From Classrooms to Paychecks*

Education and Income Inequality Across Santa Clara County Neighborhoods

Abhishek Rasikbhai Shekhada

October 25, 2025

This paper examines whether neighborhood-level differences in educational attainment are linked to household income across Santa Clara County. Using tract level data from the American Community Survey (ACS) provided by the City of San José (City of San José 2025). I apply a simple linear regression model with median household income as the response variable and the share of adults without a high school diploma as the predictor. The results show a strong negative relationship between tracts with lower educational attainment tend to have lower median household incomes. These finds highlight how educational disparities continue to shape local economic outcomes even within a wealthy region. The analysis suggests that improving education access may help reduce income inequality across neighborhoods.

Introduction

Education has long been recognized as one of the most important factors associated with people's economic opportunities. Individuals who complete more schooling typically earn higher incomes, enjoy greater job stability, and have access to a wider range of careers (Card 1999; Baum, Ma, and Payea 2003). On the other hand, communities where fewer adults complete high school often face long-term challenges such as lower wages, reduced access to resources, and fewer chances for upward mobility. These ideas motivate me to explore the relationship between education and income at the neighborhood level, where disparities are often less visible but still meaningful.

Santa Clara County, the heart of Silicon Valley, provides an especially interesting setting for this question. The county is well known for its wealth and innovation, yet not all neighborhoods share equally in this prosperity. Some tracts benefit from high levels of education and the rewards of the tech economy, while others lag behind in both schooling and earnings. Looking

*Project repository available at: https://github.com/arshekhada/Math-261A-Project-1.

at these differences within a single county can reveal how local variations in education are related to differences in household income, even in a region that is often considered wealthy overall.

Most research on this topic has focused on individuals or national trends, leaving a gap in understanding how education and income interact across neighborhoods within the same county. This paper addresses that gap by asking: *Do census tracts with a higher share of adults lacking a high school diploma also report lower median household incomes?* To answer this question, I use tract level data from the American Community Survey and apply a simple linear regression model with median household income as the outcome and the percent of adults without a high school diploma as the predictor. It's important to note that this analysis is correlational rather than causal it identifies associations between education and income across neighborhoods but does not establish that one directly causes the other.

The results show a strong negative association: as the share of adults without high school completion increases, median household income tends to decrease. This finding is important because it highlights how educational inequality correlates with local economic outcomes, even in one of the richest regions of the United States. Most studies on education and earnings focus on individuals or large-scale national trends (Card 1999; Baum, Ma, & Payea, 2003) but less is known about how these relationships play out across neighborhoods within a single county. This study helps fill that gap by examining local variations in education and income within Santa Clara County, providing a more detailed picture of inequality at the community level. The rest of this paper is organized as follows: the **Data** section describes the dataset and variables, the **Methods** section explains the regression model and its assumptions, the **Results** section presents the fitted model and interpretations, and the **References** section lists all sources used.

Data

The observational units in this study are *census tracts* within Santa Clara County, California. A census tract is a small, relatively stable subdivision created by the U.S. Census Bureau that is designed to capture populations of roughly similar size and characteristics. Each row of the dataset corresponds to one tract and the cleaned dataset used in this analysis contains 150 tracts with complete information. Focusing on tracts rather than individuals allows for a neighborhood level view of the relationship between education and income. Because the analysis uses data aggregated at the census tract level rather than for individual people, this study is ecological in nature. This means the results describe patterns across neighborhoods not for specific individuals so no individual level conclusions can be made from the analysis.

The dataset is called *Equity Index Census Tracts* and was compiled by the City of San José using the **American Community Survey (ACS) 2021 5-year estimates** (City of San José 2025). The ACS is a nationwide survey conducted by the U.S. Census Bureau that collects detailed demographic, social, housing and economic information from a representative sample of households in every community across the United States. It includes people living in both regular housing units and group quarters such as dormitories, nursing homes and shelters. Because of its scope and standardized design, the ACS provides consistent and comparable data across all census tracts. While the full dataset includes many indicators related to race, income, language and education, this project focuses on two main variables:

- **Outcome variable**: Median Household Income (INCMEDIANINCOME), which measures the midpoint household income in U.S. dollars for each tract.
- **Predictor variable**: Percent Without High School Education (EDULESSTHANHSRATIO), which captures the share of adults aged 25 or older in a tract who did not complete a high school diploma or equivalent.

