Empirical Project 1
Stories from the Atlas: Describing Data using Maps, Regressions, and Correlations
DUE: SUNDAY, JANUARY 23, 2022 (11:59 PM)

Instructions

Please submit your Empirical Project on **BBLearn**. Your submission should include two files:
1. A do-file with your STATA code
2. A word or pdf document with answers to the listed questions

This is a group project and only one submission per group is allowed.
(Same groups as your Lab groups)

The Opportunity Atlas was publicly released on October 1, 2018, and an accompanying article appeared on the front page of the *New York Times*. The Opportunity Atlas is a freely available interactive mapping tool that traces the roots of outcomes such as poverty and incarceration back to the neighborhoods in which children grew up.

Policymakers, journalists, and the public have begun to explore the Opportunity Atlas, casting new light on the geography of upward mobility in communities across the country. As an example, see Jasmine Garsd's recent analysisfor the New York City neighborhood of Brownsville in Brooklyn.

In this first empirical project, you will use the Opportunity Atlas mapping tool and the underlying data to describe equality of opportunity in your hometown and across the United States. (If you grew up outside the United States, you may select a community in which you have spent some time, such as Philadelphia, PA.)

The next page lists specific analyses and questions that you must address.

This project focuses on the following methods for descriptive data analysis. (The later empirical projects you will do in this class will be focused on causal inference and prediction).

1. *Data visualization.* Maps are a powerful way to present descriptive statistics for data with a geographic component. You will use maps to display upward mobility statistics for the Census tracts in your hometown.
2. *Regression and correlation analysis.* You will use linear regressions and correlation coefficients to quantify the statistical relationship between upward mobility and potential explanatory variables.

The Stata data file that you will use in this assignment, atlas.dta, contains an extract of the Opportunity Atlas data and several other variables, which you may use for the correlational analysis.

Group No: 4

Group Members: Swetha Nandakumar, Shalin Luitel, Jahnavi Kalyan, Arnold Gyateng

Questions

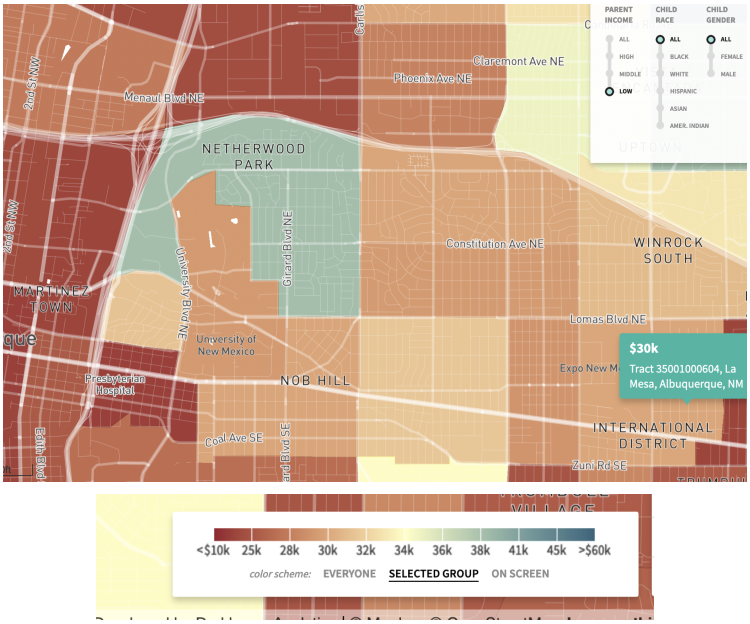
1. Start by looking up the city where you grew up on the Opportunity Atlas. Zoom in to the Census tracts around your home.

Examples for Milwaukee, WI and Los Angeles, CA are shown on the next page. Describe what you see, and what data are being visualized.

Examine the patterns for a number of different groups (e.g., lowest income children, high income children) and outcomes (e.g., earnings in adulthood, incarceration rates). Only choose one or two of these to include in your narrative.

Upon looking at the Opportunity Atlas, you can see that there is some economic diversity within Albuquerque, New Mexico, with the majority median income being 30k.

Household income varies by race and gender. We chose to analyze the white and Hispanic populations in Albuquerque due to them being the demographic majority. White households have a median income 37k, while Hispanic households have a median income of 35k. For white females, the household median income is 39k, and for Hispanic females the household median income is 36k.



