Exploratory Data Analysis on Indian School Dataset: Project Report

Kartik Soni, Manya Verma, Yash Gupta

November 15, 2023

1 Introduction

1.1 Motivation

India is a developing nation, and to transition into a developed country, the foundation must be built on education. We must systematically address the educational system's challenges as we strive for progress. One crucial aspect is the infrastructure supporting our schools, a cornerstone for providing quality education to every child.

The motivation behind this project lies in the recognition that educational infrastructure plays a pivotal role in shaping the learning environment. Disparities in infrastructure can lead to unequal educational opportunities, hindering the overall development of our nation. By delving into the intricacies of this infrastructure, we aim to uncover patterns, identify disparities, and provide valuable insights to inform strategic decision-making.

1.2 Problem Statement

The current state of educational infrastructure in India exhibits disparities that pose significant challenges to the goal of providing quality education for all. Variations in the availability of essential facilities, such as classrooms, sanitation, electricity, and furniture, among schools across different states can lead to unequal learning opportunities. Addressing these disparities is crucial for fostering an inclusive and effective education system. The principal objective of this analysis encompassed identifying discernible patterns and trends in the availability of infrastructure, facilities, etc., among educational institutions, geographically wise (From District level as well as state level).

2 About the Data-set

The data set was sourced from the **NDAP** Portal and uploaded by the Unified District Information System for Education (**UDISE**). It comprises **1,551,000** records, including **50** variables such as institution ID, geographic location, counts of classrooms, presence of amenities such as toilets, water supply, electricity, computing resources, solar panels, printers, libraries, playgrounds, internet connectivity, projectors, furniture, and medical facilities.

Existing Analyses and how this differs? In exploring our dataset is noteworthy to highlight that the literature related to our specific dataset notably limited. The absence of a substantial body of previous work adds a layer of originality to our research, allowing us to contribute novel findings and perspectives to the existing knowledge landscape.

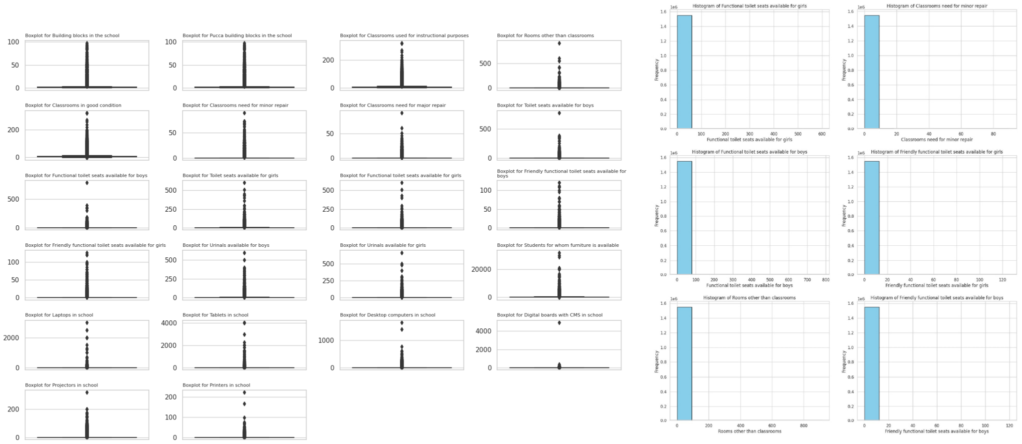
3 Data Pre-processing

1. **Examining Redundancy and NULL values:** As the data set was occupied from the NDAP portal and already preprocessed, it had no duplicate values and only a few NaN values, so we deleted them.

2. **Feature Selection:** Certain features, such as Pucca building blocks, no of boundary walls, availability of handrails, etc., were removed from the data set due to their lack of relevance and objective of our project. Evaluation of value counts across all columns aimed to discern the uniqueness and variability of each column's values. Columns had (Value Count plot) almost 98% values of the same metric, so we removed them.

3. **Outlier Mitigation:** Identification of outliers was accomplished through Box plots, followed by calculating the percentage of outliers. There were two types of outliers in our dataset. First, having more than 10% values, those were replaced with the 99th percentile value, preserving the dataset's integrity, and some others were outliers but were very low in percentage, so we removed them.

