Clustering Of Toronto Neighborhoods for an Investment Opportunity

Muhammad Zulqarnain

1 Problem description

Toronto as it the most populous city in Canada and the largest urban and metro area, with a population density of 4,149.5 people per square kilometer (10,750/sq mi) (1). The metro area of the city sprawls outward to a total surface area of 5,905.71 km2 (2,280.21 sq mi). Over the next 20 years, Toronto is expected to continue its moderate growth, surpassing 3 million by 2026, and reaching nearly 3.2 million in 2036. Foreign-born people account for nearly half of the population of Toronto. It is also reported that nearly 73% of the Canadians like to experience the ethnic food and flavors (2). There has been a recent increase the appreciation of the Afghan food(3). Keeping this statistics in mind company X is interested in opening a branch of their well-known Afghan restaurants in Toronto. They are looking for a suitable location to establish their new venture. They require some data-driven analysis of different neighborhoods to select an optimal location. The ideal location that they are interested in should be in a neighborhood popular for its eatery places and devoid of any existing Afghan restaurant. The company is also interested if they could be given more than one option, so that they can select an option by keeping in view of their economics. Let's see how data science can help in their selection process.

2 Background discussion (Abstract)

Data driven decision are becoming a norm in financial industry. Machine learning techniques can aid in making these decisions if enough and good quality data is available. In this study machine learning algorithm of *k*-mean clustering is applied for clustering of Toronto neighborhoods to identify a suitable location for a new Afghan restaurant. *k*-mean clustering is a type of unsupervised learning, mostly used for unlabeled data. The algorithm works iteratively to assign each data point to one of *k* groups based on the features that are provided. Data points are clustered based on feature similarity (4). Therefore, by using this algorithm we will be able to cluster neighborhoods with similar venues and may be able to identify more than one neighborhood suitable for the location of new restaurant. Initially the data of different boroughs is collected by scrapping a relevant webpage from Wikipedia. Then data wrangling, which involves data cleaning and transforming it to desired data frames is performed. In next step, the venue data for each of borough is collected from Foursquare and is transformed to data frames. Then top 20 venues data for each neighborhood are used in *k*-mean clustering algorithm. The resultant clusters are analyzed for good number of food venues but no Afghan restaurant. The results show that neighborhoods associated with M1E are some of the most suitable option to establish a new Afghan restaurant.

3 Data description and its usage

In this, a brief introduction of data and how it used to answer the required question is presented. Let's start we data description.

3.1 Data description

In order to cluster neighborhoods based on their popular venues, popular venues and their location-based data is required. In this project, we use the postcode as the basis to define a neighborhood. The data for the postcode and associated boroughs is obtained by scrapping relevant Wikipedia webpage (5). Data is cleaned and transformed to desired dataframes, as shown in Figure 1 below.

	Postcode	Borough	Neighbourhood
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

Figure 1: Some of the Postcodes, correspond boroughs and neigh hoods of Toronto

In next step geospatial, coordinates are assigned to these postcodes. These geospatial coordinates are used to obtain venues data from Foursquare (6). Venues are searched within the radius of 500 m form the geospatial location of these postcodes. Some of the venues associated with "The Beaches' neighborhood are shown in Figure 2.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
1	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub
2	The Beaches	43.676357	-79.293031	Starbucks	43.678798	-79.298045	Coffee Shop
3	The Beaches	43.676357	-79.293031	Upper Beaches	43.680563	-79.292869	Neighborhood
4	The Beaches	43.676357	-79.293031	Dip 'n Sip	43.678897	-79.297745	Coffee Shop

Figure 2: Venues associated with 'The Beaches' neighborhood

Now we see how we use this data to solve our problem to find an optimal location for establishing the new restaurant.

3.2 Data usage to solve problem

Now we have the venue data for each neighborhood that we have obtained from Foursquare. We initially recognized that clustering will be most suitable option for this problem. As it will provide us cluster of neighborhoods with similar characteristics. Then we will be able to identify the most suitable neighborhoods for new investment opportunity.

4 Methodology

In this section, the tools and technique of data analysis used to perform the data analysis are summarized.

