Identifying common patterns and difference between neighborhoods belonging to two different cities

Introduction

- This presentation shows the results of an investigation aimed to understand if is possible to measure how similar two cities are and how the different neighbourhoods of each city correlate.
- The cities chosen for this work are Toronto and New York.
- Foursquare location data has been used to collect the various venues belonging to the neighbourhoods of each city.
- K-Means clustering has be then used to identify categories in which to divide the neighbourhoods of each city.

Data Acquisition: Original Datasets

	Postcode	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront
3	M5A	Downtown Toronto	Regent Park
4	M6A	North York	Lawrence Heights
5	M6A	North York	Lawrence Manor

The Toronto Dataset contained each Postcode, Borough and Neighborhood of the city.
To use Foursquare, however, we need also the coordinates of each neighbourhood.

Data Acquisition: Original Datasets

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.758800	-79.320197
1	M4A	North York	Victoria Village	43.732658	-79.311189
2	M5A	Downtown Toronto	Harbourfront	43.640080	-79.380150
3	M5A	Downtown Toronto	Regent Park	43.660706	-79.360457
4	M6A	North York	Lawrence Heights	43.722778	-79.450933
5	M6A	North York	Lawrence Manor	43.722079	-79.437507

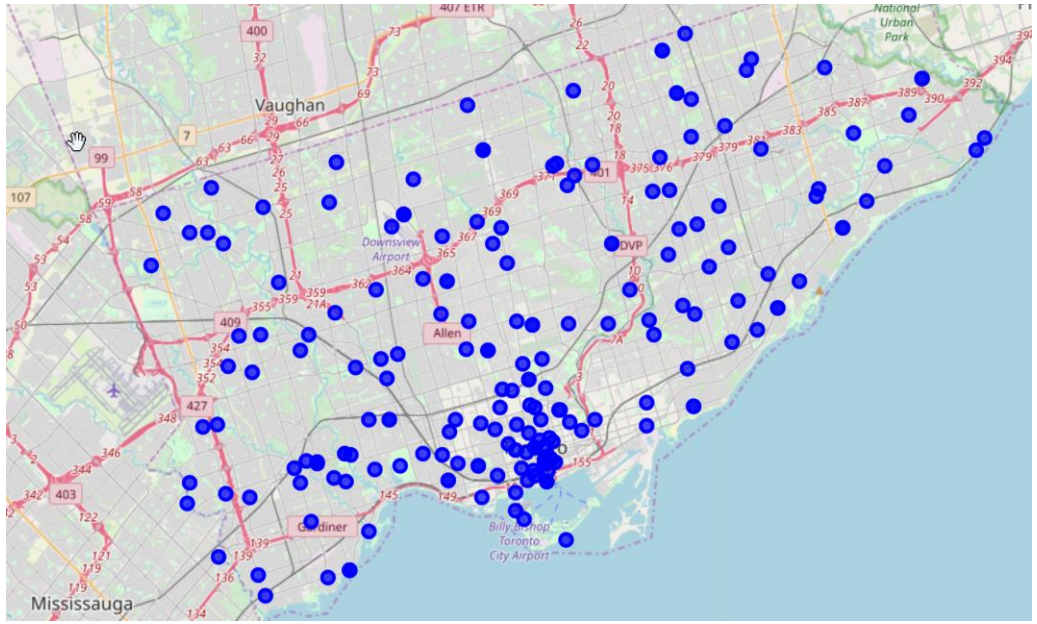
To add latitude and longitude for each neighborhood we have used Geocode

Data Acquisition : Original Datasets

	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

New York Dataset was completed with coordinates of every neighborhood so no extra action was necessary

Data Acquisition: Original Datasets



The location of each neighborhoods can be plotted using folium.
From the density of spots on the masp It can be easily noticed that New York is a larger city than Toronto.

Data Acquisition: Venues Extraction

	City	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Toronto	Parkwoods	43.7588	-79.320197	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant
1	Toronto	Parkwoods	43.7588	-79.320197	LCBO	43.757774	-79.314257	Liquor Store
2	Toronto	Parkwoods	43.7588	-79.320197	Tim Hortons	43.760668	-79.326368	Café
3	Toronto	Parkwoods	43.7588	-79.320197	A&W Canada	43.760643	-79.326865	Fast Food Restaurant
4	Toronto	Parkwoods	43.7588	-79.320197	Dollarama	43.757317	-79.312578	Discount Store

- For each dataset the Foursquare location service was used to identify venues belonging to each neighborhoods.
- A radius of 900 m around each location
- The maximum number of venues for a single location was set to 500.
- Of the fields provided by Foursquare the one used for the rest of the investigation is Venue Category.

