Clustering London boroughs with k-means

1. Introduction

How to identify the individual's living conditions of a city whether are similar to another one? However, this tough question can be answered if we treat this as quantitative analysis. Since the 1970s, the Ministry of Housing, Communities and Local Government have calculated measures of deprivation in England which is represented with the English Indices of Deprivation. And the study I explored is based on this data set.

Kidokoro et al. (2020) divided the megacity regions by different economic statistics. They applied the cluster analysis to split Osaka city by various housing conditions and supported the suggestions on developing the living levels for government or local authorities. For example, the results show the inflow of people increasing in Airin area because of international tourists. So that the local offices can focus on urban development and tourism in Airin (Kidokoro, Hsiao and Fukuda, 2020). Another research is studying emission inequality by cluster analysis. According to the results in clustering city groups, Cheng et al. (2021) propose some policy implications. As for cluster 2 (Nanjing, Wuhan, Changsha and other cities) which is one of the three clusters, those cities should be controlled on emissions, as well as decreasing the number of heavy chemical companies (Cheng et al., 2021).

My study aims to propose policy implications for local organizations to identify places where should be prioritized resources and received funding effectively. In this paper, I will use clustering analysis to split the London boroughs into several groups which have closer living conditions. Also representing the results with different ways is necessary for this paper.

2. Data

The data employed for analysis comes from The English Indices of Deprivation 2019 (IoD2019) which is released by the Ministry of Housing, Communities and Local Government (MHCLG) in September in 2019. The data collects the living conditions of all 32,844 LSOAs (Lower-layer Super Output Area) on England. And the living conditions have been described by some aspects including Income score (Income, Education, Skills and Training Score (EST), Health Deprivation and Disability Score (HDD), Crime Score (Crime), Barriers to Housing and Services Score (BHS) and Living Environment Score (Environment). Indices of Deprivation measure deprivation on a relative value rather than absolute. Also, higher scores in some respects mean higher potential risk, for example, Crime Score measures the risk of personal victimization at region level (Ministry of Housing Communities and Local Government,

2019).

The contributors of this data announced these data sets are aimed at users and analysts who are interested in specific domains of deprivation, so this data is available in this study. And this paper focuses on boroughs in London so that I selected the regions in London. Data with the selected region is represented by Tables 1.

Table 1: Seven distinct domains of deprivation with selected variables

LADs code	Income	EST	HDD	Crime	BHS	Environment
E09000001	0.06	5.36	0.67	1.66	36.27	40.37
…	…	…	…	…	…	…
E09000033	0.13	8.32	0.84	0.12	23.20	41.12

London city and 32 boroughs in London are including on the table, as well as seven average scores of various aspects. After loading and cleaning the data, the primary work is to check the characters of observed samples. And I outputted the result with descriptive statistics by Python.

Table 2: Descriptive statistics

	Income	EST	HDD	Crime	BHS	Environment
count	33.00	33.00	33.00	33.00	33.00	33.00
mean	0.13	12.70	-0.40	0.20	31.42	29.77
std	0.04	5.17	0.47	0.43	7.41	7.28
min	0.06	3.49	-1.38	-1.66	17.42	16.55
25%	0.11	8.32	-0.67	0.08	26.74	24.68
50%	0.13	12.27	-0.40	0.28	30.95	29.41
75%	0.16	16.03	-0.03	0.47	36.07	35.06
max	0.20	25.84	0.39	0.72	49.44	44.06

3. Methodology

K-means clustering in this paper is worked by Python.

3.1 Normalization of data

Before we use the k-means clustering on data, normalization of data is important, because k-means clustering is sensitive to outliers, which can result in inaccurate clusters. On table 2, no outliers can be seen on different variables, so that the method of Min-Max normalization is applied to this data which has a small range between minimum and maximum.

