Capstone Project: Battle of the Neighbourhoods (Final Assignment IBM Data Science Course)

Table of contents

- 1. Introduction: Business Problem
- 2. Data
- 3. Methodology
- 4. Analysis
- 5. Results and Discussion
- 6. Conclusion

1. Introduction: Business Problem

In this project we will try to find an optimal location for an Indian restaurant. Specifically, this report will be targeted to stakeholders interested in opening an **Indian restaurant** in **Toronto**, Canada.

Since the highest population of Indian people in Canada is in Toronto, here is on the one hand the most potential guests but on the other hand there is also the highest competition. Cause there are lots of restaurants in Toronto we will try to detect **locations that are not already crowded with Indian restaurants, but has a high population of Indian people.** Thus we expect that this is a smart business plan.

In this analysiy we try to generate a few most promising neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

Target Audience:

Business personnel who wants to invest or open an Indian restaurant in Toronto. This analysis will be a comprehensive guide to start or expand restaurants targeting the Indian crowd. Freelancer who loves to have their own restaurant as a side business. This analysis will give an idea, how beneficial it is to open a restaurant and what are the pros and cons of this business. Indian crowd who wants to find neighborhoods with lots of option for Indian restaurants.

2. Data

2.1 Data Sources

a) Most important are data about the neighborhoods in Toronto. Therefore we use the Wikipedia page “List of Postal code of Canada: M” (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M), where are all neighborhoods and boroughs listed.

b) In addition we need the longitudes and latitudes of the neighborhoods. The geographical coordinates we find in a csv file here: “https://cocl.us/Geospatial_data”

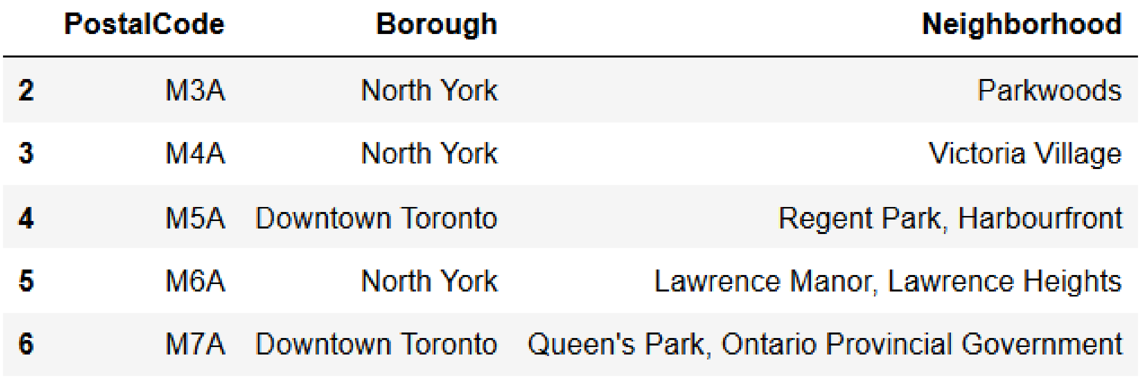
c) As described above we need information about the distribution of indian population in Toronto. The Wikipedia page “Demographics of Toronto” (https://en.m.wikipedia.org/wiki/Demographics_of_Toronto#Ethnic_diversity) is appropriated. Using this page I’m going to identify the neighborhoods which are densely populated with Indians as it might be helpful in identifying the suitable neighborhood to open a new Indian restaurant.

d) Furthermore we need the number of Indian restaurants in the neighborhoods. This information we can get by using the Foursquare’s explore API. From Foursquare API (https://developer.foursquare.com/docs), we retrieve name, category, latitude and longitude for each venue. Name: The name of the venue.

2.2 Data Cleaning and Wrangling

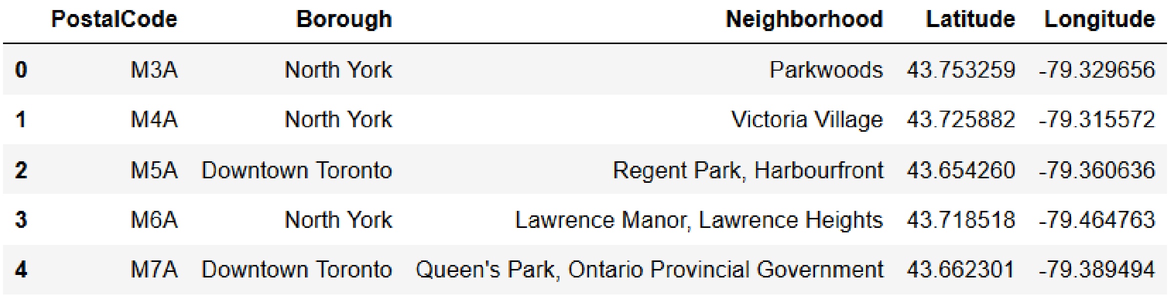
Neighborhoods of Toronto

First we web scraping the data from the Wikipedia Website “List of Postal code of Canada: M” and creating the dataframe. The Dataframe will consist of three columns: PostalCode, Borough, and Neighborhood Only the cells that have an assigned borough will be processed. Borough that is not assigned are ignored. More than one neighborhood can exist in one postal code area. If a borough but a no assigned neighborhood, then the neighborhood will be the same as the borough.
As a result, it looks like this:



