Proposing best areas to live nearby Colombo

Senake Sandanayake

May 15, 2021

1. Introduction

1.1 Background

Colombo being the commercial capital of Sri Lanka, fairly a high percentage people who seek for an employment or people who seek for a better break in his/her career or people who need to change his/her career or Sri Lankans those who are living in other parts of the world and hoping to be back will be moved to Colombo & its suburbs and will be settled down in these areas.

Following are the main factors that drive people to move to Colombo & its suburbs.

a) Most of the Headquarters of Departments, Authorities, Boards, and Companies are situated in Colombo.
b) Availability of best Government schools, best Private Schools & best International schools in Colombo and its suburbs
c) Availability of other facilities like various kinds of restaurants, shopping malls, entertainment places etc…
d) Availability of newly built apartments & housing schemes at a competitive rate.
e) High development rate compared to other areas.

When people are planning to move to Colombo & its suburbs, followings are the main requirements that they usually consider.

a) Less populated area: (As they are coming from other parts of the country which are mostly less populated, the natural preference will be to select a similar environment.)
b) Area which is within reasonable distance from city of Colombo.
c) Area where you get all the necessary facilities around you. (Restaurants, Supermarkets, banks, ATMs, shopping places etc…)

1.2 Problem statement

As a Data Scientist how can you help a person to find out an area/areas in Colombo & its suburbs that satisfy the above mention three requirements?

2. Data sources

Sri Lanka is divided into 25 districts geographically. Each district is comprised of several Divisional Secretariats (DS Divisions) which are demarcated geographically. Colombo city lies in Colombo District itself. There is one more district very close to Colombo city, which is Gampaha district, as depicted in the below map. Therefore DS Divisions of these two districts (Colombo & Gampaha) will be considered to analyze.

DS Division, Population Density for Colombo & Gampaha districts are available in following **Wikipedia** sites:

https://en.wikipedia.org/wiki/Colombo_District

https://en.wikipedia.org/wiki/Gampaha_District

Latitude & Longitude for each DS Division will be obtained using **Nominatim geodecoding** service.

Foursquare API will be used to get the nearby venues for the DS Divisions.

3. Methodology

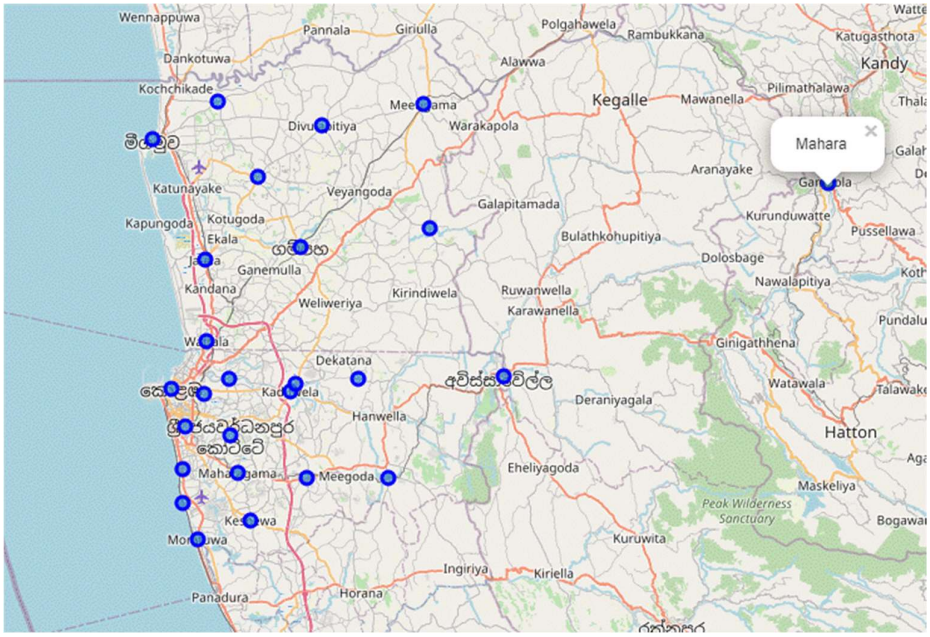
3.1 Data Visualization & Cleaning

Data scraped from two Wikipedia sites were combined into one table. Then latitude & longitude for each DS Division was added to the table using Nominatim geodecoding service. The features of the data frame were DS_Division, Population_Density, Latitude & Longitude. After that checked whether there were any missing values in the final table. No missing values were found.

