Japanese Restaurants Opening Locations Prediction and Analysis in Toronto

Archel Taneka Sutanto

February 4, 2021

1. Introduction

1.1. Background

The city of Toronto is the 4th largest city in North America. It is a home to over 6 million people. It also provides an international center of business, finance, arts, and culture. This city is recognized as one of the most multicultural and cosmopolitan cities around the world. As the population grows, there are nearly various numbers of ethnics and backgrounds among its citizens. This also means that its culture is influenced by people who resided in Toronto from foreign countries. Thus, the culture varies based on regions (Asia, Europe, Africa, etc.) All kinds of business are appearing, especially food-related industries such as restaurants. People who want to start their businesses have to think and analyze the risks before getting into it. The idea derived from this analysis will provide deep analysis and insights which can help them strategically and lead to a high income low risk scenario. Toronto and its surroundings have established a strong base in terms of cuisines. This is probably influenced by the immigrants who moved to Toronto for the past decades. Specifically, Japanese dishes like ramen and sushi have become very popular among the Canadians. Starting one could be one of a challenge, but with the correct strategy and market analysis, a business owner is already on the right track of success.

1.2. Problem

Japanese cuisines demand high standards in terms of ingredients and spices. Since Toronto is not surrounded by waters, import is the only way to acquire the best ingredients. Not only that, location also determines the fate of the business. One needs to focus on which neighborhoods

and outlets have already proven their success. High traffic locations greatly impact the number of customers who walk in into the restaurant. Perhaps if a more quiet environment is preferrable, opening the restaurant far from the cities with scenery is also not a bad idea. There are still numerous factors that have to be considered before deciding the location.

1.3. Interest

The target audience will be the people who live in the Riverdale area. Using the given borough, the main objective is to analyze and recommend which location works the best. The final decision making is based on the number of traffic intensities and surrounding venues. By providing deep and thorough analysis, this would surely draw people's and investors' attention to open a Japanese restaurant around the area.

2. Data Acquisition and Cleaning

2.1. Data Sources

A list of Canada's neighborhoods can be found in this Wikipedia page. Nonetheless, we still yet to find their precise locations (longitudes and latitudes). This is why I also used Geospatial data in this project to combine both of them into a single data with the complete information.

2.2. Data Cleaning

First, I scraped the Wikipedia page to obtain the list of Canada’s neighborhoods. This step is rather easy since Wikipedia always follows a format where the list is placed inside a table. However, there are plenty of missing values in the list. These missing values are indicated by “Not assigned” in Borough and Neighborhood columns. Various steps can be implemented in order to replace numerical values, but in this case, we can just ignore or delete the rows corresponding to that missing value. Next, we can build a dataframe object out of the cleaned dataset before feeding it into our machine learning model (Table 1).

Moving on to the Geospatial data (Table 2), it is already in a csv format. Hence, we can directly load it into a dataframe without any additional preprocessing and cleaning steps. Now, we have the cleaned data for Canada’s neighborhoods and their geospatial information. To combine both of them, I used the ‘Postal Code’ column as the pivot and merged them. In the end, we should have the finalized dataset along with the neighborhood names, latitudes, longitudes, and postal codes (Table 3).

	Postalcode	Borough	Neighborhood
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

Table 1: Toronto neighborhood dataset

	Postalcode	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Table 2: Geospatial information according to postal codes

(39, 5)

	Postalcode	Latitude	Longitude	Borough	Neighborhood
37	M4E	43.676357	-79.293031	East Toronto	The Beaches
41	M4K	43.679557	-79.352188	East Toronto	The Danforth West, Riverdale
42	M4L	43.668999	-79.315572	East Toronto	India Bazaar, The Beaches West
43	M4M	43.659526	-79.340923	East Toronto	Studio District
44	M4N	43.728020	-79.388790	Central Toronto	Lawrence Park

Table 3: The final cleaned and combined dataset from both sources

2.3. Feature Selection

Upon examining data for each feature, it is clear that the categories of the venues will be our features. If we look inside a response that is returned by the Foursquare in JSON format, there is an additional column named ‘Venue Category’ for example: trail, beaches, park, offices, restaurants, pub, etc. As we already know that only numerical data is accepted by the machine

learning model, we need to convert these values into categorical variables. I implemented the one-hot encoding (Table 4) to this column where it creates columns of venue categories and fills in the values by 1 according to its actual value and 0 for other categories. Finally, we can drop the neighborhood column, because it does not tell us about something and also does not affect the model while training.

