Capstone Project: Singapore and Kuala Lumpur Compared

Purpose

This exercise serves as the document for the peer review assignment for the IBM Data Science Professional Certificate - Applied Data Science Capstone.

Introduction

This exercise demonstrates the use of the Foursquare API and the k-means machine learning algorithm.

The objectives of this exercise was to illustrate the qualitative differences of two neighborhoods in two cities – Singapore and Kuala Lumpur, Malaysia. This exercise could presumably be helpful in two ways: a guide for potential visitors of these two cities on the kinds of places they offer and the food scenes these two places offer; and for potential restauranteurs in thinking about what kinds of dining options to offer in these two places.

Singapore and Kuala Lumpur are two cities that are often compared with each other. They are physically just 400km from each other - about four hours ride, and share several qualities: both are diverse multicultural cities where people of different ethnicities, faiths, and cultures live with each other. They have also developed their own food cultures, which often compete with each other. This exercise does not claim to be definititve in either way - merely using Foursquare data to look at how visitors interact with both cities on Foursquare, and how that data might reveal patterns about the food culture as experienced by users who check in on Foursquare.

The findings from this small limited study will still provide interesting nuggets of information for potential visitors and businesses to consider as they find out more about both cities.

Data

The data from this comes from the Foursquare API. Foursquare API calls offers information on the venue and category of the venue – which is sufficient information for the tasks to perform. The file format can be organized for further analysis in the pandas dataframe. The main fields of the Foursquare API that will be used would be:

- Venue Name
- Latitude
- Longitude
- Venue cateogry. For the geographical data, I have relied on Google searches to determine the approximate coordinates of the places of interest.

Methodology

The main tools I will be using will be K-means clustering, as I am trying to understand the qualitative nature of the various neighbourhoods within Singapore and Kuala Lumpur, and how these neighbourhoods would be distinctive in their own right. Such a description would be useful to potential visitors and restauranteurs and the places they might choose to visit and set up their businesses.

Owing to the small-scale nature of this exercise, this is certainly not a comprehensive analysis, and further data collection and analysis would be required. Nonetheless, this small demonstration would be sufficient to reveal the insights of these two cities and their constitutent neighbourhoods.

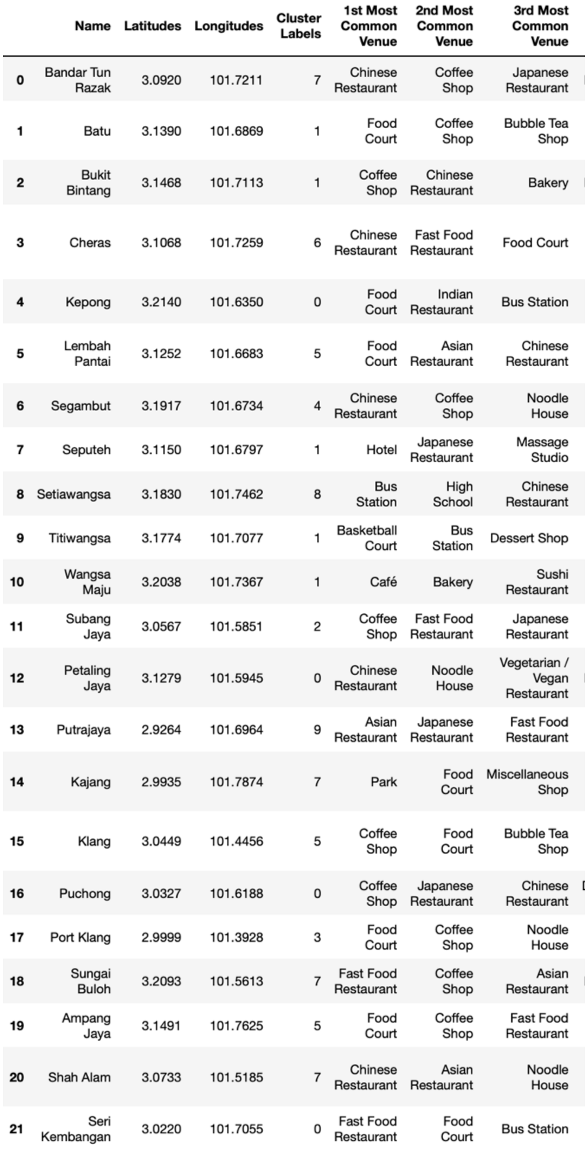
I will be collecting the names of the neighbourhoods, followed by their latitude and longitude coordinates for use in the Foursquare API. In situations where it is difficult to collect the latitude and longitude coordinates using the Geocoder API, I will manually obtain them through Google searches.

