Best cities to set-up business in South East Asia

A. Introduction

A.1. Description & Discussion of the Background

South East Asia is on its way becoming a commercial hub in Asia. Over recent years, the region's economy with the population of more than 600 million has thrived to be the world's fastest growing markets with the total combined GDP of $2.4 trillion, according to McKinsey. Every year, many international corporates are setting up their footprint in this region. However, locating the right location to setup an business (or just an office) is not an easy task for the outsider. There's too many factors that needed to be consider before deciding which cities in South East Asia that fits the most. That includes (but not limited to): ease of doing business, cost of living, Rent cost, nearby venue structure, etc...

After all, understanding how cities different to another will give the business owner valuable information to decide whether a location is best fit for business. It's also good to add other developed cities in Asia into dataset to see how similarity between SEA countries and the rest. Once the data is clustered, we expect to see clearer picture of the similarity or dissimilarity between cities.

A.2. Data Description

To consider a location is best fit for an office setup, several datasets from different sources below are used that I've found it useful for this project:

- **Cost of Living index [3]:** The site summarize all of the Cost of Living Index, Rent Index, Cost of Living Plus Rent Index and other Index that will be used in the calculation.
- **Four square API [1]:** was used to explore the nearby venues around the given location. With the free account limitation, it's max 100 venues per API call. The data will then need to be filtered out based on venue categories: restaurant, museum, etc…
- **Simplemaps**[5]: I've also used this dataset from Simplemaps to get the latitude and longitude of each cities.
- **Ease of Doing Business** [4]: how easy to setup or start a business, rank by the World Bank. I've found this very relevant data to our project. The data also includes the ranking of more categories: how easy to deal with construction permits, how easy to get Electricity or Registering property… In this project scope, I only use the overall ranking "ease of Doing Business" into our calculation. Unfortunately, the data doesn't cover all the cities in South East Asia so I've decided to drop out all of the cities that doesn't have this data. Obviously, the dropped cities has relatively small population instead.

B. Methodology

As a database, I used GitHub repository in my study. My master data which has the main components *Cities basic information database, Cost of Living Index, Ease of Doing Business ranking, Latitude* and *Longitude* information of the city.

The below merged dataframe consists of the both country latitude and longitude together with population data.

	Ranking	Name	Country	Population	city_ascii	lat	lng	country	iso2	iso3	admin_name	capital	population	id
0	1	Jakarta	Indonesia	10135030.0	Jakarta	-6.1744	106.8294	Indonesia	ID	IDN	Jakarta	primary	9125000.0	1.360771e+09
1	2	Bangkok	Thailand	8750600.0	Bangkok	13.7500	100.5166	Thailand	TH	THA	Krung Thep Maha Nakhon	primary	6704000.0	1.764069e+09
2	3	Ho Chi Minh City	Vietnam	7981411.0	Ho Chi Minh City	10.7800	106.6950	Vietnam	VN	VNM	Hồ Chí Minh	admin	5314000.0	1.704774e+09
3	4	Hanoi	Vietnam	7067000.0	Hanoi	21.0333	105.8500	Vietnam	VN	VNM	Hà Nội	primary	4378000.0	1.704414e+09
4	5	Singapore	Singapore	5399000.0	Singapore	1.2930	103.8558	Singapore	SG	SGP	NaN	primary	5183700.0	1.702341e+09
5	6	Yangon	Myanmar	5451439.0	Yangon	16.8409	96.1735	Myanmar	NaN	NaN	NaN	NaN	NaN	NaN
6	7	Surabaya	Indonesia	3457409.0	Surabaya	-7.2492	112.7508	Indonesia	ID	IDN	Jawa Timur	admin	2845000.0	1.360485e+09
7	8	Quezon City	Philippines	2936116.0	Quezon City	14.6504	121.0300	Philippines	PH	PHL	Quezon	admin	2761720.0	1.608974e+09
8	9	Nakhon Ratchasima	Thailand	2646401.0	Nakhon Ratchasima	15.0000	102.1000	Thailand	TH	THA	Nakhon Ratchasima	admin	334984.0	1.764752e+09
9	10	Bandung	Indonesia	2575478.0	Bandung	-6.9500	107.5700	Indonesia	ID	IDN	Jawa Barat	admin	2394000.0	1.360313e+09

