Capstone Final Project: *A data-based comparison between neighborhoods of Toronto and Manhattan (NY) – a simple similarity comparison of venues*

Chiranjib Dutta – July 15, 2019

Section 1: Introduction and Objective

New York (NY) City is one of the most vibrant and attractive cities in the world. There are many other cities around the globe which are compared to NY in various aspects such as economic growth standard, living standard, new business opportunities etc. There are many published articles on these kinds of comparisons based on various surveys. One of the important aspects of economic growth is the tourism industry and NY is one of the top-most places that people like to visit. However, there are other cities in the world who are also making themselves to the top of the lists people like to visit usually. A basic data-based similarity comparison of those cities considering Manhattan area as a baseline would help a lot in making the decision whether to visit the city or not easier for common people for a relaxing weekend. In this project, we choose another NY like city, Toronto in Canada to compare it with Manhattan, NY for its various venues. The idea is to compare the neighborhoods (Manhattan and Toronto) that we came across in our previous modules in the course with the data in hand and observe the similarity of the venues that the people would come across when they visit both the cities. Similar approach can be executed when we want to compare any other cities with NY and make conclusions how similar or different their neighborhoods are for a specific purpose. Note that this is just an example that is chosen to be the criterion for this project

Section 2: Data collection, filtering and pre-processing

We will use the data from our previous modules that we completed. The NY data can be accessed from the below link:

https://geo.nyu.edu/catalog/nyu_2451_34572

The above link has data set which contains data from 5 boroughs and 306 neighborhoods of the NY city. In addition, it also contains the longitude and latitude of each neighborhood. We have used this data in one of the exercises in week 3 of the course. However, similar data for Toronto is not readily available. Hence, we will execute two steps to gather similar data for Toronto neighborhood like what we have for NY city.

Step1: We scrape the data from Wikipedia as we have done this in one of the assignments in week 3. The following link contains the borough and neighborhood information for Toronto but not the longitude and latitude:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Step2: We use the geospatial data from the below link to synchronize the co-ordinates of the neighborhoods (longitude and latitude) which we already have from Step 1. This is what we have done in the same previous assignment in week3:

http://cocl.us/Geospatial_data

Then we will use the Foursquare API to explore the neighborhoods of both the cities and do a bit of clustering (K-means). Then we will compare the venues based on various categories. Finally, we will derive some results and quantify the similarity of both the cities in terms of what normally people would find in respective neighborhoods.

The details of the data processing and pre-processing are explained in the following sections

Section 3: Data analysis – Manhattan Neighborhood

Once the data is retrieved from the IBM provided source as below:

```
!wget -q -O 'newyork_data.json' https://cocl.us/new_york_dataset
print('Data downloaded!')
with open('newyork_data.json') as json_data:
    newyork_data = json.load(json_data)
```

I have investigated the key feature called "*features*" which gives all the necessary information for our analysis. Below is an example snapshot:

```
In [4]: neighborhoods_data = newyork_data['features']
        neighborhoods_data[0]   ## First example of this new data

Out[4]: {'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]}}
```

The next step is to create a Pandas data frame to work with which contains only the neighborhood of Manhattan. Below is an example data frame which contains the neighborhoods of Manhattan with their respective Latitude and Longitude.

In [7]:
```
manhattan_data = neighborhoods[neighborhoods['Borough'] == 'Manhattan'].reset_index(drop=True)
manhattan_data.head()
```

Out[7]:

	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
4	Manhattan	Hamilton Heights	40.823604	-73.949688

Then using Foursquare API (with a limit of 100), I explored the neighborhood of Manhattan

In [88]:
```
manhattan_venues_total=manhattan_venues.shape[0]
print("Size of the dataframe is " ,manhattan_venues.shape)
manhattan_venues.head(50)
```

```
Size of the dataframe is  (3324, 7)
```

Out[88]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.910660	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.910660	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.910660	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.910660	Starbucks	40.877531	-73.905582	Coffee Shop
4	Marble Hill	40.876551	-73.910660	Dunkin'	40.877136	-73.906666	Donut Shop
5	Marble Hill	40.876551	-73.910660	Blink Fitness Riverdale	40.877147	-73.905837	Gym
6	Marble Hill	40.876551	-73.910660	TCR The Club of Riverdale	40.878628	-73.914568	Tennis Stadium

It was interesting to also see how many unique venues could be found in the neighborhood:

In [16]:
```
manhattan_venues_unique=len(manhattan_venues['Venue Category'].unique())
print('There are {} uniques categories.'.format(len(manhattan_venues['Venue Category'].unique())))
```

```
There are 339 uniques categories.
```

Note that the details of the codes can be found in my ipython notebook uploaded along with this report. A k-means clustering was used to cluster the Manhattan neighborhood into 5 clusters and look at those clusters individually.

