Imran Khan
xxxx xxxx

Final Technical Report

## Title:

Quantifying segregation in New York City schools

## Abstract:

This research aimed to better quantify and subsequently visualize the extent of segregation within New York City's (NYC) school system. Results of this research provide policymakers with yet another tool for analyzing school segregation in order to address timely issues on school development, school choice, busing, and diversity within schools. This research provided an approach to determine how segregation can be characterized, and therefore quantified, in NYC schools using predictive analytics techniques, such as machine learning and clustering. The general approach for this research was to clean and analyze data, run a machine learning model, create visualizations of maps and clusters, and interpret results. Results showed the extent to which school segregation based on race and ethnicity has drastically increased, therefore concluding that schools have become increasingly less diverse over the last several years. Implications of this work possibly include fostering the creation of more diverse learning environments, increasing equity and overall quality within the NYC school system, and better understanding of how school enrollment has affected broader social phenomena such as white flight and minority concentration.

## Introduction:

Inequity and segregation within NYC schools is a well-documented and growing problem. This research sought to gain a better understanding of how various predictive analytics techniques could be used to characterize segregation and gain insights into interventions which could be explored in combating the effects of segregation as part of other diversity-promoting Department of Education (DOE) policies and programs. From an urban informatics and systems perspective, school environment and academic performance shape social and economic outcomes of individuals. These outcomes are magnified in highly dense urban settings, and thus define local communities socially, economically, and culturally.

While there have been many attempts to date to qualitatively characterize segregation in NYC schools, there have been fewer attempts to accurately quantify the extent of segregation within the largest public school system in the country. Tagging schools that are segregated or becoming more segregated may be a useful practice when considering the various policy interventions that could be implemented in order to make choice schools, zones, or districts more desirable to a larger swath of NYC's diverse population, thus increasing enrollment of diverse sets of students in a given school.

*This research set out to determine if segregation could be characterized and quantified in NYC schools.*

Despite the rich diversity that exists in America's largest urban playground, NYC has one of the most segregated school systems in the country[1]. While housing patterns contribute to the creation of monolithic communities within the City, decades of flawed policymaking have contributed to further social and economic isolation, thus exacerbating segregation based on race and class. One education policy in particular that has reinforced inequality is school choice, the option provided to families to send their kids outside of their neighborhood to a non-zoned school. Top high schools created a safe space where they would not be held accountable for maintaining diversity by instituting high stakes test-based admissions processes. Elementary and middle schools, however, have been at the forefront of much debate due to their outright negligence in striving towards creating more diverse academic environments[2].

In practice, struggling schools with low standardized test scores, are also characterized by low enrollment, low attendance, inactive parent-teacher groups, and lack of program funding. While desirability drastically falls for struggling schools in a given neighborhood, competition for admission into good schools in surrounding neighborhoods increases to the extent that non-zoned children are not able to get a seat. This often results in increased segregation. When a school has persistently struggled over a number of years, it is tagged for closure. However, school closures may also fuel further segregation within the public school system[3]. The Community and Renewal School programs were recent programs that were developed as an intervention to combat school closures. In 2014, Mayor Bill de Blasio tagged 94 schools, all of which were in the bottom five percent of lowest-performing schools in New York State, to be a part of the Renewal Schools program. Renewal Schools received additional funding for support services for students and families and were paired with a community-based organization partner to help serve various needs[4].

The education field is rich in its use of basic statistical tests and linear regression models to inform policy and practice. Because of the vast amount of data that is available, it is prudent to employ big data strategies, such as machine learning, to gain deeper insights to address segregation, inequity, and other long-standing problems within schools.

While there have been several attempts at using machine learning to explain education outcomes previously, the bulk of these studies focus on the tertiary sector and individual student outcomes. Golino and Gomes applied four machine learning techniques (classification trees, bagging, random forest, and boosting) to 77 college students in their second and third year of medical school in Brazil[5]. Amrieh, Hamtini, and Aljarah fitted machine learning models to 500 users of a learning management system (LMS) across 16 features to explain performance[6]. Bagging, boosting, and random forest were applied to reveal a strong relationship between learner's behavior (on the LMS) and student performance in elementary and secondary schools. Shahiri, Husain, and Rashid applied a number of data mining techniques to determine important attributes used in predicting college student performance, and preferred models for predicting student performance[7]. Thiele, Singleton, Pope, and Stanistreet explored the relationship between student contextual background characteristics (socio-demographic) and academic performance at a university in order to identify which characteristics were associated with students' chances of achieving a ‘good degree’[8]. The topic of this research draws on similar contextual background and socio-demographic characteristics and their relationship to school performance. This research seeks to characterize segregation by clustering temporal trends of minority enrollment by school, and to determine subsequent spatial patterns, along with feature importance for the trends characterized by clustering. Finally, one other technique applicable to this research is the time-series random forest technique. Deng et al. overcame the problem of the large feature space by employing a random forest approach for time series data, using summary statistics (mean, standard deviation, and slope) of each interval as features[9].

