Individual Project: Reproducible Data Organization

Assignment overview

This project is an opportunity to put some of the lessons we've learned into practice with a data set of your own. Your assignment is to **(1)** clean and organize a 'messy' data set and prepare metadata describing the resulting 'clean' data. The complete project requires the submission of these three items via the course Canvas website:

(1) R code that imports, cleans, and organizes, and saves 'messy' data (please make sure I have access to the original 'messy' data by either providing a link or the data themselves. If these data cannot be shared because of privacy / confidentiality concerns please let me know in advance so we can arrange an alternative means for me to review them);
(2) The resulting 'clean' (i.e., corrected and organized) data in an appropriate format (e.g., .csv, .txt)
(3) Metadata describing the corrected data set (This can either be in the form of the link to a github repository or the submissio of a file in .txt or .pdf format)

Due Date & Point Value: The assignment is due **1 May 2023** by 5 pm and is worth **500 points.**

Evaluation & Grading Rubric

Each portion of your submission will be evaluated using the point values, minimum standards, and rubric below. The individual components will be evaluated as "Meets required standards", "Moderate revisions required to meet the minimum standards", "Major revisions required to meet the minimum standards", or as "Incomplete / Unacceptable". In addition, I have noted a series of additional (optional) steps that can be taken to earn a designation of "Exceptional Work" on each section that can result in a bonus of up to 10%.

Section	Value	Meets standards	Moderate revisions	Major revisions	Incomplete / Unacceptable	Potential Bonus
Code	150	150-135	134-120	119-105	104-75	15
Cleaned Data	100	100-90	89-80	79-70	69-50	10
Metadata	250	250-225	224-200	199-175	174-125	25
Total	500	500-450	447-400	397-350	347-250	50

(1) Code (150 points): To evaluate this portion I will use your code to process the 'messy' data files and then review the resulting 'clean' ones. Remember - this not a programming class, and I am aware some of you may be programming for the first time. This is reflected in the relative weight given to the code vs. the resulting clean data set. Your code doesn't have to be elegant or sophisticated for you to get full credit. My primary concern is the outcome - does it work? It does, however, need to meet some minimum standards to ensure you and others can interpret it in the future.

The following items are the Minimum Standards required for your code:

- A header that explains what the code is for, what packages were used, and other relevant information.
- Commenting that allows a new user to understand the steps being taken
- Modularity: complex problems are broken down into smaller, logically discrete steps.
- Use of functions are used instead of repeated code chunks.
- Data are imported, corrected, reorganized, and exported without on/off commenting of code
- data are saved in a proprietary format

An evaluation of "Exceptional" requires the following:

- Meets the required minimum standards
- Adherence to an R style guide (e.g., http://adv-r.had.co.nz/Style.html) and
- Code archived and assigned a DOI
- Data, Code, and Output are organized in an Rstudio project

(2) Clean Data (100 points): Once I have processed your original 'messy' data I will review the results to see the extent to which they meet the standards we discussed in class. The Minimum Standards depend in part on the kind of data with which you are working. That said…

The Minimum Standards for most data sets are as follows:

- The data are in 'tidy' form
- Subjects have unique identifiers
- Column names are consistent, efficient, and properly formatted
- Dates adhere to a standard format
- Columns contain only a single type of data
- Missing values are identified with a consistent fixed code
- Codes are used when possible to reduce errors
- No data have leading or trailing white spaces
- The file names are informative and properly formatted

An evaluation of "Exceptional" requires the following:

- Meets the required minimum standards
- All columns are set to the appropriate data type
- Factors are ordered when appropriate
- Corrections or changes are recorded in a separate log file
- Data integrity verified with checksums or other QA/QC measures

(3) Metadata (250 pts): : A data set is only as useful as the metadata that accompanies it. This portion of the assignment is the opportunity to prepare the metadata that will accompany your clean data and ensure it is (re)usable in the future by you and others. The metadata that need to be included depend on the project and data set (e.g., if you are interviewing human subjects you obviously don't have to include taxonomic data on the focal species). Though Michener *et al.* 1997 was written with geospatial environmental data in mind, it is actually a useful checklist for other disciplines as well. Please organize your Metadata File(s) using the five classes of Data Descriptors in Michener *et al.'s* Table 1. Include the most relevant Subheadings from each of these Classes, as well as any not listed relevant to your discipline or data. I have posted a text

version of the Classes & Subheadings in Table 1 on the course website so that you don't have to enter them manually; simply delete any that aren't relevant.

The items included in Metadata vary with data set and discipline. However, here are the items that are required for Metadata files to meet minimum standards:

Class 1: Data Set Descriptors

- data set identity and identification codes
- Names and contact information of the Investigators associated with the data set, including the one to be contacted with questions
- Information on any funders of the data collection
- Brief description of the research objectives and data contents
- Keywords

Class 2: Research Descriptors

- Time Frame of Data Collection
- Ecological, socioeconomic, or historical description of the site of data collection (as appropriate)
- Study or Sampling Design:
  - Design overview
  - Temporal aspects of data collection (e.g., data collected hourly, daily, weekly)
  - Spatial aspects of data collection (e.g., specific locations of data collection; spatial structure of sampling within locations)
- Research Methods
  - Instruments used to collect data
  - References, archives, or collections used to identify samples
  - Personnel involved in Data Collection
  - Information on the precision of the sampling instruments and recorded data, if appropriate
  - Description of the focal units on which data were collected (e.g., individuals, species, populations, samples, artefacts, etc.)
  - Names of individuals that assisted with data collection, data entry, and QA/QC.
  - References to pertinent scientific and collecting permits, relevant laws, or institutional policies (e.g., IRB, IACUC)

Class 3: Information on data set status an accessibility

- Status: Dates of verification, archiving, updating, etc.
- Accessibility: storage location and medium, security, proprietary restriction, etc.
- Contact information for access or questions

Class 4: Information on data set structure, organization, and how values are to be interpreted

- File descriptors: name, size, storage mode and format,first and last columns, etc.
- Variable identity: well-defined variables with properly formatted names
- Comprehensive description of each data column, including attributes of the values (units of measurement, range, precision)

- Variable codes are listed and defined.

Class 5: Supplemental Descriptors

- Quality assurance/quality control procedures
- Description of data aquiition materials (forms, loggers)
- Information on the locations and archiving procedures of original data forms, relevant maps, photographs, videos, GIS data layers, physical specimens, field notebooks, comments, etc.
- Description of how data are archived for long-term storage and access
- Information on data set usage and attribution
- History of data set usage, including list of publications or other materials

An evaluation of "Exceptional" requires the following:

- Meets the required minimum standards for Metadata
- Metadata archived at a permanent, public repository (can be embargoed)
- Metadata file generated with Rmarkdown; file saved to Github to allow for version control

Potential Datasets

The following sites have lots of data available for download that would be suitable for this project. If you find something you are interested in using, please be sure to run it by me first.

1. UNDP: http://hdr.undp.org/en/content/download-data
2. UNICEF: https://data.unicef.org/resources/resource-type/datasets/
3. UNEP: https://www.unep.org/data-resources
4. US Census Bureau American Community Survey: https://www.census.gov/programs-surveys/acs/data.html
5. Global Forest Watch: https://www.globalforestwatch.org/