The Battle of Neighborhoods

Capstone Project

Background

- A New York based startup based on event management is planning its expansion in Toronto. However, unlike New York, the neighborhood data is not readily available on the internet.
- The company is required to put together an application that offers good organized events and offers various services like a hotel of residence, meeting halls, places of landscapes to visit, stores for shopping, restaurants and cafes.

Problem

- The company's purpose is to make a list of places of landscapes in Toronto, including the nearest restaurants, cafes and shopping stores for each place.

Interest

- The company is required to put together an application that offers good organized events and offers various services like a hotel of residence, meeting halls, places of landscapes to visit, stores for shopping, restaurants and cafes.
- So, the company's purpose is to make a list of places of landscapes in Toronto, including the nearest restaurants, cafes and shopping stores for each place.

	Postcode	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned\n
1	M2A	Not assigned	Not assigned\n
2	M3A	North York	Parkwoods\n
3	M4A	North York	Victoria Village\n
4	M5A	Downtown Toronto	Harbourfront\n

Data sources 1

- Neighborhood data are from Wikipedia for Postcode, Borough and Neighborhood.
- https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Postal Code	Latitude	Longitude
M1B	43.806686	-79.194353
M1C	43.784535	-79.160497
M1E	43.763573	-79.188711
M1G	43.770992	-79.216917
M1H	43.773136	-79.239476

Data sources 2

- Geospatial Coordinates data are from the CSV file.
- https://cocl.us/Geospatial_data

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Harbourfront, Regent Park	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Harbourfront, Regent Park	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Harbourfront, Regent Park	43.65426	-79.360636	Toronto Cooper Koo Family Cherry St YMCA Centre	43.653191	-79.357947	Gym / Fitness Center
3	Harbourfront, Regent Park	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
4	Harbourfront, Regent Park	43.65426	-79.360636	Morning Glory Cafe	43.653947	-79.361149	Breakfast Spot

Data sources 3

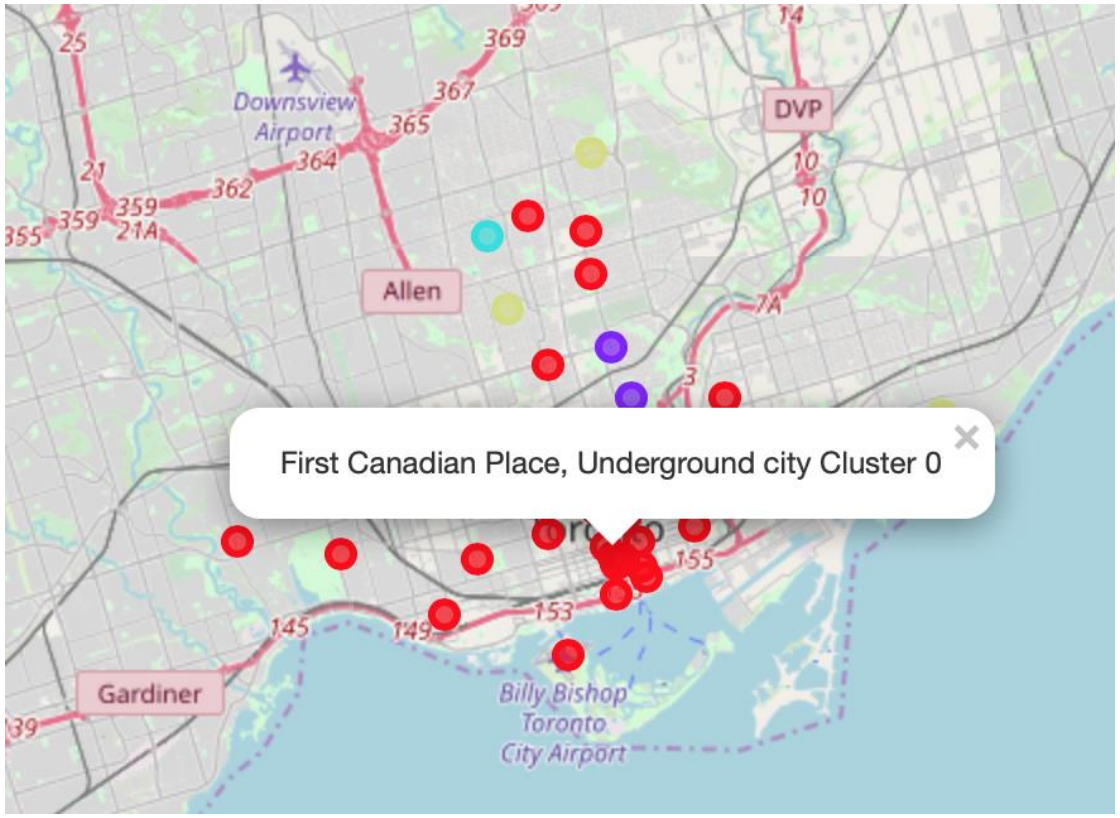
Amenity data are from Foursquare by API request.

Data cleaning

- Neighborhood data are available on the web. So obtained them by using Beautiful Soup and put them into a dataframe. Then read the CSV file with Geospatial Coordinates into another dataframe. As both of the two dataframe have postal code, I can use the postal code as keys to consolidate two dataframes into one dataframe with Neighborhood and Coordinates. Then I use the Coordinates to call Foursqare API to get Amenity data.

Exploratory Data Analysis

- Once I got the Neighborhood and Coordinates, I marked those Neighborhood in the map with blue circles.
- Then I used Foursquare API to get the nearby venues within 500 meter area. For this, I created an account on Foursquare API to get Client ID and Client Secret. I used this information to access locations on Foursquare API. I grouped them by neighborhoods and took the mean on the frequency occurrences of each venue category. This is the preprocessed data for Clustering.
  I used K-Means Clustering Method to group different venues in group. I created 4 clusters to show the futures of different neighborhood Clusters.



Result

- 4 Clusters on the map with circles in different colors.

Result

The 4 clusters are :
Cluster 1 - Most Common Venues in this cluster are related to shop, park, Café, Store, etc.

Cluster 2 - Most Common Venues in this cluster are related to Playground, park, Trail, Building, etc.

Cluster 3 - Most Common Venues in this cluster are related to shop, Garden, Yoga Studio, Dive Bar, etc.

Cluster 4 - Most Common Venues in this cluster are related to park, Trail, Health Food Store, Lake, etc.

Conclusions

- As clustered with different futures, the best place to live with family should be cluster 1 supported by the most important amenities for daily life. The other 3 clusters could be also taken into consideration because of diversity of people who prefers to those futures.