Introduction to Data Wrangling

Election Data Science

Peter Licari

2020-09-15

The fact is, most of your time will be spent cleaning data and preparing data *for* analysis, rather than doing the analysis.

Data wrangling

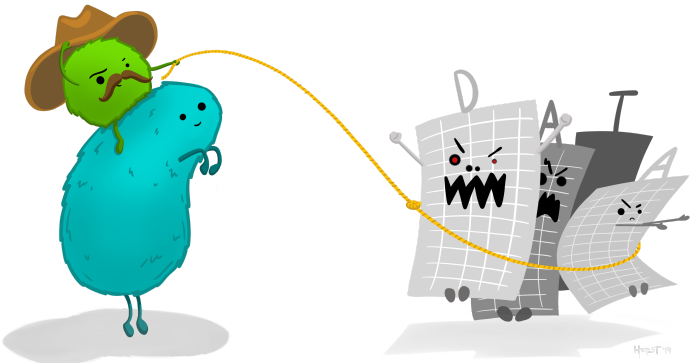

Image Credit: Allison Horst

- **Wikipedia:** "Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics."

- **Urban Dictionary:** "The act of consolidating 2 or more mutually exclusive data sets or sections of computer code, circumventing the requirement to write a shit load of complex code."

Helpful to return back to $f(x)$

- $f(1) = 4$. What are the functional steps?
  - $f(x) = 4x$
  - $f(x) = 3x + 1$
  - $f(x) = 4 \times \sum_{n=1}^{\infty} (\frac{1}{2})^n$
  - $f(x) = 4$

There are an infinite number of ways to get a particular output. The *best* ones are the ones that balance simplicity (e.g., ease of steps) and parsimony (e.g., number of steps).


tidyverse
www.rstudio.com

The Tidyverse is a series of packages that helps you wrangle, visualize, and analyze "tidy" data.

There are three criteria for tidy data:

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

— Hadley Wickham (2014)

Tidy elections data set.

Adams County Code

Search:

	SOS_VOTERID	COUNTY_NUMBER	COUNTY_ID	LAST_NAME	FIR
1	OH0023280066	01	29434	JAMES	EL
2	OH0016306559	01	23134	PISTOLE	BE
3	OH0022673930	01	33139	AYERS	NC

Showing 1 to 100 of 100 entries

Just because your data is *tidy* doesn't mean you're done wrangling.

Data wrangling is an iterative process that you repeat as new needs emerge in your exploration and analysis.

Example: Voters from "West Union"

Orignal Data | Transformation Code | New Table

Search:

	SOS_VOTERID	COUNTY_NUMBER	COUNTY_ID	LAST_NAME	FIR
1	OH0023280066	01	29434	JAMES	EL
2	OH0016306559	01	23134	PISTOLE	BE
3	OH0022673930	01	33139	AYERS	NC

Showing 1 to 100 of 100 entries

The best way to wrangle your data is to start at the **end**. What do you want it to look like? What are the steps immediately preceding that final output? How can you get from where you are to those steps?