Imperial College London

EXPLORATORY DATA ANALYSIS AND VISUALISATION

IMPERIAL COLLEGE LONDON

DEPARTMENT OF MATHEMATICS

Assessment 3 - London Schools Analysis

Author:
Tom Richardson (CID: 01349943)

Date: April 7, 2025

1 Introduction

This exploratory data analysis has been conducted for Dr Laura Gilbert CBE and her team at 10 Downing Street. It examines the current state of primary and secondary schools across London using a rich dataset containing school meta-data, performance indicators, and socioeconomic factors at both the school and neighbourhood levels. The aim of the report is to uncover patterns and potential insights, to prepare everyone for the upcoming briefing with representatives from the Department for Education and the Mayor of London's office. To accompany this written report, I have included a concise presentation that offers a high-level summary of the findings and recommendations, aimed at facilitating productive discussion during our first meeting at 10 Downing Street on the 10th May 2025.

2 Data Quality

Ensuring data quality is a critical first step in any data analysis pipeline, as it underpins the validity and reliability of all downstream insights. This involves checking for missing values and outliers, validating data sources, and confirming that data collection methods are consistent and well-documented.

Concerning validating our data, I looked at the top 30 rows and compared these against resources online such as the London Schools Atlas and the respective webpage for each school. This preliminary check reassured me that the spatial data, school meta-data, and the LSOA data was roughly in line with what we'd expect. However, during this brief analysis it became evident that there were a number of missing entries.

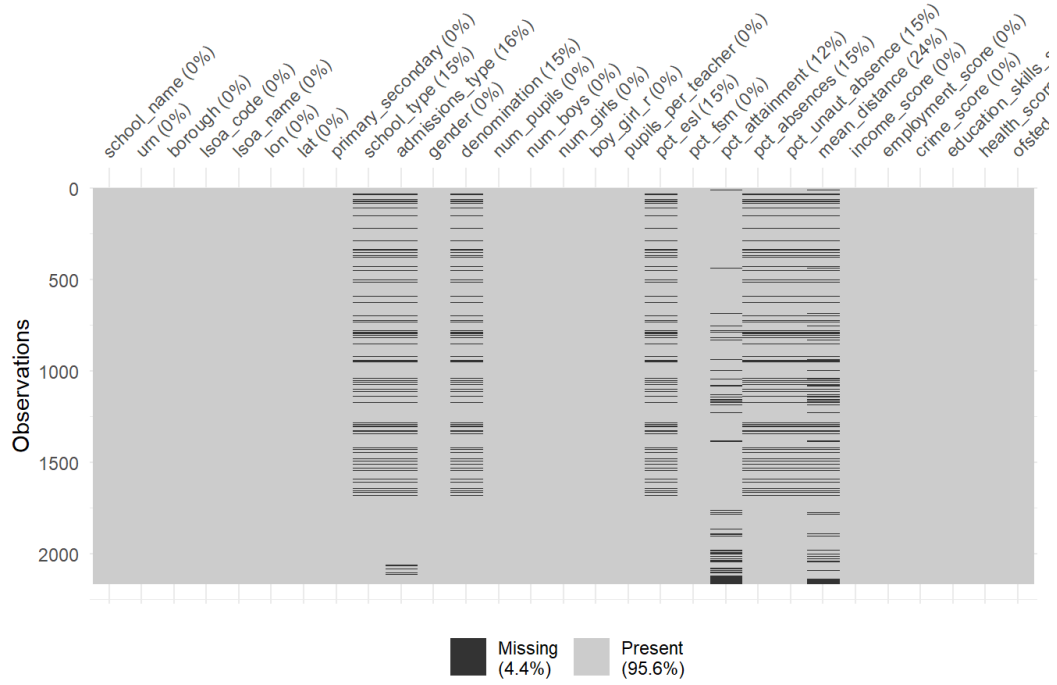


Figure 1: Missing data matrix showing presence (grey) and absence (black) of values across variables. Columns represent variables and rows represent school observations.

Most of the columns have no missing entries at all suggesting reliability for our data collection methods, or that the data had already been cleaned. We examined the missing data and found that missing

values in variables such as pct_attainment and school_type were more likely in schools with worse Ofsted ratings or higher deprivation scores. The distribution tables below supports that missingness is dependent on observed data, supporting the assumption of Missing at Random (MAR). To address the missing data, we applied multiple imputation using chained equations (MICE). Numeric variables were imputed using predictive mean matching (PMM), while categorical variables were imputed using polytomous logistic regression. Five imputed datasets were generated, and the first completed version was merged back into the original data, replacing only missing values. This allowed us to retain the full sample size while reducing potential bias from if we had simply done listwise deletion.

