Bricks in the Wall

A Machine Learning Study of Student-Teacher Ratios

Team: Diana Simonson & Elizabeth Frank

Introduction

This project investigates student-teacher ratios across public schools in the United States, with the goal of understanding the underlying structural, geographic, and economic conditions that dictate staffing disparities. While student-teacher ratio is commonly used as a proxy for educational quality it is rarely modeled as a predictive outcome, making this project both unique and significant. By combining supervised regression and unsupervised clustering we aim to uncover the essential groupings of schools and their staffing challenges.

The need for adequate staffing is deeply connected to the fundamental needs of the stakeholders. Students benefit from smaller class sizes that provide individualized attention and instruction building stronger relationships with their educators. Parents depend on schools to provide safe and consistent learning environments that support academic development. Teachers are directly affected by inadequate staffing, increased workloads, higher rates of burnout, and a systemic failure to support. Policymakers and school district leaders require reliable, data driven actionable insights to make informed decisions about funding, hiring, and real time intervention strategies.

Our study revealed that location, regional wage competitiveness, locale classification, and poverty levels were among the most predictive features. Multi-year modeling demonstrated the instability of these relationships over time, while an individual year uncovered meaningful structural divisions with variables like rural versus urban staffing patterns. Tree based models, particularly XGBoost, performed best in predicting multi year student teacher ratios with an R^2 of 0.62. Single year clustering showed how schools can be grouped not only by geography, but by shared contextual burdens, all insights that inform targeted policy responses.

Literature Review

While extensive research exists on factors that influence student achievements and educational inequality, student-teacher ratios are most often treated as background attributes and rarely addressed as predictive targets. With studies traditionally focused on test scores, graduation rates, or funding to determine school quality, research consistently shows that smaller class sizes improve academic outcomes, particularly for younger students, and low income or high need areas (Schanzenbach), reinforcing why understanding and even forecasting student-teacher ratios could offer new actionable insights into educational equity.

Overcrowded classrooms not only affect students, but also entire communities. An educated population supports the success of local economies, civic participation, and future generations. Daily stakeholders such as students, parents, and teachers are witness to what high ratios result in – teacher fatigue, and student disengagement due to less individualized attention. Recent research shows that even modest pay increases for teachers can reduce their attrition rates in prone districts where socioeconomic conditions are challenging (Kraft et al). Adequately staffed classrooms support long-term educational budget planning and allow for working families to have a safe, supervised environment for children. With accurate data analysis, providing actionable insights, policymakers are better equipped to make strategic decisions that move the needle when it comes to prioritizing the classroom structure.

This project addresses the gap in predictive modeling of student-teacher ratios by applying supervised regression and unsupervised clustering techniques. With minimal prior work modeling this specific outcome, similar machine learning methods have been used in related educational research such as dropout prediction, early warning systems, and resource allocation (Baker and Siemens). Employing these techniques to explore complex, highly dimensional datasets from NCES EDGE provides multiple perspectives and a more nuanced understanding of the underlying structure of staffing disparities.

Research indicates that rural schools often face systemic challenges related to isolation, funding formulas, and declining teacher pipelines that are difficult to capture by the urban-centric policy frameworks (Tieken). Our analysis gives special attention to these structural inequities across locales with one clustering result revealing distinct groupings among rural and urban schools, trending with literature by using structural modeling to uncover these obscure patterns in staffing.

Data

The dataset was compiled by merging four publicly available files from the National Center for Education Statistics (NCES) EDGE program. These sources include multiple years of Public School Characteristics, Public School Location, School Neighborhood Poverty Estimates, and the CWIFT teacher wage competitiveness file (National Center for Education Statistics). Depending on the year, each file contains between 98,000 to 101,000 public schools with over 150 potential attributes describing demographics, enrollment, staffing, geography, funding designations, and socioeconomic indicators.

Initial filtering removed schools that did not reflect traditional PreK-12 educational facilities. Excluding virtual schools, adult education programs, and special education institutions that frequently have specialized staffing regimens to reduce skew and preserve comparability. Additionally, schools located in US territories like Guam, Puerto Rico, Northern Mariana Islands, and the US Virgin Islands were removed due to the missing all socioeconomic data and poverty estimates.

Exploration of multi-year trends revealed a slight overall decline in student-teacher ratios, while diverging patterns between student and teacher counts indicate volatility. Charter and magnet schools were found to have higher than average ratios, with charger schools exceeding traditional schools by approximately 25%. Additionally, early analysis showed that free and reduced lunch eligibility, commonly associated with economic disadvantage, only slightly correlated with student-teacher ratios.

Methods & Results

This project involved two complementary approaches conducted by each team member from the NCES EDGE dataset. Both related sub-analysis explored factors influencing student-teacher ratios, one focused on a combination of supervised and unsupervised learning to draw out structural drivers and patterns, and one concentrated on a multi year supervised modeling and clustering to predict and evaluate staffing ratios over time. The two workflows shared many preprocessing steps and core features but each analysis was developed independently and led to distinct outcomes.

