Coursera Capstone

IBM Applied Data Science Capstone

Finding ideal hotels in Singapore

By Do Thi Kim Uyen

January, 2020

I. Introduction

For many travelers, to find out an ideal hotel for their trip to Singapore is quite important. As a heart of Asia, which is easy to reach by many transportation types, Singapore is now growing incredibly in hostibility industry. Many entertainment parks, hotels, man-made beautiful views, sightseeings.. were being built to maximize exploit the strength of Singapore in travelling.

Think as a travelers, we should find a hotel which is easy to reach to shopping mall also entertainment area, sightseeings… With a small country like Singapore and the population is quite big enough, living standard is high, and in downtown Singapore, the price is extremely high within real estate for rent, not except for hotel price. We also can rent accommodation in suburbs but in my point of view, it must be close to MRT and convenience stores. If your pocket has its limitation, tourists do not want to put much of their expenses on just accommodation. Our choices focus on entertainment and traveling experience. That is why we do not have many choices on hotel.

As a result, I come to a research on finding the ideal hotel which is close to utilities and help to match with your expected hotel expense. Particularly, the location of the hotel is the most important decision which is ideally match to our expectations on a trip to Singapore.

Busniness problem

The objective of this capstone is to analyse the best hotel location in Singapore. Using data science methodology and machine learning techniques like clustering, the porject aims to provide solutions to answer question: In Singapore, if travelers are looking for an ideal hotel to stay, which hotel would you recommend them?

Target Audience of this Project

This project is obviously useful for travelers, who want to help great travelling experience to Singapore with a suitable amount of money. This project is always timely.

Because Singapore is always a hot destination of travelers all year around. Data from Budget Direct Insurance about Singapore Tourism Statistics 2019 released recently showed that there are 7.8 millions international tourists arrived in Singapore, which is up 1.49% compared to 2018. Anyone, who desire to visit Singapore, may be really interest in this project.

II. Data

1. To solve the problem, we will need the following data:

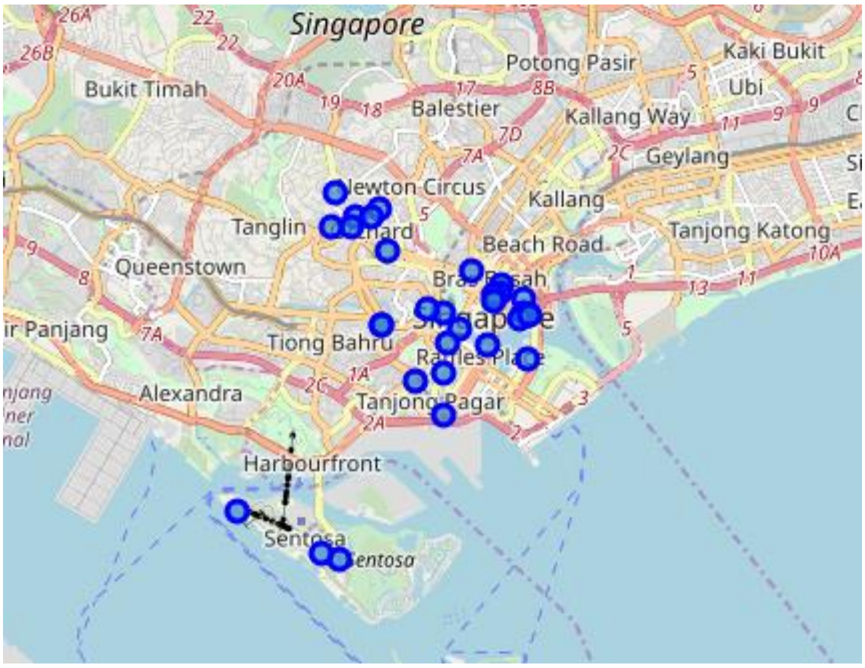
- List of neighborhood in Singapore.
- Latitude and longitude coordinates of those neighborhoods. This is required to plot the map and also to get the venue data.
- Venue data, particularly data related to hotels. We will use this data to perform clustering on the neighborhoods.

2. Sources of data and methods to extract them

This wikipedia page (https://en.wikipedia.org/wiki/List_of_hotels_in_Singapore) contains a list of hotels in Singapore, with a total 29 hotels. We will use web scraping techniques to extract data from Wikipedia page, with the help of Python Beautifulsoup and requests packages. Then we can get cooridinates of neighborhood using Geocorder package. This package give us the latitude and longitude coordinates of those neighborhood.

