Relationship Between School Demographics and Educational Outcomes: Evidence from New York City Public Schools

Abstract

Relationship Between School Demographics and Educational Outcomes: Evidence from New York City Public Schools

This paper examines the effect of specific school demographics on students' academic performance and outcomes in New York City public schools. We use variables such as racial background, location in the city, income level, and other factors to evaluate our dependent variable, SAT scores. Our findings indicate that a one-unit increase in the percentage of Asian students in a school is associated with an approximate 10-point increase in SAT scores. Additionally, when controlling for multiple effects, a 10% increase in income is associated with approximately a 4-point increase in SAT scores. Population data is found to be the least significant among the variables. These findings suggest that policies should be updated to help regulate visible inequities in the education system.

I. Introduction

In recent years, standardized testing has faced increasing scrutiny, particularly following the COVID-19 pandemic, with some universities becoming test-optional schools. The debate around standardized testing, particularly regarding SAT scores, is rooted in the question of inequality in the access to opportunities and resources.

Numerous studies have been conducted to investigate the relationship between school demographics and academic outcomes. For example, dRocco d’Este and Elias Einiö's paper, "Asian Segregation and Scholastic Achievement: Evidence from Primary Schools in New York City," found that an increase in the share of Asian students in a high school correlated with a decrease in SAT scores for

students of other races. Also, Atila Abdulkadiroglu, Weiwei Hu, and P. Pathak's paper, "Small High Schools and Student Achievement: Lottery-Based Evidence from New York City," revealed that smaller class sizes were often associated with higher acceptance rates, as students received more personalized attention.

Our paper seeks to explore the relationship between school demographics and average SAT scores in New York City public schools, drawing on a range of predictors such as income, location (borough and zip code data), population size, and a student's racial background. We wish to unify all the different studies that look at individuals' predictors of scores into one comprehensive study with clear visualisations and support. The focus on New York City is attributed to the city's impressive diversity of cultures, religions, ethnicities, and income brackets. Data is not homogenous and repetitive. Using data analysis techniques such as summary statistics, visual graphs, mapping, and regression, we aim to identify factors that impact SAT scores and provide insights into potential interventions and improvements that could be made to the education system, particularly in New York City.

In the following sections, we will present a detailed analysis of our data, before discussing the implications of our findings and suggesting potential areas for further research.

II. Data

2.1 Data source and collection method

Most of the data used in this project was obtained from the NYC Open Data portal. Highschool characteristics were provided by the New York City Department of Education, while SAT scores were provided by the College Board. New York City income information was retrieved from the US Census Bureau. Population per zip code data was retrieved from the New York Demographic website. Finally, we scraped information from a USA ESTA article to get insight into areas viewed as dangerous in NYC.

2.2 Variable description and unit of measurement

The Y variable in our study are average SAT scores, which measures a student's academic performance in a given school. The SAT scores, in this context, are marked out of 2400 (800 points per subject), which nowadays have been updated to be marked out of 1600. They evaluate the following competencies: Mathematics, English Writing and English Reading. SAT scores serve as a benchmark for educational quality and play a crucial role in determining college prospects.

The X variables include the demographic makeup of the school, such as the share of minorities, location data (borough, zip codes). It also includes median income per zip code ($USD), population per zip code data and information on neighbourhood safety. These variables were selected as they can impact the educational environment and affect student performance.

For instance, studying the different boroughs can provide information on the socioeconomic status of the school's area and the share of minorities can give insight into the diversity of the student body and cultural background, which can impact the learning environment. Unsafe neighborhoods can affect the school's environment and student performance, while population and income can give insights into the school's community and resources available to students. Is there a relationship between higher population and higher scores, possibly due to greater access to resources for supporting students?

The X variables are important for the analysis as they provide insight into the relationship between school demographics and academic performance. By understanding these factors, we can improve educational outcomes for students in NYC public schools.

2.3 Sample size, sampling method, and data cleaning/transformation

We started out originally with a dataset containing 435 values representing 435 schools in NYC. To clean the data, we dropped irrelevant columns and missing values. We merged the dataset with income and population data. We also scraped data from a website to obtain information on neighbourhoods with safety problems. All additional data was added into a new data frame, which was then used for visualisations and regressions. To improve data manipulation, percentage signs (%) were eliminated and

strings were converted to be numeric. We added summary information on the Average Overall SAT Score into a new column. Our final dataset contained 374 observations.

III. Visualization and Summary Statistics

3.1 Summary Statistics for Variables Used in Analysis

	Percent White	Percent Black	Percent Hispanic	Percent Asian	Average Score (SAT Math)	Average Score (SAT Reading)	Average Score (SAT Writing)	Average Total SAT Score	Median Income
count	374.000000	374.000000	374.000000	374.000000	374.000000	374.000000	374.000000	374.000000	374.000000
mean	8.524599	35.387166	43.929679	10.412567	432.719251	424.342246	418.286096	1275.347594	75087.473262
std	13.359205	25.367159	24.495584	14.400556	71.916833	61.884529	64.548388	194.866056	39772.439805
min	0.000000	0.000000	2.600000	0.000000	317.000000	302.000000	284.000000	924.000000	21846.000000
25%	1.300000	16.400000	20.825000	1.600000	386.000000	386.000000	382.000000	1157.000000	45616.000000
50%	2.600000	28.750000	45.300000	4.200000	414.000000	412.500000	402.500000	1226.000000	65908.000000
75%	9.375000	50.100000	63.375000	11.150000	457.250000	444.500000	436.000000	1327.000000	98177.000000
max	79.900000	91.200000	100.000000	88.900000	754.000000	697.000000	693.000000	2144.000000	250000.000000

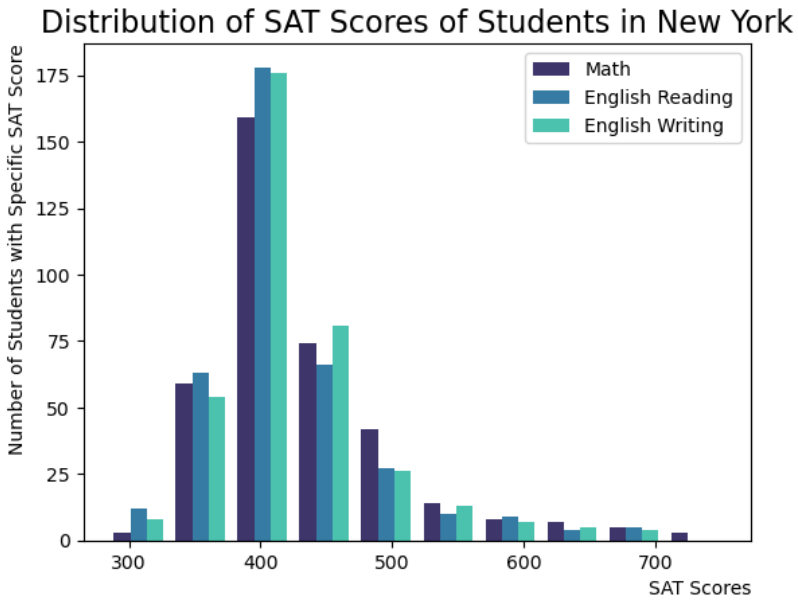
This table provides a summary of relevant data points. Turning our attention to the SAT scores, we see that the average SAT scores for different subjects' range between 418 and 432 points. Although the scores are similar, there is a slight trend towards higher math SAT scores and lower English writing SAT scores. Overall, when adding all three areas tested together, the average total SAT score has a mean of 1275 points.

Regarding the racial composition of the schools, our data set includes 374 observations, which represent 374 schools in NYC. On average, the student population is primarily composed of Hispanic students (43.9%), followed by Black students (35.3%). The remaining students are made up of Asian students (10.5%) and White students (8.58%). The high representation of Hispanic students in these New York schools is noteworthy. We will further explore whether there is a correlation between the racial background of the students and their performance on the SAT.

Finally, looking at the Median Income column, the lowest earners make about $22,000 while the highest earners make upwards of $250,000. This represents quite a significant difference. We will investigate the geographical distribution of high and low-income earners in NYC.

3.2 The Dependent Variable: Sat Scores

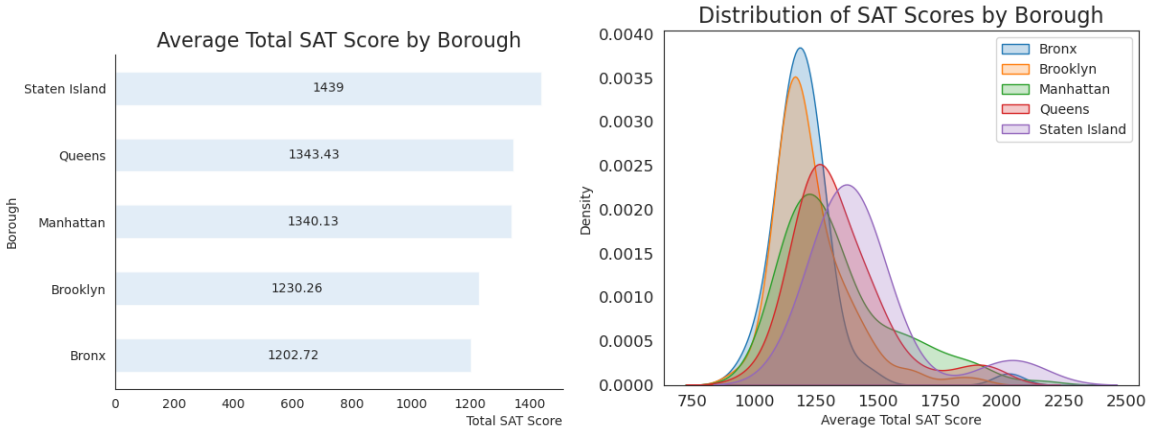
Figure 1: Distribution of SAT Scores of Students in New York



On this histogram, Figure 1, we observe how scores vary over a range of 300 to 700 points per subject and is positively skewed. Most of the data is centered around the 400-point mark. Between 160 to 175 schools score 400 points in either Mathematics, Reading or Writing. When considering scores at the 500-point mark, which represent a higher level of achievement, students performed better in mathematics than in English reading and writing. This supports the trend we observed in our summary statistic. Specifically, roughly 50 students received 500 points in mathematics, whereas 32 students achieved the same score in the other two subjects.

