The Battle of Neighborhoods: Starting a Taiwanese Bubble Tea Shop

This is to build models for segmenting neighborhoods in Toronto to explore suitable locations for starting a Taiwanese Bubble Tea Shop.

Andy Chang
August 3, 2021

Introduction

The largest city in Canada, Toronto, containing more than 2.7 million population with renowned as one of the most multicultural cities globally because of it attracts immigrants from around the world.

Business Problem

As the increasing popularity of Taiwanese bubble tea which brings attention worldwide, especially in multicultural cities like L.A, N.Y, S.F; this project is to explore the business in Toronto because of similar characteristics.

Key items in consideration:

What are most popular locations/venues?
Are there existing competition, like coffee shop?
Traffic and pedestrian information, how many people will get around in each location?
Crime rate and so on.

Goal

Identifying suitable locations with the considered items listed above, from data science perspective.

Target Audience

Business owner and stakeholder planning to expand bubble tea business, and this is how data science can be applied.

Data collection

The following data sources are to be used in the project:

1. Toronto post codes (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
2. Popular venues of a given neighborhood in Toronto (https://developer.foursquare.com/)
3. Traffic Signal Vehicle and Pedestrian Volumes, Toronto (https://ckan0.cf.opendata.inter.prod-toronto.ca/en/dataset/traffic-signal-vehicle-and-pedestrian-volumes)
4. Toronto Crime Indicator, Toronto Police Service (https://data.torontopolice.on.ca/datasets/major-crime-indicators-1)

Methodology

Analytic Approach: To approach the problem with k-means, which is a clustering technique that cluster data points based on similar characteristics. In this case, it shows audience the neighborhoods clustered in venue categories.

Data Cleaning, Preparation, Process, and Objectives

1. **Toronto post codes:** a Wikipedia page about Toronto postal code. We will scrape the page and create a data frame consisting of three columns: Postal Code, Borough, and Neighborhood. We remove any rows that do not have borough assigned. Then, we will be using the Geocoder python package to retrieve the postal code's coordinates. It will return 103 rows and 5 columns. Objective: To obtain the exact coordinates for each neighborhood based on the postal code, allowing us to explore and map the city.

