IBM APPLIED DATA SCIENCE CAPSTONE

Opening Bubble Tea Shop in Toronto: Report

By Asmar Aliyeva

October 2019

Introduction

Since Green Grotto opened at Yonge and Sheppard in 1993, the Toronto's thirst for bubble tea—that sweet, tea-based drink studded with tapioca balls, or *boba*—has really popped. Today, bubble tea shops are proliferating in the downtown core. Bubble tea in Toronto is so popular, you'll find international chains and local faves expanding to all corners of the city at a ferocious pace.

One of the most significant factors when thinking about opening a bubble tea franchise is the location of the shop. It is crucial to make sure that the café will be in an area with high traffic that will come in and peruse your store for food and drink. To open a successful bubble tea business, area that has less competition will be the best choice.

Business Problem

This capstone project will analyse neighbourhoods of Toronto city to choose the best location for opening a new Bubble Tea Shop. Data Science and Machine Learning techniques such as K-Means Clustering will be used to analyse this business problem and to find the neighbourhood with lowest competition and less concentration of Bubble Tea Shops. So, the main question of this research will be to find the location where it is most beneficial to open Bubble Tea Shop.

Target Audience

The result of this capstone project will be particularly useful for investors looking to invest in cafes and restaurants, principally, in Bubble Tea Shops in Toronto, ON. Bubble Tea Shops are getting more popular and new Shops are opening every year. Yet, there are still available and undiscovered area where investors can take advantage of.

Data:

Required Data:

1. The data regarding neighbourhoods of Toronto.
2. Latitude and longitude of neighbourhoods
3. Venue data for each neighbourhood

Data collection:

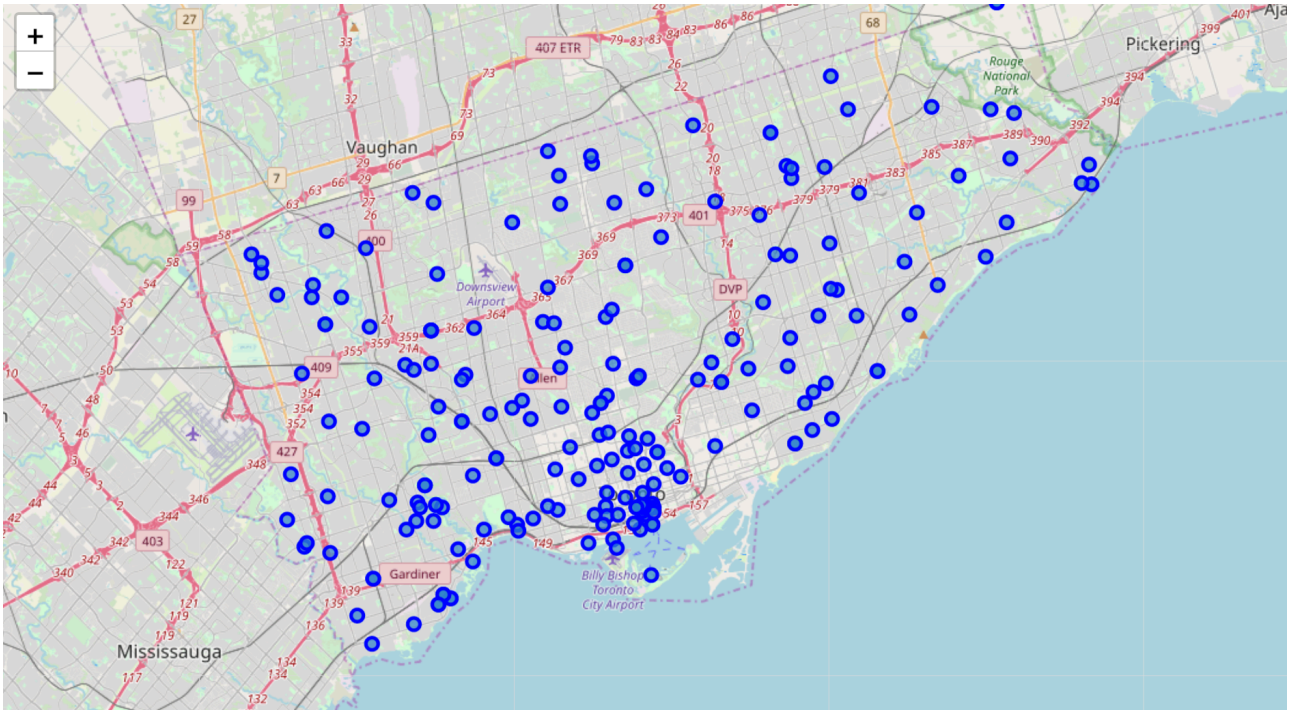
1. Data regarding neighbourhoods of Toronto will be collected from Wikipedia Page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M), which contains the list of all neighbourhoods with their corresponding postal codes and boroughs. We will use web scrapping techniques, Python Request and BeautifulSoup, to extract the data.
2. To collect corresponding latitude and longitude coordinates of neighbourhoods we will use Python Geocoder package.
3. To collect venue data of each neighbourhood we will use Foursquare API as it has the largest database of locations with over 105 million places. Along with venue name, it provides information about the category of the venue and its latitude and longitude. Among venue categories I am interested in Bubble Tea Shops.

Our final dataset will consist of name of the neighbourhood, its latitude and longitude, venue name, its corresponding latitude and longitude and the category to which the venue belongs.