## Data:

This research employed School-Level Master File (SCHMA) data from New York University’s (NYU) Research Alliance for New York City Schools, housed within the Steinhardt School of Culture, Education, and Human Development. The SCHMA is a compilation of publicly available data from the NYC DOE and the U.S. DOE that dates back to the 1995-1996 school year. This dataset is a consistent, accessible document that can be used to investigate characteristics of individual NYC schools or groups of schools, and how they have changed over time. Examples of characteristics that are included in the SCHMA include enrollment and demographic information, attendance rates, number of classes and class sizes, progress report information (when available), test scores for grades three to eight, high school graduation rates, and school expenditures data. Of the 270 features in this dataset, 140 of them were employed in temporal clustering analysis.

This research also employed spatial data from NYC Open Data, specifically shapefiles on NYC school zones, districts, and point locations. These files provided the foundation for spatial analysis for this research. The choice between utilizing either a school zone layer, school district layer, or some comparative analysis between the two would provide insight into how re-drawing boundaries could have drastic implications on school system diversity and DOE policy. Select gentrifying and non-gentrifying neighborhoods could be used for additional qualitative and quantitative analysis based on results.

## Methodology:

The methods used for this research sought to obtain interpretable models and results, useful and relevant for policymaking bodies in the education policy domain.

Three broad themes guided the application of machine-learning techniques for this research:

*Interpretability* - Interpretability is important for methods informing actions and interventions. A clear and consistent message about how and why changes are made to policies instill trust and reliability of those policies and adherence. If the models are not interpretable, policymakers are not able to provide the level of transparency required by parents, teachers, and other administrators in the school system, leading to a lack of trust and adherence. The method used for this was a decision tree classifier.

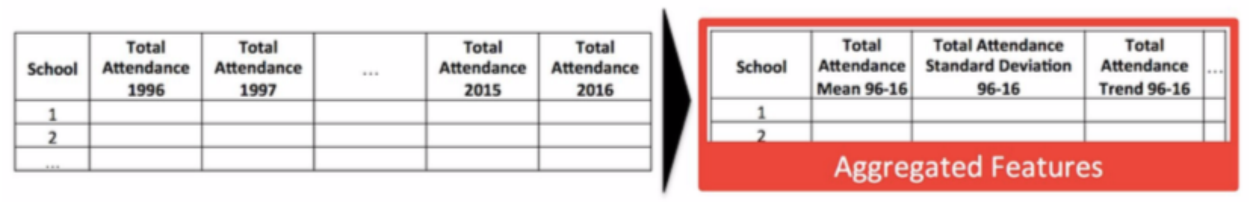
*Feature selection and dimensionality reduction* - As discussed in the data section, the primary dataset used for this research (SCHMA) has over 140 variables used for temporal analysis, for over 1900 schools. A suitable method was needed to help prioritize which features are best at supporting the response variable of interest. The method used for this is the less-interpretable random forest classifier. However, this classifier was only used to inform features which perform best on average, for which dimensionality can be reduced and resulting features included in the decision tree classifier.

*Action-oriented* - Related to the interpretability theme above, results also need to be actionable. The features included in classifiers need to inform relevant policy interventions which could be enacted. If results are not able to be implemented, it brings into question why the research was undertaken to begin with. Again, this theme lends itself to a decision tree classifier, along with an overlay assessing specific policy considerations which could be employed subsequently. If the feature is not actionable from a policy perspective, it would be removed from the list of potential features for fitting.

*Feature engineering: time series aggregation*

The SCHMA dataset was organized into a structure providing one observation per school per year. Data was converted to a time series by feature and by school. The time period for each feature varied considerably. While some features had the full 21 years (1996 through 2016) available, it was more common that subsets finishing in 2016 were available (such as 2002-2016 or 2008-2016). In cases where schools had been closed or reassigned, the observation for 2016 was not available (NaN).

For the time series random forest technique described in the introduction, annual observations were processed to develop a mean, standard deviation, and trend (first year's data subtracted from last year's data) for each school and for each feature. A number of these summary statistics were created for different lengths of time (1996-2016, 2005-2016, 2009-2016, and 2011-2016). This allowed for a smaller number of features, three features per school from the original list of variables, as represented in Figure 1 below.



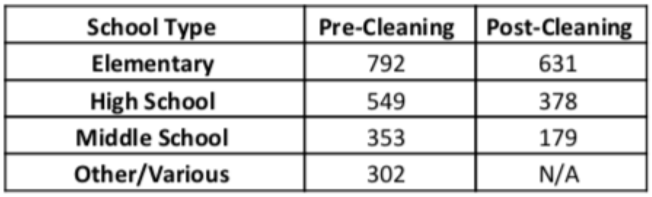
| School | Total Attendance 1996 | Total Attendance 1997 | ... | Total Attendance 2015 | Total Attendance 2016 |
|---|---|---|---|---|---|
| 1 | | | | | |
| 2 | | | | | |
| ... | | | | | |

| School | Total Attendance Mean 96-16 | Total Attendance Standard Deviation 96-16 | Total Attendance Trend 96-16 | ... |
|---|---|---|---|---|
| 1 | | | | |
| 2 | | | | |