Table 1: Ofsted Ratings Across All Data

Rating	Number	%
1	555	25.6%
2	1395	64.4%
3	206	9.5%
4	10	0.5%
Total	2166	100%

Table 2: Missing school_type data by Ofsted Rating

Rating	Number	%
1	67	20.7%
2	197	60.8%
3	55	17.0%
4	5	1.5%
Total	324	100%

We also had some clearly erroneous data within the pupil count variables. To rectify this, we assumed that the num_girls and num_boys variables were correct, and that also if it was a single gender school, then clearly there shouldn't be students of the opposite gender. Using this, we had to clean the num_girls and num_boys variables, correct num_pupils when it wasn't the sum of num_girls and num_boys, and correct boy_girl_r. We also had 40 entries where they were zero for each of these pupils variables. There were another 18 entries which had pupils_per_teacher equal to zero. For these, we removed them for our later EDA concerning pupil count as we showed them to be MCAR using Little's MCAR test. Please see attached my R markdown for verification.

For data quality, we also assessed outliers as I show in my R markdown. Once we had cleaned up the number of pupils variables, we only had a few variables with outliers, all of which are inappropriate to impute given the nature of the data i.e the mean distance for some voluntary controlled schools might be much higher than other types of schools. However, this did help identify other erroneous data such as some of the percentage variables having entries above 1.

3 Clustering

To aid my exploratory data analysis in preparation for this meeting, I decided to use clustering to provide insights into the data. I explored a variety of different clustering possibilities but decided to focus on two. Firstly, I clustered the local socioeconomic characteristics, so that we can create a rating for each LSOA. This will help provide context to the struggle that a school might be facing given the challenges of the neighbourhood and we can establish how best to tackle them on a city-wide scale.

Secondly, I explored clustering for "borderline" schools. I hypothesized that certain schools may just need assistance with a fixable problem. For this, I looked at the features pct_attainment, pct_absences, "ofsted" and pupils_per_teacher. The idea was government can provide assistance for these variables and we could use clustering to identify schools further away from their cluster; that were potentially able to get to the next level. For example, maybe they had a level 3 Ofsted report but actually have the attributes of a good level 2 Ofsted school. Then, with the correct support, the school can be assisted to that level, giving assistance to its growth over the next few years.

To visualise and validate these clusters, we employed dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE). When plotted, these reduced dimensions reveal the separability of clusters and help to uncover latent structures that may not be apparent when inspecting variables individually.

3.1 Clustering by Local Socioeconomic characteristics

We collated the five LSOA numeric variables for this clustering and prior to clustering, all variables were standardised to ensure comparability. I've included my PCA analysis to visualise the clustering structure The first two principal components captured 85.2% of the total variance, with PC1 explaining 71% and PC2 contributing a further 14.2%.

- **PC1** had strong positive loadings across all variables, particularly on `income_score`, `education_skills_score`, and `employment_score`, indicating that it can be interpreted as a general measure of socioeconomic advantage.
- **PC2** showed the highest loading on `crime_score`, which separates areas of similar deprivation based on local safety.

We used the elbow method to select $k = 3$ as the optimal number of clusters and performed k-means clustering on the standardised LSOA features. The clusters were then visualised in PCA space (Figure 2), highlighting meaningful groupings of schools based on the socioeconomic conditions of their surrounding areas. Further below, I use this clustering in my spatial exploratory data analysis,

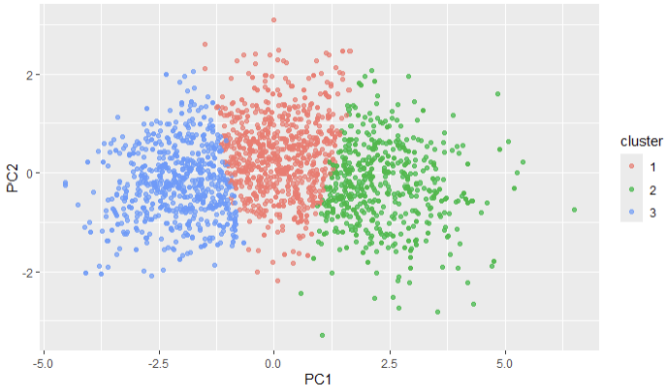


Figure 2: LSOA Clusters in PCA Space. Clusters formed using k-means on five standardised LSOA deprivation indicators.

