Exploring the Suburbs of Colombo to Open a Bakery

Modisha Jayaratne

June 13, 2019

2. Data Extraction and Preprocessing

2.1.Data Sources and Extraction

The Colombo district is divided into 13 Divisional Secretary's Divisions. Generally, these divisions are named after the main city falling within that division. The details of these divisions are available on this wikipedia page. The names of the divisions and the total population of each division were scraped from this page.

As supplementary information, the average real estate prices for the divisions of interest were included in the analysis. This information was extracted from Lanka Property web page by scraping the table "Average Land perch prices in Western Province (Q1 2018)".

To obtain the details of the popular venues of a given division, Foursquare API was utilized. The "explore" endpoint was used to obtain the venues of highest footfall retrieved in real-time.

In order to access the Foursquare location data, it was necessary to identify the coordinates of each division. This was achieved with the help of Nominatim Geocoder available in GeoPy library.

2.2.Data Preprocessing

After scraping the raw data off the websites, there were few challenges in terms of cleaning the data before proceeding in to the analysis stage.

Firstly, the focus of this analysis is on Colombo suburban area. Therefore, the Colombo and Thimbirigasyaya divisions had to be removed from the dataset as they fall within Colombo city limits. Next, Sri Jayawardenapura Kotte had to be renamed as Kotte for the purpose of obtaining Foursquare data.

In the property price data, the prices are available for major cities. Hence, the average property price in the entire division was assumed to be equal to the average property price of the main city of that division. In the event where the division name is different from that of the main city (i.e. Kotte and Seethawaka), the price of the main city (Ethul Kotte and Avissawella respectively) had to be assigned explicitly. Further, the city of Ratmalana is spelled differently as Rathmalana in property data. This had to be renamed, as the table join does not recognize the two as identical.

Below Table 1 indicates the dataset after the preprocessing steps:

	DivSec	Population	Land Price	Latitude	Longitude
0	Dehiwala	87834	3131549	6.851279	79.865977
1	Homagama	236179	271067	6.841273	80.003058
2	Kaduwela	252057	350032	6.935703	79.984331
3	Kesbewa	244062	353686	6.795740	79.940848
4	Kolonnawa	190817	825000	6.932625	79.890314
5	Maharagama	195355	1030166	6.847278	79.926608
6	Moratuwa	167160	809972	6.774682	79.882610
7	Padukka	65167	95720	6.841538	80.091647
8	Ratmalana	95162	1112115	6.815259	79.866778
9	Seethawaka	113477	164454	6.952948	80.218633
10	Kotte	107508	2010233	6.888322	79.918741

Table 1: Preprocessed Data of Colombo Suburbs