The next step, we will use Foursquare API to get the venue data of the neighborhood. Foursquared API can provide may of categories of utilities data, which is surround a hotel. We are particularly interest in entertainment area, shopping mall, sightseeing categories in order to find out which hotel is a good choice. To finish this project, I must use Data science skills and knowledges, from web scraping, working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-Means Clustering) and visualizing use Folium and then analyse to give comments on hotel choices depend on types of target audience.

We can also visualize geographic information for all these hotels by using Folium library. So I drawed a map of Singapore with these hotel location base on the latitude and longitude from crawled data.



III. Methodology

In this project, I try to find a hotel which is located in an area surround by many utilities (such as restaurants, genaral entertainments, market/ convenience store, shopping malls…). This report will target to travelers, who is interested in arrive in Singapore with a spirit of experience in culture and happiness.

First, to define the business problem in the Introduction part.

Secondly, to download the data and preprocess and explore what we got. We do the web scraping using Python requests and beautiful soup to extract the list from Wikipedia page. However, it is just a list of name. We need to get the Latitude and Longitude using Foursquared API. To to this, we can use the Geocoder package, that will allow to convert address into geographical coordinates with Latitude and Longitude. After combined the data, we will convert them into a DataFrame then using Folium to visualize the hotel locations. This step is a way to check if the coordinates we gatherd by Geocorder are correctly plotted in Singapore.

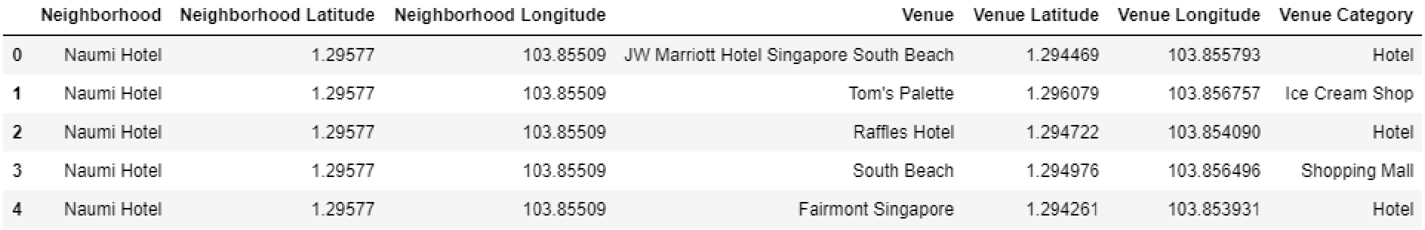
Next, we use Foursquared API to get the top 100 venues that are within a radius of 500 meters. We need to register a Foursquared Developer API in order to obtain the Foursquared ID and secret key. We then make API calls to Foursquared passing the geographical coordinates of the neighborhoods. With the data we can check all utilities near any hotels and how many utilities in total in each neighborhood and examine how many unique categories can be curated from all venues data. Then, we will analyse each neighborhood by grouping the rows by neighborhood and calculate the mean of frequency of occcurrence of each venue category. Until here, we finished cleaning up and preparing the data for clustering.

Finally, we perform clustering the data by using unsupervised learning algorithm K-Means clustering. First of this step, I clarify which number of centroids is suitable for the dataset by using the Elbow method. We choose k base on the loss and number of frequency is still in acceptable rate. Then analyse the benefits of each cluster and visualise again to have a clearly view about our solutions and recommends.