	PostalCode	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

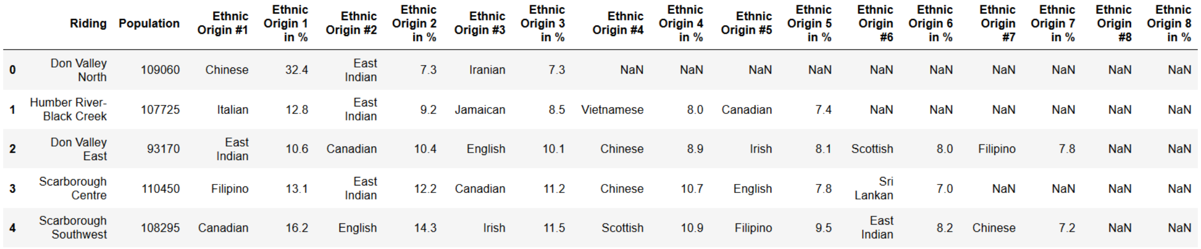
Next we combine the coordinates tot he neighborhoods by using a csv file:



	PostalCode	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

Indian Population in Toronto

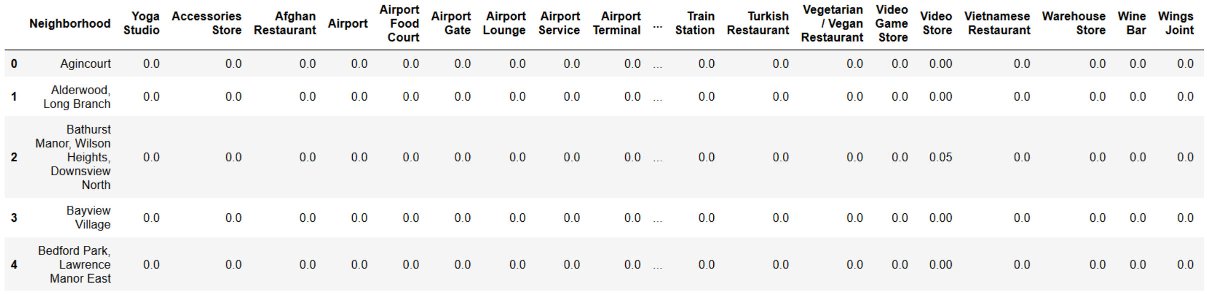
Next we extract the information about the population in each riding of Toronto from the wikipedia page “Demographics of Toronto”. Later we will assgin the ridings to the neighborhoods.



	Riding	Population	Ethnic Origin #1	Ethnic Origin 1 in %	Ethnic Origin #2	Ethnic Origin 2 in %	Ethnic Origin #3	Ethnic Origin 3 in %	Ethnic Origin #4	Ethnic Origin 4 in %	Ethnic Origin #5	Ethnic Origin 5 in %	Ethnic Origin #6	Ethnic Origin 6 in %	Ethnic Origin #7	Ethnic Origin 7 in %	Ethnic Origin #8	Ethnic Origin 8 in %
0	Don Valley North	109060	Chinese	32.4	East Indian	7.3	Iranian	7.3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Humber River-Black Creek	107725	Italian	12.8	East Indian	9.2	Jamaican	8.5	Vietnamese	8.0	Canadian	7.4	NaN	NaN	NaN	NaN	NaN	NaN
2	Don Valley East	93170	East Indian	10.6	Canadian	10.4	English	10.1	Chinese	8.9	Irish	8.1	Scottish	8.0	Filipino	7.8	NaN	NaN
3	Scarborough Centre	110450	Filipino	13.1	East Indian	12.2	Canadian	11.2	Chinese	10.7	English	7.8	Sri Lankan	7.0	NaN	NaN	NaN	NaN
4	Scarborough Southwest	108295	Canadian	16.2	English	14.3	Irish	11.5	Scottish	10.9	Filipino	9.5	East Indian	8.2	Chinese	7.2	NaN	NaN

Explore Neighborhoods in Toronto

Let us get the venues in Toronto by using Fourquare API.



	Neighborhood	Yoga Studio	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	...	Train Station	Turkish Restaurant	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnamese Restaurant	Warehouse Store	Wine Bar	Wings Joint
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0
1	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0
2	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.05	0.0	0.0	0.0	0.0
3	Bayview Village	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0
4	Bedford Park, Lawrence Manor East	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0

3. Methodology

In this project we will direct our efforts on detecting areas of Toronto that have low restaurant density, particularly those with low number of Indian restaurants.

