# IBM DATA SCIENCE

## *Capstone Project – The Battle of Neighborhoods*

## *Finding a suitable Neighborhood for Senior Citizens to live in Toronto, Canada.*

## Introduction

This Capstone project is about the senior citizens comfortness in terms of neighborhoods. The idea behind this project is that there are many neighborhoods around the city, which has their own pros and cons. Keeping in consideration only the senior citizens, this project has tried to find the best neighborhoods to live. Every person wants to retire in a place where they will be comfortable, and the basic requirements are in their vicinity.

## Business Problem

The objective of this project is to find out a suitable neighborhood for senior citizens to live in Toronto. By using data science methods and machine learning methods such as clustering, this project aims to provide solution to answer the business question: In Toronto, "WHICH IS THE BEST NEIGHBORHOOD FOR SENIOR CITIZENS TO LIVE IN TORONTO.

## Target Audience

The Real Estate Agents, Developers and Senior Citizens, who need to decide a neighborhood for their clients and themselves, respectively.

## Data

To solve this problem, the required data are:

- List of neighborhoods in Toronto, Canada.
- Latitude and Longitude of these neighborhoods.
- Venue category data related to pharmacy, park, grocery store, medical center and bank.

This will help us find the neighborhoods.

## Extracting the data

- Scrapping of Toronto neighborhoods via Wikipedia.
- Getting Latitude and Longitude data of these neighborhoods via Geocoder package.
- Using Foursquare API to get venue category data related to these neighborhoods.

## Methodology

First, to get the list of neighborhoods in Toronto, Canada the data is extracted from the Wikipedia page: ("https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M")

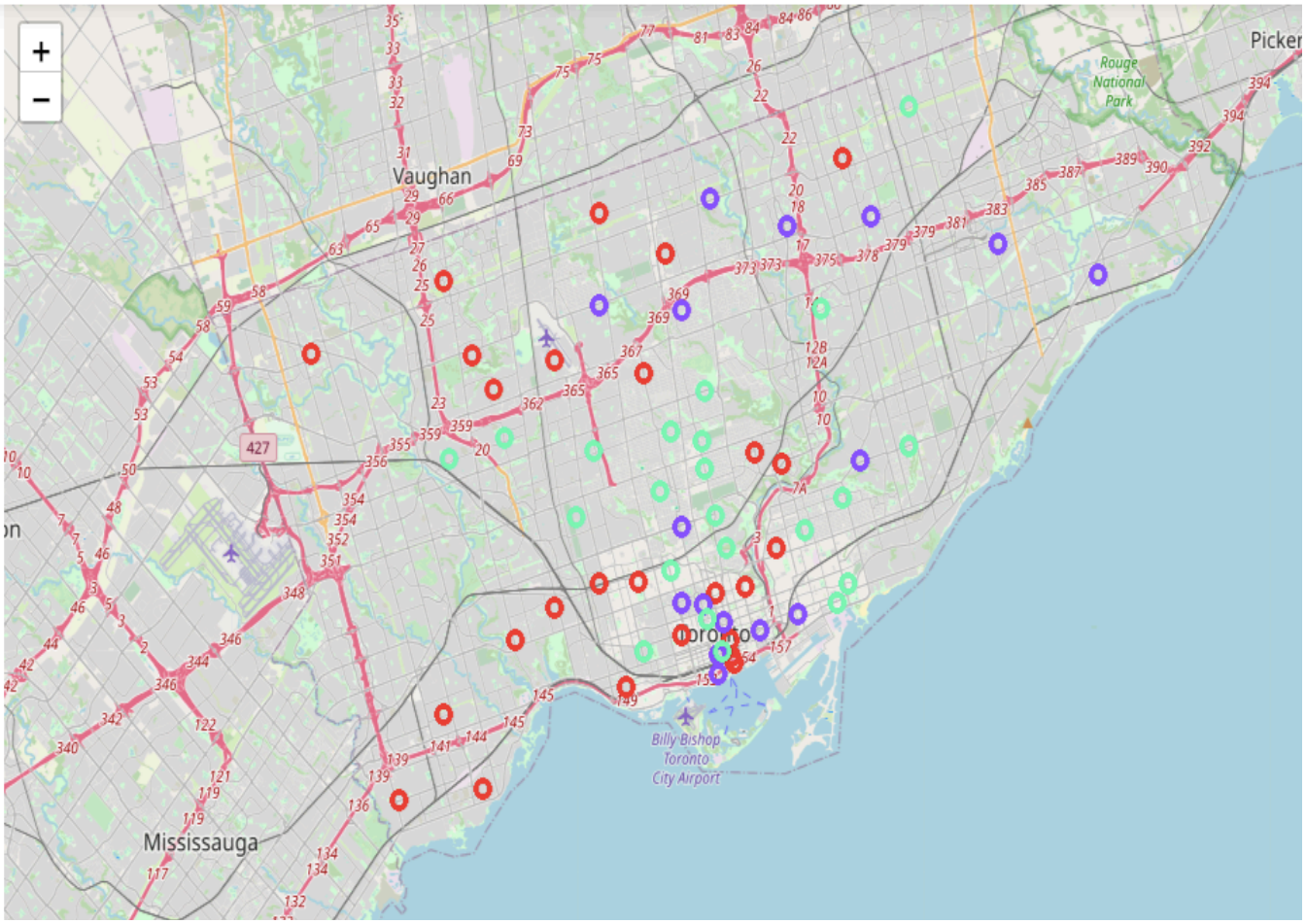
The data is then scrapped by utilizing pandas html table using scraping method, as it is easier and more convenient to pull tabular data directly from a web page into data frame.

