The Battle of Neighbourhood's

Identifying the similarity (or dissimilarity) between Neighbourhoods, and predict any business opportunities amongst themselves

Koushik Karmakar

November 28, 2019

1. Introduction

1.1. Background

- Nowadays, cities across the world (let's say financial capitals of the countries) present an array of job, business and tourism opportunities, to attract people from different parts of the globe (especially from financial capital of a competitive country).
- At the same time, people (individuals, entrepreneurs, etc) in this era of globalization, are willing to travel across the world and grab these opportunities.
- While opportunities, play a major role in decision making (to travel across cities), another important factor that one lingers to, is Neighbourhood/locality. People can definitely let go of opportunities, if the destination cities doesn't present them with the choice of their preferred Neighbourhood/locality.
- A comparative study between one's current neighbourhood and the neighbourhoods of the destination city will be of great help in the decision-making process.

1.2. Problem

This project aims to compare the neighbourhoods of few major financial capitals of the world, on the basis of the venues (e.g. Restaurants, Parks, Museums, Hotels, Stores, etc.) present in the Neighbourhood, and present two pieces of information:

i. How similar or dissimilar are the neighbourhoods of one city, compared to another
ii. What are the business opportunities that the neighbourhoods present, in terms of their similarity with another neighbourhoods (within same city or different city)

1.3. Interest

- Tourists, who want to travel these cities. They can select the Neighbourhood to live, depending on what the Neighbourhood has to present to them or as per their own taste of Neighbourhood.
- People, who are willing to relocate across different cities of the world in search of better job opportunities.
- Entrepreneurs, who are willing to expand their business (overseas or within the same city). Using this report, they can identify locations, which have appetite for their business.

2. Data Acquisition and cleaning

2.1. Data Sources

There are two pieces of information that are required here, namely:

1. Neighbourhood information: Name/details of Borough, Neighbourhood, along with its longitude and latitude
2. All different kind of venues (Restaurants, Parks, Museums, Hotels, Stores, etc.) present in the neighbourhood.

As a part of this report, we will compare New York, Toronto & Paris. Thus, we needed the above information for the three cities.

1. New York: The information with neighbourhoods of New York (along with latitude and longitude), exists as a json file on the web.
2. Toronto: The information with neighbourhood information of the city of Toronto is present on a Wikipedia page. This information didn't have the latitude and longitude details. The latitude and longitude details are present in the web.
3. Paris: The neighbourhood information of Paris (known as Arrondissement & Quartiers) is present on a Wikipedia page. The latitude and longitude details were not present readily, thus were extracted one by one from google.

The venue information for each neighbourhood, was extracted using the **explore location** option of the **PLACES** API, provided by **FOURSQUARE**.

2.2. Data Processing/Cleaning and Feature Selection

- The Neighbourhood information (Neighbourhood details Name, Postcode, etc. & location information latitude, longitude, etc.) were available in a different manner for different cities. The task here is to:
  a. Extract the information from different sources (for e.g. json file for New York, wiki page & excel sheet for Toronto, & wiki page & google for Paris)
  b. Standardize the column information (for e.g. Neighbourhoods in New York and Toronto are segregated as Neighbourhood & Borough, whereas Neighbourhoods in Paris are segregated as Arrondissement (Districts) & Quartiers).
  c. Label Neighbourhoods from different cities before they can be merged into one frame.
  d. Finally, the datasets are merged to create one dataset, and the features are Borough, Neighbourhood, City, Latitude & Longitude. The total number of Neighbourhoods is 489, and city numbers are New York 306, Toronto 103 and Paris 80.
  The Final dataset, after completing the above process looks something like:

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

- **Explore location** option of the **PLACES API**, provided by **FOURSQUARE** was used to extract venues for within 1000 metre radius of each Neighbourhood/Location.
  - ✓ The API returned a total of 33,384 venues for the 489 neighbourhoods.
  - ✓ The total number of unique categories of venues returned was 535.

  The returned information looked something like:

	Borough	Neighbourhood	City	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bronx	Wakefield	1	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Bronx	Wakefield	1	40.894705	-73.847201	Ripe Kitchen & Bar	40.898152	-73.838875	Caribbean Restaurant
2	Bronx	Wakefield	1	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
3	Bronx	Wakefield	1	40.894705	-73.847201	Ali's Roti Shop	40.894036	-73.856935	Caribbean Restaurant
4	Bronx	Wakefield	1	40.894705	-73.847201	Jackie's West Indian Bakery	40.889283	-73.843310	Caribbean Restaurant

