The Battle of the Neighborhoods

- Introduction

» Background

This project aims to compare two cities: Toronto and New York. New York is located in the United States of America while Toronto is located in Canada. New York is the most populous city of United States, hence it carters to large amount of people. In 2017, the city had an estimated population density of 28,491 inhabitants per square mile, rendering it the most densely populated of all municipalities housing over 100,000 residents in the United States Toronto is the most populous city of Canada which carters to variety of people. Both cities have boroughs which have many venues for numerous purposes. Toronto encompasses a geographical area formerly administered by many separate municipalities. Toronto is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world

» Problem Statement

As it is discussed above that both cities are densely populated therefore they entertains a large amount of crowd every day. This project will compare in terms of number of venues present in city's borough to check which city has more variety of borough as well as in large numbers. This will give an overview that which city will provide services to its population better.

- Data Section

» Toronto Dataset

The dataset for Toronto is scrapped from Wikipedia page and it is stored in the dataframe. The dataset contains three columns namely postcode, borough and neighborhood. It has 288 rows with raw data which is refined by removing unassigned borough which leaves only 102 rows for further processing.

Postcode	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Harbourfront
M6A	North York	Lawrence Heights

» New York Dataset

New York dataset is taken from New York University Website. This dataset has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood.

{"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]}},{"type":"Feature","id":"nyu_2451_34572.2","geometry":{"type":"Point","coordinates":
[-73.82993910812398,40.87429419303012]},"geometry_name":"geom","properties":{"name":"Co-op City","stacked":2,"annoline1":"Co-op","annoline2":"City","annoline3":null,"annoangle":0E-11,"borough":"Bronx","bbox":
[-73.82993910812398,40.87429419303012,-73.82993910812398,40.87429419303012]}},{"type":"Feature","id":"nyu_2451_34572.3","geometry":{"type":"Point","coordinates":
[-73.82780644716412,40.887555677350775]},"geometry_name":"geom","properties":{"name":"Eastchester","stacked":1,"annoline1":"Eastchester","annoline2":null,"annoline3":null,"annoangle":0E-
11,"borough":"Bronx","bbox":[-73.82780644716412,40.887555677350775,-73.82780644716412,40.887555677350775]}},{"type":"Feature","id":"nyu_2451_34572.4","geometry":{"type":"Point","coordinates":
[-73.90564259591682,40.89543742690383]},"geometry_name":"geom","properties":{"name":"Fieldston","stacked":1,"annoline1":"Fieldston","annoline2":null,"annoline3":null,"annoangle":0E-11,"borough":"Bronx","bbox"
[-73.90564259591682,40.89543742690383,-73.90564259591682,40.89543742690383]}},{"type":"Feature","id":"nyu_2451_34572.5","geometry":{"type":"Point","coordinates":

» Data Cleaning

Both dataset were cleaned for working on them efficiently. First the entire boroughs with not assigned values were dropped from the dataset. Then the neighborhood having unassigned values were replaced with their respective borough values. Neighborhood having same borough had their values merged into single row, so that it was simpler to work by taking borough as the center feature. Longitude and Latitude values for each borough are searched and joined with the current dataset. Latitude and Longitude value helped in plotting the borough on the map and getting desired results. Postcode column was dropped from the dataset as it was unnecessary for the further use.

Foursquare API was used to gather data about venues in the borough. Data for venues which within 500 meter radius were collected by the API and it was joined with the original dataset. Venue data consisted of name, longitude, latitude and category of the venue. Data was grouped by the borough to get total number of venue in the borough.

Dataset after cleaning looks like:

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bronx	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Bronx	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Bronx	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Bronx	Wakefield	40.894705	-73.847201	Shell	40.894187	-73.845862	Gas Station

- Methodology Section

To gather information how venues are distributed and which are the top venues of the area, one hot encoding was done on the dataset. Mean value of the distributed venues were taken to get their frequency in the borough and understand how they were scattered across the venue. This was done by grouping encoded values on basis of the borough. Frequency were mapped to their respective venue category and sorted in the ascending order of their availability in the borough.

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.

This algorithm was used as it is fast and robust. It also provides well defined results which are distinctly separated from each other. K-Means was used in this project to define cluster for the dataset. The cluster were formed by the taking the frequency of the venues in the borough and grouping them according to their similarities. The clusters were mapped on the map to get the desired output and visualization of the result.

