Coursera Capstone

IBM Applied Data Science Capstone

Opening a new restaurant in Colombo, Sri Lanka

By: Ashane Fernando

December 2019

Introduction

Background

Colombo is the de-facto - though no longer the official - capital of Sri Lanka. Nevertheless, it is still a burgeoning commercial hub with most of the economic activity in Sri Lanka being centered in this region. Over the last couple of years, there has been a dramatic rise in foreign investment in Colombo, due to the Sri Lankan government's initiative to transform Colombo into a "South Asian Las Vegas"

As a consequence, this has resulted in a rapid boom in tourism in Sri Lanka which in turn has triggered a meteoric rise in commercial development, with many restaurants, entertainment venues such as nightclubs, and hotels popping up around Colombo almost overnight. While Sri Lanka has always been a popular tourist destination, in the past it was mainly due to scenic beauty and natural attractions, and thus was mainly focused in the rural areas outside of the capital. However, at present, there is a large influx of tourists who come to Colombo solely to embrace the flourishing nightlife and sample a wide variety of cuisines from all over the globe.

While there are already many restaurants in Colombo which attract a substantial clientele, they are not sufficient to cater to the number of foreign tourists flooding into the city, in addition to the locals who already patronize these establishments. Thus, there is ample opportunity for a new restaurant to thrive, provided that they position themselves well; both in terms of location as well as business strategy.

Business Problem

There is a saying among marketing professionals that three things are important in business: location, location and location, and when it comes to selecting a location for a restaurant this still holds true. Even within the city of Colombo, not all neighborhoods are the same, and the location you select for your restaurant can determine its a success or failure.

This project aims to apply the data science methodology, data analysis, and machine learning techniques to address the business question; if someone were to open a new restaurant within Colombo, where would you recommend that they open it?

Target Audience

This project is particularly geared towards local or foreign investors or restauranteurs who would be interested in opening or investing in new restaurants in the Colombo area.

For instance, earlier this year, in recognition of the growth of the restaurant sector in Sri Lanka, several of the leading restauranteurs in Sri Lanka banded together to launch the Colombo City Restaurant Collective (CCRC) in order to act as a shared voice for restauranteurs in Colombo and to facilitate coordination and collaboration with government authorities to guide the development of the restaurant industry[1].

Thus, this project may be of interest to current restauranteurs as well, who may wish to expand further.

Data

When considering the location for a restaurant, we need to consider the other venues that surround it. In addition to the presence of other restaurants, we should also pay attention to the number of hotels and entertainment venues that are in close proximity. A large concentration of hotels would mean that there are more tourists in the area who are likely to visit your restaurant. In the same way, entertainment venues, such as bars, nightclubs and theatres would also attract more people to the area which once again translates to more potential business for your restaurant.

Thus, the primary focus would be on gathering information on the hotels, restaurants and entertainment venues within the neighborhoods of Colombo.

Data requirement

In order to address this problem, we need to collect the following data:

- The list of neighborhoods within the Colombo city limits which we could use to divide the city into distinct locales.
- Latitude and longitude values which we could use to define the locations of each of these neighborhoods.
- Venue data for each of the neighborhoods, with particular emphasis on the hotels, restaurants, and entertainment venues. We need to obtain the location coordinates as well as the type of venue.

Data Sources and Data Collection

Sri Lanka is a small country, with a land area of only 65,000km, and therefore it is not as easy to find structured geographical and location data as it would be in the case of larger and more developed countries. Nevertheless, there are still many tools we could leverage to obtain the information we need.

The Wikipedia page https://en.wikipedia.org/wiki/Postal_codes_in_Sri_Lanka contains the list of post codes for the whole of Sri Lanka. We can use web scraping techniques, such as the Beautiful Soup Python library or directly use the Pandas library to read the html directly from the website in order to obtain the list of neighborhoods in Colombo.

We can then use the Python Geocoder package, which is free to use, to obtain the latitude and longitude coordinates of each of these neighborhoods.

