DataSci Cleaning Data Lecture Notes: Components of tidy data

4 things you should have after raw data processing

1. The raw data
2. A tidy data set
3. A code book explaining each variable and its values in the tidy data set
   1. Aka the Meta-data
   2. Explains what the data is trying to say
   3. For example, Each column in the tidy data corresponds to a variable; the code book could tell the units of the variable
4. The explicit and exact processes you used to go from 1 -> 2-> 3

Raw Data Examples

1. Strange binary files that a measurement machine spits out
2. Unformatted Excel files with 10 worksheets
3. Twitter’s JSON data formatted API feed
4. Hand entered numbers collected by researchers

Evidence Raw Data is in the Correct Format

1. No software was run on the data
2. The numbers were not manipulated in any way
3. No data was removed from the set
4. The data has not been summarized in any way

Tidy Data Examples

1. Each variable measured has its own column
2. Each different observation of that variable has its own row
3. Each “kind” of variable has its own table
   1. Like a table for Facebook data and a table for Twitter data
4. Multiple table databases should include columns in each table that allows them to be linked
   1. Such as IDs and Foreign Keys in MySQL

Tidy Data Tips and Tricks

1. Include a row at the top of each file with variable names
2. Variables should have easy to understand names (AgeAtDiagnosis vs AgeDx)
3. Data should be saved one file per table (not all this subsheet crap in Excel)

The Code Book

1. Info about the varibles
   1. Include the units
2. Info about the summary choices you made
   1. Such as variable measured against monthly revenue vs quarterly revenue
3. Info about the experimental study designed used
   1. Was this observational studies, random samplings, etc etc

Code Book Tips and Tricks

1. Use a common format such as Word or Text (also markdown files)
2. Include a section called “Study Design” that has a thorough description of data collection methods
3. Include a description of each variable and its units in its own section like an index

Instruction List

1. Ideally a computer script such as R or Python should be included for reproducibility should be included
2. Include the input for the script as the original raw data
3. The output should always be processed and tidy data
4. The script has no parameters
   1. In another words, the end user should not have to put in any other parameters for your script to work
5. If a script is not possible to include as part of the instructions, write the instructions out
   1. Such as – step 1, take raw file, run version xyz of a software with parameters a, b, c.
   2. Step 2 – run software separately for each sample
   3. Step 3, take column v of output file for each sample as the corresponding row of the data set

Why an Instruction List is important

1. Misunderstanding between code writer and end user can be disastrous. Just see Comedy Central’s evening news shows.
2. Listing your steps can help others find errors