Schools in London - Exploratory Data Analysis

Saad Sharif

February 20, 2026

1 Introduction

In this report, we will perform an exploratory analysis of schools in London. The dataset we will analyse contains 2,166 observations and 29 variables. The dataset consists information related to the schools, such as school type, attainment rates, ofsted ratings, as well as data related to the region the schools are located, such as income and crime scores.
First, we will begin by analysing the quality of the information recorded in the dataset, such as identifying missing information and anomalies, and what actions can be taken to remedy these. Next we will perform spatial analysis using information such as longitude, latitude and borough, followed by an analysis of the numeric variables to identify any potential clusters. Finally, we will perform a dimension reduction technique, principle component analysis, and determine whether this can lead us to find any further clusters.

2 Data Quality and Anomalies

2.1 Missing Data

Before performing any data analysis, it is always important to check the quality of the dataset and decide whether any cleansing or imputations need to be done. In our dataset we have found 8 of the 29 variables to consist of missing information, with the variable "Mean Distance" having the most missing data in 520 observations as seen in **figure 1**.

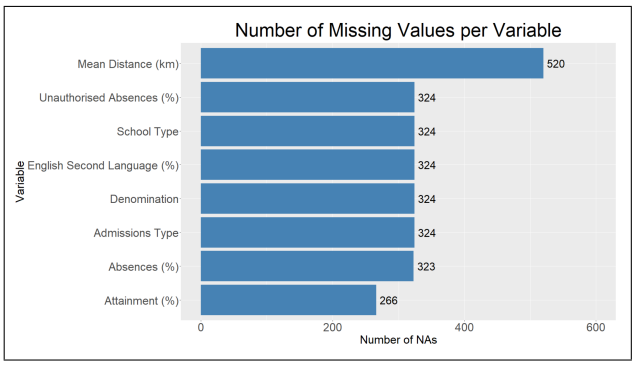

Figure 1: Number of missing values per variable

We can analyse the dataset to determine whether the missing information falls under MCAR (Missing Completely at Random), MAR (Missing at Random) or NMAR (Not Missing at Random). This can be done by using the TestMCARNormality function in R and taking the null hypothesis to be MCAR and using a significance level of 5% to determine whether to reject this null hypothesis. When carrying out this test on the dateset, we have found a significant P-value of 1.82 x 10^{-280} using Hawkin Test and 2.71 x 10^{-6} using Non-Parametric Test, therefore we can reject the null hypothesis of MCAR.

However, we still do not know whether the dataset falls under MAR or NMAR. Before finding this out, it would also be good to identify any patterns between the occurrences of missing information between these 8 variables as this could indicate any potential relationships between them. The following bar chart shows the combinations of missing information and how frequent they are in the dataset.

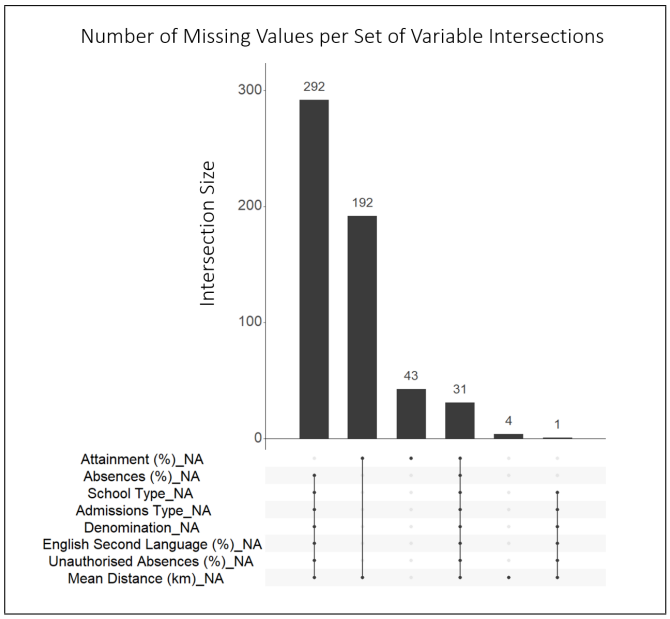

Figure 2: Number of Missing Values per Intersections

As we can see in **figure 2**, the most common intersection of variables with missing information are "Absences (%)", "School Type", "Admissions Type", "Denomination", "English Second Language (%)", "Unauthorised Absences (%)" and "Mean Distance" with 292 observations, followed by a second intersection formed by the variables "Attainment (%)" and "Mean Distance" only with 192 observations.