Next, we can use the Foursquare Developer API to obtain the venue data for each of the neighborhoods. Foursquare is one of the largest location data platforms in the world and boasts a database for over 105 million places. Its API is used by many other technology giants such as Uber and AirBnB. We can leverage the Foursquare API ourselves (which is available for free, with some limitations) to gain information on the venues present in each of the platforms, including but not limited to their geographical coordinates and the type of venue.

Methodology

Obtaining the neighborhood data

Our first order of business is to obtain the list of neighborhoods in Colombo along with their location coordinates. Therefore we need to scrape the Wikipedia page https://en.wikipedia.org/wiki/Postal_codes_in_Sri_Lanka to extract the list of neighborhoods.

When it comes to web scraping, the Beautiful Soup library may have a better performance overall compared to Pandas since it is tailor-made for that purpose, and would have been the go-to solution if the webpage was a complex one with embedded images and links. However, in this case since the Wikipedia webpage is simple and contains mainly text, I decided to simply use the Pandas library, instead of installing a new package, to directly read the html from the webpage since there wouldn't be a substantial difference between the results.

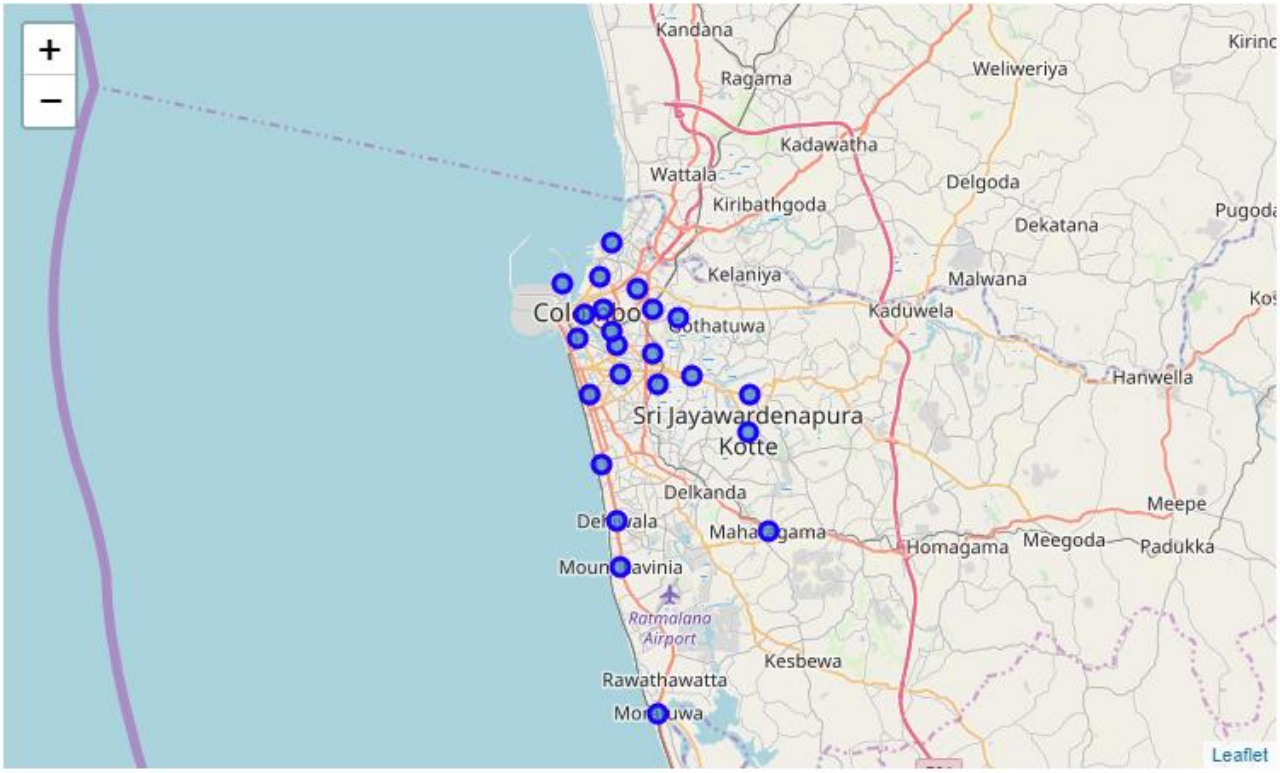
Once the list was obtained, I performed some preprocessing and data cleaning on the scraped data in order to get it into a workable format. Since the number of neighborhoods is relatively low - just 22 neighborhoods - it was more expedient to perform the data cleaning manually.

