Tidy!

Hadley Wickham
@hadleywickham
Chief Scientist, RStudio

 July 2015

Outline

- Warmups
- What is tidy data?
- Values in column names
- Multiple variables in one column
- Variable names in cells

Warmups

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	10	1996-06-21

What are the variables in this dataset?

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

disease counts

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

What are the variables in this dataset?

disease counts

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

pollution

city	particle size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

What are the variables in this dataset?

pollution

city	particle size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

How are these datasets similar?
How are they different?

x	y	z
1	1	2.5
2	3	4.6
1	3	1.7
4	4	7.2

2.5		1.7	
		4.6	
			7.2

How are these datasets similar?
How are they different?

x	y	z
1	1	2.5
2	3	4.6
1	3	1.7
4	4	7.2

key-value

2.5		1.7	
		4.6	
			7.2

matrix

How are these datasets similar?
How are they different?

spread

x	y	z
1	1	2.5
2	3	4.6
1	3	1.7
4	4	7.2

2.5		1.7	
		4.6	
			7.2

gather

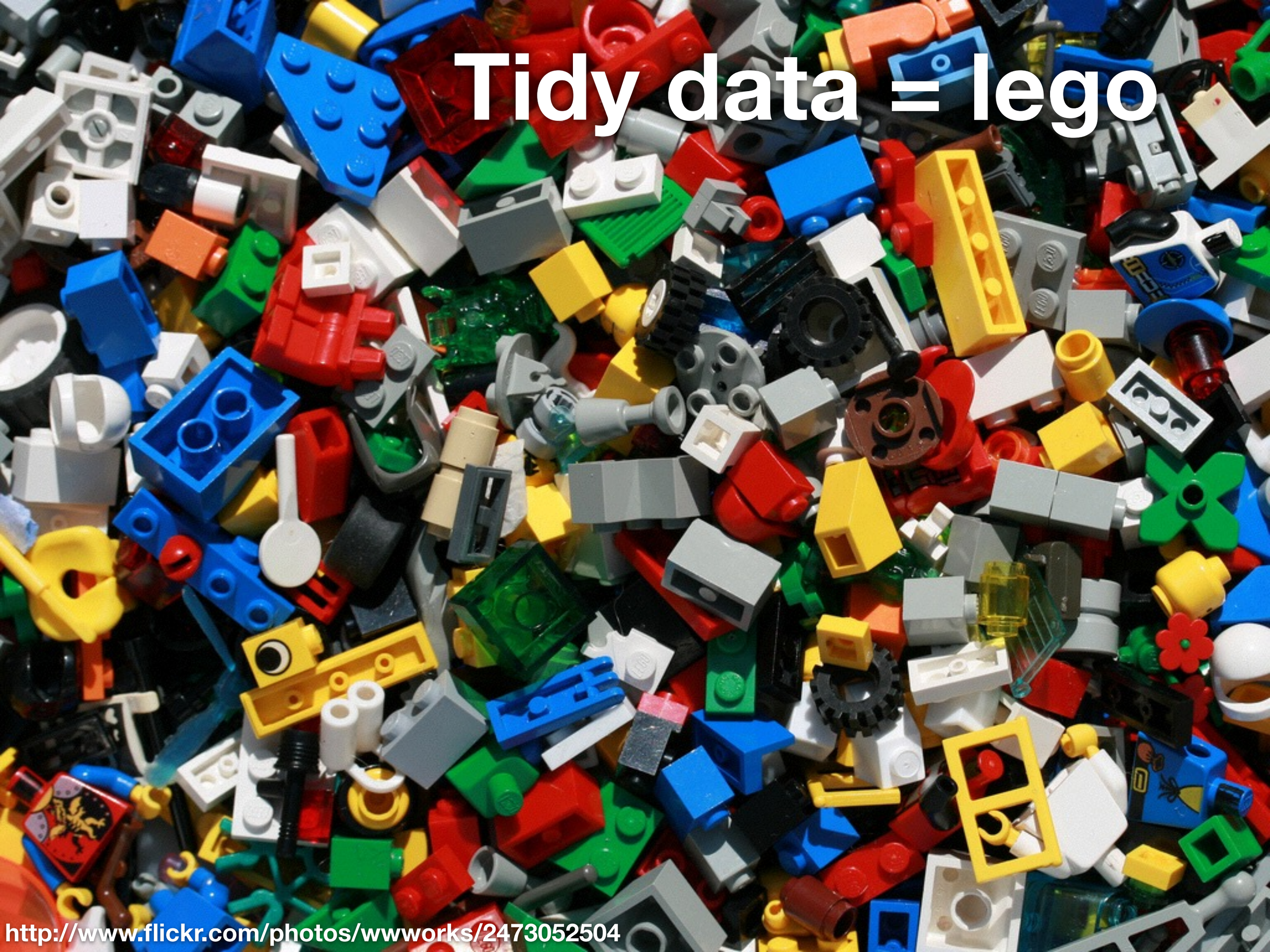
How are these datasets similar?
How are they different?

