**Applied Data Science Capstone**
**Joshua Triana**
**April 2020**

## Introduction

In this project, we want to explore a strategic location for a new coffee shop in the Toronto Area. We want to convince a client looking to invest in the Toronto Area.
Coffee shops are a desired business all around the world, so it's a “safe business” to invest in. Nevertheless, we want to analyze the best possible location in the Toronto Area in order to increase the odds.

## BUSINESS PROBLEM

The main objective is to find the best location to open a new coffee shop in Toronto, Canada. We will be aiming to apply all the knowledge gained in this specialization in order to answer the business question: In Toronto, if new clients want to open a new coffee shop, where would be the best location¡

## TARGET AUDIENCE

A client looking to open a new coffee shop in Toronto Area

## DATA SOURCES

To solve this problem, we will need below data:
● List of neighborhoods in Toronto, Canada
(https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
● Latitude and Longitude of these neighborhoods (http://cocl.us/Geospatial_data)
● Venue data related to Coffee shops (https://api.foursquare.com)

## DATA CLEANING

● Scrap of Toronto neighborhoods inWikipedia
● Using Foursquare API to get venue data
● Latitude and Longitude by using Geocoder package

## METHODOLOGY

We need to extract data from:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
Then we are going to scrap the table data to a dataframe and then merge it with the corresponding coordinates.

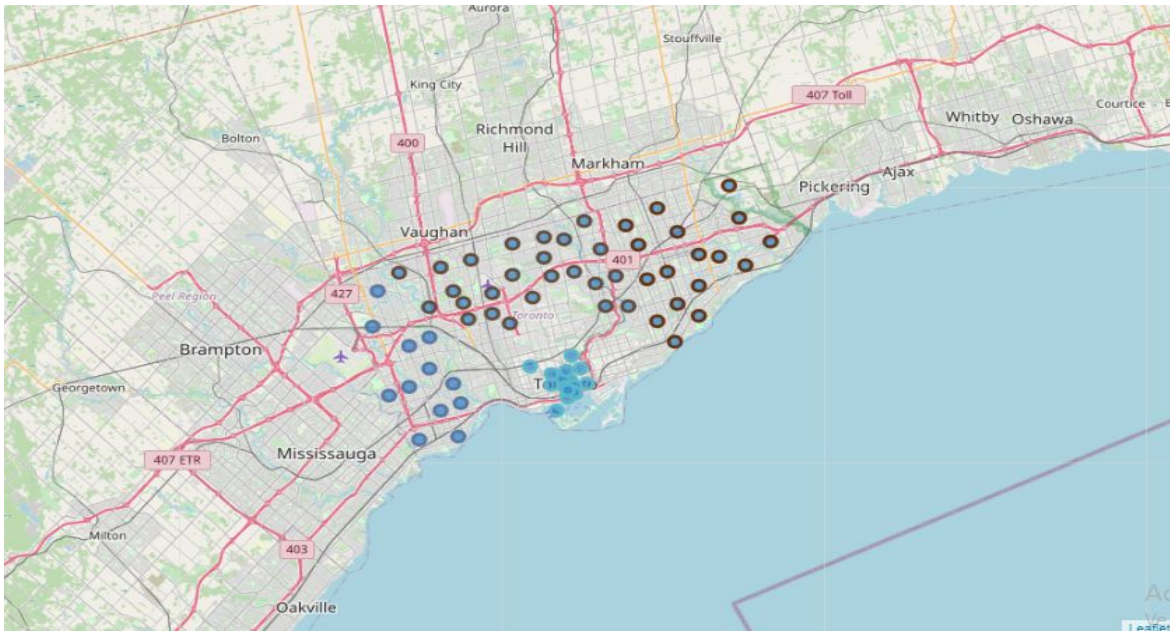
```
df_merge.head()
```

| | Postcode | Borough | Neighbourhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | M3A | North York | Parkwoods | 43.753259 | -79.329656 |
| 1 | M4A | North York | Victoria Village | 43.725882 | -79.315572 |
| 2 | M5A | Downtown Toronto | Regent Park / Harbourfront | 43.654260 | -79.360636 |
| 3 | M6A | North York | Lawrence Manor / Lawrence Heights | 43.718518 | -79.464763 |
| 4 | M7A | Downtown Toronto | Queen's Park / Ontario Provincial Government | 43.662301 | -79.389494 |

The map's visualization will be key as a first approach to understand the data.
Then, we are going to use the Foursquare API to get the venues data.
We will take the top four boroughs to work with. The idea is to put the coffee shop in key and popular locations.

```
Borough
Central Toronto      9
Downtown Toronto    19
East Toronto         5
East York            5
Etobicoke           12
Mississauga          1
North York          24
Scarborough         17
West Toronto         6
York                 5
```

The analysis of frequencies by groups will be necessary in order to apply k-means.



