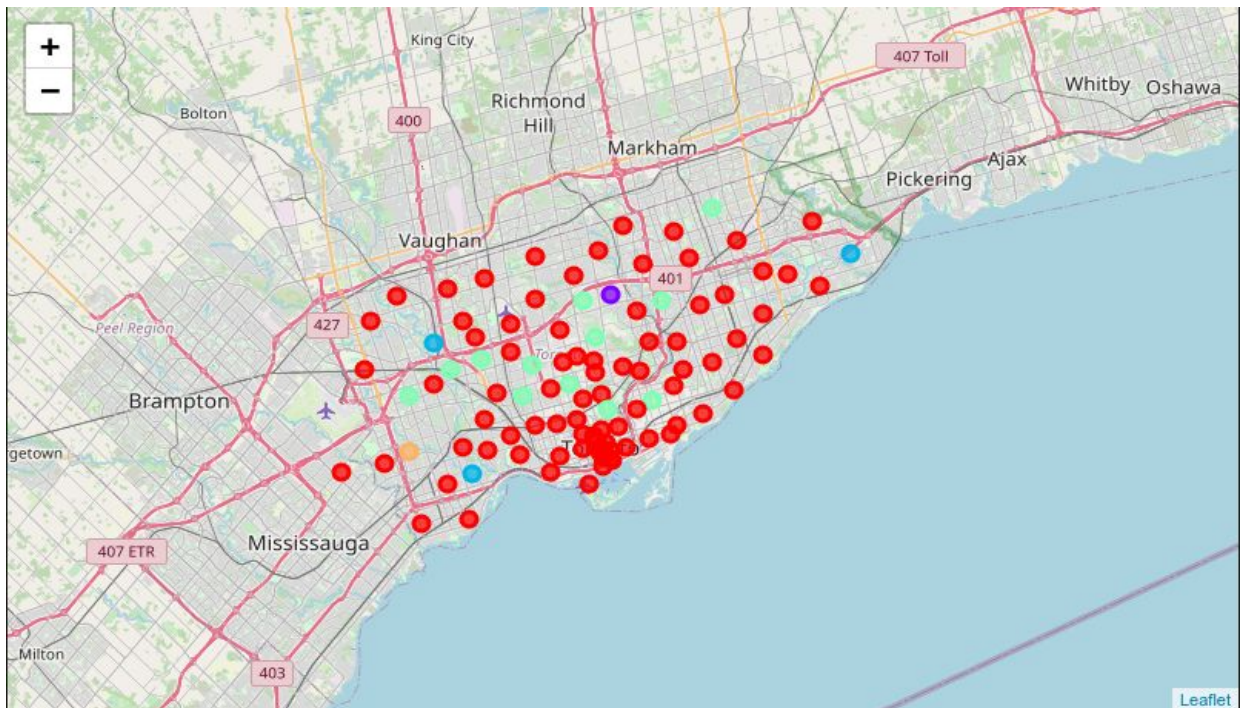


Figure 20: Clusters of Toronto City neighborhoods

Toronto and Newyork City dataset combined

After making clusters of individual cities, then we combined the dataframe of both cities to get more insights, as shown in figure below.

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

Figure 21: Combined Dataframe of Toronto and Newyork City

	Neighborhood_	1st Most Common Category	2nd Most Common Category	3rd Most Common Category	4th Most Common Category	5th Most Common Category
0	Allerton_NYC	Pizza Place	Chinese Restaurant	Spa	Deli / Bodega	Supermarket
1	Annadale_NYC	Pizza Place	American Restaurant	Diner	Train Station	Food
2	Arden Heights_NYC	Deli / Bodega	Pharmacy	Bus Stop	Coffee Shop	Pizza Place
3	Arlington_NYC	Deli / Bodega	Boat or Ferry	Bus Stop	Coffee Shop	Grocery Store
4	Arrochar_NYC	Bus Stop	Deli / Bodega	Italian Restaurant	Athletics & Sports	Supermarket