Single Year Analysis Methods & Results

The single year analysis machine learning technique was applied to understand the structural contributing factors and their influence on student-teacher ratios. The focus was not only predictive accuracy, but also interpretability and actionable insight. Several features were engineered to reduce dimensionality and better capture school context. A total grades attribute was created to reflect the number of grades attended at each school. Free lunch and reduced lunch eligibility were calculated into rates, serving as an additional proxy for economic disadvantage. Grade level enrollments were banded into groups for elementary, middle, and high school but were later dropped due to the multicollinearity with the school level variable.

Outlier handling was applied selectively and focused on the target variable. Values above 80 were excluded aftering identifying that these extreme ratios typically resulted from schools reporting one or fewer teachers despite having typical student numbers. The decision to apply targeted filtering helped preserve the nature of the data while acknowledging the presence of incorrect data entry.

Supervised learning was used to model the log transformed target variable. A reusable pipeline was defined to streamline imputation, scaling, encoding, and model fitting in a clean, reproducible technique across experiments. Multiple regression models were evaluated, including Ridge, Elastic Net, Random Forest, Gradient Boosting, Hist Gradient Boosting, and XGBoost, the extent of this list is directly related to the desperation to find a model that produced R^2 values that were worth reporting. After extensive runtimes and experimentation, it became clear that the performance had a ceiling and that I had reached its highest peak with tree based models. XGBoost and Random Forest were selected for hyperparameter tuning using grid search across maximum tree depth, number of estimators, and learning rate (where applicable). Final model performance was evaluated using R^2, RMSE, MAE, and Median

Absolute Error and models were refit for the full training data for test evaluation. Predictions and residuals were plotted for further understanding of how the models were generalizing (see Figure A-2).

Exploring structural similarities among schools beyond what regression could capture felt more important after the low predictive power of the supervised learning models for predicting student-teacher ratios. Principal Component Analysis (PCA) was used to reduce the high dimensional data and was followed by tSNE for visualization in a two dimensional space revealing clear separation of school groupings. While KMeans was explored as an option for clustering, HDBSCAN was ultimately used to identify the natural groupings. This method when compared to the tSNE labeled with the locale attribute highlighted the latent structural groupings found in the data.

Supervised regression models revealed moderate predictive power in estimating student-teacher ratios, with tree based modeling significantly outperforming linear approaches. Among the models tested Random Forest Tuned achieved the highest performance with a test $R^2 \approx 0.399$, obtained using 5-fold cross-validation for hyperparameter tuning and a final evaluation on the held out test set. Additional metrics included RMSE, MAE, and Median Absolute Error, which all indicated that while limited, it could detect some meaningful signals in the predictors (see Figure A-1).

Feature importance confirmed that structural and geographic variables were drivers of model performance (see Figure A-2). The most influential features were socioeconomic, including district teacher wage competitiveness, free lunch rate, income to poverty ratio estimates, the one hot encoded Title I unknown, which given data exploration is primarily California, and reduced lunch rate, followed by encoded locales and other categorical variables (see Figure A-4). Despite the moderate success given the type of data, the model residuals do suggest that important variation remained unexplained.

Unsupervised learning offered a valuable secondary perspective of the data given the predictive power of the modeling. Dimensionality reduction using PCA followed by tSNE revealed visually distinct groupings of schools compressed into a 2-D space. When the urban centric locale attribute was labeled on the tSNE projection, clear alignment between the local categories and visual clustered was apparent. Clustering with HDBSCAN reinforced the structural patterns, showing a strong visual alignment with the locale divisions (see Figures A-6 and A-7). This suggests that school context, particularly geographic variables, function as a latent organizing principle and shape staffing disparities beyond just demographics.

Multi-Year Analysis Methods & Results

For the multi-year (2017-2022) evaluation, 7 files (see Data Sources) were joined to create an initial data set with a shape of 269,289 x 88 containing string and numeric data. Every row represented statistics from a specific school for a specific year with 2019-2020 and 2020-2021 missing due to Covid.

The same created features used for the single year analysis were retained for the multi-year analysis. Two other created features were based on student teacher ratio indicating above or below the average, and which quadrant the student teacher ratio was in. The above or below average feature was used to map the data using shape files included with the data for that purpose, and the quadrant was used for prediction, but only one of these features was in the training data at a time during prediction experiments to avoid data leakage.

Meaningful features were selected with preference being given to numeric features that had the least NULLs. Many features were redundant or not useful (for example, the LEA (Local Education Agency) IDs were initially useful for joining tables but were not contextually meaningful. Address data for the LEAs was also dropped, but longitude and latitude data of the schools was kept for location. After dropping features and performing data analysis it made sense to exclude special education schools, and schools with a student teacher ratio of over 60. Special education schools tended to have significantly lower student teacher ratios due to the needs of that special population (Figure A-8). Excluding schools with student teacher ratios over 60 eliminated some non-traditional schools that were independent study and remote learning programs. The number of teachers was excluded to avoid data leakage.

