Clustering and comparing the Neighborhood Of

Mansoor Bari

July 27, 2021

Section 1

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 in United States of America and Toronto which is 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 businesspeople 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 (NYC) is one of the most populous cities 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.

Image 1: New York map

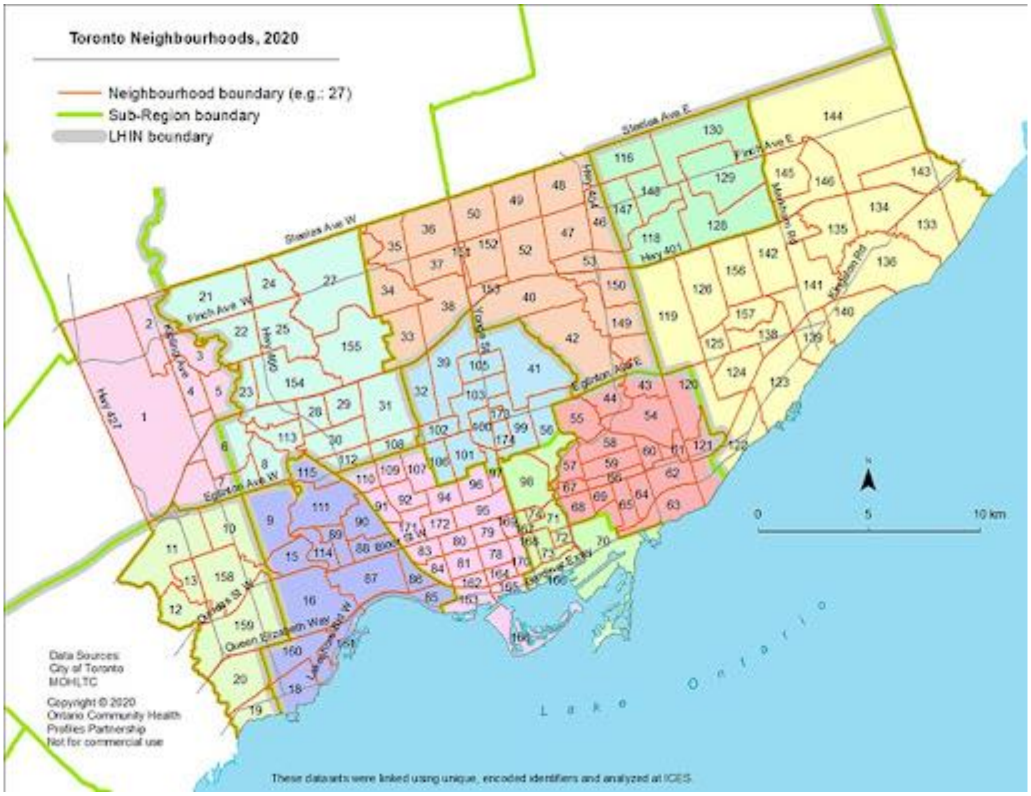


Image2: map of Toronto

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 center for business and finance, and it is considered the financial capital of Canada.

Section 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.
Venue's 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 Moriscos Seafood.
Julio C Barber Shop 2 Salon / Barbershop

Table 1: Example of the venues data

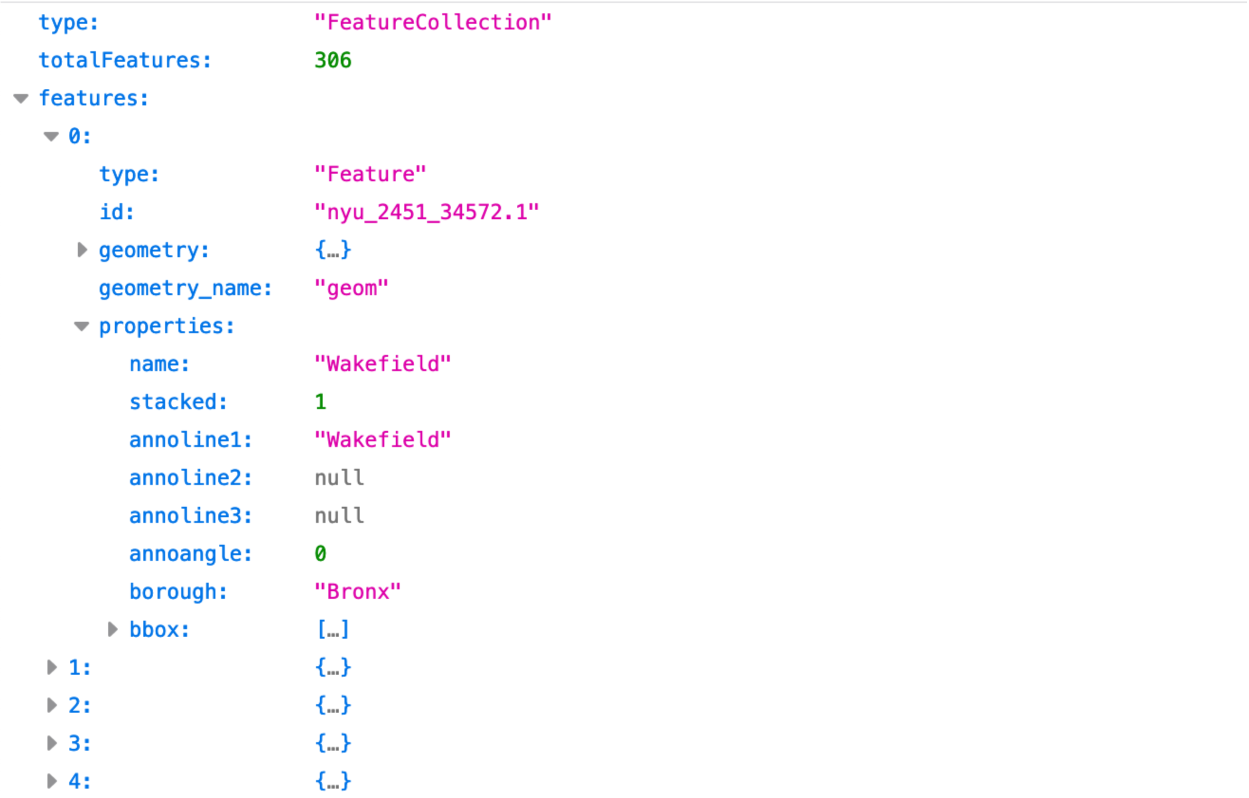
This data will be retrieved from Foursquare which is one of the world largest sources of location and venue data. Foursquare API will be utilized to get and download the data.

