Economic Factors Contributing to Variations in Housing Prices in the U.S.

Group Members: David King and Nick Pinto

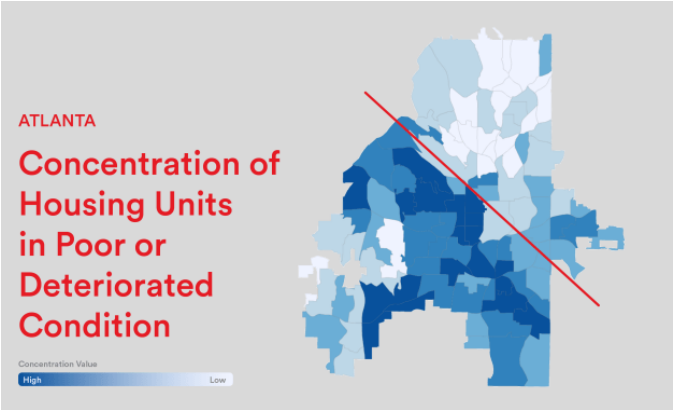
Abstract

As income inequality in the U.S. rises, a pattern of residential segregation becomes more apparent. This paper seeks to test if differences in education can account for variation in housing prices across zip codes. We have developed regression models to analyze several economic factors, mainly education, and their relationship with housing costs. We discovered a positive linear relationship between the percentage of people with higher levels of college education and higher housing prices in the zip code. These findings support the idea that education is a predictor of geographic stratification.

I. Introduction

Every city has its rich neighborhoods. All across America, there are pockets of wealth and homogeneity that stand in stark contrast to their surroundings. Atlanta, for example, is consistently ranked as one of the top cities in the country for income inequality.

Graphic 1: Atlanta Housing Map



There is a clear pattern of residential segregation, dividing in the city into two halves, but, more than that, there are smaller geographic clusters of wealth around the city. For instance, the area surrounding Georgia Tech is extraordinarily wealthy, while nearby English Avenue and Vine City have some of the highest poverty and crime rates in the city. Much of this can be attributed to the effects of redlining, a discriminatory practice in which banks refused loans and mortgages within specific geographic areas, mostly neighborhoods with high minority populations. However, in this paper, we seek to explore other factors that can lead to these geographical inequalities, mainly education.

Income inequality is growing, and it continues to result in a pattern of residential segregation. According to Wheeler and La Jeunesse (2006), the U.S. Census reports that the variance in income has increased by almost 25% between 1970 and 2000. The U.S. Census also reports that the share of poor Americans living severe poverty has risen to 45.6% (Bialik). Segregation threatens social mobility and harms the American Dream. This is an issue so important that the 10th UN Sustainable Development goal, which we have chosen to focus on, is “Reduced Inequalities.” This paper seeks to identify economic factors that lead to residential inequality, so we can better understand how inequality can be reduced.

Education is often hailed as the great equalizer. In theory, the public education system gives everyone the opportunity to succeed in America and to move upwards in society. Equal access to education, however, is not a reality. Families with more wealth can afford to give their children higher and better quality education, but, with the cost of college rising, poorer families cannot. We seek to

examine in this paper if differences in education can explain variation in housing prices. We hypothesize that higher educated people usually have higher incomes and can afford to live in more expensive areas, usually with other wealthy, well-educated individuals, creating geographical pockets of wealth.

II. Literature Review

In the first paper we examined, Guerrieri, Hartley and Hurst (2013) examined the variation in home price increases within neighborhoods across America. They noticed that neighborhoods in a city do not all experience home value appreciation together at the same rate and determined there is a systematic pattern in the variation which is linked to gentrification. There is a sorting process, where rich individuals are willing to pay more to live closer to richer neighbors, and poorer residents are less willing to pay the higher housing prices to live near the rich, so they tend to live further away. This creates an equilibrium where the rich residents are concentrated together with low income residents at the periphery. Home values are thus higher and increase faster in these wealthy residential clusters. This is consistent with other papers which find clusters of substantial income and wealth among small housing units.

