CS-GY 9223 Final Project Report - EduViz: An Interactive Statewide Education Assessment Data Analytic System

1st Dominic Zhang
New York University
Brooklyn, New York
kz2206@nyu.edu

***Abstract*—Data-driven decision making in education has grown rapidly, yet many districts and state agencies lack the capacity to turn well-documented assessment data into accessible, actionable insights. This project presents *EduViz*, an interactive analytics and visualization system built around State of Texas Assessments of Academic Readiness data from district results for grades 3 and 5 from 2022–2025. Using a time-series split, machine models were trained to forecast district outcomes and highlight districts with unstable or concerning trends. Model performance was assessed with standard metrics, and interpretability was supported through LIME to identify districts and features most influential for predictions. The accompanying Streamlit dashboard supports overview-first exploration and details-on-demand deep dives into model behavior, error patterns, agreement across models, and LIME explanations, enabling educators and policymakers to more closely examine statewide trends and identify districts that may require additional support.**

I. Introduction

Data driven decision making in education has become increasingly popular in recent years [1]. What was once thought of as a profession dependent solely on ”skillful teachers” and ”talented students”, now looks to the monumental amount of education data collected each year in the hopes of learning more about other possible variables that lead to successful and effective learning. As a high school math teacher, I gave out daily feedback forms, as well as quarterly surveys that contain both qualitative and quantitative questions. Data, when used correctly, can inform teachers both of their own pedagogical practices, as well as help them locate areas in the curriculum for improvement. For students, data helps locate areas needing improvement in their own learning and knowledge, which gives rise to clear and actionable items that they can work on.

In recent years, however, education has come under the attack of politicians, and educational budget is often one of the first budgets to be cut by the government when money is on the line. As a result, states, districts, and schools either cannot afford or choose not to engage in detailed data analytics. This has caused a large amount of well-documented and timely data to go unnoticed or underutilized. In this paper, I hope to address that need, for educators and local policymakers alike, by closely engaging in a case study in state-wide assessment data in Texas. I will showcase five different machine learning models that were trained on time series data from all districts on their STAAR (State of Texas Assessment of Academic Readiness) results for grade 3 and 5, in order to predict future trends in performance statewide and in specific districts, as well as highlight districts with sporadic performance that perhaps needs additional support from the state. A data visualization dashboard will also be introduced to support the evaluation of the performance of the different machine learning models, and their biases and struggles in performance will be highlighted and discussed. Finally, I will offer some actionable insights that result from this case study, as well as how other educational entities can easily use this work as the foundation to engage in similar and additional research studies.

II. Prior Related Work

While there is not nearly as much research in the field of education data visualization compared to other fields, likely due to a lack of corporate interest and monetary motivations, some research have been done in recent years advocating for the importance of data analysis and data visualization for educators and policymakers alike.

Dayana, Samanta, Ranganathan, Venkatachalam, and Jain argue that data analytic insights from Exploratory

Data Analaysis (EDA) alone makes its way into strategic business and decision making, and as the size of the data pool increases, we must continue to explore possible models and techniques for understanding and visualizing big data [2]. Moss describes the education system centered around standardized testings and assessments as ”a system of inputs variously defined in terms of money, cultural capital, teacher quality, and the structural features governing the organization of schools nationally and locally, and outputs,represented by student test performance data expressed in national averages and dispersions round the mean” [3]. And Wang, Zhao, and Li argues that current education analysis systems lack strong theoretical foundations, advocates for the careful design and implementation of additional education information platforms and big data visualization tools that can better assist students, teachers, and administrators [4]. Research is also being done in digital data governance, where Williamson surveys the landscape of what he calls ”digital education governance”, in an examination of Pearson Education and the massive online data bank that it uses to construct knowledge regarding education systems. He also proposes the possibility of engaging in predictive analytical work with student data, which can serve in an anticipatory role regarding student and system performance, as well as measuring the impact of instated policies [5].

