# Exploring the cities on New York and Toronto

Raju Kalidindi

**Background and Description of the problem:** I am a software engineer working for a major bank in Minneapolis, Minnesota USA. I am currently looking for Data Engineer positions and have attended a few interviews from clients at New York City and Toronto. I currently live close to Minneapolis downtown and enjoy many amenities and venues in Minneapolis. As a regular practitioner of yoga and meditation, it is important for me to relocate to neighborhoods that have these facilities to help me to continue with my interests. I would like to relocate to a city, which has similar or better facilities than Minneapolis. Since I am close to getting offers from clients located in New York and Toronto, I would like to use the skills gained in this course to compare the different venues and amenities in New York and Toronto cities.

**Description of the data that will be used to solve the problem:** Data for New York city will be downloaded from https://cocl.us/new_york_dataset. Data for Toronto postal codes will be downloaded from https://data.mongabay.com/igapo/toronto_zip_codes.htm. For the venues and amenities in New York and Toronto, we will utilize Foursquare application. We will use data wrangling and k-Means clustering to explore the data for the two cities.

**Manhattan and Toronto datasets used for analysis:** Sample of the datasets used for solving the business problem are shown below.

Manhattan Dataset

| | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|
| 0 | Manhattan | Marble Hill | 40.876551 | -73.910660 |
| 1 | Manhattan | Chinatown | 40.715618 | -73.994279 |
| 2 | Manhattan | Washington Heights | 40.851903 | -73.936900 |
| 3 | Manhattan | Inwood | 40.867684 | -73.921210 |

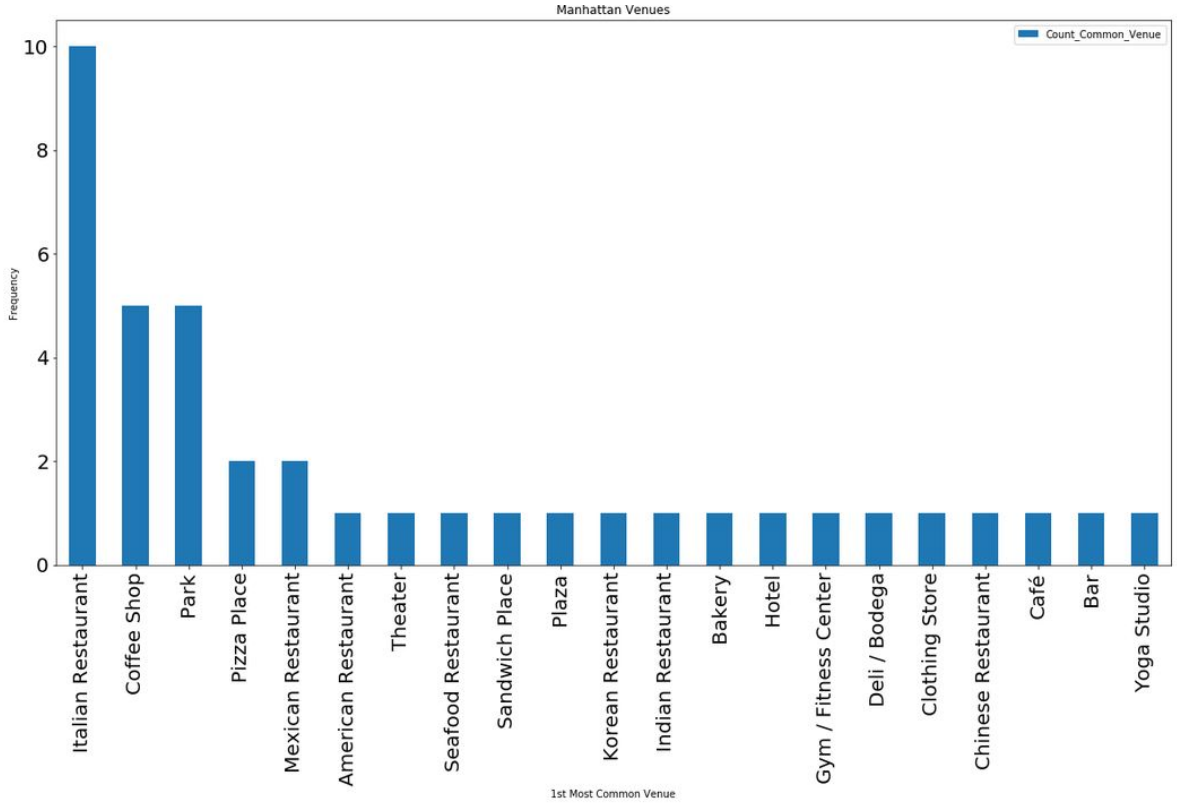
Toronto Dataset:

