Understanding Trends in School Grouping Using Clustering and a Visualization Tool

Steven Tang, Zhen Li, Zhen Gao

eMetric LLC

Paper written for the 2021 meeting of the National Council on Measurement in Education. The views expressed in this paper are solely those of the authors and they do not necessarily reflect the positions of eMetric LLC.

Correspondence concerning this paper should be addressed to Steven Tang, eMetric, 211 N Loop 1604 E, Suite 170, TX 78232. Email: steven@emetric.net.

Abstract

This paper investigates K-means, hierarchical, and density-based clustering on real testing data from hundreds of elementary schools and high schools from a single state. A "visual clustering" approach is proposed to allow stakeholders to engage with the clustering in real-time. Results from real-data analysis will be presented.

Key words: clustering, school accountability, visualization technique

Introduction

Every year, many students in the United States take end-of-year assessments. Policymakers and other stakeholders are keenly interested in understanding the results of these assessments, as high-stakes decisions are often made. In this paper, we propose a "visual clustering" approach aimed at providing invested stakeholders with additional useful information about their test results, both at the student and school levels. The purpose of clustering is to analytically determine a set of groups from data; analysts can then make interpretations and judgments about the quality and possible usefulness of the found groups. Visual clustering is scalable and can be performed in real-time, and therefore has the potential to be useful across different iterations of testing results across the country.

Few studies about clustering on educational data have been published in recent years. Beerenwinkel and von Arx (2017) applied clustering analyses to investigate the kinds of constructivist components and teaching patterns found in science education. Azarnoush et al. (2013) developed a clustering approach for segmenting the learners of online environments. Their approach can uncover subgroups within each cluster and highlight key characteristics of each cluster.

In this study, we first directly compare clustering methods to one another in terms of statistical properties by analyzing results on a set of multi-variate student test data. Then, we show the results of utilizing a visualization technique that could be used by stakeholders in the future. The research questions we seek to answer are:

1. How comparable are different clustering methods when applied to school testing data and results?
2. To what extent can cluster results be usefully visualized for general educational stakeholder usage?

Methodology

Data

The dataset in this paper comes from accountability measures collected in 3 school years (2017-18, 2018-19 and 2019-20) from 377 elementary schools and 115 high schools in a single state.

Table 1

Dataset Description

School Year	Number of Schools	Number of Variables	% Complete	Maximum Missing Rate
Elementary School				
2017-18	359	71	83%	4.46%
2018-19	377	68	79.60%	6.63%
2019-20	377	107	63.10%	26%
High School				
2017-18	112	118	31.30%	34.80%
2018-19	115	101	53%	18.30%

For school year 2017-18, the elementary school data contains 71 input variables in 9 categories. 83% of schools have complete data, and the highest missing rate of a variable is 4.46%. School year 2018-19 has 68 input variables and similar missing rates as those in school year 2017-18. For school year 2019-20, 107 variables are included. For 2019-20, new fiscal variables are added, but assessment-related variables are not available due to cancelled testing due to COVID-19. The high school data contains 118 input variables in 11 categories for 2017-18 and 101 input variables for 2018-19. For school year 2017-18, 31.3% of high schools have complete data, and the highest missing rate of a variable is 34.8%. For school year 2018-19, 53.0% of high schools have complete data. The highest missing rate of a variable is 18.3%. Mean imputation was applied to generate complete data for clustering analysis.

For a detailed list of variables for the 2017-2018 school year in this dataset, see Appendix III and Appendix IV. The full list of 2018-2019 and 2019-2020 variables can be made available by contacting the authors.

Clustering Approaches

Three clustering approaches are used and compared in this application: K-means, hierarchical, and density. K-means (Forgy, 1965) is the most widely-used clustering method. "K" indicates that the number of clusters need to be specified before clustering. "means" indicate that the groups will be defined according to the centroids of each group. K-means iteratively finds the best K centroids and assigns each observation to its nearest centroid's group. Another centroid-based approach is hierarchical agglomerative cluster analysis (Lance and Williams, 1967), which treats an individual as a cluster at the beginning and then joins similar individuals into clusters step by step. Additionally, we will also test the performance of density-based clustering (HDBSCAN, Campello, Moulavi, & Sander, 2013), which is known to be more efficient to detect arbitrary shaped clusters and outliers.

In the current study, we use "sklearn.cluster" and "hdbscan" packages in Python for clustering analysis. The number of clusters for K-means is fixed at 4. The linkage method for hierarchical clustering is "ward", while a hyperparameter "n_clusters" is set at 4 to extract a flat clustering from the dendrogram. Two hyperparameter for HDBSCAN are tuned and fixed at this level: min_cluster_size=5 and min_samples=1. The number of clusters and the percentage of schools included in HDBSCAN clusters vary across different input data sets.

Cluster results can be compared by analyzing internal clustering structure (Silhouette coefficient, Rousseeuw, 1987), looking at how similar two different clusters are (Jaccard Index, Halkidi, Batistakis, & Vazirgiannis, 2001), and performing a qualitative analysis on whether the visual representation of the clusters have intrinsic face validity.

Results

Different clustering algorithms can produce different groupings on the same set of data. Choosing which clustering approach to use requires context of the research inquiry at hand. In this section, we directly compare three clustering approaches to one another across 15 sets of variables, spanning data from both elementary schools and high schools. The 15 sets are intended to represent a sample of possible research inquiries. The 15 sets of variables are shown in detail in Appendix II.

Result 1 – Silhouette Coefficients

The Silhouette coefficient (Rousseeuw, 1987) describes internal clustering structure. The Silhouette coefficient is the mean of taking the silhouette index over all points in the data. The larger the Silhouette coefficient, the greater the distance between points within a cluster compared to points outside of that cluster. Larger Silhouette coefficients indicate larger separation between clusters relative to the average distance of points within a cluster. The formula to compute Silhouette index and Silhouette coefficient can be found in Appendix I.

Table 2 provides the Silhouette coefficients for all 15 sets of variables across the 3 clustering approaches using Euclidean distance. K-means had the highest overall average of Silhouette coefficients, achieving the highest coefficient in 8 out of the 15 variable sets. This indicates that K-means is relatively the best performer, but in 7 out of 15 sets, one of the other clustering approaches achieved a higher coefficient. HDBSCAN had interesting results that varied more greatly relative to the other approaches. For *elem_8v*, *elem_10v_ns, elem_2y_ns, elem_3y_ns*, HDBSCAN had the highest performance by a wide margin. However, in other variable sets, such as *high_all*, HDBSCAN had very poor performance. Note that HDBSCAN is reported in two columns; this is because HDBSCAN inherently incorporates a "noise" factor, where some schools may be discarded from its clustering if the school is not close enough to another school in a cluster. From these results, K-means could be a reasonable first-choice approach, given both its simplicity and its effectiveness in finding internal clustering structure. HDBSCAN may produce good results for some variable sets; it may require more careful tuning and attention and may not be applicable to every variable set.

