Finding the ideal location to open a pizzeria in Colombo, Sri Lanka

Rayantha Sinnathamby,

July 2020.

1. Introduction

1.1. Background:

Colombo Sri Lanka has a bustling food culture with a demand for great fast casual and casual dining experiences for the hundreds of thousands who pour in and out of the city for activities ranging from shopping to its attractions and its vibrant night life. The city is also the commercial hub of Sri Lanka and as a result a considerable number of those living and working in different parts of the city possess a relatively high disposable income and are willing to spend on good quality authentic dining experiences with quick service due to the fast paced lifestyles in the city.

1.2 Business Problem:

Stakeholders looking to open an authentic pizzeria in the city of Colombo want to determine the optimum location, to maximize exposure to crowds of people looking for a relatively quick and good quality bite?

1.3 Deeper Dive:

Parts of the city is undergoing and has undergone major development in the form of infrastructure with this increased urbanization, the number of attractions is also increasing rapidly which attracts even more individuals in from neighboring suburbs.

In suburbs with high activity from rapid development a strong and diverse food culture has immersed as a result of people with medium to high disposable income travelling to and from these suburbs for work and leisure activities such as shopping, theaters, parks and other entertainment centers.

In order to maximize stakeholder value from the proposed pizzeria it is important to find the busiest suburbs of the city in terms of spending. (Malls, shopping centers, parks, theaters etc), but also those where the barrier to entry for a new entrant is low.

Although some suburbs are optimal in terms of venues where people spend, or are optimal in terms of people passing by from a busy hub they are already saturated and therefore the optimal solution for the stakeholders here is a balance between busy areas with attractive venues in terms of expenditure but also not saturated especially in terms of food places that potentially sell pizza.

2. Data Acquisition and Cleaning

2.1 Data Sources:

- A list of Suburbs in Colombo Sri Lanka were obtained from https://en.wikipedia.org/wiki/Category:Suburbs_of_Colombo.

- Latitude and longitude values for the suburbs scraped from the website were obtained using the geocode add in on Google Sheets.

- The explore endpoint of the Foursquare API was used to obtain venue data for the scraped suburbs.

2.2 Data Use:

- Colombo is a relatively small city therefore it important to capture as many subunits as possible to capture venues Wikipedia had the most complete list of Colombo suburbs.

- In order to provide suburb latitude and longitude values to the Foursquare API it was necessary to geocode the data, to do this, the geocode add in on Google sheets was used because Geopy was not able to successfully geo code these locations.

