Factors Affecting College Preparedness in NYC High Schools

March 21, 2023

Ross Lauterbach

Introduction

New York is one of the most socially and economically diverse regions in the world. While there are many benefits to this, it has also contributed to disparities in college preparedness in high schools across the five boroughs. Some students are equipped with a multitude of tools to guide them through the college process confidently while others are unsure of their ability to obtain higher education due to a lack of resources. To reduce this disparity, it is critical to examine how schools can better prepare students from less affluent communities for postsecondary success.

An analysis was conducted using data from the U.S. Department of Education regarding high schools across New York State in 2017. Initially, the intention was to analyze school districts across the entire state – however, the vast diversity of districts made it difficult to draw meaningful comparisons. The focus then shifted to New York City (NYC) public schools as they present a more controlled environment with shared governance and policies. The dataset includes enrollment statistics, funding received, staffing levels, and standardized testing participation, and was analyzed to address the following questions:

1. What are the differences in NYC high schools as indicated by key metrics like funding per student, teacher-to-student ratio, and SAT/ACT participation rates? What are some effective ways to visualize disparities in these areas?

2. How does clustering NYC high schools into groups based on key metrics compare to grouping them by white student enrollment percentages?
3. Which factors most strongly predict college enrollment and are there any variables that negatively affect the predictive model?

Data

The data used in this analysis was collected from the U.S. Department of Education's Civil Rights Data Collection (2017). The dataset focuses on New York City high schools and contains approximately 1,600 rows and columns for total enrollment, enrollment by ethnicity, SAT/ACT participation rates, teacher and counselor staffing levels, and state funding. While the dataset provides more information, these specific variables were used due to their relevance to college preparedness. The main outcome variable in this analysis is SAT/ACT participation, as it serves as a proxy for college intent. The only columns containing categorical information, namely school and district names, were excluded from the analysis to focus solely on quantitative data.

Analysis

To begin the analysis, the dataset was summarized to observe how NYC high schools differ in key variables. This summary revealed substantial variation in school size, with enrollment figures ranging from 7 to 5,838 students. Such disparities suggest that ratios (e.g., funding per student) would provide a more meaningful comparison than raw totals. The summary also showed significant differences between the maximum values in key metrics and the 75th percentile, indicating that certain schools receive far more resources than others.

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Enrollment	1596	605.583	497.883	7	327.75	704.25	5838
Counselors	1588	1.777	1.974	0	1	2.09	21.39
Teachers	1588	48.301	31.977	7.12	28.15	58.935	275.69
Funding	1596	5953205.192	4439567.088	0	3308573.5	7019010.75	40740152
SAT_ACT_Takers	1588	152.02	103.125	14.65	86.388	185.342	860.39

Figure 1: Descriptive statistics for both predictor and outcome variables

This disparity raises the question of whether resource allocation is influenced by factors implicated by geographical differences and the subsequent wealth disparity. One of these factors worth exploring is the influence of racial demography on the allocation of educational resources. To explore this, a 3D scatter plot was created visualizing the relationships between funding per student, teacher-to-student ratio, and SAT/ACT participation rates. Figure 2 uses different colors to represent quartiles of white student enrollment to examine the correlation between racial composition and school resources.

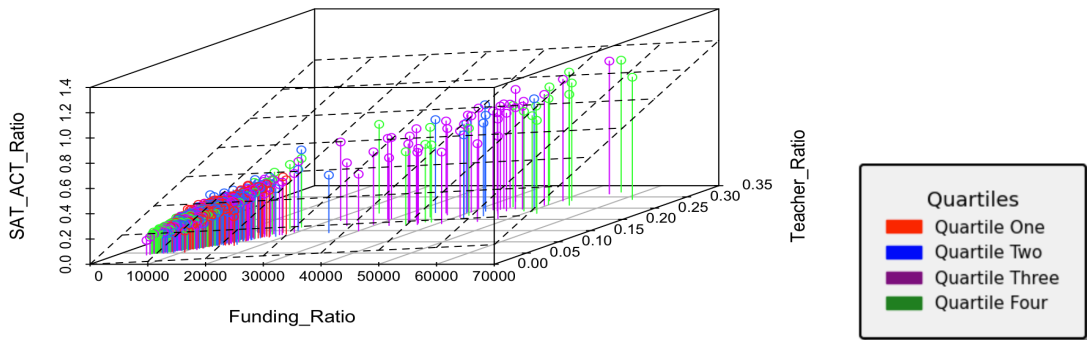


Figure 2: 3-D plot colored with quartiles of white student enrollment

There is a visible concentration of green and purple points (representing schools in the top 50% of white student enrollment) in the upper part of the plane. Conversely, red points (representing the first quartile) are entirely absent from this region. This suggests that schools with higher percentages of white students may receive more funding and better college

preparation resources. This visualization aligns with the broader discussion about how racial privilege influences educational opportunities.

