Project Report

Topic: Advanced-Data Analysis and Visualization Using Seaborn

IE6600 Computation and Visualization in Analytics

Group 18
Nishanth Manoharan (002321972)
Arun Solaiappan Valiappan (002053328)
Preeti Purnimaa Kannan (002374503)

Table of Contents

SUMMARY

The analysis of the *Public-School Characteristics 2022-23* dataset provides a comprehensive examination of public elementary and secondary schools across the United States, focusing on key attributes such as student enrollment, school locations, staffing, and administrative details. This dataset, compiled by the National Center for Education Statistics (NCES) through the *Common Core of Data (CCD)* program, serves as a vital resource for educational research and policymaking. The study began with data acquisition and inspection, ensuring a clear understanding of the dataset's structure, completeness, and potential inconsistencies. Through detailed data cleaning, missing values were handled, categorical variables were encoded, and geographical attributes were standardized to ensure accuracy. Exploratory Data Analysis (EDA) was then conducted to visualize trends, identify relationships among key variables, and detect potential outliers. Various statistical methods and data visualization techniques were employed to explore the distribution of student-teacher ratios, enrollment trends, and the demographic composition of schools.

Following the initial data exploration, advanced analytical techniques were applied to extract deeper insights. A feature importance analysis using a *Random Forest Regressor* identified the most influential factors affecting total student enrollment, highlighting the impact of school type, grade span, and location-based attributes. Clustering methods, such as *K-Means clustering*, were implemented to categorize schools based on enrollment size and student-teacher ratios, providing a structured view of school characteristics across different regions. Additionally, predictive modeling using *XGBoost* was performed to estimate student enrollment, achieving a high level of accuracy with an R^2 *score of 97.8%*. These findings offer valuable insights into the distribution of educational resources, disparities among school districts, and key factors influencing school capacity. The study's conclusions can guide policymakers, educators, and researchers in making data-driven decisions to optimize school resource allocation, enhance educational planning, and address disparities in public education across the country.

INTRODUCTION

The Public-School Characteristics 2022-23 dataset is an extensive collection of administrative, demographic, and geographic information on public elementary and secondary schools across the United States. Developed under the Common Core of Data (CCD) program by the National Center for Education Statistics (NCES) and updated annually through the Education Demographic and Geographic Estimates (EDGE) program, this dataset provides critical insights into school locations, student enrollment, staffing, and operational characteristics. The data, reported by state education agencies, includes essential details such as school names, district affiliations, grade spans, student-teacher ratios, and racial and socioeconomic distributions of students. This dataset is widely used by policymakers, researchers, and educators to analyze school performance, identify trends in student demographics, and allocate resources effectively. However, the dataset contains missing values, inconsistencies, and potential data quality issues, which require careful preprocessing before meaningful insights can be derived.

The primary objective of this project is to perform a comprehensive analysis of the dataset, beginning with data cleaning and preparation to handle missing or inaccurate information. Exploratory Data Analysis (EDA) will be conducted to identify trends, distributions, and correlations among key variables, offering insights into factors such as school size, location-based disparities, and demographic influences on student enrollment. Additionally, advanced analytical techniques, including clustering and predictive modeling using machine learning algorithms, will be applied to uncover deeper patterns and relationships within the data. By leveraging statistical and machine learning approaches, this study aims to assess the factors influencing school enrollment, highlight disparities in educational resources, and provide data-driven recommendations for improving educational planning and policy-making. The findings from this analysis can support decision-makers in addressing educational inequalities, optimizing resource distribution, and enhancing school administration strategies.

DATA ACQUISTION AND INSPECTION

The dataset was loaded and inspected for structure, data types, and completeness. The dataset contains **101,390 rows** and **77 columns**, covering various attributes of public schools, including:

- **Geographical details**: Latitude (LATCOD), Longitude (LONCOD), State (STABR)
- **School information**: School Name (SCH_NAME), District Name (LEA_NAME), Charter Status (CHARTER_TEXT), School Type
- **Demographics**: Student Enrollment (TOTAL), Grade-Level Data (PK-G12), Racial/Ethnic Distributions
- **Financial & Administrative data**: Student-Teacher Ratio (STUTERATIO), Free and Reduced-Price Lunch (TOTFRL)

Initial dataset checks included:

- **No duplicate rows** were found.
- The dataset contains various **missing and invalid values**, indicated by:
  - **"-1" or "M"** (Missing data)
  - **"-2" or "N"** (Not applicable)
  - **"-9"** (Did not meet NCES data quality standards)

	X	Y	OBJECTID	NCESSCH	SURVYEAR	STABR	LEAID	ST_LEAID	LEA_NAME	SCH_NAME	...	HIALF	HI	TRALM	TRALI
0	-86.206200	34.260200	1	10000500870	2022-2023	AL	100005	AL-101	Albertville City	Albertville Middle School	...	251.0	502.0	17.0	15.0
1	-86.204900	34.262200	2	10000500871	2022-2023	AL	100005	AL-101	Albertville City	Albertville High School	...	468.0	958.0	26.0	19.0
2	-86.220100	34.273300	3	10000500879	2022-2023	AL	100005	AL-101	Albertville City	Albertville Intermediate School	...	241.0	504.0	7.0	6.0
3	-86.221806	34.252700	4	10000500889	2022-2023	AL	100005	AL-101	Albertville City	Albertville Elementary School	...	236.0	497.0	11.0	16.0
4	-86.193300	34.289800	5	10000501616	2022-2023	AL	100005	AL-101	Albertville City	Albertville Kindergarten and PreK	...	152.0	319.0	4.0	4.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
101385	-64.917602	18.341950	101386	780003000024	2022-2023	VI	7800030	VI-001	Saint Thomas - Saint John School District	LOCKHART ELEMENTARY SCHOOL	...	30.0	63.0	1.0	2.0
101386	-64.952483	18.338742	101387	780003000026	2022-2023	VI	7800030	VI-001	Saint Thomas - Saint John School District	ULLA F MULLER ELEMENTARY SCHOOL	...	27.0	54.0	2.0	0.0
101387	-64.899024	18.354782	101388	780003000027	2022-2023	VI	7800030	VI-001	Saint Thomas - Saint John School District	YVONNE BOWSKY ELEMENTARY SCHOOL	...	22.0	59.0	0.0	1.0
101388	-64.945940	18.336658	101389	780003000033	2022-2023	VI	7800030	VI-001	Saint Thomas - Saint John School District	CANCRYN JUNIOR HIGH SCHOOL	...	62.0	136.0	0.0	1.0
101389	-64.890311	18.318230	101390	780003000034	2022-2023	VI	7800030	VI-001	Saint Thomas - Saint John School District	BERTHA BOSCHULTE JUNIOR HIGH	...	21.0	48.0	0.0	0.0

