2019

CS + Sociology: Using Big Data to Identify and Understand Educational Inequality in America (1)

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

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

Title: CS + Sociology: Using Big Data to Identify and Understand Educational Inequality in America (Week 1 of 2)

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

Date: 5/29/2019

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 1)

Description: This is the first part (first 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. 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 1

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 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 to be conducted 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:

"The Relationship Between School Spending and Student Achievement: A Review and Analysis of 35 Years of Production Function Research" (Verstegen and King 1998)

"Socioeconomic Status and Its Relationship to Educational Resources" (Sledge 2016, Thesis)

A clear and reader-friendly, but serious, overview of the relationship between the teacher-pupil relationship and student achievement: http://www.centerforpubliceducation.org/research/class-size-and-student-achievement

Another clear and user-friendly overview of research on the teacher-pupil ratio: https://www.brookings.edu/research/class-size-what-research-says-and-what-it-means-for-state-policy/

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.
2. Independent and dependent variables.
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 important ideas would be that different forms of inequality are often associated. So areas that are low income would be more likely to have lower quality educational resources. 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 may vary from place to place depending on other factors. For example in the lab data from New York is examined but data from Mississippi might be different.

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

1. Students will be able to identify and interpret available resources for gathering socioeconomic data.
2. Students will be able to compare and contrast socioeconomic indicators of public high school success throughout New York City and be able to reach conclusions regarding relevant data.
3. Students will understand how to utilize socioeconomic characteristics in a comparative context and be able to make sound judgments regarding data.
4. Students will compare accessing data using Web APIs using a computer script and via a browser interface and assess the differences.
5. Students will use 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. Students will be able to explain at a beginning level what Web APIs are and they are used.

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).

If your students have more skills than assumed here you can have them create additional graphs for states besides New York and compare them. (This can also be part of a homework assignment.) 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. For this reason the lab has students graph just the data from New York.

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

Class's Suggested Schedule

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 consider relationship between socioeconomic status and the quality of educational experiences
Socioeconomic status
Overarching questions:
What is socioeconomic status?
What is the socioeconomics status of a place? How does that differ from that for an individual?

Instructor may first ask for volunteers to answer this question. Ultimately, instructor will define socioeconomic status (SES) for students using their own definition or may draw from this definition: "Socioeconomic status refers to, among other things, how much money and education an individual and an individual's family possess. Though there are different ways to categorize SES, many sociologists often think in terms of the following: poor, working class, lower middle class, middle class, upper middle class, and upper/elite class. SES and 'social class' are synonymous with each other."

Can we think of a place, such as a neighborhood or school district as having a socioeconomic status? How does that differ from individual socioeconomic status? Briefly discuss. Key points: Yes a place can have a status, but the characteristics of a place do not necessarily apply to every person living in that place. A neighborhood that has an overall low level of college education will still have many individuals that have attended college (or are currently attending college).

The independent 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 might want to explore the possibility that the socioeconomic status of a place is related to other variables. For example, would it be reasonable to predict that school districts with low child poverty rates would have better schools than school districts with high child poverty rates?

Quality of Educational Experience
Why is this a reasonable variable to use if you are interested in schools?
What are some specific (quantitative) measures of the quality of a school?
The Instructor can either lecture about specific measures of school quality that might be used or ask students: "What are some specific measures of the quality of a school (a high school, for example)?" Instructor will elicit student responses. Some specific, quantitative measures of the quality of a high school including but not limited to: graduation rate, attendance rate, percentage of all teachers who possess proper certification, and the percentage of students who were enrolled in a postsecondary program within six months of graduation – among many other measures.

Instructor will then inform students that the dependent variable used in this lesson is the "teacher-pupil ratio". Students and instructor will then discuss what this means. Why might the teacher-pupil ratio matter? Why might it not matter? (See citations above for suggested background readings.)

Either the instructor or the students Can you create a question that asks whether or not a relationship exists between these two variables?
Instructor can elicit student responses. Ultimately instructor will identify the answers:

- Independent variable: Child poverty rate
- Dependent variable: Teacher-pupil ratio

Hypotheses

Instructor will ask students to make a hypothesis ("A hypothesis is a prediction made prior to collection of data") about the possible relationship between child poverty rate and teacher-pupil ratio in a school district. Students may first answer independently in their notebooks or in their own heads, followed by a class discussion.

The higher the child poverty in a school district, the BLANK (larger or smaller) the class size? Ask students to provide guesses to this relationship and have discussion on why they think this to be so.

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":"2056808574","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_analysis.Rmd.R"*):
(https://gist.github.com/elinw/cf310b7b2422d8562e1c24a2b8632044

Click on the "raw" button and copy and paste the full text into your empty file.

Save the file.
From this point on follow the instructions in the file.

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).

Bibliography

"Class Size: What Research Says and What it Means for State Policy." The Brookings Institution. Retrieved April 2019: https://www.brookings.edu/research/class-size-what-research-says-and-what-it-means-for-state-policy/

Sledge, C. (2016). "Socioeconomic Status and Its Relationship to Educational Resources" (Unpublished thesis).

The Center for Public Education, 2019. "Class Size and Student Achievement." Retrieved April 2019 (http://www.centerforpubliceducation.org/research/class-size-and-student-achievement).

Verstegen, D. and King, R. (1998). "The Relationship Between School Spending and Student Achievement: A Review and Analysis of 35 Years of Production Function Research." *Journal of Education Finance*, *24*(2), 243-262.

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