IBM APPLIED DATA SCIENCE CAPSTONE

Where to Open a Restaurant in Toronto

Xinyao Jie

Introduction

This project is a capstone project for IBM applied data science course by coursera. The topic I choose is to explore where to open a new restaurant in Toronto. For many customers, going to restaurant is a great way for food after working for full day. It is also a great way to relax and enjoy themselves with their friends during weekends and holidays. For restaurant owners, the central location and the large crowd provides a great distribution channel for them. Property developers are also taking advantage of this trend to open a new restaurant to cater to the demand. As a result, there are many restaurants in the city of Toronto and many more are being built. Opening new restaurants allows owners to make money by selling foods. Of course, as with any business decision, opening a new restaurant requires serious consideration and is a lot more complicated than it seems. Particularly, the location of a restaurant is one of the most important decisions that will determine whether the restaurant will be a success or a failure.

Business Problem

The objective of this capstone project is to analyze and select the best locations in the city of Toronto, Canada to open a new restaurant. 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 restaurant, where would you recommend that they open it?

Target Audience of Project

This project is particularly useful to someone who want to start their own restaurant and investors looking to open or invest in new restaurants in the city of Toronto. Also, it provides some guidance for tourists and customers about the density of restaurants in different neighborhoods.

Data

The following data is needed to solve this problem

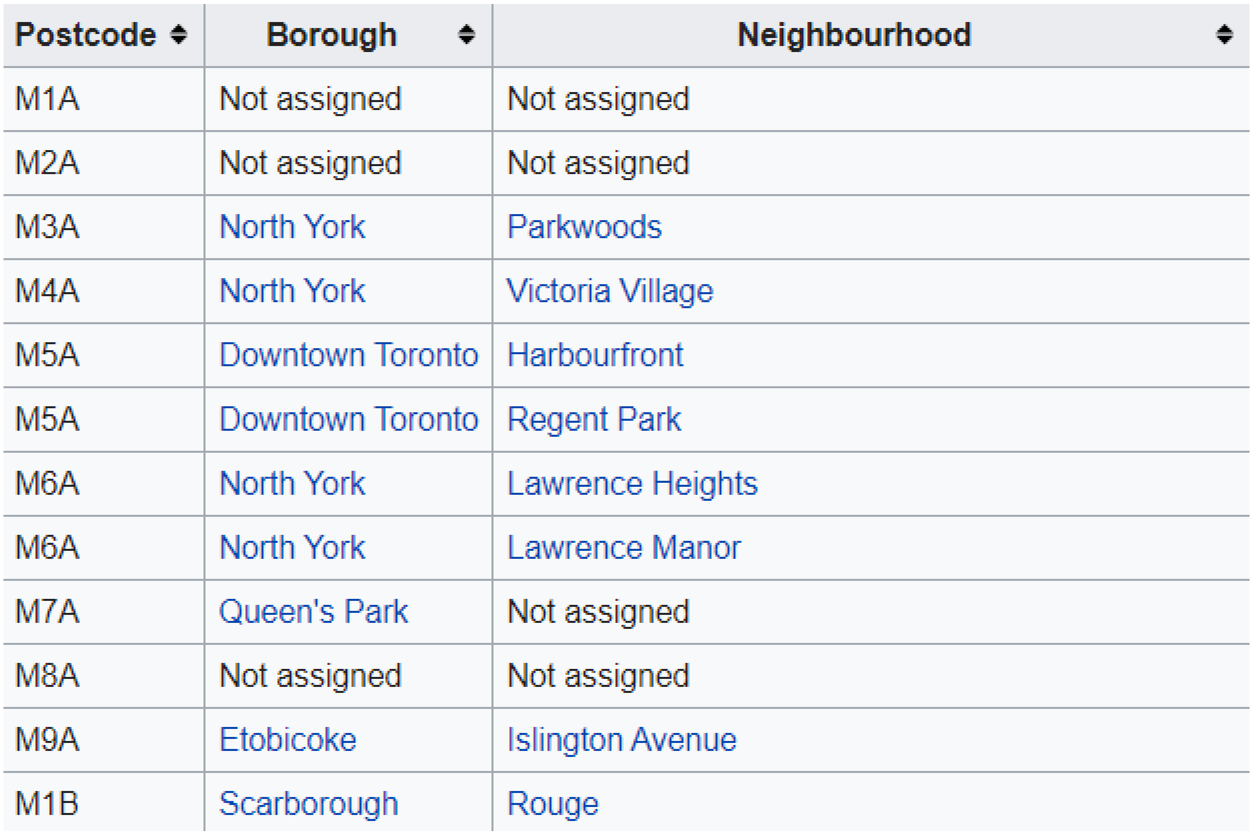
- List of neighbourhoods in Toronto. This defines the scope of this project which is confined to the city of Toronto.
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to different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