4.1 Tools and techniques used

The major set of software and libraries used are summarized in Table 1. The detailed procedure adopted is presented in next section.

Table 1: Data science tools used in this work

Tool	Description	Usage in this project
ANACONDA	– open-source distribution of the Python and R	– used to run Jupyter notebook
Jupyter Notebook	– a server-client application that allows editing and running notebook documents via a web browser	– Main programming interface
Pandas	– to create dataframes	– To create dataframes of venues etc.
Numpy	– to deal with N-dimensional arrays	– to handle arrays
Matplot	– plotting library	– used to plot the number of clusters and error
Requests	– HTTP requests	– used to send requests for HTTP data
BeutifulSoup	– for scrapping web information	– used to scrap postal code data from Wikipedia
Geocoder	– to find geospatial coordinates of addresses	– used to get geospatial coordinates from Boroughs
Folium	– to create interactive maps	– created maps of Boroughs and clusters
Sikit-Learn	– machine learning library	– used k mean algorithm

4.2 Process flow diagram

We started with importing the relevant libraries in Jupyter Notebook. The first task was to get the postal code and boroughs information of the Toronto.

Data Scraping: This information was extracted from the relevant Wikipedia which has the postal code and neighborhood information of the Toronto. Requests and BeautuifulSoup libraries were used for this purpose.

Data Transformation: Then this raw data was transformed to a dataframe by using Pandas library. Dataframe display information revealed that the scrapped dataframe needs data wrangling to transform it to desired dataframes.

Data Wrangling: This step involved deleting the rows in which borough information was absent. In addition, missing neighborhood were assigned the same name as boroughs. Now we had the cleaned dataframe with postcode, borough and neighborhood information. We need to assign geospatial coordinates to these boroughs to be able to apply spatial operations like displaying them on map and finding the venues based on their geospatial coordinates.

Geospatial coordinate assignments: Geocoder library was used to find latitude and longitude based on address of these boroughs. Then this new information is appended to the current dataframe with postcodes and other information. Now we were ready to use Foursquare location service API.

Foursquare location API: Location based venues data based on the geospatial coordinates was obtained from Foursquare. Then this data was transformed to dataframes to be further processed. Top 10 most frequent venues in ach of the boroughs were extracted from this dataframe and used for k mean clustering algorithm.

***k* mean clustering**: *k* mean clustering algorithm from Sikit-learn machine learning library was implemented to cluster boroughs with similar characteristics.

The steps involved in data usage to solve the process are summarized in the in Figure 3. The results are presented in next section.