Now that I had the list of neighborhoods, I need to find the coordinates for each of the neighborhoods. I elected to use the Geocoder API (not to be confused with the Google Geocoding API which is no longer free) for this purpose. Although the Geocoder API is a little unstable and does not always provide the required information on the first request, by using a loop to repeat the request until a response was received, I was able to obtain the location data in the form of latitude and longitude values for all the neighborhoods.

The dataframe of the neighborhood data is shown below.

	Postal Code	Neighborhood	Latitude	Longitude
0	10120	Battaramulla	6.902181	79.919578
1	00400	Bambalapitiya	6.901825	79.854683
2	00600	Wellawatte	6.874128	79.859332
3	00300	Kollupitiya	6.921812	79.865561
4	00500	Narahenpita	6.905727	79.882130
5	00800	Borella	6.917932	79.880256
6	00700	Cinnamon Gardens	6.910335	79.866994
7	00900	Dematagoda	6.936130	79.880049
8	00100	Fort	6.946338	79.843968
9	01200	Hulftsdorp	6.935853	79.860260
10	01300	Kotahena	6.948985	79.858923
11	01000	Maradana	6.926799	79.863879
12	01400	Grandpass	6.943993	79.873717
13	01500	Mutwal	6.962534	79.864048
14	01100	Pettah	6.934129	79.852892
15	00200	Slave Island	6.924657	79.850273
16	10100	Sri Jayawardenepura	6.886693	79.918738
17	10107	Rajagiriya	6.909504	79.896218
18	10350	Dehiwala	6.851279	79.865977
19	10600	Kolonnawa	6.932625	79.890314
20	10280	Maharagama	6.847278	79.926608
21	10400	Moratuwa	6.774682	79.882610
22	10370	Mount Lavinia	6.832936	79.867410

In order to visualize the locations of each of the neighborhoods, I also generated a map using the neighborhood coordinates and the Folium library for Python. This also allows us to visually confirm whether the proper coordinates have been taken for all the neighborhoods.



Obtaining Venue Data using the Foursquare API

Next, I used the Foursquare API to search for the top 100 venues that were within a radius of 1km of the neighborhood center.

This range of 1km was decided upon by first checking the distance between adjacent neighborhoods. On average, the distance between two adjacent neighborhoods is under 2km (with Moratuwa being the sole outlier with a distance of 6.6km), hence the choice of 1km would ensure that the entire city is covered, while minimizing the overlap between search zones.

In order to do this, I needed to make an API call to Foursquare with the geographical coordinates (latitude and longitude) of the neighborhood and the **venue.explore** endpoint. Foursquare would then return information on the nearby venues in JSON format from which I was able to extract venue information such as the venue name, venue category (i.e. Italian restaurant, spa etc.) and the geographical coordinates of the venue.

Grouping neighborhoods by venues

I then used the obtained data to find out what the ten most common venues in each neighborhood would be. This was achieved by grouping the list of venues by neighborhood and taking the mean of the frequency of occurrence of each type of venue. One-hot encoding was used in order to convert the categorical variable that is “venue category” into a numerical variable.

This was done in an attempt to profile the neighborhoods, which didn’t yield expected results. However, what I discovered is that in most, if not all the neighborhoods, restaurants are the most common type of venue. Additionally, a restaurant any kind invariably appears within the top three most common venues in each of the neighborhoods.

Since our interest is specifically in density of hotels, restaurants, and entertainment venues in each of the neighborhoods, I needed to filter the grouped data to include only these three types of venues. I used keywords in the venue category names to classify the venues into “Hotels”, “Restaurants” and “Entertainment” while dropping the venues that didn’t fall into any of these three categories. As before, I grouped the list of venues by neighborhood and to obtained the relative proportions of occurrence of each type of venue in each neighborhood.