Through a perhaps more philosophical and aesthetic approach, Mikhaylova and Pettersson lay out their vision that instead of simple data visualizations, the future of education is subject to a continuous invention through datafication and beautification of data, which builds an ”affective connection” with the audience and allows easier access for the general audience in extracting actionable insights from gathered and beautified data [6]. And even when there is top-down communications from the state or district regarding larger scale data findings, Pella argues for the importance of ”contextualized collaborative data inquiry” at the classroom level among teaching colleagues, rather than accepting the findings at face value [7]. My work in this paper is inspired by all of the research mentioned above, and will be an attempt to build and present a usable data visualization tool and dashboard on a case study in the state of Texas that could allow others to continue the work in making effective, accessible, and actionable data visualizations with readily available public education data that teachers and education agencies gather every year.

III. Data Used

The case study and data dashboard presented in this paper is based on the STAAR (State of Texas Assessment of Academic Readiness) results sorted by districts for grade 3 and grade 5 students from the year 2022 to 2025. STAAR is ”a standardized academic achievement test designed to measure the extent to which a student has learned and is able to apply the defined knowledge and skills in the Texas Essential Knowledge and Skills (TEKS) at each tested grade, subject, and course”, and all students enrolled in public schools and open-enrollment charter schools are required to take this test at the end of every academic year [8]. There are a variety of reasons that Texas was chosen to be the state from which data is collected and used for analysis that I feel compelled to mention here. First reason is that there is previous work done on education data visualization that uses the state of Texas as an example, which provided me a clear direction in locating the database (See citation [1]). There was much difficulty in trying to secure federal level data, which was my original intention in researching, with multiple federal level websites and databases returning the ”404 not found” error. There are both practical and political reasons why this may be happening, but that is outside of the main scope of this paper and will therefore not be discussed here. Secondly, Texas consists of the second largest k-12 student enrollment in the country, with around 5.5 million pre-k to 12 students enrolled in the year 2022 alone, which ranks second in the country, only behind California at 5.9 million [9]. Texas also possesses a very diverse student body, with 52.7% of totall enrolled students identifying as Hispanic, 60.6% identifying as economically disadvantaged, and 21.7% identifying as emergent bilingual learners [10].

Thirdly, the aforementioned student body statistics, perhaps not surprisingly, presents tension with the political and economic landscape of the state. Texas funds public and charter schools from collected property taxes, and the massive geographical size of the state means larger, more urban districts will have disproportionally more funding compared to smaller, rural districts with only a few hundred students. Though the state of Texas aims to provide funding based on an amount per student, in reality, this does not work well for a lot of smaller, special-needs schools that require additional support in both labor and workforce as well as in supplies and equipments.

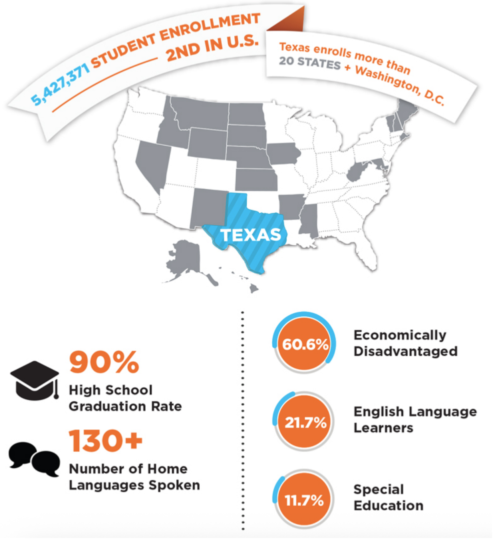

5,427,371 STUDENT ENROLLMENT
2ND IN U.S.
Texas enrolls more than
20 STATES + Washington, D.C.
TEXAS
90%
High School Graduation Rate
130+
Number of Home Languages Spoken
60.6%
Economically Disadvantaged
21.7%
English Language Learners
11.7%
Special Education

