# Project: Finding the best location in Toronto for setting up an office (HUB) for a food delivery company.

## 1. Introduction/ Business Problem

This project finds out list of all the restaurants in Toronto based on Borough, and map them to visualize the location. This project is specially made for food delivery companies who intend to know the location of all the restaurants and usually set their office close to most of the restaurants for their business purpose. So, here my target client is food delivery companies like Uber Eats, Doordash, Grubhub etc. To understand clearly here I have tried to explain a case study of "Lieferando" company, how Data Science might help them selecting the best location for their office.

Lieferando (Former name Takeaway) (https://www.lieferando.de/en/) is a Dutch dot-com company specialized in online food ordering and home delivery. It is an intermediary online portal between the customer and the restaurants, where customers can order food online from restaurant's menu, and have it delivered by the restaurants directly to their home. Foods are delivered by bikes, electric bikes or cars. At present, they are operating 10 European countries and plan to expand in Canada. They have decided to open their first office in Toronto which they will operate both as Hub and Office for the company. Here Hub means where all the delivery vehicles of the company will be parked and used for delivery during order time.

Riders usually stay at the office (Hub) and go to the restaurants when they get orders. Its important that the office must be built close to most of the restaurants so that, riders can move fast towards the restaurants and deliver the food as quick as possible. To select the office location Lieferando is willing to use scientific method rather than following any suggestion by the employees-who don't have enough knowledge of Toronto city. So, they have decided to use Data Science to find out the best location for their office.

In brief the pre-requisite of this project is: a) currently they will set only one office (HUB) in Toronto b) The HUB should be as close to as most of the neighbourhoods in Toronto c) It has to set close to city centre.

## 2. Data Collection

A list of postal codes in Canada where the first letter is M is taken from wikipedia. Postal codes beginning with M are located within the city of Toronto in the province of Ontario. Only the first three characters are listed, corresponding to the Forward Sortation Area.

Canada Post provides a free postal code look-up tool on its website, via its applications for such smartphones as the iPhone and Blackberry, and sells hard-copy directories and CD-ROMs. Many vendors also sell validation tools, which allow customers to properly match addresses and postal codes. Hard-copy directories can also be consulted in all post offices, and some libraries. We shall gather basic information of borough , postcode and neighbourhood from this webpage
(https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M).

Besides we need the geospatial data of this postcode which we can get from Geocoder Python package. The problem with this Package is you have to be persistent sometimes in order to get the geographical coordinates of a given postal code. So, one can make a call to get the latitude and longitude coordinates of a given postal code and the result would be None, and then make the call again and you would get the coordinates. So, here I used a list of geographical coordinates of each postal code from this link (: http://cocl.us/Geospatial_data).

Once we collect all the data then we can go for further processing of these data sets. We have to analyse the data, process it, drop unnecessary columns, and use important column to extract information. However, we need to know in depth information of neighourhoods of each postcode as well as borough in Toronto. For example, how many parks, gym, restaurants are there in each postcode. All of these, we would find at Fourthsquare APi website (https://developer.foursquare.com/). Foursquare is a technology company that built a massive dataset of location data. This site will be used to find out the list of restaurants in different boroughs of Toronto. Then K-Means Clustering algorithm will be used to find out the best location. .

## 3. Methodology

The whole work is divided into three sections.
Step 1: Data importing and pre processing
Step 2: Using the Final DataFrame with FourthSquare API to extract maps
Step 3: Applying K-Means Clustering algorithm to cluster the restaurants

**Step 1: Data Importing and Pre-Processing**
The basic information of borough , postcode and neighbourhood is imported from the Wikipedia page link -(https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). There we get a dataframe like below.

| | Postcode | Borough | Neighbourhood |
|---|---|---|---|
| 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 | Harbourfront |
| 5 | M6A | North York | Lawrence Heights |
| 6 | M6A | North York | Lawrence Manor |
| 7 | M7A | Queen's Park | Not assigned |
| 8 | M8A | Not assigned | Not assigned |
| 9 | M9A | Downtown Toronto | Queen's Park |

