Data Organization, Tidy Data, and Spreadsheet Best Practices

Week 2
Data Science Workshop for
NGA LTER REU Students

Review of Last Week

- Data can be fudged
  - Intentionally
  - Unintentionally (motivated reasoning)
- Data can be lost
- Data will sometimes need updating
- Data will be shared
  - At very least: with advisors and collaborators
  - Required: with scientific community
  - Maximally: with Everybody in the Whole World !!

Goals for Today

- Cover guidelines for organizing data and and your work
- Discuss Tidy Data
  - How do computers see data vs how do people see it?
- Interactive practice cleaning a messy dataset

Project Organization (slide 1)

- Encapsulate whole project in one directory.
- Separate raw data from derived data and other data summaries.
- Separate the data from the code.

- Write ReadMe files to document processing steps.

Initial Steps Toward Reproducible Research

Project Organization (slide 2)

- Choose file names carefully.
  - Replace spaces with underscores
  - Be as clear and explicit as possible

- Avoid using “final” in a file name. Nothing is ever final
  - Use “_v1”, “_v2”, etc.
  - Document what the versions are in the README

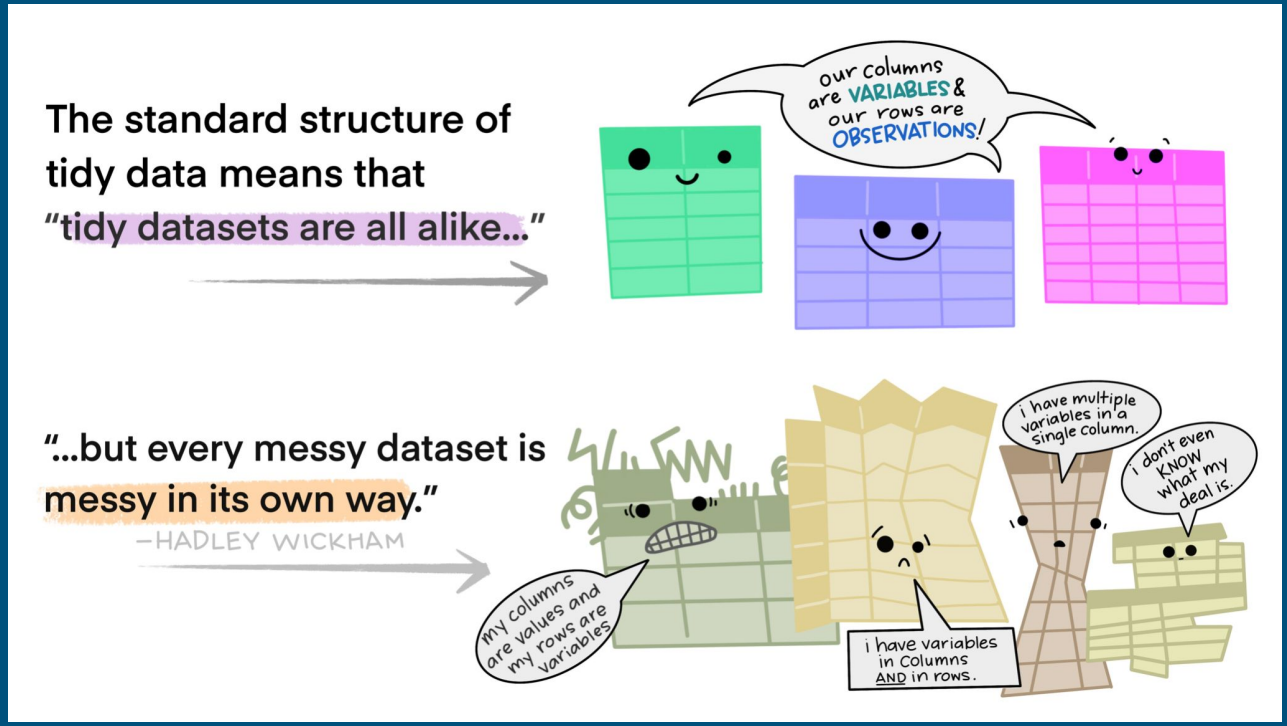
Initial Steps Toward Reproducible Research

Example Project Organization

CITATION	
README	Processing steps
LICENSE	
Requirements.txt	Description of what the code need to run
data/	Directory for original data
|-- birds_count_table.csv	
doc/	
|-- notebook.m	Directory for write-up of results
|-- manuscript.md	
|-- changelog.txt	
results/	Directory for analysis output
|-- summarized_results.csv	
src/	
|-- sightings_analysis.py	Directory for code that does the analysis
|-- runall.py	

Good enough practices in scientific computing

File Organization - Tidy Data



https://www.openscapes.org/blog/2020/10/12/tidy-data/

Tidy Data

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(10), 1 - 23. doi:http://dx.doi.org/10.18637/jss.v059.i10

Not Tidy - Column headers are values

religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

Table 4: The first ten rows of data on income and religion from the Pew Forum. Three columns, `$75-100k`, `$100-150k` and `>150k`, have been omitted

This dataset has three variables, religion, income and frequency

Tidy - Each Variable is a Column

religion	income	freq
Agnostic	<$10k	27
Agnostic	$10-20k	34
Agnostic	$20-30k	60
Agnostic	$30-40k	81
Agnostic	$40-50k	76
Agnostic	$50-75k	137
...		

Working with Spreadsheets

1. Put all your variables in columns - the thing you're measuring, like 'weight' or 'temperature'.
2. Put each observation in its own row.
3. Don't combine multiple pieces of information in one cell.
    a. Be able to use or sort that data.
    b. Units as metadata
4. Export the cleaned data to a text-based format like CSV (comma-separated values) format. This ensures that anyone can use the data, and is required by most data repositories.

Data Organization in Spreadsheets for Ecologists

Example Dataset

“The data used in the ecology lessons are observations of a small mammal community in southern Arizona. This is part of a project studying the effects of rodents and ants on the plant community that has been running for almost 40 years. The rodents are sampled on a series of 24 plots, with different experimental manipulations controlling which rodents are allowed to access which plots.

This is a real dataset that has been used in over 100 publications. We’ve simplified it just a little bit for the workshop, but you can download the full dataset and work with it ...”

Notes from Data Organization in Spreadsheets for Ecologists

Conclusions

- Good data organization is the foundation of any research project.
- Assume you are going to mess up.
  - Never modify raw data
  - Keep records of every step
  - Check every step

Finish-up

- Assignment #2 on GitHub - XML for Metadata

- Next week we will be at a this Zoom Link again

Resources

- Data Organization - organizing data in spreadsheets
- Initial Steps Toward Reproducible Research
- Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(10), 1 - 23. doi:http://dx.doi.org/10.18637/jss.v059.i10
- Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK (2017) Good enough practices in scientific computing. PLoS Comput Biol 13(6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510
- Data Carpentries: Data Organization in Spreadsheets for Ecologists