# Coursera Capstone Project

## IBM Applied Data Science

Dipak Jani

**Topics Cover in the Reports**

- Introduction
- Business Problem
- **Procedure for implementing the setting of Objectives**
  - Resources Use
  - Data Sourcing
  - Implicate Data Science and Machine Learning
- Analysis and Results
- Discussion
- Conclusion

## Introduction

This is the part of the Capstone Project for the IBM Data Science Professional Certificate.

In this document, I have to setup hypothetical business scenario using Neighbourhoods of the large city to analyse and search suitable location for the set criteria.

I have to apply Machine Learning/Data science methods to meet my set objectives described in the next section.

## Project Description

### Business Background:

XYZ. Ltd is a business in Hampshire, UK, distributing Asian food ingredients to the Local Small Shops and Supermarkets in UK. It owns its own warehouse and packing facility, sourcing ingredients mainly from South East Asia and South America.
They have food processing facility in Gujarat, India and Distribution facility in Dubai for Middle East and African markets.

XYZ is now looking for Organic Growth on the North American Continent and have identified Toronto, Canada for expansion, targeting both wholesale and retail sectors.
Toronto was selected, on bases of

a) Presence of a large Indian, Pakistani and Chinese communities
b) Stable socio-economic and political systems
c) Close Trade and Political relations with UK.

d) Close proximity and amicable trade relations with USA for future expansion into USA.

**Business Objectives:**

To select a site to set up a Supermarket where,

- There are other small and large businesses, to obtain a good foot fall.
- Ideally based around Central Toronto for easy access from all neighbourhoods of Toronto.
- With good public transport links and good Car Parking facilities.

I will use internet search to source the relevant data. I will then analyse the data and convey my interpretations and suggestions to XYZ.

## Procedure for implementing the setting of Objectives:

To set objectives in Project Description, I will employ following resources and procedures to set up the project.

**Data Sourcing:**

- Source Toronto Neighbourhood data with its location in Toronto.
  https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M,
- Geocoding data for linking Longitude and Latitude to the Neighbourhoods.
  http://cocl.us/Geospatial_data and Geocoder
- Sourcing Venue data using Foursquare.
- Internet search for Asian population in Toronto, socio/economic data of Neighbourhoods.
  Demographics of **Toronto** - Wikipedia
  https://en.wikipedia.org/wiki/Demographics_of_Toronto

**Resources Use:**

- Install Python-3 on Jupyter Notebook using Mac/Windows operating system.
- Install all standard libraries for Python.
- Install BeautifulSoup, Geocoder, Request, Folium, Jason and other software as require.
- Test, Jupyter Notebook, Foursquare and Github are functioning properly.

**Implicate Data Science and Machine Learning:**

- Set up New Jupyter Notebook.
- Import and setup all libraries and software.
- Scrape the Wikipedia for Toronto data using BeautifulSoup and transform the data to pandas dataframe
- Download Longitude and Latitude for the Neighbourhoods and merged it.
- Clean and explore the dataset for processing.

- Capture Venues and Categories data using Foursquare.
- Segment and cluster the data for evaluation.
- Draw conclusion and recommend to XYZ.

# Data

## Data Sourcing:

- Source Toronto Neighbourhood data with its location in Toronto.
  https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M,

  Use BeautifulSoup to scrape Wikipedia to collect Toronto Postal, Bourgh and Neighbourhood data

```
n [3]: url = requests.get('https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M').text
       soup = BeautifulSoup(url,'lxml')
```

```
n [4]: table_post = soup.find('table')
       fields = table_post.find_all('td')

       postcode = []
       borough = []
       neighborhood = []

       for i in range(0, len(fields), 3):
           postcode.append(fields[i].text.strip())
           borough.append(fields[i+1].text.strip())
           neighborhood.append(fields[i+2].text.strip())

       df_pc = pd.DataFrame(data=[postcode, borough, neighborhood]).transpose()
       df_pc.columns = ['Postcode', 'Borough', 'Neighborhood']
       df_pc.head()
```

ut[4]:

| | Postcode | Borough | Neighborhood |
|---|---|---|---|
| 0 | M1A | Not assigned | Not assigned |
| 1 | M2A | Not assigned | Not assigned |
| 2 | M3A | North York | Parkwoods |
| 3 | M4A | North York | Victoria Village |
| 4 | M5A | Downtown Toronto | Regent Park, Harbourfront |

