Jakarta – Seoul Sister City
Similarities and Dissimilarities

Anang Syarifudin A.

1. Introduction

1.1 Background

Jakarta is the capital of Indonesia, with population of 10 Million in area of 662.3 km2, Jakarta is a very dynamic and fast-growing urban area. Indonesia in general has been a magnet for foreign investment and Jakarta especially has become main target for foreign investment. Alongside with inter-country cooperation, many cities also form cooperation between cities known as "Sister City" to foster cooperation and accelerate economic cooperation. Jakarta has no exception also has many Sister Cities widespread across continents.

One of Jakarta's sister city is Seoul, capital of South Korea which is the object of this analysis. Seoul was chosen mainly because it has demographic similarities with Jakarta. With population of 9.7 million in area of 605.2 km2 quite the same as Jakarta. While South Korea is a developed country, Seoul as its capital supposedly more advance compare to Jakarta.

1.2 Problem

Have been cooperated as Sister City since 1984, Jakarta and Seoul need to tighten their cooperation and exploring new opportunities and learn from each other. This simple study is main objective is to answer question about what the similarities and dissimilarities between Jakarta and Seoul are, also what Jakarta Administration need to learn from Seoul.

2. Data Acquisition and Cleaning

2.1 Data Source

The main data sources for this study is Wikipedia for neighbourhood information of both cities. Jakarta neighbourhood was scrapped from this page and Seoul neighbourhood was scrapped from this page. Second source is venues of both cities from Foursquare. I also utilize Nominatim geocoder for getting latitude and longitude of the neighbourhood.

2.1 Data Cleaning

Wikipedia page that contains list of Jakarta neighbourhood has 7 tables alongside with other texts and images. First table contains information about 6 Jakarta district and the rest are table about Kecamatan (sub-district) of each district. The web page is scrapped using pandas built-in function for web scraping: read_html. This function return list contains 7 dafaframes that correspond with 7 tables.

From 7 tables I only interested in 5 tables. Since I work in sub-district level, I ignore the first dataframe. I also ignore the last dataframe since it contains sub-district of Kepulauan Seribu which is an archipelago not an urban area. Then I merge 5 dataframes into single dataframe and remove extra row containing summary of each districts. Other transformation that done is removing columns other than Kecamatan name. The result is 42 rows x 1 column dataframe.

Seoul neighbourhood can be scrapped easily from Wikipedia page, since it only contains 1 table. I just need to keep the column contains district name and remove other columns. The result is 24 rows x 1 column dataframe.

The latitude and longitude of each neighbourhood is acquired with help of Nominatim geocoder. I use name of Kecamatan or district and appended with ‘Jakarta’ or ‘Seoul’ as search queries. I also utilise RateLimiter function for avoiding API call rate limiter.

Information about venues of a neighbourhood provided by Foursquare through explore API. This API requires at least longitude, latitude and radius as input of the calls. The output of the calls is list of the venues with its attribute such as address, contacts, coordinates and categories. For free account Foursquare limits not only the total calls, total calls per day but also maximum number of venues per call. If I try to fetch all the related venues and group it by certain category, I might ended with incomplete data since I only use free account. Luckily the explore API provide information about the total result. I also can add categories as input variable of the calls.

For this analysis I need information about number of venues of the following categories: Entertainments, Educations, Outdoor & Recreations, Government Buildings, Offices, Factories, Food, Shopping & Services and Medical Facilities.

Getting data from Foursquare with free account is a bit tricky, not only I have to deal with rate limiter but also, I need to retry is the call was unsuccessful. After looping through each neighbourhood, I have a complete set of data for further processing.

3. Methodology

In order to answer the research question, I will use unsupervised machine learning method specifically Kmeans clustering. Since I goal is creating segments of neighbourhood from both cities. After acquiring and cleaning the data, I go through exploratory data analysis. In the EDA process I might get some insight about the data and its distribution. I might also need feature engineering in order to make the Kmeans algorithm works properly.

The process of segmenting neighbourhood starts with finding optimum number of clusters with elbow method. Then clustering neighbourhood to k number of clusters based on elbow method selection.

After neighbourhood cluster is created then I need to analyse the cluster property and check whether the segmentation is making sense.

I proceed the clustering process on both cities and compare the segmentations between the cities, seek the similarities and dissimilarities, make conclusion and recommendations.

4. Exploratory Data Analysis

4.1 Jakarta Neighbourhood Venues

After I got complete set of venues for each Jakarta neighbourhood, I do simple descriptive statistics and univariate analysis.

Most of the neighbourhood has not more than 26 entertainment venues, but some neighbourhood has entertainment venue more than twice of the average. Number of education venue are from 20-40 venue distributed quite normally. High variation of number of food related venue ranging as low as 4 and as high as 239. Outdoor and recreation venue is right skewed with median of 32. Please find histogram of those venue categories on fig. 1.

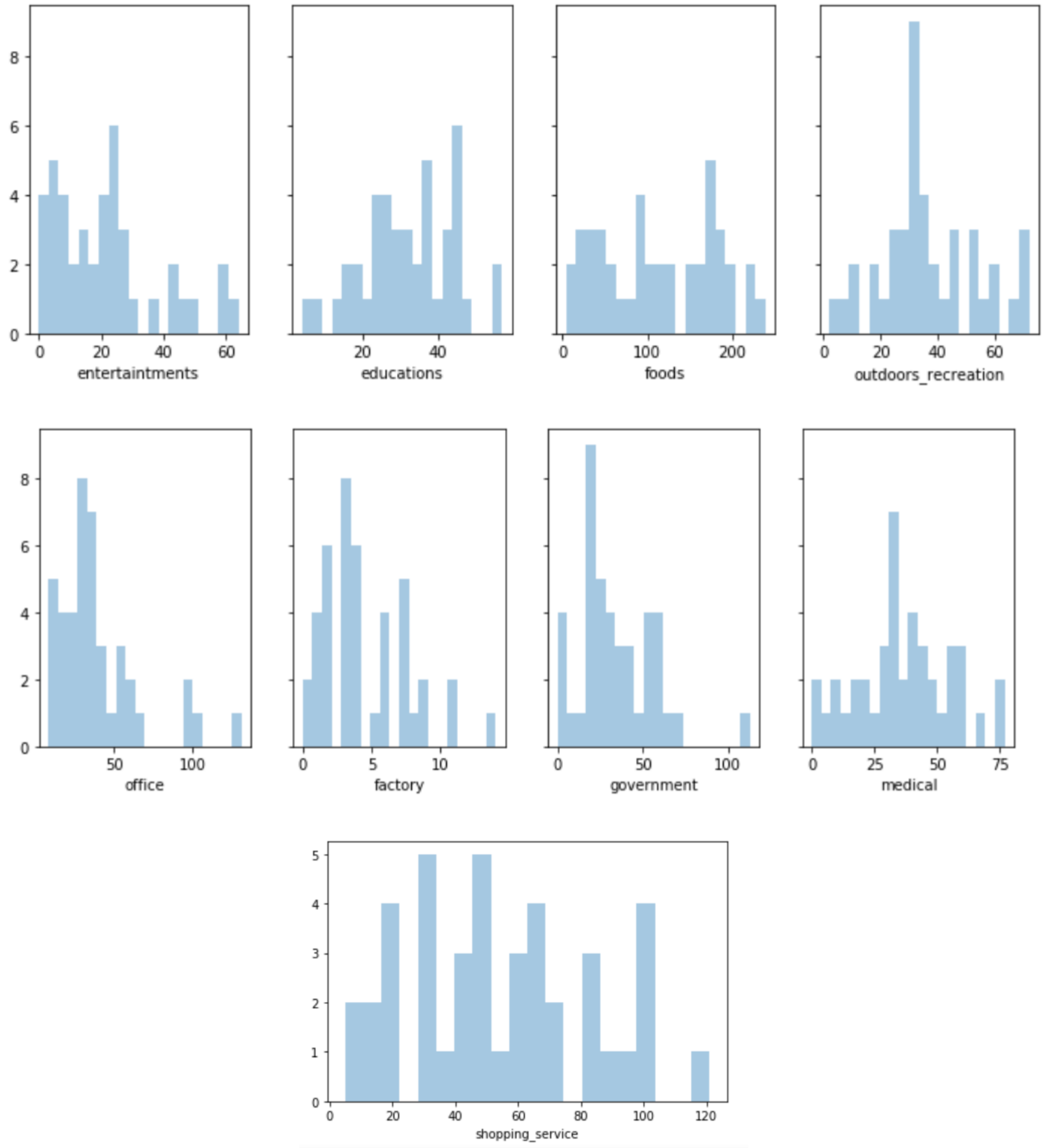
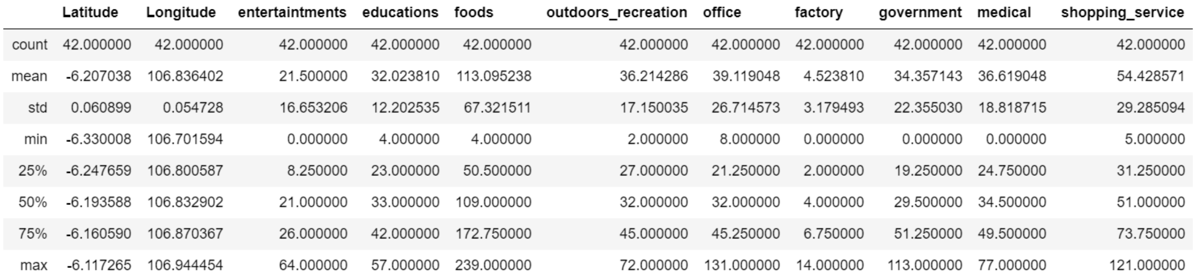


