# Identifying common patterns and difference between neighborhoods belonging to New York City and Toronto.

## 1 Introduction

This report show the result of an investigation aimed to investigate on how similar the cities of New York and Toronto are. The aim is to use Foursquare location data and clustering techniques to investigate on which are the most similar neighborhoods between the two cities and, if there are "unique" neighborhoods that do not have an equivalent neighbors in the other city. Another objective is to identify which characteristics are the most prominent ones in differentiating between neighborhoods. This can have many applications especially if it is possible to identify some basic categories on which to identify the neighborhoods of every city. For example, a family that has to move from one city to the other will be able which neighborhoods is similar to their original neighborhood.

## 2 Data

### 2.1 Data Introduction

Two datasets are involved for this investigation, one for each city. Each dataset contains the each city neighborhoods and its coordinates. Foursquare location has been used to extract the various venues in each neighborhood and their category. Once the venues categories have been extracted they can be used to study the differences among the various neighborhoods and the characteristics they have in common.

### 2.2 Data Cleaning and Acquisition

The Toronto Dataset was available on wikipedia. It contained each Postcode, Borough and Neighborhood of the city (Table 1).

| | 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 |

*Table 1: Toronto Dataset Original Format*

Geocode was used to find the latitude and longitude of each Neighborhood. Geolocator was, however, not able to identify some particular Neighborhood available in the table, for each of this entries we undertake a research using mainly google map to understand why the name was not recognized and to find the name of the neighborhood that geolocator would recognize. For each of these entries we had to slightly modify the names, for example "Hummond-Chadervale" had to be changed to simply "Chadervale" the resulting coordinates were double checked to make sure they were correct.

| | 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 |

*Table 2: Toronto Dataset after adding latitude and longitude*

The New York Dataset was available from a previous exercise of this course and was already in the format needed for this investigation.

| | 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 |

*Table 3: New York Dataset Example*

Borough and Postcode were not needed so they were dropped from the datasets. Figure 1 and 2 show the Position of the neighborhoods for each city.

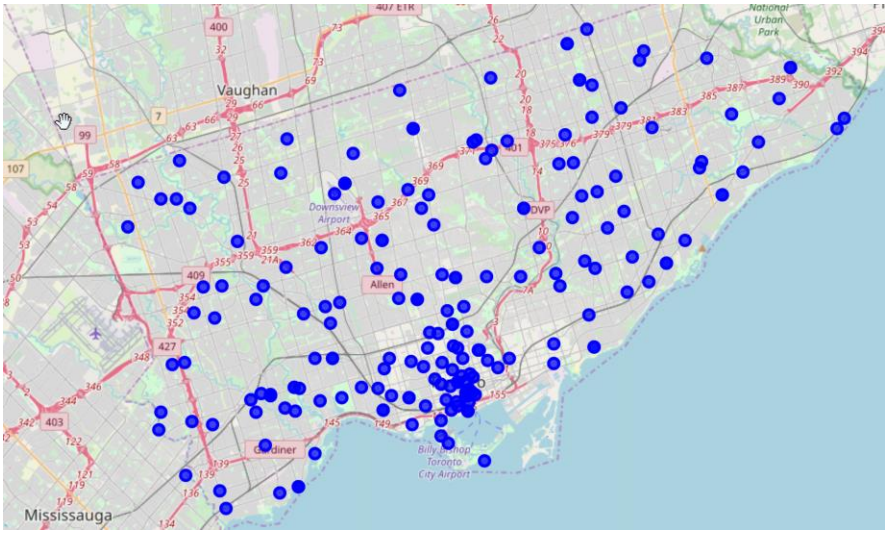


*Figure 1: Latitude and Longitude of Toronto neighborhoods*


*Figure 2: Latitude and Longitude of New York neighborhoods*

## 2.3 Venues Extraction

For each dataset we have used the Foursquare location service to identify venues belonging to each neighborhoods. We used a radius of 900 m around each location and we limited the venues to 500 for each location. The resulting datasets contained for each venue, the neighborhood to which it belonged the neighborhood coordinates, the venue name and category and the venue coordinates (example in table 4).

| | 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 |

*Table 4 : example of dataset after adding the venues.*

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.

# 3 Clustering Analysis

## 3.1 Introduction

Clustering analysis using k-means has been applied to both datasets first separately and then together. Clustering was based on the 'Venue Category' variable.

To apply clustering the datasets we had to use hot encoding so that to obtain equivalent datasets containing the venue category as first column and as many extra columns as the 'Venue Category' values. For each sample the value of a column would be 1 if the Venue belonged to that category or 0 if not. Grouping this dataset based on which neighbourhood the venue belonged to and calculating the mean of each category value gave us a measure on how each venue category was present in each neighbourhood and therefore a way of characterize the various neighbourhoods.