- There were certain venue categories, which were general as well as widely classified. For e.g.
  - ✓ There are restaurant & food venues are very well classified. Restaurant as Caribbean, Indian, French, Japanese, Brazilian, etc. (total 8752 venues). Food as Fast Food, Comfort Food, Food Trucks, etc. (total 358). But there are 410 venues, which are just marked as "Restaurant", and 52 which are just marked as "Food". We will rename the "Restaurant" & "Food" categories as "General Eating Venue".
  - ✓ There are other categories where general and classified categories co-exist (for e.g. Pub, Gastropub, Irish Pub or Market, Christmas Market, Flea Market, Fish Market, Super Market), but the total venues for these categories are very less. Thus, we will keep them as it is.
- The dataset that we now have (as seen in the above screenshot) is like, one row for each venue in a neighbourhood. Thus, we have 33,384 rows in our dataset. As our aim is to compare neighbourhoods, we need to compress this information to neighbourhood level. Thus, we:
  a. Use one hot encoding, on feature "Venue Category". This creates 534 columns (number of unique category) for each row.
  b. Then we group the information, on basis of 'Neighbourhood' and find the mean for each "Venue Category".
  c. The dataset now, looks something like this:

	Neighbourhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	...	Waterfront	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop
0	Adelaide, King, Richmond	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.01	0.0
1	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.00	0.0
2	Agincourt North, L'Amoreaux East, Milliken, St...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.00	0.0
3	Albion Gardens, Beaumond Heights, Humbergate, ...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.00	0.0
4	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.00	0.0

- This is the dataset we will use for comparing the Neighbourhoods. This dataset consists of 482 rows (neighbourhoods) & 535 features ('Neighbourhood' name, and 534 unique categories).
- **<u>Assumption</u>**:
  - ✓ The higher number of venues in a particular neighbourhood would indicate their higher popularity, amongst the residents of the neighbourhood.
  - ✓ Similarly, lesser number of venues in a particular neighbourhood, would indicate their unpopularity (or lesser popularity), amongst the residents of the neighbourhood.

3. Methodology

- The idea here is to produce a comparative study of the neighbourhoods, on basis of venues present in them.
- Thus, we are going to use the Non-supervised Machine Learning technique of Clustering, for our comparison process.
- The clustering techniques that were evaluated for the report were DBSCAN (density based), Agglomerative (Hierarchy based) & KMeans (Partition based).
- As in case of non-supervised ML techniques, evaluating the outcome of models are a bit tricky. Thus, here we chose a technique that gave us the most level of segregation.
- Let's go through each of the methods used, and see the results:

1. DBSCAN:

```
db=DBSCAN(eps=0.15,min_samples=10).fit(Neighborhood_grouped_clustering)
```

```
unique, counts = np.unique(db.labels_, return_counts=True)
for i in range(0,len(unique)):
    print('Label: {}, Count: {}'.format(unique[i],counts[i]))
```

```
Label: -1, Count: 209
Label: 0, Count: 273
```

This method seggregates 273 neighbourhoods into one category (label 0), & remaining 209 as outliers. Clearly not optimal clustering.

2. Agglomerative Clustering:

```
agglom = AgglomerativeClustering(n_clusters=3, linkage='complete').fit(Neighborhood_grouped_clustering)
```

```
agglom.labels_
unique, counts = np.unique(agglom.labels_, return_counts=True)
for i in range(0,len(unique)):
    print('Label: {}, Count: {}'.format(unique[i],counts[i]))
```

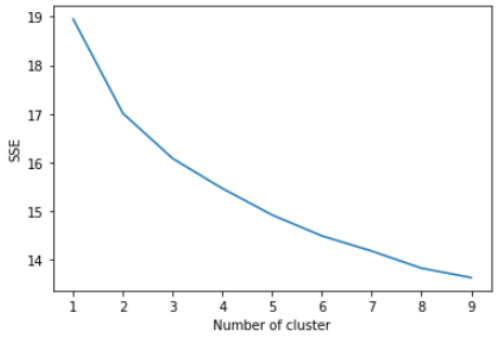
```
Label: 0, Count: 480
Label: 1, Count: 1
Label: 2, Count: 1
```

This method seggregates 480 neighbourhoods into one category (label 0), & one each into the other two category. Clearly not optimal clustering.

3. KMeans Clustering:

```
# Lets first choose a right value of k, using the Elbow Criterion Method
sse={}
for k in range(1,10):
    # run k-means clustering
    kmeans_loop = KMeans(init='k-means++',n_clusters=k, random_state=0,n_init=15).fit(Neighborhood_grouped_clustering)
    sse[k]=kmeans_loop.inertia_

plt.figure()
plt.plot(list(sse.keys()), list(sse.values()))
plt.xlabel("Number of cluster")
plt.ylabel("SSE")
plt.show()
```



It seems somewhere around 3, there is an elbow, after which the curve seems to follow the same slope

```
# set number of clusters
kclusters = 3

# run k-means clustering
kmeans = KMeans(init='k-means++',n_clusters=kclusters, random_state=0,n_init=15).fit(Neighborhood_grouped_clustering)

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

```
array([2, 1, 1, 1, 1, 1, 0, 1, 2, 1])
```

```
unique, counts = np.unique(kmeans.labels_, return_counts=True)
for i in range(0,len(unique)):
    print('Label: {}, Count: {}'.format(unique[i],counts[i]))
```

