Spring 2023

Professor Raj Chetty and Dr. Gregory Bruich
Dept. of Economics, Harvard University

Empirical Project: Stories from the Opportunity Atlas

Part 1 due date: 11:59 p.m. March 1, 2023
Part 2 due date: 11:59 p.m. April 14, 2023

PROJECT DESCRIPTION

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. In summer 2020, the Atlas de Oportunidades was released for Spain, with an accompanying article in *El País*. As shown in Lecture 1, Opportunity Atlases have now been constructed for countries on every continent, except Antarctica.

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 analysis for the New York City neighborhood of Brownsville in Brooklyn.

In this 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 use the data for Spain or select a community of your choice.)

This project consists of two parts. The first part is a structured analysis where you will conduct specific analyses and answer specific questions, focusing on the methods for descriptive data analysis that you learned in Lab 1 and Lab 2. The second part of this project is a 15 page paper or memo (double spaced including references, graphs, maps, and tables) in which you describe what you have learned from the Opportunity Atlas, and is described in more detail on the following pages.

TABLE OF CONTENTS: LIST OF FIGURES AND TABLES

Professor Raj Chetty and Dr. Gregory Bruich
Dept. of Economics, Harvard University

PART 1: EXPLORATORY DATA ANALYSIS (100 points)

1. (2 points) 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). Examples for Milwaukee, WI and Brownsville, NY are shown below.

   Save a .png of a map to share with the rest of the class using the Opportunity Atlas' "download as image" button. Join our Slack Workspace. Post your map to the **#ec50-s23-welcome** channel. Include a brief message saying your name and the community that you are showing to us. Please feel free to react (kindly and respectfully, please) to the posts of others. Also include your image in your solution write up.

2. (3 points) Now turn to the **atlas.dta** data set. An important data science skill is identifying and dealing with missing data. All your future data analyses should start in this way. (Indeed, many research groups give job applicants a data analysis task; sometimes they intentionally add missing data with the hopes that potential new hires will notice.) Missing data in Stata is a period . and in R is `NA`. Which variables have missing values? How many missing values do they have? Refer to Table 2 and Table 3 for Stata and R suggestions, respectively.

3. The variable **kfr_pooled_pooled_p25** corresponds exactly to Statistic 1: Absolute Mobility at the 25th Percentile that you calculated in Lab 2.

   a. (3 points) What are the units that these variables are measured in?

   b. (3 points) Do higher or lower values correspond to higher upward mobility?

   c. (3 points) Explain briefly why this statistic is estimated using a linear statistical model.

4. (3 points) 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?

5. (3 points) Report summary statistics (mean, standard deviation, minimum, and maximum) for **kfr_pooled_pooled_p25**. Include a table of the results in your solutions

6. (3 points) Why can **kfr_pooled_pooled_p25** be negative or above 100 in these data?

   Hint: think about the limitations of a *linear* statistical model.

7. (4 points) 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? To answer this question, compare the value of **kfr_pooled_pooled_p25** in your home Census tract to the mean in your state and the mean in the U.S. overall.

8. (4 points) 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?

9. In this question, we will explore the relationship between upward mobility and rent in your chosen community, similar to the analysis from lecture for Seattle, WA.

    a. (3 points) Using data restricted to only the Census tracts in your home county (or if you prefer your home Commuting Zone), produce a regular scatter plot of **kfr_pooled_pooled_p25** versus **rent_twobed2015** with a linear best fit line. Include an image of your graph in your solutions.

    b. (3 points) 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. (4 points) Is your home Census tract an "Opportunity Bargain"? 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.

10. How have the neighborhoods in your chosen county (or if you prefer Commuting Zone) changed over the last 20 years?

    a. (3 points) Produce scatter plots with linear best fit lines of 2010 poverty rate versus 1990 poverty rate; 2010 fraction Black versus 2000 fraction Black; 2010 fraction Hispanic versus 2000 fraction Hispanic; 2010 fraction Asian versus 2000 fraction Asian; 2010 fraction white versus 2000 fraction white. Subset the data to just Census tracts in your county or Commuting Zone. Include images of your graphs in your solutions.

    b. (4 points) Describe in words how the neighborhoods in your graphs have changed from 1990-2000 when the children whose incomes we measure in the Opportunity Atlas were young (1978-1983 birth cohorts) compared to 2010.

11. I have also included historical information from the 1930s on the Home Owners' Loan Corporation (HOLC) "Redlining" grades for 9,276 Census tracts; 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 in Part 2. It is based on Greg and Raj's analysis of these data. We want to demonstrate in a concrete example how we think about formulating and testing hypotheses.

    a. (3 points) Report averages of **kfr_pooled_pooled_p25** for A, B, C, and D grade neighborhoods, corresponding to Census tracts with **HOLC_A**, **HOLC_B**, **HOLC_C**, and **HOLC_D** greater than 0.5 and nonmissing (i.e., `HOLC_A > 0.5 & HOLC_A != .` in Stata and `HOLC_A> 0.5 & !is.na(HOLC_A)` in R). 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?

b. (3 points) 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_black2000** 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. (3 points) Next, Raj and Greg 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. (3 points) Having found that these relationships persist even holding fixed race, we can 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) vegetation; (ii) extreme summer time temperatures; and (iii) fraction of land area developed. Report means of these variables (**vegetation**, **extreme_heat**, and **developed**) 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. (3 points) Produce "Opportunity Insights style" bar graphs to visualize the means that you reported in parts a-d. Include your graphs as images in your solutions.

