COURSERA CAPSTONE PROJECT

BY SAHAR HEKMATDOUST

DECEMBER 2020

Table of Contents

INTRODUCTION

Problem Description:

Toronto, London and New York are famous tourist destinations in the world. They are diverse in many ways. All are multicultural as well as the financial hubs of their respective countries. We want to explore how much they are similar or dissimilar in aspects from a tourist point of view regarding food, accommodation, beautiful places, and many more.

Audience:

The target audience would be **tourists** and **travel agencies**. Tourists can explore neighborhoods in each city and decide which city they prefer to visit or if they have been to one of these cities before and enjoyed their visit, they can select a similar city to travel next time. Travel agencies also can recommend destinations to their customers based on customers' experience and similarity and dissimilarity between different cities.

DATA

Data sources and processing:

Boroughs and neighborhoods: For each city , we need neighborhood name and its location data

London: London has in total 32 boroughs. To explore, analyze and segment neighborhoods, longitude and latitude of each neighborhood and borough will be added. This dataset exists for free on the web: https://skgrange.github.io/www/data/london_sport.json

New York: New York has a total of 5 boroughs and 306 neighborhoods. This dataset exists for free on the web. Here is the link to the dataset: https://geo.nyu.edu/catalog/nyu_2451_34572

Toronto: For Toronto I used the table in Wikipedia for postal code and borough of each neighborhood. (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) and for the longitude and latitude of each neighborhood, I used a csv file available in: http://cocl.us/Geospatial_data

Data sources and processing:

Foursquare API

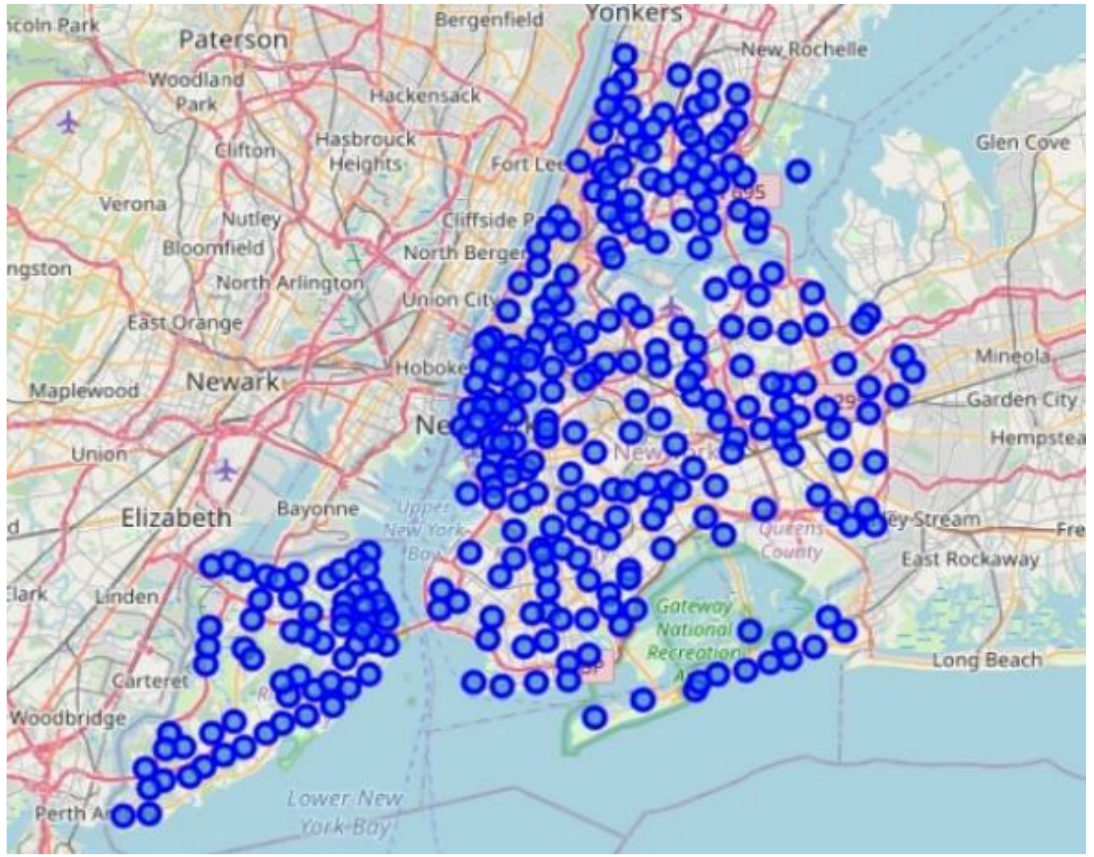
in order to explore neighborhoods and cluster them we need to search for venues in each neighborhood. **Foursquare API(utilized via the Request library in Python)** permits to provide venues information for each neighborhood in London, Toronto and New York.

