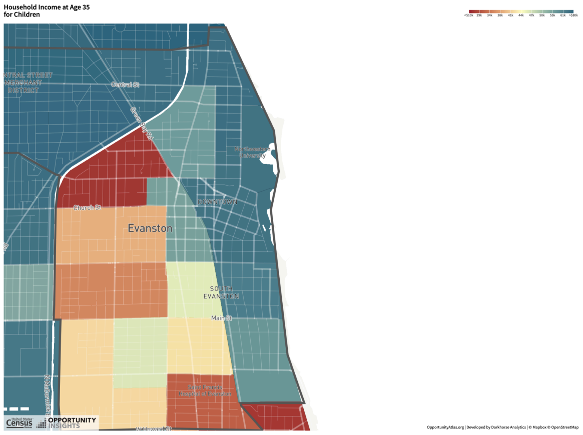
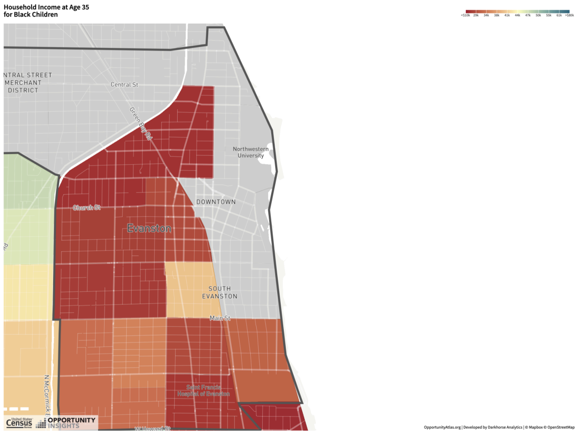
Shreya Chaturvedi
March 7th, 2024

HKS SUP 135: Using Big Data to Solve Economic and Social Problems
Project Part 1: Exploratory Data Analysis

1. **Start by looking up the community where you grew up or another community of your choice on the Opportunity Atlas. Zoom in on the map to view upward mobility at the Census tract level. Examine the spatial variation in your community for a number of different groups (e.g., race, gender, income level) and outcomes (e.g., income in adulthood, incarceration rates, teenage pregnancy rates).**




Household Income at Age 35
for Black Children
CENTRAL STREET MERCHANT DISTRICT
Central St
Northwestern University
Church St
DOWNTOWN
Evanston
SOUTH EVANSTON
Main St
N McCormick
OPPORTUNITY INSIGHTS
Census
OpportunityAtlas.org | Developed by Darkhorse Analytics | © Mapbox © OpenStreetMap

The neighborhood I have chosen to work with in this project is Evanston, IL. As in the prompt, the image on the left shows income at ages 31-37 for children of all races whereas the image on the right shows the distribution of the same metric but only for children of a black race.

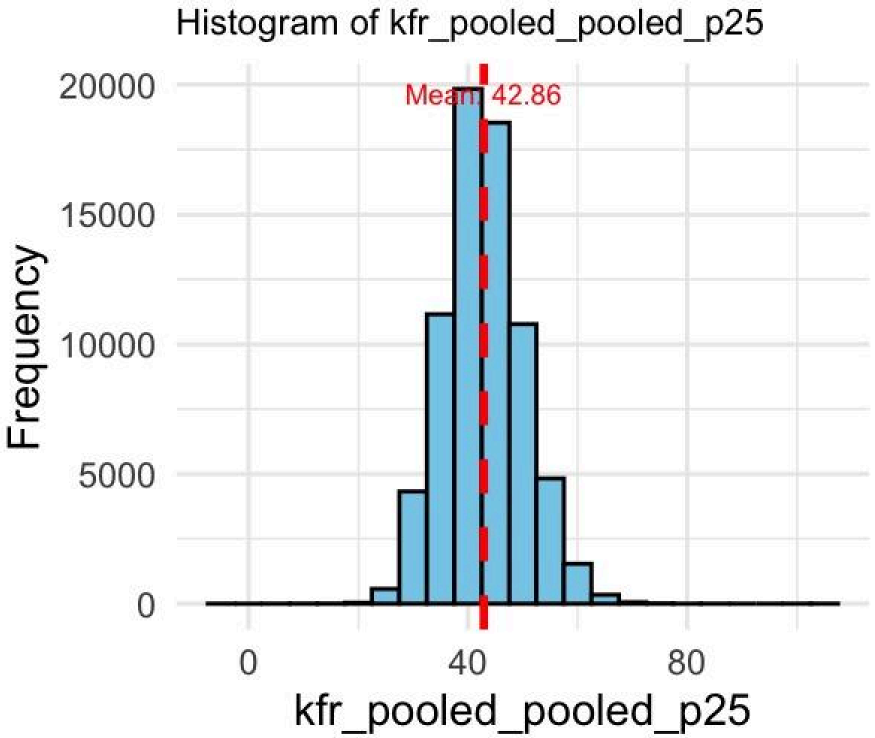
2. **Now turn to the atlas.dta data set. The variable kfr_pooled_pooled_p25 corresponds exactly to Statistic 1: Absolute Mobility at the 25th Percentile that you calculated in Lab 2.**
    a. **What are the units that this variable is measured in?**
    b. **Do higher or lower values correspond to higher upward mobility?**
    c. **Explain briefly why this statistic is estimated using a linear statistical model.**

The kfr_pooled_pooled_p25 variable corresponds to Statistic 1: Absolute Mobility at the 25th percentile. In other words, this shows the mean percentile rank in the national distribution of the children whose parents were at the 25th percentile of the income distribution. Therefore, the percentile itself does not have a unit since it is a rank measure. However, the percentile also corresponds to a particular annual income value which would be measured in US dollars.

