# The Battle of Neighbourhoods

IBM Applied Data Science Capstone Project

## 1. Background

There are fewer foreign nationals living in the United Kingdom than citizens born in other countries. About 6.2 million citizens of non-British ethnicity were living in the United Kingdom in 2019 and 9.4 million were born abroad.

The United Kingdom's migrant population is concentrated in London. About 35% of people living in the UK who were born abroad live in the capital city. Likewise, about 37% of people living in London were born outside the United Kingdom, compared to 14% in the United Kingdom as a whole.

In 2019, Canada welcomed more than 300,000 immigrants, more than one in three immigrants has chosen to settle in the Greater Toronto Area (GTA). The GTA welcomed more immigrants (118,000 newcomers) than the four Atlantic provinces, Quebec, Manitoba, Saskatchewan, and Canada's three combined territories.

Given the fact that Toronto and London are far apart from each other, resemblances between these two cities can still be found.

## 2. Business problem

Relocation can be considered to be a major decision for a individual. In addition to job opportunities, environmental and cultural shocks are also a major concern for migrants as they move to another area. It is also beneficial for them to consider a neighbourhood comparable to the one they used to live in. In this project, we will introduce a machine learning technique to cluster Toronto and London areas in order to suggest neighbourhoods that are the best choice for migrants based on surrounding facilities such as school, hospital, shops, etc.

As a result, people who read in this article will be someone who wants to move from one city to another. This recommendation method will also help stakeholders interested in citing a business in a new city.

# 3. Data

We will use the following datasets for this project:

- Toronto.csv that consists of Toronto's postcodes boroughs, neighbourhoods.

  Data source: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

- London.csv that consists of London's postcodes, boroughs, neighbourhoods, post town.

  Data source: https://en.wikipedia.org/wiki/List_of_areas_of_London

| | PostalCode | Borough | Neighbourhood |
|---|---|---|---|
| 2 | M3A | North York | Parkwoods |
| 3 | M4A | North York | Victoria Village |
| 4 | M5A | Downtown Toronto | Regent Park / Harbourfront |
| 5 | M6A | North York | Lawrence Manor / Lawrence Heights |
| 6 | M7A | Downtown Toronto | Queen's Park / Ontario Provincial Government |

Fig. 1: Toronto Dataset

| | Neighbourhood | Borough | Postcode |
|---|---|---|---|
| 0 | Abbey Wood | Bexley, Greenwich | SE2 |
| 1 | Acton | Ealing, Hammersmith and Fulham | W3 |
| 2 | Acton | Ealing, Hammersmith and Fulham | W4 |
| 3 | Angel | Islington | EC1 |
| 4 | Angel | Islington | N1 |

Fig. 2: London Dataset

- A csv file that has the geographical coordinates of each postal code for neighbourhoods in Toronto is provided.

  Data source: http://cocl.us/Geospatial_data

- The Foursquare API search tool is used to obtain information on neighbourhoods' venues and can be used to identify and compare Toronto and London neighbourhoods.

# 4. Methodology

### 4.1. Data Cleaning

For the Toronto dataset, we will ignore any rows with a Borough that is Not assigned and if a row has a borough but a Not assigned Neighbourhood, then the Neighbourhood will be the same as the borough. After cleaning the Toronto dataset, we end up with 103 rows.

For the London dataset, any rows with the same postcode are combined into one with the Borough and Neighbourhood separated with a slash. After cleaning the London dataset, we end up with 175 rows.

