Coursera Capstone Project: Applied Data Science
The Battle of Neighborhoods

Haoyuan Gao

Introduction

Neighborhoods are different, even if they are in the same city or just next to each other. When searching a new place to live, everyone has the desire to find the place that matches his or her lifestyle. For new residents and even long-time residents to a city, it is difficult to deal with everything that each unique corner offers. Toronto, with its unofficial nickname of "the city of neighborhoods", is filled with over 200 official and unofficial fabulous neighborhoods within its city boundaries, which makes it a real challenge to find an ideal neighborhood to live in Toronto.

An ideal neighborhood recommendation should match one's characteristic and lifestyle to the neighborhood's characteristics, such as demographics, economic factors, recreational opportunities, dining options, entertainment options, etc. Everyone who lives or will live in the target city is a potential customer who might need recommendation while choosing an ideal place of residence, either renting an apartment or purchasing a house to become part of the neighborhood. To succeed with a neighborhood lifestyle guide, one must explore into the neighborhood and understand its characteristics and components in order to offer its customer with ideal recommendations.

Business Problem

For new residents and even long-time residents to a city, it is difficult to deal with everything that each unique corner offers. Although finding a perfect neighborhood that meets all of a user's expectations is difficult, there is still a chance to match his or her preferences to provide optimal results. Thus, the main objective of this project is to find ideal neighborhood recommendations for a user, in which his or her lifestyle could be matched with the neighborhood's characteristics.

Methods

Data collection

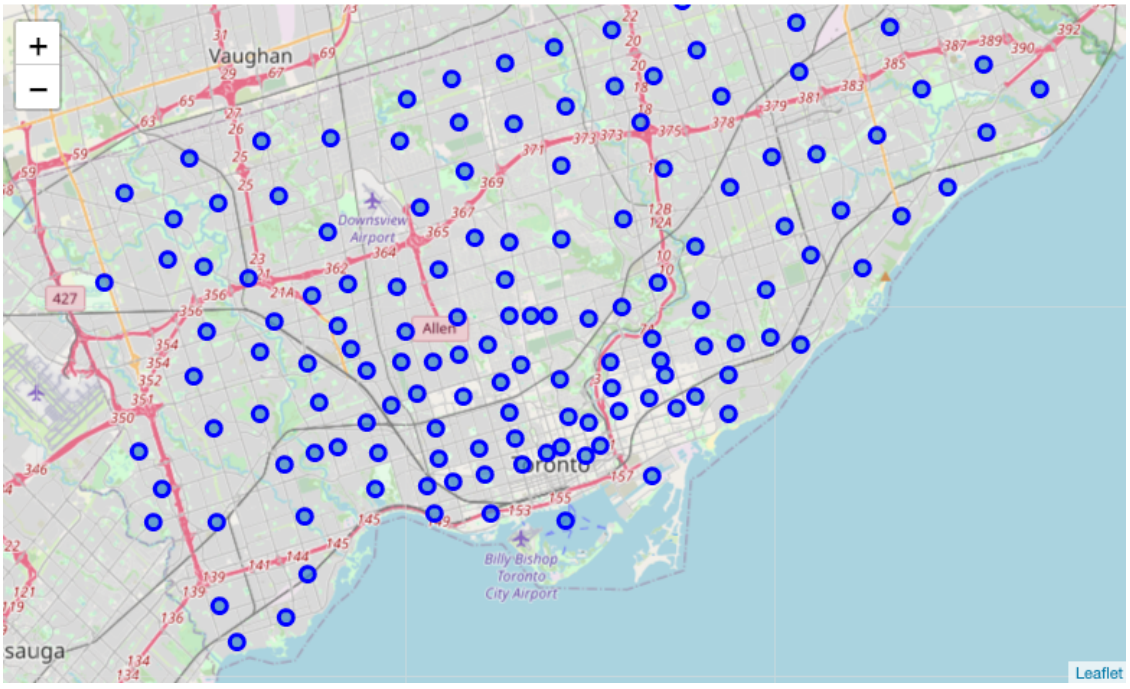
The data for this project is retrieved from multiple sources with the considerations to the accuracy.

The data of the neighborhoods in Toronto is from Toronto Open Data, including economics, demographics, crime and the boundaries of Toronto neighborhoods. Data is downloaded and pre-processed using Python packages.

The venue data is retrieved by passing required parameters to the FourSquare API. The data contains the contains the information regarding venue name, geographical coordinates of the given venues and venue category. A combined DataFrame that contains all the venue details and respective neighborhoods information will be created for further analysis.

Exploratory Analysis

Folium is used for visualizing geospatial data in this project. It takes the location parameters, latitude and longitude, and generates an interactive map around the given coordinates. All cluster visualization is generated using Folium and its Marker class.



Statistical Modeling

K-Means clustering
Cluster is the technique for grouping observations within a dataset, which allow us to identify the similarity among different observations. K-mean clustering is a commonly used unsupervised machine learning algorithm for splitting the dataset into a set of k groups, in where k is the pre-specified number of groups.

Optimal number of clusters
The Elbow Method is used for optimizing the value of k by running K-Means clustering for a range of values of k and computes the average distortion score for each cluster. The value of k ranges from 1 to 10, where the value of k stands for the number of clusters created, and the optimal cluster size is determined.

Correlation Matrix
Correlation matrix is a table that shows the correlation coefficients between the selected variables, in which each cell in the matrix shows the correlation between two variables. The correlation matrix is used to summarize data and to select variables as inputs for further advanced analysis.

