SOC 4650/5650: PS-01 - Missouri Dam Data

Christopher Prener, Ph.D.

February 5th, 2018

Directions

Using data from the `MO_HYDRO_Dams.csv` data, create a well-formatted notebook that cleans a data set of all Missouri dams in the U.S. Army Corps of Engineers' National Inventory of Dams. Your entire project folder system, including notebook output and results, should be uploaded to GitHub by Monday, February 12th at 4:15pm.

Analysis Development (Review from Lectures 01 and 02)

1. In your course folder system, find the `ProblemSets/PS-01` subdirectory.
2. Add folders within it called `docs`, `data`, and `results`.
3. Copy and paste the data file `MO_HYDRO_Dams.csv` from the course `Data` folder into `ProblemSets/PS-01/data`.
4. Using RStudio, add an R Project to the `ProblemSets/PS-01` folder.
5. In RStudio, create a new R Notebook and save it to the `ProblemSets/PS-01/docs` subdirectory you created above. Make sure it is fully set-up following the workflow we have been using.

Part 1: Cleaning the Dam Data

6. In the data load code chunk of your notebook, use the `read_csv()` function to load the data stored in `MO_HYDRO_Dams.csv`. Make sure it is saved as a tibble.
7. Begin by creating a pipeline that:
   (a) Renames variables to `camelCase` en masse,
   (b) then renames the variable `objectid` to `id`,
   (c) then renames the variable `offname` to `name`,
   (d) then renames the variable `owntype` to `ownType`,
   (e) then renames the variable `damtype` to `damType`,

(f) then renames the variable `yrcmplt` to `yearComplete`,

(g) then renames the variable `damht` to `height`,

(h) and assigns these changes back into the existing tibble.

8. Create a report of missing data across variables - do any variables have missing data?

9. Create a report of missing data across observations - do any variables have missing data?

10. Create and evaluate a duplicate observation report for the entire data frame.

11. Check to see if there are duplicates in the `id` variable, which appears like it may uniquely identify observations. Is this the case?

12. Check to see if there are duplicates in the `nidId` variable, which appears like it may uniquely identify observations. This variable is the National Inventory of Dams Identification Number. Is this the case?

13. Based on your answer to the last two questions, which variable is best to use if we want to uniquely identify observations?

14. In a pipeline, make the following two changes:

    (a) Create a subset of observations where dam height is 30 or more feet,

    (b) then remove the following variables: `maxstor`, `resarea`, and `wtrshed`,

    (c) and assign these changes to a new tibble.

15. In a pipeline, edit the following variables in your high dams subset to create a new measure and edit an existing one:

    (a) Edit the `ownType` variable's values so that they are more descriptive.[1] The following definitions apply:

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

    (b) Then edit the `yearComplete` variable, which has values of `0` included for damns whose completion years are unknown. Replace these values with `NA` (missing data) values.[2]

    (c) Then make a variable that is `TRUE` if the dam type is `RE` (an earthen dam), and `FALSE` otherwise.[3]

[1] *Hint:* Make sure you wrap your values in *double quotes*. You will need either five instances of the `mutate()` function combined with `ifelse()` **or** one instance of the `mutate()` function combined with `case_when()`.

[2] *Hint:* Make sure you **do not** wrap your reference to `NA` values in *double quotes*.

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

(d) Assign these changes back into the existing tibble containing the high dams subset.

16. Experiment with a new function - use the `janitor` package's GitHub `README` file to learn about the tabulating tools included in the package. Make a *simple* frequency table of the variable `ownType` from the high dams subset. Who owns the most high dams in Missouri?

Part 2: Plotting the Dam Data

17. Create a histogram of the variable `height` from the original tibble that contains all observations, and save this plot to your `results` subdirectory.

18. Create a bar plot of the variable `ownType` from the high dams subset:[4]

    (a) Use the same code layout as you did for the previous question, but change the geom to `geom_bar`

    (b) Change the data reference to the appropriate subset

    (c) Change the variable included in the aesthetic mapping to `ownType`

[4] If you have questions, you should check the `ggplot2` website for details on the bar geom.

Part 3: Reproducible Example

19. Note that the `statbin` warning message appears again in question 17. Using the `starwars` data in the `dplyr` package, create a reproducible example the shows this warning. You can use either the `height`, `mass`, or `birth_year` variables for a selection of Star Wars characters to create this example. Don't forget to include your `library()` functions!

    Send it to Brandon and Chris in a direct message on Slack along with a well thought out question about either the plot, the code, or the wider process use to create the plot. The `R` file you create does not have to be submitted nor do you have to include the `reprex()` function in your notebook.