IBM APPLIED DATA SCIENCE CAPSTONE

Recommending a perfect Business to start at a Tourist Hotspot

Abstract

This report contains details of the project which was completed as a part IBM Applied Data Science Capstone.

Anjana Dodampe
ahmdodampe@hotmail.com

1. Introduction

1.1 Background

Tourism, the act and process of spending time away from home in pursuit of recreation, relaxation, and pleasure has been a trending sector across the globe. Human as a social animal, always fonds of travelling, exploring new adventures. No matter which country you're from, you can always come across a group of people who always like travelling places. Tourism plays a significant role in developing economy of a country and it brings a particular country to a prominent place in global standing.

Tourism industry is important in many aspects since it creates demand and growth for many more industries. It plays an important role in generating more employments, revenues and contributing in empowering livelihood of many locals. Sri Lanka, the pearl of the Indian Ocean is a country that is immensely benefited from tourism industry.

1.2 Problem

All the benefits of tourism tend to reflect on the employment opportunities that it provides to the people of that country. The objective of this project is to analyse tourist places across districts of Sri Lanka and try to recommend the best location where they can open a business to make the best use of the opportunity.

The target audience for this project includes the people who are interested in opening a business that associates with tourism industry. This also recommends tourists, tourist attraction hotspots in a particular district in Sri Lanka.

2. Data acquisition and cleaning

2.1 Data sources

The Wikipedia page https://en.wikipedia.org/wiki/Districts_of_Sri_Lanka is the main source of data that is being used to obtain all the districts of Sri Lanka. I have used BeautifulSoup library to scrape required data from the above link. Then I've extracted the required tables from the above page and converted it into a pandas data frame.

2.2 Data cleaning

Data scraped from multiple sources were combined into one table. I ran a code to find whether there are NaN values in the dataset. As well as there were many unwanted columns in the data frame that I initially made after scraping data from the Website. So, I dropped unwanted columns and renamed columns for ease and created a new data frame. With the help of geopy library, I could get the latitude and longitude details related to particular district and merged it with the data frame I created lately. The final data set after cleaning consists of single data frame with 6 columns containing Province, latitudes & longitudes of a particular district. Other columns like Population, Population density which also been scraped from the website which can be used for further analysis.

3. Literature Review

There are certain factors within the characteristics of the population which makes the tourism industry lead to an improvement of the socio-economic conditions of the population. This will eventually result in low rates of unemployment and a higher percentage of the working population. The former improves the socioeconomic conditions of the population where as the latter helps in uplifting financial status through different tax burdens, public policies aimed at achieving a higher level of economic development.

The annual statistical report issued by Sri Lanka Tourism Development Authority provides us with the following facts:

2019
Highlights

Number of international tourists to Sri Lanka	**1,913,702**
Foreign exchange earnings	**Rs. 646,362.3 million**
	US$ 3606.9 million
Direct contribution to GDP	**4.3%*(Current price)**
Foreign exchange receipts per tourist per day	**US$ 181.2**
Average duration of stay	**10.4 nights**
Total foreign guest nights	**19,902,501**
Room occupancy rate of graded accommodation	**57.09%**
Total employment generation Direct employment Indirect employment	**402,607** 173,592 229,015
Top 5 tourist source markets to Sri Lanka	**India, United Kingdom China, Germany, France**