### 4.2. Obtain Coordinates

To acquire the latitude and longitude coordinates of every neighbourhood in Toronto and London, we are going to use a csv file with the geographic coordinates of each neighbourhood postal code in Toronto that was provided on week 3 assignment. Neighbourhood geographical coordinates in London are not provided; therefore, we will use the Geopy Library and Nominatim API. Because the Geocoder package can be very unreliable, any duplicate coordinates will be dropped.

| | City | Borough | Neighbourhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | Toronto | North York | Parkwoods | 43.753259 | -79.329656 |
| 1 | Toronto | North York | Victoria Village | 43.725882 | -79.315572 |
| 2 | Toronto | Downtown Toronto | Regent Park / Harbourfront | 43.654260 | -79.360636 |
| 3 | Toronto | North York | Lawrence Manor / Lawrence Heights | 43.718518 | -79.464763 |
| 4 | Toronto | Downtown Toronto | Queen's Park / Ontario Provincial Government | 43.662301 | -79.389494 |

Fig. 3: Toronto dataset with coordinates

| | City | Borough | Neighbourhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | London | Westminster / Camden | Covent Garden / Charing Cross / Aldwych / St G... | 51.51651 | -0.11968 |
| 1 | London | Camden / Camden and Islington | Bloomsbury / Holborn / St Pancras / King's Cross | 51.52450 | -0.12273 |
| 2 | London | Westminster | Maida Vale / Little Venice | 51.52587 | -0.19526 |
| 3 | London | Kensington and Chelsea | Holland Park | 51.50162 | -0.19173 |
| 4 | London | Ealing | Hanwell | 51.50878 | -0.33630 |

Fig. 4: London dataset with coordinates

### 4.3. Data Visualization

Now that we have a list of boroughs, neighbourhoods and their respective geographic coordinates for Toronto and London, let's use the Folium Library to plot a map of all neighbourhoods in each city. The neighbourhoods that are part of the same borough and are drawn with the same colour.

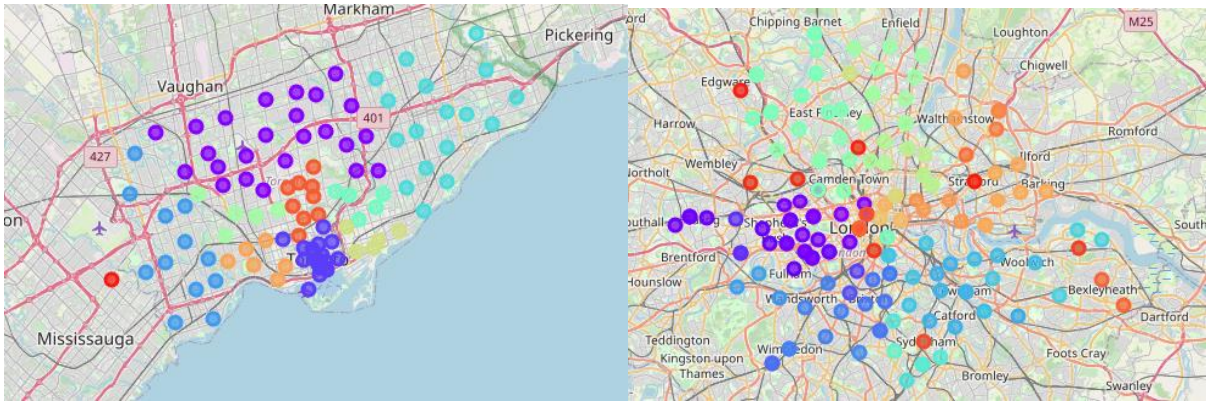



Fig. 5: Visualisation of Toronto and London

### 4.4. Foursquare API

Foursquare API offers access to vast databases for location data and venues information, including address, photos, tips, ratings, and comments. We will use the Foursquare API to locate nearby venues within 500 meters of each neighbourhood in Toronto and London. After defining the Foursquare credential and generating the API request URL, we will submit the HTTP request and receive the venues information in the json file.

| | City | Borough | Neighbourhood | Latitude | Longitude | Venue | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Toronto | North York | Parkwoods | 43.753259 | -79.329656 | Brookbanks Park | Park |
| 1 | Toronto | North York | Parkwoods | 43.753259 | -79.329656 | 649 Variety | Convenience Store |
| 2 | Toronto | North York | Parkwoods | 43.753259 | -79.329656 | Variety Store | Food & Drink Shop |
| 3 | Toronto | North York | Victoria Village | 43.725882 | -79.315572 | Victoria Village Arena | Hockey Arena |
| 4 | Toronto | North York | Victoria Village | 43.725882 | -79.315572 | Tim Hortons | Coffee Shop |

