COURSERA CAPSTONE

IBM Applied Data Science Capstone

Finding a Better Place in Etobicoke, Toronto

By- Akshay Jadhav

Introduction

The neighborhoods that appeal to you will largely be a matter of personal choice. However, a truly great neighborhood will have a few key factors: accessibility, appearance, and amenities. Your neighborhood may also dictate the size of the lot on which your house is built.In terms of accessibility, you should look for a neighborhood that is situated near your city's major routes and that has more than one point of entry. Commuting to and from work is a big part of many people's day, so a house with easy access to roads and/or public transportation will be more desirable than one that is tucked away and can only be accessed by one route.A great neighborhood should also include important amenities such as grocery stores, shops, and restaurants. Most people like to frequent places that are convenient if you have to drive a great distance to get to anything, it's likely to make your house less attractive. Schools are another important amenity even if you don't have kids, if you want to sell your home in the future, many buyers will be on the lookout for good schools. The quality of local schools and the distance from the house are both important factors to consider.

The following system will help the user to select a perfect location.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Etobicoke, Toronto to shift to new venue. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: It will help people making smart and efficient decision on selecting great neighborhood out of numbers of other neighborhoods in Etobicoke, Toranto.

Target Audience of this project

This project is particularly useful to people looking to shift to great location in the city of Etobicoke, Toranto. The number of people who change residences within the United States each he number of people who change residences within the United States each year is large: roughly 1.5 percent of the population moves between two of the four ear is large: roughly 1.5 percent of the population moves between two of the four Census regions (Northeast, Midwest, South, and West) annually, and about the census regions (Northeast, Midwest, South, and West) annually, and about the same number of individuals (roughly 1.3 percent of the population) move to a same number of individuals (roughly 1.3 percent of the population) move to a different state within the same region. These numbers come from the IRS series different state within the same region.

Data

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

- List of neighbourhoods in of Etobicoke, Toranto. This defines the scope of this project which is confined to the city of of Etobicoke,, the best city of the country.
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to daily requirements. We will use this data to perform clustering on the neighbourhoods.

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 postal codes in toronto , with a borough 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 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.

Methodology

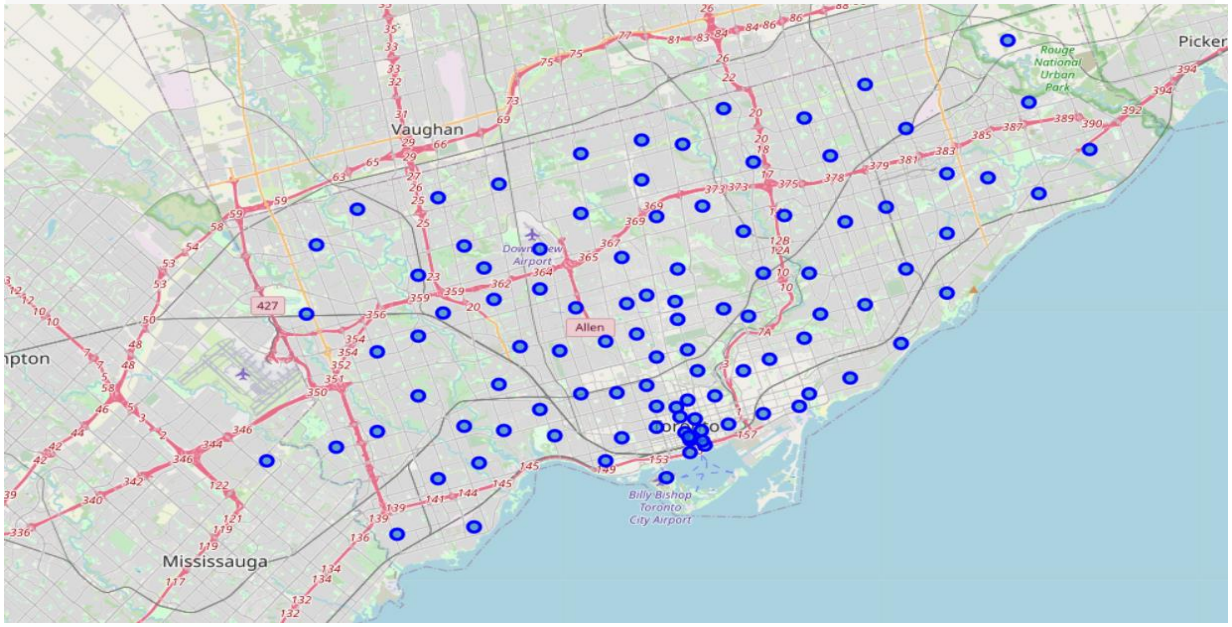
Firstly, we need to get the list of neighbourhoods & postal codes in the city of Etobicoke, Toranto. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods and postal codes 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 “top venues” data, we will filter the “top venues” 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 3 clusters based on their frequency of occurrence for “venues”. The results will allow us to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have more number of amenities. Based on the occurrence of amenities in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to choose a venue.