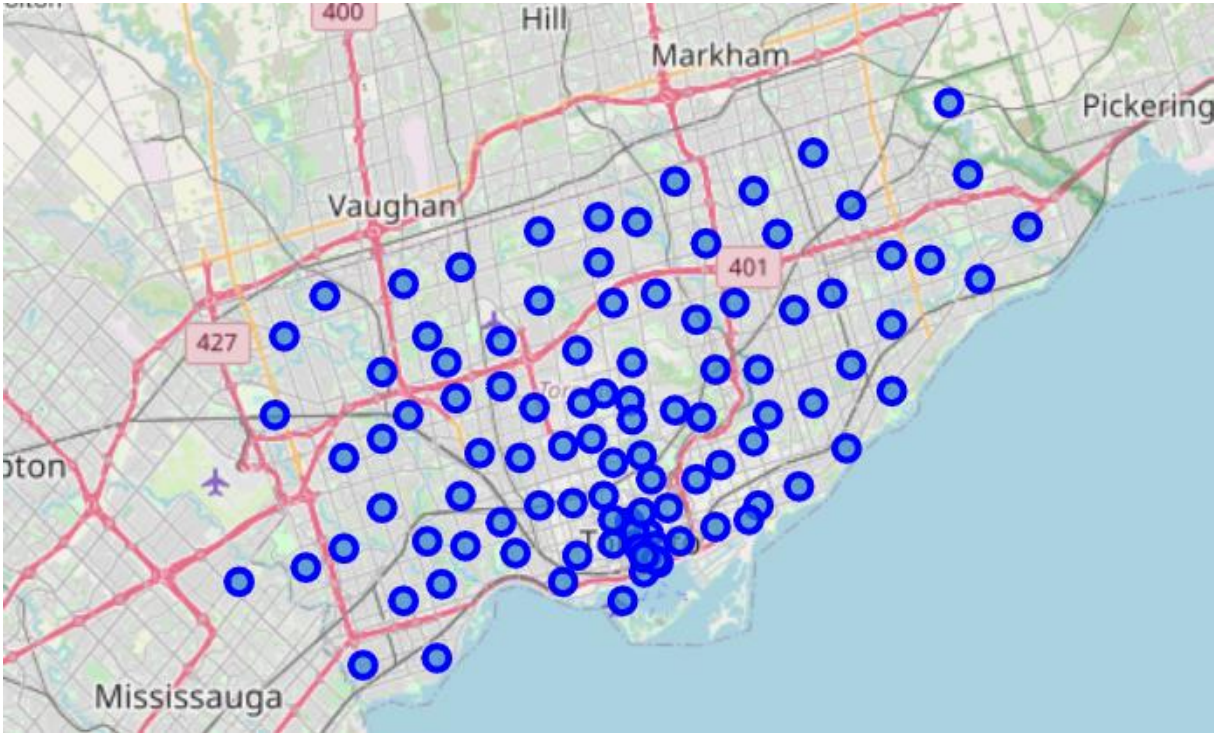
In first step we have collected the required data: We have collected all data about the neighborhoods, especially names and longitudes/latitudes. Next we have collected the population data of the ridings of Toronto. Cause we are interested in neighborhoods and not in ridings we have to match the ridings to the neighborhoods they belong to. The location and type (category) of every venue we can find in a radius of 500m from the centre of each neighborhood using Foursquare.

Second step in our analysis will be calculation and exploration of 'Indian restaurant density' across Torontos neighborhoods. We do this by creating a map, where the location of restaurants near to each other are grouped. Additionally, we take a look in the neighborhoods with Indian population. Cause this neighborhoods we are interested in, assuming that there a lot of potential guests for our restaurant.

In third and final step we will create clusters of locations that meet some basic requirements established in discussion with stakeholders: we will take into consideration locations with high Indian population, but less restaurants and search for optimal location by stakeholders. We use kmean-clustering, cause we have unlabeled data. The optimal number of clusters we determine by using the elbow method. To examine the clusters we take a deeper look into the resulting clusters. Then we decide what locations we can recommend for opening an Indian restaurant.

