# Final Report: The Battle of Neighborhoods

## Introduction

Toronto is the most populous city in Canada and the fourth most populous city in North America. It is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. Considered to be one of the most livable cities in the world, there are many beautiful parks in Toronto, which provide convenience for citizens.

This project is targeting on governments or any city construction agencies that have the power on construction of a new park. Suppose that the City of Toronto is preparing to build a new park as a place for residents' daily activities. The team leader believes that this park should be built in a place where there are fewer parks to facilitate the use of residents in that neighborhood. This project will predict the location of the new park from the perspective of data and algorithms, and provide suggestions for the location of the new park by visualizing the current parks in Toronto.

## Data Collection

The purpose of this project is to characterize the pattern of how parks in Toronto locate. To achieve this purpose, the following problems need to be solved using data and algorithms:

1. Find the information of each park in Toronto and locate them on the map
2. Utilize k-means algorithm to cluster and find the neighborhoods or areas with fewest parks, and choose that neighborhood to build the new park

### *Data Source*

The information of neighborhoods, boroughs, postal codes of Toronto can be extracted from this wikipedia webpage: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

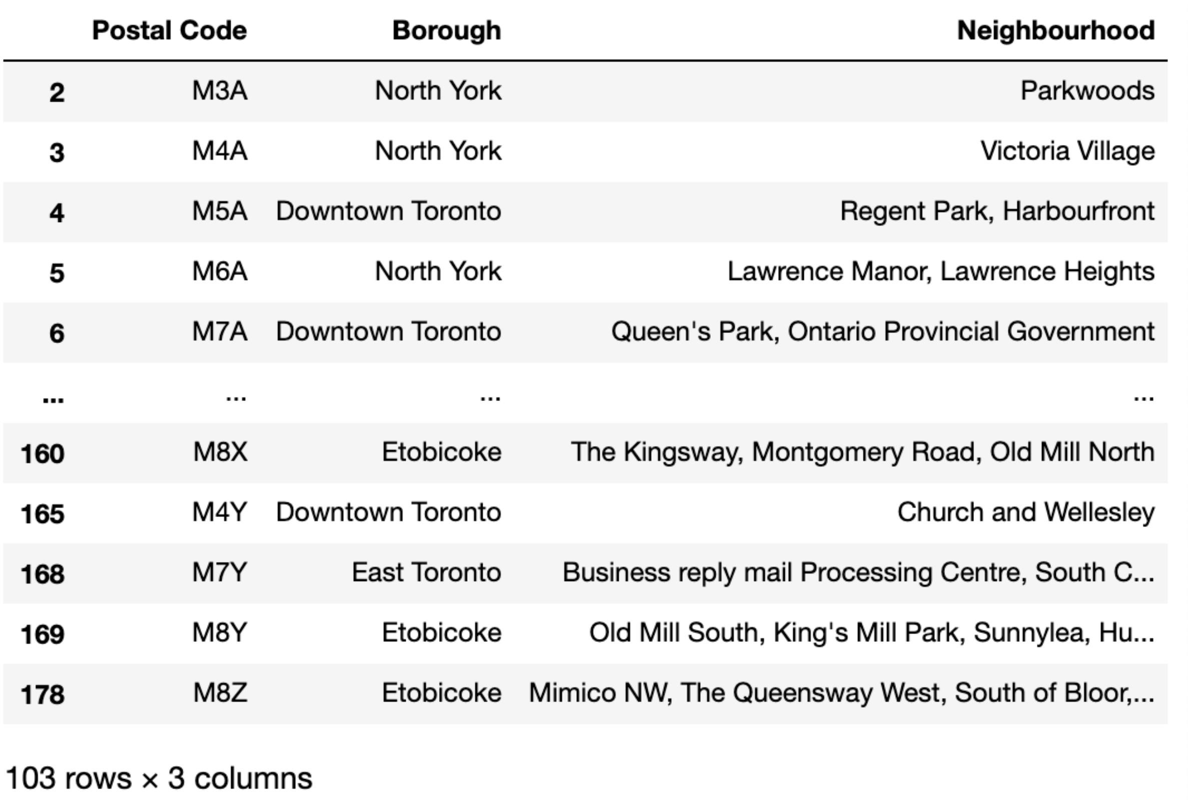
Moreover, we need data of latitude and longitude to pin down each venue in Toronto, which can be retrieved from this website: https://cocl.us/Geospatial_data. This website provides a csv file that can be directly imported into notebook.