3.3 SAT Scores and Location

Figure 2 and 3: Distribution of SAT Scores Across NYC Boroughs



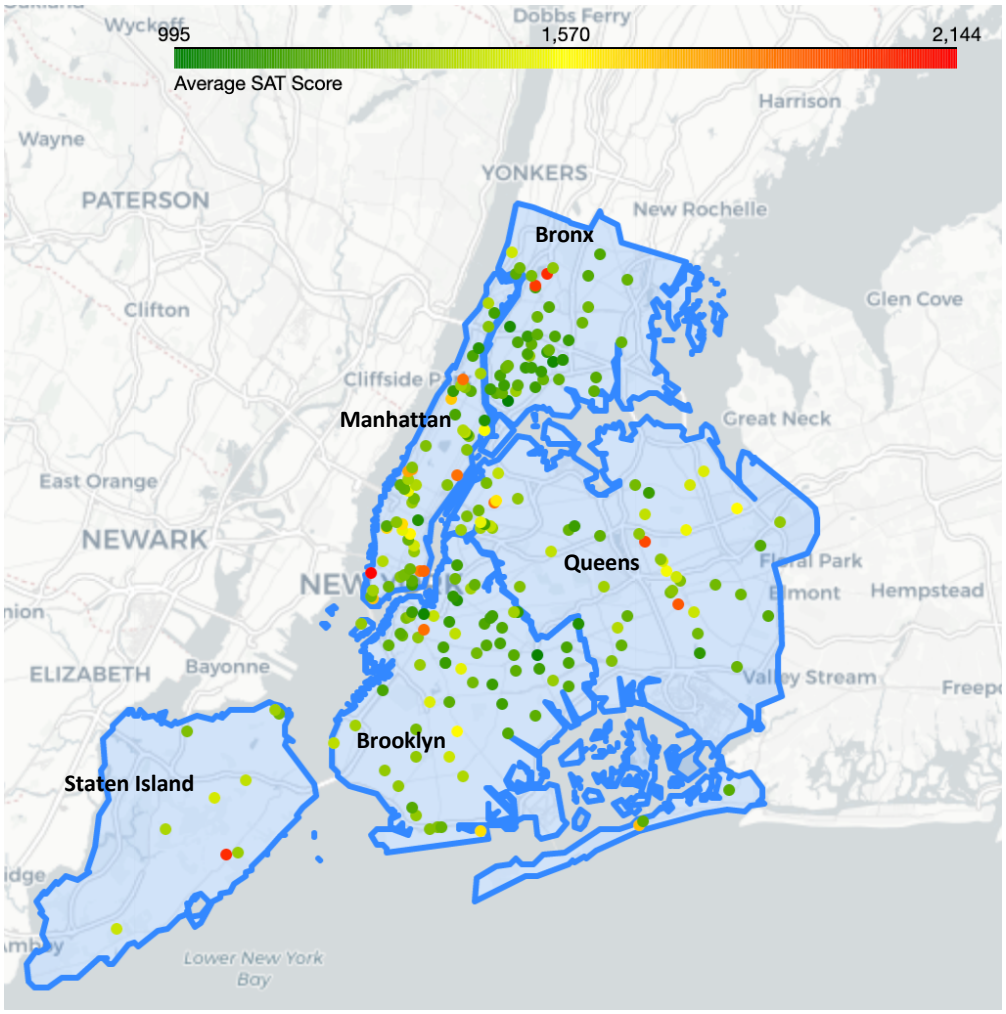
Figures 3 and 4 provide insights into the distribution of SAT scores across different boroughs in New York City. We first notice the trend towards an average overall SAT score of 1200 points. Most schools fall within the 1000-point to 1400-point range.

It is worth noting that Manhattan, known for having the highest real estate prices among the boroughs, has an average SAT score of 1240.13 points. In contrast, the Bronx and Brooklyn, which have the highest poverty rates at 24.4% and 17.8% respectively according to an article in 'The City' (1), seem to have the lowest average SAT scores at 1202.72 points and 1230.26 points respectively.

On the other hand, Staten Island and Queens have the lowest poverty rates at 10.6% and 10.3% respectively and the highest average SAT scores at 1439 points and 1343 points (1). This suggests a possible correlation between poverty rates and SAT scores in these boroughs. Staten Island's SAT score is 100 points above Queens and Manhattan, who have very similar scores, and 200 points above Brooklyn and Bronx!

There seems to be a relationship between school location and SAT scores, as schools in certain areas tend to perform consistently better or worse than others.

Figure 4: SAT Score Distribution Heatmap of NYC Boroughs

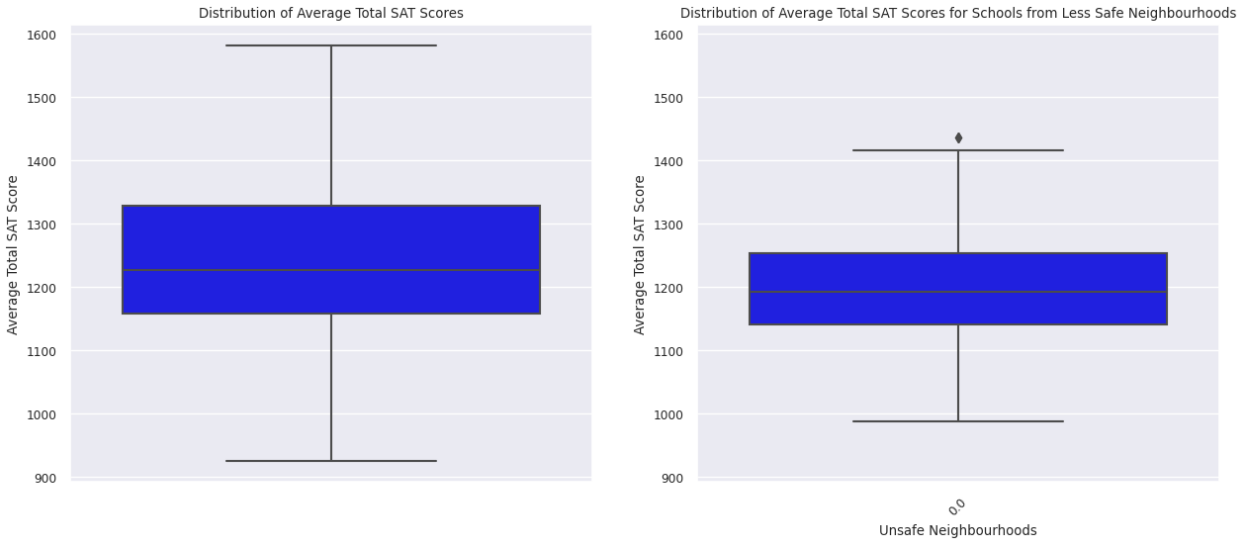


We have generated a heatmap that shows the distribution of SAT scores across NYC. This map visualizes the relationship between our Y variable, SAT scores, and our X variable, location (boroughs). "The map predominantly displays green dots, indicating an average score of around 1300 points. However, upon closer examination of individual boroughs, such as Manhattan and Queens, a concentration of yellow, orange, and red dots becomes apparent, indicating higher average scores in these regions compared to others. In fact, we observe eleven schools in Manhattan with SAT scores above 1500.

In contrast, the Bronx stands out as having the highest number of schools with SAT scores ranging from 1000 to 1300 points, with only two schools surpassing this range. These findings are consistent with our previous analysis in Figure 2, which suggests that schools in economically disadvantaged boroughs tend to perform comparatively lower than those in more affluent areas.

