SOC 4650/5650: Lab-03

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 7th, 2017. This lab uses the file `MO_HYRDO_ImpariedRiversStreams.csv`,[1] which lists all streams and rivers in Missouri listed under the Clean Water Act. You will need to clear several aspects of these data to produce a tidier dataset of bodies of water within St. Louis County that are listed under the Clean Water Act.

[1] There is a typo on this file and a companion file that lists Impaired Lakes. In both files, the word `HYDRO` is misspelled. Please fix this misspelling on both `.csv` files before proceeding.

Getting Started

1. Using Atom, create a new file, and change its language to "Stata".
2. Expand the snippet `headFull`.
3. *Without clicking anywhere*, begin using the `Tab` key to move through the fields that you are prompted to complete.[2] Fill them out as best you can. Your project name should be `listedStreams`. Review your answers with your partner, and work together to improve your responses to these fields.
4. Save your do-file as `listedStreams.do`. Your do files should always be identical to your project name.

[2] 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!

Working with Stata and Atom

5. In Stata, set the working directory to the `Lab-03` directory in your folder hierarchy.
6. Create a plan with your partner for the commands you will need to answer each of the sub-questions under question 7.
7. Test the commands for the following tasks in Stata, and then copy the *working* commands into your Atom do-file beginning under the

"Import/Open Data" header. When you copy the command in, include some narrative text below it that begins with the appropriate question number like so:

```
/***
**7a.** This is the answer to question 7a.
***/
```

Your narrative text should explain what the command accomplished and also provide an answer to the question prompt.

(a) Import the raw data using the following command:

```
import delimited fileName, varnames(1)
```

When you move the command into Atom, the import command should look like this:[3]

```
import delimited 'rawData', varnames(1)
```

(b) Drop the following variables: `businessid`, `mdnr_impsz`, `size_`, `epa_apprsz`, `unit`, `wb_epa`, `comment_`, `eventdat`, `reachcode`, `rchsmdate`, `rch_res`, `src_desc`, `feat_url`, `fmeasure`, `tmeasure`, `shape_leng`, and `shape_le_1`.

(c) Drop observations where `county_u_d` does not equal `"St. Louis"`. This will create a *subset* of your data that focuses solely on bodies of water within St. Louis County that are listed under the Clean Water Act.

(d) Test to see if the variable `wbid` uniquely identifies observations.[4]

(e) Test to see if the variable `perm_id` uniquely identifies observations. [5]

(f) Identify how many observations have missing data in the variable `perm_id`.

(g) Since the variable `perm_id` is incomplete, drop it from your dataset.

(h) Rename the variable `source_` to `source`.

(i) Rename the variable `county_u_d` to `county`.

(j) Run a duplicates report on your dataset. How many unique observations are there?

(k) If there are duplicates, remove them from the dataset. How many observations (if any) were deleted?

(l) Tabulate the `water_body` variable and note how Fee Fee Creek is labeled differently than the other observations. Fix this issue.[6]

[3] This command works because, when you filled out the fields in question 7, you entered the filename of the listed lakes data and saved in the object (or, in Stata-speak, the local macro) named `rawData`. This improves the reproducibility of your work because it limits the number of times you have to enter the same filename in. If there is a change in the filename, you can make that update once and it will be applied to any area of your code where the object `rawData` is listed.

[4] Remember not to include this command in your do-file. Only include your written answer to whether or not the variable `wbid` uniquely identifies observations.

[5] Remember not to include this command in your do-file. Only include your written answer to whether or not the variable `perm_id` uniquely identifies observations.

[6] *Hint:* Use the `rename` command.

8. 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?

   (c) How many entries exist in total for this dataset after you have finished cleaning it? Of those entries, how many unique bodies of water are included?[7]

   (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?

9. 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 `Lab-02` directory in your assignments repository.

[7] `Hint:` Use the output for question 7l to answer this question.