Figure 1: Aggregation of multiple time series features into three features

*Data cleaning*

Although the SCHMA dataset was incredibly rich with many features across multiple schools, there were gaps in data with many schools missing values for entire features or entire years. For many machine learning techniques, these values needed to either be imputed or removed. This was completed as needed for specific features to first determine whether or not these values could be imputed (aggregates, assumed to be zero, etc.) or removed. Results of data cleaning are summarized in Table 1 below.



| School Type | Pre-Cleaning | Post-Cleaning |
|---|---|---|
| Elementary | 792 | 631 |
| High School | 549 | 378 |
| Middle School | 353 | 179 |
| Other/Various | 302 | N/A |

Table 1: Total number of schools used in analysis, before and after data munging

*Categorizing segregation of minority groups*

*Time series clustering: demographic enrollment*

This research sought to classify time series of enrollment percentage of different demographic groups. Temporal clusters for both enrollment numbers and enrollment percentages for black, white, Hispanic, and special education students were fit along with drop-out rate. Examples of these outputs are shown in Figure 2 and Appendix A.

A k-means algorithm was employed to fit the clusters and assess the optimal number of clusters using the elbow method. The optimal number of clusters for each of the demographic variables was five in each case. Clustering was also undertaken on performance metrics (for English and math test scores), and the optimal number of clusters in these cases was three.

*Feature selection and decision tree visualization for demographic enrollment clusters*

A decision tree was fit to further characterize each of the demographic enrollment clusters. The methodology for fitting and tuning the random forest classifier is the same as the methodology used for the initial random forest fitting (i.e. oversampling the minority classes and tuning the hyperparameters on a validation set). The features determined to be most informative were selected as the inputs for a decision tree classifier, which was subsequently visualized.

## Results:

Employing the above data and methodology, the following results were obtained.

*Unsupervised grouping of temporal trends*

The map in Figure 2 below shows time series clustering for the black student enrollment percentage variable between 2002 and 2016. Clustered schools are represented by dots and school districts are represented by boundaries of middle school districts.

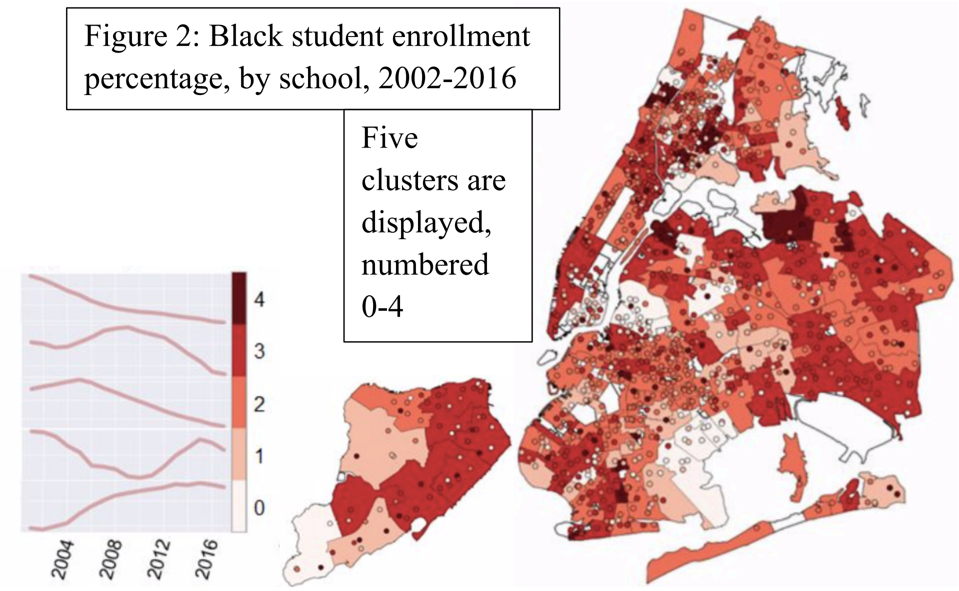


Figure 2: Black student enrollment percentage, by school, 2002-2016

Five clusters are displayed, numbered 0-4

Only one of the clusters showed an increasing trend of black student enrollment percentage (cluster 0), comprising 110 schools, approximately 10 percent of the total number of schools clustered.

The darker colors (dark red) represent clusters with black student enrollment percentage decreasing the most, while the lighter color (faint/light pink) is the single cluster with increasing black student enrollment percentage. The total number of schools in each cluster, after cleaning, is shown below in Table 2.

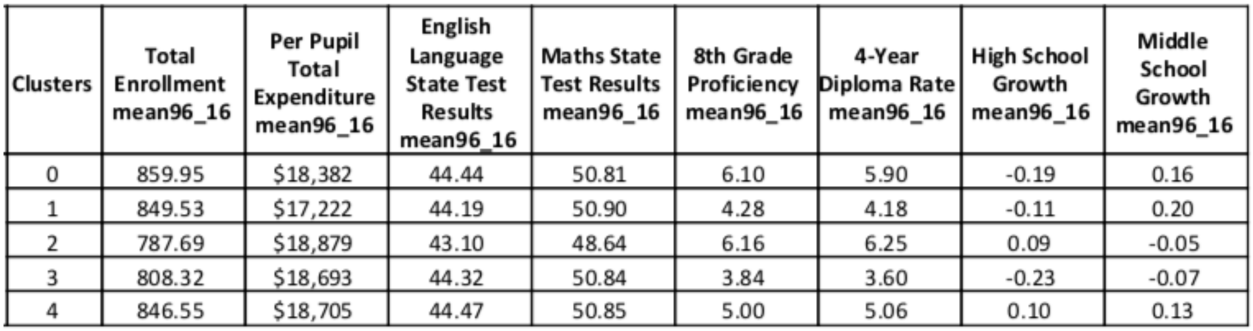
| Cluster # | Number of schools in cluster |
|---|---|
| 4 | 90 |
| 3 | 326 |
| 2 | 341 |
| 1 | 124 |
| 0 | 110 |

