IBM Applied Data Science Capstone

Recommending a Business at a particular Tourism Site

Quan Kien Minh

Contents

I. Introduction

Tourism has always been a thriving sector across the world. No matter which country you are living in, you can always come across a group of people, big or small, who always like to visit attractions. I am a big fan of adventure, and I acknowledge this fact as to how tourism plays a salient role for a traveler/explorer. Tourism is not only an important aspect of a country's economy but also for its global standing.

Why Tourism is important to any country?

The tourism industry is important for the benefits it brings and due to its role as a commercial activity that creates demand and growth for many more industries. Tourism not only contributes to more economic activities but also generates more employment, revenues, and play a significant role in development.

- Tourism activity creates demand.
- Tourism industry value chain meets & spreads demand across industries & boosts more economic activities.
- Tourism induces more consumption.

II. Business Problem

All the benefits of tourism tend to reflect on the employment opportunity which it gives to the people of that country. The objective of this project is to analyze the tourist places of a

given state in Vietnam, and try to recommend the best location where they can open a restaurant or lodging to make the best use of the opportunity.

The target audience for this project includes people who are interested in opening a restaurant, lodging, transport services, or any other similar businesses which fall within the tourism industry. This also recommends travelers' tourist venues to be visited in a given state of a country.

III. Data Anatomization

To tackle the above mentioned problem, we need to have the dataset that contains -

- All the provinces of a Vietnam.
- Latitude and longitudes of all the districts.

The major sources of data are derived from [2][3][4]. Those sources obtain all the districts of Ho Chi Minh, Ha Noi, Da Nang municipalities. We then use beautifulsoup4 package, a Python module that helps to scrape information from the web pages to extract all the tables from this Wikipedia page and convert it into a pandas dataframe. Then we use Python's geopy package to obtain the latitude and longitude of all the districts present in the dataframe.

Description of the data

The output shows the final dataset. The dataset consists of a single Dataframe with 7 columns containing Municipalities, District/Municipal City, Area (km^2), Population (person), Wards, Latitude, Longitude.

Municipalities	District/Municipal City	Area (km2)	Population (person)	Wards	Latitude	Longitude
Ho Chi Minh City	Thu Duc City	21156.0	1013795	34 wards	10.829830	106.761790
Ho Chi Minh City	District 1	772.0	142625	10 wards	10.774845	106.699350
Ho Chi Minh City	District 3	492.0	190375	12 wards	10.771551	106.698380
Ho Chi Minh City	District 4	418.0	175329	13 wards	10.759243	106.704890
Ho Chi Minh City	District 5	427.0	159073	14 wards	10.756129	106.670375

Table 1. Description of the Data

IV. Literature Review

There are specific factors within the characteristics of the population which makes the tourism industry lead to an improvement of the socio-economic conditions of the population

[1]. 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 whereas the latter helps finance, through different tax burdens, public policies aimed at achieving a higher level of economic development. It also demonstrates that countries with regressive population pyramids have greater difficulties for tourism growth to improve their socio-economic conditions.

The survey from Annual Report Tourism of Vietnam provides us with the following facts:

- According to the World Economic Forum's (WEF) 2017 Tourism Competitiveness Index, Vietnam ranks 32nd globally (out of 120 countries) in terms of the volume and attractiveness of natural and cultural resources, and 3rd within Southeast Asia.
- Vietnam is home to 8 UNESCO World Heritage sites, tied with Indonesia at the top spot for SEA.
- The country also has premier urban tourism destinations such as Hanoi, Ho Chi Minh City, and Da Nang, which have been growing rapidly for the past years as well.
- The number of international tourist arrivals has reached nearly 14.5 million in the first 10 months of 2019 (13% more than the same period last year).
- In 2018, there was a four-fold increase in the number of domestic traveler trips, from 20.5 million in 2008 to 80 million in 2018.
- Hotels & Tourism is the 4th top contributor to the country's GDP as of Q1 2019.