In [48]:
```
kclusters = 3

ny_clustering = ny_grouped.drop('Borough', 1)

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

kmeans.labels_[0:4]
```

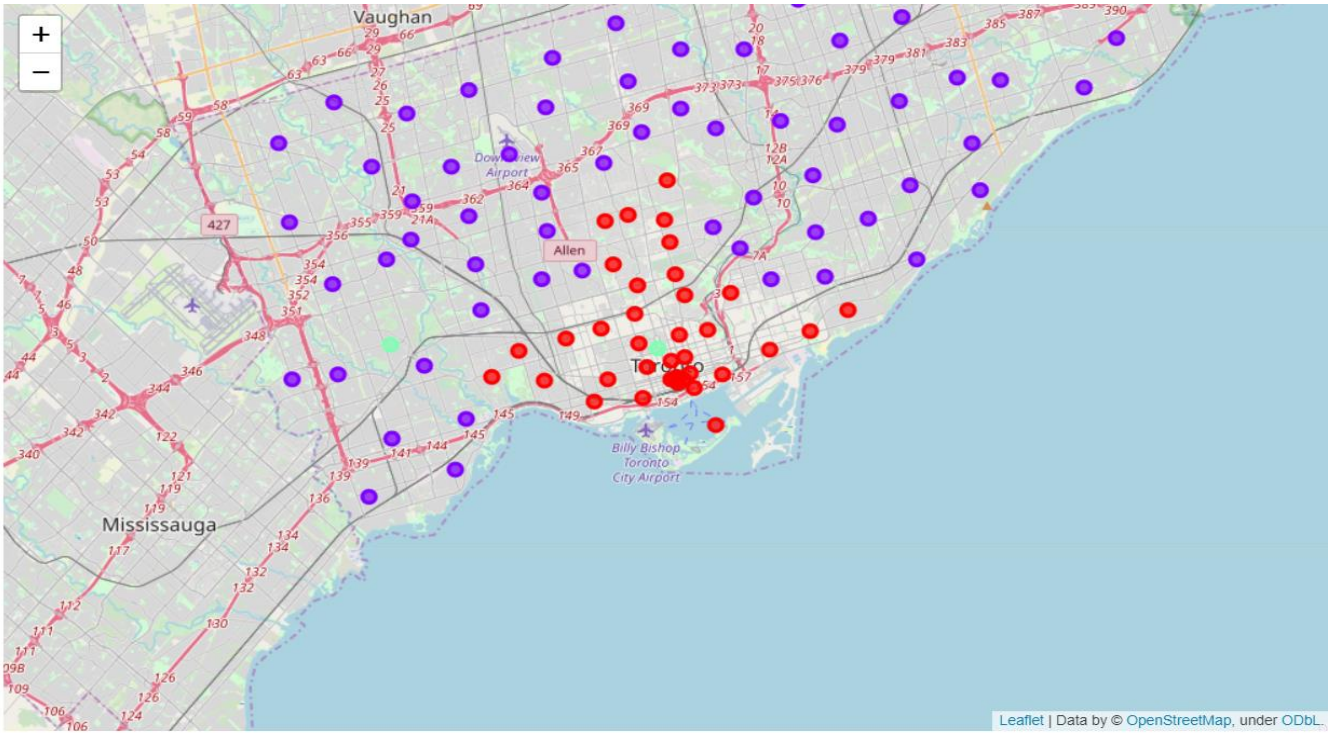
- Results

To get the desired results, clusters that were obtained by the k-means algorithm were projected on the map to get visual representation. The names of the venues located in the cluster were also seen in the cluster algorithm. Results were helpful in obtaining the conclusion.

Toronto Cluster 0: It was concentrated which means that venues are located closely and they are in lesser quantity compared to other cluster. It has Coffee shops as the most common venue among them and other venues are variable.

Toronto Cluster 1: It was scattered across the map but it consisted more venues than cluster 0. It has park as the most common venue for the cluster.

