

Coursera Capstone Project

IBM Applied Data Science Capstone

Opening a New 'Café' in Mumbai



Introduction

Cafes are one of the most important aspects of humans. Be it a date, a meeting, if you want to enjoy with friends, if you want to eat some snacks, these are the most favourite places of people. As a result, there are many cafes in the city of Mumbai and many more are required to build considering increasing popularity. Even though these businesses require less capital and space, they can fetch large profits. But what matters is the location. The location plays a vital role in deciding whether this business will fail or not.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Mumbai to open a new cafe. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question.

Target Audience of this project

This project is particularly useful to small businesses and individuals looking to open or invest in new cafes in Mumbai. Since, food industry is one of the industry which is everlasting and always in demand, many people are interested in either investing in such business or actually put their money and do this business.

Data

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

- List of neighbourhoods in Mumbai. This defines the scope of this project which is confined to the Mumbai, the financial capital city of the country of India.
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to shopping malls. We will use this data to perform clustering on the neighbourhoods

Sources of data and methods to extract them

This Wikipedia page

(https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai) contains a list of neighbourhoods in Mumbai, with a total of 42 neighbourhoods. 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 neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. 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.