

City University of New York (CUNY)

CUNY Academic Works

Open Educational Resources

Lehman College

2019

CS + Sociology: Using Big Data to Identify and Understand Educational Inequality in America (2)

Joseph Cleary

CUNY Lehman College, joseph.cleary@lehman.cuny.edu

Elin Waring

CUNY Lehman College, elin.waring@lehman.cuny.edu

[How does access to this work benefit you? Let us know!](#)

More information about this work at: https://academicworks.cuny.edu/le_oers/25

Discover additional works at: <https://academicworks.cuny.edu>

This work is made publicly available by the City University of New York (CUNY).

Contact: AcademicWorks@cuny.edu

Title: CS + Sociology: Using Big Data to Identify and Understand Educational Inequality in America (Week 2 of 2)

Author/Affiliation: Joseph Cleary/Lehman College and Elin Waring/Lehman College

Date: 5/31/2019

Material Type: Lecture + Lab

CS + Sociology

Software/Equipment Dependencies: Computers for each student and instructor with R-Studio

Prior Knowledge Needed (if any): None

Keywords: big data, education, poverty, teacher-pupil ratio, computer science, sociology

Approximate time needed: 3 hours (for Week 2)

Description: This is the second part (second week) of a two-part (two week) lesson plan (i.e. lecture + lab) combining the teaching and learning of computer science and sociology. Students will develop CS skills and behaviors including but not limited to: learning what an API is, learning how to access and utilize data on an API, and developing their R coding skills and knowledge. Students will also learn basic, but important, sociological principles such as how poverty is related to educational opportunities in America (and how this relationship varies between states). Although prior knowledge of CS and sociology is helpful, neither is necessary for student (or instructor) success on this two-week project. Three instructional hours per week (total of six hours over two weeks).

Week 2

Note to instructors: This lesson can be modified depending on your and your students' needs. For example, you may substitute two other variables of your choosing instead of using the ones suggested herein (i.e. % student poverty and teacher-pupil ratio). In addition, feel welcome to also modify this lesson by eliminating certain steps or tweaking/expanding the homework assignment in order to fulfill your class's needs.

Introduction for Instructor

This exercise explores the use of “big data” to study a social issue important in the sociology of education, which is the relationship between the socioeconomic status of a place and the quality of the education available to students in that place. This week's lesson builds upon last week's lesson that investigated the relationship between poverty and the student-teacher ratio within the state of New York. This week's lesson investigates data on this relationship in six other states, as well as the country as a whole.

This exercise is written using the R programming language, but it can easily be modified to use Python or any language that has methods to access web APIs. It could also be modified, if the instructor is willing to create it, to use a relational database. It also could be modified using a spreadsheet.

The instructions are written as though students will be using RStudioServer (which is available at no cost for classroom use). However, R and RStudio can also be installed on individual computers. The lesson can also be done using R at the command line or in the R programming environment. However this is not how most data scientists would use R.

Although data science focused, this lesson also deliberately does not rely on students having knowledge of statistics since the target audience is general education students. For students with more statistical knowledge the exercise can be made more sophisticated, for example by using a linear model to analyze the data.

The design of this lab is such that it can be used either in the context of a “CS0” course or in the context of a sociology course that wants to incorporate coding and real world data analysis.

Requirements for Instructors Prior to This Lab

1. Students must have access to R and RStudio (default installations) with the following additional packages installed: dplyr, ggplot2, knitr, curl, devtools and their dependencies. If you intend to have students install these on their personal computers you will need to provide instructions for that. These are available on the www.r-project.org/ and www.rstudio.com/ websites, respectively. We do not recommend using the same class period to do this this exercise. Potentially you could use rdrr.io/snippets/ but we have not experimented with using this with multiple students at the same time. Using Jupyter notebooks instead of RStudio is also an option (e.g rnotebook.io/)

Our instructions assume that students will be able to install a package (the use of package structure in R and how to do the installation is part of what they will be learning). This means that you must insure that they have the ability to do this.

