COURSERA CAPSTONE

IBM Data Science Module

Research to open a Mall in the city of India

Author: Swapnil, Feb 2021

Introduction

Malls are like a one-stop destination for all types of shoppers. For retailers, the central location and the large crowd at the shopping malls provides a great distribution channel to market their products and services. Developers are also taking advantage of this trend to build more malls to cater to the demand. As a result, there are many malls in India and many more are being built. Opening malls allows property developers to earn consistent rental income. Of course, the location of the mall is one of the most important decisions that will determine whether the mall will be a success or a failure.

Business Problem

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

Target Audience of this project

This project is particularly useful for developers and investors looking to open or invest in new malls in the Indian city.

Data

- List of neighbourhoods in Indian city. This defines the scope of this project which is confined to the city 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 malls. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them

This Wikipedia page (https://commons.wikimedia.org/wiki/Category:Suburbs_of_India) contains a list of neighbourhoods in India, with a total of 58 neighbourhoods. I will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautiful-soup 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.

Foursquare API will provide many categories of the venue data, we are particularly interested in the Mall 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.