The University of Sheffield

TOPIC:

Discover the association between deprivation of education and other indexes of deprivation in England with Clustering analysis, PCA, and Multi-linear Regression

Module Code	INF6027
Registration Number	230141036
Words count	3159

Abstract

Background

Education deprivation is a multifaceted problem encompassing insufficient infrastructure, economic barriers, and geographic isolation. Policy makers need to perceive level of deprivation in education to implement appropriate policies to ameliorate this discrepancy.

Aims

Discover the association between educational deprivation and factors such as financial indicators, crime, health, environment, and population.

Objectives

- Perform K-means clustering and PCA at the local authority level to evaluate the association among educational deprivation scores and other variables, including income, employment, health, crime, environment, and the population of dependent children.

- Predict educational deprivation scores using the multi-linear regression model and compare different variables selected by the correlation matrix and PCA.

Methods

Utilise the Spearman correlation coefficient to evaluate the relationship among educational deprivation scores and other variables. Group the data set by their local authority and conduct K-means clustering and PCA. Utilise multi-linear regression to predict educational deprivation scores based on variables selected by correlation coefficient and PCA.

Result

The Correlation coefficients among deprivation of education, income, employment, health, and crime are high at 0.81, 081, 0.71, and 0.52, respectively. Otherwise, factors such as environment, the population of dependent children and barriers to housing and services are less significant. In multi-linear regression, the model with variables selected by correlation matrix has better accuracy (0.70) than a model with PCA (0.68).

Conclusion

From correlation analysis, K-means clustering and PCA highlight significant associations between educational deprivation and factors such as income, employment, health, and crime. The multi-linear regression model indicates the significance between the accuracy and explainability of models while selecting independent variables.

Table of contents

1. Introduction and aim

1.1 Literuture review

Correlation among deprivation of education, financial factors, health care, and crime.

Education is a crucial component in shaping and influencing society in various ways. Its impact is profound and extends beyond individual development to the broader fabric of communities and nations. Several relevant reports are as follows:

In the United Kingdom, the father's educational level significantly influences the probability of low educational achievement, surpassing other factors. Individuals whose fathers have a low educational background are 7.5 times more prone to experiencing lower educational outcomes than those with highly educated fathers (Serafino & Tonkin, 2014). Moreover, individuals with lower educational attainment are nearly five times more likely to be in poverty compared to those with a higher level of education (Serafino & Tonkin, 2014).

In research conducted by OECD (2022), adults aged 25-34 with tertiary education, on average, had employment rates that were eight percentage points higher than their counterparts with merely upper secondary or post-secondary non-tertiary education in 2021. As Perera (2022) mentioned, by age 40, the typical worker in the United Kingdom with a degree earns two times the salary as much as someone with qualifications limited to a GCSE (General Certificate of Secondary Education) level or below. Namely, increasing accessibility to high-quality education can help one break the poverty cycle (Hart, 2019).

According to Zajacova and Lawrence (2018), in high-income countries such as the United States, it has been observed that adults with lower educational achievement endure insufficient health compared to other populations. Raghupathi and Raghupathi (2020) mentioned that this pattern is attributed to the massive health unfairness contributed by education attainment.

According to (Bradley & Green, 2020), increasing educational attainment on crime has several influences. Encouraging attending schools among crime-prone groups probably reduces the crime rate; education policies can also decrease property and violent crime. Research conducted by Anderson (2014) indicated an approximately 17 percent decrease in arrest rates related to property crime and violent crime among 16-18-year-old juveniles, which was achieved by impoving attendance at school.

The possible methodology for studying deprivation

In research conducted by Aungkulanon et al. (2017), they utilised principal component analysis and cluster analysis to discover the association between socioeconomic deprivation and mortality differentials. In their result, both methods effectively identified clusters based on socioeconomic characteristics. Moreover, to compare the index of multiple deprivation score constituent countries of the United Kingdom, Abel et al. (2016) utilised a linear regression model based on income and employment scores. This gives a simple way of obtaining comparable measures of area-based deprivation across the UK.

1.2 Overall aim and objectives

Aim

Discover the deprivation of education in England and the correlation among various factors such as financial indicators, crime, health care, living environment, and population.

Objectives

- Utilise correlation coefficient to evaluate the association among educational deprivation scores and other variables in England's different Lower Layer Super Output Areas (LSOAs).

- Conduct K-means clustering on the scale of local Authority based on educational deprivation score, financial indicators, environment, and population of dependent children.

- Measure difference by creating a multiple linear regression model based on variables selected by correlation analysis and variables selected by Principal Component Analysis in the scale of LSOA.

2. Methodology

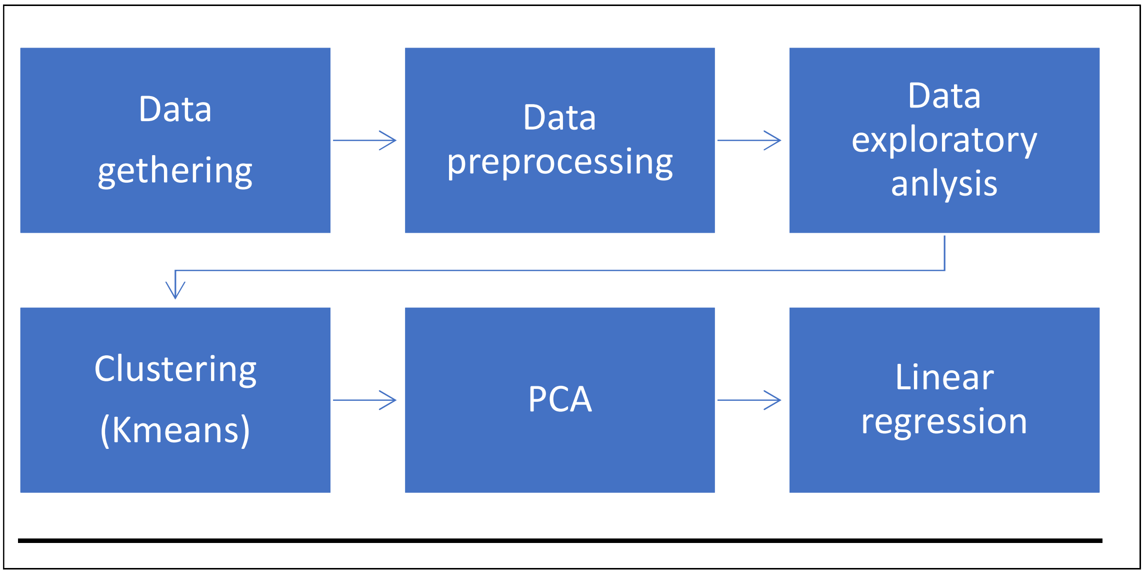


Figure 1 The discovery processes.

2.1 Data gathering

This report utilises the Consumer Data Research Centre (CDRC) Index of Multiple Deprivation (https://data.cdrc.ac.uk/dataset/index-multiple-deprivation-imd). These indices provide measures of relative deprivation across Lower Layer Super Output Areas (LSOAs), designed to improve the study of small area statistics in England. The domains are as follows:

- Income: Assesses the percentage of the population confronting hardship associated with insufficient wages.
- Employment: Evaluates the percentage of the working-age population in a region involuntarily unable to participate in the workforce.
- Education: Assesses the community's absence of educational

achievements and skills.

