The Battle of the Neighborhoods

Finding the best location to open an Indonesian Restaurant in Toronto, Canada

September 19

IBM Coursera Capstone Project

Fitria Kurniasari

Contents

1. Introduction

In the early post-World War II period, most migrants from Indonesia to Canada were Indo people of mixed Dutch and original ancestry. Many did not come directly from Indonesia, but rather went to the Netherlands and then re-migrated due to racial prejudice they faced there. Community members believe that perhaps 3,000 live in the Ontario area. Indonesians of Chinese descent formed the main group in the stream of migration which began in the late 1960s and early 1970s. They have come to comprise an estimated 80% of Canada's population of Indonesian background. 7,610 respondents to the 1991 census stated their place of birth as "Indonesia". Around half of those were settled in the Greater Toronto Area. Data from the 2006 Census suggested that 14,320 people of Indonesian ethnic origin reside in Canada (3,225 single responses, 11,095 in combination with other responses), primarily in Ontario (6,325, or 44%), British Columbia (4,640, or 32%), and Alberta (1,920, or 13%). (Wikipedia)

Business Problem

Due to the numbers of Indonesian immigrant in Greater Toronto, Canada, our client want to open Indonesian restaurant in which area. The objectives of this capstone project is to select and analyze the best location in Toronto to open a new restaurant. Using data science methodology and machine learning techniques like clustering in this project to provide solutions to answer the business question: where would you recommend to open a new Indonesian restaurant in Toronto, Canada?

2. Data

The following data is needed to solve the problem

a. List of the neighborhoods in Toronto, Canada.
b. Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map.
c. Venues data, particularly related to Indonesian Restaurants. For this project we will use the Foursquare Places API. One of the features of this API is to provide a list of venues within a specific location, based on the Latitude and longitude coordinates and a radius. The data is used to clustering on the neighborhoods.

Sources of data and methods to extract them

The Wikipedia page contains a list of neighborhoods in Toronto, Canada (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M), with a total of 4 borough and 38 neighborhoods. Web scraping techniques with the help of Python requests and beautifulsoup packages is used to extract the data from the Wikipedia page. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. We will use Foursquare API to get the venues data for those neighborhoods. Foursquare API will provide many categories of the venue data, we are particularly interested in the Indonesian Restaurant category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

3. Methodology

How to get neighborhood list

Fortunately, the list of neighborhoods in the Toronto, Canada. is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). We do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the Toronto city.

Used Foursquare API to get the top 100 venues that are within a radius of 2000 meters

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the “Indonesian Restaurant” data, we will filter the “Indonesian Restaurant” as venue category for the neighborhoods.

Clustering on the data by using k-means clustering

Lastly, we will perform clustering on the data 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 is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 8 clusters based on their frequency of occurrence for “Indonesian Restaurant”. The results will allow us to identify which neighborhoods have higher concentration of Indonesian Restaurant while which neighborhoods have fewer number of Indonesian Restaurant. Based on the occurrence of Indonesian Restaurant in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new restaurant.

4. Results

We do web scraping from Wikipedia to get the borough and neighborhoods data. After that, with used Geocoder package to convert address into geographical coordinates in the form of latitude and longitude. Drop and clean the nonessential data content NaN. After gathering the data, we will populate the data into a pandas DataFrame. The result can be seen in figure.1. Then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinate data returned by Geocoder are correctly plotted in the Toronto city (figure.2). After then, create a new DataFrame with only boroughs that contain the word Toronto (figure.3).

Use the Foursquare API to explore the neighborhoods and get the top 100 venues that are within a radius of 2000 meters. Analyze Each Neighborhood and group rows by neighborhood and by taking the mean of the frequency of occurrence of each category. Then, searching how many Indonesian Restaurant in total at Toronto. There are 7 Indonesian Restaurant in total. Create a new DataFrame for Indonesian Restaurant data only. Then, Run k-means to cluster the neighborhoods in Toronto into 8 clusters. We choose k=8 because the cluster can have good separation using this k number. Sort the results by Cluster Labels (figure.4) and plot the map based on clustering category (figure.5). Now it's time to examine each cluster in figure 6 from cluster no. 0 to no.7

As you can see in cluster no. 1 there was Indonesian Restaurant in almost every neighborhoods. And in cluster no.5 almost half of the neighborhoods have Indonesian Restaurant already. When we take a look in the plotting map, cluster no.3, no. 7 and no.4 are too close to neighborhood cluster no.5.

