**Capstone Project (Proposal)**

# The Battle of Neighborhoods

YW
6/12/2019

## 1. Introduction

### 1.1. Background

Food-In Pro, Inc. is a well-known supplier in the manufacturing of powdered food ingredients (e.g. salt, sugar, flour and yeast) and food additives (e.g. sweetener, artificial color, creamer and softener) in Canada as well as worldwide. Due to the growth of business, the company decides to open a small distribution warehouse in the city of Toronto for serving the local customers effectively. Before the financial budget of investment is evaluated, the company would first like to know if there is an optimal location/neighborhood in Toronto for the warehouse based on the needs from target and/or potential clients.

According to the sales and marketing strategies, the most clients in a city are roughly summarized and rated into five categories depending on the business scale and stability as below:

**Table 1. Categories of clients with business description**

| Client's Category | Business description | Score rate |
|---|---|---|
| Supermarket and grocery | Large and stable business | 5 |
| Restaurant | Large business but varied stability | 4 |
| Food spot, court, place, (H)house, pub, joint, diner... | Medium business and varied stability | 3 |
| Ice cream, bakery, dessert, chocolate, donut, smoothie and cafe | Small business and varied stability | 2 |
| Other food stores | Small and unstable business | 1 |

### 1.2. Problem Description

The company assigns this project to the data analyst to explore the distribution of the above categorical clients against the neighborhood in the city of Toronto. The analyst is expected to give a recommendation to the facility address searching group about the best area of neighborhood(s) to locate the warehouse in the city, according to the distribution of clients. The principles of priority for the recommendation of location are: 1) more amount of high rate clients (especially supermarkets and groceries), then 2) more amount of total clients.

## 2. Data Acquisition and Cleaning

### 2.1 Data Sources

The data of neighborhoods in the city of Toronto are acquired from Wikipedia (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) and scraped by the BeautifulSoup package. The geographical coordinates are acquired

from the course material. The location data of those categorical venues are acquired by querying Foursquare API. The keyboards sent for querying are “Supermarket”, “Restaurant”, and “Ice cream”, where the other categories can also be returned. The venue lists are acquired and combined into one dataframe.

2.2 Data Selection and Cleaning

Only “Toronto boroughs” (Downtown Toronto, East Toronto, West Toronto and Central Toronto) and their neighborhoods are selected for evaluation in this project. “Not assigned” neighborhoods under the assigned postcodes are also dropped. Venue categories are examined to exclude unreasonable ones (e.g. Pharmacy). Then, the venues are identified by venue categories as shown in **Table 2**. This is the initial dataframe for the analysis.

**Table 2. Neighborhood venues in client categories (partial view)**

| Neighbourhood | Neighbourhood Latitude | Neighbourhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|
| Ryerson | 43.657162 | -79.378937 | Metro | 43.658404 | -79.376748 | Supermarket |
| Garden District | 43.657162 | -79.378937 | Metro | 43.658404 | -79.376748 | Supermarket |
| St. James Town | 43.651494 | -79.375418 | Metro | 43.649027 | -79.373313 | Supermarket |
| Berczy Park | 43.644771 | -79.373306 | Loblaws | 43.644462 | -79.369486 | Supermarket |
| Berczy Park | 43.644771 | -79.373306 | Metro | 43.649027 | -79.373313 | Supermarket |
| Central Bay Street | 43.657952 | -79.387383 | The Market by Longo's Elizabeth | 43.655357 | -79.385115 | Supermarket |
| Central Bay Street | 43.657952 | -79.387383 | Metro | 43.660569 | -79.383768 | Supermarket |
| Christie | 43.669542 | -79.422564 | Fiesta Farms | 43.668471 | -79.420485 | Supermarket |
| Christie | 43.669542 | -79.422564 | Loblaws | 43.671807 | -79.421102 | Supermarket |
| Adelaide | 43.650571 | -79.384568 | Rabba Marché | 43.649216 | -79.386908 | Supermarket |

2.3 Feature Identification

There are over 2800 venues provided at those neighborhoods. Venue category is the principle feature that contributes to score the rates of neighborhoods which gives us the candidate(s). Neighborhood coordinates are used to map the neighborhood variables, and may contribute to search the compromised location if the candidates are not adjacent. Although there is another type of coordinates for venues which gives a more accurate information, the tiny variance gives difficulty to mapping and observing since we are targeting the neighborhoods. They are therefore dropped in the following analysis.

## 3. Methodology

I. Rank the neighborhoods by the number of venues
II. Rank the neighborhoods by scoring the rates based on **Table 1**. Explore the difference between the two ranks.
III. Rank the neighborhoods by the number of supermarkets and groceries.
IV. Run k-mean to cluster the neighborhoods to identify the top common venues.
V. Examine the ranks and clusters, and recommend the candidate(s) according to the principles in Section 1.2.