Applying Data Science to Compare the Neighborhoods of New York and Toronto

G. G. C. M. Gunasekara

February 15, 2020

1. Introduction

1.1 Background

New York is one of the largest metropolitan area in the world and most populous in the United States. On the other hand, Toronto is most populous city in Canada. Both cities are very diverse and are the financial capitals of their respective countries.

Larger metropolitan areas such as New York and Toronto are attracted by many people around the world due to their socio-economic values. Those metropolitan areas are excessively large in their geography and therefore usually divided into several boroughs and each borough consists of several neighborhoods. Diverse population in each neighborhood has a significant impact on its culture and various socio-economic aspects such as commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports.

1.2 Problem

It is a known fact that the venues around a given location add values to that location. It could be useful and meaningful if we could find out similar neighborhoods among each cities by analyzing venues surrounded by them. It would be great if we could simply illustrate group of similar neighborhood on a map so that one could make a valuable decision. However, due to its diversity it is very difficult to create a model or index to compare one neighborhood with another.

Modern data science tools have been evolved such that various unsupervised clustering techniques can be used to solve this kind of problems. Unsupervised clustering algorithms mainly focuses on creating clusters such that distance between points within the clusters are minimized and distance between inter-cluster points are maximized. Leveraging this concept together with data sources provided by popular location data providers we can find a scientific approach to group neighborhoods of different metropolitan areas based on surrounded venues.

1.3 Interest Groups

Anybody who is curious of socio-economic aspects of both New York and Toronto will be interest groups. Few of them would be;

1. Business personals who are interest on investing various ventures in those two cities.
2. Tourists who are interesting to find places which are similar to the places which they visited before in one of the cities.
3. Employees/workers/students who are migrating from one city to the other city
4. Data scientist who are curious about real-world application of data science

2. Data Acquisition and Cleaning

2.1 Data Sources

New York City data containing the neighborhoods, boroughs, latitudes and longitudes were obtained from the data source https://cocl.us/new_york_dataset whereas neighborhoods and boroughs of Toronto were scraped from Wikipedia website https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M and latitudes and longitudes of each borough of Toronto was obtained from http://cocl.us/Geospatial_data

2.2 Data Cleaning and Preparation

Using the data sources two DataFrames were created for New York and Toronto separately including Borough, Neighborhood and location of Neighborhood. However, based on the dataset New York had 306 Neighborhoods in total whereas Toronto had only 103 neighborhoods. Figure 3.1 illustrates distribution of Neighborhoods of both Cities. According to the figure New York is denser than Toronto.

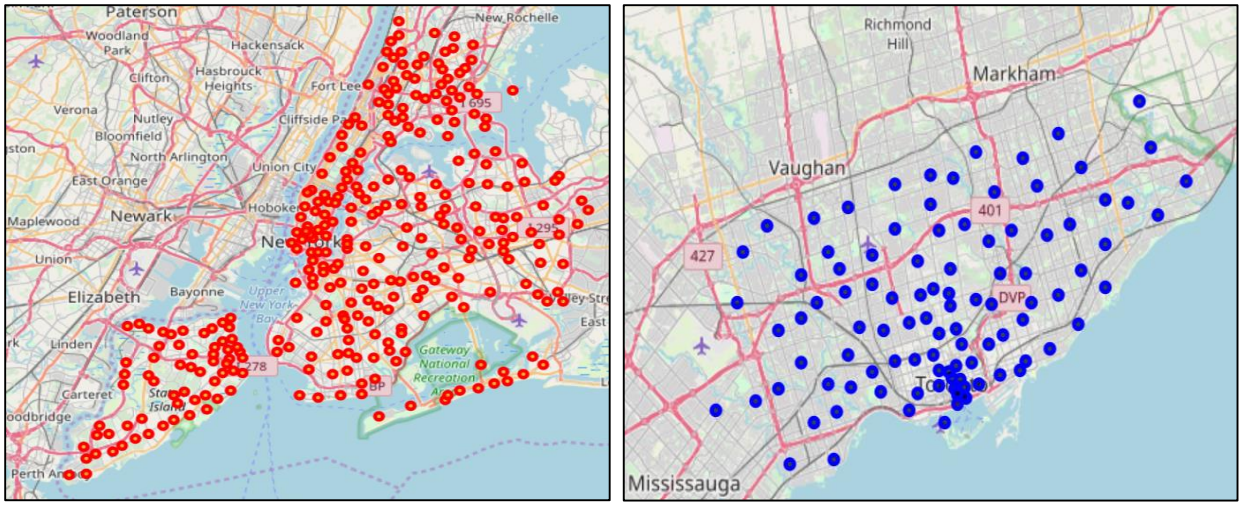

Figure 3.1 Distribution of Neighborhoods of New York and Toronto

Surrounded venues in each neighborhood was explored using the Foursquare APIs. Foursquare's *explore* endpoint was used to obtain venues and their venue categories within a radius of 500m of each neighborhood. The API request used to retrieve information was as follows

https://api.foursquare.com/v2/venues/search?&client_id=1234&client_secret=1234&v=20180605&ll=40.89470517661,-73.84720052054902&radius=500&limit=100

Client ID and client secret are the credentials used to access Foursquare data and with each query longitude and latitude should be provide to obtain the result, and radius is the radius of the circle surrounded the given location, and the query results are limits to 100 venues per each query. The Result was stored in two tables (i.e. two DataFrames) for each city in the format shown in below two figures.

