# The Battle of Neighborhoods
# A Restaurant in Toronto

Shuiqi Cheng

March 25, 2020


## 1 Introduction

Nowadays, with the requirement of consuming raising in a rapid way, we are likely to eat out frequently. For this, a lot of businessmen are now planning to open some restaurants in Toronto. More importantly, how to choose the address of these restaurants is of vital significant, which may affect the income of their business directly. Thus, in this assignment, I will use what I'v already learned to help them find a relatively **good position for them to open a restaurant**.

## 2 Analyze of the problem

- **density of the population in the area** since we are to open a restaurant, we must focus on the issue that if the area holds the enough population so that our business is more likely to trait more consumers. for this, we need to utilize the data in the following section.
- **distribution of other restaurants** in many cases, it's not enough for us just to hold a lot potential consumers. Because if lots of other restaurants opening here, we cannot make sure that the consumers would like to try ours. So, it's also very important for us to measure if there are a lot other restaurants.
- **tendency of consuming(eating out)** besides what we have just analyzed, people's opinion of consuming may vary from area to area. For this reason, we try to find the area that people are more likely to eat out use the API.
- **target audience** this project is useful for someone who is about to open a restaurant in Toronto. Also, please pay attention to the fact that since we are going to use the Foursquare API to analyse, our conclusions are time dependent.

## 3 Experimental Data

As we have analyzed, we will use the data below in our model.

- **Borough-Neighborhood information** in order to obtain the density of population, we may as well assume the neighborhood information can represent the population information in some way. We can observe the data in the link of https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. Shown as figure1
- **position information** as we have obtained the dataframe of Borough-Neighborhood information, we need to transform form them into latitude and longitude. we can simply use the Google Maps Geocoding API. As figure2

| | 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 |

Figure 1: data example 1

| | PostalCode | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | M5A | Downtown Toronto | Regent Park,Harbourfront | 43.654260 | -79.360636 |
| 1 | M5B | Downtown Toronto | Garden District, Ryerson | 43.657162 | -79.378937 |
| 2 | M5C | Downtown Toronto | St. James Town | 43.651494 | -79.375418 |
| 3 | M5E | Downtown Toronto | Berczy Park | 43.644771 | -79.373306 |
| 4 | M5G | Downtown Toronto | Central Bay Street | 43.657952 | -79.387383 |

Figure 2: data example 2

- **restaurant distribution information** in this part, we can use the foursquare API and the position information combined with the folium lib, we can visualize the position of the restaurants and shops in the area(figure3)


- **tendency of consuming(eating out)** since we just focus on the information of Toronto, it is reasonable to assume that the consuming ability is relatively the same, so that we can simplify the model and mainly focus on the position and restaurants.

# 4 Modeling Method

## 4.1 Obtain the Data Frame

In this section, I utilized the re and Beautifulsoup lib to get the data from the data source. Then, I cleaned the data for example, I dropped the rows with 'Not assigned'.
And to get the position of each neighborhood, I downloaded the csv document Geospatial Coordinates.csv. Eventually, the position data is added into our dataframe shown below.

| | PostalCode | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | M5G | Downtown Toronto | Central Bay Street | 43.657952 | -79.387383 |
| 1 | M2H | North York | Hillcrest Village | 43.803762 | -79.363452 |
| 2 | M4B | East York | Parkview Hill,Woodbine Gardens | 43.706397 | -79.309937 |
| 3 | M1J | Scarborough | Scarborough Village | 43.744734 | -79.239476 |
| 4 | M4G | East York | Leaside | 43.709060 | -79.363452 |
| 5 | M4M | East Toronto | Studio District | 43.659526 | -79.340923 |
| 6 | M1R | Scarborough | Wexford,Maryvale | 43.750072 | -79.295849 |
| 7 | M9V | Etobicoke | South Steeles,Silverstone,Humbergate,Jamestown... | 43.739416 | -79.588437 |
| 8 | M9L | North York | Humber Summit | 43.756303 | -79.565963 |
| 9 | M5V | Downtown Toronto | CN Tower,King and Spadina,Railway Lands,Harbou... | 43.628947 | -79.394420 |
| 10 | M1B | Scarborough | Malvern,Rouge | 43.806686 | -79.194353 |
| 11 | M5A | Downtown Toronto | Regent Park,Harbourfront | 43.654260 | -79.360636 |

Figure 3: data frame

## 4.2 Neighborhood Distribution

As what I have shown in the part above, in order to have more consumers, we have to observe the density of the population in each borough. In this section, we can simply take the neighborhoods as a symbol of the population. Thus, we can focus on the number of neighbors each borough holds. Then we calculate the number and plot the pie chart. We can see that the North York holds a

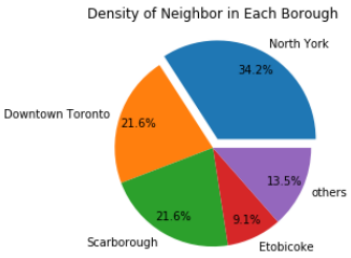


Figure 4: pie chart

relatively large population. So I select the data and use the folium lib to plot the map figure 5. with the help of joint plot figure 6, it is reasonable to divide this area into **two** clusters. Simply utilize the **KMeans**, we can find 2 centers of the