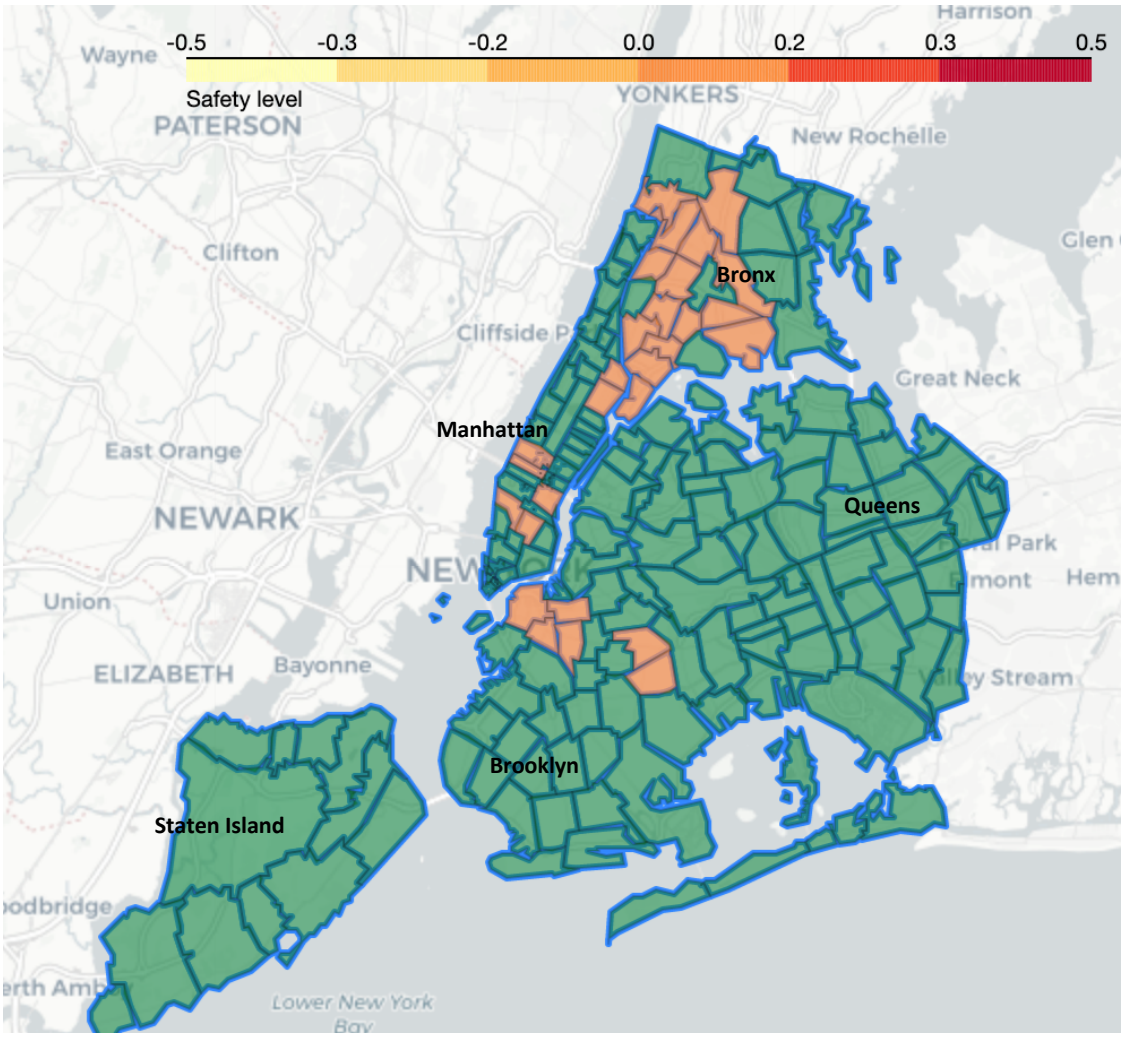
Figure 4 also sheds light on the high scores reported in Staten Island. We note that only 10 schools are represented, but they have significantly higher scores, which explains the borough's overall high average score.

Figure 5: Relationship between Neighborhood Safety and SAT Scores in NYC



By analyzing the boxplots, we observe that the median Average Total SAT Score of schools situated in less safe neighborhoods is lower than the overall median SAT score across all boroughs. While a few neighborhoods like Midtown and East Harlem have higher medians, most of the other neighborhoods have medians below the city average. This implies that students hailing from disadvantaged neighborhoods are associated with lower SAT scores compared to the city average.

Figure 6: Map of Unsafe Neighborhoods in NYC



By using this map, we gain further insight into the relationship between unsafe neighborhoods and lower SAT scores. The concentration of unsafe neighborhoods in the Bronx, Brooklyn, and Manhattan is particularly noteworthy. It's interesting to note that the Bronx, which has the lowest average SAT scores in our dataset, has a particularly high concentration of unsafe neighborhoods.

This observation suggests that the location of schools in less safe neighborhoods could be a contributing factor to lower SAT scores. These schools may face greater challenges in terms of resource

access and the socio-economic background of the students, making it more difficult for them to achieve high scores on standardized tests like the SAT.

3.4 Sat Scores and Racial Demographics

Figure 7: Correlation between SAT Scores and Race: Schools with White vs Hispanic Students

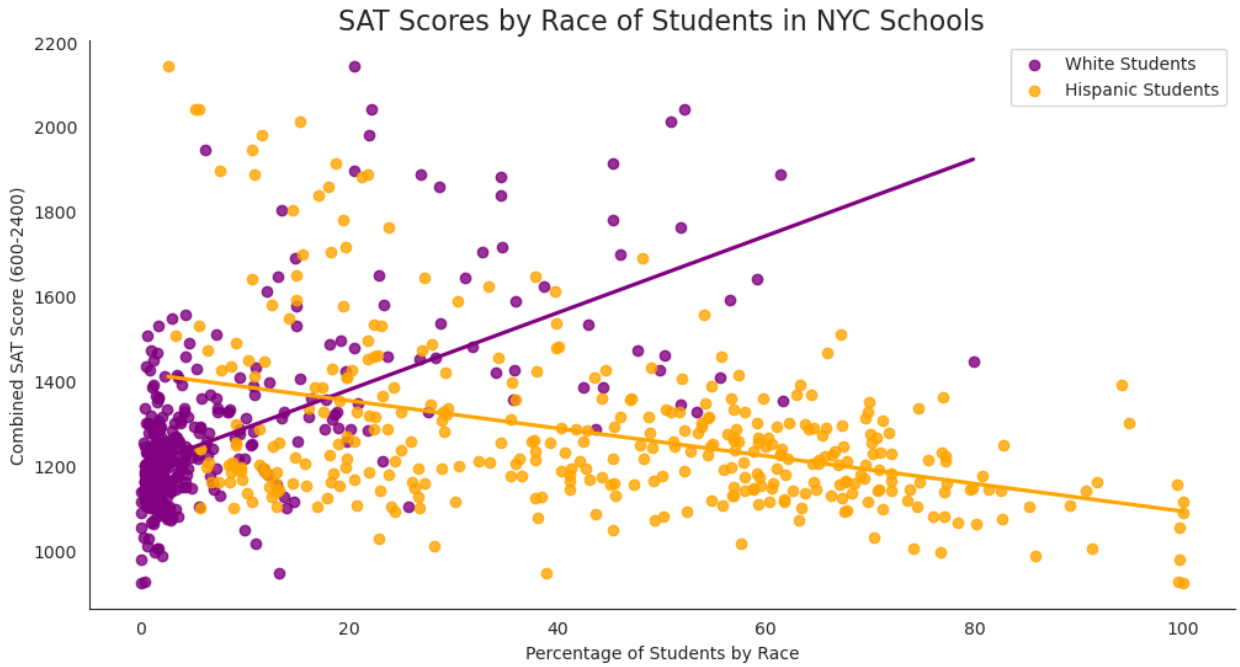
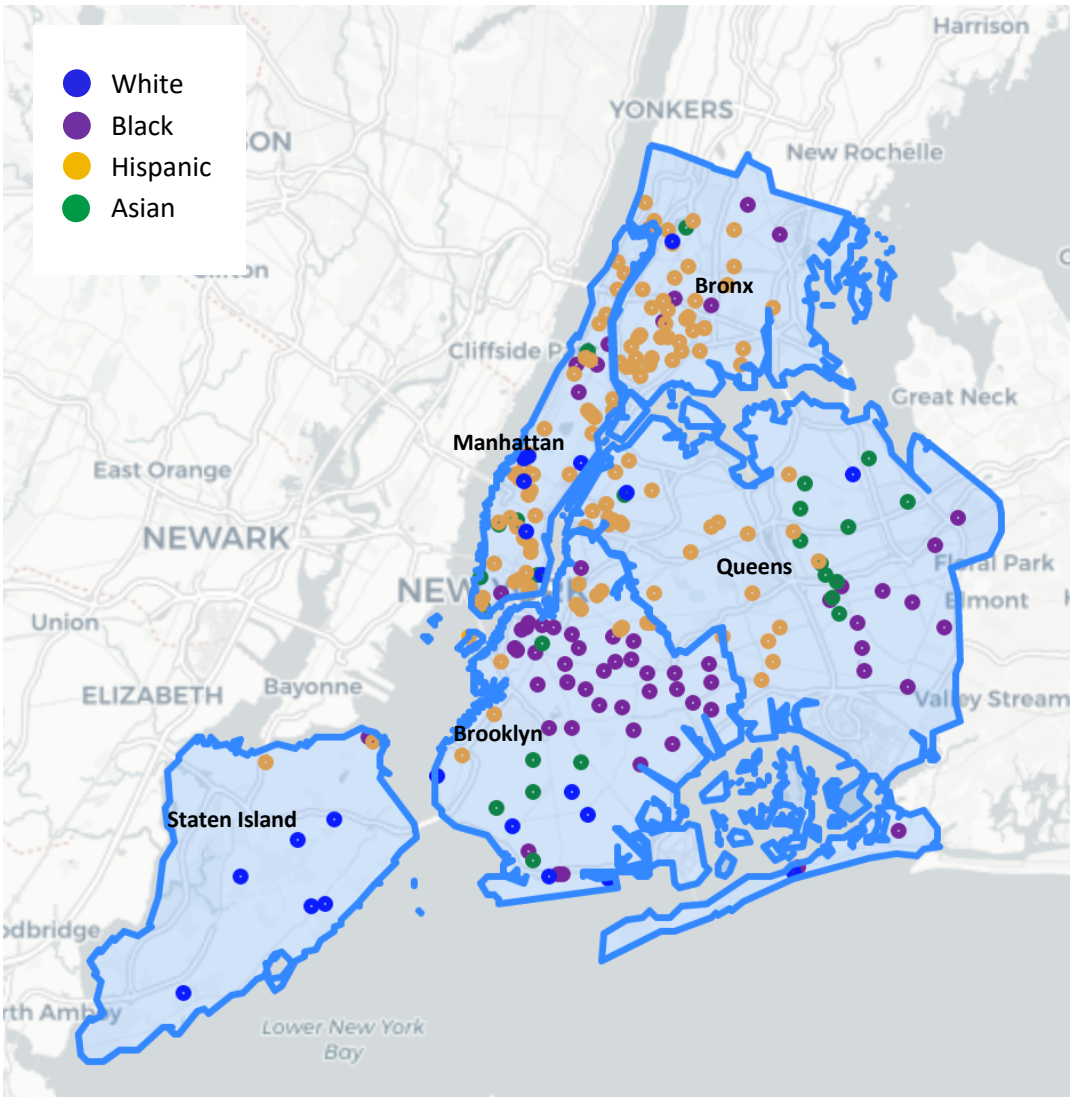


