The Problem Statement

The SAT, a test designed to measure “college readiness”, has over the years become mired in controversy. The prevailing notion is that the test amplifies existing social inequities, rather than measuring academic readiness, with those having greater access to better resources—tutors, advanced coursework—reaping higher scores. This perception is so pronounced that an SAT score is often regarded as a mere proxy for wealth.

With this historical context in mind, I set out to determine to what extent these factors play out in the largest metropolitan area in the United States. I primarily wanted to answer the following question: to what extent do factors outside a student’s control determine their SAT score? Essentially, if I could model SAT Cumulative Scores by, say, simply looking at features tied to wealth or other factors outside a student’s control, then this would speak to clear systemic biases.

That said, the project does not intend to establish a cutoff for “clear systemic biases”. Rather, the aim is to arm college admissions boards with important data that can provide a basis for how they use SAT scores when evaluating applicants.

One area I also wanted to look at closely was racial considerations. Given that the distribution of resources in the U.S. has been shown to be unfairly apportioned across racial lines, or what can be described as systemic racism, I wanted to determine how marked these effects were. Such information would enable college admissions boards to consider weighing systemic inequities when considering applicants from marginalized communities.*

Regardless of race, the SAT plays a critical part in the lives of any student looking to go to college. To provide more context around the high stakes nature of the test: the SAT is a standardized test that plays a large role in determining which U.S. college a student is admitted to. Not all colleges require this score, but many of the top-tiered schools do—schools that can open up a lifetime of opportunities that might otherwise be denied to those in lower-tiered schools. Indeed, the entire trajectory of a student’s future can often hinge on their SAT score.

The aim of this project is to determine which extrinsic factors affect SAT scores. If the SAT does indeed amplify existing social inequalities, we have some pressing questions to answer around the future of standardized testing in America.

*During the time in which this project was conducted, the Supreme Court reached an historic decision to overturn affirmative action. Just how colleges will be able to weigh extrinsic factors when determining which applicant to accept remains unknown at this time.

Data Preparation

I utilized a dataset containing 437 high schools, out of a total of 542 high schools in New York City, and the greater boroughs (those not included invite a potential selection bias). The dataset included SAT scores, average household income per zip code, and four ethnic groups as well as many other features used to determine what interaction existed between these variables and the target variable–Cumulative SAT score.

The dataset was downloaded from a csv file taken from a New York City government website and contains scores from the 2016-2017 academic year. For wealth data, I merged a csv file from the IRS website, which provided average income per zip code. This was the most nuanced income data I could find, but it did not necessarily align with the average income per high school. Therefore, while the target variable—SAT Cumulative Score—was broken down via high school, an important variable—average income—was averaged out across zip codes.

Other proxies for wealth were used, centering around the real estate market. I merged a csv file containing data around New York City housing with the main dataset, taking specific features that I felt might track with wealth and access to resources, such as latest sale price of a piece of real estate.

To further enrich the main dataset, I found a dataset that contained the student-teacher ratio per classroom, a feature missing from the main dataset, that I felt determined access to a critical resource—a teacher's bandwidth.

Once I merged these datasets, the dataset included 27 features. However, I judged many of these features extraneous to the analysis, so they were removed. These include the following:

- School ID
- Building Code
- Latitude and Longitude (given zip code already existed)
- Phone Number
- Start and End time

Once these features were removed, the dataset consisted of the following 16 features (more features were added during modeling):

- Student Enrollment
- Percent White
- Percent Black
- Percent Hispanic
- Percent Asian
- Average Score (SAT Math)
- Average Score (SAT Reading)
- Average Score (SAT Writing)
- Percent Tested

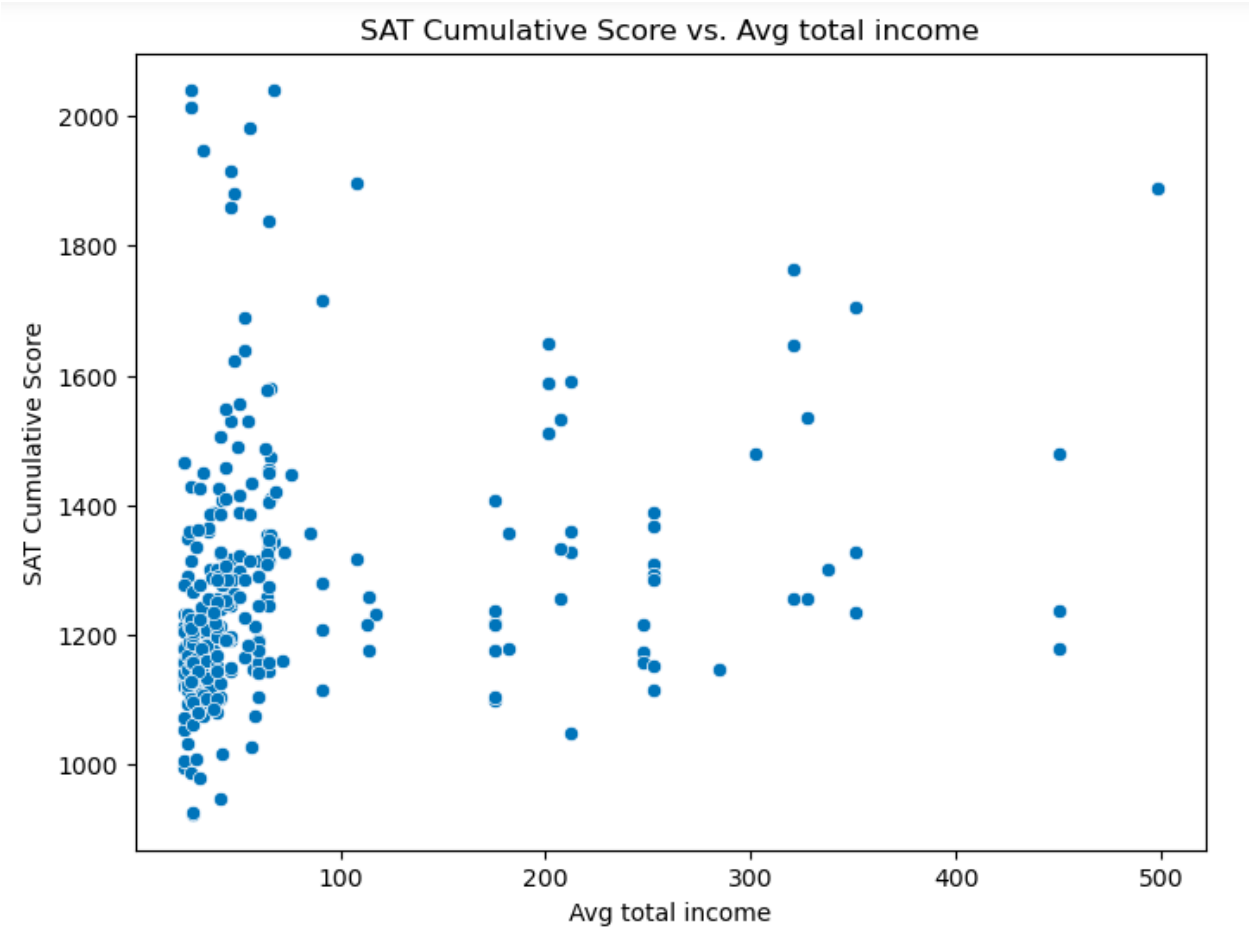
- Average Total Income
- Total Units
- Gross Square Feet
- Sale Price
- Average Price per Square Feet
- School Pupil-Teacher Ratio
- SAT Cumulative Score

For some of the schools from this dataset, they were inputted differently than how they were inputted on the main dataset, so I was unable to match them into the main dataset. This means that the dataset dropped from 435 schools to 298 schools. This might have invited a potential selection bias but assuming how they were inputted differently was likely random, I don't expect there to be a strong selection bias, if any at all.