	Neighborhood	ATM	Accessories Store	Adult Boutique	Advertising Agency	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	Alternative Healer	American Restaurant	Antique Shop	Arcade	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	Assisted Living	Athletics & Sports	Audit
0	The Beaches	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	The Beaches	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	The Beaches	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	The Beaches	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	The Beaches	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Table 4: One-hot encoding

3. Exploratory Data Analysis

As I have mentioned before, with the help from Foursquare API, we can send a “search” or “explore” request and it will send back a response in JSON format. Here, I used the “explore” request where I can get up to 10 (or more) most commonly visited venues based on the location (latitudes and longitudes) I send along with my request (Table 5).

	Neighborhood	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
0	Berczy Park	Office	Building	Parking	Tech Startup	Residential Building (Apartment / Condo)	Breakfast Spot	Korean Restaurant	Movie Theater	Assisted Living	Hotel
1	Brockton, Parkdale Village, Exhibition Place	Office	Residential Building (Apartment / Condo)	Tech Startup	Conference Room	Building	Advertising Agency	Café	Bar	Coworking Space	Convenience Store
2	Business reply mail Processing Centre, South C...	Building	Office	Light Rail Station	Convenience Store	Medical Center	Butcher	Fast Food Restaurant	Theater	Restaurant	Gym / Fitness Center
3	CN Tower, King and Spadina, Railway Lands, Har...	Airport Gate	Airport Service	Moving Target	Airport Terminal	Park	Airport Lounge	Harbor / Marina	Airport	Boat or Ferry	General Travel
4	Central Bay Street	Hospital	Hospital Ward	Medical Center	Office	Emergency Room	Coffee Shop	Pharmacy	Deli / Bodega	Food Court	Fast Food Restaurant

Table 5: Most common visited venues according to each neighborhood

From the table above, we can see several venues regarding restaurants, cafes, and bars. However, we still cannot be sure about which of them belong to which cluster. We need to feed these data into a machine learning model before we make any assumptions about the data.

4. Results

After training the K-means clustering model with 5 numbers of clusters, we can inspect the cluster labels. For viewing purposes, we can add these predictions into the previous dataset (Table 6).

	Postalcode	Latitude	Longitude	Borough	Neighborhood	Cluster Labels	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
37	M4E	43.676357	-79.293031	East Toronto	The Beaches	1	School	Park	Building	Playground	Laundry Service	Residential Building (Apartment / Condo)	Jewelry Store	Salon / Barbershop	Doctor's Office	Martial Arts School
41	M4K	43.679557	-79.352188	East Toronto	The Danforth West, Riverdale	1	Greek Restaurant	Spa	Salon / Barbershop	Miscellaneous Shop	Gym / Fitness Center	Women's Store	Office	Health Food Store	Pharmacy	Fruit & Vegetable Store
42	M4L	43.668999	-79.315572	East Toronto	India Bazaar, The Beaches West	1	Convenience Store	Park	Office	Laundry Service	Pet Store	Rental Car Location	Church	Car Wash	Light Rail Station	Pizza Place
43	M4M	43.659526	-79.340923	East Toronto	Studio District	1	Building	Office	Automotive Shop	Coffee Shop	Restaurant	Nail Salon	Spa	Seafood Restaurant	Furniture / Home Store	Moving Target
44	M4N	43.728020	-79.388790	Central Toronto	Lawrence Park	1	College Classroom	College Auditorium	School	Housing Development	Fast Food Restaurant	Building	Bus Line	Park	College Theater	Parking

Table 6: Combined cluster label predictions with the dataset

This is where latitudes and longitudes information come in handy. We can pinpoint the exact location of each venue on the map (Figure 1). Each color represents which venue belongs to which cluster. For instance, points with purple color belong to the 1st cluster, teal color for the 2nd cluster, red color for the 3rd cluster, and orange and blue for the 4th and 5th cluster respectively.