Data Acquisition: Venues Extraction

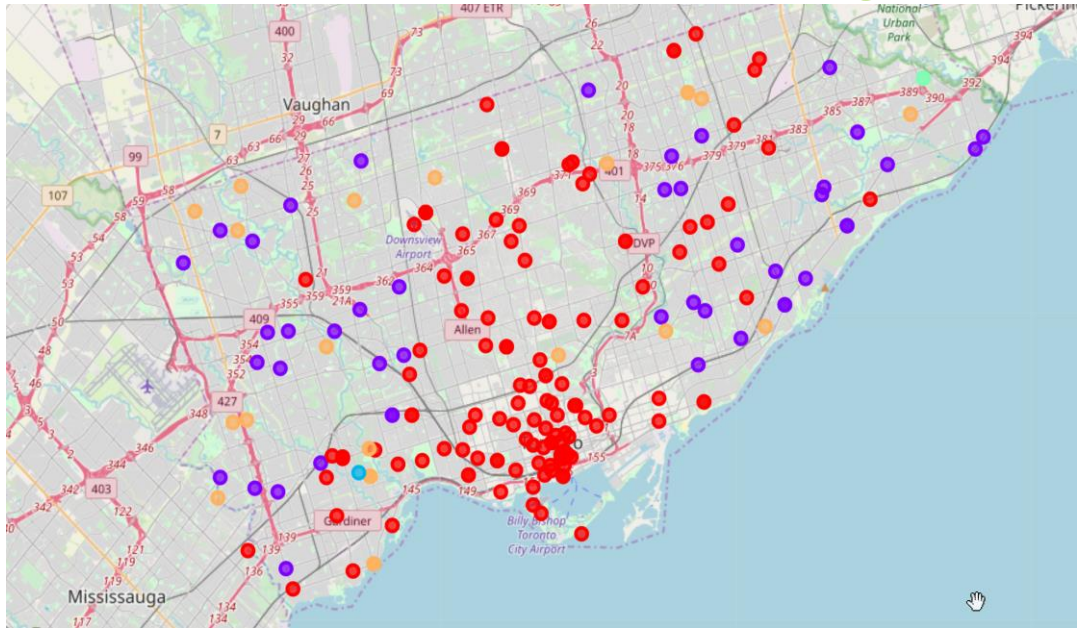
	City	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Toronto	Parkwoods	43.7588	-79.320197	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant
1	Toronto	Parkwoods	43.7588	-79.320197	LCBO	43.757774	-79.314257	Liquor Store
2	Toronto	Parkwoods	43.7588	-79.320197	Tim Hortons	43.760668	-79.326368	Café
3	Toronto	Parkwoods	43.7588	-79.320197	A&W Canada	43.760643	-79.326865	Fast Food Restaurant
4	Toronto	Parkwoods	43.7588	-79.320197	Dollarama	43.757317	-79.312578	Discount Store