Fig. 1 Histogram

Office, government building and shopping & service venue relatively has same range, while medical facilities has shorter range. The number of factories has lowest number and shorter range amongst other variables. I can summarize the analysis in table 1 which is produced by pandas describe functions.

Table 1 Descriptive Statistics Summary of Jakarta neighbourhood venues

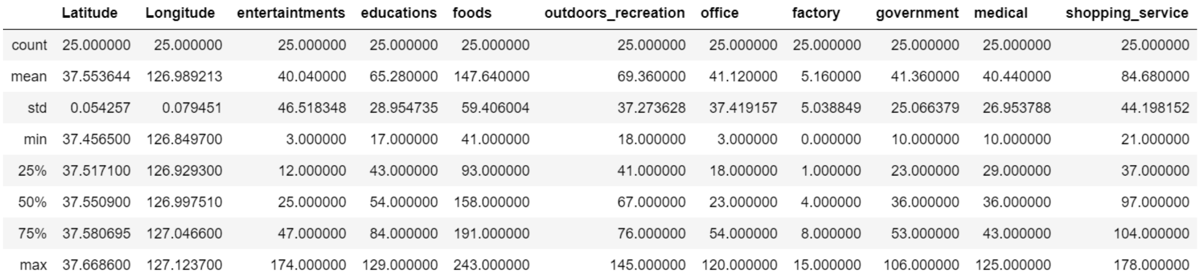


	Latitude	Longitude	entertaintments	educations	foods	outdoors_recreation	office	factory	government	medical	shopping_service
count	42.000000	42.000000	42.000000	42.000000	42.000000	42.000000	42.000000	42.000000	42.000000	42.000000	42.000000
mean	-6.207038	106.836402	21.500000	32.023810	113.095238	36.214286	39.119048	4.523810	34.357143	36.619048	54.428571
std	0.060899	0.054728	16.653206	12.202535	67.321511	17.150035	26.714573	3.179493	22.355030	18.818715	29.285094
min	-6.330008	106.701594	0.000000	4.000000	4.000000	2.000000	8.000000	0.000000	0.000000	0.000000	5.000000
25%	-6.247659	106.800587	8.250000	23.000000	50.500000	27.000000	21.250000	2.000000	19.250000	24.750000	31.250000
50%	-6.193588	106.832902	21.000000	33.000000	109.000000	32.000000	32.000000	4.000000	29.500000	34.500000	51.000000
75%	-6.160590	106.870367	26.000000	42.000000	172.750000	45.000000	45.250000	6.750000	51.250000	49.500000	73.750000
max	-6.117265	106.944454	64.000000	57.000000	239.000000	72.000000	131.000000	14.000000	113.000000	77.000000	121.000000

4.2 Seoul Neighbourhood Venue

Then number of venues has higher average than Jakarta the range is also longer. I can find the summary on table 2 below.

Table 2 Descriptive Statistics Summary of Seoul neighbourhood venues



	Latitude	Longitude	entertaintments	educations	foods	outdoors_recreation	office	factory	government	medical	shopping_service
count	25.000000	25.000000	25.000000	25.000000	25.000000	25.000000	25.000000	25.000000	25.000000	25.000000	25.000000
mean	37.553644	126.989213	40.040000	65.280000	147.640000	69.360000	41.120000	5.160000	41.360000	40.440000	84.680000
std	0.054257	0.079451	46.518348	28.954735	59.406004	37.273628	37.419157	5.038849	25.066379	26.953788	44.198152
min	37.456500	126.849700	3.000000	17.000000	41.000000	18.000000	3.000000	0.000000	10.000000	10.000000	21.000000
25%	37.517100	126.929300	12.000000	43.000000	93.000000	41.000000	18.000000	1.000000	23.000000	29.000000	37.000000
50%	37.550900	126.997510	25.000000	54.000000	158.000000	67.000000	23.000000	4.000000	36.000000	36.000000	97.000000
75%	37.580695	127.046600	47.000000	84.000000	191.000000	76.000000	54.000000	8.000000	53.000000	43.000000	104.000000
max	37.668600	127.123700	174.000000	129.000000	243.000000	145.000000	120.000000	15.000000	106.000000	125.000000	178.000000

5. Clustering and Segmentation

5.1 Scaling Numbers

Luckily, I only have numeric data for clustering, no need for advance encoding techniques and transformations. Before using Kmeans for clustering the neighbourhood, I need to transform the data in order to make the algorithm works properly. I use Standard scaler from Scikit learn package to transform both Jakarta and Seoul datasets.

5.2 Finding Optimum Cluster of Jakarta and Seoul Neighbourhood

In order to get distinguishable clusters, I have to find the optimum k number of clusters using elbow method. I iterate the process of clustering from 1 to 9 clusters and check the **Sum of Square Error (SSE)** and plot it.