12. 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 lecture). To measure air pollution, they construct "satellite-derived, high-resolution data on particulate matter smaller than 2.5 microns ($PM_{2.5}$) concentrations." They have generously provided these data, which I have merged with the Opportunity Atlas data.

a. (3 points) Using the $PM_{2.5}$ variables for 1982, 1990, 2000, and 2010, has air pollution improved or worsened on average in the United States over the last 40 years?

b. (3 points) How did $PM_{2.5}$ for your home Census tract in 1990 compare with the mean in your state and the mean in the U.S. overall? (If your home Census tract is in Hawaii or Puerto Rico, there is no pollution data; but you could pick a different area to compare for this question)

c. (3 points) Visualize the relationship between **kfr_pooled_pooled_p25** and **pm25_1990** using a binned scatter plot. Include an image of the graph in your solutions.

d. (3 points) Colmer, Voorheis, and Williams (2022) find that the *correlation coefficient* between **kfr_pooled_pooled_p25** and $PM_{2.5}$ across counties is -0.6. What is the *correlation coefficient* (not regression coefficient) between **kfr_pooled_pooled_p25** and **pm25_1990** across Census tracts? Why do you think the correlation coefficient might be smaller across tracts than across counties?

13. Using data restricted to only the Census tracts in your home county, Commuting Zone, or State (your choice), can you identify any other covariates which are strongly related to **kfr_pooled_pooled_p25**? Some examples of covariates you might examine include income inequality, fraction of children with single parents, job density, etc.

    a. (3 points) Visualize the relationship between 2 or 3 of these and **kfr_pooled_pooled_p25** using binned scatter plots. Include images of the graphs in your solutions.

    b. (3 points) Report estimated *correlation coefficients* (<u>not</u> regression coefficients) to quantify these relationships.

14. (4 points) Next, examine whether the patterns you have looked at in question 13a,b are similar by race and gender *if possible* by using the variables like **kfr_black_pooled_p25** and **kir_pooled_female_p25**. Include images of your bin scatter graphs in your solutions and also report *correlation coefficients*. You don't need to replicate 13a,b for every possible combination of race and gender, but look at 2-3 that interest you. Note also that the United States is a very segregated country; there may not be enough non-missing values to conduct separate analyses by race. If data is missing for most racial groups, then examine heterogeneity only by gender or choose a different area to examine. For example, you can study all the Census tracts in your home Commuting Zone (CZ) or State rather than county.

15. The files to submit for Part 1 of the Empirical Project are listed below. Unlike the labs, here you will submit your code to a separate Gradescope assignment from the write up. Further, when submitting the write up, you must carefully assign the pages to the specific questions.

    a. **(10 points)** Your well annotated do-file/.R script file replicating all your analyses above (with enough comments that a principal investigator on a research project would be able to follow and understand what each step of the code is doing). You can submit this to the "Project Part 1 – code" Gradescope assignment.

    b. For Stata users, a log-file with the log showing the output generated by your final do-file. You can submit this to the "Project Part 1 – code" Gradescope assignment.

    c. **(5 points)** PDF version of the solutions to the above questions to the "Project Part 1 – write up". For graphs, you can save them as .png files and insert them into the document. To receive full credit, please carefully assign each page of your PDF to the corresponding question in Gradescope. If you do not do this, you will not be able to see the comments that we write for you. There will also be a loss of points, because not assigning the pages to the questions makes it much more difficult for us to grade your assignment and delays when work can be returned to the class.

    d. From Question 1, your map from question 1 to the **#ec50-s23-welcome** Slack channel

Professor Raj Chetty and Dr. Gregory Bruich
Dept. of Economics, Harvard University

PART 2: MEMO ASSIGNMENT (100 points)

Compose a paper or memo synthesizing your exploratory data analysis and addressing the following questions. The narrative should reflect your creativity, so this part of the assignment is very open-ended. In the past the most successful narratives have combined personal experience and historical research with thorough data analysis to paint a vivid picture of their chosen areas and delve into the roots of local variation in opportunity and mobility. There is a rubric for Part 2.

Please note: In a "real-world" paper or memo on this topic, you would not include Stata or R code or copy/paste results from Stata/R inside the document (or even include the output as part of a Markdown document). *All empirical results used in the narrative should be presented in the text or, better, in nicely formatted tables and graphs.*

1. Based on what you see in your map, your exploratory analysis, and background research about the community that you have selected, state a *hypothesis* that might explain the variation in upward mobility for children who grew up in the Census tracts near your home 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 in Part 1.

2. The philosopher Karl Popper wrote, "*In so far as a scientific statement speaks about reality, it must be falsifiable; and in so far as it is not falsifiable, it does not speak about reality.*" Test your hypothesis as best you can in these data. You might start by providing some correlational evidence and then further investigate those patterns by reasoning by conditioning on race or gender. For example, Greg and Raj were interested in whether upward mobility for children born in the 1980s depends on the HOLC maps drawn in the 1930s, and then further testing whether those overall patterns were due to the causal effect of neighborhoods or compositional differences. They had a particular causal mechanism in mind.

   For this question, many covariates have been provided to you in the atlas.dta file, which are described under the "Characteristics of Census tracts" header in Table 1. You are welcome to use outside data that are not included in atlas.dta, but this is *not* required. Diane Sredl has created a research guide for our class that contains links to other data sources. You may wish to read this tutorial on how to add variables to a data set in Stata, or this tutorial on how to add variables to a data set using join in R.

3. 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?

4. Finally, describe **two** methods (or approaches) that you learned in Economics 50 since January that would allow you to further test your hypothesis more definitively. At least **one of the two** methods (or approaches) should allow you to move beyond association in order to establish causal relationships using a quasi-experimental research design.