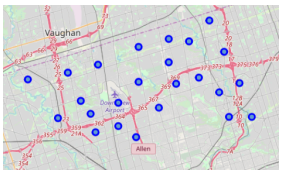


Figure 5: neighbors in North York

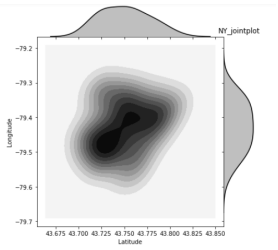


Figure 6: location joint plot


Figure 7: two center points

area, which matches the expectation of what we called **population density**. As shown, the red points are our center point based on the population density (figure7).

## 4.3 Nearby Restaurants

**First**, I apply the **Foursquare API** to obtain the nearby venues information around the neighborhoods in the North York. After selecting the restaurants out, with **one-hot** method, I can obtain the dataframe of neighborhoods with the number of restaurants in the region. Then I plot the data in a horizontal bar chart in figure 8 The reason why we need to measure the restaurants in the area

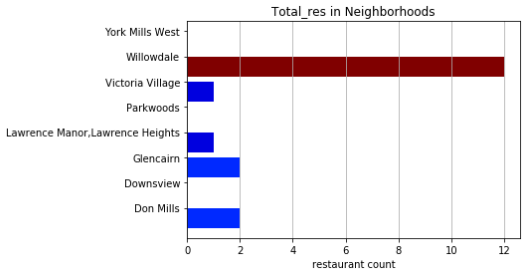


Figure 8: neighbor-restaurant barh

is that since people have almost the same tendency to eat in each restaurants when they hang out. In this case, the probability (P) of this tendency can be seemed as $\frac{1}{(restaurants+1)}$
**second**, based on what we have discussed, we can use the probability as a weight to calculate the average point of the restaurants existed. We call the point as the restaurant center. Because we want more consumers to get into our business, we cannot open our restaurant in the 'restaurant center'. The equation below can be used to calculate the coordinate of our restaurant center:

$$coordinate = \frac{\sum_{n=0}^{k} L_n \times w_n}{\sum_{n=0}^{k} w_n} \tag{1}$$

where:

- $k$ is the total neighborhoods in this borough
- $L$ is the coordinate of each neighborhood
- $w_n$ is the number of restaurants in each neighborhood

### 4.4 Co-Analyze

In this section, we are going to find the best location for us to open a restaurant, considering both the two aspects we discussed in the previous sections, as we know, Neighborhood Distribution and Nearby Restaurants.
Plotting the 'restaurant center' in the map and we can notice that it can be clustered into the right area, which can also be improved by our distance function(the yellow point in figure 9).
Thus, we will choose our position more likely in the left area. What to do next is to divide the neighbors into the 2 new groups, according to the distance towards our left center point and the restaurant center. I utilize the **K-means** function, where $k = 1$ to realize it. It is beneficial to do so, since we can find the new center of the 2 groups (the two black points in figure 9). And the center relatively far from the restaurant center is in a residential area with relatively few competitors. This point is the best position to open a restaurant (c1 in figure 9).

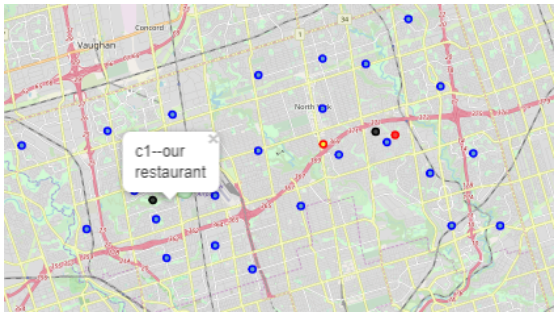


Figure 9: position choice

## 5 Results

In this model, the best position to open a restaurant in Toronto is at (43.73575162, -79.4973781). As is analyzed, we are more likely to have relatively more customers with correspondingly lower competition with other restaurants.
And we can also find in this model that areas in the right cluster may be more flourishing than the left one, in terms of catering.

## 6 Discussion

What's more, we can find that most of the restaurants are located in Willowdale in the right cluster. Oppositely, few of restaurants are clustered into the left cluster, while they have almost the same population density. This provide the opportunity for us to open the restaurant here and have more space to develop.

Interestingly, we first cluster the data into 2 circles depend on the distribution of the neighborhood since it's just like a rectangle. The truth is that the distribution of restaurants meets this well. As we can observe in the experiment data, our final location is close to that of the center created by our first cluster work.

# 7 Conclusion

Although, there's some features that we didn't take into consideration for example the income of each borough may be different and cause the different tendency to consuming. And because we have no access to gain the acreage and accurate population size of each area, we just consider the number of neighborhoods as the substitution of the population density in each borough. In terms of the existing data we've gathered, the result has relatively high credibility.
Also, since we just focus on the specific area, Toronto, the method is reliable, and the features ignored are likely to have less influence of our result.