The Cost of Living Index is also imported as below:

	Rank	City	Cost of Living Index	Rent Index	Cost of Living Plus Rent Index	Groceries Index	Restaurant Price Index	Local Purchasing Power Index
0	1	Zurich, Switzerland	128.29	61.66	96.42	127.96	124.73	126.90
1	2	Basel, Switzerland	125.54	45.76	87.38	124.99	123.11	121.47
2	3	Lausanne, Switzerland	124.02	50.64	88.92	127.26	123.61	110.52
3	4	Geneva, Switzerland	118.98	68.47	94.82	112.88	119.58	111.16
4	5	Bern, Switzerland	116.03	40.52	79.91	107.58	115.56	131.89

and merged data is shown below:

	index	Name	Country	Population	lat	lng	iso2	iso3	capital	id	Cost of Living Index	Rent Index	Groceries Index	Restaurant Price Index	Local Purchasing Power Index
0	0	Jakarta	Indonesia	10135030.0	-6.1744	106.8294	ID	IDN	primary	1.360771e+09	44.48	19.42	45.41	25.06	27.34
1	1	Bangkok	Thailand	8750600.0	13.7500	100.5166	TH	THA	primary	1.764069e+09	57.15	26.19	56.88	28.14	34.04
2	2	Ho Chi Minh City	Vietnam	7981411.0	10.7800	106.6950	VN	VNM	admin	1.704774e+09	38.20	16.74	36.30	20.71	26.64
3	3	Hanoi	Vietnam	7067000.0	21.0333	105.8500	VN	VNM	primary	1.704414e+09	39.88	11.45	39.32	19.69	28.99
4	4	Singapore	Singapore	5399000.0	1.2930	103.8558	SG	SGP	primary	1.702341e+09	81.10	63.27	66.75	58.99	88.96
9	9	Bandung	Indonesia	2575478.0	-6.9500	107.5700	ID	IDN	admin	1.360313e+09	35.71	9.69	36.86	16.95	25.78
13	16	Kuala Lumpur	Malaysia	1808922.0	3.1667	101.7000	MY	MYS	primary	1.458989e+09	42.47	15.18	40.77	27.32	67.27
15	18	Manila	Philippines	1780148.0	14.6042	120.9822	PH	PHL	primary	1.608618e+09	38.99	21.33	33.32	25.61	25.14
16	19	Chiang Mai	Thailand	1763742.0	18.8000	98.9800	TH	THA	admin	1.764663e+09	46.64	13.13	49.81	22.52	33.90
28	37	Phnom Penh	Cambodia	1242992.0	11.5500	104.9166	KH	KHM	primary	1.116261e+09	46.92	18.23	43.29	26.45	14.56
37	0	Delhi	India	46960000.0	28.6700	77.2300	IN	IND	admin	1.356873e+09	28.18	8.18	26.15	24.76	54.69
39	1	Tokyo	Japan	39800000.0	35.6850	139.7514	JP	JPN	primary	1.392686e+09	86.87	38.00	83.42	56.70	89.70
40	2	Mumbai	India	25000000.0	19.0170	72.8570	IN	IND	admin	1.356227e+09	28.84	21.02	26.86	26.61	41.68
41	3	Seoul	South Korea	24800000.0	37.5663	126.9997	KR	KOR	primary	1.410836e+09	85.50	30.45	103.69	45.88	76.81
42	4	Shanghai	China	31100000.0	31.2165	121.4365	CN	CHN	admin	1.156074e+09	50.07	35.67	52.50	36.48	54.40
43	5	Beijing	China	20700000.0	39.9289	116.3883	CN	CHN	primary	1.156229e+09	44.89	34.97	43.08	35.85	60.72
44	6	Guangzhou	China	48600000.0	23.1450	113.3250	CN	CHN	admin	1.156237e+09	41.26	17.28	42.80	27.95	63.76