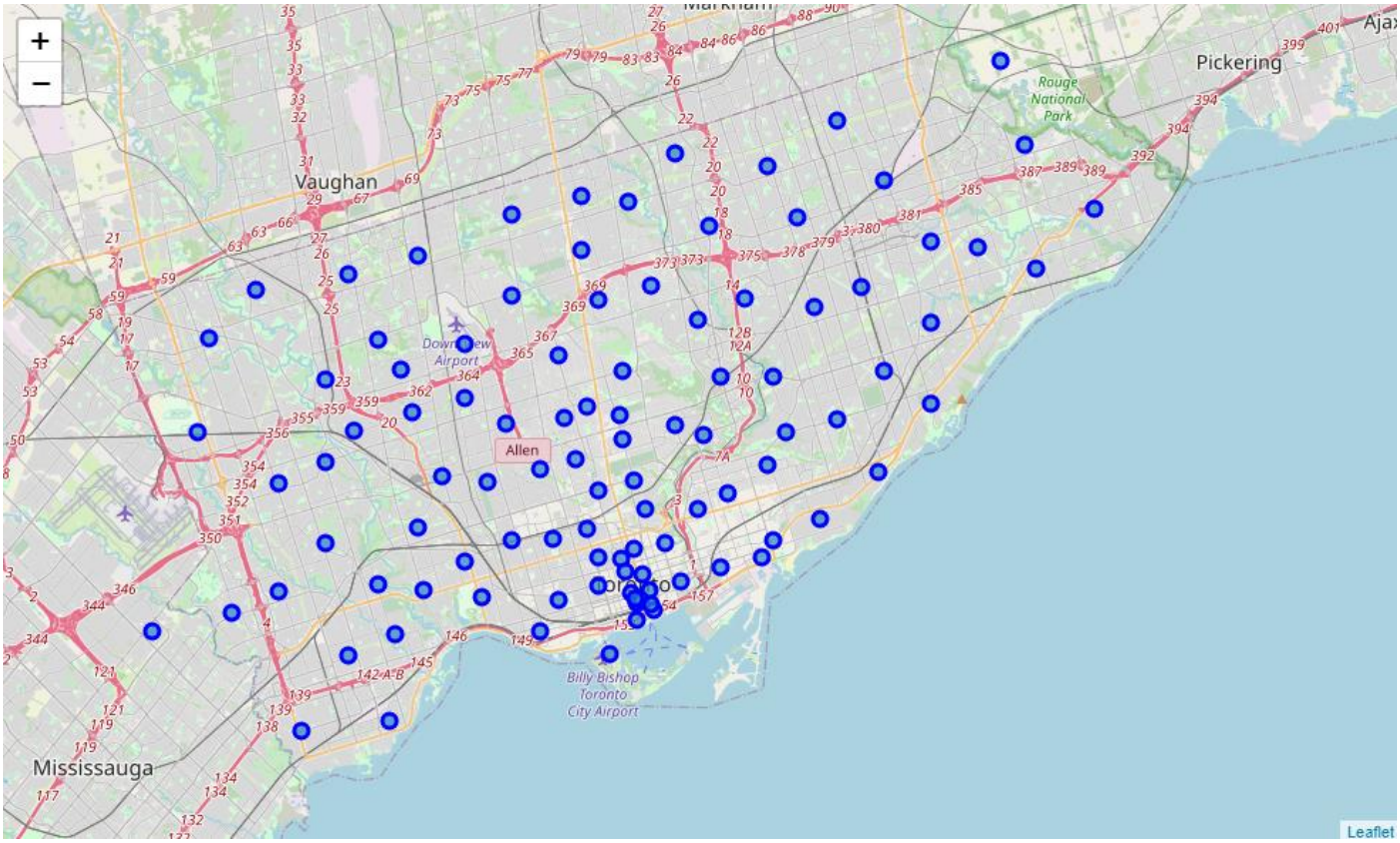

Figure 1. Map of boroughs in Toronto.

2. **Popular venues:** stored inside Foursquare Location Data, and we will use Foursquare API to access it. We utilize the postal coordinates to retrieve popular venues around a specific radius. As a result, the same venue categories will be returned to different neighborhoods. We can use this idea to cluster the neighborhoods based on their venues representing services and amenities. Objective: leverage Foursquare credentials to access the 2nd data source through its API and retrieve the popular venues along with their details, especially coffee shops.
   To implement k-means clustering, elbow method is selected to determine the number of k. As k=4 derived from elbow method with coffee shop as key, further clustering is conducted.