3.2 K-means clustering

Cluster analyses aim to divide the individuals by their similarity, also the similar characteristics are using as criteria to group those samples. And K-means clustering is one of the most popular clustering analyses. Given a group of n data points like $(x_1, x_2,\ldots, x_n)$, and an inter k ($\leq n$), this clustering algorithms is to assign those points into k groups S= $\{S_1, S_2, \ldots, S_k\}$. A point is assigned to in a particular group if it is

closer to that group's centroid than any other centroid (Piech, 2020). And this criterion of assignment can be represented as to minimize the within-cluster sum of squares (WCSS) (Kanungo et al., 2002).

Formula 1: WCSS

$$arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

Note: Where μ_i is the mean of points in S_i

3.3 Identify the quality of clusters by elbow way and silhouette analysis

When using k-means clustering, it is necessary to identify the number of clusters.
One method is called the elbow method. The idea of this method is to calculate and store all sum of squared errors (SSE) by changing the number of k. Then the values of SSE by different K will plot as a line, and finally, we can get 'elbow' which is optimal k (Gove, 2017).

Calculating the silhouette score is straight forward to determine the k. This method measures how close each point in one particular cluster is to points in near clusters (Pedregosa et al., 2011). And this measure ranges from -1 to 1, the silhouette score is approaching to 1, which means the sample is well matched to its own cluster. While the negative value indicates this sample is not well cluster.

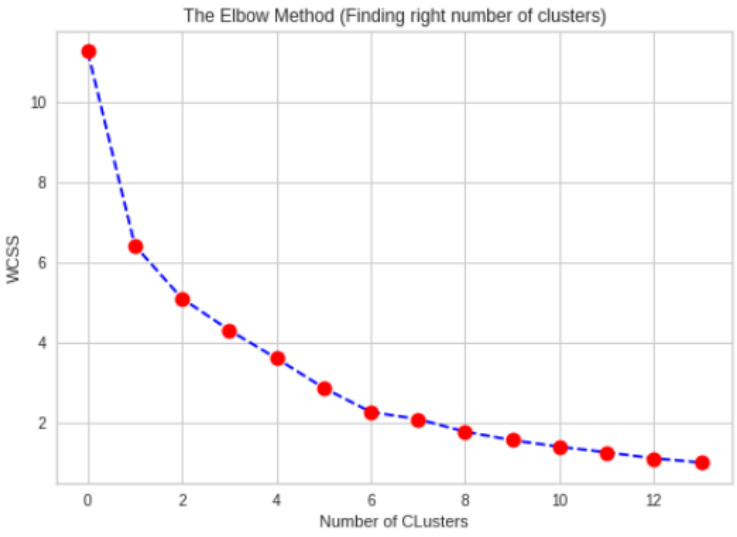
Formula 2: Silhouette score

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

4. Presentation of results

4.1 Identify the quantity of k

Figure 1: The elbow method



According to this line plot, the ‘elbow’ is hard to determine directly. Our goal is to pick a small value of k with a low value of SSE, however, it is difficult to choose the outcome with k between 2 and 3.

I used another method that is Silhouette analysis. I ran the codes by Python to calculate the silhouette score with different k by python.

Table 3: The silhouette score for k-clusters

k	The silhouette score
2	0.36
3	0.34
4	0.26
5	0.25
6	0.24

For k=2, the silhouette score is 0.36 which is closer to 1 than other scores, so that I choose 2 as the number of clusters.

4.2 Represent result of k-means clustering

Figure 2a: Cluster 0

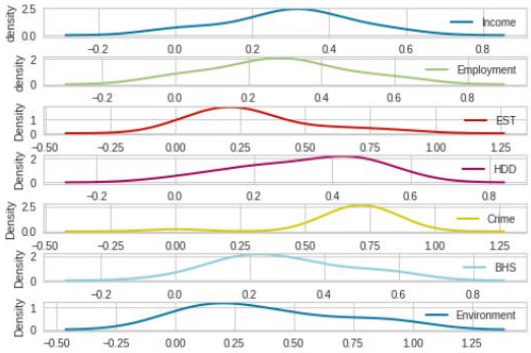
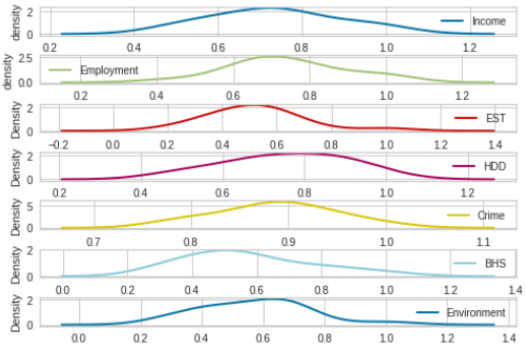


