Location Based Market Analysis to Identify Optimal Business Location in Colombo District and Suburbs

Shazly Shanawaz

IBM Data Science Capstone Report

June 12, 2020

Introduction

Background

Colombo is a vibrant and dynamic city and the commercial capital of Sri Lanka with a population of more than 3 million. Colombo is the designated hub for businesses due to the majority of corporations having head offices located and many, a large harbor with strategic positioning along the sea trade routes and an international airport located close to the heart of the city. As the commercial hub of the country, Colombo is seen as a hotspot for identifying new markets, starting new businesses and investment opportunities attracting both local and foreign businessmen and investors.

Business Problem

Selecting the optimal location to perform business operations is essential as it the business gains many opportunities when placed in a location withing close proximity of the target audience and enables the business to easily reach out to current and potential customers and manage logistics efficiently reducing operational costs. Colombo consists of many residential and commercial areas with its own unique blend of venues such as restaurants, supermarkets, consumer stores and service businesses. This makes it challenging for businesses to identify the best location to establish business operations. The project aims to perform unsupervised clustering to effectively identify and cluster similar areas based on the available location data.

Interested Parties

The research would be highly valuable for potential businesses and investors interested in starting a business in Colombo due to the competitive advantage and business value gained by a strategically located business.

Dataset, Features and Preprocessing

Data Sources

The data required for the dataset was collected through three different data sources and compiled into the final dataset. The initial data requirement was to obtain the data related to the

neighborhoods of Colombo which was extracted from the Wikipedia page containing postal codes of Sri Lanka [1]. Then the coordinates of the neighborhoods would be extracted using the GeoPy Python package [2] and finally the venue details for the coordinates will be extracted using the Foursquare API [3]. The dataset would contain the suburb name, the coordinates of the area and nearby venues (e.g. Parks, Supermarkets, Restaurants, Retail Stores, Service Businesses, Hotels) extracted using Foursquare API within the limited radius of the area.

Preprocessing

The initial dataset obtained from the Wikipedia page contained the data of the province, cities and postal codes of each district. The data of other districts except Colombo were dropped from the dataset as we would only require the data regarding cities in the Colombo district.

During the initial analysis of the dataset, it was identified that some of the city names in the dataset were outdated and replaced with new names. As this would cause potential errors when trying to obtain location data using the city names, the outdated city names were identified and replaced with the current names as to avoid any errors when calling the GeoPy package.

The data was then used in the GeoPy package to obtain the coordinates of each city, and the coordinates were combined with the existing data forming a new dataset. A map was created using folium to verify the positioning and accuracy of the location data as inaccurate data would lead to capturing of inaccurate venue data.

The coordinates data were then used to obtain location data regarding nearby venues in the location by using the Foursquare API, the results provided by the API which included venue, venue category and venue coordinates were combined into one dataset which can be used for exploratory data analysis.

Feature Selection

After the data preprocessing is done, the venue data is analyzed to check the number of venues for each neighborhood and also the unique categories from all the returned venues to get a better understanding of the dataset.

The dataset returned 107 unique venue categories and the categories were listed as columns in the dataset for feature extraction. Then the frequency of each venue category was obtained and the most common venues types were identified for each of the neighborhoods.

Table 1: Feature Selection

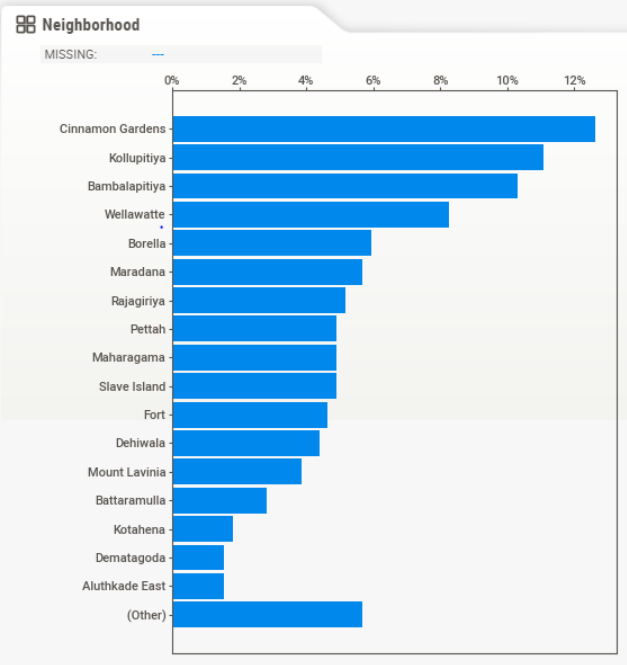
Features Kept	Features Dropped	Reason
District, City, Latitude, Longitude, Cluster Labels, Venue Category, Most Common Venues (Sorted by order)	Province, Other districts except Colombo	Out of scope and irrelevant for the purpose
	Postal Code, Venue Latitude, Venue Longitude	Dropped as not a required feature to train the data model
	Venue Name	As the frequency of venue category already depicts the same data required

Exploratory Data Analysis

Initial Analysis

For the initial exploratory data analysis, the opensource python library Sweetviz was used to create visualization graphs on the dataset in order to better understand the venue types and the distributions of venues across each neighborhood. The graphs helped understand the nature of each type of neighborhood.