Table 2

Silhouette Coefficients – Euclidean Space

	K-means (k=4)	Hierarchical (#cluster=4)	HDBSCAN_Allnodes (#cluster)	HDBSCAN_Clusters (% data included)
elem_nspf	0.19	0.15	-0.08 (4)	0.04 (68.0)
elem_6v	0.23	0.24	0.16 (3)	0.20 (90.3)
elem_8v	0.28	0.27	0.59 (2)	0.59 (100)
elem_2y_score	0.26	0.19	0.25 (3)	0.31 (89.7)
elem_3y_score	0.17	0.16	0.14 (3)	0.23 (84.1)
elem_5v_ns	0.34	0.34	-0.03 (11)	0.15 (74.1)
elem_10v_ns	0.18	0.13	0.49 (2)	0.49 (99.4)
elem_2y_ns	0.20	0.19	0.36 (2)	0.39 (96.8)
elem_3y_ns	0.19	0.2	0.34 (3)	0.40 (89.9)
high_nspf	0.37	0.34	0.26 (5)	0.40 (82.1)
high_10v_ns	0.29	0.26	0.02 (3)	0.33 (54.5)
high_10v	0.23	0.2	0.21 (2)	0.30 (86.6)
high_2y	0.22	0.19	0.26 (3)	0.29 (94.0)
elem_all	0.15	0.11	0.28 (2)	0.29 (97.2)
high_all	0.14	0.13	-0.12 (4)	0.20 (29.5)
Overall Average	0.23	0.21	0.21	0.31

Table 3 shows the Silhouette coefficients for each clustering approach using Mahalanobis space (Mahalanobis, 1936). *elem_all* and *high_all* include linearly dependent variables; these two sets are excluded from the table. Mahalanobis distance adjusts for covariance in data and is commonly used when multi-variate data is assumed to have covariance. Hierarchical is relatively the best performer, although it was only the highest score in 5 of the 15 variable sets. However, in those 5 sets, it sometimes greatly outperformed the other clustering approaches, leading to the highest overall average score. This could indicate that if a practitioner decides to use the Mahalanobis transformation, the practitioner should be aware that Hierarchical clustering can greatly outperform K-means in some circumstances. Similar to the Euclidean results, HDBSCAN's performance depends greatly on the variable set chosen. K-means seems relatively more stable than HDBSCAN across the variable sets.

Table 3

Silhouette Coefficients - Mahalanobis Space

	K-means (k=4)	Hierarchical (#cluster=4)	HDBSCAN_allnodes(#cluster)	HDBSCAN_clusters (% data included)
elem_nspf	0.08	0.04	-0.19 (9)	0.10 (27.6)
elem_6v	0.13	0.1	0.13 (2)	0.19 (86.9)
elem_8v	0.17	0.1	0.56 (2)	0.57 (99.7)
elem_2y_score	0.18	0.17	0.31 (3)	0.36 (91.0)
elem_3y_score	0.16	0.25	0.00 (2)	0.18 (56.1)
elem_5v_ns	0.3	0.29	-0.03 (9)	0.09 (79.9)
elem_10v_ns	0.12	0.11	0.48 (2)	0.49 (99.4)
elem_2y_ns	0.17	0.16	0.23 (3)	0.29 (91.3)
elem_3y_ns	0.13	0.39	0.31 (3)	0.37 (89.9)
high_nspf	0.08	0.06	0.02 (2)	0.18 (60.7)
high_10v_ns	0.11	0.28	0.12 (2)	0.37 (59.8)
high_10v	0.1	0.27	0.13 (2)	0.26 (76.8)
high_2y	0.12	0.1	-0.08 (5)	0.25 (40.5)
Overall Average	0.14	0.18	0.15	0.29

Result 2 – Jaccard Similarity Indices

The Jaccard similarity index (Halkidi et al., 2001) can be used to gauge the similarity and diversity of two clustering results. Suppose there is one cluster *a* from one clustering approach, and one cluster *b* from another clustering approach. The Jaccard similarity index is defined as:

$$Jaccard_{cluster\_a-cluster\_b} = \frac{|Cluster\_a \ \cap Cluster\_b|}{|Cluster\_a \ \cup Cluster\_b|}$$

A high Jaccard similarity index means that cluster *a* and cluster *b* have a high proportion of schools in common. This indicates that both clustering methods have a cluster with similar properties, indicating that both methods have clustered together a similar set of schools.

To compare algorithms to one another, we seek to create an "aggregate" similarity score. For example, when K=4 in K-Means clustering, there are 4 clusters. These 4 clusters can each be compared to the clusters found in HDBSCAN. If HDBSCAN found 3 clusters, then there will be a total of 12 Jaccard similarity indices. For each of the 4 K-means clusters, there are 3 Jaccard similarity indices. To create an aggregate statistic, we take the highest similarity score for each cluster and average them together, treating one of the algorithms as the "base" algorithm. KM stands for K-means, HI stands for Hierarchical, and HD stands for HDBSCAN. KM–HI indicates the average of the highest Jaccard indices for each of the KM clusters comparing to HI clusters; KM is considered the base algorithm. Conversely, HI-KM indicates the average of the highest Jaccard indices for each of the HI clusters comparing to KM clusters, with HI considered as the base algorithm. Note that this may not be symmetric if the number of clusters is not symmetric, or if there are two cluster pairs that differ in ranking depending on which algorithm is treated as the base algorithm.

Table *4* shows the aggregate Jaccard similarity scores comparing each of the 3 algorithms pairwise. KM-HI and HI-KM have the highest average aggregate Jaccard scores, indicating a relatively high level of agreement between the two approaches, which indicates that the clusters produced by both algorithms tend to be similar. In our results, the agreement reached as high as .938 between KM and HI, and reached .285 at its lowest. 11 out of the 15 variable sets had at least a .6 aggregate Jaccard similarity between KM and HI, showing that most of the time there is a high level of agreement. For HD, the results were quite different. For 12 out of the 15 variable sets, HD had lower than a 0.5 aggregate Jaccard Score with one of the other approaches. There could be two key reasons why HD differs. Firstly, HD is density based, while KM and HI are not. Secondly, HD's number of clusters is not fixed, while we fixed KM and HI to 4. The aggregate Jaccard Scores may have higher values if we were to first run HD, and then fix the number of clusters in KM and HI to the same number of clusters found in HD.

Table 4

Aggregate Jaccard Scores Between Algorithms

	KM-HI	HI-KM	KM-HD	HD-KM	HI-HD	HD-HI
elem_nspf	0.374	0.37	0.225	0.129	0.277	0.203
elem_6v	0.654	0.654	0.27	0.195	0.286	0.223
elem_8v	0.86	0.86	0.5	0.714	0.5	0.7
elem_2y_score	0.387	0.395	0.263	0.205	0.239	0.177
elem_3y_score	0.693	0.693	0.236	0.197	0.235	0.211
elem_5v_ns	0.872	0.872	0.479	0.254	0.486	0.266
elem_10v_ns	0.617	0.603	0.5	0.724	0.5	0.762
elem_2y_ns	0.71	0.71	0.247	0.241	0.28	0.296
elem_3y_ns	0.285	0.355	0.255	0.263	0.321	0.253
high_nspf	0.812	0.812	0.71	0.601	0.754	0.65
high_10v_ns	0.938	0.938	0.215	0.363	0.22	0.375
high_10v	0.712	0.712	0.385	0.486	0.377	0.472
high_2y	0.795	0.795	0.31	0.263	0.349	0.337
elem_all	0.511	0.55	0.248	0.225	0.248	0.239
high_all	0.736	0.736	0.175	0.229	0.142	0.228
Average	**0.664**	**0.670**	**0.335**	**0.339**	**0.348**	**0.359**

Result 3 – Case Study

The previous 2 results gave results for all 15 variable sets. This gives a general sense of how the 3 clustering algorithms perform relative to each other. However, clustering is usually very contextual, and the results and interpretations are often dependent on the exact variables and data distributions at hand. In this section, we analyze one of the variable sets more deeply.

The chosen variable set for this case study contains 12 variables from 2017-18 high school data. Table 5 show the descriptive statistics of the clustering variables.