2. Create a complete set of instructions for logging into RStudioServer or another environment you have chosen.

3. Prepare any printed or online instructions. We have found success with both instructions embedded in Blackboard and printed instructions.

4. For background on the sociological content we suggest the following articles for the faculty:

- Achilles, C. M. (2005). Financing Class Size Reduction. *Regional Educational Laboratory Southeast*.
- Biddle, B. J., & Berliner, D. C. (2002). What research says about small classes and their effects.
- Biddle, B. J., & Berliner, D. C. (2008). Small class size and its effects. *Schools and society: A sociological approach to education*, 3, 86-95.
- Whitehurst, G. J., & Chingos, M. M. (2011). Class Size: What Research Says and What It Means for State Policy. *Brookings Institution*.

The following three website resources are particularly helpful and user-friendly, and **it is strongly recommended to show students each of these three sites:**

- <https://www.governing.com/gov-data/education-data/state-education-spending-per-pupil-data.html>
- https://www.huffpost.com/entry/states-student-teacher-ratio_n_6083480
- <https://www.publicschoolreview.com/average-student-teacher-ratio-stats/national-data/states>

5. If you are unfamiliar with R and are a computer science or information science faculty member or are experienced with other programming languages we suggest [Advanced R](#) as a useful introduction. If you are from a discipline that has traditionally used Stata, SPSS or SAS, a helpful resource is [R for Data Science](#). For both groups there are many other useful resources available.

6. The data we will be using relies on the API provided by the Urban Institute and documented here <https://educationdata.urban.org/documentation/> and here <https://github.com/UrbanInstitute/education-data-package-r> . We suggest reading over this material and the linked documentation.

Computer Science Topics

This lesson plan is relevant to a number of topics that might be included in a CS0 level course in which the students have had a few weeks of experience writing code. Some topics are:

1. The R language.
2. The concept of data structures.
3. Data science as an area that uses computer programming skills intensively.
4. The idea of big data.
5. The idea of a web server and a client application.
6. The idea of an API and specifically a Web API.
7. JSON as a way of representing an object in text format.
8. The idea of a list.
9. The idea of functions with parameters
10. The idea of a join using a key.

It is not essential that any one of these be covered in the lecture but any of them can be emphasized in the lab by modifying it.

Sociology Topics

This lesson incorporates a number of sociological concepts that are appropriate to sociology courses at various levels. They include:

1. Social inequality (i.e. high government funding vs. limited government funding).
2. Associations/relationships between two variables (X and Y).
3. Basic statistics (use of ratio rather than raw numbers).
4. Use of official (government) data in sociology.
5. Data visualization.

These topics may be introduced in more or less depth during the lecture. The most relevant idea in this week's lesson is the concept of government funding (at the state level) and the implications of high and low government funding on educational resources (in this particular case: the student-teacher ratio). The social implication of this is that people are not merely individual actors – their life opportunities are shaped by the resources provided by their social environments. Although some individuals will succeed and some will fail in both wealthy and impoverished environments, overall there may be general patterns making success easier in some social settings than others.

If you are not experienced in teaching sociology you may not want to go further than to say that this is a general pattern that sociological research has found. However, sociologists would also say that the strength of that pattern might vary from place to place depending on other factors (i.e. New York vs. California). This may become evident when comparing data from different states in today's lab. You may want to treat this simply as a way to explore the use of big data without going into details about the theory. The topic itself (the relationship of poverty rate in a school district to the student teacher ratio in that district) should be of interest to students from many majors.

Student Learning Outcomes

Students will be able to...

1. Identify and interpret available resources for gathering socioeconomic data.
2. Compare and contrast socioeconomic indicators of public high school success throughout New York City and reach conclusions regarding relevant data.
3. Utilize socioeconomic characteristics in a comparative context and identify conclusions regarding data.
4. Access data using Web APIs (using a computer script) and access data using a browser interface, and compare these two different approaches.
5. Utilize a web-based Integrated Development Environment (IDE) to create a simple script to obtain data that will be used as part of a larger project.
6. Explain what Web APIs are and how they are used (at a basic level).

