CAPSTONE PROJECT

IBM Data Science
Professional Certificate

25/05/2020

OPENING A NEW SUPERMAKET IN COLOMBO, SRILANKA

Manoj Wickramasinghe

INTRODUCTION

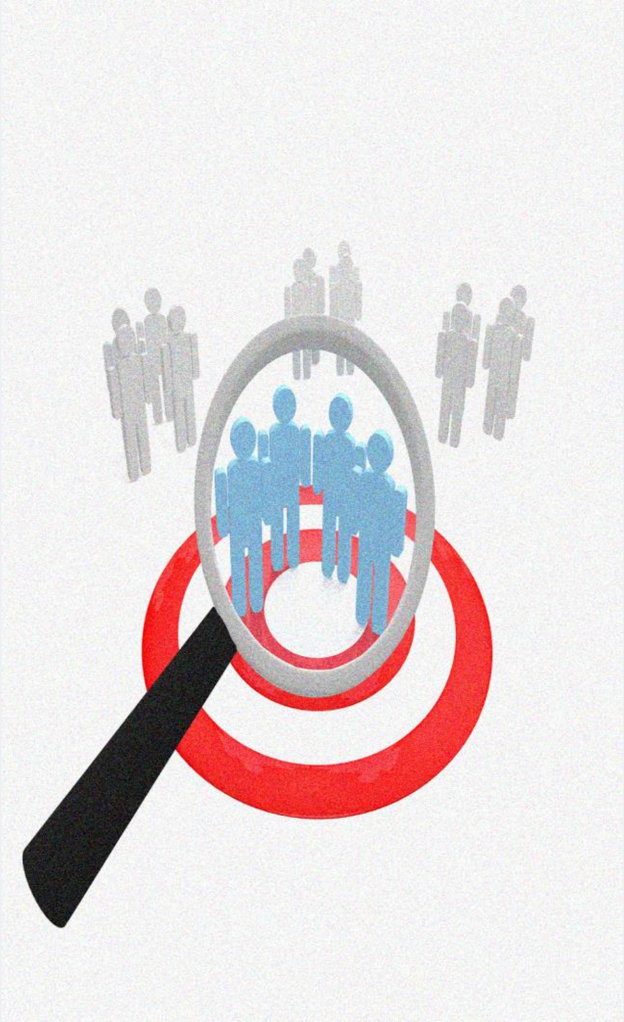
For many shoppers, visiting supermarkets is a great way to enjoy themselves while browsing for their groceries. They can do grocery shopping, which is easy and quick and due to adequate parking space, shopping becomes an easy and pleasing activity rather than boredom. Supermarkets are like a one-stop 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.*

METHODOLOGY

Firstly, we need to get the list of neighborhoods in the city of Colombo. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_Colombo). 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 wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas data frame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Colombo. 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 "Supermarket" data, we will filter the "Supermarket" 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 algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for "Supermarket". The results will allow us to identify which neighborhood have higher concentration of supermarkets while which neighborhoods have fewer number of supermarkets. Based on the occurrence of supermarkets in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new supermarkets.