   You do not have to carry out the analysis. Instead, be as specific as possible in outlining what you would do, being sure to describe any additional data that you might need to apply those methods (or approaches). For example, Colmer, Voorheis, and Williams (2022) exploit the Clean Air Act to show that air pollution has a causal effect on children's long-run outcomes.

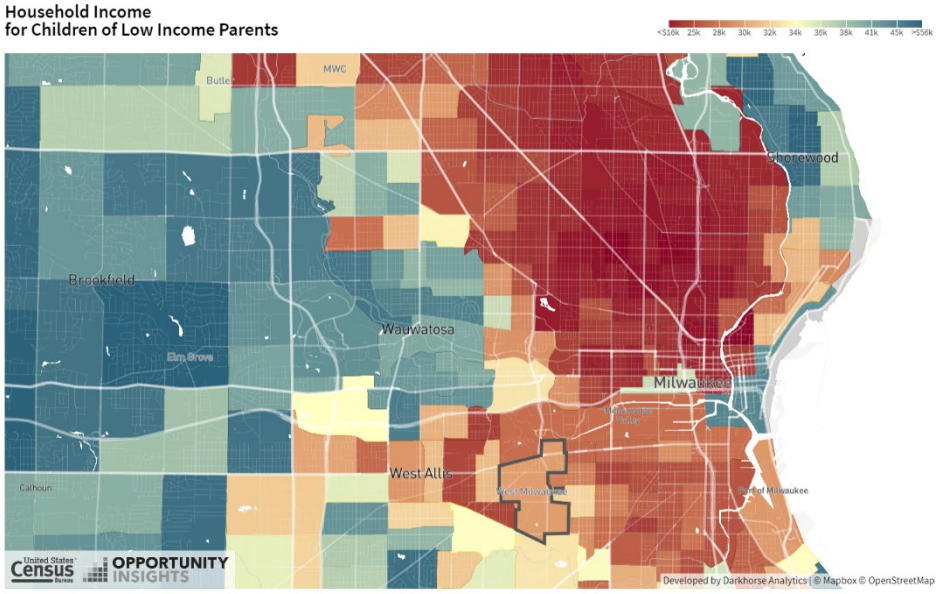
5. The deliverables for Part 2 of this Empirical Project are as follows. Unlike the labs, here you will submit your code to a separate Gradescope assignment from the write up. Further, when submitting the write up, you must assign the pages to questions in Gradescope.

    a. Your well annotated do-file/.R script replicating all your data constructions and empirical analyses for Part 2 (with enough comments that a principal investigator on a research project would be able to follow and understand what each step of the code is doing). You can submit this to the "Project Part 2 – code" Gradescope assignment. There are 10 points allocated for this under the "Analysis" rubric criteria.

    b. For Stata users, a log-file with the log showing the output generated by your final do-file. You can submit this file to the same "Project Part 2 – code" Gradescope assignment.

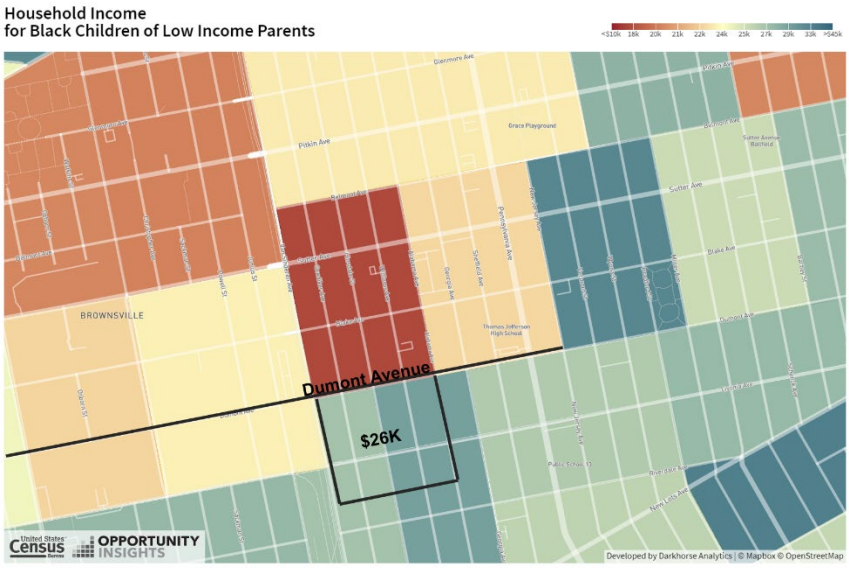
    c. A PDF of a document of your memo submitted to the "Project Part 2 – memo" Gradescope assignment. The paper should come to 15 pages (double spaced including references, graphs, maps, and tables). Please do not exceed 15 pages. There is a rubric for Part 2, which describes the minimum standards for each area under which your work will be assessed.

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
Household Income for Black Men Raised in Low-Income Households in Brownsville, NY



Notes: This figure is based on Jasmine Garsd's analysis and was discussed in Week 1.