Fig. 1: Texas students enrollment trends [11]

IV. Methodology

For this case study, since we are dealing with time series data, I decided to use the data from 2022-2024 as training data, and use 2025 as our testing and validation data. There are 1243 rows × 24 columns in total, with rows being independent school districts, and columns being qualifiers such as number of tests taken, average score, and percentage that meets/approaches/do not meet state standards. Five different machine learning models were developed in order to better learn from the three years worth of statewide public school assessments in grade 3 and 5 through the STAAR program and to forecast student performance in the future. The intention was threefold. First, to train models that learn from our data and be able to hopefully engage in predictive work regarding educational trends at the state level. Second, to use explainability techniques such as LIME and data visualization techniques to discover and isolate particular districts that can help explain the overall trends in the state, which would likely be districts that require additional support from the state level and beyond. Finally, we intend to evaluate the performance and bias in the machine learning models trained, and develop a data visualization dashboard that allows educators and policymakers alike to engage in close examination of the collected data, in order to create actionable insights that can lead to change.

To start, I conducted data cleaning and preprocessing, with one-hot-encoding done for grade 3 and grade 5, as well as all districts. As mentioned above, data was then split into train and test sets for model training. For regression models, I trained a generalized additive model (GAM), a gradient boosting regressor (GBR), and a logistic regression model. For classification models, a decision tree model and a simple neural network in the form of a fully connected multilayer perceptron (MLP) are developed. Classification in this context was set such that a district with over 70% of students passing the STAAR in a grade would be labeled as a pass, while a district with less than a 70% pass rate would be labeled as a fail. In particular, the MLP was designed with two hidden layers with ReLU activation and sigmoid nonlinearity in the final output layer for classification. I also scaled the data, introduced batch normalization and dropouts to obtain a much stronger performance. Finally, all model predictions, evaluation scores such as the accuracy, precision, and recall scores, and LIME scores were calculated for all models, and precomputed into .csv files, in order for better performance at the visualization stage.

V. Evaluation

In the following section, I will present both the validated results of the different machine learning models that were trained for predictive analysis, as well as the quality of performance of those models, and any potential performance biases that may exist in the trained models. I will also present the data dashboard that was designed and built to help users engage in data visualization and deeper explore the relevant models and findings. Finally, I will give a concrete example of the findings in this case study, involving two different districts that models identified as important in predicting statewide trends, and talk about any potential future applications and work that can be done here.

A. Results and Model Performance

Out of the three classification models, the decision tree model gives us the best F_1 score and accuracy, while the logistic regression model showed significant struggles in learning through a low accuracy in conjunction with a high recall, as seen in Table 1.

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.773	0.705	0.952	0.810
Decision Tree	0.837	0.811	0.886	0.847
Neural Network (MLP)	0.817	0.826	0.811	0.819

TABLE I: Classification performance metrics across the three classification models.

Using the performance of the logistic regression classification model as an example, I see that it is a highly sensitive model due to its high recall, but it also produced the most number of false positives. Table two contains the confusion matrix for logistic regression in units of number of districts. Advanced data visualizations through the dashboard will be shown in the following section.

	Predicted Fail	**Predicted Pass**
Actual Fail	130	91
Actual Pass	11	217

TABLE II: Confusion Matrix for Logistic Regression

One interesting pattern that emerged from the data analysis is that all three classification models produced higher accuracy in predicting third grade performance in 2025 compared to their fifth grade counterparts. With logistic regression at 84% for third grade versus 70% for fifth grade, decision tree at 87% for third grade and 81% for fifth grade, and simple MLP at 83% and 80%, respectively. This warrants additional investigation in the real world context of Texas education. One possible explanation is due to third graders likely all coming into school with a more predictable and solid foundation in what they will be tested on at the end of the year, while fifth graders are more unpredictable in their initial stages of adolescent development, and their STAAR assessments presenting to be more rigorous or difficult compared to the third grade counterpart.