IV. Analysis

Analyse Venues

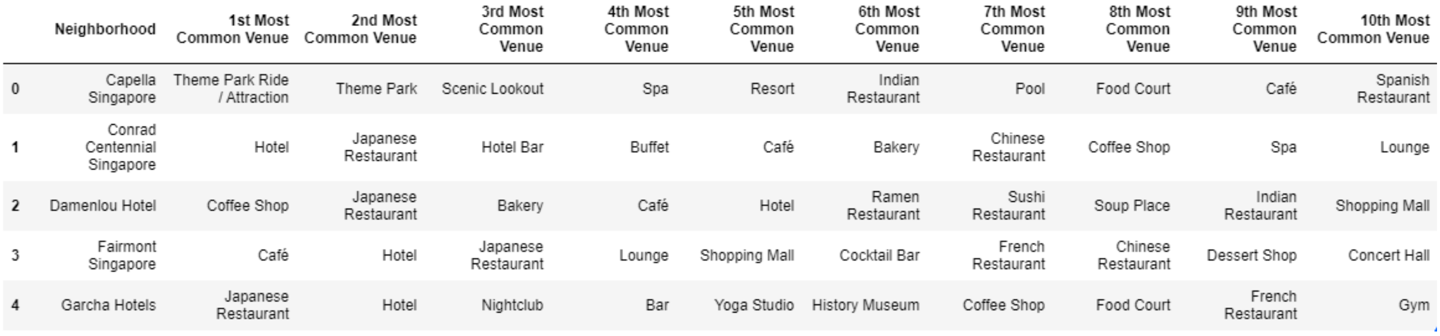
We obtained the number of existing facilities and theis type and location in every hotal area with Foursquared API with a limit 100 venues and within 500 meters. The columns Venue name, Venue Latitude and Longitude and Category information is crawled from Foursquared API.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Naumi Hotel	1.29577	103.85509	JW Marriott Hotel Singapore South Beach	1.294469	103.855793	Hotel
1	Naumi Hotel	1.29577	103.85509	Tom's Palette	1.296079	103.856757	Ice Cream Shop
2	Naumi Hotel	1.29577	103.85509	Raffles Hotel	1.294722	103.854090	Hotel
3	Naumi Hotel	1.29577	103.85509	South Beach	1.294976	103.856496	Shopping Mall
4	Naumi Hotel	1.29577	103.85509	Fairmont Singapore	1.294261	103.853931	Hotel

We can also check how many venues were returned and grouped rows by each neighborhood and taking the mean of the frequency of occurrence of each category.

Print each Neighborhood with the top 10 most common venues and put in a new DataFrame.

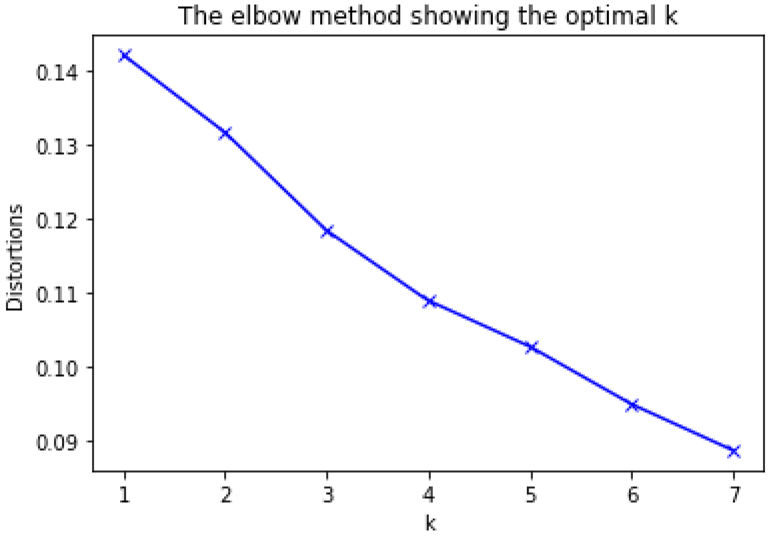


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Capella Singapore	Theme Park Ride / Attraction	Theme Park	Scenic Lookout	Spa	Resort	Indian Restaurant	Pool	Food Court	Café	Spanish Restaurant
1	Conrad Centennial Singapore	Hotel	Japanese Restaurant	Hotel Bar	Buffet	Café	Bakery	Chinese Restaurant	Coffee Shop	Spa	Lounge
2	Damenlou Hotel	Coffee Shop	Japanese Restaurant	Bakery	Café	Hotel	Ramen Restaurant	Sushi Restaurant	Soup Place	Indian Restaurant	Shopping Mall
3	Fairmont Singapore	Café	Hotel	Japanese Restaurant	Lounge	Shopping Mall	Cocktail Bar	French Restaurant	Chinese Restaurant	Dessert Shop	Concert Hall
4	Garcha Hotels	Japanese Restaurant	Hotel	Nightclub	Bar	Yoga Studio	History Museum	Coffee Shop	Food Court	French Restaurant	Gym

