Final Project Report

Adam Foster & Taylor Lehman
INFO-I513: Usable AI
GitHub Link

Applying Artificial Intelligence Concepts to Indiana Department of Education Data

We've elected to explore and analyze data within Indiana K-12 schools, centered on assessment performance. We both grew up attending Indiana schools and have backgrounds in education – one as a former teacher, the other as a current data analyst. While there is question around the weight given to assessment results and how informative they are, many factors that influence assessment results can be captured in data that is unrelated to test taking, such as school funding, free and reduced lunch numbers, attendance rates, early-year reading comprehension, and other assessment results.

Key Questions

We sought to answer a number of questions with this project, including these key questions:

- Which features (enrollment, attendance, demographics, etc.) of a school/corporation are most indicative of quality assessment results?
- Can we predict assessment results based on key features of a school/corporation?
- Does more funding within a school/corporation mean better assessment results?
- Which school corporations across the state are most alike? What defines their similarities?
- Can we use assessment predictions to recommend in-state partnerships between corporations?

Data Description

The Indiana Department of Education offers a plethora of datasets around its public (and some private) schools and corporations. These datasets include, primarily for our interests, results from SAT, ACT, ILEARN, IREAD, and I AM assessments, as well as archived ISTEP results.

In addition to assessment results, the IDOE provides data around attendance, enrollment, demographics (grade level, ethnicity, free/reduced meals, gender, special education, etc.), third-grade reading levels, college/career readiness, funding, teacher statistics, and school ratings. Most of these datasets are downloadable in Excel Workbook format.

Specifically, we cleaned and stored 15 datasets of corporation-specific data inside of a SQLite database, connected by a single variable – "Corporation ID":

- Corporation enrollment, disaggregated by grade and gender (2024)
- Corporation enrollment, disaggregated by ethnicity and free/reduced lunch (2024)
- Corporation enrollment, disaggregated by English Language Learners and Special Education (2024)
- Corporation IREAD results (2024)
- Corporation ILEARN (grades 3-8) results, English Language Arts (2024)
- Corporation ILEARN (grades 3-8) results, Mathematics (2024)
- Corporation ILEARN (grades 3-8) results, ELA & Math (2024)
- Corporation ACT results (2008-2018)
- Corporation SAT results, Evidence-Based Reading and Writing (2024)
- Corporation SAT results, Mathematics (2024)
- Corporation SAT results, EBRW & Mathematics (2024)
- Corporation third-grade reading data (2012-2023)
- Corporation chronic absenteeism (2013-2024)
- Corporation and School-Level federal ratings (2023)
- ESSA School-Level Financial Data (2022)

The most challenging datasets to load and clean were datasets with disaggregated data. First, while most of the values in the datasets were very clean, disaggregated datasets redacted values for groups where raw counts were less than 10, in order to protect the privacy of the students. While we were aware of this pattern, we never did encounter it during this project.

For every dataset, we ensured the formatting of the variable "Corporation ID" was uniform to best connect the datasets into the SQLite database. We also pulled the most recent year from each Excel workbook (conveyed in the list above). Once the database was prepared, we pulled the specific features that

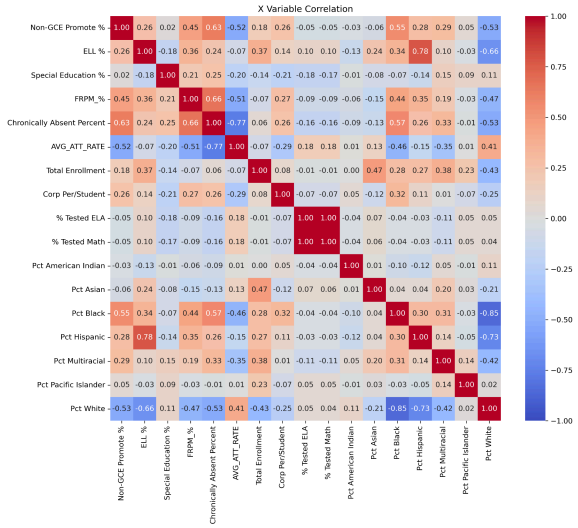
presented as little redundancy as possible. Those features are as follows.