	City	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	NYK	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	NYK	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	NYK	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	NYK	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop
4	NYK	Wakefield	40.894705	-73.847201	Shell	40.894187	-73.845862	Gas Station

Figure 3.2 New York Neighborhoods with venue categories

	City	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Toro	Rouge,Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Toro	Rouge,Malvern	43.806686	-79.194353	Interprovincial Group	43.805630	-79.200378	Print Shop
2	Toro	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	Chris Effects Painting	43.784343	-79.163742	Construction & Landscaping
3	Toro	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
4	Toro	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	Affordable Toronto Movers	43.787919	-79.162977	Moving Target

Figure 3.3 Toronto Neighborhoods with venue categories

3. Methodology

3.1 Exploratory Data Analysis

3.1.1 Most common venues in two cities

In order to get an idea about the distribution of venue categories among two cities, venues of each city were sorted based on the frequency of availability. Ten Most common venue categories were plotted. Figure 3.1.1 and Figure 3.1.2 shows the most common venues in New York and Toronto respectively.

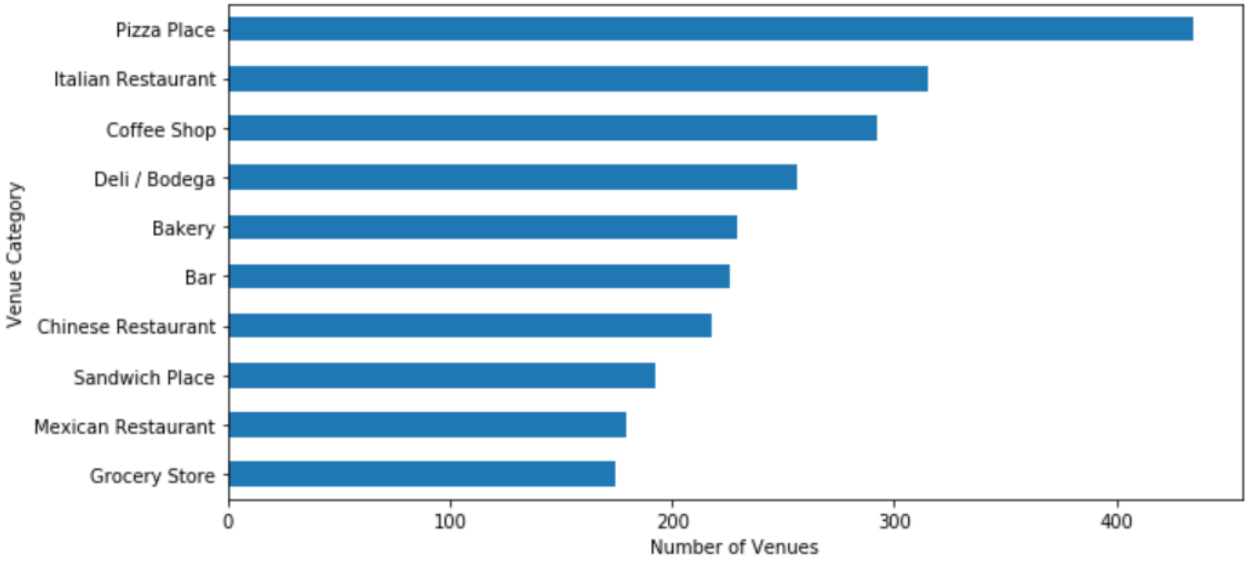

Figure 3.1.1 Top most common venues in New York

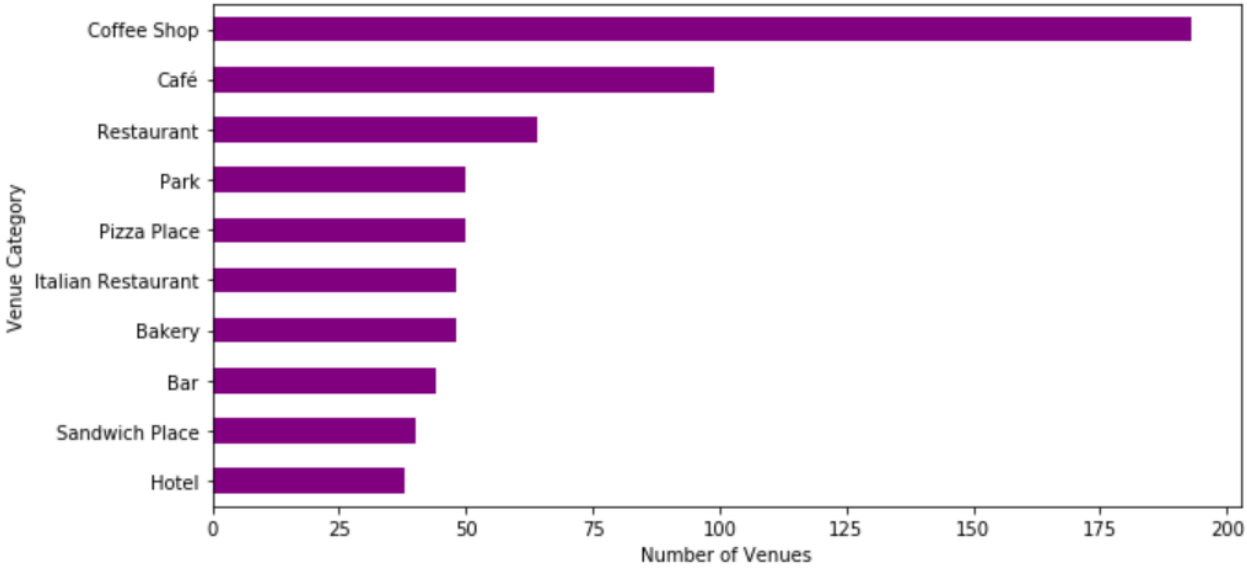

Figure 3.1.2 Top most common venues in Toronto

According to the above figures, we can see that in both cities the highest number of venues are food places.

3.1.2 Widespread venues in both cities

Identifying venues which are distributed in large number of neighborhoods are also important. Top 10 venues which are distributed amount large number of venues were plotted. The Figure 3.1.3 and Figure 3.1.4 shows the most widespread venues in New York and Toronto respectively.