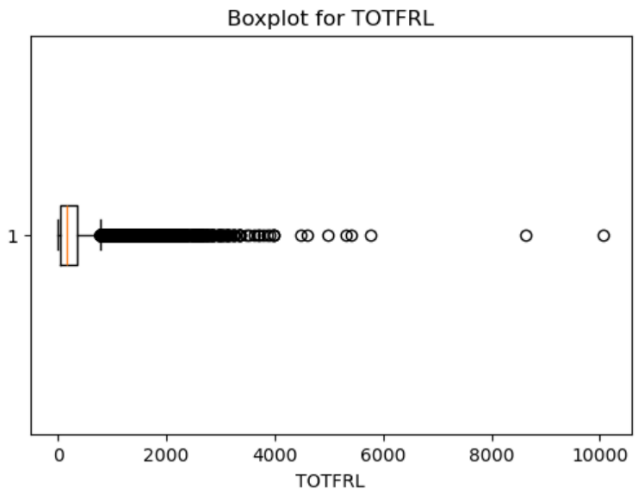
101390 rows × 77 columns

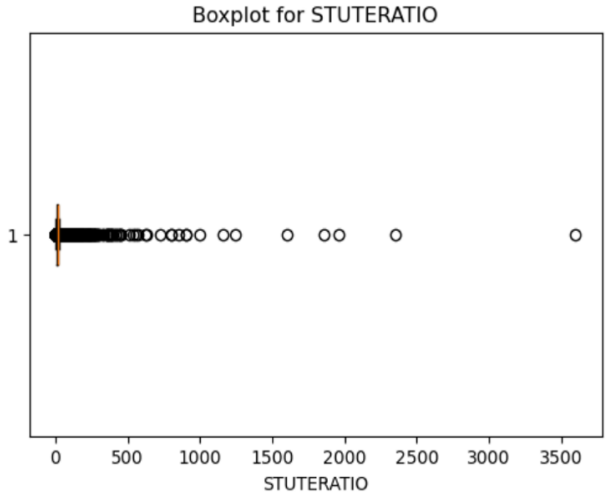
A table summarizing missing values in various columns of a dataset, showing the column names, the count of missing values, and their percentage. Some columns, like **LSTREET2** and **G13**, have nearly 100% missing values, while others have minimal data gaps.

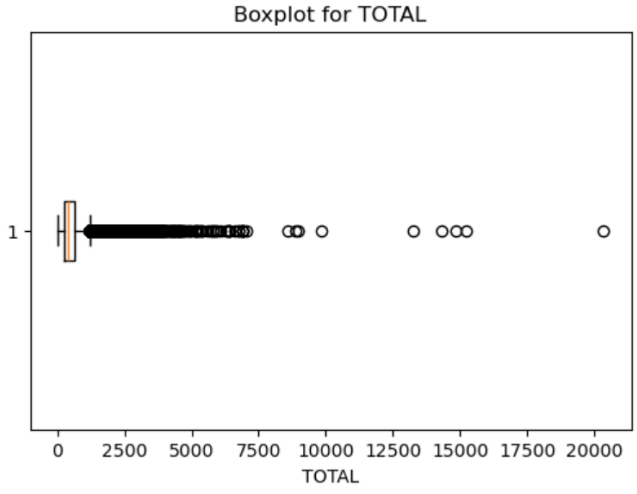
	Column Name	Missing Values	Percentage
LSTREET1	LSTREET1	1	0.000986
LSTREET2	LSTREET2	100818	99.435842
PK	PK	68998	68.052076
KG	KG	47329	46.680146
G01	G01	46978	46.333958
G02	G02	46921	46.277739
G03	G03	46931	46.287602
G04	G04	47132	46.485847
G05	G05	48376	47.712792
G06	G06	63367	62.498274
G07	G07	68166	67.231482
G08	G08	67898	66.967157
G09	G09	73289	72.284249
G10	G10	73501	72.493343
G11	G11	73502	72.494329
G12	G12	73574	72.565342
G13	G13	101257	99.868823
UG	UG	93501	92.219154
AE	AE	101207	99.819509
TOTMENROL	TOTMENROL	2480	2.446001
TOTFENROL	TOTFENROL	2480	2.446001
TOTAL	TOTAL	1671	1.648092
MEMBER	MEMBER	1671	1.648092
FTE	FTE	3853	3.800178
STUTERATIO	STUTERATIO	1814	1.789131
AMALM	AMALM	2581	2.545616
AMALF	AMALF	2579	2.543643
AM	AM	2533	2.498274
ASALM	ASALM	2492	2.457836
ASALF	ASALF	2490	2.455863
AS	AS	2484	2.449946
BLALM	BLALM	2494	2.459809
BLALF	BLALF	2497	2.462768
BL	BL	2487	2.452905
HPALM	HPALM	2608	2.572246
HPALF	HPALF	2607	2.571259
HP	HP	2561	2.525890
HIALM	HIALM	2481	2.446987
HIALF	HIALF	2480	2.446001
HI	HI	2480	2.446001
TRALM	TRALM	2487	2.452905
TRALF	TRALF	2485	2.450932
TR	TR	2484	2.449946
WHALM	WHALM	2481	2.446987
WHALF	WHALF	2481	2.446987
WH	WH	2480	2.446001

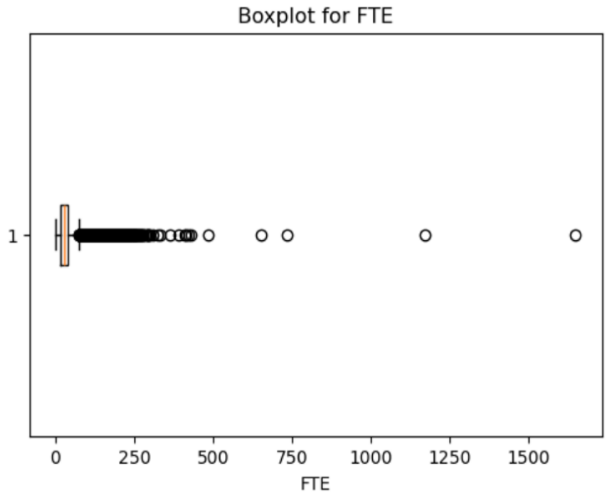
The box plots in the attached images visually confirm the presence of outliers in the **TOTFRL, TOTAL, STUTERATIO, and FTE** columns.