It will be great opportunity to open the restaurant in other clusters because it doesn't exist yet. The cluster selected should be not too close to cluster 1 and 5 example cluster no.0, no.2, and no.6. So, the businessman should go south area instead of towards the middle area.

	PostalCode	Borough	Neighborhood
0	M1B	Scarborough	Rouge, Malvern
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Figure.1

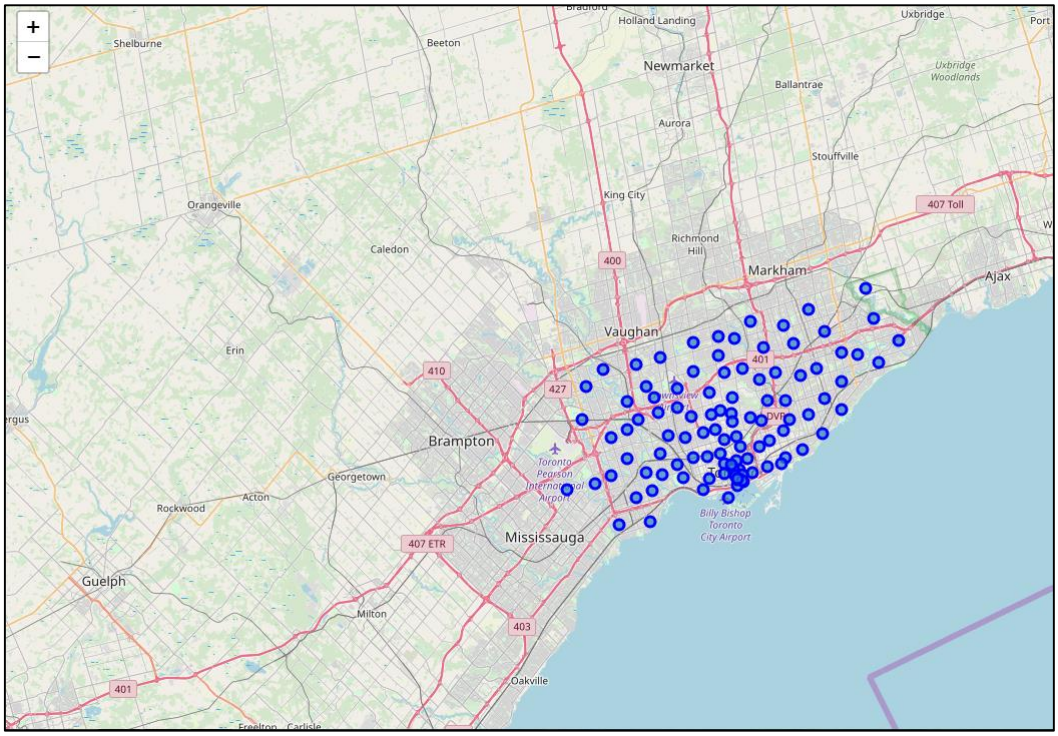

Figure.2

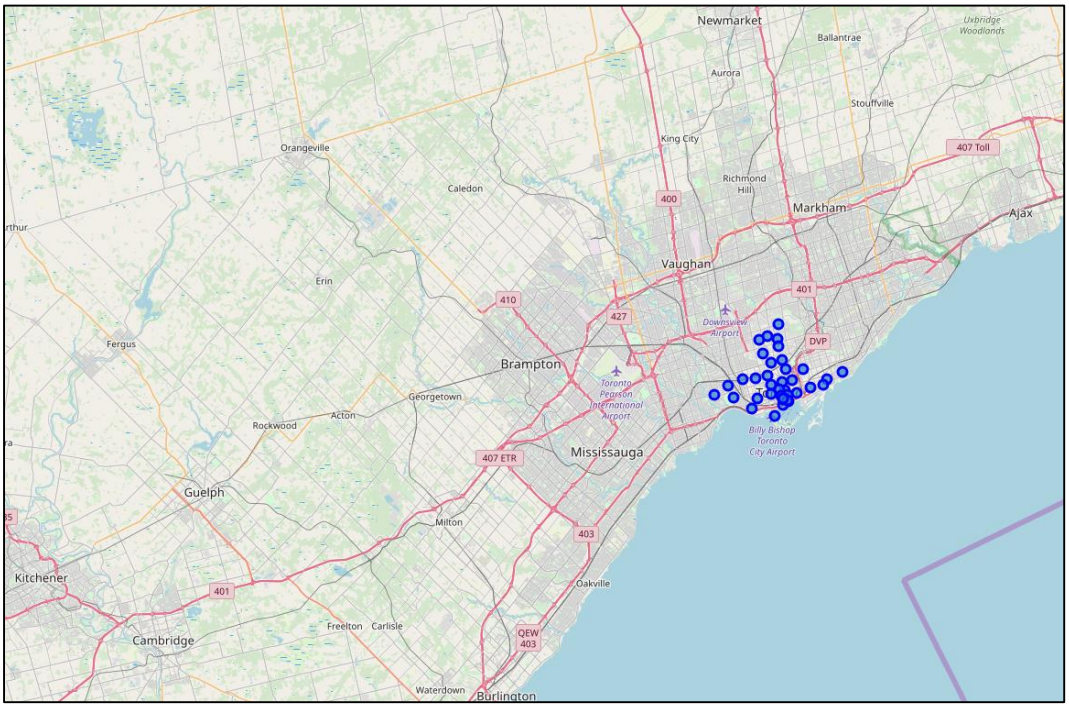

Figure.3

	PostalCode	Borough	Neighborhoods	Indonesian Restaurant
0	M4E	East Toronto	The Beaches	0.0
1	M4K	East Toronto	The Danforth West, Riverdale	0.0
2	M4L	East Toronto	The Beaches West, India Bazaar	0.0
3	M4M	East Toronto	Studio District	0.0
4	M4N	Central Toronto	Lawrence Park	0.0

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Indonesian Restaurant
28	M5W	Downtown Toronto	Stn A PO Boxes 25 The Esplanade	43.646435	-79.374846	0	0.00
27	M5V	Downtown Toronto	CN Tower, Bathurst Quay, Island airport, Harbo...	43.628947	-79.394420	0	0.00
19	M5J	Downtown Toronto	Harbourfront East, Toronto Islands, Union Station	43.640816	-79.381752	0	0.00
16	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	0	0.00