Table 5

Descriptive Stats for 12 Clustering Variables

	Valid N	%Missing	Mean	S.D.	Min	Max	Median
2017-18 Math Mean Scale Score	112	0.0%	17.9	2.6	13.5	33.8	17.7
2017-18 ELA Mean Scale Score	112	0.0%	16.4	3.1	9.6	30.8	16.2
2017-18 Science Mean Scale Score	112	0.0%	17.8	2.7	12.4	32.8	17.6
2017-18 Chronic Absenteeism Rate	111	0.9%	26.5	15.9	0.0	92.8	23.9
2017-18 4-Year Graduation Rate	112	0.0%	86.0	17.8	22.2	100.0	91.3
2017-18 Post-Secondary Preparation Participation %	112	0.0%	61.2	25.9	0.0	100.0	63.3
2017-18 Post-Secondary Preparation Completion %	112	0.0%	36.8	29.1	0.0	100.0	30.4
2017-18 % of Graduates Receiving an Advanced Diploma	112	0.0%	30.8	19.3	0.0	100.0	29.4
2017-18 # of 9th Grade Credit Sufficient Students	107	4.5%	278.9	249.0	0.0	819.0	227.0
2017-18 # of Graduates	112	0.0%	246.2	221.0	5.0	722.0	179.5
2017-18 # of Graduates Receiving a Standard Diploma	112	0.0%	171.4	168.7	0.0	572.0	107.0
2017-18 # of Graduates Receiving an Advanced Diploma	112	0.0%	74.0	74.4	0.0	322.0	57.5

In Table 5, 10 of the 12 clustering variables have complete data. 0.9% of schools have missing values on one variable (Chronic Absenteeism Rate) and 4.5% of schools have missing values on another variable (# of 9th Grade Credit Sufficient Students). The clustering variables include mean scores from summative testing, attendance, graduation, college and career readiness. All variables are standardized for clustering analysis.

Table 6, Table 7, and

Table 8 presents the Jaccard similarity indexes between each pair of clusters among the three clustering approaches. The highlighted cells are the most similar clusters. In Table 6, each of the four clusters by K-means has a corresponding most similar cluster from the four clusters by Hierarchical clustering; this correspondence is symmetrical. The Jaccard similarity indexes range from 0.619 to 0.929. In Table 7, K-means is compared to HDBSCAN. HDBSCAN determines the number of clusters algorithmically, rather than user defined like K-means. Comparing the values between Table 6 and Table 7, it appears that *cluster_1* in both cases appear to have an exact match, meaning that *cluster_1* in both Hierarchical and HDBSCAN appear to be very similar, if not identical. It seems that the 5th cluster from HDBSCAN is represented by *cluster_2* and *cluster_4* from K-means. Thus, the overall structure of the HDBSCAN clusters is comprised of *cluster_1*, *cluster_3*, and parts of clusters 2 and 4 are broken up to form a 5th cluster. This shows that HDBSCAN may recover some structures identically from K-means and Hierarchical, while still forming different structures with the remaining data.

Table 8 confirms that *cluster_1* of HDBSCAN and Hierarchical are identical, with a Jaccard similarity index of 1. *cluster_4* and *cluster_5* are the least represented from HDBSCAN, with Jaccard similarity indices of .238 and .562 respectively.

Table 6

Jaccard similarity index for each cluster of K-means and Hierarchical clustering

	Hierarchical Clustering			
K-means	cluster_1	cluster_2	cluster_3	cluster_4
cluster_1	0.857	0	0.037	0
cluster_2	0	0.929	0	0
cluster_3	0	0	0.844	0.056
cluster_4	0	0.053	0.034	0.619

Table 7

Jaccard similarity index for each cluster label of K-means and HDBSCAN

	HDBSCAN				
K-means	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5
cluster_1	0.857	0	0	0	0
cluster_2	0	0.744	0	0.167	0
cluster_3	0	0	0.738	0	0
cluster_4	0	0	0.02	0.12	0.5

Table 8

Jaccard similarity index for each cluster label of HDBSCAN and Hierarchical clustering

	Hierarchical Clustering			
HDBSCAN	cluster_1	cluster_2	cluster_3	cluster_4
cluster_1	1	0	0	0
cluster_2	0	0.69	0	0
cluster_3	0	0	0.762	0
cluster_4	0	0.238	0	0
cluster_5	0	0	0	0.562

Figure 1 shows 3 plots, whereby each dot represents a high school. The 12 clustering variables are reduced using a dimensionality reduction technique known as T-SNE (van der Maaten and Hinton, 2008). This is used to visualize each school on a 2-dimensional plot, and we can then pinpoint exactly which schools were labeled differently by different algorithms. *Cluster_1* (blue nodes) shows high consistency across all 3 approaches. *Cluster_4* (purple nodes) in K-means and Hierarchical show the least consistency between the two plots, with some schools being labelled differently on different edges of the clustering space. In HDBSCAN, several of the schools are labeled as *noise*, meaning that the clustering algorithm chooses not to assign a label to these nodes because they are not similar enough to a neighboring school. Perhaps not surprisingly, it can be clearly seen that many of the schools that K-means and Hierarchical disagree on are labeled as noise by HDBSCAN, indicating these schools are harder to categorize.