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.



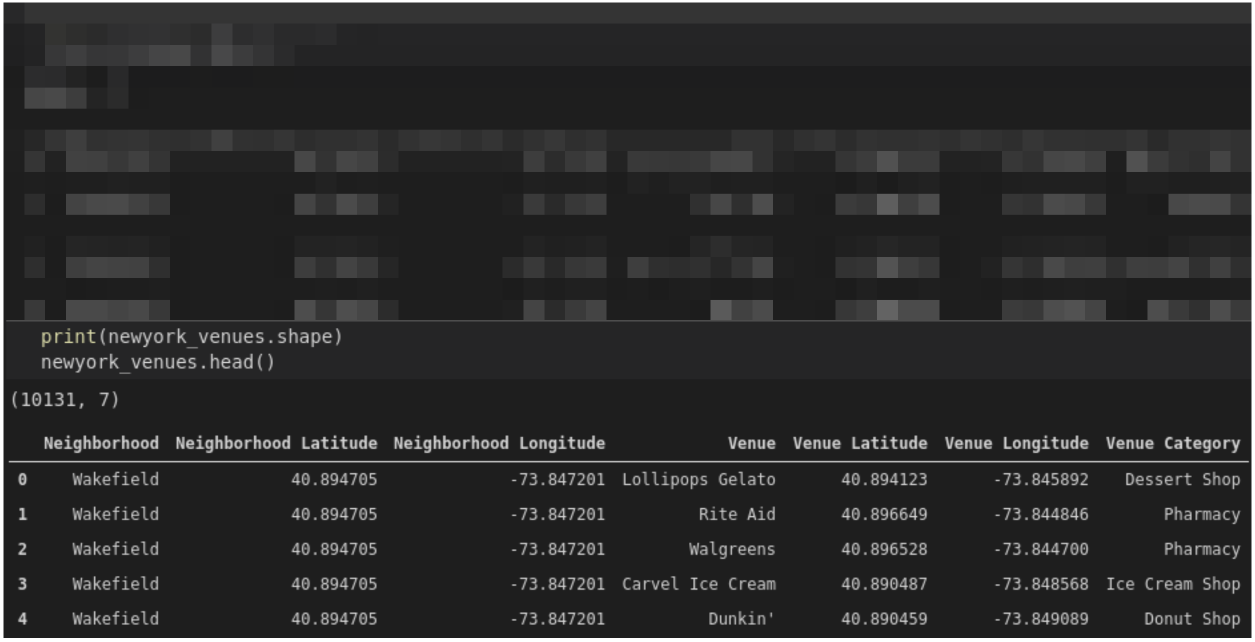
```
  type:                 "FeatureCollection"
  totalFeatures:        306
▼ features:
  ▼ 0:
      type:             "Feature"
      id:               "nyu_2451_34572.1"
    ▶ geometry:         {…}
      geometry_name:    "geom"
    ▼ properties:
        name:           "Wakefield"
        stacked:        1
        annoline1:      "Wakefield"
        annoline2:      null
        annoline3:      null
        annoangle:      0
        borough:        "Bronx"
      ▶ bbox:           […]
  ▶ 1:                  {…}
  ▶ 2:                  {…}
  ▶ 3:                  {…}
  ▶ 4:                  {…}
```

Image 3: JSON file to describe the specification

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 inturn 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.
The first five rows of the resulting dataframe of New York City are shown in the figure below.