- This data frame consists of three columns: PostalCode, Borough, and Neighbourhoods and has 287 rows.
- Only the cells that have an assigned borough is processed. Cells with a borough that is 'Not assigned' is ignored.
- More than one neighbourhood exists in one postal code area. For example, in the table on the Wikipedia page, you notice that M5A is listed twice and has two neighbourhoods: Harbourfront and Regent Park. These two rows is combined into one row with the neighbourhoods separated with a comma as shown in row 11 in the below table.

| | | Neighbourhood |
|---|---|---|
| Postcode | Borough | |
| M1B | Scarborough | Rouge, Malvern |
| M1C | Scarborough | Highland Creek, Rouge Hill, Port Union |
| M1E | Scarborough | Guildwood, Morningside, West Hill |
| M1G | Scarborough | Woburn |
| M1H | Scarborough | Cedarbrae |
| M1J | Scarborough | Scarborough Village |
| M1K | Scarborough | East Birchmount Park, Ionview, Kennedy Park |
| M1L | Scarborough | Clairlea, Golden Mile, Oakridge |
| M1M | Scarborough | Cliffcrest, Cliffside, Scarborough Village West |
| M1N | Scarborough | Birch Cliff, Cliffside West |

- If a cell has a borough but a 'Not assigned' neighbourhood, then the neighbourhood is kept same as the borough. So, for the 9th cell in the table on the Wikipedia page, the value of the Borough and the Neighbourhood columns will be Queen's Park.
- Finally, we get a dataframe of 103 rows which will be used for further analysis
- To get the longitude and latitude value of each borough in Toronto, a file from this link – (http://cocl.us/Geospatial_data) is downloaded.
  Then a new dataframe for Geospatial Data is created as below

| | Postal Code | Latitude | Longitude |
|---|---|---|---|
| 0 | M1B | 43.806686 | -79.194353 |
| 1 | M1C | 43.784535 | -79.160497 |
| 2 | M1E | 43.763573 | -79.188711 |
| 3 | M1G | 43.770992 | -79.216917 |
| 4 | M1H | 43.773136 | -79.239476 |

- Next Latitude and Longitude columns of Geospatial Data is concatenated with our Toronto Postcode data frame. Final dataframe look like below:

| | Postcode | Borough | Neighbourhood | Postal Code | Latitude | Longitude |
|---|---|---|---|---|---|---|
| 0 | M1B | Scarborough | Rouge, Malvern | M1B | 43.806686 | -79.194353 |
| 1 | M1C | Scarborough | Highland Creek, Rouge Hill, Port Union | M1C | 43.784535 | -79.160497 |
| 2 | M1E | Scarborough | Guildwood, Morningside, West Hill | M1E | 43.763573 | -79.188711 |
| 3 | M1G | Scarborough | Woburn | M1G | 43.770992 | -79.216917 |
| 4 | M1H | Scarborough | Cedarbrae | M1H | 43.773136 | -79.239476 |
| 5 | M1J | Scarborough | Scarborough Village | M1J | 43.744734 | -79.239476 |
| 6 | M1K | Scarborough | East Birchmount Park, Ionview, Kennedy Park | M1K | 43.727929 | -79.262029 |
| 7 | M1L | Scarborough | Clairlea, Golden Mile, Oakridge | M1L | 43.711112 | -79.284577 |
| 8 | M1M | Scarborough | Cliffcrest, Cliffside, Scarborough Village West | M1M | 43.716316 | -79.239476 |
| 9 | M1N | Scarborough | Birch Cliff, Cliffside West | M1N | 43.692657 | -79.264848 |

- This dataframe will be used to extract information for Toronto city neighbourhoods.

**Step 2: Using the Final DataFrame with FourthSquare API to Extract Maps**

- Now, lets create a map of Toronto using the value of latitude and longitude. For this, geopy.geocoders and folium libraries are installed at the notebook. First, a map of Toronto with neighbourhoods on top is created. Here neighbourhoods are shown corresponding to each borough of the dataframe.