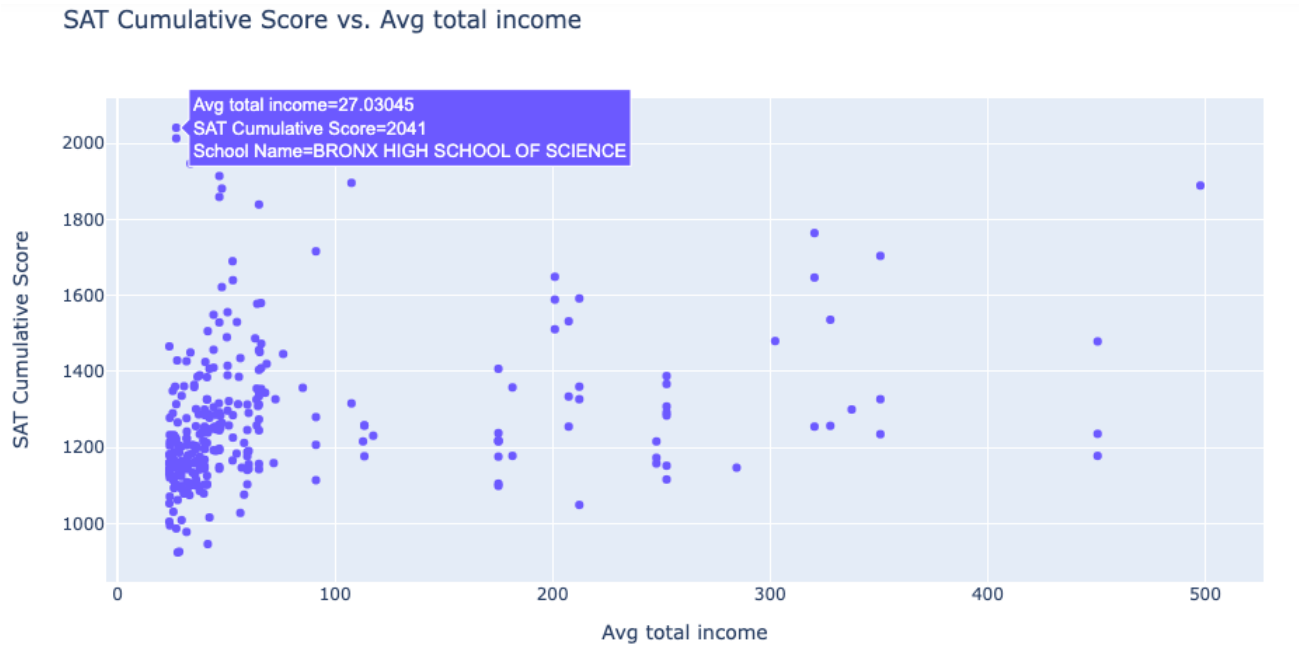
Finally, it should be noted that the original dataset only contained scores broken down by the three test sections—math, verbal, and writing—not a cumulative score, which I considered most important. I created an additional column for SAT Cumulative Score (noted in the list of 16 features above), and I used this as the target variable for the project.

Exploratory Data Analysis

1. Average income



The first thing I determined was the correlation between income and SAT scores (.24) using a correlation matrix. To graph these findings, I used the scatterplot above. What is striking is that the highest scores also have the lowest income (the lowest scores, perhaps unsurprisingly, also fall into the lowest income bracket, where resources are likely scant.) But how then to account for the highest scores existing at the lowest income level?



Using plotly, I searched out these extreme outliers to reveal an interesting trend. Many of New York's most competitive high schools—Bronx High School of Science (shown above) has an acceptance rate lower than that of Harvard—are located in zip codes with the lowest total income.

This created a confound: the average household income of students did not match—and at times were starkly at odds with—the average household income for the zip code where a school was located.

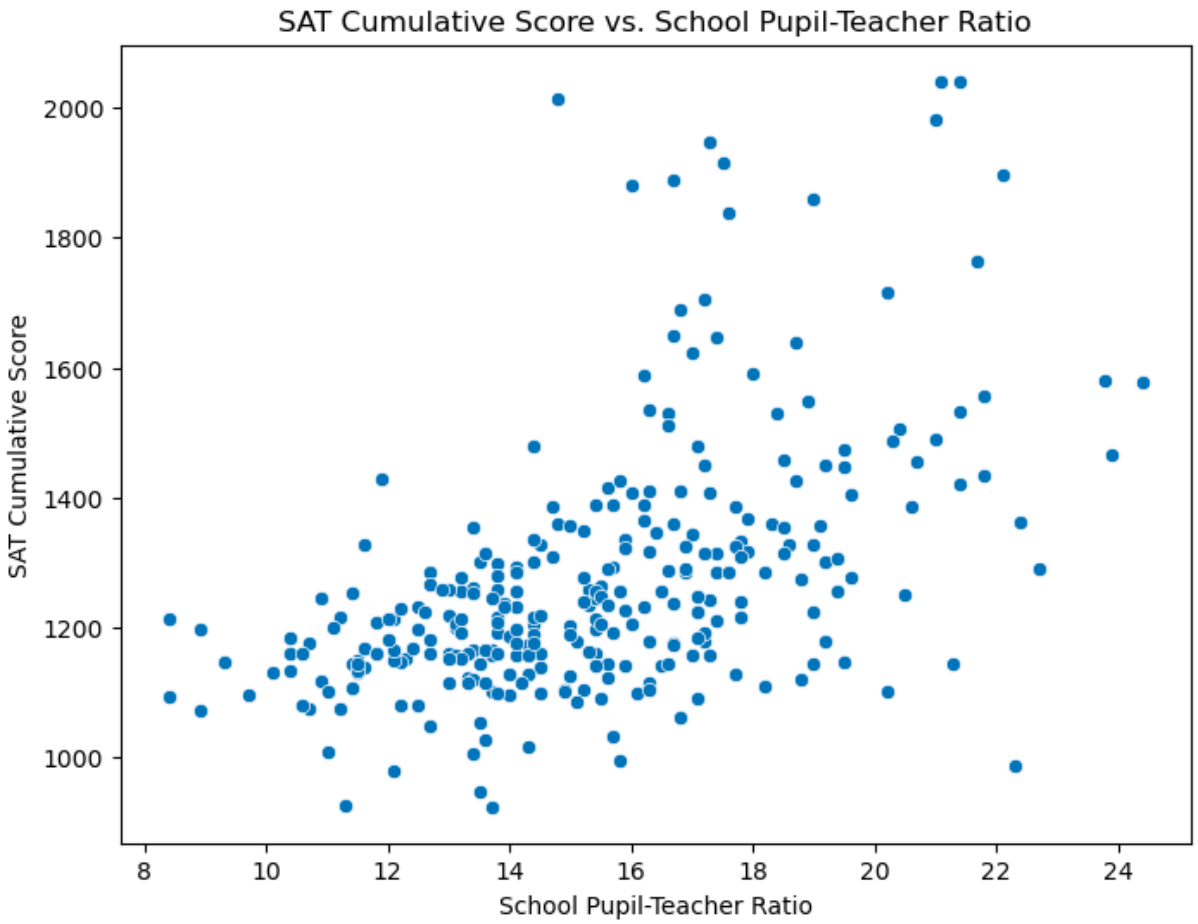
Given these specialized schools fall into a few different categories, including public secondary schools, specialized schools, technical schools, it is difficult to identify each one in the dataset and remove them from the dataset to focus on those public schools that are in the same zip code where a student lives. The reason for this phenomenon is that New York City has a process called a "Choice-Based System", whereby students can apply to any high school in the five boroughs.