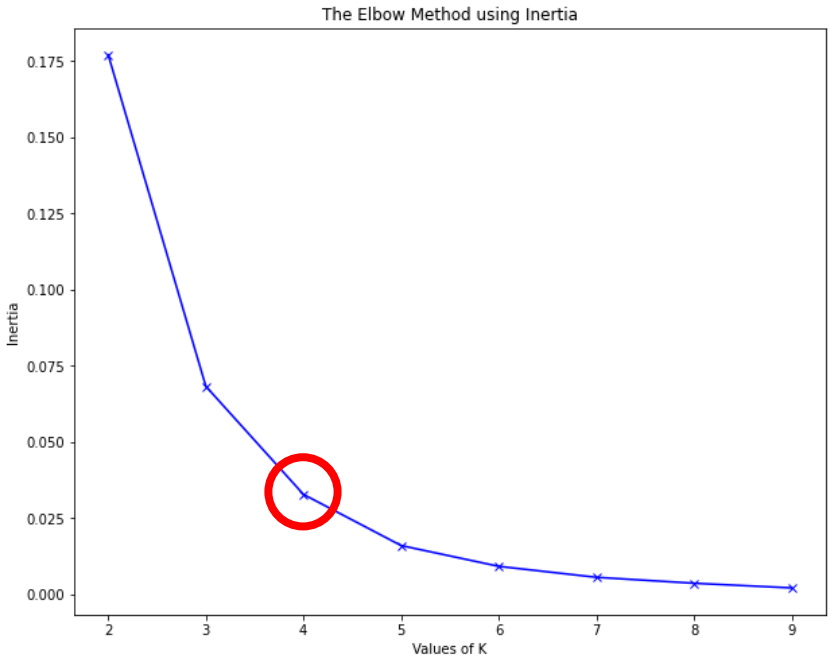

Figure 2. k-value derived by the elbow method

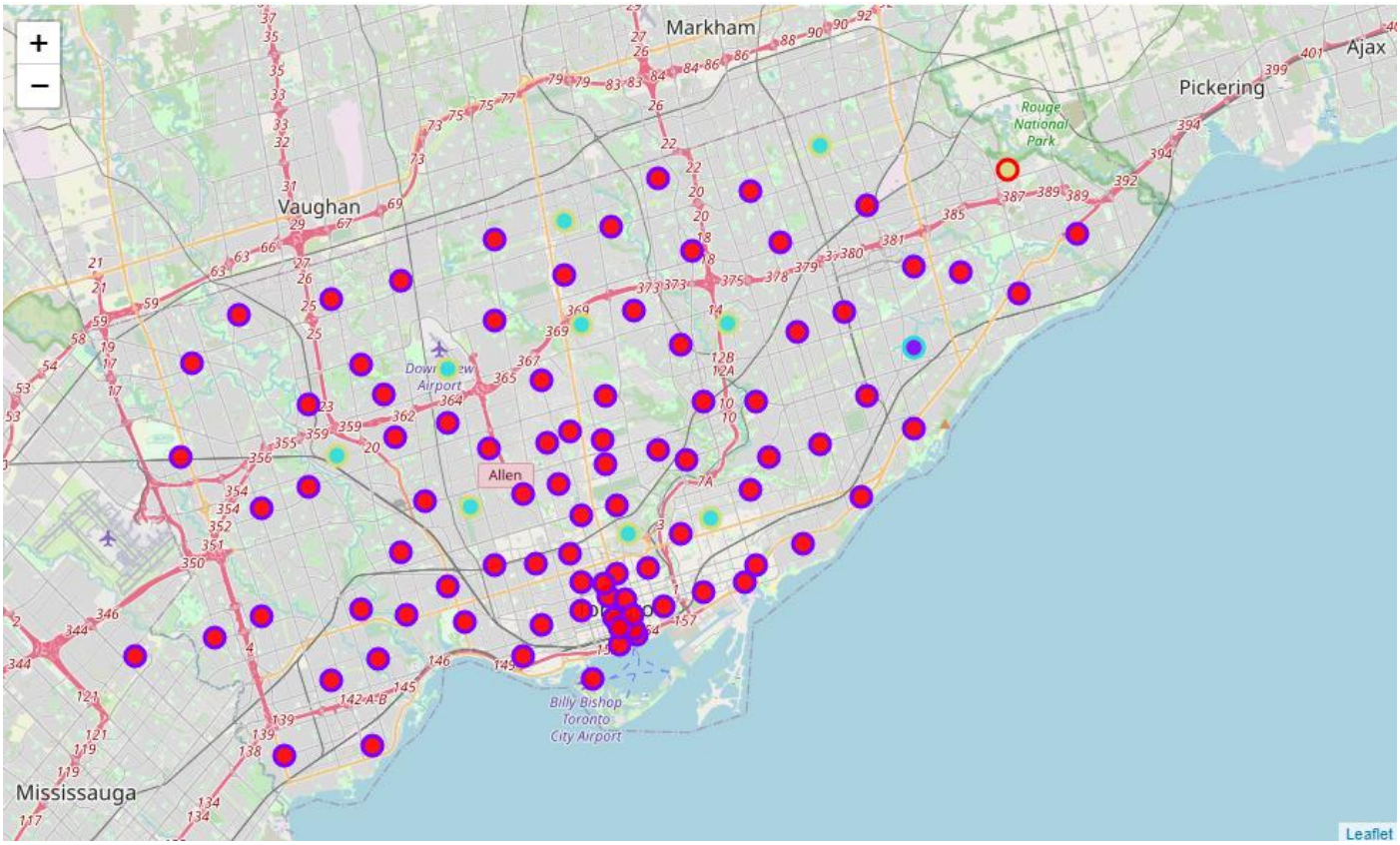

Figure 3. Result based on venue with k-means clustering.

Clustered venues:

Cluster 1 (Red): Coffee shops, restaurants (highly competitive)

Cluster 2 (Blue): Convenient stores, restaurants (less competitive)

Cluster 3 (Green): Fast food, stores (less competitive)

Cluster 4 (Red circle): Ball Park, stores (tea truck may be considered in playing season)

3. **Vehicle and Pedestrian Volumes:** It contains 2280 rows and 11 columns. The rows represent the intersection that each main road has. The data is typically collected between 7:30 a.m. and 6:00 p.m. at intersections where there are traffic signals. Each intersection holds vehicle and pedestrian volumes data, along with its coordinates.
   We will focus on 5 columns; those are Main, 8 Peak Hr Pedestrian Volume, 8 Peak Hr Vehicle Volume, Latitude, and Longitude. We will use these features to diagnose each main road's characteristics and locate the busiest main roads in the city. Objective is to analyze pedestrian traffic.
   Highlighted locations on map indicates the vehicle traffic more than 12,000 per hour, and more than 1,200 pedestrians per hour during peak hour.