- Geocoding data for linking Longitude and Latitude to the Neighbourhoods. Merged Geocode data with the Toronto Data
  http://cocl.us/Geospatial_data and Geocoder

| | Postcode | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | M1B | Scarborough | Malvern, Rouge | 43.806686 | -79.194353 |
| 1 | M1C | Scarborough | Rouge Hill, Port Union, Highland Creek | 43.784535 | -79.160497 |
| 2 | M1E | Scarborough | Guildwood, Morningside, West Hill | 43.763573 | -79.188711 |
| 3 | M1G | Scarborough | Woburn | 43.770992 | -79.216917 |
| 4 | M1H | Scarborough | Cedarbrae | 43.773136 | -79.239476 |

- Sourcing Venue data using Foursquare
  Use Foursquare API and Geocode to scape Venues data and merged with the Central Toronto data.

| | Postcode | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | M4N | Central Toronto | Lawrence Park | 43.728020 | -79.388790 |
| 1 | M4P | Central Toronto | Davisville North | 43.712751 | -79.390197 |
| 2 | M4R | Central Toronto | North Toronto West, Lawrence Park | 43.715383 | -79.405678 |
| 3 | M4S | Central Toronto | Davisville | 43.704324 | -79.388790 |
| 4 | M4T | Central Toronto | Moore Park, Summerhill East | 43.689574 | -79.383160 |

- Internet search for Asian population in Toronto, socio/economic data of Neighbourhoods.
  Demographics of **Toronto** - Wikipedia

  https://en.wikipedia.org/wiki/Demographics_of_Toronto

## Method for Cleaning the data:

Check how many venues were returned for each neighbourhood.
Apply One Hot for making dataset suitable for analysis.

| | Neighborhood | American Restaurant | BBQ Joint | Bagel Shop | Bank | Breakfast Spot | Brewery | Burger Joint | Bus Line | Café | ... | Supermarket | Sushi Restaurant | Swim School | Tennis Court | Rest |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Davisville | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.000 | 0.028571 | 0.00 | 0.000000 | 0.057143 | ... | 0.000000 | 0.057143 | 0.000000 | 0.0 | 0.( |
| 1 | Davisville North | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.125 | 0.000000 | 0.00 | 0.000000 | 0.000000 | ... | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.( |
| 2 | Forest Hill North & West, Forest Hill Road Park | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.000 | 0.000000 | 0.00 | 0.000000 | 0.000000 | ... | 0.000000 | 0.250000 | 0.000000 | 0.0 | 0.( |
| 3 | Lawrence Park | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.000 | 0.000000 | 0.00 | 0.333333 | 0.000000 | ... | 0.000000 | 0.000000 | 0.333333 | 0.0 | 0.( |
| 4 | Moore Park, Summerhill East | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.000 | 0.000000 | 0.00 | 0.000000 | 0.000000 | ... | 0.000000 | 0.000000 | 0.000000 | 1.0 | 0.( |
| 5 | North Toronto West, Lawrence Park | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.000 | 0.000000 | 0.00 | 0.000000 | 0.052632 | ... | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.( |
| 6 | Roselawn | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.000 | 0.000000 | 0.00 | 0.000000 | 0.000000 | ... | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.( |

### Top 10 venues for each neighbourhood

| | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Davisville | Dessert Shop | Pizza Place | Sandwich Place | Gym | Italian Restaurant | Café | Sushi Restaurant | Coffee Shop | Pharmacy | Park |
| 1 | Davisville North | Hotel | Sandwich Place | Park | Breakfast Spot | Gym | Department Store | Food & Drink Shop | Garden | Flower Shop | Fried Chicken Joint |
| 2 | Forest Hill North & West, Forest Hill Road Park | Trail | Park | Jewelry Store | Sushi Restaurant | Yoga Studio | Fried Chicken Joint | Farmers Market | Fast Food Restaurant | Flower Shop | Food & Drink Shop |
| 3 | Lawrence Park | Bus Line | Park | Swim School | Yoga Studio | Fast Food Restaurant | Flower Shop | Food & Drink Shop | Fried Chicken Joint | Garden | Gas Station |
| 4 | Moore Park, Summerhill East | Tennis Court | Yoga Studio | Gas Station | Farmers Market | Fast Food Restaurant | Flower Shop | Food & Drink Shop | Fried Chicken Joint | Garden | Gourmet Shop |