Figure 7 reveals a correlation between SAT scores and the racial background of a student. Schools with more white students have higher SAT scores - modelled by the purple line - while schools with more Hispanic students have lower SAT scores – fitted by the orange line. Despite a noticeable cluster of points close to the origin and some variance in the data, the correlation is evident.

Our analysis of the data brings to light a disheartening phenomenon. Most schools that predominantly comprise Hispanic students perform worse than schools with a more modest representation of white students. This trend may be attributed to the fact that schools with a higher proportion of minorities generally receive less funding. The disparity observed here underscores the potential unfairness

of standardized testing as an evaluation tool, given that schools with limited resources often produce less favorable outcomes than those that are more well-equipped.

Figure 8: Map of NYC Boroughs: Racial Background by Zip Code and SAT Scores



For Figure 3, we generated a map that visually represents schools in New York City, with color-coded markers indicating the predominant racial demographic of each school. It serves as one of our X variables and helps us better understand its relationship with our Y variable, SAT scores.

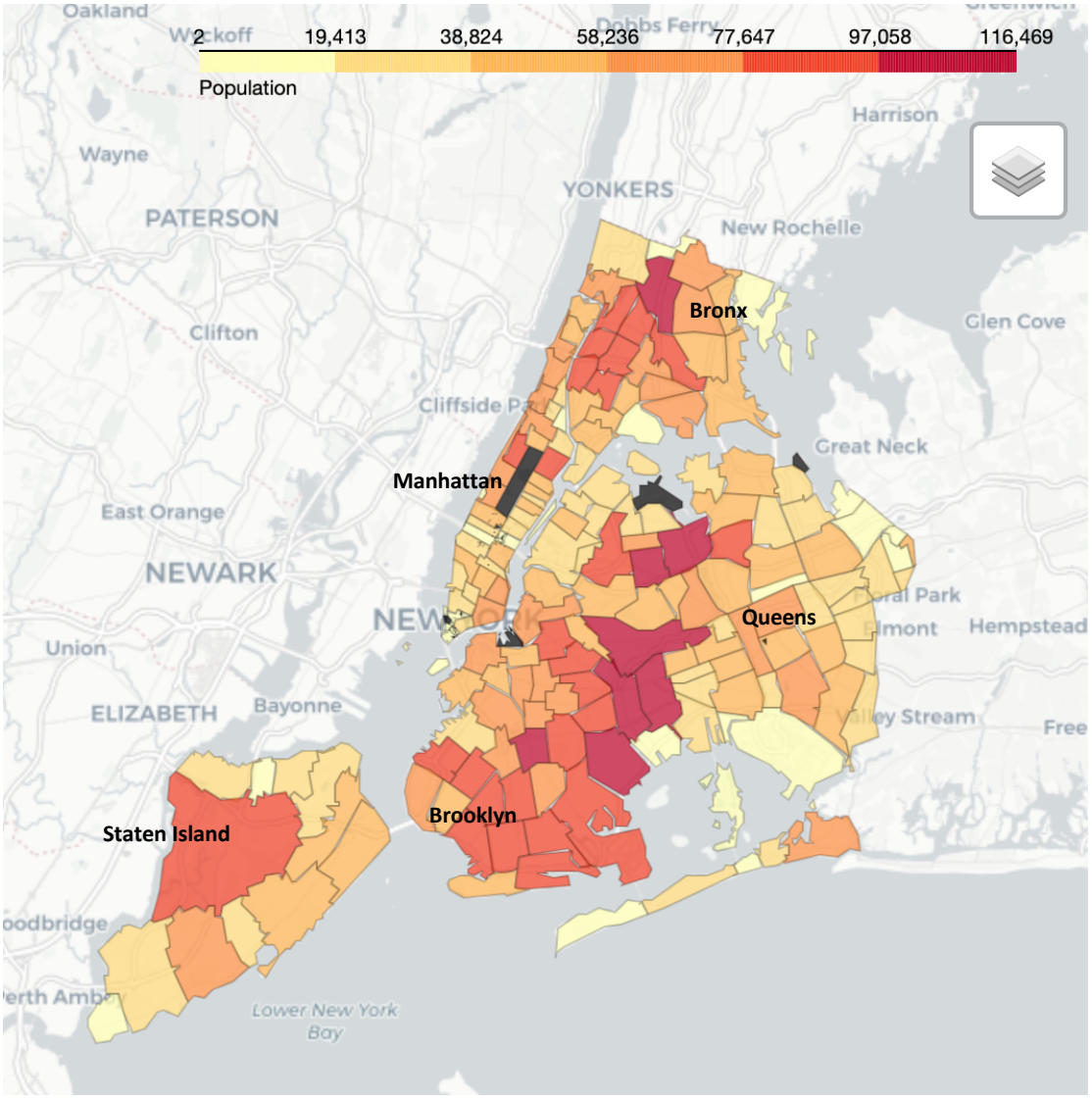
We observe that Staten Island has the highest concentration of predominantly white schools. As evident in Figures 2, 3, and 4, it also had notably high scores compared to other boroughs. Furthermore, based on our regression analysis in Figure 7, schools with predominantly white students tended to perform better. Could the racial composition of these schools in Staten Island influence testing, or are there other factors at play?

However, ethnic group distribution varies significantly across other boroughs. In Manhattan and the Bronx, most schools have a major Hispanic student population, while Brooklyn has a higher concentration of schools with black students. Queens has a more diverse mix of predominantly Black, Hispanic, and Asian student populations.

Although it might be tempting to compare schools with a certain ethnic makeup to their SAT scores, it's essential to consider that individual student performance is influenced by various factors, including socioeconomic status, access to resources, and educational opportunities. These factors can differ significantly within schools, regardless of their predominant racial demographic.