Understanding the distribution of venue data for each of the neighborhood.



TOP CATEGORIES		
Cinnamon Gardens	49	13%
Kollupitiya	43	11%
Bambalapitiya	40	10%
Wellawatte	32	8%
Borella	23	6%
Maradana	22	6%
Rajagiriya	20	5%
Pettah	19	5%
Maharagama	19	5%
Slave Island	19	5%
Fort	18	5%
Dehiwala	17	4%
Mount Lavinia	15	4%
Battaramulla	11	3%
Kotahena	7	2%
Dematagoda	6	2%
Aluthkade East	6	2%
(Other)	22	6%
ALL	388	100%

Here we are able to see the distribution of the venues among the neighborhoods in Colombo, where Cinnamon gardens have the highest venue locations followed by Kollupitiya and Bambalapitiya.

Understanding the distribution of the venue category in the dataset

Venue Category

21	5%	Bakery
15	4%	Clothing Store
14	4%	CafÃ©
14	4%	Asian Restaurant
14	4%	Hotel
13	3%	Restaurant
12	3%	Coffee Shop
11	3%	Chinese Restaurant
10	3%	Fast Food Restaurant
9	2%	Bus Station
9	2%	Pizza Place
8	2%	Women's Store
8	2%	Train Station
8	2%	Bookstore
7	2%	Department Store
7	2%	Indian Restaurant
7	2%	Cosmetics Shop
6	2%	Seafood Restaurant
6	2%	Vegetarian / Vegan Restaurant
6	2%	Supermarket
6	2%	Gym
6	2%	Platform
6	2%	Bar
5	1%	Sri Lankan Restaurant
5	1%	IT Services
5	1%	Convenience Store
5	1%	Pub
5	1%	Shopping Mall
5	1%	Movie Theater
4	1%	Food Court
4	1%	Men's Store
4	1%	Cricket Ground
4	1%	Dessert Shop
4	1%	Market
4	1%	Theater
3	1%	Bubble Tea Shop
3	1%	Electronics Store
3	1%	Casino
3	1%	Pool
3	1%	Italian Restaurant
3	1%	Office
3	1%	Thai Restaurant
3	1%	Boutique
3	1%	Nightclub
2	1%	Multiplex
2	1%	Juice Bar
2	1%	Mediterranean Restaurant
2	1%	Athletics & Sports
2	1%	Grocery Store
2	1%	Pool Hall
2	1%	Playground
2	1%	Cocktail Bar
2	1%	Spa
2	1%	Golf Course
2	1%	History Museum
2	1%	Art Gallery
2	1%	Furniture / Home Store
2	1%	Concert Hall
2	1%	Japanese Restaurant
2	1%	Ice Cream Shop
52	13%	(Other)

Here we are able to identify analyze the types of venues in the dataset with bakery, clothing stores, cafes and Asian restaurants being the most common types of venues in most neighborhoods. Also, it was identified that a considerable portion of the venue categories was