Clustering of Manhattan Neighborhood

Step 1 : Run *k*-means to cluster the neighborhood into 5 clusters.

Step 2: Create a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood.

Step 3: Visualize the new cluster

In [21]:
```
# set number of clusters
kclusters = 5

manhattan_grouped_clustering = manhattan_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(manhattan_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

Out[21]:
```
array([2, 1, 1, 2, 1, 2, 2, 0, 1, 1], dtype=int32)
```

Below is an example of how each cluster looks like with the top 10 venues for the neighborhoods in that cluster:

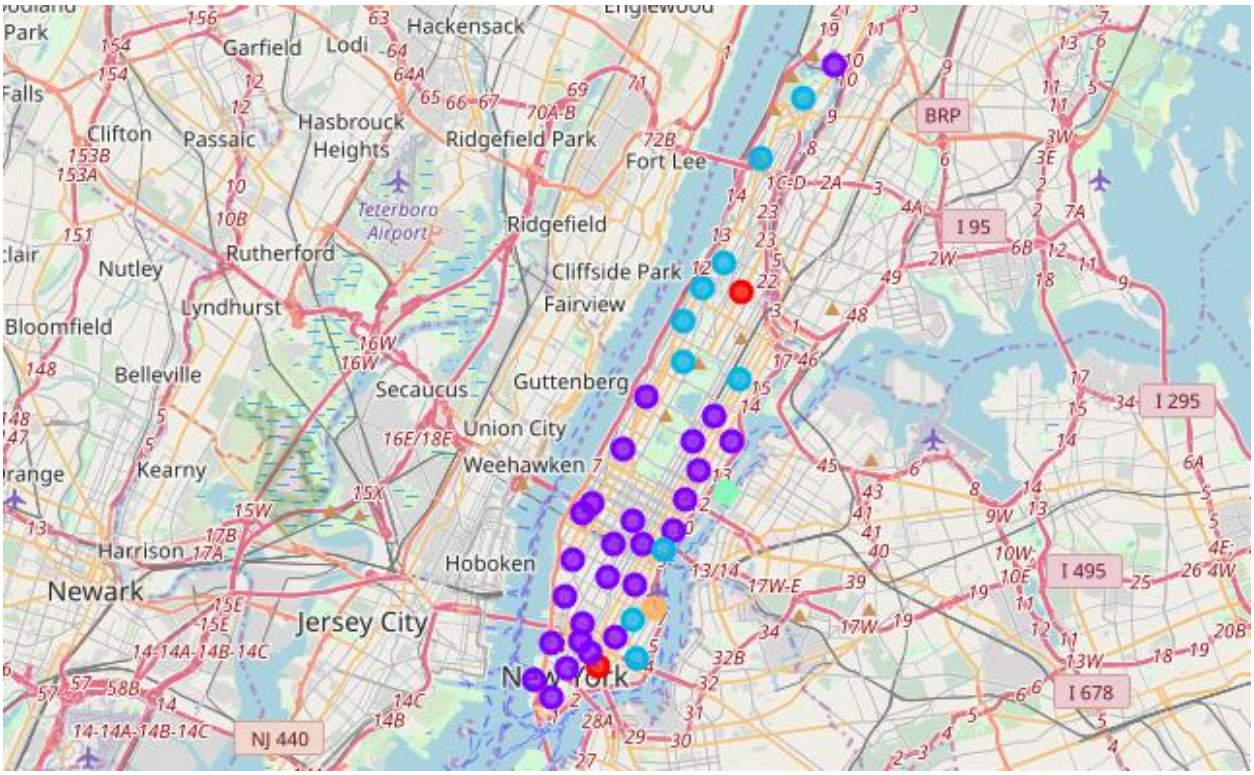
Cluster1 of Manhattan