In another paper, Fryer Jr. and Katz (2013) explore the factors that can allow children to escape the gravitational pull of poverty. Racial inequality remain a consistent regularity, and the researchers examined whether high-quality schools alone weaken the cycle of intergenerational poverty for those living in high-poverty areas. They analyzed data from the MTO housing mobility experiment, where families in high poverty areas received vouchers to live in private rental housing or Section-8 housing. The results suggested that investments in school quality are effective in decreasing persistent economic and educational inequalities and for reducing risky behaviors, while neighborhood improvements work to reduce mental and physical health inequalities. The vital policy question that results is how to generate systematic large-scale improvements in school and teacher quality for low-income students growing up in high-poverty neighborhoods.

The final paper we examined, authored by Kearney and Levine (2016), is slightly different from the others; its tests whether higher rates of income inequality are correlated with higher high school dropout rates. The researchers sought to discover if income inequality leads to the perception of a lower rate of return of investment in a student's own human capital. Although education is a key pathway in which an individual may obtain a middle-class life or higher, this may seem less true in areas of high income inequality. According to their results, a greater gap might contribute to a heightened sense of economic marginalization. Students, especially boys, are more likely to drop out if they live in areas with high income inequality, controlling for individual and family demographics.

These papers provided an important foundation for our own analysis. We knew from Guerrieri, Hartley and Hurst (2013) that there are geographic patterns of income inequality, in the form of wealthy residential clusters. Based on Fryer Jr. and Katz (2013) and Kearney and Levine (2016), we hypothesized that education is a predictor of these patterns. Because education can reduce and is intertwined with income inequality, we looked at differences in education to further explore how residential segregation occurs and how income inequality happens. We are also contributing to the literature by performing quantitative analysis on thousands of zip codes across the U.S. We will create a model that incorporates other factors as well, including race, to attempt to explain variations in housing prices.

III. Data

Our dependent variable is the logarithm of median household price in a zip code. Our independent variables that we have chosen are split into two types. The first is the variables of interest, which are the ones that we have decided are the best measures of education for a zip code. They include percent of residents who are high school dropouts by zip code, percent of residents with only a high school diploma by zip code, percent of residents with a bachelor's degree by zip code, and percent of residents with a graduate degree by zip code. Our other independent variables are control variables, included to prevent violation of Gauss-Markov Assumptions. In other words, if we left out these variables, it would violate assumption MLR.4 because these variables have an effect on the median household price, and they are correlated with the other independent variables of interest.

Some variables were considered but not included because excluding these variables did not violate any of the Gauss-Markov assumptions--that is either they had little effect on median household prices, or they had very low collinearity with levels of education. Variables such as percentage of the population that is male/female and total population were excluded because they were not significantly correlated with median household prices.

For the purposes of this research, we have compiled three linear regressions. The first is a simple linear regression that looks at how percentage of the population that are high school dropouts affects median household prices. The second is a multiple linear regression that looks at how the three different levels of education affect median household prices. The final is a multiple linear regression that expands on the first multiple linear regression to remove violations of the Gauss-Markov assumptions.

The data on our independent variables was obtained from the American Community Survey on Educational Attainment, which included a random sample of several thousand Americans and contained data by zip code. Some of the observations had to be dropped because they did not include data on median income, which would have skewed our simple linear regression and first multiple linear

regression. The data on median household prices by zip code was obtained through Zillow Housing Data. Some basic information about the variables that we are using is contained in the table below.

Table 1: Summary Statistics

Variable	Observations	Mean	Std. Dev.	Min.	Max.
Median Household Price	7190	265373.7	225758.4	32600	4892500
Per HS Dropouts	7190	42.17292	9.186326	4.2	78.6
Per w/ HS Diploma	7190	28.47993	9.195567	1.3	61.2
Per w/ Bach Degree	7190	18.37558	8.244074	0	52.3
Per w/ Grad Degree	7190	10.97157	7.533541	0	64.7
Median Income	7190	38138.45	11064.3	10690	111151
Percent Asian Only	7190	4.327745	7.262386	0	64.7925
Percent White Only	7190	68.31158	19.56906	1.189918	96.45502
Percent Black Only	7190	9.943952	14.91144	0	86.61782
Percent Hispanic Only	7190	10.92057	14.14587	0	87.09892

Our three regressions follow the Gauss-Markov assumptions to a varying degree, with our second multiple linear regression model following them most closely.