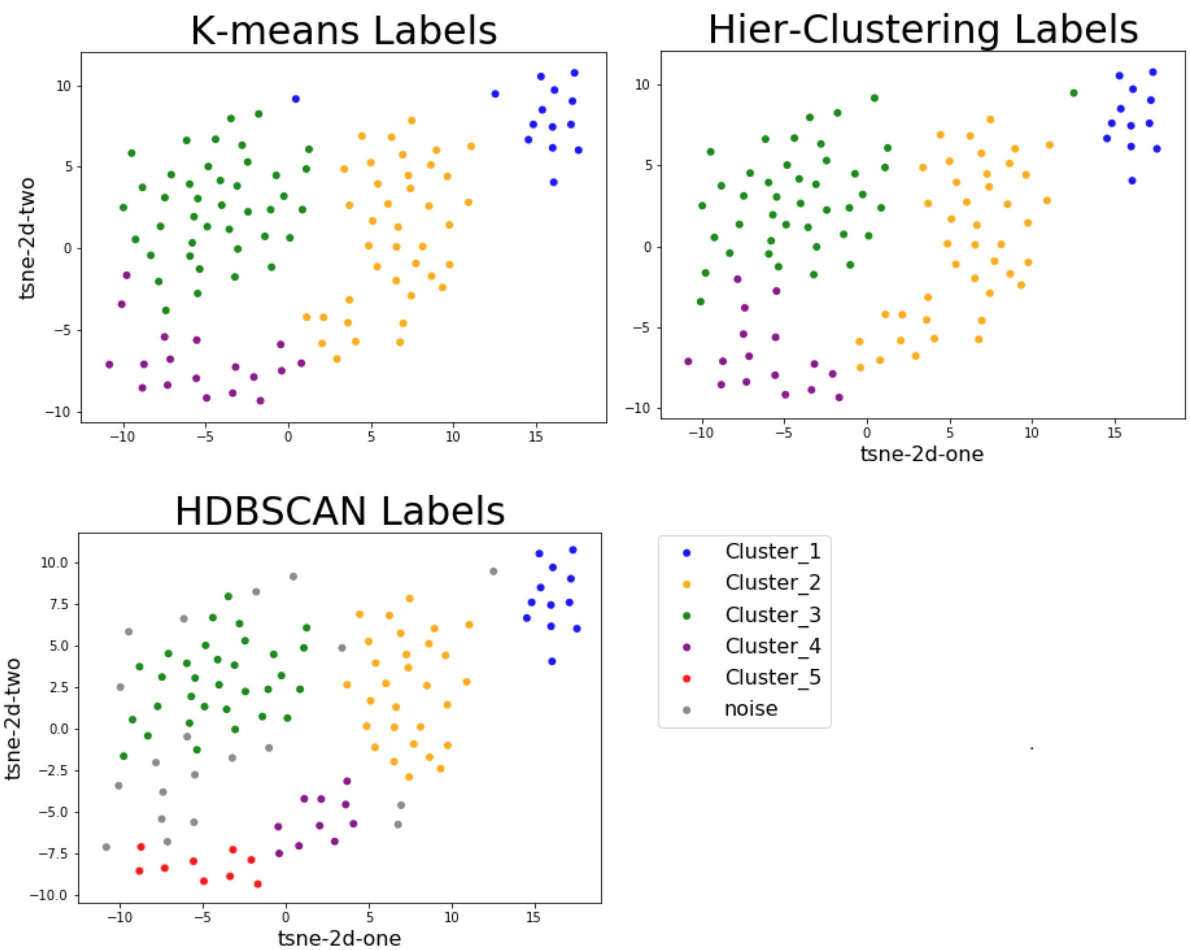

Figure 1

T-SNE plot of all schools by different cluster labels

Result 4 – Visualization Methods

One of the motivating factors behind this research paper is to eventually enable users interested in educational measurement to make use of clustering in real time. A crucial benefit of clustering algorithms is that they can often be visualized in real time, meaning that users can make custom reports on the fly to suit their exact needs. In this section, we look at two ways of visualizing results.

Visualization Method 1

We introduce a tree-map-based visualization aiming to provide an easy-to-understand clustering result with rich information. There are two types of nodes in the visualization - the cluster node, and the school node. The cluster nodes are represented by squares with dashed borders, and, the school nodes are represented by circles where the color represents the school

Star Rating, and the size of the node indicates the N count of students. The summary information of a cluster is shown on the top of the cluster node. The visualization algorithm takes the input of the clustering result of a clustering algorithm, and generates the graph in GML (Graph Modeling Language) format. The layout of nodes is processed with the SBGN (https://sbgn.github.io/sbgn) algorithm, which aims to provide a standardized graphical notation for molecular and system-biology applications that describe biological pathways/networks.

The legend for the following 3 figures is:

The star rating is an overall measure (determined by the state's department of education) of how well the school is doing. The schools are colored by their star rating in the following plots. *Figure 2*, *Figure 3*, and *Figure 4* show examples of using this visualization technique for K-means, HDBSCAN, and Hierarchical respectively. Schools were clustered on the 12 variables from 2017-18 high school data (as shown in Table 5). The size of each node corresponds to the overall number of students enrolled at that school. By looking at the number of schools, average star ratings, and average group values of clustering variables listed in the headers of each cluster block, the user can form a hypothesis about how to characterize each found cluster.

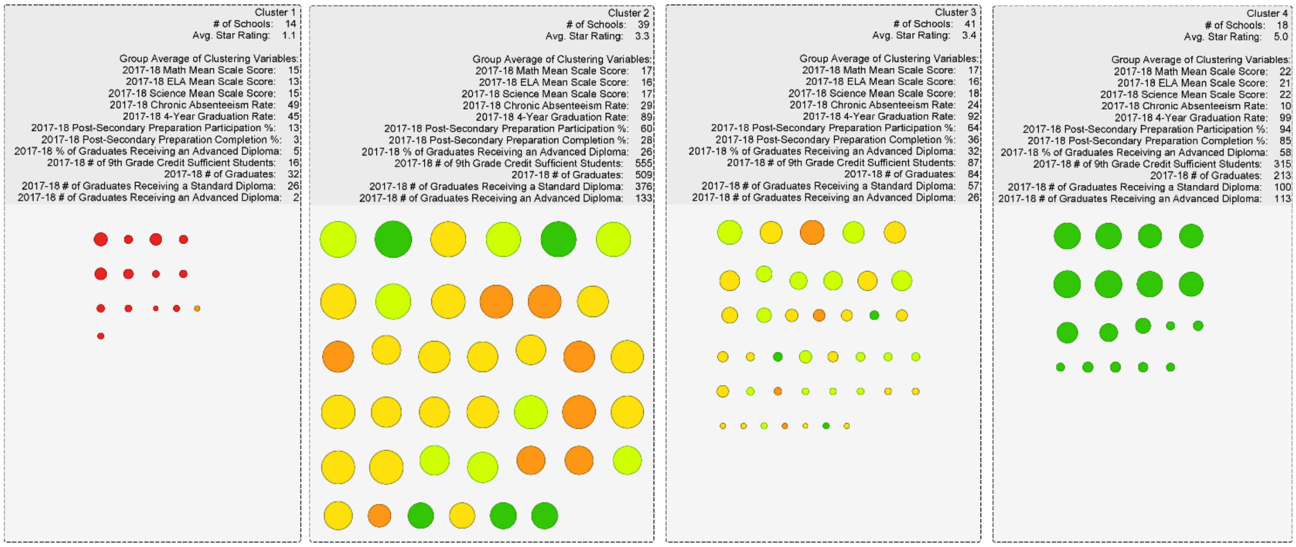

Figure 2

K-Means Visualization Example

In Figure 2, all 112 high schools are grouped into 4 clusters using the K-means algorithm. The first group contains almost all 1-star schools, while the 4th group contains only 5-star schools. The mean scale scores for Math, ELA, and Science of schools in *Cluster_1* is obviously lower than the mean scale scores of schools in *Cluster_4*. Both *Cluster_2* and *Cluster_3* contains schools with various star ratings. By looking at the averages of clustering variables, we find that the average values of # of graduates differ significantly between the two clusters.