3.5 Sat Scores and Population

Figure 9: Exploring Population Density and SAT Scores Across NYC Boroughs

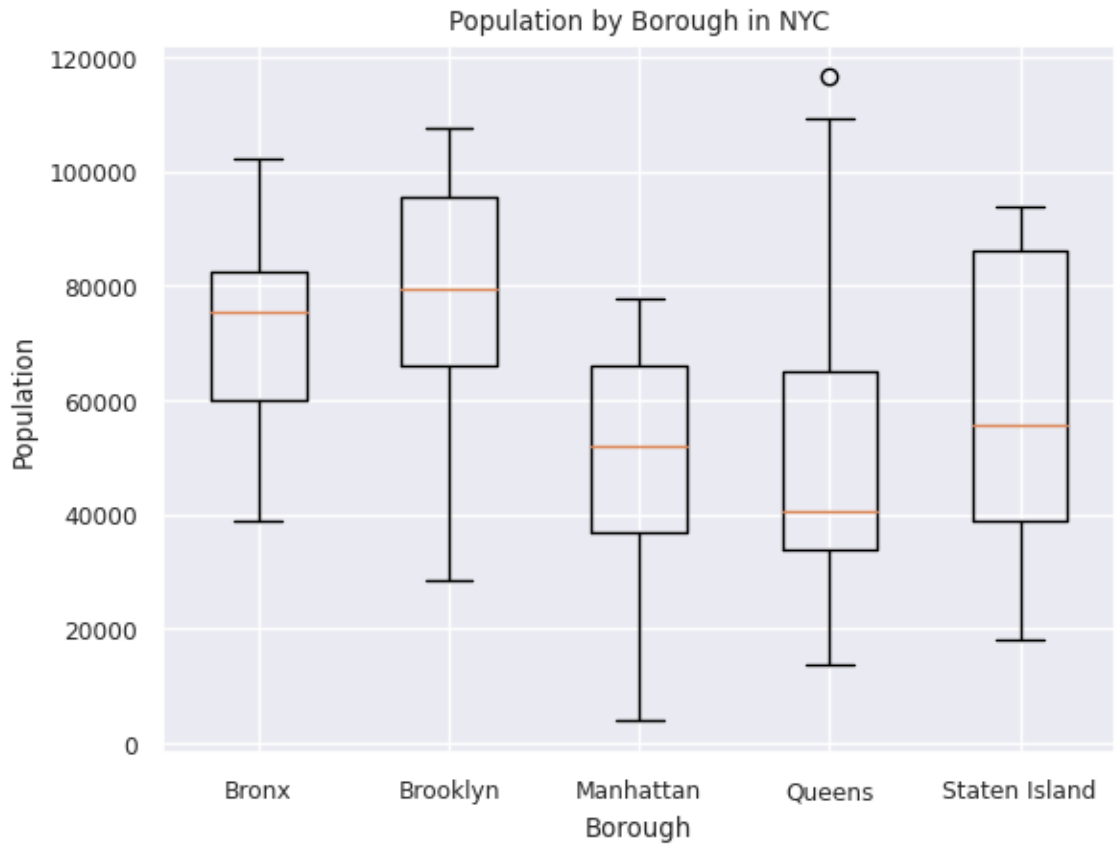


Here we study population distribution in NYC. The visualization clearly shows that boroughs such as the Bronx and Brooklyn have a larger population than the other boroughs, which is indicated by the darker red areas. In contrast, the east of Queens and the south of Manhattan seem to have a much smaller population than the rest of New York.

This observation is interesting, as we had previously hypothesized that higher population could lead to higher SAT scores due to the availability of more resources. However, the data shows that the

boroughs with the highest average SAT scores, such as Staten Island, Manhattan, and Queens, are less densely populated than areas in Bronx and Brooklyn.

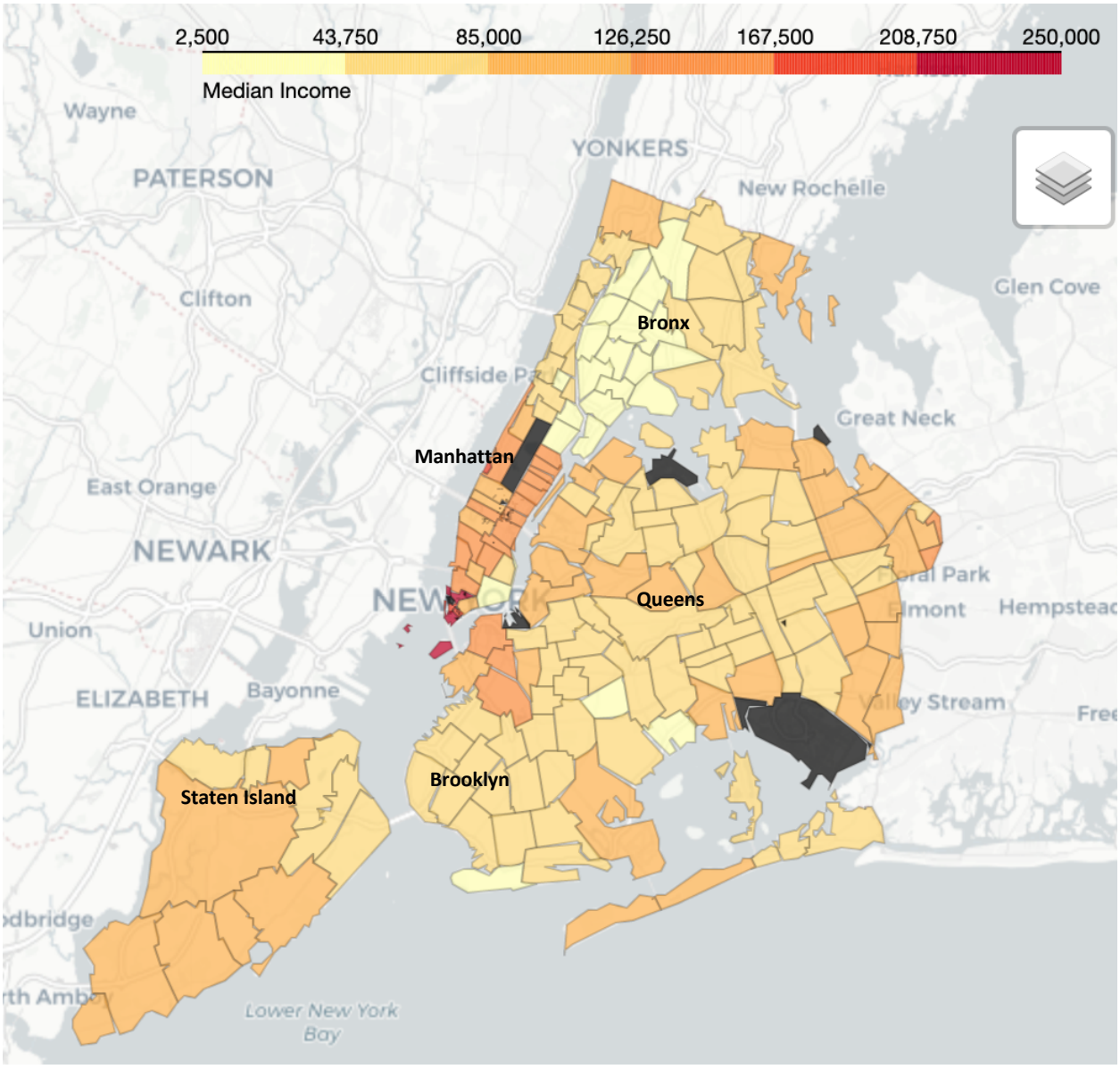
Figure 10: Boxplot of the Population by Borough in NYC



This boxplot acts as an extension of the map to clearly visualize population size. Brooklyn and Bronx have a higher median population than the other neighbourhoods, about 80,000 and 75,000 people respectively.

3.6 Sat Scores and Income

Figure 11: Exploring Income Distribution and SAT Scores Across NYC Boroughs



By examining the income distribution across NYC using Figure 11, we observe that income is generally evenly dispersed throughout the city. However, there is a clear disparity between the boroughs of Manhattan and the Bronx. Bronx has a higher concentration of zip codes with lower incomes ranging from low-income to moderate-income levels, while Manhattan has more zip codes with high-income levels, including median incomes well above $200,000, particularly in the southern tip of the borough.

Throughout our analysis, we have found that the Bronx has a higher population, lower income brackets, and lower SAT scores, with a population consisting mostly of Hispanic and Black individuals. On the other hand, Manhattan has a lower population, higher income brackets, a more diverse population, and higher SAT scores. Therefore, the income disparity between these two boroughs may be one of the contributing factors to the differences in academic performance between the schools in these areas.

3.7 Summary of Section Findings

From our visualizations in this section, we have identified several relationships between our dependent and independent variables. Overall, SAT scores in our analysis varied between 1200 and 1300 points. As we investigated factors that may influence higher or lower scores, we found that location played a significant role. Specifically, the Bronx and Brooklyn showed a trend towards lower scores, while Queens, Manhattan, and Staten Island trended higher.

Further analysis incorporating unsafe neighborhood data revealed that many of these neighborhoods were concentrated in the Bronx, potentially contributing to the lack of resources and challenges in achieving academic success in that area. Additionally, we observed differences in income distribution, with some areas, such as Manhattan and the Bronx, showing significant disparities. This finding aligns with previous research, including a study by Atila Abdulkadiroglu, Weiwei Hu, and P. Pathak, which has demonstrated the impact of resources and individual attention on student performance.

Population density also emerged as a factor, with areas with higher populations tending to show lower scores. This could be attributed to limited resources and individual attention available to students in densely populated areas. Furthermore, racial demographics were found to be spread out across boroughs, but a correlation was observed between higher scores and White and Asian students.

In conclusion, our analysis revealed that location, neighborhood safety, income distribution, population density, and racial demographics are factors that may influence SAT scores in New York City schools.