4. Methodology

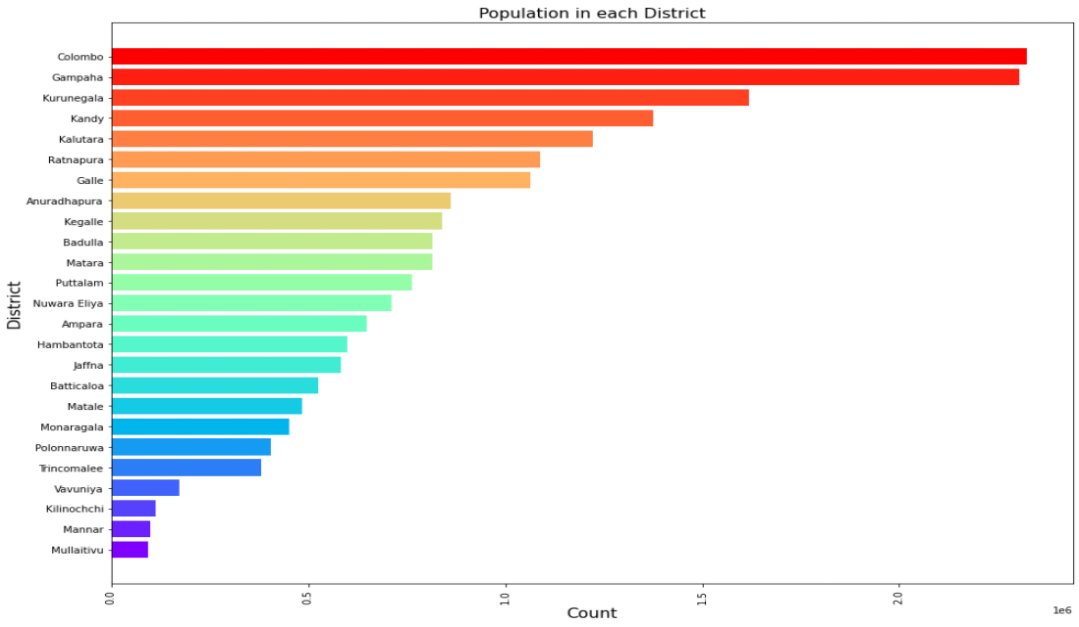
As mentioned in the previous section, data is collected through web scraping the Wikipedia page https://en.wikipedia.org/wiki/Districts_of_Sri_Lanka. Then with the use of Geopy API latitudes and longitudes of all districts of the country are obtained. After performing required data cleaning and necessary refinements to the data set, the final data set is taken as shown below.

	District	Province	Population	Population Density	District_latitude	District_longitude
0	Mullaitivu	Northern	92238	38 (98)	9.269853	80.814535
1	Mannar	Northern	99570	53 (140)	8.977244	79.913779
2	Kilinochchi	Northern	113510	94 (240)	9.384007	80.408722
3	Vavuniya	Northern	172115	92 (240)	8.759352	80.500078
4	Trincomalee	Eastern	379541	150 (390)	8.576425	81.234495
5	Polonnaruwa	North Central	406088	132 (340)	7.939536	81.000339
6	Monaragala	Uva	451058	82 (210)	6.902200	81.347838
7	Matale	Central	484531	248 (640)	7.472045	80.623431
8	Batticaloa	Eastern	526567	202 (520)	7.735603	81.694196
9	Jaffna	Northern	583882	629 (1,630)	9.665093	80.009303