Table 2: Number of schools by cluster for black student enrollment percentage

Spatial trends are evident across districts, particularly cluster 0, clustered around lower East Manhattan, and outer Brooklyn.

It is important to note that the analysis conducted used percentage rather than absolute number in order to normalize data per school. Four of the five clusters saw drops in black student enrollment percentage, and this may indicate some diversification of these schools. However, the observation that that only 10 percent of schools saw black student enrollment percentage increasing suggests that black students are leaving most schools to move to predominantly black schools.

The table below shows that various features, along with clusters for black enrollment, most appear to be consistent (per pupil expenditure, enrollment, and test results). However, there are some variations, although they do not appear to be immediately meaningful. It is worth noting that there appears to be differences in the growth features (difference in performance at the beginning and end of middle and high school, split according to deciles, and the change in these deciles over time). The growth features indicate that clusters 0 and 1 see the highest growth in middle school, while these same clusters see some of the lowest growth in high school. Aggregated features for black student enrollment clusters are shown in Table 3 below.



| Clusters | Total Enrollment mean96_16 | Per Pupil Total Expenditure mean96_16 | English Language State Test Results mean96_16 | Maths State Test Results mean96_16 | 8th Grade Proficiency mean96_16 | 4-Year Diploma Rate mean96_16 | High School Growth mean96_16 | Middle School Growth mean96_16 |
|---|---|---|---|---|---|---|---|---|
| 0 | 859.95 | $18,382 | 44.44 | 50.81 | 6.10 | 5.90 | -0.19 | 0.16 |
| 1 | 849.53 | $17,222 | 44.19 | 50.90 | 4.28 | 4.18 | -0.11 | 0.20 |
| 2 | 787.69 | $18,879 | 43.10 | 48.64 | 6.16 | 6.25 | 0.09 | -0.05 |
| 3 | 808.32 | $18,693 | 44.32 | 50.84 | 3.84 | 3.60 | -0.23 | -0.07 |
| 4 | 846.55 | $18,705 | 44.47 | 50.85 | 5.00 | 5.06 | 0.10 | 0.13 |

Table 3: Means of various features for black student enrollment clusters

*Classifying trends in black student enrollment*

The decision tree (Figure 3) and corresponding confusion matrix (Figure 4) below highlight results in classifying trends in black student enrollment. After performing feature selection by a random forest technique, a decision tree was fit on the most informative features, as interpretability was critical in this case. The decision tree classifier gave a result of 0.4657 for in-sample accuracy and 0.4170 for out-of-sample accuracy. The decision tree fit well for clusters 0, 2, and 3 as referenced below. It was observed that with decreasing black student enrollment, overall school enrollment increased.

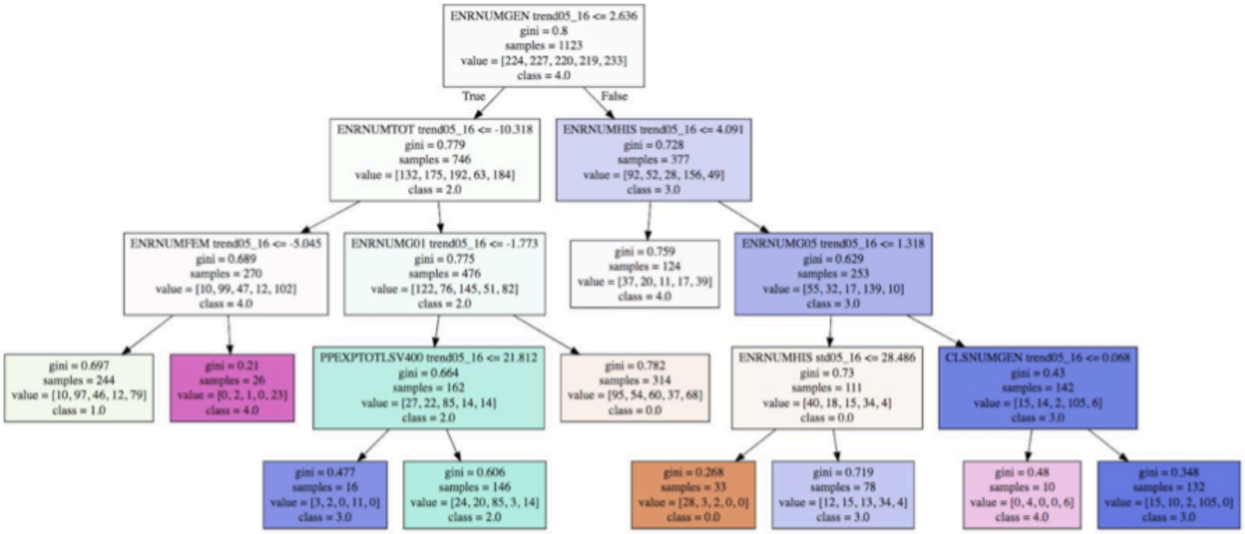
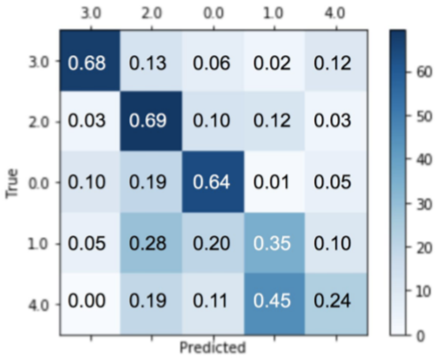