- **TOTFRL (Total Free and Reduced Lunch Students)** and **TOTAL (Total Enrollment)**: These box plots show a concentration of values close to zero, with numerous extreme outliers extending beyond 10,000 and even 20,000. This suggests that while most schools have manageable enrollments, some records contain significantly larger numbers, which could indicate data anomalies or special cases (e.g., very large schools or reporting errors).
- **STUTERATIO (Student-Teacher Ratio)**: This box plot exhibits many outliers beyond a reasonable range (~10-30). Some schools have extreme ratios, possibly due to incorrect data entry or unique circumstances (e.g., virtual schools or schools with part-time faculty affecting calculations).
- **FTE (Full-Time Equivalent Teachers)**: The distribution is highly skewed, with many small values and a long tail of outliers beyond 500-1500. Schools reporting abnormally high FTE values may need further verification.









The image presents a **data consistency check** table, validating key dataset attributes:

1. **Number of Unique State Codes**:
   - The dataset contains **56 unique state codes**, which may suggest invalid or additional entries beyond the **50 US states + Washington D.C. + territories**.
   - Further verification is needed to ensure correctness.
2. **Valid Zip Code Lengths**:
   - Expected **5-digit** zip codes are dominant (**93,357 entries**).
   - However, there are **7,153** zip codes with **4 digits** and **880** with **3 digits**, which are likely incomplete or incorrect.
   - Cleaning or filling missing digits may be required.
3. **Enrollment Sum Consistency**:
   - The total enrollment count across all schools is **44,040**.
   - This should match the sum of individual grade enrollments (**G01-G12**), ensuring no missing or misreported values.
4. **Latitude & Longitude Validity**:
   - Both **Latitude Validity** and **Longitude Validity** are marked as **False**, meaning some geographic coordinates are **out of valid US bounds** or missing.
   - Valid latitudes for the US should be between **24.396308° and 49.384358°**, and longitudes between **-125.000000° and -66.934570°**.
   - These errors might require correction by identifying and filtering out incorrect records.

	Check	Result
0	Number of Unique State Codes	56
1	Valid Zip Code Lengths	{5: 93357, 4: 7153, 3: 880}
2	Enrollment Sum Consistency	44040
3	Latitude Validity	False
4	Longitude Validity	False

DATA CLEANING AND PREPARATION

Handling Missing Data:

- Some columns, such as **LSTREET2, G13, and AE**, had over **99% missing values** and were **dropped** to improve data quality.
- **Enrollment data (PK-G12)** showed **40-70% missing values**, requiring imputation.
- **Numeric missing values** were filled using the **median** to prevent bias from extreme values.
- **Categorical missing values** were replaced with the **mode** (most frequently occurring value).

Handling Invalid Data:

- **Latitude (X) & Longitude (Y)** were checked against valid **U.S. geographical ranges**:
  - **Valid latitude: 24 to 50**
  - **Valid longitude: -125 to -65**
  - Records outside this range were filtered out.
- **Zip codes** were standardized to a **5-digit format** to ensure consistency.
- The **TOTAL enrollment** column was recomputed as the **sum of individual grade-level enrollments** to maintain consistency.

Data Type Conversions:

- Some numeric columns stored as **object types** were converted into **proper numerical formats**.
- **Categorical features** were encoded using **Label Encoding (LabelEncoder)** to facilitate model training.

	X	Y	OBJECTID	NCESSCH	SURVYEAR	STABR	LEAID	ST_LEAID	LEA_NAME	SCH_NAME	...	HIALF	HI	TRALM	
0	0.666645	0.391394	0.00000	0.000000e+00	0	0	0.000000	67	450	2212	...	0.064641	0.072817	0.009918	0.
1	0.666667	0.391474	0.00001	1.724083e-12	0	0	0.000000	67	450	2209	...	0.120525	0.138961	0.015169	0
2	0.666404	0.391922	0.00002	1.551672e-11	0	0	0.000000	67	450	2210	...	0.062065	0.073107	0.004084	0
3	0.666374	0.391091	0.00003	3.275752e-11	0	0	0.000000	67	450	2208	...	0.060778	0.072092	0.006418	0
4	0.666869	0.392587	0.00004	1.286164e-09	0	0	0.000000	67	450	2211	...	0.039145	0.046272	0.002334	0.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
100396	0.619406	0.333837	0.99996	9.999990e-01	0	23	0.999999	1282	1216	5290	...	0.000000	0.000000	0.000000	0.
100397	0.295244	0.422911	0.99997	9.999991e-01	0	30	0.999999	1161	6894	33044	...	0.000000	0.000000	0.000000	0.
100398	0.554906	0.703204	0.99998	9.999993e-01	0	10	0.999999	1236	10324	51033	...	0.000000	0.000000	0.000000	0.
100399	0.134111	0.371734	0.99999	9.999995e-01	0	3	0.999999	1251	11380	55307	...	0.000000	0.000000	0.000000	0.
100400	0.505120	0.416564	1.00000	1.000000e+00	0	34	1.000000	1137	7963	39461	...	0.000000	0.000000	0.000000	0.

99604 rows × 59 columns

Key Observations from the Image:

- The **X (longitude) and Y (latitude) columns** contain decimal values, suggesting improper formatting.
- **Several columns (e.g., LEAID, SCH_NAME, etc.)** have floating-point values, which may require type conversion.
- The **last few rows contain values close to 1.000000**, possibly due to rounding errors, requiring normalization.