METHODOLOGY

Data preparation:

New York

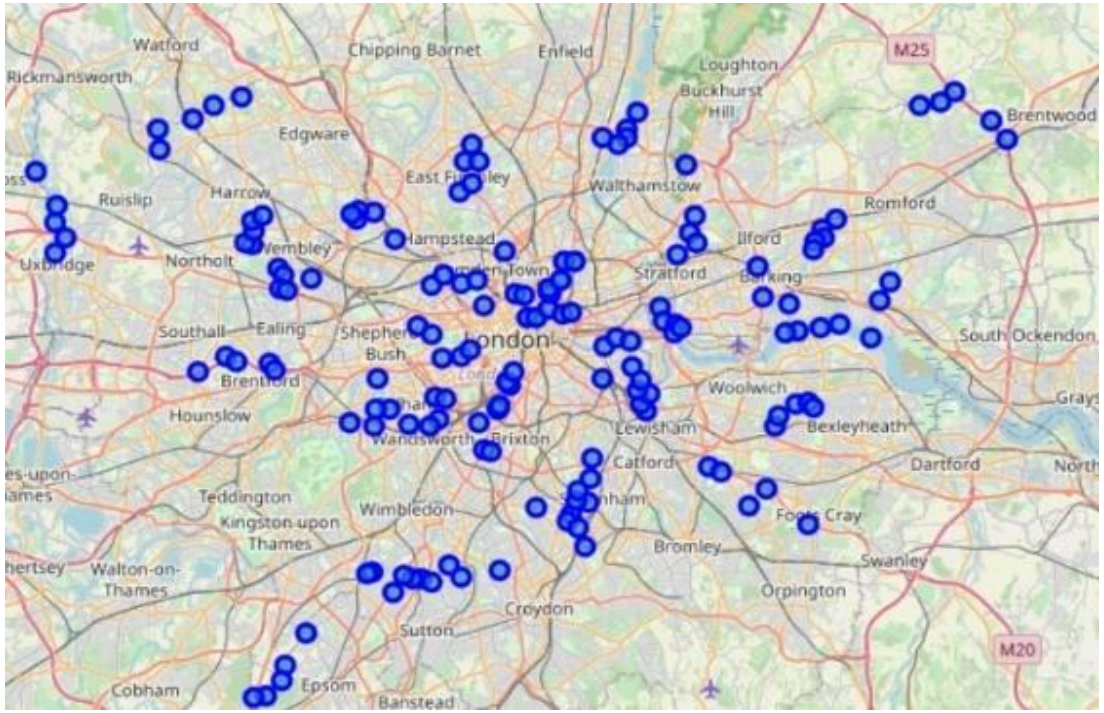
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585



Data preparation:

London

	Neighborhood	Latitude	Longitude
0	Bromley0	51.442884	0.031639
1	Bromley1	51.440465	0.041526
2	Bromley2	51.423211	0.063333
3	Bromley3	51.431508	0.076946
4	Bromley4	51.413598	0.109226



Data preparation:

Toronto

	Neighborhood	Latitude	Longitude
0	Parkwoods	43.753259	-79.329656
1	Victoria Village	43.725882	-79.315572
2	Regent Park, Harbourfront	43.654260	-79.360636
3	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Exploratory data analysis

Using foursquare location data: Top 100 venues in each neighborhood that are within a radius of 500 meters are explored in London, Toronto and New York. Then the resulting data frames are combined into one data frame .

