# HKS SUP-135 Project Part 1: Exploratory Data Analysis

Matt Khinda

2/25/2023

## Question 1: Exploring the Opportunity Atlas (Washington, DC)

This screenshot from the Opportunity Atlas website shows household income at age 35 for children of low-income parents in Washington, DC. The census tract where I grew up (tract 11001000600) is highlighted with a black outline.

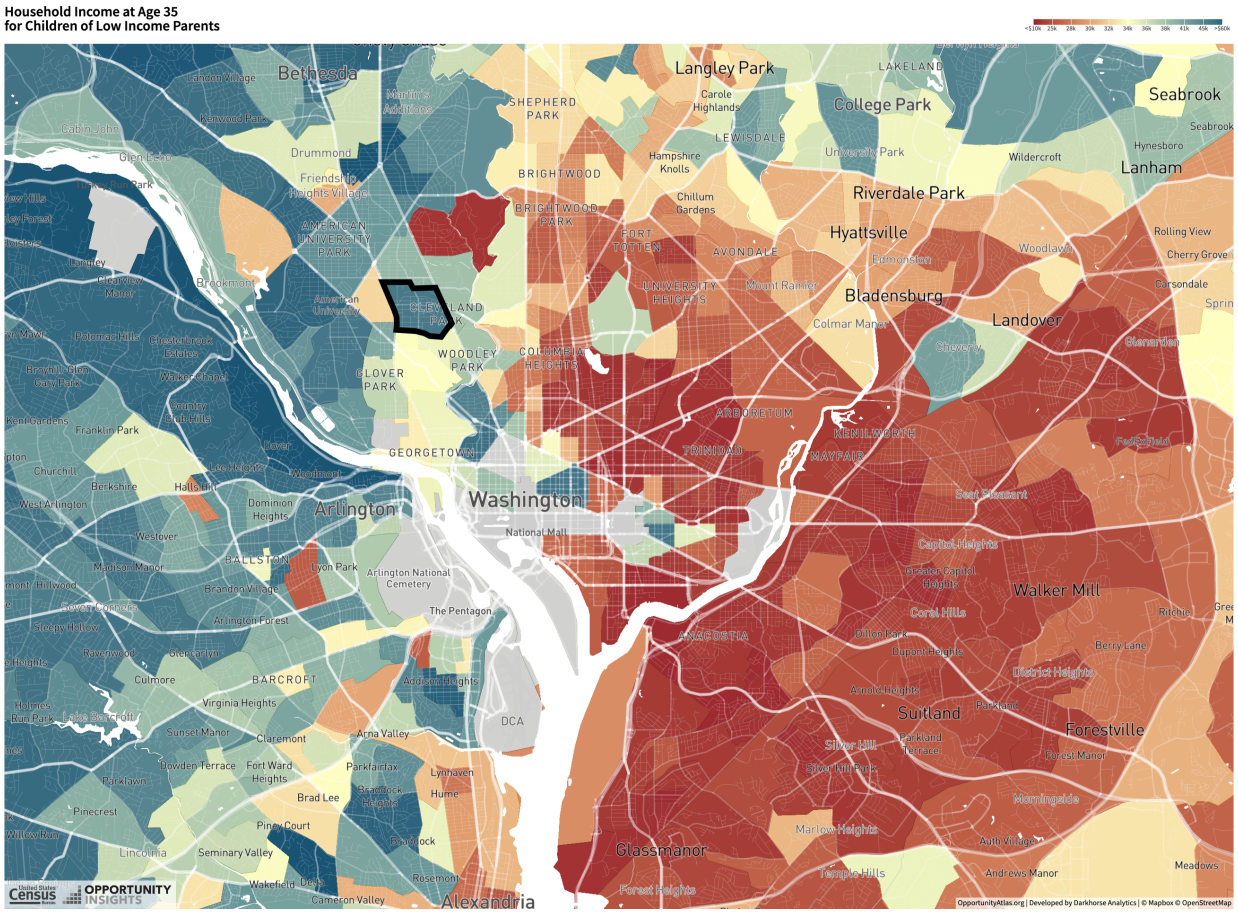



Figure 1: Screenshot of Washington, DC from the Opportunity Atlas (www.opportunityatlas.org)

## Question 2: Identifying Missing Data

There are 1745870 missing values in the Opportunity Atlas dataset.
They appear in the following variables:

| variable | missing_vals |
|---|---|
| state | 0 |
| county | 0 |
| tract | 0 |
| tract_name | 0 |
| cz | 0 |
| czname | 0 |
| kfr_pooled_pooled_p25 | 1189 |
| kfr_natam_pooled_p25 | 71464 |
| kfr_asian_pooled_p25 | 57759 |
| kfr_black_pooled_p25 | 39111 |
| kfr_hisp_pooled_p25 | 35581 |
| kfr_white_pooled_p25 | 5221 |
| kir_pooled_female_p25 | 1554 |
| kir_pooled_male_p25 | 1512 |
| kir_natam_female_p25 | 72350 |
| kir_asian_female_p25 | 65467 |
| kir_black_female_p25 | 47861 |
| kir_hisp_female_p25 | 47319 |
| kir_white_female_p25 | 7983 |
| kir_natam_male_p25 | 72347 |
| kir_asian_male_p25 | 65112 |
| kir_black_male_p25 | 48140 |
| kir_hisp_male_p25 | 47796 |
| kir_white_male_p25 | 7683 |
| jail_pooled_pooled_p25 | 1322 |
| jail_natam_pooled_p25 | 71790 |
| jail_asian_pooled_p25 | 59738 |
| jail_black_pooled_p25 | 42014 |
| jail_hisp_pooled_p25 | 38468 |
| jail_white_pooled_p25 | 5816 |
| jail_pooled_female_p25 | 1723 |
| jail_pooled_male_p25 | 1725 |
| jail_natam_female_p25 | 72505 |
| jail_asian_female_p25 | 66571 |
| jail_black_female_p25 | 49929 |
| jail_hisp_female_p25 | 49276 |
| jail_white_female_p25 | 8717 |
| jail_natam_male_p25 | 72572 |
| jail_asian_male_p25 | 66368 |
| jail_black_male_p25 | 51527 |
| jail_hisp_male_p25 | 50725 |
| jail_white_male_p25 | 8623 |
| HOLC_A | 63923 |
| HOLC_B | 63923 |
| HOLC_C | 63923 |
| HOLC_D | 63923 |
| pm25_1982 | 1296 |

| variable | missing_vals |
|---|---|
| pm25_1990 | 1296 |
| pm25_2000 | 1296 |
| pm25_2010 | 1296 |
| vegetation | 2531 |
| extreme_heat | 2531 |
| developed | 2531 |
| hhinc_mean2000 | 893 |
| mean_commutetime2000 | 882 |
| frac_coll_plus2000 | 852 |
| frac_coll_plus2010 | 202 |
| foreign_share2010 | 916 |
| med_hhinc1990 | 882 |
| med_hhinc2016 | 436 |
| popdensity2000 | 726 |
| poor_share2010 | 262 |
| poor_share2000 | 880 |
| poor_share1990 | 872 |
| share_white2010 | 84 |
| share_black2010 | 84 |
| share_hisp2010 | 84 |
| share_asian2010 | 1250 |
| share_black2000 | 827 |
| share_white2000 | 827 |
| share_hisp2000 | 827 |
| share_asian2000 | 2146 |
| gsmn_math_g3_2013 | 1109 |
| rent_twobed2015 | 16592 |
| singleparent_share2010 | 631 |
| singleparent_share2000 | 910 |
| singleparent_share1990 | 999 |
| traveltime15_2010 | 256 |
| emp2000 | 851 |
| mail_return_rate2010 | 652 |
| ln_wage_growth_hs_grad | 21563 |
| jobs_total_5mi_2015 | 888 |
| jobs_highpay_5mi_2015 | 888 |
| popdensity2010 | 5 |
| ann_avg_job_growth_2004_2013 | 2531 |
| job_density_2013 | 736 |

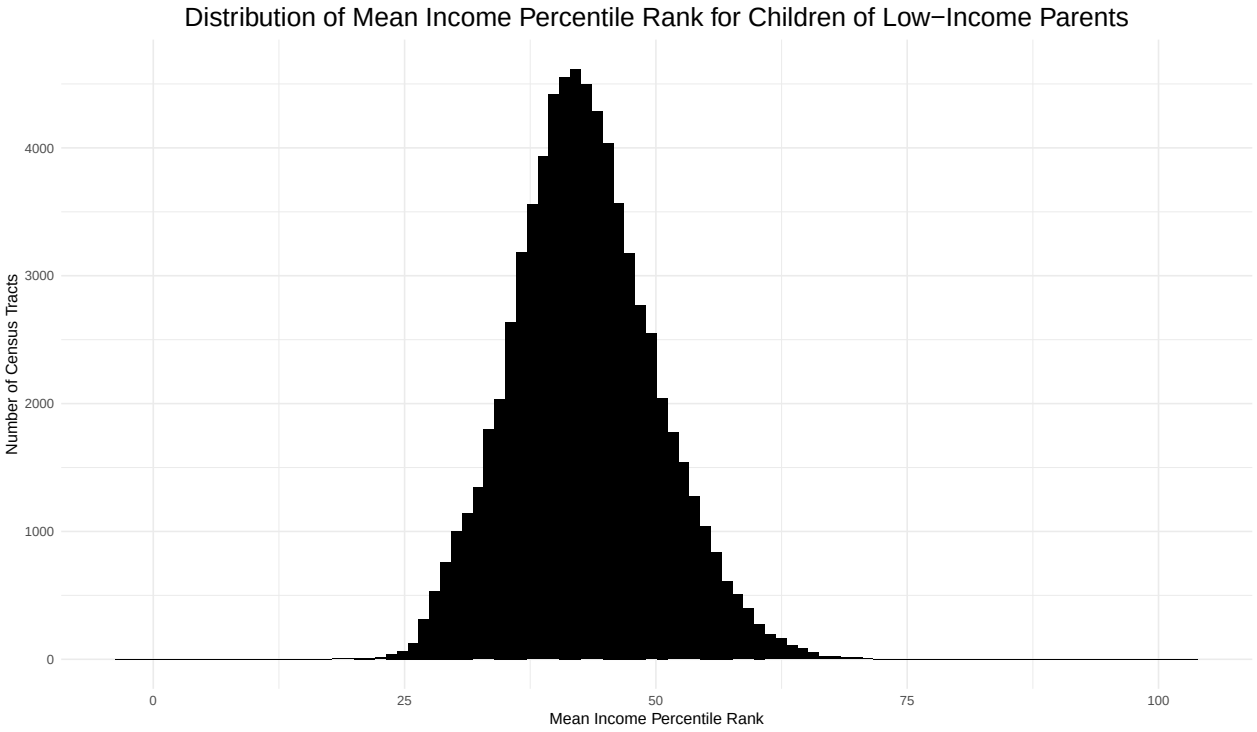
## Question 3: Absolute Mobility at the 25th Percentile

**3a: Units** In the Opportunity Atlas dataset, Absolute Mobility at the 25th percentile (kfr_pooled_pooled_p25) is measured as a percentile rank within the national household income distribution as measured in 2014-2015 tax data. In this sense, it does not measure an absolute amount of income in dollars but rather shows how the average income in that tract for people born to low-income parents compares to the rest of the country.

