CAPSTONE PROJECT

IBM Data Science
Professional Certificate

25/05/2020

OPENING A NEW SUPERMAKET IN COLOMBO, SRILANKA

Manoj Wickramasinghe

INTRODUCTION

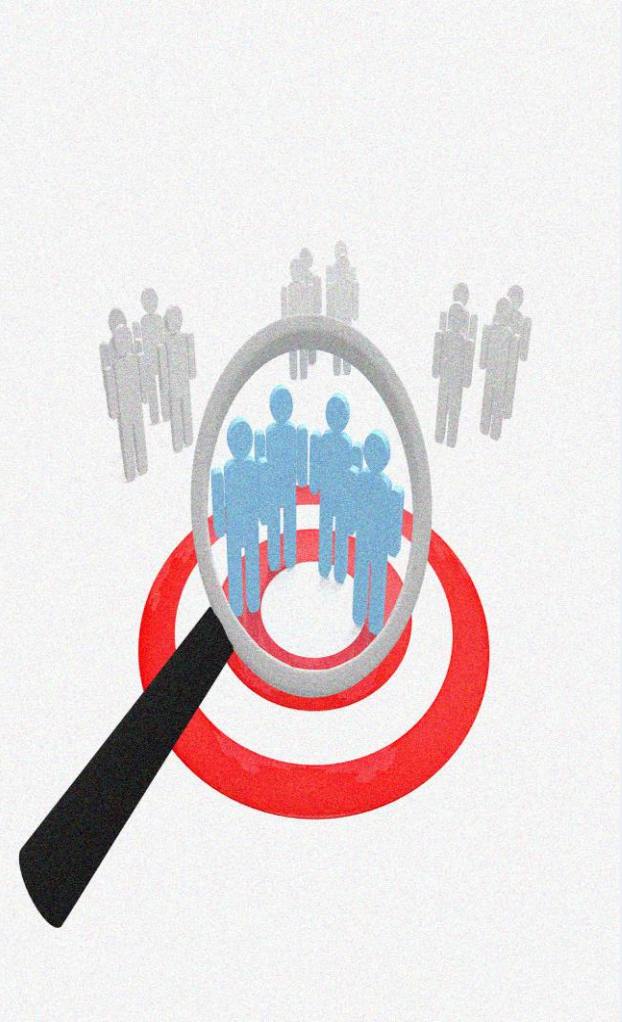
For many shoppers, visiting supermarkets is a great way to enjoy themselves while browsing for their groceries. They can do grocery shopping, which is easy and quick and due to adequate parking space, shopping becomes an easy and pleasing activity rather than boredom. Supermarkets are like a one-stop destination for all types of shoppers which will guarantee freedom of selection. For retailers, the central location and the large crowd at the supermarket provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more supermarkets to cater to the demand. As a result, there are many supermarkets in the city of Colombo and many more are being built. Opening supermarkets allows property developers to earn consistent rental income and as with many other business decisions, opening a new supermarket requires serious consideration and is lot more complicated than it seems. Particularly, the location of the supermarket is one of the most important decisions that will determine whether the market will be a success or a failure.

BUSINESS PROBLEM

The objective of this capstone project is to analyze and select the best locations in the city of Colombo, Sri Lanka to open a new supermarket. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Colombo, Sri Lanka, if a property developer is looking to open a new supermarket, where should they open it?

TARGET AUDIENCE OF THIS PROJECT

This project is particularly useful to property developers and investors looking to open or invest in new supermarkets in the capital of Sri Lanka. This project is timely as the city is currently suffering from oversupply of supermarkets. Data from the National Property Information Centre (NAPIC) released last year showed that the an additional 15 per cent will be added to existing market space, and the agency predicted that total occupancy may dip below 86 per cent. The local newspaper The Daily Mirror also reported in March last year that the true occupancy rates in markets and malls may be as below as 40 per cent in some areas, quoting a Financial Time (FT) article cataloguing the country's continued obsession with building more supermarket space despite chronic oversupply.



SOURCE OF DATA AND METHOD OF EXTRACTION

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_Colombo) contains a list of neighborhoods in Colombo, with a total of 67 neighborhoods. We will use web scrapping techniques to extract the data from Wikipedia page, with the help of pandas and beautifulsoup packages. Then we will get the geographical coordinates of the neighborhood 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 databases 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 Supermarket 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 skill, 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 we used.

DATA

To develop a model to cluster neighborhoods, we need the following data:

➢ List of neighborhoods in Colombo. This defines the scope of this project which is confined to the city of Colombo.

➢ *Latitude and longitude coordinates of those neighborhoods. This require in order to plot the map and to get venue data.*

➢ *Venue data, particularly data related to supermarkets. We will use this data to perform clustering on the neighborhoods.*