Analysis of Districts of Bangkok

Rob Egrot

2021

Contents

1 Problem Definition and background

1.1 Problem Statement

The goal is to support businesses and other organizations in Bangkok by providing district profiles, organizing districts into groups, and analyzing the makeup of venues within each group. We aim to use this analysis to answer two related but distinct questions:

(a) Given a proposed business venture or community project, for example opening a certain kind of restaurant, or creating a new public space, in which kinds of districts are these venues common/uncommon? Answering this question will allow business owners and community organizers to both target their projects to areas where they are likely to be successful, and also to identify potentially underserved communities.

(b) Given a district, what kind of venues are common/uncommon? Answering this question will allow potential developers to better understand what opportunities are available in a specific area.

1.2 Background on Bangkok

Thailand is classified as an upper middle income country by the World Bank [1]. It is one of the world's leading exporters of rice [2], and agriculture accounts for 8.6% of GDP [3], and employs 31% of the Thai workforce [4]. As Thailand continues industrializing these percentages decrease, with an increasing share of GDP being taken by service and industry [5]. In particular, Thailand is major regional a center of production in the automotive and electronics industries, particularly for Japanese companies, and also had, before Covid-19, a thriving tourism industry [6].

An estimated 10 million people live in Bangkok, the capital and by far the most populous city in the country [7]. This is roughly 14% of the total population of 69.63 million. Despite rapid growth over the past 30 years, Thailand is one of the most unequal countries in the world in terms of wealth distribution [8]. This inequality is evident at the national level in the divide between urban and rural communities, with some malls in Bangkok consuming more electricity than whole provinces in poorer regions [9], and also within Bangkok itself, as aside from the luxurious malls just mentioned, an estimated 23.7% of the urban population in Thailand lived in slums in 2018 [10]. The majority of these urban slum dwellers live in Bangkok, often in the shadows of the high end malls and apartments frequented by the wealthier citizens.

Bangkok itself is divided into 50 districts, each of which is divided into between 2 and 12 subdistricts. Our analysis will be at the district level, as data for individual subdistricts is difficult to obtain. Bangkok is governed by the Bangkok Metropolitan Administration (BMA), and of course the national government.

2 Data

To address the problem described in the previous section we use data on the 50 districts of Bangkok. We collected data from multiple sources in 7 broad categories.:

1) Basic district data.

2) Population data.

3) Community data.

4) Data on schools.

5) Data on new businesses.

6) Location of rapid transit stations.

7) Venue data.

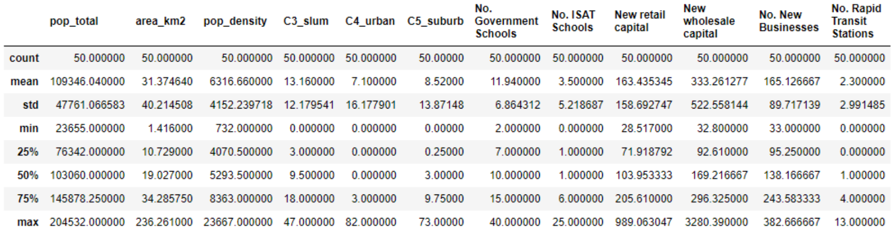
Basic statistical properties of the feature list, plus population and area information, are given in Figure 1.

The data was used in two main stages as follows:

(i) Data from categories 1) - 6) was used to build profiles of each of the 50 districts of Bangkok, and to arrange these districts into groups using unsupervised clustering techniques.

(ii) Data from category 7) (venue data) was used to build a venue profile for each of the clusters. These profiles were used to obtain insights relevant to questions (a) and (b) above.

The idea is that categories 1) - 6) provide a broad socioeconomic portrait of the districts, from which it should be possible to distinguish meaningful groupings. Details of data collection and sources can be found in Section 7. Following data collection and preparation (see Section 8 for details), we obtained two tables. The first contained (normalized and scaled) socioeconomic data, and the second (normalized and scaled) venue data. The correlations between features in these tables can be found in the form of heatmaps in Figures 2 and 3 respectively.



	pop_total	area_km2	pop_density	C3_slum	C4_urban	C5_suburb	No. Government Schools	No. ISAT Schools	New retail capital	New wholesale capital	No. New Businesses	No. Rapid Transit Stations
count	50.000000	50.000000	50.000000	50.000000	50.000000	50.00000	50.000000	50.000000	50.000000	50.000000	50.000000	50.000000
mean	109346.040000	31.374640	6316.660000	13.160000	7.100000	8.52000	11.940000	3.500000	163.435345	333.261277	165.126667	2.300000
std	47761.066583	40.214508	4152.239718	12.179541	16.177901	13.87148	6.864312	5.218687	158.692747	522.558144	89.717139	2.991485
min	23655.000000	1.416000	732.000000	0.000000	0.000000	0.00000	2.000000	0.000000	28.517000	32.800000	33.000000	0.000000
25%	76342.000000	10.729000	4070.500000	3.000000	0.000000	0.25000	7.000000	1.000000	71.918792	92.610000	95.250000	0.000000
50%	103060.000000	19.027000	5293.500000	9.500000	0.000000	3.00000	10.000000	1.000000	103.953333	169.216667	138.166667	1.000000
75%	145878.250000	34.285750	8363.000000	18.000000	3.000000	9.75000	15.000000	6.000000	205.610000	296.325000	243.583333	4.000000
max	204532.000000	236.261000	23667.000000	47.000000	82.000000	73.00000	40.000000	25.000000	989.063047	3280.390000	382.666667	13.000000