**3b: Interpreting values** In this variable, higher values signify greater economic outcomes and upward mobility. For example, a value of 99 signifies that the average income in that tract for someone born to parents at the 25th income percentile falls in the top 1% of incomes in the nation. Conversely, a value of 1 would indicate that the average income in that tract for that same group falls in the bottom 1%.

**3c: Rationale for using a linear model** A linear model is used to construct this statistic in order to account for the potentially small sample size of observations right at the 25th percentile in a given tract. The arithmetic mean of those limited observations may be higher or lower than the overall trend simply due to variance. Instead, the linear model is fit to the relationship between parents' and childrens' incomes at all percentiles, which provides greater accuracy when measuring the expected outcomes for a given child born to parents at the 25th income percentile in that particular tract.

## Question 4: Histogram of Absolute Mobility at the 25th Percentile




## Question 5: Summary Statistics for Absolute Mobility at the 25th Percentile

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. Std. Dev     NA's
##   -3.286   38.070   42.520   42.858   47.350  103.349    7.126     1189
```

## Question 6: Extreme Values

One of the limitations of a linear model is that it establishes a uniform and continous relationship between the predictor variable and the response variable. As such, it does not account for or respect particular upper or lower bounds — in this case the 100th and 0th percentile respectively. So, there may be cases where a certain input returns a value above the 100th or below the 0th percentile, as we see in the summary statistics for the kfr_pooled_pooled_p25 variable above.