```
Label: 0, Count: 76
Label: 1, Count: 237
Label: 2, Count: 169
```

This method provides good segregation. Thus we will go ahead with this method.

- Thus, we choose KMeans Clustering as a basis for our Comparative Study.

4. Results

We will look at the results in this section, from **two perspectives**. The outcomes/recommendation from the results will be discussed in Discussion section.

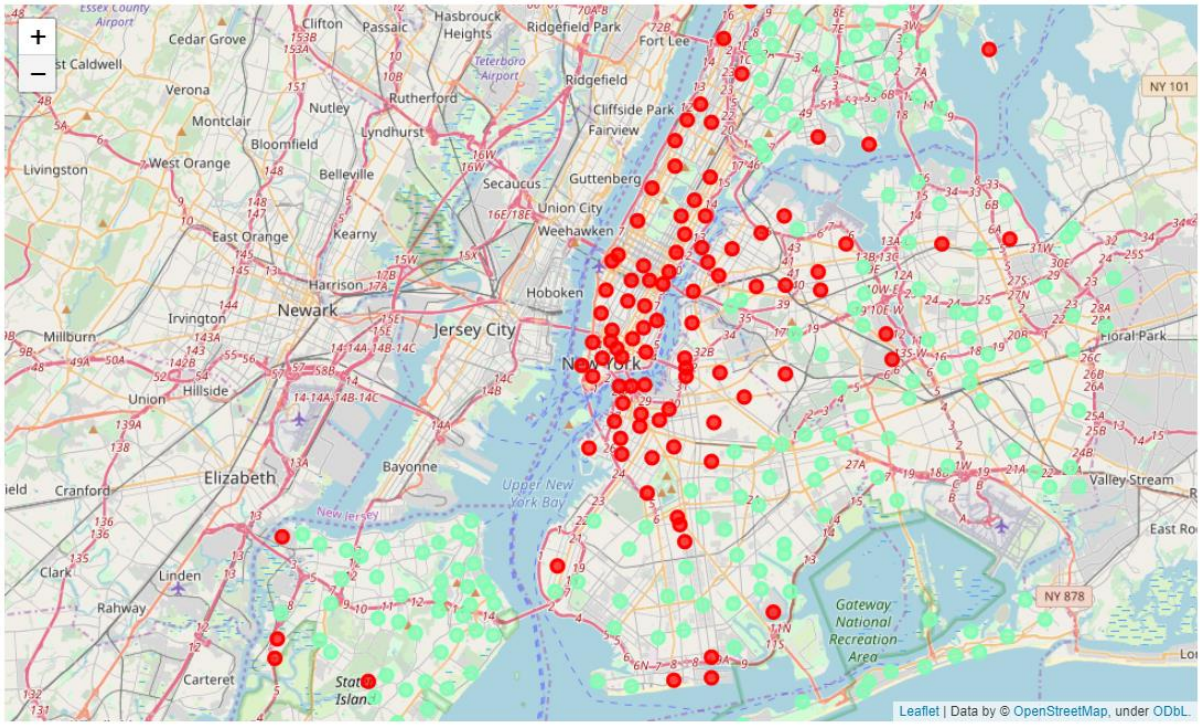
4.1. Similarity or Dissimilarity between the neighbourhoods

- The neighbourhoods within the three cities are segregated into three groups (or clusters). The cluster counts looks like:

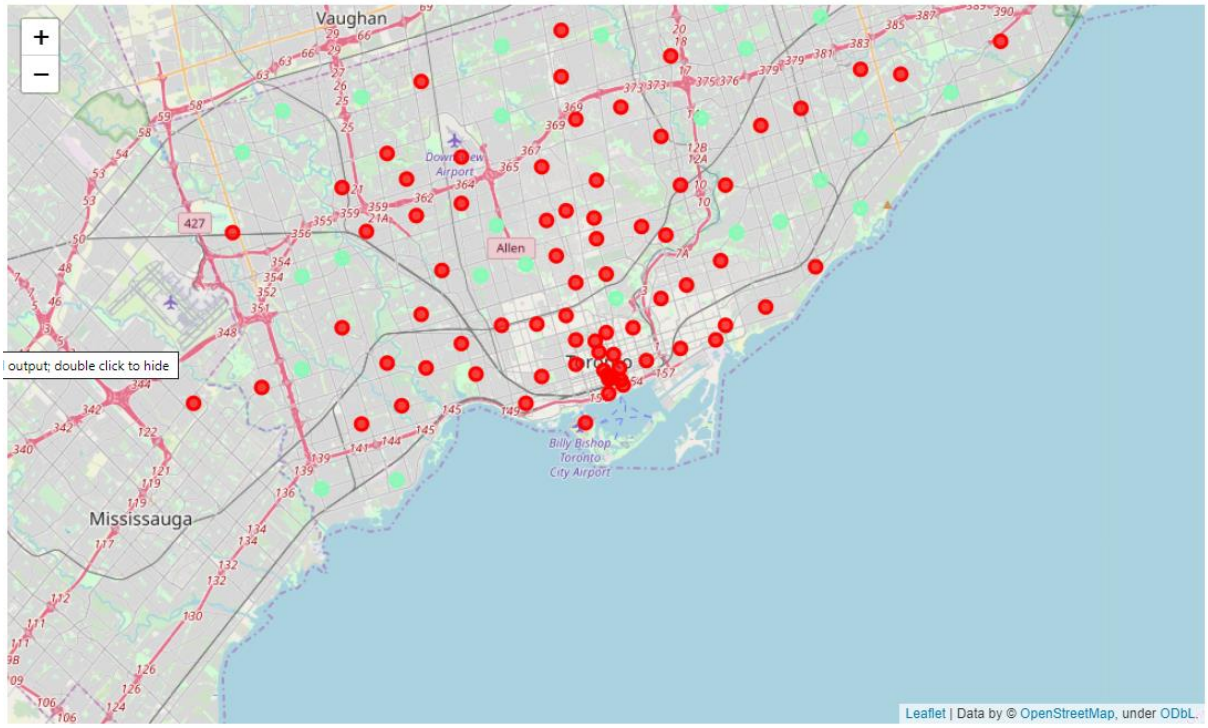
Cluster Label	New York	Toronto	Paris
0	0	0	76
1	211	28	1
2	95	74	3

- Let's look at the **map of these cities**, with the neighbourhood's colour coded as per their segregations.
  Colour Coding: Cluster 0 – Dark Purple, Cluster 1 – Fluorescent Green, Cluster 2 - Red

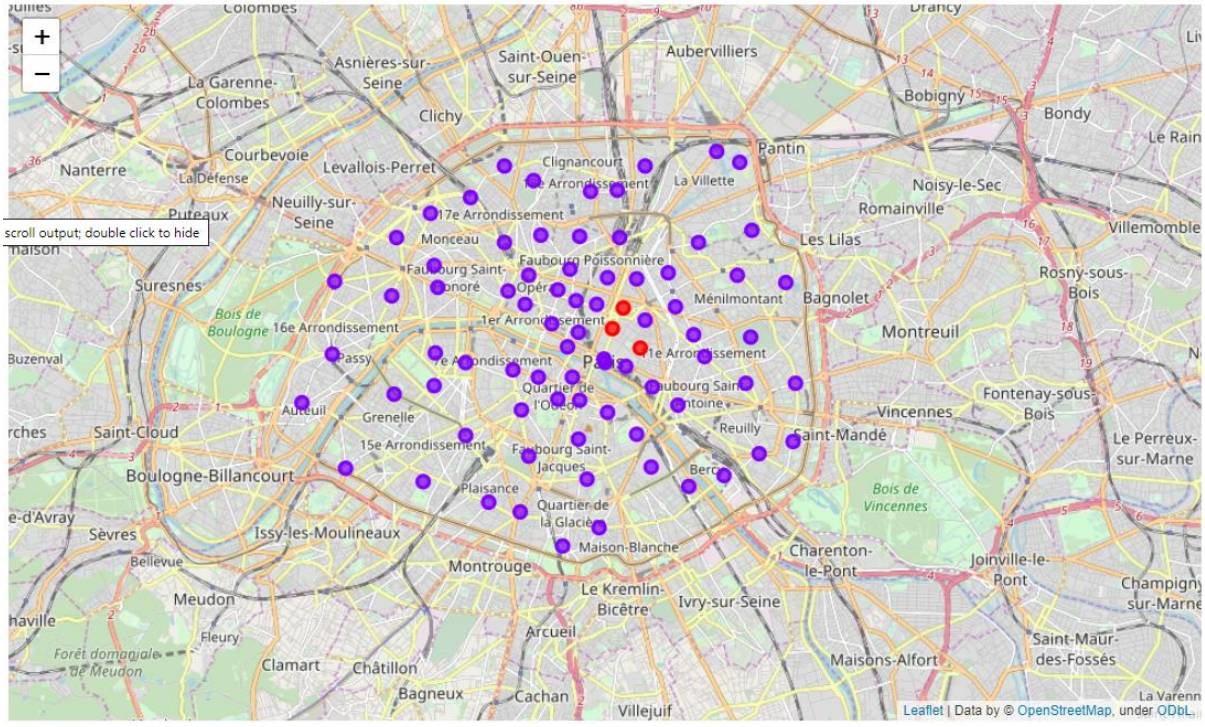
New York:



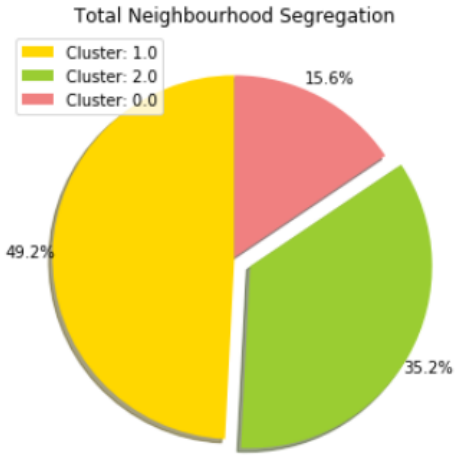
Toronto:

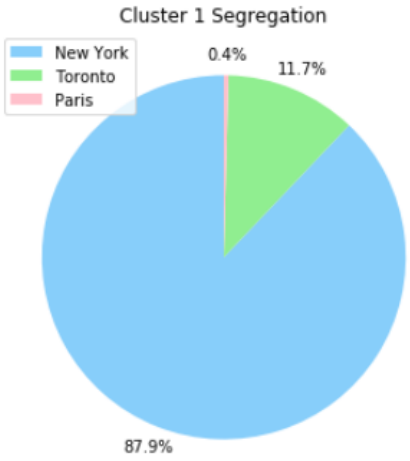


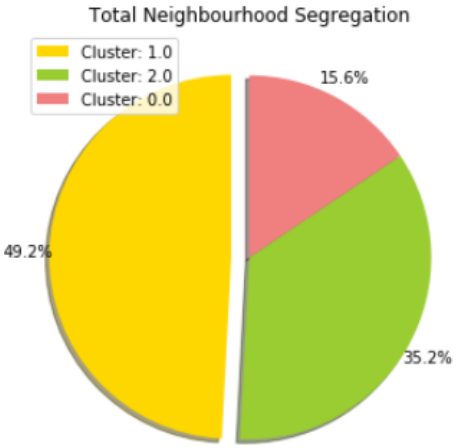
Paris:

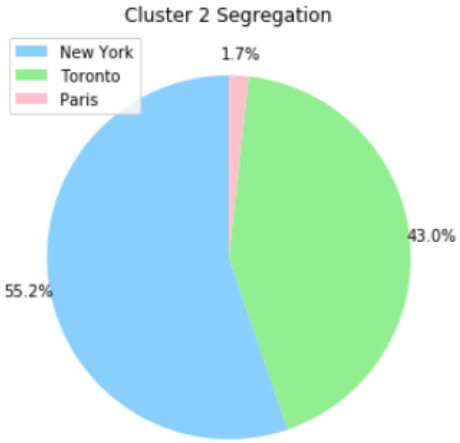


- Let's talk about the **similarity or dissimilarity between the neighbourhoods**:
  1. It seems New York & Toronto are much similar to each other. Both the cities are divided into two types of neighbourhood (cluster 1 & 2), and has good distribution of neighbourhoods amongst them. As can be seen below:

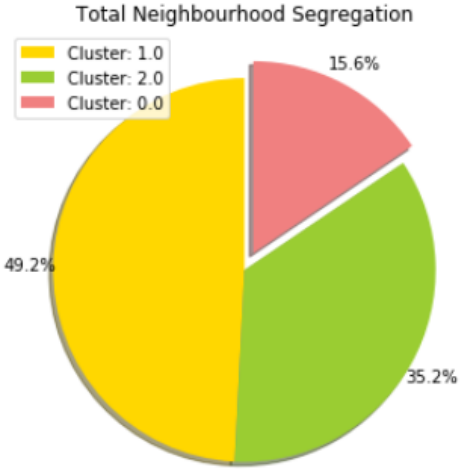


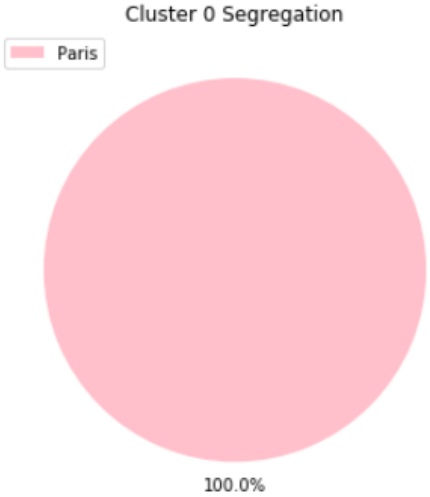






  2. Paris seems to be different from both New York & Toronto (as it is clustered, almost entirely into cluster 0). There are few neighbourhoods in Paris which are like the neighbourhood of New York & Toronto (cluster 1 & 2).





- The entire list of neighbourhoods, with clusters labels are present in the excel sheet. City Label: 1 New York, 2 Toronto, 3 Paris. We will call this sheet as ***similarity/dissimilarity matrix***.

4.2. Business Opportunities that each neighbourhood cluster has to offer

4.2.1. Information we have:

- We have two pieces of information for **each neighbourhood cluster**, namely:
  - ✓ **total number (or presence)** of each venue in a neighbourhood cluster &

Cluster Labels	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	Airport Service	Airport Terminal	Alsatian Restaurant	American Restaurant	...	Whisky Bar	Wine Bar	Wine Shop	Winery	Wings Joint
0	3.0	0.0	0.0	13.0	0.0	0.0	0.0	0.0	1.0	17.0	...	0.0	159.0	31.0	0.0	1.0
1	11.0	0.0	2.0	9.0	0.0	0.0	1.0	1.0	0.0	144.0	...	0.0	8.0	29.0	0.0	27.0
2	9.0	4.0	3.0	7.0	2.0	3.0	6.0	1.0	0.0	222.0	...	12.0	114.0	132.0	1.0	12.0

  - ✓ **mean (of presence)** of each venue in neighbourhood cluster