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	City
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop	Newyork
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy	Newyork
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy	Newyork
3	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop	Newyork
4	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop	Newyork

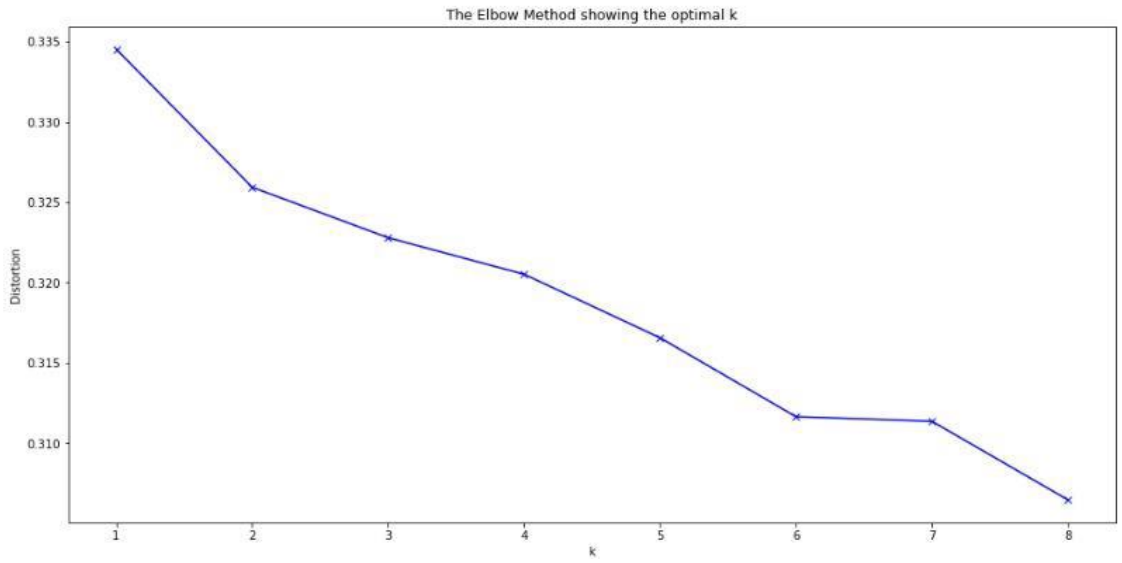
Machine Learning

One-hot encoding is used in order to transform Categorical Data into Numerical Data for Machine Learning algorithms. The resulting data frame:

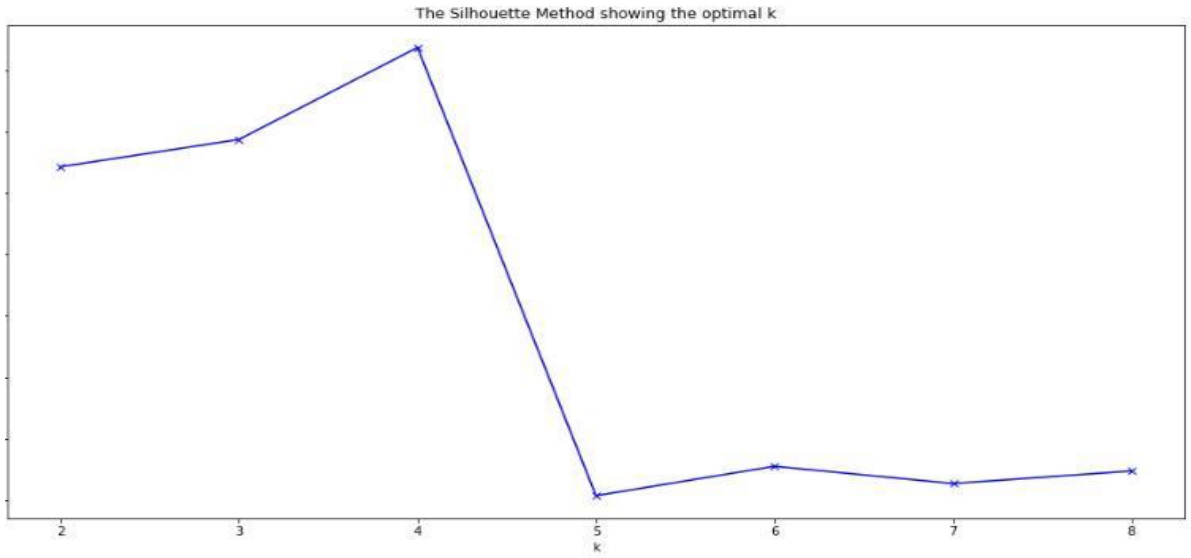
	Neighborhood	City	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	...	V
0	Wakefield	Newyork	0	0	0	0	0	0	0	0	...	
1	Wakefield	Newyork	0	0	0	0	0	0	0	0	...	
2	Wakefield	Newyork	0	0	0	0	0	0	0	0	...	
3	Wakefield	Newyork	0	0	0	0	0	0	0	0	...	
4	Wakefield	Newyork	0	0	0	0	0	0	0	0	...	

Machine Learning:

K-means Clustering: Finding the best K using two methods:



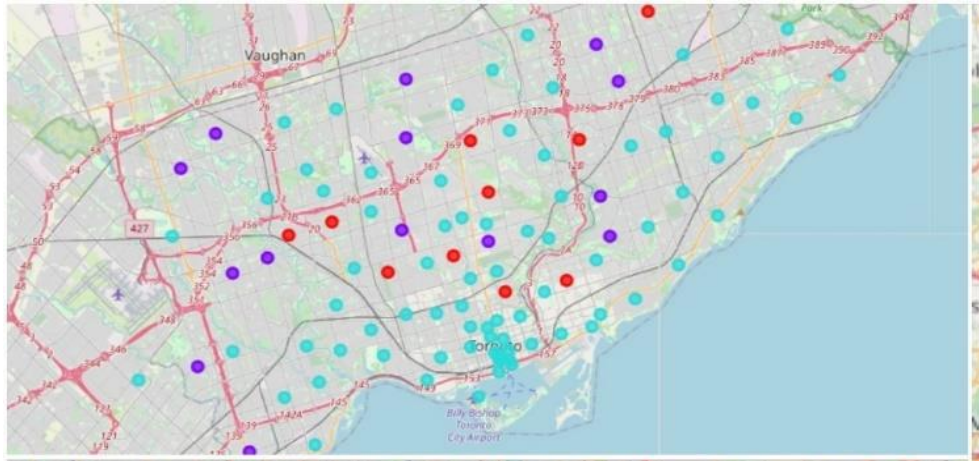
Elbow method:

