Data analysis for choosing a new location for a store

Introduction

The similarity of neighborhoods has been an important consideration in decision in choosing business location and city planning. The availability of large location data makes data analysis and visualization possible. Statistical methods and machine learning approaches enables decision making based on quantitative analysis, which is more evidence-informed, robust, objective and comprehensive.

Assume you are a store owner in Toronto in one of the neighborhoods and you want to change location for a new shop. The decision you make would base on the use of data making the best of the available information.

Data

This report is using Foursquare location data and postal code information from Wikipedia. Location data enables the visualization of results, as well as the analysis and interpretation of different venue categories, to facilitate decision making and can analyze data on a large scale, within a short time, relatively low computational cost and financial cost.

Methodology

Python was used for data analysis. K-mean clustering was used to cluster similar neighborhoods. A wide range of statistical methods were used to explore data structure and foster interpretation. Folium package was used to visualize data on an interactive map, facilitating analysis as well as increasing the comprehensibility of the data.

(1)Data loading and preparation

Numpy, pandas,json, geopy, requests, matplotlib, Scikitlearn and Folium were used for data loading and preparation. This is the data frame after scraping from Wikipedia combined with coordinate data, respectively:

	Postal Code	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

(2)Data exploratory analysis

The coordinate of Toronto was acquired from geolocator, a folium map of Toronto was created using Folium and markers were added.

The neighbourhood data was acquired and explored using Foursquare location data. After exploration of the first neighbourhood, 2 venues were returned by Foursquare. Here is the first five venues of all venues in Toronto returned:

- Parkwoods
- Victoria Village
- Regent Park, Harbourfront
- Lawrence Manor, Lawrence Heights
- Queen's Park, Ontario Provincial Government

The venues were returned for each neighbourhood, including venue and venue category , along with geological coordinates. The first five neighbourhoods are shown as below:

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Agincourt	5	5	5	5	5	5
Alderwood, Long Branch	7	7	7	7	7	7
Bathurst Manor, Wilson Heights, Downsview North	21	21	21	21	21	21
Bayview Village	4	4	4	4	4	4

Each neighbourhood was analysed, the 3 most common venues for each neighbourhood were explored. Here's the first three neighbourhoods:

```
----Agincourt----
            venue  freq
0          Lounge   0.2
1  Breakfast Spot   0.2
2    Skating Rink   0.2


----Alderwood, Long Branch----
         venue  freq
0  Pizza Place  0.29
1     Pharmacy  0.14
2          Gym  0.14


----Bathurst Manor, Wilson Heights, Downsview North----
                venue  freq
0         Coffee Shop  0.10
1                Bank  0.10
2  Chinese Restaurant  0.05
```

The top 10 venues for each neighbourhood were also explored. Here's the first three:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Agincourt	Lounge	Skating Rink	Latin American Restaurant	Breakfast Spot	Clothing Store	Drugstore	Discount Store	Distribution Center	Dog Run	Doner Restaurant
1	Alderwood, Long Branch	Pizza Place	Gym	Pharmacy	Coffee Shop	Sandwich Place	Pub	Women's Store	Dog Run	Dim Sum Restaurant	Diner
2	Bathurst Manor, Wilson Heights, Downsview North	Coffee Shop	Bank	Mobile Phone Shop	Bridal Shop	Sandwich Place	Diner	Restaurant	Deli / Bodega	Supermarket	Middle Eastern Restaurant

(3)Machine Learning and clustering

K-means was used to cluster neighborhoods and each neighbourhood was labelled, as shown below:

	Neighbourhood	Clusters labels
0	Parkwoods	0.0
1	Victoria Village	1.0
2	Regent Park, Harbourfront	1.0
3	Lawrence Manor, Lawrence Heights	1.0

Results

There are 9, 88, 1, 1, 1 boroughs in each cluster. Some of the neighbourhoods in each cluster are shown below, according to the order of the labels. For the first two clusters, only the first three neighborhoods are shown:

First cluster: Caledonia-Fairbanks, East Toronto, Broadview North (Old East York).