Cluster Labels	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	Airport Service	Airport Terminal	Alsatian Restaurant	American Restaurant	...	Weight Loss Center	Whisky Bar	Wine Bar	
0	0.000400	0.000000	0.000000	0.001733	0.000000	0.000000	0.000000	0.000000	0.000133	0.002266	...	0.000000	0.000000	0.021197	0.0
1	0.000872	0.000000	0.000158	0.000713	0.000000	0.000000	0.000079	0.000079	0.000000	0.011410	...	0.000317	0.000000	0.000634	0.0
2	0.000679	0.000302	0.000226	0.000528	0.000151	0.000226	0.000452	0.000075	0.000000	0.016740	...	0.000075	0.000905	0.008596	0.0

  Both piece of information represents same thing, but in a different manner.

- We have the **mean (of presence)** of each venue in an **individual neighbourhood.** This is the dataset, that we had used for classification

	Neighbourhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	...	Waterfront	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop
0	Adelaide, King, Richmond	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.01	0.0
1	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.00	0.0
2	Agincourt North, L'Amoreaux East, Milliken, St...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.00	0.0
3	Albion Gardens, Beaumond Heights, Humbergate, ...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.00	0.0
4	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.00	0.0

4.2.2. Let's look at how we can use this information:

- Let's first look at what each neighbourhood cluster has to offer, in terms of the **total number of venues present**,
  Or in another word,
  Let's look at 30 most famous venues (in order of their presence) in each neighbourhood cluster
  - ✓ **Cluster 0**:

  - ✓ **Cluster 1**:

✓ **Cluster 2**:

- For a **venue category** in a **cluster**, if the **neighbourhood mean** is **much less** than the **neighbourhood cluster mean,** then there is an **opportunity to grow the business** in that **particular neighbourhood.**
- The reason we say**, "much less"** is because of the **assumption** we had made in data section about presence of venue in a neighbourhood and its popularity/unpopularity.
- **For our report**:
  - ✓ We compared only for top 30 venue categories in each neighbourhood cluster
  - ✓ We will present an opportunity if the **neighbourhood mean** is **less than half of** the **neighbourhood cluster mean**. Below is the snapshot of the result:

	Neighbourhood	Venue Category	Venue Rank	Neighbourhood Mean	Cluster Mean	Opportunity	Cluster Label
137	Chaussée-d'Antin	French Restaurant	1	0.050000	0.135449	0.017724	0
340	La Chapelle	French Restaurant	1	0.066667	0.135449	0.001058	0
402	La Villette	French Restaurant	1	0.060000	0.135449	0.007724	0
532	Parc Montsouris	French Restaurant	1	0.042105	0.135449	0.025619	0
0	Amérique	Hotel	2	0.030000	0.068391	0.004195	0