I’ve also imported the Ease of Doing Business Index (last column) into the main dataframe

	index_x	Name	Country	Population	lat	lng	iso2	iso3	capital	id	Cost of Living Index	Rent Index	Groceries Index	Restaurant Price Index	Local Purchasing Power Index	index_y	Ease of Doing Business
0	0	Jakarta	Indonesia	10135030.0	-6.1744	106.8294	ID	IDN	primary	1.360771e+09	44.48	19.42	45.41	25.06	27.34	118	73.0
1	1	Bangkok	Thailand	8750600.0	13.7500	100.5166	TH	THA	primary	1.764069e+09	57.15	26.19	56.88	28.14	34.04	170	21.0
2	2	Ho Chi Minh City	Vietnam	7981411.0	10.7800	106.6950	VN	VNM	admin	1.704774e+09	38.20	16.74	36.30	20.71	26.64	121	70.0
3	3	Hanoi	Vietnam	7067000.0	21.0333	105.8500	VN	VNM	primary	1.704414e+09	39.88	11.45	39.32	19.69	28.99	121	70.0
4	4	Singapore	Singapore	5399000.0	1.2930	103.8558	SG	SGP	primary	1.702341e+09	81.10	63.27	66.75	58.99	88.96	189	2.0
9	9	Bandung	Indonesia	2575478.0	-6.9500	107.5700	ID	IDN	admin	1.360313e+09	35.71	9.69	36.86	16.95	25.78	118	73.0
13	16	Kuala Lumpur	Malaysia	1808922.0	3.1667	101.7000	MY	MYS	primary	1.458989e+09	42.47	15.18	40.77	27.32	67.27	179	12.0
15	18	Manila	Philippines	1780148.0	14.6042	120.9822	PH	PHL	primary	1.608618e+09	38.99	21.33	33.32	25.61	25.14	96	95.0
16	19	Chiang Mai	Thailand	1763742.0	18.8000	98.9800	TH	THA	admin	1.764663e+09	46.64	13.13	49.81	22.52	33.90	170	21.0
28	37	Phnom Penh	Cambodia	1242992.0	11.5500	104.9166	KH	KHM	primary	1.116261e+09	46.92	18.23	43.29	26.45	14.56	47	144.0
37	0	Delhi	India	46960000.0	28.6700	77.2300	IN	IND	admin	1.356873e+09	28.18	8.18	26.15	24.76	54.69	128	63.0
38	0	Delhi	India	46960000.0	37.4306	-120.7759	US	USA	NaN	1.840019e+09	28.18	8.18	26.15	24.76	54.69	128	63.0
39	1	Tokyo	Japan	39800000.0	35.6850	139.7514	JP	JPN	primary	1.392686e+09	86.87	38.00	83.42	56.70	89.70	162	29.0
40	2	Mumbai	India	25000000.0	19.0170	72.8570	IN	IND	admin	1.356227e+09	28.84	21.02	26.86	26.61	41.68	128	63.0

Python folium library was used to visualize geographic details of each cities. The circle radius represent the population of each cities.

Next, I utilized the Foursquare API to explore the cities and segment them. I designed the limit as 100 venue and the radius 2000 meter for each borough from their given latitude and longitude information. Below is a head of the list Venues name, category, latitude and longitude informations from Foursquare API.

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Bandung	100	100	100	100	100	100
Bangkok	100	100	100	100	100	100
Beijing	100	100	100	100	100	100
Chiang Mai	100	100	100	100	100	100
Delhi	171	171	171	171	171	171
Guangzhou	100	100	100	100	100	100
Hanoi	100	100	100	100	100	100
Ho Chi Minh City	100	100	100	100	100	100
Hong Kong	100	100	100	100	100	100
Jakarta	100	100	100	100	100	100
Kuala Lumpur	100	100	100	100	100	100
Manila	100	100	100	100	100	100
Mumbai	100	100	100	100	100	100

Note that with Foursquare API free account, the maximum resulted venue count per call is at 100. While coverage radius of 2000m which is not too coarse, the maximum venues called is already maxed out at 100. This can be solved be narrowing down the radius while having more calls per each metropolitan areas in each cities. To do this, we shall need to do another study on each cities and re-select the areas that best fit our requirement. Then as mentioned, run Foursquare API call multiple times and merge data later on. Eventually, we would be able to acquire more data that can help different between cities. In above snapshot, all of the cities have the same venue counts.