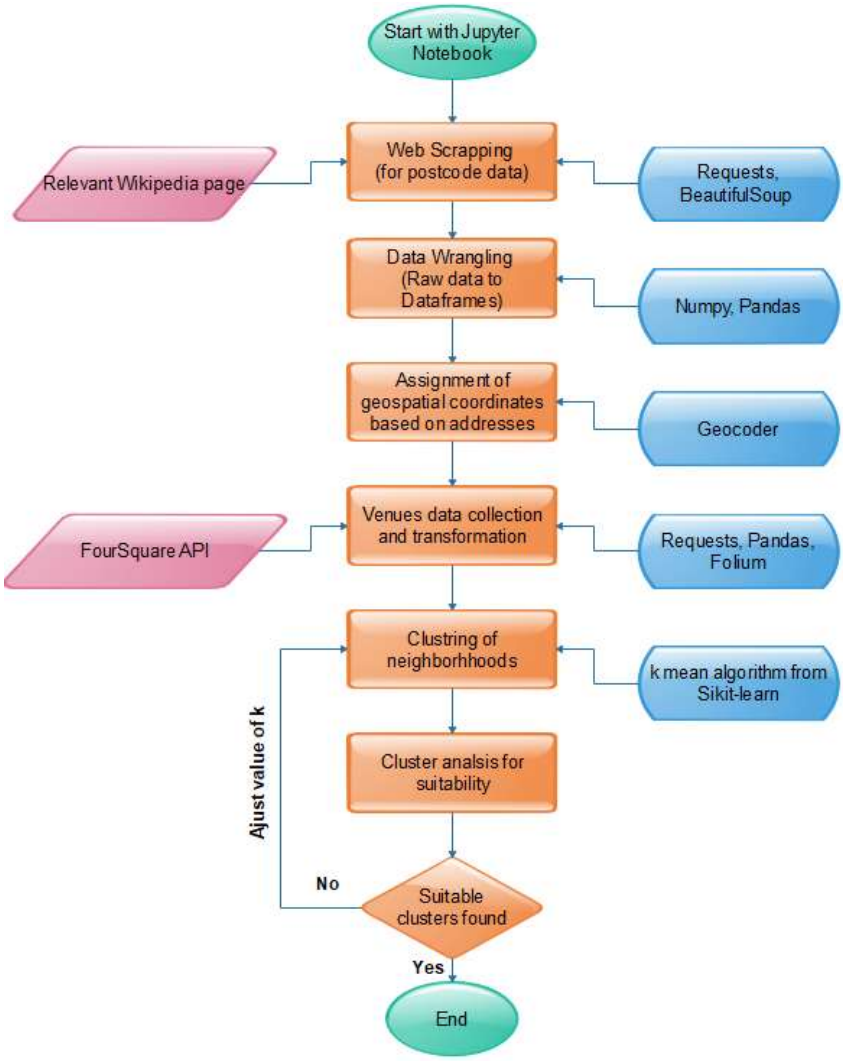


Figure 3: High-level process flow diagram of data usage

5 Results and Discussions

The Toronto boroughs are shown in Figure 4. Ninety-nine boroughs are used in clustering analysis.

Figure 4: Map of Toronto neighborhoods

As number of cluster is user input to the *k* mean clustering algorithm, we check the results sensitivity of the value of k by using the elbow method. Ideally, after a certain threshold value of k, the slope of the curve shown should become zero. However, we can observe in Figure 5, that slope does not become zero even when the number of clusters becomes equal to the number of boroughs. Therefore, in this case a reasonable value of 11 is selected, knowing that the clusters will not be unique and there will be some overlap of venues among different clusters..

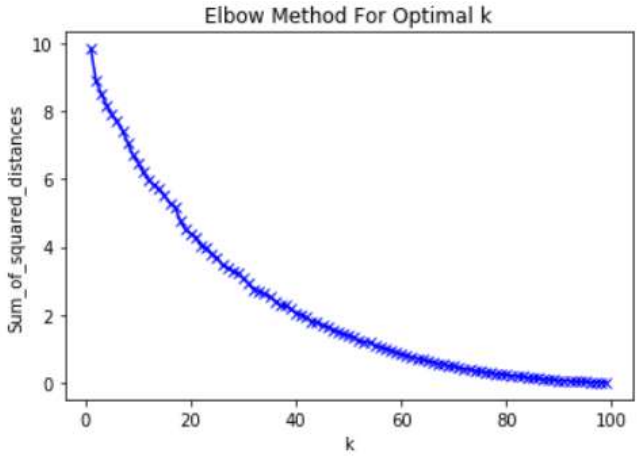


Figure 5: Plot of elbow method

The resultant clusters are shown in Figure 6. We can observe from the map that clustering facilitate in analyzing multiple locations with similar venues. Thus based on economics a suitable location for the restaurant can be selected.

