Coursera Capstone project: Battle of Neighborhoods

1 INTRODUCTION

1.1 Background:

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. Tourism industry is important for the benefits it brings and due to its role as a commercial activity that creates demand and growth for many more industries. Tourism not only contributes towards more economic activities but also generates more employment, revenues and play a significant role in development. Many countries such as Turkey, France and Italy depend heavily on tourism industry for their expanses.

1.2 Interest:

Knowing what makes tourists choose their travel destination is crucial information for anyone working in the travel business. Therefore, for anyone who relies on tourists and tourism, understanding the consumer behavior is essential. In this project I will focus on venues such as restaurants, hotels, parks, cafes, cinemas and so on in London, Toronto and New York and cluster their neighborhoods in order to understand the similarities and differences between these cities. Therefor 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.

2 DATA

This project will analyze venues of the city of Toronto, New York and London.

The data below will be used for this analysis.

2.1 Boroughs and neighborhoods

2.1.1 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. I used this website: https://skgrange.github.io/www/data/london_sport.json

2.1.2 New York:

New York 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. Luckily, this dataset exists for free on the web. Here is the link to the dataset: https://geo.nyu.edu/catalog/nyu_2451_34572

2.1.3 Toronto:

For Toronto I used the table in Wikipedia for postal code and borough of each neighborhood. (link to the Wikipedia page: 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

2.2 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, New York.

3 METHODOLOGY:

3.1 Business understanding:

The main aim of this project is to cluster and segment neighborhoods in 3 different cities to check for similarity and difference between them.

3.2 Data Preparation:

3.2.1 London:

This data is contained in a JSON file that is downloaded from link:

https://skgrange.github.io/www/data/london_sport.json

This data is also transformed into Pandas data frame in python. It is cleaned containing only the name of neighborhoods and location of each neighborhood.

	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

3.2.2 Toronto:

The data used is available in a table in Wikipedia containing name of neighborhoods and boroughs of Toronto plus postal codes of each neighborhood. Then I used the CSV file to add latitude and longitude of each neighborhood to my data frame.

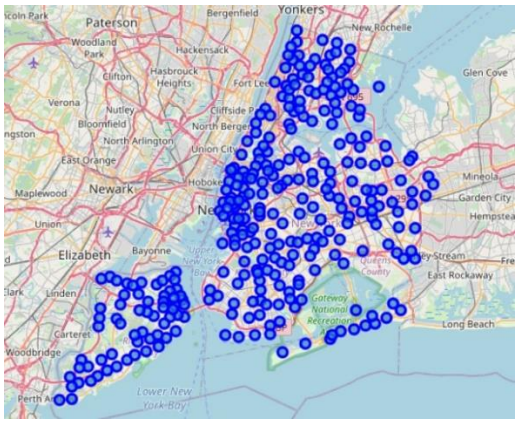
	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

3.2.3 New York:

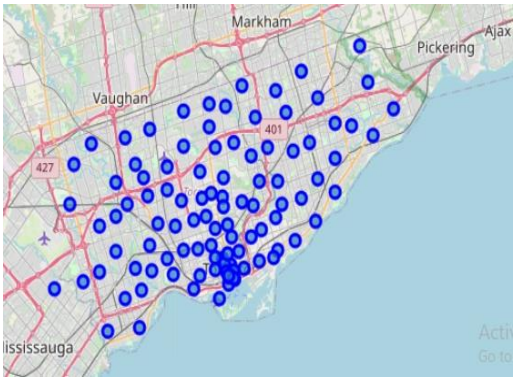
For accessing NY neighborhoods, the json file mentioned earlier is used and cleaned by extracting only name of boroughs and neighborhoods and their corresponding latitude and longitude.

	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

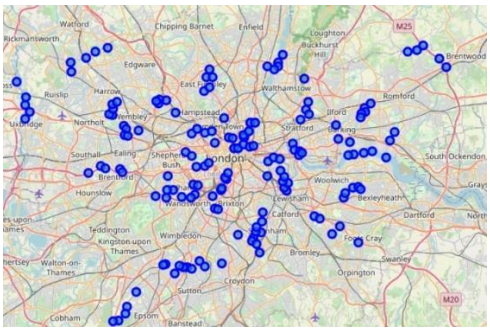
For better understanding, map of these cities with neighborhoods superimposed on them is provided.