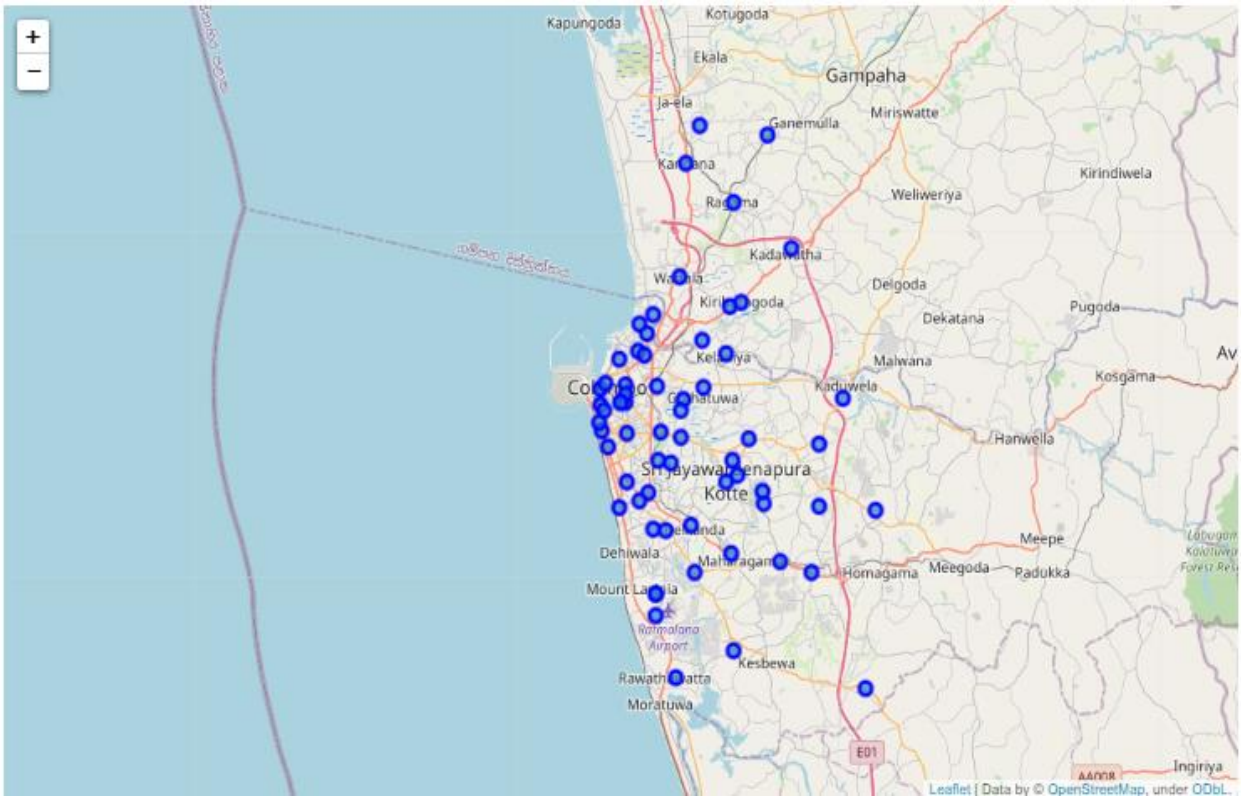
- The four square API's explore endpoint was used to obtain as much venue data as possible in these vicinities in order to do a satisfactory clustering a radius of 1000m was defined about the suburb Lat, Lon values and a maximum of 100 venues per suburb was specified.

3. Methodology

3.1 Data Extraction and Cleaning

- The data was scraped from Wikipedia using the requests.get() method and parsed using beautiful soup. The web elements on the page was used to create the data frame specifying the suburbs of the city.

- This data was saved into a csv and geocoded on Google sheets due to the Geopy library failing to geocode these locations successfully. The geocoded data read into the notebook as a data frame and plotted using Folium.The map obtained is as follows.



- Using the Foursquare API credentials and specifying a version the Explore endpoint was used to obtain 100 locations within a 1000m radius of the venue. The head of the returned dataframe is as follows.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	distance
0	Fort (Colombo)	6.933701	79.850030	& Co.	6.934104	79.842830	English Restaurant	0.044558
1	Kollupitiya	6.912815	79.850681	168 Seafood Palace	6.909465	79.850689	Seafood Restaurant	0.370509
2	Kaduwela, Western Province	6.929061	79.982775	177 Bus Halt	6.934589	79.983762	Bus Station	0.611334
3	Union Place	6.922958	79.852357	1864 Restaurant	6.919898	79.846697	Restaurant	0.338468
4	Boralesgamuwa	6.840989	79.901719	1979 Restaurant	6.834730	79.905004	Chinese Restaurant	0.692194

- The initial length of this data frame was, 1827 rows note here that venues were duplicated since the 1km radius meant overlapping of suburbs especially for suburbs closer to the center of the city. In order to fix this, the geopy libraries' distance method was used to obtain the distance from each venue to its suburb by row. The data frame was sorted by venue then ascending distance and duplicates were dropped keeping just the first occurrence of the venue. The new length of the data frame was 1160.

3.2 Data Transformation and variable selection

- In order to separate suburbs by composition the venue categories were one hot encoded, and suburb added back.