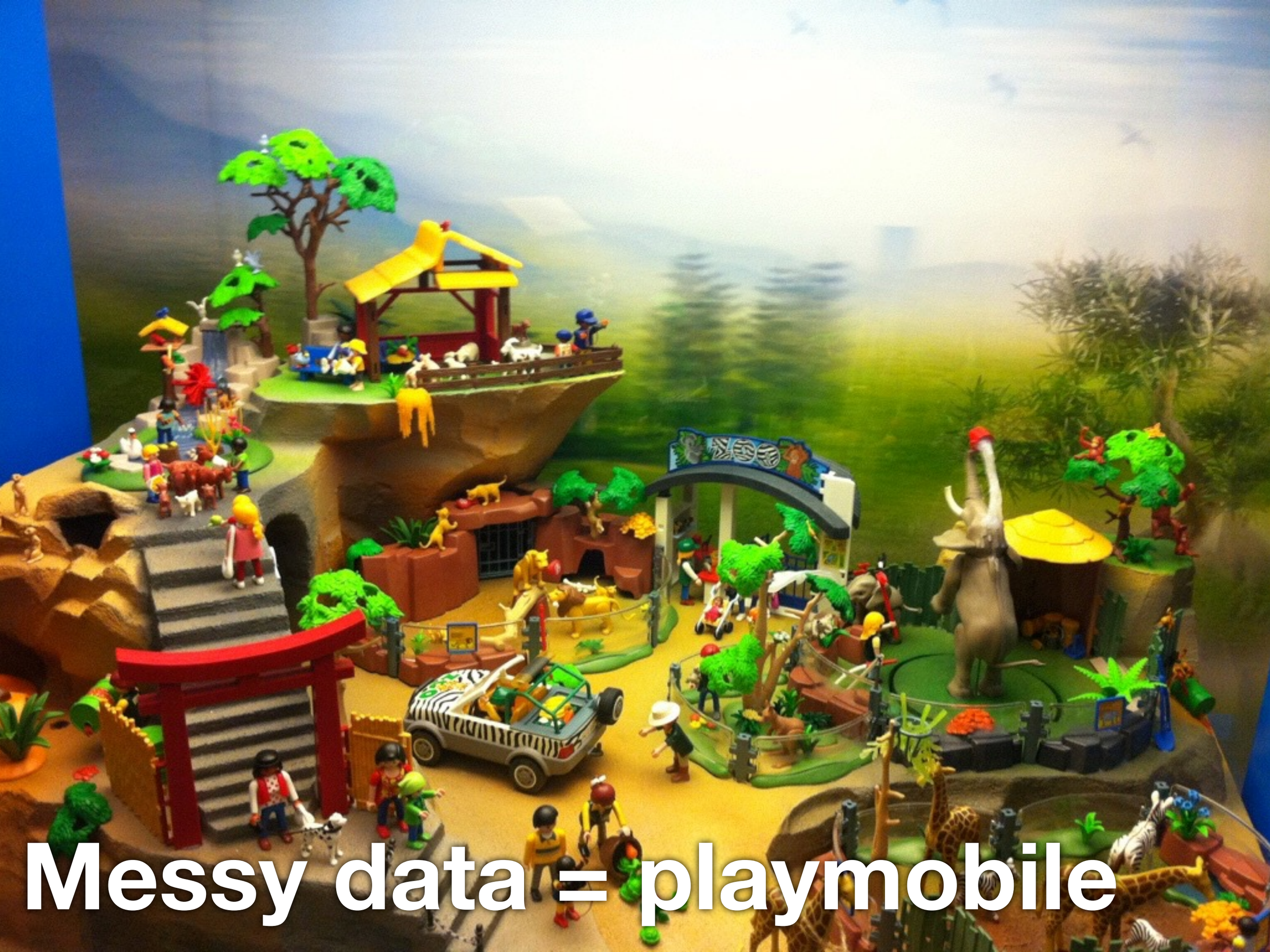
x	y	z
W	a	2.5
X	c	4.6
W	c	1.7
Z	d	7.2

	a	b	c	d
W	2.5		1.7	
X			4.6	
Y				
Z				7.2

Tidy data

Storage	Meaning
Rows	Observations
Columns	Variables
One data frame	One data set


Tidy data = lego
http://www.flickr.com/photos/wwworks/2473052504


Messy data = playmobile

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

disease counts

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

pollution

city	particle size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

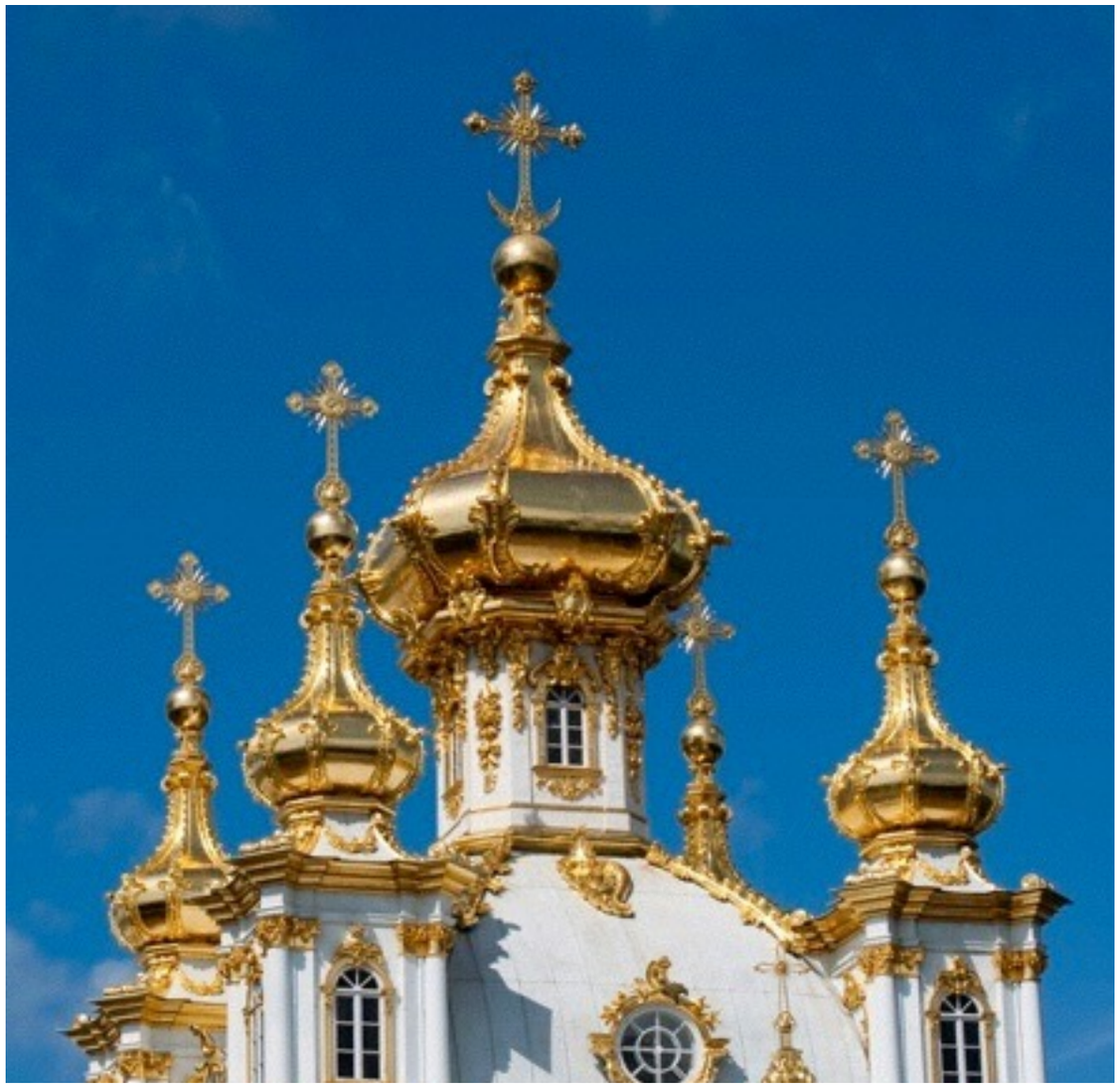
Tidying data

- A surprisingly small set of verbs are needed to turn many types of messy data into tidy data
- We’re going to use functions from the tidyr package
- (If you’ve done this in the past you might have used reshape2)

```
library(readr)
library(tidyr)
library(dplyr)
library(ggplot2)
```

Values in column names

Income distribution within U.S. religious groups



- Collected by Pew Research Center
- Examines the relationship between income and religion in the US
- i.e, which religions have the wealthiest adherents?

```
Source: local data frame [18 x 11]
                  religion <$10k $10-20k $20-30k $30-40k $40-50k $50-75k
1                 Agnostic    27      34      60      81      76     137
2                  Atheist    12      27      37      52      35      70
3                 Buddhist    27      21      30      34      33      58
4                 Catholic   418     617     732     670     638    1116
5       Don’t know/refused    15      14      15      11      10      35
6         Evangelical Prot   575     869    1064     982     881    1486
7                    Hindu     1       9       7       9      11      34
8  Historically Black Prot   228     244     236     238     197     223
9        Jehovah’s Witness    20      27      24      24      21      30
10                  Jewish    19      19      25      25      30      95
11            Mainline Prot   289     495     619     655     651    1107
12                   Mormon    29      40      48      51      56     112
13                   Muslim     6       7       9      10       9      23
14                 Orthodox    13      17      23      32      32      47
15          Other Christian     9       7      11      13      13      14
16             Other Faiths    20      33      40      46      49      63
17   Other World Religions     5       2       3       4       2       7
18             Unaffiliated   217     299     374     365     341     528
Variables not shown: $75-100k (int), $100-150k (int), >150k (int), Don't
   know/refused (int)
```

```
pew <- read_csv("tidy/pew.csv")
```