Venue Rank indicates the ranking of the venue in top 30 venues in a particular cluster
Opportunity indicates the difference between **neighbourhood cluster mean (divided by two) and neighbourhood mean** for a particular Venue Category.

- There are 6838 business opportunities identified, and entire list can be found here. We will call this sheet has ***business opportunities matrix***.

5. Discussion:

In this section, we will try to visit **some** examples, on how one can use the ***similarity/dissimilarity matrix*** *and* ***business opportunities matrix***.

5.1. Similarity/dissimilarity matrix:

- Let's say, I am a resident of Neighbourhood Wakefield in Borough Bronx, and:
  - I want to relocate within city to Manhattan:
    - ✓ Neighbourhood Wakefield is a part of Cluster 1, so this will be my first filter
    - ✓ As I want to move within city to Manhattan, I will select City as 1 (New York) & Borough as Manhattan.
    - ✓ Thus my options are:

Borough	Neighbourhood	City	Latitude	Longitude	Cluster Labels
Manhattan	Marble Hill	1	40.87655078	-73.91065966	1

  - I want to relocate within city to Brooklyn:
    - ✓ Neighbourhood Wakefield is a part of Cluster 1, so this will be my first filter
    - ✓ As I want to move within city to Manhattan, I will select City as 1 (New York) & Borough as Manhattan.
    - ✓ Thus my options are:

Borough	Neighbourhood	City	Latitude	Longitude	Cluster Labels
Brooklyn	Bensonhurst	1	40.6110089	-73.99517998	1
Brooklyn	Sunset Park	1	40.64510295	-74.01031619	1
Brooklyn	Gravesend	1	40.59526001	-73.97347088	1
Brooklyn	Manhattan Terrace	1	40.61443251	-73.95743841	1
Brooklyn	East Flatbush	1	40.64171777	-73.93610256	1
Brooklyn	Kensington	1	40.64238196	-73.98042111	1
Brooklyn	Brownsville	1	40.66394994	-73.91023536	1
Brooklyn	Cypress Hills	1	40.68239101	-73.87661596	1
Brooklyn	East New York	1	40.6699257	-73.88069864	1
Brooklyn	Starrett City	1	40.64758905	-73.8793697	1
Brooklyn	Canarsie	1	40.63556433	-73.9020927	1
Brooklyn	Flatlands	1	40.63044604	-73.92911303	1
Brooklyn	Coney Island	1	40.57429256	-73.98868296	1
Brooklyn	Bath Beach	1	40.5995187	-73.99875221	1
Brooklyn	Borough Park	1	40.63313051	-73.99049823	1
Brooklyn	Dyker Heights	1	40.61921946	-74.01931376	1
Brooklyn	Gerritsen Beach	1	40.59084843	-73.93010171	1
Brooklyn	Marine Park	1	40.60974778	-73.93134404	1

    - ✓ ** Only few options are shown here, refer to excel sheet for exact neighbourhood information.
    - ✓ The options here are too many. Venue information, on basis of popularity can be further used to make a sound decision.
  - I want to visit Paris for a short trip, and want to live in a neighbourhood similar to mine:
    - ✓ Neighbourhood Wakefield is a part of Cluster 1, so this will be my first filter
    - ✓ As I want to travel to Paris, I will select City as 3 (Paris)