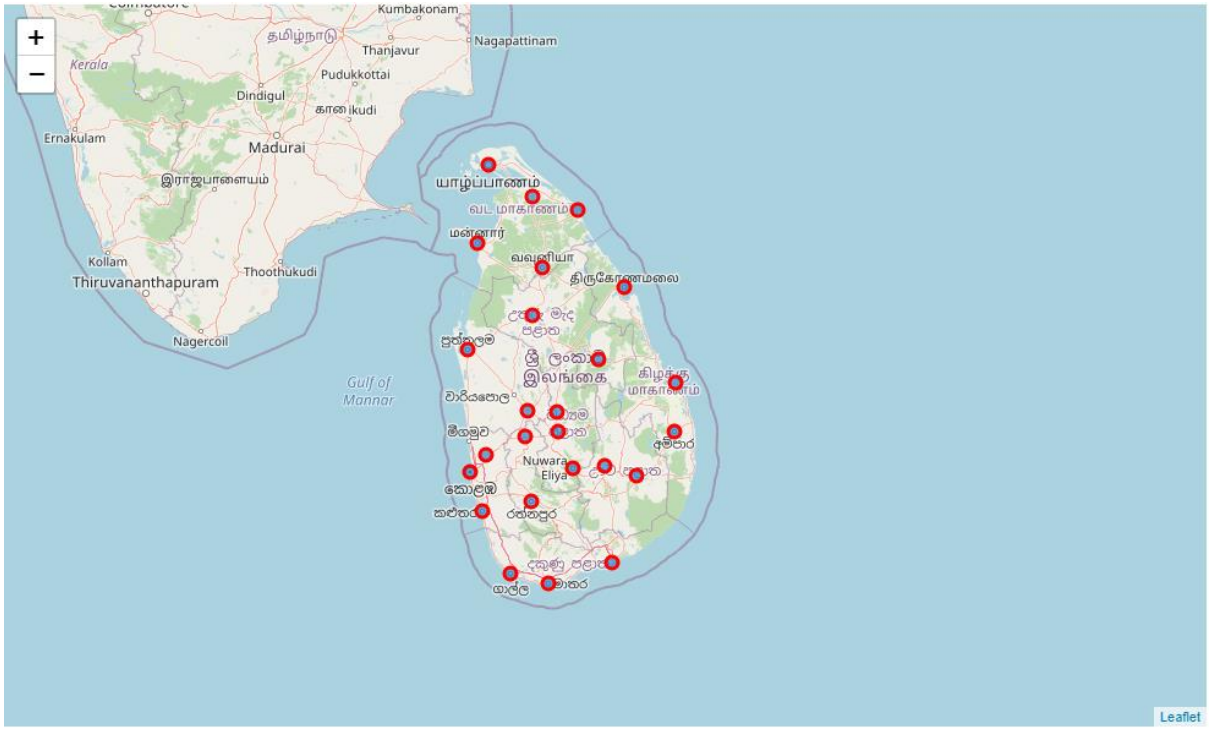
4.1 Exploratory Data Analysis

There are 25 districts that have been retrieved from the webpage and stored in the data frame as the main data set for this project.

As mentioned in the literature review, there can be some impacts of the population of a district on tourism. The below graph shows the population in each district.



With the aid of the dataset, a visualization with all districts in Sri Lanka can be obtained as shown below.



Using the Foursquare API, we acquire only the categories which are related to tourism for tourist category and which are related to tourist services for employment opportunities to people separately.

