

Spreadsheets:

Best Practices

Bianca Peterson (PhD)

Conquest Analytics and Training

29 August 2021

Introduction

The following lesson is based on the [Data Carpentry Ecology lesson](#)

Introduction

- Good data organization is the foundation of any project
- Structure data the way that computers need it
- Much of your time will be spent in this 'data wrangling' stage

We can use spreadsheets for:

- Data entry
- Organizing data
- Subsetting and sorting data
- Statistics
- Plotting, but...

Problems with spreadsheets

- The graphical, drag and drop nature of spreadsheet programs makes it very difficult, if not impossible, to replicate your steps
- When doing calculations in a spreadsheet, it's easy to accidentally apply a formula to incorrect cells
- Difficult to troubleshoot errors if you have interrelated spreadsheet data scattered across multiple locations
- Time-consuming "manual" work - unfit for agile business practices
- Hard to consolidate spreadsheets distributed throughout an organisation
- Prone to trivial human errors, e.g. copy-pasting, cell entry, and range specification

Common mistakes

[1] Using multiple tables on a single sheet

- Humans can (usually) interpret these things, but computers don't view information the same way
- You're drawing false associations between things for the computer, which sees each row as an observation
- You're potentially using the same field name in multiple places, which makes it harder to clean the data up into a usable form
- We have to set up our data for the computer to understand it in order to analyze it fast and efficiently

Common mistakes

[2] Using multiple sheets

- The computer will not be able to see connections in the data
- You are more likely to accidentally add inconsistencies to your data when recording data in a new tab every time
- This will add an extra step for yourself, because data will need to be combined before analyzing it
- Ask yourself: Can I avoid adding a new tab by adding another column to the original spreadsheet?
- The data sheet may get very long - freeze column headers so that they remain visible

Common mistakes

[3] Not filling in zeros

- There's a difference between a zero and a blank cell
- Zero = actual data | blank cell = no measurement
- Missing/null/unknown values can cause problems with subsequent calculations or analyses

Common mistakes

[4] Using problematic null values

Null values	Problems	Compatibility	Recommendation
0	Indistinguishable from a true zero		Never use
blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently.	R, Python, SQL	Best option
999, -999	Not recognized as null by many programs without user input. Can be inadvertently entered into calculations.		Avoid
NA, na	Can also be an abbreviation (e.g., North America), can cause problems with data type (turn a numerical column into a text column). NA is more commonly recognized than na.	R	Good option
N/A	An alternate form of NA, but often not compatible with software		Avoid
NULL	Can cause problems with data type	SQL	Good option
None	Can cause problems with data type	Python	Avoid
No data	Can cause problems with data type, contains a space		Avoid

Article: Nine simple ways to make it easier to (re)use your data by White et al., 2013
<https://peerj.com/preprints/7/>

Common mistakes

[5] Using formatting to convey information

Date	Code	Sex	Weight	
1/8/2014	NA			
1/9/2014	DM	M	87	
1/10/2014	DM	M	98	
1/11/2014	OL			
1/12/2014	PE	M	77	
1/13/2014	DM	M	72	
1/14/2014	PF	M	84	
1/15/2014	OT	F	56	
1/16/2014	DS	M	91	
1/17/2014	OX	F	52	
1/18/2014	NA	F	63	
1/19/2014	PF	F	59	
1/20/2014	DM			
1/21/2014	PE	F	66	
1/22/2014	OT	F	68	
	Device wasn't calibrated correctly			

Date	Code	Sex	Weight	Calibrated
1/8/2014	NA			
1/9/2014	DM	M	87	Y
1/10/2014	DM	M	98	Y
1/11/2014	OL			
1/12/2014	PE	M	77	Y
1/13/2014	DM	M	72	Y
1/14/2014	PF	M	84	Y
1/15/2014	OT	F	56	Y
1/16/2014	DS	M	91	Y
1/17/2014	OX	F	52	Y
1/18/2014	NA	F	63	N
1/19/2014	PF	F	59	Y
1/20/2014	DM			
1/21/2014	PE	F	66	N
1/22/2014	OT	F	68	Y