Your turn

What are the variables in this data set?
Discuss with your neighbours for one minute.

```
Source: local data frame [18 x 11]
                  religion <$10k $10-20k $20-30k $30-40k $40-50k $50-75k
1                Agnostic    27      34      60      81      76     137
2                 Atheist    12      27      37      52      35      70
3                Buddhist    27      21      30      34      33      58
5       Don’t know/refused    15      14      15      11      10      35
6         Evangelical Prot   575     869    1064     982     881    1486
7                    Hindu     1       9       7       9      11      34
Variables not shown: $75-100k (int), $100-150k (int), >150k (int), Don't
   know/refused (int)
```

```
# Fixing this problem is easy. We gather
# all the columns that aren't variables into
# a pair of variables: religion and n

pew %>% gather(income, n, -religion)
```

Gathering data

```
pew %>% gather(income, n, -religion)
```

```
head(raw)
     religion <$10k $10-20k $20-30k $30-40k $40-50k $50-75k $75-100k $100-150k >150k Don't know
1    Agnostic    27      34      60      81      76     137      122       109    84         96
2     Atheist    12      27      37      52      35      70       73        59    74         76
3    Buddhist    27      21      30      34      33      58       62        39    53         54
4    Catholic   418     617     732     670     638    1116      949       792   633       1489
5  Don't know    15      14      15      11      10      35       21        17    18        116
6 Evangelical   575     869    1064     982     881    1486      949       723   414       1529
```

Gathering data

data set to gather

```
pew %>% gather(income, n, -religion)
```

```
head(raw)
     religion <$10k $10-20k $20-30k $30-40k $40-50k $50-75k $75-100k $100-150k >150k Don't know
1    Agnostic    27      34      60      81      76     137      122       109    84         96
2     Atheist    12      27      37      52      35      70       73        59    74         76
3    Buddhist    27      21      30      34      33      58       62        39    53         54
4    Catholic   418     617     732     670     638    1116      949       792   633       1489
5  Don't know    15      14      15      11      10      35       21        17    18        116
6 Evangelical   575     869    1064     982     881    1486      949       723   414       1529
```

Gathering data

variable in columns

```
pew %>% gather(income, n, -religion)
```

```
head(raw)
     religion <$10k $10-20k $20-30k $30-40k $40-50k $50-75k $75-100k $100-150k >150k Don't know
1    Agnostic    27      34      60      81      76     137      122       109    84         96
2     Atheist    12      27      37      52      35      70       73        59    74         76
3    Buddhist    27      21      30      34      33      58       62        39    53         54
4    Catholic   418     617     732     670     638    1116      949       792   633       1489
5  Don't know    15      14      15      11      10      35       21        17    18        116
6 Evangelical   575     869    1064     982     881    1486      949       723   414       1529
```

Gathering data

variable in cells

```
pew %>% gather(income, n, -religion)
```

```
head(raw)
     religion <$10k $10-20k $20-30k $30-40k $40-50k $50-75k $75-100k $100-150k >150k Don't know
1    Agnostic    27      34      60      81      76     137      122       109    84         96
2     Atheist    12      27      37      52      35      70       73        59    74         76
3    Buddhist    27      21      30      34      33      58       62        39    53         54
4    Catholic   418     617     732     670     638    1116      949       792   633       1489
5  Don't know    15      14      15      11      10      35       21        17    18        116
6 Evangelical   575     869    1064     982     881    1486      949       723   414       1529
```

Gathering data

variables to gather

```
pew %>% gather(income, n, -religion)
```

```
head(raw)
     religion <$10k $10-20k $20-30k $30-40k $40-50k $50-75k $75-100k $100-150k >150k Don't know
1    Agnostic    27      34      60      81      76     137      122       109    84         96
2     Atheist    12      27      37      52      35      70       73        59    74         76
3    Buddhist    27      21      30      34      33      58       62        39    53         54
4    Catholic   418     617     732     670     638    1116      949       792   633       1489
5  Don't know    15      14      15      11      10      35       21        17    18        116
6 Evangelical   575     869    1064     982     881    1486      949       723   414       1529
```

```
# Ways of selecting variables:

# all except x
-x

# from a to z
a:z

# individually named
a, d, e, f
```

	religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k	$75-100k
1	Agnostic	27	34	60	81	76	137	122
2	Atheist	12	27	37	52	35	70	73
3	Buddhist	27	21	30	34	33	58	62
4	Catholic	418	617	732	670	638	1116	949
5	Don’t know	15	14	15	11	10	35	21
6	Evangelical	575	869	1064	982	881	1486	949

	religion	income	n
1	Agnostic	<$10k	27
2	Atheist	<$10k	12
3	Buddhist	<$10k	27
4	Catholic	<$10k	418
5	Don’t know	<$10k	15
6	Evangelical	<$10k	575

