# Coursera Capstone Project

## IBM Applied Data Science Project

## Predicting a New location for "Cafe Coffee day" shop in Bangalore, India

### Introduction

Café Coffee Day (CCD) is an Indian café chain. It is a subsidiary of Coffee Day Enterprises Limited. Coffee Day serves 1.8 billion cups of coffee annually. Café Coffee Day(CCD) Global Limited Company is a Chikkamagaluru-based business which grows coffee in its own estates of 20,000 acres. It is the largest producer of arabica beans in Asia, exporting to various countries including the U.S., Europe, and Japan  
CCD is assumed to be a best place to hang out with friends where you can enjoy conversations over signature coffee brews, cold beverages, bites, desserts. There are around 130 outlets in the Bangalore Location of existence of these CCD outlets enhance the profit or loss of the company.

#### **Business Problem**

The objective of this Capstone project is to analyze and get the best location for the CCD to open a new outlet in Bangalore , India .  
Using Data science methodology and machine learning techniques like Clustering , segmentation , this project aims to provide solutions to answer the business problem  
i.e. in the city of Bangalore , if a CCD owner wishes to open a New Outlet , which is the best suitable place ?

#### **Target Audience**

This project is mainly useful for the CCD owners as well as competitors of CCD as to where the new outlet has to be started to have a good business.  
This project is timely as the city is currently suffering from oversupply of CCD shops

### Data

To solve the mentioned Problem, we need the following

* List the Neighborhoods of Bangalore. This defines the scope of this project which is confined to the city Bangalore, a city in Karnataka state.
* Latitude and Longitude of this Neighborhood. This is required to plot the map and also get the venue data
* Venue data i.e. particular data related to Cafe Coffee shops. We will use this data to perform clustering on the neighborhoods

#### **Sources of Data and methods to extract them**

This Wikipedia page <https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Bangalore> contains a list of neighborhoods in Bangalore with a total of neighborhoods .We will use web scraping techniques to extract the data from the Wikipedia page , with the help of Python request and beautifulspoup packages .Then we will get the geographical coordinates of neighborhood using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

We will use Foursquare API to get the venue data for these neighborhoods. Foursquare has one of the largest Databases of 105+ million places. Foursquare API will provide many categories of the Venue data but we are mainly interest in the CCD Coffee Shop in order to help us to solve the business problem This project will make use of many Data Science skills from web scraping, working with APR (Foursquare), data cleaning , data Wrangling to machine learning (K means Clustering) and data Visualization (Folium)  
In the next section, Methodology section we will elaborate on the data analysis , machine learning technique.

### Methodology

Firstly, we need to get the list of neighborhoods in the city of Bangalore, India. Fortunately, the list is available in the Wikipedia Page ()

We will do Web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data.

However, this is just a list of names. We need to get the geographical Coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the Geocoder Package that will allow us to convert address into geographical coordinates in the form of latitude and longitudes. After gathering the data, we will populate the data in to pandas Dataframe and then visualize the neighborhoods in a map using folium package.

This allows us to perform sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Bangalore.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the “Coffee Shop” data, we will filter the “Coffee Shop” as venue category for the neighborhoods.

Lastly we will perform clustering on the data by using k –means clustering. K-means clustering algorithm identifies k number of centroids and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithm and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 4 clusters based on their frequency of the occurrence for “Coffee Shop”. The results will allow us to identify which neighborhood have higher concentration of Coffee Shops in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new Coffee shops.

### Results:

The results from the –means clustering show that we can categorize the neighborhoods into 4 clusters on the frequency of occurrence for Coffee shops

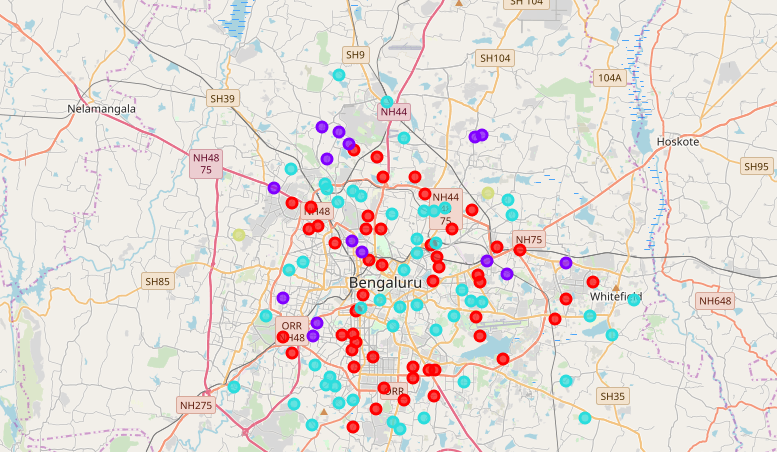
Cluster 0: Neighborhoods with high number of Coffee Shops

Cluster 1: Neighborhoods with moderate number of Coffee Shops

Cluster 2: Neighborhoods with high number of Coffee Shops

Cluster 3: Neighborhoods with low number of Coffee Shops.

The results of clustering are visualized in the map below



### Discussion

As observations noted from the map in the Results section, most of the Coffee shops are concentrated in the central are of Bangalore, with the highest number in cluster 0 & 2 and moderate number in Cluster 1. On other hand, cluster 3 has very low number of Coffee shops. This represents a great opportunity and high potential areas to open a new Coffee Shops as there is little competition. This project recommends property developers to capitalize on these findings to open new shops in neighborhoods in cluster 3. Lastly property developers are advised to avoid neighborhoods in cluster 0 and 2 which are highly concentration of shops and suffering intense competition

### Conclusions

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 4 clusters based on their similarities and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new shop. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhood in cluster 3 are the most preferred locations to open a new Coffee shop. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decision to open a new Coffee shop

### References

<https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Bangalore>