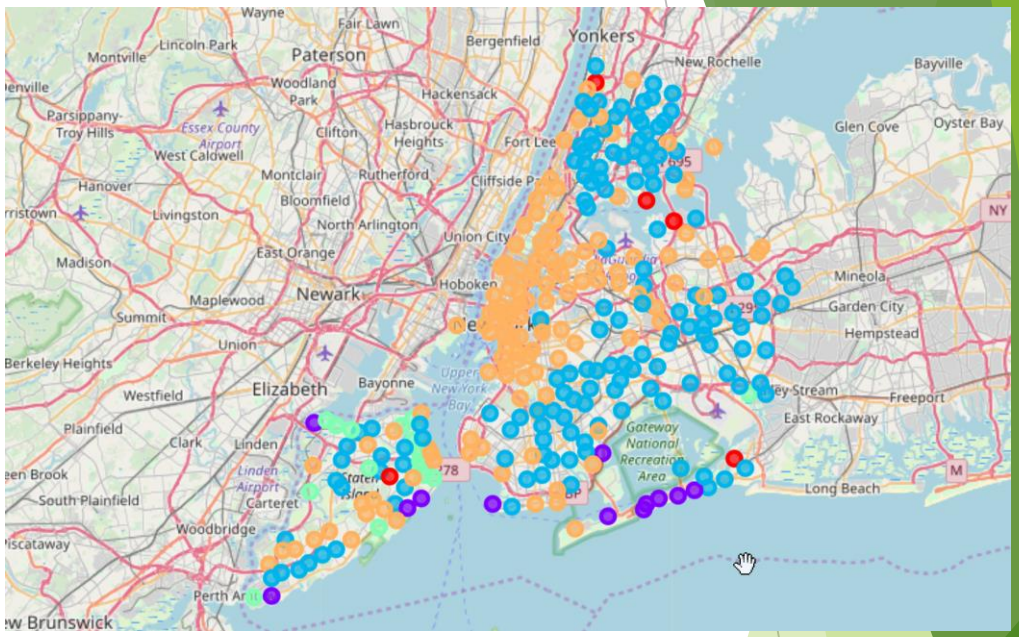
- Both datasets together have 26346 samples.
- New York has 462 different venue categories and Toronto has 353.
- Of these categories 322 are common to both cities. Others however exist only in one city, for example 'Czech Restaurant' is only in New York.
- We will see how this will affect the clustering of the datasets.

K-means Clustering: Introduction

- Clustering was first applied to the two cities separately as a first test.
- Then the two cities datasets were merged into one dataset and clustering was applied to this merged dataset.
- To apply clustering hot encoding had to be used on the “venue category” field, ending up with a column of 0 and 1 for each venue category.
- After hot encoding the mean of each field was calculated per each neighbourhood obtaining a measure of how much each venue category was represented in each neighbourhood.

K-means Clustering: Introduction

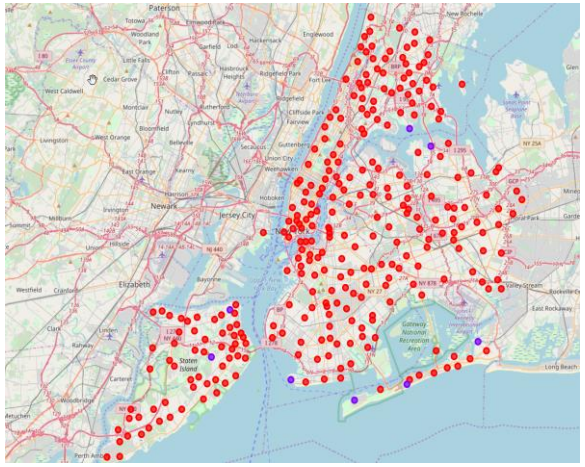


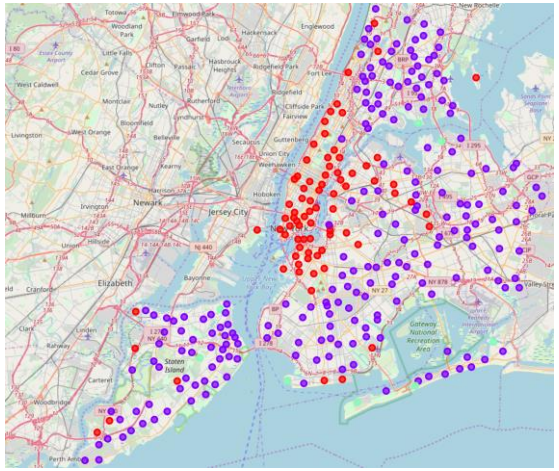


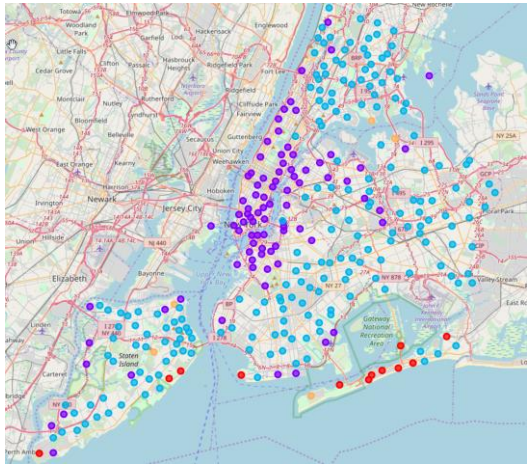
- Clustering was first applied to the two cities separately as a first test.
- We can notice in both cities a geographical correlation with the clusters with one category predominant in the city centre.