2. Now turn to the atlas.dta data set. How does average upward mobility, pooling races and genders, for children with parents at the 25th percentile (kfr pooled_p25) in your home Census tract compare to mean (population-weighted, using count_pooled) upward mobility in your state and in the U.S. overall? Do kids where you grew up have better or worse chances of climbing the income ladder than the average child in America?

Hint: The Opportunity Atlas website will give you the tract, county, and state FIPS codes for your home address. For example, searching for "Lynwood Road, Verona, New Jersey" will display Tract 34013021000, Verona, NJ. The first two digits refer to the state code, the next three digits refer to the county code, and the last 6 digits refer to the tract code. In Stata, listing this observation can be done as follows:

list kfr_pooled_p25 if state == 34 & county == 013 & tract == 021000

Census tract NM = 35001001700
list kfr_pooled_p25 if state == 35 & county == 001 & tract == 001700
Income = 30224.92
National average = 34443.48

3. What is the standard deviation of upward mobility (population-weighted) in your home county? Is it larger or smaller than the standard deviation across tracts in your state? Across tracts in the country? What do you learn from these comparisons?

```
. sum kfr_pooled_p25 if state == 35 & county == 001

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
kfr_poole~25 |        153    32073.18    5865.688   14214.86   55482.66

. sum kfr_pooled_p25 if state == 35

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
kfr_poole~25 |        497    32299.19     5935.82   14214.86   55482.66

. sum kfr_pooled_p25

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
kfr_poole~25 |     72,011    34443.48    8169.155          0   105732.4
```

The standard deviation in my home county (Bernalillo County) and state (New Mexico) is $5865 and $5935 per year respectively. However, the standard deviation of USA is

much higher at $8169 per year. We can see a lot of variations in the distribution in USA, which shows higher spread of income between rich and poor. Lower standard deviation in this case means, on average, people do not have high spread of income (similar income range).

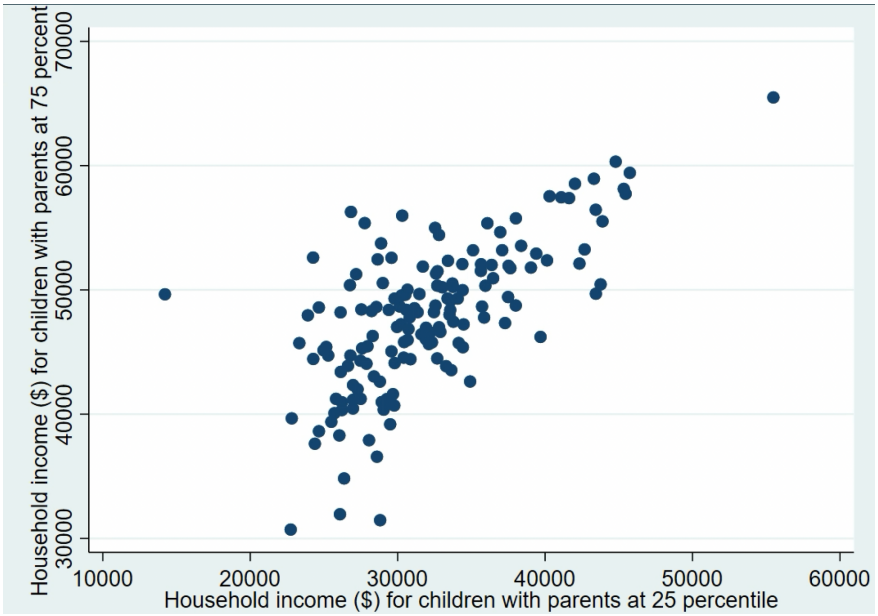
4.
Now let's turn to downward mobility: repeat questions (2) and (3) looking at children who start with parents at the 75th and 100th percentiles. How do the patterns differ?

```
. sum kfr_pooled_p75 if state == 35 & county == 001

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
kfr_poole~75 |        153    47888.08    5834.451   30719.96   65488.57

. sum kfr_pooled_p75 if state == 35

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
kfr_poole~75 |        497    48760.01    7989.179   15617.24   75343.16

. sum kfr_pooled_p75

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
kfr_poole~75 |     72,012    51500.78    9491.954          0     137454

.
```

