The Battle of Neighbourhood's

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

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.

Problem

- This study 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:
    1. How similar or dissimilar are the neighbourhoods of one city, compared to another
    2. What are the business opportunities that the neighbourhoods present, in terms of their similarity with another neighbourhoods (within same city or different city)

Target Audience

- 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 has appetite for their business.

Data Acquisition and Cleaning

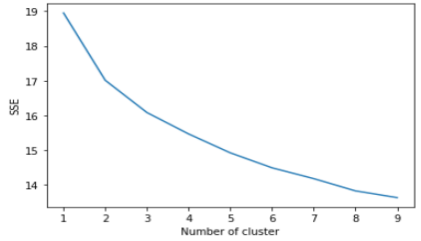
- There are two pieces of information that are required here, namely:
  1. Neighbourhood information: Name/details of Borough, Neighbourhood, along with its longitude and latitude
  2. All different kind of venues (Restaurants, Parks, Museums, Hotels, Stores, etc) present in the neighbourhood.
- As a part of this report, we will compare the cities of New York, Toronto & Paris.
- The Neighbourhood information had to be scrapped from multiple different sources (Wikipedia, google, etc).
  - The total number of Neighbourhoods are 489, and city numbers are New York 306, Toronto 103 and Paris 80.
- The venue information for each neighbourhood was extracted using explore location option of the PLACES API, provided by FOURSQUARE.
  - The API returned a total of 33,384 venues for the 489 neighbourhoods.
  - The total number of unique categories of venues returned, were 535.
- The venues information was then compressed to Neighbourhood level by using One Hot Encoding method, and rows were grouped to find mean at Neighbourhood Level.
  - This dataset consists of 482 rows (neighbourhoods) & 535 features ('Neighbourhood' name, and 534 unique categories).
- **Assumption**:
  - 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.

Comparison using ML

- Unsupervised Machine Learning Algorithm K-Means, was used compare the neighbourhoods, and cluster them into groups.
- The optimum number of clusters were found using the Elbow Criterion Method, which came out to be 3, as per the below snippet.
- The algorithm was able to successfully cluster the 482 neighbourhoods into three groups/clusters, with good spread across each cluster.
- In the next section, we will evaluate the results of clustering from two perspective:
  - Similarity (or Dissimilarity) of Neighbourhoods
  - Business opportunities that each neighbourhood cluster present

```
# 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])
```

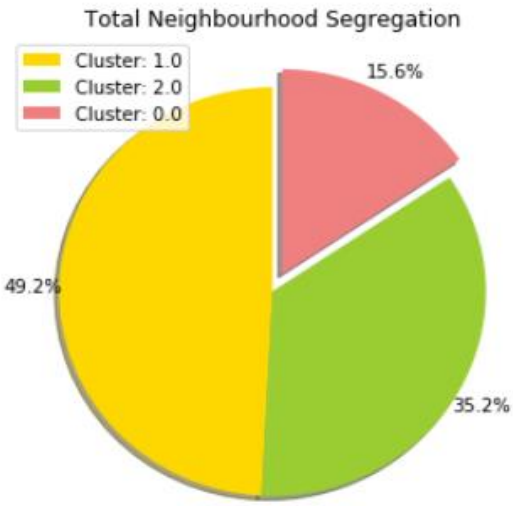
```
unique, counts = np.unique(kmeans.labels_, return_counts=True)
for i in range(0,len(unique)):
    print('Label: {}, Count: {}'.format(unique[i],counts[i]))
```

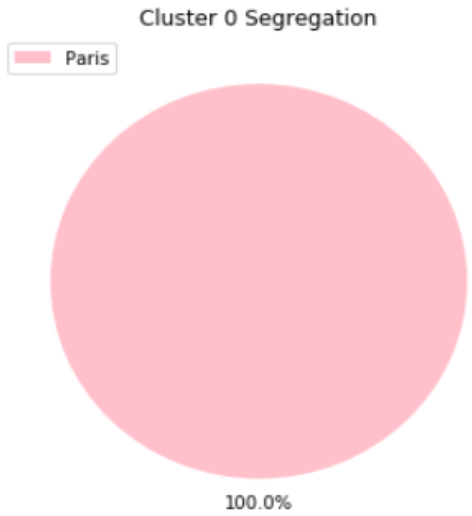
```
Label: 0, Count: 76
Label: 1, Count: 237
Label: 2, Count: 169
```

