Beyond Curriculums and Teachers: Examining Objective Factors Influencing Academic Performance

By Feifan Liu

Selecting the right school for their children is a crucial decision for parents, as it can have a significant impact on their educational outcomes. In this study, we examined objective factors such as the average environment, racial makeup, living standards, and population density to understand their impact on academic performance. Our analysis of 1270 schools in New York City revealed that a supportive environment, strong family-community ties, effective school leadership, and trust are key components of a positive school environment that is associated with better ELA proficiency. However, race was found to have no direct impact on academic performance, but environmental factors such as living conditions may explain the differences in performance between different racial groups. Additionally, we found that students from areas with higher population densities tend to perform better in school. Our findings can guide government subsidies for schools to improve educational outcomes.

1. Introduction

In today's world, academic success is highly valued, and studying is one of the many paths to achieving it. However, the factors that influence academic performance in students are still unclear, making it difficult for parents to improve their children's academic achievements. The aim of this project is to investigate the factors that impact students' academic performance beyond just curriculums and teachers. Besides, with so many schools available, selecting a suitable one for their children is a crucial decision for parents. Therefore, this project also seeks to provide insights into the factors that affect student performance to assist parents in making an informed decision about selecting the most appropriate school for their children.

To explore the factors that impact student performance, I will analyze a dataset containing information on 1270 schools in New York City (Data Science for Good: PASSNYC, 2018). This dataset, published by PASSNYC, a non-profit organization committed to expanding educational opportunities for underserved students, includes details on school addresses, student performance, teacher development, environment rating, racial makeup and other relevant factors.

Mainly, I will analyze two types of data: the population makeup of schools (including Percent Asian, Percent Black/Hispanic, and Percent White) and the school environment rating (including Supportive Environment, Effective School Leadership, Strong Family-Community Ties, and Trust). I will use these as independent variables and choose average ELA Proficiency as my dependent variable.

From the dataset, we can confirm that a higher environment rating (Supportive Environment, Effective School Leadership, Strong Family-Community Ties, and Trust) enhances students' performance. However, racial makeup does not affect students' performance. The performance distinction of different racial group is likely due to environmental factors rather than inherent differences in the students themselves. To further explore the impact of environment on student performance, I imported a real property dataset published by NYC Open Data to analyze the average property prices of each area (Real

Property Income and Expense Form Non-compliance List, 2018). The analysis revealed a positive relationship between property prices and student performance. Moreover, I discovered that the performance of Black/Hispanic students is limited due to their poorer family environment.

Lastly, I scraped a democratic table from a Wikipedia page(Wikipedia contributors, 2023) and imported a democratic dataset from NYC Open Data (Demographic Statistics by Zip Code, 2022) to investigate the relationship between population density and student performance. The analysis revealed that students living in higher density areas tend to perform better in school.

To further investigate the effect of the independent variables on the dependent variable (average ELA proficiency), I utilized both OLS and IV-2SLS regression models to analyze the influence of each variable. The regression model results revealed some contradictory findings compared to our previous analysis. For instance, the increase in trust or population density might lead to a decrease in students' academic performance. Moreover, Strong Family-Community Ties was found to have no significant influence on the average ELA proficiency in the regression model. To further deepen our analysis and minimize the errors of the regression model, I employed a regression tree approach. The results of the regression tree indicate that to obtain the lowest possible errors, we may want to segment the data into groups based on the percent Black/Hispanic and average house price.

Numerous studies have examined the factors that could potentially impact academic achievement, but most have focused on subjective characteristics. For instance, Edgerton and McKechnie investigated students' perceptions of their physical school environment could affect their academic performance (2023). Similarly, Shi and Ko explored the impact of school and family psychological environments on the academic self-identity and self-efficacy of university students who major in English education(2023). Furthermore, Steinmayr et al examined the role of students' motivation in their academic achievement(2019).

My project aims to investigate the impact of objective characteristics on student performance, which are factors that parents can control when selecting a suitable school for their children. This extends the existing research that has mainly focused on subjective characteristics, such as students' perceptions and motivations. By examining objective factors such as school environment, living standards, and population density, this project seeks to provide valuable insights into how parents can make informed decisions when choosing a school for their children.

I will begin by introducing and explaining the key variables used in our analysis.

2. Background

2.1 The meaning of variables

2.1.1 The Meaning of Dependent Variable

Average ELA Proficiency: ELA, also known as English Language Arts Performance, reflects student's level of understanding by ranging students from 1 (insufficient) to 4 (more than sufficient).

2.1.2 The Meaning of Independent Variable

Each variable is separated into one of the four groups (Not Meeting Target, Meeting Target, Approaching Target, and Exceeding Target) based on their grades.

Supportive Environment: How well the school creates a culture where students feel encouraged, challenged, and safe?

Effective School Leadership: How successfully can school leadership distribute leadership to fulfil their objective while inspiring the school community with a clear educational vision?

Strong Family-Community Ties: How well the relationship is between the school and families?

Trust: Whether there is mutual respect and trust between parents, teachers, administrators, and students.

Percent Asian / (Black/Hispanic) / White: Percentage of the students of the school that is Asian / (Black/Hispanic) / White.

In summary, these independent variables are typical characteristics of the school that parents are usually paying attention to. I want to find out if parents want their children to get high grades in ELA, whether they should put effort into these aspects.

I want to analyze whether the school's environment, leadership, the relationship between family and school and race will influence the students' ELA grades.

2.2 Data

	Percent Asian	Percent Black / Hispanic	Percent White	Supportive Environment %	Effective School Leadership %	Strong Family-Community Ties %	Trust %	Average ELA Proficiency
count	1167.000	1167.000	1167.000	1167.000	1167.000	1167.000	1167.000	1167.000
mean	0.118	0.728	0.134	0.886	0.814	0.829	0.903	2.531
std	0.177	0.296	0.202	0.059	0.095	0.057	0.055	0.359
min	0.000	0.030	0.000	0.650	0.340	0.650	0.620	1.810
25%	0.010	0.480	0.010	0.840	0.760	0.790	0.870	2.250
50%	0.040	0.900	0.030	0.890	0.830	0.830	0.910	2.450
75%	0.145	0.960	0.170	0.930	0.880	0.870	0.940	2.750
max	0.950	1.000	0.920	1.000	0.990	0.990	1.000	3.930

As we can see from the table that after we clean the useless data, there are 1167 remaining schools.

The first three columns (percent Asian, Percent Black/Hispanic, and Percent White) describe the students' makeup. These data help us to find out whether students' performance is related to their race.

This is connected to our topic since I want to discuss whether race might influence students' performance. The value of each data means the proportion of the students that is Asian/Black/Hispanic/White.

The following four columns (Supportive Environment, Effective School Leadership, Strong Family-Community Ties, and Trust) describe the overall environment grades of the school. The higher the value is, the better the environment is. As we can see from the data that the grades of all the independent variables are between 0 and 1. This is reasonable since the range of each independent variable is 0 to 1.

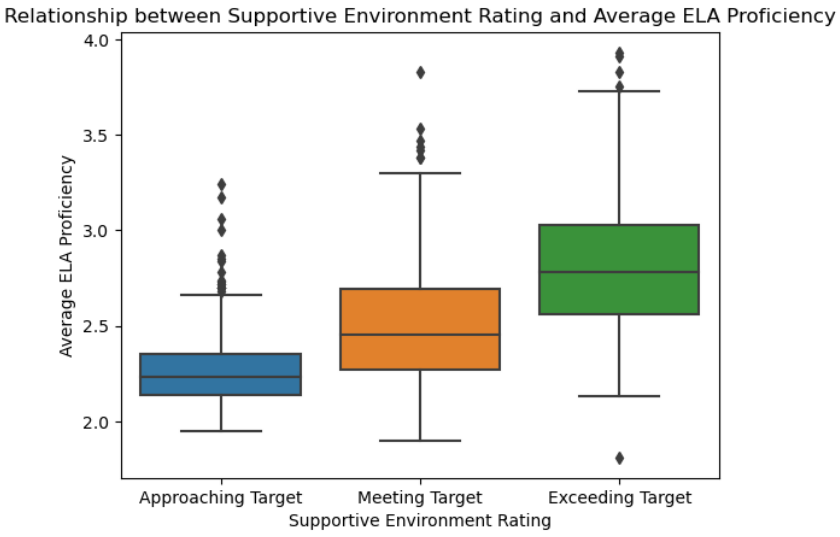
The dependent variable (Average ELA Proficiency) as the next column varies from 1.81 to 3.93. This is also reasonable since we can only get 1-4 points in the ELA exam. This variable is considered our objective grading of students' performance. The higher the score is, the better the students' performances are.

In the following part, we will check whether the increase of the variables from the 1st – the 7th column will increase the value from the 8th column. In other words, I want to check whether the value of the environment ratings and students' makeup will impose any influence on Average ELA Proficiency. Therefore, I will investigate whether the increase in the value of the environment ratings and students' makeup will increase the Average ELA Proficiency value.

3. Analysis

3.1 Box plot of Average ELA proficiency and Average Environment

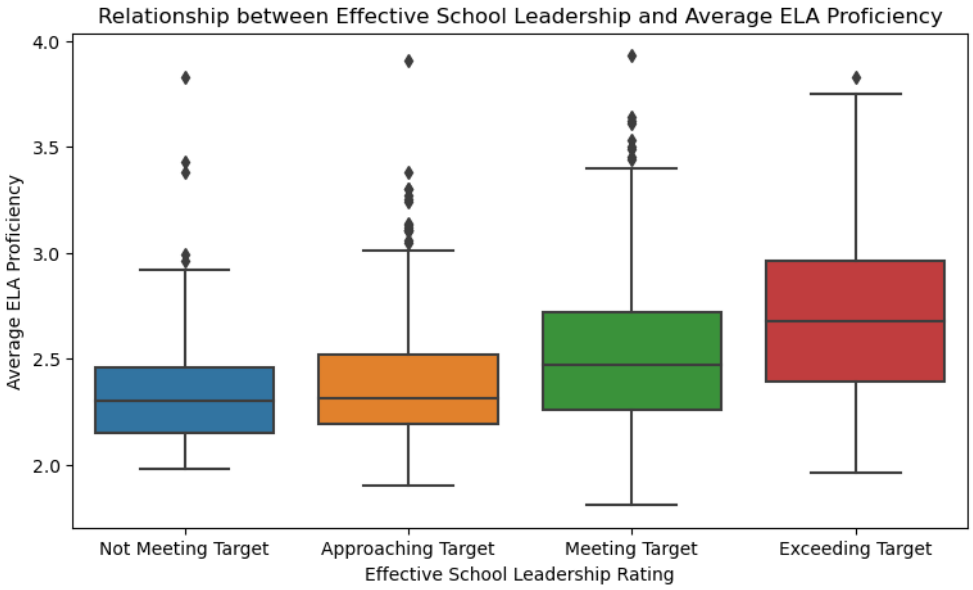
3.1.1 Box plot of Average ELA proficiency and Supportive Environment



This graph shows the relationship between supportive environment rating and Average ELA Proficiency. We can see from the graph that the mean of the Average ELA Proficiency is lowest when Supportive Environment Rating is Approaching Target. Then, the mean increases when Supportive Environment Rating is Meeting Target. While Supportive Environment Rating is Exceeding Target, we get the highest Average ELA Proficiency.

