The Battle of Neighborhoods

INTRODUCTION

Sri Lanka is one of the most attractive tourist destinations in the world. Many people from all around the world travel to Sri Lanka to explore Sri Lanka for its tourist attractions.

Most of the tourists coming to Sri Lanka is trying to experience the authentic Sri Lankan food. Since most of the time tourists travel all-around the country if they can find areas with more authentic food, they can plan the trip considering that factor as well.

Thus, the goal I want to achieve with this project is to recommend to tourists visiting Sri Lanka. The areas which has the most probability of finding authentic Sri Lankan food.

DATA

For the project to make sense we need to come up with areas and the restaurants that can be found in the areas. Since the tourists take a broad look at Sri Lanka the plan is to do the analysis based on district level of Sri Lanka.

Data Source 1:

Wikipedia – District Names of Sri Lanka, Province, Population, Area size

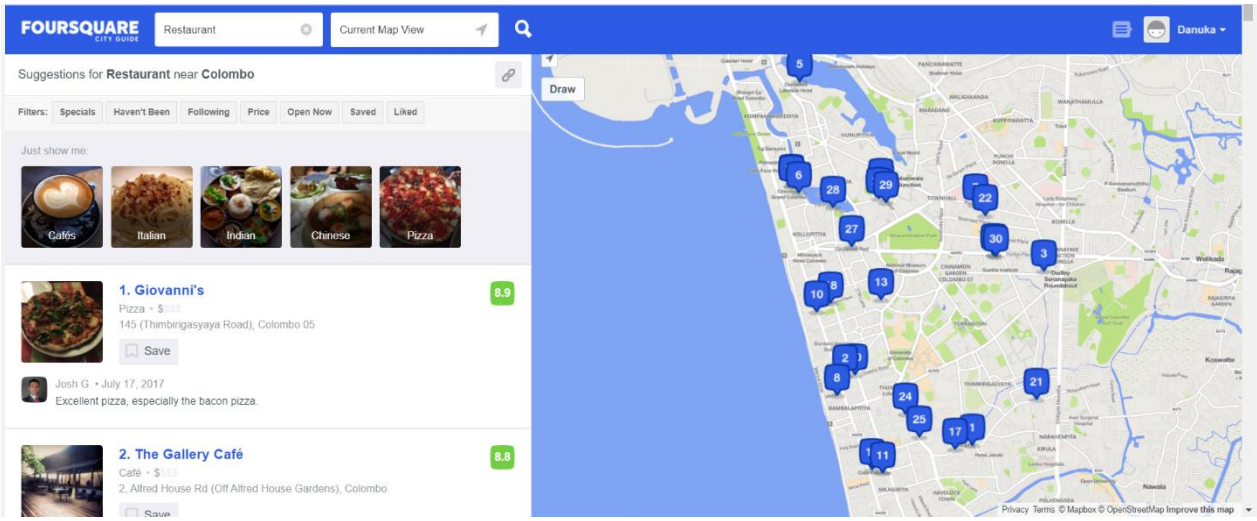
In the page, https://en.wikipedia.org/wiki/Districts_of_Sri_Lanka there is a table that contain the information about population, area, density, etc in district level of Sri Lanka.

Data Source 2:

Foursquare – Venues in each district, Restaurants in each District

The Foursquare APIs can be used to find location about places near a location.

