IBM Coursera Capstone Project

**Exploring Neighborhoods in NYC**

Identifying an ideal neighborhood for Opening an Indian Restaurant

horizontal line

# Introduction

New York is one of the famous places in the world. It is diverse in many ways. It is multicultural as well as the financial hub of the USA. Today Tourism is one of the pillars of the economy and the people most often visit NYC which is rich in heritage and developed enough from a foreign perspective, like a friendly environment. Every city is unique in it’s own way and gives something new. And now the information is so common regarding the location of every place around the world on your fingertips which make it easier to explore.

This data science capstone project deals with the process of leveraging location data acquired from data providers such as Foursquare to explore the neighborhoods within a targeted city and create clustering models. Using K-means clusters, similar locations with minimum distance shall be grouped into clusters. It is the simplest form of unsupervised machine learning algorithm and it helps in grouping similar data points. In this project we will go through a step by step process to make a decision whether it is a good idea to open an Indian restaurant. We analyze the neighborhoods in NYC to identify the most profitable area since the success of the restaurant depends on the people and ambience. Since we already know that NYC shelters a greater number of Indians it is a good idea to start the restaurant here, but we just need to make sure whether it is a profitable idea or not. If so, where we can place it, so it yields more profit to the owner.

## Data Exploration

The data for this project will be extracted, processed and analysed by integrating the borough information for New York City extracted from the web and venue related information acquired through Foursquare API.The data extraction from web shall be done using the web scraping libraries for python such as Beautiful Soup. After extracting the html page the information shall be converted into a data frame using the pandas python library. Using the pandas library, the data will be cleaned and processed to prepare a final data frame for analysis. In order to render my data onto a map, I will be using the Folium library. Also, to create clusters of similar regions of interest, I will be using k-means clustering technique.

### Data Sources

1. I am using the "<https://cocl.us/new_york_dataset>" page to get all the information about the neighborhoods present in New York City. This page has the borough & the name of all the neighborhoods, Latitude and Longitude present in New York city.
2. Then “<https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City>” page to get all the population details of the neighborhoods.
3. To get location and other information about various venues in New York City I’m using Foursquare’s explore API. Using the Foursquare’s explore API (which gives venues recommendations), I’m fetching details about the venues present in New York City and collecting their names, categories and locations (latitude and longitude). From Foursquare API (<https://developer.foursquare.com/docs>), I retrieved the following for each venue:

Name: The name of the venue.

Id: Venue Id

Category: The category type as defined by the API.

Latitude: The latitude value of the venue.

Longitude: The longitude value of the venue.