    - ✓ Thus my options are:

Borough	Neighbourhood	City	Latitude	Longitude	Cluster Labels
2nd arrondissement	Gaillon	3	49.16104	1.34016	1

5.2. Business Opportunities Matrix:

- Let's first look at the snapshot of the matrix:

Borough	Neighbourhood	City	Venue Category	Venue Rank	Neighbourhood Mean	Cluster Mean	Opportunity	Cluster Lab
9th arrondissement(Called "de l'OpÃ©	ChaussÃ©e-d'Antin	3	French Restaurant	1	0.05	0.135448607	0.017724303	0
18th arrondissement(Called "des Butt	La Chapelle	3	French Restaurant	1	0.066666667	0.135448607	0.001057637	0
19th arrondissement(Called "des Butt	La Villette	3	French Restaurant	1	0.06	0.135448607	0.007724303	0
14th arrondissement(Called "de l'Obs	Parc Montsouris	3	French Restaurant	1	0.042105263	0.135448607	0.02561904	0
19th arrondissement(Called "des Butt	AmÃ©rique	3	Hotel	2	0.03	0.068390881	0.004195441	0
16th arrondissement(Called "de Passy	Auteuil	3	Hotel	2	0.01754386	0.068390881	0.016651581	0
20th arrondissement(Called "de MÃ©	Belleville	3	Hotel	2	0	0.068390881	0.034195441	0
18th arrondissement(Called "des Butt	Clignancourt	3	Hotel	2	0	0.068390881	0.034195441	0
19th arrondissement(Called "des Butt	Combat	3	Hotel	2	0	0.068390881	0.034195441	0
3rd arrondissement(Called "du Templ	Enfants-Rouges	3	Hotel	2	0.03	0.068390881	0.004195441	0
11th arrondissement(Called "de Popin	Folie-MÃ©ricourt	3	Hotel	2	0.01	0.068390881	0.024195441	0

- <u>Few important fields</u>:
  - ✓ Venue Category: The Business that we are talking about
  - ✓ Venue Rank: Popularity ranking of a business, in a cluster neighbourhood
  - ✓ Opportunity: The difference between neighbourhood cluster mean (divided by two) and neighbourhood mean for a particular Venue Category. The higher the value, the better chance of the business succeeding.
- Let's say, I run a coffee shop in in Eastchester, Bronx and I want to expand my business within the same Borough. What are my options of the neighbourhood?
  - ✓ So if I filter on Borough = 'Bronx' & Venue Category = 'Coffee Shop', and sort by Opportunity in descending order, I see below options:

Borough	Neighbourhood	Ci	Venue Category	Venue Rai	Neighbourhood Mea	Cluster Mei	Opportuni	Cluster Lab
Bronx	City Island	1	Coffee Shop	1	0	0.052782386	0.026391193	2
Bronx	Clason Point	1	Coffee Shop	1	0	0.052782386	0.026391193	2
Bronx	High Bridge	1	Coffee Shop	1	0	0.052782386	0.026391193	2