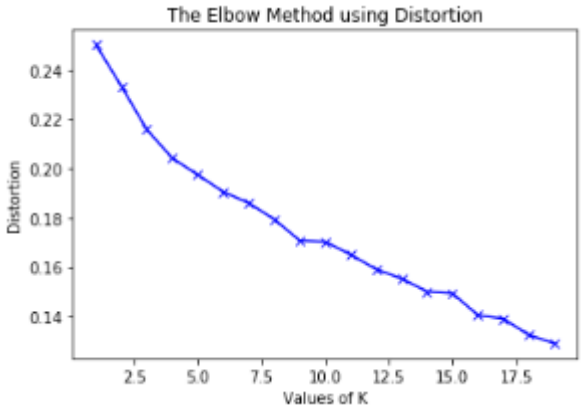
- Of the 189 venue categories in the found on the one hot encoded data frame 110 were selected based on, if they were venues that were busy, and people visited for leisure and/ or entertainment. Secondly, if they were venues that served food i.e. restaurants, pubs etc.

- The above data frame was grouped by suburb and mean calculated for each cell in order to normalize value distribution within columns. Below shows an extract of the data frame.

	Neighborhood	Accessories Store	Airport	Arcade	Art Gallery	Arts & Crafts Store	Asian Restaurant	BBQ Joint	Bakery	Bar	Beach	Bistro	Boat or Ferry	Boutique	Bowling Alley	Breakfast Spot
0	Athurugiriya	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
1	Bambalapitiya	0.0	0.0	0.0	0.027778	0.0	0.013889	0.0	0.041667	0.0	0.0	0.0	0.0	0.013889	0.0	0.0
2	Battaramulla	0.0	0.0	0.0	0.000000	0.0	0.058824	0.0	0.294118	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
3	Batuwatta	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
4	Bloemendhal	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0

3.3 Selecting K in K- Means

- K- means was used to fit the above data frame for values from 1,20 and the distortion function was calculated and plotted against K to determine the elbow. As seen below even after a few random initializations the plot descended smoothly.

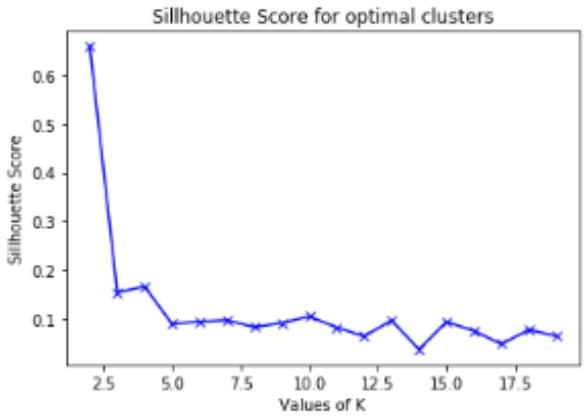


- As such the Sillhouette score was calculated in order determine number of clusters. The silhouette score is a measure of cluster quality and is calculated as follows,

  $(x-y)/\max(x,y)$

Where x is the average distance between a point to other points in its own cluster, and y the average distance between a point and points in another cluster, from the above equation smaller the average intra cluster distance and higher the inter cluster distance better will be the Silhouette score.

This metric was plotted from k= 2 to 20.



- Since this too did not give a clear indicator different values of K were experimented with and visualized. In order to visualize the clusters PCA was used to reduce the number of dimensions and Plotly was used to plot the clusters based on the first three principle components.

- After a few rounds of experimentation with the elbow plot, silhouette score and visualizations of the 3 largest principle components the optimal K was decided to be 8.

- The plot below shows the three largest clusters visualized on the principle component axes.

- The bin counts by cluster was as follows,

Cluster	Suburbs
Cluster 6	23
Cluster 0	20
Cluster 4	11
Cluster 2	3
Cluster 7	1
Cluster 5	1
Cluster 3	1
Cluster 1	1

3.4 Cluster Analysis

- For the three largest clusters two metrics were calculated by suburb,

    - Food Total- Total presence of food related venues.
    - Pizza Total- Total presence of food related venues that also potentially sell pizza.

- The two metrics were divided in order to determine a ratio. The motivation for the above two metrics is as follows. Within each cluster due to venue similarity it would be fair to compare their demand for food outlets since, the same factors contribute to demand. Assuming that the presence of food outlet by suburb portrays demand in a saturated market venues that may sell pizza will have a somewhat equal presence; therefore the goal was given similar composition, which suburbs had a high demand for food related venues lacking a similar presence of venues with pizza.

