IBM Data Science Capstone Project

Picking the right location for a new restaurant in Colombo, Sri Lanka

By Tobi Taiwo

January, 2020

Introduction

This project is about using data science tool set on a real-life problem and demonstrating the creation of value by applying the learned skills. It particularly addresses the business decision of “Where is the best location to set up a new restaurant in the city of Colombo, Sri Lanka that will yield high profitability”.

Business Problem

The main objective behind this project is to select the ideal locations in the city of Colombo in Sri Lanka to to open up a new restaurant. The project aims at using Data Science methodology and machine learning techniques to find out the best possible neighborhood for starting up a new restaurant that will lead to a high amount of sales and increased profitability compared to other neighborhoods in Colombo.

Target Audience

This project is beneficial to individuals or investors looking at starting up or investing money into building new restaurants in the city of Colombo in Sri Lanka. The project serves as a form of guideline when it comes to choosing the most viable location to open up a new restaurant that will potentially lead to high amounts of sales and increased profitability for the restaurant stakeholders.

Data

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

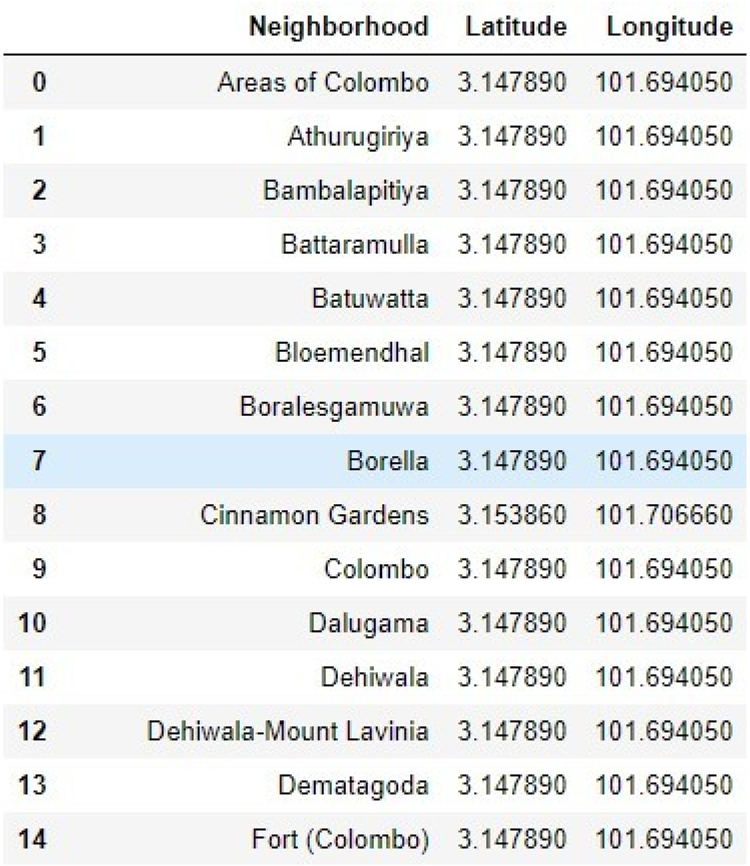
- List of neighbourhoods in Colombo, Sri Lanka. This defines the scope of this project which is confined to the city of Colombo, the capital city of the country of Sri Lanka.
- Geo-coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to restaurants. This will be obtained from Foursquare through an API.

Sources of data and methods to extract them

This Wikipedia page “ https://en.wikipedia.org/wiki/Category:Suburbs_of_Colombo “ contains a total of 67 neighbourhoods in Colombo. Web scraping techniques will be used to extract the data, as well as Python requests and beautifulsoup packages. Python Geocoder package will be used to get the geographical coordinates of the neighbourhoods. Then, we will use Foursquare API to get the venue data for those neighbourhoods.