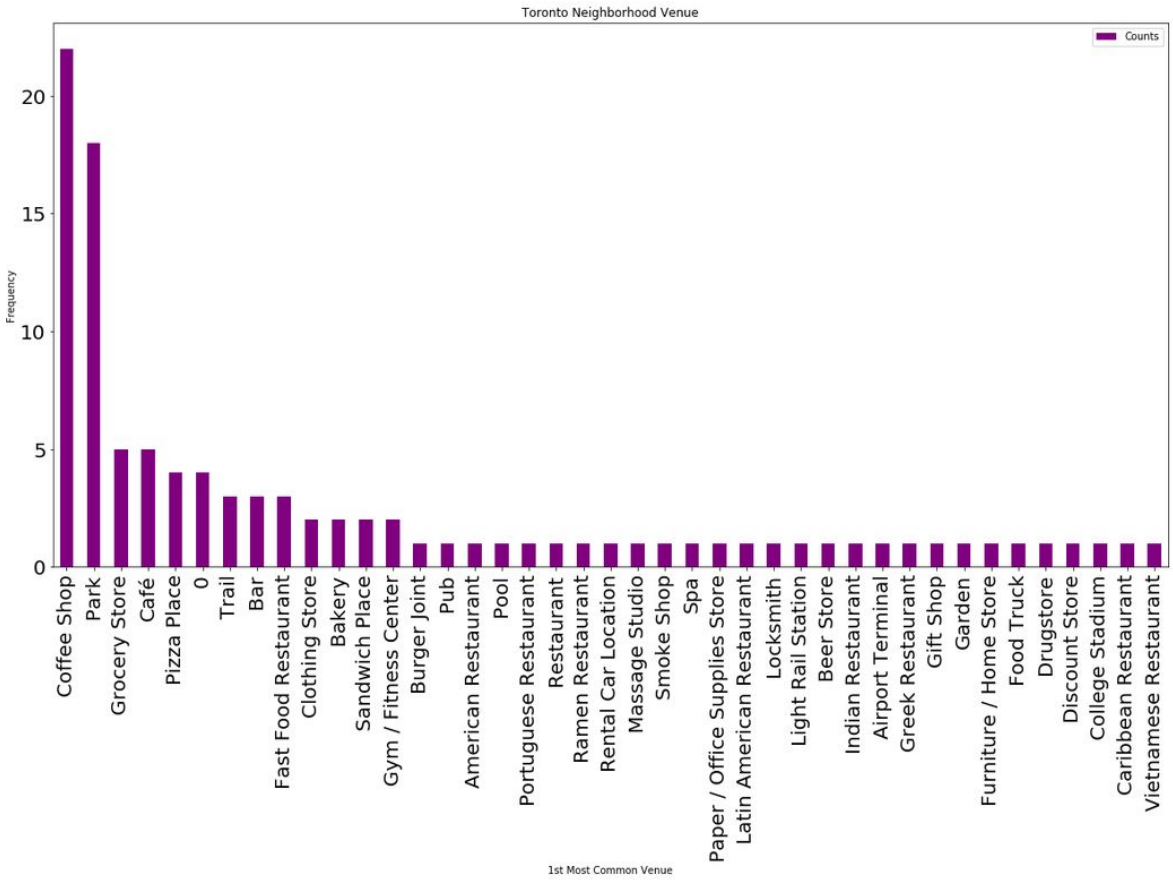
| | PostalCode | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | M1B | Scarborough | Rouge, Malvern | 43.806696 | -79.194353 |
| 1 | M1C | Scarborough | Highlan Creek, Rouge Hill, Port Union | 43.784535 | -79.160497 |
| 2 | M1E | Scarborough | Guildwood, Morningside, West Hill | 43.763573 | -79.188711 |
| 3 | M1G | Scarborough | Woburn | 43.7770992 | -79.216917 |

The above datasets have the names of the neighborhoods of the above cities and the latitude and longitude coordinates.