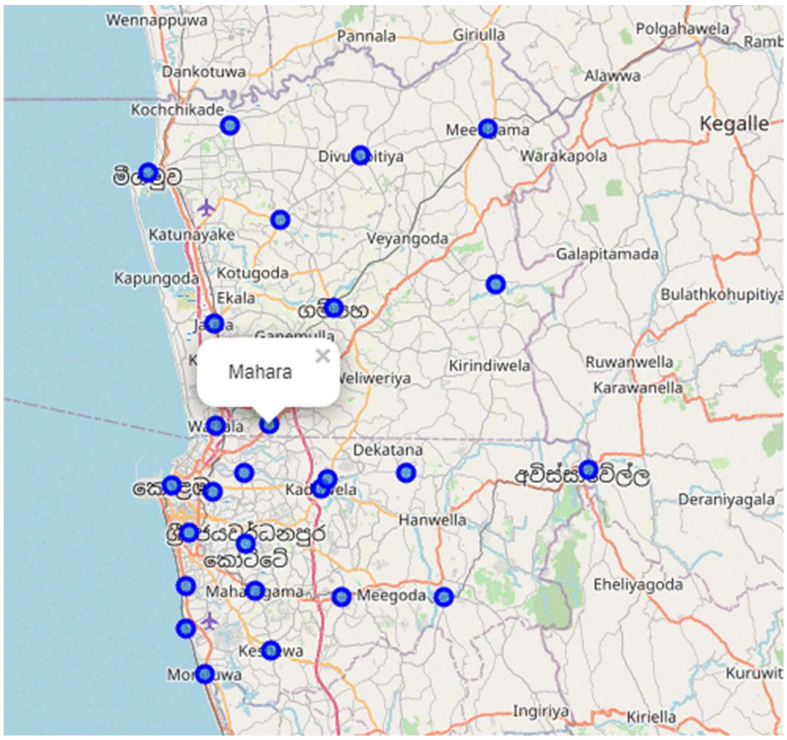
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26 entries, 0 to 25
Data columns (total 4 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   DS_Division         26 non-null     object
 1   Population_Density  26 non-null     object
 2   Latitude            26 non-null     float64
 3   Longitude           26 non-null     float64
dtypes: float64(2), object(2)
memory usage: 960.0+ bytes
```

The data type of Population_Density feature was object type and it was converted to float.

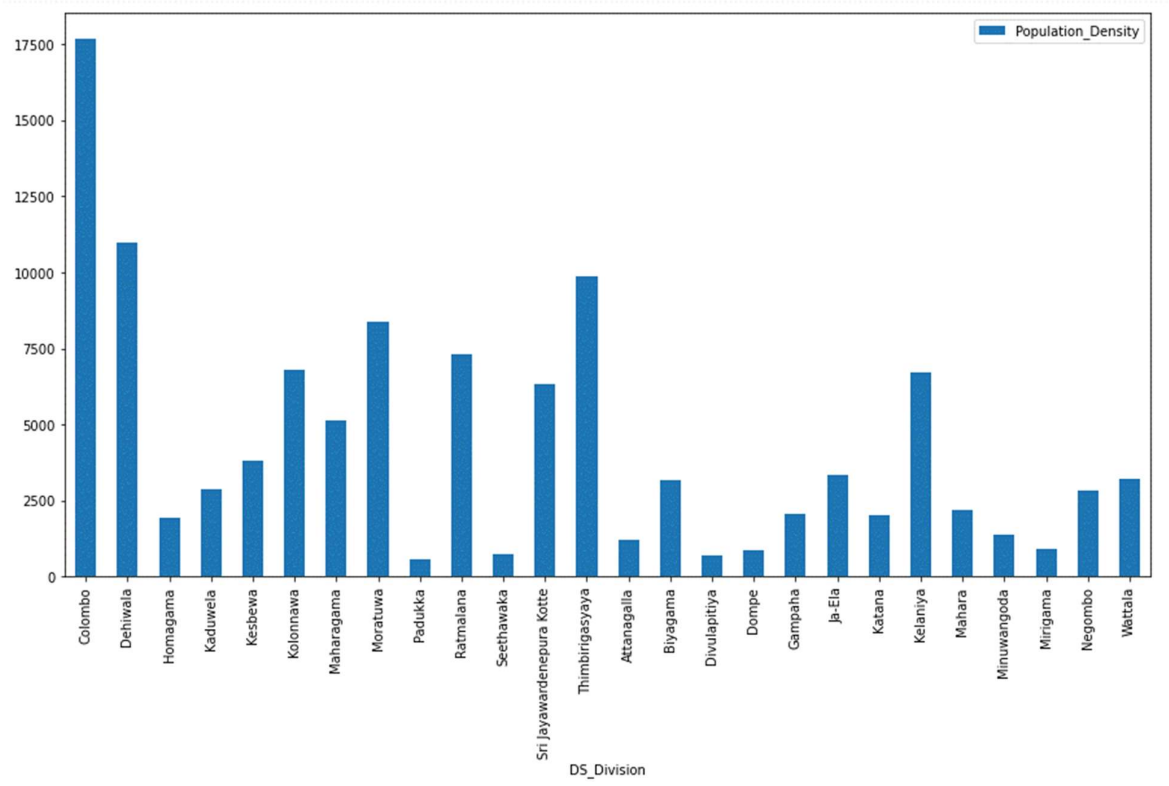
DS divisions were plotted on a map to see how the DS divisions were scattered around Colombo city as depicted below.



It was observed that geo coordinates for Mahara DS division is wrong. This is because there are two cities in the name of Mahara in Sri Lanka. This was corrected by using the same geodecoding service by giving both DS name and the District as the input. See the below map.



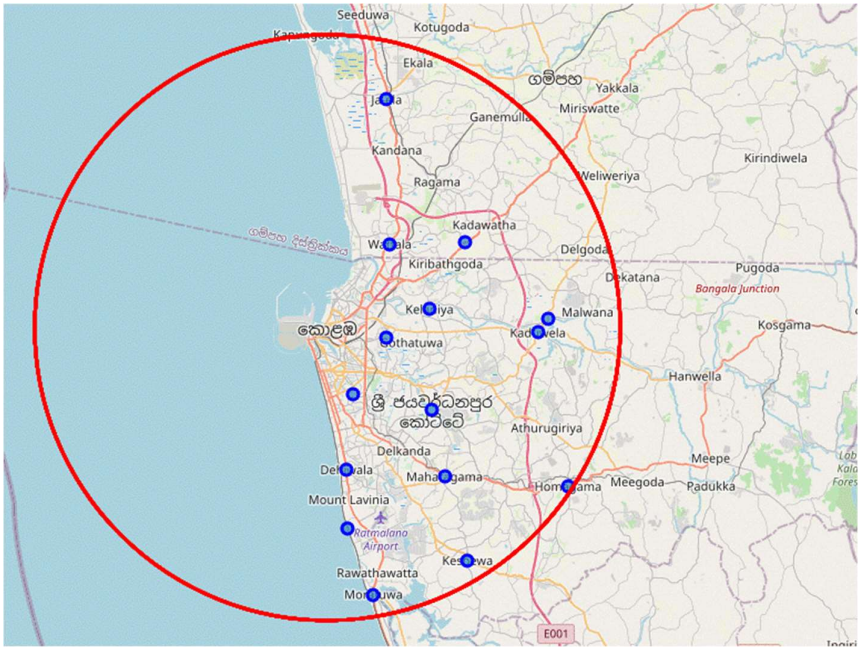
Then Population_Density was plotted against the DS_Division.



It was observed that population density in Colombo DS Division is extremely high compared to the other divisions. Therefore it was dropped from the data set as the final goal is to find less populated areas.

Distance from the Colombo city to all the DS visions were calculated using geo coordinates and new column was added to the data frame. Then DS Divisions which are more than 20km away from Colombo city were dropped. In Sri Lanka, it is quite common a person to travel 20km from his/her residence to work place on a daily basis.

Below diagram verify that final DS Divisions lies within 20km distance from the Colombo city.



Final table ended up with fifteen DS Divisions.

	DS_Division	Population_Density	Latitude	Longitude	Dis_To_Colombo
0	Dehiwala	10979.0	6.851279	79.865977	9.8
1	Homagama	1952.0	6.841273	80.003058	19.7
2	Kaduwela	2864.0	6.935703	79.984331	14.4
3	Kesbewa	3813.0	6.795740	79.940848	18.6
4	Kolonnawa	6815.0	6.932625	79.890314	4.1
5	Maharagama	5141.0	6.847278	79.926608	12.9