4. **Feature Engineering:** While digging into the data, we noticed something interesting. There weren't many traditional computers in many schools, but plenty of tablets and laptops. This made sense, especially in smaller villages where full-fledged computers might be expensive. Instead, more affordable educational tablets seemed to



(a) Outlier Detection using Box-Plots

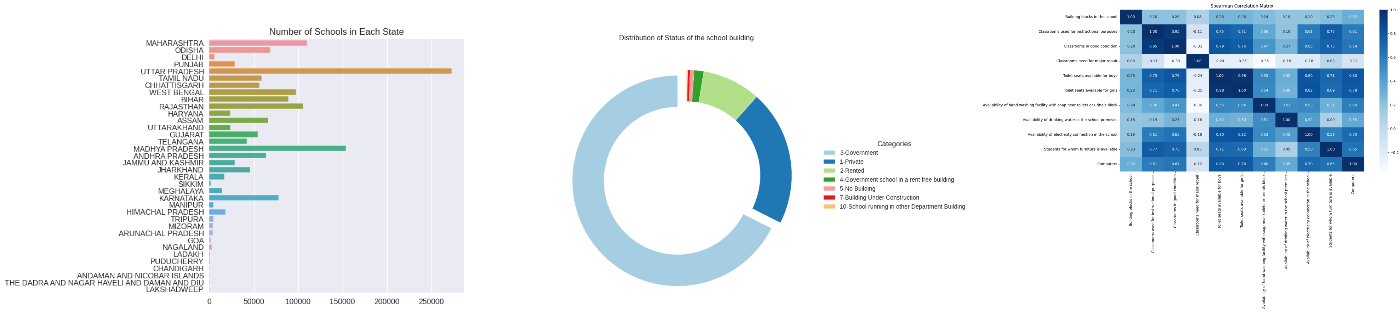
(b) Heavily Biased Values

Figure 1: Data Pre-processing with the aid of visualization

be more common. To make sense of this, we decided to group various technology-related aspects like 'laptops,' 'tablets,' 'computers,' and 'learning management systems' into a single category called 'computers.' Categorical variables within the data are transformed into a numeric format to facilitate their integration into a generalized model.

4 Exploratory Data Analysis

Through EDA we got some basic insights about our data that how the 1.5 Million schools are distributed across India and in which states. Then we saw almost 70% of the schools are under government, 15% private and rest other . Apart from this we plotted the initial correlationplot with the help of corr() function and got to know that some of our columns are highly correlated so we dropped such columns with more than 60% correlation. The assessment of columns was conducted through meticulous scrutiny, employing multiple methods:



(a) State-wise Distribution of Schools

(b) Category Distribution of Schools

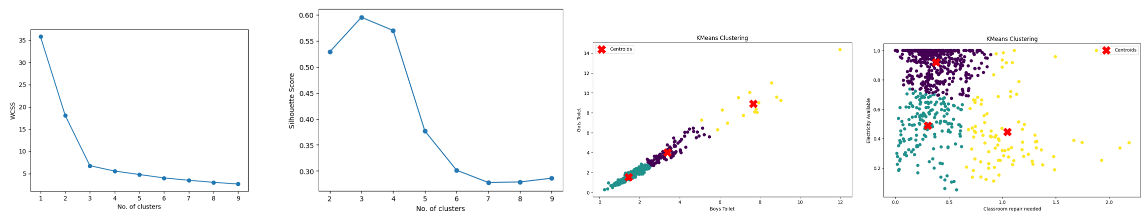
(c) Heat-map of Correlation

5 Clustering Analysis

5.1 K-Means Clustering

The data set encompasses data from 36 States and Union Territories of India. The 'computers' feature ranges between 5 and 98.5, while other features span a more comprehensive coverage, mostly above 40 to 100, occasionally surpassing 100. Moreover, distinct disparities exist in the mean values of different features. This variability in ranges among features might lead to bias, especially in clustering algorithms utilizing Euclidean Distance for calculations. To mitigate this, we intend to standardize all features between 0 and 1 using MinMaxScaling().