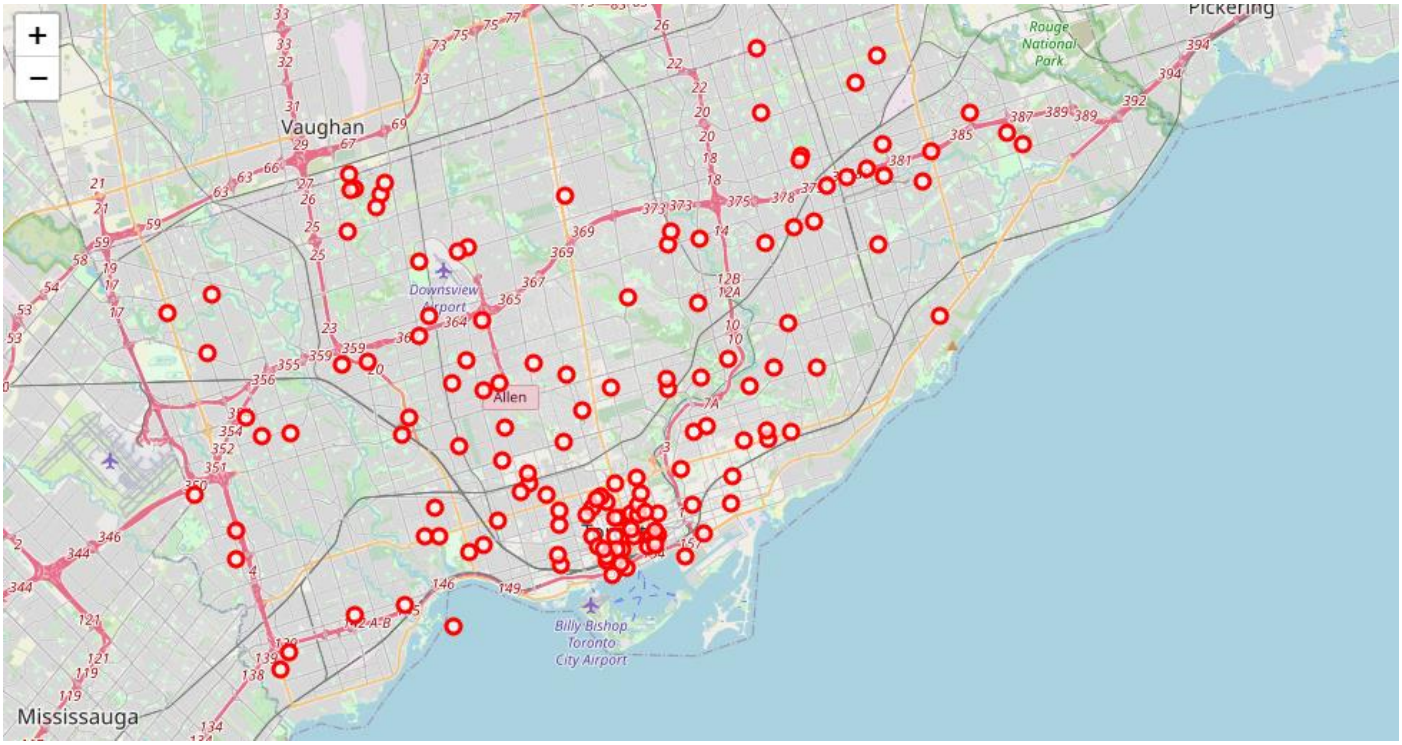

Figure 4. The locations with high vehicle and pedestrian traffic in peak hours.

4. **Toronto Crime indicator:** It contains 242,879 rows and 29 columns. The rows represent crime incidents that reported from 2014 to 2020. It has 5 Major Crime Indicators (MCIs) scattered to 17 divisions and 140 listed neighborhoods. We will group the data based on division and get statistics about crime rates. Objective is to analyze Crime rate.