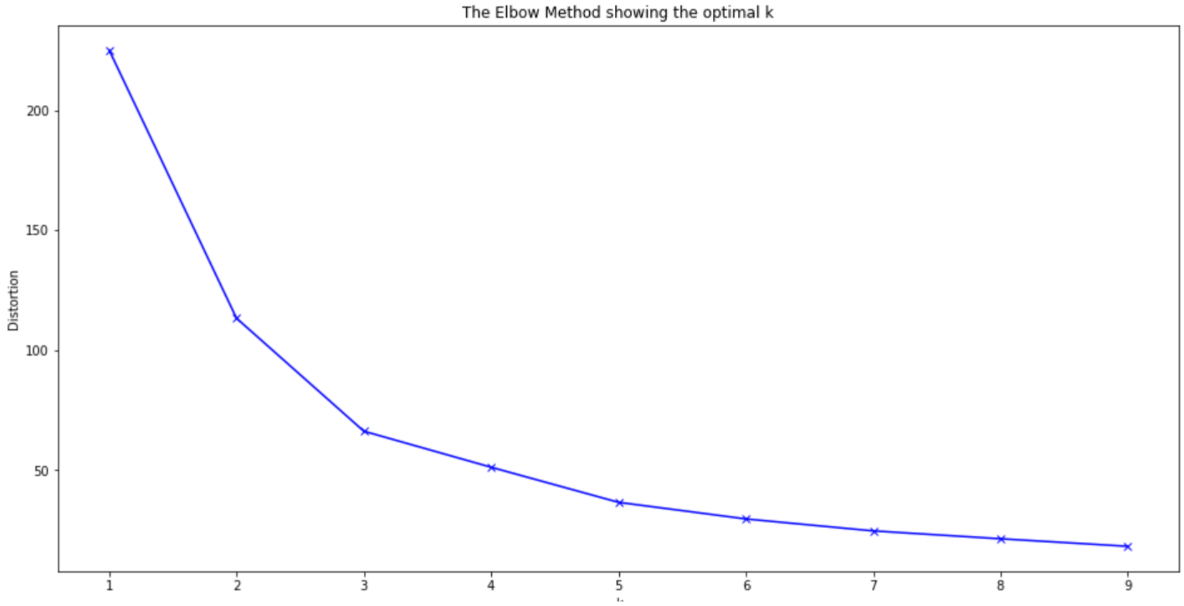


Fig. 2 Elbow Chart of Jakarta Neighbourhood Clusters

From the chart that shown on figure 2 I can see that the “elbow” is at k=3. But after I tried to cluster using this k, the difference between clusters is not so obvious. So, I repeat the process with k=4.

I also find the optimum cluster for Seoul neighbourhood with same method.

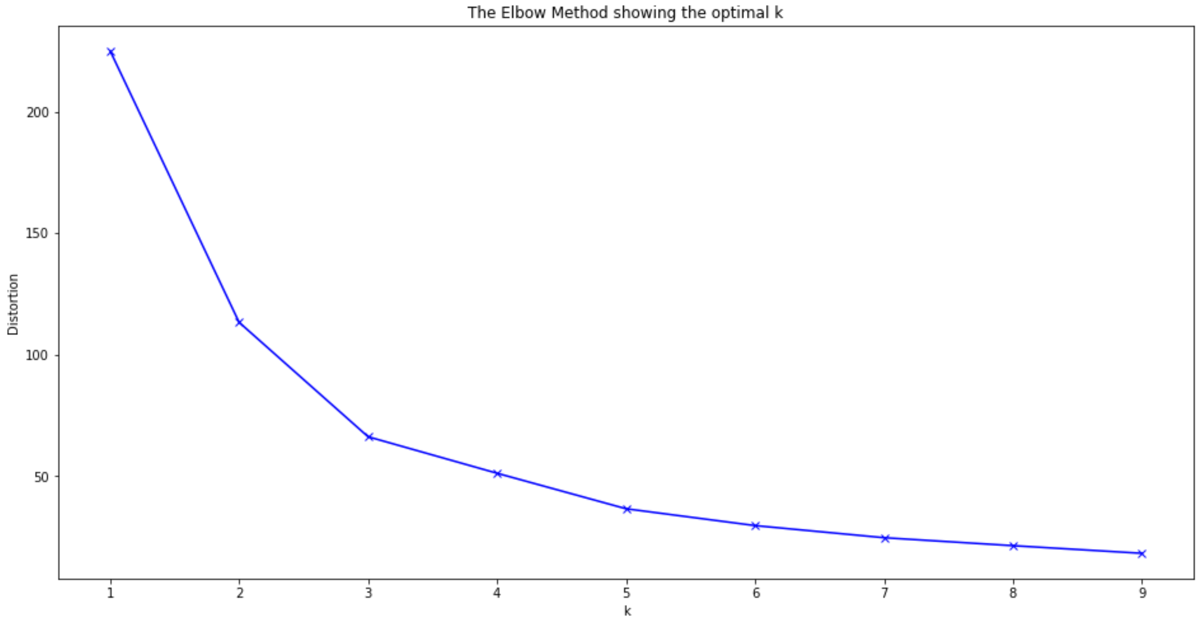


Fig. 5 Elbow Chart of Seoul Neighbourhood Clusters

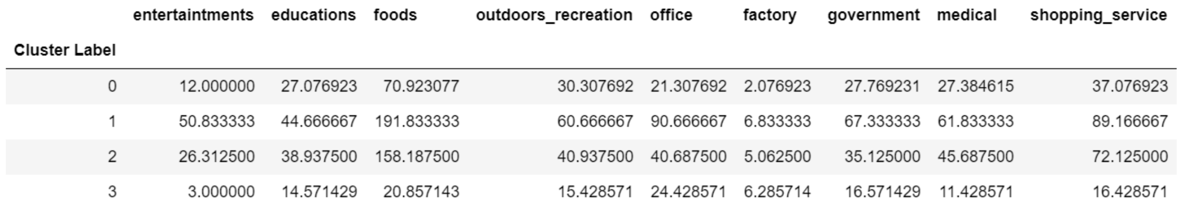
The result is quite the same with Jakarta, I can see that the “elbow” is at k=3. But after I tried to cluster using this k, the difference between clusters is not so obvious. So, I repeat the process with k=4.

6. Results

6.1 Jakarta Neighbourhood Clusters and Segmentation

As the result of the Kmeans clustering, I got 13 neighbourhoods segmented as cluster 0, 6 neighbourhoods of cluster 1, 16 cluster 2 and 7 cluster 3. Summary of each clusters can be found in the following table.

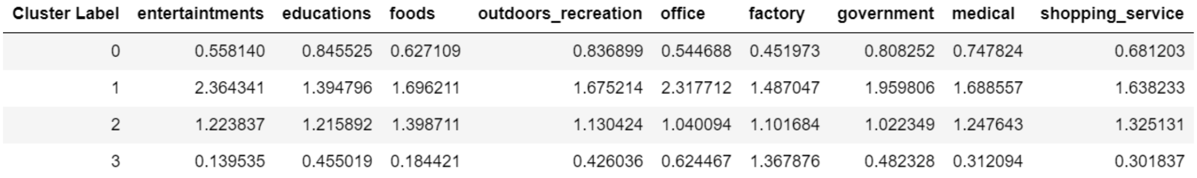
Table 3 Jakarta Neighbourhood Clusters Summary



	entertaintments	educations	foods	outdoors_recreation	office	factory	government	medical	shopping_service
Cluster Label									
0	12.000000	27.076923	70.923077	30.307692	21.307692	2.076923	27.769231	27.384615	37.076923
1	50.833333	44.666667	191.833333	60.666667	90.666667	6.833333	67.333333	61.833333	89.166667
2	26.312500	38.937500	158.187500	40.937500	40.687500	5.062500	35.125000	45.687500	72.125000
3	3.000000	14.571429	20.857143	15.428571	24.428571	6.285714	16.571429	11.428571	16.428571

I can compare average number of venue per category for each cluster to see difference cluster characteristics. But it is still hard enough compare, I need to compare each number with total population average as a baseline. The result is index instead of raw number then I plot it as a radar chart.