Higher values of this variable would correspond to higher upward mobility since it would show that given the same starting point (parent at income percentile 25), the children of that neighborhood can reach a higher rank or earn more income compared to their peers.

Finally, linear models are used to estimate this statistic because they are a powerful and interpretable tool, as they can efficiently estimate the relationship between parental income and children's future income. The linearity assumption approximates this relationship well per economic research, and the models allow researchers to isolate the effect of parental income while controlling for confounders. Though the basic model is linear, transformations and interactions also provide flexibility to capture nuances.

3. **Produce a histogram of kfr_pooled_pooled_p25 using all Census tracts in the U.S. to get a sense of what the data look like, in density units as you did in Lab 1. Include an image of your graph in your solutions. What do you see?**



The histogram of absolute mobility for all the census tracts in the US is shown above. The mean of this distribution is 42.86, which means that on average, the child of a parent at income rank 25 can expect to reach income rank ~43 in the US. Next, most of the distribution lies between 20 and 80 percentiles, with very few observations outside of that range. The distribution also seems to be very slightly skewed towards the left, as is expected given the income inequality in the US.

4. **Report summary statistics (mean, standard deviation, minimum, and maximum) for kfr_pooled_pooled_p25. Include a table of the results in your solutions. Are there any observations with missing values for this variable?**

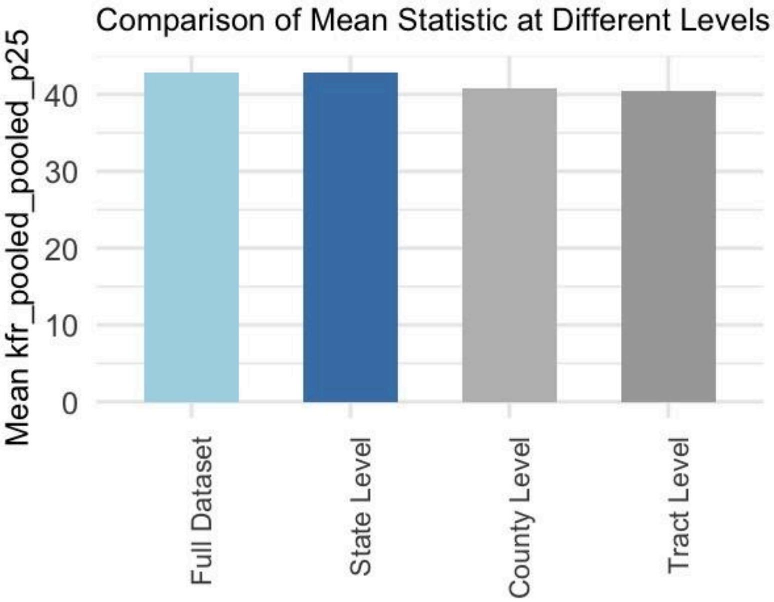
Statistic	**Value**
Mean	42.85
Standard Deviation	7.12
Minimum	-3.285
Maximum	103.34
Count of NAs	1189

5. **Why can kfr_pooled_pooled_p25 be negative or above 100 in these data? Hint: think about the limitations of a linear statistical model.**

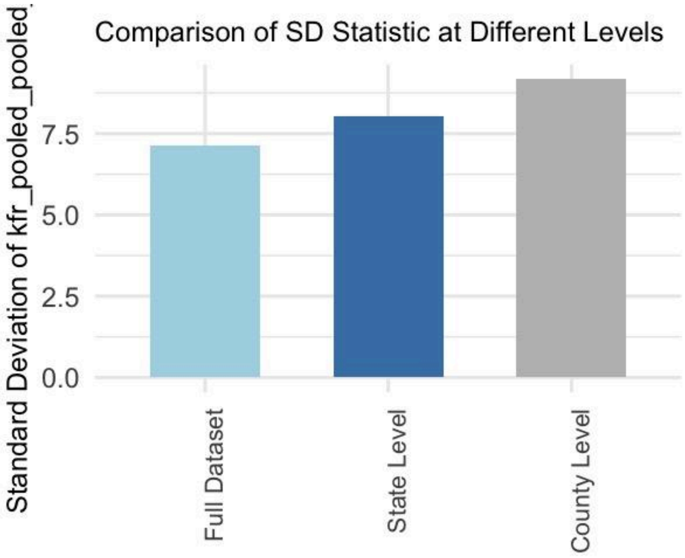
   A limitation of linear statistical models is that they can produce values outside of the expected range due to their inherent assumption of a constant rate of change across the predictor variables, which may not accurately reflect complex, non-linear relationships in real-world data - for instance, resulting in values at 103 percentile, which does not make sense at all. These models extrapolate based on the observed relationships, leading to potentially unrealistic predictions when applied to data points that are outside or at the extremes of the observed data range.