**Methodology:** The datasets are processed and any rows with missing information are filtered out from the datasets. The top ten most common venues for each neighborhood are identified. Next the K-means algorithm is applied to classify the venues into clusters. The logarithm is then used to assign labels to each of the clusters and finally the first most common venue for each neighborhood is identified.

**Results:** From the resulting dataframes, the data is grouped for the first most common venue and then we count the top four most common venues among neighborhoods. We plot the most common venues on a bar chart to visualize the same. We also find neighborhoods corresponding to the top four venues of my interests, for each of the city in order to predict which neighborhood will work better for me as I prefer neighborhoods close to a gym and fitness center.






The above two plots show the most common venues in both the cities.

**Discussions:** Based on the above data, I can list neighborhoods which are close to the venues of my interests. See details below:

```
In [87]: #Getting the neighborhoods of Manhattan for the top 4 most common venues
         # list_top_venues based on my interests
         list_top_venues = ['Park','Boat or Ferry','Yoga Studio', 'Gym / Fitness Center']
         print('----------------------------------------------------')
         for venue in list_top_venues:
             print('The nearest neighborhood to the ' + venue + ' is : ')
             locator = manhattan_new.loc[manhattan_new['1st Most Common Venue'] == venue]
             print(locator['Neighborhood'].values)
             print('----------------------------------------------------')

         ----------------------------------------------------
         The nearest neighborhood to the Park is :
         ['Morningside Heights' 'Battery Park City' 'Tudor City']
         ----------------------------------------------------
         The nearest neighborhood to the Boat or Ferry is :
         ['Stuyvesant Town']
         ----------------------------------------------------
         The nearest neighborhood to the Yoga Studio is :
         ['Flatiron']
         ----------------------------------------------------
         The nearest neighborhood to the Gym / Fitness Center is :
         ['Civic Center']
         ----------------------------------------------------
```

```
In [85]: #Getting the neighborhoods for the top 4 most common venues
         # list_top_venues based on my interests
         list_top_venues = ['Gym / Fitness Center','Park','Trail', 'Pool']
         print('----------------------------------------------------')
         for venue in list_top_venues:
             print('The nearest neighborhood to the ' + venue + ' is : ')
             locator = toronto_new_df.loc[toronto_new_df['1st Most Common Venue'] == venue]
             print(locator['Neighborhood'].values)
             print('----------------------------------------------------')

         ----------------------------------------------------
         The nearest neighborhood to the Gym / Fitness Center is :
         ['Don Mills North' 'Downsview Northwest']
         ----------------------------------------------------
         The nearest neighborhood to the Park is :
         ["Agincourt North, L'Amoreaux East, Milliken, Steeles East"
          'York Mills West' 'Parkwoods' 'CFB Toronto, Downsview East'
          'East Toronto' 'Lawrence Park' 'Moore Park, Summerhill East' 'Rosedale'
          'Forest Hill North, Forest Hill West' 'Caledonia-Fairbanks'
          'Downsview, North Park, Upwood Park' 'Weston']
         ----------------------------------------------------
         The nearest neighborhood to the Trail is :
         ['The Beaches' 'Humewood-Cedarvale']
         ----------------------------------------------------
         The nearest neighborhood to the Pool is :
         ['Hillcrest Village']
         ----------------------------------------------------
```

From the above results, I can decide which neighborhoods in Manhattan or Toronto have venues of my interests close by.

**Conclusion:** In this Capstone project, we analyzed the neighborhoods of Manhattan, New York and Toronto, CA cities to determine the most common venues in both cities. We could also list neighborhoods which were close to the venues of my interests.