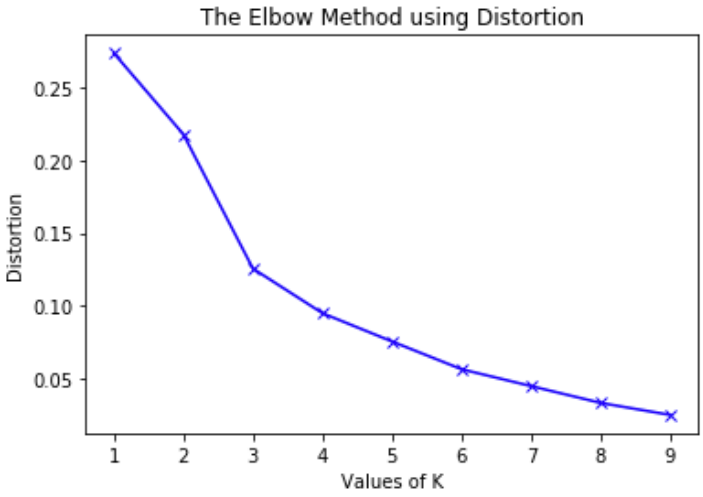
	Neighborhood	Entertainment	Hotel	Restaurant
0	Bambalapitiya	0.138462	0.123077	0.738462
1	Battaramulla	0.000000	0.000000	1.000000
2	Borella	0.428571	0.071429	0.500000
3	Cinnamon Gardens	0.311475	0.065574	0.622951
4	Dehiwala	0.111111	0.000000	0.888889
5	Dematagoda	0.000000	0.000000	1.000000
6	Fort	0.500000	0.500000	0.000000
7	Grandpass	0.500000	0.000000	0.500000
8	Hulftsdorp	0.125000	0.125000	0.750000
9	Kollupitiya	0.260870	0.086957	0.652174
10	Kolonnawa	0.000000	0.000000	1.000000
11	Kotahena	0.333333	0.000000	0.666667
12	Maharagama	0.000000	0.000000	1.000000
13	Maradana	0.318182	0.045455	0.636364
14	Moratuwa	0.000000	0.000000	1.000000
15	Mount Lavinia	0.176471	0.176471	0.647059
16	Mutwal	1.000000	0.000000	0.000000
17	Narahenpita	0.285714	0.000000	0.714286
18	Pettah	0.238095	0.095238	0.666667
19	Rajagiriya	0.058824	0.000000	0.941176
20	Slave Island	0.188406	0.173913	0.637681
21	Sri Jayawardenepura	0.222222	0.000000	0.777778
22	Wellawatte	0.116279	0.116279	0.767442

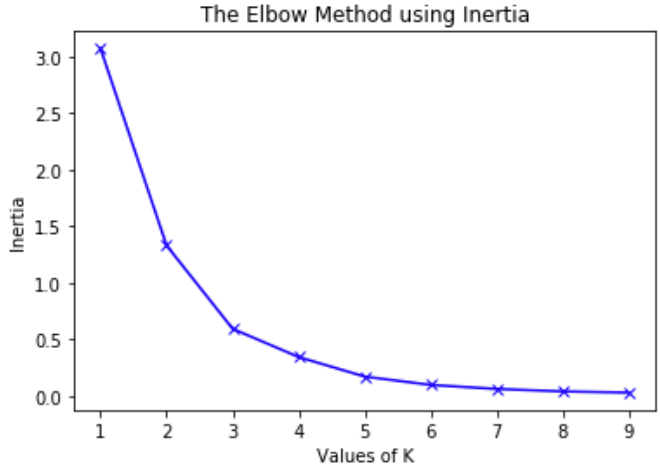
Clustering Neighborhoods based on the proportion of hotels, restaurants and entertainment venues.

In order to cluster the neighborhoods, I elected to use K-means clustering. K-means is a type of unsupervised learning algorithm that is used to cluster datapoints based on common features. This is one of the simpler algorithms and is computationally inexpensive, making it a good fit for our particular application.