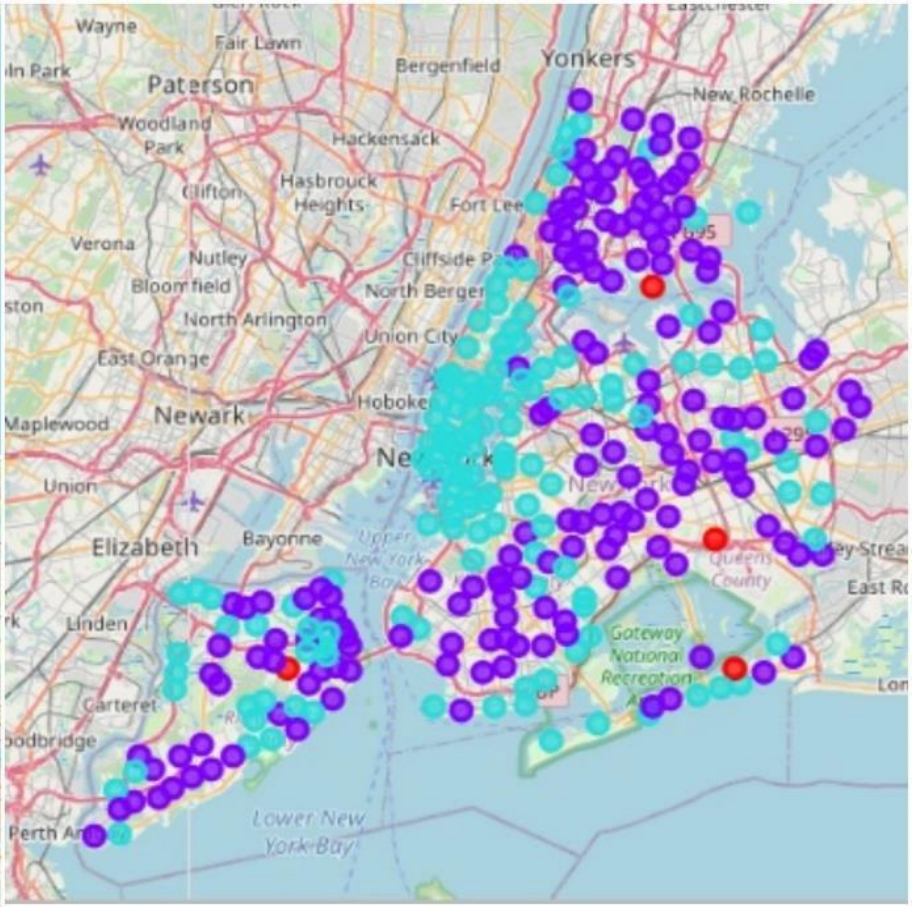


Silhouette method

Machine learning

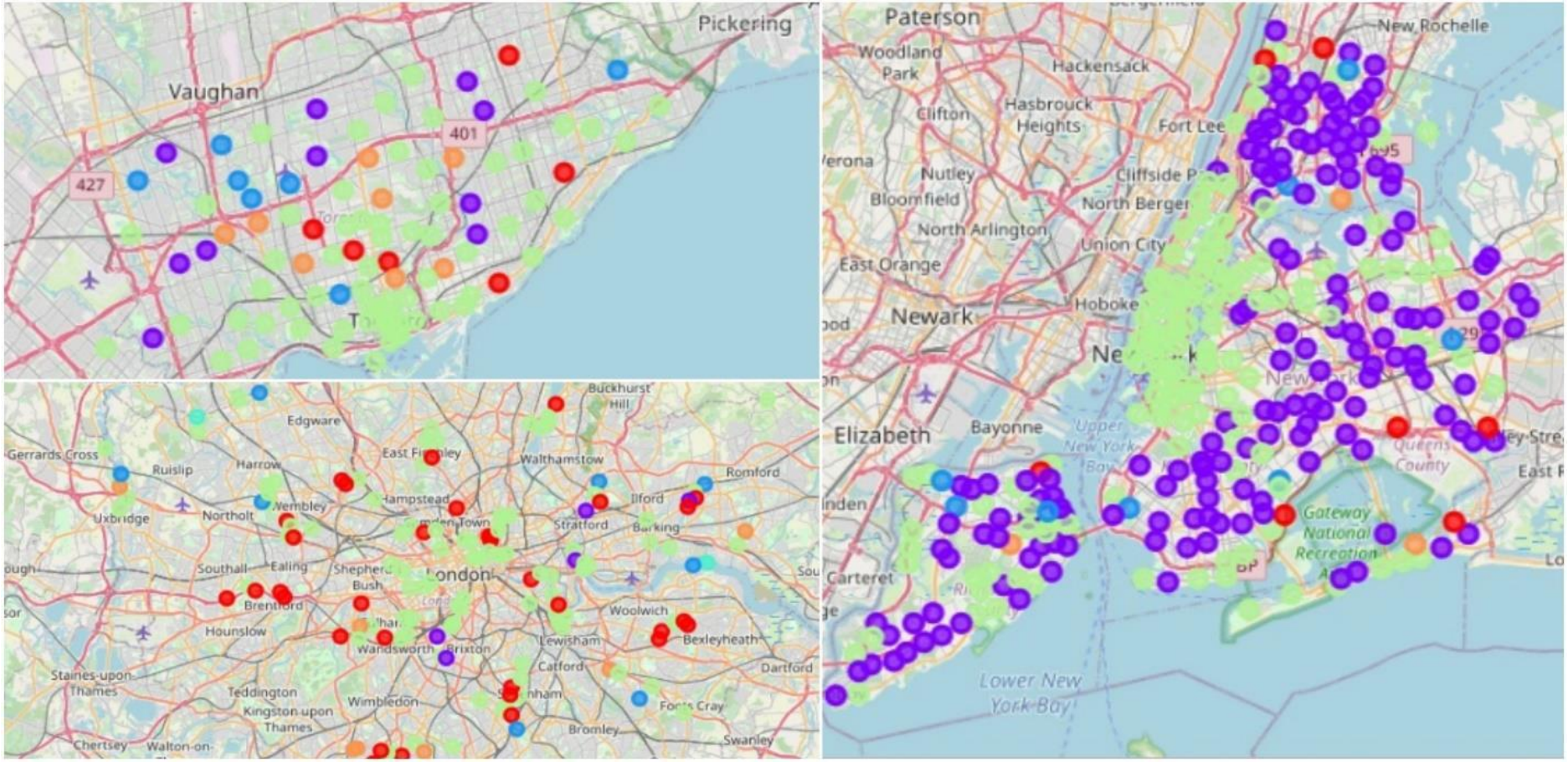
K=4





Machine learning

K=6



RESULTS

K=4

Cluster 0(red): parks & Playground, Construction & Landscaping, Exhibits and farms

Cluster 1(purple): Bus station, pharmacy, Bakery and grocery