V. Methodology

The first step is to collect the data. This is done by scraping the Wikipedia pages [2][3][4]. Then we use geopy API to get the latitude and longitude of all the districts of the country. There existed some missing values in the dataset which were removed. The final dataset has 7 columns containing Municipalities, District/Municipal City, Area (km^2), Population (person), Wards, Latitude, Longitude:

Municipalities	District/Municipal City	Area (km2)	Population (person)	Wards	Latitude	Longitude
Ho Chi Minh City	Thu Duc City	21156.0	1013795	34 wards	10.829830	106.761790
Ho Chi Minh City	District 1	772.0	142625	10 wards	10.774845	106.699350
Ho Chi Minh City	District 3	492.0	190375	12 wards	10.771551	106.698380
Ho Chi Minh City	District 4	418.0	175329	13 wards	10.759243	106.704890
Ho Chi Minh City	District 5	427.0	159073	14 wards	10.756129	106.670375

Table 2. Data after being cleaned

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

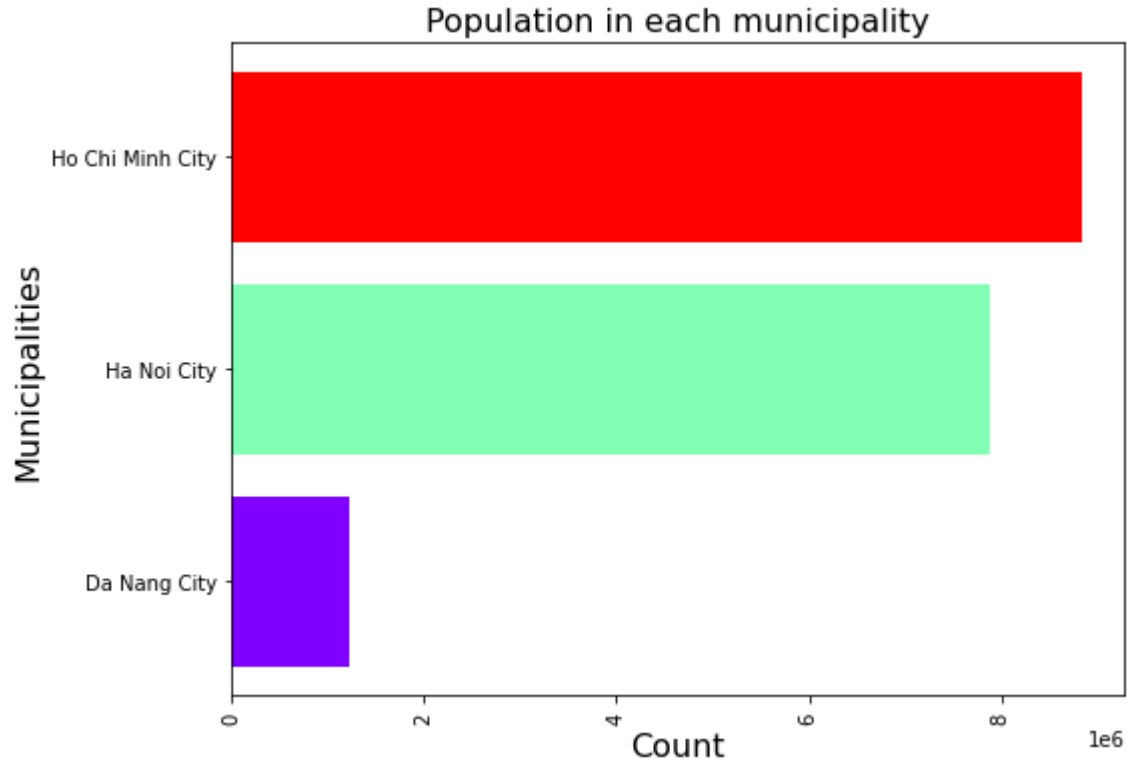


Figure 1. Population in each municipality

The user can enter the state of his choice among the given states. Here Ho Chi Minh City is taken as a choice. A visualization with all the districts of the given municipality will be displayed as shown below:

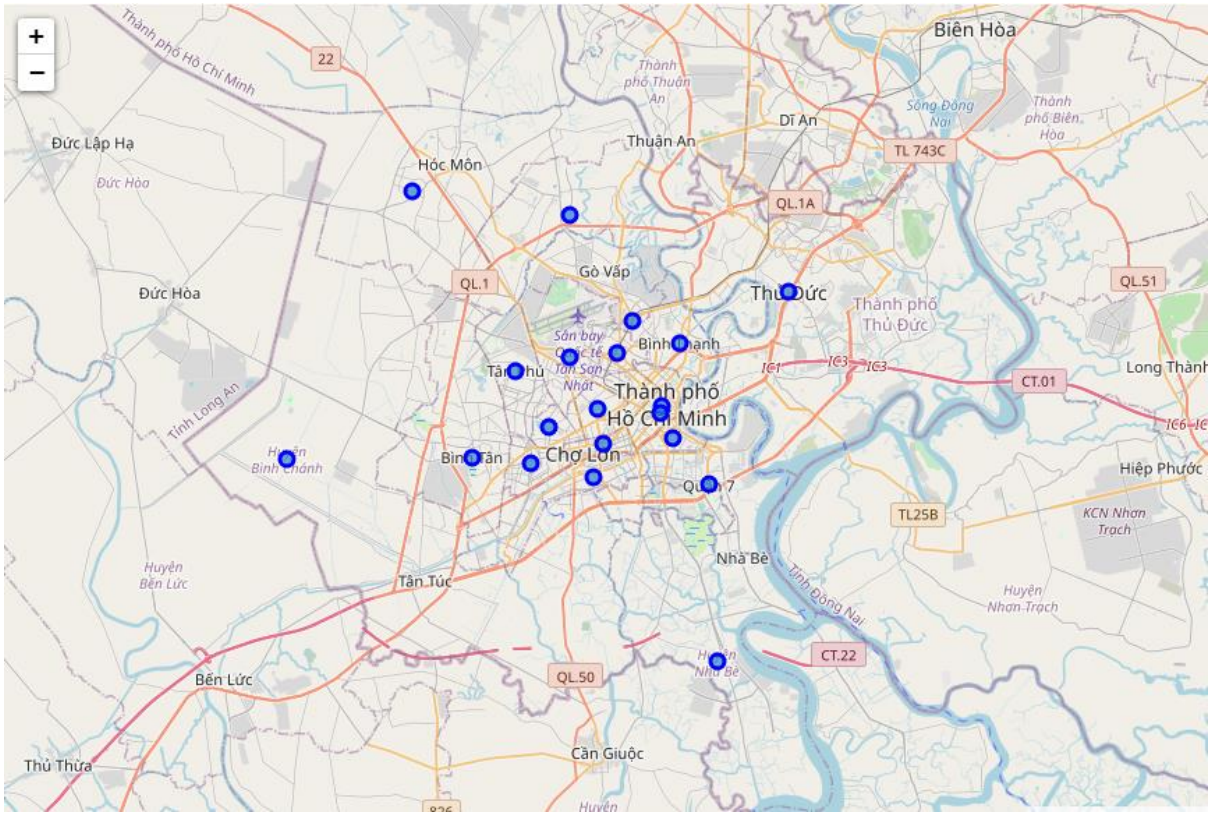


Figure 2. Visualization of districts in Ho Chi Minh City

Using the Foursquare API, we acquire only the categories which are related to tourism for tourists' category and which are related to tourist services for employment opportunities to people separately. The former includes Arts & Entertainment, Nightlife Spot, Outdoors & Recreation, whereas the latter includes Food, Shop & Service, Travel & Transport services. The next step is to obtain the nearby tourist venues within a radius of 25km. This gives us multiple tourist spots if there are in a particular district. We visualize a bar graph by plotting District with the count to obtain the number of venues in each district. The visualization can be shown like the following:

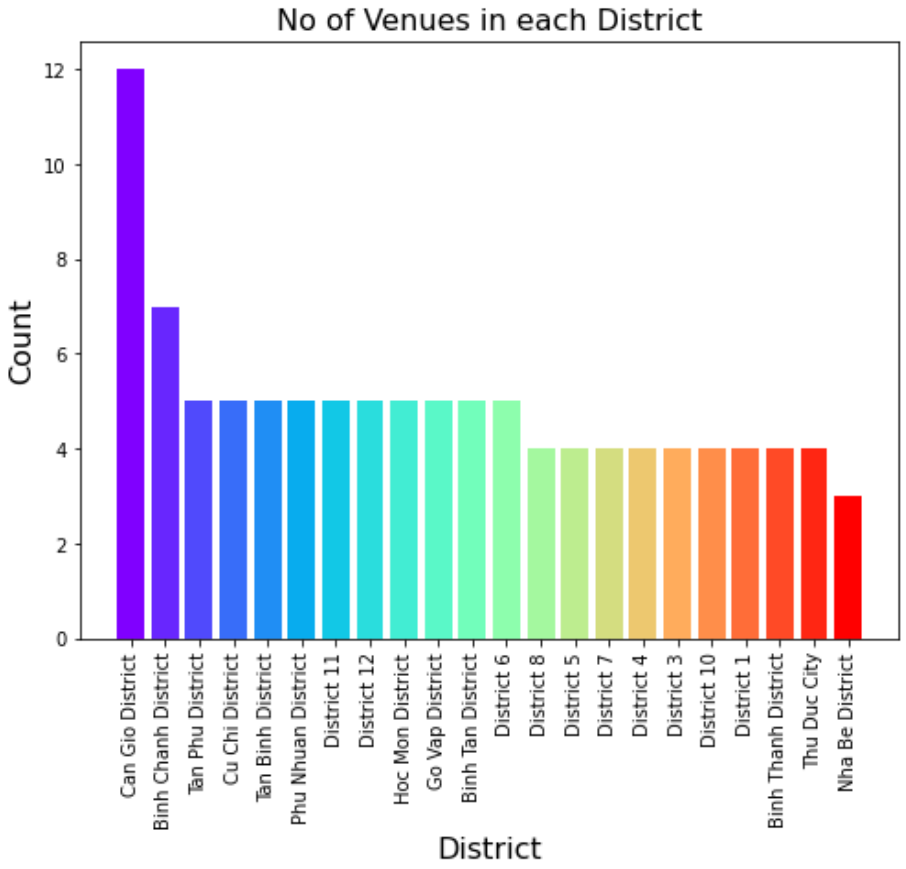


Figure 3. Venues in each District

We then organize the unique venue categories obtained and create a one-hot encoding to analyze each district. This results in a data-frame that displays the most common venue category in a particular district. The result is as shown the following:

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
Binh Chanh District	Theme Park	Public Art	Park	Historic Site	Gun Range	Garden	Brewery	Campground
Binh Tan District	Theme Park	Public Art	Park	Historic Site	Gun Range	Garden	Campground	Brewery
Binh Thanh District	Theme Park	Public Art	Park	Historic Site	Gun Range	Garden	Campground	Brewery
Can Gio District	Brewery	Beach	Theme Park	Public Art	Park	Historic Site	Gun Range	Garden
Cu Chi District	Theme Park	Public Art	Park	Historic Site	Gun Range	Garden	Brewery	Campground

Table 3. The common venues in each district

We then aggregate all the venues which belong to the particular category in a particular district.

District	Venue Category	Venue
Binh Chanh District	Art Gallery	Artinus 3D Painting Gallery
Binh Chanh District	Brewery	Pasteur Street Brewing Company, Winking Seal Beer Co.
Binh Chanh District	Garden	Tùng Sơn Thạch Hoa Viên - Rin Rin Park
Binh Chanh District	Park	Công viên Gia Định (Gia Dinh Park), Cau Ca Thanh Long
Binh Chanh District	Public Art	Saigon Outcast

Table 4. Venue categories in each district

After obtaining the most common venue categories in all the 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
Binh Chanh District	Công viên Gia Định (Gia Dinh Park), Cau Ca Thanh Long	Pasteur Street Brewing Company, Winking Seal Beer Co.	Saigon Outcast	Tùng Sơn Thạch Hoa Viên - Rin Rin Park	Artinus 3D Painting Gallery
Binh Tan District	Saigon Outcast	Công viên Gia Định (Gia Dinh Park)	Tùng Sơn Thạch Hoa Viên - Rin Rin Park	Pasteur Street Brewing Company	Artinus 3D Painting Gallery
Binh Thanh District	Saigon Outcast	Công viên Gia Định (Gia Dinh Park)	Tùng Sơn Thạch Hoa Viên - Rin Rin Park	Pasteur Street Brewing Company	

Table 5. The most common venue categories in all the districts

This gives an idea to a person as to where he could start his business in a particular district. But still, he can be not sure or have any idea as to what type of business he could open up at a given tourist venue. So to make sure that his business attracts many tourists as possible, we then attempt to 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 500 meters.

Venue	Business	BLatitude	BLongitude	Business Category
Saigon Outcast	Mr. Singh Indian Restaurant	10.813892	106.726555	Indian Restaurant
Pasteur Street Brewing Company	B3 - Steakhouse & Craft Beer	10.775190	106.702492	Steakhouse
Pasteur Street Brewing Company	O Lé	10.774772	106.699524	Spanish Restaurant
Pasteur Street Brewing Company	Takashimaya	10.773194	106.701075	Department Store
Pasteur Street Brewing Company	Liberty Central Saigon Citypoint Hotel	10.774758	106.700795	Hotel
Pasteur Street Brewing Company	Boa cafe	10.775238	106.702770	Coffee Shop
Pasteur Street Brewing Company	The Old Compass Cafe	10.774816	106.700685	Café
Pasteur Street Brewing Company	Le Bourgeois @Continental	10.776408	106.702731	French Restaurant
Pasteur Street Brewing Company	Shin Coffee	10.775219	106.703208	Café
Pasteur Street Brewing Company	Pizza 4P's Saigon Center	10.773303	106.700806	Pizza Place
Pasteur Street Brewing Company	L'Usine: Cafe, Bistro & Lifestyle Shop	10.775994	106.703186	Café
Pasteur Street Brewing Company	Park Hyatt Saigon	10.777574	106.703609	Hotel