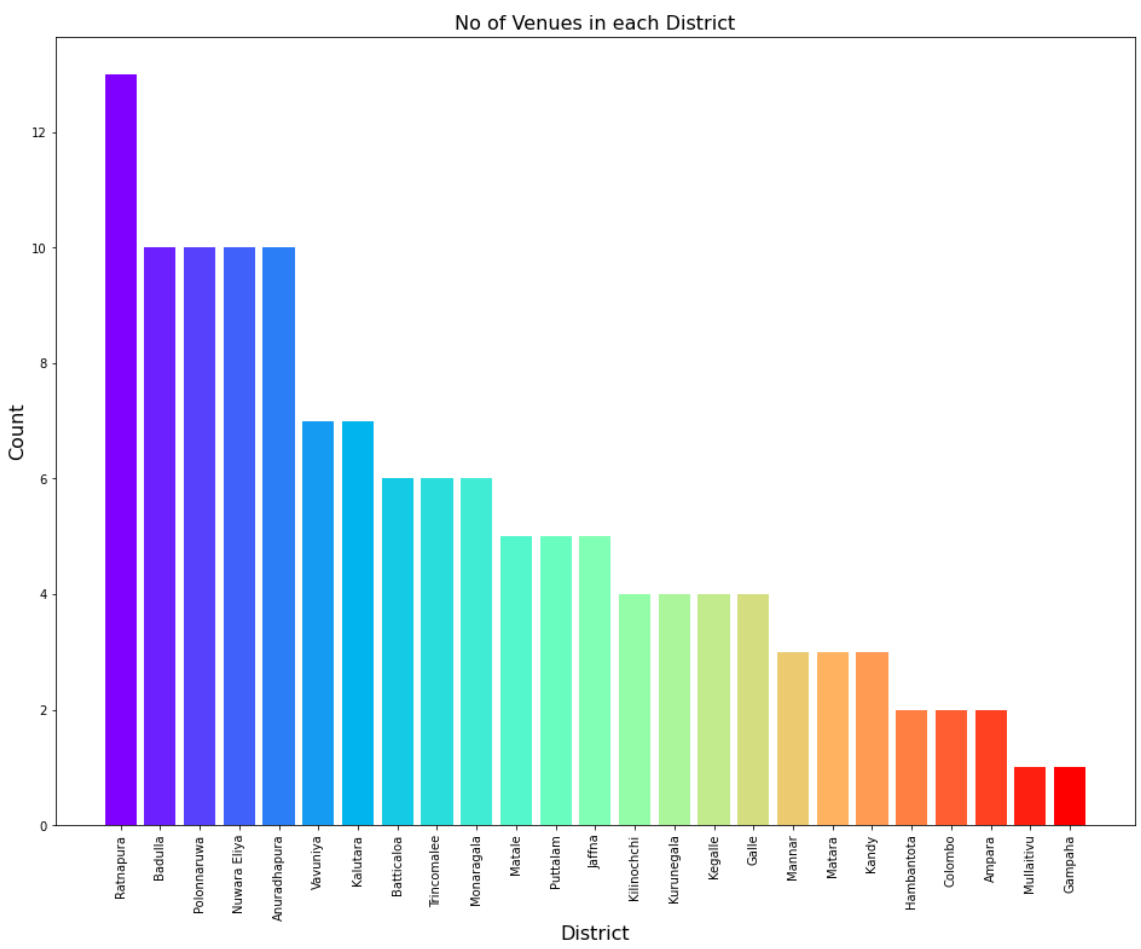
```
all_categories = {}

for i in range(categories):
    venues = category_results['response']['categories'][i]['name']
    all_categories[venues] = i

print(all_categories)
```

```
{'Arts & Entertainment': 0, 'College & University': 1, 'Event': 2, 'Food': 3, 'Nightlife Spot': 4, 'Outdoors & Recreation': 5,
'Professional & Other Places': 6, 'Residence': 7, 'Shop & Service': 8, 'Travel & Transport': 9}
```

The next step is to obtain the nearby tourist venues within a radius of 50km. This provides us with multiple tourist spots in a particular district. We can visualize this in a bar graph by plotting Districts v/s Count to obtain the number of venues in each district as shown below.



As the next step, we organize the unique venue categories obtained and create a one-hot encoding to analyse each district. These results is depicted in a data frame that displays the most common venue category in a particular district. The results are as follows.

	District	1st Most Common Venue Category	2nd Most Common Venue Category	3rd Most Common Venue Category	4th Most Common Venue Category	5th Most Common Venue Category	6th Most Common Venue Category	7th Most Common Venue Category	8th Most Common Venue Category	9th Most Common Venue Category	10th Most Common Venue Category	11th Most Common Venue Category	12th Most Common Venue Category	13th M Comm Ve Categ
0	Ampara	Beach	Park	Zoo	National Park	Botanical Garden	Bridge	Campground	Castle	Garden	Historic Site	Lake	Lighthouse	Moun
1	Anuradhapura	Historic Site	National Park	Lake	Zoo	Beach	Botanical Garden	Bridge	Campground	Castle	Garden	Lighthouse	Mountain	Muse
2	Badulla	Scenic Lookout	National Park	Bridge	Historic Site	Mountain	Waterfall	Zoo	Museum	Beach	Botanical Garden	Campground	Castle	Gar
3	Batticaloa	Beach	Pool	Park	Zoo	Museum	Botanical Garden	Bridge	Campground	Castle	Garden	Historic Site	Lake	Lightho
4	Colombo	Water Park	Pool	Zoo	Museum	Beach	Botanical Garden	Bridge	Campground	Castle	Garden	Historic Site	Lake	Lightho

Then we aggregate all the venues which belong to the particular category in a particular district. The aggregated results are as follows.

