IBM Data Science Capstone Project

Singapore Venues Cluster Analysis

by Agratama Arfiano

1. Introduction

a. Background

Singapore is a small country where the center of economic development in Southeast Asia happens, despite its limited land area. Urban Redevelopment Authority as a government agency that planned area zoning in all Singapore is rigid about how zoning area works. Therefore, the zoning area is relatively neatly arranged with a lot of raw data that publicly available on the internet.

b. Business Problem

An international private commercial developer wants to expand their hotel business to Singapore. Before they started a new project, they want to do a comprehensive analysis on the property around Singapore to find the center of tourist area that potential to be their new development location. They also want to analyze all type of amenities that already available in every area within Singapore to understand where are certain type of amenities located, while able to bring new unique amenities to those areas.

Taking into account the type of facilities at which the development will be done, the intent is to find a characteristic on certain areas for the development program to adapt to the surrounding facilities.

c. Audience

While this project is addressed to the private developer, anyone who wants to start a business within Singapore could be benefitted from the analysis report that helped them to categorized what kind of business that common in the certain areas.

2. Data Acquisition

Before start the analysis, we need to know the list of data that needed and where the data will be sourced:

- **List of neighborhood in Singapore**

  We can find the list of areas by scraping the data from Wikipedia

  https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore#List_of_Planning_Areas

- **Coordinates of areas in Singapore**

  We will used the geopy modules to search each of the areas to find the coordinates

- **List of all venues on each area**

  List of all venues will be obtained from Foursquare API

3. Methodology

a. Use of Data

First, we need to collect all data and merging it into one dataset. After cleaning and exploring the data, we will use clustering machine learning approach using K-means technique to create a cluster of venues with silhouette score for choosing the optimal number of clusters.

b. Data Preprocessing

We start our analysis by creating a list of Singapore Planning areas by scraping it from Wikipedia using BeautifulSoup library and dropped all irrelevant data such as area size, population, and density. From the dataset, we can see 55 Planning Areas/ Neighborhood and categorized into 5 Region.

	Planning Areas	Region
0	Ang Mo Kio	North-East
1	Bedok	East
2	Bishan	Central
3	Boon Lay	West
4	Bukit Batok	West
5	Bukit Merah	Central
6	Bukit Panjang	West
7	Bukit Timah	Central
8	Central Water Catchment	North
9	Changi	East

Figure 1. Table of Planning Areas & Region

After that, we attached geo-coordinates of each area from Wikipedia using geopy library. We can see on the new dataset that consists of new column of longitude and latitude coordinates of each Planning Areas.

	Planning Areas	Region	Latitude	Longitude
0	Ang Mo Kio	North-East	1.370080	103.849523
1	Bedok	East	1.323976	103.930216
2	Bishan	Central	1.350986	103.848255
3	Boon Lay	West	1.338550	103.705812
4	Bukit Batok	West	1.349057	103.749591
5	Bukit Merah	Central	1.270439	103.828318
6	Bukit Panjang	West	1.379149	103.761413
7	Bukit Timah	Central	1.354690	103.776372
8	Central Water Catchment	North	1.375708	103.801743
9	Changi	East	1.351080	103.990064

Figure 2. Table of Coordinates on each Planning Areas

Next, we visualize the representation of each Planning Areas into Singapore map using folium library

Figure 3. Plot map of every Planning Areas coordinates

Next step we explore venues on each Planning Areas using Foursquare API. Then we got location-based data on different sorts of venues. Foursquare indexes each location through the longitude and latitude of that location and provides the data based on the endpoints selected. The data from foursquare is received in JSON format. Therefore, we need to do some data cleaning to make it become a usable dataset.