Both variables are continuous: income is reported in dollars and the education variable is expressed as a proportion between 0 and 1. Some data preparation was needed before analysis. I removed columns that contained mostly missing values, such as income broken down by race, and I excluded derived equity score variables to avoid circularity since some of them already combine education and income. I also dropped rows with missing values for the two main variables. After cleaning, the dataset contained 150 complete tracts and 38 useful variables, although only two are used in the regression. Median household incomes in the cleaned data range from about $41,000 to $250,000, with some values top-coded at the ACS limit. The share of adults without a high school diploma ranges from close to 0% in some tracts to about 45% in others, showing substantial variation across neighborhoods.

The ACS is widely considered the most reliable source for tract-level socioeconomic data because it uses large, continuous samples and standardized data collection across all U.S. regions. Its five-year estimates strike a good balance between precision and stability, making it especially suitable for small area studies like this one. However, because it is still a survey, it remains subject to sampling error and nonresponse bias. For example, smaller tracts may have less precise estimates and income is self-reported, which can introduce measurement error. Alternative or additional sources could include California Open Data, San Francisco Open Data or the Harvard Dataverse, which also provide socioeconomic indicators at different geographic scales. Still, the ACS remains the best choice for obtaining consistent, reliable data across all tracts in the county. To illustrate key characteristics of the data, I generated summary tables and visualizations such as histograms of income and education and a scatterplot of the two variables, which already suggested a strong negative association that motivated the regression analysis.

Methods

To examine the connection between education and income, I used a **simple linear regression** (SLR) model. This method estimates how changes in one explanatory variable are associated with changes in a response variable. In my case, the explanatory variable is the **percent of**

adults without a high school diploma in each census tract and the response variable is the **median household income** of the tract.

Formally, the model can be written as:

$$\text{Income}_i = \beta_0 + \beta_1(\%\text{NoHS})_i + \varepsilon_i$$

where: -

- Income_i: The median household income in tract i (measured in U.S. dollars)
- $(\%\text{NoHS})_i$: The fraction of adults aged 25 or older without a high school diploma in tract i
- β_0: The intercept which represents the predicted income when the predictor equals zero
- β_1: The slope which shows the expected change in income when the predictor increases by one unit (interpreted as about $3,097 per one percentage point increase)
- ε_i: The random error term, which represents unobserved factors that affect income but are not included in the model. It is assumed to have a mean of zero, constant variance and to be independent across tracts.

Like all regression models, this one relies on several assumptions about the error term ε_i:

1. **Linearity** – The expected value of income is a linear function of the predictor, $E(\text{Income}_i) = \beta_0 + \beta_1(\%\text{NoHS})_i$.

2. **Independence** – The error terms ε_i are independent across census tracts.

3. **Constant variance (homoscedasticity)** – The variance of the error term, $\text{Var}(\varepsilon_i)$, is the same for all tracts..

4. **Normality** – The error terms (ε_i) are normally distributed with a mean of zero.

In practice, these assumptions cannot be observed directly, so I checked them using the **residuals** (the sample estimates of ε_i). I examined scatterplots and diagnostic plots of the residuals to confirm that the linear model appeared reasonable for this tract level analysis. While real world data rarely fit every assumption perfectly, the model provided a good overall fit for the purpose of this project.

This study has several **limitations**. The model is bivariate, meaning it only considers one predictor. Other important factors such as housing costs or employment sectors may also influence income but are not included here. In addition, the data are aggregated at the census tract level, so the results cannot be interpreted as applying to individual households. Finally, because the ACS is survey-based, there is some sampling error, particularly in tracts with smaller populations.