IV. Regression Results

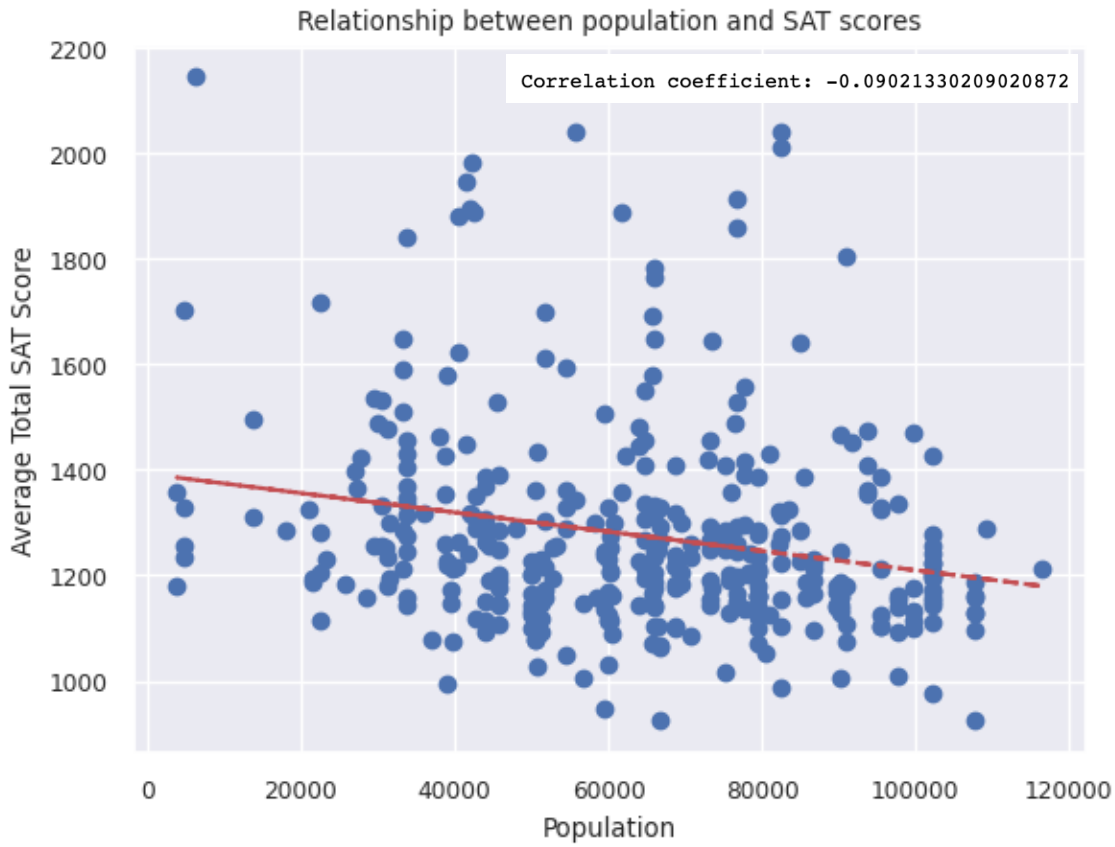
4.1 Description of Regression Models Used

To better understand the factors influencing SAT scores, we have selected four of our variables for regression analysis: boroughs, racial demographics, population, and median income. The identified factors were chosen based on their potential relevance to SAT scores in the context of our study. We will create regression graphs, models, and tables to interpret hypothesized relationships.

Our analysis suggests a possible linear relationship between SAT scores and these X variables. The regression graphs below (Figure 12 and 13) reveal a linear trend between SAT scores and median income, as well as between SAT scores and population. However, the relationship with median income is slightly heteroskedastic, so we will take the log transformation of that X variable to eliminate this effect. Population also shows a scatter around the OLS line, so we may consider logging that variable as well.

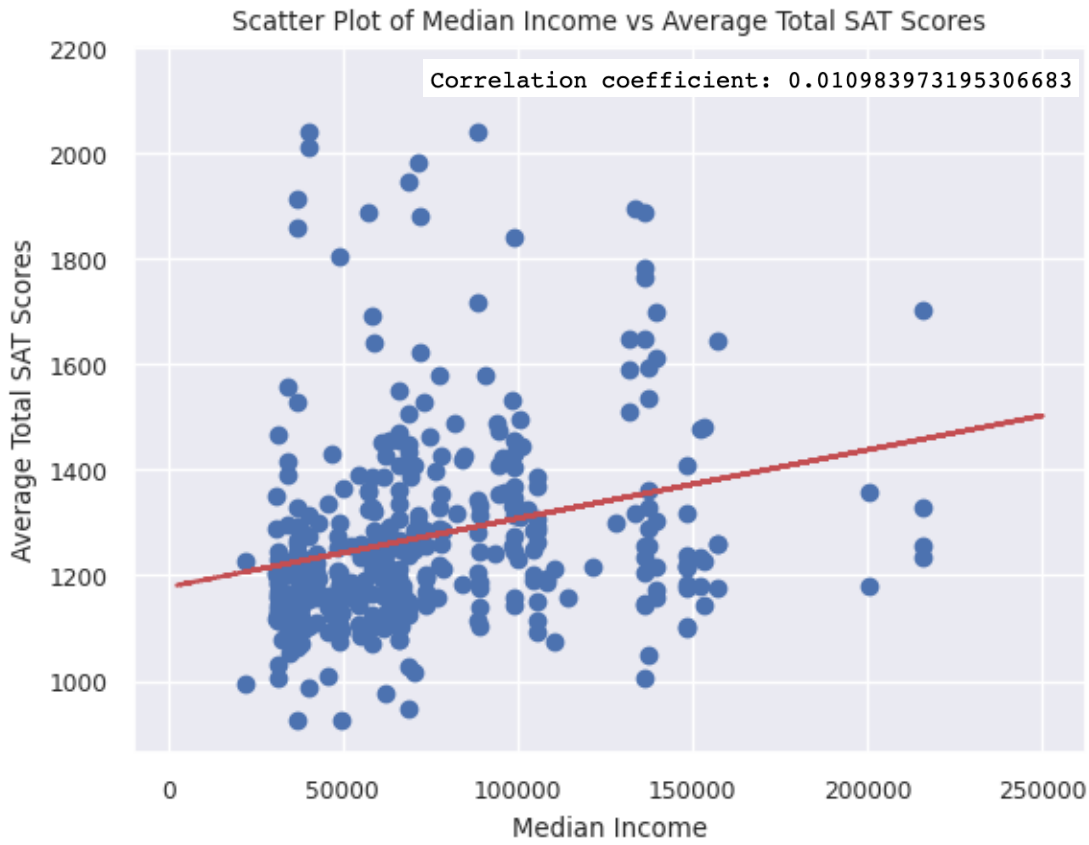
While we acknowledge that potential non-linearities and endogeneity issues could complicate our analysis, the theory supports a linear relationship between SAT scores and our X variables.

Figure 12: Exploring the Relationship Between Population and SAT Scores



This scatterplot displays a regression line that exhibits a slight negative slope from the Y axis at an intercept of around 1400, indicating a weak negative relationship between population and SAT scores. As population increases, average total SAT scores tend to decrease slightly. The correlation coefficient between population and SAT scores is also weakly negative, with a value of -0.09021330209020872. This correlation does not imply causation. However, it is interesting to observe that the areas with lower average SAT scores tend to have higher populations, as seen in Figure 9.

Figure 13: Exploring the Relationship Between Income and SAT Scores



Based on the scatter plot graph and the correlation coefficient, we can see that there is a weak positive relationship between average SAT scores and median income. Income is a weak predictor of average SAT scores in this dataset. Our previous analysis has identified other predictors, such as location and race, which have a greater significance on SAT scores.

4.2 Presentation of Regression Results

Figure 14: SAT scores, Population Logged, and Median Income Logged

	Dependent variable:Average Total SAT Score		
	Model 1	Model 2	Model 3
	(1)	(2)	(3)
const	2154.656***	96.605	796.747*
	(205.449)	(219.492)	(414.471)
ln(Median Income)		106.174***	84.436***
		(19.751)	(22.507)
ln(Population)	-80.187***		-41.841**
	(18.714)		(21.043)
Observations	374	374	374
R^2	0.047	0.072	0.082
Adjusted R^2	0.044	0.070	0.077
Residual Std. Error	190.484 (df=372)	187.964 (df=372)	187.222 (df=371)
F Statistic	18.360*** (df=1; 372)	28.897*** (df=1; 372)	16.540*** (df=2; 371)
Note:			*p<0.1; **p<0.05; ***p<0.01

$$\widehat{satscores}_i = 796.747 + 84.436\ income_i - 41.841\ population_i + \varepsilon$$

In this figure, we regress population and income individually on SAT scores before combining them together in Model 3. Although the R^2 increases slightly, it remains quite low. Overall, we gather that the variation in our independent variables does not account for much variation in our dependent variable. However, our modeling is still useful for understanding relationships between our variables. For Model 3, the predicted Average Total SAT Score when Median Income and Population are allowed to vary is 796.747 points. Furthermore, when controlling for population, an increase of 10% in one's income is associated with an 8.4-point increase in SAT scores. However, when controlling for income, a 10% increase in a zip codes population is associated with about a 4.2-point decrease in SAT scores. It confirms

what we have previously observed in our visualizations; increases in income and decreases in population are associated with higher SAT scores. We cannot infer causation, but a relationship does exist. Moreover, in line with the previous studies we mentioned earlier, there are theories that support the idea that areas with more resources and personal specialization for students tend to have better outcomes.

Figure 15: SAT Scores, Racial Demographics, and Median Income

	Dependent variable:Average Total SAT Score	
	Model 1	Model 2
	(1)	(2)
Percent Asian	17.972***	11.844***
	(0.489)	(1.627)
Percent Black	12.237***	6.278***
	(0.193)	(1.524)
Percent Hispanic	11.226***	5.547***
	(0.174)	(1.451)
Percent White	18.902***	12.455***
	(0.539)	(1.719)
ln(Median Income)		52.183***
		(13.244)
Observations	374	374
R^2	0.989	0.990
Adjusted R^2	0.989	0.989
Residual Std. Error	135.019 (df=370)	132.445 (df=369)
F Statistic	8443.869*** (df=4; 370)	7023.394*** (df=5; 369)
Note:		*p<0.1; **p<0.05; ***p<0.01

$$\widehat{satscores}_i = 96.605 + 11.844\ Asian_i + 6.278\ Black_i + 5.547\ Hispanic_i + 12.455\ White_i + 52.183\ Income_i + \varepsilon$$

Figure 15 regresses the different racial compositions found in NYC public schools with SAT scores. In Model 2, we observe a high R^2 of 0.990. It tells us that 99% of variation in SAT scores can be explain by variation in race and income. When controlling for income, an increase of one percentage point in the proportion of Asian or White students is associated with about a 12-point increase in SAT scores. For Hispanic and Black students, the increase is around 6 points, half the magnitude.