After data cleaning, we convert the JSON format into pandas DataFrame. The data will consist of up to 100 venues for each Planning Areas within 1000m of the center coordinates that we have earlier.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Ang Mo Kio	1.37008	103.849523	Old Chang Kee	1.369094	103.848389	Snack Place
1	Ang Mo Kio	1.37008	103.849523	FairPrice Xtra	1.369279	103.848886	Supermarket
2	Ang Mo Kio	1.37008	103.849523	MOS Burger	1.369170	103.847831	Burger Joint
3	Ang Mo Kio	1.37008	103.849523	Face Ban Mian 非板面 (Ang Mo Kio)	1.372031	103.847504	Noodle House
4	Ang Mo Kio	1.37008	103.849523	NTUC FairPrice	1.371507	103.847082	Supermarket
5	Ang Mo Kio	1.37008	103.849523	Bangkok Street Mookata	1.365688	103.853186	BBQ Joint
6	Ang Mo Kio	1.37008	103.849523	A&W	1.369541	103.849043	Fast Food Restaurant
7	Ang Mo Kio	1.37008	103.849523	ST31 Coffee Shop	1.367478	103.848334	Coffee Shop
8	Ang Mo Kio	1.37008	103.849523	Aramsa ~ The Garden Spa	1.362292	103.847602	Spa
9	Ang Mo Kio	1.37008	103.849523	Bishan - Ang Mo Kio Park	1.362219	103.846250	Park

Figure 4. Table of Singapore venues

We also check how many venues have been collected for each Planning Areas. For our case, we need to drop one Planning Areas, which is the Northern Island areas, as the areas don't have any Foursquare venues.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Ang Mo Kio	90	90	90	90	90	90
Bedok	95	95	95	95	95	95
Bishan	69	69	69	69	69	69
Boon Lay	63	63	63	63	63	63
Bukit Batok	45	45	45	45	45	45
Bukit Merah	57	57	57	57	57	57
Bukit Panjang	55	55	55	55	55	55
Bukit Timah	30	30	30	30	30	30
Central Water Catchment	5	5	5	5	5	5
Changi	30	30	30	30	30	30

Figure 5. Table of venues count

Before doing further analysis, we will be focusing on the venues categories. We use the one-hot encoding method to create a dummy variable for each category, so the data can be prepared for further machine learning analysis. The result of one-hot encoding is the following dataset with the top ten most common venues for each district.

	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	Ang Mo Kio	Food Court	Coffee Shop	Fast Food Restaurant	Chinese Restaurant	Supermarket	Japanese Restaurant	Grocery Store	Dessert Shop	Bubble Tea Shop	Bus Stop
1	Bedok	Coffee Shop	Café	Chinese Restaurant	Food Court	Supermarket	Sandwich Place	Japanese Restaurant	Noodle House	Asian Restaurant	Fast Food Restaurant
2	Bishan	Food Court	Coffee Shop	Seafood Restaurant	Chinese Restaurant	Asian Restaurant	Bubble Tea Shop	Thai Restaurant	Café	BBQ Joint	Grocery Store
3	Boon Lay	Japanese Restaurant	Asian Restaurant	Fast Food Restaurant	Chinese Restaurant	Dessert Shop	Food Court	Coffee Shop	Café	Indian Restaurant	Karaoke Bar
4	Bukit Batok	Food Court	Coffee Shop	Fast Food Restaurant	Chinese Restaurant	Grocery Store	Malay Restaurant	Supermarket	Sandwich Place	Stadium	Lottery Retailer
5	Bukit Merah	Bus Station	Clothing Store	Japanese Restaurant	Coffee Shop	Multiplex	Food Court	Fast Food Restaurant	Chinese Restaurant	Department Store	Toy / Game Store
6	Bukit Panjang	Coffee Shop	Asian Restaurant	Fast Food Restaurant	Supermarket	Sushi Restaurant	Indonesian Restaurant	Bus Station	Park	Shopping Mall	Gym
7	Bukit Timah	Trail	Scenic Lookout	Rest Area	Nature Preserve	Park	Vegetarian / Vegan Restaurant	Lake	Tourist Information Center	Australian Restaurant	Arts & Crafts Store
8	Central Water Catchment	Café	Gift Shop	Reservoir	Food Court	Business Service	Farm	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant
9	Changi	Airport	Airport Lounge	Airport Service	Bubble Tea Shop	Border Crossing	Road	Men's Store	Supermarket	Gift Shop	General Entertainment

Figure 6. Top 10 common venues

c. Clustering

After the dataset is ready, we could perform clustering with unsupervised machine learning techniques based on K-means. By using K-means technique, we need to find the optimal number of clusters that will be used for further analysis. To speed up the iteration process, we used silhouette score to find out the best score for each number of clusters.