- Health: Gauges the likelihood of premature mortality and the impact on quality of life due to inadequate physical or mental health.

- Crime: Evaluates the potential for personal and material victimisation at the local level.

- Barriers to Housing & Services: Examines the challenges regarding physical and financial access to housing and local services.

- Living Environment: Evaluates the overall quality of the local environment, encompassing both indoor and outdoor aspects.

Variables name	**Type**	**Definition**
LAdnm	categorical	Local authority
lsoa11nmw	categorical	Lower Layer Super Output Areas
IncScore	numerical	Income Score
EmpScore	numerical	Employment Score
EduScore	numerical	Education, Skills, and Training Score
HDDScore	numerical	Health Deprivation and Disability Score
CriScore	numerical	Crime Score
BHSScore	numerical	Barriers to Housing and Services Score
EnvScore	numerical	Living Environment Score
S_ Depchi	numerical	Share of Dependent Children population in the total population

Table 1: Table of variables

2.2 Data preprocessing

The detection of missing value

Addressing missing values is crucial before data analysis, as neglecting or excluding them could lead to biased or uninformed conclusions. Various suggestions have been put forth in the literature regarding treating missing values(Emmanuel et al., 2021). It is crucial to scan missing values in our data set by using the built-in function summary in R.

Data transformation

One of the research questions in this report aims to compare differences among local authorities in England. Nonetheless, in this data set, the smallest unit is the Lower Layer Super Output Area (LSOA), the lowest geographical area level for census statistics. Summarise local authorities' data using pipeline, group_by, summarise, mean, and sum methods in the dplyr package. Moreover, divide the number of dependent children by the total population to show the proportion of dependent children.

2.3 Data Exploratory Analysis

Distribution

Investigating data distribution is essential because it involves the possibility of the whole population and which metric should be applied. To acquire more information, form the distribution of education deprivation score, utilise methods such as geom_histogram, geom_boxplot, and geom_densit in the ggplot2 package.

Correlation

To investigate the correlations among different variables to educational deprivation score, utilise the cor function, which is built-in function in R, to calculate correlation coefficients and p-value. Two significant points need to be considered. First, selecting a method for calculating correlation coefficients depends on the data type. The Spearman rank order correlation coefficient is computed when variables are not a normal distribution. The coefficient is calculated from −1 to +1 (Puth et al., 2015). Second, interpret each correlation with the p-value. When the p-value is lower than 0.05, it demonstrates statistical significance between the two variables. Statistical significance helps filter out numerous spurious and inconsequential correlations that might arise randomly in the presence of small sample sizes (Komaroff, 2020). After calculating the correlation, utilise the corrplot package for visualisation.

Cluster analysis by K-means

To distinguish the similarity among local authorities, implement the K-means algorithm, a technique to address the difficulty in sorting the collection of patterns(Bagirov et al., 2023). In the clustering process, it relies on assessing the distance between observations. Data points that are close or similar are grouped in the same cluster, while distant or dissimilar ones are placed in separate clusters(Gupta & Chandra, 2022). However, there is a crucial decision in choosing the appropriate number for the cluster, known as the value of k in the K-means algorithm. The effectiveness of the k-means algorithm depends on the chosen value of k, which must be specified for any clustering analysis to occur. Placing different k values in the clustering process will generate distinct results(Ahmed et al., 2020).

Determine k value in k-means (Silhouette Coefficient)

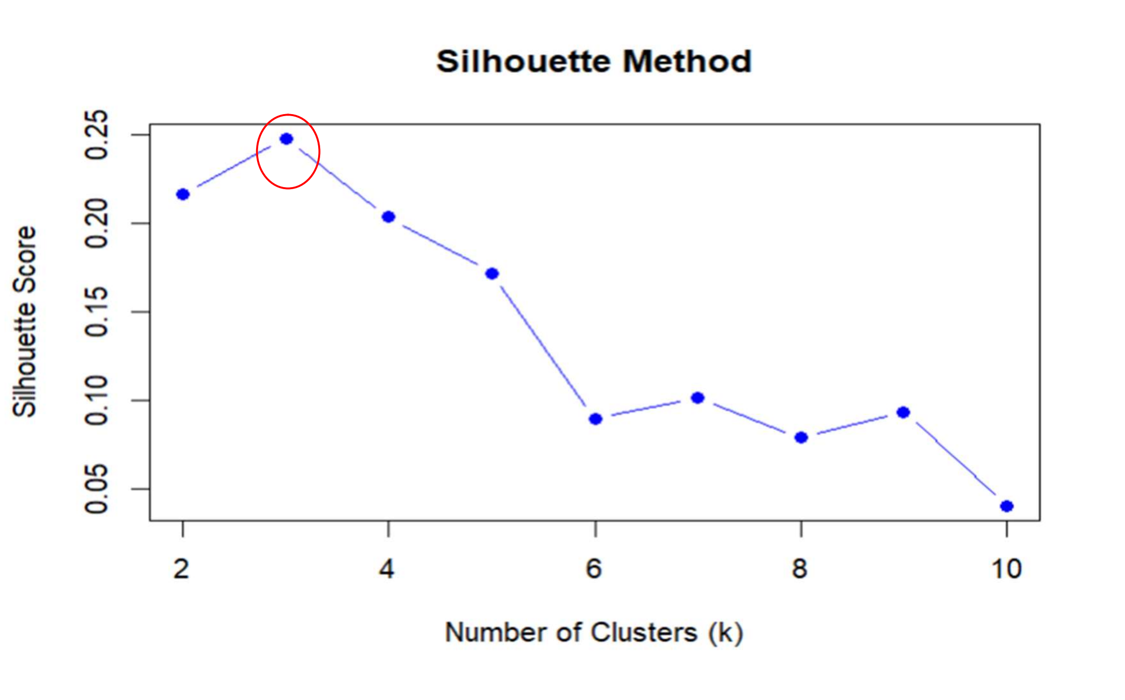


Figure 2: Silhouette Coefficient

The silhouette coefficient serves the purpose of identifying well-separated clusters. This coefficient is computed within the range of -1 to 1; A value close to 1 indicates effective clustering, while a silhouette coefficient approaching -1 signifies misclassification(Bagirov et al., 2023). It was based on the Silhouette Coefficient; when the k-value equals 3, it will perform better in clustering.

Implementing K-mean in R

First, Utilise the kmean method in cluster package and silhouette function to find appropriate clusters iteratively from 2 to 10. Second, use the plot function to create visualisation.

2.4 Principal Component Analysis (PCA)

This report implements PCA to catch crucial information about each variable and assist in visualising the 2-dimensional graph. PCA is a method for reducing dimensionality and converting data from a high dimension into a lower dimension. In the meantime, it conserves essential information and lowers the influence of the noise in the data set (Sanguansat, 2012). The PCA can be used in the prcomp function from R.

2.5 Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) is a statistical method used to forecast the outcome of a response variable by incorporating several explanatory variables. MLR aims to formulate a model that represents the linear association between the independent variables (x) and the dependent variable (y) under examination (Zhang et al., 2019).