Prelab Options for Instructors

Optionally, the instructor can install the Urban Institute education data package as a shared package prior to the lab. This will save some time, but students will not learn how to install packages.

```
install.packages('devtools') # if necessary
devtools::install_github('UrbanInstitute/education-data-package-r')
```

Optionally, the instructor can create the data sets prior to the class. This will save the time needed to download the data for each student. If the data are stored as an r data frame (rda) in an accessible location students will be able to skip the portions of the lab that have them do this. This will potentially save a considerable amount of time. Alternatively the students could save the data to a file using the `save(dataframe, file="path/to/filename.rda")` command. This way they will only download the file once instead of each time the full script is run.

In the template provided school district level data is used. This is enough data to make it obvious that using a spreadsheet or visual inspection of data will not be reasonable. However, if desired, state level data with its smaller number of observations may be used.

In this lab we are using a specific set of two variables, one from each data set. Other variables may be chosen, although consideration should be given to whether there are reasonable hypotheses about the relationship between them.

There are many interesting features of the data being used. For example <https://www.census.gov/programs-surveys/saipe/guidance-geographies/same-name/2017-2018.html> illustrates a common data science problem, the fact that names are not unique identifiers.

When the students first go into RStudio you may want to have them explore the environment by trying “R as calculator” (doing some simple arithmetic with R).

Students will create additional graphs for states besides New York and compare them. These can be stand-alone or they can be done using “small multiples” in a grid structure using facets. <http://www.cookbook-r.com/Graphs/> is a good source for information about how to create specific kinds of graphs.

Please note that graphing the entire data set can be slow and challenging because of the large number of data points. The graphs produced are very dense and do not show many individual points. The lab instructs students to produce data sets and graphs for seven states. After this, students are finally instructed to graph all national data on the relationship under examination (poverty and student-teacher ratio). Again, keep in mind (and inform students of this fact) that this step may be tedious given the large amount of data being retrieved.

Instructors should explore various options before using this lab and modify it as desired.

The entire prelab section below can be done before the start of the actual lab, either as homework or in the lecture period. RStudio supports the use of git and Github or BitBucket. We are not documenting that here, but there are resources on the RStudio website that explain how to do this.

Lecture

Introduction

- How do we get data from the web?
- How do the applications that let us access data on the web work?

- Government and organizations collect vast quantities of statistical data every year for various purposes. Today we will be exploring how web based requests for data work and how they can be used to explore sociological variables.
- In sociology it is common to do analysis of data. In news articles and school readings you may often see data visualizations that are created using sociological approaches. Today we are going to look at the use of data to explore variables about education.

Sociology Background Materials

This lesson is focused on having students further investigate the relationship between socioeconomic status and the quality of educational experiences across multiple states (as well as the country as a whole).

Socioeconomic Status - Overarching Questions

- Review questions:
 - o What is socioeconomic status?
 - o What is the socioeconomic status of a place?
 - o How does that differ from that for an individual?
- New questions:
 - o Why do some states have better public education systems than others?
 - o Why do some states provide more public funding to their education systems than others?

Instructor may first ask for volunteers to answer these questions.

Instructor will remind students that the first variable used in this lesson is the federal government's estimate of the percent of people ages 5 to 17 who are in households that are below the federal poverty level. We will call this the X-variable. Instructor will then remind students that the second variable used in this lesson is the "teacher-pupil ratio" (the Y-variable)[1]. Instructor and students will then briefly discuss the nature of the relationship between these two variables by focusing on results in last week's lesson (data from the state of New York). What did the ggplot look like?

Quality of Educational Experiences From State-to-State

The instructor may provide further review of last week's lesson as needed. If students appear to have a good grasp on why these two variables are related, you may move on to the next topic that is the focus of week 2's lesson:

- 1) Why do some states have better public schools than other states?
- 2) How does the relationship between X and Y (percentage of students living in poverty and the student-teacher ratio) vary from state-to-state? What might explain this variation? What does this mean?