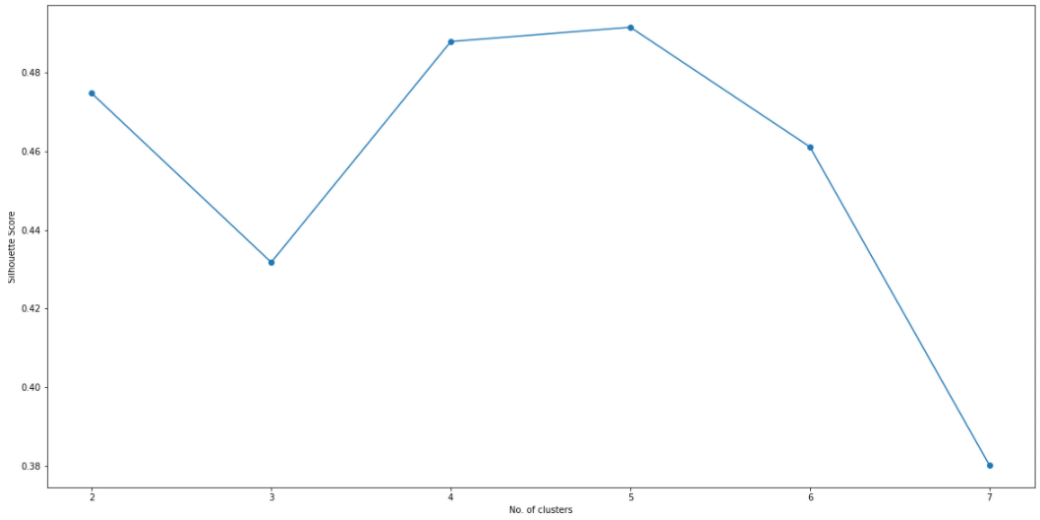

Figure 7. Silhouette Scores

From the graph, we can see that the optimal number cluster is 5, where the score is the highest. In the next step, we run the K-means clustering algorithm with the parameter of 5 as the number of clusters.

Next, we insert the cluster labels into our current dataset

	Neighborhood	Latitude	Longitude	Cluster Labels	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	Ang Mo Kio	1.370080	103.849523	0	Food Court	Coffee Shop	Fast Food Restaurant	Chinese Restaurant	Supermarket	Japanese Restaurant	Grocery Store	Dessert Shop	Bubble Tea Shop	Bus Stop
1	Bedok	1.323976	103.930216	0	Coffee Shop	Café	Chinese Restaurant	Food Court	Supermarket	Sandwich Place	Japanese Restaurant	Noodle House	Asian Restaurant	Fast Food Restaurant
2	Bishan	1.350986	103.848255	0	Food Court	Coffee Shop	Seafood Restaurant	Chinese Restaurant	Asian Restaurant	Bubble Tea Shop	Thai Restaurant	Café	BBQ Joint	Grocery Store
3	Boon Lay	1.338550	103.705812	0	Japanese Restaurant	Asian Restaurant	Fast Food Restaurant	Chinese Restaurant	Dessert Shop	Food Court	Coffee Shop	Café	Indian Restaurant	Karaoke Bar
4	Bukit Batok	1.349057	103.749591	0	Food Court	Coffee Shop	Fast Food Restaurant	Chinese Restaurant	Grocery Store	Malay Restaurant	Supermarket	Sandwich Place	Stadium	Lottery Retailer
5	Bukit Merah	1.270439	103.828318	0	Bus Station	Clothing Store	Japanese Restaurant	Coffee Shop	Multiplex	Food Court	Fast Food Restaurant	Chinese Restaurant	Department Store	Toy / Game Store
6	Bukit Panjang	1.379149	103.761413	0	Coffee Shop	Asian Restaurant	Fast Food Restaurant	Supermarket	Sushi Restaurant	Indonesian Restaurant	Bus Station	Park	Shopping Mall	Gym
7	Bukit Timah	1.354690	103.776372	2	Trail	Scenic Lookout	Rest Area	Nature Preserve	Park	Vegetarian / Vegan Restaurant	Lake	Tourist Information Center	Australian Restaurant	Arts & Crafts Store
8	Central Water Catchment	1.375708	103.801743	0	Café	Gift Shop	Reservoir	Food Court	Business Service	Farm	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant
9	Changi	1.351080	103.990064	3	Airport	Airport Lounge	Airport Service	Bubble Tea Shop	Border Crossing	Road	Men's Store	Supermarket	Gift Shop	General Entertainment

