### **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

Graphical user interface

Description automatically generated

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.

A picture containing graphical user interface

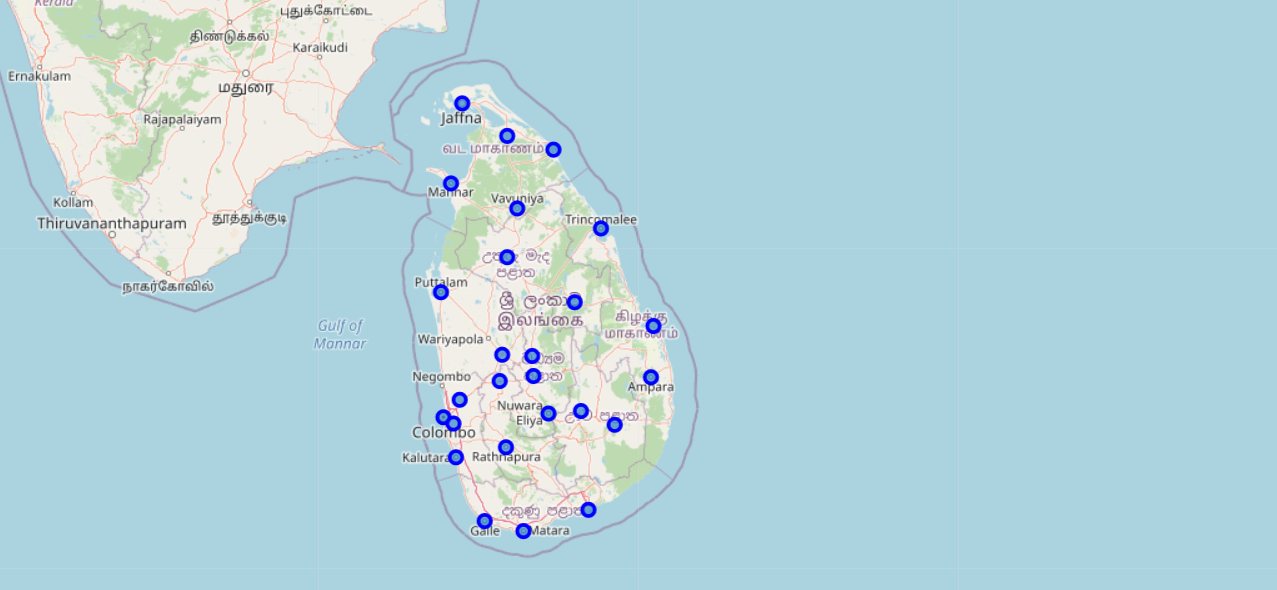
Description automatically generated

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.

A picture containing table

Description automatically generated

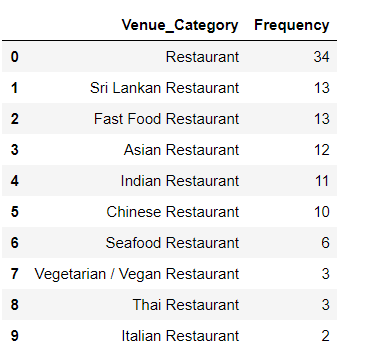
The point of interests (Districts)can be found as below,

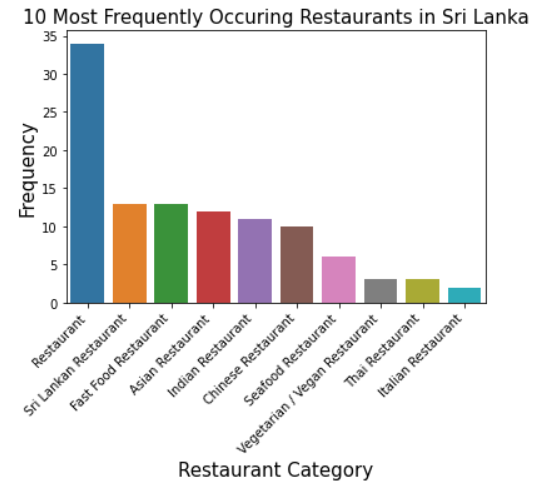


Sri Lanka map with district centers pointed

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,





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.

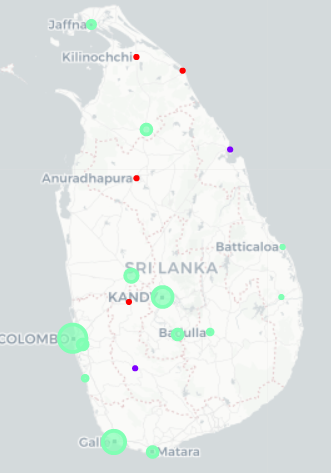
RESULTS

Graphical user interface, text, application

Description automatically generated

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.

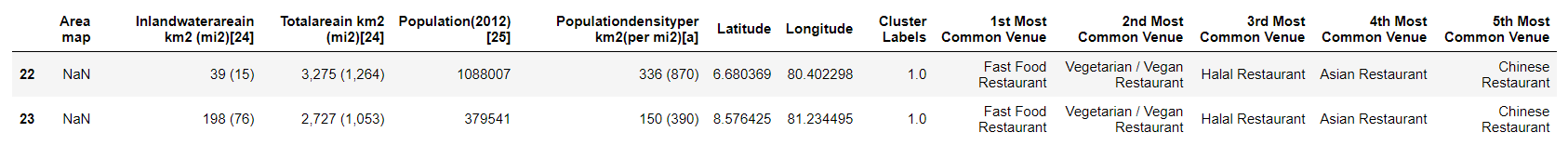
Cluster 1

A picture containing text

Description automatically generated

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



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

Graphical user interface

Description automatically generated

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!.