- **Non-GCE Promote %:** The percentage of third-grade students that are students without a Good Cause Exemption, did not pass IREAD, and were promoted to the fourth grade
- **ELL %:** The percentage of total students who are classified as English Language Learners
- **Special Education %:** The percentage of total students that are students who are classified as Special Education students
- **Free/Reduced Price Meals (FRPM) %:** The percentage of total students that qualify for free or reduced-price lunches
- **Chronically Absent %:** The percentage of total students who are deemed chronically absent
- **Avg. Attendance Rate:** The mean of the attendance rates at schools within corporations
- **Corp Per/Student:** The number of dollars the corporation budgeted per student in the school year 2021-22
- **Total Enrollment:** Total students within the corporation
- **Pct American Indian:** The percentage of total students that identify as American Indian
- **Pct Asian:** The percentage of total students that identify as Asian
- **Pct Black:** The percentage of total students that identify as Black
- **Pct Hispanic:** The percentage of total students that identify as Hispanic
- **Pct Multiracial:** The percentage of total students that identify as Multiracial
- **Pct Pacific Islander:** The percentage of total students that identify as Native Hawaiian or other Pacific Islander
- **Pct White:** The percentage of total students that identify as White
- **% Tested ELA:** The percentage of total qualifying students tested for English Language Arts on ILEARN
- **% Tested Math:** The percentage of total qualifying students tested for Math on ILEARN

We then selected two dependent variables we hope to learn about through their dependencies on the features listed above:

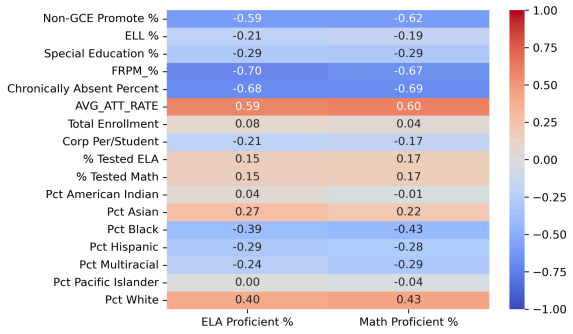
- **ELA Proficient %:** The percentage of ILEARN test-takers in a given corporation that tested proficiently in English Language Arts during the 2023-24 school year
- **Math Proficient %:** The percentage of ILEARN test-takers in a given corporation that tested proficiently in Mathematics during the 2023-24 school year

Below is a correlation map for the X features to quickly investigate how the features interact.

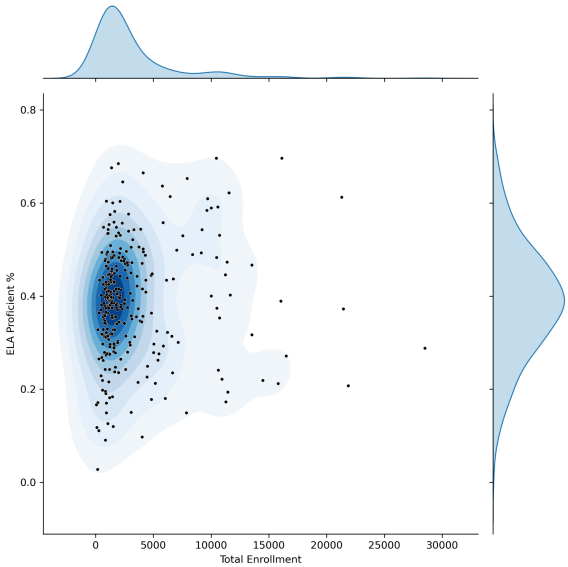


There are a few features (Non-GCE Promote %, FRPM_%, and Chronically Absent Percent) that have stronger correlation coefficients across the entire set than others, and Pct White and Pct Black have the strongest correlations among the ethnicity features, inversely related to non-ethnicity features. These will all be features we keep close consideration of throughout our analysis.

This same pattern stayed true when comparing correlation coefficients between X features and target variables (Math Proficiency % and ELA Proficiency %). We dig into the three non-ethnicity features (Non-GCE Promote %, FRPM_%, and Chronically Absent Rate) in the Results section.