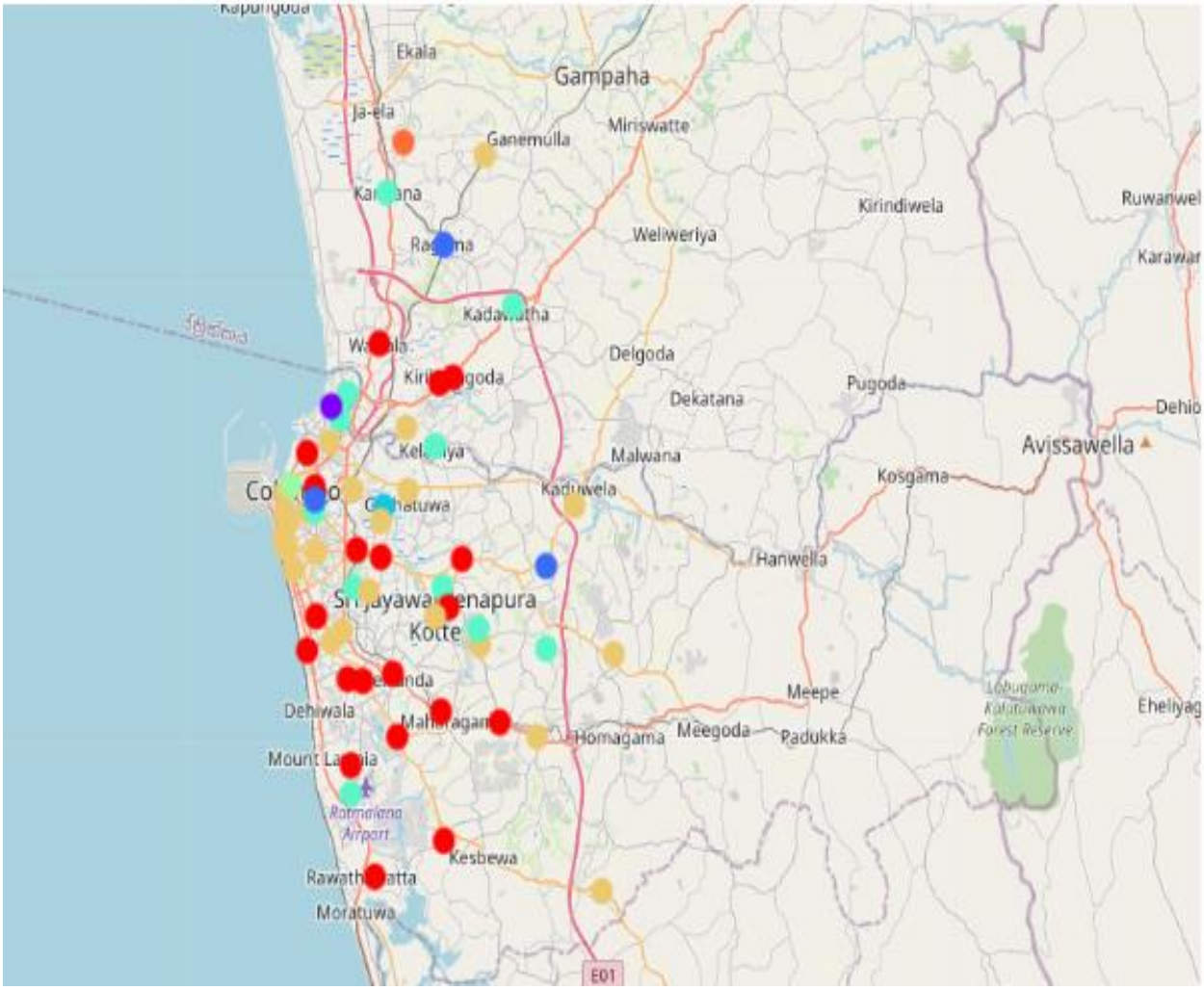
- Scatter plots using the food total, pizza total and ration by major cluster were used to determine best location, given the other known properties of the clusters. The ratio was used as the bubble size to indicate by how much the presence of food as larger than the presence of Pizza as this would help visually indicate suburbs, since larger bubble sizes are more desirable in this instance.