Sources of data

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) contains a list of neighbourhoods in Toronto, with a total of 103 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python Pandas packages. Then we will get the geographical coordinates of the neighbourhoods using a csv. File I already downloaded. I use Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods to get that data. 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 Restaurant category in order to help us to solve the business problem put forward. Notice there are mant kinds of restaurant like Asian Restaurant, African Restaurant, etc. We will consider all of them in one piece. 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).

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Toronto from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. The list is as follows.



Postcode	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Harbourfront
M5A	Downtown Toronto	Regent Park
M6A	North York	Lawrence Heights
M6A	North York	Lawrence Manor
M7A	Queen's Park	Not assigned
M8A	Not assigned	Not assigned
M9A	Etobicoke	Islington Avenue
M1B	Scarborough	Rouge

We will do web scraping using Python Pandas packages to extract the list of neighbourhoods data. Then we do some cleanups for this form to transform it to the form we can use. Then, 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, I use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. Considering this package having some drawbacks that you have to be persistent sometimes in order to get the geographical coordinates of a given postal code. You can make a call to get the latitude and longitude coordinates of a given postal code and the result would be None, and then make the call again and you would get the coordinates. So, in order to make it clear and efficient, I use it to download the coordinates first and store them into a csv file. When I need to use it, I upload it. 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 are correct and have a first impression on the neighborhoods.

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 analyze 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 analyzing the restaurant data, we will filter the venue with "Restaurant" in their category names for the neighbourhoods because the categories have different type of restaurants. We need them all.

