DATA 227

Tidy Data

2022-09-29

Motivating Tidy Data

"No matter the project, whether in government, academia, or industry, we always ended up with the same problem: We needed to clean the data before we could do the data science." Jeff Leek, simplystatistcs.org

Tidy Data 1

“Tidy datasets provide a standardized way to link the structure of a dataset (its physical layout) with its semantics (its meaning).”

Why ensure that your data is tidy?

- A general advantage to picking one consistent way of storing data.
  - If you have a consistent data structure, it’s easier to learn the tools that work with it because they have an underlying uniformity.
  - “Tidy datasets and tidy tools work hand in hand to make data analysis easier, allowing you to focus on the interesting domain problem, not on the uninteresting logistics of data”

Wickham, H. . (2014). Tidy Data. Journal of Statistical Software, 59(10), 1–23. https://doi.org/10.18637/jss.v059.i10

Data Semantics

- A dataset is a collection of *values*, usually either numbers (if quantitative) or strings (if qualitative).
- Values are organized in two ways. Every value belongs to a variable and an observation.
    - A *variable* contains all values that measure the same underlying attribute (like height, temperature, duration) across units.
    - An *observation* contains all values measured on the same unit (like a person, or a day, or a race) across attributes.

Wickham, H. . (2014).

Tidy Data 2

There are three interrelated rules which make a dataset tidy:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

Wickham, H. . (2014).

Common Problems in Untidy Data 1

1. Column headers are values, not variable names (shows up frequently in tabular data).

```
##    Row  Col  Hgt90  Hgt96  ...       Deer97  Cover95  Fert  Spacing
## 0    1    1    NaN    NaN  ... -2147483648        0     0        15
## 1    1    2   14.0  284.0  ...           1        2     0        15
## 2    1    3   17.0  387.0  ...           0        1     0        15
## 3    1    4    NaN    NaN  ... -2147483648        0     0        15
## 4    1    5   24.0  294.0  ...           0        2     0        15
## 5    1    6   22.0  310.0  ...           0        1     0        15
## 6    1    7   18.0  318.0  ...           0        0     0        15
## 7    1    8   32.0  328.0  ...           0        1     0        15
##
## [8 rows x 15 columns]
```

Common Problems in Untidy Data 2

- Multiple variables are stored in one column.

```
##                                 Name  ...                        Value
## 0  'C' A CATERING AND EVENT COMPANY  ...  4300-4304 N CENTRAL AVE 
## 1  'C' A CATERING AND EVENT COMPANY  ...                        60634
## 2  'C' A CATERING AND EVENT COMPANY  ...                   Restaurant
## 3  'C' A CATERING AND EVENT COMPANY  ...                     No Entry
## 4  'C' A CATERING AND EVENT COMPANY  ...  4300-4304 N CENTRAL AVE 
## 5  'C' A CATERING AND EVENT COMPANY  ...                        60634
## 6  'C' A CATERING AND EVENT COMPANY  ...                   Restaurant
## 7  'C' A CATERING AND EVENT COMPANY  ...                         Pass
##
## [8 rows x 4 columns]
```

Common Problems in Untidy Data 3

- Variables are stored in both rows and columns.

```
##                       V1   V2   V3   V4   V5  ... V59 V60 V61
## 0  MX000017004195504TMAX  310  310  310  320  ...             -214
## 1  MX000017004195504TMIN  150  150  160  150  ...             -214
## 2  MX000017004195504PRCP    0    0    0    0  ...             -214
## 3  MX000017004195505TMAX  310  310  310  300  ...             -214
## 4  MX000017004195505TMIN  200  160  160  150  ...             -214
## 5  MX000017004195505PRCP    0    0    0    0  ...             -214
## 6  MX000017004195506TMAX  300  290  280  270  ...             -214
## 7  MX000017004195506TMIN  160  160  150  140  ...             -214
##
## [8 rows x 63 columns]
```

Common Problems in Untidy Data 4

- Multiple types of observational units are stored in the same table.

```
## [’artist’, ’track’, ’date.entered’, ’wk1’, ’wk2’, ’wk3’, ’wk4’]
## [’wk6’, ’wk7’, ’wk8’, ’wk9’, ’wk10’, ’wk11’, ’wk12’, ’wk13’]
## [’wk14’, ’wk15’, ’wk16’, ’wk17’, ’wk18’, ’wk19’, ’wk20’, ’wk21’]
## [’wk22’, ’wk23’, ’wk24’, ’wk25’, ’wk26’, ’wk27’, ’wk28’, ’wk29’]
## [’wk30’, ’wk31’, ’wk32’, ’wk33’, ’wk34’, ’wk35’, ’wk36’, ’wk37’]
## [’wk38’, ’wk39’, ’wk40’, ’wk41’, ’wk42’, ’wk43’, ’wk44’, ’wk45’]
## [’wk46’, ’wk47’, ’wk48’, ’wk49’, ’wk50’, ’wk51’, ’wk52’, ’wk53’]
## [’wk54’, ’wk55’, ’wk56’, ’wk57’, ’wk58’, ’wk59’, ’wk60’, ’wk61’]
```

