# Final Report

# Capstone Project – The Battle of Neighborhoods Finding a Better Place for immigrants moving to Toronto, Canada

## 1. Introduction

Where to live in a new city is one of the most daunting tasks and depends at a big part, a matter of one's preferences and lifestyle. This project tries to create a content-based recommendation system for assisting immigrants in choosing a neighborhood based on the way they rank a number of lifestyle categories. For testing this system, we will use Toronto/Canada as an example city.

Toronto, is a multicultural city that continues to attract a big number of expats from different countries in the world. Its neighborhoods are constantly evolving, and boundaries can become blurred and disputed, so the pure geographically defined Neighborhoods are not the best benchmark for suitability to a new expat that doesnt know the city and help him choose where to live.

A recommendation system suggesting suitable neighborhoods for new expats, could be a value-added feature to existing recommendation platforms as Tripadvisor, Foursquare, Time Out etc. where the user can be guided on a suitable Neighborhood for him and relative venues can be suggested around this area basis his profile. This system could also have an application in real estate agency services where the agent can provide a more personalized recommendation on property to a client at a new city. Business owners, marketers and city designers would also benefit in redesigning their models around "lifestyle clusters" that can be created from developing this system rather than focusing purely on the standardized center/suburb layout.

## 2. Data Sources

For the purpose of this project we used data from two sources: We scrapped the Neighborhood/area data for Toronto from Wikipedia to create a list of Neighborhoods with geographical coordinates and then using the Foursquare (online venue recommendation platform) API we explored the venues listed in Toronto, which helped us create the lifestyle categories we needed.

**Links to the data**:
List of Neighborhoods organized by Postal Codes: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
List of Coordinates: https://cocl.us/Geospatial_data
For the venue categories: **Foursquare API** https://developer.foursquare.com/docs/api/venues/explore

## 3. Methodology

**Clustering Approach:**
To find the best neighborhoods, we decided to explore neighborhoods, segment them, and group them into clusters to find similar neighborhoods. To be able to do that, we need to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm.

**Using K-Means Clustering Approach**

### K-means Clustering

Preprocessing the data.

In [56]:
```
from sklearn.preprocessing import StandardScaler
X = Toronto_grouped_new.values[:,1:].astype(float)
X = np.nan_to_num(X)
cluster_dataset = StandardScaler().fit_transform(X)
cluster_dataset
```

Out[56]:
```
array([[ 0.39487793, -0.74915845, -0.52822975, ..., -0.63863091,
        -0.24607752,  5.32995767],
       [-1.05411213,  2.60901663,  0.40111789, ..., -0.63863091,
        -0.24607752, -0.26010574],
       [-0.22611781,  0.37023325, -0.52822975, ...,  0.54511916,
        -0.24607752, -0.26010574],
        ...,
       [ 0.87787462, -0.74915845, -0.52822975, ..., -0.63863091,
        -0.24607752, -0.26010574],
       [-1.05411213, -0.74915845,  0.40111789, ..., -0.63863091,
         1.64822465,  2.93421621],
       [-1.05411213, -0.74915845, -0.52822975, ..., -0.63863091,
        -0.24607752, -0.26010574]])
```

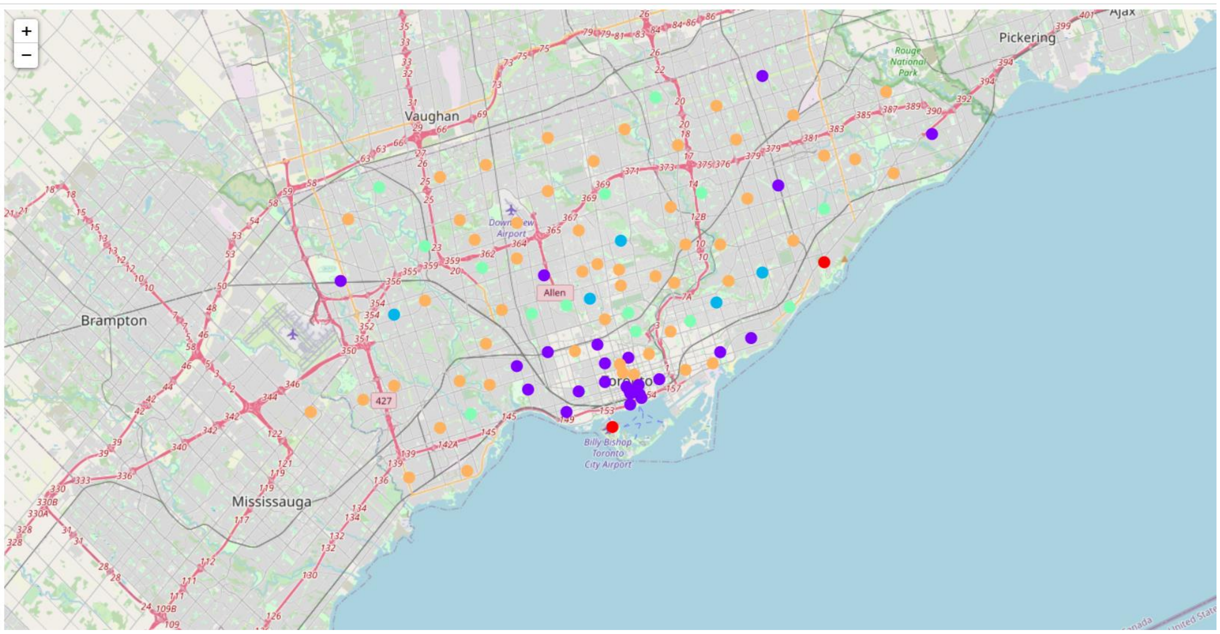
In [57]:
```
from sklearn.cluster import KMeans
num_clusters = 5
k_means = KMeans(init="k-means++", n_clusters=num_clusters, n_init=12)
k_means.fit(cluster_dataset)
labels = k_means.labels_

print(labels)
```

```
[4 4 4 4 4 1 3 1 4 0 3 4 4 4 4 1 4 0 1 4 4 4 4 1 4 1 3 4 4 1 2 4 1 2 4 1 1
```