That said, the correlation between SAT scores and average income per household was .24, which is very close to the nationwide correlation (per Google) of 0.22. If anything, there is a more pronounced correlation between SAT scores and average income per household that is lost in the noise of the mismatch between a student's household income and the average income of the zip code where they attend school.

It should be noted that other proxies for wealth that I used in the dataset, and which were focused around real estate, all yielded weak correlations,

2. School Pupil-Teacher Ratio

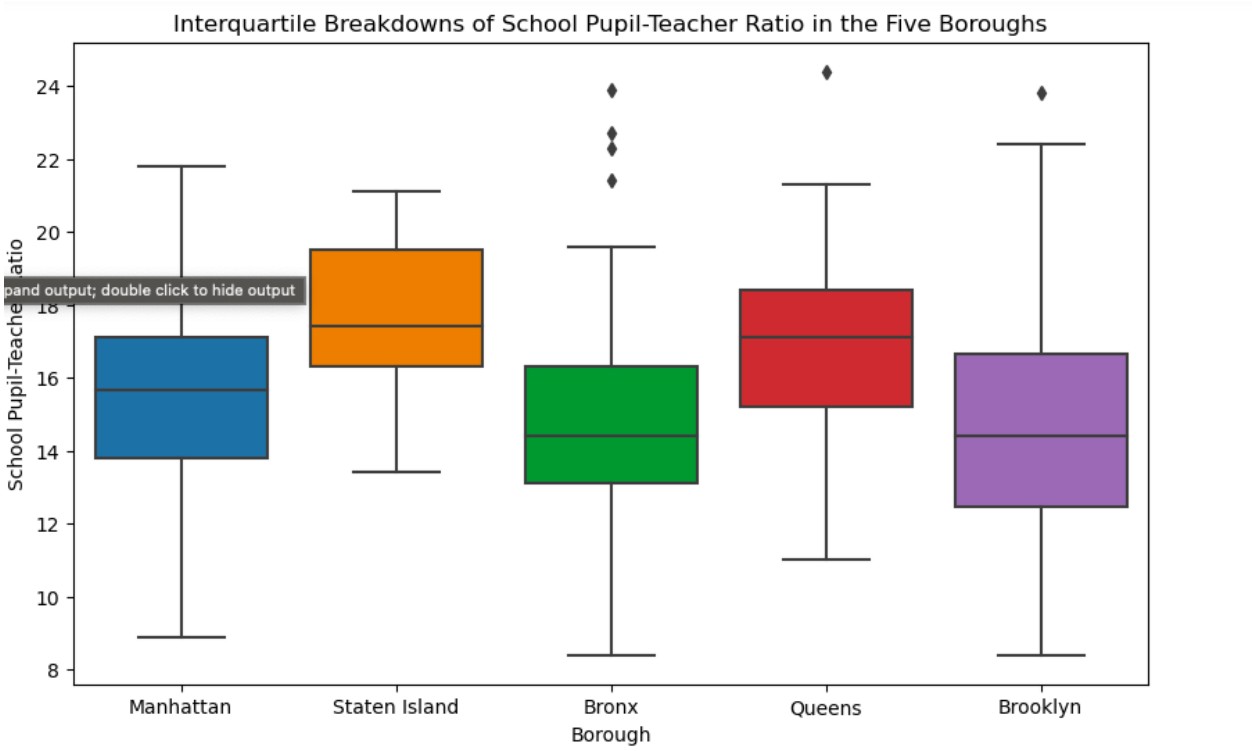
Given the above, I looked at other salient correlations from the data and the most surprising result was that there was a relatively strong positive correlation of 0.53 between SAT scores and school pupil-teacher ratio. One would assume that this effect would be reversed: smaller class sizes lead to better learning outcomes.



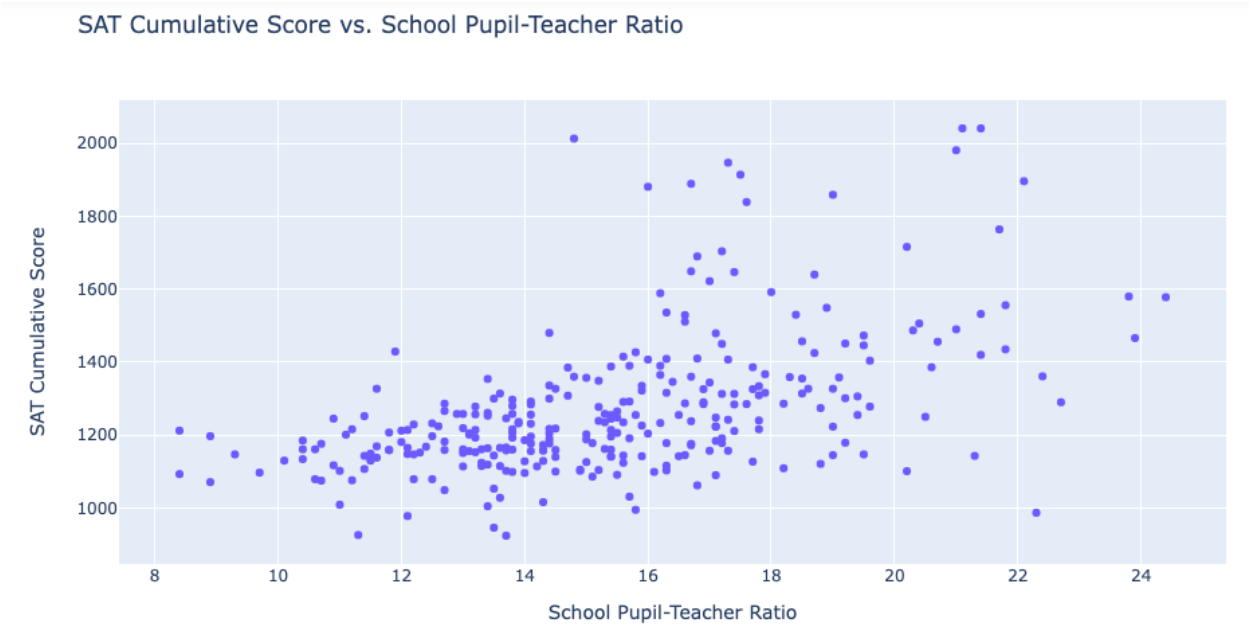
This paradox became clear when I used plotly to identify the specific schools that corresponded to each data point. Once again, the highly selective preparatory schools in which students' SAT scores are high, also have large classroom sizes, around 22:1 to 24:1. For example, the Bronx High School of Science has an acceptance rate of approximately 3% and is tied at first for the

