Exploring the Suburbs of Colombo to Open a Bakery

Modisha Jayaratne

June 13, 2019

1. Introduction

1.1.Background

The district of Colombo in Sri Lanka is the most densely populated district in the island. As per the statistics released in 2012 by the Department of Census and Statistics in Sri Lanka, the total population of the district exceeds 2.31 million. Of this total population, 0.75 million live in the city of Colombo, and the remaining 1.56 million live in suburban areas. Regardless of the fact that majority of the population lives outside the city center, most of the commercial venues are restricted to Colombo city limits.

1.2.Problem

Even though there is a recent trend among corporates to open their new offices outside the hustle and bustle of the city center - mostly owing to the skyrocketing real estate prices, the dining options offered to the work force of those offices are not growing at a similar pace. For breakfast or a quick snack, they have to depend on mobile boutiques that sell food items of questionable quality. Those who are conscious of quality, rely on food delivery mobile apps that pick up food from high-end bakeries in Colombo city and deliver to the office. Considering the traffic jams during the daytime, the delivery usually takes more than half an hour, meaning the order should be placed well in advance, and also it is not ideal for a quick bite.

Further, as the real estate prices in the city of Colombo rise exponentially, new residents looking to settle in Colombo district tend to explore the suburbs where property prices are more affordable. They are constantly looking for good quality baked goods as a convenient breakfast for kids or to indulge a sudden craving. However, the bakery industry in the Colombo suburban areas is not growing at a rate proportional to the rising demand in these areas.

1.3.Objective and Target Audience

This analysis was performed with the aim of addressing the aforementioned problem. The main objective was to analyze the Colombo suburban areas and to identify the most favorable location to open a new bakery.

The main target audience of this analysis is the bakery owners who are interested in opening their next outlet in Colombo suburban area, but unsure of the ideal location.

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

3. Methodology

3.1.Exploratory Data Analysis

With the preprocessed dataset in place, the Foursquare API was called to obtain the most popular venues in real-time. For a given division, an area within 5km radius from the main city was considered and a maximum of 100 venues were retrieved. This returned a total of 637 venues across all 11 divisions. The distribution of the number of venues across the divisions is indicated in Table 2.

	DivSec	Venue
0	Dehiwala	100
1	Homagama	38
2	Kaduwela	29
3	Kesbewa	25
4	Kolonnawa	100
5	Kotte	100
6	Maharagama	100
7	Moratuwa	31
8	Padukka	7
9	Ratmalana	100
10	Seethawaka	7

Table 2: Number of Venues

The Foursquare data was further analyzed to identify the top ten most common venue categories for each division. Table 3 shows the result of this analysis.

	DivSec	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Dehiwala	Bakery	Restaurant	Clothing Store	Asian Restaurant	Café	Pizza Place	Coffee Shop	Women's Store	Cosmetics Shop	Beach
1	Homagama	Supermarket	Convenience Store	Chinese Restaurant	Bakery	Gym / Fitness Center	Shopping Mall	Pizza Place	Restaurant	Bus Station	Asian Restaurant
2	Kaduwela	Pizza Place	Restaurant	Asian Restaurant	Supermarket	Gym	Fast Food Restaurant	Diner	Movie Theater	College Cafeteria	Snack Place
3	Kesbewa	Bus Station	Grocery Store	Gym / Fitness Center	Department Store	Chinese Restaurant	Shopping Mall	Snack Place	Flea Market	Clothing Store	Pizza Place
4	Kolonnawa	Restaurant	Dessert Shop	Pub	Bakery	Café	Italian Restaurant	Hotel	Sri Lankan Restaurant	Seafood Restaurant	IT Services
5	Kotte	Bakery	Gym	Restaurant	Convenience Store	Café	Supermarket	Clothing Store	Asian Restaurant	Coffee Shop	Fast Food Restaurant
6	Maharagama	Supermarket	Bakery	Gym	Convenience Store	Chinese Restaurant	Pizza Place	Asian Restaurant	Restaurant	Café	Bookstore
7	Moratuwa	Clothing Store	Restaurant	Train Station	Chinese Restaurant	Pizza Place	Resort	Food Court	Supermarket	Juice Bar	Fast Food Restaurant
8	Padukka	Bakery	Resort	Train Station	Bus Station	Tea Room	Shopping Mall	Women's Store	Fair	College Cafeteria	Comfort Food Restaurant
9	Ratmalana	Restaurant	Pizza Place	Clothing Store	Shopping Mall	Asian Restaurant	Bakery	Chinese Restaurant	Fast Food Restaurant	Department Store	Convenience Store
10	Seethawaka	Resort	Restaurant	Pizza Place	Comfort Food Restaurant	Café	Convenience Store	Bus Station	Cocktail Bar	Coffee Shop	College Cafeteria

