Data Wrangling in R

Andrew Frewin

afrewin@uoguelph.ca

Data wrangling is the process of transforming data into a format appropriate for a particular task

Data is seldom in a useable form

consider this statement and think about data you have collected yourself or been given

Was this data fit for purpose?

What steps did you have to take to make it useable?

Data is seldom in a useable form

General cases

1. new variables
2. new arrangement
3. combining data

How have **you** manipulated data?

Spreadsheets…

1. spreadsheets are for data entry and storage

2. analysis and visualization should happen separately

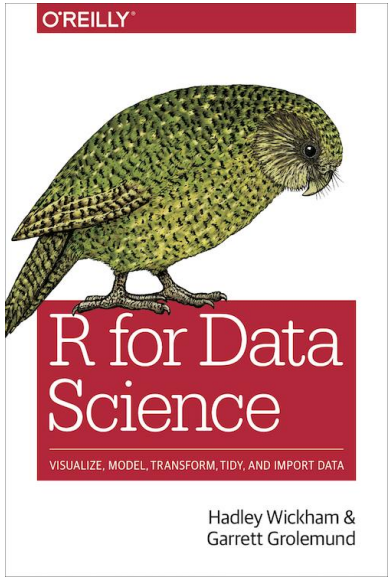
➢ “reduces the risk of **contaminating** or destroying data”

“Data organization in spreadsheets” – Broman & Woo 2017 PeerJ

https://peerj.com/preprints/3183/

Tidy data

1. Each variable must have its own column
2. Each observation must have its own row
3. Each value must have its own cell



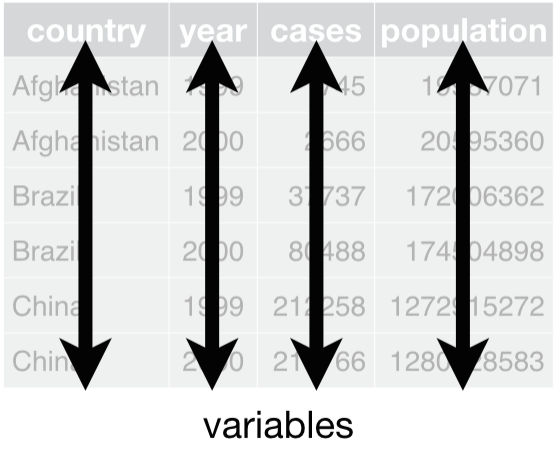
Content from this excellent book.

Why tidy data?

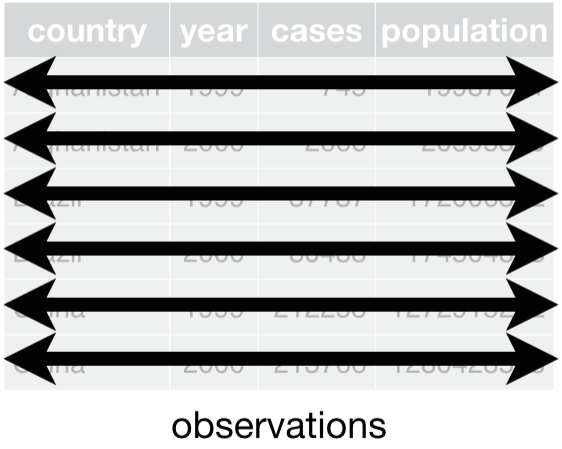
- easier to apply tools to data with a similar structure
- variables in columns, “allows R’s … nature to shine”

1. Each **variable** must have its own column
2. Each **observation** must have its own row
3. Each **value** must have its own cell

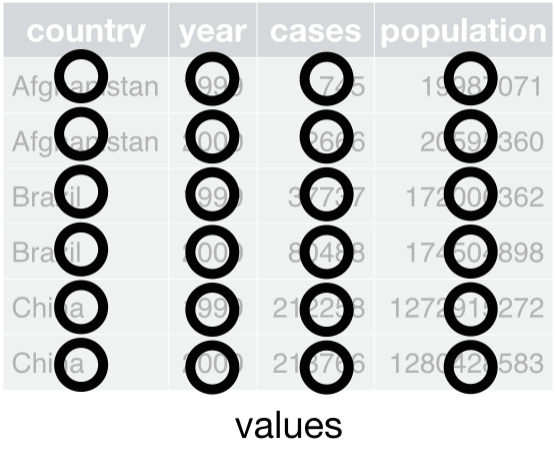
country	year	cases	populations
Afghanistan	1999	745	19987071
Afghanistan	2000	266	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	127291572
China	2000	213766	1280428583



variables



observations



values

library(tidyverse)

-library(dplyr)

filter() – pick observations by their values

select() – pick variables by their names

group_by() – creates a grouping structure

summarize() – summarize many values

-library(tidyr)

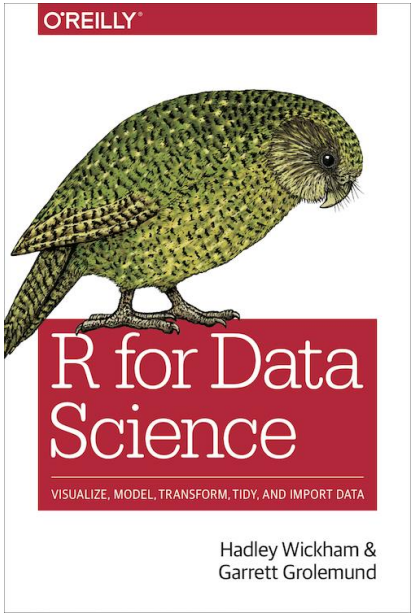
spread() – two columns to multiple columns

- makes long data wider

gather() – combines multiple columns

- makes wide data longer

For more information:



- Chapter 3: Data Transformations with dplyr
- Chapter 9: Tidy Data with tidyr

Data organization in spreadsheets
https://peerj.com/preprints/3183/

European Spreadsheet Risk Interest Group
http://www.eusprig.org/