6. **Do kids where you grew up have better or worse chances of climbing the income ladder than the average child in America? In your home state? In your home county? To answer this question, compare the value of kfr_pooled_pooled_p25 in your home Census tract to the means in your state, county, and in the U.S. overall. Create an "Opportunity Insights style" bar graph visualizing these four values. Include an image of your graph in your solutions.**

   The required graph is shown below. As is seen here, the mean statistic is almost the same in all four cases, with only a small decrease as we move to the county and tract levels from the full dataset and state levels.

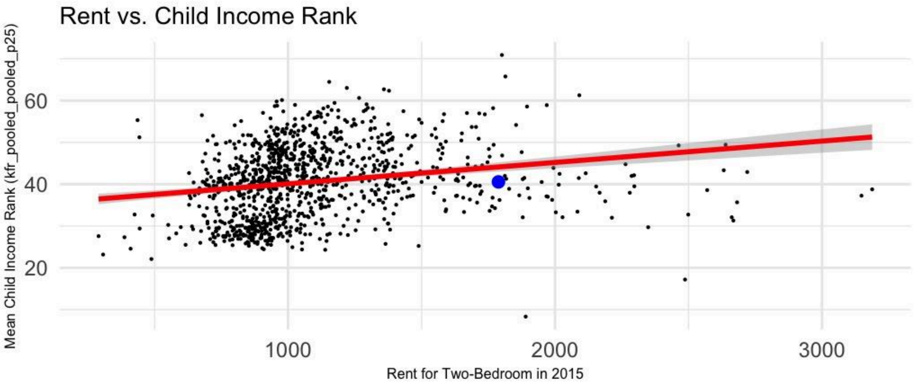


7. **What is the standard deviation of kfr_pooled_pooled_p25 in your home county? Is it larger or smaller than the standard deviation across tracts in your state? Across tracts in the entire U.S.? What do you learn by comparing the standard deviations of kfr_pooled_pooled_p25 for your county, state, and U.S. overall? Create an "Opportunity Insights style" bar graph visualizing these three values. Include an image of your graph in your solutions.**



The required graph is shown above. As is seen, the standard deviation increases as we move to a smaller subset of the data - from the US to only IL and then from IL to the county - each has a higher standard deviation. This could be explained by the smaller number of observations in each subsequent subset of the data as one possible justification.

8. **In this question, we will explore the relationship between upward mobility and rent in your chosen community, similar to the analysis from Professor Chetty's lecture for Seattle, WA shown below.**
    a. **Replicate the figure Professor Chetty showed in lecture using data restricted to only the Census tracts in your home county (or if you prefer your home Commuting Zone) instead of Seattle, WA. (If you are from Seattle, you can look at another community of your choice). Produce a regular scatter plot of kfr_pooled_pooled_p25 (y-axis) versus rent_twobed2015 (x-axis) with a linear best fit line. Include an image of your graph in your solutions.**
    b. **Do neighborhoods with better outcomes for low-income children have higher or lower rent in general? Explain clearly what you see in your scatter plot that leads you to your conclusions.**
    c. **Is your home Census tract an "Opportunity Bargain" as defined by Professor Chetty in lecture? What are some other communities in your county that are "Opportunity Bargains"? Explain clearly what you see in your scatter plot that leads you to your conclusions.**



The image below shows the scatter plot of the mean child income rank with the mean rent for a two-bedroom. My neighborhood is highlighted in blue, and the red line shows the linear best fit.

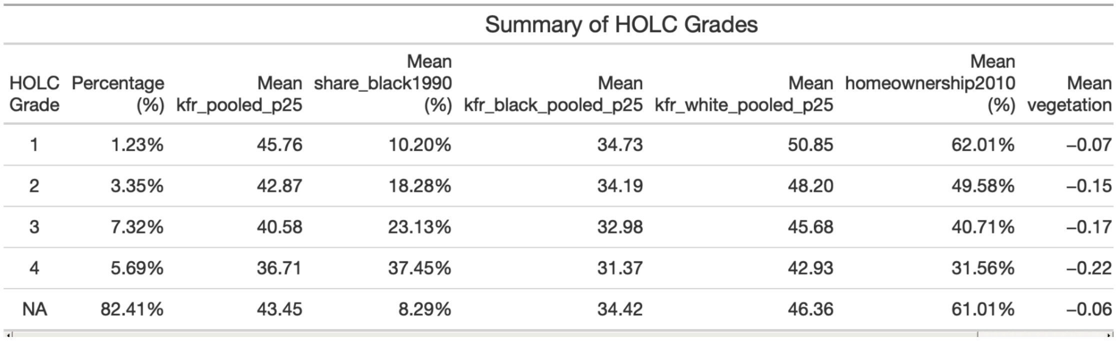
As is seen from the graph, the mean rent is almost independent of the child's income, with only a very slight upward slope. In other words, there is very little relation between child income rank and two-bedroom rent. As a consequence, it is possible to be in a much better neighborhood and pay the same or even lesser amount of rent.