For the first assumption, each of our models is in some form of $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \mu$, where y is the logarithm of median household price of the zip code, x_1 is percent of people 25+ in the zip code with a high school diploma, x_2 is the percent of people 25+ in the zip code with a bachelor's degree, x_3 is the percent of people 25+ in the zip code with a graduate's degree, x_4 is the median income of the zip code, x_5 is the percentage of the population that is asian only in the zip code, x_6 is the percentage that is white only, x_7 is the percentage that is black only, x_8 is the percentage that is hispanic only, and μ is the unobserved random error. In the first model, x_1 is percent of people 25+ without a high school diploma. In all three regressions, all β's are linear, therefore our models are linear in parameters.

For the second assumption, the surveys that we obtained the data from were conducted randomly, and therefore all three models follow this assumption.

For the third assumption, no variable is constant throughout all data points. In addition, there is no perfect collinearity among independent variables. Below is a correlation table of our variables of interest.

Table 2: Correlation Table

	Per w/ HS Diploma	Per w/ Bach Degree	Per w/ Grad Degree	Per w/ Grad Degree	Percent Asian Only	Percent White Only	Percent Black Only	Percent Hispanic Only
Per w/ HS Diploma	1							
Per w/ Bach Degree	-0.8051	1						
Per w/ Grad Degree	-0.7539	0.8181	1					
Median Income	-0.5558	0.7586	0.7518	1				
Percent Asian Only	-0.3848	0.3370	0.2741	0.2908	1			
Percent White Only	.137121 3	0.1261	0.0744	0.2350	-0.3940	1		
Percent Black Only	0.0773	-0.2139	-0.1377	-0.2609	-0.7755	-0.0854	1	
Percent Hispanic Only	-0.1451	-0.1813	-0.1941	-0.2492	0.1593	-0.2866	-0.0432	1

Since the correlation coefficient does not equal 1 for any field, no independent variable is a linear combination of the others, and the assumption is not violated. An important thing to note is that the high multicollinearity among some of the variables, mainly those related to education, could lead to high variances in our estimators.

For the fourth assumption, it is followed in only the second multiple linear regression. In the simple linear regression and our first multiple linear regression, this assumption is violated because there are variables that affect the median household prices of a zip code that are contained in the μ term that are also affected by our independent variable. As explained above in the section for independent variables that were not included, we believe that with the second linear regression model, there is no information contained in μ term that is affected by any of our x variables and also affects median household prices.

For the fifth assumption, it is hard to detect homoscedasticity. In our simple linear regression, this assumption may be violated because the variance of μ probably changes given different values for the independent variable. If there is a higher percentage of the population with a high school diploma, it could cause a greater variance in other levels in education (contained in μ). In our first multiple linear regression, this assumption may be violated because the variance of μ probably changes given different values for the independent variable. If there is a higher percentage of the population with a degree, it could cause a greater variance in median income (contained in μ). In our second multiple linear regression model, we have to show the variance of μ does not change given different values for the independent variable, or that $Var(y|x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8) = \sigma^2$. This might be difficult because the variance of housing prices could change given different values of each of our independent variables. For the purposes of this regression, we are assuming that the data exhibits homoscedasticity.

Our second multiple linear regression model follows MLR.1-MLR.5, so we can assume that the ordinary least squares estimators for $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$, $\hat{\beta}_5$, $\hat{\beta}_6$, $\hat{\beta}_7$, and $\hat{\beta}_8$ are the best linear unbiased estimators (BLUEs) for β_0, β_1, β_2, β_3, β_4, β_5, β_6, β_7, and β_8 respectively. Our first two models violated one or more of the Gauss-Markov assumptions, so their estimators are not the BLUEs.

