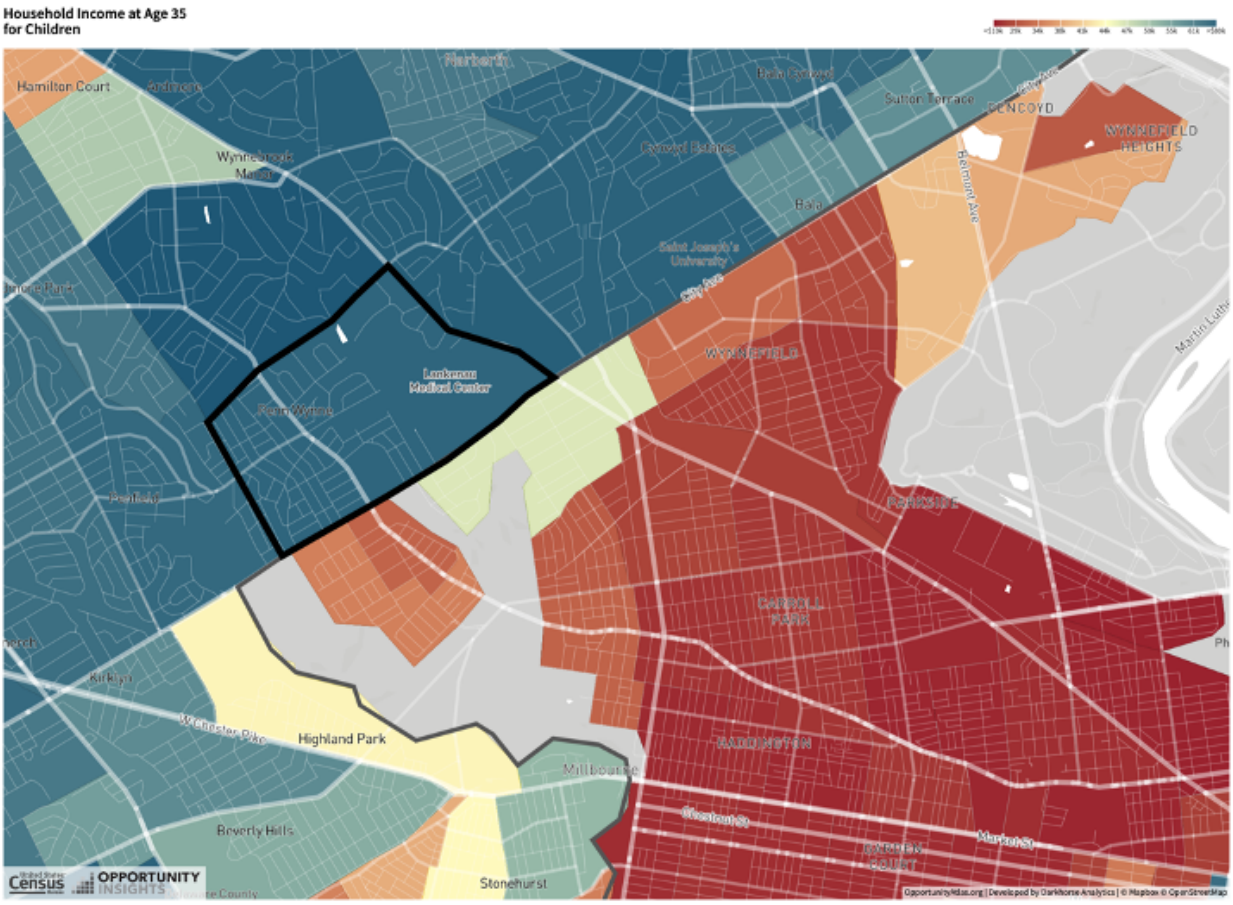
Stories from the Atlas: Describing Data using Maps, Regressions, and Correlations

Ben Warren

6/8/2022



The stark contrast that exists across the boundary between the city of Philadelphia and the surrounding suburbs on the Opportunity Atlas is evident. Within the city limits, every available tract is marked with a color from yellow to deep red, indicating a lower-than-median household income at the age of 35. The highest household income at age 35 for a tract in the city is 46k, in the Overbrook district (light yellow in Figure 1). Most other tracts on the Philadelphia side are between 20k and 30k, much lower than the national median. On the other hand, nearly every tract on the suburban side, particularly in Lower Merion (the author's more recent hometown), is colored in deep blue, indicating a household income at 35 much higher than the median. In Penn Wynne, the household income at 35 is 70k, a full $24k higher than the highest on the Philadelphia side.

One important statistic in determining the social and economic health of a neighborhood is the incarceration

rate of its residents. In the case of West Philadelphia and its surrounding suburbs, incarceration rates are, predictably, higher in Philadelphia, ranging from 1.3% in Overbrook to a stunning 8.9% in the Carroll Park neighborhood. Meanwhile, in my hometown of Penn Wynne, incarceration rate remains under 1%. However, unlike with household income at age 35, there are some neighborhoods in the suburbs that cross into the less desirable range of the scale (below-average for household income, above-average for incarceration rate). Neighborhoods in nearby Upper Darby reach 1.2% and 1.5%, while the Ardmore neighborhood of Lower Merion reaches 2%, on par with several of the Philadelphia neighborhoods. Based on this information, it seems important to not simply note the discrepancies in the data across the city line, but also within the suburban communities themselves. For example, two tracts of the Ardmore neighborhood, those with incarceration rates above 1%, reveal household incomes of 38k and 49k, just above and just below the median. In contrast, the one other tract listed under the town of Ardmore boasts a far above-average 61k median household income. Clearly, there are factors at play within the suburban township of Lower Merion that allow lower incomes and higher incarceration rates to exist within very specific areas.

Household income is the most prominent statistic in the Opportunity Atlas, and it appears to show a clear message about affluence across the boundary between Philadelphia and its suburbs: the city is poorer; the suburbs are wealthier. However, when several variables are considered, the message changes. When household income at 35 is considered just for Black people whose parents were 50^{th} percentile for income, the neighborhood of Penn Wynne, where Lower Merion borders Philadelphia, boasts worse outcomes than many West Philadelphia neighborhoods, with only a 25k median household income for those individuals. This remains the case for the children of parents in the 25^{th} percentile for income, who fare even worse in Penn Wynne. These individuals average a 19k household income at 35, again less than most of their West Philadelphia counterparts. When all Black individuals are considered, Penn Wynne is simply comparable to most of West Philadelphia. It holds true that a Black child growing up across the Philadelphia border in Penn Wynne is no better off than one growing up in many nearby West Philadelphia neighborhoods, and indeed, is likely to fair worse unless the income of their parents is far above the median. This presents several intriguing questions. First, is this the case for all children? Or just Black children? This is easily answered: white children from Penn Wynne whose parents were middle-income make, as a median, 62k, more than double their Black counterparts, and far more than any corresponding West Philadelphia neighborhood. This pattern repeats for children with low-income parents. **So, why does Penn Wynne act as a bastion of economic mobility for white children, but not for Black children?**

The data presented in the Opportunity Atlas tool are taken from the 2000 and 2010 censuses and the American Community Surveys from 2005-2015. The recent nature of these data is misleading however, since only the *outcomes* are measured from recent years. The birth cohort being studied is from 1978-1983, children who grew up across the decades of the 80s and 90s. The outcomes studied in more recent years are dependent upon the nature of these neighborhoods 25 or 30 years ago. Thus, it seems likely that in many neighborhoods there has been significant change in factors that impact outcomes in adulthood, such as the opportunity for education, poverty rates, job growth, and crime rates. Though these statistics are unlikely to have undergone complete transformations, there are certainly locales where they have changed a lot.