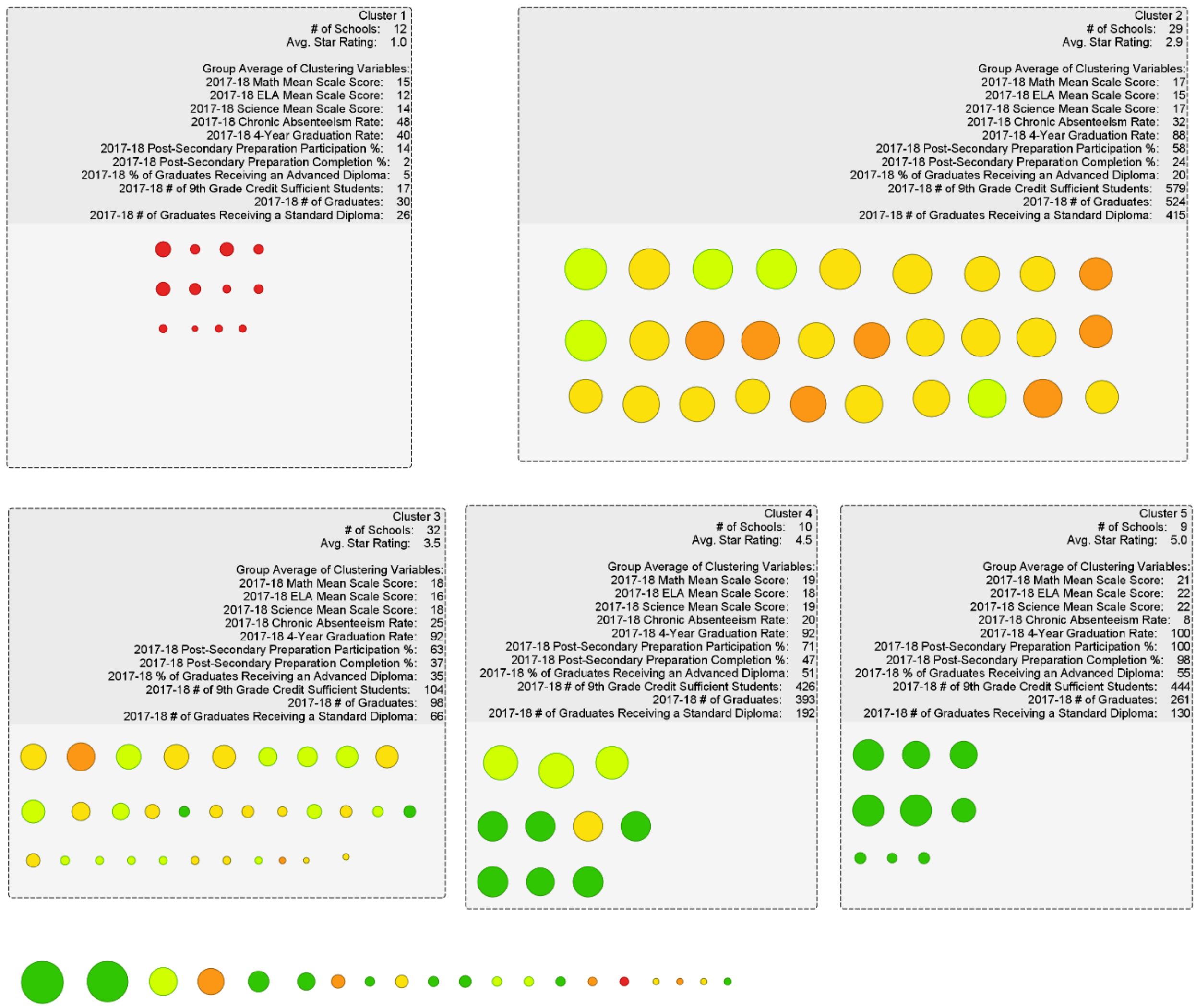

Figure 3

HDBSCAN Visualization Example

In this example, five school groups are identified by HDBSCAN clustering, while 20 out of the 112 high schools are not clustered into any group. The "Noise" schools are lined up at the bottom of the plot. Not surprisingly, the 12 schools in *Cluster_1* are all 1-star schools, while the 9 schools in in *Cluster_4* are all 5-star schools. The majority of schools in *Cluster_2* are 2-star and 3-star schools with a relatively large school size, but also include some 4-star schools. The average star rating of *Cluster_2* is 2.9. *Cluster_3* has schools with smaller school size, including 4-star schools, 3-star schools, and a few 2-star schools. *Cluster_4* is similar to *Cluster_5*, with schools whose mean scale scores are slightly lower.

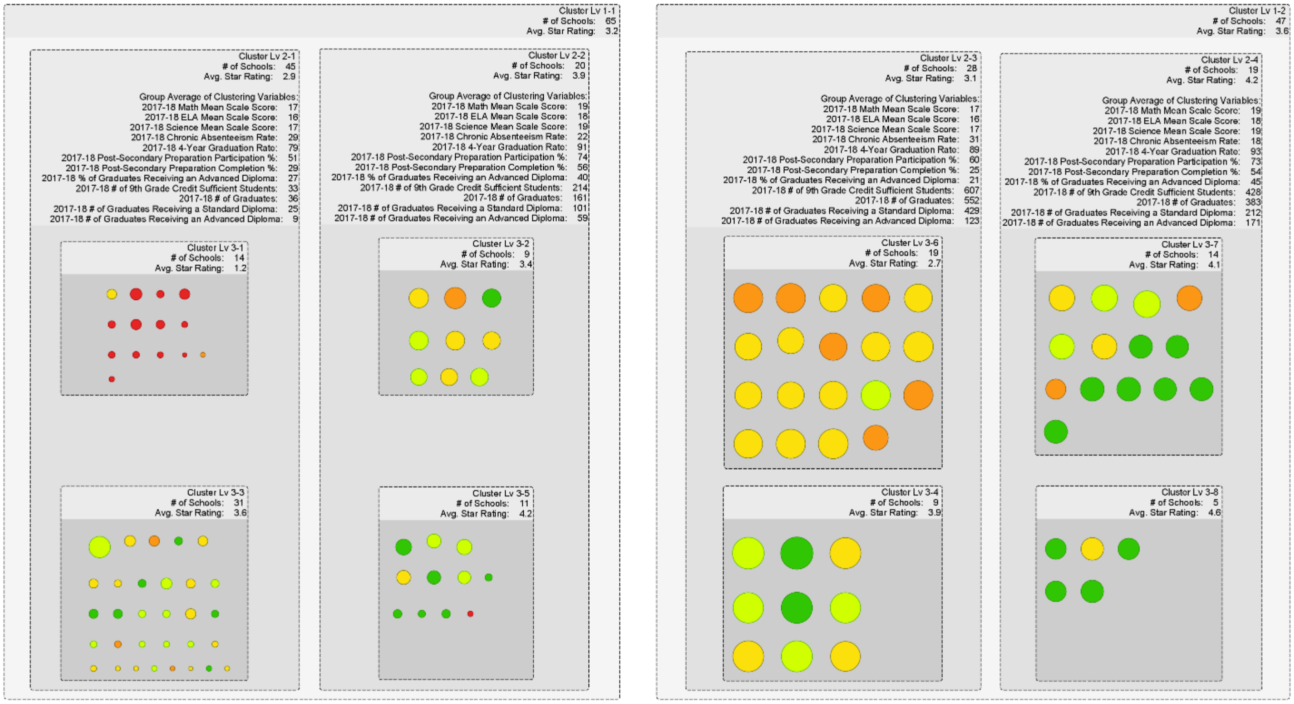

Figure 4

Hierarchical Clustering Visualization Example

The visualization of hierarchical clustering put schools in nested boxes. In this example, there are 3 levels of clusters. The highest level contains two clusters; the middle level contains 4 clusters; and the lowest level contains 8 clusters. Simply speaking, schools are divided into two smaller groups at each level.

The advantage of hierarchical clustering visualization is to identify smaller groups of similar schools within a large cluster. In this example, we could see that Cluster lv 2-1 contains schools with relatively small school sizes. In addition, these schools could be further clustered into two subgroups, one with higher star ratings and one with lower star ratings.

Visualization Method 2

As part of our research, we developed an "Analytics Lab" tool where users can actively select which variables they want clustered, even from multiple years. *Figure 5* shows a truncated screenshot of the options selection page. Users can select up to 10 variables they want clustered. Using the case study from earlier, we select the first 10 out of the 12 variables to visualize in our clustering engine.