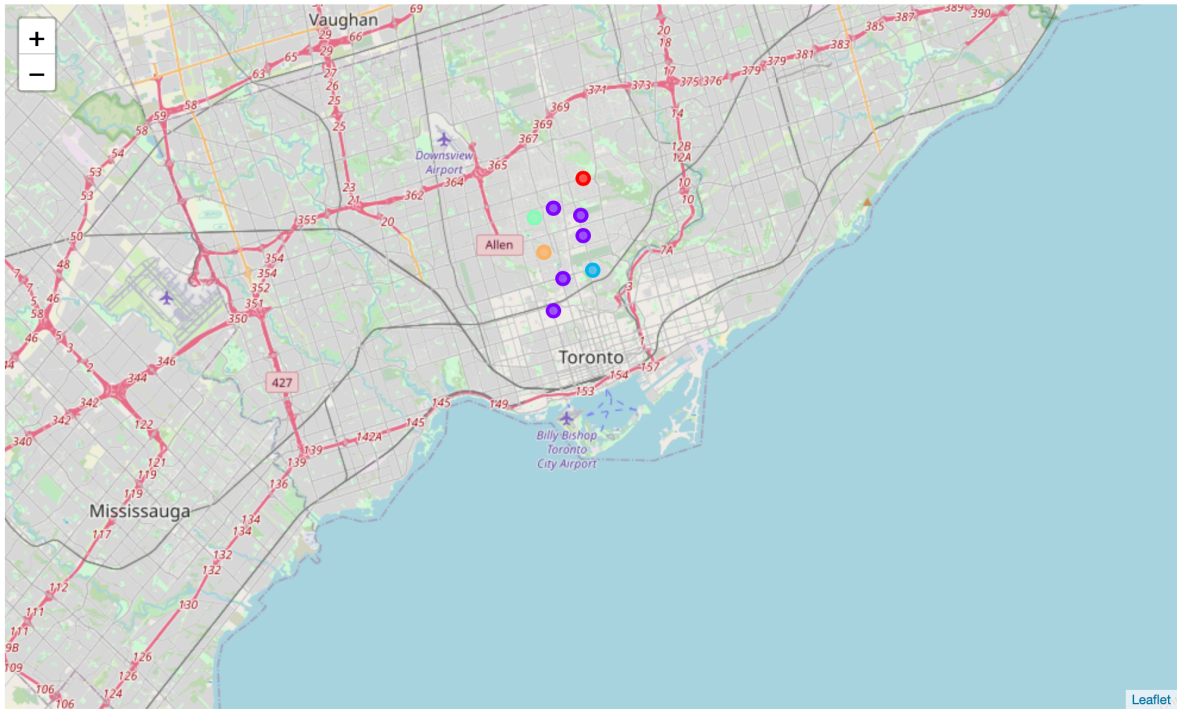
### k-means to cluster the neighbourhood into 5 clusters

K-Means Cluster algorithm used to the venue data to get cluster classification for optimising the Neighbourhood selection.

```
# set number of clusters
kclusters = 5

centrepark_grouped_clustering = centrepark_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(centrepark_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
array([1, 1, 4, 0, 2, 1, 3, 1, 1], dtype=int32)
```

```
map_clusters
```




**Foursquare Query for Supermarket and Carpark data**

To meet the additional requirement of the problem, use Foursquare Query for the Proff of Concept.

The Query was run on the following two category search.

1. Supermarket to establish how many and type of supermarkets are in Central Toronto and their location in the Neighborhood.
2. Availability of the parking is important for attracting the prospective customers.

See below the code for query use and its output.