One method to determine whether a variable is MAR is by comparing the distribution of other variables in the dataset between cases with and without the missing values in the set of variables of interest.

With respect to intersection 1, when we compare the distribution in our 292 observations compared to all other observations in our data set for the other attributes, there are

no obvious differences. Therefore, the missingness of data in this intersection of variables do not seem to be related to another attribute.

On the other hand when we do the same analysis for intersection 2, and compare the distribution of other attributes within and outside these 192 observations, we can see some differences. For example in **figure 3**, when observing the distribution of the "Number of Pupils" and "Free School Meals (%)", we can see that all the quartiles are significantly lower when there is missing information in "Attainment (%)" and "Mean Distance" in comparison to when these attributes do have information. As a result, we can make a reasonable assumption that the dataset is MAR.

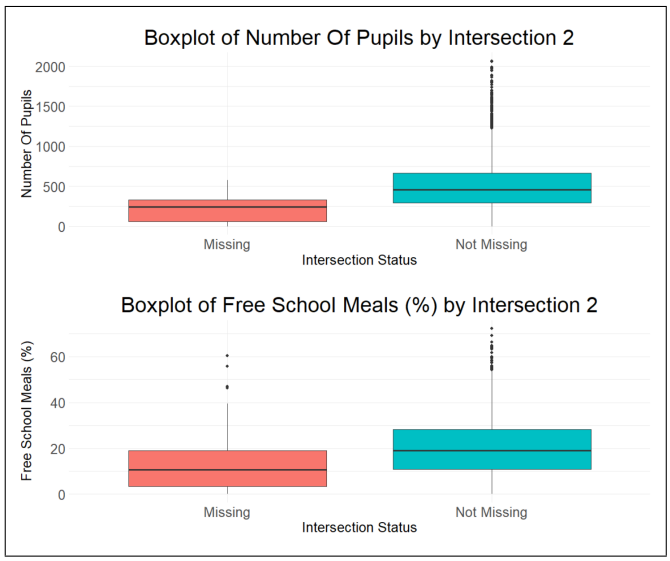

Figure 3: Varying Distributions by Intersection 2

2.2 Anomalies

In addition to missing data, several outliers were identified in various attributes which may need to be further addressed. Some examples of extreme outliers have been shown in **figure 4**. The first boxplot shows an observation with a "Attainment (%)" of 1.55, which would not make sense as we would expect this percentage variable to only have values ranging from 0 to 1 . Another example of an outlier is in the second boxplot where we can see an observation with "Absences (%)" of 0.7, which although is in the range 0 to 1, is quite extreme. The histogram in the third visual shows observations which have a "Pupils Per Teacher" of 0, which we would expect to be impossible since that would mean a school has no pupils at all.

Other issues identified were the way in which "Boy Girl Ratio" was being calculated, which appeared to be "Numner of Boys" / "Number of Girls". his seemed problematic as it would heavily skew the attribute and cause extreme values. A better calculation for this variable would be "Number of Boys" / ("Number of Boys" + "Number of Girls") as this would only range from 0 to 1. Other issues found are the "Number of Pupils" not equating to the sum of "Number of Boys" + "Number of Girls".