Out[31]:


- We see that this picture doesn't provide any significant information of maps of restaurants of each borough in Toronto either. So, data frame needs to further process.
- Let's find the number of neighbourhoods associated with each borough.

```
DataFrame['Borough'].value_counts()
```

```
North York          24
Downtown Toronto    19
Scarborough         17
Etobicoke           11
Central Toronto      9
West Toronto         6
East York            5
York                 5
East Toronto         5
Queen's Park         1
Mississauga          1
Name: Borough, dtype: int64
```

- So, borough- North York has the highest number of neighbourhood associated with it. That means highest number of public places like gym, restaurants, parks are associated with it. For any area the number of restaurants is proportional to the number of neighbourhoods in that area. That means we can assume that maximum number of restaurants may be found at the borough North York. However, still we don't know the exact number of restaurants of North York.
- To get the exact list of restaurants in North York I am using Forthsquare API. This site gives the detailed information of every neighbourhoods staying in North York along with longitude and latitude of each neighbourhood.

- By using the Fourthsquare API data, all the neighbourhood of North York is extracted. These neighbourhoods are framed according to their name, categories and latitude and longitude as shown at above dataframe.

| | name | categories | lat | lng |
|---|---|---|---|---|
| 0 | Earl Bales Park | Park | 43.753043 | -79.436228 |
| 1 | Archers Arena | Recreation Center | 43.754101 | -79.467200 |
| 2 | Wolfie's Deli | Deli / Bodega | 43.754875 | -79.442438 |
| 3 | Starbucks | Coffee Shop | 43.758597 | -79.466252 |
| 4 | Nordstrom | Clothing Store | 43.726076 | -79.449335 |
| 5 | True North Climbing | Climbing Gym | 43.745507 | -79.474332 |
| 6 | Kinka Izakaya | Japanese Restaurant | 43.760161 | -79.409827 |
| 7 | LCBO | Liquor Store | 43.732717 | -79.454717 |
| 8 | Bagel Plus | Restaurant | 43.755395 | -79.440686 |
| 9 | Crate & Barrel | Furniture / Home Store | 43.726584 | -79.452661 |

- Then this dataframe is filtered by only ‘restaurant’, since we need only the list of restaurants of North York,

| | name | categories | lat | lng |
|---|---|---|---|---|
| 6 | Kinka Izakaya | Japanese Restaurant | 43.760161 | -79.409827 |
| 8 | Bagel Plus | Restaurant | 43.755395 | -79.440686 |
| 13 | RH Courtyard Café | Restaurant | 43.724874 | -79.455536 |
| 15 | Auberge du Pommier | French Restaurant | 43.746962 | -79.407879 |
| 23 | Konjiki Ramen | Ramen Restaurant | 43.766998 | -79.412222 |

Above dataframe shows the first five rows of our dataframe filtered by restaurant. Now, lets plot these restaurant on map.

```
map_North_York_Restaurants
```

So, these are the locations of restaurants that lies on North York.