Measuring the delta of several important statistics for economic mobility would provide a clearer picture of the way neighborhoods have changed over time for the purposes of this research. The changes in median income, job growth, housing prices, access to education, and incarceration rates would provide a solid picture of this change. Perhaps an index could be created, taking the above measures into consideration, with weights assigned to the most important factors.

Household income ($) at age 31-37 for children with parents at the 25th percentile of the national income distribution (my home tract): $55,845.74

Household income ($) at age 31-37 for children with parents at the 25th percentile of the national income distribution (mean, all tracts): $34,443.48

There is clearly more opportunity for upward mobility in my home tract than in the average tract in the U.S. The adult income of children with parents at the same 25th percentile mark in my home tract is over $20,000/yr higher than that same metric for children from the average tract. This implies that the opportunity for economic mobility is quite high.

I will qualify this by saying that the mean household income in my tract for 2000 was $144,336, far higher than the average tract, with a 2010 poverty rate of 3.8%. It may be the case then, that this number represents a very small sample size, so it may be less telling than it seems.

In my home county, the standard deviation is 8736.658.

In my home state, the standard deviation is 8013.031. This is smaller than my county, but by less than 10%.

In the country, the standard deviation between tracts is 8169.155. This is smaller than in my county, but larger than in my state.

It seems that in Montgomery County, there is significant variability between the levels of economic mobility possible between tracts. This makes sense - some areas of the county are much more well off than others. For example, the standard deviation of the 2000 mean income in the county is $37,620.74, a significant amount. At the same time, the large standard deviation in both the state as a whole and the nation imply that there is a rift in social mobility across census tracts across the nation.

Household income ($) at age 31-37 for children with parents at the 75th percentile of the national income distribution (my home tract): $67,812.67

Household income ($) at age 31-37 for children with parents at the 75th percentile of the national income distribution (mean, all tracts): $51,500.78

Household income ($) at age 31-37 for children with parents at the 100th percentile of the national income distribution (my home tract): $78,572.56

Household income ($) at age 31-37 for children with parents at the 100th percentile of the national income distribution (mean, all tracts): $69,699.34

One point that sticks out immediately is that the adult household income for children whose parents were in the 25th percentile of income for my home tract is *greater* than the same metric for children who grew up in the 75th percentile in the rest of the country ($55,845 to $51,500). This is pretty remarkable. The economic mobility possible in my home tract appears unique based on this fact (though as I mentioned in my answer to #3, the sample size is small).

Otherwise, the household income statistics don't reveal too much more than they did for 25th percentile. The only distinction I see is that the difference between my tract and the nation at-large is greater for 25th percentile than 75th percentile.

25th percentile: $55,845.74/$34,443.48 = 1.6 times higher

75th percentile: $67,812.67/$51,500.78 = 1.3 times higher

100th percentile: $78,572.56/$69,699.34 = 1.12 times higher

There appears to be a smaller difference as you increase the childhood income percentile.

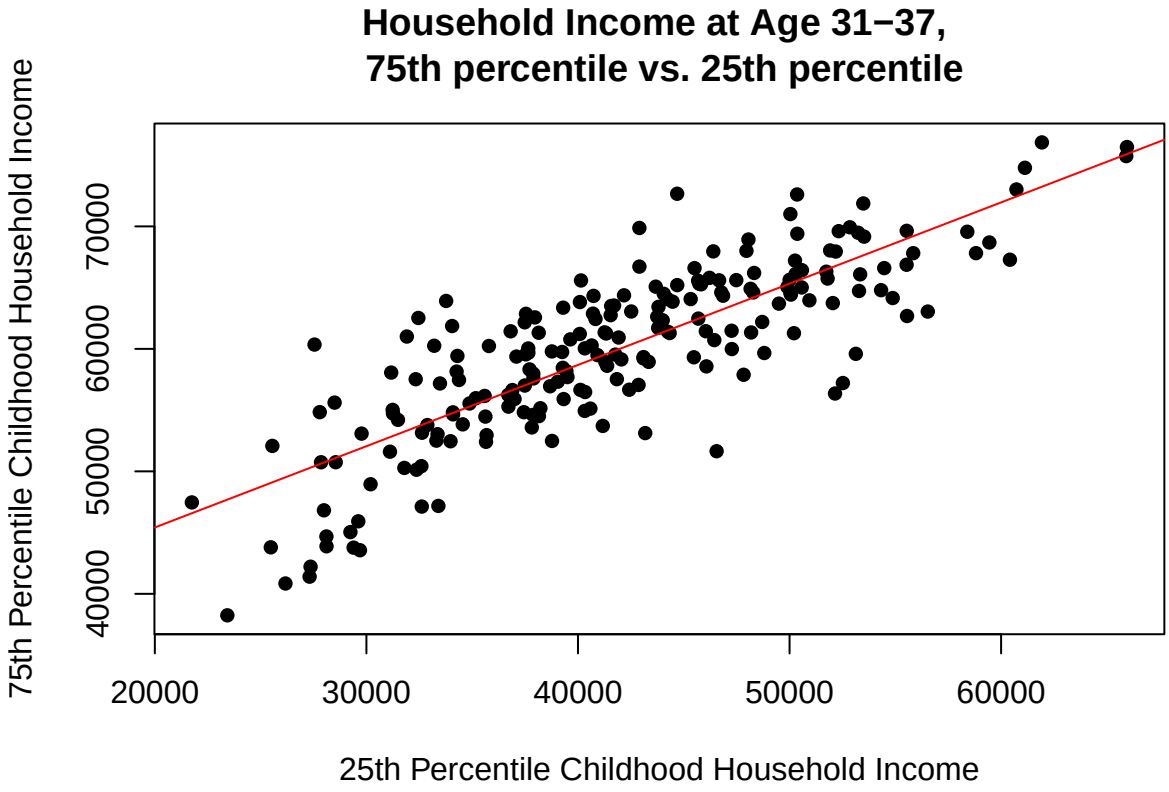
75th percentile:

At the 100th percentile the standard deviation follows the same pattern that it does at the 75th percentile, with the county level demonstrating far less deviation than the state and nation. It's reasonable to assume that in a larger geographic area, there will be more variation, however, at the national level there is less deviation than at the state level, so this impact may level off above a certain geographic area.

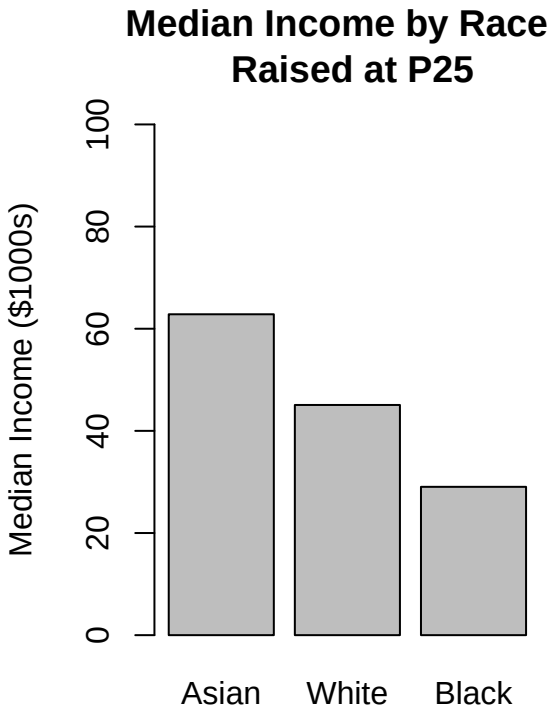
A linear regression shows that there is a clear correlation between strong outcomes for low-income (25th percentile) children and high-income (75th percentile) children, with an R2 of 0.6756 (so nearly 70% of variance is explained by the independent variable) for a p-value of effectively 0.