Cluster 2(light blue): bars, pubs, restaurants, Hotels, banks and plazas

Cluster 3(creamy): Golf Course and zoos

K=6

Cluster 0(red): parks & Playground, Farm and Factory

Cluster 1(purple): Bus station, pharmacy, Gym

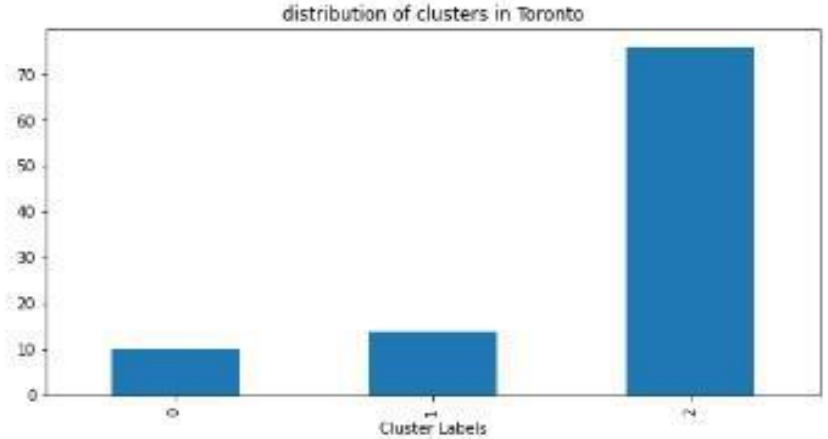
Cluster 2(light blue): Grocery Store, Playground