Determining Number of Clusters: The Elbow Method is employed to choose the suitable number of clusters for the K-Means Clustering algorithm. It identifies the point on the graph where the Within Cluster Sum of Squares (WCSS)



(a) Elbow Curve (b) Silhouette Score Curve (c) K-Means Clustering (d) K-Means Clustering

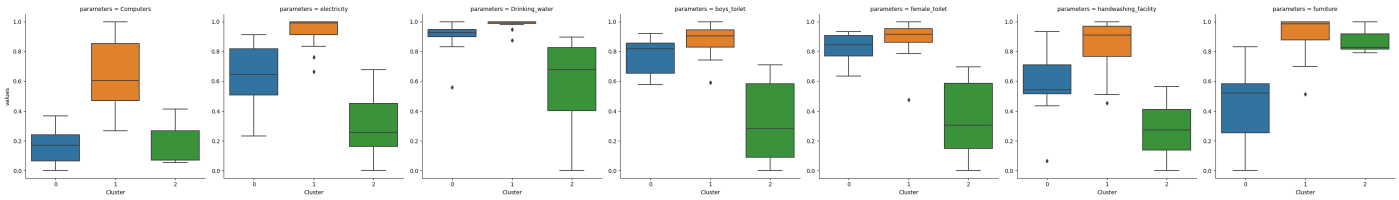


Figure 4: Box-Plot for clusters

exhibits a slower rate of decrease, resembling an elbow bend. Analyzing the WCSS values and the Scree Plot suggests that 2 or 3 clusters might be optimal.

Upon analyzing the WCSS values and the Silhouette score, it's evident that 2 or 3 clusters could be the optimal choice. To select the precise number of clusters between 2 and 3, we'll utilize specific metrics:

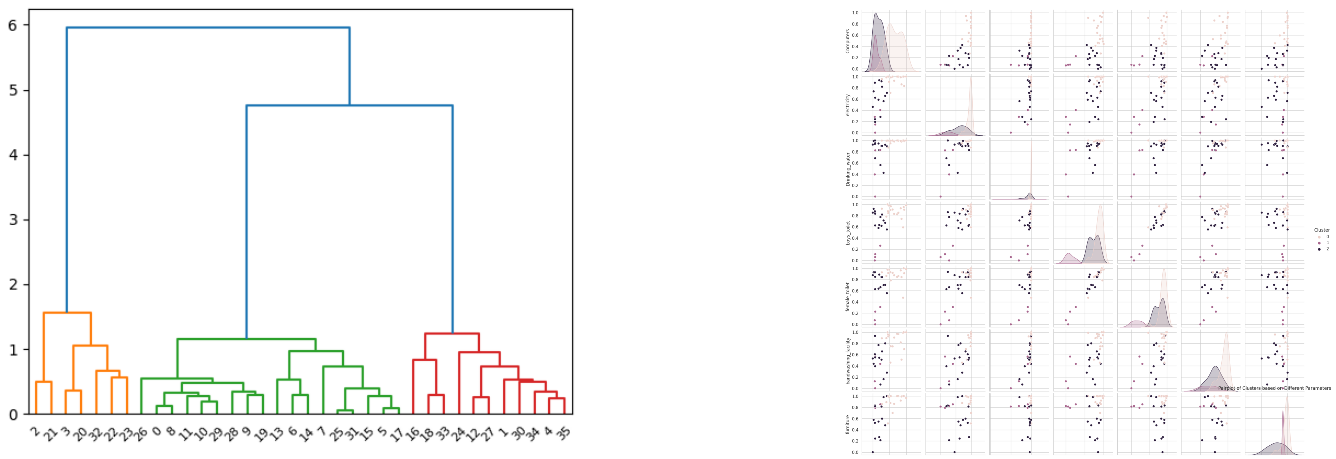
1. Silhouette Score: Ranges from -1 to 1. Higher values signify better-formed clusters. Closer to 1 indicates well-separated clusters, 0 suggests points on cluster borders, and negative values denote misclassified points.

2. Calinski-Harabasz Index: Measures cluster density. A higher score indicates denser clusters, portraying better clustering. It starts from 0 with no upper limit, showing cluster density.

5.2 Hierarchical Clustering

We employed the Agglomerative Clustering technique, a form of Hierarchical Clustering, for further analysis. This approach is characterized by a bottom-up strategy, initially treating each data point as an individual cluster and then iteratively merging the closest points based on a designated distance metric until a single cluster is formed.

The visualization of Hierarchical Clustering through a dendrogram indicated the optimal number of clusters to be 3.



(a) Dendogram (b) Pair-Plots of Clustered Variables

Cluster 1 showcases the broadest range of infrastructure components, especially in computers, electricity, water, and toilets, signifying it as the cluster with the most comprehensive infrastructure. Cluster 0 closely follows, while Cluster 2 presents the narrowest spectrum of infrastructure attributes.

Enrollment ratios vary significantly across clusters, with Cluster 2 having the highest ratio and Clusters 0 and 1 showing relatively similar percentages. However, Cluster 1 exhibits more variability in its higher range of values.