```
In [24]: m_c1 = manhattan_merged.loc[manhattan_merged['Cluster Labels'] == 0, manhattan_merged.columns[[1] + list(range(5, manhattan_merged.shape[1]))]]
         m_c1
```

Out[24]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	Washington Heights	Café	Mobile Phone Shop	Bakery	Grocery Store	Coffee Shop	Mexican Restaurant	Liquor Store	Latin American Restaurant	New American Restaurant	Park
3	Inwood	Mexican Restaurant	Café	Lounge	Pizza Place	Park	Bakery	Caribbean Restaurant	Chinese Restaurant	Restaurant	Frozen Yogurt Shop
4	Hamilton Heights	Coffee Shop	Mexican Restaurant	Pizza Place	Café	Yoga Studio	Indian Restaurant	Sushi Restaurant	Sandwich Place	Deli / Bodega	Liquor Store
7	East Harlem	Mexican Restaurant	Deli / Bodega	Bakery	Latin American Restaurant	Thai Restaurant	Gas Station	Donut Shop	Liquor Store	Coffee Shop	Cocktail Bar

A clustering of Manhattan can be seen in the below map (using Folium) :



Section 4: Data analysis – Toronto Neighborhood

In principle, the approach for the Toronto neighborhood exploration should be the same. However, the data for this study is not as simplified as for the data for New York City neighborhood. As mentioned in the earlier section (Section 2), the data is first scraped from Wikipedia, cleaned up and pre-processed for longitude and latitude.

	Postalcode	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront, Regent Park	43.654260	-79.360636
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763
4	M7A	Queen's Park	Queen's Park	43.662301	-79.389494
5	M9A	Etobicoke	Islington Avenue	43.667856	-79.532242
6	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
7	M3B	North York	Don Mills North	43.745906	-79.352188

Then similar to the Manhattan neighborhood data analysis, using Foursquare API, we can do a neighborhood exploration for different venues:

In [44]:
```
print(toronto_venues.shape)
toronto_venues_total=toronto_venues.shape[0]
toronto_venues.head()
```

```
(2264, 7)
```

Out[44]:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	KFC	43.754387	-79.333021	Fast Food Restaurant
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

Again, we can get the number of unique venues in the Toronto neighborhood:

Unique catagories :

In [46]:
```
toronto_venues_unique=len(toronto_venues['Venue Category'].unique())
print('There are {} uniques categories.'.format(len(toronto_venues['Venue Category'].unique())))
```

```
There are 275 uniques categories.
```

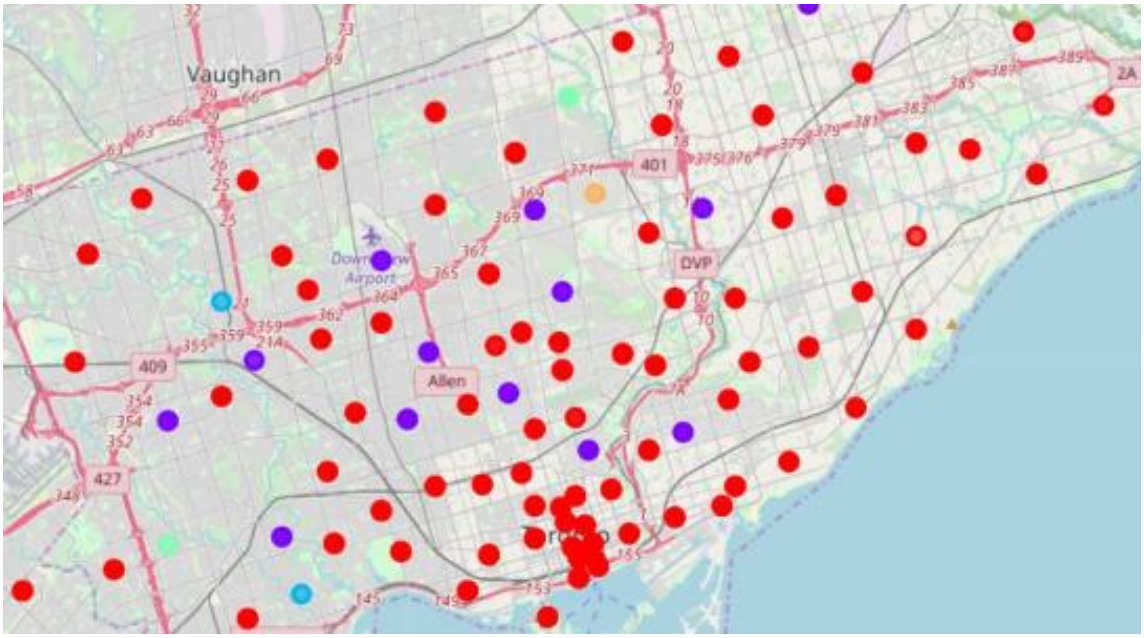
A k-means clustering was used to cluster the Toronto neighborhood into 5 clusters and look at those clusters individually similar to what we did for Manhattan neighborhood. Below is a snapshot of an example cluster with top 10 venues of different neighborhood

Cluster 1 of Toronto

