Inara Yusubova

IBM Data Science

Capstone project

Acknowledgements

That IBM Data Science course for giving me professional knowledge in this field. Although this capstone project is not that complicated, I made my best to deliver answer to the problem.

1. Introduction

This project, I am creating a hypothetical scenario for a concept that there may not be enough Indian Restaurants in Toronto Area. Therefore, it might be a great opportunity for an entrepreneur who is based in Canada. As the Indian food is popular among Asian community, so this entrepreneur might think of opening its business in areas where asian community resides.

With the purpose in mind, finding the location to open such a restaurant is one of the most important decisions for this entrepreneur and I am designing this project to help him find the most suitable location.

2. Objective

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new Indian Restaurant in Toronto, Canada. By using data science methods and tools along with machine learning algorithms such as clustering, this project aims to provide solutions to answer the business question: *In Toronto, if an entrepreneur wants to open an Indian Restaurant, where should they consider opening it?*

3. Data

I will use explanatory data analysis to explore my data.

To solve this problem, we will need below data:

- List of neighborhoods in Toronto, Canada
- Latitude and Longitude of these neighborhoods
- Venue data related to Indian restaurants. This will help us find the neighborhoods that are more suitable to open an Indian Restaurant

4. Overall process

Scrapping of Toronto neighborhoods via Wikipedia

- Getting Latitude and Longitude data of these neighborhoods via Geocoder package
- Using Foursquare API to get venue data related to these neighborhoods

5. Methodology

First, I need to get the list of neighborhoods in Toronto, Canada. This is possible by extracting the list of neighborhoods from Wikipedia:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

This is my main dataframe:

	Postalcode	Borough	Neighborhood	Latitude	Longitude
0	M4E	East Toronto	The Beaches	43.676357	-79.293031
1	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188
2	M4L	East Toronto	India Bazaar, The Beaches West	43.668999	-79.315572
3	M4M	East Toronto	Studio District	43.659526	-79.340923
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790

I did the web scraping by utilizing pandas HTML table scraping method as it is easier and more convenient to pull tabular data directly from a web page into the data frame. However, it is only a list of neighborhood names and postal codes. I need to get their coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates, I tried using Geocoder Package but it was not working so I used the CSV file provided by IBM team to match the coordinates of Toronto neighborhoods. After gathering these coordinates, I visualize the map of Toronto using Folium package to verify whether these are correct coordinates. Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius.

Foursquare API

I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude, and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues.

Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later. Here, I made a justification to specifically look for “Indian restaurants”.