highest average cumulative SAT score of 2041 in New York City, yet it has a School Pupil-Teacher Ratio of 21.4, which is amongst the highest in the city.

Just as these preparatory schools lead to the fact that average household income does not necessarily match the zip code where students go to school, so too do they not align with School Pupil-Teacher Ratio. Below, we can see the outliers in the Bronx, Queens, and Brooklyn, three boroughs that typically have smaller class sizes.



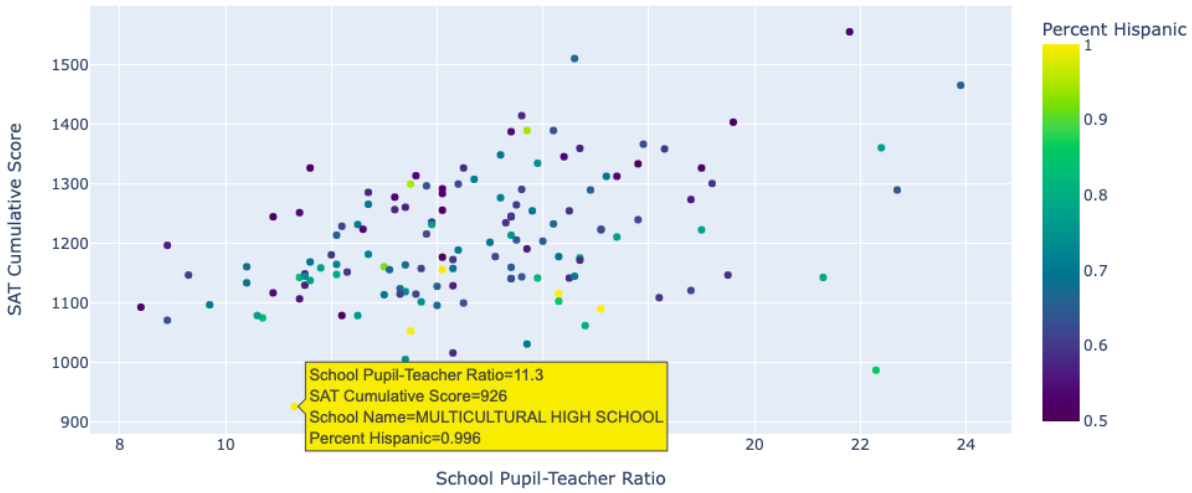
The schools with small classroom sizes, regardless of which borough they are in, tend to have lower SAT scores. Again, using plotly and a scatterplot, I was able to home in on specific schools to identify some patterns.



- Many of these schools are ELL, or English Language Learner, focused.
    - Given the SAT is difficult, even for native speakers, explains these school's low average scores.
    - These classroom sizes are small (some as low as
- Schools preparing students for a specific career (automotive, business, health) tend to have small class sizes (some as low as 8.9 students per classroom)
    - These schools are not as likely to be focused on college prep work.

3. Racial Dimensions

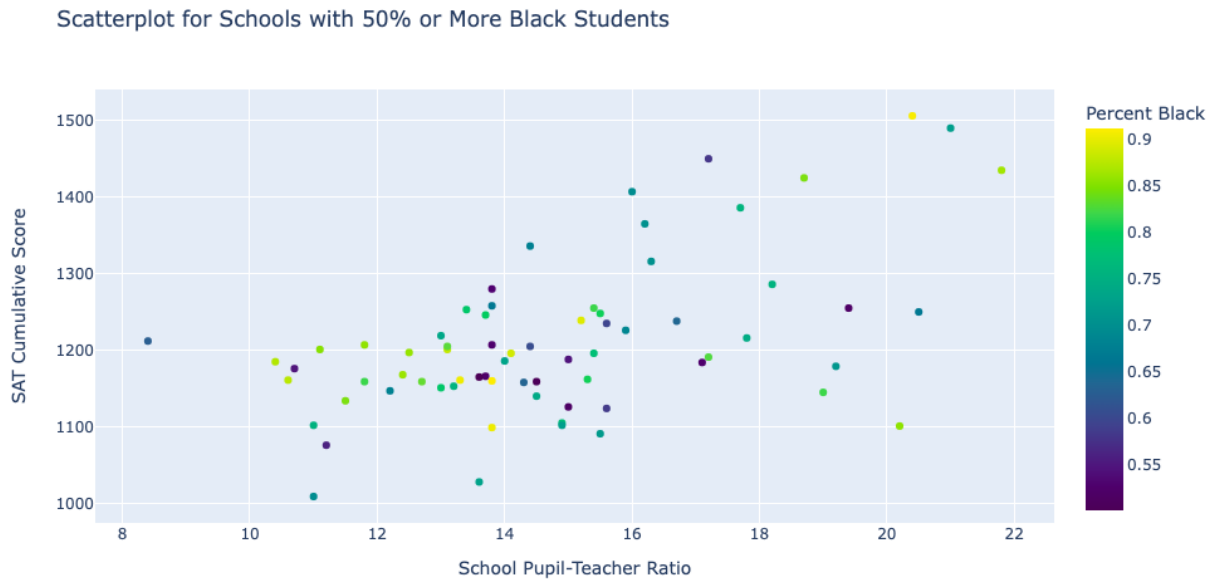
After discovering that many of the schools that had average SAT scores on the low end of the spectrum were international and multicultural schools (or ELL schools), I wanted to explore a little more deeply the connection between race and average SAT score in New York City. To do so, I analyzed schools in which Hispanic students comprised over 50% of the student body, to potentially account for a language barrier.



The school with the lowest average SAT score also had nearly 100% Hispanic students. The school's mission is "dedicated to helping English Language Learners (ELLs) develop their English proficiency."