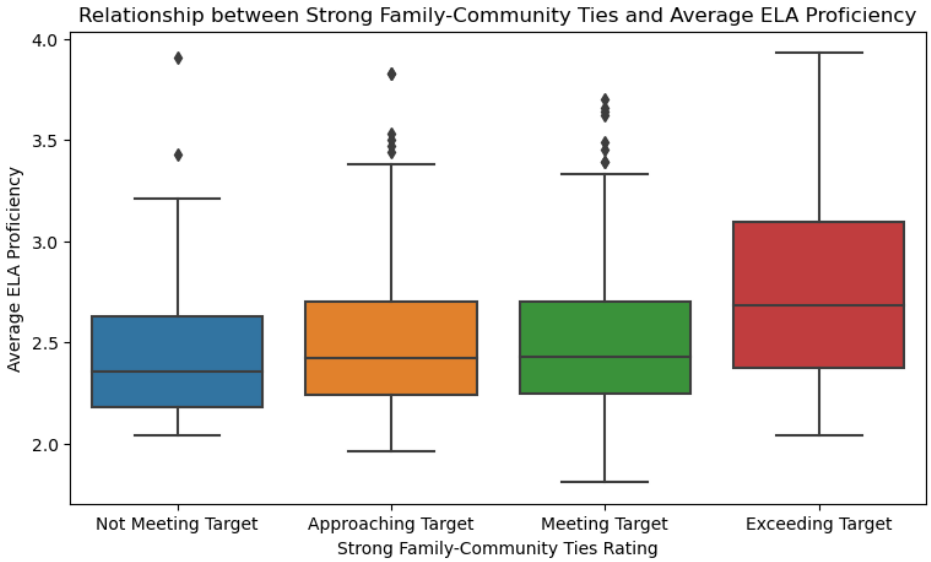
As a result, we can know that with the increase of the supportive environment rating, the average ELA Proficiency also increases.

3.1.2 Box plot of Average ELA proficiency and Effective School Leadership



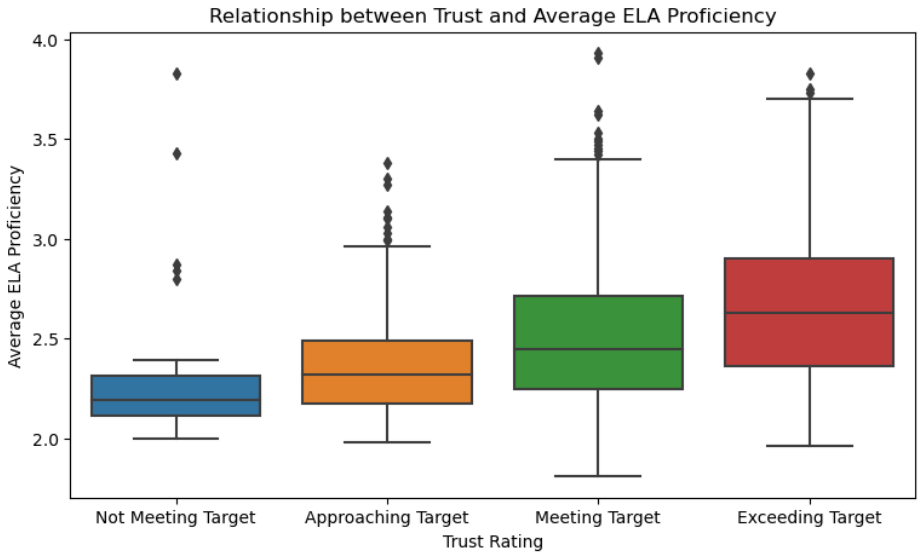
The graph displays a positive correlation between Effective School Leadership Rating and the Average ELA Proficiency. The plot demonstrates that as the Effective School Leadership Rating increases, the mean of Average ELA Proficiency also increases. However, for the groups Not Meeting Target and Approaching Target, the mean values are almost the same. Therefore, additional investigation may be necessary to explore the relationship between Effective School Leadership Rating and the Average ELA Proficiency in further detail.

3.1.3 Box plot of Average ELA proficiency and Strong Family-Community



The graph also indicates a positive association between Strong Family-Community Ties Rating and Average ELA Proficiency. The plot reveals that the group, Exceeding Target, has a significantly higher mean, while the Not Meeting Target group has the lowest mean. However, the Approaching Target and Meeting Target groups have almost the same mean. Therefore, to explore the relationship between Strong Family-Community Ties Rating and Average ELA Proficiency in further detail, a scatter plot will be used.

3.1.4 Box plot of Average ELA proficiency and Trust

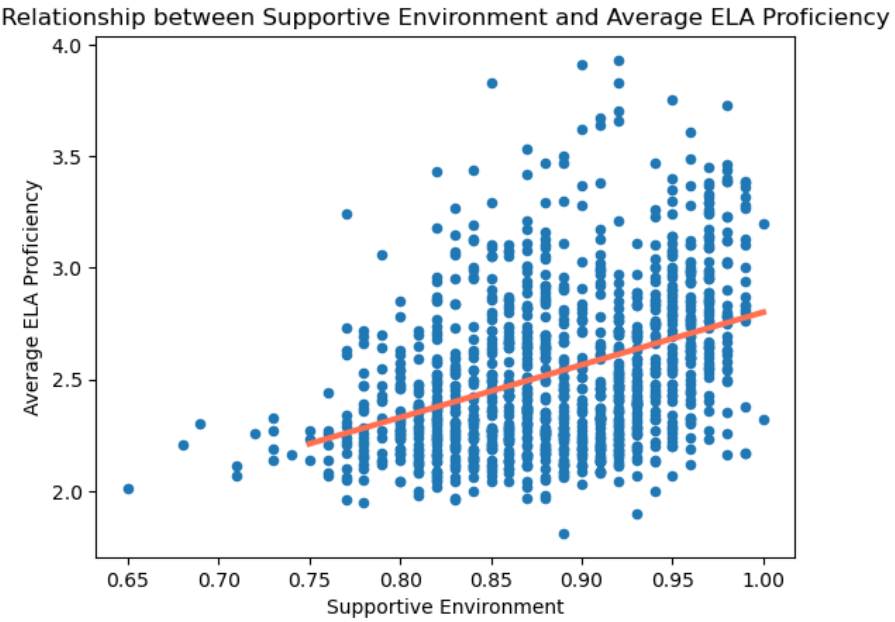


The graph demonstrates a clear rise in the mean of Average ELA Proficiency with increasing Trust Rating. The mean gradually increases from the Not Meeting Target group to the Approaching Target group, then to the Meeting Target group, and finally receives the highest mean in the Exceeding Target group. Consequently, we can conclude that there is a positive correlation between Trust Rating and Average ELA Proficiency.

All four graphs demonstrate a positive correlation between the independent variables and the dependent variable. However, the Effective School Leadership Rating graph and Strong Family-Community Ties Rating graph exhibit some instances where the mean of Average ELA Proficiency is the same for different groups. Therefore, to further investigate the relationship between the independent and dependent variables, a scatter plot will be utilized.

3.2 Scatter plot of Average ELA proficiency and Average Environment

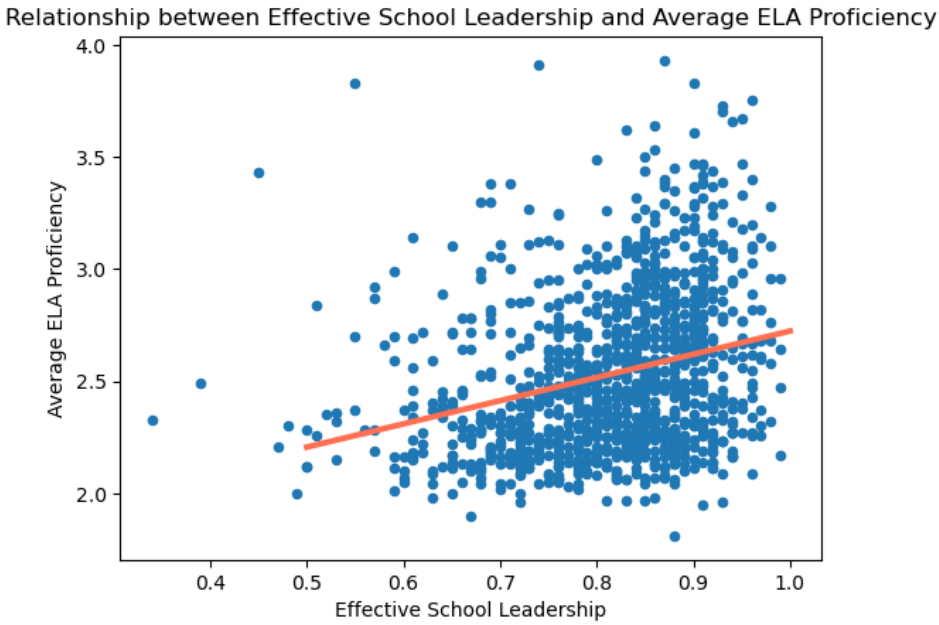
3.2.1 Scatter plot of Average ELA proficiency and Supportive Environment



