Applied Data Science

Capstone Project - The Battle of Neighborhoods (Week 2)

By Michael Oduor Opondo

Toronto Vs New York Neighborhood Types

Table of Contents

Applied Data Science....1
Capstone Project - The Battle of Neighborhoods (Week 2)....1
Toronto Vs New York Neighborhood Types....1
Introduction/Business Problem section:....1
Data Description:....2
Use of the Foursquare API:....2
Python Libraries & Models to be used:....2
K-Means Clustering Approach:....3
Methodology & Workflow:....3
Results:....6
Conclusion:....6

Introduction/Business Problem section:

This project was manly to assist people intending to or already exploring the best of the two major cities between New York City and Toronto, based on facilities available in their neighborhoods, using location data to make better and informed decisions on selecting the best neighborhoods. This is informed by the fact that such people are moving to and from different locations and would like to make decisions on whether to settle in Toronto or New York and have the need to explore and research for great locations to settle their families, based on factors including best schools locations, hospitals, malls, amongst other amenities.

The aim of this project is hence to develop an analysis of main determining features features for migrations to one of the either major cities of Toronto, Ontario (Canada) or New York City (USA), creating a better awareness of the neighborhood amenities for the end user, before moving.

Achieving this purpose necessitates the comparison of the neighborhoods of the two major cities and determine how similar or dissimilar they are to each other.

What would be compared in the two major cities include but not limited to the neighborhood types (which of the two is well defined and uniform.

Toronto remains a popular migration destination in Canada, located in the province of Ontario in Canada. Having attracted different groups from different walks of life, Toronto is home to diversity and multicultural nature of population make up. The main Neighborhood of interest in Toronto is Scarborough.

New York City on the other hand also represents a diverse and multicultural make up and would be easily assumed to be similar to Toronto, but dissimilarities exist within the two major cities hence the need for this comparison. The main neighborhood of interest in New York is Manhattan.

Data Description:

The crux of this project is based on the analysis of the boroughs and neighborhoods in both Toronto and New York Cities, to help properly segment the neighborhoods and explore them. We therefore essentially needed a data-set that contains such boroughs and neighborhoods that exist in each city as well as the the latitude and longitude coordinates of each neighborhood.

The Foursquare API provides the prime data gathering source as it has a database of millions of places, especially their places API which provides the ability to perform location search, location sharing and details about a business.

The New York City has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a data set that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the the latitude and longitude coordinates of each neighborhood. Luckily, this data-set exists for free on the web. Feel free to try to find this data-set on your own, but here is the link to the data-set: https://geo.nyu.edu/catalog/nyu_2451_34572

Unlike in the case of New York, the neighborhood data for Scarborough, Toronto is not readily available on the internet, which presents an interesting fact about the field of data science that each project can be challenging in its unique way hence the need to learn to be agile and refine the skill to learn new libraries and tools quickly depending on the project.

The main data source available for this project is the Wikipedia page from the link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Use of the Foursquare API:

To achieve the ends of this project, there's the need to gain in-depth data about the different neighborhoods of New York and Toronto cities with the choice of data source being the Foursquare's API location data.

This is informed by the fact that Foursquare's API has an in-depth data about locations which include, but not limited to pictures, name of venues, menus(where necessary) and locations. The output data obtained from the Foursquare API had venue information with specified distances of the longitude and latitude of the postcodes. The information obtained per venue as follows:

- Neighborhoods themselves;
- Neighborhood Latitudes;
- Neighborhood Longitudes;
- Venues and their names;
- Venue Latitudes;
- Venue Longitudes;
- Venue Categories.

The geographical coordinate of New York City are 40.7127281, -74.0060152 while the geographical coordinate of Toronto city are 43.6534817, -79.3839347.

Python Libraries & Models to be used:

To achieve the ends of this project, the following Python Modules would be used:

1. Beautiful Soup and Requests: For web scrapping, automation & handling HTTP requests;
2. Pandas: For creating and manipulating dataframes;
3. Matplotlib: ForCharts and data Plotting;

4. Folium: For visualiing the neighborhoods cluster distribution of using interactive leaflet map;
5. Scikit Learn: For importing k-means clustering;
6. Geocoder: For retrieving Location Data;
7. XML: To separate data from presentation and XML stores data in plain text format;
8. JSON: For Handling JSON files.