When holding racial demographics constant, a 10% increase in income is associated with around a 5.2-poiont increase in scores. These values are highly statistically significant and moderately economically significant, indicating that both one's racial background and income play a role in determining their test scores.

Figure 16: SAT Scores, Boroughs, and Median Income

	Dependent variable:Average Total SAT Score	
	Model 1	Model 2
	(1)	(2)
Bronx	-89.334***	-67.503***
	(16.327)	(16.829)
Brooklyn	-77.167***	-69.626***
	(15.692)	(15.575)
Queens	14.975	19.779
	(19.203)	(18.979)
Staten Island	141.594***	135.210***
	(41.868)	(41.339)
const	1291.006***	679.352***
	(11.858)	(136.753)
ln(Median Income)		54.289***
		(12.094)
Observations	714	714
R^2	0.098	0.123
Adjusted R^2	0.093	0.117
Residual Std. Error	155.516 (df=709)	153.457 (df=708)
F Statistic	19.275*** (df=4; 709)	19.867*** (df=5; 708)
Note:		*p<0.1; **p<0.05; ***p<0.01

$$\widehat{satscores_i} = 679.352 - 67.503\ Bronx_i - 69.626\ Brooklyn_i + 19.779\ Queens_i + 135.210\ Staten\ Island_i + 54.289\ Income_i + \varepsilon$$

Figure 16 displays regression results for borough data and income data on SAT scores. Dummy variables were created for Bronx, Brooklyn, Queens, and Staten Island as independent variables, with Manhattan serving as the reference category.

For Model 1, the intercept reveals that the omitted variable, Manhattan, has an average SAT score of approximately 1300 points. Bronx and Brooklyn have negative coefficients, indicating that, on average, students in these boroughs score between 77 and 89 points lower than students in Manhattan. Queens, on the other hand, shows similar scores to Manhattan, while Staten Island has scores approximately 141 points higher. However, the model's explanatory power is limited, as only a small amount of variation in SAT scores is explained by the X variables, as evidenced by the low R^2 value of 0.098.

In Model 2, when income is added as a controlling variable, the R^2 value increases to 0.123, though it remains relatively low. When controlling for all the different boroughs, a 10% increase in income is associated with a 5.4-point increase in SAT scores. Furthermore, when controlling for income, students in Bronx and Brooklyn still score lower than students in Manhattan, but the difference is reduced to around 70 points, compared to the previous difference of almost 90 points. All of these findings are statistically significant, except for Queens, indicating that the borough in which a student lives, and their income background appear to influence testing. Moreover, these results are highly economically significant, with a discrepancy of over 200 points in scores between Brooklyn and Staten Island.

Figure 17: SAT Scores, Boroughs, Racial Demographics, Median Income, Population

	Dependent variable:Average Total SAT Score		
	Model 1	Model 2	Model 3
	(1)	(2)	(3)
Bronx		-67.503***	-19.639***
		(16.829)	(6.172)
Brooklyn		-69.626***	-37.322***
		(15.575)	(5.746)
Percent Asian			10.428***
			(0.671)
Percent Black			7.096***
			(0.640)
Percent Hispanic			6.792***
			(0.622)
Percent White			13.097***
			(0.712)
Queens		19.779	-8.881
		(18.979)	(8.067)
Staten Island		135.210***	-5.160
		(41.339)	(16.000)
const	96.605	679.352***	
	(219.492)	(136.753)	
ln(Median Income)	106.174***	54.289***	40.317***
	(19.751)	(12.094)	(4.083)
ln(Population)			4.695
			(3.597)
Observations	374	714	1,942
R^2	0.072	0.123	0.994
Adjusted R^2	0.070	0.117	0.994
Residual Std. Error	187.964 (df=372)	153.457 (df=708)	93.417 (df=1932)
F Statistic	28.897*** (df=1; 372)	19.867*** (df=5; 708)	33490.666*** (df=10; 1932)
Note:			*p<0.1; **p<0.05; ***p<0.01

$$\widehat{satscores_i} = 1275 - 19.639\ Bronx_i - 37.322 Brooklyn_i - 8.881\ Queens_i - 5.160\ Staten\ Island_i +$$
$$+10.428\ Asian_i + 7.096\ Black_i\ +\ 6.792\ Hispanic_i +\ 13.097\ White_i + 40.317\ Income_i +$$
$$4.695\ Population +\ \varepsilon$$

In this final figure, we include all previous independent variables together. Let us focus mainly on Model 3. In this case, the R^2 is extremely high, meaning 94% of variation in SAT scores is explained by all our independent variables.

First, we notice for borough variables, when holding all else constant, a student in the Bronx scores between 20 and 38 points lower than a student in Manhattan. These results are significant at a 1% level. Queens and Staten Island has positive coefficients in Model 1, but in Model 3, these coefficients become insignificant. This indicates that the effect of these boroughs on SAT scores are not statistically significant after accounting for other variables in the model.

Regarding racial demographics, when controlling for income, location, and population, there is still a positive relationship where SAT scores are higher for students from White and Asian backgrounds. For example, a one-unit increase in the percentage of Asian students in a school is associated with about a 10-point increase in SAT scores.

Median income shows similar effects to what we had seen before. When controlling for all else, a 10% increase in income is associated with about a 4-point increase in SAT scores.

The coefficient for population is not statistically significant. Therefore, when considering all other variables, population is not a significant indicator for scores in our model.

V. Conclusion

Our study analyzed the relationship between SAT scores and school demographics in New York City public schools. We focused mainly on the borough data, racial demographics, population data, and median income data as key variables. Our findings provide insights into the factors that may influence standardized testing scores, such as SAT scores, and shed light on potential ways to improve educational outcomes.

Our analysis revealed that the average SAT scores were primarily centered around 400 points, per subject, with some schools performing better than others. We also found a correlation between the racial composition of the schools and the students' performance on the SAT, with white students tending to score higher than students of an ethnic minority.

Our analysis of the correlation heatmap showed that the students' racial background played a significant role in their SAT scores. Specifically, we observed that Percent White and Percent Asian were positively correlated with higher SAT scores, while Percent Hispanic and Percent Black were negatively correlated with lower SAT scores.

Additionally, our examination of different maps led us to conclude that income and location were also significant factors in students' performance on the SAT. The scatter plot showed a weak positive relationship between median income and average SAT scores, indicating that income is a weak predictor of SAT scores in this dataset. However, our map analysis demonstrated that schools located in poorer boroughs tend to perform worse on the SAT than those in more affluent areas, suggesting that location may be a more significant factor in students' performance.

We build on existing research, such as studies that have examined the relationship between educational achievement and various factors. For instance, we have referred several times to Rocco d'Este and Elias Einiö's paper on Asian segregation and scholastic outcomes. Our paper focuses more on effects

to high school students and all different factors that could influence their performances. We support the findings of Rocco and Elias' paper by finding positive correlations between racial composition and SAT scores, as schools having higher proportion of Asian and White students generally score higher than schools with mainly Hispanic and Black students.

Additionally, our analysis suggests relationships between SAT scores and median income and population, with higher incomes and smaller populations associated with higher scores. This aligns with previous research that has shown correlation between income, resources, and outcomes. However, we do acknowledge that potential non-linearities and endogeneity issues could complicate our analysis and further research could be done explore these relationships more in depth.

In future research, it would be interesting to investigate variables linked to our independent variables. What leads to different income levels? What leads to different boroughs having different infrastructure and development? Other potential factors such as school resources, class sizes and teaching quality could provide a more comprehensive understanding of underlying mechanisms driving our findings. It would also be interesting to use more qualitative research methods as opposed to observed data, such as interview and focus groups, to provide deeper insights and perspectives.

In conclusion, our study contributes to the growing body of literature on the relationship between SAT scores and school demographics, providing important insights into the factors that may influence educational outcomes in New York City public schools. Our findings suggest that addressing disparities in racial demographics and income levels may be key in creating a more equitable educational system. Further research is needed to better understand the underlying mechanisms driving these relationships and to inform evidence-based policies and interventions aimed at improving educational opportunities for all students.

Appendix

VI. Machine Learning

A regression tree was also created in this study, but we did not want it to be a main part of the paper, so we are including it in the appendix.

6. 1 Description of the Objective Function

$$\widehat{satscores_i} = \beta 0 + \beta 1 \text{race} + \beta 2 \text{income} + \varepsilon$$

$$\frac{1}{N}\sum_{i=1}^{N}(satscores_i) - (\beta 0 + \beta 1\, race_i + \beta 2 \log(income_i))^2$$

Our objective function aims to minimize the mean squared error between SAT scores and two selected X variables: racial demographics and income. These variables were chosen based on our observations from regression trees, where models with racial demographics as independent variables showed higher R squared values and lower standard errors. Additionally, incorporating the income variable in different models improved their accuracy. Therefore, including these two variables for a regression tree seems like the right fit.