Consequently, we categorized the clusters as follows: 0: Good Infrastructure, Basic Facilities, Limited Resources 1: Best Infrastructure, Advanced Facilities, Adequate Resources 2: Inadequate Infrastructure, Intermediate Facilities, Abundant Resources

The visualization also depicts the states colour-coded based on the clusters they are associated with.

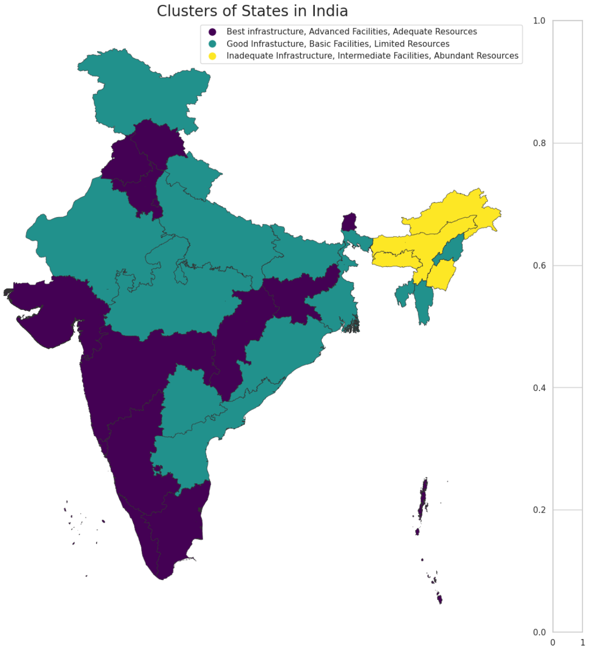
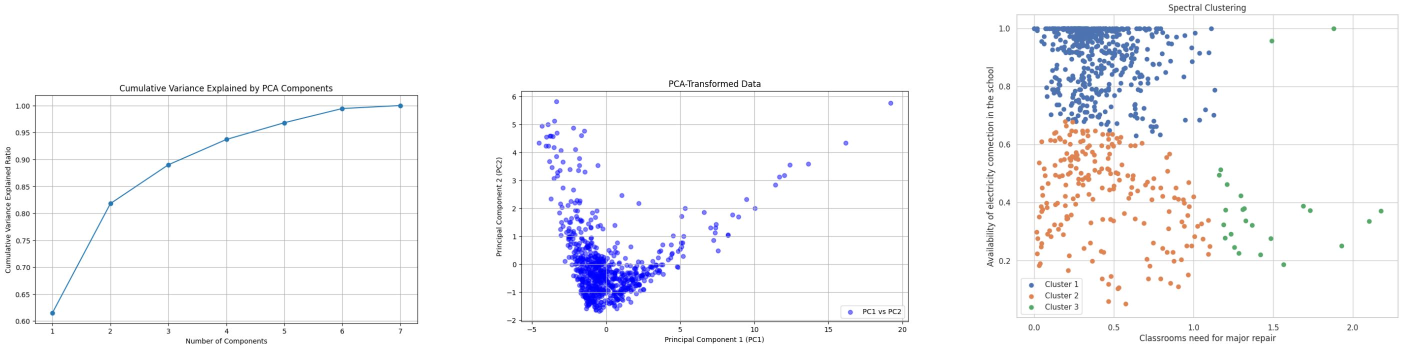

Clusters of States in India
● Best infrastructure, Advanced Facilities, Adequate Resources
● Good Infrastucture, Basic Facilities, Limited Resources
● Inadequate Infrastructure, Intermediate Facilities, Abundant Resources
1.0
0.8
0.6
0.4
0.2
0.0
0
1

Figure 6: Colour-Coded States

6 Dimension Reduction

6.1 Principle Component Analysis

As we explored educational infrastructure, PCA aided in uncovering the key factors influencing the overall scenario. For instance, it helped us understand the interplay between variables like 'Computers,' 'Electricity,' 'Drinking Water,' and more. Eventually using Dimensionality reduction we can see from the plots that our almost 85% of the data variance can be explained using 3 principal components only and that in itself is a major feat as now we know that for further studies it's easier to just consider those 3 components and go with it.