Figure 1: Basic statistical properties of district properties

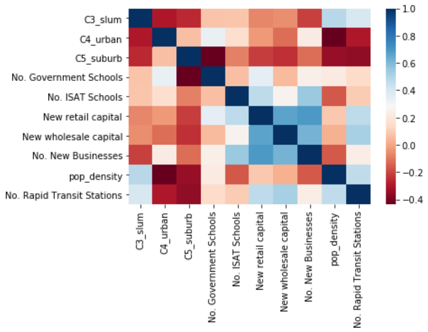


Figure 2: Socioeconomic data correlation heatmap

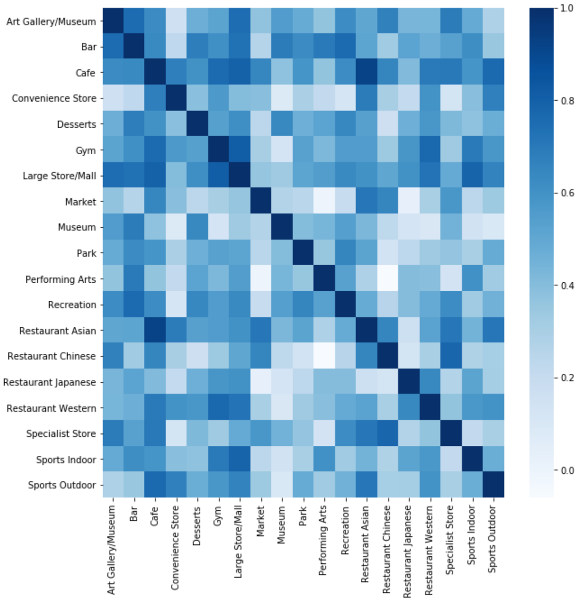
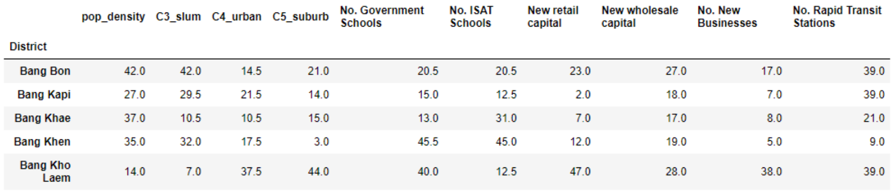


Figure 3: Venue correlation heatmap

3 Methodology

We used three clustering techniques to group the 50 districts of Bangkok into clusters, the techniques used being DBSCAN, Agglomerative Clustering and K-Medoids. Investigation revealed Euclidean distance to not be very robust with our data, with the average fraction of distance to furthest point divided by distance to nearest point being low. This relates to known issues with nearest neighbour style algorithms in high dimensional data. To mitigate this problem we used either the Manhattan or Canberra metrics when implementing our chosen clustering algorithms.

For DBSCAN and Agglomerative Clustering we worked with the normalized, scaled socioeconomic data exactly as described above, but for K-Medoids we instead obtained new data by *ranking* each district in each category. This ranking was based on the original socioeconomic data, and so assigned to each district a score between 1 and 50 for each feature, with 1 representing having the highest score for that feature (see Figure 4). Ties were handled using the averaging method.



District	pop_density	C3_slum	C4_urban	C5_suburb	No. Government Schools	No. ISAT Schools	New retail capital	New wholesale capital	No. New Businesses	No. Rapid Transit Stations
Bang Bon	42.0	42.0	14.5	21.0	20.5	20.5	23.0	27.0	17.0	39.0
Bang Kapi	27.0	29.5	21.5	14.0	15.0	12.5	2.0	18.0	7.0	39.0
Bang Khae	37.0	10.5	10.5	15.0	13.0	31.0	7.0	17.0	8.0	21.0
Bang Khen	35.0	32.0	17.5	3.0	45.5	45.0	12.0	19.0	5.0	9.0
Bang Kho Laem	14.0	7.0	37.5	44.0	40.0	12.5	47.0	28.0	38.0	39.0

Figure 4: Ranked districts (lower rank means higher score)

Prior to clustering we used PCA to significantly reduce dimensionality. After some experimentation, we decided to reduce down to only two dimensions. This only explained between 55% and 62% of the variance, but made for more robust clusters. We used the Canberra metric for DBSCAN and Agglomerative Clustering, and the Manhattan metric for K-Medoids. The reason for using the Canberra metric was that it scales down large differences between large values, allowing for looser clusters further from the origin. This was not necessary for the rank based K-Medoids clustering, so we just used Manhattan. DBSCAN found 3 clusters plus noise, so we used 3 as our target number of clusters in Agglomerative Clustering and K-Medoids. The sizes of the clusters produced by each method are given in Figure 5

	DBSCAN	Agglomerative	K-Medoids
0	24	25	21
1	8	19	14
2	9	6	15

Figure 5: The sizes of clusters produced by each method. DBSCAN also classified 9 districts as noise.

We used Rand and adjusted Rand scores to test the similarity between the three clusterings produced (see Figure 13 later), and analyzed the socioeconomic profile of each cluster for each clustering using box plots. We used this to identify 3 reasonably distinct clusters of district within Bangkok (see the next section). Finally, we studied the distributions of venues in each cluster, again using box plots.

4 Results

4.1 Clustering

After using PCA to reduce down to 2 dimensions, our clustering algorithms produced the scatter plots in Figures 6, 7 and 8. Remember that we are not using the Euclidean metric, so we cannot apply our usual intuitions about space and distance to these groupings. The corresponding geographical clusterings can be seen in Figures 9, 10 and 11. Figure 12 provides the dendrogram associated with our agglomerative clustering.

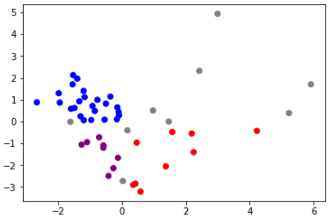


Figure 6: Scatter plot of DBSCAN clustering. Noise is in grey.

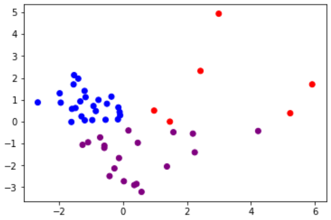


Figure 7: Scatter plot of Agglomerative Clustering.

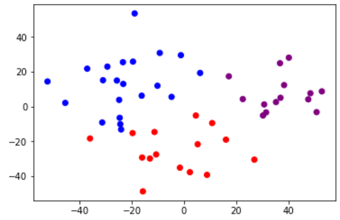


Figure 8: Scatter plot of K-Medoids clustering. The shape is different as this is based on rank data.

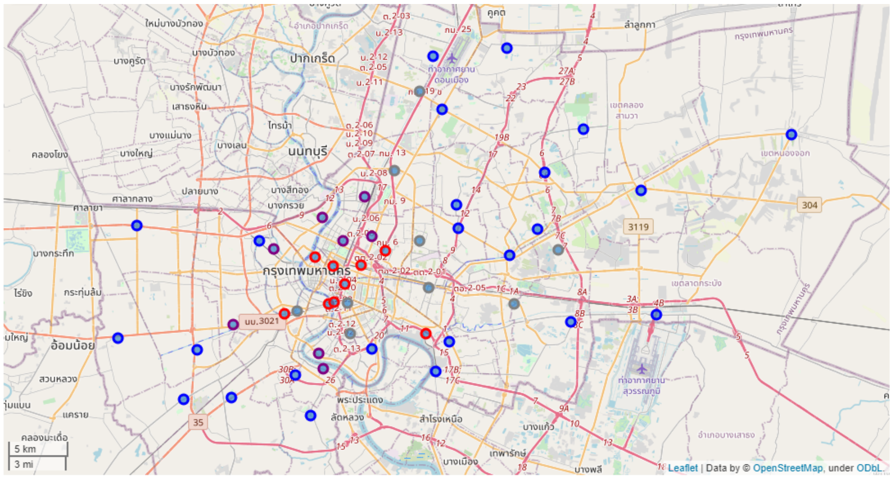


Figure 9: DBSCAN clustering.