*search_query = 'Super Market'*
*radius = 1000*
*print(search_query + ' .... OK!')*
*Super Market .... OK!*

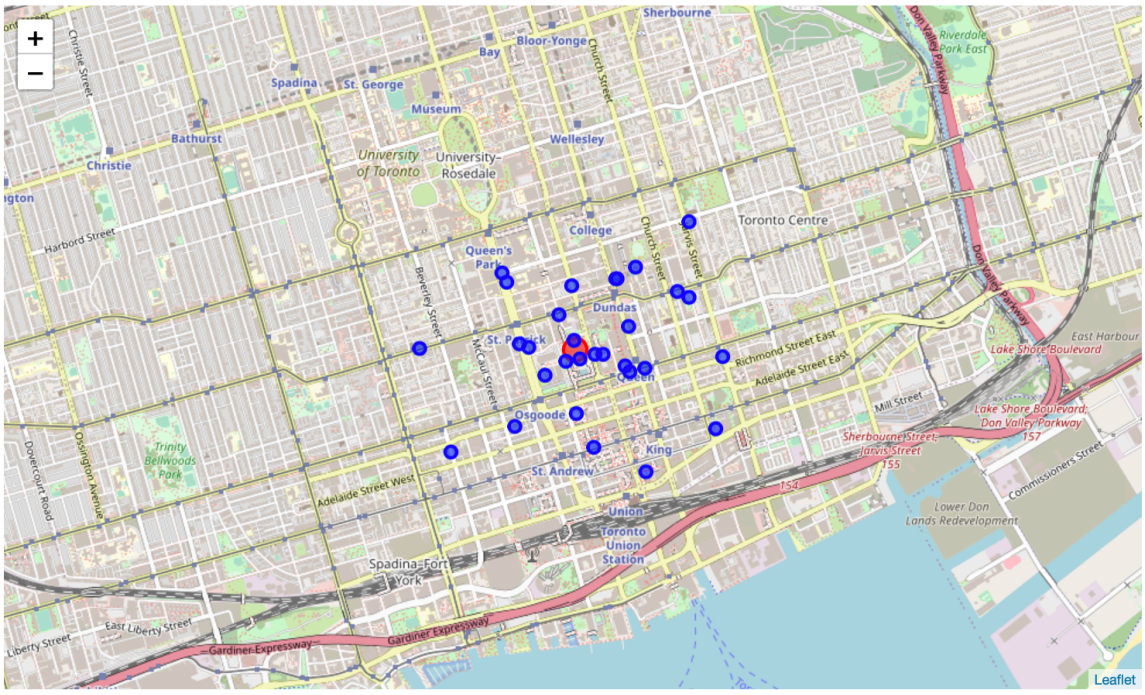
**Supermarket List and Map**

```
dataframe_filtered.name
```

```
0                                  The Supermarket
1                      Starbucks (Market at Longo's)
2                                  City of Toronto
3                                   Farmers Market
4                                Superior Amusements
5                             Marcello's Market and Deli
6      Superior Plus Interlock and Construction Inc.
7             Mike's Independent City Market Toronto
8                    The Market by Longo's Elizabeth
9                                  The 888 Market
10                                   Super wraps
11                           Super Premium Ice Cream
12                             The Market by Longo’s
13                                Super Ordinary Lab
14                             The Market By Longo's
15                               Super Dollar & More
16                                     Super Stop
17               Superior Court of Justice Courthouse
18                                     Super Dollar
19                   The Farmers’ Market at Ryerson
20                Richtree Natural Market Restaurants
21                         Market Square Shoe Repair
22                                     Hasty Market
23                                  Foodwares Market
24                                     Market Spell
25                  Sick Kids Hospital Farmer's Market
26                                  Jarvis Market
27             Nathan Phillips Square Farmer's Market
28                               SickKids Farmers Market
29                                  Tiny Flower Market
Name: name, dtype: object
```

```
# display map
Carpark_map
```

0]:



## Analysis and Results

I use following information from the Data Analysis and Machine Learning for the analysis and the results.

- Analysis of the Venues and Categories.
- Analysis of K-Means Clustering
- Foursquare query search

```
----Forest Hill North & West, Forest Hill Road Park----
                venue  freq
0        Jewelry Store  0.25
1                Trail  0.25
2                 Park  0.25
3     Sushi Restaurant  0.25
4  American Restaurant  0.00


----Lawrence Park----
                venue  freq
0         Swim School  0.33
1                Park  0.33
2            Bus Line  0.33
3  American Restaurant  0.00
4  Rental Car Location  0.00
```

| | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Davisville | Dessert Shop | Pizza Place | Sandwich Place | Gym | Italian Restaurant | Café | Sushi Restaurant | Coffee Shop | Pharmacy | Park |
| 1 | Davisville North | Hotel | Sandwich Place | Park | Breakfast Spot | Gym | Department Store | Food & Drink Shop | Garden | Flower Shop | Fried Chicken Joint |
| 2 | Forest Hill North & West, Forest Hill Road Park | Trail | Park | Jewelry Store | Sushi Restaurant | Yoga Studio | Fried Chicken Joint | Farmers Market | Fast Food Restaurant | Flower Shop | Food & Drink Shop |
| 3 | Lawrence Park | Bus Line | Park | Swim School | Yoga Studio | Fast Food Restaurant | Flower Shop | Food & Drink Shop | Fried Chicken Joint | Garden | Gas Station |
| 4 | Moore Park, Summerhill East | Tennis Court | Yoga Studio | Gas Station | Farmers Market | Fast Food Restaurant | Flower Shop | Food & Drink Shop | Fried Chicken Joint | Garden | Gourmet Shop |

**Cluster 2**

```
]: centrepark_merged.loc[centrepark_merged['Cluster Labels'] == 1, centrepark_merged.columns[[1] + list(range(5, centrepark_merged.shape[1]))]]
```

]:

| | Borough | Cluster Labels | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Central Toronto | 1 | Hotel | Sandwich Place | Park | Breakfast Spot | Gym | Department Store | Food & Drink Shop | Garden | Flower Shop | Fried Chicken Joint |
| 2 | Central Toronto | 1 | Coffee Shop | Clothing Store | Yoga Studio | Spa | Fast Food Restaurant | Metro Station | Mexican Restaurant | Park | Diner | Gym / Fitness Center |
| 3 | Central Toronto | 1 | Dessert Shop | Pizza Place | Sandwich Place | Gym | Italian Restaurant | Café | Sushi Restaurant | Coffee Shop | Pharmacy | Park |
| 5 | Central Toronto | 1 | Pub | Coffee Shop | American Restaurant | Sports Bar | Pizza Place | Liquor Store | Restaurant | Light Rail Station | Skating Rink | Fried Chicken Joint |
| 8 | Central Toronto | 1 | Café | Sandwich Place | Coffee Shop | Indian Restaurant | Park | Pharmacy | Pizza Place | Pub | Liquor Store | Middle Eastern Restaurant |

## Foursquare Query:

- There are 29 Super Markets are there in Central Toronto, and list suggest there are mainly Western (Local), and Chines Super Markets.
- There is both public and private car parks.

**Discussion**

## Cluster Analysis and interpretation

---

**Amont the total 5 clusters, Cluster 2 has the maximum number of Venues and Categories gives good choice and selection.**

**Line 33 Top 5 most common Venues are in Summer Hill West, The Annex and Davisille.**

**Line 35 Top 10 Venues with good Footfall are Davisville, Davisville North, Summer Hill Westand and The Annex.**

**Cluster Analysis based on the Postcode also confirn the above listed Neighbourhoods are best for setup Supermarket.**

---

- **Based on the analysis and results, following Neighborhoods are suitable for setting the Asian Ingredients supermarket.**
- **The selection is based on the following criteria,**
  - **Good mix of the venue categories bring all type of the customers.**
  - **Neighborhood with mainly shops, café, pub, restaurants bring higher footfall.**

| Classification | Neighbourhoods |
|---|---|
| Best | Davisville, Summer Hill West, The Annex |
| Ok | Lawrence Park |
| Avoid | Moore Park |

**Conclusion**

- Toronto is the capital city of Canada, and population of about 6 million.
- About 12 % of the population is Asian, excluding Chines.
- Central Toronto borough with nine Neighborhood has good selection of the venues and its categories bring customers for the shopping.
- It appears, with the growing Asian population there are not many Asian ingredient supermarket in Central Toronto.
- Scope for both wholesale and retail business from same location.
- There is good public transportation facility to Central Toronto from surrounding Boroughs and Neighborhoods.

**The above points suggest Central Toronto is the ideal location for the South Asian Ingredient supermarket**

**Additional Requirements**

Although not within the scope of this project, if XYZ will require following additional resources.

- More Data Analysis and ML for setup SQL for Products, Order and dispatch records.
- Setup Cloud or local server.
- ML regression for predicting sales and marketing.
- Web scare for Canadian food regulation and food label.
- Product Life and expire date.