Bronx	Allerton	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Castle Hill	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Claremont Village	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Concourse	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Country Club	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Eastchester	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Edenwald	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Melrose	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Morris Heights	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Morrisania	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Mount Eden	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Mount Hope	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Olinville	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Pelham Bay	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Pelham Gardens	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Soundview	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Unionport	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	University Heights	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Wakefield	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Williamsbridge	1	Coffee Shop	14	0	0.018302829	0.009151414	1
Bronx	Woodlawn	1	Coffee Shop	14	0	0.018302829	0.009151414	1

  - ✓ Cluster 2 (i.e. Neighbourhoods City Island, Clason Point, & High Bridge) is where coffee shop is liked more (Venue Rank = 1), and the Opportunity is higher.
- Let's say, I run a chain of coffee shops in New York City, and want to expand the business in Toronto or Paris. What are my options of the neighbourhood?
  - ✓ So if I filter on Venue Category = 'Coffee Shop', city = '2' or '3', and sort output by Venue Rank (ascending) & Opportunity (descending), below are my options:

Borough	Neighbourhood	Ci	Venue Category	Venue Ra	Neighbourhood Mea	Cluster Me	Opportuni	Cluster Lab
Scarborough	Birch Cliff, Cliffside West	2	Coffee Shop	1	0	0.052782386	0.026391193	2
North York	Downsview Central	2	Coffee Shop	1	0	0.052782386	0.026391193	2
North York	Emery, Humberlea	2	Coffee Shop	1	0	0.052782386	0.026391193	2
Scarborough	Highland Creek, Rouge Hill, Port Union	2	Coffee Shop	1	0	0.052782386	0.026391193	2
Etobicoke	Humber Bay, King's Mill Park, Kingsway Park South East, Mimico NE, Old Mill	2	Coffee Shop	1	0	0.052782386	0.026391193	2
North York	Silver Hills, York Mills	2	Coffee Shop	1	0	0.052782386	0.026391193	2
19th arrondissement(Called "des Buttes-(	AmÃ©rique	3	Coffee Shop	11	0	0.0179976	0.0089988	0
16th arrondissement(Called "de Passy")	Auteuil	3	Coffee Shop	11	0	0.0179976	0.0089988	0
12th arrondissement(Called "de Reuilly")	Bel-Air	3	Coffee Shop	11	0	0.0179976	0.0089988	0
16th arrondissement(Called "de Passy")	Chaillot	3	Coffee Shop	11	0	0.0179976	0.0089988	0
20th arrondissement(Called "de MÃ©nilm	Charonne	3	Coffee Shop	11	0	0.0179976	0.0089988	0
18th arrondissement(Called "des Buttes-I	Grandes-CarriÃ¨res	3	Coffee Shop	11	0	0.0179976	0.0089988	0
16th arrondissement(Called "de Passy")	La Muette	3	Coffee Shop	11	0	0.0179976	0.0089988	0
11th arrondissement(Called "de Popincou	La Roquette	3	Coffee Shop	11	0	0.0179976	0.0089988	0