Methodology:

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. Next, 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 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 and make sure that geographical coordinates data returned by Geocoder are correctly plotted in the city of Toronto (Figure 1).
Next, we will use Foursquare API to get the top 100 venues that are within a radius of 1500 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 use OneHot Encoding to



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 “Bubble Tea Shop” data, we will filter the “Bubble Tea Shop” 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 5 clusters based on their frequency of occurrence for “Bubble Tea Shop”. The results will allow us to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighbourhoods, it will

help us to answer the question as to which neighbourhoods are most suitable to open new shopping malls.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 5 clusters:

• Cluster 0: Neighbourhoods with almost no Bubble Tea Shop

• Cluster 1: Neighbourhoods with the highest number of Bubble Tea Shops

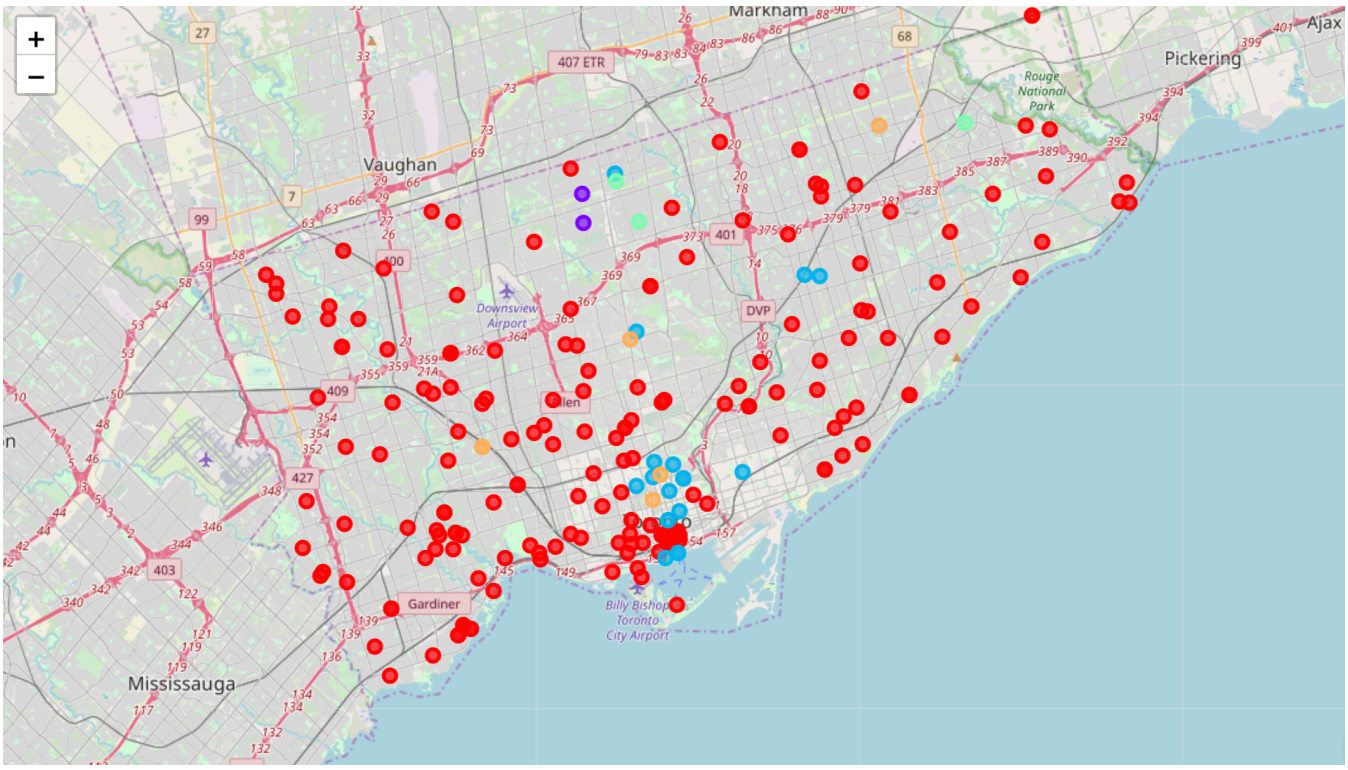
• Cluster 2: Neighbourhoods with moderate concentration of Bubble Tea Shops

• Cluster 3: Neighbourhoods with high concentration Bubble Tea Shops

• Cluster 4: Neighbourhoods with moderate concentration of Bubble Tea Shops

In other words, the raking of clusters based on the concentration of Bubble Tea Shops (from lowest to highest concentration) will be as following:

Cluster 0 → Cluster 2→ Cluster 4→ Cluster 3→ Cluster 1



Red – Cluster 0, Blue – Cluster 2, Purple – Cluster 1, Green – Cluster 3, Orange – Cluster 4

Discussions:

As observations noted from the map in the Results section, most of the bubble tea shops are concentrated in Cluster 1 and 3, and moderate number in cluster 2 and 4. On the other hand, cluster 0 has very low number to no bubble tea shops in the neighbourhoods. This represents a great opportunity and high potential areas to open tea shops as there is very little to no competition. Meanwhile, tea shops in cluster 1 are likely suffering from intense competition due to oversupply and high concentration. This project recommends property developers to capitalize on these findings to open new bubble tea shops in neighbourhoods in cluster 0 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open bubble tea shops in neighbourhoods in cluster 2 with moderate competition. Lastly, property developers are advised to avoid neighbourhoods in cluster 1 which already have high concentration of tea shops and are 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 5 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 bubble tea shop. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 0 are the most preferred locations to open a new shopping mall. 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 bubble tea shop.