- Finally, the properties of the clusters and best suited locations within the clusters were compared against each other to determine ideal location.

4. Results

4.1 Cluster Distribution

- The cluster distribution map is as follows,



Here the 3 main clusters, cluster 6, 0 and 4 are seen in yellow, red, and light blue, respectively. Barring a few seemingly random allocations, yellow seems to capture the major suburbs of the business hub of the city. Red suburbs are inner as well as slightly outer suburbs capturing the busiest transportation routes in the city. The light blue suburbs even though mixed in with the slightly outer suburbs and containing major travel routes are both areas which are highly residential and also home to many businesses. E.g. of this include Battaramulla, Narahenpita, Kelaniya and Ratmalana.

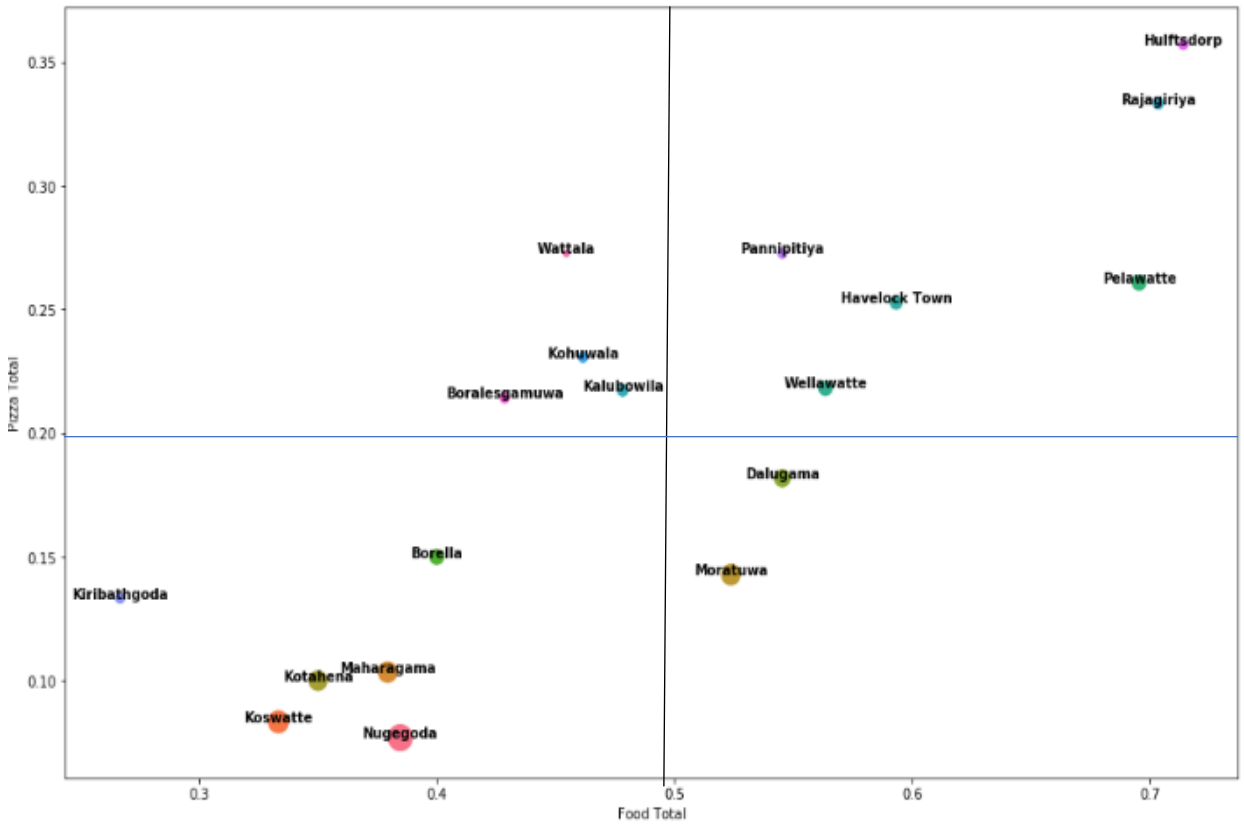
4.2 Cluster 0

- Once the two metrics and ratio were calculated an extract of cluster 0 looked as follows,