In the case of my neighborhood, I would argue that the community is NOT an opportunity bargain since it is possible to move towards the left and pay much lesser to have access to the same level of opportunity (or even have access to higher opportunity by paying lesser with a move towards the top-left of the graph).

9. I have also included historical information from the 1930s on the Home Owners' Loan Corporation (HOLC) "Redlining" grades for each Census tract; the grades are missing for Census tracts that have no overlap with neighborhoods that were rated by HOLC. HOLC "Redlining" grades were used to restrict access to home mortgage financing, and were explicitly tied to racial and ethnic composition of neighborhoods in a very discriminatory manner, as shown both in historical documents digitized by the Digital Scholarship Lab at the University of Richmond and empirically by Aaronson, Hartley, and Mazumder (2021) and many others. For more background on the HOLC Redlining, see Aaronson, Hartley, and Mazumder (2021) and (optionally) the Vox video "Does My Neighborhood Determine My Future?" Warning: some of the language in the video may be upsetting.
   This question is meant to help walk you through an example of how you might use the richness of the Opportunity Atlas data to test your own hypothesis about upward mobility. It is based on Dr. Bruich and Prof. Chetty's analysis of these data. We want to demonstrate in a concrete example how we think about formulating and testing hypotheses, and illustrate the power of "reasoning by conditioning."
   a. Report averages of kfr_pooled_pooled_p25 for A, B, C, and D grade neighborhoods. What do you conclude about whether upward mobility for children born in the 1980s differs depending on their Census tract's HOLC grade from the 1930s? Be careful of missing data.
   b. The relationship you've demonstrated in the previous question could reflect compositional differences across neighborhoods (e.g., HOLC D grade neighborhoods are still highly segregated even today). As a first step in assessing whether this explanation is consistent with the data, report averages of share_black1990 for A, B, C, and D grade neighborhoods to document any differences in racial composition. What do you conclude about whether racial composition is a potential confounding variable?
   c. Next, Dr. Bruich and Prof. Chetty asked if the relationship between upward mobility and HOLC grade is still there holding fixed race. To do this, report averages of kfr_black_pooled_p25 and kfr_white_pooled_p25 for each of the HOLC grades. Explain clearly why racial composition cannot be a confounder in this analysis. What do you conclude?
   d. Having found that these relationships persist even holding fixed race, we next consider causal mechanisms. Inspired by a recent New York Times story about research by Hoffman, Shandas, and Pendleton (2020), one hypothesis is that HOLC "Redlining" led to underinvestment in D Grade neighborhoods: fewer public goods like parks and green spaces, more asphalt, and hotter temperatures even today. To assess this theory, I have also included data on (i) homeownership rates (ii) vegetation; and (iii) extreme summer time

temperatures. Report means of these variables (homeownership2010, vegetation, and extreme_heat) for each of the HOLC grades. What do you conclude about what might be driving the differences in mobility across HOLC grades that you documented earlier?

e. **Produce "Opportunity Insights style" bar graphs to visualize the means that you reported in parts a, b, c, and d. Include your graphs as images in your solutions.**

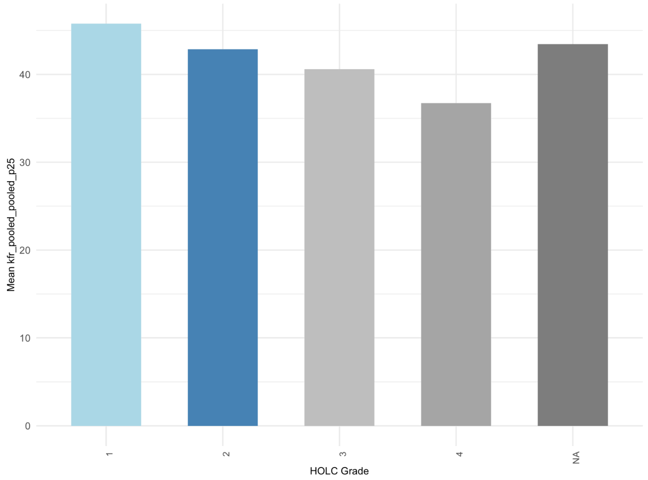


Summary of HOLC Grades

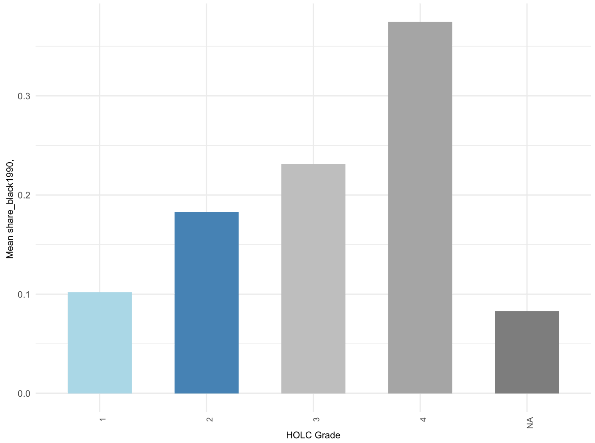
HOLC Grade	Percentage (%)	Mean kfr_pooled_p25	Mean share_black1990 (%)	Mean kfr_black_pooled_p25	Mean kfr_white_pooled_p25	Mean homeownership2010 (%)	Mean vegetation
1	1.23%	45.76	10.20%	34.73	50.85	62.01%	−0.07
2	3.35%	42.87	18.28%	34.19	48.20	49.58%	−0.15
3	7.32%	40.58	23.13%	32.98	45.68	40.71%	−0.17
4	5.69%	36.71	37.45%	31.37	42.93	31.56%	−0.22
NA	82.41%	43.45	8.29%	34.42	46.36	61.01%	−0.06