We can analyze a lot of venue categories:

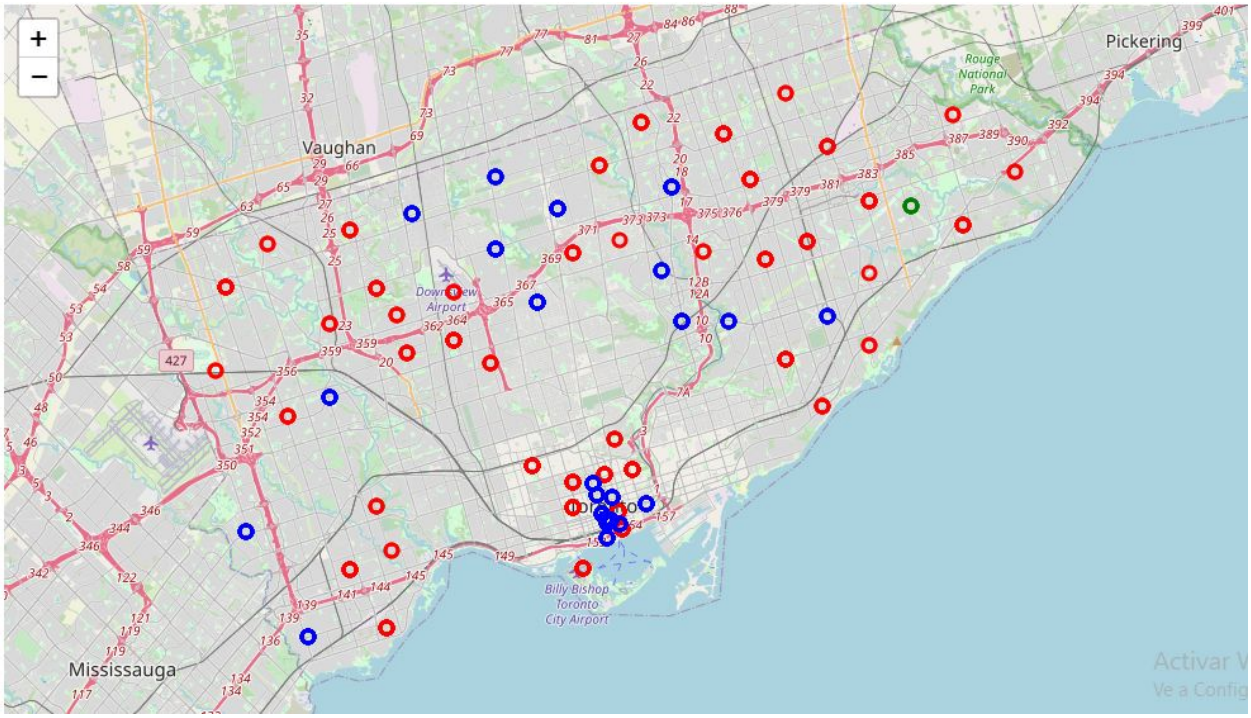
```
array(['Park', 'Food & Drink Shop', 'Hockey Arena', 'Coffee Shop',
       'Portuguese Restaurant', 'Pizza Place', 'Bakery',
       'Distribution Center', 'Spa', 'Breakfast Spot', 'Restaurant',
       'Historic Site', 'Pub', 'Farmers Market', 'Chocolate Shop',
       'Dessert Shop', 'Theater', 'Performing Arts Venue',
       'Gym / Fitness Center', 'French Restaurant', 'Café',
       'Mexican Restaurant', 'Event Space', 'Yoga Studio',
       'Ice Cream Shop', 'Shoe Store', 'Art Gallery', 'Cosmetics Shop',
       'Electronics Store', 'Bank', 'Beer Store', 'Hotel',
       'Health Food Store', 'Wine Shop', 'Antique Shop', 'Boutique',
       'Furniture / Home Store', 'Vietnamese Restaurant',
       'Clothing Store', 'Accessories Store', "Women's Store",
       'Arts & Crafts Store', 'Miscellaneous Shop', 'Italian Restaurant',
       'Beer Bar', 'Creperie', 'Burrito Place', 'Diner', 'Hobby Shop',
       'Discount Store', 'Fried Chicken Joint', 'Burger Joint',
       'Juice Bar', 'Sandwich Place', 'Gym', 'College Auditorium', 'Bar',
       'Music Venue', 'Fast Food Restaurant',
       'Construction & Landscaping', 'Caribbean Restaurant',
       'Japanese Restaurant', 'Baseball Field', 'Comic Shop', 'Plaza',
       'Tea Room', 'Ramen Restaurant', 'Thai Restaurant', 'Movie Theater',
       'Steakhouse', 'Shopping Mall', 'American Restaurant',
       'College Rec Center', 'Gastropub', 'Bookstore', 'Tanning Salon',
       'Middle Eastern Restaurant', 'Modern European Restaurant',
       'Seafood Restaurant', 'Lake', 'Department Store',
```