The coordinates will then be fed into the Foursquare Places API calls to collect information of the venues around the neighbourhoods. For ease of analysis I will collect up a limit of 50 places for neighbourhoods in these two cities.

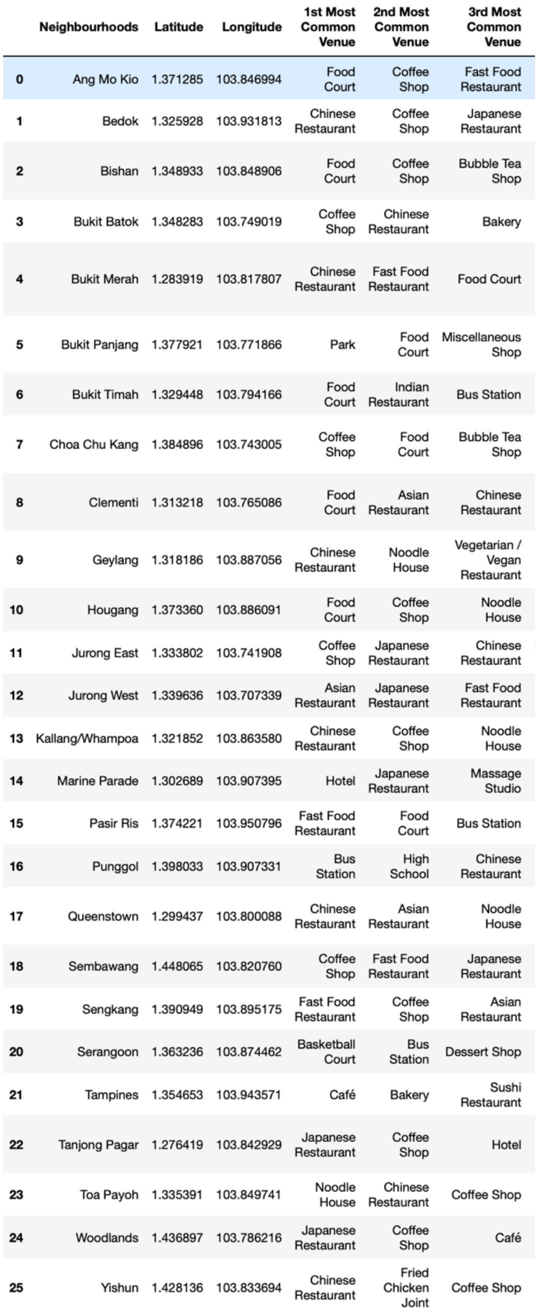
I will be collecting the coordinates for Singapore and Kuala Lumpur neighbourhoods/districts. I will then call the Foursquare API and obtain the venues for both places. I will convert the venues data into vectors through one-hot encoding. I will then cluster the neighbourhoods/districts in both places, and attempt to interpret the findings. I will also take a look at the most popular venues for both cities. All of these together, will provide an illustrative view of the food scenes in both places, and inform potential visitors and business owners as they make decisions on where to visit/put a restaurant in.

Results

Top 3 Most Common Venues in both Cities



	Name	Latitudes	Longitudes	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Bandar Tun Razak	3.0920	101.7211	7	Chinese Restaurant	Coffee Shop	Japanese Restaurant
1	Batu	3.1390	101.6869	1	Food Court	Coffee Shop	Bubble Tea Shop
2	Bukit Bintang	3.1468	101.7113	1	Coffee Shop	Chinese Restaurant	Bakery
3	Cheras	3.1068	101.7259	6	Chinese Restaurant	Fast Food Restaurant	Food Court
4	Kepong	3.2140	101.6350	0	Food Court	Indian Restaurant	Bus Station
5	Lembah Pantai	3.1252	101.6683	5	Food Court	Asian Restaurant	Chinese Restaurant
6	Segambut	3.1917	101.6734	4	Chinese Restaurant	Coffee Shop	Noodle House
7	Seputeh	3.1150	101.6797	1	Hotel	Japanese Restaurant	Massage Studio
8	Setiawangsa	3.1830	101.7462	8	Bus Station	High School	Chinese Restaurant
9	Titiwangsa	3.1774	101.7077	1	Basketball Court	Bus Station	Dessert Shop
10	Wangsa Maju	3.2038	101.7367	1	Café	Bakery	Sushi Restaurant
11	Subang Jaya	3.0567	101.5851	2	Coffee Shop	Fast Food Restaurant	Japanese Restaurant
12	Petaling Jaya	3.1279	101.5945	0	Chinese Restaurant	Noodle House	Vegetarian / Vegan Restaurant
13	Putrajaya	2.9264	101.6964	9	Asian Restaurant	Japanese Restaurant	Fast Food Restaurant
14	Kajang	2.9935	101.7874	7	Park	Food Court	Miscellaneous Shop
15	Klang	3.0449	101.4456	5	Coffee Shop	Food Court	Bubble Tea Shop
16	Puchong	3.0327	101.6188	0	Coffee Shop	Japanese Restaurant	Chinese Restaurant
17	Port Klang	2.9999	101.3928	3	Food Court	Coffee Shop	Noodle House
18	Sungai Buloh	3.2093	101.5613	7	Fast Food Restaurant	Coffee Shop	Asian Restaurant
19	Ampang Jaya	3.1491	101.7625	5	Food Court	Coffee Shop	Fast Food Restaurant
20	Shah Alam	3.0733	101.5185	7	Chinese Restaurant	Asian Restaurant	Noodle House
21	Seri Kembangan	3.0220	101.7055	0	Fast Food Restaurant	Food Court	Bus Station