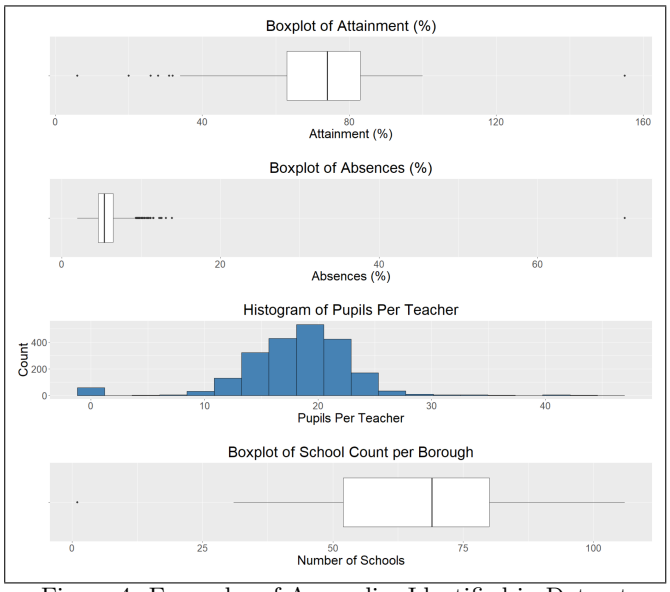

Figure 4: Examples of Anomalies Identified in Dataset

2.3 Data Imputation

Now that missing data, anomalies and other data quality issues have been identified, these can be addressed using different techniques. In our analysis, we have decided to replace all extreme outliers with NA so that they are no different to being missing information. Additionally, any variables with issues have been recalculated where possible, such as the "Boy Girl Ratio" and the "Number of Pupils.

Regarding dealing with the missing information, due to the dataset being categorised as MAR, we have decided that the appropriate method to apply data imputation is using Predictive Mean Matching on Numeric variables and Multinomial Regression on the Categorical Variables, leveraging the MICE package in R. The aim of these techniques is to estimate the missing information using the data we have already captured in the other observations.

City of London borough has been excluded from this reports analysis as this borough only consisted of one school, while all other boroughs have an average of 68 schools, and therefore this school is treated as an outlier.

3 Spatial Data Analysis

Now that our dataset has been imputed, we will move on to analysing the spatial information such as "Longitude" and "Latitude" attributes of each school in London. In this section we will also briefly explore the variable "Ofsted Rating" which is a useful indicator of overall school performance.

3.1 Longitude v Latitude plots

Each school has an associated longitude and latitude attribute which represents their geographic location. Using this it is relatively straightforward to make a geospatial plot as demonstrated in **figure 5**. We can also show other variables, such as "Ofsted Rating", by using different colours for each school.

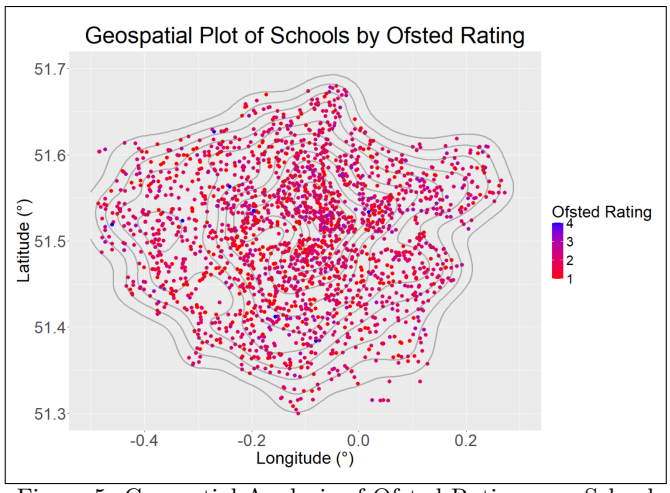

Figure 5: Geospatial Analysis of Ofsted Ratings per School

Unfortunately, this geospatial plot can be quite difficult to identify patterns due the level of granularity being used. It would be easier to understand and identify patterns if the data can be aggregated by categories.

3.2 Borough Maps

To mitigate against the issue of high granularity in longitude versus latitude plots, we will utilise another attribute in the dataset, "Borough", which also represents geospatial information, as it represents the region of London each school is located. Shape files of London boroughs have been utilised to aid us in creating these visuals.