Common mistakes

[6] Merging cells

- Computers need tidy data:
 - Each column should contain only one variable
 - Each row should contain only one observation
 - Each cell should contain a single value (whether it's text or numbers)
- Solution: restructure your data in such a way that you will not need to merge cells to organize your data

Common mistakes

[7] Placing comments or units in cells

- The computer can't do calculations when cells contain comments or units (text)
- Solution: create another field to add notes or units to cells

Common mistakes

[8] Entering more than one piece of information in a cell

- It's difficult/impossible to do counts/calculations
- Solution: create another column to record all the information you need
- It's easy to combine data later if needed

Common mistakes

[9] Using problematic field names

Avoid
Maximum Temp (°C)
precmm
Mean growth/year
M/F
w.
Cell Type
1st Obs

Common mistakes

[9] Using problematic field names (continued)

- Spaces can be misinterpreted by parsers that use whitespace as delimiters
- Some programs don't like field names that are text strings that start with numbers
- Underscores (_) are a good alternative to spaces
- Consider writing names in camel case (like this: `ExampleFileName`) or snake case (like this: `Example_File_Name`) to improve readability
- Do not include spaces, special characters of any kind, and if you want to add numbers, make sure it's not at the beginning
- Be descriptive

Common mistakes

[10] Using special characters in data

- Examples:
 - line breaks when writing longer text in a cell
 - when copying data in from other applications such as Word
 - using fancy non-standard characters (such as left- and right-aligned quotation marks)
- Solution: avoid adding characters such as newlines, tabs, and vertical tabs; only use text and spaces

Common mistakes

[11] Inclusion of metadata in data table

- Example: adding legends to the top or bottom of your data table to explain column meaning, units, exceptions, etc.
- Solution: recording data about your data ("metadata") is essential, but should not be contained in the data file itself
- Your future self and other people may want to examine or use your data to understand your findings, verify your findings, replicate your results, design a similar study, or archive your data for access and re-use
- Store metadata as a separate plain text file in the same directory as your data file

Common mistakes

[12] Storing dates in a single column

- Not best practice to store dates in a single column
- How does Excel read dates? [DEMO]
- According to your computer's settings...
- Excel is unable to parse dates from before 1899-12-31 (referred to as the 1900 date system) - be careful if you're working with historical data
- Alternatively, Excel may use the 1904 date system where a different serial number is assigned
- Be careful when exporting data from Excel - dates may be ~4 years off!
- Solution: Store MONTH, DAY and YEAR in separate columns

Exercise

Download and open the `survey_data_spreadsheet_messy.xlsx` data file

According to best practices...

- Identify things that are correct
- Identify things that need to be changed/corrected

How do we fix it?

- Make sure your data is tidy (not pretty!)
- Implement quality assurance to only allow valid values [DEMO]
- Sort your data - bad values often sort to the bottom or top of the column (sort data by each field, one at a time)
- Remember to expand your sort to prevent data corruption, i.e. all the data in one row should move together
- Conditional formatting - color code values by some criteria or lowest to highest
- This makes it easy to scan your data for outliers

Exporting data

Storing data in Excel default file format (*.xls or .xlsx*) isn't a good idea, because...

- It's proprietary format - in future it might become inconvenient/impossible to open the file
- Other spreadsheet software may not be able to open files in this format
- Different versions of Excel may handle data differently, leading to inconsistencies
- More journals and grant agencies are requiring data to be deposited in a data repository, and most of them don't accept Excel format
- Use tabs as delimiter (*.tsv file*) when the data contains commas

Things to keep in mind

- Never modify your raw data - always make a copy before making any changes
- Keep track of all the steps you take to clean your data
- Be careful when using commas - Excel might import data incorrectly from CSV files
- Use . for decimal separator and , for thousands separator (set Options in Excel)