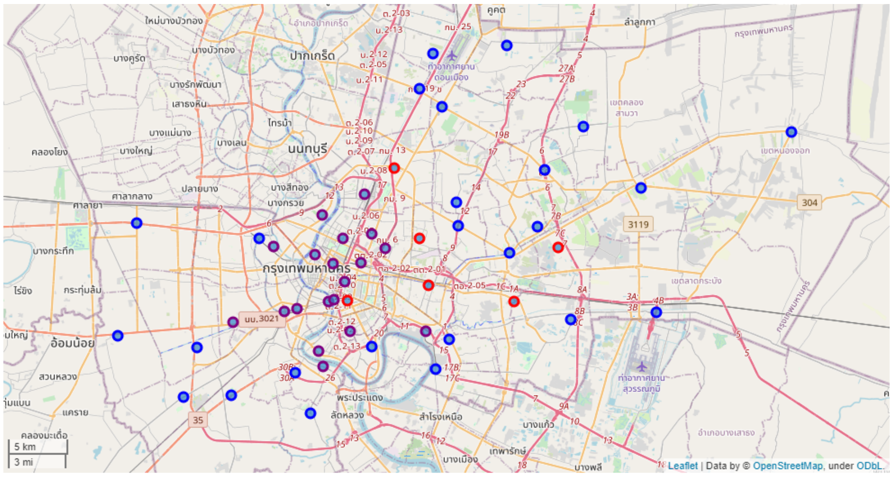


Figure 10: Agglomerative Clustering.

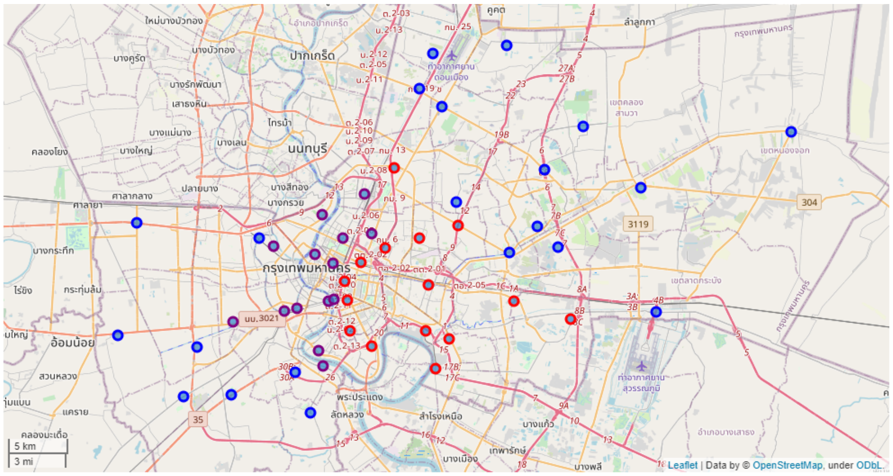


Figure 11: K-Medoids clustering.

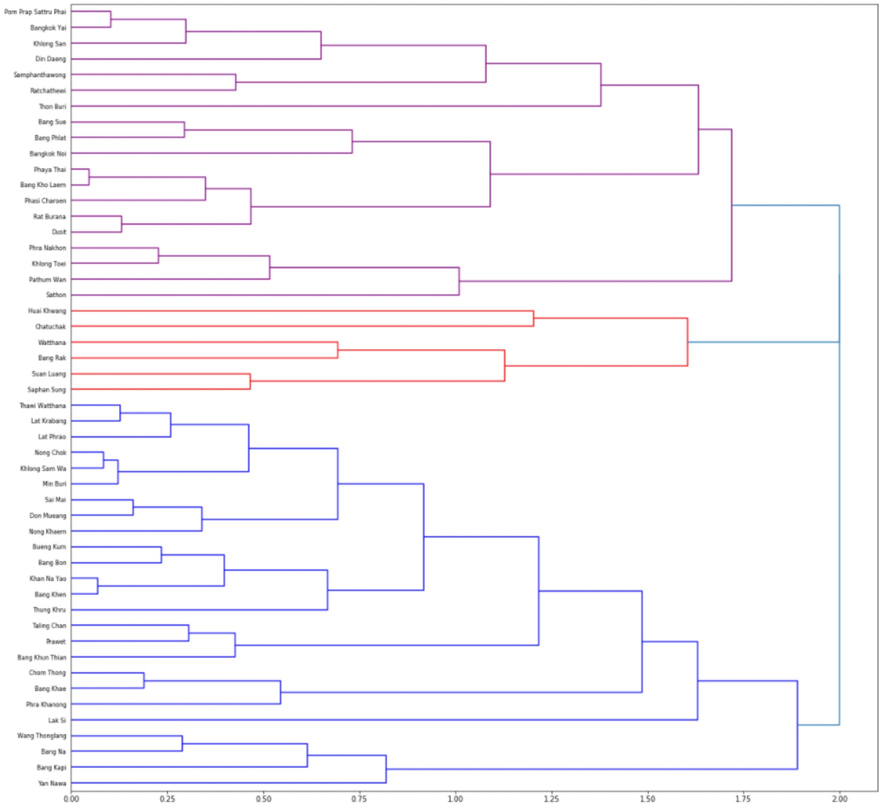


Figure 12: Agglomerative Clustering dendrogram.

From the geographical maps it is obvious that our 3 clusterings have quite a lot in common, though the agreements are far from perfect. This is born out by calculating adjusted Rand scores to measure similarity. The results of this are presented in Figure 13. Note that I don't know exactly what probability model sklearn uses for the calculation of the expected Rand score from a random clustering, so I'm not 100% sure that the calculated adjusted Rand scores here are exactly right.

	DBSCAN	Agglomerative	K-Medoids
DBSCAN	1	0.82	0.55
Agglomerative	0.82	1	0.53
K-Medoids	0.55	0.53	1

Figure 13: Adjusted Rand scores for each pair of clusterings. Comparison between the results of DBSCAN and the others is done using only non-noise districts, so is significantly higher than it would be if noise were included.