Most importantly, the venue data can be acquired using Foursquare API, a platform that provides geographical data.

## Methodology

### *Web Scrapping*

Wikipedia webpage has provided a detailed dataset that includes Postal Code, Borough and corresponding Neighborhood of Toronto. Utilizing *requests* library will help scrapping data from the website. After we get the raw data, we need to use *pandas* package to clean the data, since the rows with Borough showing “Not Assigned” is not the data we need. After cleaning data, we can have 103 valid rows of data.

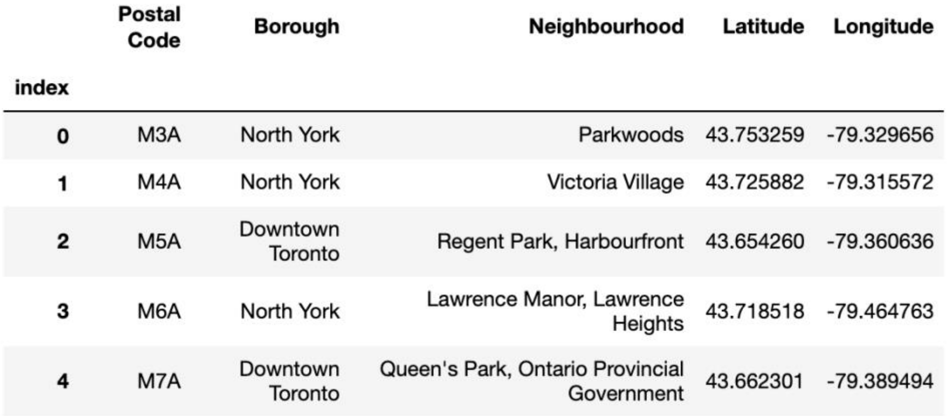


| | Postal Code | Borough | Neighbourhood |
|---|---|---|---|
| 2 | M3A | North York | Parkwoods |
| 3 | M4A | North York | Victoria Village |
| 4 | M5A | Downtown Toronto | Regent Park, Harbourfront |
| 5 | M6A | North York | Lawrence Manor, Lawrence Heights |
| 6 | M7A | Downtown Toronto | Queen's Park, Ontario Provincial Government |
| ... | ... | ... | ... |
| 160 | M8X | Etobicoke | The Kingsway, Montgomery Road, Old Mill North |
| 165 | M4Y | Downtown Toronto | Church and Wellesley |
| 168 | M7Y | East Toronto | Business reply mail Processing Centre, South C... |
| 169 | M8Y | Etobicoke | Old Mill South, King's Mill Park, Sunnylea, Hu... |
| 178 | M8Z | Etobicoke | Mimico NW, The Queensway West, South of Bloor,... |