3.2 Clustering transition schools

Similarly to the above, we have our PCA graph of this clustering to showcase our methodology and validate the clustering. We used k-means clustering again rather than hierarchical, and chose k=3 based on preliminary analysis. To identify outliers with each cluster, we computed the Euclidean distance from each school to its cluster center in the PCA space.

Schools in the top 15% furthest away from their cluster center were labelled as transition schools. These institutions represent those at risk of declining performance or those improving more rapidly than their peers. By visualising these on the PCA plot, we gain insight into which schools might warrant support first as it would be more efficient.

This approach provides a systematic, data-driven method for flagging schools that are behaving differently from the typical patterns in their peer group, and thus might benefit from strategic intervention. I have flagged these schools and we can discuss possible intervention methods at our meeting in May.

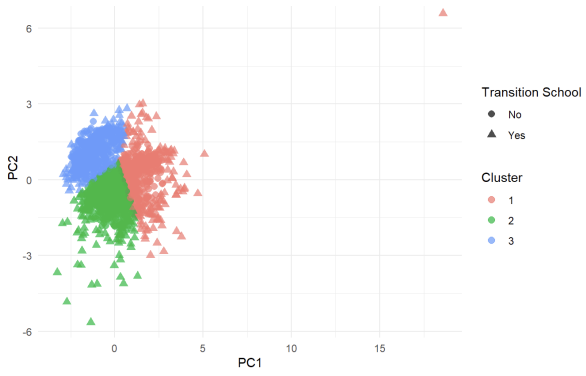


Figure 3: Transition Schools Identified in PCA Space with Transition Schools Highlighted. Transition schools = top 15% furthest from cluster centers

4 Exploratory Spatial Analysis

This section showcases several interesting trends we found within the data that we should discuss further at 10 Downing Street. Firstly, we looked at the LSOA clusters that we developed earlier, helping us identify which neighbourhoods in London had the least beneficial environment to schools within it.

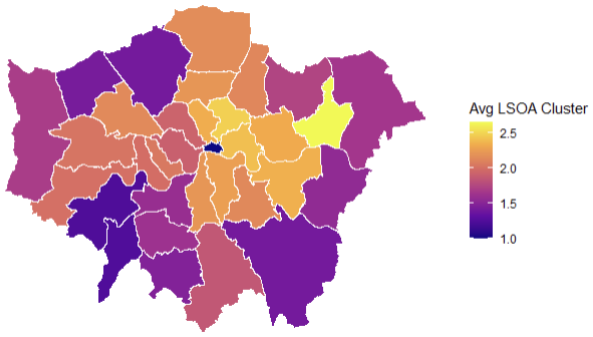
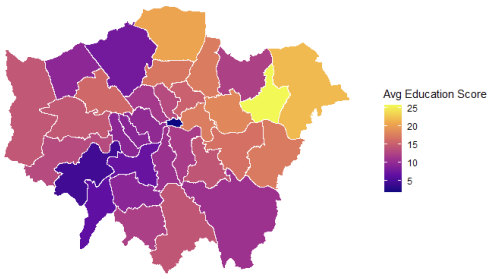


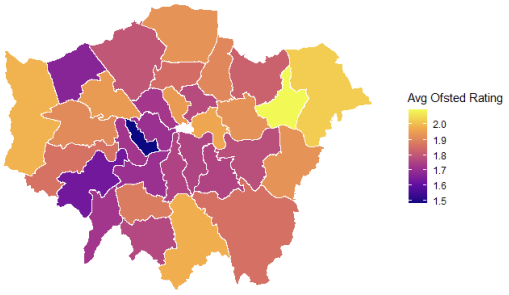
Figure 4: Average LSOA Cluster per Borough.

As you can see in the Choropleth (Figure 4) we tend to see East London boroughs having a higher average, where Barking and Dagenham has the highest average at over 2.5.

The outer boroughs tended to have a lower average, except for the northern boroughs such as Enfield and Haringey. The lowest average was the City of London but there was only one entry here. Figure 5(a) is another choropleth map that compliments Figure 4 as it highlights the East vs West outcomes in attainment and skills of the local population. Included in the R markdown is a breakdown of each of the LSOA choropleth maps which we can discuss about in person; the team may be more interested to explore Crime Score by borough given that knife crime is prolific in the media at the moment.



((a)) Average Education Skills Score per Borough.



((b)) Average Ofsted Rating per Borough.