See also the confusion matrix for Agglomerative Clustering and K-Medoids in Figure 14. Note that there is never any confusion between clusters 0 and 1.

In other words, if one of these two methods assigns a district to Cluster 0, the other never assigns it to Cluster 1 (or vice versa). So the difficult districts seem to be those borderline between 2 and 0, and between 2 and 1.

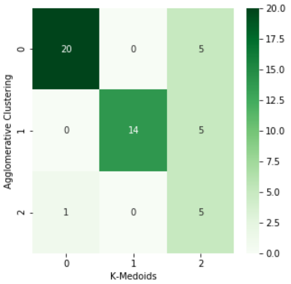


Figure 14: Confusion matrix for Agglomerative Clustering and K-Medoids.

Figures 16, 17 and 18 provide the distributions of features for each cluster in each clustering. From these distributions we extract the categories in Figure 15. Note that DBSCAN and Agglomerative clustering do not necessarily produce 'spherical' clusters (this is one of their selling points), so we should not expect to able to characterize the clusters they produce in terms of high or low values for particular features. Nevertheless, in our case the clusters are reasonably spherical, so we can extract the characterizations described.

- Cluster 0 (relatively suburban): Members of this cluster seem to be characterized by a low number of rapid transit links, and generally average or a little below average business investment. In addition there is below average population density. This trend is quite strong for all 3 clustering techniques.

- Cluster 1 (urban relatively poor): This cluster is characterized by high population density, and a high proportion of the population living in slums as opposed to urban or suburban community types. The distinction between Clusters 1 and 2 is weakest for DBSCAN, but becomes quite strong for agglomerative Clustering and K-Medoids.

- Cluster 2 (urban relatively affluent): This cluster is similar to Cluster 1, but is distinguished by relatively high capital investment, and better rapid transport links.

- Noise (DBSCAN only): As you might expect, for all features the 'Noise' category displays a high degree of variance, either in terms of wide box plots or with the presence of fairly extreme outliers. This 'cluster' seems to contain those districts which have an unusually high value for some feature. For example, unusually high investment, or a large number of ISAT schools.

Figure 15: Analyzing the clusters

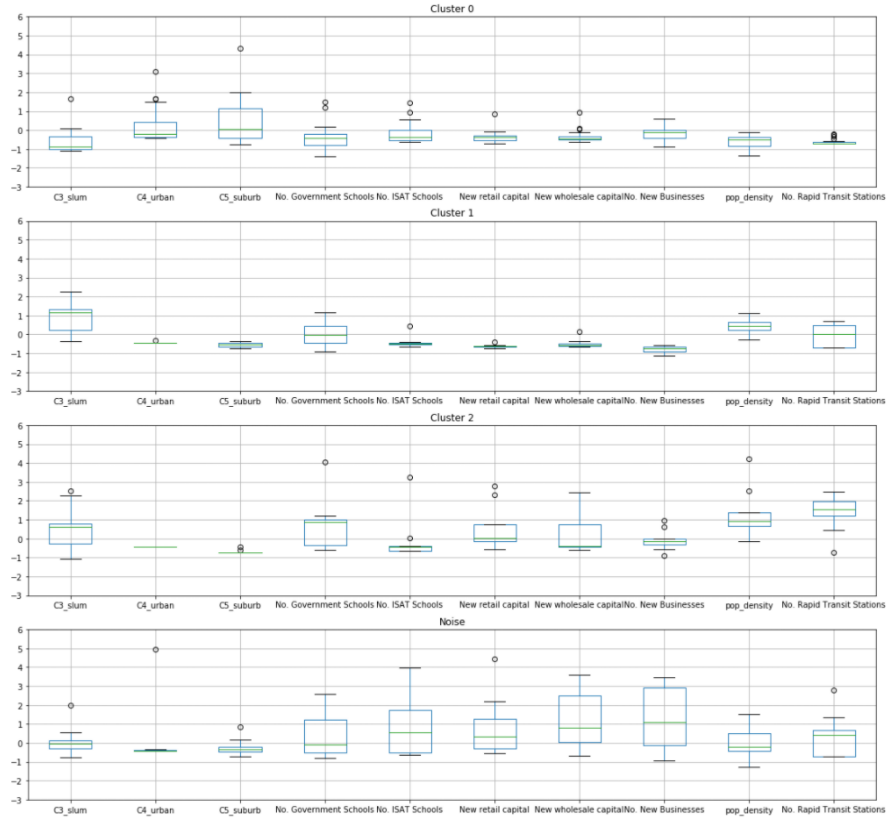


Figure 16: Distributions of features for each cluster obtained by DBSCAN.

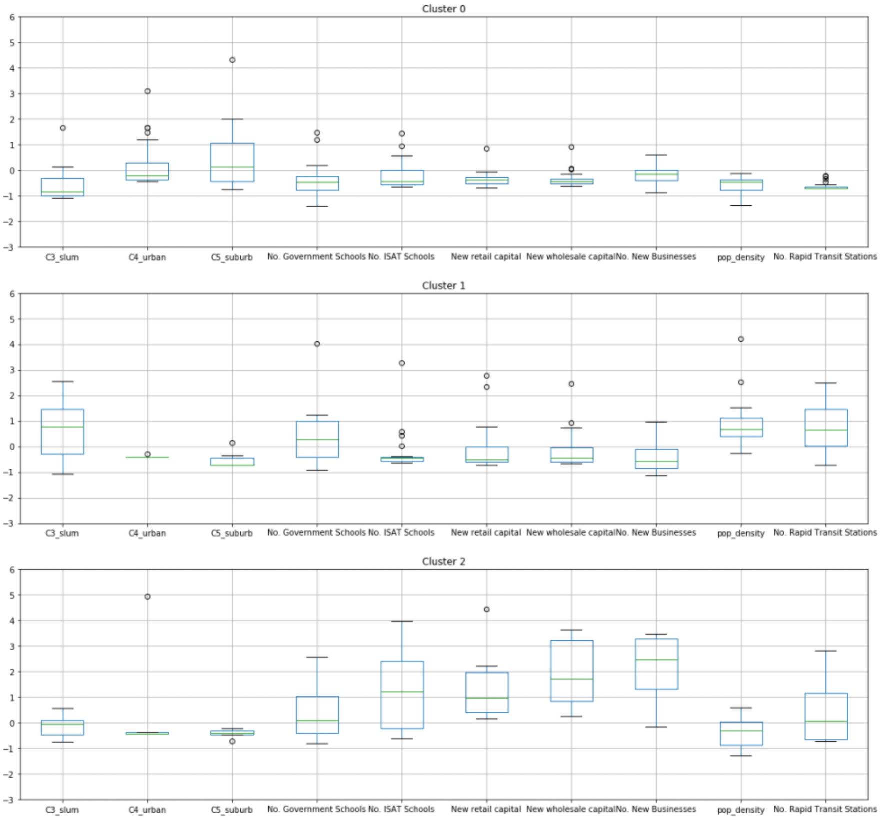


Figure 17: Distributions of features for each cluster obtained by Agglomerative Clustering.