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

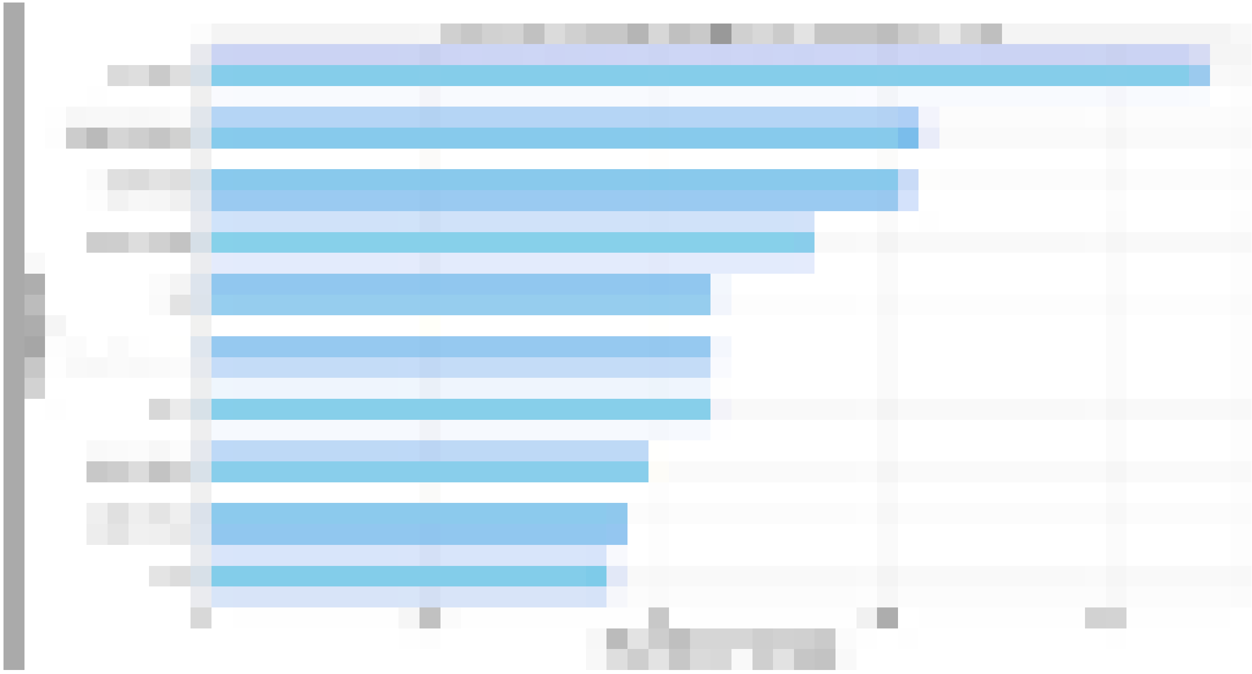
```
(10131, 7)

   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 6: Venue dataframe for New York City
Data 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 New York City.



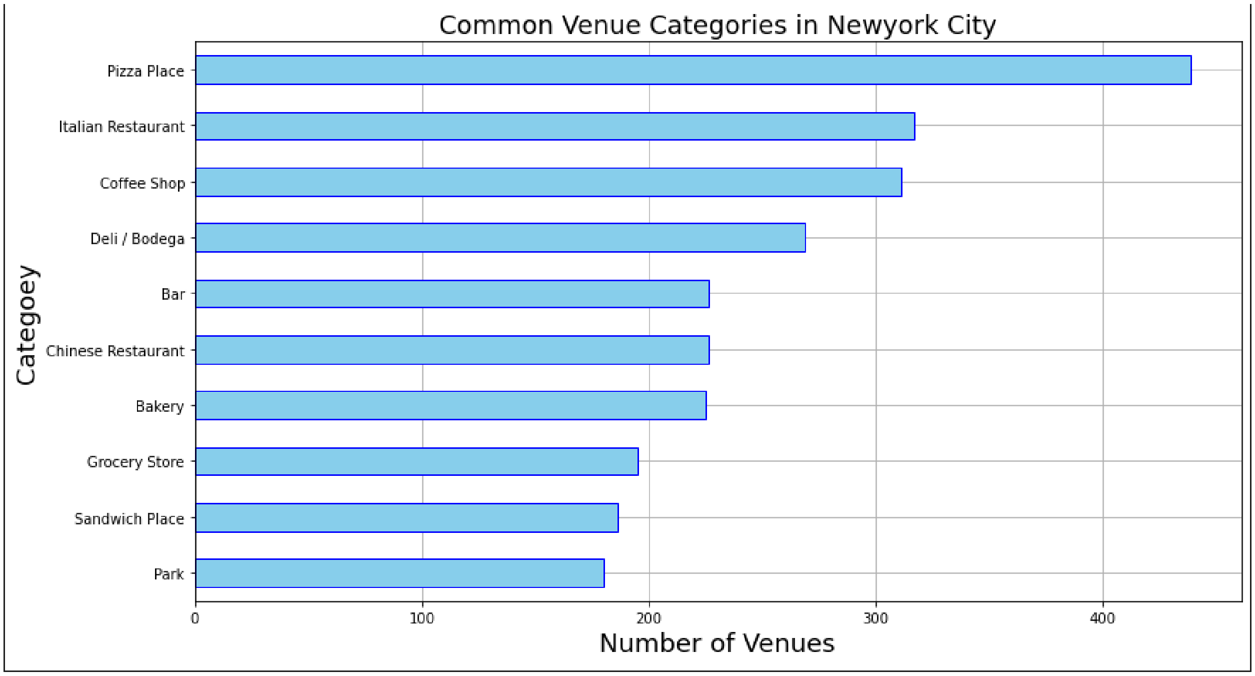
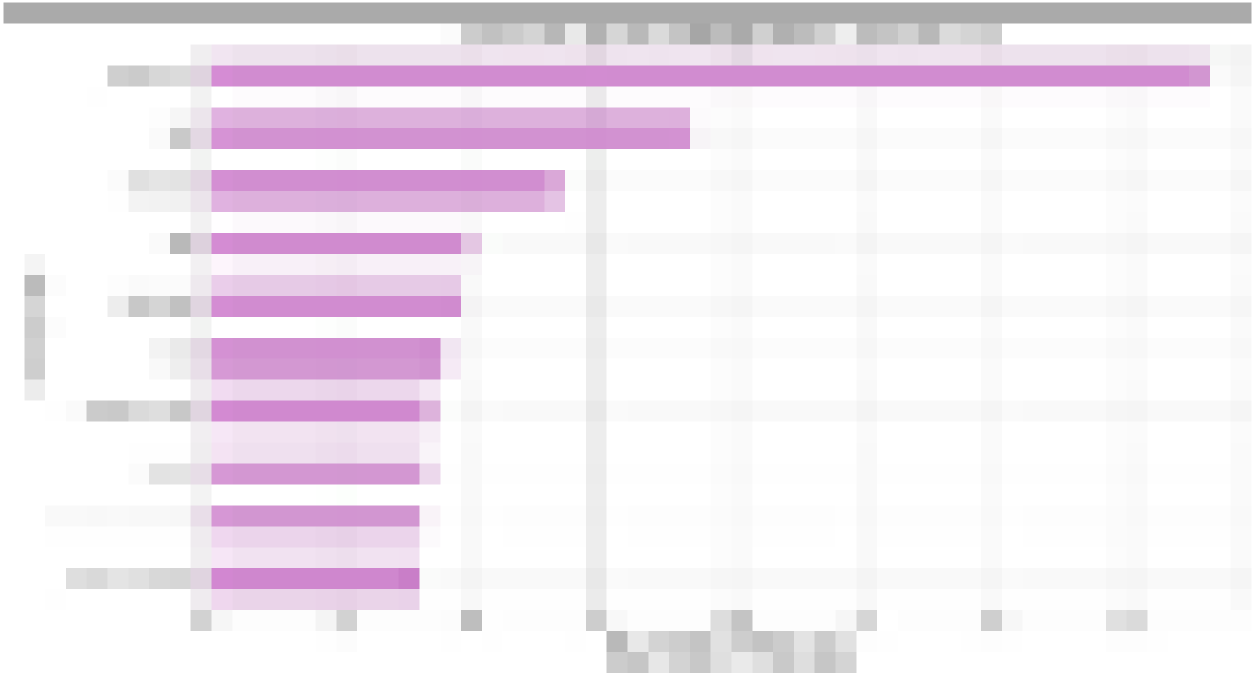


Figure 7: Common Venues for New York City



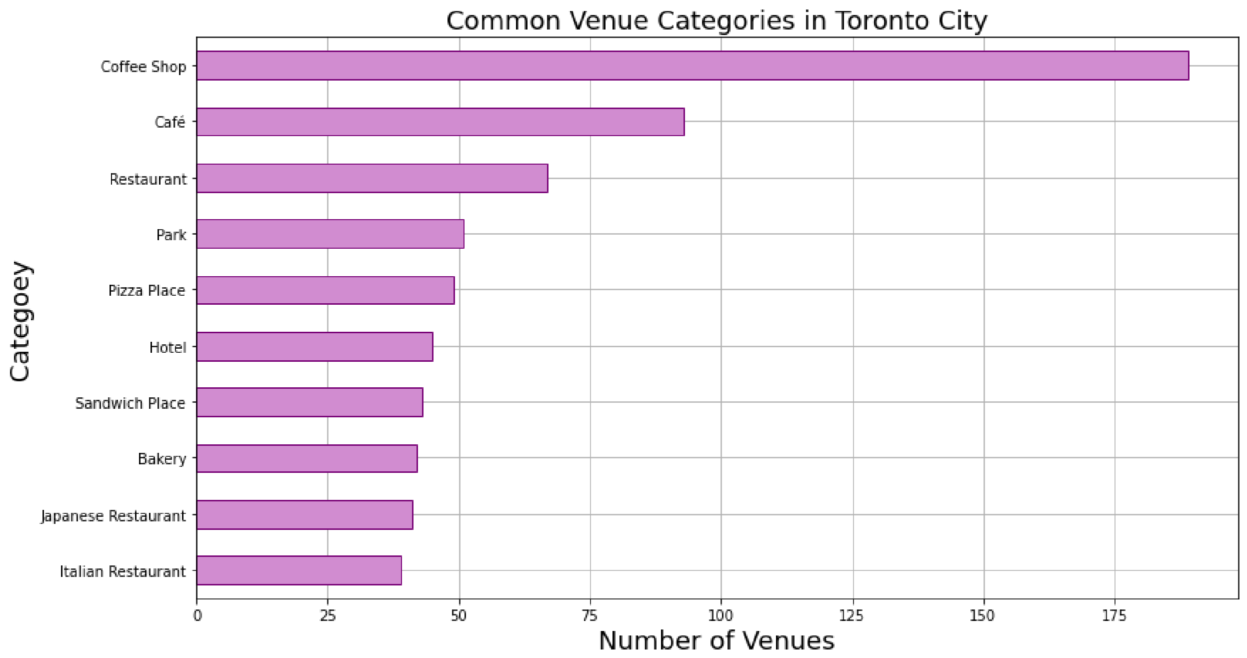


Figure 8: Common Venues for Toronto City

We can see that the most common categories for New York 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.