From the summary table presented here, we can see that the proportion of NAs is relatively large in the dataset, with the NAs taking up a sizeable majority. Within the other grads, the shares are more comparable. Further,

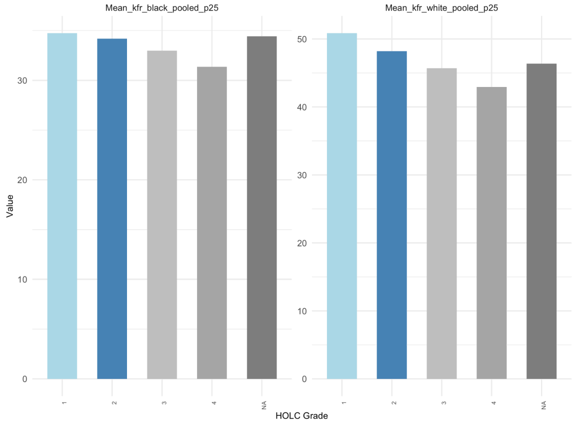
1. Averages of kfr_pooled_pooled_p25 decrease as we move from grade A to D, showing a lower mobility in the subsequent grades.



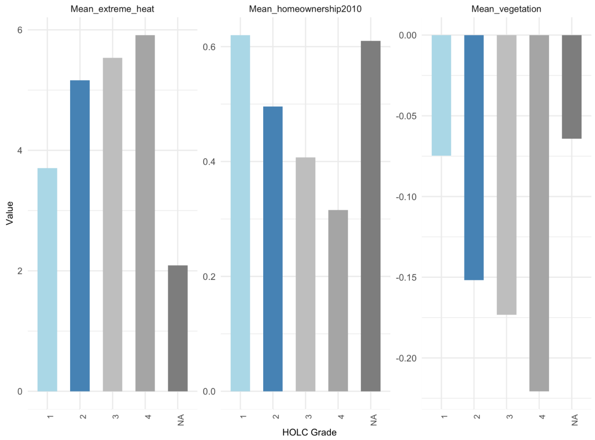
2. The share of black people also subsequentlyy increases across the grades, confirming the expected negative relationship between mobility and fraction of black populations.



3. Next, averages of kfr_black_pooled_p25 and kfr_white_pooled_p25 for each of the HOLC grades show that even within a particular race, the mobility pattern still holds. This confirms that acial composition cannot be a confounder in this analysis since the movement is in the same direction.

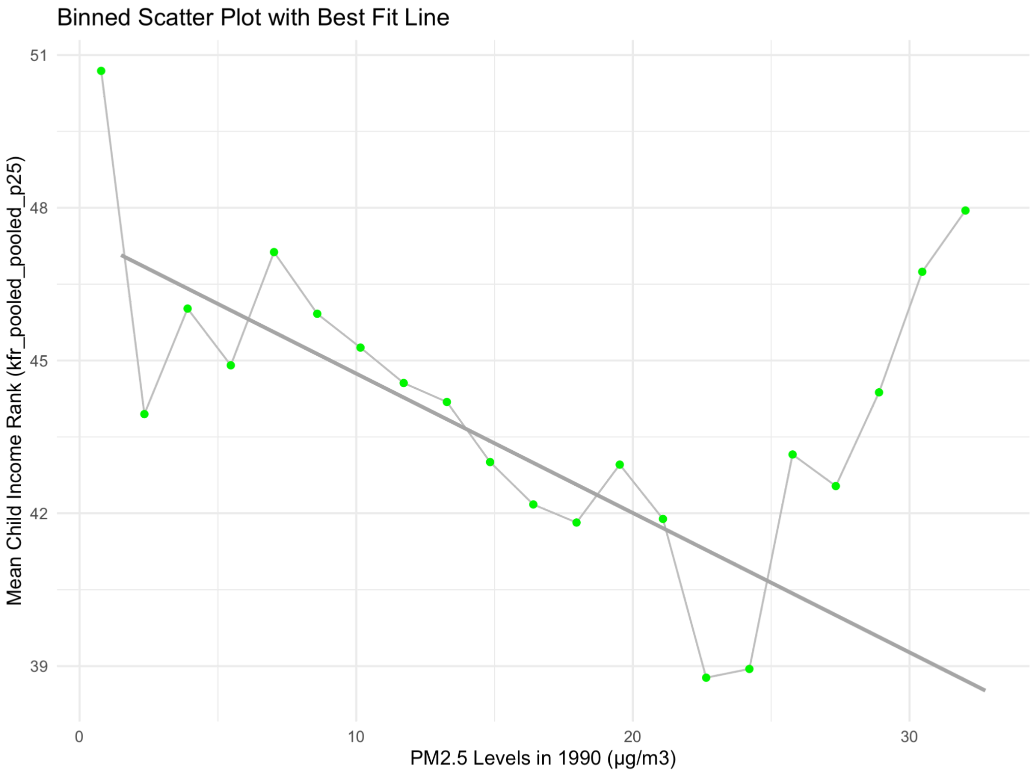


4. Finally, looking at homeownership2010, vegetation, and extreme_heat variables across the grade, we find that as we move through the grades (from A to D), the mean extreme heat increases , the home ownership decreases, and the vegetation decreases.