In this part, the experiment will be conducted in two parts separately. The first part will utilise the variables selected from the correlation matrix. The other one will calculate the score based on the principal Components 1 and 2. All the processes will be executed with the lm function in R.

3. Results and discussion

3.1 Data Exploratory Analysis

Distribution of dependent variables (education deprivation score in LSOAs)

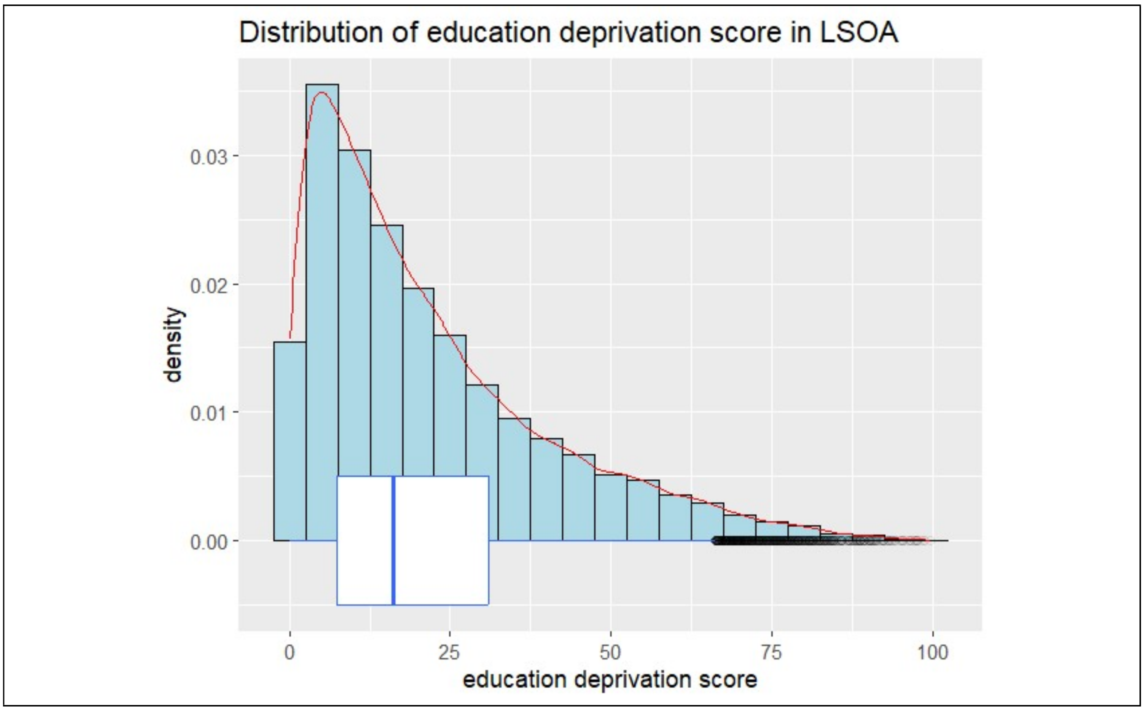


Figure 3: Distribution of educational deprivation score

As shown in Figure 3, the distribution of education deprivation scores in LSOAs is not a normal distribution but a right-skewed distribution. It also indicates that at least 75 percent of LSOAs response score among 7.360 and 30.907.

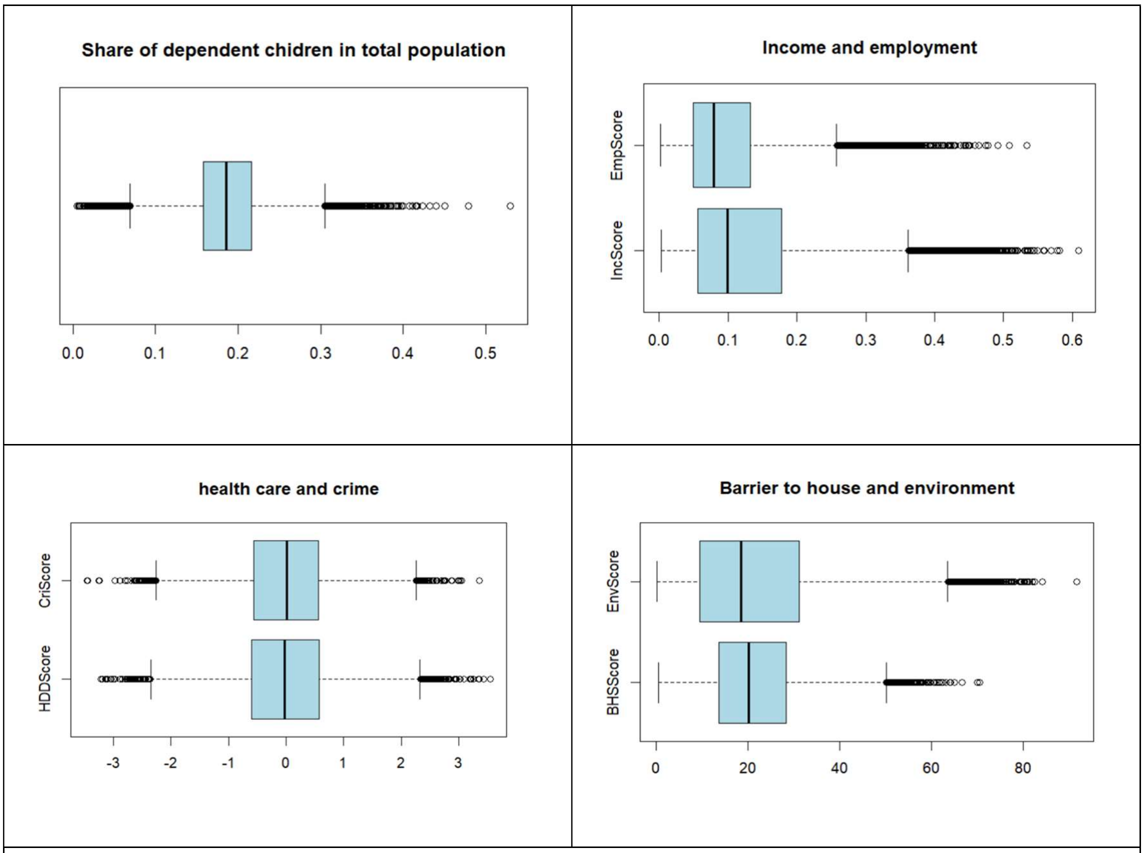
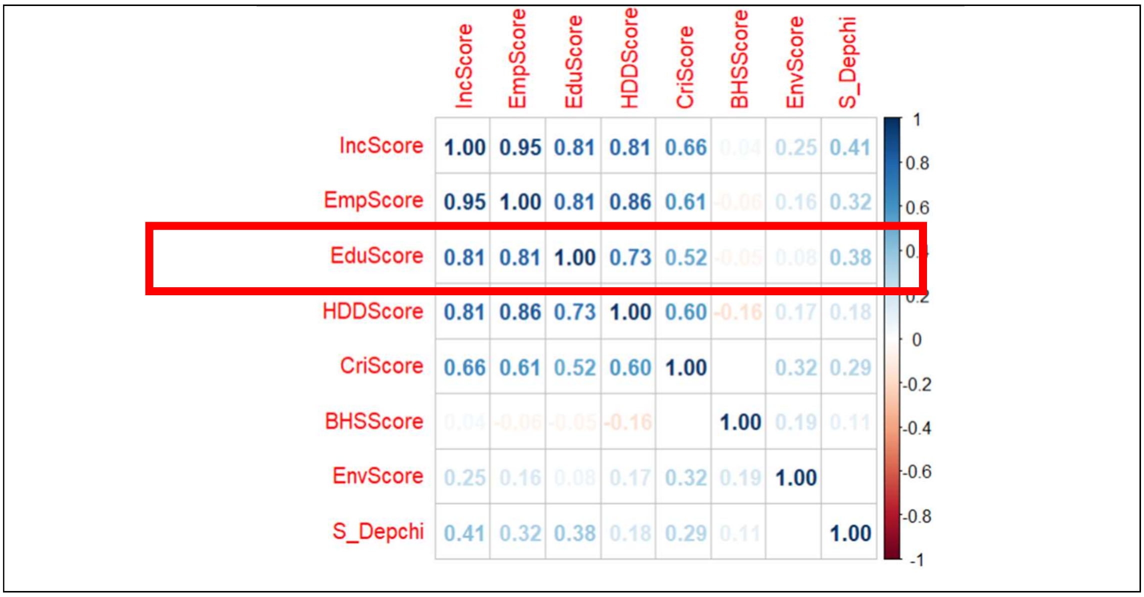