ENRNUMGEN trend05_16 <= 2.636 gini = 0.8 samples = 1123 value = [224, 227, 220, 219, 233] class = 4.0
True
False
ENRNUMTOT trend05_16 <= -10.318 gini = 0.779 samples = 746 value = [132, 175, 192, 63, 184] class = 2.0
ENRNUMHIS trend05_16 <= 4.091 gini = 0.728 samples = 377 value = [92, 52, 28, 156, 49] class = 3.0
ENRNUMFEM trend05_16 <= -5.045 gini = 0.689 samples = 270 value = [10, 99, 47, 12, 102] class = 4.0
ENRNUMG01 trend05_16 <= -1.773 gini = 0.775 samples = 476 value = [122, 76, 145, 51, 82] class = 2.0
gini = 0.759 samples = 124 value = [37, 20, 11, 17, 39] class = 4.0
ENRNUMG05 trend05_16 <= 1.318 gini = 0.629 samples = 253 value = [55, 32, 17, 139, 10] class = 3.0
gini = 0.697 samples = 244 value = [10, 97, 46, 12, 79] class = 1.0
gini = 0.21 samples = 26 value = [0, 2, 1, 0, 23] class = 4.0
PPEXPTOTLSV400 trend05_16 <= 21.812 gini = 0.664 samples = 162 value = [27, 22, 85, 14, 14] class = 2.0
gini = 0.782 samples = 314 value = [95, 54, 60, 37, 68] class = 0.0
ENRNUMHIS std05_16 <= 28.486 gini = 0.73 samples = 111 value = [40, 18, 15, 34, 4] class = 0.0
CLSNUMGEN trend05_16 <= 0.068 gini = 0.43 samples = 142 value = [15, 14, 2, 105, 6] class = 3.0
gini = 0.477 samples = 16 value = [3, 2, 0, 11, 0] class = 3.0
gini = 0.606 samples = 146 value = [24, 20, 85, 3, 14] class = 2.0
gini = 0.268 samples = 33 value = [28, 3, 2, 0, 0] class = 0.0
gini = 0.719 samples = 78 value = [12, 15, 13, 34, 4] class = 3.0
gini = 0.48 samples = 10 value = [0, 4, 0, 0, 6] class = 4.0
gini = 0.348 samples = 132 value = [15, 10, 2, 105, 0] class = 3.0

Figure 3: Classification tree characterizing black student enrollment percentage trends



| | 3.0 | 2.0 | 0.0 | 1.0 | 4.0 |
|---|---|---|---|---|---|
| 3.0 | 0.68 | 0.13 | 0.06 | 0.02 | 0.12 |
| 2.0 | 0.03 | 0.69 | 0.10 | 0.12 | 0.03 |
| 0.0 | 0.10 | 0.19 | 0.64 | 0.01 | 0.05 |
| 1.0 | 0.05 | 0.28 | 0.20 | 0.35 | 0.10 |
| 4.0 | 0.00 | 0.19 | 0.11 | 0.45 | 0.24 |

Figure 4: Confusion matrix for black student enrollment percentage classifier

**Conclusions:**

A model was successfully developed to try to understand the nature of segregation in schools and the features that these segregated schools may be experiencing. The clustering of racial groups by enrollment trends provides a useful and easy way to compare different trends throughout the City, and when looking at features across these clusters, there are few easily identifiable insights or trends.

One of the major issues with a dataset of this size is the large number of features that can dilute the impact of important features. The use of a combination of random forest to identify a smaller subset of features for selection before being input into the decision tree allows for a decent level of predictability with clear interpretability.

While the decision tree model for black student enrollment does not appear to provide many insights, it does indicate that while schools are growing in enrollment overall, black student enrollment is decreasing across the City. This is further compounding the already existing issue of segregation and continuing the generational discrimination against minority students and families. The analysis developed a classification approach to understand features highly correlated with segregation. These indicators could be compared with various other socio-economic factors that could potentially lead to development of policy that improves school environments, reduces inequity and segregation, and promotes development of long-neglected neighborhoods.