Second cluster: Victoria Village, Regent Park, Harbourfront

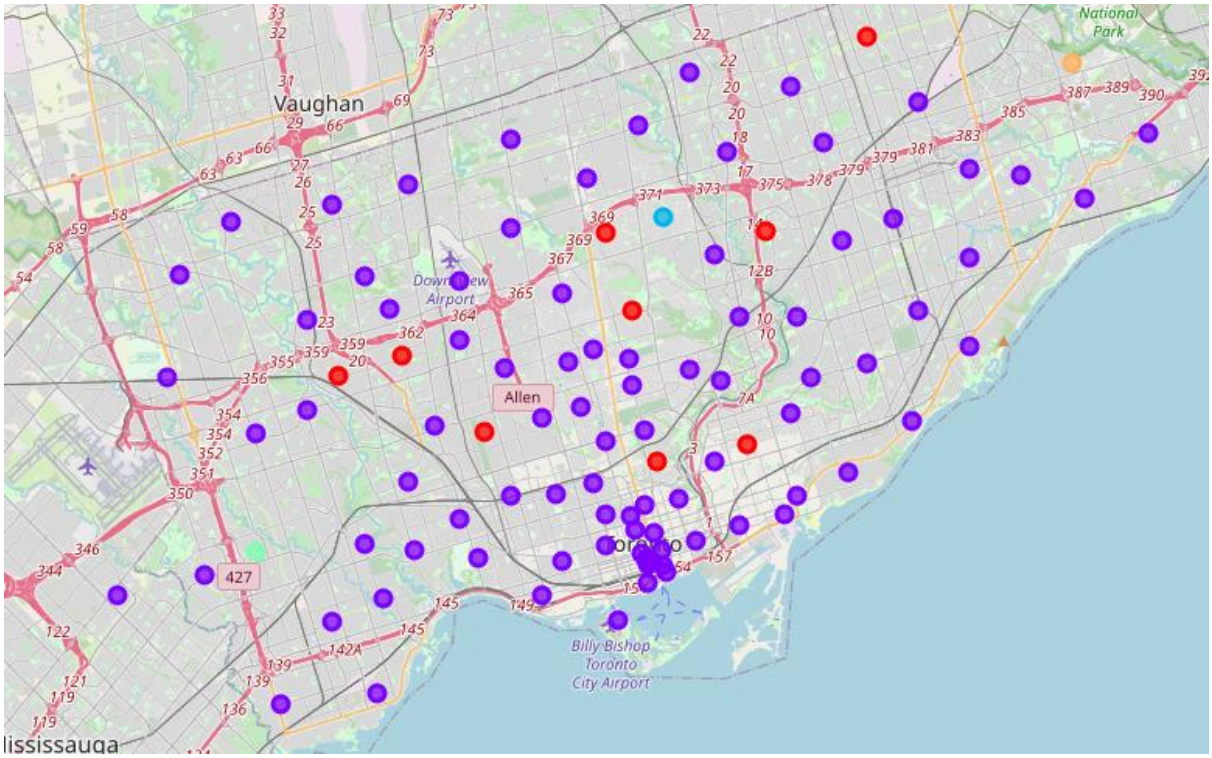
Third cluster: York Mills, Silver Hills

Fourth cluster: West Deane Park, Princess Gardens, Martin Grove, Islington and Cloverdale

Fifth cluster: Malvern and Rouge. Fast food restaurant is the most common venue.

Discussion

Clusters are visualized on folium map represented by different colors:



Most neighborhoods are concentrated in the second cluster with the three out of five clusters having only one borough. This may suggest the high similarity between neighborhoods. The division of different functional areas in the city may not be distinct to each other to a great extent. As is shown in the map, the second cluster covers most of the city while the other clusters are embedded inside, with no clear boundaries between clusters.

Thus, if a change of location is envisioned, based on where the shop is originally located, the decision might be different. If it is located in the neighborhoods of the second cluster, it should not be a problem moving to most of the places in the city (only taking account of venue categories). This could also be the same for if it is located in the neighborhoods of the first cluster. Moving from York Mills, Silver Hills, West Deane Park, Princess Gardens, Martin Grove, Islington, Cloverdale, Malvern and Rouge could be more challenged. However, considering the embedding distribution of the neighborhoods in different clusters, and relative close distance, even being located in the neighborhoods of last three clusters could still be within customers' travel range and have enough ability to attract customers. Nevertheless, many other factors including prices, downtown/uptown, subjective consideration etc. are also of great importance.

Conclusion

1. Change of location should be based on where the shop is originally located.
2. Most neighborhoods are concentrated in the second cluster.
3. Other factors also play an important part.