The components of tidy data

Jeffrey Leek, Assistant Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

"Tidy data is your target
during data processing and
cleaning

The four things you should have *(when you've finished going from a raw data set to a tidy data set)*

1. The raw data. *Files*
2. A tidy data set *Finished product*
3. A code book describing each variable and its values in the tidy data set. *Metadata*
4. An explicit and exact recipe you used to go from 1 -> 2,3.

3) Describe variables, units of variables
4) Very important, large focus in this class
    ↳ We will be using R scripts

The raw data

- The strange binary file your measurement machine spits out
- The unformatted Excel file with 10 worksheets the company you contracted with sent you
- The complicated JSON data you got from scraping the Twitter API
- The hand-entered numbers you collected looking through a microscope

You know the raw data is in the right format if you

1. Ran no software on the data
2. Did not manipulate any of the numbers in the data
3. You did not remove any data from the data set
4. You did not summarize the data in any way

https://github.com/jtleek/datasharing

One component of your data pipeline

Different levels of "raw" data

For your purposes with a particular dataset it is the rawest form of data that you had access to

The tidy data

1. Each variable you measure should be in one column
2. Each different observation of that variable should be in a different row
3. There should be one table for each "kind" of variable
4. If you have multiple tables, they should include a column in the table that allows them to be linked

Some other important tips

· Include a row at the top of each file with variable names.

· Make variable names human readable AgeAtDiagnosis instead of AgeDx

· In general data should be saved in one file per table.

https://github.com/jtleek/datasharing

4/7

) Fit nice in a
⌐Target/end goal
data frame
3] Ex. one for twitter, one for
facebook, etc
4] Column ID

~ very useful

~ ex. Don't do equivalent of having multiple Excel
spreadsheets/pages in one file
↳ One table per R script

The code book

1. Information about the variables (including units!) in the data set not contained in the tidy data
2. Information about the summary choices you made
3. Information about the experimental study design you used

Some other important tips

· A common format for this document is a Word/text file.

· There should be a section called "Study design" that has a thorough description of how you collected the data.

· There must be a section called "Code book" that describes each variable and its units.

https://github.com/jtleek/datasharing

Markdown file

↳ A piece that's often missing

How you picked what obervations to collect, what you extracted out of the database, what did you exclude, etc.

The instruction list

- Ideally a computer script (in R :-), but I suppose Python is ok too...)

- The input for the script is the raw data

- The output is the processed, tidy data

- There are no parameters to the script — *you fix everything, so end user doesn't have to do anything to exactly replicate your tidy dataset*

In some cases it will not be possible to script every step. In that case you should provide instructions like:

1. Step 1 - take the raw file, run version 3.1.2 of summarize software with parameters a=1, b=2, c=3

2. Step 2 - run the software separately for each sample

3. Step 3 - take column three of outputfile.out for each sample and that is the corresponding row in the output data set

https://github.com/jtleek/datasharing

↳ should be able to exactly replicate your tidy data set from the raw data set that you used

↳ Markdown file
↳ Be extremely detailed

Why is the instruction list important?

Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff

Thomas Herndon*   Michael Ash   Robert Pollin

April 15, 2013

→ grad student who got ahold of the excel file they used and found lots of errors in their data processing
↳ called paper and subsequent political decisions (re: austerity) into serious question

http://www.colbertnation.com/the-colbert-report-videos/425748/april-23-2013/austerity-s-spreadsheet-error