Professor Raj Chetty and Dr. Gregory Bruich
Dept. of Economics, Harvard University

RUBRIC FOR PART 2

Concerns Deductions for areas that could be improved	**Criteria** Minimum Standards for Empirical Project Up to 80 points for meeting minimum standards	**Exceeds Standards** Up to 20 points for excellent work
	Subject Matter Expertise: Discussion of the data, methods, and background information for the chosen community are without any major factual errors (15 points)	
	Analysis: student successfully retrieves relevant data, conducted appropriate analyses, and the analyses are reproducible based on a *well commented* .do file or .R script (10 points)	
	Interpretation: The interpretation of the data analysis is scientific, objective, accurate, and avoids confirmation bias[1] (15 points)	
	Caveats: Major limitations of the data and analyses are recognized (10 points)	
	Written report: Presentation of data and writing is clear, project is visually appealing, and without major typographical errors. All empirical results are either in text, formatted tables, or graphs. (10 points)	
	Discussion of two methods, at least one of which uses a quasi-experimental research design, is logical. Data requirements for the proposed analyses are correct and clear. (20 points)	

[1] Confirmation bias is "the seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand" (Nickerson 1998)

Professor Raj Chetty and Dr. Gregory Bruich
Dept. of Economics, Harvard University

DATA DESCRIPTION, FILE: atlas.dta

The data consist of *N* = 73,199 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.

The data contain historical information from the 1930s on the Home Owners' Loan Corporation (HOLC) "Redlining" grades for 9,276 Census tracts; the grades are missing for Census tracts that have no overlap with neighborhoods that were rated by HOLC. These data were generously provided by Professors Daniel Aaronson and Daniel Hartley at the Federal Reserve Bank of Chicago. For more background on the HOLC Redlining, see Aaronson, Hartley, and Mazumder (2021).

The data also contain air pollution measures in 1982, 1990, 2000, and 2010 constructed from satellite-derived, high-resolution data on particulate matter smaller than 2.5 microns ($PM_{2.5}$) concentrations. There are 1,215 Census tracts that do not have $PM_{2.5}$. These Census tracts consist primarily of areas outside North America (e.g., Hawaii, Puerto Rico) that were not included in the original satellite-derived, high-resolution data from Meng et al. (2019). These data were generously provided by Professors Jonathan Colmer at the University of Virginia and John Voorheis at the U.S. Census Bureau. See Colmer, Voorheis, and Williams (2022) for more details on the data.

Table 1
Definitions of Variables in atlas.dta

Variable (1)	Description (2)	Obs. (3)
1. Geographic identifiers (See the Census Bureau's Understanding Geographic Identifiers for information on FIPS Codes)		
tract	Six-digit tract 2010 FIPS code. A Census tract roughly corresponds to a neighborhood: roughly 4,000 people live in a Census tract.	73,199
county	Three-digit county 2010 FIPS code. Each state in the U.S. is divided into Counties, parishes, or boroughs. Each county is assigned a three digit numeric identifier. A County is much larger than a Census tract. There are 3,242 counties in the United States.	73,199
state	Two-digit state 2010 FIPS code. Each state is assigned a two digit numeric identifier.	73,199
tract_name	String variable consisting of the name of the Census tract.	73,199
cz	Five-digit 1990 commuter zone code. Commuting zones are geographical aggregations of counties that are similar to metro areas but cover the entire U.S., including rural areas. Commuting zones are meant to consist of local labor markets where people both live and work.	73,199
czname	String variable consisting of the name of the commuting zone.	73,199
2. Home Owners' Loan Corporation (HOLC)'s historical "Redlining" grades *(Calculations by me based on data provided by Aaronson, Hartley, and Mazumder (2021))*		
HOLC_A	Fraction of Census Tract's total area that overlaps with neighborhoods rated by HOLC as Grade A	9,276

HOLC_B	Fraction of Census Tract's total area that overlaps with neighborhoods rated by HOLC as Grade B	9,276
HOLC_C	Fraction of Census Tract's total area that overlaps with neighborhoods rated by HOLC as Grade C	9,276
HOLC_D	Fraction of Census Tract's total area that overlaps with neighborhoods rated by HOLC as Grade D	9,276
	Note: Shares may add up to more than 100%. Not all cities had HOLC maps drawn, in which case missing values are assigned (". " in stata, NA in R)	
3. Environmental Data, largely derived from high resolution satellite data.		
pm25_1982	Concentration of ambient fine particulate matter ($PM_{2.5}$) in 1982 as constructed by Colmer et al. (2022).	71,903
pm25_1990	Concentration of ambient fine particulate matter ($PM_{2.5}$) in 1990 as constructed by Colmer et al. (2022)..	71,903
pm25_2000	Concentration of ambient fine particulate matter ($PM_{2.5}$) in 2000 as constructed by Colmer et al. (2022).	71,903
pm25_2010	Concentration of ambient fine particulate matter ($PM_{2.5}$) in 2010 as constructed by Colmer et al. (2022).	71,903
vegetation	Normalized Difference Vegetation Index (NDVI) relative to baseline NDVI. The variable comes from Benz and Burney (2021). Normalized Difference Vegetation Index (NDVI) quantifies vegetation by measuring the difference between near-infrared (which vegetation strongly reflects) and red light (which vegetation absorbs). Benz and Burney (2021) take the difference between the NDVI and a baseline NDVI measure and then aggregate to Census tracts. See their paper for more details.	70,668
extreme_heat	Extreme summer daytime surface urban heat (°C relative to baseline). The variable comes from Benz and Burney (2021). They measure summertime land surface temperatures (LSTs) between June 1 and August 31 for the 5 years 2010–2014. Benz and Burney (2021) take the difference between the LST and a baseline LST measure and then aggregate to Census tracts. See their paper for more details.	70,668
developed	Fraction of land area developed which is the total developed land area divided by the total land area in a Census tract. The variable comes from Benz and Burney (2021) and is constructed from the USGS National Land Cover Database.	70,668
4. Characteristics of Census tracts		
hhinc_mean2000	Mean Household Income 2000 (2015 dollars). Obtained from 2000 Decennial Census.	72,306
mean_commutetime2000	Average Commute Time of Working Adults in 2000 (minutes). Mean commute time for workers over 16 years old in the tract, as measured in the 2000 Decennial Census.	72,317
frac_coll_plus2000	Fraction of Residents w/ a College Degree or More in 2000	72,347

