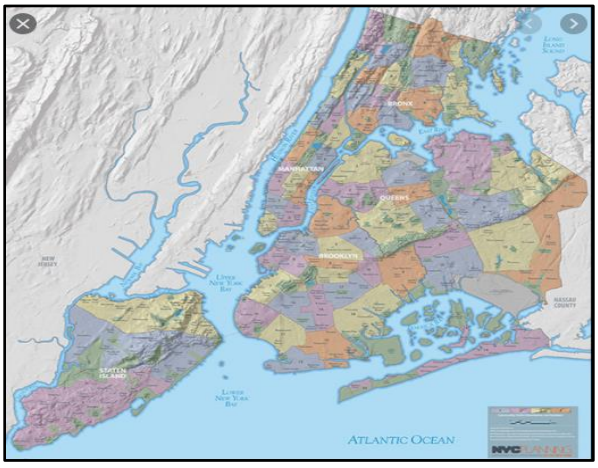
Harsh Chovatia August 2020

Coursera- IBM Data Science Professional Certificate

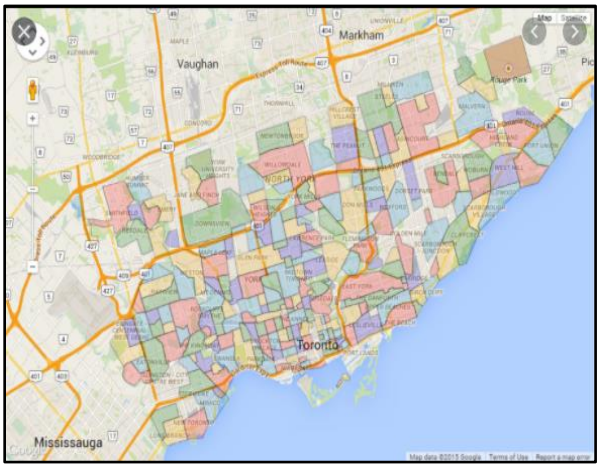
Applied Data Science Capstone Project

Introduction to Business Problem

A Fortune 500 company is looking to move its headquarters to either Toronto or New York City. The company wants insight into the neighborhoods and local businesses in the cities so that its employees may have the optimum living standards and quality of life. This project will analyze and explore the similarities and dissimilarities between certain neighborhoods in the two cities i.e. Toronto, Canada and New York City, New York. and determine which neighborhoods best fit the culture of the Fortune 500 company's employees.



VS



Data

Data The data used for this project will be acquired from the respective cities Wikipedia website pages. The datasets consist of the postal codes, neighborhood names, latitude, and longitude information for each neighborhood. Foursquare API search feature will be used to collect neighborhood venue information. Details about local venues and locality will be provide insight into the qualities of a neighborhood. In addition to Foursquare, various python packages will be used to create maps and machine learning models to further provide insights into our neighborhood battle project.

We will be using the following datasets from the following websites:

Toronto Neighborhoods - https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
Toronto Latitude and Longitude - http://cocl.us/Geospatial_data
New York City neighborhood data- https://geo.nyu.edu/catalog/nyu_2451_34572

Also, we will be using the FourSquare API to explore the surrounding locations and venues to assist with the intended analysis.

Methodology

Python Packages:

- Pandas - Library for Data Analysis
- NumPy – Library to handle data in a vectorized manner
- JSON – Library to handle JSON files
- Geopy – To retrieve Location Data
- Requests – Library to handle http requests
- Matplotlib – Python Plotting Module
- Scikit-learn – Python machine learning Library
- Folium – Map rendering Library