Every combination in the original data set is preserved

	religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k	$75-100k
1	Agnostic	27	34	60	81	76	137	122
2	Atheist	12	27	37	52	35	70	73
3	Buddhist	27	21	30	34	33	58	62
4	Catholic	418	617	732	670	638	1116	949
5	Don’t know	15	14	15	11	10	35	21
6	Evangelical	575	869	1064	982	881	1486	949

	religion	income	n
1	Agnostic	<$10k	27
2	Atheist	<$10k	12
3	Buddhist	<$10k	27
4	Catholic	<$10k	418
5	Don’t know	<$10k	15
6	Evangelical	<$10k	575

Every combination in the original data set is preserved

	religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k	$75-100k
1	Agnostic	27	34	60	81	76	137	122
2	Atheist	12	27	37	52	35	70	73
3	Buddhist	27	21	30	34	33	58	62
4	Catholic	418	617	732	670	638	1116	949
5	Don’t know	15	14	15	11	10	35	21
6	Evangelical	575	869	1064	982	881	1486	949

	religion	income	n
1	Agnostic	<$10k	27
2	Atheist	<$10k	12
3	Buddhist	<$10k	27
4	Catholic	<$10k	418
5	Don’t know	<$10k	15
6	Evangelical	<$10k	575

Every combination in the original data set is preserved

	religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k	$75-100k
1	Agnostic	27	34	60	81	76	137	122
2	Atheist	12	27	37	52	35	70	73
3	Buddhist	27	21	30	34	33	58	62
4	Catholic	418	617	732	670	638	1116	949
5	Don’t know	15	14	15	11	10	35	21
6	Evangelical	575	869	1064	982	881	1486	949

	religion	income	n
1	Agnostic	<$10k	27
2	Atheist	<$10k	12
3	Buddhist	<$10k	27
4	Catholic	<$10k	418
5	Don’t know	<$10k	15
6	Evangelical	<$10k	575

Every combination in the original data set is preserved

	religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k	$75-100k
1	Agnostic	27	34	60	81	76	137	122
2	Atheist	12	27	37	52	35	70	73
3	Buddhist	27	21	30	34	33	58	62
4	Catholic	418	617	732	670	638	1116	949
5	Don’t know	15	14	15	11	10	35	21
6	Evangelical	575	869	1064	982	881	1486	949

	religion	income	n
1	Agnostic	<$10k	27
2	Atheist	<$10k	12
3	Buddhist	<$10k	27
4	Catholic	<$10k	418
5	Don’t know	<$10k	15
6	Evangelical	<$10k	575

Every combination in the original data set is preserved

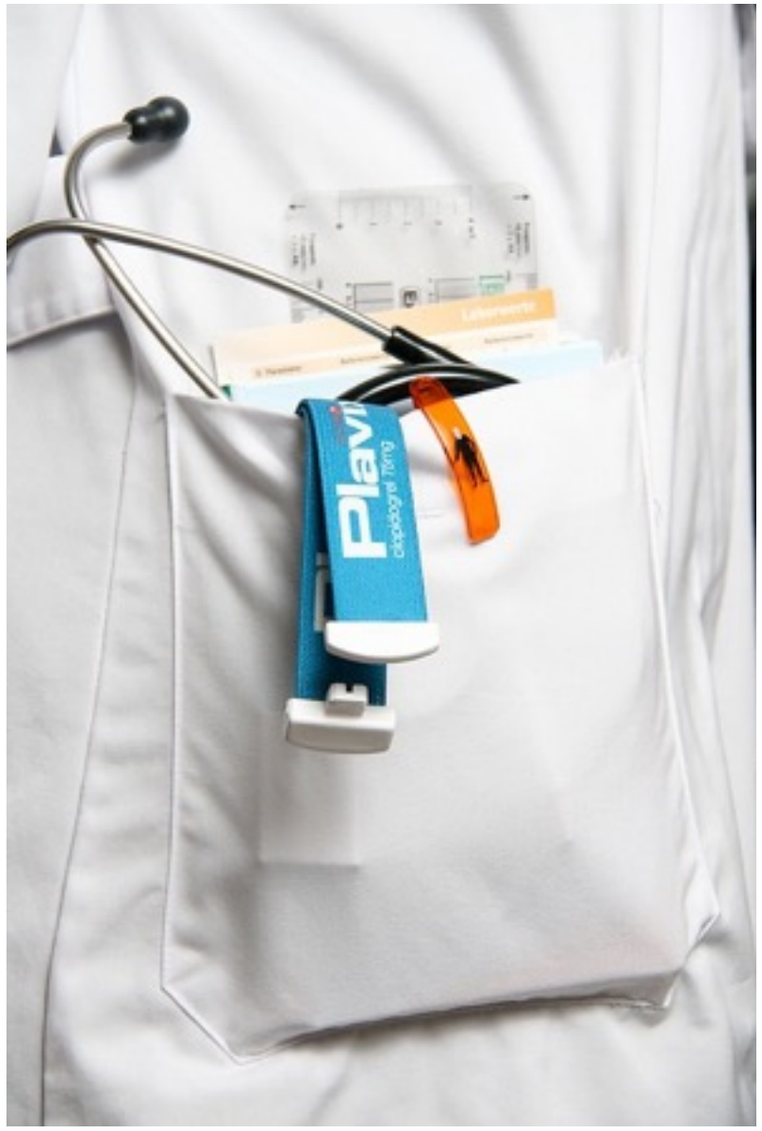
	religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k	$75-100k
1	Agnostic	27	34	60	81	76	137	122
2	Atheist	12	27	37	52	35	70	73
3	Buddhist	27	21	30	34	33	58	62
4	Catholic	418	617	732	670	638	1116	949
5	Don’t know	15	14	15	11	10	35	21
6	Evangelical	575	869	1064	982	881	1486	949