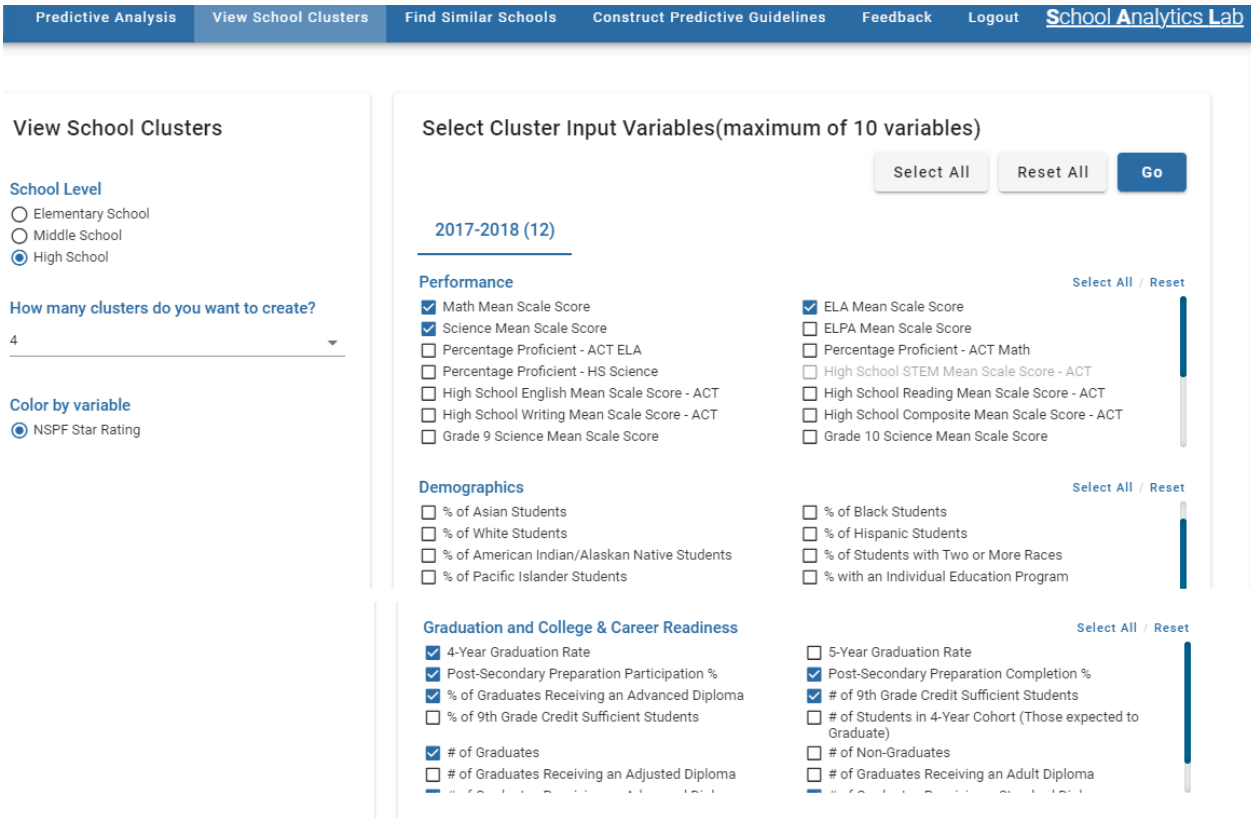

Figure 5

School Analytics Lab Variable Selection Page

Figure 6 shows a screenshot of the clustering visualization using the 10 high school variables. Additional information can be made available by contacting the authors.

The goal of the clustering visualization is to let users see their data succinctly, notice trends, and be able to quickly dive deeper into the data by interactively selecting clusters, nodes, and searching for individual schools. Outliers and trends stand out using this interactive framework. Any combination of variables can be chosen, allowing for continuous exploration. As the field of educational measurement continues to obtain more and more data, tools that enable interesting and useful explorations have increasing value, and we hope that the research in this paper sparks interest in applying clustering visualizations to better understand data.

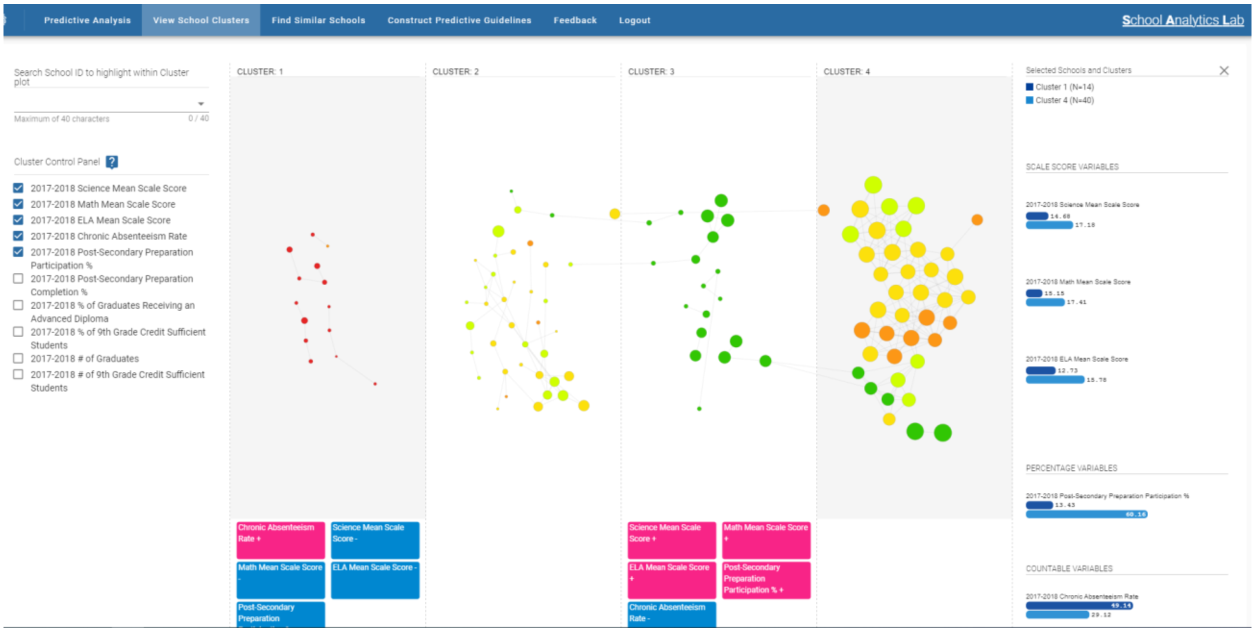

Figure 6

School Analytics Lab Clustering Visualization Screen Shot

Discussion and Conclusion

In this study, we explore different ways to cluster and visualize elementary and high schools in a state. First, three clustering approaches are compared using 15 sets of school data variables. K-means was found to have the highest overall Silhouette scores in Euclidean space, but K-means was not universally always the highest, with the other two clustering approaches, HDBSCAN and Hierarchical clustering, sometimes achieving the best result depending on the variable set chosen. Second, the degree of similarity between all 3 clustering approaches was compared pairwise. We found that K-means and Hierarchical clustering had stronger agreement compared to HDBSCAN. This is expected because HDBSCAN is a density-based clustering approach, which has a different way of defining clusters from the centroid-based clustering approaches. However, our analysis is somewhat limited since we did not fix the number of clusters to be equal, so it is almost a given that HDBSCAN would have lower agreement if HDBSCAN has a varying number of clusters. Future work can fix the number of clusters to be consistent in comparison. Third, we looked more deeply at a particular case study, particularly exploring how the “additional” cluster from HDBSCAN is comprised of parts of the other clusters from both K-means and Hierarchical approaches. A T-SNE plot was generated to give a visual representation of when the algorithms agreed and disagreed, especially showing how the “noise” component of HDBSCAN often coincides with disagreements between K-means and Hierarchical. Finally, A “visual clustering” tool is proposed. We showed two examples of how the clustering approaches described in this paper could be visualized, potentially for widespread use cases where many practitioners can create interactive visual plots to perform exploratory data analysis with.

It is our hope to continue this line of clustering visualization research with the purpose of making clustering analysis available to stakeholders interested in exploring educational measurement data to help inform decision making. Clustering tools, particularly refined to the needs of educational assessment and measurement needs, could see possible use whenever practitioners need to understand how groups are forming relative to variables of interest.

Reference

Azarnoush, B., Bekki, J. M., Runger, G.C., Bernstein, B.L., Atkinson, R.K. (2013) Toward a framework for learner segmentation. *Journal of Educational Data Mining*, 5(20).

Beerenwinkel, A. & Von Arx, M. (2017). Contructivism in practice: an exploratory study of teaching patterns and student motivation in physics classrooms in Finland, Germany and Switzerland. *Research in Science Education*, 47, pp. 237–255

Campello R.J.G.B., Moulavi D., Sander J. (2013) Density-Based Clustering Based on Hierarchical Density Estimates. In: Pei J., Tseng V.S., Cao L., Motoda H., Xu G. (eds) *Advances in Knowledge Discovery and Data Mining*. *PAKDD 2013.* Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21, 768–9

Halkidi M, Batistakis Y, Vazirgiannis M. (2001). On clustering validation techniques. J. *Intell. Inf. Syst.* 17:107–145.

