Singapore Travel Guide

AN EXERCISE IN CLUSTERING SINGAPORE NEIGHBOURHOODS

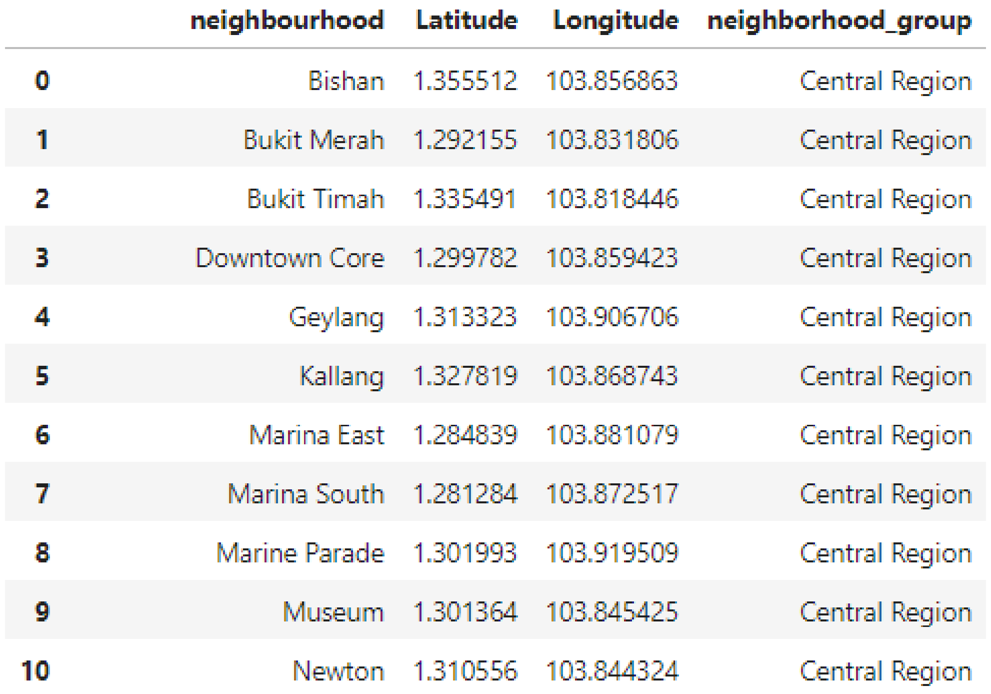
- BY ASHWIN MOHANANATH SYAMALA

Business Case

- There are numerous places to visit in Singapore falling under different categories that makes it hard for the average traveller to narrow down on the places that they would like to visit.
- Apart from the actual venues, often travellers have to think about accommodation which is an important part of any travellers itinerary.
- Travel agents and Tour operators can capitalise on this need if they can get tailored recommendations for travellers.
- K-means Clustering algorithm is used for this endeavour to cluster neighbourhoods in Singapore based on the Neighbourhoods, Venue categories and the no of Airbnb listings near the venues.

About the Data: Singapore neighbourhoods

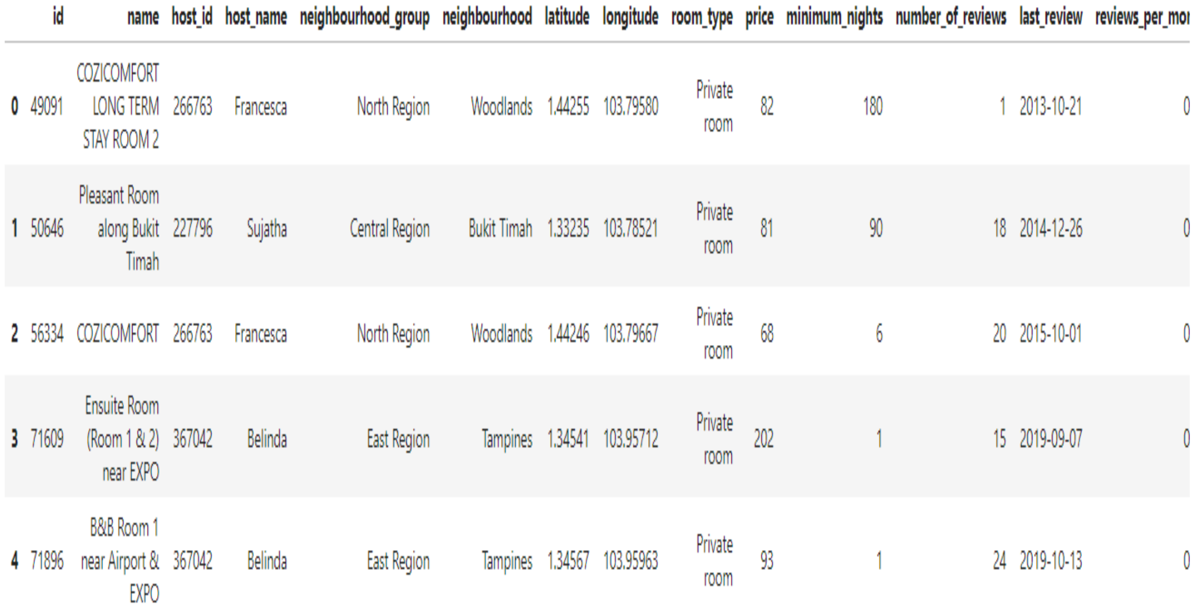
- Neighbourhood information of Singapore and the Number of Airbnb listings in each neighbourhood is available from the source **Inside Airbnb (http://insideairbnb.com/get-the-data.html)**
- The Neighbourhood dataset has information about the Neighbourhood group, Neighbourhood name, Neighbourhood Coordinates.



	neighbourhood	Latitude	Longitude	neighborhood_group
0	Bishan	1.355512	103.856863	Central Region
1	Bukit Merah	1.292155	103.831806	Central Region
2	Bukit Timah	1.335491	103.818446	Central Region
3	Downtown Core	1.299782	103.859423	Central Region
4	Geylang	1.313323	103.906706	Central Region
5	Kallang	1.327819	103.868743	Central Region
6	Marina East	1.284839	103.881079	Central Region
7	Marina South	1.281284	103.872517	Central Region
8	Marine Parade	1.301993	103.919509	Central Region
9	Museum	1.301364	103.845425	Central Region
10	Newton	1.310556	103.844324	Central Region

About the Data: Airbnb listings

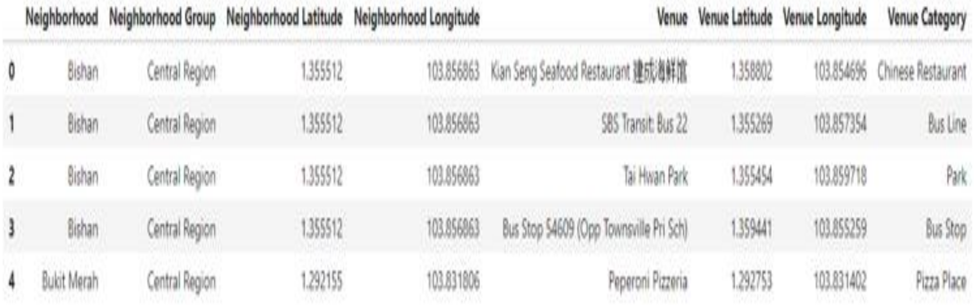
- Neighbourhood information of Singapore and the Number of Airbnb listings in each neighbourhood is available from the source **Inside Airbnb** (**http://insideairbnb.com/get-the-data.html**)
- The Airbnb dataset has information about the rental properties listed in the neighbourhoods along with the coordinates of the listings.
- The Airbnb dataset has also other information relating to the properties listed such as price, availability, review etc.