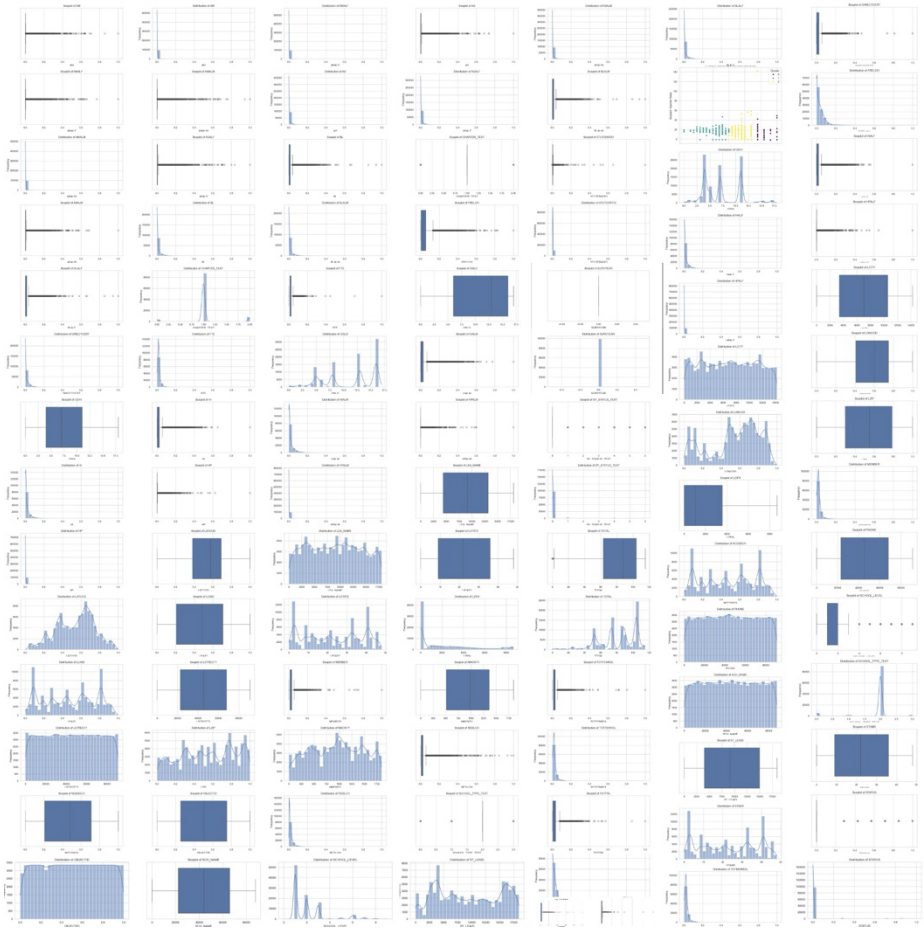
EXPLORATORY DATA ANALYSIS (EDA)

Data Distribution (Histograms):

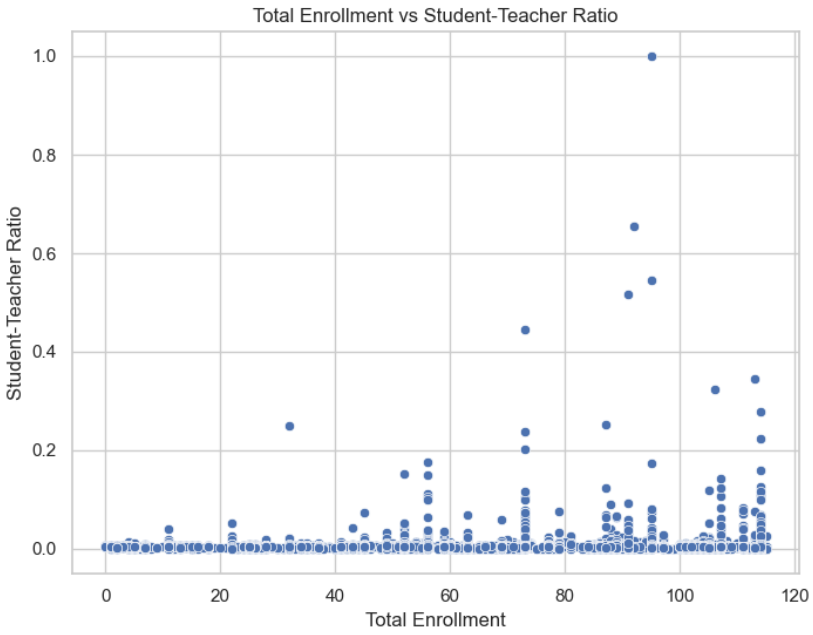
- **Normal vs. Skewed Distribution**: Some numerical columns might exhibit normal distribution (bell-shaped curves), while others could be **skewed** (right or left). This helps determine whether transformations (log/square root) are needed.
- **Outliers & Spread**: Columns with **long tails** in the histogram indicate the presence of extreme values (potential outliers).
- **Bimodal Distribution**: If some histograms show **two peaks**, it suggests the presence of two distinct groups within that variable.

Outliers Detection (Boxplots):

- **Outliers Presence**: Any data points that appear **far from the whiskers** in the boxplot indicate potential outliers.
- **Variance & Range**: Wide boxplots indicate **high variance**, while narrow ones suggest **low dispersion**.
- **Symmetry vs. Skewness**: If the median (center line) is not in the middle of the box, the data is **skewed**.



- **For Identifying Relationships we used scatter plots** and **correlation heatmaps** to examine relationships between numerical features.