Initial data exploration showed student teacher ratios trending lower over the years (Figure A-8) with student attendance dropping off at the same time the number of teachers was on the rise (Figure A-9). A map of the United States (Figure A-11) built using shape files that came with the data set revealed an uneven distribution of student teacher ratios across the country, validating the need to explore the factors involved.

Initially, the data was split into two dataframes, a numeric dataframe and a string frame, with the exception of a PK that matched a corresponding PK for the same original row in the numeric frame so they could be put back together accurately. Racial data was not included. One hot encoding was used on all the features in the string frame. The dataframes were merged, and a RandomForestRegression model was used to try to predict student teacher ratios. The results were so poor that using the mean of the student teacher ratio would've been more accurate. The Mean Squared Error was 564.90 and R-squared -.29.

Switching the model to XGBoost raised the R-squared value to .06. Dropping a few redundant columns to avoid too much emphasis on location, numeric encoding with meaningful scales rather than using one-hot encoding, scaling data by taking the square root of high range features, and getting rid of outliers all improved performance. At this point, the R-squared was .62. CV was used to evaluate overtraining which indicated the true model performance was closer to an R-squared of .54.

Clustering (Figure A-10) was used on a 5,000 row sample that had been reduced to 2D using tSNE. The clustering was used to evaluate features that might be important according to XGBoost or our own hypothesis. It was surprising to see that longitude seemed to vary significantly in the clusters, while latitude was not that different, but this echoed the findings of the sorted XGBoost importance with longitude being listed first, and latitude being listed 4th. Teacher salary didn't seem very differentiated in the clustering, and was listed 15th out of the 22 features. When Longitude and Latitude were removed, the model performed at .42 R-squared.

When racial data was introduced, that raised the accuracy of the model only 1% all together. Location, primarily longitude, and urbanization are the two top factors for the model. The importance of longitude was also clearly seen in the clustering.

A Keras Sequential neural network was also run on a 5,000 record sample of the data, rotating through three different optimizers - ADAM, RMSprop and SGD using ReLU. Batch Normalization was used on the two dense layer sandwich having 64 and 32 neurons, and linear activation for the output. These all performed pretty close, and echoed the CV results, but was slow, even though Dropout was employed. The XGBoost model is much better suited for this, especially with the full data set.

Discussion

This project's goal was to better understand the underlying structure and contextual drivers of student-teacher ratios in public PreK-13 schools across the United States. We aimed to build predictive models, but the larger goal was interpretability and pattern recognition to support stakeholder decision making. In that sense, both sub-projects achieved meaningful outcomes even if model performance was modest.

For the single year analysis, the regression models, despite extensive preprocessing, experimentation, and tuning offered limited predictive power with Random Forest achieving an R^2 of 0.399. Regardless, the analysis did identify structural and socioeconomic factors that influence levels like wage competitiveness, locale, poverty estimates, and free lunch rates. The revelation came during unsupervised clustering when the latent groupings across locales were revealed, indicating that geography and school context do operate as foundational organizing principles. These insights go beyond the traditional focus on test scores or the common demographics and provide valuable information for framing policy and interventions.

The multi year analysis further contextualized these relationships by examining how they shifted over time. With the XGBoost model achieving a R^2 of 0.54, it highlighted the challenge of model staffing levels consistently over years. With the inclusion of longitudinal data creating some volatility, it also emphasized persistent disparities. Longitude and urbanization ranked among the most influential features, which was mirrored in the clustering analysis and supported by visualization. The exploration of different modeling strategies, including neural networks, underscores the challenge of scaling complex models such as large educational datasets and showcasing the practicality of tree based models.

The two sub-analysis together point to critical insights, student-teacher ratios are not only about the characteristics of students or schools, but possibly more structural, locational, and economic origins. These findings should resonate with stakeholder concerns. Parents and teachers are already deeply aware of how local economics and their geography are a large component to the educational quality of their students. Despite that, policymakers and district leaders often appear to lack awareness of these patterns and seem to lean on aggregated statistics and raw data, information that prevents the more granular data driven insights that should guide meaningful decisions.

Based on the findings of the project as a whole, several key observations have emerged. Geographic areas including some specific states and urban centers had concentrated areas of higher student teacher ratios, an indication of enrollment demands and local cost of living pressures, secondarily supported by corresponding higher wage competitiveness scores that should be prioritized for staffing support and possible resource allocation (see Figure A-5). Structural clusters identified would require a more nuanced intervention as they go beyond state and district boundaries, but still underscore the need for salary adjustments in underserved areas to improve not only teacher retention, but recruitment. Multi-year modeling has shown that historical patterns should be a resource for forecasting and invaluable in long term staffing plans. Implementing educational dashboards to monitor trends and indicators of underlying contextual patterns would be beneficial to district and school leaders to prioritize preventative measures versus reactive. Additionally, investing in the educational national data infrastructure to reduce reporting gaps and systemic data entry errors would strengthen its reliability by policy makers looking to make data driven decisions.

In reflection of the goals of the project, it is acknowledged that the predictive accuracy was limited. The larger picture of uncovering structural disparities and providing insights to stakeholders was meaningfully achieved. The combination of single year and multi-year, supervised and unsupervised learning allowed for a deeper perspective to the standard descriptive metrics frequently used. Confidently, this project ultimately demonstrates how machine learning can effectively reveal hidden inequalities, guide resource allocation, and highlight where intervention is most needed.