New York



Toronto



London

3.3 Exploratory Data Analysis:

3.3.1 Using foursquare location data:

In this section, top 100 venues in each neighborhood that are within a radius of 500 meters will be explored in London, Toronto and New York. Then the resulting data frames will be combined in a data frame containing neighborhood name and its location, venues' name and location and the name of city. The resulting data frame is shown below.

	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

Our foursquare API calls resulted 15835 different venues in total and 505 unique venues categories.

```
print('There are {} uniques categories.'.format(len(three_cities['Venue Category'].unique())))
```

```
There are 505 uniques categories.
```

3.4 Machine learning:

3.4.1 one-hot encoding:

Then to analyze the data we gathered; we need a technique in which Categorical Data is transformed into Numerical Data for Machine Learning algorithms. This technique is called **One hot encoding**. For each of the neighborhoods, individual venues were turned into the frequency at how many of those Venues were located in each neighborhood. In each row, only one column is 1 and the others are 0.

```
three_cities_onehot = pd.get_dummies(three_cities[['Venue Category']], prefix="", prefix_sep="")
cols = three_cities_onehot.columns.tolist()
cols.remove('Neighborhood')
#cols.remove('City')
# add neighborhood column back to dataframe
three_cities_onehot['Neighborhood'] = three_cities['Neighborhood']
three_cities_onehot['City']=three_cities['City']
# move neighborhood column to the first column
fixed_columns = ['Neighborhood','City'] + cols
three_cities_onehot = three_cities_onehot[fixed_columns]

three_cities_onehot
```

	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	...	

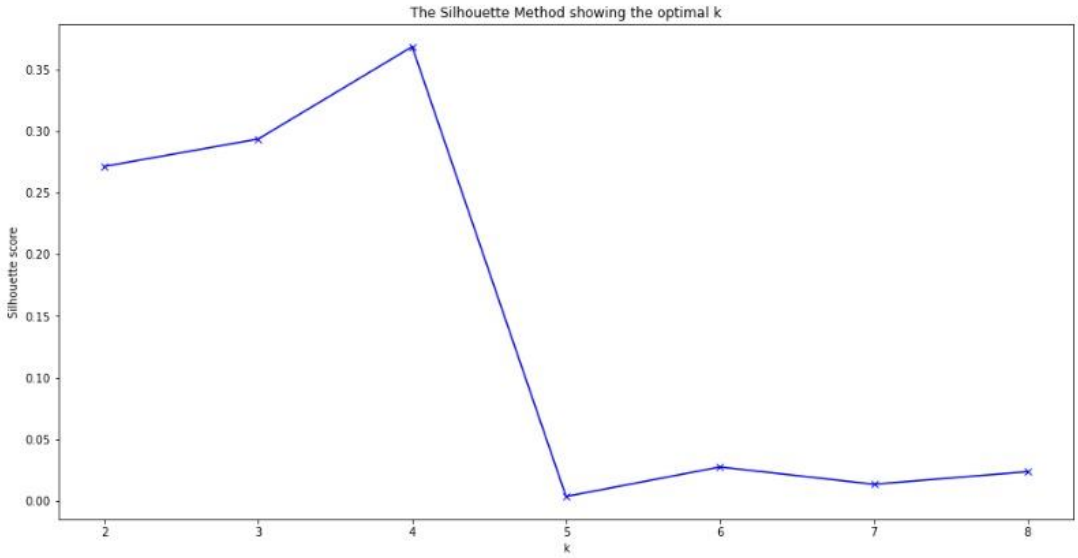
Next, we group rows by neighborhood and by taking the mean of the frequency of occurrence of each category

	Neighborhood	City	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	...
0	Agincourt	Toronto	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
1	Alderwood, Long Branch	Toronto	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
2	Allerton	Newyork	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
3	Annadale	Newyork	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
4	Arden Heights	Newyork	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...

3.4.2 K-means Clustering:

***k*-means clustering** is a method of vector quantization, originally from signal processing, that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. In this project, K-means clustering method is used and in order to find the best K, I used elbow and Silhouette method to find the optimum value for K (number of clusters).