4. Analysis

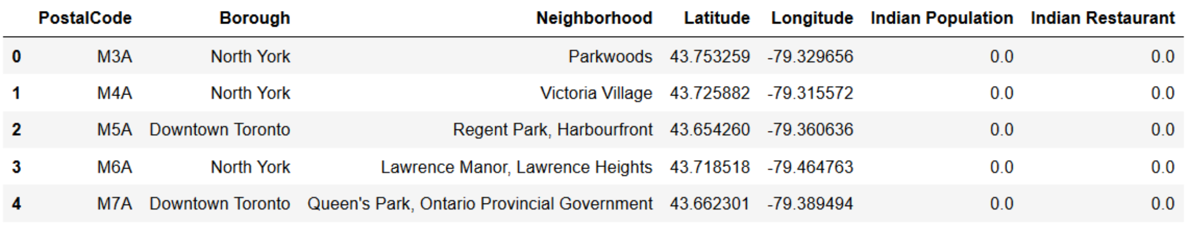
Create map of Toronto neighborhoods



Picture 1: Neighborhoods of Toronto

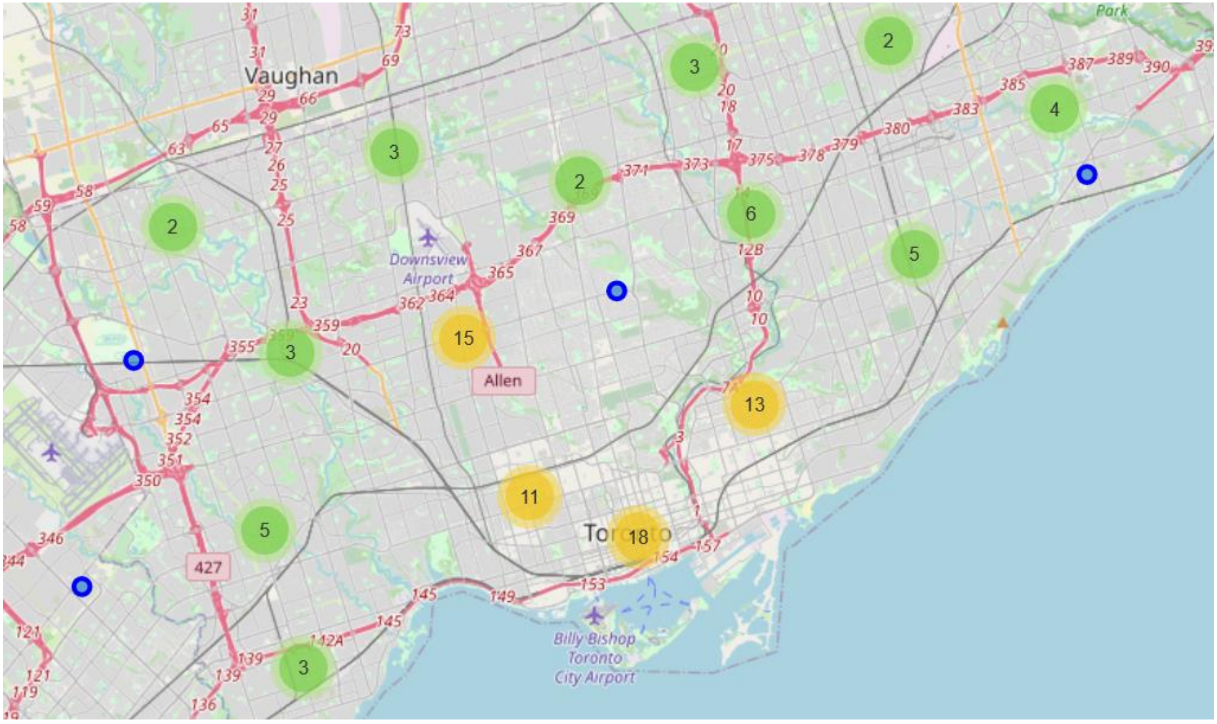
Relationship between neighborhood and Indian restaurants

First we will extract the Neighborhood and Indian Restaurant column from the above toronto dataframe for further analysis and merge it with the neighborhood dataframe:



	PostalCode	Borough	Neighborhood	Latitude	Longitude	Indian Population	Indian Restaurant
0	M3A	North York	Parkwoods	43.753259	-79.329656	0.0	0.0
1	M4A	North York	Victoria Village	43.725882	-79.315572	0.0	0.0
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	0.0	0.0
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	0.0	0.0
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	0.0	0.0

For identifying the dense of Indian restaurants in the neighborhoods we visualize the distribution of Indian Neighborhoods in the following map.

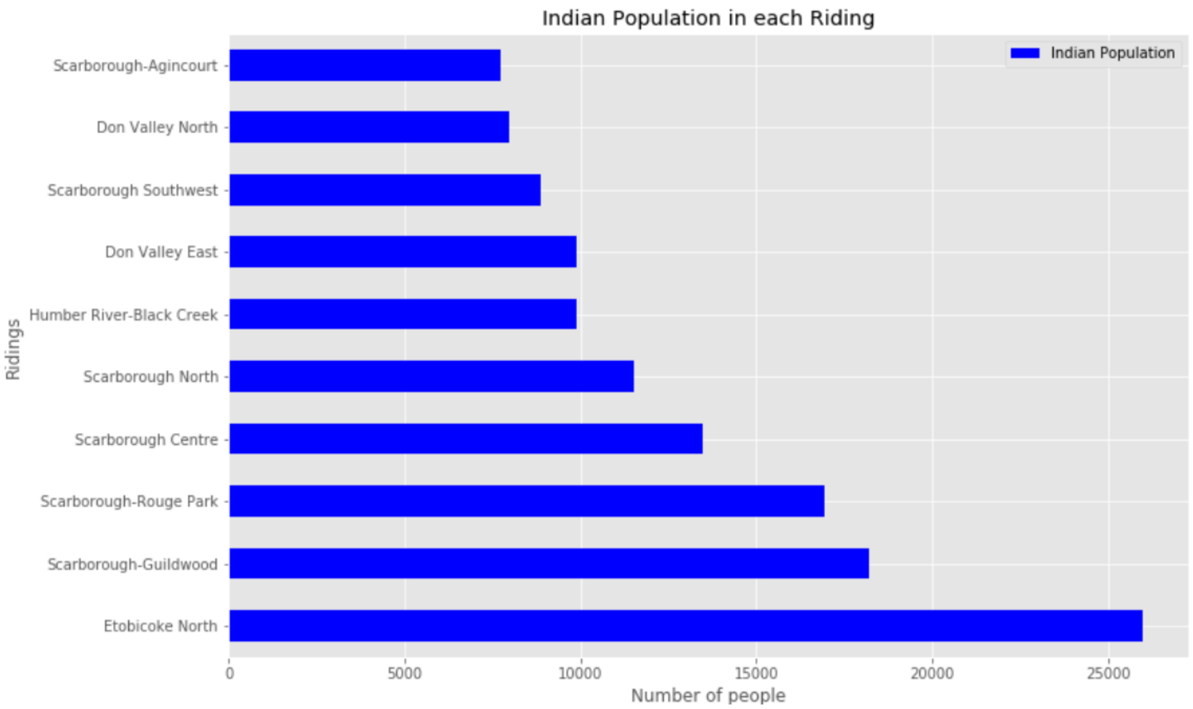


Picture 2: Dense of Indian restaurants in Toronto

As we can see there is a higher dense of Indian restaurants in the centre of Toronto, whereas the other areas has a much smaller dense.

