Clustering of Similar Neighborhoods in Different Cities

Aditya Dutta

March 2020

1. Introduction

1.1. Purpose

The Purpose of this report is to explore and investigate the distribution of different venues in 2 of the most populous and popular cities in U.S. and Canada: New York City and Toronto.

1.2. Background

Every city has its own characteristics and various venues spread all over the city. Every city has some common places where people visit often and enable other tourists to explore the city. NYC and Toronto collectively welcome close to 100 million tourists every year. Despite having many dissimilarities, it is possible to take into consideration the similar venues of different cities, segment them and group the neighborhood according to Venue category. This can give a fair enough idea and reference to help decide when people consider moving out of a city to another and tourists who come to enjoy the stay.

The City of New York, known as New York City (NYC) is the most populous city in the United States. With an estimated 2018 population of 8,398,748 distributed over about 784 km^2. In 2018, NYC welcomed over 65.2 million tourists.

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 6,082,000 as of 2016 distributed over 630.2 km^2. In 2018, Toronto welcomed over 27.5 million visitors.

1.3. Method of Investigation

All the data has been extracted from secondary sources regarding the Borough and Neighborhood Venues along with location data.

2. Data Acquisition & Cleaning

2.1. Data Sources

All the data has been extracted from secondary sources and has been processed to make it ready for analysis. The analysis uses 2 sets of data consisting of NYC and Toronto Neighborhood Location data and other 2 sets having Venue spots around these neighborhoods.

2.2. Data Cleaning

The NYC and Toronto Neighborhood locations were extracted from geolocation json data files. The NYC data was extracted from .json files. All the variables under 'features' was extracted consisting of Borough. Neighborhood, Latitude and Longitude. The Borough and Neighborhood data of Toronto was extracted by python's web scraping package 'Beautiful Soup' from URL, which was followed by extracting the location data from URL and finally merging them on common key 'Postcode'. The Toronto data had to be cleaned by clearing the Not Assigned Postcodes and replacing the Not Assigned Neighborhoods by the assigned Borough Value.

After obtaining all the required location data, we needed to figure out the venue spots near the Neighborhood locations. For this, we used the Foursquare API which grants access to an enormous database consisting of venues from all over the world including rich variety of info such as address, tips, photos and comments. With an already logged in credentials the Foursquare API can be accessed by giving the CLIENT ID and 'CLIENT_SECRET'.

Using Foursquare API, different venues were extracted. One Hot Encoding was performed on the datasets and grouped together. Following this, the NY and Toronto venues data was merged on common venue spots which left us with 250 venue spots per neighborhood.

After this the data analysis has been performed.