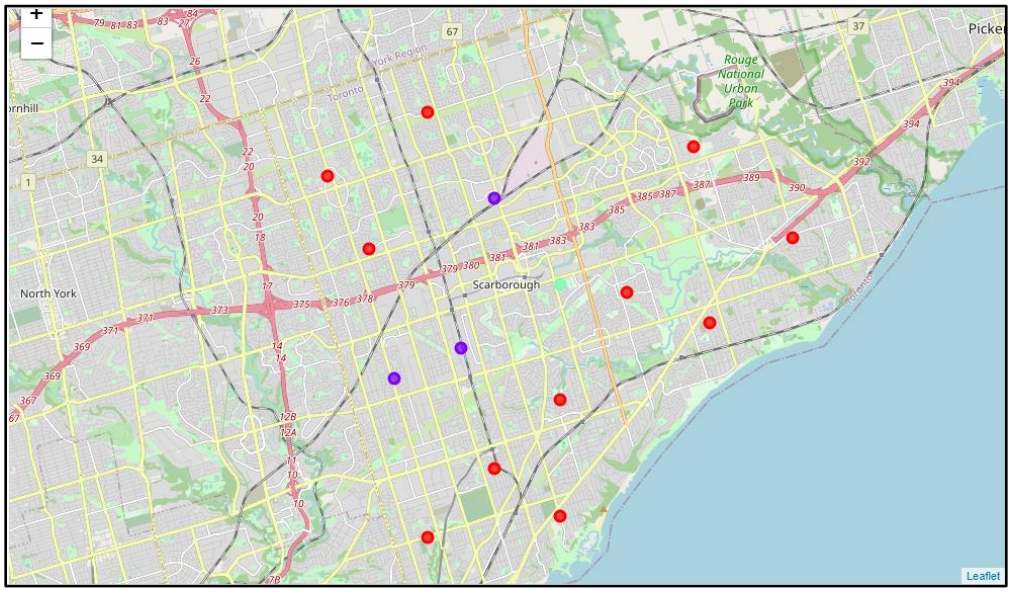
Workflow:

1. Download, Extract and Transform the dataset for the information on Postal codes, Boroughs and Neighbourhoods of Toronto, CA.
2. Merge the latitude and longitude information for the various neighbourhoods of Toronto using the available dataset and visualize the neighbourhoods over Folium maps.
3. Execute HTTP requests to the Foursquare API server using the latitude and longitude information for the neighbourhoods of Toronto city to explore the venues for each neighbourhood.
4. For detailed analysis and simplicity of understanding, we will restrict the scope of this project only for one Borough of Toronto i.e. Scarborough.
5. Using the Foursquare API search feature would be enabled to collect the nearby places of the Scarborough neighbourhood. We will prepare a data frame listing the top 10 venues in each of the Scarborough neighbourhoods.
6. Based on the venue information for each neighbourhood, we will perform K-means Clustering and divide the Scarborough neighbourhoods into 3 clusters. These 3 clusters can then be visualized on a Folium map.
7. For the next part of the project, to draw a comparison between Toronto, CA and New York city, USA, we perform similar set of operations as mentioned in **Steps 1-6** for the New York city data.
8. We will use the borough of Queens, for the analysis. We will explore the venues for the neighbourhoods. Prepare the data frames with the venue information.
9. For New York city, Queens borough information we perform the K-means clustering and divide the neighbourhoods into 5 clusters since the data is more compared to Scarborough.
10. Extensive comparative analysis of two randomly picked neighbourhoods would be carried out to derive the desirable insights from the outcomes. The clusters from each of those two chosen neighbourhoods would be analysed individually collectively and comparatively to derive the conclusions.

