Capstone Project

Sining LIU

1. Introduction

In this project for the capstone course for Applied Data Science, we shall cluster and compare neighbourhoods in Toronto and New York City on the basis of popular venues, major crime indicators and area of the neighbourhood. At the end of the exercise, a desirable output shall be a table of similar neighbourhoods across New York and Toronto. This information shall be useful for anyone who is doing business in any of these cities and wants to expand to the other city. It shall also be useful for professionals who are looking to change jobs within New York or Toronto or from one city to another.

1.1 Background

The City of New York, usually called either New York City (NYC) or simply New York (NY), is the most populous city in the United States. With an estimated 2018 population of 8,398,748 distributed over a land area of about 302.6 square miles (784 km2), New York is also the most densely populated major city in the United States.Located at the southern tip of the state of New York, the city is the center of the New York metropolitan area, the largest metropolitan area in the world by urban landmass and one of the world's most populous megacities. A global power city, New York City has been described as the cultural,financial and media capital of the world, and exerts a significant impact upon commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports.

Similarly, Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 2,731,571 in 2016. The diverse population of Toronto reflects its current and historical role as an important destination for immigrants to Canada. Its varied cultural institutions, which include numerous museums and galleries, festivals and public events, entertainment districts, national historic sites, and sports activities, attract over 43 million tourists each year. Toronto is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

1.2 Problem Description

Both cities have a large and diverse population of both Toronto and New York, including. Every year hundreds of thousands of immigrants, businessmen and professionals visit, migrate to or settle in these cities for work, livelihood and tourism. Due to the large area, several neighbourhoods, income differences, and variations in quality of life from one neighbourhood to another, it is often a tedious task to find neighbouhoods suitable to one's preferences. Businessmen often need information on neighbourhood to reloacte or to open new businesses. Further, anyone moving from New York to Toronto or vie versa would want to move to more or less similar neighbourhood. Also many people prefer to avoid crime-prone areas of a city whether for residence or business. Therefore the crime data for each neighbourhood is also relevant to categorization of similar neighbourhoods. Therefore the problem is to group neighbourhoods within and across New York and Toronto and categorize them based on popular venues, businesses, area and crime rate.

1.3 Target Audience

1. Businesses looking for expansion in New York and Toronto
2. Professional looking for relocation
3. Students looking for relocation
4. House buyers

1.4 Success criteria

A good categorization of neighbourhoods between Toronto and New York and their prominent features.

2. Data

For our project we will need the following data for both Toronto and New York:

1. Major Crime Indicators for each neighbourhood/precinct
2. Area, Latitude and Longitutde for each neighbourhood/precinct
3. List of popular venues for each neighbourhood/precinct

2.1 Data for Toronto

For crime statistics of Toronto we shall use the data provided here. (https://services.arcgis.com/S9th0jAJ7bqgIRjw/arcgis/rest/services/Neighbourhood_MCI/FeatureServer/0/query?where=1%3D1&outFields=*&returnGeometry=false&outSR=4326&f=json) This dataset provides a number of features including, neighbourhood name, neighbouhood id, major crime indicators from 2014-2018, average of major crimes for last 5 years, area and length of boundaries of each neighbourhood and population. For our analysis, we shall retain the data on

Neighbourhood
Assault_AVG
AutoTheft_AVG
BreakandEnter_AVG
Robbery_AVG
TheftOver_AVG
Homicide_AVG
Shape__Area

For latitude and longitude of each neighbourhood, we shall fetch the data for each neighbourhood using **'Here'** geocoder from geopy library in Python.
The data of venues and venue categories for each neighbourhood will be fetched using **Foursquare API**.

2.2 Data for New York

For crime statistics of New York we shall use the data provided here. (https://www1.nyc.gov/assets/nypd/downloads/excel/analysis_and_planning/historical-crime-data/seven-major-felony-offenses-by-precinct-2000-2018.xls) This dataset provides a number of features including precinct id, major crimes from 2001-2018. For our analysis, we shall retain the data on

Precinct id

Crime figures from 2014 to 2018

Using **dataset cleaning and feature engineering**, we shall compute the average of major crimes for each precinct and rename the columns to align the dataset with toronto dataset. For further analysis, each precinct shall be treated as a neighbourhood.

For location data of New York Precincts, open data available at this link (https://data.cityofnewyork.us/api/views/kmubvria/rows.csv?accessType=DOWNLOAD) will be used. This dataset provides the boundary data and shape_area for each precinct. The centroid (latitude and longitude) of each precinct will be calculated by extracting the boundary data from above dataset and used as neighbourhood latitude and longitude. Shape_area feature will be used as it is.

During coding, the performance of several free geocoders in fetching location data for precincts was observed to be poor as many precincts could not be located by free geocoders available in geopy.

The data of venues and venue categories for each neighbourhood will be fetched using **Foursquare API**.