	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_mon
0	49091	COZICOMFORT LONG TERM STAY ROOM 2	266763	Francesca	North Region	Woodlands	1.44255	103.79580	Private room	82	180	1	2013-10-21	0
1	50646	Pleasant Room along Bukit Timah	227796	Sujatha	Central Region	Bukit Timah	1.33235	103.78521	Private room	81	90	18	2014-12-26	0
2	56334	COZICOMFORT	266763	Francesca	North Region	Woodlands	1.44246	103.79667	Private room	68	6	20	2015-10-01	0
3	71609	Ensuite Room (Room 1 & 2) near EXPO	367042	Belinda	East Region	Tampines	1.34541	103.95712	Private room	202	1	15	2019-09-07	0
4	71896	B&B Room 1 near Airport & EXPO	367042	Belinda	East Region	Tampines	1.34567	103.95963	Private room	93	1	24	2019-10-13	0

About the Data: Foursquare Data

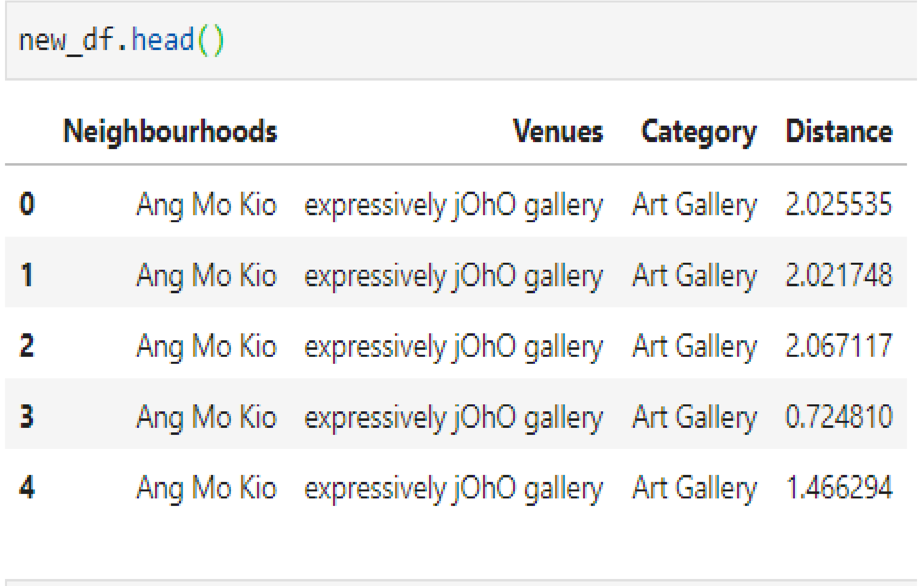
- Foursquare API calls to explore venues surrounding Singapore neighbourhoods is made and loaded into a dataframe.
- Results will contain name of the venue, the venue category, and the coordinates of the venues.



	Neighborhood	Neighborhood Group	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bishan	Central Region	1.355512	103.856863	Kian Seng Seafood Restaurant 建成海鲜馆	1.358802	103.854696	Chinese Restaurant
1	Bishan	Central Region	1.355512	103.856863	SBS Transit: Bus 22	1.355269	103.857354	Bus Line
2	Bishan	Central Region	1.355512	103.856863	Tai Hwan Park	1.355454	103.859718	Park
3	Bishan	Central Region	1.355512	103.856863	Bus Stop 54609 (Opp Townsville Pri Sch)	1.359441	103.855259	Bus Stop
4	Bukit Merah	Central Region	1.292155	103.831806	Peperoni Pizzeria	1.292753	103.831402	Pizza Place

Methodology: Haversine Distance Calculation

- Using the Haversine formula, distance between two geographical coordinates on earth could be found.
- A custom defined function is written to loop through all the listings in the Airbnb dataset and find the distance between the coordinates of the Neighbourhood and the coordinate of the Airbnb listing.
- Result is stored in a new dataframe.
- Information on Haversine formula can be found in https://en.wikipedia.org/wiki/Haversine_formula

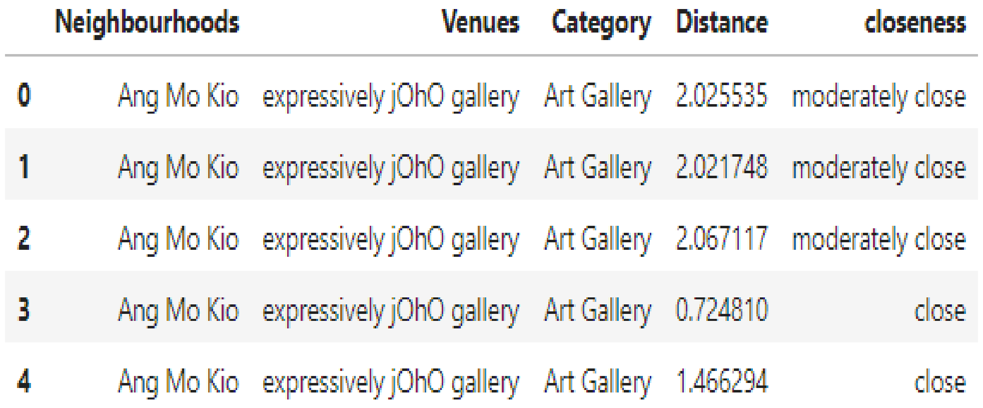


```
new_df.head()
```

	Neighbourhoods	Venues	Category	Distance
0	Ang Mo Kio	expressively jOhO gallery	Art Gallery	2.025535
1	Ang Mo Kio	expressively jOhO gallery	Art Gallery	2.021748
2	Ang Mo Kio	expressively jOhO gallery	Art Gallery	2.067117
3	Ang Mo Kio	expressively jOhO gallery	Art Gallery	0.724810
4	Ang Mo Kio	expressively jOhO gallery	Art Gallery	1.466294