Relationship between neighborhood and Indian population

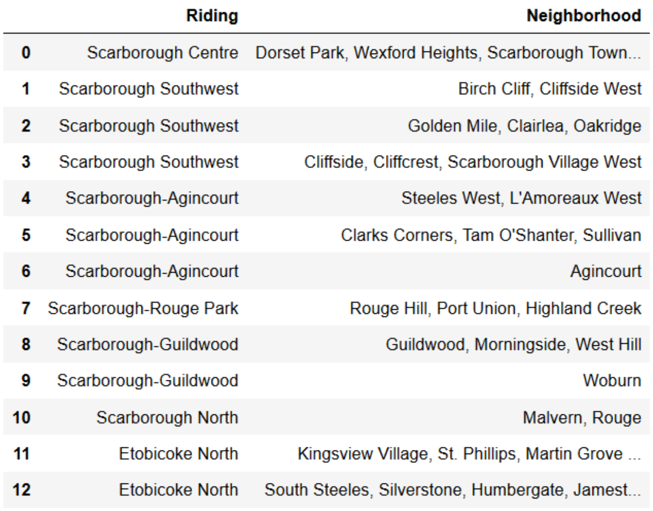
All in alll there are 130k Indian people living in Toronto. But where exactly do they live? We create a bar chart (see below) to visualize the number of Indian people in the ridings. In all other ridings there is no Indian population. Don't forget later we assign each riding to the according neighborhood.Accordingly to our business idea, these ridings are potential locations for our Indian restaurants. But to make a smart decision we first have to connect the population datawith the dense of Indian restaurant data.



Picture 3: Indian population in each riding in Toronto

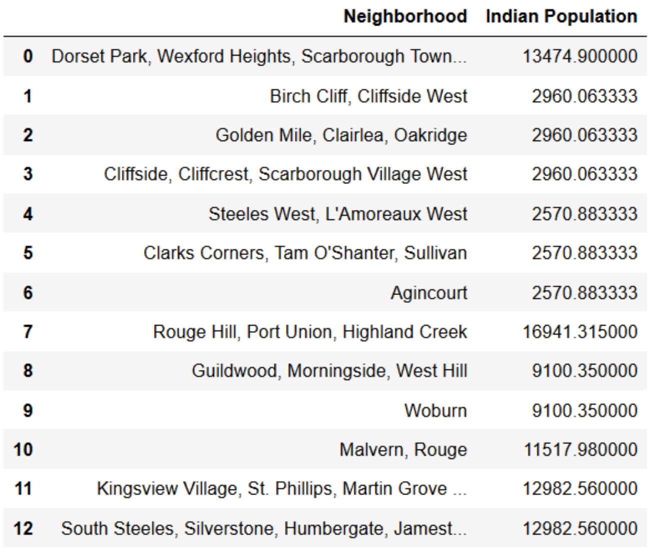
Relationship between Indian poplation and Indian restaurant

Cause the population dataset is assigned to Ridings, but our other datasets use Neighborhoods, we need to replace the Ridings with the belonging Neighboorhoods. Cause there are more than one neighborhood assigned to a riding, we have to correct the population data. We can do this by dividing the popu lation data with the number of neighborhoods in a riding.



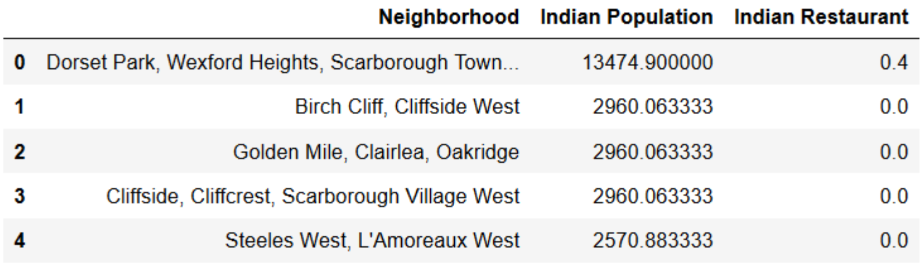
	Riding	Neighborhood
0	Scarborough Centre	Dorset Park, Wexford Heights, Scarborough Town...
1	Scarborough Southwest	Birch Cliff, Cliffside West
2	Scarborough Southwest	Golden Mile, Clairlea, Oakridge
3	Scarborough Southwest	Cliffside, Cliffcrest, Scarborough Village West
4	Scarborough-Agincourt	Steeles West, L'Amoreaux West
5	Scarborough-Agincourt	Clarks Corners, Tam O'Shanter, Sullivan
6	Scarborough-Agincourt	Agincourt
7	Scarborough-Rouge Park	Rouge Hill, Port Union, Highland Creek
8	Scarborough-Guildwood	Guildwood, Morningside, West Hill
9	Scarborough-Guildwood	Woburn
10	Scarborough North	Malvern, Rouge
11	Etobicoke North	Kingsview Village, St. Phillips, Martin Grove ...
12	Etobicoke North	South Steeles, Silverstone, Humbergate, Jamest...