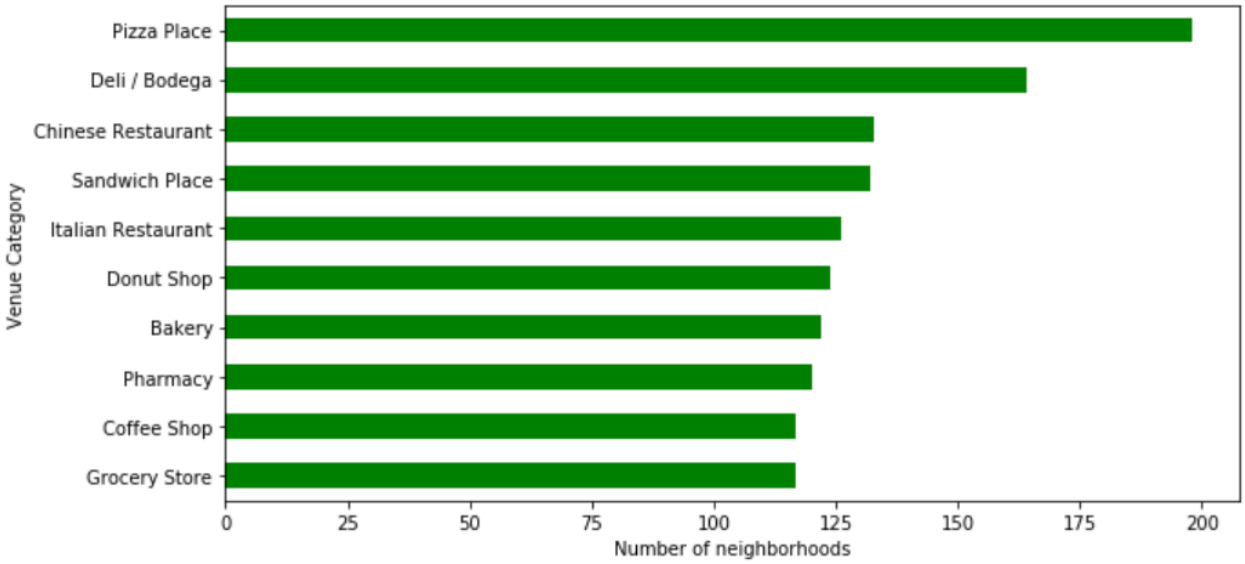

Figure 3.1.3 Top most widespread venues in New York

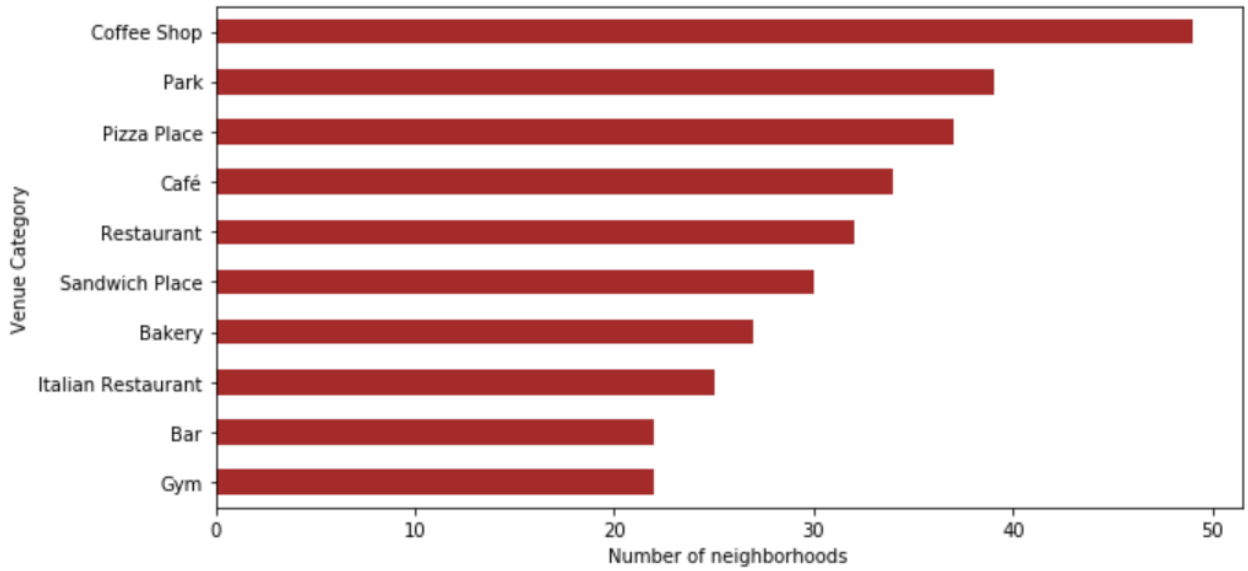

Figure 3.1.4 Top most widespread venues in Toronto

According to above two figures it is clear that most widespread venues of both cities are also food places. Above two analyses were carried out on entire cities and it gives a macroscopic view on two cities. However, our goal is to analyze similarities between neighborhoods of each city thus the analysis should

be carried out in microscopic scale. In order to group neighborhoods of both cities based on their similarities found based on the venues surrounded them, we can apply k-means clustering algorithm on neighborhoods using venue categories as features.

3.2 Clustering Neighborhoods

K-means clustering algorithm is an unsupervised clustering algorithm which partitions data points into clusters such that the distance between data points within the cluster is minimized and the distance between inter-cluster points are maximized. The algorithm partitions the given dataset in to K independent clusters, where K should be given prior to initialize the algorithm. Finding best K is a challenge associate with K-mean clustering algorithm.

3.2.1 Feature Selection and Reformat Dataframes

In order to apply k-means clustering algorithm using venue categories as features, first two DataFrames should be reformat such that venue categories become features against each neighborhood. In order to do that, one-hot encoding was applied on venue categories column and then get the mean by grouping on neighborhood. So it gives a new Dataframe averaging the number of venues in each neighborhood. A new field was introduced as City in order to distinguish New York and Toronto data.

	City	Neighbourhood	Accessories Store	Afghan Restaurant	Airport Terminal	American Restaurant	Antique Shop	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Baby Store	Bagel Shop	Bakery	Bank	Bar	Baseb Fie
0	NYK	Allerton	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.032258	0.0	0.0	0
1	NYK	Annadale	0.0	0.0	0.0	0.153846	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.076923	0.0	0.0	0
2	NYK	Arden Heights	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.0	0
3	NYK	Arlington	0.0	0.0	0.0	0.125000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.0	0
4	NYK	Arrochar	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.052632	0.0	0.0	0.105263	0.000000	0.0	0.0	0

Figure 3.2.1 Reformatted DataFrame to apply K-means clustering algorithm

3.2.2 Finding common venue categories in both cities

We can see that most of the venues can be find out in both cities. However, some venues can only be found in one city and the other city might not have that venue. But in order to apply K-means clustering, we need to make feature set of both cities identical. This could be obtain using two ways. One way is that we can include missing venues on the other dataframe with zero values. However, since such kind of venues are very few in number in a particular neighborhood we can simply remove those venues from the dataframes so that removing it will not impact on clustering.