Categorising Distance into Bins

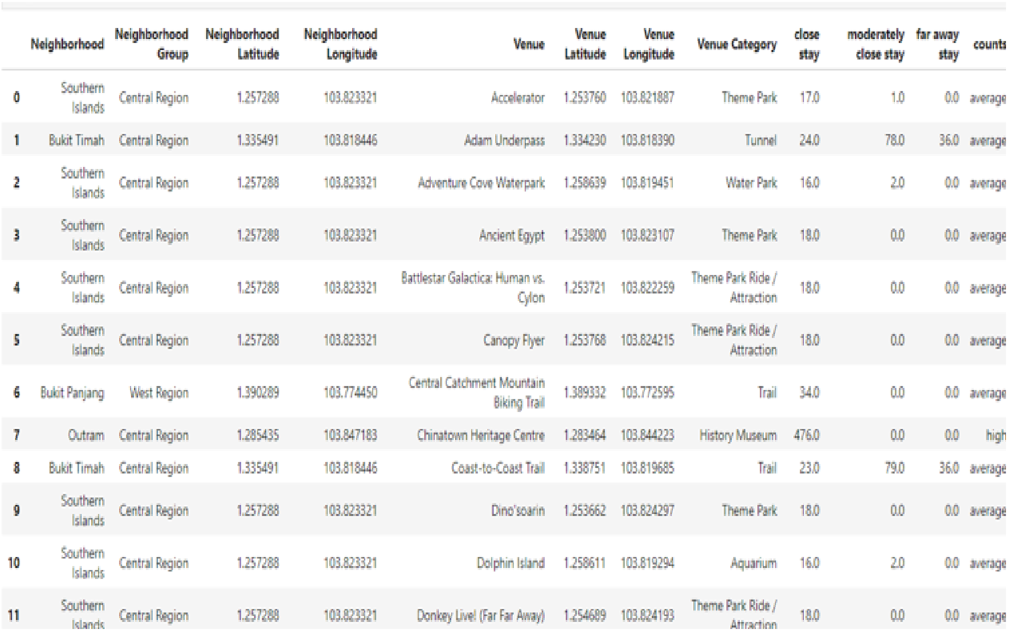
- Using Cut method of Pandas Dataframes, the distance information is categorised into 3 bins.
- The 3 bins are close, moderately close and far defined using the bin edges.
- Distance in the range of 0 – 2 km is close, between 2 and 4 is moderately close and between 4 and 7 km is far.



	Neighbourhoods	Venues	Category	Distance	closeness
0	Ang Mo Kio	expressively jOhO gallery	Art Gallery	2.025535	moderately close
1	Ang Mo Kio	expressively jOhO gallery	Art Gallery	2.021748	moderately close
2	Ang Mo Kio	expressively jOhO gallery	Art Gallery	2.067117	moderately close
3	Ang Mo Kio	expressively jOhO gallery	Art Gallery	0.724810	close
4	Ang Mo Kio	expressively jOhO gallery	Art Gallery	1.466294	close

Airbnb listings count and binning

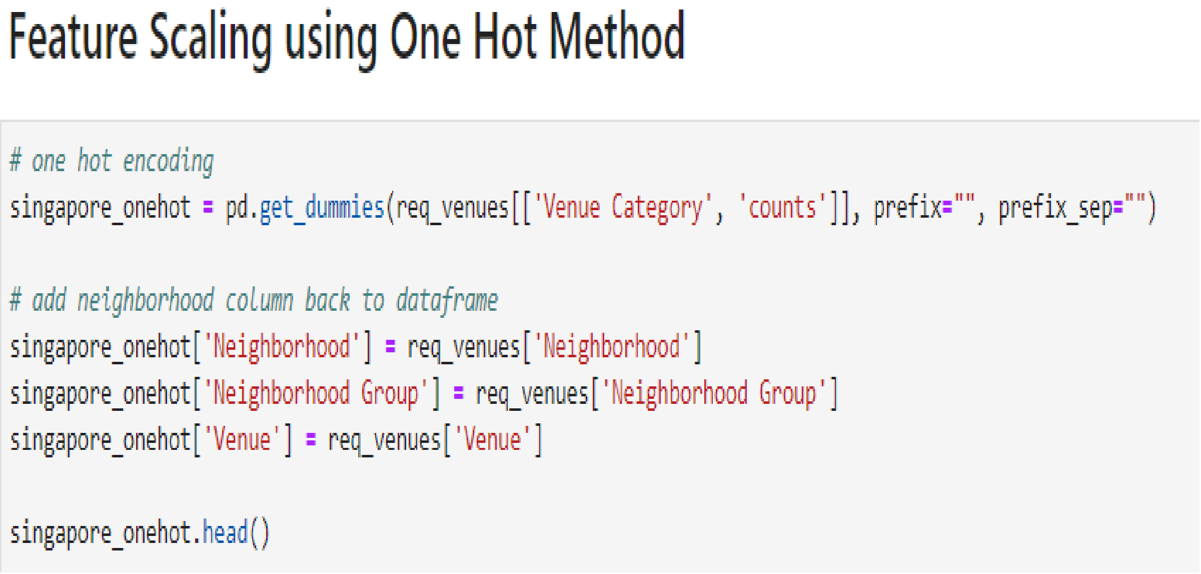
- Counts of Airbnb listings in each distance category is found and added to the dataframe in similarly named columns.
- Similar binning method is used to categorise the no of listings in “close stay column” as “low”, “average” or “high”
- “High” means high no of listings within 2km from venue and similar interpretation for “average” and “low” categories.



	Neighborhood	Neighborhood Group	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	close stay	moderately close stay	far away stay	counts
0	Southern Islands	Central Region	1.257288	103.823321	Accelerator	1.253760	103.821887	Theme Park	17.0	1.0	0.0	average
1	Bukit Timah	Central Region	1.335491	103.818446	Adam Underpass	1.334230	103.818390	Tunnel	24.0	78.0	36.0	average
2	Southern Islands	Central Region	1.257288	103.823321	Adventure Cove Waterpark	1.258639	103.819451	Water Park	16.0	2.0	0.0	average
3	Southern Islands	Central Region	1.257288	103.823321	Ancient Egypt	1.253800	103.823107	Theme Park	18.0	0.0	0.0	average
4	Southern Islands	Central Region	1.257288	103.823321	Battlestar Galactica: Human vs. Cylon	1.253721	103.822259	Theme Park Ride / Attraction	18.0	0.0	0.0	average
5	Southern Islands	Central Region	1.257288	103.823321	Canopy Flyer	1.253768	103.824215	Theme Park Ride / Attraction	18.0	0.0	0.0	average
6	Bukit Panjang	West Region	1.390289	103.774450	Central Catchment Mountain Biking Trail	1.389332	103.772595	Trail	34.0	0.0	0.0	average
7	Outram	Central Region	1.285435	103.847183	Chinatown Heritage Centre	1.283464	103.844223	History Museum	476.0	0.0	0.0	high
8	Bukit Timah	Central Region	1.335491	103.818446	Coast-to-Coast Trail	1.338751	103.819685	Trail	23.0	79.0	36.0	average
9	Southern Islands	Central Region	1.257288	103.823321	Dino'soarin	1.253662	103.824297	Theme Park	18.0	0.0	0.0	average
10	Southern Islands	Central Region	1.257288	103.823321	Dolphin Island	1.258611	103.819294	Aquarium	16.0	2.0	0.0	average
11	Southern Islands	Central Region	1.257288	103.823321	Donkey Live! (Far Far Away)	1.254689	103.824193	Theme Park Ride / Attraction	18.0	0.0	0.0	average