Similarly, the other population groupings and their models do not currently provide clear insights as these models could be tweaked and studied further. Further study could use the idea of creating these clusters around specific features in schools, and then use anomaly detection to analyze across clusters to determine what schools are outperforming their peers. This may assist with identifying features that make these schools different. For instance, if we were to cluster along socio-economic lines, we could identify anomalous schools based on academic performance (both good and bad), and identify which features make those schools anomalous, as well as providing the DOE with a list of schools to use as a model for other peer schools.

Education policy is a wicked problem in that it is socially complex, often subject to unforeseen consequences, potentially multi-causal, and difficult to change or reverse the outcome once implemented. Machine learning provides a set of powerful analysis tools for handling the type of big data that has become the norm in the education space and may prove useful for the DOE. Diversity is important in order to achieve better social and societal outcomes, particularly in education environments where youth are being developed intellectually, emotionally, and physically. All policy interventions at the DOE should include frameworks for increasing diversity and decreasing segregation, especially for black students and others that have been marginalized for far too long. Indeed, increased social cohesion and social inclusion within NYC would provide for better implementation when policy interventions are put through for almost every social and economic policy issue area.

*Limitations of the analysis*

*Variability in schools* - In general, schools are quite typical and standard institutions across elementary, middle, and high schools. However, there were multiple cases where schools had variation in start and stop periods - e.g. a K-8 school versus an elementary school, and elementary school that goes through grade 5 versus an elementary school that goes through grade 6. There were also a few special or unusual programs. Due to the large and complex nature of the dataset and due to time constraints, the cleaning process meant that some of these more complex schools lost some of the features which could ultimately impact the model's performance and effectiveness, as well as the loss of some specificity, granularity, and potentially informative features. More time could be spent cleaning the datasets and imputing values to reduce this loss.

*Incomplete datasets* - Although the SCHMA is an incredibly rich dataset with a comprehensive list of features for most schools in NYC over an extended period, multiple data points needed to be removed due to incomplete features, with some considerably important features needing to be removed from the analysis all together. While the amount of features and schools were still significant enough to conduct analysis, the loss of data may have implications in terms of the accuracy of the results, particularly as there may have been specific characteristics related to equity for schools that either had missing data or did not provide data to the Research Alliance.

*Schools as opposed to individual students* - The analysis was conducted on a school level basis. While this allows for reasonably good interpretability and provides a good overview of the education system, the purpose of education is to improve outcomes for individual students. Schools are complex structures, and any analysis conducted at a school level should be taken in cautiously. While school level information is useful to diagnose issues and identify policy tweaks, ultimately it is not granular enough to provide detailed insights on individual student outcomes, which are the most important goals when addressing challenges in education. More complex analysis, such as network analysis, could also provide insight into and quantifying of social interactions.

*Masked data may introduce biases in reporting* - The SCHMA masks records where low sample size is recorded. An example of this are the counts of minority enrollment at schools. The masked records could potentially introduce biases into any analysis undertaken, particularly temporal trends which may have removed schools with increasing minority enrollments because those enrollments are smaller or below a threshold.

## References:

1. New York Times Editorial Board (2017). Confronting Segregation in New York City Schools. Retrieved from https://www.nytimes.com/2017/05/15/opinion/school-segregation-nyc.html.
2. Kemple, J. (2016). School Closures In New York City. Retrieved from https://www.educationnext.org/school-closures-in-new-york-city-did-students-do-better/.
3. Kemple, J. (2015). High School Closures in New York City: Impacts on Students' Academic Outcomes, Attendance, and Mobility. Retrieved from https://steinhardt.nyu.edu/scmsAdmin/media/users/sg158/PDFs/hs_closures/HighSchoolClosuresinNewYorkCity_ResearchAllianceforNYCSChools_pdf.pdf.
4. Kolodner, M. (2017). The Convoluted Path to Improving New York City's Schools. Retrieved from https://www.theatlantic.com/education/archive/2017/02/the-renewal-school-gamble/515985/.
5. Golino, H. and Gomes, C. (2014). Four Machine Learning Methods to Predict Academic Achievement of College Students: A Comparison Study. Retrieved from https://pdfs.semanticscholar.org/d70d/3c08181b742ae95395f745a2c276673a0c75.pdf.
6. Amrieh, E., Hamtini, T., and Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. Retrieved from https://www.researchgate.net/profile/Ibrahim_Aljarah/publication/307968552_Mining_Educational_Data_to_Predict_Student's_academic_Performance_using_Ensemble_Methods/links/57d4634a08ae5f03b4915f4d/Mining-Educational-Data-to-Predict-Students-academic-Performance-using-Ensemble-Methods.pdf.
7. Shahiri, A., Husain, W., and Rashid, N. (2015). A Review on Predicting Student's Performance using Data Mining Techniques. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877050915036182/pdf?md5=97065b113bee2528b075858bc1267466&pid=1-s2.0-S1877050915036182-main.pdf.
8. Thiele, T., Singleton, A., Pope, D., and Stanistreet, D. (2016). Predicting students' academic performance based on school and socio-demographic characteristics. Retrieved from https://www.tandfonline.com/doi/pdf/10.1080/03075079.2014.974528?needAccess=true.
9. Deng, H., Runger, G., Tuv, E., and Vladimir, M. (2013) A time series forest for classification and feature extraction. Retrieved from https://arxiv.org/pdf/1302.2277.pdf.