Before I proceeded with the K-means clustering, I needed to determine the optimal value for k (number of clusters). For this, I used the elbow method, which plots the cost function produced by different values of k and uses the point where the curve begins to become linear as the ideal value for k.

Two metrics were used for the cost function; inertia and distortion, and plotted it against the k-value (plots given below). Based on the plots, I decided to go with a k value of 3.





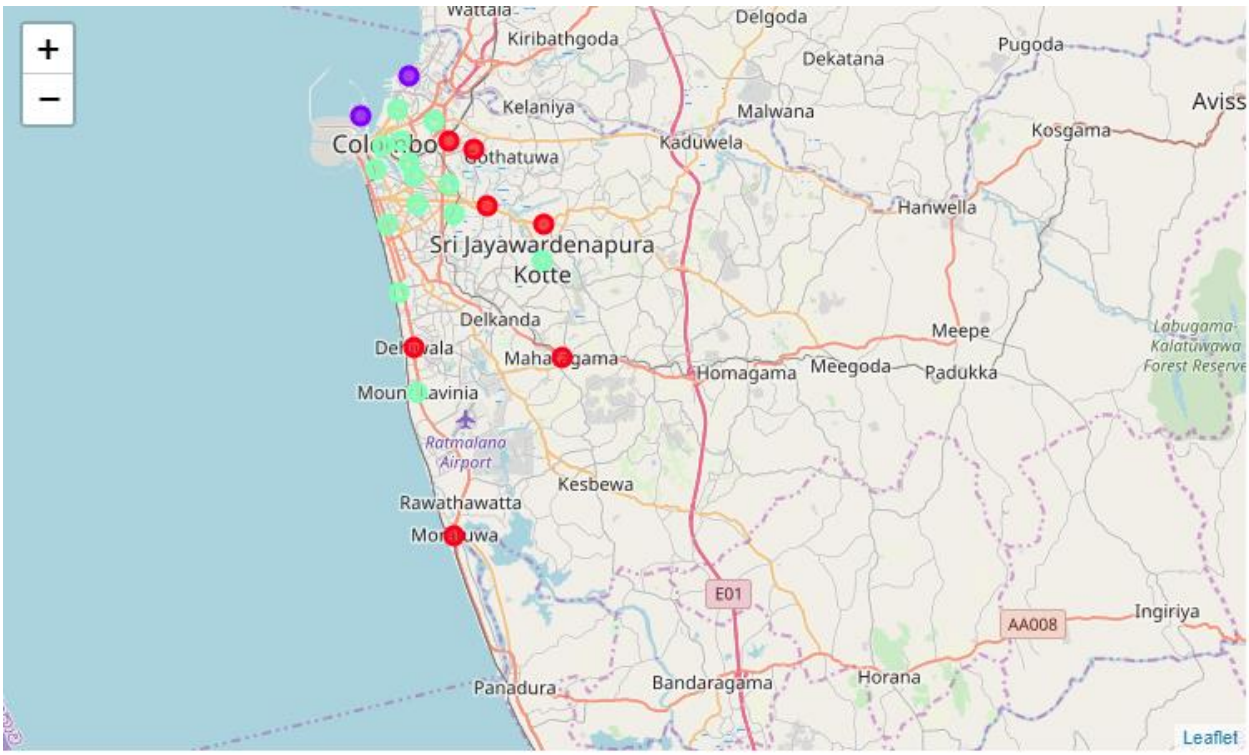
So, I ran the K-means algorithm with the number of clusters set to 3, and clustered the neighborhoods based on the features “Hotels”, “Restaurants” and “Entertainment”

I then marked the neighborhoods once more on the map (generated using Folium), but this time the points are color-coded based on the cluster to which it belongs. This will be discussed in the next section.

Results

The results of the k-means clustering are given below.