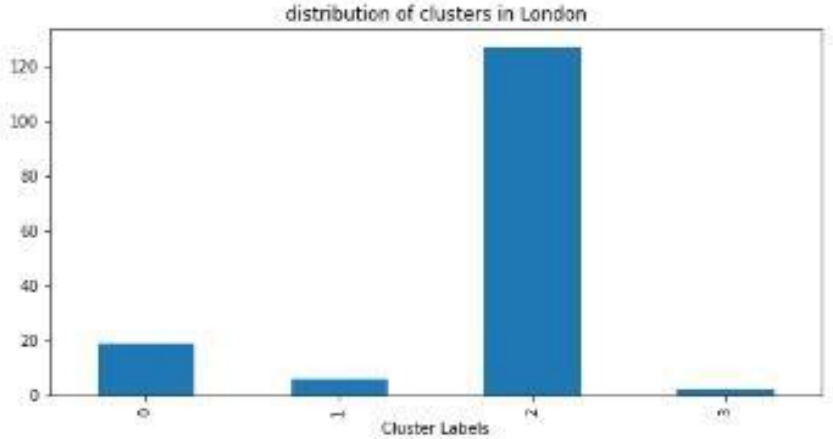
Cluster 3(creamy): Golf Course and zoo

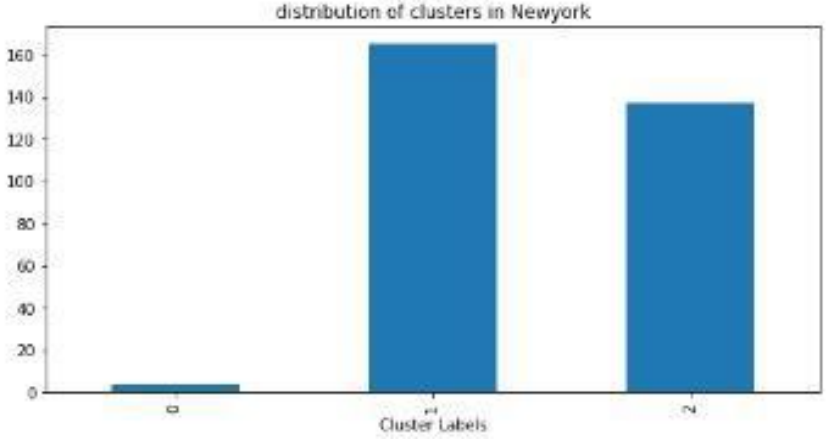
Cluster 4(light green): Bar, Pub, Restaurant, Cafe, Pharmacy

Cluster 5(orange): Park, Exhibit and Event Space

Distribution of clusters:

