Capstone Project – Battle of the Neighborhoods Chan Jang – April 2020

Introduction

The purpose of this project is to utilize data analysis within python to compare and contrast the two neighborhoods. The selected neighborhoods include North York, Toronto and Bronx, New York. A Fortune 500 company that is located in North York is interested in expanding into a second headquarters. The senior staff have chosen the city of New York as their primary location. After conducting the data analysis, we will compare and contrast the two cities to see if the Bronx location is a sensible choice.

Data

Web-scraping tools allow us to utilize data stored online. In this project, we will be using Wikipedia's list of neighborhoods in each city. Coordinates for each neighborhood can be accessed through the GeoPy API. Lastly, the venues in each location can be identified through the Foursquare API. These tools will be used to collect information to conduct analyses in python including cleaning/transforming data, visualizing data into maps (Folium), and conducing unsupervised machine learning: k-means clustering.

The following websites were used for this project:
Toronto Neighborhoods with postal code M:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
Coordinates for Toronto neighborhoods:
https://cocl.us/Geospatial_data
New York Neighborhood Data:
https://geo.nyu.edu/catalog/nyu_2451_34572
New York coordinates for each neighborhood: Geolibrary within python

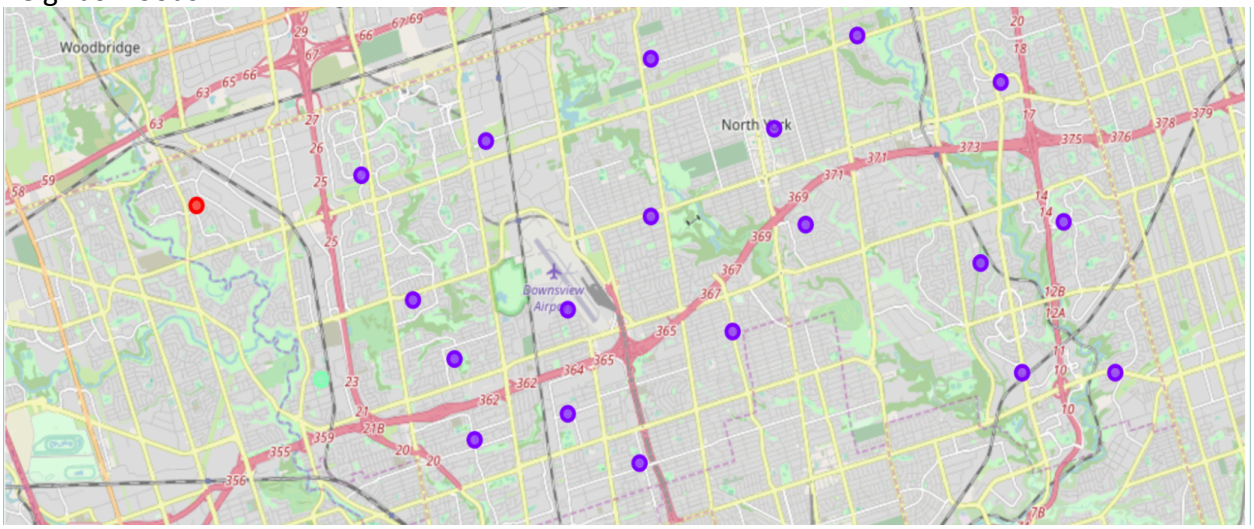
Methodology

1. Using Foursquare's developer application, HTTP requests were generated using the postal codes of the North York neighborhoods (latitude and longitude).
2. The Foursquare API allowed us to generate information about nearby venue demographics. The neighborhood parameters were set to a 500-meter radius.
3. Visualization of the obtained data was processed through python's Folium library.
4. Extensive comparative analysis of the two neighborhoods (North York/Bronx) were conducted using python's scientific libraries Pandas, NumPy, and Scikit-learn

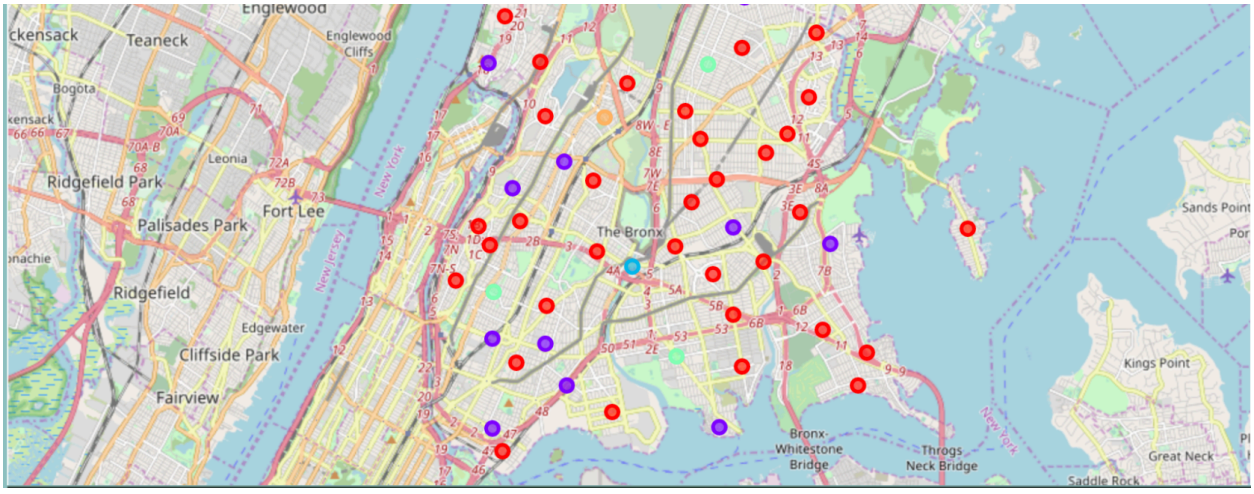
5. Unsupervised machine learning algorithm K-Means clustering was applied to generate clusters of venues categories.

Results

North York in Toronto, Canada: Three clusters were chosen to be used in the k-means clustering algorithm (Cluster 0, 1, and 2). Cluster 0 and 2 showed one neighborhood. Cluster 1 showed 20 neighborhoods.



Bronx in New York, USA: Five clusters were chosen to be used in the k-means clustering algorithm (Cluster 0, 1, 2, 3, 4). Cluster 0, 1, 2, 3, and 4 showed 37, 11, 1, 3, and 1 neighborhood(s) respectively.



Discussion

Toronto showed to have 10 boroughs and 103 neighborhoods. The coordinates for Toronto, CA are 43.6534817, -79.3839347. Within North York, there were 240 venues in 24 neighborhoods. There were 198 distinct venues in 101 categories.

Bronx, New York showed to have 885 distinct venues in 169 categories. The latitude and longitude of Bronx, NY are 40.8466508, -73.8785937.

The Bronx borough showed a much higher density of venues and categories.

Conclusion

Based on the information gather and analyzed, we can say for certain the Bronx borough has significantly different features when compared to North York. Although the types of venues can be grouped to be similar, the sheer amount of venues within the geographical region is a cause for concern. Further analyses are recommended to gain a holistic approach for an expansion location. Venue density cannot alone rule out a decision to expand into the Bronx borough and other factors must be evaluated. These factors include employee morale, availability of other locations, profit margins, and international laws for multination corporations.