**<u>Supplementary materials:</u>**

The complete GitHub repository, including code, data, and documentation, for this project can be found at https://github.com/buildwelllabs/machine.learning.project.

Appendix A: Additional figures and tables

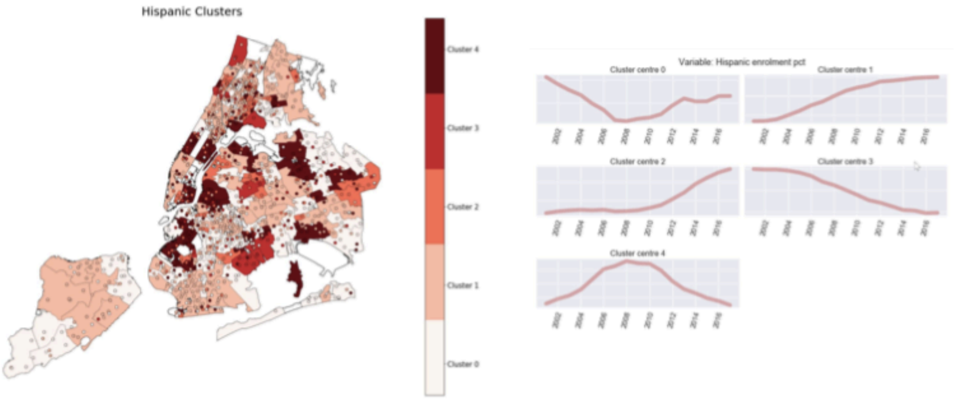


Figure 5: Hispanic student enrollment percentage, by school, 2002-2016

| Clusters | Total Enrollment mean96_16 | Per Pupil Total Expenditure mean96_16 | English Language State Test Results mean96_16 | Maths State Test Results mean96_16 | 8th Grade Proficiency mean96_16 | 4-Year Diploma Rate mean96_16 | High School Growth mean96_16 | Middle School Growth mean96_16 |
|---|---|---|---|---|---|---|---|---|
| 0 | 736.07 | $18,315 | 46.75 | 52.35 | 5.85 | 6.63 | 0.82 | 0.11 |
| 1 | 869.89 | $18,852 | 40.23 | 46.48 | 5.67 | 5.32 | -0.35 | -0.02 |
| 2 | 752.77 | $19,770 | 43.79 | 49.49 | 5.33 | 4.92 | -0.41 | 0.12 |
| 3 | 834.91 | $15,845 | 50.64 | 57.20 | 4.98 | 5.07 | 0.10 | 0.18 |
| 4 | 898.94 | $18,834 | 43.02 | 49.40 | 4.94 | 5.04 | 0.11 | 0.05 |

Table 4: Means of various features for Hispanic student enrollment clusters

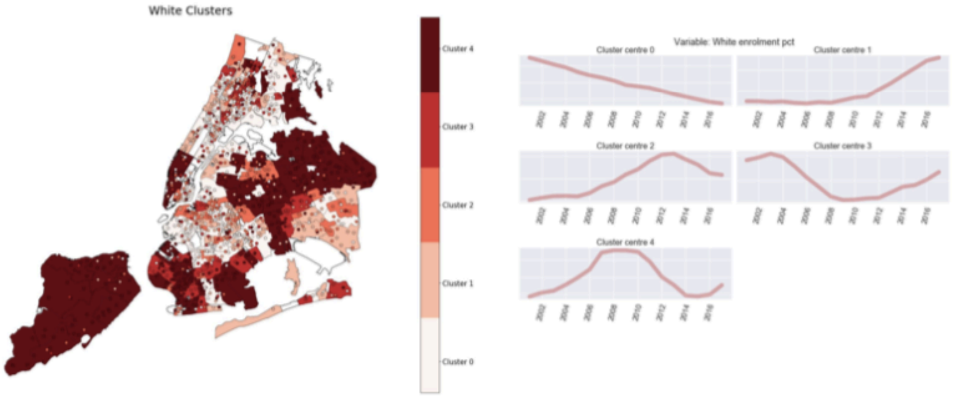
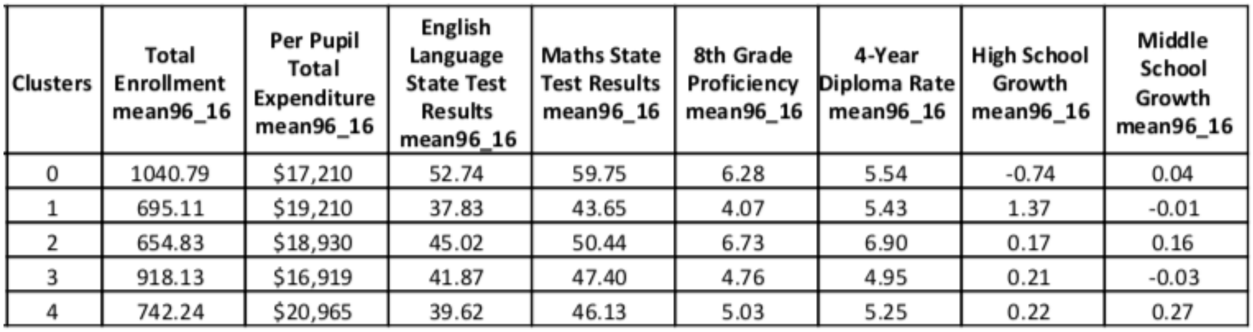