Table 4 Jakarta Neighbourhood Clusters Index



Cluster Label	entertaintments	educations	foods	outdoors_recreation	office	factory	government	medical	shopping_service
0	0.558140	0.845525	0.627109	0.836899	0.544688	0.451973	0.808252	0.747824	0.681203
1	2.364341	1.394796	1.696211	1.675214	2.317712	1.487047	1.959806	1.688557	1.638233
2	1.223837	1.215892	1.398711	1.130424	1.040094	1.101684	1.022349	1.247643	1.325131
3	0.139535	0.455019	0.184421	0.426036	0.624467	1.367876	0.482328	0.312094	0.301837

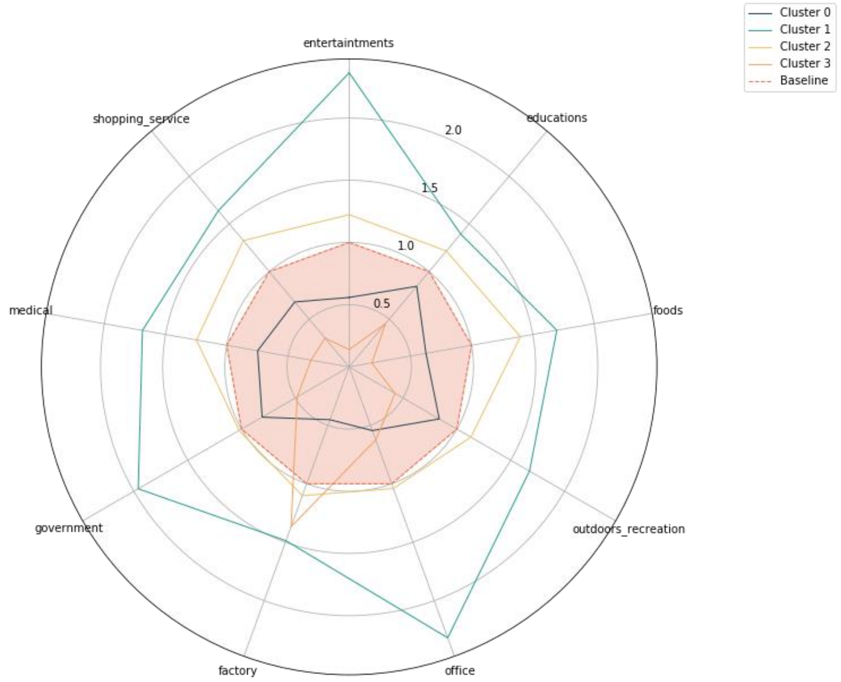
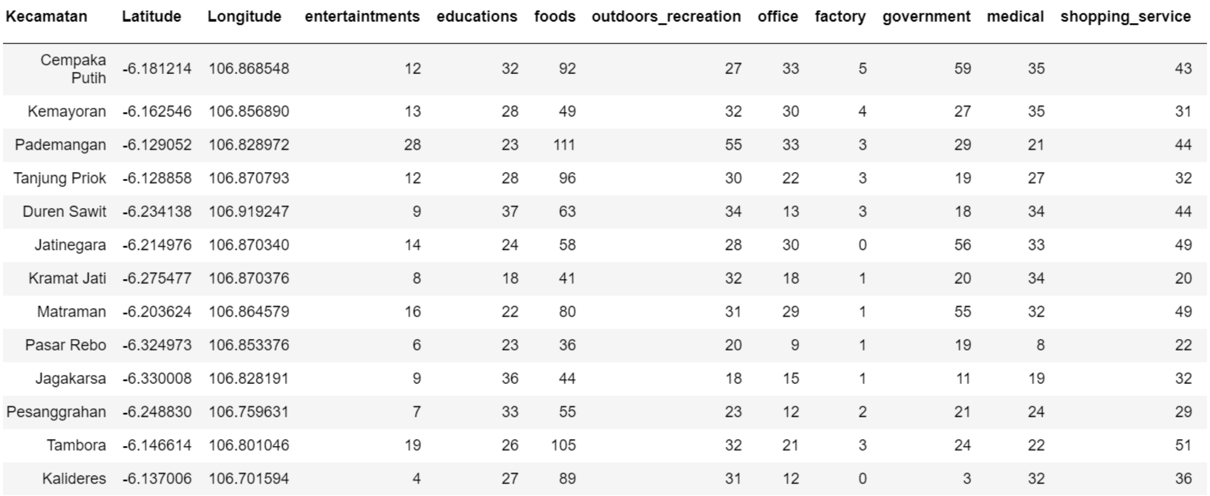


Fig. 3 Jakarta Neighbour Cluster Radar Chart

I can clearly see difference between clusters. There are cluster that over index, another under index cluster, average cluster and cluster that under index in all categories but factory category.

Then I examine member of each clusters to confirm whether the segmentation make sense.

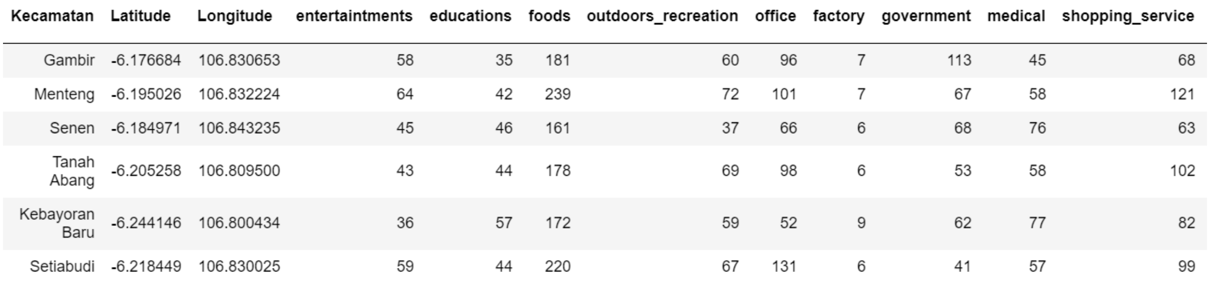
Table 5 Jakarta Neighbourhood Cluster 0



Kecamatan	Latitude	Longitude	entertaintments	educations	foods	outdoors_recreation	office	factory	government	medical	shopping_service
Cempaka Putih	-6.181214	106.868548	12	32	92	27	33	5	59	35	43
Kemayoran	-6.162546	106.856890	13	28	49	32	30	4	27	35	31
Pademangan	-6.129052	106.828972	28	23	111	55	33	3	29	21	44
Tanjung Priok	-6.128858	106.870793	12	28	96	30	22	3	19	27	32
Duren Sawit	-6.234138	106.919247	9	37	63	34	13	3	18	34	44
Jatinegara	-6.214976	106.870340	14	24	58	28	30	0	56	33	49
Kramat Jati	-6.275477	106.870376	8	18	41	32	18	1	20	34	20
Matraman	-6.203624	106.864579	16	22	80	31	29	1	55	32	49
Pasar Rebo	-6.324973	106.853376	6	23	36	20	9	1	19	8	22
Jagakarsa	-6.330008	106.828191	9	36	44	18	15	1	11	19	32
Pesanggrahan	-6.248830	106.759631	7	33	55	23	12	2	21	24	29
Tambora	-6.146614	106.801046	19	26	105	32	21	3	24	22	51
Kalideres	-6.137006	106.701594	4	27	89	31	12	0	3	32	36

Cluster 0 is under-indexing in all categories, I can confirm that member of this cluster is sub-urban area which typically has less facility. Kecamatan on northern and eastern part of Jakarta are the member of this cluster.