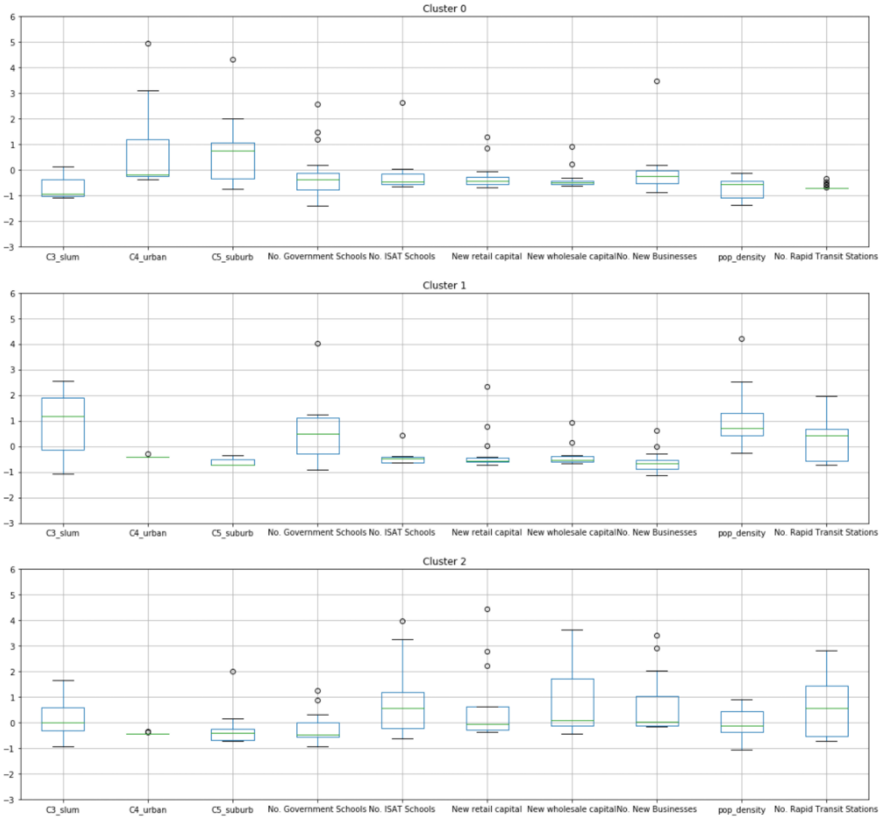


Figure 18: Distributions of features for each cluster obtained by K-Medoids.

4.2 Venue distributions

We investigate venue distribution for the clusters obtained using K-Medoids, as with this clustering every district is assigned a cluster, and each cluster has at least 14 districts in it (see Figure 5). In addition, distinctions between clusters as defined in Figure 15 are reasonably clear with this clustering (see Figure 18). The distributions of venue types for these clusterings are given in Figure 19. We can see some clear distinction between our clusters. Cluster 0 has less (per person) of basically everything, which is consistent with Cluster 0 being composed of more peripheral regions. Cluster 1 and Cluster 2 are similar, but Cluster 2 generally has more venues in most categories, and has significantly more performing arts venues, Western and Japanese Restaurants and indoor sports venues. This supports the characterization of Cluster 1 as being relatively poorer, and Cluster 2 as being relatively richer. It is notable however that all clusters have significant outliers in all categories, and Cluster 1 in particular has some of the most extreme.

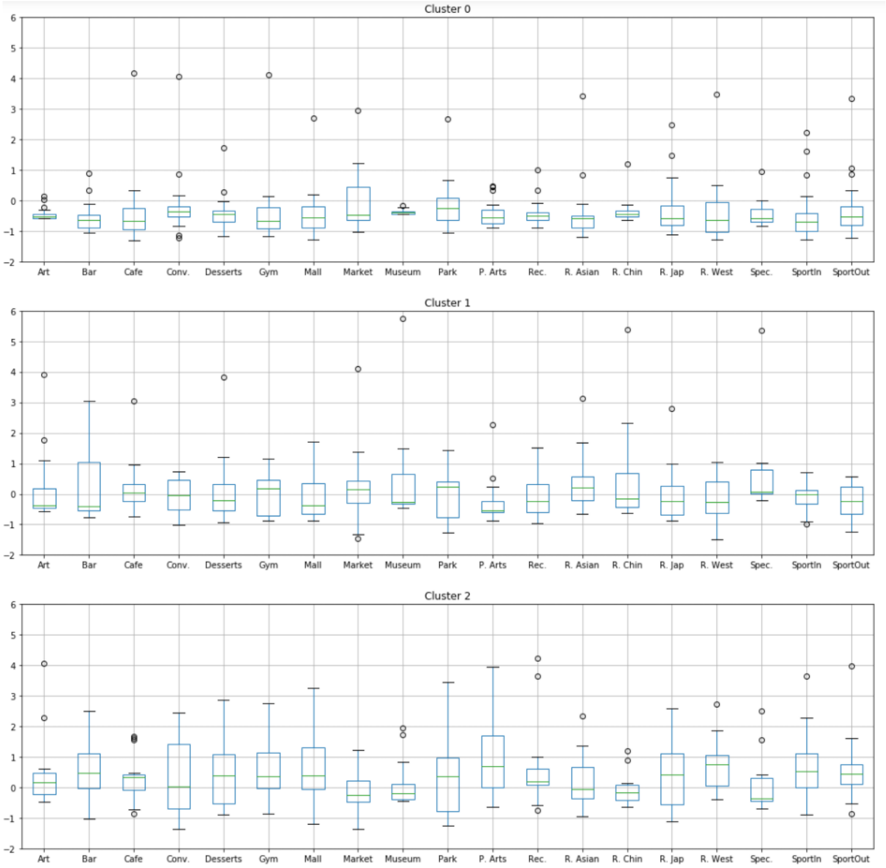


Figure 19: Distributions of venues for K-Medoids clusterings.