	District	Venue Category	Venue
0	Ampara	Beach	Kattankudy Beach
1	Ampara	Park	Mahathma Gandhi Park
2	Anuradhapura	Historic Site	Ruwanvelisaya Temple (රුවන්වැලිසෑය), Sri Maha Bodhi (ජය ශ්‍රී මහා බෝධිය), Anuradhapura Sacred City, Mihinthale, Jetavana Stupa, Samadhi Buddha Image, Aukana Temple, Kuttam Pokuna (Twin Ponds)
3	Anuradhapura	Lake	Elephant Pond
4	Anuradhapura	National Park	Wilpattu National Park

After obtaining the most common venue categories in all districts, we replace the categories with the venues if they are present in the district.

	District	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	Ampara	Kattankudy Beach	Mahathma Gandhi Park								
1	Anuradhapura	Ruwanvelisaya Temple (රුවන්වැලිසෑය), Sri Maha Bodhi (ජය ශ්‍රී මහා බෝධිය), Anuradhapura Sacred City, Mihinthale, Jetavana Stupa, Samadhi Buddha Image, Aukana Temple, Kuttam	Wilpattu National Park	Elephant Pond							

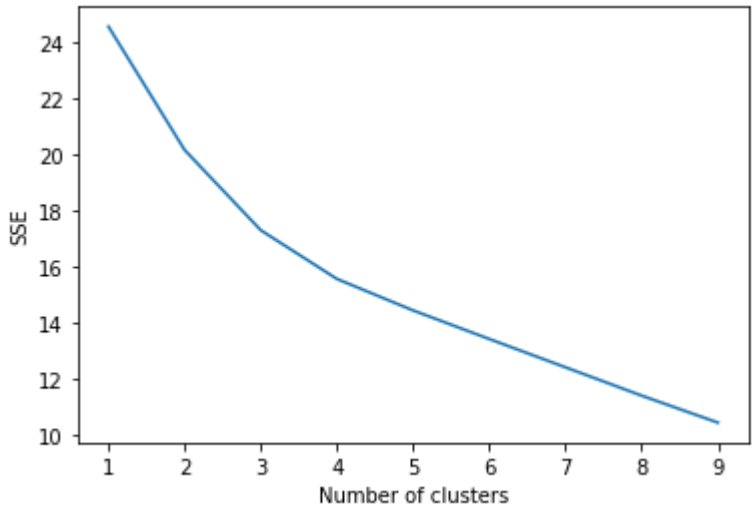
This provides an idea to a person as to where he could start his business in a particular district. But still, he can't decide or have any idea as to what type of business he could open up at a given tourist venue. So, in order to make sure that a business attracts many tourists as possible, we should find the most sought business at the tourist spot. So, then we acquire the top businesses which are being established at the tourist venue within the range of 500m.

	Venue	Business	BLatitude	BLongitude	Business Category
5	Mannar Fort	choice hotel	8.976538	79.913078	Fast Food Restaurant
15	Mullaitiv Beach	Mullai Cafe	9.272246	80.818185	Sri Lankan Restaurant
23	Mihinthale	Chamy Restuarent	8.358825	80.511776	Restaurant
24	Mihinthale	Prami Restaurant	8.359149	80.511500	Restaurant
28	Anuradhapura Sacred City	Margosa Lake Resort	8.352442	80.433730	Hotel
29	Anuradhapura Sacred City	Chef Roshan	8.354131	80.433591	Sri Lankan Restaurant
44	Fort Fredrick	INOX	8.573595	81.240960	Cosmetics Shop
45	Fort Fredrick	Rich bag choice 230, n.c road trincomalee	8.573595	81.240960	Sporting Goods Shop
48	Uppuveli Beach	Café on the 18th	8.613747	81.215965	Café
49	Uppuveli Beach	Rice❤Curry	8.616524	81.216377	Sri Lankan Restaurant