IV. Results

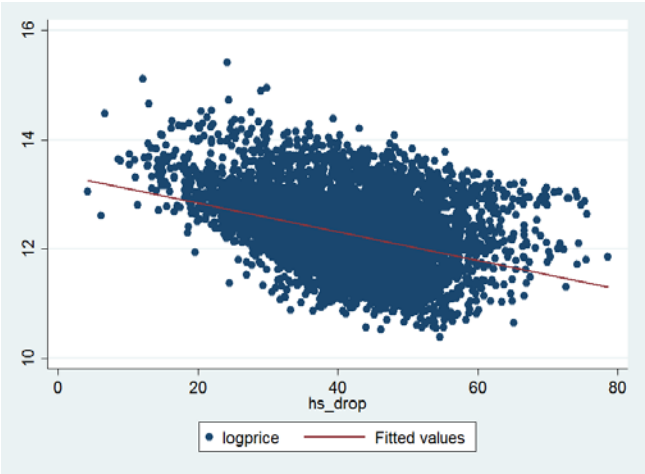
The results of the simple linear regression are summarized in the table below.

Table 3: Simple Regression

Dependent Variable:	log(Median Housing Price)
Per HS Dropouts	-.0262079 *** (-34.15)
Intercept	13.36691*** (403.50)
R^2	0.1396
# of observations	7190

The following is a scatter plot of the percentage of people age 25 and up with a high school diploma in the zip code and the log of the median housing price in the zip code.

Graphic 2: Price-HS Dropout Scatter Plot



This regression can be written as the equation $\hat{y}_i = 13.36691 - .0262079x_i$, where $\hat{y}_i$ is the expected logarithm median household price for the zip code given the independent variable, and x_i is the percent of people age 25 and up with a high school diploma in the zip code. Relevant analysis of the regression can be found in the following paragraphs.

According to the R^2 term, .1396 of the deviation from the mean can be explained by the correlation with percent of population that are high school dropouts. This is modest but significant, which makes sense because this variable will likely affect median household prices, but there are many other factors at play.

At 0 percent of people aged 25 and up with a high school diploma, expected median household price is $638,521.10. If other factors are held constant, an increase in 1% of people over the age of 25 without a high school diploma will lower expected median household price by 2.62%. This makes sense because a decrease in the level of education should decrease housing prices.

The t statistic for the x variable has a very high magnitude, and the p-value is basically 0. This means that the x variable is significant to an extremely high level of confidence (above 99%).

The results of the first multiple linear regression are summarized in the table below.

Table 4: First Multiple Regression Results

Dependent Variable:	log(Median Housing Price)
Per w/ HS Diploma	-.0278117*** (-25.82)
Per w/ Bach Degree	.0219789*** (16.01)
Per w/ Grad Degree	.0026078* (1.92)
Intercept	12.62123*** (255.03)
R^2	0.4494
# of observations	7190

This regression can be written as the equation $\hat{y}_i = 12.62123 - .0278117x_{1i} + .0219789x_{2i} + .0026078x_{3i}$, where $\hat{y}_i$ is the expected logarithm median household price for the zip code given the independent variables, x_{1i} is the percent of people age 25 and up with a high school diploma in the zip code, x_{2i} is the percent of people age 25 and up with a bachelor's degree in the zip code, and x_{3i} is the percent of people age 25 and up with a graduate degree in the zip code. Relevant analysis of the regression can be found in the following paragraphs.

According to the R^2 term, .4494 of the deviation from the mean can be explained by the correlation with the independent variables. This is larger than the simple linear regression model, which makes sense because we are factoring in other levels of education.

At a value of 0 for all independent variables, the expected median household price is $302,921.81. Holding other factors constant, an increase in 1% of people over the age of 25 with a high school diploma will lower expected median household price by 2.78%; an increase in 1% of people over the age of 25 with a bachelor's degree will increase expected median household price by 2.20%; and an increase in 1% of people over the age of 25 with a graduate degree will increase expected median household price by .26%.

It makes sense that the higher percentage of people with a bachelor's degree and graduate degree will cause median household price to increase because median income will likely be higher. It is interesting to note that an increase in the percentage of people in the zip code with a high school degree (but not a college education) actually lowers expected housing prices. This could signal that a simple high school education does not have the same effect on housing prices as a college education.