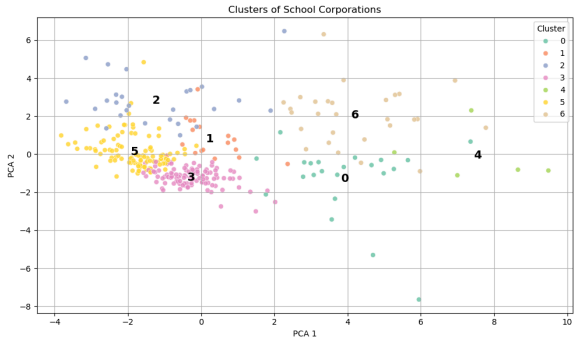


We also plotted out relationships between each feature and one target variable – ELA Proficiency %. Below is the relationship between Total Enrollment and ELA proficiency, a largely independent relationship, but these plots were insightful throughout the process and will appear again below.



Cluster Analysis

Given the number of features included in the dataset, this was a good project to deploy clustering and determine similar corporation profiles for classification. We chose to use KMeans clustering to classify the corporations in the state, and we decided the number of clusters via the Elbow Method (inertia), Silhouette scores, and Davies-Bouldin scores. These results showed us that seven clusters are the most appropriate number for this dataset.



Note: This PCA visualization was constructed before clusters were coded with letters. They are corresponding to the order of the alphabet (e.g. 0=A, 1=B, 2=C, etc.).

The visualized clusters seem effectively distributed, with only a few outliers clustered outside of the outlier cluster (Cluster E, shown as 4 above). Those outliers will be classified alone by the sub-clustering.

We were able to single out the centroids of each cluster to understand which corporations were the best references for each group.

Given the anticipated variance within clusters, because some clusters are tighter than others, we chose to create sub-clusters within every cluster, except Cluster E, a set of five collective outliers. We determined the number of sub-clusters within each cluster by using the Elbow Method, Silhouette scores, and Davies-Bouldin scores for each cluster. This resulted in 36 total sub-clusters.

Distribution of Corporations by Cluster and Sub-Cluster

From here, we calculated the mean for each feature within each cluster and sub-cluster and determined each average's percentile among all of the state's corporations to normalize for easier comparison within and between clusters and sub-clusters. The clusters can be defined as follows:

- **Cluster A:** Tightly packed moderately sized cluster with high rates of non-GCE promotions, free/reduced lunches, chronic absenteeism, and non-White students; Relatively high enrollment and extremely poor ILEARN results on average; *Centroid corporation: Anderson Community School Corporation*
- **Cluster B:** Smaller cluster but with seven sub-clusters and moderate average ILEARN

results; Most defined by high rates of Hispanic students and ELL students but also lower rates of free/reduced lunches than correlation charts suggest; *Centroid corporation: Warsaw Community Schools*

- **Cluster C:** Moderately sized cluster of large schools with low rates of White students, free/reduced lunches, and chronic absenteeism; the highest-performing cluster on ILEARN assessments; *Centroid corporation: Center Grove Community School Corporation*
- **Cluster D:** One of the two very large clusters (with Cluster F) but much lower performing on ILEARN than the other; When compared to Cluster F, has a higher rate of free/reduced lunches, much higher rate of chronic absenteeism, and higher rates of special education students and non-GCE promotions; *Centroid corporation: Spencer-Owen Community Schools*
- **Cluster E:** Group of five outlier schools with very high rates of free/reduced lunches, non-GCE promotions, chronically absent students, and Black students; Extremely poor performances on ILEARN and the highest average funding; *Centroid corporation: Gary Lighthouse Charter School*
- **Cluster F:** Very large cluster with low rates of free/reduced lunches, chronic absenteeism, and non-GCE promotions, while also owning moderate rates of students across all ethnicities; Second-highest performing cluster on ILEARN; *Centroid corporation: Hamilton Heights School Corporation*
- **Cluster G:** A moderately sized cluster with high rates of ELL students, non-GCE promotions, free/reduced lunches, chronic absenteeism, and mostly Hispanic and Black students; Very low average scores on ILEARN; *Centroid corporation: Hamilton Heights School Corporation*

The sub-clusters within the clusters were extremely helpful for intra-cluster analysis, which is shown in the results below.