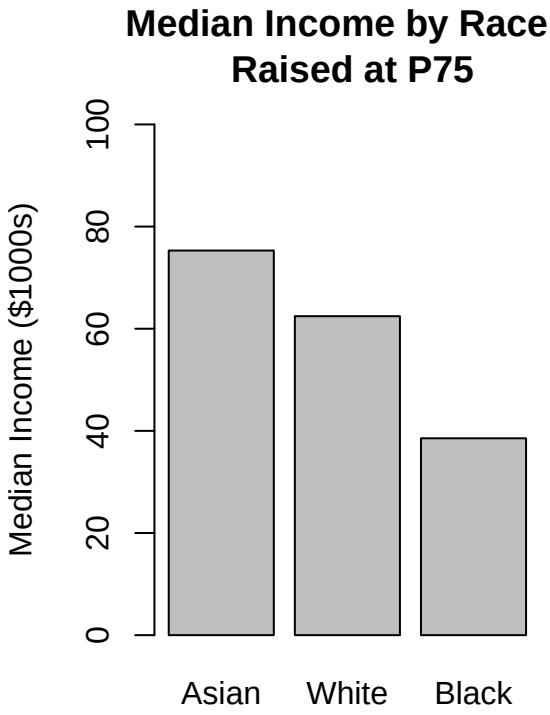
```
##
## Call:
## lm(formula = kfr_pooled_p75 ~ kfr_pooled_p25, data = home_county)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11390.3  -2350.4    151.7   2674.7  10866.1
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.215e+04  1.363e+03   23.59   <2e-16 ***
## kfr_pooled_p25 6.635e-01  3.187e-02   20.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4026 on 208 degrees of freedom
## Multiple R-squared:  0.6756, Adjusted R-squared:  0.6741
## F-statistic: 433.3 on 1 and 208 DF,  p-value: < 2.2e-16
```



Wealth and economic mobility in the US is a highly racialized phenomenon (for more on this, read The Upshot's analysis here). So, it is pertinent to extend this analysis along racial lines.

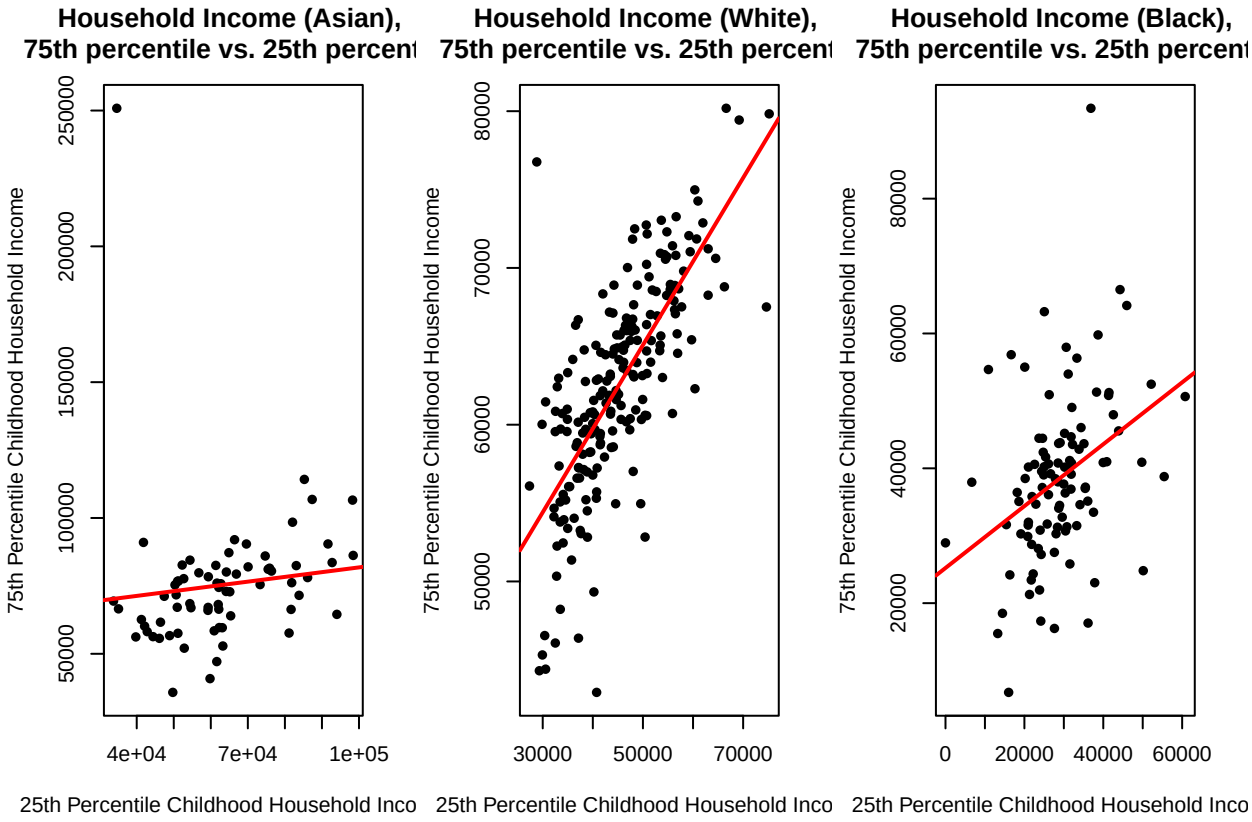




Overall, the mean Asian child raised in either the 25th or the 75th percentile in Montgomery County has a significantly higher median income than either a similar white or Black child. Black children have the lowest median income in comparison to their peers, with a median income under half of Asian children raised in similar economic situations.

When we run regressions for each of the major racial groups, the relationship between household income at the 25th and 75th percentile is the strongest among white families (with a p-value of almost 0 and an R-squared around 0.5) and weakest among Asian families (with a p-value of 0.37 and an R-squared of almost 0).

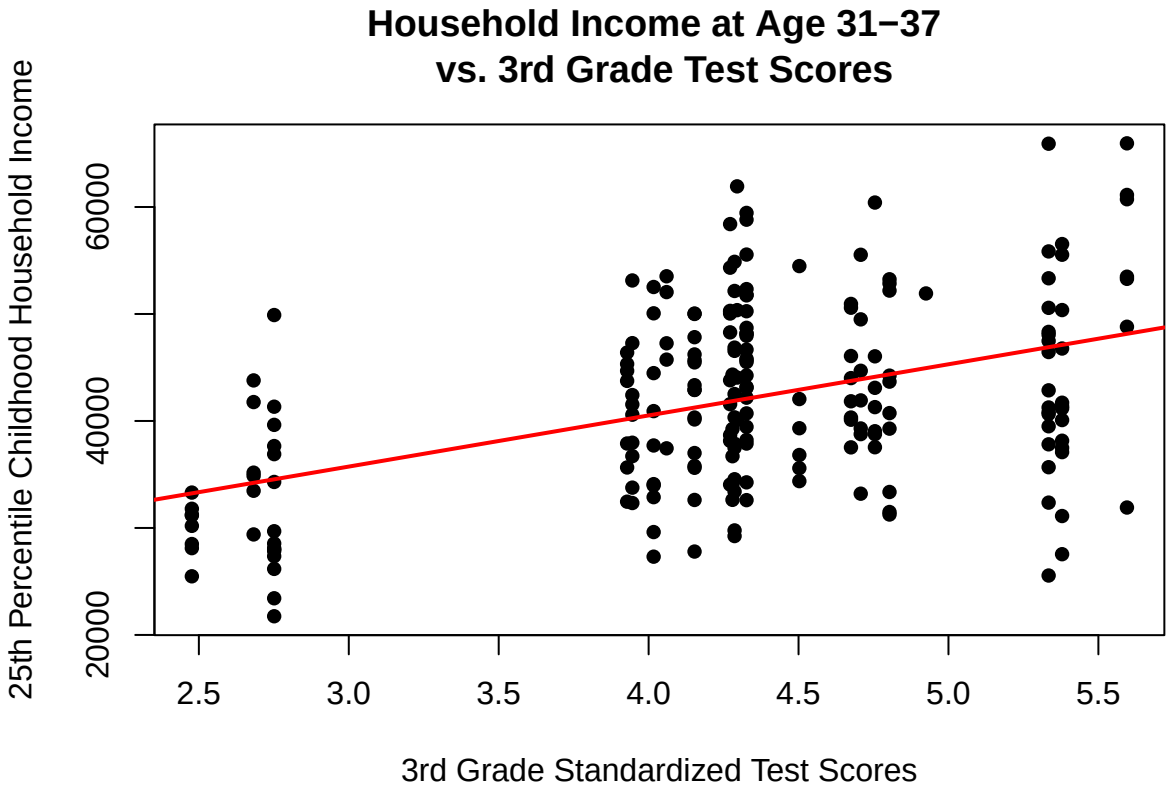
Black families lie in between. These data would suggest that in communities where white (and to a lesser extent Black) families succeed at a low socioeconomic status, they also succeed at a higher socioeconomic status, while this does *not* hold for Asian communities, though in Montgomery County, Asian families have the highest median incomes coming from both the first and third quartile of childhood income.