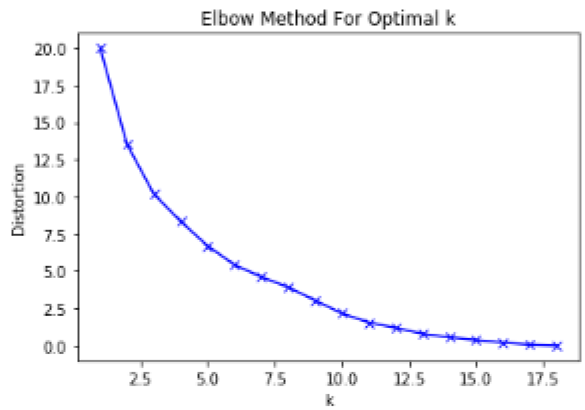
In summary there are 268 unique categories resulted from Foursquare. However, many of the categorie s must be dropped down as it's not related to our project. Only venue with categories below are selecte d into the main dataset.

- "Airport", "Building", "Station", "Cafeteria", "Convention Center", "Hotel", "University"

Below is the result of the most common venue per cities.

	Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Bandung	Hotel	Convention Center	Cafeteria	Building	Airport
1	Bangkok	Hotel	Convention Center	Cafeteria	Building	Airport
2	Beijing	Hotel	Convention Center	Cafeteria	Building	Airport
3	Chiang Mai	Hotel	Convention Center	Cafeteria	Building	Airport
4	Delhi	Hotel	Airport	Convention Center	Cafeteria	Building

Combining all of the factors, the data now is ready to be clustered using K-means algorithm. K-Means algorithm is one of the most common cluster method of unsupervised learning. Next, I will run K-means loops to see the optimum cluster number for the dataset.



As you can see on the above figure, the Elblow plot doesn't show the clear cut for the best K-means determination. This can be explained due to the insufficient dataset from the Foursquare API where most of the cities venue reach max to 100 and hence there's no clear difference between cities.

With the given insufficient dataset that couldn't get us the clear-cut optimum, however we will go with 8 cluster (distortion value around 5.0) as the optimum number.

Here is my merged table with cluster labels for each cities.