Limitations

While this project revealed meaningful insights, several inherent limitations were unveiled when tackling education data and predictive modeling. Even with extensive filtering and thorough exploration, the quality and structure of the data inevitably contained inconsistencies and missing values for not only one offs but for entire states. The reporting gaps are reflective of a system that relies solely on self reporting with apparently limited validation protocols.

Context specific metrics that are difficult or impossible to measure like, instructional quality, appropriate school capacity, and class level variation. Baker and Inventado explain that educational datasets frequently depend on proxy variables and this introduces ambiguity and drastically reduces model precision (Baker and Inventado). The use of these stand in measurements for economic needs like, free lunch eligibility, Title I, and even CWIFT are unable to fully capture the real time changes in communities. And when predictive ceilings occur, as found in this project, it's not only an indication of the difficulty of predicting educational factors, but possibly more reflective that the predictive model is lacking key attributes.

The complexity of applying machine learning to educational data not only lies in the data, but in the application of the insights uncovered. Clustering analysis in this project indicated meaningful structural patterns, but interpreting real time recommendations for districts requires far more granularity which results in limited policy actions. The school reforms often fail not due to poor ideas, but the translation of research into equitable, scalable, policy (Slavin). This gap results in generalized notions rather than budgetary action and changes in policy.

Future Work

Looking forward from the findings and limitations in the project presents several avenues for future research. Incorporating more granular data, mid year staffing reports, transfer rates, and budget data could provide higher precision in modeling and allow more dynamic forecasting. Real time teacher movement data, including attrition and hiring rates could also improve short term predictive trends and allow for a more proactive policy response. The continued exploration of the clustering by incorporating the results into more interpretable methods like decision trees on cluster labels could provide insights into the groupings and also how they might evolve over time.

While the goal of this analysis was to deviate from the traditional educational predictive outcomes like student achievement, dropout, and teacher turnover, including or combining these aspects could lead to a richer understanding of the nuance that drives educational patterns. Ultimately, translating the results into actionable dashboards or district specific decision tools and resources for the policymakers and leaders is the most critical next step. Flagging real time staffing risks or socioeconomic shifts can allow for intervention instead of response to systemic changes that impede the ultimate goal which is allowing for the optimal environment in which children learn and thrive.

Work Cited

Schanzenbach, Diane Whitmore. *Class Size and Student Outcomes: Research and Policy Implications*. National Education Policy Center, 2014, nepc.colorado.edu/publication/brief-class-size.

Kraft, Matthew A., et al. "Teacher Pay and Teacher Retention: Evidence from a Policy Intervention in Virginia." *Journal of Policy Analysis and Management*, vol. 40, no. 3, 2021, pp. 784–819. Wiley Online Library, https://doi.org/10.1002/pam.22255.

Baker, Ryan S., and George Siemens. "Educational Data Mining and Learning Analytics." *British Journal of Educational Technology*, vol. 45, no. 3, 2014, pp. 480–490. Wiley Online Library, https://doi.org/10.1111/bjet.12136.

Tieken, Mara Casey. "Why Rural Schools Matter." *The Rural Educator*, vol. 35, no. 1, 2014, pp. 1–4. https://doi.org/10.35608/ruraled.v35i1.234.

National Center for Education Statistics (NCES) EDGE Data Portal. https://data-nces.opendata.arcgis.com/

Baker, Ryan S. J. d., and Paul S. Inventado. "Educational Data Mining and Learning Analytics." *Learning Analytics*, edited by Johan Larusson and Brandon White, Springer, 2014, pp. 61–75.

Slavin, Robert E. "Evidence-Based Reform in Education: What Will It Take?" *European Educational Research Journal*, vol. 7, no. 1, 2008, pp. 124–128. https://doi.org/10.2304/eerj.2008.7.1.124

Data Sources

Public_School_Characteristics_2017-18.csv

Public_School_Characteristics_2018-19.csv

Public_School_Characteristics_2021-22.csv

EDGE_ACS_CWIFT2017_LEA1718.csv

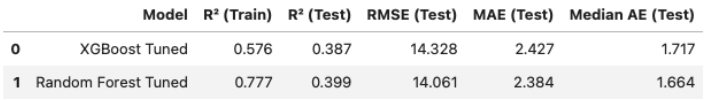
EDGE_ACS_CWIFT2018_LEA1819.csv

EDGE_ACS_CWIFT2019_LEA1920.csv

EDGE_ACS_CWIFT2021_LEA2122.csv

Appendix A

Figure A-1 : Tune Random Forest Evaluation Metrics



	Model	R² (Train)	R² (Test)	RMSE (Test)	MAE (Test)	Median AE (Test)
0	XGBoost Tuned	0.576	0.387	14.328	2.427	1.717
1	Random Forest Tuned	0.777	0.399	14.061	2.384	1.664