We have decided to work with a Coffee shop, since it is the clients requirements, but we can work with a lot if we want to.

Kmeans is a popular and very used machine learning algorithm and is appropriate to this problem. We get two clusters for coffee shops and finally analyze the data in order to get the best cluster to open the new coffee shop

## RESULT

K-means clustering show that we can categorize Toronto neighborhoods into 2 clusters




```
In [88]: to_merged.loc[(to_merged['Cluster Labels'] ==0) & (to_merged['Venue Category'] == 'Coffee Shop') ]
```

Out[88]:

| | Neighborhood | Coffee Shop | Cluster Labels | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|---|---|
| 32 | Lawrence Manor / Lawrence Heights | 0.062500 | 0 | 43.718518 | -79.464763 | Tim Hortons | 43.719427 | -79.467995 | Coffee Shop |
| 30 | Kensington Market / Chinatown / Grange Park | 0.064516 | 0 | 43.653206 | -79.400049 | Tim Hortons | 43.653303 | -79.405457 | Coffee Shop |
| 30 | Kensington Market / Chinatown / Grange Park | 0.064516 | 0 | 43.653206 | -79.400049 | I Deal Coffee | 43.655058 | -79.403254 | Coffee Shop |
| 30 | Kensington Market / Chinatown / Grange Park | 0.064516 | 0 | 43.653206 | -79.400049 | Dark Horse Espresso Bar | 43.650564 | -79.397018 | Coffee Shop |
| 30 | Kensington Market / Chinatown / Grange Park | 0.064516 | 0 | 43.653206 | -79.400049 | Little Pebbles | 43.654883 | -79.400264 | Coffee Shop |
| 51 | Steeles West / L'Amoreaux West | 0.062500 | 0 | 43.799525 | -79.318389 | Tim Hortons | 43.799102 | -79.318715 | Coffee Shop |
| 50 | St. James Town / Cabbagetown | 0.069767 | 0 | 43.667967 | -79.367675 | Jetfuel Coffee | 43.665295 | -79.368335 | Coffee Shop |
| 50 | St. James Town / Cabbagetown | 0.069767 | 0 | 43.667967 | -79.367675 | Tim Hortons | 43.667169 | -79.368849 | Coffee Shop |
| 50 | St. James Town / Cabbagetown | 0.069767 | 0 | 43.667967 | -79.367675 | Tim Hortons | 43.665786 | -79.368284 | Coffee Shop |
| 49 | St. James Town | 0.058824 | 0 | 43.651494 | -79.375418 | Everyday Gourmet (Teas & Coffees) | 43.648757 | -79.371645 | Coffee Shop |
| 49 | St. James Town | 0.058824 | 0 | 43.651494 | -79.375418 | Fahrenheit Coffee | 43.652384 | -79.372719 | Coffee Shop |
| 49 | St. James Town | 0.058824 | 0 | 43.651494 | -79.375418 | Versus Coffee | 43.651213 | -79.375236 | Coffee Shop |
| 49 | St. James Town | 0.058824 | 0 | 43.651494 | -79.375418 | Aroma Espresso Bar | 43.648224 | -79.376352 | Coffee Shop |
| 49 | St. James Town | 0.058824 | 0 | 43.651494 | -79.375418 | Tim Hortons | 43.647955 | -79.373833 | Coffee Shop |
| 11 | Church and Wellesley | 0.066667 | 0 | 43.665860 | -79.383160 | Rooster Coffee House | 43.669654 | -79.379871 | Coffee Shop |
| 11 | Church and Wellesley | 0.066667 | 0 | 43.665860 | -79.383160 | The Drink | 43.664167 | -79.380149 | Coffee Shop |
| 11 | Church and Wellesley | 0.066667 | 0 | 43.665860 | -79.383160 | Croissant Tree | 43.669575 | -79.382331 | Coffee Shop |
| 7 | CN Tower / King and Spadina / Railway Lands / ... | 0.071429 | 0 | 43.628947 | -79.394420 | Balzac's Coffee Roasters | 43.631392 | -79.395952 | Coffee Shop |
| 5 | Berczy Park | 0.054545 | 0 | 43.644771 | -79.373306 | Starbucks | 43.644285 | -79.369771 | Coffee Shop |
| 5 | Berczy Park | 0.054545 | 0 | 43.644771 | -79.373306 | Everyday Gourmet (Teas & Coffees) | 43.648757 | -79.371645 | Coffee Shop |
| 5 | Berczy Park | 0.054545 | 0 | 43.644771 | -79.373306 | Mos Mos | 43.641640 | -79.377552 | Coffee Shop |
| 10 | Christie | 0.055556 | 0 | 43.669542 | -79.422564 | Starbucks | 43.671530 | -79.421400 | Coffee Shop |
| 11 | Church and Wellesley | 0.066667 | 0 | 43.665860 | -79.383160 | Piedmont Coffee Bar | 43.668286 | -79.382520 | Coffee Shop |
| 11 | Church and Wellesley | 0.066667 | 0 | 43.665860 | -79.383160 | Starbucks | 43.664980 | -79.380510 | Coffee Shop |

```
In [89]: to_merged.loc[(to_merged['Cluster Labels'] ==1) & (to_merged['Venue Category'] == 'Coffee Shop') ]
```

