# Best location to open new Chinese Restaurant in Toronto.

Aibek Satbayev

26.07.2021

Table of contents

[Best location to open new Chinese Restaurant in Toronto. 1](#_Toc78194515)

[1. Introduction/Business Problem. 1](#_Toc78194516)

[2. Data acquisition and cleaning. 1](#_Toc78194517)

[2.1. Data sources. 1](#_Toc78194518)

[2.2 Data cleaning. 1](#_Toc78194519)

[3. Exploratory Data Analysis. 2](#_Toc78194520)

[4. Clustering of Neighborhoods. 2](#_Toc78194521)

[5. Conclusion. 2](#_Toc78194522)

## Introduction/Business Problem.

Owner of the global chain of Chinese restaurants wants to open brand new restaurant in Toronto as a part of the strategic development plan in North America.

A restaurant's location is one the most important factors for its success. So, it is critical to find the best location.

He requested our firm to propose the best Toronto neighborhood to establish a Chinese restaurant.

## Data acquisition and cleaning.

### Data sources.

Following data will be used for analysis:

* The Toronto's crime rates data. (https://open.toronto.ca/dataset/neighbourhood-crime-rates/)
* Geographical coordinates of each neighborhood: (https://www.openstreetmap.org).
* The Foursquare API will be used to explore neighborhood venues in Toronto. We will check presence of Chinese restaurants in each neighborhood to check possible competition.
* The Toronto's census data for its social demographic characteristics (https://bit.ly/3airrOJ). We will obtain spending power, total population and distribution of Chinese population.

### 2.2 Data loading and cleaning.

Toronto census data was downloaded from <http://map.toronto.ca/wellbeing/> into the ‘welbeing\_toronto.csv’ file. Total population, Chinese population and Household Income for each neighborhood were selected. Data was imported into ‘df\_census’ dataframe. I decided to select neighborhoods with more than average total population and Chinese population values. Based on statistical info of dataframe I found that mean value of total population was 20 000 and for Chinese population was 2 000. After applying these filters to neighborhoods data, there left 25 neighborhoods out of 140. I dropped ‘Combined Indicators’ column because it was with empty values.

Toronto crime rates data was downloaded from <https://open.toronto.ca/dataset/neighbourhood-crime-rates/> into the ‘neighbourhood-crime-rates.csv’ file. Data was imported into ‘df\_crime’ dataframe. I decided to summarize only 2020 year’s all crime types’ rates into one ‘TotalCrimeRate2020’ feature to have more relevant and more complex information. I dropped all the rest columns of dataframe and left only Neighborhood Name and ID, TotalCrimeRate2020.

I merged ‘df\_census’ and ‘df\_crime’ dataframes based in Neighborhood ID to add crime rate data into the Neighborhoods census data.

I tried to obtain 25 selected Neighborhoods’ coordinates through <https://www.openstreetmap.org> , but could not do it for 5 Neighborhoods. The reason was that in that map the names of those Neighborhoods were different. So I changed the names to be able to obtain coordinates.

After obtaining coordinates for each Neighborhood I added it into Neighborhoods dataframe.

## Exploratory Data Analysis.

I created a map to visualize selected Neighborhoods.

I used Foursquare API to explore top 100 venues in each selected Neighborhood. All explored venues were stored into ‘toronto\_venues’ dataframe. I analyzed each Neighborhood with One Hot encoding. I grouped the rows by Neighborhood and by taking the mean of the frequency of occurrence of Chinese restaurants and saved the result in ‘toronto\_grouped’ dataframe.

I visualized frequency distribution of Chinese restaurants in each Neighborhood using Bar graph. And found that only 12 Neighborhoods out of 25 had Chinese restaurants. ‘Milliken’ and ‘Steeles’ Neighborhoods had the most density of Chinese restaurants. Hence, they are not advisable for business operations to be taken place there.

I visualized Chinese population distribution in each Neighborhood using Bar graph. And found that there were Top 5 Neighborhoods the most populated by Chinese. They were: ‘Milliken’, ‘Steeles’, ‘Willowdale East’, ‘Agincourt North’ and ‘L’Amoureaux’. Hence, they are mostly advisable for business operations to be taken place there.

I visualized Household income in each Neighborhood using Bar graph. And found that Household income is more or less equally distributed by 25 Neighborhoods. With one exception – ‘Rouge’ had outstanding and highest value comparing to others. Hence, it is advisable for business operations to be taken place there.

I visualized Total population distribution in each Neighborhood using Bar graph. And found that there were Top 6 Neighborhoods by Total population. They were: ‘Toronto Islands’, ‘Woburn’, ‘Willowdale East’, ‘Rouge’, ‘L’Amoureaux’ and ‘Malvern’. Hence, they are advisable for business operations to be taken place there.

I visualized Total population distribution in each Neighborhood using Bar graph. And found that there were Top 2 Neighborhoods with the highest Total crime rates. They were: ‘Church-Yonge Corridor’ and ‘Bay Street Corridor‘. Hence, they are not advisable for business operations to be taken place there.

## Clustering of Neighborhoods.

I finalized dataframe for clustering – added ‘Chinese Restaurant’ feature to ‘df\_hoods’.

I normalized the dataset. But why do we need normalization in the first place? Normalization is a statistical method that helps mathematical-based algorithms to interpret features with different magnitudes and distributions equally. We use StandardScaler() to normalize our dataset. Now he had ‘df\_normalized’ dataframe with 5 normalized features: ‘Household Income’, ‘Total Population’, ‘Chinese Population’, ‘Crime Rate’ and ‘Chinese Restaurants Density’.

So, how can we choose right value for K? The general solution is to reserve a part of your data for testing the accuracy of the model. Then chose k =3, use the training part for modeling, and calculate the accuracy of prediction using all samples in your test set. Repeat this process, increasing the k, and see which k is the best for your model.

After analysing using elbow method using distortion score & Squared error for each K value, looks like K = 6 is the best value. I clustered the Toronto Neighborhood Using K-Means with K = 6.

I added found Cluster label into ‘df\_hoods\_final’ dataframe and visualized Clustered neighborhoods on the Map.

I analyzed all six clusters and made a following profiling:

* Cluster 0. It consists of neighborhoods with low total population and low competition rate. **Not recommended.**
* Cluster 1. It consists of neighborhoods with high number of Chinese population and low crime rate. But competition rate is very high. **Not recommended.**
* Cluster 2. It consists of neighborhoods with average number of Chinese population, low competition rate, and average Crime rate. **Recommended.**
* Cluster 3. It consists of neighborhoods with comparatively low household income, zero competition rate, and with high Crime rate. **Not recommended.**
* Cluster 4. It consists of neighborhoods with highest household income, average number of Chinese population, zero competition rate, and low Crime rate. **Recommended.**
* Cluster 5. It consists of neighborhoods with high competition rate and low total population. . **Not recommended.**

## Conclusion.

* 'Willowdale East' from Cluster 2 is highly recommended neighborhood to open Chinese restaurant due to high number of Chinese population, low crime rate and zero competition. In addition to that it has more than average total population and household income.
* Second choice is 'Rouge' from Cluster 4 due to highest household income, low crime rate and zero competition.