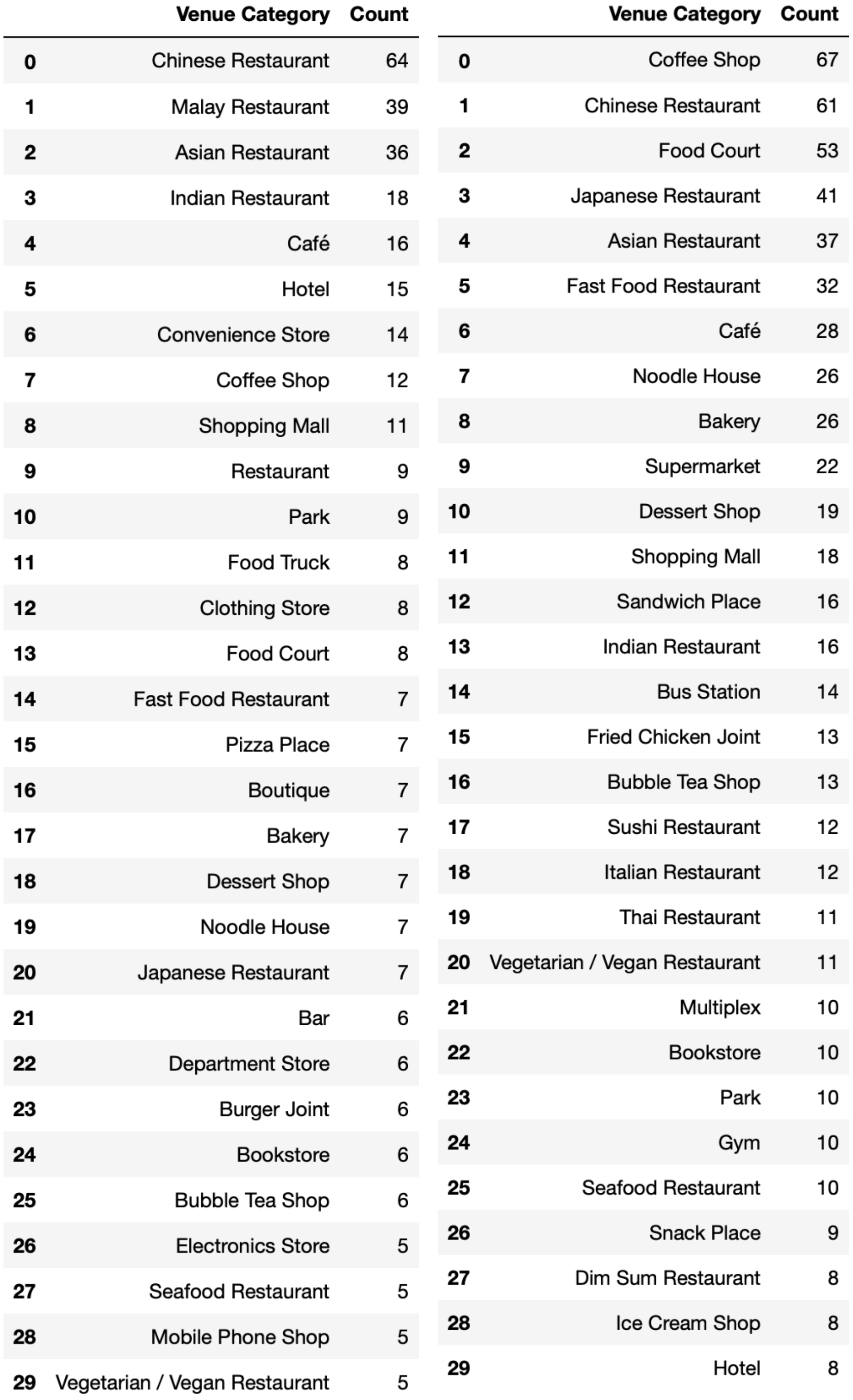
	Neighbourhoods	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Ang Mo Kio	1.371285	103.846994	Food Court	Coffee Shop	Fast Food Restaurant
1	Bedok	1.325928	103.931813	Chinese Restaurant	Coffee Shop	Japanese Restaurant
2	Bishan	1.348933	103.848906	Food Court	Coffee Shop	Bubble Tea Shop
3	Bukit Batok	1.348283	103.749019	Coffee Shop	Chinese Restaurant	Bakery
4	Bukit Merah	1.283919	103.817807	Chinese Restaurant	Fast Food Restaurant	Food Court
5	Bukit Panjang	1.377921	103.771866	Park	Food Court	Miscellaneous Shop
6	Bukit Timah	1.329448	103.794166	Food Court	Indian Restaurant	Bus Station
7	Choa Chu Kang	1.384896	103.743005	Coffee Shop	Food Court	Bubble Tea Shop
8	Clementi	1.313218	103.765086	Food Court	Asian Restaurant	Chinese Restaurant
9	Geylang	1.318186	103.887056	Chinese Restaurant	Noodle House	Vegetarian / Vegan Restaurant
10	Hougang	1.373360	103.886091	Food Court	Coffee Shop	Noodle House
11	Jurong East	1.333802	103.741908	Coffee Shop	Japanese Restaurant	Chinese Restaurant
12	Jurong West	1.339636	103.707339	Asian Restaurant	Japanese Restaurant	Fast Food Restaurant
13	Kallang/Whampoa	1.321852	103.863580	Chinese Restaurant	Coffee Shop	Noodle House
14	Marine Parade	1.302689	103.907395	Hotel	Japanese Restaurant	Massage Studio
15	Pasir Ris	1.374221	103.950796	Fast Food Restaurant	Food Court	Bus Station
16	Punggol	1.398033	103.907331	Bus Station	High School	Chinese Restaurant
17	Queenstown	1.299437	103.800088	Chinese Restaurant	Asian Restaurant	Noodle House
18	Sembawang	1.448065	103.820760	Coffee Shop	Fast Food Restaurant	Japanese Restaurant
19	Sengkang	1.390949	103.895175	Fast Food Restaurant	Coffee Shop	Asian Restaurant
20	Serangoon	1.363236	103.874462	Basketball Court	Bus Station	Dessert Shop
21	Tampines	1.354653	103.943571	Café	Bakery	Sushi Restaurant
22	Tanjong Pagar	1.276419	103.842929	Japanese Restaurant	Coffee Shop	Hotel
23	Toa Payoh	1.335391	103.849741	Noodle House	Chinese Restaurant	Coffee Shop
24	Woodlands	1.436897	103.786216	Japanese Restaurant	Coffee Shop	Café
25	Yishun	1.428136	103.833694	Chinese Restaurant	Fried Chicken Joint	Coffee Shop