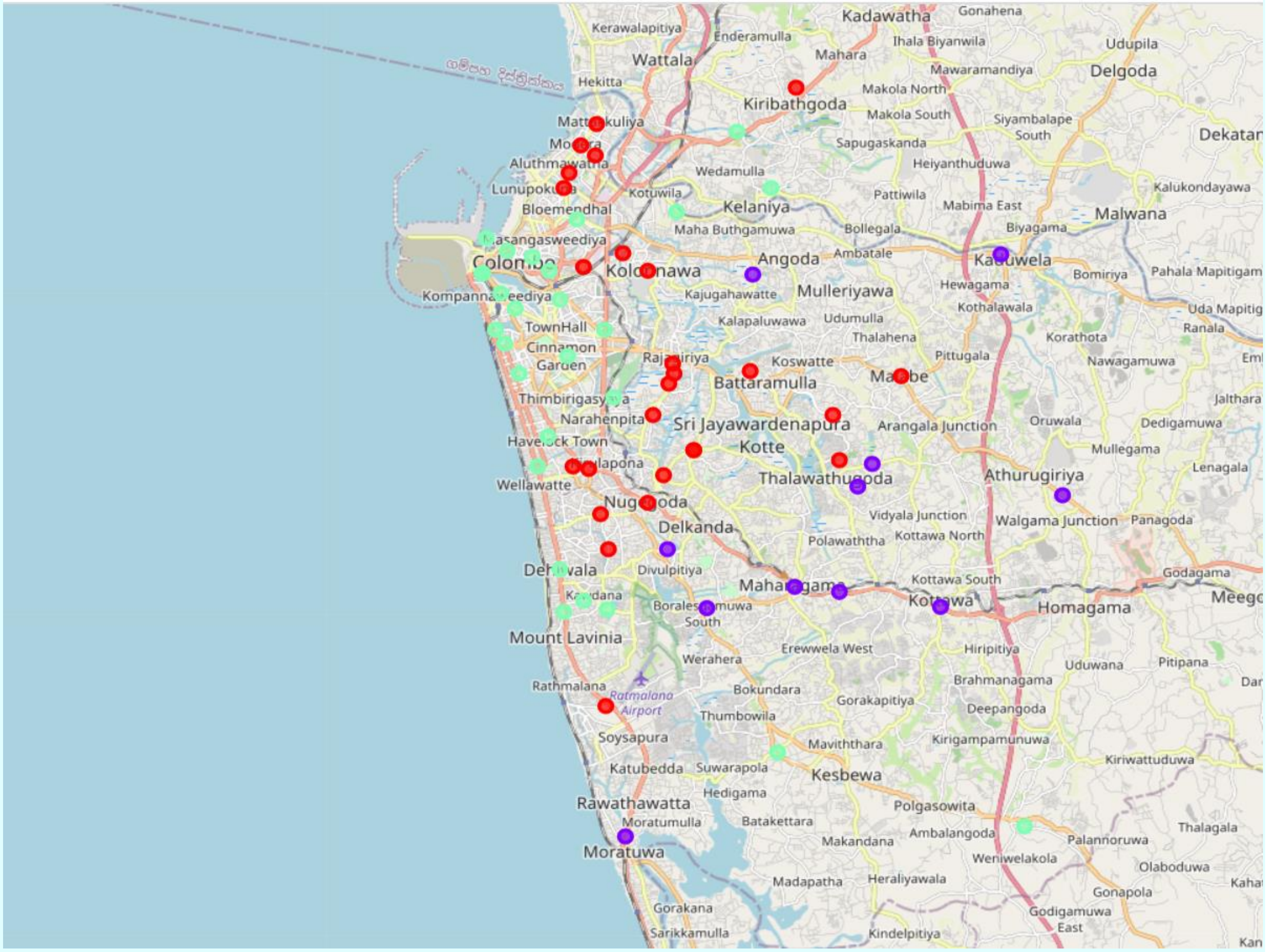
RESULTS

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for "Supermarket":

➢ Cluster 0: Neighborhoods with moderate number of supermarkets.

➢ Cluster 1: Neighborhoods with high concentration of supermarkets.

➢ Cluster 2: Neighborhoods with low number to no existence of supermarkets.



Cluster 0

Cluster 1

Cluster 2

DISCUSSIONS

As observations noted from the map in the Results section, most of the supermarkets are concentrated in the Central area of Colombo city, with the highest number in cluster 1 and moderate number in cluster 0. On the other hand, cluster 2 has very low number to no shopping mall in the neighborhoods. This represents a great opportunity and high potential areas to open new supermarkets as there are very little to no competition from existing markets. Meanwhile, supermarkets in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of supermarkets. From another perspective, the results also show that the oversupply of supermarkets mostly happened in the Central area of the city, with the coastal area having very few supermarkets. Therefore, this project recommends property developers to capitalize on these findings to open new supermarkets in neighborhoods in cluster 2 which has little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new supermarkets in neighborhoods in cluster 0 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 1 which already have high concentration of supermarkets and suffering from intense competition.

LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

In this project, we only consider one factor i.e. frequency of occurrence of supermarkets, there are other factors such as population and income of residents that could influence the location decision of a new supermarkets. However, to the best knowledge of this researcher such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new supermarket. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

CONCLUSION

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 3 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 supermarket. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 2 are the most preferred locations to open a new supermarket. 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 decisions to open a new supermarket.

REFERENCES

Category: Suburbs in Colombo.
Wikipedia. Retrieved from
https://en.wikipedia.org/wiki/Category:Suburbs_of_Colombo

Foursquare Developers Documentation.
Foursquare. Retrieved from
https://developer.foursquare.com/docs/

THANK YOU