The second part of the analysis delves deeper into racial inequalities in education. K-means clustering was applied to explore whether groupings of NYC high schools based on resource metrics align with the racial demographic breakdown indicated by the percentage of white student enrollment. This approach allowed us to investigate the extent to which resource disparities correlate with demographic composition. The ideal choice for the number of clusters would be four due to easy comparison with the quartiles of white student enrollment. Figure 3 shows the total within-cluster sum of squares, which indicates the deviation from each respective cluster center. This plot, which measures the error for each number of clusters, confirms that four clusters are a reasonable choice with regard to error and overfitting. Figure 4 shows the result of k-means clustering, with each color representing a different cluster. Comparing this to the previous plot reveals similarities in how schools are grouped, reinforcing the idea that racial demographics play a role in resource distribution.

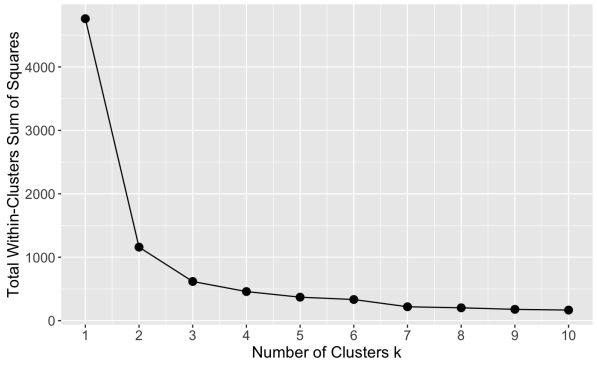


Figure 3: Total Sum of Squares for varying amounts of clusters

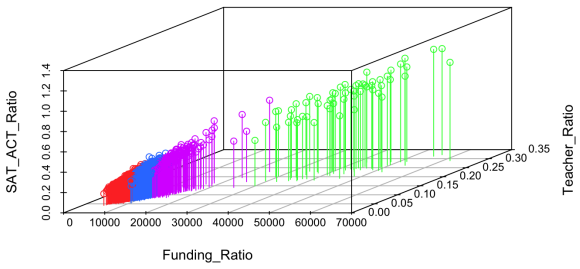


Figure 4: 3-D plot colored by cluster number

A heat map was created (Figure 5) comparing the clusters with the quartiles of white student enrollment to further illustrate the relationship between these classifications. The lighter the shade of blue, the higher the concentration of schools in that cell. The heatmap shows that clusters tend to align with the corresponding or adjacent quartiles, showing that the natural clustering of schools based on resources and participation rates has a tangible relationship with racial demographics. In fact, 96.3% of clusters align with the corresponding or adjacent quartiles. This finding supports the hypothesis that white students are generally at an advantage when it comes to college preparedness.

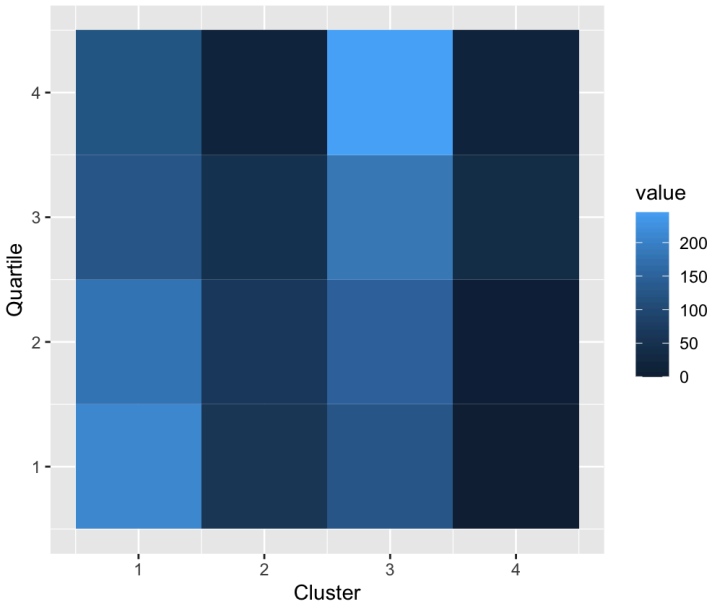


Figure 5: Heatmap of cluster and quantile alignment

The final part of the analysis involves building a predictive model to estimate SAT/ACT participation rates based on several factors. Using a multi-linear regression, the model included the following predictor variables: teacher-to-student ratio, counselor-to-student ratio, white student percentage, and funding per student. The results are summarized in Figure 6.