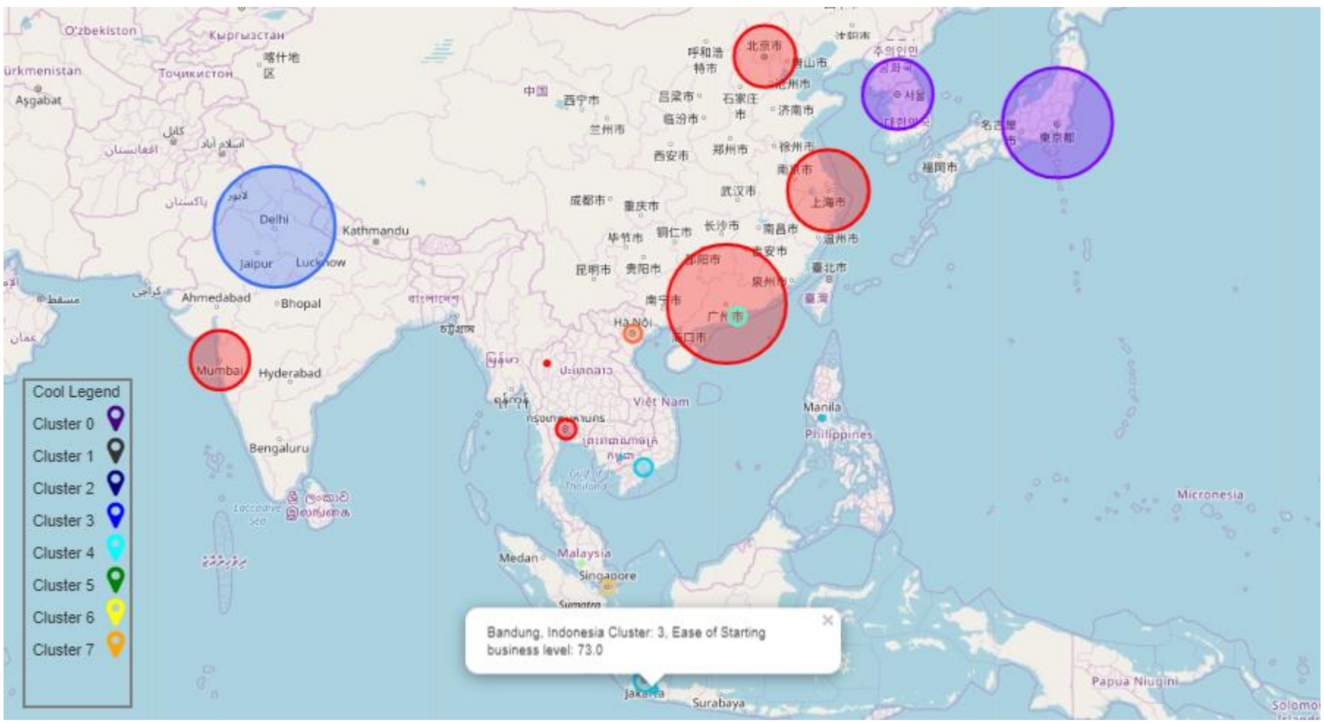
	Cluster Labels	index_x	Name	Country	Population	lat	lng	iso2	iso3	capital	id	Cost of Living Index	Rent Index	Groceries Index	Restaurant Price Index	Local Purchasing Power Index	index_y	Ease of Doing Business	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	0	1	Bangkok	Thailand	8750600.0	13.7500	100.5166	TH	THA	primary	1.764069e+09	57.15	26.19	56.88	28.14	34.04	170	21.0	Hotel	Convention Center	Cafeteria	Building	Airport
16	0	5	Beijing	China	20700000.0	39.9289	116.3883	CN	CHN	primary	1.156229e+09	44.89	34.97	43.08	35.85	60.72	160	31.0	Hotel	Convention Center	Cafeteria	Building	Airport
15	0	4	Shanghai	China	31100000.0	31.2165	121.4365	CN	CHN	admin	1.156074e+09	50.07	35.67	52.50	36.48	54.40	160	31.0	Hotel	Convention Center	Cafeteria	Building	Airport
13	0	2	Mumbai	India	25000000.0	19.0170	72.8570	IN	IND	admin	1.356227e+09	28.84	21.02	26.86	26.61	41.68	128	63.0	Hotel	Convention Center	Cafeteria	Building	Airport
8	0	19	Chiang Mai	Thailand	1763742.0	18.8000	98.9800	TH	THA	admin	1.764663e+09	46.64	13.13	49.81	22.52	33.90	170	21.0	Hotel	Convention Center	Cafeteria	Building	Airport
17	0	6	Guangzhou	China	48600000.0	23.1450	113.3250	CN	CHN	admin	1.156237e+09	41.26	17.28	42.80	27.95	63.76	160	31.0	Hotel	Convention Center	Cafeteria	Building	Airport
14	1	3	Seoul	South Korea	24800000.0	37.5663	126.9997	KR	KOR	primary	1.410836e+09	85.50	30.45	103.69	45.88	76.81	186	5.0	Hotel	Convention Center	Cafeteria	Building	Airport
12	1	1	Tokyo	Japan	39800000.0	35.6850	139.7514	JP	JPN	primary	1.392686e+09	86.87	38.00	83.42	56.70	89.70	162	29.0	Hotel	Convention Center	Cafeteria	Building	Airport
11	2	0	Delhi	India	46960000.0	37.4306	-120.7759	US	USA	NaN	1.840019e+09	28.18	8.18	26.15	24.76	54.69	128	63.0	Hotel	Airport	Convention Center	Cafeteria	Building