- Red: Cluster 0
- Purple: Cluster 1
- Lime green: Cluster 2



Cluster 0:

	Neighborhood	Entertainment	Hotel	Restaurant	Cluster Labels
1	Battaramulla	0.000000	0.0	1.000000	0
19	Rajagiriya	0.058824	0.0	0.941176	0
4	Dehiwala	0.111111	0.0	0.888889	0
5	Dematagoda	0.000000	0.0	1.000000	0
10	Kolonnawa	0.000000	0.0	1.000000	0
12	Maharagama	0.000000	0.0	1.000000	0
14	Moratuwa	0.000000	0.0	1.000000	0

Cluster 1:

	Neighborhood	Entertainment	Hotel	Restaurant	Cluster Labels
6	Fort	0.5	0.5	0.0	1
16	Mutwal	1.0	0.0	0.0	1

Cluster 2:

	Neighborhood	Entertainment	Hotel	Restaurant	Cluster Labels
0	Bambalapitiya	0.138462	0.123077	0.738462	2
20	Slave Island	0.188406	0.173913	0.637681	2
18	Pettah	0.238095	0.095238	0.666667	2
17	Narahenpita	0.285714	0.000000	0.714286	2
15	Mount Lavinia	0.176471	0.176471	0.647059	2
11	Kotahena	0.333333	0.000000	0.666667	2
21	Sri Jayawardenepura	0.222222	0.000000	0.777778	2
9	Kollupitiya	0.260870	0.086957	0.652174	2
8	Hulftsdorp	0.125000	0.125000	0.750000	2
7	Grandpass	0.500000	0.000000	0.500000	2
3	Cinnamon Gardens	0.311475	0.065574	0.622951	2
2	Borella	0.428571	0.071429	0.500000	2
13	Maradana	0.318182	0.045455	0.636364	2
22	Wellawatte	0.116279	0.116279	0.767442	2

Observations

- In the neighborhoods in Cluster 0, we see that there are little to no hotels or entertainment venues, compared to the number of restaurants in the area.
- In the neighborhoods in Cluster 1, there are no restaurants to be found, but hotels and entertainment venues are present.
- In the neighborhoods in Cluster 2, although there is still a high proportion of restaurants to other venues, it still has a considerable number of hotels and entertainment venues

The significance of these observations will be discussed in the next section.

Discussion

As mentioned before, in the neighborhoods belonging to Cluster 0, we see that there are little to no hotels or entertainment venues to be found, while there is a significant number of restaurants in the area. Therefore, it would not be advisable to locate a new restaurant there, as it would face heavy competition. It may be interesting to investigate why there is such a high concentration of restaurants in these areas, despite the lack of hotels and entertainment venues, but that is a task for another day.

At first glance, the neighborhoods in Cluster 1 appear to be the ideal place for a restaurant since they contain hotels and entertainment areas, but no restaurants. However, upon closer inspection, I found that out of the two neighborhoods in this cluster, one has only a single hotel and an entertainment venue, while the other has just a single hotel. Thus, these neighborhoods would not be suitable either.

Finally, we come to Cluster 2. While these neighborhoods also have a high proportion of restaurants in comparison to other venues, they still contain a considerable number of hotels and entertainment venues as well. Thus, if we were to start a restaurant, it should be within these neighborhoods.

Limitations and Suggestions for Future Research

In this project, we only consider one factor, i.e. the types of venues around the location, when deciding the location of the restaurant. However, there are many other factors, such as the population, the income levels of residents, the local crime rate that may affect the planned location of the restaurant. This is highlighted particularly in the case of neighborhood cluster 0 which shows a high concentration of restaurants although there are little to none other venues in the vicinity. However, to my best knowledge, this data is not available to the neighborhood level. If such data were to become available, then the model could be further refined.

Conclusion

In the course of this project, I have identified a business problem; namely the determination of the best location for a new restaurant; and followed the data science methodology by specifying the data required, extracting and preparing the data, selecting and implementing a suitable model (K-means clustering) and finally providing an answer to the business problem in the form of a recommendation to the prospective stakeholders. Therefore, to answer the business question posed in the introduction section, the best location to open a new restaurant would be in the neighborhoods in Cluster 2.

References

[1] http://www.ft.lk/business/Leading-restaurateurs-unite-to-serve-up-Colombo-City-Restaurant-Collective/34-686048