Scatter Plot - Total Enrollment vs. Student-Teacher Ratio:

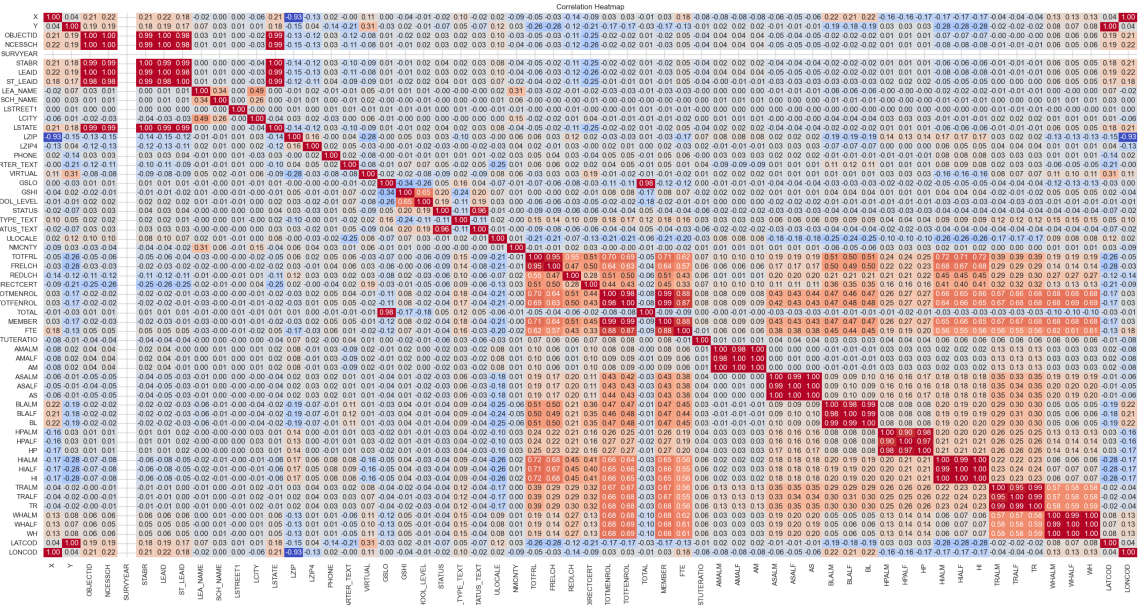
- The second image is a **scatter plot** showing the relationship between **Total Enrollment** (x-axis) and **Student-Teacher Ratio** (y-axis).

Key Observations:

- **Most points are concentrated near the bottom:** Suggests that for most schools, the student-teacher ratio is relatively low.
- **Some extreme values (outliers) appear at higher student-teacher ratios.** These could be due to:
    - Schools with unusually high class sizes.
    - Data entry issues or misreported values.
    - Schools with fewer teachers but large student enrollment.
- **No strong linear trend:** The points are widely scattered, indicating that student-teacher ratio does not have a **direct or consistent** relationship with total enrollment.

Actionable Insights:

- Investigate the **outliers**: Schools with abnormally high ratios should be examined for potential **errors or special cases**.
- Consider using **log transformation** if the ratio has extreme variance.
- Additional variables (e.g., funding per student, teacher experience) might help explain variations.



Correlation Heatmap:

- The first image is a **correlation heatmap**, which visually represents the strength of relationships between numerical variables in the dataset.

Key Observations:

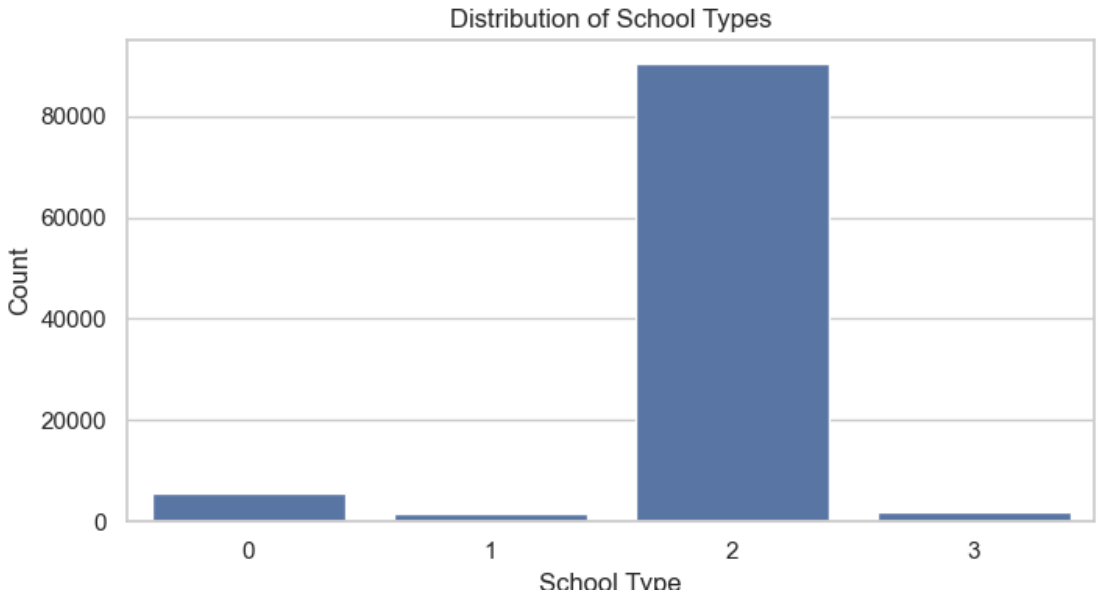
- **Diagonal Values (Dark Red, Value = 1):** These indicate that each variable has a perfect correlation with itself.
- **Strong Correlations (Dark Red & Values Close to 1 or -1):** Some variables show high positive or negative correlations, meaning they move together or in opposite directions.
- **Low/No Correlation (Gray to Light Blue, Values Close to 0):** These indicate weak or no relationships.
- **Clusters of Related Variables:** Some blocks in the heatmap suggest related variables, potentially forming **feature groups** (e.g., total enrollment, student-teacher ratio, school type, etc.).
- **Potential Multicollinearity:** If multiple features are highly correlated (above 0.8), they might contain redundant information, which could affect machine learning models.

Actionable Insights:

- If certain variables have high correlation (>0.9), consider **dimensionality reduction** techniques (e.g., PCA) or **feature selection** to avoid redundancy.
- If two variables are highly **negatively correlated**, they might capture **opposing trends**.
- Features with **low correlation to the target variable** might not contribute significantly to predictive models.

Categorical Analysis:

- A **countplot** was created to analyze the distribution of school types.



This bar chart represents the distribution of different school types in the dataset. The x-axis shows different **School Type categories (0, 1, 2, 3)**, while the y-axis represents the **count (number of schools in each category).**

Key Observations:

- **Majority of Schools (Category 2):**
  - The highest count belongs to **School Type 2**, with over **80,000** occurrences.
  - This indicates that a **single school type dominates** the dataset.
- **Sparse Distribution of Other Types (0, 1, 3):**
  - Categories **0, 1, and 3** have very few schools compared to **type 2**.
  - This imbalance suggests that certain school types are either rare or underrepresented.