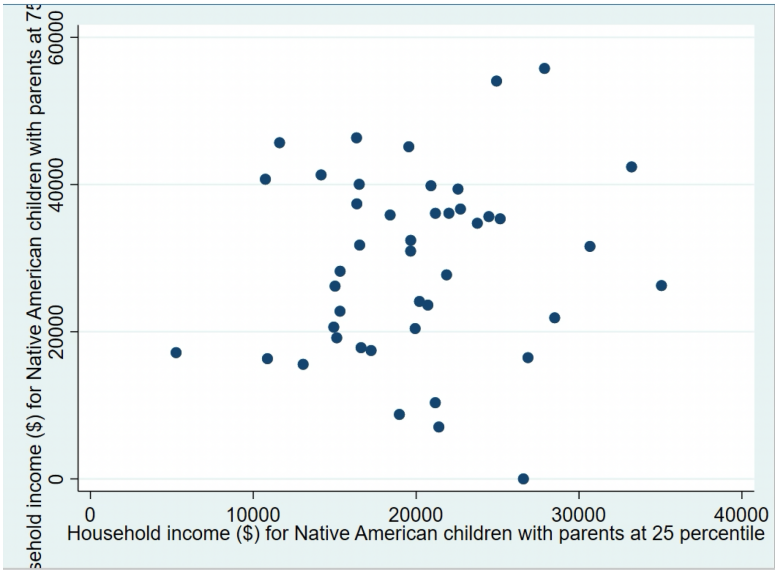
```
. sum kfr_pooled_p100 if state == 35 & county == 001

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
kfr_pool~100 |        153    64148.62     10137.9   33605.98    97300.2

. sum kfr_pooled_p100 if state == 35

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
kfr_pool~100 |        497    66377.85    16593.56   13103.33   171343.2

. sum kfr_pooled_p100

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
kfr_pool~100 |     71,968    69699.34    18074.19          0     980579
```

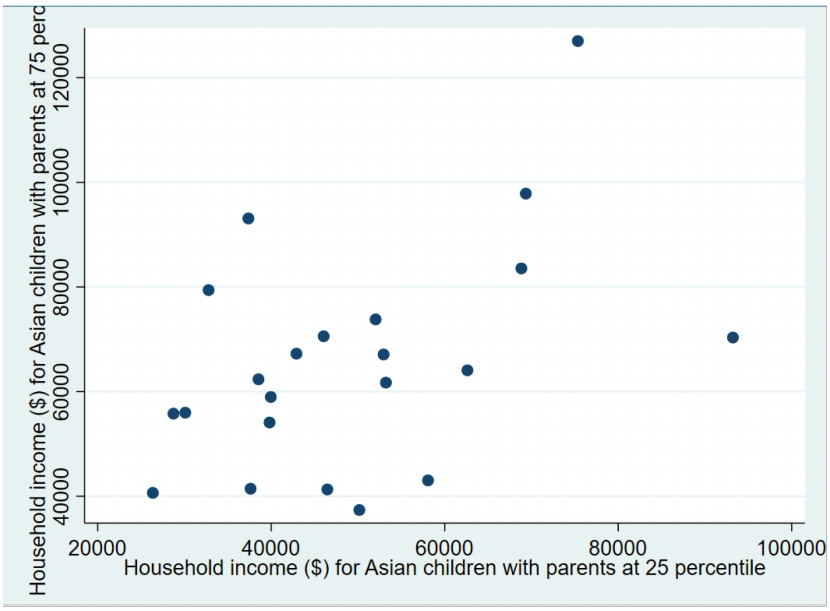
5. Using a linear regression, estimate the relationship between outcomes of children at the 25th and 75th percentile for the Census tracts in your home county. Generate a scatter plot to visualize this regression (refer to Table 2 for guidance). Do areas where children from low-income families do well generally have better outcomes for those from high-income families, too?