Fig. 6: Venue Names and Categories in Each Neighbourhoods in Toronto

| | City | Borough | Neighbourhood | Latitude | Longitude | Venue | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | London | Westminster / Camden | Covent Garden / Charing Cross / Aldwych / St G... | 51.51651 | -0.11968 | Scarfes Bar | Hotel Bar |
| 1 | London | Westminster / Camden | Covent Garden / Charing Cross / Aldwych / St G... | 51.51651 | -0.11968 | Rosewood London | Hotel |
| 2 | London | Westminster / Camden | Covent Garden / Charing Cross / Aldwych / St G... | 51.51651 | -0.11968 | The Hoxton Holborn | Hotel |
| 3 | London | Westminster / Camden | Covent Garden / Charing Cross / Aldwych / St G... | 51.51651 | -0.11968 | Sir John Soane's Museum | History Museum |
| 4 | London | Westminster / Camden | Covent Garden / Charing Cross / Aldwych / St G... | 51.51651 | -0.11968 | Lincoln's Inn Fields | Park |

Fig. 7: Venue Names and Categories in Each Neighbourhoods in London

After merging the Toronto and London venues dataset into a single dataframe, we will change all types of restaurants such as French Restaurant, Pizza place into Restaurant. We will then generate one-hot vectors for venue categories and group the dataset by neighbourhood and by taking the mean of the frequency of occurrence of each category. By looking at the grouped venues, we find that there are similar venues names; for example, ‘Bar' and 'Hotel bar' are basically in the same category ‘Bar’. We will clean up these columns for better analysis. Finally, we create a new dataframe and display the top 10 venues for each neighbourhood.

| | Neighbourhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Acton | Restaurant | Cafeteria | Pub | Grocery Store | Bakery | Bookstore | Supermarket | Park | Sandwich Place | Organic Grocery |
| 1 | Agincourt | Lounge | Restaurant | Breakfast Spot | Skating Rink | Clothing Store | Flea Market | Field | Film Studio | Fish & Chips Shop | Fish Market |
| 2 | Alderwood / Long Branch | Restaurant | Gym | Pub | Cafeteria | Pharmacy | Sandwich Place | Skating Rink | Pool | Fabric Shop | Farmers Market |
| 3 | Anerley / Penge | Supermarket | Hotel | Grocery Store | Restaurant | Convenience Store | Fish Market | Farmers Market | Field | Film Studio | Fish & Chips Shop |
| 4 | Angel / Hackney | Cafeteria | Restaurant | Bar | Pub | Breakfast Spot | Burger Joint | Yoga Studio | Costume Shop | Donut Shop | School |
| 5 | Archway / Upper Holloway | Restaurant | Cafeteria | Grocery Store | Pub | Sandwich Place | Farmers Market | Park | Gastropub | Hotel | Bar |

Fig. 8: Top 10 venues for each neighbourhood in Toronto and London

### 4.5. Machine Learning Models

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 centres or cluster centroid), serving as a prototype of the cluster.

We apply K-Means to segment and cluster all the neighbourhoods in Toronto and London based on the similarity of the venue types. We use the Elbow method to determine the right value of K. We run the clustering with different values of K and calculate the sum of squared errors (SSE). The right K is determined by the elbow point on the line chart.