	Neighborhood	Food Total	Pizza Total	Food Pizza Ratio
42	Nugegoda	0.384615	0.076923	5.000000
28	Koswatte	0.333333	0.083333	4.000000
34	Maharagama	0.379310	0.103448	3.666667
39	Moratuwa	0.523810	0.142857	3.666667
29	Kotahena	0.350000	0.100000	3.500000

From a map of the city and prior knowledge of some suburbs it is evident that all the top (by sorting on ratio) suburbs in this cluster are inner and slightly outer suburbs of Colombo just outside the heart of the city they see heavy traffic and are also residential, and happen to be suburbs containing some of the busiest transportation routes in the city.

The scatter plot for cluster 0 is as follows,



Looking at the 2x2 scatter plot it could be seen that in this cluster two suburbs fall into the desired quadrant, where one suburb has an especially low presence of places that may sell pizza and is a potential market opportunity.

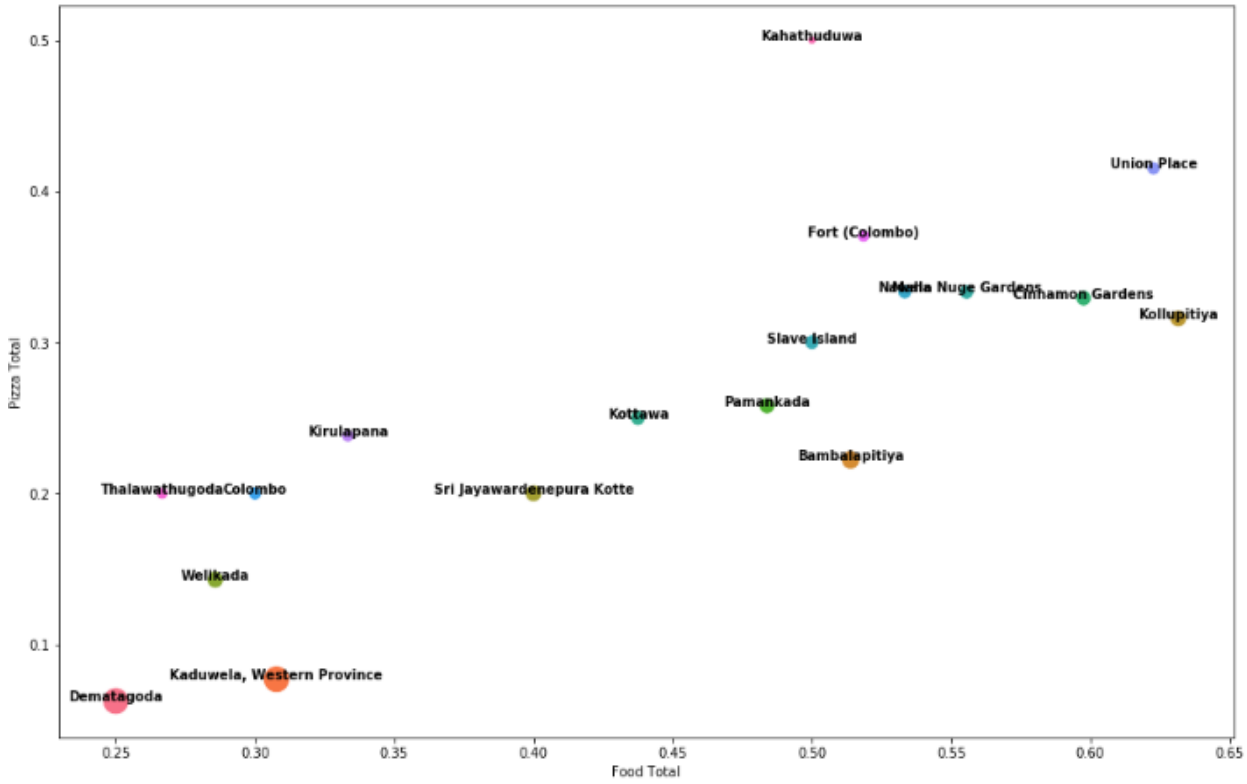
4.3 Cluster 6

- When sorted by the food to pizza ratio the top suburbs were as follows,

	Neighborhood	Food Total	Pizza Total	Food Pizza Ratio
11	Dematagoda	0.250000	0.062500	4.0000
18	Kaduwela, Western Province	0.307692	0.076923	4.0000