The relationship between household income at the 25th and 75th percentile is the strongest among white families (with a p-value of almost 0 and an R-squared around 0.5) and weakest among Asian families (with a p-value of 0.37 and an R-squared of almost 0). Black families lie in between. These data would suggest that in communities where white (and to a lesser extent Black) families succeed at a low socioeconomic status, they also succeed at a higher socioeconomic status, while this does *not* hold for Asian communities, though in Montgomery County, Asian families have the highest median incomes coming from both the first and third quartile of childhood income.

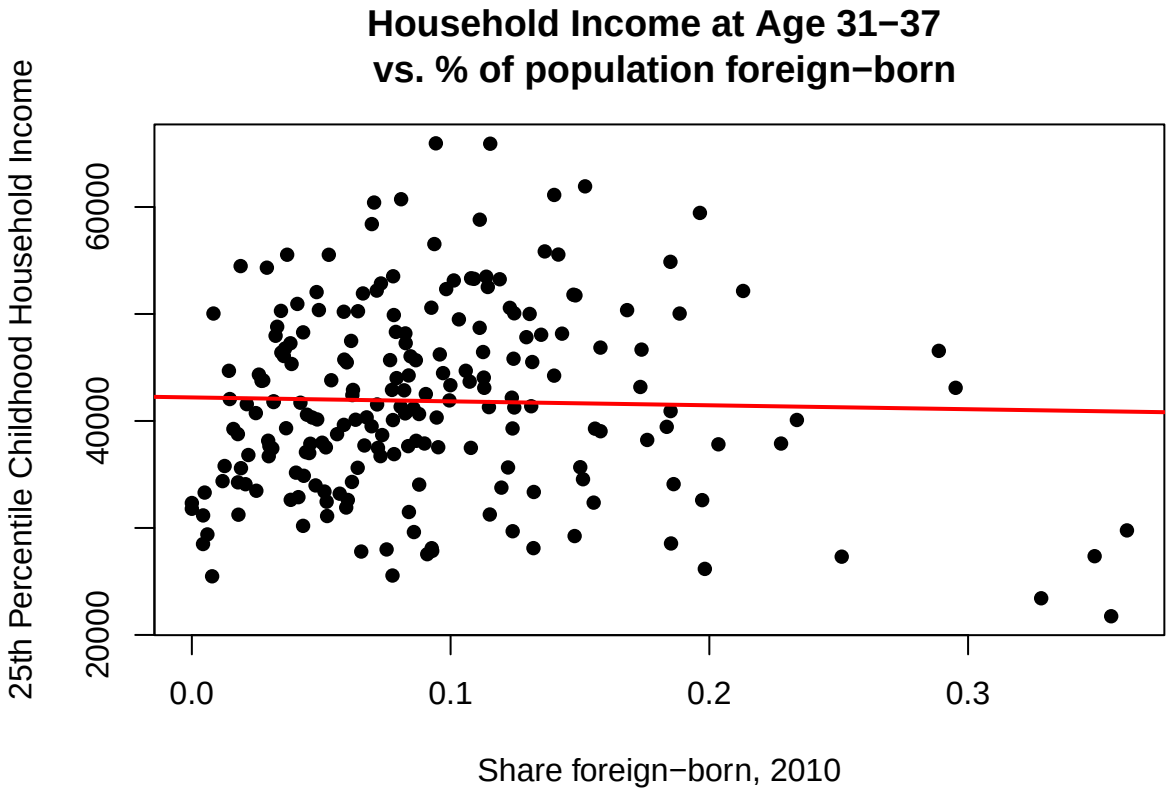
In a regression analysis, it is important to consider any covariates that might impact the results. I run through some of the possible covariates for this analysis here.

Standardized Test Scores in 3rd grade: Low correlation (0.19 R squared)



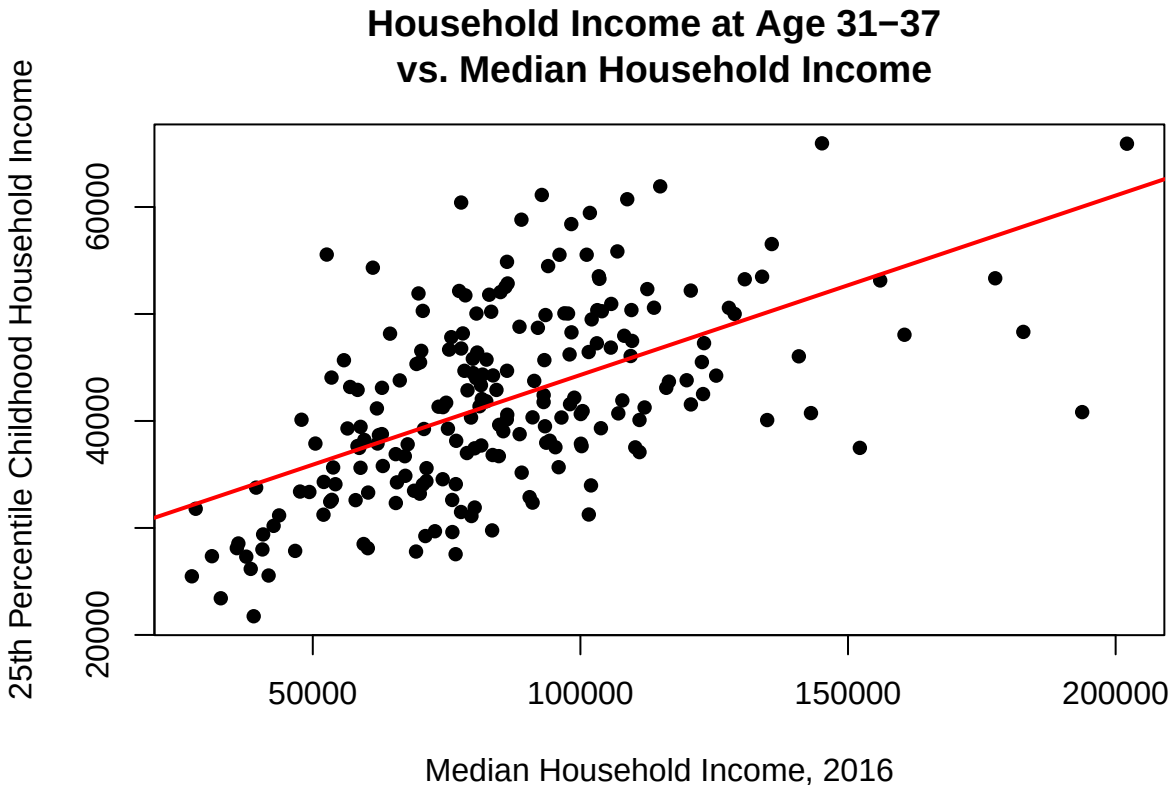

```
##
## Call:
## lm(formula = kfr_pooled_p25 ~ gsmn_math_g3_2013, data = home_county)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21342.6  -5735.8   -339.2   5883.5  19998.2
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          21393       2973   7.196 1.12e-11 ***
## gsmn_math_g3_2013     4781        682   7.011 3.28e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7887 on 207 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1919, Adjusted R-squared:  0.188
## F-statistic: 49.16 on 1 and 207 DF,  p-value: 3.284e-11
```