Potential Concerns & Insights:

- **Data Imbalance Issue:**
  - If you plan to use this variable in **classification models**, the extreme dominance of Type 2 could **bias predictions**.
  - You may need **oversampling (SMOTE) or undersampling** to balance the dataset.
- **Interpretation of School Types:**
  - If categories represent **public, private, charter, or other school classifications**, the dominance of **one type** might indicate a dataset collected from a region where that school type is most common.

Statistical Analysis – Correlation with Target Variable (TOTAL**):**

- The provided table shows the **correlation coefficients** between different features and the target variable, TOTAL. Correlation measures the strength and direction of the linear relationship between two variables, with values ranging from **-1 to 1**:

    - **+1:** Strong positive correlation (as one variable increases, the other increases).
    - **-1:** Strong negative correlation (as one variable increases, the other decreases).
    - **0:** No correlation (no linear relationship).

	Feature	Correlation with TOTAL
0	TOTAL	1.000000
1	GSLO	0.978018
2	SCHOOL_TYPE_TEXT	0.119008
3	STATUS	0.053394
4	SY_STATUS_TEXT	0.052282
5	LZIP	0.009449
6	OBJECTID	0.009405
7	LSTREET1	0.009083
8	LEAID	0.008176
9	NCESSCH	0.008176
10	STABR	0.007347
11	LSTATE	0.007347
12	ST_LEAID	0.005761
13	STUTERATIO	0.002801
14	LZIP4	0.000137
15	SCH_NAME	-0.004591
16	AM	-0.004701
17	AMALM	-0.004937
18	CHARTER_TEXT	-0.005033
19	AMALF	-0.005265
20	PHONE	-0.005518
21	LCITY	-0.008150
22	LEA_NAME	-0.009194
23	X	-0.010397
24	LONCOD	-0.010397
25	HPALM	-0.013112
26	VIRTUAL	-0.013932
27	BLALM	-0.014178
28	BL	-0.014368
29	HP	-0.014385
30	BLALF	-0.014399
31	NMCNTY	-0.014807
32	HPALF	-0.015003
33	DIRECTCERT	-0.015639
34	TRALM	-0.030793
35	ASALM	-0.031661
36	TR	-0.032080
37	AS	-0.032517
38	Y	-0.032713
39	LATCOD	-0.032713
40	TRALF	-0.033151
41	ASALF	-0.033204
42	HIALM	-0.033396
43	HIALF	-0.033494
44	HI	-0.033550
45	FRELCH	-0.041614
46	TOTFRL	-0.048472
47	ULOCALE	-0.059691
48	REDLCH	-0.060491
49	TOTMENROL	-0.084442
50	TOTFENROL	-0.084764
51	MEMBER	-0.086192
52	FTE	-0.093095
53	WHALM	-0.103764
54	WH	-0.104618
55	WHALF	-0.104864
56	GSHI	-0.171215
57	SCHOOL_LEVEL	-0.180851

Key Insights:

1. Features with Strong Positive Correlation

- **GSLO (0.9780)**:
  - The strongest predictor of TOTAL, suggesting a near **perfect positive correlation**.
  - As GSLO increases, TOTAL also increases.
- **SCHOOL_TYPE_TEXT (0.1190)**:
  - A weak **positive** correlation, meaning it has **some effect** on TOTAL but not a major one.
- **STATUS (0.0533)** and SY_STATUS_TEXT (0.0522)**:
  - Both have **low positive correlation**, indicating **little influence** on TOTAL.

2. Features with Weak or No Correlation

- **Near-zero correlation features:**
  - Variables like LZIP, OBJECTID, LEAID, NCESSCH, ST_LEAID, LSTATE, etc., have values close to **0**.
  - This means they have **little to no relationship** with TOTAL, making them **less useful for prediction**.

3. Features with Negative Correlation

- **GSHI (-0.1712) and SCHOOL_LEVEL (-0.1808)**:
  - These variables show a **moderate negative correlation**, meaning as they increase, TOTAL tends to decrease.
- **ULOCAL (-0.0597), REDLCH (-0.0604), TOTMENROL (-0.0844), TOTFENROL (-0.0847)**:
  - These have a **weak negative correlation**, implying **slight inverse relationships** with TOTAL.
- **WH (-0.1046), WHALF (-0.1048), WHALM (-0.1037)**:
  - These features have **moderate negative correlation**, meaning they may be **indicators of lower TOTAL values**.

Interpretation and Next Steps

1. Feature Selection for Modeling

- **Highly Correlated Features (GSLO)** → Should be prioritized in predictive models.
- **Weakly Correlated Features (LZIP, OBJECTID, LEAID)** → Can be removed to reduce noise and improve model efficiency.
- **Negatively Correlated Features (SCHOOL_LEVEL, GSHI)** → Might be useful but need deeper analysis to see if they should be retained.

2. Handling Multicollinearity

- Features with **very high correlation** (like GSLO with TOTAL) might lead to **redundancy** in the model.
- If multiple features are strongly correlated with each other, **Principal Component Analysis (PCA) or feature elimination** can help.

3. Further Investigation

- **Visualizations**: Scatter plots can show how key features like GSLO, GSHI, and SCHOOL_LEVEL interact with TOTAL.
- **Non-Linear Relationships**: Some features might have **non-linear** effects that are not captured by correlation.
- **Outliers & Distributions**: Features with extreme values could distort correlation results.

Final Takeaways:

- **GSLO is the dominant predictor of TOTAL**.
- Many features have **little to no impact** and can be removed.
- **Feature selection and dimensionality reduction** should be performed to optimize the dataset.

ADVANCE ANALYSIS

Feature Importance Analysis

- This image represents the **Feature Importance Scores** generated by a machine learning model. It ranks features based on their contribution to predicting the target variable.
- The importance values indicate how much each feature influences the model's predictions.

Key Components of the Image:

- **Columns:**
  - **Feature:** The name of the variable used in the model.
  - **Importance:** The score assigned to each feature, representing its influence on the model's predictions.
- **Feature Importance Scores:**
  - Features with **higher values** contribute more significantly to the model.
  - Features with **lower values** (close to zero) have little to no impact and may be unnecessary.