K-Means Clustering Approach:

There exisists different possible clustering models for clustering but for purposes of this project, we intend to present the model that is considered the one of the simplest model among them which is the K-Means Clustering Approach. Despite its simplicity, this approach is vastly used for clustering in many data science applications, especially useful if you need to quickly discover insights from unlabeled data

Methodology & Workflow:

1. First, we imported the boroughs and neighborhood list of Toronto from Wikipedia and converted it to data frame using pandas package in python.

Out[7]:

	Postalcode	Borough	Neighborhood
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
8	M9A	Etobicoke	Islington Avenue, Humber Valley Village
9	M1B	Scarborough	Malvern, Rouge
11	M3B	North York	Don Mills
12	M4B	East York	Parkview Hill, Woodbine Gardens
13	M5B	Downtown Toronto	Garden District, Ryerson

2. Another data set comprised of location data of neighborhood and boroughs was then imported in .csv format and then converted to data frame.

Out[9]:

	Postalcode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
5	M9A	Etobicoke	Islington Avenue, Humber Valley Village
6	M1B	Scarborough	Malvern, Rouge
7	M3B	North York	Don Mills
8	M4B	East York	Parkview Hill, Woodbine Gardens
9	M5B	Downtown Toronto	Garden District, Ryerson

3. After Cleaning the data sets, the two tables were then merged to get the final Toronto neighborhood data set.
4. Next, the Geo location data of New York was then imported in .json format.
5. Then the Neighborhoods, Boroughs and their corresponding latitude and longitude were filtered out and converted to a data frame.
6. Having done this, with the output data of the neighborhood location data for each city, using Foursquare API, all venues data will be imported into two data frame for each neighborhood in Toronto and New York.
7. To keep the comparison of the two cities manageable, only a number of the most common venues shall be taken into account and the rest dropped.