Share of population foreign-born: No correlation




```
##
## Call:
## lm(formula = kfr_pooled_p25 ~ foreign_share2010, data = home_county)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19147.6  -6357.2   -518.5   6207.9  24138.9
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          42199       1036   40.72   <2e-16 ***
## foreign_share2010    -3667       9174   -0.40     0.69
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8754 on 208 degrees of freedom
## Multiple R-squared:  0.0007676,  Adjusted R-squared:  -0.004036
## F-statistic: 0.1598 on 1 and 208 DF,  p-value: 0.6898
```

Median Household Income: strong correlation (0.32 R squared, miniscule p-value)

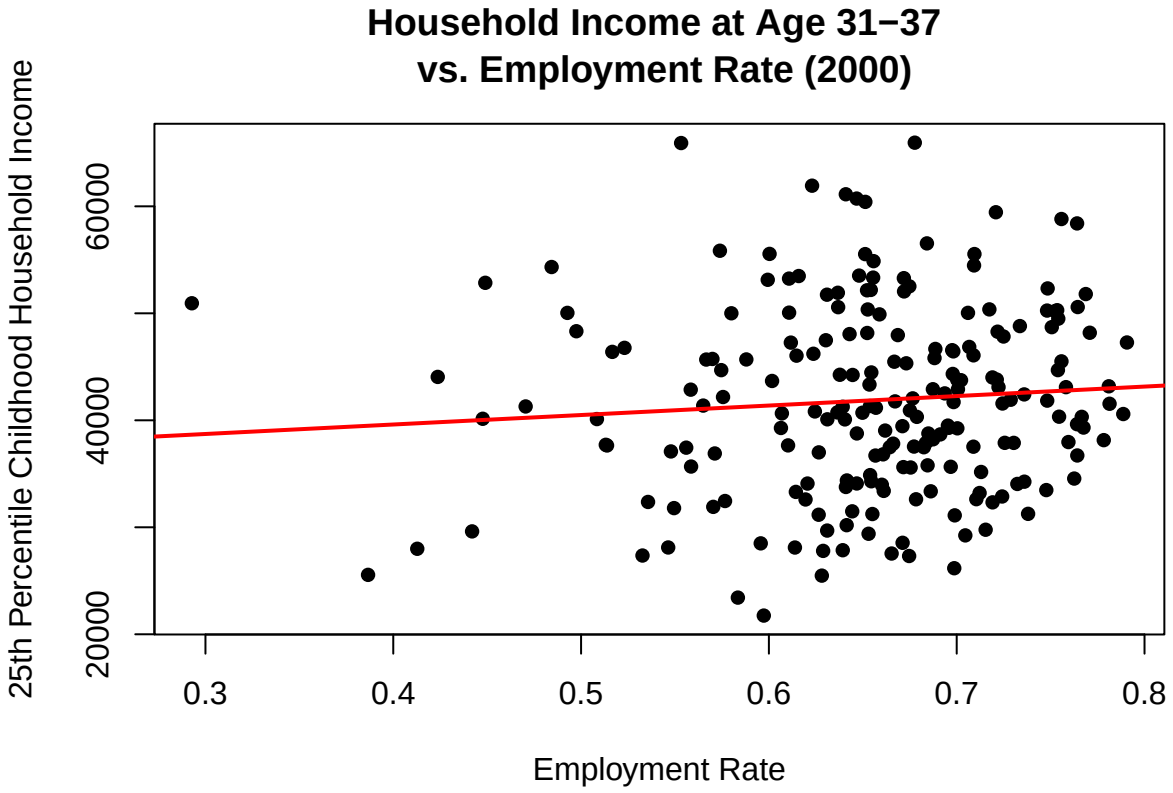



```
##
## Call:
## lm(formula = kfr_pooled_p25 ~ med_hhinc2016, data = home_county)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19194.1  -5086.4   -796.4   4877.5  19853.4
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.753e+04  1.531e+03  17.985   <2e-16 ***
## med_hhinc2016 1.677e-01  1.693e-02   9.903   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7220 on 208 degrees of freedom
## Multiple R-squared:  0.3204, Adjusted R-squared:  0.3172
## F-statistic: 98.07 on 1 and 208 DF,  p-value: < 2.2e-16
```

Both standardized test scores and overall median household income showed positive correlations with the median household income at P25. This implies that stronger schools and academic achievement, as well as being surrounded by overall stronger socioeconomic conditions might have a positive impact on economic mobility in Montgomery County.

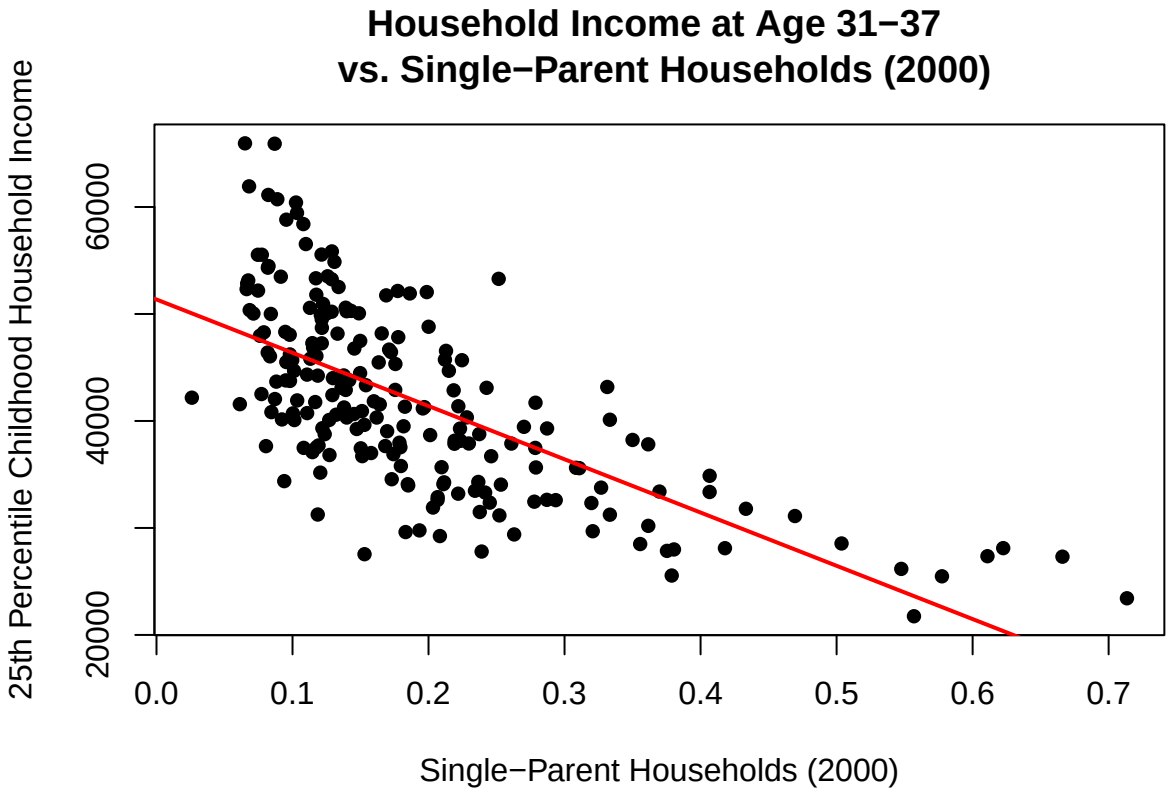
The largest predictors of economic mobility so far appear to be race, standardized test scores, and overall socioeconomic status of the census tract. Let's consider a few more factors:

Employment Rate: No significant correlation



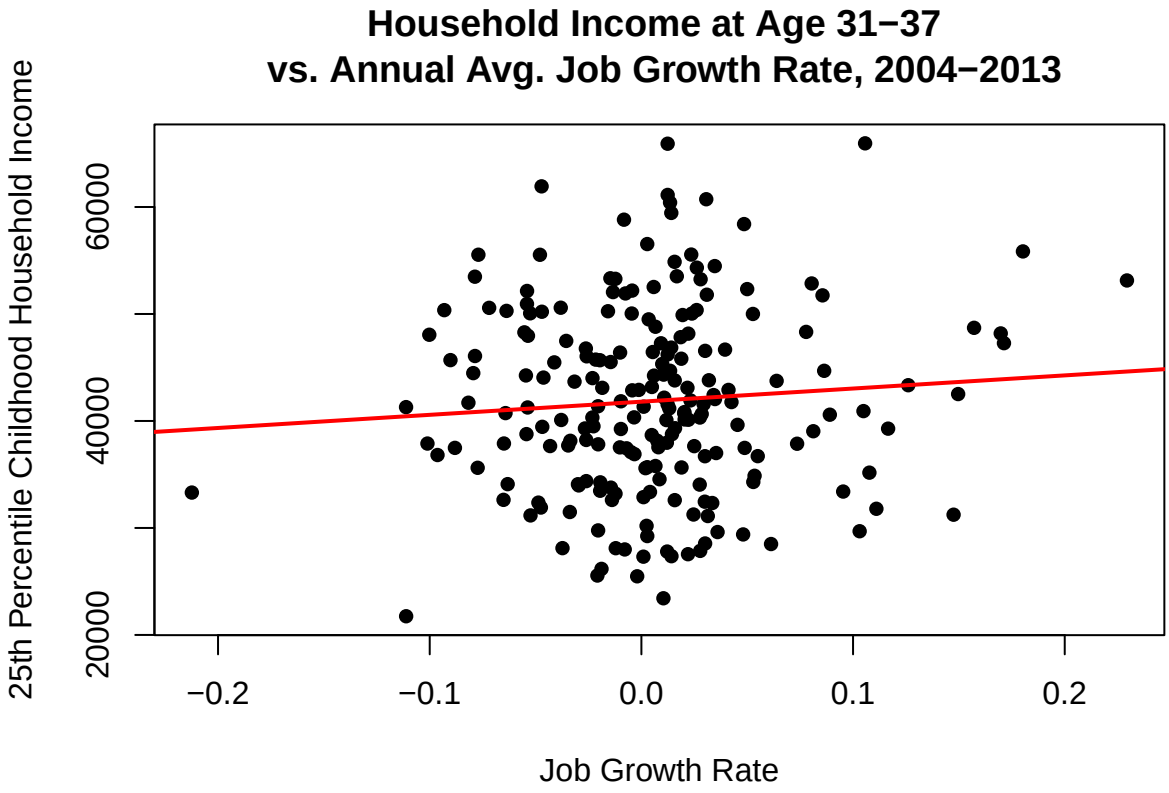

```
##
## Call:
## lm(formula = kfr_pooled_p25 ~ emp2000, data = home_county)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19606.9  -6442.9   -572.6   5939.7  24949.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    36063       4888   7.377 3.79e-12 ***
## emp2000         8861       7412   1.195    0.233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8728 on 208 degrees of freedom
## Multiple R-squared:  0.006824,   Adjusted R-squared:  0.002049
## F-statistic: 1.429 on 1 and 208 DF,  p-value: 0.2333
```

Share of single-parent households: Strong correlation (for each 1990, 2000, 2010 - but strongest for 2000)



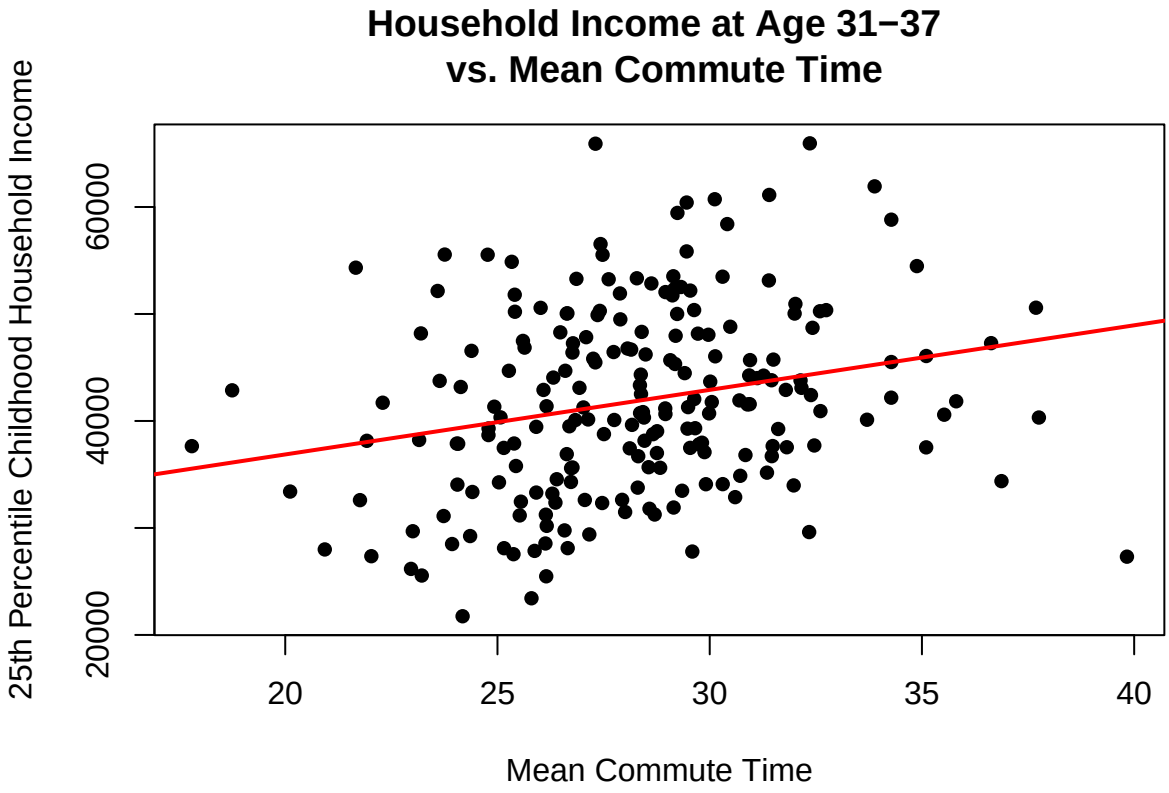

```
##
## Call:
## lm(formula = kfr_pooled_p25 ~ singleparent_share2000, data = home_county)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16192.5  -4925.6   -485.4   4528.2  18887.8
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              51352.6      846.4   60.67   <2e-16 ***
## singleparent_share2000 -49780.4     3773.7  -13.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6462 on 208 degrees of freedom
## Multiple R-squared:  0.4555, Adjusted R-squared:  0.4529
## F-statistic:   174 on 1 and 208 DF,  p-value: < 2.2e-16
```