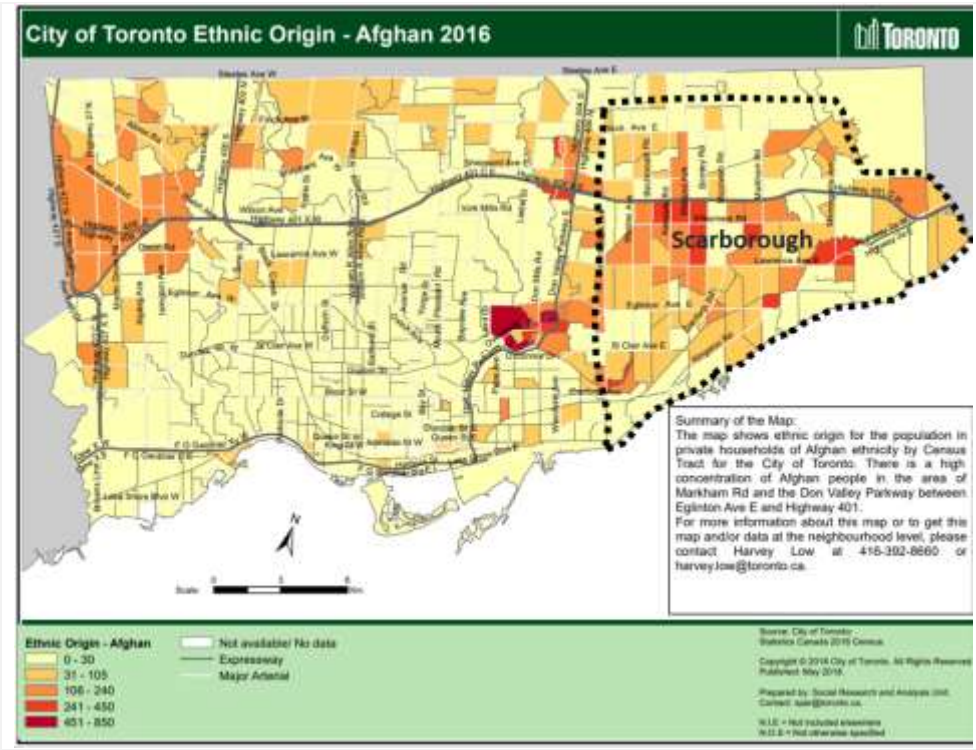
Figure 6: Cluster shown on the map

We now analyze unique venues of each cluster, so that the desired cluster is selected for further analysis. The cluster and their main venue type are summarized in Table 2 below. Cluste-1 has the desired features of having eatery venues as most frequent venues.

Table 2: Cluster and their prominent venues

Cluster	Cluster Label	Description	Borough
1	0	Eatery	Multiple
2	1	Mixed	Multiple
3	2	Mixed	West Toronto
4	3	Mixed	North York, Etobicoke, Scarborough
5	4	Mixed	Multiple
6	5	Mixed	Downtown
7	6	Mixed	West Toronto, York
8	7	Mixed	Downtown
9	8	Mixed	Downtown
10	9	Mixed	Downtown
11	10	Mixed	North York

Now we analyze the distribution population with Afghan origin in Toronto. A map of the population density of people with Afghan origin is shown in Figure 7 (a), with more concentration in Scarborough than other parts.



(a) Aghna origin ethinic densty (taken from (7))

(b) Existing Afgan Resturants in Scarborough (taken from (8)

Figure 7: (a)Map of Aghna origin ethinic densty, (b) Existing Afgan Resturants

We also display the exiting Afghan restaurant locations, shown in Figure 7 (b). Scarborough neighborhoods Guildwood, Morningside, West Hill, that are associated with postcode M1E seems to be one of the many suitable places amongst the cluster-1 associated areas, shown in