	Number of people aged 25 or older who have a bachelor's degree, master's degree, professional school degree, or doctorate degree, divided by the total number of people aged 25 or older in a tract. We use the 2000 Census to obtain the estimate for 2000, and the 2006-2010 ACS to obtain the estimate for 2010.	
frac_coll_plus2010	Fraction of Residents w/ a College Degree or More in 2010 Number of people aged 25 or older who have a bachelor's degree, master's degree, professional school degree, or doctorate degree, divided by the total number of people aged 25 or older in a tract. We use the 2000 Census to obtain the estimate for 2000, and the 2006-2010 ACS to obtain the estimate for 2010.	72,997
foreign_share2010	Share of Population Born Outside the U.S. in 2006-2010 ACS Number of foreign born residents in the 2010 Census divided by the sum of native and foreign born residents. Obtained from the ACS 2006-2010	72,283
med_hhinc1990	Median Household Income in 1990 (2015 dollars) Median household income. The data for 1990 is measured in the 1990 Census, while the data for 2016 comes from the 2012-2016 American Community Survey.	72,317
med_hhinc2016	Median Household Income in 2016 (2015 dollars) Median household income. The data for 1990 is measured in the 1990 Census, while the data for 2016 comes from the 2012-2016 American Community Survey.	72,763
popdensity2000	Population Density (per square mile) in 2000 Number of residents per square mile, calculated by dividing the total tract level population in the Decennial Census from 2000 and 2010 with tract land area given in square miles from the 2010 Census Gazetteer Files	72,473
poor_share2010	Share Below Poverty Line 2006-2010 ACS Share of individuals in the tract below the federal poverty line, measured in the decennial Census of the relevant year for the 1990 and 2000 estimates, and measured in the 2006-2010 ACS for the 2010 estimate.	72,937
poor_share2000	Poverty Rate 2000. Share of individuals in the tract below the federal poverty line, measured in the decennial Census of the relevant year for the 1990 and 2000 estimates, and measured in the 2006-2010 ACS for the 2010 estimate.	72,319
poor_share1990	Poverty Rate 1990. Share of individuals in the tract below the federal poverty line, measured in the decennial Census of the relevant year for the 1990 and 2000 estimates, and measured in the 2006-2010 ACS for the 2010 estimate.	72,327
share_white2010	Share White 2010 Racial Shares in the decennial Census. All races (except Hispanic) exclude Hispanics and Latinos	73,115

share_black2010	Share Black 2010 Racial Shares in the decennial Census. All races (except Hispanic) exclude Hispanics and Latinos	73,115
share_hisp2010	Share Hispanic 2010 Racial Shares in the decennial Census. All races (except Hispanic) exclude Hispanics and Latinos	73,115
share_asian2010	Share Asian 2010 Racial Shares in the decennial Census. All races (except Hispanic) exclude Hispanics and Latinos	71,949
share_black2000	Share Black 2000 Racial Shares in the decennial Census. All races (except Hispanic) exclude Hispanics and Latinos	72,372
share_white2000	Share White 2000 Racial Shares in the decennial Census. All races (except Hispanic) exclude Hispanics and Latinos	72,372
share_hisp2000	Share Hispanic 2000 Racial Shares in the decennial Census.	72,372
share_asian2000	Share Asian 2000 Racial Shares in the decennial Census. All races (except Hispanic) exclude Hispanics and Latinos	71,053
gsmn_math_g3_2013	Average School District Level Standardized Test Scores in 3rd Grade in 2013 (grade equivalent units – 3 would be right at third grade level, 1 would be first grade level, etc.) Mean 3rd grade math test scores in 2013. Obtained from the Stanford Education Data Archive and measured at the district level. We create a crosswalk from districts to tracts by weighting by the proportion of land area that a given school district covers in a tract.	72,090
rent_twobed2015	Median Rent for Two-Bedroom Apartment in 2015 (dollars). The median gross rent for renter-occupied housing units with two bedrooms that pay cash rent (from the 2011-2015 ACS)	56,607
singleparent_share2010	Share of Single-Headed Households with Children 2006-2010 ACS The number of households with females heads (and no husband present) or male heads (and no wife present) with own children under 18 years old present divided by the total number of households with own children present (1990 and 2000 estimates are from the decennial Census, and the 2010 estimate is from the 2006-2010 estimate).	72,568
singleparent_share1990	Share of Single-Headed Households with Children 1990 The number of households with females heads (and no husband present) or male heads (and no wife present) with own children under 18 years old present divided by the total number of households with own children present	72,200