List of venue categories common to both cities were obtain. All together there were 243 common venue categories and two DataFrames were filtered based on common categories. Two Dataframes were concatenated on common venue categories so that we can apply k-means algorithm on both cities at once.

3.2.3 Finding the 10 most common venue categories in each neighborhood

The dataset of neighborhoods was having 243 number of venue categories. However, based on the frequency of the occurrence of each venue in a particular neighborhood we could sort and find out most 10 common venues in each Neighborhood which will give us an understanding about which type of Neighborhood is this. Following python codes was used to find the top most common venues in neighborhood

```
## function to return dataframe with top most common venues
def get_top_venues(venue_grouped, num_top_venues = 10):

    indicators = ['st', 'nd', 'rd']

    # create columns according to number of top venues
    columns = ['City','Neighbourhood']
    for ind in np.arange(num_top_venues):
        try:
            columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
        except:
            columns.append('{}th Most Common Venue'.format(ind+1))

    # create a new dataframe
    neighborhoods_venues_sorted = pd.DataFrame(columns=columns)
    neighborhoods_venues_sorted['City'] = venue_grouped['City']
    neighborhoods_venues_sorted['Neighbourhood'] = venue_grouped['Neighbourhood']

    for ind in np.arange(venue_grouped.shape[0]):
        neighborhoods_venues_sorted.iloc[ind, 2:] = return_most_common_venues(venue_grouped.iloc[ind, :], num_top_venues)

    return(neighborhoods_venues_sorted)
```

```
NYK_venue_sorted=get_top_venues(NYK_grouped, num_top_venues=10)
Toro_venue_sorted=get_top_venues(Toro_grouped, num_top_venues=10)
Two_cities_venues_sorted=get_top_venues(Two_cities_grouped, num_top_venues=10)
```

Figure 3.2.3 Code used to list 10 top most common venues in neighborhoods

Figure 3.2.3 illustrates sorted ten most common venue categories in first 5 neighborhoods of the dataset.

	City	Neighbourhood	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
0	NYK	Allerton	Pizza Place	Supermarket	Deli / Bodega	Chinese Restaurant	Cosmetics Shop	Bakery	Fast Food Restaurant	Breakfast Spot	Fried Chicken Joint	Electronics Store
1	NYK	Annadale	Pizza Place	American Restaurant	Train Station	Pharmacy	Pub	Diner	Restaurant	Sports Bar	Bakery	Cosmetics Shop
2	NYK	Arden Heights	Pizza Place	Home Service	Coffee Shop	Bus Stop	Pharmacy	Filipino Restaurant	Field	Fast Food Restaurant	Farmers Market	Drugstore
3	NYK	Arlington	Grocery Store	Coffee Shop	Home Service	American Restaurant	Deli / Bodega	Bus Stop	Boat or Ferry	Intersection	Event Space	Falafel Restaurant
4	NYK	Arrochar	Bus Stop	Deli / Bodega	Bagel Shop	Italian Restaurant	Sandwich Place	Cosmetics Shop	Pharmacy	Mediterranean Restaurant	Pizza Place	Hotel

Figure 3.2.3 Ten most common venue categories in each neighborhood of New York

3.2.4 Applying K-mean clustering

K-mean clustering algorithm requires number of clusters (K value) before the initialization. In order to find the best K value, clustering was done recursively for K values from 2 to 10 and the distribution of neighborhoods of each cities among the cluster labels were obtain. Below codes were used to get the distribution of neighborhoods among each cluster label.

```
# set number of clusters
for kclusters in range(11):
    if(kclusters==0 or kclusters==1):
        continue
    clustering = Two_cities_grouped.drop(['City','Neighbourhood'], axis=1)
    # run k-means clustering
    kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(clustering)
    Two_cities_clustered = Two_cities_venues_sorted.loc[:]
    Two_cities_clustered.insert(0, 'Cluster Labels', kmeans.labels_)

    print(Two_cities_clustered.groupby('Cluster Labels')['City'].value_counts().unstack(level=1),"\n")
```

Figure 3.2.4 Code use to find out best k value

Below table illustrates the distribution of neighborhoods among cluster labels for different k values.