23	M5P	Central Toronto	Forest Hill North, Forest Hill West	43.696948	-79.411307	1	0.01
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790	1	0.00
5	M4P	Central Toronto	Davisville North	43.712751	-79.390197	1	0.01
6	M4R	Central Toronto	North Toronto West	43.715383	-79.405678	1	0.01
7	M4S	Central Toronto	Davisville	43.704324	-79.388790	1	0.01
22	M5N	Central Toronto	Roselawn	43.711695	-79.416936	1	0.01
0	M4E	East Toronto	The Beaches	43.676357	-79.293031	2	0.00
37	M7Y	East Toronto	Business Reply Mail Processing Centre 969 Eastern	43.662744	-79.321558	2	0.00
1	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188	2	0.00
2	M4L	East Toronto	The Beaches West, India Bazaar	43.668999	-79.315572	2	0.00
3	M4M	East Toronto	Studio District	43.659526	-79.340923	2	0.00
33	M6K	West Toronto	Brockton, Exhibition Place, Parkdale Village	43.636847	-79.428191	3	0.00
32	M6J	West Toronto	Little Portugal, Trinity	43.647927	-79.419750	3	0.00
30	M6G	Downtown Toronto	Christie	43.669542	-79.422564	3	0.00
26	M5T	Downtown Toronto	Chinatown, Grange Park, Kensington Market	43.653206	-79.400049	3	0.00
25	M5S	Downtown Toronto	Harbord, University of Toronto	43.662696	-79.400049	3	0.00

