**Clustering Neighborhoods in Toronto**

Xiqiao Liu IBM-Capstone

**Define the problem and background**

Tom lives in Highland Creek, which is a neighborhood in eastern Toronto, Ontario. He gets a new job in Etobicoke at the western part of Toronto. Without a car, the commute between the two places must be painful and it would take 2-hour，three transfers for one way. So Tom decide to move to a new place in Etobicoke to mitigate his pressure. His wants to find a neighborhood similar to Highland Creek.

**Data Source**   
Data are collected from Foursquare. We will use information of venues to cluster the neighborhoods in Etobicode as well as Highland creek in order to find similar neighborhoods to Highland creek.

A screenshot of a cell phone

Description automatically generated

Data from Foursquare in JSON format

{'id': '4c38d4ab18e72d7fb01d19f5',

'name': 'Colonel Danforth Park',

'location': {'lat': 43.7775073613354,

'lng': -79.16430271940706,

'labeledLatLngs': [{'label': 'display',

'lat': 43.7775073613354,

'lng': -79.16430271940706}],

'distance': 840,

'cc': 'CA',

'city': 'Toronto',

'state': 'ON',

'country': 'Canada',

'formattedAddress': ['Toronto ON', 'Canada']},

'categories': [{'id': '4bf58dd8d48988d1e7941735',

'name': 'Playground',

'pluralName': 'Playgrounds',

'shortName': 'Playground',

'icon': {'prefix': 'https://ss3.4sqi.net/img/categories\_v2/parks\_outdoors/playground\_',

'suffix': '.png'},

'primary': True}],

'referralId': 'v-1575558004',

'hasPerk': False}

The venue categories of an area as example:

['General College & University',

'Coworking Space',

'Pharmacy',

'Miscellaneous Shop',

'Middle School',

'Fast Food Restaurant',

'Office',

'Pet Store',

'Spa',

'Financial or Legal Service',

'Trade School',

'Mexican Restaurant',

'Hobby Shop',

'Fast Food Restaurant',

'Miscellaneous Shop',

'College Arts Building',

'Liquor Store',

'Café',

'Medical Center',

'Office',

'Office',

'Office',

'Office',

'Office',

'Office']

After changing the categorical variables into dummy variables, columns of venue categories with 0 or 1 will be added to the original data frame.

A screenshot of a cell phone

Description automatically generated

**Build Model**

To find similar neighborhoods, I clustered the neighborhoods based on the venue categories within each area. K-modes model is applied here because the variables are categorical. K-means has better performance for numerical data, however, Euclidean Distance can’t appropriately find the dissimilarity of categorial data points. After creating a K-modes model, I tuned the parameters by changing the number of K. The optimal K in this case is 6 which has the lowest cost.

A picture containing knife

Description automatically generated



Then, displaying clusters on the map using different color points.

A close up of a map

Description automatically generated

**Conclusion**

Therefore, I recommend Tom to move to the following neighborhoods which resemble Highland Creek that of venue categories.

* Humber Bay Shores
* Mimico South New Toronto
* The Kingsway Montgomery Road
* Old Mill North
* Humber Bay
* King's Mill Park
* Kingsway Park South East
* Mimico NE
* Old Mill South
* The Queensway East
* Royal York South East
* Sunnylea