However, it is only a list of neighborhood names and postal codes. Foursquare is used to pull the list of venues near these neighborhoods. To get the coordinates, The coordinates are retrieved from http://cocl.us/Geospatial_data.

After gathering all these coordinates the Foursquare API is used to pull the list of top 100 Venue within 500 meters radius. A Foursquare developer account is needed in order to obtain ClientID and API key to pull the data. From Foursquare the names, categories, latitude and longitude of the venues are extracted. The extracted data related to pharmacy, park, grocery store, medical center and bank categories, are considered as the basic requirements for the Senior Citizens to live. Each neighborhood is analyzed by grouping the rows by neighborhood and taking the occurrence of each venue category.

At last, the clustering method, k-means clustering is performed. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. The neighborhoods in Toronto are Clustered into 3 clusters based on the occurrence of the venue categories. Based on the results (the concentration of clusters), this project is able to recommend the ideal location for senior citizens to reside.

# Results




The results from k-means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on the categories in each neighborhood:

Cluster 0: Most of the requirements for senior citizens are met in Cluster 0 which has 22 Neighborhoods.

Cluster 1: All the neighborhoods (16) in Cluster 1 has Bank with few other requirements.

Cluster 2: all the neighborhoods (22) in Cluster 2 has Parks with few other requirements.

The results are visualized in the above map with Cluster 0 in red color, Cluster 1 in purple color and Cluster 2 in light green color.

## Recommendations

Most of the requirements for senior citizens are met in Cluster 0 which has 22 Neighborhoods. All the neighborhoods (16) in Cluster 1 has Bank and all the neighborhoods (22) in Cluster 2 has Parks. Looking at nearby venues, it seems Cluster 0 might be a good location as most of the requirements are met in these neighborhoods. Therefore, this project recommends the neighborhoods suitable for the senior citizens to live in Toronto. The Real Estate brokers, Agents and developers can look into the report and see and determine which neighborhood is suitable for their clients in Toronto.

## Limitations and Suggestions for Future Research

This project has taken into consideration 5 categories i.e. pharmacy, park, grocery store, medical center and bank. There are many other factors that can be taken into consideration such as entertainment, crime rate, housing prices, that could influence the decision of a neighborhood. However, to put all these data into this project is not possible to do within a short time frame for this capstone project. Future research can take into consideration of these factors.

## Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing the machine learning by utilizing k-means clustering and providing recommendation to the stakeholder.

## References

List of neighborhoods in Toronto: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Foursquare Developer Documentation: https://developer.foursquare.com/docs

All codes for this project can be found here: https://github.com/sarawgiabhilasha/Coursera_Capstone/blob/master/IBM%20final%20capstone%20project%202020.ipynb