**Step 3: Applying K-Means Clustering Algorithm to Cluster the Restaurants**

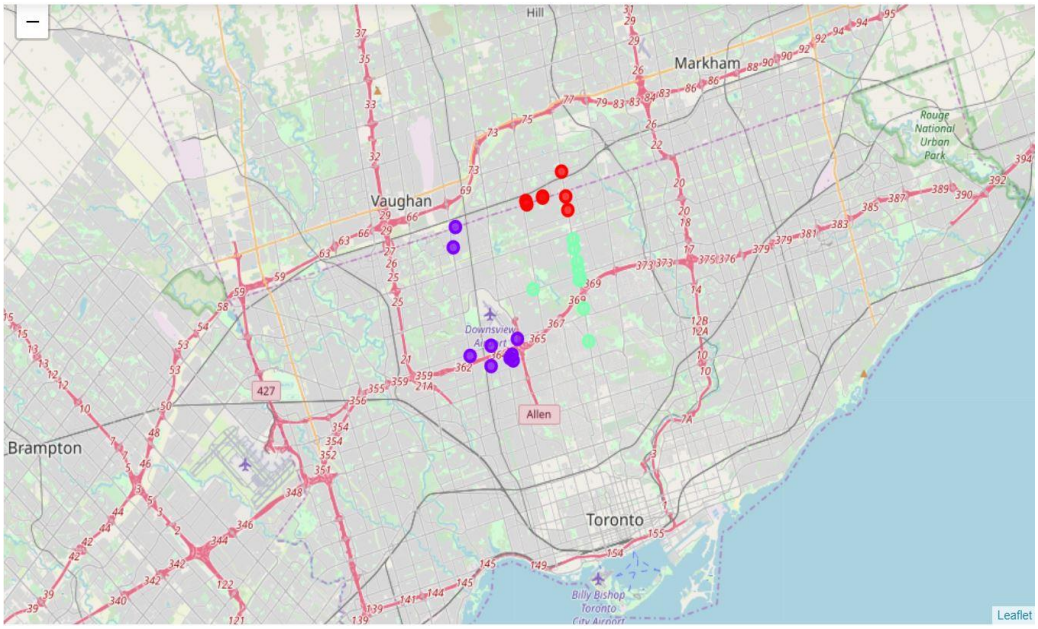
- Now that we have differentiate all the restaurants in North York, we shall apply K-Means Clustering algorithm to cluster the restaurants, and find out the best cluster to set office (HUB).

```
# set number of clusters
kclusters = 3


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

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

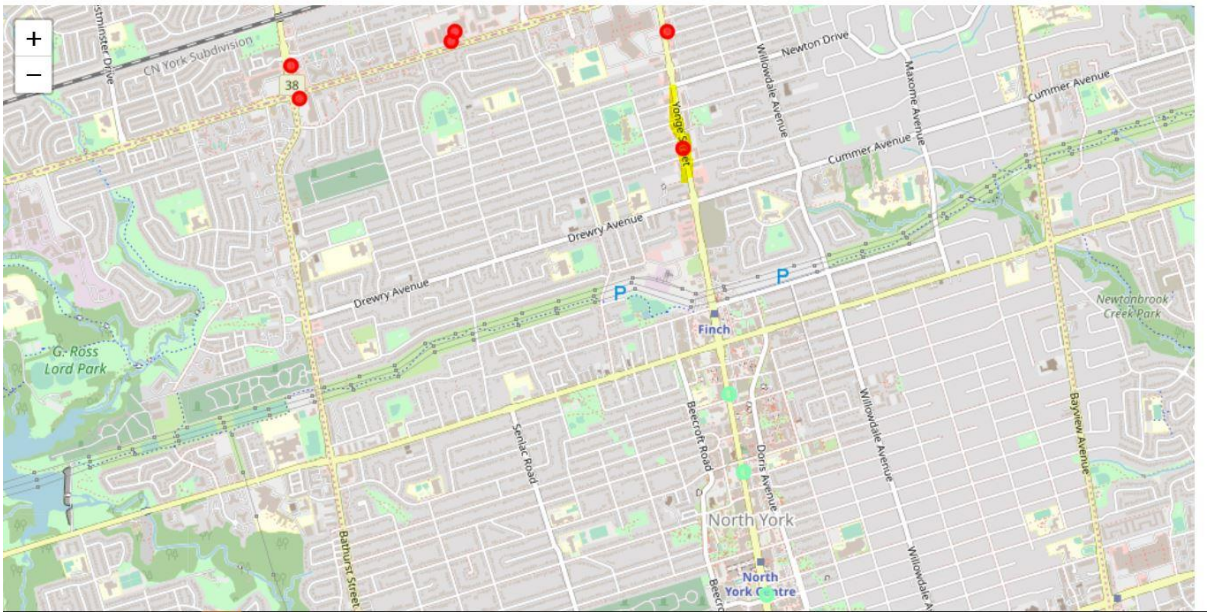
- After executing K-Means Clustering algorithm, the output was visualized on Toronto map using folium library.

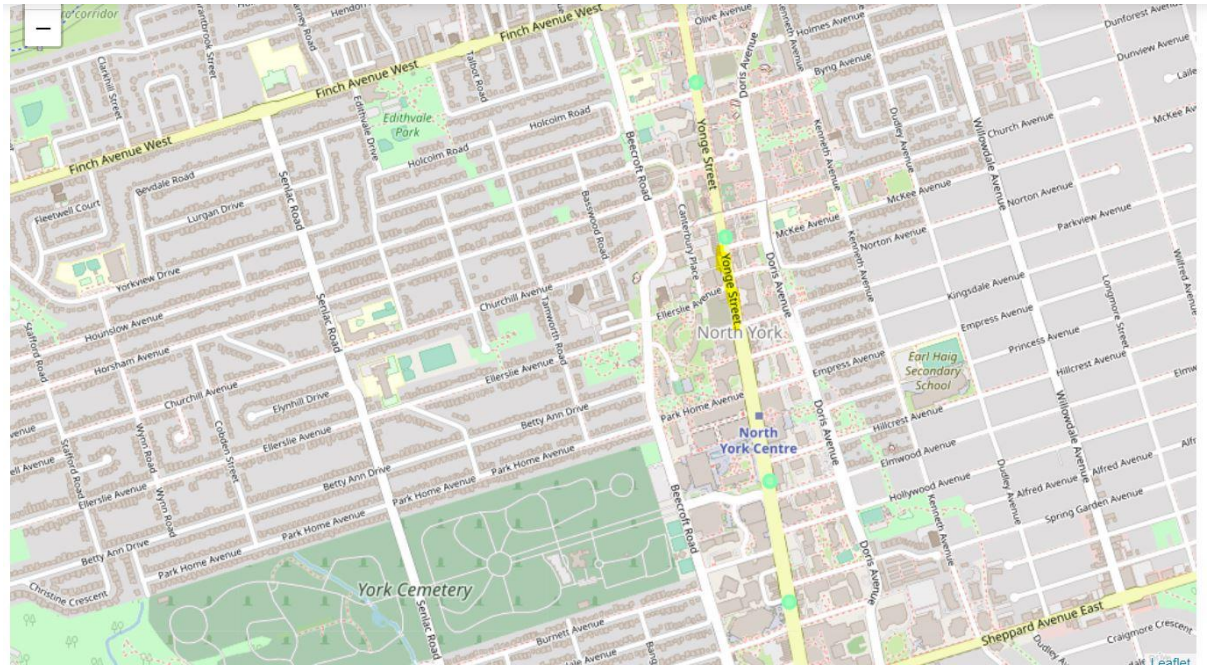



- Above figure shows 3 clusters of restaurants. The figure shows that green cluster is the biggest cluster.

## 4. Results

If we observe previous map carefully, we see that most of the restaurants locate within some common places (mostly along a straight line). After zoom in the map, we see that most of the restaurants lie on **Yonge Street, North York.**




Food delivery companies usually try to set up their HUB close to most of the restaurants.

## 5. Discussion

a) This project strongly depends on data from Wikipedia and the site “FourthSquare API”. As we saw many of the neighbourhoods associated with the postcode were not assigned due to unavailability of information, I had to drop these rows. As a result, some of the data were lost while data pre-processing- which harms the project accuracy.

b) Since I am using only the free package on FourthSquare, I couldn’t call the API more than 2500 times a day. That is why, while calling the venues of North York I have to set limit. If there were no limit then I might have got more list of neighbourhoods which finally would have helped to find more accurate results.

## 6. Conclusion

This project gives an idea how data science could be used to find out an appropriate business location for any company. However, success of any data science project highly depends on the availability and the reliability of data. While doing the project, I spent most of the time doing data cleaning- as this is the most important part of the process and, accuracy of the result highly relies on data cleaning. This project can be developed further by analysing more postcodes and boroughs of Toronto. For that we need a concrete source of list of postcodes associated with detailed neighourhoods. Also, we need to buy Fourthsquare API package to call API enough number of times from the site.