The t statistic for the x_1 and x_2 variables have very high magnitudes, and the p-values are basically 0. This means that these variables are significant to an extremely high level of confidence (above 99%). On the other hand, the t statistic for the x_3 variable has a moderate magnitude, and the p-value is .055. This means that this variable is significant but to a lesser level of confidence (94.5%).

The results of the second multiple linear regression are summarized in the table below.

Table 5: Second Multiple Regression Results

Dependent Variable:	log(Median Housing Price)
Per w/ HS Diploma	-.0135873*** (-12.43)
Per w/ Bach Degree	.0100981*** (7.65)
Per w/ Grad Degree	-.002533** (-2.09)
Median Income	.0000239*** (33.33)
Percent Asian Only	.016546 *** (15.35)
Percent White Only	-.0050365*** (-7.21)
Percent Black Only	-.0100239*** (-12.86)
Percent Hispanic Only	.0099633*** (23.15)
Intercept	11.84342*** (171.70)
R^2	0.6577
# of observations	7190

This regression can be written as the equation $\hat{y}_i = 11.84342 - .0135873x_{1i} + .0100981x_{2i} - .002533x_{3i} + 0000239x_{4i} + .0165464x_{5i} - .0050365x_{6i} - .0100239x_{7i} + .0099633x_{8i}$,where $\hat{y}_i$ is the expected logarithm of median household price for the zip code given the independent variables, x_{1i} is the percent of

people age 25 and up with a high school diploma in the zip code, x_{2i} is the percent of people age 25 and up with a bachelor's degree in the zip code, x_{3i} is the percent of people age 25 and up with a graduate degree in the zip code, x_{4i} is the median income of the zip code, x_{5i} is the percentage of the population that is asian only in the zip code, x_{6i} is the percentage that is white only, x_{7i} is the percentage that is black only, and x_{8i} is the percentage that is hispanic only. Relevant analysis of the regression can be found in the following paragraphs.

According to the R^2 term, .6577 of the deviation from the mean can be explained by the correlation with the independent variables. This makes sense because we have included many independent variables that are correlated with median household prices.

At a value of 0 for all independent variables, the expected median household price is $139,165.62. Holding other factors constant, an increase in 1% of people over the age of 25 with a high school diploma will lower expected median household price by 1.36%; an increase in 1% of people over the age of 25 with a bachelor's degree will increase expected median household price by 1.01%; an increase in 1% of people over the age of 25 with a graduate degree will lower expected median household price by .25%; an increase in median income by $1 will increase expected median household price by .002%; an increase in 1% percent of the population that are asian only in the zip code will increase expected median household price by 1.65%; an increase in 1% percent of the population that are white only in the zip code will decrease expected median household price by .50%; an increase in 1% percent of the population that are black only in the zip code will decrease expected median household price by 1.00%; an increase in 1% percent of the population that are hispanic only in the zip code will increase expected median household price by 1.00%.

The effect of education on housing prices is somewhat similar but to a lesser magnitude than our second model (the only difference is that percent of the population with a graduate degree is now slightly negatively correlated with housing prices). It is important to note that this is the increase holding median income constant. In our previous models that did not include median income, we have assumed that education affects housing prices through median income.

The magnitude and direction of our other coefficients in the data is somewhat unimportant since we are specifically looking at education. Although, it is important to note that these other variables play a significant part in determining housing prices because the R^2 value dramatically increased, and the coefficients of each of the education variables decreased in magnitude. This implies that our other variables are perhaps even better indicators than education in terms of determining housing prices, but this is not to say that education is not an important factor.

The t statistic for all variables besides x_3 have very high magnitudes, and the p-values are basically 0. This means that these variables are significant to an extremely high level of confidence (above 99%). On the other hand, the t statistic for the x_3 variable has a moderate magnitude, and the p-value is .037. This means that this variable is significant but to a lesser level of confidence (96.3%).