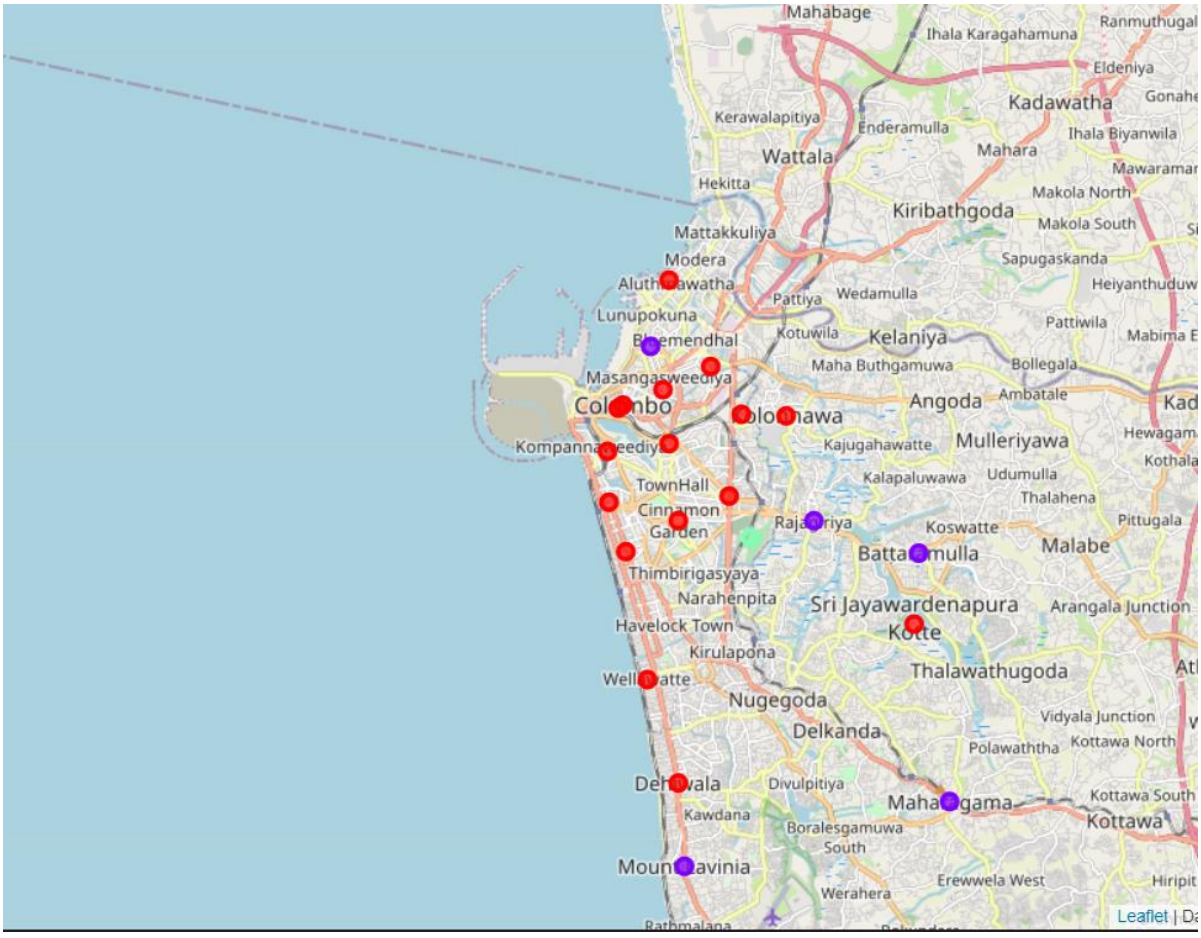
labelled under “Other” which could include many types of businesses not defined in the category list.

Model and Techniques

When training the model, I used the frequency of the venue category for each neighborhood and the most common venue category of the given neighborhood as the features

K – Means Clustering

The K – Means Clustering model was selected due to its ease in training clustering models with frequency distribution and achieve reasonable accuracy scores. The frequency of the venue category was used to find out the most common venues available in each neighborhood and was used in the training model.



Discussion and Conclusion

Results

- The first cluster consists of the consumer target market with almost all businesses targeted at residential consumers, with the most common venues being restaurants, coffee shops, supermarkets, consumer stores and other recreational venues for customers. This cluster group is best suited for businesses looking to sell their end product or services to everyday consumers.

- The second cluster consists of a combination of consumer market and business market with restaurants, super markets and also IT companies, electronic stores and other service companies, with a mix of businesses which cater to consumers and also businesses which cater to business users and industries. These neighborhoods are optimal if the business looks to sell its products or services to both target markets

- The third cluster consists of the business and industrial target market where most of the businesses cater towards other businesses or industries, and have relatively a smaller number of consumer-focused businesses compared to the other clusters

Observations

In the study I have analyzed and identified the neighborhoods which share similar features in terms of venue location and the possibility of clustering into groups based on the target market. This model can be very helpful for potential businesses and investors can identify the best area with their target audience to perform business operations.

Future

The dataset in the study contained details regarding venues in the proximity of the neighborhoods which makes the model identify the potential market with fairly accurate results, however I’ve noticed that there are more factors which affects selecting a strategic location for business, which includes the resources and suppliers in the area and the logistics involved in the business which would make the model even more relevant and would help improve the models performance.

Reference

[1] Postal Codes in Sri Lanka, Wikipedia: https://en.wikipedia.org/wiki/Postal_codes_in_Sri_Lanka

[2] GeoPy Geocoder Python: https://geopy.readthedocs.io/en/stable/

[3] Foursquare API: https://developer.foursquare.com/docs/api-reference/venues/search/