## Question 7: Home Tract Comparison of Absolute Mobility at the 25th Percentile

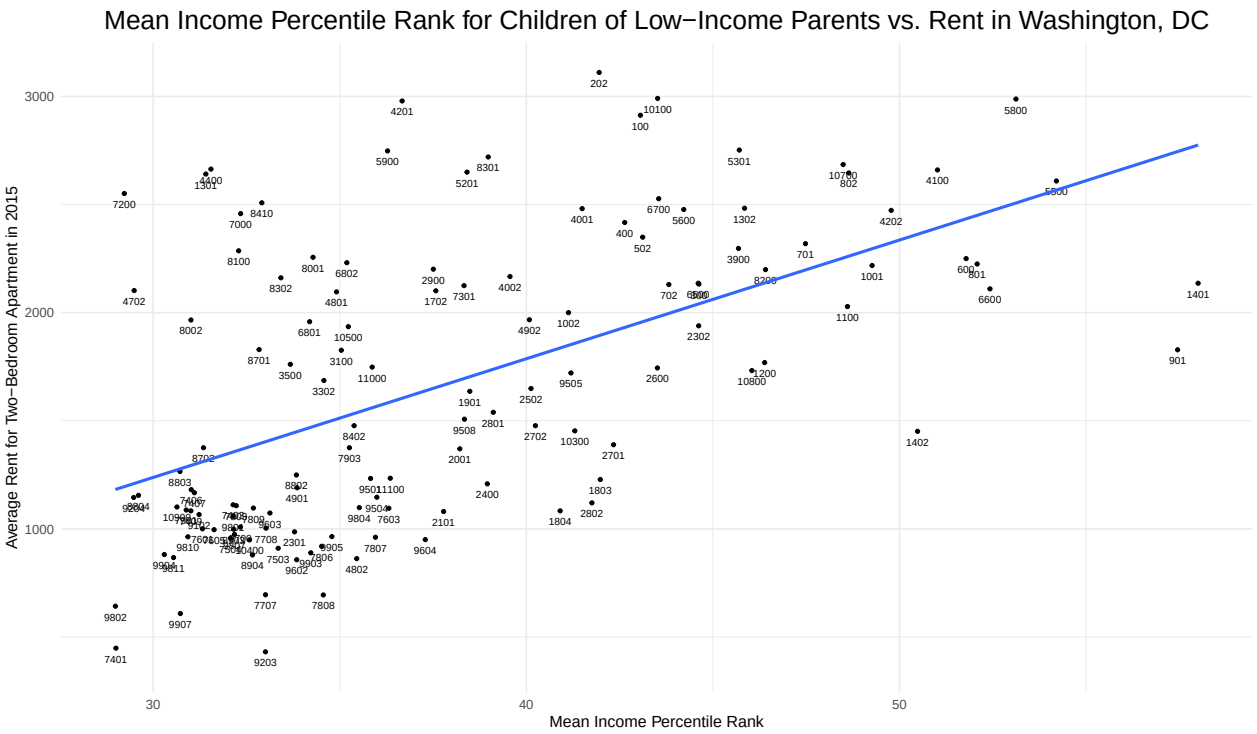
In my home tract, the average income percentile rank reached by someone born to parents at the 25th percentile of income is 51.78702, which is higher than both the Washington, DC average of 37.31235 and national average of 42.85813.

## Question 8: Home County Standard Deviation for Absolute Mobility at the 25th Percentile

The standard deviation of kfr_pooled_pooled_p25 in my home county is 6.406956, which is also the standard deviation for the state because the District of Columbia only contains one county. This is less than the standard deviation of kfr_pooled_pooled_p25 in the national dataset of 7.126422. From this measure we can conclude that the distribution of outcomes is more narrowly concentrated in DC than in the United States as a whole.

## Question 9: Upward Mobility & Rent

### 9a: Plotting County-Level Absolute Mobility at the 25th Percentile vs Rent

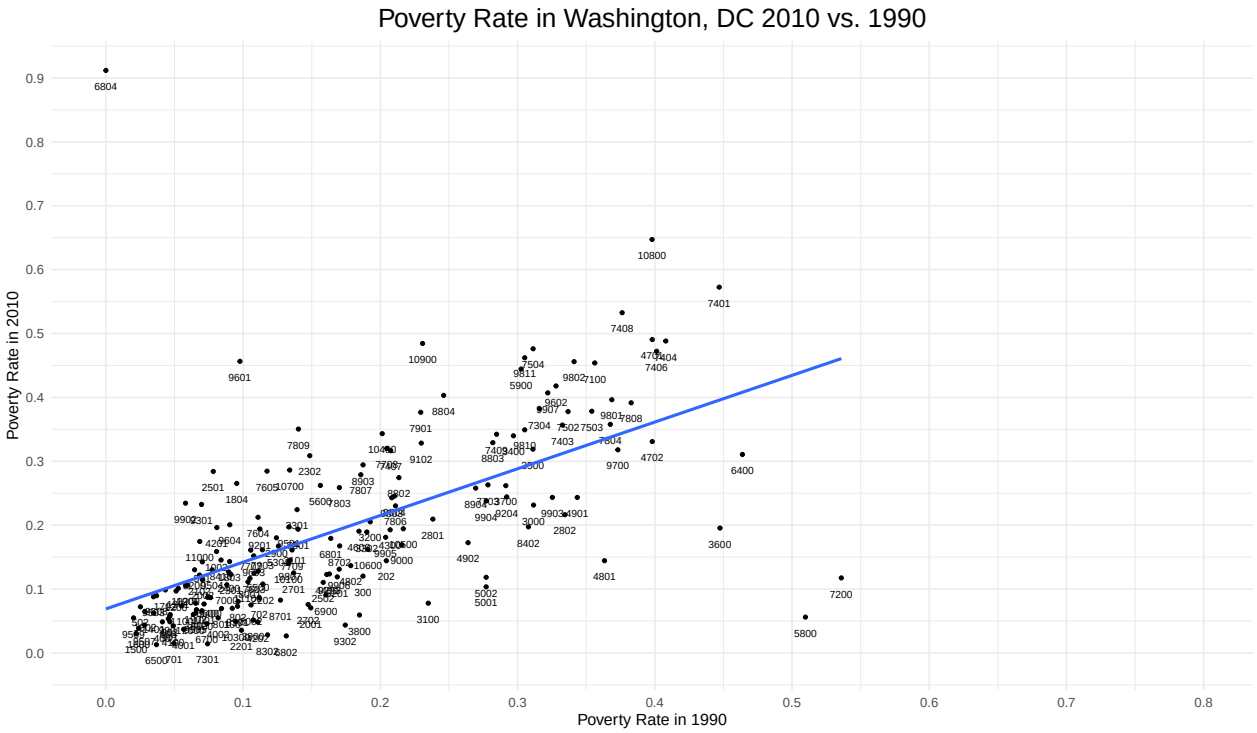


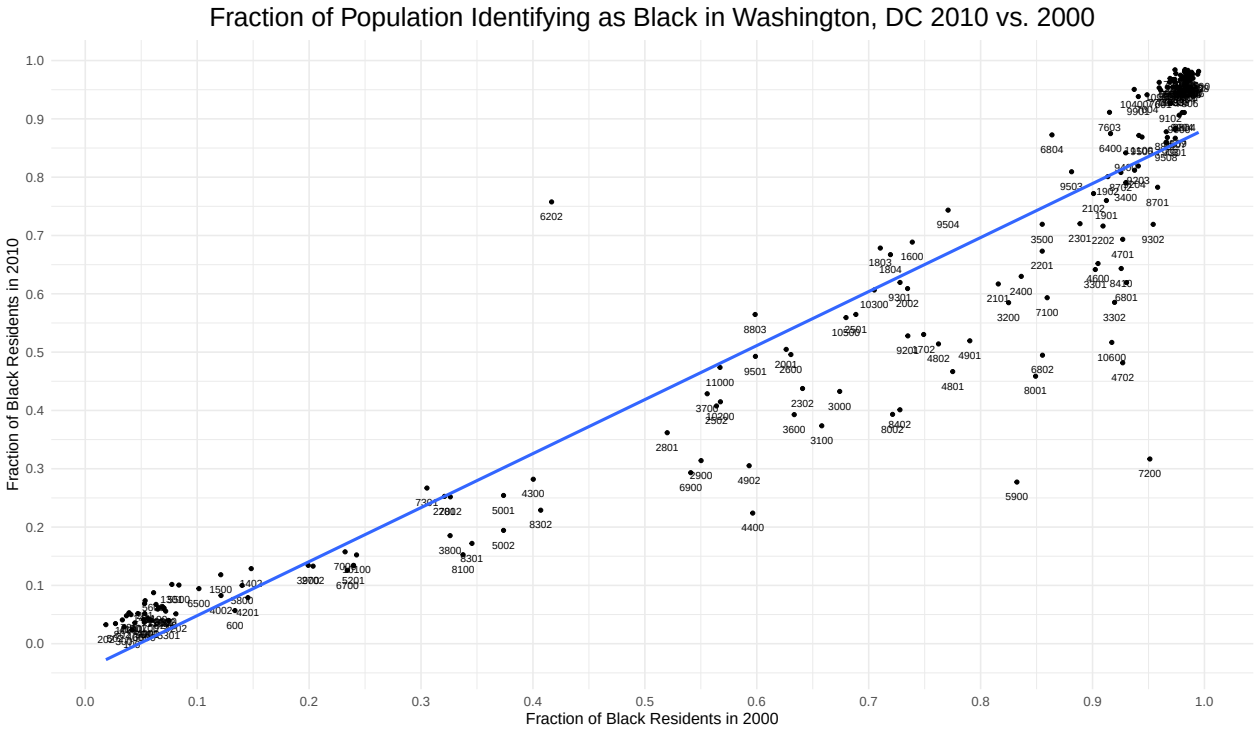

**9b: Reflections** The line of best fit in the plot above shows that there is a positive correlation between rent and mean income percentile rank, meaning that rent is typically higher in tracts with better economic outcomes for children of low-income parents. However, it is apparent that there is quite a lot of variance at the individual tract level. There are some tracts (tract 202) where rent is significantly higher than predicted based on the observed economic outcomes, while other tracts (tract 1402) cost significantly less than expected for their economic outcomes.

