**Where to make Sushi**

Levllyn Rocha

15th May 2020

# Introduction

## Background

Asian cuisine is much sought after in Toronto. Since the city is huge location there could be a lot of areas that could be a possible location of a new restaurant. In the restaurant business location matters a lot and since historically the suburbs are not an ideal location this study focusses on a part of the city area.

There are many ways to make a decision on location. This study aims at using cluster analysis on venue / location data to narrow down our choices. Note that the study aims to only broadly define possible locations and further narrowing down would depend on a lot more factors not considered here

## Problem

A budding restauranteur is all set to take his idea of a Japanese restaurant to fruition. This analysis focuses specifically on West Toronto for the optimum location to set up a Japanese Restaurant based on the areas most frequented for eating and locations of existing restaurants.

## Interest

This analysis would be beneficial to budding restaurateurs who have the same use-case of opening a Japanese Restaurants. The logic used could of course be expanded or tweaked to fit the appropriate needs such as other cuisines, other locations or even other business models should the right data be available.

# Data

## Data Sources

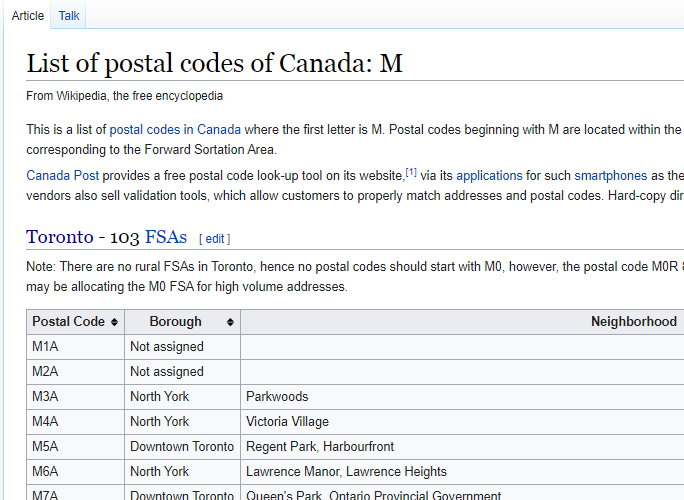
There are 4 primary sources of data both publically available related to

* + The neighborhoods and boroughs in Toronto. Their location and postal code is a necessity.
  + Coordinate data for Each Postal code
  + Details of the most visited places in Toronto where we will be using data from FourSquare API
  + Geopy to get Coordinates where required

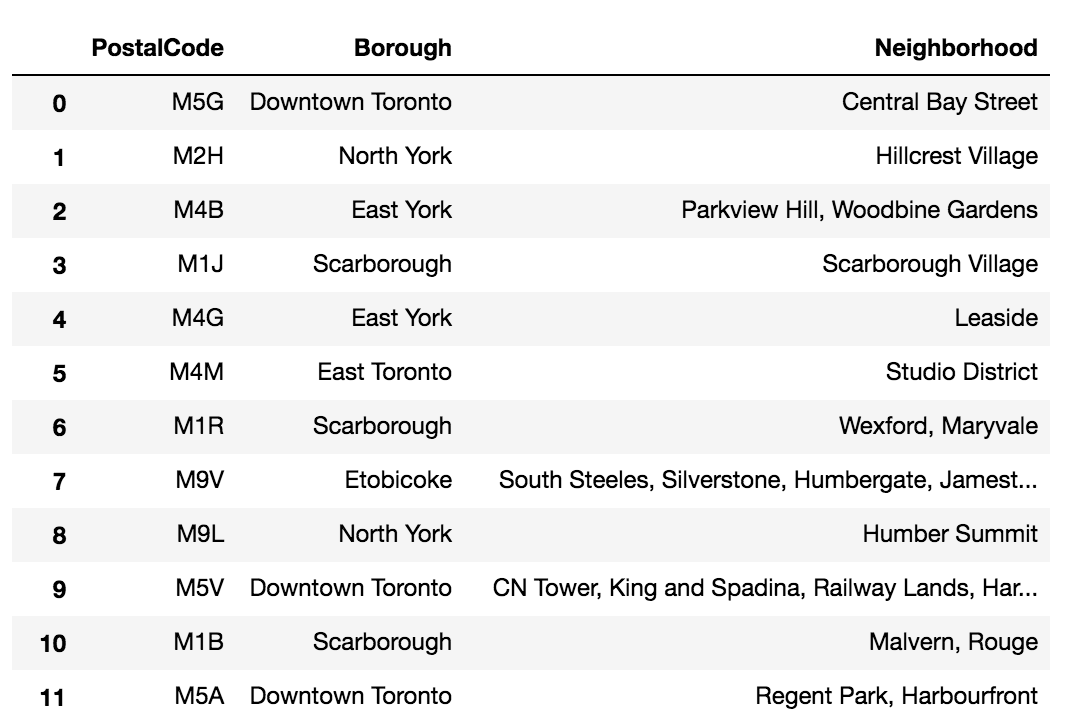
1. Postal Areas with Neighborhoods & Buroughs