10	2	0	Delhi	India	46960000.0	28.6700	77.2300	IN	IND	admin	1.356873e+09	28.18	8.18	26.15	24.76	54.69	128	63.0	Hotel	Airport	Convention Center	Cafeteria	Building
0	3	0	Jakarta	Indonesia	10135030.0	-6.1744	106.8294	ID	IDN	primary	1.360771e+09	44.48	19.42	45.41	25.06	27.34	118	73.0	Hotel	Convention Center	Cafeteria	Building	Airport
9	3	37	Phnom Penh	Cambodia	1242992.0	11.5500	104.9166	KH	KHM	primary	1.116261e+09	46.92	18.23	43.29	26.45	14.56	47	144.0	Hotel	Convention Center	Cafeteria	Building	Airport
5	3	9	Bandung	Indonesia	2575478.0	-6.9500	107.5700	ID	IDN	admin	1.360313e+09	35.71	9.69	36.86	16.95	25.78	118	73.0	Hotel	Convention Center	Cafeteria	Building	Airport
2	3	2	Ho Chi Minh City	Vietnam	7981411.0	10.7800	106.6950	VN	VNM	admin	1.704774e+09	38.20	16.74	36.30	20.71	26.64	121	70.0	Hotel	Convention Center	Cafeteria	Building	Airport
7	3	18	Manila	Philippines	1780148.0	14.6042	120.9822	PH	PHL	primary	1.608618e+09	38.99	21.33	33.32	25.61	25.14	96	95.0	Hotel	Convention Center	Cafeteria	Building	Airport
18	4	8	Hong Kong	Hong Kong	7300000.0	22.3050	114.1850	HK	HKG	NaN	1.344983e+09	77.22	79.57	75.94	54.36	65.32	188	3.0	Hotel	Convention Center	Cafeteria	Building	Airport
6	5	16	Kuala Lumpur	Malaysia	1808922.0	3.1667	101.7000	MY	MYS	primary	1.458989e+09	42.47	15.18	40.77	27.32	67.27	179	12.0	Hotel	Building	Convention Center	Cafeteria	Airport
4	6	4	Singapore	Singapore	5399000.0	1.2930	103.8558	SG	SGP	primary	1.702341e+09	81.10	63.27	66.75	58.99	88.96	189	2.0	Hotel	Building	Convention Center	Cafeteria	Airport
3	7	3	Hanoi	Vietnam	7067000.0	21.0333	105.8500	VN	VNM	primary	1.704414e+09	39.88	11.45	39.32	19.69	28.99	121	70.0	Hotel	Cafeteria	Convention Center	Building	Airport