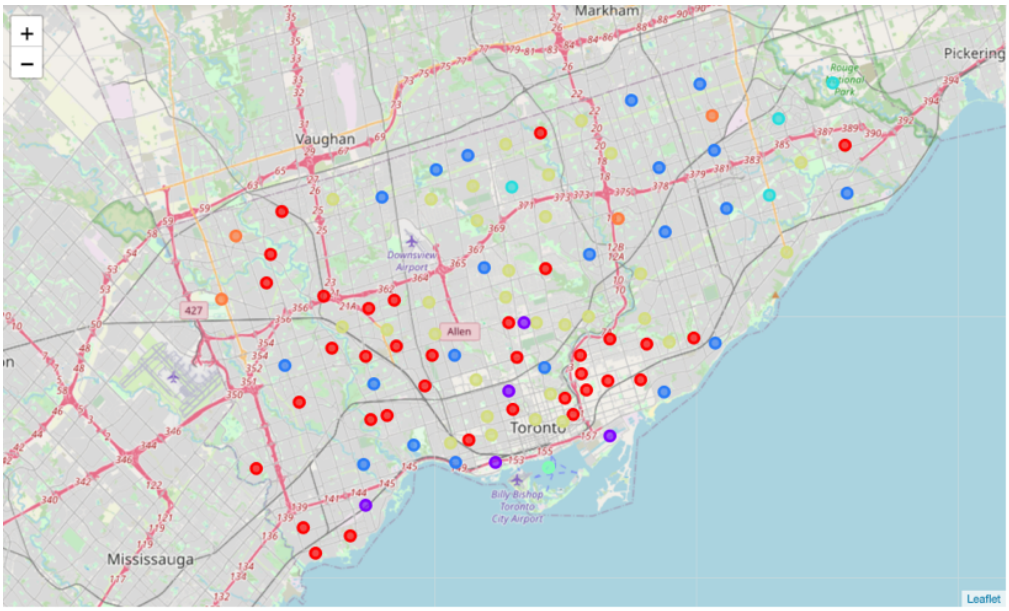
Results

The dataset used in this analysis contains 140 observations and 28 variables: neighborhood (Neighborhood), total business establishments (Business), TransUnion debt load risk score (Debt Risk Score), real estate sale prices (Home Prices), total local employment (Local Employment), social assistance recipients (Social Assistance Recipients), number of population in each age group (Seniors 65 and over, Child 0-14, Youth 15-24, Adults 25-64), number of population of foreign language groups (Language - Chinese, Language - Italian, Language - Korean, Language - Persian, Language - Portuguese, Language - Russian, Language - Spanish, Language - Tagalog, Language - Tamil, Language - Urdu), geolocation information (Latitude and Longitude), crime data (Total Major Crime Incidents) and top 5 most common venues (Most Common Venue).

Dataset contained missing geolocation values were excluded from the analysis. The demographics data is divided into age groups and language groups for further analysis. Since most of the neighborhoods have Chinese-speaking population as the majority foreign language speaking group, the K-Means clustering result on language group is excluded for further analysis; and the data of the language groups is converted to categorical data by calculating the top foreign population group (represented as Language Group) for later.

The K-Means clustering analysis could help to identify the components of similar neighborhoods, the percentage of different age groups that live in the neighborhoods and the type of venue components that existing in the neighborhoods, in order to suggest similar community components with the map locations provided as follows:

	Cluster Labels	Neighborhood	Seniors 55 and over	Child 0-14	Youth 15-24	Adults 25-54	Longitude	Latitude
0	6	West Humber-Clairville	22.442473	16.523427	14.970890	46.063210	-79.596356	43.716180
1	6	Mount Olive-Silverstone-Jamestown	17.270479	22.304670	14.593336	45.831515	-79.587259	43.746868
31	6	Parkwoods-Donalda	23.450061	16.650770	12.862788	47.036381	-79.330180	43.755033
91	6	Agincourt North	28.472222	13.070988	12.808642	45.648148	-79.266712	43.805441



A correlation matrix is created by using the economics and crime data to view the correlations between different economic factors and safety. To be more specific, total number of registered business has strong positive linear relationships on the number of total major crime incidents and local employment; and the number of total major crime incidents has moderate positive relationships on the number of local employment and social assistance recipients.

Conclusion

This project aims on providing residential recommendation among 140 neighborhoods in the city of Toronto by applying statistical modeling on the demographics, economics, crime, geolocation and the nearby venues data of the neighborhoods. The neighborhoods with similar components, such as age groups, similar common venues are recognized as identical places. By using provided user information regards of residential preferences, this project could provide ideal recommendations to user. Further application should be focus on creating user profiles by pre-define different lifestyle characteristics and the associated neighborhoods components.

References

Social Development, Finance & Administration (2014, December 31). Wellbeing Toronto - Economics, Open Data Dataset. Retrieved November 11, 2020, from: https://open.toronto.ca/dataset/wellbeing-toronto-economics/

Social Development, Finance & Administration (2017, February 28). Wellbeing Toronto - Demographics, Open Data Dataset. Retrieved November 11, 2020, from: https://open.toronto.ca/dataset/wellbeing-toronto-economics/

Social Development, Finance & Administration (2020, November 2). Neighborhoods, Open Data Dataset. Retrieved November 11, 2020, from: https://open.toronto.ca/dataset/neighbourhoods/

Social Development, Finance & Administration (2017, February 28). Wellbeing Toronto - Safety, Open Data Dataset. Retrieved November 11, 2020, from: https://open.toronto.ca/dataset/wellbeing-toronto-safety/