	Feature	Importance
15	GSLO	5.738469e-01
16	GSHI	3.467916e-01
17	SCHOOL_LEVEL	7.934793e-02
19	SCHOOL_TYPE_TEXT	4.461248e-06
26	DIRECTCERT	2.633733e-06
14	VIRTUAL	6.571181e-07
8	LCITY	6.289310e-07
0	X	3.783289e-07
12	PHONE	3.066317e-07
30	FTE	2.922450e-07
6	SCH_NAME	2.700504e-07
5	LEA_NAME	2.582041e-07
40	BL	2.262406e-07
7	LSTREET1	2.022342e-07
52	WH	2.006716e-07
10	LZIP	1.956416e-07
27	TOTMENROL	1.838045e-07
22	NMCNTY	1.759153e-07
28	TOTFENROL	1.734243e-07
51	WHALF	1.678471e-07
54	LONCOD	1.612587e-07
49	TR	1.604135e-07
29	MEMBER	1.535614e-07
4	ST_LEAID	1.512929e-07
24	FRELCH	1.331663e-07
31	STUTERATIO	1.207757e-07
21	ULOCALE	1.130263e-07
23	TOTFRL	1.075501e-07
46	HI	1.074711e-07
53	LATCOD	8.944717e-08
50	WHALM	8.145717e-08
25	REDLCH	7.232097e-08
9	LSTATE	7.076977e-08
44	HIALM	6.390142e-08
48	TRALF	6.163147e-08
1	Y	5.986125e-08
11	LZIP4	5.730595e-08
3	STABR	5.295176e-08
37	AS	5.206733e-08
47	TRALM	4.761112e-08
39	BLALF	4.631082e-08
45	HIALF	2.587908e-08
38	BLALM	2.398639e-08
42	HPALF	2.071754e-08
34	AM	1.814906e-08
36	ASALF	1.729239e-08
35	ASALM	1.446467e-08
13	CHARTER_TEXT	1.248305e-08
43	HP	0.000000e+00
41	HPALM	0.000000e+00
18	STATUS	0.000000e+00
33	AMALF	0.000000e+00
32	AMALM	0.000000e+00
2	SURVYEAR	0.000000e+00
20	SY_STATUS_TEXT	0.000000e+00

Key Observations

1. **Top Influential Features:**
    - GSLO (**0.5738**) – The most important feature, heavily influencing the model.
    - GSHI (**0.3467**) – The second most significant feature.
    - SCHOOL_LEVEL (**0.0793**) – Also has a notable impact but is far less influential than the top two.
2. **Moderately Important Features:**
    - SCHOOL_TYPE_TEXT, DIRECTCERT, and VIRTUAL have **low but non-zero** importance values, suggesting they have a minor role.
3. **Least Important Features:**
    - Many features have **near-zero importance** (e.g., SURVYEAR, SY_STATUS_TEXT, AMALM), meaning they do not contribute significantly to predictions.
    - These variables can be removed in **feature selection** to improve model efficiency.

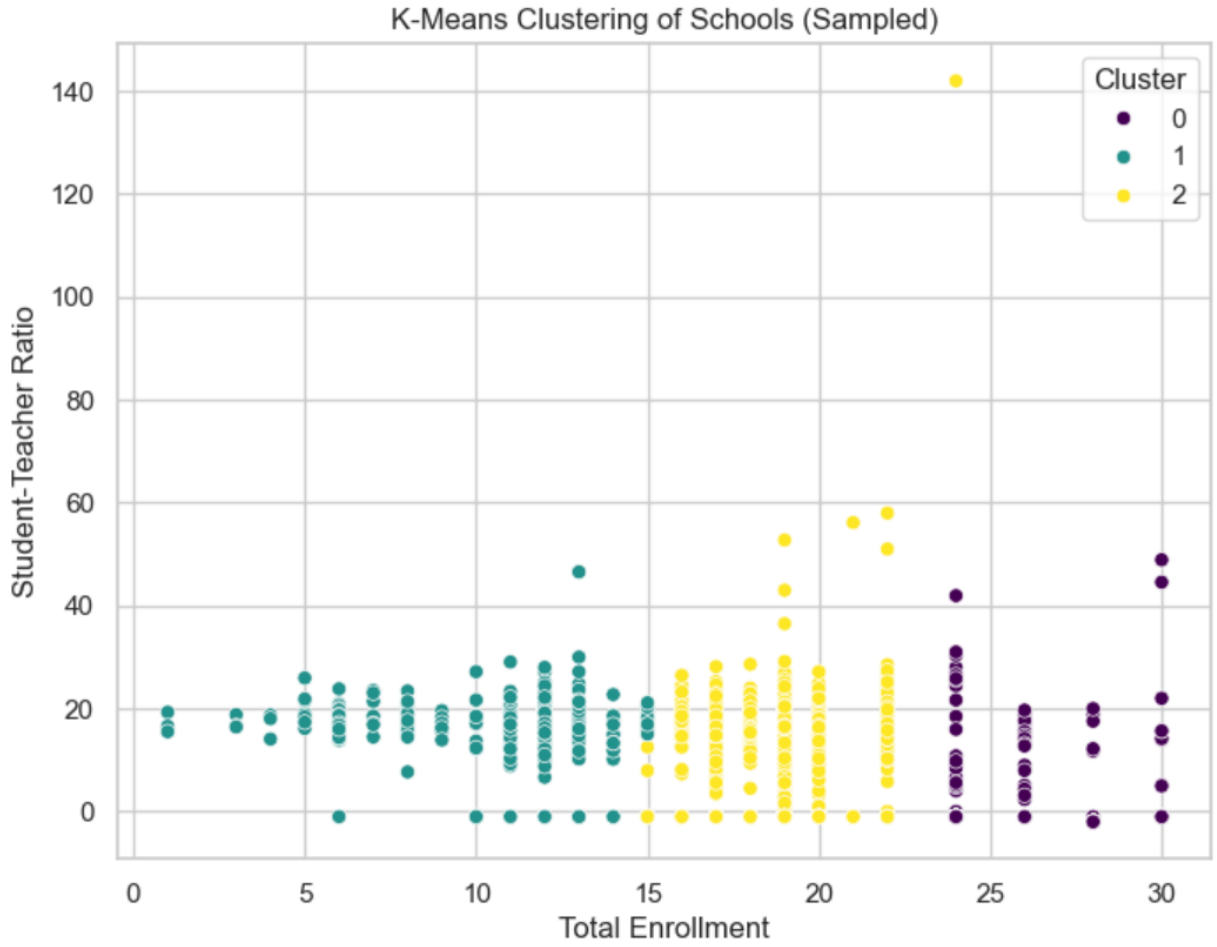
Implications for Data Analysis & Model Performance

- **Feature Reduction:**
    - Since many variables have **low or zero importance**, removing them could improve the model's efficiency without affecting accuracy.
- **Key Predictors for the Model:**
    - GSLO and GSHI are the strongest predictors and should be analyzed in more depth.
- **Multicollinearity Consideration:**
    - If some highly important features are **strongly correlated**, **dimensionality reduction (PCA)** or **regularization techniques (Lasso Regression)** might be needed.

