**Capstone Project - The Battle of Neighborhoods**

# Introduction

We want to open new shop in Toronto, Ontario, Canada.

The main problem is, where the best location to open a new shop is and what best shop in certain locations of Toronto is?

In order to solve this problem, the data about the shops of each neighbourhood will be collected to show what are the current shops at each locations, how many similar shops, demand of the shops, what type of shops and more.

# Data

The data that consisted of postal codes, neighbourhood and borough of Canada will be collected.

The number of shops of each neighbourhood will be collected through Foursquare API by combining with the data that consisted of postal codes, neighbourhood and borough of Canada.

After combining the data, the data will be showed what are the shops at each neighbourhood and what type of shops are high demand.

Postal codes, neighbourhood and borough of Canada: [***https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M***](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

# Methodology

In order to make the final result more accurate, some exploratory data analysis will be processing such as cleaning and modification.

1. Exploratory Data Analysis
   1. Data Cleaning

The cells that having “Not assigned” under borough will be ignored and process with the cells that have an assigned borough.

The neighbourhoods that having missing values will be assigned the same borough to the neighbourhood for that column.

* 1. Data Modification

There are many neighbourhood that having same postal code area which means that more than one neighbourhood can exist in one postal code area.

For instances, postal code area M5A has two neighbourhoods which is Harbour front and Regent Park. The neighbourhoods that having the same postal code area will be combined together with a comma and put in the same postal code area.

Moreover, in order to utilize the foursquare location data, two new variables which is latitude and the longitude coordinates of each neighbourhood need to add into the cleaned data.

* 1. Machine Learning

The machine learning method that will be using is K means clustering. This is because this method able to cluster the result into group which is like what are the shops at that location.

By looking at the cluster, the user can understand it easily and make the best decision.

# Result and Discussion

The final result is in the table format which listed out all the top 10 most common shops in that location by the postal code area, borough, neighbourhood, longitude and latitude.