Data Cleaning Checklist

Nick Hagerty

September 22, 2022

1. **Convert file formats**, as necessary.
2. **Import data and wrangle into a tidy layout.**
3. **Remove irrelevant, garbage, or empty** columns and rows.
4. **Identify the primary key**, or define a surrogate key.
5. **Resolve duplicates** (remove true duplicates, or redefine the primary key).
6. **Understand the definition, origin, and units** of each variable, and document as necessary.
7. **Rename variables** as necessary, to be succinct and descriptive.
8. **Understand patterns of missing values.**
    - Find out why they're missing.
    - Make sure they are not more widespread than you expect.
    - Convert other intended designations (i.e., -1 or -999) to NA.
    - Distinguish between missing values and true zeros.
9. **Convert to numeric** when variables are inappropriately stored as strings. Correct typos as necessary.
10. **Convert to date/time** format where appropriate.
11. **Recode binary variables** as 0/1 as necessary. (Often stored as "Yes"/"No" or 1/2.)
12. **Convert to factors** when strings take a limited set of possible values.
13. **Make units and scales consistent.** Avoid having in the same variable:
    - Some values in meters and others in feet.
    - Some values in USD and others in GBP.
    - Some percentages as 40% and others as 0.4.
    - Some values as millions and others as billions.
14. **Perform logical checks on quantitative variables:**
    - Define any range restrictions each variable should satisfy, and check them.
    - Correct any violations that are indisputable data entry mistakes.
    - Create a flag variable to mark remaining violations.
15. **Clean string variables.** Some common operations:
    - Make entirely uppercase or lowercase
    - Remove punctuation
    - Trim spaces (extra, starting, ending)
    - Ensure order of names is consistent
    - Remove uninformative words like "the" and "a"
    - Correct spelling inconsistencies (consider text clustering packages)

16. **Literally look at your data** tables regularly, throughout the entire process, to spot issues you haven't thought of.
17. **Save your clean data** to disk before further manipulation (merging dataframes, transforming variables, restricting the sample). Think of the whole wrangling/cleaning/analysis pipeline as 2 big phases:
    - Taking messy data from external sources and making a nice, neat table that you are likely to use for multiple purposes in analysis.
    - Taking that nice, neat table and doing all kinds of new things with it.

Guidelines that apply throughout:

- Record all steps in a script.
- Never overwrite the original raw data file.
- Whenever possible, make changes to values ONLY by logical conditions on one or more substantive variables – not by observation ID, another key, or (even worse) row number. You want the changes you make to be rule-based, for 2 reasons:
  - So that they're general – able to handle upstream changes to the data.
  - So that they're principled – no one can accuse you of cherry-picking.