**9c: Identifying "Opportunity Bargains"** In order to determine if a particular tract is an "opportunity bargain" we can compare the rent predicted by the linear model and the reported rent for that tract. If reported rent is below predicted rent then we can say the tract is in fact an "opportunity bargain." My home tract (600) is an "opportunity bargain" with the predicted average rent for a two bedroom apartment being \$2433.88 and the measured average rent being \$2250 or a bargain of \$183.88. Some other tracts that are "opportunity bargains" as shown in the plot above are tract 1420, tract 1401, and tract 901 — all with high economic opportunity and comparatively low rents.

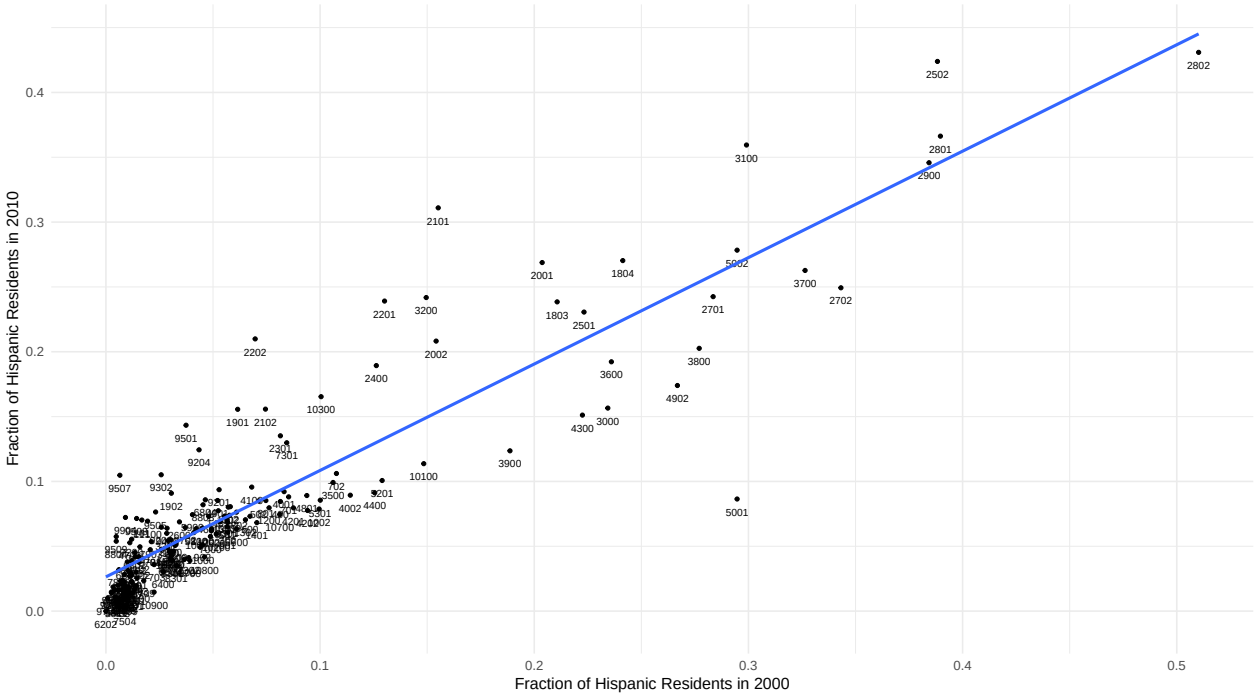
## Question 10: Change Over the Past 20 Years

### 10a: Comparison Scatter Plots

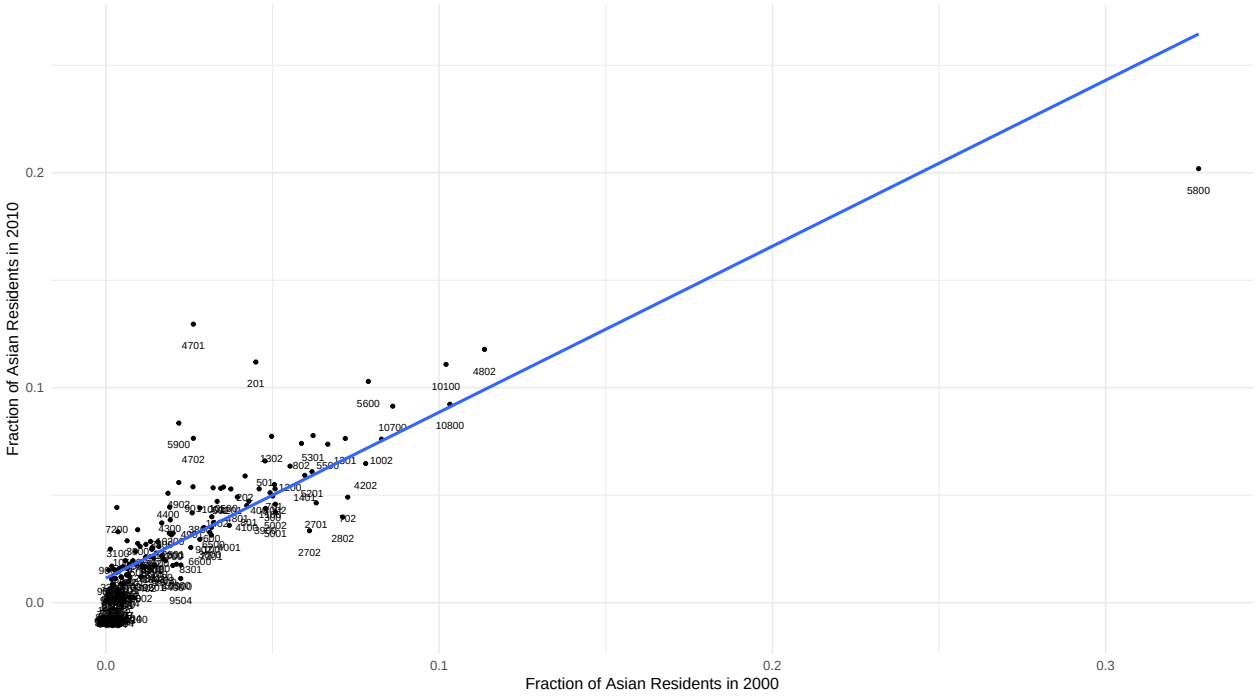


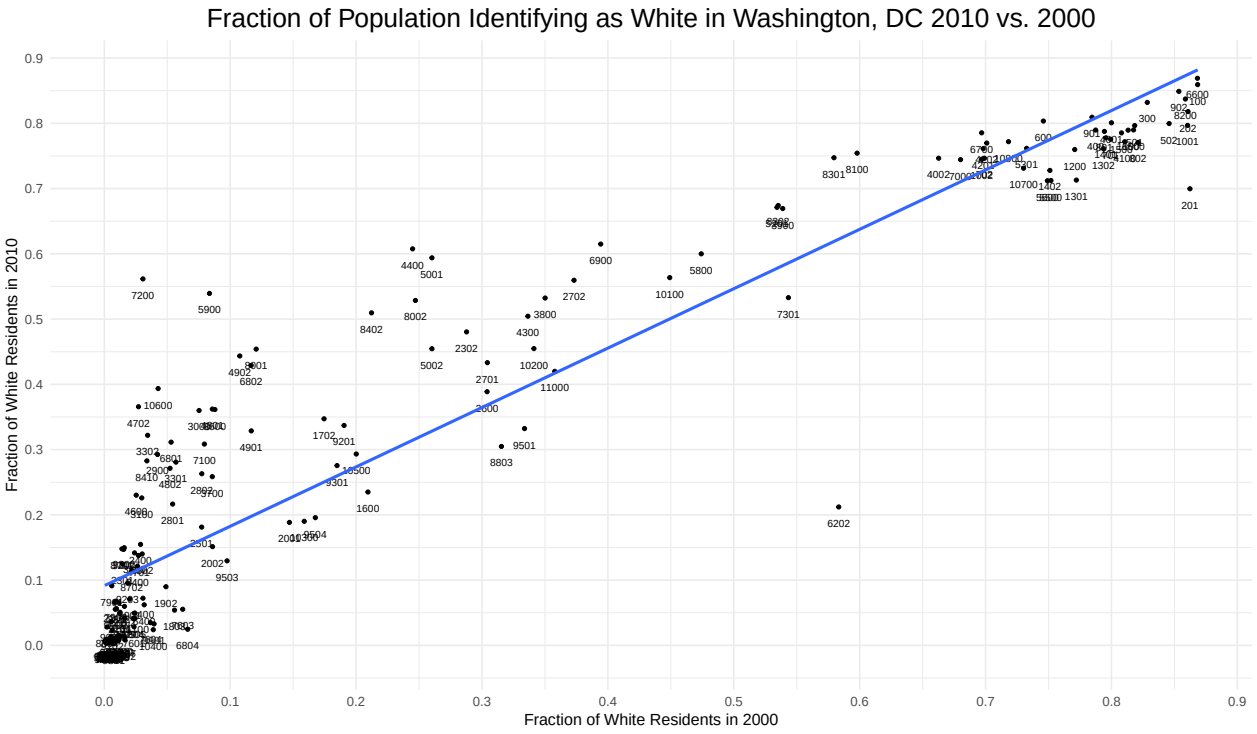



Fraction of Population Identifying as Hispanic in Washington, DC 2010 vs. 2000




Fraction of Population Identifying as Asian in Washington, DC 2010 vs. 2000






**10b: Describing Change Over Time** In the scatter plots above, a number of interesting economic and demographic trends are apparent. First, we notice in the comparison of poverty that tracts with a poverty rate of 0.25 or less in 1990 seem to have increased that share by 2010 on average, while those with a poverty rate higher than 0.25 in 1990 have decreased their share of the population living below the poverty line on average. Demographically, nearly all tracts have decreased their fraction of Black residents between 2000-2010 except for a cluster of tracts with a nearly entirely Black population which speaks to the intense and deepening racial segregation of Washington, DC. A parallel trend can be seen in the change in fraction of White residents between 2000-2010 which has largely grown on average, which may allude to gentrification or homogenization of once racially diverse or more significantly minority neighborhoods.