Here I am going to utilize this in district level



An example search on foursquare for a restaurant in Sri Lanka

Methodology

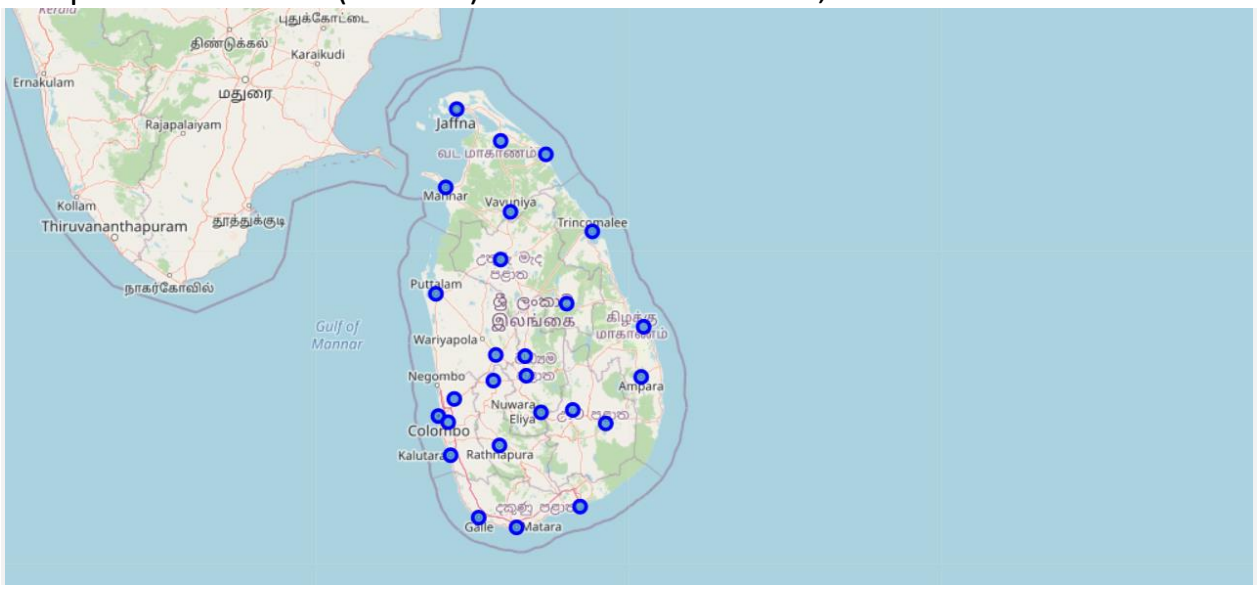
Main objective of the project is to give an insight for the tourists about locations, so as the first step the location areas were obtained by scraping the Wikipedia web page. **pandas HTML read function was used for this.**

	District	Area map	Province	Districtcapital	Landareain km2 (mi2)[24]	Inlandwaterareain km2 (mi2)[24]	Totalareain km2 (mi2)[24]	Population(2012)[25]	Populationdensityper km2(per mi2)[a]
0	Ampara	NaN	Eastern	Ampara	4,222 (1,630)	193 (75)	4,415 (1,705)	649402	154 (400)
1	Anuradhapura	NaN	North Central	Anuradhapura	6,664 (2,573)	515 (199)	7,179 (2,772)	860575	129 (330)
2	Badulla	NaN	Uva	Badulla	2,827 (1,092)	34 (13)	2,861 (1,105)	815405	288 (750)
3	Batticaloa	NaN	Eastern	Batticaloa	2,610 (1,010)	244 (94)	2,854 (1,102)	526567	202 (520)
4	Colombo	NaN	Western	Colombo	676 (261)	23 (8.9)	699 (270)	2324349	3,438 (8,900)

Then geopy package with Nominatim was used to get the geographical coordinates of each district and the data were append as Latitude and Longitude to the data frame.