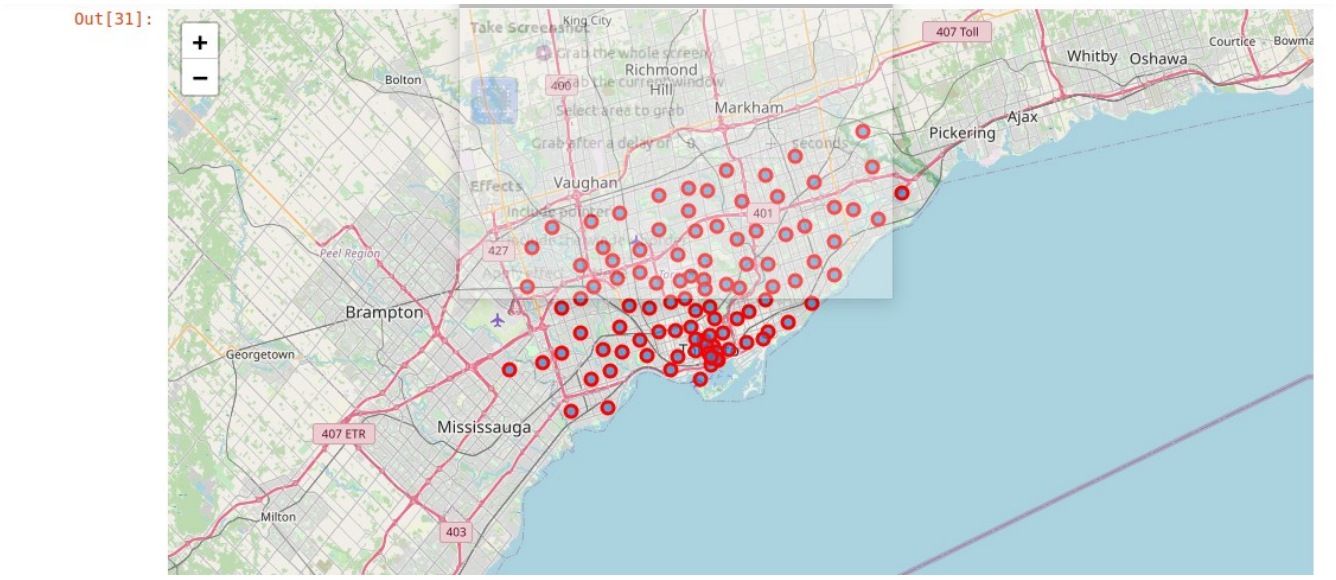

Figure 1: Toronto Map

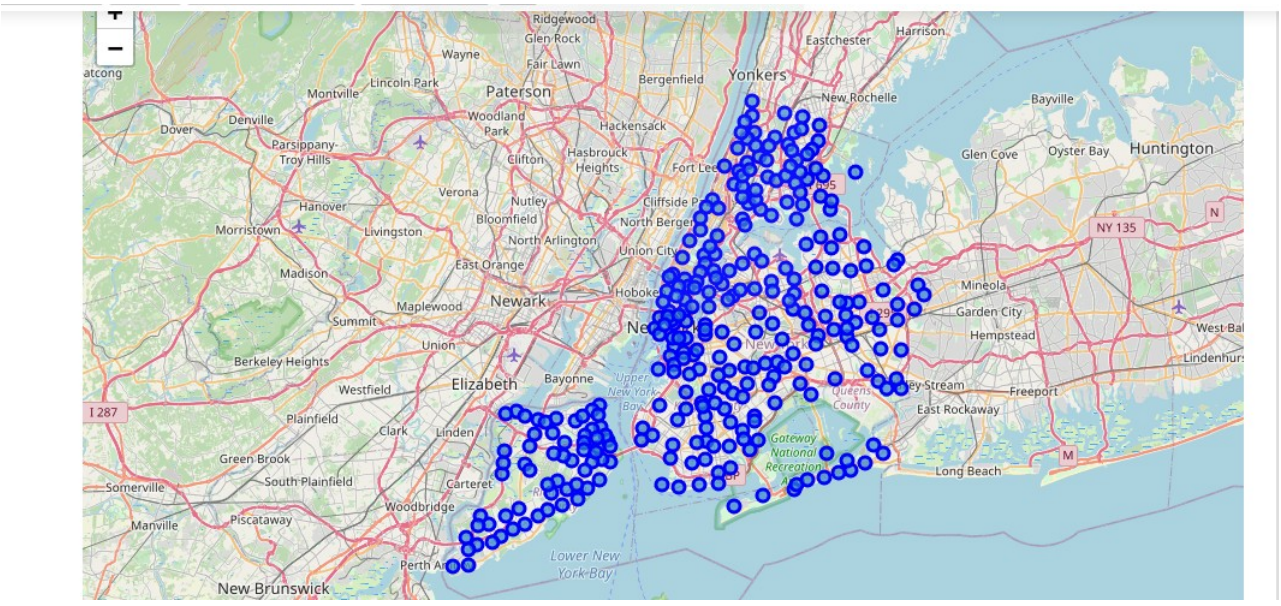

Figure 2: New York Map

8. Using the Sci-kit learn module of Python, We'll then introduce the K-Means Clustering model, taking clusters of 5 for both cities, and the labeled neighborhood data was plotted in a map using folium package.