Instructor now introduces students to a third variable, “Z”, which refers to “place”. In this investigation, place refers to any of the 50 states, the nation as a whole, and two additional areas that are part of the United States (Puerto Rico and Washington D.C.). **Instructor can now have a discussion with students about each of the above questions.**

Hypotheses (optional)

Instructor may now ask students to create a hypothesis (“A hypothesis is a prediction made prior to collection of data”) about the possible relationship between X, Y, and Z. As an example:

“We are more likely to observe a clear positive relationship between X (*percentage of students living in poverty*) and Y (*student-teacher ratio*), where increases in X are associated with increases in Y, in states (and areas) that have larger populations.”

Or:

“We are more likely to observe a clear positive relationship between X (*percentage of students living in poverty*) and Y (*student-teacher ratio*), where increases in X are associated with increases in Y, in states (and areas) that devote less state spending toward public education.”

Instructor and students will briefly discuss students’ rationales behind their proposed hypotheses that attempt to explain why some states/areas experience a stronger connection between student poverty and the student-teacher ratio.

Lab 1: R

Prelab

We will be using data found at this website: <https://educationdata.urban.org>

Find the Common Core of Data

- According to the description what data does it contain?
- What federal agency collects the Common Core data?

Find the SAIPE data

- According to the description what does SAIPE stand for?
- What federal agency collects the SAIPE data?

Now let’s look at what the data that is returned to our application from the Urban Institute looks like: <https://ed-data-portal.urban.org/api/v1/schools/ccd/directory/2013/>

Possibly wild looking, depending on your browser and operating system. But the data are actually highly structured and designed to be read by a computer application, not a human, but they are also somewhat human readable. This link is showing the Common Core Directory data for 2013. "count":102815 is an example of a key value pair, where "count" is the key and 102815. In this case there are 102815 records in the data set. JSON organizes data using a small number of punctuation marks, specifically : (colon), {} (curly brackets), [] (square brackets) and , (comma). If you look carefully at the page you will see that these follow very well defined patterns.

In case your browser hides the JSON with a nice display of the data, here is data for the first school in the file.

```
{ "year":2013,"ncessch":"010000200277","school_id":"00277","school_name":"SEQUOYAH SCH - CHALKVILLE CAMPUS","leaid":"0100002","lea_name":"ALABAMA YOUTH SERVICES","state_leaid":"210","seasch":"0020","street_mailing":"P O BOX 9486","city_mailing":"BIRMINGHAM","state_mailing":"AL","zip_mailing":"35220","street_location":"RT 2 OLD SPRINGVILLE RD","city_location":"PINSON","state_location":"AL","zip_location":"36126","phone":"205680 8574","fips":1,"latitude":33.673727,"longitude":-86.628716,"csa":null,"cbsa":null,"urban_centric_locale":21,"county_code":"01073","school_level":3,"school_type":4,"school_status":1,"lowest_grade_offered":7,"highest_grade_offered":12,"bureau_indian_education":0,"title_i_status":-1,"title_i_eligible":-1,"title_i_schoolwide":-1,"charter":-2,"magnet":0,"shared_time":0,"virtual":0,"teachers_fte":-1,"free_lunch":-1,"reduced_price_lunch":-1,"free_or_reduced_price_lunch":-1,"elem_cedp":0,"high_cedp":1,"middle_cedp":1,"ungrade_cedp":0,"enrollment":-1 }
```

JSON is a very efficient and compact way to send data, but we need ways for humans to read and nicely display the results and to do data analysis.

Lab

Login to RStudio using instructions provided by the instructor (these will potentially differ on various campuses).

If you have never used RStudio before, look around the environment. Try typing some simple math in the console.

To access the data we want, we need to install a special R package for working with the Urban Institute data.