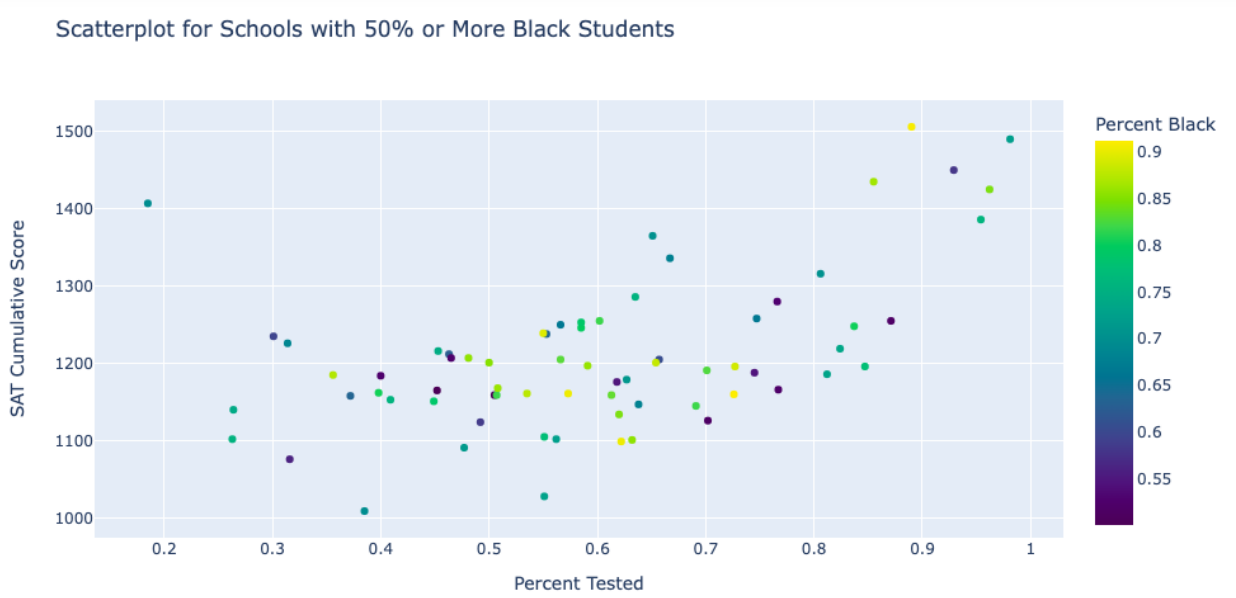
While almost all the yellow dots (nearly 100% Hispanic) are ELL institutes, this is not necessarily the case as you get closer to 50% enrolled Hispanic. Nonetheless, some language barrier could still exist even if students aren't enrolled in a school that explicitly bills itself as a language learning institute.

This phenomenon likely accounts for some of the observed correlation of -0.43 for percent Hispanic and SAT Cumulative Score. Again, the observed classroom size (School Pupil-Teacher Ratio) tends to be smaller in these institutes likely because the focus is on language acquisition.



Diving further into the racial dimension of the dataset, I created a similar scatterplot for black students, using plotly once again to identify a specific school to see whether any patterns emerge. Of the ten lowest scoring institutes, seven of them are specialty or vocational schools (automotive, aviation, sports management, etc.) The focus on specific vocational skills could account for the small classroom sizes, as well as SAT scores, given that the curriculum is likely not heavy on college preparation.

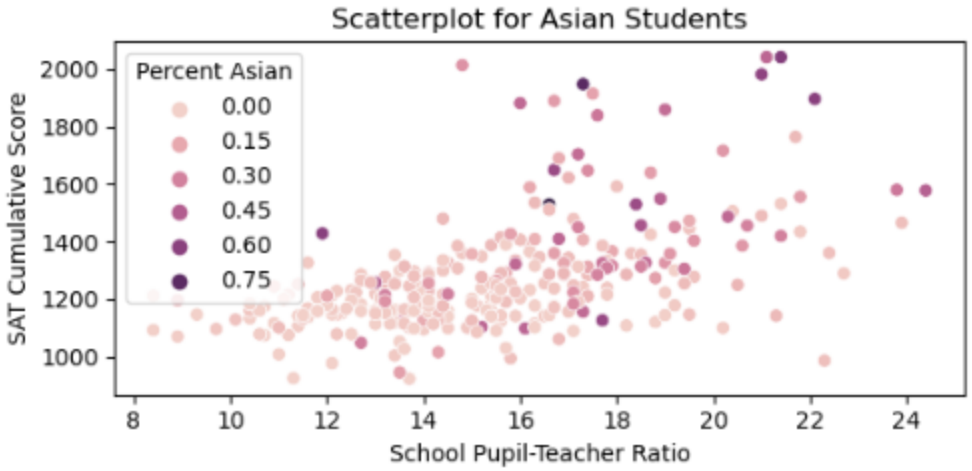
To explore this a little further, I delved into the percentage of students tested. I was curious what percent of students at these schools actually took the SAT. Percent Tested also has one of the highest correlations with SAT Cumulative Score, at 0.59 (I'll discuss Percent Tested in the next section.)

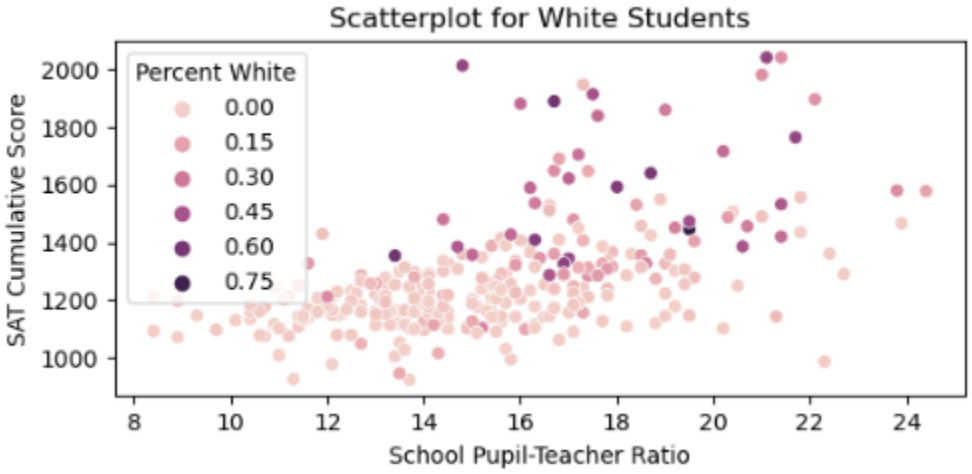


Four of the top five institutes for highest % tested have the words “college”, “preparatory”, “academy” in them. This is perhaps not surprising given that their focus is preparing students for college.

The vocational schools, on the other hand, had a much lower percentage tested, many were around the 50% mark and a few were close to 25%. This highlights the fact that these schools likely do not put much emphasis or resources on SAT and college preparation, given their focus on career specific skills in which college might not be necessary.

For white and asian students, the data tell a different story. Here, I used a similar scatterplot showing when a specific race comprises over 50% of a school. I included the target variable, SAT Cumulative Score, and School Pupil-Teacher Ratio, the latter being employed to show how white and asian students tend to be overrepresented in the elite preparatory schools.