6	Moratuwa	8358.0	6.774682	79.882610	18.5
7	Ratmalana	7320.0	6.815259	79.866778	13.8
8	Sri Jayawardenepura Kotte	6324.0	6.888322	79.918741	9.1
9	Thimbirigasyaya	9874.0	6.897905	79.869608	4.9
10	Biyagama	3167.0	6.944490	79.990557	15.1
11	Ja-Ela	3353.0	7.079377	79.890763	16.2
12	Kelaniya	6735.0	6.950125	79.917271	7.1
13	Mahara	2205.0	6.991303	79.939375	11.1
14	Wattala	3228.0	6.989871	79.892709	7.1

3.2 Exploring Nearby Venues

Using **Foursquare API** name of the venue & category of the venue of all nearby venues that are within the radius of 2km was explored for each DS Division. Maximum number of venues for each DS Division was set to 100. Then a new data frame was created as follows.

	DS_Division	Latitude	Longitude	Venue	Venue Category
0	Dehiwala	6.851279	79.865977	Al-Ameen Traders	Women's Store
1	Dehiwala	6.851279	79.865977	Yum Yum Fine Foods	Candy Store
2	Dehiwala	6.851279	79.865977	Beach House Bistro	Beach Bar
3	Dehiwala	6.851279	79.865977	Srina Palace	Cosmetics Shop
4	Dehiwala	6.851279	79.865977	Salon Anoma	Cosmetics Shop

Then one hot encoding was used to create a new data frame with columns containing venue categories as depicted below.

	DS_Division	ATM	Airport	Arcade	Art Gallery	Asian Restaurant	Athletics & Sports	BBQ Joint	Badminton Court	Bakery	Bar	Basketball Court	Beach	Beach Bar	Bookstore	Boutique	Breakf Sp
0	Dehiwala	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	Dehiwala	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	Dehiwala	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
3	Dehiwala	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	Dehiwala	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

By grouping rows on DS Division in above data frame and taking the mean of the frequency of occurrence of each category, a new data frame was created as below.

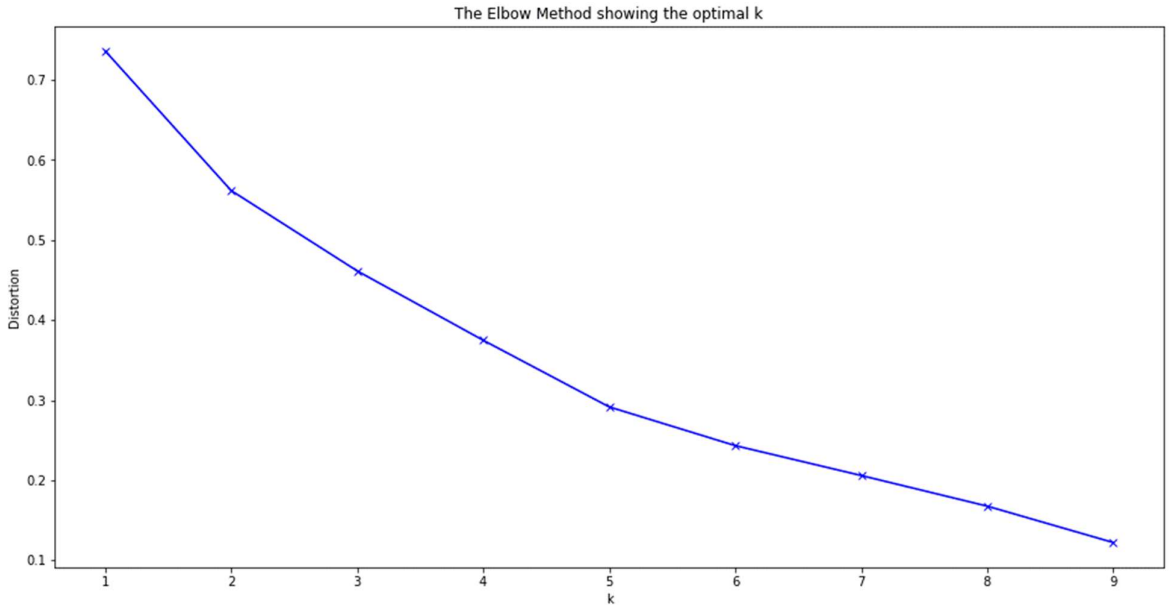
	DS_Division	ATM	Airport	Arcade	Art Gallery	Asian Restaurant	Athletics & Sports	BBQ Joint	Badminton Court	Bakery	Bar	Basketball Court	Beach	Beach Bar	E
0	Biyagama	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.090909	0.090909	0.00	0.000000	0.000000	
