Intro to Reproducibile Data Cleanup in R

1. Set up an RStudio Project and download the relevant packages

1. File -> New Project
2. Name the Project as follows: `las_demo`
3. Create the following three folders:

   ```
   data_raw
   data_clean
   code
   ```

   You can create these in the folder using your operating systems “create folder” option ***or*** you can create within R studio using the `files` tab

4. Install libraries, load libraries, and run commands

```
# Use the tab to install package `dadjokeapi`.
# You could also do the following:
install.packages("dadjokesapi")
install.packages("tidyverse")
# load the library
library(dadjokeapi)
# run a command
groan()
```

2. Download the data we will be using in class

1. Open the messy data file `demo_data.csv` by following this link
2. the data will open as a tab in your web browser in `.csv` format; save them to the `data_raw` folder by going to ‘File’ on the menu bar of your web browser and selecting ‘Save page as’ from the drop-down menu.
3. save the file to the `data_raw` folder.

3. Data Cleaning: Overview

1. *Characteristics of clean data set include:*
   - Free of duplicate rows/values
   - Error-free (correct misspellings, eliminate special characters)
   - correct data type for analysis
   - outliers identified and dealt in the correct way
   - “tidy” data structure
2. *Take your data from `messy` to `clean` in 5 steps*
   - Familiarize yourself with the data set

- Check for structural errors
- Check for data irregularities
- Decide how to deal with missing values

4. Data Cleaning: Practice

1. Review the Data Set

1. Review the `.csv` file
2. What things do you see that need to be corrected?
3. Make a list of the what you think needs to be corrected and the steps necessary to identify and implement each correction. Some of the things to look out for include:
    - Numeric values stored as character data types
    - Factors stred as characters
    - Duplicate rows
    - Spelling mistakes
    - inconsistent formatting (eg., codes, capitalizations)
    - White spaces
    - Missing data
    - Zeros instead of null values
    - Special characters (e.g. commas in numeric values instead of decimals)
    - column headings with spaces between words or that start with numerals
4. It is often useful to make an outline of the different steps. Note that there might be differetn ways to do the same thing, so an outline will help figure out which is best. For instance:

    Option 1
    1. Import table 1
    2. Correct column headings in Table 1
    3. Import table 2
    4. Correct column headings in Table 2
    5. Bind Table 1 and Table 2 Together

    is less efficient than

    Option 2
    1. Import table 1
    2. Import table 2
    3. Bind Table 1 and Table 2 Together
    4. Correct the column headings in the Table

5. KEY MESSAGES

1. Keep raw data raw. Always.

2. Remember: the # symbol (hash tag) allows you to make comments in the script. Be sure to Annotate. Lots.

Tools & Resources

1. These introductions to R and R Studio were made by Professor Ethan White (UF-WEC). They are a good overview of some R basics.
    - Intro to R and RStudio.
    - Navigate R and RStudio web page
    - Intro to R Packages
    - Expressions and Variables in R
2. The Carpentries' R workshops (self-paced or taught in-person) are excellent, I use many of their materials in class:
    - R for Social Scientists
    - Data Analysis and Visualization in R for Ecologists
3. Software Carpentry lesson on Project Management with R Studio
4. Hadley Wickham wrote a book on using the tidyverse and the online version is FREE. This is a phenomenal resource on using R to import, tidy, and visualize data.
5. RStudio Cheat Sheets: help with commands for using the different `tidyverse` packages, RStudio shortcuts and tricks, help with R commands, and more. You definitely want the ones for Data Import, Work with Strings, Factors, Data Transformation, and Base R.
6. Where and How to ask for help
    - Hadley Wickham's advice on how to write a good reproducible example for getting help with R
    - how to post good questions on StackOverflow
    - The UF R-users listserv is *very* user friendly and a great place to post requests for help.
7. Ten simple rules for biologists learning to program
8. Lot's more on the course's 'Resources' page

Additional (interesting) Reading

1. Lewis, Keith P., Eric Vander Wal, and David A. Fifield. 2018. Wildlife biology, big data, and reproducible research. Wildlife Society Bulletin 42(1): 172-179.
2. White EP, Baldridge E, Brym ZT, Locey KJ, McGlinn DJ, Supp SR. 2013. Nine simple ways to make it easier to (re)use your data. Ideas in Ecology and Evolution. 6(2):1-10.
3. The humanities have a 'reproducibility' problem

4. The humanities do not need a replication drive
5. Reproducible Research: A primer for the social sciences
6. Replicability and replication in the humanities
7. Towards reproducible science in the digital humanities
8. The possibility and desirability of replication in the humanities
9. Reproducible Research: A Retrospective

For when you feel more comfortable with R and programming

1. Bryan, J. (2018). Excuse me, do you have a moment to talk about version control? The American Statistician, 72(1), 20-27.
2. Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. Good enough practices in scientific computing. PLoS Computational Biology 13, no. 6 (2017): e1005510.