Figure 4 box chart of independent variables

First, most of the share or dependent children in LSOAs lie in 0.02 to 0.03. Second, most income and employment scores lie between 0 and almost 0.4. Third, most health and crime scores lie slightly lower than -2 and 2.5. Fourth, the environment scores lie between 0 and 65, and barrier to house scores lie between 0 and 50.

Correlation analysis in LSOAs



	IncScore	EmpScore	EduScore	HDDScore	CriScore	BHSScore	EnvScore	S_Depchi
IncScore	1.00	0.95	0.81	0.81	0.66	0.04	0.25	0.41
EmpScore	0.95	1.00	0.81	0.86	0.61	-0.06	0.16	0.32
EduScore	0.81	0.81	1.00	0.73	0.52	-0.05	0.08	0.38
HDDScore	0.81	0.86	0.73	1.00	0.60	-0.16	0.17	0.18
CriScore	0.66	0.61	0.52	0.60	1.00		0.32	0.29
BHSScore	0.04	-0.06	-0.05	-0.16		1.00	0.19	0.11
EnvScore	0.25	0.16	0.08	0.17	0.32	0.19	1.00	
S_Depchi	0.41	0.32	0.38	0.18	0.29	0.11		1.00

Figure 5: Correlation among variables

Correlation analysis aims to comprehend and evaluate the linear or nonlinear association between two continuous variables. The sign of the correlation coefficient indicates a positive or negative relationship between two variables, and the value demonstrates the strength of the association. The absolute value between 0.5 and 0.8 illustrates a moderate correlation, and a value above 0.8 means a strong association between two variables (Zou et al., 2003).

Figure 5 demonstrates the correlation among different variables toward the educational deprivation scores. First, the income and employment scores show the same correlation coefficient (0.81), indicating a strong and positive relation toward the educational deprivation score. Second, the correlation coefficient between health and crime scores is 0.73 and 0.52, respectively. These two variables are suggested to be moderately positively correlated to educational deprivation scores. Third, the living environment score and barriers to housing and services score are correlated weakly with the educational deprivation score.

3.1 Discussion

From the distributions in Figure 1,2 among these variables, it is inferred that they are not normally distributed. This result also influences the selection of correlation coefficients in correlation tests. According to Puth et al. (2015), Spearman coefficients are suitable for calculating the correlation between variables that are not normally distributed. In addition, there are outliers within the data. However, as Noble et al. (2019) mentioned, the 2019 indices of deprivation build upon the earlier versions of the release and have been carefully designed to guarantee the strength and dependability of the resulting datasets and reports. Based on this reason, this report did not remove the outliers. The existence of an outlier could affect the regression coefficients and lead to a different model (Dhakal, 2017).

The correlation coefficients shown in Figure 5 imply that financial factors might have a significant correlation to deprivation of education. According to (Checchi, 2001), Educational attainment plays a vital role in income inequality; people with better educational backgrounds tend to have decent salaries in their work. In addition, health and crime rates might also influence inequality in education.

3.2 Cluster analysis and Principal Component Analysis for local authorities

Cluster analysis

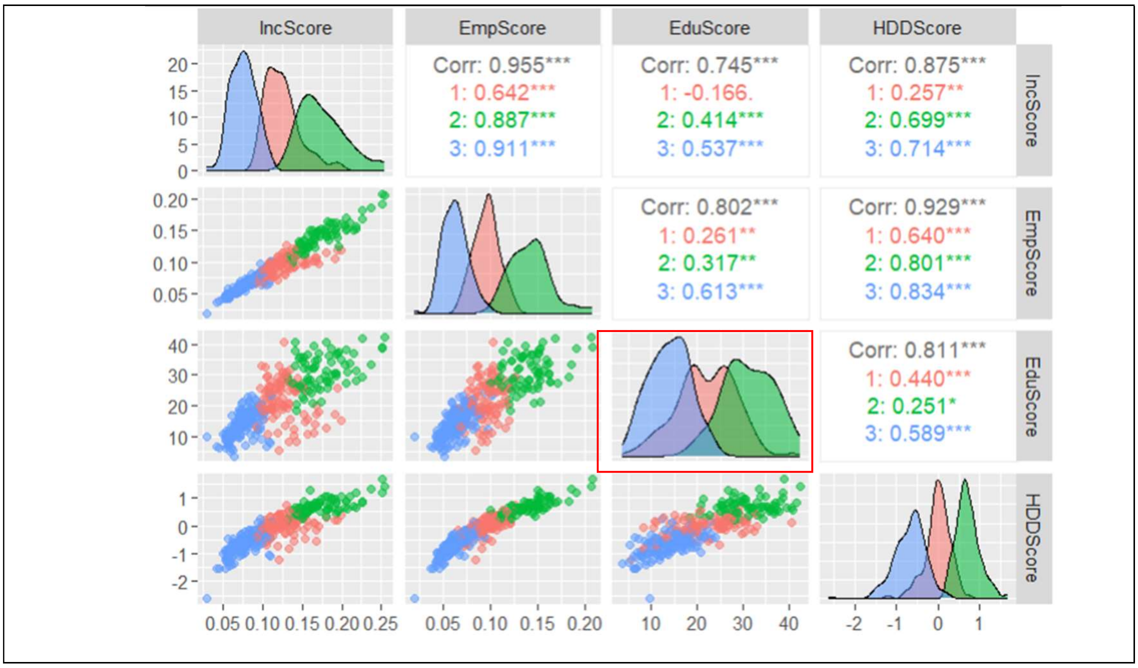


Figure 6 Scatter plot matrix in clusters for education, income, employment, health

The density plots in Figure 6 illustrate that most local authorities in Cluster 2 have higher deprivation scores in education, which range between 15 to 40; meanwhile, they also have comparatively higher scores in income, employment, and health than in clusters 1 and 3. On the contrary, local authorities in Cluster 3 have smaller distributions, which range between 0 and 25 in educational deprivation scores, lower than those in Cluster 2. Furthermore, scatter plots in Figure 6 show that income and employment are significantly associated with education scores because education scores increase with higher income and employment scores. Nonetheless, the boundary between education and health is slightly ambiguous.

