SOC 4650/5650: PS-01

Christopher Prener, Ph.D.

January 31st, 2017

Directions

Please complete all steps below. Your final do-file, log-file, and markdown file with answers should be uploaded to your GitHub assignment repository by 4:20pm on Tuesday, February 14th, 2017. This lab uses the file `MO_HYDRO_Dams.csv`, which lists all Missouri dams in the U.S. Army Corps of Engineers' National Inventory of Dams. You will need to clean several aspects of these data to produce a tidier dataset of large dams in the state.

Cleaning the Dam Data

1. Using Atom, construct a well-formatted do-file using the `headFull` snippet. Be sure to edit the appropriate lines in the template that detail the name and purpose of the file.[1] You will want to name your project `largeDams`, and save your file as `largeDams.do`.
2. Your do-file should *successfully* accomplish the following tasks. It should include narrative text that explain what the command accomplished and, if applicable, provides an answer to the question prompt.
   (a) Import the raw data into Stata.[2]
   (b) Create a table that lists all of the variables in the dataset. How many variables are there?
   (c) Remove the following variables from the dataset: `maxstor`, `resarea`, and `wtrshed`.
   (d) List the first ten observations for the variables `objectid` (object ID number), `nid_id` (National Inventory of Dams ID number), and `offname` (Dam facility name). Which of these variables might uniquely identify observations?
   (e) Test each of these three variables to see if they uniquely identify observations, and include *only* the test that confirms that works without generating an error (because the variable does not uniquely identify observations).
   (f) Check and see if there are any duplicate observations. If there are duplicate observations, drop them. If there are not duplicate

[1] If you lose the `Tab` functionality, you need to edit lines 3, 9, 44, 46, 47, 50, and 70. Note that these line numbers may change if you enter multiple lines of text in, for example, the description area. Look for `/*prompt*/` and replace all of that with your own text. Do not leave the `/*` and `*/` behind!

[2] See the Week-03 jotter for details on this process.

observations, why do you think that some of the three variables from 2d do not uniquely identify observations?

(g) Rename the variable `offname` (Dam facility name). In your narrative, include a justification for the new name you have selected.

(h) Create a frequency table that summarizes the values for the variable `owntype` (Dam owner). Make sure you use the option that will include missing data in the table. Which category has the *most* dams?

(i) Since these values are not specific, use the `replace` command to add more descriptive values.[3] The following definitions apply:

```
F = Federal
L = Local Government
P = Private
S = State
U = Public Utility
```

[3] *Hint:* Make sure you wrap your values in *double quotes*.

(j) Create a frequency table that summarizes the values for the variable `yrcmplt` (year completed). Make sure you use the option that will include missing data in the table. What do you think the value `"0"` means?

(k) Calculate descriptive statistics for the variable `damht` (height of dam in feet). What is the average height?

(l) Create a subset of the data that only includes dams that are *higher* than average by dropping those dams that are *shorter* than average. How many dams are *higher* than average?

(m) Create a frequency table that summarizes the values for the variable `damtype` (type of construction). Make sure you use the option that will include missing data in the table. What is the most common type of construction material? To address this question, use the following information for the National Inventory of Dams data dictionary:

```
Code indicating the type of dam. Codes used are as follows:
RE = Earth; ER = Rockfill; PG = Gravity; CB = Buttress;
VA = Arch; MV = Multi-Arch; CN = Concrete; MS= Masonry;
ST = Stone; TC = Timber Crib; OT = Other.

Codes are concatenated if the dam is a combination of
several types.
```

3. At the end of your do-file, also answer these questions:

   (a) Use the Data Editor (Browse) mode to scan your dataset. Are variables stored in unique columns that measure only a single concept?

   (b) Does each observation form a row with no duplicates? How might we account for the fact that some of the identification variables tested above *do not* uniquely identify observations?

   (c) How many entries exist in total for this dataset after you have finished cleaning it?[4]

   (d) What is the observational unit in this dataset?

   (e) Overall, can we consider this dataset to be a tidy dataset?

   (f) What values could be more clearly defined?

4. Save your do-file in Atom, and close the application. Open the do-file in Stata and execute it to create your final directories that contain a copy of your code as well as your log file, raw data, imported data, and your Markdown output. Your log file and Markdown output should be submitted into the `PS-01` directory in your assignments repository.

[4] *Hint:* Use the table created in 2m to answer this question.