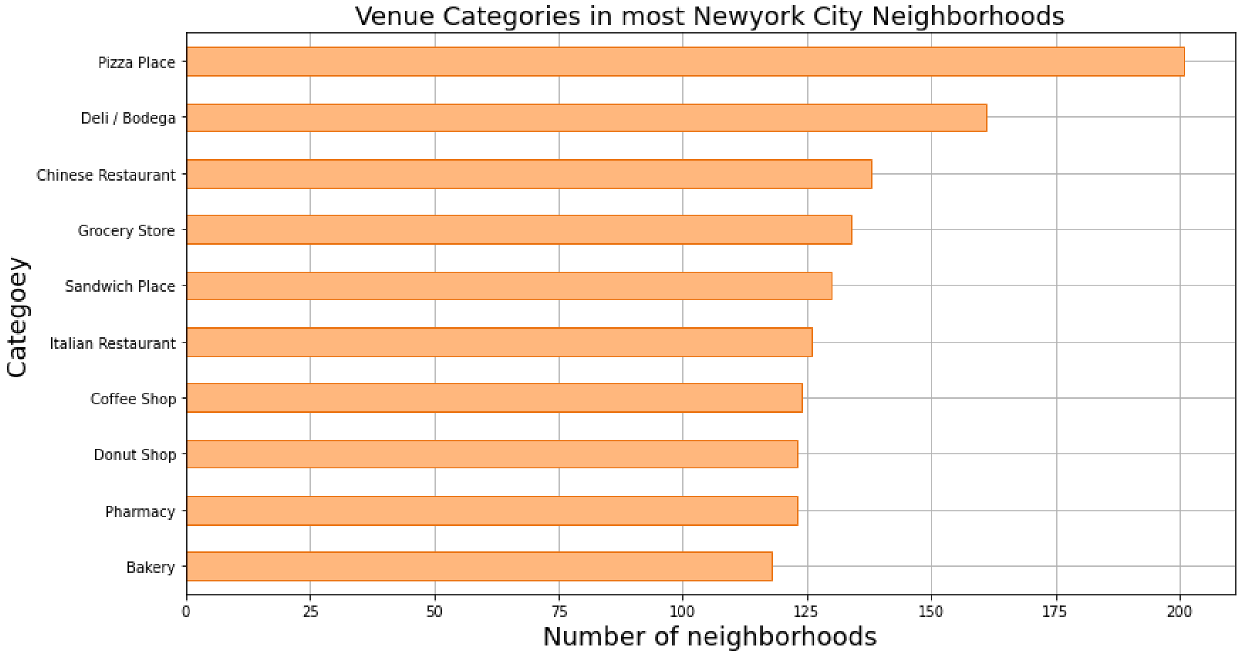
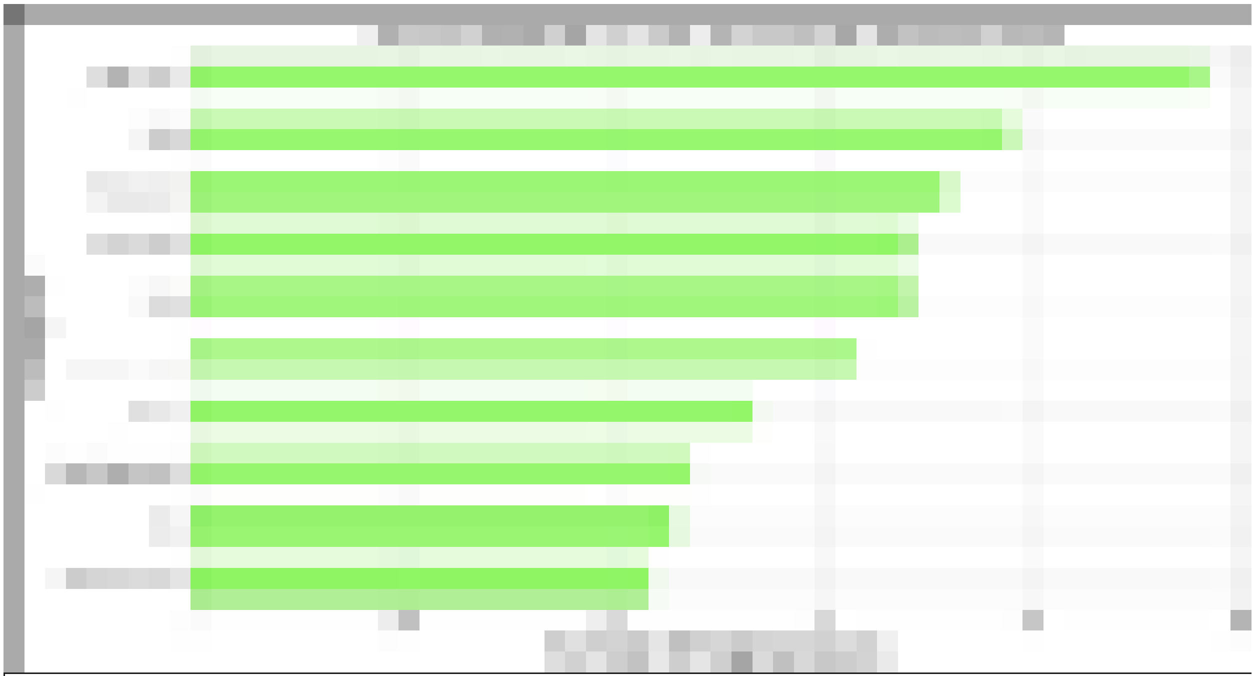


Figure 9: Most widespread venue categories for New York city



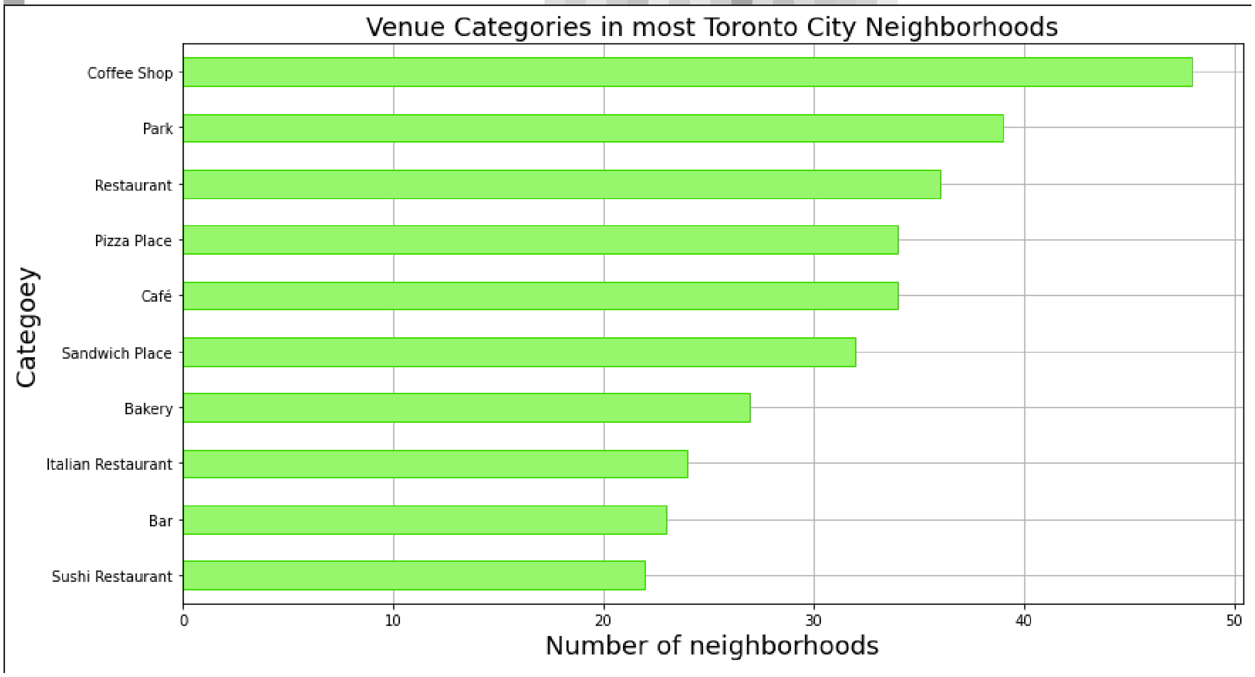
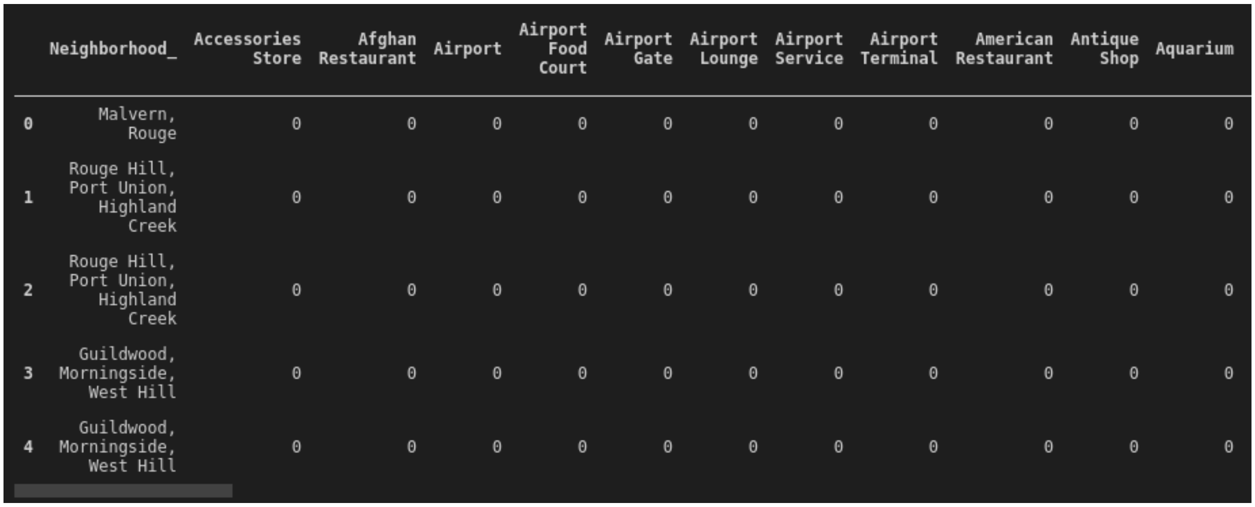


Figure 10: Most widespread venue categories in Toronto city

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 New York 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 New York and Toronto Dataset.

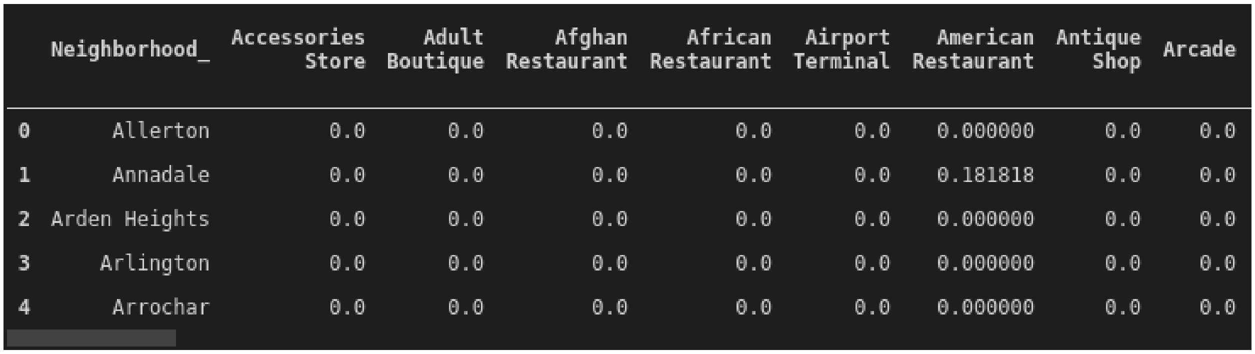
Figure 11: one hot encoding on Newyork data