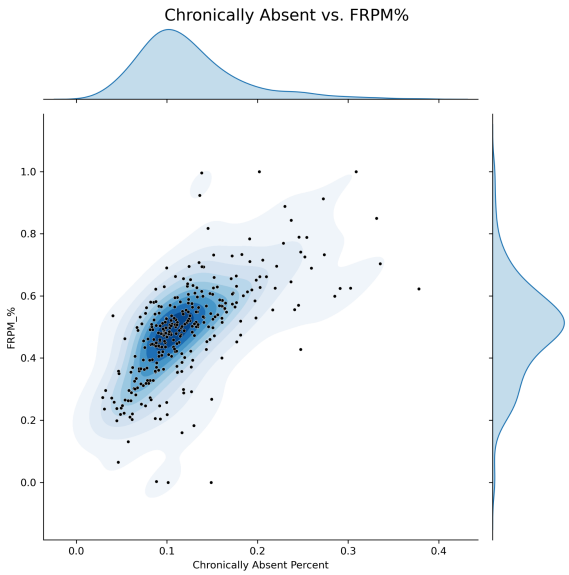
Results

Corporations budget funding in unique ways, making it difficult for direct comparison, so we simply used the feature Corp Per/Student from ESSA reports in 2022, just to offer an idea of this relationship. There is not as much variance between corporation funding as one might assume, and in fact, the corporations receiving the most funding are still struggling on ILEARN. After clustering and sub-clustering, there are 10 sub-clusters that average funding in the top-20th percentile, and eight of them have assessment results beneath the 35th percentile.

Relationships between significant features

Referring to the features correlation chart, we noticed important relationships between the following variables: Chronically Absent Percent vs. Free/Reduced and Paid Meals % (FRPM_%) and Chronically Absent Percent vs. Non-GCE Promote % (students being moved on after third grade despite not passing IREAD and not possessing an exemption). Not only did these features show stronger correlation coefficients with each other, but they also showed the strongest correlation with the target variables as well.

First, the relationship between Chronically Absent Percent and FRPM_% suggests that as the rate of chronically absent students increases, so does the rate of students in need of free/reduced lunches.



The Pearson correlation coefficient between the two features is 0.66, and the p-value is 0.00, suggesting statistical significance.

To qualify for the top-25% of Chronically Absent Percent, a corporation must have 14.69% of its student body deemed chronically absent, and for the top-25% of FRPM_%, 57.84% of its student body

must be receiving free/reduced lunches. The chart above shows that 48 of the 306 corporations within the state (15.7%) are in the top-quarter of both metrics, while 58 other corporations are in the top-quarter of either metric but not both.

When the corporations are labeled with an absent-frpm attribute for top-25% of neither variable, only absenteeism, only free/reduced lunches, or both variables labelled as 0, 1, 2, and 3, respectively, there is a clear correlation between this relationship and other features in the dataset.

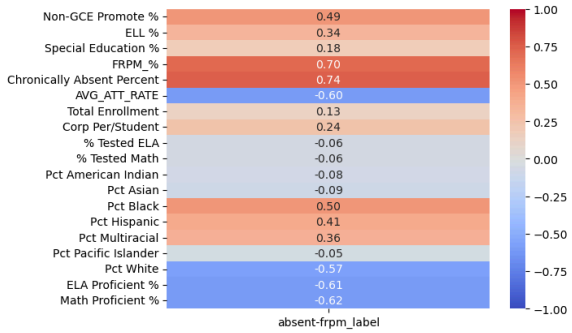
Table. Mean Rates of Corporations by absent-frpm label

absent-frpm label	Pct Black	Pct White	ELL %	ELA Proficient %	Math Proficient %
0	0.027	0.850	0.032	0.439	0.456
1	0.055	0.793	0.045	0.321	0.319
2	0.084	0.665	0.097	0.322	0.313
3	0.254	0.484	0.093	0.249	0.228

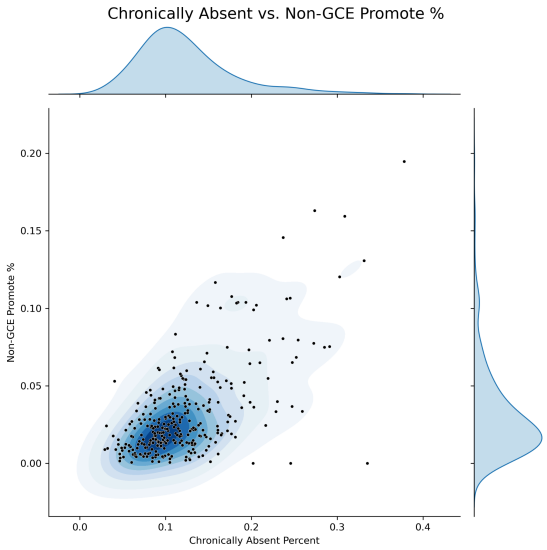
Note: 0=in top-25% of neither variable, 1=in top-25% of only Chronically Absent Percent, 2=in top-25% of only FRPM_%, 3=in top-25% of both variables