Table 6. The nearby business of each venue

We then perform similar one hot encoding and analyze each venue to get the top businesses 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
Beach	Hotel	Seafood Restaurant	Café	Restaurant	Diner	Video Game Store	Department Store	Food Truck	Flower Shop	Flea Market
Biển Tân Thành	Seafood Restaurant	Market	Video Game Store	Convenience Store	Food Truck	Flower Shop	Flea Market	Electronics Store	Eastern European Restaurant	Diner
Cu Chi Tunnels (Ben Duoc)	Motorcycle Shop	Video Game Store	French Restaurant	Food Truck	Flower Shop	Flea Market	Electronics Store	Eastern European Restaurant	Diner	Department Store

Table 7. The most common business in a particular venue

We then 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.

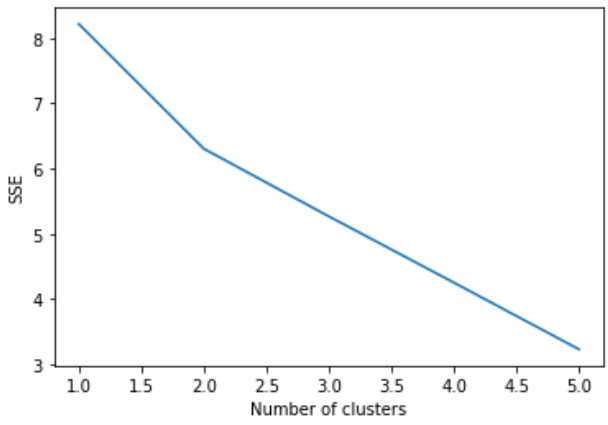
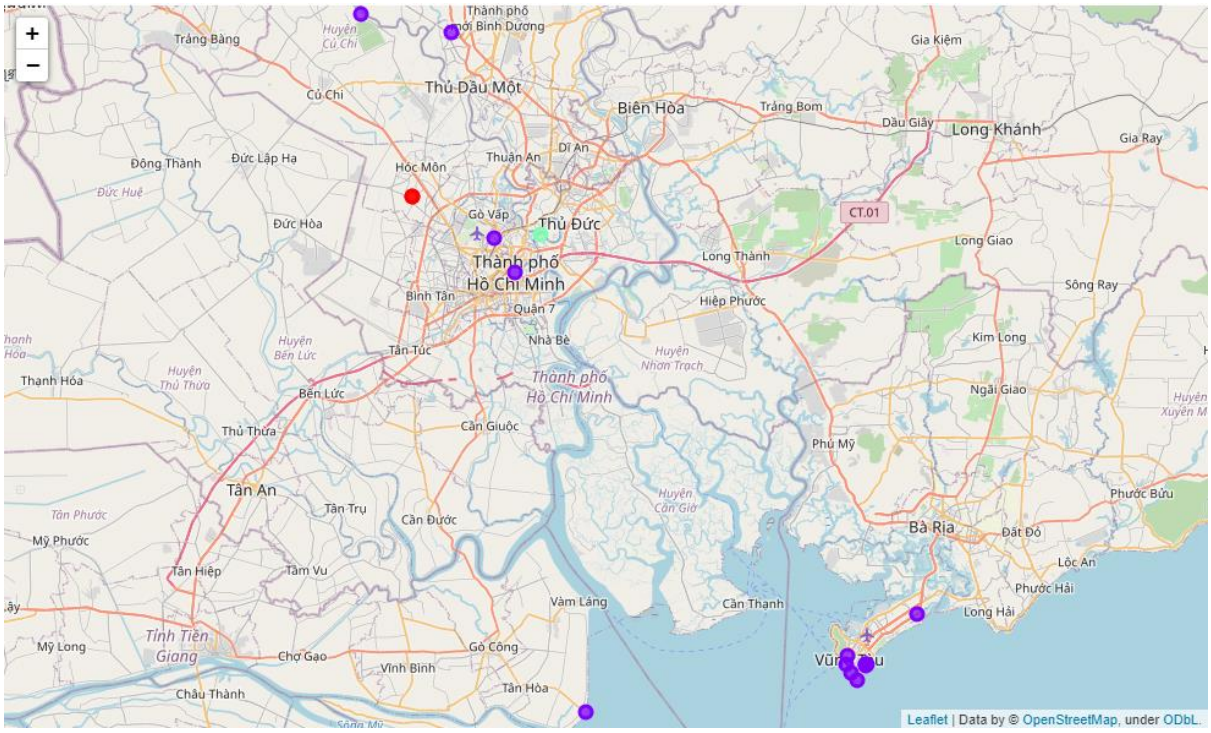