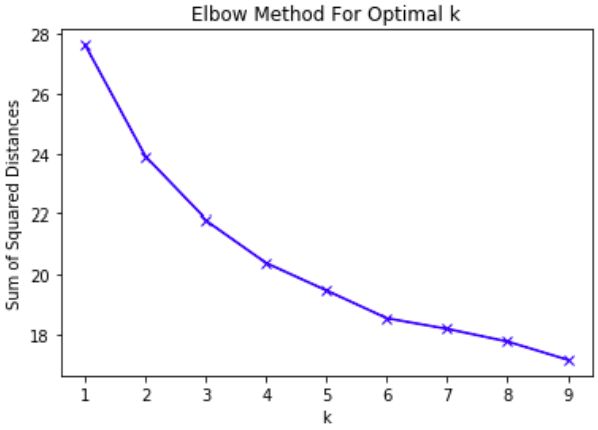


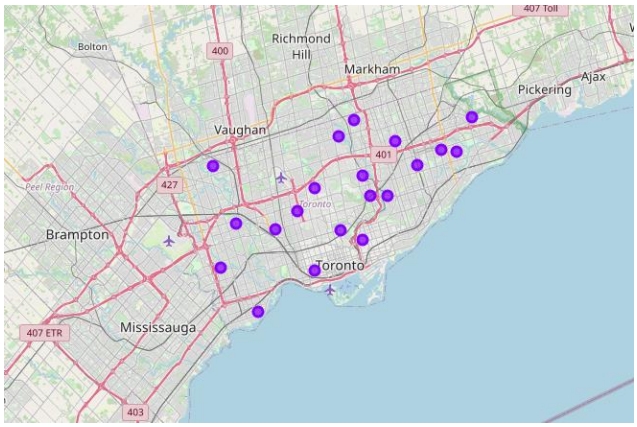

Fig. 9: Elbow Method For Optimal K

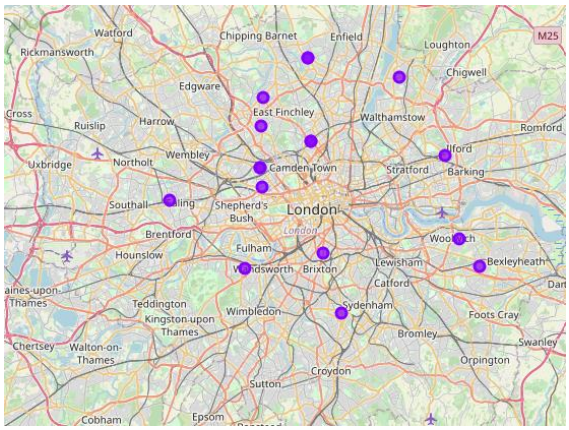
The optimal K is not clear, it could be 4 or 6. After testing the two values of K, when K=6 we get one to two clusters out of the six with one single neighbourhood which isn't useful in the scope of this project. In conclusion the value of K is 4.

## 5. Results

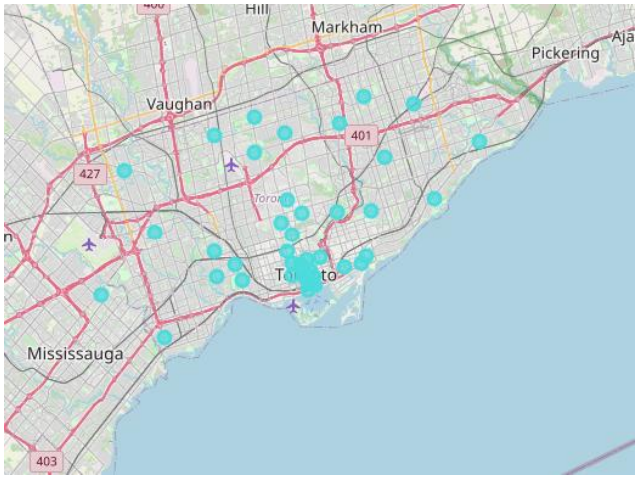
Folium maps of each cluster are created to facilitate the analysis. Similar neighbourhoods in Toronto and London are plotted with the same colour.