Map of Toronto

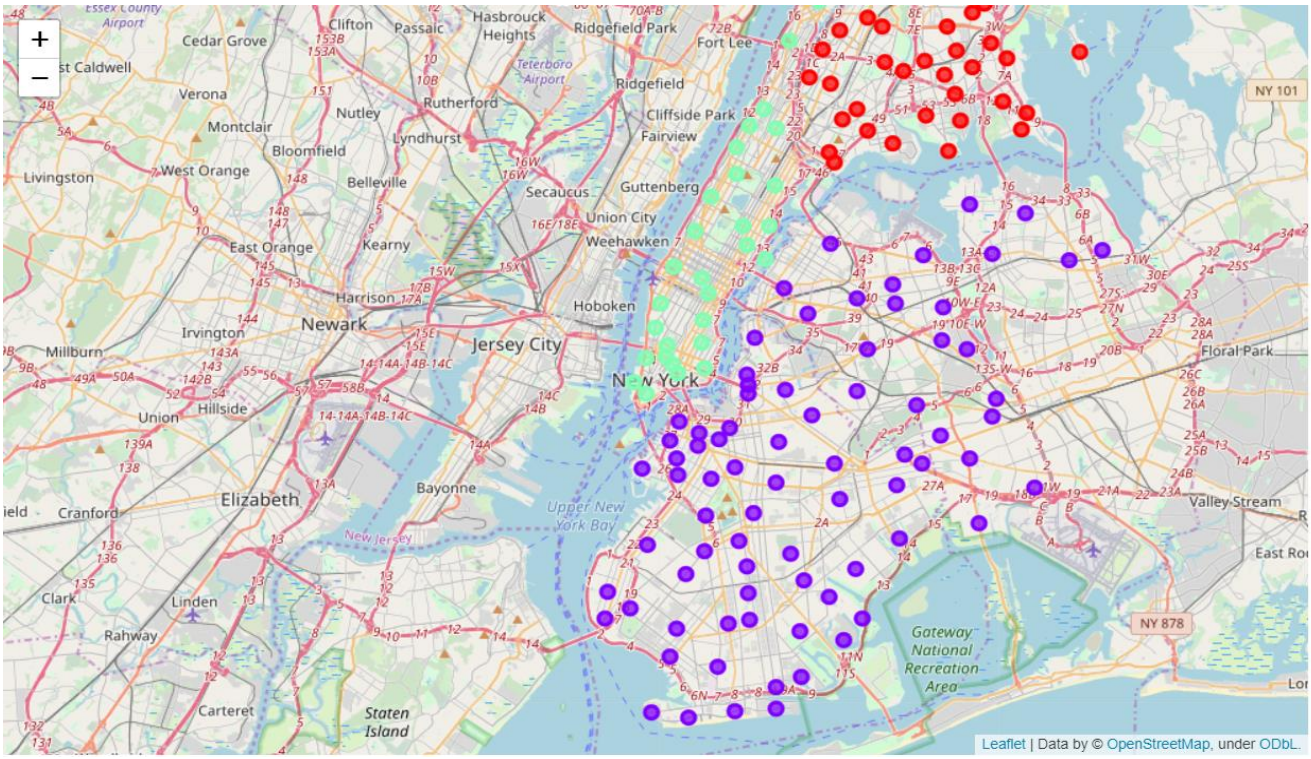


New York Cluster 0: It is smallest cluster among the three clusters and it has Pizza Place and Bodega as the two most common venues among borough.

New York Cluster 1: It is evenly distributed across map and it has more boroughs than cluster 0. It has coffee shops as the most common venue.

New York Cluster 2: It was the biggest cluster among the three clusters. It also has Pizza place as the most common venue but Bakery is the second most common venue unlike cluster 0.

Map of New York



- Conclusion

From above results, we can conclude that New York has more variety of venues in greater quantity than Toronto. While Pizza place has been common venue in the New York which is present in large quantity and closely located to each other. This suggests that Pizza is the one of the favorite food for the people in New York.
In Toronto, Coffee place has been the most common place which is suggested by the cluster and it is seconded by the park in the neighborhoods.

- Future Works

This can be better researched and explained by using other features for the dataset like using venue ratings which will provide better overview. By using venue rating, it is provide clear understanding about the venues and also clustering can be better on its basis. Other features can also be used for comparison and feature selection can be used for the best outcome of the project.