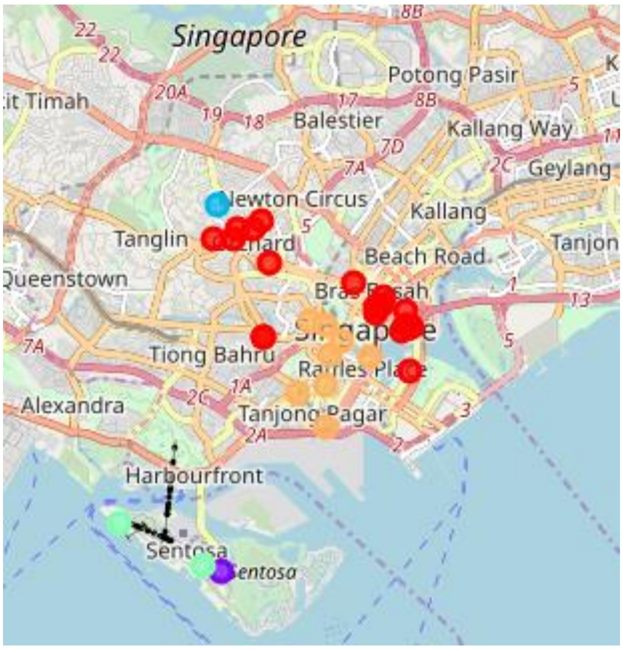
Cluster Neighborhood

According to the above data, using K-Means algorithm to cluster hotel and analysis the advantages of each Neighborhood to help visitor choose the best suitable hotal area for themselves.

First of this step, using Elbow method to define the best k value. From the graph, I chose 5 cluster for optimum k of K-Means.



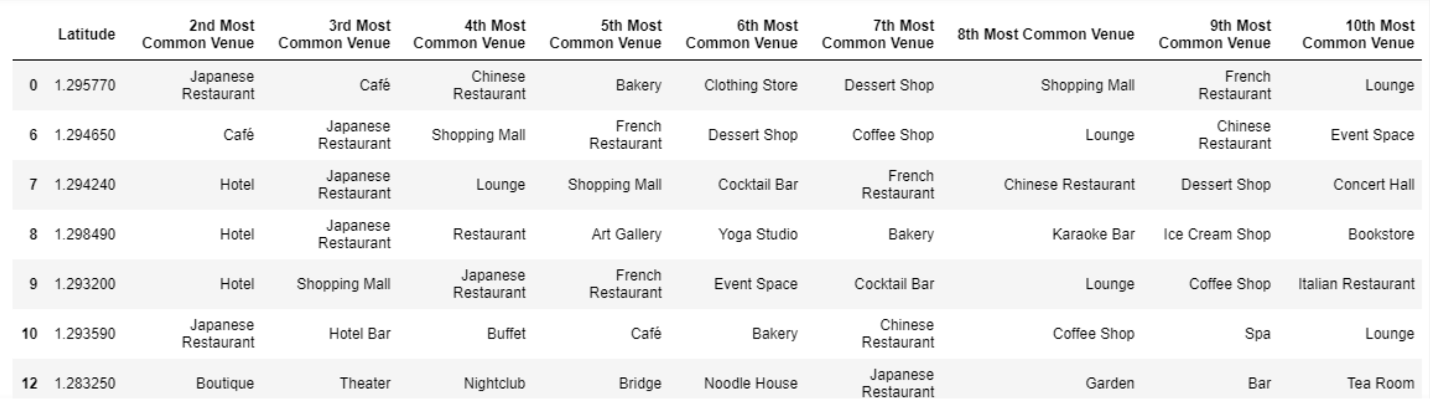
Then, using Folium again to visualise our 5 clusters, which should be a starting point for tourists to explore and search for ideal hotel area for their trip to Singapore.



Examine Cluster

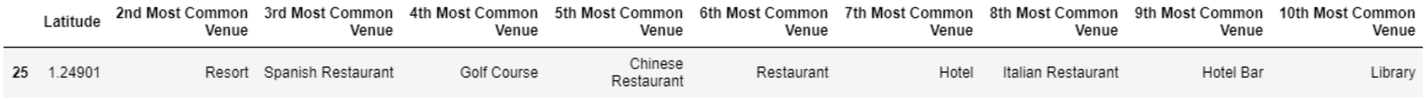
Applied K-Means Algorithm, all affordable hotels were devide into 5 clusters.