And here we see the adapted population of each of these neighborhood:



	Neighborhood	Indian Population
0	Dorset Park, Wexford Heights, Scarborough Town...	13474.900000
1	Birch Cliff, Cliffside West	2960.063333
2	Golden Mile, Clairlea, Oakridge	2960.063333
3	Cliffside, Cliffcrest, Scarborough Village West	2960.063333
4	Steeles West, L'Amoreaux West	2570.883333
5	Clarks Corners, Tam O'Shanter, Sullivan	2570.883333
6	Agincourt	2570.883333
7	Rouge Hill, Port Union, Highland Creek	16941.315000
8	Guildwood, Morningside, West Hill	9100.350000
9	Woburn	9100.350000
10	Malvern, Rouge	11517.980000
11	Kingsview Village, St. Phillips, Martin Grove ...	12982.560000
12	South Steeles, Silverstone, Humbergate, Jamest...	12982.560000

To finish our dataframe for the analysis we combine all data for our decision in one dataframe. So we have to add the Indian Restaurant column.



	Neighborhood	Indian Population	Indian Restaurant
0	Dorset Park, Wexford Heights, Scarborough Town...	13474.900000	0.4
1	Birch Cliff, Cliffside West	2960.063333	0.0
2	Golden Mile, Clairlea, Oakridge	2960.063333	0.0
3	Cliffside, Cliffcrest, Scarborough Village West	2960.063333	0.0
4	Steeles West, L'Amoreaux West	2570.883333	0.0

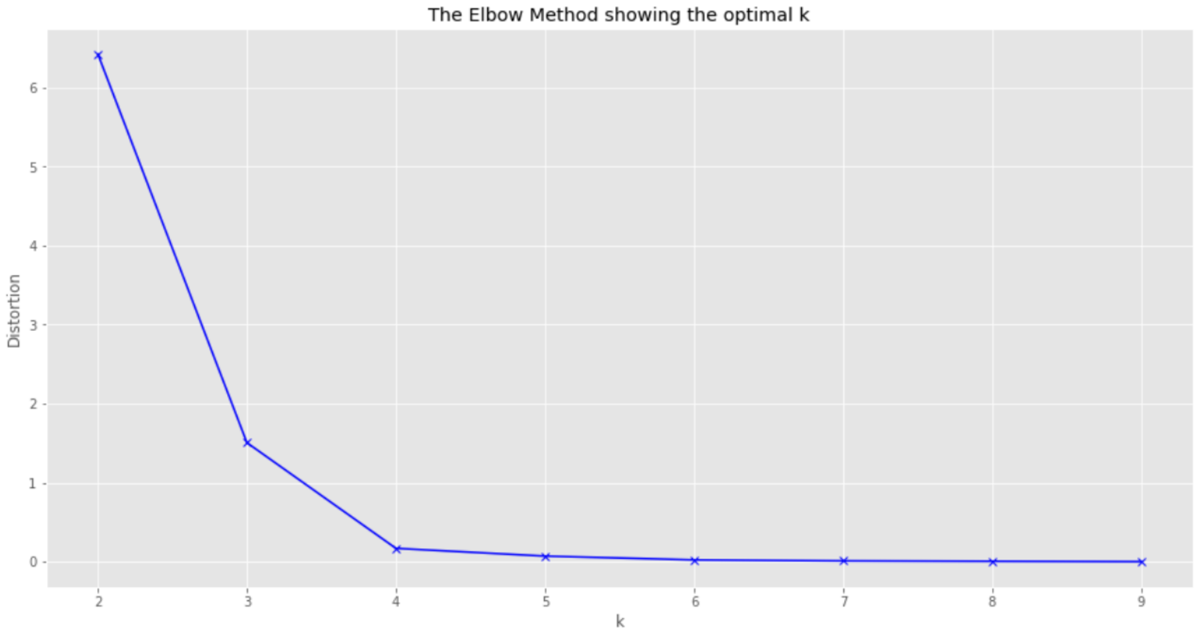
Unfortunately most of the neighborhoods have not an Indian community and also a Indian restaurants, so we have not enough information to infer a connection between both variables. On the other hand that are good news for our business plan, cause that means we have less competition in the neighborhoods with high dense of Indian people.

Predictive Modeling

In the predictive modelling we are going to use clustering techniques since this is analysis of unlabelled data. K-Means clustering is used to perform the analysis of the data

Clustering Neighborhoods of Toronto:

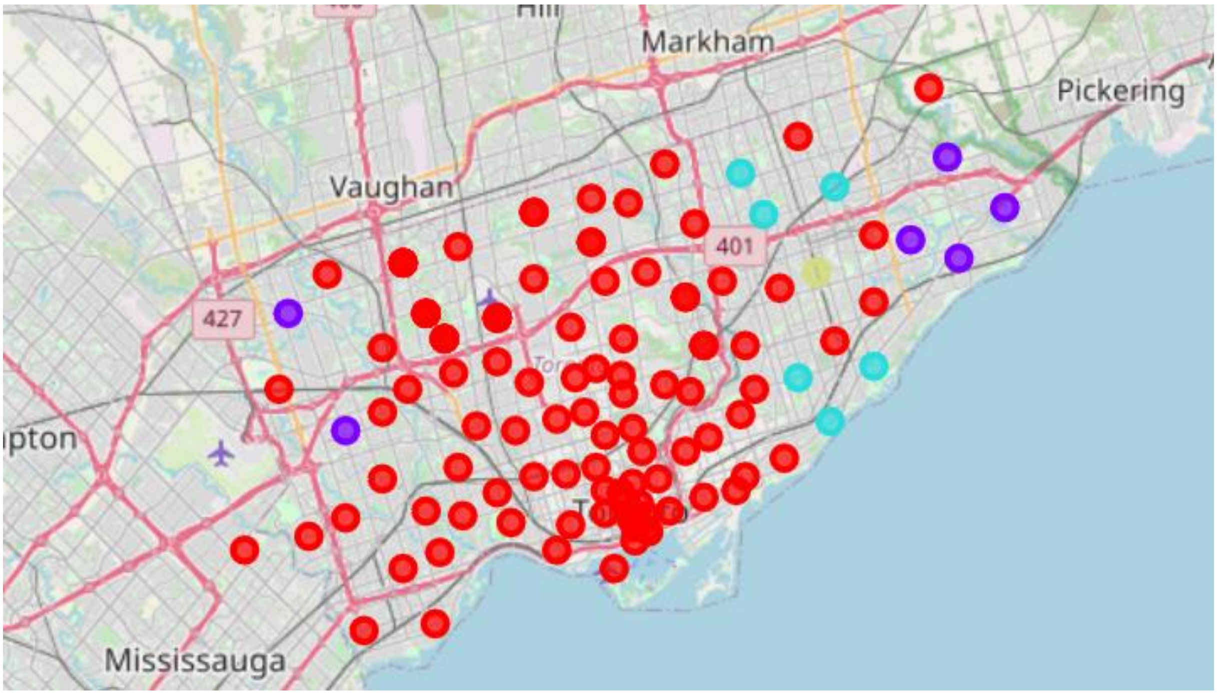
First step in K-means clustering is to identify best K value meaning the number of clusters in a given dataset. To do so we are going to use the elbow method on the Toronto dataset with Indian restaurant percentage.



Picture 4: Choosing the optimal k with the elbow method

After analysing using elbow method using distortion score for each K value, looks like K = 4 is the best value.

Clustering the Toronto neighborhoods leads to following map:

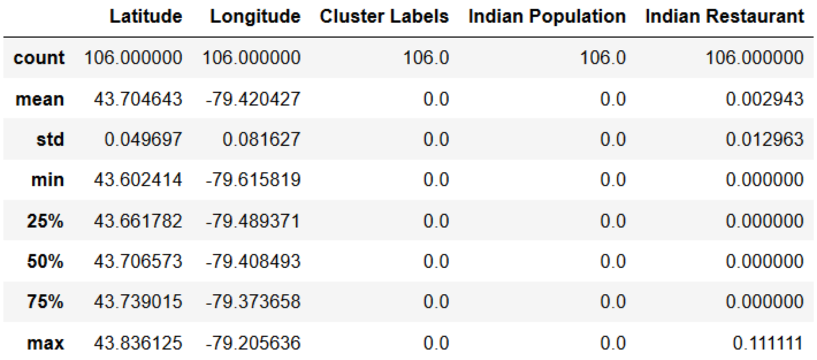


Picture 5: Clusters of Toronto

Now would like to take closer look to the clusters and determine which cluster is appropiated for our plan to open an Indian restaurant.

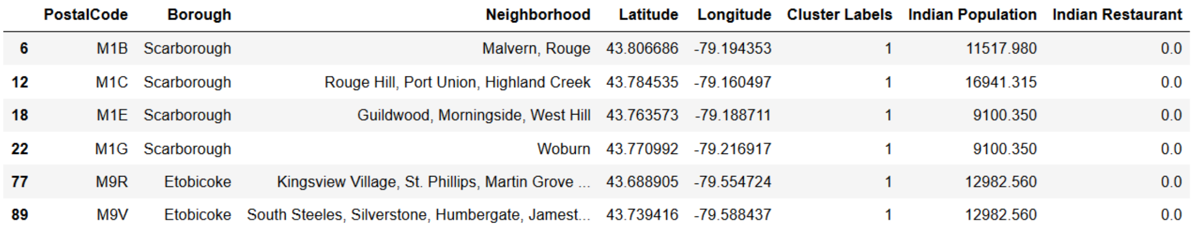
Examing the Clusters:

Cluster 0: contains all the neighborhoods which has no Indian Population and sparse number of Indian restaurants. It is shown in red color in the map and represents about 75% of all neighborhoods.



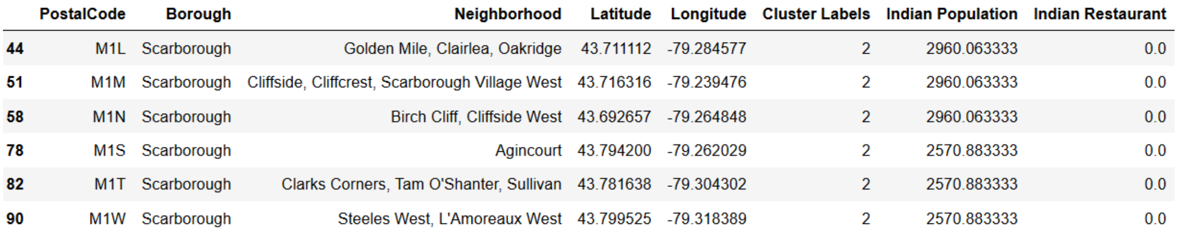
	Latitude	Longitude	Cluster Labels	Indian Population	Indian Restaurant
count	106.000000	106.000000	106.0	106.0	106.000000
mean	43.704643	-79.420427	0.0	0.0	0.002943
std	0.049697	0.081627	0.0	0.0	0.012963
min	43.602414	-79.615819	0.0	0.0	0.000000
25%	43.661782	-79.489371	0.0	0.0	0.000000
50%	43.706573	-79.408493	0.0	0.0	0.000000
75%	43.739015	-79.373658	0.0	0.0	0.000000
max	43.836125	-79.205636	0.0	0.0	0.111111

Cluster 1: contains all the neighborhoods which has high Indian Population but no Indian restaurants. It is shown in purple color in the map.