Out[89]:

| | Neighborhood | Coffee Shop | Cluster Labels | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|---|---|
| 9 | Central Bay Street | 0.187500 | 1 | 43.657952 | -79.387383 | Jimmy's Coffee | 43.658421 | -79.385613 | Coffee Shop |
| 9 | Central Bay Street | 0.187500 | 1 | 43.657952 | -79.387383 | Starbucks | 43.659456 | -79.390411 | Coffee Shop |
| 9 | Central Bay Street | 0.187500 | 1 | 43.657952 | -79.387383 | Hailed Coffee | 43.658833 | -79.383684 | Coffee Shop |
| 9 | Central Bay Street | 0.187500 | 1 | 43.657952 | -79.387383 | Neo Coffee Bar | 43.660140 | -79.385870 | Coffee Shop |
| 44 | Richmond / Adelaide / King | 0.082474 | 1 | 43.650571 | -79.384568 | HotBlack Coffee | 43.650364 | -79.388669 | Coffee Shop |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 25 | Harbourfront East / Union Station / Toronto Is... | 0.120000 | 1 | 43.640816 | -79.381752 | Mos Mos | 43.641640 | -79.377552 | Coffee Shop |
| 25 | Harbourfront East / Union Station / Toronto Is... | 0.120000 | 1 | 43.640816 | -79.381752 | Lavazza Espression | 43.639537 | -79.381763 | Coffee Shop |
| 25 | Harbourfront East / Union Station / Toronto Is... | 0.120000 | 1 | 43.640816 | -79.381752 | Pilot Coffee Roasters | 43.645018 | -79.380415 | Coffee Shop |
| 25 | Harbourfront East / Union Station / Toronto Is... | 0.120000 | 1 | 43.640816 | -79.381752 | Balzac's Coffee | 43.644373 | -79.383065 | Coffee Shop |
| 14 | Commerce Court / Victoria Hotel | 0.100000 | 1 | 43.648198 | -79.379817 | Starbucks | 43.647261 | -79.378599 | Coffee Shop |

118 rows × 9 columns

- Cluster 0: Neighborhoods with the less number of coffee shops(24)
- Cluster 1: Neighborhoods with more coffee shops(118)

This results recommends the client to open the coffee shop around cluster 0, with less competition