Table 6 Jakarta Neighbourhood Cluster 1



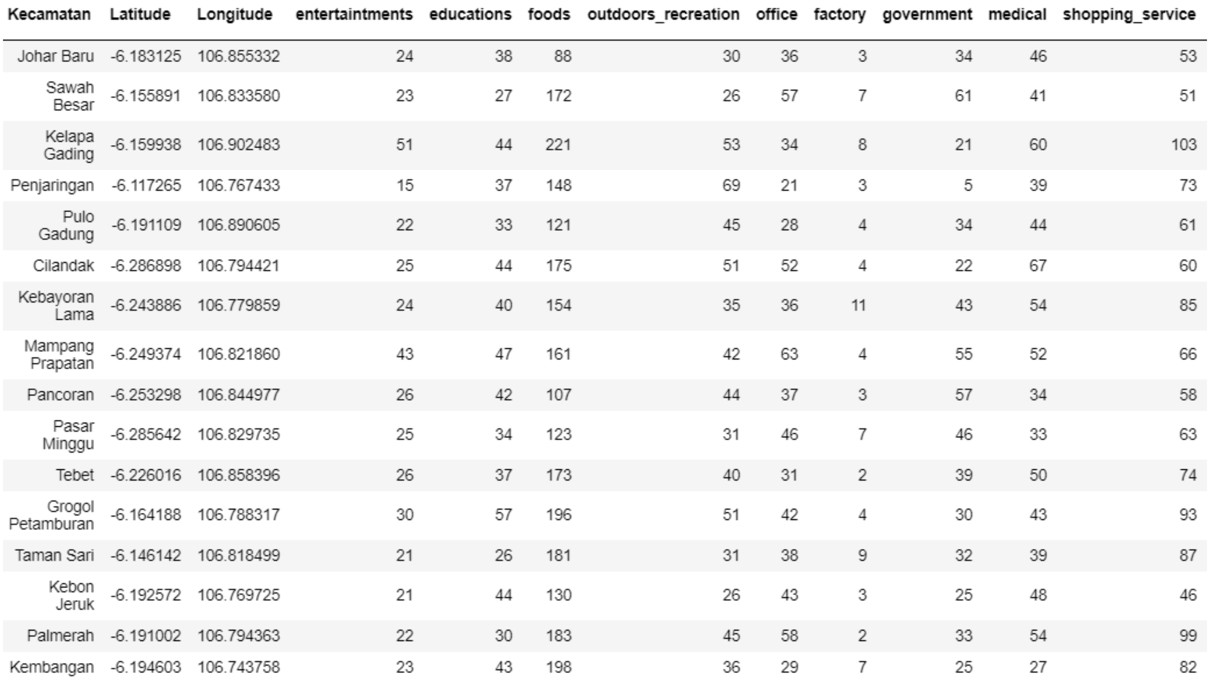
Kecamatan	Latitude	Longitude	entertaintments	educations	foods	outdoors_recreation	office	factory	government	medical	shopping_service
Gambir	-6.176684	106.830653	58	35	181	60	96	7	113	45	68
Menteng	-6.195026	106.832224	64	42	239	72	101	7	67	58	121
Senen	-6.184971	106.843235	45	46	161	37	66	6	68	76	63
Tanah Abang	-6.205258	106.809500	43	44	178	69	98	6	53	58	102
Kebayoran Baru	-6.244146	106.800434	36	57	172	59	52	9	62	77	82
Setiabudi	-6.218449	106.830025	59	44	220	67	131	6	41	57	99

All member of the cluster 1 are located in the center of the city and has more facility compared to other area. This is the most active cluster compare the other clusters. Government buildings and commercial are also concentrated within these cluster.

Old settlement in southern and western part of Jakarta and has average number of venues for each category are grouped in cluster 2. This cluster has most member compare to other cluster which is made it perfect sense.

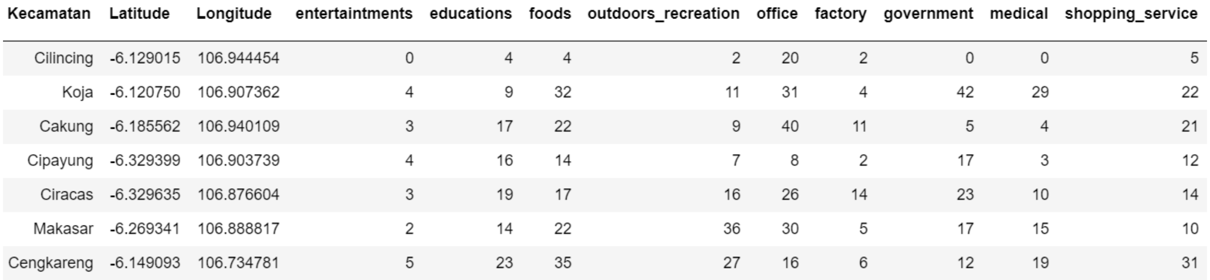
The last cluster, cluster 3 consists of industrial area. Generally, this area has less facility other than factories.

Table 7 Jakarta Neighbourhood Cluster 2



Kecamatan	Latitude	Longitude	entertaintments	educations	foods	outdoors_recreation	office	factory	government	medical	shopping_service
Johar Baru	-6.183125	106.855332	24	38	88	30	36	3	34	46	53
Sawah Besar	-6.155891	106.833580	23	27	172	26	57	7	61	41	51
Kelapa Gading	-6.159938	106.902483	51	44	221	53	34	8	21	60	103
Penjaringan	-6.117265	106.767433	15	37	148	69	21	3	5	39	73
Pulo Gadung	-6.191109	106.890605	22	33	121	45	28	4	34	44	61
Cilandak	-6.286898	106.794421	25	44	175	51	52	4	22	67	60
Kebayoran Lama	-6.243886	106.779859	24	40	154	35	36	11	43	54	85
Mampang Prapatan	-6.249374	106.821860	43	47	161	42	63	4	55	52	66
Pancoran	-6.253298	106.844977	26	42	107	44	37	3	57	34	58
Pasar Minggu	-6.285642	106.829735	25	34	123	31	46	7	46	33	63
Tebet	-6.226016	106.858396	26	37	173	40	31	2	39	50	74
Grogol Petamburan	-6.164188	106.788317	30	57	196	51	42	4	30	43	93
Taman Sari	-6.146142	106.818499	21	26	181	31	38	9	32	39	87
Kebon Jeruk	-6.192572	106.769725	21	44	130	26	43	3	25	48	46
Palmerah	-6.191002	106.794363	22	30	183	45	58	2	33	54	99
Kembangan	-6.194603	106.743758	23	43	198	36	29	7	25	27	82

Table 8 Jakarta Neighbourhood Cluster 3



Kecamatan	Latitude	Longitude	entertaintments	educations	foods	outdoors_recreation	office	factory	government	medical	shopping_service
Cilincing	-6.129015	106.944454	0	4	4	2	20	2	0	0	5
Koja	-6.120750	106.907362	4	9	32	11	31	4	42	29	22
Cakung	-6.185562	106.940109	3	17	22	9	40	11	5	4	21
Cipayung	-6.329399	106.903739	4	16	14	7	8	2	17	3	12
Ciracas	-6.329635	106.876604	3	19	17	16	26	14	23	10	14
Makasar	-6.269341	106.888817	2	14	22	36	30	5	17	15	10
Cengkareng	-6.149093	106.734781	5	23	35	27	16	6	12	19	31

The plot of the areas and corresponding clusters can be found in the figure 4 below.