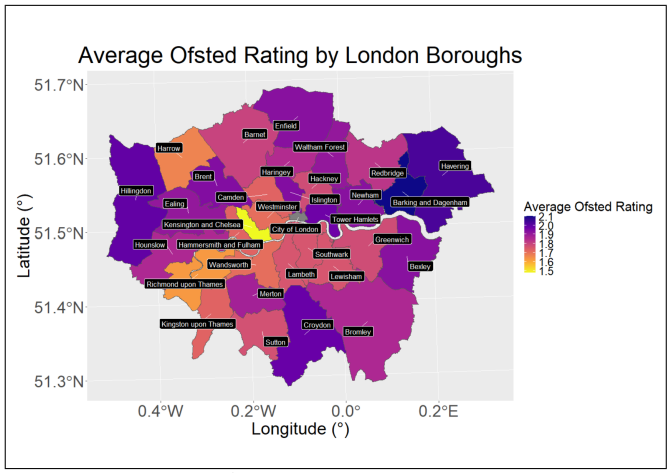

Figure 6: Geospatial Analysis of Mean Borough Ofsted Ratings

In **figure 6** we have a visual showing the boroughs of London and a colour scheme representing the average ofsted ratings.

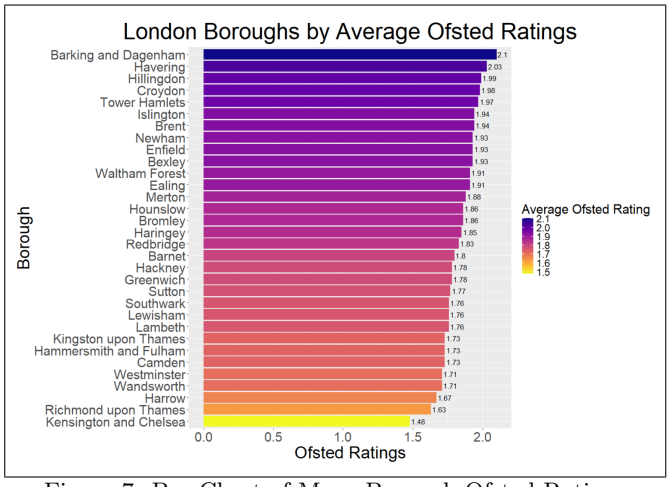

Figure 7: Bar Chart of Mean Borough Ofsted Ratings

We also have a bar chart in **figure 7** showing these average ofsted ratings in descending order. We can see that Barking and Dagenham, Havering and Hillingdon have the three worst average ratings of 2.1, 2.03 and 1.99 respectively. On the other hand, Kensington and Chelsea, Richmond upon Thames and Harrow have the three best ratings in London with 1.48, 1.63 and 1.67 respectively.

4 Analysis of Clusters

In this section, we will briefly explore the numeric attributes and find any patterns that could indicate clustering in the data. We will also attempt to see whether any clustering behaviours can be linked to other variables present in the dataset, such as any categorical attributes.

After computing correlation plots of 12 numeric attributes against each other, one of the pair of variables which stands out the most is "Crime Score" against "Number of Pupils". We can see the density plot for this pair of variables in **figure 8** which shows two clusters that are close proximity.