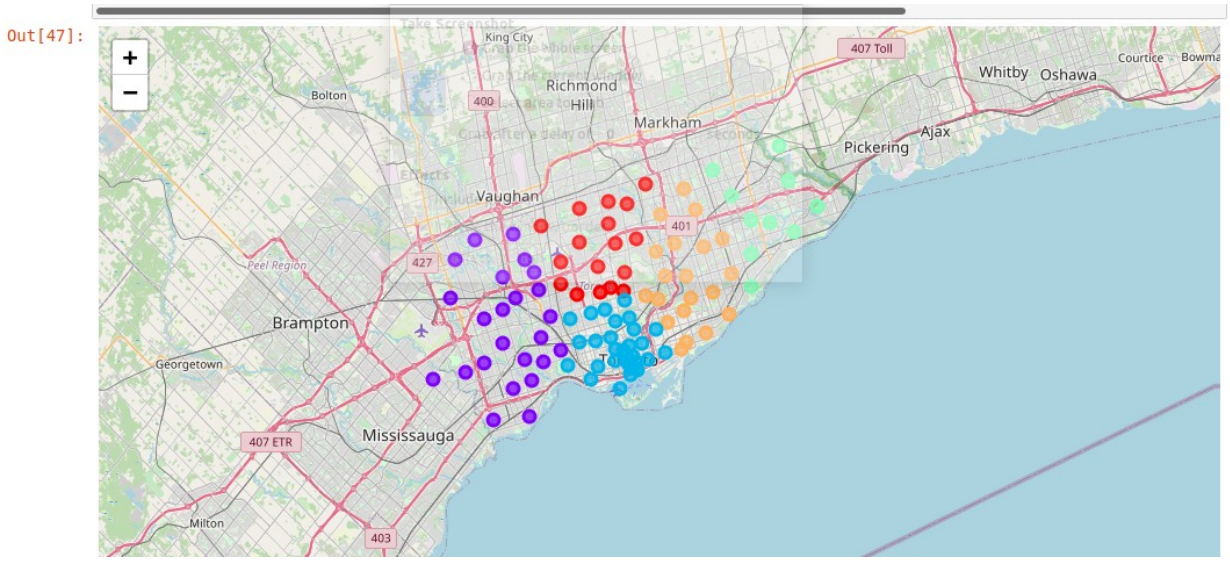

Figure 3: Toronto Cluster Map

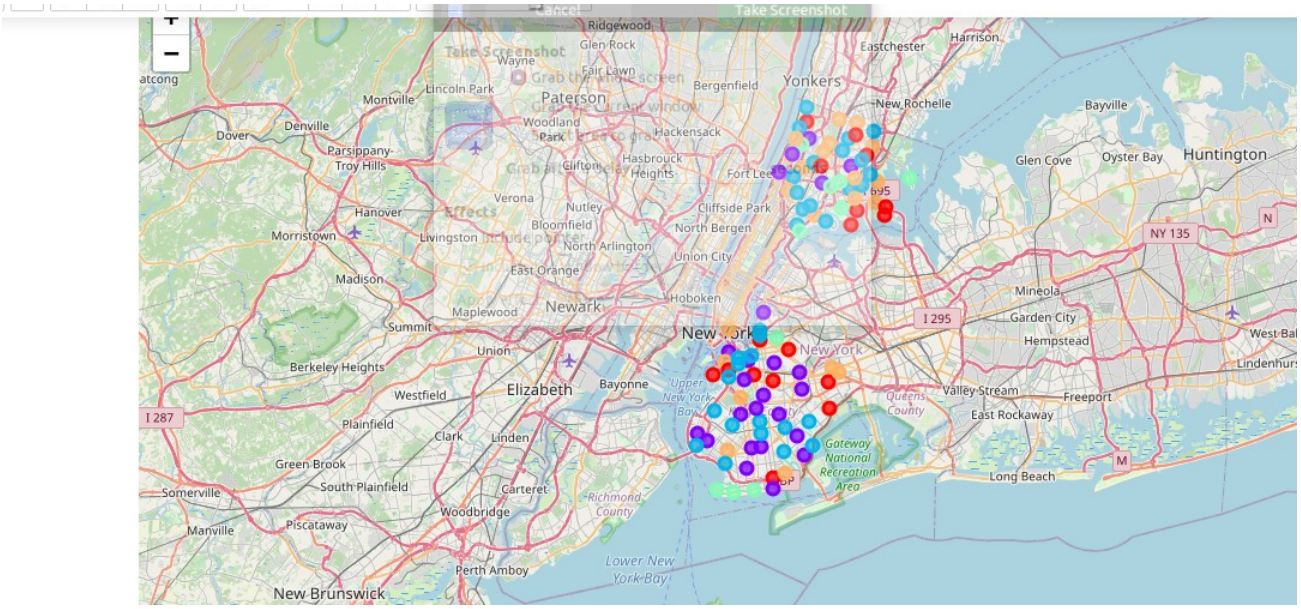

Figure 4: New York Cluster Map

9. From there we'll then be able to make the necessary comparisons which form the basis of this project.

Results:

It can be seen that Toronto has five big clusters (about 25% of the neighborhoods) and a smaller one with no insignificant clusters compared to them. These are based on Boroughs, Neighborhoods and Postal codes.

For New York, there is one big (83%) and two mid size clusters. Other two clusters are insignificant compared to them.

Conclusion:

In conclusion, with regard to neighborhood types, Toronto seems to have more defined and uniform neighborhood types while New York has much more varieties.