	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

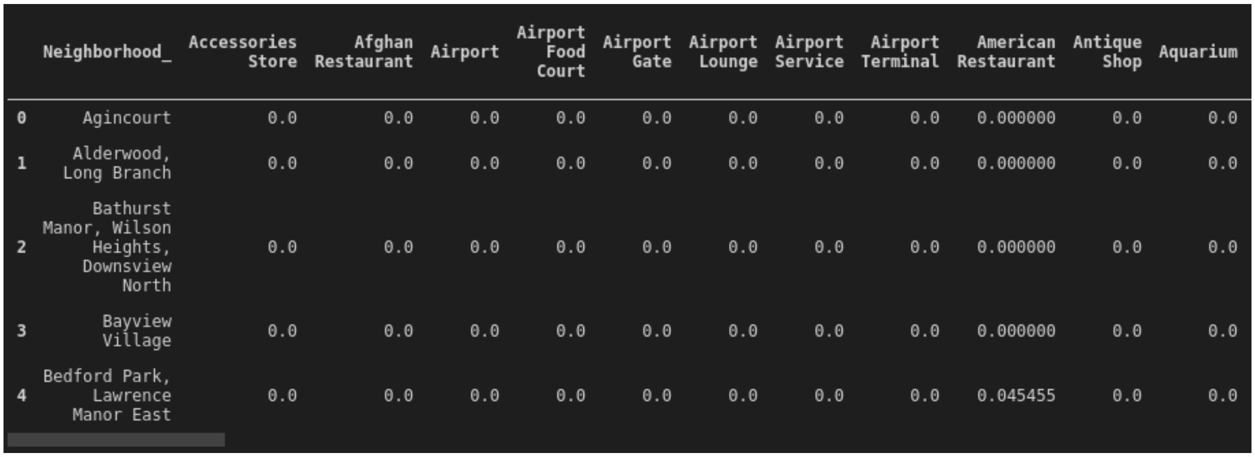
Figure 12: one hot encoding on Toronto data

The next step 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 13: Average or mean datafame of New York 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 14: Average or mean datafame of Toronto City

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

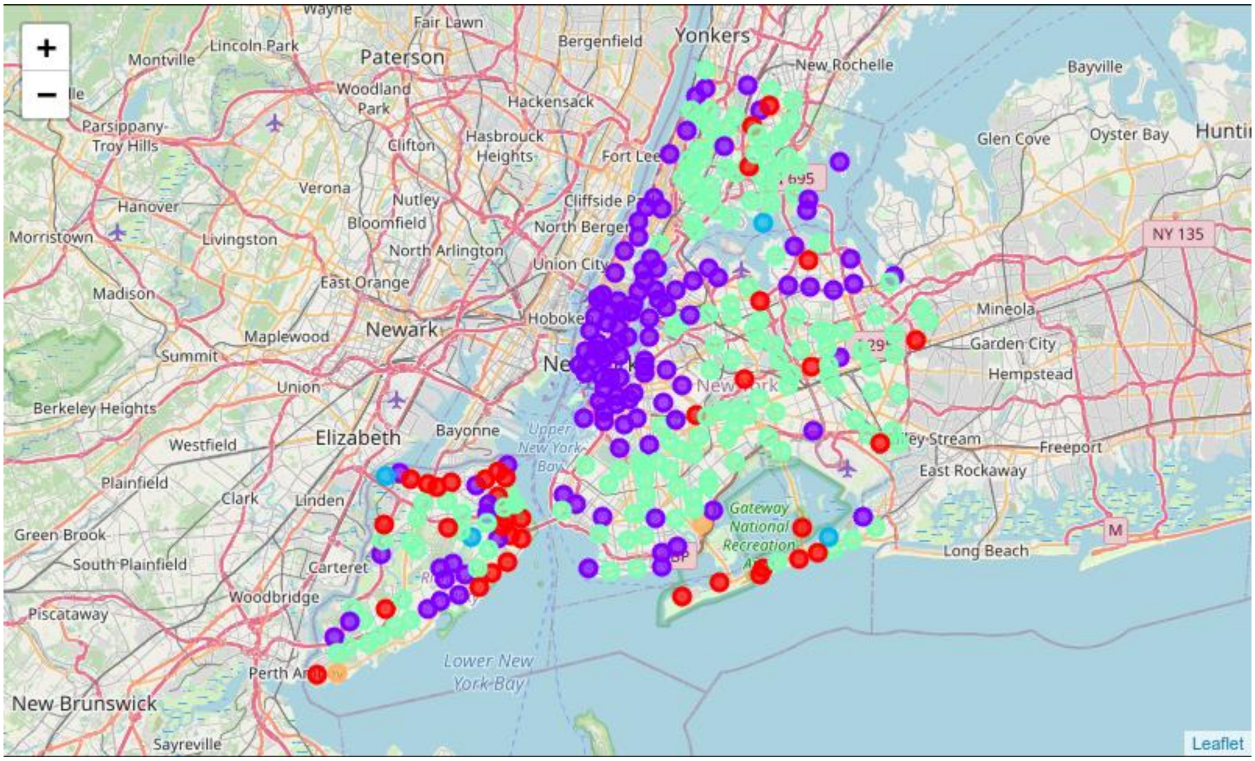


Figure 15: Clusters of New York City neighborhoods

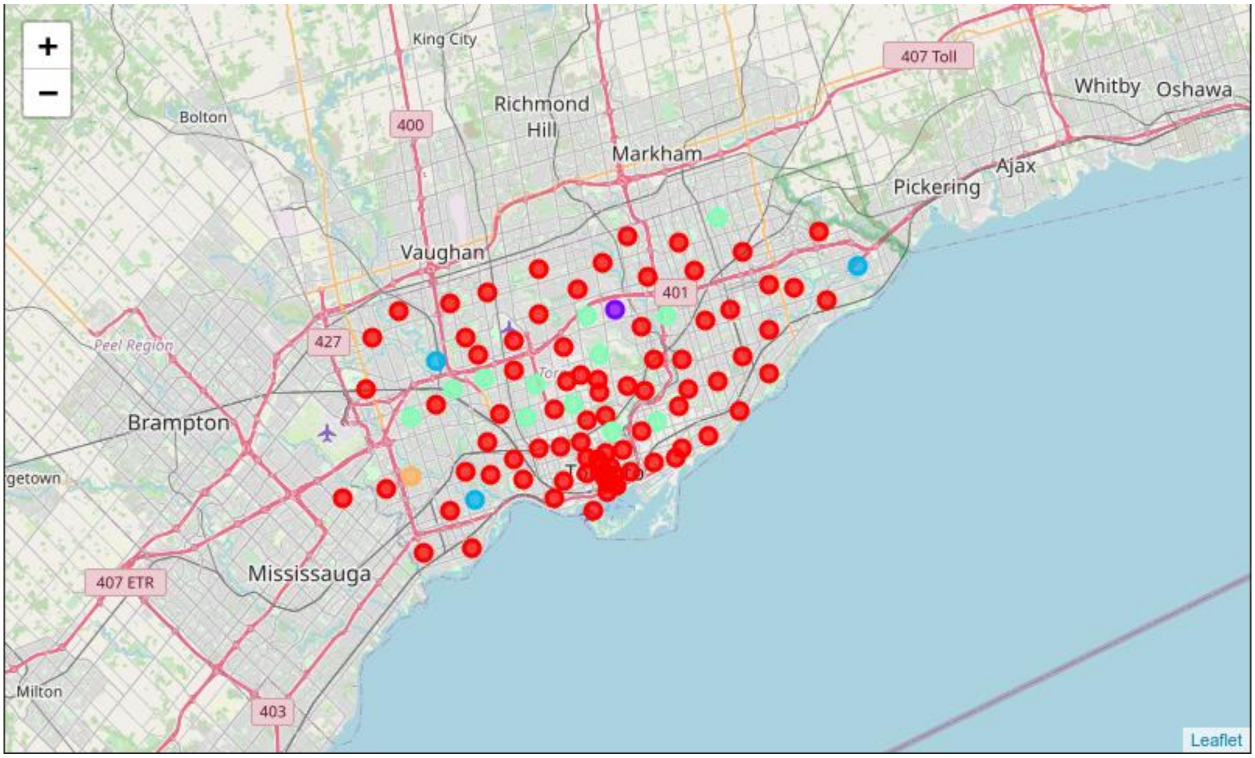
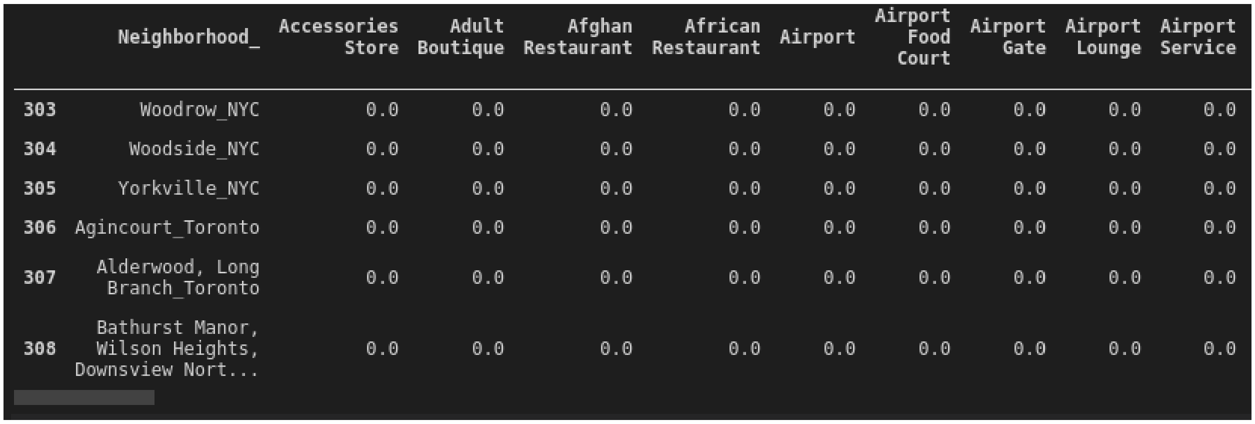