	(1990 and 2000 estimates are from the decennial Census, and the 2010 estimate is from the 2006-2010 estimate).	
singleparent_share2000	Share of Single-Headed Households with Children 2000 The number of households with female heads (and no husband present) or male heads (and no wife present) with own children under 18 years old present divided by the total number of households with own children present (1990 and 2000 estimates are from the decennial Census, and the 2010 estimate is from the 2006-2010 estimate).	72,289
traveltime15_2010	Share of Working Adults w/ Commute Time of 15 Minutes Or Less in 2006-2010 ACS Share of workers 16 years and over who do not work at home whose commute is shorter than 15 minutes. Measured in the 2006-2010 ACS.	72,943
emp2000	Employment Rate 2000, defined as: (number employed / pop 16+ in and out of labor force) The rate of employment computed as total employed population (the sum of employed females and employed males) divided by the total population 16 years and over. Obtained from 2000 Decennial Census	72,348
mail_return_rate2010	Census Form Rate Return Rate 2010, defined as: (# mail forms completed in 2010/number of valid housing units expected to receive a form) The 2010 Census return rate, measured as the number of 2010 Census mail forms completed and returned over the number of valid occupied housing units where a Census form was expected to be delivered for mail return to Census. Taken from the Census planning database.	72,547
ln_wage_growth_hs_grad	Log wage growth for HS Grad., 2005-2014 (wage earnings_hs_grad/(weekly_hours_total*52); ln(wage_2005_2009) – ln(wage_2010_2014)) Wage growth for high school graduates. Wages are constructed by dividing the average high school graduate annual earnings by the product of overall average weekly hours worked and 52. High school graduate wage growth is then computed as the difference in logarithms between high school graduate wages in 2010-2014 and school graduate wages in 2005-2009. Wages are measured in the 2005-2009 and 2010-2014 American Community Surveys. Log differences are approximately equal to percentage changes measured in decimals.	51,636
jobs_total_5mi_2015	Number of Primary Jobs within 5 Miles in 2015 (total number of jobs in LEHD within 5 miles of tract centroid) Total number of jobs in own and neighboring tracts whose centroids fall within a radius of 5 miles from own tract centroid. Constructed using information from the Workplace Area Characteristics (WAC) data files in the	72,311

	LEHD Origin-Destination Employment Statistics (LODES) provided by the Census Bureau.	
jobs_highpay_5mi_2015	Number of High-Paying (>USD40,000 annually) Jobs within 5 Miles in 2015 (number of jobs in LEHD within 5 miles of tract centroid paying $3,333 / month or more) Number of jobs with earnings greater than $3,333 per month in own and neighboring tracts whose centroids fall within a radius of 5 miles from own tract centroid. Constructed using LODES - WAC data files provided by the Census Bureau.	72,311
popdensity2010	Population Density (per square mile) in 2010	73,194
ann_avg_job_growth_2004_2013	Average Annual Job Growth Rate 2004-2013, defined as: (exp(1/9*(log(Jobs in tract in 2013)-log(Jobs in tract in 2004)))-1) Average annualized job growth rate over the time period 2004 to 2013. Constructed using LODES - WAC data files provided by the Census Bureau. Data unavailable for Massachusetts and Washington D.C.	70,668
job_density_2013	Job Density (in square miles) in 2013, defined as: (number of jobs in the LEHD in tract / land area in sq mi). Number of jobs per square mile in each tract. Constructed using LODES - WAC data files provided by the Census Bureau.	72,463
4. Measures of Upward Mobility from the Opportunity Atlas by race and gender		
4a. Statistic 1 Absolute Mobility at the 25th Percentile using Household Income as Income Concept		
kfr_pooled_pooled_p25	Pooling all racial/ethnic groups and genders	72,010
kfr_natam_pooled_p25	Native American, pooling all genders	1,735
kfr_asian_pooled_p25	Asian, pooling all genders	15,440
kfr_black_pooled_p25	Black, pooling all genders	34,088
kfr_hisp_pooled_p25	Hispanic, pooling all genders	37,618
kfr_white_pooled_p25	White, pooling all genders	67,978
4b. Statistic 1 Absolute Mobility at the 25th Percentile using Individual Income as Income Concept		
kir_pooled_female_p25	Female, pooling all racial groups	71,645
kir_pooled_male_p25	Male, pooling all racial groups	71,687
kir_natam_female_p25	Native American, female	849
kir_asian_female_p25	Asian, female	7,732
kir_black_female_p25	Black, female	25,338
kir_hisp_female_p25	Hispanic, female	25,880
kir_white_female_p25	White, female	65,216
kir_natam_male_p25	Native American, male	852
kir_asian_male_p25	Asian, male	8,087
kir_black_male_p25	Black, male	25,059
kir_hisp_male_p25	Hispanic, male	25,403
kir_white_male_p25	White, male	65,516
4c. Fraction incarcerated on April 1st, 2010 for children from families at the 25th percentile		
jail_pooled_pooled_p25	Pooling all racial/ethnic groups and genders	71,877
jail_natam_pooled_p25	Native American, pooling all genders	1,409
jail_asian_pooled_p25	Asian, pooling all genders	13,461

jail_black_pooled_p25	Black, pooling all genders	31,185
jail_hisp_pooled_p25	Hispanic, pooling all genders	34,731
jail_white_pooled_p25	White, pooling all genders	67,383
jail_pooled_female_p25	Female, pooling all racial groups	71,476
jail_pooled_male_p25	Male, pooling all racial groups	71,474
jail_natam_female_p25	Native American, female	694
jail_asian_female_p25	Asian, female	6,628
jail_black_female_p25	Black, female	23,270
jail_hisp_female_p25	Hispanic, female	23,923
jail_white_female_p25	White, female	64,482
jail_natam_male_p25	Native American, male	627
jail_asian_male_p25	Asian, male	6,831
jail_black_male_p25	Black, male	21,672
jail_hisp_male_p25	Hispanic, male	22,474
jail_white_male_p25	White, male	64,576

Note: This table describes the variables included in the atlas.dta file. Note that the race and ethnicity splits are coarse and obscure heterogeneity within these broad categories. The reason is that the Census Bureau adheres to the Office of Management and Budget's 1997 race and ethnicity standards, which specify five major race groups: White, Black or African American ("Black" here), American Indian or Alaska Native, Asian, and Native Hawaiian or Other Pacific Islander, and define two ethnic groups (Hispanic and non-Hispanic).