Figure A-2 : Prediction and Residuals from Tuned Random Forest

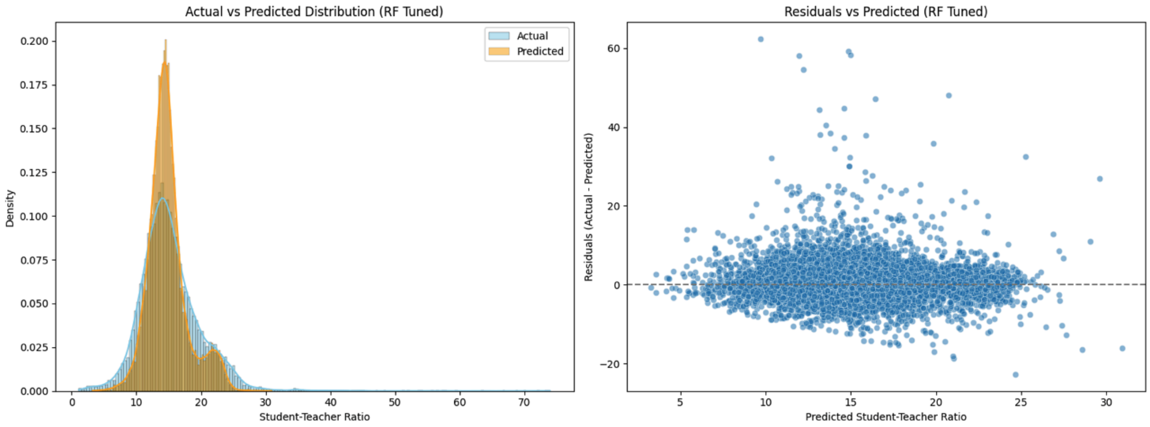


Figure A-3 : Feature Importance from Tuned Random Forest

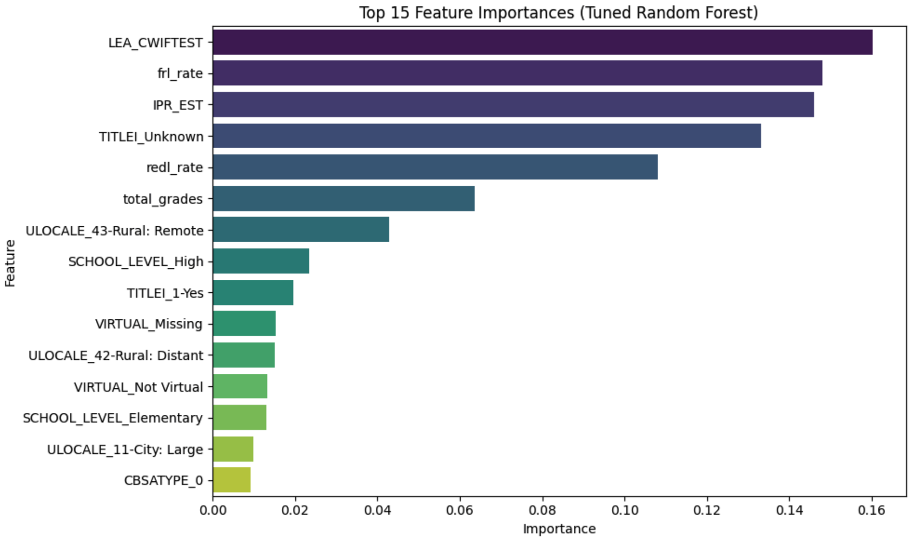


Figure A-4 : Distribution of Schools by Title I Status by State

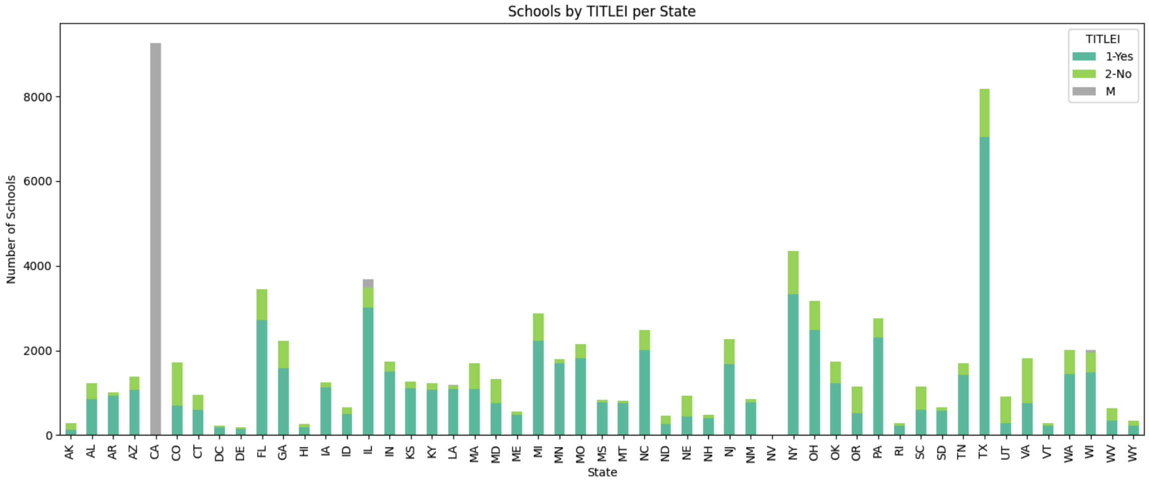


Figure A-5 : Distribution of Schools by Student-Teacher Ratio (between 10-30) and CWIFT