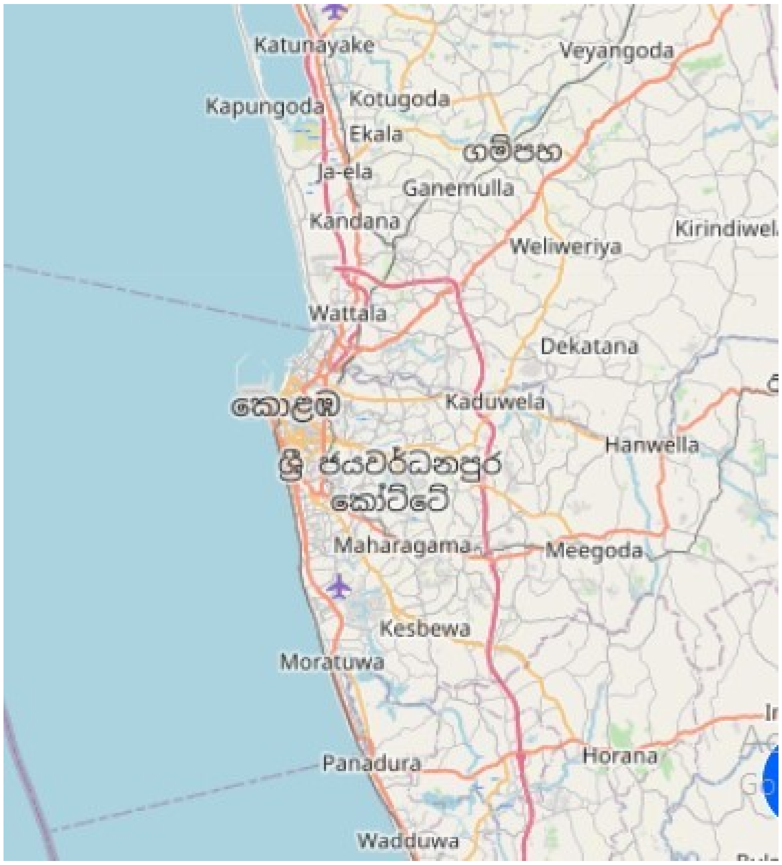
Methodology

The data consisting of the list of neighborhoods in the city of Colombo in Sri Lanka was extracted from this Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_Colombo). We use the Geocoder package to convert address into geographical coordinates in form of latitude and longitude coordinates. After performing this task, we get the following table we use in a pandas dataframe format.



	Neighborhood	Latitude	Longitude
0	Areas of Colombo	3.147890	101.694050
1	Athurugiriya	3.147890	101.694050
2	Bambalapitiya	3.147890	101.694050
3	Battaramulla	3.147890	101.694050
4	Batuwatta	3.147890	101.694050
5	Bloemendhal	3.147890	101.694050
6	Boralesgamuwa	3.147890	101.694050
7	Borella	3.147890	101.694050
8	Cinnamon Gardens	3.153860	101.706660
9	Colombo	3.147890	101.694050
10	Dalugama	3.147890	101.694050
11	Dehiwala	3.147890	101.694050
12	Dehiwala-Mount Lavinia	3.147890	101.694050
13	Dematagoda	3.147890	101.694050
14	Fort (Colombo)	3.147890	101.694050

Then use the Folium package to visualize the neighborhoods in form of a map.



In the next step of the analysis, the districts were explored in greater detail. It means venues were collected for each district via Foursquare API. The data from Foursquare is received in json format. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing this, we are also preparing the data for use in clustering. Since we are analyzing the “Restaurant” data, we will filter the “Restaurant” as venue category for the neighbourhoods.

