The Battle of Neighborhoods

Finding a Suitable Neighborhood

NAME: K SAI RAKESH

1 .Introduction

In this project unsupervised machine learning K-means clustering algorithm is used for analysis. Unsupervised learning is a part of machine learning where no response variable is present to provide guidelines in the learning process and data is analyzed by algorithms itself to identify the trends. K-means clustering algorithm is one of the most popular and simplest algorithms. Collection of data points aggregated together because of certain similarities is referred to as clusters. The target number k refers to the number of centroids required in the dataset. A centroid is the imaginary or real location representing the center of the cluster. Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares. While keeping the centroids as small as possible, K-means algorithm identifies k number of centroids, and then allocates every data point present to the nearest cluster. The 'means' in the K-means refers to averaging of the data which is finding the centroid. Unsupervised K-means clustering is an extensively used algorithm for data cluster analysis.

2. Problem description

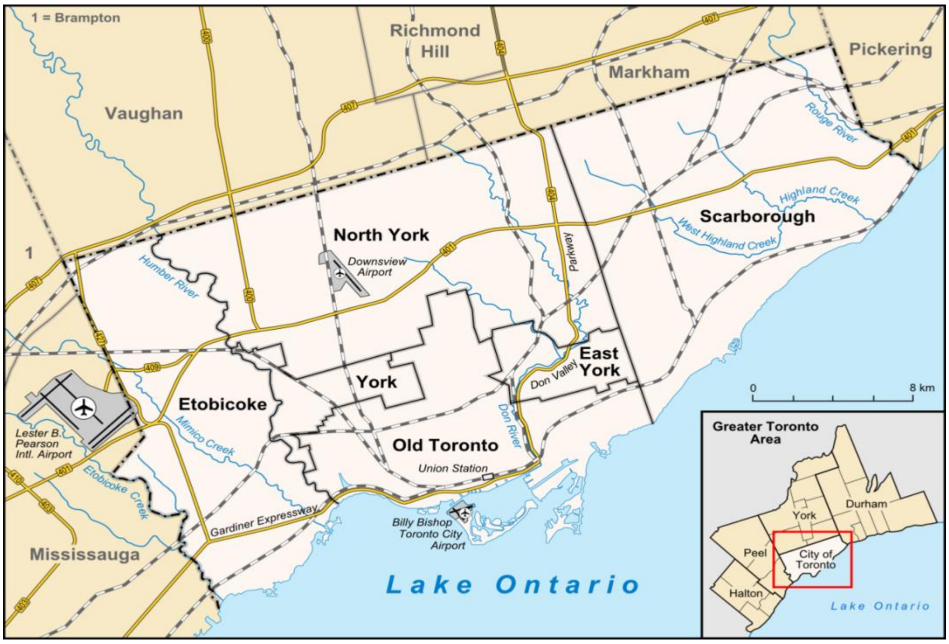
To understand the similarities and differences of neighborhoods between Queens borough in New York City and Scarborough borough in Toronto, also selecting the best and suitable neighborhood for a X company to open a new branch based on the different types of venues in the two places considering the quality of life and also the optimum living standards for its employees assuming the other conditions for opening a new branch of the company are met in the places under study.

3. Objective of the study

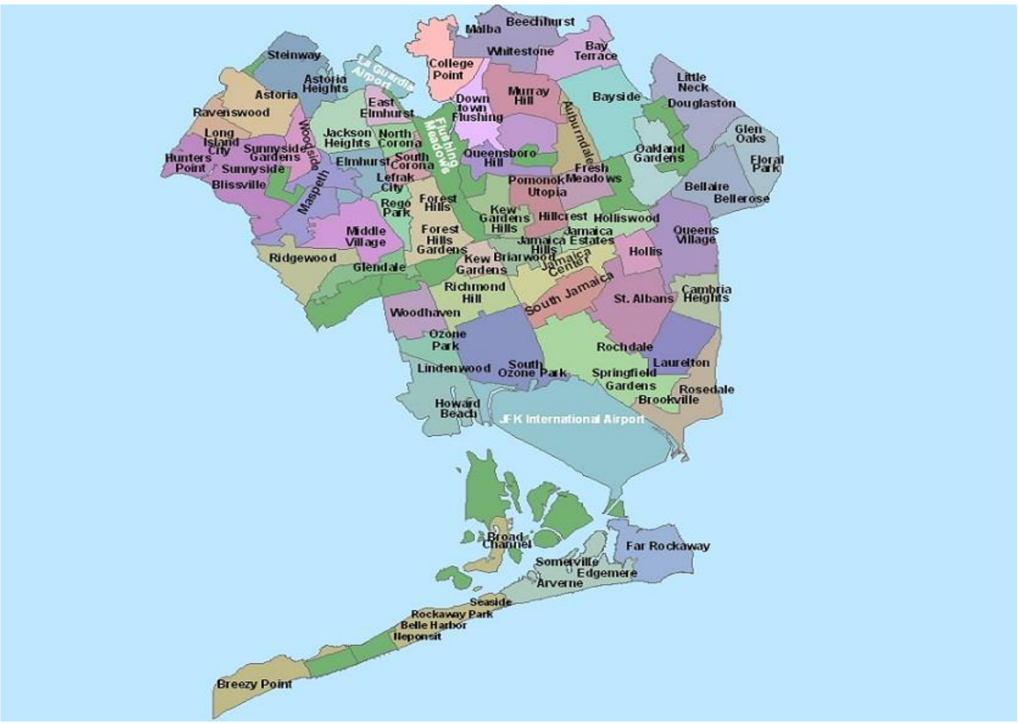
Using Foursquare collect the top venues of the neighbourhoods under study and then use the unsupervised k-means clustering algorithm to categorise the venues of the neighbourhoods under study. Obtain insights into the neighbourhoods by identifying the similarities and differences between Scarborough and Queens neighbourhoods.

Locations under study:

Scarborough borough in Toronto



Queens borough in New York City



4. About the Data:

The data which is used in the project is web-scrapped from Wikipedia websites pages of respective cities, Foursquare API and various python packages were used for creating maps and also machine learning algorithms to further analyse the problem.

The following datasets were acquired from these Websites:

>Neighbourhoods of Toronto:

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

>Latitude and Longitude data of Toronto:

http://cocl.us/Geospatial_data.

> Neighbourhoods of New York City:

https://geo.nyu.edu/catalog/nyu_2451_34572.

> Latitude and Longitude data of New York City:

Python Geolibrar

5. Methodology

> The data is web-scraped from various websites , the data further gets cleaned and processed into dataframes.

> After sorting the data the Foursquare API search feature would be used to locate and collect the places of the neigborhoods under study.

> The Python visualization library such as Folium would be used to visualize the neighbourhood clusters over a map.

> Unsupervised machine learning K-means clustering algorithm would be used to form different clusters of groups of places located in and around the neighborhoods under study.

> The derived clusters from both the neighborhoods will be further analysed individually for drawing out required conclusions.

Python packages used:

Pandas - (Library for data analysis),**Sklearn** - (Machine Learning library), **NumPy**- (Library used to work with arrays), **JSON** – (Used to handle JSON files), **Geopy** – (For retrieving data locations) ,**Requests** – (For handling https requests), **Folium** – (Map rendering library), **Mathplotlib** – (Library for plotting)