By looking at the sample math test questions released, the first third grade question asks which algebraic expression with only addition is equivalent to the given number, while the initial fifth grade question provides a linear equation in $y = mx + b$ form, with three points on the line graphed on 2D plane, and asks for additional possible points on the line – a much higher order level of reasoning and understanding of both algebra and geometry is required here.

For the two regression models, the Generalized Additive Model and the Gradient Boosting Regressor, I calculate the mean absolute error, the root mean square error, and the R^2 value. Both regression models struggled in learning and making predictions, with GBR only yielding us an R^2 score of 0.417, showing very moderate fit and explanation of variation in the data. See Table three below. I also calculated the mean residual for both GAM and GBR. GAM shows signs of underprediction, with a mean residual of -2.33, while GBR shows signs of overprediction, with a mean residual of 2.09.

	MAE	**RMSE**	R^2
GAM	5.421	7.240	0.637
GBR	7.075	9.168	0.417

TABLE III: Error values for regression models

In terms of error analysis of the two regression models, they are examined against both grade level as well as by district types. GBR overpredicts fifth grade performance in 2025 and GAM underpredicts third grade performance, both by approximately 3 to 4 percentage points. Looking at district breakdowns, a similar pattern emerges. Both models underpredicted on larger, more affluent districts in TX while overpredicting on smaller, more rural districts. One possible explanation could be the lack of samples and data points from smaller districts and charter schools. Parents are given the option to opt their child out of the STAAR testing, by receiving an automatic zero on that year's exam, while still staying in the class.

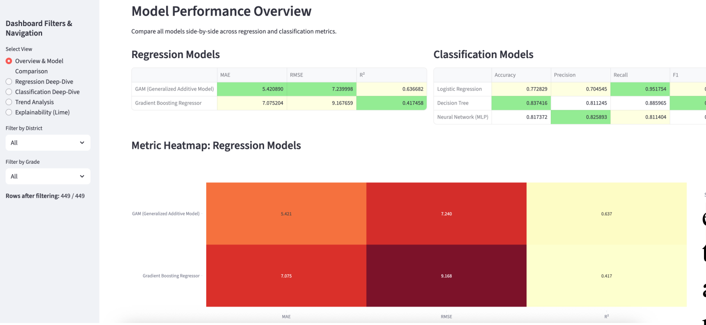
In general, all models were able to predict and clearly identify the top and bottom performing districts in the state. Average regression disagreement was calculated to be 5.43%, and average classification agreement was 79.1% With Coppell ISD being the highest performing district, it has an average score of 90, which is known to be one of the most expensive areas to live in the state of Texas, mainly due to the performance of Coppell, as well as its proximity to DFW airport. On the other hand, George I Sanchez Charter, a school that accepts mainly struggling and failing students from nearby districts, was predicted to be the lowest performing district, with an average score of 10. Additional visualizations will be provided in the visualization dashboard subsection below, and additional real world connections and analysis will be discussed in the final subsection.

B. *The Data Visualization Dashboard*

The data visualization dashboard is designed and built using Streamlit, which provides a simple and accessible framework for building interactive web-based data apps, and Plotly was used in favor of matplotlib to create more detailed and interactive plots.

The dashboard follows an overview-first and details on demand approach, and offers various options and filters to the user from the beginning. The user is able to choose among the options of: ”Overview & Model Comparison”, ”Regression Deep-Dive”, ”Classification Deep-Dive”, ”Trend Analysis”, and ”Explainability (LIME)”.

On the landing page, the user is given all filtering and deep dive choices on the tool bar on the left hand side of the screen, while the main panel displayed performance overviews of models, which includes all of the values of sklearn metrics: accuracy, precision, F1 score, etc.



Dashboard Filters & Navigation

Select View

- ● Overview & Model Comparison
- ○ Regression Deep-Dive
- ○ Classification Deep-Dive
- ○ Trend Analysis
- ○ Explainability (Lime)

Filter by District: All

Filter by Grade: All

Rows after filtering: 449 / 449

Model Performance Overview

Compare all models side-by-side across regression and classification metrics.