```
In [59]: t_c1 = toronto_merged.loc[toronto_merged['Cluster Labels_TO'] == 0, toronto_merged.columns[[1] + list(range(5, toronto_merged.shape[1]))]]
         t_c1
```

	Neighbourhood Latitude	Venue Longitude	Venue Category	Cluster Labels_TO	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
3	43.725882	-79.315635	Hockey Arena	0	0	Portuguese Restaurant	Coffee Shop	French Restaurant	Hockey Arena	Intersection	Empanada Restaurant	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant
4	43.725882	-79.313103	Coffee Shop	0	0	Portuguese Restaurant	Coffee Shop	French Restaurant	Hockey Arena	Intersection	Empanada Restaurant	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant
5	43.725882	-79.312785	Portuguese Restaurant	0	0	Portuguese Restaurant	Coffee Shop	French Restaurant	Hockey Arena	Intersection	Empanada Restaurant	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant
6	43.725882	-79.317410	French	0	0	Portuguese	Coffee Shop	French	Hockey Arena	Intersection	Empanada	Ethiopian	Electronics	Eastern	Dumpling

A clustering of Toronto can be seen in the below map (using Folium) :



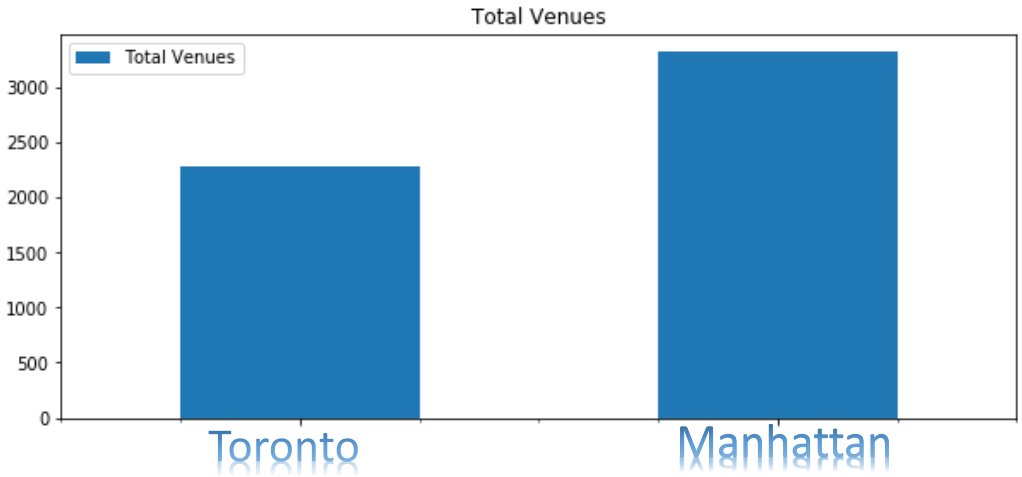
Now that, we are equipped with the right format and information of both the datasets (Manhattan and Toronto), we will do some simple comparison of various venues and try to get an idea of how these two cities are similar or different in some specific aspect. Note that although we have done the clustering and all as part of this exercise and the previous ones, we will not be exclusively use the clusters to do basic comparison. Instead, we will do an overall comparison based on different venue types.

Section 5: Comparative Analysis (Manhattan Neighborhood Vs Toronto Neighborhood):

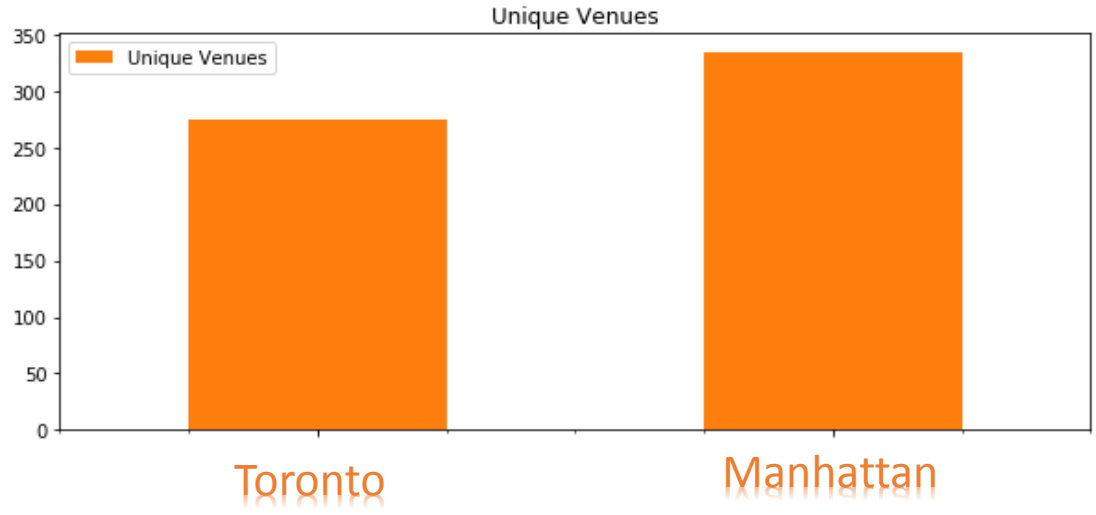