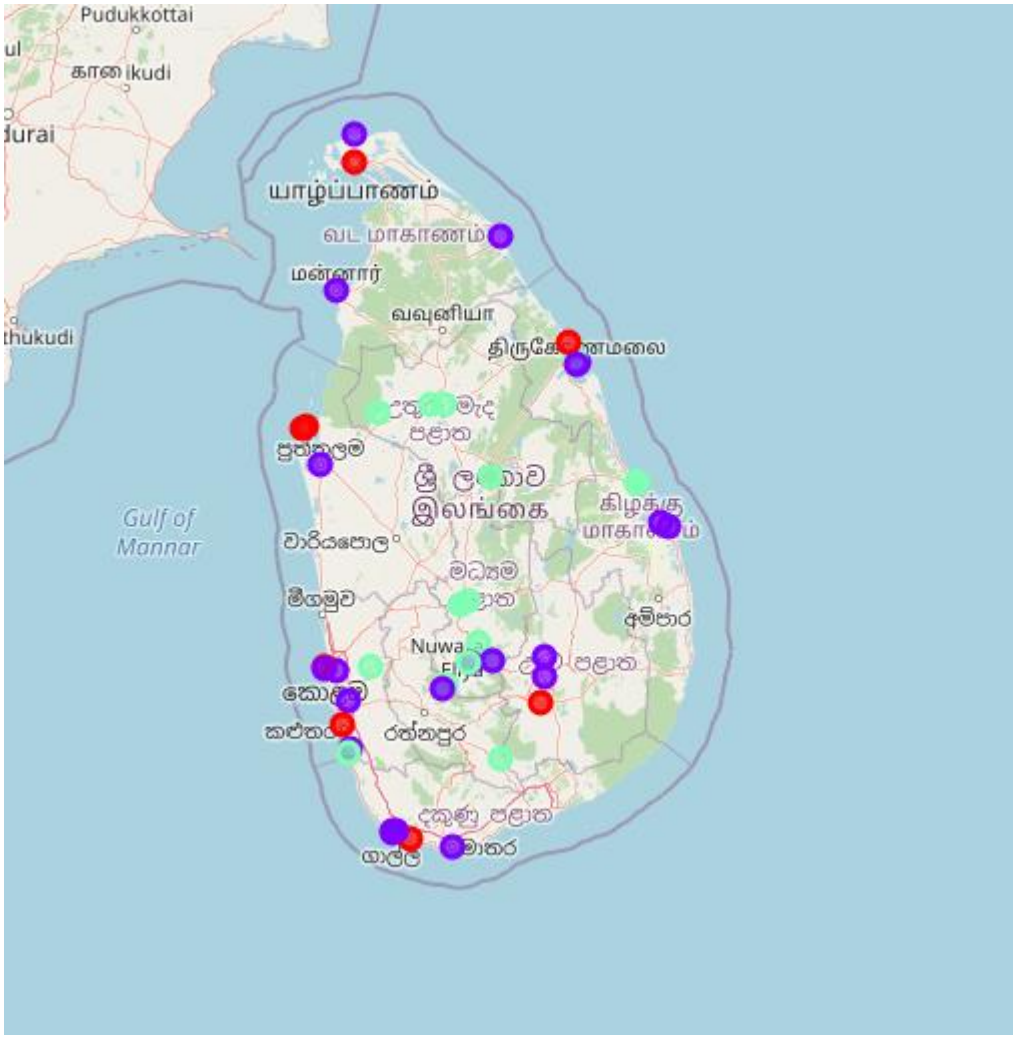
Then we perform similar one hot encoding and analyse each venue to get the top business at a venue.