C. Results

As mentioned, my objectives of the project is to show the similarity between cities by clustering between cities within South East Asia and versus several reference cities in Asia. The “similarity” of discussed factors including the similarity of nearby venues and other several key indices will help the business owner have the overall understanding how similar the targeted cities to the reference cities.

I’ve used the Folium map and update the color of the bubble with the cluster number. To make it easy for reader to identify the cluster, the additional HTML legend section were added on the bottom left corner of the map. Here’s the details:

- The radius represents the population of the cities
- Cluster number was presented in the different colors
- Pop-up information with: City name, Cluster label, population data, Ease of doing business Index.



Let’s have a look on each of the Cluster below to evaluate the similarity between cities

- Cluster 0:

	Cluster Labels	index_x	Name	Country	Population	lat	lng	iso2	iso3	capital	id	Cost of Living Index	Rent Index	Groceries Index	Restaurant Price Index	Local Purchasing Power Index	index_y	Ease of Doing Business	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	0	1	Bangkok	Thailand	8750600.0	13.7500	100.5166	TH	THA	primary	1.764069e+09	57.15	26.19	56.88	28.14	34.04	170	21.0	Hotel	Convention Center	Cafeteria	Building	Airport
8	0	19	Chiang Mai	Thailand	1763742.0	18.8000	98.9800	TH	THA	admin	1.764663e+09	46.64	13.13	49.81	22.52	33.90	170	21.0	Hotel	Convention Center	Cafeteria	Building	Airport
13	0	2	Mumbai	India	25000000.0	19.0170	72.8570	IN	IND	admin	1.356227e+09	28.84	21.02	26.86	26.61	41.68	128	63.0	Hotel	Convention Center	Cafeteria	Building	Airport
15	0	4	Shanghai	China	31100000.0	31.2165	121.4365	CN	CHN	admin	1.156074e+09	50.07	35.67	52.50	36.48	54.40	160	31.0	Hotel	Convention Center	Cafeteria	Building	Airport
16	0	5	Beijing	China	20700000.0	39.9289	116.3883	CN	CHN	primary	1.156229e+09	44.89	34.97	43.08	35.85	60.72	160	31.0	Hotel	Convention Center	Cafeteria	Building	Airport
17	0	6	Guangzhou	China	48600000.0	23.1450	113.3250	CN	CHN	admin	1.156237e+09	41.26	17.28	42.80	27.95	63.76	160	31.0	Hotel	Convention Center	Cafeteria	Building	Airport

- Cluster 1:

	Cluster Labels	index_x	Name	Country	Population	lat	lng	iso2	iso3	capital	id	Cost of Living Index	Rent Index	Groceries Index	Restaurant Price Index	Local Purchasing Power Index	index_y	Ease of Doing Business	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
12	1	1	Tokyo	Japan	39800000.0	35.6850	139.7514	JP	JPN	primary	1.392686e+09	86.87	38.00	83.42	56.70	89.70	162	29.0	Hotel	Convention Center	Cafeteria	Building	Airport
14	1	3	Seoul	South Korea	24800000.0	37.5663	126.9997	KR	KOR	primary	1.410836e+09	85.50	30.45	103.69	45.88	76.81	186	5.0	Hotel	Convention Center	Cafeteria	Building	Airport

- Cluster 2:

	Cluster Labels	index_x	Name	Country	Population	lat	lng	iso2	iso3	capital	id	Cost of Living Index	Rent Index	Groceries Index	Restaurant Price Index	Local Purchasing Power Index	index_y	Ease of Doing Business	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
10	2	0	Delhi	India	46960000.0	28.6700	77.2300	IN	IND	admin	1.356873e+09	28.18	8.18	26.15	24.76	54.69	128	63.0	Hotel	Airport	Convention Center	Cafeteria	Building

- Cluster 3:

	Cluster Labels	index_x	Name	Country	Population	lat	lng	iso2	iso3	capital	id	Cost of Living Index	Rent Index	Groceries Index	Restaurant Price Index	Local Purchasing Power Index	index_y	Ease of Doing Business	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	3	0	Jakarta	Indonesia	10135030.0	-6.1744	106.8294	ID	IDN	primary	1.360771e+09	44.48	19.42	45.41	25.06	27.34	118	73.0	Hotel	Convention Center	Cafeteria	Building	Airport
2	3	2	Ho Chi Minh City	Vietnam	7981411.0	10.7800	106.6950	VN	VNM	admin	1.704774e+09	38.20	16.74	36.30	20.71	26.64	121	70.0	Hotel	Convention Center	Cafeteria	Building	Airport
5	3	9	Bandung	Indonesia	2575478.0	-6.9500	107.5700	ID	IDN	admin	1.360313e+09	35.71	9.69	36.86	16.95	25.78	118	73.0	Hotel	Convention Center	Cafeteria	Building	Airport
7	3	18	Manila	Philippines	1780148.0	14.6042	120.9822	PH	PHL	primary	1.608618e+09	38.99	21.33	33.32	25.61	25.14	96	95.0	Hotel	Convention Center	Cafeteria	Building	Airport
9	3	37	Phnom Penh	Cambodia	1242992.0	11.5500	104.9166	KH	KHM	primary	1.116261e+09	46.92	18.23	43.29	26.45	14.56	47	144.0	Hotel	Convention Center	Cafeteria	Building	Airport

- Cluster 4:

	Cluster Labels	index_x	Name	Country	Population	lat	lng	iso2	iso3	capital	id	Cost of Living Index	Rent Index	Groceries Index	Restaurant Price Index	Local Purchasing Power Index	index_y	Ease of Doing Business	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
18	4	8	Hong Kong	Hong Kong	7300000.0	22.305	114.185	HK	HKG	NaN	1.344983e+09	77.22	79.57	75.94	54.36	65.32	188	3.0	Hotel	Convention Center	Cafeteria	Building	Airport

