Spring 5-31-2019

CS+Sociology: Global Inequality Lab 1

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

Janet Michello
CUNY La Guardia Community College, jmichell@lagcc.cuny.edu

	Title: Global Inequality Lab 1: CS+Sociology
	Author: Janet Michello, Elin Waring
	Date: 5/31/2019
	Material Type: Lesson and Lab Plan
	Software/Equipment Dependencies: R, RStudio (optonal); internet access required.
	Prior Knowledge Needed (if any): None
Science	**Keywords:** Inequality, Global, R, Web API, CIA Factbook, Sociology, Computer
	Approximate time needed: 1-3 hours depending one instructor choices
	Description: These materials include background for the instructor and a lab that engages student in an analysis of global inequality while learning and using the R language (a programming language for statistics). Students ultimately write a function to access country level data from the CIA World Factbook.

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

CS + Sociology: Global Inequality

Introduction for Instructor

This exercise explores the use of "big data" to study a social issue important in the discipline of sociology and to understand the nature and extent of global inequality. It uses country level data to help students explore variations across the world within the context of the United States.

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 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 or by adding visualization.

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, listviewer, 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. They should help non-social science faculty understand the meaning of some important terms and the reasons why students are investigating specific national characteristics. For example, one of the articles states the variance for level of education when measuring occupational attainment and how education is frequently used as a predictive variable because of its ease of use rather than its explanatory power. Other articles address the changing nature of work and how it contributes to: inequality/social stratification, dispersion of wealth and power, food production, and poverty rates. In addition, other important sociological factors are: the relationship between lifestyle and longevity, the societal impact of mortality and fertility rates, urbanization, global warming, pollution, population changes, and the impact of technology and economic factors.

   Collins, R. (2004). Lenski's Power Theory of Economic Inequality: A Central Neglected Question in Stratification Research, *Sociological Theory*, Vol. 22 (2): 219-228. (See attached)

   Lam, D. (2011). How the World Survived the Population Bomb: Lessons From 50 Years of Extraordinary Demographic History, *Demography*, 48:1231–1262. (See attached)

   Nau, C. & G. Firebaugh. (2012). A New Method for Determining Why Length of Life is More Unequal in Some Populations Than in Others, *Demography,* 49:1207–1230. (See attached)

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 use 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 is from the CIA world factbook. Exploring the CIA World Factbook illustrates some of the important issues and possibilities for this data and instructors should review both the Factbook site and this document. The data access and raw documented data are here https://github.com/iancoleman/cia_world_factbook_api and the json file is here https://raw.githubusercontent.com/iancoleman/cia_world_factbook_api/master/data/factbook.json

Computer Science Content

These materials are primarily focused on the "+x" aspect of this lab. The Computer Science content, assuming that students have had 2-3 weeks of a CS0 style class includes:

7. Introduction of the R language.
8. Introduction to data science as part of the larger set of tasks that computer programming is applied to.
9. Introduction of the idea of an API and in particular a Web API.
10. Introduction of JSON which provides a way to visually represent a data object.
11. Introduction of the idea of a tree.
12. Introduction of the idea of metadata.
13. Introduction of the idea of a function in R (the following functions are used in lab 1: library(), file(), readLines(), fromJSON(), names(), jsonedit())
14. Literate programming

Instructors can choose to focus on one or more of these as desired. None are crucial except the introduction to R.

The most likely unusual problem that students will have is having to run jsonedit() in the console while the remainder of the code is run notebook style in an RMarkdown document

Student Learning Outcomes

1. Students will be able to identify and interpret available resources for gathering socio-demographic data.
2. Students will be able to compare and contrast socio-economic indicators of nations throughout the world and be able to reach conclusions regarding relevant data.
3. Students will understand how to utilize socio-demographic 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.