Feature Scaling: One Hot Encoding

- One Hot encoding is implemented on Venues categories and the counts column from previous screenshot to convert categorical variables in 0's and 1's.

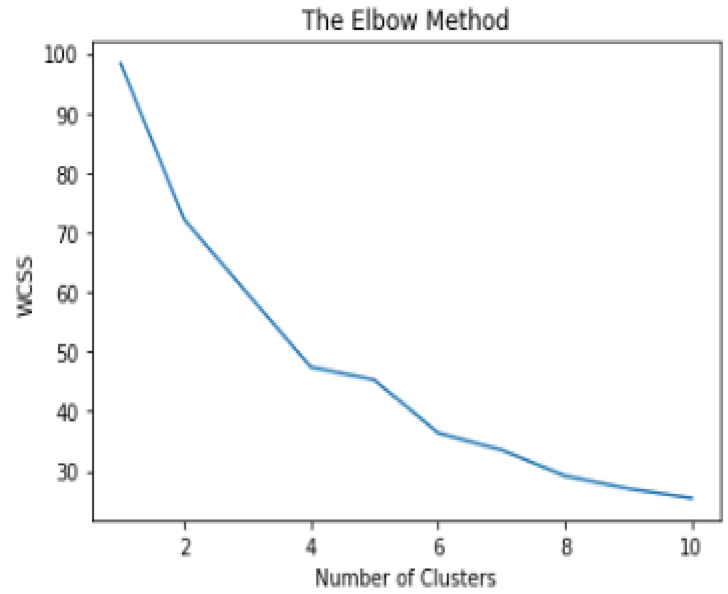


Feature Scaling using One Hot Method

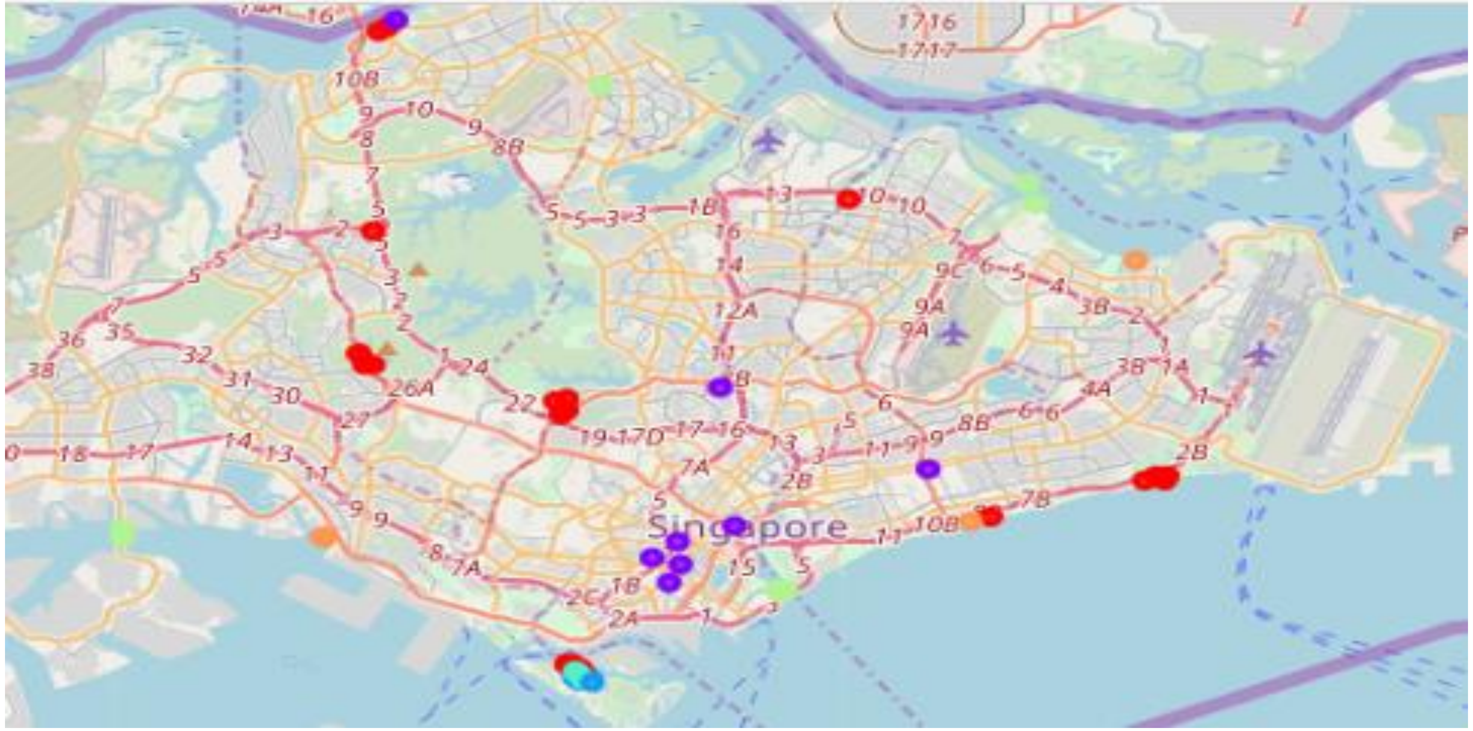
```
# one hot encoding
singapore_onehot = pd.get_dummies(req_venues[['Venue Category', 'counts']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
singapore_onehot['Neighborhood'] = req_venues['Neighborhood']
singapore_onehot['Neighborhood Group'] = req_venues['Neighborhood Group']
singapore_onehot['Venue'] = req_venues['Venue']

singapore_onehot.head()
```