## Question 11: Redlining

**11a: HOLC Grades and Upward Mobility** There is a clear correlation between a higher HOLC grade and greater economic mobility in a given census tract in the United States. Tracts with a majority of the area graded A had a mean kfr_pooled_pooled_p25 value of 44.0140, those majority graded B had a mean value of 42.4677, those majority graded C had a mean value of 39.8727, and those majority graded D had a mean value of 36.1583.

**11b: HOLC Grades and Racial Composition of Tracts** The share of Black residents in a census tract is correlated with HOLC grades. Tracts with a majority of the area graded A had a mean share_black2000 value of 0.2008, those majority graded B had a mean value of 0.2886, those majority graded C had a mean value of 0.3334, and those majority graded D had a mean value of 0.4659. It is fair to say then that this could be a confounding variable in determining the effect of HOLC grades on economic outcomes since we also know that economic outcomes in tracts that are highly segregated or have a high share of Black residents tend to be below average.

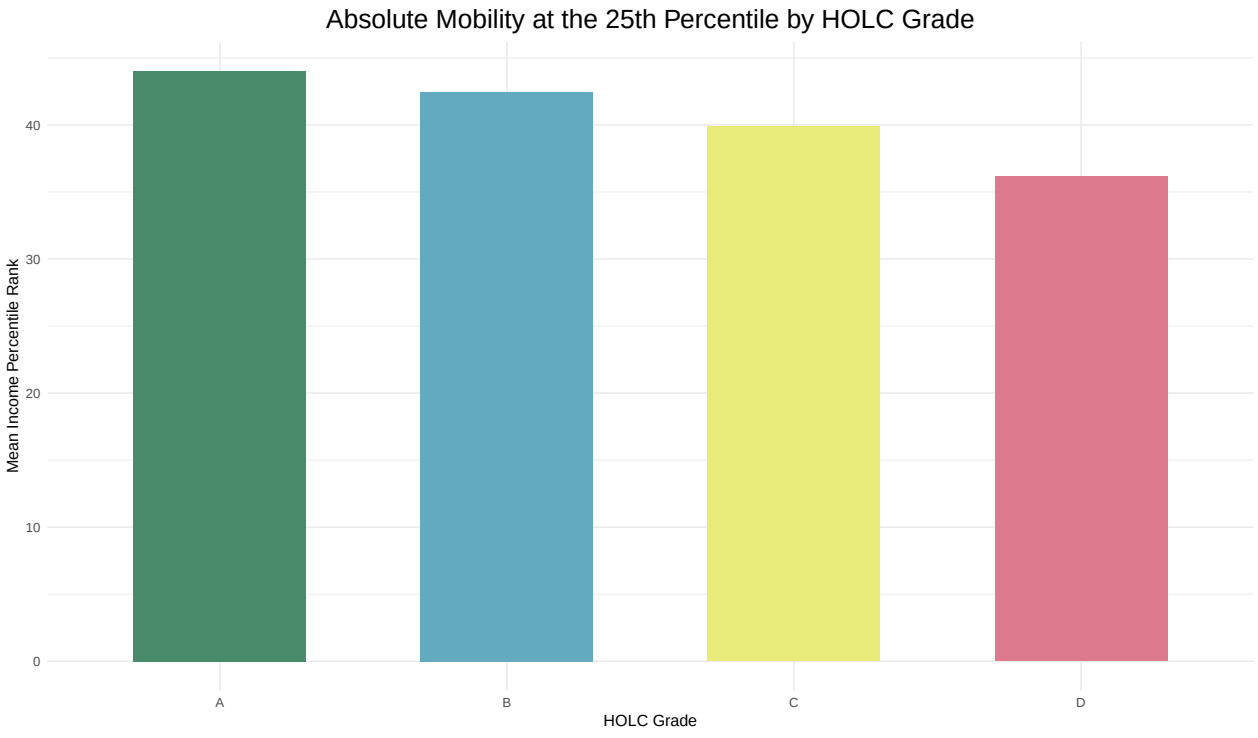
**11c: HOLC Grades and Upward Mobility Disaggregated by Race** By analyzing measures of Absolute Mobility at the 25th Percentile disaggregated by race we can remove this potential confounding

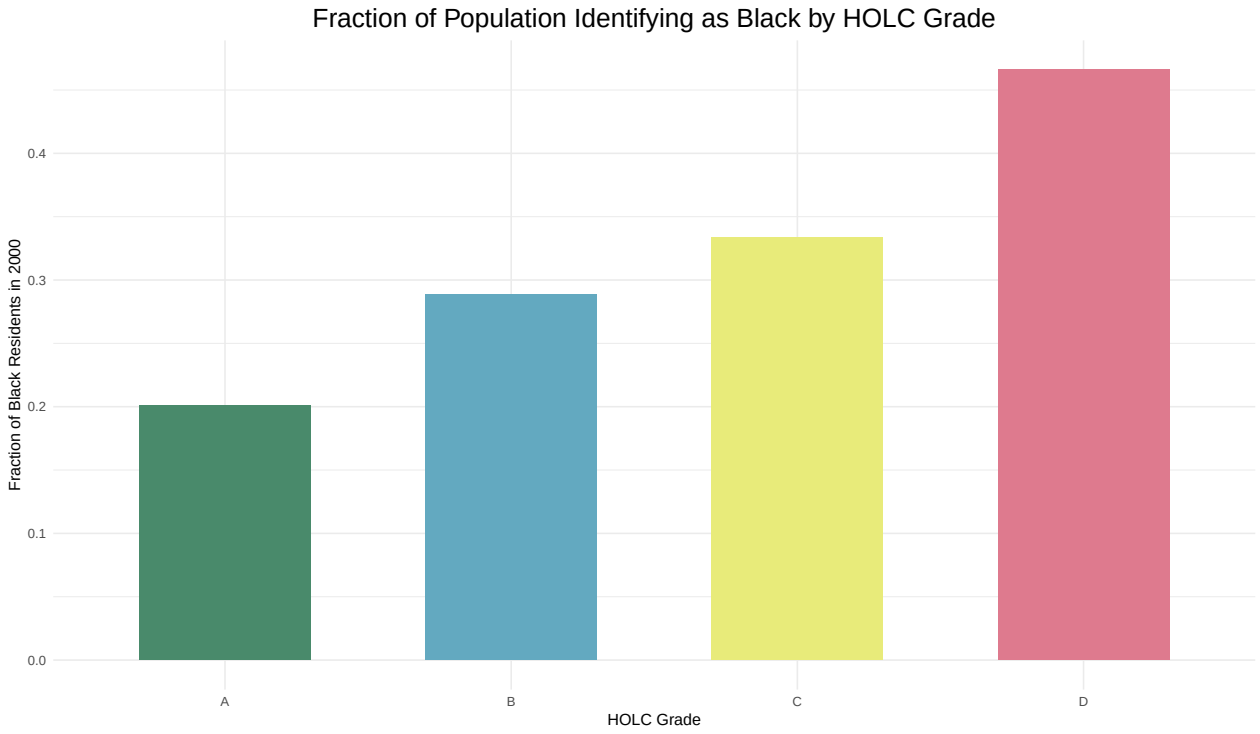
effect and start to better understand the direct relationship between HOLC grades and economic mobility. Because we are comparing the same mobility measure in the same tracts for both White and Black residents, we can then determine if the direction and magnitude of correlation between HOLC grades and outcomes is similar for both groups. Tracts with a majority of the area graded A had a mean kfr_black_pooled_p25 value of 34.4376 and kfr_white_pooled_p25 value of 50.3644, those majority graded B had mean values of 34.5071 and 48.7535 respectively, those majority graded C had mean values of 