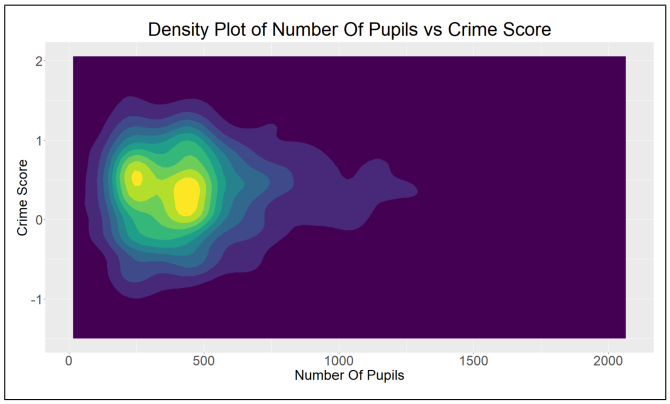

Figure 8: Density Plot of Number of Pupils against Crime Score

It would be useful to find other variables that may be related to the clustering behaviour we see in this density plot. A scatter plot of "Crime Score" against "Number of Pupils" is shown in **figure 9**, along with colour representing the attribute "School Type". In this plot we can see the centroid of each school type also plots to help us visually see how these different categories are positioned on average against crime and number of pupils.

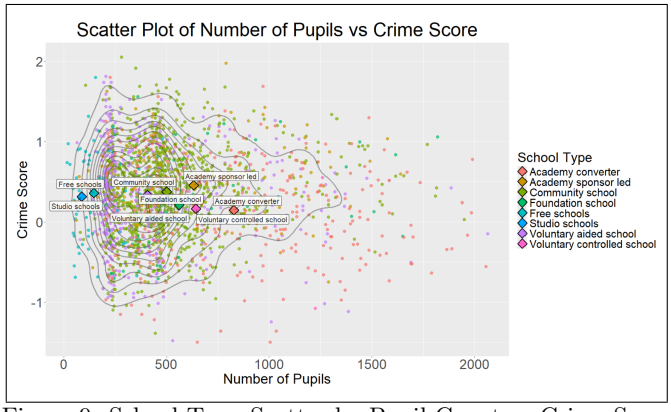

Figure 9: School Type Scatter by Pupil Count vs Crime Score

We can see Free schools and Studio schools are on the lower end of the number of pupils count, followed by voluntary aided schools, which is the second most common type of school. These school types could be contributing to the first smaller cluster in the density plot. On the other hand, Community schools, which is the most common type of school, appears to be mostly centered around where the second larger cluster is located in the density plot. Academy converters, the third most common type of school, appears to have the largest average number of pupils and seems to be scattered across the right section of the plot.

We have also plotted the same scatter plot but with the "School Level" attribute as the colour scheme instead in **figure 10**. Here we can see a large difference in the position of the centroids, where Primary schools generally have a lower number of pupils than secondary schools.

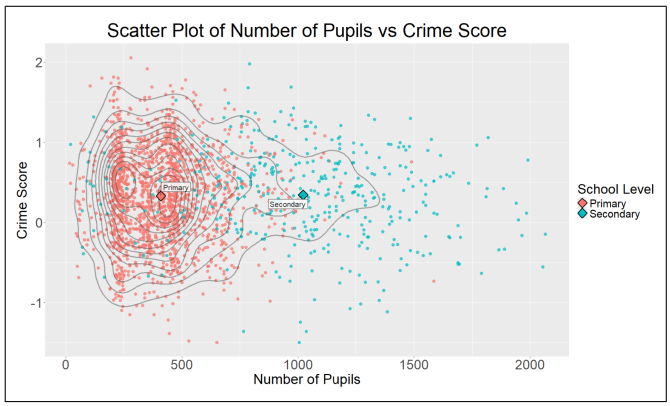

Figure 10: School Level Scatter by Pupil Count vs Crime Score

5 Dimension Reduction

In this section we will use dimension reduction techniques on our dataset to see whether this assists us in finding clusters. Dimension reduction is an important method as it allows us to simplify the dataset whilst retaining majority of the information.

5.1 Principal Component Analysis

One prominent method of dimension reduction is Principal Component Analysis (PCA), which we will use in our analysis. As we apply this technique to our 29-variable data set, we compute the principal components (which are linear combinations of these variables) in descending order of the proportion of the dataset variance they retain.

In **figure 11** we plot a scatter graph of the first principal component against the second principal component which overall capture 51% of he datasets variance.