Table 3: Top Ten Most Common Venues

3.2.Clustering Divisions

K-means clustering is an unsupervised learning algorithm that helps to identify structure in the data otherwise less obvious to the naked eye. This algorithm was used for the analysis with the intention of uncovering the hidden similarities between different divisions and grouping them appropriately. The "Yellowbrick" library was utilized to determine the optimal k value based on the "elbow method".

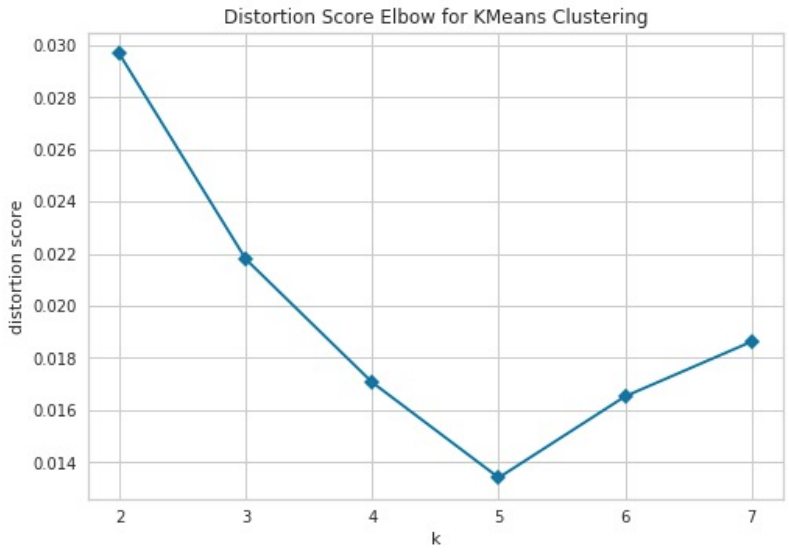

Figure 1: Elbow Method

The above graph in Figure 1 suggested using k=5 would be optimal. K-means clustering was thus performed on the dataset using the "Scikit" library, and the resulting clusters were populated on a map using the "Folium" library to understand the results better. This map is shown in Figure 2.

