Tidy data

Data are tidy if...

Each variable forms a column

Each case forms a row

Each value of a variable is stored in a cell

No footnotes

No units

Why do we care?

Tidy data sets are easy to manipulate, visualize and model.

Every guest Jon Stewart ever had on The Daily Show
Is the data set tidy?

YEAR	GoogleKnowlege_Occupation	Show	Group	Raw_Guest_List
2015	us president	7/21/15	Politician	Barack Obama
2015	actor	7/22/15	Acting	Jake Gyllenhaal
2015	Writer	7/23/15	Media	Ta-Nehisi Coates
2015	author	7/27/15	Media	David McCullough
2015	actor	7/28/15	Acting	Tom Cruise
2015	biographer	7/29/15	Media	Doris Kearns Goodwin
2015	director	7/30/15	Media	J. J. Abrams
2015	stand-up comedian	8/3/15	Comedy	Amy Schumer
2015	actor	8/4/15	Acting	Denis Leary
2015	comedian	8/5/15	Comedy	Louis C.K.

Child mortality rate (per 1,000 born), 1800-2015
Is the data set tidy?

Under five mortality	1800	1801	1802	1803	1804	1805	1806	1807	1808
Abkhazia	NA	NA	NA	NA	NA	NA	NA	NA	NA
Afghanistan	468.58	468.58	468.58	468.58	468.58	468.58	469.98	469.98	469.98
Akrotiri and Dhekelia	NA	NA	NA	NA	NA	NA	NA	NA	NA
Albania	375.20	375.20	375.20	375.20	375.20	375.20	375.20	375.20	375.20
Algeria	460.21	460.21	460.21	460.21	460.21	460.21	460.21	460.21	460.21
American Samoa	NA	NA	NA	NA	NA	NA	NA	NA	NA
Andorra	NA	NA	NA	NA	NA	NA	NA	NA	NA
Angola	485.68	485.68	485.68	485.68	485.68	485.68	485.68	485.68	485.68
Anguilla	NA	NA	NA	NA	NA	NA	NA	NA	NA
Antigua and Barbuda	473.60	469.77	465.97	462.20	458.47	454.76	451.08	447.43	443.82

Note: `NA` **denotes a missing value**

Commodity prices, 2003 vs. 2009

Is the data set tidy?

	bigmac2009	bread2009	rice2009	bigmac2003	bread2003	rice2003
Amsterdam	19	10	11	16	9	9
Athens	30	13	27	21	12	19
Auckland	19	19	13	19	19	9
Bangkok	45	43	27	50	42	25
Barcelona	21	17	8	22	19	10
Berlin	19	10	17	16	10	16
Bogota	58	36	21	93	48	16
Bratislava	62	23	25	54	20	28
Brussels	19	13	11	18	11	12
Bucharest	42	27	44	79	22	21

Tidy datasets are all alike but every messy dataset is messy in its own way.

— **Hadley Wickham**