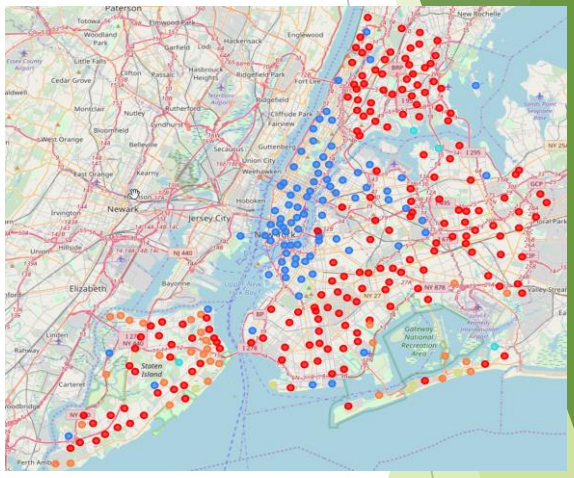
K-means Clustering: Both Cities

Clustering on both cities was undertaken using different values of k.









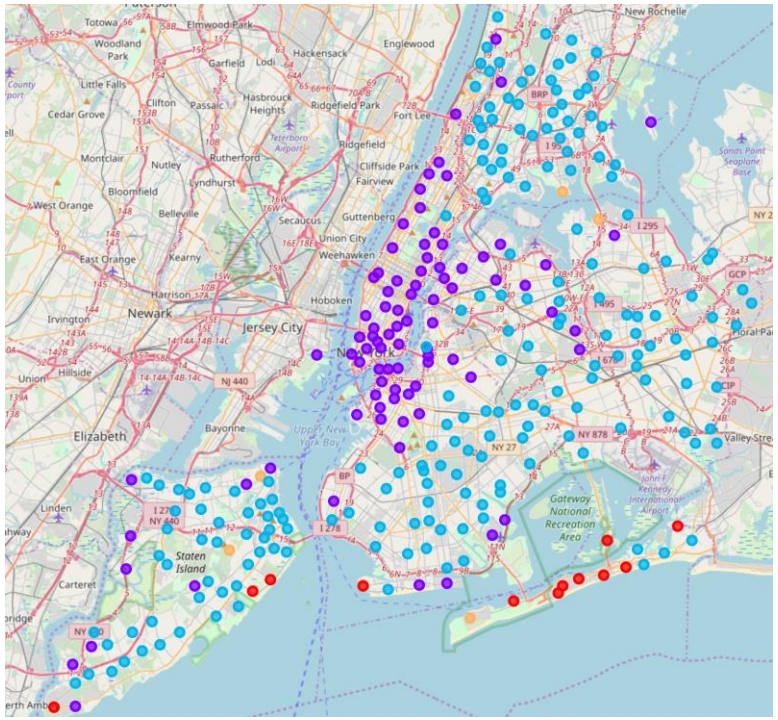
With k=2 we can see that the majority of location belong to one category only.