Common Problems in Untidy Data 5

- A single observational unit is stored in multiple tables or files–these tables and files are often split up by another variable, so that each represents a single year, person, or location, etc.
    - As long as the format for individual records is consistent, this is an easy problem to fix.
- For example, consider the Household Pulse Survey run by the US Census Bureau.
    - "Designed to quickly and efficiently deploy data collected on how people's lives have been impacted by the coronavirus pandemic."
    - You can visit the Household Pulse Survey website to download multiple files, one for every survey collection period, that could be combined to examine trends over time.

Data Cleaning

1. Familiarize yourself with the data set.
2. Check for structural errors (more on this in a bit).
3. Check for data irregularities (use plots, summary statistics, etc.).
4. Decide how to deal with missing values (beyond the scope of this class).
5. Document data versions and changes made.

Source: Towards Data Science Blog

Familiarize yourself with the data set.

- Domain knowledge, domain knowledge, domain knowledge! Have a good grasp on what your variables mean, which are important, and which need cleaning.
  - Ask as many questions of your collaborator(s) as you need.
  - Do this sooner, rather than later!
- General steps:
  - Check the file size.
  - Check the dimensions of your dataset.
  - Check the first few rows of your dataset and the storage types of the variables to make sure you have read it in correctly.

Check for structural errors.

- Structural Errors: faulty data types, non-unique ID numbers, mislabeled variables, string inconsistencies, etc. It's hard to give you a full list, since there are so many ways a dataset can be messy!
- General steps
  - Check the names of the columns in your dataset and rename them if necessary.
  - Check the storage formats of the variables.
  - Look for duplicate rows.
  - Check variations in a column (e.g., `"female"`, `"Female"`, `"F"`, etc.)

Check for data irregularities.

- Data Irregularities: accuracy concerns like invalid values and outliers.
- Invalid Values: Values that don't make sense.
  - Individual Values: E.g., 19 pounds for a woman over 18, ages under 15 for a dataset of employees
  - Across Columns: E.g., Sexual orientation, gender, and household partners
- Outliers:
  - Repeat analysis with and without the point, and see if the results change.

Missing Values.

There are two types of missing values:

1. Explicitly, i.e. flagged with `None`, `NaN`, `NA`, etc. Strategizing about missing values is beyond the scope of this class.
2. Implicitly, i.e. simply not present in the data.

- One bad habit is entering a year in an early row and leaving later rows empty because they should be filled with the value (and just are not, for some reason).

We like explicit values, a lot! It's much more helpful to know if something was left missing intentionally.

General Advice

If you discover any of these errors in a data file, never modify it directly, but instead write code to correct the value and explain why you made the fix (R for Data Science).

Tools in R

- `tidyr` is a package that makes it easy to “tidy” your data.
  - Helpful commands in `tidyr` include `pivot_longer()`, `pivot_wider()`, `separate()`, `unite()`, `complete()`, and `fill()`.
- `dplyr` is a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges.
  - The most common “verbs” in `dplyr` are `mutate()`, `select()`, `filter()`, `summarise()`, and `arrange()`.
  - These all combine naturally with `group_by()`, which allows you to perform any operation “by group”.
  - Other helpful `dplyr` functions: `rownames_to_columns()`, `columns_to_rownames()`, `has_rownames()`, `remove_rownames()`, `bind()`, `bind_cols()`, `bind_rows()`, `left_join()`, `right_join()`, `inner_join()`, `full_join()`, `count()`, `tally()`.

Source 1: https://www.rstudio.com/blog/introducing-tidyr/, Source 2: https://dplyr.tidyverse.org/

Tools in Python

- You can convert the key `dplyr` verbs to Python code using `pandas`– see this tutorial.

Additional Tools

Don't discount additional tools!

- Excel. Lots of statisticians look down on Excel, but many people use it, so it's good to have a working knowledge. Plus, there are some useful procedures!
  - Text-to-Data (demo)
  - Write-protecting (instructions in Data Organization in Spreadsheets)
  - Data-Validation (same as above)
- Library Carpentry, OpenRefine
- SQL

Tidy Data 3

Some people use the term “messy” to refer to non-tidy data. This is an oversimplification: there are lots of useful and well-founded data structures that are not tidy data.

1. Alternative representations may have substantial performance or space advantages.
2. Specialized fields have evolved their own conventions for storing data that may be quite different to the conventions of tidy data.

- Expression sets for expression data.
- Summarized experiments for a variety of genomic experiments.
- Granges Lists for genomic intervals.
- Corpus objects for corpora of texts.
- `igraph` objects for graphs.