Regression Models

	MAE	RMSE	R^2
GAM (Generalized Additive Model)	5.420890	7.239998	0.636682
Gradient Boosting Regressor	7.075204	9.167659	0.417458

Classification Models

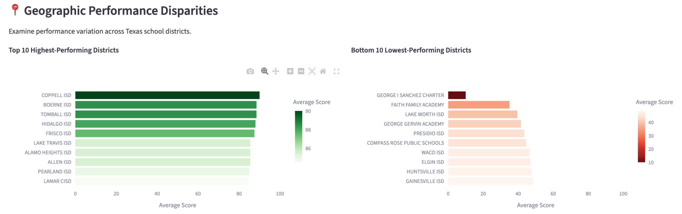
	Accuracy	Precision	Recall	F1
Logistic Regression	0.772829	0.704545	0.951754	0.809701
Decision Tree	0.837416	0.811245	0.885965	0.846960
Neural Network (MLP)	0.817372	0.825893	0.811404	0.818584

Metric Heatmap: Regression Models

	MAE	RMSE	R^2
GAM (Generalized Additive Model)	5.421	7.240	0.637
Gradient Boosting Regressor	7.075	9.168	0.417

Fig. 2: The landing page display of the dashboard

As mentioned in the above subsection, the dashboard also displays the top 10 and bottom 10 performing districts predicted and validated with 2025 data. A data table with the corresponding districts are also provided below, including the average score in each district, the calculated standard deviation, as well as the pass rate of the STAAR assessment in 2025. See figure 3 below for the bar graph display of the top and bottom performing districts..

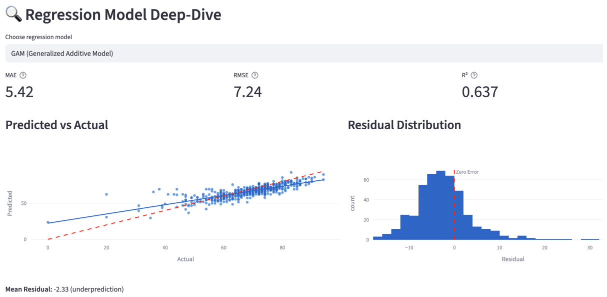


📍 Geographic Performance Disparities

Examine performance variation across Texas school districts.

Fig. 3: The bar graph dispaly of the top and bottom 10 district performances in 2025

Moving onto the deep dive pages for each type of models, regression deep dive allows users to examine either the Generalized Additive Model or the Gradient Boosting Regressor more closely from a drop-down menu selection. Each model's performance and evaluation with its mean absolute error, root mean square error, and the R^2 value are listed, as well as scatter graph with trend lines on all true vs. predicted data points, as well as a vertical bar graph of the residual distribution of predictions. See figure 4 below.



🔍 Regression Model Deep-Dive

Choose regression model: GAM (Generalized Additive Model)

MAE: 5.42

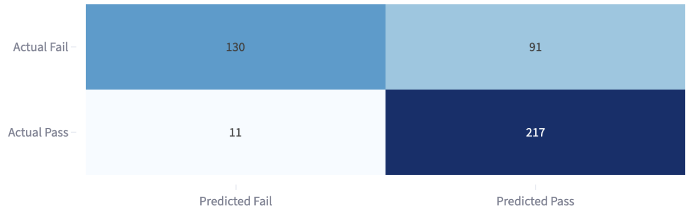
RMSE: 7.24

R^2: 0.637

Mean Residual: -2.33 (underprediction)

Fig. 4: Regression deep dive example for GAM

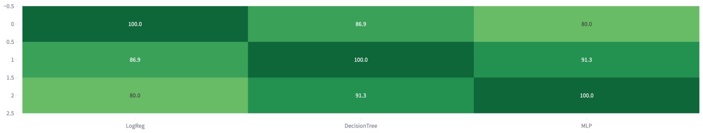
The dashboard also highlights the biases and differences in each model's performance, as mentioned in the results subsection. Color-coded horizontal bar graphs are used to display difference in mean residuals in the prediction by grade level (grade 3 versus 5), while mean residuals for all districts is displayed in a horizontal diverging bar graph. Similar features are displayed for the classification deep dive, where users can choose among the classification models trained and see similar information displayed. In this section, confusion matrices are also displayed to highlight the breakdown of all predictions.