Figure 8. Table with cluster labels

We also use folium to help us visualize how the venues cluster spread across Singapore

Figure 9. Plot map with 5 different clusters

4. Results

By exploring at each cluster dataset, we can understand the characteristic of each cluster of venues

Cluster 0

	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	Ang Mo Kio	Food Court	Coffee Shop	Fast Food Restaurant	Chinese Restaurant	Supermarket	Japanese Restaurant	Grocery Store	Dessert Shop	Bubble Tea Shop	Bus Stop
1	Bedok	Coffee Shop	Café	Chinese Restaurant	Food Court	Supermarket	Sandwich Place	Japanese Restaurant	Noodle House	Asian Restaurant	Fast Food Restaurant
2	Bishan	Food Court	Coffee Shop	Seafood Restaurant	Chinese Restaurant	Asian Restaurant	Bubble Tea Shop	Thai Restaurant	Café	BBQ Joint	Grocery Store
3	Boon Lay	Japanese Restaurant	Asian Restaurant	Fast Food Restaurant	Chinese Restaurant	Dessert Shop	Food Court	Coffee Shop	Café	Indian Restaurant	Karaoke Bar
4	Bukit Batok	Food Court	Coffee Shop	Fast Food Restaurant	Chinese Restaurant	Grocery Store	Malay Restaurant	Supermarket	Sandwich Place	Stadium	Lottery Retailer
6	Bukit Panjang	Coffee Shop	Asian Restaurant	Fast Food Restaurant	Supermarket	Sushi Restaurant	Indonesian Restaurant	Bus Station	Park	Shopping Mall	Gym
11	Choa Chu Kang	Coffee Shop	Fast Food Restaurant	Food Court	Café	Noodle House	Supermarket	Chinese Restaurant	Gym	Sandwich Place	Thai Restaurant
12	Clementi	Food Court	Coffee Shop	Chinese Restaurant	Asian Restaurant	Indian Restaurant	Dessert Shop	Bakery	Supermarket	Café	Gym
14	Geylang	Chinese Restaurant	Noodle House	Asian Restaurant	Food Court	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Supermarket	Steakhouse	Seafood Restaurant	Dessert Shop
15	Hougang	Coffee Shop	Food Court	Chinese Restaurant	Fast Food Restaurant	Asian Restaurant	Supermarket	Café	Vegetarian / Vegan Restaurant	Playground	Metro Station

Figure 10. Cluster 0

Cluster 0 is the biggest cluster with 27 neighborhoods and dominated by Food Court, Coffee Shop & Thematic Restaurant

Cluster 1

	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
10	Changi Bay	Boat or Ferry	Harbor / Marina	Pizza Place	Yoga Studio	Food	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant	Fish & Chips Shop
51	Western Islands	Resort	Harbor / Marina	Boat or Ferry	Scenic Lookout	Yoga Studio	Filipino Restaurant	Exhibit	Farm	Farmers Market	Fast Food Restaurant
52	Western Water Catchment	Gun Range	Gym	Train Station	Yoga Studio	Fish & Chips Shop	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant	Flea Market

Figure 11. Cluster 1

Cluster 1 contain 3 neighborhoods and dominated by venues type towards waterfront activities such as Boat/Ferry, Resort, Pool, Harbor / Marina.

