**Exploring the Suburbs of Colombo to Open a Bakery**

**Modisha Jayaratne**

**June 12, 2019**

# Data Acquisition and Preparation

## 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](https://en.wikipedia.org/wiki/Colombo_District" \t "_blank). 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](https://www.lankapropertyweb.com/house_prices.php" \t "_blank) 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](https://developer.foursquare.com/" \t "_blank) was utilized. The "explore" endpoint was used to obtain the venues of highest footfall at the time of executing the command.

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](https://geopy.readthedocs.io/en/stable/" \l "nominatim" \t "_blank) Geocoder available in GeoPy package.

## Data Cleaning

After scraping the raw data off the websites, there were few challenges in terms of cleaning 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 table scraped from wikipedia 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 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 would not recognize the two as identical.

Below figure indicates the dataset after the preprocessing steps:



Figure 1: Preprocessed Data of Colombo Suburbs