This Scatter plot indicates a positive correlation between Supportive Environment and Average ELA Proficiency. Moreover, as both the scatter plot and the box plot indicate the correlation, we can be assured that there is a high probability that Supportive Environment does affect Average ELA

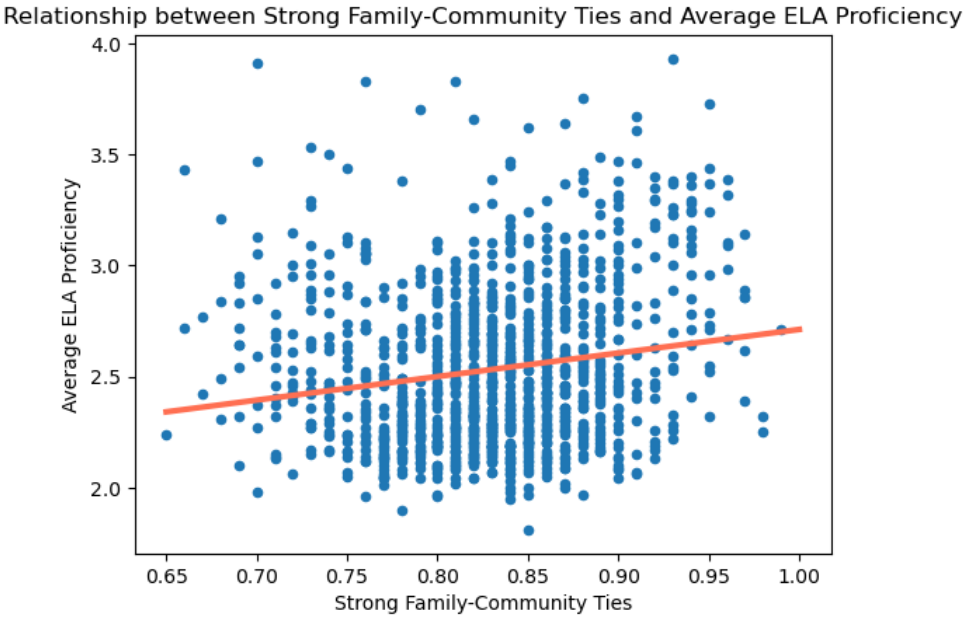
Proficiency. With the increase of the grade of Supportive Environment, Average ELA Proficiency also increases.

3.2.2 Scatter plot of Average ELA proficiency and Effective School Leadership



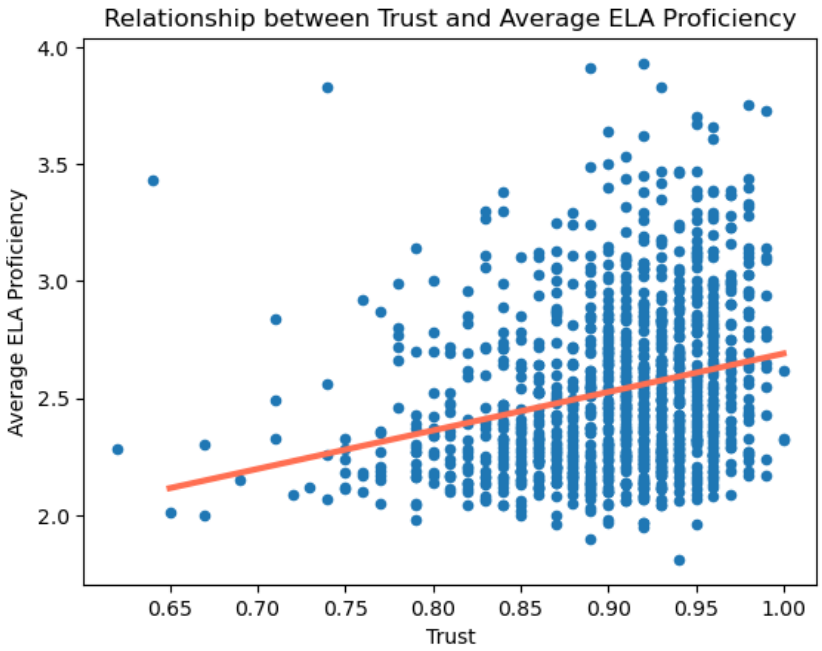
From this graph, it can be observed that the Average ELA Proficiency increases with an increase in the grade of Effective School Leadership. Furthermore, the boxplot confirms a positive relationship between these variables. Hence, it can be inferred that there is a high likelihood of a positive correlation between Effective School Leadership and Average ELA Proficiency.

3.2.3 Scatter plot of Average ELA proficiency and Strong Family-Community



This graph also illustrates a positive correlation between Strong Family-Community Ties and Average ELA Proficiency. The boxplot of Strong Family-Community Ties also shows a positive indication. Combining both graphs, it can be concluded that there is a high likelihood of a positive correlation between Strong Family-Community Ties and Average ELA Proficiency.

3.2.4 Scatter plot of Average ELA proficiency and Trust

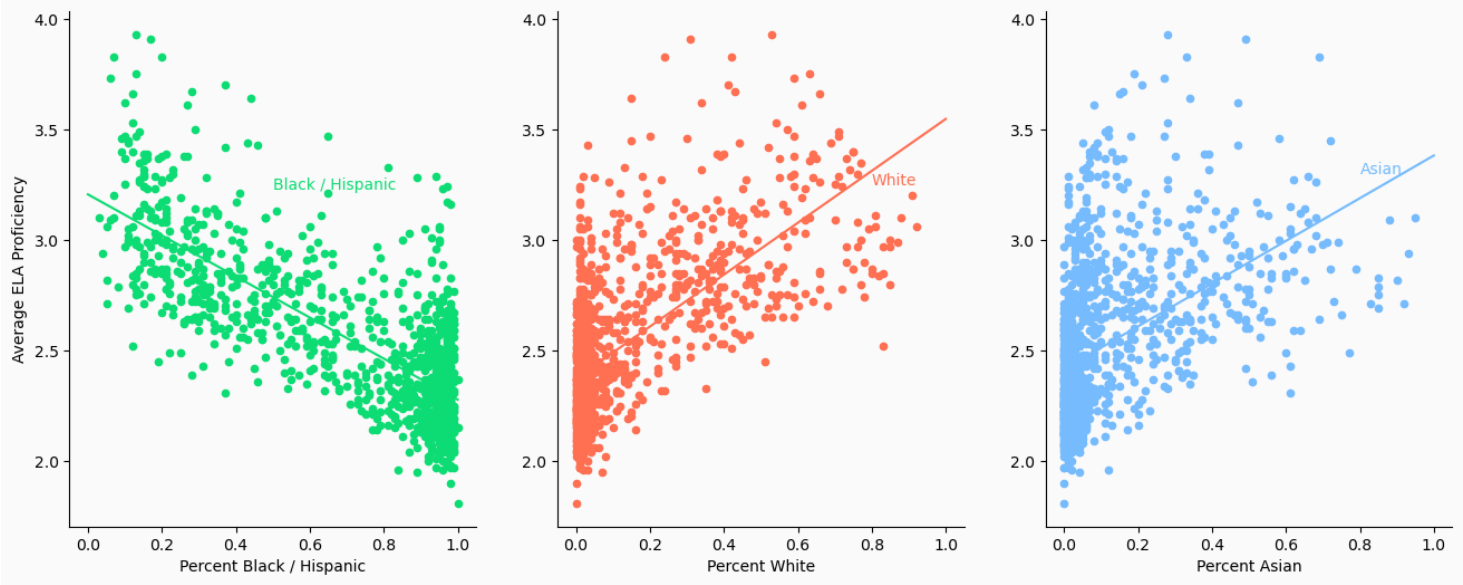


It can be observed that there is an incline in Average ELA Proficiency with an increase in the grade of Effective School Leadership. Moreover, both the boxplot and scatter plot demonstrate a positive correlation, indicating that there is a high likelihood of Trust having a positive influence on student performance (Average ELA Proficiency).

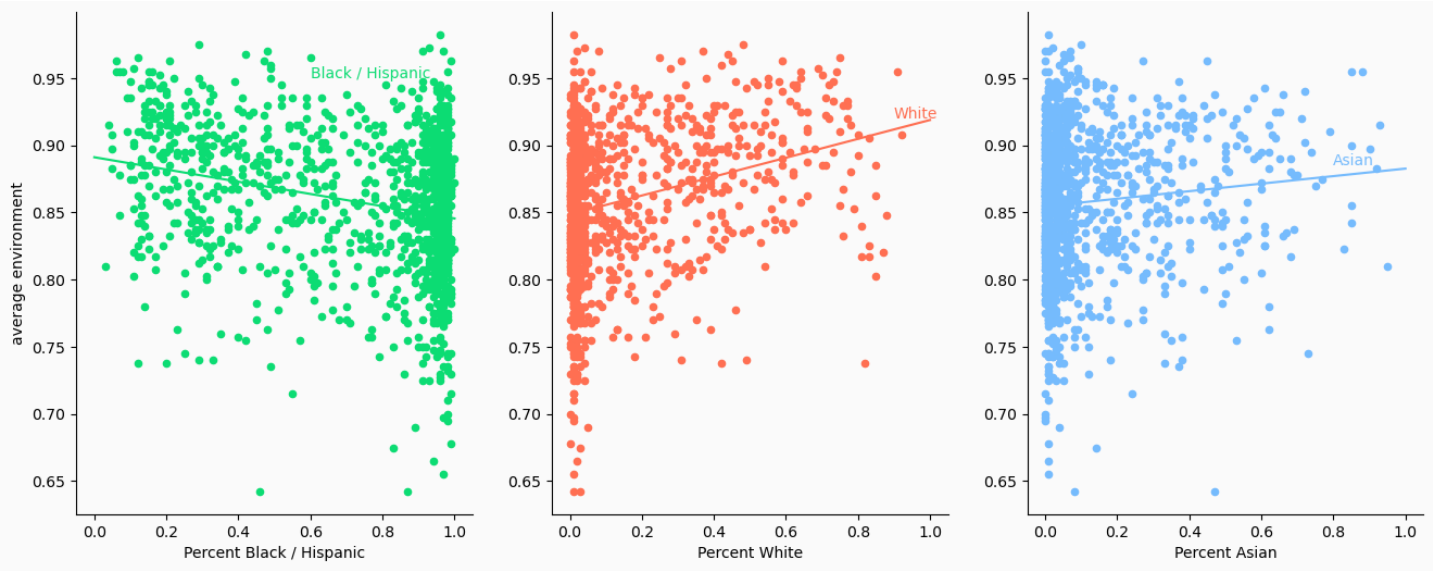
3.3 Racial makeup's influence

In this project, I want to discover what might impact students' academic performance except for curriculums and teachers. I want to further discover whether the racial makeup will influence student's performance. Maybe students from different groups will have distinct grades on ELA Proficiency. I am not saying that there is an intelligence difference between races. Even people with the same intelligence could perform differently according to their experiences. I am trying to find out what might experiences

caused by their race influence their performance. The same race might result in the same experiences. For example, it is common for Asian parents to have a high expectation of their children's grades while neglecting their extracurricular activities. Thus, Asian students might have the same experience as being forced to stay at home studying instead of staying outdoors to do some exercises. Therefore, I am using race to represent students' home environment.



