

## **Capstone project: Find the optimal location to start a restaurant in Toronto**

### **1. Introduction**

Dongjun is a new immigrant from Korea to the Toronto city, and he would like to start a Korean restaurant. According to South Korea's Ministry of Foreign Affairs and Trade, there were 240,942 ethnic Koreans or people of Korean descent in Canada as of 2017. With its diverse demographic of Toronto, we intend to help Dongjun find the optimal location to start a Korean restaurant in the city of Toronto.

### **2. Collecting and Processing Data**

First, we need to obtain the neighborhood and its geospatial data of Toronto from a publicly available database, e.g., <https://open.toronto.ca/dataset/neighbourhood-profiles/>. We download the CSV file and uploaded to IBM cloud.

To open it, we have to use "project.get\_file" The credential is hidden on purpose to protect my privacy.

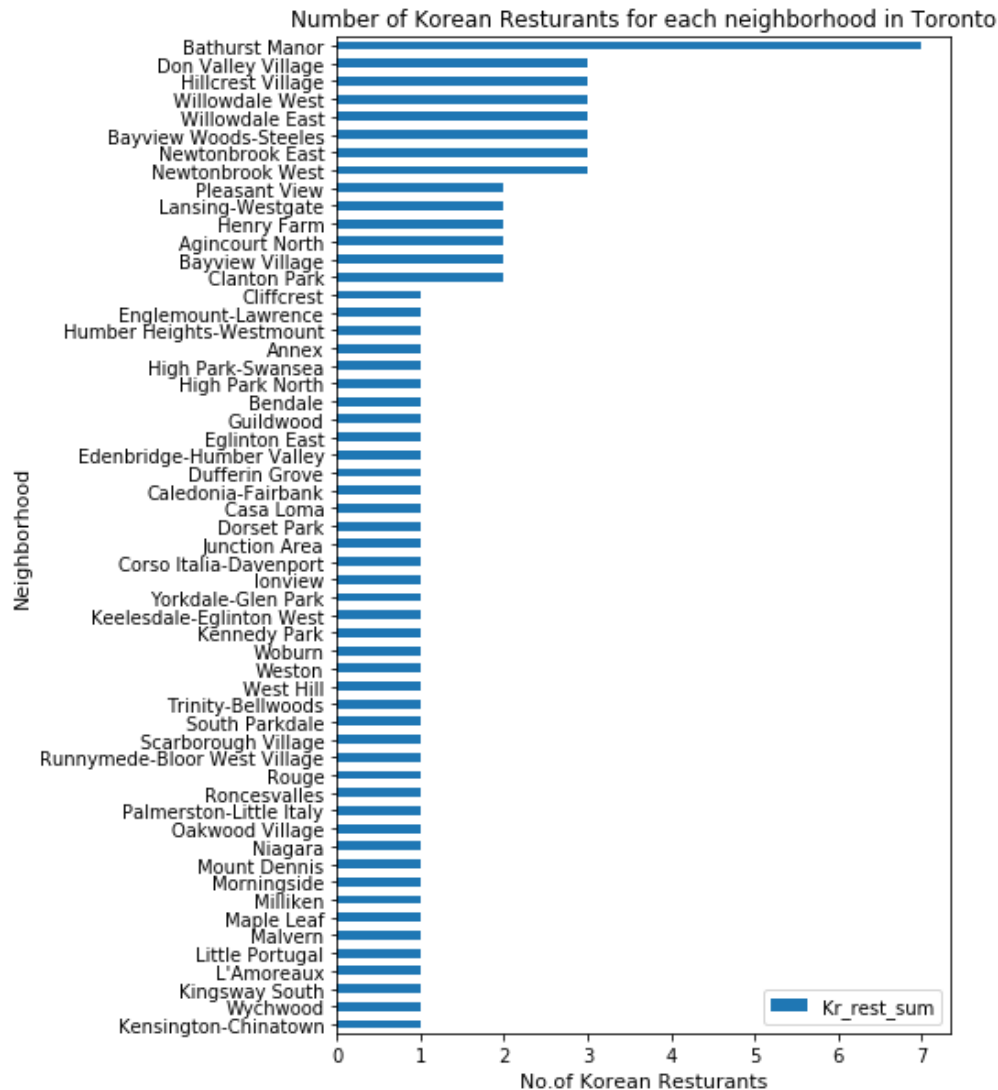
We would like to know our potential customers for Dongjun's business. Hence, we would focus on residents with Korean origin (who can speak Korean, or who characterized themselves of Korean ancestry) in a neighborhood. Since the size of dataframe is huge, only several key data were selected including average individual income and origin related to Korean based on the name of the column in the CSV file. We then cleaned the data, e.g. remove commas and change the datatype of number from object to integer for future analysis.

Here we propose an index to quantify how "Korean" related a neighborhood is. We give different weights to different column. We assume the weight represents "how many times a resident would like to visit a Korean restaurant per month". For example, Using Korean at work has a weight as 16, since it indicates the business need for Korean customers. (almost 4 times per week). Speaking Korean at home is assumed to have a weight of 8. (almost twice per week). Korean as mother tongue is assumed to have a weight of 8. (almost twice per week). Visible minority population, Ethnic origin as Korean all have a weight as 4. (almost once per week). Knowledge of Korean have weights as 1 (almost once per month). We then create a new column called "Korean index", the higher the index, the more Korean related residents in this neighborhood. We can also normalize this index with total population of the neighborhood and get "Normalized index". We then sort the table by normalized index, then found out the Newton brook East has the highest normalized index. The income of each neighborhood is also an important factor, we list average individual income of a neighborhood as a column in the Table.

We then Use Nominatim as geocoder, because geocoder.google keep returning None. Nominatim did a better job than google but coordinates for some neighborhood still cannot be obtained. 35 of 140 neighborhood do not have coordinate. We here just neglect those neighborhoods due to limitation of time to find the coordinate manually. We then use the Foursquare API to obtain detailed information on existing Korean restaurants in each neighborhood with known coordinates and find its competitors in each neighborhood. Lastly, we will combine data from foursquare with demographic data to find a solution to Dongjun's question.

### **3. Results and Discussion**

Based on the results return from Foursquare API, 56 neighborhoods at least have one Korean restaurant. Bathurst Manor has 7 Korean restaurants as shown in the bar charts.



We join it with previous demographic dataframe 'toronto\_kr' obtained from City of Toronto by column "Neighbourhood". Now we can use Korean index, Kr\_rest\_sum and income to do the data analysis. Ideally, we want to find a neighborhood with high Korean\_index, high income, but low Kr\_rest\_sum. But also, can accept a neighborhood with high Korean\_index, high income, also high Kr\_rest\_sum. We would like to avoid a neighbourhood with low Korean\_index, low income, also low Kr\_rest\_sum. Hence, we can plot all the neighborhood with these three values as coordinates in the 3D space. We normalize the data and visualize them using scatter plot in 3D as shown below. Here we found an apparent outlier, Willowdale East, which has very high Korean index, relatively good average income, but very low Korean restaurant number as 1. Hence it could be our option.