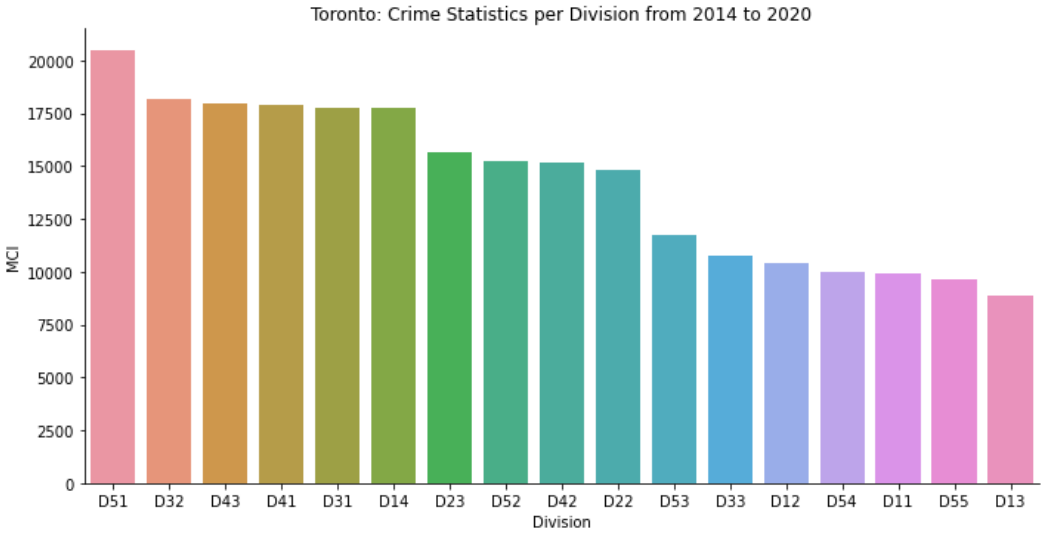

Figure 5. Toronto: Crime statistics per division from 2014 to 2020

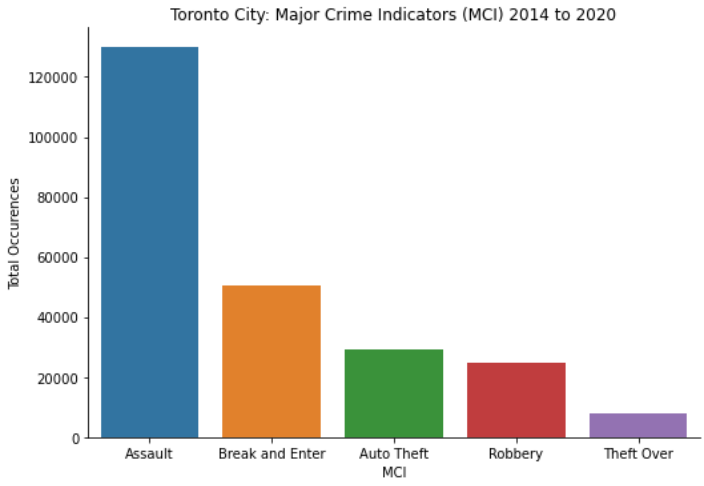

Figure 6. Toronto: Major crime indicators (MCI), 2014 to 2020

A summary in crime rate:

1. High crime rate: D51, D32, D43, D41, D31, D14
2. Middle crime rate: D23, D52, D42, D22
3. Low crime rate: D53, D33, D12, D54, D11, D55, D13

Hence the suitable locations are where crime rate are low, referred to Toronto police wiki:

1. Central Toronto: D11, D12, D53
2. East York: D12, D53, D54, D55
3. York: D13

Assault is the highest incidents within the 7 years. The divisions are categorized below based on incident numbers:

Results

The goal of this project is to identify suitable locations to open a Taiwanese bubble tea shop in Toronto. It would be subjective to define "suitable", yet we may line up the consideration in the following through analysis results:

Safety

The location with lowest crime rate can be considered.

Accessibility and Demographics

1. Accessibility: pedestrian vehicle traffics are the keys to select a location. Central Toronto, East York, and York are identified that meets traffic volume requirement. Knowing how and why to reach the shop location are also crucial, like parking lots, bus stations, visibility, and convenience. This requires further team discussion.
2. Demographic: As threshold on traffic set to filter out high traffic and pedestrian, it remains upcoming identification to clarify whether target demographic exists. Hence further investigation of target demographic is required through team discussion.

Neighborhood Business

Cluster 1: Coffee shops, restaurants (highly competitive)
Cluster 2: Convenient stores, restaurants (less competitive)
Cluster 3: Fast food, stores (less competitive)
Cluster 4: Ball Park, stores (tea truck may be considered in playing season)

Conclusion

Finding a suitable location to start bubble tea business, like other business, is challenging with various uncertainty. Throughout the analysis, it is believed helpful to gain meaningful insights to see more clearly, although further works on discussion and more analysis will be required. Hope this is helpful in dealing with similar cases.