TABLE 2:
Suggested Stata Code

STATA command	Description
*clear the workspace clear all version 17 *change directory and open data set cd "C:\Users\gbruich\Ec50\Project 1\" use atlas.dta, clear *Display all variables in the data describe *Report detailed information on all variables codebook	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. The describe and codebook commands will report information on what is included in the data set loaded into memory. The codebook command will report information on the number of missing observations for each variable.
*Summary stats sum yvar *Summary stats for Wisconsin sum yvar if state == 55 *Summary stats for Milwaukee County sum yvar if state == 55 & county == 079	These commands report means and standard deviations for yvar. 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.
*Report correlation coefficient corr yvar xvar if state == 55 & county == 079	These commands show how to estimate correlation coefficients between the variables yvar and xvar
*Draw scatter plot (command all goes on one line) twoway (scatter yvar xvar if state == 55 & county == 079) (lfit yvar xvar if state == 55 & county == 079) *Save scatter plot graph export figure1.png, replace *Draw scatter plots with points labelled using tract ID twoway (lfit yvar xvar if state == 55 & county == 079) (scatter yvar xvar if state == 55 & county == 079, mcolor(gray%50) msize(small) msymbol(circle_hollow) mlabel(tract) mlabsize(vsmall) mlabcolor(gray%50)) *Save scatter plot graph export figure1_labels.png, replace	In the first block of code, this pair of commands first draws a scatter plot of yvar against xvar. The second line saves the graph as a .png file. The second block of code shows how to label each point in the scatter plot using the tract FIPS code. It also changes the color and uses partial transparency to make the plot more clear. Also see this tutorial on graphs in Stata.
*Install binscatter ssc install binscatter *Draw plot binscatter yvar xvar if state == 55 & county == 079 *Save graph graph export binscatter.png, replace	The first command installs binscatter, which only has to be done once. The second line draws a binned scatter plot of yvar against xvar, restricting the graph to observations in a particular state and county.
*Histogram histogram yvar graph export histogram_yvar.png, replace *Histogram, changing number of bins to 50 histogram yvar, bin(50) graph export histogram_yvar.png, replace	These commands create and save histograms of a variable "yvar" which is a placeholder for the name of a variable in your data set. The first line creates a histogram (letting Stata decide how many bins to use). The second line saves the graph as a .png file. The second block of code changes the options by adding ", bin(50)" which will override the default binning and group the data into 50 buckets. The easiest way to draw graphs in Stata is to use the drop down menu, as described here.

```
*Opportunity Insights style bar graph
clear all
set obs 4

gen x = _n
gen y = .
replace y = 40 in 1
replace y = 30 in 2
replace y = 20 in 3
replace y = 10 in 4

#delimit ;
twoway
(bar y x if x == 4, barwidth(0.6) fcolor(navy) lcolor(navy))
(bar y x if x == 3, barwidth(0.6) fcolor(navy) lcolor(navy))
(bar y x if x == 2, barwidth(0.6) fcolor(navy) lcolor(navy))
(bar y x if x == 1, barwidth(0.6) fcolor(navy) lcolor(navy))
, legend(off)
xtitle("HOLC grade")
ytitle("Title for y-axis variable")
graphregion(color(white)) bgcolor(white)
xsc(range(0.7 4.3)) ylab(0(10)55,nogrid)
xlab(1 "A: Best" 2 "B: Still Desirable" 3 "C: Declining" 4 "D:
Hazardous")
;
#delimit cr

graph export bar_graph_v1.png, replace
```

These commands show how to draw an Opportunity Insights style bar graph for four categories, building on Lab 3 and Lab 4. We start by clearing the workspace. We will make a data set with four observations using `set obs 4`.

Then we create a variable x that numbers the observations 1, 2, 3 and 4. Next we create a variable `y` that equals the height of the bars that we want to plot. The sample code puts 40 for the first observation, 30 for the second, 20 for the third, and 10 for the fourth.

Then we use the #delimit command to reset the character that marks the end of a command to a semi colon `;` and later set it back to a carriage return `cr`. We do this because the options for the graph are quite complicated and spill over onto multiple lines.

Everything from `twoway` through the semi colon in red is one command. We create the graph by overlaying four `bar` type twoway graphs. Most of the other lines are just formatting options. The ones in purple are what you might want to change: the colors for the bars, the y-axis label, and the range and increments on the y-axis of the graph. Everything else can stay the same.

The `graph export` command saves the graph.

```
*close any possibly open log-files
cap log close

*start a log file
log using milwaukee.log, replace

*commands go here

*close and save log file
log close
```

These commands show how to start and close a log file, which will save a text file of all the commands and output that appears on the command window in stata.

The first line is short for “capture log close” which will close any open log files, and otherwise just proceed to the next step.

Then the “log using milwaukee.log replace” starts the log file and changes the default in two ways. First, it changes the file type to have a .log file extension, which creates a plain text log file (which is readable in Gradesope so is important!). Second, it also adds the “, replace” option which will save over any other log file that has the same name. This is usually what you want.

The rest of your lab code can go below the “log using milwaukee.log, replace” line. At the end of your do-file you can include the last line which is “log close” which will close and save the log-file.

TABLE 3:
Suggested R Code

R command	Description

#clear the workspace
rm(list=ls()) # removes all objects from the environment
cat('\014') # clears the console
 	
#Install and load haven package
if (!require(haven)) install.packages("haven"); library(haven)
 	
#Change working directory and load stata data set
setwd("C:/Users/gbruich/Ec 50/Project 1")
atlas <- read_dta("atlas.dta")
 	
#Report detailed information on all variables
summary(atlas)
	This sequence of commands shows how to open Stata datasets in R. The first block of code clears the work space. The second block of code installs and loads the "haven" package. The third block of code changes the working directory to the location of the data and loads in atlas.dta. To change the working directory in R Studio, you can also use the drop down menu. Go to session -> set working directory -> choose working directory.
	