33.2393 and 46.3272 respectively, and those majority graded D had mean values of 31.6186 and 44.1159 respectively. From this comparison we can conclude that there is a correlation between HOLC grades and economic outcomes for both Black and White residents, where outcomes increase as the HOLC grade increases (A being the highest), though the magnitude of the effect is much larger for the White population than the Black population.

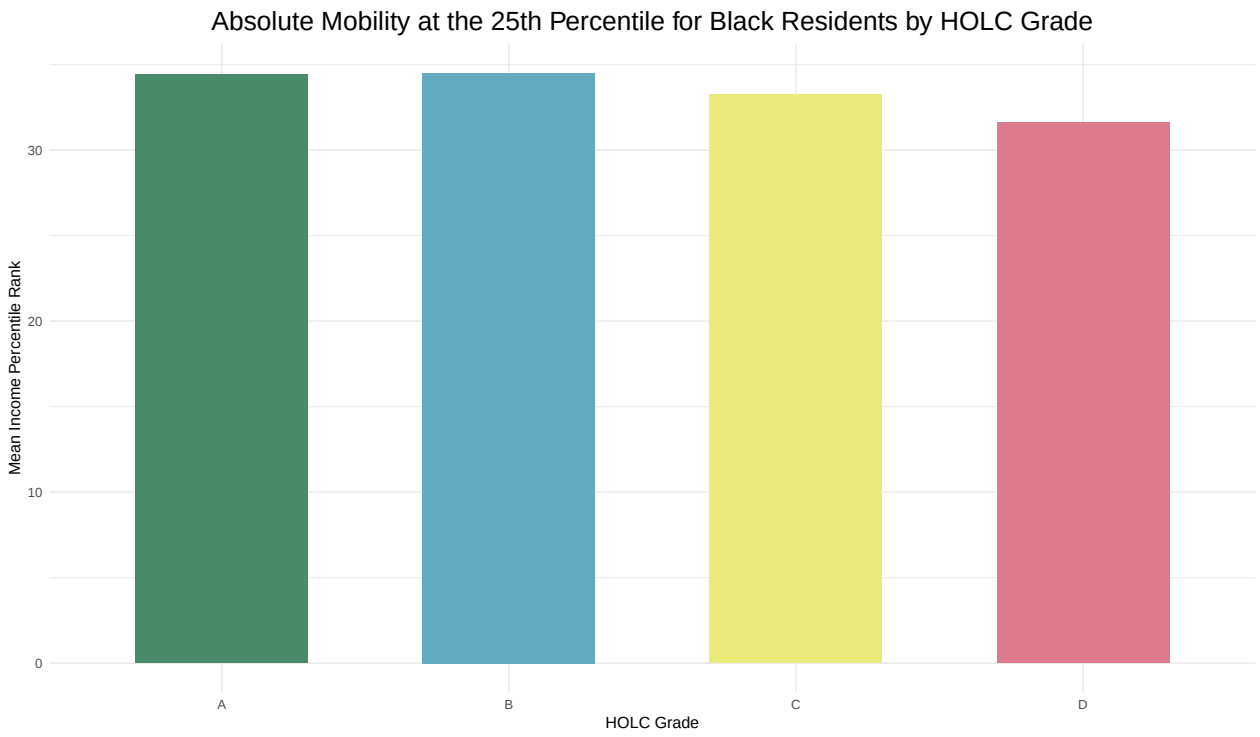
**11d: HOLC Grades and Environmental Characteristics** In keeping with Hoffman, Shandas, and Pendleton's hypothesis, these data show that negative environmental characteristics do correlate with lower HOLC grades. The mean values for vegetation, extreme heat, and fraction of land area developed by HOLC grade are displayed here:

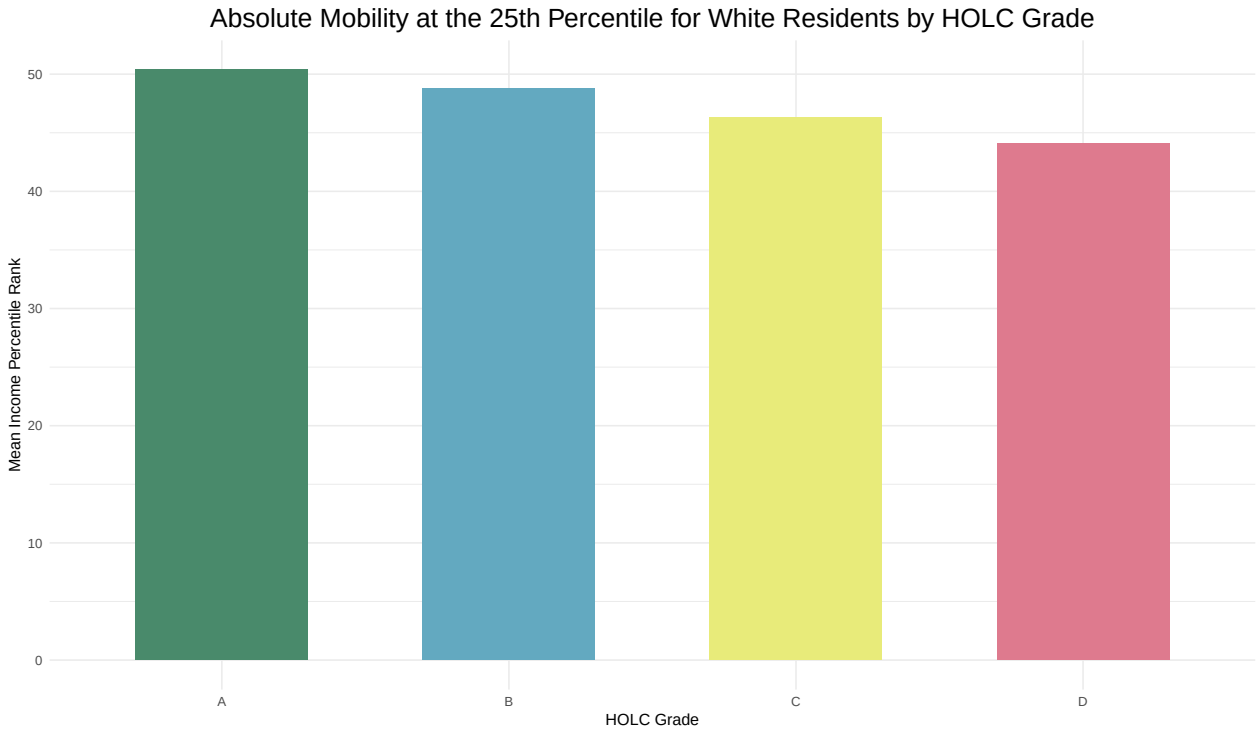
| HOLC_grade | vegetation | extreme_heat | developed |
|---|---|---|---|
| A | -0.0801636 | 4.223889 | 0.8917042 |
| B | -0.2048198 | 5.734555 | 0.9401400 |
| C | -0.2110291 | 5.764688 | 0.9425283 |
| D | -0.2633288 | 6.387577 | 0.9441439 |

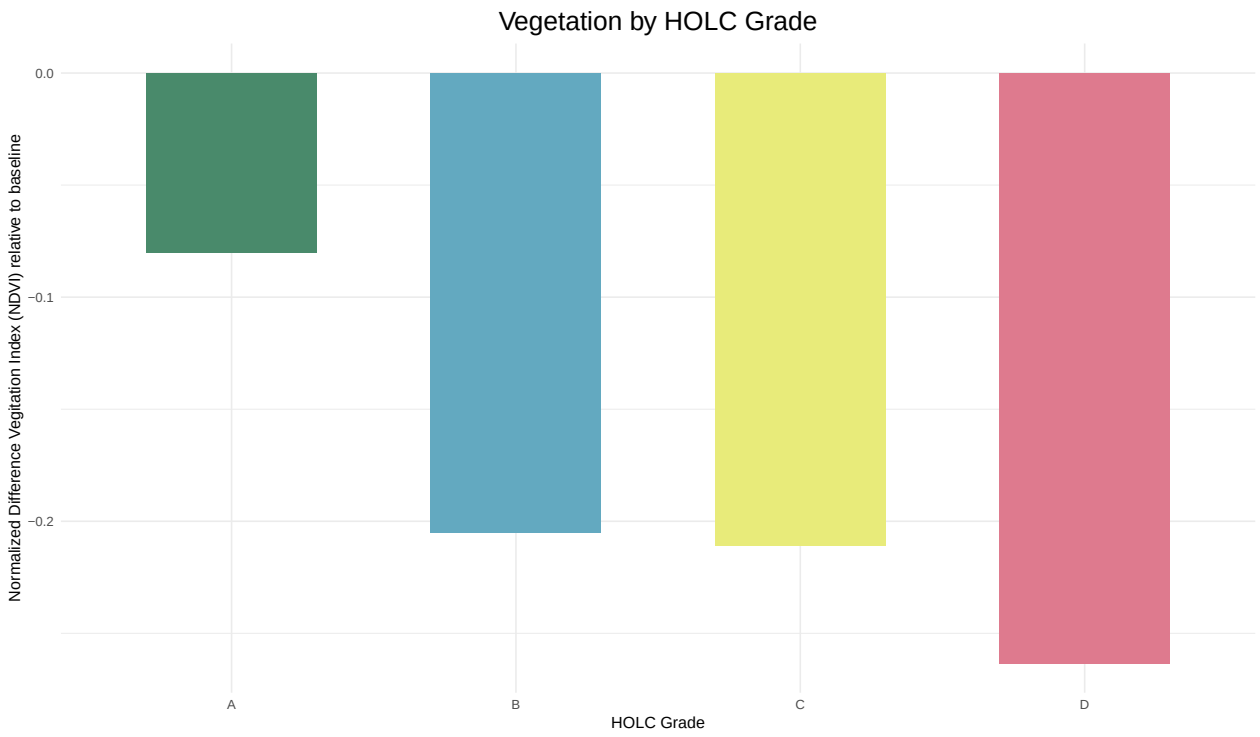
**11e: Bar Charts of Characteristics by HOLC Grades**

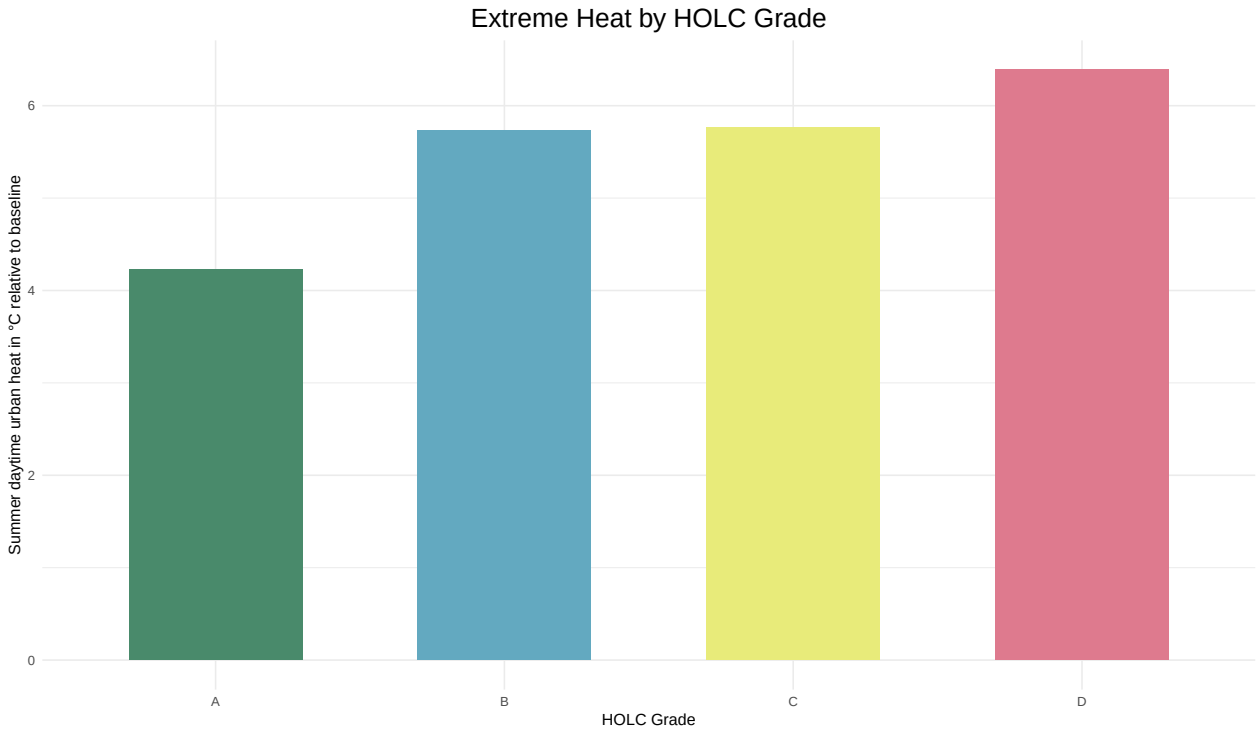


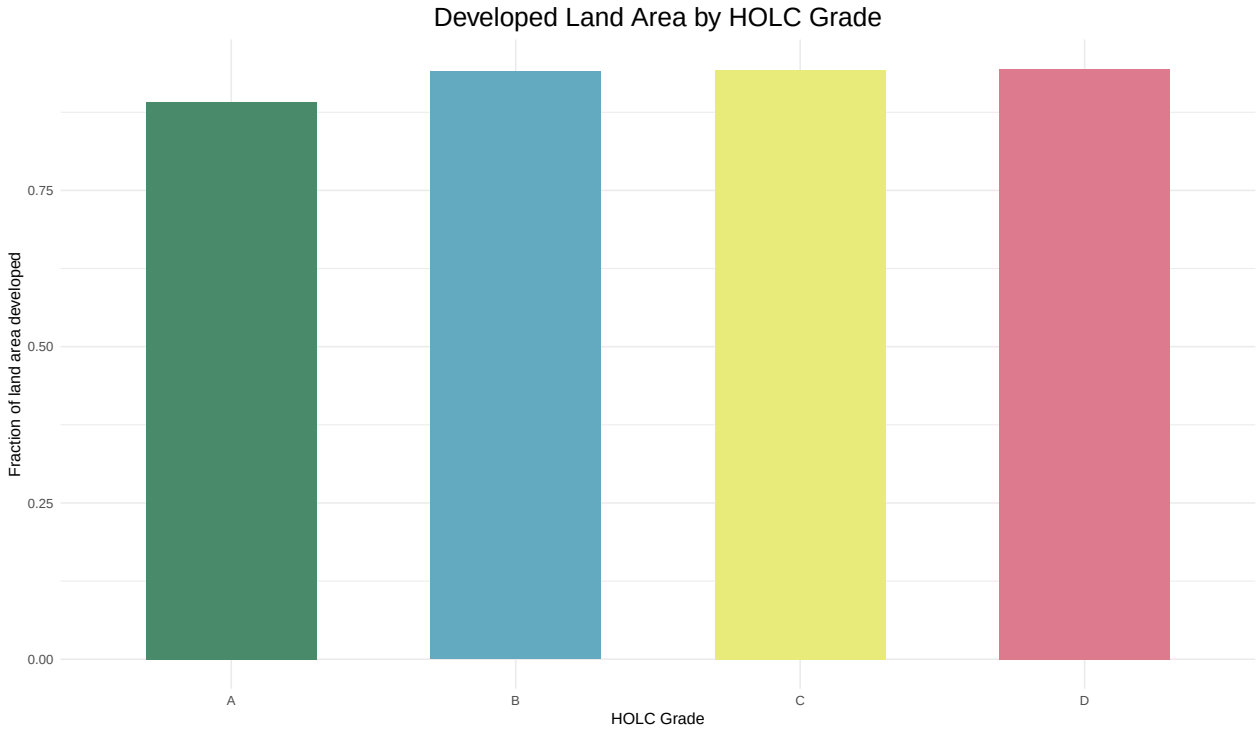

Fraction of Population Identifying as Black by HOLC Grade
Fraction of Black Residents in 2000
0.0
0.1
0.2
0.3
0.4
A
B
C
D
HOLC Grade


Absolute Mobility at the 25th Percentile for Black Residents by HOLC Grade
Mean Income Percentile Rank
0
10
20
30
A
B
C
D
HOLC Grade


Absolute Mobility at the 25th Percentile for White Residents by HOLC Grade
Mean Income Percentile Rank
50
40
30
20
10
0
A
B
C
D
HOLC Grade


Vegetation by HOLC Grade
Normalized Difference Vegitation Index (NDVI) relative to baseline
0.0
−0.1
−0.2
A
B
C
D
HOLC Grade



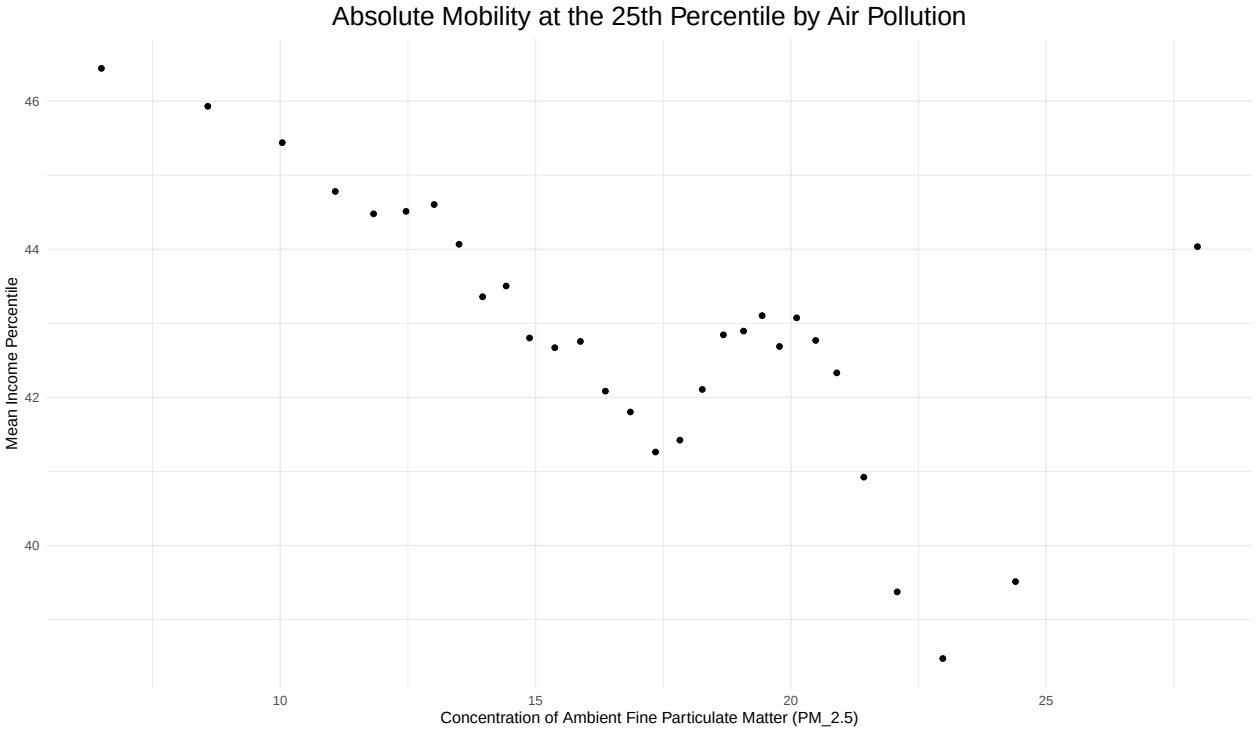