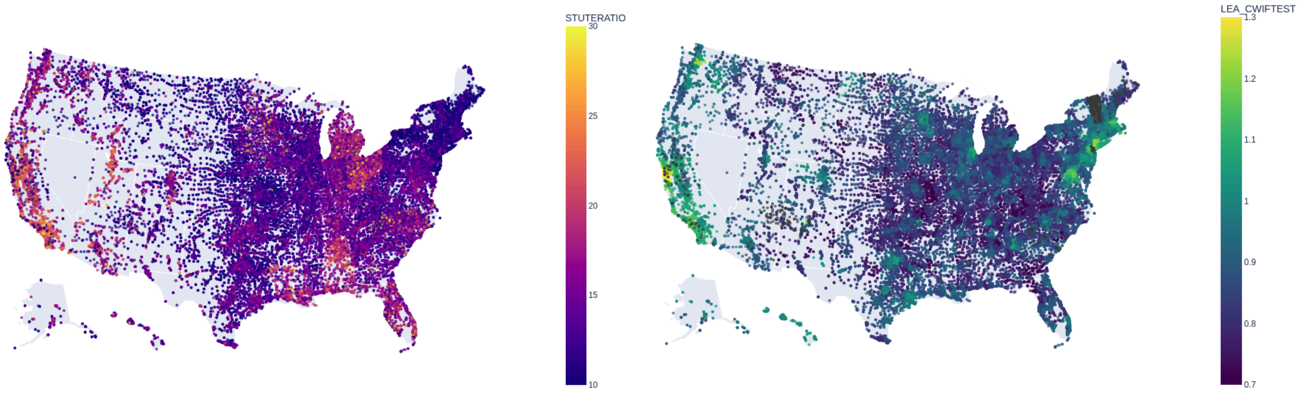


Figure A-6 : tSNE Projection of Schools Labeled by Locale and HDBSCAN Cluster

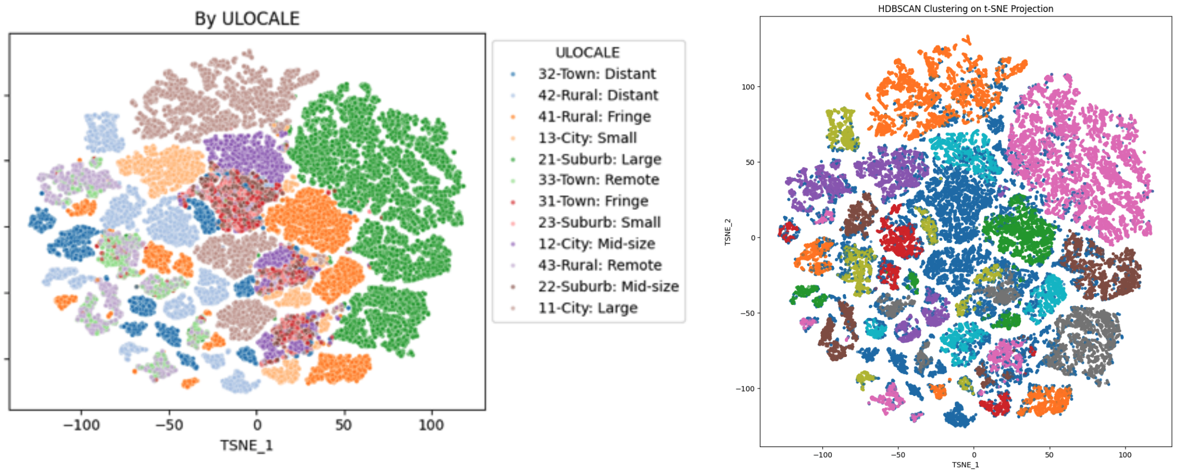


Figure A-7 : Distribution of Urban Centric Locale by HDBSCAN Cluster

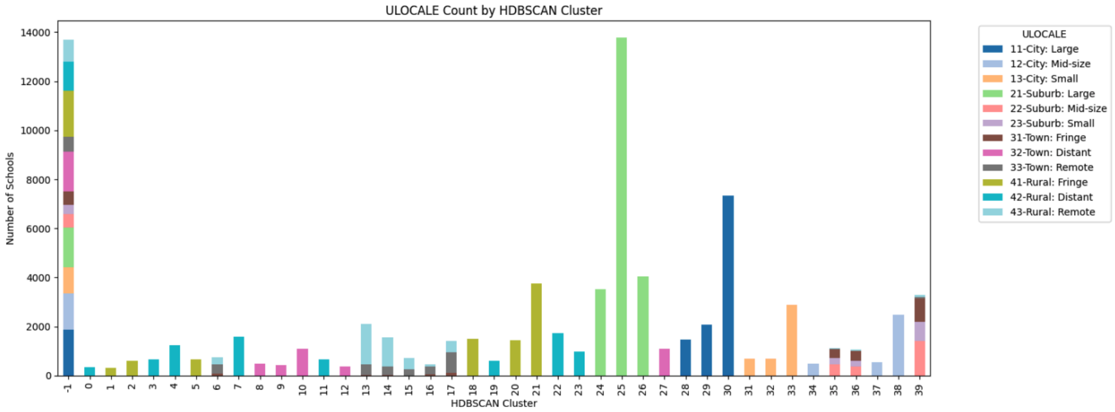
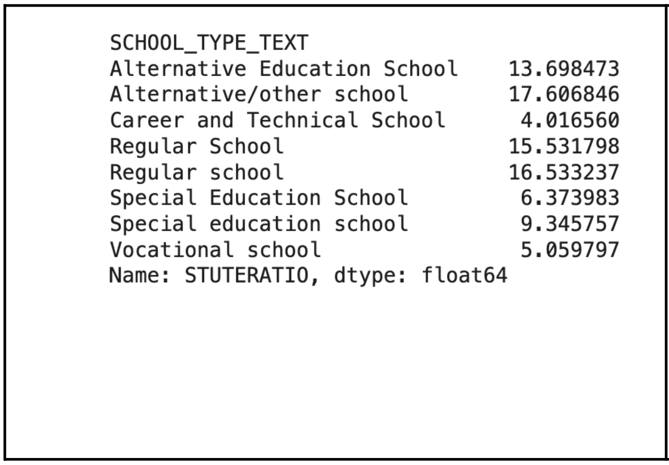


Figure A-8 : Different types of schools average different student teacher ratios, but overall, the student teacher ratios are falling between 2017 and 2022



```
SCHOOL_TYPE_TEXT
Alternative Education School    13.698473
Alternative/other school        17.606846