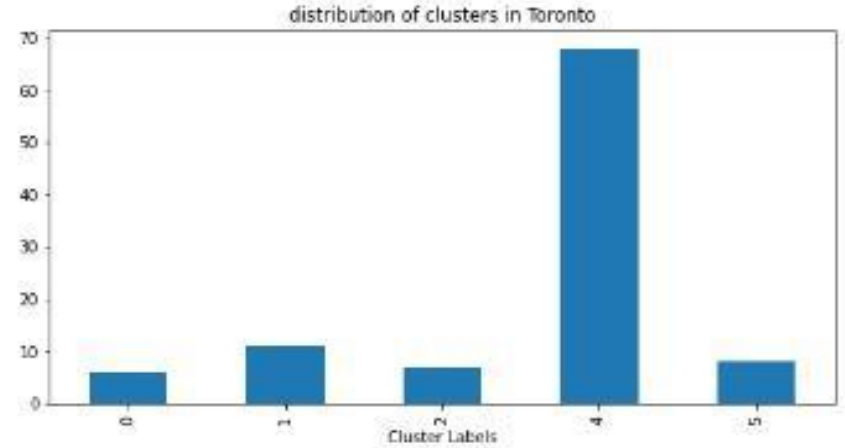


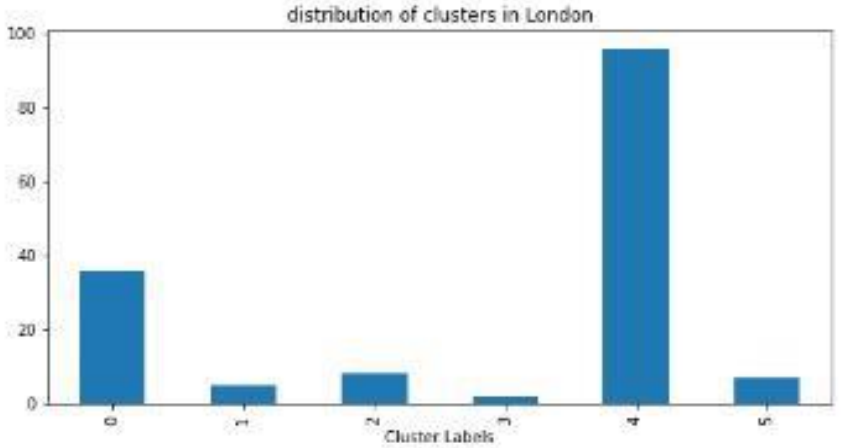


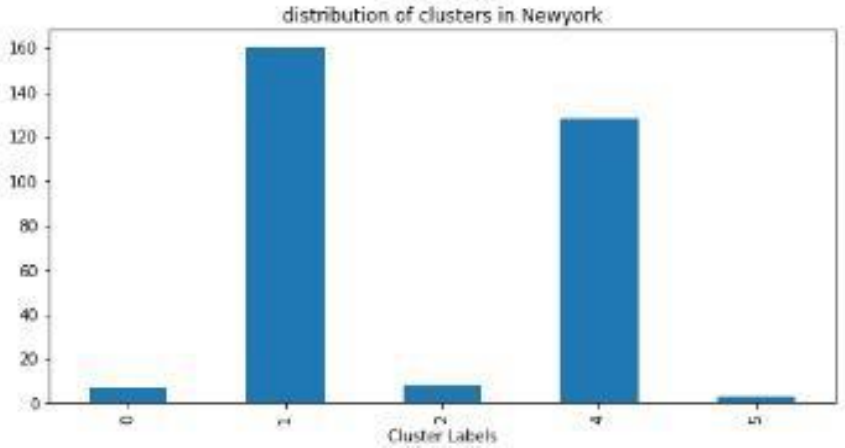


K=4

Distribution of clusters:







K=6

K=4

- In all 3 cities central and downtown neighborhoods mostly belong to cluster 2.
- In London and New York outer neighborhoods belong to cluster 0 in contrast to Toronto that has red neighborhoods in central areas of it.
- London is the only city with all 4 clusters as New York and Toronto do not have any neighborhood in cluster 3.
- In London and Toronto most neighborhoods are in cluster 2, while in New York cluster 1 is the dominant one.

K=4

- Although the majority of restaurants, bars and pubs are in cluster 2, they can be significantly seen in other clusters as well .
- Cities are located near either sea, river or lake which makes the neighborhoods near these areas similar.
- **All in all, it can be resulted from the number of neighborhoods in each cluster and their location in each city that these three cities are similar to each other in total but London and Toronto are more alike.**

DISCUSSION & CONCLUTION

- This project would have had better results if there were more data in terms of industrial places within the area, traffic access and allowance of more venues exploration with the Foursquare.
- The results also could potentially vary if we use some other clustering techniques like DBSCAN.
- All above analysis is depended on the adequacy and accuracy of Foursquare data.