## Question 12: Air Pollution & Upward Mobility

**12a: Change in Pollution Over Time** Over the past forty years air pollution has trended downwards from 20.4198 PM_2.5 in 1982 to 16.8441 in 1990 to 12.5006 in 2000 and finally to 9.2864 in 2010.

**12b: Tract and National Pollution Comparison** The air pollution in my home tract in 1990 was 22.0052 which was higher than the national average of 16.8441 for that same year.

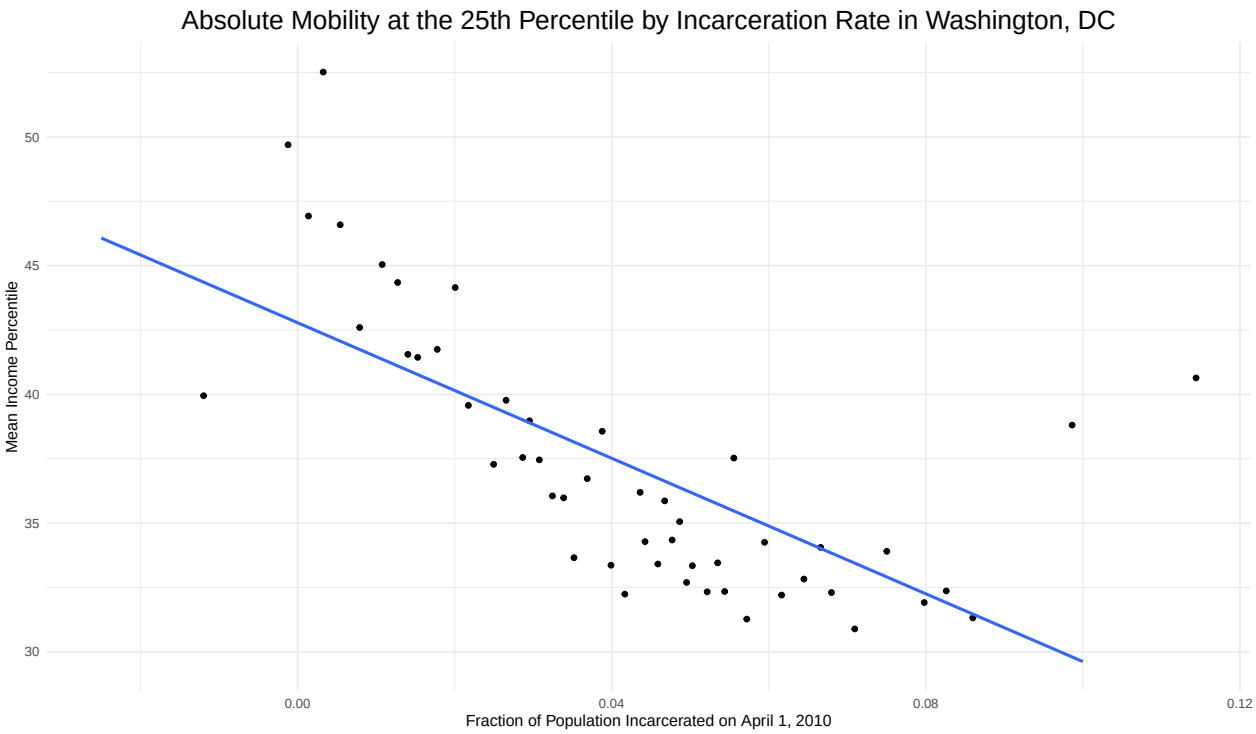
**12c: Pollution and Upward Mobility Binned Scatter Plot**

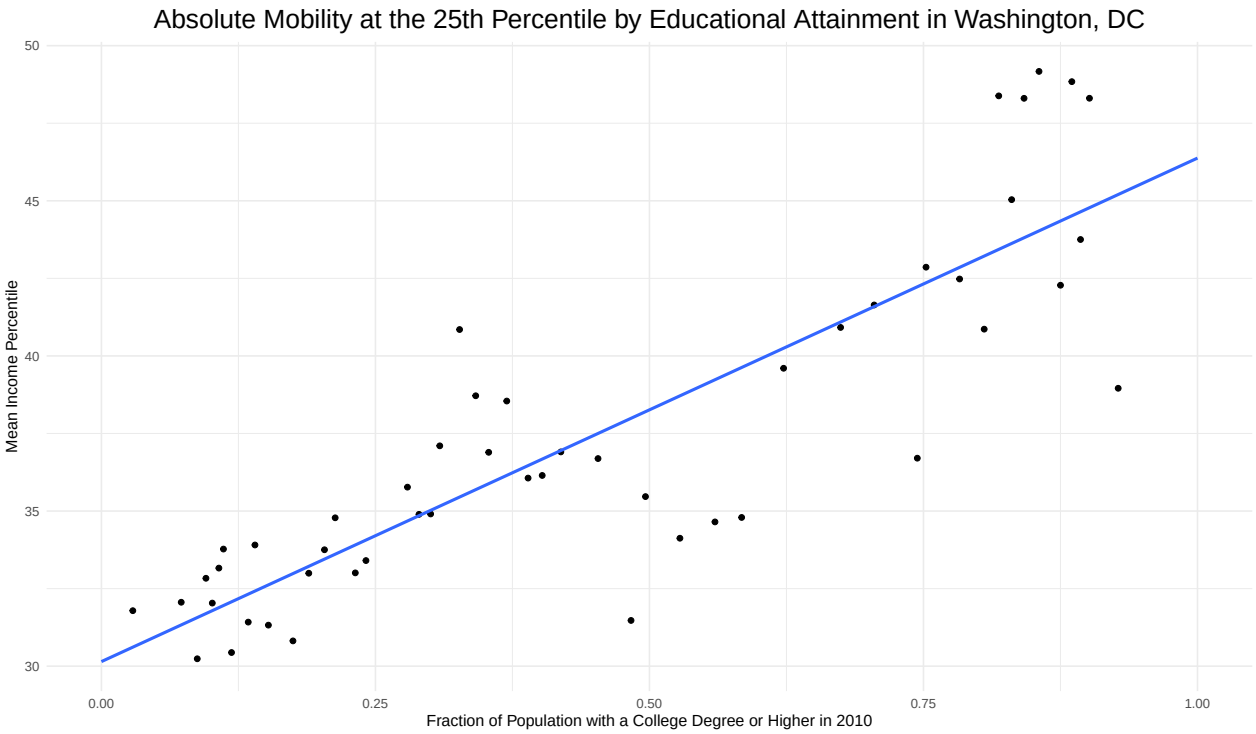



**12d: Pollution and Upward Mobility Correlation Coefficient** The correlation coefficient between kfr_pooled_pooled_p25 and pm25_1990 across census tracts is -0.1837, which is smaller than the county-level coefficient of -0.6 found by Colmer, Voorheis, and Williams. One possible explanation for this could be that there is greater variance in the tract level data than in the county level data, and as sample variance increases the estimated correlation coefficient becomes smaller.

## Question 13: Other Covariates

### 13a: Covariate Binned Scatter Plots



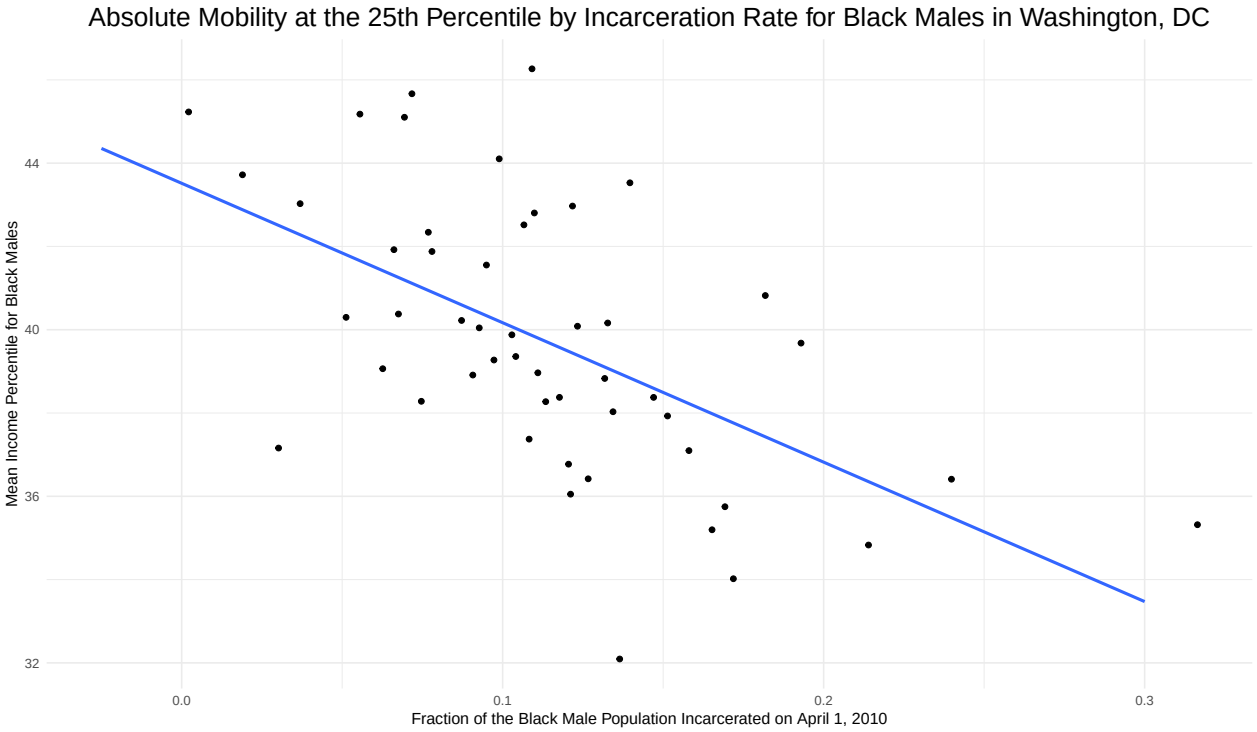



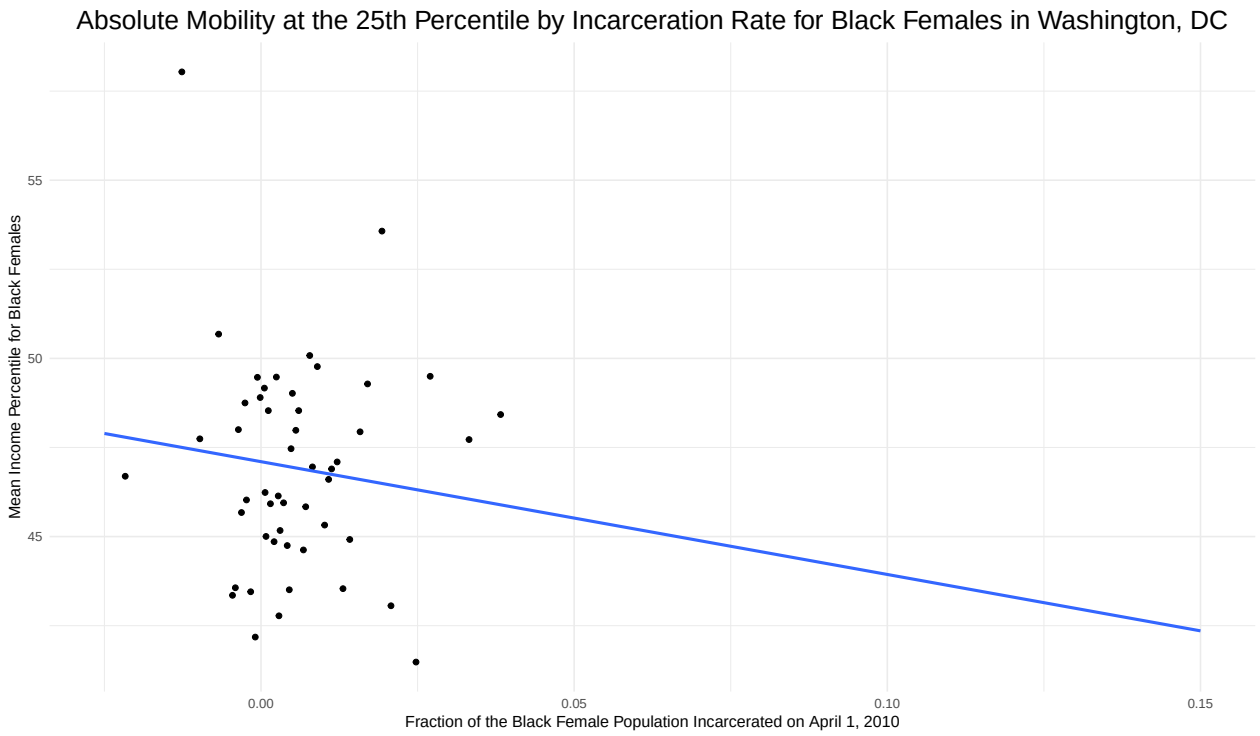
**13b: Covariate Coefficients** The two covariate relationships identified are with the jail_pooled_pooled_p25 variable which has a correlation coefficient of -0.549 and with the frac_coll_plus2010 variable with a correlation coefficient of 0.719.

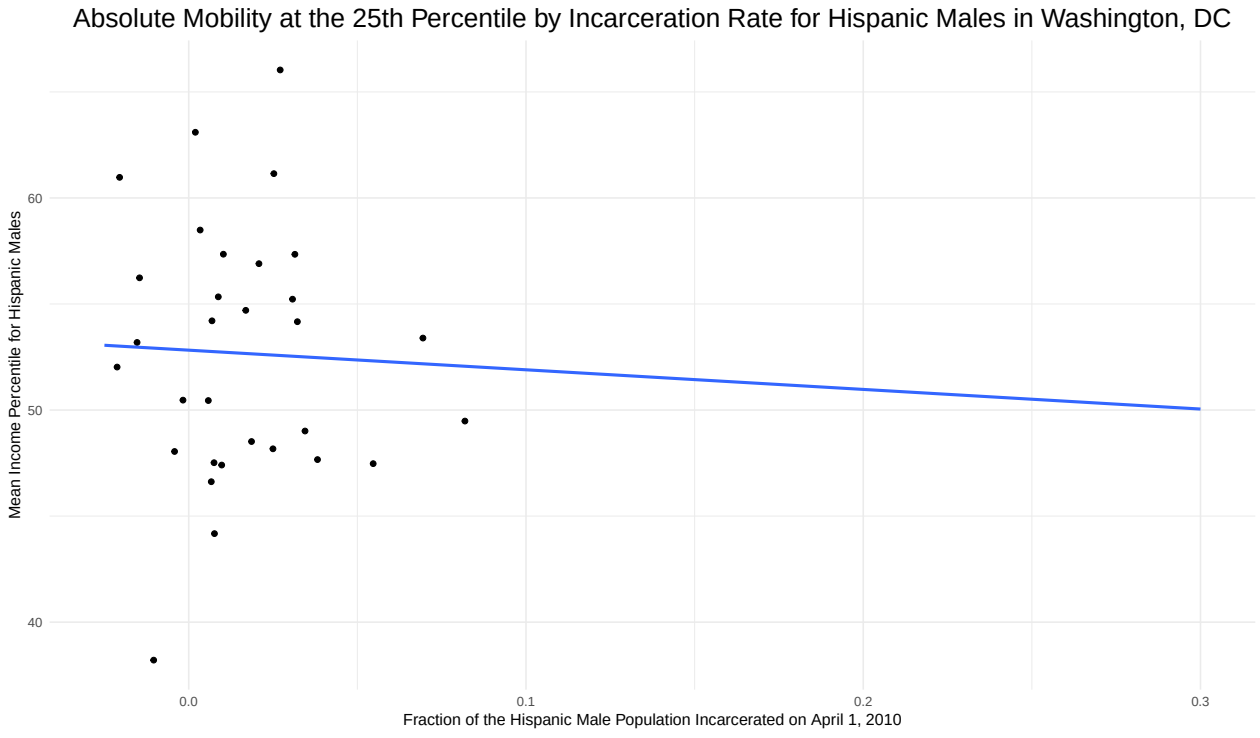
## Question 14: Other Covariates by Race and Gender

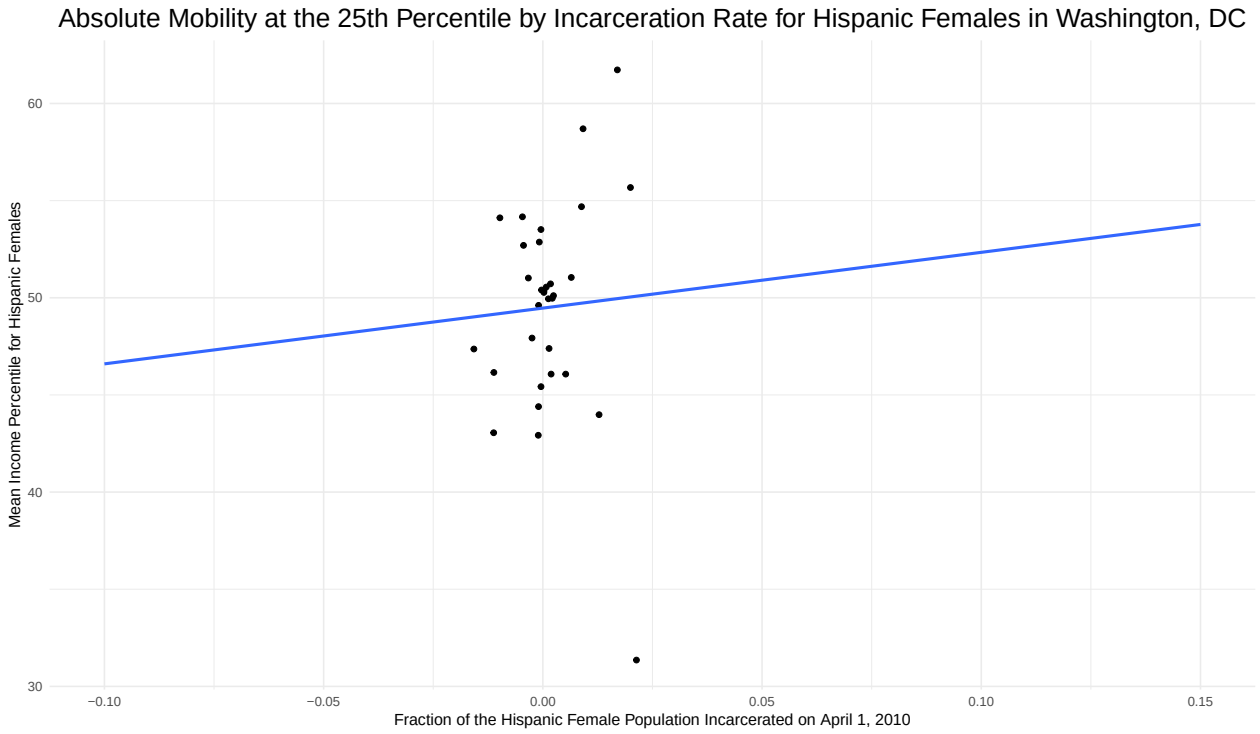
Of the two covariates identified, only the fraction of population incarcerated has the corresponding subset data by race and gender. Looking at the relationship between incarceration rate and economic outcomes for Black males, Black females, Hispanic males, Hispanic females, White males, and