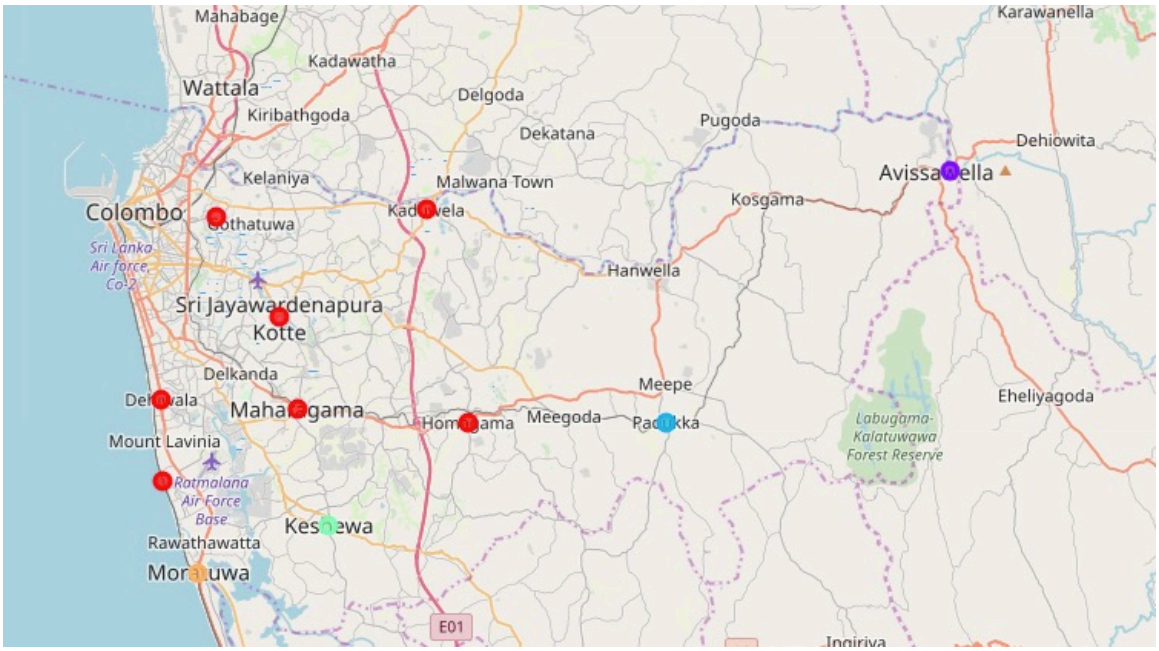

Figure 2: Visualization of Division Clusters

3.3.Focused Analysis on Bakeries

As the key focus of this project is the bakery business, the data on bakeries were filtered out for further analysis. A visual representation of the bakeries spread across the 11 divisions is indicated below in Figure 3.

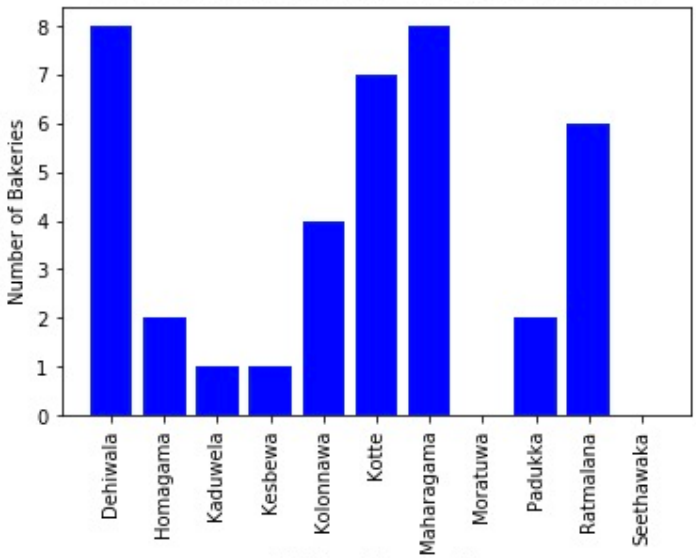

Figure 3: Number of Bakeries in each Division

The summary of the above data is tabulated below in Table 4 as a classification of divisions purely considering the number of popular bakeries in each division.

	Number of Bakeries	DivSec
0	0-1	Kaduwela, Kesbewa, Moratuwa, Seethawaka
1	2-3	Homagama, Padukka
2	4-5	Kolonnawa
3	6-7	Kotte, Ratmalana
4	8-9	Dehiwala, Maharagama

Table 4: Classification of Divisions based on Number of Bakeries

This data was replicated on a choropleth map using the "Folium" library to better understand the variability of the number of bakeries across the geographical areas. In order to achieve that, an approximate geojson map of divisions in Colombo district had to be created using geojson.io tool, as this was not readily available online. The resulting choropleth map is shown in Figure 4.