These elite preparatory schools do not represent the racial diversity of New York City. To quantify this diversity, I used the dataset for this project (298 out of 537 schools in New York City are represented) to yield the following estimates:

```
Black Students: 34.77%
Hispanic Students: 45.32%
Asian Students: 10.81%
White Students: 9.09%
```

On the other hand, the racial makeup of students at two of the elite preparatory schools are as follows:

The Bronx High School of Science

Black Students: 3.7%
Hispanic Students: 8.2%
Asian Students: 61.4%
White Students: 21.5%

Staten Island Technical Institute
Black Students: 0.7%
Hispanic Students: 4.1%
Asian Students: 60.7%
White Students: 32.5%

The racial diversity of New York City's high school students is starkly at odds with the racial diversity of some of the top schools. One possible explanation might be that there is already an equal access to resources starting at the grade or middle school level. These inequities might then be further amplified by the placement tests at these elite preparatory schools. The SAT scores, then, could be construed as yet another test in a long line of tests that largely reflect a long history of unequal access to resources. But without data both around earlier placement tests and the types of resources afforded in middle school, before students opt into a high school, it is difficult to conclude more.

That said, there might be more complexity involved. For instance, there could be certain social and cultural factors that account for the high percentage of white and asian students that go beyond mere wealth (the correlation with average income is .23 and .02, respectively). The focus on supplementary education–tutors, cram schools–are a few examples. But again, without access to that kind of data, it is hard to draw any conclusions. And even with such data, correlations do not imply causation.

Modeling

The EDA highlighted the fact that there are several significant contributing factors to a student's SAT score that are beyond their control.

- Students use a "Choice-Based System" to select which high school in New York City to go to.
    - Many of these schools already have selection criteria, so a student's choice is actually limited.
    - The more selective the school, the more likely it is to invest resources that lead to higher SAT score outcomes (the Percent Tested feature points to this).
- Many students choose to go to vocational schools or English Language Learner schools, both of which have a focus that might not be aligned with college preparation.

In modeling the data, I set off with the following assumption: given that the features in the dataset are beyond a student's control, the more efficient a model is at predicting the target variable, the more this speaks to potential biases and inequities in a test. This logic might be a little inverted from those cases where we want to find the best model; here, the objective is to provide college admissions boards with data around how an SAT score can come down to unequal access to resources. In this context, a highly efficient model would argue against using the SAT as a crucial part of the admissions process.

Steps taken in the Machine Learning process

Preprocessing

- To ensure that the features in the dataset were quantifiable and continuous, I removed the following: City, State, School ID, School Name, Building Code, Street Address, Phone Number, Start Time, End Time.
- I also dropped the columns Average Score (SAT Math), Average Score (SAT Reading), and Average Score (SAT Writing) because the target variable is a composite of the three and this could lead to data leakage.
- I used one hot encoding on Borough and Zip Code.

Feature Engineering

To improve the predictive power of my model, I introduced some engineered features.

1. SAT scores per student
    - Rationale: this is a way to normalize SAT scores across schools of varying sizes. For example, a school with 2,000 students and a score of 1800 might differ from a school with 200 students and that same score.
2. Enrollment ratio interaction
    - Rationale: there could be an interaction between student enrollment and another relevant variable. For instance, pairing this with classroom size might highlight some nuanced relationships.
3. Tested-enrollment interaction
    - Rationale: it is possible that schools with a large student body and high percentage of test takers perform differently from schools with a large student body and a low percentage of test takers.
4. Diversity index
    - Rationale: in quantifying diversity, the aim is to gauge to what extent this factor might have on the target variable. I used a threshold of at least 15% student enrollment for each race to qualify a school as diverse - 1, vs. not diverse - 0.
5. High achievers
    - Rationale: determining which schools had "high" SAT scores—the top 10% of schools with the highest average SAT were included—can provide some potential insights in how this feature interacts with other variables.

I did an initial performance evaluation and realized, based on a test R^2 of 1.0, that there may be data leakage. I reviewed the engineered features and realized that two of these features—% high achievers and SAT score per student—contained the target variable, thereby potentially allowing the algorithm to derive the target variable via one of these features. Therefore, I removed % high achievers and SAT scores per student, during preprocessing.

Performance Modeling

I began by using multiple regression models to look at the interactions between different features. I chose this as a starting point to get a baseline for models before moving on to hyperparameter tuning and cross folds validation, i.e., other techniques that might optimize the performance explored using the more simple tools of multiple linear regression.

In order to explore the interaction between variables, I explored a variety of combinations, focusing on those variables that had high coefficients in the matrix/heatmap. I first started with two variables and, if the results looked promising, added more features to the multiple regression model. I used ‘School Pupil-Teacher Ratio”, given its high correlation, and then coupled that with another feature. One of these first couplings that stood out for its performance was ‘School Pupil-Teacher Ratio” coupled with ‘Diversity Index’, which yielded the following:

Training MAE: 114.23648006552924
Test MAE: 88.4527216374366
Training MSE: 26588.65502526878
Test MSE: 14798.110239279624
Training R^2: 0.35248739437742693
Test R^2: 0.3795080298116601

Next, I started adding more features to whichever two I chose, focusing on those that had given the strongest performance metrics when coupled. For instance, I added ‘Tested-Enrollment’ and ‘Enrollment-Ratio Interaction’ (both of which had performed well when coupled with ‘School Pupil-Teacher Ratio) to ‘Diversity Index’ and ‘School Pupil-Teacher Ratio’ to get the following:

----- Evaluation using 'School Pupil-Teacher Ratio', 'Tested-Enrollment Interaction', 'Enrollment-Ratio Interaction', and 'Diversity Index' -----
Training MAE: 96.59443441978308
Test MAE: 81.9883408935996
Training MSE: 18648.075330488577
Test MSE: 14869.571372816657
Training R^2: 0.5458640598550313
Test R^2: 0.3765116296752029

While strong, the Training R^2 was slightly higher than when using just the two variables. This trend played out in other iterations: adding more features to features that had performed well in isolation did not necessarily improve, and oftentimes impaired, performance. To illustrate, when I combined seven features with high coefficients, many of which had performed well when coupled with ‘School Pupil-Teacher Ratio’, led to numbers that suggested overfitting.

----- Evaluation using 'School Pupil-Teacher Ratio', 'Tested-Enrollment Interaction', 'Enrollment-Ratio Interaction', 'Diversity Index', 'Percent White', 'Percent Asian', and 'Diverse School' -----
Training MAE: 82.82962182276907
Test MAE: 88.77483601766082
Training MSE: 12835.043938359153
Test MSE: 17124.98691244702
Training R^2: 0.6874286143504118
Test R^2: 0.28194095753194903

Ultimately, the model that proved best amongst 15 multiple linear regression models was ‘School Pupil-Teacher Ratio’ coupled with ‘Diversity Index’ (see numbers above). This is potentially revelatory in that two indications of “systemic fairness” were the two features that led to the strongest model. Again, the stronger the model, the more likely SAT scores are determined by systemic factors outside of a student’s control. I will return to and elaborate on this in the conclusion section.