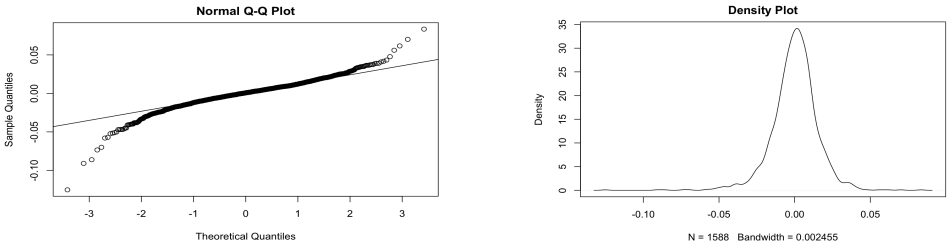
Coefficients	Estimate	Std. Error	T-Value	P-Value
Intercept	0.001864	0.0016	1.165	0.24418
Counselor/Student	0.05478	0.1601	0.342	0.7322
Teacher/Student	2.824	0.03682	76.692	< 2e-16
White Student %	-0.005637	0.002137	-2.638	0.00841
Funding/Student	0.00000272	0.000000184	12.347	< 2e-16

Figure 6: Multivariable regression table utilizing four predictor variables

The model suggests that teacher-to-student ratio and funding per student were strong predictors of SAT/ACT participation, with significant p-values for the corresponding hypothesis test. The white student percentage also revealed a significant relationship with the response variable but surprisingly yielded a negative relationship. This negative coefficient aligns with the clustering findings, where resource-based groupings were found to correlate with white student enrollment. The results suggest that, even after controlling for resource metrics, white student percentage is inversely related to SAT/ACT participation. This may indicate underlying systemic disparities beyond racial demography alone. Interestingly, the counselor-to-student ratio did not significantly impact the model, which was surprising given the role counselors play in college advising. The multiple R-squared value was 0.9817, indicating that the model explains a substantial portion of the variance in SAT/ACT participation. However, because of the presence of outliers, this would need to be further investigated. Residual analysis shows a symmetric of residuals with skew (Figures 7 and 8). The skew observed at the higher end may be due to extreme outliers in school funding, as a few schools received dramatically more resources than others. Normalization techniques could be employed for improvement in performance, although they were avoided for interpretability during the upcoming presentation.

To assess the quality of the model, I conducted a residual analysis using Q-Q and density plots (Figures 7 and 8). The Q-Q plot showed that the residuals are not perfectly normally distributed, with deviations at the tails of the plot. This suggests that there are some outliers that pull the regression line, particularly for schools at the extremes of the funding distribution, which was reflected in the high R-squared value. The slight right skew in the density plot further highlights this issue. The skew could be attributed to a few schools receiving either very high or very low levels of funding, creating an imbalance in how well the model fits across all schools.

Figures 7 and 8: Q-Q and density plot for residual analysis



While these deviations are not large enough to undermine the overall fit of the model, they do indicate areas where future refinements could be made. For instance, introducing additional variables that capture more information about budgetary breakdowns or student demographics might help address the residual outliers. Despite these limitations, the analysis confirms that teacher-to-student ratio and funding per student are the most critical factors in predicting college enrollment.

Conclusion

This analysis revealed significant disparities in college preparedness across NYC high schools, particularly along racial and socioeconomic lines. Together, the clustering and regression analyses reveal that resource allocation disparities in NYC high schools are closely aligned with racial demographics. Teacher-student ratios and funding per student emerge as the strongest predictors of college preparedness, although lots of confounding is taking place, as funding seems to be the underlying predictor. The inverse relationship between white student enrollment and ACT/SAT participation introduces great nuance in this issue of college preparedness. It seems that the issue of disparity in college preparedness is not directly determined by race but perhaps by class. Students who live in poor neighborhoods tend to have to go to schools that are underfunded. These schools have fewer available resources, such as funding and teachers per student, and hence are motivated to participate in the ACT/SAT less. However, students in wealthy schools tend to have the funding to hire more teachers and spend more on each student. This, in turn, can increase student participation in the ACT/SAT. Furthermore, as the analysis indicated, the wealthiest schools tend to be disproportionately populated by white students. White students, in a sense, are simply more likely to be able to go to better schools. Historic racial tensions in the United States' complicated past have caused class differences to be formed, and these differences have helped white people far more than any other demographic group. Another explanation for the findings is the nature of the ACT/SAT as a part of the college application process. The ACT/SAT costs an amount of money that can be out of reach for many low-income students. Not factoring in the cost of tutoring or study materials, registration for the SAT can cost upwards of $70 for registration for a single test. If a student needs to take the test three or four times in order to get a score worth the investment, this can

cost families nearly $250. The cost of tutoring and study materials is also extremely high and are very closely tied to success in these tests, meaning that higher class individuals are more likely to be able to make the investment to successfully participate in these tests.

Future analysis could benefit from more granular data, such as funding sources (state vs. federal) and student performance on standardized tests. Additionally, qualitative data on student experiences and aspirations would provide a more comprehensive understanding of the factors influencing college preparedness. If this study were to continue with the current data, addressing multicollinearity would be the first investigation, alongside utilizing other robust forms of regression that consider interaction effects. Addressing these disparities is crucial for creating a more equitable education system that prepares all students for success, regardless of their background.

References

Civil Rights Data Collection (2017) https://ocrdata.ed.gov/flex/Reports.aspx?type=school (Enrollment, Funding, and Took SAT/ACT Reports)