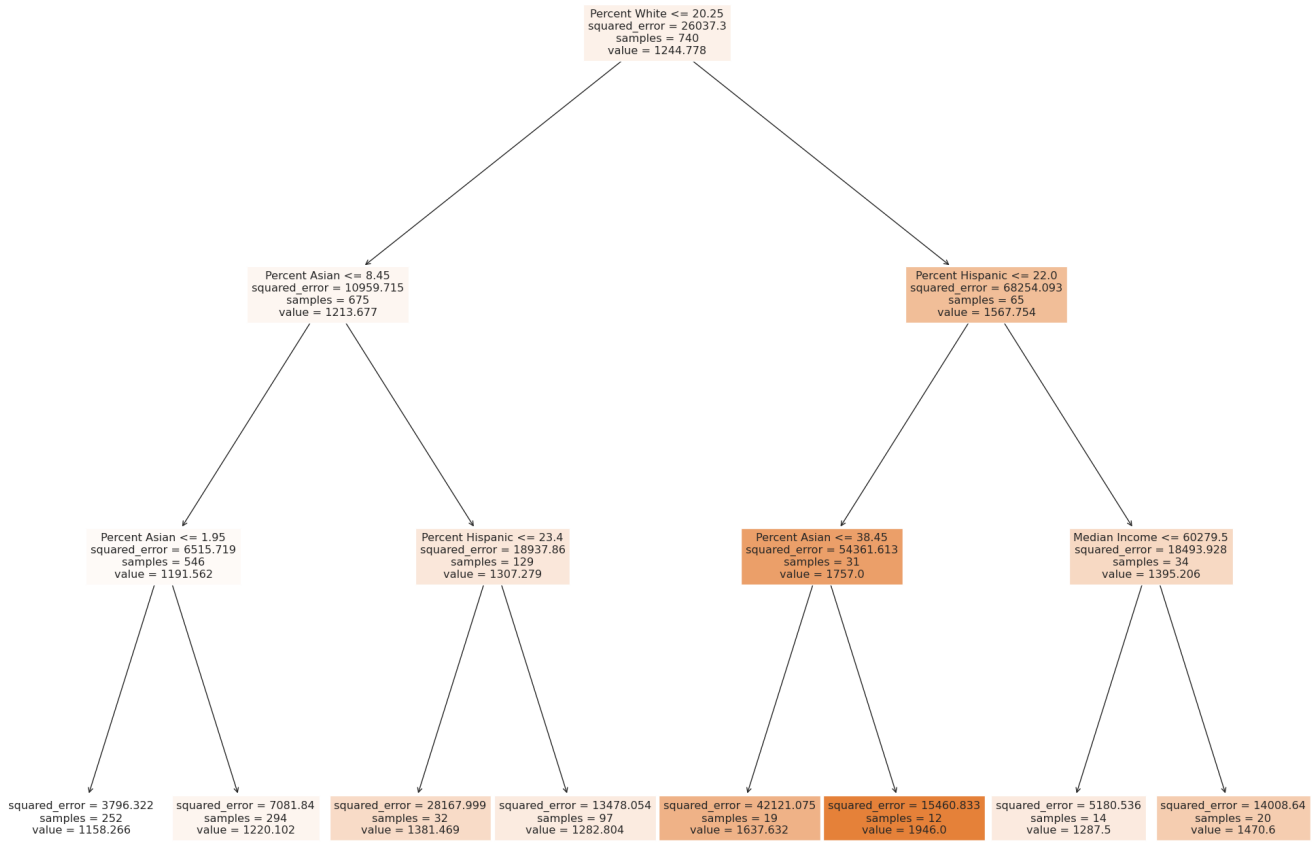
To regulate the error in our model, we can use various techniques and parameters to reduce it. The first parameter we consider is the maximum tree depth, which limits the number of levels in the tree. A shallower tree with less depth can help prevent overfitting and excessive complexity. In our regression, we choose a depth of 3 levels, as 2 may not capture enough information and 4 can result in an overcrowded and illegible tree.

The second parameter we implement is the minimum sample leaf size, which specifies the minimum number of samples required for a node to be considered a leaf. A larger sample size can result

in a simpler tree with more generalized data, while a smaller sample size may lead to a more complex tree. In our case, we choose a minimum sample leaf size of 5.

Lastly, we also set a specific alpha value, typically ranging from 0 to 1. In our case, we set $\alpha = 0.8$. A higher alpha indicates stronger regularization, resulting in more aggressive pruning of values that do not fit well in the context. If we were to decrease the alpha, our tree would have more branches and leaves.

Figure 18: Regression Tree Analysis: SAT Scores, Racial Demographics, Median Income



6.2 Regression Interpretation

The first node of the tree is based on the independent variable "Percent White." For a sample size of 380 schools, when splitting the data based on the proportion of white students in a school being less than or equal to 20.25%, the predicted score that satisfies this condition is about 1244 points. Moving on to the second level of the tree, if a school has a population with less than 20% white students and less than 8.5% Asian students, the predicted scores are around 1225 points. For the final nodes on one split of the tree, with a smaller sample size of 254 schools, when a school has less than 2.45% Asian students, the predicted scores tend to be around 1193 points.

In the other split from the main node, when schools have less than 22% Hispanic students, which means they are comprised of other races besides Hispanic, the predicted scores tend to be much higher, around 1590 points.

When considering the addition of income to the tree, if the median income of the school's neighborhood is less than $60,000 the predicted scores trend towards 1287 points. However, if the median income is higher than $60,000, the predicted scores trend towards 1470 points, indicating a significant economic difference.

6.3 A Comparison of Regression and Regression Tree Results

Comparing to our regression results, the regression tree provides a more visual representation of our findings. While the regression table offers more information such as t-statistics and p-values, the regression tree summarizes all the information in a concise manner. Furthermore, the tree can capture non-linear relationships among variables, which can be challenging to do directly with regression tables. For instance, in our case, we observed that taking the logarithm of income and population resulted in more significant findings, and instrumental variables could have also been used. However, the tree instantly captures these non-linear relationships.

The tree allows us to directly visualize how schools with lower percentages of White and Asian students tend to have lower scores, compared to schools with lower percentages of Hispanic students, which tend to have higher scores. Additionally, the relationship between income and SAT scores is also captured. Also, regression trees are capable of capturing interaction effects, such as the effect of racial demographics on SAT scores depending on one's income level. These effects may not be captured in a regression table.

The predictive aspect of the tree is particularly interesting. Starting with large samples of schools, the tree progressively narrows down the samples at each level, leading to more specific inferences, while considering the measure of error.

Let's delve into precision. In our regression table titled "SAT scores, Racial Demographics, and Median Income," we observe different coefficient estimates for each racial category. However, what we cannot determine is if some effects dominate others. Are the effects of each individual race equally significant on SAT scores? Students of Black and Hispanic backgrounds experience a lower increase in their scores compared to students of White and Asian backgrounds, but which group is affected more by their ethnicity?

Upon examining the tree, we notice that the variable "Percent Black" does not appear in any of the nodes, which could indicate that its effects on SAT scores are not as strong as the other variables. On the other hand, "Percent Asian" and "Percent Hispanic" are used multiple times for splitting in the tree, suggesting that they are considered important for partitioning the data and capturing different patterns. Similarly, "Percent White" is an important predictor, as it is the first variable used to partition the data and may have a greater impact on the overall performance of the model.

References:

Census.gov. (n.d.). United States Census Bureau. Retrieved from https://www.census.gov/

Average SAT Scores for NYC Public Schools. (n.d.). Kaggle. Retrieved from https://www.kaggle.com/datasets/nycopendata/high-schools

Top 10 Most Dangerous Neighborhoods in New York City. (n.d.). USA ESTA Online. Retrieved from https://usaestaonline.com/most-dangerous-neighborhoods-in-new-york-city

Zip Codes in New York. (n.d.). New York Demographics. Retrieved from https://www.newyorkdemographics.com/zip_codes_by_population

"The City". Poverty Rates by NYC Community District: 2015-2019. Retrieved from https://www.thecity.nyc/data/poverty-rates-by-nyc-community-district-2015-2019-09414/

d'Este, R., & Einiö, E. (Year). Asian Segregation and Scholastic Achievement: Evidence from Primary Schools in New York City. IZA Discussion Paper No. 11682. Retrieved from https://docs.iza.org/dp11682.pdf

Abdulkadiroglu, A., Hu, W., & Pathak, P. (Year). Small High Schools and Student Achievement: Lottery-Based Evidence from New York City. Journal of Applied Econometrics, 31(1), 113-137.

Graham, A. E., & Husted, T. A. (Year). Understanding state variations in SAT scores. Journal of Education Finance, 45(2), 241-259

d'Este, R., & Einiö, E. (Year). Asian Segregation and Scholastic Achievement: Evidence from Primary Schools in New York City. Education Finance and Policy, 15(4), 511-533. Retrieved from https://www.nber.org/system/files/working_papers/w19576/w19576.pdf