As corporations are increasingly stricken by chronic absenteeism and/or free/reduced lunch necessity, their non-White populations (particularly Black populations) increase, while White populations decrease. English Language Learning rates also increase. Both ELA and Math proficiency rates on ILEARN dwindle. Only 20 of the 106 schools with a non-zero absent-frpm label reached the mean in percentage of students proficient in ELA or Math. Three of those schools were in the top-25% of both variables.

The correlation coefficients for these particular features and the absent-frpm label are shown along with the rest of the features in the dataset below.



We performed the same exercise with the relationship between Chronically Absent Percent and Non-GCE Promote % and found a similar relationship.



To qualify for the top-25% of Chronically Absent Percent, a corporation must have 14.69% of its student body deemed chronically absent, and for the top-25% of Non-GCE Promote %, 3.98% of its promoted third graders must be promoted without reaching proficiency in IREAD without exemption. The chart above shows that 43 of the 306 corporations within the state (14.1%) are in the top-quarter of both metrics, while 68 other corporations are in the top-quarter of either metric but not both.

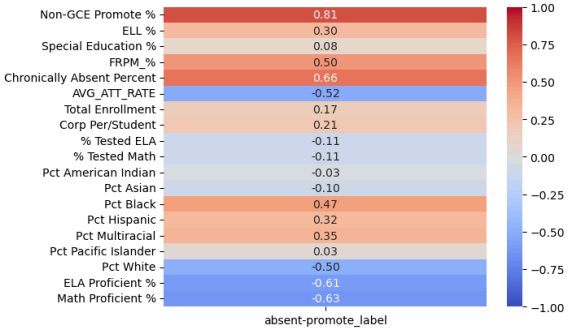
When the corporations are labeled with an absent-promote label (0, 1, 2, 3) for top-25% of neither variable, only absenteeism, only non-GCE promotions, or both variables, there is a clear correlation between this relationship and other features in the dataset.

Table. Mean Rates of Corporations by absent-promote label

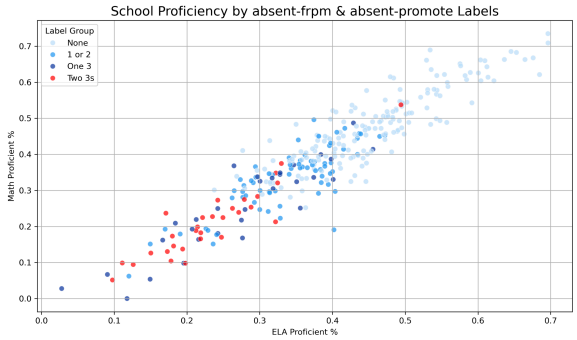
absent-promote label	Pct Black	Pct White	ELL %	ELA Proficient %	Math Proficient %
0	0.029	0.839	0.035	0.442	0.460
1	0.077	0.762	0.041	0.313	0.310
2	0.069	0.753	0.068	0.324	0.308
3	0.259	0.472	0.102	0.247	0.225

Note: 0=in top-25% of neither variable, 1=in top-25% of only Chronically Absent Percent, 2=in top-25% of only Non-GCE Promote %, 3=in top-25% of both variables

This is much like the absent-FRPM relationship, where, as the conditions for each variable increase, so do the average rates of non-White students and English Language Learners, while ILEARN proficiency rates decline. The correlation coefficients also look very similar to what was shown above.



There are 30 schools that were labeled with 3 for both labels, meaning they exist in the top-25% for all three variables Chronically Absent Percent, FRPM_%, and Non-GCE Promote %, and given that these three features have some of the strongest correlation coefficients with ILEARN proficiency rates, it's not surprising that the chart below shows most of these corporations near the bottom.