1	Bambalapitiya	0.513889	0.222222	2.3125
26	Kollupitiya	0.631579	0.315789	2.0000
54	Sri Jayawardenepura Kotte	0.400000	0.200000	2.0000

Apart from Kaduwela the rest of the suburbs present here are very central locations in Colombo and are prominent hubs that people travel to and from for a wide range of reasons.

The scatter plot for this cluster is as follow,



As it would be expected the suburbs of Colombo city (inner suburbs of Colombo) show a high presence of both food as well as places that sell Pizza, based on the distribution there could be some potential in Bambalapitiya as well as the countries capital Sri Jayawardenapura Kotte. It was noted although that even here that even the potential opportunities in this cluster show a much higher presence of competition that the previous cluster.

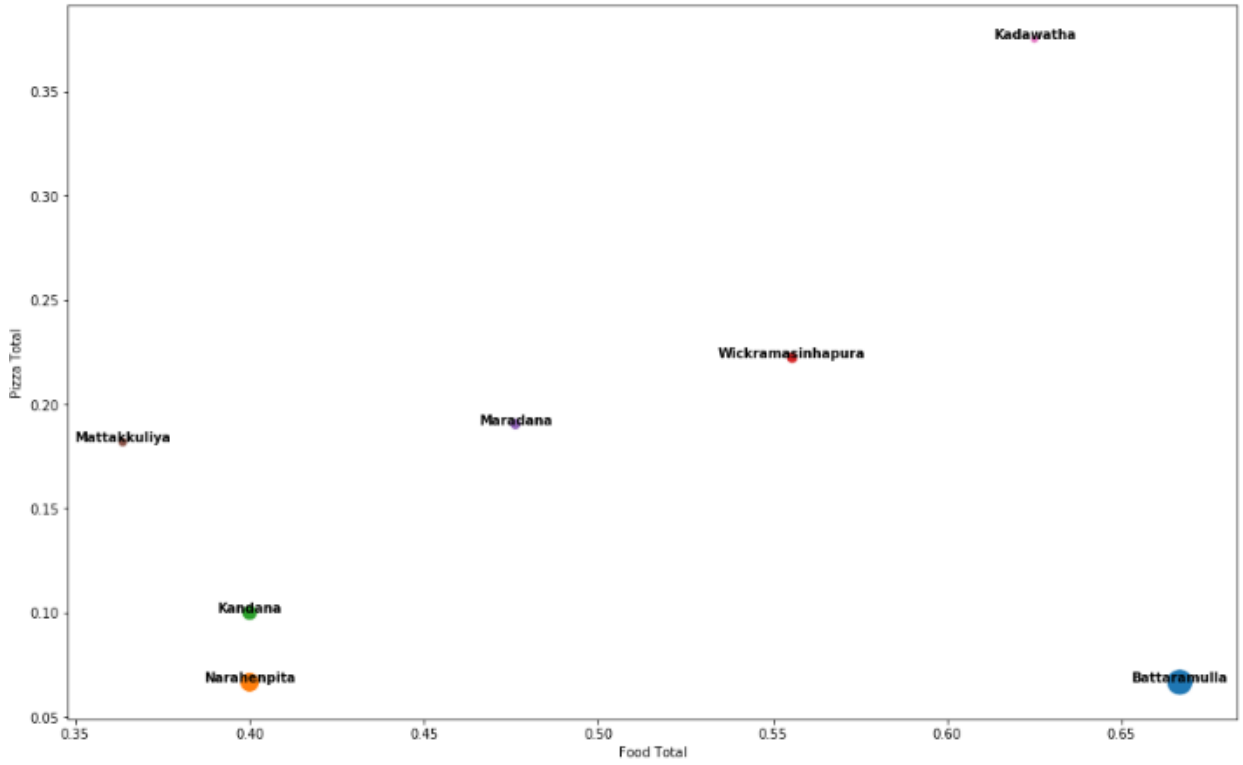
4.4 Cluster 4

- In this cluster not all suburbs intuitively made sense as a group, but looking at a majority of the suburbs it made sense these were areas that were both residential and also housing many business entities and not being in the central there also have more leisure based venues and many food places. The sorting for this cluster is as follows,