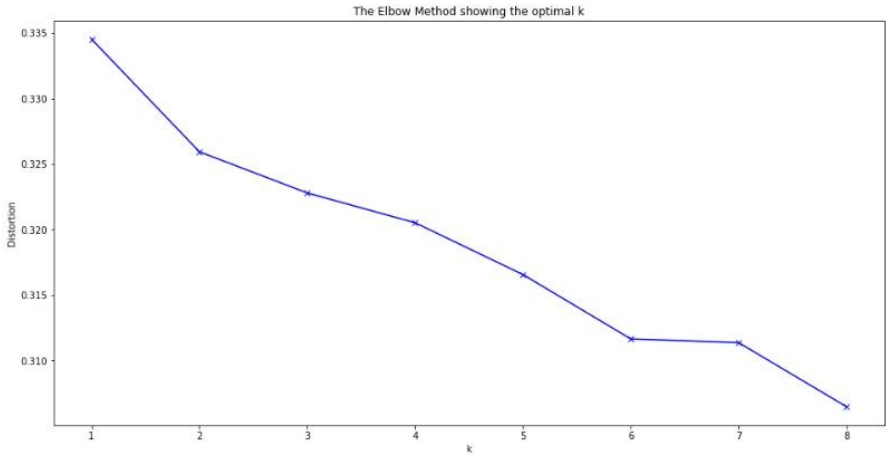
Silhouette method: The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation). The range of the Silhouette value is between +1 and -1. A high value is desirable and indicates that the point is placed in the correct cluster. If many points have a negative Silhouette value, it may indicate that we have created too many or too few clusters. The Silhouette score can be easily calculated in Python using the metrics module of Sklearn library. Here is the result of this method:



As it is mentioned before, a high Silhouette Score is desirable. The Silhouette Score reaches its global maximum at the optimal k. This should ideally appear as a peak in the Silhouette Value-versus-k plot. There is a clear peak at k = 4. Hence, therefore it is optimal.

Elbow method: This is probably the most well-known method for determining the optimal number of clusters. It is also a bit naive in its approach. We calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose

the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow.



As it is clear from the graph, there are bends in K=6.Therefore I will run this algorithm for K values of 4 and 6.

Clustering:

I used K-means clustering to segment neighborhoods and I chose k=4 and k=6 to have 4 and 6 clusters respectively. Clustering will be based on venue categories and neighborhoods with similar venues categories will be in the same cluster. Accordingly, we can compare neighborhoods in 3 cities after clustering them.

```
kclusters = 4

#three_cities_grouped_clustering = three_cities_grouped.drop('Neighborhood', 1)
#three_cities_grouped_clustering = three_cities_grouped_clustering.drop('City', 1)
three_cities_grouped_clustering
# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(three_cities_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:30]
```

```
array([2, 1, 1, 1, 1, 2, 1, 2, 2, 1, 2, 0, 2, 3, 2, 2, 2, 2, 2, 1, 1, 2,
       2, 2, 1, 2, 2, 2, 1, 2], dtype=int32)
```

K-means clustering for K=4

Now that we have cluster number of each neighborhood, we can add ‘Cluster Label’ column to our dataset and visualize each city with its clustered neighborhoods. The number of neighborhoods in each cluster is a below:

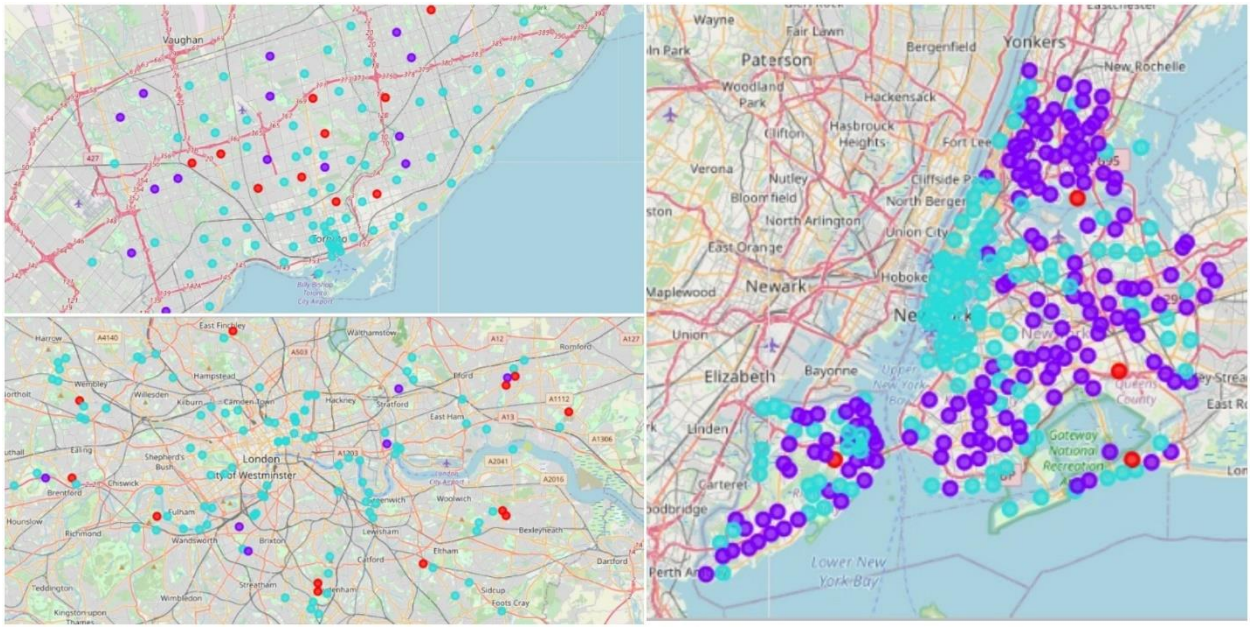
Cluster Labels	Neighborhood	Latitude	Longitude	City
0	33	33	33	33
1	185	185	185	185
2	340	340	340	340
3	2	2	2	2