	The easiest way to open a Stata data set in R Studio is to use the drop down menu. Go to file, then import data set, and finally browse to locate the file you want to open. This option will be available after you install the haven package.
	
	The summary command will report information on what is included in the data set loaded into memory, including information on the number of missing observations NAs for each variable.

# summary stats, unweighted 	
summary(atlas$yvar)
mean(atlas$yvar, na.rm=TRUE)
sd(atlas$yvar, na.rm=TRUE)
	These commands show how to calculate unweighted summary statistics.

## subset observations to Wisconsin wisconsin <- subset(atlas,state == 55) ## subset observations to Milwaukee County milwaukee <- subset(atlas,state == 55 & county == 079)	These commands show how to subset the data to observations in only Wisconsin and in only Milwaukee county.

#Report correlation coefficients
cor(milwaukee$yvar, milwaukee$xvar)
	These commands show how to estimate correlation between the variables yvar and xvar

# Install and load ggplot2 package 	
install.packages("ggplot2")
library(ggplot2)
 	
# Draw scatter plot with linear fit line 	
ggplot(data = milwaukee) + geom_point(aes(x = xvar1, y = yvar)) +
  geom_smooth(aes(x = xvar, y = yvar), method = "lm", se = F)
 	
#Save graph as .png
ggsave("milwaukee_scatter.png")
 	
#Add labels to scatter plot
ggplot(data = milwaukee) + geom_point(aes(x = xvar1, y = yvar)) +
  geom_smooth(aes(x = xvar, y = yvar), method = "lm", se = F) +
  geom_text(aes(y =xvar, x =yvar, label=tract), check_overlap = TRUE, size = 3)
 	
#Save graph as .png
ggsave("milwaukee_scatter_labels.png")
 	
(continued on next page)
	These commands show how to draw a scatter plot of yvar against xvar. The geom_smooth part of the code adds an OLS regression line. The last line saves the graph as a .png file. See this page for more examples.
	
	The second block of code shows how to label each point in the scatter plot using the tract FIPS code. It uses the geom_text() layer. See this page for more examples.

# Install statar 	
install.packages("statar")
library(statar)
 	
#Draw binscatter plot with linear best fit line
ggplot(milwaukee, aes(x = xvar1, y = yvar)) +
  stat_binmean(n = 20) +
  stat_smooth(method = "lm", se = FALSE)
 	
#Save graph as a .png
ggsave("milwaukee_binscatter.png")
	These commands show how to draw a binned scatter plot of yvar against xvar. The last line saves the graph as a .png file. See this page for more examples.

#Histogram in base R
png("histogram_ yvar.png")
hist(atlas$yvar, probability = T)
dev.off()
 	
#Histogram using ggplot
if (!require(tidyverse)) install.packages("tidyverse"); library(tidyverse)
if (!require(ggplot2)) install.packages("ggplot2"); library(ggplot2)
 	
ggplot(atlas) + geom_histogram(aes(x=yvar, y=..density..))
ggsave("histogram_yvar.png")
 	
#Use 50 bins, overriding default
ggplot(atlas) + geom_histogram(aes(x=yvar, y=..density..), bins = 50)
	These commands create and save histograms of a variable "yvar" which is a placeholder for the name of a variable in your data set. The first line creates a histogram (letting R decide how many bins to use) using base R.
	
	The second block of code shows how to do this using ggplot. First start by installing the tidyverse library. Then use ggplot to draw the graph. The ggsave() line saves the graph as a .png file. See this page for more examples.
	
	The last line overrides the default to show 50 bins by adding the bins = 50 option.

#Bar graph
#Load tidyverse library
if (!require(tidyverse)) install.packages("tidyverse"); library(tidyverse)
 	
#Create a data frame with two columns
#Column 1 is the height of the two bars (in blue)
#Column 2 is the group names (in red)
df <- data.frame(c(**0.4, 0.3, 0.2, 0.1**),
  c("Grade A", "Grade B", "Grade C", "Grade D"))
 	
# Change name of 1st column of df to "yvar" 	
names(df)[1] <- **"yvar"**
 	
# Change name of 2nd column of df to "Group" 	
names(df)[2] <- **"HOLC"**
 	
# Bar graph displaying results 	
ggplot(data=df, aes(x=HOLC, y=**yvar**, fill=**HOLC**)) +
  geom_bar(stat="identity", show.legend = TRUE, width=.6) +
  scale_fill_manual(values=c("navy", "navy", "navy", "navy")) +
  labs(y = **"Title for y-axis variable"**, x = "")
 	
ggsave("fig1.png")
	These commands show how to draw an Opportunity Insights style bar graph for four categories, building on Lab 3 and Lab 4. We start by loading the tidyverse package.
	
	We will make a data frame with four observations and two columns. Column 1 is the height of the two bars (in blue font). Column 2 is the group names (in red font). For the first observation, I fill in 0.4 for the height of the first bar. The other values correspond to the height of the other bars.
	
	Then we give the first column the name "yvar" and the second column the name "HOLC" so that we can refer to these in the ggplot command.
	
	We use the geom_bar plot type in ggplot. The "identity" option says to plot the numbers in the data frame as is, as opposed to plotting some statistic computed for the data frame. The scale_fill_manual() code changes the color of the bars.