We have applied this analysis first to the separate datasets. In figure 3 we see the clustering result for the city of Toronto using 5 clusters.

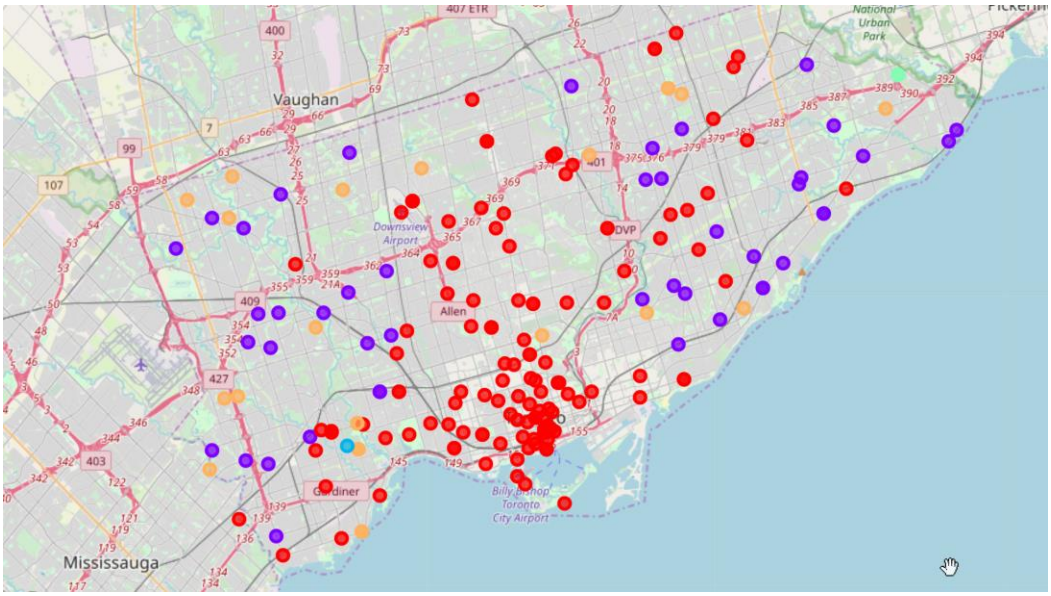


*Figure 3 : Clustering analysis result on Toronto datasets.*

We can see that red category is the most prevalent especially in the center of the city.

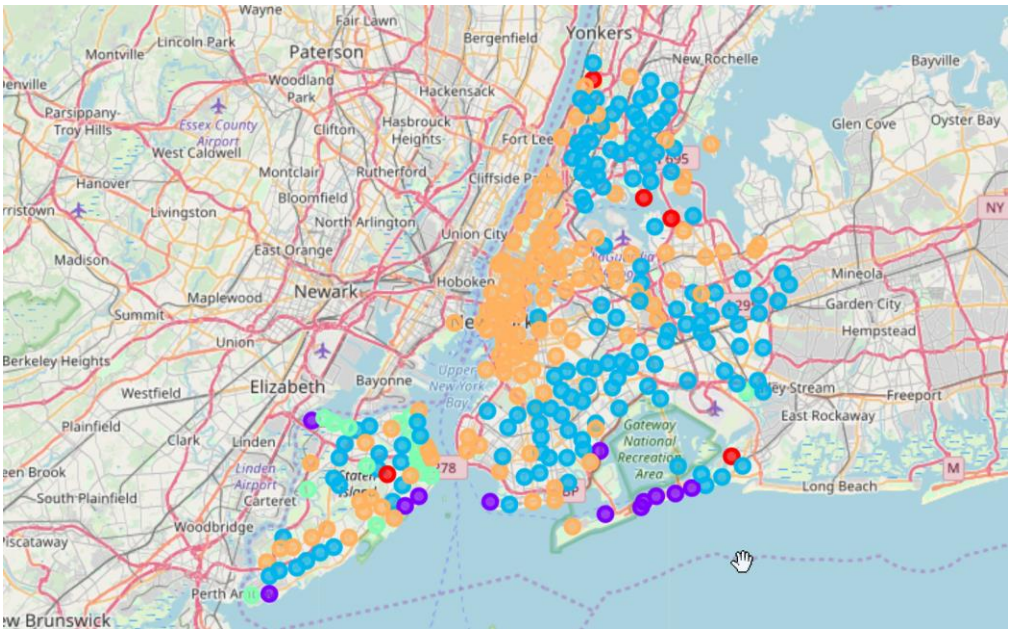


*Figure 4: Clustering analysis result on Toronto datasets.*

In figure 4 we have the clustering result for New York, to notice that since the clustering has been done separately the red category in figure 4 has nothing related to the red category in figure 3. On New York results we can we can see once again as one cluster (the orange) seems to have a geographical correlation.