1	Dehiwala	0.000000	0.000000	0.011905	0.00	0.071429	0.011905	0.011905	0.000000	0.083333	0.011905	0.00	0.023810	0.011905	
2	Homagama	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	
3	Ja-Ela	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	
4	Kaduwela	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.100000	0.00	0.000000	0.000000	
5	Kelaniya	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.166667	0.000000	0.00	0.000000	0.000000	
6	Kesbewa	0.000000	0.000000	0.000000	0.00	0.052632	0.000000	0.000000	0.000000	0.157895	0.000000	0.00	0.000000	0.000000	
7	Kolonnawa	0.083333	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	
8	Mahara	0.000000	0.000000	0.000000	0.00	0.041667	0.000000	0.000000	0.000000	0.083333	0.000000	0.00	0.000000	0.000000	
9	Maharagama	0.000000	0.000000	0.000000	0.00	0.054054	0.000000	0.000000	0.027027	0.108108	0.000000	0.00	0.000000	0.000000	
10	Moratuwa	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.062500	0.00	0.000000	0.000000	
11	Ratmalana	0.000000	0.000000	0.000000	0.00	0.038462	0.000000	0.000000	0.000000	0.115385	0.000000	0.00	0.038462	0.000000	
12	Sri Jayawardenepura Kotte	0.000000	0.000000	0.000000	0.00	0.029851	0.000000	0.000000	0.000000	0.104478	0.014925	0.00	0.000000	0.000000	
13	Thimbirigasyaya	0.000000	0.000000	0.000000	0.01	0.020000	0.000000	0.000000	0.000000	0.040000	0.010000	0.01	0.000000	0.000000	
14	Wattala	0.000000	0.027778	0.000000	0.00	0.027778	0.000000	0.000000	0.000000	0.111111	0.000000	0.00	0.000000	0.000000	

Above data frame was used to cluster the DS Divisions as per their nearby venue categories.

Another data frame was created with top 10 venue categories for each DS Division which would be later use to profile each cluster.

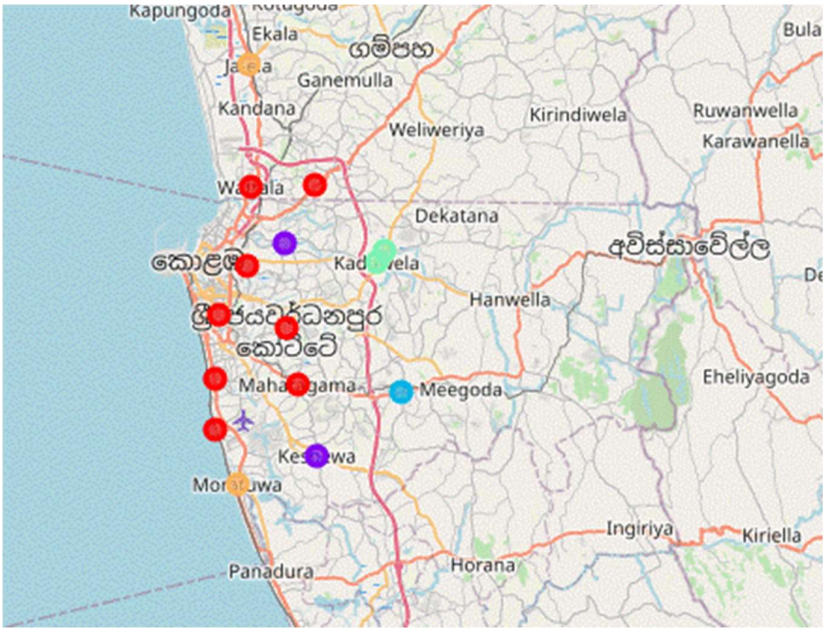
3.3 Clustering DS Divisions

K-Means clustering was used to cluster the DS Divisions. **Elbow method** was used to determine the optimum k value as depicted below.



As per the graph, optimum value of k was **5**. Therefore K-Means clustering were performed with k value of 5.

Below map shows the DS Division for each cluster.