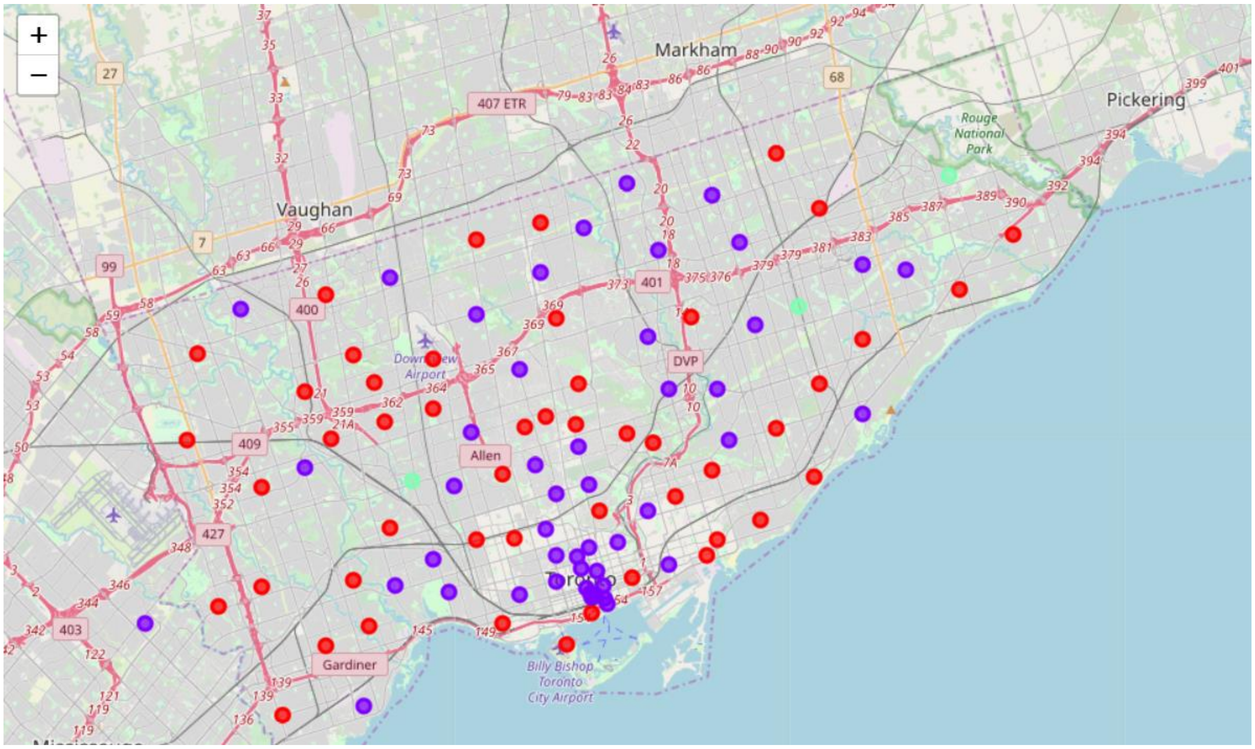
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine 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 Restaurant. The results will allow us to identify which neighbourhoods have higher concentration of restaurants while which neighbourhoods have fewer number of restaurants. Based on the occurrence of restaurants in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to start a new restaurant.

Result

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 Restaurant:

- Cluster 0: Neighbourhoods with low frequency of restaurants. Marked as red dots.
- Cluster 1: Neighbourhoods with mid frequency of restaurants. Marked as purple dots.
- Cluster 2: Neighbourhoods with high frequency of restaurants. Marked as green dots.

We visualize the clustering result below in the map.



Discussion

As observations noted from the map in the Results section, we need to acknowledge that this frequency is an absolute value but relative to the venue density. That means even a neighborhood clustered as label 3 do not need to have more restaurant in absolute value that a neighborhood clustered as label 0. The frequency is relative to venue density and venue density can be seemed as related to population density. In neighborhood having high population density, even there are many restaurants, there may also be a lack of restaurant compared to the population. In neighborhood having low population density, even there are only a few restaurants, there may also a high competition among restaurants.

After we acknowledge that, we can find those neighborhoods labelled as 3 do not need to locate in the center of Toronto city. That is because population density there is also high. Most neighborhoods having a high competition among restaurants is not close to city center. Neighborhoods in cluster 2 are likely suffering from intense competition. We definitely don’t want to open our restaurant in that place where there are not many customers but enough restaurants.

Meanwhile, neighborhoods in cluster 0 has very low number to no restaurants in the neighborhoods. This represents a great opportunity and high potential areas to open new restaurant as there is very little to no competition from existing malls. These neighborhoods are perfect place to start a new restaurant because literally you will face no change here.

From another perspective, the results also show that neighborhoods in central city area are mostly clustered as label 1. That means restaurants there will face some competition but it’s still profitable. Restaurants owners with unique menus or high quality of cuisine will stand out from the competition and can potentially make a large amount of money because of high population density.

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. restaurant owners and investors regarding the best locations to start a new restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is:

1. The neighborhoods in cluster 0 are the most preferred locations to open a new restaurant. There are little competition and you will make money.
2. The neighborhoods in cluster 1 are okay to open a new restaurant. Also, in central city area, owners with unique competition advantages can make a lot of money.
3. The neighborhoods in cluster 2 are definitely not good places to open a new restaurant.

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 restaurant.