Methodology

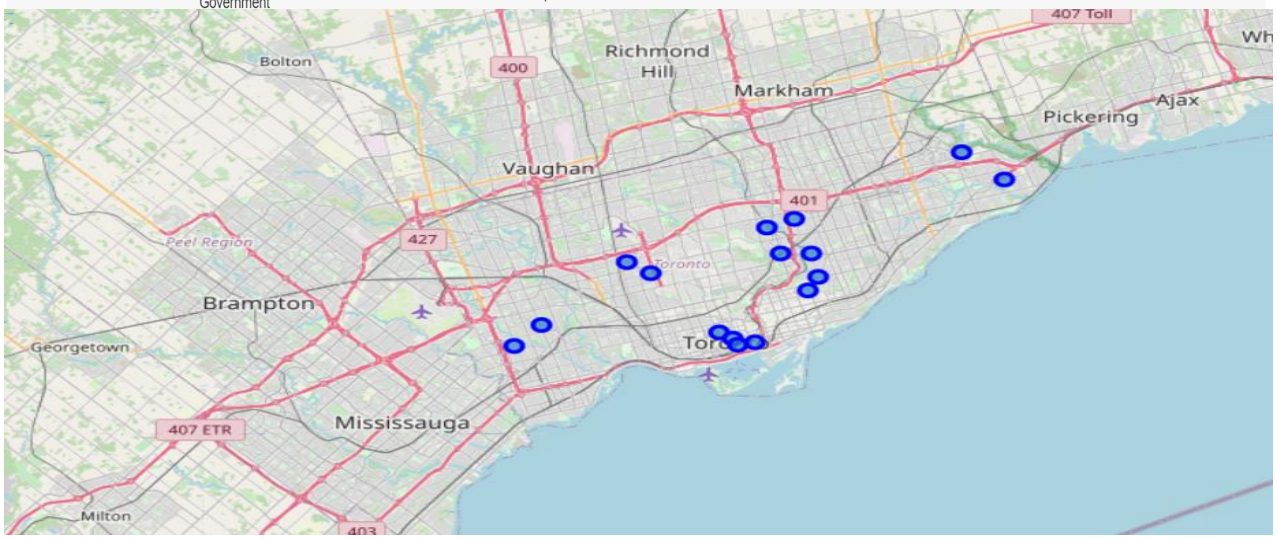
	Postal Code	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
7	M8A	Not assigned	Not assigned
8	M9A	Etobicoke	Islington Avenue, Humber Valley Village
9	M1B	Scarborough	Malvern, Rouge
10	M2B	Not assigned	Not assigned
11	M3B	North York	Don Mills
12	M4B	East York	Parkview Hill, Woodbine Gardens
13	M5B	Downtown Toronto	Garden District, Ryerson
14	M6B	North York	Glencairn



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 “best places”:

	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Cluster Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	M3A	North York	Parkwoods	43.753259	-79.329656	2	Park	Pet Store	Construction & Landscaping	Food & Drink Shop	Bed & Breakfast	Women's Store	Dog Run	Doner Restaurant	Donut Shop	Drugstore
3	M4A	North York	Victoria Village	43.725882	-79.315572	1	Hockey Arena	French Restaurant	Playground	Coffee Shop	Pizza Place	Park	Café	Sporting Goods Shop	Portuguese Restaurant	Gym
4	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	1	Coffee Shop	Park	Café	Theater	Restaurant	Pub	Bakery	Thai Restaurant	Performing Arts Venue	Breakfast Spot
5	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	1	Clothing Store	Accessories Store	Coffee Shop	Vietnamese Restaurant	Fast Food Restaurant	Miscellaneous Shop	Furniture / Home Store	Women's Store	Seafood Restaurant	Men's Store
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	1	Coffee Shop	Sandwich Place	Café	Italian Restaurant	Burrito Place	Japanese Restaurant	Park	Diner	Beer Bar	Bookstore



Discussion

As observations noted from the map in the Results section, most of the best venues are concentrated in the area of Etobicoke, Toranto city, with the highest number in cluster 1 and moderate number in cluster 0. On the other hand, cluster 2 has very low number of venues in the neighbourhoods. This represents a great opportunity and high potential areas to choose as there is very good facilities. Meanwhile, amenities in cluster 2 are likely suffering from less number of venues.

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 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e.to people who are willing to choose a prfect location. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 are the most preferred locations to choose a location .

References

Postal Codes: Toronto. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur

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