	Predicted Fail	Predicted Pass
Actual Fail	130	91
Actual Pass	11	217

Fig. 5: Confusion matrix display for logistic regression.

In ”Trend Analysis”, additional visualizations on model performances are shown to better help users evaluate each model. Displays include average regression disagreement average classification agreement, a GBR versus GAM agreement graph, a classification models agreement matrix, as well as residual trends by grade and model.



	LogReg	DecisionTree	MLP
0	100.0	86.9	80.0
1	86.9	100.0	91.3
2	80.0	91.3	100.0

Fig. 6: The agreement matrix display for classification models.

Finally, the ”Explainability (Lime)” page displays and helps users understand why each regression and classification model made the specific predictions they did. Top 5 lime explainers are listed for each regression model, both locally and globally, and their lime importance scores are displayed in a table. These results highlight specific districts that had really unpredictable performance and therefore likely led to the specific predictions that the model made, hence the high lime importance score. See figure 7 below for an example of the top 5 lime explainers by our Generalized Additive Model.

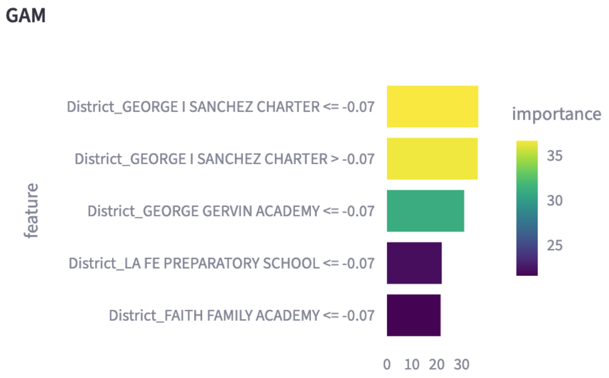


Fig. 7: Top 5 lime explainers for GAM.

C. Real World Connections and Future Work

As mentioned above, our models were able to reliably predict statewide STAAR assessment performances by students. In particular the top lime explainers provides us with interesting insights by highlighting specific districts that performed inconsistently and caused the model to make specific decisions more than other districts. For example, one feature that all models consistently flagged up was the George I Sanchez Charter School. The visualization dashboard correctly identifies this as a school that requires closer attention. Indeed, looking into George I Sanchez Charter School shows that 89.9% of students were considered at risk of dropping out of school, 68.1% of students were enrolled in bilingual and english language learning programs, and 49.4% of all students are classified as chronic absentees, meaning they have missed at least ten percent of the school year [12]. This school is known to accept at-risk students from other nearby public school districts and charter schools, which further diminishes its passing rate on the end of year statewide standardized assessment. This tool, therefore, can very much be used by state policymakers who are interested in looking at the performances of different districts and identifying at risk districts that need additional support.

In the future, much additional work can be done and much improvements can be implemented to the system. For example, I omitted dimensionality reduction in the analysis stage of data and machine learning models. Our main reason behind this decision is that I wanted to focus on lime for explainability and interpretability rather than engaging in principle component analysis or other dimensionality reduction techniques that would reduce interpretability. Another idea to implement would be to combine GeoJSON files of all school districts in the state and create choropleth maps with embedded information that I have calculated and analyzed elsewhere in the visualization dashboard. For example, since Texas funds public school districts through property tax collections, geographical features of districts will certainly have a correlation to the funding and thus performance of the district. And Texas is a state of significant size and contains a huge amount of urban as well as extremely rural districts that would yield interesting analyses and results.