Figure.4

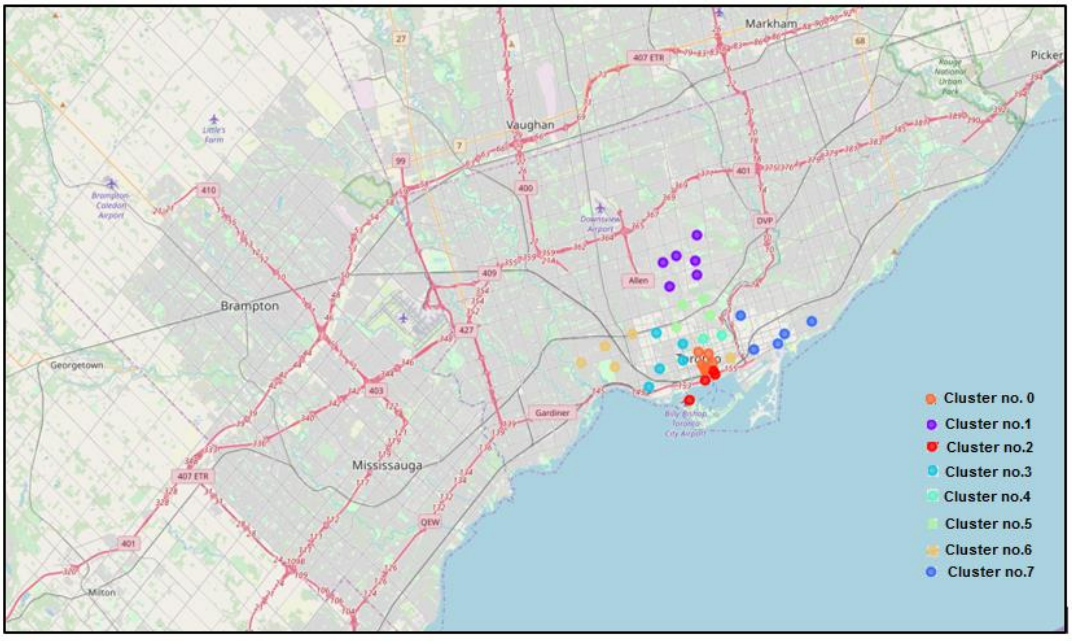

Figure.5

In [35]:
```
#cluster 1
toronto_merged.loc[toronto_merged['Cluster Labels'] == 0]
```

Out[35]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Indonesian Restaurant
16	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	0	0.0
19	M5J	Downtown Toronto	Harbourfront East, Toronto Islands, Union Station	43.640816	-79.381752	0	0.0
27	M5V	Downtown Toronto	CN Tower, Bathurst Quay, Island airport, Harbo...	43.628947	-79.394420	0	0.0
28	M5W	Downtown Toronto	Stn A PO Boxes 25 The Esplanade	43.646435	-79.374846	0	0.0

Cluster 1 ¶

In [64]:
```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 1]
```

Out[64]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Indonesian Restaurant
23	M5P	Central Toronto	Forest Hill North, Forest Hill West	43.696948	-79.411307	1	0.01
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790	1	0.00
5	M4P	Central Toronto	Davisville North	43.712751	-79.390197	1	0.01
6	M4R	Central Toronto	North Toronto West	43.715383	-79.405678	1	0.01
7	M4S	Central Toronto	Davisville	43.704324	-79.388790	1	0.01
22	M5N	Central Toronto	Roselawn	43.711695	-79.416936	1	0.01

Cluster 2

In [65]:
```
toronto_merged.loc[toronto_merged['Cluster Labels']== 2]
```

Out[65]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Indonesian Restaurant
0	M4E	East Toronto	The Beaches	43.676357	-79.293031	2	0.0
37	M7Y	East Toronto	Business Reply Mail Processing Centre 969 Eastern	43.662744	-79.321558	2	0.0
1	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188	2	0.0
2	M4L	East Toronto	The Beaches West, India Bazaar	43.668999	-79.315572	2	0.0
3	M4M	East Toronto	Studio District	43.659526	-79.340923	2	0.0

