# Finding a Suitable Sushi Restaurant in Toronto

**Dezhen Kong**

**2020/06/05**

# 1. Introduction

## 1.1 Background

Sushi has been a well-known Asia food for years and is popular in Toronto, therefore, it has a huge market in Toronto. The idea of this project is derived from there maybe not enough sushi restaurant in some areas of Toronto which cannot satisfy the need of the public. By investigating the tendency of sushi restaurant distribution, this project serves as a guide for people who want to open a sushi restaurant.

## 1.2 Objective

In this project, I am aiming to create a hypothetical scenario for people who want to open a new sushi restaurant in Toronto.

## 1.3 Target audience

The target audience in this project is people who want to open a new sushi restaurant in Toronto.

# 2. a acquisition and cleaning

To solve this problem, the following data are used.

- List of neighborhoods in Toronto
- The geographical information (latitude and longitude) of these neighborhoods.
- Venue data related to the sushi restaurant

## 2.1 Data acquisition

- Scrapping the list of neighborhoods in Toronto via Wikipedia.
- Getting latitude and longitude data via Geocoder package.
- Using Foursquare API to get venue data related to these neighborhoods.

## 2.2 Data cleaning

There are several problems in Neighborhoods data:

First, the neighborhood data downloaded has a lot of boroughs that is “not assigned”. Therefore, I only use cells with assigned boroughs in this project.

Second, more than one neighborhood can exist in one postal code area. Therefore, I merge two rows with the same postal code into one row with the neighborhoods separated with a comma.

Third, some cell has a borough but a not assigned neighborhood. In this project, these neighborhoods will be the same as the borough.

After fixing these problems, the neighborhood data is ready to merge with the geographical location data.

## 3. Methodology

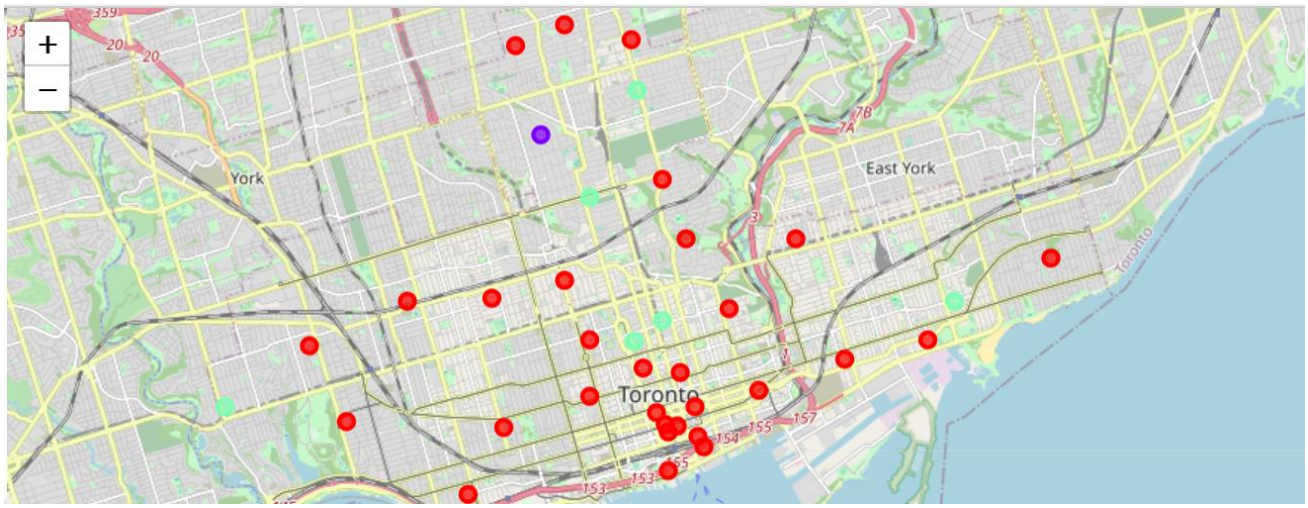
First, I extracting the list of neighborhoods from Wikipedia ("https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:M "). I did the web scraping by utilizing the pandas html table scraping method as it is easier and more convenient to pull tabular data directly from a web page into a dataframe. Since it is only a list of neighborhood names and postal codes but without coordinates which is vital in finding the venues by Foursquare and creating the distribution map, I further get the coordinates by using the csv file provided by IBM team to match the coordinates of Toronto neighborhoods.

After gathering all these coordinates, I visualized the map of Toronto using the Folium package to verify whether these are correct coordinates.

Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. From Foursquare, I am able to obtain the names, categories, latitude and longitude of the venues. With this data, I sort different venues by using sort_value function to find the popular venue in this city, which is helpful to identify industry that is marketable and less competitive. After that, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later. Here, I made a justification to specifically look for "Sushi Restaurants".

Lastly, I performed the clustering method 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 it is highly suited for this project as well. I have clustered the neighborhoods in Toronto into 3 clusters based on their frequency of occurrence for "Sushi Restaurant". Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

## 4. Results



The results from k-means clustering show that we can categorize Toronto neighborhoods into clusters based on number of sushi restaurants are in each neighborhood:

- Cluster 0(red circle): Neighborhood with no Thai restaurants
- Cluster 1(purple circle): Neighborhoods with little restaurants
- Cluster 2(light green circle): Neighborhoods with high number of Thai restaurants

## 5. Discussion

Most sushi restaurants are in Cluster 2 which is around Wellesley, Davisville, and The Beaches West areas. Except for those areas, it's very hard to find a sushi restaurant. Therefore, by looking at the nearby venue, it looks like the area around Ossington, Dufferin, and Lansdowne are lack of a sushi restaurant. Therefore, this project recommends the entrepreneur to open a sushi restaurant in these locations with little or no competition but still have a market.

It should be noted that I only take into consideration of one factor: the distribution of sushi restaurants in each neighborhood in this project. In fact, factors should be taken into consideration such as residents' preference, population density, the income of residents, rent that could influence the decision to open a new restaurant. Future research is needed to take into consideration these factors.

## 6. Conclusion

In this project, we have gone through the process of identifying an economic problem that finds a suitable place for opening a sushi restaurant, specifying the data required, extracting and preparing the data, performing the machine learning by utilizing k-means clustering and providing recommendation to the stakeholder. Finally, it concluded that the area around Ossington, Dufferin, and Lansdowne are promising candidates to be a new place for a sushi restaurant.