Cluster Labels	Neighborhood	Latitude	Longitude	City
0	49	49	49	49
1	176	176	176	176
2	23	23	23	23
3	2	2	2	2
4	292	292	292	292
5	18	18	18	18

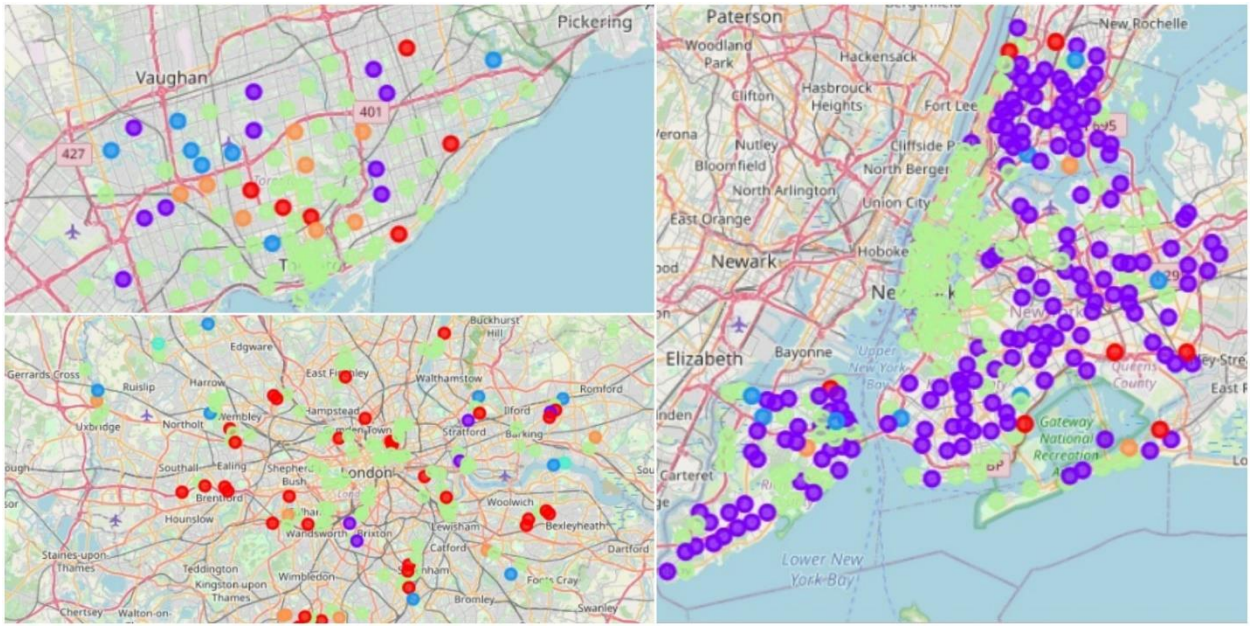
The number of neighborhoods in each cluster for K=4 and K=6

It is clear that for k=4 most neighborhoods are in cluster number 2 and 1 respectively, and for K=6 most neighborhoods go to cluster 4 and 1

For better understanding, I used Folium library for visualization.