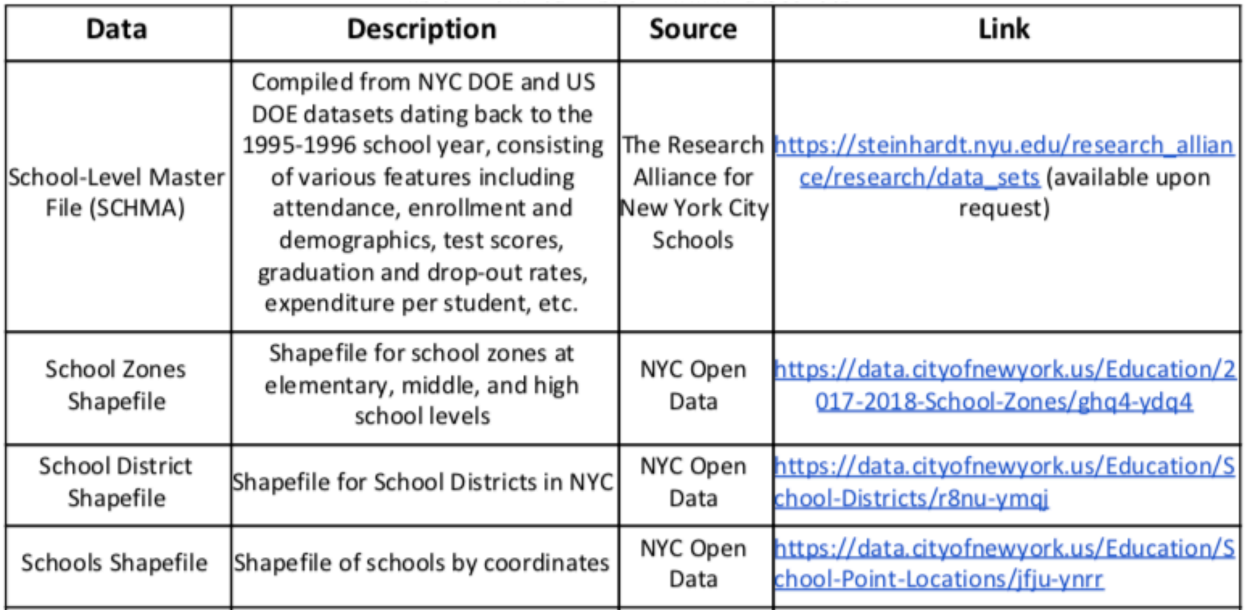
Figure 6: White student enrollment percentage, by school, 2002-2016



| Clusters | Total Enrollment mean96_16 | Per Pupil Total Expenditure mean96_16 | English Language State Test Results mean96_16 | Maths State Test Results mean96_16 | 8th Grade Proficiency mean96_16 | 4-Year Diploma Rate mean96_16 | High School Growth mean96_16 | Middle School Growth mean96_16 |
|---|---|---|---|---|---|---|---|---|
| 0 | 1040.79 | $17,210 | 52.74 | 59.75 | 6.28 | 5.54 | -0.74 | 0.04 |
| 1 | 695.11 | $19,210 | 37.83 | 43.65 | 4.07 | 5.43 | 1.37 | -0.01 |
| 2 | 654.83 | $18,930 | 45.02 | 50.44 | 6.73 | 6.90 | 0.17 | 0.16 |
| 3 | 918.13 | $16,919 | 41.87 | 47.40 | 4.76 | 4.95 | 0.21 | -0.03 |
| 4 | 742.24 | $20,965 | 39.62 | 46.13 | 5.03 | 5.25 | 0.22 | 0.27 |

Table 5: Means of various features for white student enrollment clusters

Appendix B: Data information



| Data | Description | Source | Link |
|---|---|---|---|
| School-Level Master File (SCHMA) | Compiled from NYC DOE and US DOE datasets dating back to the 1995-1996 school year, consisting of various features including attendance, enrollment and demographics, test scores, graduation and drop-out rates, expenditure per student, etc. | The Research Alliance for New York City Schools | https://steinhardt.nyu.edu/research_alliance/research/data_sets (available upon request) |
| School Zones Shapefile | Shapefile for school zones at elementary, middle, and high school levels | NYC Open Data | https://data.cityofnewyork.us/Education/2017-2018-School-Zones/ghq4-ydq4 |
| School District Shapefile | Shapefile for School Districts in NYC | NYC Open Data | https://data.cityofnewyork.us/Education/School-Districts/r8nu-ymqj |
| Schools Shapefile | Shapefile of schools by coordinates | NYC Open Data | https://data.cityofnewyork.us/Education/School-Point-Locations/jfju-ynrr |

Table 6: Detailed data dictionary