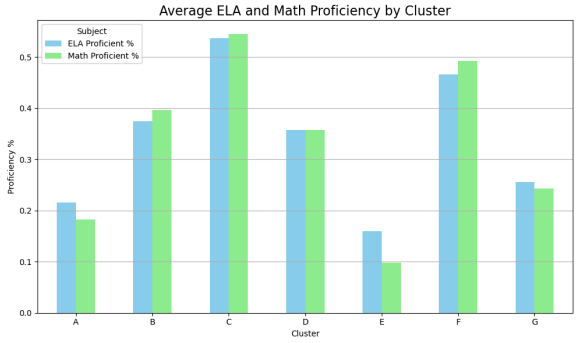


Clustering Analysis Results

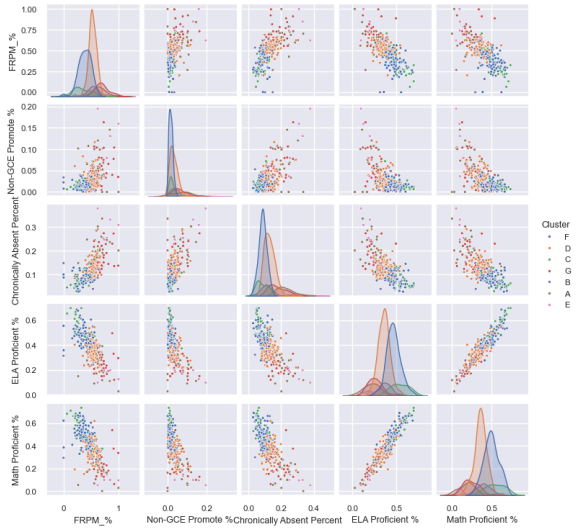
We also sought to explore the question "Which school corporations across the state are most alike? What defines their similarities?" through cluster analysis.

Classifying with KMeans clustering helped us gain significantly deeper insight into the highest and lowest performing corporations and the patterns that often lead toward success, within the features included in our dataset.

Most simply, the chart below conveys the ELA Proficiency % and the Math Proficiency % of each cluster. An interesting observation is that the corporations with high results on ILEARN tended to excel more in Math than in ELA, while those with lower rates of proficiency were stronger in ELA than Math.

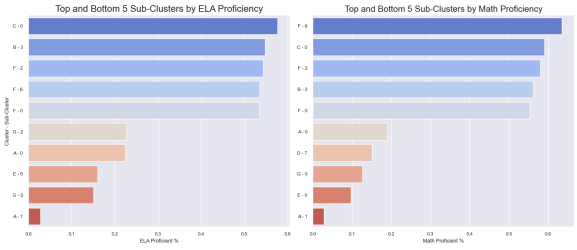


Using the three strongest correlated features with proficiency rates (Non-GCE Promote %, FRPM_%, and Chronically Absent Percent) explored previously, we can quickly see which clusters are in best position to excel in ILEARN – Clusters F, C, and D.



In the chart above, we also see the sizes of each cluster in relation to each other by analyzing the KDE charts. Clusters F and D have far larger curves than the other clusters.

The chart below demonstrates the granular detail that can be gained through splitting the clusters into sub-clusters, as the top-5 and bottom-5 sub-clusters are shown in terms of ILEARN proficiency. We will be discussing sub-cluster C0 momentarily.

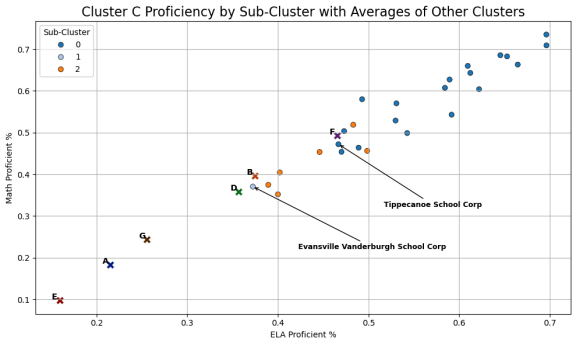


One helpful function of clustering the schools in this manner is encouraging collaboration or inspiration between corporations. For example, one cluster that is particularly interesting is Cluster C, which averages the highest ILEARN proficiency rates in the state. However, the average population of White students in these corporations is in the 23rd percentile of the state, which runs counter to the correlation map between features and proficiency rates. When split into sub-clusters, the variance within the cluster begins to take shape, as 19 schools (including its representative centroid school, Center Grove) fit into one of the highest performing sub-clusters in the state, while six fit into a second, lower-performing sub-cluster, and a single school (Evansville Vanderburgh) sits alone in a, relatively speaking, low-performing sub-cluster (Cluster C1).