A comparison of different venues is done between Manhattan and Toronto neighborhoods. The idea of the study is to see which of the cities (neighborhoods) would have easily accessible common venues for a weekend getaway and so on to become more effective. To do so, the strategy would be to first find out the total number of venues (and unique venues) in both the neighborhoods to sort of have an idea of the livelihood and things to do accessibility of the respective neighborhoods. Then few specific venues are selected to go to one step further as far as entertainment and health resources are concerned in both the cities and a comparison is done. Following are the details of the comparative analysis:

When we compared the total venues in both the neighborhoods, we found that Manhattan area has 3324 venues while Toronto has 2264 venues. However, if we look at only unique venues, the numbers are kind of comparable. Manhattan has 339 unique venues while Toronto has 275. The following bar diagram clearly shows the comparison:

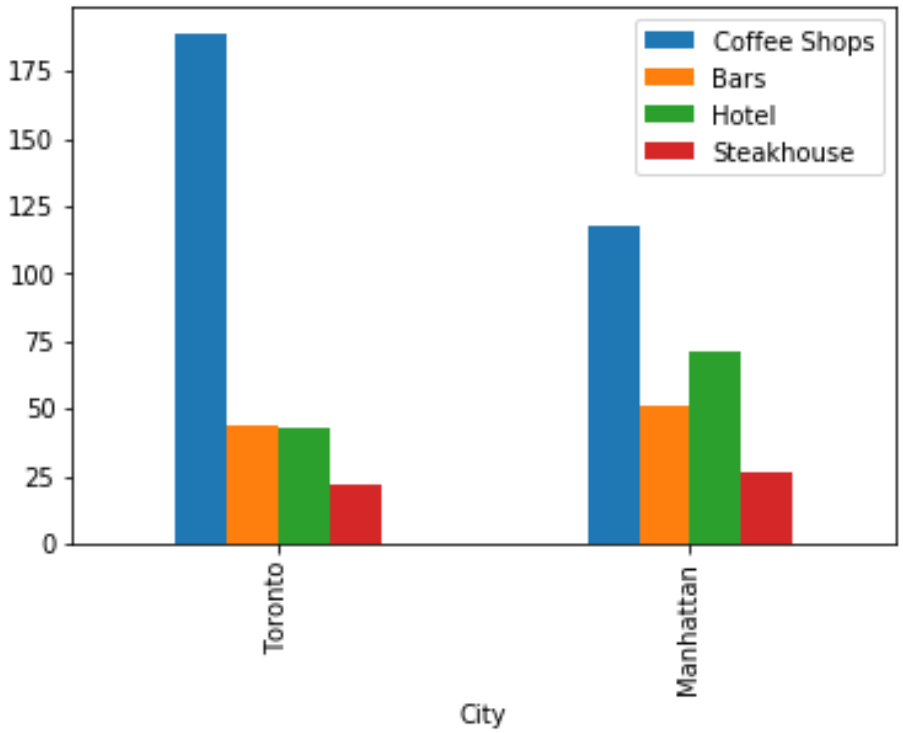
Total Venues:



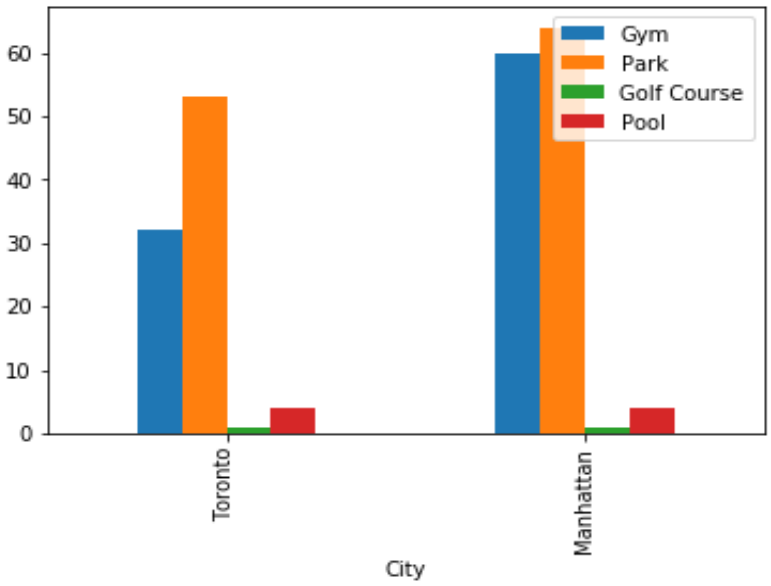
Unique Venues:



Then a little digging into both the datasets reveals a lot of information about the two neighborhoods. For instance, if we compare the usual hang out or entertainment places such as coffee shops, bars, hotels and steakhouses, we can see both neighborhoods have enough such places. However, Manhattan has higher number of such places except the coffee houses and that too with a smaller area compared to Toronto. Following plot shows this behavior exclusively:



Another instance would be to compare the some of the commonly used exercise places such as Gym, Park etc. Below plot shows the comparison very clearly between the two neighborhoods:



Section 6: Final outcome of the study

Let's refresh our objective of the study quickly before we wrap up with the conclusions. We intended to compare two neighborhoods (Manhattan and Toronto) to see which one would offer a better or a slight edge when it comes to deciding to visit and spend few hours or couple of days for a weekend getaway. It