SOC 4650/5650: Lab 2-1 - Clean Water Act Data Cleaning

Christopher Prener, Ph.D.

Spring 2022

Directions

Using data from the `data/lab-2-1/` folder available in the `module-2-data-cleaning` repository, clean up the data on Clean Water Act issues with streams and rivers in Missouri. Your entire project folder system should be uploaded to GitHub by Monday, February 21st at 4:15pm.

Analysis Development

The goal of this section is to create a self contained project directory with all of the data, code, map documents, results, and documentation a project needs.

a. **Clone** the `module-2-data-cleaning` repository if you have not already done so.

b. Rename the folder in your assignments repository's `Labs` directory from `Lab-04` to `Lab-2-1`.

c. Create a project folder system with all of the necessary components, and drag the lab data from `module-2-data-cleaning/data/lab-2-1/` into your RStudio Project's `data/` subdirectory.

d. Create a `README.md` text file (File ▷ New File ▷ Text File). **In addition,** add a quick description of your project and outline the key directories and files that are included.

e. Create a well-formatted RMarkdown document for your data cleaning efforts.

f. Load the `.csv` file containing the lab data.

Part 1: Data Wrangling

1. Begin by creating a pipeline that:

   (a) Renames variables to `snake_case` en masse using the `clean_names()` function,

   (b) renames the variable `eventdat` to `date`,
   (c) and rename the variable `county_u_d` to `county`.
2. Next create a missing variable summary using `miss_var_summary()`.
3. Create a duplicate observation report. How many duplicates are there in total? How many actual unique observations are there (i.e. if you removed all of the duplicates but kept a single observation for each unique case)?[1]
4. Check to see if there are duplicates in the `perm_id` variable, which appears like it may uniquely identify observations. Is this the case? If it is not, how many duplicate instances are there? If there are more than twenty, remove this code chunk and its output from your notebook to keep its length short and document in your narrative what your findings were.
5. In a pipeline, make the following two changes:
   (a) Create a subset of observations where `county` is equal to `St. Louis`.
   (b) Then keep only the following variables: `yr`, `wbid`, `water_body`, and `pollutant`, and `source`.
   (c) Assign these changes to a new tibble.
6. In a pipeline, edit the following variables in your St. Louis subset to create a new measure and edit an existing one:
   (a) Edit the `water_body` variable for observations that have the value `Gravois Creek tributary`. Change these values to `Gravois Cr. tributary` so that they match how the word "Creek" is abbreviated in the other observations.
   (b) Then make the a similar change for values `Twomile Creek`.
   (c) Then make the a similar change for values `Watkins Creek tributary`
   (d) Then create a new variable named `ecoli` that is `TRUE` if the `pollutant` is `Escherichia coli (W)` and `FALSE` otherwise.
   (e) Assign these changes back into the existing tibble containing the St. Louis subset.

[1] Look at the `dupe_count` variable that is created in your output. Remember that if you duplicate report is long, it should not be included in your notebook! Just document the results.

Analysis Development Follow-up

Don't forget to knit your document when you are done! Also be sure to go back and update your `README.md` file with any changes to your project's organization or contents.