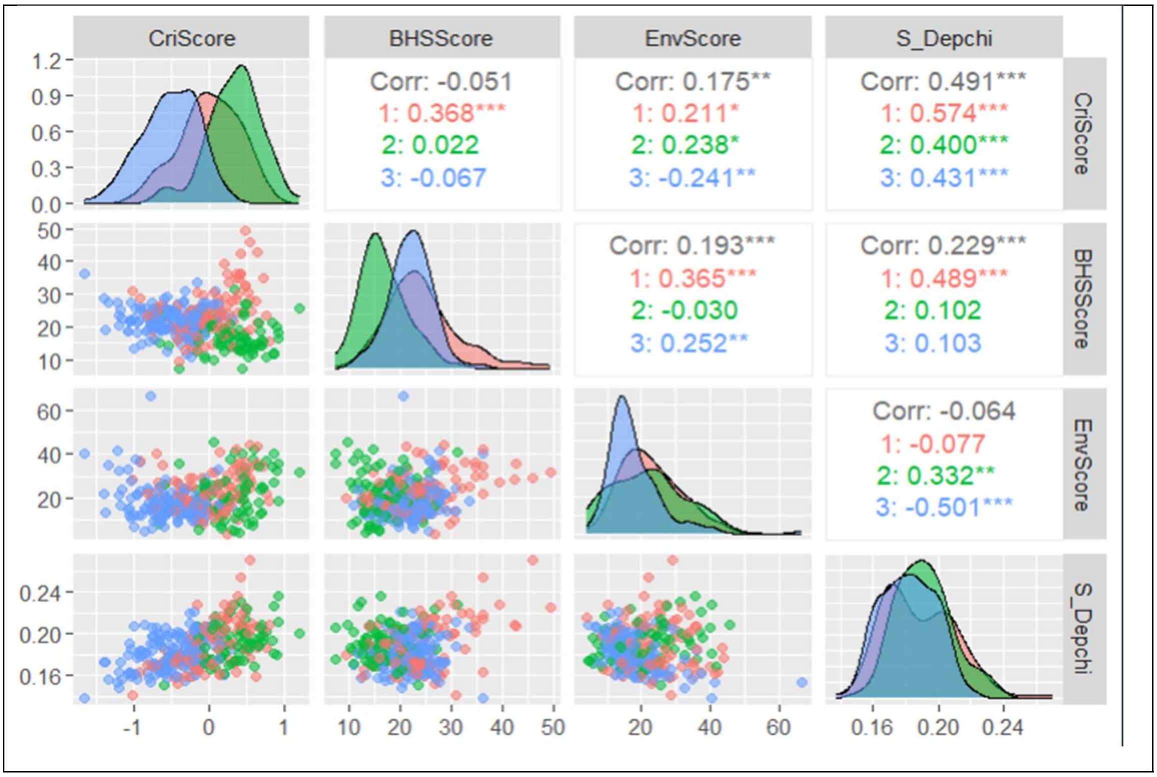


Figure 7 Scatter plot matrix for different clusters with health and crime factors

Figure 7 indicates cluster 2 has a higher density, between 0 and 1 in crime scores, than other clusters. However, cluster 2 has marginally lower Barriers to Housing & Services scores than other clusters.

Principal Component Analysis

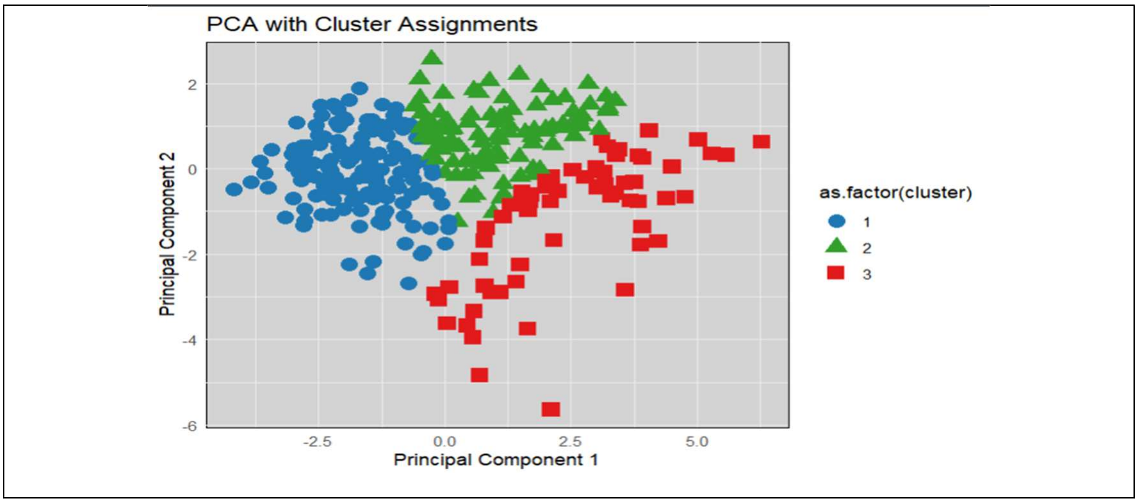


Figure 8 PCA visualisation with cluster

Figure 8 demonstrates that most local authorities in cluster 3 have a higher value in principal component (PC) 1 and less value in PC2. However, most local authorities in clusters 1 and 2 both have a higher value of PC2. The difference between clusters 1 and 2 is that cluster 1 has more negative PC1.

Variables	PC 1	PC 2
IncScore	0.4674875	-0.1062773
EmpScore	0.4695270	0.1007717
EduScore	0.4156735	0.1490627
HDDScore	0.4569765	0.1519071
CriScore	0.3616180	-0.3558651
BHSScore	-0.1544235	-0.6083336
EnvScore	0.1154101	-0.2494053
S_Depchi	0.1070239	-0.6118349

Table 2 PCA rotation

In Table 2, PC1 has high loadings related to income, employment, education, health, and crime score. Nonetheless, In PC2, Barriers to Housing and the proportion of dependent children have a more significant impact than employment, education, health, and crime scores.