13th arrondissement(Called "des Gobelin	Maison-Blanche	3	Coffee Shop	11	0	0.0179976	0.0089988	0
14th arrondissement(Called "de l'Observ	Montparnasse	3	Coffee Shop	11	0	0.0179976	0.0089988	0
15th arrondissement(Called "de Vaugirar	Necker	3	Coffee Shop	11	0	0.0179976	0.0089988	0
14th arrondissement(Called "de l'Observ	Parc Montsouris	3	Coffee Shop	11	0	0.0179976	0.0089988	0
17th arrondissement(Called "des Batigno	Plaine Monceau	3	Coffee Shop	11	0	0.0179976	0.0089988	0
20th arrondissement(Called "de MÃ©nilm	PÃ¨re-Lachaise	3	Coffee Shop	11	0	0.0179976	0.0089988	0
20th arrondissement(Called "de MÃ©nilm	Saint-Fargeau	3	Coffee Shop	11	0	0.0179976	0.0089988	0
6th arrondissement(Called "du Luxembou	Saint-Germain-des-PrÃ©s	3	Coffee Shop	11	0	0.0179976	0.0089988	0
1st arrondissement(Called "du Louvre")	Saint-Germain-l'Auxerrois	3	Coffee Shop	11	0	0.0179976	0.0089988	0
Etobicoke	Albion Gardens, Beaumond Heights, Humbergate, Jamestown, Mount Olive	2	Coffee Shop	14	0	0.018302829	0.009151414	1
North York	Bayview Village	2	Coffee Shop	14	0	0.018302829	0.009151414	1
Scarborough	Cliffcrest, Cliffside, Scarborough Village West	2	Coffee Shop	14	0	0.018302829	0.009151414	1
Etobicoke	Cloverdale, Islington, Martin Grove, Princess Gardens, West Deane Park	2	Coffee Shop	14	0	0.018302829	0.009151414	1
2nd arrondissement(Called "de la Bourse	Gaillon	3	Coffee Shop	14	0	0.018302829	0.009151414	1
Etobicoke	Humber Bay Shores, Mimico South, New Toronto	2	Coffee Shop	14	0	0.018302829	0.009151414	1
North York	Humber Summit	2	Coffee Shop	14	0	0.018302829	0.009151414	1

- ✓ The options highlighted in blue are best bet, as the Venue Rank is higher and Opportunity is more. The neighbourhoods selected are from City 2 i.e. Toronto. For Paris, the options where City is 3 can be reviewed.

- As I had mentioned earlier, these are few examples of how the matrices can be used. The opportunities are limitless.

6. Conclusion:

- As a part of this report, I have tried to compare neighbourhood of three cities (New York, Toronto & Paris), on the basis of the venues present in the cities.
- I have used K-Means clustering algorithm to cluster the cities, and was able to cluster them into 3 groups/clusters. The results are presented as Similarity/dissimilarity matrix and can be found here. This information can be used by tourists or peoples who would like to relocate to or explore neighbourhoods within these three cities.
- The neighbourhoods within clusters were further compared, to identify if there are any business opportunities present within a neighbourhood. The results are presented as Business opportunities matrix and can be found here. This information can be used by entrepreneurs, who are willing to expand their business (overseas or within the same city).