Neighborhoods colored by number of clusters for K=4



Neighborhoods colored by number of clusters for K=6

4 Results:

I created a data frame of each neighborhood and its most common venues and cluster label. After exploring clusters, it was clear that each cluster mainly referred to which venues. here is the result:

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

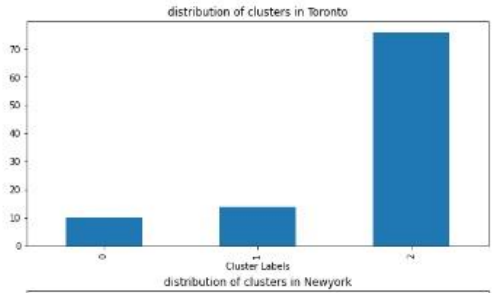
Cluster 2(light blue): Grocery Store, Playground

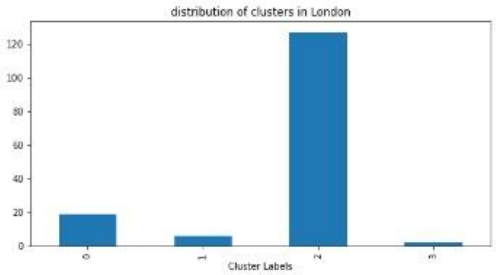
Cluster 3(creamy): Golf Course and zoo

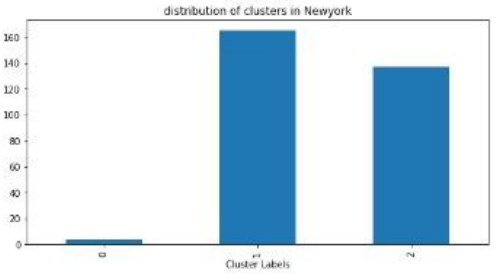
Cluster 4(light green): Bar, Pub, Restaurant, Cafe, Pharmacy

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

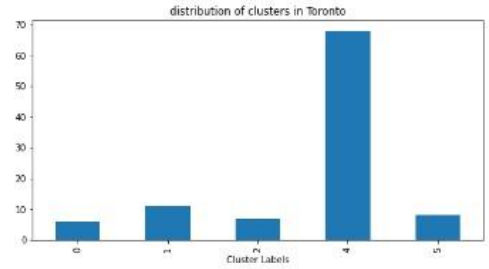
Here is a bar chart, depicting each city with number of neighborhoods in every cluster.

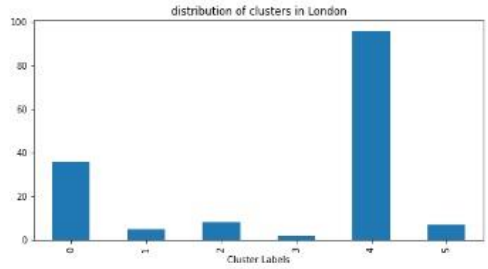


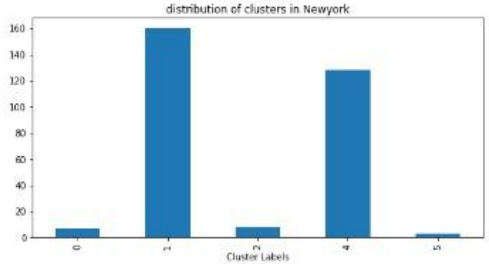




K=4







K=6

I only discussed my findings related to K=4 in detail for simplicity. Looking at the map for K=4, it can be seen that:

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 even 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.

Although the majority of restaurants, bars and pubs are in cluster 2, they can be significantly seen in other clusters as well.

They 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.

5 Discussion and Conclusion

In a fast-moving world, there are many real-life problems or scenarios where data can be used to find solutions to those problems. Like seen in the example above, data was used to cluster neighborhoods in London, Toronto and New York based on the most common venues in their major districts. These cities are the most significant city of their corresponding country and as a result, they are rich in industry, culture and tourism infrastructure. The results can help a traveler to have a better perspective of each city and its neighborhoods.

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 (limited venues for free calls). Furthermore, this results also could potentially vary if we use some other clustering techniques like

DBSCAN. As a final note, all of the above analysis is depended on the adequacy and accuracy of Foursquare data. A more comprehensive analysis and future work would need to incorporate data from other external databases.