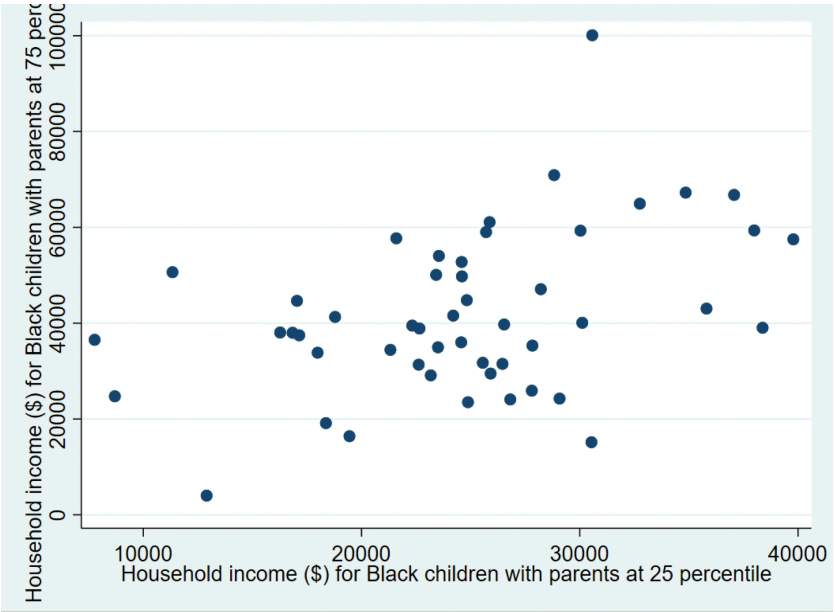
Yes, children from low-income families do generally have better outcomes in high-income areas.

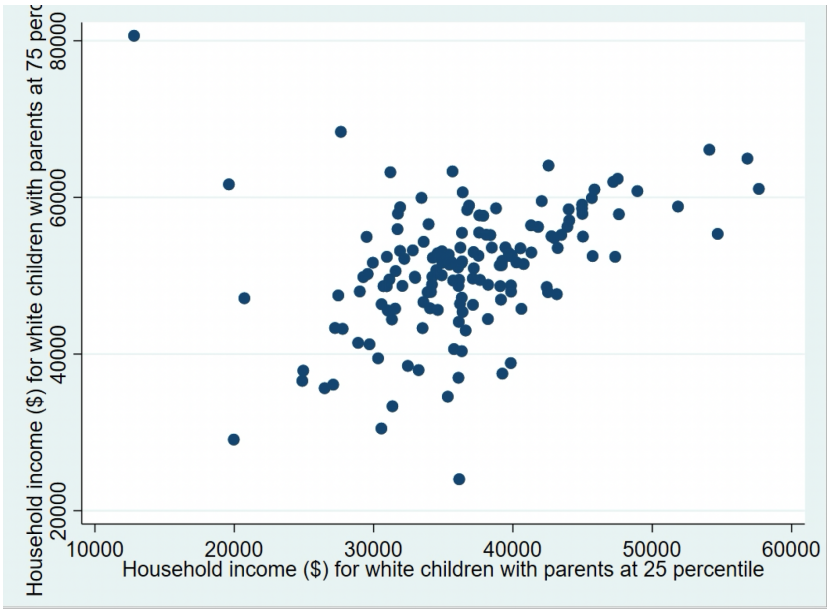


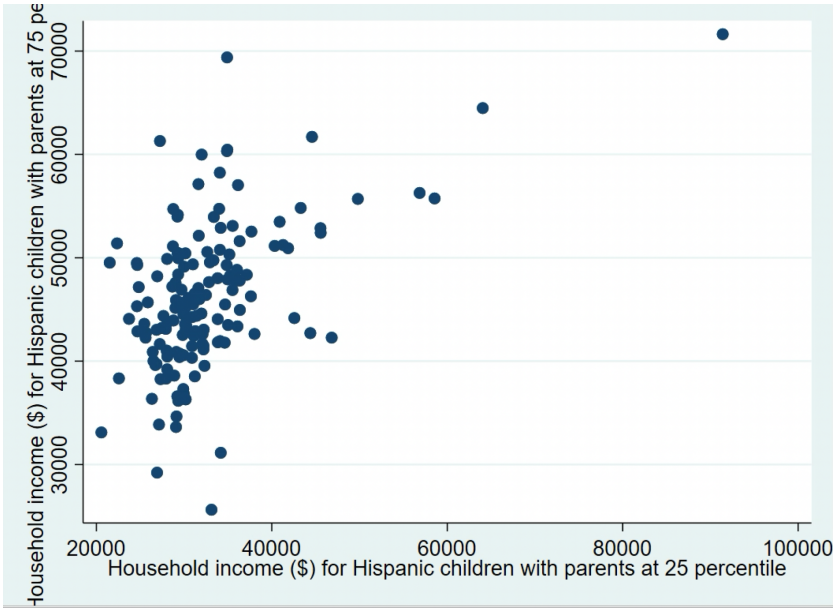
6. Next, examine whether the patterns you have looked at above are similar by race. If there is not enough racial heterogeneity in the area of interest (i.e., data is missing for most racial groups), then choose a different area to examine.