All analyses were carried out in **R** (R Core Team 2025) using the packages *readr* (Wickham, Hester, and Bryan 2024), *ggplot2* (Wickham 2016), *scales* (Wickham and Seidel 2022), *dplyr* (Wickham et al. 2023), and *broom* (Robinson and Hayes 2023). The workflow was designed to be reproducible: I cleaned the raw dataset, saved a cleaned version, and then fit the regression model on that file. Although this model captures a clear association between education and income, it does not include other potential predictors such as housing costs, industry mix, or demographic composition, which may also influence neighborhood income levels. Future extensions could incorporate these variables in a multiple regression framework to see whether the effect of education remains strong when controlling for other factors. Including additional predictors would also help address potential omitted variable bias and improve the model's explanatory power.

I validated the model fit by reviewing standard regression diagnostics and confirming that the main conclusions remained stable.

Results

The fitted regression model shows a clear negative relationship between education and income at the neighborhood level. The estimated equation is:

$$\widehat{\text{Income}} = 183,922 - 309,721 \times (\%\text{NoHS})$$

Here, the intercept of about \$183,922 represents the predicted median household income for a tract where all adults have at least a high school diploma. The slope coefficient is negative and large in magnitude. When interpreted on the percentage scale, it means that for every one percentage point increase in adults without a high school diploma, the median household income is predicted to decrease by about **\$3,097**.

The model explains nearly half of the variation in tract-level incomes, with an R^2 value of about 0.49. The slope coefficient was tested using a t-test, which determines whether the estimated slope is significantly different from zero.

- **Null hypothesis (H_0):** $\beta_1 = 0$ (no linear relationship between education and income)
- **Alternative hypothesis (H_1):** $\beta_1 \neq 0$ (a linear relationship exists between education and income)

The very small p-value ($p < .001$) provides strong evidence against the null hypothesis, suggesting that the negative association between education and income is statistically significant. In other words, census tracts with higher shares of adults lacking a high school diploma tend to have substantially lower median household incomes.

For this test to be valid, the main regression assumptions linearity, independence, constant variance, and normality of the errors (ε_i) should be reasonably satisfied. I checked these assumptions informally using diagnostic plots and by visually inspecting the residuals. The residuals appeared roughly evenly spread around zero and showed no clear nonlinear patterns, suggesting that the linear model was appropriate.