Results

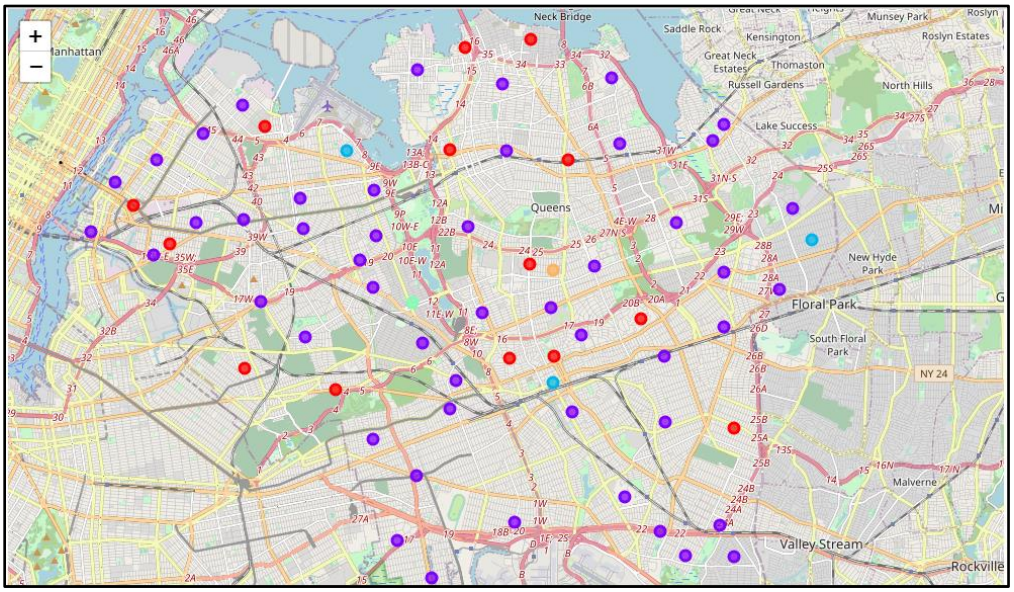
Scarborough Borough in Toronto, Canada

I use k-means to group the neighbourhoods in Scarborough into 3 clusters. Cluster_0 has 12 neighbourhoods and the most common venues are stadiums, banks, bus stations, parks, restaurants, fast food joints and breakfast spots. Cluster_1 has 3 neighbourhoods, and the most common venues are international cuisine restaurants, skating rinks, and general entertainment venues. Cluster_2 has 1 neighbourhood, and the most common venues are Asian restaurants and sports stores.



Queens Borough in New York City, USA

I used k-means to group the Queens borough into 5 clusters. Cluster_0 has 15 neighbourhoods and consist of many international cuisine restaurants and grocery stores. Cluster_1 has 61 neighbourhoods and the most common venue is a shopping mall along with some other important venues like pizza places, deli shops, and Chinese restaurants, banks, entertainment places. Cluster_2 has 3 neighbourhoods and the most common venue are donut shops and international cuisine restaurants, clothing stores. Cluster_3 has 1 neighbourhood and the most common venues are gyms and donut shops. Cluster_4 has 1 neighbourhood and the most common venues are the beach and a bakery.



Discussion

Toronto has 11 boroughs and 103 neighbourhoods. The geographical coordinate of Toronto, Canada are 43.7170226, -79.4197830350134. For our project we investigated in detail in one of the boroughs from Toronto i.e. Scarborough. In Scarborough borough, all its neighbourhoods collectively had 79 distinct venues in 52 categories.

New York City has 5 boroughs and 306 neighbourhoods. The geographical coordinate of New York City is 40.7308619, -73.9871558. For the scope of this project, we investigated in detail in one of the boroughs from New York city i.e. Queens. Foursquare Found 2097 distinct venues all together from the 81 neighbourhoods of the Queens borough.

Many of the neighbourhoods are homogenous and are very similar to each other. Both Scarborough and Queens borough consist of neighbourhood cluster that contain majority of the neighbourhoods, and the remaining cluster had 1-5 neighbourhoods. Queens borough had a significant a greater number of neighbourhoods and venues than Scarborough. Also, from our clusters, it was evident that the neighbourhoods from Queens had a better span of venues available and the number of places and options was also large. So, people from diverse backgrounds living in these parts of the city can easily adapt and find the places of their interest and necessity to lead a healthy lifestyle.

Conclusion

In conclusion, based on the quantity of venues and variety of venues, I would choose Queens over Scarborough as a choice to relocate the headquarters of the Fortune 500 company. Queens offers way more in choices for restaurants, gyms, grocery stores, and extracurricular activities for individuals and families of the company employees. It can be assumed that people staying in Queens can lead comfortable lifestyle.