VI. Conclusion

Education data analysis and data visualization is still an emerging field of research. Education is a complex social and deeply political field, where analysis and other quantitative work should never be done in a numerical and technical vacuum. Conducting data analysis with context-specific knowledge is crucial, and as demonstrated in this paper, data visualization techniques can be used in conjunction with machine learning techniques to produce a useful tool that can be beneficial to school teachers and state level education policymakers alike. While there exists a wide variety of factors that would impact a school or district's capacity to engage in data analysis using professional subscription based software (such as Tableau), educators are eager to engage with data that can better inform their teaching and the impact of education policies. Open source software and data visualization tools like the dashboard developed in this paper could be easily adapted by individual districts or states for analysis. In particular, lime explainability was a crucial tool in highlighting districts that performed unpredictably and require additional support from the state. Even with one semester of graduate level study in computer science and machine learning, this project was able to be developed and established as a foundation for much more additional work to come.

Acknowledgment

I want to thank our TA Pari, in particular, for all of his support throughout the process of this project. I had to change the topic of my final project due to technical difficulties quite late in the timeline, and Pari was kind enough to give me a second chance and allowing me to engage in this project on education data that I am personally very passionate about. Our late night zoom calls in the week leading up to the end were extremely formative and beneficial to the quality of this final product, none of which I could've done without him. For my first computer science/data science project after just learning how to code earlier this year, I think it is a decent starting point for much more learning and work to come.

References

[1] Taylor, L.; Gupta, V.; Jung, K.Leveraging Visualization and Machine Learning Techniques in Education: A Case Study of K-12 State Assessment Data. Multimodal Technol. Interact. 2024, 8, 28. https://doi.org/10.3390/mti8040028

[2] Dayana, B.D.; Samanta, A.; Ranganathan, N.; Venkatachalam, K.; Jain, N. A comprehensive approach to visualize industrial data set to meet business intelligence requirements using statistical models and big data analytics. Int. J. Recent Technol. Eng. 2019, 7, 1437–1443.

[3] Moss, G. The Rise of Data in Education Systems: Collection, visualization and use. Lond. Rev. Educ. 2014, 12, 154–155.

[4] Wang, P.; Zhao, P.; Li, Y. Design of Education Information Platform on Education Big Data Visualization. Wirel. Commun. Mob.Comput. 2022, 2022, 6779105

[5] Williamson, B. Digital education governance: Data visualization, predictive analytics, and 'real-time' policy instruments. J. Educ. Policy 2016, 31, 123–141.

[6] Mikhaylova, T., & Pettersson, D. (2025). The timeless beauty of data: inventing educational pasts, presents and futures through data visualisation. Critical Studies in Education, 66(2), 142–158.

[7] Pella, S. What should count as data for data driven instruction? Toward contextualized data-inquiry models for teacher education and professional development. Middle Grades Res. J. 2012, 7, 57–75.

[8] Texas Education Agency. 2024. "STAAR — Texas Education Agency." Texas.gov. September 10, 2024. https://tea.texas.gov/student-assessment/staar.

[9] National Center for Education Statistics. 2024. "COE - Public School Enrollment." Nces.ed.gov. May 2024. https://nces.ed.gov/programs/coe/indicator/cga/public-school-enrollment.

[10] Division of Research and Analysis Office of Operations, Texas Education Agency. 2022. Enrollment in Texas Public Schools 2021-22.

[11] Raise Your Hand Texas, "2023 Texas Education by the Numbers." Accessed: Feb. 10, 2025. [Online]. Available: https://www.raiseyourhandtexas.org/2023-texas-education-by-the-numbers/

[12] "George I Sanchez Charter." 2015. Texas Public Schools. Dec. 8, 2015. https://schools.texastribune.org/districts/george-i-sanchez-charter/.