10. **Many students are passionate about the environment and climate change, but may not realize that economists and other social scientists study these issues in their research. Recently, Colmer, Voorheis, and Williams (2022) show that one of the strongest correlates with upward mobility across counties in the Opportunity Atlas is air pollution (in addition to the top five discussed in the video "Using Big Data to Understand Upward Mobility: Correlational Analysis"). To measure air pollution, they construct "satellite-derived, high-resolution data on particulate matter smaller than 2.5 microns (PM2.5) concentrations." They have generously provided these data, which I have merged with the Opportunity Atlas data.**
    a. **Note that currently, the EPA's National Ambient Air Quality Standards (NAAQS) for annual average PM2.5 levels is 12 micrograms per cubic meter of air (µg/m3). What fraction of Census Tracts in the U.S. were above this threshold in 1982? 1990? 2010? Be careful of missing values. Has air pollution worsened or improved over time? Explain your answer clearly.**
    b. **Visualize the relationship between kfr_pooled_pooled_p25 and pm25_1990 using a binned scatter plot for the entire U.S. Include an image of the graph in your solutions.**
    c. **Colmer, Voorheis, and Williams (2022) find that the correlation coefficient between kfr_pooled_pooled_p25 and PM2.5 across counties is −0.6. Collapse the data to create a county level data set. The grouping variables should be county and state. Compute and report the correlation coefficient (not regression coefficient) between kfr_pooled_pooled_p25 and pm25_1990 in the collapsed data for the entire U.S.**
    d. **Returning to the uncollapsed data, what is the correlation coefficient (not regression coefficient) between kfr_pooled_pooled_p25 and pm25_1990 across Census tracts for the entire U.S.? Why do you think the correlation coefficient might be smaller across tracts than across counties? This is a thinking question.**

After discarding NA values from consideration and calculating the fraction of Census Tracts above the PM2.5 threshold for the years we have data for, we find that over time the average fraction of counties over the threshold has gone from 92.9% ro 11.9%. This indicates that over time air pollution has improved.



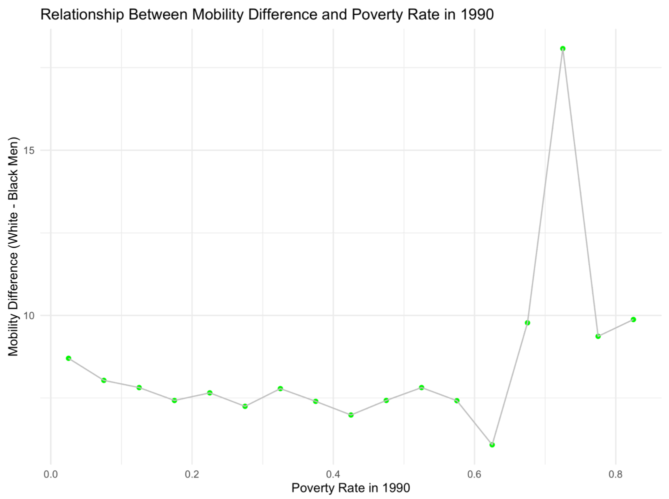
The correlation coefficient is -0.602 between the child income rank at 25 percentile and the 1990 pm levels when calculated at the county and state level. When the same metric is calculated at tract levels as in the original data, the value comes out to be -0.1837.

One possible explanation for the smaller value could be that aggregation can mask within-group variability and local spatial heterogeneity present at finer geographic scales such as tracts. Averaging the data within counties also smooths out extremes and outliers that contribute to the analysis at the tract level. This loss of granular variability and spatial nuance when moving from tracts to counties can obscure detailed relationships, leading to weaker observed correlations compared to the more precise tract-level data.