White females yeilds important information about the unequal impacts of certain factors along demographic lines. The correlation was strongest for Black males with a correlation coefficient of -0.395. Hispanic females had a correlation coefficient of 0.144, though looking at the plot it is easy to tell that there is no real relationship between these variables. For White males, Hispanic males, Black females, and White females the correlation coefficients were weak (between -0.1 and 0.1).

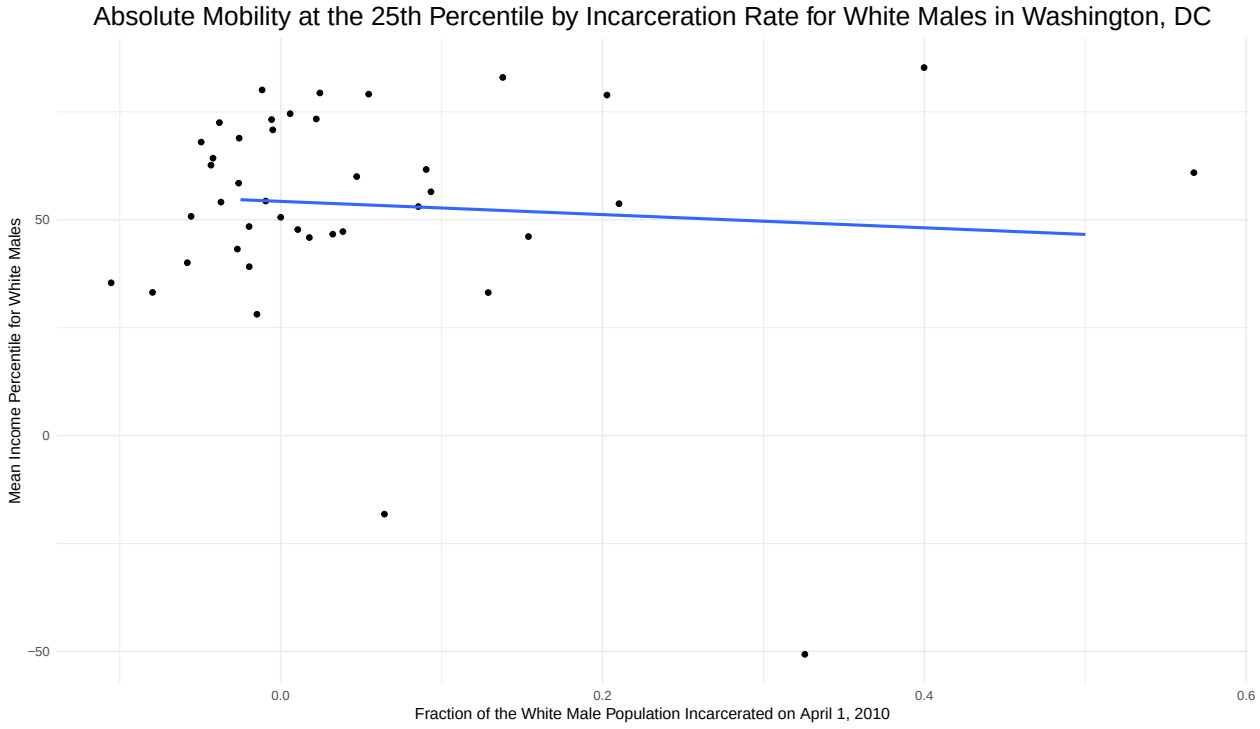
Binned scatter plots visualizing these relationships are presented below:

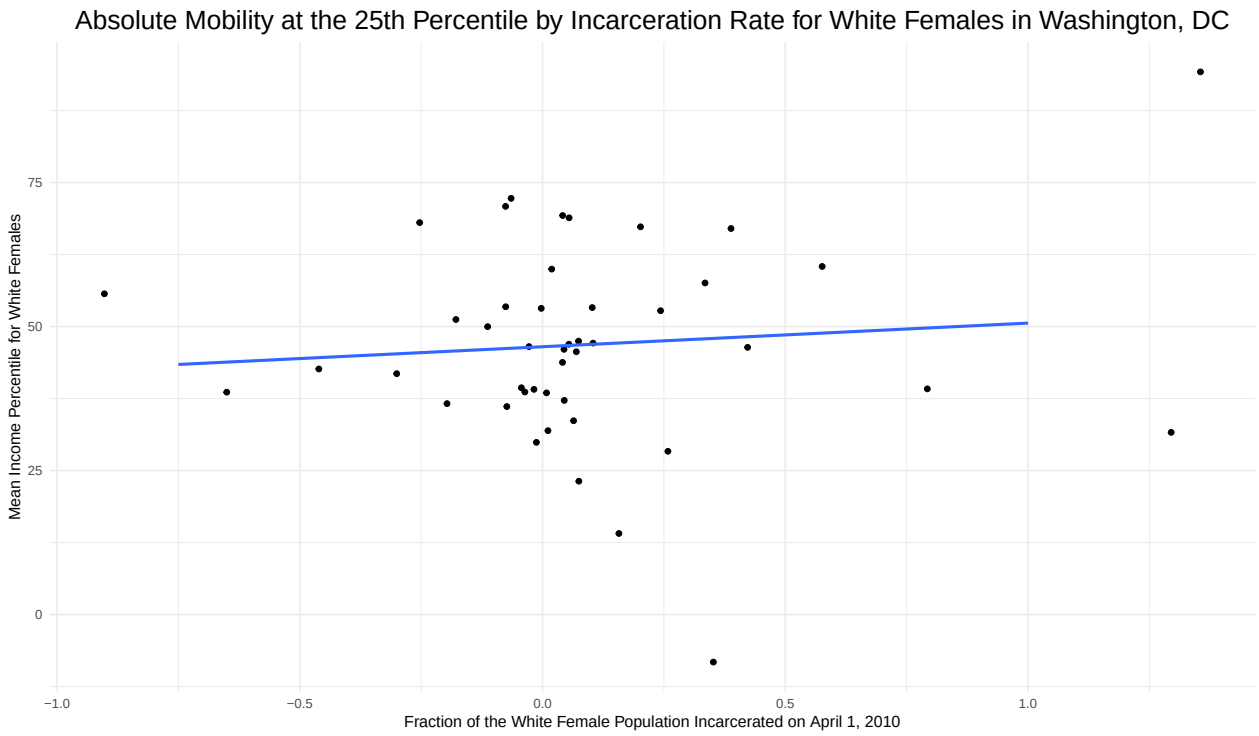



Absolute Mobility at the 25th Percentile by Incarceration Rate for Black Females in Washington, DC
Mean Income Percentile for Black Females
55
50
45
0.00
0.05
0.10
0.15
Fraction of the Black Female Population Incarcerated on April 1, 2010


Absolute Mobility at the 25th Percentile by Incarceration Rate for Hispanic Males in Washington, DC
Mean Income Percentile for Hispanic Males
60
50
40
0.0
0.1
0.2
0.3
Fraction of the Hispanic Male Population Incarcerated on April 1, 2010


Absolute Mobility at the 25th Percentile by Incarceration Rate for Hispanic Females in Washington, DC
Mean Income Percentile for Hispanic Females
60
50
40
30
−0.10
−0.05
0.00
0.05
0.10
0.15
Fraction of the Hispanic Female Population Incarcerated on April 1, 2010


Absolute Mobility at the 25th Percentile by Incarceration Rate for White Males in Washington, DC
Mean Income Percentile for White Males
50
0
−50
0.0
0.2
0.4
0.6
Fraction of the White Male Population Incarcerated on April 1, 2010


Absolute Mobility at the 25th Percentile by Incarceration Rate for White Females in Washington, DC
Mean Income Percentile for White Females
75
50
25
0
−1.0
−0.5
0.0
0.5
1.0
Fraction of the White Female Population Incarcerated on April 1, 2010