Comparing Sub-Clusters within Cluster C

Cluster	C	C	C
sub_cluster	0	1	2
Non-GCE Promote %	32.67974	85.94771	34.64052
ELL %	73.85621	91.17647	62.0915
Special Education %	19.28105	74.18301	90.84967
FRPM_%	10.45752	54.24837	36.60131
Pct American Indian	65.68627	61.11111	57.84314
Pct Asian	96.07843	52.28758	61.11111
Pct Black	78.43137	77.45098	98.03922
Pct Hispanic	72.54902	87.5817	79.41176
Pct Multiracial	79.41176	59.80392	75.1634
Pct Pacific Islander	66.33987	95.42484	82.67974
Pct White	24.18301	100	71.24183
Chronically Absent Percent	11.76471	22.87582	16.99346
AVG_ATT_RATE	83.66013	45.75163	57.84314
Total Enrollment	91.17647	13.0719	39.86928
% Tested ELA	72.22222	99.34641	94.44444
% Tested Math	71.89542	59.47712	62.7451
Corp Per/Student	27.45098	59.15033	62.4183
ELA Proficient %	93.79085	59.47712	50.65359
Math Proficient %	90.84967	42.48366	66.99346

There is a clear divide between these schools, especially geographically, but they seem to be bonded by high enrollments, low rates of White students, high rates of ELL students, and high average attendance rates, while performing fairly high on ILEARN (Evansville Vanderburgh is in the 42nd percentile for both subjects). The chart below shows Cluster C broken down into its sub-clusters and plotted by proficiency compared to the average corporation of each other cluster.



What is shown is that even having been clustered together into one of the smaller clusters, there is still a

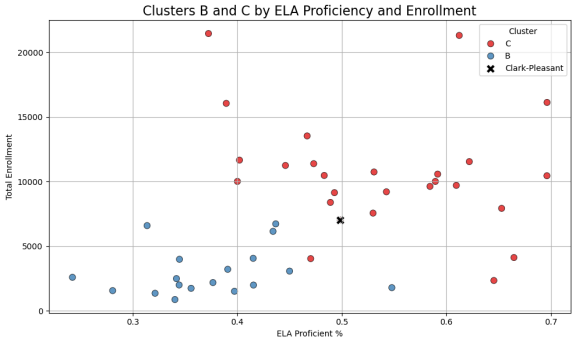
near-30% difference in rate of proficient students. Even being within the same cluster, though, one corporation might be surrounded by an entirely different community and/or have a different relationship with that community. The best places to begin looking for potential partners between corporations would most likely be the schools within the same cluster and slightly higher proficiency rates. Looking at this chart in particular, a school that might be fitting for Evansville Vanderburgh to study is Tippecanoe, which is a lower-performing Cluster C0 corporation with a similar profile at scale but is half the size. A particular area Tippecanoe could possibly aid Evansville Vanderburgh is in Non-GCE Promote rates, where Evansville Vanderburgh far outpaces the other schools in the cluster, with 7.2%. Tippecanoe is more than 50% lower than that figure, as are all of the other schools.

Additionally, once clusters are defined and understood, collaboration can happen outside of clusters. Continuing with Cluster C, Sub-Clusters 0 and 1 could most likely assist much of Cluster B, as both clusters are ethnically and economically most similar to each other and Cluster C has found slightly better performance on ILEARN. More specifically, Cluster C could first connect with Sub-Cluster 5 of Cluster B (four corporations), which has higher enrollments, low free/reduced lunch rates, and lower rates of White students. Clusters C0 and B5 would be productive collaborators, it appears, in an effort to increase Cluster B5's proficiency and decrease its Non-GCE Promotion rate.