K-Means Clustering: Find Best K

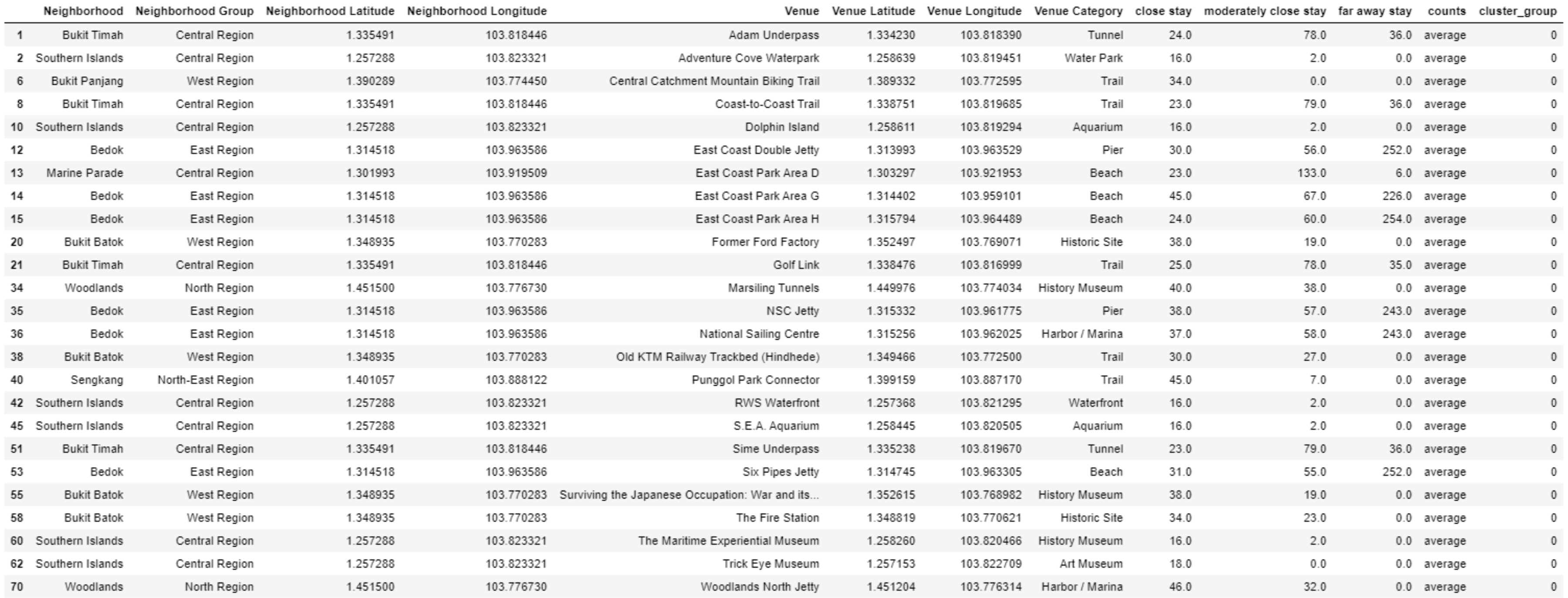
- Elbow Method is used to find the best value of K to run the K-means algorithm.
- The best value of K is determined to be 6.