Figure 4. The graph used to find K clusters

VI. Results and Discussion



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

The results show that the most common business in cluster one at the respective venues are Hotel. So Hotel are popular in these tourist venues and opening up a similar one can attract many tourists. This is because these venues are adjacent to Vung Tau beach, tourists would like to stay there several days to relax and experience the flavor of fresh seafood. Thus this could be a nice opportunity to open up a business at that locality.

Cluster 0

Whereas in cluster two the most sought business is the Hotel, Seafood Restaurants, and Cafeterias. This is clearly visible in the map above. The green clusters at the seaside clearly indicate that opening a seafood restaurant would help a person make the best use of the opportunity. Also, there are some green clusters in the middle of the map, which indicates Hotels and Cafeterias would be the best business at that tourist spot.

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
Pasteur Street Brewing Company	Hotel	Coffee Shop	Café	Spa	Italian Restaurant	French Restaurant	Massage Studio	Asian Restaurant	Burger Joint	BBQ Joint
Công viên Gia Định (Gia Dinh Park)	Café	Convenience Store	Flea Market	Electronics Store	Department Store	French Restaurant	Food Truck	Flower Shop	Eastern European Restaurant	Diner
Vung Tau Beach	French Restaurant	Video Game Store	Halal Restaurant	Food Truck	Flower Shop	Flea Market	Electronics Store	Eastern European Restaurant	Diner	Department Store
Dog Racing Stadium	Hotel	French Restaurant	Breakfast Spot	Café	Restaurant	Department Store	Food Truck	Flower Shop	Flea Market	Electronics Store
Hải Đăng Vũng Tàu	Hotel	Café	Australian Restaurant	Video Game Store	Diner	French Restaurant	Food Truck	Flower Shop	Flea Market	Electronics Store
Vung Tau Front Beach	Asian Restaurant	Hotel	BBQ Joint	Indian Restaurant	Coffee Shop	Diner	Food Truck	Flower Shop	Flea Market	Electronics Store

Table 8. Cluster 0

Cluster 1

In middle of city, tourists always like to experience the flavor of different dishes available at a particular location and so this could be a nice opportunity to open up a business at that locality.

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
Saigon Outcast	Indian Restaurant	Video Game Store	Department Store	French Restaurant	Food Truck	Flower Shop	Flea Market	Electronics Store	Eastern European Restaurant	Diner

Table 9. Cluster 1

Cluster 2

Finally, in cluster three BBQ Joint has been given a top priority.

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
Tùng Sơn Thạch Hoa Viên - Rin Rin Park	BBQ Joint	Video Game Store	Department Store	French Restaurant	Food Truck	Flower Shop	Flea Market	Electronics Store	Eastern European Restaurant	Diner

Table 10. Cluster 2

VII. Conclusion

In this project, an attempt has been made to make use of the Foursquare API to get the famous tourist locations situated in a particular district of a municipality. K-means clustering algorithm has been used to cluster these tourist spots based on exploring the frequency of the businesses that are present which could help us indicate a business opportunity that could be established in the locality so that the business could attract as many tourists as possible.

Future possible research could make use of other significant factors which includes the foot traffic where the tourists are likely to bypass the area with a high traffic area, competition etc. The number of similar businesses also could impact the new business being established, accessibility, and average business rates that could be incurred for a particular business. These above-mentioned factors could help the system make the analysis more accurate.

VIII. References

[1] https://www.educationaltravelasia.org/economy-tourism-vietnams-timeless-charm/

[2] https://vi.wikipedia.org/wiki/Th%C3%A0nh_ph%E1%BB%91_H%E1%BB%93_Ch%C3%AD_Minh

[3] https://en.wikipedia.org/wiki/Hanoi

[4] https://en.wikipedia.org/wiki/Da_Nang