	Venue	1st Most Common Business	2nd Most Common Business	3rd Most Common Business	4th Most Common Business	5th Most Common Business	6th Most Common Business	7th Most Common Business	8th Most Common Business	9th Most Common Business	10th Most Common Business
0	9 Arch Bridge - Demodara	Café	Juice Bar	Tourist Information Center	Breakfast Spot	Dumpling Restaurant	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Burger Joint
1	Adam's Peak (Sri Pada) Footpath	Hotel	Sri Lankan Restaurant	Café	Tourist Information Center	Bookstore	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Burger Joint
2	Anuradhapura Sacred City	Sri Lankan Restaurant	Hotel	Tourist Information Center	Bookstore	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Café	Burger Joint
3	Beruwala Lighthouse	Hotel	Spa	BBQ Joint	Tourist Information Center	Breakfast Spot	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Café
4	China Fort	Jewelry Store	Flea Market	Market	Fast Food Restaurant	Dumpling Restaurant	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Café

After that, we use the K-means clustering algorithm to group the businesses into clusters that aim to partition ‘n’ observations into k clusters in which each observation belongs to the cluster. Here elbow method is used to determine the optimum value of k to perform K-means clustering. The graph obtained is shown below.



5. Results



The colours purple, green and represents cluster 0,1 and 2 respectively.

The results depict that the most common business in that can be found in each 3 clusters.

6. Discussion

From the results shown in the first cluster (cluster 0), the most sought business at the respective venues is Fast Food Restaurants, Indian Restaurants, Asian Restaurants. This is clearly visible in the map. The purple cluster around the beach sides around the island clearly indicate that opening a seafood restaurant would help a person make the best of the opportunity as majority of the foreign tourists attracts beach side.

264	St. Clair's Falls Viewing Gallery	Indian Restaurant	Tourist Information Center	Breakfast Spot	Dumpling Restaurant	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Café	Burger Joint
267	Gregory Lake park	Fast Food Restaurant	Hotel	Café	Breakfast Spot	Dumpling Restaurant	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Burger Joint
288	Puttalam Lagoon	BBQ Joint	Tourist Information Center	Breakfast Spot	Dumpling Restaurant	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Café	Burger Joint
305	Matara Beach	Fast Food Restaurant	Asian Restaurant	Café	Restaurant	Breakfast Spot	Dumpling Restaurant	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop
324	Galle Fort	Sri Lankan Restaurant	Café	Restaurant	Mediterranean Restaurant	Arts & Crafts Store	Hotel	Burger Joint	Bistro	Dumpling Restaurant	Breakfast Spot
482	Galle Lighthouse	Restaurant	Sri Lankan Restaurant	Café	Mediterranean Restaurant	Arts & Crafts Store	Hotel	Burger Joint	Bistro	Dumpling Restaurant	Breakfast Spot
502	Rumassala Beach	Indian Restaurant	Restaurant	Hotel	Boat or Ferry	Fast Food Restaurant	Tourist Information Center	Buffet	Diner	Dessert Shop	Cosmetics Shop

According to the second cluster (cluster 1), it clearly shows that the hotels are prominent business to start at relevant tourist attraction site. As it seems many people prefers to enjoy their tour accommodating and having dines at same place that is much closer to the place they are visiting.