Table 3.2.4 Distribution of neighborhoods for different K values

K=2		
City	New York	Toronto
Cluster Labels		
0	7	15
1	294	83

K=3		
City	New York	Toronto
Cluster Labels		
0	36	4
1	5	12
2	260	82

K=4		
City	New York	Toronto
Cluster Labels		
0	40	1
1	7	14
2	2	2
3	252	81

K=5		
City	New York	Toronto
Cluster Labels		
0	159	25
1	102	56
2	5	13
3	---	3
4	35	1

K=6		
City	New York	Toronto
Cluster Labels		
0	91	51
1	32	11
2	---	2
3	3	---
4	7	14
5	168	20

K=7		
City	New York	Toronto
Cluster Labels		
0	2	1
1	5	12
2	125	44
3	6	18
4	32	1
5	---	2
6	131	20

K=8		
City	New York	Toronto
Cluster Labels		
0	75	15
1	13	1
2	103	61
3	16	2
4	---	2
5	1	---
6	89	2
7	4	15

K=9		
City	New York	Toronto
Cluster Labels		
0	28	11
1	90	55
2	150	13
3	---	1
4	2	---
5	3	14
6	---	2
7	26	2
8	2	---

K=10		
City	New York	Toronto
Cluster Labels		
0	217	78
1	2	1
2	2	---
3	1	---
4	---	2
5	2	---
6	1	3
7	5	12
8	---	1
9	71	1

According to the table, when the number of clusters increases (i.e K value increases) the number of neighborhoods fall in to particular cluster drops and some clusters are created with very few data points from a single city neighborhood. When we compare the distribution of data points from two cities in the formed clusters k=5 gives good enough partitioning of the neighborhoods.

3.2.5 Clustering with k=5

Since k=5 gives the good enough partitioning of the neighborhoods, Clusters were obtained by using k=5 and result was merged with the original dataset such that Borough, Neighborhood and location data also available in line with cluster labels.

	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	City	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
0	Scarborough	Rouge,Malvern	43.806686	-79.194353	0	Toro	Fast Food Restaurant	Yoga Studio	Cosmetics Shop	Flea Market	Fish Market	Fish & Chips Shop	Filipino Restaurant	Field
1	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	1	Toro	Bar	Construction & Landscaping	Dumpling Restaurant	Flea Market	Fish Market	Fish & Chips Shop	Filipino Restaurant	Field
2	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711	0	Toro	Spa	Intersection	Rental Car Location	Electronics Store	Mexican Restaurant	Medical Center	Breakfast Spot	Pizza Place
3	Scarborough	Woburn	43.770992	-79.216917	1	Toro	Coffee Shop	Korean Restaurant	Yoga Studio	Eastern European Restaurant	Flea Market	Fish Market	Fish & Chips Shop	Filipino Restaurant
4	Scarborough	Cedarbrae	43.773136	-79.239476	0	Toro	Athletics & Sports	Fried Chicken Joint	Gas Station	Thai Restaurant	Bank	Bakery	Caribbean Restaurant	Fast Food Restaurant

Figure 3.2.5 Neighborhoods with their cluster labels

4. Result

4.1 Distribution of neighborhoods among each cluster

Distribution of neighborhoods from each city among five clusters were as shown in the table below. Out of five clusters 0, 1, 2 are having neighborhoods from both cities thus those clusters contain similar neighborhoods from both cities. However, cluster 3 is having no neighborhoods from New York and cluster 4 is having only one neighborhood from Toronto. Therefore, those two clusters contain neighborhoods which are unique to each city.

Table 4.1 Distribution of neighborhoods among clusters

	City	
Cluster Label	**New York**	**Toronto**
0	159	25
1	102	56
2	5	13
3	---	3
4	35	1

4.2 Distribution of clustered neighborhoods in both cities

Below two figures illustrates the distribution of clustered neighborhoods in both cities. There are five different clusters and they are marked with five different colors on the map. Neighborhoods with same colors fall in to same cluster, thus we can conclude that they are similar to each other in one way or the other.