Lance, G. N., and Williams, W. T. (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems. Comput. J. 9, 373–380

Mahalanobis, P.C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India.* 2(1): 49–55.

Rousseeuw, P. J. *(1987).* Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis*. Computational and Applied Mathematics. 20: 53–65.*

van der Maaten, L.J.P. and Hinton, G.E. (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research.* 9:2579-2605,

Appendix I

Silhouette index Formula (Rousseeuw, 1987)

Suppose schools have been clustered into *K* clusters: $a_1, \ldots a_m, \ldots, a_K$.

For data point $i \in a_m$ (data point i in the cluster a_m), let

$$d_w(i) = \frac{1}{|a_m| - 1} \sum_{j \in a_m\, i \neq j} distance(i, j)$$

be the mean distance between i and all other data points in the same cluster, where $distance(i, j)$ is the distance between data points i and j in the cluster a_m. In one word, $d_w(i)$ measures how well i is assigned to its cluster (the smaller the value, the better the assignment).

Next, let *k* be any value in (1,…, *K*) except for m. The mean dissimilarity of point i to cluster a_k is defined as the mean of the distance from i to all points in a_k. For each data point $i \in a_m$:

$$d_b(i) = min(\frac{1}{|a_k|} \sum_{j \in a_k} distance(i, j))$$

$d_b(i)$ is the smallest mean distance of i to all points in any other cluster ($a_k \neq a_m$). The cluster with this smallest mean dissimilarity is said to be the "neighboring cluster" of i because it is the next best fit cluster for point i.

At last, Silhouette index for data point i is computed as following:

$if |a_m| > 1$:

$$silhouette\_index(i) = \begin{cases} 1 - \frac{d_w(i)}{d_b(i)}, & if\ d_w(i) < d_b(i) \\ 0, & if\ d_w(i) = d_b(i) \\ \frac{d_b(i)}{d_w(i)} - 1, & if\ d_w(i) > d_b(i) \end{cases}$$

$if\ |a_m| = 1$:

$$silhouette\_index(i) = 0,$$

From the above definition it is clear that

$$-1 \leq silhouette\_index(i) \leq 1$$

The mean of $silhouette\_index(i)$ over all points of a cluster is a measure of how tightly grouped all the points in the cluster are. The Silhouette coefficient is the mean of taking the silhouette index over all points in the data.

Appendix II

15 Sets of Clustering Variables

	Set Name	Variables
1	elem_nspf	1. Math Mean Scale Score 2. ELA Mean Scale Score 3. Science Mean Scale Score 4. Percent Proficient - Read By Grade 3 5. Math Gap % 6. ELA Gap % 7. Math Growth (MGP) 8. ELA Growth (MGP) 9. English Language Proficiency Growth (MGP) 10. Chronic Absenteeism Rate
2	elem_6v	1. Math Mean Scale Score 2. ELA Mean Scale Score 3. Science Mean Scale Score 4. Average Daily Attendance 5. PPE Leadership % 6. # Of Computers per Student
3	elem_8v	1. Percentage Proficient Math 2. Percentage Proficient ELA 3. Percentage Proficient Science 4. % English Learners 5. % FRL 6. PPE Operations % 7. PPE Instruction % 8. Teacher Average Daily Attendance
4	elem_2y_score	1. 2017-18 Math Score 2. 2017-18 PPE Leadership % 3. 2017-18 # of Incidents of Violence to Other Students 4. 2018-19 ELA Score 5. 2018-19 % Hispanic Students 6. 2018-19 # of Computers per Student
5	elem_3y_score	1. 2017-18 Chronic Absenteeism 2. 2017-18 # of Incidents of Violence to Staff 3. 2017-18 Total # of Long-Term Substitute Teachers 4. 2018-19 Science Mean Scale Score 5. 2018-19 Overall Total Spending Per Pupil 6. 2018-19 Transiency Rate 7. 2019-20 Federal – Overall Total Spending Per Pupil 8. 2019-20 # of Teachers Teaching Out of Field

		9. 2019-20 # of Inexperienced Teachers
6	elem_5v_ns	1. Professional Development Funding 2. Average Daily Attendance 3. Transiency Rate 4. # of Teach Coachers per Student 5. % of Elementary Classes Not Taught by Highly Qualified Teachers
7	elem_10v_ns	1. % Male 2. % Asian 3. % FRL 4. % IEP 5. PPE Instruction % 6. Student/Teacher Ratio 7. Transiency Rate 8. # Of Incidents of Violence to Other Students 9. Total # of Short-Term Substitute Teachers 10. Teacher Average Daily Attendance
8	elem_all	All Variables from 2018 Elementary School Data
9	elem_2y_ns	1. 2017-18 Student Teacher Ratio 2. 2017-18 Chronic Absenteeism Rate 3. 2017-18 # Of Mobile Learning Devices 4. 2018-19 Student/Teacher Ratio – 4th grade 5. 2018-19 PPE Instruction Support % 6. 2018-19 % of Students with Two or More Races
10	elem_3y_ns	1. 2017-18 Transiency Rate 2. 2017-18 PPE Operations % 3. 2017-18 Overall Total Spending Per Pupil 4. 2018-19 % of Black Students 5. 2018-19 Student/Teacher Ratio – 5th Grade 6. 2018-19 # of Teach Coaches per Student 7. 2019-20 Total # of Long-Term Substitute Teachers 8. 2019-20 State/Local – Instruction Spending Per Pupil – Personnel 9. 2019-20 % of Computers 5 Years or Newer
11	high_nspf	1. Math Mean Scale Score 2. ELA Mean Scale Score 3. Science Mean Scale Score 4. Chronic Absenteeism Rate 5. 4-Year Graduation Rate 6. Post-Secondary Preparation Participation % 7. Post-Secondary Preparation Completion % 8. % of Graduates Receiving an Advanced Diploma 9. # of 9th Grade Credit Sufficient Students

12		10. # of Graduates 11. # of Graduates Receiving a Standard Diploma 12. # of Graduates Receiving an Advanced Diploma
12	high_10v_ns	1. Dropout Rate 2. # of Math Classes Not Taught by Highly Qualified Teachers 3. # of Long-Term Substitute Teachers – ELA 4. 4-Year Graduation Rate 5. 5-Year Graduation Rate 6. # of Graduates Receiving an Adult Diploma 7. Average Class Size: Math 8. Average Class Size: English 9. Transiency Rate, % of Pacific Islander Students 10. % of Students Receiving Free or Reduced-Price Lunch
13	high_10v	1. Percentage Proficient – ACT Math 2. Percentage Proficient – HS Science 3. Grade 9 Science Mean Scale Score 4. Math Mean Scale Score 5. ELA Mean Scale Score 6. % of Students Eligible for FRL 7. % of Students Receiving FRL 8. PPE Instruction % 9. Chronic Absenteeism Rates 10. Teacher Average Daily Attendance
14	high_2y	1. 2017-18 Percentage Proficient - ACT ELA 2. 2017-18 % of English Learners 3. 2017-18 Chronic Absenteeism Rate 4. 2017-18 # of Incidents of Violence to Other Students 5. 2017-18 # of Computers 6. 2018-19 Star Rating 7. 2018-19 Interest in Arts – ACT 8. 2018-19 Interest in Science and Technology – ACT 9. 2018-19 # of Bullying/Cyber Bullying Incidents Reported 10. 2018-19 Grade 11 Dropout Rate
15	high_all	All Variables from 2018 High School Data

- Note: if no year is specified for a variable, the default data set is from 2017-18 school year.