	religion	income	n
1	Agnostic	<$10k	27
2	Atheist	<$10k	12
3	Buddhist	<$10k	27
4	Catholic	<$10k	418
5	Don’t know	<$10k	15
6	Evangelical	<$10k	575

Every combination in the original data set is preserved

Multiple variables in one column

Tuberculosis



http://www.flickr.com/photos/diekatrin/4299075534/

- Collected by World Health Organization
- counts of TB cases by country, year, and demographic group

```
tb <- read_csv("tidy/tb.csv")
```

```
Source: local data frame [5,769 x 22]

   iso2 year m_04 m_514 m_014 m_1524 m_2534 m_3544 m_4554 m_5564 m_65 m_u f_04
1    AD 1989   NA    NA    NA     NA     NA     NA     NA     NA   NA  NA   NA
2    AD 1990   NA    NA    NA     NA     NA     NA     NA     NA   NA  NA   NA
3    AD 1991   NA    NA    NA     NA     NA     NA     NA     NA   NA  NA   NA
4    AD 1992   NA    NA    NA     NA     NA     NA     NA     NA   NA  NA   NA
5    AD 1993   NA    NA    NA     NA     NA     NA     NA     NA   NA  NA   NA
6    AD 1994   NA    NA    NA     NA     NA     NA     NA     NA   NA  NA   NA
7    AD 1996   NA    NA     0      0      0      4      1      0    0  NA   NA
8    AD 1997   NA    NA     0      0      1      2      2      1    6  NA   NA
9    AD 1998   NA    NA     0      0      0      1      0      0    0  NA   NA
10   AD 1999   NA    NA     0      0      0      1      1      0    0  NA   NA
11   AD 2000   NA    NA     0      0      1      0      0      0    0  NA   NA
12   AD 2001   NA    NA     0     NA     NA      2      1     NA   NA  NA   NA
13   AD 2002   NA    NA     0      0      0      1      0      0    0  NA   NA
14   AD 2003   NA    NA     0      0      0      1      2      0    0  NA   NA
15   AD 2004   NA    NA     0      0      0      1      1      0    0  NA   NA
16   AD 2005    0     0     0      0      1      1      0      0    0   0    0
17   AD 2006    0     0     0      1      1      2      0      1    1   0    0
18   AD 2007   NA    NA    NA     NA     NA     NA     NA     NA   NA  NA   NA
..  ...  ...  ...   ...    ...    ...    ...    ...    ...    ...  ... ...  ...
Variables not shown: f_514 (int), f_014 (int), f_1524 (int), f_2534 (int),
  f_3544 (int), f_4554 (int), f_5564 (int), f_65 (int), f_u (int)
```

Your turn

What are the variables in this data set?
Discuss with your neighbours for one minute.

```
Source: local data frame [5,769 x 22]

   iso2 year m_04 m_514 m_014 m_1524 m_2534 m_3544 m_4554 m_5564 m_65 m_u f_04
1    AD 1989   NA    NA    NA     NA     NA     NA     NA     NA   NA  NA   NA
2    AD 1990   NA    NA    NA     NA     NA     NA     NA     NA   NA  NA   NA
3    AD 1991   NA    NA    NA     NA     NA     NA     NA     NA   NA  NA   NA
4    AD 1992   NA    NA    NA     NA     NA     NA     NA     NA   NA  NA   NA
5    AD 1993   NA    NA    NA     NA     NA     NA     NA     NA   NA  NA   NA

..  ...  ...  ...   ...   ...    ...    ...    ...    ...    ...  ... ...  ...
Variables not shown: f_514 (int), f_014 (int), f_1524 (int), f_2534 (int),
  f_3544 (int), f_4554 (int), f_5564 (int), f_65 (int), f_u (int)
```

```
# Need to fix this in two steps. First start
# by gathering non-variable columns:

tb %>%
  gather(demographic, cases, m_04:f_u, na.rm = TRUE)
```

```r
# Next separate demographic into sex and age

tb %>%
  gather(demographic, cases, m_04:f_u, na.rm = TRUE) %>%
  separate(demographic, c("sex", "age"))
```

```r
# Finish with a little tidying up

tb %>%
  gather(demographic, cases, m_04:f_u, na.rm = TRUE) %>%
  separate(demographic, c("sex", "age")) %>%
  rename(country = iso2) %>%
  arrange(country, year, sex, age)
```

Your turn

`"tidy/population.csv"` contains matching population data. Read it into R and tidy in the same way.

