# Coursera IBM Data Science Specialization Capstone Project

# ‘Determining the best location to open a café in Toronto’

## 1. Introduction

- Background

Toronto is one of the biggest cities in the world along with being a major business hub in Canada as well as the world. Thus, the city also attracts talent from all across the globe, right from students to business professionals. For a person who has moved to the city of Toronto quite recently, i.e, who hasn’t lived there for too long to completely know the culture, opening a business based on only gut feeling can be misleading. At the same time, a long time resident of the city might misinterpret his understanding of the emerging trends in the city. It is thus best to make the decision of opening a business based on data.

- Problem

A person wants to open a cafe in the city of Toronto. Seemingly a straightforward decision, Toronto is a big city with a vibrant culture as well as financial hub of Canada and thus has intense competition for opening and running a business. It is also imperative for any business owner that the investment made reaps a decent level of returns. It is thus advantageous that the decision of deciding a place for opening the cafe happens by utilising the right data.

- Interest

The results of the same analysis can also be used by any other business owner. Since the model will segregate the location, we can optimally decide where he/she should set up the new business to maximise possible success of the business.

## 2. Data

The data required for the model will be imported from the Wikipedia page of Toronto giving us the neighbourhood and borough names. The page can be found here: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. The location data along with the nearby places will be queried using Foursquare API. The location data becomes especially crucial in the case of our problem statement, that of opening a café. The possible footfall of the café may be highly influenced by a lot of factors such as proximity to office spaces/residential places/educational hotspots or being in an upscale locality. This data can be readily explored using the API and will help us reach a solution. Moreover, the Foursquare API will provide us data regarding the locations of existing cafes and that will also help in deciding the ideal location for opening a new one.

- Data Acquisition and cleaning:

The data for the neighbourhoods can be readily imported from the Wikipedia page having the postal code, borough names and neighbourhood names in Toronto.

The data has some ‘Not assigned’ values and we will be removing the rows from the dataframe. At the same time, we will require the location co-ordinates of the neighbourhoods. Since the geopy library seldom has errors in returning the co-ordinate values, we will be using a csv file to import the coordinates.

3. Methodology

The code begins with importing the required libraries. The same are listed below along with the purpose of importing them.

- pandas, numpy: required for data wrangling and analysis
- sklearn: required for building the Machine Learning algorithm
- folium: to generate maps
- matplotlib: required for adding visualisations
- json: required for working with json file
- requests: used for getting results from the Foursquare API

Using pandas, we will import the neighbourhood names from the Wikipedia page. Since the page has 3 tables, indexing and importing the right dataframe does the trick. By using keyword argument, the rows having ‘not assigned’ values are removed.

Next, the csv file containing the location data for the neighbourhoods is imported and the data is merged with the neighbourhood names to create a new dataframe. This will be the dataframe on which we will work ahead.

The Foursquare API credentials are defined and imported and the relevant parameters such as radius, limit which all form part of the arguments to be passed to the API are defined. Since our problem statement requires us to determine a place to open a new café, we will use the ‘search’ rather than ‘explore’ option of the Foursquare API. Using the search option we can specifically target areas having cafes to map them. The explore option returns us overall places in the defined area and isn’t of direct use.

The venue IDs to be passed to the search url are found on this page: https://developer.foursquare.com/docs/build-with-foursquare/categories/

We will use category ID for café as ‘4bf58dd8d48988d16d941735’, found in the above link.

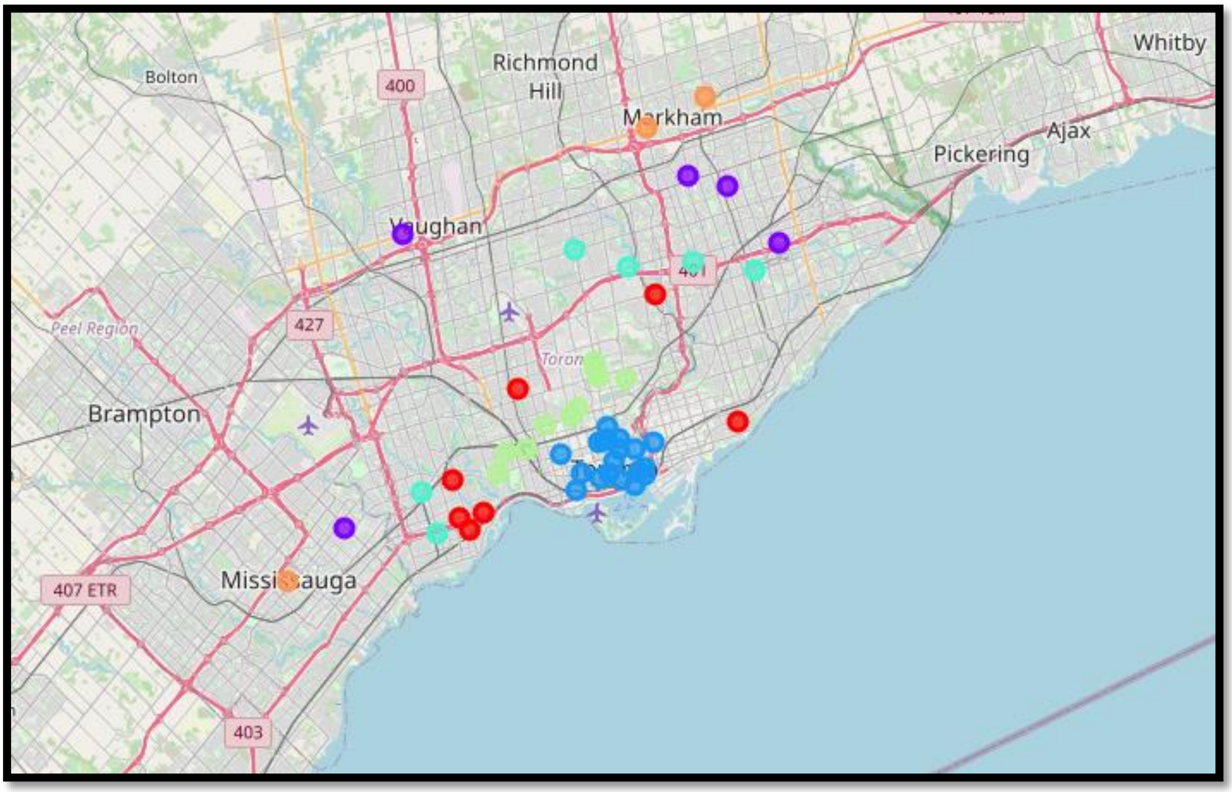
The results are defined in a json file and are converted to a dataframe.