Source: Jeff Leek, simplystatistics.org

Creating Data 1

- Be consistent.
    - Use a consistent data layout in multiple files (make it tidy!).
    - Use consistent codes for categorical variables.
    - Use a consistent fixed code for any missing values.
    - Use consistent variable names.
    - Use consistent subject identifiers.
    - Use consistent file names.
    - Use a consistent format for all dates.
    - Use consistent phrases in your notes.
    - Be careful about extra spaces within cells.

“Whatever you do, do it consistently. Entering and organizing your data in a consistent way from the start will prevent you and your collaborators from having to spend time harmonizing the data later.”

Creating Data 2

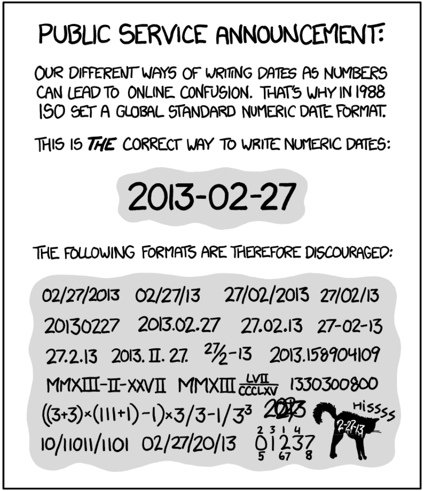
- Choose good names for things.
  - As a general rule, don't use spaces, either in variable names or file names
  - Where you might uses paces, use underscores or perhaps hyphens.
  - Be careful about extraneous spaces at the beginning or end of a variable name.
  - Avoid special characters ($,@,%,#,&,*,(,),!,/, etc.), except for underscores and hyphens.
  - The main principle in choosing names, whether for variables or for file names, is short, but meaningful.
  - Finally, never include "final" in a file name.

Creating Data 3

- Write dates like YYYY-MM-DD.

"When entering dates, we strongly recommend using the global ISO 8601 standard, YYYY-MM-DD, such as 2013-02-27."

If you like, you can read about some horror stories coming from spreadsheet errors. Many of them involve incorrect date formatting.


PUBLIC SERVICE ANNOUNCEMENT:
OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.
THIS IS THE CORRECT WAY TO WRITE NUMERIC DATES:
2013-02-27
THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:
02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. 27/2-13 2013.158904109
MMXIII-II-XXVII MMXIII LVII/CCCLXV 1330300800
((3+3)×(111+1)-1)×3/3-1/3³ 2013
10/11011/1101 02/27/20/13 01237
HISSSS
2-27-13

Creating Data 4

- Create a data dictionary. It is helpful to have a separate file that explains what all of the variables are.
- Such a “data dictionary” might contain:
  - The exact variable name as in the data file.
  - A version of the variable name that might be used in data visualizations.
  - A longer explanation of what the variable means.
  - The measurement units.
  - Expected minimum and maximum values (helpful for data cleaning or identifying outliers).

This is part of the metadata that you will want to prepare: information about the data.

Creating Data 5

- Don't leave any cells empty, and use some common code for missing data.
- Don't include calculations in the raw data.
- Don't use font color or highlighting as data.
- Make backups.
- Use data validation to avoid data entry errors.
- Save the data in plain text (usually, a .csv file).

These tips come from Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. The American Statistician, 72(1), 2-10.

Final Advice 1

Adventures in Education Data Wrangling

- Know your data!
  - *I spent several hours reading through column headers and looking at how the data was organized before starting to actually wrangle the data. This time spent familiarizing myself with the data made it easier to understand what I ultimately needed to do with the data, as well as articulate my questions to others when something wasn't working correctly.*
- Be tenacious!
  - *Being confident in my ability to figure things out helped keep the impending sense of panic at bay [. . .] It helped to remind myself that with a couple of good search terms and 20 minutes of reading, the answer can almost always be found.*

Final Advice 2

- Know when to ask for help.
    - *I got to a point in the wrangling where I knew what I needed to do with the data, but I couldn't find an answer online that I both understood and could get to work. After working at it for 30 minutes, I reached out to a couple of friends who are more fluent in R than I am, and they had the issue sorted within minutes.*
- Recognize when you're headed down a rabbit hole.
    - *This is where I'll spend all of my time if I'm not careful, because it's so darn easy to do. When I found myself reading through a weird online argument about coalesce vs. coalesce2, I knew that I had taken a wrong turn about 30 clicks back and needed to re-focus on the task at hand.*

Sources, again

- Adventures in Education Data Wrangling
- Data Organization in Spreadsheets
- dplyr
- R for Data Science: R Markdown Workflow)
- R for Data Science: Tidy Data
- simplystatistics.org
- Spreadsheet Horror Stories
- Towards Data Science