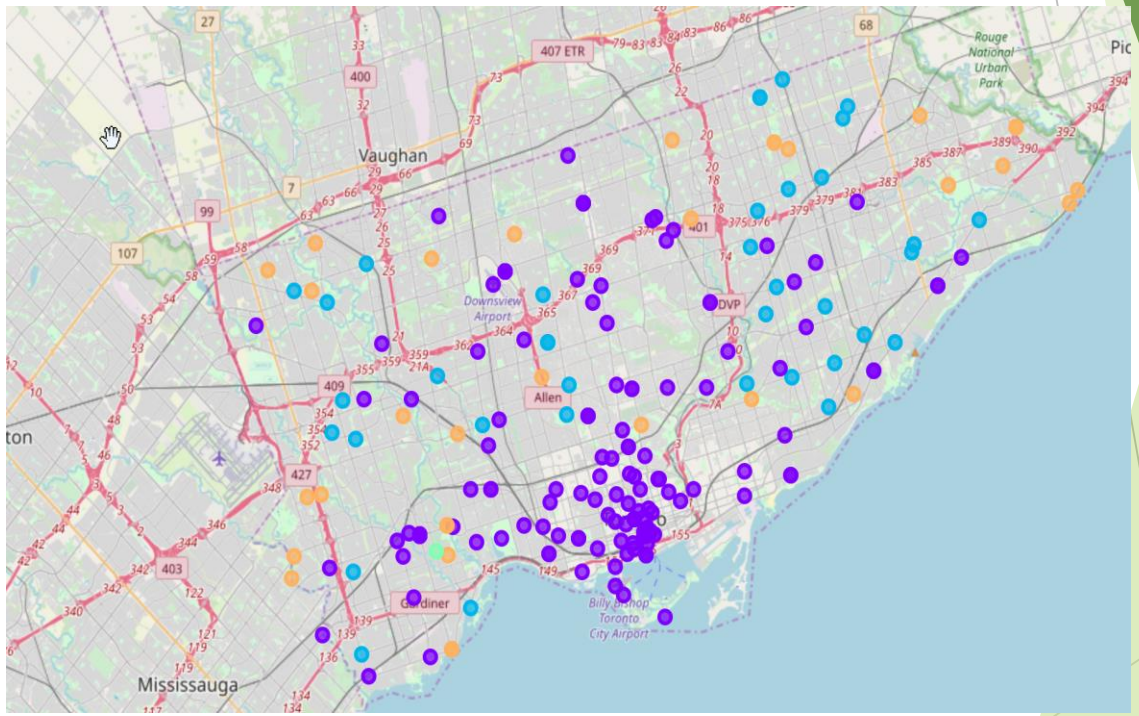
With k=3 we can see that the majority of neighbourhoods are now divided in two clusters.

With k=5 and k=7 few neighbourhoods change cluster while the majority still belong to one of two divisions that were identified already with k=3

K-means Clustering: Both Cities

Let’s focus on the results obtained using k=5.



