Coursera Capstone

IBM Applied Data Science Capstone

Select Neighborhood in Toronto to Open New Restaurant

May 2020

Introduction

Business Problem:

For restaurants to be successful, location is one of the most determining factors. A good location can bring foot traffic to a new restaurant, which is crucial in the early stages. Location is also a factor that is difficult/costly to change in the future, so it is very important for the business owner to make an informed, calculated decision.

Target Audience:

This project aims to help business owners to select the most appropriate neighbourhood to open a new restaurant in the city. Toronto will be used for demonstration.

To help the owners make better decisions, we will first retrieve information about venues by neighbourhood from Foursquare, and then use machine learning kmeans clustering to group neighbourhood into clusters. Then we will examine the traits of each cluster and choose one that is most ideal for new restaurants.

Dataset

When choosing a location for a new restaurant, there are two important factors to consider: foot traffic and competition.

To measure these two factors, we will use the following metrics:
1. List of Toronto neighbourhoods, which we will extract from Wikipedia
2. Postal codes, Latitudes and longitudes of each neighbourhood, which we can get from the csv file provided in previous week's assignment

We will then retrieve the following data using Foursquare API
3. number of restaurants within 1000m of radius from the center of the neighbourhood
4. % of venues that are restaurants within 1000m of radius from the center of the neighbourhood
5. number of tips given to the restaurants
6. average rating scores of the restaurants

For example, for Parkwoods neighbourhood, the data will look like:
1.Parkwoods
2.M3A, 43.753259 -79.329656
3.34
4.24.65%
5.56
6.6.87

Methodology

1. Data Scraping

First, we'll get the list of Toronto neighbourhoods from Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). Using Pandas read_html, we could scrape the first table of the page and save it to a dataframe.

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

Next, need to get the latitudes and longitudes for each neighbourhood using data provided by IBM(https://cocl.us/Geospatial_data) and merge the two dataframes into one.

	Postal Code	Borough	Neighborhood	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. Acquiring Data from Foursquare

In this step, we will get all the nearby venue information within the radius of 1000m from the center of each neighbourhood. We will retrieve the following information by running the 'explore' requests with Foursquare API:

1) Venue name
2) Venue latitude, longitude
3) Venue category
4) Venue ID

We are still missing data on ratings and tips. So next we will use the 'venues' requests to get the data for each venue.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	ID	Rating	Number of Tips
0	Parkwoods	43.753259	-79.329656	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant	4b8991cbf964a520814232e3	8.8	16
1	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park	4e8d9dcdd5fbbbb6b3003c7b	7.2	4
2	Parkwoods	43.753259	-79.329656	Tim Hortons	43.760668	-79.326368	Café	57e286f2498e43d84d92d34a	7.0	1
3	Parkwoods	43.753259	-79.329656	A&W	43.760643	-79.326865	Fast Food Restaurant	58a8dcaa6119f47b9a94dc05	6.8	1
4	Parkwoods	43.753259	-79.329656	Bruno's valu-mart	43.746143	-79.324630	Grocery Store	4bafa285f964a5203a123ce3	6.6	4

We've acquired all data now. Next, we need to use data cleaning and wrangling techniques to prepare the data.

3. Data Cleaning and Wrangling

Firstly, many venues have not received any ratings. For these venues, assigning 0 rating scores will not be appropriate as it will downwardly screw the rating data. We will assign the average rating of all venues in the same neighbourhood to these venues.

Secondly, let's use onehot encoding to transform the dataframe and calculate the metrics that will be used in clustering:

1) Restaurant count: number of restaurants in each neighbourhood
   To get this data, we will create a list of columns that contain the word 'Restaurant", then create a column indicating whether each venue belongs to the restaurant category. A groupby function is then used to summarize the total number of restaurants in each neighbourhood.
2) Restaurant blend: this is the percentage of venues that restaurants represent in each neighbourhood
   Calculation: =count of restaurant/count of total venues
3) Restaurant rating: the average rating of all restaurants in each neighbourhood
4) Restaurant number of tips: total number of tips given to all restaurants in each neighbourhood

The dataframe looks like below:

Neighborhood	Restaurant Count	Restaurant Blend	Restaurant Number of Tips	Restaurant Rating
Agincourt	15	0.500000	0.0	0.0
Alderwood, Long Branch	1	0.033333	0.0	0.0
Bathurst Manor, Wilson Heights, Downsview North	4	0.133333	0.0	0.0
Bayview Village	4	0.266667	0.0	0.0
Bedford Park, Lawrence Manor East	11	0.366667	0.0	0.0