Career and Technical School      4.016560
Regular School                  15.531798
Regular school                  16.533237
Special Education School         6.373983
Special education school         9.345757
Vocational school                5.059797
Name: STUTERATIO, dtype: float64
```

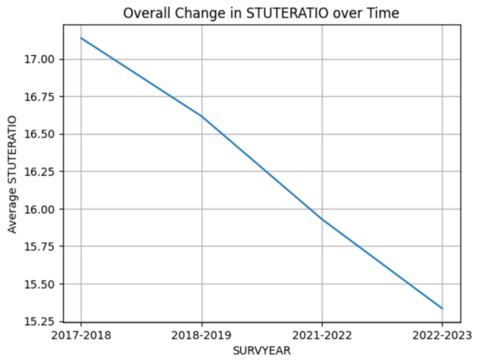


Figure A-9 Student attendance over time has decreased while teachers have increased

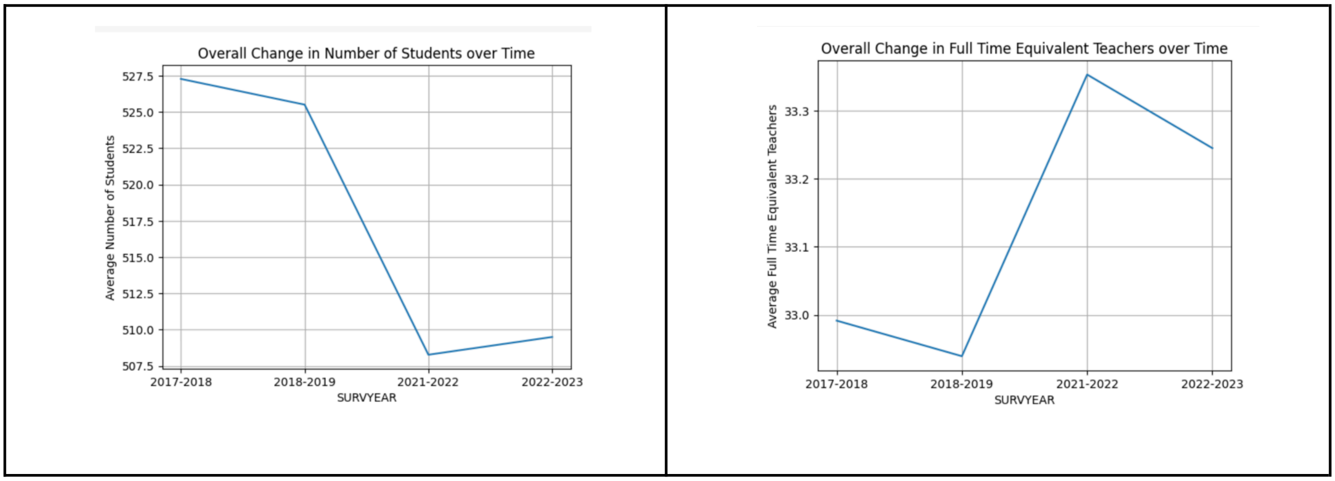


Figure A-10 : Clustering shows three distinct groups with longitude distinctions evident