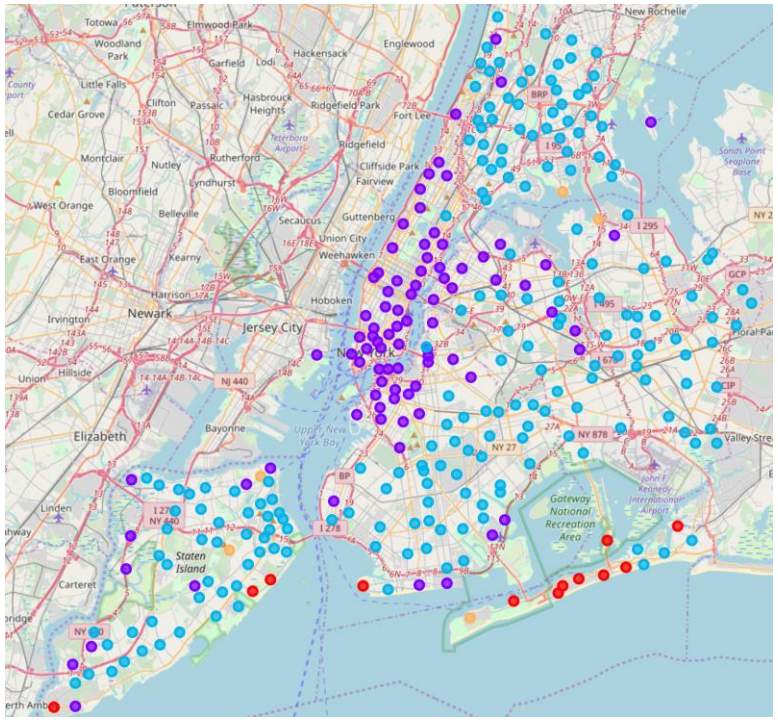


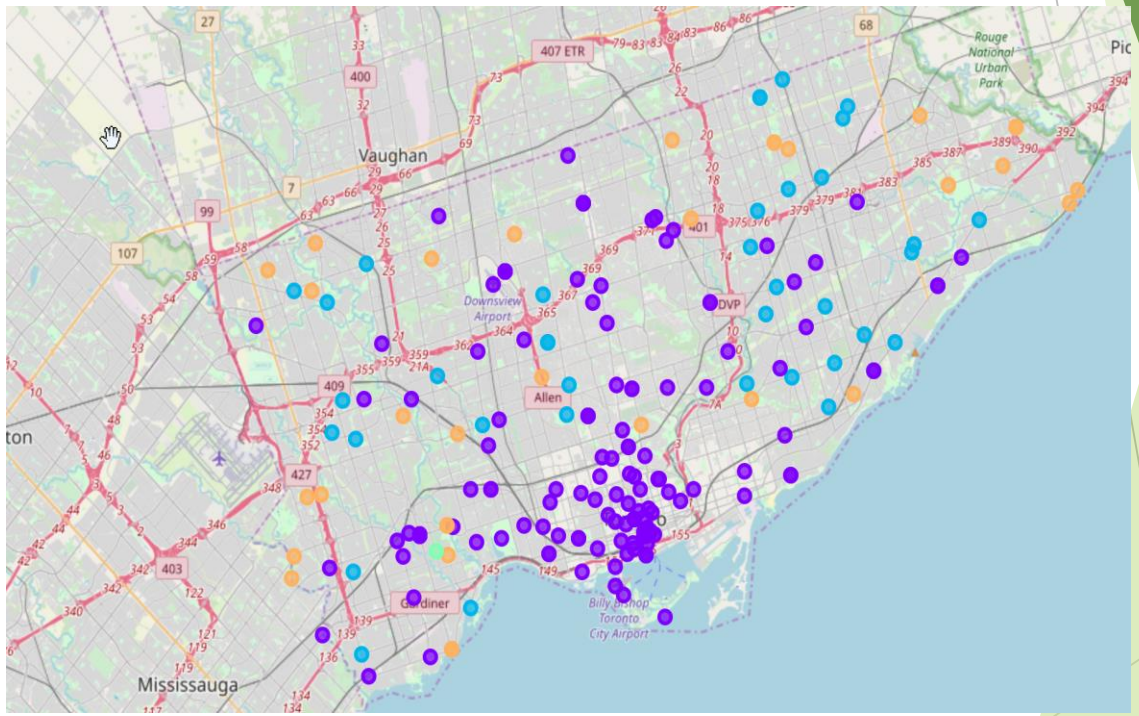
- Cluster 0
- Cluster 1
- Cluster 2
- Cluster 3
- Cluster 4

- Cluster 1 seem to be concentrated in center of the city (we easily see a correlation with the island of Manhattan in New York).
- This is an indication that clustering is working correctly.
- The geographical correlation is higher in New York than in Toronto. This can be an indication of real differences between the two cities in term of venues density and locations.

K-means Clustering: Both Cities

Let’s focus on the results obtained using k=5.





- Cluster 0
- Cluster 1
- Cluster 2
- Cluster 3
- Cluster 4

- To notice also that of the 5 clusters only 3 are in both cities whilst cluster 3 (green) is only in Toronto and cluster 0 (red) is only in New York. This does not surprise as we know there were venues categories that did not belong to both datasets.

K-means Clustering: Both Cities

If we focus on cluster 0 plotting the most common venues we notice that the neighbourhoods belonging to the cluster all have “beach” among the most common venues. This does not happen in Toronto and supports the validity of the cluster results.

	City	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
15	New York	0	Playground	Home Service	Indian Restaurant	Beach	Athletics & Sports
20	New York	0	Beach	Spa	Deli / Bodega	Trail	Pharmacy
32	New York	0	Deli / Bodega	Chinese Restaurant	Park	Beach	Metro Station
125	New York	0	Beach	Surf Spot	Burger Joint	Donut Shop	Taco Place
173	New York	0	Baseball Field	Beach	Chinese Restaurant	Dessert Shop	Bagel Shop
187	New York	0	Beach	Park	Pizza Place	Bus Stop	Zoo Exhibit
235	New York	0	Beach	Pizza Place	Deli / Bodega	Ice Cream Shop	BBQ Joint
236	New York	0	Beach	Pizza Place	Deli / Bodega	Donut Shop	Ice Cream Shop
241	New York	0	Baseball Field	Theater	Beach	Boat or Ferry	Irish Pub
245	New York	0	Beach	Supermarket	Home Service	Park	Spa
252	New York	0	Pier	Beach	American Restaurant	Playground	Athletics & Sports
271	New York	0	Italian Restaurant	Beach	Deli / Bodega	Chinese Restaurant	Ice Cream Shop

Comments and Conclusions

- In this presentation we have shown that clustering techniques can be used to measure similarity between neighboorhoods of different cities.
- This simple example was using only venues close to the centre of each neighboords (900m).
- A more accurate analysis should include all the venues belonging to each neighbours and not only, other metrics like house value and crime statistics for examples should be used.