## Introduction/ Business Problem:

In the lab, we have practiced with data of New York and in the assignment of week 3, we have worked with the data given on Wikipedia of postal codes of Canada. In this assignment, I would like to address the similarity between two cities which are Toronto and New York. By using clustering, I will address the similarity between two cities in this report.

## Data:

For this problem, I have used the data of Postal codes in Canada which is available on the link given below:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The city of Toronto is compared with the New York city. For that data, Foursquare API is used.

The data contains details about the places which are around the neighborhood. Such places include café, restaurants, parks, hotels, and so on.

## Methodology:

The data of Postal codes in Canada is preprocessed first as it was in the form of table in Wikipedia page. So, to retrieve that table in dataframe format, the library named BeautifulSoup is used. Also, the values that are not assigned are being removed from the dataset. Here, we are using only the data of city of Toronto, so it was retrieved from the table. Another city which is compared here is New York. To get that data Foursquare API is used.

Folium library is used here for graphical representation here. Foursquare API is used to get nearby venues around the neighborhood. To find the similarity between two cities I have used clustering technique. To implement it, K-means algorithm is used. Here, I have 7 different number of clusters.

## Results:

Form the results, it is observed that most common venues of Toronto are coffee shop, café and restaurants. Whereas in New York, the most common places are restaurants, Bar and parks. Thus, if investors like to invest in Toronto, opening a coffee shop is the best option. Similarly, in New York area, opening an Italian Restaurant is the best option.

## Discussion:

This data contains the selection of most common venues based on the most common venues in neighborhood. It does not have data about the frequency of venue visited, cost or the environment. So, it is difficult for investors to decide the place to invest from the given data. Other factors should also be considered to get an optimal place.

## Conclusion:

Both the cities have almost similar venues in neighborhood. Although, the most common venue is varied in the district of the cities. Also, the issue here is that the investors can not decide place based on most common venues of neighborhoods. The better option to choose it based on feedback from people about which place people like most and other similar factors.