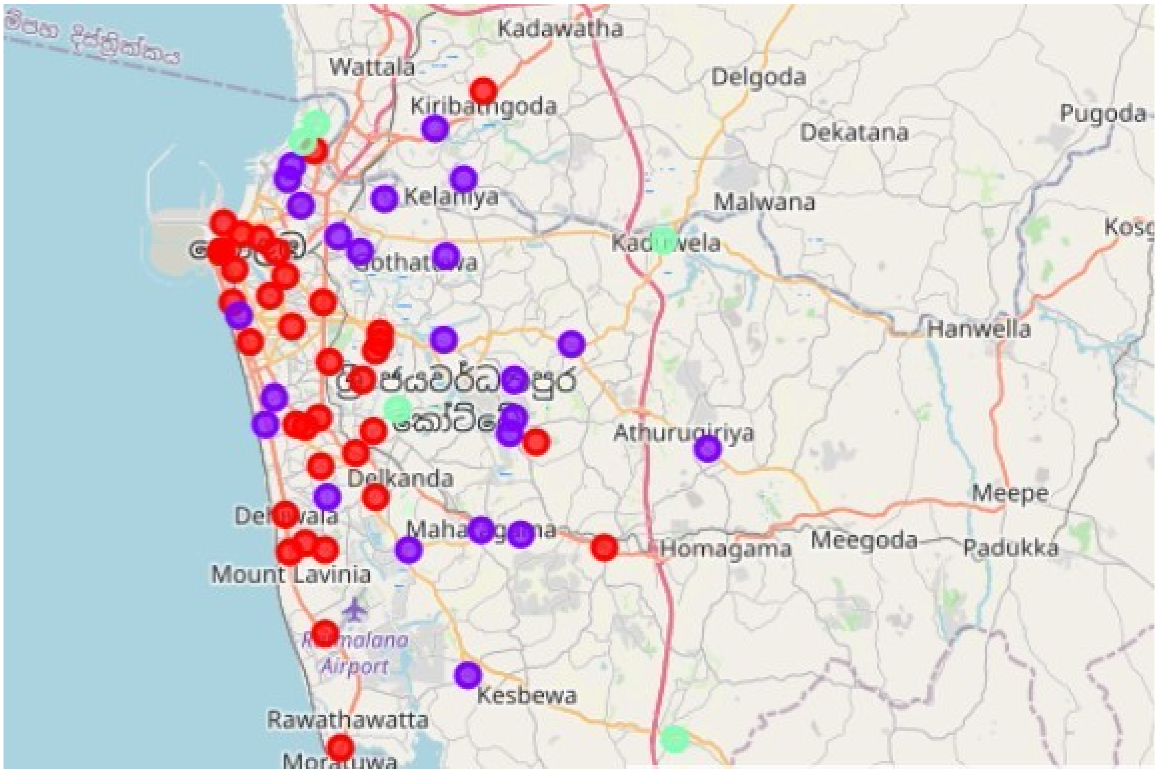
Lastly, we perform clustering on the data by using K-means clustering. For K-means clustering, we need to decide on the number of clusters that we want to use. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. The results will allow us to identify which neighbourhoods have higher concentration of restaurants and the ones with fewer number of restaurants. Based on the occurrence of restaurants in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to set up new restaurants.

Results

From the results from the k-means clustering, we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Restaurant”:

- Cluster 0: These neighbourhoods have a moderate number of restaurants.
- Cluster 1: These neighbourhoods have a high number of restaurants.
- Cluster 2: These neighbourhoods have a low number of restaurants.

We can visualize the clusters on the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.



Discussion

Based on our observation on the clusters, we can advise the restaurant owner to consider the districts from Cluster 2 as an ideal location for setting up a new restaurant. These are the districts where there is little competition thereby leading to high sales potential. Cluster 1 has a high concentration of restaurants. This won’t be an ideal location to open up a new restaurant as the competition will be fierce. Whereas, Cluster 0 presents a moderate number of restaurants, opening a restaurant here might be feasible as the competition wouldn't be so fierce but as a newly opening restaurant, it might not be the best idea as customers might not quickly start coming around. On the other hand, Cluster 2 has very little restaurants available, these are good locations to start up a restaurant as the market is very scarce. Therefore, there is a high potential of making a lot of sales in these areas. Based on this project, it's recommended that Cluster 2 be considered as the most viable location to start up a new restaurant.

Conclusion

This paper discussed the process of coming up with an answer to the business question that was raised in the introduction section. It went through the process of identifying the business problem, specifying the required data, extracting and preparing the data, performing machine learning by clustering the data into clusters based on similarities, and then giving recommendations to the relevant stakeholders. The analysis was performed based on the toolset of data science and relied heavily on the use of Python and Python libraries such as Pandas, Scikit, Folium. 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 setting up a new restaurant.

References

The Jupyter notebook of the analysis can be found on GitHub.
https://github.com/misstobitaiwo/Coursera_Capstone/blob/master/My%20Capstone%20Project.ipynb