Household income ($) for Asian children with parents at 75 perc
40000
60000
80000
100000
120000
20000
40000
60000
80000
100000
Household income ($) for Asian children with parents at 25 percentile


Household income ($) for Black children with parents at 75 perc
0
20000
40000
60000
80000
100000
10000
20000
30000
40000
Household income ($) for Black children with parents at 25 percentile


Household income ($) for white children with parents at 75 perc
20000
40000
60000
80000
10000
20000
30000
40000
50000
60000
Household income ($) for white children with parents at 25 percentile



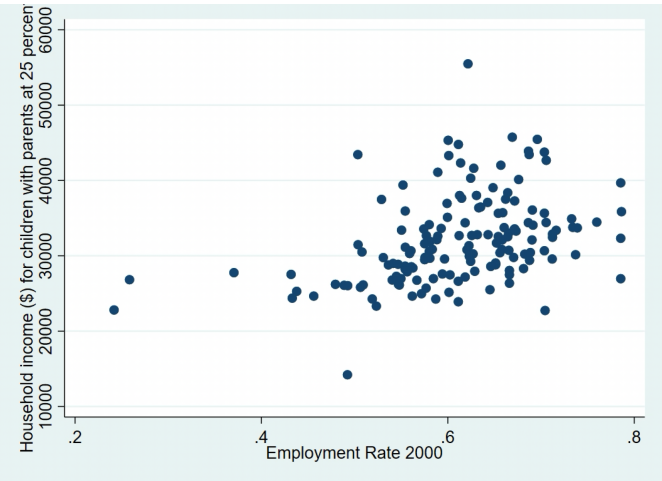
When analyzing the outcomes by race:

- For Native American children, the scatterplot shows varied outcomes, with some experiencing growth within high-income neighborhoods, while others do not.
- For Asian children, the scatter plot demonstrates that they are the demographic minority and they do not provide us significant data in terms of household income.
- For Black children, the results are varied and one cannot say if a high-income neighborhood equivocates growth in this demographic.
- For White children, the scatterplot shows a similar relationship between parents income at 25 and 75 percentile.
- For Hispanic children, the scatter plot is skewed towards low-income parents at the 25th percentile, however for the 75th percentile, there is a large scale of different incomes.
- Overall, the household incomes of children with parents in the 25th percentile versus the 75th percentile have varied outcomes. When the household incomes of children with parents in the 25th percentile and 75 percentile were analyzed by individual races, there was a lot of variation in their relationship.

7. Using the Census tracts in your home county, can you identify any covariates which help explain some of the patterns you have identified above? Some examples of covariates you might examine include housing prices, income inequality, fraction of children with single parents, job density, etc. For 2 or 3 of these, report estimated correlation coefficients along with their 95% confidence intervals.

- x-axis: employment rate in 2000
- y axis: income in 25th percentile

Code used: twoway(scatter kfr_pooled_p25 emp2000 if state == 35 & county == 001)



All the other variables were skewed heavily to either x-axis or y-axis. Even if we got a high correlation for the observation, it was misleading and did not span the graph. Therefore, we chose the employment rate in 2000 (x-axis) and pooled income in the 25th percentile (y-axis) as our variables even if its correlation was 0.3924 (slight positive correlation).

8. Putting together all the analyses you did above, what have you learned about the determinants of economic opportunity where you grew up? Identify one or two key lessons or takeaways that you might discuss with a policymaker or journalist if asked about your hometown. Mention any important caveats to your conclusions; for example, can we conclude that the variable you identified as a key predictor in the question above has a causal effect (i.e., changing it would change upward mobility) based on that analysis? Why or why not?

After conducting different types of data analysis on Albuquerque, New Mexico, it would be crucial for a policymaker or journalist to know the correlations between income, race, and gender. While there is some positive association between low-income households experiencing growth in high-income neighborhoods, it doesn't change the socioeconomic realities of gender and race. Only after tackling issues such as the gender pay gap would upward mobility be truly achievable.