	Venue	1st Most Common Business	2nd Most Common Business	3rd Most Common Business	4th Most Common Business	5th Most Common Business	6th Most Common Business	7th Most Common Business	8th Most Common Business	9th Most Common Business	10th Most Common Business
23	Mihinthale	Restaurant	Tourist Information Center	Bookstore	Dumpling Restaurant	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Café	Burger Joint
28	Anuradhapura Sacred City	Sri Lankan Restaurant	Hotel	Tourist Information Center	Bookstore	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Café	Burger Joint
48	Uppuveli Beach	Hotel	Sri Lankan Restaurant	Seafood Restaurant	BBQ Joint	Café	Buffet	Tourist Information Center	Breakfast Spot	Diner	Dessert Shop
107	Sigiriya Rock (සීගිරිය)	RV Park	Bookstore	Dumpling Restaurant	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Café	Burger Joint	Buffet
113	Sigiryia Rock Fresco	RV Park	Bookstore	Dumpling Restaurant	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Café	Burger Joint	Buffet
134	Kandy View Point	Hotel	Fast Food Restaurant	Sri Lankan Restaurant	Souvenir Shop	Food Court	Shopping Mall	Diner	Breakfast Spot	Dessert Shop	Cosmetics Shop
152	Royal Botanic Gardens	Hotel	Coffee Shop	Tourist Information Center	Breakfast Spot	Dumpling Restaurant	Diner	Dessert Shop	Cosmetics Shop	Café	Burger Joint
161	Orchid House - Royal Botanical Gardens	Hotel	Coffee Shop	Tourist Information Center	Breakfast Spot	Dumpling Restaurant	Diner	Dessert Shop	Cosmetics Shop	Café	Burger Joint
173	Passekudah Bay	Hotel	Restaurant	Tourist Information Center	Bookstore	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Café	Burger Joint
179	Pasikudah	Hotel	Asian Restaurant	Restaurant	Tourist Information Center	Breakfast Spot	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Café
194	Maalu Maalu	Hotel	Breakfast Spot	Asian Restaurant	Tourist Information Center	Dumpling Restaurant	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Café

Finally, the second cluster (cluster 2) also provides a similar result as the second cluster.

Cluster 2

```
nearby_business_merged.loc[nearby_business_merged['Cluster Labels'] == 2, nearby_business_merged.columns[[0] + list(range(4, near
```

	Venue	1st Most Common Business	2nd Most Common Business	3rd Most Common Business	4th Most Common Business	5th Most Common Business	6th Most Common Business	7th Most Common Business	8th Most Common Business	9th Most Common Business	10th Most Common Business
65	Nilaveli Beach	Hotel	Tourist Information Center	Breakfast Spot	Dumpling Restaurant	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Café	Burger Joint
93	Pidurangala Rock	Hotel	Tourist Information Center	Breakfast Spot	Dumpling Restaurant	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Café	Burger Joint
131	Diyaluma Falls	Hotel	Tourist Information Center	Breakfast Spot	Dumpling Restaurant	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Café	Burger Joint
219	Dutch Fort	Hotel	Tourist Information Center	Breakfast Spot	Dumpling Restaurant	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Café	Burger Joint
292	Sri Lanka Kite	Hotel	Tourist Information Center	Breakfast Spot	Dumpling Restaurant	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Café	Burger Joint
298	Cocodance kitesurfing resort	Seafood Restaurant	Tourist Information Center	Bookstore	Dumpling Restaurant	Diner	Dessert Shop	Cosmetics Shop	Coffee Shop	Café	Burger Joint

7. Conclusion

In this project I attempted to make the use of Foursquare API to get the famous tourist locations situated in a particular district in Sri lanka. K-means clustering algorithm has been used to cluster these tourist spots based on exploring the frequency of the businesses that are already available which could help us indicate a business opportunity that could be established in the tourist hotspots so that business could attract as many tourists as possible.

Future possible research could make the use of other significant factors which includes the finding number of similar businesses that could impact the new business being established resulting competition and demand for a particular business, accessibility, and average business rates that could be incurred for a particular business. All these above-mentioned factors could help the system to perform analysis more accurately.

8. References

[1]"Districts of Sri Lanka", *En.wikipedia.org*, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Districts_of_Sri_Lanka. [Accessed: 31- Dec- 2020].

[2]*Sltda.gov.lk*, 2020. [Online]. Available: https://sltda.gov.lk/storage/common_media/AAnnual%20Statistical%20Report%20new%202109%20Word3889144215.pdf. [Accessed: 31- Dec- 2020].

[3]*S3-api.us-geo.objectstorage.softlayer.net*, 2020. [Online]. Available: https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/DP0701EN/sample_submission/Predicting_the_Improvement_of_NBA_players_Report.pdf. [Accessed: 31- Dec- 2020].