### 3.2 Clustering analysis on both cities together

The two datasets have been merged in one single dataset and the same methodology described in session 3.1 has been applied. We have tested various values of k (Figure 5).

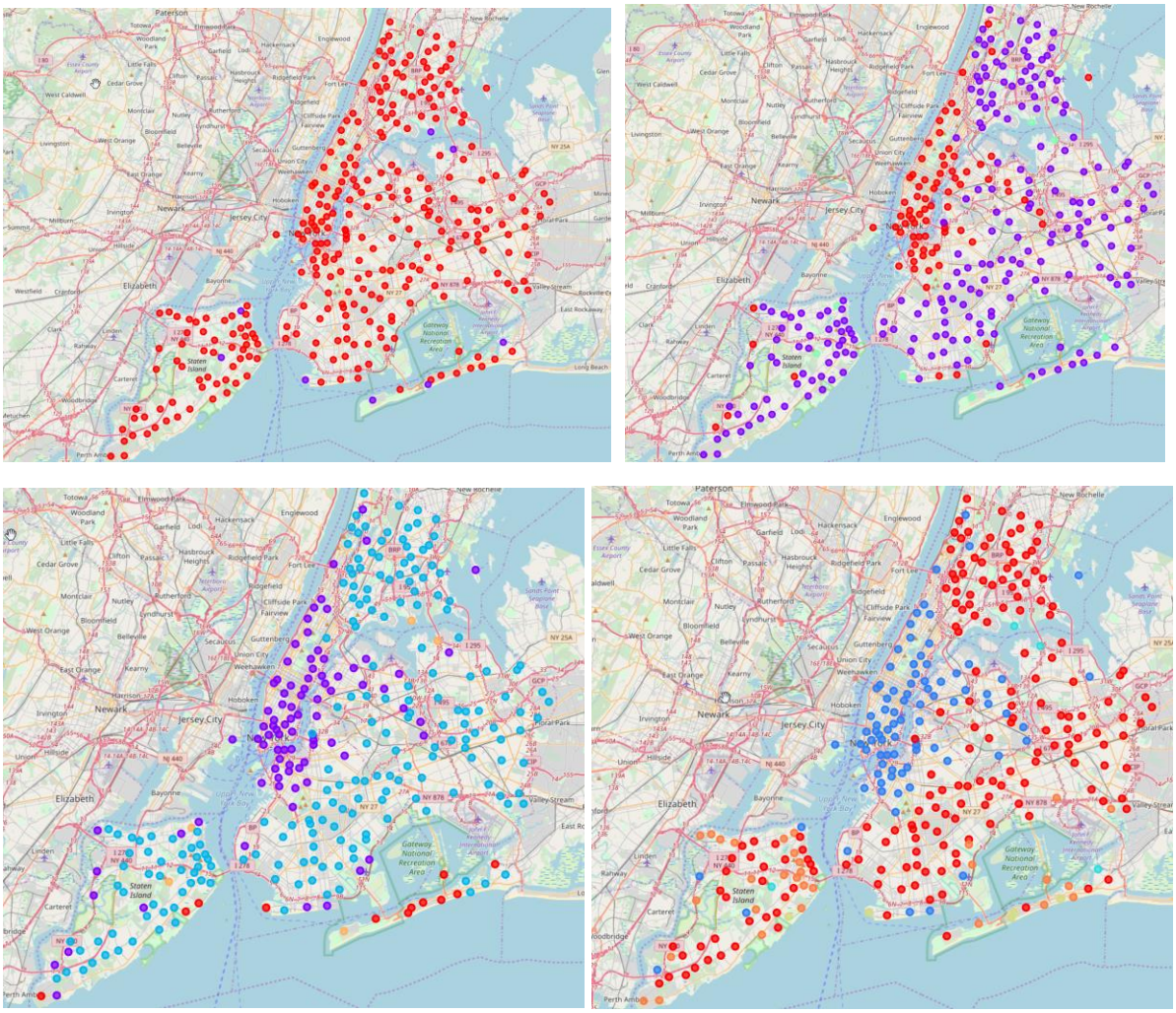

*Figure 5: Examples of clustering results using different values of k (2,3,5 and 7).*

After testing different values we have decided to focus on k=5. Looking at the results of clustering with a k value of 5 we can see in both figure 6 and 7 that 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

also an indication that clustering is working. The geographical correlation is higher in New York than in Toronto.