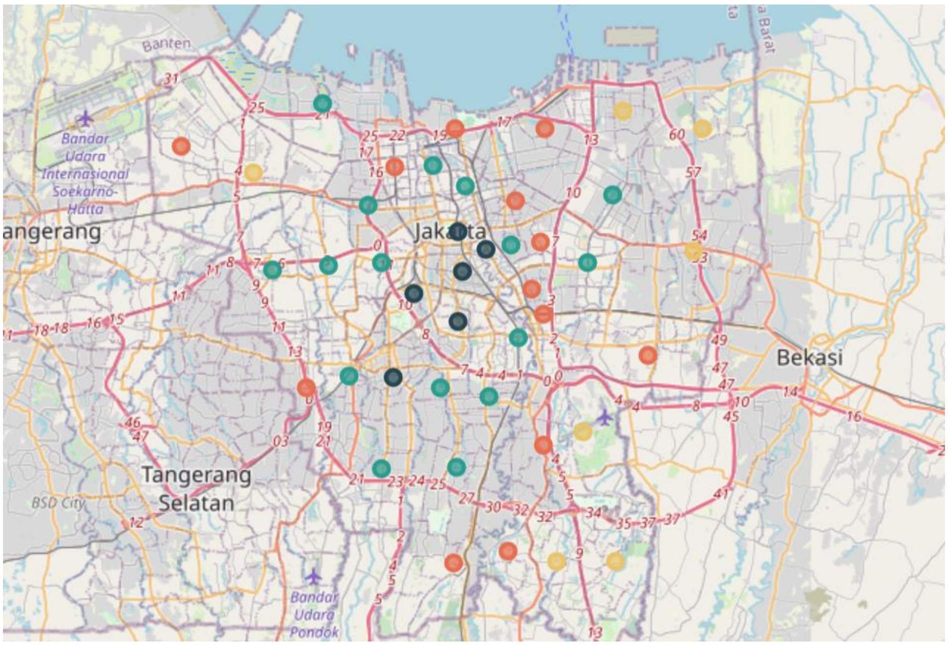
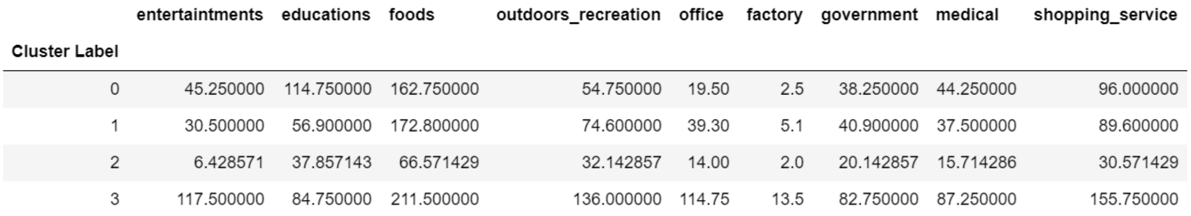


Fig. 4 Jakarta Neighbourhood Clusters

6.2 Seoul Neighbourhood Clusters and Segmentations

After clustering Seoul neighbourhoods with Kmeans, I got 4 neighbourhoods segmented as cluster 0, 10 neighbourhoods of cluster 1, 7 cluster 2 and 4 cluster 3. Summary of each clusters can be found in the table 9.

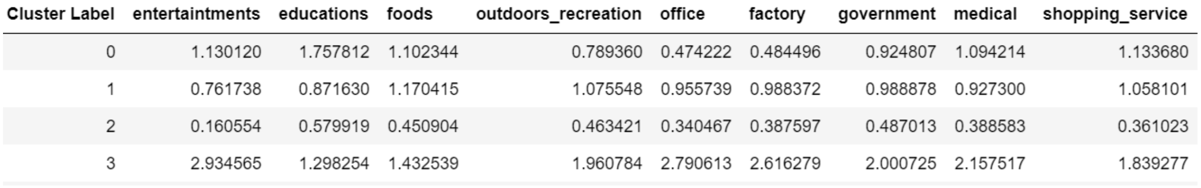
Table 9 Seoul Neighbourhood Clusters Summary



	entertaintments	educations	foods	outdoors_recreation	office	factory	government	medical	shopping_service
Cluster Label									
0	45.250000	114.750000	162.750000	54.750000	19.50	2.5	38.250000	44.250000	96.000000
1	30.500000	56.900000	172.800000	74.600000	39.30	5.1	40.900000	37.500000	89.600000
2	6.428571	37.857143	66.571429	32.142857	14.00	2.0	20.142857	15.714286	30.571429
3	117.500000	84.750000	211.500000	136.000000	114.75	13.5	82.750000	87.250000	155.750000

I have to compare it to the entire population in order to get the index as shown in table 10, then plot it into a radar chart.

Table 10 Seoul Neighbourhood Clusters Index



Cluster Label	entertaintments	educations	foods	outdoors_recreation	office	factory	government	medical	shopping_service
0	1.130120	1.757812	1.102344	0.789360	0.474222	0.484496	0.924807	1.094214	1.133680
1	0.761738	0.871630	1.170415	1.075548	0.955739	0.988372	0.988878	0.927300	1.058101
2	0.160554	0.579919	0.450904	0.463421	0.340467	0.387597	0.487013	0.388583	0.361023
3	2.934565	1.298254	1.432539	1.960784	2.790613	2.616279	2.000725	2.157517	1.839277

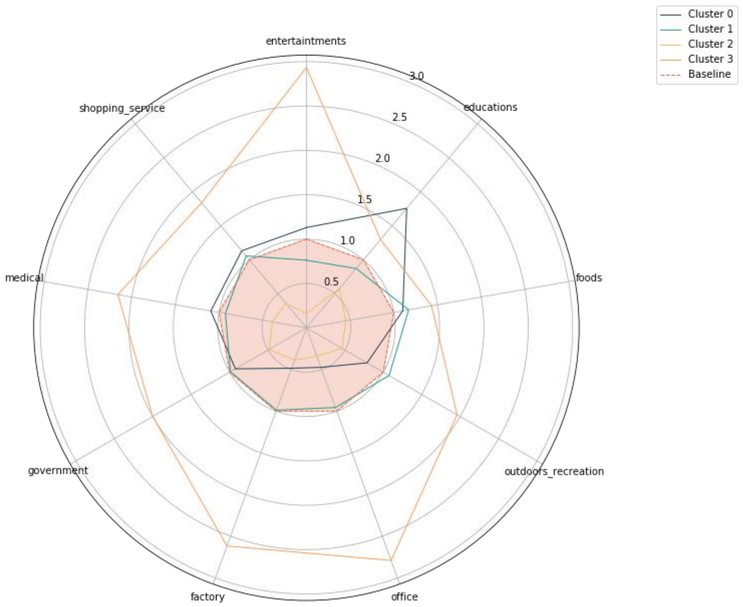
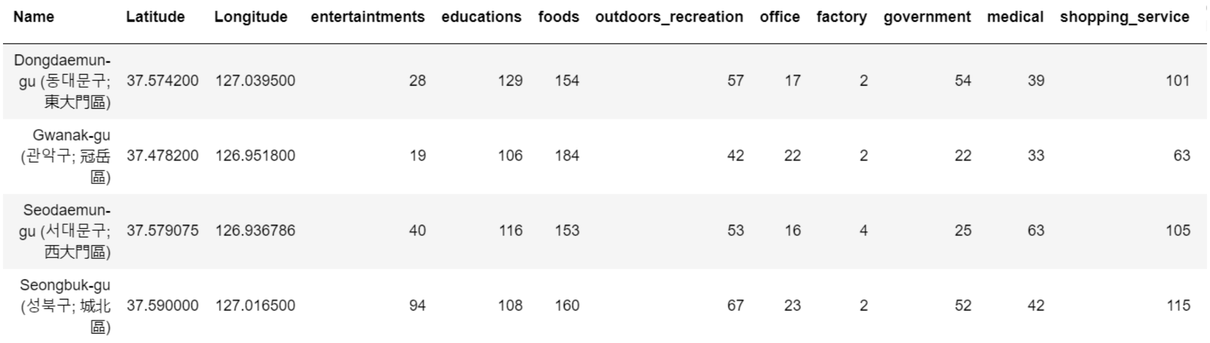


Fig. 5 Seoul Neighbour Cluster Radar Chart

If I examine cluster 0, the characteristic is it has over-indexing on education venues. Member of this cluster are area in second ring of the city.