Figure 16: Clusters of Toronto City neighborhoods

Toronto and New York City dataset combined

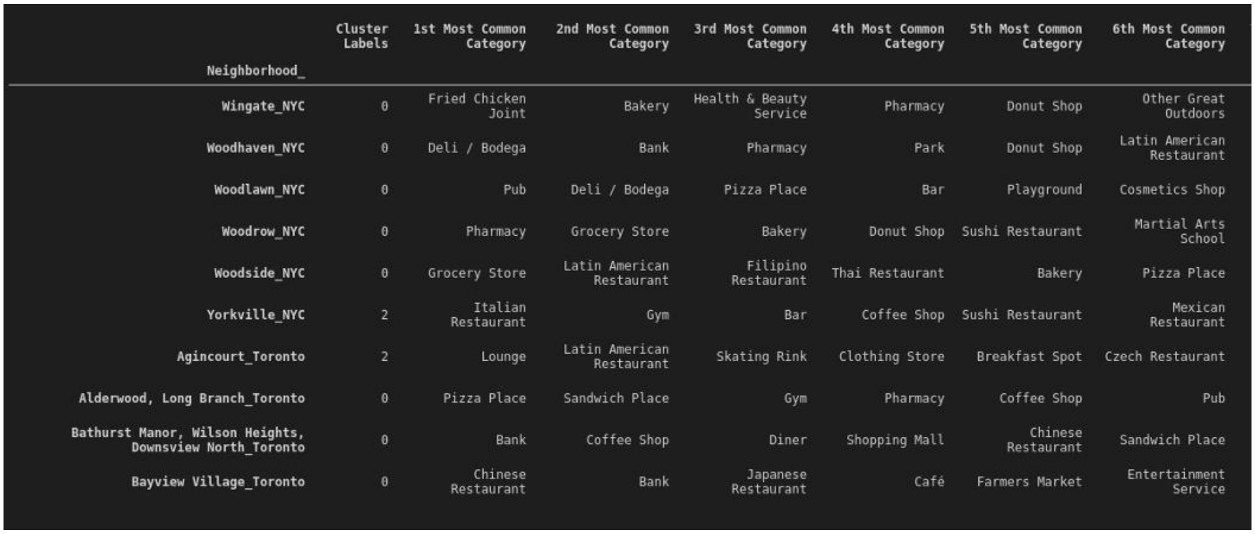
After making clusters of individual cities, then we combined the dataframe of both cities to get more insights, as shown in figure below.



	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

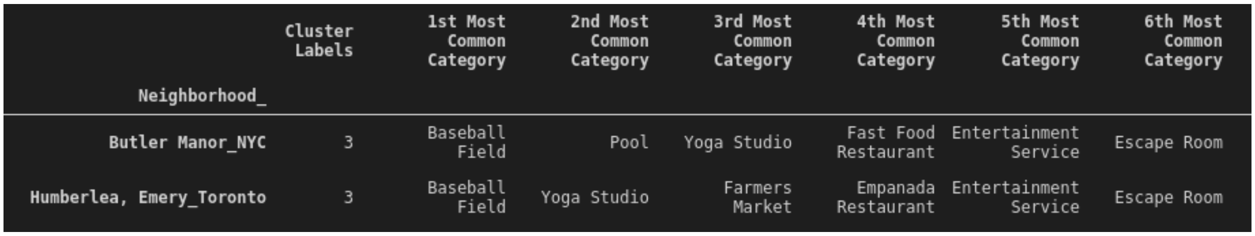
Figure 17: Combined Dataframe of Toronto and New York City

Then make clusters from new combined dataframe and 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.



	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