Results: Folium Map Visualisation

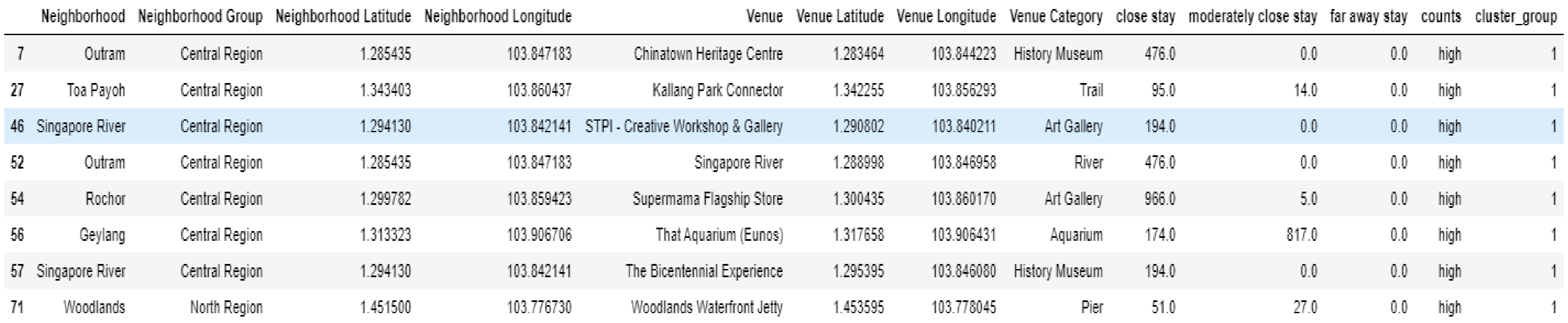


Results: Cluster 0



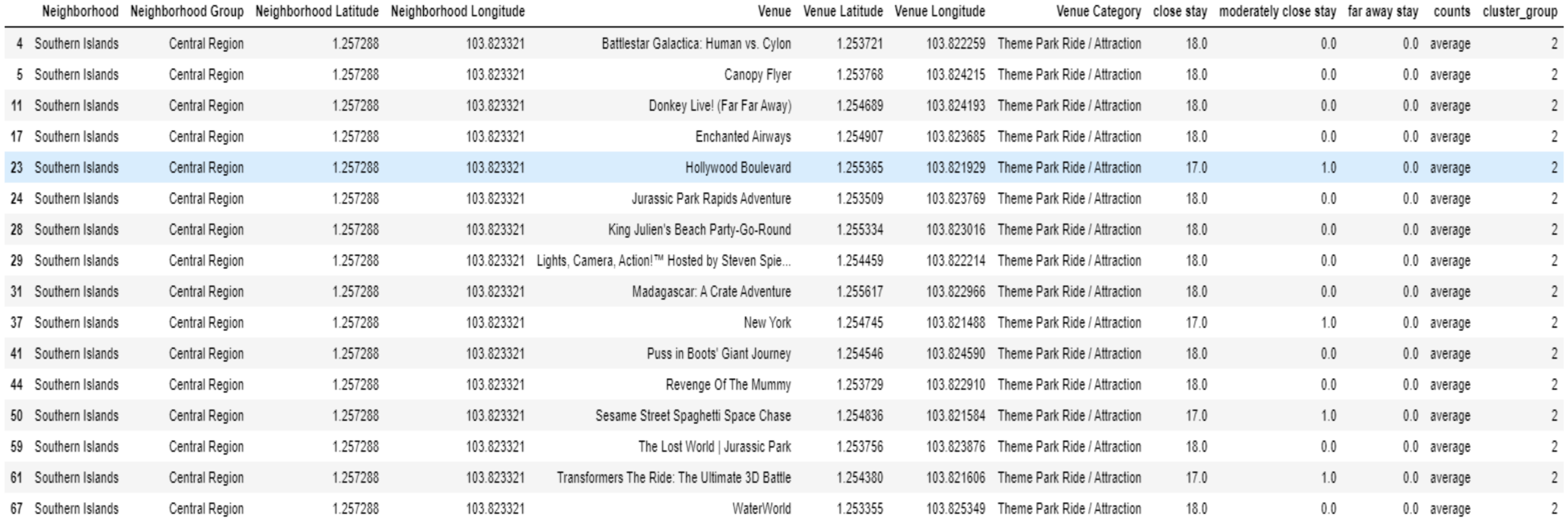
	Neighborhood	Neighborhood Group	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	close stay	moderately close stay	far away stay	counts	cluster_group
1	Bukit Timah	Central Region	1.335491	103.818446	Adam Underpass	1.334230	103.818390	Tunnel	24.0	78.0	36.0	average	0
2	Southern Islands	Central Region	1.257288	103.823321	Adventure Cove Waterpark	1.258639	103.819451	Water Park	16.0	2.0	0.0	average	0
6	Bukit Panjang	West Region	1.390289	103.774450	Central Catchment Mountain Biking Trail	1.389332	103.772595	Trail	34.0	0.0	0.0	average	0
8	Bukit Timah	Central Region	1.335491	103.818446	Coast-to-Coast Trail	1.338751	103.819685	Trail	23.0	79.0	36.0	average	0
10	Southern Islands	Central Region	1.257288	103.823321	Dolphin Island	1.258611	103.819294	Aquarium	16.0	2.0	0.0	average	0
12	Bedok	East Region	1.314518	103.963586	East Coast Double Jetty	1.313993	103.963529	Pier	30.0	56.0	252.0	average	0
13	Marine Parade	Central Region	1.301993	103.919509	East Coast Park Area D	1.303297	103.921953	Beach	23.0	133.0	6.0	average	0
14	Bedok	East Region	1.314518	103.963586	East Coast Park Area G	1.314402	103.959101	Beach	45.0	67.0	226.0	average	0
15	Bedok	East Region	1.314518	103.963586	East Coast Park Area H	1.315794	103.964489	Beach	24.0	60.0	254.0	average	0
20	Bukit Batok	West Region	1.348935	103.770283	Former Ford Factory	1.352497	103.769071	Historic Site	38.0	19.0	0.0	average	0
21	Bukit Timah	Central Region	1.335491	103.818446	Golf Link	1.338476	103.816999	Trail	25.0	78.0	35.0	average	0
34	Woodlands	North Region	1.451500	103.776730	Marsiling Tunnels	1.449976	103.774034	History Museum	40.0	38.0	0.0	average	0
35	Bedok	East Region	1.314518	103.963586	NSC Jetty	1.315332	103.961775	Pier	38.0	57.0	243.0	average	0
36	Bedok	East Region	1.314518	103.963586	National Sailing Centre	1.315256	103.962025	Harbor / Marina	37.0	58.0	243.0	average	0
38	Bukit Batok	West Region	1.348935	103.770283	Old KTM Railway Trackbed (Hindhede)	1.349466	103.772500	Trail	30.0	27.0	0.0	average	0
40	Sengkang	North-East Region	1.401057	103.888122	Punggol Park Connector	1.399159	103.887170	Trail	45.0	7.0	0.0	average	0
42	Southern Islands	Central Region	1.257288	103.823321	RWS Waterfront	1.257368	103.821295	Waterfront	16.0	2.0	0.0	average	0
45	Southern Islands	Central Region	1.257288	103.823321	S.E.A. Aquarium	1.258445	103.820505	Aquarium	16.0	2.0	0.0	average	0
51	Bukit Timah	Central Region	1.335491	103.818446	Sime Underpass	1.335238	103.819670	Tunnel	23.0	79.0	36.0	average	0
53	Bedok	East Region	1.314518	103.963586	Six Pipes Jetty	1.314745	103.963305	Beach	31.0	55.0	252.0	average	0
55	Bukit Batok	West Region	1.348935	103.770283	Surviving the Japanese Occupation: War and its...	1.352615	103.768982	History Museum	38.0	19.0	0.0	average	0
58	Bukit Batok	West Region	1.348935	103.770283	The Fire Station	1.348819	103.770621	Historic Site	34.0	23.0	0.0	average	0
60	Southern Islands	Central Region	1.257288	103.823321	The Maritime Experiential Museum	1.258260	103.820466	History Museum	16.0	2.0	0.0	average	0
62	Southern Islands	Central Region	1.257288	103.823321	Trick Eye Museum	1.257153	103.822709	Art Museum	18.0	0.0	0.0	average	0
70	Woodlands	North Region	1.451500	103.776730	Woodlands North Jetty	1.451204	103.776314	Harbor / Marina	46.0	32.0	0.0	average	0

Results: Cluster 1



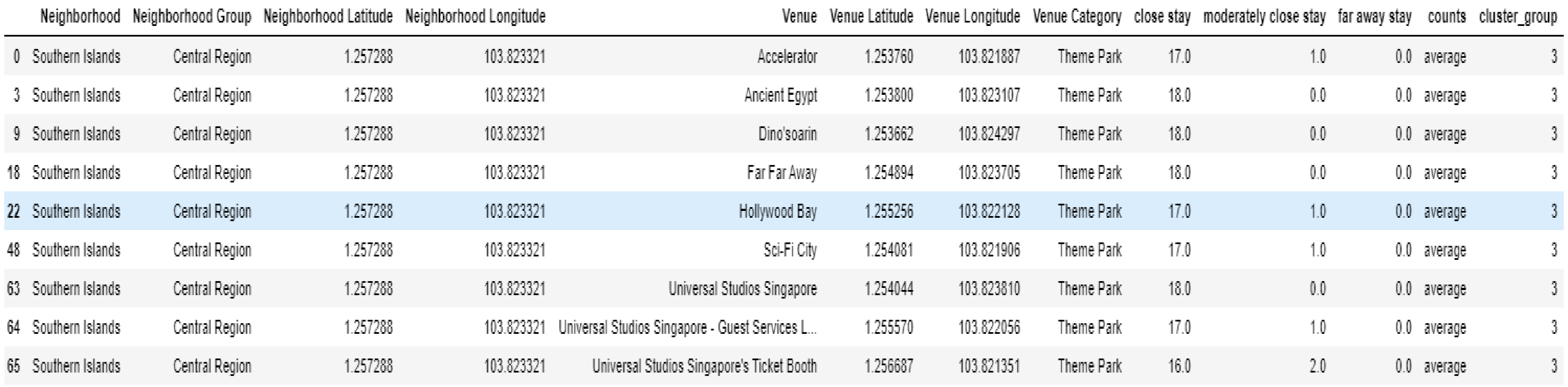
	Neighborhood	Neighborhood Group	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	close stay	moderately close stay	far away stay	counts	cluster_group
7	Outram	Central Region	1.285435	103.847183	Chinatown Heritage Centre	1.283464	103.844223	History Museum	476.0	0.0	0.0	high	1
27	Toa Payoh	Central Region	1.343403	103.860437	Kallang Park Connector	1.342255	103.856293	Trail	95.0	14.0	0.0	high	1
46	Singapore River	Central Region	1.294130	103.842141	STPI - Creative Workshop & Gallery	1.290802	103.840211	Art Gallery	194.0	0.0	0.0	high	1
52	Outram	Central Region	1.285435	103.847183	Singapore River	1.288998	103.846958	River	476.0	0.0	0.0	high	1
54	Rochor	Central Region	1.299782	103.859423	Supermama Flagship Store	1.300435	103.860170	Art Gallery	966.0	5.0	0.0	high	1
56	Geylang	Central Region	1.313323	103.906706	That Aquarium (Eunos)	1.317658	103.906431	Aquarium	174.0	817.0	0.0	high	1
57	Singapore River	Central Region	1.294130	103.842141	The Bicentennial Experience	1.295395	103.846080	History Museum	194.0	0.0	0.0	high	1
71	Woodlands	North Region	1.451500	103.776730	Woodlands Waterfront Jetty	1.453595	103.778045	Pier	51.0	27.0	0.0	high	1