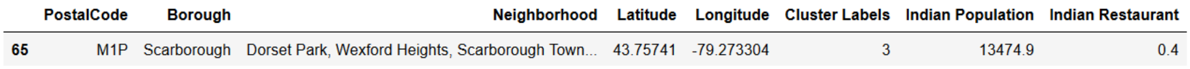
	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Population	Indian Restaurant
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353	1	11517.980	0.0
12	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	1	16941.315	0.0
18	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	1	9100.350	0.0
22	M1G	Scarborough	Woburn	43.770992	-79.216917	1	9100.350	0.0
77	M9R	Etobicoke	Kingsview Village, St. Phillips, Martin Grove ...	43.688905	-79.554724	1	12982.560	0.0
89	M9V	Etobicoke	South Steeles, Silverstone, Humbergate, Jamest...	43.739416	-79.588437	1	12982.560	0.0

Cluster 2: contains all the neighborhoods which has a small Indian Population but no Indian restaurants. It is shown in blue color in the map.



	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Population	Indian Restaurant
44	M1L	Scarborough	Golden Mile, Clairlea, Oakridge	43.711112	-79.284577	2	2960.063333	0.0
51	M1M	Scarborough	Cliffside, Cliffcrest, Scarborough Village West	43.716316	-79.239476	2	2960.063333	0.0
58	M1N	Scarborough	Birch Cliff, Cliffside West	43.692657	-79.264848	2	2960.063333	0.0
78	M1S	Scarborough	Agincourt	43.794200	-79.262029	2	2570.883333	0.0
82	M1T	Scarborough	Clarks Corners, Tam O'Shanter, Sullivan	43.781638	-79.304302	2	2570.883333	0.0
90	M1W	Scarborough	Steeles West, L'Amoreaux West	43.799525	-79.318389	2	2570.883333	0.0

Cluster 3: contains the neighborhoods which has the highest Indian Population and a high dense of Indian restaurants. It is shown in green color in the map.



	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Population	Indian Restaurant
65	M1P	Scarborough	Dorset Park, Wexford Heights, Scarborough Town...	43.75741	-79.273304	3	13474.9	0.4

Results and Discussion:

Results:

In this report we tried to recommend neighborhoods for opening a succesful Indian restaurant in Toronto. We have dicovered that the 130k Indian people in Toronto live in 13 neighborhoods. That means they not very spread out in the city. It is more than there are strong communities. Assuming that Indian people prefer Indian food, this areas we should prefer for opening our restaurant. These neighborhoods all are located in Scarborough and Etobicoke. The neighborhood with most Indian restaurant is in Scarborough, but the most other neighborhoods with Indian restaurants are in the central area of Toronto.

By clustering the neighborhoods we have identify 4 clusters. More than 75% of the neighborhoods has no Indian population and only a sparse amount of Indian restaurants. This areas are not very interisting for our plans. Among the group of neighbourhoods with a high density of Indian people, one neighbourhood with a very high density of Indian restaurants stands out. This neighborhood forms a separate cluster and is also not interisting for us. But therefore the other neighborhoods with high amount of Indian people has no Indian restaurants yet. Thus this are the areas we should prefer! So we can recommend neighborhoods in Scarborough and Etobicoke. Our favourite are the neighborhoods Rouge Hill, Port Union and Highland Creek, cause here we can find most Indian people. Here we find the best opportunities for our business: Most Indian people, No Indian Restaurants. Cluster 3 contains all the neighborhoods which has a small Indian Population but no Indian restaurants, this areas would be only interesting for us, if we cannot open a restaurant in the preferable neighborhoods cause of any reason.

Discussion:

According to this analysis, Scarborough borough will provide least competition for the new upcoming Indian restaurant as there is very little Indian restaurants spread or no Indian restaurants in neighborhoods. Also looking at the population distribution looks like it is densely populated with Indian crowd which helps the new restaurant by providing high customer visit possibilty. So, definitely this region could potentially be a perfect place for starting a quality Indian restaurants. Some of the drawbacks of this analysis are — the clustering is completely based only on data obtained from Foursquare API. Also the Indian population distribution in each neighborhood is also based on the 2016 census which is not up-to date. Thus population distribution would have definitely changed by 2019 given 3 years gap in the data. Since population distribution of Indian crowd in each neighborhood & number of Indian restaurants are the major feature in this analysis and it is not fully up-to date data, this analysis is definitely not far from being conclusory & it has lot of areas where it can be improved. In a further analysis it would be wise to include much more location data, e.g. from Foursquare. Unfortunately Foursquare queries are limited in the free account. However, it certainly provides us with some good insights, preliminary information on possibilites & a head start into this business problem by setting the step stones properly. Furthermore, this may also potentially vary depending on the type of clustering techniques that we use to examine the data.

Conclusion:

In a nutshel the result of our analysis is that we can highly recommend the neighborhoods Rouge Hill, Port Union and Highland Creek for opening an Indian restaurant in Toronto. Final decission on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.