DS Divisions and top 10 venue categories for each cluster was obtained.

Cluster -1 (Red Circles):

	DS_Division	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	Clothing Store	Asian Restaurant	Restaurant	Pizza Place	Fast Food Restaurant	Supermarket	Seafood Restaurant	Cosmetics Shop	Shopping Mall
4	Kolonnawa	Convenience Store	ATM	IT Services	Bus Stop	Gym	Juice Bar	Fruit & Vegetable Store	Office	Department Store	Park
5	Maharagama	Supermarket	Bakery	Diner	Gym / Fitness Center	Asian Restaurant	Gym	Burger Joint	Pizza Place	Chinese Restaurant	Movie Theater
7	Ratmalana	Train Station	Bakery	Restaurant	Supermarket	Clothing Store	Bookstore	Pharmacy	Pizza Place	Department Store	Pub
8	Sri Jayawardenepura Kotte	Bakery	Chinese Restaurant	Fast Food Restaurant	Convenience Store	Coffee Shop	Juice Bar	Track	Supermarket	Restaurant	Food
9	Thimbirigasyaya	Café	Bakery	Dessert Shop	Restaurant	Pub	Bookstore	Coffee Shop	Hotel	Office	Chinese Restaurant
13	Mahara	Restaurant	Supermarket	Bus Station	Pizza Place	Clothing Store	Bakery	Ice Cream Shop	Grocery Store	Hotel Bar	Gym
14	Wattala	Bakery	Train Station	Clothing Store	Restaurant	Supermarket	Toll Booth	Food	Fast Food Restaurant	Pizza Place	Comfort Food Restaurant

Cluster -2 (Violet Circles):

	DS_Division	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
3	Kesbewa	Bakery	Grocery Store	Clothing Store	Fast Food Restaurant	Bus Station	Asian Restaurant	Department Store	Chinese Restaurant	Restaurant	Supermarket
12	Kelaniya	Bus Station	Bakery	Chinese Restaurant	Supermarket	Grocery Store	Convenience Store	Department Store	Pizza Place	Fruit & Vegetable Store	Dessert Shop

Cluster-3(Blue Circles):

	DS_Division	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
1	Homagama	Train Station	Hotel	Supermarket	Clothing Store	Bus Station	Women's Store	Donut Shop	Flower Shop	Fast Food Restaurant	Electronics Store

Cluster-4(Green Circles):

	DS_Division	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
2	Kaduwela	Boutique	Bar	Chinese Restaurant	Restaurant	Bus Station	Pizza Place	Supermarket	Pharmacy	Grocery Store	Gym