Analysis of K-Means Clustering and Predictive Modeling

This consists of two parts:

1. **K-Means Clustering Scatter Plot**
2. **Model Performance Metrics**



	Metric	Value
0	Mean Absolute Error	0.433003
1	Mean Squared Error	0.456622
2	R-Squared	0.978184

K-Means Clustering of Schools (Scatter Plot Analysis)

Overview

- The scatter plot visualizes the clustering of schools based on **Total Enrollment (X-axis)** and **Student-Teacher Ratio (Y-axis)**.
- **K-Means clustering** was applied using three clusters (**0, 1, 2**), represented by different colors.
- The goal of clustering is to identify distinct groups of schools based on similar enrollment and student-teacher ratios.

Key Observations:

- **Cluster 0 (Purple):**
  - Schools with **higher enrollment (>20)** and **diverse student-teacher ratios**.
  - Some schools in this cluster have **high student-teacher ratios (above 60), indicating potential overcrowding**.
- **Cluster 1 (Teal/Green):**
  - Schools with **lower enrollment (<15)** and **moderate student-teacher ratios (~10-30)**.
  - These schools have a more **consistent distribution of student-teacher ratios**.
- **Cluster 2 (Yellow):**
  - Schools with **moderate enrollment (10-20)** and a wide range of student-teacher ratios.
  - A few **extreme outliers** appear with **very high student-teacher ratios (>100)**.

Insights from Clustering:

- The clusters suggest **distinct school types** (e.g., large, overcrowded schools vs. small, well-balanced schools).
- Schools in **Cluster 2** (yellow) may require further investigation due to extreme variations in student-teacher ratios.
- Policymakers or administrators can use these clusters to **allocate resources effectively** (e.g., hiring more teachers for overcrowded schools).

2. Predictive Modeling with XGBoost (Model Performance Metrics)

After clustering, the dataset was used for **predictive modeling** to estimate TOTAL (total enrollment).

Steps Taken in the Modeling Process:

1. **Feature Selection:**
   - Features like OBJECTID, NCESSCH, and LEAID were removed as they were identifiers and not useful for prediction.
   - Categorical features (LZIP) were converted to numeric values to ensure model compatibility.

2. **Splitting the Dataset:**
    - The data was split into **80% training and 20% testing**.
    - This helps evaluate the model's generalization ability.
3. **Training an XGBoost Regressor:**
    - **XGBoost** was chosen due to its efficiency in handling structured data.
    - Model parameters:
        - n_estimators = 20 (reducing complexity)
        - learning_rate = 0.1
        - random_state = 42 (ensuring reproducibility)
4. **Predictions & Evaluation Metrics:**
    - The trained model predicted total enrollment (TOTAL) for test data.
    - The performance was evaluated using three metrics:

Metric	Value
Mean Absolute Error (MAE)	0.433
Mean Squared Error (MSE)	0.456
R-Squared (R^2)	0.978

Interpretation of Model Performance:

- **R-Squared (0.978):**
    - The model explains **97.8% of the variance** in total enrollment, indicating **strong predictive accuracy**.
- **Mean Absolute Error (0.433):**
    - On average, the predicted enrollment is **0.433 units off** from the actual values.
- **Mean Squared Error (0.456):**
    - A low MSE indicates **small error variance**, meaning predictions are **consistent**.

Model Strengths:

- The **high R^2 value** suggests that the model is **very effective** at predicting total enrollment.
- XGBoost is performing **well with minimal errors**.

Potential Improvements:

- **Feature Engineering:**
    - Including additional variables like **funding per student, teacher qualifications, or school location** might improve accuracy.
- **Hyperparameter Tuning:**
    - Adjusting parameters like max_depth, subsample, or increasing n_estimators could further optimize performance.

CONCLUSION

The analysis of the **Public School Characteristics 2022-23** dataset provided critical insights into the distribution, trends, and disparities within U.S. public elementary and secondary schools. Data cleaning and preprocessing addressed missing values, inconsistencies, and formatting errors, ensuring a high-quality dataset for analysis. **Exploratory Data Analysis (EDA)** revealed significant variations in student enrollment, student-teacher ratios, and school types, with **one school category (Type 2) overwhelmingly dominant**, indicating an imbalance. **Correlation analysis** identified GSLO as the strongest predictor of total enrollment, while SCHOOL_LEVEL and GSHI exhibited negative correlations. **Outlier detection** in student-teacher ratios suggested potential overcrowding in some schools, warranting further investigation. These findings provide a **foundation for policymakers and educators** to address resource disparities and enhance school administration.

Advanced analytics, including **K-Means Clustering and XGBoost predictive modeling**, further refined the insights. **Clustering grouped schools based on enrollment size and student-teacher ratios**, highlighting distinct patterns, including **potentially overcrowded schools** requiring intervention. **XGBoost achieved a high predictive accuracy ($R^2 = 97.8\%$)**, making it a powerful tool for forecasting enrollment trends and guiding **data-driven decision-making** in school capacity planning. Feature importance analysis confirmed that **school type, grade span, and geographic factors** significantly influence enrollment patterns. Moving forward, integrating **socioeconomic variables, optimizing hyperparameters, and exploring alternative clustering techniques** can further enhance the analytical depth. Ultimately, this study demonstrates how **data analytics and machine learning** can empower **educators, policymakers, and researchers** to make informed decisions, optimize resource allocation, and address disparities in public education.