Cluster 2

	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
7	Bukit Timah	Trail	Scenic Lookout	Rest Area	Nature Preserve	Lake	Military Base	Australian Restaurant	Park	Arts & Crafts Store	Vegetarian / Vegan Restaurant
43	Southern Islands	Trail	Dry Cleaner	Yoga Studio	Flea Market	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant	Fish & Chips Shop	Food

Figure 12. Cluster 2

Cluster 2 is the smallest cluster with only 2 neighborhoods and dominated by venues type towards natural & physical activities such as hiking trail, scenic lookout, park, lake, etc.

Cluster 3

	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
5	Bukit Merah	Bus Station	Clothing Store	Food Court	Coffee Shop	Japanese Restaurant	Multiplex	Fast Food Restaurant	Chinese Restaurant	Department Store	Toy / Game Store
9	Changi	Airport	Airport Lounge	Airport Service	Border Crossing	Bubble Tea Shop	Road	Jewelry Store	General Entertainment	Sandwich Place	Movie Theater
13	Downtown Core	Hotel	Waterfront	Italian Restaurant	Event Space	Japanese Restaurant	Boutique	Lounge	Buffet	Dim Sum Restaurant	Theater
16	Jurong East	Chinese Restaurant	Café	Japanese Restaurant	Coffee Shop	Food Court	Shopping Mall	Clothing Store	Bus Station	Korean Restaurant	Multiplex
19	Lim Chu Kang	Farm	Zoo Exhibit	Vegetarian / Vegan Restaurant	Theme Park Ride / Attraction	Fish & Chips Shop	Exhibit	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant
20	Mandai	Cosmetics Shop	Speakeasy	Halal Restaurant	Trail	Chinese Restaurant	Yoga Studio	Fish & Chips Shop	Farm	Farmers Market	Fast Food Restaurant
21	Marina East	Scenic Lookout	Garden	Racetrack	Seafood Restaurant	Art Gallery	Golf Course	Botanical Garden	Waterfront	Bridge	Park
22	Marina South	Japanese Restaurant	Garden	Hotel	Boutique	Coffee Shop	Scenic Lookout	Waterfront	Bridge	Boat or Ferry	Theater
23	Marine Parade	Chinese Restaurant	Noodle House	Hotel	Indian Restaurant	Asian Restaurant	Coffee Shop	Japanese Restaurant	Bar	Multiplex	Massage Studio
24	Museum	Hotel	Japanese Restaurant	Café	Shopping Mall	Coffee Shop	Restaurant	Bubble Tea Shop	Movie Theater	Korean Restaurant	Steakhouse

Figure 13. Cluster 3

Cluster 3 contain 20 neighborhoods and dominated by facilities that support tourist activities such as Airport, Hotel, Cafe, Restaurant, Boutique, etc.

Cluster 4

	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
8	Central Water Catchment	Café	Gift Shop	Reservoir	Business Service	Food Court	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant	Fish & Chips Shop
37	Seletar	Café	Gastropub	Airport Service	Airport Terminal	Harbor / Marina	Resort	Boat or Ferry	Yoga Studio	Flea Market	Fast Food Restaurant
44	Straits View	Cruise Ship	Diner	Café	Pier	Food	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant	Fish & Chips Shop

Figure 14. Cluster 4

Cluster 4 contain 3 neighborhoods and dominated by venues type towards cafe and cruise ship-oriented activities

5. Discussion & Recommendation

With every neighborhood, the cluster has its unique category of venues. We advised the developer to consider the neighborhood from cluster 3 as a potential location for their new hotel project. These are the neighborhood where tourist hotspot located with their supporting amenities such as hotel and restaurant are frequent.

6. Conclusion

This project resulting in information that answers the audience's business problem. The analysis was performed based on the toolset of data science and relied heavily on the use of Python and Python libraries such as Pandas, Scikit, Folium to name a few. Data was collected from different types of sources and in different formats. For analysis, the machine learning clustering technique was used. The output of the analysis provided a thorough for the business problem.