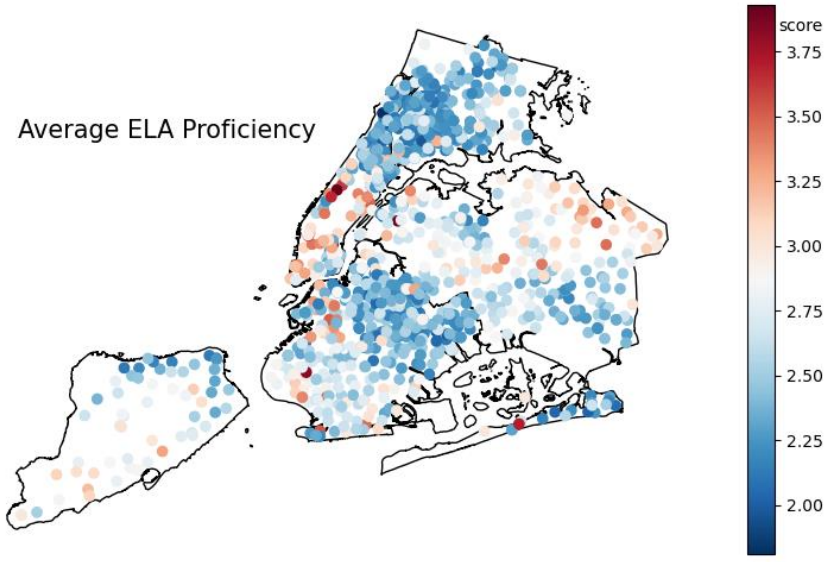
We can see from the graph that as the percentage of the Black/Hispanic population increases, students' average ELA Proficiency decreases. In contrast, as the percentage of Whites or Asians increases, students' performance increases. This shows that Whites and Asians usually perform better than Blacks/Hispanics. However, the problem is that we do not know if this performance distinction is caused by the students themselves or the students' environment. For example, a school that has a higher Black/Hispanic percentage might have a lower grade in the school's environment grades. As a result, I decide to make another graph to illustrate the relationship between average environment grades (the mean of Supportive Environment, Effective School Leadership, Strong Family-Community Ties, and Trust) and the percentage of racial makeup.



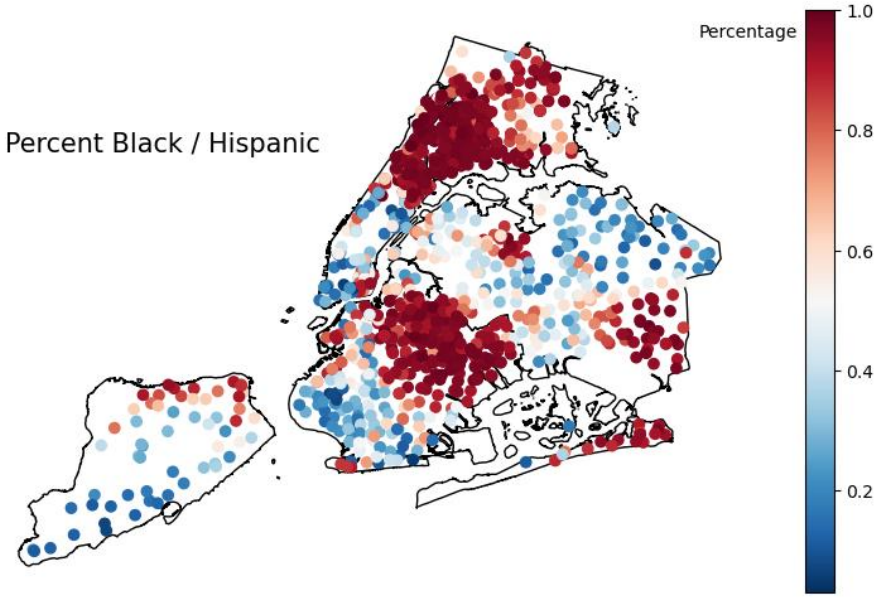
From these 3 graphs, we can see that the value of average environment is higher when the percentage of Black/Hispanic is lower or the percentage of White or Asian is higher. In other words, Whites and Asian usually have a better environment, when Black/Hispanic have a poorer environment. Consequently, the performance distinction of different races might cause by the environmental differences. Thus, we can not have any conclusion on whether the racial makeup will influence students' performance.

3.4 Maps

In order to further investigate the relationship between racial makeup and students' performance, I decide to draw maps and compare the difference between maps. Thus, firstly, I import the map of New York (Borough Boundaries | NYC Open Data, n.d.) and draw the Average ELA Proficiency map.



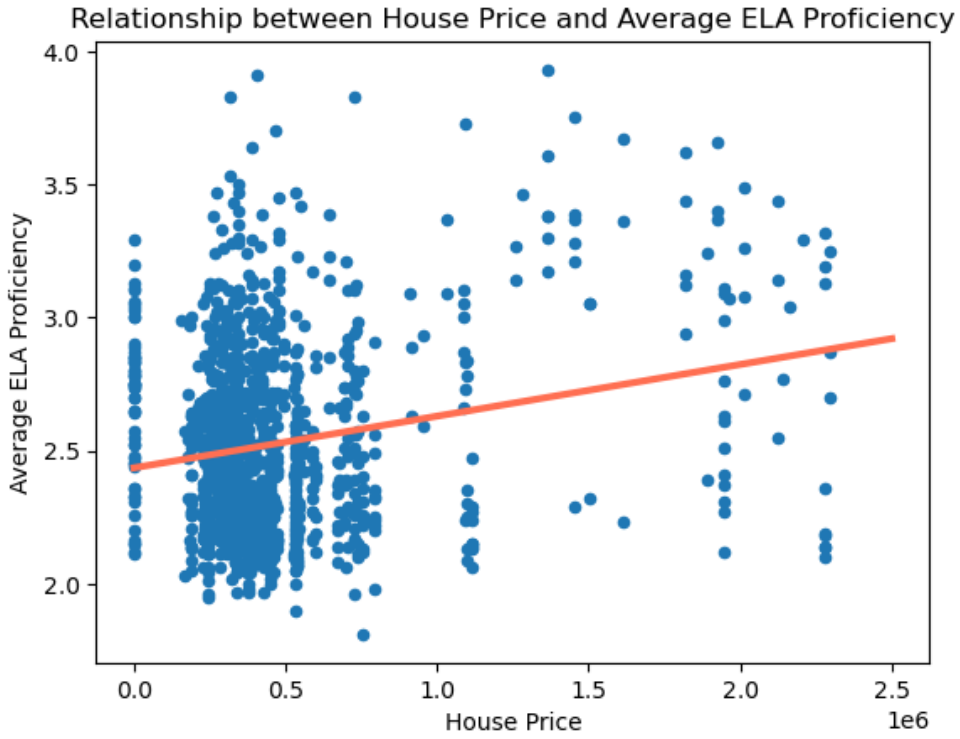
This is the plot of New York, where our data are located. Each point on the graph represents a school in the dataset. The color of the point is determined by the value of average ELA proficiency for each school. The higher the value is, the redder the point is. In other words, a red point means this school has a high average ELA proficiency, while a blue point means this school has a low average ELA proficiency. We can see from the plot that blue points are more concentrated than red points. Most of the blue points are in one of the three parts of New York, while red points are more scattered on the map. In other words, if schools are concentrated in one area, they will tend to have lower grades.



The percentage of the Black / Hispanics determines the color of the points. Same as the map above, the higher the percentage is, the redder the point is. Thus, a school that has many Black / Hispanic students will appear as a red point on the map.

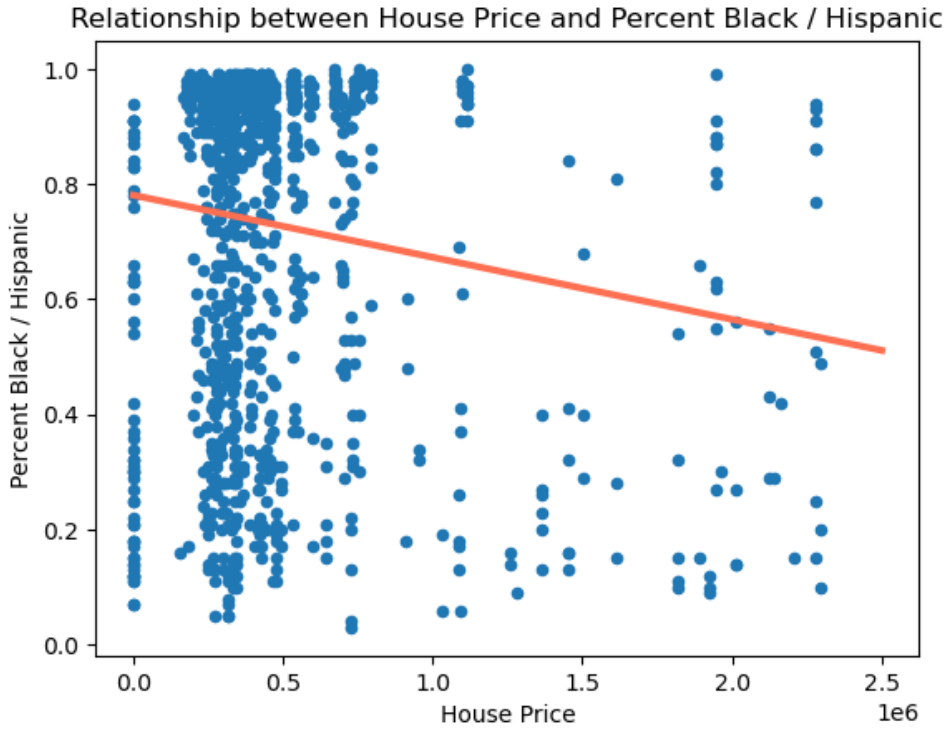
From this graph, we can see that most schools that have a larger fraction of Black / Hispanic students tend to concentrate, while schools with a lower fraction of Black / Hispanic students are scattered. This is similar to the map of average ELA proficiency, where schools with lower average ELA proficiency are concentrated. This also matches the result we get above that a higher fraction of Black / Hispanic students might lead to a lower average ELA proficiency.