V. Extensions

It is important that we test to see if our measures of education have a "joint significance" in determine median household prices. In order to do that, we conduct an F-test with a restricted model that does not include our first three variables from our second multiple linear regression. Important information for our F-test is summarized in the table below:

Table 6: F-Test Values

R^2 of Unrestricted Model	.6577
R^2 of Restricted Model	.6175
Numerator Degrees of Freedom	3
Denominator Degrees of Freedom	7181
F-Statistic	281.267
p-value of F-Statistic	0.000

The F-Statistic for the education variables has an extremely high magnitude, and the p-value is basically 0, so we can conclude with a very high degree of certainty (over 99%) that the education variables are jointly significant.

In our analysis, we also looked at the different types of relationships our independent variables had with housing prices. Below is a scatterplot of median income with the log of the median housing price.

Graphic 3: Price-Median Income Scatter Plot

It is visually apparent that the relationship is not completely linear and that there is a somewhat parabolic shape. In order to investigate this further, we added median income squared to our third model.

Table 7: Different Functional Form

Dependent Variable:	log(Median Housing Price)
Per w/ HS Diploma	-.0146245*** (-13.59)
Per w/ Bach Degree	.0032439** (2.37)
Per w/ Grad Degree	.0020026 (1.63)
Median Income	.0000557*** (26.25)
Median Income Squared	-3.30e-10*** (-15.91)
Percent Asian Only	.0170979*** (16.12)
Percent White Only	-.0051163*** (-7.46)
Percent Black Only	-.0094277*** (-12.29)
Percent Hispanic Only	.0107894*** (25.32)

Intercept	11.24228*** (144.84)
R^2	0.6694
# of observations	7190

This regression can be written as the equation $\hat{y}_i = 11.24228 - .0146245x_{1i} + .0032439x_{2i} + .0020026x_{3i} + .0000557x_{4i} - 3.30e\text{-}10x_{5i} + .0170979x_{6i} - .0051163x_{7i} - 0094277x_{8i} + .0107894x_{9i}$, where $\hat{y}_i$ is the expected logarithm of median household price for the zip code given the independent variables, x_{1i} is the percent of people age 25 and up with a high school diploma in the zip code, x_{2i} is the percent of people age 25 and up with a bachelor's degree in the zip code, x_{3i} is the percent of people age 25 and up with a graduate degree in the zip code, x_{4i} is the median income of the zip code, x_{5i} is the median income of the zip code squared, x_{6i} is the percentage of the population that is asian only in the zip code, x_{7i} is the percentage that is white only, x_{8i} is the percentage that is black only, and x_{9i} is the percentage that is hispanic only. Relevant analysis of the regression can be found in the following paragraphs.

According to the R^2 term, .6694 of the deviation from the mean can be explained by the correlation with the independent variables. This makes sense because we have included many independent variables that are correlated with median household prices.

At a value of 0 for all independent variables, the expected median household price is $76,288.69. Holding other factors constant, an increase in 1% of people over the age of 25 with a high school diploma will lower expected median household price by 1.46%; an increase in 1% of people over the age of 25 with a bachelor's degree will increase expected median household price by .32%; an increase in 1% of people over the age of 25 with a graduate degree will increase expected median household price by .20%; an increase in median income by $1 will increase expected median household price by .005%; an increase in median income square by $1 will increase lower expected median household price by 3.3e-8%; an increase in 1% percent of the population that are asian only in the zip code will increase expected median household price by 1.71%; an increase in 1% percent of the population that are white only in the zip code will decrease expected median household price by .51%; an increase in 1% percent of the population that are black only in the zip code will decrease expected median household price by .94%; an increase in 1% percent of the population that are hispanic only in the zip code will increase expected median household price by 1.08%.

The main differences between the last model and this one are that the magnitude of the coefficient of the percent of population with a bachelor's degree is much smaller, and there is again a positive (albeit small) correlation of median household prices with percent of the population with a graduate degree. Most

of the other variables were barely changed by the addition of the median income squared variable. This could indicate that median income had a much larger play in the median household price than we initially thought.