Challenge: can you combine the two datasets to compute a rate? (n/population)

Variable names in cells

Weather data

http://www.flickr.com/photos/76708317@N02/7024035011

- Daily temperatures in Cuernavaca, Mexico for 2010
- `1 - 31`, days of month
- `tmax`, `tmin`, maximum and minimum temperatures

```
weather <- read_tsv("tidy/weather.txt", na = ".")
```

```
   year month element  1   2   3  4   5  6   7   8  9  10  11 12  13  14  15
1  2010     1    tmax NA  NA  NA NA  NA NA  NA  NA NA  NA  NA NA  NA  NA  NA
2  2010     1    tmin NA  NA  NA NA  NA NA  NA  NA NA  NA  NA NA  NA  NA  NA
3  2010     2    tmax NA 273 241 NA  NA NA  NA  NA NA  NA 297 NA  NA  NA  NA
4  2010     2    tmin NA 144 144 NA  NA NA  NA  NA NA  NA 134 NA  NA  NA  NA
5  2010     3    tmax NA  NA  NA NA 321 NA  NA  NA NA 345  NA NA  NA  NA  NA
6  2010     3    tmin NA  NA  NA NA 142 NA  NA  NA NA 168  NA NA  NA  NA  NA
7  2010     4    tmax NA  NA  NA NA  NA NA  NA  NA NA  NA  NA NA  NA  NA  NA
8  2010     4    tmin NA  NA  NA NA  NA NA  NA  NA NA  NA  NA NA  NA  NA  NA
9  2010     5    tmax NA  NA  NA NA  NA NA  NA  NA NA  NA  NA NA  NA  NA  NA
10 2010     5    tmin NA  NA  NA NA  NA NA  NA  NA NA  NA  NA NA  NA  NA  NA
11 2010     6    tmax NA  NA  NA NA  NA NA  NA  NA NA  NA  NA NA  NA  NA  NA
12 2010     6    tmin NA  NA  NA NA  NA NA  NA  NA NA  NA  NA NA  NA  NA  NA
13 2010     7    tmax NA  NA 286 NA  NA NA  NA  NA NA  NA  NA NA  NA 299  NA
14 2010     7    tmin NA  NA 175 NA  NA NA  NA  NA NA  NA  NA NA  NA 165  NA
15 2010     8    tmax NA  NA  NA NA 296 NA  NA 290 NA  NA  NA NA 298  NA  NA
16 2010     8    tmin NA  NA  NA NA 158 NA  NA 173 NA  NA  NA NA 165  NA  NA
17 2010    10    tmax NA  NA  NA NA 270 NA 281  NA NA  NA  NA NA  NA 295 287
18 2010    10    tmin NA  NA  NA NA 140 NA 129  NA NA  NA  NA NA  NA 130 105
..  ...   ...     ... .. ... ... .. ... .. ... ... .. ... ... .. ... ... ...
Variables not shown: 16 (int), 17 (int), 18 (lgl), 19 (lgl), 20 (lgl), 21
  (lgl), 22 (lgl), 23 (int), 24 (lgl), 25 (int), 26 (int), 27 (int), 28 (int),
  29 (int), 30 (int), 31 (int)
```

Your turn

What are the variables in this data set?
Discuss with your neighbours for one minute.

```
   year month element  1   2   3  4   5  6   7   8  9  10  11 12  13  14  15
1  2010     1    tmax NA  NA  NA NA  NA NA  NA  NA NA  NA  NA NA  NA  NA  NA
2  2010     1    tmin NA  NA  NA NA  NA NA  NA  NA NA  NA  NA NA  NA  NA  NA
3  2010     2    tmax NA 273 241 NA  NA NA  NA  NA NA  NA 297 NA  NA  NA  NA
4  2010     2    tmin NA 144 144 NA  NA NA  NA  NA NA  NA 134 NA  NA  NA  NA
5  2010     3    tmax NA  NA  NA NA 321 NA  NA  NA NA 345  NA NA  NA  NA  NA
6  2010     3    tmin NA  NA  NA NA 142 NA  NA  NA NA 168  NA NA  NA  NA  NA
7  2010     4    tmax NA  NA  NA NA  NA NA  NA  NA NA  NA  NA NA  NA  NA  NA
..  ...   ...     ... .. ... ... .. ... .. ... ... .. ... ... .. ... ... ...
Variables not shown: 16 (int), 17 (int), 18 (lgl), 19 (lgl), 20 (lgl), 21
  (lgl), 22 (lgl), 23 (int), 24 (lgl), 25 (int), 26 (int), 27 (int), 28 (int),
  29 (int), 30 (int), 31 (int)
```

```
weather %>%
  gather(day, value, `1`:`31`, na.rm = TRUE) %>%

# Source: local data
#
#   year month element day value
# 1  2010    12    tmax   1   299
# 2  2010    12    tmin   1   138
# 3  2010     2    tmax   2   273
# 4  2010     2    tmin   2   144
# 5  2010    11    tmax   2   313
# 6  2010    11    tmin   2   163
# 7  2010     2    tmax   3   241
#..  ...   ...     ... ...   ...

# Which columns aren't variables?
```