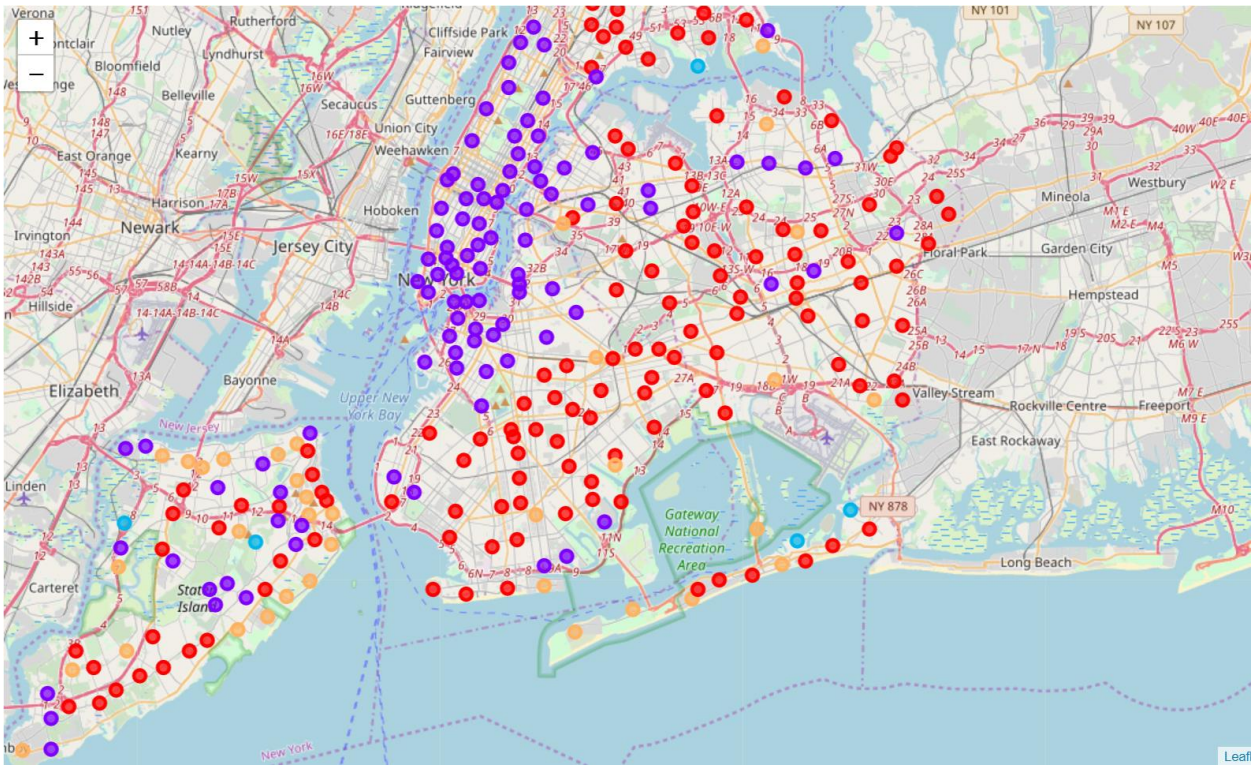

Figure 4.1 Clustered neighborhoods in New York

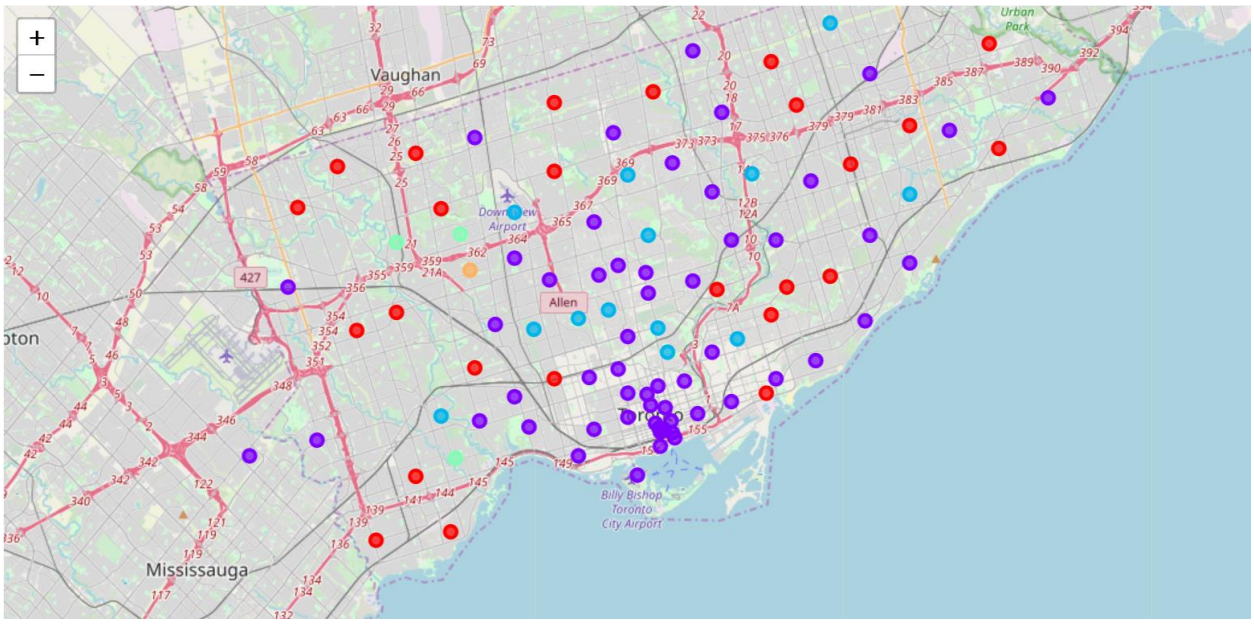

Figure 4.1 Clustered neighborhoods in Toronto

5. Discussion

5.1 Cluster Analysis

Our analysis resulted in 5 different clusters. For the sake of understanding each cluster generated by algorithm, we will find most common venues in each cluster. Below code snippet was used to get the top most venues in each cluster. The result is shown in Figure 5.2

```
D={}
for clust_no in Two_cities_clustered['Cluster Labels'].value_counts().index:
    df=Two_cities_clustered[Two_cities_clustered['Cluster Labels']==clust_no].iloc[:,3:] #extract requied portion from the dat
    Items=[df.loc[i,j] for i in df.index for j in df.columns] #
    Freq={}
    for item in Items:
        Freq[item]=Items.count(item)
    #sort the venur, freq pairs based on freq and take first most 10
    Top_venue=sorted(Freq.items(), key = lambda kv:(kv[1], kv[0]), reverse=True)[:10]
    total=len(Items)
    b=pd.DataFrame(Top_venue)[0].to_list()
    c=pd.DataFrame(Top_venue)[1].to_list()
    avg=[x/total*100 for x in c]
    D["Cluster" + str(clust_no)]=b
    D["Cluster" + str(clust_no) + " venues %"]=avg

pd.DataFrame(D)
```

Figure 5.1 The code used to get top most venues in each cluster

	Cluster0	Cluster0 venues %	Cluster1	Cluster1 venues %	Cluster4	Cluster4 venues %	Cluster2	Cluster2 venues %	Cluster3	Cluster3 venues %
0	Pizza Place	5.550239	Coffee Shop	6.217949	Bus Stop	8.235294	Park	10.000	Fish & Chips Shop	10.000000
1	Donut Shop	3.971292	Italian Restaurant	4.679487	Field	7.058824	Field	9.375	Filipino Restaurant	10.000000
2	Fast Food Restaurant	3.875598	Café	3.846154	Fast Food Restaurant	7.058824	Fast Food Restaurant	9.375	Field	10.000000
3	Deli / Bodega	3.444976	Pizza Place	3.653846	Farmers Market	6.470588	Farmers Market	9.375	Fast Food Restaurant	10.000000
4	Pharmacy	3.253589	Bar	3.525641	Filipino Restaurant	5.294118	Falafel Restaurant	9.375	Farmers Market	10.000000
5	Sandwich Place	3.157895	Bakery	2.564103	Falafel Restaurant	5.294118	Filipino Restaurant	6.875	Construction & Landscaping	10.000000
6	Chinese Restaurant	3.157895	Yoga Studio	2.243590	Yoga Studio	3.529412	Event Space	6.875	Falafel Restaurant	6.666667
7	Bank	2.870813	Park	2.243590	Playground	3.529412	Yoga Studio	6.250	Event Space	6.666667
8	Bakery	2.200957	American Restaurant	2.179487	Fish & Chips Shop	3.529412	Donut Shop	4.375	Drugstore	6.666667
9	Farmers Market	2.153110	Fast Food Restaurant	1.987179	Event Space	3.529412	Dog Run	4.375	Park	3.333333