	Neighborhood	Area map	Province	Districtcapital	Landareain km2 (mi2) [24]	Inlandwaterareain km2 (mi2) [24]	Totalareain km2 (mi2) [24]	Population(2012) [25]	Populationdensityper km2(per mi2) [a]	Latitude	Longitude
0	SriLanka-Ampara	NaN	Eastern	Ampara	4,222 (1,630)	193 (75)	4,415 (1,705)	649402	154 (400)	7.291123	81.672395
1	SriLanka-Anuradhapura	NaN	North Central	Anuradhapura	6,664 (2,573)	515 (199)	7,179 (2,772)	860575	129 (330)	8.334985	80.410610
2	SriLanka-Badulla	NaN	Uva	Badulla	2,827 (1,092)	34 (13)	2,861 (1,105)	815405	288 (750)	6.989820	81.056943
3	SriLanka-Batticaloa	NaN	Eastern	Batticaloa	2,610 (1,010)	244 (94)	2,854 (1,102)	526567	202 (520)	7.735603	81.694196
4	SriLanka-Colombo	NaN	Western	Colombo	676 (261)	23 (8.9)	699 (270)	2324349	3,438 (8,900)	6.934997	79.853846
5	SriLanka-Galle	NaN	Southern	Galle	1,617 (624)	35 (14)	1,652 (638)	1063334	658 (1,700)	6.032814	80.214955
6	SriLanka-Gampaha	NaN	Western	Gampaha	1,341 (518)	46 (18)	1,387 (536)	2304833	1,719 (4,450)	7.092560	79.995140
7	SriLanka-Hambantota	NaN	Southern	Hambantota	2,496 (964)	113 (44)	2,609 (1,007)	599903	240 (620)	6.124913	81.124256
8	SriLanka-Jaffna	NaN	Northern	Jaffna	929 (359)	96 (37)	1,025 (396)	583882	629 (1,630)	9.665093	80.009303
9	SriLanka-Kalutara	NaN	Western	Kalutara	1,576 (608)	22 (8.5)	1,598 (617)	1221948	775 (2,010)	6.583522	79.961251