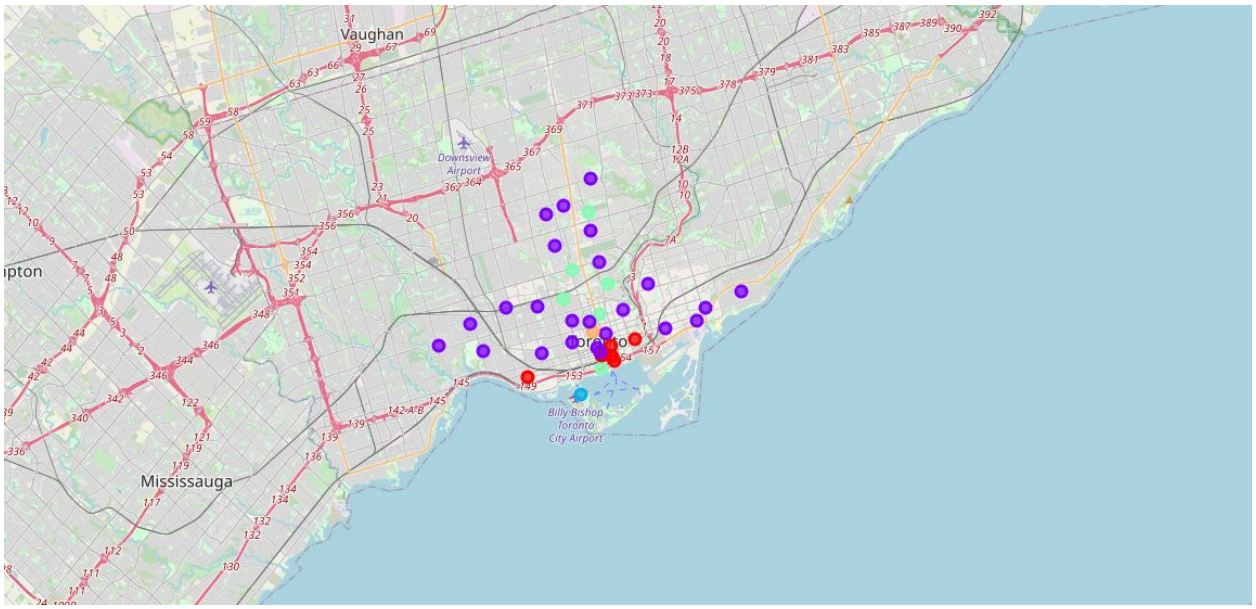

Figure 1: The clustering results plotted on a map.

In general, we can divide each cluster according to its top 10 most common venues. Cluster one is where offices and buildings run their businesses. Most of the venues consist of office, building, tech startup, and similar services. The next cluster has the most venues which consists mostly restaurants, cafes, and shops. We can also analyze that there are several services like laundries, gyms, health & beauties, bars, and lounges. Moving on to the next cluster, we can see it from the map that the location is pretty far away from the city center. Apparently, this is the only airport that Toronto has (Toronto Pearson International Airport). The following cluster focuses on living places like residential buildings/condo. Finally, the last cluster is for health and

medical purposes. We can see here that the most common visited venues are hospitals, hospital wards, medical centers, emergency rooms, and pharmacies.

5. Discussion

Naturally, if someone wants to open a restaurant, more crowded places are preferred rather than the quiet ones. This is normal because it attracts people who usually pass by the same street everyday. According to the results, the 1st and 2nd cluster suit our purposes better than the rest of the clusters. Specifically, I would recommend the 2nd cluster (restaurants & cafes). Even on the map, we can see that these places are located nearby with each other with fair distances. Office workers who are finding a place to eat in their lunch break for example, can be our target customers due to the close distance between office and restaurant clusters. There is no need to consider opening a restaurant inside the hospital and airport clusters since it is located far away from the city without similar venues near them.

6. Conclusion

In this report, I predicted and analyzed the best preferred location to open a Japanese restaurant around Toronto neighborhoods in Canada based on the number of most common venues from each neighbor that can be obtained from the Foursquare API. Through clustering algorithms like K-means, data can be clustered (grouped) together according to their features. In this case, neighbors with similar venues are clustered while other neighbors with dissimilar features belong to other clusters. In addition, utilizing the Foursquare API, I obtained each venue's position (latitude and longitude), then plotted it into a map for easier visualization. Looking at the visualization for each cluster, we can draw a conclusion on the best location to open a Japanese restaurant in order to maximize profit and long-term success.