Results: Cluster 2



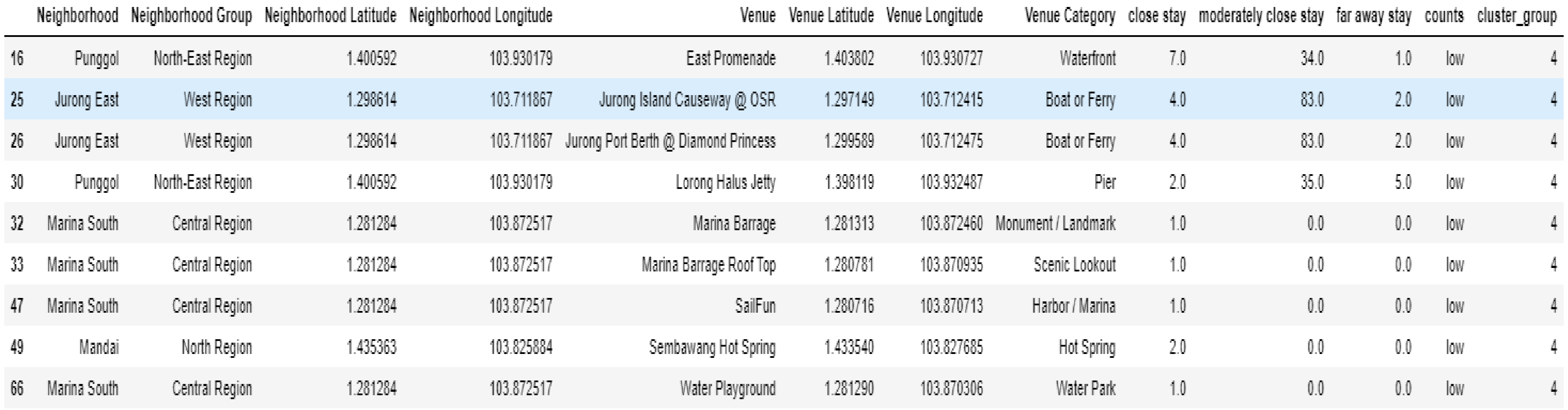
	Neighborhood	Neighborhood Group	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	close stay	moderately close stay	far away stay	counts	cluster_group
4	Southern Islands	Central Region	1.257288	103.823321	Battlestar Galactica: Human vs. Cylon	1.253721	103.822259	Theme Park Ride / Attraction	18.0	0.0	0.0	average	2
5	Southern Islands	Central Region	1.257288	103.823321	Canopy Flyer	1.253768	103.824215	Theme Park Ride / Attraction	18.0	0.0	0.0	average	2
11	Southern Islands	Central Region	1.257288	103.823321	Donkey Live! (Far Far Away)	1.254689	103.824193	Theme Park Ride / Attraction	18.0	0.0	0.0	average	2
17	Southern Islands	Central Region	1.257288	103.823321	Enchanted Airways	1.254907	103.823685	Theme Park Ride / Attraction	18.0	0.0	0.0	average	2
23	Southern Islands	Central Region	1.257288	103.823321	Hollywood Boulevard	1.255365	103.821929	Theme Park Ride / Attraction	17.0	1.0	0.0	average	2
24	Southern Islands	Central Region	1.257288	103.823321	Jurassic Park Rapids Adventure	1.253509	103.823769	Theme Park Ride / Attraction	18.0	0.0	0.0	average	2
28	Southern Islands	Central Region	1.257288	103.823321	King Julien's Beach Party-Go-Round	1.255334	103.823016	Theme Park Ride / Attraction	18.0	0.0	0.0	average	2
29	Southern Islands	Central Region	1.257288	103.823321	Lights, Camera, Action!™ Hosted by Steven Spie...	1.254459	103.822214	Theme Park Ride / Attraction	18.0	0.0	0.0	average	2
31	Southern Islands	Central Region	1.257288	103.823321	Madagascar: A Crate Adventure	1.255617	103.822966	Theme Park Ride / Attraction	18.0	0.0	0.0	average	2
37	Southern Islands	Central Region	1.257288	103.823321	New York	1.254745	103.821488	Theme Park Ride / Attraction	17.0	1.0	0.0	average	2
41	Southern Islands	Central Region	1.257288	103.823321	Puss in Boots' Giant Journey	1.254546	103.824590	Theme Park Ride / Attraction	18.0	0.0	0.0	average	2
44	Southern Islands	Central Region	1.257288	103.823321	Revenge Of The Mummy	1.253729	103.822910	Theme Park Ride / Attraction	18.0	0.0	0.0	average	2
50	Southern Islands	Central Region	1.257288	103.823321	Sesame Street Spaghetti Space Chase	1.254836	103.821584	Theme Park Ride / Attraction	17.0	1.0	0.0	average	2
59	Southern Islands	Central Region	1.257288	103.823321	The Lost World | Jurassic Park	1.253756	103.823876	Theme Park Ride / Attraction	18.0	0.0	0.0	average	2
61	Southern Islands	Central Region	1.257288	103.823321	Transformers The Ride: The Ultimate 3D Battle	1.254380	103.821606	Theme Park Ride / Attraction	17.0	1.0	0.0	average	2
67	Southern Islands	Central Region	1.257288	103.823321	WaterWorld	1.253355	103.825349	Theme Park Ride / Attraction	18.0	0.0	0.0	average	2