5 Discussion

Clustering algorithms often perform poorly with high dimensional data, and extreme dimensionality reduction comes with an associated loss of information, so it is perhaps surprising that we are able to extract 3 reasonably well defined and intuitive clusters from our mix of socioeconomic indicators. However, the boundaries between our clusters are not very sharp, and each cluster contains districts with properties that could potentially place it in another category according to our intuitive definitions from Figure 15. It is also striking how well defined the clusters appear on the geographical map (see Figure 11 for example).

It is interesting that these rather loose categories obtained from socioeconomic features seem to be identifiable from Foursquare venue data. We should be cautious, however, as given that we are looking at 19 venue types, it is expected that we would see some distinction just through random variation. The patterns we see in Figure 19 make intuitive sense. I.e. we would expect peripheral regions to have fewer venues in most categories than central regions, and we would expect relatively wealthy districts to contain more ‘upscale’ venues such as Western and Japanese restaurants, and possibly performing arts venues too (which is what we do in fact appear to see).

Nevertheless, the risk when looking for patterns in data as we do here is that it is very likely that we will find them, whether or not they reflect underlying patterns in nature. Our observations here should serve as a starting point for further investigation.

6 Conclusions and further work

As discussed in the previous section, while potentially interesting, our observations here should serve as a prompt for further investigation. Some possible directions are listed now.

- The data from the Thai government is often confusing and incomplete. More work obtaining and understanding socioeconomic data could provide a better selection of starting features for cluster analysis.
- The district level is possibly too broad. A better investigation might use subdistrict data, though this is more difficult as subdistrict level data is harder to find.
- More in depth statistical analysis on the observations of venue distributions for identified clusters should be performed. In particular, it would be good to provide some support beyond appeals to intuition for the idea that the observed distinctions are not purely down to chance.
- Some feature to venue analysis. E.g. how do socioeconomic features relate to numbers of specific venues? Can we predict cluster membership from venue data? These could be investigated by, for example, regression and classification algorithms respectively.

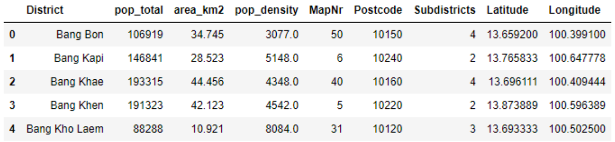
7 Appendix A: Details of data collection

The details of data collection are as follows:

1) **Basic district data:** This was scraped from Wikipedia [11]. This contains district names in both Thai and Romanized forms, along with geographical coordinates, post codes, population, no. of subdistricts, and map numbers marking locations on an associated map (these map numbers correspond to district codes in official documents, as we shall see later). We do not use the Thai district names, and we replace the population figures with more up to date ones obtained from official data.

2) **Population data:** The Bangkok Metropolitan Administration (BMA) makes various datasets available at [12]. We use the data at [13] to obtain more

recent district population counts, along with values for the areas and population densities of the districts. Some district areas are missing, so we added values obtained from Wikipedia where necessary. Unfortunately the BMA do not provide Romanized names, so these were added separately. There are several different ways to Romanize Thai words, but it is convenient to use the same method as used by Wikipedia everywhere so we can easily merge dataframes using district names. The BMA data also includes subdistricts, but we dropped these rows from the data table as we do not need this information.



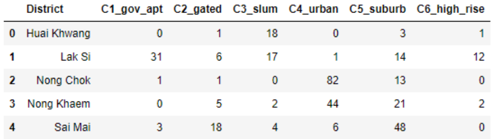
	District	pop_total	area_km2	pop_density	MapNr	Postcode	Subdistricts	Latitude	Longitude
0	Bang Bon	106919	34.745	3077.0	50	10150	4	13.659200	100.399100
1	Bang Kapi	146841	28.523	5148.0	6	10240	2	13.765833	100.647778
2	Bang Khae	193315	44.456	4348.0	40	10160	4	13.696111	100.409444
3	Bang Khen	191323	42.123	4542.0	5	10220	2	13.873889	100.596389
4	Bang Kho Laem	88288	10.921	8084.0	31	10120	3	13.693333	100.502500

Figure 20: Basic population and district data

3) **Community data:** The BMA publishes data on types of community in each district here: [14]. This counts for each district the number of communities in 6 categories. Unfortunately, it is not clear from the column headings provided what the categories are. Moreover, the provided headings seem to be wrong (they look like attempts to Romanize Thai, but e.g. 'salam' should probably be 'slum', and should be associated with community type 3, not 6, and so on). Correct headings can be worked out by comparing with the data in [15, 2.6] and doing a little translating. A full report on communities in Bangkok is available as [16], but unfortunately this is from a different year so the figures don't match exactly, and there is no English version available. The correct community types are as follows:

- $C1$ = apartments for government officials and police/military.
- $C2$ = gated community.
- $C3$ = slum.
- $C4$ = dense urban housing (not slum).
- $C5$ = suburbs.
- $C6$ = high rise apartments.

Another complication is that district names are in Thai, but they appear in the same order as those for population data above, so providing Romanizations is straightforward as we can reuse the old method.



	District	C1_gov_apt	C2_gated	C3_slum	C4_urban	C5_suburb	C6_high_rise
0	Huai Khwang	0	1	18	0	3	1
1	Lak Si	31	6	17	1	14	12
2	Nong Chok	1	1	0	82	13	0
3	Nong Khaem	0	5	2	44	21	2
4	Sai Mai	3	18	4	6	48	0

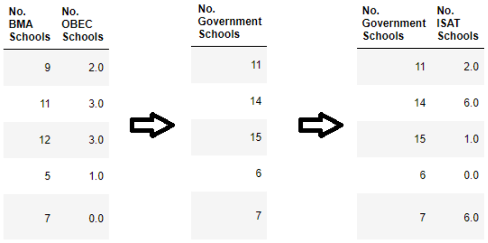
Figure 21: Community data

4) **Data on schools:** We make a simple count of the number of schools in each district. This is complicated by the fact that in Bangkok there are three different governing bodies for schools. Most schools are registered with the BMA itself, but there are also schools registered with the national Office of the Basic Education Commission (OBEC), and also 'international' schools registered with the International Schools Association of Thailand (ISAT). Data on BMA schools was obtained directly from [17], while data on OBEC schools was obtained directly from [18]. For the purpose of this analysis we do not distinguish between e.g. high school and elementary school.

   For the BMA schools, the 'dcodes' field corresponds to map numbers obtained during step 1), so it was straightforward synthesizing this data with that already obtained. The list of OBEC schools did not provide this, so districts had to be obtained manually. BMA and OBEC schools were combined into a single category 'government schools', as we are not aware of a relevant difference between the two categories.