Figure 5.2 Most common venues in each cluster with their percentage of availability

The difference between each cluster can be seen from the above figure. Though same venue categories are available in multiple clusters their contribution or to each cluster is significantly different. Some of the observations that can be obtained from the above figure are.

- While there is a contribution of 6% from pizza places in cluster 0, there is only 3% of contribution to cluster 1 from pizza places. On the other hand, Coffee shops gives 6% contribution to cluster 1 while giving less than 2% significance to cluster 0
- While Chinese Restaurants having 3% of significance in cluster 0, similar contribution is there from Italian Restaurants to Cluster 1
- Pharmacies and Banks could be seen only in Cluster 0 with a significant of 3% which indicates those are more business oriented or urban places which are suitable for Business peoples and Employees
- Cluster 2 which is having 9% contribution from Parks, 7% of Fast Food restaurants and 6% of Yoga studios are more suitable for leisure and relaxation.

6. Conclusion

This report describes how we used modern data science tools to cluster neighborhoods of New York and Toronto which are largest metropolitans in the world with vast diversification. The k-mean clustering algorithm was used to cluster neighborhoods of these largest metropolitans based on the venues/venues categories surrounded by their neighborhoods. The venues in each neighborhood were obtained from Foursquare API using the explore endpoint.

Though it is hard to give a name to the resulted clusters it was clearly shown the difference between each cluster based on the percentage of significance of each venue in each cluster. Folium library was used to illustrate clustered neighborhoods on a maps.

6.1 Future Direction

In this project only venue categories were used as features in order to cluster the neighborhoods of both cities. However, in reality two venues with same category will not be the same due to various reasons. For example, a pizza place in Toronto might not be the same as a pizza place in New York due to customer preferences, cultural impacts and so many other reasons. But the algorithm treats both with same weight. Therefore, if we could include other relevant information associates with venues the result could be improved better.

Even though we were succeeded in clustering neighborhoods with unsupervised clustering algorithm there is no way to cross check the performance of resulted clusters which is a drawback to improve the method.