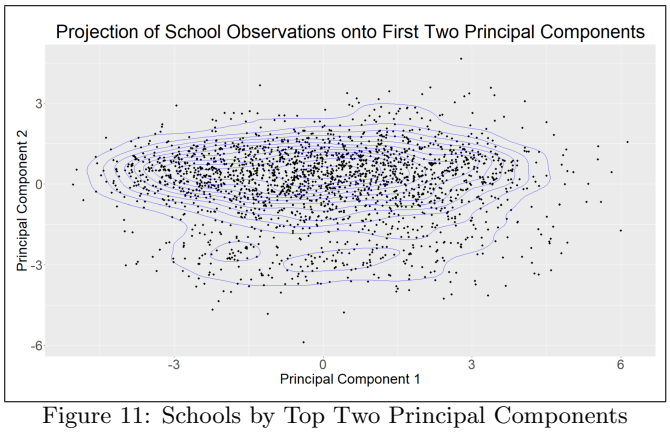

Figure 11: Schools by Top Two Principal Components

In **figure 12**, we can see that the first three principal components results in 35%, 51% and 60% of the variance respectively. We would need 6 principal components to retain more than 80% of the dataset variance.

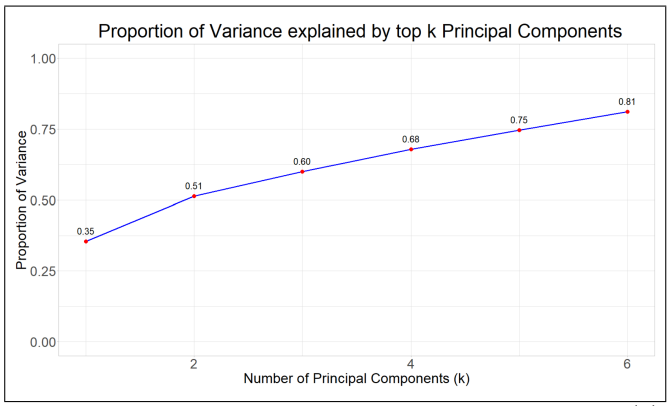

Figure 12: Variance by number of Principal Components (k)

5.2 K-Means Clustering

K-Means clustering requires specifying the number of clusters, K, in advance. To help determine the best K, we use techniques like the silhouette score, which compares the average distance within clusters to the distance to the nearest cluster. Scores range from -1 to 1, where 1 indicates strong clustering, 0 is weak, and negative suggest misclassifications.

We applied this to our dataset using different numbers of principal components and found that with two components, the best K is 3, giving an average silhouette score of 0.41. This suggests a moderate clustering result, as scores above 0.5 are considered good, and above 0.7 strong.

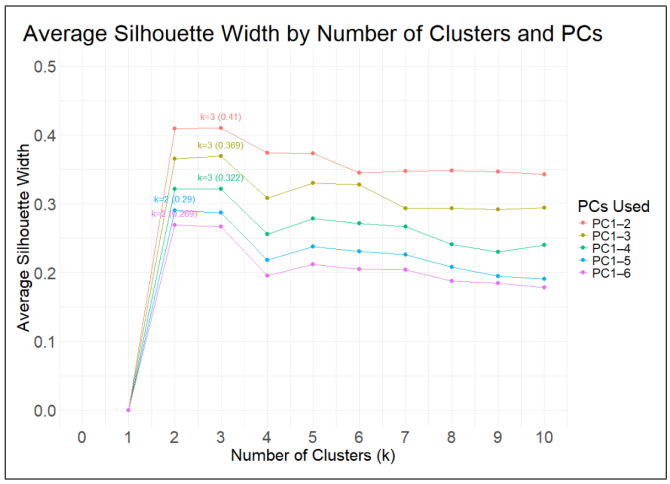

Figure 13: Silhouette by Number of PCs and Clusters

We can see in **figure 13** as we increase the number of principal components, the average silhouette score begins to drop and eventually when we introduce the fifth principal component, the optimal number of clusters drops to 2.