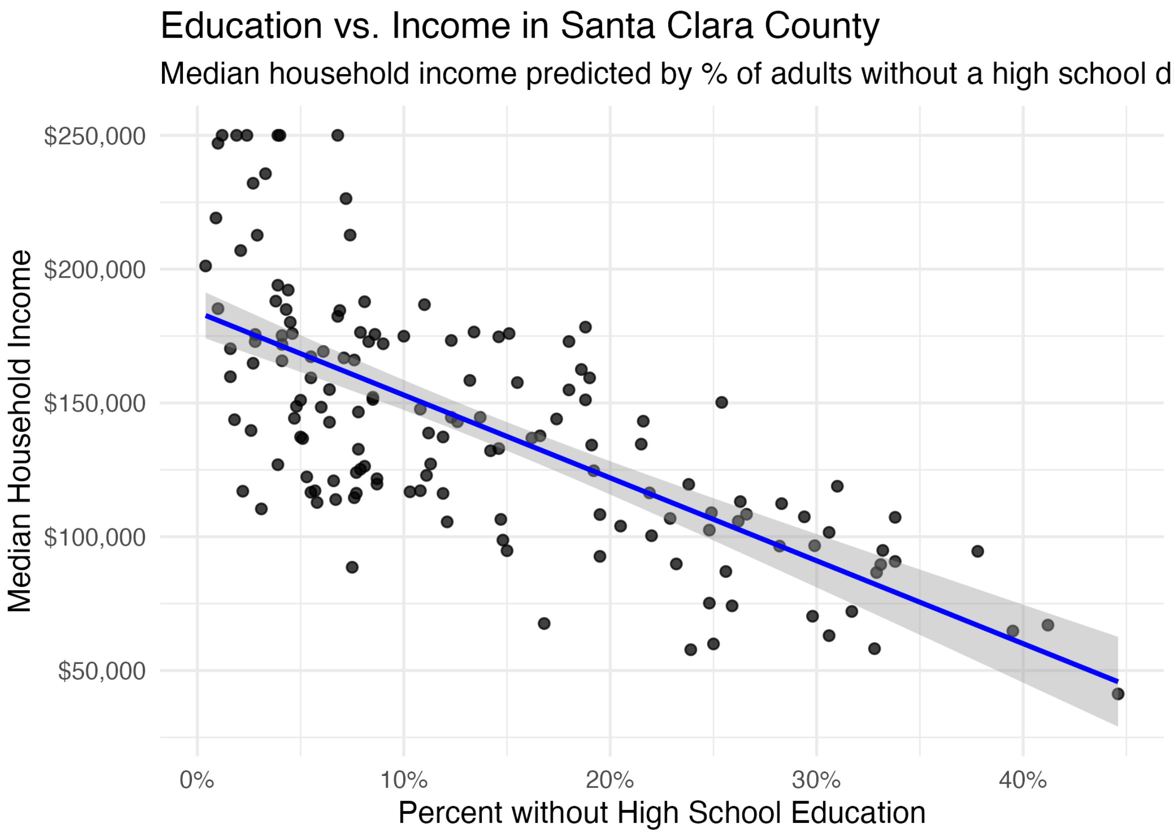

Figure 1: Education vs. Income in Santa Clara County. Relationship between educational attainment and median household income across Santa Clara County census tracts. Each point represents one census tract. The fitted regression line (blue) with a 95% confidence band shows that neighborhoods with a higher share of adults without a high school diploma tend to have lower median household incomes.

Parameter	Estimate	Std. Error	t-statistic	p-value
Intercept	183922	4431	41.51	2.01E-83
Percent without High School (ratio)	-309721	25995	-11.91	2.27E-23

Results from a simple linear regression predicting median household income from the percent of adults without a high school diploma. The intercept is approximately \$183,922, and the slope is –309,721 ($p < .001$) meaning that income decreases by about \$3,097 for every additional percentage point of adults without a high school diploma.

Discussion

The results from this analysis show a clear and statistically significant negative relationship between education and income across neighborhoods in Santa Clara County. In simple terms, areas with more adults who did not finish high school tend to have lower median household incomes. This directly answers the main research question and supports the idea that education is closely related to local economic outcomes. Even though Santa Clara County is known for being wealthy overall, the results show that not all communities share that prosperity equally.

I found it especially interesting how consistent the pattern was even small differences in education seemed to line up with noticeable changes in income. This suggests that improving education access and completion rates could be an effective way to reduce neighborhood level income inequality. These findings are also in line with earlier studies showing that education strongly influences earnings and job opportunities (Card 1999; Baum, Ma, and Payea 2003). In the context of a place like Silicon Valley, where most jobs require advanced skills these gaps in education can have even larger effects on local income differences.

Of course, this analysis has some limitations. The model only looks at one variable, so it doesn't consider other factors that might affect income like housing costs, job types or transportation access. Because the data come from the census tract level, the results describe neighborhood trends not individual outcomes. There is also some uncertainty in survey data like the American Community Survey, which can introduce small errors or sampling variation. Still, the relationship between education and income here is strong enough that these limitations are unlikely to change the main takeaway.

For future work, it would be useful to include more predictors in a multiple regression model to see how education interacts with other social and economic factors. It may also be interesting to examine whether these patterns have changed over time such as whether the education income gap is widening or narrowing across different parts of the county. Overall, this simple regression provides a clear picture of how education remains closely linked to economic opportunity in local communities.

References

Baum, Sandy, Jennifer Ma, and Kathleen Payea. 2003. "Higher Education and Future Earnings: Evidence from Longitudinal Data." *Review of Higher Education* 26 (3): 361–80. https://doi.org/10.1353/rhe.2003.0012.

Card, David. 1999. "The Causal Effect of Education on Earnings." In *Handbook of Labor Economics*, 3:1801–63. Elsevier. https://doi.org/10.1016/S1573-4463(99)03011-4.

City of San José. 2025. "Equity Index Census Tracts (ACS 2021 5-Year Estimates)." Data set. https://gisdata-csj.opendata.arcgis.com/datasets/CSJ::equity-index-census-tracts.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, and Alex Hayes. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wickham, Hadley, and Dana Seidel. 2022. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.