Using this dataframe, we will build our K means cluster.

## 4. Model

We built the model using the K means clustering algorithm and will be using the distance parameter to cluster the data points. In the model, the init parameter is set to 12,i.e, the process will be repeated 12 times to find an optimum solution. Also, we will be using the k++ parameter to smartly select the initial cluster centroids.

Based on the input parameters, the model returns us clusters which we have visualised using folium.



## 5. Recommendations:

It can be seen upon visualisation that many existing cafes are nearer to the coast of Lake Ontario and in proximity of the University of Toronto. On closer inspection, it can also be seen that many existing cafes are alongside or in very close proximity of the major roadways. On the other hand, since the density of cafes is quite high on one side of the university of Toronto while one side is conspicuously less dense with only a couple of cafes.

Since proximity to national highway, sea and university seem to be a crowd puller, Our friend who wants to open his new café business can look into a location which is nearby to a national highway or a learning place which doesn’t go too much towards the sea side (thereby opening a business in the already crowded area) but still isn’t too far away.

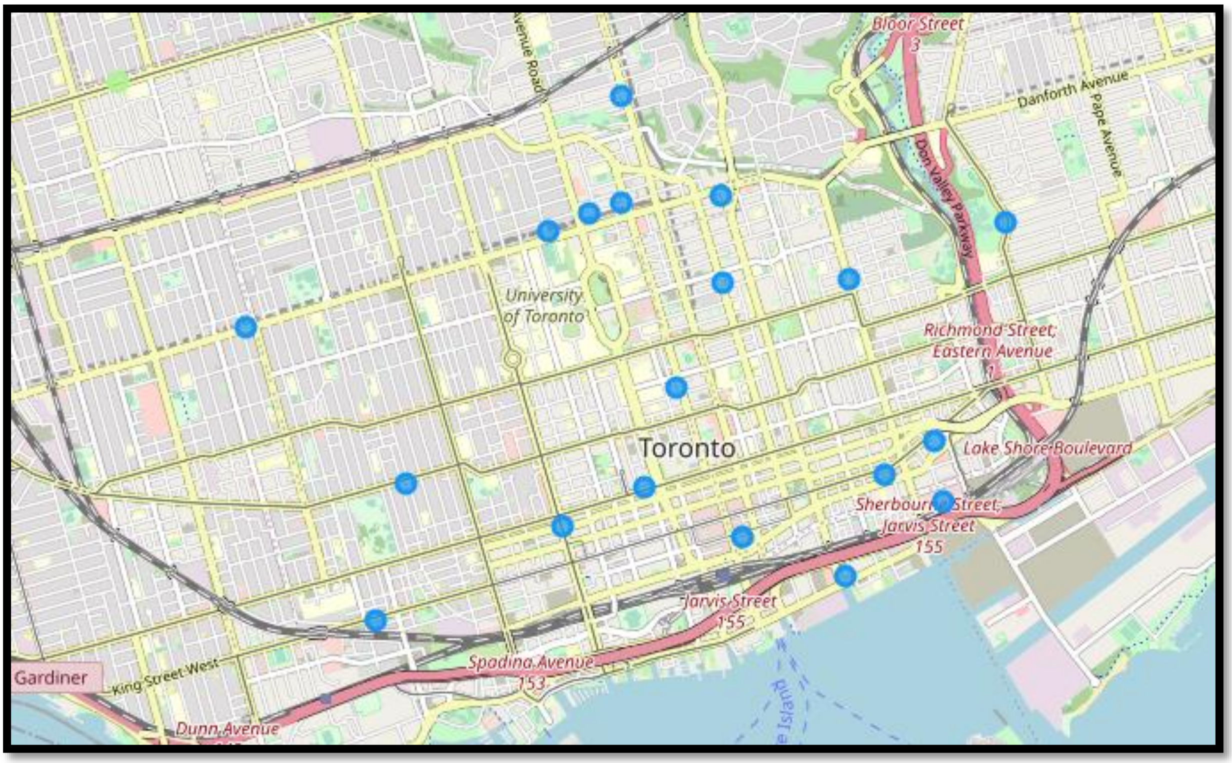


Fig: Decent number of cafes around the University of Toronto and close to the coast

## 6. Conclusion

A lot of other factors such as time of day, the quality offered by the café as well as proximity of influential/upscale neighbourhoods can be considered for the study. But, the present analysis does give a fair idea about the possible locations for opening the business. Utilising the location data as well as the 'trending' feature of the Foursquare API (which returns real time results and thus might change at every API request made), further insight can be taken to even figure out a specific type of café(having specific menu/seating style etc) which is presently trending.