   Data on ISAT schools was scraped from [19], and synthesized with previous data using postcodes. When these could not be scraped properly or were incorrect they were added manually. There is a small problem here as sometimes the same postcode is used for more than one district. We chose to count ISAT schools with postcodes used for multiple districts as occurring in *all* districts with that postcode, so several ISAT schools get double counted. This is not ideal, but it seems unlikely to cause big problems.

5) **Data on new businesses:** The BMA makes data on new business registrations available. These come in two categories, retail [15, 4.11], and wholesale [15, 4.12]. In each category there is data for the number of new business registrations and total new capital for the years 2017-2019. We record the averages for each district.



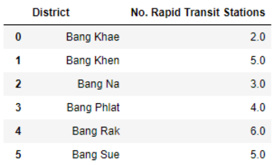
No. BMA Schools	No. OBEC Schools
9	2.0
11	3.0
12	3.0
5	1.0
7	0.0

No. Government Schools
11
14
15
6
7

No. Government Schools	No. ISAT Schools
11	2.0
14	6.0
15	1.0
6	0.0
7	6.0

Figure 22: Adding school counts

6) **Location of rapid transit stations:** There are 3 rapid transit rail systems operating in Bangkok. These are the BTS, the MRT and the Airport Rail Link. We scraped data from the links on the Wikipedia page [20] to count the number of rapid transit stations in each district. Some stations are along the border of two districts. In these cases we counted them for both districts.



	District	No. Rapid Transit Stations
0	Bang Khae	2.0
1	Bang Khen	5.0
2	Bang Na	3.0
3	Bang Phlat	4.0
4	Bang Rak	6.0
5	Bang Sue	5.0

Figure 23: Counting rapid transit stations

7) **Venue data:** We obtained data on venues in each district using the Foursquare API. We calculated a search radius for each district such that the corresponding circle has the same area as that district. Then for each district center we found the top 50 (this is the maximum allowed by the API) venues within the corresponding search radius in each of the categories 'food', 'drinks', 'coffee', 'shops', 'arts', 'outdoors', 'sights'. We drop duplicates created by some venues appearing in more than one category. Some venues are still duplicated as they get counted in two different districts due to overlapping circles (since the districts are not actually circular in reality). Some venues may be missed due to being outside the scope of all the approximating circles, and also possibly due to the limit on search results. This is not ideal, but again it should not cause big problems.

8 Appendix B: Data preparation

8.1 Socioeconomic data

Preliminary investigation showed a high degree of correlation between number of new retail businesses and number of new wholesale businesses, so these were combined into a single feature 'No. New Businesses'.

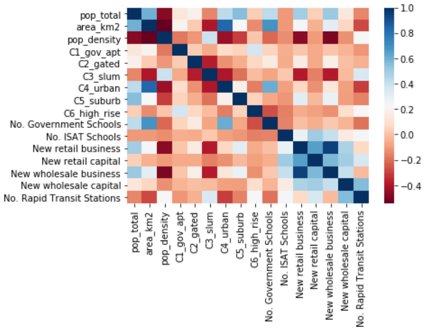
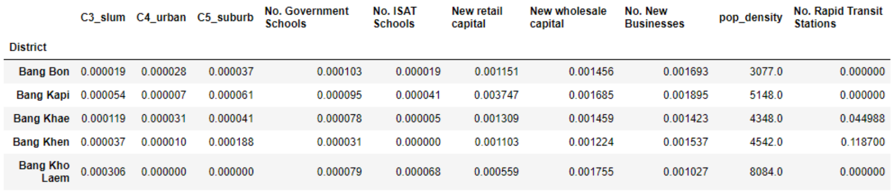


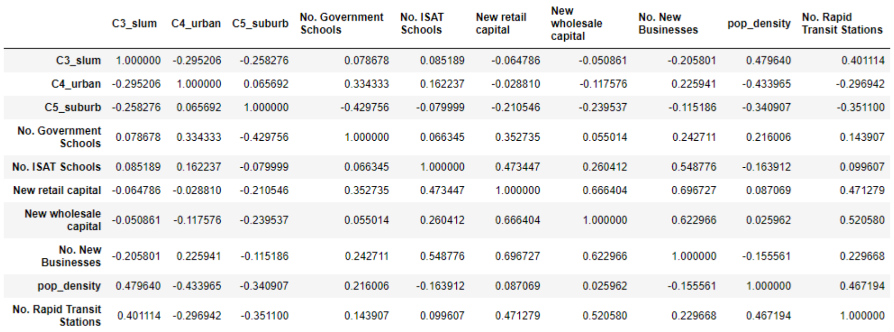
Figure 24: Correlations between original features

In addition, community types C1, C2 and C6 did not seem to add much, and are also a little unclear, so we dropped those features. Finally, we normalized all features not relating explicitly to population or area with respect to population or area. Specifically, 'No. Rapid Transit Stations' was normalized using ‘area_km2’ for each district, and the rest using ‘pop_total’, again for each district. Having normalized in this way, we dropped the ‘area_km2’ and ‘pop_total’ features.



	C3_slum	C4_urban	C5_suburb	No. Government Schools	No. ISAT Schools	New retail capital	New wholesale capital	No. New Businesses	pop_density	No. Rapid Transit Stations
District										
Bang Bon	0.000019	0.000028	0.000037	0.000103	0.000019	0.001151	0.001456	0.001693	3077.0	0.000000
Bang Kapi	0.000054	0.000007	0.000061	0.000095	0.000041	0.003747	0.001685	0.001895	5148.0	0.000000
Bang Khae	0.000119	0.000031	0.000041	0.000078	0.000005	0.001309	0.001459	0.001423	4348.0	0.044988
Bang Khen	0.000037	0.000010	0.000188	0.000031	0.000000	0.001103	0.001224	0.001537	4542.0	0.118700
Bang Kho Laem	0.000306	0.000000	0.000000	0.000079	0.000068	0.000559	0.001755	0.001027	8084.0	0.000000

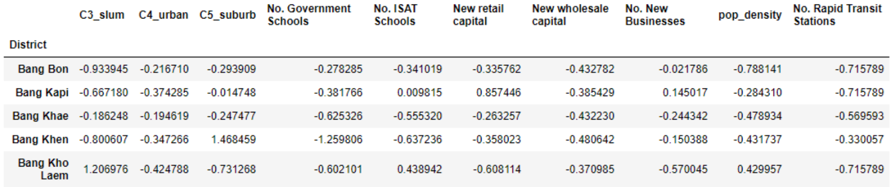
Figure 25: Final feature list (normalized) for clustering