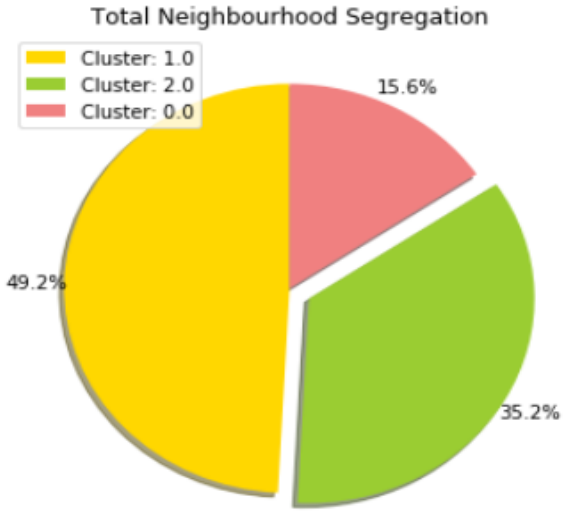
This method provides good segregation. Thus we will go ahead with this method.

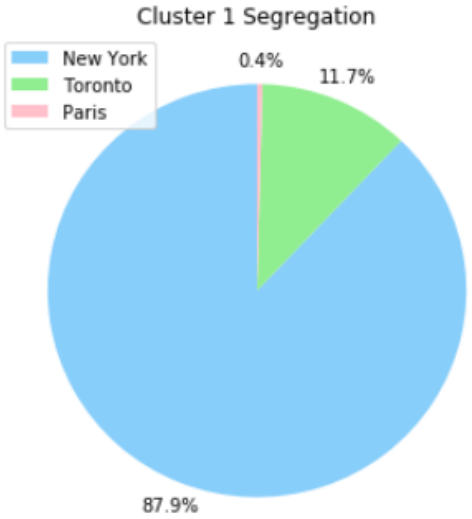
Similarity (or Dissimilarity)

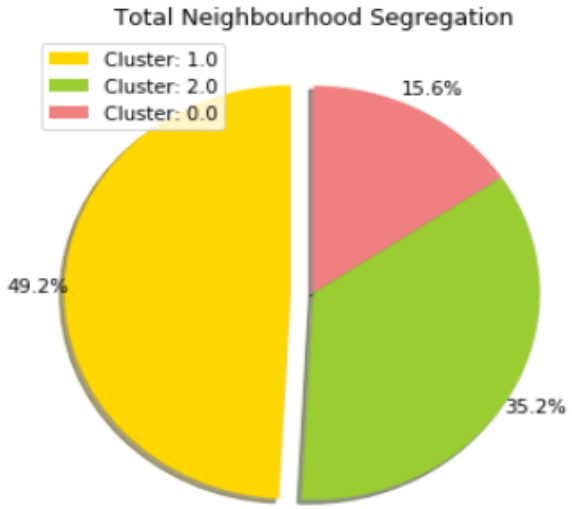
- Let’s visualize the cluster on a Pie Chart