Nevertheless, there might be other reasons that cause the correlation between the average ELA proficiency and the fraction of Black / Hispanics. For example, Black / Hispanics are poorer, so they can only afford schools with low average environmental grades, which might further lead to a lower average ELA proficiency. Thus, for the next part, I want to check whether there is a relationship between the money owned by the residents and the fraction of Black / Hispanics. As property owned by the residents can reflect their living standard. Consequently, I will introduce the data on the housing price in New York and calculate the average property price of each zipcode area (Real Property Income and Expense Form Non-compliance List, 2018).



We can see from the graph that the red line is upward sloping. In other words, there is a positive relationship between the House price and Average ELA Proficiency. As we are using the average house price to represent the money that holds by the residents, we can say that the students from a richer family might be able to preform better in school.

In addition to examining whether Black students perform worse due to their family's income, we also want to explore the relationship between the percentage of Black/Hispanic residents and average house prices. Specifically, we want to see how the average house price varies in areas with higher percentages of Black/Hispanic residents.



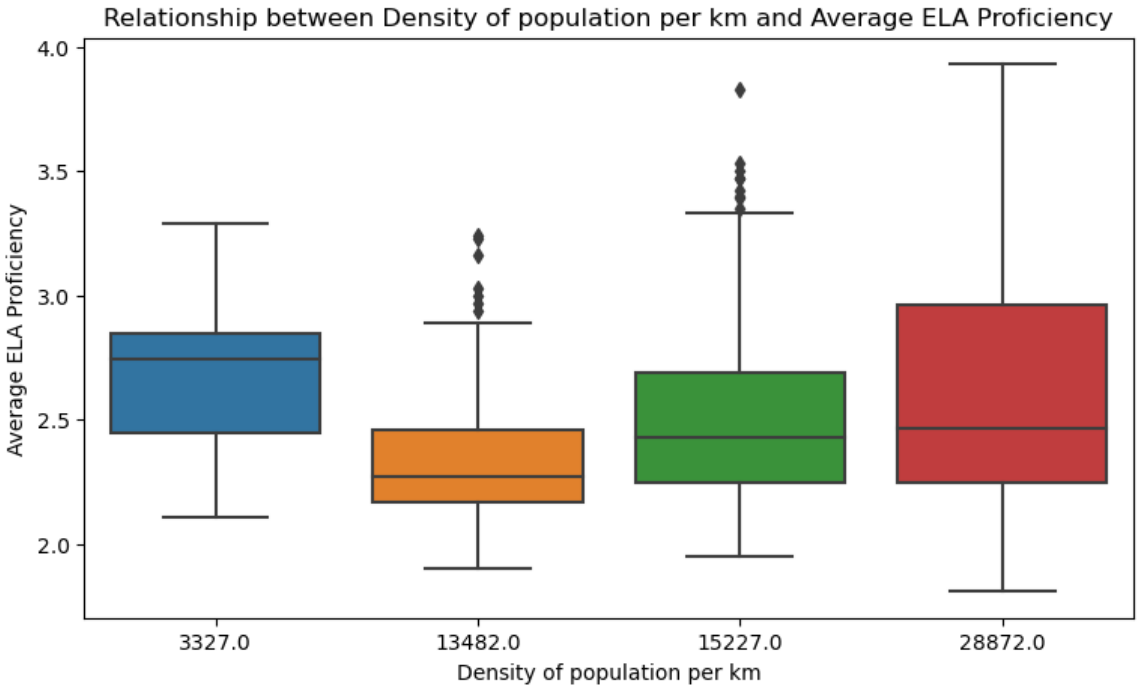
From the graph, we can observe that there is an inverse relationship between house price and the percentage of Black/Hispanic residents. As house prices increase, the percentage of Black/Hispanic residents decreases. This suggests that Black/Hispanic individuals are more likely to come from lower-income families, which may in turn influence their academic performance. This finding is also consistent with the earlier plot showing that schools with higher percentages of Black/Hispanic students tend to have poorer environments. It is possible that this is due to the fact that Black/Hispanic students who come from lower-income families cannot afford to attend more expensive schools with better environments.

3.5 Influence of Population Density

In this next phase, I will investigate whether population density plays a role in student success.

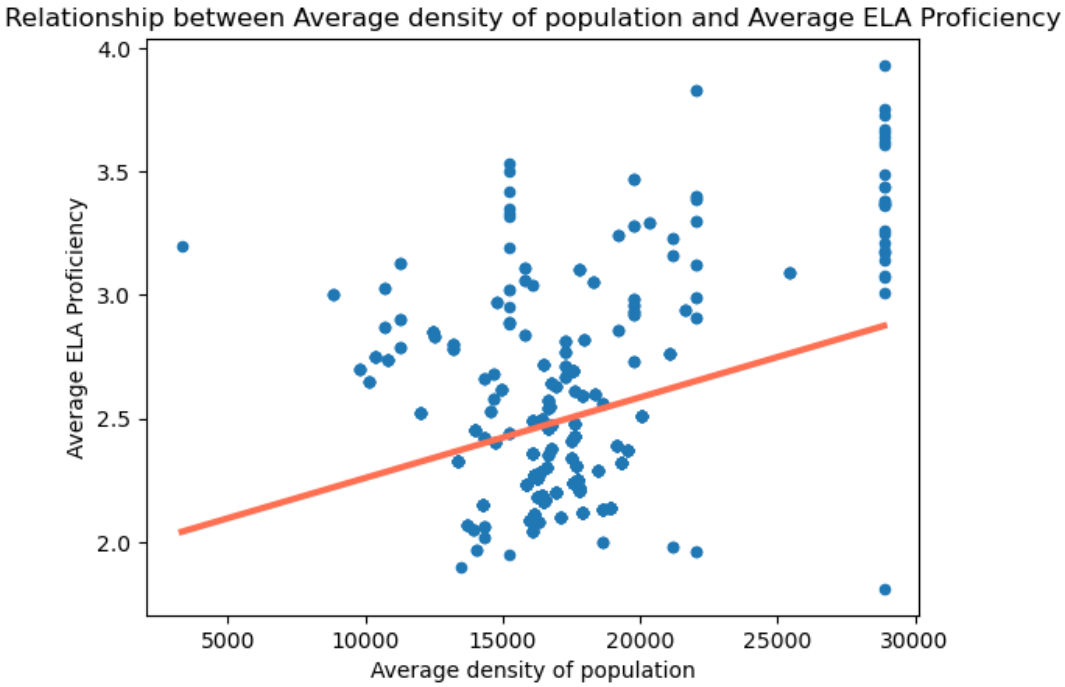
The quality of life in high-density and low-density populations differs across a range of factors, including access to resources, job opportunities, social connections, and cultural norms. Schools in high-density areas tend to offer a diverse range of social connections and services, albeit with limited personal space. Conversely, schools in low-density areas may provide more personal space and a stronger sense of community, along with access to nature, but fewer services. The density of population may thus influence students' behavior, living styles, and study habits in their neighborhoods.

To explore this hypothesis, I will gather demographic information about New York City from Wikipedia contributors (2023). Then, for each school in the original dataset, I will add columns including the area's population and population density. Finally, I will analyze whether there is a correlation between the average ELA Proficiency and the population and population density values.



The box plot reveals that, apart from the location with the lowest population density per km, there is a noticeable increase in average ELA proficiency as population density rises. However, due to the limited number of density of population data points, further analysis is needed to establish the precise relationship between population density and average ELA proficiency.

To generate a more informative scatter plot, some adjustments to the data are necessary. Specifically, a new column will be created to display the mean population density for each level of average ELA proficiency. For example, a row with an average ELA proficiency of 3.3 will have a new column that indicates the average population density when the average ELA proficiency is 3.3. This modification will enable the creation of a more meaningful and informative scatter plot.