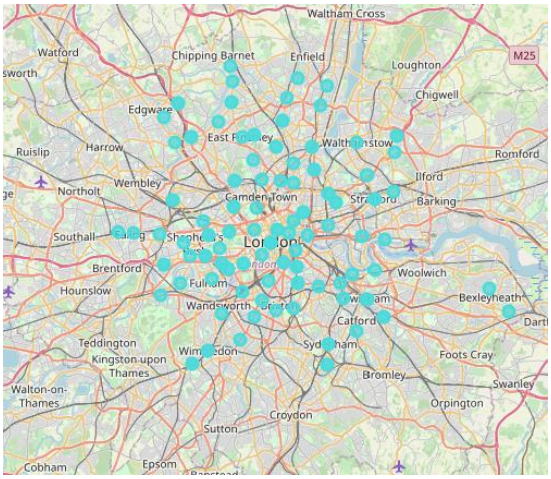
The number of neighbourhoods in cluster 1 is 37



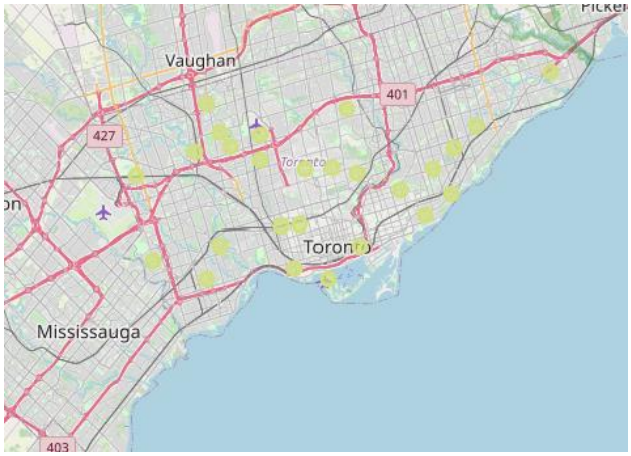



The number of neighbourhoods in cluster 2 is 170



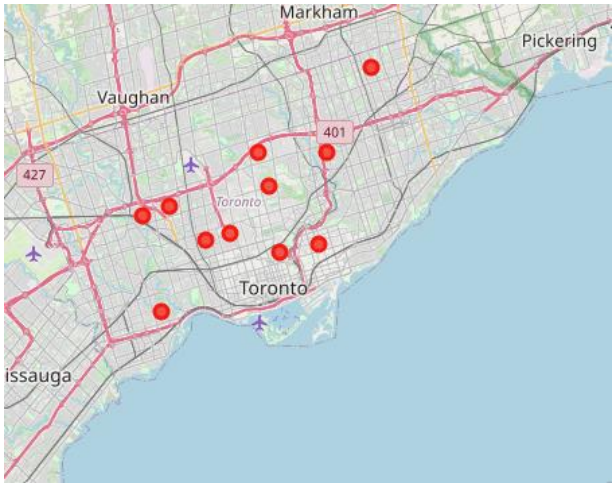




The number of neighbourhoods in cluster 3 is 55





The number of neighbourhoods in cluster 4 is 13




## 6. Discussion

From the results, we can conclude that for those who prefer to settle down in a residential area where surrounded by parks, grocery stores, pharmacy and restaurants, clusters 1 and 3 would be the best choice. While for those who prefer to live in a more crowded area where have access to a variety of venues, cluster 2 would be the best choice.

However, from the results, we notice that the majority of the neighbourhoods in lie into cluster 2 which represents 62% of the dataset. Also, cluster 4 represents less than 5% with only 2 neighbourhood from London. This problem is due to the unreliable locations provided by the Geocoder package, even though, we tried to address this issue by merging any duplicate postcodes. This was the best solution given the tools as we tested grouping and merging by neighbourhoods and boroughs, but the results did contain the greatest number of outliers.

We believe that this model will perform better if the location data for London is more accurate.

## 7. Conclusion

In this project, we built a clustering model to group the similar neighbourhoods in two big cities, Toronto and London, based on their nearby venues. This system can be applied to any other cities. Nowadays, People in the world are moving more frequently than before, we hope this analysis helps people to make better decisions depending on their needs when choosing the neighbourhood in the destination city. Furthermore, this research can be useful for real estate business if combined with Price Paid data, or city security solution if combined with crime data, etc.