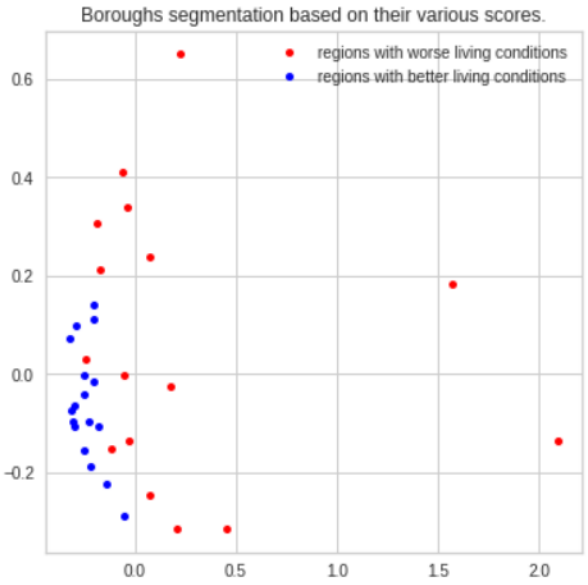
Figure 2b: Cluster 1



Those density plots describe the characters of two clusters, and cluster 1 has higher scores than cluster 0 in seven district domains. So, we can define the cluster 1 as more deprived regions, on the contrary, boroughs in cluster 0 have better living conditions.

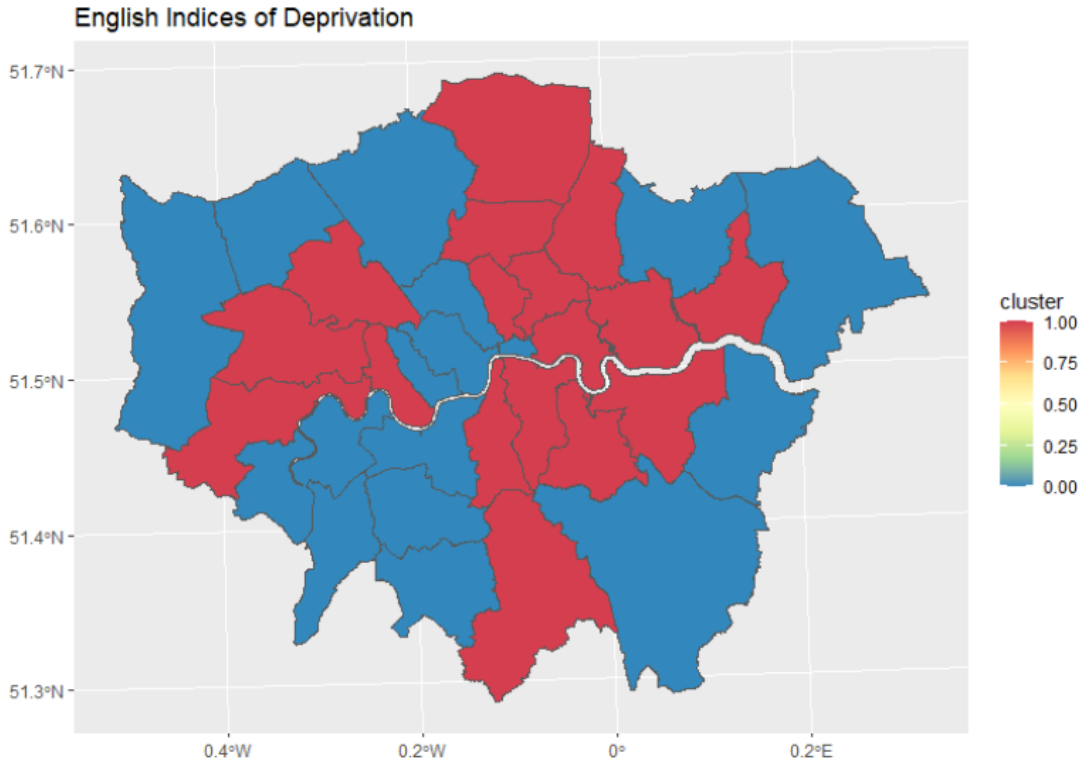
I used PCA to transform data to 2 dimensions for visualization of two clusters.