can also be used in a more specific way to decide for a better tourist spot. Equipped with the datasets of both New York City and Toronto, an analysis has been performed using the techniques we have learnt in the previous modules in the course. Below are few key points that can be derived from the comparative analysis:

- Both Manhattan and Toronto neighborhood have very diverse and dynamic environment which make both a very nice place to spend the weekend or in general a nice tourist place.

- Manhattan has 31% more total venues compared to Toronto. One notable aspect of it is Manhattan has an area of ~33 square miles while Toronto has an area of ~243 square miles. This tells us that Manhattan area has a much higher venue density, meaning, most of these venues are easily accessible if one can browse around on foot. In other words, the venues are comparatively closer in Manhattan area than the Toronto area. In addition, Manhattan has 18% more unique venues than Toronto which makes Manhattan more interesting or lucrative compared to Toronto.

- One interesting observation is Manhattan has more Bars, Steakhouses and Hotels in the smaller area compared to the bigger Toronto area. But at the same time Toronto has ~40% more coffee shops than the Manhattan area. So, this can be a very good indication of what to expect in the respective area.

- Another important outcome is Manhattan has more fitness centers and parks in a smaller area compared to Toronto. At the same time, both the neighborhoods have the same number of pools and Golf course

- Although both neighborhoods have their own characteristics, I would put Manhattan at a slightly higher position as a lively tourists' spot. However, as a weekend gateway or couple of days of relaxing days out, I would put Toronto at a slightly higher position because with all the necessary venues in it, it is a bit less densely populated and it seems to be more peaceful than Manhattan.