Annual Avg. Job Growth Rate: No correlation



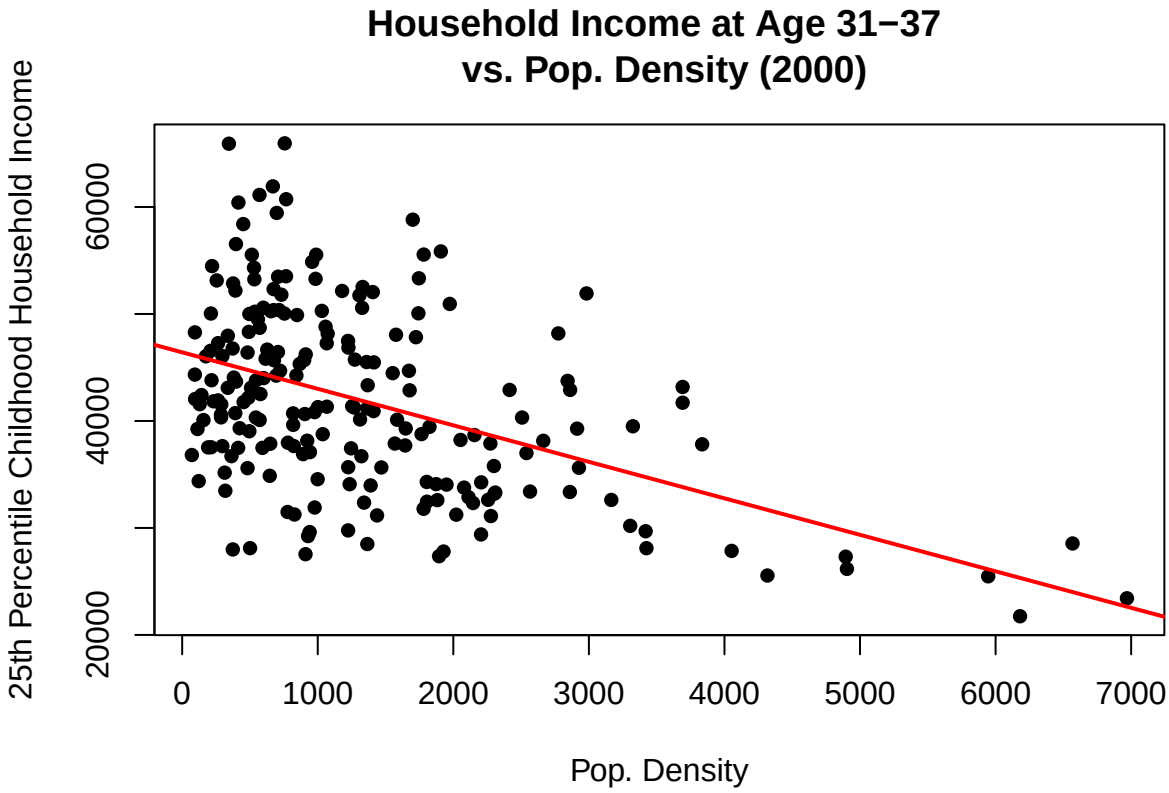

```
##
## Call:
## lm(formula = kfr_pooled_p25 ~ ann_avg_job_growth_2004_2013, data = home_county)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18687.6  -6140.9   -519.7   5742.6  23961.2
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    41801.6      604.7  69.127   <2e-16 ***
## ann_avg_job_growth_2004_2013  12287.1    10633.0   1.156    0.249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8730 on 208 degrees of freedom
## Multiple R-squared:  0.006379,   Adjusted R-squared:  0.001602
## F-statistic: 1.335 on 1 and 208 DF,  p-value: 0.2492
```

Commute time: Weak correlation



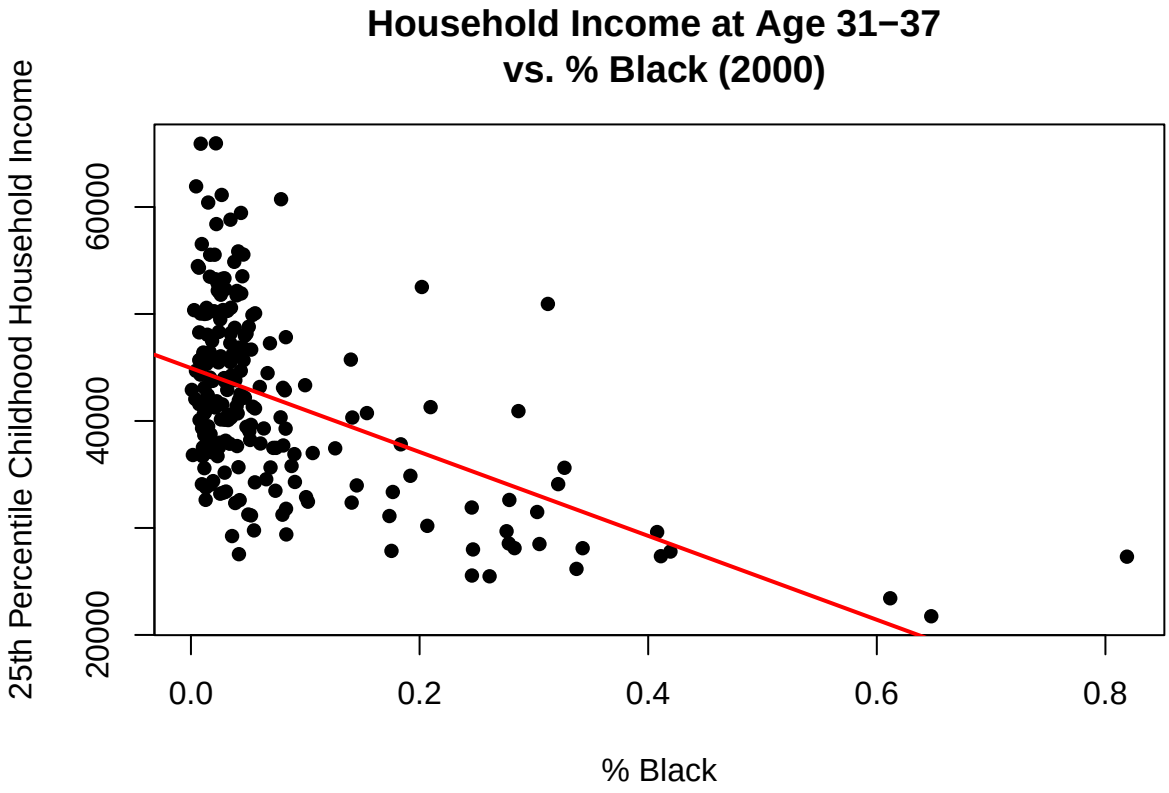

```
##
## Call:
## lm(formula = kfr_pooled_p25 ~ mean_commutetime2000, data = home_county)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21523  -6313  -1057   5793  24634
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           24798.5     4812.1   5.153 5.93e-07 ***
## mean_commutetime2000    603.6      168.9   3.573 0.000439 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8501 on 208 degrees of freedom
## Multiple R-squared:  0.05782,    Adjusted R-squared:  0.05329
## F-statistic: 12.76 on 1 and 208 DF,  p-value: 0.000439
```