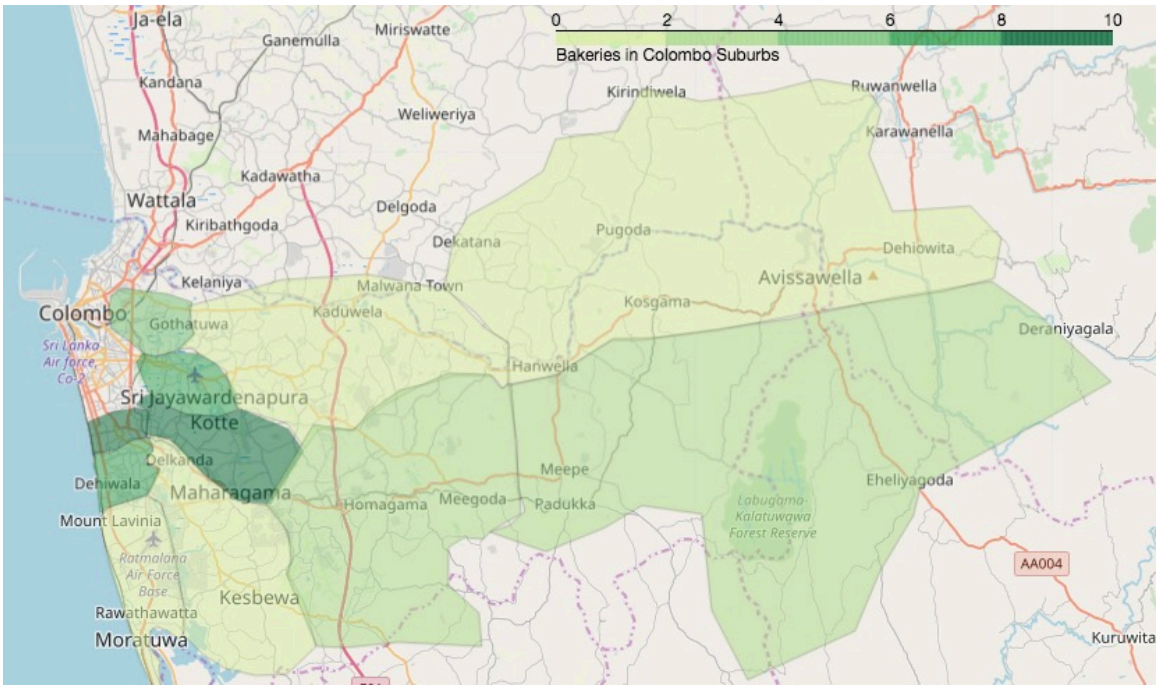

Figure 4: Choropleth Map

However, it is understood that arriving at conclusions merely relying on the number of bakeries is misleading. It is important to calculate a per capita figure in order to be able to compare the divisions as the population varies from one division to another. The population data that was extracted initially were used to calculate the number of bakeries for 100,000 people in each division. When two divisions are similar in terms of the per capita bakeries, property price was used as the tiebreaking criteria. Therefore, to derive the most preferable to least preferable division, the dataset was first sorted by the per

capita bakeries figure, and then by the land price in the ascending order. The final results are in Table 5.

	DivSec	No of Bakeries	Population	Bakeries per 100,000	Land Price
0	Seethawaka	0	113477	0.0	164454
1	Moratuwa	0	167160	0.0	809972
2	Kaduwela	1	252057	0.4	350032
3	Kesbewa	1	244062	0.4	353686
4	Homagama	2	236179	0.8	271067
5	Kolonnawa	4	190817	2.1	825000
6	Padukka	2	65167	3.1	95720
7	Maharagama	8	195355	4.1	1030166
8	Ratmalana	6	95162	6.3	1112115
9	Kotte	7	107508	6.5	2010233
10	Dehiwala	8	87834	9.1	3131549

Table 5: Preferential Sequence of Divisions

In order to visualize the results more easily, following line plot in Figure 5 was generated. For the two variables to fall in the same spectrum, the land price was normalized by dividing it by 100,000.

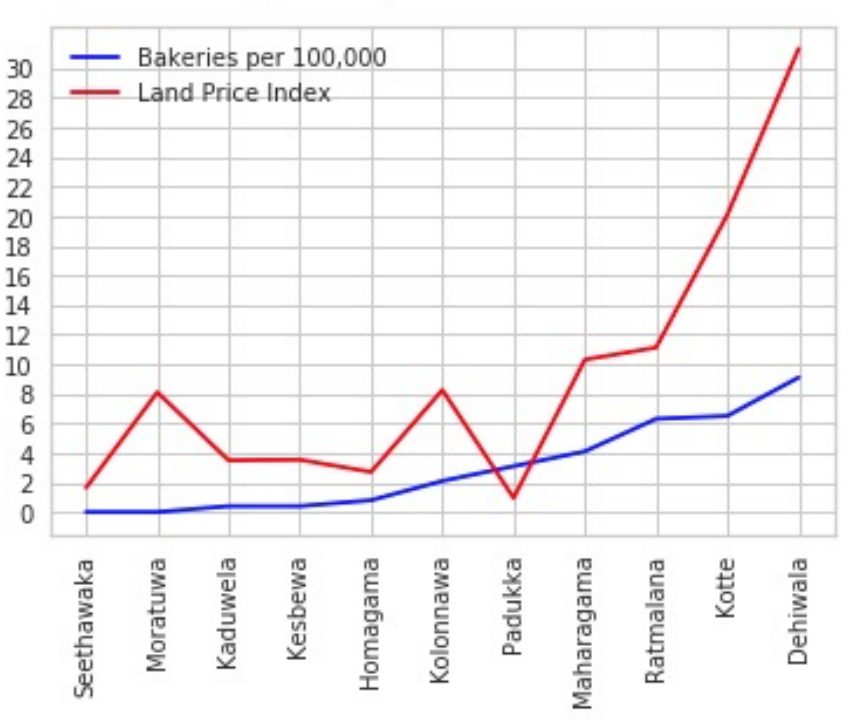

Figure 5: Per Capita Bakeries vs Land Prices

4. Results

The outcome of the clustering exercise is summarized in the Figure 6 below. It should be emphasized that clustering was based on the top ten most common venues in the divisions.