103 rows × 3 columns

### *Latitude and Longitude Data*

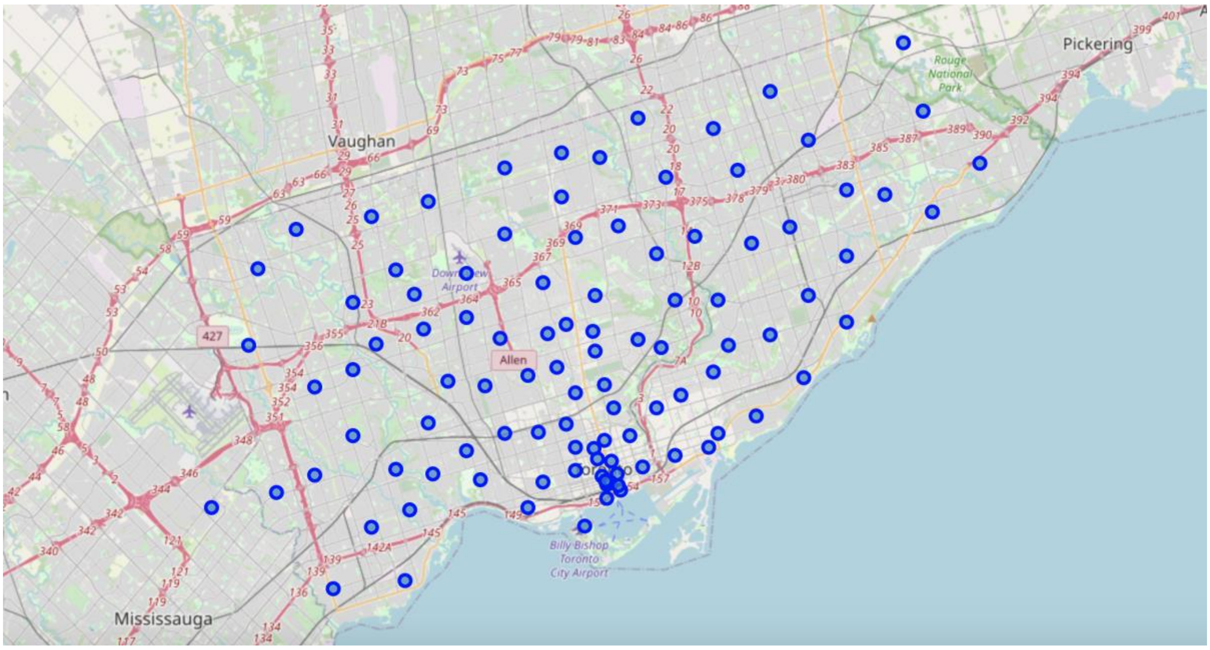
To further process the data, we need geographical data of each site, i.e. the latitude and longitude. The geographical data can be retrieved online (shown in Data Source section). By using *geocoder* library, we are able to append the geographical data to the data frame.



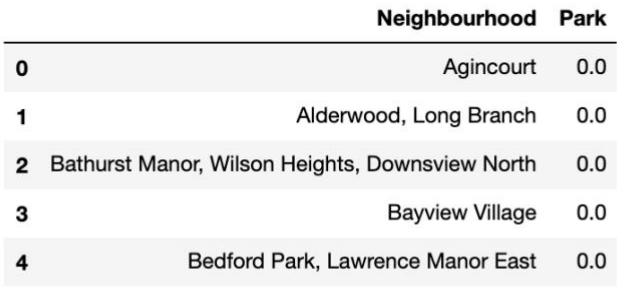
| | Postal Code | Borough | Neighbourhood | Latitude | Longitude |
|---|---|---|---|---|---|
| index | | | | | |
| 0 | M3A | North York | Parkwoods | 43.753259 | -79.329656 |
| 1 | M4A | North York | Victoria Village | 43.725882 | -79.315572 |
| 2 | M5A | Downtown Toronto | Regent Park, Harbourfront | 43.654260 | -79.360636 |
| 3 | M6A | North York | Lawrence Manor, Lawrence Heights | 43.718518 | -79.464763 |
| 4 | M7A | Downtown Toronto | Queen's Park, Ontario Provincial Government | 43.662301 | -79.389494 |

### *Data Preprocessing*

Our dataset includes 10 boroughs and 103 different postal code of Toronto. We use *geopy* library to get the latitude and longitude values of Toronto, and then create a map of Toronto with neighborhoods superimposed on top to have a quick look.



After initializing Foursquare API, we are able to retrieve data of specific location, i.e. parks in Toronto. With the geographical data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we can group rows by neighborhood and take the mean of occurrence of each category. Since we only focus on the data of parks, we create a new data frame, which is prepared for the subsequent modeling.