	Neighborhood	Food Total	Pizza Total	Food Pizza Ratio
2	Battaramulla	0.666667	0.066667	10.0
40	Narahenpita	0.400000	0.066667	6.0
21	Kandana	0.400000	0.100000	4.0
60	Wickramasinhapura	0.555556	0.222222	2.5
36	Maradana	0.476190	0.190476	2.5

It is immediately evident that the ratio for both Narahenpita and Battaramulla are much higher than the other clusters that were examined.

- Upon closer examination using the scatterplot,



in this cluster Battaramulla is the only best suburb for consideration since it has a very high demand for food places and not enough presence in terms of places selling pizza.

5. Discussion

- In this application of clustering localities based on venue types using K means it was not immediately intuitive what the best number for K was or if the clustering would be feasible. The way this handled was by using standard evaluation metrics such as the elbow plot and silhouette scores, but also by plotting the clusters for different instances.

- Since only 3 dimensions can be visualized in a plot it was important to use PCA even though the explained variance of the first three components was between 10%- 20%. Through this method while variable selection based on the business problem clusters that made sense could be derived.

- The variables picked were almost entirely based on the business problem definition and might not necessarily be the ideal combination of variables for this clustering.

- The idea that the presence of food related venues relates to demand is an assumption but could be justified on the basis that even if not all businesses are enjoying the same level of success the market is filled to that extent because there is a potential demand in that locality.

- The basis for picking food venues that may sell pizza is entirely based on prior encounters with different kinds of food venues and has a bias since the types of venues picked could change from person to person based on their experience and may overstate or understate the presence of potential competitors.

- The reason for creating a presence of pizza variable was to understand the degree to which food related venues in a suburb looked like against the food related venues containing pizza. The hypothesis here is that since pizzerias are as widely popular as any other type of restaurant in a market that is full the pizza variable trails close to the food variable.

- The reason for analyzing these variables within clusters is that similarity of landscape of the neighborhoods need to be established before comparing them for demand since the other variables (types of locations) are vastly different.

- This analysis could potentially be a lot better with variables that could indicate financial conditions in these areas, such as income, average prices of the same good in the different localities, housing prices. revenue of firms operating in these localities etc.

- Based on what was available and the analysis based on venue distribution (from Four Square) it is quite evident that cluster 6 locations are not ideal, therefore in establishing the new pizzeria stake holders should not look to open in the heart of the city.

- Targeting busy routes that are also residential might be a good strategy and from the quadrant visualization in cluster 0 Moratuwa or Dalugama are good locations.

- Cluster 4 containing areas booming with both business and are residential like Battaramulla and Narahenpita seems to indicate a very open market for pizza especially in Battaramulla where the presence of food related venues is very high with little presence of pizza venues.

6. Conclusion

The purpose of this project was to use the landscape of venue distribution across different suburbs of Colombo to identify the optimal location to open a pizzeria. From all the venue categories found it was determined the best variables to use were one's indicative of busy areas, with venues that would attract spending for entertainment leisure etc and spending on food (Presence of restaurants).

After this the data was clustered and similarities within clusters were identified based on location and preexisting knowledge of the city. Proxy variables were created for demand and supply i.e. the food and pizza variables and visualizations were used to identify suburbs that would ensure best entry for the new pizzeria.

Finally based on the cluster and the calculated metrics it seemed as though in terms of a suburb Battaramulla might be the ideal location for this, although suburbs like Moratuwa could also still be considered; with this study could safely move to the next step of where in Battaramulla this needs to be to increase exposure to people looking to eat something quick or even stop by for a meal.