# **Battle of the Neighborhoods - Capstone Project**

## **Introduction:**

It is common for people to relocate to a new city or country for work. However, moving is not an easy task. Careful considerations need to be made considering the available options and accordingly a decision needs to be taken based on the pros & cons.

As part of this project, the neighborhoods in the city of New York in the US & Toronto in Canada are analyzed and segmented into clusters. Neighborhoods classified in the same cluster are similar & have similar features. Features are defined based on the venues & the venue categories which are available within a specified neighborhood radius. This will be further elaborated in the methodology section. Neighborhoods classified in different clusters are dissimilar & have dissimilar features. No comparison has been made between similar and dissimilar cluster of neighborhoods to suggest which one is better. The intention is only to cluster the neighborhoods in both cities into segments based on their venue / venue categories.

The objective here is for anyone considering to move from New York to Toronto or vice versa to identify the cluster that his neighborhood belong to and then find a similar cluster in Toronto which will offer him the equivalent features of his existing neighborhood.

This allows the user to benefit from machine learning clustering algorithms & take informed decision fully leveraging on the available data on the neighborhoods in the two major cities of US & Canada.

## **Data**

In order to accomplish the objective, geo data for both cities is required. The data for the cities of new York and Toronto have been sourced from 2 different sources. The data for both sources is available publicly. The data sources are defined below.

1. Toronto:
   1. List of postal codes, boroughs and neighborhoods obtained from Wikipedia. URL: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>
   2. List of postal codes along with their corresponding geo data (latitude & longitude) is obtained using the URL: <https://cocl.us/Geospatial_data>
2. New York:
   1. List of boroughs, neighborhoods along with their corresponding geo data (latitude & longitude) is obtained using the URL: <https://cocl.us/new_york_dataset>

Once the data is obtained, the four square API has been used to procure the venue details such as name & category of venues available in the neighborhoods.

Based on the data & the required goal a machine learning based clustering algorithm (k-means) will be used to process the data. The geo data for the city of New York & Toronto will be loaded into a panda’s data frame. The data frame used will essentially have 4 columns *namely* ***boroughs, neighborhood, latitude and longitude***.

### **New York & Toronto Geo Data**

The data for the city of New York is available as a JSON object which can be formatted based on the required data set format and loaded in the data set directly. The Toronto geo data however would be managed differently. Unlike the New York date, Toronto’s latitude and longitudes and available based on postal codes. Another data source provides us the borough & neighborhood based on a corresponding postal code. Therefore the Toronto data needs to be preprocessed and transformed before it can be used. As part of the project the 2 sources of Toronto data sets have been uploaded in two different data sets and then merged based on the index column of postal code which is the key to merge both data sets. Once the data is merged the Toronto data set contains 5 columns namely ***postal code***, ***boroughs, neighborhood, latitude and longitude.*** The postal code column is then dropped and data for both the Toronto and New York data set is copied to a new data set for consolidation. The data is also cleaned before it’s used in the ML model. Boroughs without any description are removed so the data is not polluted which may lead to misleading results.

Once the data from both sources is pre-processed and transformed it can be used for the ML clustering analysis. The final data set has 4 columns and includes boroughs and neighborhoods from both cities of New York & Toronto. The final data set’s composition is as below;

|  |  |  |
| --- | --- | --- |
| Serial | Column Name | Type |
| 1 | Borough | Str |
| 2 | Neighborhood | Str |
| 3 | Latitude | Float |
| 4 | Longitude | Float |

In order to use the data in our machine learning based clustering algorithm we would need to explore the venues available in the neighborhood. This is to be done using the four square API. With the four square API our notebook will be connected using a four square developer account & the venues in the specified radius of the neighborhoods will be extracted for the full data set and then used by our ML based clustering algorithm.

How venues will be used for classification would be covered in the **METHODOLOGY** section.