Figure 1
Household Income in Adulthood for Children Raised in Low-Income Households in Milwaukee, WI

Notes: This figure shows household income at ages 31-37 for low income children who grew up in Census tracts near Milwaukee, WI. The image was saved from www.opportunity-atlas.org by first searching for "Milwaukee, WI" and then clicking on the "download as image" button.

Figure 2
Incarceration Rates for Black Men Raised in the Lowest-Income Households in Los Angeles, CA

Notes: This figure is from the non-technical summary of the Opportunity Atlas and was discussed in Lecture 2.

DATA DESCRIPTION, FILE: atlas.dta

The data consist of n = 73,278 U.S. Census tracts. For more details on the construction of the variables included in this data set, please see Chetty, Raj, John Friedman, Nathaniel Hendren, Maggie R. Jones, and Sonya R. Porter. 2018. "The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility." NBER Working Paper No. 25147.

Table 1
Definitions of Variables in atlas.dta

Variable name	**Label**	**Obs.**
(1)	(2)	(3)
1. Geographic identifiers		
tract	Tract FIPS Code (6-digit) 2010	73,278
county	County FIPS Code (3-digit)	73,278
state	State FIPS Code (2-digit)	73,278
cz	Commuting Zone Identifier (1990 Definition)	72,473
2. Characteristics of Census tracts		
hhinc_mean2000	Mean Household Income 2000	72,302
mean_commutetime2000	Average Commute Time of Working Adults in 2000	72,313
frac_coll_plus2010	Fraction of Residents with a College Degree or More in 2010	72,993
frac_coll_plus2000	Fraction of Residents with a College Degree or More in 2000	72,343
foreign_share2010	Share of Population Born Outside the U.S.	72,279

med_hhinc2016	Median Household Income in 2016	72,763
med_hhinc1990	Median Household Income in 1999	72,313
popdensity2000	Population Density (per square mile) in 2000	72,469
poor_share2010	Poverty Rate 2010	72,933
poor_share2000	Poverty Rate 2000	72,315
poor_share1990	Poverty Rate 1990	72,323
share_black2010	Share black 2010	73,111
share_hisp2010	Share Hispanic 2010	73,111
share_asian2010	Share Asian 2010	71,945
share_black2000	Share black 2000	72,368
share_white2000	Share white 2000	72,368
share_hisp2000	Share Hispanic 2000	72,368
share_asian2000	Share Asian 2000	71,050
gsmn_math_g3_2013	Average School District Level Standardized Test Scores in 3rd Grade in 2013	72,090
rent_twobed2015	Average Rent for Two-Bedroom Apartment in 2015	56,607
singleparent_share2010	Share of Single-Headed Households with Children 2010	72,564
singleparent_share1990	Share of Single-Headed Households with Children 1990	72,196
singleparent_share2000	Share of Single-Headed Households with Children 2000	72,285
traveltime15_2010	Share of Working Adults w/ Commute Time of 15 Minutes Or Less in 2010	72,939
emp2000	Employment Rate 2000	72,344
mail_return_rate2010	Census Form Rate Return Rate 2010	72,547

ln_wage_growth_hs_grad	Log wage growth for HS Grad., 2005-2014	51,635
jobs_total_5mi_2015	Number of Primary Jobs within 5 Miles in 2015	72,311
jobs_highpay_5mi_2015	Number of High-Paying (>USD40,000 annually) Jobs within 5 Miles in 2015	72,311
nonwhite_share2010	Share of People who are not white 2010	73,111
popdensity2010	Population Density (per square mile) in 2010	73,194
ann_avg_job_growth_2004_2013	Average Annual Job Growth Rate 2004-2013	70,664
job_density_2013	Job Density (in square miles) in 2013	72,463
3. Measures of Upward Mobility from the Opportunity Atlas		
kfr_pooled_p25	Household income ($) at age 31-37 for children with parents at the 25th percentile of the national income distribution	72,011
kfr_pooled_p75	Household income ($) at age 31-37 for children with parents at the 75th percentile of the national income distribution	72,012
kfr_pooled_p100	Household income ($) at age 31-37 for children with parents at the 100th percentile of the national income distribution	71,968
kfr_natam_p25	Household income ($) at age 31-37 for Native American children with parents at the 25th percentile of the national income distribution	1,733
kfr_natam_p75	Household income ($) at age 31-37 for Native American children with parents at the 75th percentile of the national income distribution	1,728
kfr_natam_p100	Household income ($) at age 31-37 for Native American children with parents at the 100th percentile of the national income distribution	1,594
kfr_asian_p25	Household income ($) at age 31-37 for Asian children with parents at the 25th percentile of the national income distribution	15,434