The t statistic for all variables besides x_2 and x_3 have very high magnitudes, and the p-values are basically 0. This means that these variables are significant to an extremely high level of confidence (above 99%). On the other hand, the t statistic for the x_2 variable has a moderate magnitude, and the p-value is .018. This means that this variable is significant but to a lesser level of confidence (98.2%). Finally, the t statistic for the x_3 variable has a fairly low magnitude, and the p-value is .102. This means that this variable is significant only to a 89.8% confidence.

VI. Conclusion

If we revisit our original hypothesis that education leads to housing segregation through higher median household prices we get some interesting results. We can observe, as in our first regression model, that the amount of high school dropouts in a zip code is correlated with lower housing prices. This confirms our original idea that education creates residential clusters of wealth. However, as we discovered in our multiple regressions, an increase in percentage of the population over 25 with a high school diploma (holding other factors constant) actually decreased expected median household prices. This could be explained because a high school education is provided by the government for free. This leads us to conclude that the public education system does have an effect in reducing inequalities. Although the wealthy can send their kids to private schools, a public high school education puts students on a somewhat level playing field. However, an increase in the population with a bachelor's degree is associated with higher housing prices. A college education, which can lead to a higher income, is often exorbitantly expensive, and some are not able to afford it. This supports our hypothesis, because those who can afford to pursue higher levels of education are in turn "rewarded" with higher incomes and can afford to live in areas with higher median household prices.

Other variables are still important in creating a complete picture. We can see that race as an effect on where individuals live and who they live with, and higher median income, which is intertwined with education, also is correlated with a higher median housing price. But we have showed, as we have set out to, that education is an important piece of the puzzle.

Using this information, we can work towards UN Sustainable Development Goal #10, by increasing access to education. We have provided evidence to support that differences in education is one of the roots of inequality. By reducing differences in education, we may be able to begin further reducing inequalities.

VI. References and Appendix

References

Bialik, Kristen. (2017). *Americans deepest in poverty lost more ground in 2016. Pew Research Center.* Retrieved 16 October 2017, from http://www.pewresearch.org/fact-tank/2017/10/06/americans-deepest-in-poverty-lost-more-ground-in-2016/

Fryer, R., & Katz, L. (2013). Achieving Escape Velocity: Neighborhood and School Interventions to Reduce Persistent Inequality. *The American Economic Review, 103*(3), 232-237. Retrieved 22 November 2017, from http://www.jstor.org/stable/23469735

Guerrieri, V., Hartley, D., & Hurst, E. (2013). Endogenous Gentrification and Housing Price Dynamics. *Journal of Public Economics,* 45-60. Retrieved 22 November 2017, from http://www.sciencedirect.com/science/article/pii/S0047272713000297

Kearney, M., & Levine, P. (2016). Income Inequality, Social Mobility, and the Decision to Drop Out of High School. *Brookings Papers on Economic Activity*, 333-380. Retrieved 22 November 2017, from http://www.jstor.org/stable/43869027

STATA Output - Simple Regression Model:

```
. regress logprice hs_drop
```

Source	SS	df	MS	Number of obs	=	7,190
				F(1, 7188)	=	1165.98
Model	416.69379	1	416.69379	Prob > F	=	0.0000
Residual	2568.82082	7,188	.357376297	R-squared	=	0.1396
				Adj R-squared	=	0.1395
Total	2985.51461	7,189	.415289277	Root MSE	=	.59781

logprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hs_drop	-.0262079	.0007675	-34.15	0.000	-.0277125	-.0247034
_cons	13.36691	.0331272	403.50	0.000	13.30197	13.43184

STATA Output - Multiple Regression Model 1:

```
. regress logprice mo_per_hs mo_per_bach mo_per_grad
```

Source	SS	df	MS	Number of obs	=	7,190
				F(3, 7186)	=	1955.15
Model	1341.72075	3	447.240251	Prob > F	=	0.0000
Residual	1643.79386	7,186	.228749493	R-squared	=	0.4494
				Adj R-squared	=	0.4492
Total	2985.51461	7,189	.415289277	Root MSE	=	.47828

logprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mo_per_hs	-.0278117	.0010769	-25.82	0.000	-.0299228	-.0257006
mo_per_bach	.0219789	.0013725	16.01	0.000	.0192884	.0246694
mo_per_grad	.0026078	.0013561	1.92	0.055	-.0000506	.0052661
_cons	12.62123	.0494883	255.03	0.000	12.52422	12.71824