Results: Cluster 3



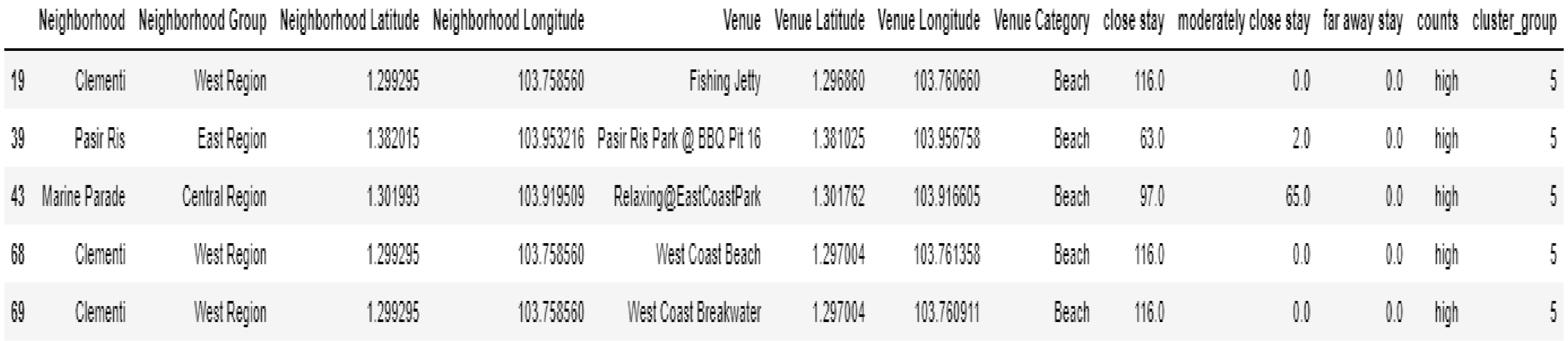
	Neighborhood	Neighborhood Group	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	close stay	moderately close stay	far away stay	counts	cluster_group
0	Southern Islands	Central Region	1.257288	103.823321	Accelerator	1.253760	103.821887	Theme Park	17.0	1.0	0.0	average	3
3	Southern Islands	Central Region	1.257288	103.823321	Ancient Egypt	1.253800	103.823107	Theme Park	18.0	0.0	0.0	average	3
9	Southern Islands	Central Region	1.257288	103.823321	Dino'soarin	1.253662	103.824297	Theme Park	18.0	0.0	0.0	average	3
18	Southern Islands	Central Region	1.257288	103.823321	Far Far Away	1.254894	103.823705	Theme Park	18.0	0.0	0.0	average	3
22	Southern Islands	Central Region	1.257288	103.823321	Hollywood Bay	1.255256	103.822128	Theme Park	17.0	1.0	0.0	average	3
48	Southern Islands	Central Region	1.257288	103.823321	Sci-Fi City	1.254081	103.821906	Theme Park	17.0	1.0	0.0	average	3
63	Southern Islands	Central Region	1.257288	103.823321	Universal Studios Singapore	1.254044	103.823810	Theme Park	18.0	0.0	0.0	average	3
64	Southern Islands	Central Region	1.257288	103.823321	Universal Studios Singapore - Guest Services L...	1.255570	103.822056	Theme Park	17.0	1.0	0.0	average	3
65	Southern Islands	Central Region	1.257288	103.823321	Universal Studios Singapore's Ticket Booth	1.256687	103.821351	Theme Park	16.0	2.0	0.0	average	3

Results: Cluster 4



	Neighborhood	Neighborhood Group	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	close stay	moderately close stay	far away stay	counts	cluster_group
16	Punggol	North-East Region	1.400592	103.930179	East Promenade	1.403802	103.930727	Waterfront	7.0	34.0	1.0	low	4
25	Jurong East	West Region	1.298614	103.711867	Jurong Island Causeway @ OSR	1.297149	103.712415	Boat or Ferry	4.0	83.0	2.0	low	4
26	Jurong East	West Region	1.298614	103.711867	Jurong Port Berth @ Diamond Princess	1.299589	103.712475	Boat or Ferry	4.0	83.0	2.0	low	4
30	Punggol	North-East Region	1.400592	103.930179	Lorong Halus Jetty	1.398119	103.932487	Pier	2.0	35.0	5.0	low	4
32	Marina South	Central Region	1.281284	103.872517	Marina Barrage	1.281313	103.872460	Monument / Landmark	1.0	0.0	0.0	low	4
33	Marina South	Central Region	1.281284	103.872517	Marina Barrage Roof Top	1.280781	103.870935	Scenic Lookout	1.0	0.0	0.0	low	4
47	Marina South	Central Region	1.281284	103.872517	SailFun	1.280716	103.870713	Harbor / Marina	1.0	0.0	0.0	low	4
49	Mandai	North Region	1.435363	103.825884	Sembawang Hot Spring	1.433540	103.827685	Hot Spring	2.0	0.0	0.0	low	4
66	Marina South	Central Region	1.281284	103.872517	Water Playground	1.281290	103.870306	Water Park	1.0	0.0	0.0	low	4

Results: Cluster 5



	Neighborhood	Neighborhood Group	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	close stay	moderately close stay	far away stay	counts	cluster_group
19	Clementi	West Region	1.299295	103.758560	Fishing Jetty	1.296860	103.760660	Beach	116.0	0.0	0.0	high	5
39	Pasir Ris	East Region	1.382015	103.953216	Pasir Ris Park @ BBQ Pit 16	1.381025	103.956758	Beach	63.0	2.0	0.0	high	5
43	Marine Parade	Central Region	1.301993	103.919509	Relaxing@EastCoastPark	1.301762	103.916605	Beach	97.0	65.0	0.0	high	5
68	Clementi	West Region	1.299295	103.758560	West Coast Beach	1.297004	103.761358	Beach	116.0	0.0	0.0	high	5
69	Clementi	West Region	1.299295	103.758560	West Coast Breakwater	1.297004	103.760911	Beach	116.0	0.0	0.0	high	5

Conclusion:

- From the results, it can be seen that each cluster has some unique property such as some might be best suitable for a particular type of Venue category.
- Some others might be better just because it has a high number of Airbnb rental listings.
- As an example, if a traveller is looking for places that has the best trail, it can be suggested to go for the Neighbourhoods in Cluster 0 which has relatively more Trail venues and an average number of Airbnb listings.
- For travellers, who prefer beach venues, Cluster 5 can be suggested as it has an high no of Airbnb listings close to those venues.
- Hence Travel agents can create specific itineraries which might suggest the best places to visit for travellers based on their needs and also based on the number of Airbnb listings.