Figure 22: Common categories for combined neighborhoods of both Cities

Clustering on combined dataset of both cities

```
# the number of clusters
kclusters = 5
newyork_toronto_grouped_clustering = newyork_toronto_grouped.drop('Neighborhood_', 1)
# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(newyork_toronto_grouped_clustering)
# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]

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

Figure 23: Code used to perform K-means clustering on combined datareme

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.

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
Wingate_NYC	0	Fried Chicken Joint	Bakery	Health & Beauty Service	Pharmacy	Donut Shop	Other Great Outdoors
Woodhaven_NYC	0	Deli / Bodega	Bank	Pharmacy	Park	Donut Shop	Latin American Restaurant
Woodlawn_NYC	0	Pub	Deli / Bodega	Pizza Place	Bar	Playground	Cosmetics Shop
Woodrow_NYC	0	Pharmacy	Grocery Store	Bakery	Donut Shop	Sushi Restaurant	Martial Arts School
Woodside_NYC	0	Grocery Store	Latin American Restaurant	Filipino Restaurant	Thai Restaurant	Bakery	Pizza Place
Yorkville_NYC	2	Italian Restaurant	Gym	Bar	Coffee Shop	Sushi Restaurant	Mexican Restaurant
Agincourt_Toronto	2	Lounge	Latin American Restaurant	Skating Rink	Clothing Store	Breakfast Spot	Czech Restaurant
Alderwood, Long Branch_Toronto	0	Pizza Place	Sandwich Place	Gym	Pharmacy	Coffee Shop	Pub
Bathurst Manor, Wilson Heights, Downsview North_Toronto	0	Bank	Coffee Shop	Diner	Shopping Mall	Chinese Restaurant	Sandwich Place
Bayview Village_Toronto	0	Chinese Restaurant	Bank	Japanese Restaurant	Café	Farmers Market	Entertainment Service

Figure 24: Clusters and most common categories of neighborhoods of both Cities

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
Butler Manor_NYC	3	Baseball Field	Pool	Yoga Studio	Fast Food Restaurant	Entertainment Service	Escape Room
Humberlea, Emery_Toronto	3	Baseball Field	Yoga Studio	Farmers Market	Empanada Restaurant	Entertainment Service	Escape Room

Figure 25: Data included in 4th cluster

In the figure above we can see how similar are the neighborhoods of 2 cities i.e. Both Bulter Manor and Emery have the baseball field, Yoga Studio and Entertainment Service as the most common Venue categories in Newyork and Toronto respectively

Results and Analysis of Clusters

Cluster 1		Cluster 2	
	% of venues		% of venues
Pizza Place	6.29871	Park	46.1538
Deli / Bodega	4.06573	Convenience Store	7.69231
Chinese Restaurant	3.20202	Playground	5.12821
Pharmacy	3.09669	Pool	5.12821
Donut Shop	2.9703	Bakery	5.12821
Bank	2.82284	Trail	2.5641
Grocery Store	2.82284	Boat or Ferry	2.5641

Cluster 3	
	% of venues
Coffee Shop	5.77617
Italian Restaurant	3.34269
Café	3.03517
Bar	2.94157
Pizza Place	2.52708
American Restaurant	2.07247
Bakery	2.07247

Cluster 4		Cluster 5	
	% of venues		% of venues
Baseball Field	57.1429	Boat or Ferry	66.6667
Pool	28.5714	Grocery Store	33.3333
Construction & Landscaping	14.2857		

Figure 26: 5 clusters of combined data of both cities

The differences between the clusters can be seen from the figure; each cluster distinguishably has a different distribution of common venue categories than other clusters.

Here are some key points that are clearly visible from above figure

1. 1st cluster majorly consist of Pizza Place, Deli/Bodega and Chinese Restaurant and these make about 14% of Venues
2. 2nd cluster consists of Park, Convenience Store and Playground as most common categories
3. Italian restaurants appear as the most common categories of the 3rd cluster only.
4. The 4th cluster has 57% of Baseball Field as the most common categories.
5. Boat or Ferry is most common in 5th cluster

In the figure below we have tried to get the insight of clusters by plotting the bar graph between the number of neighbourhoods and clusters on the x axis. Also in the graph below blue bars are representing the Newyork City and orange bars are representing Toronto City.

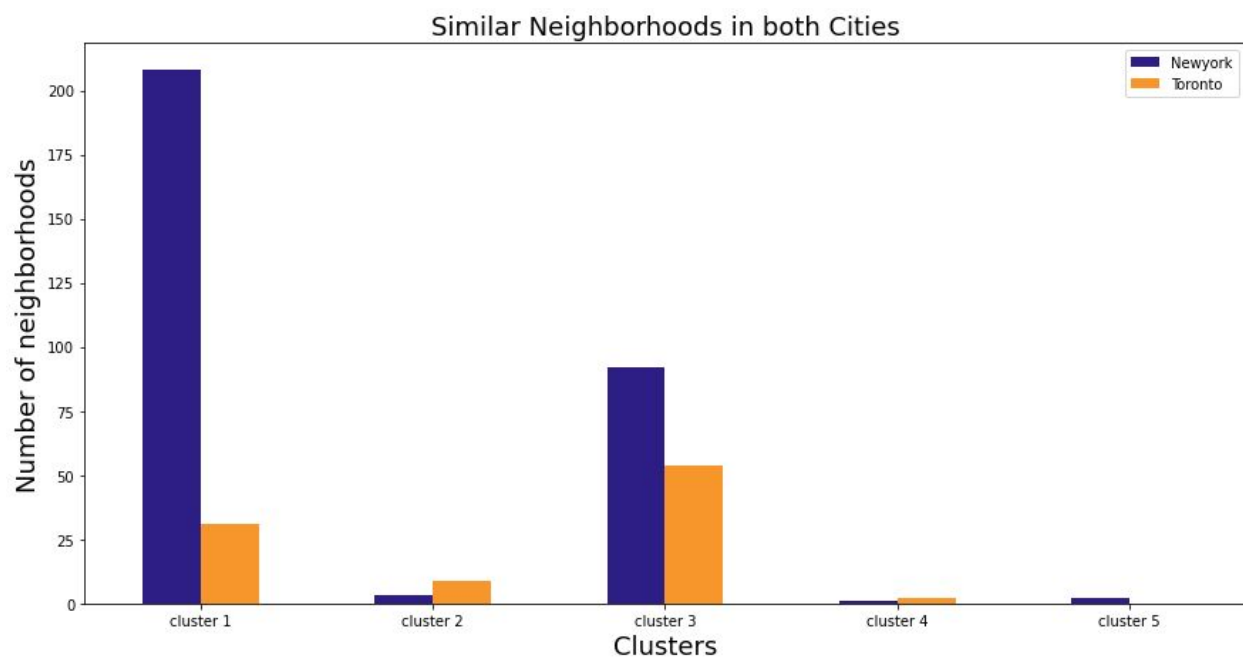


Figure 27: Neighborhoods of both Cities in each cluster

Conclusions

In this project, the areas of New York City and Toronto were clustered into various groups dependent on the classifications (kinds) of the venues in these areas. The outcomes demonstrated that there are venue categories that are more common in certain groups than the others; Also these most common venue categories contrast from one cluster to the next. So using this information one can make decisions and be able to find similar neighbourhoods in the new City. It will also help business people to get answers to questions like 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.. In the future if a more profound investigation is performed considering more viewpoints, it may bring about finding various styles in each cluster dependent on the most common categories in the cluster.