- Cluster 5:

	Cluster Labels	index_x	Name	Country	Population	lat	lng	iso2	iso3	capital	id	Cost of Living Index	Rent Index	Groceries Index	Restaurant Price Index	Local Purchasing Power Index	index_y	Ease of Doing Business	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
6	5	16	Kuala Lumpur	Malaysia	1808922.0	3.1667	101.7	MY	MYS	primary	1.458989e+09	42.47	15.18	40.77	27.32	67.27	179	12.0	Hotel	Building	Convention Center	Cafeteria	Airport

- Cluster 6

	Cluster Labels	index_x	Name	Country	Population	lat	lng	iso2	iso3	capital	id	Cost of Living Index	Rent Index	Groceries Index	Restaurant Price Index	Local Purchasing Power Index	index_y	Ease of Doing Business	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
4	6	4	Singapore	Singapore	5399000.0	1.293	103.8558	SG	SGP	primary	1.702341e+09	81.1	63.27	66.75	58.99	88.96	189	2.0	Hotel	Building	Convention Center	Cafeteria	Airport

- Cluster 7

	Cluster Labels	index_x	Name	Country	Population	lat	lng	iso2	iso3	capital	id	Cost of Living Index	Rent Index	Groceries Index	Restaurant Price Index	Local Purchasing Power Index	index_y	Ease of Doing Business	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
3	7	3	Hanoi	Vietnam	7067000.0	21.0333	105.85	VN	VNM	primary	1.704414e+09	39.88	11.45	39.32	19.69	28.99	121	70.0	Hotel	Cafeteria	Convention Center	Building	Airport

D. Discussion

With the clustering results specifically, we can see a clear similarity as well as dissimilarity between the Cluster. Hongkong, Seoul, Japan, Singapore (Cluster 4, 5, 6) stand out while most of the cities in South East Asia and India fall into the same clusters (see cluster 0 and cluster 3). Bangkok, Chiang Mai, Beijing, Mumbai are more or less the same. This result reflects correctly the reality. In another words, the business owner who already know how difficult to set up a business in Jakarta Indonesia shall expect to see the same difficulty level in Ho Chi Minh City, Vietnam.

In reality, the comparison will be much more complicated and involved much more other factors. Even within a city, different areas already have different structures, cost of living, and different business setup process. The business analysis shall need to study in details including survey-ing the areas to narrow down the area of interest. Once the areas are picked, he or she shall re-run the Foursquare API call with multiple different locations around the area in order to achieve more venue information.

In this studying, I used the K-Means algorithm as part of this clustering study. The nature of K-means algorithm treats all of the features fairly and sets weights of all features equally when evaluating dissimilarity. In fact, experiment results show that a meaningful clustering phenomenon often occurs in a subspace defined by some specific features. Hence the K-means with weight shall be the better methodology for the next project development.

F. Conclusion

As a result, it's clear to see the cluster between cities in South East Asia and other cities in Asia. The business owner can use this clustering result to consider a location that best fits the requirement. Either it's in Singapore with all of the best indices but costly or either Bangkok, Thailand which is pretty similar to Beijing, China or even Ho Chi Minh City that similar to Manilla, Indonesia.

I ended the study by visualizing the data and clustering information of South East Asia Cities. I hope you find this study useful for you.

Thank you,

Viet Kieu
kieuxuanhoangviet@gmail.com

G. References:

[1] Foursquare API *https://developer.foursquare.com/docs/api/venues/explore*

[2] Metropolitan areas in Asia: *https://en.wikipedia.org/wiki/List_of_metropolitan_areas_in_Asia*

[3] Cost of living Index:
https://www.numbeo.com/cost-of-living/region_rankings.jsp?title=2019&region=142

[4] Ease of doing business
https://www.doingbusiness.org/en/rankings

[5] Longitude and Latitude of all cities: *https://simplemaps.com*