The point of interests (Districts)can be found as below,

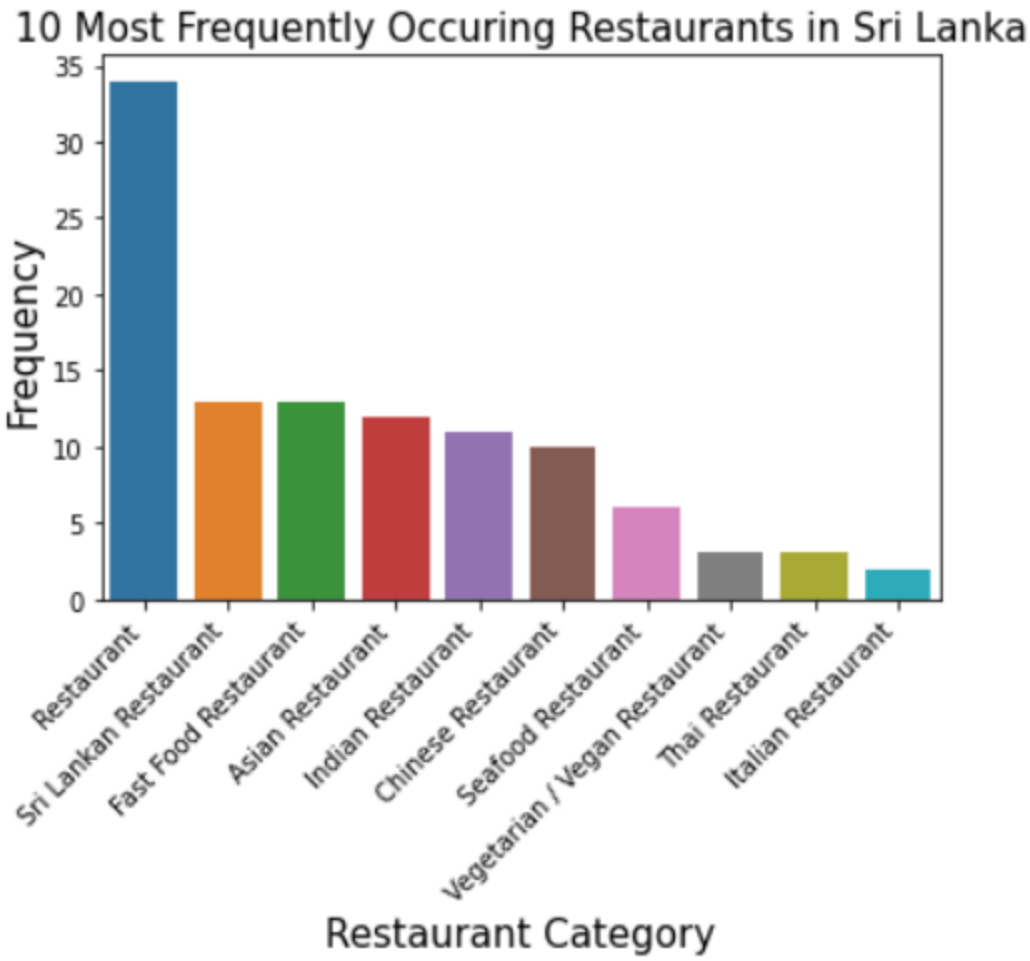


Sri Lanka map with district centers pointed

In order to identify the restaurants located in these areas the Foursquare APIs were used. The search for a maximum of 1000 venues in a 2km radius for each district center was done, which was resulted in 570 venues from 25 districts. In 570 venues there are 118 restaurants. They could be further categorize as restaurants which serve 19 different types of foods such as Sri Lanka, Chinese, Indian and Seafood.

The distribution of the 10 most frequently observed restaurants in the city out of 19 can be visualized as follows,

	Venue_Category	Frequency
0	Restaurant	34
1	Sri Lankan Restaurant	13
2	Fast Food Restaurant	13
3	Asian Restaurant	12
4	Indian Restaurant	11
5	Chinese Restaurant	10
6	Seafood Restaurant	6
7	Vegetarian / Vegan Restaurant	3
8	Thai Restaurant	3
9	Italian Restaurant	2



The Restaurant category was assumed to be Sri Lankan restaurants since the local restaurants usually cater authentic Sri Lankan food,

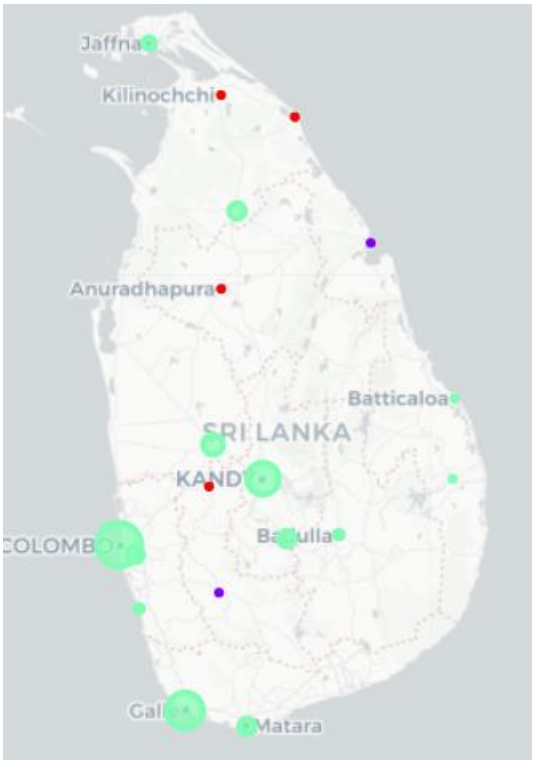
Then one-hot encoding was used and then the clustering was done on the data. With k=3.

RESULTS

	Neighborhood	Area map	Province	Districtcapital	Landareain km2 (mi2) [24]	Inlandwaterareain km2 (mi2)[24]	Totalareain km2 (mi2) [24]	Population(2012) [25]	Populationdensityper km2(per mi2)[a]	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th M Comn Ve
0	SriLanka-Ampara	NaN	Eastern	Ampara	4,222 (1,630)	193 (75)	4,415 (1,705)	649402	154 (400)	7.291123	81.672395	2.0	Asian Restaurant	Vegetarian / Vegan Restaurant	Halal Restaurant	Chinese Restaurant	Dump Restau
1	SriLanka-Anuradhapura	NaN	North Central	Anuradhapura	6,664 (2,573)	515 (199)	7,179 (2,772)	860575	129 (330)	8.334985	80.410610	0.0	Sri Lankan Restaurant	Asian Restaurant	Vegetarian / Vegan Restaurant	Halal Restaurant	Chin Restau
2	SriLanka-Badulla	NaN	Uva	Badulla	2,827 (1,092)	34 (13)	2,861 (1,105)	815405	288 (750)	6.989820	81.056943	2.0	Restaurant	Asian Restaurant	Fast Food Restaurant	Vegetarian / Vegan Restaurant	H Restau
3	SriLanka-Batticaloa	NaN	Eastern	Batticaloa	2,610 (1,010)	244 (94)	2,854 (1,102)	526567	202 (520)	7.735603	81.694196	2.0	Indian Restaurant	Restaurant	Halal Restaurant	Asian Restaurant	Chin Restau
4	SriLanka-Colombo	NaN	Western	Colombo	676 (261)	23 (8.9)	699 (270)	2324349	3,438 (8,900)	6.934997	79.853846	2.0	Seafood Restaurant	Indian Restaurant	Restaurant	Chinese Restaurant	Fast F Restau