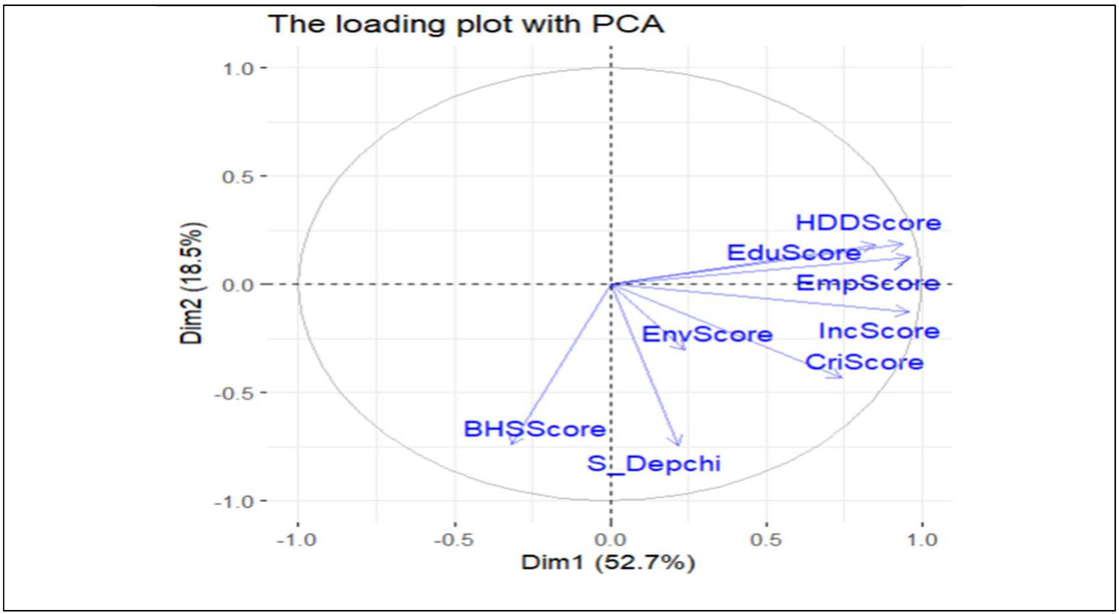


Figure 9 Similarities among the variables

Figure 9 indicates that educational deprivation scores have a stronger correlation to income, employment, health, and crime scores than index: environment, the share of dependent children, and barrier of house.

3.2 Discussion

By Conducting a K-means clustering analysis, it is inferred that cluster 3 has higher educational deprivation scores. In the meantime, it also has the potential to have higher deprivation scores in income, employment, health, and crime. Research conducted by Ferguson and Michaelsen (2015) also had similar results, namely that deprivation of education is highly correlated with financial factors. Moreover, according to Zajacova and Lawrence (2018), the completion of education is associated with better health due to the accumulation of extra knowledge and skills.

In principal component analysis, it is suggested that results are similar to the output computed by Kmeans clustering. Compared to the result from K-means clustering, PCA provides more concise clusters by reducing dimension and noise from the data. In k-means clustering, not all selected variables demonstrate differences, especially the scores in barrier to house and service, environment, and share of dependent children.

3.3 Multi-linear regression analysis

This section compares the performance of multivariate linear regression between different variables selected by different methods. One selects the variables by correlation matrix, and the other utilises PCA. The results are as follows:

Model	Variables	P-value (Coefficients)	P-Value (model)	Adjusted *R2* Value	MSE	F-statistic
A	IncScore	< 2e-16	< 2.05e-14	0.7071	102.8302	1.394e+04
	EmpScore	< 2e-16				
	HDDScore	< 2e-16				
	CriScore	< 2e-16				
B	PCA1	< 2e-16	< 2.2e-16	0.6893	111.8775	2.549e+04
	PCA2	< 2e-16				

Table 3 Result of Multivariate linear regression

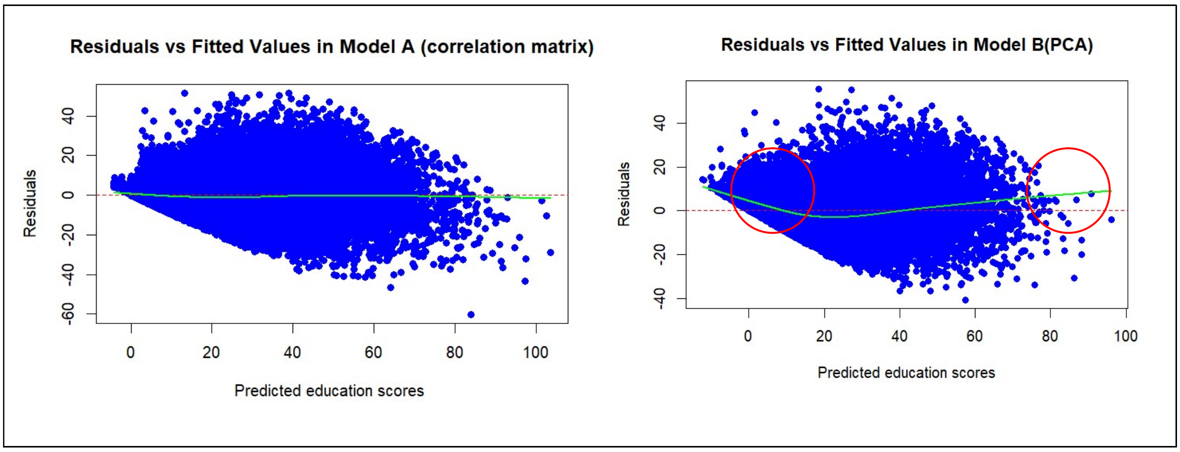


Figure 10 Scatter plots with lowess line about residuals in models

Evaluate performance between model A and model B.

The P-value for each coefficient is lower than 0.05 in models A and B. It is suggested that selected variables in both models are statistically significant and contribute to model prediction.

Adjusted R-squared is a crucial metric for evaluating the accuracy of a regression model. Additionally, it considers both the sample size and model simplicity (Plonsky & Ghanbar, 2018). The Adjusted $R2$ value of model A is around 0.70, which is slightly higher than model B.

The Mean Squared Error (MSE) measures how close a regression line is to a set of data points. Model A has a smaller MSE (102.8302) than the MSE (111.8775) in model B.

The higher F-statistic in Model B (2.549e+04) suggests that model B better fits the data than model A (Sureiman & Mangera, 2020). Moreover, additional predictors or complexity in Model B contribute significantly to explaining the variability in the response variable.

The report by Schützenmeister et al. (2012) suggests that robust linear models must follow both normality and constant variance assumptions. To evaluate homoscedasticity between the two models, figure 10 demonstrates that the trend of residuals in model B fluctuates between 0 and 20. However, the trend of residuals in model A is approximately at 0. According to Trenkler (1998), homoscedasticity describes that the variability should not increase with the magnitude of the predicted values. It is inferred that the homoscedasticity of model A is better than model B.

3.3 Discussion

The Adjusted $R2$ value, MSE, and homoscedasticity indicate that model A performs better than model B. However, Model B has a higher F-test score than Model A. It is inferred that variables used in model B contribute crucially to explaining the variability.

From the result in Table 3, it is suggested that the level of deprivation of education can be predicted by deprivation in income, employment, health, and crime. However, explaining the extent of deprivation of these possible factors in this report is challenging. We must also consider the context behind the educational deprivation, not just accuracy. The same problem is also shown in the research conducted by Ferguson and Michaelsen (2015), which suggests that a significant portion of the impact can be attributed to the whole range of deprivation rather than a few factors.