4. K Means Clustering

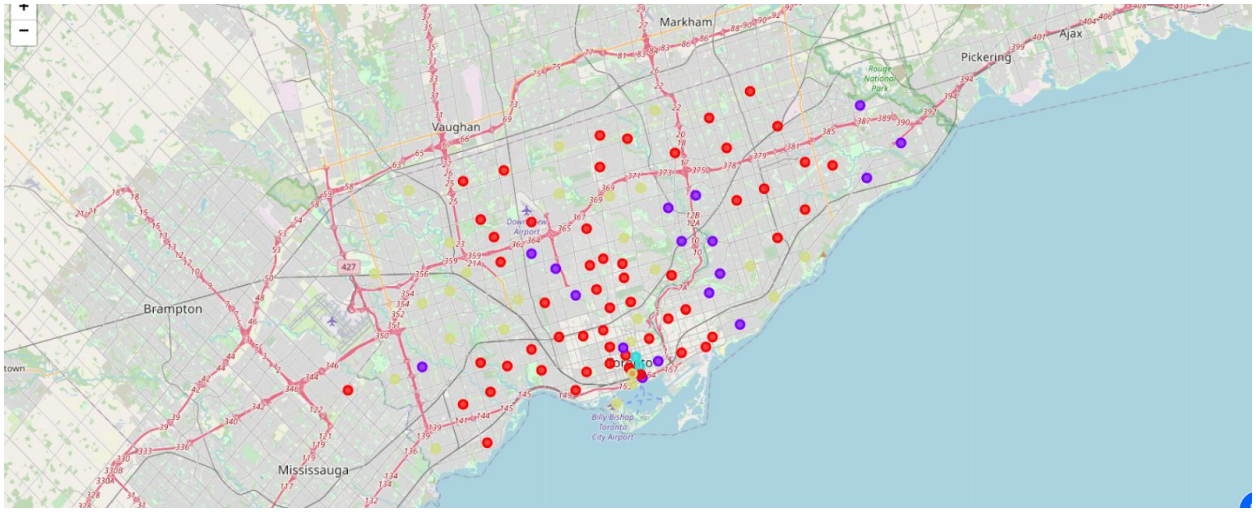
We will use K Means Clustering method to cluster the neighbourhoods into 4 clusters based on the metrics mentioned above. The results will help us have a better understanding of the characteristics of each neighbourhood.

The summary below shows the basic information of each cluster:

Cluster Labels	Restaurant Count	Restaurant Blend	Restaurant Number of Tips	Restaurant Rating	Latitude	Longitude
0	8.964286	0.301516	0.000000	0.000000	43.702568	-79.395410
1	6.470588	0.206481	12.223704	7.149935	43.712823	-79.342320
2	7.500000	0.250000	67.205357	8.329464	43.654328	-79.377177
3	1.777778	0.084035	0.000000	0.000000	43.702519	-79.443866

5. Data Visualization

The last step is to visualize the grouping using Folium.
Red dots represent cluster 0, purple dots represent cluster 1, blue dots represent cluster 2, yellow dots represent cluster 3.



Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 4 clusters:

- Cluster 0: Neighbourhoods with a decent amount of restaurants but no ratings and tips, indicating that customers have low level of motivation to endorse the restaurants.
- Cluster 1: Neighbourhoods with a decent amount of restaurants and average tips amount and rating.
- Cluster 2: Neighbourhoods decent amount of restaurants and a high volume of tips and high rating.
- Cluster 3: Neighbourhoods with a minimum representation of restaurants. No tips and ratings.
-

Cluster Labels	Restaurant Count	Restaurant Blend	Restaurant Number of Tips	Restaurant Rating	Latitude	Longitude
0	8.964286	0.301516	0.000000	0.000000	43.702568	-79.395410
1	6.470588	0.206481	12.223704	7.149935	43.712823	-79.342320
2	7.500000	0.250000	67.205357	8.329464	43.654328	-79.377177
3	1.777778	0.084035	0.000000	0.000000	43.702519	-79.443866

Discussions

As mentioned in the dataset section, foot traffic and competition are the two most important factors to be considered when selecting restaurant locations. Ideally, we would want to open the restaurant at a place that has a lot of foot traffic and relatively low competition.

- Restaurant count: more restaurants mean a high level of competition but also indicate more foot traffic due to the economy of scale effect.
- Restaurant blend: higher blends indicate greater competition as most venues fall into the same category.
- Restaurant rating: high ratings indicate high qualities of existing restaurants and a higher level of competition
- Restaurant number of tips: more tips given by users indicate that customers are more likely to give feedback and recommendations, increasing the online presence of the restaurants.

Conclusions

Cluster 1 and 2 seem to be the most suitable areas to open a new restaurant, with healthy foot traffic suggested by the restaurant count, restaurant blend.

Out of these two clusters, cluster 1 is better than cluster 2 for three reasons:

1) Restaurants represent lower percentage of all venues in the neighbourhood, which means less competition is present.
2) Existing restaurants have lower ratings, also suggesting less competition.
3) Cluster 2 neighbourhoods are located in the center of the city. The prime real estate usually cost more to rent.

Based on the above reasoning, business owners should consider opening new restaurants in neighbourhoods in cluster 1.