Figure 3: Use PCA to visualization



To achieve our goal to identify which specific boroughs should be prioritized resources and received funding effectively, I also applied those two clusters into a map.

Figure 4: London is separated by two clusters



5 Discussion of results

I have regrouped all 33 regions by k-means clustering, the following table shows more details about those two clusters.

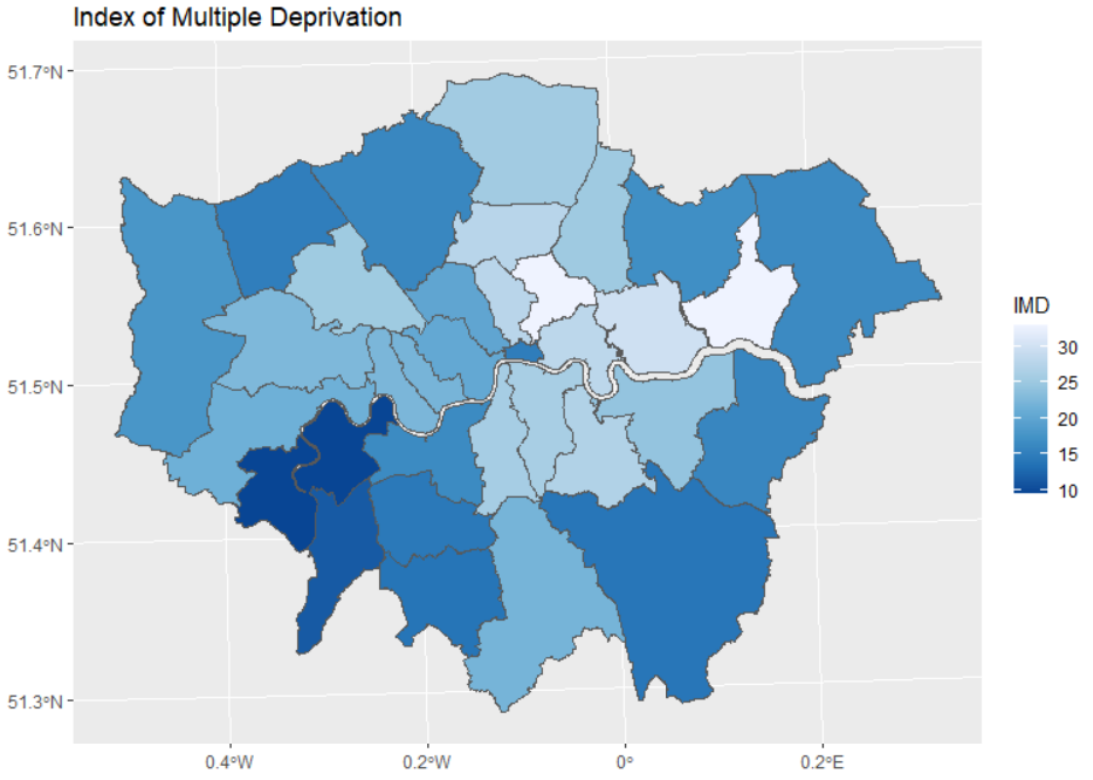
Table 4: The average scores of various indices by two clusters

cluster	N	Income	Employment	EST	HDD	Crime	BHS	Environment
0	16.00	0.29	0.28	0.29	0.34	0.67	0.29	0.38
1	17.00	0.74	0.74	0.52	0.75	0.89	0.58	0.57
total	33.00	0.52	0.52	0.41	0.55	0.78	0.44	0.48