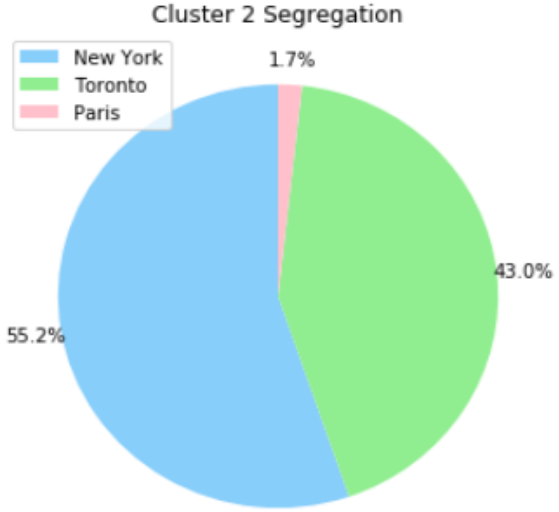






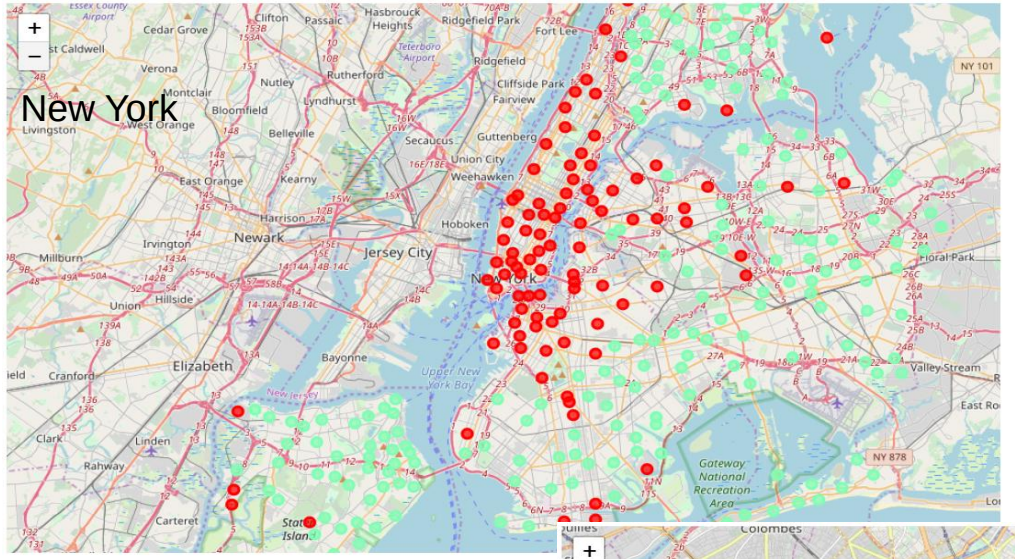


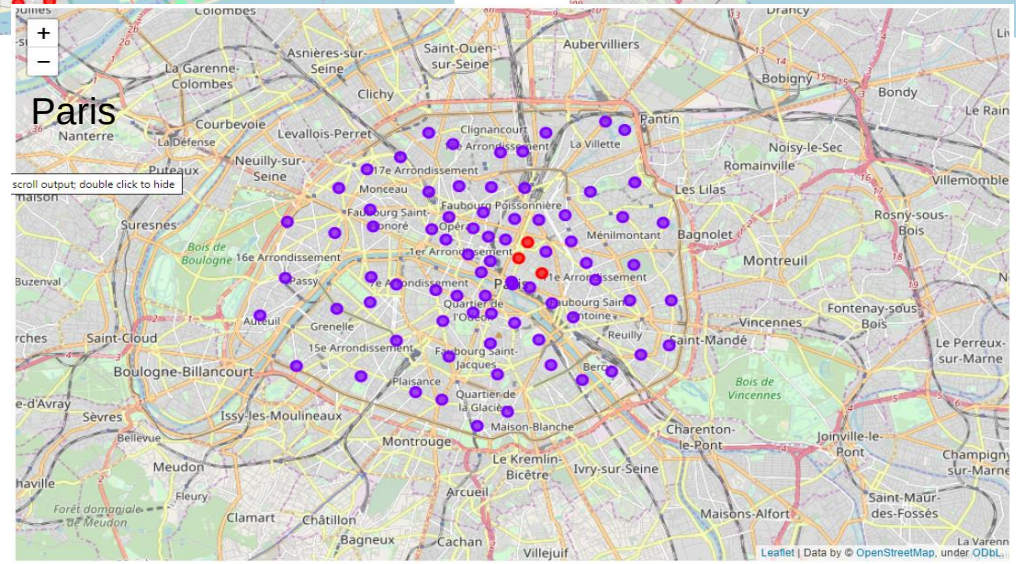




Similarity (or Dissimilarity)

- Lets visualize the neighbourhoods colour coded (as per the clusters) on the map





Colour Coding:
Cluster 0: Dark Purple
Cluster 1: Fluorescent Green
Cluster 2: Red

Similarity (or Dissimilarity)

- 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.
- 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 neighbourhoods of New York & Toronto (cluster 1 & 2).
- The results of clustering will be called similarity/dissimilarity matrix, and can be found here.

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

Business Opportunities

- Let’s visualize the most popular venues in each neighbourhood

Cluster 0

Cluster 1

Cluster 2

Business Opportunities

- In the last slide, the venues in the neighbourhood cluster were ranked (popularity), on the basis of their presence in the neighbourhood cluster (cluster mean value). For e.g. Coffee Shop is the most popular venue in Cluster 2.
- If there is a neighbourhood in a neighbourhood cluster, where the presence of a venue (neighbourhood mean value) is much lesser than presence of venue in neighbourhood cluster (cluster mean value), then there is an opportunity for the business in said neighbourhood. For e.g.
  - Bayside, Queens is a part of Neighbourhood Cluster 2.
  - Coffee shop is the most popular venue in Neighbourhood Cluster 2, with cluster mean value of 0.05278.
  - The presence of Coffee Shops in Bayside, Queens is 0.02 (neighbourhood mean value).
  - As the neighbourhood mean value of 0.02, is much lesser than the cluster mean value of 0.05278 for venue ‘Coffee Shop’, there is an opportunity here to grow the business.
- There are 6838 such business opportunities identified, and they can be found here.

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. 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 cluster. The results are presented as Business opportunities matrix. This information can be used by entrepreneurs, who are willing to expand their business (overseas or within the same city).