	C3_slum	C4_urban	C5_suburb	No. Government Schools	No. ISAT Schools	New retail capital	New wholesale capital	No. New Businesses	pop_density	No. Rapid Transit Stations
C3_slum	1.000000	-0.295206	-0.258276	0.078678	0.085189	-0.064786	-0.050861	-0.205801	0.479640	0.401114
C4_urban	-0.295206	1.000000	0.065692	0.334333	0.162237	-0.028810	-0.117576	0.225941	-0.433965	-0.296942
C5_suburb	-0.258276	0.065692	1.000000	-0.429756	-0.079999	-0.210546	-0.239537	-0.115186	-0.340907	-0.351100
No. Government Schools	0.078678	0.334333	-0.429756	1.000000	0.066345	0.352735	0.055014	0.242711	0.216006	0.143907
No. ISAT Schools	0.085189	0.162237	-0.079999	0.066345	1.000000	0.473447	0.260412	0.548776	-0.163912	0.099607
New retail capital	-0.064786	-0.028810	-0.210546	0.352735	0.473447	1.000000	0.666404	0.696727	0.087069	0.471279
New wholesale capital	-0.050861	-0.117576	-0.239537	0.055014	0.260412	0.666404	1.000000	0.622966	0.025962	0.520580
No. New Businesses	-0.205801	0.225941	-0.115186	0.242711	0.548776	0.696727	0.622966	1.000000	-0.155561	0.229668
pop_density	0.479640	-0.433965	-0.340907	0.216006	-0.163912	0.087069	0.025962	-0.155561	1.000000	0.467194
No. Rapid Transit Stations	0.401114	-0.296942	-0.351100	0.143907	0.099607	0.471279	0.520580	0.229668	0.467194	1.000000

Figure 26: Correlations between final (normalized) features

Finally, features were scaled using StandardScaler.



	C3_slum	C4_urban	C5_suburb	No. Government Schools	No. ISAT Schools	New retail capital	New wholesale capital	No. New Businesses	pop_density	No. Rapid Transit Stations
District										
Bang Bon	-0.933945	-0.216710	-0.293909	-0.278285	-0.341019	-0.335762	-0.432782	-0.021786	-0.788141	-0.715789
Bang Kapi	-0.667180	-0.374285	-0.014748	-0.381766	0.009815	0.857446	-0.385429	0.145017	-0.284310	-0.715789
Bang Khae	-0.186248	-0.194619	-0.247477	-0.625326	-0.555320	-0.263257	-0.432230	-0.244342	-0.478934	-0.569593
Bang Khen	-0.800607	-0.347266	1.468459	-1.259806	-0.637236	-0.358023	-0.480642	-0.150388	-0.431737	-0.330057
Bang Kho Laem	1.206976	-0.424788	-0.731268	-0.602101	0.438942	-0.608114	-0.370985	-0.570045	0.429957	-0.715789

Figure 27: Normalized and scaled features

8.2 Venue data

The initial trawl of Foursquare data produced 11403 venues, after dropping duplicated from the table. These came in 323 unique categories. Examination of the category list revealed a large amount of redundancy, with several categories being unclear and or needing merging with others. New venue categories were formed by manually creating a dictionary mapping old categories to newer, more clear and inclusive ones. This resulted in 115 unique venue categories. To reduce noise we dropped all categories with a total venue count of less than 100. This resulted in a final tally of 19 venue categories. Counts in these categories were normalized by dividing by $\frac{\text{district population}}{100000}$ (the 100000 on the denominator is to produce more human-readable numbers). Correlations between the new venue categories are given below. Note that the category 'Restaurant Asian' is used for restaurants either identifying themselves as Thai or generically as Asian. The logic of this is that Asian restaurants that do not otherwise distinguish themselves are likely to be Thai, seeing as this is Bangkok.

From the correlation heatmap in Figure 3 we see that all venue types are positively correlated with each other (after normalizing with respect to population). This makes sense, as more populous areas are likely to contain more of everything.

References

[1] [Online]. Available: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

[2] [Online]. Available: https://www.statista.com/statistics/255947/top-rice-exporting-countries-worldwide-2011/

[3] [Online]. Available: https://data.worldbank.org/indicator/NV.AGR.TOTL.ZS?locations=TH&year_high_desc=true

[4] [Online]. Available: https://www.statista.com/statistics/332341/employment-by-economic-sector-in-thailand/

[5] [Online]. Available: https://www.statista.com/statistics/331893/share-of-economic-sectors-in-the-gdp-in-thailand/#:~:text=ShareofeconomicsectorsintheGDPinThailand2019&text=In2019,theshareof,sectorcontributedabout58.59percent

[6] [Online]. Available: https://aseanup.com/business-thailand/

[7] [Online]. Available: https://www.cia.gov/the-world-factbook/countries/thailand/#people-and-society

[8] [Online]. Available: https://worldpopulationreview.com/country-rankings/wealth-inequality-by-country

[9] [Online]. Available: https://coconuts.co/bangkok/news/bangkoks-luxury-malls-use-more-energy-some-provinces-infographic/

[10] [Online]. Available: https://data.worldbank.org/indicator/EN.POP.SLUM.UR.ZS?locations=TH

[11] [Online]. Available: https://en.wikipedia.org/wiki/List_of_districts_of_Bangkok

[12] [Online]. Available: http://data.bangkok.go.th/en/dataset/?page=1

[13] [Online]. Available: http://data.bangkok.go.th/dataset/birthdead/resource/4ea00caf-2e96-40f7-8c2f-b00c36d376b3

[14] [Online]. Available: http://data.bangkok.go.th/dataset/e57c4e2e-77c8-4994-b9d8-b2d3a67aad82/resource/a41b3558-a441-4bb2-bcde-d51cca91f4c9

[15] [Online]. Available: http://www.bangkok.go.th/pipd/page/sub/16647/-2562

[16] [Online]. Available: http://www.bangkok.go.th/upload/user/00000103/KorPorChor/FinalReport.pdf

[17] [Online]. Available: http://data.bangkok.go.th/en/dataset/bmaschool/resource/3bd7d9f2-61cf-4484-b78c-9573df17fceb

[18] [Online]. Available: http://data.bangkok.go.th/en/dataset/becschools/resource/f0dc0de6-5934-48b5-a11d-8ae6b0b241a1

[19] [Online]. Available: https://www.isat.or.th/member-school

[20] [Online]. Available: https://en.wikipedia.org/wiki/List_of_rapid_transit_stations_in_Bangkok