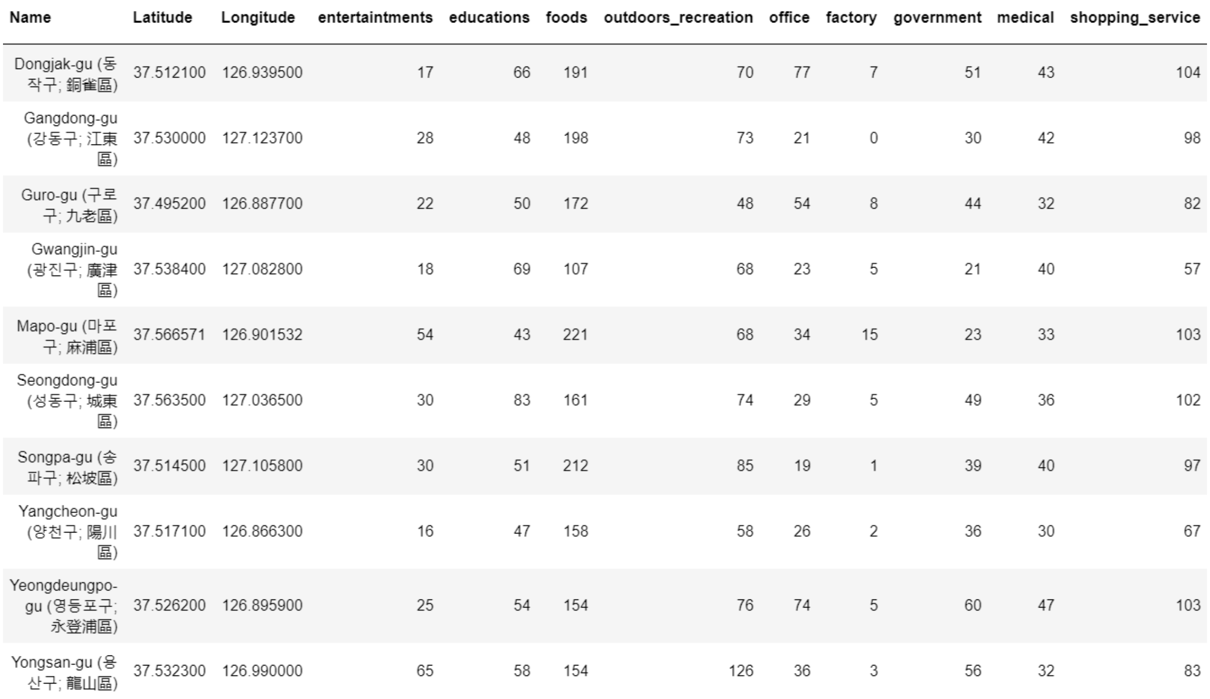
Table 11 Seoul Neighbourhood Cluster 0



Name	Latitude	Longitude	entertaintments	educations	foods	outdoors_recreation	office	factory	government	medical	shopping_service
Dongdaemun-gu (동대문구; 東大門區)	37.574200	127.039500	28	129	154	57	17	2	54	39	101
Gwanak-gu (관악구; 冠岳區)	37.478200	126.951800	19	106	184	42	22	2	22	33	63
Seodaemun-gu (서대문구; 西大門區)	37.579075	126.936786	40	116	153	53	16	4	25	63	105
Seongbuk-gu (성북구; 城北區)	37.590000	127.016500	94	108	160	67	23	2	52	42	115

Cluster 1 is where most of the neighbourhoods fall into, these are the average neighbourhood. It spreads across the city.

Table 12 Seoul Neighbourhood Cluster 1

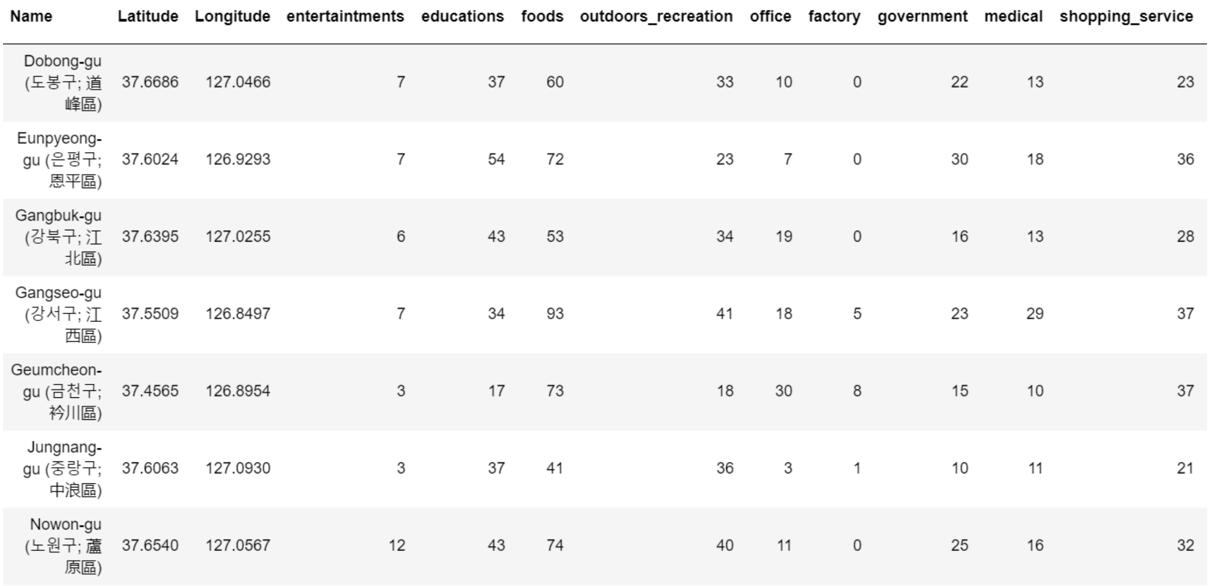


Name	Latitude	Longitude	entertaintments	educations	foods	outdoors_recreation	office	factory	government	medical	shopping_service
Dongjak-gu (동작구; 銅雀區)	37.512100	126.939500	17	66	191	70	77	7	51	43	104
Gangdong-gu (강동구; 江東區)	37.530000	127.123700	28	48	198	73	21	0	30	42	98
Guro-gu (구로구; 九老區)	37.495200	126.887700	22	50	172	48	54	8	44	32	82
Gwangjin-gu (광진구; 廣津區)	37.538400	127.082800	18	69	107	68	23	5	21	40	57
Mapo-gu (마포구; 麻浦區)	37.566571	126.901532	54	43	221	68	34	15	23	33	103
Seongdong-gu (성동구; 城東區)	37.563500	127.036500	30	83	161	74	29	5	49	36	102
Songpa-gu (송파구; 松坡區)	37.514500	127.105800	30	51	212	85	19	1	39	40	97
Yangcheon-gu (양천구; 陽川區)	37.517100	126.866300	16	47	158	58	26	2	36	30	67
Yeongdeungpo-gu (영등포구; 永登浦區)	37.526200	126.895900	25	54	154	76	74	5	60	47	103
Yongsan-gu (용산구; 龍山區)	37.532300	126.990000	65	58	154	126	36	3	56	32	83

Cluster 2 mostly consists of sub-urban areas which located in the outer part of the city. This cluster has under-indexing for all categories.