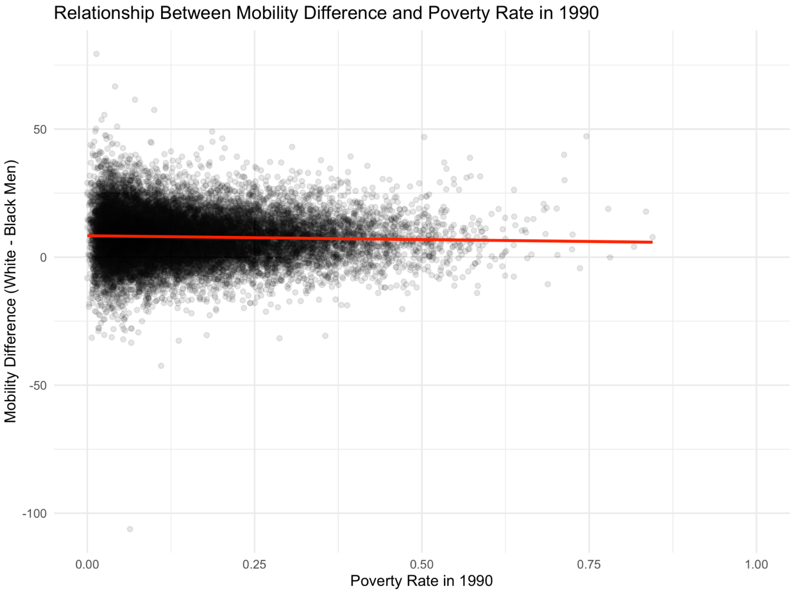
11. **In this question, we will replicate some key results documented by Chetty, Hendren, Jones, and Porter (2020) on racial disparities in upward mobility.**
    a. **Generate a new variable that equals the difference in Absolute Mobility at the 25th Percentile for white and Black men, defined as kir_white_male_p25 minus kir_black_male_p25.**
    b. **In what fraction of Census tracts do white men have higher levels of upward mobility than Black men? (Be careful of missing values).**
    c. **Visualize the relationship between the variable you generated in part (a) and the poverty rate in 1990 (poor_share1990) using a binned scatter plot. Include the image of the graph in your solution.**
    d. **Are racial disparities between white and Black men bigger or smaller in more affluent neighborhoods? Explain clearly what you see in your binned scatter plot that leads you to your Conclusion.**
    e. **Formulate a hypothesis that might explain the result you documented in the previous question. This is a thinking question.**

The variable is created in the accompanying R Code. The calculated fraction value for tracks where white men have higher levels of upward mobility is 82.72%. That is, in 82.72% of tracts, white men have higher upward mobility, all else equal.

The relationship between the mobility difference and poverty rate is in the scatter plot below.



Since this binned scatterplot was not very clear, I also created a regulat scatterplot below:



This scatterplot shows a balanced distribution or in other words almost the same expected mobility difference across poverty rates on average (shown by the points symmetrically along the red line). This shows that the expected value of difference in mobility in both more and less affluent neighborhoods is approximately the same. However, there is greater variability in poorer neighborhoods.

A hypothesis to explain the potential relationship between neighborhood affluence and racial disparities in mobility might involve access to resources and opportunities. In more affluent neighborhoods, the hypothesis could be that there is less competition for resources, leading to smaller racial disparities as all residents may benefit from the wealth of the neighborhood. Conversely, in less affluent neighborhoods, the competition for limited resources might exacerbate racial disparities due to systemic biases and discrimination that limit opportunities for Black men relative to white men. Another hypothesis could involve social capital and networks, which may be more homogeneous in terms of race in affluent areas, possibly leading to smaller disparities within those networks. However, these hypotheses would need to be tested with data that clearly differentiates the mobility outcomes by race across different levels of neighborhood affluence.

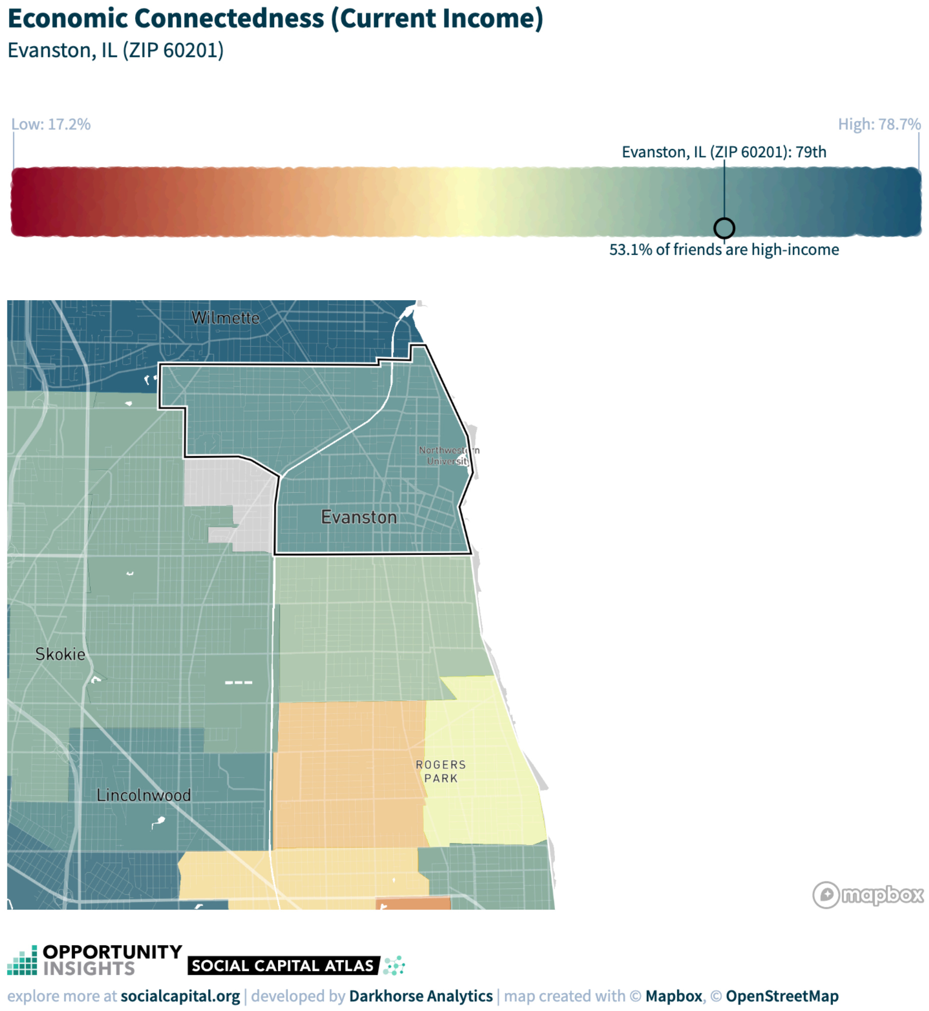
12. **Next we will replicate some key results from Chetty et al. (2022a, 2022b) on the relationship between social capital and upward mobility. We will combine the social capital measures constructed using the Facebook data with the data on upward mobility. A complication is that the Social Capital data is only available by Zip code and not by Census tract.**
    a. **Start by merging the atlas.dta data set with the cross-walk data set zip_tracts_xwalk.dta that connects each Census tract with one or more zip codes.**
    b. **Now collapse the data to obtain a data set containing the weighted average of kfr_pooled_pooled_p25 for each zip code. The grouping variable is zip and the weight is zpoppct, the percent of the zip code's population contained in each Census tract.**
    c. **Next merge the collapsed mobility data set with the social capital data Social_capital_zip.dta.**
    d. **(3 points) Calculate the correlation coefficient (not regression coefficient) between kfr_pooled_pooled_p25 and each of the following measures of social capital: (i) economic connectedness (ec_zip), cohesiveness (clustering_zip), and civic engagement (civic_organizations_zip).**
    e. **Discuss the results you documented in (d). Which measure of social capital is most strongly related to upward mobility?**