Instructions for students

Instructor Prep for Lab 1

The lab asks students to find a number of statistics by extracting them from a JSON string using the R $ notation for lists. The answers for these are (prefixed with countries$-----$data$ where ---- is the country name as found early in the lab):

people$population$total
people$population$global_rank
people$age_structure$0_to_14 $percent
people$age_structure$15_to_24$percent
people$age_structure$25_to_54$percent
people$age_structure$55_to_64$percent
people$age_structure$65_and_older$percent
peoplesex_ratiototal_population$value

people$birth_rate$births_per_1000_population
people$infant_mortality_rate$value
people$urbanization$rate_of_urbanization$value
people$school_life_expectancy$male$value
people$school_life_expectancy$female$value
people$literacy$total_population$value
people$literacy$total_population$value
people$life_expectancy_at_birth$total_population$value
people$adult_obesity$percent_of_adults
economy$unemployment_rate$annual_values$value$0
economy$labor_force$by_occupation$agriculture
economy$labor_force$by_occupation$industry
economy$labor_force$by_occupation$manufacturing
economy$labor_force$by_occupation$other_services

The expectation should not be that students will obtain all of these in the first lab. In the second lab students will complete all of them by learning to write a function that selects the desired data for any statistic in the fact book.

Optionally, selection of the countries can be done as a pre-lab assignment or the instructor may assign countries (potentially randomly), or all students may use the same countries. This may depend on the coding experience of the class and the length of time available for the lab.

Optionally as a post lab assignment have the students locate all of the desired statistics. Instructors should decide on the appropriate number of statistics to request (and make specific judgements on the importance of each one). For some students this may take a while, but the goal should be for them to identify the repetitive nature of building up the strings, thus providing motivation for creating a less time intensive method.

Lab 1

Pre-lab/warm up
You will select one developing nation and a developed nation in order to contrast specific characteristics with the United States. Since students may have difficulty deciding which are developed nations and which are developing ones, you can go the website of the Population Reference Bureau which lists the average per capita income (GNI PPP) for all world nations in American dollars. To assist in making your selection, consider countries with a GNI of less than $7,000 as developing and over $25,000 as developed. You can also view the World Bank website which lists the richest and poorest nations in the world. Another way of thinking about this is to consider countries that are "well off" compared to those which are

relatively poor. Consider this as a continuum with very rich such as the United States and Japan on one end and Afghanistan and Botswana on the other, with many, many nations in the middle. You will then make your selection of a developed nation and a developing nation to contrast with the United States.

According to the United Nations, these are the economically developed countries of the world:

> Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden, United Kingdom, Bulgaria, Croatia, Cyprus, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Romania, Slovakia, Slovenia, Iceland, Norway, Switzerland, Australia, Canada, Japan, New Zealand, United States, Canada.

We will use the R programming language to access the data for the U.S. and your two countries.

Examine the data.
Visit this page:
https://raw.githubusercontent.com/iancoleman/cia_world_factbook_api/master/data/factbook.json

This is the data from the CIA world factbook in JSON form. Although it may look complex, if you examine it closely you will see patterns, where {} (curly brackets) are used to group together sections of the data and : is used to separate the label for a piece of data and its value. These are called key-value pairs. An example near the top of the file is

"name": "World"

The key is “name” and the value is “World”.

JSON is one of the most commonly used formats for sending data over the web. The data goes from a server to an application. The page full of JSON may look very hard to read to a human but it is very readable to a computer. This is called an Application Programming Interface or API.

Lab

For this lab we will use RStudio to use the R language and access these data.

Follow the instructions from your instructor for logging into RStudio.

Creating your RMarkdown file

Create a new file by typing at the command line:

file.create(“world_factbook.rmd”)

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

Click on the file to open it.

Go to this link

https://gist.github.com/elinw/f63d8b2ea2406fd6807185638ebc4e77

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

Save the file as “world_fact_book.Rmd”.

After saving you should see an icon on the top of the file that says “Knit” and has a picture of a ball of yarn. If the icon doesn’t appear save again and double check that you have used .Rmd at the end of the name.

Follow the instructions in the file, including answering all the questions directly in the file.

When you have completed this lab, click on the knit button and knit to Word.

Save your results.

Post Lab

What patterns did you notice in how each of the individual data items were requested?

How did each of the individual requests differ from the other requests?

Hand in your results following the instructions given to you.