Cluster 3

In [66]:
```
toronto_merged.loc[toronto_merged['Cluster Labels']== 3]
```

Out[66]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Indonesian Restaurant
33	M6K	West Toronto	Brockton, Exhibition Place, Parkdale Village	43.636847	-79.428191	3	0.0
32	M6J	West Toronto	Little Portugal, Trinity	43.647927	-79.419750	3	0.0
30	M6G	Downtown Toronto	Christie	43.669542	-79.422564	3	0.0
26	M5T	Downtown Toronto	Chinatown, Grange Park, Kensington Market	43.653206	-79.400049	3	0.0
25	M5S	Downtown Toronto	Harbord, University of Toronto	43.662696	-79.400049	3	0.0

Cluster 4

In [67]:
```
toronto_merged.loc[toronto_merged['Cluster Labels']== 4]
```

Out[67]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Indonesian Restaurant
12	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	4	0.0
11	M4X	Downtown Toronto	Cabbagetown, St. James Town	43.667967	-79.367675	4	0.0

Cluster 5

In [68]:
```
toronto_merged.loc[toronto_merged['Cluster Labels']== 5]
```

Out[68]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Indonesian Restaurant
24	M5R	Central Toronto	The Annex, North Midtown, Yorkville	43.672710	-79.405678	5	0.00
9	M4V	Central Toronto	Deer Park, Forest Hill SE, Rathnelly, South Hi...	43.686412	-79.400049	5	0.01
10	M4W	Downtown Toronto	Rosedale	43.679563	-79.377529	5	0.00
8	M4T	Central Toronto	Moore Park, Summerhill East	43.689574	-79.383160	5	0.01

Cluster 6

```
In [69]: toronto_merged.loc[toronto_merged['Cluster Labels']== 6]
```

Out[69]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Indonesian Restaurant
36	M6S	West Toronto	Runnymede, Swansea	43.651571	-79.484450	6	0.0
31	M6H	West Toronto	Dovercourt Village, Dufferin	43.669005	-79.442259	6	0.0
13	M5A	Downtown Toronto	Harbourfront, Regent Park	43.654260	-79.360636	6	0.0
34	M6P	West Toronto	High Park, The Junction South	43.661608	-79.464763	6	0.0
35	M6R	West Toronto	Parkdale, Roncesvalles	43.648960	-79.456325	6	0.0

Cluster 7

```
In [70]: toronto_merged.loc[toronto_merged['Cluster Labels']== 7]
```

Out[70]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Indonesian Restaurant
21	M5L	Downtown Toronto	Commerce Court, Victoria Hotel	43.648198	-79.379817	7	0.0
20	M5K	Downtown Toronto	Design Exchange, Toronto Dominion Centre	43.647177	-79.381576	7	0.0
17	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383	7	0.0
29	M5X	Downtown Toronto	First Canadian Place, Underground city	43.648429	-79.382280	7	0.0
15	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	7	0.0
14	M5B	Downtown Toronto	Ryerson, Garden District	43.657162	-79.378937	7	0.0
18	M5H	Downtown Toronto	Adelaide, King, Richmond	43.650571	-79.384568	7	0.0

Figure.6

5. Discussion

This model will be great if there is additional data of numbers Indonesian immigrant in each Toronto neighborhoods and their ethics. So, the model will be more accurate and what type of Indonesian Restaurant can be configure like Javanese, Sundanese, Bataknese, Medan, or Padang Restaurant.

6. Conclusion

Most of the Indonesian Restaurant are concentrated in the central Toronto, at cluster 1 and cluster 5 with the same number. Almost every neighborhood at cluster 1 has indonesian restaurant, while cluster 5 is only 2 neighborhood. So, it will be great opportunity to open the restaurant in other clusters because it doesn't exist yet. The cluster selected should be not too close to cluster 1 and 5 example cluster no.0, no.2, and no.6, so the businessman should go south area instead of towards the middle area. Lastly, businessman are advised to avoid neighborhoods in cluster 1 and 5 which already have Indonesian Restaurant and suffering from intense competition.