1 isn't a valid variable name (it doesn't start with a letter) so we need to use backticks

```
weather %>%
  gather(day, value, `1`:`31`, na.rm = TRUE) %>%
  spread(element, value)

# Source: local data frame [33 x 5]
#
#    year month day tmax tmin
# 1   2010     1  30  278  145
# 2   2010     2   2  273  144
# 3   2010     2   3  241  144
# 4   2010     2  11  297  134
# 5   2010     2  23  299  107
# 6   2010     3   5  321  142
# 7   2010     3  10  345  168
# 8   2010     3  16  311  176
# ..   ...    ... ...  ...  ...
```

How are these datasets similar?
How are they different?

spread

x	y	z
1	1	2.5
2	3	4.6
1	3	1.7
4	5	7.2

2.5		1.7	
		4.6	
			7.2

gather

Every combination of values is retained

year	month	day	element	value
2010	1	30	tmax	278
2010	1	30	tmin	145
2010	2	2	tmax	273
2010	2	2	tmin	144
2010	2	3	tmax	241
2010	2	3	tmin	144

year	month	day	tmax	tmin
2010	1	30	278	145
2010	2	2	273	144
2010	2	3	241	144
2010	2	11	297	134
2010	2	23	299	107
2010	3	5	321	142

Every combination of values is retained

year	month	day	element	value
2010	1	30	tmax	278
2010	1	30	tmin	145
2010	2	2	tmax	273
2010	2	2	tmin	144
2010	2	3	tmax	241
2010	2	3	tmin	144

year	month	day	tmax	tmin
2010	1	30	278	145
2010	2	2	273	144
2010	2	3	241	144
2010	2	11	297	134
2010	2	23	299	107
2010	3	5	321	142

Every combination of values is retained

year	month	day	element	value
2010	1	30	tmax	278
2010	1	30	tmin	145
2010	2	2	tmax	273
2010	2	2	tmin	144
2010	2	3	tmax	241
2010	2	3	tmin	144

year	month	day	tmax	tmin
2010	1	30	278	145
2010	2	2	273	144
2010	2	3	241	144
2010	2	11	297	134
2010	2	23	299	107
2010	3	5	321	142

Every combination of values is retained

year	month	day	element	value
2010	1	30	tmax	278
2010	1	30	tmin	145
2010	2	2	tmax	273
2010	2	2	tmin	144
2010	2	3	tmax	241
2010	2	3	tmin	144

year	month	day	tmax	tmin
2010	1	30	278	145
2010	2	2	273	144
2010	2	3	241	144
2010	2	11	297	134
2010	2	23	299	107
2010	3	5	321	142

Every combination of values is retained

year	month	day	element	value
2010	1	30	tmax	278
2010	1	30	tmin	145
2010	2	2	tmax	273
2010	2	2	tmin	144
2010	2	3	tmax	241
2010	2	3	tmin	144

year	month	day	tmax	tmin
2010	1	30	278	145
2010	2	2	273	144
2010	2	3	241	144
2010	2	11	297	134
2010	2	23	299	107
2010	3	5	321	142

Every combination of values is retained

year	month	day	element	value
2010	1	30	tmax	278
2010	1	30	tmin	145
2010	2	2	tmax	273
2010	2	2	tmin	144
2010	2	3	tmax	241
2010	2	3	tmin	144

year	month	day	tmax	tmin
2010	1	30	278	145
2010	2	2	273	144
2010	2	3	241	144
2010	2	11	297	134
2010	2	23	299	107
2010	3	5	321	142

titanic2

Characteristics and fate of passengers on the Titanic.

```
titanic2 <- read_csv("tidy/titanic2.csv")
```

```
head(titanic2)
# Source: local data frame [32 x 5]
#
#    class   age     fate gender   n
# 1    1st adult perished   male 118
# 2    1st adult survived   male  57
# 3    1st child perished   male   0
# 4    1st child survived   male   5
# 5    2nd adult perished   male 154
# 6    2nd adult survived   male  14
# 7    2nd child perished   male   0
# 8    2nd child survived   male  11
# ..    ...    ...       ...     ... ...
```

Your turn

Make a tidy version of this data.

```
titanic2 <- read_csv("tidy/titanic2.csv")
```

Then compute the survival rate for each class, gender and age.

```
titanic2 %>%
  gather(gender, n, male:female) %>%
  spread(fate, n) %>%
  mutate(rate = survived / (survived + perished))
```

Where next

Tidy data paper
www.jstatsoft.org/v59/i10/

Manipulatr mailing list
https://groups.google.com/group/manipulatr