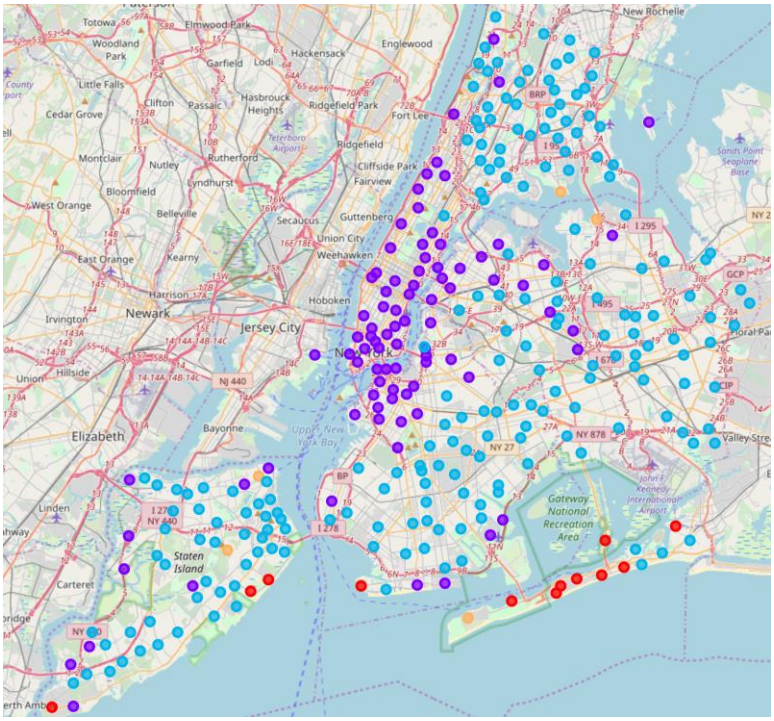


*Figure 6: Clustering results for the city of New York (k=5). Cluster 0=Red. Cluster 1= Violet. Cluster 2=Blue. Cluster 3=Green. Cluster 4=Orange.*

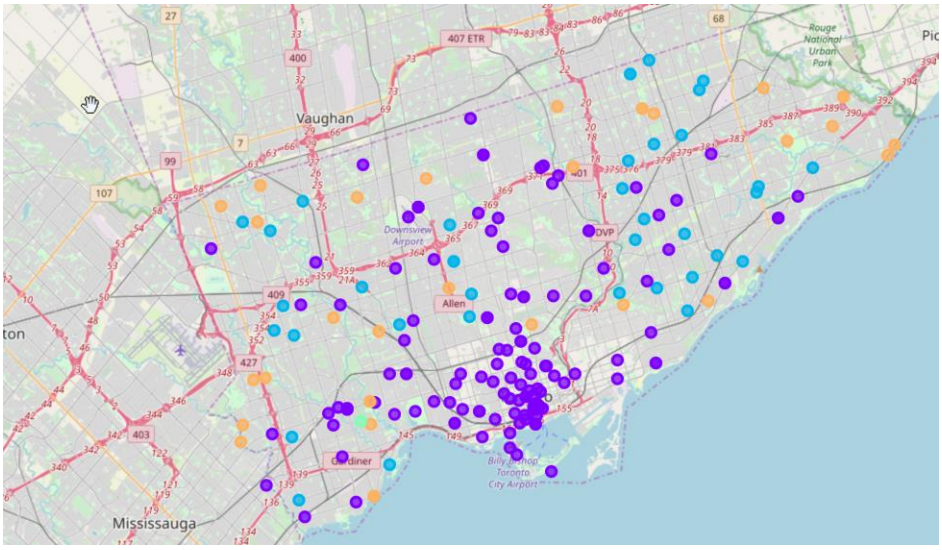


*Figure 7: Clustering results for the city of Toronto (k=5). Cluster 0=Red. Cluster 1= Violet. Cluster 2=Blue. Cluster 3=Green. Cluster 4=Orange.*

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. Focusing on cluster 0 we can see that these neighborhoods all have 'beach' as one of the most common venues (table 5), this does not happen in Toronto. This characteristic once again supports the validity of the clustering result.

| | 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 |

*Table 5: Neighborhoods belonging to Cluster 0 and their most common venues.*

## 4 Conclusions

We have build a dataset containing common venues for each neighbourhood of two cities, New York and Toronto. Using the venue category we have applied a clustering algorithm to identify neighbourhoods in one city that correlate with neighbourhoods in the other city. The resulting cluster show a geographical correlation. However we need to take into consideration that the dataset has been collecting the venues closer to the coordinate associated to each neighbourhood (ray of 900m) and not all the venues belonging to it.

A more complex clustering method will be useful in providing a high correlation map between the two cities.