After performing the outlined data manipulations, these are the values we get:

1. Economic Connectedness Correlation (cor_kfr_ec): 0.69
2. Cohestiveness Correlation (cor_kfr_clustering): -0.07
3. Civic Engagement Correlation (cor_kfr_civic): -0.01

This means that the economic connectedness is most strongly correlated to upward mobility given the highest magnitude of the three measures.

13. **Look up the community of your choice on the Social Capital Atlas. Examine the spatial variation in the three social capital metrics for your community by zip code (i.e., connectedness, cohesiveness, and civic engagement). Save a .png of a map to share with the rest of the class using the Social Capital Atlas' "download as image" button. Post your map to the #ec50-s24-welcome channel. Include a brief message saying which measures of social capital your community is high or low in. Please feel free to react (kindly and respectfully, please) to the posts of others. Please also include your image and description in your solution write-up.**


Economic Connectedness (Current Income)
Evanston, IL (ZIP 60201)
Low: 17.2%
High: 78.7%
Evanston, IL (ZIP 60201): 79th
53.1% of friends are high-income
Wilmette
Northwestern University
Evanston
Skokie
ROGERS PARK
Lincolnwood
mapbox
OPPORTUNITY INSIGHTS
SOCIAL CAPITAL ATLAS
explore more at socialcapital.org | developed by Darkhorse Analytics | map created with © Mapbox, © OpenStreetMap

The community I have chosen is Evanston, IL. This neighborhood is high in measures of economic connectedness (79th percentile) and civic engagement (91st percentile). It is moderately low in the measure of cohesiveness (35th percentile).

14. **Based on what you see in your maps in questions 1 and 13, your exploratory analysis above, and background knowledge about the community that you have selected, state a preliminary hypothesis that might explain the variation in upward mobility in the maps you selected. Your hypothesis must include discussion of possible causal mechanisms, like the underinvestment in public goods mechanism for the HOLC example. Your hypothesis should also not be anachronistic: keep in mind that the children in the Opportunity Atlas data were young children in the 1980s and 1990s. You can modify your hypothesis for Part 2, but this will give you an opportunity to receive some feedback.**

Hypothesis: Northwestern University's presence increased upward mobility in Evanston
Evanston, IL ranks highly on economic connectedness and civic engagement, suggesting a tight-knit community. Northwestern University was likely a key driver of this engaged culture, expanding opportunities for local youth.

Possible Causal Mechanisms:

1. Networks and Mentorship: Northwestern provided access to social capital through networks of students, faculty, and alumni. These connections offered mentorship and guidance to inspire local kids' aspirations.
2. Exposure to Careers and Education: Interactions with the Northwestern community exposed youth to diverse career paths and educational opportunities to enhance their mobility.
3. Economic and Work Experiences: University connections opened doors for internships, jobs, and other hands-on professional experiences to boost skills.
4. Civic and Community Engagement: Northwestern's service-oriented culture bled into the broader Evanston community, creating an actively engaged environment and empowering youth.
5. Educational Enrichment: Academic and tutoring programs offered by Northwestern supplemented local student's education and helped bridge opportunity gaps.