(KL on the left, Singapore on the right)

Singapore Clusters:
The main way the algorithm has clustered is by the most common category of venues in each cluster. The main way the clusters have differed is by the difference in food shops. The biggest cluster is Cluster 3, which contains several housing estates, and other amenities.

KL Clusters:
The KL clusters revolve around restaurants - Chinese/Japanese/Asian/Fast Food, and coffee shops.

Popular Venues for both Cities



	Venue Category	Count
0	Chinese Restaurant	64
1	Malay Restaurant	39
2	Asian Restaurant	36
3	Indian Restaurant	18
4	Café	16
5	Hotel	15
6	Convenience Store	14
7	Coffee Shop	12
8	Shopping Mall	11
9	Restaurant	9
10	Park	9
11	Food Truck	8
12	Clothing Store	8
13	Food Court	8
14	Fast Food Restaurant	7
15	Pizza Place	7
16	Boutique	7
17	Bakery	7
18	Dessert Shop	7
19	Noodle House	7
20	Japanese Restaurant	7
21	Bar	6
22	Department Store	6
23	Burger Joint	6
24	Bookstore	6
25	Bubble Tea Shop	6
26	Electronics Store	5
27	Seafood Restaurant	5
28	Mobile Phone Shop	5
29	Vegetarian / Vegan Restaurant	5

	Venue Category	Count
0	Coffee Shop	67
1	Chinese Restaurant	61
2	Food Court	53
3	Japanese Restaurant	41
4	Asian Restaurant	37
5	Fast Food Restaurant	32
6	Café	28
7	Noodle House	26
8	Bakery	26
9	Supermarket	22
10	Dessert Shop	19
11	Shopping Mall	18
12	Sandwich Place	16
13	Indian Restaurant	16
14	Bus Station	14
15	Fried Chicken Joint	13
16	Bubble Tea Shop	13
17	Sushi Restaurant	12
18	Italian Restaurant	12
19	Thai Restaurant	11
20	Vegetarian / Vegan Restaurant	11
21	Multiplex	10
22	Bookstore	10
23	Park	10
24	Gym	10
25	Seafood Restaurant	10
26	Snack Place	9
27	Dim Sum Restaurant	8
28	Ice Cream Shop	8
29	Hotel	8

(KL on the left, Singapore on the right)

While both cities have similar constitutes in their venue category frequency counts, they differ significantly in the overall order. In both cities, restaurants and food outlets occupy the top spots, in Singapore supermarkets make an appearance, highlighting the high residential density there.