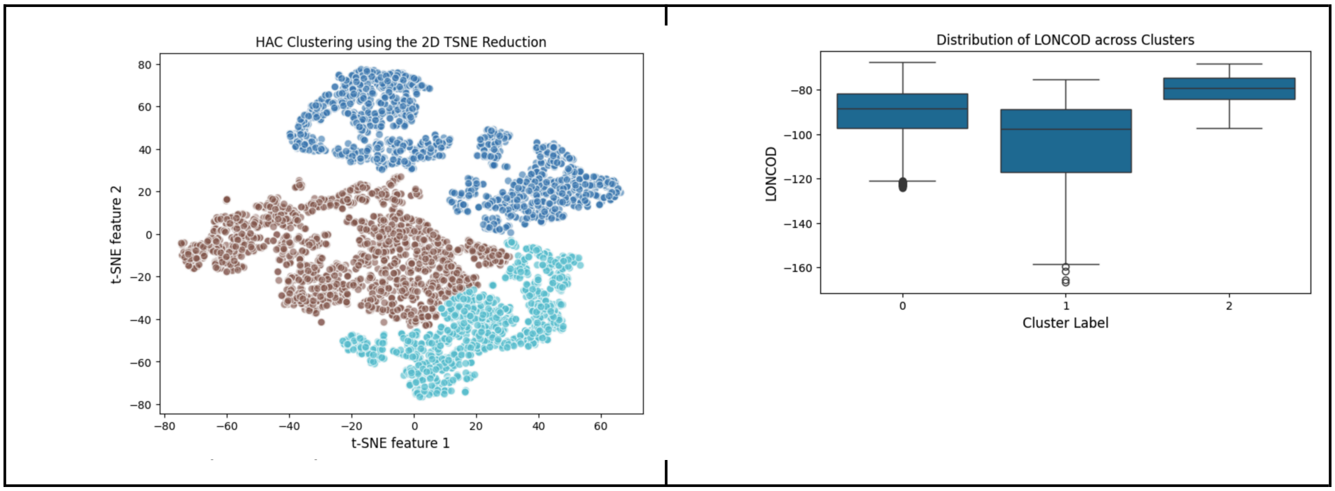
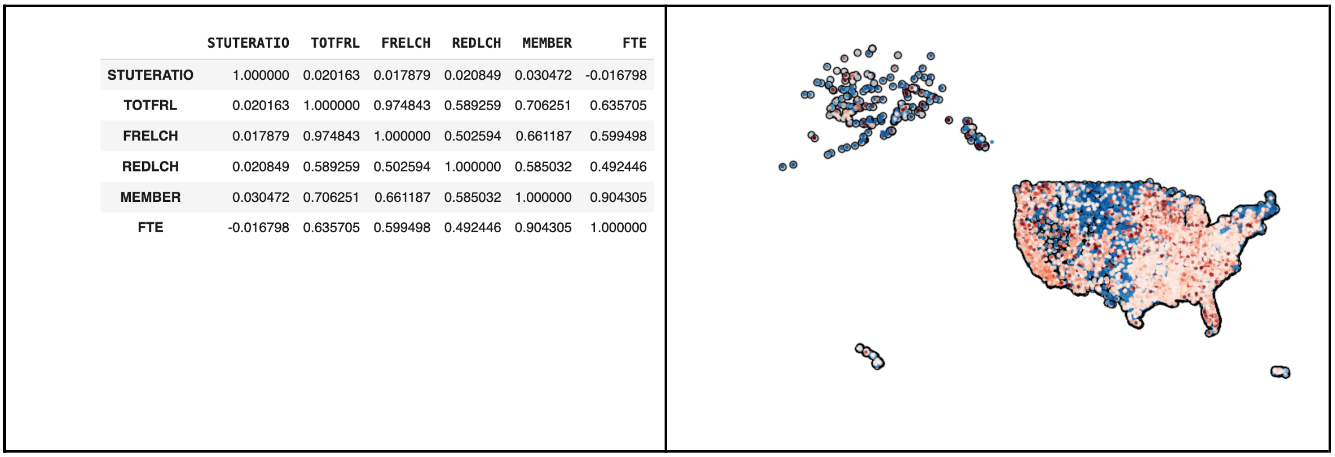


Figure A-11 Free and reduced student lunches, a measure of poverty, are not highly correlated with student teacher ratios. Red shows areas of above average student teacher ratios while blue is below.



	STUTERATIO	TOTFRL	FRELCH	REDLCH	MEMBER	FTE
STUTERATIO	1.000000	0.020163	0.017879	0.020849	0.030472	-0.016798
TOTFRL	0.020163	1.000000	0.974843	0.589259	0.706251	0.635705
FRELCH	0.017879	0.974843	1.000000	0.502594	0.661187	0.599498
REDLCH	0.020849	0.589259	0.502594	1.000000	0.585032	0.492446
MEMBER	0.030472	0.706251	0.661187	0.585032	1.000000	0.904305
FTE	-0.016798	0.635705	0.599498	0.492446	0.904305	1.000000