Appendix III

All Variables from 2017-18 Elementary School Data

Performance
Math Mean Scale Score
ELA Mean Scale Score
Science Mean Scale Score
ELPA Mean Scale Score
Percentage Proficient - Math
Percentage Proficient - ELA
Percentage Proficient - Science
Percent Proficient - Read By Grade 3
Math Gap %
ELA Gap %
Math Growth (MGP)
ELA Growth (MGP)
English Language Proficiency Growth (MGP)
Star Rating
Demographics
% of Male Students
% of Female Students
% of Asian Students
% of Black Students
% of White Students
% of Hispanic Students
% of American Indian/Alaskan Native Students
% of Students with Two or More Races
% of Pacific Islander Students
% with an Individual Education Program
% of English Learners
% of Students Eligible for Free or Reduced Price Lunch
% of Students Receiving Free or Reduced Price Lunch
% of Students Eligible for Free or Reduced Price Breakfast
% of Students Receiving Free or Reduced Price Breakfast
Financial
Overall Total Spending Per Pupil
Per Pupil Expenditures - Instruction $
Per Pupil Expenditures - Instruction Support $
Per Pupil Expenditures - Operations $
Per Pupil Expenditures - Leadership $
Per Pupil Expenditures - Instruction %
Per Pupil Expenditures - Instruction Support %
Per Pupil Expenditures - Operations %
Per Pupil Expenditures - Leadership %
Professional Development Funding
Enrollment & Attendance
Average Daily Attendance

Total Enrollment
Student/Teacher Ratio
Student/Teacher Ratio - Kindergarten
Student/Teacher Ratio - 1st Grade
Student/Teacher Ratio - 2nd Grade
Student/Teacher Ratio - 3rd Grade
Student/Teacher Ratio - 4th Grade
Student/Teacher Ratio - 5th Grade
Transiency Rate
Chronic Absenteeism Rate
Discipline
of Incidents of Violence to Other Students
of Incidents of Violence to Staff
of Bullying/Cyber Bullying Incidents Reported
Technology
of New Computers
of Computers
of Old Computers
of Mobile Learning Devices
of IT Technicians per Computer
of Tech Coaches per Student
of Computers per Student
of New Computers per Student
of Old Computers per Student
% of Computers 5 Years or Newer
Substitute Teachers & Paraprofessionals
Total # of Long Term Substitute Teachers
Total # of Short Term Substitute Teachers
of Paraprofessionals Employed
of Paraprofessionals Not NCLB Qualified
% of Paraprofessionals Not NCLB Qualified
Teacher Information
Teacher Average Daily Attendance
Core Subject Classes Not Taught by Highly Qualified Teachers
of Elementary Classes Not Taught By Highly Qualified Teachers
% of Elementary Classes Not Taught By Highly Qualified Teachers

Appendix IV

All Variables from 2017-18 High School Data

Performance
Math Mean Scale Score
ELA Mean Scale Score
Science Mean Scale Score
ELPA Mean Scale Score
Percentage Proficient - ACT ELA
Percentage Proficient - ACT Math
Percentage Proficient - HS Science
High School STEM Mean Scale Score - ACT
High School English Mean Scale Score - ACT
High School Reading Mean Scale Score - ACT
High School Writing Mean Scale Score - ACT
High School Composite Mean Scale Score - ACT
Grade 9 Science Mean Scale Score
Grade 10 Science Mean Scale Score
High School ELA Mean Scale Score - ACT
High School Math Mean Scale Score - ACT
High School Science Mean Scale Score - ACT
High School Grades in Natural Science - ACT
Interest in Science and Technology - ACT
Interest in Arts - ACT
Interest in Social Service - ACT
Interest in Administration and Sales - ACT
Interest in Business Operations - ACT
Interest in Technical - ACT
Star Rating
Demographics
% of Male Students
% of Female Students
% of Asian Students
% of Black Students
% of White Students
% of Hispanic Students
% of American Indian/Alaskan Native Students
% of Students with Two or More Races
% of Pacific Islander Students
% with an Individual Education Program
% of English Learners
% of Students Eligible for Free or Reduced Price Lunch
% of Students Receiving Free or Reduced Price Lunch
% of Students Eligible for Free or Reduced Price Breakfast
% of Students Receiving Free or Reduced Price Breakfast
Financial
Overall Total Spending Per Pupil

Per Pupil Expenditures - Instruction $
Per Pupil Expenditures - Instruction Support $
Per Pupil Expenditures - Operations $
Per Pupil Expenditures - Leadership $
Per Pupil Expenditures - Instruction %
Per Pupil Expenditures - Instruction Support %
Per Pupil Expenditures - Operations %
Per Pupil Expenditures - Leadership %
Professional Development Funding
Enrollment & Attendance
Average Daily Attendance
Chronic Absenteeism Rate
Total Enrollment
Transiency Rate
Average Class Size: English
Average Class Size: Math
Average Class Size: Science
Average Class Size: Social Studies
Discipline
of Incidents of Violence to Other Students
of Incidents of Violence to Staff
of Bullying/Cyber Bullying Incidents Reported
Technology
of New Computers
of Computers
of Old Computers
of Mobile Learning Devices
of IT Technicians per Computer
of Tech Coaches per Student
of Computers per Student
of New Computers per Student
of Old Computers per Student
% of Computers 5 Years or Newer
Core Subject Classes Not Taught By Highly Qualified Teachers
of Core Classes Not Taught By Highly Qualified Teachers
of English Classes Not Taught By Highly Qualified Teachers
of Math Classes Not Taught By Highly Qualified Teachers
of Science Classes Not Taught By Highly Qualified Teachers
of Social Studies Classes Not Taught By Highly Qualified Teachers
of Foreign Language Classes Not Taught By Highly Qualified Teachers
of Arts Classes Not Taught By Highly Qualified Teachers
% of Core Classes Not Taught By Highly Qualified Teachers
% of English Classes Not Taught By Highly Qualified Teachers
% of Math Classes Not Taught By Highly Qualified Teachers
% of Science Classes Not Taught By Highly Qualified Teachers
% of Social Studies Classes Not Taught By Highly Qualified Teachers
% of Foreign Language Classes Not Taught By Highly Qualified Teachers

% of Science Arts Classes Not Taught By Highly Qualified Teachers
Substitute Teachers & Paraprofessionals
Total # of Long Term Substitute Teachers
Total # of Short Term Substitute Teachers
of Long Term Substitute Teachers - Math
of Short Term Substitute Teachers - Math
of Long Term Substitute Teachers - Science
of Short Term Substitute Teachers - Science
of Long Term Substitute Teachers - Social Studies
of Short Term Substitute Teachers - Social Studies
of Long Term Substitute Teachers - ELA
of Short Term Substitute Teachers - ELA
of Paraprofessionals Employed
of Paraprofessionals Not NCLB Qualified
% of Paraprofessionals Not NCLB Qualified
Teacher Information
Teacher Average Daily Attendance
Dropout Rates
Dropout Rate
Grade 9 Dropout Rate
Grade 10 Dropout Rate
Grade 11 Dropout Rate
Grade 12 Dropout Rate
Graduation and College & Career Readiness
4-Year Graduation Rate
5-Year Graduation Rate
Post-Secondary Preparation Participation %
Post-Secondary Preparation Completion %
% of Graduates Receiving an Advanced Diploma
of 9th Grade Credit Sufficient Students
% of 9th Grade Credit Sufficient Students
of Students in 4-Year Cohort (Those expected to Graduate)
of Graduates
of Non-Graduates
of Graduates Receiving an Adjusted Diploma
of Graduates Receiving an Adult Diploma
of Graduates Receiving an Advanced Diploma
of Graduates Receiving a Standard Diploma
of Students Receiving High School Equivalency