Coursera

IBM DATA SCIENCE CAPSTONE

by Khyati Jha

Feb 2020

Opening a new Bakery in the Torronto city

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Toronto, Canada to open a new Bakery. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Toronto, Canada, if someone is looking to open a new Bakery, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful to property developers and investors looking to open or invest in Bakery in the capital city of Toronto, Canada.

To solve the problem, we will need the following data:

- List of neighborhoods in Toronto. This defines the scope of this project which is confined to the city.
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to Bakery. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them:

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) contains a list of neighbourhoods in Toroto, with a total of 70 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers.
Foursquare API will provide many categories of the venue data, we are particularly interested in the Bakery 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 use.

Method:

Firstly, we need to get the list of neighbourhoods in the city of Toronto. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
. We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods 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. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods 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 city of Toronto.

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 neighbourhoods 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 neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood 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 analysing the “Bakeries” data, we will filter the “Bakeries” as venue category for the neighbourhoods.

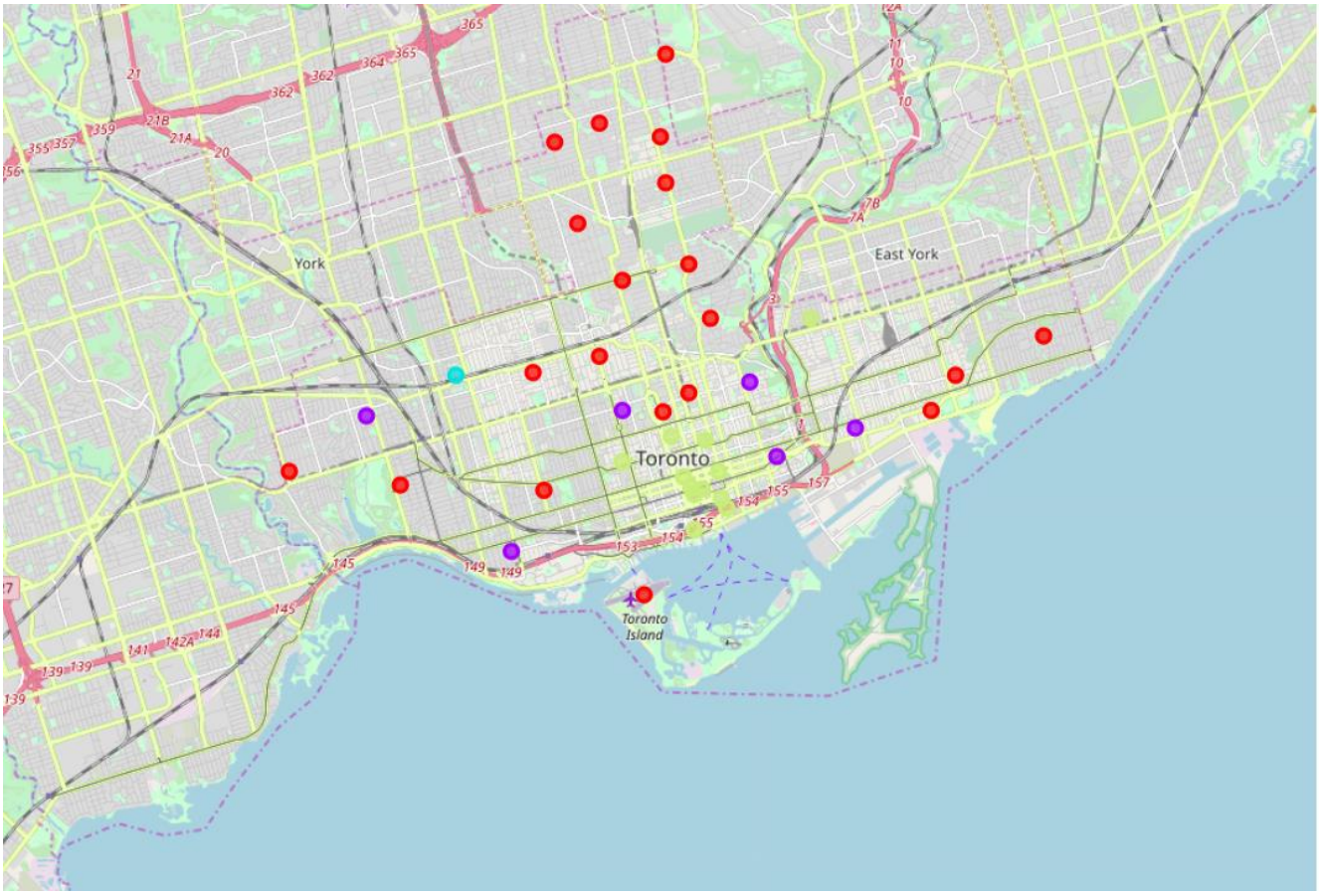
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 4 clusters based on their frequency of occurrence for “Bakeries”. The results will allow us to identify which neighbourhoods have higher concentration of Bakery while which neighbourhoods have fewer number of Bakeries. Based on the occurrence of Bakeries in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Bakery.