STATA Output - Multiple Regression Model 2:

```
. regress logprice mo_per_hs mo_per_bach mo_per_grad inc_tot w_per b_per a_per h_per
```

Source	SS	df	MS	Number of obs	=	7,190
				F(8, 7181)	=	1724.74
Model	1963.58469	8	245.448086	Prob > F	=	0.0000
Residual	1021.92992	7,181	.142310252	R-squared	=	0.6577
				Adj R-squared	=	0.6573
Total	2985.51461	7,189	.415289277	Root MSE	=	.37724

logprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mo_per_hs	-.0135873	.001093	-12.43	0.000	-.0157299	-.0114447
mo_per_bach	.0100981	.0013195	7.65	0.000	.0075114	.0126847
mo_per_grad	-.002533	.0012126	-2.09	0.037	-.0049101	-.0001559
inc_tot	.0000239	7.16e-07	33.33	0.000	.0000225	.0000253
w_per	-.0050365	.0006982	-7.21	0.000	-.0064052	-.0036679
b_per	-.0100239	.0007795	-12.86	0.000	-.0115519	-.0084959
a_per	.0165464	.0010783	15.35	0.000	.0144327	.0186602
h_per	.0099633	.0004303	23.15	0.000	.0091198	.0108068
_cons	11.84342	.0689769	171.70	0.000	11.70821	11.97864

STATA Output - F-Test Restricted Model:

```
. regress logprice inc_tot a_per w_per b_per h_per
```

Source	SS	df	MS	Number of obs	=	7,190
				F(5, 7184)	=	2319.31
Model	1843.48615	5	368.697231	Prob > F	=	0.0000
Residual	1142.02846	7,184	.158968326	R-squared	=	0.6175
				Adj R-squared	=	0.6172
Total	2985.51461	7,189	.415289277	Root MSE	=	.39871

logprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
inc_tot	.0000352	5.03e-07	70.01	0.000	.0000343	.0000362
a_per	.0172833	.0011219	15.41	0.000	.0150841	.0194825
w_per	-.0080054	.0007124	-11.24	0.000	-.0094019	-.0066088
b_per	-.0124252	.0008144	-15.26	0.000	-.0140216	-.0108288
h_per	.0113087	.0004187	27.01	0.000	.0104879	.0121295
_cons	11.38956	.0609526	186.86	0.000	11.27008	11.50905

STATA Output - Different Functional Form:

```
.  regress logprice mo_per_hs mo_per_bach mo_per_grad inc_tot inc_tot_squared w_per b_per a_per h_per
```

Source	SS	df	MS	Number of obs	=	7,190
				F(9, 7180)	=	1615.01
Model	1998.36523	9	222.040581	Prob > F	=	0.0000
Residual	987.149384	7,180	.137485987	R-squared	=	0.6694
				Adj R-squared	=	0.6689
Total	2985.51461	7,189	.415289277	Root MSE	=	.37079

logprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mo_per_hs	-.0146245	.0010763	-13.59	0.000	-.0167344	-.0125147
mo_per_bach	.0032439	.0013667	2.37	0.018	.0005648	.005923
mo_per_grad	.0020026	.0012255	1.63	0.102	-.0003998	.004405
inc_tot	.0000557	2.12e-06	26.25	0.000	.0000516	.0000599
inc_tot_squared	-3.30e-10	2.07e-11	-15.91	0.000	-3.70e-10	-2.89e-10
w_per	-.0051163	.0006863	-7.46	0.000	-.0064616	-.003771
b_per	-.0094277	.0007671	-12.29	0.000	-.0109314	-.0079241
a_per	.0170979	.0010604	16.12	0.000	.0150192	.0191766
h_per	.0107894	.0004261	25.32	0.000	.0099541	.0116247
_cons	11.24228	.0776209	144.84	0.000	11.09012	11.39444