As for two clusters, it can be seen clearly that the average scores of various indices in cluster 1 are much higher than those in cluster 0. In cluster 0, 16 regions are included, which are City of London, Barnet, Bexley, Bromley, Camden, Harrow, Havering, Hillingdon, Kensington & Chelsea, Kingston upon Thames, Merton, Redbridge, Richmond upon Thames, Sutton, Wandsworth and Westminster. On the other hand, 17 boroughs of Barking & Dagenham, Brent, Croydon, Ealing, Enfield, Greenwich, Hackney, Hammersmith & Fulham, Haringey, Hounslow, Islington, Lambeth, Lewisham, Newham, Southwark, Tower Hamlets and Waltham Forest are in cluster 1 which areas are more 'deprived'.

In addition, in the report of English Indices of Deprivation, a summary index has mentioned by MHCLG, which is Index of Multiple Deprivation (IMD) (Ministry of Housing Communities and Local Government, 2019). This index is comprised of seven mentioned distinct domains of deprivation with appropriate weights. And the most deprived area has the largest number of IMD. In Figure 5, the map of London was filled with Index of Multiple Deprivation by R.

Figure 5: Index of Multiple Deprivation in London boroughs



The ‘dark blues’ can be observed on the outer of London, which are less deprived than other ‘light blues’ boroughs. And ‘the red patterns’ of cluster 1 in Figure 4 is showed with ‘lighter blues’ in the map of IMD. So the output in the Index of Multiple Deprivation is consistent with the results of clustering analysis.

6. Conclusions

In this paper, I divided London into two clusters by k-means clustering analysis. The criteria of classification are to find the similarity on seven distinct domains of deprivation. The government should put more attention on the development of areas where are more ‘deprivation’. For instance, the boroughs in cluster 1 (i.e., Barking & Dagenham, Hackney, Newham and so on) have priority to receive the resource and funding. However, some limitations exist in this research. Due to the diversity of all seven variables, the result of clustering assigned boroughs only in two clusters. We can only see that the overall scores in cluster 1 are higher than cluster 2, but the similarity of some specific domains is unknown. If the clustering shows more groups, the local organizations will have more targeted to support the distinct boroughs with specific aspects. For example, the local authorizes should attach great importance to education on the boroughs where is belongs to ‘higher scores in Education’ cluster.

Student number:20051170
Word counts:1674

Reference:

Cheng, S., Fan, W., Zhang, J., Wang, N., Meng, F. and Liu, G. (2021). 'Multi -sectoral determinants of carbon emission inequality in Chinese clusterin g cities'. Energy. Elsevier Ltd, 214, p. 118944. doi: 10.1016/j.energy.202 0.118944.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R. an d Wu, A. Y. (2002). 'An efficient k-means clustering algorithms: Analy sis and implementation'. IEEE Transactions on Pattern Analysis and Ma chine Intelligence, 24 (7), pp. 881–892. doi: 10.1109/TPAMI.2002.10176 16.

Kidokoro, T., Hsiao, H. and Fukuda, R. (2020). 'Study on the polarization to megacity regions and the urban divide: Focusing on the case of Nishina ri Ward, Osaka City, Japan'. Japan Architectural Review, 4 (1). doi: 10. 1002/2475-8876.12189.

Ministry of Housing Communities and Local Government. (2019). 'National Sta tistics English indices of deprivation 2019'. Ministry of Housing, Comm unities & Local Government, 2019. Available at: https://www.gov.uk/gov ernment/statistics/english-indices-of-deprivation-2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A. Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). 'Scikit-learn: Machine Learning in Python'. JMLR 12, pp. 2825-2830, 2011.

Piech, C. (2020). K Means [online] Available at: https://stanford.edu/~cpiech/cs2 21/handouts/kmeans.html [Accessed 17 Jan. 2021].