Results:

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Bakery”:

- Cluster 0: Neighborhood with almost no bakeries.
- Cluster 1: Neighborhood with good number of Bakeries
- Cluster 2: Neighborhood with highest concentration of Bakeries
- Cluster 3: Neighborhood with Low concentration of Bakeries

Map representing 4 clusters:



Cluster 0:

	Neighborhoods	Bakery	clusters	PostalCode	Borough	Latitude	Longitude
29	Roselawn	0.0	0	M5N	Central Toronto	43.711695	-79.416936
17	Forest Hill North, Forest Hill West	0.0	0	M5P	Central Toronto	43.696948	-79.411307
22	Lawrence Park	0.0	0	M4N	Central Toronto	43.728020	-79.388790
23	Little Portugal, Trinity	0.0	0	M6J	West Toronto	43.647927	-79.419750
28	Rosedale	0.0	0	M4W	Downtown Toronto	43.679563	-79.377529
13	Deer Park, Forest Hill SE, Rathnelly, South Hi...	0.0	0	M4V	Central Toronto	43.686412	-79.400049
12	Davisville North	0.0	0	M4P	Central Toronto	43.712751	-79.390197
11	Davisville	0.0	0	M4S	Central Toronto	43.704324	-79.388790
37	The Beaches West, India Bazaar	0.0	0	M4L	East Toronto	43.668999	-79.315572
24	Moore Park, Summerhill East	0.0	0	M4T	Central Toronto	43.689574	-79.383160
8	Christie	0.0	0	M6G	Downtown Toronto	43.669542	-79.422564
25	North Toronto West	0.0	0	M4R	Central Toronto	43.715383	-79.405678
35	The Annex, North Midtown, Yorkville	0.0	0	M5R	Central Toronto	43.672710	-79.405678
26	Parkdale, Roncesvalles	0.0	0	M6R	West Toronto	43.648960	-79.456325
4	CN Tower, Bathurst Quay, Island airport, Harbo...	0.0	0	M5V	Downtown Toronto	43.628947	-79.394420
3	Business Reply Mail Processing Centre 969 Eastern	0.0	0	M7Y	East Toronto	43.662744	-79.321558
27	Queen's Park	0.0	0	M7A	Downtown Toronto	43.662301	-79.389494
36	The Beaches	0.0	0	M4E	East Toronto	43.676357	-79.293031
9	Church and Wellesley	0.0	0	M4Y	Downtown Toronto	43.665860	-79.383160
30	Runnymede, Swansea	0.0	0	M6S	West Toronto	43.651571	-79.484450

cluster 1:

	Neighborhoods	Bakery	clusters	PostalCode	Borough	Latitude	Longitude
21	High Park, The Junction South	0.043478	1	M6P	West Toronto	43.661608	-79.464763
19	Harbourfront	0.065217	1	M5A	Downtown Toronto	43.654260	-79.360636
34	Studio District	0.048780	1	M4M	East Toronto	43.659526	-79.340923
5	Cabbagetown, St. James Town	0.044444	1	M4X	Downtown Toronto	43.667967	-79.367675
2	Brockton, Exhibition Place, Parkdale Village	0.047619	1	M6K	West Toronto	43.636847	-79.428191
18	Harbord, University of Toronto	0.057143	1	M5S	Downtown Toronto	43.662696	-79.400049

cluster 2:

	Neighborhoods	Bakery	clusters	PostalCode	Borough	Latitude	Longitude
15	Dovercourt Village, Dufferin	0.142857	2	M6H	West Toronto	43.669005	-79.442259

cluster 3:

	Neighborhoods	Bakery	clusters	PostalCode	Borough	Latitude	Longitude
31	Ryerson, Garden District	0.020000	3	M5B	Downtown Toronto	43.657162	-79.378937
33	Stn A PO Boxes 25 The Esplanade	0.021053	3	M5W	Downtown Toronto	43.646435	-79.374846
32	St. James Town	0.030000	3	M5C	Downtown Toronto	43.651494	-79.375418
0	Adelaide, King, Richmond	0.030000	3	M5H	Downtown Toronto	43.650571	-79.384568
16	First Canadian Place, Underground city	0.020000	3	M5X	Downtown Toronto	43.648429	-79.382280
14	Design Exchange, Toronto Dominion Centre	0.020000	3	M5K	Downtown Toronto	43.647177	-79.381576
10	Commerce Court, Victoria Hotel	0.020000	3	M5L	Downtown Toronto	43.648198	-79.379817
7	Chinatown, Grange Park, Kensington Market	0.024691	3	M5T	Downtown Toronto	43.653206	-79.400049
6	Central Bay Street	0.024390	3	M5G	Downtown Toronto	43.657952	-79.387383
1	Berczy Park	0.036364	3	M5E	Downtown Toronto	43.644771	-79.373306
20	Harbourfront East, Toronto Islands, Union Station	0.020000	3	M5J	Downtown Toronto	43.640816	-79.381752
38	The Danforth West, Riverdale	0.024390	3	M4K	East Toronto	43.679557	-79.352188

Discussion:

As observations noted from the map in the Results section, most of the Bakries are concentrated in the central area of Toronto, with the highest number in cluster 2 and moderate number in cluster 01 On the other hand, cluster 0 has very low number to no Bakeryin the neighbourhoods. This represents a great opportunity and high potential areas to open new Bakerys as there is very little to no competition from existing malls. Meanwhile, Bakery in cluster 2 are likely suffering from intense competition due to high concentration of Bakery. From another perspective, the results also show that the oversupply of bakerymostly happened in the central area of the city, with the suburb area still have very few Bakeries. Therefore, this project recommends property developers to capitalize on these findings to open new Bakery in neighbourhoods in cluster 1 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new Bakery in neighbourhoods in cluster 0 with moderate competition. Lastly, property developers are advised to avoid neighbourhoods in cluster 2 which already have high concentration of Bakerys and suffering from intense competition.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 4 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new Bakery. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 and 3 are the most preferred locations to open a new Bakery. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new Bakery.

Refrecnce:

Category:Suburbs in Toronto. *Wikipedia*. Retrieved from (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

Foursquare Developers Documentation.*Foursquare*. Retrieved from https://developer.foursquare.com/docs