Cluster B & C Percentiles of Mean Corporations

Cluster	B	C
sub_cluster	5	0
Non-GCE Promote %	75.1634	32.67974
ELL %	73.52941	73.85621
Special Education %	50.98039	19.28105
FRPM_%	21.56863	10.45752
Pct American Indian	51.30719	65.68627
Pct Asian	80.71895	96.07843
Pct Black	81.69935	78.43137
Pct Hispanic	91.50327	72.54902
Pct Multiracial	60.13072	79.41176
Pct Pacific Islander	47.38562	66.33987
Pct White	16.66667	24.18301
Chronically Absent Percent	55.55556	11.76471
AVG_ATT_RATE	50.3268	83.66013
Total Enrollment	83.00654	91.17647
% Tested ELA	71.89542	72.22222
% Tested Math	70.58824	71.89542
Corp Per/Student	50.3268	27.45098
ELA Proficient %	57.51634	93.79085
Math Proficient %	55.88235	90.84967

One key difference between Clusters B and C, though, is that Cluster B comprises much smaller schools, which can make collaboration challenging. Below is how ELA proficiency compares to Total Enrollment for Clusters C and B.



Cluster C's enrollment and proficiency are clearly greater than Cluster B's, but there are several Cluster C schools below the 8,000-students line that could potentially aid Cluster B in its proficiency. Most of those smaller Cluster C schools don't necessarily match Cluster B in ethnicity or economics, but Clark-Pleasant Community School Corporation (from Cluster C2 and marked in the chart above) southeast of Indianapolis is very close, in its free/reduced lunch rate, rate of White students, and chronic absentee rate.

Model-Based Performance Predictions

Following the completion of our cluster analysis, we developed a predictive modeling system that estimates district-level ELA and Math proficiency rates based on operational and demographic variables. The intent was to move from descriptive segmentation (e.g., which cluster a district falls into) to predictive analytics that can estimate how a district might perform given its specific characteristics.

We trained three types of regression models: Linear Regression, Decision Tree Regressor, and Random Forest Regressor. These models were built using a train-test split from our cleaned dataset, with the goal of capturing the relationship between 17 input variables and proficiency rates.

To ensure consistency, linear models were trained on scaled data using StandardScaler, while tree-based models (which are not sensitive to feature scaling) were trained on raw input values. We evaluated each model using RMSE and R^2, which showed that the Random Forest model achieved strong performance with minimal overfitting and maintained outputs within realistic bounds.

We then extended the models into a practical application: a tool that allows users to input values for all 17 variables and receive predicted proficiency rates from all three models. This enabled a way to simulate how a hypothetical district — or a real one being analyzed — might perform academically given its current structure.

To test the system, we used example inputs modeled after both strong and weak clusters identified earlier in the analysis. For example, we tested a strong district profile with low chronic absenteeism (11%), low FRPM (10%), and a per-student expenditure of $11,500. In contrast, we also input a weaker profile that matched characteristics seen in Cluster C3: high FRPM (82%), low attendance (86%), and a high percentage of special education and ELL students.

Across multiple test samples, Random Forest consistently outperformed the other models in both realism and accuracy. It rarely predicted values outside of 0–1, and it captured nonlinear interactions that the linear model missed. The Decision Tree was more interpretable but prone to overfitting on local patterns. The Linear Regression model was the most efficient, but it occasionally extrapolated outside of bounds, predicting percentages above 100% or below 0% in some extreme inputs — particularly when simulating very strong or weak districts.

The final takeaway is that this modeling approach pairs well with the clustering analysis. While the clusters help us group and compare districts by similarity, the predictive model allows us to test "what-if" scenarios: how might a district's performance change if its attendance improves? What happens if FRPM drops by 10%? This tool can support decision-making and resource planning by providing a fast and interpretable way to estimate likely outcomes.

Looking ahead

We only scratched the surface with this project, so below are some possibilities if we were to continue this work:

- We only used 18 variables for a discussion that is extremely complex. We could have added available (and not readily available) data that could have steered the project in another direction: disaggregated test results by ethnicity and gender, teacher-to-student ratio, SAT test results, ACT test results (only results before 2019 were available through IDOE), location median household income, and graduation rates.
- We primarily focused on how clusters can work together, within and without, to lift nearby corporations, but we didn't do too much investigating around why the two largest clusters – Cluster D (105 corporations) and Cluster F (102) – differed so significantly given their sizes. This would require digging into data not readily available to us.
- One particular school, Indiana Math and Science Academy, has an extremely interesting profile, in that it breaks from all understood trends within these features to achieve successful proficiency rates on ILEARN. Doing more investigation into what that school does could be helpful.
- We only scratched the surface of the work that can be done with cluster analysis. Given more time, we could compare centroids over time or how schools move between clusters over time. We could create a system where corporations are flagged once they fall into a cluster or sub-cluster and use the sub-clusters to measure growth within clusters.