Cluster 1. Contains 18 hotels, top 10 common venues mainly conclude Shopping Mall, Restaurant, Barkery, Bar, Buffet,.. This cluster may in the downtown center of Singapore, which has many suitable destination for tourists to explore food, culture and enjoy shopping.



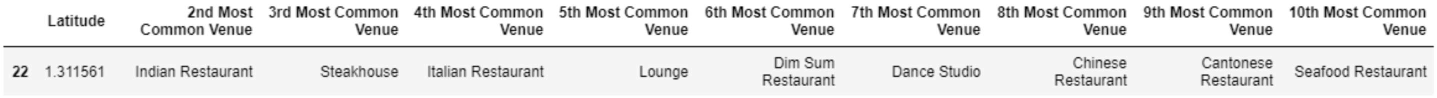
	Latitude	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	1.295770	Japanese Restaurant	Café	Chinese Restaurant	Bakery	Clothing Store	Dessert Shop	Shopping Mall	French Restaurant	Lounge
6	1.294650	Café	Japanese Restaurant	Shopping Mall	French Restaurant	Dessert Shop	Coffee Shop	Lounge	Chinese Restaurant	Event Space
7	1.294240	Hotel	Japanese Restaurant	Lounge	Shopping Mall	Cocktail Bar	French Restaurant	Chinese Restaurant	Dessert Shop	Concert Hall
8	1.298490	Hotel	Japanese Restaurant	Restaurant	Art Gallery	Yoga Studio	Bakery	Karaoke Bar	Ice Cream Shop	Bookstore
9	1.293200	Hotel	Shopping Mall	Japanese Restaurant	French Restaurant	Event Space	Cocktail Bar	Lounge	Coffee Shop	Italian Restaurant
10	1.293590	Japanese Restaurant	Hotel Bar	Buffet	Café	Bakery	Chinese Restaurant	Coffee Shop	Spa	Lounge
12	1.283250	Boutique	Theater	Nightclub	Bridge	Noodle House	Japanese Restaurant	Garden	Bar	Tea Room

Cluster 2. Contains just 1 hotel location, which is near retaurant, resort and golf course… This cluster is suitable for visitor who love to relax and enjoy a soft ourdoor activity.



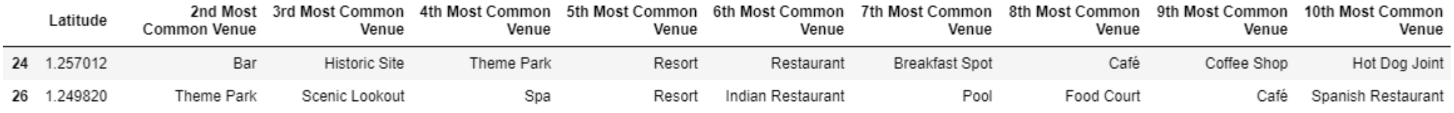
	Latitude	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
25	1.24901	Resort	Spanish Restaurant	Golf Course	Chinese Restaurant	Restaurant	Hotel	Italian Restaurant	Hotel Bar	Library

Cluster 3. Also contains only 1 hotel locations and mainly surround by restaurants. Any travelers enjoying in food from many cultures can choose to be here for their trip, as they are easy to reach any restaurants nearby their hotels.



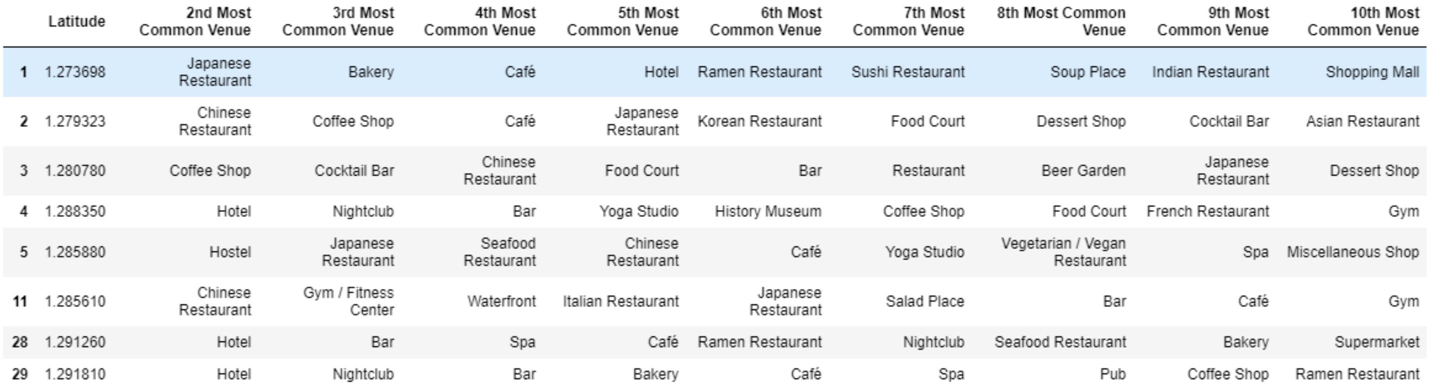
	Latitude	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
22	1.311561	Indian Restaurant	Steakhouse	Italian Restaurant	Lounge	Dim Sum Restaurant	Dance Studio	Chinese Restaurant	Cantonese Restaurant	Seafood Restaurant

