Assignment 1: Working with Data

PSCI 8357, Spring 2016
January 14, 2016

This assignment must be turned in by the start of class on **Thursday, January 21**. You must follow the instructions for submitting an assignment.

Background

You will be working with the dataset `capability-data.csv`, a modified version of the National Material Capabilities dataset from the Correlates of War project.

```
capability_data <- read.csv("capability-data.csv")
dim(capability_data)
```

```
## [1] 198 109
```

```
head(names(capability_data), 10)
```

```
##  [1] "ccode"     "irst.1990" "irst.1991" "irst.1992" "irst.1993"
##  [6] "irst.1994" "irst.1995" "irst.1996" "irst.1997" "irst.1998"
```

```
tail(names(capability_data), 10)
```

```
##  [1] "upop.1998" "upop.1999" "upop.2000" "upop.2001" "upop.2002"
##  [6] "upop.2003" "upop.2004" "upop.2005" "upop.2006" "upop.2007"
```

Each row of the dataset is a country. The first column, `ccode`, gives its Correlates of War country code. You can look up which countries have which codes in the COW State System Membership dataset. Each subsequent column gives how much of a particular military capability component a state possessed in a particular year. There are six of these components:

- `irst`: Iron and steel production, in thousands of tons
- `milex`: Military expenditures, in thousands USD
- `milper`: Military personnel, in thousands
- `pec`: Primary energy consumption, in thousands of coal-ton equivalents

- `tpop`: Total population, in thousands
- `upop`: Urban population, in thousands

Missing data are coded with the value `-9`.

Main Task

You will tidy this ugly dataset, and along the way you will calculate some commonly used transformations of the data. Your goal is to transform `capability-data.csv` into a dataset where each row is a country-year pair. The dataset should contain nine columns:

- `ccode`: The COW country code of the observation
- `year`: The year of the observation
- `irst`, `milex`, `milper`, `pec`, `tpop`, and `upop`: The country's holding of each military component in the given year, expressed as the share of the world total of that component in that year. If we let $x_{i,t}$ denote country i's raw total of the component at time t (i.e., the value in `capability-data.csv`), and denote the set of all countries by J, then what you want to calculate is
$$s_{i,t} = \frac{x_{i,t}}{\sum_{j \in J} x_{j,t}}.$$
If you run into missing values when calculating the annual sum, just drop them. If $x_{i,t}$ itself is missing, then treat the share as missing, recording it as `NA`.
- `cinc`: The country's Composite Index of National Capabilities in the given year. The CINC score is defined as the average of the country's shares of the six individual components:
$$\text{CINC}_{i,t} = \frac{1}{6} \sum_{k=1}^{6} s_{i,t}^{(k)}.$$
If some but not all of the individual components are missing, just take the average of those that are observed. If all of them are missing, then treat the CINC score as missing, recording it as `NA`.

I strongly recommend using the **dplyr** and **tidyr** packages to perform the cleaning and transformation.

The first few rows of your final output should look something like the following. The values here are made up, so don't expect yours to match exactly.

```
##   ccode year irst milex milper  pec tpop upop     cinc
## 1     2 1990 0.25  0.10   0.05 0.15 0.00 0.00 0.091667
## 2     2 1991 0.20  0.05   0.20 0.00 0.25 0.10 0.133333
## 3     2 1992 0.10  0.20   0.00 0.10 0.15 0.25 0.133333
```

Your R script should save your final output in a CSV file called `final-data.csv`. You don't have to commit this file to your Git repository (though you may if you choose).

Your grade will reflect not only whether you get the right answer, but also whether your code is readable and follows the best practices laid out in "Best Practices for Scientific Computing" and the other readings from this week. The same will hold true for the rest of the assignments this semester.

Weekly Visualization Challenge

How best to communicate empirical findings visually could be—should be—an entire course of its own. So, unfortunately, we don't have the time to cover data visualization in a systematic way. As a cheap ad hoc substitute, each weekly assignment will end with a "visualization challenge" asking you to make a graph of some slice of the assignment results.

You will receive credit for the visualization portion of the assignment as long as you make a good faith effort. But whoever turns in the *best* visualization each week, as chosen by me, will receive a $5 Starbucks gift card. We will discuss each week's winning entry and what makes it good, in hopes that we will all absorb some good taste in visualization by osmosis, despite the scant coverage elsewhere in the curriculum. I recommend, but do not require, using the **ggplot2** package to make graphics.

For this week's challenge, we'll start simple: Track the changes over time in the CINC scores of the five countries with a U.N. Security Council veto.