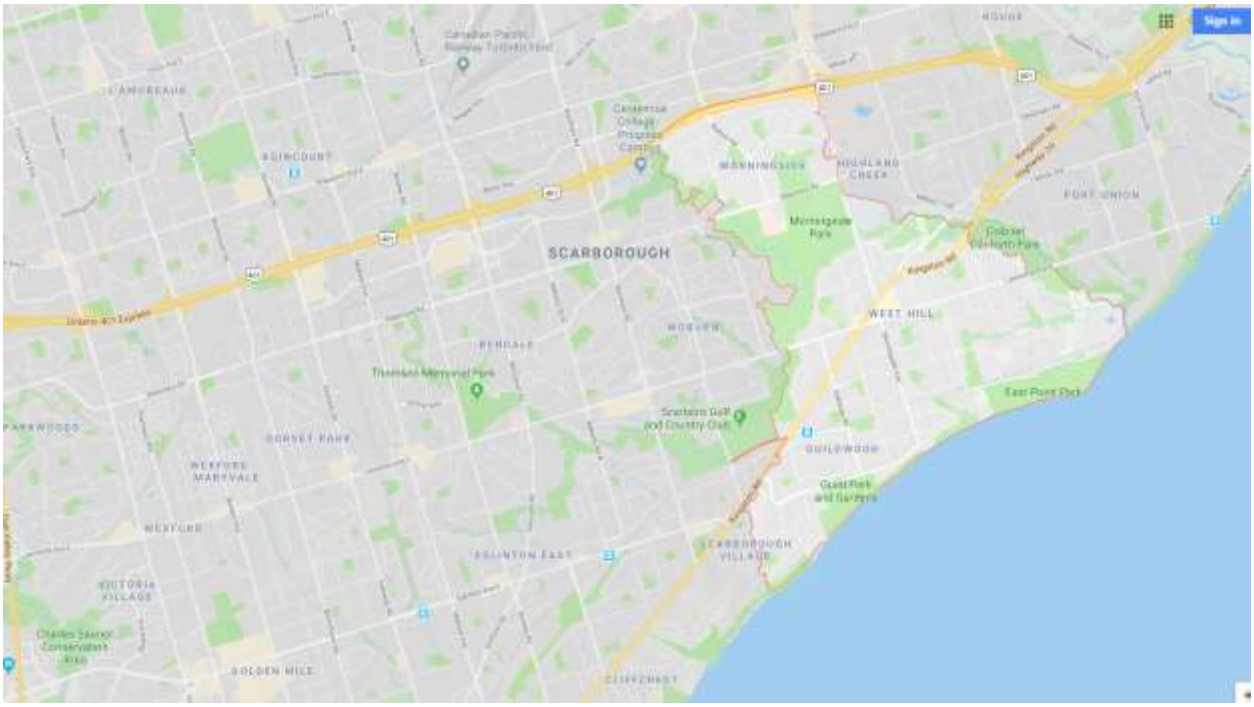


Figure 8: Neighborhoods associated with postcode M1E shown by red boundary (image taken from Google Maps(8))

6 Assumptions and Their Implications

- In searching boroughs for venues, a radius of 700 meter was considered from the specified geospatial coordinates of that postcode. As we move away from downtown, this may not be a suitable searching radius, as venue concentration decreases as we move away from downtown.

7 Conclusions

a) Unsupervised machine leaning technique of *k* mean clustering is applied to make data driven decision to invest in a new Afghan restaurant in the neighborhoods of Toronto.

b) Data obtained from different sources is cleaned and transformed to required data formats

c) Eleven clusters were used to identify the cluster having the required unique set of features for investment opportunity.

d) The results show that multiple options are available for company X, as a set of neighborhoods exits with the desired unique .

e) One set of neighborhoods associated with postcode M1E was identified as one of the suitable location.

8 Acknowledgements

The author is thankful to all of the people involved in developing freely available data analysis tools, from Anaconda distribution to all of the freely available libraries. I am especially thankful to "*IBM Professional Data Science Course"* instructors for arranging well-structured and comprehensive courses.

9 References

1. Toronto Population 2019 [Available from: http://worldpopulationreview.com/world-cities/toronto-population/.
2. Canadians are craving ethnic foods 2019 [Available from: http://www.canadiangrocer.com/top-stories/canadians-increasingly-receptive-to-ethnic-foods-65179.
3. WHERE TO EAT THIS WEEKEND: OTTAWA 2019 [Available from: https://www.foodbloggersofcanada.com/where-to-eat-this-weekend-ottawa/.
4. Garbade MJ. Understanding K-means Clustering in Machine Learning 2019 [Available from: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.
5. Wikipedia. List of postal codes of Canada 2019 [Available from: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.
6. FourSquare. Places by Foursquare 2019 [Available from: https://foursquare.com/.
7. Toronto Social Atlas 2016 Maps 2019 [Available from: https://www.toronto.ca/wp-content/uploads/2018/06/97b3-ct16_TOR_EthnicOrigin_Afghan.pdf.
8. Afghan resturants in Scarborough 2019 [Available from: https://www.google.com/maps/search/afghan+restaurant/@43.7502459,-79.2625375,13z/data=!4m2!2m1!6e5ahl=en.