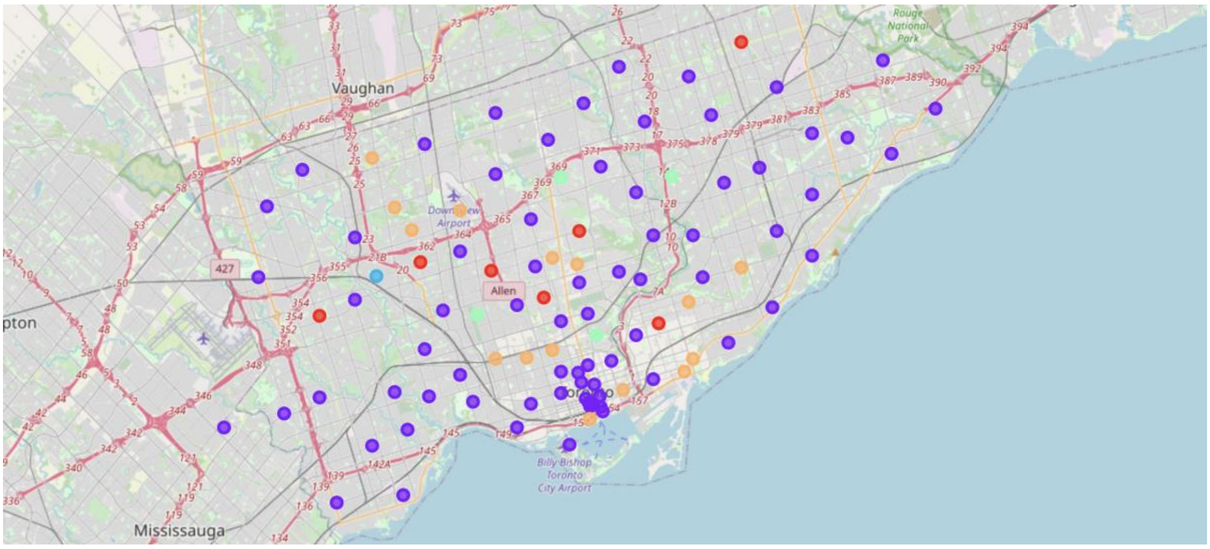
| | Neighbourhood | Park |
|---|---:|---:|
| **0** | Agincourt | 0.0 |
| **1** | Alderwood, Long Branch | 0.0 |
| **2** | Bathurst Manor, Wilson Heights, Downsview North | 0.0 |
| **3** | Bayview Village | 0.0 |
| **4** | Bedford Park, Lawrence Manor East | 0.0 |

### *Modeling*

In this case, we perform clustering algorithm on the data by using k-means clustering. K-means clustering identifies k number of centroids and allocates data points to the nearest cluster. It’s a very widely used unsupervised machine learning algorithm and can be applied to our case to characterize the pattern of park distribution in Toronto. We will cluster the neighborhoods in 5 based on the frequency of occurrences for parks.

## Results

The results of our k-means clustering show  that we can categorize the neighborhoods into 5 clusters:



- Cluster 0 (in red): 7 neighborhoods are in cluster 0, showing moderate number of parks
- Cluster 1 (in purple): 73 neighborhoods are in cluster 1, showing the highest number of parks
- Cluster 2 (in blue): only 1 neighborhood is in cluster 2, showing the lowest number of parks.
- Cluster 3 (in green): 4 neighborhoods in cluster 3, showing low number of parks.
- Cluster 4 (in orange): 15 neighborhoods are in cluster 4, showing high number of parks.

## Discussion

Based on the results of our modeling, we can observe that cluster 2 shows the lowest number of parks, and includes only 1 neighborhood, which is Weston. This represents this neighborhood can be a target area. If the government or other construction agencies have interests in building a new park, this site can be considered as a beneficial location.

## Conclusion

In this project, we explored the potential site for building a new park in Toronto. Suppose that the City of Toronto is preparing to build a new park as a place for residents' daily activities. This project predicted the location of the new park from the perspective of data and algorithms, and provide suggestions for the location of the new park by visualizing the current parks in Toronto. By applying k-means clustering algorithm, we identified a neighborhood with lowest number of parks – Weston, which can be considered as a suitable and potential site for the construction.