Population Density: moderate correlation




```
##
## Call:
## lm(formula = kfr_pooled_p25 ~ popdensity2000, data = home_county)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17149.1  -5675.0   -860.7   5268.9  22112.6
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    46415.3966   800.9445  57.951  < 2e-16 ***
## popdensity2000    -3.4114     0.4472  -7.629 8.36e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7741 on 208 degrees of freedom
## Multiple R-squared:  0.2186, Adjusted R-squared:  0.2149
## F-statistic:  58.2 on 1 and 208 DF,  p-value: 8.356e-13
```

Share black: moderate correlation




```
##
## Call:
## lm(formula = kfr_pooled_p25 ~ share_black2000, data = home_county)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15749  -5532  -1146   5102  21863
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      44941.0      626.4  71.744  < 2e-16 ***
## share_black2000 -39204.5     4499.4  -8.713 9.32e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7496 on 208 degrees of freedom
## Multiple R-squared:  0.2674, Adjusted R-squared:  0.2639
## F-statistic: 75.92 on 1 and 208 DF,  p-value: 9.318e-16
```

Considering previously-analyzed strong factors (race, standardized test scores, and overall socioeconomic status of a tract) and new ones (share of single-parent households, population density, and commute time), a clearer picture begins to emerge of economic mobility in Montgomery County.

Children from low-income households are more likely to advance socioeconomically if their schools are high-achieving, their neighbors are well-off, families in their communities tend to

have more than 1 parent in a household, workers near them don't have to commute long distances, and their neighborhoods are not hugely dense in population. Finally, if their communities are less Black, or they themselves are Asian or white, they are more likely to be economically mobile.

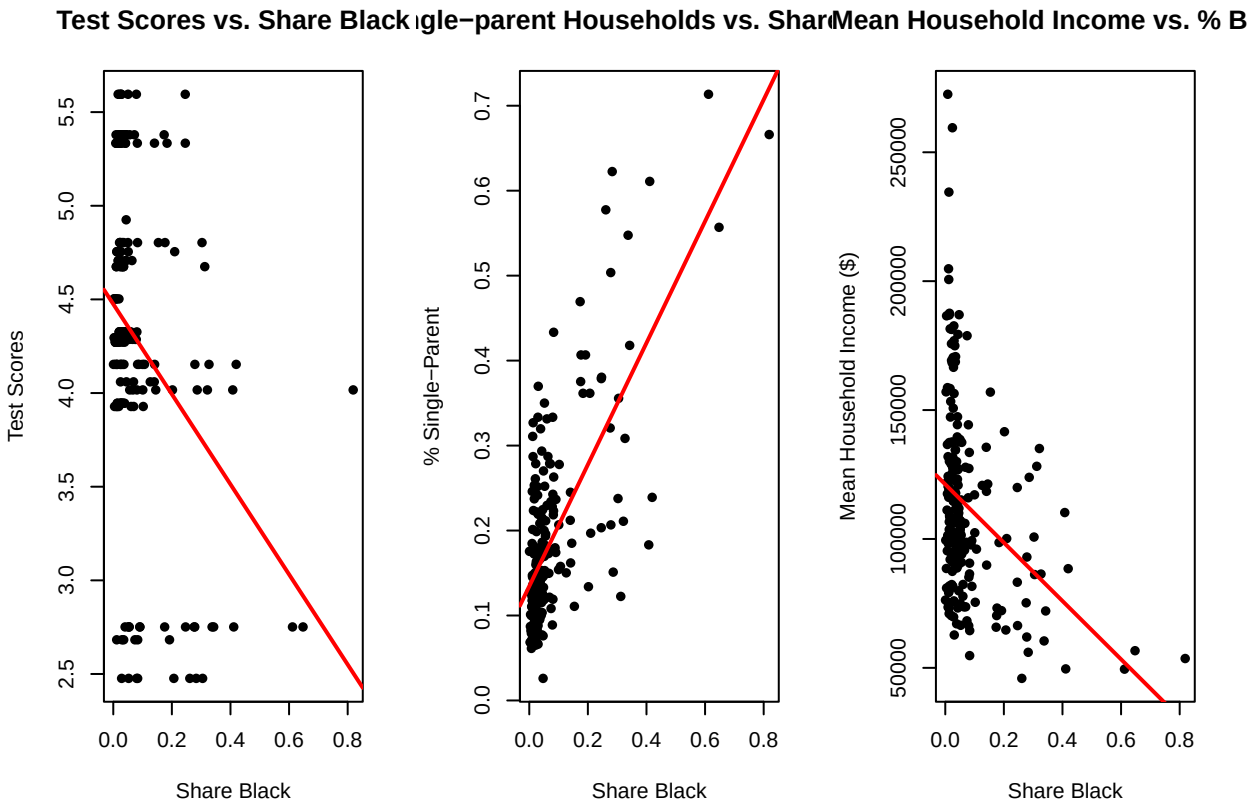
Interestingly, some factors that we might assume to have an impact, such as job growth and employment rate, do not. I would guess that this is because many people who live in Montgomery County may not work in the tract where they live. It is largely residential, with many residents working either in the major commercial areas in the county (typically not in the same tract where they live) or in the city of Philadelphia, which is not included in the county statistics. Either way, this obfuscates any kind of correlation between job statistics and economic mobility by tract.

It is concerning (but not surprising) to see that race is a major factor in economic mobility in Montgomery County. My intuition says that race may be correlated with some of the other factors that affect economic mobility. I check this assumption below:

Share black vs. standardized test scores, single-parent household, overall socioeconomic status:

```
##
## Call:
## lm(formula = gsmn_math_g3_2013 ~ share_black2000, data = home_county)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92974 -0.26407 -0.05638  0.38567  1.71116
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.47505    0.06322   70.79  < 2e-16 ***
## share_black2000 -2.40563    0.45300   -5.31 2.81e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7541 on 207 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1199, Adjusted R-squared:  0.1156
## F-statistic:  28.2 on 1 and 207 DF,  p-value: 2.812e-07
```

```
##
## Call:
## lm(formula = singleparent_share2000 ~ share_black2000, data = home_county)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24321 -0.05480 -0.01927  0.04627  0.28543
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.134431   0.007121   18.88   <2e-16 ***
## share_black2000 0.715800   0.051149   13.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08521 on 208 degrees of freedom
## Multiple R-squared:  0.4849, Adjusted R-squared:  0.4825
## F-statistic: 195.8 on 1 and 208 DF,  p-value: < 2.2e-16
```




```
##
## Call:
## lm(formula = hhinc_mean2000 ~ share_black2000, data = home_county)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -57017 -24854  -4987  16468 152284
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       121177       2956  40.990  < 2e-16 ***
## share_black2000  -113103      21234  -5.326 2.59e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35380 on 208 degrees of freedom
## Multiple R-squared:   0.12,  Adjusted R-squared:  0.1158
## F-statistic: 28.37 on 1 and 208 DF,  p-value: 2.592e-07
```

The results of the above regressions demonstrate that there is a significant correlation between the proportion of a tract's population that is Black and its mean test scores, proportion of single-parent households, and mean household income. Proportion of single-parent households in particular is strongly correlated to share of Black population, with an R-squared of almost 0.5.

This appears to explain some of the correlation between Black population and a lack of socioeconomic mobility. Tracts in Montgomery County where there are more Black families are disproportionately poor,

have lower test scores, and more single-parent households than other tracts. Each of these characteristics is correlated with worse economic mobility outcomes.