Cluster labels for districts with most common types of restaurants

In this table, we see that cluster labels assigned by the k-means clustering algorithm. The clusters are visualized on the map as follows,



On this map, it can be observed 3 different colors of points on district centers. Each color represents a different cluster. Now we will inspect these clusters in more detail and try to give a name for each one.

Cluster 1

	Area map	Inlandwaterareain km2 (mi2)[24]	Totalareain km2 (mi2)[24]	Population(2012)[25]	Populationdensityper km2(per mi2)[a]	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	NaN	515 (199)	7,179 (2,772)	860575	129 (330)	8.334985	80.410610	0.0	Sri Lankan Restaurant	Asian Restaurant	Vegetarian / Vegan Restaurant	Halal Restaurant	Chinese Restaurant
11	NaN	8 (3.1)	1,693 (654)	840648	499 (1,290)	7.253201	80.345413	0.0	Sri Lankan Restaurant	Vegetarian / Vegan Restaurant	Halal Restaurant	Asian Restaurant	Chinese Restaurant
12	NaN	74 (29)	1,279 (494)	113510	94 (240)	9.384007	80.408722	0.0	Vegetarian / Vegan Restaurant	Sri Lankan Restaurant	Halal Restaurant	Asian Restaurant	Chinese Restaurant
18	NaN	202 (78)	2,617 (1,010)	92238	38 (98)	9.269853	80.814535	0.0	Sri Lankan Restaurant	Vegetarian / Vegan Restaurant	Halal Restaurant	Asian Restaurant	Chinese Restaurant

Most of the districts labeled as cluster 1(0 in code and table) have Sri Lankan Restaurants as the most common restaurant type. So we can assume that this cluster represents **Sri Lanka Restaurants**.

Cluster 2

	Area map	Inlandwaterareain km2 (mi2)[24]	Totalareain km2 (mi2)[24]	Population(2012)[25]	Populationdensityper km2(per mi2)[a]	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
22	NaN	39 (15)	3,275 (1,264)	1088007	336 (870)	6.680369	80.402298	1.0	Fast Food Restaurant	Vegetarian / Vegan Restaurant	Halal Restaurant	Asian Restaurant	Chinese Restaurant
23	NaN	198 (76)	2,727 (1,053)	379541	150 (390)	8.576425	81.234495	1.0	Fast Food Restaurant	Vegetarian / Vegan Restaurant	Halal Restaurant	Asian Restaurant	Chinese Restaurant

All of the districts labeled as cluster 2(1 in code and table) have Fast Food Restaurants as the most common restaurant type. So we can assume that this cluster represents **Fast Food Restaurants**.