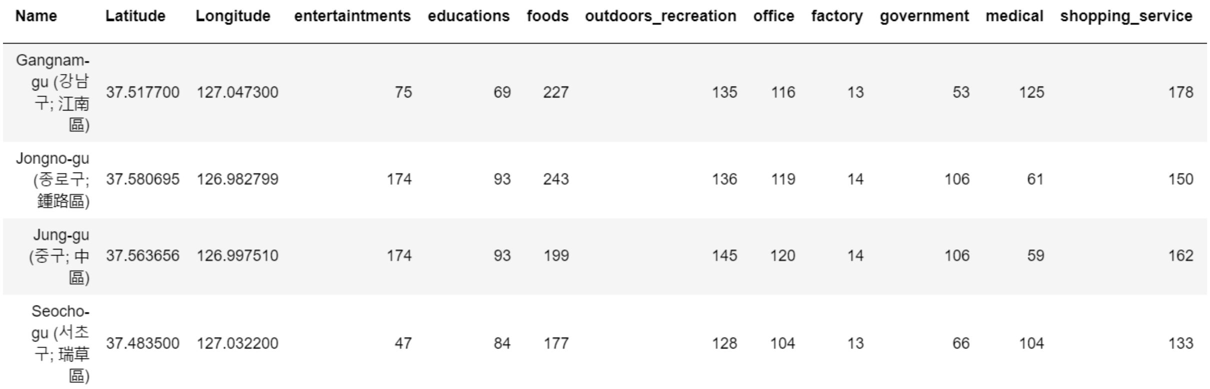
The most active cluster is cluster 3 which has more venues for all categories compared to other clusters. The member of this cluster including the famous Gangnam district.

Table 13 Seoul Neighbourhood Cluster 2



Name	Latitude	Longitude	entertaintments	educations	foods	outdoors_recreation	office	factory	government	medical	shopping_service
Dobong-gu (도봉구; 道峰區)	37.6686	127.0466	7	37	60	33	10	0	22	13	23
Eunpyeong-gu (은평구; 恩平區)	37.6024	126.9293	7	54	72	23	7	0	30	18	36
Gangbuk-gu (강북구; 江北區)	37.6395	127.0255	6	43	53	34	19	0	16	13	28
Gangseo-gu (강서구; 江西區)	37.5509	126.8497	7	34	93	41	18	5	23	29	37
Geumcheon-gu (금천구; 衿川區)	37.4565	126.8954	3	17	73	18	30	8	15	10	37
Jungnang-gu (중랑구; 中浪區)	37.6063	127.0930	3	37	41	36	3	1	10	11	21
Nowon-gu (노원구; 蘆原區)	37.6540	127.0567	12	43	74	40	11	0	25	16	32

Table 14 Seoul Neighbourhood Cluster 3



Name	Latitude	Longitude	entertaintments	educations	foods	outdoors_recreation	office	factory	government	medical	shopping_service
Gangnam-gu (강남구; 江南區)	37.517700	127.047300	75	69	227	135	116	13	53	125	178
Jongno-gu (종로구; 鍾路區)	37.580695	126.982799	174	93	243	136	119	14	106	61	150
Jung-gu (중구; 中區)	37.563656	126.997510	174	93	199	145	120	14	106	59	162
Seocho-gu (서초구; 瑞草區)	37.483500	127.032200	47	84	177	128	104	13	66	104	133

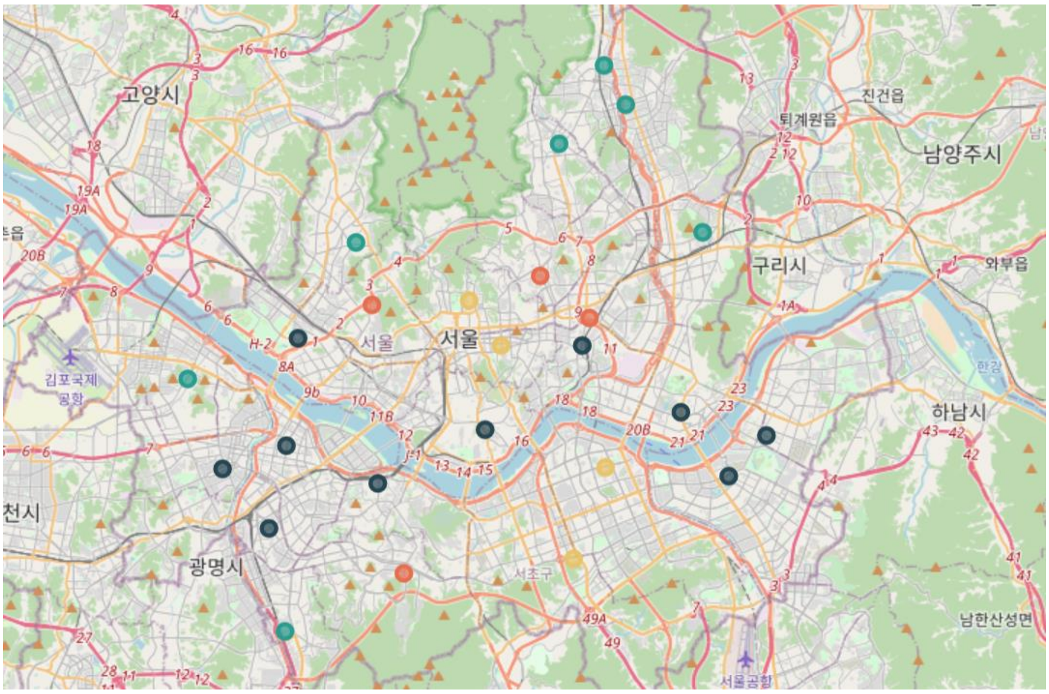


Fig. 6 Seoul Neighbourhood Clusters

7. Discussion

From above segment comparison I found that Jakarta and Seoul have many similarities. There are same number of clusters with almost the same characteristics. City centers are the most active one, has more venues that other area. There are also neighbour with less venue, usually in the outer parts of the city.

Although many similarities there are dissimilarities, which is Jakarta has factories area which located outside of the city and generally has less venue. While Seoul has cluster that has more education venues. Although is I examine closer those education cluster also industrial area according to Wikipedia. This might indicate that industrial area of Seoul is more well managed and provide more education facility to the workers.

I can also see that Seoul in general has more venue than Jakarta. It might the result of more foursquare active users is Seoul or indeed Seoul is more advance than Jakarta.

This simple study is very limited to data provided by Foursquare, which is a crowdsource data. The quality of the data is heavily depended on the contributors of the data, be it the one who entry the venue or the one that edited or corrected the data. To increase the quality of the data I have to acquire the data from more credible source, i.e. from local administration or government body.

The other aspect that can be improved is adding more venue categories or break down the category into more granular category. For example, education can be broken down into college, high school, elementary school and so on. Offices can be dissected into more specific category, etc.

8. Conclusion

In this study I tried to find similarities and dissimilarities between Jakarta and Tokyo using data from Foursquare. By utilising unsupervised machine learning specifically Kmeans, Jakarta and Seoul neighbourhood can be clustered and segmented into 4 clusters which has distinctive characteristics.

While the two sister-city has same number of clusters which has many similarities, they also have dissimilarities such as industrial cluster that only found in Jakarta and education cluster that only found in Seoul. In general Seoul also has more venues than Jakarta.

The finding in this study can be used by Jakarta administration as suggestion what can be learned from Seoul to improve Jakarta public facilities and can lead to further study considering some improvements such as more credible data source and more granular study.