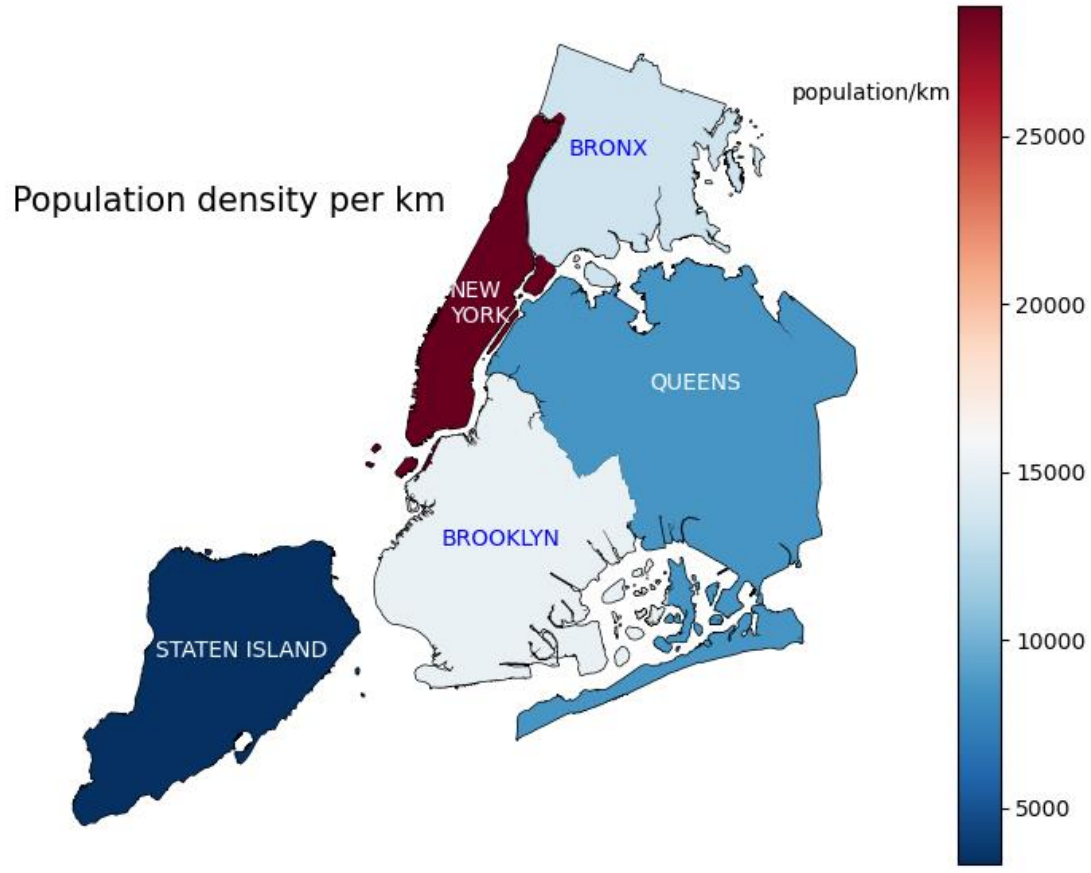


The scatter plot reveals that as the mean population density increases, so does the average ELA proficiency. This suggests that students living in higher-density environments tend to perform better. This may be attributed to the availability of more resources, job opportunities, social connections, and cultural norms in such areas.

However, the limited range of population density values in our current dataset hinders a more thorough analysis of this relationship. Thus, obtaining a new dataset that includes population density by zip code would be advantageous. In the following section, we will introduce a new dataset containing demographic information for each zip code area in New York City (Demographic Statistics by Zip Code, 2022).

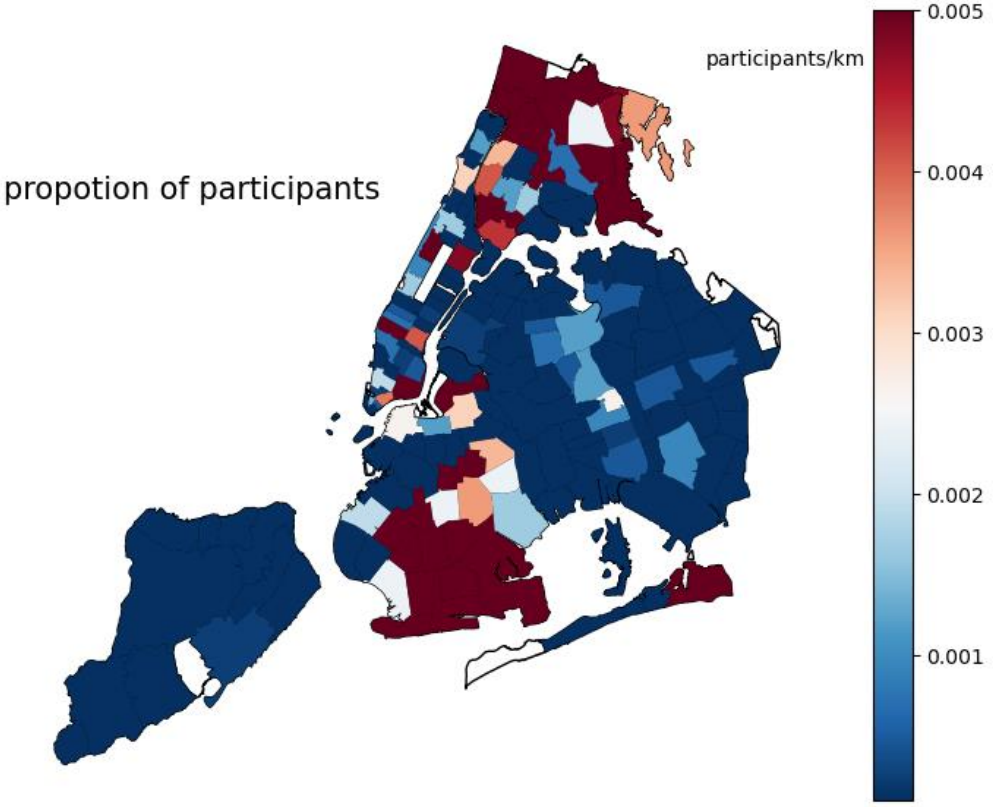
The new dataset includes information about residence in New York, including their genders, races, and other details. However, I will only be using the number of people in each zip code area for my analysis. I will use this number to calculate the proportion of the total number of people that each zip code area represents, which will be further used to represent the density of each area.

The problem is that the new dataset only includes 4168 people, which indicates that it does not represent the entire population of New York City. Therefore, before using the proportion of people in each zip code area to represent population density, it is necessary to verify its accuracy. This can be achieved by calculating the proportion of people in each zip code area and comparing it to the actual population distribution. To do this, we will plot the population distribution for each borough area and each zip code area and compare the two graphs to ensure that the dataset accurately reflects the population distribution of New York City.



This graph displays the population density distribution for the five borough areas of New York City. The redder areas indicate a higher population density, while darker blue areas have a lower population density.

It is apparent from the graph that Staten Island (the borough area in the bottom left corner) has the lowest population density, while Queens (the area in the bottom right corner) has the second-lowest. The area in the Bronx (upper right corner of the graph) has a slightly larger population density, followed by Brooklyn (the area in the bottom left corner of the right side of the graph). The largest population concentration is located in the New York (upper left corner of the right side of the graph).

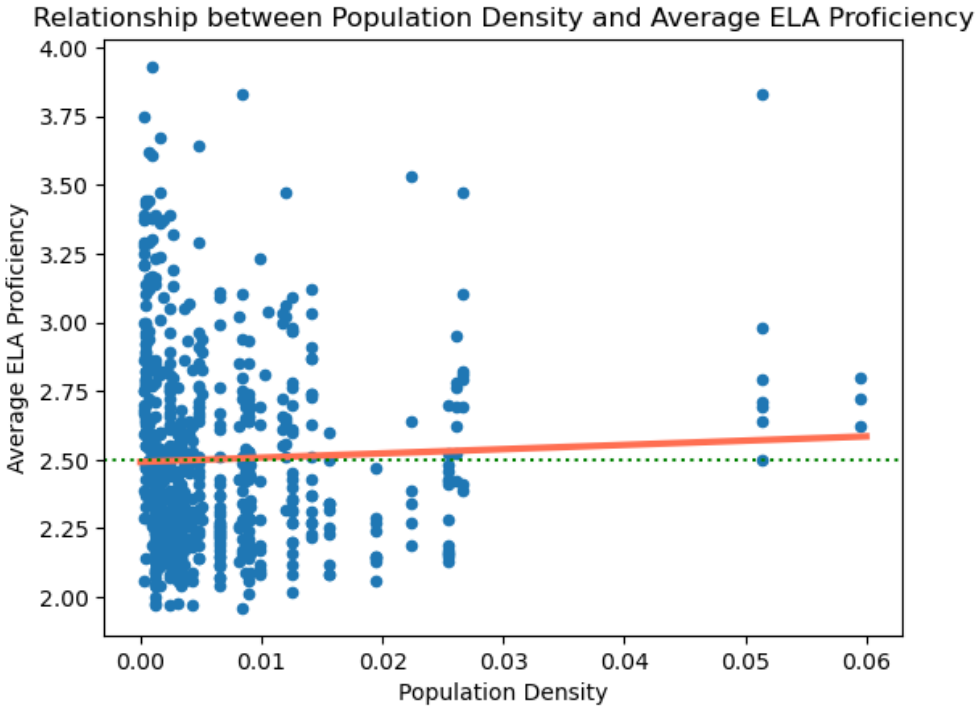


This graph displays the proportion of people in each zip code area of New York City, with red areas indicating a higher proportion and darker blue areas indicating a lower proportion.

It is apparent from the graph that areas with low population density in the borough graph also have a low proportion of people in the zip code graph. Similarly, areas with higher population density in the borough graph also have some red areas in the zip code graph, indicating a higher proportion of

people. This suggests that the proportion of people may be a useful proxy for population density in each zip code area.

Therefore, in the following analysis, the proportion of people will be used to represent the population density of each zip code area.



The scatter plot depicts the relationship between population density and average ELA proficiency, with a green horizontal line drawn at y=2.5 to aid in visualizing the trend. The red solid line and the line drawn in part 3.5 both show an upward slope, indicating a positive correlation between population density and average ELA proficiency. This suggests that students living in higher density environments may have a greater likelihood of performing well in English compared to those living in lower density environments. However, since the slope is small, we can conclude that the density of population may only have a slight influence on students' performance.

4. Regression

4.1 OLS Regression

In the previous section, we found that a school's environment, represented by supportive environment, strong family-community ties, effective school leadership, and trust, as well as the living standard (represented by house price) and the population density (represented by the proportion of people), can positively influence students' average performance in ELA. In order to investigate further, we will perform a regression analysis to determine if there is a linear relationship between these independent variables and the average ELA proficiency.

Additionally, we will investigate the impact of race on students' performance, although it may be influenced by their living standard. For now, we will include it in the model to gain a more accurate understanding of the relationship between these factors and academic achievement and to identify the factors that are most strongly correlated with student performance.