**Map of Toronto:**

Out[66]:




## 4. Results

**K-Means Clustering:** The algorithm gave us 5 clusters according to the 12 categories created. The cluster with the most Neighborhoods was the one that included Downtown Toronto which was expected considering that the venues in a city are more condensed around its center. Kmeans did help identify area clusters based on the venue concentration and give us a first distinction on the areas as i.e where is the busiest locations and which ones are more of residential ones (i.e downtown vs suburbs).

**Recommendation System:** As mentioned we do not have previous user history of preferences or knowledge of the city of Toronto and its Neighborhoods to verify the suitability of the suggestions, so we are dealing with cold start users where it is not possible to make a comparative evaluation. To measure our results we cross-referenced the recommendations given with the top 5 venue dataset we created earlier in our process.

```
In [82]: compare=neighbourhoods_venues_sorted.loc[neighbourhoods_venues_sorted['Neighbourhood'].isin(result['Neighbourhood'])]
         compare
```

Out[82]:

| | Neighbourhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|---|
| 45 | Lawrence Manor, Lawrence Heights | Shopping | Personal_care | Health_Fitness | Restaurants | Leisure |
| 61 | Parkwoods | Shopping | Kids_Friendly | Leisure | Transportation | Food_markets |
| 62 | Queen's Park, Ontario Provincial Government | Shopping | Restaurants | Kids_Friendly | Fast_Food | Nightlife |
| 63 | Regent Park, Harbourfront | Shopping | Nightlife | Food_markets | Culture | Kids_Friendly |
| 85 | Victoria Village | Restaurants | Shopping | Fast_Food | Leisure | Transportation |

By mapping the recommended neighborhoods, we can see that out of the 5 recommended neighborhoods, 4 belong to the same cluster as calculated by K-means and 1 in a different one. A qualitative analysis of the recommended results is a more challenging task since to properly evaluate the suitability of the neighborhoods basis the user preferences there should be a sampling process where a number of Toronto residents would participate as users and evaluate the recommendations basis their knowledge of the neighborhoods.

## 5. Discussion

Having a tool to suggest a neighborhood for an expat about to relocate to a new country or city without prior knowledge of the area would reduce significantly the time and effort needed for research by limiting the options to a number of recommended areas. The system presented showed a good performance in choosing neighborhoods that score highest on user rankings however it lacks other qualitative characteristics.

## 6. Conclusion

Relocation is a big challenge for everyone. In this project we tried to visualize Toronto Neighborhoods as clusters created basis common features of a number of lifestyle categories and create a recommendation system for suggesting the top 5 Neighborhoods a new visitor/expat could select basis the importance each has to him. Such a system with necessary refinement and development as mentioned in the discussion section could be scaled to include all the major cities globally where a platform with records of user profiles and preferences could provide personalized recommendations for each user.

The more data were gathered the more the possibility to cluster the neighborhoods across major cities into areas that concentrate the interest of residents with specific lifestyle preferences, as i.e families, bachelors, foreign students, people into fitness etc. Such a platform could be of tremendous use to city designers, perspective regional business owners and marketers who could customize their products and services basis the “lifestyle” group of each Neighborhood and build their business model around smaller cluster centers rather than clutter the current urban centers.