DATA MANAGEMENT PROJECT

LAS 6292 (2022)

Assignment overview

Your project this semester is to:

Clean and organize a 'messy' data set, prepare metadata describing the resulting 'clean' data, and prepare a data management plant for your thesis research.

The assignment is **due is due 30 April 2021** by 5 pm and has a **total value of 800 points.** The complete project requires the submission of these items via the course Canvas website:

Part 1: R code that imports, cleans, and organizes the 'messy' data and then saves the resulting 'clean' data. Please make sure I have access to the original ('messy') data, either by 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.

Part 2: The resulting 'clean' data in txt, csv, or some other non-proprietary format

Part 3: A link granting access to a .txt file of Metadata describing these data.

Part 4: A link granting access to a Data Management Plan for your thesis research (acceptable formats for the file include .pdf, .txt, and .docx)

Required Content & 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	100	100-90	89-80	79-70	69-50	10
Cleaned Data	275	275-248	247-220	219-192	191-138	27
Metadata	275	275-248	247-220	219-192	191-138	27
DMP	150	150-135	134-120	119-105	104-75	15
Total	800					79

Rubric for Part 1 (Code, 100 points)

Part 1: Code. 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

Rubric for Part 2: Clean Data

Once I have processed your '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

Rubric for Part 3: Metadata (275 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 minimum standards for most Metadata files will include:

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 institu-

tional 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 acquisition 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

Rubric for Part 4: DMP (150 pts)

The Data Management Plan (DMP) is a critical document describing the data to be collected for a research project, how it will be stored and managed, and the investigator with primary responsibility for its management. Many funding agencies, including NSF and NIH, now require a DMP with all grant applications. I ***strongly*** recommend you prepare your DMP using the template from the online DMP Tool best suited to your research. Not all DMPs include the same information, and you should refer to the course materials to decide what to include and for examples from different disciplines.

That said, the Minimum Standard for DMPs - regardless of discipline - include the following:

Administrative Information

- Project title
- Researcher name and contact information
- Details of any relevant institutional policies (e.g., IRB, IP, IAUCUC)
- Names of funders that supported the data collection

- Who is using the data
- Who is responsible for managing the data?
- Who will ensure that the data management plan is carried out?

Information on Data Collection

- The purpose of research for which the data are being collected.
- The kind of samples and data that were collected
- How collected and how often
- Format of raw data (paper, digital, image, audio)
- How much data: number of samples, number and size of files, total size of digital archive.
- Reproducibility of collection or analysis and if collection used standard methods
- Metadata files, code books, or other documentation needed by other researchers to use and interpret data, including how archived

Information on data formats and standards, storage, and backup

- Data formats and if (a) standard for the field and (b) open or proprietary.
- Repository in which data will be archived
- Short- and long-term data storage and preservation (physical, digital) procedures
- Plans for regular data backup
- Plans to ensure security of private/restricted data
- Dinancial costs related to data archiving or storage (if appropriate)
- Plans to ensure long-term data use (i.e., storage media, file formats, etc.)
- Any tools or software are required to read or view the data

Information on data sharing and access policies

- How personal or sensitive information have been removed to ensure privacy protection.
- Who holds intellectual property rights for the data and other information created by the project and if there any patent- or technology-licensing-related restrictions on data sharing.
- Whether re-use, redistribution, or the creation of new tools, services, data sets, or products will be permitted and if commercial use is allowed.
- Any embargoes on the data
- The attribution of credit to individuals and institutions, including funders.
- The length of time the data will be retained (if not permanently archived)

An evaluation of "Exceptional" requires the following:

- Meets the required DMP minimum standards
- DMP file generated with Rmarkdown; file saved to Github to allow for version control