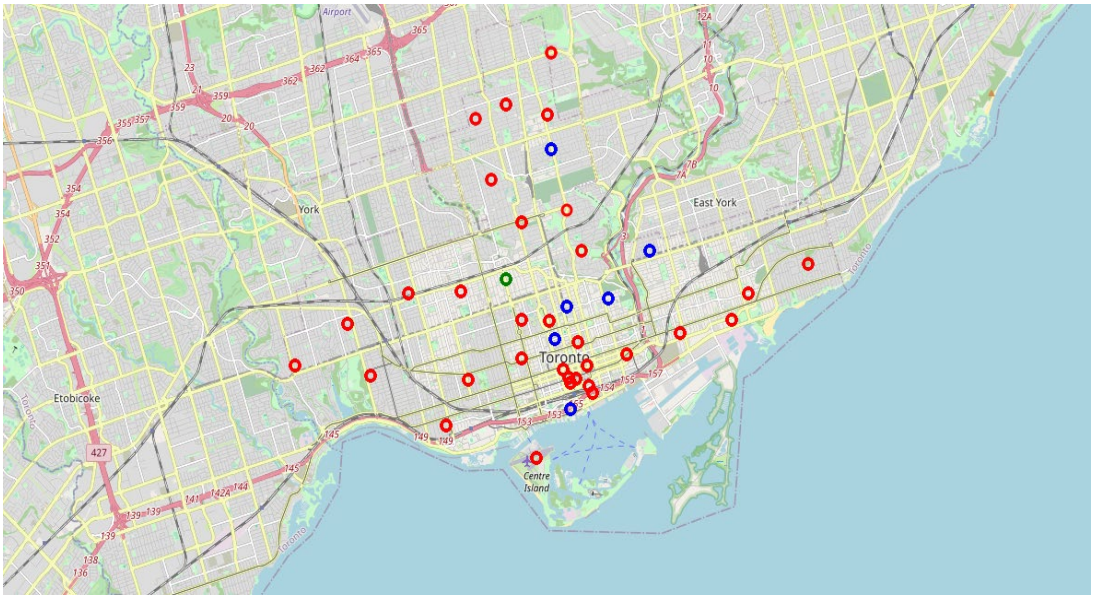
Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Berczy Park	57	57	57	57	57	57
Brockton, Parkdale Village, Exhibition Place	22	22	22	22	22	22
Business reply mail Processing Centre	18	18	18	18	18	18
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport	16	16	16	16	16	16
Central Bay Street	62	62	62	62	62	62
Christie	17	17	17	17	17	17
Church and Wellesley	78	78	78	78	78	78
Commerce Court, Victoria Hotel	100	100	100	100	100	100
Davisville	38	38	38	38	38	38
Davisville North	7	7	7	7	7	7
Dufferin, Dovercourt Village	19	19	19	19	19	19
First Canadian Place, Underground city	100	100	100	100	100	100
Forest Hill North & West	4	4	4	4	4	4
Garden District, Ryerson	100	100	100	100	100	100
Harbourfront East, Union Station, Toronto Islands	100	100	100	100	100	100
High Park, The Junction South	22	22	22	22	22	22
India Bazaar, The Beaches West	22	22	22	22	22	22
Kensington Market, Chinatown, Grange Park	56	56	56	56	56	56
Lawrence Park	3	3	3	3	3	3
Little Portugal, Trinity	42	42	42	42	42	42
Moore Park, Summerhill East	1	1	1	1	1	1

Lastly, I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well.

I have clustered the neighborhoods in Toronto into 3 clusters based on their frequency of occurrence for "Indian food". Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

	Neighborhoods	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	...	Theme Restaurant	Toy / Game Store	Trail	Train Station	Vegetarian / Vegan Restaurant	Video Game Store	Vietnamese Restaurant	Wine Bar	Women's Store	Yoga Studio
0	Berczy Park	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	0.000000	...	0.000000	0.000000	0.00000	0.00	0.017544	0.000000	0.000000	0.000000	0.000000	0.000000
1	Brockton, Parkdale Village, Exhibition Place	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	0.000000	...	0.000000	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	Business reply mail Processing Centre	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	0.000000	...	0.000000	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.055556
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	0.0625	0.0625	0.0625	0.125	0.125	0.125	0.000000	0.000000	...	0.000000	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	Central Bay Street	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	0.000000	...	0.000000	0.000000	0.00000	0.00	0.016129	0.000000	0.000000	0.000000	0.000000	0.016129
5	Christie	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	0.000000	...	0.000000	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
6	Church and Wellesley	0.012821	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.012821	0.000000	...	0.012821	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.025641
7	Commerce Court, Victoria Hotel	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.040000	0.000000	...	0.000000	0.000000	0.00000	0.00	0.020000	0.000000	0.000000	0.010000	0.000000	0.000000
8	Davisville	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	0.000000	...	0.000000	0.026316	0.00000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
9	Davisville North	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	0.000000	...	0.000000	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10	Dufferin, Dovercourt Village	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	0.000000	...	0.000000	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

6. Results



The results from k-means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on how many Indian restaurants are in each neighborhood:

- Cluster 0: Neighborhoods with less number of Indian restaurants.
- Cluster 1: Neighborhoods with no Indian restaurants.
- Cluster 2: Neighborhoods with more number of Indian restaurants

The results are visualized in the above map with Cluster 0 in green, Cluster 1 in blue, Cluster 2 in red.

7. Recommendations

Most of the Indian restaurants are in cluster 2 which is around Central Bay Street, Church and Wellesley, Berczy Park, Union Station, Richmond, lowest in Cluster 1 areas which are in North Toronto West and Parkade areas. Also, there are good opportunities to open near St James Town, Cabbagetown.

Looking at nearby venues it seems cluster 0 might be a good location as there are not a lot of Indian restaurants in these areas.

Therefore, this project recommends the entrepreneur to open an authentic Indian restaurant in these locations.