Neighborhood_							
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 18: Clusters and most common categories of neighborhoods of both Cities



	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
Neighborhood_							
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 19: Data included in 4th cluster

In the figure above we can see how similar the neighborhoods of 2 cities are i.e. Both Butler Manor and Emery have the baseball field, Yoga Studio and Entertainment Service as the most common Venue categories in New York and Toronto respectively.
In the figure below we have tried to get the insight of clusters by plotting the bar graph between the number of neighborhoods and clusters on the x axis. Also in the graph below blue bars are representing the New York City and orange bars are representing Toronto City.

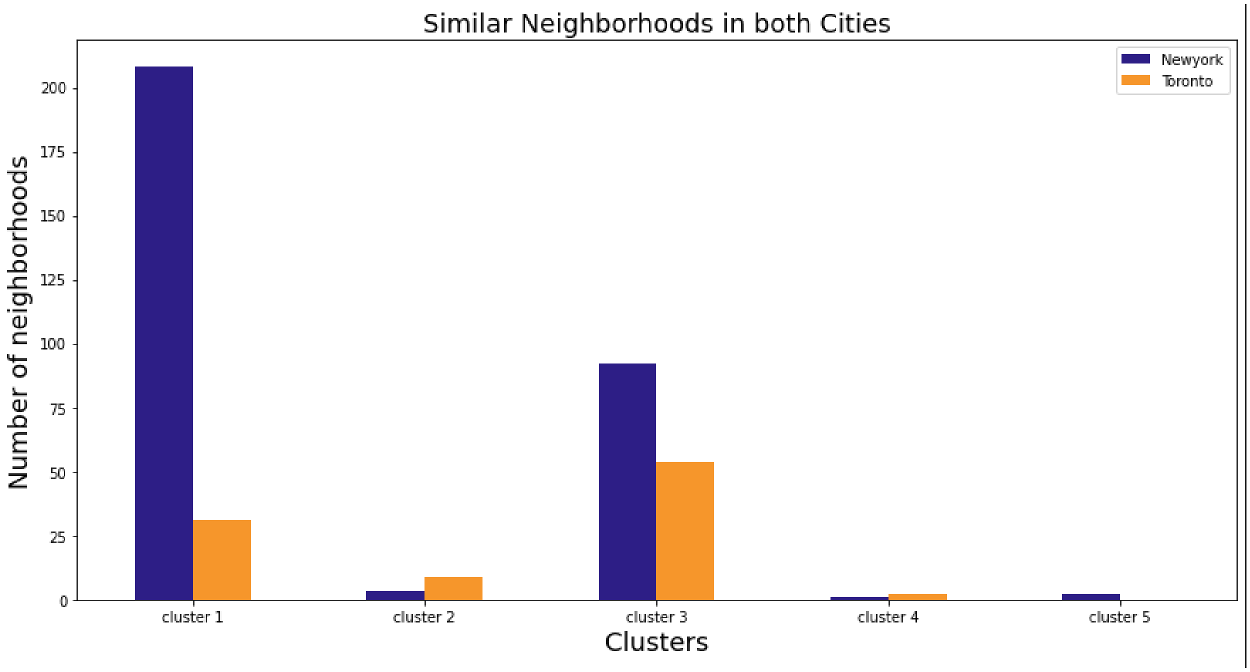


Figure 20: Neighborhoods of both Cities in each cluster

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 neighborhoods in the new City. It will also help businesspeople 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 businesspeople 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