Cluster 4. Contains only 2 hotels, srrouns by park, drink stores, resort, spa.. same as cluster 2. This cluster may be a good choice for tourist who love relaxing and soft and slow activities.



	Latitude	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
24	1.257012	Bar	Historic Site	Theme Park	Resort	Restaurant	Breakfast Spot	Café	Coffee Shop	Hot Dog Joint
26	1.249820	Theme Park	Scenic Lookout	Spa	Resort	Indian Restaurant	Pool	Food Court	Café	Spanish Restaurant

Cluster 5. Contains 8 hotels locations with many uitilities nearby like Restaurant, coffee shop, museum, yoga and gym, also nightclub and spa,.. same as cluster 1, this cluster may be around downtown Singapore, which related many entertainment places for active tourists.



	Latitude	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
1	1.273698	Japanese Restaurant	Bakery	Café	Hotel	Ramen Restaurant	Sushi Restaurant	Soup Place	Indian Restaurant	Shopping Mall
2	1.279323	Chinese Restaurant	Coffee Shop	Café	Japanese Restaurant	Korean Restaurant	Food Court	Dessert Shop	Cocktail Bar	Asian Restaurant
3	1.280780	Coffee Shop	Cocktail Bar	Chinese Restaurant	Food Court	Bar	Restaurant	Beer Garden	Japanese Restaurant	Dessert Shop
4	1.288350	Hotel	Nightclub	Bar	Yoga Studio	History Museum	Coffee Shop	Food Court	French Restaurant	Gym
5	1.285880	Hostel	Japanese Restaurant	Seafood Restaurant	Chinese Restaurant	Café	Yoga Studio	Vegetarian / Vegan Restaurant	Spa	Miscellaneous Shop
11	1.285610	Chinese Restaurant	Gym / Fitness Center	Waterfront	Italian Restaurant	Japanese Restaurant	Salad Place	Bar	Café	Gym
28	1.291260	Hotel	Bar	Spa	Café	Ramen Restaurant	Nightclub	Seafood Restaurant	Bakery	Supermarket
29	1.291810	Hotel	Nightclub	Bar	Bakery	Café	Spa	Pub	Coffee Shop	Ramen Restaurant

V. Results and Discussion

Although there are many hotels and homestay in Singapore, I just choosing 30 hotel to explore the suitable hotel area for travelers. Many of them is in downtown Singapore, which is suitable for tourist to reach famous views and entertainment destinations.

As mentioned above in examine clusters, I already gave comments for each cluster and recommend for types of tourists. So that, they can consider my comments to choose hotel area, suitable with their needs for a trip to the heart country of Asia.

VI. Conclusion

Purpose of this project is to find ideal hotels for travelers to Singapore. An ideal hotel is to allow tourist to reach their needed destinations to enjoy and explore Singapore in their trips. Target audience is international tourists, but might be person who in business trip can interest in. As the result showed that all of hotels in dataset has surround by many utilities such as restaurants, markets, coffee shop, parks, bar, resort…

The Foursquared location data was leveraged to compared each hotel location to provide believable suggestions for tourists. With the help of K-Means Clustering algorithm., all hotels were clustered into 5 groups. And each group’s benefits canbe clarified easily.

Further analysis can be done with these 5 cluster, but may use additional dataset related to Singapore facilities to explore more in details based on tourists’ needs.