Spatial Distribution of Clusters for both cities

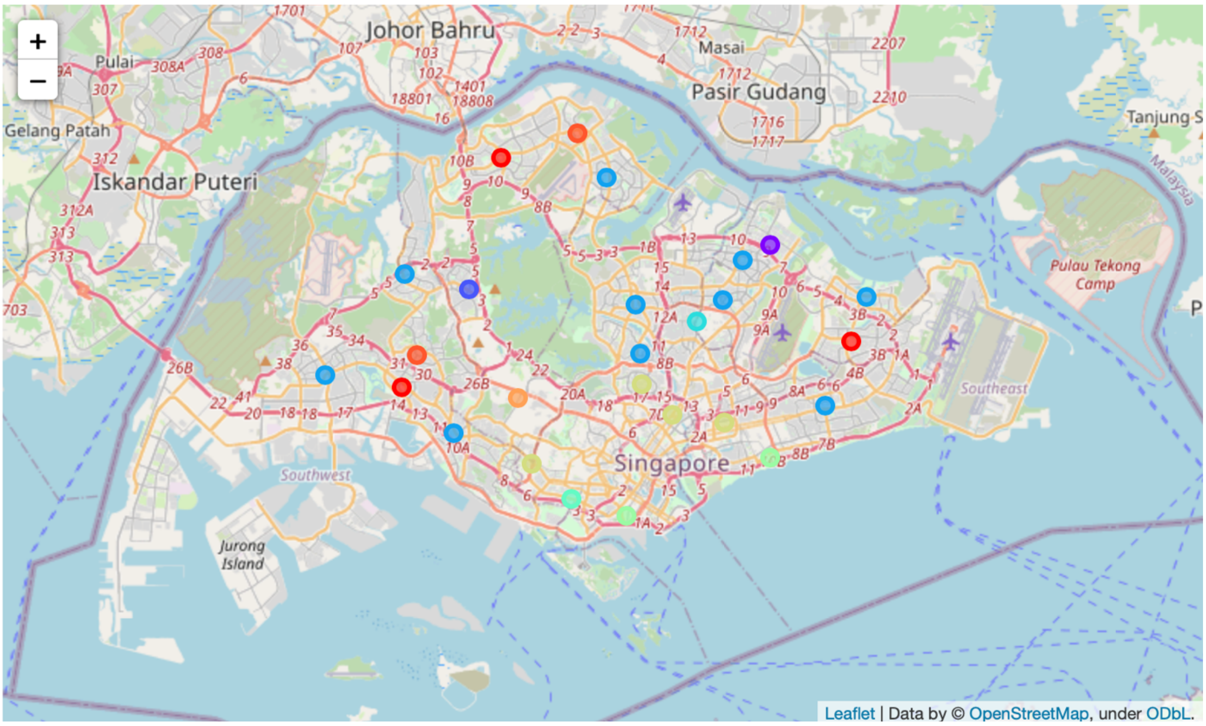


Figure 1 Distribution of Clusters for Singapore Neighbourhoods

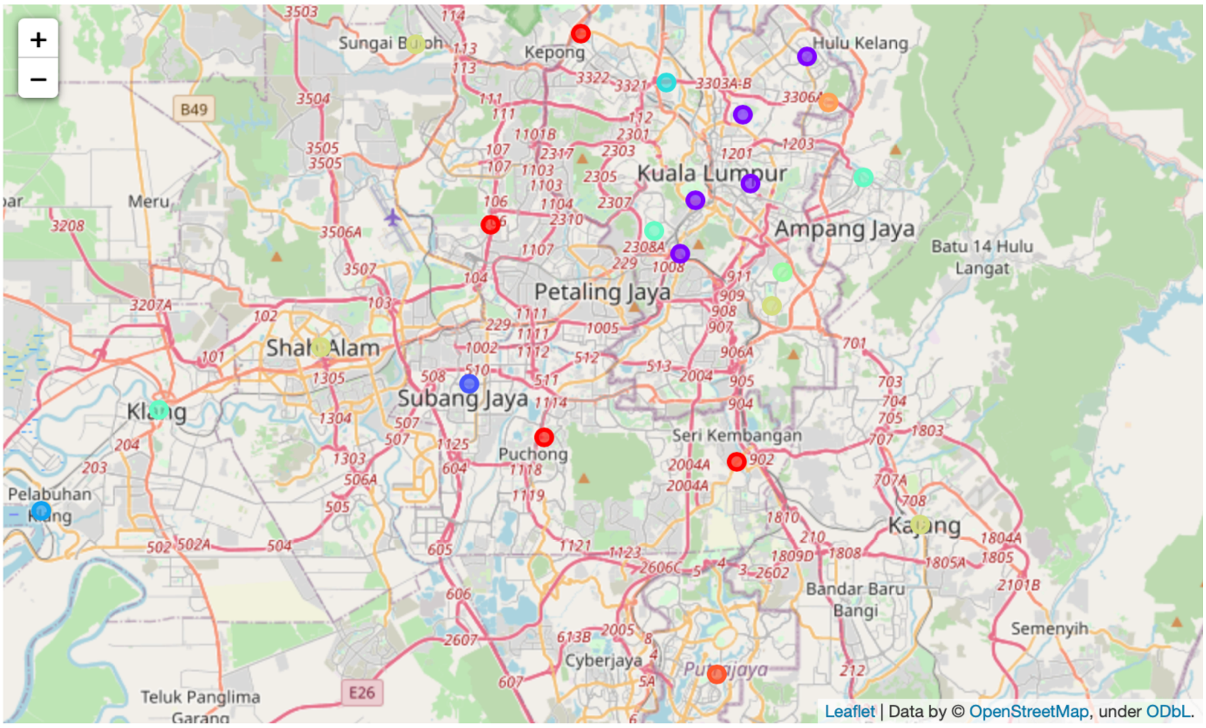


Figure 2 Distribution of Clusters for KL Districts/Neighbourhoods

KL's clusters are more spread out, owing to the larger spatial extent. Singapore’s clusters are more dense, owing to Singapore’s smaller size.

Discussion

Here are the main observations from this clustering analysis:

1. KL has 582 total venues from 163 categories. Singapore's figure was 908 venues from 149 categories. This alone would already be interesting.
2. Singapore's clusters are dominated by housing estates, with possibly Marine Parade and Tanjong Pagar as more unusual neighbourhoods.
3. KL's clusters are more heterogenous, with more obviously different neighbourhoods.
4. Singapore's common places has slightly higher number of food outlets, edging out Malaysia - by about 22 to 19. This suggests that the food landscape in dense Singapore might be more competitive than KL. This is what this dataset suggests, and refinements of this finding should be undertaken.
5. A visitor that likes food places should consider Singapore owing to the higher placements of food places, although KL could be more interesting as a location with more distinct clusters.
6. A potential restauranteur can consider KL to avoid competition, although there is a trade off when it comes to density. KL's sprawling nature means more traveling time for consumers when compared to Singapore.

Conclusion

This concludes the report for the IBM Data Science Professional Certificate - Applied Science Capstone module.

I have formulated objectives and utilised data sources relevant for the project - via Geopy, and from Google's API. I have used the API calls from Foursquare's Places API, and collected venue names, and venue categories, and processed them for analysis.

I have made use of k-means clustering, an unsupervised machine learning classification algorithm to classify the neighbourhoods. This approach allows for an initial exploration into the data collected that can be further refined through subsequent investigation ("Now we know what to look out for").