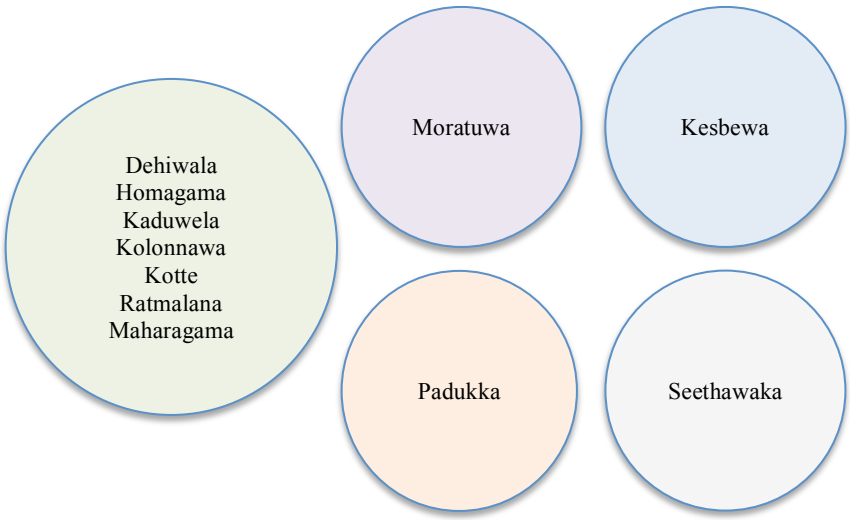

Figure 6: Division Clusters

However, when studying the choropleth map in Figure 4, it is apparent that the bakeries resonate with the clustering results only to a certain extent. This implies that bakeries alone do not follow the same pattern when all the venues are considered together.

This pattern changed further when the per capita bakery count was introduced into the mix. Seethawaka, Moratuwa, Kaduwela, Kesbewa and Homagama are the top five divisions to consider where the number of bakeries per 100,000 people is less than one.

Seethawaka and Moratuwa divisions seem most favorable to open a bakery, as both divisions do not have any bakeries among top ten trending venues. When the cost of real estate is added as a second criterion, Seethawaka is clearly on the lead, since property prices in Moratuwa are drastically higher than in Seethawaka. If the investment in real estate is a concern, one might even prefer Kaduwela, Kesbewa and Homagama divisions to Moratuwa division. This is acceptable, as shown in Figure 5; the per capita bakeries figure does not vary much compared to the variability in real estate prices.

5. Discussion

This analysis is solely based on the data acquired from external sources as listed in section 2. Thus, the reliability of the analysis is directly dependent on the reliability of the data, and it has not been verified as part of this study.

The main limitation that should be noted is that there could be many venues that are not captured by Foursquare. Sri Lanka is a small country, and has many venues that are yet to appear online. If these locations were to be included in the analysis, the outcome would have been considerably different.

The population statistics were as per the census collected in the year 2012. This data may have changed significantly over the past seven years. In addition, according to the source web page, the land prices are from the first quarter of 2018. The real estate prices in Sri Lanka are rising at an exponential rate. Hence, this data may also be outdated as more than a year has passed to date.

Further, in the analysis, only the popular venues, population and the land price of each division were considered. However, in order for a bakery owner to explore the prospects of a new bakery, there may be many other aspects to consider. Some of these factors would include, the number of workplaces and schools in close proximity, availability of vacant commercial property, preference for rental/leased spaces and the quality of infrastructure. In order to arrive at a concrete conclusion, a comprehensive analysis should be conducted with more data pertaining to these other dynamics.

6. Conclusion

In this study, the Colombo suburban area was studied with a view of identifying the optimal division to open a new bakery. The popular venues and the population of each division were the main factors included in the study. In addition, the land prices were used as a supplementary factor in a tiebreaking scenario. With these three factors, the conclusion was that Seethawaka is the most favorable division to open a bakery. However, this conclusion is subject to limitations such as the reliability of data and the limited factors considered.