IBM Applied Data Science Capstone

Opening a Coffee shop in suburbs of Chennai, India

By: Ashok Venkatachalam

September 2020

Introduction

Worldwide, experts estimate that people drink about 2.5 billion cups of coffee a day. Sales in the ready-to-drink market—which includes coffee shops—are forecast to grow by 67 percent between now and 2022. Additionally, coffee and other ready-to-drink shops show incredible resilience in volatile markets, helping to eliminate some of the uncertainty associated with small business ownership.

Business Problem

The objective of this capstone project is to analyze and select the best locations in the suburbs of Chennai to open a coffee shop. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the suburbs of Chennai, if investor is looking to open a new coffee shop, where would you recommend that they open it?

This project is timely as the city is currently short of coffee shop and the lifestyle of the people is changing and people are preferring to go to coffee shop for meeting and get together with friends.

Target Audience of this project

This project is particularly useful to investors looking to open or invest in new coffee shop in the suburbs of Chennai.

Data

To solve the problem, we will need the following data:

- List of neighborhoods in Chennai. This defines the scope of this project which is confined to the city of Chennai, the metropolitan city in the south India.

- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to 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:Suburbs_of_Chennai) contains a list of neighborhoods in Chennai, with a total of 65 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers.
Foursquare API will provide many categories of the venue data, we are particularly interested in the Coffee Shop category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.