In order to improve performance of this model, I used Lasso and Ridge regularization. In addition to the model that employed ‘School Pupil-Teacher Ratio’, I also included one other model that was relatively promising, given the following performance:

----- Evaluation using 'School Pupil-Teacher Ratio' and 'Tested-Enrollment Interaction' -----
Training MAE: 112.47517210627686
Test MAE: 76.4930696203487
Training MSE: 26772.781677073228
Test MSE: 15366.575661652258
Training R^2: 0.3480033643292295
Test R^2: 0.35567199776375225

Going forward, I will refer to these numbers directly above as Set A and the model with ‘School Pupil-Teacher Ratio’ and ‘Diversity Index’ as Set B. This nomenclature is also reflected in the code.

Set B performed better than Set A for both regularization techniques and, more importantly, performed better than before applying regularization.

----- Evaluation for Ridge Regression Set B -----

Training MAE: 112.95359119921665
Test MAE: 85.56733960671772
Training MSE: 26716.638927364533
Test MSE: 14660.341637710986
Training R^2: 0.349370606790954
Test R^2: 0.38528473437968724

----- Evaluation for Lasso Regression Set B -----
Training MAE: 113.36751382537511
Test MAE: 86.37545236213354
Training MSE: 26641.18188786237
Test MSE: 14653.48837228359
Training R^2: 0.3512082094908341
Test R^2: 0.38557209513714275

Ensemble Techniques

At this point, given these numbers and improved performance, I moved on to more complex methods, such as Ensemble techniques, with the knowledge that they might not be appropriate given my small dataset (n = 297) and relatively small feature size.

Indeed, when Random Forest was applied to the Dataset, there was severe overfitting (R^2 was typically >0.95). This trend persisted when I used a Randomized CV Search to optimize hyperparameters.

Perhaps this result is not too surprising; again the dataset was small and assigning multiple folds and splits will likely result in overfitting. Even with hyperparameter tuning, the performance was as follows:

----- Random Forest with Best Hyperparameters Evaluation -----
Training MAE: 27.173650282030614
Test MAE: 76.9892694063927
Training MSE: 1432.545747779314
Test MSE: 13124.72575530118
Training R^2: 0.9651132624445811
Test R^2: 0.4496738562960646

While the MSE is low, even compared to the best multiple regression models, the training R^2 indicates overfitting to a degree that does not make this method viable.

I also tried techniques to reduce potential overfitting: Gradient Boosting and K nearest neighbors. However, both resulted in numbers that exacerbated the overfitting problem, suggesting that the issue was impervious to interventions such as these, most likely because of the small dataset.

Cross-validation

With this in mind, I returned to the multiple regression models that performed best—Set A and Set B, with and without regularization. Given that these models were so far the best performing, I wanted to obtain an even more reliable estimate of model performance. Therefore, I used cross-validation, with the intention of achieving even more robust results.

Using cross-validation with five folds, I achieved the following results with Set B:

R^2 for each fold: [0.74263963 0.47306598 0.66593018 0.66488597 0.71023562]
Mean R^2: 0.6513514764858449
Std R^2: 0.09378945646672339

Negative MSE for each fold: [-10857.76534505 -9283.46667227 -14778.21378512
-18908.06024149
 -12479.36388463]
Mean Negative MSE: -13261.373985712295
Std Negative MSE: 3358.526151105917

This indicates that the model can account for approximately 65% of the variance between the two variables ‘School Pupil-Teacher Ratio’ and ‘Diversity Index’, and ‘SAT Cumulative Score.’

I ran cross validation on Set A, with and without regularization, and Set B, with regularization, and the numbers for Set B (directly above), without regularization performed the most optimally. None, however, achieved nearly such strong performance as Set B without regularization, making this the clear “winner.”