Figure 5: Education and Ofsted Score Comparisons by Borough

Concerning the performance of schools, I have also included a choropleth map of the average Ofsted rating per borough. Here we can once again see the Barking and Dagenham is an outlier; it is the only borough with an average Ofsted rating significantly over 2. A distinction about this graph though is that boroughs just below the river seem to perform best, with this trend not being as clear in the previous two graphs. The top performing borough was Kensington and Chelsea, excluding the City of London borough which I have removed from the conditional formating. This removal has helped improve the comparison of the other boroughs.

Additionally, I have included a Box plot of Attainment by Borough. This shows the range, lower quartile, upper quartile and median of each of the boroughs, along with any outliers. Once again, we see that Barking and Dagenham are lowest, and that Kensington and Chelsea, and the City of London are significantly higher than the others. This graph was included as it perfectly visualises that there is a significant number of schools below 50% in percentage attainment. This should be an urgent problem for Mayor of London's team and should be addressed promptly.

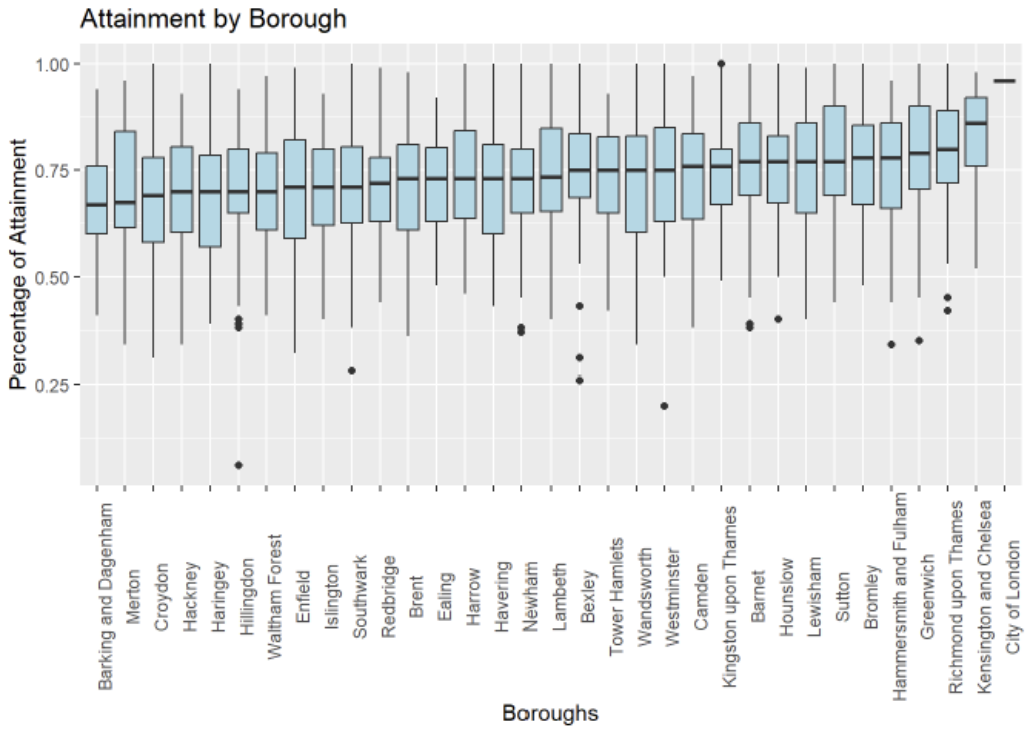


Figure 6: Distribution of Percentage of Attainment by Borough. Boroughs are ordered by median attainment.

5 Conclusion

This report has explored the landscape of London's schools using the dataset provided. Beginning with data quality checks and imputations, we ensured a clean analytical base for our exploration. Through clustering techniques and dimensionality reduction, we identified patterns of similarity across schools and boroughs, allowing us to highlight not only consistent groupings but also those schools which stand out as borderline cases - those least typical of their peers.

We showed spatial insights through choropleth maps, revealing stark spatial divides in educational outcomes and deprivation levels. Notably, boroughs east of the capital tended to have lower attainment and higher deprivation scores, while pockets in West London and just south of the Thames exhibited stronger Ofsted ratings and educational outcomes. The removal of extreme outliers like the City of London helped enhance these comparisons.

The use of PCA and K-means clustering allowed us to summarize the multifaceted data into intuitive insights. These techniques not only supported our understanding of broader trends but also enabled us to pinpoint boroughs and schools that may warrant further policy attention.

Overall, this initial analysis provides a strong foundation for our meeting on the 10th May where we will discuss recommendations for London's education system. Further work could build on these findings by incorporating temporal data, funding levels, or pupil progress metrics to assess changes over time.