4. Conclusion

Summary

First, in correlation analysis, it is inferred that factors such as income, employment, health and crime are significantly correlated to deprivation of education. On the other hand, the influence of the environment, barriers to housing and services, and the population have less association with educational deprivation.

Second, from K-means clustering and PCA, a cluster with higher educational deprivation scores is also highly deprived of income, employment, and health.

Finally, in the multi-linear regression model, predict educational deprivation using variables associated with educational deprivation score and variables processed by PCA. It is suggested that even though income, employment, health, and crime significantly contributed to the prediction of education deprivation scores. However, it is also necessary to consider the context of the problem, not just accuracy.

Weakness and Further Research

First, the data has been statically processed. The calculation of each score is not revealed in the documentation. The bias might exist in the data. The outcome of the score could be ambiguous. To refine this situation, it might add more variables, such as the number of educational qualifications in the area or how many schools are there. Doing so might increase the performance of linear regression and cluster models.

Second, from the distributions of all variables in this report, it is indicated that there are several outliers within the data. However, these outliers are not removed because this data set is reported as processed data. These outliers might have a specific meaning in reflecting actual inequality. However, no matter the multi-linear regression model, PCA and K-means clustering are sensitive to outliers.

5. Reference list

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295.

Aungkulanon, S., Tangcharoensathien, V., Shibuya, K., Bundhamcharoen, K., & Chongsuvivatwong, V. (2017). In Thailand, area-level socioeconomic deprivation and mortality differentials: results from principal component analysis and cluster analysis. *International journal for equity in health*, *16*, 1-12.

Bagirov, A. M., Aliguliyev, R. M., & Sultanova, N. (2023). Finding compact and well-separated clusters: Clustering using silhouette coefficients. *Pattern Recognition*, *135*, 109144.

Bradley, S., & Green, C. H. (2020). The economics of education : a comprehensive overview.

Checchi, D. (2001). Education, inequality and income inequality. *LSE STICERD Research Paper*(52).

Dhakal, C. P. (2017). Dealing with outliers and influential points while fitting regression. *Journal of Institute of Science and Technology*, *22*(1), 61-65.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*(1), 1-37.

Ferguson, N. T., & Michaelsen, M. M. (2015). Money changes everything? Education and regional deprivation revisited. *Economics of Education Review*, *48*, 129-147.

Gupta, M. K., & Chandra, P. (2022). Effects of similarity/distance metrics on k-means algorithm with respect to its applications in IoT and multimedia: A review. *Multimedia Tools and Applications*, *81*(26), 37007-37032.

Hart, C. S. (2019). Education, inequality and social justice: A critical analysis applying the Sen-Bourdieu Analytical Framework. *Policy Futures in Education*, *17*(5), 582-598.

Komaroff, E. (2020). Relationships between p-values and Pearson correlation coefficients, type 1 errors and effect size errors, under a true null hypothesis. *Journal of Statistical Theory and Practice*, *14*(3), 49.

Noble, S., McLennan, D., Noble, M., Plunkett, E., Gutacker, N., Silk, M., & Wright, G. (2019). The English indices of deprivation 2019. *CLG Ministry of Housing, Editor. London*.

OECD. (2022). *How does educational attainment affect participation in the labour market?* https://doi.org/doi:https://doi.org/10.1787/0c0b63ae-en

Perera, V. (2022). Little progress on attainment gap could have lifelong impact on disadvantaged children. *Nuffield Foundation*. https://www.nuffieldfoundation.org/news/little-progress-on-attainment-gap-could-have-lifelong-impact-on-disadvantaged-children

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *The Modern Language Journal*, *102*(4), 713-731.

Puth, M.-T., Neuhäuser, M., & Ruxton, G. D. (2015). Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits. *Animal Behaviour*, *102*, 77-84.

Raghupathi, V., & Raghupathi, W. (2020). The influence of education on health: an empirical assessment of OECD countries for the period 1995–2015. *Archives of Public Health*, *78*(1), 1-18.

Sanguansat, P. (2012). *Principal component analysis*. BoD–Books on Demand.

Serafino, P., & Tonkin, R. (2014). Intergenerational transmission of disadvantage in the UK & EU.

Sureiman, O., & Mangera, C. M. (2020). F-test of overall significance in regression analysis simplified. *Journal of the Practice of Cardiovascular Sciences*, *6*(2), 116-122.

Zajacova, A., & Lawrence, E. M. (2018). The relationship between education and health: reducing disparities through a contextual approach. *Annual review of public health*, *39*, 273-289.

Zhang, Z., Li, Y., Li, L., Li, Z., & Liu, S. (2019). Multiple linear regression for high efficiency video intra coding. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

Zou, K. H., Tuncali, K., & Silverman, S. G. (2003). Correlation and simple linear regression. *Radiology*, *227*(3), 617-628.