We will be using data from Wikipedia on Postal codes beginning with M in Canada which covers all area in the main city of Toronto

[https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M,](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) will be used in order to obtain the data that is in the table of postal codes



To transform the data into a *pandas*dataframe like the one shown below:



While there are different website scraping libraries and packages in Python. For scraping the above table, we will be using pandas to read the table into a pandas dataframe.

1. Coordinate data for postal codes is got from the below location:

<http://cocl.us/Geospatial_data>

This is stored into a CSV to obtain and subseauently added to the data fra

1. Venue Data from Foursquare

Foursquare is a technology company that built a massive dataset of location data. They actually crowd-sourced their data and had people use their app to build their dataset and add venues and complete any missing information they had in their dataset. Currently its location data is the most comprehensive out there, and quite accurate that it powers location data for many popular services like Apple Maps, Uber, Snapchat, Twitter and many others

Calls to the API return structured data like below:



1. JSON file from Geopy to get coordinate data



## Data Cleaning & Selection

Basic cleansing steps like analyzing and removal of NULLS is required.

Data from Postal code is taken as the base 🡪 Coordinate data is added to this from COCL.US site 🡪 This is used to pull venue data from each location.

Since we are focusing on West Toronto, the CN Tower is taken as border all coordinate to the West of CN Tower are retained.

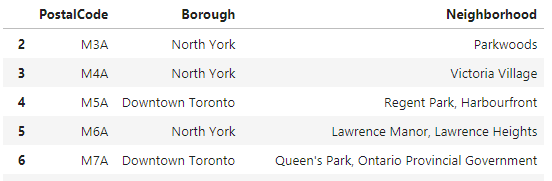
One-hot encoding is done from the venues obtained from Foursquare which is the base of the analysis.

The data is changed into frequency and modeled so we get the top 10 venues for each location after the require grouping is done.

# Methodology

## Reading the Postal Code Data

Data was read using Pandas, with html5lib as flavor. Blank data was dropped as required. The initial dataframe is as below:

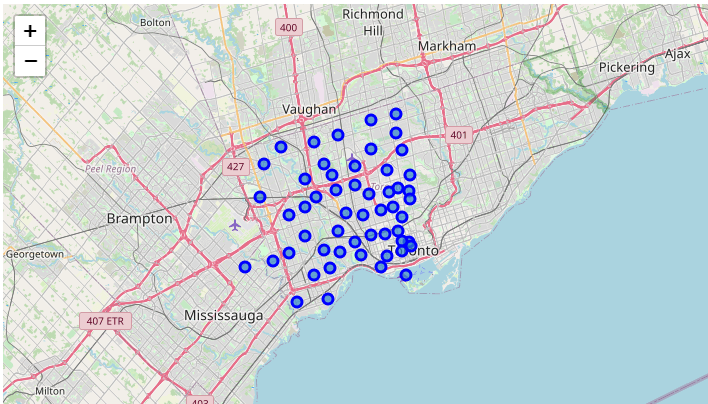


Using Geospatial data, Latitude & Longitude are added to this.

## Refining the data to only Western half of Toronto

The next step is to pick out only West Toronto with that being the target area of this analysis. This was achieved by deleting data east of the coordinates of CN Tower

Plotting data we have”



## Acquiring Venue data from Foursquare & Processing

Venue of top 100 places within 500m of the locations are procured with calls to the Foursquare API.