kfr_asian_p75	Household income ($) at age 31-37 for Asian children with parents at the 75th percentile of the national income distribution	15,360
kfr_asian_p100	Household income ($) at age 31-37 for Asian children with parents at the 100th percentile of the national income distribution	13,480
kfr_black_p25	Household income ($) at age 31-37 for Black children with parents at the 25th percentile of the national income distribution	34,086
kfr_black_p75	Household income ($) at age 31-37 for Black children with parents at the 75th percentile of the national income distribution	34,049
kfr_black_p100	Household income ($) at age 31-37 for Black children with parents at the 100th percentile of the national income distribution	32,536
kfr_hisp_p25	Household income ($) at age 31-37 for Hispanic children with parents at the 25th percentile of the national income distribution	37,611
kfr_hisp_p75	Household income ($) at age 31-37 for Hispanic children with parents at the 75th percentile of the national income distribution	37,579
kfr_hisp_p100	Household income ($) at age 31-37 for Hispanic children with parents at the 100th percentile of the national income distribution	35,987
kfr_white_p25	Household income ($) at age 31-37 for white children with parents at the 25th percentile of the national income distribution	67,978
kfr_white_p75	Household income ($) at age 31-37 for white children with parents at the 75th percentile of the national income distribution	67,968
kfr_white_p100	Household income ($) at age 31-37 for white children with parents at the 100th percentile of the national income distribution	67,627
3. Counts of number of children under 18 in 2000 (to calculate weighted summary statistics)		

count_pooled	Count of all children	72,451
count_white	Count of White children	72,451
count_black	Count of Black children	72,451
count_asian	Count of Asian children	72,451
count_hisp	Count of Hispanic children	72,451
count_natam	Count of Native American children	72,451

Note: This table describes the variables included in the atlas.dta file.

Table 2

STATA Hints

STATA command	**Description**
**clear the workspace* *clear* *set more off* *cap log close* **change working directory and open data set* *cd "C:\Users\gbruich\Ec1152\Projects\"* *use atlas.dta*	This code shows how to clear the workspace, change the working directory, and open a Stata data file. To change directories on either a mac or windows PC, you can use the drop down menu in Stata. Go to file -> change working directory -> navigate to the folder where your data is located. The command to change directories will appear; it can then be copied and pasted into your .do file.
**Summary stats* *sum yvar [aw = count_pooled]* **Summary stats for Wisconsin* *sum yvar if state == 55 [aw = count_pooled]* **Summary stats for Milwaukee County* *sum yvar if state == 55 & county == 079 [aw = count_pooled]* (Last two lines all go on one line in Stata)	These commands report means and standard deviations for *yvar*, weighted by the variable *count_pooled*. The first line calculates these statistics across the full sample. The second line calculates these statistics for observations in Wisconsin. The third line calculates these statistics for observations in Milwaukee County.
reg yvar xvar1 xvar2 xvar3, robust	This command estimates an OLS regression of *yvar* against *xvar*1, *xvar*2, and *xvar*3, using heteroskedasticity-robust standard errors.

**Report correlation coefficients* **Method 1* *sum yvar* *gen y_std = (yvar - r(mean))/r(sd)* *sum xvar* *gen x_std = (xvar - r(mean))/r(sd)* *reg y_std x_std , robust* **Method 2* *corr yvar xvar*	These commands show two methods for estimating correlation coefficients. The first block of code shows how to first generate standardized versions of the variables *yvar* and *xvar* by subtracting from each its mean and then dividing each by its variance (which are stored temporally by Stata as r(mean) and r(sd)). The last line reports an OLS regression of these transformed variables, with heteroskedasticity robust standard errors. The second method is to use the *corr* command, which does not report standard errors.
twoway (scatter yvar xvar) *graph export figure1.png, replace*	This pair of commands first draws a scatter plot of *yvar* against *xvar*. The second line saves the graph as a .png file. Also see this tutorial on graphs in Stata.