10	Biyagama	Pharmacy	Bakery	Breakfast Spot	Pizza Place	Supermarket	Restaurant	Chinese Restaurant	Bar	Grocery Store	Gym

Cluster-5(Yellow Circles):

	DS_Division	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
6	Moratuwa	Pizza Place	Chinese Restaurant	Train Station	Supermarket	Bar	Restaurant	Fast Food Restaurant	Food	Lake	Tech Startup
11	Ja-Ela	Pizza Place	Train Station	Supermarket	Platform	Clothing Store	Department Store	Restaurant	Seafood Restaurant	Movie Theater	Fast Food Restaurant

4. Results

Goal was to find DS Divisions with fairly modern neighborhoods with less populated and lies within reasonable distance to Colombo City.

Out of the five clusters, cluster one mark with red circles would be the DS divisions with modern neighborhood compared to other clusters. Below table shows the DS divisions in that particular cluster with population density and distance to Colombo city.

DS_Division	Population_Density	Latitude	Longitude	Dis_To_Colombo
Dehiwala	10979.0	6.851279	79.865977	9.8
Kolonnawa	6815.0	6.932625	79.890314	4.1
Maharagama	5141.0	6.847278	79.926608	12.9
Ratmalana	7320.0	6.815259	79.866778	13.8
Sri Jayawardenepura Kotte	6324.0	6.888322	79.918741	9.1
Thimbirigasyaya	9874.0	6.897905	79.869608	4.9
Mahara	2205.0	6.991303	79.939375	11.1
Wattala	3228.0	6.989871	79.892709	7.1

From above, one could easily choose a DS division depending upon his/her priority of population density and distance to Colombo city.

5. Conclusion

In this study, I analyzed the neighborhoods around the DS Divisions which are within 20kms from the Colombo city to find out the DS Divisions which are of modern neighborhoods. Also I compared the population density and distance to Colombo city so that one could easily select a DS division depending upon his/her priority.

6. Further directions

It was observed that venues explored were lower than the actuals. This will definitely give the better clustering output. Therefore it needs to be addressed. Also I don't consider the land prices in each DS divisions which I think, will be useful for recommending an area.