* + For each neighborhood we get a json file
  + Data is appended to a dataframe
  + After grouping by Neighborhood, we use one hot encoding to break up the venues
  + From this we the frequency of the top 10 venues in that location

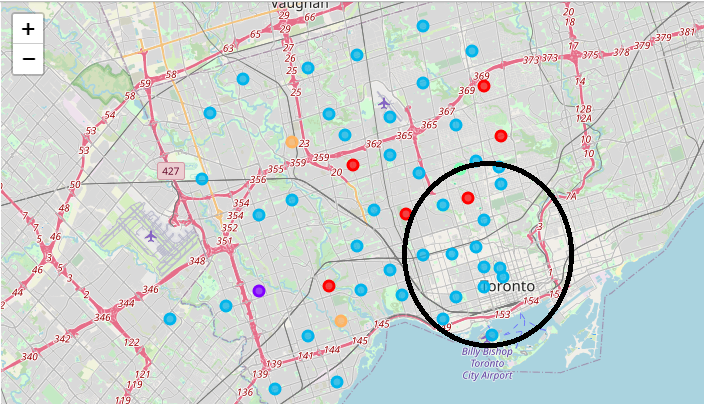
The data is now ready for Cluster Analysis

## Cluster Analysis

For this analysis, to determine the different area cluster we turn to k-means

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. It is popular for cluster analysis in data mining. k-means clustering minimizes within-cluster variances (squared Euclidean distances),

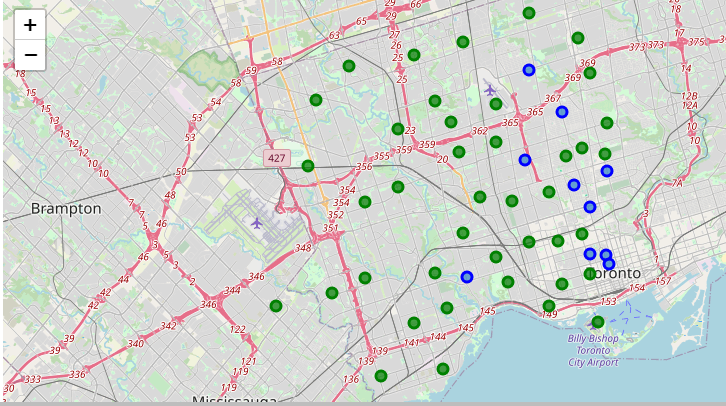
With trial and error we pick 5 as the optimum number of clusters for the analysis.



Post using kmeans, we generate our output visually. On analysis of the 5 clusters venues cluster 2 (light blue) seems to be the best neighborhoods for restaurant. The circled area shows the highest concentration of neighborhoods.

## Drilling down to Existing Japanese Restaurants

Picking out only cluster 2, we delve further into where existing restaurants are present.



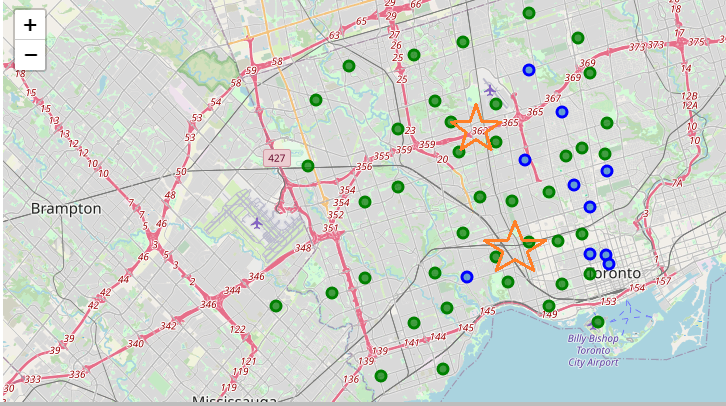
The blue points show existing Japanese / Sushi Restaurants in Cluster 2.

Green points are cluster 2 neighborhoods without Japanese Restaurants.

# Results section

The K means clustering clearly shows Cluster 2 has being the main cluster for restaurants. The South East corner of the Grid has a particular density around restaurants making it a good choice for your new restaurant.

However looking at the distribution of existing Japanese Restaurants in the grid we can visually see that there are already a plethora of Japanese restaurants in this area which would be unnecessary competition.



## Recommendation

Neighborhoods in the marked areas may have a better bet as they are also popular for restaurants and there will be a lack of competition with other Japanese outlets.

* + These areas are also sufficiently close the neighborhoods where cluster 2 is dense and popular for restaurants
  + Being more West may also make your restaurant first choice for the entire western part of West Toronto as it will be one of the closest to them.

# Conclusion

In this study we used venue popularity data with Machine learning techniques like k mean to come up with a proposition for where the optimum location for a restaurant would be

Many other factors would need to be taken into consideration like supply chain, parking, rents, labor, the scale of your restaurant to match with the area etc.

We hope this model will be of use to future restaurant owner and that a lot of Sushi will be made for us to enjoy!