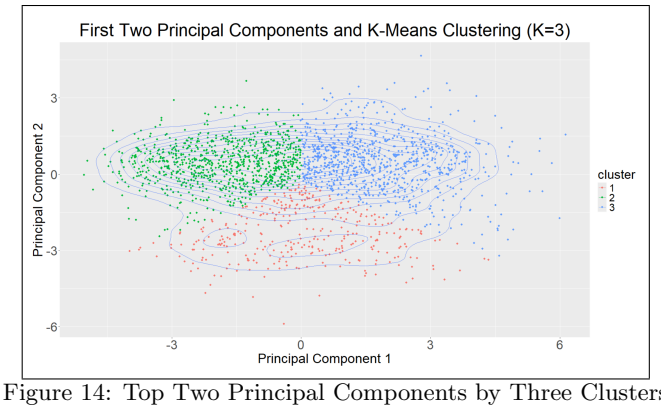

Figure 14: Top Two Principal Components by Three Clusters

figure 14 shows a scatter plot of the data after reducing the dimensions to two principal components, we using K=3 for our K-Means Clustering to visually observe the clusters. We can see there is weak structure present but the boundaries are not sharp and therefore the clusters could exist but are quite weak and may be overlapping.

In **figure 15** and In **figure 16** we have the same PCA plot with the attributes of "School Level" and the "School Type" to see whether these contribute to the clustering behaviour.

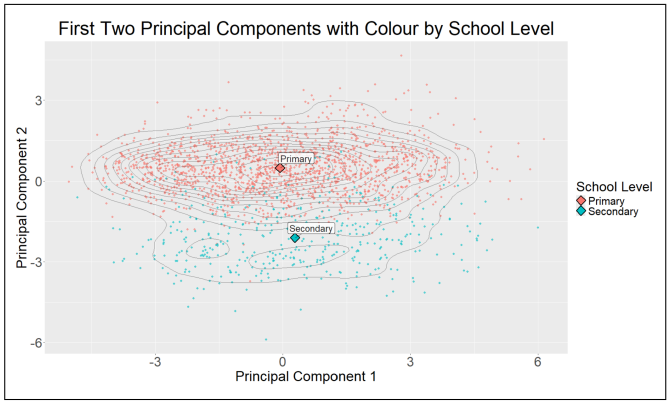

Figure 15: Top Two Principal Components by School Level

In **figure 15** we can see after using PCA, that the attribute "School Level" is likely one of the main contributors to the segmentation of cluster 1 and may be correlated to the numeric variables that contribute to the second principal component.

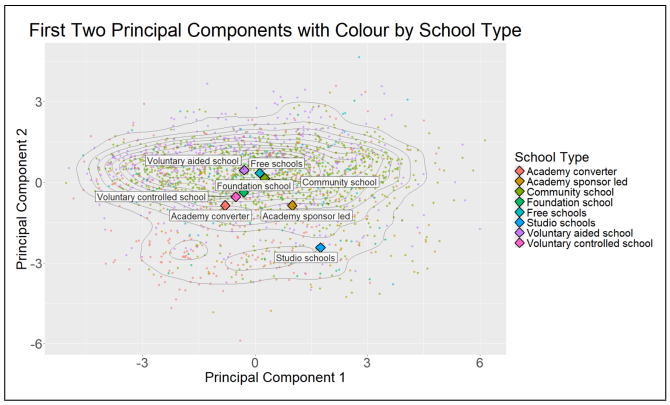

Figure 16: Top Two Principal Components by School Type

We can also see in **figure 16** "School Type" may also be another variable influencing where schools are positioned in the PCA plot but not as strongly as the School Level variable. It is likely other numeric variables also influence where a school is located in the PCA plot, such as "Income Score", "Employment Score" and "Health Score" which we have found to have the top three scores in the first principal component.

Overall, whilst this report followed a brief exploratory analysis, we conclude there are some relationships between variables and clustering behaviour present in the provided dataset, These patterns and trends appear to be complex and would require further analysis to gain meaningful insights into what influences performance across all the schools in London.