(a) Cumulative Variance after PCA

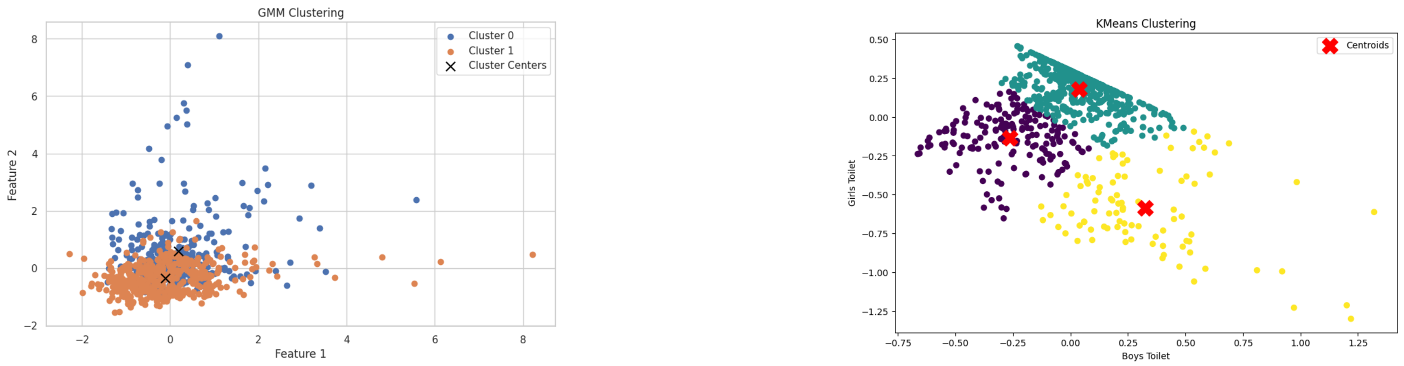
(b) After PCA 2 components

(c) Spectral Plot

Figure 7: PCA Transformed Data

6.2 Multi-Dimensional Scaling

After applying PCA, we can see decent transformation, but our data is non-linear, so we apply MDS to visualise non-linear data better. MDS allowed us to visualize the underlying structure of our data in a way that maintains the original pairwise relationships. This was particularly valuable as we sought a nuanced understanding of how different states and clusters relate to each other in the educational landscape. Through the spectral clustering we can see that it clusters it into fairly good clusters and we can understand that our dataset being non linear spectral clustering handled it well.



(a) GMM Clustering

(b) K-Means clustering after MDS

Figure 8: Clustering Comparison

7 Conclusion and Self-reflection

Our journey through this project has been both challenging and enlightening. One of the initial hurdles we faced was the overwhelming nature of the high-dimensional dataset. The high dimension of the data made us question the feasibility of obtaining meaningful insights within a reasonable timeframe. It was at this juncture that we contemplated changing the dataset, seeking a more manageable route.

However, instead of taking the easy road, we decided to delve into the literature to explore techniques for handling large-dimensional data. This exploration led us to discover powerful tools such as PCA and MDS. These techniques addressed the challenge of dimensionality and offered a deeper understanding of the underlying structures within the data.

We found intriguing insights as we progressed from fundamental analyses to exploring each dimension meticulously. The visualization of our findings on the Indian map was a pivotal moment. Notably, the cluster analysis illuminated a distinct segment in northeastern India, characterized by abundant resources but lacking in school infrastructure. This revelation prompted a targeted recommendation to focus on improving infrastructure facilities in this region. We analysed the data and found North eastern states even though having the resources the infrastructure is the weak point so government should focus over there.

Navigating the challenges posed by the dataset's complexity became an invaluable learning experience. Moreover, the process reinforced our commitment to thorough exploration, emphasizing the need to understand each dimension before drawing conclusions.

References

1. Unified District Information System for Education (2018-19). *Facility-School infrastructure and facility related data.* Retrieved from `https://ndap.niti.gov.in/dataset/7016`

2. Varthana (2023). *How School Infrastructure Impacts Your Child's Learning Journey.* Retrieved from `https://varthana.com/school/how-school-infrastructure-impacts-your-childs-learning-journey/#:~:text=According%20to%20the%20District%20Information,and%2092.3%25%20have%20electricity%20connections.`

3. Dr.Ambika N (2022). *A Study on School Infrastructure in India* . Retrieved from `https://ijrpr.com/uploads/V3ISSUE12/IJRPR8657.pdf`

4. Hanagodimath S. V. (2011). *Education Infrastructure and Education Status Indices in India: An Inter-State Analysis.* Retrieved from `https://www.ijrssh.com/admin/upload/15%20S%20V%20Hanagodimath%2001214.pdf`