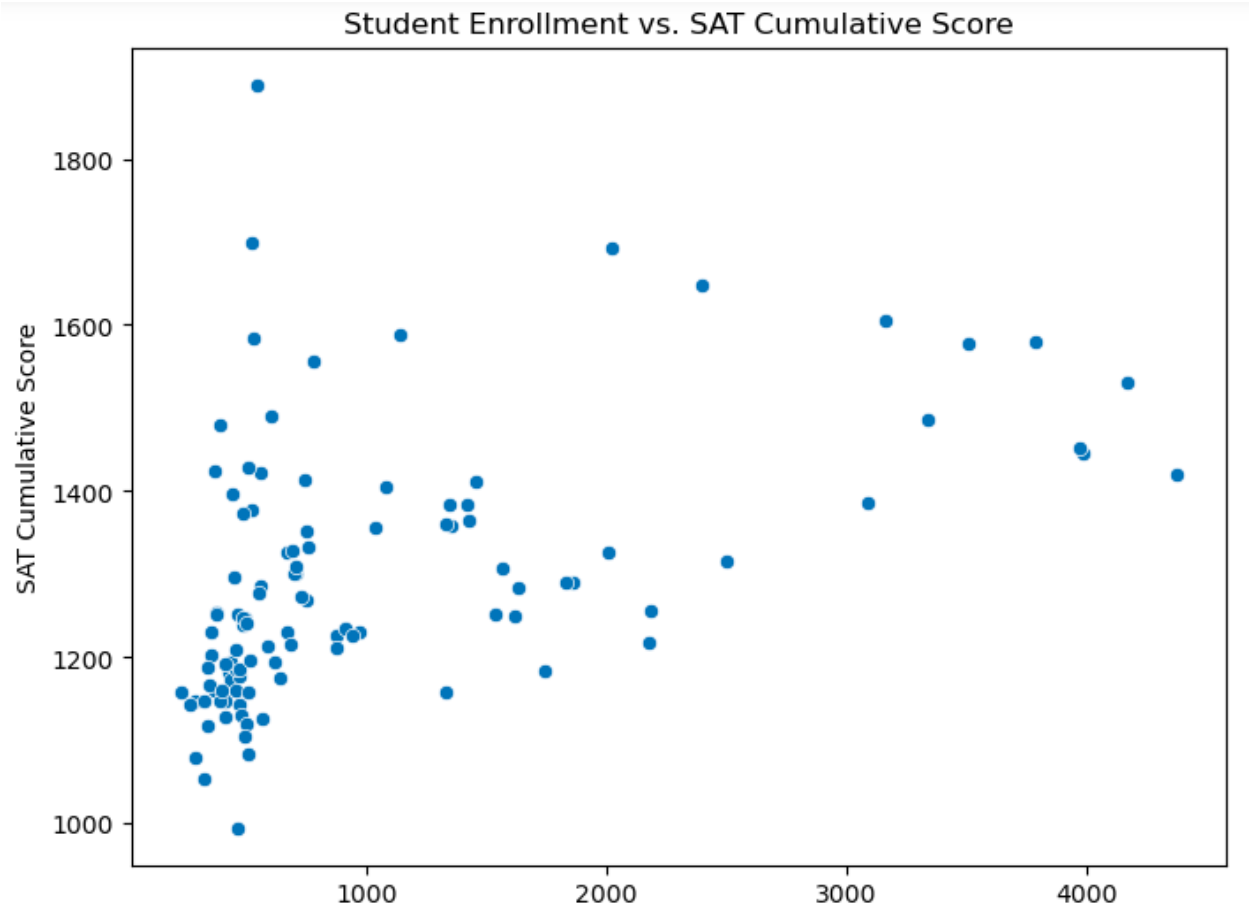
Takeaways

- The model that worked best relied on only two features—’School Pupil-Teacher Ratio’ and ‘Diversity Index’.
- More complex methods resulted in overfitting, with most cases yielding an R^2 in the test set that was >0.95.
- Hyperparameter tuning and regularization only had incremental improvements in performance or no improvement at all.
- The technique that worked best to improve the model was cross validation. However, this was only the case in one of the four instances attempted, the “winning” model.

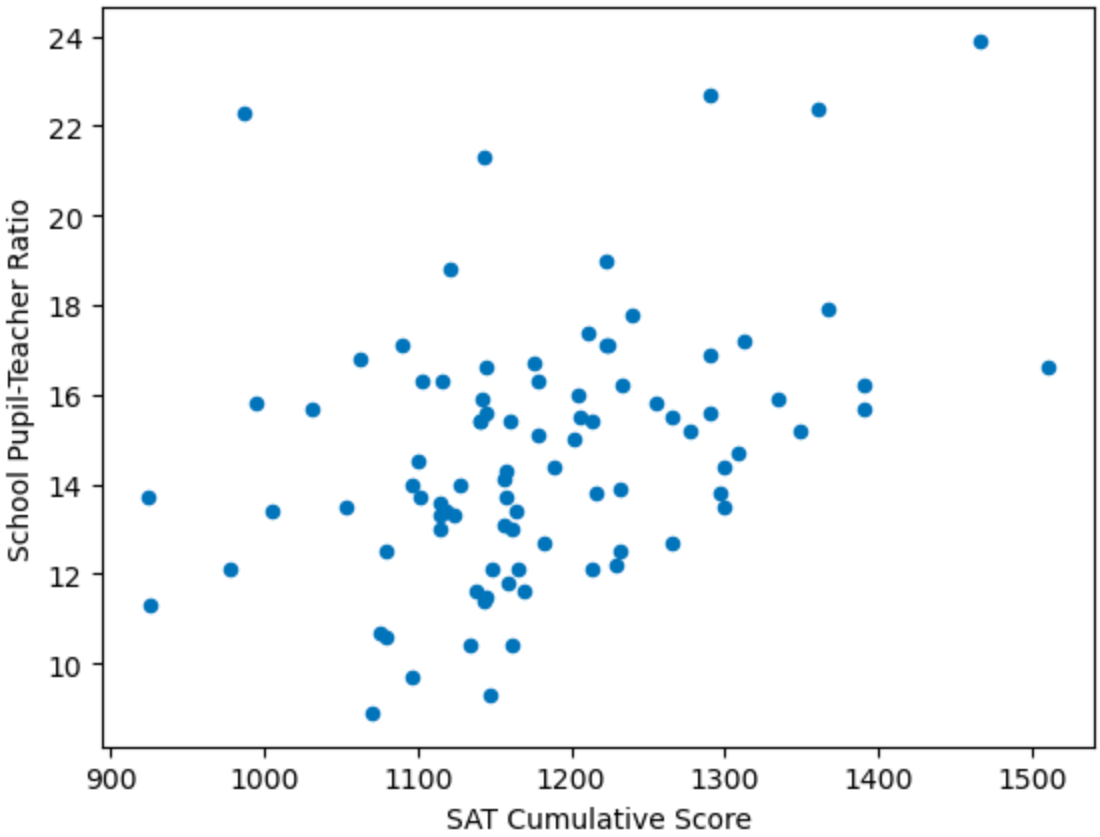
The first metric that stood out immediately, because I assumed it would be highly correlated, was the connection between average household income and SAT Cumulative Score. The connection here was in the same direction as expected but more moderate, at a .24. To calibrate my own assumptions on this correlation, and to get a quantifiable baseline, I googled this and learned that research had pinned the Pearson coefficient for income and SAT scores at around .22. It seems, then, that an SAT score could not exactly be “bought”, but that income does seem to tip, albeit slightly, the balance in favor of the wealthy.

Granted, there might exist a host of other systemic factors (e.g., smaller classrooms) that were factors beyond a student’s control. With this in mind, I looked first at the connection of Total Student Enrollment and the target value–SAT Cumulative Score. One might think that smaller schools would translate to smaller classroom sizes, which in turn would translate to more

attention and focus allocated to the individual student. The correlation of .38, however, showed that there was a somewhat moderate effect in the opposite direction.



One theory was that it might be possible that these large schools had vast amounts of resources and were thus able to have more teachers per student. The connection between School-Pupil Teacher Ratio and SAT Cumulative Score, however, showed the exact opposite: the more students in the classroom, the higher the SAT Cumulative Score.



The correlation was .53, which is significant, especially because one would assume the correlation be the inverse, or something closer to -.53. In fact, research has shown that the ideal classroom size for learning is 13-17 students per classroom; at 22-25 students per classroom, learning outcomes begin to suffer. In the scatterplot above, we essentially see the exact opposite effect.

To account for this paradox, I first looked at schools that were outliers, in other words schools that followed the standard assumption. Next, and at those schools that fell along the regression line. Those that fell along the regression line were essentially going against what research had said, and could therefore offer up some useful insights.

A pattern emerged: schools that fit the trend that buck conventional wisdom, as well as research, tend to fall into two categories: either trade schools or institutes that teach foreigners whose first language is not English. (I'll refer to the latter, going forward, as ESL institutions.) One example of these ESL institutes is Multicultural High School, which has small classroom sizes of 11.3, and has the second lowest SAT Cumulative Score of all the schools in the dataset at 926. This should perhaps not be surprising, given how rigorous the language demands of the SAT are, including the math section, a substantial portion of which is 50-60 words-long problem solving questions. The Pan American International High School (924) and the International High (946) School at Prospect Heights have similar classroom sizes and SAT scores.

Next, are trade schools, such as the Automotive High School and the Business Sports School, both of which are highly specialized skills and as a result likely have small classroom sizes. However, given that the focus is likely more on the specialization and less on general academic skills, these schools also have low SAT Cumulative Scores, albeit higher SAT Cumulative Scores (at 1076 and 1152, respectively.)

This phenomenon, however, does not account for the highest scoring schools, which have the largest size classrooms. Even if trade schools and ESL institutions have small classroom sizes, it does not necessarily explain the paradox of the big classroom and high score that upends research and prevailing wisdom.

Close inspection reveals that these schools are