First, a regression will be run using all independent variables, including average house price, average environment, and density of population, Percent Black / Hispanic. For simplicity, the average environment will be used to represent the influence of supportive environment, strong family-community ties, effective school leadership, and trust. Once it is confirmed that there is a linear relationship between the average environment and the average ELA proficiency, the average environment will be separated into smaller parts for further investigation.

In other words, I expect my linear regression model to be:

$$\text{Average ELA Proficiency} = \beta_0 + \beta_1 * \text{Percent Black / Hispanic} + \beta_2 * \text{average house price} + \beta_3 * \text{density of population} + \epsilon$$

β_0: the intercept of the linear regression line on the y-axis (all independent variables = 0)

β_1: the coefficient for Percent Black / Hispanic.

β_2: the coefficient for average house price.

β_3: the coefficient for density of population.

ϵ: a random error term

Luckily, we have all the p-value smaller than 0.01. In other words, they all have significant effects on the average ELA proficiency. The positive coefficients of 1.054 and 0.086 for average environment and average house price, respectively, further suggest a positive correlation between these variables and the average ELA proficiency. However, the coefficients of Percent Black / Hispanic and density of population is negative. In other words, there is a negative correlation between these two variables and the average ELA proficiency.

	Dependent variable:Average ELA Proficiency
	(1)
Percent Black / Hispanic	-0.870***
	(0.023)
average environment	1.054***
	(0.126)
average house price	0.086***
	(0.017)
const	2.230***
	(0.114)
density of population	-3.285***
	(0.901)
Observations	1,149
R^2	0.617
Adjusted R^2	0.616
Residual Std. Error	0.222 (df=1144)
F Statistic	461.456*** (df=4; 1144)
Note:	*p<0.1; **p<0.05; ***p<0.01

Additionally, the R-squared value of 61.7% suggests that more than half of the variation in average ELA proficiency can be explained by our independent variables. Although this is a high value, it still indicates room for further improvement.

Therefore, out linear regression for now is:

$$\text{Average ELA Proficiency} = 2.230 - 0.87 * \text{Percent Black / Hispanic} + 0.086 * \text{average house price} - 3.285 * \text{density of population} + \epsilon$$

For the next steps, I plan to investigate the impact of smaller parts of the average environment in more detail to identify which specific factors have the strongest correlation with academic achievement.

In other words, my model 2 will be:

Average ELA Proficiency

$$= \beta_0 + \beta_1 * \text{Effective School Leadership} + \beta_2 * \text{Strong Family-Community Ties} + \beta_3 * \text{Supportive Environment} + \beta_4 * \text{Trust} + \epsilon$$

β_0: the intercept of the linear regression line on the y-axis (all independent variables = 0)

β_1: the coefficient for Effective School Leadership.

β_2: the coefficient for Strong Family-Community Ties.

β_3: the coefficient for Supportive Environment.

β_4: the coefficient for Trust.

ϵ: a random error term

	Dependent variable:Average ELA Proficiency
	(1)
Effective School Leadership %	1.306***
	(0.202)
Strong Family-Community Ties %	-0.249
	(0.193)
Supportive Environment %	2.524***
	(0.209)
Trust %	-1.735***
	(0.385)
const	1.002***
	(0.204)
Observations	1,149
R^2	0.184
Adjusted R^2	0.181
Residual Std. Error	0.324 (df=1144)
F Statistic	64.428*** (df=4; 1144)
Note:	*p<0.1; **p<0.05; ***p<0.01

The table clearly indicates that both Supportive Environment, Effective School Leadership and Trust have p-values lower than 0.01, indicating their significant effects on the average ELA proficiency. Nonetheless, the p-value for Strong Family-Community Ties is greater than 0.1, so we cannot conclude any influence it has on the Average ELA Proficiency.

Therefore, for the sequential model, I will exclude Strong Family-Community Ties and re-include the average house price, density of population, and Percent Black / Hispanic.

My new model 3 will be: Average ELA Proficiency $= \beta_0 + \beta_1 *$ Effective School Leadership $+ \beta_2 *$ Percent Black / Hispanic $+ \beta_3 *$ Supportive Environment $+ \beta_4 *$ Trust $+ \beta_5 *$ average house price $+ \beta_6 *$ density of population $+ \epsilon$

β_0: the intercept of the linear regression line on the y-axis(all independent variables = 0)

β_1: the coefficient for Effective School Leadership.

β_2: the coefficient for Percent Black / Hispanic.

β_3: the coefficient for Supportive Environment.

β_4: the coefficient for Trust.

β_5: the coefficient for average house price.

β_6: the coefficient for density of population.

ϵ: a random error term

Dependent variable:Average ELA Proficiency

	(1)
Effective School Leadership %	0.862***
	(0.139)
Percent Black / Hispanic	-0.858***
	(0.025)
Supportive Environment %	0.510***
	(0.146)
Trust %	-0.831***
	(0.263)
average house price	0.089***
	(0.017)
const	2.721***
	(0.143)
density of population	-3.221***
	(0.900)
Observations	1,149
R^2	0.621
Adjusted R^2	0.619
Residual Std. Error	0.221 (df=1142)
F Statistic	311.227*** (df=6; 1142)
Note:	*p<0.1; **p<0.05; ***p<0.01

The table indicates that all p-values are below 0.01, providing strong evidence of a relationship between the independent variables and the average ELA proficiency. The positive slopes for Effective School Leadership, Supportive Environment, and Average House Price are 0.862, 0.510, and 0.089, respectively, indicating a positive correlation between these variables and the average ELA proficiency. However, the slope for Percent Black / Hispanic, Trust and density of population are negative at -0.858, -0.831, -3.221, suggesting that an increase in these variables could lead to poorer average ELA proficiency.

Furthermore, the R-squared value has increased to 0.621, indicating a better fit of the model to the dataset.

Besides, The F-statistic of 311.227 is quite high, indicating a good fit of the model. Moreover, the p-value being less than 0.01 suggests that the probability of obtaining such a high F-statistic by chance alone is very low, typically less than 1%. This indicates strong evidence that at least one of the independent variables in the model is significantly related to the dependent variable.

Therefore, my model is:

$$\begin{aligned}\text{Average ELA Proficiency} &= 2.721 + 0.862 * \text{Effective School Leadership} + 0.858 * \text{Percent Black / Hispanic} \\ &+ 0.51 * \text{Supportive Environment} - 0.831 * \text{Trust} + 0.089 * \text{average house price} \\ &- 3.221 * \text{density of population} + \epsilon\end{aligned}$$

Additionally, we noted in project 2 that the impact of Percent Black / Hispanic on the average ELA proficiency may be influenced by their living standard, which could make it an endogenous variable affected by the Average House Price. Therefore, we plan to estimate an IV-2SLS regression.

4.2 IV-2SLS Regression

My new model will be:

$$\begin{aligned}\text{Average ELA Proficiency} &= \beta_0 + \beta_1 * \text{Effective School Leadership} + \beta_2 * \widehat{PercentBlack/Hispanic} + \beta_3 \\ &* \text{Supportive Environment} + \beta_4 * \text{Trust} + \beta_5 * \text{average house price} + \epsilon\end{aligned}$$

β_0: the intercept of the linear regression line on the y-axis(all independent variables = 0)

β_1: the coefficient for Effective School Leadership.

$\widehat{PercentBlack/Hispanic}$: the predicted value of Percent Black / Hispanic that depends on the value of density of population.

β_2: the coefficient for $\widehat{PercentBlack/Hispanic}$

β_3: the coefficient for Supportive Environment.

β_4: the coefficient for Trust.

β_5: the coefficient for average house price.

ϵ: a random error term

Dep. Variable:	Average ELA Proficiency	**R-squared:**	0.1598
Estimator:	IV-2SLS	**Adj. R-squared:**	0.1561
No. Observations:	1149	**F-statistic:**	300.93
Date:	Tue, Apr 04 2023	**P-value (F-stat)**	0.0000
Time:	11:14:13	**Distribution:**	chi2(5)
Cov. Estimator:	unadjusted		

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	4.7439	0.5807	8.1696	0.0000	3.6058	5.8820
Supportive Environment %	-1.4633	0.5908	-2.4769	0.0133	-2.6211	-0.3054
Effective School Leadership %	0.4836	0.2351	2.0574	0.0396	0.0229	0.9443
Trust %	0.0073	0.4625	0.0159	0.9873	-0.8992	0.9139
density of population	-5.0083	1.4109	-3.5497	0.0004	-7.7737	-2.2430
Percent Black / Hispanic	-1.7786	0.2530	-7.0297	0.0000	-2.2745	-1.2827

The table shows that after considering Percent Black / Hispanic as endogenous and estimating an IV-2SLS regression, the p-values for Trust become greater than 0.05, suggesting that these variables are no longer statistically significant. Therefore, for clarity, we will exclude this variable and run the regression again.

Dep. Variable:	Average ELA Proficiency	**R-squared:**	0.1600
Estimator:	IV-2SLS	**Adj. R-squared:**	0.1571
No. Observations:	1149	**F-statistic:**	283.87
Date:	Tue, Apr 04 2023	**P-value (F-stat)**	0.0000
Time:	11:14:22	**Distribution:**	chi2(4)
Cov. Estimator:	unadjusted		

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	4.7457	0.6278	7.5588	0.0000	3.5152	5.9762
Supportive Environment %	-1.4611	0.5130	-2.8482	0.0044	-2.4665	-0.4556
Effective School Leadership %	0.4869	0.1112	4.3780	0.0000	0.2689	0.7048
density of population	-5.0067	1.3924	-3.5958	0.0003	-7.7358	-2.2777
Percent Black / Hispanic	-1.7784	0.2451	-7.2559	0.0000	-2.2588	-1.2980

The p-values for all variables are now below 0.01, indicating that they all have a significant effect on the Average ELA proficiency. However, the IV-2SLS model has a lower R-squared value of 0.1600 compared to the previous OLS regression model, meaning that a smaller proportion of the data can be explained by the model. Moreover, the value of F-statistics also decrease. Therefore, I would prefer the previous OLS model with a higher R-squared value.

