Capstone Project - The Battle of Neighborhoods

June 30, 2019

Introduction

Ranked as one of the top countries to live in according to its unprecedented quality of life, public education system as well as medical facilities, Canada has been seeing a massive rise in the number of new immigrants over the past decades. When it comes to the place for newcomers to settle down, Toronto is always among the first destinations to consider.

The objective of this project is to provide some guidance to those new immigrants who will be looking for a suitable neighborhood in Toronto for them to settle down, based on a comparative analysis on various features and amenities across different neighborhoods. It is understood that immigrants will pick their own neighborhood based on different criteria depending on their own characteristics and preferences. In this project, the analysis is aiming for those families with a particular focus on the income level and education background of the residents living in the neighborhood. A high income and education level is normally a reflection of the safety of the community and relatively good manners of the residents. Some other common amenities such as restaurants and grocery stores nearby are also taken into account in this analysis. This project can always be modified and customized for those families who consider other attributes as driving factors to decide on which neighborhood to settle down.

Data

The data used in this analysis is gathered from the following various sources.

- **Neighborhood location information of Toronto :**

  Location information of Toronto neighborhoods is available from the previous assignment in this course, which includes the postal code, longitude and latitude information for each neighborhood in the city of Toronto.

- **Neighborhood profile of Toronto :**

  The neighborhood profile is obtained from the Census of Population. The profile collects data about age and sex, language, immigration and internal migration, ethnocultural diversity, housing, education, income, and labour, among which population, education, and income information is of interest in this analysis. Due to the fact that the Census is held across Canada every 5 years, data to be used is from the most recent Census in year 2016.

  Data source publicly available : https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#8c732154-5012-9afe-d0cd-ba3ffc813d5a

- **Venue information of interest :**

  Venue information of interest (i.e. restaurants and grocery stores) is obtained using the Foursquare API based on the longitude and latitude coordinates of the neighborhood.

Methodology

The methodology adopted in this analysis is shown as the following steps.

Step 1 : Identify the neighborhoods of Toronto in which residents have high income and education background to construct the pool of neighborhoods of interest for the analysis

As mentioned in the previous section, a high income / education level normally reflects a relatively safe community and good manners of the residents. The population of each neighborhood, the average individual income and the number of residents in that neighborhood who have a university Bachelor's degree and above, are extracted from the neighborhood profile dataset. Because each community has different population size, the percentage of residents with a Bachelor's degree and above is used, which is calculated as the number of residents with a Bachelor's degree and above divided by the population of that neighborhood. Then, the income and education information is normalized by dividing the maximum value across all neighborhoods.

This analysis targets those neighborhoods that have both high income and education level, therefore, a weighted average between the normalized income and education data for each neighborhood is computed. The pool of neighborhoods of interest for the remaining analysis is constructed by selecting those with the highest weighted average values between income and education. In this analysis, 12 neighborhoods are chosen.

Step 2 : Locate those neighborhoods selected from previous step

The postal codes for those neighborhoods of interested will be looked up within the neighborhood location dataset created from the previous assignment in this course. The latitude and longitude of each neighborhood will then be obtained by calling the geocoder library. A dataset will be created and saved for later analysis containing the neighborhood name, the borough that it belongs to, postal code, and its latitude and longitude information.

Step 3 : Connect to Foursquare and retrieve venue data of interest within each neighborhood

After the neighborhood dataset is created, venue information within each neighborhood is collected by connecting to the Foursquare API. The radius for hunting venues is set to be 1 kilometer from the center of each neighborhood. Since this analysis has a particular focus on amenities about places for food, such as restaurants, grocery and convenience stores, some post-processing is required, where only those venues of interest are extracted from the venue dataset.

Once the venue information is all gathered and post-processed, the column of Venue Category will be one-hot encoded so that different venues will have different feature columns, which will be used for subsequent machine learning and statistical analysis.

Step 4 : Apply machine learning technique (K-Means Clustering) to analyze the data

In this step, one of the machine learning techniques, i.e. K-Means Clustering, is applied to the dataset, where neighborhoods are clustered. The value of "K" is selected to be 5, which is deemed to be able to cover the complexity of the problem. After clustering, each neighborhood is assigned to one of the 5 cluster groups.¶

Step 5 : Make decisions on the most suitable neighborhood based on statistical indicators

The final step is to determine the most suitable neighborhood by comparing the sum score of all venues for each cluster. The cluster with the highest score is identified, and the neighborhoods within that cluster are returned as the most suitable communities to choose.

Results

This section presents the analysis results. By applying one of the machine learning technique K-means clustering and summing up the total venue score for each of the clusters, it can be seen that Cluster 5 has the highest venue score of 54, and Clusters 3 and 1 follow closely with the second / third highest scores of 53 and 50, respectively. Cluster 2 has the lowest venue score. Based on the results, Clusters 5, 3, and 1 have the most accessibility of various places for food, including all kinds of restaurants, joints, and convenience / grocery stores.

Cluster	Venue Total Sum Score
5	54.0
3	53.0
1	50.3
4	38.5
2	7.2

As the final step, the neighbourhoods that belong to each cluster are identified as shown below.

Neighborhood	Cluster
York Mills	1
Rosedale	5
Moore Park	2
Forest Hill South	1
Lawrence Park South	2
Waterfront Communities	4
The Islands	1
Annex	2
Leaside	4
Bay Street Corridor	2
Bedford Park	3
The Beaches	2

Discussions

Based on the results, it can been seen that Cluster 5 has only one neighborhood, Rosedale located in downtown Toronto, which has the highest venue score, followed by Bedford Park within Cluster 3, and York Mills, Forest Hill South, and The Islands from Cluster 1 closely. These neighborhoods are therefore recommended to those new immigrants who are looking for a high income / education community with a good accessibility to places to eat.

Conclusions

It can be concluded that the following neighborhoods are determined to be suitable for those new immigrants who are seeking for a community with a relatively high income and education level, and a good accessibility to places for food within the neighborhood:

- Rosedale
- Bedford Park
- York Mills
- Forest Hill South
- The Islands

This study presented a comparative analysis on the neighborhoods of Toronto to provide some guidance to new immigrants who are deciding on the most suitable community for them to settle down. The analysis targets particularly those immigrants who are looking for a neighborhood with a high income and education level, as well as a good accessibility to places for food such as restaurants and convenience stores. This study can be customized for other immigrants that take into account other criteria and driving factors. Due to complexity, the analysis presented did not consider housing availability and prices within the neighborhood. The future study may include these factors and other considerations to achieve a more comprehensive scope.