Data Cleaning Checklist

Nick Hagerty

September 2024

1. **Convert file formats** as necessary, and import your data.
2. **Structure data into tidy format** if not already.
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. **Convert variable formats** as necessary:
    - Numeric variables – may be inappropriately stored as strings when there are typos.
    - Dates and times – store in date or time format.
    - Binary variables – code as 0/1 integers (not "Yes"/"No" or 1/2).
    - Factors – use when strings take a limited set of possible values.
    - ID variables – store as integers or character, not numeric.
    - Strings of digits – store as character, not numeric.
9. **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.
10. **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.
11. **Enforce logical conditions 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.
12. **Clean string variables** if necessary. 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)

13. **Save your clean data** to disk before further manipulation (merging data, transforming variables, restricting the sample). Think of the whole wrangling/cleaning/analysis pipeline as 2 distinct phases:

    a. Taking messy data from external sources and making a nice, neat table that you are likely to use for multiple purposes in analysis.
    b. Taking that nice, neat table and doing all kinds of new things with it.

Guidelines that apply throughout:

1. **When editing values, identify observations by substantive logical conditions** rather than by observation ID 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.

2. **Record all steps in a script.**

3. **Never overwrite the original raw data file.**

4. **Look at your data** every step of the way, to spot issues you haven't thought of, and to make sure you're actually doing what you think you're doing.