Cluster 3

	Area map	Inlandwaterareain km2 (mi2)[24]	Totalareain km2 (mi2)[24]	Population(2012) [25]	Populationdensityper km2(per mi2)[a]	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	NaN	193 (75)	4,415 (1,705)	649402	154 (400)	7.291123	81.672395	2.0	Asian Restaurant	Vegetarian / Vegan Restaurant	Halal Restaurant	Chinese Restaurant	Dumpling Restaurant
2	NaN	34 (13)	2,861 (1,105)	815405	288 (750)	6.989820	81.056943	2.0	Restaurant	Asian Restaurant	Fast Food Restaurant	Vegetarian / Vegan Restaurant	Halal Restaurant
3	NaN	244 (94)	2,854 (1,102)	526567	202 (520)	7.735603	81.694196	2.0	Indian Restaurant	Restaurant	Halal Restaurant	Asian Restaurant	Chinese Restaurant
4	NaN	23 (8.9)	699 (270)	2324349	3,438 (8,900)	6.934997	79.853846	2.0	Seafood Restaurant	Indian Restaurant	Restaurant	Chinese Restaurant	Fast Food Restaurant
5	NaN	35 (14)	1,652 (638)	1063334	658 (1,700)	6.032814	80.214955	2.0	Restaurant	Sri Lankan Restaurant	Indian Restaurant	Asian Restaurant	Chinese Restaurant
8	NaN	96 (37)	1,025 (396)	583882	629 (1,630)	9.665093	80.009303	2.0	Vegetarian / Vegan Restaurant	Asian Restaurant	Indian Restaurant	Sri Lankan Restaurant	American Restaurant
9	NaN	22 (8.5)	1,598 (617)	1221948	775 (2,010)	6.583522	79.961251	2.0	Restaurant	Seafood Restaurant	Vegetarian / Vegan Restaurant	German Restaurant	Asian Restaurant
10	NaN	23 (8.9)	1,940 (750)	1375382	716 (1,850)	7.293092	80.635077	2.0	Restaurant	Fast Food Restaurant	Indian Restaurant	Chinese Restaurant	Sri Lankan Restaurant
13	NaN	192 (74)	4,816 (1,859)	1618465	350 (910)	7.487046	80.364908	2.0	Restaurant	Sri Lankan Restaurant	Asian Restaurant	Chinese Restaurant	Fast Food Restaurant
16	NaN	13 (5.0)	1,283 (495)	814048	641 (1,660)	5.947822	80.548292	2.0	Restaurant	Fast Food Restaurant	Asian Restaurant	Vegetarian / Vegan Restaurant	Halal Restaurant
19	NaN	35 (14)	1,741 (672)	711644	417 (1,080)	6.973886	80.767127	2.0	Sri Lankan Restaurant	Restaurant	Indian Restaurant	Italian Restaurant	Thai Restaurant
24	NaN	106 (41)	1,967 (759)	172115	92 (240)	8.759352	80.500078	2.0	Restaurant	Asian Restaurant	Vegetarian / Vegan Restaurant	Indian Restaurant	Seafood Restaurant
25	NaN	2,905 (1,122)	65,610 (25,330)	20359439	325 (840)	6.877829	79.934959	2.0	Chinese Restaurant	Asian Restaurant	Restaurant	Thai Restaurant	Vegetarian / Vegan Restaurant

Most of the restaurants in this category under first common venue is labeled as “Restaurants”. With this fact, we can consider these areas have common type of restaurants rather than special types.

So the conclusion can be obtained that in areas of cluster1 we can guarantee that there a authentic Sri Lankan restaurants where as in areas in cluster3 there is a good probability of having authentic Sri Lankan restaurants where as in other areas we cannot make sure. So this result can be easily used for the planning stage of the trip for tourists.

DISCUSSION

In this project, I was able to utilize most of the concepts learned like data cleaning, scraping, handling, analysis, and getting results with machine learning algorithms.

Although the amount of data in the Foursquare API was limited for Sri Lanka. If the data was enough it was possible to go another level deep and do the analysis in sub district level which will be furthermore convenient for tour planning.

CONCLUSION

Overall, the target according to the problem declaration was achieved successfully, this now can be used to get a input for tour planning for tourists. As improvements if we have more data, we can add a hierarchical clustering methodology for more accurate clustering.

If the data limitation was addressed using another source with more samples and develop this as a web app for result showing I think this project will have a real-world application and value as well.

I hope you enjoyed my capstone project!.