At the command prompt, type this command.

```
devtools::install_github('UrbanInstitute/education-data-package-r')
```

(You can see the instructions to do this on the Urban Institute page)

Creating your RMarkdown file

Create a new file by typing at the command line:

```
file.create("education_analysis.rmd")
```

You should then see the file in the Files tab on the right of your screen.

Click on the file to open it. Notice that the f in file must be lower case.

Go to this link (which corresponds with the attachment titled, "education_anaysis_lab2.Rmd"):

<https://gist.github.com/elinw/be3252de719c51273a89f2acdcd7c38e>

****IMPORTANT:*** *If you are starting with this lesson – which is to say if the class did not complete Week 1’s lesson – then have students go to this website instead (which corresponds with the attachment titled, “education_analysis_base.Rmd”):*

<https://gist.github.com/elinw/b001a9bb3b95df857d864dcb4f95fd37>

Now click on the “raw” button and copy and paste the full text into your empty file (follow this instruction regardless of which of the above two links you instruct students to click/utilize).

Save the file.

From this point on follow the instructions in the file.

GG Plots

The ggplot2 package, created by Hadley Wickham, offers a powerful graphics language for creating elegant and complex plots. Its popularity in the R community has exploded in recent years. Origanlly based on Leland Wilkinson's The Grammar of Graphics, ggplot2 allows you to create graphs that represent both univariate and multivariate numerical and categorical data in a straightforward manner. Grouping can be represented by color, symbol, size, and transparency. The creation of trellis plots (i.e., conditioning) is relatively simple.

Mastering the **ggplot2** language can be challenging (see the **Going Further** section below for helpful resources). There is a helper function called **qplot()** (for quick plot) that can hide much of this complexity when creating standard graphs. For more information on GG Plots (including but not limited to coding instructions), visit:

<https://www.statmethods.net/advgraphs/ggplot2.html>

Source: <https://www.statmethods.net/advgraphs/ggplot2.html>

Post lab

Instruct students to knit/download their work and save the results to their desktops. (Some students may need help with this if they do not fully understand the difference between the web and their individual computer.)

Encourage students to save a copy of the work in cloud storage or to email a copy to themselves. If the class is using RStudioServer the results will also save on their account.

Collect the results in the desired format (print, upload to Blackboard, other submission methods).

Homework (Optional)

Students can think about/research what differences in states there are that might explain the differences in the relationships.

*Note to instructor (use the following information as you see fit, i.e. do share, or do not share, this information with students before announcing this optional homework assignment): *How might variations in state funding (a.k.a. “per-pupil-funding”) affect the relationship between poverty and the student-teacher ratio, if at all?*

Bibliography

Achilles, C. M. (2005). Financing Class Size Reduction. *Regional Educational Laboratory Southeast*.

Biddle, B. J., & Berliner, D. C. (2002). What research says about small classes and their effects.

Biddle, B. J., & Berliner, D. C. (2008). Small class size and its effects. *Schools and society: A sociological approach to education*, 3, 86-95.

Governing, 2018. “Education Spending Per Student By State.” Retrieved May 15, 2019 (<https://www.governing.com/gov-data/education-data/state-education-spending-per-pupil-data.html>).

Huffington Post Life, 2017. “These Are The States With The Most Students For Every Teacher.” Retrieved May 15, 2019 (https://www.huffpost.com/entry/states-student-teacher-ratio_n_6083480).

Public School Review, 2019. “Average Public School Student: Teacher Ratio.” Retrieved May 15, 2019 (<https://www.publicschoolreview.com/average-student-teacher-ratio-stats/national-data/states>).

Whitehurst, G. J., & Chingos, M. M. (2011). Class Size: What Research Says and What It Means for State Policy. *Brookings Institution*.

This OER material was produced as a result of the CS04ALL CUNY OER project.

Creative Commons License



This work is licensed under a [Creative Commons Attribution-Noncommercial-Share Alike 4.0 License](https://creativecommons.org/licenses/by-nc-sa/4.0/).