# IBM Data Science Capstone Project

# Opening a new shopping mall in Toronto, Canada


# By: Milind Hingane


## Table of Contents:

- ## Introduction: -

The project aims to identify venues in Toronto, Canada based on their latitude and Longitude. In this notebook, we will identify various venues in the city of Toronto, Canada using Foursquare API to help property developers to select the location that suits them the best. So, to decide which location is suitable for the shopping mall we required to identify factors such as location cost, it will be profitable or not and so many.

- **Business Problem: -**

The objective of this IBM data science capstone project is to analyze and select the best location for opening a shopping mall in the city of Toronto, Canada. So, for that, we are using the data science methodology and machine learning technique such as clustering, this project provides a solution to the business question.

Business question: -In the city of Toronto, Canada if a property developer is lolling to open a new shopping mall, where would you recommend that they should open it.

- **Integrated Audience: -**

This project is particularly useful to property developers and investors to open or invest in a new shopping mall in the city of Toronto, Canada.

- ## Data Discussion:

  - ### Required Data:

    - We will collect the list of neighborhoods in the city of Toronto using the Wikipedia page(https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
    - Longitude and latitude coordinates of the neighborhoods. This data is required to plot the map and to get venue data.
    - Venue data is related to a shopping mall. We will use this data to perform clustering on the neighborhoods.

  - ### Data collection:

    - To collect the geographical coordinates of the respective postal code or neighborhood I used the CSV file that has the geographical coordinates of each postal code (Neighborhood): (http://cocl.us/Geospatial_data)
    - Using this CSV file to create the following data Frame:

| | Borough | Neighborhood | PostalCode |
|---|---|---|---|
| 0 | North York | Parkwoods | M3A |
| 1 | North York | Victoria Village | M4A |
| 2 | Downtown Toronto | Regent Park / Harbourfront | M5A |
| 3 | North York | Lawrence Manor / Lawrence Heights | M6A |
| 4 | Downtown Toronto | Queen's Park / Ontario Provincial Government | M7A |
| 5 | Etobicoke | Islington Avenue | M9A |
| 6 | Scarborough | Malvern / Rouge | M1B |
| 7 | North York | Don Mills | M3B |
| 8 | East York | Parkview Hill / Woodbine Gardens | M4B |
| 9 | Downtown Toronto | Garden District, Ryerson | M5B |

- Creating a map of Toronto by latitude and longitude values using map folium function.


- Foursquare API:

Using Foursquare API to fetch the nearest venue locations so that we can use them for the cluster. Foursquare API returns the nearest venue in radius and gets corresponding Venue Name, Venue Latitude, Venue Longitude, Venue Category, respectively.

| | Neighborhood | Latitude | Longitude | VenueName | VenueLatitude | VenueLongitude | VenueCategory |
|---|---|---|---|---|---|---|---|
| 0 | Parkwoods | 43.753259 | -79.329656 | Allwyn's Bakery | 43.759840 | -79.324719 | Caribbean Restaurant |
| 1 | Parkwoods | 43.753259 | -79.329656 | Donalda Golf & Country Club | 43.752816 | -79.342741 | Golf Course |
| 2 | Parkwoods | 43.753259 | -79.329656 | Brookbanks Park | 43.751976 | -79.332140 | Park |
| 3 | Parkwoods | 43.753259 | -79.329656 | Island Foods | 43.745866 | -79.346035 | Caribbean Restaurant |
| 4 | Parkwoods | 43.753259 | -79.329656 | Galleria Supermarket | 43.753520 | -79.349518 | Supermarket |
| 5 | Parkwoods | 43.753259 | -79.329656 | LCBO | 43.757774 | -79.314257 | Liquor Store |
| 6 | Parkwoods | 43.753259 | -79.329656 | Graydon Hall Manor | 43.763923 | -79.342961 | Event Space |
| 7 | Parkwoods | 43.753259 | -79.329656 | The Captain's Boil | 43.754986 | -79.349524 | Seafood Restaurant |
| 8 | Parkwoods | 43.753259 | -79.329656 | Me Va Me Kitchen Express | 43.754957 | -79.351894 | Mediterranean Restaurant |
| 9 | Parkwoods | 43.753259 | -79.329656 | Menchie's | 43.754764 | -79.350013 | Frozen Yogurt Shop |

- ## Methodology:

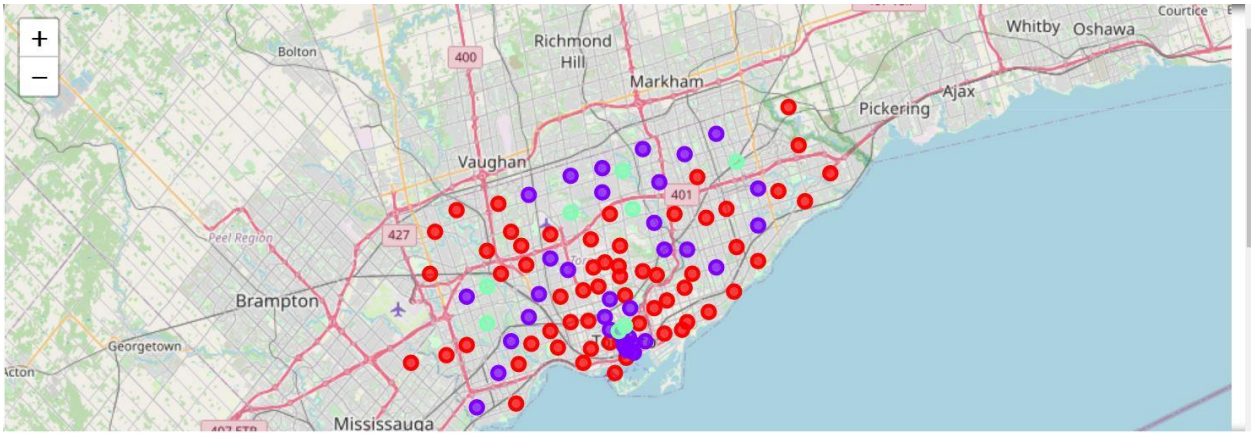
Firstly, we will collect the list of neighborhoods in the city of Toronto using the Wikipedia page(https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). Then we will do web scrapping using the beautiful soup and python request to extract the list of data from the page. So, this is just a name along with the postal code. We need geographic data such as longitude and latitude of respective data so that we used a CSV file that has the geographic coordinates of each postal code and Neighborhoods (http://cocl.us/Geospatial_data). Then we consolidated two data, respectively. After collecting all the data, we can visualize the map of Toronto using the folium package.

Secondly, we used the foursquare API to get the 100 venues in the radius of 2000meters in Toronto city. For that, we need to make API calls to foursquare passing in the geographic coordinates of the neighborhoods in a loop. Foursquare API will return all the data in JSON format. We need to extract the Venue Name, Venue Latitude, Venue Longitude, Venue Category. With this data, we can also check how many unique categories received. Then we will analyze each neighborhood by grouping the rows and taking the mean frequency occurrence of each category and preparing this data for the clustering. As we want to analyze the shopping mall data, we will filter the venue like a shopping mall in neighborhoods.

Lastly, we will perform the clustering of the data by using k-means clustering. It is one of the most popular and simplest unsupervised machine learning techniques suitable for this project. We will now cluster neighborhoods into three clusters based on their frequency of occurrence for the shopping mall. The result will give clarity in which neighborhood has the highest concentration of shopping malls and which have a lower number of a shopping mall. Based on the occurrence of the neighborhood it will help us to answer the question as to which neighborhood are most suitable for opening the shopping malls.

- **Results and discussion:**

The results of the k- means clustering as follows: -




Cluster 0 (Red Color): Neighborhood has no existence of the shopping mall. Cluster 1 (Purple Color): Neighborhood has a moderate number of a shopping mall. Cluster 2 (Mint Green): Neighborhood has a high number of a shopping mall.

So, there are some neighborhoods have high mean frequency occurrence as 0.039. As we observed the map, we can say that more locations are available to build the shopping mall. Cluster 0 has zero number of shopping malls in the neighborhood. This will be a great opportunity to open a new shopping mall as there is no competition from the existing mall. Meanwhile, Cluster 1 and cluster 2 have more amount of shopping malls so they have competition due to the oversupply. Therefore, this project recommends property developers to open new shopping malls in cluster 0 to gain the maximum amount of profit as compared to the other location.

For future reference instead of taking the mean frequency occurrence of a shopping mall, we can take other factors such as the income of residents and population that could help the property developer to decide the location of a new shopping mall. Also, in this project, we used a primary package with a limitation of making API call in foursquare and returned values. So, for future purposes, we can use a paid account to avoid such limitations.

- Conclusion:

In this project, we've seen the method of identifying the business problem, data extraction, data preparation, and performing machine learning techniques like clustering and transform the information into three clusters based on their similarities and providing recommendations to the relevant property developer regarding the best location to open a mall. Also, the project recommended the property developer to open a new mall in cluster zero to get a maximum amount of profit with less amount of competition.