Therefore, I will consider my linear regression model as

$$\begin{aligned}\text{Average ELA Proficiency} &= 2.721 + 0.862 * \text{Effective School Leadership} + 0.858 * \text{Percent Black / Hispanic} \\ &+ 0.51 * \text{Supportive Environment} - 0.831 * \text{Trust} + 0.089 * \text{average house price} \\ &- 3.221 * \text{density of population} + \epsilon\end{aligned}$$

4.3 Machine learning

As our linear regression model is:

Average ELA Proficiency

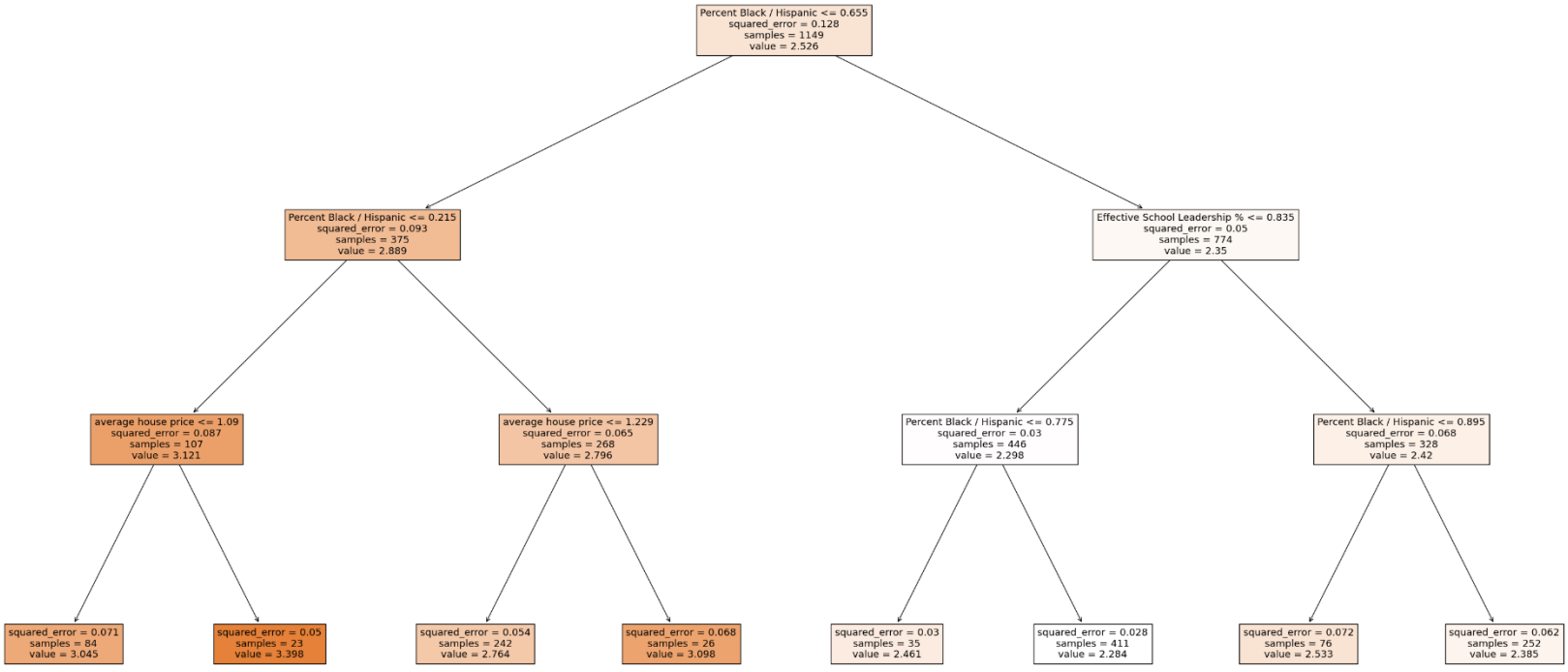
$$= 2.721 + 0.862 * \text{Effective School Leadership} + 0.858 * \text{Percent Black / Hispanic}$$
$$+ 0.51 * \text{Supportive Environment} - 0.831 * \text{Trust} + 0.089 * \text{average house price}$$
$$- 3.221 * \text{density of population} + \epsilon$$

In the following regression tree part, I decide to choose only the independent variables in this model to predict our dependent variable (average ELA proficiency).

The regression tree is trying to minimize Mean Squared Error (MSE) by splitting the data into two parts.

Each split can be written as: $\min_{j,s} \left[\sum_{i:x_{i,j} \le s, x_i \in R1} (y_i - \widehat{y_{R1}})^2 + \sum_{i:x_{i,j} > s, x_i \in R2} (y_i - \widehat{y_{R2}})^2 \right]$

where y_i is the actual value of the dependent variable (Average ELA Proficiency) for the i-th observation, $\widehat{y_{R1}}$ and $\widehat{y_{R2}}$ is the predicted value of the dependent variable for each split based on the regression tree.



The regression tree that I generated reveals that the initial split can be expressed as:

$$\min_{j,s}\left[\sum_{i:PercentBlack/Hispanic_i \le 0.655, PercentBlack/Hispanic \in R1}(y_i - \widehat{y_{R1}})^2 + \sum_{i:PercentBlack/Hispanic_i > 0.655, PercentBlack/Hispanic \in R2}(y_i - \widehat{y_{R2}})^2\right]$$

where y_i denotes the actual value of the dependent variable, i.e., Average ELA Proficiency, for the i-th observation, and $\widehat{y_{R1}}$ and $\widehat{y_{R2}}$ denote the predicted values of the dependent variable for each split based on the regression tree.

The algorithm determined that splitting the data by the variable Percent Black / Hispanic at the threshold of 0.655 results in the lowest mean squared error (MSE) at the first step.

Furthermore, I found that the minimum leaf size of my regression tree is 23, which occurs when the sample has a Percent Black / Hispanic value of less than or equal to 0.215 and an average house price greater than 1.09. The maximum depth of the tree is set to 3, and these two parameters control the complexity of the tree.

Increasing the maximum tree depth could potentially result in a lower MSE for the regression tree. In other words, we might have lower possible errors by increasing tree depth.

At the same time the Mean Squared Error for linear regression is 0.048537, while the Mean Squared Error for regression tree: 0.048355. In other words, difference between two Mean Squared Error: 0.000182. Upon examining the results, it can be observed that the difference between the Mean Squared Error values for the two models is minimal. Nonetheless, the Mean Squared Error value for the regression tree is lower than that of the linear regression model, indicating that the predictions made by the regression tree are more accurate and closer to the actual values. As a result, the regression tree model is a better fit for the data and can more accurately predict the dependent variable.

Furthermore, the linear regression tree provides us with valuable information on how to partition our data into smaller groups in order to minimize errors. This allows for more targeted investigations into student performance. However, linear regression provides an overall prediction for the entire sample group, which is more convenient but less precise.

5. Conclusion

Selecting a suitable school for their children is a crucial decision for parents, as it can have a significant impact on their educational outcomes. However, this task can be challenging due to the multitude of variables that may affect students' academic performance. Therefore, understanding the factors that influence academic performance beyond just curriculums and teachers is an important question to address. While most previous projects have focused on subjective factors that may influence student performance, I aim to examine objective factors, such as the average environment, racial makeup, living standards, and population density.

After analyzing 1270 schools in New York City, we can conclude that a school's environment has a positive correlation with average ELA proficiency, with supportive environment, strong family-community ties, effective school leadership, and trust being the key components. However, race does not have a direct impact on academic performance, but environmental factors such as living conditions may explain the differences in performance between different racial groups. For example, students from poorer environments, such as Black/Hispanic students, may have limited access to better learning conditions, which may lead to poorer academic performance. Additionally, we found that population density may also influence academic performance, with students from areas with higher population densities performing better in school.

Moving forward, further research can explore these factors in greater depth, and investigate other factors that may contribute to academic success, such as student motivation, teacher quality, or extracurricular activities. This could help to build a more comprehensive understanding of the factors that

contribute to academic success and guide government subsidies for schools to improve educational outcomes.

In addition, our regression model results contradict some of our previous analysis. We found that trust and population density may have a negative impact on academic performance, with trust being an inefficient independent variable that may decrease the efforts put into other areas. Moreover, Strong Family-Community Ties had no significant impact on academic performance. We also discovered that race directly impacts academic performance, rather than indirectly through living standards. Further investigation could focus on exploring the underlying mechanisms of these findings.

Furthermore, it is important to note that our analysis only considered linear regression and regression tree models. Future studies could incorporate non-linear models to provide a more nuanced understanding of the relationships between various factors and academic performance.

Reference

1. Borough Boundaries | NYC Open Data. (n.d.). NYC Open Data. https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm
2. Data Science for Good: PASSNYC. (2018). [Dataset]. PASSNYC. https://www.kaggle.com/datasets/passnyc/data-science-for-good?resource=download
3. Demographic Statistics By Zip Code. (2022). [Dataset]. NYC OpenData. https://data.cityofnewyork.us/City-Government/Demographic-Statistics-By-Zip-Code/kku6-nxdu
4. Edgerton, E., & McKechnie, J. (2023). The relationship between student's perceptions of their school environment and academic achievement. Frontiers in Psychology. https://doi.org/10.3389/fpsyg.2022.959259
5. Real Property Income and Expense Form non-compliance list. (2018). [Dataset]. NYC OpenData. https://data.cityofnewyork.us/City-Government/Real-Property-Income-and-Expense-Form-non-complian/wvts-6tdf
6. Shi, Y., & Ko, Y. C. (2023). A Study on the Influence of Family and School Psychological Environment on Academic Self-Efficacy and Self-Identity of English Education Major University Students. Participatory Educational Research (PER). https://doi.org/http://dx.doi.org/10.17275/per.23.6.10.1
7. Steinmayr, R., Weidinger, A. F., Schwinger, M., & Spinath, B. (2019). The Importance of Students' Motivation for Their Academic Achievement – Replicating and Extending Previous Findings. Frontiers in Psychology, 10. https://doi.org/10.3389/fpsyg.2019.01730
8. Wikipedia contributors. (2023, March 24). Demographics of New York City. Wikipedia. https://en.wikipedia.org/wiki/Demographics_of_New_York_City
9. Zip Code Boundaries. (2018). [Dataset]. NYC OpenData. https://data.cityofnewyork.us/Business/Zip-Code-Boundaries/i8iw-xf4u