6. Appendix

```
# Install and load the foreign package
library(factoextra)
library(RColorBrewer)
library(GGally)
library(cowplot)
library(cluster)
library(corrplot)
library(tidyverse)
library(foreign)
library(ggplot2)
library(dplyr)

##[1]extract data  (Read the DBF file into a data)##
#[1-1]Read data of England 2019
dbf_file <- "IMD_2019.dbf"
Imd_2019_data <- read.dbf(dbf_file)
##[2]Data preprocess##
#[2-1-1]Select required variables
IMD_2019_LSOA<-
Imd_2019_data[grepl("IncScore|EmpScore|EduScore|HDDScore|CriScore|BH
SScore|EnvScore|LADnm|lsoa11nmw|TotPop|DepChi",
colnames(Imd_2019_data))]
IMD_2019_LSOA$S_Depchi<-
IMD_2019_LSOA$DepChi/IMD_2019_LSOA$TotPop
#[2-1-2]View missing value (No missing V)
summary (IMD_2019_LSOA)
#[2-1-3]Transforming group by mean and sum for clustering
IMD_2019_LA<-IMD_2019_LSOA %>%
group_by(LADnm)%>%
summarise(
  IncScore = mean(IncScore),
  EmpScore = mean(EmpScore),
  EduScore = mean(EduScore),
  HDDScore = mean(HDDScore),
  CriScore = mean(CriScore),
  BHSScore = mean(BHSScore),
  EnvScore = mean(EnvScore),
  S_Depchi = sum(DepChi)/sum(TotPop),
  )
##[3]Data exploratory analysis##
#[3-1]Descriptive analysis
#box chart for independent variables
#Population
boxplot(IMD_2019_LSOA$S_Depchi, col = "lightblue", main = "Share of
dependent chidren in total population",horizontal = TRUE)
#score
boxplot(IMD_2019_LSOA[c("IncScore","EmpScore")], col = "lightblue",
main = "Income and employment",horizontal = TRUE)
boxplot(IMD_2019_LSOA[c("HDDScore","CriScore")], col = "lightblue",
main = "health care and crime",horizontal = TRUE)
boxplot(IMD_2019_LSOA[c("BHSScore","EnvScore")], col = "lightblue",
main = "Barrier to house and environment",horizontal = TRUE)


#Distribution of Education Score in LSOA
ggplot(IMD_2019_LSOA, aes(x = EduScore)) +
geom_histogram(binwidth = 5, aes(y = ..density..), fill = "lightblue", color =
"black") +
geom_boxplot(width = 0.01,
            position = position_dodge(100),
            colour = "#3366FF",
            outlier.colour = "black", outlier.shape = 1 ,outlier.alpha = 0.1)+
geom_density(color = "RED")+
labs(title = "Distribution of education deprivation score in LSOA", x =
"Education deprivation score", y = "density")

#[3-2]Correlation matrix
#compare with different Score
IMD_2019_score_pop<-IMD_2019_LSOA[grepl("Score|S_Depchi",
colnames(IMD_2019_LSOA))]
#continuous value ustilising pearson
#p-value
cor_matrix <- cor(IMD_2019_score_pop,method = "spearman")
res1 <- cor.mtest(IMD_2019_score_pop, conf.level = .95)
corrplot(cor_matrix, p.mat = res1$p, sig.level = .05,method = "number")
#[3-3]clustering analysis
cluster_ds_1 <-
IMD_2019_LA[grep("IncScore|EmpScore|EduScore|HDDScore|CriScore|BHS
Score|EnvScore|S_Depchi", colnames(IMD_2019_LA))]
scaled_cluster_ds_1<-scale(cluster_ds_1)

silhouette_scores <- sapply(2:10, function(k) {
  kmeans_result <- kmeans(cluster_ds_1, centers = k)
  silhouette_avg <- silhouette(kmeans_result$cluster, dist(scaled_cluster_ds_1))
  return(mean(silhouette_avg[, 3]))
})

# silhouette_scores plot
plot(2:10,
     silhouette_scores,
     type = "b", pch = 19, col = "blue",
     xlab = "Number of Clusters (k)", ylab = "Silhouette Score", main =
"Silhouette Method")
#clustering
kmeans_result <- kmeans(scaled_cluster_ds_1, centers = 3)
cluster_assignments <- kmeans_result$cluster
cluster_centers <- kmeans_result$centers
cluster_ds_1$cluster<-as.factor(cluster_assignments)

#Scater plot matrix 1
ggpairs(cluster_ds_1,
        columns = 1:4,
        aes(color = cluster,
            alpha = 0.5))
#Scater plot matrix 2
ggpairs(cluster_ds_1,
        columns = c(5,6,7,8),
        aes(color = cluster,
           alpha = 0.5))
#PCA
pca<-prcomp(IMD_2019_LA[,2:9],scale=TRUE)
IMD_2019_LA.pca<-data.frame(
LA=IMD_2019_LA$LADnm,
PC1=pca$x[,1],
PC2=pca$x[,2]
)
# show pca rotation
pca$rotation[, 1:2]
#Graph of the variables
fviz_pca_var(pca, col.var = "blue",
            repel=TRUE,
            alpha=0.5,
            title="The loading plot with PCA")
IMD_2019_LA$cluster<-kmeans_result$cluster
IMD_2019_LA$pca1<-pca$x[,1]
IMD_2019_LA$pca2<-pca$x[,2]

ggplot(IMD_2019_LA, aes(x = pca1, y = pca2, color =as.factor(cluster),shape
= as.factor(cluster))) +
  geom_point(size = 4) +
  labs(title = "PCA with Cluster Assignments", x = "Principal Component 1", y
= "Principal Component 2") +
  theme_minimal() +
  scale_color_manual(values = c("#1f78b4", "#33a02c", "#e31a1c"))+ #
Custom color palette
  theme(panel.background = element_rect(fill = "lightgray"))


#////////[4-1]Predictive analysis\\\\\\\\\\\\
##Model A Variables selected by correlation matrix
ML_dataset<-IMD_2019_LSOA %>% select(c(3:7))
#scale data
ML_dataset[c("IncScore","EmpScore","HDDScore","CriScore")] <-
```

```
scale(ML_dataset[c("IncScore","EmpScore","HDDScore","CriScore")])
#split data set into train data and test data
sample <- sample(c(TRUE, FALSE), nrow(ML_dataset), replace=TRUE,
prob=c(0.7,0.3))
IMD_Score_train  <- ML_dataset[sample, ]
IMD_Score_test   <- ML_dataset[!sample, ]
#train the model
mod_IMDScore <- lm(
formula=EduScore ~ .,
data=ML_dataset)
#Check model detail
summary(mod_IMDScore)
#predict with test data
predictions <- predict(mod_IMDScore, newdata = IMD_Score_test)
#model evaluation
mse <- mean((predictions - IMD_Score_test$EduScore)^2)
cat("Mean Squared Error (MSE):", mse, "\n")
rmse <- sqrt(mse)
cat("Root Mean Squared Error (RMSE):", rmse, "\n")
mae <- mean(abs(predictions - IMD_Score_test$EduScore))
cat("Mean Absolute Error (MAE):", mae, "\n")

#residual plot
plot(fitted(mod_IMDScore), residuals(mod_IMDScore), pch = 16, col = rgb(0,
0, 1, 0.3),
     xlab = "Predicted education scores ", ylab = "Residuals",
     main = "Residuals vs Fitted Values in Model A (correlation matrix)")
abline(h = 0, col = "red", lty = 2)
lines(lowess(fitted(mod_IMDScore), residuals(mod_IMDScore)), col = "green",
lwd = 2)

## Model b Use PCA
ML_dataset_2<-IMD_2019_LSOA %>% select(c(3:9,12))
ML_dataset_2_without_index<-ML_dataset_2 %>% select(-c(EduScore))
# Conducting PCA
ML_pca<-prcomp(ML_dataset_2_without_index,scale=TRUE)
ML_dataset_pca<-data.frame(
  EduScore=ML_dataset_2$EduScore,
  PC1=ML_pca$x[,1],
  PC2=ML_pca$x[,2]
)
# Spilt data set
sample <- sample(c(TRUE, FALSE), nrow(ML_dataset_pca), replace=TRUE,
prob=c(0.7,0.3))
Score_train  <- ML_dataset_pca[sample, ]
Score_test   <- ML_dataset_pca[!sample, ]
# train model
mod_IMDScore_pca <- lm(
  formula=EduScore ~ .,
  data=Score_train)

# View model's detail
summary(mod_IMDScore_pca)
# predict by test data
predictions <- predict(mod_IMDScore_pca, newdata = Score_test)
# evaluate model
mse <- mean((predictions - Score_test$EduScore)^2)
cat("Mean Squared Error (MSE):", mse, "\n")
rmse <- sqrt(mse)
cat("Root Mean Squared Error (RMSE):", rmse, "\n")
mae <- mean(abs(predictions - Score_test$EduScore))
cat("Mean Absolute Error (MAE):", mae, "\n")
# Residual plot
plot(fitted(mod_IMDScore_pca), residuals(mod_IMDScore_pca), pch = 16, col
=rgb(0, 0, 1, 0.3),
     xlab = "Predicted education scores", ylab = "Residuals",
     main = "Residuals vs Fitted Values in Model B(PCA)")
abline(h = 0, col = "red", lty = 2)
lines(lowess(fitted(mod_IMDScore_pca), residuals(mod_IMDScore_pca)), col
= "green", lwd = 2)
```